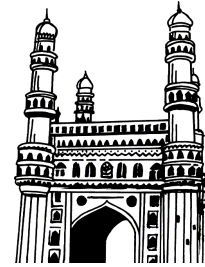


Rahul's ✓
Topper's Voice



BBA

III Year V Sem

Latest 2024 Edition

BUSINESS ANALYTICS

- ☞ Study Manual
- ☞ Important Questions
- ☞ Short Questions and Answers
- ☞ Choose the correct Answers
- ☞ Fill in the blanks
- ☞ One Mark Answers
- ☞ Solved Model Papers

- by -

WELL EXPERIENCED LECTURER

₹. 169/-



Rahul Publications™

Hyderabad. Cell : 9391018098, 9505799122

All disputes are subjects to Hyderabad Jurisdiction only

BBA

III Year V Sem

BUSINESS ANALYTICS

Inspite of many efforts taken to present this book without errors, some errors might have crept in. Therefore we do not take any legal responsibility for such errors and omissions. However, if they are brought to our notice, they will be corrected in the next edition.

© No part of this publications should be reporduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher

Price ₹. 169-00

Sole Distributors :

Cell : 9391018098, 9505799122

VASU BOOK CENTRE

Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.

Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.

Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

BUSINESS ANALYTICS

C O N T E N T S

STUDY MANUAL

Important Questions	IV - VIII
Unit - I	1 - 30
Unit - II	31 - 80
Unit - III	81 - 122
Unit - IV	123 - 152
Unit - V	153 - 178

SOLVED MODEL PAPERS

Model Paper - I	179 - 180
Model Paper - II	181 - 182
Model Paper - III	183 - 184

SYLLABUS

UNIT - I

INTRODUCTION TO BUSINESS ANALYTICS:

Definition of Business Analytics, Categories of Business Analytical methods and models, Business Analytics in practice, Big Data - Overview of using Data, Types of Data- Business decision modeling.

UNIT - II

DESCRIPTIVE ANALYTICS:

Overview of Description Statistics (Central Tendency, Variability), Data Visualization -Definition, Visualization Techniques – Tables, Cross Tabulations, charts, Data Dashboards using Advanced Ms-Excel or SPSS.

UNIT - III

PREDICTIVE ANALYTICS:

Trend Lines, Regression Analysis – Linear & Multiple, Predictive modeling, forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Data mining and business intelligence, Data mining for business, Classification, Association, Cause Effect Modeling.

UNIT - IV

PRESCRIPTIVE ANALYTICS:

Overview of Linear Optimization, Non Linear Programming Integer Optimization, Cutting Plane algorithm and other methods, Decision Analysis – Risk and uncertainty methods - Text analytics Web analytics.

UNIT - V

Unit – V: PROGRAMMING USING R:

R Environment, R packages, Reading and Writing data in R, R functions, Control Statements, Frames and Subsets, Managing and Manipulating data in R.

Contents

UNIT - I

Topic	Page No.
1.1 Introduction to Business Analytics	1
1.1.1 Definition of Business Analytics	1
1.2 Categories of Business Analytical Methods and Models	2
1.3 Business Analytics in Practice	8
1.4 Big Data - Overview of using Data	11
1.5 Types of Data	17
1.6 Business Decision Modeling	19
➤ Short Questions and Answers	24 - 27
➤ Choose the Correct Answers	28 - 28
➤ Fill in the Blanks	29 - 29
➤ One Mark Answers	30 - 30

UNIT - II

2.1 Overview of Description Statistics (Central Tendency, Variability)	31
2.2 Data Visualization - Definition, Visualization Techniques	45
2.3 Tables, Cross Tabulations, charts, Data Dashboards using Advanced Ms-Excel or SPSS.	51
➤ Short Questions and Answers	73 - 76
➤ Choose the Correct Answers	77 - 78
➤ Fill in the Blanks	79 - 79
➤ One Mark Answers	80 - 80

UNIT - III

3.1 Introduction	81
3.1.1 Trend Lines	83
3.2 Regression Analysis – Linear and Multiple	86
3.3 Predictive Modelling	88
3.4 Forecasting Techniques	90
3.5 Data Mining-Definition-Approaches in Data Mining	95

Topic	Page No.
3.6 Data Exploration & Reduction	102
3.7 Data Mining and Business Intelligence	108
3.8 Data Mining for Business	109
3.9 Classification, Association	110
3.10 Cause Effect Modeling	111
➤ Short Questions and Answers	114 - 119
➤ Choose the Correct Answers	120 - 120
➤ Fill in the Blanks	121 - 121
➤ One Mark Answers	122 - 122
UNIT - IV	
4.1 Prescriptive Analytics	123
4.1.1 Introduction	123
4.2 Overview of Linear Optimization	124
4.3 Non Linear Programming Integer Optimization	133
4.4 Cutting Plane algorithm and other methods	135
4.5 Decision Analysis	136
4.6 Risk and Uncertainty Methods	138
4.7 Text Analytics Web Analytics	140
➤ Short Questions and Answers	146 - 148
➤ Choose the Correct Answers	149 - 150
➤ Fill in the Blanks	151 - 151
➤ One Mark Answers	152 - 152
UNIT - V	
5.1 R Environment	153
5.2 R Packages	154
5.3 Reading and Writing data in R	157
5.4 R functions	158
5.5 Control Statements	160
5.6 Frames and Subsets	165

Topic	Page No.
5.7 Managing and Manipulating data in R	167
➤ Short Questions and Answers	172 - 175
➤ Choose the Correct Answers	176 - 176
➤ Fill in the Blanks	177 - 177
➤ One Mark Answers	178 - 178

Important Questions

UNIT - I

1. What is business analytics? Explain the importance of business analytics.

Ans :

Refer Unit-I, Q.No. 1

2. Discuss various types of business analytics

Ans :

Refer Unit-I, Q.No. 4

3. List and explain various business analytics techniques used in Business.

Ans :

Refer Unit-I, Q.No. 7

4. What are the applications of business analytics

Ans :

Refer Unit-I, Q.No. 8

5. What is Big data? Discuss various types of Big Data.

Ans :

Refer Unit-I, Q.No. 9

6. What are the advantages of Big data.

Ans :

Refer Unit-I, Q.No. 11

7. Discuss various types of data.

Ans :

Refer Unit-I, Q.No. 12

8. What is decision modeling? What are the elements of decision modeling.

Ans :

Refer Unit-I, Q.No. 14

UNIT - II

1. What is descript analytics? Give any two examples.

Ans :

Refer Unit-II, Q.No. 1

2. What is Frequency distribution? What are the types of frequency distribution?

Ans :

Refer Unit-II, Q.No. 5

3. Explain Central tendency in point of distributions.

Ans :

Refer Unit-II, Q.No. 7

4. What is Data Visualization? What are the benefits of data visualization.

Ans :

Refer Unit-II, Q.No. 9

5. What are the applications of data visualization.

Ans :

Refer Unit-II, Q.No. 11

6. List and explain various data visualization techniques

Ans :

Refer Unit-II, Q.No. 13

7. What is mean by table? How can we create table in Excel.

Ans :

Refer Unit-II, Q.No. 14

8. How to create cross tab in Excel?

Ans :

Refer Unit-II, Q.No. 16

9. Explain the process of inserting a chart in Excel.

Ans :

Refer Unit-II, Q.No. 18

10. What are the uses of Data Dashboard?

Ans :

Refer Unit-II, Q.No. 20

UNIT - III

1. What is mean by Predictive analytics?

Ans :

Refer Unit-III, Q.No. 1

2. What is a Trend Line? Explain the need of trend lines?

Ans :

Refer Unit-III, Q.No. 4

3. What is Regression Analysis?

Ans :

Refer Unit-III, Q.No. 7

4. What is Predictive Modelling?

Ans :

Refer Unit-III, Q.No. 12

5. Discuss various forecasting techniques.

Ans :

Refer Unit-III, Q.No. 17

6. What is Data mining? Give a brief overview on data mining.

Ans :

Refer Unit-III, Q.No. 18

7. What is Data Exploration?

Ans :

Refer Unit-III, Q.No. 22

8. What is business intelligence?

Ans :

Refer Unit-III, Q.No. 26

9. How to use Data mining for business analytics.

Ans :

Refer Unit-III, Q.No. 28

10. What is classification in data mining ?

Ans :

Refer Unit-III, Q.No. 29

11. Discuss in detail about cause and effect modelling.

Ans :

Refer Unit-III, Q.No. 32

UNIT - IV

1. What is prescriptive analytics? And list out the benefits of prescriptive analytics.

Ans :

Refer Unit-IV, Q.No. 1

2. What is Linear optimization? What are the applications of linear optimization.

Ans :

Refer Unit-IV, Q.No. 3

3. What is mean by non linear optimization?

Ans :

Refer Unit-IV, Q.No. 7

4. What is Cutting plane algorithm?

Ans :

Refer Unit-IV, Q.No. 9

5. What is decision there analysis ?

Ans :

Refer Unit-IV, Q.No. 10

6. What is Risk Analysis?

Ans :

Refer Unit-IV, Q.No. 11

7. What are the steps involved in risk analysis.

Ans :

Refer Unit-IV, Q.No. 13

8. What is text analytics? What are the benefits of text analytics.

Ans :

Refer Unit-IV, Q.No. 16

UNIT - V

1. What is R Programming ? What re the features of R ?

Ans :

Refer Unit-V, Q.No. 1

2. Explain in detail about R Packages.

Ans :

Refer Unit-V, Q.No. 3

3. What are the functions are useful for reading data into R and writing data to files.

Ans :

Refer Unit-V, Q.No. 5

4. What is mean by function ? What are the components of function?

Ans :

Refer Unit-V, Q.No. 8

5. What are the conditional statements supported by R.

Ans :

Refer Unit-V, Q.No. 11

6. What is mean by frame ? How can we create a frame in R?

Ans :

Refer Unit-V, Q.No. 14

7. How do you manipulate data in R?

Ans :

Refer Unit-V, Q.No. 17

UNIT I

Introduction to Business Analytics:

Definition of Business Analytics, Categories of Business Analytical methods and models, Business Analytics in practice, Big Data - Overview of using Data, Types of Data- Business decision modeling.

1.1 INTRODUCTION TO BUSINESS ANALYTICS

1.1.1 Definition of Business Analytics

Q1. What is business analytics? Explain the importance of business analytics.

Ans : (Imp.)

Business analytics is the process of gathering and processing all of your business data, and applying statistical models and iterative methodologies to translate that data into business insights. Most importantly, you need to be able to translate that data into what your customers want. And attempting to guess what your customers want is not enough anymore.

Business analytics combines processes, skills, and technologies to collect, analyze and present historical performance of the business, based on data, with the goal of driving business planning. Business analytics essentially uses large amounts of historical business data for the purpose of providing insights to evaluate a business. Business analytics reveal previously unknown insights or identify unanticipated issues to generate new business value. Business analytics can be leveraged to create new processes for solving business challenges and increase efficiency, productivity, and revenue.

Importance of Business Analytics

- Business analytics can transform raw data into more valuable inputs to leverage this information in decision making.
- With Business Analytics tools, we can have a more profound understanding of primary and secondary data emerging from their activities. This helps businesses refine their procedures further and be more productive.

- To stay competitive, companies need to be ahead of their peers and have all the latest toolsets to assist their decision making in improving efficiency as well as generating more profits.

Q2. What are the advantages and disadvantages of Business Analytics.

Ans :

Following are the advantages of Business analytics

1. Make better business decisions

Business analysts understand the organization's goals and use analytics to guide data-driven business decisions.

The past data of the company, current market situation, and product performance are leveraged to predict future trends and accordingly design strategies.

2. Monitor progress and performance

Business analytics can be used to track the progress of the organization over the years. It can also identify the performance of a product or a strategy in the market.

Based on these reports, it can be deduced what is working well for the organization and what isn't. As a result, updated decisions and methodologies can be implemented to improve the stats.

3. Reduce risks:

One main advantage of business analytics is its ability to mitigate risks. It helps in tracking the mistakes made by the organization in the past and understanding the factors that led to their occurrence.

With this knowledge, analysis is done to predict the probability of the reoccurrence of similar risks in the upcoming future, and therefore, the corresponding measures can be taken to prevent the same.

4. Enhance customer experience

All successful businesses have figured out the secret to success – making their customers happy! Organizations today, identify their customer base, understand their needs and behaviors and correspondingly cater to them.

This is possible because of the statistical tools and models used in business analytics.

Disadvantages

1. Lack of Commitment

The business Analytics process can be extremely costly as well as time-consuming. Although the solutions can be easily achieved, the time and cost factors leave people feeling disinterested and therefore less trusting.

This eventually leads to the complete failure of the business.

2. Low-Quality Data

Organizations have a lot of data. But the real question is how much of this data is actually correct and accessible.

Having poorly constructed, heavily complicated or insufficient data is a huge limitation and can hinder the business analytics processes.

3. Privacy Concerns

Companies collect customer data to analyze it and make better business decisions. But this can lead to a breach of the customer's privacy.

There have been instances when one company shares its collected user data with another company for mutual benefit.

This data can be used against a particular user in any way possible.

Therefore, it is essential for organizations to collect only vital information and work on maintaining the security and confidentiality of the data collected.

Q3. How business analytics benefits the organizations?

Ans :

Business analytics' specific purpose is bringing information and expertise to help guide business decisions and create an "Information Advantage." With business analytics, you can:

- Turn big data into insights
- Build statistical models to make projections about a business
- Pitch ideas to optimize performance
- Advise management around decision-making
- Leverage data to influence business outcomes

Business analytics enable you to choose opportunities with the highest propensity for success, calculate the strategy that would deliver the best return for the business, and it can help prepare your company for any upcoming changes or market trends that are on the horizon.

Business analytics can help you understand the environment you are in, how you can become more competitive, streamline decision making and as a result, increase revenues and decrease risk.

Previously, only companies with high levels of risk, like insurance companies would apply business analytics. However, since data has revolutionized the world, business analytics has become a necessity for all businesses to remain competitive and succeed in a digital world.

1.2 CATEGORIES OF BUSINESS ANALYTICAL METHODS AND MODELS

Q4. Discuss various types of business analytics

Ans :

(Imp.)

Following are the types of business analytics

1. Descriptive Analytics

It analyses historical data to determine the response of a unit over a set of given variables. It tracks key performance indicators (KPIs) for a better understanding of the present state of a business.

It involves the following steps:

- Deciding which business metrics will effectively evaluate performance against objectives
- Identifying required data as per the current business state
- Collecting and preparing data using various processes like depublication, transformation, and cleansing.
- Analyzing data for patterns to measure performance
- Presenting data in charts and graphs to make it understandable for non-analytics experts

Examples of Descriptive Analytics

Summarizing past events, exchange of data, and social media usage

Reporting general trends

2. Diagnostic Analytics

Diagnostic Analytics is one of those business analytics types that help understand why things happened in the past. Using drill-downs, data mining, data discovery, and correlations, you can comprehend the driving factors.

This advanced analytics method is usually employed as a preceding step of Descriptive Analytics to find the reasoning behind certain results in finance, marketing, cybersecurity, and more.

Examples of Diagnostic Analytics

- Examining market demand
- Identifying technical issues
- Explaining customer behavior
- Improving organization culture

3. Predictive Analytics

It considers historical data trends for determining the probability of particular future outcomes. It uses several techniques like data mining, machine learning algorithms, and statistical modeling to forecast the likelihood of events.

Predictive analytics helps improve business areas, including customer service, efficiency, fraud detection and prevention, and risk management. It allows you to grow the most profitable customers, improve the operations of businesses, and determine customer responses and cross-sell opportunities.

Examples of Predictive Analytics

- Predicting customer preferences
- Detection of employee intentions
- Recommending products
- Predicting staff and resources

4. Prescriptive Analytics

Prescriptive analytics generates recommendations to handle similar future situations relying on past performances. It employs several tools, statistics, and ML algorithms for the available internal data and external data.

It gives you insights into what may happen, when, and why.

Examples of Prescriptive Analytics

- Tracking fluctuating manufacturing prices
- Improving equipment management
- Suggest the best course of action
- Price modeling
- Evaluating rates of readmission
- Identifying testing

5. Cognitive Analytics

Combining Artificial Intelligence and Data Analytics, Cognitive Analytics is one of the newest types of business analytics. It looks at the available data in the knowledge base and discovers the best solutions for the questions posed.

Cognitive analytics covers multiple analytical techniques to analyze large data sets and monitor customer behavior patterns and emerging trends.

Examples of Cognitive Analytics

- Tapping unstructured data sources such as images, text documents, emails, and social posts.

Q5. What are the statistical methods used in business analytics?

Ans :

Some popular statistical methods used in Business Analytics

(i) Sampling

It is a process of taking a small set of observations (sample) from a large population. It is a common tool used in any type of data analysis.

Some of the sampling methods are random sampling, stratified sampling and cluster sampling. Sometime due to time constraints or it could be similarities in data – we could not analysis the whole data. So in such circumstance, we can go for sampling.

(ii) Correlation Analysis

It is used to study the closeness of the relationship between two or more variables i.e. the degree to which the variables are associated with each other. Suppose in a manufacturing firm, they want the relation between–

- Demand & supply of commodities.
- Production volume & the efficiency of machinery equipment.

(iii) Regression Analysis

It is a commendable statistical technique used in business analytics. It helps to predict the value of future outcomes by using the past data.

This method helps in forecasting the data and Time-series Analysis. The different types of regression analysis are

- Linear Regression
- Multiple Regression
- Logistic Regression
- Poisson Regression...

(iv) Graphical Analysis

Here, the data are presented in the form of graphs or diagrams. When we presented data through diagrams and graphs – it looks more convincing & appealing.

Thus provide the meaningful outlook of a data. Some of the popular graphical tools used are:

- Histogram
- Bar chart
- Pareto chart
- Scatter plot...

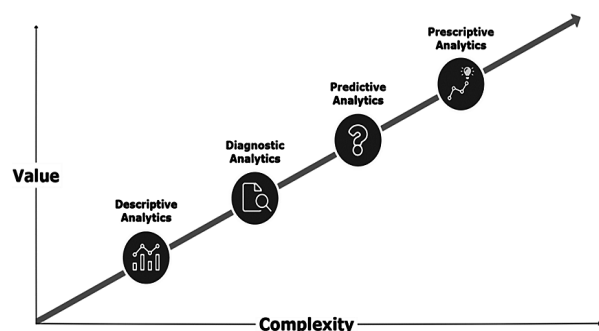
Q6. What are the analytical methods used in business analytics ?

Ans :

The four main analytical models organisations can deploy are:

1. descriptive
2. diagnostic
3. predictive
4. prescriptive.

As you move from descriptive to prescriptive analytics, each model offers increasing value to an organisation. But, at the same time, they increase in complexity.



1. Descriptive Analytics

Descriptive analytics answer the question: What happened.

This is the most common type of analytics found in business. It generally uses historical data from a

For example:

- How many sales did we make in the last week/day/hour?
- Which customers required the most help from our customer service team?
- How many people viewed our website?
- Which product had the most defects?single internal source to pinpoint when an event occurred.

Descriptive analytics are often displayed on dashboards and in reports, which are convenient ways to consume data and inform decisions.

Descriptive analytics account for most of the statistics we use, including basic aggregation (e.g. count or sum of values filtered from a column or data), averages, and percentage changes.

2. Diagnostic Analytics

Diagnostic analytics help us to answer the next question: Why did it happen.

To do this, analysts dive deeper into an organization's historical data, combining multiple sources in search of patterns, trends, and correlations.

Diagnostics analytics used for following reasons

- **Identify anomalies:** Analysts use the results from descriptive analysis to identify areas that need further investigation and raise questions that can't be answered by simply looking at the data.

For example: Why have sales increased in a region that had no change in marketing?

- **Drill down into data:** To explain anomalies, analysts must find patterns outside existing data sets to identify correlations.

They might need to use techniques such as data mining, and use data from external sources.

➤ **Determine causal relationships:**

Having identified anomalies and searched for patterns that could be correlated, analysts use more advanced statistical techniques to determine whether these are related.

Traditionally, data analysts performed diagnostic analytics manually, but as data volume, variety, and velocity increase, fully manual analysis is no longer feasible. Instead, modern diagnostic analytics solutions employ machine-learning techniques to augment the analyst's skills.

Computers can process vast amounts of data and recognise patterns, detect anomalies, and expose 'unusual' events, and they can apply analytical techniques from a portfolio of algorithms to identify drivers of change and determine causation.

3. Predictive Analytics

As an organisation increases its analytical maturity and embarks on predictive analytics, it shifts its focus from understanding historical events to creating insights about a current or future state. Predictive analytics is at the intersection of classical statistical analysis and modern artificial intelligence (AI) techniques. It tries to answer the question: What will happen next

It's impossible to predict exactly what will happen in the future, but by employing predictive analytics, organisations identify the likelihood of possible outcomes and can increase the chance of taking the best course of action. We see predictive analytics used in many sectors.

For Example

Aerospace – Predictive analytics are used to predict the effect of specific maintenance operations on aircraft reliability, fuel use, availability, and uptime.

Financial services – Predictive analytics are used to develop credit-risk models and forecast financial market trends.

Manufacturing – Predictive analytics are used to predict the location and rate of machine failures, and to optimise ordering and delivery of raw materials based on projected future demands.

Online retail – Systems monitor customer behaviour, and predictive models determine whether providing additional product information or incentives will increase the likelihood of a sale.

Simple predictive models can be created using tools such as Excel or Tableau. As these models start to accommodate more variables, with more complex relationships, these analytics become the responsibility of data scientists.

4. Prescriptive Analytics

Prescriptive analytics is the most complex type of analytics. It combines internal data, external sources, and machine-learning techniques to provide the most effective outcomes.

In prescriptive analytics, a decision-making process is applied to descriptive and predictive models to find the combinations of existing conditions and possible decisions that are likely to have the most effect in the future.

This process is both complex and resource intensive but, when done well, can provide immense value to an organisation.

Applications of prescriptive analytics include:

- risk management
- improving healthcare
- guided marketing, selling and pricing

Q7. List and explain various business analytics techniques used in Business.

Ans : (Imp.)

The top 10 different business analysis techniques are listed below.

1. SWOT Analysis

The four components of SWOT are as follows:

Strength: The elements of a project or business that offer it an edge over rivals.

Weakness: When compared to the competitors or other projects, the business's characteristics disadvantage the project or company.

Opportunities: The project or business could take advantage of environmental factors.

Threats: Environment-related factors that could obstruct the project or enterprise.

It is a complete analysis carried out by a business analyst, considering internal and external aspects, such as strengths and weaknesses, threats, and opportunities.

For a business analyst, a SWOT analysis is a four-quadrant analysis where the data are the responses for each quadrant. Each of the questions in the quadrants is answered by techniques used by the business analyst.

2. Business Process Modeling

Process improvement is the focus of business process modeling and is one of the most important business analysis investigation techniques.

Although it is a legacy process, it is frequently used as a business analysis technique to understand or assess the gaps between the current and new business processes that the organization chooses.

There are four steps in this technique:

- Plan strategically
- Analyses of business models
- Describing and creating the procedure
- Technical evaluation of advanced business solutions

This method is preferred by many businesses, particularly the IT sector, because it is a clear, concise way to illustrate the steps of the execution process and demonstrate how it will function in various roles.

3. MOST (Mission, Objectives, Strategies, and Tactics) Analysis

MOST Analysis



MOST is a strong framework for business analysis, and it's regarded as one of the greatest modeling techniques business analysis uses for figuring out what an organization can do and why.

This method entails conducting a thorough internal analysis of the organization's objectives and methods for achieving them. The abbreviation means:

- **Mission:** What is the goal of the company?
- **Objectives:** What are the main objectives that support completing the mission?
- **Strategies:** What are the possible options for attaining the goals?
- **Tactics:** What procedures will the organization use to implement the strategies?

4. Pestle Analysis

While determining how to best address environmental concerns when making business choices, business analysts use the PESTLE model (sometimes referred to as PEST). These factors include:

- **Political:** Financial assistance, subsidies, official programs, and regulations.
- **Economic:** Interest rates, the cost of energy, and labor.

- **Sociological:** life, culture, the media, and population.
- **Technological:** New technologies for information and communication systems.
- **Legal:** Employment standards and local and federal government restrictions.
- **Environmental:** waste, recycling, pollution, and climate.

Analysts can better predict how these aspects will affect the organization's story by evaluating and studying them. Therefore, analysts can more easily build plans to address problems due to this understanding.

5. CATWOE

CATWOE compiles the opinions of various stakeholders onto a single, unified platform, identifying the key actors and winners. This method is a common analysis technique in business to carefully assess how each suggested action would impact the various parties. The abbreviation means:

- **Customers:** Who gains from the company's operations?
- **Actors:** Who are the participants in the process?
- **Transformation:** What kind of transformation occurs at the system's core?
- **World View:** What is the overall situation like, and what effects does it have?
- **Owner:** Who is the system's owner, and what is their connection?
- **Environmental:** What are restrictions, and how do they affect the solution?

6. Brainstorming

There is nothing better than a well-conducted, old-fashioned brainstorming session to produce fresh concepts, pinpoint the sources of a problem, and develop solutions to challenging business issues.

The practice of group brainstorming is frequently incorporated into other approaches like PESTLE and SWOT.

7. Non-functional Requirements Analysis

This technique is used for any project where a technology solution is replaced, modified, or created from scratch.

The analysis identifies and records the qualities required for a brand-new or changed system and frequently deals with needs like data storage or performance. Typically, non-functional requirement analysis covers the following-

- Reliability
- Logging
- Security
- Performance

Non-Functional Requirement Analysis is frequently carried out during the Analysis phase of a project and executed during the Design phase.

8. MoSCoW

MoSCoW sets priorities by providing a framework that compares each demand to the others. Its long form is (Must or Should, Could or Would). Due to the process, you are compelled to consider if a particular piece is necessary.

9. Six Thinking Hats

This business analysis method directs a group's thinking by urging them to consider various viewpoints and ideas. The "six hats" include-

- **White:** Concentrate on your reasoning and data.
- **Red:** Makes use of feelings, instincts, and intuition.
- **Black:** Think about possible bad outcomes and what could go wrong.
- **Yellow:** Keep an upbeat attitude and concentrate on the positives.

- **Green:** Displays originality.
- **Blue:** Process management and consideration of the big picture.

The six thinking hats method is frequently used in conjunction with brainstorming to guide the team's thought processes and get them to consider various points of view.

10. Mind Mapping

The tool, the mind map, is helpful for brainstorming. They are frequently employed by business analysts in idea identification, root cause analysis, and requirement identification with stakeholders. Their straightforward graphic style emphasizes the relationships between ideas and subjects.

They can also support business analysts in identifying issues and potential fixes. Any business analyst has to use mind maps.

Making a mind map involves:

- Placing the primary organizational structure or issues as the focus point.
- The primary components are connected to the center through branches that are labeled to reflect different areas or issues,
- The key components should be identified using just one phrase (ideally).
- The core elements can branch out into second-level branches. These stand for more intricate components.
- If more detail is needed, other levels can be added.

1.3 BUSINESS ANALYTICS IN PRACTICE

Q8. What are the applications of business analytics

Ans : (Imp.)

Business analytics applications show tremendous growth in the relevant sector compared to earlier. Analytics comprises of all the valuable data, computer-based models, and statistical analysis.

With Analytics, appropriate decisions are made successfully for the growth of the company in the future, also getting the companies ready to overcome the upcoming challenges. By these techniques, most problems will be solved already before the issue arises.

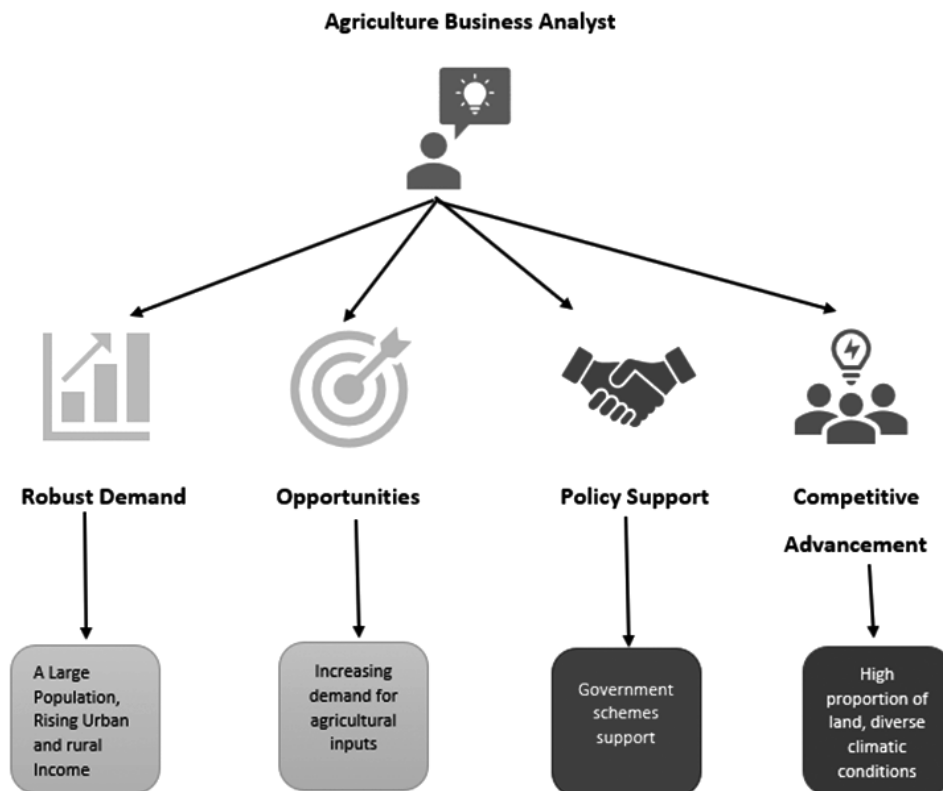
Here, an index of industry application is taken into account for showing the role of business analyst in different domains.

1. Agriculture Business Analytics
2. Stock Marketing
3. Finance Marketing
4. Manufacturing Industry
5. Medical Methodology
6. Customer Relation Management
7. Bond Marketing
8. Human Resources

1. Agricultural Business Analytics

A large part of the Indian economy depends on agriculture. This sector has a major contribution to the economic growth of the country. A business analyst can assure the availability of crops on time, crop production, quality and quantity of seeds, the effect of climate change, monsoon changes, rainwater storage, crop loss, fertilizer requirements, wind direction, floods, and draughts risk management, etc. can be controlled by predictions of business analysts.

For example, predictions on soil parameters can be analyzed by previous data and implemented on crop yield.



Agriculture lacks organizational attention and support from bank loans and other welfare schemes. A business analyst can also process the bank loan and farmers' welfare scheme for increasing agricultural inputs.

It can give opportunities to farmers for individual income and growth. Another advancement is to give insight into irrigation, sowing, harvest, and an area of land for the crop.

2. Stock Marketing

Business analyst improves the performance of the organization in terms of business process and profit by analyzing the variance in the market and updating the changing price or fluctuation in stock trends.

After analyzing the changes taking place in the stock market, he can provide the price list of shares of an item, relevant information regarding shareholder's tracing of audience related to this market, etc.

The business analyst also makes flexible strategies and plans for future investment and earnings. For example, stocks are continuous trends in the market and change very frequently, so he works on the continuous trend dataset to implement strategies.

Business Analytics is surely a vast field to cover, many business analysts are now using machine learning algorithms and natural language processing techniques in order to predict the growth or shrinking of stocks. Although the accuracy depends upon the quality of data being trained.

3. Manufacturing Industry

Business analysts seek to tackle the best possible items for manufacturing and supplying in the market. They investigate the detailed dataset to analyze and implement in business development.

For example, they describe the status to the supplier and product management team regarding the most manufactured product.

The role of technology like IoT in the manufacturing industry shows the importance of technology and its uses.

A business analyst making use of such technologies can tell the highest number of customers for a specific product or service, the product's performance, its demand elasticity (quality and quantity-wise), and product advertisement.

4. Finance Marketing

Business Analysts utilize various analytical techniques and approaches to improve financially relevant issues such as fraud detection, risk mitigation, product pricing, marketing campaign optimization, financial planning, forecasting, etc.

These issues can be controlled by a business analyst. For example, a business analyst can find out the number of customers who are not making payments on time. Similarly, loan defaulters can be traced through a graphical representation that shows the defaulter's age, gender, name, customer ID, etc.

5. Medical Methodology

In the medical or healthcare department, the Business analyst makes predictions about the stock of medicine available in the hospital or medical store, the shipment of medicines in the local market predictions related to disease, impacts of different medicines on same diseases, appointment and availability of doctor, arranging slots for patients, to a medicine available for cure.

For example, allotment of free slots to the patient considering the doctor's working hours, duties of working staff in the hospital, etc.

Production of medicines can also be optimized by the business analyst. He proposes the strategies regarding production costs of medicines, areas of production and stock available, low cost, and high yield preparation methods.

6. Customer Relation Management

Business Analyst takes the required steps for a strong and healthy customer relationship with the organization. He helps to assemble emotional links with customers.

The business analyst helps in increasing productivity according to customer demand and variation in products with consumptions. He utilizes data to maintain the involvement of end to end-user and improves internal and external factors for better customer service and experience.

7. Bond Marketing

The bond market is not much behind the stock market in terms of sellers and buyers using online interfaces to buy and sell bonds. It gives people all the data they want and need to create data-driven trading strategies.

Businesses are solving their liquidity crisis by enabling people to access more information than they used to have before. This helps them make their own reports and ideally make choices on the suitable bond for them.

Business analysts have started to gain almost the same amount of preference and attention given to bond traders.

8. Human Resources

The application of business analytics in human resources can be understood by understanding its role in HR analytics.

The analysis of employees' behavior has been recognized as HR analytics. Ultimately, business analysis plays a significant role in HR analytics as it provides crucial insights into the performance of a project.

1.4 BIG DATA - OVERVIEW OF USING DATA

Q9. What is Big data? Discuss various types of Big Data.

Ans :

(Imp.)

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Following are some of the Big Data examples

The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured

1. Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.

However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Examples of Structured Data

An 'Employee' table in a database is an example of Structured Data

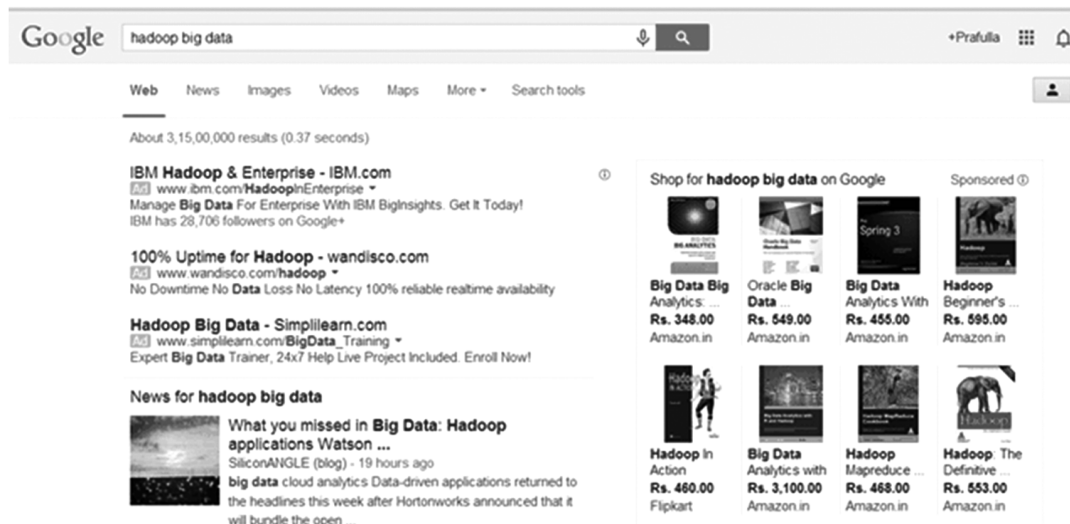
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

2. Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples of Un-structured Data



3. Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples of Semi-structured Data

```
<rec> <name>Prashant Rao</name> <age>35</age> </rec>
<rec> <name>Seema R.</name> <age>41</age> </rec>
```

```
<rec><name>Satish Mane</name><age>29</age></rec>  
<rec><name>Subrato Roy</name><age>26</age></rec>  
<rec><name>Jeremiah J.</name></rec>
```

Personal data stored in an XML file-

Q10. What are the characteristics of Big Data.

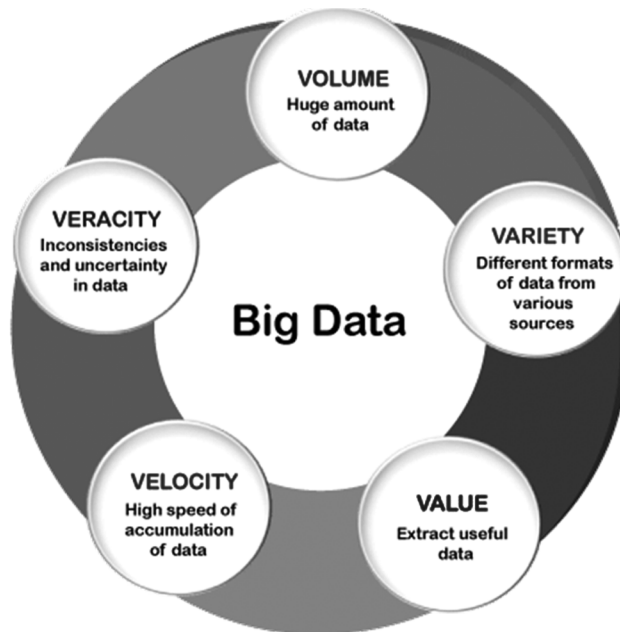
Ans :

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. The data flow would exceed 150 exabytes per day before replication.

There are five v's of Big Data that explains the characteristics.

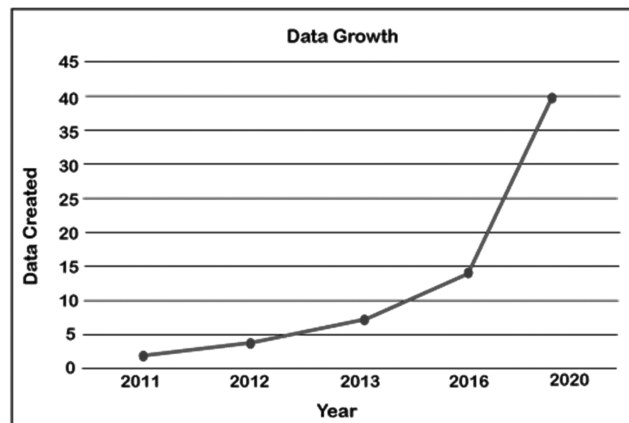
5 V's of Big Data

1. Volume
2. Veracity
3. Variety
4. Value
5. Velocity

**1. Volume**

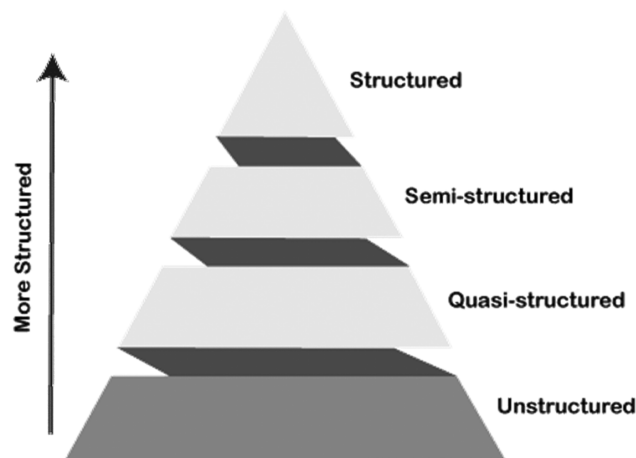
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more.

Facebook can generate approximately a billion messages, 4.5 billion times that the "Like" button is recorded, and more than 350 million new posts are uploaded each day. Big data technologies can handle large amounts of data.



2. Variety

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources. Data will only be collected from databases and sheets in the past, But these days the data will comes in array forms, that are PDFs, Emails, audios, SM posts, photos, videos, etc.



The data is categorized as below:

- (i) **Unstructured Data:** All the unstructured files, log files, audio files, and image files are included in the unstructured data. Some organizations have much data available, but they did not know how to derive the value of data since the data is raw.
- (ii) **Quasi-structured Data:** The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of activities.

- (iii) **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data. It is stored in relations, i.e., tables.
- (iv) **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

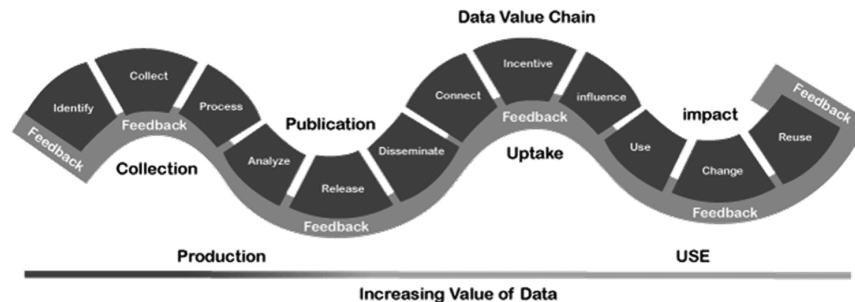
3. Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, Facebook posts with hashtags.

4. Value

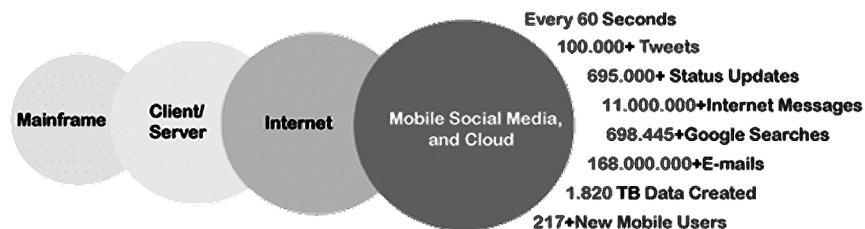
Value is an essential characteristic of big data. It is not the data that we process or store. It is valuable and reliable data that we store, process, and also analyze.



5. Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in real-time. It contains the linking of incoming data sets speeds, rate of change, and activity bursts. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.



Q11. What are the advantages of Big data.

Ans :

(Imp.)

Advantages of Big Data

1. Making wiser decisions

Businesses use big data to enhance B2B operations, advertising, and communication. Big data is primarily being used by many industries, such as travel, real estate, finance, and insurance, to enhance decision-making. Businesses can use big data to accurately predict what customers want and don't want, as well as their behavioural tendencies because it reveals more information in a usable format.

Big data provides business intelligence and cutting-edge analytical insights that help with decision-making. A company can get a more in-depth picture of its target market by collecting more customer data.

2. Cut back on the expense of business operations

According to surveys done by New Vantage and Syncsort (now Precisely), big data analytics has helped businesses significantly cut their costs. Big data is being used to cut costs, according to 66.7% of survey participants from New Vantage.

Moreover, 59.4% of Syncsort survey participants stated that using big data tools improved operational efficiency and reduced costs.

3. Detection of Fraud

Financial companies especially use big data to identify fraud. To find anomalies and transaction patterns, data analysts use artificial intelligence and machine learning algorithms.

These irregularities in transaction patterns show that something is out of place or that there is a mismatch, providing us with hints about potential fraud.

For credit unions, banks, and credit card companies, fraud detection is crucial for identifying account information, materials, or product access.

By spotting frauds before they cause problems, any industry, including finance, can provide better customer service.

For instance, using big data analytics, banks and credit card companies can identify fraudulent purchases or credit cards that have been stolen even before the cardholder becomes aware of the issue.

4. A rise in productivity

A survey by Syncsort found that 59.9% of respondents said they were using big data analytics tools like Spark and Hadoop to boost productivity.

They have been able to increase sales and improve customer retention as a result of this rise in productivity. Modern big data tools make it possible for data scientists and analysts to analyse a lot of data quickly and effectively, giving them an overview of more data.

5. Enhanced customer support

As part of their marketing strategies, businesses must improve customer interactions. Since big data analytics give businesses access to more information, they can use that information to make more specialised, highly personalised offers to each individual customer as well as more targeted marketing campaigns.

Social media, email exchanges, customer CRM (customer relationship management) systems, and other major data sources are the main sources of big data. As a result, it provides businesses with access to a wealth of data about the needs, interests, and trends of their target market.

Big data also enables businesses better to comprehend the thoughts and feelings of their clients to provide them with more individualised goods and services. Providing a personalised experience can increase client satisfaction, strengthen bonds with clients, and, most importantly, foster loyalty.

6. Enhanced speed and agility

Increasing business agility is a big data benefit for competition. Big data analytics can assist businesses in becoming more innovative and adaptable in the marketplace. Large customer data sets can be analysed to help businesses gain insights ahead of the competition and more effectively address customer pain points.

Additionally, having a wealth of data at their disposal enables businesses to assess risks, enhance products and services, and improve communications. Additionally, big data assists businesses in strengthening their business tactics and strategies, which are crucial in coordinating their operations to support frequent and quick changes in the industry.

7. Greater innovation

Innovation is another common benefit of big data, and the NewVantage survey found that 11.6 per cent of executives are investing in analytics primarily as a means to innovate

and disrupt their markets. They reason that if they can glean insights that their competitors don't have, they may be able to get out ahead of the rest of the market with new products and services.

1.5 TYPES OF DATA

Q12. Discuss various types of data.

Ans : (Imp.)

There are different types of data in Statistics, that are collected, analysed, interpreted and presented. The data are the individual pieces of factual information recorded, and it is used for the purpose of the analysis process.

The two processes of data analysis are interpretation and presentation. Statistics are the result of data analysis. Data classification and data handling are important processes as it involves a multitude of tags and labels to define the data, its integrity and confidentiality.

The data is classified into majorly four categories:

1. Nominal data
2. Ordinal data
3. Discrete data
4. Continuous data

Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical.

The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense.

Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

1. Nominal Data

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value.

Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

2. Ordinal Data

Ordinal data/variable is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category.

Quantitative or Numerical Data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing.

Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

3. Discrete Data

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

Example: Number of students in the class

4. Continuous Data

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

Example: Temperature range

Q13. Discuss in detail about various data collection methods.

Ans :

Data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem.

The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organization process.

The main sources of the data collections methods are "Data". Data can be classified into two types, namely primary data and secondary data. The primary importance of data collection in any research or business process is that it helps to determine many important things about the company, particularly the performance.

So, the data collection process plays an important role in all the streams. Depending on the type of data, the data collection method is divided into two categories namely,

1. Primary Data Collection methods
2. Secondary Data Collection methods

1. Primary Data Collection Methods

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types. They are

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Let us discuss the different methods performed to collect the data under these two data collection methods.

(i) Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures.

This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

(ii) Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable.

This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

(iii) Observation Method

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation

- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

(iv) Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person.

The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.

- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

(v) Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire.

The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

(vi) Schedules

This method is similar to the questionnaire method with a slight difference.

The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove mis-understandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

2. Secondary Data Collection Methods

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

- Government publications
- Public records
- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data includes

- Diaries
- Letters
- Unpublished biographies, etc.

1.6 BUSINESS DECISION MODELING

Q14. What is decision modeling? What are the elements of decision modeling.

Ans : (Imp.)

Decision models are visual representation of a process which show how data and knowledge are merged together to make a particular business decision.

The model is connected to business processes, business rules, resources and performance measures in the organizations.

Decision models can also be used for both simple and complex decisions.

Simple decision models use a single decision table or decision tree to show how a set of business rules work on a set of data components to produce a decision.

There are some elements of decision modeling which need to be included to make it complete and they include:

All decision modelling approaches involve three main elements, which are:

- decision.
- information.
- knowledge.

Q15. Discuss various types of decision making models.

Ans :

Decision making models are tools that help individuals and organizations make better choices. There are a variety of different models available, but all share a common goal: to improve the quality of decisions.

