JEPPIAAR ENGINEERING COLLEGE

DEPARTMENT OF MANAGEMENT STUDIES

BATCH 2022 – 2024

I YEAR / I SEMESTER

BA4206 : BUSINESS ANLYTICS

COURSE MATERIAL

Anna University Chennai

Regulation 2021

JEPPIAAR ENGINEERING COLLEGE

DEPARTMENT OF MANAGEMENT STUDIES

VISION

To build Jeppiaar Engineering College as an institution of academic excellence in technology and management education, leading to become a world class university..

MISSION

- To excel in teaching and learning, research and innovation by promoting the principles of scientific analysis and creative thinking.
- To participate in the production, development and dissemination of knowledge and interact with national and international communities.
- To equip students with values, ethics and life skills needed to enrich their lives and enable them to contribute for the progress of society.
- To prepare students for higher studies and lifelong learning, enrich them with the practical skills necessary to excel as future professionals and entrepreneurs for the benefit of Nation's economy.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOS):

MBA programme curriculum is designed to prepare the post graduate students

- To have a thorough understanding of the core aspects of the business.
- To provide the learners with the management tools to identify, analyze and create business opportunities as well as solve business problems.
- To prepare them to have a holistic approach towards management functions.
- To inspire and make them practice ethical standards in business.

PROGRAMME OUTCOMES (POS)

On successful completion of the programme,

1. Ability to apply the business acumen gained in practice.
2. Ability to understand and solve managerial issues.
3. Ability to communicate and negotiate effectively, to achieve organizational and individual goals.
4. Ability to understand one's own ability to set achievable targets and complete them.
5. Ability to fulfill social outreach
6. Ability to take up challenging assignments

COURSE OBJECTIVE:

Acquire a reasonable knowledge in accounts analysis and evaluate financial statements

COURSE OUTCOMES:

1. A thorough grounding of financial accounting concepts
2. Preparation of financial statement analysis
3. Understand the management and cost accounting techniques
4. Apply the management and cost accounting techniques for decision making
5. Assess the accountancy standards of practices in India

CO-PO MATRIX

CO	PO1	PO2	PO3	PO4	PO5	PO6
CO1	3	3	0	2	0	2
CO2	3	3	0	3	0	2
CO3	2	3	0	2	0	2
CO4	3	3	0	2	0	2
CO5	3	3	0	3	0	2
Average	2.8	3	0	2.4	0	2

Unit -1

INTRODUCTION TO BUSINESS ANALYTICS (BA) 9

Business Analytics - Terminologies, Process, Importance, Relationship with Organisational Decision Making, BA for Competitive Advantage.

PART- A

1. What is a data set?
   It is a simple collection of data or a data file.

2. What is a database?
   It is a collection of data files that contain information on people, locations, and so on.

3. What are computer clouds?
   It is hardware and software used for data remote storage, retrieval, and computational functions.

4. What is data warehousing
   It is a collection of databases used for reporting and data analysis store data.

5. What is Big data?
   It describes the collection of data sets that are so large and complex that software systems are hardly able to process them.

6. What is Analytics?
   It can be defined as a process that involves the use of statistical techniques (measures of central tendency, graphs, and so on), information system software (data mining, sorting routines), and operations research methodologies (linear programming) to explore, visualize, discover and communicate patterns or trends in data. Analytics convert data into useful information. Example of the use of analytics is the weather measurements collected and converted into statistics, which in turn predict weather patterns.

7. Describe the three steps of the business analytics process.
   Descriptive
   Predictive
   Prescriptive

8. What is Descriptive analytics?
   The application of simple statistical techniques that describes what is contained in a dataset or database. Example: An Age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.

9. What is predictive analytics?

An application of advanced statistical information software or operations research methods to identify predictive variables and build predictive models to identify trends and

relationships not readily observed in a descriptive analysis. Example: Multiple regression is used to show the relationship (or lack of relationship) between age, weight and exercise on diet food sales. Knowing that relationships exist helps explain why one set of independent variables influences dependent variables such as business performance.

15.What is Prescriptive analytics?

An application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make best use of allocable resources. Example: A department store has a limited advertising budget to target customers. Liner programming models can be used to optimally allocate the budget to various advertising media.

16. Define business analytics
(BA) can be defined as a process beginning with business related data collection and consisting of sequential application of descriptive, predictive, and prescriptive major analytic components, the outcome of which supports and demonstrates business decision-making and organizational performance.

17.Define Business intelligence (BI)
It can be defined as a set of processes and technologies that convert data into meaningful and useful information for business purposes. While some believe that BI is a broad subject that encompasses analytics, business analytics, and information systems. focused on collecting, storing, and exploring large database organizations for information useful to decision-making and planning. BI involves storing an organization's data in computer cloud storage or in data warehouses.

18. What is the difference between business performance role in analytics, BA & BI?

Business Analytics	**Business Intelligence**	**Analytics**
What is happening and what will be happening? What is the best strategy to deal with it?	What is happening, how and what have we done in the past to deal with it?	What is happening and what will be happening?

19. Describe four data classification measurement scales.
Categorical data
Ordinal data
Interval data
Ratio data

20. What is categorical data?
Data that is grouped by one or more characteristics. Categorical data usually involves cardinal numbers counted or expressed as percentages. Example 1: Product markets that can be characterized by categories of "high-end" products or "low-income"producs, based on dollar

sales> It is common to use this term to apply to data sets that contain items identified by categories as well as observations summarized in cross-tabulations or contingency tables.

21. What is ordinal data?
Data that is ranked or ordered to show relational preference. Ex:1: Football team rankings not based on points scored but on wins. Example 2: Ranking of business firms based on product quality.

22. What is interval data?
Data that is arranged along a scale where each value is equally distant from others. It is ordinal data. Example 1: A temperature gauge. Example 2: A survey instrument using a Likert scale (that is, 1, 2, 3, 4, 5, 6, 7), where 1 to 2 is perceived as equidistant to the interval from 2 to 3, and so on. Note: In ordinal data, the ranking of firms might vary greatly from first place to second, but in interval data, they would have to be relationally proportional.

23. What is ratio data?
Data expressed as a ratio on a continuous scale. Example 1: The ratio of firms with green manufacturing programs is twice that of firms without such a program.

24. What is the difference between Data analytics and business analytics?

	Business Analytics	**Data Analytics**
Goal	Focuses on identifying trends in the organization that can be optimized to improve overall business planning and performance. Supports continuous improvement in technology and processes. Seeks to arrive at a single version of the truth.	Benefits come from recognizing patterns in a dataset and making accurate predictions based on events.
Data	Data sources are defined in advance based on project goals.	Analysis is more ad hoc with data sources added on the fly as correlations are uncovered.
Approach	Involves defining the goals and requirements for programs and projects. More retrospective and descriptive.	Typically, more predictive and prescriptive. Strives to answer specific questions and discover new insights for competitive advantage.
Team members	CIO, CDO, analytics manager, business analyst, data warehouse engineer	Data analyst, line of business manager

25. What is the difference between business analytics and business intelligence?

BASIS FOR COMPARISON	Business Intelligence	Business Analytics
Definition	Analyses past and present to drive current business needs	Analyses past data to drive current business
Usage	To run current business operations	To change business operations and improve productivity
Ease of Operations	For current business operations	For future business operations
Tools	SAP Business Objects, QlikSense, TIBCO, PowerBI, etc.,	Word processing, Google docs, MS Visio, MS Office Tools, etc.,
Applications	Apply to all large-scale companies to run current business operations	Applies to companies where future growth and productivity as its goal
Field	Comes under Business Analytics	Contains Data warehouse, information management, etc.,

26.What are Digital analytics?
It is a term that describes any source of data that is conveyed using digital sources. Examples of these new sources of data-based analytics include text analytics and unstructured data analytics.

27. What is Text analytics?
It can be defined as a set of linguistic, statistical, and computer-based techniques that model and structure the information content from textual sources.

28. What is text data mining?

It is a search process in databases to find patterned text material that provides useful information. Also referred to as text data mining, text analytics uses data mining software to look into databases to find and validate the kinds of information on which predictions can be made.

29. What is social media data?
One example of technology driven data is social media data. Social media can be defined as interactions or communications among people or communities, usually performed on a technology platform, involving the sharing, creating, discussing, and modifying of communicated verbal or electronic content.

30. List the two global social platforms.
Twitter and Facebook.

31. The methodologies or technologies used in the purveyance of social media data can include any means of distribution of verbal or other types of communications, including, but not limited to, photographs or pictures, video, Internet forums, web logs, discussion forums, social blogs, wikis, social networks, and podcasts. These sources of data are the basis of social media analytics, on which the analytics information can aid in learning new types of social media behavior and information.

32.What is Mobile analytics?
They can be defined as any data secured from mobile devices, such as smartphones, iPhones, iPads, and Web browsers. These are all mobile technologies used to obtain digital data from the interaction of people.

33.What is structured data?
When data is placed in databases and can be logically filed, accessed, referenced, and used, it is known as structured data.

34.What is unstructured data?
When data or information, either digital or nondigital, cannot be put into a database or has no predefined structure, it is known as unstructured data. Examples of unstructured data include images, text, and other data that, for one reason or another, cannot be placed in a logically searchable database based on content. Much of the data contained in emails and on the Web is unstructured.

35. Explain the application of mobile analytics?
The fact that they are mobile and move from location to location with the user differentiates the type of information available to the analytics analysts. For example, the mobile technology allows analysts to not only track what a potential customer might talk about on the use of a product (such as in social media analytics), but track movements of where the customer makes decisions on products. That can help explain why those decisions are made. For example, mobile technology might reveal the location of a purchaser of hair spray to have been physically located

near an area where billboards are used for hair spray advertising, thus helping to reveal the possible connection and effectiveness of a billboard promotion.

36. What is strategic planning?
Strategic planning is the art of creating specific business strategies, implementing them, and evaluating the results of executing the plan, in regard to a company's overall long-term goals or desires.

37.What is operational planning?
It is usually performed by lower-level managers to implement short range goals

38. What is tactical planning?
It is usually performed by middle-level managers to develop implementation tactics for strategy planning. These are usually middle –range goals such as acquiring additional firms to support the growth strategy.

PART- B

1.Why does each step in the business analytics process have a past, present, and future dimension?

	Time Period		
Step in BA	**Past**	**Present**	**Future**
1. Descriptive	What happened in the past?	What is happening now based on the past?	What will appear to happen based on the past?
2. Predictive	How did it happen in the past? Why did it happen in the past?	What possible trends exist in the data that can predict or forecast what course of action should be taken now?	What is the range and likelihood of possible outcomes that can happen if the current trends or forecasts are allowed?
3. Prescriptive	How best can we leverage what we know from the trends and forecasts?	How can we optimally apply resources to maximize the business performance outcomes in the future?	How can we continuously apply BA in the future to optimize upcoming business performance outcomes?

In this illustrative case scenario a local credit union offers a series of packaged homeowner loans that are periodically marketed by running a promotional campaign in a variety of media (print ads, television commercials, radio spots). The idea is to bring in new customers to make home loans that fit one of the packaged deals. Halfway through the marketing program, the credit union does not know if the business generated is due to the promotional campaign or just a result of their normal business cycle. To clarify, the credit union undertakes a BA analysis.

