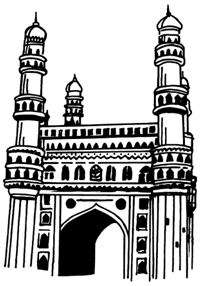
Rahul's ✔

Topper's Voice

AS PER CBCS SYLLABUS



MBA

II Year III Sem

(Osmania University)

Latest 2021-22 Edition

BUSINESS ANALYTICS

- ☞ **Study Manual**
- ☞ **Short Questions and Answers**
- ☞ **Multiple Choice Questions**
- ☞ **Fill in the blanks**
- ☞ **Solved Model Papers**
- ☞ **Solved Previous Question Paper**

Price
` .199-00

- by -

Well Experienced Lecturer

Rahul Publications™

Hyderabad. Ph : 66550071, 9391018098

MBA

II Year III Sem

(Osmania University)

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.

Price ` 199-00

Sole Distributors : ☎ : 66550071, Cell : 9391018098

VASU BOOK CENTRE

Shop No. 3, Beside Gokul Chat, Koti, Hyderabad.

Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.
Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

CONTENTS

BUSINESS ANALYTICS

STUDY MANUAL

SOLVED MODEL PAPERS

SOLVED PREVIOUS QUESTION PAPER

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.

UNIT - II

Descriptive Analytics

Over view of Description Statistics (Central Tendency, Variability), Data Visualization-Definition, Visualization Techniques - Tables, Cross Tabulations, charts, Data Dash boards using Ms-Excel or SPSS.

UNIT - III

Predictive Analytics

Trend Lines, Regression Analysis –Linear & Multiple, Forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Classification, Association, Cause Effect Modeling.

UNIT - IV

Prescriptive Analytics

Overview of Linear Optimization, Non Linear Program-ming Integer Optimization, Cutting Plane algorithm and other methods, Decision Analysis - Risk and uncertainty methods

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

Topic **Page No.**

UNIT - I

UNIT - II

Topic Page No.

UNIT - V

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.

1.1 Definition of Business analytics

Q1. What is Business analytics ?

Ans :

Business analytics is the practice of iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.

Definition of business analytics

According to Schaer (2018) - "allows your business to make predictive analysis rather than reacting to changes in data".

According to Gabelli School of Business (2018)- "involves applying models, methods, and tools to data, producing insights that lead to informed business decisions"

According to Wells (2008) - "the application of logic and mental processes to find meaning in data"

According to Lynda (2018) - "allows us to learn from the past and make better predictions for the future".

Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods. Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling, and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (predict), and what is the best outcome that can happen (optimize).

Business analytics. Abbreviated as BA, business analytics is the combination of skills, technologies, applications and processes used by organizations to gain insight in to their business based on data and statistics to drive business planning

Q2. What is Data for business Analytics.

Ans :

A business analytics is used to gain insights that inform business decisions and can be used to automate and optimize business processes. Data-driven companies treat their data as a corporate asset and leverage it for a competitive advantage. Successful business analytics depends on data quality, skilled analysts who understand the technologies and the business, and an organizational commitment to data-driven decision-making.

Business analytics examples

- Business analytics techniques break down into two main areas. The first is basic business intelligence. This involves examining historical data to get a sense of how a business department, team or staff member performed over a particular time. This is a mature practice that most enterprises are fairly accomplished at using.
- The second area of business analytics involves deeper statistical analysis. This may mean doing predictive analytics by applying statistical algorithms to historical data to make a prediction about future performance of a product, service or website design change. Or, it could mean using other advanced analytics techniques, like cluster analysis, to group customers based on similarities across several data points. This can be helpful in targeted marketing campaigns, for example.

Business analytics tools come in several different varieties:

1. Data visualization tools
2. Business intelligence reporting software
3. Self-service analytics platforms
4. Statistical analysis tools
5. Big data platforms

- Self-service has become a major trend among business analytics tools. Users now demand software that is easy to use and doesn't require specialized training. This has led to the rise of simple-to-use tools from companies such as Tableau and Qlik, among others. These tools can be installed on a single computer for small applications or in server environments for enterprise-wide deployments. Once they are up and running, business analysts and others with less specialized training can use them to generate reports, charts and web portals that track specific metrics in data sets

- Once the business goal of the analysis is determined, an analysis methodology is selected and data is acquired to support the analysis. Data acquisition often involves extraction from one or more business systems, data cleansing and integration into a single repository, such as a data warehouse or data mart. The analysis is typically performed against a smaller sample set of data.

- Analytics tools range from spreadsheets with statistical functions to complex data mining and predictive modeling applications. As patterns and relationships in the data are uncovered, new questions are asked, and the analytical process iterates until the business goal is met.

- Deployment of predictive models involves scoring data records - typically in a database - and using the scores to optimize real-time decisions within applications and business processes. BA also supports tactical decision-making in response to unforeseen events. And, in many cases, the decision-making is automated to support real-time responses.

1.2 Categories of Business Analytical Methods and Models

Q3. Explain the different business analytical methods.

Ans :

There are four types in business analytics

(i) **Prescriptive:** This type of analysis reveals what actions should be taken. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps.

(ii) **Predictive:** An analysis of likely scenarios of what might happen. The deliverables are usually a predictive forecast.

(iii) **Diagnostic:** A look at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.

(iv) **Descriptive:** What is happening now based on incoming data. To mine the analytics, you typically use a real-time dashboard and/or email reports.

1. **Prescriptive analytics :** It is really valuable, but largely not used. Where big data analytics in general sheds light on a subject, prescriptive analytics gives you a laser-like focus to answer specific questions. For example, in the health care industry, you can better manage the patient population by using prescriptive analytics to measure the number of patients who are clinically obese, then add filters for factors like diabetes and LDL cholesterol levels to determine where to focus treatment. The same prescriptive model can be applied to almost any industry target group or problem.

2. **Predictive analytics :** It use big data to identify past patterns to predict the future. For example, some companies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc. Properly tuned predictive analytics can be used to support sales, marketing, or for other types of complex forecasts.

3. **Diagnostic analytics :** They are used for discovery or to determine why something happened. For example, for a social media marketing campaign, you can use descriptive analytics to assess the number of posts, mentions, followers,

fans, page views, reviews, pins, etc. There can be thousands of online mentions that can be distilled into a single view to see what worked in your past campaigns and what didn't.

4. **Descriptive analytics :** Descriptive analysis (or) data mining are at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight. A simple example of descriptive analytics would be assessing credit risk; using past financial performance to predict a customer's likely financial performance. Descriptive analytics can be useful in the sales cycle, for example, to categorize customers by their likely product preferences and sales cycle.

Q4. Explain the different models in Business Analytics?

Ans :

An analytical model is simply a mathematical equation that describes relationships among variables in a historical data set. The equation either estimates or classifies data values. In essence, a model draws a "line" through a set of data points that can be used to predict outcomes. What is a business analysis model?

Simply put, a business analysis model outlines the steps a business takes to complete a specific process, such as ordering a product or on boarding a new hire. Process modeling (or mapping) is key to improving process efficiency, training, and even complying with industry regulations.

Because there are many different kinds of processes, organizations, and functions within a business, BAs employ a variety of visual models to map and analyze data.

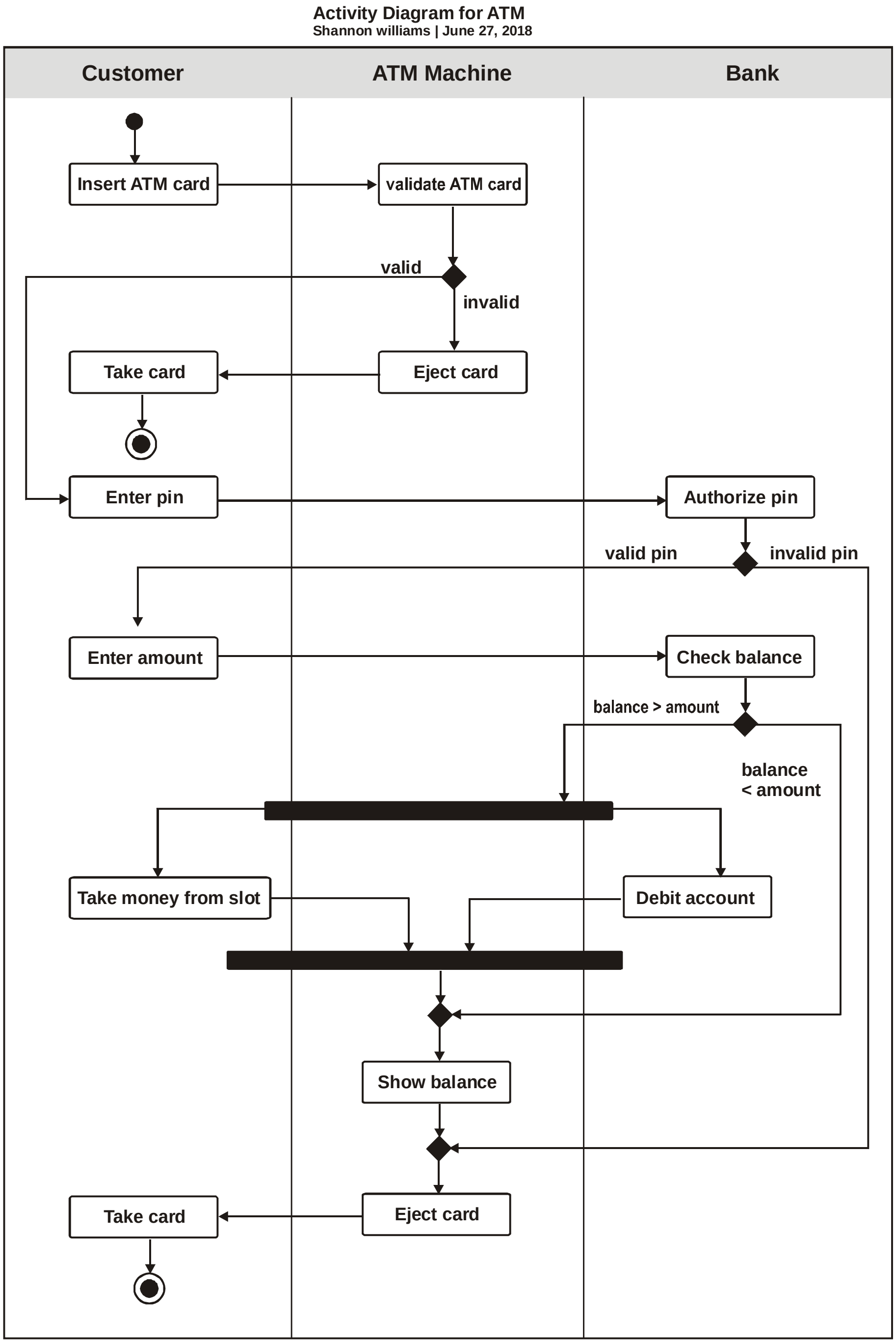
The following the different models:

1. **Activity diagrams**

Activity diagrams are a type of UML behavioural diagram that describes what needs to happen in a system. They are particularly useful for communicating process and procedure to stakeholders from both the business and development teams.

A Business analytical might use an activity diagram to map the process of logging in to a website or completing a transaction like withdrawing or depositing money

Activity diagram for ATM

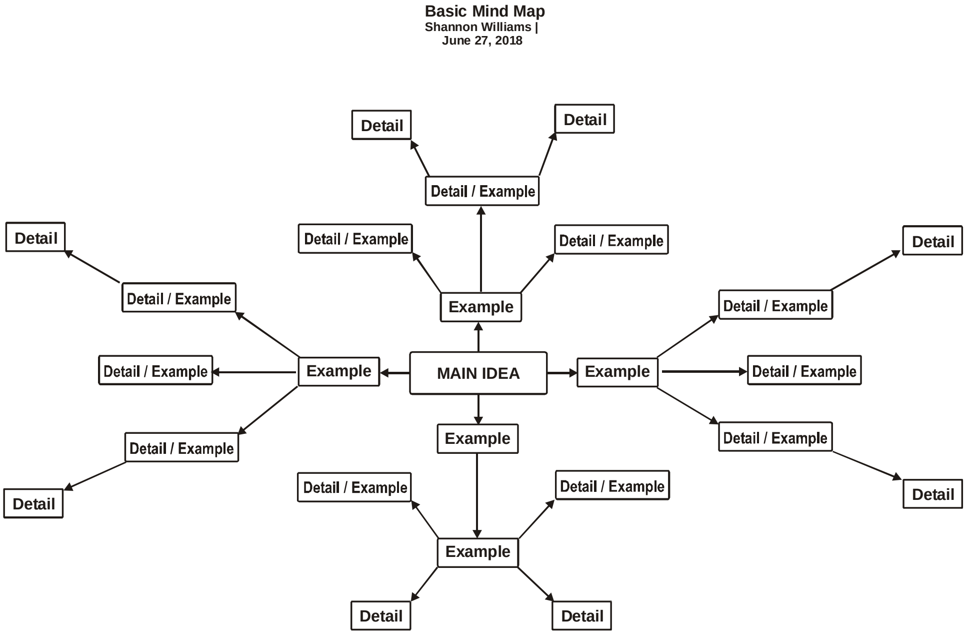


2. Feature mind maps

Business diagrams aren't just for late-stage analysis or documentation. They are also useful during a project's initial brainstorming phase. Feature mind maps help BAs organize the sometimes messy brainstorm process so that ideas, concerns, and requests are clearly captured and categorized.

This visual ensures initial details and ideas don't fall through the cracks so you can make informed decisions about project direction, goals, and scope down the line.

Basic mind map



3. Product roadmaps

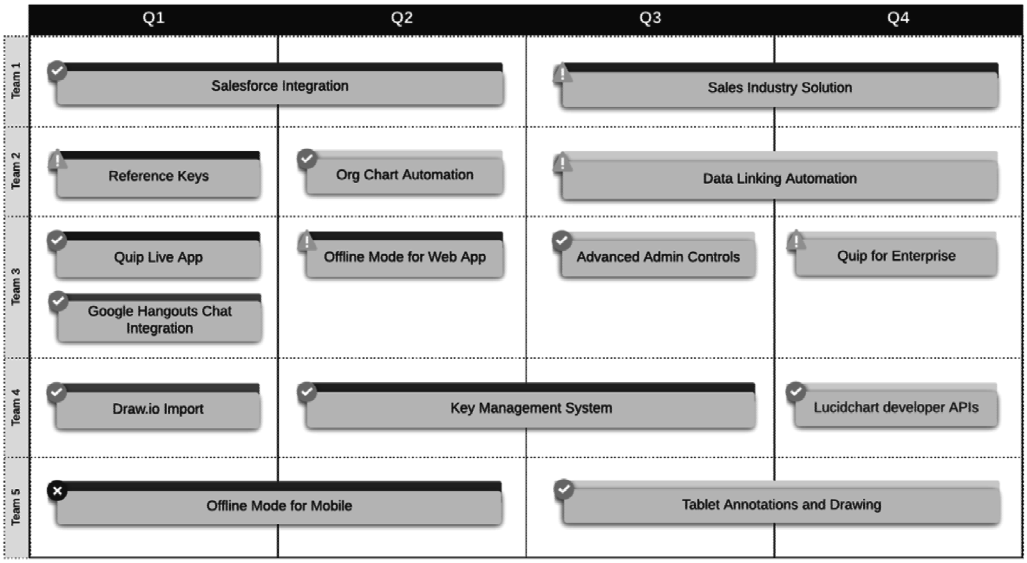
Product (or feature) roadmaps outline the development and launches of a product and its features. They are a focused analysis of a product's evolution, which helps developers and other stakeholders focus on initiatives that add direct value to the user.

The beauty of product roadmaps lies in their flexibility and range of applications. BAs can create different product roadmaps to illustrate different information, including:

- Maintenance and bug fixes
- Feature releases
- High-level strategic product goals

While product roadmaps are commonly used internally by development teams, they are also useful resources for other groups like sales.

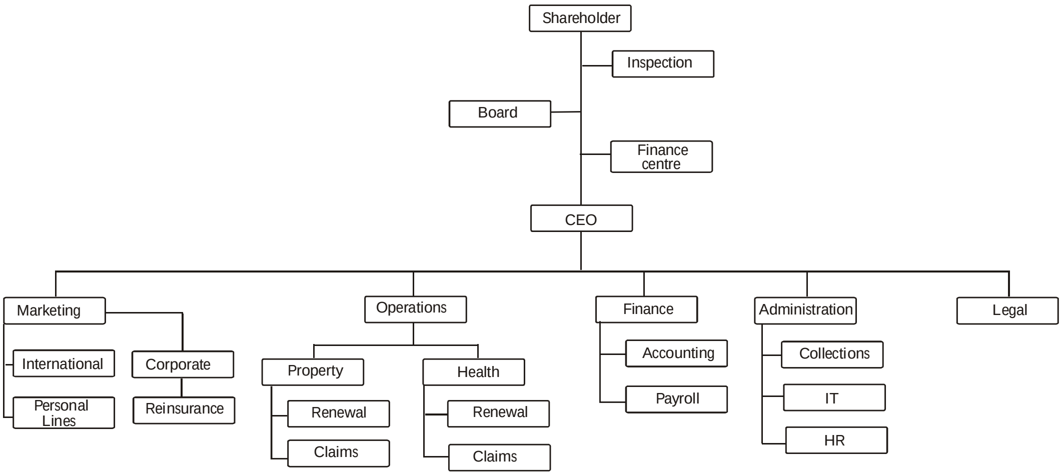
A defined product outline and schedule helps sales stay on the same page as the developers so they can deliver accurate, updated information to their prospects and clients. Because of their versatility and broad applications across teams and organizations, product roadmaps are a core part of an analyst's toolbox.



4. Organizational charts

Organizational charts outline the hierarchy of a business or one of its departments or teams. They are especially helpful reference charts for employees to quickly understand how the company is organized and identify key stakeholders and points of contact for projects or queries.

Additionally, organizational charts prove useful for stakeholder analysis and modeling new groupings and teams following organizational shifts.



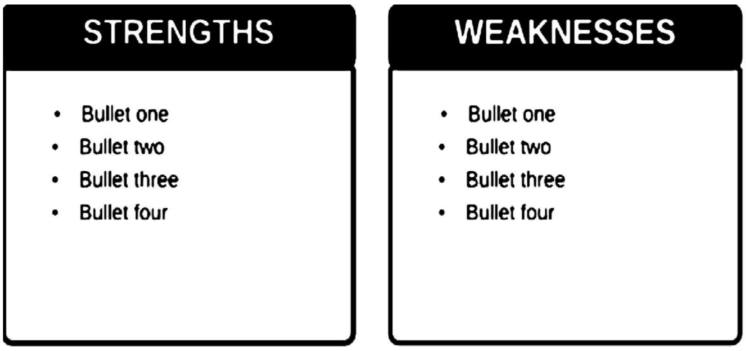
5. SWOT analysis

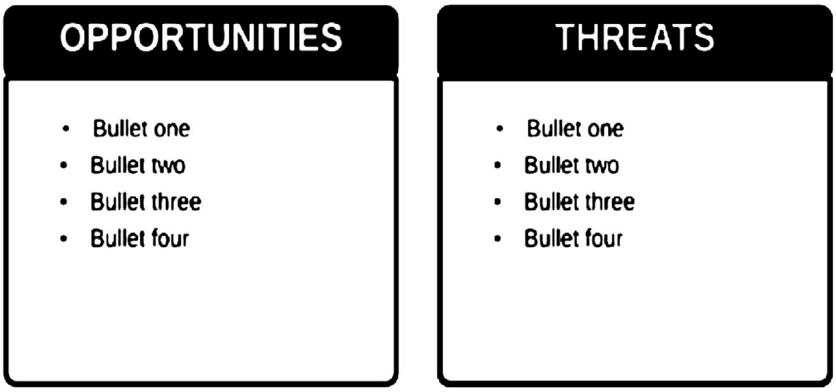
The SWOT analysis is a fundamental tool in a Business analytics. SWOT stands for strengths, weaknesses, opportunities, and threats. A SWOT analysis evaluates a business's strengths and weaknesses and identifies any opportunities or threats to that business.

SWOT analysis helps stakeholders make strategic decisions regarding their business. The goal is to capitalize on strengths and opportunities while reducing the impact of internal or external threats and weaknesses.

From a visual modeling perspective, SWOT analysis is fairly straight forward. A typical model will have four boxes or quadrants-one for each category-with bulleted lists outlining the respective results.

SWOT Analysis





6. User interface wireframe

Another essential business diagram is the UI wireframe. Software development teams use wireframes (also called mockups or prototypes) to visually outline and design a layout for a specific screen. In other words, wireframes are the blueprints for a website or software program. They help stakeholders assess navigational needs and experience for a successful practical application.

Wireframes range from low-fidelity to high-fidelity prototypes. Low-fidelity wireframes are the most basic outlines, showing only the bare-bones layout of the screen. High-fidelity wireframes are typically rendered in the later planning stages and will include specific UI elements (e.g., buttons, drop-down bars, text fields, etc.) and represent how the final implementation should look on the screen.

Website Design Wireframe (Click on image to modify online)

7. Process flow diagram

A process flow diagram (PFD) is typically used in chemical and process engineering to identify the basic flow of plant processes, but it can also be used in other fields to help stakeholders understand how their organization operates.

A PFD is best used to:

- Document a process.
- Study a process to make changes or improvements.
- Improve understanding and communication between stakeholders.

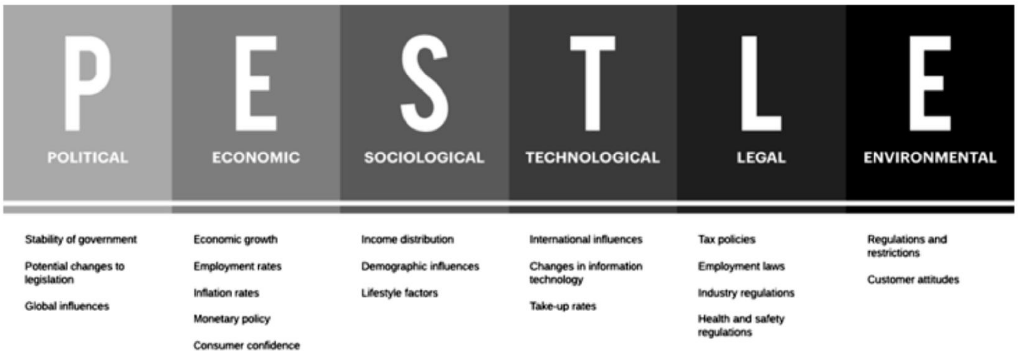
These diagrams focus on broad, high-level systems rather than annotating minor process details.

8. PESTLE analysis

A PESTLE analysis often goes hand-in-hand with a SWOT analysis. PESTLE evaluates external factors that could impact business performance. This acronym stands for six elements affecting business: political, economic, technological, environmental, legal, and sociological.

PESTLE analysis assesses the possible factors within each category, as well as their potential impact, duration of effect, type of impact (i.e., negative or positive), and level of importance.

This type of business analysis helps stakeholders manage risk, strategically plan and review business goals and performance, and potentially gain an advantage over competitors.



9. Entity-relationship diagram

An entity-relationship diagram (ER diagram) illustrates how entities (e.g., people, objects, or concepts) relate to one another in a system. For example, a logical ER diagram visually shows how the terms in an organization's business glossary relate to one another.

ER diagrams comprise three main parts:

- Entities
- Relationships
- Attributes

Attributes apply to the entities, describing further details about the concept. Relationships are where the key insights from ER diagrams arise. In a visual model, the relationships between entities are illustrated either numerically or via crow's foot notation.

These diagrams are most commonly used to model database structures in software engineering and business information systems and are particularly valuable tools for Business analytics in those fields.

Q5. Discuss briefly about role of business analytics in current business environment.

Ans :

1. Financial Analytics

Organizations use predictive models for forecasting future financial performance for constructing financial instruments like derivatives and assessing the risk involved in investment projects and portfolios. They also use prescriptive models for creating optimal capital budgeting plans for constructing optimal portfolios of investments and allocating assets. Addition to this, simulation is also used for ascertaining risk in the financial sector.

Example : GE Asset Management utilizes optimization models of analytics to make investment decisions of cash received from various sources. The approximate benefit obtained from using optimization models over a five-year period was $ 75 million.

2. Marketing Analytics

Business analytics is used in marketing for obtaining a better understanding of consumer behaviours by using the scanned data and social networking data. It leads to efficient use of advertising budgets, improved demand forecasting, effective pricing strategies, increased product line management and improved customer loyalty and satisfaction. Marketing analytics has gained much interest due to the data generated from social media.

Example : NBC Universal utilizes a predictive model every year to aid the annual up front market. An upfront market is a period in ending of May when every TV network sells most of the on-air advertisements for the upcoming season of television. The results of forecasting model are utilized by more than 200 NBC sales for supporting sales and pricing decisions.

3. Human Resource (HR) Analytics

HR function utilizes analytics to ensure that the organization consists of the employees with required skills to meet its needs, to ensure that it achieves its diversity goals and to ensure that it is hiring talent of the highest quality and also offering an environment which retains it.

Example : Sears Holding Corporation (SHC) owners of Roebuck Company, retailers Kmart and Sears. They made a team of HR analytics inside the corporate HR function. They apply predictive and descriptive analytics for tracking and influencing retention of employees and for supporting employee hiring.

4. Health Care Analytics

Health care organizations utilize analytics to provide more effective treatment and control cost. They use descriptive, prescriptive and predictive analytics for improving patient flow, staff and facility scheduling, purchasing and control of inventory. However, prescriptive analytics is specially used for the purpose of treatment and diagnosis. It is the most important proven utility of analytics.

Example : Memorial Sloan-Kettering Cancer Center along with Georgia Institute of Technology created a real-time prescriptive model for determining the optimal placement of radio active seeds for prostate cancer treatment. The results led to requirements of 20-30% lesser seeds and less invasive and faster procedure.

5. Supply Chain Analytics

Analytics is used by logistics and supply chain management to achieve efficiency. The entire spectrum of analytics is utilized by them. Various organizations such as UPS and FedEx apply analytics for efficient delivery of goods. Analytics helps them in optimal sorting of goods, staff and vehicle scheduling and vehicle routing, which helps in increasing the profitability. Analytics enable better processing control, inventory and more effective supply chains.

Example : ConAgra Foods utilized the prescriptive and predictive analysis for a better plan capacity utilization by incorporation of inherent uncertainty in pricing of commodities. ConAgra Foods attained a 100% return on investment in just three months.

6. Analytics for Government and Non-Profit Organizations

Government and non-profit organizations apply analytics for driving out inefficiencies and increasing the accountability and effectiveness of programs. During the period of Word War II, advanced analytics was first applied by the

English and U.S. Military. Analytics applicability is very extensive in government agencies from elections to tax collections. Non-profit organizations utilize analytics for ensuring the accountability and effectiveness to their clients and donors.

Examples : The New York State Department incorporated with IBM for using prescriptive analytics in developing a more efficient tax collection approach.

Catholic Relief Services (CRS) is a non-profit organization which is the official international humanitarian agency of the U.S. Catholic community. This offer helps to the victims of both human-made and natural disasters. It also offers various other services through its agricultural, educational and health programs. It utilizes analytical spread sheet model for helping in the annual budget allocation based on the effects of its relief programs and efforts in various countries.

7. Sports Analytics

Analytical applicability in area of sports became popular when a renowned author Michael Lewis published Money ball in the year 2003. The book explained how the athletics of Oakland applied an analytical approach for evaluating players for assembling a competitive team with a limited budget. Analytics is used for evaluation of on-field strategy which is a common thing in professional sports. Analytics is also used in off-the-field decisions to ensure customer satisfaction.

Example : Professional sports teams utilize analytics for assessing players for the amateur drafts and for decision making of contract negotiations offered to the players. Various franchises across many major sports utilize prescriptive analytics for adjusting the ticket prices throughout the season for reflecting the potential demand and relative attractiveness for every game.

8. Web Analytics

The analysis of internet activity including visits of users to social media sites like Linkedln and Facebook and other websites is called web analytics. It plays a vital role in sales or promotions of products and services. Through internet various leading companies utilize advanced and descriptive analytics by applying them to the data gathered from online experiments for determining the best way for configuration of ads, websites and proper utilization of social networks to promote products and services. Online experimentation is exposing various group of visitors to varied versions of a website and then tracking the results. These experiments can be conducted without risking the overall business disruption of the company due to several number of internet users. But these experiments have proven to be invaluable as they enable the company for using trial-and-error method of determine statistically the reasons for differences in the sales and website traffic.

Q6. Explain the various challenges in business analytics.

Ans :

- **Executive Ownership:** Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
- **IT Involvement:** Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
- **Available Production Data vs. Cleansed Modeling Data :** Watch for technology infrastructure that restrict available data for historical modeling, and know the difference between historical data for model development and real-time data in production.
- **Project Management Office (PMO)** : The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.
- **End user Involvement and Buy-In :** End users should be involved in adopting Business Analytics and have a stake in the predictive model.
- **Change Management :** Organizations should be prepared for the changes that Business Analytics bring to current business and technology operations.
- **Explain ability vs. the "Perfect Lift"**: Balance building precise statistical models with being able to explain the model and how it will produce results.

1.3 Business Analytics in Practice

Q7. Explain the role of Business Analytics in Best Practices.

Ans :

Adopting and implementing Business Analytics is not something a company can do overnight. But, if a company follows some best practices for Business Analytics, they will get the levels of insight they seek and become more competitive and successful. We list some of the most important best practices for Business Analytics here, though your organization will need to determine which best practices are most fitting for there needs.

- Know the objective for using Business Analytics. Define the business use case and the goal ahead of time.
- Define the criteria for success and failure.
- Select the methodology and be sure to know the data and relevant internal and external factors
- Validate models using to predefined success and failure criteria

Business Analytics is critical for remaining competitive and achieving success. When they get BA best practices in place and get buy-in from all stakeholders, the organization will benefit from data-driven decision making.

1.4 Big Data - Overview of Data

Q8. What is Big Data?

Ans :

- Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too bigor it moves too fast or it exceeds current processing capacity.
- Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

Technologies in big data

We can categories them into two (storage and Querying/Analysis).

- Apache Hadoop. Apache Hadoop is a Java based free software framework that can effectively store large amount of data in a cluster. ...
- Microsoft HDInsight. ...
- NoSQL. ...
- Hive. ...
- Sqoop. ...

- PolyBase. ...
- Big data in EXCEL. ...
- Presto
- Some of the most common of those big data challenges include the following:
- Dealing with data growth. ...
- Generating insights in a timely manner. ...
- Recruiting and retaining big data talent. ...
- Integrating disparate data sources. ...
- Validating data. ...
- Securing big data. ...
- Organizational resistance

Q9. Explain the evolution of Big Data.

Ans :

The evolution of big data is discussed below,

(i) 1970s and before

(ii) 1980s and 1990s

(iii) 2000s and beyond

(i) 1970s and before : The data generation and storage of 1970s and before is fundamentally primitive and structured. This era is termed as the era of mainframes, as it stores the basic data.

(ii) 1980s and 1990s : In 1980s and 1990s the evolution of relational data bases took please. The relational data utilization is complex and thus this era comprises of data intensive applications.

(iii) 2000s and beyond : The World Wide Web (www) and the Internet of Things (IOT) have an aggression of structured, unstructured and multimedia data. The data driven is complex and unstructured.

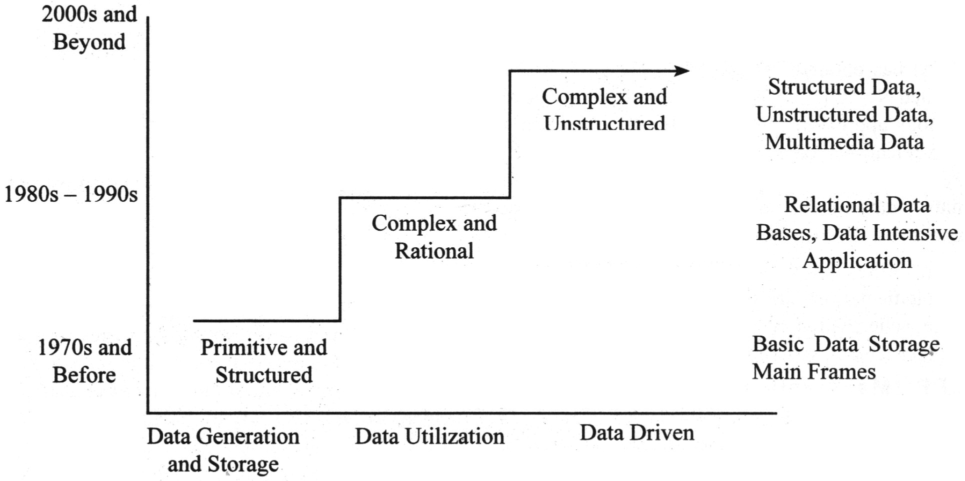

2000s and Beyond
1980s – 1990s
1970s and Before
Primitive and Structured
Complex and Rational
Complex and Unstructured
Structured Data, Unstructured Data, Multimedia Data
Relational Data Bases, Data Intensive Application
Basic Data Storage Main Frames
Data Generation and Storage
Data Utilization
Data Driven

Fig.: The Evolution of Big Data

Q10. Explain the dimensions of big data?

Ans :

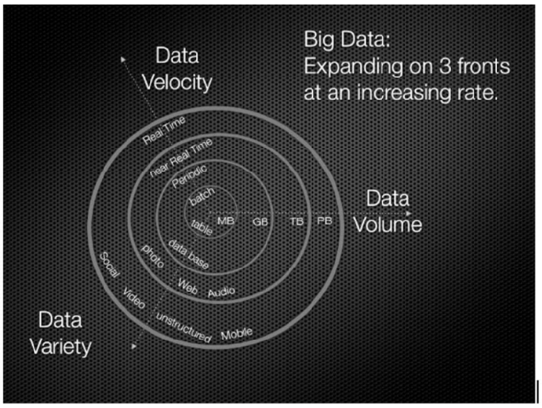
Big data refers to datasets whose size is beyond the ability of typical database software tool to capture, store, managed and analyze. Big data is data that goes beyond the traditional limits of data along four dimensions:

i) Data Volume

ii) Data Variety

iii) Data Velocity

iv) Variability



i) **Data Volume:** Data Volume can be measured by quality of transactions, events and amount of history. Big Data isn't just a description of raw volume. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

ii) **Data Variety :** It is the assortment of data. Traditionally data, especially operational data, is structured as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Wide variety of data: Internet data (Social media, Social Network-Twitter, Face book),

iii) **Data Velocity** : It is the measure of how fast the data is coming in. Remember our Facebook example. 250 billion images may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day. So that 250 billion number from last year will seem like a drop in the bucket in a few months. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

iv) **Variability** : The increase in the range of values typical of a large data set - and value, which addresses the need for valuation of enterprise data.

Q11. Explain the relationship of big data with other areas?

Ans :

Big data models Improve Operational Efficiencies Increase Revenues Achieve Competitive Differentiation Reduce risks and costs Sell to microtrends Offer new services Save time Enable self service Seize market share Lower complexity Improve customer experience Incubate new ventures Enable self service Detect fraud

I) Digital Marketing.

II) Financial Services

III) Big data and Advances in health care

IV) Advertising

I) Digital Marketing

Introduction Database Marketers, Pioneers of Big Data Big Data & New School of Marketing Cross Channel Life cycle Marketing Social and Affiliate Marketing

Empowering marketing with Social Intelligence Introduction Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want.

Digital \marketing is easy when consumers interact with corporate‘ primary platform (i.e.. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (e.g., Face book, Twitter, Google +). One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (i.e., There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions. Big data on the web will completely transform a company's ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organi- zations will be able to make hundreds and thousands of smart decisions every day.

II) Financial Services

Fraud & Big Data - Fraud is intentional deception made for personal gain or to damage another individual. - One of the most common forms of fraudulent activity is credit card fraud. - Social media and mobile phones are forming new frontiers fraud. - Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the following three Vs : 1. High volume: Years of consumer records and transactions (150 billion + 2. records per year). 3. High velocity: Dynamic transactions and social media info. 4. High variety: Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

III) Big data and Healthcare

Big data promises enormous revolution in healthcare, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine. In addition to saving and improving lives, Big Data has the potential to transform the entire healthcare system by replacing guesswork and institution with objective data-driven science.-The healthcare industry now has huge amount of data: from biological data such as gene expression, Special Needs Plans

(SNPs), proteomics, metabolomics, and next-generation gene sequence data etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and Health Information Exchanges (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices. - In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science

IV) Advertising and Big Data

Big Data is changing the way advertisers address three related needs. (i) How much to spend on advertisements. (ii) How to allocate amount across all the marketing communication touch points. (iii) How to optimize advertising effectiveness. Given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction. Reach, Resonance, and Reaction Reach: First part of reach is to identify the people who are most volumetrically responsive to their advertising and then answer questions such as what do those people watch? What do they do online? How to develop media plan against intended audience. The second part of reach is delivering advertisements to the right audience. That is, to understand if we are actually reaching our desired audience. If we think about the online world, it’s a world where we can deliver 100 million impressions but we never really know for sure who our campaign was actually delivered to. If our intended audience is women aged 18 to 35, of our 100 million impressions, what percentage of impressions were actually delivered to the intended audience? What was the reach, what was the frequency, what was the delivery against the intended audience? Resonance: If we know whom we want to reach and we’re reaching them efficiently with your media spend, the next question is, are our ads breaking through? Do people know they’re from our brand? Are they changing attitudes? Are they making consumers more likely to want to buy our brand? This is what is called resonance. Reaction: Advertising must drive a behavioral reaction or it isn’t really working. We have to measure the actual behavioral impact.

Q12. What are the categories which come under Big Data?

Ans :

Big data works on the data produced by various devices and their applications. Below are some of the fields that are involved in the umbrella of Big Data.

1. **Black Box Data:** It is an incorporated by flight crafts, which stores a large sum of information, which includes the conversation between crew members and any other communications (alert messages or any order passed) by the technical grounds duty staff.
2. **Social Media Data:** Social networking sites such as Face book and Twitter contains the information and the views posted by millions of people across the globe.
3. **Stock Exchange Data:** It holds information (complete details of in and out of business transactions) about the 'buyer' and 'seller' decisions in terms of share between different companies made by the customers.
4. **Power Grid Data:** The power grid data mainly holds the information consumed by a particular node in terms of base station.
5. **Transport Data:** It includes the data's from various transport sectors such as model, capacity, distance and availability of a vehicle.
6. **Search Engine Data:** Search engines retrieve a large amount of data from different sources of database.

Q13. Explain the importance of big data.

Ans :

The importance of big data is how you utilize the data which you own. Data can be fetched from any source and analyze it to solve that enable us in terms of

1. Cost reductions
2. Time reductions,
3. New product development and optimized offerings, and
4. Smart decision making.

Advantage of Big data Business Models:

Big data models Improve Operational Efficiencies Increase Revenues Achieve Competitive Differentiation Reduce risks and costs Sell to microtrends Offer new services Save time Enable self service Seize market share Lower complexity Improve customer experience Incubate new ventures Enable self service Detect fraud.

Industry Examples of Big data:

- Digital Marketing.
- Financial Services
- Big data and Advances in health care
- Pioneering New Frontiers in medicine
- Advertising

Industry Examples of Big Data

I) Digital Marketing

Introduction Database Marketers, Pioneers of Big Data Big Data & New School of Marketing Cross Channel Life cycle Marketing Social and Affiliate Marketing Empowering marketing with Social Intelligence Introduction Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want. Digital \marketing is easy when consumers interact with corporate‘ primary platform (i.e., The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (e.g., Face book, Twitter, Google +). One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (i.e., There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions. Big data on the web will completely transform a company’s ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organizations will be able to make hundreds and thousands of smart decisions every da

II) Financial Services

i) Fraud & Big Data - Fraud is intentional deception made for personal gain or to damage another individual. - One of the most common forms of fraudulent activity is credit card fraud. - Social media and mobile phones are forming new frontiers fraud. - Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the following three Vs : 1. High volume: Years of consumer records and transactions (150 billion + 2. records per year). 3. High velocity: Dynamic transactions and social media info. 4. High variety: Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

III) Big data and Healthcare

Big data promises enormous revolution in healthcare, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine. - In addition to saving and improving lives, Big Data has the potential to transform the entire healthcare system by replacing guesswork and institution with objective data-driven science. - The healthcare industry now has huge amount of data: from biological data such as gene expression, Special Needs Plans (SNPs), proteomics, metabolomics, and next-generation gene sequence data etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and Health Information Exchanges (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices. - In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science

IV) Advertising and Big Data

Big Data is changing the way advertisers address three related needs. (i) How much to spend on advertisements. (ii) How to allocate amount across all the marketing communication touch points. (iii) How to optimize advertising effectiveness. Given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction. Reach, Resonance, and Reaction Reach: First part of reach is to identify the people who are most volumetrically responsive to their advertising and then answer questions such as what do those people watch? What do they do online? How to develop media plan against intended audience. The second part of reach is delivering advertisements to the right audience. That is, to understand if we are actually reaching our desired audience. If we think about the online world, it’s a world where we can deliver 100 million impressions but we never really know for sure who our campaign was actually delivered to. If our intended audience is women aged 18 to 35, of our 100 million impressions, what percentage of impressions were actually delivered to the intended audience? What was the reach, what was the frequency, what was the delivery against the intended audience? Resonance: If we know whom we want to reach and we're reaching them efficiently with your media spend, the next question is, are our ads breaking through? Do people know they're from our brand? Are they changing attitudes? Are they making consumers more likely to want to buy our brand? This is what is called resonance. Reaction: Advertising must drive a behavioral reaction or it isn't really working. We have to measure the actual behavioral impact.

Q14. Explain the life cycle of big data.

Ans :

Big Data Life Cycle

In today's big data context, the previous approaches are either incomplete or suboptimal. For example, the SEMMA methodology disregards completely data collection and pre-processing of different data sources. These stages normally constitute most of the work in a successful big data project.

A Big Data Analytics Cycle can be described by the following stages:

1. Business Problem Definition
2. Research
3. Human Resources Assessment
4. Data Acquisition
5. Data Mugging
6. Data Storage
7. Exploratory Data Analysis
8. Data Preparation for Modeling and Assessment
9. Modeling
10. Implementation

1. Business Problem Definition

This is a point common in traditional **BI** and big data analytics lifecycle. Normally, it is a non-trivial stage of a big data project to define the problem and evaluate correctly how much potential gain it may have for an organization. It seems obvious to mention this, but it has to be evaluated what are the expected gains and costs of the project.

2. Research

Analyze what other companies have done in the same situation. This involves looking for solutions that are reasonable for your company, even though it involves adapting other solutions to the resources and requirements that your company has. In this stage, a methodology for the future stages should be defined.

3. Human Resources Assessment

Once the problem is defined, it is reasonable to continue analyzing if the current staff is able to complete the project successfully. Traditional **BI** teams might not be capable to deliver an optimal solution to all the stages. So, it should be considered before starting the project if there is a need to outsource a part of the project or hire more people.

4. Data Acquisition

This section is key in a big data life cycle; it defines which type of profiles would be needed to deliver the resultant data product. It is a non-trivial step of the process; it normally involves gathering unstructured data from different sources. To give an example, it could involve writing a crawler to retrieve reviews from a website. This involves dealing with text, perhaps in different languages normally requiring a significant amount of time to be completed.

5. Data Mugging

Once the data is retrieved, for example, from the web, it needs to be stored in an easy-to-use format. To continue with the reviews examples, let's assume the data is retrieved from different sites where each has a different display of the data.

Suppose one data source gives reviews in terms of rating in stars. Therefore, it is possible to read this as a mapping for the response variable $y \in \{1, 2, 3, 4, 5\}$. Another data source gives reviews using two arrows system, one for up voting and the other for down voting. This would imply a response variable of the form $y \in$ {positive, negative}.

In order to combine both the data sources, a decision has to be made in order to make these two response representations equivalent. This can involve converting the first data source response representation to the second form, considering one star as negative and five stars as positive. This process often requires a large time allocation to be delivered with good quality.

6. Data Storage

Once the data is processed, it sometimes needs to be stored in **a** database. Big data technologies offer plenty of alternatives regarding this point. The most common alternative is using the Hadoop File System for storage that provides users a limited version of SQL, known as HTVE Query Language. This allows most analytics task to be done in similar ways as would be done in traditional BI data warehouses, from the user perspective. Other storage options to be considered are Mongo DB, Redis and SPARK.

This stage of the cycle is related to the human resources knowledge in terms of their abilities to implement different architectures. Modified versions of traditional data warehouses are still being used in large-scale applications. For example, Teradata and IBM offer SQL databases that can handle terabytes of data; open source solutions such as postgre SQL and MySQL are still being used for large-scale applications.

Even though there are differences in how the different storages work in the background, from the client side, most solutions provide a SQL API. Hence, having a good understanding of SQL is still a key skill to have for big data analytics.

This stage a priori seems to be the most important topic; in practice, this is not true. It is not even an essential stage. It is possible to implement a big data solution that would be working with real-time data. So, in this case, we only need to gather data to develop the model and then implement it in real time. So, there would not be a need to formally store the data at all.

7. Exploratory Data Analysis

Once the data has been cleaned and stored in a way that insights can be retrieved from it, the data exploration phase is mandatory. The objective of this stage is to understand the data. This is normally done with statistical techniques and also plotting the data. This is a good stage to evaluate whether the problem definition makes sense or is feasible.

8. Data Preparation for Modeling and Assessment

This stage involves reshaping the cleaned data retrieved previously and using statistical pre-processing for missing values imputation, outlier detection, normalization, feature extraction and feature selection.

9. Modelling:

The prior stage should have produced several data sets for training and testing, e.g., a predictive model. This stage involves trying different models and looking forward to solving the business problem at hand. In practice, it is normally desired that the model would give some insight into the business. Finally, the best model or combination of models is selected evaluating its performance on a left-out data set.

10. Implementation

In this stage, the data product developed is implemented in the data pipeline of the company. This involves setting up a validation scheme while the data product is working in order to track its performance. For example, in case of implementing a predictive model, this stage would involve applying the model to new data and once the response is available, evaluate the model.

Q15. Explain the users of big data?

Ans :

The people who are using Big Data know better that, what Big Data is. Let’s look at some such industries:

- **Healthcare:** Big Data has already started to create a huge difference in the healthcare sector. With the help of predictive analytics, medical professionals and HCPs are now able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring - all powered by Big Data and AI - are helping change lives for the better.

- **Academia:** Big Data is also helping enhance education today. Education is no more limited to the physical bounds of the classroom - there are numerous online educational courses to learn from. Academic institutions are investing in digital courses powered by Big Data technologies to aid the all-round development of budding learners.

- **Banking:** The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/ debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.
- **Manufacturing:** According to TCS 2013 Global Trend Study, the most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality. In the manufacturing sector, Big Data helps create a transparent infrastructure, thereby predicting uncertainties and incompetencies that can affect the business adversely.
- **IT:** One of the largest users of Big Data, IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity and minimize risks in business operations. By combining Big Data technologies with ML and AI, the IT sector is continually powering innovation to find solutions even for the most complex of problems.