Some common Decision making models include the

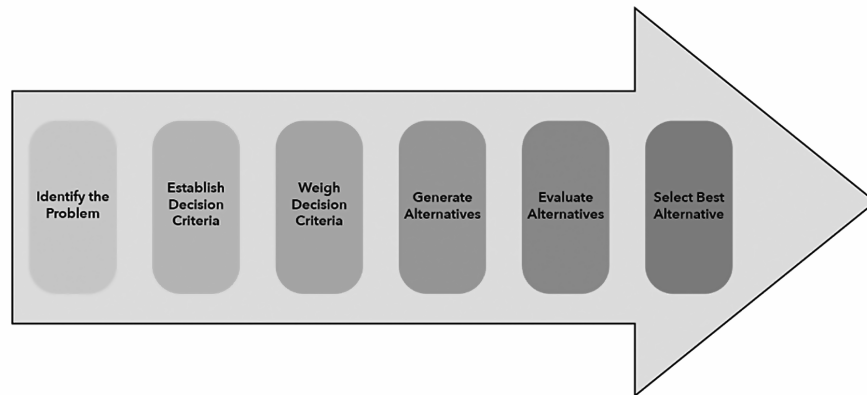
1. Rational decision making model,
2. The bounded rationality model (the satisficing model),
3. Vroom-Yetton decision model,
4. Recognition-primed decision making model,
5. The intuitive decision making model.

Each of these models has its own strengths and weaknesses, but all can be useful in the right situation. The key is to understand when to use each model and how to apply it in a way that will produce the best results. With the right decision making model, organizations and individuals can make choices that are more informed, efficient, and effective.

1. Rational Decision Making Model

The rational decision making model is a popular choice for organizations because it is based on logical reasoning and objective analysis. This model starts with a problem or opportunity, then collects data and information about the situation. Once all relevant information has been gathered, it is analyzed to identify possible decision alternatives and related course of action / solutions.

From there, the pros and cons of each solution are considered and a final decision is made. The picture below represents how do we arrive at the a final decision based on rational decision making.



This model works well when there is time to gather all relevant information and when all stakeholders are in agreement about the best course of action. However, it can be difficult to implement in fast-paced environments or when there are multiple stakeholders with different objectives. Data scientists can help in this phase in terms of performing data discovery related to different decisions or hypotheses. The insights derived from data discovery can then be used to arrive at the final decision and related course of action.

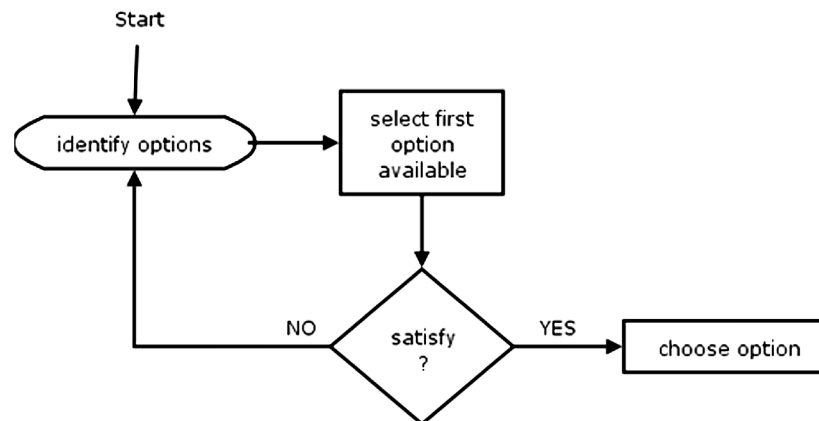
The best course of action and related outcome is laid out as a hypothesis. Hypothesis is then formulated formally. The following represents the next steps about how decision gets driven by the data & insights. These steps are followed in other decision making models as well described in this post.

- Identify the KPIs (leading & lagging) which will be used to measure the output / outcome and validate the hypothesis regarding actions resulting in desired outcome. These KPIs also become the insights called actionable insights.
- Create analytics solution such as a dashboard using Excel or data visualization tools such as QlikSense, Tableau, Google data studio etc
- Perform action and bring the data related to KPIs on dashboard
- Validate the hypothesis about whether the decision driving the action is resulting in desired outcome

2. Bounded Rationality Model (Satisficing Model)

The satisficing model is a simpler version of the rational decision making model. It is also termed as bounded rationality model. It is based on the idea that decision makers cannot always make perfect decisions because they are limited by time, resources, and information.

In this model, decision makers set a goal or criteria for what is considered an acceptable solution. Once this criteria is met, the search for possible solutions stops. Otherwise, the search for different options continue. The picture below represents this idea.



This model works well when time is limited or when there is a large amount of information to consider. However, it can lead to sub-optimal decisions if the criteria for an acceptable solution is not set correctly. This allows decision makers to make decisions quickly without getting bogged down in too much analysis.

Once the decision is made based on the satisficing model, you can follow the steps mentioned in previous section to come up with KPIs, gather the data, build analytical solutions, analyze insights, perform action and validate the hypotheses (action – outcome) based on KPIs.

3. Vroom-Yetton-Jago Decision Making Model

The Vroom-Yetton-Jago decision making model is a tool that can be used to determine the best course of action in a given situation. This model starts with a series of questions about the situation, then uses the answers to these questions to identify the most appropriate decision making style.

This model is useful when there is need to make decisions in a variety of different situations. It helps to ensure that the right decision making style is used for each situation, which can lead to better results.

4. Intuitive Decision Making Model

Intuitive decision making is a process that relies on unconscious pattern recognition and previous experience to make decisions. Intuitive decision making is often contrasted with rational or logical decision making, which relies on deliberate, step-by-step reasoning. Intuitive decision making is a process that relies on gut feelings and instinctual judgments to make decisions.

This type of decision making is often used in situations where there is not enough time to analyze all of the available options. Intuitive decision making can also be useful when the available information is overwhelming or uncertain.

Some research has suggested that Intuitive decision making is more efficient than rational decision making, as it relies on patterns that have been learned through experience. Intuitive decision making can also lead to impulsive decisions that are not well thought-out. Intuitive decision making is typically faster and more efficient than rational decision making, but it can also lead to errors, biases, and bad decisions.

5. Recognition-Primed Decision Making Model

The recognition-primed decision making model suggests that humans use both logic and intuition when making decisions. This model starts with a recognition of the situation, then uses past experience to generate possible solutions.

These solutions are then evaluated based on their feasibility and potential outcomes. A solution is finally selected and implemented in practice. If the solution does not work, the decision maker tries out alternate solutions until a successful one is found.

This model works well when there is time pressure and when there are multiple stakeholders with different objectives. Decision makers need to be experienced in the domain to make effective decisions using this model. Data scientists can help by providing data-driven insights that can be used to evaluate different solutions.

Q16. What are the advantages and disadvantages of using decision models.

Ans :

Advantages

- Decision models are easy to use and explain to the stakeholders.
- They are a useful tool in impact analysis.
- They have multiple points of views which can be added to the diagrams.
- They can help break down complex decisions into simpler ones.
- They can be used for grouping business rules and enabling them for reuse.
- The models work for both simple decision making process like manual decisions or complex decision making like rules-based automation.

Disadvantages

- You would need to use a second diagram style when modelling business processes that contain decisions, thus making them more complex.
- Only those rules required by known decisions are shown and this may lead to misconceptions on the number of business rules that are needed.
- They may cause the stakeholders to believe that their decision models have been standardized.
- They consider the enterprise as a whole and thus increase the number of stakeholder involved in its approval.
- The business terminology must be clearly described to avoid issues related to the automated decisions.

Short Questions and Answers

1. What is Descriptive Analytics?

Ans :

It analyses historical data to determine the response of a unit over a set of given variables. It tracks key performance indicators (KPIs) for a better understanding of the present state of a business.

It involves the following steps:

- Deciding which business metrics will effectively evaluate performance against objectives
- Identifying required data as per the current business state
- Collecting and preparing data using various processes like depublication, transformation, and cleansing.
- Analyzing data for patterns to measure performance
- Presenting data in charts and graphs to make it understandable for non-analytics experts

2. What is Diagnostic Analytics?

Ans :

Diagnostic Analytics is one of those business analytics types that help understand why things happened in the past. Using drill-downs, data mining, data discovery, and correlations, you can comprehend the driving factors.

This advanced analytics method is usually employed as a preceding step of Descriptive Analytics to find the reasoning behind certain results in finance, marketing, cybersecurity, and more.

Examples of Diagnostic Analytics

- Examining market demand
- Identifying technical issues
- Explaining customer behavior
- Improving organization culture

3. What is Predictive Analytics?

Ans :

It considers historical data trends for determining the probability of particular future outcomes. It uses several techniques like data mining, machine learning algorithms, and statistical modeling to forecast the likelihood of events.

Predictive analytics helps improve business areas, including customer service, efficiency, fraud detection and prevention, and risk management. It allows you to grow the most profitable customers, improve the operations of businesses, and determine customer responses and cross-sell opportunities.

Examples of Predictive Analytics

- Predicting customer preferences
- Detection of employee intentions
- Recommending products
- Predicting staff and resources

4. What is Prescriptive Analytics?

Ans :

Prescriptive analytics generates recommendations to handle similar future situations relying on past performances. It employs several tools, statistics, and ML algorithms for the available internal data and external data.

It gives you insights into what may happen, when, and why.

Examples of Prescriptive Analytics

- Tracking fluctuating manufacturing prices
- Improving equipment management
- Suggest the best course of action
- Price modeling
- Evaluating rates of readmission
- Identifying testing

5. What is Structured Data?*Ans :*

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Examples of Structured Data

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

6. What is Nominal Data?*Ans :*

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

7. What is Ordinal Data?*Ans :*

Ordinal data/variable is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category.

8. What are the advantages and disadvantages of using decision models.*Ans :***Advantages**

- Decision models are easy to use and explain to the stakeholders.
- They are a useful tool in impact analysis.
- They have multiple points of views which can be added to the diagrams.
- They can help break down complex decisions into simpler ones.
- They can be used for grouping business rules and enabling them for reuse.
- The models work for both simple decision making process like manual decisions or complex decision making like rules-based automation.

Disadvantages

- You would need to use a second diagram style when modelling business processes that contain decisions, thus making them more complex.
 - Only those rules required by known decisions are shown and this may lead to misconceptions on the number of business rules that are needed.
 - They may cause the stakeholders to believe that their decision models have been standardized.
 - They consider the enterprise as a whole and thus increase the number of stakeholder involved in its approval.
 - The business terminology must be clearly described to avoid issues related to the automated decisions.
-

9. What is Business Analytics?*Ans :*

Business analytics is the process of gathering and processing all of your business data, and applying statistical models and iterative methodologies to translate that data into business insights. Most importantly, you need to be able to translate that data into what your customers want. And attempting to guess what your customers want is not enough anymore.

Business analytics combines processes, skills, and technologies to collect, analyze and present historical performance of the business, based on data, with the goal of driving business planning. Business analytics essentially uses large amounts of historical business data for the purpose of providing insights to evaluate a business.

10. What is Unstructured Data?*Ans :*

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples of Un-structured Data

Google search results for "hadoop big data".

Search bar: hadoop big data

Navigation: Web, News, Images, Videos, Maps, More, Search tools

Results: About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
www.ibm.com/HadoopnEnterprise
Manage **Big Data** For Enterprise With IBM BigInsights. Get it Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
www.wandisco.com/hadoop
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
www.simplilearn.com/BigData_Training
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) · 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for hadoop big data on Google

Sponsored

Product	Price	Source
Big Data Big Analytics ...	Rs. 348.00	Amazon.in
Oracle Big Data	Rs. 549.00	Amazon.in
Big Data Analytics With Spring 3	Rs. 455.00	Amazon.in
Hadoop Beginner's ...	Rs. 595.00	Amazon.in
Hadoop in Action	Rs. 460.00	Flipkart
Big Data Analytics with ...	Rs. 3,100.00	Amazon.in
Hadoop Mapreduce ...	Rs. 468.00	Amazon.in
Hadoop: The Definitive ...	Rs. 553.00	Amazon.in

Choose the Correct Answers

1. _____ is a branch of advanced analytics that makes predictions about future outcomes using historical data. [a]
(a) Predictive analytics (b) Diagnostic
(c) Prescriptive (d) Descriptive
2. _____ is a form of data analytics that uses past performance and trends to determine what needs to be done to achieve future goals. [c]
(a) Predictive analytics (b) Diagnostic
(c) Prescriptive (d) Descriptive
3. _____ is a statistical interpretation used to analyze historical data to identify patterns and relationships. [d]
(a) Predictive analytics (b) Diagnostic
(c) Prescriptive (d) Descriptive
4. Qualitative data, also known as the _____. [a]
(a) Categorical (b) Numerical
(c) Ordinal (d) Continuous
5. _____ data is one of the types of qualitative information which helps to label the variables without providing the numerical value. [a]
(a) Nominal (b) Ordinal
(c) Discrete (d) Continuous
6. _____ data/variable is a type of data that follows a natural order. [b]
(a) Nominal (b) Ordinal
(c) Discrete (d) Continuous
7. _____ information contains only a finite number of possible values. [c]
(a) Nominal (b) Ordinal
(c) Discrete (d) Continuous
8. _____ data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range. [d]
(a) Nominal (b) Ordinal
(c) Discrete (d) Continuous
9. _____ data is also known as numerical data which represents the numerical value. [b]
(a) Qualitative (b) Quantitative
(c) Ordinal (d) Nominal
10. _____ is a combination of structured, semistructured and unstructured data. [-]
(a) Big data (b) Small data
(c) Semi data (d) Natural Data

Fill in the Blanks

1. _____ is the process of transforming data into insights to improve business decisions.
2. _____ can show “what happened” and is the foundation of data insights.
3. _____ addresses “why things happened.”
4. Businesses use _____ to “see the future” and predict “what is likely to happen.
5. _____ analytics driven by AI (Artificial Intelligence) systems, helps companies make decisions and determine “what they should do next.”
6. _____ can be a valuable resource when approaching an important strategic decision.
7. _____ and _____ tools can help break down the numbers and models so that the human eye can easily grasp what is being presented.
8. _____ helps organizations make better decisions by providing data-driven insights into customer needs, competitor strategies, and other market trends.
9. By using _____, businesses are better equipped to optimize their processes and allocate resources more efficiently.
10. _____ for business analytics sorts through large datasets using databases, statistics, and machine learning to identify trends and establish relationships

ANSWERS

1. Business analytics
2. Descriptive analytics
3. Diagnostic analytics
4. predictive analytics
5. Prescriptive analytics
6. Business analytics
7. Visualization, reporting
8. Business analytics
9. Analytics
10. Data mining

One Mark Answers

1. Nominal Data

Ans :

Nominal data is a type of data that represents discrete units which is why it cannot be ordered and measured. They are used to label variables without providing any quantitative value. Also, they have no meaningful zero.

2. Ordinal Data

Ans :

Ordinal values represent discrete as well as ordered units. Unlike nominal, here the ordering matters. However, there is no consistency in the relative distance between the adjacent categories. And, similar to nominal data, ordinal data also don't have a meaningful zero.

3. Interval Data

Ans :

It represents ordered data that is measured along a numerical scale with equal distances between the adjacent units. These equal distances are also referred to as intervals.

4. Ratio Data

Ans :

Ratio data are also ordered with the same difference between the individual units. However, they also have a meaningful zero so they cannot take negative values.

5. Qualitative Data

Ans :

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical.

UNIT II

DESCRIPTIVE ANALYTICS

Overview of Description Statistics (Central Tendency, Variability), Data Visualization -Definition, Visualization Techniques – Tables, Cross Tabulations, charts, Data Dashboards using Advanced Ms-Excel or SPSS.

2.1 OVERVIEW OF DESCRIPTION STATISTICS (CENTRAL TENDENCY, VARIABILITY)

Q1. What is descript analytics? Give any two examples.

Ans : (Imp.)

Descriptive analytics is the process of using current and historical data to identify trends and relationships. It's sometimes called the simplest form of data analysis because it describes trends and relationships but doesn't dig deeper.

Descriptive analytics is relatively accessible and likely something your organization uses daily. Basic statistical software, such as Microsoft Excel or data visualization tools, such as Google Charts and Tableau, can help parse data, identify trends and relationships between variables, and visually display information.

Descriptive analytics is especially useful for communicating change over time and uses trends as a springboard for further analysis to drive decision-making.

Following are the examples of descriptive analytics.

1. Traffic and Engagement Reports

One example of descriptive analytics is reporting. If your organization tracks engagement in the form of social media analytics or web traffic, you're already using descriptive analytics.

These reports are created by taking raw data-generated when users interact with your website, advertisements, or social media content-and using it to compare current metrics to historical metrics and visualize trends.

For example, you may be responsible for reporting on which media channels drive the most traffic to the product page of your company's website. Using descriptive analytics, you can analyze the page's traffic data to determine the number of users from each source. You may decide to take it one step further and compare traffic source data to historical data from the same sources. This can enable you to update your team on movement; for instance, highlighting that traffic from paid advertisements increased 20 percent year over year.

The three other analytics types can then be used to determine why traffic from each source increased or decreased over time, if trends are predicted to continue, and what your team's best course of action is moving forward.

2. Financial Statement Analysis

Another example of descriptive analytics that may be familiar to you is financial statement analysis. Financial statements are periodic reports that detail financial information about a business and, together, give a holistic view of a company's financial health.

There are several types of financial statements, including the balance sheet, income statement, cash flow statement, and statement of shareholders' equity. Each caters to a specific audience and conveys different information about a company's finances.

Financial statement analysis can be done in three primary ways: vertical, horizontal, and ratio.

Vertical analysis involves reading a statement from top to bottom and comparing each item to those above and below it. This helps determine relationships between variables. For instance, if each line item is a percentage of the total, comparing them can provide insight into which are taking up larger and smaller percentages of the whole.

Horizontal analysis involves reading a statement from left to right and comparing each item to itself from a previous period. This type of analysis determines change over time.

Finally, ratio analysis involves comparing one section of a report to another based on their relationships to the whole. This directly compares items across periods, as well as your company's ratios to the industry's to gauge whether yours is over- or underperforming.

Each of these financial statement analysis methods are examples of descriptive analytics, as they provide information about trends and relationships between variables based on current and historical data.

3. Demand Trends

Descriptive analytics can also be used to identify trends in customer preference and behavior and make assumptions about the demand for specific products or services.

Streaming provider Netflix's trend identification provides an excellent use case for descriptive analytics. Netflix's team - which has a track record of being heavily data-driven - gathers data on users' in-platform behavior. They analyze this data to determine which TV series and movies are trending at any given time and list trending titles in a section of the platform's home screen.

Not only does this data allow Netflix users to see what's popular - and thus, what they might enjoy watching - but it allows the Netflix team to know which types of media, themes, and actors are especially favored at a certain time. This can drive decision-making about future original content creation, contracts with existing production companies, marketing, and retargeting campaigns.

Q2. What are the benefits and challenges of descriptive analytics.

Ans :

There are several benefits of descriptive analytics.

i) Simple Analysis

Descriptive analysis doesn't require great expertise or experience in statistical methods or analytics.

ii) Many Tools Available

Many apps make this function a plug-and-play form of analysis.

It Answers Most Common Business Performance Questions

Most stakeholders and salespeople want simple answers to basic questions such as "How are we doing?" or "Why did sales drop?" Descriptive analytics provides the data to effectively and efficiently answer those questions.

Challenges to Descriptive Analytics

Like any other tool, descriptive analysis is not without problems. There are three significant challenges for organizations wanting to use descriptive analytics.

i) It is a Blunt Tool without Insight

The descriptive analysis examines the relationship between a handful of variables, and that is all. It simply describes what is happening. Organizations must ensure that users understand what descriptive analytics will provide.

ii) It Tells an Organization What, Not Why

Descriptive analysis reports events as they happened, not why they happened or what could happen next. The organization will need to run the full analytics suite entirely to grasp a situation.

iii) Can Measure the Wrong Thing

If the incorrect metrics are used, the analysis is useless. Organizations must analyze what they want to measure and why. Thought must be put into this process and matched with the outcomes that current data can provide.

iv) Poor Data Quality

While vast amounts of data can be collected, it will not produce accurate results if it is not helpful or full of errors. After an organization decides on the metrics it requires, the data must be checked to ensure it can provide this information. Once it is ascertained that it will provide the relevant information, the data must be thoroughly cleansed. Erroneous data, duplicates, and missing data fields must be resolved.

Q3. What are the steps involved in Descriptive Analytics.

Ans :

Applying descriptive analytics generally starts with defining the metrics you want to produce and culminates with presenting them in the desired format. Here are the steps to follow to generate your own descriptive analytics.

- 1. State business metrics:** The first step is to identify the metrics that you want to generate. These should reflect key business goals of each group or of the company overall. For example, a growth-oriented company might focus on measuring quarterly increases in revenue, while the company's accounts receivable group might want to track days sales outstanding and other metrics that reflect how long it takes to collect money from customers.
- 2. Identify data required:** Locate the data you need to produce the desired metrics. At some companies, the data may be scattered across multiple applications and files. Companies that use ERP systems may already have most or all of the data they need in their systems' databases. Some metrics may also require data from external sources, such as industry benchmarking databases, e-commerce websites and social media platforms.
- 3. Extract and prepare data:** If the data comes from multiple sources, extracting, combining and preparing the data for analysis is a time-consuming yet vital step to ensure accuracy. This step may involve data cleansing to eliminate inconsistencies and errors in data from different sources, as well as transforming data into a format suitable for analysis tools. Advanced forms of data analytics employ a process called data modeling to help prepare, structure and organize company information. Data modeling is a framework within information systems to define and format data.

- 4. Analyze data:** Companies can use a variety of tools to apply descriptive analytics, from spreadsheets to business intelligence (BI) software. Descriptive analytics often involves applying basic mathematical operations to one or more variables. For example, sales managers may want to track the average revenue per sale or monthly revenue from new customers. Executives and financial specialists may seek to monitor financial metrics such as gross profit margin, which is the ratio of gross profit to sales.

- 5. Present data:** Presenting data in compelling visual forms, such as pie charts, bar charts and line graphs, often makes it easier for stakeholders to understand. However, some people, including finance specialists, may prefer to see information presented as numbers and tables.

Q4. What is mean by descriptive statistics? and list out the types of descriptive statistics.

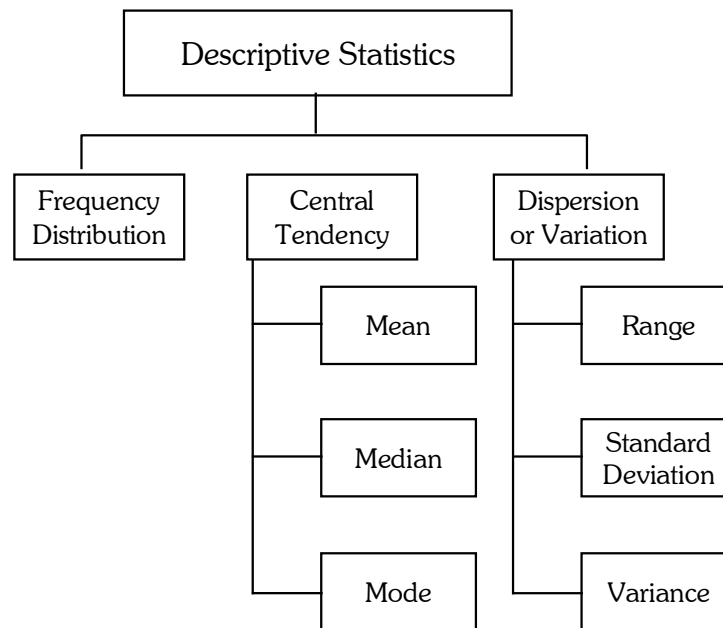
Ans :

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

There are 3 main types of descriptive statistics:

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** or dispersion concerns how spread out the values are.



1. Frequency distribution

A data set is made up of a distribution of values, or scores. In tables or graphs, you can summarize the frequency of every possible value of a variable in numbers or percentages. This is called a **frequency distribution**.

Example:

You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:

- Go to a library
- at a theater
- Visit a national park

Your data set is the collection of responses to the survey. Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

- a) Simple frequency distribution table:** For the variable of gender, you list all possible answers on the left hand column. You count the number or percentage of responses for each answer and display it on the right hand column.

Gender	Number
Male	182
Female	235
Other	27

From this table, you can see that more women than men or people with another gender identity took part in the study.

- b) **Grouped distribution table:** In a grouped frequency distribution, you can group numerical response values and add up the number of responses for each group. You can also convert each of these numbers to percentages.

Library visits in the past year Percent

0–4	6%
5–8	20%
9–12	42%
13–16	24%
17+	16%

From this table, you can see that most people visited the library between 5 and 16 times in the past year.

2. Measures of central tendency

Measures of central tendency estimate the center, or average, of a data set. The mean, median and mode are 3 ways of finding the average.

Here we will demonstrate how to calculate the mean, median, and mode using the first 6 responses of our survey.

a) Mean

The **mean**, or M , is the most used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called N .

Mean number of library visits

Data set 15, 3, 12, 0, 24, 3

Sum of all values $15 + 3 + 12 + 0 + 24 + 3 = 57$

Total number of responses $N = 6$

Mean Divide the sum of values by N to find M : $57/6 = 9.5$

b) Median

The **median** is the value that's exactly in the middle of a data set.

To find the median, order each response value from the smallest to the biggest. Then, the median is the number in the middle. If there are two numbers in the middle, find their mean.

Median number of library visits

Ordered data set 0, 3, 3, 12, 15, 24

Middle numbers 3, 12

Median Find the mean of the two middle numbers: $(3 + 12)/2 = 7.5$

c) Mode

The **mode** is simply the most popular or most frequent response value. A data set can have no mode, one mode, or more than one mode.

To find the mode, order your data set from lowest to highest and find the response that occurs most frequently.

Mode number of library visits

Ordered data set 0, 3, 3, 12, 15, 24

Mode Find the most frequently occurring response: **3**

3. Measures of Variability

Measures of variability give you a sense of how spread out the response values are. The range, standard deviation and variance each reflect different aspects of spread.

Range

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.

Range of visits to the library in the past year

Ordered data set: 0, 3, 3, 12, 15, 24

Range: $24 - 0 = \mathbf{24}$

Standard deviation

The standard deviation (s or SD) is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by $N - 1$.
6. Find the square root of the number you found.

Standard deviations of visits to the library in the past year In the table below, you complete **Steps 1 through 4**.

Raw data	Deviation from mean	Squared deviation
15	$15 - 9.5 = 5.5$	30.25
3	$3 - 9.5 = -6.5$	42.25
12	$12 - 9.5 = 2.5$	6.25
0	$0 - 9.5 = -9.5$	90.25
24	$24 - 9.5 = 14.5$	210.25
3	$3 - 9.5 = -6.5$	42.25
$M = 9.5$	Sum = 0	Sum of squares = 421.5

Step 5: $421.5/5 = 84.3$

Step 6: $\sqrt{84.3} = 9.18$

From learning that $s = 9.18$, you can say that on average, each score deviates from the mean by 9.18 points.

Variance

The variance is the average of squared deviations from the mean. Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

To find the variance, simply square the standard deviation. The symbol for variance is s^2 .

Variance of visits to the library in the past year

Data set: 15, 3, 12, 0, 24, 3

$s = 9.18$

$s^2 = 84.3$

Q5. What is Frequency distribution? What are the types of frequency distribution?

Ans :

(Imp.)

The frequency of a value is the number of times it occurs in a dataset. A frequency distribution is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions

There are four types of frequency distributions:

- **Ungrouped frequency distributions:** The number of observations of each **value** of a variable.
 - You can use this type of frequency distribution for categorical variables.
- **Grouped frequency distributions:** The number of observations of each **class interval** of a variable. Class intervals are ordered groupings of a variable's values.
 - You can use this type of frequency distribution for quantitative variables.
- **Relative frequency distributions:** The proportion of observations of each value or class interval of a variable.
 - You can use this type of frequency distribution for **any type of variable** when you're more interested in **comparing frequencies** than the actual number of observations.
- **Cumulative frequency distributions:** The sum of the frequencies less than or equal to each value or class interval of a variable.
 - You can use this type of frequency distribution for ordinal or quantitative variables when you want to understand how often observations fall below certain values.

Q6. Explain in detail about Measures of central tendency.

Ans :

Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

- **Mode:** the most frequent value.
- **Median:** the middle number in an ordered dataset.
- **Mean:** the sum of all values divided by the total number of values.

Mean

The arithmetic mean of a dataset (which is different from the geometric mean) is the sum of all values divided by the total number of values. It's the most commonly used measure of central tendency because all values are used in the calculation.

Example: Finding the mean

Participant	1	2	3	4	5
Reaction time (milliseconds)	287	345	365	298	380

First you add up the sum of all values:

$$\Sigma x = 287 + 345 + 365 + 298 + 380 = 1,675$$

Then you calculate the mean using the formula

$$\frac{\Sigma x}{n}$$

There are 5 values in the dataset, so $n = 5$.

$$\bar{x} = \frac{1675}{5} = 335$$

Mean (x): 335 milliseconds

Median

The median of a dataset is the value that's exactly in the middle when it is ordered from low to high.

Example: Finding the median You measure the reaction times of 7 participants on a computer task and categorize them into 3 groups: slow, medium or fast.

Participant	1	2	3	4	5	6	7
Speed	Medium	Slow	Fast	Fast	Medium	Fast	Slow

To find the median, you first order all values from low to high. Then, you find the value in the middle of the ordered dataset - in this case, the value in the 4th position.

Ordered dataset Slow Slow Medium Medium Fast Fast Fast

Median: Medium

In larger datasets, it's easier to use simple formulas to figure out the position of the middle value in the distribution. You use different methods to find the median of a dataset depending on whether the total number of values is even or odd.

Median of an odd-numbered dataset

For an odd-numbered dataset, find the value that lies at the $\frac{(n+1)}{2}$ position, where n is the number of values in the dataset.

Example

You measure the reaction times in milliseconds of 5 participants and order the dataset.

Reaction time (milliseconds) 287 298 345 365 380

The middle position is calculated using, $\frac{(n+1)}{2}$ where $n = 5$.

That means the median is the 3rd value in your ordered dataset.

Median: 345 milliseconds

Median of an even-numbered dataset

For an even-numbered dataset, find the two values in the middle of the dataset: the values at the $\frac{n}{2}$ and $\left(\frac{n}{2}\right) + 1$ positions. Then, find their mean.

Example: You measure the reaction times of 6 participants and order the dataset.

The middle positions are calculated using $\frac{n}{2}$ and $\left(\frac{n}{2}\right) + 1$, where $n = 6$.

$$\frac{6}{2} = 3$$

$$\left(\frac{6}{2}\right) + 1 = 4$$

That means the middle values are the 3rd value, which is **345**, and the 4th value, which is **357**.

To get the median, take the mean of the 2 middle values by adding them together and dividing by 2.

$$\frac{(345 + 357)}{2} = 351$$

Median: 351 milliseconds

Mode

The **mode** is the most frequently occurring value in the dataset. It's possible to have no mode, one mode, or more than one mode.

To find the mode, sort your dataset numerically or categorically and select the response that occurs most frequently.

Example

Finding the mode In a survey, you ask 9 participants whether they identify as conservative, moderate, or liberal.

To find the mode, sort your data by category and find which response was chosen most frequently.

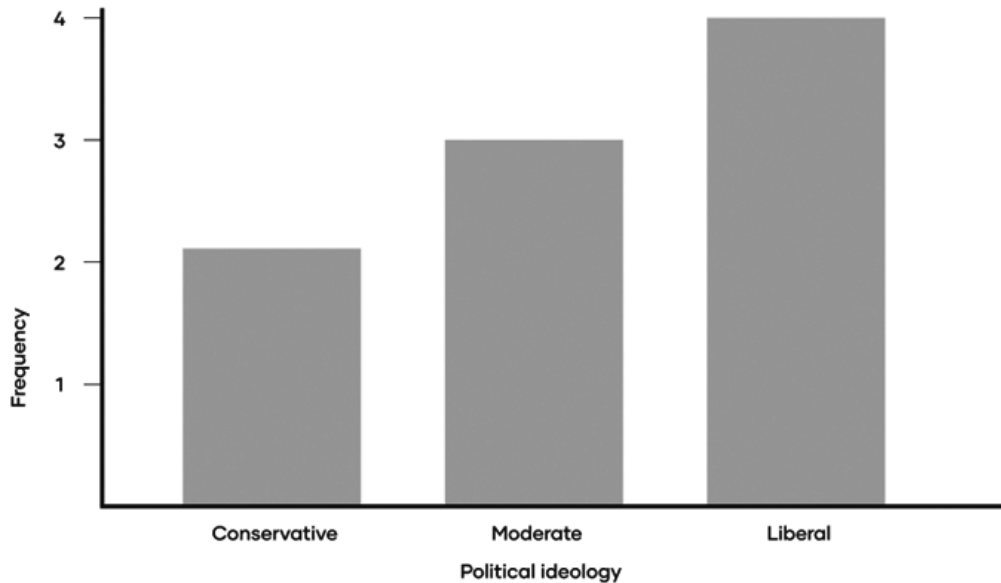
To make it easier, you can create a frequency table to count up the values for each category.

Political ideology	Frequency
Conservative	2
Moderate	3
Liberal	4

Mode: Liberal

The mode is easily seen in a bar graph because it is the value with the highest bar.

Frequency distribution: Political ideology



The 3 main measures of central tendency are best used in combination with each other because they have complementary strengths and limitations. But sometimes only 1 or 2 of them are applicable to your dataset, depending on the level of measurement of the variable.

- The mode can be used for any level of measurement, but it's most meaningful for nominal and ordinal levels.
- The median can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.
- The mean can only be used on interval and ratio levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

Levels of measurement	Examples	Measure of central tendency
Nominal	Ethnicity	Mode
	Political ideology	
Ordinal	Level of anxiety	Mode
	Income bracket	
Interval and ratio	Median	Mode
	Reaction time	
	Test score	
	Temperature	
	Median	
	Mean	

Q7. Explain Central tendency in point of distributions.*Ans :***(Imp.)**

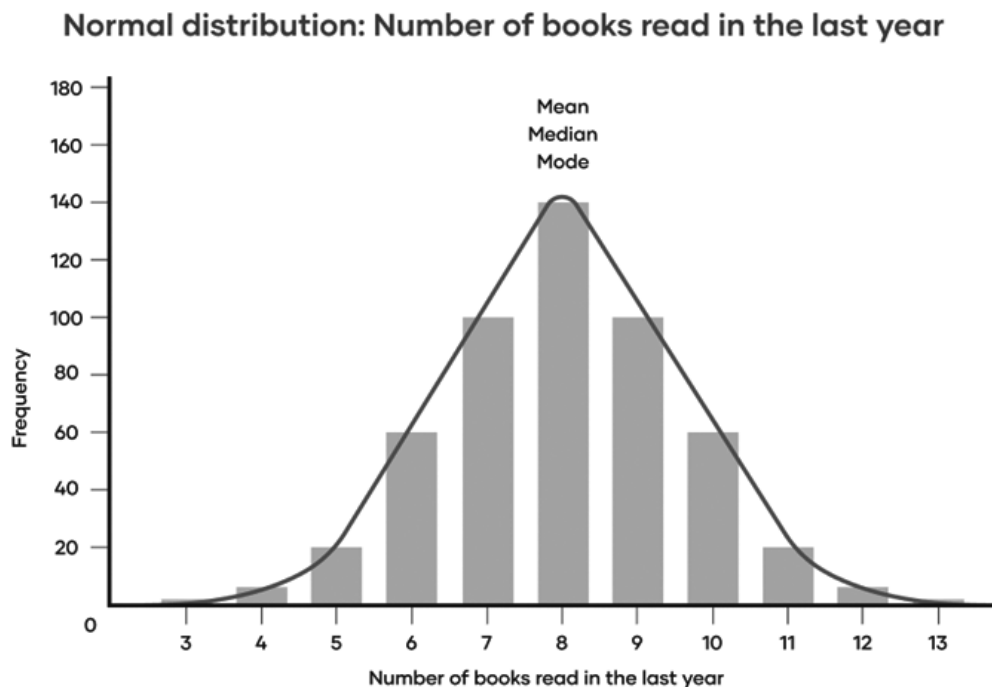
A dataset is a distribution of n number of scores or values.

Normal distribution

In a normal distribution, data is symmetrically distributed with no skew. Most values cluster around a central region, with values tapering off as they go further away from the center. The mean, mode and median are exactly the same in a normal distribution.

Example: Normal distribution You survey a sample in your local community on the number of books they read in the last year.

A histogram of your data shows the frequency of responses for each possible number of books. From looking at the chart, you see that there is a normal distribution.



The mean, median and mode are all equal; the central tendency of this dataset is 8.

Skewed distributions

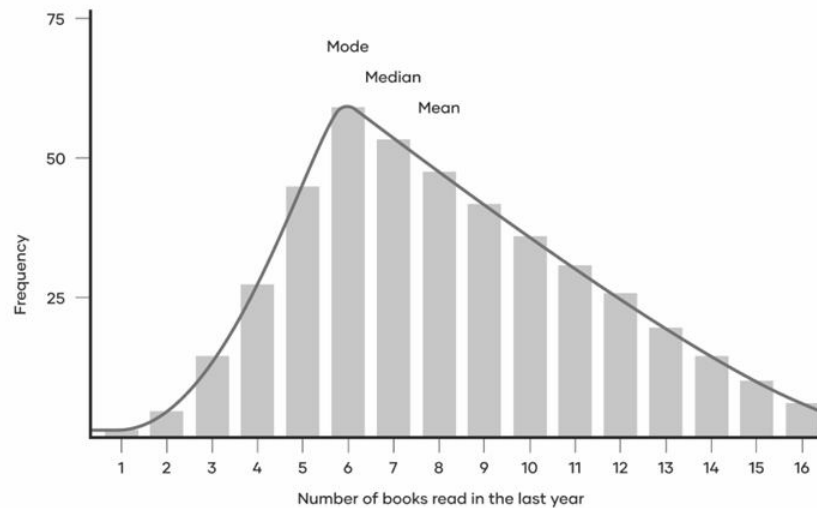
In skewed distributions, more values fall on one side of the center than the other, and the mean, median and mode all differ from each other. One side has a more spread out and longer tail with fewer scores at one end than the other. The direction of this tail tells you the side of the skew

In a positively skewed distribution, there's a cluster of lower scores and a spread out tail on the right. In a negatively skewed distribution, there's a cluster of higher scores and a spread out tail on the left.

a) Positively skew distribution

- In this histogram, your distribution is skewed to the right, and the central tendency of your dataset is on the lower end of possible scores.
- In a positively skewed distribution, $\text{mode} < \text{median} < \text{mean}$.

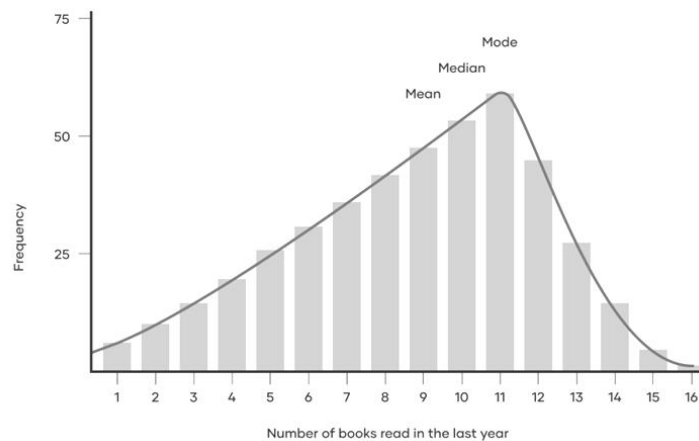
Positively skewed distribution: Number of books read in the last year

**b) Negitively skewed distribution**

In this histogram, your distribution is skewed to the left, and the central tendency of your dataset is towards the higher end of possible scores.

In a negatively skewed distribution, $\text{mean} < \text{median} < \text{mode}$.

Negitively skewed distribution: Number of books read in the last year

**Q8. Explain in detail about variance.**

Ans :

The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean.

Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

Different formulas are used for calculating variance depending on whether you have data from a whole population or a sample.

Population variance

When you have collected data from every member of the population that you're interested in, you can get an exact value for population variance.

The population variance formula looks like this:

Formula

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

Explanation

- σ^2 = population variance
- Σ = sum of
- x = each value
- μ = population mean
- N = number of values in the population

Sample variance

When you collect data from a sample, the sample variance is used to make estimates or inferences about the population variance.

The sample variance formula looks like this:

Formula

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

Explanation

- S^2 = sample variance
- Σ = sum of....
- X = each value
- \bar{x} = sample mean
- n = number of values in the population

With samples, we use $n - 1$ in the formula because using n would give us a biased estimate that consistently underestimates variability. The sample variance would tend to be lower than the real variance of the population.

Reducing the sample n to $n - 1$ makes the variance artificially large, giving you an unbiased estimate of variability: it is better to overestimate rather than underestimate variability in samples.

Steps for calculating the variance by hand

The variance is usually calculated automatically by whichever software you use for your statistical analysis. But you can also calculate it by hand to better understand how the formula works.

There are five main steps for finding the variance by hand. We'll use a small data set of 6 scores to walk through the steps.

Data set

46 69 32 60 52 41

Step 1: Find the mean

To find the mean, add up all the scores, then divide them by the number of scores.

Mean (\bar{x})

$$\bar{x} = (46 + 69 + 32 + 60 + 52 + 41) \div 6 = \mathbf{50}$$

Step 2: Find each score's deviation from the mean

Subtract the mean from each score to get the deviations from the mean.

Since $\bar{x} = 50$, take away 50 from each score.

Score	Deviation from the mean
46	$46 - 50 = \mathbf{-4}$
69	$69 - 50 = \mathbf{19}$
32	$32 - 50 = \mathbf{-18}$
60	$60 - 50 = \mathbf{10}$
52	$52 - 50 = \mathbf{2}$
41	$41 - 50 = \mathbf{-9}$

Step 3: Square each deviation from the mean

Multiply each deviation from the mean by itself. This will result in positive numbers.

Squared deviations from the mean

$$(-4)^2 = 4 \times 4 = \mathbf{16}$$

$$19^2 = 19 \times 19 = \mathbf{361}$$

$$(-18)^2 = -18 \times -18 = \mathbf{324}$$

$$10^2 = 10 \times 10 = \mathbf{100}$$

$$2^2 = 2 \times 2 = \mathbf{4}$$

$$(-9)^2 = -9 \times -9 = \mathbf{81}$$

Step 4: Find the sum of squares

Add up all of the squared deviations. This is called the sum of squares.

Sum of squares

$$16 + 361 + 324 + 100 + 4 + 81 = \mathbf{886}$$

Step 5: Divide the sum of squares by $n - 1$ or N

Divide the sum of the squares by $n - 1$ (for a sample) or N (for a population).

Since we're working with a sample, we'll use $n - 1$, where $n = 6$.

Variance

$$886 \div (6 - 1) = 886 \div 5 = \mathbf{177.2}$$

2.2 DATA VISUALIZATION - DEFINITION, VISUALIZATION TECHNIQUES

Q9. What is Data Visualization? What are the benefits of data visualization.

Ans :

(Imp.)

Data visualization is the representation of information and data using charts, graphs, maps, and other visual tools. These visualizations allow us to easily understand any patterns, trends, or outliers in a data set.

Data visualization also presents data to the general public or specific audiences without technical knowledge in an accessible manner. For example, the health agency in a government might provide a map of vaccinated regions.

The purpose of data visualization is to help drive informed decision-making and to add colorful meaning to an otherwise bland database.

Benefits of data visualization

Data visualization can be used in many contexts in nearly every field, like public policy, finance, marketing, retail, education, sports, history, and more. Here are the benefits of data visualization:

- **Storytelling:** People are drawn to colors and patterns in clothing, arts and culture, architecture, and more. Data is no different colors and patterns allow us to visualize the story within the data.
- **Accessibility:** Information is shared in an accessible, easy-to-understand manner for a variety of audiences.
- **Visualize relationships:** It's easier to spot the relationships and patterns within a data set when the information is presented in a graph or chart.
- **Exploration:** More accessible data means more opportunities to explore, collaborate, and inform actionable decisions.

Q10. Discuss various types of data visualization.

Ans :

Visualizing data can be as simple as a bar graph or scatter plot but becomes powerful when analysing.

Here are some common types of data visualizations:

- i) **Table:** A table is data displayed in rows and columns, which can be easily created in a Word document or Excel spreadsheet.
- ii) **Chart or graph:** Information is presented in tabular form with data displayed along an x and y axis, usually with bars, points, or lines, to represent data in comparison. An infographic is a special type of chart that combines visuals and words to illustrate the data.
 - **Gantt chart:** A Gantt chart is a bar chart that portrays a timeline and tasks specifically used in project management.
 - **Pie chart:** A pie chart divides data into percentages featured in “slices” of a pie, all adding up to 100%.
- iii) **Geospatial visualization:** Data is depicted in map form with shapes and colors that illustrate the relationship between specific locations, such as a choropleth or heat map.
- iv) **Dashboard:** Data and visualizations are displayed, usually for business purposes, to help analysts understand and present data.

Q11. What are the applications of data visualization.**Ans :** (Imp.)

Using data visualization tools, different types of charts and graphs can be created to illustrate important data. These are a few examples of data visualization in the real world:

- **Data science:** Data scientists and researchers have access to libraries using programming languages or tools such as Python or R, which they use to understand and identify patterns in data sets. Tools help these data professionals work more efficiently by coding research with colors, plots, lines, and shapes.
- **Marketing:** Tracking data such as web traffic and social media analytics can help marketers analyze how customers find their products and whether they are early adopters or more of a laggard buyer. Charts and graphs can synthesize data for marketers and stakeholders to better understand these trends.
- **Finance:** Investors and advisors focused on buying and selling stocks, bonds, dividends, and other commodities will analyze the movement of prices over time to determine which are worth purchasing for short- or long-term periods. Line graphs help financial analysts visualize this data, toggling between months, years, and even decades.
- **Health policy:** Policymakers can use choropleth maps, which are divided by geographical area (nations, states, continents) by colors. They can, for example, use these maps to demonstrate the mortality rates of cancer or ebola in different parts of the world.

Q12. List out the tools which are helpful for data visualization.**Ans :**

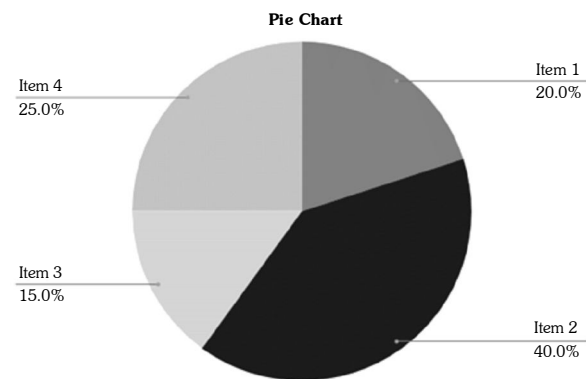
There are plenty of data visualization tools out there to suit your needs. Before committing to one, consider researching whether you need an open-source site or could simply create a graph using Excel or Google Charts. The following are common data visualization tools that could suit your needs.

- Tableau
- Google Charts
- Dundas BI
- Power BI
- Jupyter
- Infogram
- ChartBlocks
- D3.js
- FusionCharts
- Grafana

Q13. List and explain various data visualization techniques**Ans :** (Imp.)

The type of data visualization technique you leverage will vary based on the type of data you're working with.