Ex: Credit Union example for BA Analysis

	Time Period		
Step in BA	**Past**	**Present**	**Future**
1. Descriptive	Based on graphics results, past ad campaigns resulted in a moderate increase in new loans.	Based on sorting of loan activities, new homeowner loans are experiencing just a moderate increase in new loan applications.	Based on histogram of loans to date, there is no discernible pattern, just uniform new loan sales that are constant over time. No business cycle impact is observed to alter loan patterns.
2. Predictive	Statistically, correlations have revealed in the past that marketing promotions will increase new loans, but why they generate new loans depends on how the promotion campaign invests its funding.	Utilizing multiple regression, the model predicts that a greater allocation in funds for television commercials and print ads will be more effective in generating new loans than investing in radio spots.	Utilizing variance statistics from the regression model, a confidence interval can estimate the number of new loans possible if a reallocation of promotion funds is implemented.
3. Prescriptive	Reallocating marketing budget funds from radio spots to television commercials and print ads is required to more effectively reach the target audience.	Given the constrained resource of funding, a linear programming model is used to optimally allocate the marketing budget in dollars to maximize the promotional outcome for new loans.	Continuous tracking over a period of time of the new loan applications caused by the promotion campaign needs to be monitored and mapped against the predicted outcomes suggested in the analysis.

The answers to the questions raised in the credit union example are typical of any business organization problem□solving or opportunity-seeking quest. The answers were not obtained by just using statistics, computer search routines, or operations research methodologies, but rather were a result of a sequential BA process. The informational value of the answers in this scenario suggests a measurable and precise course of action for the management of the credit union to follow. By continuously applying BA as a decision support system, firms have come to see not only why they need BA, but also how BA can become a strategy to achieve competitive advantage. Kiron et al. (2012) reported in a survey on business through the year 2012 that firms

applying business analytics permit the organization to have better access to data for decision-making and offer a competitive advantage.

2. What is a competitive advantage, and how is it related to BA?

Business analytics is the process of gathering data, measuring business performance, and producing valuable conclusions that can help companies make informed decisions on the future of the business, through the use of various statistical methods and techniques.

Analytics has become one of the most important tools at an organization's disposal. When data and analytics work hand in hand, the benefits become obvious. Companies can leverage data to improve cost savings, redefine processes, drive market strategy, establish competitive differentiators and, perhaps most importantly, build exceptional and truly personalized customer experience.

Business analytics for organisations is becoming a competitive advantage and is now necessary to apply business analytics, particularly its subset of predictive business analytics. The use of business analytics is a skill that is gaining mainstream value due to the increasingly thinner margin for decision error. It is there to provide insights, predict the future of the business and inferences from the treasure chest of raw transactional data, that is internal and external data that many organizations now store (and will continue to store) as soft copy.

Business analytics enables differentiation. It is primarily about driving change. Business analytics drives competitive advantage by generating economies of scale, economies of scope, and quality improvement. Taking advantage of the economies of scale is the first way organizations achieve comparative cost efficiencies and drive competitive advantage against their peers. Taking advantage of the economies of scope is the second-way organizations achieve comparative cost efficiencies and drive competitive advantage against their peers.

Business analytics improves the efficiency of business operations. The efficiencies that accumulate when a firm embraces big data technology eventually contributes to a ripple effect of

increased production and reduced business costs. In the modern world, the vast quantities of data produced by corporations make their study and management practically impossible.

One can make the case that increasing the primary source of attaining a competitive advantage will be an organization's competence in mastering all flavours of analytics. If your management team is analytics-impaired, then your organization is at risk. Predictive business analytics is arguably the next wave for organizations to successfully compete. This will result not only from being able to predict outcomes but also to reach higher to optimize the use of their resources, assets and trading partners. It may be that the ultimate sustainable business strategy is to foster analytical competency and eventually mastery of analytics among an organization's workforce.

Analytics gives companies an insight into their customers' behaviour and needs. It also makes it possible for a company to understand the public opinion of its brand, to follow the results of various marketing campaigns, and strategize how to create a better marketing strategy to nurture long and fruitful relationships with its customers.

Business analytics helps organisations to know where they stand in the industry or a particular niche provides the company with the needed clarity to develop effective strategies to position itself better in the future.

For a company to remain competitive in the modern marketplace that requires constant change and growth, it must stay informed on the latest industry trends and best practices. Not only does business analytics provide the needed knowledge for companies to survive in today's constantly changing business environment, but it also makes room for growth and improvement, providing a detailed look into various opportunities and challenges that companies face on a day-to-day basis.

4. How does BA help achieve sustainability?

Using sustainability analytics, companies can reduce how they use resources, increasing their resilience to price fluctuations and supply constraints. As well as hedge their resources, this also helps to anticipate changes in demand, and supply, so that when prices fall, their supplies can be priced appropriately.

We want to empower our Board members and business leaders to focus on sustainability and implement plans that drive efficiency. Developing sustainable products or services by strategically planning and planning. Consumer choice can be triggered by sustainable marketing of products and services.

Water, sanitation, and sustainable energy should be universally accessible. Inclusive education and decent work can generate development opportunities. Creating cities and communities that can be sustainably grown and consumed requires innovation and resilient infrastructure.

While data analysis can often lead to increased efficiency, it can also lead to uncovering business opportunities that might otherwise go unnoticed, like identifying untapped segments of customers. Thus, greater intelligence becomes part of the process, resulting in endless growth and profitability possibilities.

Listed companies rated sustainability by Sustainalytics for the sake of their environmental, social, and corporate governance functions. Sustainalytics acquired the remaining 56% of shares in order to become the sole owner for the company.

Increase the earnings Earning more money with the use of sustainability will enhance your bottom line as well. As a result, sustainability businesses increase in earning money by focusing on reducing costs, improving their reputation, and attracting more customers with their commitment to sustainability. Being sustainable does not come exclusively from PR; we should achieve actual outcomes when the company has an authentic vision and drive towards its own growth trajectory. By using business analytics, companies can make better, more informed decisions and reach a variety of goals at the same time. By leveraging data, businesses are able to: Better understand consumers. Create a competitive edge by learning about their rivals. An

advantage of Sustainalytics' Impact Metrics is it helps pension funds and asset managers analyze and report on investments' impacts by utilizing internationally recognized sustainable development goals (SDGs) and other impact themes.

5. Explain different types of competitive advantages and their relationship to business analytics.

Type of Competitive Advantage	Description	Ways BA Can Help Achieve the Competitive Advantage
Price Leadership	From a marketing standpoint, offer products or services at the lowest cost to customers in the industry, while making acceptable profit for the company.	Identify main competitors; monitors, reports, and accurately forecasts competitive prices so firm can keep lowest cost profile while maintaining and measuring profit margins.
Sustainability	To ensure the firm's resource usage in a way that seeks balance to hurt neither the environment nor the bottom line of a firm's profitability.	Identify areas where resource reallocations are needed to avoid damaging the environment, suggest ways to reallocate the resources, and help allocate them optimally to achieve the best possible balance.
Operations Efficiency	Improve the internal business operations and activities over competitors, lessening the cost to the customer. That reduced cost, if passed on to customers, can provide a lower price advantage based on efficiency.	Identify operation areas needing correction or modification and suggest alternatives to improve efficiency. Also, this can be useful in selecting which alternative to use to maximize business performance.
Service Effectiveness	Make customer transactions easier or more pleasurable than with other firms. This improves the service characteristics of the firm while lowering the time it takes to get services to the customer, thus enhancing customer value.	Obtain customer opinions on problem service areas needing fixing; explain why the fix is needed, suggest alternatives to the fix, improve the effectiveness of the service operations, and measure and report improvements.
Innovation	Introduce completely new or notably better products or services with the intention of disrupting competitors' businesses by obsoleting the current market entries with break-through product offerings.	Obtain and validate customer ideas and suggestions on new products or enhancements in current products. Monitor customer reactions and suggest refinements as new products are introduced to customers. Monitor performance on new products and report results.
Product Differentiation	Provide customers with a variety of products, services, or features that competitors are not yet offering or are unable to offer.	Identify new products not offered by competitors, suggest new services to offer, forecast potential of new products for profitability measurement.

Table 2.3 Ways BA Can Help Achieve a Competitive Advantage

6. Explain how business analytics can help an organization achieve a competitive advantage.

As many in business know, competitive advantage refers to what makes a company's goods or services superior to others. Why should customers buy from you instead of from someone else? Companies have tried to build on this advantage in many ways like reducing costs, developing new products and improving the overall customer experience. However, what many overlook (surprisingly) to this day is the staggering value and benefits created by the direct relationship between data analytics and competitive advantage. This linkage has the potential to improve and build on existing advantages, thereby strengthening a company's market position (or sustaining it, if it's already at the top). Following are the benefits of business analytics which can help a organisation to achieve competitive advantage.

a. Discover new ways to cut costs

Organisations use cost-cutting to improve their competitive advantage by lowering production costs and passing on the benefits to consumers. By merging data analytics and competitive advantage-focused strategies, businesses can leverage their monumental data assets given that they provide organisations with new ways of cost-cutting. Analytics can reveal inefficiencies and flaws that would be impossible to capture through conventional means. Thus, organisations can develop new cost-cutting measures focusing on removing inefficiencies and cutting down waste in resources. An excellent example is Airbus. The aviation corporation used data analytics to improve manufacturing, development and maintenance processes. The change in processes reduced the number of operational interruptions disrupting flight schedules and made better use of resources when designing a new aircraft. Data analytics also provided real-time data which was also expected to cut fuel usage by 15%.

b. Better deployment of resources

To maintain a competitive advantage, organisations have to be smarter and more creative in resource management. Data analytics can help in this area by helping organisations deploy resources intelligently. Big data provides insights into the granular operations of a company that are easily missed with other means. When these insights are placed in the hands of business-minded individuals, they can identify ways to make better use of their resources without resorting to cost-cutting measures. To achieve this goal, the zoo used analytics to study historical attendance data, ticket sales and real-time weather. The insights from the study allowed them to

anticipate attendance and deploy resources accordingly. Thus, more staff were called in during peak attendance and less when attendance was down.

c. Personalised customer service

The internet has changed the customer-retailer relationship. Thanks to Amazon and Netflix, customers have grown accustomed to personalised service and now expect organisations to provide a similar level of customer service. Those who don't will lose competitive advantage.

Fortunately, data analytics allows organisations to provide personalised customer service. Analytics platforms can capture and breakdown customer purchase history into a detailed individual profile. Analytics goes beyond the basic demographic information – it can even go as far as to depict what the customer will buy, and under what circumstances. Thus, paving the way for personalised customer campaigns in brand awareness and engagement.

Amazon and Netflix are the best examples of organisations using data and analytics to provide personalised customer service to their customers. Using past purchases as a barometer, Amazon provides a series of recommendations that are catered specifically to that customer. Something that can only be achieved with data analytics because they have a shopper base consisting of millions. The result? A service that is almost unparalleled and incredibly difficult for competitors to compete with.

Key takeaways

Competitive advantage is vital for any organisation to survive. To compete, organisations have to discover new ways to cut costs, deploy resources more effectively, and develop ways to reach out to each of their customers individually. All of this can be achieved with data analytics. Data analytics provides brand new insights not found through other tools. These are insights that can pave the way for other organisations to make better use of their resources, cut costs and personalise their services to strengthen their advantage.

7. Explain the importance of business analytics for a business organization. [or] Explain why business analytics are important in identifying new business initiatives.