1.5 Data - Types of Data

Q16. What is Data and Explain the various types of data ?

Ans :

- Data are basic values or facts. Computers use many different types of data stored in digital format, such as text, numbers and multimedia. Data are organized in database tables, and database management systems are used to work with large databases.
- Data is a set of values of subjects with respect to qualitative or quantitative variables. Data and information are often used interchangeably; however data becomes information when it is viewed in context or in post-analysis.

i) Structured

By structured data, we mean data that can be processed, stored and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc. will be present in an organized manner.

ii) Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data.

	Structured Data	Unstructured Data
Characteristics	➢ Pre-defined data models ➢ Usually text only ➢ Easy to search	➢ No pre-defined data model ➢ May be text, images, sound, video or other formats ➢ Difficult to search
Resides in	➢ Relational databases ➢ Data warehouses	➢ Applications ➢ NoSQL databases ➢ Data warehouses ➢ Data lakes
Generated by **Typical applications**	**Humans or Machines** ➢ Airline reservation systems ➢ Inventory control ➢ CRM systems ➢ ERP systems	**Humans or Machines** ➢ Word processing ➢ Presentation software ➢ Email clients ➢ Tools for viewing or editing media
Examples	➢ **Dates** ➢ Phone numbers ➢ Social security numbers ➢ Credit card numbers ➢ Customer names ➢ Addresses ➢ Product names and numbers ➢ Transaction information	➢ **Text Files** ➢ Presentation software ➢ Email messages ➢ Audio files ➢ Video files ➢ Images ➢ Surveillance imagery

iii) Semi-structured

Semi-structured data pertains to the data containing both the formats mentioned above, i.e., structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

Short Question and Answers

1. What is Business analytics?

Ans :

Business analytics is the practice of iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.

Definition of business analytics

According to Schaer (2018) - "allows your business to make predictive analysis rather than reacting to changes in data".

According to Gabelli School of Business (2018)- "involves applying models, methods, and tools to data, producing insights that lead to informed business decisions"

According to Wells (2008) - "the application of logic and mental processes to find meaning in data"

According to Lynda (2018) - "allows us to learn from the past and make better predictions for the future".

2. Different business analytical methods.

Ans :

i) Prescriptive: This type of analysis reveals what actions should be taken. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps.

ii) Predictive: An analysis of likely scenarios of what might happen. The deliverables are usually a predictive forecast.

iii) Diagnostic: A look at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.

iv) Descriptive: What is happening now based on incoming data. To mine the analytics, you typically use a real-time dashboard and/or email reports.

3. What is Big Data?

Ans :

- Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too bigor it moves too fast or it exceeds current processing capacity.
- Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

4. Dimensions of big data?

Ans :

i) Data Volume

Data Volume can be measured by quality of transactions, events and amount of history. Big Data isn't just a description of raw volume. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

ii) Data Variety

It is the assortment of data. Traditionally data, especially operational data, is structured as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Wide variety of data: Internet data(Social media ,Social Network-Twitter, Face book),

iii) Data Velocity

It is the measure of how fast the data is coming in. Remember our Facebook example. 250 billion images may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day. So that 250 billion number from last year will seem like a drop in the bucket in a few months. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

iv) Variability

The increase in the range of values typical of a large data set - and value, which addresses the need for valuation of enterprise data.

5. Various Challenges in Business Analytics

Ans :

- **Executive Ownership:** Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
- **IT Involvement:** Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
- **Available Production Data vs. Cleansed Modeling Data :** Watch for technology infrastructure that restrict available data for historical modeling, and know the difference between historical data for model development and real-time data in production.
- **Project Management Office (PMO)** : The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.
- **End user Involvement and Buy-In :** End users should be involved in adopting Business Analytics and have a stake in the predictive model.
- **Change Management :** Organizations should be prepared for the changes that Business Analytics bring to current business and technology operations.
- **Explainability vs. the "Perfect Lift"**: Balance building precise statistical models with being able to explain the model and how it will produce results.

6. Digital Marketing

Ans :

Introduction Database Marketers, Pioneers of Big Data Big Data & New School of Marketing Cross Channel Life cycle Marketing Social and Affiliate Marketing Empowering marketing with Social Intelligence Introduction Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want.

Digital \marketing is easy when consumers interact with corporate‘ primary platform (ie. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (eg. Face book, Twitter, Google +). One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (ie. There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions. Big data on the web will completely transform a company’s ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organizations will be able to make hundreds and thousands of smart decisions every day.

Choose the Correct Answer

1. Facebook Tackles Big Data with ________ based on Hadoop. [a]

   (a) Project Prism (b) Prism

   (c) Project Data (d) Project Bid

2. All of the following accurately describe Hadoop, EXCEPT: [b]

   (a) Open Source (b) Real-time

   (c) Java-based (d) Distributed computing approach

3. What are the main components ________ of Big Data? [d]

   (a) Map Reduce (b) HDFS

   (c) YARN (d) All of these

4. ________ has the world's largest Hadoop cluster. [c]

   (a) Apple (b) Datamatics

   (c) Facebook (d) None of the mentioned

5. According to analysts, for what can traditional IT systems provide a foundation when they are integrated with Big Data technologies like Hadoop? [a]

   (a) Big Data management and data mining

   (b) Data warehousing and business intelligence

   (c) Management of Hadoop clusters

   (d) Collecting and storing unstructured data

6. What are the five V's of Big Data? [d]

   (a) Volume (b) Velocity

   (c) Variety (d) All the above

7. What are the different features of Big Data Analytics? [d]

   (a) Open Source (b) Scalability

   (c) Data Recovery (d) All the above

8. __________ Data refers to the data that lacks any specific form [b]

(a) Structured data (b) Unstructured data

(c) Both (d) None of the above

9. __________ is the last stage is Big data life cycle. [a]

(a) Implementation (b) Datastorage

(c) Data Mugging (d) Research

10. __________ analyze what other companies have done in the same situations. [d]

(a) Implementation (b) Datastorage

(c) Data Mugging (d) Research

Fill in the blanks

1. ____________ refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.

2. ____________ tools range from spreadsheets with statistical functions to complex data mining and predictive modeling applications.

3. ____________ analytics it is really valuable, but largely not used.

4. An ____________ model is simply a mathematical equation that describes relationships among variables in a historical data set.

5. ____________ analytics can be useful in the sales cycle,

6. ____________ (or feature) roadmaps outline the development and launches of a product and its features.

7. ____________ outline the hierarchy of a business or one of its departments or teams.

8. PFD stands for ____________.

9. ____________ can be measured by quality of transactions, events and amount of history.

10. ____________ models Improve Operational Efficiencies Increase Revenues Achieve Competitive Differentiation Reduce risks and costs

Answers

1. Business analytics (BA)
2. Analytics
3. Prescriptive
4. Analytical
5. Descriptive
6. Product
7. Organizational charts
8. Process flow Diagram

9 Data Volume

10. Big data

UNIT II

DESCRIPTIVE ANALYTICS

Over view of Description Statistics (Central Tendency, Variability), Data Visualization-Definition, Visualization Techniques - Tables, Cross Tabulations, charts, Data Dash boards using Ms-Excel or SPSS.

2.1 Overview of Description Statistics

Q1. What is Statistics

Ans :

Statistics is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.

According to Prof Horace Secrist: Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.

Descriptive statistics employs a set of procedures that make it possible to meaningfully and accurately summarize and describe samples of data. In order for one to make meaningful statements about psychological events, the variable or variables involved must be organized, measured, and then expressed as quantities. Such measurements are often expressed as measures of central tendency and measures of variability.

Q2. Explain briefly about Descriptive Statistics?

Ans :

Descriptive Statistics

Descriptive statistics is used to summarize data and make sense out of the raw data collected during the research. Since the data usually represents a sample, then the descriptive statistics is a quantitative description of the sample.

The level of measurement of the data affects the type of descriptive statistics. Nominal and ordinal type data (often termed together as categorical type data) will differ in the analysis from interval and ratio type data (often termed together as continuous type data).

Descriptive statistics for categorical data

Contingency tables (or frequency tables) are used to tabulate categorical data. A contingency table shows a matrix or table between independent variables at the top row versus a dependent variable on the left column, with the cells indicating the frequency of occurrence of possible combination of levels. (check SPSS for examples).

Descriptive statistics for continuous data

There are two the two aspects of descriptive statistics used for continuous type data. They are;

- Central tendency
- Variability of the data

2.1.1 Measures of Central Tendency

Q3. Explain briefly about Measures of Central Tendency ?

Ans :

It refers to a number (statistic) that best characterizes the group as a whole" (Sommer & Sommer, 1997). It is generally referred to as the average. The three measures of central tendency, the mean, median, and mode, describe a distribution of data and are an index of the average, or typical, value of a distribution of scores

The three types of averages are:

1. Mean
2. Median
3. Mode

1. MEAN (M)

It is the arithmetic average (sum of all score divided by the number of cases) The mean, the arithmetic average of all scores under consideration, is computed by dividing the sum of the scores by the number of scores.

The sample mean of the values

2. MEDIAN

It is the midpoint of a distribution of data. Half the scores fall above and half below the median. The three measures of central tendency, the mean, median, and mode, describe a distribution of data and are an index of the average, or typical, value of a distribution of scores.

The median is the point at which 50% of the observations fall below and 50% above or, in other words, the middle number of a set of numbers arranged in ascending or descending order. (If the list includes an even number of categories, the median is the arithmetic average of the middle two numbers.) Based on the data in Table , the full list of each student's study hours would be written 10, 9, 9, 9, 8, 8, 8, 8, and so on. If the list were written out in full, it would be clear that the middle two numbers of the 40 entries are 6 and 6, which average 6. So the median of the hours studied is 6.

3. MODE

It is the single score that occurs most often in a distribution of data. The mode is the number that appears most often. Based on the data in Table , the mode of the number of hours studied is also 6 (8 students studied for 6 hours, so 6 appears 8 times in the list, more than any other number).

2.1.2 Measures of Variability

Q4. Explain the various ways of Measure of variability?

Ans :

There are many ways to describe variability including :

(i) Range

(ii) Interquartile Range (IQR)

(iii) Variance

(iv) Standard Deviation

(i) Range

Range = Maximum – Minimum

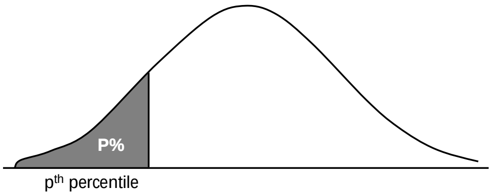
(a) Easy to calculate

(b) Very much affected by extreme values (ranges is not a resistant measure of variability).

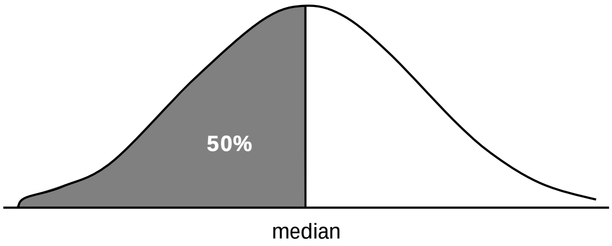
(ii) Interquartile Range (IQR)

In order to talk about interquartile range, we need to first talk about percentiles.

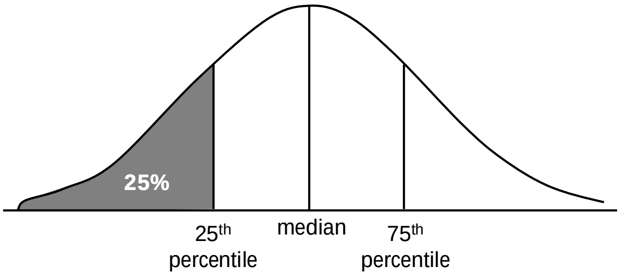
The path percentile of the data set is a measurement such that after the data are ordered from smallest to largest, at most, p% of the data are at or below this value and at most, (100 – p)% at or above it.



Thus, the median is the 50th percentile. Fifty percent or the data values fall at or below the median.



Also, Q_1 = lower quartile = the 25th percentile and Q_3 = upper quartile = the 75th percentile.



Interquartile Range

It is the difference between upper and lower quartiles and denoted as IQR.

IQR = $Q_3 - Q_1$ = upper quartile - lower quartile = 75th percentile - 25th percentile.

Details about how to compute IQR will be given in Lesson 2.3.

Note: IQR is not affected by extreme values. It is thus a resistant measure of variability.

(iii) Variance

Two vending machines A and B drop candies when a quarter is inserted. The number of pieces of candy one gets is random. The following data are recorded for six trials at each vending machine:

Pieces of candy from vending machine A:

1, 2, 3, 3, 5, 4

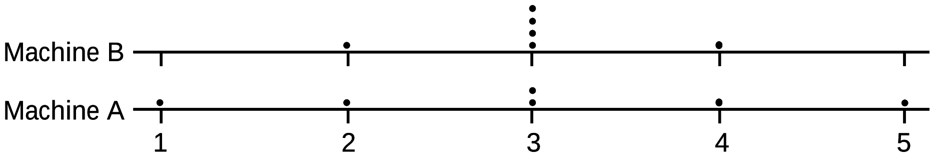
Mean = 3, Median = 3, Mode = 3

Pieces of candy from vending machine B:

2, 3, 3, 3, 3, 4

Mean = 3, Median = 3, Mode = 3

Dotplots for the pieces of candy from vending machine A and vending machine B:



They have the same center, but what about their spreads? One way to compare their spreads is to compute their standard deviations. In the following section, we are going to talk about how to compute the sample variance and the sample standard deviation for a data set.

Variance is the average squared distance from the mean.

Population variance is defined as:

$\alpha^2 = \Sigma i = 1N (yi - \mu) / 2N$

In this formula μ is the population mean and the summation is over all possible values of the population. N is the population size.

The sample variance that is computed from the sample and used to estimate α^2 is:

$s2 = \Sigma i = 1n (yi - \bar{y})2n - 1$

Why do we divide by n – 1 instead of by n? Since μ is unknown and estimated by $\bar{y}$, the y_i's tend to be closer to $\bar{y}$ than to μ. To compensate, we divide by a smaller number, n – 1.

Sample Variance

It is the common default calculations used by software. When asked to calculate the variance or standard deviation of a set of data, assume - unless otherwise instructed - this is sample data and therefore calculating the sample variance and sample standard deviation.

Examples

Let's find S2 for the data set from vending machine A: 1, 2, 3, 3, 4, 5

$\bar{y}$ = 1 + 2 + 3 + 3 + 4 + 56 = 3

s_2 = (y1 − $\bar{y}$)2 + +(yn − $\bar{y}$) 2n − 1

= (1 − 3)2 + (2 − 3)2 + (3 − 3)2 +(3 − 3)2 + (4 − 3)2 + (5 − 3) 26

−1 = 2

Calculate S^2 for the data set from vending machine B yourself and check that it is smaller than the S^2 for data set A. Work out your answer first, then click the graphic to compare answers.

(iv) Standard Deviation

The population standard deviation is notated by σ and found by $\sigma = \sigma^2 - \surd$ has the same unit as y_i's. This is a desirable property since one may think about the spread in terms of the original unit.

σ is estimated by the sample standard deviation *s* :

s = *s*2 − √

For the data set *A*,

s = 2 − Š = 1.414 pieces of candy.

2.2 Data Visualization

Q5. What is Data visualization? Explain the importance of Data Visualization ?

Ans :

It is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

- Today's data visualization tools go beyond the standard charts and graphs used in Microsoft Excel spreadsheets, displaying data in more sophisticated ways such as info graphics, dials and gauges, geographic maps, spark lines, heat maps, and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

Importance of data visualization

- Data visualization has become the de facto standard for modern business intelligence (BI). The success of the two leading vendors in the BI space, Tableau and Qlike - both of which heavily emphasize visualization - has moved other vendors toward a more visual approach in their software. Virtually all BI software has strong data visualization functionality.
- Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. They are typically easier to operate than traditional statistical analysis software or earlier versions of BI software. This has led to a rise in lines of business implementing data visualization tools on their own, without support from IT.
- Data visualization software also plays an important role in big data and advanced analytics projects. As businesses accumulated massive troves of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.
- Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

Examples of data visualization

Data visualization tools can be used in a variety of ways. The most common use today is as a BI reporting tool. Users can set up visualization tools to generate automatic dash boards that track company performance across key performance indicators and visually interpret the results.

Many business departments implement data visualization software to track their own initiatives. For example, a marketing team might implement the software to monitor the performance of an email campaign, tracking metrics like open rate, click-through rate and conversion rate.

Q6. How data visualization works?

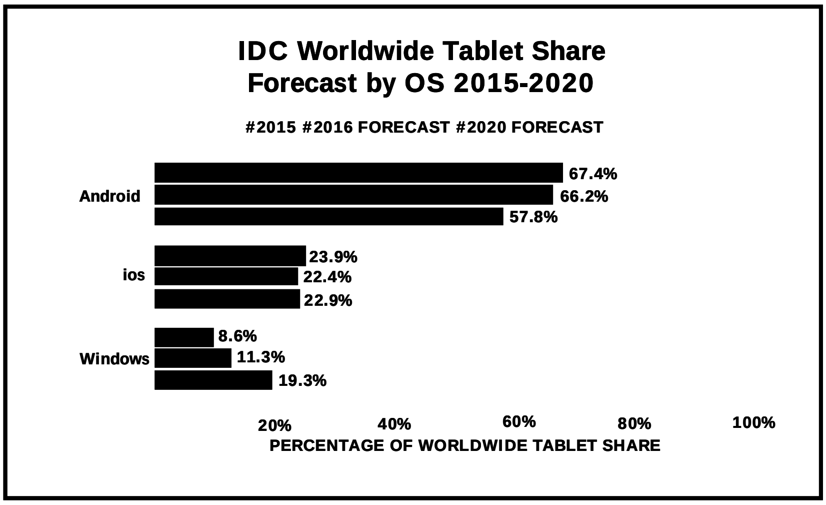
Ans :

Most of today's data visualization tools come with connectors to popular data sources, including the most common relational databases, Hadoop and a variety of cloud storage platforms. The visualization software pulls in data from these sources and applies a graphic type to the data.

Data visualization software allows the user to select the best way of presenting the data, but, increasingly, software automates this step. Some tools automatically interpret the shape of the data and detect correlations between certain variables and then place these discoveries into the chart type that the software determines is optimal.

Typically, data visualization software has a dashboard component that allows users to pull multiple visualizations of analyses into a single interface, generally a web portal.

Below is a chart forecasting tablet sales by operating system.



Q7. Explain the uses of data visualization.

Ans :

- By using data visualization, it became easier for business owners to understand their large data in a simple format.
- The visualization method is also time saving. So, businesses does not have to spend much time to make a report or solve a query. They can easily do it in a less time and in a more appealing way.
- Visual analytics offers a story to the viewers. By using charts and graphs or images, a person can easily exposure the whole concept as well the viewers will be able to understand the whole thing in an easy way.
- The most complicated data will look easy when it gets through the process of visualization. Complicated data report gets converted into a simple format. And it helps people to understand the concept in an easy way.
- With the visualization process, it gets easier to the business owners to understand their product growth and market competition in a better way.

2.3 Data Visualization Techniques

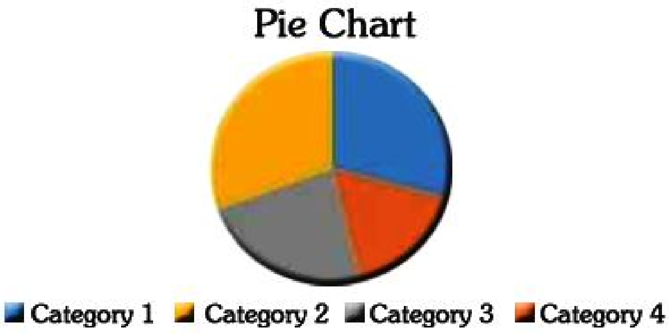
Q8. What are the techniques of Data Visualization?

Ans :

The data visualization techniques are Diagrams, charts, graphs.

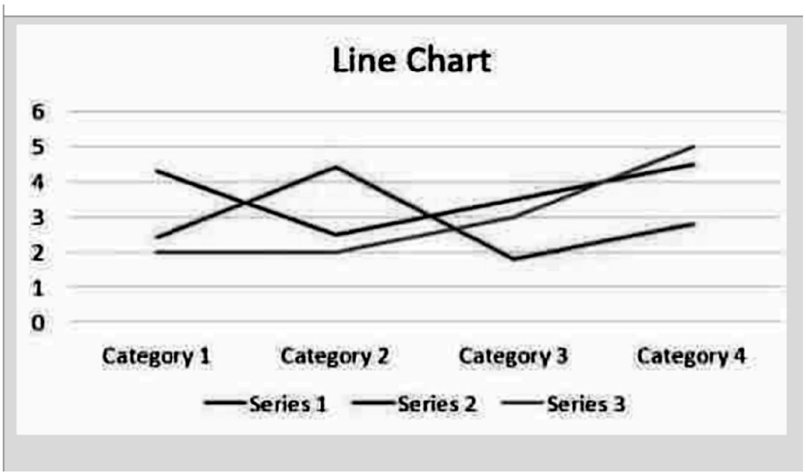
Most widely used forms of data visualization are presented below:

1. Pie Chart



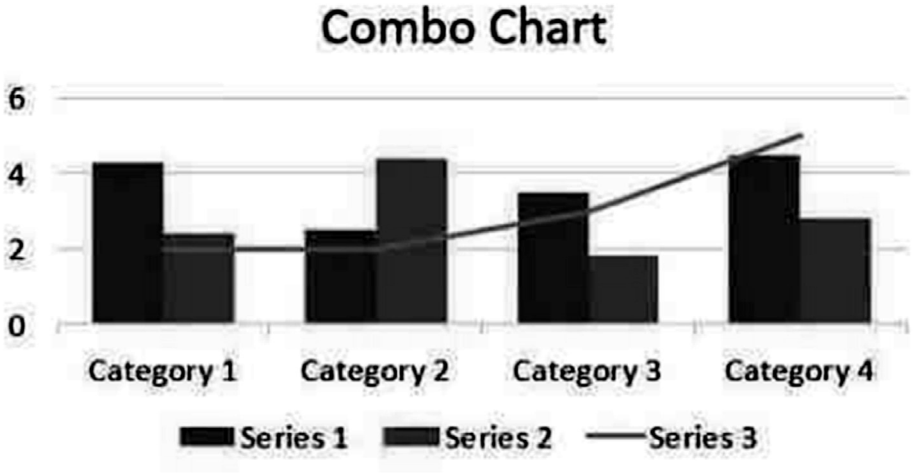
Pie Charts : Pie Charts are one of the common popular techniques. It also comes under data visualization techniques in excel. However, to some people, it can be hard to understand the chart while comparing to the line and bar type chart.

2. Line Chart



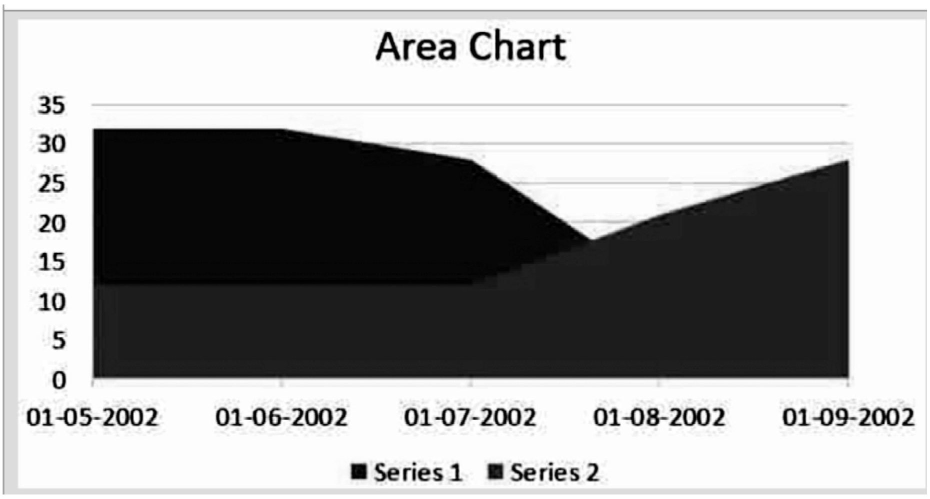
To make your data simple and more appealing you can simply use the line charts technique. Line chart basically displays the relationship between two patterns. Also, it is one of the most used techniques world wide.

3. Combo Chart



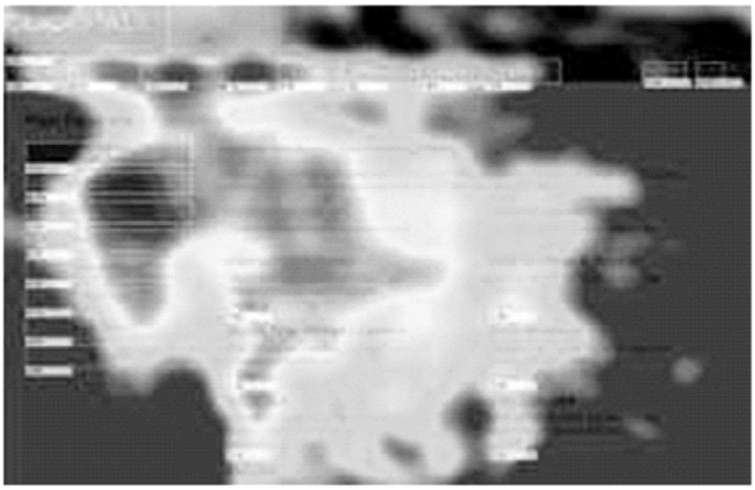
Bars charts are also one of the most commonly used techniques when it comes to comparing two different patterns. The bar charts can display the data in a horizontal way or in a vertical way. It all depends on your needs.

4. Area Chart



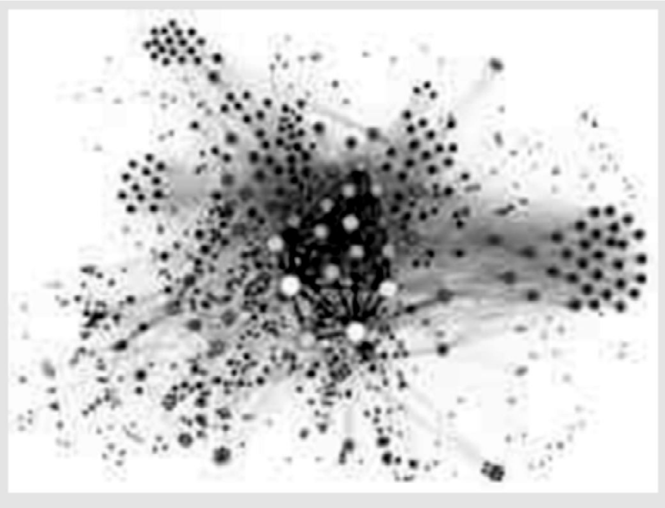
An area chart or area graph is similar to a line chart but provides graphically quantitative data. The areas can be filled with colour, hatch, pattern. This chart is generally used when comparing quantities which is depicted by area.

5. Heat Map



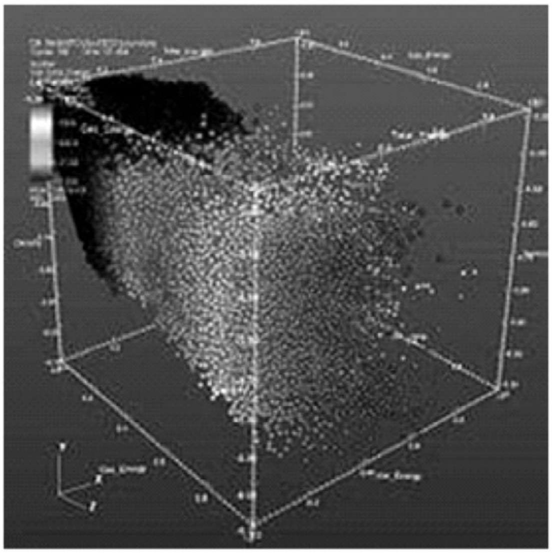
This type of chart is widely used by websites, mobile application makers, research institutes etc. These maps shows the concentration of activity/entity over a particular area.

6. Network Diagram



This is a powerful tool for finding out connections & correlations. It highlights and bridges the gaps. Shows strongly one activity is connected to other.

7. Scattered 3 D Plot



As the image shows it shows the distribution of entity in a 3 dimensional nature. It can be considered as showing location and concentration of gases in a box with different colours assigned to each gas.

2.3.1 Tables

Q9. How "data visualization technique –Tables" can display data analysis reports using Ms.Excel?

Ans :

Data analysis reports using Ms.Excel can display in a number of ways. However, if the data analysis results can be visualized as charts that highlight the notable points in the data, the audience can quickly grasp what they want to project in the data. It also leaves a good impact on the presentation style.

Here you will get to know how to use Excel charts and Excel formatting features on charts that enable you to present your data analysis results with emphasis.

Visualizing Data with Charts

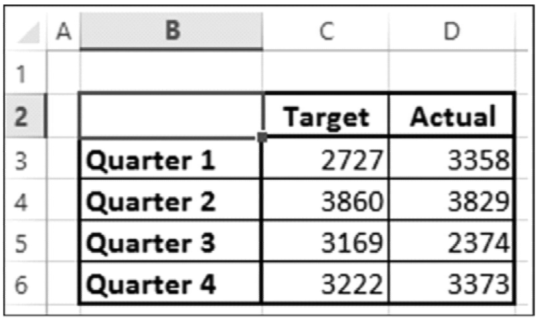
In Excel, charts are used to make a graphical representation of any set of data. A chart is a visual representation of the data, in which the data is represented by symbols such as bars in a Bar Chart or lines in a Line Chart. Excel provides you with many chart types and you can choose one that suits your data or you can use the Excel Recommended Charts option to view charts customized to your data and select one of those.

Refer to the Tutorial Excel Charts for more information on chart types.

In this chapter, you will understand the different techniques that you can use with the Excel charts to highlight your data analysis results more effectively.

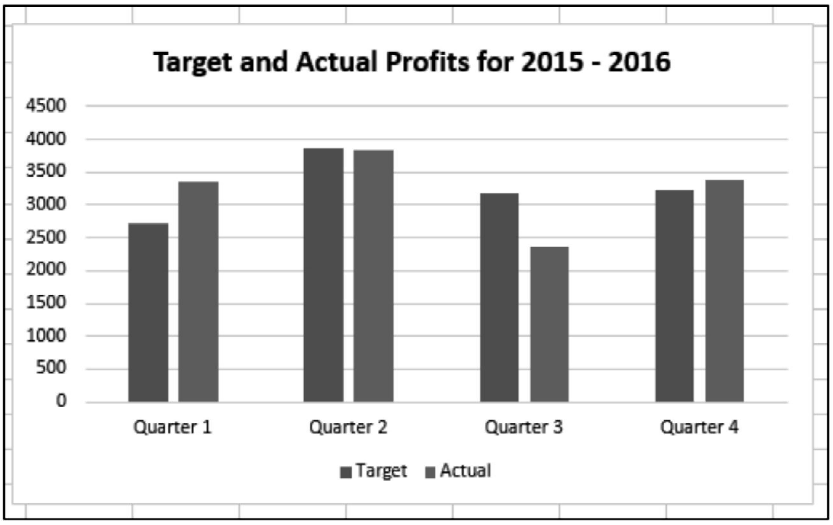
Creating Combination Charts

Suppose you have the target and actual profits for the fiscal year 2015-2016 that you obtained from different regions.



	A	B	C	D
1				
2			Target	Actual
3		Quarter 1	2727	3358
4		Quarter 2	3860	3829
5		Quarter 3	3169	2374
6		Quarter 4	3222	3373

We will create a Clustered Column Chart for these results.

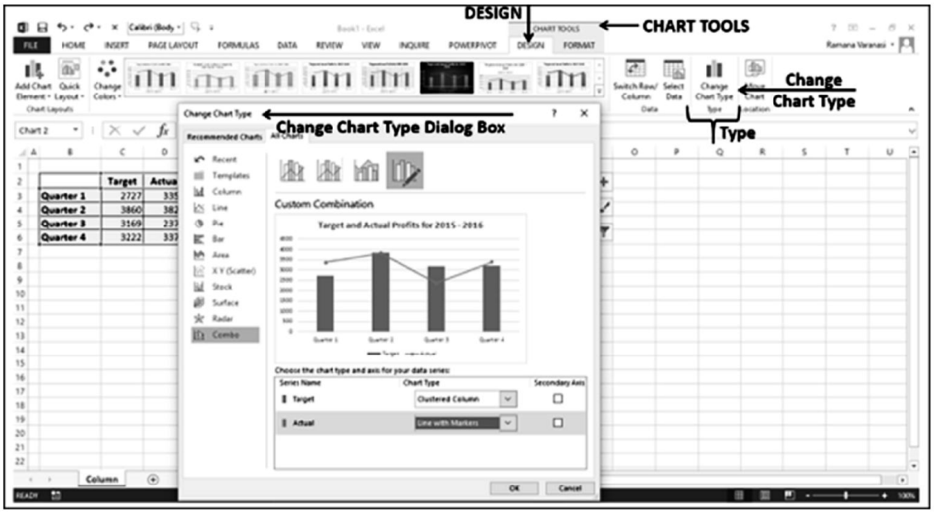


As you observe, it is difficult to visualize the comparison quickly between the targets and actual in this chart. It does not give a true impact on your results.

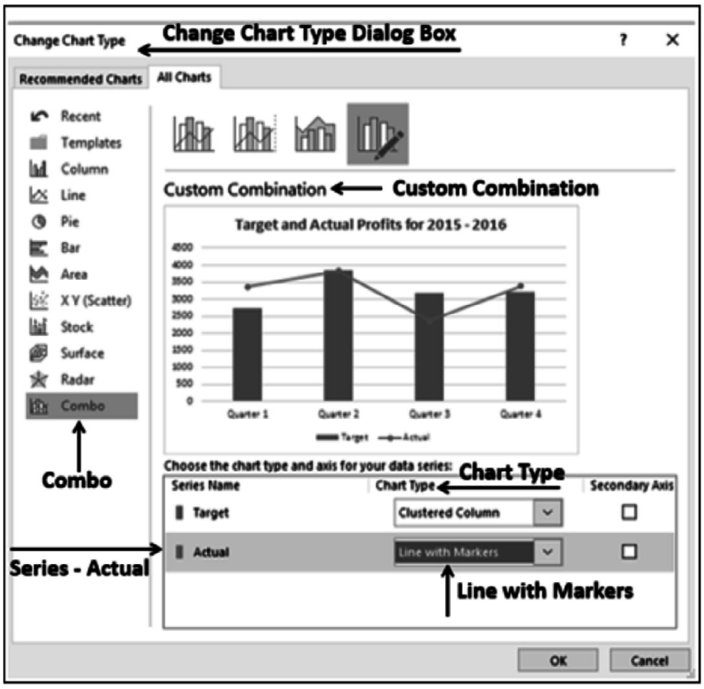
A better way of distinguishing two types of data to compare the values is by using Combination Charts. In Excel 2013 and versions above, you can use Combo charts for the same purpose.

Use Vertical Columns for the target values and a Line with Markers for the actual values.

- Click the DESIGN tab under the CHART TOOLS tab on the Ribbon.
- Click Change Chart Type in the Type group. The Change Chart Type dialog box appears.


DESIGN
CHART TOOLS
Change Chart Type
Type
Change Chart Type Dialog Box
Custom Combination
Target and Actual Profits for 2015 - 2016

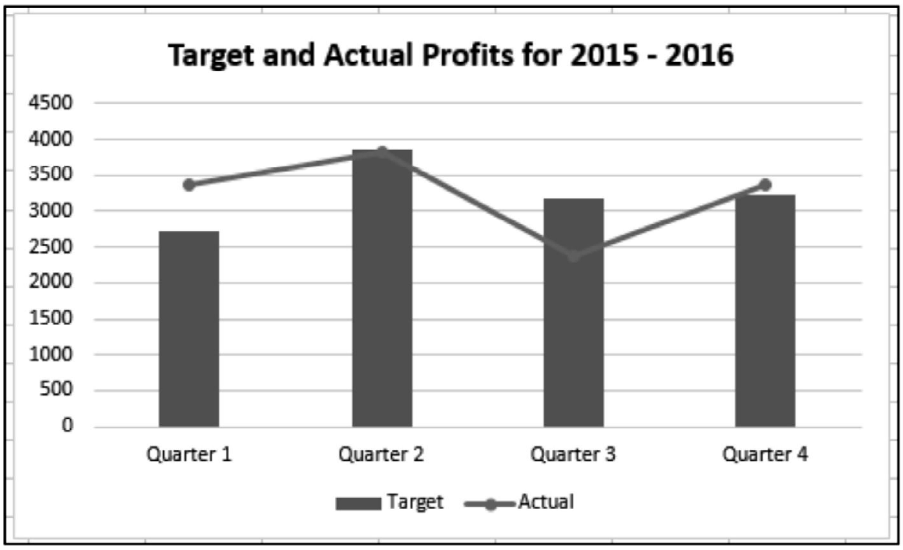
- Click Combo.
- Change the Chart Type for the series Actual to Line with Markers. The preview appears under Custom Combination.
- Click OK.



Choose the chart type and axis for your data series:

Series Name	Chart Type	Secondary Axis
Target	Clustered Column	□
Actual	Line with Markers	□

Your Customized Combination Chart will be displayed.

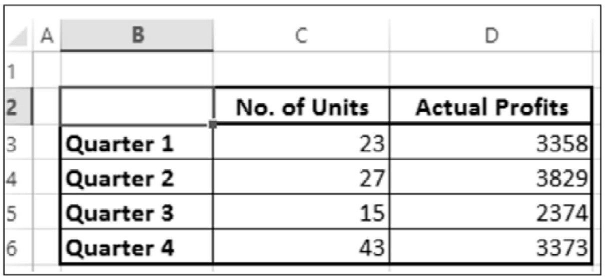


As you observe in the chart, the Target values are in Columns and the Actual values are marked along the line. The data visualization has become better as it also shows you the trend of your results.

However, this type of representation does not work well when the data ranges of your two data values vary significantly.

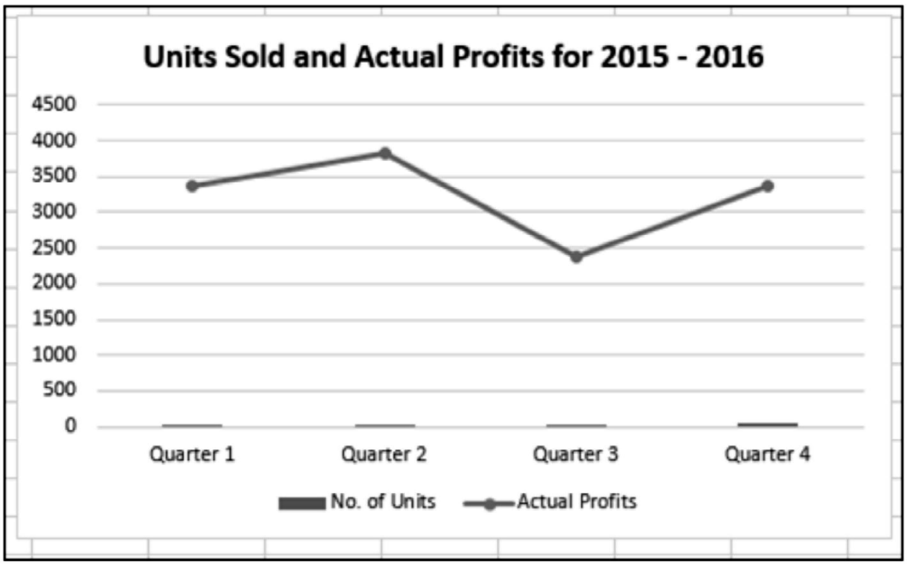
Creating a Combo Chart with Secondary Axis

Suppose you have the data on the number of units of your product that was shipped and the actual profits for the fiscal year 2015-2016 that you obtained from different regions.



	A	B	C	D
1				
2			No. of Units	Actual Profits
3		Quarter 1	23	3358
4		Quarter 2	27	3829
5		Quarter 3	15	2374
6		Quarter 4	43	3373

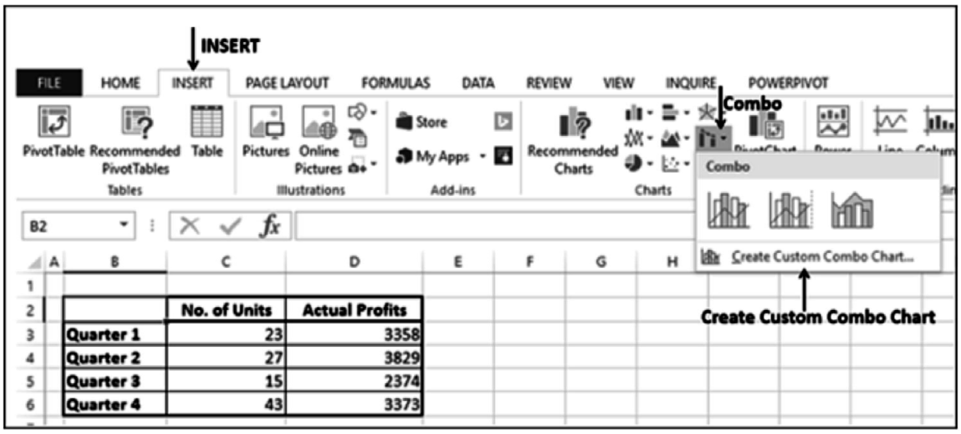
If you use the same combination chart as before, you will get the following:



In the chart, the data of **No. of Units** is not visible as the data ranges are varying significantly.

In such cases, you can create a combination chart with secondary axis, so that the primary axis displays one range and the secondary axis displays the other.

- Click the INSERT tab.
- Click Combo in Charts group.
- Click Create Custom Combo Chart from the drop-down list.



INSERT

Combo

Create Custom Combo Chart

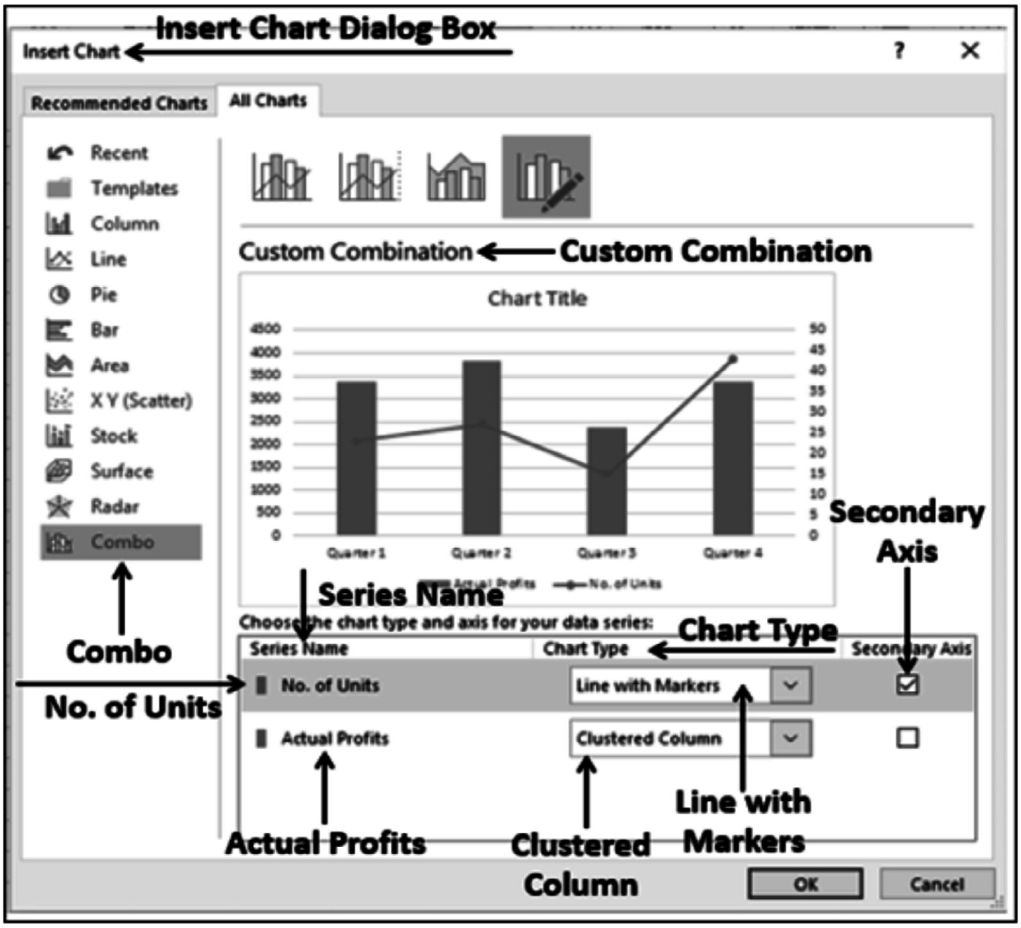
	No. of Units	Actual Profits
Quarter 1	23	3358
Quarter 2	27	3829
Quarter 3	15	2374
Quarter 4	43	3373

The Insert Chart dialog box appears with Combo highlighted.

For Chart Type, choose

- Line with Markers for the Series No. of Units
- Clustered Column for the Series Actual Profits
- Check the Box Secondary Axis to the right of the Series No. of Units and click OK.

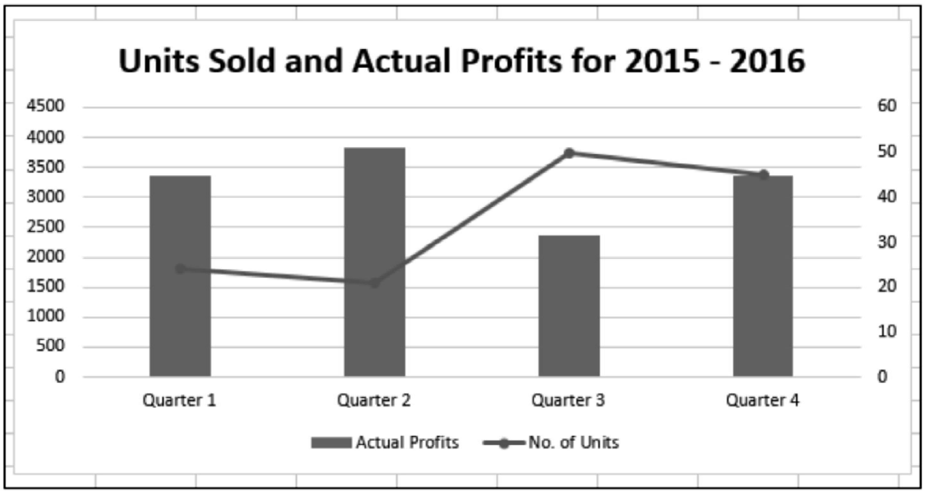
A preview of your chart appears under Custom Combination.



Choose the chart type and axis for your data series:

Series Name	Chart Type	Secondary Axis
No. of Units	Line with Markers	☑
Actual Profits	Clustered Column	□

Your Combo chart appears with Secondary Axis.