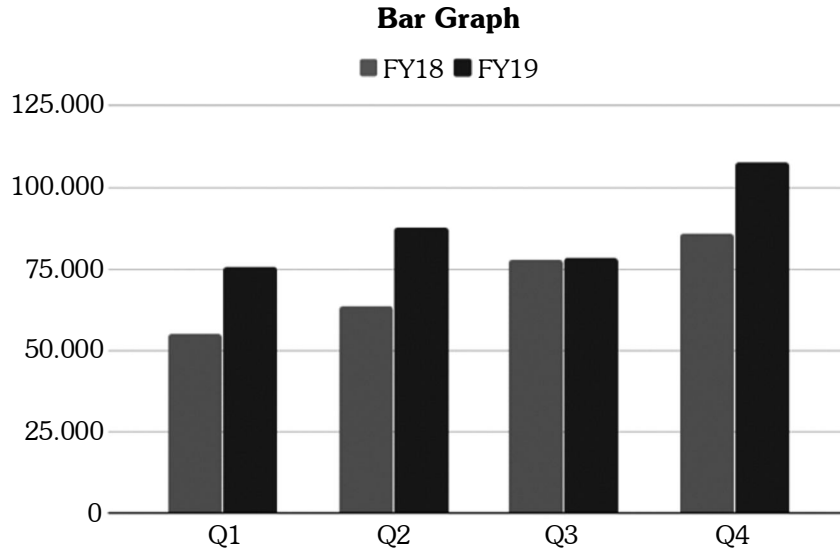
Here are some important data visualization techniques to know:

1. Pie Chart

Pie charts are one of the most common and basic data visualization techniques, used across a wide range of applications. Pie charts are ideal for illustrating proportions, or part-to-whole comparisons.

Because pie charts are relatively simple and easy to read, they're best suited for audiences who might be unfamiliar with the information or are only interested in the key takeaways. For viewers who require a more thorough explanation of the data, pie charts fall short in their ability to display complex information.

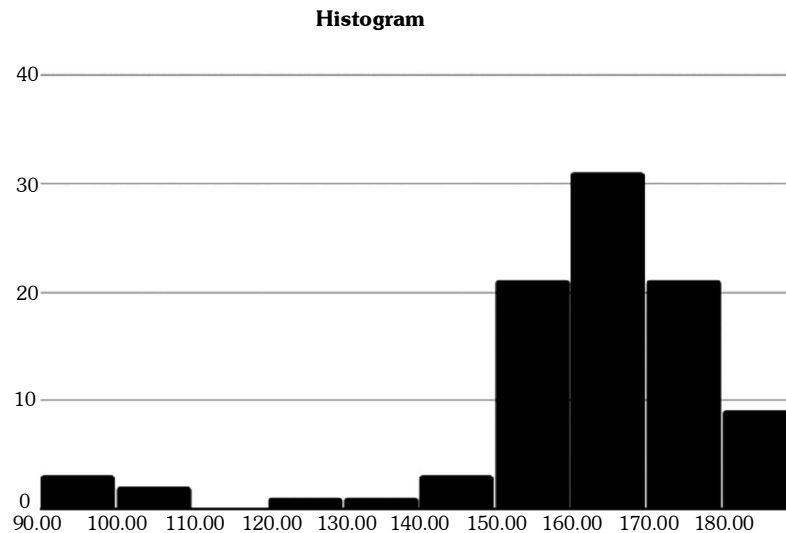
2. Bar Chart



The classic bar chart, or bar graph, is another common and easy-to-use method of data visualization. In this type of visualization, one axis of the chart shows the categories being compared, and the other, a measured value. The length of the bar indicates how each group measures according to the value.

One drawback is that labeling and clarity can become problematic when there are too many categories included. Like pie charts, they can also be too simple for more complex data sets.

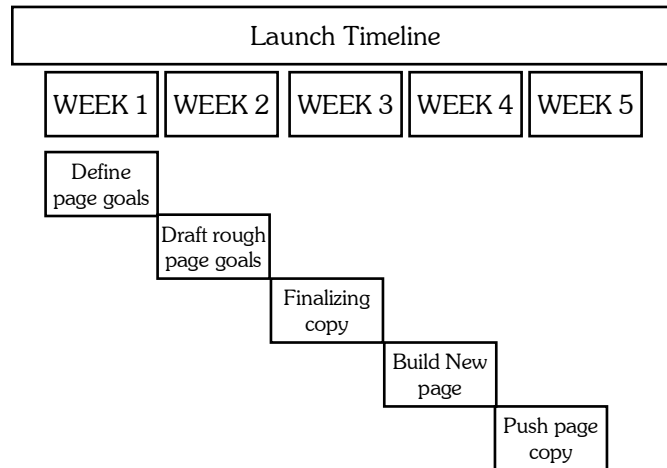
3. Histogram



Unlike bar charts, histograms illustrate the distribution of data over a continuous interval or defined period. These visualizations are helpful in identifying where values are concentrated, as well as where there are gaps or unusual values.

Histograms are especially useful for showing the frequency of a particular occurrence. For instance, if you'd like to show how many clicks your website received each day over the last week, you can use a histogram. From this visualization, you can quickly determine which days your website saw the greatest and fewest number of clicks.

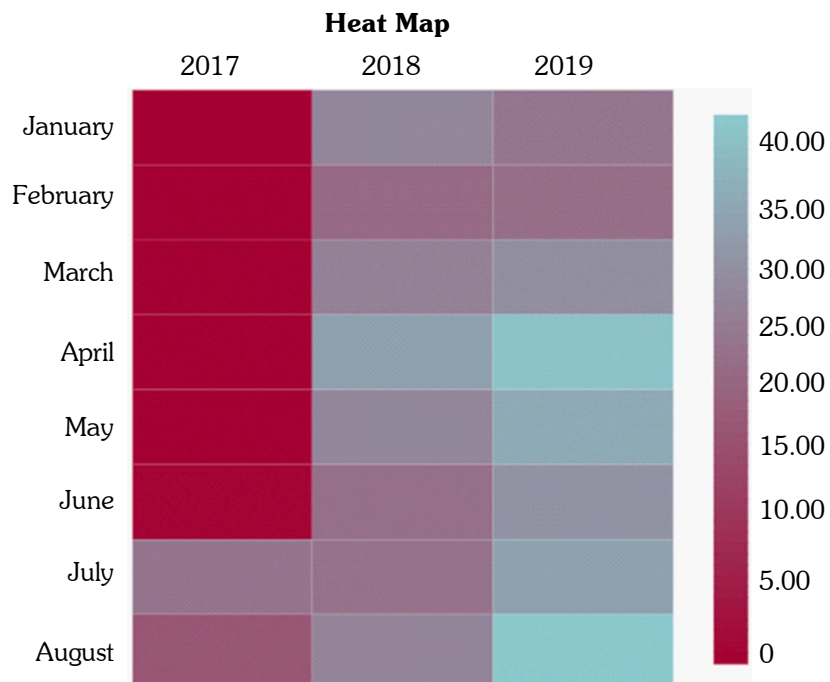
4. Gantt Chart



Gantt charts are particularly common in project management, as they're useful in illustrating a project timeline or progression of tasks. In this type of chart, tasks to be performed are listed on the vertical axis and time intervals on the horizontal axis. Horizontal bars in the body of the chart represent the duration of each activity.

Utilizing Gantt charts to display timelines can be incredibly helpful, and enable team members to keep track of every aspect of a project. Even if you're not a project management professional, familiarizing yourself with Gantt charts can help you stay organized.

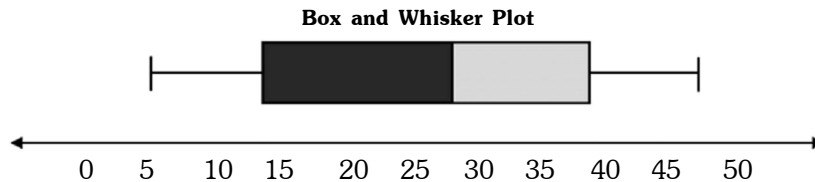
5. Heat Map



A heat map is a type of visualization used to show differences in data through variations in color. These charts use color to communicate values in a way that makes it easy for the viewer to quickly identify trends. Having a clear legend is necessary in order for a user to successfully read and interpret a heatmap.

There are many possible applications of heat maps. For example, if you want to analyze which time of day a retail store makes the most sales, you can use a heat map that shows the day of the week on the vertical axis and time of day on the horizontal axis. Then, by shading in the matrix with colors that correspond to the number of sales at each time of day, you can identify trends in the data that allow you to determine the exact times your store experiences the most sales.

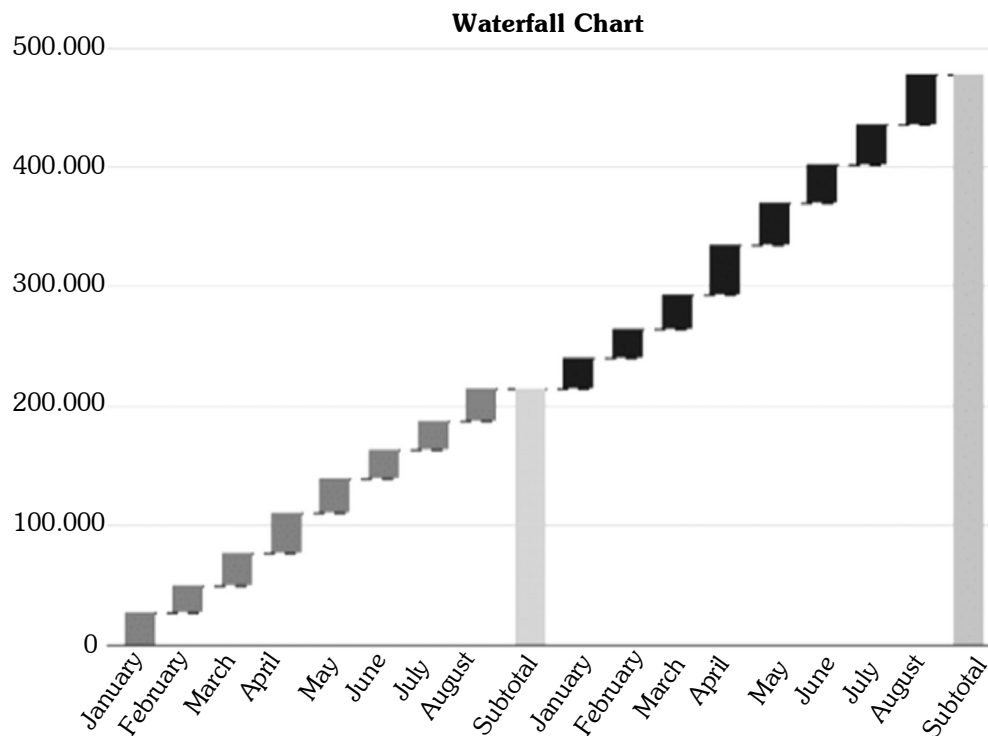
6. A Box and Whisker Plot



A box and whisker plot, or box plot, provides a visual summary of data through its quartiles. First, a box is drawn from the first quartile to the third of the data set. A line within the box represents the median. “Whiskers,” or lines, are then drawn extending from the box to the minimum (lower extreme) and maximum (upper extreme). Outliers are represented by individual points that are in-line with the whiskers.

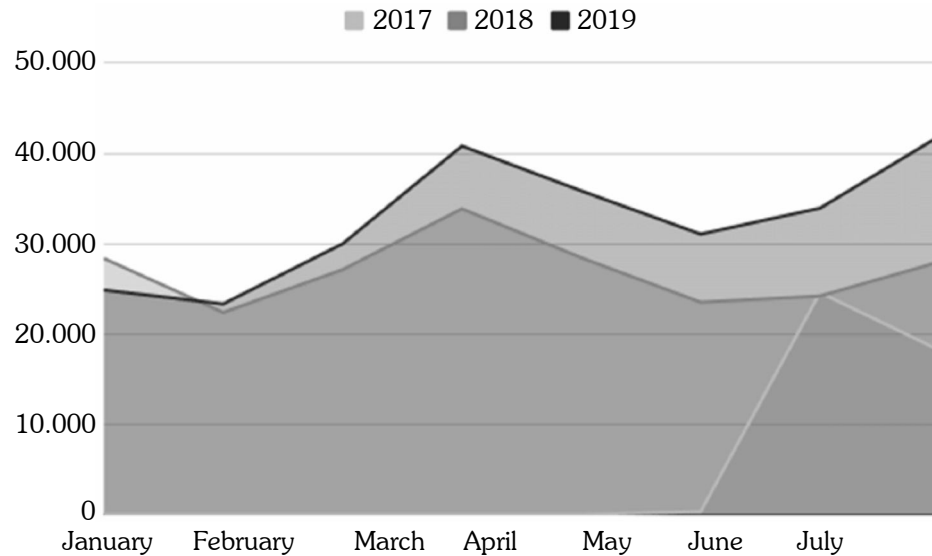
This type of chart is helpful in quickly identifying whether or not the data is symmetrical or skewed, as well as providing a visual summary of the data set that can be easily interpreted.

7. Waterfall Chart



A waterfall chart is a visual representation that illustrates how a value changes as it's influenced by different factors, such as time. The main goal of this chart is to show the viewer how a value has grown or declined over a defined period. For example, waterfall charts are popular for showing spending or earnings over time.

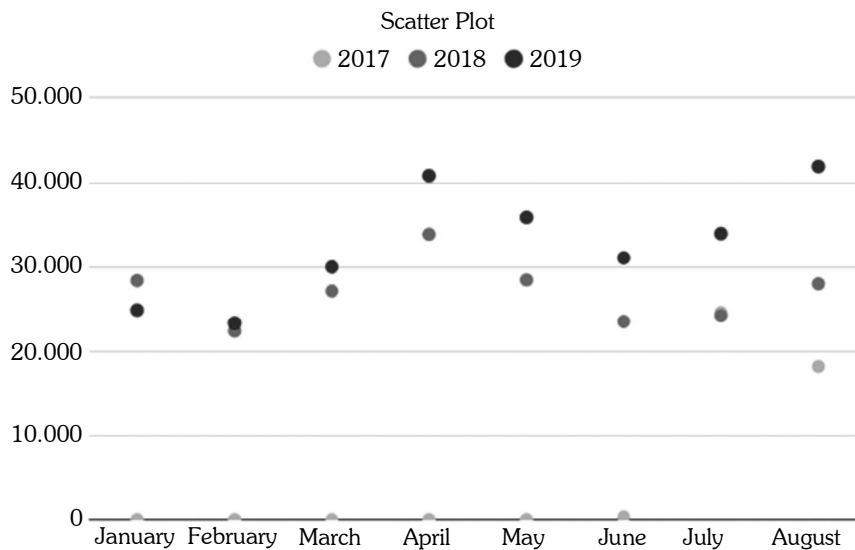
8. Area Chart



An area chart, or area graph, is a variation on a basic line graph in which the area underneath the line is shaded to represent the total value of each data point. When several data series must be compared on the same graph, stacked area charts are used.

This method of data visualization is useful for showing changes in one or more quantities over time, as well as showing how each quantity combines to make up the whole. Stacked area charts are effective in showing part-to-whole comparisons.

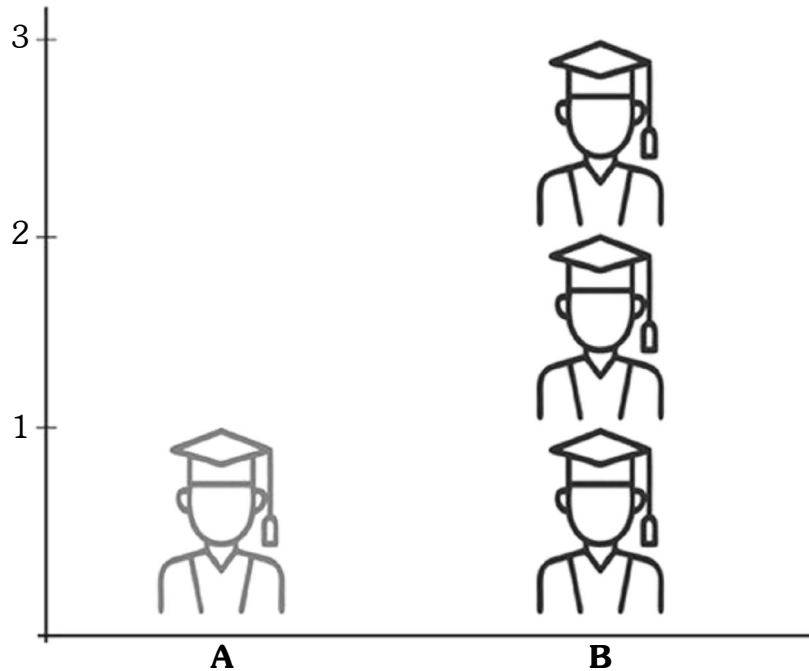
9. Scatter Plot



Another technique commonly used to display data is a scatter plot. A scatter plot displays data for two variables as represented by points plotted against the horizontal and vertical axis. This type of data visualization is useful in illustrating the relationships that exist between variables and can be used to identify trends or correlations in data.

Scatter plots are most effective for fairly large data sets, since it's often easier to identify trends when there are more data points present. Additionally, the closer the data points are grouped together, the stronger the correlation or trend tends to be.

10. Pictogram Chart



Pictogram charts, or pictograph charts, are particularly useful for presenting simple data in a more visual and engaging way. These charts use icons to visualize data, with each icon representing a different value or category. For example, data about time might be represented by icons of clocks or watches. Each icon can correspond to either a single unit or a set number of units (for example, each icon represents 100 units).

In addition to making the data more engaging, pictogram charts are helpful in situations where language or cultural differences might be a barrier to the audience's understanding of the data.

2.3 TABLES, CROSS TABULATIONS, CHARTS, DATA DASHBOARDS USING ADVANCED MS-EXCEL OR SPSS.

Q14. What is mean by table? How can we create table in Excel.

Ans :

(Imp.)

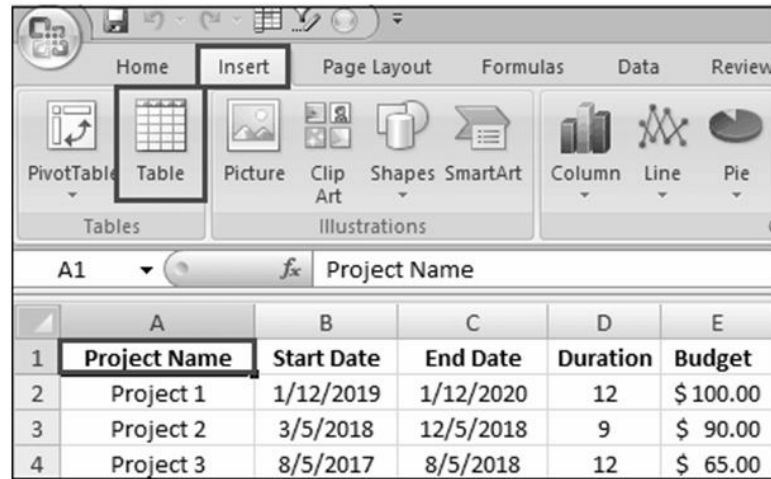
A table is a powerful feature to group your data in Excel. You can consider a table as a specific set of rows and columns in a spreadsheet. You can have multiple tables on the same sheet. **Tables** allow you to analyze your data quickly and easily in **Excel**.

Following are the steps involved to create tables in Excel.

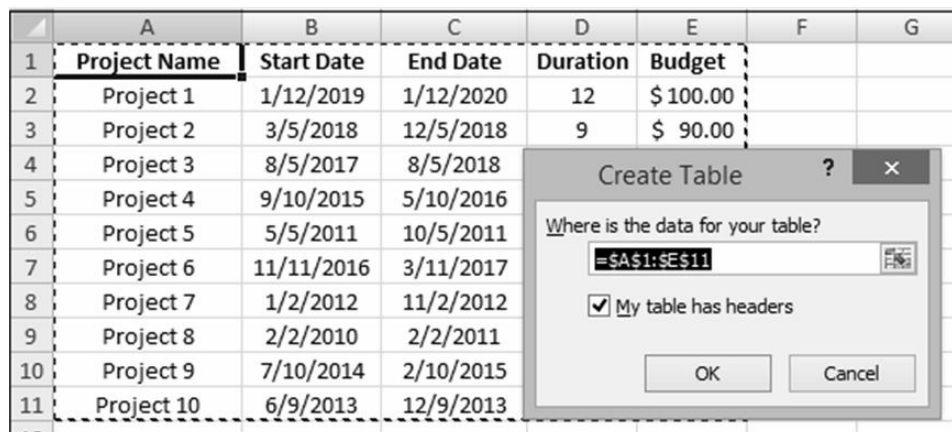
Step 1: Select any cell within your data set.

Step 2: Go on the **Insert** tab and click on the **Table** button in the **Tables** group.

Or you can press the Ctrl + T shortcut key to insert a table.



Step 3: The Create Table dialog box appears with all the data selected for you automatically. You can adjust the range and if you want the first row of data to become the table headers, make sure the My table has headers box is selected.



Step 4: Now click on the OK button.

As a result, Excel converts your range of data into a true table with the default style.

	A	B	C	D	E
1	Project Name	Start Date	End Date	Duration	Budget
2	Project 1	1/12/2019	1/12/2020	12	\$ 100.00
3	Project 2	3/5/2018	12/5/2018	9	\$ 90.00
4	Project 3	8/5/2017	8/5/2018	12	\$ 65.00
5	Project 4	9/10/2015	5/10/2016	8	\$ 87.00
6	Project 5	5/5/2011	10/5/2011	5	\$ 90.00
7	Project 6	11/11/2016	3/11/2017	4	\$ 110.00
8	Project 7	1/2/2012	11/2/2012	11	\$ 66.00
9	Project 8	2/2/2010	2/2/2011	12	\$ 70.00
10	Project 9	7/10/2014	2/10/2015	7	\$ 72.00
11	Project 10	6/9/2013	12/9/2013	6	\$ 95.00
12					

Here are some essential points which need to remember when you create a table in excel.

- Prepare and clean your data before creating a table, remove all blank rows, give each column a unique name, and make sure each row contains information about one record.
- When a table is inserted, Excel retains all formatting that you currently have. For best results, you may want to remove some of the existing formatting, e.g., background colors, so it does not conflict with a table style.
- You are not limited to just one table per sheet. You can have as many as needed. It stands to reason to insert at least one blank row and one blank column between a table and other data for better readability.

Q15. What is cross Tabulation? What are the benefits of cross tabulation.

Ans :

Cross tabulation (crosstab) is a useful analysis tool commonly used to compare the results for one or more variables with the results of another variable. It is used with data on a nominal scale, where variables are named or labeled with no specific order.

Crosstabs are basically data tables that present the results from a full group of survey respondents as well as subgroups. They allow you to examine relationships within the data that might not be obvious when simply looking at total survey responses.

Benefits of cross tabulation

With cross tabulation, you can examine your data in a variety of ways to achieve a deeper understanding of groups within your respondents.

i) Reduce confusion when analyzing data

Analyzing large datasets can be difficult. Finding relevant, actionable insights within large amounts of data is even more complicated. Crosstabs simplify and divide data into subgroups for ease of interpretation—they show percentages and frequencies that may change when contrasted with variables in other categories. By making datasets more manageable at scale, fewer errors will result.

ii) More granular data points

Using crosstabs, you can examine the relationships between one or more variables, which leads to insights on a more granular level. These insights could go unnoticed without crosstabs, lost in a sea of data, or require additional work to reveal. Use multiple filters to dig even deeper into data to uncover more details.

iii) Actionable insights

Using crosstabs simplifies datasets so that you can make quick comparisons between them. This means faster insights for creating new marketing strategies guided by the data. You are also able to watch for global trends across survey responses and take action accordingly.

iv) Clarity of interpretation

When you use crosstabs, datasets are simplified and divided into subgroups. The resulting clean data is in a more digestible format and easily viewed and used by research professionals and team members without analytics training.

Q16. How to create cross tab in Excel?*Ans :***(Imp.)**

A crosstab is a table that summarizes the relationship between two categorical variables.

The following step-by-step example explains how to create a crosstab in Excel.

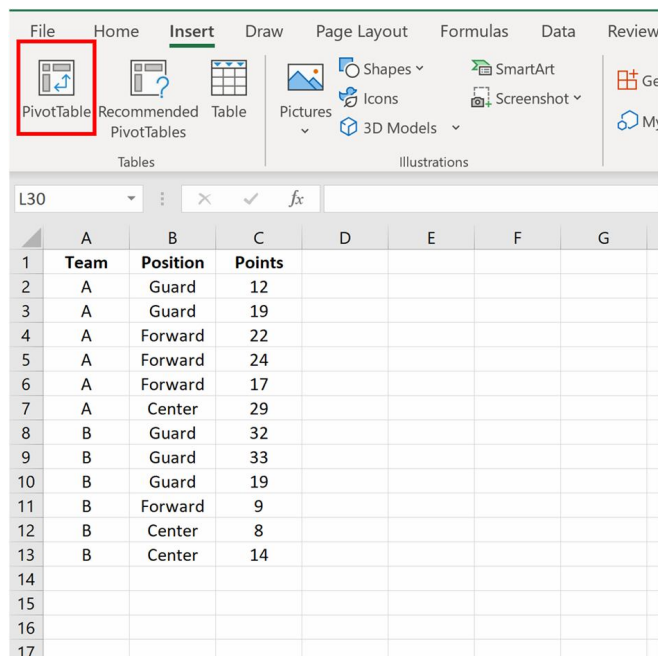
Step 1: Enter the Data

First, let's enter the following dataset into Excel:

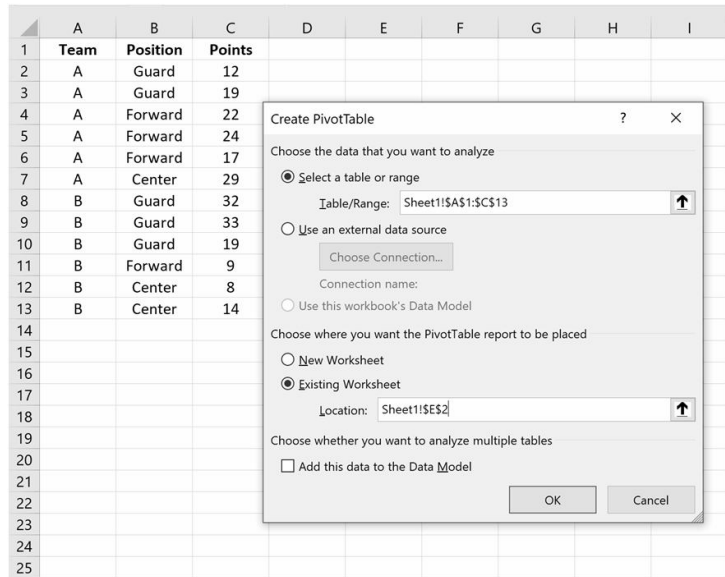
	A	B	C	D	E	F	G
1	Team	Position	Points				
2	A	Guard	12				
3	A	Guard	19				
4	A	Forward	22				
5	A	Forward	24				
6	A	Forward	17				
7	A	Center	29				
8	B	Guard	32				
9	B	Guard	33				
10	B	Guard	19				
11	B	Forward	9				
12	B	Center	8				
13	B	Center	14				
14							
15							
16							
17							
18							
19							
20							
21							
22							

Step 2: Create the Crosstab

Next, click the Insert tab along the top ribbon and then click the PivotTable button.



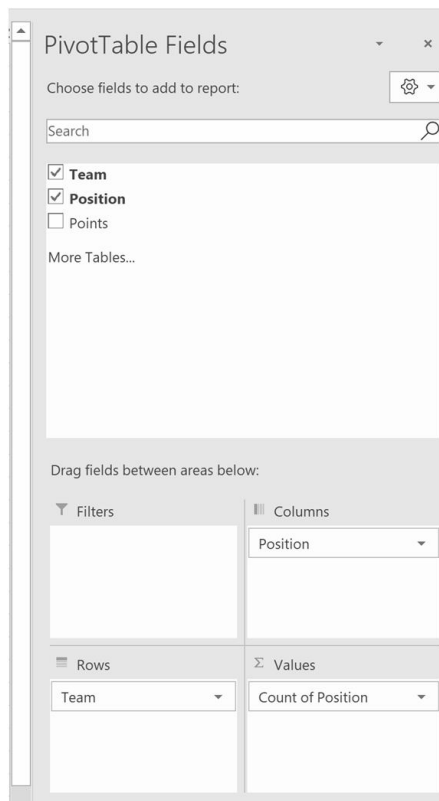
In the new window that appears, select the range that contains the data as the Table/Range and choose any cell you'd like in the Existing Worksheet to place the crosstab. We'll choose cell E2:



Step 3: Populate the Crosstab with Values

Once you click OK, a new window on the right side of the screen will appear.

Drag the Team variable to the Rows area, the Position variable to the Columns area, then the Position variable again to the Values area as follows:



Once you do so, the following crosstab will appear in the cell that you specified:

	A	B	C	D	E	F	G	H	I
1	Team	Position	Points						
2	A	Guard	12		Count of Position	Column Labels			
3	A	Guard	19		Row Labels	Center	Forward	Guard	Grand Total
4	A	Forward	22		A		1	3	2
5	A	Forward	24		B		2	1	3
6	A	Forward	17		Grand Total		3	4	5
7	A	Center	29						
8	B	Guard	32						
9	B	Guard	33						
10	B	Guard	19						
11	B	Forward	9						
12	B	Center	8						
13	B	Center	14						
14									
15									
16									
17									
18									

Step 4: Interpret the Crosstab

Here's how to interpret the values in the crosstab:

Row Totals

- A total of **6** players are on team A
- A total of **6** players are on team B

Column Totals

- A total of **3** players have a position of Center
- A total of **4** players have a position of Forward
- A total of **5** players have a position of Guard

Individual Cells

- **1** player has a position of Center on team A
- **3** players have a position of Forward on team A
- **2** players have a position of Guard on team A
- **2** players have a position of Center on team B
- **1** player has a position of Forward on team B
- **3** players have a position of Guard on team B

Q17. List out various types of charts supported by Excel.

Ans :

Excel provides you different types of charts that suit your purpose. Based on the type of data, you can create a chart. You can also change the chart type later.

Excel offers the following major chart types:

- Column Chart
- Line Chart

- Pie Chart
- Doughnut Chart
- Bar Chart
- Area Chart
- XY (Scatter) Chart
- Bubble Chart
- Stock Chart
- Surface Chart
- Radar Chart
- Combo Chart

Each of these chart types have sub-types. In this chapter, you will have an overview of the different chart types and get to know the sub-types for each chart type.

Column Chart

A Column Chart typically displays the categories along the horizontal (category) axis and values along the vertical (value) axis. To create a column chart, arrange the data in columns or rows on the worksheet.

A column chart has the following sub-types:

- Clustered Column.
- Stacked Column.
- 100% Stacked Column.
- 3-D Clustered Column.
- 3-D Stacked Column.
- 3-D 100% Stacked Column.
- 3-D Column.

Line Chart

Line charts can show continuous data over time on an evenly scaled Axis. Therefore, they are ideal for showing trends in data at equal intervals, such as months, quarters or years.

In a Line chart

- Category data is distributed evenly along the horizontal axis.

- Value data is distributed evenly along the vertical axis.

To create a Line chart, arrange the data in columns or rows on the worksheet.

A Line chart has the following sub-types:

- Line
- Stacked Line
- 100% Stacked Line
- Line with Markers
- Stacked Line with Markers
- 100% Stacked Line with Markers
- 3-D Line

Pie Chart

Pie charts show the size of items in one data series, proportional to the sum of the items. The data points in a pie chart are shown as a percentage of the whole pie. To create a Pie Chart, arrange the data in one column or row on the worksheet.

A Pie Chart has the following sub-types:

- Pie
- 3-D Pie
- Pie of Pie
- Bar of Pie

Doughnut Chart

A Doughnut chart shows the relationship of parts to a whole. It is similar to a Pie Chart with the only difference that a Doughnut Chart can contain more than one data series, whereas, a Pie Chart can contain only one data series.

A Doughnut Chart contains rings and each ring representing one data series. To create a Doughnut Chart, arrange the data in columns or rows on a worksheet.

Bar Chart

Bar Charts illustrate comparisons among individual items. In a Bar Chart, the categories are organized along the vertical axis and the values are organized along the horizontal axis. To create a Bar Chart, arrange the data in columns or rows on the Worksheet.

A Bar Chart has the following sub-types:

- Clustered Bar
- Stacked Bar
- 100% Stacked Bar
- 3-D Clustered Bar
- 3-D Stacked Bar
- 3-D 100% Stacked Bar

Area Chart

Area Charts can be used to plot the change over time and draw attention to the total value across a trend. By showing the sum of the plotted values, an area chart also shows the relationship of parts to a whole. To create an Area Chart, arrange the data in columns or rows on the worksheet.

An Area Chart has the following sub-types:

- Area
- Stacked Area
- 100% Stacked Area
- 3-D Area
- 3-D Stacked Area
- 3-D 100% Stacked Area

XY (Scatter) Chart

XY (Scatter) charts are typically used for showing and comparing numeric values, like scientific, statistical, and engineering data.

A Scatter chart has two Value Axes:

- Horizontal (x) Value Axis
- Vertical (y) Value Axis

It combines x and y values into single data points and displays them in irregular intervals, or clusters. To create a Scatter chart, arrange the data in columns and rows on the worksheet.

Bubble Chart

A Bubble chart is like a Scatter chart with an additional third column to specify the size of the bubbles it shows to represent the data points in the data series.

A Bubble chart has the following sub-types:

- Bubble
- Bubble with 3-D effect

Stock Chart

As the name implies, Stock charts can show fluctuations in stock prices. However, a Stock chart can also be used to show fluctuations in other data, such as daily rainfall or annual temperatures.

To create a Stock chart, arrange the data in columns or rows in a specific order on the worksheet. For example, to create a simple high-low-close Stock chart, arrange your data with High, Low, and Close entered as Column headings, in that order.

A Stock chart has the following sub-types:

- High-Low-Close
- Open-High-Low-Close
- Volume-High-Low-Close
- Volume-Open-High-Low-Close

Surface Chart

A Surface chart is useful when you want to find the optimum combinations between two sets of data. As in a topographic map, colors and patterns indicate areas that are in the same range of values.

To create a Surface chart:

- Ensure that both the categories and the data series are numeric values.
- Arrange the data in columns or rows on the worksheet.

A Surface chart has the following sub-types:

- 3-D Surface
- Wireframe 3-D Surface
- Contour
- Wireframe Contour

Radar Chart

Radar charts compare the aggregate values of several data series. To create a Radar chart, arrange the data in columns or rows on the worksheet.

A Radar chart has the following sub-types:

- Radar
- Radar with Markers
- Filled Radar

Combo Chart

Combo charts combine two or more chart types to make the data easy to understand, especially when the data is widely varied. It is shown with a secondary axis and is even easier to read. To create a Combo chart, arrange the data in columns and rows on the worksheet.

A Combo chart has the following sub-types:

- Clustered Column – Line
- Clustered Column – Line on Secondary Axis
- Stacked Area – Clustered Column
- Custom Combination

Q18. Explain the process of inserting a chart in Excel.

Ans :

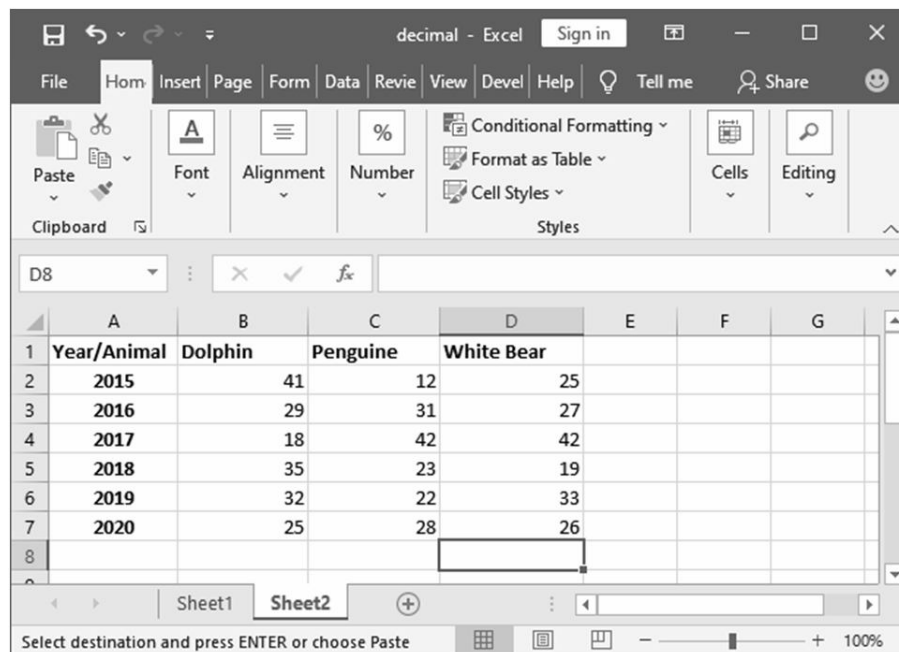
(Imp.)

Excel enables easy to use user interface using which you can easily insert a required chart for your data. You need to follow few simple steps, Excel > Insert tab > Chart section > choose a chart.

Following are the steps to insert a chart in Excel.

Step 1

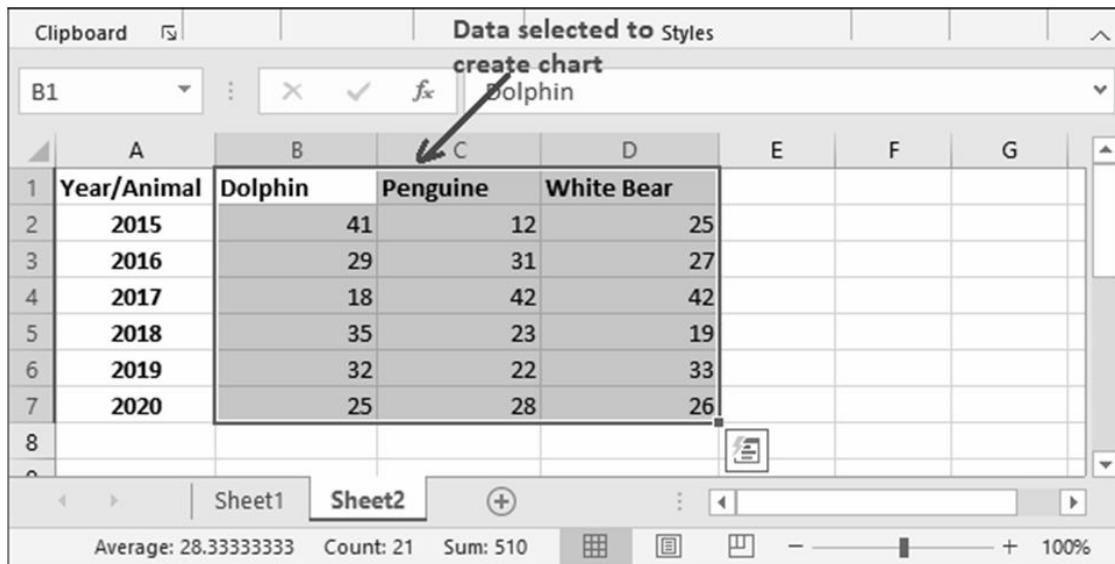
We have the following dataset (Animal population rate for six years from 2015-2020) for which you want to create a chart in Excel.



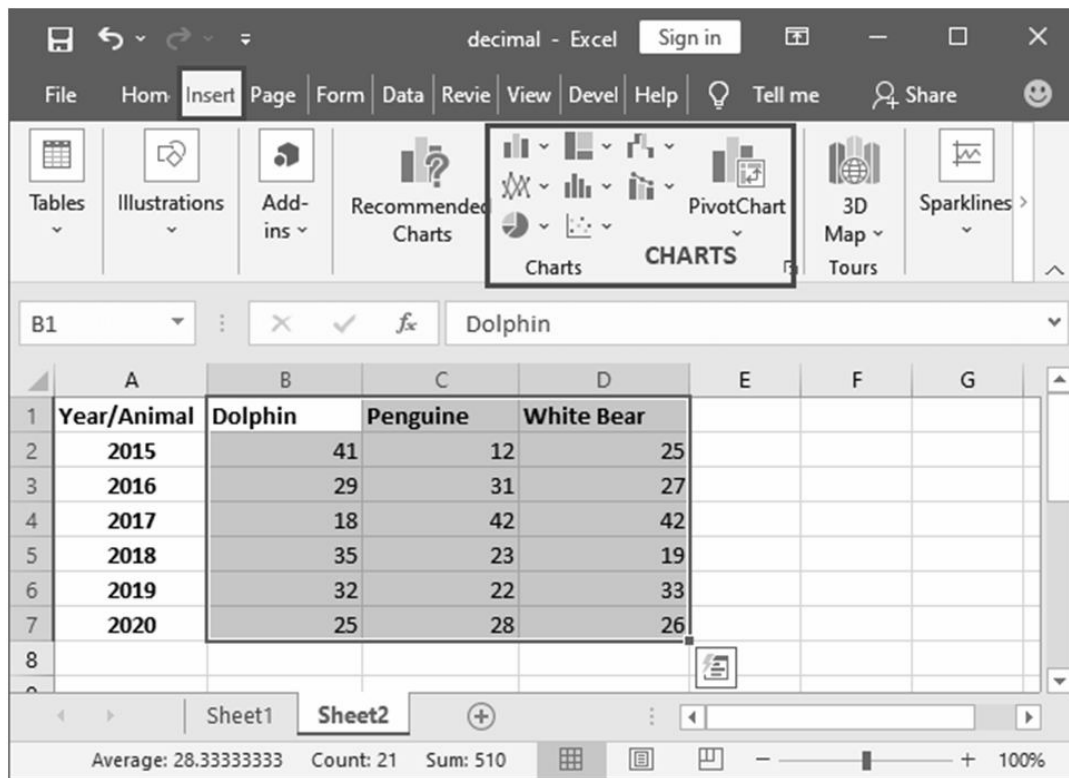
Year/Animal	Dolphin	Penguin	White Bear
2015	41	12	25
2016	29	31	27
2017	18	42	42
2018	35	23	19
2019	32	22	33
2020	25	28	26

Step 2

Select the data, including column header and row label for which you want to create a chart. This data will be the source data for your chart.

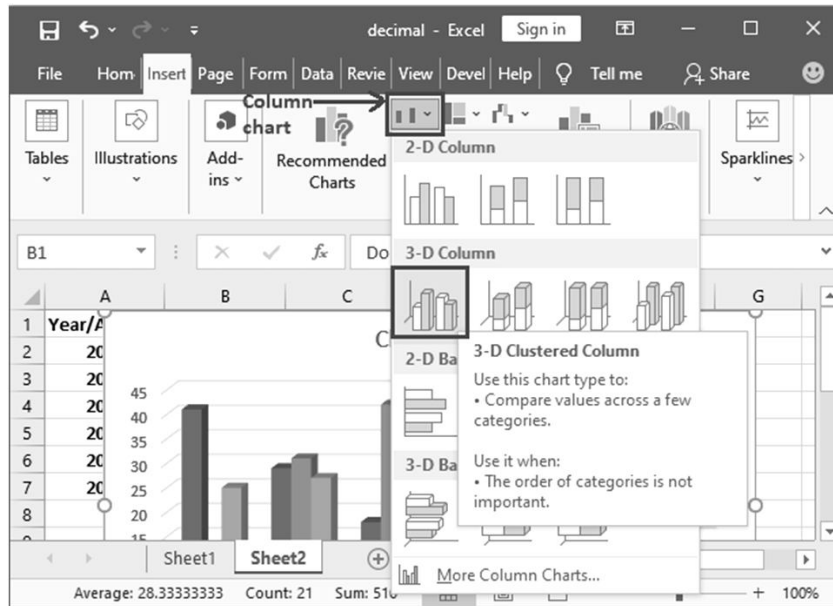
**Step 3**

Navigate to the Insert tab in the Excel header, where you will see a charts section that contains a list of all these charts.

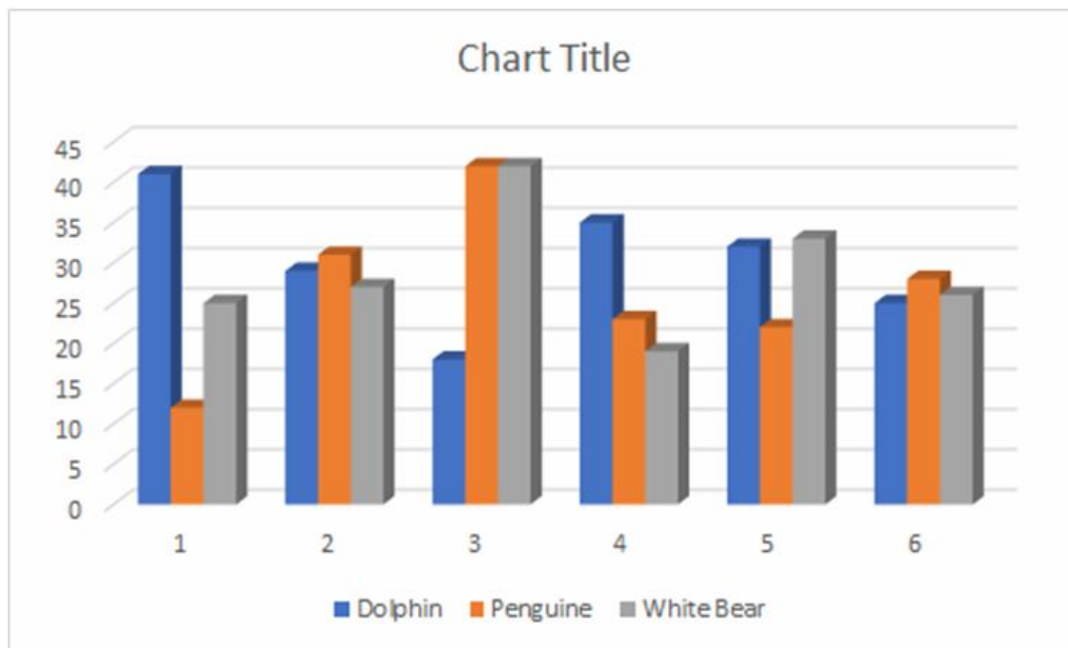


Step 4

Choose a chart from here according to your data. We have chosen a 3D Column chart containing vertical bars for your data.

**Step 5**

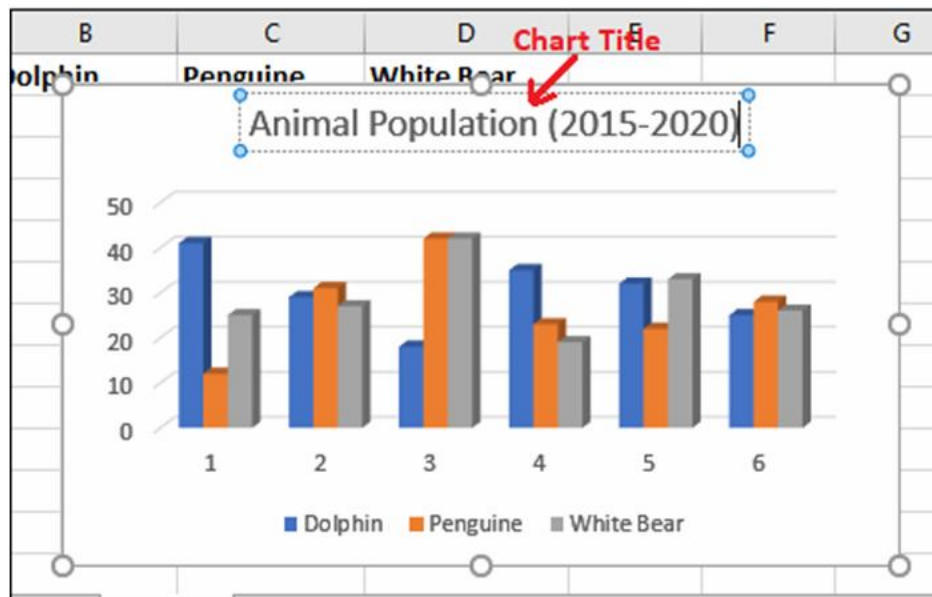
The selected chart is inserted into your Excel worksheet. Initially, the chart looks like this for the data selected in step 2.



Currently, this chart does not have a valid title, clear values for analysis, and more. You can set up all these things in your chart by modifying it.