1. Enhance Customer Experience

With the variety of options available, customers are spoiled for choice. To ensure businesses can retain their customer base, they turn to analytics. For example, companies can analyse a customer's interaction on their website and past purchasing habits. Based on

this data, they can analyse patterns and make improvements to their website performance. This can be as simple as sending a push notification prompting products that customers haveadded to their shopping cart. This will result in overall better customer experience, and eventually, loyalty.

2. Make Informed Decisions

Businesses often outsource a few of their processes to enhance their efficiency. When it comes to selecting vendors for such activities, they need to know which one will bring more profits. Analytics can help them evaluate supplier performance based on customer ratings, order fulfillment speed, quality, etc. This data will help them decide which one works best for their business.

3. Reduce Employee Turnover

Every year, companies have to bear an extensive cost due to employee on boarding and attrition. To save time and costs, HR professionals can use analytics tools to examine the likelihood of an employee aligning with a companies culture, tracking their performance and how satisfied they are with the new role, and similar factors. Once you know such answers, it will be easier to recognize employees who will stick to your company for a long time.

4. Improve Efficiency

Efficiency is not always limited to employees. Businesses can also analyse other resources to learn more about their performance. For example, a grocery store chain was able to reduce refrigeration costs by merely analysing the temperatures of in-store coolers. It was found that the refrigerators were being kept several degrees lower than necessary, which increased power usage. So, by increasing the temperature, power costs went down without affecting safe food storage. Business owners can learn from such examples and use data to make their resources efficient.

5. Identify Frauds

Finance companies have begun using analytics to reduce fraud. One way they do this is by using data to identify potentially fraudulent purchases, based on the analysis of customers previous transactions. These companies also use predictive analytics to look at customer profiles and gauge the level of risk. This helps rate the risk that a particular customer presents and use this analysis to prevent losses, and builds stronger customer relationships.

6.Cut Manufacturing Costs

One company that has outranked everyone when it comes to using analytics to reduce manufacturing costs is Intel. Initially, this tech giant would perform 19,000 tests on each chip being manufactured. With the advent ofpredictive analysis, Intel was able to determine which chips need, which tests before their launch. By using thedata collected from all of that testing, it has been able to save almost $3 million.

7.Make The Most Of Your Investment

Earlier, business people would rely on traditional marketing techniques to gain and retain customers. However, the internet has changed this completely. Now search marketing has become the most efficient and inexpensive way for businesses to find leads. By using business

intelligence, marketers can craft perfect campaigns and strategies, eventually increasing chances of higher ROI.

8.Improved Advertising

Advertising is expensive; hence marketers must know how to get the best return on investment. This is why theyuse analytical methods such as A/B and C split-testing. When it comes to advertising online, all landing pages, pop-ups, and even product descriptions are evaluated and tweaked to ensure maximum results. Even the wayproducts are positioned on the website is assessed to identify the best location to drive more engagement and sales.

9.Better Product Management

When it comes to retail companies, they have more than a thousand products to offer. So, how do they decide which ones to release at which time? Yes, the answer is data analytics. Such businesses analyse which are the most popular products depending on the region and season. This data is then used to target the right product at the right time, which eventually has a positive impact on sales.

10. Tackle Problems

Whenever a problem arises, a business often pauses the current operations, which leads to a huge loss. To prevent such situations, business analysts help the organization to make an informed decision by providing informationthat can help identify potential risks and avoid any occurrence of loss. These professionals can use the raw data to detect a malfunction in the existing system, and thus help business owners to fix it at the earliest.

11.Accelerate Through Uncertainty

Look at the current business scenario. Owing to the lockdown across the globe, the business environment is as uncertain as it gets. Almost nobody has an idea when things will get back to normal, and the corporate world will be allowed to resume its operations. At such times, data analytics can be used to resolve supply chain issues, introduce crisis management solutions, optimize costs, and more.

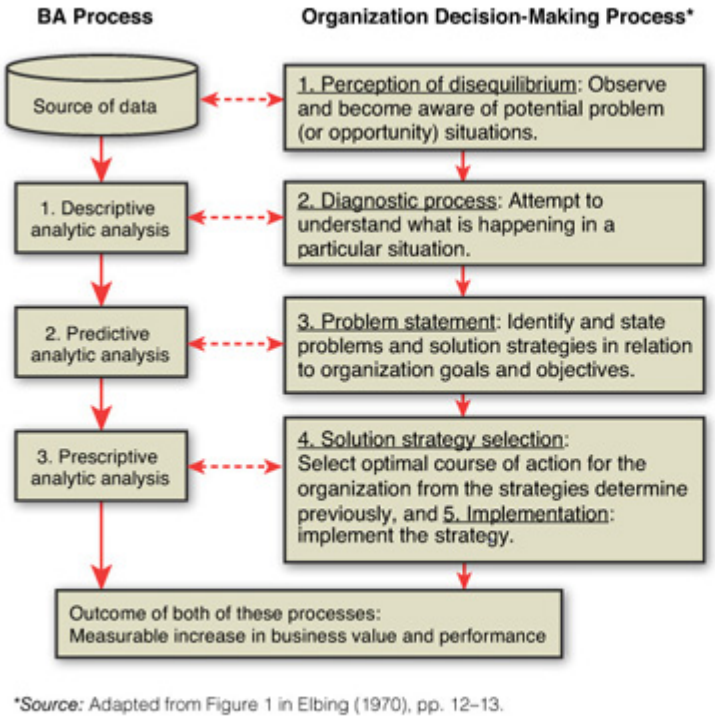
12.Conduct a Competitor Analysis

Today, almost every business has a clear idea of its competitors. An effective way to get ahead of them is by understanding what they are up to, their strategies, USPs, etc. By gathering this data by conducting a SWOT analysis, you can get a preview of how your business is performing as compared to your competitors.

8. Explain the relationship of the business analytics process with the organization decision-making process.

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities, can be

viewed as a problem of strategy choice requiring a solution. It should come as no surprise that the BA process described in Section 1.2 closely parallels classic organization decision-making processes. As depicted in Figure 1.2, the business analytic process has an inherent relationship to the steps in typical organization decision-making processes.



*Source: Adapted from Figure 1 in Elbing (1970), pp. 12–13.

Figure 1.2 Comparison of business analytics and organization decision-making processes

The *organization decision-making process* (ODMP) developed by Elbing (1970) and presented in Figure 1.2 is focused on decision making to solve problems but could also be applied to finding opportunities in data and deciding what is the best course of action to take advantage of them. The five-step ODMP begins with the perception of disequilibrium, or the awareness that a problem exists that needs a decision. Similarly, in the BA process, the first step is to recognize that databases may contain information that could both solve problems and find opportunities to improve business performance. Then in Step 2 of the ODMP, an exploration of the problem to determine its size, impact, and other factors is undertaken to diagnose what the problem is. Likewise, the BA descriptive analytic analysis explores factors that might prove useful in solving problems and offering opportunities. The ODMP problem statement step is similarly structured

to the BA predictive analysis to find strategies, paths, or trends that clearly define a problem or opportunity for an organization to solve problems. Finally, the ODMP's last steps of strategy selection and implementation involve the same kinds of tasks that the BA process requires in the final prescriptive step (make an optimal selection of resource allocations that can be implemented for the betterment of the organization).

The decision-making foundation that has served ODMP for many decades parallels the BA process. The same logic serves both processes and supports organization decision-making skills and capacities.

Unit -2

UNIT II MANAGING RESOURCES FOR BUSINESS ANALYTICS 9

Managing BA Personnel, Data and Technology. Organisational Structures aligning BA. Managing Information policy, data quality and change in BA.

1. Who are administrators?
   They are the people manage servers, their load balancing, installation, and configurations. They manage reports from computer portals, manage dispatchers and perform trouble shooting for technology. They are also incharge of user authorization and authentication for security.
2. Who are designers?

They are responsible for building reports using relational data models, as well as enhancing, customizing and managing professional reports.

3. Who are developers?
   They are responsible for application for analytics, data warehousing,model building, use of operations research and statistical methodologies and real time monitoring of data flows to users.
4. Who are solution experts?
   They are experts analyze,plan, design, deploy,and operate BA applications using an appropriate methodology and development approach. This requires knowledge in many differing BA Software applications,including statistical, information system and operation research methods.
5. Who are technical specialists?
   They are responsible for installation and configuration of BA and BI applications.

6. List the competency required by BA Personnel?
   Business, Analytic, and information system.

7. List the roles performed in business competency by BA Personnel?
   Leadership, people related management and communication skills
   Managing BA projects (prioritizing, scheduling)

Managing BA Processes (rules, procedure)
Determine project requirement
Training

8. List the types of Internal Data ?
   Billing and Remainder systems
   Business
   Customer
   CRM
   HR
   Information from ERP
   Product
   Production
   Questionnaires
   Web Logs

9. List the types of External Data?
   Customer satisfaction – revenue, profit
   Customer demographics- income level, market size
   Competition – market share
   Economic – population statistics

10. What is data quality?
Data quality can be defined as data that serves the purpose for which it is collected.

11.What is Data privacy?
It refers to the protection of shared data such that access is permitted only to those users for whom it is intended.

12. List the General information technology infrastructure.
Computer hardware
Computer software
Networking and telecommunication technology
Data management technology

13. What is data management technology?
It includes database management systems, data warehouses, data marts and online analytical processing as well as data, text and web mining technologies.
14. What is DBMS?
Database management systems (DBMS) is a data management technology software that permits firms to centralize data, manage it efficiently, and provide access to stored data by application programs. DBMS usually serves as an interface between application programs and the physical data files of structured data.

15.What is database content management?

Database Content Item	Description
Purpose	Why the database exists, including any additional reports or analyses used in leveraging the data.
Time	Window of time period when the data is collected or will be useful.
Source	Internal (auditing, accounting, and so on) and external (customers, and so on) sources.
Schematics	Diagrams illustrating the connections between tables and other data files.
Cost	Expense of collecting data, including purchasing prices.
Availability of Data	Window of time when the data may be available.
Collection Techniques	Methods of collection, including observation, data mining, census, and focus groups.
Collection Tools	Web, customer generated, e-survey, and so on.

16. What is a data warehouse?

Data warehouses are databases that store current and historical data of potential interest to decision makers. What a data warehouse does is make data available to anyone who needs access to it. In a data warehouse, the data is prohibited from being altered. Data warehouses also provide a set of query tools, analytical tools, and graphical reporting facilities.

17. What is data mart?

Data marts are focused subsets or smaller groupings within a data warehouse. A data mart can be constructed more quickly and at lower cost than enterprise-wide data warehouses to concentrate effort in areas of greatest concern.

18. Online analytical processing (OLAP) is software that allows users to view data in multiple dimensions. For example, employees can be viewed in terms of their age, sex, geographic location, and so on. OLAP would allow identification of the number of employees who are age 35, male, and in the western region of a country. OLAP allows users to obtain online answers to ad hoc questions quickly, even when the data is stored in very large databases.

19. Data mining is the application of a software, discovery driven process that provides insights into business data by finding hidden patterns and relationships in big data or large databases and inferring rules from them to predict future behavior.