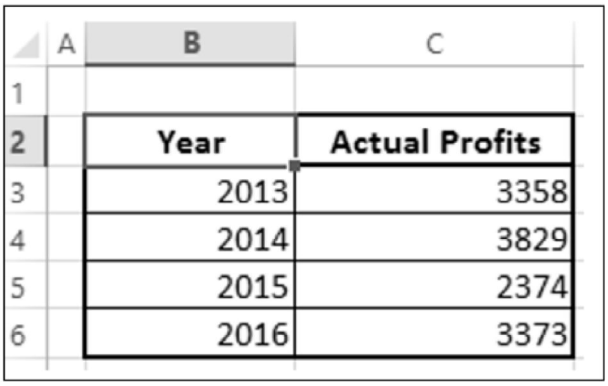


You can observe the values for Actual Profits on the primary axis and the values for No. of Units on the secondary axis.

A significant observation in the above chart is for Quarter 3 where No. of Units sold is more, but the Actual Profits made are less. This could probably be assigned to the promotion costs that were incurred to increase sales. The situation is improved in Quarter 4, with a slight decrease in sales and a significant rise in the Actual Profits made.

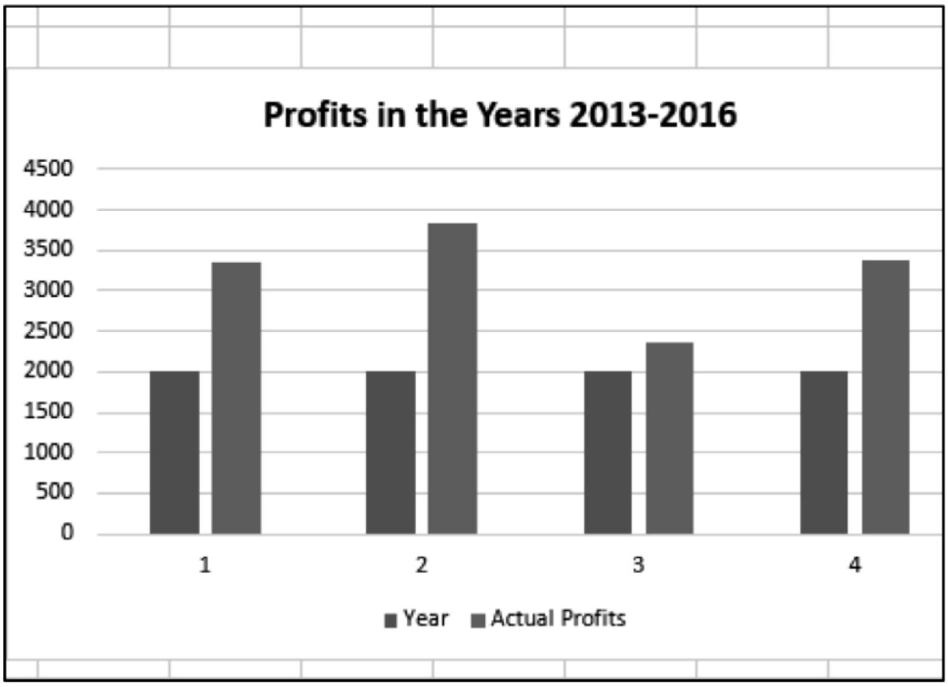
Discriminating Series and Category Axis

Suppose you want to project the Actual Profits made in Years 2013-2016.



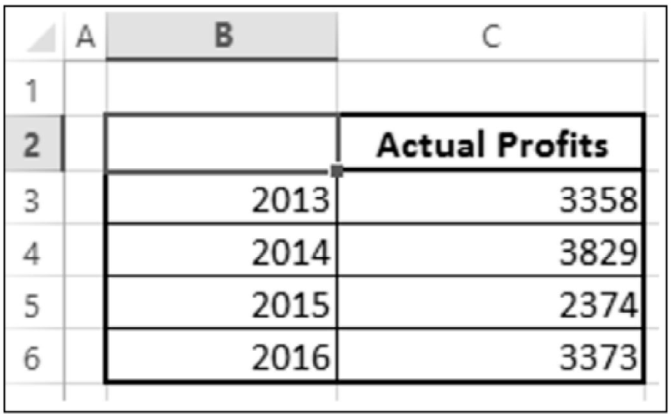
	A	B	C
1			
2		Year	Actual Profits
3		2013	3358
4		2014	3829
5		2015	2374
6		2016	3373

Create a clustered column for this data.



As you observe, the data visualization is not effective as the years are not displayed. You can overcome this by changing year to category.

Remove the header year in the data range.



	A	B	C
1			
2			Actual Profits
3		2013	3358
4		2014	3829
5		2015	2374
6		2016	3373

Now, year is considered as a category and not a series. Your chart looks as follows -

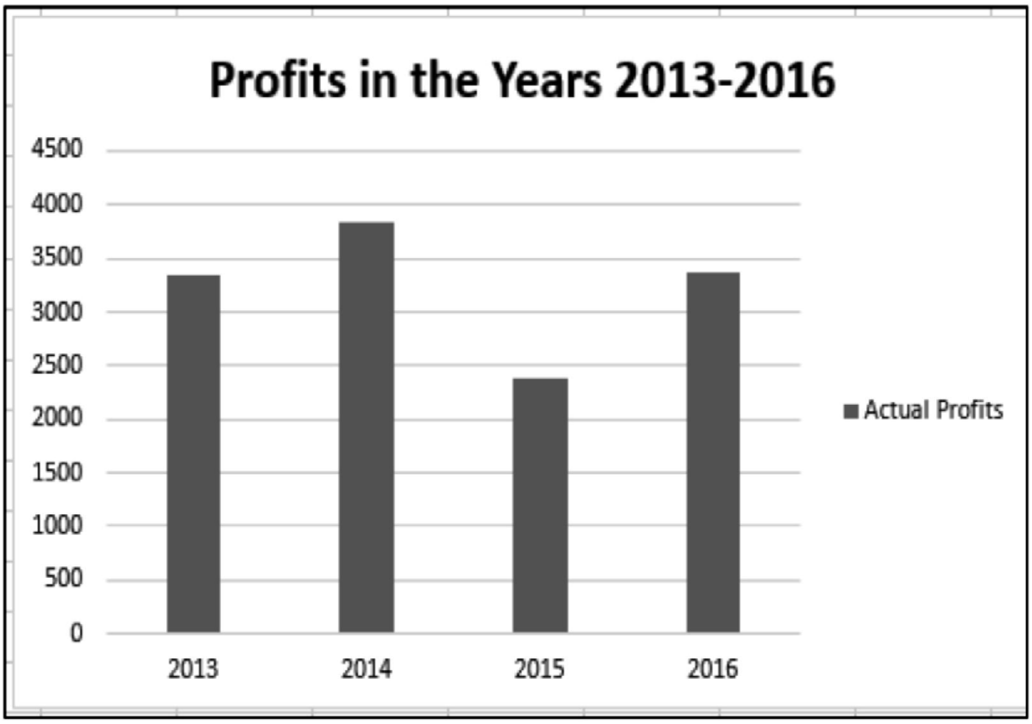


Chart Elements and Chart Styles

Chart Elements give more descriptions to your charts, thus helping visualizing your data more meaningfully.

- Click the Chart

Three buttons appear next to the upper-right corner of the chart "

- ➕ Chart Elements
- 🖌 Chart Styles
- ▽ Chart Filters

For a detailed explanation of these, refer to Excel Charts tutorial.

- Click Chart Elements.
- Click Data Labels.

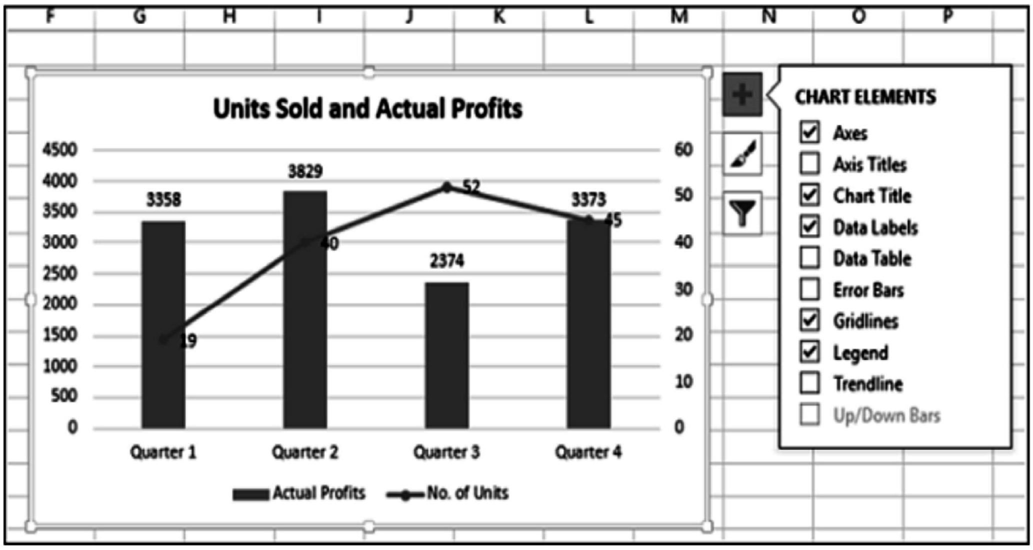
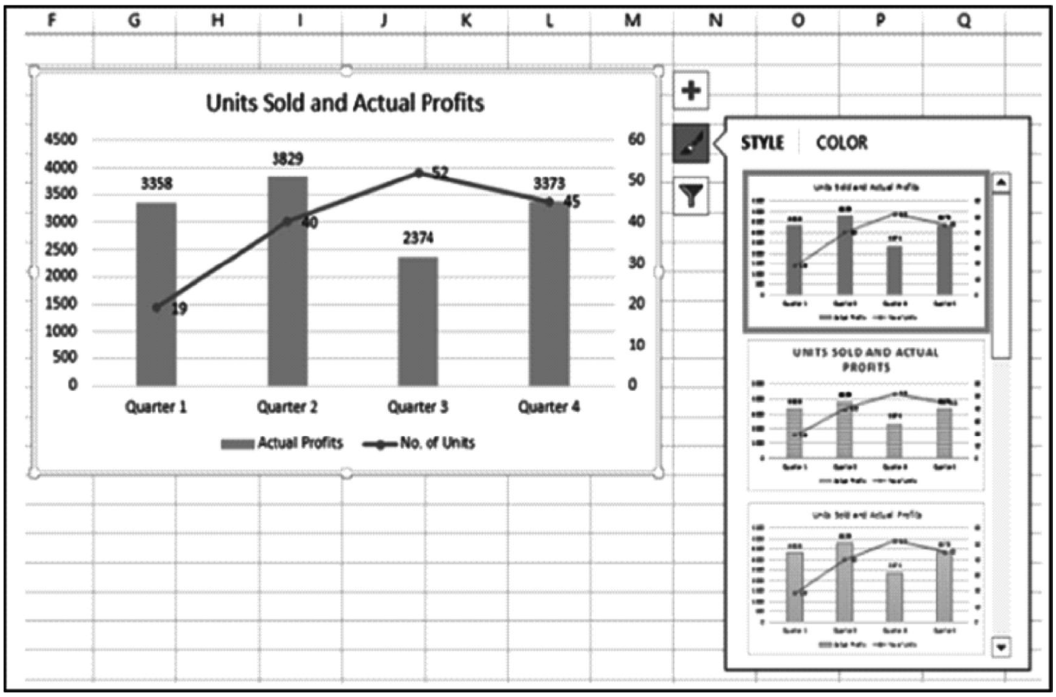


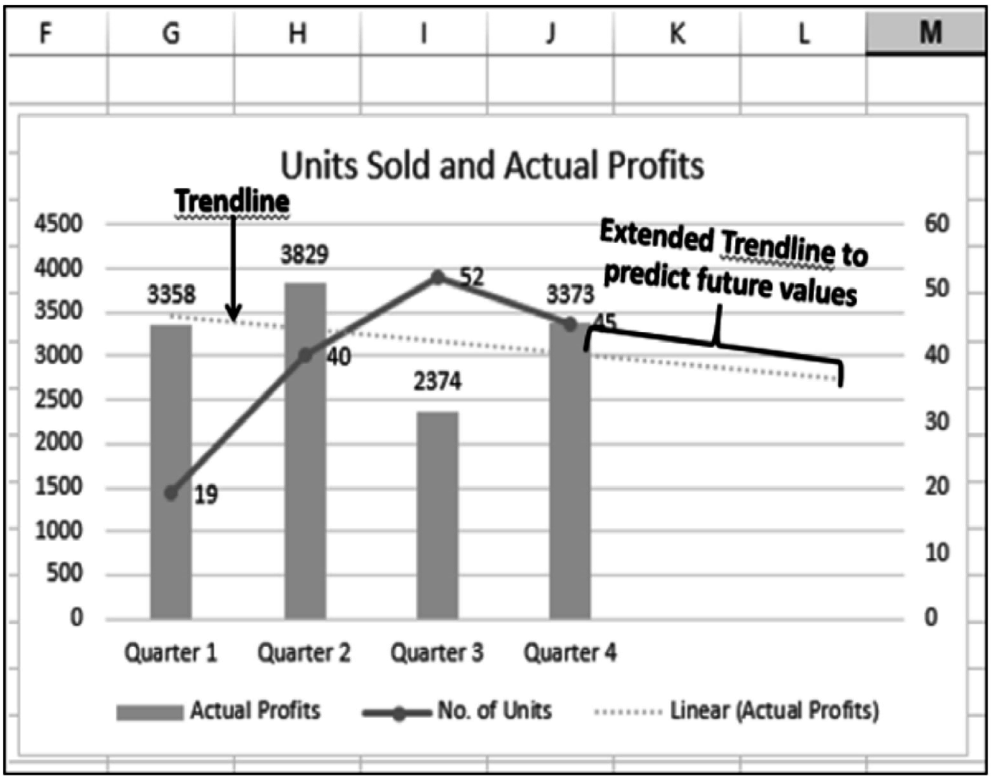
CHART ELEMENTS

- ☑ Axes
- □ Axis Titles
- ☑ Chart Title
- ☑ Data Labels
- □ Data Table
- □ Error Bars
- ☑ Gridlines
- ☑ Legend
- □ Trendline
- □ Up/Down Bars

- ➢ Click Chart Styles
- ➢ Select a Style and Color that suits your data.



You can use Trendline to graphically display trends in data. You can extend a Trendline in a chart beyond the actual data to predict future values.



2.3.2 Cross Tabulations

Q10. Explain briefly about "Cross tabulations charts" by using Ms. Excel?

Ans :

Cross tabulation is usually performed on categorical data that can be divided into mutually exclusive groups.

An example of categorical data is the region of sales for a product. Typically, region can be divided into categories such as geographic area (North, South, Northeast, West, etc) or state (Andhra Pradesh, Rajasthan, Bihar, etc). The important thing to remember about categorical data is that a categorical data point cannot belong to more than one category.

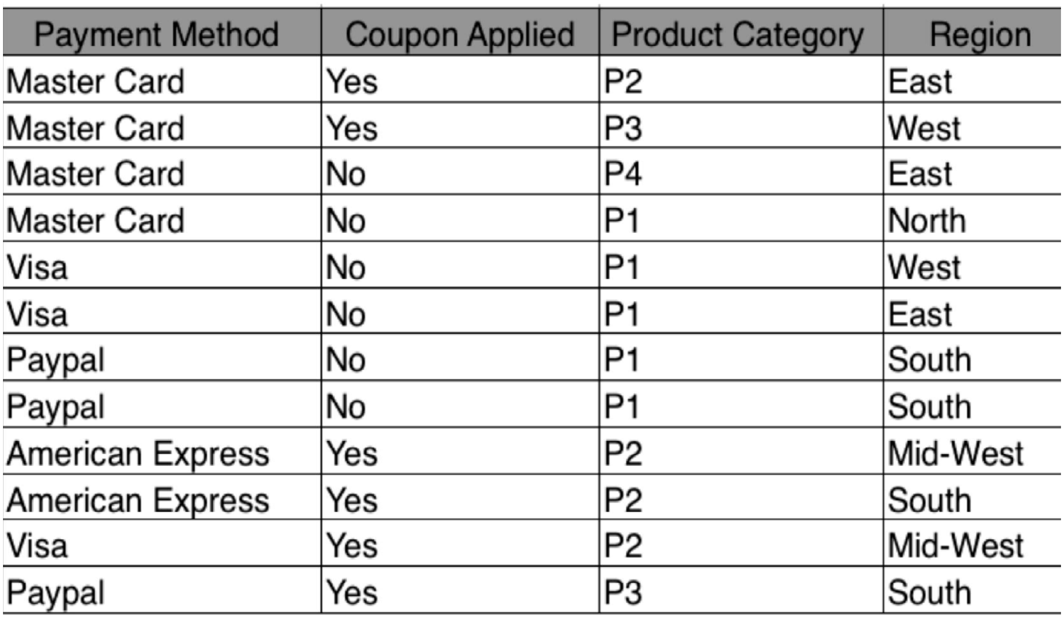
Cross tabulations are used to examine relationships within data that may not be readily apparent. Cross tabulation is especially useful for studying market research or survey responses. Cross tabulation of categorical data can be done with through tools such as SPSS, SAS, and Microsoft Excel.

An example of cross tabulation

"No other tool in Excel gives you the flexibility and analytical power of a pivot table." –Bill Jalen

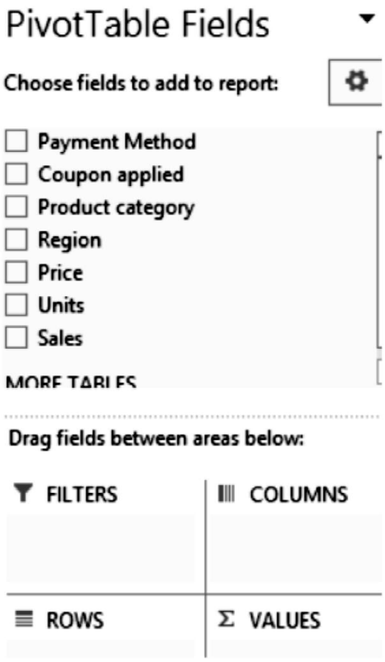
One simple way to do cross tabulations is Microsoft Excel's pivot table feature. Pivot tables are a great way to search for patterns as they help in easily grouping raw data.

Consider the below sample data set in Excel. It displays details about commercial transactions for four product categories. Let's use this data set to show cross tabulation in action.



Payment Method	Coupon Applied	Product Category	Region
Master Card	Yes	P2	East
Master Card	Yes	P3	West
Master Card	No	P4	East
Master Card	No	P1	North
Visa	No	P1	West
Visa	No	P1	East
Paypal	No	P1	South
Paypal	No	P1	South
American Express	Yes	P2	Mid-West
American Express	Yes	P2	South
Visa	Yes	P2	Mid-West
Paypal	Yes	P3	South

This data can be converted to pivot table format by selecting the entire table and inserting a pivot table in the Excel file. The table can correlate different variables row-wise, column-wise, or value-wise in either table format or chart format.



PivotTable Fields

Choose fields to add to report:

- □ Payment Method
- □ Coupon applied
- □ Product category
- □ Region
- □ Price
- □ Units
- □ Sales

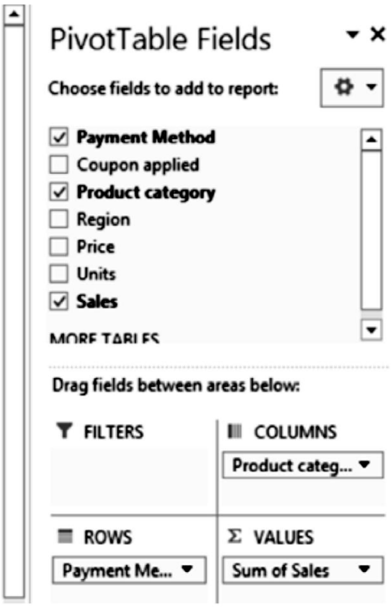
MORE TABLES

Drag fields between areas below:

FILTERS | COLUMNS

ROWS | VALUES

Let’s use cross tabulation to check the relation between the type of payment method (i.e. visa, MasterCard, PayPal, etc) and the product category with respect to the region of sales. We can select these three categories in the pivot table.



PivotTable Fields

Choose fields to add to report:

- ☑ **Payment Method**
- □ Coupon applied
- ☑ **Product category**
- □ Region
- □ Price
- □ Units
- ☑ **Sales**

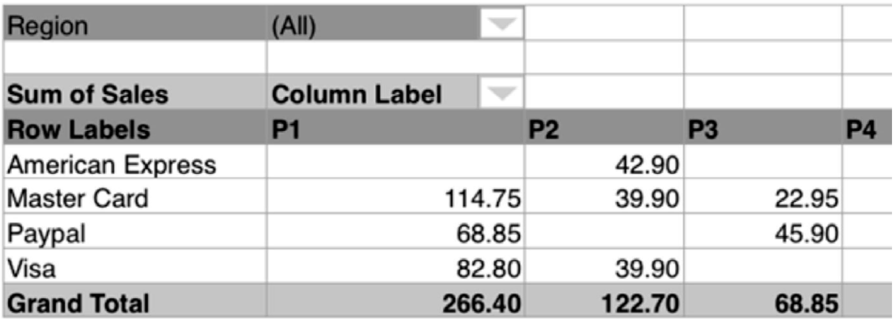
MORE TABLES

Drag fields between areas below:

FILTERS | COLUMNS: Product categ...

ROWS: Payment Me... | VALUES: Sum of Sales

Then the results appear in a pivot table:



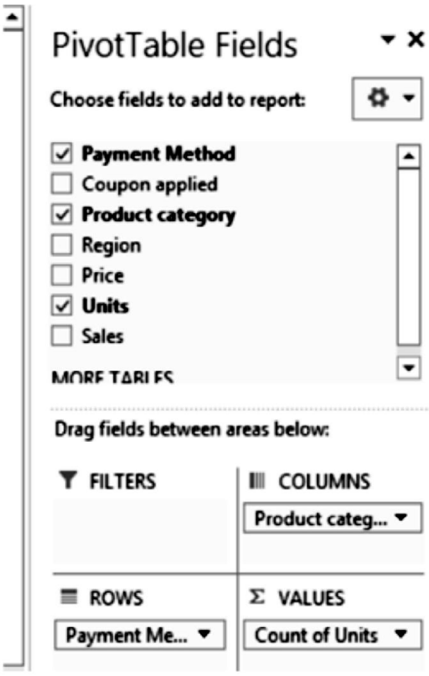
Region	(All)			
Sum of Sales	**Column Label**			
Row Labels	**P1**	**P2**	**P3**	**P4**
American Express		42.90		
Master Card	114.75	39.90	22.95	
Paypal	68.85		45.90	
Visa	82.80	39.90		
Grand Total	**266.40**	**122.70**	**68.85**	

Cross tabulation 1: Relation between payment method and the total amount of sales in product category with respect to region in which products sold

It is now clear that the highest sales were done for P1 using Master Card. Therefore, we can conclude that the MasterCard payment method and product P1 category is the most profitable combination.

Similarly, we can use cross tabulation and find the relation between the product category and the payment method type with regard to the number of transactions.

This can be done by grouping the payment method, product category, and units sold:



PivotTable Fields

Choose fields to add to report:

☑ **Payment Method**
☐ Coupon applied
☑ **Product category**
☐ Region
☐ Price
☑ **Units**
☐ Sales

MORE TABLES

Drag fields between areas below:

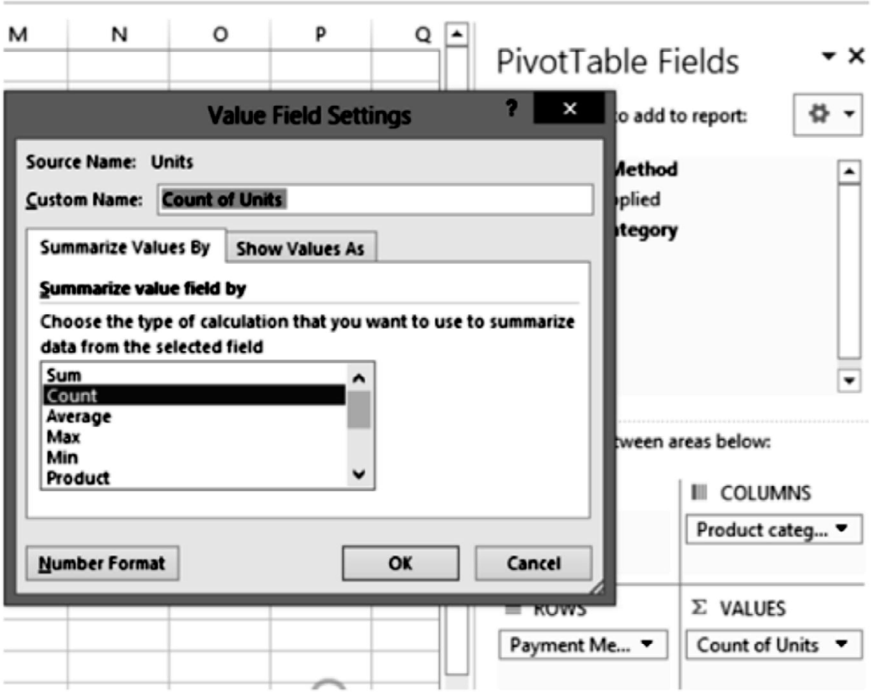
FILTERS

COLUMNS: Product categ...

ROWS: Payment Me...

VALUES: Count of Units

By default, Excel's pivot table aggregates values as a sum. Summing the units will give us the total number of units sold. Since we want to compare the number of transactions instead of the number of units sold, we need to change the Value Field Setting from Sum to Count for Units.



Value Field Settings

Source Name: Units

Custom Name: Count of Units

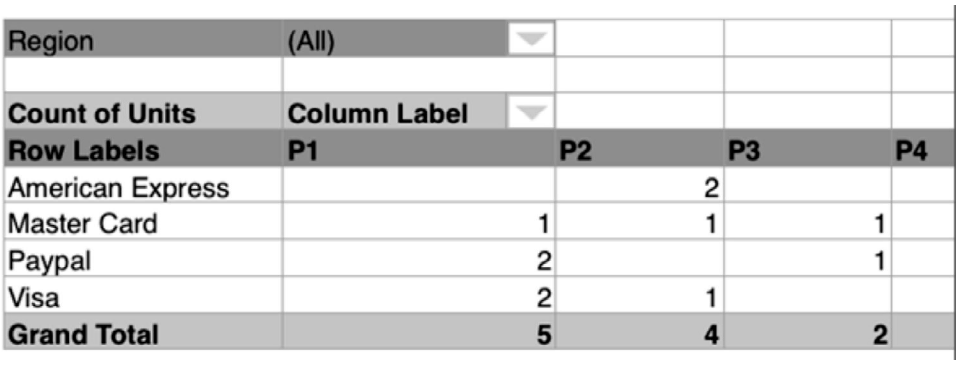
Summarize Values By | Show Values As

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

Sum
Count
Average
Max
Min
Product

The results of this pivot table mapping is as shown below. This is a cross tabulation analysis of 3 variables - it analyses the correlation between the payment method and payment category according to the number of transactions.



Region	(All)			
Count of Units	**Column Label**			
Row Labels	**P1**	**P2**	**P3**	**P4**
American Express		2		
Master Card	1	1	1	
Paypal	2		1	
Visa	2	1		
Grand Total	**5**	**4**	**2**	

Cross tabulation 2: Relation between payment method and total number of transactions in the product category with respect to region of sales

For all regions, we can observe that the highest selling category of products was P1 and the highest number of transactions was done using Master Card. We can also see the preferred payment method in each of the product categories. For example, American Express is the preferred card for P2 products.

Q11. Explain the benefits of cross tabulation ?

Ans :

i) Eliminates confusion while interpreting data

Raw data can be difficult to interpret. Even for small data sets, it is all too easy to derive wrong results by just looking at the data. Cross tabulation offers a simple method of grouping variables, which minimizes the potential for confusion or error by providing clear results.

ii) Helps in deriving innumerable insights

As we observed in our example, cross tabulation can help us derive great insights from raw data. These insights are not easy to see when the raw data is formatted as a table. Since cross tabulation clearly maps out relations between categorical variables, researchers can gain better and deeper insights — insights that otherwise would have been overlooked or would have taken a lot of time to decode from more complicated forms of statistical analysis.

iii) Offers data points to chart out a course of action

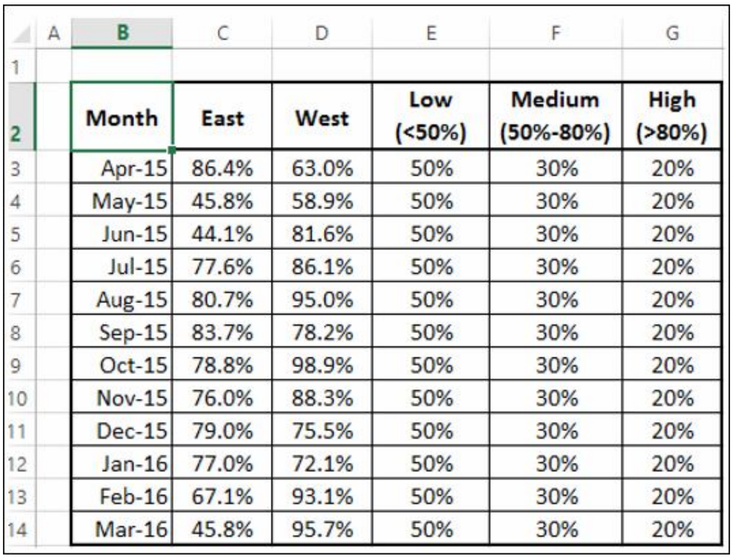
Cross tabulation makes it easier to interpret data, which is beneficial for researchers who have limited knowledge of statistical analysis. With cross tabulation, people do not need statistical programming to correlate categorical variables. The clarity offered by cross tabulation helps professionals evaluate their current work and chart out future strategies.

Q12. Explain briefly about band chart?

Ans :

You might have to present customer survey results of a product from different regions. Band Chart is suitable for this purpose. A Band Chart is a Line Chart with an added shaded area to display the upper and lower boundaries of groups of data.

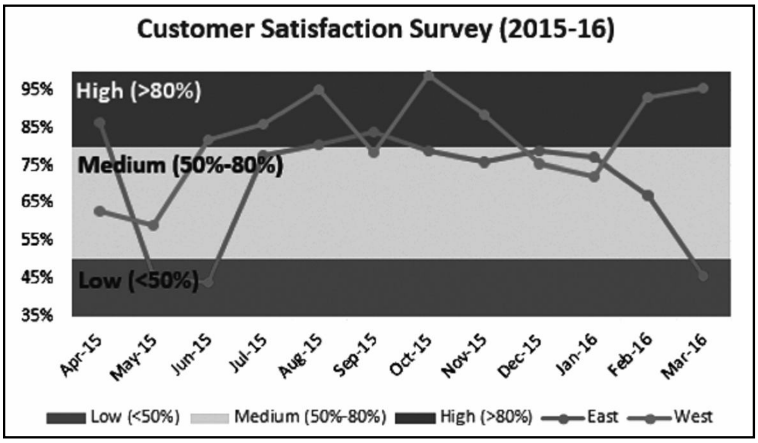
Suppose your customer survey results from the east and west regions, month-wise are:



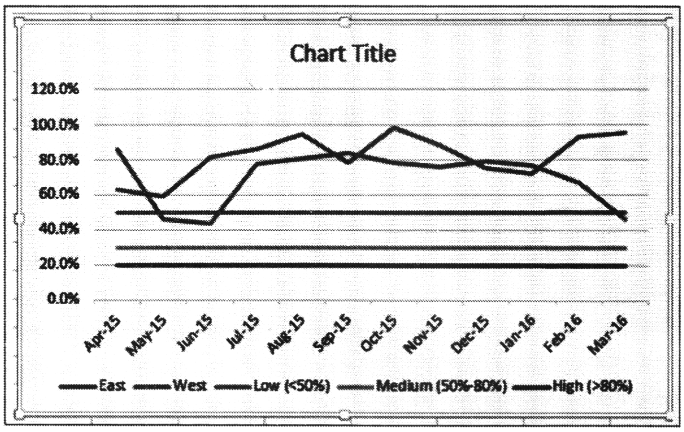
Month	East	West	Low (<50%)	Medium (50%-80%)	High (>80%)
Apr-15	86.4%	63.0%	50%	30%	20%
May-15	45.8%	58.9%	50%	30%	20%
Jun-15	44.1%	81.6%	50%	30%	20%
Jul-15	77.6%	86.1%	50%	30%	20%
Aug-15	80.7%	95.0%	50%	30%	20%
Sep-15	83.7%	78.2%	50%	30%	20%
Oct-15	78.8%	98.9%	50%	30%	20%
Nov-15	76.0%	88.3%	50%	30%	20%
Dec-15	79.0%	75.5%	50%	30%	20%
Jan-16	77.0%	72.1%	50%	30%	20%
Feb-16	67.1%	93.1%	50%	30%	20%
Mar-16	45.8%	95.7%	50%	30%	20%

Here, in the data < 50% is Low, 50% - 80% is Medium and > 80% is High.

With Band Chart, you can display your survey results as follows:

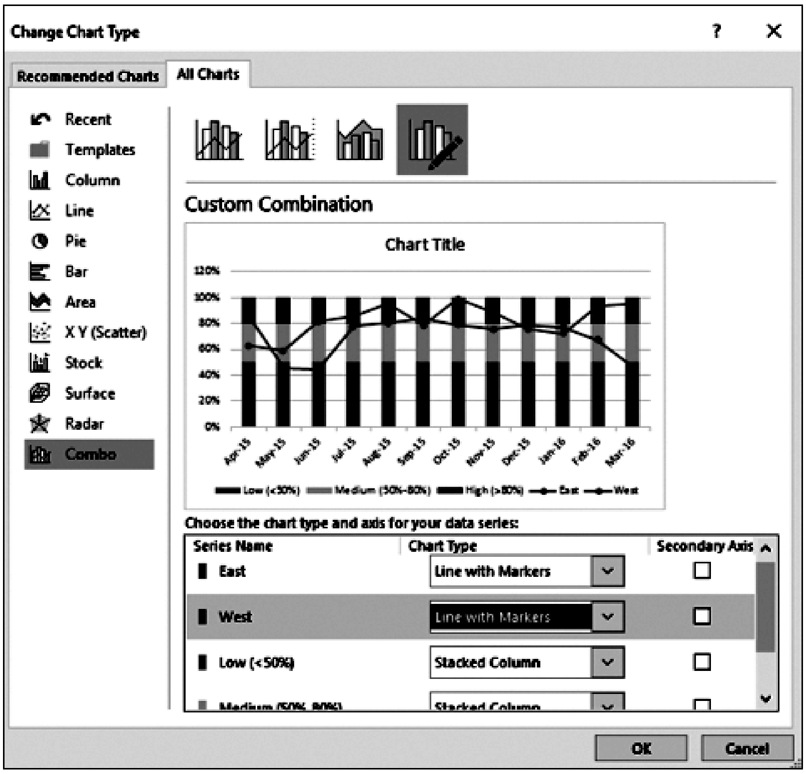


Create a Line Chart from your data.



Change the chart type to:

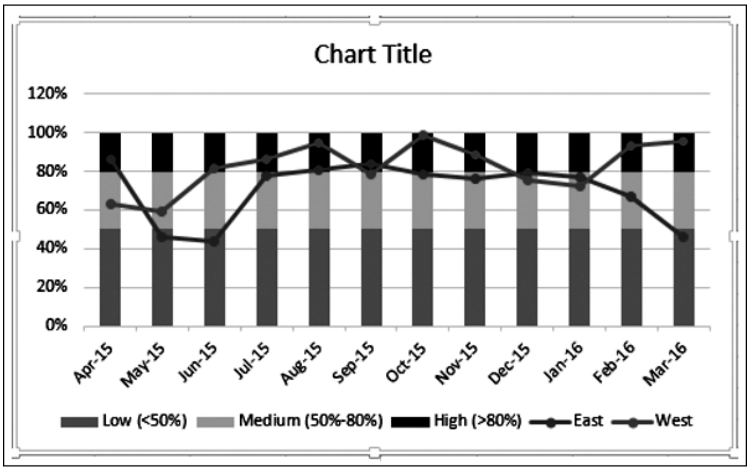
- East and West Series to Line with Markers.
- Low, Medium and High Series to Stacked Column.



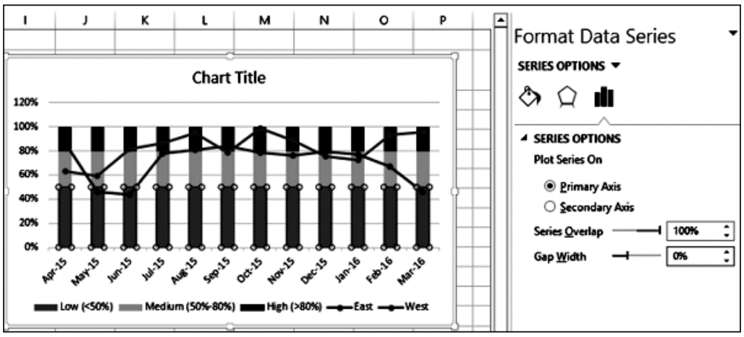
Choose the chart type and axis for your data series:

Series Name	Chart Type	Secondary Axis
East	Line with Markers	□
West	Line with Markers	□
Low (<50%)	Stacked Column	□
Medium (50% 80%)	Stacked Column	□

Your chart looks as follows:



- Click on one of the columns.
- Change gap width to 0% in Format Data Series.



Format Data Series

SERIES OPTIONS

SERIES OPTIONS

Plot Series On

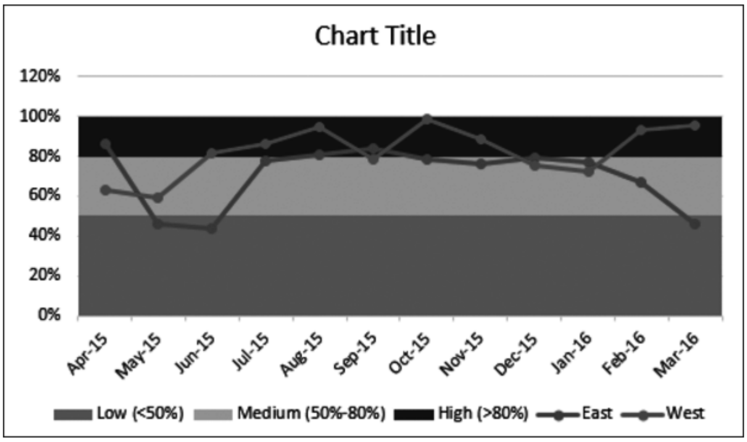
● Primary Axis

○ Secondary Axis

Series Overlap 100%

Gap Width 0%

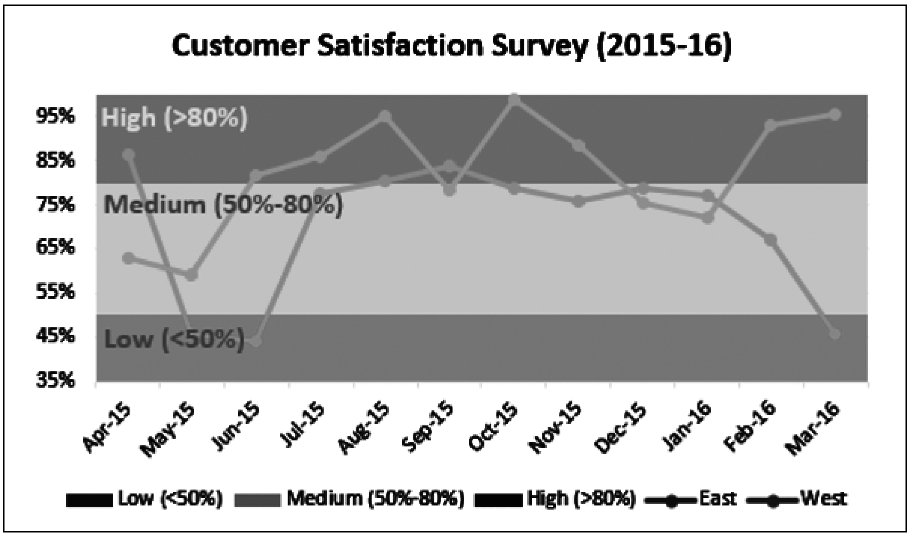
You will get Bands instead of columns.



To make the chart more presentable:

- Add Chart Title.
- Adjust Vertical Axis range.
- Change the colors of the bands to Green-Yellow-Red.
- Add Labels to bands.

The final result is the Band Chart with the defined boundaries and the survey results represented across the bands. One can quickly and clearly make out from the chart that while the survey results for the region West are satisfactory, those for the region East have a decline in the last quarter and need attention.



Q13. Explain briefly about Gantt Chart.

Ans :

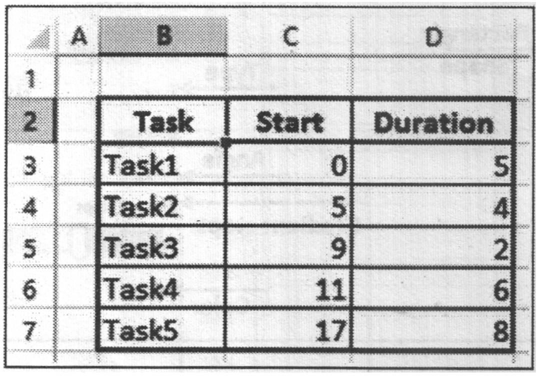
A Gantt Chart is a chart in which a series of horizontal lines shows the amount of work done in certain periods of time in relation to the amount of work planned for those periods.

In Excel, you can create a Gantt Chart by customizing a Stacked Bar Chart type so that it depicts tasks, task duration and hierarchy. An Excel Gantt Chart typically uses days as the unit of time along the horizontal axis.

Consider the following data where the column:

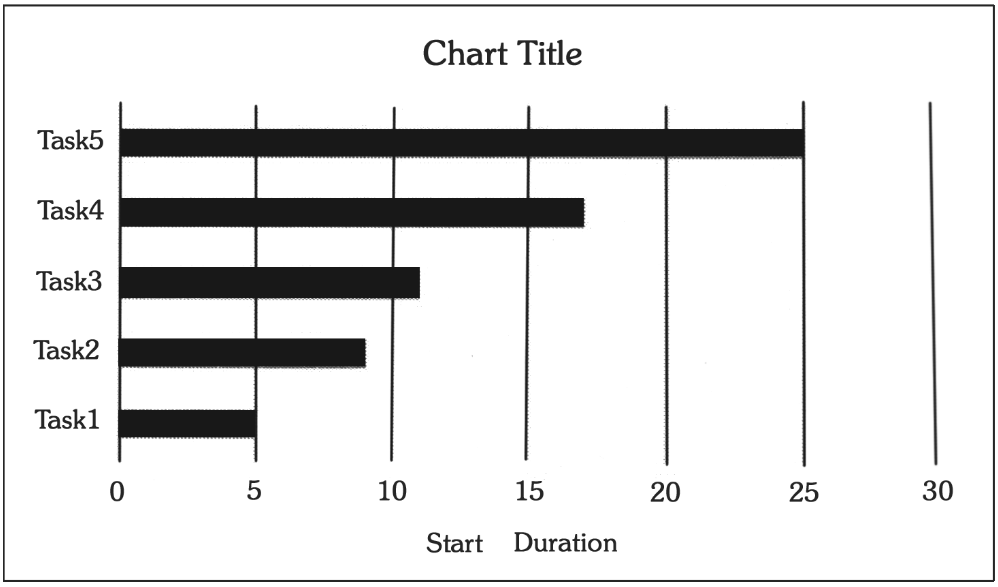
- Task represents the Tasks in the project
- Start represents number of days from the Start Date of the project
- Duration represents the duration of the Task

Note that Start of any Task is Start of previous Task + Duration. This is the case when the Tasks are in hierarchy.

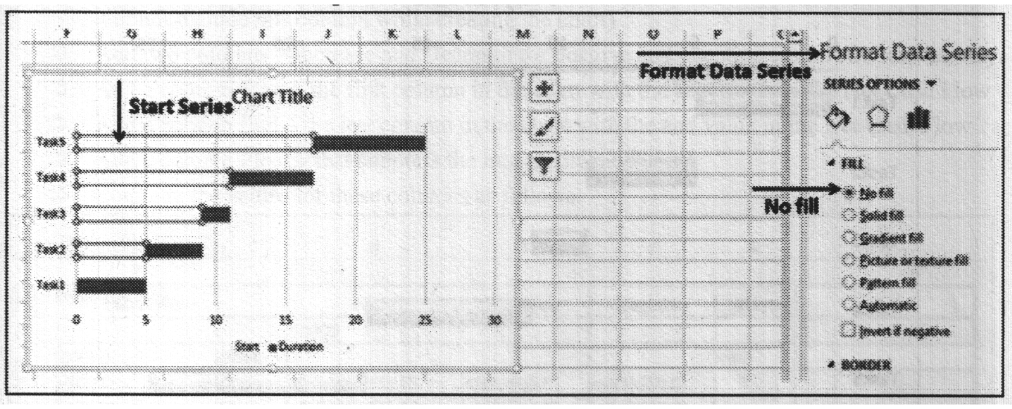


Task	Start	Duration
Task1	0	5
Task2	5	4
Task3	9	2
Task4	11	6
Task5	17	8

- Select the data
- Create Stacked Bar Chart



- ➢ Right-click on Start Series
- ➢ In Format Data Series options, select No fill



Format Data Series

SERIES OPTIONS ▾

◢ FILL

● No fill

○ Solid fill

○ Gradient fill

○ Picture or texture fill

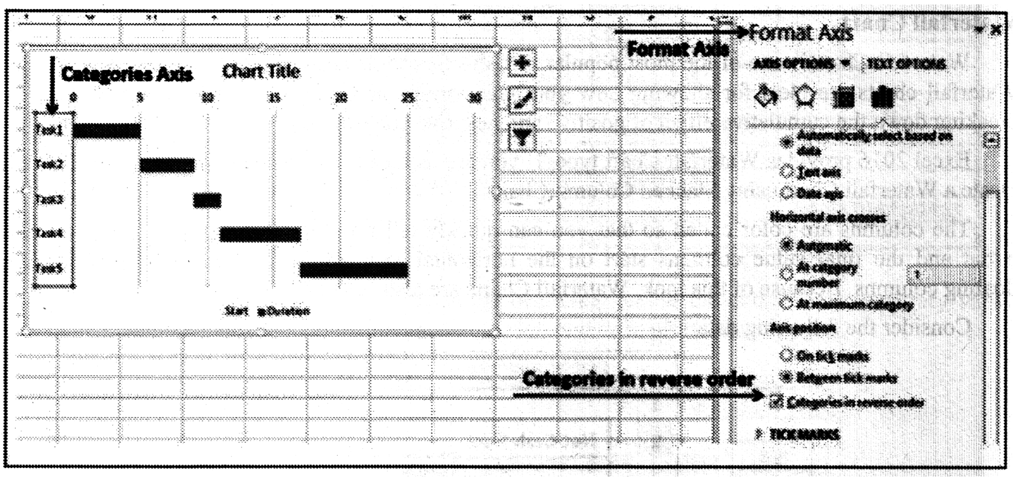
○ Pattern fill

○ Automatic

□ Invert if negative

◢ BORDER

- ➢ Right-click on Categories Axis
- ➢ In Format Axis options, select Categories in reverse order.