Step 6

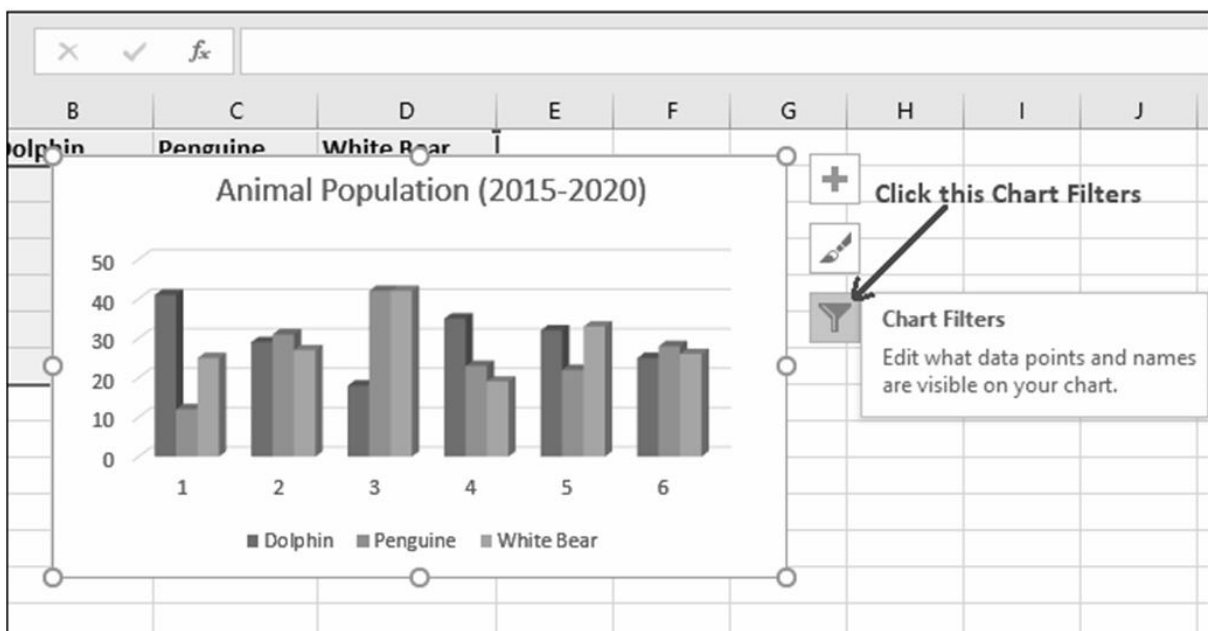
Double-tap on the Chart Title to make it editable and then provide a new valid title that suites to it.



Here, Blue color vertical bar is representing to **Dolphin**, Orange vertical bar to **Penguin**, and Grey vertical bar to **White Bear** population count.

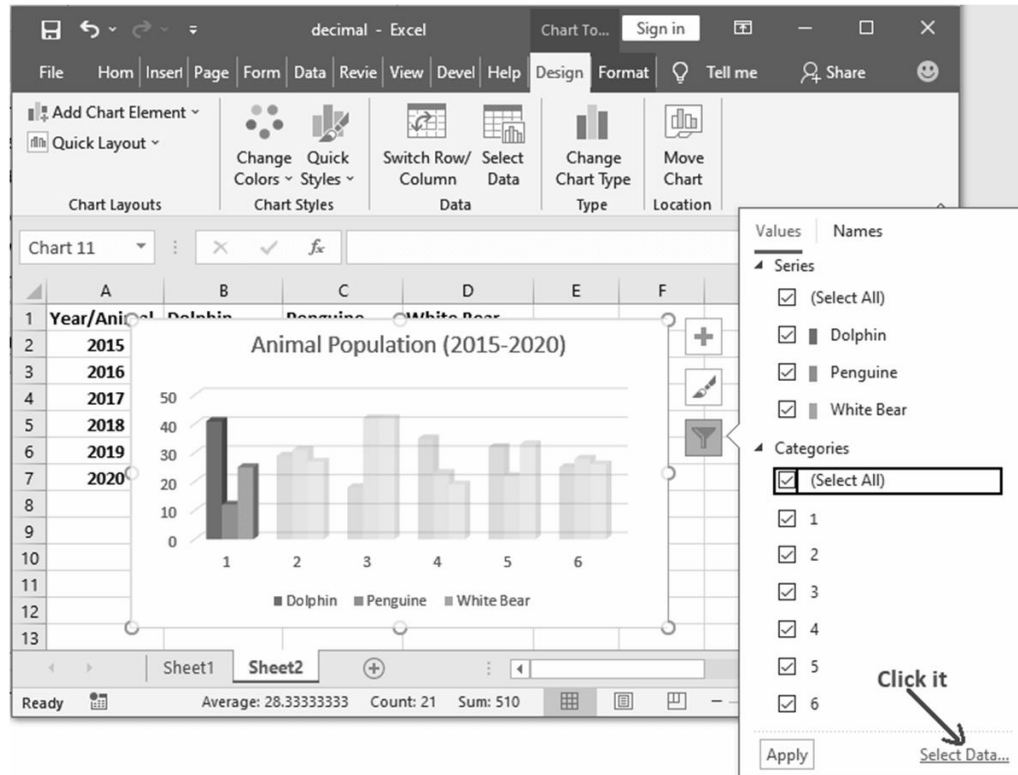
Step 7

You can also define each vertical bar for its year so that the user can easily analyze the values. Click on the **Chart Filters** icon here.

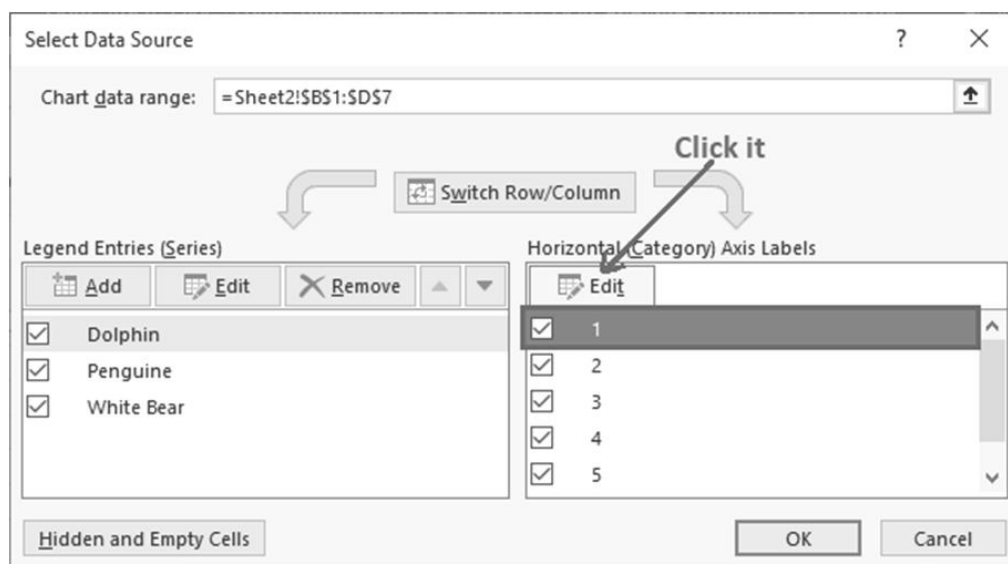


Step 8

Click on the **Select data** present at the bottom of the list to replace the years 2015 for each vertical bar.

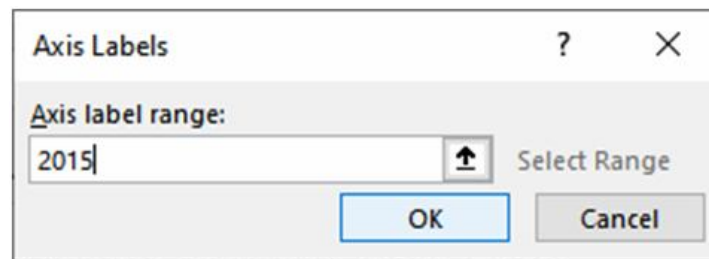
**Step 9**

Here, select the number 1 to replace it with year 2015 and click the Edit button.

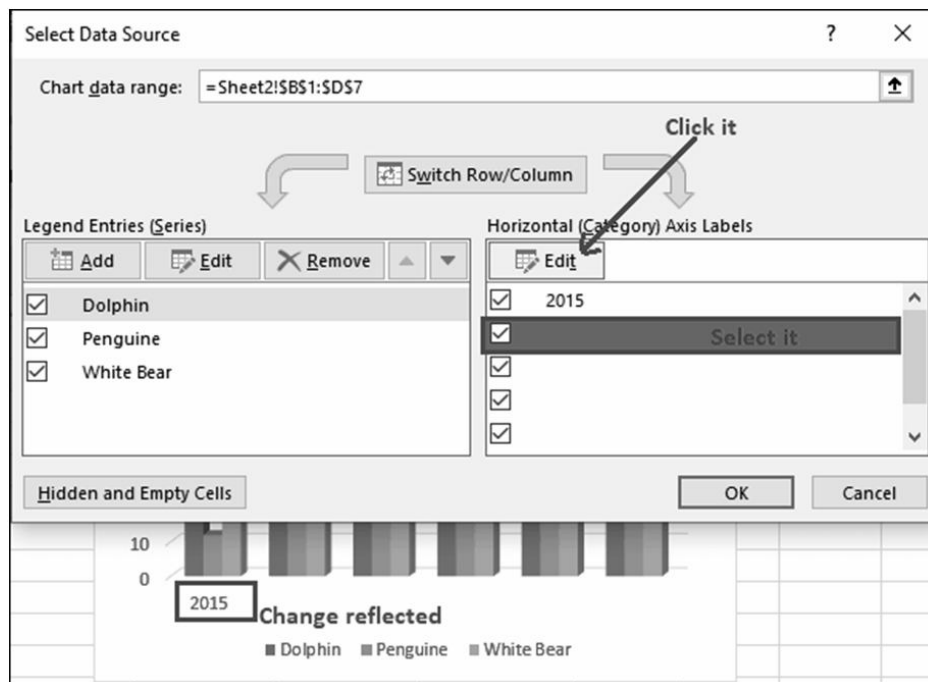


Step 10

Enter the year and click **OK**.

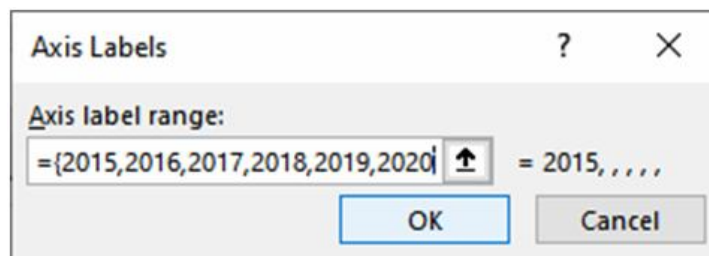
**Step 11**

The year 2015 will immediately reflect on the chart, and all other become blank. Now, to put all other years for each vertical bar, click one more time here.

**Step 12**

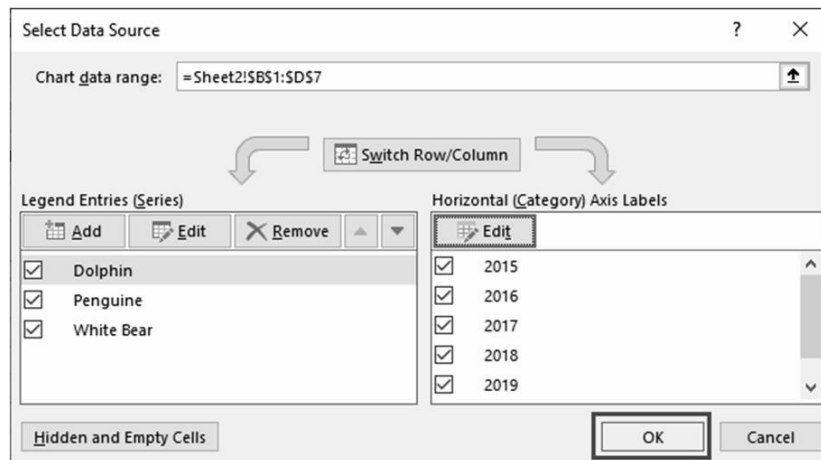
Add more years from 2016 to 2020 inside curly braces separated by a comma and click OK.

{2015,2016,2017,2018,2019,2020}

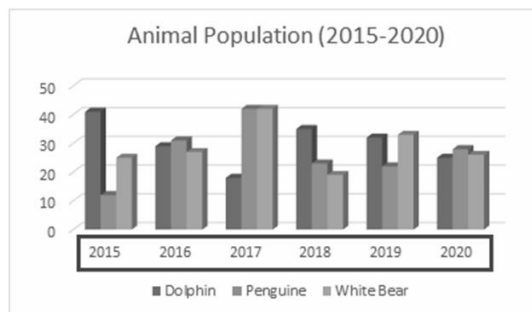


Step 13

All values are now added. So, click OK.

**Step 14**

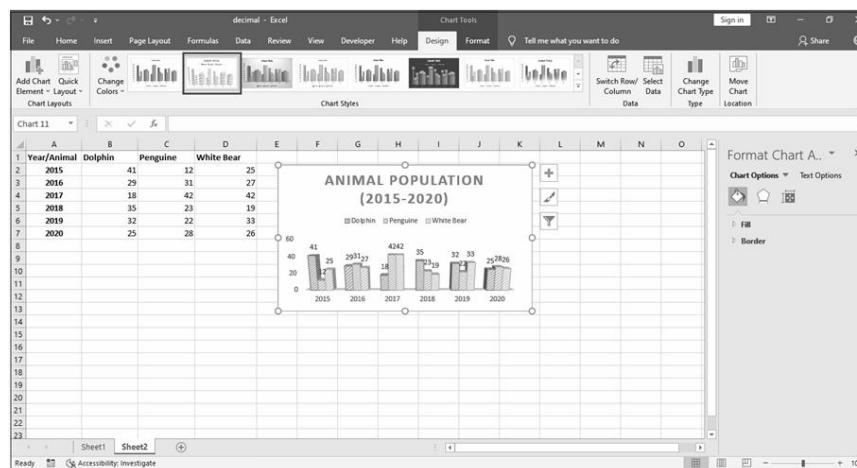
See the charts that the years are reflected on the chart correspond to each vertical bar.

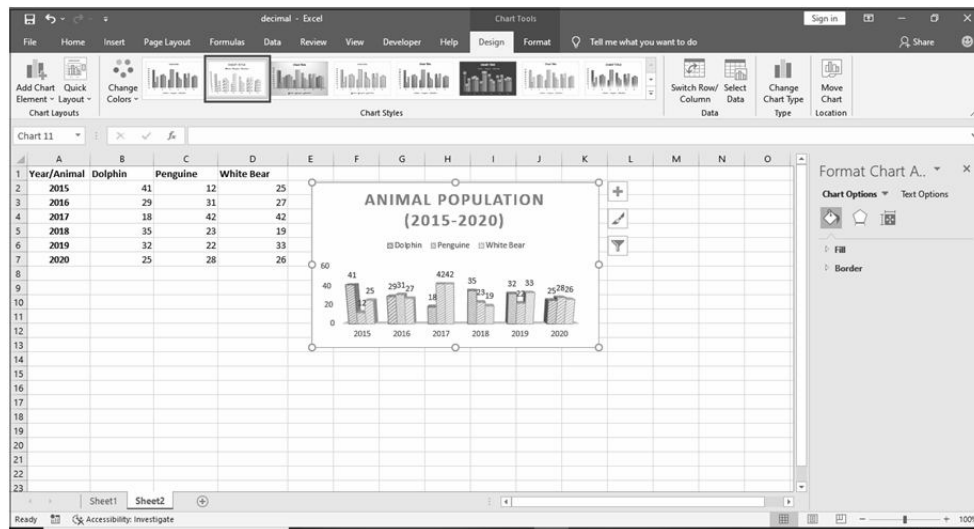


You can see that the exact value is not defined at the end of each bar. Only the graph is showing. Excel enables the users to choose a detailed bar.

Step 15

Choose another chart style for the Column chart for detailed description from the **Chart Style** in the ribbon. We have chosen **Style 2**.





Q19. What is mean by Data Dashboard?

Ans :

A dashboard is a way of displaying various types of visual data in one place. Usually, a dashboard is intended to convey different, but related information in an easy-to-digest form. And oftentimes, this includes things like key performance indicators (KPI)s or other important business metrics that stakeholders need to see and understand at a glance.

Dashboards are useful across different industries and verticals because they're highly customizable. They can include data of all sorts with varying date ranges to help you understand: what happened, why it happened, what may happen, and what action you should take. And since dashboards use visualizations like tables, graphs, and charts, others who aren't as close to the data can quickly and easily understand the story it tells or the insights it reveals.

Q20. What are the uses of Data Dashboard?

Ans :

(Imp.)

The main use of a dashboard is to show a comprehensive overview of data from different sources. Dashboards are useful for monitoring, measuring, and analyzing relevant data in key areas. They take raw data from many sources and clearly present it in a way that's highly tailored to the viewer's needs—whether you're a business leader, line of business analyst, sales representative, marketer, and more.

Use dashboards to measure things like:

- Customer metrics
- Financial information
- Sales information
- Web analytics
- Manufacturing information
- Human resources data
- Marketing performance
- Logistics information

Q21. List out the points which guide in creating data dash board?

Ans :

There are many different solutions to help you build dashboards: Tableau, Excel, or Google Sheets. But at a basic level, here are important steps to help you build a dashboard:

1. **Define your audience and goals:** Ask who you are building this dashboard for and what do they need to understand? Once you know that, you can answer their questions more easily with selected visualizations and data.
2. **Choose your data:** Most businesses have an abundance of data from different sources. Choose only what's relevant to your audience and goal to avoid overwhelming your audience with information.
3. **Double-check your data:** Always make sure your data is clean and correct before building a dashboard. The last thing you want is to realize in several months that your data was wrong the entire time.
4. **Choose your visualizations:** There are many different types of visualizations to use, such as charts, graphs, maps, etc. Choose the best one to represent your data. For example, bar and pie charts can quickly become overwhelming when they include too much information.
5. **Use a template:** When building a dashboard for the first time, use a template or intuitive software to save time and headaches. Carefully choose the best one for your project and don't try to shoehorn data into a template that doesn't work.
6. **Keep it simple:** Use similar colors and styles so your dashboard doesn't become cluttered and overwhelming.
7. **Iterate and improve:** Once your dashboard is in a good place, ask for feedback from a specific person in your core audience. Find out if it makes sense to them and answers their questions. Take that feedback to heart and make improvements for better adoption and understanding.

Q22. Discuss various types of Dashboards.

Ans :

(Imp.)

Following are the types of Dash boards.

i) Business dashboards

Companies can't make solid decisions without data, which is where business dashboards come into play. They can host all kinds of different data, from sales, finance, management, marketing, human resources, and more. They're designed to give managers and directors the data needed to make strategic plans and refine ideas.

ii) Executive dashboards

An executive dashboard is a specific type of business dashboard meant to visualize crucial metrics for the executive team. Usually, the data is high level, but gives leaders transparency into critical business activity and performance to help them make more informed decisions, better plan, and assess effectiveness.

iii) KPI dashboards

Arguably one of the most important is the key performance indicator (KPI) dashboard—used by subject matter experts, executives, or laypeople. They visually display the performance of key data points at a glance, revealing progression toward key goals. The most important part of a KPI dashboard is to know what your KPIs are and the best way to measure them.

iv) Project Dashboards

When running and/or managing a large project, this dashboard is a useful tool to track its progress and share that with your team and other key stakeholders. It offers a complete view of the project status, insights, and main data.

v) Performance dashboards

The versatile performance dashboard can track everything from overall business performance to the performance of individual campaigns. It's useful for marketing, finance, advertising, human resources, and other business groups.

vi) Website dashboards

When tracking site performance, creating a website dashboard is useful. It tracks data like overall traffic, total users, active users, e-commerce activity, sales, and revenue. Whether your organization maintains a simple or more complex site, this dashboard offers an integrated, clear view of your metrics.

vii) Operations dashboards

This is a common type of business dashboard. Unlike the high-level dashboards previously mentioned, these are hyper-focused on helping you run the business day-to-day and give users an end-to-end view of daily operations.

viii) Industry dashboards

Since dashboards are versatile and customizable to the needs of the user and business, they're a common tool across different industries. Industries that rely heavily on data analysis to make decisions (e.g., healthcare, sales, and marketing) use these to help with their decision-making and problem-solving.

Here are some common industry dashboards:

i) Healthcare Dashboards

The healthcare industry deals with large amounts of critical data: hospital admission and discharge rates, costs, staff allocations, insurance claims, appointment attendance, no-show rates, and more. Healthcare dashboards keep that information accessible, understandable, and secure so clinicians, office administrators, and other healthcare staff can focus on improving patient outcomes.

ii) Marketing Dashboards

Marketers of all levels deal with an overabundance of data due to the complex nature of online tracking and analytics. That's why they rely on dashboards to streamline analysis and discover key insights that help enhance campaigns or performance. These dashboards include data such as return on investment (ROI), churn rate, retention, lead numbers, cost per lead, revenue, goal completions, and more.

iii) Retail Dashboards

E-commerce and brick-and-mortar stores deal with complex data about inventory and profit, which is why many managers and owners utilize retail dashboards. They commonly house data such as the number of sales, net profit, inventory, foot traffic, or employee turnover and performance to help retailers understand areas such as performance, customer engagement, and how to improve service.

iv) Sales Dashboards

The sales process is complex with many steps and people involved. A sales dashboard can vary depending on this process and the main KPIs, but oftentimes includes data like the number of leads, open cases, opportunities, contact, closed deals, lost opportunities, and revenue, which reveals to sales staff if they're fulfilling, exceeding, or underperforming against different goals.

Q22. List and explain Dashboard best practices.

Ans :

Best practices for building and using dashboards vary across companies (and person to person). Keep in mind the goal you hope to achieve with the dashboard and who will look at it, as well as these considerations for creating a good dashboard:

- 1. KPIs:** Don't overwhelm your audience with data. Choose only the most relevant data for and present it in a way that makes sense.
- 2. Elements:** Ensure you choose the correct charts, graphs, and tables for each piece of data. The best visual enhances understanding.

3. **Design:** Make sure your dashboard is easy to understand at a glance by organizing the data and using a consistent color scheme.
4. **Labels:** Be concise and clearly label every piece of information.
5. **Interactivity:** Use interactive elements as needed. This allows people to drill further into data or shows variability.

Q23. How can we create Data Dashboard using Excel.

Ans :

(Imp.)

1. Create a layout for your Excel Dashboard

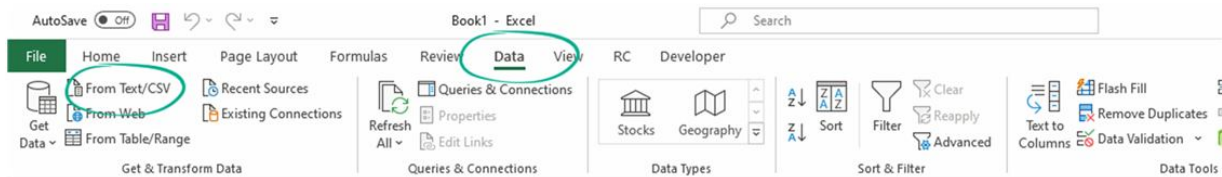
Create a proper draft! You can use paper and pencil, but we prefer Microsoft Excel to create mockups. We have used simple, grouped shapes.

The parts of the workbook structure: Mostly, you use three worksheets for an Excel dashboard.

- **Data:** you can store the raw data tables here
- **Dashboard Tab:** the main dashboard Worksheet
- **Calculation:** make the calculations on this Worksheet

2. Get your data into Excel

- To create an Excel Dashboard, you need to choose data sources. If the data is in Excel, you are lucky and can jump to the next step. If not, you have to use external data sources.



- Go to the **Data tab** and **pick one of the import options**. It's easy to import data into an Excel workbook. In the example, you are **using a CSV file** to create the initial dataset for our dashboard.

3. Clean Raw Data

Our raw data is in Excel. **Now you can start the data cleansing process. There are many tricks to clean and consolidate data.**

- Sort data to see extremes and peaks
- **Remove duplicates** to avoid errors
- Change the text to lower, upper, or proper case
- Remove leading and trailing spaces

How do we remove leading and trailing spaces from raw data? First, go to the Formula bar and **apply the TRIM function**. Now copy the formula down. Finally, use **data cleansing add-ins** to avoid issues and clean your data faster and easier.

Use an Excel Table and Filter Data

You don't have cleaned input in this phase, but you already have data on a worksheet. What will be the next step?

First, you must **check if the required information is in a tabular format**. The tabular format means that every data point lives in one cell, for example, the city's name, address, or phone number. If it is in a tabular format, you should convert it into an **Excel table and select the data range**.

Choose a table or use the **insert table shortcut, Ctrl + T, on the Insert Tab**.

The screenshot shows an Excel worksheet with a data range from D2 to J18. A 'Create Table' dialog box is open, asking 'Where is the data for your table?'. The range '\$D\$1:\$I\$18' is entered, and the checkbox 'My table has headers' is checked. The background data is as follows:

OrderDate	Region	Rep	Item	Units	Unit Cost	Total
1/6/19	East	Jones	Pencil	95	1.99	189.05
1/23/19	Central	Kivell	Binder	50	19.99	999.50
2/9/19	Central	Jardine	Pencil	36	4.99	179.64
2/26/19	Central	Gill	Pen	27	19.99	539.73
3/15/19	West	Sorvino	Pencil	56	2.99	167.44
4/1/19	East	Jones	Binder	60	4.99	299.40
4/18/19	Central			75	1.99	149.25
5/5/19	Central			90	4.99	449.10
5/22/19	West			32	1.99	63.68
6/8/19	East			60	8.99	539.40
6/25/19	Central			90	4.99	449.10
7/12/19	East			29	1.99	57.71
7/29/19	East			81	19.99	1,619.19
8/15/19	East			35	4.99	174.65
9/1/19	Central	Smith	Desk	2	125.00	250.00
9/18/19	East	Jones	Pen Set	16	15.99	255.84
10/5/19	Central	Morgan	Binder	28	8.99	251.72

In this case, we don't have headers. Excel will automatically insert headers into the first row. If you need more data, you can only expand the table and not lose the formulas.

5. Analyze, Organize, Validate and Audit your Data

You took you through the method that converts raw data into a structure capable of creating a dashboard.

You can use **Excel formulas** and various methods to help us move forward. However, it would help if you had creativity rather than knowing all the formulas to make a useful dashboard.

So you'll use these **functions** and tools to build the Excel Dashboard in Excel: **XLOOKUP, IF, SUMIF, COUNTIF, ROW, NAME MANAGER**.

Excel grants great auditing tools to help you find and fix Workbook or Worksheet issues.

Use Microsoft Excel Inquiry to visualize which cells in your Worksheet contribute to a formula error. This step should cut down the time spent on the usual validation procedures.

The screenshot shows the 'Workbook Analysis Report' window for the file 'Excel-project-dashboard2.xlsx'. The 'Items' list on the left includes 'Summary', 'Workbook', and 'Formulas'. Under 'Formulas', 'With errors (1)' is selected. The 'Results' table on the right shows the following data:

Item	Sheet Name	Cell Address	Formula	Value
1	calc	C7	=INDEX(\$B\$49:\$B\$100,MATCH(C6,\$...	#N/A

At the bottom, there is an 'Export' section with an 'Item detail limit' of 250000 and an 'Excel Export...' button.

So, before we start creating a chart, you have to validate the data.

If you want to analyze your data quickly, use the **Quick Analysis Tool**.

6. Choose the right chart type for your Excel dashboard

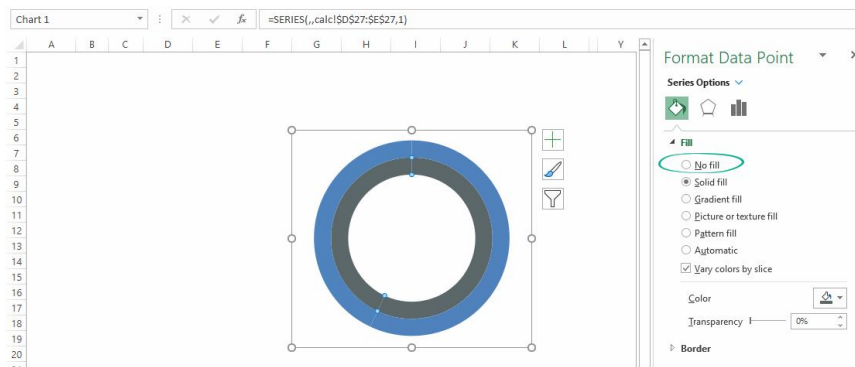
Now you have an organized, cleaned, and error-free data set; it's time to **choose the proper chart**.

7. Select the data and build your chart

We have cleaned and grouped data in this phase and just picked the chart or graph for the data. It's time to select the data! As you learned, **the combo chart requires two doughnut charts** and a simple formula.

Select the **'Calculation'** tab (which contains filtered data and calculated fields). Highlight the range of what you want to display. In the example, you use two values to show the **Acceptance Ratio**.

The actual value comes from the **Data tab**. After that, then calculate the **reminder value** using this simple formula. **In this case, 75%**. Next, select the 'Calculation' tab. Cell E23 will show the actual value. The second cell, **E24, contains a simple formula and displays the remainder value as 100%.**



Make sure that the value in the source cell is in percentage format! Okay, now we select the 'Actual Value' and 'Reminder Value' data. Next, open the 'Insert Chart' dialog to create a custom combo chart to preview and choose different chart types. Furthermore, you can move the data series to the secondary axis.

Select the inserted chart and press Control + C to duplicate the chart.

8. Improve your charts

Now you have a chart that fits your data. It looks great, but you can improve your Excel dashboard to the next level! First, clean up the chart to remove the background, title, and borders from the chart area. Next, select the reminder value section of the outer ring.

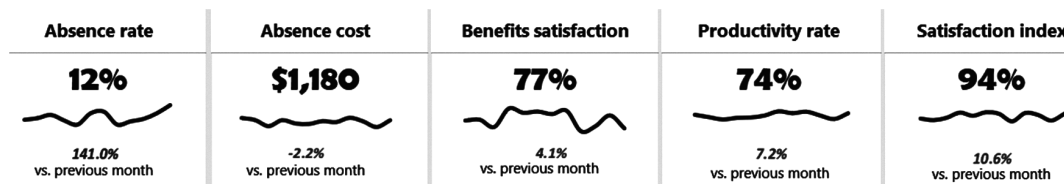


Right-click, then choose **Format Data Point**. Use the **'No fill'** option. Let's see the inner ring, select the actual value section, and apply the 'No fill' option. Adjust the doughnut hole size if you want. Insert a Text Box and remove the background and border.

9. Create a Dashboard Scorecard

Your Excel dashboard is almost ready. You need only a few components to create a scorecard:

- Label,
- Actual value,
- Annual trendline,
- Variance (between the selected and the last month)



Because you need a little bit more space, merge the cells. Select the cells to place the components and click the **'Merge cells'** button. Now, link the label name from the 'Data' sheet. If you change the name of the value on the 'Data' sheet, the widget label will reflect it. Now link the data from the 'Data' sheet to a 'Dashboard' sheet.

Go to the formula tab, enter an equal sign, and select the 'Data' sheet value. Next, use yearly Data on the 'Data' sheet and insert a line chart to create a trendline. To highlight the variance, use a little trick. Go to the 'Calculation' sheet and create a helper table.

Short Question and Answers

1. What is descript analytics?

Ans :

Descriptive analytics is the process of using current and historical data to identify trends and relationships. It's sometimes called the simplest form of data analysis because it describes trends and relationships but doesn't dig deeper.

Descriptive analytics is relatively accessible and likely something your organization uses daily. Basic statistical software, such as Microsoft Excel or data visualization tools, such as Google Charts and Tableau, can help parse data, identify trends and relationships between variables, and visually display information.

Descriptive analytics is especially useful for communicating change over time and uses trends as a springboard for further analysis to drive decision-making.

2. What is descript analytics?

Ans :

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

There are 3 main types of descriptive statistics:

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** or dispersion concerns how spread out the values are.

3. What is mean? Give an example.

Ans :

Mean

The **mean**, or M , is the most used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called N .

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	$15 + 3 + 12 + 0 + 24 + 3 = 57$
Total number of responses	$N = 6$
Mean	Divide the sum of values by N to find M : $57/6 = 9.5$

4. What is median? Explain with a simple example.*Ans :***Median**

The **median** is the value that's exactly in the middle of a data set.

To find the median, order each response value from the smallest to the biggest. Then, the median is the number in the middle. If there are two numbers in the middle, find their mean.

Median number of library visits

Ordered data set 0, 3, 3, 12, 15, 24

Middle numbers 3, 12

Median Find the mean of the two middle numbers: $(3 + 12)/2 = 7.5$

5. Write the steps to calculate standard deviation.*Ans :*

The standard deviation (s or SD) is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by $N - 1$.
6. Find the square root of the number you found.

6. What is Frequency distribution? What are the types of frequency distribution?*Ans :*

The frequency of a value is the number of times it occurs in a dataset. A frequency distribution is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions

There are four types of frequency distributions:

- **Ungrouped frequency distributions:** The number of observations of each **value** of a variable.
 - You can use this type of frequency distribution for categorical variables.
- **Grouped frequency distributions:** The number of observations of each **class interval** of a variable. Class intervals are ordered groupings of a variable's values.
 - You can use this type of frequency distribution for quantitative variables.
- **Relative frequency distributions:** The proportion of observations of each value or class interval of a variable.

- You can use this type of frequency distribution for **any type of variable** when you're more interested in **comparing frequencies** than the actual number of observations.
- **Cumulative frequency distributions:** The sum of the frequencies less than or equal to each value or class interval of a variable.
- You can use this type of frequency distribution for ordinal or quantitative variables when you want to understand how often observations fall below certain values.

7. What is Normal Distribution?

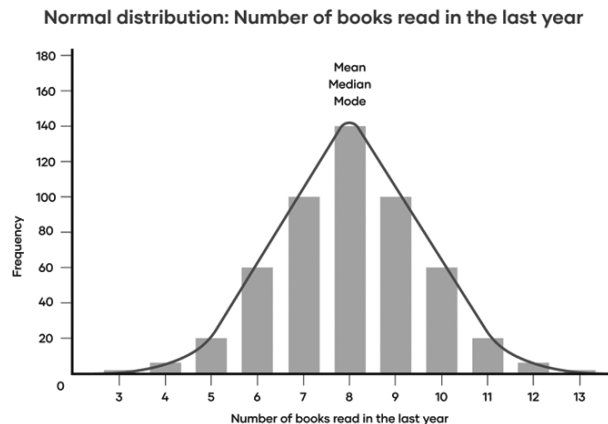
Ans :

Normal distribution

In a normal distribution, data is symmetrically distributed with no skew. Most values cluster around a central region, with values tapering off as they go further away from the center. The mean, mode and median are exactly the same in a normal distribution.

Example: Normal distribution You survey a sample in your local community on the number of books they read in the last year.

A histogram of your data shows the frequency of responses for each possible number of books. From looking at the chart, you see that there is a normal distribution.



The mean, median and mode are all equal; the central tendency of this dataset is 8.

8. Write the sample variance formula.

Ans :

Sample variance

When you collect data from a sample, the sample variance is used to make estimates or inferences about the population variance.

The sample variance formula looks like this:

Formula

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

Explanation

- S^2 = sample variance
- Σ = sum of....
- X = each value
- \bar{x} = sample mean
- n = number of values in the population

With samples, we use $n - 1$ in the formula because using n would give us a biased estimate that consistently underestimates variability. The sample variance would tend to be lower than the real variance of the population.

Reducing the sample n to $n - 1$ makes the variance artificially large, giving you an unbiased estimate of variability: it is better to overestimate rather than underestimate variability in samples.

9. What is data visualization?

Ans :

Data visualization is the representation of information and data using charts, graphs, maps, and other visual tools. These visualizations allow us to easily understand any patterns, trends, or outliers in a data set.

Data visualization also presents data to the general public or specific audiences without technical knowledge in an accessible manner. For example, the health agency in a government might provide a map of vaccinated regions.

The purpose of data visualization is to help drive informed decision-making and to add colorful meaning to an otherwise bland database.

10. Benefits of data visualization

Ans :

Data visualization can be used in many contexts in nearly every field, like public policy, finance, marketing, retail, education, sports, history, and more. Here are the benefits of data visualization:

- **Storytelling:** People are drawn to colors and patterns in clothing, arts and culture, architecture, and more. Data is no different colors and patterns allow us to visualize the story within the data.
- **Accessibility:** Information is shared in an accessible, easy-to-understand manner for a variety of audiences.

- **Visualize relationships:** It's easier to spot the relationships and patterns within a data set when the information is presented in a graph or chart.
- **Exploration:** More accessible data means more opportunities to explore, collaborate, and inform actionable decisions.

11. What is mean by cross tabulation?

Ans :

Cross tabulation (crosstab) is a useful analysis tool commonly used to compare the results for one or more variables with the results of another variable. It is used with data on a nominal scale, where variables are named or labeled with no specific order.

Crosstabs are basically data tables that present the results from a full group of survey respondents as well as subgroups. They allow you to examine relationships within the data that might not be obvious when simply looking at total survey responses.

12. What is mean by Data Dashboard?

Ans :

A dashboard is a way of displaying various types of visual data in one place. Usually, a dashboard is intended to convey different, but related information in an easy-to-digest form. And oftentimes, this includes things like key performance indicators (KPI)s or other important business metrics that stakeholders need to see and understand at a glance.

Dashboards are useful across different industries and verticals because they're highly customizable. They can include data of all sorts with varying date ranges to help you understand: what happened, why it happened, what may happen, and what action you should take. And since dashboards use visualizations like tables, graphs, and charts, others who aren't as close to the data can quickly and easily understand the story it tells or the insights it reveals.

Choose the Correct Answers

1. Descriptive analytics: [b]
 - (a) can predict risk and find relationships in data not readily apparent with traditional analyses.
 - (b) helps companies classify their customers into segments to develop specific marketing campaigns.
 - (c) helps detect hidden patterns in large quantities of data to group data into sets to predict behavior.
 - (d) can use mathematical techniques with optimization to make decisions that take into account the uncertainty in the data.
2. A manager at Gampco Inc. wishes to know the company's revenue and profit in its previous quarter. Which of the following business analytics will help the manager? [c]
 - (a) prescriptive analytics
 - (b) normative analytics
 - (c) descriptive analytics
 - (d) predictive analytic
3. What is a database? [a]
 - (a) a collection of related files
 - (b) simply a collection of data
 - (c) a data file holding a single file
 - (d) flat files used to store data
4. Roger wants to compare values across categories using vertical rectangles. Which of the following charts must Roger use? [b]
 - (a) Line chart
 - (b) Clustered column chart
 - (c) Pie chart
 - (d) Stacked column chart
5. Which of the following charts provides a useful means for displaying data over time? [d]
 - (a) Scatter chart
 - (b) A doughnut chart
 - (c) Pie chart
 - (d) Line chart
6. Observations consisting of pairs of variable data are required to construct a _____ chart. [b]
 - (a) doughnut
 - (b) scatter
 - (c) radar
 - (d) line
7. A graphical depiction of a frequency distribution for numerical data in the form of a column chart is called a _____. [c]
 - (a) cartogram
 - (b) correlogram
 - (c) histogram
 - (d) dendogram
8. Which of the following represents the proportion of the total number of observations that fall at or below the upper limit of each group? [d]
 - (a) Percentile
 - (b) Pareto chart
 - (c) Frequency distribution
 - (d) Cumulative relative frequency

9. Which of the following is true about cross-tabulation? [a]
- (a) All subcategories together must constitute the complete data set.
 - (b) A cross-tabulation table is often called a latent class model.
 - (c) Each observation can be classified into many subcategories.
 - (d) The table displays the number of categorical variables between two observations.
10. Which of the following can be used to quickly create cross-tabulations? [c]
- (a) Frequency distribution
 - (b) COUNTIF function
 - (c) Pivot Table
 - (d) Sort & Filter

Fill in the Blanks

1. The difference between the first and third quartiles is referred to as the _____.
2. The _____ measures the degree of asymmetry of observations around the mean.
3. In statistics, _____ refers to the peakedness or flatness of a histogram.
4. _____ is a measure of the linear association between two variables, X and Y.
5. _____ is a measure of the linear relationship between two variables, X and Y, which does not depend on the units of measurement.
6. The _____ of a random variable corresponds to the notion of the mean, or average, for a sample.
7. _____ involves selecting items from a population so that every subset of a given size has an equal chance of being selected.
8. _____ are statistical errors that are caused due to the sample not representing the target population adequately.
9. A trader who wants to predict short-term movements in stock prices is likely to use _____ analytics
10. Observations consisting of pairs of variable data are required to construct a _____ chart.

ANSWERS

1. Inter Quartile Range
2. Coefficient of skewness
3. Kurtosis
4. Covariance
5. Correlation
6. Expected value
7. Simple random sampling
8. Nonsampling errors
9. Predictive
10. Scatter

One Mark Answers

1. Central tendency

Ans :

A measure of central tendency is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

2. Variability

Ans :

The variance is the average of squared deviations from the mean. A deviation from the mean is how far a score lies from the mean.

3. Range

Ans :

The difference between the highest and lowest values.

4. Data Visualization

Ans :

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

5. Cross-tabulation

Ans :

Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyze categorical (nominal measurement scale) data.

UNIT III

PREDICTIVE ANALYTICS:

Trend Lines, Regression Analysis – Linear & Multiple, Predictive modeling, forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Data mining and business intelligence, Data mining for business, Classification, Association, Cause Effect Modeling.

3.1 INTRODUCTION

Q1. What is mean by Predictive analytics?

Ans : (Imp.)

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.

Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. Predictive analytics is often associated with big data and data science.

Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.

Organizations can use historic and current data to forecast trends and behaviors seconds, days, or years into the future with a great deal of precision.

The workflow for building predictive analytics frameworks follows five basic steps :

i) Define the problem

A prediction starts with a good thesis and set of requirements. For instance, can a predictive analytics model detect fraud? Determine optimal inventory levels for the holiday shopping season? Identify potential flood levels from severe weather? A distinct problem to solve will help determine what method of predictive analytics should be used.

ii) Acquire and organize data

An organization may have decades of data to draw upon, or a continual flood of data from customer interactions. Before predictive analytics models can be developed, data flows must be identified, and then datasets can be organized in a repository such as a data warehouse like BigQuery.

iii) Pre-process data

Raw data is only nominally useful by itself. To prepare the data for the predictive analytics models, it should be cleaned to remove anomalies, missing data points, or extreme outliers, any of which might be the result of input or measurement errors.

iv) Develop predictive models

Data scientists have a variety of tools and techniques to develop predictive models depending on the problem to be solved and nature of the dataset. Machine learning, regression models, and decision trees are some of the most common types of predictive models.

v) Validate and deploy results

Check on the accuracy of the model and adjust accordingly. Once acceptable results have been achieved, make them available to stakeholders via an app, website, or data dashboard.

Q2. Discuss various predictive analytics techniques.

Ans :

Predictive analytics incorporates a variety of data analysis approaches, including data mining, machine learning, and others. The following are the techniques used in predictive analytics.

i) Decision Trees

A decision tree is an analytics methodology based on Machine Learning that uses data mining algorithms to forecast the potential risks and benefits of undertaking certain options. It is a visual chart that resembles an upside-down tree that depicts the prospective result of a decision. When used for analytics, it can solve all forms of classification problems and answer difficult issues.

ii) Neural Networks

Neural networks are biologically inspired data processing systems that use historical and present data to forecast future values. Their architecture allows them to identify complicated connections buried in data in a way that replicates the pattern detecting systems of the human brain.

They are widely used for image recognition and patient diagnosis and comprise many layers that accept data (input layer), compute predictions (hidden layer), and provide output (output layer) in the form of a single prediction.

iii) Text Analytics

Text Analytics is used when a company wants to anticipate a numerical number. It is built on approaches from statistics, machine learning, and linguistics. It assists in predicting the themes of a document and analyzes words used in the supplied form.

iv) Regression Model

A regression method is crucial for the organization when it comes to estimating a numerical number, such as how long it will take a target audience to return to an airline reservation before purchasing, or how much money someone would spend on vehicle payments over a specific length of time.

Q3. What are the applications of predictive analytics in business scenario.

Ans :

Predictive analytics is used in a wide variety of ways by companies worldwide. Adopters from

diverse industries such as banking, healthcare, commerce, hospitality, pharmaceuticals, automotive, aerospace, and manufacturing get benefitted from the technology.

Here are a few examples of how businesses are using predictive analytics:

i) Customer Service

Businesses may better estimate demand by utilizing advanced and effective analytics and business intelligence. Consider a hotel company that wants to estimate how many people will stay in a certain area this weekend so that they can guarantee they have adequate employees and resources to meet demand.

ii) Higher Education

Predictive analytics applications in higher education include enrollment management, fundraising, recruiting, and retention. Predictive analytics offers a significant advantage in each of these areas by offering intelligent insights that would otherwise be neglected.

- (a) A prediction algorithm can rate each student and tell administrators ways to serve students during the duration of their enrollment using data from a student's high school years.
- (b) Models can give crucial information to fundraisers regarding the optimal times and strategies for reaching out to prospective and current donors.

iii) Supply Chain

Forecasting is an important concern in manufacturing because it guarantees that resources in a supply chain are used optimally. Inventory management and the shop floor, for example, are critical spokes of the supply chain wheel that require accurate forecasts to function.

Predictive modeling is frequently used to clean and improve the data utilized for such estimates. Modeling guarantees that additional data, including data from customer-facing activities, may be consumed by the system, resulting in a more accurate prediction.

iv) Insurance

Insurance firms evaluate policy applicants to assess the chance of having to pay out for a future claim based on the existing risk pool of comparable policyholders, as well as previous occurrences that resulted in payments. Actuaries frequently utilize models that compare attributes to data about previous policyholders and claims.

v) Software Testing

Predictive analytics can help you enhance your operations throughout the full software testing life cycle.

Simplify the process of interpreting massive volumes of data generated during software testing by using that data to model outcomes.

You can keep your release schedule on track by monitoring timelines and utilizing predictive modeling to estimate how delays will affect the project.

By identifying these difficulties and their causes, you will be able to make course corrections in individual areas before the entire project is delayed.

3.1.1 Trend Lines**Q4. What is a Trend Line? Explain the need of trend lines?**

Ans : (Imp.)

A trendline is a straight line that connects two or more price points and then extends into the future to act as a line of support and resistance.

In simple words, a trendline is a line that we draw on our chart by connecting the swing highs and swing lows during a 'Trending Market'.

- Trendlines are important for both trend identification and confirmation
- Trendline represent the psychology between the bulls and the bears
- A gap-up or gap-down through a trendline offers the best trade setup

- Reaction on price will often occur at or near a trendline
- A trendline can determine whether the market is optimistic or pessimistic

Q5. What is Upward and Downward trend?

Ans :

Following are the different types of trends in the forex market

1. Sideways trends (range bound)
2. Uptrend (higher lows)
3. Downtrend (lower highs)

1. Sideways Trends

Sideways trends indicates that a currency movement is range-bound between levels of support and resistance. It usually occurs when the market does not have a sense of direction and ends up consolidating most of the time in this range only.

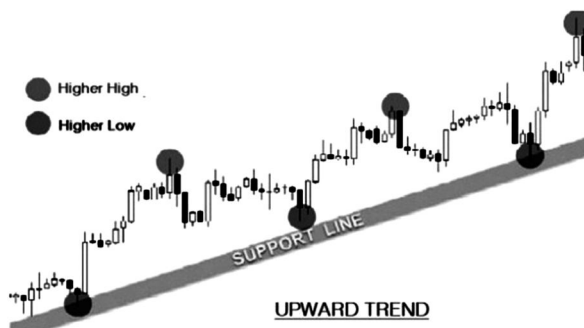
To identify if it is a sideways trend, traders often draw horizontal lines connected by the highs and lows of the price, which then form resistance and support levels. Clearly, market participants are not sure of which way the market will move and there will be LITTLE or NO rate of price change.