20. List the types of Information obtainable with data mining technology.

Types of Information	Description	Example
Associations	Occurrences linked to a single event.	An ad in a newspaper is associated with greater sales.
Classification	Recognizes patterns that describe the group an item belongs to by examining previous classified existing items and by inferring a set of rules that guide the classification process.	Identify customers who are likely to need more customer service than those who need less.
Clustering	Similar to classification when no groups have yet been defined, helps to discover different groupings within data.	Identify groups that can be differentiated within a single, large group of customers. An example would be identifying tea drinkers who choose that beverage from others offered in flight from an airline.
Forecasting	Predicts values that can identify patterns in customer behavior.	Estimate the value of a future stream of dollar sales from a typical customer.
Sequence	Links events over time.	Identify a link between a person who buys a new house and subsequently will buy a new car within 90 days.

21. What is text mining?

Text mining is a software application used to extract key elements from unstructured data sets, discover patterns and relationships in the text materials, and summarize the information.

22. What is web mining?

Web mining seeks to find patterns, trends, and insights into customer behavior from users of the Web. Marketers, for example, use BA services like Google Trends www.google.com/trends/) and Google Insights for Search (http://google.about.com/od/i/g/google-insights-forsearch.htm) to track the popularity of various words andphrases to learn what consumers are interested in and what they are buying.

23. What is organizational structure?

Organizations are hierarchical, with senior managers making the strategic planning decisions, middlelevel managers making tactical planning decisions, and lower-level managers making operational planning decisions. Within the hierarchy, other organizational structures exist to support the development and existence of groupings of resources like those needed for BA. These additional structures include programs, projects, and teams.

24. What is program in organizational structure?

A program in this context is the process that seeks to create an outcome and usually involves managing several related projects with the intention of improving organizational performance. A program can also be a large project.

25. What is project?

A project tends to deliver outcomes and can be defined as having temporary rather than permanent social systems within or across organizations to accomplish particular and clearly defined tasks, usually under time constraints. Projects are often composed of teams.

26. What is a team?
A team consists of a group of people with skills to achieve a common purpose. Teams are especially appropriate for conducting complex tasks that have many interdependent subtasks.

PART- B
1. Why is leadership an important skill set for individuals looking to make a career in BA?
2. Why is categorizing data from its sources important in BA?
3. What is data quality, and why is it important in BA?
4. What is the difference between a data warehouse and a datamart?
5. Explain why a centralized business analytics (BA) organization structure has advantages over other structures.

The relationship of programs, projects, and teams with a business hierarchy is presented in Figure 4.1. Within this hierarchy, the organization's senior managers establish a BA program initiative to mandate the creation of a BA grouping within the firm as a strategic goal. A BA program does not always have an end-time limit. Middle-level managers reorganize or break down the strategic BA program goals into doable BA project initiatives to be undertaken in a fixed period of time. Some firms have only one project (establish a BA grouping) and others, depending on the organization structure, have multiple BA projects requiring the creation of multiple BA groupings. Projects usually have an end-time date in which to judge the successfulness of the project. The projects in some cases are further reorganized into smaller assignments, called BA team initiatives, to operationalize the broader strategy of the BA program. BA teams may have a long-standing time limit (for example, to exist as the main source of analytics for an entire organization) or have a fixed period (for example, to work on a specific product quality problem and then end).

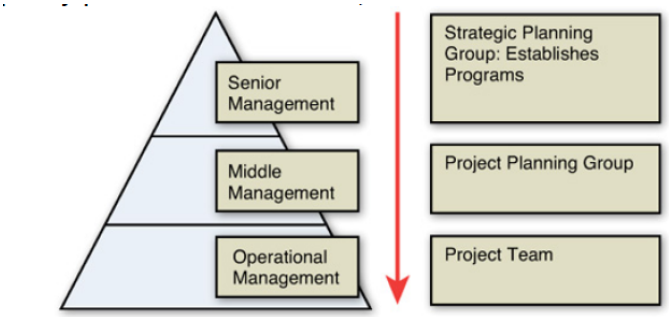


Figure 4.1 Hierarchal relationships program, project, and team planning

In summary, one way to look at the alignment of BA resources is to view it as a progression of assigned planning tasks from a BA program, to BA projects, and eventually to BA teams for implementation. As shown in Figure 4.1, this hierarchical relationship is a way to examine how firms align planning and decision-making workload to fit strategic needs and requirements. BA organization structures usually begin with an initiative that recognizes the need to use and develop some kind of program in analytics. Fortunately, most firms today recognize this need. The question then becomes how to match the firm's needs within the organization to achieve its

strategic, tactical, and operations objectives within resource limitations. Planning the BA resource allocation within the organizational structure of a firm is a starting place for the alignment of BA to best serve a firm's needs.

Aligning the BA resources requires a determination of the amount of resources a firm wants to invest. The outcome of the resource investment might identify only one individual to compute analytics for a firm. Because of the varied skill sets in information systems, statistics, and operations research methods, a more common beginning for a BA initiative is the creation of a BA team organization structure possessing a variety of analytical and management skills.
(We will discuss BA teams in Section 4.1.2.) Another way of aligning BA resources within an organization is to use a project structure. Most firms undertake projects, and some firms actually use a project structure for their entire organization. For example, consulting firms might view each client as a project (or product) and align their resources around the particular needs of that client. A project structure often necessitates multiple BA teams to deal with a wider variety of analytic needs. Even larger investments in BA resources might be required by firms that decide to establish a whole BA department containing all the BA resources for a particular organization. Although some firms create BA departments, the departments don't have to be large. Whatever the organization structure that is used, the role of BA is a staff (not line management) role in their advisory and consulting mission for the firm. In general, there are different ways to structure an organization to align its BA resources to serve strategic plans. In organizations where functional departments are structured on a strict hierarchy, separate BA departments or teams have to be allocated to each functional area, as presented in Figure 4.2. This functional organization structure may have the benefit of stricter functional control by the VPs of an organization and greater efficiency in focusing on just the analytics within each specialized area. On the other hand, this structure does not promote the cross-department access that is suggested as a critical success factor for the implementation of a BA program.

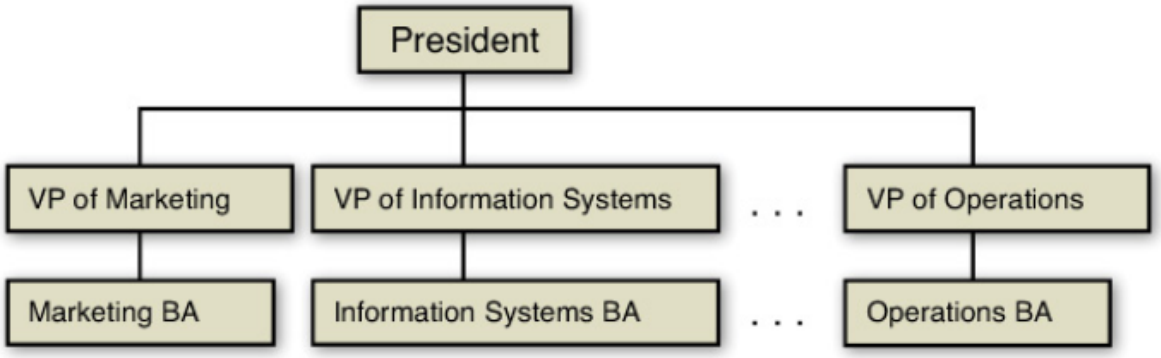


Figure 4.2 Functional organization structure with BA

The needs of each firm for BA sometimes dictate positioning BA within existing organization functional areas. Clearly, many alternative structures can house a BA grouping. For example, because BA provides information to users, BA could be included in the functional area of management information systems, with the chief information officer (CIO) acting as both the

director of information systems (which includes database management) and the leader of the BA grouping. An alternative organizational structure commonly found in large organizations aligns resources by project or product and is called a matrix organization. As illustrated in Figure 4.3, this structure allows the VPs some indirect control over their related specialists, which would include the BA specialists but also allows direct control by the project or product manager. This, similar to the functional organizational structure, does not promote the cross department access suggested for a successful implementation of a BA program.

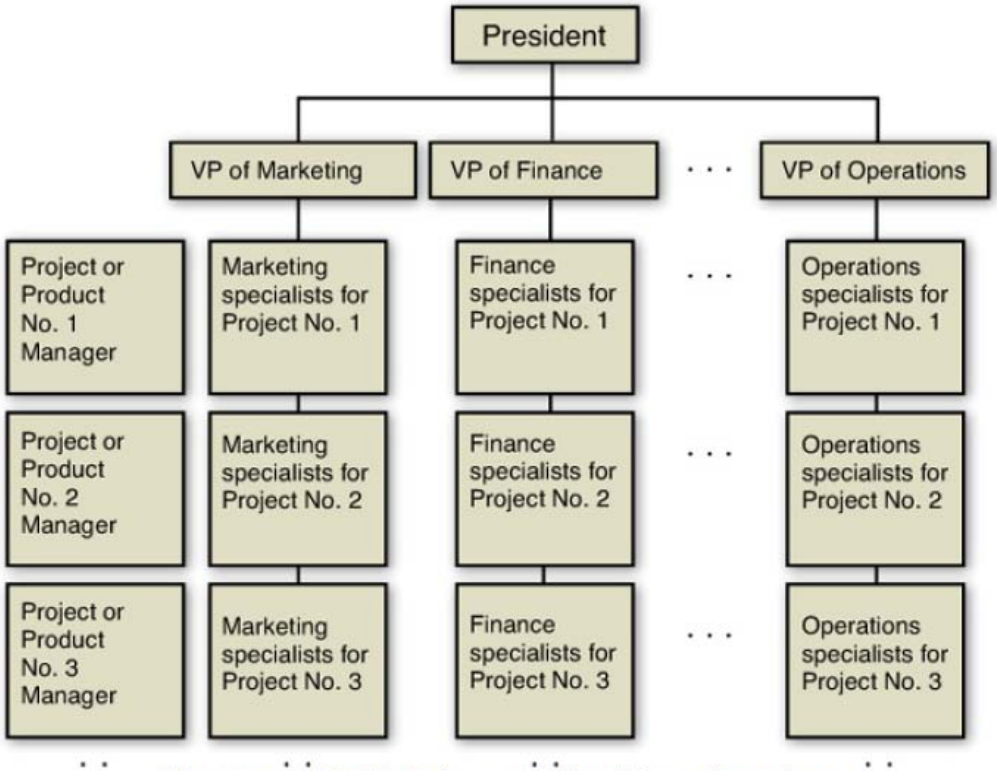


Figure 4.3 Matrix organization structure

The literature suggests that the organizational structure that best aligns BA resources is one in which a department, project, or team is formed in a staff structure where access to and from the BA grouping of resources permits access to all areas within a firm, as illustrated in Figure 4.4The dashed line indicates a staff (not line management) relationship. This centralized BA organization structure minimizes investment costs by avoiding duplications found in both the functional and the matrix styles of organization structures. At the same time, it maximizes information flow between and across functional areas in the organization. This is a logical structure for a BA group in its advisory role to the organization. Bartlett (2013, pp. 109–110) suggests other advantages of a centralized structure like the one in Figure 4.4. These include a reduction in the filtering of information traveling upward through the organization, insulation from political interests, breakdown of the siloed functional area communication barriers, a more central platform for reviewing important analyses that require a broader field of specialists, analyticbased group decision-making efforts, separation of the line management leadership from potential clients (for example, the VP of marketing would not necessarily come between the BA

group working on customer service issues for a department within marketing), and better connectivity between BA and all personnel within the area of problem solving.

solving.