Format Axis

AXIS OPTIONS ▾ TEXT OPTIONS

● Automatically select based on data

○ Text axis

○ Date axis

Horizontal axis crosses

● Automatic

○ At category number 1

○ At maximum category

Axis position

○ On tick marks

● Between tick marks

☑ Categories in reverse order

▸ TICK MARKS

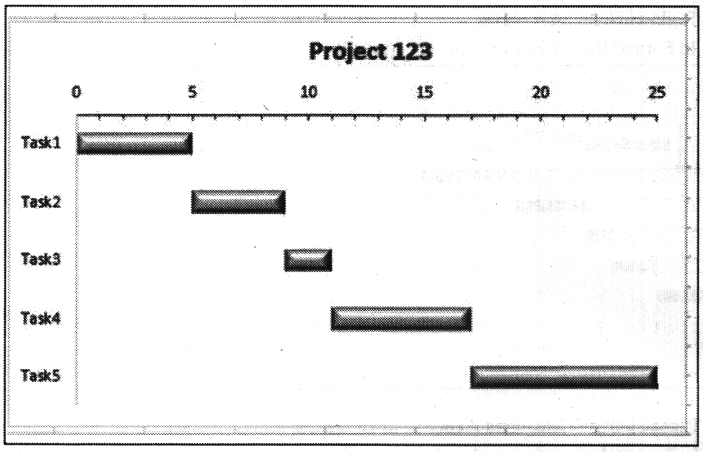
In Chart Elements, deselect:

- ➢ Legend
- ➢ Gridlines

Format the Horizontal Axis to:

- ➢ Adjust the range
- ➢ Major Tick Marks at 5 day intervals

- Minor Tick Marks at 1 day intervals Format Data Series to make it look impressive Give a Chart Title.



Q14. Explain briefly about pivot charts.

Ans :

PivotCharts are used to graphically summarize data and explore complicated data.

A PivotChart shows data series, categories and chart axes the same way a standard chart does. Additionally, it also gives you interactive filtering controls right on the chart so that you can quickly analyze a subset of your data.

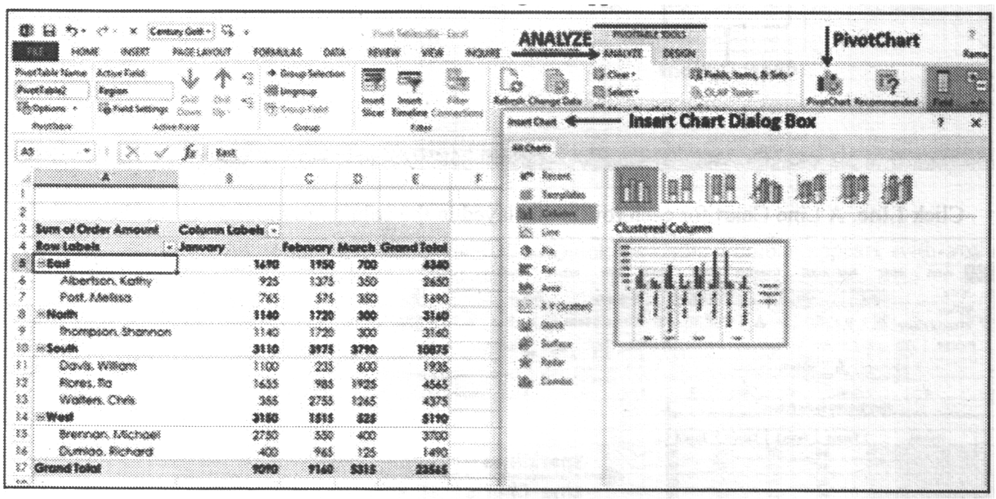
PivotCharts are useful when you have data in a huge PivotTable, or many complex worksheet data that includes text and numbers. A PivotChart can help you make sense of this data.

You can create a PivotChart from:

- A PivotTable
- A Data Table as a standalone without PivotTable **PivotChart from PivotTable**

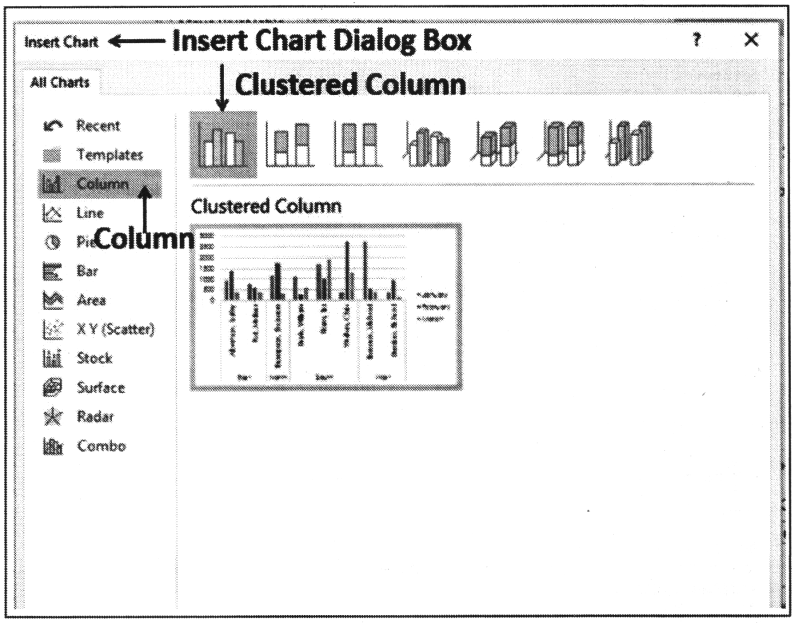
To create a PivotChart, follow the steps given below:

- Click the PivotTable.
- Click ANALYZE under PIVOTTABLE TOOLS on the Ribbon.
- Click on PivotChart. The Insert Chart dialog box appears.



Sum of Order Amount	Column Labels			
Row Labels	January	February	March	Grand Total
East	1690	1950	700	4340
Albertson, Kathy	925	1375	350	2650
Post, Melissa	765	575	350	1690
North	1140	1720	300	3160
Thompson, Shannon	1140	1720	300	3160
South	3110	3975	3790	10875
Davis, William	1100	235	600	1935
Flores, Tia	1655	985	1925	4565
Walters, Chris	355	2755	1265	4375
West	3150	1515	525	5190
Brennan, Michael	2750	550	400	3700
Dumlao, Richard	400	965	125	1490
Grand Total	9090	9160	5315	22565

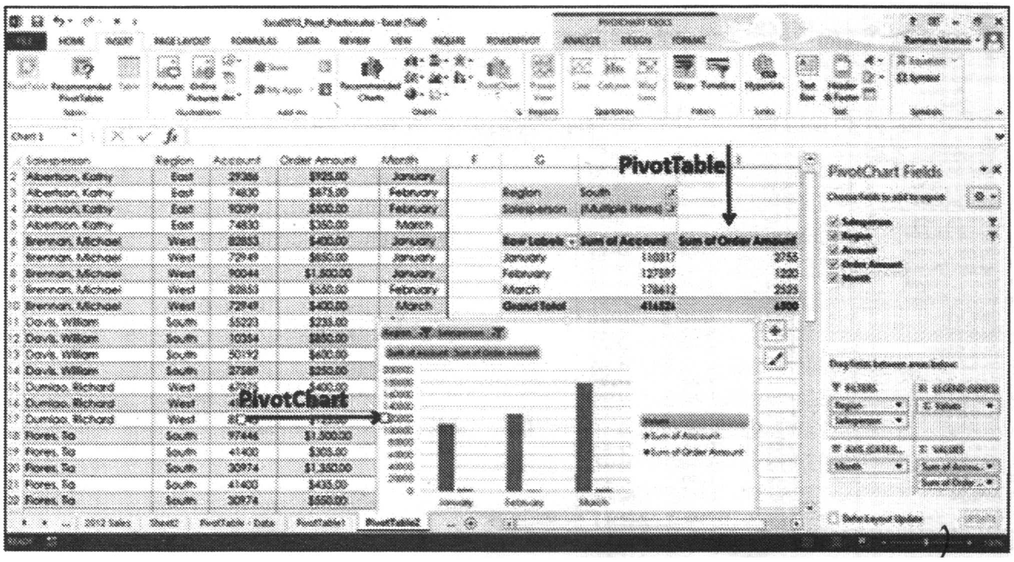
Select Clustered Column from the option Column.



Click OK. The PivotChart is displayed.

The PivotChart has three filters - Region, Salesperson and Month.

- Click the Region Filter Control option. The Search Box appears with the list of all Regions. Check boxes appear next to Regions.



	Salesperson	Region	Account	Order Amount	Month
2	Albertson, Kathy	East	29386	$925.00	January
3	Albertson, Kathy	East	74830	$875.00	February
4	Albertson, Kathy	East	90099	$300.00	February
5	Albertson, Kathy	East	74830	$350.00	March
6	Brennan, Michael	West	82853	$400.00	January
7	Brennan, Michael	West	72949	$650.00	January
8	Brennan, Michael	West	90044	$1,500.00	January
9	Brennan, Michael	West	82853	$550.00	February
10	Brennan, Michael	West	72949	$400.00	March
11	Davis, William	South	55223	$235.00	
12	Davis, William	South	10364	$850.00	
13	Davis, William	South	50192	$400.00	
14	Davis, William	South	37589	$350.00	
15	Dumlao, Richard	West	47076	$400.00	
16	Dumlao, Richard	West	[illegible]	[illegible]	
17	Dumlao, Richard	West	[illegible]	[illegible]	
18	Flores, Tia	South	97446	$1,300.00	
19	Flores, Tia	South	41400	$305.00	
20	Flores, Tia	South	30974	$1,350.00	
21	Flores, Tia	South	41400	$435.00	
22	Flores, Tia	South	30974	$550.00	

Region	South
Salesperson	(Multiple Items)

Row Labels	Sum of Account	Sum of Order Amount
January	110817	2755
February	127597	1220
March	178612	2525
Grand Total	416526	6500

2.3.3 Data Dashboards using MS. Excel (or) Spss

Q15. What is Data dashboard and explain different types dashboards.

Ans :

A data dashboard is an information management tool that visually tracks, analyzes and displays key performance indicators (KPI), metrics and key data points to monitor the health of a business, department or specific process. They are customizable to meet the specific needs of a department and company. Behind the scenes, a dashboard connects to your files, attachments, services and API's, but on the surface displays all this data in the form of tables, line charts, bar charts and gauges. A data dashboard is the most efficient way to track multiple data sources because it provides a central location for businesses to monitor and analyze performance. Real-time monitoring reduces the hours of analyzing and long line of communication that previously challenged businesses.

In the present terms, a dashboard can be defined as a data visualization tool that displays the current status of metrics and key performance indicators (KPIs) simplifying complex data sets to provide users with at a glance awareness of current performance.

Dashboards consolidate and arrange numbers and metrics on a single screen. They can be tailored for a specific role and display metrics of a department or an organization on the whole.

Dashboards can be static for a one-time view, or dynamic showing the consolidated results of the data changes behind the screen. They can also be made interactive to display the various segments of large data on a single screen.

Types of Dashboards

Dashboards can be categorized based on their utility as follows:

1. Strategic Dashboards
2. Analytical Dashboards
3. Operational Dashboards
4. Informational Dashboards

1. Strategic Dashboards

Strategic dashboards support managers at any level in an organization for decision-making. They provide the snapshot of data, displaying the health and opportunities of the business, focusing on the high level measures of performance and forecasts.

- ➢ Strategic dashboards require to have periodic and static snapshots of data (e.g., daily, weekly, monthly, quarterly and annually). They need not be constantly changing from one moment to the next and require an update at the specified intervals of time.
- ➢ They portray only the high level data not necessarily giving the details.
- ➢ They can be interactive to facilitate comparisons and different views in case of large data sets at the click of a button. But, it is not necessary to provide more interactive features in these dashboards.

The following screenshot shows an example of an executive dashboard, displaying goals and progress.

2. Analytical Dashboards

Analytical dashboards include more context, comparisons, and history. They focus on the various facets of data required for analysis.

Analytical dashboards typically support interactions with the data, such as drilling down into the underlying details, and hence should be interactive.

Examples of analytical dashboards include Finance Management dashboard and Sales Management dashboard.

3. Operational Dashboards

Operational dashboards are for constant monitoring of operations. They are often designed differently from strategic or analytical dashboards and focus on monitoring of activities and events that are constantly changing and might require attention and response at a moment's notice. Thus, operational dashboards require live and up-to-date data available at all times, and hence should be dynamic.

An example of an operation dashboard could be a support-system dashboard, displaying live data on service tickets that require an immediate action from the supervisor on high-priority tickets.

4. Informational Dashboards

Informational dashboards are just for displaying figures, facts and/or statistics. They can be either static or dynamic with live data but not interactive. For example, flights arrival/departure information dashboard in an airport.

Q16. What are the steps to create interactive Excel Dash Board?

Ans :

Create Interactive Excel Dashboard

Most of us probably rely on our trusted MS Excel dashboard for the day to day running of our businesses, but like many, we struggle to turn that data into something that will actually interest people and want them to know more about it. It is a comprehensive as well as complete visual report or analysis of your project which can be shared with other people concerned. Creating a excel dashboard can be tedious, time consuming as well as difficult if you do not have the proper knowledge about how to go about doing it. But fret now, that's where we enter.

Dashboards are not native to Excel, as they can be created on PowerPoint as well. Excel Dashboards offer a more dynamic approach to presenting data as compared to the more liner as well as un moving nature of PowerPoint dashboards.

An interactive dashboard in excel is basically slices of visualization which enables your data to tell a story. A dashboard is only useful if they are dynamic, easy to use as well as compatible with the PC you are using. Before making a dashboard, you need to consider the decisions the end user will make based on the data, the look and feel of it. You will also need to keep in mind how much they are familiar with the data and how much context they have. For example, a monthly report for your boss who is already familiar with everything is going to look very different from the one which you make to pitch a new idea to a potential client.

Another thing to remember is that the data should be the star of the excel dashboard. There is no need to clutter the screen with unnecessary components, so keeping it simple is the best way to go. You will also want to strike a perfect balance between making it look striking (so that it holds your audience's attention), but not so stylized so that it takes away from the data to be presented. When we tell a story, we must always consider the tastes and distastes of the audience and adapt our presentation accordingly. For example, if you are presenting to a very formal organisation, you should do your best to keep the excel dashboard as simple as possible, without compromising on subdued attractiveness.

Armed with the right knowledge about how to go on about creating a stunning excel dashboard, you can create a excel dashboard of your own without it being tedious or difficult! We provide you with a step by step analysis below:

1. **Bringing in data**

   Sure, Excel is very useful and flexible. But to create a excel dashboard you cannot just paste some data and add a few charts. You need to maintain it, update it and you must impose some kind of structure to that data. Usually you don't have to enter the data directly into the spreadsheet. You can copy paste the data but the best option is to bring in data via an external source. You can use it to connect your excel dashboard to Access or Oracle. A good practice is to limit the amount of data you bring in. As we've seen before, data can be brought in with two basic structures: a flat file and a pivot table. A flat file is generally smaller, where as a pivot table is a large file (As a thumb rule). Both have their pros and cons which one must figure out only through experience.

2. Select a background

Select an appropriate background which will bring your excel dashboard appear attractive without taking focus away from the data. Your data should be the star. You can go with subdued shades like blue, grey and black or you can take it up a notch like orange, green and purple. It is your choice, but keep in mind the audience you will be presenting it to. I suggest you stick to subdued hues if it is for official purposes.

3. Manage your data and link it to your excel dashboard

If you are using a pivot table, use the GETPIVOTDATA function. If you use a flat file, there are a number of formulae you can use like DSUM, DGET, VLOOKUP, MATCH, INDEX or even a dew math formulas like SUM, SUMIF, etc

But be careful here, do not punch in formula after formula. Fewer formulas mean a safer and a more reliable excel dashboard which is also easier to maintain. You can automatically reduce the formula number by using pivot tables.

Also, another important point is that you should name all your ranges. Always, always document your work. Simplify your work by making your excel dashboard formulas cleaner.

4. Use Dynamic Charting

Dashboards that a user can't interact with don't make much sense. All your excel dashboards should have controls which will enable you to change the markets, product details as well as other nitty critters. What is most important is that the user must be able to be in complete charge of his or her own excel dashboard and make changes whenever and wherever they want.

If you are creating interactive charts, you will need dynamic ranges. You can do this by using the OFFSET() function. You can also add a few cool things to your excel dashboard like greeting the user and selecting the corresponding profile when they open the excel dashboard. All this can be done using macros. All you need to do is record a macro, add a FOR NEXT or a FOR EACH loop. If you have never recorded a macro before, there are a large number of sites online which give you perfectly tailored macros as per your needs.

5. Design your excel dashboard report

If you are still using Excel 2003 or 2007, their default charts are not very attractive so I suggest you avoid them like the plague, but make sure to use acceptable formats. Excel 2010 and 2013 are a lot better but they still need some work. Keep this in mind, a chart is used to discover actionable patterns in the data and you should do your best to bring out most of it. This also means that you should remove all the jazzy, glittery stuff which adds no value to your excel dashboard. What you can do instead is create a hierarchy of focus and contextual data that is relevant, and create a form of basic interact if not much.

6. Dashboard Storytelling

Storytelling which is pregnant with data is the best kind that there is. With better access to data and better tools to make a point, we are able to recover a lot of data types. However, even though data is good, it is great, but you must not reveal all of it at once. When deciding how to make a excel dashboard, start by reviewing the purpose of the said dashboard. The goal shouldn't be to overwhelm the audience with data, but to provide data in such a form so that it gives them the insight you want them to have. I think this is true for all data based projects.

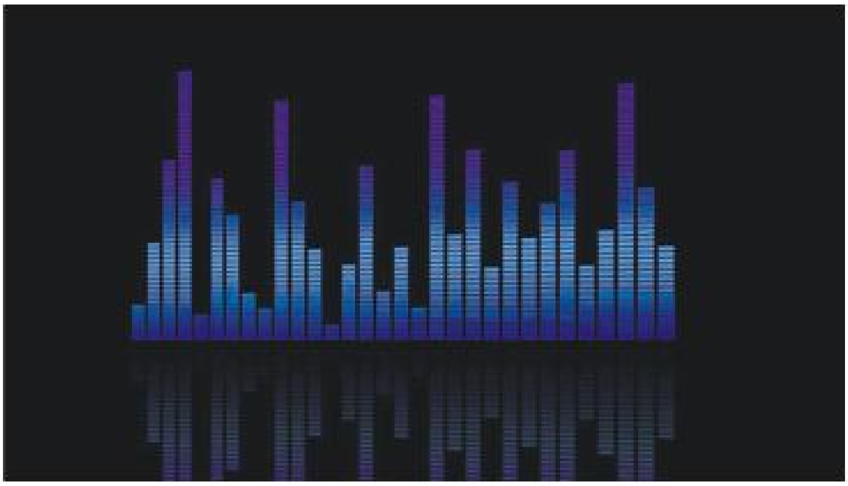
Let your audience explore the data on their own by offering them their own filters and controls. This is where interactive visuals come in the picture. If you are a newcomer to interactive excel dashboards, you can still spot trends and learn how to build up a stunning dashboard. If you are a pro at it, you can drill down deeper into the data for better charts.

7. Select the right kind of chart type

Before we decide which chart to use in our excel dashboard, let us have a review of all the charts used in dashboards and when to use what.

a) Bar Charts

Bar charts as we all know are bars on the x axis. One of the most common misgiving about excel dashboards is that the more is better; the truth is, that is seldom true. Bar charts are simple and very effective. They are particularly useful to compare one concept to another as well as trends.

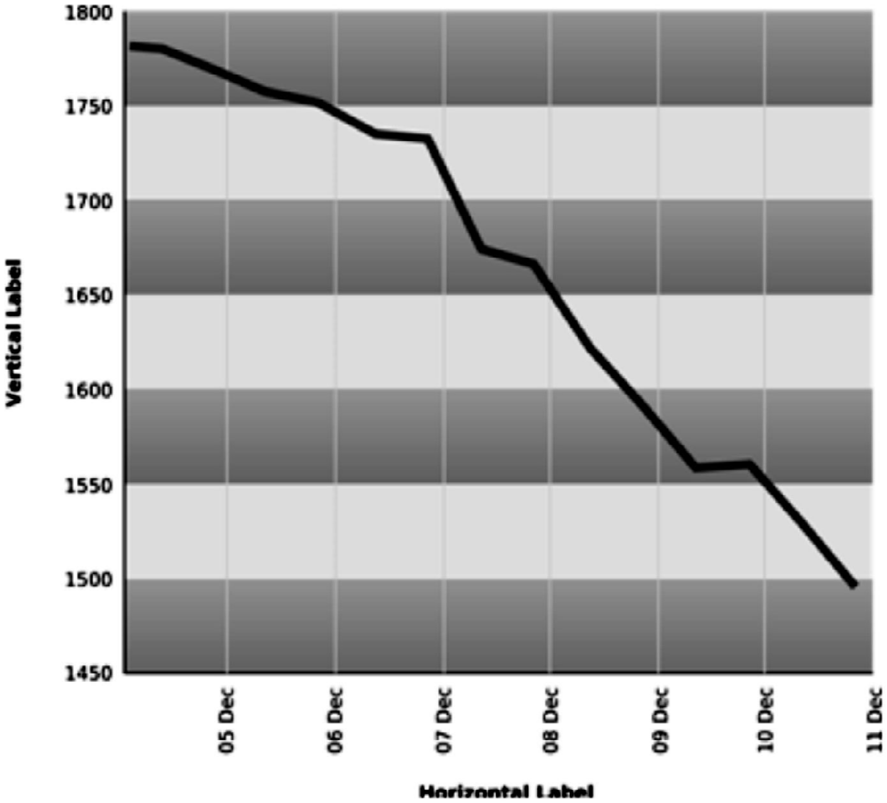


(b) Pie Charts

These charts, in my personal opinion, should be used very carefully and sparingly. Well, no matter how you feel about pie charts, you should only use them when you need a graph representing proportions of a whole. Use with extreme frugality.

(c) Line Charts

These are one of my favorites. They are so simplistic. These charts include a serious of data points that are connected by a line. These are best used to show developments over a certain period of time.



(d) Tables

Tables are great if you have detailed information with different measuring units, which may be difficult to represent through other charts or graphs.

(e) Area charts

Area charts are very useful for multiple data series, which may or may not be related to each other (partially or wholly). They are also useful for an individual series that represents a physically countable set.

So choose wisely, and you will be good.

8. Colour theory

Colours in a excel dashboard make it livelier as opposed to the drab and overused grey, black and white. I could write an entire book on how colour theory works, but well, that's already dine. You must know which colours work together and which do not. For example, you cannot pair bright pink and red together unless you want an assault on the eyes. One thing you must keep in mind while selecting a colour coding, that 8% of men and 0.5% or women are colour blind.

Most people can perceive a colour, but cannot correctly distinguish between two shades of the same colour. These people can perceive changes in brightness though, just like me and you. Avoid having shades that overlap, like the example I gave above. That would not only look ugly, but also be completely useless for users we discussed above.

Q17. What are the benefits of data dash boards.

Ans :

Dashboards allow managers to monitor the contribution of the various departments in the organization. To monitor the organization's overall performance, dashboards allow you to capture and report specific data points from each of the departments in the organization, providing a snapshot of current performance and a comparison with earlier performance.

Benefits of dashboards include the following:

- Visual presentation of performance measures
- Ability to identify and correct negative trends
- Measurement of efficiencies/inefficiencies
- Ability to generate detailed reports showing new trends
- Ability to make more informed decisions based on collected data
- Alignment of strategies and organizational goals
- Instant visibility of all systems in total
- Quick identification of data outliers and correlations
- Time-saving with the comprehensive data visualization as compared to running multiple reports.

Short Question and Answers

1. What is Statistics

Ans :

Statistics is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.

According to Prof Horace Secrist

Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.

Descriptive statistics employs a set of procedures that make it possible to meaningfully and accurately summarize and describe samples of data. In order for one to make meaningful statements about psychological events, the variable or variables involved must be organized, measured, and then expressed as quantities. Such measurements are often expressed as measures of central tendency and measures of variability.

2. Descriptive Statistics

Ans :

Descriptive Statistics

Descriptive statistics is used to summarize data and make sense out of the raw data collected during the research. Since the data usually represents a sample, then the descriptive statistics is a quantitative description of the sample.

The level of measurement of the data affects the type of descriptive statistics. Nominal and ordinal type data (often termed together as categorical type data) will differ in the analysis from interval and ratio type data (often termed together as continuous type data).

Descriptive statistics for categorical data

Contingency tables (or frequency tables) are used to tabulate categorical data. A contingency table shows a matrix or table between independent variables at the top row versus a dependent variable on the left column, with the cells indicating the frequency of occurrence of possible combination of levels. (check SPSS for examples).

Descriptive statistics for continuous data

There are two the two aspects of descriptive statistics used for continuous type data. They are;

- Central tendency
- Variability of the data.

3. What is Data visualization?

Ans :

It is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

- Today's data visualization tools go beyond the standard charts and graphs used in Microsoft Excel spreadsheets, displaying data in more sophisticated ways such as info graphics, dials and gauges, geographic maps, spark lines, heat maps, and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

4. Cross Tabulations

Ans :

Cross tabulation is usually performed on categorical data - data that can be divided into mutually exclusive groups.

An example of categorical data is the region of sales for a product. Typically, region can be divided into categories such as geographic area (North, South, Northeast, West, etc) or state (Andhra Pradesh, Rajasthan, Bihar, etc). The important thing to remember about categorical data is that a categorical data point cannot belong to more than one category.

Cross tabulations are used to examine relationships within data that may not be readily apparent. Cross tabulation is especially useful for studying market research or survey responses. Cross tabulation of categorical data can be done with through tools such as SPSS, SAS, and Microsoft Excel.

5. Benefits of cross tabulation ?

Ans :

i) Eliminates confusion while interpreting data

Raw data can be difficult to interpret. Even for small data sets, it is all too easy to derive wrong results by just looking at the data. Cross tabulation offers a simple method of grouping variables, which minimizes the potential for confusion or error by providing clear results.

ii) Helps in deriving innumerable insights

As we observed in our example, cross tabulation can help us derive great insights from raw data. These insights are not easy to see when the raw data is formatted as a table. Since cross tabulation clearly maps out relations between categorical variables, researchers can gain better and deeper insights — insights that otherwise would have been overlooked or would have taken a lot of time to decode from more complicated forms of statistical analysis.

iii) Offers data points to chart out a course of action

Cross tabulation makes it easier to interpret data, which is beneficial for researchers who have limited knowledge of statistical analysis. With cross tabulation, people do not need statistical programming to correlate categorical variables. The clarity offered by cross tabulation helps professionals evaluate their current work and chart out future strategies.

6. What is data dashboard.

Ans :

A data dashboard is an information management tool that visually tracks, analyzes and displays key performance indicators (KPI), metrics and key data points to monitor the health of a business, department or specific process. They are customizable to meet the specific needs of a department and company. Behind the scenes, a dashboard connects to your files, attachments, services and API's, but on the surface displays all this data in the form of tables, line charts, bar charts and gauges. A data dashboard is the most efficient way to track multiple data sources because it provides a central location for businesses to monitor and analyze performance. Real-time monitoring reduces the hours of analyzing and long line of communication that previously challenged businesses.

7. Importance of data visualization

Ans :

- Data visualization has become the de facto standard for modern business intelligence (BI). The success of the two leading vendors in the BI space, Tableau and Qlike - both of which heavily emphasize visualization - has moved other vendors toward a more visual approach in their software. Virtually all BI software has strong data visualization functionality.
- Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. They are typically easier to operate than traditional statistical analysis software or earlier versions of BI software. This has led to a rise in lines of business implementing data visualization tools on their own, without support from IT.
- Data visualization software also plays an important role in big data and advanced analytics projects. As businesses accumulated massive troves of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.
- Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

8. Explain the uses of data visualization.

Ans :

- By using data visualization, it became easier for business owners to understand their large data in a simple format.
- The visualization method is also time saving. So, businesses does not have to spend much time to make a report or solve a query. They can easily do it in a less time and in a more appealing way.
- Visual analytics offers a story to the viewers. By using charts and graphs or images, a person can easily exposure the whole concept as well the viewers will be able to understand the whole thing in an easy way.
- The most complicated data will look easy when it gets through the process of visualization. Complicated data report gets converted into a simple format. And it helps people to understand the concept in an easy way.
- With the visualization process, it gets easier to the business owners to understand their product growth and market competition in a better way.

9. Gantt Chart

Ans :

A Gantt Chart is a chart in which a series of horizontal lines shows the amount of work done in certain periods of time in relation to the amount of work planned for those periods.

In Excel, you can create a Gantt Chart by customizing a Stacked Bar Chart type so that it depicts tasks, task duration and hierarchy. An Excel Gantt Chart typically uses days as the unit of time along the horizontal axis.

Consider the following data where the column:

- Task represents the Tasks in the project
- Start represents number of days from the Start Date of the project
- Duration represents the duration of the Task.

10. What are the benefits of data dash boards.

Ans :

Benefits of dashboards include the following:

- Visual presentation of performance measures
- Ability to identify and correct negative trends
- Measurement of efficiencies/inefficiencies
- Ability to generate detailed reports showing new trends
- Ability to make more informed decisions based on collected data
- Alignment of strategies and organizational goals
- Instant visibility of all systems in total
- Quick identification of data outliers and correlations
- Time-saving with the comprehensive data visualization as compared to running multiple reports.

Choose the Correct Answer

1. An ideal measure of central tendency is [a]

   (a) Arithmetic mean (b) Moving average

   (c) Median (d) Harmonic Mean

2. Mathematical average is called [a]

   (a) Arithmetic mean (b) Geometric mean

   (c) Mode (d) None of these

3. Sum of deviations of the items is zero from [a]

   (a) Mean (b) Median

   (c) Mode (d) Geometric mean

4. A ________ is a characteristic that takes different values at different times, places or situations. [d]

   (a) Attributes (b) Data

   (c) Statistics (d) Variable

5. Measure of central tendency ________ [d]

   (a) Mean (b) Mode

   (c) Median (d) All

6. Which of the following measures of central tendency will always change if a single value in the data changes? [a]

   (a) Mean (b) Median

   (c) Mode (d) All of these

7. If a positively skewed distribution has a median of 50, which of the following statement is true? [c]

   (a) Mean is greater than 50 (b) Mean is less than 50

   (c) Both (a) and (c) (d) Mode is greater than 50

8. If the variance of a dataset is correctly computed with the formula using (n - 1) in the denominator, which of the following option is true? [c]

   (a) Dataset is a sample

   (b) Dataset is a population

   (c) Dataset could be either a sample or a population

   (d) Dataset is from a census

9. The difference between the highest and the lowest value of the observations in a data is called: [b]

   (a) Mean (b) Range

   (c) Total frequency (d) Sum of observation

10. The range of the data: 6,14,20,16,6,5,4,18,25,15, and 5 is [b]

   (a) 4 (b) 21

   (c) 25 (d) 20

Fill in the blanks

1. __________ is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.

2. __________ statistics is used to summarize data and make sense out of the raw data collected during the research.

3. __________ It is the midpoint of a distribution of data.

4. __________ stands for business integer.

5. __________ reports using Ms.Excel can display in a number of ways.

6. __________ is usually performed on categorical data that can be divided into mutually exclusive groups.

7. __________ Charts are used to graphically summarize data and explore complicated data.

8. KPI stands for __________ .

9. __________ Represent the information in Rows and Columns.

10. __________ allow managers to monitor the contribution of the various departments in the organization.

Answers

1. Statistics
2. Descriptive
3. Median
4. BI
5. Data analysis
6. Cross tabulation
7. Pivot
8. Key perofrmance indicators
9. Tables
10. Dashboards

UNIT III

PREDICTIVE ANALYTICS

Trend Lines, Regression Analysis –Linear & Multiple, Forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Classification, Association, Cause Effect Modeling.

3.1 Trend Lines

Q1. What is predictive analysis ?

Ans :

Definition : Predictive Analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. It does not tell you what will happen in the future.

Predictive analytics is the branch of the Advanced Analytics which is used to make predictions about unknown future events. It uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future.

Predictive Analytics uses historical data to predict future events. Typically, historical data is used to build a mathematical model that captures important trends. That predictive model is then used on current data to predict what will happen next, or to suggest actions to take for optimal outcomes.

Q2. How predictive analysis works ?

Ans :

Predictive Analytics is the process of using data analytics to make predictions based on data. This process uses data along with analysis, statistics and machine learning techniques to create a predictive model for forecasting future events.

The term “predictive analytics” describes the application of a statistical or machine learning technique to create a quantitative prediction about the future. Frequently, supervised machine learning techniques are used to predict a future value (How long can this machine run before requiring maintenance?) or to estimate a probability (How likely is this customer to default on a loan?).

Predictive Analytics starts with a business goal to use data to reduce waste, save time or cut costs. The process harnesses heterogeneous, often massive, data sets into models that can generate clear, actionable outcomes to support achieving that goal, such as less material waste, less stocked inventory, and manufactured product that meets specifications.

Q3. Explain briefly about Trend Analysis.

Ans :

A trend line is a straight line connecting a number of points on a graph. It is used to analyze the specific direction of a group of values set in a presentation. There are two kinds of trend lines, an uptrend with values going higher, and a downtrend where the direction of the line gradually drops to the lower values.

i) Predicting the Future

Trend lines allow businesses to see the difference in various points over a period of time. This helps foretell the possible path the values will take in the future. This can help reveal perform Predictive Trend Line. The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/metrics based on historical data. See the referenced wiki page on the regression analysis for more information business departments, such as sales.

By knowing how to add a trend line to your presentation, you can create a graphical representation of the values you have computed. This will enable the user to easily comprehend and analyze the message you are trying to imply.

ii) Predictive Trend Line

The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/metrics based on historical data. See the referenced wiki page on the regression analysis for more information.

3.2 Regression Analysis

Q4. Explain the concept of regression analysis.

Ans :

Regression takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents. His major conclusion was that the off-springs of unusually tall persons tend to be shorter than their parents while children of usually short parents tend to be shorter than their parents.

In a sense, the successive generations of off-springs from tall persons "regress" downward toward the height of the population. While the reverse is true originally for short families. But the distribution of heights for the total population, continues to have the same variability from generation to generation.

Regression Variables

i) **Independent Variable (Regressor or Predictor or Explanatory).** The variable which influences the values of the other variable or which is used for prediction of the value of the other variable is called independent variable.

ii) **Dependent Variable (Regressed or Explained Variable).** The variable whose value is influenced or is to be predicted is called dependent variable.

Regression test generates lines of regression of the two variables which helps in estimating the values. Lines of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x. Similarly, line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Types of Regression

a) **Simple Regression.** The regression analysis confined to the study of only two variables at a time is termed as simple regression.

b) **Multiple Regression.** The regression analysis for studying more than two variables at a time is termed as multiple regression.

c) **Linear Regression.** If the regression curve is a straight line, the regression is termed as linear regression. The equation of such a curve is the equation of a straight line i.e., first degree equation in variables x and y.

d) **Nonlinear Regression.** If the curve of the regression is not a straight line, the regression is termed as curved or non-linear regression. The regression equation will be a functional relation between variables x and y involving terms in x and y of degree more than one.

Applications / Utility of Regression Test

Regression lines or equations are useful in the predictions of values of one variable for a specified value of the other variable.

Example

i) For pharmaceutical firms which are interested in studying the effect of new drugs in patients, regression test helps in such predictions.

ii) When price and demand are related, we can estimate or predict the future demand for a specified price.

iii) When crop yield depends on the amount of rainfall, then regression test can predict crop yield for a particular amount of rainfall.

iv) If advertising expenditure and sales are related, then regression analysis helps in estimating the advertising expenditure for a required amount of sales (or) sales expected for a particular advertising expenditure.

v) When capital employed and profits earned are related, the test can be used to predict profits for a specified amount of capital invested.

Q5. Explain the limitations of regression analysis.

Ans :

Limitations of Regression Analysis

Some of the limitations of regression analysis are as follows :

1. Regression analysis assumes that linear relationship exists among the related variables. But in the ‘area of social sciences, linear relationship may not exist among the related variables.

2. When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between ‘them. These statistic conditions do not exist in social sciences, so, this assumption “minimizes the use’ of regression analysis in social science.

3. The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

Even though, there are many limitations of regression 'technique, it is still regarded as a very useful statistical tool for estimating or predicting the value of dependent variable.

Q6. Explain about regression equation.

Ans :

Regression is mainly concerned with the estimation of unknown value of one variable from the known value of other variable of the given observations. For doing so, there must be a relation between two variables. This relationship is mathematically expressed in the form of equation known as "Regression Equation " or " Estimating Equation".

The regression equation which states and explains the linear relationship between two variables is known as 'Linear Regression Equation'. Basically, as there are two regression lines, there would be two regression equations i.e.,

1. Regression equation of K on X and
2. Regression equation of X on Y.

The regression equation of Y on X is considered for predicting the value of Y when a specific value of X is given. Whereas the regression equation of X on Y is used for predicting the unknown value of X when a specific value of Y is given.

Formation of Regression Equations

There are two ways of forming regression equations as follows,

a) Normal equation and

b) Regression coefficient.

Formation of Regression Equation through Normal Equation

Generally, the situations where perfect linear relationship exists between the two variables X and Y, usually there would be two regression lines and when there are two regression lines, there would be two regression equations as follows,

1. The regression equation of Y on X is denoted as $Y. = a + bX$.
2. The regression equation of X on Y is denoted as $X = a + bY$.

In the above equations 'a' and 'b' are two unknown constants which ascertains the positions of the regression line. Therefore, these constants are known as parameters of the regression lines.

The parameter 'a' ascertains the level of a fitted line, whereas 'b' ascertains the slope of the line. Y_C and X_C are the symbols stating and showing the values of Y and X calculated from the relationship for given X or Y.

Regression Equation of Y on X

$Y = a + bX$

By applying the least square principle, the values of 'a' and 'b' are determined in such a way $Y_C = a + bX$ is minimum.

The normal equation for determining the value of a and b are,

$\Sigma y = Na + b\Sigma x$ ...(1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ ...(2)

Regression Equation of X on Y

$Y_c = a + by$

The normal equation for obtaining the values of a and b are,

$\Sigma x = Na + b\Sigma y$ ...(1)

$\Sigma xy = a\Sigma y + b\Sigma y^2$ ... (2)

After calculating the values of N, Σx, Σy, Σx^2, "Σy^2, substitute them in regression equation Y on X and X on Y for ascertaining the values of a and b. Lastly, by substituting the values of a and b in regression equation, the required best fitting straight line is obtained.

b) Regression Coefficients

To estimate values of population parameter β_0 and β_1, under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as :

$$\hat{y} = a + bx$$

where

y = estimated average (mean) value of dependent variable y for a given value of independent variable *x*.

a or b_0 = y - intercept that represents average value of $\hat{y}$

b = slope of regression line that represents the expected change in the value of y for unit change in the value of *x*

To determine the value of ŷ for a given value of *x*, this equation requires the determination of two unknown constants a (intercept) and *b* (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable y for a given value of independent variable *x*.

The particular values of *a* and *b* define a specific linear relationship between *x* and *y* based on sample data. The coefficient '*a*' represents the level of fitted line (i.e., the distance of the line above or below the origin) when *x* equals zero, whereas coefficient '*b*' represents the slope of the line (a measure of the change in the estimated value of *y* for a one-unit change in).

The regression coefficient 'b' is also denoted as :

➢ b_{yx} (regression coefficient of y on x) in the regression line, y = a + bx

➢ b_{xy} (regression coefficient of x on y) in the regression line, x = c + dy.

Q7. Discuss briefly about simple regression.

Ans :

Simple regression represents the relationship between two variables where one of them is independent variables *'X'* and other variable is dependent variable 'Y'.

The relationship between two variables can be of three types. They are,

i) Linear Relationship

The graph of linear relationship between two variables looks as follows,

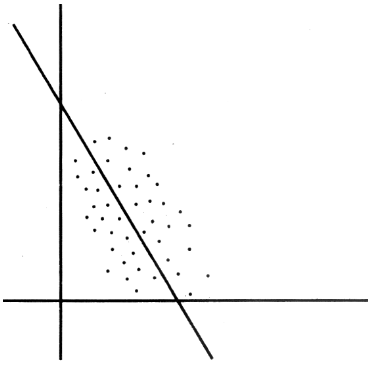


Fig. : Linear Relationship

ii) Non-Linear Relationship

The graph of non-linear relationship between two variables looks as follows,

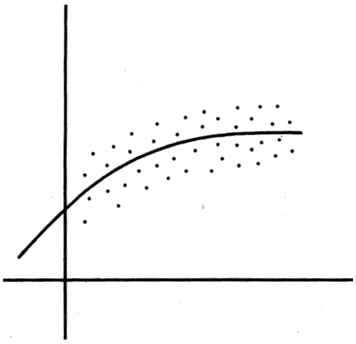


Fig. : Non-linear Relationship

iii) No Relationship

The graph of no relationship between two variables looks as follows,

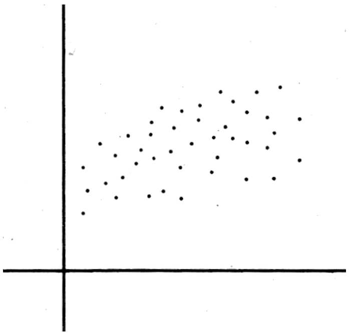


Fig. : No Relationship

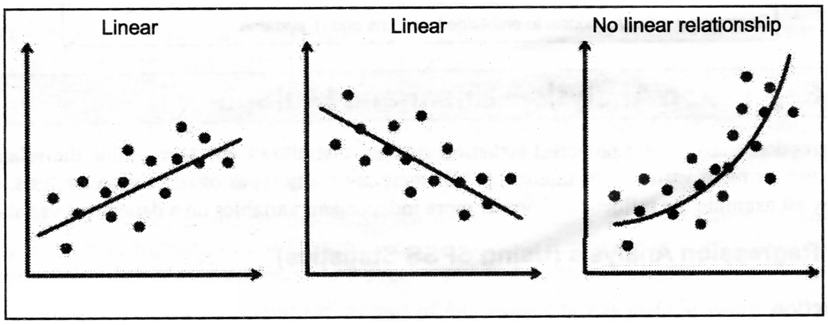
3.2.1 Linear Regression

Q8. Explain the assumptions of simple linear regressions ?

Ans :

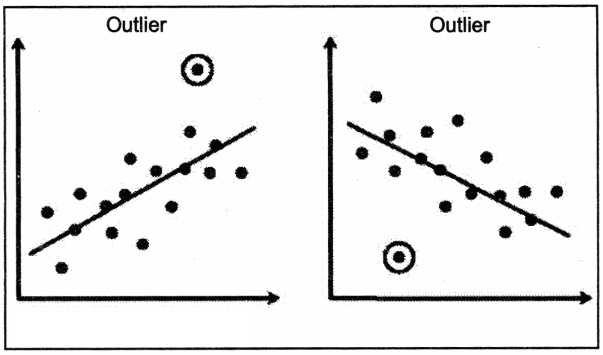
- **Assumption #1:** Two variables should be measured at the continuous level (i.e., they are either interval or ratio variables). Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article: Types of Variable.

- **Assumption #2:** There needs to be a linear relationship between the two variables. Whilst there are a number of ways to check whether a linear relationship exists between your two variables, we suggest creating a scatterplot using SPSS Statistics where you can plot the dependent variable against your independent variable and then visually inspect the scatterplot to check for linearity. Your scatteiplot may look something like one of the following:



If the relationship displayed in your scatterplot is not linear, you will have to either run a non-linear regression analysis, perform a polynomial regression or "transform" your data, which you can do using SPSS Statistics. We show you how to: (a) create a scatterplot to check for linearity when carrying out linear regression using SPSS Statistics; (b) interpret different scatterplot results; and (c) transform your data using SPSS Statistics if there is not a linear relationship between your two variables.

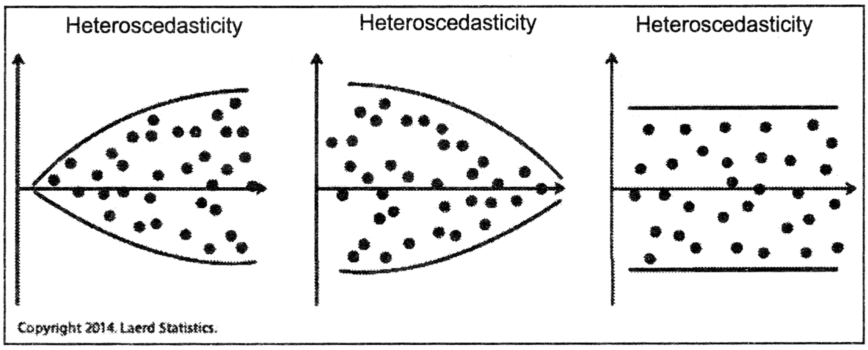
- **Assumption #3:** There should be no significant outliers. An outlier is an observed data point that has a dependent variable value that is very different to the value predicted by the regression equation. As such, an outlier will be a point on a scatterplot that is (vertically) far away from the regression line indicating that it has a large residual, as highlighted below:



The problem with outliers is that they can have a negative effect on the regression analysis (e.g., reduce the fit of the regression equation) that is used to predict the value of the dependent (outcome) variable based on the independent (predictor) variable. This will change the output that SPSS Statistics produces and reduce the predictive accuracy of your results. Fortunately, when using SPSS Statistics to run a linear regression on your data, you can easily include criteria to help you detect possible outliers. We: (a) show you how to detect outliers using "case-wise diagnostics", which is a simple process when using SPSS Statistics; and (b) discuss some of the options you have in order to deal with outliers.

- **Assumption #4:** We should have independence of observations, which you can easily check using the Durbin-Watson statistic, which is a simple test to run using SPSS Statistics.

- **Assumption #5:** Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line. Whilst we explain more about what this means and how to assess the homoscedasticity of your data take a look at the three scatterplots below, which provide three simple examples: two of data that fail the assumption (called heteroscedasticity) and one of data that meets this assumption (called homoscedasticity):



Copyright 2014. Laerd Statistics.

Whilst these help to illustrate the differences in data that meets or violates the assumption of homoscedasticity, real-world data can be a lot more messy and illustrate different patterns of heteroscedasticity. Therefore, we explain: (a) some of the things you will need to consider when interpreting your data; and (b) possible ways to continue with your analysis if your data fails to meet this assumption.

- **Assumption #6:** Finally, you need to check that the residuals (errors) of the regression line are approximately normally distributed Two common methods to check this assumption include using either a histogram (with a superimposed normal curve) or a Normal P-P Plot. Again, we: (a) show you how to check this assumption using SPSS Statistics, whether you use a histogram (with superimposed normal curve) or Normal P-P Plot; (b) explain how to interpret these diagrams; and (c) provide a possible solution if your data fails to meet this assumption.