**2. Uptrend**

An uptrend signifies that the market is heading in the upward direction, creating a bullish market. It indicates the price rallies often with intermediate periods of consolidation or movement (small downward move) against the major (prevailing) trend.

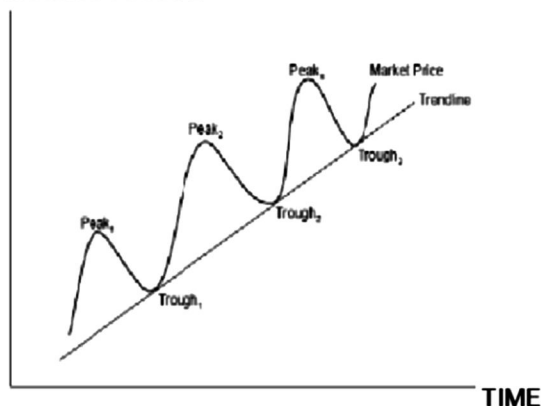
An upward trend continues until there is some breakdown in the charts (going down below

some major support areas). If the market trend is upwards, we need to be cautious on taking short position (against the overall market trend) on some minor correction in the market.



Another way to figure an upward trend of market or currency price is shown below “

CURRENCY PRICE



UPWARD TREND

Above the primary waves move the currency pair (USD/INR) in the direction of the broader trend (upward move), and secondary waves act as corrective phases (minor correction in currency, downward) of the primary waves (upward).

3. Downward Trend

A downward trend in the forex market is characterized by a price decline in the currency pair (USD/INR), with slight upward swing for a period of consolidation against the prevailing trend (downward trend). Unlike upward trend, a downward trend results in a negative rate of price change over time. In a chart,

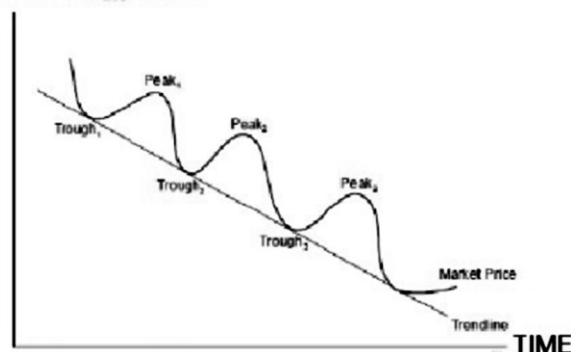
the price movements indicating a downtrend form a sequence of lower peaks and lower lows.

As currency is always traded in pair, the downtrend in forex market is not much affected as other financial markets. In case of downtrend of a currency pair (USD/INR), the fall in price of USD gives way to a rise in price of INR. It means something is always going up even in times of financial or economical downtrend.



Another way to look at the downward trend figure is in the form of primary (major trend) and secondary (minor correction) wave, as shown in the diagram below.

CURRENCY RATE



DOWNTREND

In the above figure, the primary wave (downtrend) moves the currency pair in the direction of the broader trend (downward trend), and secondary waves (uptrend) act as corrective phases of the primary waves (downtrend).

Q6. Discuss various types of Trend Lines.*Ans :*

In modern technical analysis, trendlines are one of the most often utilized and misrepresented (not used correctly) techniques. There are many types of trendlines available in the financial market like an uptrend, downtrend, counter-trendline, diagonal, inner/outer trendline, and many more. But majorly there are three types of trendlines and all the others come under these.

1. The Standard Trendline
2. The Parallel trendline
3. The Supplement Trendline

Types**1. The Standard Trendline**

- Standard trendline consist of uptrend lines or downtrend lines or both.
- Standard uptrend lines act as an support where buyers stepped in and bid the market higher.
- Standard downtrend lines act as an resistance where sellers stepped in and counter the bounce in the price.
- The correct way to draw a standard uptrend line is to connect higher lows price points.
- The correct way to draw a standard down-trend line is to connect lower high price points.

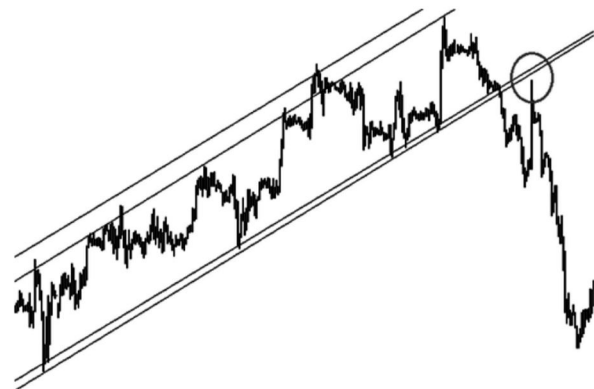
**2. The Parallel Trendline**

- The parallel trendline is used to identify a channel between the highs and the lows.
- If drawn correctly, you can notice the prices are confined between those two parallel lines.
- The bigger the channel, the more significant it is.

- The upper limit of the dominant area act as resistance.
- The lower limit of the dominant area act as support.
- The parallel trendlines can be drawn during an uptrend, downtrend or sideways market.

**3. The Supplementary Trendline**

- The supplementary trendlines are used to draw trendlines during the short period of time.
- Supplementary trendlines are used to enhance the analysis for shorter timeframe (generally used with other type of technical analysis).
- If you are beginner, better to avoid the trade based on supplementary trendlines.
- The correct way to draw a supplementary trendline is to connect the day's low of at least two previous days to the yesterday's high point.

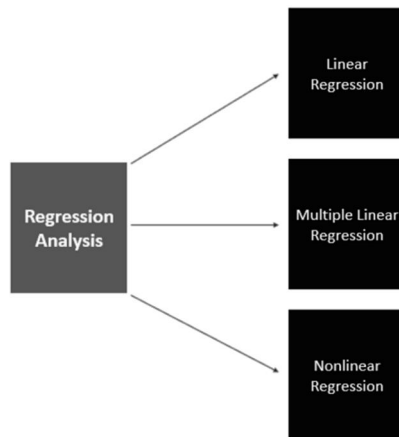


3.2 REGRESSION ANALYSIS – LINEAR AND MULTIPLE

Q7. What is Regression Analysis?

Ans : (Imp.)

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.



Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Linear regression analysis is based on six fundamental assumptions :

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Q8. What is simple Linear Regression?

Ans:

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- Y – Dependent variable
- X – Independent (explanatory) variable
- a – Intercept
- b – Slope
- ϵ – Residual (error)

Q9. What is multiple linear regression.

Ans :

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$$

Where:

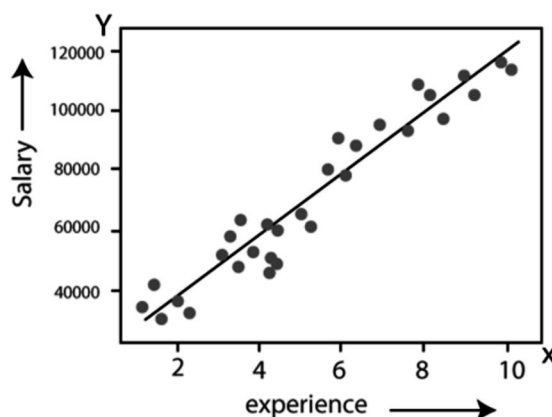
- Y – Dependent variable
- X_1, X_2, X_3 – Independent (explanatory) variables
- a – Intercept
- b, c, d – Slopes
- ϵ – Residual (error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

- **Non-collinearity:** Independent variables should show a minimum correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Q10. Discuss in detail about linear regression.*Ans :***Linear Regression**

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience.



- Below is the mathematical equation for Linear regression:

$$Y = aX + b$$

Here, Y = dependent variables (target variables),

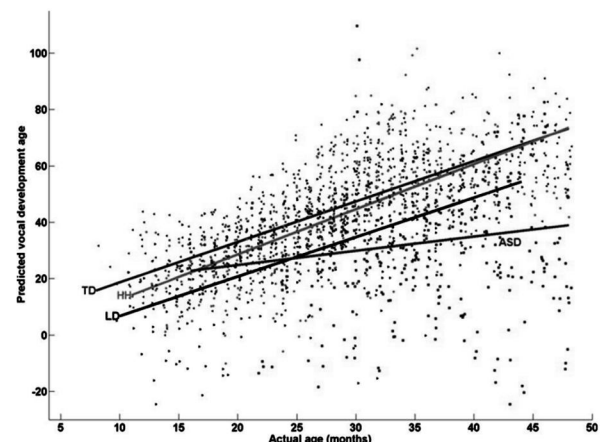
X = Independent variables (predictor variables), a and b are the linear coefficients

Some popular applications of linear regression are :

- Analyzing trends and sales estimates
- Salary forecasting
- Real estate prediction
- Arriving at ETAs in traffic.

Q11. What is multiple linear regression.*Ans :*

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

**Multiple Linear Regression Formula**

$$y_i = \beta_0 + \beta_1 \times i_1 + \beta_2 \times i_2 + \dots + \beta_p \times i_p + \epsilon$$

Where:

- y_i is the dependent or predicted variable
- β_0 is the y-intercept, i.e., the value of y when both x_1 and x_2 are 0.
- β_1 and β_2 are the regression coefficients representing the change in y relative to a one-unit change in x_1 and x_2 , respectively.
- β_p is the slope coefficient for each independent variable
- ϵ is the model's random error (residual) term.

3.3 PREDICTIVE MODELLING

Q12. What is Predictive Modelling?

Ans : (Imp.)

Predictive modeling is a predictive analysis tool. It is widely used by companies to determine the viability of a new venture, project, or proposal. It applies statistical and analytical tools for analyzing current data and historical data and determines future outcomes.

- Predictive modeling uses known results to create, process, and validate a model that can be used to make future predictions.
- Regression and neural networks are two of the most widely used predictive modeling techniques.
- Companies can use predictive modeling to forecast events, customer behavior, and financial, economic, and market risks.

Q13. Discuss various predictive models.

Ans :

Several different types of predictive modeling can be used to analyze most datasets to reveal insights into future events.

1. Classification Models

Classification models use machine learning to place data into categories or classes based on criteria set by a user. There are several types of classification algorithms, some of which are:

- **Logistic regression:** An estimate of an event occurring, usually a binary classification such as a yes or no answer.
- **Decision trees:** A series of yes/no, if/else, or other binary results placed into a visualization known as a decision tree.
- **Random forest:** An algorithm that combines unrelated decision trees using classification and regression.
- **Neural networks:** Machine learning models that review large volumes of data for correlations that emerge only after millions of data points are reviewed.

- **Naïve Bayes:** A modeling system based on Bayes' Theorem, which determines conditional probability.

2. Clustering Models

Clustering is a technique that groups data points. It is assumed by analysts that data in similar groups should have the same characteristics, and data in different groups should have very different properties. Some popular clustering algorithms are:

- **K-Means:** K-means is a modeling technique that uses groups to identify central tendencies of different groups of data.
- **Mean-Shift:** In mean-shift modeling, the mean of a group is shifted by the algorithm so that "bubbles," or maxima of a density function, are identified. When the points are plotted on a graph, data appear to be grouped around central points called centroids.
- **Density-based Spatial Clustering With Noise (DBSCAN):** DBSCAN is an algorithm that groups data points together based on an established distance between them. This model establishes relationships between different groups and identifies outliers.

3. Outlier Models

A dataset always has outliers (values outside its normal values). For instance, if you had the numbers 21, 32, 46, 28, 37, and 299, you can see the first five numbers are somewhat similar, but 299 is too far from the others. Thus, it is considered an outlier. Some algorithms used to identify outliers are:

- **Isolation Forest:** An algorithm that detects few and different data points in a sample.
- **Minimum Covariance Determinant (MCD):** Covariance is the relationship of change between two variables. The MCD measures the mean and covariance of a dataset that minimizes the influence outliers have on the data.
- **Local Outlier Factor (LOF):** An algorithm that identifies nearest neighboring data points and assigns scores, allowing those furthest away to be identified as outliers.

4. Time Series Models

Commonly used before other types of modeling, time series modeling uses historical data to forecast events. A few of the common time series models are:

- **ARIMA:** The autoregressive integrated moving average model uses autoregression, integration (differences between observations), and moving averages to forecast trends or results.
- **Moving Average:** The moving average uses the average of a specified period, such as 50 or 200 days, which smooths out fluctuations.

Q14. List out the techniques used in predictive modelling.

Ans :

The following techniques are used in predictive modeling:

1. **Linear Regression:** When two continuous variables depict a linear relationship, a linear regression can be used to determine the value of the dependent variable—based on the independent variable.
2. **Multiple Regression:** This is similar to linear regression, except the value of the dependent variable is evaluated by analyzing multiple independent variables.
3. **Logistic Regression:** It is used for ascertaining dependent variables when the data set is large—requiring categorization.
4. **Decision Tree:** This method is commonly used for data mining. A flowchart representing an inverted tree is formulated. Here the internal node splits into branches that list out two or more possible decisions, and each decision is further subdivided—to show other possible outcomes. This technique helps in selecting the best option.
5. **Random Forest:** It is a popular regression and classification model. It is used for solving machine learning algorithms. It comprises multiple decision trees—not correlated to each other. These decision trees collectively facilitate the analysis.

6. **Boosting:** As the name suggests, this method facilitates learning from the results of other models—decision tree, logistic regression, neural network, and support vector machine.

7. **Neural Networks:** It is a problem-solving mechanism used in machine learning and artificial intelligence. It develops a set of algorithms for a computational learning system. These algorithms comprise three layers—input, processing, and output.

Q15. What are the advantages and disadvantages of predictive modelling.

Ans :

Advantages

- Easy to generate actionable insights: Predictive modeling allows you to view information about your data that you might not see otherwise, enabling you to make more informed decisions.
- Can test different scenarios: Data can be manipulated or changed to test various scenarios to assess the influence changes might have on your data and models.
- Increases decision-making speed: Decisions can be reached much faster because millions of data points can be analyzed much quicker, and future trends or circumstances can be theorized within minutes or hours.

Disadvantages

- Computations can be inexplicable: You may not be able to interpret the results once you create a predictive model.
- Bias due to human input: Bias is introduced into modeling because humans are involved in setting parameters and criteria.
- High learning curve: Learning to create predictive models and/or interpret the results can be a lengthy process because you have to understand statistics, learn the jargon, and possibly even learn to code in Python or R.

Q16. Explain any four applications of predictive modelling.*Ans :***1. Insurance Sector**

Insurance companies use various predictive techniques to evaluate premium values, maximize profits, identify frauds, and improve claim settlement processes. For instance, a vehicular insurance company analyzes vehicles' conditions and applies various algorithms to determine the applicable premium amount.

2. Finance and Banking Industry

Before extending a loan, banks use prediction models to review borrowers' credit scores—to verify credibility, background, and previous defaults. It helps predict the chances of fraud, misrepresentation, and risks involved with a particular client.

3. Marketing and Retail Sector

When a business runs a marketing campaign, it uses predictive modeling techniques to anticipate campaign success. Predictive analysis also gauges target audience and future sales. In the retail sector, predictive analyses provide forecasts based on which businesses decide the required inventory for each certain product. Projections help decide how much stock volume is required to meet future demands—pertaining to a particular product.

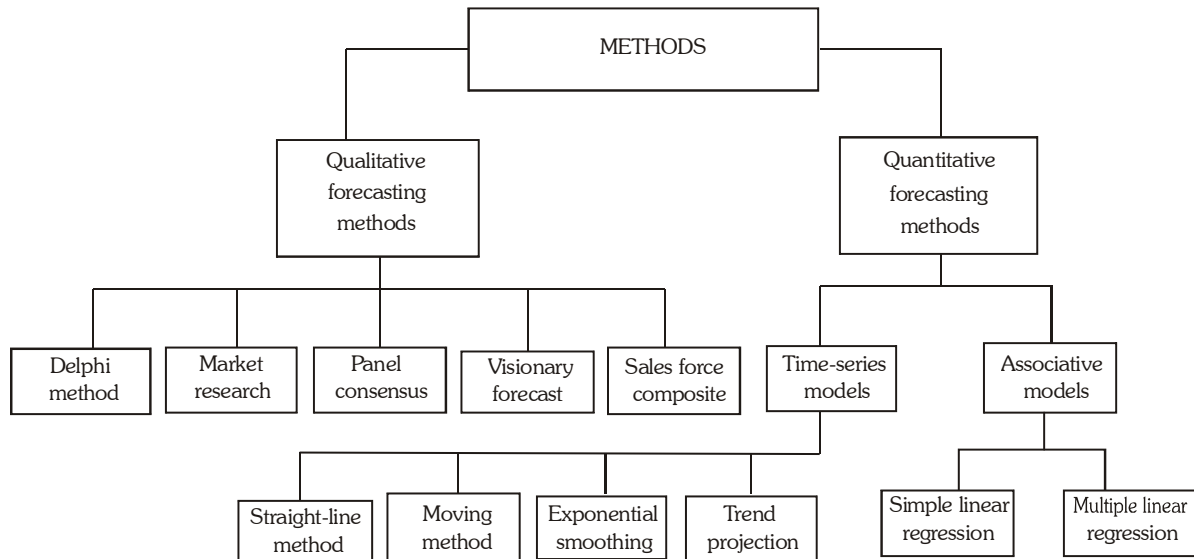
4. Weather Forecast

Predictive modeling methods like a decision tree and linear regression forecast weather changes and natural calamities—thunderstorms, cyclones, and tsunamis. These models can ascertain the wind direction and wind speed of storms. Thus, these models are used to alert inhabitants of an area.

3.4 FORECASTING TECHNIQUES**Q17. Discuss various forecasting methods.****(OR)****Discuss various forecasting techniques.***Ans :***(Imp.)**

Businesses can choose between different types of forecasting methods, including:

1. Quantitative forecasting methods, and
2. Qualitative forecasting methods.

Forecasting methods**1. Quantitative methods of forecasting**

In order to make realistic and accurate forecasts, quantitative methods include mathematical processes such as:

- Algebra,
- Permutations and combinations,
- Set theory,
- Matrix algebra, and
- Integration.

In addition, mathematical techniques such as linear programming, dynamic programming, and inventory control can also help decision-makers guide their business strategies.

Quantitative methods also include statistical processes such as:

- Standard deviation,
- One factor analysis of variance,
- Multi-factor analysis of variance,
- Two sample t-test for equal means,
- Autocorrelation, and
- Hypothesis test.

So, rather than basing the results on opinion and intuition — quantitative methods implement readily available data to interpret results. These methods are usually used to make short-term predictions by analyzing older, raw data.

Finally, quantitative methods can be further divided into:

(i) Time-series forecasting models

Examine past patterns in the data in order to predict future patterns. Examples of time-series models include: straight-line method, moving average, exponential smoothing, and trend projection.

(ii) Associative models (causal models)

The variable that is being forecasted is associated with other variables, thus, the projections are built on that relationship. Examples of causal models include: simple linear and multiple linear regression.

(i) Time-series forecasting models

Time-series is a popular forecasting model which explores past company behavior to forecast future company behavior (consumer behavior, sales behavior, etc.). This type of forecasting model uses historical data in terms of hours, weeks, months, and years to come at a point in the future based on these past values.

Time-series uses information gathered over several years to analyze sales velocity based on the business needs. Based on those figures, you can create future forecasts using mathematical formulas. There are several models of completing time-series forecasting which will help you formulate future estimations.

The subtypes below are all examples of time-series forecasting models:

- (a) Straight-line method,
- (b) Moving average model,
- (c) Exponential smoothing model, and
- (d) Trend projection model.

Let's find out more about each one of them.

(a) Straight-line method

The straight-line method is a time-series forecasting model that provides estimates about future revenues by taking into consideration past data and trends.

For this type of model, it's important to find the growth rate of sales, which will be implemented in the calculations.

(b) Moving average model

The moving average model is similar to the straight-line forecasting, except that it's often used to predict short-term trends (such as daily, monthly, quarterly, or half-yearly intervals). Companies use the moving average model when they need to forecast sales, revenue, profit, or other important business metrics.

With regards to calculating future revenue, for example, the model focuses on observing the past and current revenues (i.e., the average number of the revenues in a given time period) to predict future outcomes. This type of forecasting model is useful when calculating the performance of a specific metric within a certain time limit.

(c) Exponential smoothing model

Similar to the moving average, exponential smoothing is another time-series forecasting model which can be used to predict new values by using a set of weighted averages based on past observations.

Exponential smoothing helps to predict the future by using past company data. The weights start declining exponentially with past observations in order to predict the upcoming period. To put it simply, if the observation is a more recent one — the associated weight is higher. This means that more weight is given to recent values instead of past values.

New forecasts are predicted by including the past forecast and the percentage of value (the difference between the current and the past forecast). The idea behind this model is to attribute importance to more recent values in the series — when observations become older, past values get exponentially smaller.

(d) Trend projection model

The trend projection model works best in situations where you could work out the future influence of certain variables (dependent or independent) based on its past behavior. The model examines past events in order to identify patterns and trends that could recur frequently.

Trend projection can be used to forecast future activity since it considers that all factors involved in past trends will continue in the future as well. The model requires long and reliable time-series data which is arranged in chronological order for evaluation.

By identifying the patterns of trends, the company will be able to get a vision of the future. Consequently, once the trend has been identified, it will be able to predict future demand.

(ii) Associative (causal) forecasting models

According to associative or causal forecasting models, the forecasted variable is interconnected to other variables in the business system. Therefore, the forecast projections rely on these associations.

Associative models are an advanced way to forecast your sales because they implement specific mathematical calculations to identify the connection between different variables that can affect your business activity.

The subtypes below are all examples of causal models:

- (a) Simple-linear regression model,
- (b) Multiple linear regression model.

(a) Simple linear regression model

The simple linear regression is a type of associative forecasting model that provides a more detailed context to your forecast by examining how the independent variable is correlated to the dependent variable.

The dependent variable is the predicted value (e.g. sales), and the independent variable (e.g. profit) guides the expected value of the dependent variable.

Simple linear regression can be visualized on a graph by portraying one metric on the X axis, and the other one on the Y axis.

Here's the formula for calculating simple linear regression:

$$Y = BX + A$$

Y = dependent variable (predicted value)

B = the slope of the regression line (measure of its steepness i.e the ratio of the rise to the run, or rise divided by the run)

X = independent variable

A = Y-intercept (the point on the Y-axis by which the slope of the line sweeps)

Apart from identifying the relationship between sales and profits, it can also demonstrate the rate of increase and how that rate varies in order to help you find ways to maximize profit.

Calculating the simple linear regression is a tedious process, so you might want to use statistical programs to help you analyze the data.

(b) Multiple linear regression model

As its name suggests, the multiple linear regression model follows the same approach — i.e. makes the same assumptions as the simple linear regression — except that it applies it to a number of different business variables.

So, when business performance is influenced by more than one variable, this model allows you to explore the relationship between two or more independent variables and one dependent variable. This will help you get a clear picture of the situation and a more accurate forecast.

(2) Types of qualitative forecasting models

Another method of forecasting is qualitative forecasting.

Qualitative forecasting methods are different from quantitative methods because they're subjective and intuitive in nature. They are based on the opinion and judgment of consumers and experts.

In addition, qualitative methods use factors such as demand trends and seasonality to create more accurate forecasts.

They can only be used if past data isn't readily available.

All the forecasting models below belong to the category of qualitative forecasting methods.

(a) Delphi method

The Delphi method is a type of forecasting model that involves a small group of relevant experts who express their judgment and opinion on a given problem or situation. The expert opinions are then combined with market orientation to come up with results and develop an accurate forecast.

(b) Market research model

Market research is a qualitative forecasting model that evaluates the performance of a business's products and services by interviewing potential customers about them. Their reactions and responses are recorded, and then they're analyzed in order to come up with a sales forecast.

This model can be performed by staff members or third-party agencies (specialized in market research) by:

- Telephone,
- Opinion poll,
- Personal interviews, or
- Questionnaires.

Some examples of market research strategies include:

- Focus groups,
- Consumer surveys, or
- Product testing.

(c) Panel consensus model

Panel consensus (also called expert opinion) is a qualitative forecasting approach where experts or employees from all levels of an organization (from low-level to top-level) discuss a product or service. The members act like a focus group, expressing their thoughts and recommendations in order to develop a forecast.

Anyone can speak up during the discussion, however, sometimes lower-level employees may feel intimidated to express their opinion due to their lack of market knowledge. This is one of the drawbacks of this model.

(d) Visionary forecast model

The visionary forecasting model is based on personal opinions, judgements, and insights of a relevant and experienced individual. The projections are backed up by data, information, and facts in order to predict future scenarios. When available, historical analogies can also be used to hypothesize potential future forecasts.

In other words, the 'visionary' prophecies a set of future events by examining past events and developments. Therefore, this model is subjective and non-scientific in nature, and is solely based on an individual's guesswork and imagination.

(e) Sales force composite model

Another reliable qualitative forecasting model is sales force composite where the input of sales staff is used to estimate future sales. When estimating future demand, the company may decide to collect information from the salesperson that would help in determining customer's needs and predicting the sales in a

certain region and given time period.

According to the sales force composite model, the sales agent better understands the needs of the customers since they interact with them on a regular basis. This information will help in adjusting business operations in order to meet the client's needs and maximize sales.

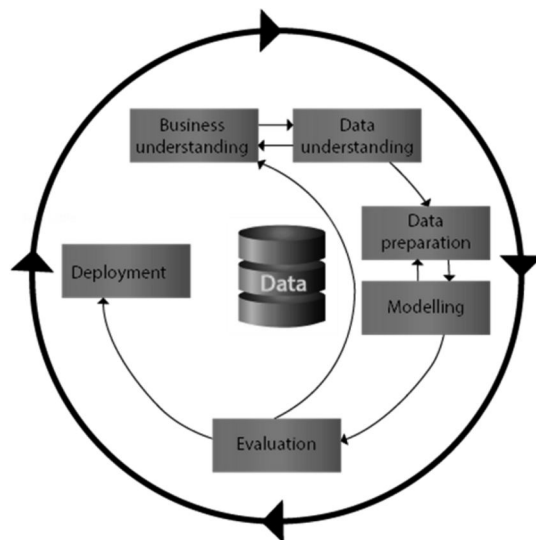
3.5 DATA MINING-DEFINITION-APPROACHES IN DATA MINING

Q18. What is Data mining? Give a brief overview on data mining.

Ans :

(Imp.)

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.



Overview of the data mining process.

Almost all businesses use data mining, and it's important to understand the data mining process and how it can help a business make decisions.

(i) Business understanding

The first step to successful data mining is to understand the overall objectives of the

business, then be able to convert this into a data mining problem and a plan. Without an understanding of the ultimate goal of the business, you won't be able to design a good data mining algorithm. For example, a supermarket may want to use data mining to learn more about their customers. The business understanding is that a supermarket is looking to find out what their customers are buying the most.

(ii) Data understanding

After you know what the business is looking for, it's time to collect data. There are many complex ways that data can be obtained from an organization, organized, stored, and managed. Data mining involves getting familiar with the data, identifying any issues, getting insights, or observing subsets. For example, the supermarket may use a rewards program where customers can input their phone number when they purchase, giving the supermarket access to their shopping data.

(iii) Data Preparation

Data preparation involves getting the information production ready. This is the biggest part of data mining. It is taking the computer-language data, and converting it into a form that people can understand and quantify. Transforming and cleaning the data for modeling is key for this step.

(iv) Modeling

In the modeling phase, mathematical models are used to search for patterns in the data. There are usually several techniques that can be used for the same set of data. There is a lot of trial and error involved in modeling.

(v) Evaluation

When the model is complete, it needs to be carefully evaluated and the steps to make the model need to be reviewed, to ensure it meets the business objectives. At the end of this phase, a decision about the data mining results will be made. In the supermarket example, the data mining results will provide a list of what the customer has purchased, which is what the business was looking for.

(vi) Deployment

This can be a simple or complex part of data mining, depending on the output of the process. It can be as simple as generating a report, or as complex as creating a repeatable data mining process to happen regularly.

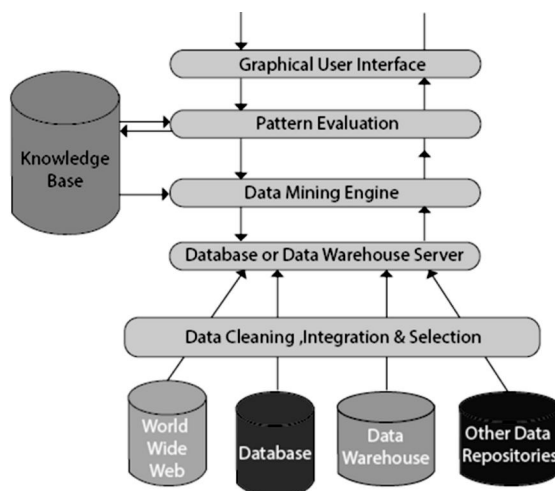
Q19. Draw and explain data mining architecture.

Ans :

The data mining process involves several components, and these components constitute a data mining system architecture.

Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.

**Data Source**

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data

mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable.

Q20. Discuss various data mining techniques.

Ans :

Data mining techniques in business analytics

(i) Classification

This data mining technique is more complex, using attributes of data to move them into discernable categories, helping you draw further conclusions. Supermarket data mining may use classification to group the types of groceries customers are buying, like produce, meat, bakery items, etc. These classifications help the store learn even more about customers, outputs, etc.

(ii) Clustering

This technique is very similar to classification, chunking data together based on their similarities. Cluster groups are less structured than classification groups, making it a more simple option for data mining. In the supermarket example, a simple cluster group could be food and non-food items instead of the specific classes.

(iii) Association rules

Association in data mining is all about tracking patterns, specifically based on linked variables. In the supermarket example, this may mean that many customers who buy a specific item may also buy a second, related item. This is how stores may know how to group certain food items together, or in online shopping they may show “people also bought this” section.

(iv) Regression analysis

Regression is used to plan and model, identifying the likelihood of a specific variable. The supermarket may be able to project price points based on availability, consumer demand, and their competition. Regression helps data mining by identifying the relationship between variables in a set.

(v) Anomaly/outlier detection

For many data mining cases, just seeing the overarching pattern might not be all you need. Data needs to be able to identify and understand the outliers in your data as well. For example, in the supermarket if most of the shoppers are female, but one week in February is mostly men, you’ll want to investigate that outlier and understand what is behind it.

Q21. List and explain data mining methods.

Ans :

There are many methods used for Data Mining, but the crucial step is to select the appropriate form from them according to the business or the problem statement. These methods help in predicting the future and then making decisions accordingly. These also help in analyzing market trends and increasing company revenue.

Some Methods are:

1. Association
2. Classification
3. Clustering Analysis
4. Prediction

5. Sequential Patterns or Pattern Tracking
6. Decision Trees
7. Outlier Analysis or Anomaly Analysis
8. Neural Network

1. Association

It is used to find a correlation between two or more items by identifying the hidden pattern in the data set and hence also called relation analysis. This method is used in market basket analysis to predict the behavior of the customer.

2. Classification

This data mining method is used to distinguish the items in the data sets into classes or groups. It helps to predict the behaviour of entities within the group accurately. It is a two-step process:

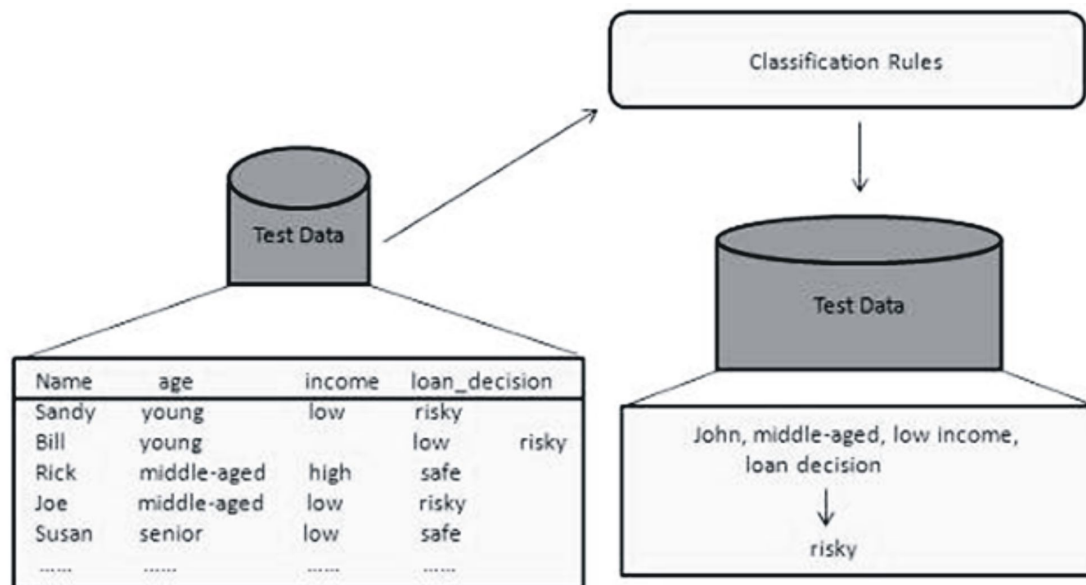
➤ Learning step (training phase)

In this, a classification algorithm builds the classifier by analyzing a training set.

➤ Classification step

Test data are used to estimate the accuracy or precision of the classification rules.

For example, a banking company uses to identify loan applicants at low, medium or high credit risks. Similarly, a medical researcher analyzes cancer data to predict which medicine to prescribe to the patient.

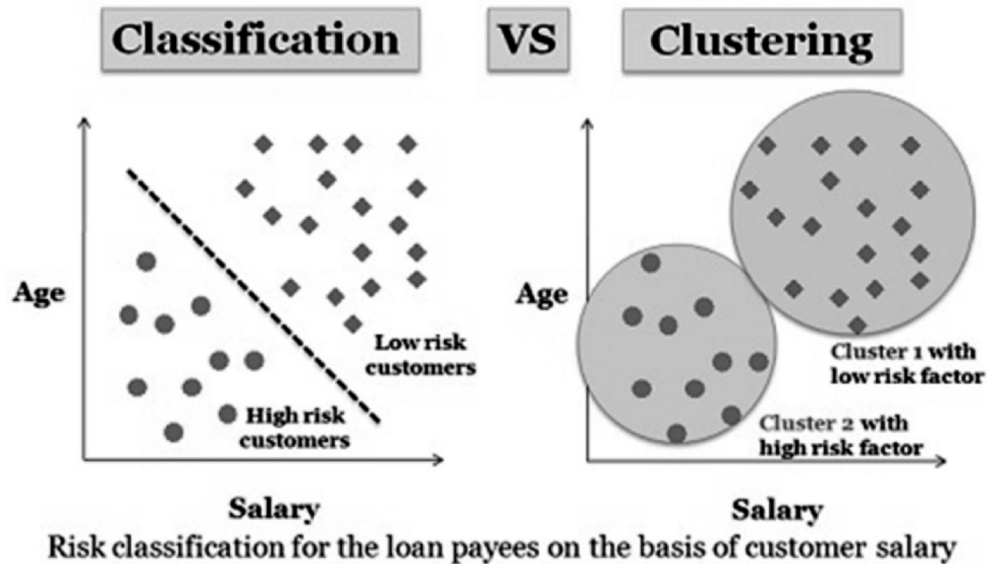


3. Clustering Analysis

Clustering is almost similar to classification, but in this cluster are made depending on the similarities of data items. Different groups have dissimilar or unrelated objects. It is also called data segmentation as it partitions huge data sets into groups according to the similarities.

Various clustering methods are used

- Hierarchical Agglomerative methods
- Grid-Based Methods
- Partitioning Methods
- Model-Based Methods
- Density-Based Methods



4. Prediction

This method is used to predict the future based on the past and present trends or data set. Prediction is mostly used to combine other mining methods such as classification, pattern matching, trend analysis, and relation.

For example, if the sales manager would like to predict the amount of revenue that each item would generate based on past sales data. It models a continuous-valued function that indicates missing numeric data values.

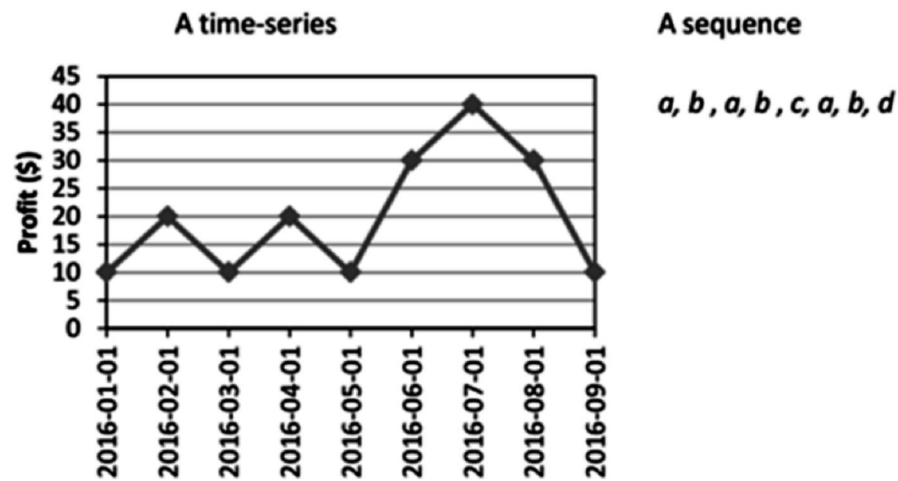


5. Sequential patterns or Pattern tracking

This method is used to identify patterns that frequently occur over a certain period of time.

For example, a clothing company's sales manager sees that sales of jackets seem to increase just before the winter season, or sales in bakery increase during Christmas or New Year's eve.

Let's look at an example with a graph.

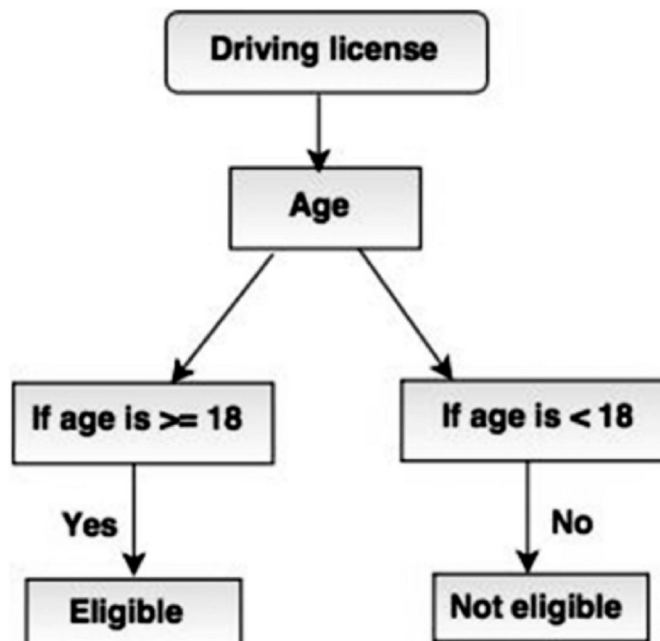


A time-series (left) and a sequence (right)

6. Decision Trees

A decision tree is a tree structure (as its name suggests), where

- Each internal node represents a test on the attribute.
- Branch denotes the result of the test.
- Terminal nodes hold the class label.
- The topmost node is the root node which has a simple question that has two or more answers. Accordingly, the tree grows, and a flow chart like structure is generated.



Decision tree

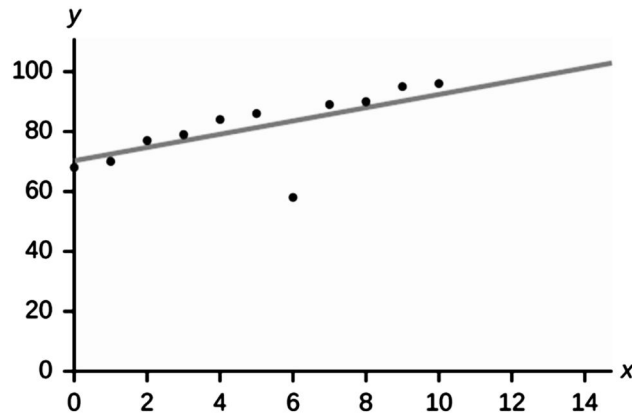
7. Outlier Analysis or Anomaly Analysis

This method identifies the data items that do not comply with the expected pattern or expected behaviour. These unexpected data items are considered as outliers or noise. They are helpful in many domains like credit card fraud detection, intrusion detection, fault detection etc. This is also called Outlier Mining.

For example, let's assume the graph below is plotted using some data sets in our database.

So the best fit line is drawn. The points lying nearby the line show expected behaviour while the end far from the line is an Outlier.

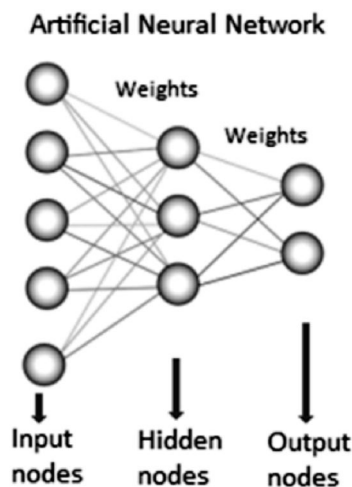
This would help to detect the anomalies and take possible actions accordingly.



8. Neural Network

This method or model is based on biological neural networks. It is a collection of neurons like processing units with weighted connections between them. They are used to model the relationship between inputs and outputs. It is used for classification, regression analysis, data processing etc. This technique works on three pillars-

- Model
- Learning Algorithm (supervised or unsupervised)
- Activation function



3.6 DATA EXPLORATION & REDUCTION**Q22. What is Data Exploration?***Ans :***(Imp.)**

Data exploration refers to the initial step in data analysis. Data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, to understand the nature of the data better.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

Data without a question is simply information. Asking a question of data turns it into an answer. Data with the right questions and exploration can provide a deeper understanding of how things work and even enable predictive abilities.

R and Python are the most common languages used for exploration; the former works best for statistical learning while the latter lends itself well to machine learning. Coding is not necessary for data exploration through no-code platforms.

The exploration process is also increasingly important to working with Geographic Information Systems (GIS) since so much of today's data is location-enriched.

Data exploration typically follows three steps:**1. Understand the Variables**

The basis for any data analysis begins with an understanding of variables. A quick read of column names is a good place to start. A closer look at data catalogues, field descriptions, and metadata can offer insight into to what each field represents and help discover missing or incomplete data.

2. Detect Any Outliers

Outliers or anomalies can derail an analysis and distort the reality of a dataset, so it's important to identify them early on. Data visualization, numerical methods, interquartile ranges, and hypothesis testing are the most common ways of detecting outliers. A boxplot, histogram, or scatterplot, for example, makes it easy to spot points far outside the standard range, while a z-score informs how far from the mean a data point is. Once found, an analyst can investigate, adjust, omit, or ignore the outliers. No matter the choice, the decision should be noted in the analysis.

3. Examine Patterns and Relationships

Plotting a dataset in a variety of ways makes it easier to identify and examine the patterns and relationships among variables. For example, a business exploring data from multiple stores may have information on location, population, temperature, and per capita income. To estimate sales for a new location, they need to decide which variables to include in their predictive model.

Q23. What are the differences between data exploration and data mining.*Ans :*

S.No.	Data Mining	S.No.	Data Exploration
1.	Data mining is also named knowledge web discovery in databases, extraction, data /pattern analysis, and information harvesting.	1.	Data Exploration is used interchangeably with exploration, web scraping, web crawling, data retrieval, data harvesting, etc.
2.	Data mining studies are mostly on structured data.	2.	Data Exploration usually retrieves data out of un structured or poorly structured data sources.
3.	Data mining aims to make available data more useful for generating insights.	3.	Data Exploration is to collect data and gather them into a place where they can be stored or further processed.
4.	Data mining is based on mathematical methods to reveal patterns or trends.	4.	Data Exploration is based on programming languages or data Exploration tools to crawl the data sources.
5.	The purpose of data mining is to find facts that are previously unknown or ignored,	5.	Data Exploration deals with existing information.
6.	Data mining is much more complicated and requires large investments in staff training.	6.	Data Exploration can be extremely easy and cost-effective when conducted with the right tool.

Q24. What is Data Reduction? What are the benefits of data reduction?*Ans :*

Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Data reduction does not affect the result obtained from data mining. That means the result obtained from data mining before and after data reduction is the same or almost the same.

Benefits of Data Reduction

The main benefit of data reduction is simple: the more data you can fit into a terabyte of disk space, the less capacity you will need to purchase. Here are some benefits of data reduction, such as:

- Data reduction can save energy.
- Data reduction can reduce your physical storage costs.
- And data reduction can decrease your data center track.

Q25. List and explain the data reduction methods.

(OR)

List and explain data reduction techniques.

Ans :

Here are the following techniques or methods of data reduction in data mining, such as:

1. Dimensionality Reduction

Whenever we encounter weakly important data, we use the attribute required for our analysis. Dimensionality reduction eliminates the attributes from the data set under consideration, thereby reducing the volume of original data. It reduces data size as it eliminates outdated or redundant features. Here are three methods of dimensionality reduction.

i. Wavelet Transform

In the wavelet transform, suppose a data vector A is transformed into a numerically different data vector A' such that both A and A' vectors are of the same length. Then how it is useful in reducing data because the data obtained from the wavelet transform can be truncated. The compressed data is obtained by retaining the smallest fragment of the strongest wavelet coefficients. Wavelet transform can be applied to data cubes, sparse data, or skewed data.

ii. Principal Component Analysis

Suppose we have a data set to be analyzed that has tuples with n attributes. The principal component analysis identifies k independent tuples with n attributes that can represent the data set.

In this way, the original data can be cast on a much smaller space, and dimensionality reduction can be achieved. Principal component analysis can be applied to sparse and skewed data.

iii. Attribute Subset Selection

The large data set has many attributes, some of which are irrelevant to data mining or some are redundant. The core attribute subset selection reduces the data volume and dimensionality. The attribute subset selection reduces the volume of data by eliminating redundant and irrelevant attributes.

The attribute subset selection ensures that we get a good subset of original attributes even after eliminating the unwanted attributes. The resulting probability of data distribution is as close as possible to the original data distribution using all the attributes.

2. sNumerosity Reduction

The numerosity reduction reduces the original data volume and represents it in a much smaller form. This technique includes two types parametric and non-parametric numerosity reduction.

i. Parametric

Parametric numerosity reduction incorporates storing only data parameters instead of the original data. One method of parametric numerosity reduction is the regression and log-linear method.

➤ **Regression and Log-Linear**

Linear regression models a relationship between the two attributes by modeling a linear equation to the data set. Suppose we need to model a linear function between two attributes.

$$y = wx + b$$

Here, y is the response attribute, and x is the predictor attribute. If we discuss in terms of data mining, attribute x and attribute y are the numeric database attributes, whereas w and b are regression coefficients.

Multiple linear regressions let the response variable y model linear function between two or more predictor variables.

Log-linear model discovers the relation between two or more discrete attributes in the database. Suppose we have a set of tuples presented in n -dimensional space. Then the log-linear model is used to study the probability of each tuple in a multidimensional space.

Regression and log-linear methods can be used for sparse data and skewed data.

ii. Non-Parametric

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric. There are at least four types of Non-Parametric data reduction techniques, Histogram, Clustering, Sampling, Data Cube Aggregation, and Data Compression.

➤ Histogram

A histogram is a graph that represents frequency distribution which describes how often a value appears in the data. Histogram uses the binning method to represent an attribute's data distribution. It uses a disjoint subset which we call bin or buckets.

A histogram can represent a dense, sparse, uniform, or skewed data. Instead of only one attribute, the histogram can be implemented for multiple attributes. It can effectively represent up to five attributes.

➤ Clustering

Clustering techniques groups similar objects from the data so that the objects in a cluster are similar to each other, but they are dissimilar to objects in another cluster.

How much similar are the objects inside a cluster can be calculated using a distance function. More is the similarity between the objects in a cluster closer they appear in the cluster.

The quality of the cluster depends on the diameter of the cluster, i.e., the max distance between any two objects in the cluster.

The cluster representation replaces the original data. This technique is more effective if the present data can be classified into a distinct clustered.