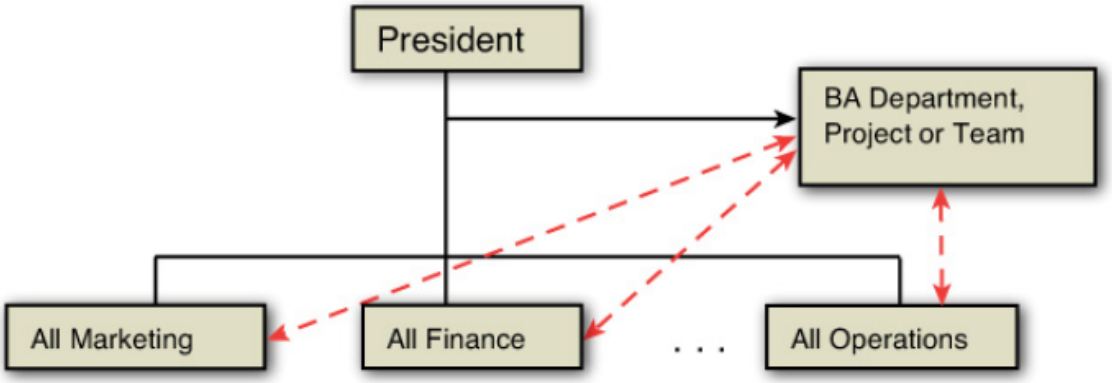


Figure 4.4 Centralized BA department, project, or team organization structure

In summary, the organizational structure that a firm may select for the positioning of their BA grouping can either be aligned within an existing organizational structure, or the BA grouping can be separate, requiring full integration within all areas of an organization. While some firms may start with a number of small teams to begin their BA program, other firms may choose to start with a full-sized BA department. Regardless of the size of the investment in BA resources, it must be aligned to allow maximum information flow between and across functional areas to achieve the most benefits BA can deliver.

7. Describe reasons why BA initiatives fail.

Reason	Description
Lack of Executive Sponsorship	Senior executive failure to recognize the value of BA and its importance eventually leads to a reduction in resources and eventual failure.
Limited Context Perception	There is an incorrect perception that analytics must be applied within a particular functional area in order to have the necessary validity to be applied to that area. Example: Financial regression analysis can only be applied correctly in the context of the finance area.
Belief of Physical Proximity	There is misperception that it takes physical proximity of the BA grouping in the business application area to be valid.
Lack of Leadership in BA Groupings	Without an advocate leader in the organization, as well as leaders in BA projects and teams to move the analysis to achieve desired goals, the entire BA effort will lead to eventual failure.
Lack of support	Without support for needed personnel, collecting data and technology to process the data will lead to failure.
Lack of Collaboration Across All Organizational Groups	Analytics that solve problems across multiple, functional areas are more likely to be accepted and successful than those that lack the cross-over into multiple organizational groups.
Lack of Skilled and Human Resources	BA departments, projects, or teams that don't have the skilled personnel to deal with the execution of analysis will eventually cause the failure of BA.
Inability to Delegate Responsibility	There is a desire to delegate responsibility to solve problems locally (a matter of trusting your own) rather than seeking help throughout the organization. This impedes the flow of problem solving efforts by an external BA department and impedes communication of information needed to successfully apply BA.
Lack of Integrated Processes	Information that is stored in silos and not shared makes it more difficult for BA analysis to succeed.

Table 4.1 Reasons for BA Initiative and Organization Failure

1. Describe typical BA team roles and reasons for their failures.

When it comes to getting the BA job done, it tends to fall to a BA team. For firms that employ BA teams the participants can be defined by the roles they play in the team effort. Some of the roles BA team participants undertake and their typical background are presented in table 4.2

Title or Function	Role Description	Background or Skills of Participant
Analytics Modeler	Develop and maintain predictive and forecasting models to provide insight.	Statistics, operation research, analytic modeling.
Analytics Process Designer	Develop and enforce reusable processes to reduce BA execution time.	Management consultant, process mapping, systems design.
Analytics Analyst	Respond to BA inquiries from functional areas within the firm to gain insight.	Reporting, problem solving, communicating, and providing customer service.
BA Team Head	Provide leadership to BA team, define strategies and tactics to ensure improved business performance, and interface with management.	BA manager or administrator.
Business Domain Expert	Provide business experience to ensure relevance of insight, help interpret business measures and the meaning of data.	Business experience in the area where the problem or opportunity exists.
Data Manager	Ensure data availability and access while minimizing costs.	Data modeling or warehousing, experience in data quality processes.
Implementation Specialist	Ensure rapid and robust model deployment to reduce time in interface.	Information system and data warehousing expertise, enterprise architecture experience.
Monitoring Analyst	Identify, establish, and enforce common analytics to be used to measure value and optimize effort.	Management and BA expert, predictive and financial modeling, process design, and team mentoring.

*Source: Adapted from Stubbs (2013), pp.137–149; Stubbs (2011) Table 3.3; Laursen and Thorlund (2010), p.15.

Table 4.2 BA Team Participant Roles*

Aligning BA teams to achieve their tasks requires collaboration efforts from team members and from their organizations. Like BA teams, collaboration involves working with people to achieve a shared and explicit set of goals consistent with their mission. BA teams also have a specific mission to complete. Collaboration through teamwork is the means to accomplish their mission. Team members' need for collaboration is motivated by changes in the nature of work (no more silos to hide behind, much more open environment, and so on), growth in professions (for example, interactive jobs tend to be more professional, requiring greater variety in expertise sharing), and the need to nurture innovation (creativity and innovation are fostered by collaboration with a variety of people sharing ideas). To keep one's job and to progress in any business career, particularly in BA, team members must encourage working with other members

inside a team and out. For organizations, collaboration is motivated by the changing nature of information flow (that is, hierarchical flows tend to be downward, whereas in modern organizations, flow is in all directions) and changes in the scope of business operations (that is, going from domestic to global allows for a greater flow of ideas and information from multiple sources in multiple locations).

How does a firm change its culture of work and business Operations to encourage collaboration? One way to affect the culture is to provide the technology to support a more open, cross-departmental information flow. This includes email, instant messaging, wikis (collaboratively edited works, like Wikipedia), use of social media and networking through Facebook and Twitter, and encouragement of activities like collaborative writing, reviewing, and editing efforts. Other technology supporting collaboration includes webinars, audio and video conferencing, and even the use of iPads to enhance face-to-face communication. These can be tools that change the culture of a firm to be more open and communicative. Reward systems should be put into place to reward team effort. Teams should be rewarded for their performance, and individuals should be rewarded for performance in a team. While middle-level managers build teams, coordinate their work, and monitor their performance, senior management should establish collaboration and teamwork as a vital function.

Reasons for failure

Despite the collaboration and best of intentions, BA teams sometimes fail. There are many reasons for this, but knowing some of the more common ones can help managers avoid them. Some of the more common reasons for team failure are presented in Table 4.3. They also represent issues that can cause a BA program to become unaligned and unproductive.

misaligned and unproductive.

Reason for Failure	Descriptions
Lack of Communication	It is not enough to come up with valuable information for decision-making and to find business opportunities in data. That information must be shared with users, clients, and everyone within a firm for benefit to come from it. It is only when analytics show a tangible and beneficial outcome that they are considered business analytics (BA). If those results are not communicated on a continual basis, BA teams can be perceived to provide less value to the organization.
Failure to Deliver	Not every BA team will be able to deliver valued information if the team lacks the ability or resources to deliver needed answers and information. The greater the number of BA team failures, the greater are the chances that the team will be eliminated.
Lack of Justification	BA teams require resource allocations. Those allocations come from other departments that supposedly benefit from BA contributions. Without the role of BA and its potential contributions to a firm being clearly spelled out, users might not associate the ongoing efforts of a BA team as being worth the money spent on them.
Fail to Provide Value	BA teams have to sell their roles and suggested solutions or ideas. Without a clear understanding of value for potential users, the team faces a hard sell.
Inability to Prove Success	BA teams need to document and measure the impact of their ideas and suggestions. Without that proof, potential users might not support future BA efforts.

*Source: Adapted from Flynn (2008) pp. 99–106 and Stubbs (2011) p. 89.

Table 4.3 Reasons for BA Team Failures*

2. Explain why establishing an information policy is important.

The information policy specifies organizational rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying all types of information and data. It defines the specific procedures and accountabilities that identify which users and organizational units can share information, where the information can be distributed, and who is responsible for updating and maintaining the information. In small firms, business owners might establish the employed in businesses. It is specifically focused on promoting data privacy, data security, data quality, and compliance with government regulations. Such information policy, data administration, and data governance must be in place to guard and ensure data is managed for the betterment of the entire organization. These steps are also important in the creation of database management systems information policy. For larger firms, data administration may be responsible for the specific policies and procedures for data management (Siegel and Shim, 2003, p. 280). Responsibilities could include developing the information policy, planning data collection and storage, overseeing database design, developing the data dictionary, as well as monitoring how information systems specialists and end user groups use data. A more popular term for many of the activities of data administration is data governance, which includes establishing policies and processes for managing the availability, usability, integrity, and security of the data

11. Describe how data can be scrubbed.

Business analytics, if relevant, is based on data assumed to be of high quality. Data quality refers to accuracy, precision, and completeness of data. High-quality data is considered to correctly reflect the real world in which it is extracted. Poor quality data caused by data entry errors, poorly maintained databases, out-of-date data, and incomplete data usually leads to bad decisions and undermines BA within a firm. Organizationally, the database management systems (DBMS, mentioned in Chapter 3) personnel are managerially responsible for ensuring data quality. Because of its importance and the possible location of the BA department outside of the management information systems department (which usually hosts the DBMS), it is imperative that whoever leads the BA program should seek to ensure data quality efforts are undertaken.

Ideally, a properly designed database with organizationwide data standards and efforts taken to avoid duplication or inconsistent date elements should have high-quality data. Unfortunately, times are changing, and more organizations allow customers and suppliers to enter data into databases via the Web directly. As a result, most of the quality problems originate from data input such as misspelled names, transposed numbers, or incorrect or missing codes. An organization needs to identify and correct faulty data and establish routines and procedures for editing data in the database. The analysis of data quality can begin with a data quality audit, where a structured survey or inspection of accuracy and level of completeness of data is undertaken. This audit may be of the entire database, just a sample of files, or a survey of end users for perceptions of the data quality. If during the data quality audit files are found that have errors, a process called data cleansing or data scrubbing is undertaken to eliminate or repair data. Some of the areas in a data file that should be inspected in the audit and suggestions on how to correct them are presented in

Table 4.6.

Data Inspection Items	Description and Cleansing/Scrubbing Recommendation
Current Data	Check to make sure the data is current. If it is out of date, remove it.
Completeness	Check to see if there is missing data. If more than 50% is missing, remove the entire file from the database.
Relevance	Check to see if the data is no longer relevant for the purpose for which it was collected. If it's no longer relevant, consider removing it from the database.
Duplication	Check to see if duplicate data files exist in the database. Remove duplicate data.
Outliers	Check for extreme values (outliers) in quantitative data files for possible errors in data coding. Remove from the data file any suspected of being in error, or repair the data.
Inconsistent Values	If data fields contain both characters and real numbers data where only characters or numbers should be, explore repairing the data.
Coding	If suspicious or unknown coding of data exists in data files, remove from the database or repair the coding of data.