You can check assumptions #2, #3, #4, #5 and #6 using SPSS Statistics. Assumptions #2 should be checked first, before moving onto assumptions #3, #4, #5 and #6. We suggest testing the assumptions in this order because assumptions #3, #4, #5 and #6 require you to run the linear regression procedure in SPSS Statistics first. So, it is easier to deal with these after checking assumption #2. Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a linear regression might not be valid.

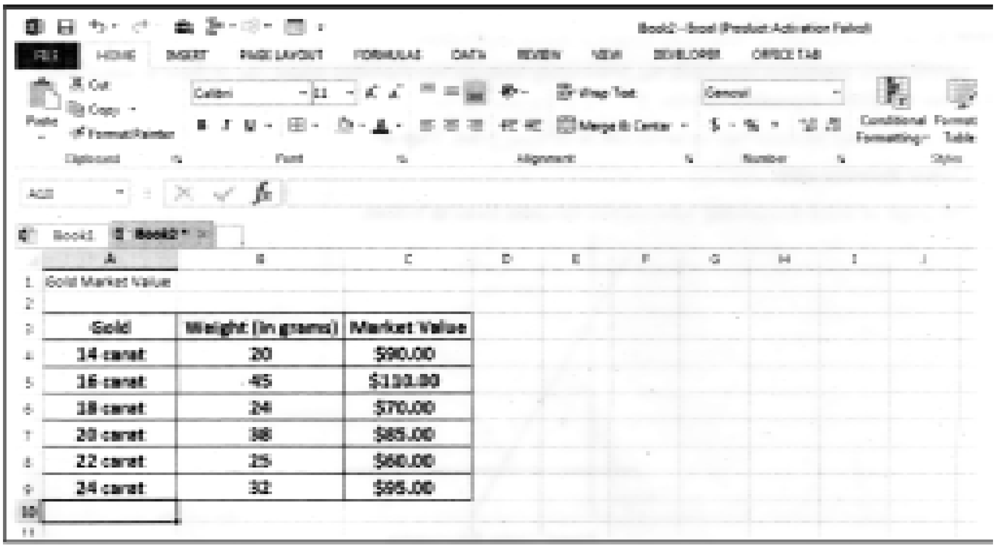
Q9. Explain the concept of simple linear regression by using MS Excel.

Ans :

In Microsoft Excel, the information regarding statistical properties of regression analysis are provided by the software tools of regression analysis. The regression tool can be used not only for simple regression but, also for multiple regression.

The steps to be followed for generating regression analysis output are as follows,

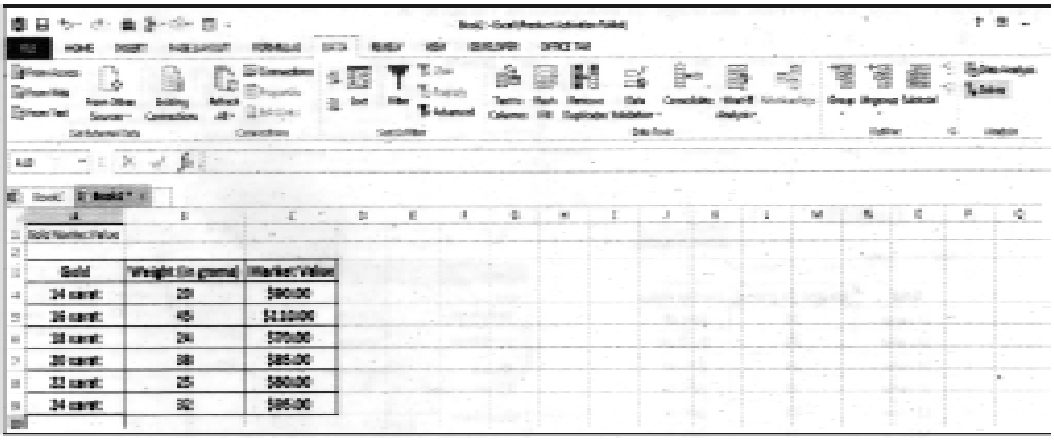
1. Select the data wherein user want to apply regression.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

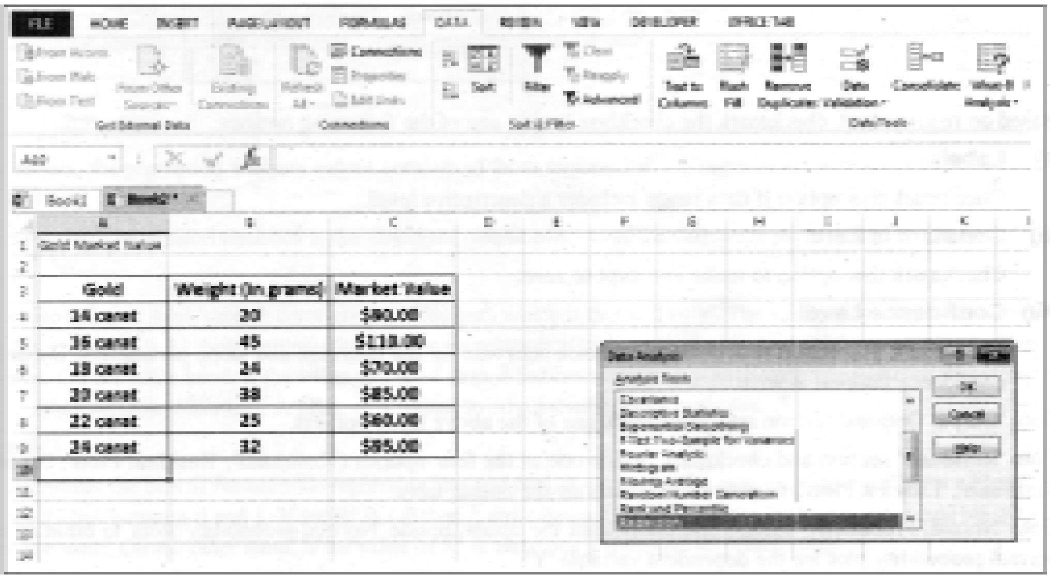
2. Click on ‘Data’ tab present on excel ribbon and the click ‘Data Analysis’ command present under’ Analysis’ group.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$80.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

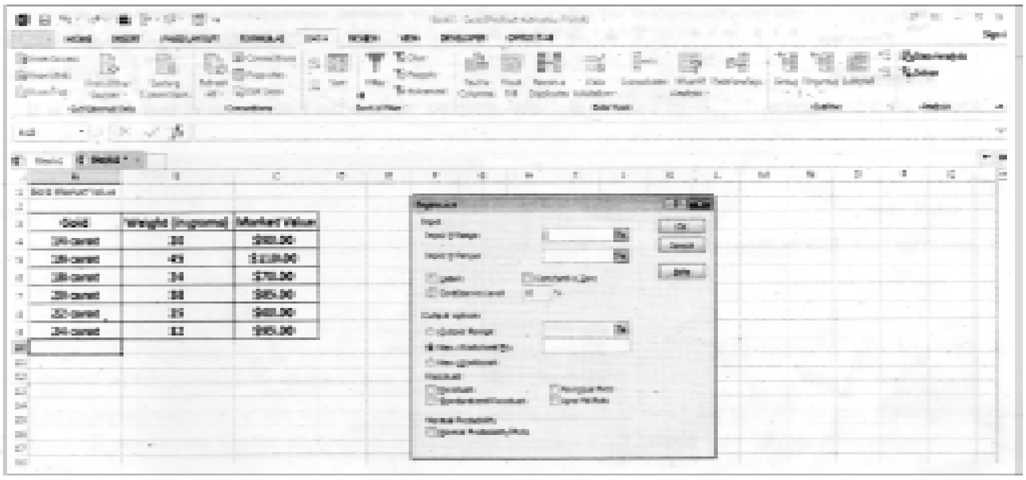
3. As a result, ‘Data Analysis’ dialog box will be displayed. Goto ‘Analysis Tools’ section and select ‘Regression’ option from the menu list and then click “OK” button.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$80.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

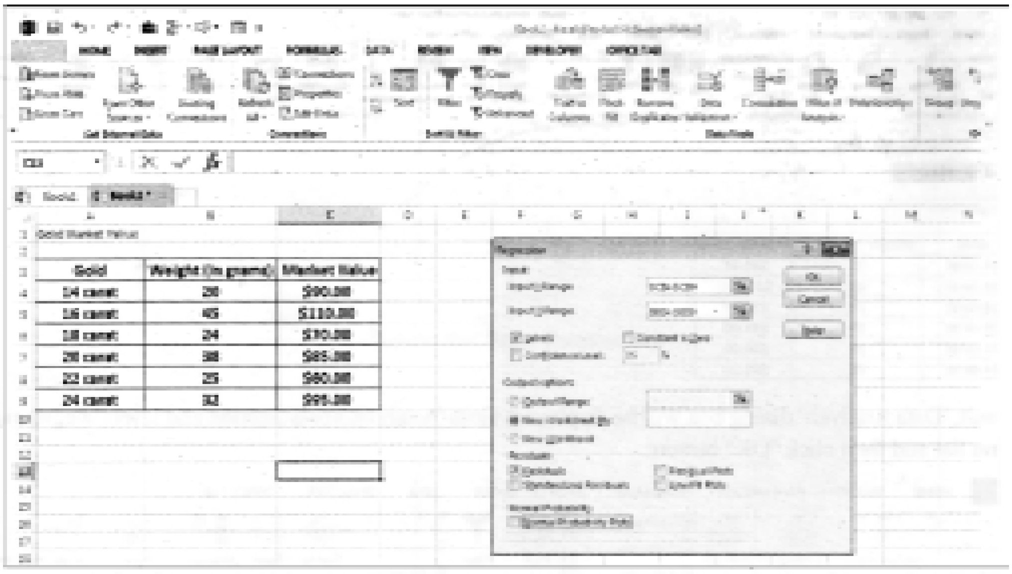
4. As a result, 'Regression' window appears on screen.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

5. In the 'Regression' dialog box, goto 'Input Y Range' field and provide the range of dependent variable 'Y'. Similarly, Goto 'Input X Range' field and provide the range of independent variable 'X'.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

6. Based on requirement, checkmark the checkbox beside one of the following options.

   i) **Labels :** Checkmark this option if data range includes a descriptive level.

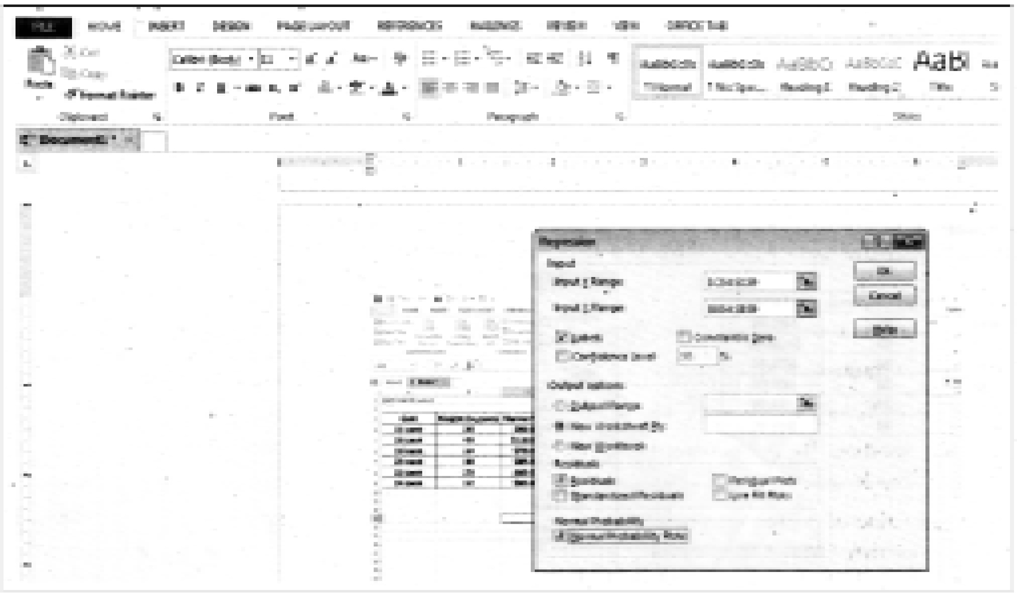
   ii) **Constant is Zero :** Checkmark this option to make intercept to zero.

   iii) **Confidence Level :** Checkmark this option to include confidence intervals for the intercept and slope parameters. By default, the confidence interval is 95%.

7. Goto 'Output Options' section and checkmark one of the above three options.

8. Goto 'Residuals' section and checkmark beside one of the four options ('Residuals', 'Residual Plots', 'Standardized Residuals', 'Line Fit Plots') to provide residuals on the output table.

9. Goto 'Normal Probability' section and checkmark the option beside 'Normal probability plots' to build or construct normal probability plot for the dependent variable 'Y'.



Regression

Input

Input Y Range: [illegible]

Input X Range: [illegible]

☑ Labels ☐ Constant is Zero

☐ Confidence Level 95 %

Output options

○ Output Range

● New Worksheet Ply

○ New Workbook

Residuals

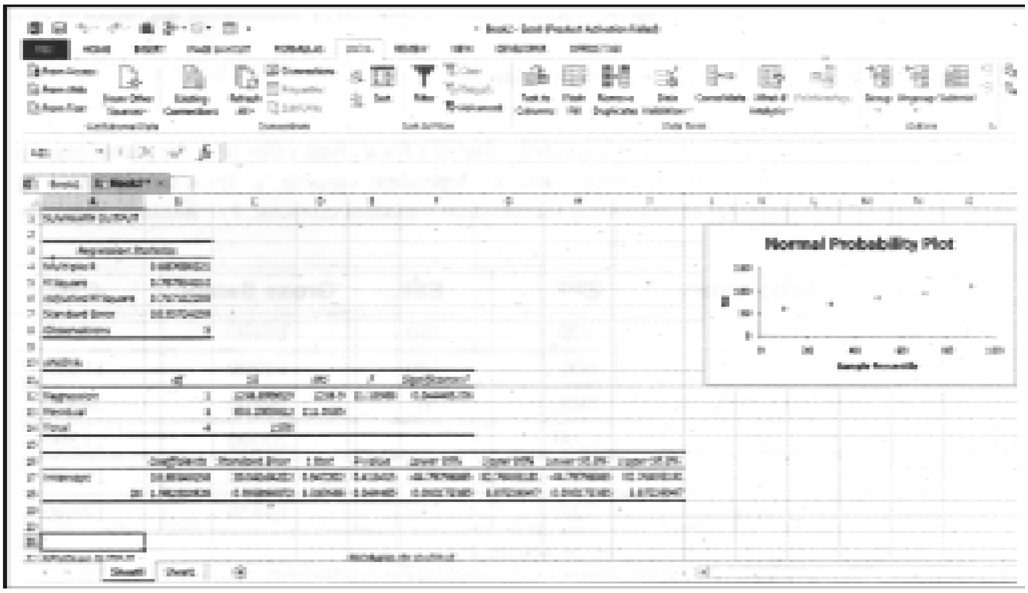
☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☑ Normal Probability Plots

10. Click on "OK" button. As a result, the regression analysis output will be displayed on the screen.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	[illegible]
R Square	[illegible]
Adjusted R Square	[illegible]
Standard Error	[illegible]
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	[illegible]	[illegible]	[illegible]	[illegible]
Residual	3	[illegible]	[illegible]		
Total	4	[illegible]			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

As shown, the regression analysis output consists of three regions namely regression statistics, Annova and unlabelled section.

The region of regression statistics in the displayed output consists of the following parameters,

i) Multiple R

It is also referred to as 'sample correlation coefficient', which is denoted by 'r'. The value of multiple R' lies between - 1 and +1. If the value of r is +1 then it represents positive correlation, which means that if one variable increases another variable also increases. On the other hand, if the value of r is -1 then it indicates negative correlation, which means that if one variable decreases another variable decreases. A value of 'r' equal to zero indicates no correlation.

ii) R-Square(R^2)

It determines the best fit between the regression line and data. R-sqaure is also referred to as 'coefficient of determination'. The value of R^2 lies between 0 and 1. If the R^2 is 1.0 then it indicates perfect fit where in each and every data point falls on the regression line itself. On the other hand, if the value of R^2 is zero then it indicates no relationship.

iii) Adjusted R Square

It refers to a statistical measure that includes in the model not only the sample size, but also the number of independent variables for modifying the value of R^2.

iv) Standard Error

It is also referred to as 'standard error' of the estimate, which is denoted by 'S_{YX}'. It is responsible for describing the variability in 'Y'.

3.2.2 Multiple Regression

Q10. Define multiple regression ? Explain its assumptions.

Ans :

Multiple regression also allows you to determine the overall fit (variance explained) of the model and me relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

- **Assumption #1:** Your dependent variable should be measured on a continuous scale (i.e., it is either an interval or ratio variable). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article: Types of Variable. If your dependent variable was measured on an ordinal scale, you will need to carry out ordinal regression rather than multiple regression. Examples of ordinal variables include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-point scale explaining how much a customer liked a product, ranging from "Not very much" to "Yes, a lot"). You can access our SPSS Statistics guide on ordinal regression here.

- **Assumption #2:** You have two or more independent variables which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable). For examples of continuous and ordinal variables, see the bullet above. Examples of nominal variables include gender (e.g., 2 groups: male and female), ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth. Again, you can learn more about variables in our article: Types of Variable. If one of your independent variables is dichotomous and considered a moderating variable, you might need to run a Dichotomous moderator analysis.

- **Assumption #3:** You should have independence of observations (i.e., independence of residuals), which you can easily check using the Durbin-Watson statistic, which is a simple test to run using SPSS Statistics. We explain how to interpret the result of the Durbin- Watson statistic, as well as showing you the SPSS Statistics procedure required, in our enhanced multiple regression guide.

- **Assumption #4:** There needs to be a linear relationship between: (a) the dependent variable and each of your independent variables, and (b) the dependent variable and the independent variables collectively. Whilst there are a number of ways to check for these linear relationships, we suggest creating scatterplots and partial regression plots using SPSS Statistics, and then visually inspecting these scatterplots and partial regression plots to check for linearity. If the relationship displayed in your scatterplots and partial regression plots are not linear, you will have to either run a non-linear regression analysis or “transform” your data, which you can do using SPSS Statistics. In our enhanced multiple regression guide, we show you how to: (a) create scatterplots and partial regression plots to check for linearity when carrying out multiple regression using SPSS Statistics; (b) interpret different scatterplot and partial regression plot results; and (c) transform your data using SPSS Statistics if you do not have linear relationships between your variables.

- **Assumption #5:** Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line. We explain more about what this means and how to assess the homoscedasticity of your data in our enhanced multiple regression guide. When you analyze your own data, you will need to plot the studentized residuals against the unstandardized predicted values. In our enhanced multiple regression guide, we explain: (a) how to test for homoscedasticity using SPSS Statistics; (b) some of the things you will need to consider when interpreting your data; and (c) possible ways to continue with your analysis if your data fails to meet this assumption.

- **Assumption #6:** Your data must not show multicollinearity, which occurs when you have two or more independent variables that are highly correlated with each other. This leads to problems with understanding which independent variable contributes to the variance explained in the dependent variable, as well as technical issues in calculating a multiple regression model. Therefore, in our enhanced multiple regression guide, we show you: (a) how to use SPSS Statistics to detect for multicollinearity through an inspection of correlation coefficients and Tolerance/VIF values; and (b) how to interpret these correlation coefficients and Tolerance/VIF values so that you can determine whether your data meets or violates this assumption.

- **Assumption #7:** There should be no significant outliers, high leverage points or highly influential points. Outliers, leverage and influential points are different terms used to represent observations in your data set that are in some way unusual when you wish to perform a multiple regression analysis. These different classifications of unusual points reflect the different impact they have on the regression line. An observation can be classified as more than one type of unusual point. However, all these points can have a very negative effect on the regression equation that is used to predict the value of the dependent variable based on the independent variables. This can change the output that SPSS Statistics produces and reduce the predictive accuracy of your results as well as the statistical significance. Fortunately, when using SPSS Statistics to run multiple regression on your data you can detect possible outliers, high leverage points and highly influential points. In our enhanced multiple regression guide, we: (a) show you how to detect outliers using "case- wise diagnostics" and "studentized deleted residuals", which you can do using SPSS Statistics, and discuss some of the options you have in order to deal with outliers; (b) check for leverage points using SPSS Statistics and discuss what you should do if you have any; and (c) check for influential points in SPSS Statistics using a measure of influence known as Cook's Distance, before presenting some practical approaches in SPSS Statistics to deal with any influential points you might have.

- **Assumption #8:** Finally, you need to check that the residuals (errors) are approximately normally distributed (we explain these terms in our enhanced multiple regression guide). Two common methods to check this assumption include using: (a) a histogram (with a superimposed normal curve) and a Normal P-P Plot; or (b) a Normal Q-Q Plot of the studentized residuals. Again, in our enhanced multiple regression guide, we: (a) show you how to check this assumption using SPSS Statistics, whether you use a histogram (with superimposed normal curve) and Normal P-P Plot, or Normal Q-Q Plot; (b) explain how to interpret these diagrams; and (c) provide a possible solution if your data fails to meet this assumption.

Q11. Discuss briefly the concept of multiple regression using excel.

Ans :

Multiple Regression

The regression analysis for studying more than two variables at a time is termed as multiple regression.

A linear regression model that contains multiple independent variables is referred to as multiple regression model. An equation for multiple regression model is given by,

$$\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \in}$$

In the above equation, β_0, β_1 specifies population parameters, X_1, X_2, ... X_p specifies independent-variables, Y defines dependent variable and '$\in$' defines error term.

The expected value of 'y' for a given value of V can be calculated using the above equation if parameter values of $\beta_0, \beta_1, . . ., \beta_q$ are known. On the other hand, if parameter values are not known then they must be calculated using the sample data.

The estimated regression equation for multiple linear regression can be attained by substituting the values of sample statistics b_0, b_1, ... , b_p in β_0, β_1, ... , β_p.

The estimated regression equation in multiple regression model is,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

In the above equation, y refers to point estimator of expected value of y for a given value of x, the partial regression coefficients b_0, b_1, ... ,b_p indicates the change in the mean value of dependent variable 'y' for a unit increase in the independent variables, while holding the values of remaining independent variables constant. For instance, consider the following excel file containing salary details of employees.

Employee	Dept	Basic Salary	EPF	ESI	Gross Salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EE	6000	720	350	13070	14040

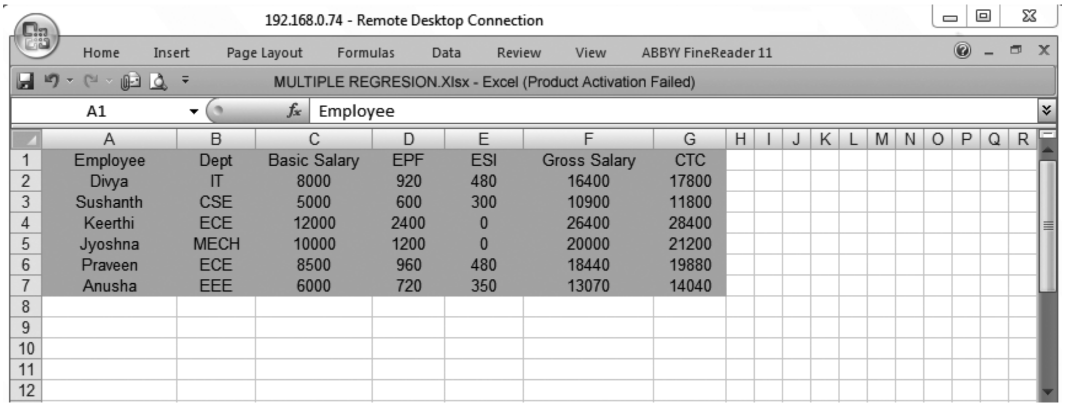
In the above table, the multiple regression model can be written as,

$$\text{CTC} = b_0 + b_x \text{ Basic Salary} + b_2 \text{ EPF} + b_3 \text{ ESI} + b_4 \text{ Gross Salary}$$

Therefore, b, indicates the change in the mean value of CTC for a unit increase in the associated independent variable 'EPF' while holding all remaining independent variables 'Basic Salary', 'EPF', 'ESI' and 'Gross Salary Constant like simple linear regression, multiple linear regression also follows the least squares technique for estimating both intercept and slope coefficients.

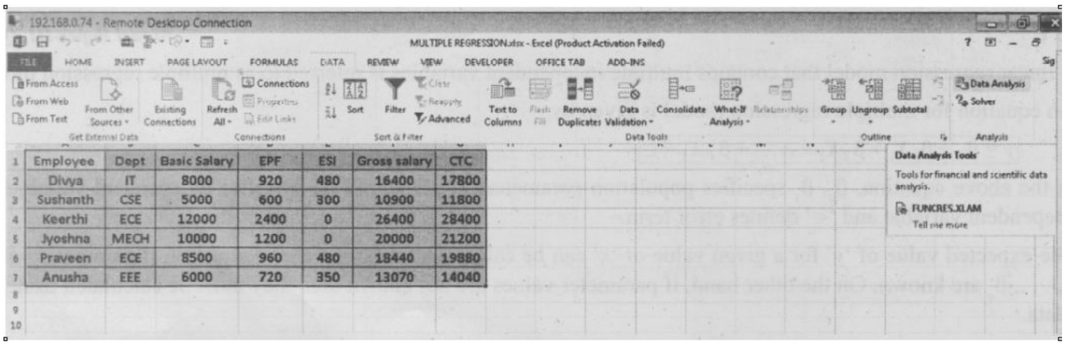
The steps to be followed for generating regression analysis output in case of multiple linear regression are given below,

1. Select the data wherein user want to apply regression.


192.168.0.74 - Remote Desktop Connection
Home
Insert
Page Layout
Formulas
Data
Review
View
ABBYY FineReader 11
MULTIPLE REGRESION.Xlsx - Excel (Product Activation Failed)
A1
Employee

	A	B	C	D	E	F	G
1	Employee	Dept	Basic Salary	EPF	ESI	Gross Salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

2. Click on 'Data' tab present on excel ribbon and the click 'Data Analysis' command present under 'Analysis' group.



Data Analysis Tools

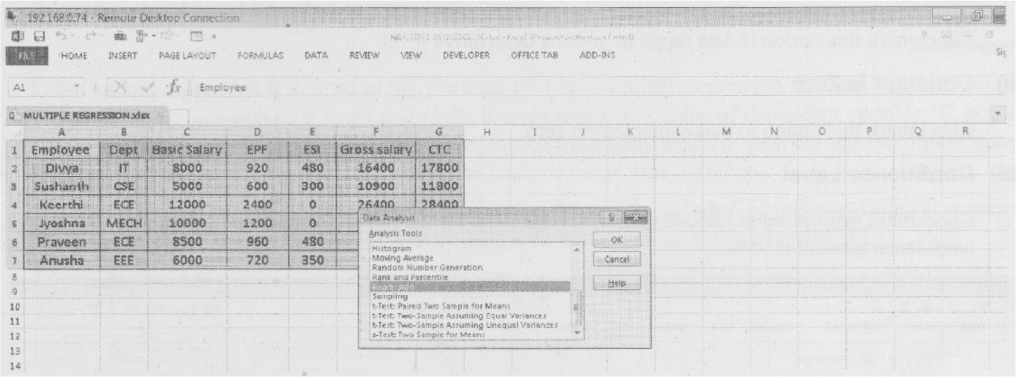
Tools for financial and scientific data analysis.

FUNCRES.XLAM

Tell me more

	Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

3. As a result, 'Data Analysis' dialog box will be displayed. Goto 'Analysis Tools' section and select 'Regression' option from the menu list and then click "OK" button.



192.168.0.74 - Remote Desktop Connection

A1 Employee

MULTIPLE REGRESSION.xlsx

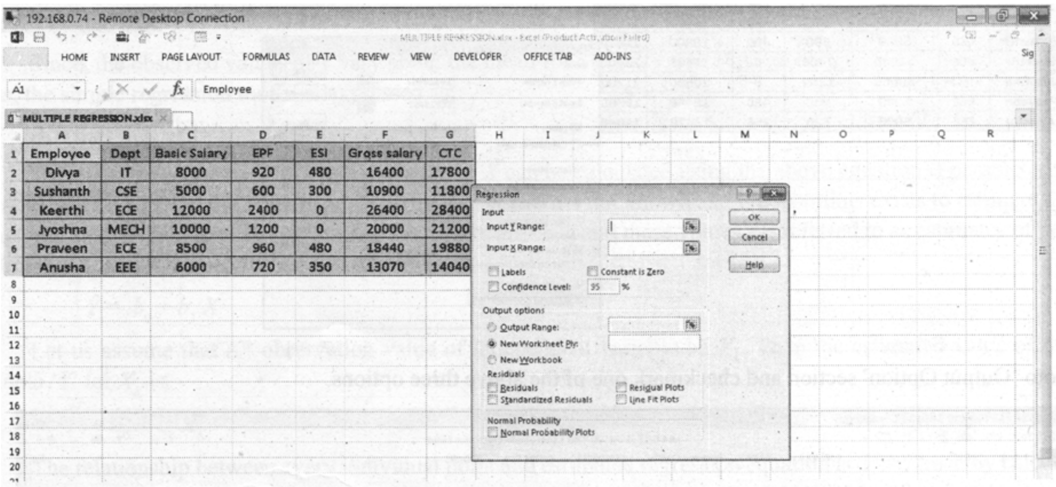
	A	B	C	D	E	F	G
1	Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0		
6	Praveen	ECE	8500	960	480		
7	Anusha	EEE	6000	720	350		

Data Analysis

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

4. As a result, 'Regression' window appears on screen.



192.168.0.74 - Remote Desktop Connection

A1 Employee

MULTIPLE REGRESSION.xlsx

	A	B	C	D	E	F	G
1	Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

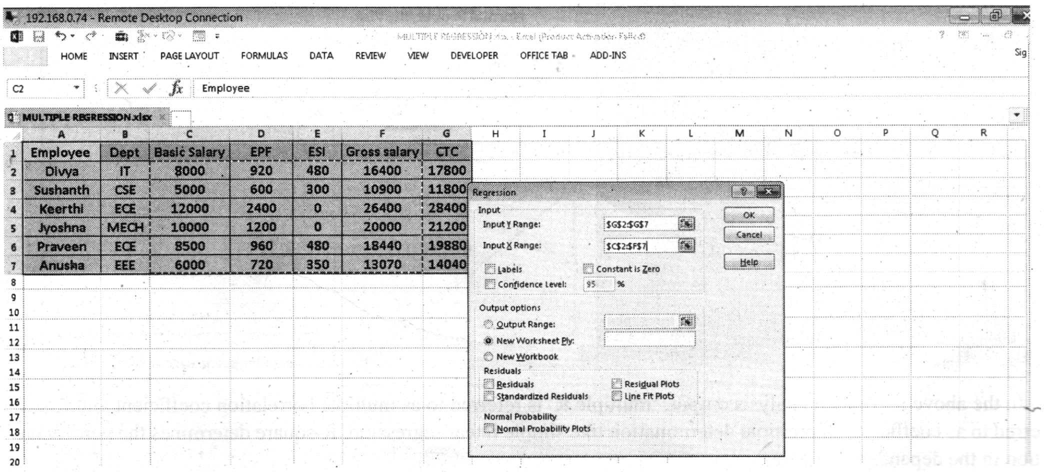
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

5. In the 'Regression', dialog box, Goto 'Input Y Range' field and provide the range of dependent variable Y. Similarly, Goto 'Input X Range' field and provide the entire range of independent variable ˡJC.


192.168.0.74 - Remote Desktop Connection
MULTIPLE REGRESSION.xlsx - Excel (Product Activation Failed)
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
DEVELOPER
OFFICE TAB
ADD-INS
MULTIPLE REGRESSION.xlsx

C2 | Employee

Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

Regression

Input
Input Y Range: $G$2:$G$7
Input X Range: $C$2:$F$7
□ Labels □ Constant is Zero
□ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

Residuals
□ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots

Normal Probability
□ Normal Probability Plots

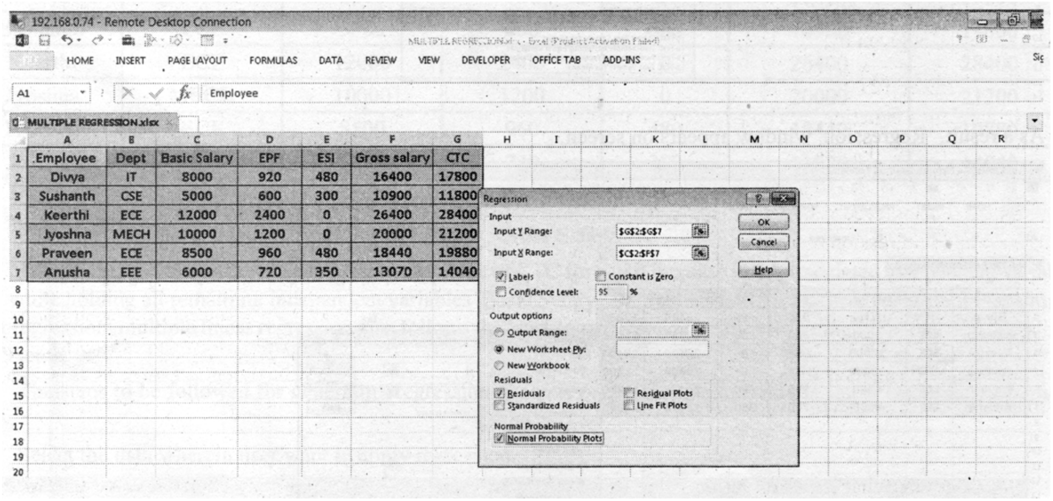
OK Cancel Help

6. Based on requirement, checkmark the checkbox beside one of the following options.

   (i) Labels: Checkmark this option if data range includes a descriptive level.

   (ii) Constant is Zero: Checkmark this option to make intercept to zero.

   (iii) Confidence Level: Checkmark this option to include confidence intervals for the intercept and slope parameters. By default, the confidence interval is 95%.


192.168.0.74 - Remote Desktop Connection
MULTIPLE REGRESSION.xlsx - Excel (Product Activation Failed)
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
DEVELOPER
OFFICE TAB
ADD-INS
MULTIPLE REGRESSION.xlsx

A1 | Employee

Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

Regression

Input
Input Y Range: $G$2:$G$7
Input X Range: $C$2:$F$7
☑ Labels □ Constant is Zero
□ Confidence Level: 95 %

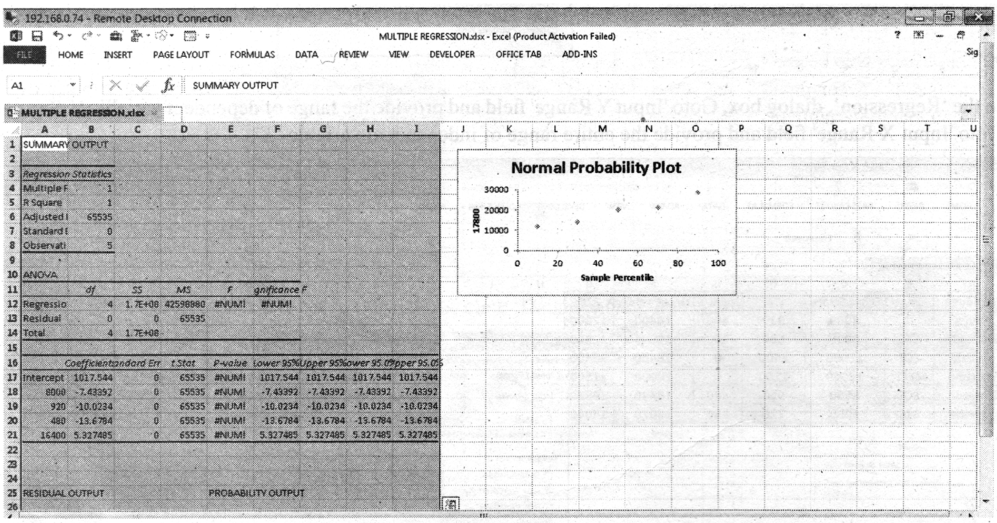
Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

Residuals
☑ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots

Normal Probability
☑ Normal Probability Plots

OK Cancel Help

7. Goto ‘Output Option’ section and checkmark one of the above three options.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	1							
5	R Square	1							
6	Adjusted I	65535							
7	Standard E	0							
8	Observati	5							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	4	1.7E+08	42598880	#NUM!	#NUM!			
13	Residual	0	0	65535					
14	Total	4	1.7E+08						
15									
16		*Coefficient*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	1017.544	0	65535	#NUM!	1017.544	1017.544	1017.544	1017.544
18	8000	-7.43392	0	65535	#NUM!	-7.43392	-7.43392	-7.43392	-7.43392
19	920	-10.0234	0	65535	#NUM!	-10.0234	-10.0234	-10.0234	-10.0234
20	480	-13.6784	0	65535	#NUM!	-13.6784	-13.6784	-13.6784	-13.6784
21	16400	5.327485	0	65535	#NUM!	5.327485	5.327485	5.327485	5.327485
22									
23									
24									
25	RESIDUAL OUTPUT				PROBABILITY OUTPUT				
26									

In the above regression analysis output, ‘multiple R’ is referred to as multiple correlation coefficient and R square is referred to as coefficient of multiple determination like simple linear regression, R-square determines the percentage of variation in the dependent variable.

3.3 Forecasting Techniques

Q12. Explain briefly about Forecasting techniques.

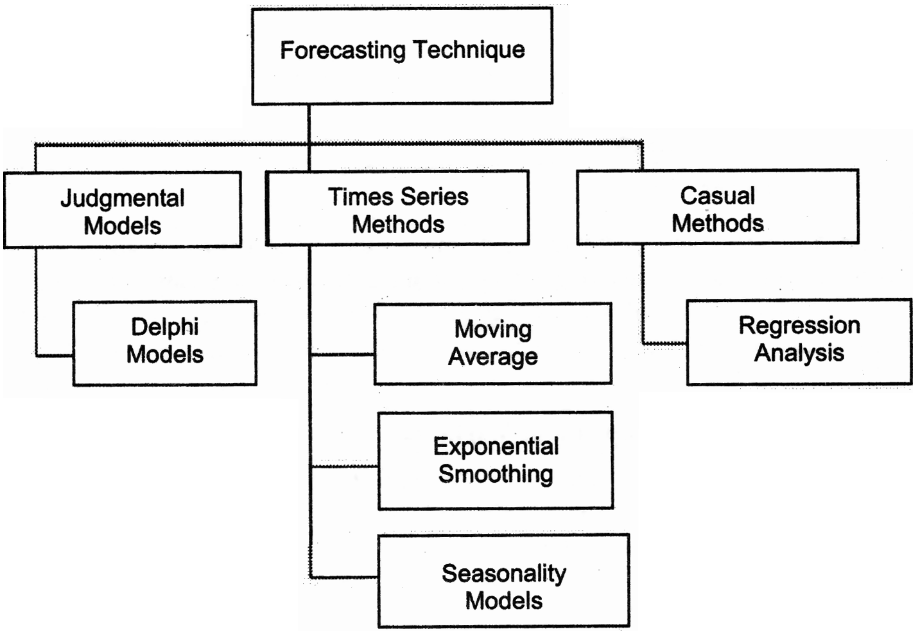
Ans :

Forecasting Analytics is a Quantitative forecasting, which focuses on data for generating numerical forecasts, is an important component of decision making in a wide range of areas and across many business functions, including economic forecasting, workload projections, sales forecasts and transportation demand.

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated.

Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date. Prediction is a similar, but more general term. Both might refer to formal statistical methods employing time series, cross-sectional or longitudinal data, or alternatively to less formal judgmental methods. Usage can differ between areas of application: for example, in hydrology the terms "forecast" and "forecasting" are sometimes reserved for estimates of values at certain specific future times, while the term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period.

Risk and uncertainty are central to forecasting and prediction; it is generally considered good practice to indicate the degree of uncertainty attaching to forecasts. In any case, the data must be up to date in order for the forecast to be as accurate as possible. In some cases the data used to predict the variable of interest is itself forecasted.



The forecasting method you select is a function of multiple qualities about your item. Is demand steady, cyclical or sporadic? Are there seasonal trends? Are trends strong or limited? Is the item new? Each item being forecast has a somewhat unique history (and future), and therefore an optimal method. A method that accurately forecasts one data set might prove inaccurate for another.

Determining the optimal forecast method is a rather complex science, especially across a large product line. This may be nearly impossible using only spreadsheets.

However, sophisticated forecasting software can within seconds test multiple methods for each item to determine which method will give you the most accurate results.

Specific Forecasting Methods

1. Moving Averages
2. Exponential Smoothing
3. Regression Analysis Models
4. Hybrid Forecasting Methods

1. Moving Averages

Moving average methods take the average of past actuals and project it forward. These methods assume that the recent past represents the future. As a result, they work best for products with relatively little change — steady demand, no seasonality, limited trends or cycles and no significant demand shifts. Many companies apply this method because it is simple and easy to use. However, since few products actually behave in this way, it tends to be less useful than more specialized methods.

2. Exponential Smoothing

Exponential smoothing is a more advanced form of time series forecasting. Unlike moving averages, exponential smoothing methods can capture trends and recurring patterns. They accomplish this by:

- Emphasizing the more recent data (as opposed to a moving average which weights all data equally), and
- Smoothing out fluctuations, which are often caused by pure randomness in the data (or “noise” in the system).

Forecasters determine the forecast weights, controlling how fast or slow the model responds to demand changes in your actuals. Not all exponential smoothing methods can handle seasonality or other recurring patterns.

Exponential smoothing forecasting methods include:

(i) Simple exponential smoothing

(ii) Holt’s linear method

(iii) Winters’ multiplicative season

(iv) Winters’ additive season

3. Regression Analysis Models

Many companies use regression models to determine the relationship between demand and demand drivers. They are especially useful for seeing trends and seasonality.

Regression analysis methods include:

(i) Linear regression

(ii) Hyperbolic trend

(iii) Logarithmic trend

(iv) Square root trend

(v) Quadratic trend

4. Hybrid Forecasting Methods

Hybrid forecasting methods combine regression, data smoothing and other techniques to produce forecasts that can compensate for the weaknesses of individual methods. For example, some forecasting methods are great at short-term forecasting, but cannot capture seasonality.

Hybrid forecasting methods include:

- Vanguard Dampened Trend - a powerful hybrid model that simultaneously detects all trends, cycles and seasonality in historical data and responds with the most accurate exponential smoothing method. Vanguard Dampened Trend is available across all Vanguard business forecasting applications
- Log Theta
- Theta

Causal / Econometric Forecasting Methods

Some forecasting methods try to identify the underlying factors that might influence the variable that is being forecast. For example, including information about climate patterns might improve the ability of a model to predict umbrella sales. Forecasting models often take account of regular seasonal variations. In addition to climate, such variations can also be due to holidays and customs: for example, one might predict that sales of college football apparel will be higher during the football season than during the off season.

Several informal methods used in causal forecasting do not rely solely on the output of mathematical algorithms, but instead use the judgment of the forecaster. Some forecasts take account of past relationships between variables: if one variable has, for example, been approximately linearly related to another for a long period of time, it may be appropriate to extrapolate such a relationship into the future, without necessarily understanding the reasons for the relationship.

Causal methods include

- Regression analysis includes a large group of methods for predicting future values of a variable using information about other variables. These methods include both parametric (linear or non-linear) and non-parametric techniques.
- Autoregressive moving average with exogenous inputs

Judgmental Methods

Judgmental forecasting methods incorporate intuitive judgement, opinions and subjective probability estimates. Judgmental forecasting is used in cases where there is lack of historical data or during completely new and unique market conditions. Artificial intelligence methods

- Artificial neural networks
- Group method of data handling
- Support vector machines

Often these are done today by specialized programs loosely labelled :

- Data mining
- machine learning
- Pattern recognition

Other methods

- Simulation
- Prediction market
- Probabilistic forecasting and Ensemble forecasting.

3.4 Data Mining

3.4.1 Definition of Datamining

Q13. Explain data Mining.

Ans :

Definition

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

- Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
- The difference between data analysis and data mining is that data analysis is to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.
- The term "data mining" is in fact a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence.
- The book Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

Q14. Explain the steps involved in data mining.

Ans :

Steps

1. **Problem Definition**

   A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

   In the problem definition phase, data mining tools are not yet required.

2. **Data Exploration**

   Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

   In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3. **Data Preparation**

   Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

   In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. **Modeling**

   Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

   In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Evaluation : Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

- Does the model achieve the business objective?
- Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

5. Deployment

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

The Intelligent Miner products assist you to follow this process. You can apply the functions of the Intelligent Miner products independently, iteratively, or in combination.

The following figure shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model.

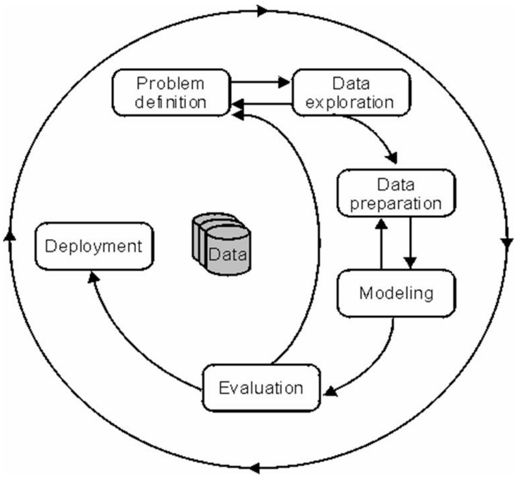


Fig.1: The CRISP DM process model

IM Modeling helps you to select the input data, explore the data, transform the data, and mine the data. With IM Visualization you can display the data mining results to analyze and interpret them. With IM Scoring, you can apply the model that you have created with IM Modeling.

Q15. Explain the scope of data mining ?

Ans :

Scope of Data Mining

1. Data mining process the work in such a manner that it allows business to more proactive to grow substantially.
2. It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.
3. It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.
4. It includes tree-shaped structure to understand the hierarchy of data and representation of the set of information described in the database.
5. Brings a genetic way of classification of different sets of data items to view the data in the quick glance.

Q16. Explain the various techniques used in data mining.

Ans :

1. **Classification:** This analysis is used to retrieve important and relevant information about data and metadata. This data mining method helps to classify data in different classes.
2. **Clustering:** Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.
3. **Regression:** Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. **Association rules:** This data mining technique helps to find the association between two or more items. It discovers a hidden pattern in the data set.

5. **Outer detection:** This type of data mining technique refers to observation of data items in the data set which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier Mining.

6. **Sequential patterns:** This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. **Prediction:** Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

Q17. Explain merits and demerits of data mining.

Ans :

Benefits of Data Mining

- ➢ Data mining technique helps companies to get knowledge-based information.
- ➢ It helps organizations to make the profitable adjustments in operation and production.
- ➢ It is a cost-effective and efficient solution compared to other statistical data applications.
- ➢ It helps with the decision-making process.
- ➢ It facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
- ➢ It can be implemented in new systems as well as existing platforms.
- ➢ It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Disadvantages of Data Mining

- There are chances of companies selling useful information of their customers to other companies for money. For example, American Express has sold credit card purchasesof their customers to the other companies.
- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate. Hence, it can cause serious consequences in certain conditions.