➤ Sampling

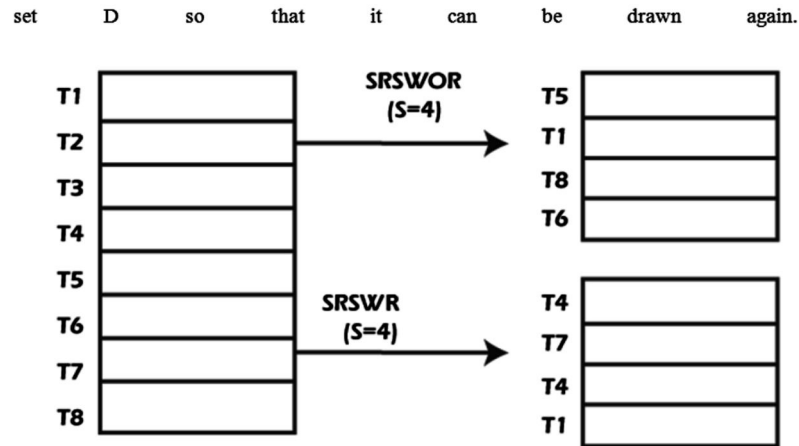
One of the methods used for data reduction is sampling, as it can reduce the large data set into a much smaller data sample. Below we will discuss the different methods in which we can sample a large data set D containing N tuples:

a. Simple random sample without replacement (SRSWOR) of sizes

In this s , some tuples are drawn from N tuples such that in the data set D ($s < N$). The probability of drawing any tuple from the data set D is $1/N$. This means all tuples have an equal probability of getting sampled.

b. Simple random sample with replacement (SRSWR) of sizes

It is similar to the SRSWOR, but the tuple is drawn from data set D, is recorded, and then replaced into the data set D so that it can be drawn again.

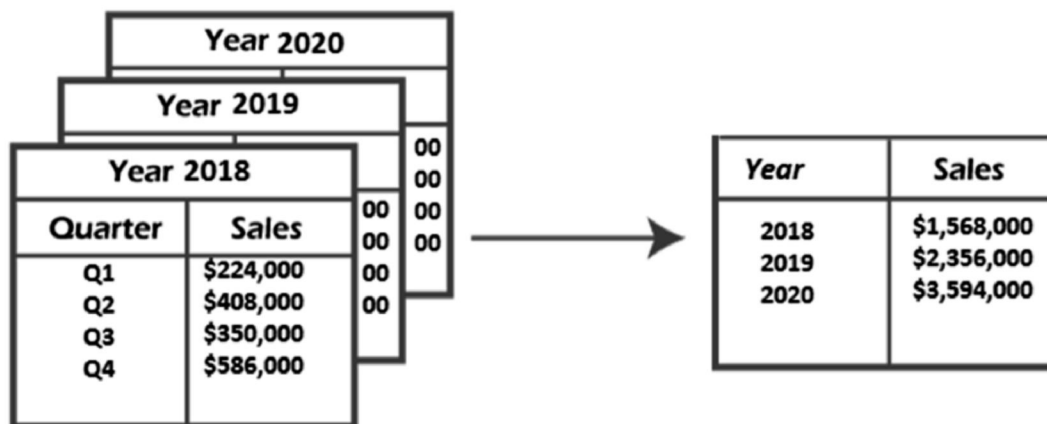


- c. Cluster sample:** The tuples in data set D are clustered into M mutually disjoint subsets. The data reduction can be applied by implementing SRSWOR on these clusters. A simple random sample of size s could be generated from these clusters where $s < M$.
- d. Stratified sample:** The large data set D is partitioned into mutually disjoint sets called 'strata'. A simple random sample is taken from each stratum to get stratified data. This method is effective for skewed data.

3. Data Cube Aggregation

This technique is used to aggregate data in a simpler form. Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction.

For example, suppose you have the data of All Electronics sales per quarter for the year 2018 to the year 2022. If you want to get the annual sale per year, you just have to aggregate the sales per quarter for each year. In this way, aggregation provides you with the required data, which is much smaller in size, and thereby we achieve data reduction even without losing any data.

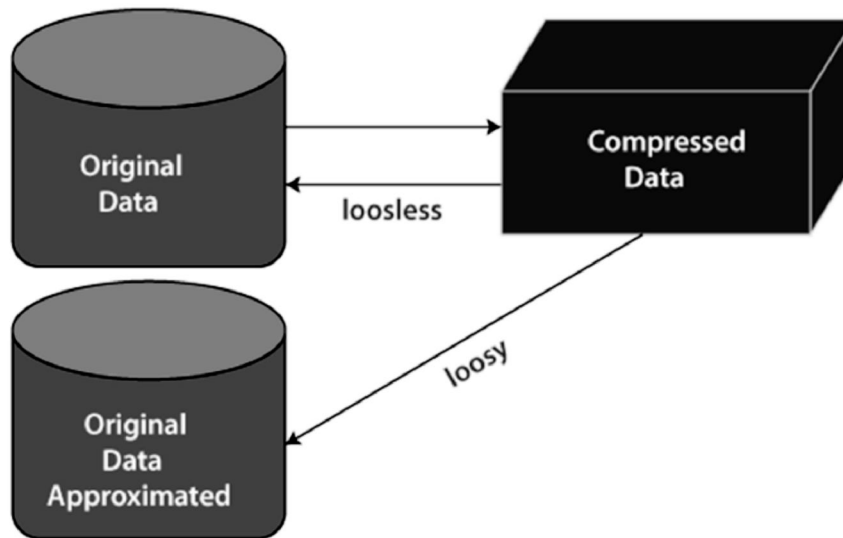


Aggregated Data

The data cube aggregation is a multidimensional aggregation that eases multidimensional analysis. The data cube present precomputed and summarized data which eases the data mining into fast access.

4. Data Compression

Data compression employs modification, encoding, or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form. Data that can be restored successfully from its compressed form is called Lossless compression. In contrast, the opposite where it is not possible to restore the original form from the compressed form is Lossy compression. Dimensionality and numerosity reduction method are also used for data compression.



This technique reduces the size of the files using different encoding mechanisms, such as Huffman Encoding and run-length Encoding. We can divide it into two types based on their compression techniques.

i. Lossless Compression

Encoding techniques (Run Length Encoding) allow a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.

ii. Lossy Compression

In lossy-data compression, the decompressed data may differ from the original data but are useful enough to retrieve information from them. For example, the JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image. Methods such as the Discrete Wavelet transform technique PCA (principal component analysis) are examples of this compression.

5. Discretization Operation

The data discretization technique is used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes with labels of small intervals. This means that mining results are shown in a concise and easily understandable way.

i. Top-down discretization

If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat this method up to the end, then the process is known as top-down discretization, also known as splitting.

ii. Bottom-up discretization

If you first consider all the constant values as split-points, some are discarded through a combination of the neighborhood values in the interval. That process is called bottom-up discretization.

3.7 DATA MINING AND BUSINESS INTELLIGENCE
Q26. What is business intelligence?*Ans :***(Imp.)**

Business Intelligence refers to a set of processes and technologies that convert raw data into usable and meaningful information to make profitable business decisions. It is an umbrella term that combines data mining, data tools, business analytics, data visualization, infrastructure, and best practices to offer quick-to-digest data summaries and aid an organization in making more data-driven decisions. BI serves enterprises to unlock sales and marketing potential, and innovate new business capabilities.

BI is used to drive change with an organization, and help eliminate its inefficiencies by swiftly adapting to changing market dynamics. Business Intelligence systems are primarily data-driven Decision Support Systems or DSS.

Business Intelligence techniques**Data analysis visualization**

Data analysis visualization is all about how you visualize your data. It presents data on dashboards and utilizes customized metrics related to the business to make better decisions based on facts.

Reporting

Business intelligence tools are used for reporting information gathering from all the sources and process it to enable better reporting and financial decision making with a rational mind.

Predictive Analytics

Predictive analytics is all about how do you know a strategy will work? The fact is you don't know, and if you know, not 100 percent. However, with business intelligence, you can create an evidence-based decision to drive business further. Business intelligence enables you to make a reasonable prediction of the latest trends and customer behaviors that impact the organization's overall development.

Q27. What are the differences between Data Mining and Data Intelligence.*Ans :*

S.No.	Business Intelligence	Data Mining
1.	Business intelligence (BI) refers to a technology-driven process that transforms the data into actionable information. Organizations have a huge flow of data coming from their customer end.	The term data mining itself explains its meaning, and it is the mining of important information, patterns, and trends.
2.	The data-driven decisions help in decision-making purposes in the organization.	It finds the answer to a problem in the business.

3.	It has a large dataset processed on relational databases.	It has small data sets processed on a small portion of data.
4.	It represents results on dashboards and reports by charts and graphs with KPI's	It identifies the solution for a problem to be represented as one of the KPI's in dashboards or reports.
5.	It depends on a small scale of past data, and there is no intelligence involved; business management has to decide based on the information.	It focused on a specific problem in business management on small-scale data using various algorithms to determine the solution.
6.	The quality of the solutions is volumetric and presents the accurate solution using data visualizations.	It uses algorithms to identify accurate patterns for a problem and identifies the blind spots.
7.	It shows price value, total cost, profit, etc.	It identifies a solution for a problem creating new KPI's for business intelligence.

3.8 DATA MINING FOR BUSINESS

Q28. How to use Data mining for business analytics.

Ans :

(Imp.)

Data mining for business analytics is the process of extracting valuable information from a vast amount of available corporate or consumer data.

The process of mining data typically consists of 3 broad stages:

1. Pre-processing

This stage involves both data preparation and the initial exploration of available data.

- In data preparation, the data is cleansed, any missing values in the data need to be corrected, or data that remains incomplete may need to be discarded. The idea is to make sure that the data is consistent and can give an overall complete picture when processed.
- At this stage, identification and classification of the various types of data is also accomplished.
- Data integration refers to the process by which data from different sources and disparate formats is brought together, standardized and validated, with duplicates removed – all while ensuring the reliability of the data.
- The next process is data selection, which ensures that you select only that data that is relevant to the goals you have identified for your data mining project. Irrelevant data, or data that is not useful for your purpose, is filtered out at this stage.
- Data transformation is the process of changing the format, structure, or values of data. It can reshape data without changing its actual content. It may involve converting data types into certain standardized formats to improve its compatibility with the rest of the data, adjusting dates and time formats, and renaming tables, columns, etc.

2. Post-processing

This stage involves the creation of data models, validation of these models, and the monitoring of their performance.

- Data modeling is the process by which a model for the data is created, much like an architect's plan for a building. This model serves as a representation of the how the data will be stored in the database and defines the relational tables, primary and foreign keys, and stored procedures. There are three types of data models: conceptual, logical, and physical. A conceptual model aims at establishing the entities and attributes of the data, and their relationships. A logical data model defines the structure of the data elements and establishes the relationships between them. A physical data model describes the database-specific implementation of the data model.
 - The actual process of data mining starts at this stage of data modeling. After assessing the entirety of the data across the enterprise, a comprehensive data model is designed to support the business.
 - Furthermore, data models could be of 3 broad types: descriptive, predictive, or prescriptive models.
 - Descriptive modeling is a technique that describes or summarizes raw data from the past.
 - Predictive modeling is a statistical data analytics technique that is used to predict future behavior. This technique relies on historical data and compares it to the current data to arrive at a prediction of future outcomes.
 - Prescriptive modeling, on the other hand, allows users to prescribe or recommend a number of different possible actions and guide companies towards a solution.
- 3. Deployment:** The next stage involves applying the data model to new data in order to arrive at some predictions and reveal actionable insights. The initial results are evaluated and the model is tested on a variety of different data sets. The results are reviewed for any inconsistencies and reiterations are done until results returned are satisfactory. Once the final model has been validated and verified, operationalization begins. At this stage, visualizations are worked upon since storytelling and visual reporting are very critical in the process. After thorough User Acceptance Testing, the data environment is ready to roll out to beta users, business managers, and executive leaders.

3.9 CLASSIFICATION, ASSOCIATION

Q29. What is classification in data mining?

Ans :

(Imp.)

Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones. Types of Classification Techniques in Data Mining

Before we discuss the various classification algorithms in data mining, let's first look at the type of classification techniques available. Primarily, we can divide the classification algorithms into two categories:

1. Generative
2. Discriminative

Here's a brief explanation of these two categories:

1. Generative

A generative classification algorithm models the distribution of individual classes. It tries to learn the model which creates the data through estimation of distributions and assumptions of the model. You can use generative algorithms to predict unseen data.

A prominent generative algorithm is the Naïve Bayes Classifier.

2. Discriminative

It's a rudimentary classification algorithm that determines a class for a row of data. It models by using the observed data and depends on the data quality instead of its distributions.

Logistic regression is an excellent type of discriminative classifiers.

Q30. What are the applications of Datamining

Ans :

Applications of Classification of Data Mining Systems

- Marketers use classification algorithms for audience segmentation. They classify their target audiences into different categories by using these algorithms to devise more accurate and effective marketing strategies.
- Meteorologists use these algorithms to predict the weather conditions according to various parameters such as humidity, temperature, etc.
- Public health experts use classifiers for predicting the risk of various diseases and create strategies to mitigate their spread.
- Financial institutions use classification algorithms to find defaulters to determine whose cards and loans they should approve. It also helps them in detecting fraud.

Q31. What is association rule in datamining?

Ans :

The Association rule is a learning technique that helps identify the dependencies between two data items. Based on the dependency, it then maps accordingly so that it can be more profitable. Association rule furthermore looks for interesting associations among the variables of the dataset. It is undoubtedly one of the most important concepts of Machine Learning and has been used in different cases such as association in data mining and continuous production, among others.

Types Of Association Rules In Data Mining

There are typically four different types of association rules in data mining. They are

- Multi-relational association rules
- Generalized Association rule
- Quantitative Association Rules

➤ Multi-Relational Association Rule

Also known as MRAR, multi-relational association rule is defined as a new class of association rules that are usually derived from different or multi-relational databases. Each rule under this class has one entity with different relationships that represent the indirect relationships between entities.

➤ Generalized Association Rule

Moving on to the next type of association rule, the generalized association rule is largely used for getting a rough idea about the interesting patterns that often tend to stay hidden in data.

➤ Quantitative Association Rules

This particular type is actually one of the most unique kinds of all the four association rules available. What sets it apart from the others is the presence of numeric attributes in at least one attribute of quantitative association rules. This is in contrast to the generalized association rule, where the left and right sides consist of categorical attributes.

3.10 CAUSE EFFECT MODELING

Q32. Discuss in detail about cause and effect modelling.

Ans :

(Imp.)

There are many different types of causal patterns in the world. Below are six patterns that are embedded in many concepts. Causality in the real world seldom falls into one neat pattern or another. The patterns often work together or different parts of a system entail different patterns—making the causality even more complex!

➤ **Linear Causality**

Cause precedes effect; sequential pattern. Direct link between cause and effect. Has a clear beginning and a clear ending. Effect can be traced back to one cause. One cause and one effect; additional causes or effects turn this pattern into domino causality

➤ **Domino Causality**

Sequential unfolding of effects over time. An extended linear pattern that results in direct and indirect effects. Typically has a clear beginning and a clear ending. Can be branching where there is more than one effect of a cause (and these may go on to have multiple effects and so on.). Branching forms can be traced back to “stem” causes. Anticipating outcomes involves deciding how far to trace effects. Short-sightedness can lead to unintended effects.

➤ **Cyclic Causality**

One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but eventually impacts the first thing). Involves a repeating pattern. Involves feedback loops. May be sequential or may be simultaneous. Typically no clear beginning or ending (Sometimes you can look back in time to a beginning but often that results in the classic ‘which came first, the chicken or the egg’ problem.).

➤ **Spiraling Causality**

One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but eventually impacts the first thing) with amplification or de-amplification of effects. Involves feedback loops. It is sequential as each event is a reaction to the one before it. Often a clear beginning and ending. It is difficult to anticipate outcomes of later feedback loops during earlier feedback loops.

➤ **Relational Causality**

Two things work in relation to each other to cause an outcome. It often involves two variables in comparison to each other. There may be a relationship of balance, equivalence, similarity or there may be a relationship of difference. If one thing changes, so does the relationship, therefore so does the outcome. If two things change but keep the same relationship, the outcome doesn’t change.

➤ **Mutual Causality**

Two things impact each other. The impact can be positive for both, negative for both, or positive for one and negative for the other. The causes and effects are often simultaneous, but can be sequential. May be event-based or may be a relationship over time (such as the moss and the algae in lichen).

Q33. What are the steps involved In cause and effect diagram?

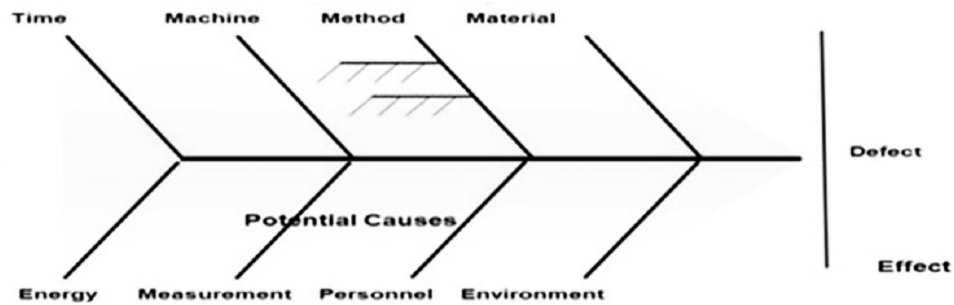
Ans :

(Imp.)

Cause and Effect Diagram

It helps uncover potential root causes by providing structure to cause identification effort. It is also called as fishbone or Ishikawa diagram. It helps in ensuring new ideas being generated during brainstorming by not overlooking any major possible cause.

It should be used for cause identification after clearly defining the problem. It is also useful as a cause - prevention tool by brainstorming ways to maintain or prevent future problems.



They break problems down into small-size pieces and displays possible causes in a graphical manner. They display how various causes interact with each other and uses brainstorming rules when generating ideas. A fishbone diagram development consists of brainstorming, prioritizing and development of an action plan.

Developing Cause and Effect Diagram – It involves the following steps

- Name the problem or effect of interest. Be as specific as possible.
- Write the problem at the head of a fishbone “skeleton”
- Decide the major categories for causes and create the basic diagram on a flip chart or whiteboard.
- Typical categories include the manpower, machines, materials, methods, measurements and environment
- Brainstorm for more detailed causes and create the diagram either by working through each category or open brainstorming for any new input.
- Write suggestions onto self-stick notes and arrange in the fishbone format, placing each idea under the appropriate categories.
- Review the diagram for completeness.
- Eliminate causes that do not apply
- Brainstorm for more ideas in categories that contain fewer items
- Discuss the final diagram. Identify causes which are most critical for follow-up investigation

Short Question and Answers

1. What is mean by Predictive analytics.

Ans :

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.

Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. Predictive analytics is often associated with big data and data science.

Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.

2. What is a Trend Line? Explain the need of trend lines.

Ans :

A trendline is a straight line that connects two or more price points and then extends into the future to act as a line of support and resistance.

In simple words, a trendline is a line that we draw on our chart by connecting the swing highs and swing lows during a 'Trending Market'.

3. What is Paralled Trendline.

Ans :

The Parallel Trendline

- The parallel trendline is used to identify a channel between the highs and the lows.
- If drawn correctly, you can notice the prices are confined between those two parallel lines.
- The bigger the channel, the more significant it is.
- The upper limit of the dominant area act as resistance.
- The lower limit of the dominant area act as support.
- The parallel trendlines can be drawn during an uptrend, downtrend or sideways market.



4. What are the fundamental assumptions of Regression Analysis.*Ans :*

Linear regression analysis is based on six fundamental assumptions :

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
 2. The independent variable is not random.
 3. The value of the residual (error) is zero.
 4. The value of the residual (error) is constant across all observations.
 5. The value of the residual (error) is not correlated across all observations.
 6. The residual (error) values follow the normal distribution.
-

5. What is simple Linear Regression.*Ans :*

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- Y – Dependent variable
 - X – Independent (explanatory) variable
 - a – Intercept
 - b – Slope
 - ϵ – Residual (error)
-

6. What is Linear Regression.*Ans :*

- Linear regression is a statistical regression method which is used for predictive analysis.
 - It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
 - It is used for solving the regression problem in machine learning.
 - Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
-

7. What is multiple linear regression.*Ans :*

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

8. What is Predictive Modelling.*Ans :*

Predictive modeling is a predictive analysis tool. It is widely used by companies to determine the viability of a new venture, project, or proposal. It applies statistical and analytical tools for analyzing current data and historical data and determines future outcomes.

- Predictive modeling uses known results to create, process, and validate a model that can be used to make future predictions.
 - Regression and neural networks are two of the most widely used predictive modeling techniques.
 - Companies can use predictive modeling to forecast events, customer behavior, and financial, economic, and market risks.
-

9. What is Decision Tree.*Ans :*

This method is commonly used for data mining. A flowchart representing an inverted tree is formulated. Here the internal node splits into branches that list out two or more possible decisions, and each decision is further subdivided to show other possible outcomes. This technique helps in selecting the best option.

10. What are the advantages of Predictive Modeling.*Ans :***Advantages****➤ Easy to generate actionable insights**

Predictive modeling allows you to view information about your data that you might not see otherwise, enabling you to make more informed decisions.

➤ Can test different scenarios

Data can be manipulated or changed to test various scenarios to assess the influence changes might have on your data and models.

➤ Increases decision-making speed

Decisions can be reached much faster because millions of data points can be analyzed much quicker, and future trends or circumstances can be theorized within minutes or hours.

11. Define Data Mining.*Ans :*

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.

12. What is Data Exploration.*Ans :*

Data exploration refers to the initial step in data analysis. Data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, to understand the nature of the data better.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

Data without a question is simply information. Asking a question of data turns it into an answer. Data with the right questions and exploration can provide a deeper understanding of how things work and even enable predictive abilities.

13. What are the differences between data exploration and data mining.

Ans :

S.No.	Data Mining	S.No.	Data Exploration
1.	Data mining is also named knowledge web discovery in databases, extraction, data /pattern analysis, and information harvesting.	1.	Data Exploration is used interchangeably with exploration, web scraping, web crawling, data retrieval, data harvesting, etc.
2.	Data mining studies are mostly on structured data.	2.	Data Exploration usually retrieves data out of un structured or poorly structured data sources.
3.	Data mining aims to make available data more useful for generating insights.	3.	Data Exploration is to collect data and gather them into a place where they can be stored or further processed.
4.	Data mining is based on mathematical methods to reveal patterns or trends.	4.	Data Exploration is based on programming languages or data Exploration tools to crawl the data sources.
5.	The purpose of data mining is to find facts that are previously unknown or ignored,	5.	Data Exploration deals with existing information.
6.	Data mining is much more complicated and requires large investments in staff training.	6.	Data Exploration can be extremely easy and cost-effective when conducted with the right tool.

14. What is Data Reduction? What are the benefits of data reduction?

Ans :

Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Data reduction does not affect the result obtained from data mining. That means the result obtained from data mining before and after data reduction is the same or almost the same.

Benefits of Data Reduction

The main benefit of data reduction is simple: the more data you can fit into a terabyte of disk space, the less capacity you will need to purchase. Here are some benefits of data reduction, such as:

- Data reduction can save energy.
- Data reduction can reduce your physical storage costs.
- And data reduction can decrease your data center track.

15. What is business intelligence.*Ans :*

Business Intelligence refers to a set of processes and technologies that convert raw data into usable and meaningful information to make profitable business decisions. It is an umbrella term that combines data mining, data tools, business analytics, data visualization, infrastructure, and best practices to offer quick-to-digest data summaries and aid an organization in making more data-driven decisions. BI serves enterprises to unlock sales and marketing potential, and innovate new business capabilities.

BI is used to drive change with an organization, and help eliminate its inefficiencies by swiftly adapting to changing market dynamics. Business Intelligence systems are primarily data-driven Decision Support Systems or DSS.

16. What are the differences between Data Mining and Data Intelligence.*Ans :*

S.No.	Business Intelligence	Data Mining
1.	Business intelligence (BI) refers to a technology-driven process that transforms the data into actionable information. Organizations have a huge flow of data coming from their customer end.	The term data mining itself explains its meaning, and it is the mining of important information, patterns, and trends.
2.	The data-driven decisions help in decision-making purposes in the organization.	It finds the answer to a problem in the business.
3.	It has a large dataset processed on relational databases.	It has small data sets processed on a small portion of data.
4.	It represents results on dashboards and reports by charts and graphs with KPI's	It identifies the solution for a problem to be represented as one of the KPI's in dashboards or reports.
5.	It depends on a small scale of past data, and there is no intelligence involved; business management has to decide based on the information.	It focused on a specific problem in business management on small-scale data using various algorithms to determine the solution.
6.	The quality of the solutions is volumetric and presents the accurate solution using data visualizations.	It uses algorithms to identify accurate patterns for a problem and identifies the blind spots.
7.	It shows price value, total cost, profit, etc.	It identifies a solution for a problem creating new KPI's for business intelligence.

17. What is classification in data mining.*Ans :*

Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones. Types of Classification Techniques in Data Mining.

Before we discuss the various classification algorithms in data mining, let's first look at the type of classification techniques available. Primarily, we can divide the classification algorithms into two categories:

1. Generative
 2. Discriminative
-

18. What are the applications of Datamining.*Ans :*

Applications of Classification of Data Mining Systems

- Marketers use classification algorithms for audience segmentation. They classify their target audiences into different categories by using these algorithms to devise more accurate and effective marketing strategies.
- Meteorologists use these algorithms to predict the weather conditions according to various parameters such as humidity, temperature, etc.
- Public health experts use classifiers for predicting the risk of various diseases and create strategies to mitigate their spread.
- Financial institutions use classification algorithms to find defaulters to determine whose cards and loans they should approve. It also helps them in detecting fraud.

Choose the Correct Answers

1. _____ involves predicting a response with meaningful magnitude, such as quantity sold, stock price, or return on investment. [a]
(a) Regression (b) Summarization
(c) Clustering (d) Classification
2. Which of the following involves predicting a categorical response? [d]
(a) Regression (b) Summarization
(c) Clustering (d) Classification
3. _____ is proprietary tool for predictive analytics. [b]
(a) R (b) SAS
(c) SSAS (d) EDR
4. Which of the following is not a major data analysis approaches? [b]
(a) Business Intelligence
(b) Predictive Intelligence
(c) Data Mining
(d) Text Analytics
5. The Process of describing the data that is huge and complex to store and process is known as _____. [b]
(a) Analytics mining (b) Big data
(c) Data cleaning (d) None of the above
6. Data Analysis is a process of? [d]
(a) inspecting data (b) transforming data
(c) cleaning data (d) All of the above
7. To glean insights from the data, many analysts and data scientists rely on _____. [b]
(a) Data warehouse (b) Data visualization
(c) Data mining (d) All of the above
8. Correlation is the relationship between two variables _____. [a]
(a) Two (b) One
(c) Zero (d) None
9. Data Analysis is a process of _____. [d]
(a) Data Cleaning (b) Transforming of data
(c) Inspecting data (d) All of the above
10. Data Analytics uses _____ to get insights from data. [a]
(a) Statistical methods (b) Statistical figures
(c) Numerical aspects (d) None of the mentioned above

Fill in the blanks

1. The Process of describing the data that is huge and complex to store and process is known as _____.
2. A good data analytics solution includes a viable self-service _____.
3. To glean insights from the data, many analysts and data scientists rely on _____.
4. For each value of the _____ distribution of the dependent variable must be normal.
5. Least Square Method uses _____ Regression.
6. Linear Regression is the supervised machine learning model in which the model finds the best fit _____ between the independent and dependent variable.
7. _____ is an essential process in which the intelligent methods are applied to extract data patterns.
8. The classification or mapping of a class using a predefined class or group is called _____.
9. _____ are the data objects that don't comply with the general model or behaviour of the available data.
10. _____ is the out put of KDD.

ANSWERS

1. Big Data
2. Data Wrangling
3. Data Visualization
4. Independent variable
5. Linear
6. Linear Line
7. Data Mining
8. Data Discrimination
9. Outlier analysis
10. Useful information

One Mark Answers

1. Trend Line

Ans :

Trend lines are easily recognizable lines that traders draw on charts to connect a series of prices together or show some data's best fit.

2. Regression

Ans :

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

3. Data Mining

Ans :

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis.

4. Data exploration

Ans :

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more.

5. Business Intelligence

Ans :

Business intelligence (BI) is software that ingests business data and presents it in user-friendly views such as reports, dashboards, charts and graphs.

UNIT IV

PREScriptive ANALYTICS

Overview of Linear Optimization, Non Linear Programming Integer Optimization, Cutting Plane algorithm and other methods, Decision Analysis – Risk and uncertainty methods - Text analytics Web analytics.

4.1 PRESCRIPTIVE ANALYTICS

4.1.1 Introduction

Q1. What is prescriptive analytics? And list out the benefits of prescriptive analytics.

Ans : (Imp.)

Prescriptive analytics is a statistical method that focuses on finding the ideal way forward or action necessary for a particular scenario, based on data. Prescriptive analytics uses both descriptive and predictive analytics but the focus here remains on actionable insights rather than data monitoring. The input of prescriptive analytics is the outcome of predictive analytics algorithms. You not only predict what the future holds, but you leverage that prediction to take the best course of action for the future. A more formal definition is that prescriptive analytics is a statistical approach utilized to generate recommendations and aid decision-making based on the computational outcomes of algorithmic models. Following are the benefits of prescriptive analytics.

i) Revenue Generation

Prescriptive analytics can help a business understand what their customers want to buy and why. These outcomes can be arrived at with detailed and timely information on customers and their purchasing journeys. This will help managers accelerate their sales cycles and be able to find and open up new avenues for cross and up-selling.

ii) Gross Margin Management

The prescriptive analytics models provide insights into the optimal mix of products that an organization should focus its attention on. The model for this can be created based on current as well as anticipated market conditions and customer

purchase patterns. It will ensure higher business productivity and profitability.

Expense Reduction

With the right algorithmic model, a company will be able to ensure that they have better inventory management systems in place. This will help in reduction of costs for long term stock storage. It also brings down the number of manual processes and costs involved. An organization will also have better control over their expenses and transparency across the board.

Q2. What are the challenges of prescriptive analytics.

Ans :

Challenges with Prescriptive Analytics

Prescriptive analytics is powerful but it does present unique challenges. Here is a look at the top five issues you may encounter:

i) Difficult to Define a Fitness Function

To optimize results, every prescriptive analytical model requires a fitness function (how 'fit' the solution is for the problem) to be well defined. A fitness function forms the base to help obtain the ideal set of solutions. However, arriving at this function can be difficult because it requires an in-depth understanding of the business from multiple angles. The best approach to handling this is to involve business partners early on to ensure that the algorithms you create are accurate to business outcomes.

ii) Human bias in models

Unfortunately, one of the biggest inhibitors to the growth of prescriptive analytics is that most

models are human written meaning they have an inherent bias. In fact, most discussions on prescriptive analytics talk about this unfortunate fact. What this means is that the algorithms are set up in a certain way, not based on data but based on a domain expert's opinions. One of the future fixes for this would be to generate models using machine learning based on the data that is flowing in. That would be an ideal way to cross out any potential human bias.

iii) Complex Constraints

Parameters need to be in place to be able to build a prescriptive analytical model that functions towards finding a range of solutions. Often there are constraints on these parameters. This happens when the solution it arrives at cannot be achieved. This could happen because of a negative length or because of a business rule that doesn't allow a price change beyond a certain amount. There are two ways to handle this – make sure the optimizer knows of these rules or have them coded into the fitness function.

4.2 OVERVIEW OF LINEAR OPTIMIZATION

Q3. What is Linear optimization? What are the applications of linear optimization.

Ans : (Imp.)

Linear optimization is a method applicable for the solution of problems in which the objective function and the constraints appear as linear functions of the decision variables. The constraint equations may be in the form of equalities or inequalities. In other words, linear optimization determines the way to achieve the best outcome (for example, to maximize profit or to minimize cost) in a given mathematical model and given some lists of requirements represented as linear equations.

Applications

Linear optimization can be applied to numerous fields, in business or economics situations, and also in solving engineering problems. It is useful in modeling diverse types of problems in planning, routing, scheduling, assignment and design.

i) Petroleum refineries

One of the early industrial applications of linear optimization has been made in the petroleum refineries. An oil refinery has a choice of buying crude oil from different sources with different compositions at different prices. It can manufacture different products, such as diesel fuel, gasoline and aviation fuel, in varying quantities. A mix of the purchased crude oil and the manufactured products is sought that gives the maximum profit.

ii) Manufacturing firms

The sales of a firm often fluctuate, therefore a company has various options. It can either build up an inventory of the manufactured products to carry it through the period of peak sales, or to pay overtime rates to achieve higher production during periods of high demand. Linear optimization takes into account the various cost and loss factors and arrive at the most profitable production plan.

iii) Food-processing industry

Linear optimization has been used to determine the optimal shipping plan for the distribution of a particular product from different manufacturing plants to various warehouses.

iv) Telecommunications

The optimal routing of messages in a communication network and the routing of aircraft and ships can also be determined by linear optimization method.

Q4. What are the characteristics of linear optimization.

Ans :

The characteristics of a linear optimization problem are:

1. The objective function is of the minimization type
2. All the constraints are of the equality type
3. All the decision variables are non-negative

Any linear optimization problem can be expressed in the standard form by using the following transformation:

- 1) The maximization of a function $f(x_1, x_2, \dots, x_n)$ is equivalent to the minimization of the negative of the same function.

For example

Minimize

$$f = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$$

is equivalent to

Maximize

$$f_2 = -c_1 x_1 - c_2 x_2 - \dots - c_n x_n$$

Consequently, the objective function can be stated in the minimization form in any linear optimization problem.

- 2) If a constraint appears in the form of a “less than or equal to” type of inequality as

$$a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kn}x_n \leq b_k$$

it can be converted into the equality form by adding a non-negative slack variable x_{n+1} as follows:

$$a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kn}x_n + x_{n+1} = b_k$$

Similarly, if the constraint is in the form of a “greater than or equal to” type of inequality, it can be converted into the equality form by subtracting the surplus variable.

- 3) In most engineering optimization problems, the decision variables represent some physical dimensions, hence the variables will be non-negative.

However, a variable may be unrestricted in sign in some problems. In such cases, an unrestricted variable (which can take a positive, negative or zero value) can be written as the difference of two non-negative variables.

Thus, if x_j is unrestricted in sign, it can be written as $x_j = x'_j - x''_j$

where

$$0 \leq x'_j \leq \text{and } 0 \leq x''_j$$

It can be seen that x_j will be negative, zero or positive, depending on whether x'_j is greater than, equal to, or less than x''_j .

Q5. Explain Linear optimization problem using Ms-Excel.

Ans :

As stated in the Linear Optimization section example above, there are three categories of information needed for solving an optimization problem in Excel: an objective function, constraints, and decision variables.

We will use the following example to demonstrate another application of linear optimization. We will be optimizing the profit for Company X's trucking business.

To reach capacity, Company X must move 100 tons of cargo per day by truck. Company X's trucking fee is \$250/ton. Besides the weight constraint, the company can only move 50,000 ft³ of cargo per day due to limited volume trucking capacity. The following amounts of cargo are available for shipping each day:

Cargo Weight (tons)	Weight (tons)	Volume (ft./ton)
1	30	550
2	40	800
3	50	400

Maximize the profit for Company X Set up this problem:

Objective Function (\$/week)

$$\text{Profit} = 250 * (\text{Cargo1} + \text{Cargo2} + \text{Cargo3})$$

Decision Variables (freight in tons)

Cargo 1

Cargo 2

Cargo 3

Constraints

$$\text{Weight} : \text{Cargo 1} + \text{Cargo2} + \text{Cargo3} \leq 100$$

$$\text{Volume} : 550 * \text{Cargo1} + 800 * \text{Cargo2} + 400 * \text{Cargo3} \leq 50000$$

$$\text{Amount 1} : \text{Cargo 1} \leq 30$$

$$\text{Cargo 1} \geq 0$$

$$\text{Amount 2} : \text{Cargo 2} \leq 40$$

$$\text{Cargo 2} \geq 0$$

$$\text{Amount 3} : \text{Cargo 3} \leq 50$$

$$\text{Cargo 3} \geq 0$$

Set Up the Problem Using Excel

Solver is an Add-in for Microsoft Excel. It will be used to optimize Company X's profit. If 'Solver' is not on the 'Tools' menu in Excel, then use the following steps to enable it:

For Windows 2007

- Click on the Office button at the top left corner of the screen. Click on the "Excel Options" button on the bottom right of the menu.
- Select "Add-ins." Make sure that "Excel Add-ins" is selected in the "Manage" drop down list. Click "Go."
- A new window will appear entitled "Add-ins." Select "Solver Add-in" by checking the box. Click "Go."
- A Configuration window will appear. Allow Office to install the Add-in.
- The solver has been successfully installed. (See Windows Help for more instruction.)

Use the figure below to set up your Excel worksheet.

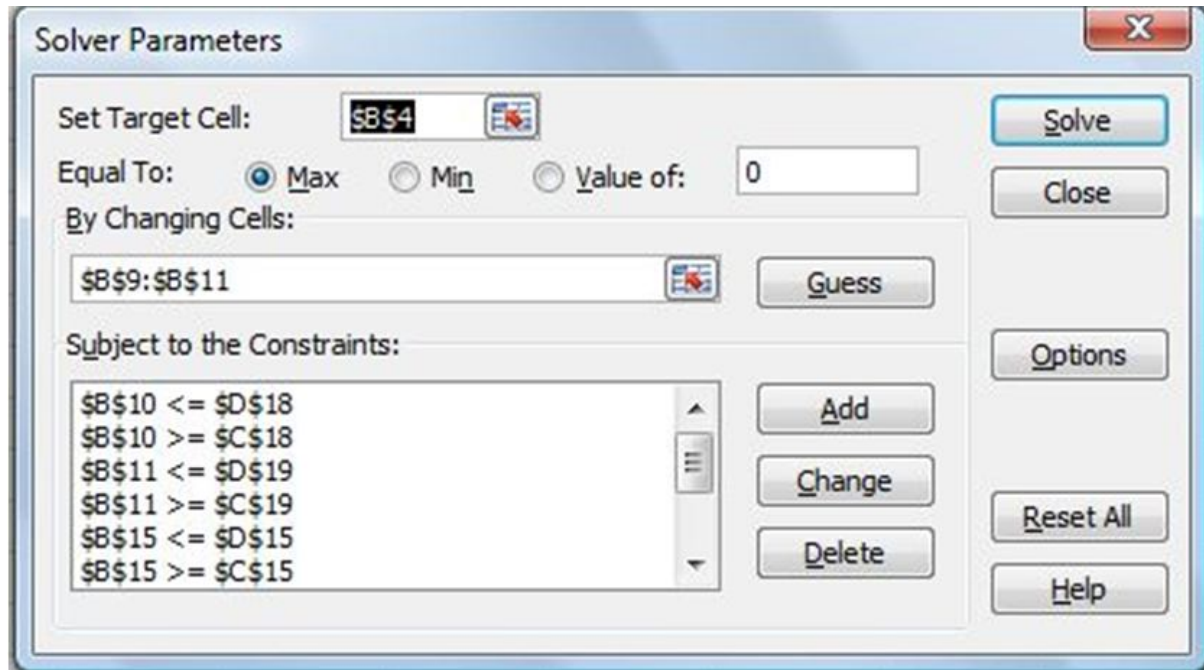
	A	B	C	D
1				
2				
3	Objective			
4	Profit	0		
5				
6	Variables			
7				
8				
9	Cargo1	0		
10	Cargo2	0		
11	Cargo3	0		
12				
13				
14	Constraints	Formula	Min	Max
15	Weight	0	0	100
16	Volume	0	0	50000
17	Amount1	0	0	30
18	Amount2	0	0	40
19	Amount3	0	0	50

Enter in the following formulas to the cells as shown below:

	A	B	C	D
1				
2				
3	Objective			
4	Profit	=250*SUM(B9:B11)		
5				
6	Variables			
7				
8				
9	Cargo1	0		
10	Cargo2	0		
11	Cargo3	0		
12				
13				
14	Constraints	Formula	Min	Max
15	Weight	=SUM(B9:B11)	0	100
16	Volume	=SUM(B9:B11)	0	50000
17	Amount1	=SUM(B9:B11)	0	30
18	Amount2	=SUM(B9:B11)	0	40
19	Amount3	=SUM(B9:B11)	0	50

Running Solver

- Click on the “Data” tab and select “Solver”. A dialog box will appear.
- Enter the parameters as shown in the figure below.



Detailed steps are as follows:

- In “Set Target Cell,” enter the cell corresponding to the company’s profit (B4).
- Select “Max” under “Equal To.”
- Click on the “Options” tab and check the “Assume Linear Model” box.
- For “By Changing Cells,” select the cells in column B corresponding to the cargo amounts (B9, B10, B11).
- To add constraints, select “Add” under “Subject to the Constraints” A dialog box will open.
- In the “Cell Reference:” field, enter the cell location of the decision value that is subject to constraint (i.e. B9).
- Use the pull-down menu in the middle to select the appropriate inequality relation (i.e. <=)
- In the “Constraint:” field, enter the cell location of the constraint value (i.e. D17).
- Continue to click the “Add.”
- Repeat above steps until all of the constraints are entered. Then click “OK.”
- When all the proper settings have been entered, click “Solve.”
- A “Solver Results” box will appear. Select “Keep Solver Solution” and click “OK.”

The solved worksheet is below.

	A	B	C	D
1				
2				
3	Objective			
4	Profit	24218.75		
5				
6	Variables			
7				
8				
9	Cargo1	30		
10	Cargo2	16.875		
11	Cargo3	50		
12				
13				
14	Constraints	Formula	Min	Max
15	Weight	96.875	0	100
16	Volume	50000	0	50000
17	Amount1	0	0	30
18	Amount2	0	0	40
19	Amount3	0	0	50

Excel's solver program allows us to analyze how our profit would change if we had an alteration in our constraint values. These values can change due to a variety of reasons such as more readily available resources, technology advancements, natural disasters limiting resources, etc.

Looking at Example 1 above, we will now walk through the steps on how to create a sensitivity report.

After clicking "solve" in excel, a solver results dialogue box appears as seen below.

The screenshot shows the Microsoft Excel interface with the Solver Results dialog box open. The dialog box indicates that a solution has been found and all constraints and optimality conditions are satisfied. It offers options to keep the solver solution or restore original values, and includes buttons for OK, Cancel, Save Scenario..., and Help. A Reports section on the right shows 'Answer', 'Sensitivity', and 'Limits' as available reports.

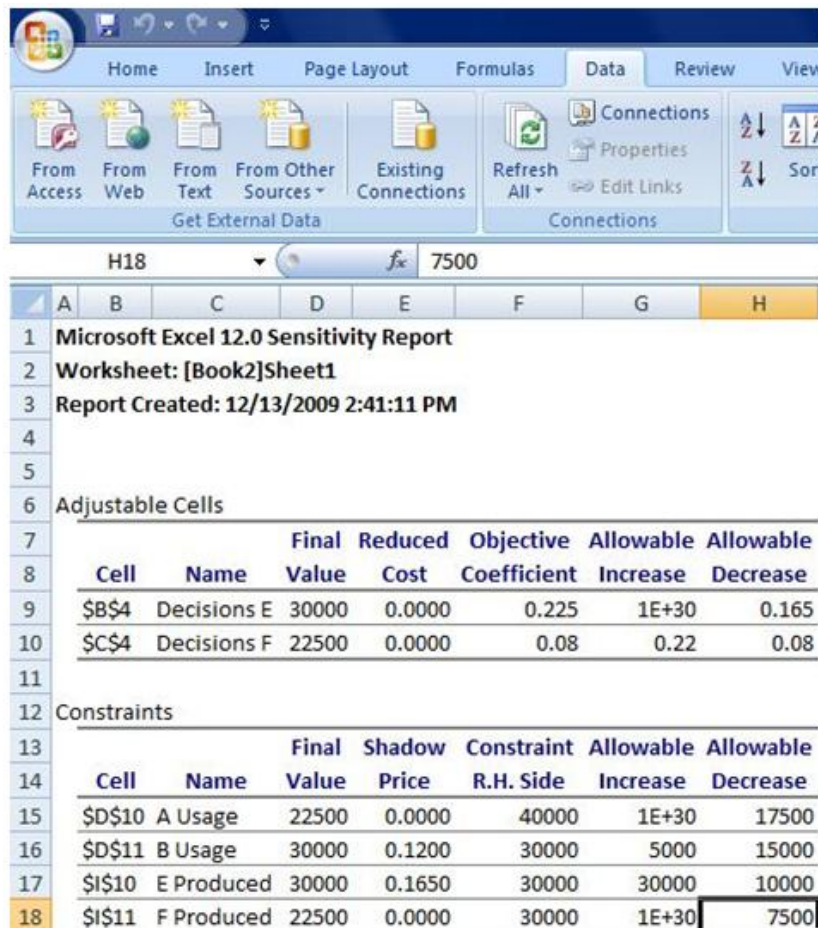
The worksheet in the background displays the following data:

	A	B	C	E	F
1	Profit	8000			
4	Decisions	30000	22500		
5	Profit / lb	0.225	0.08		
6	Fixed Cost	350	200		

Constraints:			(LHS)		(RHS)
			Usage	Inequality	Constraints
A	1	1	22500	<=	40000
B	1	2	30000	<=	30000
Total	2	3			

			(LHS)		(RHS)
			Produced	Inequality	Constraints
E			30000	<=	30000
F			22500	<=	30000

There is a list of three options on the right; answer, sensitivity, and limits. Select the sensitivity option before clicking ok. A new tab will be generated in the worksheet titled "sensitivity 1." A view of the sensitivity report within the tab is seen below. As you can see, two tables are generated. For this example, resource A and product F are non-binding as shown with a shadow price of 0 and an infinite allowable increase. The allowable decrease is the amount the capacity changes until the final value is reached. Past this point the constraint would become a binding constraint. For the constraining variables (resource B and product E), their constraints are binding. Regarding resource B, if its constraint was increased by up to 5000 or decreased by up to 15000, this would have a linear effect on profit within this range. For each unit increase or decrease, the profit will change by 12 cents per unit, respectively. The same is true if our capacity for product E changes with its allowable values and shadow price on the table.



The screenshot shows the Microsoft Excel 12.0 Sensitivity Report. The report is titled "Microsoft Excel 12.0 Sensitivity Report" and "Worksheet: [Book2]Sheet1". It was created on 12/13/2009 at 2:41:11 PM. The report is divided into two main sections: "Adjustable Cells" and "Constraints".

Adjustable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$B\$4	Decisions E	30000	0.0000	0.225	1E+30	0.165
\$C\$4	Decisions F	22500	0.0000	0.08	0.22	0.08

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
\$D\$10	A Usage	22500	0.0000	40000	1E+30	17500
\$D\$11	B Usage	30000	0.1200	30000	5000	15000
\$I\$10	E Produced	30000	0.1650	30000	30000	10000
\$I\$11	F Produced	22500	0.0000	30000	1E+30	7500

Q6. Explain Linear optimization problem using simplex method.

Ans :

Instead of solving linear optimization problems using graphical or computer methods, we can also solve these problems using a process called the Primal Simplex Algorithm. The Primal Simplex Algorithm starts at a Basic Feasible Solution (BFS), which is a solution that lies on a vertex of the subspace contained by the constraints of the problem. In the Graph in Example 1, this subspace refers to the shaded region of the plot. Essentially, after determining an initial BFS, the Primal Simplex Algorithm moves through the boundaries from vertex to vertex until an optimal point is determined.

The basic procedure is the following

1. Find a unit basis.
2. Set-up the problem in standard form using a canonical tableau.
3. Check optimality criterion.
4. If criterion passes, then stop, solution has been found.
5. Select an entering variable among the eligible variables.
6. Perform pivot step.
7. Go back to 1.