Table 4.6 Quality Data Inspection Items and Recommendations

12. Explain what change management involves and what its relationship is to BA.

Change management is defined as an approach for transitioning the organization (individuals, teams, projects, departments) to a changed and desired future state (Laudon and Laudon, 2012, pp. 540–542). Change management is a means of implementing change in an organization, such as adding a BA department (Schermerhorn 2001, pp. 382–390). Changes in an organization can be either planned (a result of specific and planned efforts at change with direction by a change leader) or unplanned (spontaneous changes without direction of a change leader). The application of BA invariably will result in both types of changes because of BA's specific problem-solving role (a desired, planned change to solve a problem) and opportunity finding exploratory nature (i.e., unplanned new knowledge opportunity changes) of BA. Change management can also target almost everything that makes up an organization (see Table 4.7).

Change Target	Description
Culture	This represents the changing values and norms of the individuals and groups that make up the organization. BA has to sell itself in some situations, build trust, and alter decision-making. It often requires a different culture of thinking about decision-making.
Organization Structure	This is the changing organizational lines of authority and communication. The cross-departmental nature of BA positions may provide information that changes the organization and alters relationships and tasks.
Personnel	BA information about the need for human resource changes in attitudes and skills can mandate changes that permit an organization to achieve higher business performance levels.
Tasks	BA analysis might find that some job designs, specifications, and descriptions that employees perform need to have their objectives and goals changed to achieve higher business performance levels.
Technology	BA analysis might find information system technology used in the design and workflow that integrate employees and equipment into operating systems and require change to achieve higher business performance levels.

*Source: Adapted from Figure 7 in Schniederjans and Cao (2002), pp. 261.

It is not possible to gain the benefits of BA without change. The intent is change that involves finding new and unique information on which change should take place in people, technology systems, or business conduct. By instituting the concept of change management within an organization, a firm can align resources and processes to more readily accept changes that BA may suggest. Instituting the concept of change management in any firm depends on the unique characteristics of that firm. There are, though, a number of activities in common with successful change management programs, and they apply equally to changes in BA departments, projects, or teams. Some of these activities that lead to change management success are presented as best practices in Table 4.8.

Best Practice	Description
Champion	Change is scary business for some, and a strong leader for change can champion the change effort, calming fears and explaining the need for change. The champion also helps direct efforts, motivate change, and keep the change activities on track.
Clearly Stated Goals	Any type of change should be clearly defined, including what the changes are, which personnel have to change, and what the processes involve and how they affect technology. This would also include deadlines needed to keep the change effort on track.
Good Communication	To avoid resistance to change (a natural norm to anything that is new), it is useful to help those facing the change understand its value through effective and repeated communications, keeping them informed on progress and easing fears.
Measured Performance	Any goals stated prior to the launch of change can be used to measure performance during the changeover period. Seeing business performance improve with changes can motivate further change and support by those impacted.
Senior Management Support	Critical to all BA departments, projects, or teams is the need for senior management to support change efforts. Sometimes that support is in direct dollars, and sometimes it's in lending authority to get resources needed for BA work.

Table 4.8 Change Management Best Practices

Structuring a BA department, undertaking a BA project, or setting up a BA team within an organization can largely determine successfulness in aligning resources to achieve information-sharing goals. In this chapter, several organization structures (functional, matrix, and centralized) were discussed as possible homes for BA resource groupings. The role of BA teams as an important organizational resource aligning tool was also presented. In addition, this chapter discussed reasons for BA organization and team failures. Other managerial issues included in this chapter were establishing an information policy, outsourcing business analytics, ensuring data quality, measuring business analytics contribution, and managing change. Once a firm has set up the internal organization for a BA department, program, or project, the next step is to undertake BA. In the next chapter, we begin the first of the three chapters devoted to detailing how to undertake the three steps of the BA process.

Discussion Questions

1. The literature in management information systems consistently suggests that a decentralized approach to resource allocation is the most efficient. Why then do you think the literature in BA suggests that the opposite —a centralized organization—is the best structure?

2. Why is collaboration important to BA?

3. Why is organization culture important to BA?

4. How does establishing an information policy affect BA?

5. Under what circumstances is outsourcing BA good for the development of BA in an organization?
6. Why do we have to measure BA contributions to an organization?
7. How does data quality affect BA?
8. What role does change management play in BA?
9.Explain why business analytics are important in solving business problems.
10.Explain why business analytics are important in identifying new business initiatives.
11.Describe the kinds of questions business analytics can help answer.
12.Explain how business analytics can help an organization achieve a competitive advantage.
13.Explain different types of competitive advantages and their relationship to business analytics.
14.Explain the importance of business analytics for a business organization.
15.Describe the differences between BA programs, projects, and teams and how they are used to align BA resources in firms.

UNIT-3

UNIT III DESCRIPTIVE ANALYTICS 9

Introduction to Descriptive analytics - Visualising and Exploring Data - Descriptive Statistics - Sampling and Estimation - Probability Distribution for Descriptive Analytics - Analysis of Descriptive analytics

1. What is descriptive analytics?

Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question "What happened?" (or What is happening?), characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives. ***Descriptive analytics*** is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.

2.How does descriptive analytics work?

Data aggregation and **data mining** are two techniques used in descriptive analytics to discover historical data. Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts. Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning. Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.

3.list the examples of Descriptive analytics

- Traffic and Engagement Reports
- Financial Statement Analysis
- Financial statements are periodic reports that detail financial information about a business and, together, give a holistic view of a company's financial health.
- Demand Trends

4. What is visualization of data?

Visualization allows business users to recognize relationships between the data, providing greater meaning to it. Exploring these patterns helps users focus on specific areas that require attention in the data, so that they can identify the significance of those areas to drive their business forward.

5. *What are the Features of Data Visualization?*

- It is interactive and exposed trends,
- It contributes a viewpoint, narrates a story and describes the process,
- It applies animation and real images,
- It fixes data into meaning and conserves time,
- It grants access to raw data and extracts meaningful, knowledgeable insights.

6. What is Descriptive Statistics?

Descriptive statistics are **brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population**. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

7.What is meant by variate?

A variate or random variable is a quantity or attribute whose value may vary from one unit of investigation to another. For example, the units might be headache sufferers and the variate might be the time between taking an aspirin and the headache ceasing.

8.List the types of variate

- Qualitative or nominal; described by a word or phrase (e.g. blood group, colour).
- Quantitative; described by a number (e.g. time till cure, number of calls arriving at a telephone exchange in 5 seconds). Quantitative data can be: Discrete: the variate can only take one of a finite or countable number of values• (e.g. a count) Continuous: the variate is a measurement which can take any value in an interval• of the real line (e.g. a weight).
- Ordinal; this is an "in-between" case. Observations are not numbers but they can be ordered (e.g. much improved, improved, same, worse, much worse).

3. What is frequency?

The frequency of a value is the number of observations taking that value. A frequency table is a list of possible values and their frequencies.

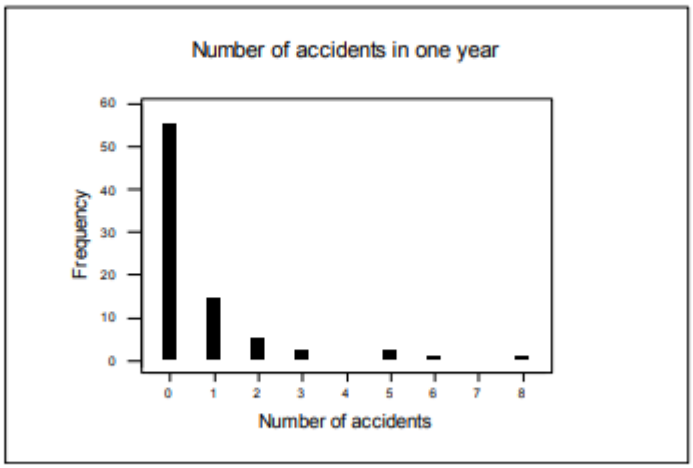
10.The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below. Construct a frequency table and draw a bar chart.

2	0	0	1	0	3	0	6	0	0	8	0	2	0	1
5	1	0	1	1	2	1	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	5	1	0	0	0	0	0	0	0	0	1	1
0	3	0	0	1	1	0	0	0	2	0	1	0	0	0
0	0	0	0	0										

Solution

Number of accidents	Tallies	Frequency
0	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	55
1	𝍸 𝍸 𝍸	14
2	𝍸	5
3	||	2
4		0
5	||	2
6	|	1
7		0
8	|	1

Barchart



4. What is Continuous data- histograms?

When the variate is continuous, we do not look at the frequency of each value, but group the values into intervals. The plot of frequency against interval is called a histogram. Be careful to define the interval boundaries unambiguously.

5. Draw a frequency table for the following data

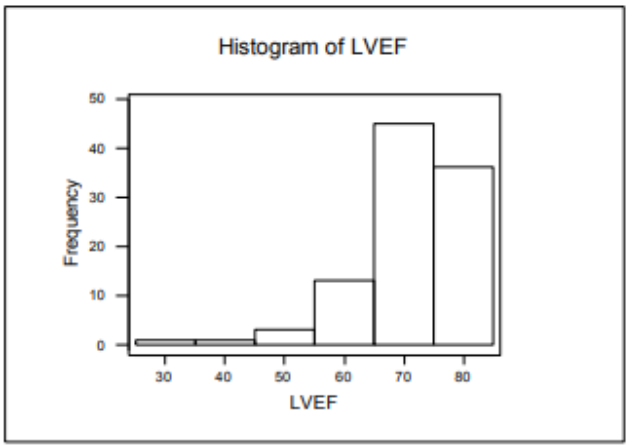
There are the left ventricular ejection fractions (LVEF) for a group of 99 heart transplant patients. Construct a frequency table and histogram.

62	64	63	70	63	69	65	74	67	77	65	72	65
77	71	79	75	78	64	78	72	32	78	78	80	69
69	65	76	53	74	78	59	79	77	76	72	76	70
76	76	74	67	65	79	63	71	70	84	65	78	66
72	55	74	79	75	64	73	71	80	66	50	48	57
70	68	71	81	74	74	79	79	73	77	80	69	78
73	78	78	66	70	36	79	75	73	72	57	69	82
70	62	64	69	74	78	70	76					

Frequency table

LVEF	Tallies	Frequency
24.5 - 34.5	|	1
34.5 - 44.5	|	1
44.5 - 54.5	|||	3
54.5 - 64.5	|||| |||| |||	13
64.5 - 74.5	|||| |||| |||| |||| |||| |||| |||| |||| ||||	45
74.5 - 84.5	|||| |||| |||| |||| |||| |||| |||| |	36

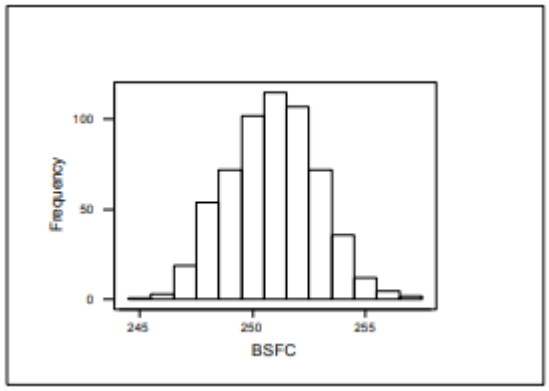
Histogram



Note: if the interval lengths are unequal, the heights of the rectangles are chosen so that the area of each rectangle equals the frequency i.e. height of rectangle = frequency ÷ interval length.