3.5 Approaches in Data Mining

Q18. Explain the various approaches for data mining with Micro Strategy.

Ans :

i) **Scoring the database:** Records are scored in batches and saved as tables or columns.

ii) **Database does the scoring:** The database scores records in response to queries.

iii) **MicroStrategy does the scoring:** MicroStrategy scores records using metrics and reports.

While MicroStrategy supports all three approaches, each has positive and negative aspects. The next sections describe each approach in detail.

i) Scoring the Database

In this approach, records are scored and inserted into the database either as new tables or as new columns in existing tables. Most often, a third-party scoring engine receives a result set and scores the records. Then the scores are added to the database. Once they are part of the database, MicroStrategy attributes or metrics can reference those scores, just like any other data in the database. Historically, this approach has been the most common. Its pros and cons are described below.

Pros

- Since an external scoring engine performs the scoring calculation, model complexity and performance is hidden within the scoring engine. Thus, the scoring process does not require any database resources and does not impact other business intelligence work.
- At run time, data is simply read from the database without having to calculate the score on the fly. Scoring on the fly can slow analysis especially if millions of scores are involved.
- MicroStrategy can use this approach by just creating metrics or attributes for the scored data.

Cons

- This approach requires database space and the support of a database administrator.
- New records that are inserted after the batch scoring are not scored.
- Updating the model or scores requires more database and database administrator overhead.
- In many companies, adding or updating information in the enterprise data warehouse is not done easily or whenever desired. The cross functional effort required to score the database limits the frequency of scoring and prevents the vast majority of users from trying new models or changing existing ones.

This approach is really no different than adding other entities to a MicroStrategy project. For more information, see the Project Design Guide.

ii) Database does the scoring

In this approach, data mining features of the database system are used to perform the scoring. Nearly all major databases have the ability to score data mining models. The most common approach persists the model in the database and then generates scores by using extensions to the SQL queries processed by the database to invoke the model. A key feature of this approach is that the model can be scored in a system that is different from the data mining tool that developed the model.

- The model can be saved in the database as a Predictive Model Markup Language (PMML) object, or, less frequently, in some form of executable code. For more information on PMML, see PMML overview.

- Persisting the model in this way is possible since the sophisticated algorithms needed to create the model are not required to score them. Scoring simply involves mathematical calculations on a set of inputs to generate a result.

- The ability to represent the model and score it outside of the model creation tool is relatively new, but more companies are adopting this approach. Its advantages and disadvantages are described below.

Pros

- Scores can be calculated on the fly even if new records are added.

- Updating the model is easier than in the Score the database option.

- This approach requires less database space than the score the database option.

- When the database supports accessing its data mining features via SQL, MicrovStrategy can take advantage of this approach using its SQL Engine.

Cons

- This approach requires support from a database administrator and application knowledge of the database's data mining tool. However, the database administrator usually does not have this knowledge.

- The database data mining tool is typically an additional cost.

iii) MicroStrategy does the scoring

- In this approach, predictive models are applied from within the Business Intelligence platform environment, without requiring support from the database and from database administrators to implement data mining models. This direct approach reduces the time required, the potential for data inconsistencies, and cross-departmental dependencies.

- MicroStrategy Data Mining Services uses enterprise data resources without significantly increasing the overhead. MicroStrategy Data Mining Services allows sophisticated data mining techniques to be applied directly within the business intelligence environment. Just as the other approaches, it also has advantages and disadvantages, as described below:

Pros

- MicroStrategy stores the predictive model in its metadata as a predictive metric that can be used just like any other metric.
- Scores can be done on the fly even if new records are added.
- The predictive model can be viewed in MicroStrategy Developer.
- The predictive model is easily updated using MicroStrategy Developer.
- This approach does not require database space or support from a database administrator.
- MicroStrategy can take advantage of this approach by using the Analytical Engine.

Cons

- This approach does not take advantage of the database data mining features.
- Predictor inputs need to be passed from the database to Intelligence Server. For large result sets, databases typically handle data operations more efficiently than moving data to MicroStrategy and scoring it there.

3.6 Data Exploration and Reduction

Q19. What is data exploration ? Explain the Steps of Data Exploration and Preparation

Ans :

Data exploration is an informative search used by data consumers to form true analysis from the information gathered. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis.

Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats. Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data.

Steps of Data Exploration and Preparation

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With my personal estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

1. Variable Identification

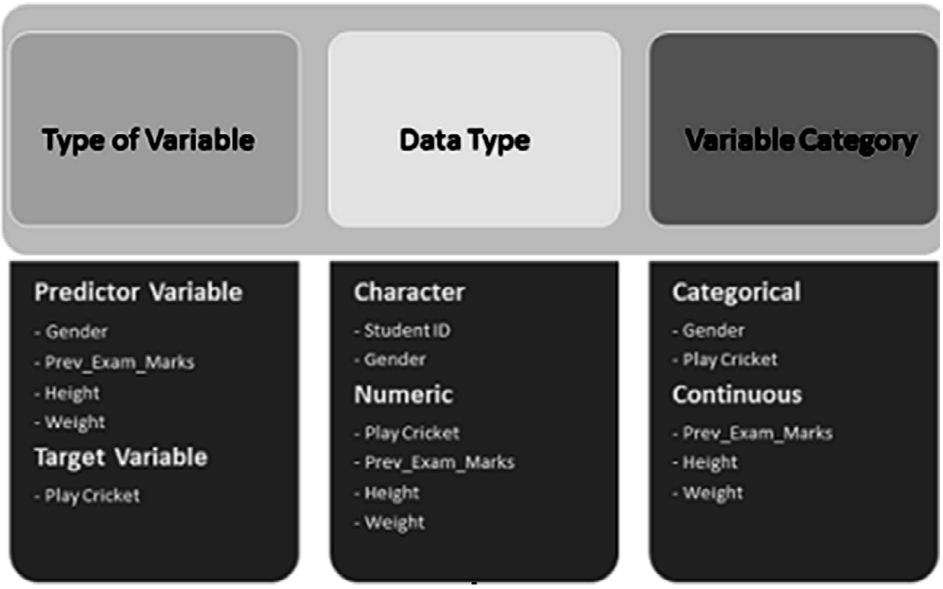
First, identify Predictor (Input) and Target (output) variables. Next, identify the data type and category of the variables.

Let's understand this step more clearly by taking an example.

Example

Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables. Below, the variables have been defined in different category:

Student ID	Gender	Prev Exam Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



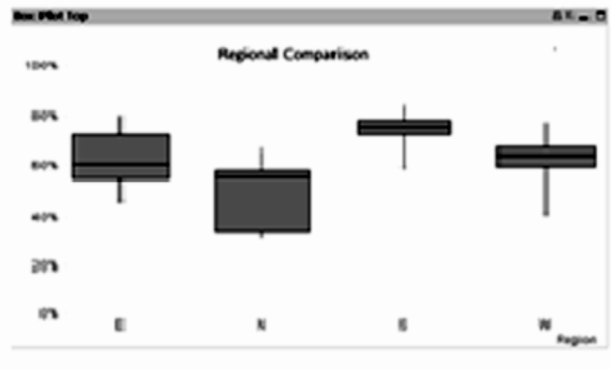
Type of Variable	Data Type	Variable Category
Predictor Variable - Gender - Prev_Exam_Marks - Height - Weight **Target Variable** - Play Cricket	**Character** - Student ID - Gender **Numeric** - Play Cricket - Prev_Exam_Marks - Height - Weight	**Categorical** - Gender - Play Cricket **Continuous** - Prev_Exam_Marks - Height - Weight

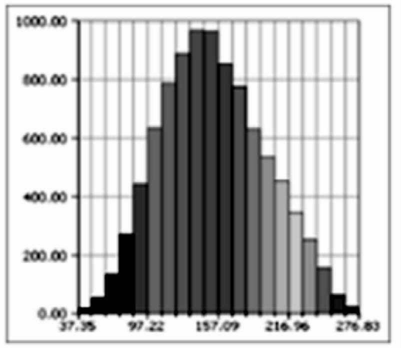
2. Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

(i) Continuous Variables: In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





Note: Univariate analysis is also used to highlight missing and outlier values. In the upcoming part of this series, we will look at methods to handle missing and outlier values. To know more about these methods, you can refer course descriptive statistics from Udacity.

(ii) Categorical Variables: For categorical variables, we'll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

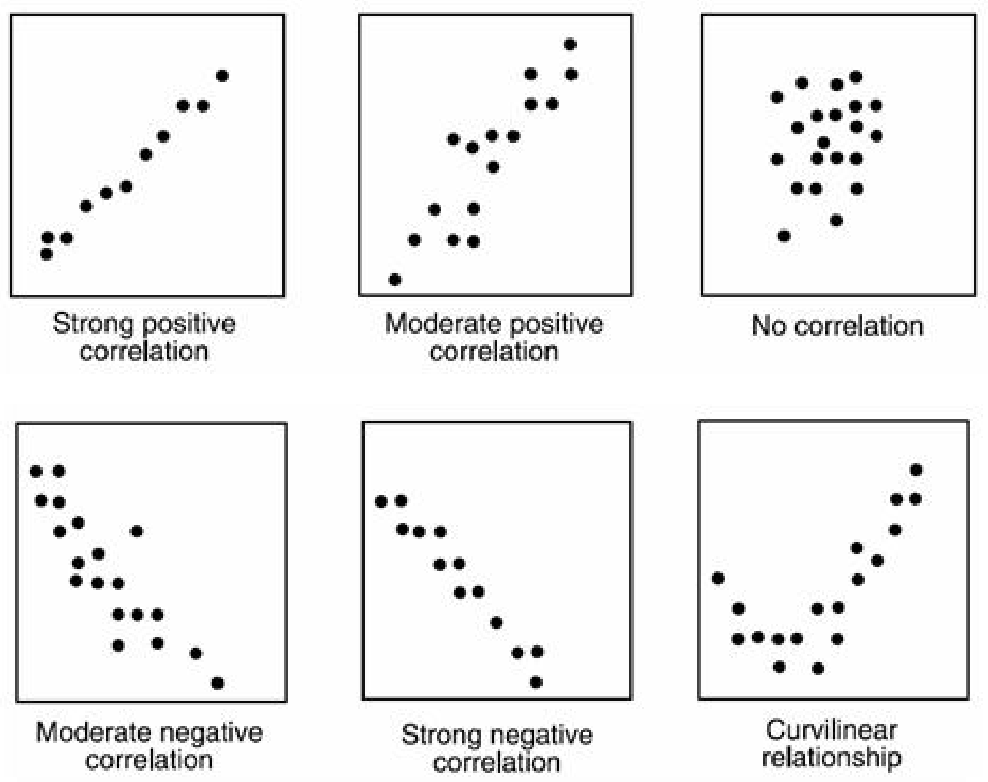
Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail:

Continuous and Continuous

While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Strong positive correlation

Moderate positive correlation

No correlation

Moderate negative correlation

Strong negative correlation

Curvilinear relationship

Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

- –1: perfect negative linear correlation
- +1:perfect positive linear correlation and
- 0: No correlation

Correlation can be derived using following formula:

Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation. These function returns Pearson Correlation value to identify the relationship between two variables:

X	65	72	78	65	72	70	65	68
Y	72	69	79	69	84	75	60	73

Metrics	**Formula**	**Value**
Co-Variance (X,Y)	= COVAR(E6:L6,E7:L7)	18.77
Variance (X)	= VAR.P(E6:L6)	18.48
Variance (Y)	= VAR.P(E7:L7)	45.23
Correlation	= G10/SQRT(G11*G12)	0.65

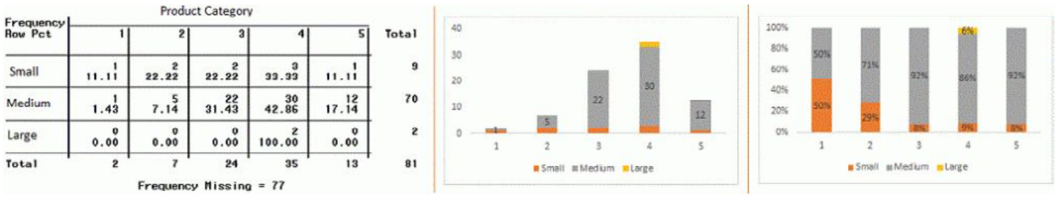
In above example, we have good positive relationship(0.65) between two variables X and Y.

Categorical and Categorical

To find the relationship between two categorical variables, we can use following methods:

- **Two-way table:** We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

- **Stacked Column Chart:** This method is more of a visual form of Two-way table.



Product Category

Frequency Row Pct	1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77

Chi-Square Test

This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent

Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence. The chi-square test statistic for a test of independence of two categorical variables is found by:

$X^2 = \Sigma(O - E)^2 / E$ where O represents the observed frequency. E is the expected frequency under the null hypothesis and computed by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{Sample size}}$$

From previous two-way table, the expected count for product category 1 to be of small size is 0.22. It is derived by taking the row total for Size (9) times the column total for Product category (2) then dividing by the sample size (81). This is procedure is conducted for each cell. Statistical Measures used to analyze the power of relationship are:

- Cramer's V for Nominal Categorical Variable
- Mantel-Haenszed Chi-Square for ordinal categorical variable.

Different data science language and tools have specific methods to perform chi-square test. In SAS, we can use Chisq as an option with Procfreq to perform this test.

Categorical and Continuous

While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

Z-Test/ T-Test

Either test assess whether mean of two groups are statistically different from each other or not. If the probability of Z is small then the difference of two averages is more significant. The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

$$Z = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_3^2}{N_1 + N_2 - 2}$$

Where

- $\overline{X}_1, \overline{X}_{2\text{: Averages}}$
- $S_1^2, S^2_{2\text{:Variances}}$
- $N_1, N_{2\text{: Counts}}$
- t : has t distribution with N_1 + $N_2 - 2$ degree of freedom

ANOVA

It assesses whether the average of more than two groups is statistically different.

Example

Suppose, we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not. This can be done by comparing the weights of the 5 groups of 4 men each.

Till here, we have understood the first three stages of Data Exploration, Variable Identification, Uni-Variate and Bi-Variate analysis. We also looked at various statistical and visual methods to identify the relationship between variables.

Now, we will look at the methods of Missing values Treatment. More importantly, we will also look at why missing values occur in our data and why treating them is necessary.

3.7 Data Reduction

Q20. Explain the Data Reduction in Data mining.

Ans :

Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs.

(a) Data reduction can be achieved using several different types of technologies. The best-known data reduction technique is data deduplication, which eliminates redundant data on storage systems.

(b) The duplication process typically occurs at the storage block level. The system analyzes the storage to see if duplicate blocks exist, and gets rid of any redundant blocks. The remaining block is shared by any file that requires a copy of the block. If an application attempts to modify this block, the block is copied prior to modification so that other files that depend on the block can continue to use the unmodified version, thereby avoiding file corruption. Some storage arrays track which blocks are the most heavily shared.

(c) Those blocks that are shared by the largest number of files may be moved to a memory- or flash storage-based cache so they can be read as efficiently as possible.

(d) Data reduction strategies applied on huge data set. Complex data and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

(e) Data reduction techniques can be applied to obtain a reduces data should be more efficient yet produce the same analytical results.

Q21. Explain the Strategies for data reduction ?

Ans :

The following are the Strategies for reduction

1. **Data cube aggregation,** where aggregation operations are applied to the data in the construction of a data cube.

2. **Attribute subset selections**, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed,

3. **Dimensionality reduction**, where encoding mechanism are used to reduce the data set size.

4. **Numerosity reductions**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models or non parametric method such as clustering, sampling, and the use of histograms.

5. **Discretization and concept hierarchy generation,** where raw data values for attributes are replaced by range or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.

Q22. Explain the techniques used in Data Reduction.

Ans :

i) Dimensionality Reduction

Dimensionality Reduction ensures the reduction of the number of attributesorrandom variables in the data set. Dimensionality reduction is a series of techniques in machine learning and statistics to reduce the number of random variables to consider. It involves feature selection and feature extraction. Dimensionality reduction makes analyzing data much easier and faster for machine learning algorithms without extraneous variables to process, making machine learning algorithms faster and simpler in turn.

ii) Sample Numerosity Reduction

Replaces the original data by an alternative smaller data representation This is a technique of choosing smaller forms or data representation to reduce the volume of data.

These techniques may be parametric or nonparametric.

a) Parametric

For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)

Example: Log-linear models, which estimate discrete multidimensional probability distributions.

b) Nonparametric

Nonparametric methods are used for storing reduced representations of the data include histograms, clustering, and sampling.

Regression and Log-Linear Models

- Regression and log-linear models can be used to approximate the given data.
- In (simple) linear regression, the data are modeled to fit a straight line.
- Multiple linear regression is an extension of (simple) linear regression, which allows a response variable y to be modeled as a linear function of two or more predictor variables.
- Log-linear models approximate discrete multidimensional probability distributions.
- Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.
- This allows a higher-dimensional data space to be constructed from lower dimensional spaces.
- Log-linear models are therefore also useful for dimensionality reduction and data smoothing
- Regression and log-linear models can both be used on sparse data, although their application may be limited.
- While both methods can handle skewed data, regression does exceptionally well. Regression can be computationally intensive when applied to high dimensional data, whereas log-linear models show good scalability for up to 10 or so dimensions.

iii) Cardinality Reduction

Transformations applied to obtain a reduced representation of the original data.

The term cardinality refers to the uniqueness of data values contained in a particular column (attribute) of a database table. The lower the cardinality, the more duplicated elements in a column. Thus, a column with the lowest possible cardinality would have the same value for every row. SQL databases use cardinality to help determine the optimal query plan for a given query.

3.8 Data Classification

Q23. Explain the Classification of Data Mining.

Ans :

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

- Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups.
- Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups.
- For example, we can apply classification in the application that given all records of employees who left the company, predict who will probably leave the company in a future period. In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.
- In technical term, classification in data mining defines as assigning an object to a certain class based on its similarity to previous examples of other objects.
- The classification process comes under the predictive method. With classification, new samples of data are classified into known classes.
- The classification is the initial process of data mining and use algorithms like decision trees, Bayesian classifiers. For classification the data required must be already labeled one.

Examples of classification are:

1. A marketing manager of a company needs to analyze the customer with available profile that who will buy a new computer.
2. A bank officer wants to predict that which loan applicants are risky or which are safe.

- A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.
- In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on.
- Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.
- The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values:
- Example, high credit rating or low credit rating. Multi-class targets have more than two values: for example, low, medium, high, or unknown credit rating
- In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target.
- Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.
- Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

Q24. Explain various issues relating to data classification.

Ans :

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities:

- **Data Cleaning:** Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

- **Relevance Analysis:** Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- **Data Transformation and Reduction:** The data can be transformed by any of the following methods:

  (i) Normalization: The data is transformed using normalization. It involves scaling all values for given attribute in order to make them fall within a small specified range. It is used when in the learning step, the neural networks or the methods involving measurements are used.

  (ii) Generalization: The data can also be transformed by generalizing it to the higher concept. For this purpose, we can use the concept hierarchies.

3.9 Data Association

Q25. Explain the Association in Data Mining ?

Ans :

- It is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions.
- Association analysis is about discovering relationship among huge data sets. Just like the famous market basket analysis which gives a relationship between {Diapers -> beer}. It says that whenever a person buys diapers he/she also buys beer.
- Besides market basket , association rules can be applied to Bioinformatics , web mining and medical analysis.

There are two key issues that need to be addressed while applying this ;

- First detecting the pattern
- Some of the detected patterns can be spurious and may be happening only by chance.
- The strength of a association rule can be measured in terms of support and confidence.

- Support determines how often a rule is applicable to the data set while confidence determines how frequently items in Y appear in transactions that contain X.
- Use packages like a rules, a rules CBA , a rules Sequences in R

  Ex :

  - library ("a rules")
  - data ("Adult")
- rules <- apriori(Adult, parameter = list(supp = 0.5, conf = 0.9, target = "rules"))

3.10 Cause Effect Modeling

Q26. Explain cause effect modelling in Data mining.

Ans :

- Time series data can be used to extract delayed relationship between two variables, for example, "CO_2 emission occurring at a place might cause air pollution at another place after some delay". These lagged relationships signify the time lag between the cause–effect parameters..
- A system such as mechanical, biological or social-economic system consists of independent components. These components influence one another to maintain their activity for the existence of a system in order to achieve the goal of the system
- The system changes behavior when a component is changed or removed significantly. This motivates us to find the reason or cause behind fault and discover the cause parameters in explaining the interactions among the components of a system or process.
- The causal discovery indicates not only that the indicators are correlated, but also how changing a cause variable is expected to induce a change in an effect variable. For example, with analyzed cause–effect relationships, we can predict potential effects before taking any actions (causes), which is useful in preventing inaccurate decision or policy making in the social-economical system.
- Time series data can be used to extract delayed relationship between two variables, for example, "CO2 emission occurring at a place might cause air pollution at another place after some delay".

- These lagged relationships signify the time lag between the cause–effect parameters. Identifying lagged relationships between socioeconomic processes is challenging due to the presence of various complex dependencies in the data.
- This dependency among the various parameters has enabled us to identify relationships among different domain parameters in time series data.
- The cause–effect relationship for time series prediction is a step towards extracting the various existing causal relations between different domain, such as employment, education, agriculture and rural development etc.
- It has also emerged in economics and social sciences such as to improve the economic development and growth of a country and to study the impact of climate change.

Q27. Explain the process of cause and effect analysis.

Ans :

The following are the steps to solve a problem with Cause and Effect Analysis:

Step 1: Identify the Problem

First, write down the exact problem you face. Where appropriate, identify who is involved, what the problem is, and when and where it occurs.

Then, write the problem in a box on the left-hand side of a large sheet of paper, and draw a line across the paper horizontally from the box. This arrangement, looking like the head and spine of a fish, gives you space to develop ideas.

Example: In this simple example, a manager is having problems with an uncooperative branch office.

Step 2: Work Out the Major Factors Involved

Next, identify the factors that may be part of the problem. These may be systems, equipment, materials, external forces, people involved with the problem, and so on.

Try to draw out as many of these as possible. As a starting point, you can use models such as the McKinsey 7S Framework (which offers you Strategy, Structure, Systems, Shared Values, Skills, Style and Staff as factors that you can consider) or the 4Ps of Marketing (which offers Product, Place, Price, and Promotion as possible factors).

Brainstorm any other factors that may affect the situation.

Then draw a line off the “spine” of the diagram for each factor, and label each line.

Step 3: Identify Possible Causes

Now, for each of the factors you considered in Step 2, brainstorm possible causes of the problem that may be related to the factor.

Show these possible causes as shorter lines coming off the "bones" of the diagram. Where a cause is large or complex, then it may be best to break it down into sub-causes. Show these as lines coming off each cause line.

Example: For each of the factors he identified in step 2, the manager brainstorms possible causes of the problem, and adds these to his diagram.

Step 4: Analyze Your Diagram

By this stage, you should have a diagram showing all of the possible causes of the problem that you can think of.

Depending on the complexity and importance of the problem, you can now investigate the most likely causes further. This may involve setting up investigations, carrying out surveys, and so on. These will be designed to test which of these possible causes is actually contributing to the problem.

Example: The manager has now finished his analysis. If he had not looked at the problem this way, he might have dealt with it by assuming that people in the branch office were "being difficult"?

Short Question and Answers

1. What is predictive analysis ?

Ans :

Predictive Analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. It does not tell you what will happen in the future.

Predictive analytics is the branch of the Advanced Analytics which is used to make predictions about unknown future events. It uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future.

Predictive Analytics uses historical data to predict future events. Typically, historical data is used to build a mathematical model that captures important trends. That predictive model is then used on current data to predict what will happen next, or to suggest actions to take for optimal outcomes.

2. Regression analysis.

Ans :

Regression takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents. His major conclusion was that the off-springs of unusually tall persons tend to be shorter than their parents while children of usually short parents tend to be shorter than their parents.

In a sense, the successive generations of off-springs from tall persons “regress” downward toward the height of the population. While the reverse is true originally for short families. But the distribution of heights for the total population, continues to have the same variability from generation to generation.

Regression Variables

i) **Independent Variable (Regressor or Predictor or Explanatory).** The variable which influences the values of the other variable or which is used for prediction of the value of the other variable is called independent variable.

ii) **Dependent Variable (Regressed or Explained Variable).** The variable whose value is influenced or is to be predicted is called dependent variable.

Regression test generates lines of regression of the two variables which helps in estimating the values. Lines of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x. Similarly, line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

3. Limitations of Regression Analysis

Ans :

Some of the limitations of regression analysis are as follows :

i) Regression analysis assumes that linear relationship exists among the related variables. But in the 'area of social sciences, linear relationship may not exist among the related variables.

ii) When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between 'them. These statistic conditions do not exist in social sciences, so, this assumption "minimizes the use' of regression analysis in social science.

iii) The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

4. Moving Averages

Ans :

Moving average methods take the average of past actuals and project it forward. These methods assume that the recent past represents the future. As a result, they work best for products with relatively little change - steady demand, no seasonality, limited trends or cycles and no significant demand shifts. Many companies apply this method because it is simple and easy to use. However, since few products actually behave in this way, it tends to be less useful than more specialized methods.

5. Data Mining

Ans :

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

- Data mining is the analysis step of the “knowledge discovery in databases” process, or KDD.Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
- The difference between data analysis and data mining is that data analysis is to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.
- The term “data mining” is in fact a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence.

6. Scope of Data Mining

Ans :

i) Data mining process the work in such a manner that it allows business to more proactive to grow substantially.

ii) It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.

iii) It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.

iv) It includes tree-shaped structure to understand the hierarchy of data and representation of the set of information described in the database.

v) Brings a genetic way of classification of different sets of data items to view the data in the quick glance.

7. Benefits of Data Mining

Ans :

- Data mining technique helps companies to get knowledge-based information.
- It helps organizations to make the profitable adjustments in operation and production.
- It is a cost-effective and efficient solution compared to other statistical data applications.

- It helps with the decision-making process.
- It facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms.
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

8. Disadvantages of Data Mining

Ans :

- There are chances of companies selling useful information of their customers to other companies for money. For example, American Express has sold credit card purchasesof their customers to the other companies.
- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate. Hence, it can cause serious consequences in certain conditions.

9. What is data exploration?

Ans :

Data exploration is an informative search used by data consumers to form true analysis from the information gathered. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis.

Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats. Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data.

10. Data Reduction

Ans :

Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs.

(a) Data reduction can be achieved using several different types of technologies. The best-known data reduction technique is data deduplication, which eliminates redundant data on storage systems.

(b) The duplication process typically occurs at the storage block level. The system analyzes the storage to see if duplicate blocks exist, and gets rid of any redundant blocks. The remaining block is shared by any file that requires a copy of the block. If an application attempts to modify this block, the block is copied prior to modification so that other files that depend on the block can continue to use the unmodified version, thereby avoiding file corruption. Some storage arrays track which blocks are the most heavily shared.

(c) Those blocks that are shared by the largest number of files may be moved to a memory- or flash storage-based cache so they can be read as efficiently as possible.

(d) Data reduction strategies applied on huge data set. Complex data and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

11. Association in Data Mining

Ans :

- It is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions.

- Association analysis is about discovering relationship among huge data sets. Just like the famous market basket analysis which gives a relationship between {Diapers -> beer}. It says that whenever a person buys diapers he/she also buys beer.

- Besides market basket , association rules can be applied to Bioinformatics , web mining and medical analysis.

Choose the Correct Answer

1. ____________ is an essential process where intelligent methods are applied to extract data patterns. [b]

   (a) Data warehousing (b) Data mining

   (c) Text mining (d) Data selection

2. Which of the following is not a data mining functionality? [c]

   (a) Characterization and Discrimination

   (b) Classification and regression

   (c) Selection and interpretation

   (d) Clustering and Analysis

3. ____________ is the process of finding a model that describes and distinguishes data classes or concepts. [b]

   (a) Data Characterization (b) Data Classification

   (c) Data discrimination (d) Data selection

4. Strategic value of data mining is ____________. [c]

   (a) Cost-sensitive (b) Work-sensitive

   (c) Time-sensitive (e) Technical-sensitive

5. ____________ is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. [a]

   (a) Data Characterization (b) Data Classification

   (c) Data discrimination (d) Data selection

6. If the two series move in reverse directions and the variations in their values are always proportionate, it is said to be: [c]

   (a) Negative correlation (b) Positive correlation

   (c) Perfect negative correlation (d) Perfect positive correlation

7. If one item is fixed and unchangeable and the other item varies, the correlation coefficient will be: [c]

   (a) Positive (b) Negative

   (c) Zero (d) Undecided

8. A process by which we estimate the value of dependent variable on the basis of one or more independent variables is called: [b]

   (a) Correlation (b) Regression

   (c) Residual (d) Slope

9. The slope of the regression line of Y on X is also called the: [d]

   (a) Correlation coefficient of X on Y

   (b) Correlation coefficient of Y on X

   (c) Regression coefficient of X on Y

   (d) Regression coefficient of Y on X

10. In simple linear regression, the numbers of unknown constants are: [b]

    (a) One (b) Two

    (c) Three (d) Four

Fill in the blanks

1. ________ Analytics is the process of using data analytics to make predictions based on data.
2. The regression analysis confined to the study of only two variables at a time is termed as ________ regression.
3. The regression analysis for studying more than two variables at a time is termed as ________ regression.
4. ________ average methods take the average of past actuals and project it forward.
5. Exponential smoothing is a more advanced form of ________ forecasting.
6. ________ is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
7. A data mining project starts with the understanding of the ________ problem.
8. ________ experts build the data model for the modeling process.
9. ________ is an informative search used by data consumers to form true analysis from the information gathered.
10. ________ Analytics is the practice of extracting information from existing data sets in order to determine patterns.

Answers

1. Predictive
2. Simple
3. Multiple
4. Moving
5. Time series
6. Data mining
7. Business
8. Domain
9. Data exploration
10. Predictive

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

4.1 Overview of Linear Optimization

Q1. What is linear programming problem (LPP)? States the mathematical formulation of LPP.

Ans :

In 1947, George dantzig and his associates, while working in the U.S. department of Air Force, observed that a large of military programming and planning problems could be formulated as maximizing/minimizing a linear form of profit/cost function whose variables were restricted to values satisfying a system of linear constraints.

A linear form is meant a thematical expression of the type $a_1x_1 + a_2x_2 + ... + a_nx_n$, where $a_1, a_2, ..., a_n$ are constants and $x_1, x_2, ..., x_n$ are variables. The term 'Programming' refers to the process of determining a particular programme or plan of action. So Linear Programming (L.P.) is one of the most important optimization (maximization/minimization) techniques developed in the field of Operations Research.

The general LPP calls for optimizing (maximizing/minimizing) a linear function of variables called the '*Objective Function*' subject to a set of linear equations and / or inequalities called the '*constraints*' or *restrictions*.

Mathematical Model of LPP

In order to find the values of n decision variables $x_1, x_2, ..., x_n$ to maximize or minimize the objective function

$$z = c_1x_1 + c_2x_2 + c_3x_3 + ... + c_nx_n$$

and also satisfy m-constraints :

$$\left.\begin{array}{l}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n (\leq = \text{ or } \geq)\ b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2j}x_j + \ldots + a_{2n}x_n (\leq = \text{ or } \geq)\ b_2 \\
\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad\quad \vdots \\
a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n\ (\leq \ \ = \text{or } \geq)\ b_i \\
\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad\quad \vdots \\
a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n (\leq = \text{ or } \geq)\ b_m
\end{array}\right\}$$

where constraints may be in the form of any inequality ($\leq$ or $\geq$) or even in the form of an equation ($=$) and finally satisfy the non-negativity restrictions

$$x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_j \geq 0,\ \ldots,\ x_n \geq 0.$$

Formulation of LPP

The formulation of linear programming problem as a mathematical model involves the following basic steps :

Step 1

Find the key-decision to be made from the study of the solution. (In this connection, looking for variables helps considerably).

Step 2

Identify the variables and assume symbols x_1, x_2 ... for variable quantities noticed in step 1.

Step 3

Express the possible alternatives mathematically in terms of variables. The set of feasible alternatives generally in the given situation is :

$$\{(x_1, x_2)\ ;\ x_1 > x_2 > 0\}$$

Step 4

Mention the objective quantitatively and express it as a linear function of variables.

Step 5

Express the constraints also as linear equalities / inequalities in terms of variables.

Q2. State the advantages and limitations of linear programming problem.

Ans :

Advantages

1. It helps in organization and study of the information in the same way that the scientific approach to the problem requires.
2. With LP the execute builds into his planning a true reflection of the limitations and restrictions under which he must operate.

3. Once a basic plan is arrived at through LP, it can be reevaluated for changing conditions.
4. Highlighting of bottlenecks in the production process is the striking advantages of this technique.
5. It provides flexibility in analyzing a variety of multidimensional problems.

Limitations of LP

Inspite of wide area of applications, some limitations are associated with linear programming techniques. These are stated below :

1. In some problems objective functions and constraints are not linear. Generally, in real life situations concerning business and industrial problems constraints are not linearly created to variables.
2. There is no guarantee of getting integer valued solutions, for example, in finding out how may men and machines would be required to perform a particular job, rounding off the solution to the nearest integer will not give an optimal solution. Integer programming deals with such problems.
3. Linear programming model does not take into consideration the effect of time and uncertainty. Thus the model should be defined in such a way that any change due to internal as well as external factors can be incorporated.
4. Sometimes large-scale problems cannot be solved with linear programming techniques even when the computer facility is available. Such difficulty may be removed by decomposing the main problem into several small problems and then solving them separately.
5. Parameters appearing in the model are assumed to be constant. But, in real life situations they are neither constant not deterministic.
6. Linear programming deals with only single objective, whereas in real life situations problems come across with multiobjectives.

Q3. State the assumptions and applications of LPP.

Ans :

Assumptions of LPP

1. Proportionality

A primary requirement of linear programming problem is that the objective function and every constraint function must be *linear*. Roughly speaking, it simply means that if 1 kg of a product costs Rs. 2, then 10 kg will cost Rs. 20. If a steel mill can produce 200 tons in 1 hour, it can produce 1000 tons in 5 hours.

Intuitively, linearity implies that the product of variables such as $x_1\ x_2$, powers of variables such as x_3^2, and combination of variables such as $a_1x_1 + a_2 \log x_2$, are not allowed.

2. Additivity

Additivity means if it takes t_1 hours on machine G to make product a and t_2 hours to make product B, then the time on machine G devoted to produce A and B both is $t_1 + t_2$, provided the time required to change the machine from product A to B is negligible.

Then additivity may not hold, in general. If we mix several liquids of different chemical composition, then the total volume of the mixture may not be the sum of the volume of individual liquids.

3. Multiplicativity

It requires :

(a) It takes one hour to make a single item on a given machine, it will take 10 hours to make 10 such items.

(b) The total profit from selling a given number of units is the unit profit times the number of units sold.

4. Divisibility

It means that the fractional levels of variables must be permissible besides integral values.

5. Deterministic

All the parameters in the linear programming models are assumed to be known exactly. While in actual practice, production may depend upon change also.

Applications of LPP

1. Personnel Assignment Problem

Suppose we are given m persons, n-jobs, and the expected productivity c_{ij} of *i*th person on the *j*th job. We want to find an assignment of persons $x_{ij} \geq 0$ for all i and j, to n jobs so that the average productivity of person assigned is maximum.

2. Transportation Problem

We suppose that m factories (called sources) supply n warehouses (called destinations) with a certain product. Factory F_i (i = 1, 2, ..., m) produces a_i units (total or per unit time) and warehouse W_j (j = 1, 2, 3 ..., n) requires b_j units. Let the decision variables x_{ij}, be the amount shipped from factory F_i to warehouse W_j. The objective is to determine the number of units transported from factory F_i to warehouse W_j. The objective is to determine the number of units transported from factory F_i to warehouse W_j so that the total transportation cost $\sum_{i=1}^{m} \sum_{i=1}^{n} c_{ij} x_{ij}$ is minimized.

3. Efficiencing on Operation of System of Dams

In this problem, we determine variations in water storage of dams which generate power so as to maximize the energy obtained from the entire system. The physical limitations of storage appear as inequalities.

4. Agricultural Applications

Linear programming can be applied in agricultural planning for allocating the limited resource such as acreage, labour, water, supply and working capital, etc. so as to maximize the net revenue.

5. Military Applications

These applications involve the problem of selecting an air weapon system against gurillas so as to keep them pinned down and simultaneously minimize the amount of aviation gasoline used, a variation of transportation problem that maximizes the total tonnage of bomb dropped on a set of targets, and the problem of community defence against disaster to find the number of defence units that should be used in the attack in order to provide the required level of protection at the lowest possible cost.

6. Marketing Management

Linear programming helps in analyzing the effectiveness of advertising campaign and time based on the available advertising media. It also helps travelling salesman in finding the shortest route for his tour.

7. Manpower Management

Linear programming allows the personnel manager to analyses personnel policy combinations in terms of their appropriateness for maintaining a steady-state flow of people into through and out of the firm.

8. Physical Distribution

Linear programming determines the most economic and efficient manner of locating manufacturing plants and distribution centres for physical distribution.

Q4. What are the requirements of linear programming problem ?

Ans :

1. Decision variables and their relationship

The decision (activity variables refer to candidates (products, services, projects etc.) that are competiting with one another for sharing the given limited resources. These variables are usually inter-related in terms of utilization of resources and need simultaneous solutions. The relationship among these variables should be linear.

2. Well defined objective function

A linear programming problem must have a clearly defined objective function to optimize which may be either to maximize contribution by utilizing available resources, or it may be to produce at the lowest possible cost by using a limited amount of productive factors. It should be expressed as a linear function of decision variables.

3. Presence of constraints or restrictions

There must be limitations on resources (like production capacity, manpower, time, machines, markets, etc.) which are to be allocated among various competing activities. These must be capable of being expressed as linear equalities or inequalities in terms of decision variables.

4. Alternative courses of action

There must be alternative courses of action. For example, it must be possible to make a selection between various combinations of the productive factors such as men, machines, materials, markets, etc.

5. Non-negative restrictions

All decision variables must assume non-negative values as negative values of physical quantities is an impossible situation. If any of the variables is unrestricted in sign, a trick can be employed which enforces non-negativity changing the original information of the problem.

Example

Rahul and Co. manufacturers two brands of products namely Shivnath and Harinath. Both these models have to under go the operations on three machines lathe, milling and grinding. Each unit of Shivnath gives a profit of Rs. 45 and requires 2 hours on lathe, 3 hours on milling and 1 hour on grinding. Each unit of Harinath can give a profit of Rs. 70 and requires 3, 5, and 4 hours on lathe, milling and grinding respectively. Due to prior commitment, the use of lathe hours are restricted to a maximum of 70 hours in a week. The operators to operate milling machines are hired for 110 hours / week. Due to scarce availability of skilled man power for grinding machine, the grinding hours are limited to 100 hours / week. Formulate the data into an LPP.

Sol :

Step 1 : Selection of Variables

In the above problem, we can observe that the decision is to be taken on how many products of each brand is to be manufactured. Hence the quantities of products to be produced per week are the decision variables.

Therefore we assume that the number of units of product Shivnath brand produced per week = x_1.

The number of units of product of Harinath brand produced per week = x_2.

Step 2 : Setting Objective

In the given problem the profits on the brands are given.

Therefore objective function is to maximize the profits.

Now, the profit on each unit of Shivnath brand = Rs. 45.

Number of units of Shivnath to be manufactured = x_1

$\therefore$ The profit on x_1 units of Shivnath brand = $45\ x_1$

Similarly, the profit on each unit of Harinath brand = Rs. 70

Number of units of Harinath brand to be manufactured = x_2

$\therefore$ The profit on x_2 units of Harinath brand = $70\ x_2$

The total profit on both brands = $45\ x_1 + 70\ x_2$

This total profit (say z) is to be maximized

Hence, the objective function is to Maximize $z = 45x_1 + 70x_2$

Step 3 : Identification of Constraint Set

In the above problem, the constraints are the availability of machine hours.

1. Constraint on Lathe Machine

Each unit of Shivnath brand requires 2 hours / week

So x_1 units of Shivnath brand requires $2x_1$ hours / week.

Each unit of Harinath brand requires 3 hours / week and so x_2 units of Harinath brand require $3x_2$ hours / week.

Total lathe hours utilized for both the brands is $2x_1 + 3x_2$ and this cannot exceed 70 hours / week

$\therefore \quad 2x_1 + 3x_2 \leq 70$

(Constraint on availability of lathe hours due to prior commitment)

2. Constraint on Milling Machine

Milling hours required for each unit of Shivnath brand = 3 hours / week.

$\therefore$ For x_1 units = $3x_1$ hours / week

Milling hours required for each unit of Harinath brand = 5 hours / week.

$\therefore$ For x_2 units = $5x_2$

Total milling hours = $3x_1 + 5x_2$

This can not be more than 110

$\therefore \quad 3x_1 + 5x_2 \leq 110$

(Constraint on availability of milling machine hours due to hiring)

3. Constraint on Grinding Machine

One unit of Shivnath needs one hour / week and x_1 units need x_1 hours / week.

One unit of Harinath needs 4 hours / week and x_2 units need $4x_2$ hours / week.

Total grinding hours = $x_1 + 4x_2$ and this cannot be greater than 100 hours.

$\therefore \quad x_1 + 4x_2 \leq 100$

(Constraint on availability of grinding hours due to scarcity of skilled labour)

4. Writing Conditions of Variables

Both x_1 and x_2 are the number of products to be produced. There can not exist any negative production. Therefore x_1 and x_2 can not assume any negative values (i.e., non negative)

Mathematically

$x_1 \geq 0$ and $x_2 \geq 0$

Step 5 : Summary

Maximize $Z = 45x_1 + 70x_2$

Subject to $2x_1 + 3x_2 \leq 70$

$3x_1 + 5x_2 \leq 110$

$x_1 + 4x_2 \leq 100$

$x_1 \geq 0$ and $x_2 \geq 0$.

Q5. Describe the steps involved in graphical solution to linear programming models.

Ans :

Simple linear programming problems of two decision variables can be easily solved by graphical method. The outlines of graphical procedure are as follows :

Step 1 : Consider each inequality-constraint as equation.

Step 2 : Plot each equation on the graph, as each one will geometrically represent a straight line.

Step 3 : Shade the feasible region. Every point on the line will satisfy the equation of the line. If the inequality-constraint corresponding to that lines is '$\leq$', then the region below the line lying in the first quadrant (due to non-negativity of variables) is shaded. For the inequality-constraint with '$\geq$' sign, the region above the line in the first quadrant is shaded. The points lying in common region will satisfy all the constraints simultaneously. The common region thus obtained is called the *feasible region.*

Step 4 : Choose the convenient value of z (say = 0) and plot the objective function line.

Step 5 : Pull the objective function line until the extreme points of the feasible region. In the maximization case, this line will stop farthest from the origin and passing through at least one corner of the feasible region. In the minimization case, this line will stop nearest to the origin and passing through at least one corner of the feasible region.

Step 6 : Read the coordinates of the extreme point(s) selected in *step 5* and find the maximum or minimum (as the case may be) value of z.

Q6. Write the computational procedure for simplex method.

(OR)

Explain the procedure for simplex method.

Ans :

Simplex algorithm was originally proposed by G.B Dantzig in 1948.

It starts at a basic level of the problem.

At each step it projects the improvement in the objective function over its previous step. Thus, the solution becomes optimum when no further improvement is possible on the objective function.

Simplex Algorithm

The algorithm goes as follows :

Step 1: Formulation of LPP

- Selection of decision variables
- Setting of objective function
- Identification of constraint set
- Writing the conditions of variables.

Step 2: Convert constraints into equality form.

- Add slack variable if constraints is $\leq$ type.
- Subtract surplus and add an artificial variable if the constraint is $\geq$ type.
- Add an artificial variable if constraint is exact (=) type.

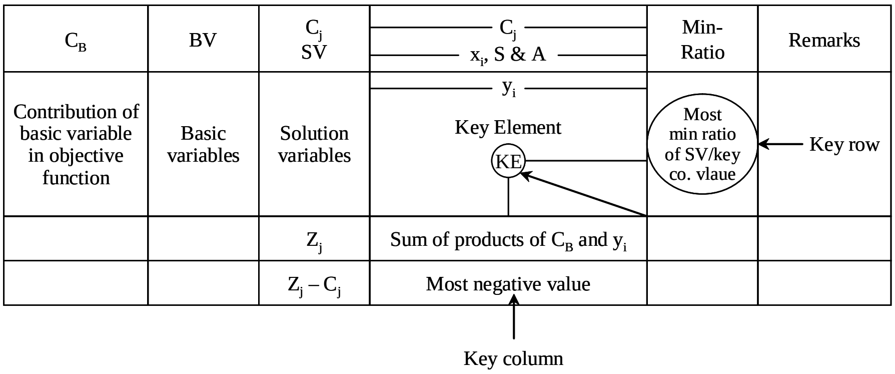
Step 3: Find, Initial Basic Feasible Solution (IBFS)

- If m non identical equations have n variables (m < n) including all decision, slack / surplus and artificial variables, we get m number of variables basic and (n – m) variables non basic (i.e., equated to zero).
- First make all decision variables (and surplus) as non basic i.e., equate to zero to identity the IBFS.
- Find solution values for basic variables.

Step 4: Construct, Initial Simplex Tableau as given above with the following notations.

- C_B : Coefficient of basic variable in the objective functions
  (or contribution of basic variables)
- BV : Basic variabels (form IBFS)

- SV : Solution value (from IBFS)
- C_j : Contribution of j^{th} variables or coefficient of each variable (j^{th}) in objective function.
- $Z_j – C_j$: Net contribution.



C_B	BV	C_j SV	C_j x_i, S & A	Min-Ratio	Remarks
Contribution of basic variable in objective function	Basic variables	Solution variables	y_i Key Element KE	Most min ratio of SV/key co. vlaue	
		Z_j	Sum of products of C_B and y_i		
		$Z_j – C_j$	Most negative value		

Step 5: Find ‘out going’ and ‘incoming’ variables.

- Find Z_n by summation of products of C_B and y_i for each column
- Computed $Z_j – C_j$ value for each column
- To find key column use most negative value of $Z_j – C_j$
- Variable in key column is ‘in coming variable’ or ‘entering’ variable.
- The variable of key row is ‘out going’ or ‘existing’ variable.
- Find the minimum ratio of solution value to corresponding key column value to identify key row.
- The cross section of key column and key row is key element with which the next iteration is carried out.

Step 6: Re-write next tableau as per given set of rules.

- Replace the existing variable from the basis with the entering variable along with its coefficient (or contribution).
- You have to make key element as unity (i.e., 1) and other element in the key column as zeros.
- To make key element as unity, divide the whole key row by the key element. This is supposed as the new row in the place of key row in the next iteration table.
- To find other rows of next iteration table, use this new row. By appropriate adding or subtracting entire new row in the old rows, make other elements of the key column as zeros.