For simplicity, we will make the following assumptions

1. The optimum lies on a vertex and is not unbounded on an extreme half-line.
2. The constraints are equations and not also inequalities.
3. In the case that the constraints are inequalities, slack variables will need to be introduced. Although the process is not very different in this case, we will ignore this to make the algorithm slightly less confusing.
4. Decision variables are required to be nonnegative.
5. The problem is a minimization problem. To turn a maximization problem into a minimization problem, multiply the objective function by -1 and follow the process to solve a minimization problem.

We will begin with the following example

Objective Function: Minimize $z = -x_5 - 8x_6$

Subject to the constraints

$$x_1 - x_5 + x_6 = 2$$

$$x_2 + x_5 + x_6 = 1$$

$$x_3 + 2x_5 + x_6 = 5$$

$$x_4 + x_6 = 0$$

$$x_i \geq 0$$

First we begin by finding a unit basis

A shortcut method to finding this unit basis is putting numbers in for each variable so that every constraint equation is satisfied.

In this case, setting $x_4 = 0$, $x_5 = 0$, and $x_6 = 0$ will satisfy the final equation and also set the values for x_1 , x_2 , and x_3 to 2, 1, and 5, respectively. Remember, these decision variables must be nonnegative.

Set up the canonical tableau in the following form:

x_1	x_2	x_3	x_4	x_5	x_6	$-z$	b
1	0	0	0	-1	1	0	2
0	1	0	0	1	1	0	1
0	0	1	0	2	1	0	5
0	0	0	1	0	1	0	0
0	0	0	0	-1	-8	1	0

As you can see, the first four rows correspond to the constraints, while the final row corresponds to the objective function. The “b” column corresponds to the right hand side (RHS) of the constraints. As you can see, the “-z” column is on the left hand side (LHS) of the equation, rather than the RHS.

First, we should perform pivot steps so that the tableau corresponds to the unit basis we found earlier. By performing pivot steps on x_1 , x_2 , x_3 , and x_4 , we will reach the feasible point where $(x_1, x_2, x_3, x_4, x_5, \text{ and } x_6) = (2, 1, 5, 0, 0, 0)$. Because x_4 , x_5 , and x_6 all equal zero, the pivot step on x_4 can actually be done on x_5 or x_6 but in this example, we used x_4 . These pivot steps can be performed on any row as long as they are all different rows. In this example, we performed pivot steps on $(x_1, 1)$, $(x_2, 2)$, $(x_3, 3)$, $(x_4, 4)$ using the Pivot and Gauss-Jordan Tool at people.hofstra.edu/Stefan_Waner/RealWorld/tutorialsf1/scriptpivot2.html. To use this tool, place the cursor on the cell that you wish to pivot on, and press “pivot”.

After four pivot steps, the tableau will look like this:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	0	0	0	-1	1	0	2
0	1	0	0	1	1	0	1
0	0	1	0	2	1	0	5
0	0	0	1	0	1	0	0
0	0	0	0	-1	-8	1	0

As you can see, this is identical to the initial tableau, as x_1 , x_2 , x_3 , and x_4 were set up such that an initial feasible point was already chosen.

The optimality criterion states that if the vector in the bottom left of the tableau is all positive, then an optimal solution exists in the “b” column vector, with the value at the bottom of the “b” column vector as the negative of the value of the objective function at that optimal solution. If this is not true, then a pivot step must be performed. In this example, clearly, a pivot step must be performed.

Next, we need to choose an entering variable. We want to choose an entering variable that has a negative element in the bottom row, meaning that the objective value could be improved if that variable was nonzero in the solution. So, we will choose in this example. Now, we must calculate ratios of each RHS coefficient divided by the coefficient of the entering variable in that row. In this case, the vector corresponding to this calculation would equal $(2/-1, 1/1, 5/2, 0/0)$. We cannot pivot on a zero element, so we cannot pivot on the fourth row. We want to keep the RHS positive, so we cannot pivot on the first row. We must choose the minimum nonnegative ratio to remain at a feasible solution, so we choose the second row in the x_5 column, which has a ratio of $1/1$.

After the pivot step

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	1	0	0	0	2	0	3
0	1	0	0	1	1	0	1
0	-2	1	0	0	-1	0	3
0	0	0	1	0	1	0	0
0	1	0	0	0	7	1	1

As we can see, x_6 has a negative coefficient in the bottom row, indicating the same step must be repeated on that column. We calculate ratios for that column, and get: $(3/2, 1/1, 3/-1, 0/1)$. Consequently, we choose to pivot on the fourth row because it corresponds to the minimum nonnegative ratio of 0.

After another pivot step

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	1	0	-2	0	0	0	3
0	1	0	-1	1	0	0	1
0	-2	1	1	0	0	0	3
0	0	0	1	0	1	0	0
0	1	0	7	0	0	1	1

Because the bottom row is all positive, we are now at an optimal solution. To understand this final tableau, we look at each column for variables that only have one “1” in the column. If the column has only one “1”, the RHS value in that row is the value of that variable. In this case, $x_1 = 3$, $x_3 = 3$, and $x_5 = 3$. Any variable that does not have just a single “1” in the column is equal to zero. So, the optimal solution is $(x_1, x_2, x_3, x_4, x_5, \text{ and } x_6) = (3, 0, 3, 0, 1, 0)$, and the optimal value is $z = -1$.

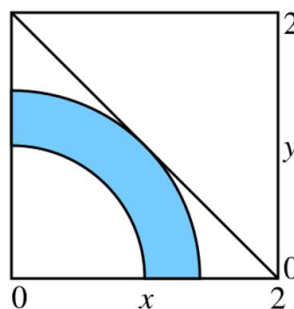
4.3 NON LINEAR PROGRAMMING INTEGER OPTIMIZATION

Q7. What is mean by non linear optimization?

Ans :

(Imp.)

Nonlinear programming is the process of solving optimization problems that concern some of the nonlinear constraints or nonlinear objective functions. It involves minimizing or maximizing a nonlinear objective function subject to bound constraints, linear constraints, nonlinear constraints, etc. These constraints can be inequalities or equalities. In addition, nonlinear programming helps in analyzing design tradeoffs, selecting optimal designs, computing optimal trajectories and portfolio optimization and model calibration in computation finance.



There are two types of nonlinear programming as follows.

i) Unconstrained Nonlinear Programming

Unconstrained nonlinear programming involves finding a vector x that is a local minimum to the nonlinear scalar function $f(x)$. Quasi-Newton, Nelder Mead, and Trust-region are some common unconstrained nonlinear programming algorithms.

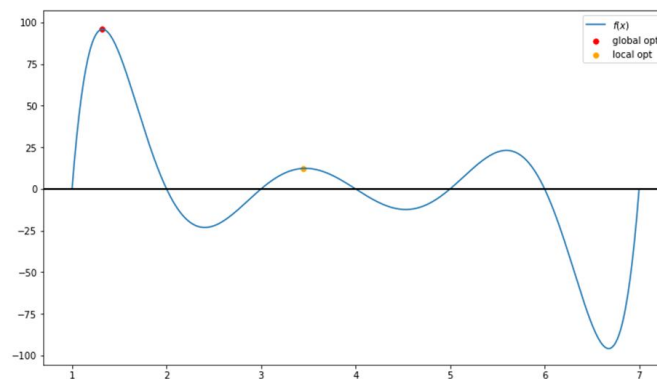
ii) Constrained Nonlinear Programming

Constrained nonlinear programming involves finding a vector x that minimizes a nonlinear function $f(x)$ subject to one or more constraints. Interior-point, sequential quadratic programming, and trust region reflective are some common constrained nonlinear programming algorithms.

Q8. Discuss in detail about non linear optimization*Ans :*

In many optimization models the objective and/or the constraints are nonlinear functions of the decision variables. Such an optimization model is called a nonlinear programming (NLP) model.

- When you solve an LP model, you are mostly guaranteed that the solution obtained is an optimal solution and a sensitivity analysis with shadow price and reduced cost is available.
- When you solve an NLP model, it is very possible that you obtain a suboptimal solution. This is because a nonlinear function can have local optimal solutions where
- A local optimum is better than all nearby points that are not the global optimal solution
- A global optimum is the best point in the entire feasible region



However, convex programming, as one of the most important type of nonlinear programming models, can be solved to its optimality.

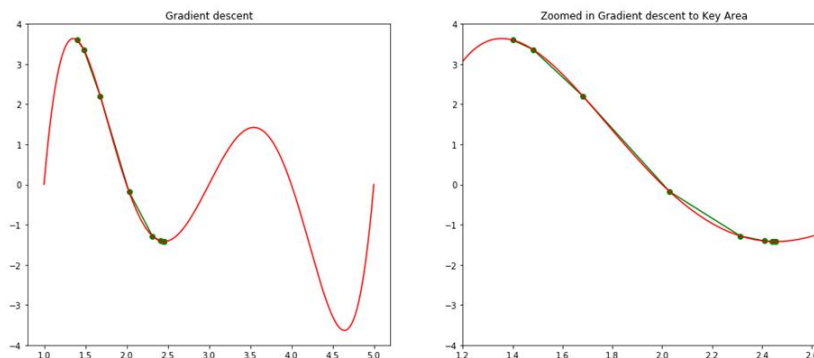
Gradient Decent Algorithm

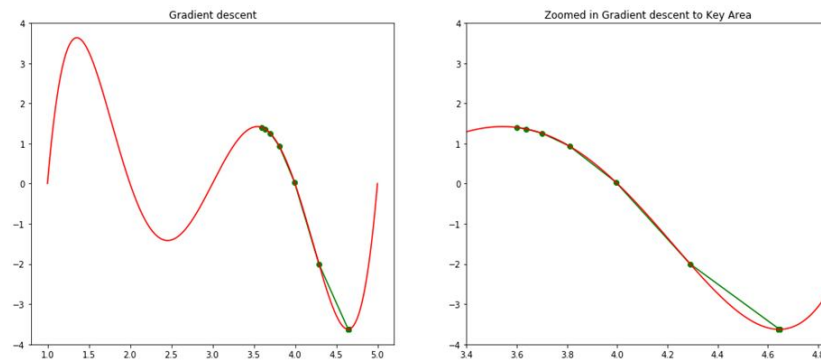
The gradient decent algorithm performs as the core part of many general purpose algorithms for solving NLP models. Consider a function

$$f(x) = (x - 1) \times (x - 2) \times (x - 3) \times (x - 4) \times (x - 5)$$

The gradient decent algorithm is typically used to find a local minimum with given initial solution. In our visual demonstration, a global optimal is found by using 3.6 as initial solution.

As it is usually impossible to identify the best initial solution, multistart option is often used in NLP solver, which will randomly select a large amount of initial solutions and run solver parallelly.





4.4 CUTTING PLANE ALGORITHM AND OTHER METHODS

Q9. What is Cutting plane algorithm?

Ans :

(Imp.)

A cutting plane algorithm is generally used to search for valid inequalities that cut-off the noninteger solutions in two cases, when the set of constraints in our integer programming model is too large, and when the inequality constraints in the original integer programming model are not sufficient to yield an integer solution. The basic idea of the cutting plane method is iteratively refining the search region by introducing linear inequalities, known as cuts, maintaining the original feasible region. If the mathematical model is linear, an extreme or corner point in the feasible region can be the optimal solution, which may or may not be integer solution. If it is not integer, a linear inequality, in other words a cut, can be found to cut away a part of the feasible region, so as to separate the optimum solution. Therefore, a new separation problem is introduced and a cut is added to the relaxed LP model, which makes the existing noninteger solution no longer in the feasible region. This cutting process is repeated until the optimal solution found is also an integer solution.

The steps of the cutting plane algorithm can be explained as follows:

1. The relaxed integer programming problem (the problem with continuous variables instead of discrete/integer variables) is solved.
2. Stop, if all variables in resulting solution have integer values, which means that it is the optimum.
3. Otherwise, generate a cut, that is, a constraint in the form of a linear inequality, which is satisfied by all feasible integer solutions.
4. Add this new constraint to the model, resolve the problem, and go back to step 2.

For example, you might want to maximize your profit and minimize your environmental impact, or you might have several budget or capacity constraints. In that case, you need to modify your problem formulation and your cutting plane method accordingly.

One way to handle multiple objectives is to use a weighted sum approach, where you assign a weight to each objective and combine them into a single objective function.

For example, if you have two objectives, f_1 and f_2 , you can write your objective function as $w_1 \cdot f_1 + w_2 \cdot f_2$, where w_1 and w_2 are the weights that reflect your preferences or trade-offs. Then, you can apply the cutting plane method to this single objective function as usual.

Another way to handle multiple objectives is to use a lexicographic approach, where you rank your objectives by their importance and solve them sequentially.

For example, if you have two objectives, f_1 and f_2 , and f_1 is more important than f_2 , you can first solve the problem with f_1 as the objective function and obtain an optimal solution x^* . Then, you can add a constraint that $f_1(x) = f_1(x^*)$, and solve the problem with f_2 as the objective function and obtain another optimal solution y^* . Then, you can compare x^* and y^* and choose the one that satisfies your criteria. You can apply the cutting plane method to each subproblem as usual.

One way to handle multiple constraints is to use a slack variable approach, where you introduce a variable that measures the amount of violation or deviation from a constraint.

For example, if you have a constraint of the form $ax \leq b$, you can write it as $ax + s = b$, where s is the slack variable that represents the difference between b and ax . Then, you can add a penalty term to your objective function that depends on s , such as $p*s$ or $p*s^2$, where p is a positive constant that reflects the cost of violating the constraint. Then, you can apply the cutting plane method to this modified problem as usual.

Another way to handle multiple constraints is to use a feasibility cut approach, where you add a cut that ensures that all the constraints are satisfied.

For example, if you have a set of constraints of the form $Ax \leq b$, you can add a cut of the form $x'(Ax - b) \geq 0$, where x' is a vector of positive constants that reflects the importance or priority of each constraint. Then, you can apply the cutting plane method to this augmented problem as usual.

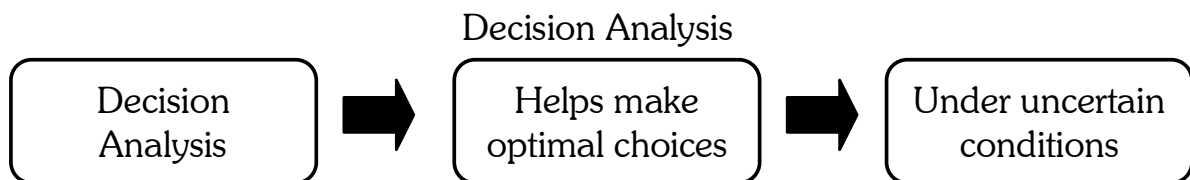
4.5 DECISION ANALYSIS

Q10. What is decision there analysis ?

Ans :

(Imp.)

Decision analysis is a technique that aims to give a rational foundation for management choices made under ambiguity. Decision analysis is a methodical, statistical, and expansive approach to making complex judgments. It is a normative technique for choosing between acts with unknown results.



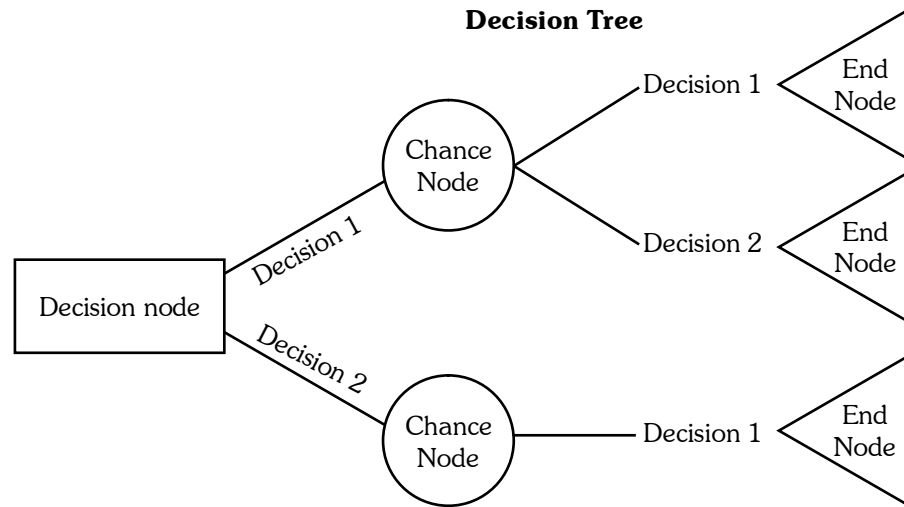
In brief, decision analysis is a normative technique for choosing between acts with unknown results. Such outcome uncertainty can be represented by a **probability distribution** for variables representing the most significant outcomes of the acts under consideration. Hence, a utility function characterizes the relative preference of the decision maker for the many possible outcomes. It incorporates the decision maker's **risk aversion**. Then, a logical decision maker should choose the action that maximizes a certain mathematical combination of the determined probabilities and utilities.

Decision Tree

Decision analysis is a technique that aims to give a logical framework for making one management choice out of several options under ambiguity. Moreover, this approach is based on a structure known as a decision tree.

Mainly, a decision tree has two sorts of nodes: choice nodes and chance nodes. A choice node is a node where a decision must be taken, whereas a chance node is a node where an unexpected consequence

is achieved. Hence, movement from node to node in the tree indicates the passage of time. Therefore travel from node to node represents either the need to make a decision or the realization of a random conclusion.



A decision-analytic model, often a decision tree, is the core instrument of decision analysis. It gives a graphical representation of the sequences of events that might occur following various options (or acts). It also considers the consequences associated with each pathway.

Choice models can integrate the probability of the underlying (actual) states of nature when calculating the distribution of potential outcomes for a specific decision. The decisionmaker is unaware of these possibilities, yet they are very essential.

The term “model” has different connotations in various contexts. Similarly, the defining characteristic of a decision model is that its purpose is to facilitate decision-making and not to establish truth claims.

Evidently, the objective of statistical research designs is to collect data. However, the objective of decision analytic studies is to analyze evidence, albeit, in the absence of a systematic and formal strategy that supports the decision maker in processing the (sometimes divergent and complicated) data, the processing is conducted less formally.

Example

A decision analyst is asked to consider and evaluate the option of installing a new machine in the production department; hence to come to a decision, the analyst decides to use the **decision analysis tree** technique.

Moreover, the decision analyst knows that the value of installing a new machine depends on the chance that the capacity will require expansion in the future. Accordingly, it influences the probabilities of true and false positive results and true and false negative findings. The relative values attributed to these various outcomes establish the usefulness of a machine. For instance, how detrimental is it to fail to install new machinery without increasing capacity relative to not installing a machine and having the need to expand in the future?

Also, the analyst considers if the knowledge acquired from the decision tree alters the earlier choice that would be taken. Likewise, does this alteration results in better future outcomes is also considered. Particularly in circumstances where various data inputs from a range of research are relevant to a specific decision-making context, decision analysis approaches have been proven to be very beneficial.

4.6 RISK AND UNCERTAINTY METHODS

Q11. What is Risk Analysis?

Ans : (Imp.)

The term risk analysis refers to the assessment process that identifies the potential for any adverse events that may negatively affect organizations and the environment. Risk analysis is commonly performed by corporations (banks, construction groups, health care, etc.), governments, and nonprofits. Conducting a risk analysis can help organizations determine whether they should undertake a project or approve a financial application, and what actions they may need to take to protect their interests. This type of analysis facilitates a balance between risks and risk reduction. Risk analysts often work in with forecasting professionals to minimize future negative unforeseen effects.

Q12. Discuss various types of risk analysis methods.

Ans :

Types of Risk Analysis

i) Risk-Benefits

Many people are aware of a cost-benefit analysis. In this type of analysis, an analyst compares the benefits a company receives to the financial and non-financial expenses related to the benefits. The potential benefits may cause other, new types of potential expenses to occur. In a similar manner, a risk-benefit analysis compares potential benefits with associated potential risks. Benefits may be ranked and evaluated based on their likelihood of success or the projected impact the benefits may have.

ii) Needs Assessment

A needs risk analysis is an analysis of the current state of a company. Often, a company will undergo a needs assessment to better understand a need or gap that is already known. Alternatively, a needs assessment may be done if management is not aware of gaps or deficiencies. This analysis lets the company know where they need to spending more resources in.

iii) Business Impact Analysis

In many cases, a business may see a potential risk looming and wants to know how the situation may impact the business. For example, consider the probability of a concrete worker strike to a real estate developer. The real estate developer may perform a business impact analysis to understand how each additional day of the delay may impact their operations.

iv) Root Cause Analysis

Opposite of a needs analysis, a root cause analysis is performed because something is happening that shouldn't be. This type of risk analysis strives to identify and eliminate processes that cause issues. Whereas other types of risk analysis often forecast what needs to be done or what could be getting done, a root cause analysis aims to identify the impact of things that have already happened or continue to happen.

Q13. What are the steps involved in risk analysis.

Ans : (Imp.)

Step #1: Identify Risks

The first step in many types of risk analysis is to make a list of potential risks you may encounter. These may be internal threats that arise from within a company, though most risks will be external that occur from outside forces. It is important to incorporate many different members of a company for this brainstorming session as different departments may have different perspectives and inputs.

A company may have already addressed the major risks of the company through a SWOT analysis. Although a SWOT analysis may prove to be a launching point for further discussion, risk analysis often addresses a specific question while SWOT analysis are often broader. Some risks may be listed on both, but a risk analysis should be more specific when trying to address a specific problem.

Step #2: Identify Uncertainty

The primary concern of risk analysis is to identify troublesome areas for a company. Most often, the riskiest aspects may be the areas that are undefined. Therefore, a critical aspect of risk analysis is to understand how each potential risk has uncertainty and to quantify the range of risk that uncertainty may hold.

Step #3: Estimate Impact

Most often, the goal of a risk analysis is to better understand how risk will financially impact a company. This is usually calculated as the risk value, which is the probability of an event happening multiplied by the cost of the event.

Step #4: Build Analysis Model(s)

The inputs from above are often fed into an analysis model. The analysis model will take all available pieces of data and information, and the model will attempt to yield different outcomes, probabilities, and financial projections of what may occur. In more advanced situations, scenario analysis or simulations can determine an average outcome value that can be used to quantify the average instance of an event occurring.

Step #5: Analyze Results

With the model run and the data available to be reviewed, it's time to analyze the results. Management often takes the information and determines the best course of action by comparing the likelihood of risk, projected financial impact, and model simulations. Management may also request to see different scenarios run for different risks based on different variables or inputs.

Step #6: Implement Solutions

After management has digested the information, it is time to put a plan in action. Sometimes, the plan is to do nothing; in risk acceptance strategies, a company has decided it will not change course as it makes most financial sense to simply live with the risk of something happening and dealing with it after it occurs. In other cases, management may want to reduce or eliminate the risk.

Q14. What are the advantages and dis-advantages of risk analysis?

Ans :

Advantages

- May aid in minimizing losses due to management preemptively forming a risk plan
- May allow management to quantify risks and assign dollars to future events
- May protect company resources, produce better processes, and mitigate overall risk

Disadvantages

- Relies heavily on estimates, so it may be difficult to perform for certain risks
- Can not predict unpredictable, black swan events
- May underestimate risk magnitude or occurrence, leading to overconfident operations

Q15. What are the differences between Risk and Uncertainty

Ans :

(Imp.)

Sl.No.	Basis for Comparison	Risk	Uncertainty
1.	Meaning	The probability of winning or losing something worthy is known as risk.	Uncertainty implies a situation the future events are not known.
2.	Ascertainment	It can be measured	It cannot be measured.
3.	Outcome	Chances of outcomes are known.	The outcome is unknown.
4.	Control	Controllable	Uncontrollable
5.	Minimization	Yes	No
6.	Probabilities	Assigned	Not assigned

4.7 TEXT ANALYTICS WEB ANALYTICS

Q16. What is text analytics? What are the benefits of text analytics.

Ans : (Imp.)

Text analytics combines a set of machine learning, statistical and linguistic techniques to process large volumes of unstructured text or text that does not have a predefined format, to derive insights and patterns. It enables businesses, governments, researchers, and media to exploit the enormous content at their disposal for making crucial decisions. Text analytics uses a variety of techniques – sentiment analysis, topic modelling, named entity recognition, term frequency, and event extraction.

Benefits of Text Analytics

There are a range of ways that text analytics can help businesses, organizations, and event social movements:

- Help businesses to understand customer trends, product performance, and service quality. This results in quick decision making, enhancing business intelligence, increased productivity, and cost savings.
- Helps researchers to explore a great deal of pre-existing literature in a short time, extracting what is relevant to their study. This helps in quicker scientific breakthroughs.
- Assists in understanding general trends and opinions in the society, that enable governments and political bodies in decision making.
- Text analytic techniques help search engines and information retrieval systems to improve their performance, thereby providing fast user experiences.
- Refine user content recommendation systems by categorizing related content.

Q17. What are the steps involved in text analytics.

Ans :

Text analytics is a sophisticated technique that involves several pre-steps to gather and cleanse the

unstructured text. There are different ways in which text analytics can be performed. This is an example of a model workflow.

1. Data gathering

Text data is often scattered around the internal databases of an organization, including in customer chats, emails, product reviews, service tickets and Net Promoter Score surveys. Users also generate external data in the form of blog posts, news, reviews, social media posts and web forum discussions. While the internal data is readily available for analytics, the external data needs to be gathered.

2. Preparation of data

Once the unstructured text data is available, it needs to go through several preparatory steps before machine learning algorithms can analyze it. In most of the text analytics software, this step happens automatically. Text preparation includes several techniques using natural language processing as follows:

➤ Tokenization

In this step, the text analysis algorithms break the continuous string of text data into tokens or smaller units that make up entire words or phrases. For instance, character tokens could be each individual letter in this word: F-I-S-H. Or, you can break up by subword tokens: Fishing. Tokens represent the basis of all natural language processing. This step also discards all the unwanted contents of the text, including white spaces.

➤ Part-of-speech-tagging

In this step, each token in the data is assigned a grammatical category like noun, verb, adjective, and adverb.

➤ Parsing

Parsing is the process of understanding the syntactical structure of the text. Dependency parsing and constituency parsing are two popular techniques used to derive syntactical structure.

➤ Lemmatization and stemming

These are two processes used in data preparation to remove the suffixes and affixes associated with the tokens and retain its dictionary form or lemma.

➤ **Stopword removal**

This is the phase when all the tokens that have frequent occurrence but bear no value in the text analytics. This includes words such as 'and', 'the' and 'a'.

3. Text analytics

After the preparation of unstructured text data, text analytics techniques can now be performed to derive insights. There are several techniques used for text analytics. Prominent among them are text classification and text extraction.

➤ **Text classification**

This technique is also known as text categorization or tagging. In this step, certain tags are assigned to the text based on its meaning. For example, while analyzing customer reviews, tags like "positive" or "negative" are assigned. Text classification often is done using rule-based systems or machine learning-based systems. In rule-based systems, humans define the association between language pattern and a tag. "Good" may indicate positive review; "bad" may identify a negative review.

➤ **Text extraction**

This is the process of extracting recognizable and structured information from the unstructured input text. This information includes keywords, names of people, places and events. One of the simple methods for text extraction is regular expressions. However, this is a complicated method to maintain when the complexity of input data increases. Conditional Random Fields (CRF) is a statistical method used in text extraction. CRF is a sophisticated but effective way of extracting vital information from the unstructured text.

Q18. Discuss the techniques used for text analytics

Ans :

There are several techniques related to analyzing the unstructured text. Each of these techniques is used for different use case scenarios.

1. Sentiment analysis

Sentiment analysis is used to identify the emotions conveyed by the unstructured text. The input text includes product reviews, customer interactions, social media posts, forum discussions, or blogs. There are different types of sentiment analysis. Polarity analysis is used to identify if the text expresses positive or negative sentiment. The categorization technique is used for a more fine-grained analysis of emotions - confused, disappointed, or angry.

Use cases of sentiment analysis

- Measure customer response to a product or a service
- Understand audience trends towards a brand
- Understand new trends in consumer space
- Prioritize customer service issues based on the severity
- Track how customer sentiment evolves over time

2. Topic modelling

This technique is used to find the major themes or topics in a massive volume of text or a set of documents. Topic modeling identifies the keywords used in text to identify the subject of the article.

Use cases of topic modeling

- Large law firms use topic modeling to examine hundreds of documents during large litigations.
- Online media uses topic modeling to pick up trending topics across the web.
- Researchers use topic modeling for exploratory literature review.
- Businesses can determine which of their products are successful.
- Topic modeling helps anthropologists to determine the emergent issues and trends in a society based on the content people share on the web.

3. Named Entity Recognition (NER)

NER is a text analytics technique used for identifying named entities like people, places, organizations, and events in unstructured text. NER extracts nouns from the text and determines the values of these nouns.

Use cases of named entity recognition

- NER is used to classify news content based on people, places, and organizations featured in them.
- Search and recommendation engines use NER for information retrieval.
- For large chain companies, NER is used to sort customer service requests and assign them to a specific city, or outlet.
- Hospitals can use NER to automate the analysis of lab reports.

4. Term frequency – inverse document frequency

TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term's importance to the document. This technique uses an inverse document frequency factor to filter out frequently occurring yet non-insightful words, articles, propositions, and conjunctions.

5. Event extraction

This is a text analytics technique that is an advancement over the named entity extraction. Event extraction recognizes events mentioned in text content, for example, mergers, acquisitions, political moves, or important meetings. Event extraction requires an advanced understanding of the semantics of text content. Advanced algorithms strive to recognize not only events but the venue, participants, date, and time wherever applicable. Event extraction is a beneficial technique that has multiple uses across fields.

Q19. What is web analytics ? What is the importance of web analytics.

Ans :

Web Analytics is the methodological study of online/offline patterns and trends. It is a technique that you can employ to collect, measure,

report, and analyze your website data. It is normally carried out to analyze the performance of a website and optimize its web usage.

We use web analytics to track key metrics and analyze visitors' activity and traffic flow. It is a tactical approach to collect data and generate reports.

Importance of Web Analytics

We need Web Analytics to assess the success rate of a website and its associated business. Using Web Analytics, we can

- Assess web content problems so that they can be rectified
- Have a clear perspective of website trends
- Monitor web traffic and user flow
- Demonstrate goals acquisition
- Figure out potential keywords
- Identify segments for improvement
- Find out referring sources

Q20. Explain web analytics process.

Ans :

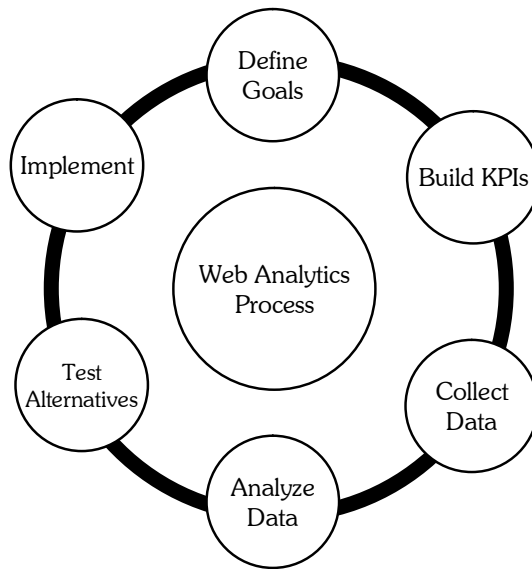
(Imp.)

The primary objective of carrying out Web Analytics is to optimize the website in order to provide better user experience. It provides a data-driven report to measure visitors' flow throughout the website.

Take a look at the following illustration. It depicts the process of web analytics.

- Set the business goals.
- To track the goal achievement, set the Key Performance Indicators (KPI).
- Collect correct and suitable data.
- To extract insights, Analyze data.
- Based on assumptions learned from the data analysis, Test alternatives.
- Based on either data analysis or website testing, Implement insights.

Web Analytics is an ongoing process that helps in attracting more traffic to a site and thereby, increasing the Return on Investment.



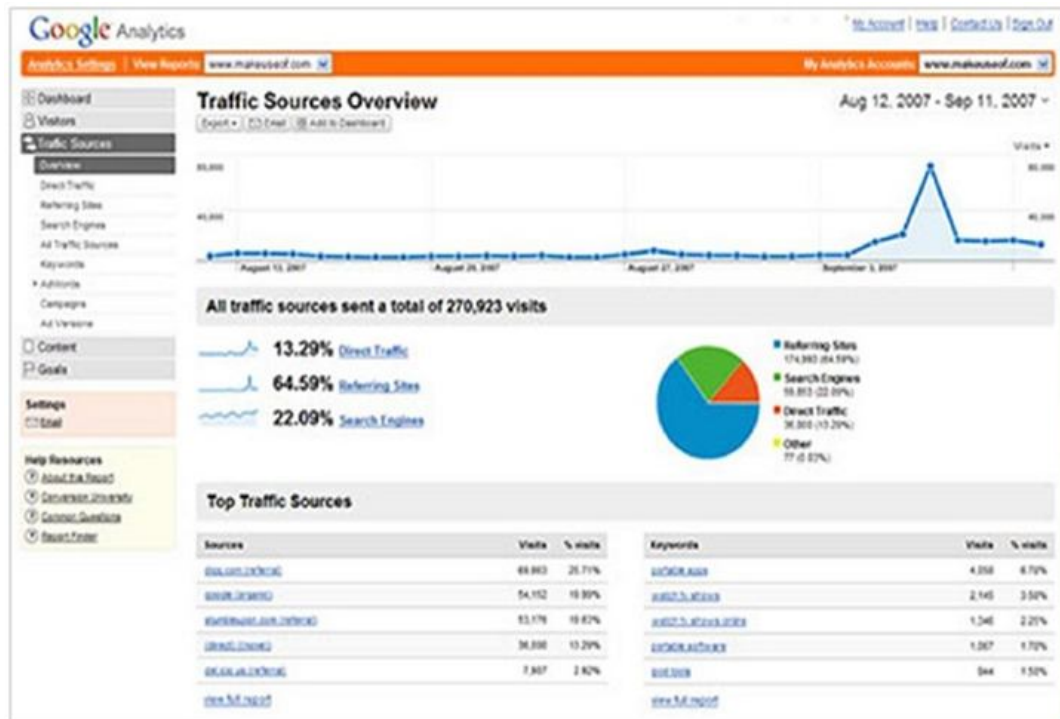
Q21. Discuss various google analytics.

Ans:

(Imp.)

Google Analytics

Google Analytics is a freemium analytic tool that provides a detailed statistics of the web traffic. It is used by more than 60% of website owners.



Google analytics helps you to track and measure visitors, traffic sources, goals, conversion, and other metrics (as shown in the above image). It basically generates reports on:

- Audience Analysis
- Acquisition Analysis
- Behavior Analysis
- Conversion Analysis

Let us discuss each one of them in detail.

1. Audience Analysis

As the name suggests, audience analysis gives you an overview of the audience who visit your site along with their session history, page-views, bounce rate, etc. You can trace the new as well as the returning users along with their geographical locations. You can track:

- The age and gender of your audience under Demographics.
- The affinity reach and market segmentation under Interests.
- Language and location under Geo.
- New and returning visitors, their frequency, and engagement under Behavior.
- Browsers, Operating systems, and network of your audience in Technology.
- Mobile device info under Mobile.
- Custom variable report under Custom. This report shows the activity by custom modules that you created to capture the selections.
- Benchmarking channels, locations, and devices under Benchmarking. Benchmarking allows you to compare your metrics with other related industries. So, you can plot what you need to incur in order to overtake the market.
- Flow of user activity under Users flow to see the path they took on your website.

2. Acquisition Analysis

Acquisition means 'to acquire.' Acquisition analysis is carried out to find out the sources from where your web traffic originates. Using acquisition analysis, you can:

- Capture traffic from all channels, particular source/medium, and from referrals.
- Trace traffic from AdWords (paid search).
- See traffic from search engines. Here, you can see Queries, triggered landing pages, and geographical summary.
- Track social media traffic. It helps you to identify networks where your users are engaged. You can see referrals from where your traffic originates. You can also have a view of your hub activity, bookmarking sites follow-up, etc. In the same tab, you can have a look at your endorsements in details. It helps you measure the impact of social media on your website.
- See which plug-ins gave you traffic.
- Have a look at all the campaigns you built throughout your website with detailed statistics of paid/organic keywords and the cost incurred on it.

3. Behavior Analysis

Behavior analysis monitors users' activities on a website. You can find behavioral data under the following four segments:

- **Site Content:** It shows how many pages were viewed. You can see the detailed interaction of data across all pages or in segments like content drill-down, landing pages, and exit pages. Content drill-down is breaking up of data into sub-folders. Landing page is the page where the user lands, and exit page is where the user exits your site. You can measure the behavioral flow in terms of content.
- **Site Speed:** Here, you can capture page load time, execution speed, and performance data. You can see how quickly the browser can parse through the page. Further, you can measure page timings, user timings, and get speed suggestion. It helps you to know where you are lagging.
- **Site Search:** It gives you a full picture of how the users search across your site, what they normally look for, and how they arrive at a particular landing page. You can analyze what they search for before landing on your website.
- **Events:** Events are visitors' actions with content, which can be traced independently. Example "downloads, sign up, log-in, etc.

4. Conversion Analysis

Conversion is a goal completion or a transaction by a user on your website. For example, download, checkout, buy, etc. To track conversions in analytics, you need to define a goal and set a URL that is traceable.

- **Goals:** Metrics that measure a profitable activity that you want the user to complete. You can set them to track the actions. Each time a goal is achieved, a conversion is added to your data. You can observe goal completion, value, reverse path, and goal flow.
- **Ecommerce:** You can set ecommerce tracking to know what the users buy from your website. It helps you to find product performance, sale performance, transactions, and purchase time. Based on these data, you can analyze what can be beneficial and what can incur you loss.
- **Multi-channel funnels:** Multi-channel funnels or MCF reports the source of conversion; what roles the website plays, referrals' role in that conversion; and what all slabs did when users pass through landing page to conversion. For example, a user searched for a query on Google search page, he visited the website, but did not convert. Later on, he directly typed your website name and made a purchase. All these activities can be traced on MCF.
- **Attribution:** Attribution modeling credits sales and conversions to touch points in conversion tracking. It lets you decide what platforms or strategy or module is the best for your business. Suppose a person visited your website through AdWords ad and made no purchase. A month later, he visits via a social platform and again does not buy. Third time, he visited directly and converted. Here, the last interaction model will credit direct for the conversion, whereas first interaction model will assign credit to paid medium. This way, you can analyze what module should be credited for a conversion.

Short Question and Answers

1. What is prescriptive analytics?

Ans :

Prescriptive analytics is a statistical method that focuses on finding the ideal way forward or action necessary for a particular scenario, based on data. Prescriptive analytics uses both descriptive and predictive analytics but the focus here remains on actionable insights rather than data monitoring. The input of prescriptive analytics is the outcome of predictive analytics algorithms. You not only predict what the future holds, but you leverage that prediction to take the best course of action for the future. A more formal definition is that prescriptive analytics is a statistical approach utilized to generate recommendations and aid decision-making based on the computational outcomes of algorithmic models. Following are the benefits of prescriptive analytics.

2. What is Linear optimization?

Ans :

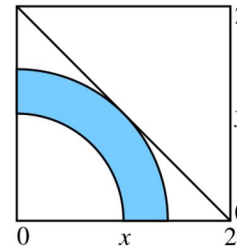
Linear optimization is a method applicable for the solution of problems in which the objective function and the constraints appear as linear functions of the decision variables. The constraint equations may be in the form of equalities or inequalities. In other words, linear optimization determines the way to achieve the best outcome (for example, to maximize profit or to minimize cost) in a given mathematical model and given some lists of requirements represented as linear equations.

3. What is mean by non linear optimization?

Ans :

Nonlinear programming is the process of solving optimization problems that concern some of the nonlinear constraints or nonlinear objective functions. It involves minimizing or maximizing a nonlinear objective function subject to bound constraints, linear constraints, nonlinear constraints, etc. These constraints can be inequalities or equalities. In addition, nonlinear programming helps

in analyzing design tradeoffs, selecting optimal designs, computing optimal trajectories and portfolio optimization and model calibration in computation finance.



4. Discuss in detail about non linear optimization

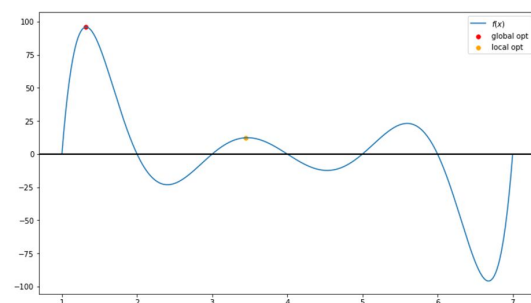
Ans :

In many optimization models the objective and/or the constraints are nonlinear functions of the decision variables. Such an optimization model is called a nonlinear programming (NLP) model.

- When you solve an LP model, you are mostly guaranteed that the solution obtained is an optimal solution and a sensitivity analysis with shadow price and reduced cost is available.
- When you solve an NLP model, it is very possible that you obtain a suboptimal solution.

This is because a nonlinear function can have local optimal solutions where

- A local optimum is better than all nearby points that are not the global optimal solution
- A global optimum is the best point in the entire feasible region

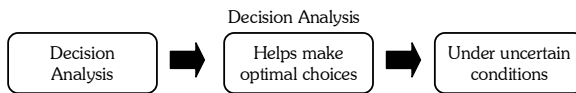


However, convex programming, as one of the most important type of nonlinear programming models, can be solved to its optimality.

5. What is decision there analysis ?

Ans :

Decision analysis is a technique that aims to give a rational foundation for management choices made under ambiguity. Decision analysis is a methodical, statistical, and expansive approach to making complex judgments. It is a normative technique for choosing between acts with unknown results.



In brief, decision analysis is a normative technique for choosing between acts with unknown results. Such outcome uncertainty can be represented by a **probability distribution** for variables representing the most significant outcomes of the acts under consideration. Hence, a utility function characterizes the relative preference of the decision maker for the many possible outcomes. It incorporates the decision maker's **risk aversion**. Then, a logical decision maker should choose the action that maximizes a certain mathematical combination of the determined probabilities and utilities.

6. What is Risk Analysis?

Ans :

The term risk analysis refers to the assessment process that identifies the potential for any adverse events that may negatively affect organizations and the environment. Risk analysis is commonly performed by corporations (banks, construction groups, health care, etc.), governments, and nonprofits. Conducting a risk analysis can help organizations determine whether they should undertake a project or approve a financial application, and what actions they may need to take to protect their interests. This type of analysis facilitates a balance between risks and risk reduction. Risk analysts often work in with forecasting professionals to minimize future negative unforeseen effects.

7. What are the advantages and disadvantages of risk analysis?

Ans :

Advantages

- May aid in minimizing losses due to management preemptively forming a risk plan
- May allow management to quantify risks and assign dollars to future events
- May protect company resources, produce better processes, and mitigate overall risk

Disadvantages

- Relies heavily on estimates, so it may be difficult to perform for certain risks
- Can not predict unpredictable, black swan events
- May underestimate risk magnitude or occurrence, leading to overconfident operations

8. What is web analytics ? What is the importance of web analytics.

Ans :

Web Analytics is the methodological study of online/offline patterns and trends. It is a technique that you can employ to collect, measure, report, and analyze your website data. It is normally carried out to analyze the performance of a website and optimize its web usage.

We use web analytics to track key metrics and analyze visitors' activity and traffic flow. It is a tactical approach to collect data and generate reports.

Importance of Web Analytics

We need Web Analytics to assess the success rate of a website and its associated business. Using Web Analytics, we can

- Assess web content problems so that they can be rectified
- Have a clear perspective of website trends
- Monitor web traffic and user flow
- Demonstrate goals acquisition
- Figure out potential keywords
- Identify segments for improvement
- Find out referring sources

9. What are the differences between Risk and Uncertainty

Ans :

Sl.No.	Basis for Comparison	Risk	Uncertainty
1.	Meaning	The probability of winning or losing something worthy is known as risk.	Uncertainty implies a situation the future events are not known.
2.	Ascertainment	It can be measured	It cannot be measured.
3.	Outcome	Chances of outcomes are known.	The outcome is unknown.
4.	Control	Controllable	Uncontrollable
5.	Minimization	Yes	No
6.	Probabilities	Assigned	Not assigned

10. Explain web analytics process.

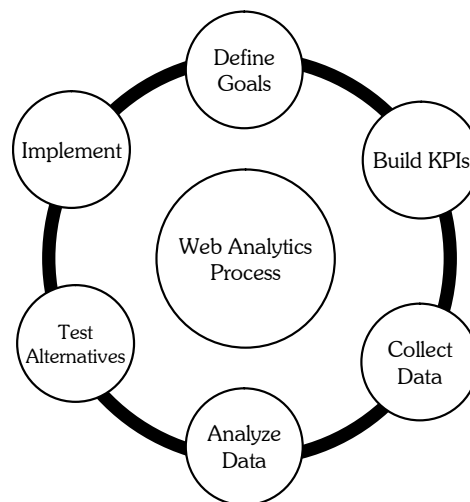
Ans :

The primary objective of carrying out Web Analytics is to optimize the website in order to provide better user experience. It provides a data-driven report to measure visitors' flow throughout the website.

Take a look at the following illustration. It depicts the process of web analytics.

- Set the business goals.
- To track the goal achievement, set the Key Performance Indicators (KPI).
- Collect correct and suitable data.
- To extract insights, Analyze data.
- Based on assumptions learned from the data analysis, Test alternatives.
- Based on either data analysis or website testing, Implement insights.

Web Analytics is an ongoing process that helps in attracting more traffic to a site and thereby, increasing the Return on Investment.



Choose the Correct Answers

1. Prescriptive analytics makes the use of machine learning to help _____ to decide a course of action based on a computer program's predictions. [a]
(a) Business organizations (b) System development
(c) Employees development (d) All of the mentioned above
2. The ability to model prices on a variety of factors allows them to make _____ about production, storage, and new discoveries. [c]
(a) No decisions (b) Better decisions
(c) Unpredictable things (d) None of the mentioned above
3. Amongst which of the following is / are the goodness of prescriptive analytics, [d]
(a) Exhausting valuable resources on housing data that does not inform business decisions
(b) Spending time sifting through unutilized data sets
(c) Missing out on unique revenue streams and insights
(d) All of the mentioned above
4. Prescriptive analytics is a process that _____ and provides instant recommendations on how to optimize business practices. [a]
(a) Analyzes data (b) Data storage
(c) Data ingestion (d) None of the mentioned above
5. Prescriptive analytics utilizes business rules, artificial intelligence, and _____ to simulates various approaches to these numerous outcomes. [a]
(a) Algorithms (b) Flowchart
(c) System flow (d) None of the mentioned above
6. Effective _____ prescriptive data tools can help businesses using informed data to create the processes and for managing and analyzing data anytime and anywhere. [a]
(a) Cloud-based (b) Data warehouse
(c) System ingestion (d) All of the mentioned above
7. Amongst which of the following is /are the techniques that are used for predictive analytics, [d]
(a) Linear Regression (b) Time series analysis and forecasting
(c) Data Mining (d) All of the mentioned above
8. Which of the following words best describes Prescriptive analytics? [a]
(a) Action (b) Prediction
(c) Diagnostics (d) Exploration

9. What is the main difference between Prescriptive and Predictive analytics? [b]
- (a) None
 - (b) Assisting decision making with suggested actions vs predicting the future
 - (c) Using sophisticated tools for analytics in prescriptive analytics
 - (d) Linear regressions vs optimization
10. Prescriptive analytics utilizes business rules, artificial intelligence, and _____ to simulates various approaches to these numerous outcomes. [a]
- (a) Algorithms
 - (b) Flow chart
 - (c) System flow
 - (d) None of the above

Fill in the Blanks

1. _____ may be defined as the problem of maximizing or minimizing a linear function that is subjected to linear constraints.
2. NLP stands for _____.
3. In mathematics, _____ is the process of solving an optimization problem where some of the constraints or the objective function are nonlinear.
4. An _____ problem is a mathematical optimization or feasibility program in which some or all of the variables are restricted to be integers.
5. _____ works by relaxing the integer variables and solving the resulting linear programming problem, which is called the continuous relaxation.
6. _____ is a systematic, quantitative, and visual approach to making strategic business decisions.
7. _____ is the process of using computer systems to read and understand human-written text for business insights.
8. _____ refers to the process of collecting website data and then processing, reporting, and analyzing it to create an online strategy for improving the website experience.
9. _____ converts raw data into actionable insights.
10. _____ is the use of advanced processes and tools to analyze data and content to recommend the optimal course of action or strategy moving forward.