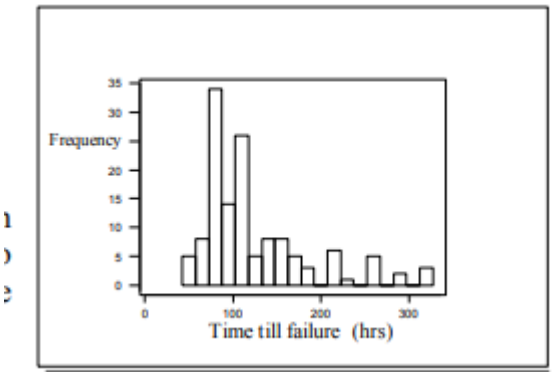
6. What is Normally distributed data?

The histogram is bell-shaped, like the probability density function of a Normal distribution. It appears, therefore, that the data can be modelled by a Normal distribution. (Other methods for checking this assumption are available.) Similarly, the histogram can be used to see whether data look as if they are from an Exponential or Uniform distribution.



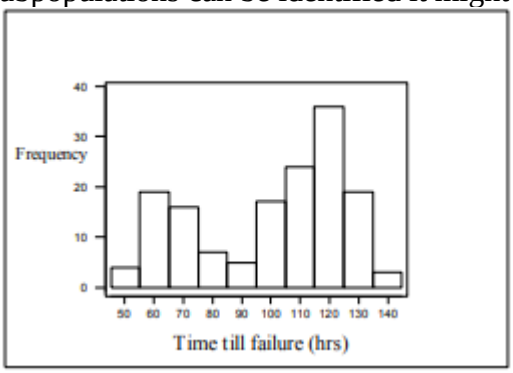
7. What is meant by Very skew data?

The relatively few large observations can have an undue influence when comparing two or more sets of data. It might be worthwhile using a transformation e.g. taking logarithms.



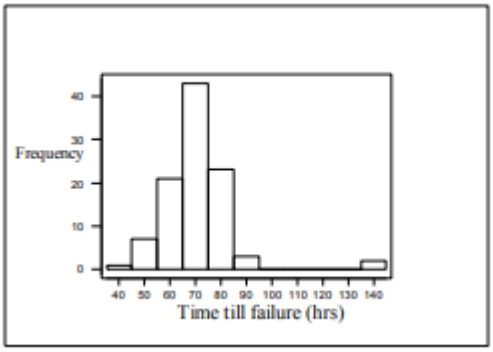
8. What is meant by Bimodality?

This may indicate the presence of two subpopulations with different characteristics. If the subpopulations can be identified it might be better to analyse them separately.



9. What is meant by Outliers?

The data appear to follow a pattern with the exception of one or two values. You need to decide whether the strange values are simply mistakes, are to be expected or whether they are correct but unexpected. The outliers may have the most interesting story to tell.

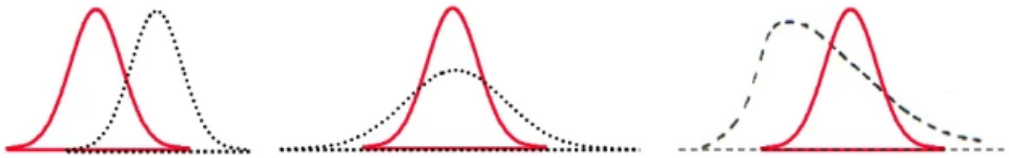


16.What are the measures of descriptive statistics?

- Descriptive Statistics

The following measures are used to describe a data set:

- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.



17. What is measures of position?

A measure of position **determines the position of a single value in relation to other values in a sample or a population data set**. Unlike the mean and the standard deviation, descriptive measures based on quantiles are not sensitive to the influence of a few extreme observations. It comprises of

Mean: The average value in a dataset.
Median: The middle value in a dataset.
Mode: The most frequently occurring value(s) in a dataset.

18. Define Mean.
Mean is the average of the given set of values. It denotes the equal distribution of values for a given data set. Mean = (Sum of all the observations/Total number of observations)
Example of **Mean:** Marketers often calculate the mean revenue earned per advertisement so they can understand how much money their company is making on each ad.

19.What is arithmetic mean?
Arithmetic mean is the total of the sum of all values in a collection of numbers divided by the number of numbers in a collection. It is calculated in the following way:

$$\text{Arithmetic mean} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

20.Geometric mean is an *n*th root of the product of all numbers in a collection. The formula for the geometric mean is:

$$\text{Geometric mean} = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

21. Compute Arithmetic mean and geometric mean.

Jim wants to find a stock for investment. He is a big fan of Apple Inc. He knows that the company has strong financials. However, to ensure that this investment will bring him a substantial return, he has decided to check how the stock performed in the past. He decides to find the average price of Apple's share price for the past five months. He gathered the monthly company's stock prices from January 2018 to June 2018 and found the monthly returns. The stock prices and returns are summarized in the table below:

	Stock Price	Return (%)	Return (decimal)
December	167.90	N/A	
January	166.11	-1.07%	0.99
February	176.72	6.39%	1.06
March	167.14	-5.42%	0.95
April	164.63	-1.50%	0.98
May	186.15	13.07%	1.13
June	185.11	-0.56%	0.99

Solution:
The formula used for the calculation would be the following:

$$\text{Arithmetic mean} = \frac{-1.07\% + 6.39\% - 5.42\% - 1.50\% + 13.07\% - 0.56\%}{6} = 1.82\%$$

The geometric mean is equal to:

$$\text{Geometric mean} = \sqrt[6]{0.99 \times 1.06 \times 0.95 \times 0.98 \times 1.13 \times 0.99} = 1.0164 \text{ or } 1.64\%$$

22. What is median?
Median is the middle number in a sorted list of numbers. To determine the median value in a sequence of numbers, the numbers must first be sorted, or arranged, in value order from lowest to highest or highest to lowest.

- Imagine that you have an e-commerce Site. Ex: Meesho
- You are interested in offering the product that best suits them.
- In order to do this, you have asked your clients for their personal financial situation
- You find out that:
- 98% of your customers have a monthly wage between 10000 and 30000.

23. Define mode.
The mode is the value that appears most frequently in a data set.
Example:3, 3, 6, 9, **16, 16, 16**, 27, 27, 37, 48

- Imagine you own a handmade jewelry on-line business.
- You are planning to launch a new product, but you are not sure whether it should be a necklace or a bracelet.
- After a Market study, you conclude that:
- **60% of people Who is buying handmade jewelry buy necklaces**.
- 20% buy bracelets.
- You then **decide to launch a new necklace**

24.What are measures of spread?

Measures of spread describe how similar or varied the set of observed values are for a particular variable (data item). Measures of spread include the **range, quartiles and the interquartile range, variance and standard deviation**.

25. Define range.

The range is **the size of the smallest interval (statistics) which contains all the data and provides an indication of statistical dispersion**. It is measured in the same units as the data. Since it only depends on two of the observations, it is most useful in representing the dispersion of small data sets.

26.List descriptive statistics with examples.

Statistics	Computation (in Data Set)	Application Area	Example	Application Notes
N or Count	Number of values.	Any.	Sample size of a company's transactions during a month.	Useful in knowing how many items were used in the statistics computations.
Sum	Total of the values in the entire data set.	Any.	Total sales for a company.	Useful in knowing the total value.
Mean	Average of all values.	Any.	Average sales per month.	Useful in capturing the central tendency of the data set.
Median	Midpoint value in the data set arranged from high to low.	Finding the midpoint in the distribution of data.	Total income for citizens of a country.	Useful in finding the point where 50 percent of the data is above and below.
Mode	Most common value in the data set.	Where values are highly repeated in the data set.	Fixed annual salaries where a limited number of wage levels are used.	Useful in declaring a common value in highly repetitive data sets.
Maximum/ Minimum	Largest and smallest values, respectively.	To conceptualize the spread of the data's distribution.	Largest and smallest sales in a day.	Useful in providing a scope or end points in the data.

27. When to use Stratified Random Sampling?

- Stratified random sampling is an extremely productive method of sampling in situations where the researcher intends to focus only on specific strata from the available population data. This way, the desired characteristics of the strata can be found in the survey sample.

- Researchers rely on this sampling method in cases where they intend to establish a relationship between two or more different strata. If this comparison is conducted using simple random sampling, there is a higher likelihood of the target groups being not equally represented.
- Samples with a population which are difficult to access or contact, can be easily be involved in the research process using the stratified random sampling technique.
- The accuracy of statistical results is higher than simple random sampling since the elements of the sample and chosen from relevant strata. The diversification within the strata will be much lesser than the diversification which exists in the target population. Due to the accuracy involved, it is highly probable that the required sample size will be much lesser and that will help researchers in saving time and efforts

28.Distinguish Cluster sampling from stratified sampling

Cluster sampling	Stratified sampling
Elements of a population are randomly selected to be a part of groups (clusters).	The researcher divides the entire population into even segments (strata).
Members from randomly selected clusters are a part of this sample.	Researchers consider individual components of the strata randomly to be a part of sampling units.
Researchers maintain homogeneity between clusters.	Researchers maintain homogeneity within the strata.
Researchers divide the clusters naturally.	The researchers or statisticians primarily decide the strata division.
The key objective is to minimize the cost involved and enhance competence.	The key objective is to conduct accurate sampling, along with a properly represented population.

29. Distinguish probability sampling from non-probability sampling

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

30. What is quota sampling?

Quota sampling is one of the methods of non-probability sampling methods in which the researcher generates a sample involving individuals that represent the population. Here, the researcher will choose the individual based on specific qualities and traits.

31. What are the two types of quota sampling?

The two types of quota sampling are:
Controlled quota sampling
Uncontrolled quota sampling

32.What are the advantages of quota sampling?

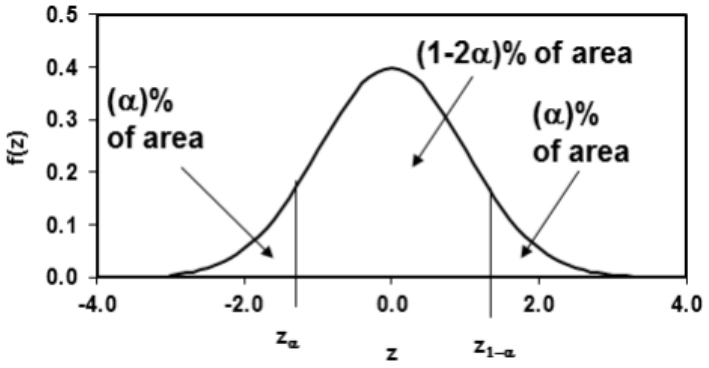
The advantages of quota sampling are:
Saves money
Saves time
Convenient for research purposes
Exactly represent the population of interest

33.Mention a few characteristics of quota sampling?

A few characteristics of quota sampling are:
The sample represents the whole population
It aims to achieve the best representation of respondents in the final sample.
During the research process, the researcher divides the population into subgroups.

34. Define confidence interval.

A confidence interval is some subset of random variable space with which someone can say something like, “I am 95% sure that the true population mean is between μlow and μhi .” In this section, we discuss how a confidence interval is defined and calculated. The confidence interval is defined by a percent. This percent is called (1-2α). So if α=0.05, then you would have a 90% confidence interval. The concept of a confidence interval is illustrated in graphical terms in Figure 6.4.



Confidence interval = (sample statistic) ± [(confidence coefficient) × (standard error of the estimate)]

35.What is meant by sample statistic in Confidence interval?
The sample statistic in the confidence interval can be any measure or proportion from a sample that is to be used to estimate a population parameter, such as a measure of central tendency like a mean.