Step 7: Check whether all the values of $Z_j - C_j$ are positive. If all are positive, the optimal solution is reached. Write the solution values and find Z_{opt}. (i.e., Z_{max} or Z_{min} as the case may be).

If $Z_j - C_j$ values are still negative, again choose most negative among these and go to step 5 and repeat the iteration till all the values of $Z_j - C_j$ become positive.

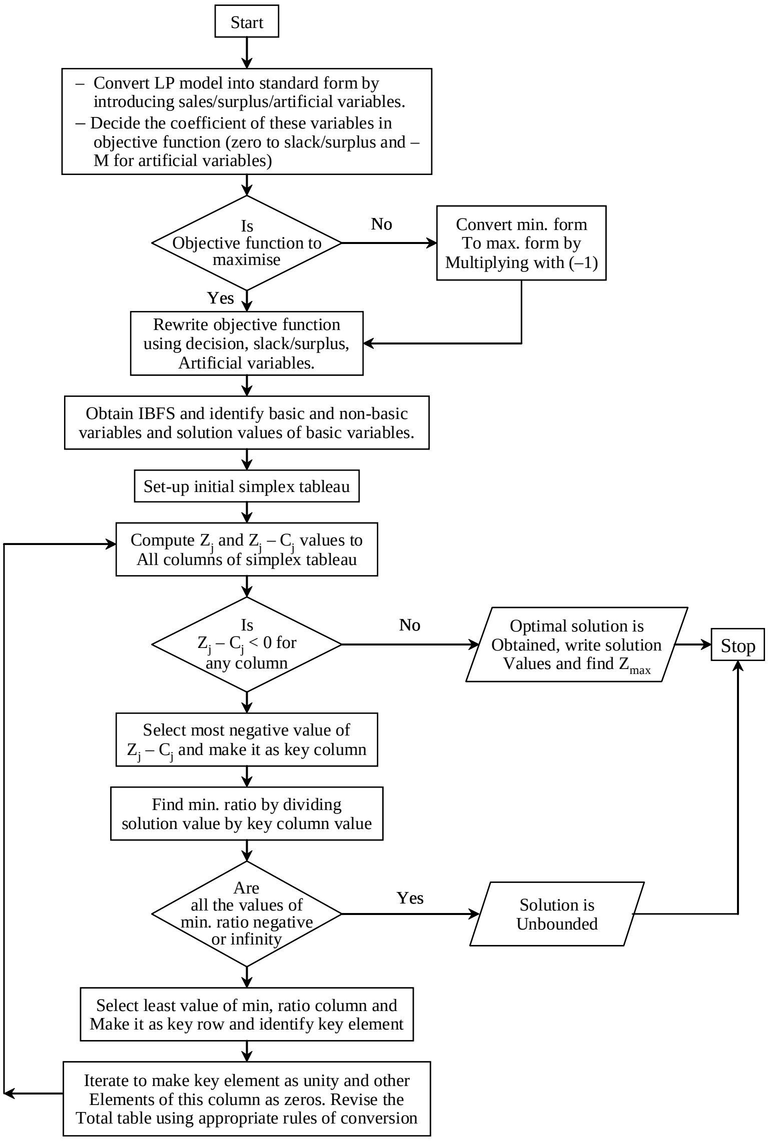

Start
– Convert LP model into standard form by introducing sales/surplus/artificial variables.
– Decide the coefficient of these variables in objective function (zero to slack/surplus and – M for artificial variables)
Is Objective function to maximise
No
Convert min. form To max. form by Multiplying with (–1)
Yes
Rewrite objective function using decision, slack/surplus, Artificial variables.
Obtain IBFS and identify basic and non-basic variables and solution values of basic variables.
Set-up initial simplex tableau
Compute Z_j and $Z_j – C_j$ values to All columns of simplex tableau
Is $Z_j – C_j < 0$ for any column
No
Optimal solution is Obtained, write solution Values and find Z_{max}
Stop
Select most negative value of $Z_j – C_j$ and make it as key column
Find min. ratio by dividing solution value by key column value
Are all the values of min. ratio negative or infinity
Yes
Solution is Unbounded
Select least value of min, ratio column and Make it as key row and identify key element
Iterate to make key element as unity and other Elements of this column as zeros. Revise the Total table using appropriate rules of conversion

Fig.: Flow Chart or Simplex Method

4.2 Non-linear Programming Integer Optimization

Q7. Explain Non Linear Programming?

Ans :

- Non Linear Programming (NLP). If an LP problem is feasible then, at least in theory, it can always be solved because. We know the solution is a "corner point": a point where lines or planes intersect.
- There are a finite number of possible solution points. The simplex algorithm will find that point. Also, a very informative sensitivity analysis is relatively easy to obtain for LP problems. But in many interesting, real-world problems, the objective function may not be a linear function, or some of the constraints may not be linear constraints
- Optimization problems that involve nonlinearities are called nonlinear programming (NLP) problems. Many NLPs do not have any constraints. They are called unconstrained NLPs. Solutions to NLPs are found using search procedures.
- Solutions are more difficult to determine, compared to LPs. One problem is difficulty in distinguishing between a local and global minimum or maximum point.
- Nonlinear programming is the process of solving an optimization problem where some of the constraints or the objective function are nonlinear.
- An optimization problem is one of calculation of the extrema (maxima, minima or stationary points) of an objective function over a set of unknown real variables and conditional to the satisfaction of a system of equalities and inequalities, collectively termed constraints.
- It is the sub-field of mathematical optimization that deals with problems that are not linear.
- In mathematics, nonlinear programming is the process of solving an optimization problem where some of the constraints or the objective function are nonlinear.
- It is the sub-field of mathematical optimization that deals with problems that are not linear.

General Form of Non-linear Programming Problems

Max f(x)

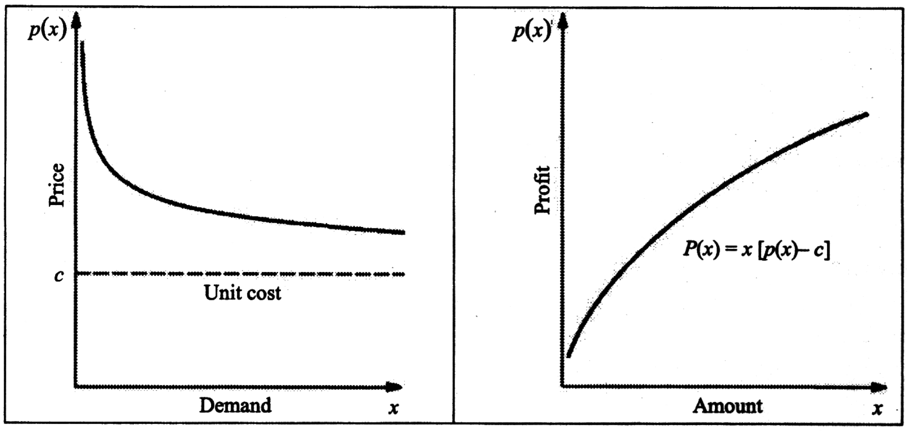
S.T. gi(x) $\leq$ b_i for i = 1, ... , m

x$\geq$0

No algorithm that will solve every specific problem fitting this format is available.

Example - The Product-Mix Problem with Price Elasticity

- The amount of a product that can be sold has an inverse relationship to the price charged, *i.e.*, the relationship between demand and price is an inverse curve.



The firm's profit from producing and selling x units is the sales revenue xp(x) minus the production costs, i.e., P(x) = xp(x) – cx.

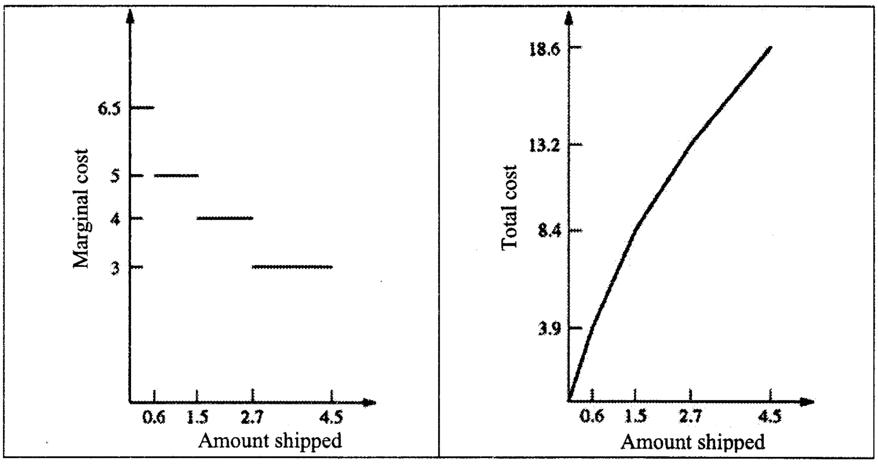
If each of the firm's products has a similar profit function, say, Pfixj) for producing and selling Xj units of product j, then the overall objective function is

$$f(x) = \sum_{j-1} p_j(x_j), \text{ a sum of non-linear functions.}$$

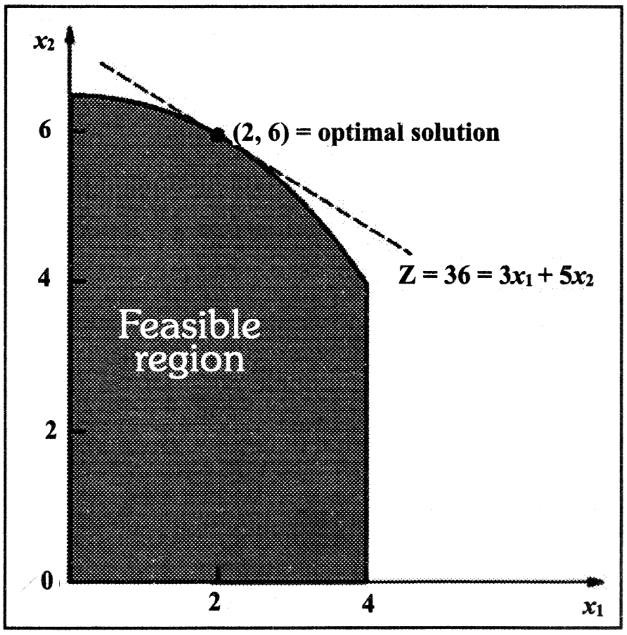
- Non-linearities also may arise in the g,(x) constraint function.

An Example - The Transportation Problem with Volume Discounts

- Determine an optimal plan for shipping goods from various sources to various destinations, given supply and demand constraints.
- In actuality, the shipping costs may not be fixed. Volume discounts sometimes are available for large shipments, which cause a piece-wise linear cost function.



Graphical Illustration of Non-linear Programming Problems



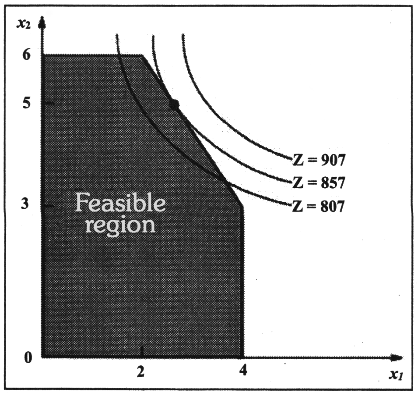
Max $Z = 3x_1 + 5x_2$

S.T. $x_1 \le 4$

$9x_1^2 + 5x_2^2 \le 216$

$X_1, X_2 \ge 0$

- The optimal solution is no longer a CPF anymore. (Sometimes, it is; sometimes, it isn't). But, it still lies on the boundary of the feasible region.
- We no longer have the tremendous simplification used in LP of limiting the search for an optimal solution to just the CPF solutions.
- What if the constraints are linear; but the objective function is not?



$\text{Max } Z = 126x_1 - 9x^2 + 182x - 13x^2$

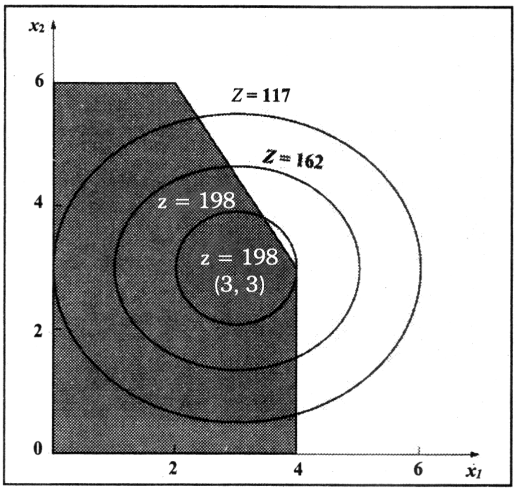
S.T. $x_1 \leq 4$

$2x_2 \leq 12$

$3x_1 + 2x_2 \leq 18$

$x_1, x_2 \geq 0$

What if we change the objective function to $54x_1 - 9x^2 + 78x - 13x^{27}$



- The optimal solution lies inside the feasible region.
- That means we cannot only focus on the boundary of feasible region. We need to look at the entire feasible region.

Constrained Optimization with Equality Constraints

- Consider the problem of finding the minimum or maximum of the function f(x), subject to the restriction that x must satisfy all the equations:

  $g_1(x) = b_1$

  ...

  g m(x) = bm

Example

Max f $(x_1, x_2) = x_1^2 + 2x_2$

S.T. $g(x_1, x_2) = x_1^2 + x_2^2 = 1$

A classical method is the method of Lagrange multipliers.

- The Lagrangian function $h(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i [g_i(x) - b_i]$, where $(\lambda_1, \lambda_2, \lambda_m)$ are called Lagrange multipliers.
- For the feasible values of x, gi(x) – $b_i = 0$ for all i, so $h(x, \lambda) = f(x)$.
- The method reduces to analyzing h(x, X) by the procedure for unconstrained optimization.
- Set all partial derivative to zero
- Notice that the last m equations are equivalent to the constraints in the original problem. So, only feasible solutions are considered.

4.3 Cutting Planning Algorithm and Methods

Q8. Explain the Cutting Plane Method ?

Ans :

- In mathematical optimization, the cutting-plane method is any of a variety of optimization
- Methods that iteratively refine a feasible set or objective function by means of linear inequalities, termed cuts. Such procedures are commonly used to find integer solutions to mixed integer linear programming (MILP) problems, as well as to solve general, not necessarily differentiable convex optimization problems. The use of cutting planes to solve MILP was introduced by Ralph E. Gomory.
- Cutting plane methods for MILP work by solving a non-integer linear program, the linear relaxation of the given integer program.
- The theory of Linear Programming dictates that under mild assumptions (if the linear program has an optimal solution, and if the feasible region does not contain a line), one can always find an extreme point or a corner point that is optimal.

- The obtained optimum is tested for being an integer solution. If it is not, there is guaranteed to exist a linear inequality that separates the optimum from the convex hull of the true feasible set.
- Finding such an inequality is the separation problem, and such an inequality is a cut. A cut can be added to the relaxed linear program.
- Then, the current non-integer solution is no longer feasible to the relaxation. This process is repeated until an optimal integer solution is found.

We start by solving the LP relaxation to get a lower bound for the minimum objective value.

We assume the final simplex tableau is given, the basic variables having columns with coefficient 1 in one constraint row and 0 in other rows. The solution can be read from this form: when the non - basic variables are 0, the basic variables have the values on right-hand side (RHS). The objective function row is of the same form, with its basic variable f.

If the LP solution is fractional, i.e., not integer, at least one of the RHS values is fractional. We proceed by appending to the model a constraint that cuts away a part of the feasible set so that no integer solutions are lost.

Take a row i from the final simplex tableau, with a fractional RHS d. Denote by x_{j0} the basic variable of this row and N the index set of non-basic variables.

Row i as an equation:

$$x_{jo} + \sum_{j \in N} w_{ij} x_j = d$$

Denote by Idm, the largest integer, i.e., #d (the whole part of d, if d is positive). Because all variables are non-negative,

$$\sum_{j \in N} [w_{ij}] x_j \, \# \sum_{j \in N} w_{ij} x_j$$

$$x_{jo} + \sum_{j \in N} [w_{ij}] x_j \, \# \, d$$

Left hand side is integer

$$x_{j0} \sum_{j \in N} [w_{ij}] x_j \, \# \, 1dm$$

From the first and last formula, it follows that

$$d - idm \# \sum_{j \in N} (w_{ij} - [w_{ij}])x_j \# idm$$

If we denote the fractional parts by symbols r = d – Idm and f_{ij} = w_{ij} – IWijm,, we get a cut constraint or a cutting plane in the solution space:

$$\sum_{j \in N} f_{ij} x_j \$ r$$

Equation form, using a slack variable s_i:

$$-\sum_{j \in N} f_{ij} x_j + s_i = -r$$

This equation is of basic form, with basic variable s_i = – r.

The resulting simplex tableau is optimal but infeasible, and we apply the dual simplex method until all variables are non-negative.

The cut constraints do not cut out any feasible integer points and they pass through at least one integer point.

The next cutting plane algorithm operates with a simplex tableau.

Q9. Explain cutting Plane Algorithm ?

Ans :

- ➢ Cutting-plane methods for general convex continuous optimization and variants are known under various names: Kelley's method, Kelley–Cheney–Goldstein method, and bundle methods.
- ➢ They are popularly used for non-differentiable convex minimization, where a convex objective function and its subgradient can be evaluated efficiently but usual gradient methods for differentiable optimization can not be used.
- ➢ This situation is most typical for the concave maximization of Lagrangian dual functions.
- ➢ Another common situation is the application of the Dantzig–Wolfe decomposition to a structured optimization problem in which formulations with an exponential number of variables are obtained.
- ➢ Generating these variables on demand by means of delayed column generation is identical to performing a cutting plane on the respective dual problem.

Q10. Explain the other methods in cutting plane algorithm.

Ans :

The following are the methods.

1. Cutting planes
2. Localization methods

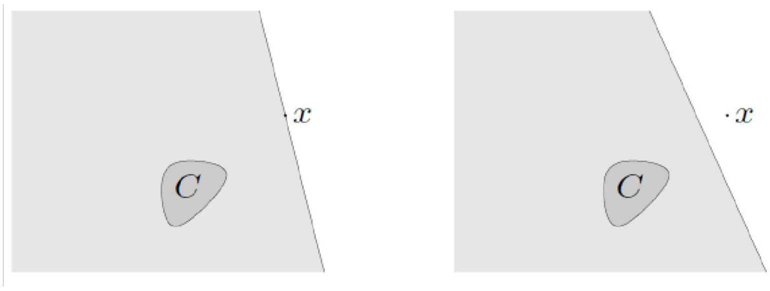
1. Cutting-plane

Oracle provides a black-box description of a convex set C

- When queried at x, oracle either asserts $x \in C$ or returns $a \neq 0$, b with

$$a^T x \geq b, \quad a^T z \leq b \quad \forall z \in C$$

- $a^T z = b$ defines a cutting plane, separating x and C
- Cut is neutral if $a^T x = b$: query point is on boundary of half-space
- Cut is deep if $a^T x > b$: query point in interior of half-spaces that is cut.



2. Localization method

- Goal: Find a point in convex set C described by cutting-plane oracle
- Algorithm: choose bounded set P_0 containing C; repeat for $k \geq 1$:
- Choose a point $x^{(k)}$ in P_{k-1} and query the cutting-plane oracle at $x^{(k)}$.
- if $x^{(k)} \in C$, return $x^{(k)}$; else, add cutting plane $a_k^T z \leq b_k$ to P_{k-1}:

$$P_k = P_{k-1} \cap \{z \mid a_k^T z \leq b_k\}$$

- Termine if $P_k = f$

Variation: to keep P_k simple, choose $P_k \supseteq P_{k-1} \cap \{z \mid a_k^T z \leq b_k\}$.

4.4 Decision Analysis

Q11. Explain the term decision analysis?

Ans :

The term decision analysis was coined in 1964 by Ronald A. Howard, professor of management science and engineering at Stanford University. Decision analysis refers to a systematic, quantitative and interactive approach to addressing and evaluating important choices confronted by organisations in the private and public sector. Decision analysis is interdisciplinary and draws on theories from the fields of psychology, economics, and management science. It utilises a variety of tools which include models for decision-making under conditions of uncertainty or multiple objectives; techniques of risk analysis and risk assessment; experimental and descriptive studies of decision-making behaviour; economic analysis of competitive and strategic decisions; techniques for facilitating decision-making by groups; and computer modeling software and expert systems for decision support.

Example

If XYZ real estate development company were deciding whether or not to build a new shopping center in a location, they might examine several pieces of input to aid in their decision-making process. These might include traffic at the proposed location on various days of the week at different times, the popularity of similar shopping centers in the area, financial demographics and spending habits of the area population, local competition, and preferred shopping habits of the area population. All of these items could be put into a decision analysis program and different simulations could be run that would help XYZ company make their decision about the shopping center.

Q12. Explain Decision Making and criteria for decision making.

Ans :

- Normally in decision we have two or more then two alternative from which to be choose one.
- The consequence of the alternative depends upon the outcomes of some future random event (or states of nature).
- Outcomes of such decision are usually unexpected but possible to predict by information carry chosen one alternative.

Criteria For Decision Making

If a decision maker carryout information about an alternative upon it become easy to takes decision.

Like converting available information into a measure of desirability.

An appropriate one criteria for decision can be: $\Rightarrow$ Probability estimates of future outcomes. And willingness of decision maker to take risk.

4.4.1 Decision Making under Risk and Uncertainity

Q13. Explain about decision making under uncertainty.

Ans :

In decision-making under uncertainty the probabilities associated with occurrence of the different states of nature are not given. The decision maker has to determine the expected payoff for the courses of action or strategies as the probabilities associated with the occurrence of states of nature are not given. The decision maker has number of criteria available and has to select one among them. The selection depends upon the attitude of the decision maker and the policy of an organization.

Decision-making under uncertainty has various criteria such as,

1. **Criterion of Pessimism or Maximin**

   Maximin initially identifies the worst possible outcome for each course of action i.e., maximum loss or minimum outcome that would occur under each decision alternative and then choosing the best out of the worst outcome (i.e., maximum payoff) is order to select the optimal course of action or strategy.

2. **Criterion of Optimism or Maximax**

   Maximax is totally reverse of maximin operations research criterion of pessimism. Maximax identifies the best possible outcome (maximum payoff associated with each course of action and then choose the maximum of the maximum value in order to select the optimal course of action or strategy.

3. **Minimax Regret Criterion**

   Minimax regret criterion is useful in identifying the regret (or opportunity loss) which is associated with each states of nature if a specific course of action is undertaken.

   For each conditional profit or the cost value (payoff) regret value is calculated by taking the difference between the maximum payoff under a state of nature and the payoff resulting from each course of action under that state is calculated. It can be shown in an equational form as, Regret payoff = Maximum payoff from a course of action-Payoff

Once the value is obtained, find the highest regret value for each course of action and then select that course of action with the minimum regret values. The regret is quite similar to the EOL (Expected Opportunity Loss) which is also known as conditional opportunity loss.

4. **Hurwicz Criterion or Criterion of Realism**

A rational decision maker should not be either completely optimistic or pessimistic. Hurwicz introduced the idea of coefficient of optimism. Let the coefficient of optimism be a then,

$$0 \leq \alpha \leq 1$$

a) If a is close to 1, the decision-maker is optimistic about the future.

b) If a is close to zero, the decision-maker is pessimistic about the future. According to Hurwicz, select the strategy that maximizes.

H = a (Maximum payoff in column) + (1 - a) (Minimum payoff in column)

5. **Criterion of Rationality or Baye's or Laplace Criterion**

Laplace criterion is based on the principle of equal likelihood or insufficient reason. According to this principle, as probabilities of future states of nature is unknown, there is no reason to consider any one outcome more likely than the other i.e., all outcomes must be considered equally likely. With outcomes, each outcome will thus have a probability of *1ln.* With the help of these probabilities such a course of action must be chosen which has the highest expected loss.

As all the outcomes are weighted equally, the average outcome for each course of action must be calculated i.e., add for each course of action the payoffs for all outcomes (states of nature), and then divide it by the number of courses of action.

Example

A food product company is contemplating the introduction of a revolutionary new product with new packaging to replace the existing product at much price (S_1) or a moderate change in the composition of the existing product with a new packaging at a small increase in price (S_2) or a small change in the composition of the existing except the word 'new' with a negligible increase in price (S_3). The three possible states of nature of events are, (i) High increase in sales (N_2), (ii) No change in sales (N_2) and (iii) Decrease in sales (N_3). The marketing department of the company worked out the payoffs in terms of yearly net profits for each course of action for these events (expected sales). This is represented in the following table,

States of Nature	**Courses of Action**		
	S_1	**S_2**	**S_3**
N_1	**7,00,000**	**5,00,000**	**3,00,000**
N_2	**3,00,000**	**4,50,000**	**3,00,000**
N_3	**1,50,000**	**0**	**3,00,000**

Which strategy should the company choose on the basis of,

a) Maximin criterion

b) Maximax criterion

c) Minimax regret criterion

d) Laplace criterion.

Sol :

a) The given table is again reproduced as table (1) and includes an extra row indicating the worst or minimum outcome associated with each course of action (strategy).

States of Nature	**Courses of Action**		
	S_1	**S_2**	**S_3**
N_1	7,00,000	5,00,000	3,00,000
N_2	3,00,000	4,50,000	3,00,000
N_3	1,50,000	0	3,00,000
Minimum payoff	1,50,000	0	3,00,000

Table(1)

Since this is associated with the worst possible outcomes of decrease in sales worth Rs. 3,00,000 the optimal course of action (or strategy) is obtained S_3 applying the maximin criterion.

b) Including an extra row representing the maximum payoff associated with each course of action and then applying the criterion of maximax, the optimal course of action is S_1, since this associates with it a maximum outcomes of Rs. 7,00,000 as shown in table (2).

c) The minimum value among the maximum regret as shown in table (3) is zero and this corresponds to course of action S_1.

States of Nature	Courses of Action		
	S_1	S_2	S_3
N_1	7,00,000	5,00,000	3,00,000
N_2	3,00,000	4,50,000	3,00,000
N_3	1,50,000	0	3,00,000
Maximum payoff	7,00,000	5,00,000	3,00,000

Table (2)

State of Nature	Courses of Action		
	S_1	S_2	S_3
N_1	7,00,000 – 7,00,000 = 0	7,00,000 – 5,00,000 = 2,00,000	7,00,000 – 3,00,000 = 4,00,000
N_2	4,50,000-3,00,000 = 1,50,000	4,50,000-4,50,000 = 0	4,50,000-3,00,000 = 1,50,000
N_3	3,00,000-1,50,000	3,00,000-0 = 3,00,000	3,00,000-3,00,000 = 0
Maximum regret	1,50,000	3,00,000	4,00,000

Table (3)

d) Here it is assumed that each course of action has a probability of occurrence equal to 1/3. Therefore, expected returns can be obtained as shown in table (4).

Course of Action	Expected Return
S_1	1/3 (7,00,000 + 3,00,000 + 1,50,000) = 3,83,333.33
S_2	1/3 (5,00,000 + 4,50,000 + 0) = 3,16,666.66
S_3	1/3 (3,00,000 + 3,00,000 + 3,00,000) = 3,00,000

Table (4)

Thus, Laplace criterion suggest that the executive should choose the strategy 5.

Q14. Explain about decision making under risk.

Ans :

Decision-making under risk assumes the long-run relative frequency of the states of nature occurrence to be given and besides this it also enumerates several states of nature. The state of natures information is probabilistic in nature i.e., the decision maker cannot predict which outcome will occur as a result of selecting a particular course of action. As each course of action results in more than one outcome, it is not easy to calculate the exact monetary payoffs or outcomes for the various combination of courses of action and states of nature.

The decision maker with the help of the past records or experience assigns probabilities to the likely possible occurrence of each state of nature. Once the probability distribution of the states of nature is known, then the best course of action must be selected which yields the highest expected payoffs.

The most widely used criteria for evaluating the alternative courses of action, is the Expected Monetary Value (EMV) which is also called as expected utility. The objective of decision-making under this condition is to optimize the expected payoff.

Example

An electrical manufacturing company has seen its business expanded to the point where it needs to increase production beyond its existing capacity. It has narrowed the alternatives to two approaches to increase the maximum production capacity, (a) Expansion, at a cost of Rs. 8 million, or (b) Mod-ernization at a cost of Rs. 5 million. Both approaches would require the same amount of time for implementation. Management believes that over the required payback period, demand will either be high or moderate. Since high demand is considered to be somewhat less likely than moderate demand, the probability of high demand has been setup at 0.35.

If the demand is high, expansion would gross an estimated additional Rs.12 million but modernization only an additional Rs. 6 million, due to lower maximum production capability. On the other hand, if the demand is moderate, the comparable figures would be Rs. 7 million for expansion and Rs. 5 million for modernization.

a) Calculate conditional profit in relation to various action and outcome combinations and states of nature.

b) If company wishes to maximize its expected monetary value, then it should modernize or expand?

c) Calculate the EVPI.

d) Construct the conditional opportunity loss table and also calculate EOL.

Sol :

a) Defining the state of nature of outcome (over which the company has no control) and course of action (company's possible decision).

Let,

States of nature : 0_1 = High demand, O_2 = Moderate demand

Courses of action : S_1 = Expand, S_2 = Modernize

Since the probability that the demand is high (outcome O_1) is estimated to 0.35, the probability of moderate demand (outcome O_2 must be (1 – 0.35) = 0.65. The calculations for conditional profit values are as follows,

	Course of Action	
State of Nature Oj	**S_1 (Expand)**	**S_2 (Modernize)**
O_1 (high demand)	12 – 8 = 4	6 – 5 = 1
O_2 (moderate demand)	7 – 8 = –1	5 – 5 = 0

Table (1) : Conditional Profit (Million Rs.)

b) The payoff table (1) can be rewritten as follows along with the given probabilities of states of nature.

		Course of Action	
State of Nature Oj	**Probability P(Oj)**	**S_1 (Expand)**	**S_2 (Modernize)**
O_1 (high demand)	0.35	4	1
O_2 (moderate demand)	0.65	–1	0

Table (2) : Conditional Profit (Million Rs.)

The calculation of EMVs for courses of action 5, and S., are given below,

EMV(S_1) = (0.35)(4) + (0.65)(–1) = 1.40 – 0.65 = Rs. 0.75 million

EMV(S_2) = (0.35)(1) + (0.65)(0) = 0.35 = 0.35 million

To maximize EMV, the company must expand course of action. The EMV of the optimal course of action is generally denoted by EMV*. Therefore,

EMV* = EMV(S_1) Rs. 0.75 million

c) To compute EVPI, we shall first calculate EPPL. For calculating EPPI, we choose the optimal course of action for each state of nature, multiply its conditional profit by the given probability to get weighted profit and then sum these weights as shown in the table (3).

State of Nature Oj	Probability P(Oj)	Optimal Course of Action	Conditional Profit	Weighted Profits
O_1	0.35	S_1	4	4 × 0.35 = 1.40
O_2	0.65	S_2	0	0 × 0.65 = 0
			EPPI	1.40

Table (3): Profit of Optimal Course of Action

The optimal EMV* is Rs. 0.75 million corresponding to the course of action S_1. Then,

EVPI = EPPI – EMV (S_1) = 1.40 – 0.75 = Rs. 0.65 million

Alternately, if the company could get a perfect information (for forecast) of demand (high or moderate) it should consider paying upto 0.65 million for an information.

The expected value of perfect information in business helps in getting and absolute upper bound on the amount that should be spent to get additional information on which a given decision is based.

d) The opportunity loss value are shown below.

State of Nature	Probability P(Oj)	Conditional Profit (Rs. million) Course of Action		Loss (Rs.million) Courses of Action	
O_j		S_1	S_2	S_1	S_2
O_1	0.35	4	1	0	3
O_2	0.65	–1	0	1	0

Table (4): Conditional Opportunity Loss Table

The conditional opportunity loss values may be explained as, if outcome O_1 occurred, then the maximum profit of Rs.4 million would be achieved by selecting course of action Thus, the choice of S_1 would result in zero opportunity loss, as it is the best decision if outcome O_1 occurs. If course of action S_2 were chosen with a payoff of one million, then this would result in a opportunity loss of 4 – 13 millions. If the outcome O_2 occurred, then the best course of action would be with zero loss. Thus, no opportunity loss would be associated with the choice of S_2. But, if S_1 were chosen, then the opportunity loss would be 0 – (–1) = Rs. 1 million. That is, the company would have Rs. 1 million worse off in that situation, if it had chosen course of action S_1.

Using the given forecast of probabilities associated with each state of nature $P(O_1) = 0.35$ and $P(O_2) = 0.65$, the expected opportunity losses for the two courses of action are,

EOL(S_1) = 0.35(0) + 0.65(1) = Rs. 0.65 million

EOL(S_2) = 0.35(3) + 0.65(0) = Rs. 0.05 million

Since decision maker seeks to minimize the expected opportunity loss, he must select course of action S_1 to produce the smallest expected opportunity loss.

Short Question and Answers

1. What is linear programming problem.

Ans :

In 1947, George dantzig and his associates, while working in the U.S. department of Air Force, observed that a large of military programming and planning problems could be formulated as maximizing/minimizing a linear form of profit/cost function whose variables were restricted to values satisfying a system of linear constraints.

A linear form is meant a thematical expression of the type $a_1x_1 + a_2x_2 + ... + a_nx_n$, where $a_1, a_2, ..., a_n$ are constants and $x_1, x_2, ..., x_n$ are variables. The term 'Programming' refers to the process of determining a particular programme or plan of action. So Linear Programming (L.P.) is one of the most important optimization (maximization/minimization) techniques developed in the field of Operations Research.

The geneal LPP calls for optimizing (maximizing/minimizing) a linear function of variables called the '*Objective Function*' subject to a set of linear equations and / or inequalities called the '*constraints*' or *restrictions*.

2. Limitations of Linear Programming

Ans :

i) In some problems objective functions and constraints are not linear. Generally, in real life situations concerning business and industrial problems constraints are not linearly created to variables.

ii) There is no guarantee of getting integer valued solutions, for example, in finding out how may men and machines would be required to perform a particular job, rounding off the solution to the nearest integer will not give an optimal solution. Integer programming deals with such problems.

iii) Linear programming model does not take into consideration the effect of time and uncertainty. Thus the model should be defined in such a way that any change due to internal as well as external factors can be incorporated.

iv) Sometimes large-scale problems cannot be solved with linear programming techniques even when the computer facility is available. Such difficulty may be removed by decomposing the main problem into several small problems and then solving them separately.

v) Parameters appearing in the model are assumed to be constant. But, in real life situations they are neither constant not deterministic.

vi) Linear programming deals with only single objective, whereas in real life situations problems come across with multiobjectives.

3. Advantages of linear programming problem.

Ans :

i) It helps in organisation and study of the information in the same way that the scientific approach to the problem requires.

ii) With LP the execute builds into his planning a true reflection of the limitations and restrictions under which he must operate.

iii) Once a basic plan is arrived at through LP, it can be re-evaluated for changing conditions.

iv) Highlighting of bottlenecks in the production process is the striking advanages of this technique.

v) It provides flexibility in analysing a variety of multi-dimensional problems.

4. Explain Non Linear Programming?

Ans :

- Non Linear Programming (NLP). If an LP problem is feasible then, at least in theory, it can always be solved because. We know the solution is a “corner point”: a point where lines or planes intersect.
- There are a finite number of possible solution points. The simplex algorithm will find that point. Also, a very informative sensitivity analysis is relatively easy to obtain for LP problems. But in many interesting, real-world problems, the objective function may not be a linear function, or some of the constraints may not be linear constraints
- Optimization problems that involve nonlinearities are called nonlinear programming (NLP) problems. Many NLPs do not have any constraints. They are called unconstrained NLPs. Solutions to NLPs are found using search procedures.
- Solutions are more difficult to determine, compared to LPs. One problem is difficulty in distinguishing between a local and global minimum or maximum point.
- Nonlinear programming is the process of solving an optimization problem where some of the constraints or the objective function are nonlinear.

5. Cutting Plane Method

Ans :

- In mathematical optimization, the cutting-plane method is any of a variety of optimization
- Methods that iteratively refine a feasible set or objective function by means of linear inequalities, termed cuts. Such procedures are commonly used to find integer solutions to mixed integer linear programming (MILP) problems, as well as to solve general, not necessarily differentiable convex optimization problems. The use of cutting planes to solve MILP was introduced by Ralph E. Gomory.
- Cutting plane methods for MILP work by solving a non-integer linear program, the linear relaxation of the given integer program.
- The theory of Linear Programming dictates that under mild assumptions (if the linear program has an optimal solution, and if the feasible region does not contain a line), one can always find an extreme point or a corner point that is optimal.

6. Decision analysis

Ans :

The term decision analysis was coined in 1964 by Ronald A. Howard, professor of management science and engineering at Stanford University. Decision analysis refers to a systematic, quantitative and interactive approach to addressing and evaluating important choices confronted by organisations in the private and public sector. Decision analysis is interdisciplinary and draws on theories from the fields of psychology, economics, and management science. It utilises a variety of tools which include models for decision-making under conditions of uncertainty or multiple objectives; techniques of risk analysis and risk assessment; experimental and descriptive studies of decision-making behaviour; economic analysis of competitive and strategic decisions; techniques for facilitating decision-making by groups; and computer modeling software and expert systems for decision support.

Choose the Correct Answer

1. For a linear programming equations, convex set of equations is included in region of ________. [a]

   (a) feasible solutions (b) disposed solutions

   (c) profit solutions (d) loss solutions

2. In linear programming, objective function and objective constraints are ________ . [b]

   (a) solved (b) linear

   (c) quadratic (d) adjacent

3. In graphical solutions of linear inequalities, solution can be divided into ________ . [b]

   (a) one subset (b) two subsets

   (c) three subsets (d) four subsets

4. Objective of linear programming for an objective function is to ________.[a]

   (a) maximize or minimize (b) subset or proper set modeling

   (c) row or column modeling (d) adjacent modeling

5. Linear programming is used to optimize mathematical procedure and is ________. [a]

   (a) Subset of mathematical programming

   (b) Dimension of mathematical programming

   (c) Linear mathematical programming

   (d) All of above

6. One of the models of operation research includes [c]

   (a) LPP (b) Networking

   (c) Game theory (d) Transportation

7. The graphical method of LP problem uses ________ [d]

   (a) Constraint equations (b) Objective function equations

   (c) Linear equations (d) All of the above

8. Alternative solutions exist in an LP model when ________ [a]

   (a) Objective function equation is parallel to one of the constraints

   (b) One of the constraints is redundant

   (c) Two constraints are parallel

   (d) Adding another constraint/variable.

9. One of the following is an assumption of linear programming model [c]

   (a) $\geq$ or $\leq$ constraints (b) Maximize profits

   (c) Divisibility (d) Minimize cost

10. If a non-redundant constraint is removed from a LP problem, then [a]

   (a) Feasible region will become larger

   (b) Feasible region will become smaller

   (c) Solution will become infeasible

   (d) The solution is unbounded

Fill in the blanks

1. The geneal LPP calls for optimizing (maximizing/minimizing) a linear function of variables called the ________.
2. LPP stands for ________.
3. Simplex algorithm was originally proposed by ________.
4. The best use of linear programming technique is to find an optimal use of ________.
5. ________ programming is the process of solving an optimization problem where some of the constraints or the objective function are nonlinear.
6. MILP stands for ________.
7. ________ methods for general convex continuous optimization.
8. The term decision analysis was coined in ________.
9. ________ is interdisciplinary and draws on theories from the fields of psychology, economics, and management science.
10. EMV stands for ________.

Answers

1. Objective Function
2. Linear Programming Problem
3. G.B Dantzig in 1948
4. Money, manpower and machine
5. Nonlinear
6. Mixed integer linear programming
7. Cutting-plane
8. 1964
9. Decision analysis
10. Expected Monetary Value

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 Programming Using R

Q1. What is R? Explain the features of R?

Ans :

R is a programming language developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithm, linear regression, time series, statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational task, C, C++ and Fortran codes are preferred.

R is not only entrusted by academic, but many large companies also use R programming language, including Uber, Google, Airbnb, Facebook and so on.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modeling and communicate the results

- **Program**: R is a clear and accessible programming tool
- **Transform**: R is made up of a collection of libraries designed specifically for data science
- **Discover**: Investigate the data, refine your hypothesis and analyze them
- **Model**: R provides a wide array of tools to capture the right model for your data
- **Communicate**: Integrate codes, graphs, and outputs to a report with R Markdown or build Shiny apps to share with the world

Features of R

As R is a leading programming language. There are so many features of **R programming** which makes it important to learn. Let's discuss them one by one.

Statistical Features of R

1. **R has some topical relevance**

- It is free, open source software.
- R is available under free software Foundation.

2. **R has some statistical features**

- **Basic Statistics :** Mean, variance, median.
- **Static graphics :** Basic plots, graphic maps.
- **Probability distributions :** Beta, Binomial.

Any Doubt yet in Why Learn R programming? Please Comment.

Programming Features of R

1. **R has some topical relevance**

- Data inputs such as data type, **importing data**, keyboard typing.
- Data Management such as data variables, operators.

2. **R has some programming features**

- **Distributed Computing** – Distributed computing is an open source, high-performance platform for the R language. It splits tasks between multiple processing nodes to reduce execution time and analyze large datasets.
- **R packages** – **R packages** are a collection of **R functions**, compiled code and sample data. By default, **R installs** a set of packages during installation.

Q2. Explain the basic tips for using R?

Ans :

- R is command-line driven. It requires you to type or copy-and-paste commands after a command prompt (>) that appears when you open R. After typing a command in the R console and pressing **Enter** on your keyboard, the command will run. If your command is not complete, R issues a continuation prompt (signified by a plus sign: +). Alternatively you can write a script in the script window, and select a command, and click the **Run** button.

- R is case sensitive. Make sure your spelling and capitalization are correct.
- Commands in R are also called functions. The basic format of a function in R is: function.name(argument, options).
- The up arrow (^) on your keyboard can be used to bring up previous commands that you've typed in the R console.
- The $ symbol is used to select a particular column within the table (e.g., table$column).
- Any text that you do not want R to act on (such as comments, notes, or instructions) needs to be preceded by the # symbol (a.k.a. hash-tag, comment, pound, or number symbol). R ignores the remainder of the script line following

  For example: Plot(x, y) # This text will not affect the plot function because of the comment.

Q3. What is R environment.

Ans :

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- An effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either on-screen or on hardcopy, and
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via *packages*. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

Q4. Explain the various types of operators in R program.

Ans :

1. Arithmetic Operators
2. Relational Operators
3. Logical Operators
4. Assignment Operators
5. Miscellaneous Operators

1. Arithmetic Operators

Following table shows the arithmetic operators supported by R language. The operators act on each element of the vector.

Operator	Description	Example

+	Adds two vectors	v <- c(2, 5.5, 6)
		t <- c(8, 3, 4)
		print(v+t)
		It produces the following result:
		[1] 10.0 8.5 10.0

–	Subtracts second vector from the first	v <- c(2, 5.5, 6)
		t <- c(8, 3, 4)
		print(v-t)
		It produces the following result:
		[1]-6.0 2.5 2.0

*	Multiplies both vectors	v <- c(2, 5.5, 6)
		t <- c(8, 3, 4)
		print(v*t)
		It produces the following result:
		[1] 16.0 16.5 24.0

/	Divides the first vector with the second	v <- c(2,5.5, 6)
		t<- c(8, 3,4) print(v/t)
		When we execute the above code, it produces the following result:
		[1] 0.250000 1.833333 1.500000

%%	Gives the remainder of the first vector with the second	v <- c(2,5.5,6)
		t <- c(8, 3, 4)
		print(v%%t)
		it produces the following result -
		[1] 2.0 2.5 2.0

% / %	The result of division of first vector with second (quotient)	v <- c(2,5.5,6)
		t <- c(8, 3,4)
		print(v%/%t)
		It produces the following result:
		[1] 0 1 1

^	The first vector raised to the exponent of, second vector	v <- c(2,5.5,6)
		t<- c(8, 3,4)
		print(v ^ t)
		It produces the following result:
		[1] 256.000 166.375 1296.000

2. Relational Operators

Following table shows the relational operators supported by R language. Each element of the first vector is compared with the corresponding element of the second vector. The result of comparison is a Boolean value.

Operator	Description	Example

>	Checks if each element of the first vector is greater than the corresponding element of the second vector.
	It produces the following result:
		v <- c(2,5.5,6,9)
		t <_ c(8,2.5,14,9)
		print(v>t)
		[1] FALSE TRUE FALSE FALSE

<	Checks if each element of the first vector is less than the corresponding element of the second vector.
	It produces the following result:
		v <- c(2,5.5,6,9)
		t <- c(8,2.5,14,9)
		print(v < t)
		[1] TRUE FALSE TRUE FALSE

= =	Checks if each element of the first vector is equal to the corresponding element of the second vector.
	It produces the following result:
		v <- c(2,5.5,6,9)
		t <- c(8,2.5,14,9)
		print(v = = t)
		[1]FALSE FALSE FALSE TRUE

<=	Checks if each element of the first vector is less than or equal to the corresponding element of the second vector.	v <- c(2,5.5,6,9)
		t <- c(8,2.5,14,9)
		print(v<=t)
		It produces the following result:
		[1] TRUE FALSE TRUE TRUE

>=	Checks if each element of the first vector is greater than or equal to the corresponding element of the second	v <- c(2,5.5,6,9)
		t <- c(8,2.5,14,9)
		print(v>=t)
		It produces the following result:
		[1] FALSE TRUE FALSE TRUE

!=	Checks if each element of the first vector is unequal to the corresponding element of the second vector.	v<-c(2,5.5,6,9)
		t <- c(8,2.5,14,9)
		print(v!=t)
		It produces the following result:
		[1] TRUE TRUE TRUE FALSE

3. Logical Operators

Following table shows the logical operators supported by R language. It is applicable only to vectors of type logical, numeric or complex. All numbers greater than 1 are considered as logical value TRUE.

Each element of the first vector is compared with the corresponding element of the second vector. The result of comparison is a Boolean value.

Operator	Description	Example

&	It is called Element-wise Logical AND operator. It combines each element of the first vector with the corresponding element of the second vector and gives a output TRUE if both the elements are TRUE.	v <- c(3, l, TRUE, 2 + 3i)
		t <- c(4, l, FALSE, 2 +3i)
		print(v&t)
		It produces the following result:
		[1] TRUE TRUE FALSE TRUE

|	It is called Element-wise Logical OR operator. It combines each element of the first vector with the corresponding element of the second vector and gives a output TRUE if one the elements is TRUE.	v <- c(3,0,TRUE,2+2i)
		t <- c(4,0,FALSE,2+3i)
		print(v|t)
		It produces the following result:
		[1] TRUE FALSE TRUE TRUE

!	It is called Logical NOT operator. It takes each element of the vector and gives the opposite logical value.	v <- c(3,0,TRUE,2+2i)
		print(!v)
		It produces the following result:
		[1] FALSE TRUE FALSE FALSE

The logical operator && and || considers only the first element of the vectors and gives a vector of single element as output.