ANSWERS

1. Linear optimization
2. Non Linear Programming
3. Nonlinear programming (NLP)
4. Integer programming
5. Cutting plane method
6. Decision Analysis
7. Text analysis
8. Web analytics
9. Data analytics
10. Prescriptive analytics

One Mark Answers

1. Prescriptive Analytics

Ans :

Prescriptive analytics is the use of advanced processes and tools to analyze data and content to recommend the optimal course of action or strategy moving forward.

2. Decision Analysis

Ans :

Decision analysis is a systematic, quantitative, and visual approach to making strategic business decisions. Decision analysis uses a variety of tools and also incorporates aspects of psychology, management techniques, and economics.

3. Text Analytics

Ans :

Text Analytics is the process of using computer systems to read and understand human-written text for business insights.

4. NLP

Ans :

Nonlinear programming (NP) involves minimizing or maximizing a nonlinear objective function subject to bound constraints, linear constraints, or nonlinear constraints, where the constraints can be inequalities or equalities.

5. Cutting plane Method

Ans :

Cutting plane method works by relaxing the integer variables and solving the resulting linear programming problem, which is called the continuous relaxation.

UNIT V

PROGRAMMING USING R

R Environment, R packages, Reading and Writing data in R, R functions, Control Statements, Frames and Subsets, Managing and Manipulating data in R.

5.1 R ENVIRONMENT

Q1. What is R Programming ? What re the features of R ?

Ans : (Imp.)

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R was initially written by **Ross Ihaka** and **Robert Gentleman** at the Department of Statistics of the University of Auckland in Auckland, New Zealand. R made its first appearance in 1993.

- A large group of individuals has contributed to R by sending code and bug reports.
- Since mid-1997 there has been a core group (the “R Core Team”) who can modify the R source code archive.

Features of R

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R.

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

Q2. Explain the procedure of R Local environment setup.

Ans :

If you are still willing to set up your environment for R, you can follow the steps given below.

1. Windows Installation

You can download the Windows installer version of R from R-3.2.2 for Windows (32/64 bit) and save it in a local directory.

As it is a Windows installer (.exe) with a name “R-version-win.exe”. You can just double click and run the installer accepting the default settings. If your

Windows is 32-bit version, it installs the 32-bit version. But if your windows is 64-bit, then it installs both the 32-bit and 64-bit versions.

After installation you can locate the icon to run the Program in a directory structure “R\R3.2.2\bin\i386\Rgui.exe” under the Windows Program Files. Clicking this icon brings up the R-GUI which is the R console to do R Programming.

2. Linux Installation

R is available as a binary for many versions of Linux at the location R.Binaries.

The instruction to install Linux varies from flavor to flavor. These steps are mentioned under each type of Linux version in the mentioned link. However, if you are in a hurry, then you can use **yum** command to install R as follows:

```
$ yum install R
```

Above command will install core functionality of R programming along with standard packages, still you need additional package, then you can launch R prompt as follows:

```
$ R
```

R version 3.2.0 (2015-04-16) – “Full of Ingredients”

Copyright (C) 2015 The R Foundation for Statistical Computing

Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type ‘license()’ or ‘licence()’ for distribution details.

R is a collaborative project with many contributors.

Type ‘contributors()’ for more information and ‘citation()’ on how to cite R or R packages in publications.

Type ‘demo()’ for some demos, ‘help()’ for on-line help, or ‘help.start()’ for an HTML browser interface to help.

Type ‘q()’ to quit R.

```
>
```

Now you can use install command at R prompt to install the required package. For example, the following command will install **plotrix** package which is required for 3D charts.

```
> install.packages("plotrix")
```

5.2 R PACKAGES

Q3. Explain in detail about R Packages.

Ans :

(Imp.)

R packages are a collection of R functions, compiled code and sample data. They are stored under a directory called **"library"** in the R environment. By default, R installs a set of packages during installation. More packages are added later, when they are needed for some specific purpose. When we start the R console, only the default packages are available by default. Other packages which are already installed have to be loaded explicitly to be used by the R program that is going to use them. Below is a list of commands to be used to check, verify and use the R packages.

Check Available R Packages

Get library locations containing R packages

```
.libPaths()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

```
[2] "C:/Program Files/R/R-3.2.2/library"
```

Get the list of all the packages installed

```
library()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

Packages in library 'C:/Program Files/R/R-3.2.2/library':

base	The R Base Package
boot	Bootstrap Functions (Originally by Angelo Canty for S)
class	Functions for Classification
cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
codetools	Code Analysis Tools for R
compiler	The R Compiler Package
datasets	The R Datasets Package
foreign	Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...
graphics	The R Graphics Package
grDevices	The R Graphics Devices and Support for Colours and Fonts
grid	The Grid Graphics Package
KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)
lattice	Trellis Graphics for R
MASS	Support Functions and Datasets for Venables and Ripley's MASS
Matrix	a Sparse and Dense Matrix Classes and Methods
methods	Formal Methods and Classes
mgcv	Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation
nlme	Linear and Nonlinear Mixed Effects Models
nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models
parallel	Support for Parallel computation in R
rpart	Recursive Partitioning and Regression Trees
spatial	Functions for Kriging and Point Pattern Analysis
splines	Regression Spline Functions and Classes
stats	The R Stats Package
stats4	Statistical Functions using S4 Classes
survival	Survival Analysis
tcltkTc	I/Tk Interface
tools	Tools for Package Development
utils	The R Utils Package

Get all packages currently loaded in the R environment

```
search()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

```
[1] ".GlobalEnv"      "package:stats"   "package:graphics"
[4] "package:grDevices" "package:utils"   "package:datasets"
[7] "package:methods" "Autoloads"       "package:base"
```

Q4. How to install new packages?

Ans :

Install a New Package

There are two ways to add new R packages. One is installing directly from the CRAN directory and another is downloading the package to your local system and installing it manually.

Install directly from CRAN

The following command gets the packages directly from CRAN webpage and installs the package in the R environment. You may be prompted to choose a nearest mirror. Choose the one appropriate to your location.

```
install.packages("Package Name")
# Install the package named "XML".
install.packages("XML")
```

Install package manually

Go to the link [R Packages](#) to download the package needed. Save the package as a **.zip** file in a suitable location in the local system.

Now you can run the following command to install this package in the R environment.

```
install.packages(file_name_with_path, repos = NULL, type = "source")
```

```
# Install the package named "XML"
install.packages("E:/XML_3.98-1.3.zip", repos = NULL, type = "source")
```

Load Package to Library

Before a package can be used in the code, it must be loaded to the current R environment. You also need to load a package that is already installed previously but not available in the current environment.

A package is loaded using the following command “

```
library("package Name", lib.loc = "path to library")
# Load the package named "XML"
install.packages("E:/XML_3.98-1.3.zip", repos = NULL, type = "source")
```

5.3 READING AND WRITING DATA IN R

Q5. What are the functions are useful for reading data into R and writing data to files.

Ans : (Imp.)

Functions for Reading Data into R

There are a few very useful functions for reading data into R.

1. **read.table()** and **read.csv()** are two popular functions used for reading tabular data into R.
2. **readLines()** is used for reading lines from a text file.
3. **source()** is a very useful function for reading in R code files from a another R program.
4. **dget()** function is also used for reading in R code files.
5. **load()** function is used for reading in saved workspaces
6. **unserialize()** function is used for reading single R objects in binary format.

Functions for Writing Data to Files

There are similar functions for writing data to files

1. **write.table()** is used for writing tabular data to text files (i.e. CSV).
2. **writeLines()** function is useful for writing character data line-by-line to a file or connection.
3. **dump()** is a function for dumping a textual representation of multiple R objects.
4. **dput()** function is used for outputting a textual representation of an R object.
5. **save()** is useful for saving an arbitrary number of R objects in binary format to a file.
6. **serialize()** is used for converting an R object into a binary format for outputting to a connection (or file).

Q6. What is the use of read.table () ?

Ans :

Reading Data Files with read.table()

The read.table() function is one of the most commonly used functions for reading data in R. TO get the help file for read.table() just type **read.table** in R console.

The read.table() function has a few important arguments:

- file, the name of a file, or a connection
- header, logical indicating if the file has a header line
- sep, a string indicating how the columns are separated
- colClasses, a character vector indicating the class of each column in the dataset
- nrows, the number of rows in the dataset. By default read.table() reads an entire file.
- comment.char, a character string indicating the comment character. This defaults to "#". If there are no commented lines in your file, it's worth setting this to be the empty string.
- skip, the number of lines to skip from the beginning
- stringsAsFactors, should character variables be coded as factors? This defaults to TRUE because back in the old days, if you had data that were stored as strings, it was because those strings represented levels of a categorical variable. Now we have lots of data that is text data and they don't always represent categorical variables. So you may want to set this to be FALSE in those cases. If you always want this to be FALSE, you can set a global option via options(stringsAsFactors = FALSE). I have never seen so much heat generated on discussion forums about an R function argument than the stringsAsFactors argument.

Q7. Discuss the basic use of readLines(), WriteLine() functions in R.

Ans :

readLines() and writeLines() function in R:

readLines() function is mainly used for reading lines from a text file and writeLines() function is useful for writing character data line-by-line to a file or connection. Check the following example to deal with readLines() and writeLines().

You can create a more descriptive representation of an R object by using the **dput()** or **dump()** functions. Unlike writing out a table or CSV file, dump() and dput() preserve the metadata, so that another user doesn't have to specify it all over again.

5.4 R FUNCTIONS

Q8. What is mean by function ? what are the components of function?

Ans : (Imp.)

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.

In R, a function is an object so the R interpreter is able to pass control to the function, along with arguments that may be necessary for the function to accomplish the actions.

The function in turn performs its task and returns control to the interpreter as well as any result which may be stored in other objects.

Function Definition

An R function is created by using the keyword **function**. The basic syntax of an R function definition is as follows “

```
function_name <- function(arg_1, arg_2, ...) {
    Function body
}
```

Function Components

The different parts of a function are:

- **Function Name:** This is the actual name of the function. It is stored in R environment as an object with this name.
- **Arguments:** An argument is a placeholder. When a function is invoked, you pass a value to the argument. Arguments are optional; that is, a function may contain no arguments. Also arguments can have default values.
- **Function Body:** The function body contains a collection of statements that defines what the function does.
- **Return Value:** The return value of a function is the last expression in the function body to be evaluated.

Q9. Discuss various types of functions supported by R.

Ans :

R has many **in-built** functions which can be directly called in the program without defining them first. We can also create and use our own functions referred as **user defined** functions.

Built-in Function

Simple examples of in-built functions are **seq()**, **mean()**, **max()**, **sum(x)** and **paste(...)** etc. They are directly called by user written programs. You can refer most widely used R functions.

Create a sequence of numbers from 32 to 44.

```
print(seq(32,44))
```

Find mean of numbers from 25 to 82.

```
print(mean(25:82))
```

Find sum of numbers from 41 to 68.

```
print(sum(41:68))
```

When we execute the above code, it produces the following result:

```
[1] 32 33 34 35 36 37 38 39 40 41 42 43 44
```

```
[1] 53.5
```

```
[1] 1526
```

User-defined Function

We can create user-defined functions in R. They are specific to what a user wants and once created they can be used like the built-in functions. Below is an example of how a function is created and used.

Create a function to print squares of numbers in sequence.

```
new.function<-function(a){
  for(iin1:a){
    b <- i ^ 2
  }
  print(b)
}
```

Calling a Function

Create a function to print squares of numbers in sequence.

```
new.function<-function(a){
  for(iin1:a){
    b <- i ^ 2
  }
  print(b)
}
```

Call the function new.function supplying 6 as an argument.

```
new.function(6)
```

When we execute the above code, it produces the following result:

```
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
```

Calling a Function without an Argument

Create a function without an argument.

```
new.function<-function(){
```

```
  for(iin1:5){
    print(i ^ 2)
  }
}
```

Call the function without supplying an argument.

```
new.function()
```

When we execute the above code, it produces the following result:

```
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
```

Q10. How can we pass arguments to the functions in R.

Ans :

Calling a Function with Argument Values (by position and by name)

The arguments to a function call can be supplied in the same sequence as defined in the function or they can be supplied in a different sequence but assigned to the names of the arguments

Create a function with arguments.

```
new.function<-function(a,b,c){
  result <- a * b + c
  print(result)
}
```

Call the function by position of arguments.

```
new.function(5,3,11)
```

Call the function by names of the arguments.

```
new.function(a = 11, b = 5, c = 3)
```

When we execute the above code, it produces the following result:

```
[1] 26
[1] 58
```

Calling a Function with Default Argument

We can define the value of the arguments in the function definition and call the function without supplying any argument to get the default result. But we can also call such functions by supplying new values of the argument and get non default result.

```
# Create a function with arguments.
new.function<-function(a =3, b =6){
  result <- a * b
  print(result)
}

# Call the function without giving any argument.
new.function()

# Call the function with giving new values of the argument.
new.function(9, 5)
```

When we execute the above code, it produces the following result:

```
[1] 18
[1] 45
```

5.5 CONTROL STATEMENTS

Q11. Discuss in detail about Control statements in R.

(OR)

Discuss in detail about decision making statements supported by R.

(OR)

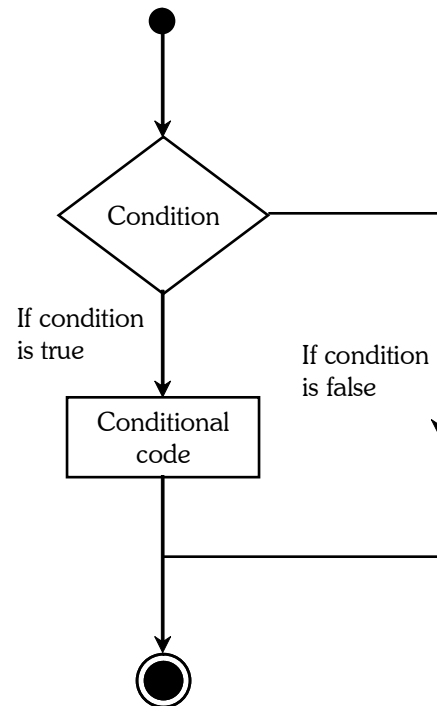
What are the conditional statements supported by R.

Ans :

(Imp.)

Decision making structures require the programmer to specify one or more conditions to be evaluated or tested by the program, along with a statement or statements to be executed if the condition is determined to be **true**, and optionally, other statements to be executed if the condition is determined to be **false**.

Following is the general form of a typical decision making structure found in most of the programming languages



R provides the following types of decision making statements. Click the following links to check their detail.

Sr. No.	Statement & Description
1.	if statement An if statement consists of a Boolean expression followed by one or more statements.
2.	if...else statement An if statement can be followed by an optional else statement, which executes when the Boolean expression is false.
3.	switch statement A switch statement allows a variable to be tested for equality against a list of values.

1. If Statement

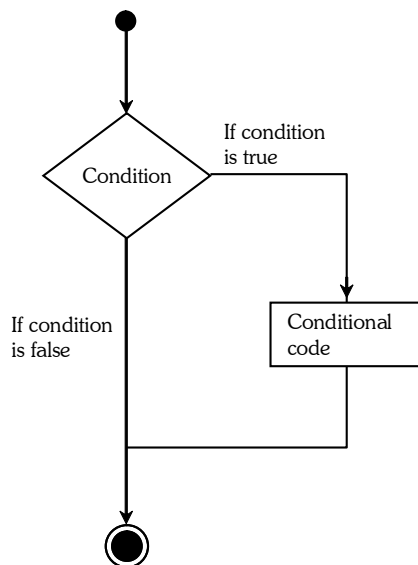
An **if** statement consists of a Boolean expression followed by one or more statements.

Syntax

The basic syntax for creating an **if** statement in R is:

```
if(boolean_expression) {
    // statement(s) will execute if the boolean
    // expression is true.
}
```

If the Boolean expression evaluates to be **true**, then the block of code inside the if statement will be executed. If Boolean expression evaluates to be **false**, then the first set of code after the end of the if statement (after the closing curly brace) will be executed.

Flow Diagram**Example**

```
x <- 30L
if(is.integer(x)){
    print("X is an Integer")
}
```

When the above code is compiled and executed, it produces the following result “

[1] “X is an Integer”

2. If Else statement

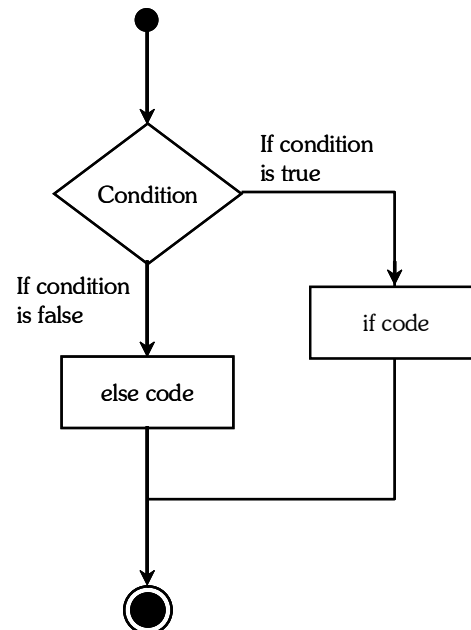
An **if** statement can be followed by an optional **else** statement which executes when the boolean expression is false.

Syntax

The basic syntax for creating an **if...else** statement in R is:

```
if(boolean_expression) {
    // statement(s) will execute if the boolean
    // expression is true.
} else {
    // statement(s) will execute if the boolean
    // expression is false.
}
```

If the Boolean expression evaluates to be **true**, then the **if block** of code will be executed, otherwise **else block** of code will be executed.

Flow Diagram**Example**

```
x <- c("what","is","truth")
if("Truth"%in% x){
    print("Truth is found")
}else{
    print("Truth is not found")
}
```

When the above code is compiled and executed, it produces the following result:

```
[1] "Truth is not found"
```

Here "Truth" and "truth" are two different strings.

The if...else if...else Statement

An **if** statement can be followed by an optional **else if...else** statement, which is very useful to test various conditions using single if...else if statement.

When using **if**, **else if**, **else** statements there are few points to keep in mind.

- An **if** can have zero or one **else** and it must come after any **else if**'s.
- An **if** can have zero to many **else if**'s and they must come before the **else**.
- Once an **else if** succeeds, none of the remaining **else if**'s or **else**'s will be tested.

Syntax

The basic syntax for creating an **if...else if...else** statement in R is:

```
if(boolean_expression 1) {  
    // Executes when the boolean expression 1 is true.  
} else if( boolean_expression 2) {  
    // Executes when the boolean expression 2 is true.  
} else if( boolean_expression 3) {  
    // Executes when the boolean expression 3 is true.  
} else {  
    // executes when none of the above condition is true.  
}
```

Example

```
x <- c("what", "is", "truth")  
if("Truth"%in%x){  
  print("Truth is found the first time")  
}elseif("truth"%in%x){  
  print("truth is found the second time")  
}else{  
  print("No truth found")  
}
```

When the above code is compiled and executed, it produces the following result:

```
[1] "truth is found the second time"
```

3. Switch statement

A **switch** statement allows a variable to be tested for equality against a list of values. Each value is called a case, and the variable being switched on is checked for each case.

Syntax

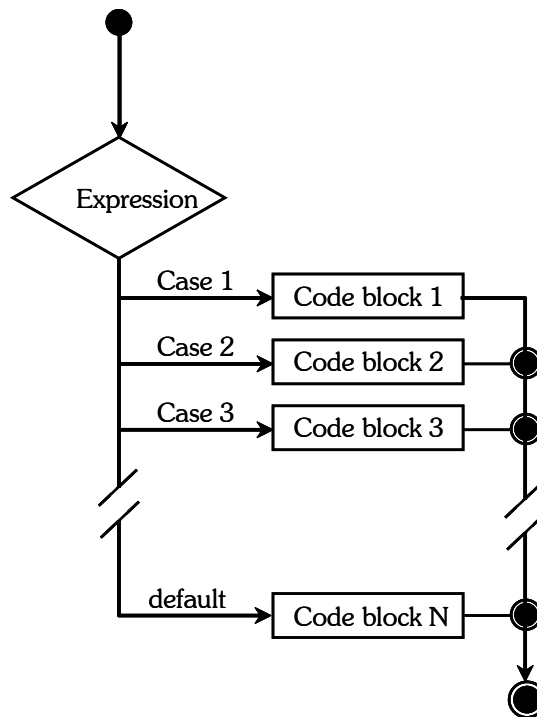
The basic syntax for creating a switch statement in R is:

```
switch(expression, case1, case2, case3....)
```

The following rules apply to a switch statement “

- If the value of expression is not a character string it is coerced to integer.
- You can have any number of case statements within a switch. Each case is followed by the value to be compared to and a colon.
- If the value of the integer is between 1 and nargs()”1 (The max number of arguments) then the corresponding element of case condition is evaluated and the result returned.
- If expression evaluates to a character string then that string is matched (exactly) to the names of the elements.
- If there is more than one match, the first matching element is returned.
- No Default argument is available.
- In the case of no match, if there is a unnamed element of ... its value is returned. (If there is more than one such argument an error is returned.)

Flow Diagram



Example

```
x <-switch(
3,
"first",
```

```
"second",  
"third",  
"fourth"  
)
```

```
print(x)
```

When the above code is compiled and executed, it produces the following result “

```
[1] "third"
```

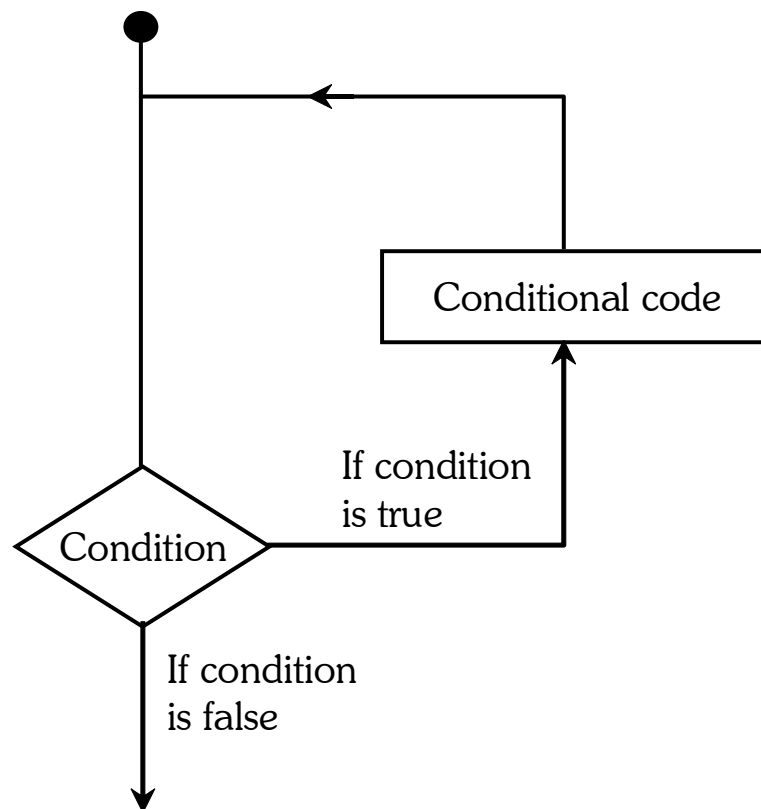
Q12. List and explain Looping statements supported by R.

Ans :

There may be a situation when you need to execute a block of code several number of times. In general, statements are executed sequentially. The first statement in a function is executed first, followed by the second, and so on.

Programming languages provide various control structures that allow for more complicated execution paths.

A loop statement allows us to execute a statement or group of statements multiple times and the following is the general form of a loop statement in most of the programming languages “



R programming language provides the following kinds of loop to handle looping requirements. Click the following links to check their detail.

Sr. No.	Loop Type & Description
1	repeat loop Executes a sequence of statements multiple times and abbreviates the code that manages the loop variable.
2	while loop Repeats a statement or group of statements while a given condition is true. It tests the condition before executing the loop body.
3	for loop Like a while statement, except that it tests the condition at the end of the loop body.

Q13. What are the loop control statements supported by R?

Ans :

Loop control statements change execution from its normal sequence. When execution leaves a scope, all automatic objects that were created in that scope are destroyed.

R supports the following control statements. Click the following links to check their detail.

Sr. No.	Control Statement & Description
1	break statement Terminates the loop statement and transfers execution to the statement immediately following the loop.
2	Next statement The next statement simulates the behavior of R switch.

5.6 FRAMES AND SUBSETS

Q14. What is mean by frame ? How can we create a frame in R?

Ans :

(Imp.)

A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

Following are the characteristics of a data frame.

- The column names should be non-empty.
- The row names should be unique.
- The data stored in a data frame can be of numeric, factor or character type.
- Each column should contain same number of data items.

Create Data Frame

```
# Create the data frame.
emp.data<-data.frame(
emp_id= c (1:5),
emp_name= c("Rick","Dan","Michelle","Ryan","Gary"),
```

```

salary = c(623.3,515.2,611.0,729.0,843.25),
start_date=as.Date(c("2012-01-01","2013-09-23","2014-11-15","2014-05-11", "2015-03-27")),
stringsAsFactors= FALSE
)
# Print the data frame.
print(emp.data)

```

When we execute the above code, it produces the following result -

emp_id	emp_name	salary	start_date
1	Rick	623.30	2012-01-01
2	Dan	515.20	2013-09-23
3	Michelle	611.00	2014-11-15
4	Ryan	729.00	2014-05-11
5	Gary	843.25	2015-03-27

Q15. What is the use of str(), summary() in data frames ?

Ans :

Str()

The structure of the data frame can be seen by using str() function.

```
str(emp.data)
```

When we execute the above code, it produces the following result -

'data.frame': 5 obs. of 4 variables:

\$ emp_id: int 1 2 3 4 5

\$ emp_name: chr "Rick" "Dan" "Michelle" "Ryan" ...

\$ salary: num 623 515 611 729 843

\$ start_date: Date, format: "2012-01-01" "2013-09-23" "2014-11-15" "2014-05-11" ...

Summary()

The statistical summary and nature of the data can be obtained by applying summary() function.

```
print(summary(emp.data))
```

When we execute the above code, it produces the following result -

emp_id	emp_name	salary	start_date
Min.:1	Length: 5	Min.:515.2	Min.:2012-01-01
1st Qu.:2	Class :character	1st Qu.:611.0	1st Qu.:2013-09-23
Median:3	Mode:character	Median:623.3	Median:2014-05-11
Mean:3		Mean:664.4	Mean :2014-01-14
3rd Qu.:4		3rd Qu.:729.0	3rd Qu.:2014-11-15
Max. :5		Max. :843.2	Max. :2015-03-27

Q16. How can we extract data from data frame ?*Ans :*

Extract specific column from a data frame using column name. # Extract Specific columns.

```
result <- data.frame(emp.data$emp_name, emp.data$salary)
print(result)
```

When we execute the above code, it produces the following result -

emp.data	emp_name	emp.data.salary
1	Rick	623.30
2	Dan	515.20
3	Michelle	611.00
4	Ryan	729.00
5	Gary	843.25

5.7 MANAGING AND MANIPULATING DATA IN R**Q17. How do you manipulate data in R?***Ans :***(Imp.)**

In order to manipulate the data, R provides a library called dplyr which consists of many built-in methods to manipulate the data. So to use the data manipulation function, first need to import the dplyr package using library(dplyr) line of code. Below is the list of a few data manipulation functions present in dplyr package.

Function Name	Description
filter()	Produces a subset of a Data Frame.
distinct()	Removes duplicate rows in a Data Frame
arrange()	Reorder the rows of a Data Frame
select()	Produces data in required columns of a Data Frame
rename()	Renames the variable names
mutate()	Creates new variables without dropping old ones.
transmute()	Creates new variables by dropping the old.
summarize()	Gives summarized data like Average, Sum, etc.

Q18. What is the use of filter() method in R?*Ans:***filter() method**

The filter() function is used to produce the subset of the data that satisfies the condition specified in the filter() method. In the condition, we can use conditional operators, logical operators, NA values, range operators etc. to filter out data. Syntax of filter() function is given below-

```
filter(dataframeName, condition)
```

Example

In the below code we used filter() function to fetch the data of players who scored more than 100 runs from the “stats” data frame.

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))
# fetch players who scored more
# than 100 runs
filter(stats, runs>100)
```

Output

	player	runs	wickets
1	B	200	20
2	C	408	NA

Q19. What is the use of distinct() method in R.

Ans :

The distinct() method removes duplicate rows from data frame or based on the specified columns. The syntax of distinct() method is given below-

```
distinct(dataframeName, col1, col2,..., .keep_all=TRUE)
```

Example

Here in this example, we used distinct() method to remove the duplicate rows from the data frame and also remove duplicates based on a specified column.

```
# import dplyr package
library(dplyr)

# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D', 'A', 'A'),
                    runs=c(100, 200, 408, 19, 56, 100),
                    wickets=c(17, 20, NA, 5, 2, 17))

# removes duplicate rows
distinct(stats)

#remove duplicates based on a column
distinct(stats, player, .keep_all = TRUE)
```


Output:

	player	runs	wickets
1	A	100	17
2	B	200	20
3	C	408	NA
4	D	19	5
5	A	56	2

	player	runs	wickets
1	A	100	17
2	B	200	20
3	C	408	NA
4	D	19	5

Q20. What is the use of arrange() method.

Ans :

In R, the arrange() method is used to order the rows based on a specified column. The syntax of arrange() method is specified below-`arrange(dataframeName, columnName)`

Example:

In the below code we ordered the data based on the runs from low to high using arrange() function.

```
# import dplyr package
library(dplyr)

# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))

# ordered data based on runs
arrange(stats, runs)
```

Output

	player	runs	wickets
1	D	19	5
2	A	100	17
3	B	200	20
4	C	408	NA

Q21. What is the use of select() method in R.*Ans :***select() method**

The select() method is used to extract the required columns as a table by specifying the required column names in select() method. The syntax of select() method is mentioned below-

```
select(dataframeName, col1,col2,...)
```

Example:

Here in the below code we fetched the player, wickets column data only using select() method.

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))
# fetch required column data
select(stats, player,wickets)
```

Output

	player	wickets
1	A	17
2	B	20
3	C	NA
4	D	5

Q22. What is the use of rename() method in R?*Ans :***(Imp.)**

rename() method

The rename() function is used to change the column names. This can be done by the below syntax-
rename(dataframeName, newName=oldName)

Example

In this example, we change the column name “runs” to “runs_scored” in stats data frame.

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))
# renaming the column
rename(stats, runs_scored=runs)
```

Output

	player	runs_scored	wickets
1	A	100	17
2	B	200	20
3	C	408	NA
4	D	19	5

Q23. What are the uses of mutate() and transmute() methods in R?*Ans :***mutate() & transmute() methods**

These methods are used to create new variables. The mutate() function creates new variables without dropping the old ones but transmute() function drops the old variables and creates new variables. The syntax of both methods is mentioned below-

```
mutate(dataframeName, newVariable=formula)
```

```
transmute(dataframeName, newVariable=formula)
```

Example

In this example, we created a new column avg using mutate() and transmute() methods.

```
# import dplyr package
```

```
library(dplyr)
```

```
# create a data frame
```

```
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, 7, 5))
```

```
# add new column avg
```

```
mutate(stats, avg=runs/4)
```

```
# drop all and create a new column
```

```
transmute(stats, avg=runs/4)
```

Output

	player	runs	wickets	avg
1	A	100	17	25.00
2	B	200	20	50.00
3	C	408	7	102.00
4	D	19	5	4.75

avg

```
1 25.00
```

```
2 50.00
```

```
3 102.00
```

```
4 4.75
```

Short Questions and Answers

1. What are the features of R-Programming?

Ans :

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R.

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

2. What is mean by function? Write the syntax of function in R?

Ans :

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.

In R, a function is an object so the R interpreter is able to pass control to the function, along with arguments that may be necessary for the function to accomplish the actions.

The function in turn performs its task and returns control to the interpreter as well as any result which may be stored in other objects.

Function Definition

An R function is created by using the keyword **function**. The basic syntax of an R function definition is as follows “

```
function_name <- function(arg_1, arg_2, ...) {
    Function body
}
```

3. What is mean by user defined function?

Ans :

User-defined Function

We can create user-defined functions in R. They are specific to what a user wants and once created they can be used like the built-in functions. Below is an example of how a function is created and used.

Create a function to print squares of numbers in sequence.

```
new.function <- function(a){
  for(i in 1:a){
    b <- i^2
    print(b)
  }
}
```

4. Draw and explain the flow chart of If Else statement.

Ans :

If Else statement

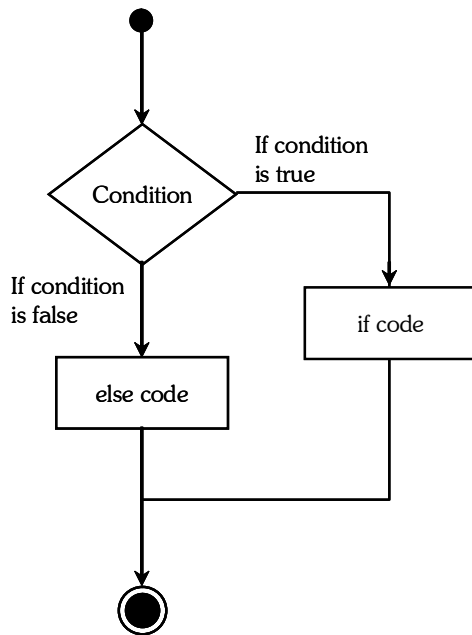
An **if** statement can be followed by an optional **else** statement which executes when the boolean expression is false.

Syntax

The basic syntax for creating an **if...else** statement in R is:

```
if(boolean_expression) {
    // statement(s) will execute if the boolean
    expression is true.
} else {
    // statement(s) will execute if the boolean
    expression is false.
}
```

If the Boolean expression evaluates to be **true**, then the **if block** of code will be executed, otherwise **else block** of code will be executed.

Flow Diagram**Example**

```

x <- c("what", "is", "truth")
if("Truth"%in% x){
  print("Truth is found")
}else{
  print("Truth is not found")
}
  
```

When the above code is compiled and executed, it produces the following result:

5. What is the use of switch statement?

Ans :

A **switch** statement allows a variable to be tested for equality against a list of values. Each value is called a case, and the variable being switched on is checked for each case.

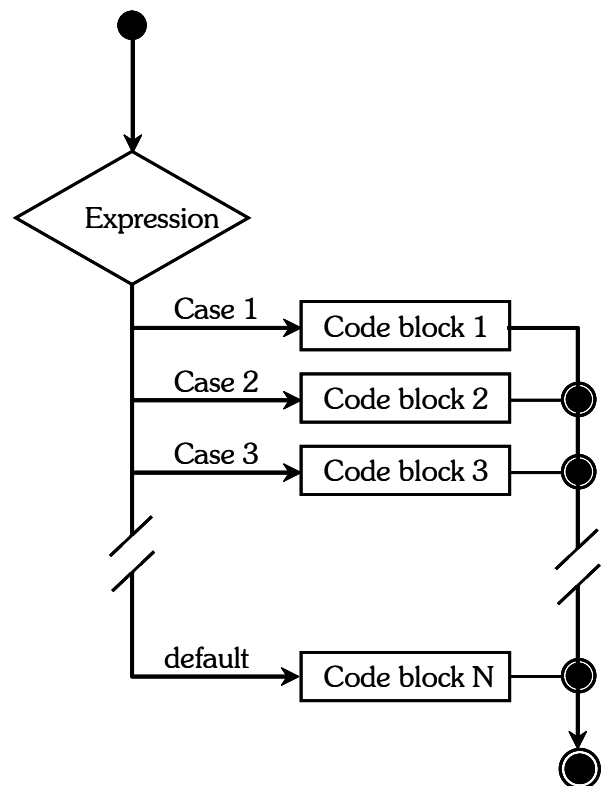
Syntax

The basic syntax for creating a switch statement in R is:

```
switch(expression, case1, case2, case3....)
```

The following rules apply to a switch statement -

- If the value of expression is not a character string it is coerced to integer.
- You can have any number of case statements within a switch. Each case is followed by the value to be compared to and a colon.
- If the value of the integer is between 1 and nargs() then the corresponding element of case condition is evaluated and the result returned.
- If expression evaluates to a character string then that string is matched (exactly) to the names of the elements.
- If there is more than one match, the first matching element is returned.
- No Default argument is available.
- In the case of no match, if there is a unnamed element of ... its value is returned. (If there is more than one such argument an error is returned.)

Flow Diagram

Example

```
x <-switch(
3,
"first",
"second",
"third",
"fourth"
)
print(x)
```

When the above code is compiled and executed, it produces the following result “

6. What are the loop control statements supported by R?

Ans :

Loop control statements change execution from its normal sequence. When execution leaves a scope, all automatic objects that were created in that scope are destroyed.

R supports the following control statements. Click the following links to check their detail.

Sr. No.	Control Statement & Description
1	break statement Terminates the loop statement and transfers execution to the statement immediately following the loop.
2	Next statement The next statement simulates the behavior of R switch.

7. How do you manipulate data in R?

Ans :

In order to manipulate the data, R provides a library called dplyr which consists of many built-in methods to manipulate the data. So to use the data manipulation function, first need to import the dplyr package using library(dplyr) line of code. Below is the list of a few data manipulation functions present in dplyr package.

Function Name	Description
filter()	Produces a subset of a Data Frame.
distinct()	Removes duplicate rows in a Data Frame
arrange()	Reorder the rows of a Data Frame
select()	Produces data in required columns of a Data Frame
rename()	Renames the variable names
mutate()	Creates new variables without dropping old ones.
transmute()	Creates new variables by dropping the old.
summarize()	Gives summarized data like Average, Sum, etc.

8. What is the use of arrange() method?*Ans :*

In R, the arrange() method is used to order the rows based on a specified column. The syntax of arrange() method is specified below-`arrange(dataframeName, columnName)`

Example:

In the below code we ordered the data based on the runs from low to high using arrange() function.

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))
# ordered data based on runs
arrange(stats, runs)
```

Output

	player	runs	wickets
1	D	19	5
2	A	100	17
3	B	200	20
4	C	408	NA

9. What is the use of rename() method in R?*Ans :*

rename() method

The rename() function is used to change the column names. This can be done by the below syntax-`rename(dataframeName, newName=oldName)`

Example

In this example, we change the column name “runs” to “runs_scored” in stats data frame.

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),
                    runs=c(100, 200, 408, 19),
                    wickets=c(17, 20, NA, 5))
# renaming the column
rename(stats, runs_scored=runs)
```

Output

	player	runs_scored	wickets
1	A	100	17
2	B	200	20
3	C	408	NA
4	D	19	5

Choose the Correct Answers

1. What is R? [a]
(a) A statistical programming language (b) A spreadsheet program
(c) A web development language (d) An operating system
2. Which of the following is the correct syntax for assigning a value to a variable in R? [a]
(a) var = 10 (b) 10 = var
(c) var == 10 (d) var := 10
3. R was created by? [c]
(a) Ross Ihaka (b) Robert Gentleman
(c) Both A and B (d) Ross Gentleman
4. Which of the following is true about R? [-]
(a) R is a well-developed, simple and effective programming language
(b) R has an effective data handling and storage facility
(c) R provides a large, coherent and integrated collection of tools for data analysis.
(d)
5. R files has an extension _____. [c]
(a) .S (b) .RP
(c) .R (d) .SP
6. How many atomic vector types does R have? [d]
(a) 3 (b) 4
(c) 5 (d) 6
7. R language is a dialect of which of the following languages? [a]
(a) s (b) c
(c) sas (d) matlab
8. Point out the wrong statement? [b]
(a) Setting up a workstation to take full advantage of the customizable features of R is a straightforward thing
(b) q() is used to quit the R program
(c) R has an inbuilt help facility similar to the man facility of UNIX
(d) Windows versions of R have other optional help systems also
9. R made its first appearance in? [c]
(a) 1992 (b) 1995
(c) 1993 (d) 1994
10. R allows integration with the procedures written in the? [a]
(a) C (b) Ruby
(c) Java (d) Basic

Fill in the blanks

1. A _____ is a series of data elements of the same basic type.
2. A two-dimensional data structure used to bind the vectors from the same length, known as the _____.
3. A _____ is a generic form of a matrix. It is a combination of lists and matrices.
4. An _____ is a symbol that tells the compiler to perform specific mathematical or logical manipulations.
5. A _____ is a set of statements organized together to perform a specific task.
6. _____ are the R objects which contain elements of different types like “ numbers, strings, vectors and another list inside it.
7. A _____ is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.
8. An array is created using the _____ function.
9. An _____ consists of a Boolean expression followed by one or more statements.
10. R is a _____ and software environment for statistical analysis, graphics representation and reporting.

ANSWERS

1. Vector
2. Matrix.
3. Data frame
4. Operator
5. Function
6. Lists
7. Data frame
8. Array()
9. if statement
10. Programming language

One Mark Answers

1. Function

Ans :

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.

2. Decision Making

Ans :

Decision making structures require the programmer to specify one or more conditions to be evaluated or tested by the program, along with a statement or statements to be executed if the condition is determined to be true, and optionally, other statements to be executed if the condition is determined to be false.

3. Operator

Ans :

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations.

4. List

Ans :

Lists are the R objects which contain elements of different types like “ numbers, strings, vectors and another list inside it. A list can also contain a matrix or a function as its elements. List is created using `list()` function.

5. Frame

Ans :

A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

FACULTY OF MANAGEMENT
BBA V - Semester (CBCS) Examination
Model Paper - I
BUSINESS ANALYTICS

Time : 3 Hours]

[Max. Marks : 80

PART - A (5 × 4 = 20 Marks)

[Short Answer Type]

ANSWERS

Note : Answer any five of the following questions.

- | | |
|---|--------------------|
| 1. What is Business Analytics? | (Unit-I, SQA-9) |
| 2. What is Diagnostic Analytics? | (Unit-I, SQA-2) |
| 3. What is descript analytics? | (Unit-II, SQA-1) |
| 4. Benefits of data visualization | (Unit-II, SQA-10) |
| 5. What is a Trend Line? Explain the need of trend lines. | (Unit-III, SQA-2) |
| 6. What is mean by non linear optimization? | (Unit-IV, SQA-3) |
| 7. What are the features of R-Programming? | (Unit-V, SQA-1) |
| 8. What are the applications of Datamining. | (Unit-III, SQA-18) |

PART - B (5 × 12 = 60 Marks)

[Essay Answer Type]

Note : Answer all the questions using the internal choice.

- | | |
|---|----------------------|
| 9. (a) Discuss various types of business analytics | (Unit-I, Q.No. 4) |
| OR | |
| (b) What are the characteristics of Big Data. | (Unit-I, Q.No. 10) |
| 10. (a) What is mean by descriptive statistics? and list out the types of descriptive statistics. | (Unit-II, Q.No. 4) |
| OR | |
| (b) List and explain Dashboard best practices. | (Unit-II, Q.No. 22) |
| 11. (a) Discuss in detail about linear regression. | (Unit-III, Q.No. 10) |
| OR | |
| (b) What is Data Exploration? | (Unit-III, Q.No. 22) |

12. (a) What is Linear optimization? What are the applications of linear optimization. (Unit-IV, Q.No. 3)

OR

- (b) Discuss in detail about non linear optimization (Unit-IV, Q.No. 8)

13. (a) What is R Programming ? What re the features of R ? (Unit-V, Q.No. 1)

OR

- (b) What are the functions are useful for reading data into R and writing data to files. (Unit-V, Q.No. 5)

FACULTY OF MANAGEMENT**BBA V - Semester (CBCS) Examination****Model Paper - II****BUSINESS ANALYTICS**

Time : 3 Hours]

[Max. Marks : 80

PART - A (5 × 4 = 20 Marks)**[Short Answer Type]****ANSWERS****Note : Answer any five of the following questions.**

- | | |
|--|-------------------|
| 1. What is Predictive Analytics? | (Unit-I, SQA-3) |
| 2. What is Structured Data? | (Unit-I, SQA-5) |
| 3. What is median? Explain with a simple example. | (Unit-II, SQA-4) |
| 4. What is mean by Data Dashboard? | (Unit-II, SQA-12) |
| 5. What is Paralled Trendline. | (Unit-III, SQA-6) |
| 6. What is decision there analysis ? | (Unit-IV, SQA-5) |
| 7. Draw and explain the flow chart of If Else statement. | (Unit-V, SQA-4) |
| 8. What is Risk Analysis? | (Unit-IV, SQA-6) |

PART - B (5 × 12 = 60 Marks)**[Essay Answer Type]****Note : Answer all the questions using the internal choice.**

- | | |
|---|----------------------|
| 9. (a) What are the analytical methods used in business analytics ? | (Unit-I, Q.No. 6) |
| OR | |
| (b) Discuss various types of data. | (Unit-I, Q.No. 12) |
| 10. (a) Explain in detail about variance. | (Unit-II, Q.No. 8) |
| OR | |
| (b) Explain the process of inserting a chart in Excel. | (Unit-II, Q.No. 18) |
| 11. (a) Discuss various predictive models. | (Unit-III, Q.No. 13) |
| OR | |
| (b) What is business intelligence? | (Unit-III, Q.No. 26) |

12. (a) What is Cutting plane algorithm? (Unit-IV, Q.No. 9)

OR

(b) What are the steps involved in risk analysis. (Unit-IV, Q.No. 13)

13. (a) Discuss various types of functions supported by R. (Unit-V, Q.No. 9)

OR

(b) Discuss in detail about decision making statements supported by R. (Unit-V, Q.No. 11)

FACULTY OF MANAGEMENT**BBA V - Semester (CBCS) Examination****Model Paper - III****BUSINESS ANALYTICS**

Time : 3 Hours]

[Max. Marks : 80

PART - A (5 × 4 = 20 Marks)**[Short Answer Type]****ANSWERS****Note : Answer any five of the following questions.**

- | | |
|--|--------------------|
| 1. What are the advantages and disadvantages of using decision models. | (Unit-I, SQA-8) |
| 2. What is Nominal Data? | (Unit-I, SQA-6) |
| 3. What is Normal Distribution? | (Unit-II, SQA-7) |
| 4. Write the steps to calculate standard deviation. | (Unit-II, SQA-5) |
| 5. What is Decision Tree. | (Unit-III, SQA-9) |
| 6. What is web analytics ? What is the importance of web analytics. | (Unit-IV, SQA-8) |
| 7. What is mean by function? Write the syntax of function in R? | (Unit-V, SQA-2) |
| 8. What are the differences between Data Mining and Data Intelligence. | (Unit-III, SQA-16) |

PART - B (5 × 12 = 60 Marks)**[Essay Answer Type]****Note : Answer all the questions using the internal choice.**

- | | |
|--|----------------------|
| 9. (a) What are the applications of business analytics. | (Unit-I, Q.No. 8) |
| OR | |
| (b) Discuss various types of decision making models. | (Unit-I, Q.No. 15) |
| 10. (a) How to create cross tab in Excel? | (Unit-II, Q.No. 16) |
| OR | |
| (b) Explain Central tendency in point of distributions. | (Unit-II, Q.No. 7) |
| 11. (a) What is Data mining? Give a brief overview on data mining. | (Unit-III, Q.No. 18) |
| OR | |
| (b) Discuss in detail about cause and effect modelling. | (Unit-III, Q.No. 32) |

12. (a) What are the steps involved in text analytics. (Unit-IV, Q.No. 17)

OR

(b) What is decision there analysis ? (Unit-IV, Q.No. 10)

13. (a) What is mean by frame ? How can we create a frame in R? (Unit-V, Q.No. 14)

OR

(b) What is the use of filter() method in R? (Unit-V, Q.No. 18)