Operator	Description	Example

&&	It is called Logical AND operator. It takes first element of both the vectors and gives the TRUE only if both are TRUE.	v <- c(3,0,TRUE,2+2i)
		t <-c(1,3,TRUE,2+3i)
		print(v&&t)
		It produces the following result:
		[1] TRUE

||	It is called Logical OR operator. It takes first element of both the vectors and gives the TRUE if one of them is TRUE.	v <- c(0,0,TRUE,2+2i)
		t <-c(0,3,TRUE,2+3i)
		print(v||t)
		It produces the following result:
		[1] FALSE

Assignment Operators

These operators are used to assign values to vectors.

Operator	Description	Example

<–
or
=
or
<<–
	It is called Left Assignment.	v1 <- c(3,l,TRUE,2+3i)
		v2 <<-c(3,l,TRUE,2+3i)
		v3 = c(3,l,TRUE,2+3i)
		print(vl)
		print(v2)
		print(v3)
		It produces the following result:
		[1] 3+0i 1+Oi 1+Oi 2+3i
		[1] 3+0i 1+Oi 1+Oi 2+3i
		[1] 3+0i 1+Oi 1+Oi 2+3i

–> or –>>	It is called Right Assignment.	c(3, l, TRUE, 2+3i) -> v1 c(3,l,TRUE,2+3i) –>> v2 print(vl) print(v2) It produces the following result: [1] 3+0i 1+Oi 1+Oi 2+3 i [1] 3+0i 1+Oi 1+Oi 2+3i

Q5. What Is A Package?

Ans :

A package is a suitable way to organize your own work and, if you want to, share it with others. Typically, a package will include code (not only R code!), documentation for the package and for the functions inside, some tests to check everything works as it should, and data sets.

The basic information about a package is provided in the DESCRIPTION file, Which means a file contains basic information about the package where you can find out what the package does, who the author is, what version the documentation belongs to, the date, the type of license its use, and the package dependencies.

Besides finding the DESCRIPTION files such as cran.r-project.org or stat.ethz.ch, you can also access the description file inside R with the command package Description (“package”), via the documentation of the package help(package = “package”), or online in the repository of the package.

5.2 R PACKAGE

Q6. What is R package?

Ans :

R packages are a collection of R functions, complied 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.

All the packages available in R language are listed at R Packages.

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        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
tcltk	Tcl/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"
```

Q7. How to Install a R New Package.

Ans :

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.

(i) 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")
```

ii) 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")
```

iii) 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

Q8. How the data can be read and write in R?

Ans :

Reading Data in R

For reading, (importing) data into R following are some functions.

- read.table(), and read.csv(), for reading tabular data
- readLines() for reading lines of a text file
- source() for reading in R code files (inverse of dump)
- dget() for reading in R code files (inverse of dput)
- load() for reading in saved workspaces.

Writing Data in R

Following are few functions for writing (exporting) data to files.

- write.table(), and write.csv() exports data to wider range of file format including csv and tab-delimited.
- writeLines() write text lines to a text-mode connection.
- dump() takes a vector of names of R objects and produces text representations of the objects on a file (or connection). A dump file can usually be sourced into another R session.
- dput() writes an ASCII text representation of an R object to a file (or connection) or uses one to recreate the object.
- save() writes an external representation of R objects to the specified file.

Reading data files with read.table()

The read.table() function is one of the most commonly used functions for reading data into R. It 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 data set
- nrows, the number of rows in the dataset
- comment.char, a character string indicating the comment character
- skip, the number of lines to skip from the beginning
- stringsAsFactors, should character variables be coded as factors?

read.table() and read.csv() Examples

```
> data<-read.table(“foo.txt”)
> data<-read.table(“D:\\datafiles\\mydata.txt”)
> data<-read.csv(“D:\\datafiles\\mydata.csv”)
```

R will automatically skip lines that begin with a #, figure out how many rows there are (and how much memory needs to be allocated). R also figure out what type of variable is in each column of the table.

Writing data files with write.table()

Following are few important arguments usually used in write.table() function.

- x, the object to be written, typically a data frame
- file, the name of the file which the data are to be written to
- sep, the field separator string
- col.names, a logical value indicating whether the column names of x are to be written along with x, or a character vector of column names to be written
- row.names, a logical value indicating whether the row names of x are to be written along with x, or a character vector of row names to be written
- na, the string to use for missing values in the data

write.table() and write.csv() Examples

```
> x <- data.frame(a = 5, b = 10, c = pi)
> write.table(x, file = “data.csv”, sep = “,”)
> write.table(x, “c:\\mydata.txt”, sep = “\t”)
> write.csv(x, file = “data.csv”).
```

5.4 R Functions

Q9. What is R Function? Explain the components of R Functions.

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.

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
}
```

Components of R

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.

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.

Q10. Explain different types of functions?

Ans :

i) 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 frm 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
```

ii) 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)
}
}
```

iii) Calling a Function

```
  # Create a function to print squares of numbers in sequence.
  new.function <- function(a) {
       for(i in 1: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
```

iv) Calling a Function without an Argument

```
# Create a function without an argument.
new.function <- function() {
     for(i in 1:5) {
          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
```

v) 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
```

vi) 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
```

vii) Lazy Evaluation of Function

Arguments to functions are evaluated lazily, which means so they are evaluated only when needed by the function body.

```
# Create a function with arguments.
    new.function <- function(a, b) {
    print(a ^ 2)
    print(a)
    print(b)
}
```

# Evaluate the function without supplying one of the arguments.

new.function(6)

When we execute the above code, it produces the following result “

[1] 36

[1] 6

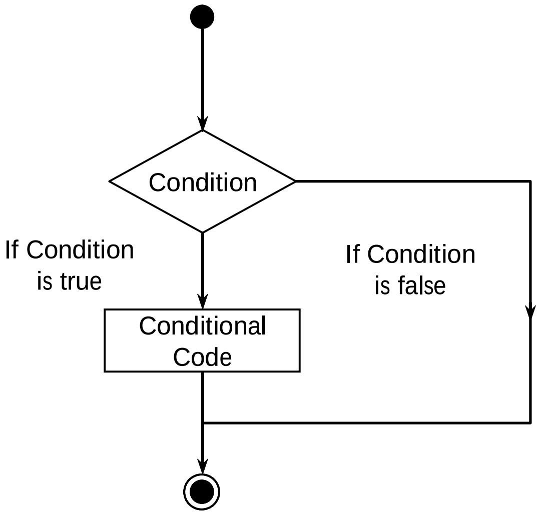
Error in print(b) : argument “b” is missing, with no default.

5.5 Control Statements

Q11. Explain briefly about control statements.

Ans :

Looping is similiar to creating functions in that they are merely a means to automate a certain multi step process by organizing sequences of R expressions. R consists of several loop control statements which allow you to perform repetititve code processes with different intentions and allow these automated expressions to naturally respond to features of your data. Consequently, learning these loop control statements will go a long ways in reducing code redundancy and becoming a more efficient data wrangler.



- **if** statement for conditional programming
- **if...else** statement for conditional programming

- for loop to iterate over a fixed number of iterations
- while loop to iterate until a logical statement returns FALSE
- repeat loop to execute until told to break
- break/next arguments to exit and skip interations in a loop.

Q12. Explain briefly about if statement.

Ans :

The conditional if statement is used to test an expression. If the test_expression is TRUE, the statement gets executed. But if it's FALSE, nothing happens.

```
# syntax of if statement
if (test_expression) {
      statement
}
```

The following is an example that tests if any values in a vector are negative. Notice there are two ways to write this if statement; since the body of the statement is only one line you can write it with or without curly braces. I recommend getting in the habit of using curly braces, that way if you build onto if statements with additional functions in the body or add an else statement later you will not run into issues with unexpected code procedures.

```
        x <- c(8, 3, -2, 5)
# without curly braces
if(any(x < 0)) print("x contains negative numbers")
## [1] "x contains negative numbers"
# with curly braces produces same result
if(any(x < 0)){
      print("x contains negative numbers")
}
## [1] "x contains negative numbers"
# an if statement in which the test expression is FALSE
# does not produce any output
```

```
y <- c(8, 3, 2, 5)
if(any(y < 0)){
      print("y contains negative numbers")
}.
```

Q13. Explain briefly about if.... else statement?

Ans :

The conditional if...else statement is used to test an expression similar to the if statement. However, rather than nothing happening if the test_expression is FALSE, the else part of the function will be evaluated.

```
# syntax of if...else statement
if (test_expression) {
      statement 1
} else {
      statement 2
}
```

The following extends the previous example illustrated for the if statement in which the if statement tests if any values in a vector are negative; if TRUE it produces one output and if FALSE it produces the else output.

```
# this test results in statement 1 being executed
x <- c(8, 3, -2, 5)
if(any(x < 0)){
      print("x contains negative numbers")
} else{
      print("x contains all positive numbers")
}
## [1] "x contains negative numbers"
# this test results in statement 2 (or the else statement) being executed
y <- c(8, 3, 2, 5)
```

```
if(any(y < 0)){
       print(“y contains negative numbers”)
} else{
       print(“y contains all positive numbers”)
}
## [1] “y contains all positive numbers”
```

Simple if...else statements, as above, in which only one line of code is being executed in the statements can be written in a simplified alternative manner. These alternatives are only recommended for very short if...else code:

```
x <- c(8, 3, 2, 5)
# alternative 1
if(any(x < 0)) print(“x contains negative numbers”) else print(“x contains all positive numbers”)
## [1] “x contains all positive numbers”
# alternative 2 using the ifelse function
ifelse(any(x < 0), “x contains negative numbers”, “x contains all positive numbers”)
## [1] “x contains all positive numbers”
```

We can also nest as many if...else statements as required (or desired). For example:

```
# this test results in statement 1 being executed
x <- 7
if(x >= 10){
       print(“x exceeds acceptable tolerance levels”)
} else if(x >= 0 & x < 10){
       print(“x is within acceptable tolerance levels”)
} else {
        print(“x is negative”)
}
## [1] “x is within acceptable tolerance levels”
```

Q14. Define loop? Explain different kinds of loops in R programming.

Ans :

A loop statement allows us to execute a statement or group of statements multiple times. The following is the general form of a loop statement in most of the programming languages:

- **repeat loop:** Executes a sequence of statements multiple times and abbreviates the code that manages the loop variable.
- **while loop:** Repeats a statement or group of statements while a given condition is true. It tests the condition before executing the loop body.
- **for loop:** Like a while statement, except that it tests the condition at the end of the loop body.

Q15. Explain briefly about loop statement?

Ans :

The for loop is used to execute repetitive code statements for a particular number of times. The general syntax is provided below where i is the counter and as i assumes each sequential value defined (1 through 100 in this example) the code in the body will be performed for that ith value.

```
# syntax of for loop
for(i in 1:100) {
      <do stuff here with i>
}
```

For example, the following for loop iterates through each value (2010, 2011, …, 2016) and performs the paste and print functions inside the curly brackets.

```
for (i in 2010:2016){
      output <- paste("The year is", i)
      print(output)
}
## [1] "The year is 2010"
## [1] "The year is 2011"
## [1] "The year is 2012"
## [1] "The year is 2013"
## [1] “The year is 2014”
## [1] “The year is 2015”
## [1] “The year is 2016”
```

If you want to perform the for loop but have the outputs combined into a vector or other data structure than you can initiate the output data structure prior to the for loop. For instance, if we want to have the previous outputs combined into a single vector x we can initiate x first and then append the for loop output to x.

```
x <- NULL
for (i in 2010:2016){
      output <- paste(“The year is”, i)
      x <- append(x, output)
}
x
## [1] “The year is 2010” “The year is 2011” “The year is 2012” “The year is 2013”
## [5] “The year is 2014” “The year is 2015” “The year is 2016”
```

However, an important lesson to learn is that R is not efficient at *growing* data objects. As a result, it is more efficient to create an empty data object and *fill* it with the for loop outputs. In the previous example we *grew* x by appending new values to it. A more efficient practice is to initiate a vector (or other data structure) of the right size and fill the elements. In the example that follows, we create the vector x of the right size and then fill in each element within the for loop. Although this inefficiency is not noticed in this small example, when you perform larger repetitions it will become noticable so you might as well get in the habit of *filling* rather than *growing*.

```
x <- vector(mode = “numeric”, length = 7)
counter <- 1
for (i in 2010:2016){
      output <- paste(“The year is”, i)
      x[counter] <- output
      counter <- counter + 1
}
```

```
x
## [1] "The year is 2010" "The year is 2011" "The year is 2012" "The year is 2013"
## [5] "The year is 2014" "The year is 2015" "The year is 2016"
```

Another example in which we create an empty matrix with 5 rows and 5 columns. The for loop then iterates over each column (note how *i* takes on the values 1 through the number of columns in the my.mat matrix) and takes a random draw of 5 values from a poisson distribution with mean *i* in column *i*:

```
my.mat <- matrix(NA, nrow = 5, ncol = 5)
for(i in 1:ncol(my.mat)){
    my.mat[, i] <- rpois(5, lambda = i)
}
my.mat
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    0    2    1    7    1
## [2,]    1    2    2    3    9
## [3,]    2    1    5    6    6
## [4,]    2    1    5    2   10
## [5,]    0    2    2    2    4
```

Q16. Explain briefly about while loop?

Ans :

While loops begin by testing a condition. If it is true, then they execute the statement. Once the statement is executed, the condition is tested again, and so forth, until the condition is false, after which the loop exits. It's considered a best practice to include a counter object to keep track of total iterations

```
# syntax of while loop
counter <- 1
while(test_expression) {
    statement
    counter <- counter + 1
}
```

while loops can potentially result in infinite loops if not written properly; therefore, you must use them with care. To provide a simple example to illustrate how similiar for and while loops are:

```
counter <- 1
while(counter <= 10) {
    print(counter)
    counter <- counter + 1
}
# this for loop provides the same output
counter <- vector(mode = “numeric”, length = 10)
for(i in 1:length(counter)) {
    print(i)
}
```

The primary difference between a for loop and a while loop is: a for loop is used when the number of iterations a code should be run is known where a while loop is used when the number of iterations is not known. For instance, the following takes value x and adds or subtracts 1 from the value randomly until x exceeds the values in the test expression. The output illustrates that the code runs 14 times until x exceeded the threshold with the value 9.

```
counter <- 1
x <- 5
set.seed(3)
while(x >= 3 && x <= 8 ) {
    coin <- rbinom(1, 1, 0.5)
    if(coin == 1) { ## random walk
        x <- x + 1
    } else {
        x <- x - 1
    }
    cat(“On iteration”, counter, “, x =”, x, ‘\n’)
    counter <- counter + 1
}
```

```
## On iteration 1 , x = 4
## On iteration 2 , x = 5
## On iteration 3 , x = 4
## On iteration 4 , x = 3
## On iteration 5 , x = 4
## On iteration 6 , x = 5
## On iteration 7 , x = 4
## On iteration 8 , x = 3
## On iteration 9 , x = 4
## On iteration 10 , x = 5
## On iteration 11 , x = 6
## On iteration 12 , x = 7
## On iteration 13 , x = 8
## On iteration 14 , x = 9
```

Q17. Explain briefly about repeat loop?

Ans :

A repeat loop is used to iterate over a block of code multiple number of times. There is test expression in a repeat loop to end or exit the loop. Rather, we must put a condition statement explicitly inside the body of the loop and use the break function to exit the loop. Failing to do so will result into an infinite loop.

```
# syntax of repeat loop
counter <- 1
repeat {
    statement

    if(test_expression){
        break
    }
    counter <- counter + 1
}
```

For example ,say we want to randomly draw values from a uniform distribution between 1 and 25. Furthermore, we want to continue to draw values randomly until our sample contains at least each integer value between 1 and 25; however, we do not care if we've drawn a particular value multiple times. The following code repeats the random draws of values between 1 and 25 (in which we round). We then include an if statement to check if all values between 1 and 25 are present in our sample. If so, we use the break statement to exit the loop. If not, we add to our counter and let the loop repeat until the conditional ifstatement is found to be true. We can then check the counter object to assess how many iterations were required to reach our conditional requirement.

```
counter <- 1
x <- NULL
repeat {
     x <- c(x, round(runif(1, min = 1, max = 25)))

     if(all(1:25 %in% x)){
          break
     }

     counter <- counter + 1
}
counter
## [1] 75
```

break/next arguments

The break argument is used to exit a loop immediately, regardless of what iteration the loop may be on. break arguments are typically embedded in an if statement in which a condition is assessed, if TRUE break out of the loop, if FALSE continue on with the loop. In a nested looping situation, where there is a loop inside another loop, this statement exits from the innermost loop that is being evaluated.

In this example, the for loop will iterate for each element in x; however, when it gets to the element that equals 3 it will break out and end the for loop process.

```
x <- 1:5
for (i in x) {
      if (i == 3){
            break
            }
      print(i)
}
## [1] 1
## [1] 2
```

The next argument is useful when we want to skip the current iteration of a loop without terminating it. On encountering next, the R parser skips further evaluation and starts the next iteration of the loop. In this example, the forloop will iterate for each element in x; however, when it gets to the element that equals 3 it will skip the for loop execution of printing the element and simply jump to the next iteration.

```
x <- 1:5
for (i in x) {
      if (i == 3){
            next
            }
      print(i)
}
## [1] 1
## [1] 2
## [1] 4
## [1] 5
```

5.6 Frames and Subsets

Q18. Explain briefly about data frame in R?

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.

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.

i) 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),
         Join_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:

S.No.	emp_id	emp_name	salary	Join_date
1	1	Rick	623.30	2012-01-01
2	2	Dan	515.20	2013-09-23
3	3	Michelle	611.00	2014-11-15
4	4	Ryan	729.00	2014-05-11
5	5	Gary	843.25	2015-03-27

Get the Structure of the Data Frame

The structure of the data frame can be seen by using **str()** function.

```
# 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),
Join_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15",
                           "2014-05-11", "2015-03-27")),
stringsAsFactors = FALSE
)
# Get the structure of the data frame.
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
$ Join_date: Date, format: "2012-01-01" "2013-09-23" "2014-11-15" "2014-05-11" ...
```

ii) Summary of Data in Data Frame

The statistical summary and nature of the data can be obtained by applying **summary()** function.

```
# 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),
      Join_date = as.Date(c(“2012-01-01”, “2013-09-23”, “2014-11-15”,
                                  “2014-05-11”, “2015-03-27”)),
  stringsAsFactors = FALSE
)
# Print the summary.
print(summary(emp.data))
```

When we execute the above code, it produces the following result “

emp_id	emp_name	salary	Join_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

iii) Extract Data from Data Frame

Extract specific column from a data frame using column name.

```
# 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),
Join_date = as.Date(c(“2012-01-01”,”2013-09-23",”2014-11-15",
                         ”2014-05-11", “2015-03-27”)),
stringsAsFactors = FALSE
)
# 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

Extract the first two rows and then all columns

```
# 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),
    Join_date = as.Date(c(“2012-01-01”, “2013-09-23”, “2014-11-15”,
                                 “2014-05-11”, “2015-03-27”)),
   stringsAsFactors = FALSE
)
# Extract first two rows.
result <- emp.data[1:2,]
print(result)
```

When we execute the above code, it produces the following result”

emp_id	emp_name	salary	start_date
1 1	Rick	623.3	2012-01-01
2 2	Dan	515.2	2013-09-23

Extract 3rd and 5th row with 2nd and 4th column

```
# 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),
  Join_date = as.Date(c(“2012-01-01”, “2013-09-23”, “2014-11-15”, “2014-05-11”,
   “2015-03-27”)),
  stringsAsFactors = FALSE
)
# Extract 3rd and 5th row with 2nd and 4th column.
result <- emp.data[c(3,5),c(2,4)]
print(result)
```

When we execute the above code, it produces the following result

emp_name	Join_date
3 Michelle	2014-11-15
5 Gary	2015-03-27

iv) Expand Data Frame

A data frame can be expanded by adding columns and rows.

Add Column

Just add the column vector using a new column name.

```
# 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),
  Join_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
     "2015-03-27")),
   stringsAsFactors = FALSE
)
# Add the "dept" coulmn.
emp.data$dept <- c("IT","Operations","IT","HR","Finance")
v <- emp.data
print(v)
```

When we execute the above code, it produces the following result “

S.No.	emp_id	emp_name	salary	Join_date	dept
1	1	Rick	623.30	2012-01-01	IT
2	2	Dan	515.20	2013-09-23	Operations
3	3	Michelle	611.00	2014-11-15	IT
4	4	Ryan	729.00	2014-05-11	HR
5	5	Gary	843.25	2015-03-27	Finance

Add Row

To add more rows permanently to an existing data frame, we need to bring in the new rows in the same structure as the existing data frame and use the **rbind()** function.

In the example below we create a data frame with new rows and merge it with the existing data frame to create the final data frame.

```
# Create the first 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),
  Join_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
                              "2015-03-27")),
   dept = c("IT","Operations","IT","HR","Finance"),
   stringsAsFactors = FALSE
)
# Create the second data frame
emp.newdata <- data.frame (
   emp_id = c (6:8),
   emp_name = c("Rasmi","Pranab","Tusar"),
   salary = c(578.0,722.5,632.8),
   Join_date = as.Date(c("2013-05-21","2013-07-30","2014-06-17")),
   dept = c("IT","Operations","Fianance"),
   stringsAsFactors = FALSE
)
# Bind the two data frames.
emp.finaldata <- rbind(emp.data,emp.newdata)
print(emp.finaldata)
```

When we execute the above code, it produces the following result "

S.No.	emp_id	emp_name	salary	Join_date	dept
1	1	Rick	623.30	2012-01-01	IT
2	2	Dan	515.20	2013-09-23	Operations
3	3	Michelle	611.00	2014-11-15	IT
4	4	Ryan	729.00	2014-05-11	HR
5	5	Gary	843.25	2015-03-27	Finance
6	6	Rasmi	578.00	2013-05-21	IT
7	7	Pranab	722.50	2013-07-30	Operations
8	8	Tusar	632.80	2014-06-17	Fianance

Q19. What is mean by subsets in R ? Explain in detail?

Ans :

Subsetting Data

R has powerful indexing features for accessing object elements. These features can be used to select and exclude variables and observations. The following code snippets demonstrate ways to keep or delete variables and observations and to take random samples from a dataset.

Selecting (Keeping) Variables

```
# select variables v1, v2, v3
myvars <- c(“v1”, “v2”, “v3”)
newdata <- mydata[myvars]
# another method
myvars <- paste(“v”, 1:3, sep=””)
newdata <- mydata[myvars]
# select 1st and 5th thru 10th variables
newdata <- mydata[c(1,5:10)]
```

To practice this interactively, try the selection of data frame elements exercises in the Data frames chapter of this introduction to R course.

Excluding (DROPPING) Variables

```
# exclude variables v1, v2, v3
myvars <- names(mydata) %in% c(“v1”, “v2”, “v3”)
newdata <- mydata[!myvars]
# exclude 3rd and 5th variable
newdata <- mydata[c(-3,-5)]
# delete variables v3 and v5
mydata$v3 <- mydata$v5 <- NULL
```

Selecting Observations

```
# first 5 observations
newdata <- mydata[1:5,]
# based on variable values
newdata <- mydata[ which(mydata$gender=='F'
& mydata$age > 65), ]
# or
attach(mydata)
newdata <- mydata[ which(gender=='F' & age > 65),]
detach(mydata)
```

Selection using the Subset Function

The **subset( )** function is the easiest way to select variables and observations. In the following example, we select all rows that have a value of age greater than or equal to 20 or age less then 10. We keep the ID and Weight columns.

```
# using subset function
newdata <- subset(mydata, age >= 20 | age < 10,
select=c(ID, Weight))
```

In the next example, we select all men over the age of 25 and we keep variables weight *through* income (weight, income and all columns between them).

```
# using subset function (part 2)
newdata <- subset(mydata, sex=="m" & age > 25,
select=weight:income)
```

To practice the **subset()** function, try this this interactive exercise. on subsetting data.tables.

Random Samples

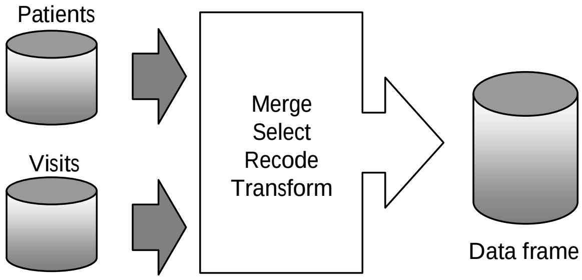
Use the **sample( )** function to take a **random sample of size n** from a dataset.

```
# take a random sample of size 50 from a dataset mydata
# sample without replacement
mysample <- mydata[sample(1:nrow(mydata), 50,
replace=FALSE),]
```

5.7 Managing and Manipulating Data in R

Q20. How data can be managed in R?

Ans :



Once you have access to your data, you will want to massage it into useful form. This includes creating new variables (including recoding and renaming existing variables), sorting and merging datasets, aggregating data, reshaping data, and subsetting datasets (including selecting observations that meet criteria, randomly sampling observeration, and dropping or keeping variables).

Each of these activities usually involve the use of R's built-in operators (arithmetic and logical) and functions (numeric, character, and statistical). Additionally, you may need to use control structures (if-then, for, while, switch) in your programs and/or create your own functions. Finally you may need to convert variables or datasets from one type to another (e.g. numeric to character or matrix to data frame).

This section describes each task from an R perspective.

Q21. How to manipulate data in R?

Ans :

Data Manipulation In R

Data structures provide the way to represent data in data analytics. We can manipulate data in R for analysis and visualization.

Before we start playing with data in R, let us see how to import data in R and ways to export data from R to different external sources like SAS, SPSS, text file or CSV file.

One of the most important aspects of computing with data Data Manipulation in R and enable its subsequent analysis and visualization. Let us see few basic data structures in R:

(a) Vectors in R

These are ordered a container of primitive elements and are used for 1-dimensional data.

Types – integer, numeric, logical, character, complex

(b) Matrices in R

These are Rectangular collections of elements and are useful when all data is of a single class that is numeric or characters.

Dimensions – two, three, etc.

(c) Lists in R

These are ordered a container for arbitrary elements and are used for higher dimension data, like customer data information of an organization. When data cannot be represented as an array or a data frame, list is the best choice. This is so because lists can contain all kinds of other objects, including other lists or data frames, and in that sense, they are very flexible.

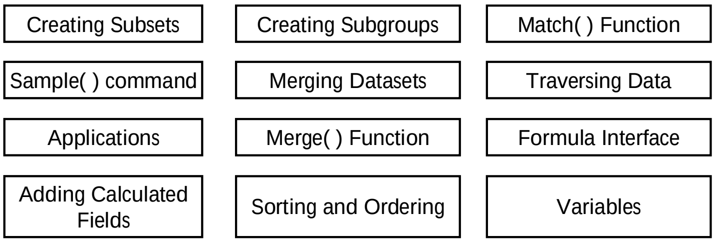
Q22. What is Data Manipulation and Data Processing?

Ans :

In this R Programming we can learn **Data manipulation in R** and data processing with R. Moreover, we will see three subset operators in R and how to perform R data manipulation like subsetting in R, sorting and merging of data in R programming language. Also, we will learn data structures in R, how to create subsets in R and usage of R sample() command, ways to create R data subgroups or bins of data in R. Along with this, we will look at different ways to combine data in R, how to merge data in R, sorting and ordering data in R, ways to traverse data in R and formula interface in R. At last, this R Data Manipulation topics will provide you complete tutorial on ways for manipulating and processing data in R.

So, let's start Data Manipulation in R.

Data Manipulation in R



Creating Subsets

Creating Subgroups

Match( ) Function

Sample( ) command

Merging Datasets

Traversing Data

Applications

Merge( ) Function

Formula Interface

Adding Calculated Fields

Sorting and Ordering

Variables

Data Manipulation In R Tutorial | R Processing Data

Short Question and Answers

1. What is R? Explain the features of R?

Ans :

R is a programming language developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithm, linear regression, time series, statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational task, C, C++ and Fortran codes are preferred.

R is not only entrusted by academic, but many large companies also use R programming language, including Uber, Google, Airbnb, Facebook and so on.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modeling and communicate the results

- **Program**: R is a clear and accessible programming tool
- **Transform**: R is made up of a collection of libraries designed specifically for data science
- **Discover**: Investigate the data, refine your hypothesis and analyze them
- **Model**: R provides a wide array of tools to capture the right model for your data
- **Communicate**: Integrate codes, graphs, and outputs to a report with R Markdown or build Shiny apps to share with the world

Features of R

As R is a leading programming language. There are so many features of **R programming** which makes it important to learn. Let's discuss them one by one.

Statistical Features of R

i) R has some topical relevance

- It is free, open source software.
- R is available under free software Foundation.

ii) R has some statistical features

- **Basic Statistics :** Mean, variance, median.
- **Static graphics :** Basic plots, graphic maps.
- **Probability distributions :** Beta, Binomial.

Any Doubt yet in Why Learn R programming? Please Comment.

Programming Features of R

i) R has some topical relevance

- Data inputs such as data type, **importing data**, keyboard typing.
- Data Management such as data variables, operators.

ii) R has some programming features

- **Distributed Computing** – Distributed computing is an open source, high-performance platform for the R language. It splits tasks between multiple processing nodes to reduce execution time and analyze large datasets.
- **R packages** – **R packages** are a collection of **R functions**, compiled code and sample data. By default, **R installs** a set of packages during installation.

2. R environment

Ans :

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- An effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either on-screen or on hardcopy, and
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

3. Reading Data in R

Ans :

For reading, (importing) data into R following are some functions.

- read.table(), and read.csv(), for reading tabular data
- readLines() for reading lines of a text file

- source() for reading in R code files (inverse of dump)
- dget() for reading in R code files (inverse of dput)
- load() for reading in saved workspaces.

4. Writing Data in R

Ans :

Following are few functions for writing (exporting) data to files.

- write.table(), and write.csv() exports data to wider range of file format including csv and tab-delimited.
- writeLines() write text lines to a text-mode connection.
- dump() takes a vector of names of R objects and produces text representations of the objects on a file (or connection). A dump file can usually be sourced into another R session.
- dput() writes an ASCII text representation of an R object to a file (or connection) or uses one to recreate the object.
- save() writes an external representation of R objects to the specified file.

5. R 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. 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.

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
}
```

6. Components of R

Ans :

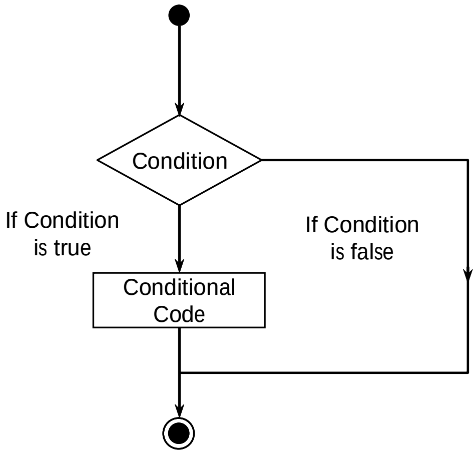
The different parts of a function are:

- **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.

7. Control statements.

Ans :

Looping is similiar to creating functions in that they are merely a means to automate a certain multi step process by organizing sequences of R expressions. R consists of several loop control statements which allow you to perform repetititve code processes with different intentions and allow these automated expressions to naturally respond to features of your data. Consequently, learning these loop control statements will go a long ways in reducing code redundancy and becoming a more efficient data wrangler.



- **if** statement for conditional programming
- **if...else** statement for conditional programming
- for loop to iterate over a fixed number of iterations
- while loop to iterate until a logical statement returns FALSE
- repeat loop to execute until told to break
- break/next arguments to exit and skip interations in a loop.

8. Data frame in R

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.

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.

Q9. What is Data Manipulation and Data Processing?

Ans :

In this R Programming we can learn **Data manipulation in R** and data processing with R. Moreover, we will see three subset operators in R and how to perform R data manipulation like subsetting in R, sorting and merging of data in R programming language. Also, we will learn data structures in R, how to create subsets in R and usage of R sample() command, ways to create R data subgroups or bins of data in R. Along with this, we will look at different ways to combine data in R, how to merge data in R, sorting and ordering data in R, ways to traverse data in R and formula interface in R. At last, this R Data Manipulation topics will provide you complete tutorial on ways for manipulating and processing data in R.

Choose the Correct Answer

1. In 2004, ________ purchased the S language from Lucent for $2 million. [a]

   (a) Insightful (b) Amazon

   (c) IBM (d) All of the mentioned

2. Which will be the output of following code? [c]

   x - 3

   Switch (6, 2+2, mean (1:10), morm(5))

   (a) 10 (b) 1

   (c) NULL (d) All of the mentioned

3. ________ programming language is a dialect of S. [c]

   (a) B (b) C

   (c) R (d) K

4. Which of the following is primary tool for debugging? [a]

   (a) debug 0 (b) trace 0

   (c) browser 0 (d) All of the mentioned

5. Point out the wrong statement: [d]

   (a) R is a language for data analysis and graphics

   (b) K is language for statistical modelling and graphics

   (c) One key limitation of the S language was that it was only available in a commercial package, S-PLUS

   (d) None of the mentioned

6. ________ is used to skip an iteration of a loop. [a]

   (a) next (b) skip

   (c) group (d) All of the mentioned

7. Which of the following may be used for linear regression? [c]

   (a) X %*% Y (b) solve (A)

   (c) solve (A,B) (d) All of the mentioned

8. Point out the WRONG statement: [a]

   (a) Early versions of the S language contain functions for statistical modeling.

   (b) The book "Programming with Data" by John Chambers documents S version of the language.

   (c) In 1993, Bell Labs gave StatSci (later Insightful Corp.) an exclusive license to develop and sell the S language.

   (d) All of the mentioned

9. Functions are defined using the ________ directive and are stored as R objects [a]

   (a) function 0 (b) funct 0

   (c) Functions 0 (d) All of the mentioned

10. In 1991, R was created by Ross Ihaka and Robert Gentleman in the Department of Statistics at the University of ________. [d]

    (a) John Hopkins (b) California

    (c) Harvard (d) Auckland

11. Which of the following adds marginal sums to an existing table ? [b]

    (a) par() (b) prop.table()

    (c) addmargins() (d) All of the mentioned

12. Which of the following lists names of variables in a data.frame ? [a]

    (a) quantile() (b) names()

    (c) barchart() (d) All of the mentioned

Fill in the blanks

1. ________ is a programming language developed by Ross Ihaka and Robert Gentleman in 1993.
2. The term ________ is intended to characterize it as a fully planned and coherent system.
3. R packages are a collection of R functions, complied code and sample data. They are stored under a directory called ________.
4. There are ________ ways to add new R packages.
5. A ________ is a set of statements organized together to perform a specific task.
6. ________ is similar to creating functions in that they are merely a means to automate a certain multi step process by organizing sequences of R expressions.
7. ________ begin by testing a condition.
8. A ________ loop is used to iterate over a block of code multiple number of times.
9. ________ structures provide the way to represent data in data analytics.
10. ________ language is made up of a collection of libraries designed specifically for data science.

Answers

1. R
2. Environment
3. Library
4. Two
5. Function
6. Looping
7. While loops
8. Repeat
9. Data
10. R

FACULTY OF MANAGEMENT

BBA III Year - VI Semester

Model Paper - I

BUSINESS ANALYTICS

Time : 3 Hours] [Max. Marks : 80

PART - A (5 × 4 = 20 Marks)

[Short Answer type]

ANSWERS

1.	a)	What is Business analytics?	**(Unit-I, SQA 1)**
	b)	What is Statistics	**(Unit-II, SQA 1)**
	b)	What is data dashboard.	**(Unit-II, SQA 6)**
	d)	Regression analysis.	**(Unit-III, SQA 2)**
	e)	Benefits of Data Mining	**(Unit-III, SQA 7)**
	f)	What is linear programming problem.	**(Unit-IV, SQA 1)**
	g)	R environment	**(Unit-V, SQA 2)**
	h)	Writing Data in R	**(Unit-V, SQA 4)**

PART - B (5 × 12 = 60 Marks)

[Essay Answer type]

Answer all the questions using the internal choice

2. a) Explain the different business analytical methods. **(Unit-I, Q.3)**

Or

b) What is Data and Explain the various types of data ? **(Unit-I, Q.16)**

3. a) What is Data visualization? Explain the importance of Data Visualization ? **(Unit-II, Q.5)**

Or

b) What is Data dashboard and explain different types dashboards. **(Unit-II, Q.15)**

4. a) Explain the concept of regression analysis. **(Unit-III, Q.4)**

Or

b) Explain the techniques used in Data Reduction. **(Unit-III, Q.22)**

5. a) What is linear programming problem (LPP)? States the mathematical formulation of LPP. **(Unit-IV, Q.1)**

Or

b) Explain about decision making under uncertainty. **(Unit-IV, Q.13)**

6. a) What is R? Explain the features of R? **(Unit-V, Q.1)**

Or

b) How to manipulate data in R? **(Unit-V, Q.21)**

FACULTY OF MANAGEMENT

BBA III Year - VI Semester

Model Paper - II

BUSINESS ANALYTICS

Time : 3 Hours] [Max. Marks : 80

PART - A (5 × 4 = 20 Marks)

[Short Answer type]

ANSWERS

1. a) What is Big Data? **(Unit-I, SQA 3)**

   b) Descriptive Statistics **(Unit-II, SQA 2)**

   b) What is Data visualization? **(Unit-II, SQA 3)**

   d) Data Mining **(Unit-III, SQA 5)**

   e) Association in Data Mining **(Unit-III, SQA 9)**

   f) Cutting Plane Method **(Unit-IV, SQA 5)**

   g) R Function **(Unit-V, SQA 5)**

   h) Reading Data in R **(Unit-V, SQA 3)**

PART - B (5 × 12 = 60 Marks)

[Essay Answer type]

Answer all the questions using the internal choice

2. a) What is Business analytics ? **(Unit-I, Q.1)**

   Or

   b) Explain the importance of big data. **(Unit-I, Q.13)**

3. a) Explain briefly about "Cross tabulations charts" by using Ms. Excel? **(Unit-II, Q.10)**

   Or

   b) Explain briefly about pivot charts. **(Unit-II, Q.14)**

4. a) Explain briefly about Forecasting techniques. **(Unit-III, Q.12)**

   Or

   b) Explain the steps involved in data mining. **(Unit-III, Q.14)**

5. a) State the advantages and limitations of linear programming problem. **(Unit-IV, Q.2)**

Or

b) Explain Non Linear Programming? **(Unit-IV, Q.7)**

6. a) Explain the various types of operators in R program. **(Unit-V, Q.4)**

Or

b) Explain different types of functions? **(Unit-V, Q.10)**

FACULTY OF MANAGEMENT

BBA III Year - VI Semester

Model Paper - III

BUSINESS ANALYTICS

Time : 3 Hours] [Max. Marks : 80

PART - A (5 × 4 = 20 Marks)

[Short Answer type]

ANSWERS

1. a) Various Challenges in Business Analytics **(Unit-I, SQA 5)**

b) Cross Tabulations **(Unit-II, SQA 4)**

b) Gantt Chart **(Unit-II, SQA 9)**

d) Limitations of Regression Analysis **(Unit-III, SQA 3)**

e) Data Reduction **(Unit-III, SQA 10)**

f) Decision analysis **(Unit-IV, SQA 6)**

g) Control statements **(Unit-V, SQA 7)**

h) Data frame in R **(Unit-V, SQA 8)**

PART - B (5 × 12 = 60 Marks)

[Essay Answer type]

Answer all the questions using the internal choice

2. a) Discuss briefly about role of business analytics in current business environment. **(Unit-I, Q.5)**

Or

b) Explain the life cycle of big data. **(Unit-I, Q.14)**

3. a) Explain the benefits of cross tabulation. **(Unit-II, Q.11)**

Or

b) What are the benefits of data dash boards. **(Unit-II, Q.17)**

4. a) Explain the various techniques used in data mining. **(Unit-III, Q.16)**

Or

b) Explain the various approaches for data mining with Micro Strategy. **(Unit-III, Q.18)**

5. a) What are the requirements of linear programming problem ? **(Unit-IV, Q.4)**

Or

b) Explain about decision making under risk. **(Unit-IV, Q.14)**

6. a) Explain briefly about control statements. **(Unit-V, Q.11)**

Or

b) Explain briefly about data frame in R? **(Unit-V, Q.18)**

FACULTY OF MANAGEMENT

B.B.A VI-Semester (CBCS) Examination

MAY - 2019

BUSINESS ANALYTICS

Time: 3 Hours Max. Marks : 80

PART – A (5 × 4 = 20 Marks)

(Short Answer Type)

Note: Answer all the questions.

ANSWERS

1. Answer any five of the following questions in not exceeding 20 lines each..

(a) Define data and data types with examples. **(Unit-I, Q.No. 16)**

(b) Explain business analytics in practice with examples. **(Unit-I, Q.No. 7)**

(c) Explain types of charts. **(Unit-II, Q.No. 8)**

(d) Prepare a Dash Board for daily sales report using MS-Excel.

Ans :

Dashboards are made up of tables, charts, gauges, and numbers. They can be used in any industry, for almost any purpose. For example, you could make a project dashboard, financial dashboard, marketing dashboard, and more.

(i) How to Bring Data into Excel

Before creating dashboards in Excel, you need to import the data into Excel. You can copy and paste the data, or if you use CommCare, you can create an Excel Connection to your export. But, the best way is to use ODBC (or Live Data Connector). ODBC can connect your apps to Excel, passing real-time data from your app to Excel. As data is updated in your app, your Excel dashboard will also be updated to reflect the latest information. This is a perfect option if you track and store data in another place, and prefer creating a dashboard in Excel. Data can be imported two different ways: in a flat file or a pivot table.

(ii) Set Up Your Excel Dashboard File

Once you have added your data, you need to structure your workbook. Open a new Excel Workbook and create two to three sheets (two to three tabs). You could have one sheet for your dashboard and one sheet for the raw data (so you can hide the raw data). This will keep your Excel workbook organized. In this example, we'll have two tabs.

(iii) Create a Table with Raw Data

(a) In the Raw Data sheet, import or copy and paste your data. Make sure the information is in a tabular format. This means that each item or data point lives in one cell.

(b) In this example, we're adding columns for Project Name, Timeline, Number of Team Members, Budget, Risks, Open Tasks, and Pending Actions.

(iv) Analyze the Data

Before building the dashboard, take some time to look at your data and figure out what you want to highlight. Do you need to display all the information? What kind of story are you trying to communicate? Do you need to add or remove any data?

Once you have an idea of your dashboard's purpose, think about the different tools you can use. Options include:

- Excel formulas like SUMIF, OFFSET, COUNT, VLOOKUP, GETPIVOTDATA and others
- Pivot tables
- Excel tables
- Data validation
- Auto-shapes
- Named ranges
- Conditional formatting

(v) Build the Dashboard

Add a Gantt Chart

We'll add a Gantt chart to visually show your project timeline.

(1) Go to your Dashboard sheet and click *Insert*.

(2) In the *Charts* section, click the bar chart icon and select the second option.

(e) Explain the role of multiple regressions in demand forecasting. **(Unit-III, Q.No. 10)**

(f) Explain the concept of risk and uncertainty. **(Unit-IV, Q.No. 13)**

Ans :

Risk is uncertainty that a future event with a favourable outcome will occur. In other words, risk is the probability that an investment will not perform as expected and the investor will lose the money invested in the project. All business decisions and opportunities are based on this concept that future performance and returns are uncertain and rely on many uncontrollable variables.

Risk is inherent in any investment. Risk may relate to loss of capital, delay in repayment of capital, non-payment of return or variability of returns. The risk of an investment is determined by the investments, maturity period, repayment capacity, nature of return commitment and so on.

Risk implies future uncertainty about deviation from expected earnings or expected outcome. Risk measures the uncertainty that an investor is willing to take to realize a gain from an investment.

(g) Explain built in functions and user defined functions in r^2. **(Unit-V, Q.No. 10)**

(h) Write the code for creating and calling a function to print squares of numbers sequence using R. **(Unit-V, Page 203, (iv) Point)**

PART – B (5 × 12 = 60 Marks)

(Essay Answer Type)

Note: Answer all the questions using the internal choice.

2. (a) Explain categories of Business Analytical Methods and models with examples. **(Unit-I, Q.No. 3, 4)**

OR

(b) Explain the role of big data in competing food apps Swiggy and Zomato.

Ans :

As the online food ordering trend is becoming more and more prominent in India, the food delivery platforms like Swiggy and Zomato are growing their user base at an exponential rate.

1. Swiggy

According to a report, the number of user interactions on Swiggy has grown exponentially from 2 billion in October 2017 to a massive 40 billion in January 2019. To keep up with the massive growth the company looks up to Artificial intelligence as a solution to many of the problems. Head of the Engineering and Data Science Team at Swiggy, Dale Vaz says "AI is critical for us to sustain our growth,".

Artificial intelligence helps Swiggy distinguish the dishes from images classifying them as vegan or non-vegan dishes. Natural Language Processing can greatly help the platform in serving wider geography without having to consider linguistic boundaries that enables search using colloquial terms which customers could use to obtain accurate results.

2. Zomato

As seen by Swiggy as its arch-rival in the food delivery market, Zomato doesn't seem to be backing down too. Recently the company had raised Rs 284 crore from a US investor Glade Brook Capital Partners as part of its strategy to acquire more market share from its rivals. Last month Zomato made a claim of achieving a 28 million monthly order run rate as of December compared to the 21 million in October which also helps the company in projecting future order volume. The platforms Gold Subscription package also claims to have worked out for the company, bringing on board 7 lakh members and over 6,000 restaurant partners, up from 6 lakh members and 4,000 restaurants.

It was in the late December that Zomato acquired Lucknow-based startup TechEagle Innovations looking forward to establishing a drone-based delivery network in India.

Zomato's Founder and CEO Deepinder Goyal said in a press release -"We believe that robots powering the last-mile delivery is an inevitable part of the future and hence is going to be a significant area of investment for us."

3. (a) Explain any six data visualization techniques and their importance in Business Analysis. **(Unit-II, Q.No. 8)**

OR

(b) Explain the process of charts in spss by clearly mentioning the path. **(Unit-II, Q.No. 16)**

4. (a) Explain cause and effect modelling with a hypothetical example. **(Unit-III, Q.No. 26)**

OR

(b) Explain the approaches in Data Mining. **(Unit-III, Q.No. 18)**

5. (a) How do you draft a proposal for benefiting from linear optimization? Give an example. **(Unit-IV, Q.No. 1)**

OR

(b) Explain the concept of Decision Analysis. **(Unit-IV, Q.No. 11)**

6. (a) Explain R environment and any five R packages and their uses. **(Unit-V, Q.No. 2, 3, 6)**

OR

(b) Using the example given explain the remaining data types in R.

Data type	Example	Verify
1. Logical	True, False	V ← True
2. Numeric	?	?
3. Integer	?	?
4. Complex	?	?
5. Character	?	?
6. Raw	?	?

Ans :

Data type	Example	Verify
1. Logical	True, False	V ← True
2. Numeric	True, False	V ← True
3. Integer	True, False	V ← True
4. Complex	True, False	V ← False
5. Character	True, False	V ← False
6. Raw	True, False	V ← False