36. What is meant by confidence coefficient?
The confidence coefficient is set as a percentage to define the degree of confidence to accurately identify the correct sample statistic. The larger the confidence coefficient, the more likely the population mean from the sample will fall within the confidence interval.

37. What is Frequency Theory of probability?
It basically states that, through a large number of trials, the relative frequency outcome of an event, A, can be used to determine the probability of A, represented here as P(A). This probability is based on
experiential observations, and it is assumed that the experiment accurately represents the possible behaviors that are being observed. The probabilities that are determined using this approach are based on past observations that are converted into probabilities to be used in the future. Because the probabilities are based on past events, we sometimes refer to these probabilities as being posteriori probabilities.

38. Using Frequency Theory, what is the probability that a food server will be tipped an amount that will fall in the tip class interval of $8 to $9.99 based on the following collection of data from a food server's experiences at the restaurant?

Number of Customers

Tips	Who Tipped
$6 to $7.99	50
$8 to $9.99	120
$10 to $12.99	30
Total	200

Answer: The relative frequency of the class interval of $8–$9.99 is 120 out of 200. If we assume the 200 tip experiments recorded are sufficient to accurately describe all tip behavior of the customers for the specific server, Frequency Theory holds that the probability of a tip between $8 and $9.99 being given to that server is 0.60, or 60 percent (120 / 200 = 0.60).

39.What is Principle of insufficient Reason?
The Principle of Insufficient Reason states that each possible outcome in an experiment is equally probable if there is no evidence to challenge the assumption. Alternatively, if there is no reason to prefer one outcome over another, each outcome is equally likely or has the same probability. To determine probabilities using this principle, we must logically abstract the decimal value probabilities by dividing the frequency of occurrence by the total number of

possible outcomes. Because there is no evidence to support that any of the possible outcomes is any more probable to occur than the rest, each will be given an equal probability.

40. A stock broker must select one stock to invest in out of a sample of four stocks. The stock broker is unfamiliar with the four sample stocks.
What is the probability that the stock broker will pick the best stock out of the four?
Answer: Assuming that there is only one "best" stock in the sample of four, and because there is one chance out of four to
obtain it, the probability of choosing the best one out of four is 0.25 (1 / 4 = 0.25). In this problem, each of the four is equally likely to be the best stock, because no additional information on the stocks is available.

41. What is subjective approach to probability?
In the subjective approach to probability, estimation of probabilities is based on personal opinion or judgment. Under this approach, we assume that an individual's subjective judgment may be as accurate as or better than any other objective approach. BA analysts, whose subjective expertise is combined with the objective frequency information provided by a database, can combine these sources to greatly improve probability estimates for decision-making.

42. List the parametric tests used in BA.

Test Statistic	Application Area	Access to SPSS Function	Access to Excel Function
F-Test Two Sample for Variances	Compares the variances from two samples to see if they are from the same probability distribution	Analyze > Compare Means > One-Way ANOVA	Data Analysis > F-Test Two Sample for Variances
t-test: Paired Two Sample Means	Compares the mean values from two samples to see if they come from the same probability distribution	Analyze > Compare Means > Paired Samples t-test	Data Analysis > t-test: Paired Two Sample Means
Z-test: Two Sample Means	Compares the mean values from two populations to see if they have the same probability distribution	For samples sizes above 30: Analyze > Compare Means > Paired Samples t-test	Data Analysis > Z-test: Two Sample Means
ANOVA: Single Factor	Compares the variance between and within two or more samples to see if the samples are drawn from the same probability distribution	Analyze > Compare Means > One-Way ANOVA	Data Analysis > ANOVA: Single Factor

Table A.2 Common Parametric Statistical Tests and Software Access Information

43. List non-parametric test

Test Statistic	Application Area
Binomial Test	This test compares the observed frequencies of the two categories of a dichotomous variable to the frequencies that are expected under a binomial distribution with a specified probability parameter (α).
Chi-Square	This test (used often for a goodness-of-fit test) compares the observed and expected frequencies in each category of a distribution to test that all categories contain the same proportion of values or to test that each category contains a user-specified proportion of values.
Kolmogorov-Smirnov (One-Way)	Multiple versions of this test can be applied to differing comparative analyses. The one-way procedure compares the observed cumulative distribution function for a variable with any specified theoretical distribution.
Wilcoxon Signed-Rank	Multiple versions of this test can be applied to differing comparative analyses. This test is applicable when two samples (two populations) are related (not independent). The test is designed to compare some n matched pairs of ranked or ordinal data from two populations.
Run	A *run* is a sequence of like observations. This tests whether the order of occurrence of two values of a variable is random. A sample with too many or too few runs suggests that the sample is not random.

Table A.4 Common Nonparametric Statistical Tests and Software Access Information

Part- B

1. Discuss the Importance of Data Visualization

Data visualization is essential to assist businesses in quickly identifying data trends, which would otherwise be a hassle. The pictorial representation of data sets allows analysts to visualize concepts and new patterns. With the increasing surge in data every day, making sense of the quintillion bytes of data is impossible without Data Proliferation, which includes data visualization. Every professional industry benefits from understanding their data, so data visualization is branching out to all fields where data exists. For every business, information is their most significant leverage. Through visualization, one can prolifically convey their points and take advantage of that information. A dashboard, graph, infographics, map, chart, video, slide, etc. all these mediums can be used for visualizing and understanding data. Visualizing the data enable decision-makers to interrelate the data to find better insights and reap the importance of data visualization, which are:

1. Analyzing the Data in a Better Way

Analyzing reports helps business stakeholders focus on the areas that require attention. The visual mediums help analysts understand the key points needed for their business. Whether it is a sales report or a marketing strategy, a visual representation of data helps companies increase their profits through better analysis and better business decisions.

2. Faster Decision Making

Humans process visuals better than any tedious tabular forms or reports. If the data communicates well, decision-makers can quickly take action based on the new data insights, accelerating decision-making, and business growth simultaneously.

3. Making Sense of Complicated Data

Data visualization allows business users to gain insight into their vast amounts of data. It benefits them to recognize new patterns and errors in the data. Making sense of these patterns helps the users pay attention to areas that indicate red flags or progress. This process, in turn, drives the business ahead.

2. What is data visualization? Explain the Purpose of Data Visualization?

The science of data visualization

The science of data visualization comes from an understanding of how humans gather and process information. Daniel Kahn and Amos Tversky collaborated on research that defined two different methods for gathering and processing information. System 1 focuses on thought processing that is fast, automatic and unconscious. This method is frequently used in day-to-day life and helps accomplish:

- reading the text on a sign;
- solving simple math problems, like 1+1;
- identifying where a sound is coming from;
- riding a bike; and
- determining the difference between colors.

System 2 focuses on slow, logical, calculating and infrequent thought processing. This method is used in one of the following situations:

- reciting a phone number;
- solving complex math problems, like 132 x 154;
- determining the difference in meaning between multiple signs standing side by side; and
- understanding complex social cues.

Data visualization tools and vendors

Data visualization tools can be used in a variety of ways. The most common use today is as a business intelligence (BI) reporting tool. Users can set up visualization tools to generate automatic dashboards that track company performance across key performance indicators (KPIs)

and visually interpret the results.The generated images may also include interactive capabilities, enabling users to manipulate them or look more closely into the data for questioning and analysis. Indicators designed to alert users when data has been updated or when predefined conditions occur can also be integrated.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.As data visualization vendors extend the functionality of these tools, they are increasingly being used as front ends for more sophisticated big data environments. In this setting, data visualization software helps data engineers and scientists keep track of data sources and do basic exploratory analysis of data sets prior to or after more detailed advanced analyses.The biggest names in the big data tools marketplace include Microsoft, IBM, SAP and SAS. Some other vendors offer specialized big data visualization software; popular names in this market include Tableau, Qlik and Tibco.

The purpose of data visualization is pretty clear. It is to make sense of the data and use the information for the organization's benefits. That said, data is complicated, and it gains more value as and when it gets visualized. Without visualization, it is challenging to quickly communicate the data findings and identify patterns to pull insights and interact with the data seamlessly. Data scientists can find patterns or errors without visualization. However, it is crucial to communicate data findings and identify critical information from them. And for this, interactive data visualization tools make all the difference.

- Data visualization strengthens the impact of messaging for your audiences and presents the data analysis results in the most persuasive manner. It unifies the messaging systems across all the groups and fields within the organization.
- Visualization lets you comprehend vast amounts of data at a glance and in a better way. It helps to understand the data better to measure its impact on the business and communicates the insight visually to internal and external audiences.
- Decisions can't be made in a vacuum. Available data and insights enable decision-makers to aid decision analysis. Unbiased data without inaccuracies allows access to the right kind of information and visualization to represent that information and keep it relevant.

Data visualization has the potential to solve many business issues. All businesses must incorporate data visualization tools and reap transformative benefits in their critical areas of operations.

3. Explain in detail the types of Data Visualization with a neat sketch. Also illustrate with examples of data visualization.

Pie Chart

A common, but limited, visualization used to show how a few dimensions compare to one another and the whole.

Bar Chart

Used to categorize elements based on size. Can be ordered or unordered based on the nature of the dimensions.

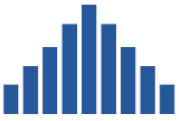
Line Chart

Best used to show trends across time intervals. Multiple lines can be used to compare categories within a dimension.

Treemaps

A treemap breaks the whole into its parts using a quantitative measure to determine the size of each square.

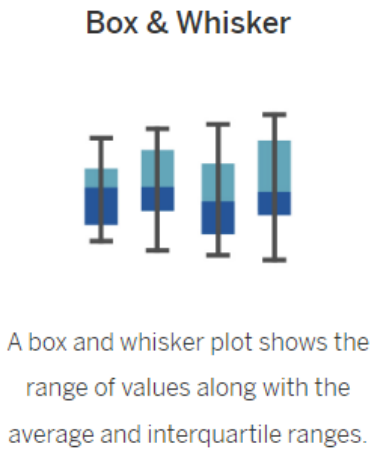
Histogram



Histograms split a single continuous measure into bins, or groups, to analyze distribution.

Bullet Graph

A bullet graph takes a bar chart and adds additional elements to create a compact way to analyze performance against a goal or threshold.



Common general types of data visualization:

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

1. Bar Graphs

Bar graphs include data in the form of multiple categories, we can either make individual graphs for each and every category or keep it in a single form through including multiple bars as one for each category at each time label. These bars could be assigned side by side or accumulated on top of each other. In order to make bar graph more effectively and easy to read, outline includes orders of the bars should be chronological, fix time frames label at one axis and label other quantities on other axes, data should not be placed from most to least or least to most but must be in chronology. If you want to analyze data over time or the data is assembled in multiple categories such as various industries, variety of food, and the progress of a company in the past 5 years, etc, a Bar Graph is the best choice.

2. **Line Chart**

Line graphs are also used for presenting data over time or classified data by category as bar graphs. The only difference is that line graphs allow for refinement. If you want to present data over very long time periods or continuously changing data, the line graph could be a solid choice to consider. Most of the time it happens, we clearly don't know how to fill data accurately in the

time duration for which data is available, in that condition we are drawing nothing other than a straight line. Though the rate of progress or decay between time duration is not linear up to a remarkable extent, so line graphs must be used very delicately to avoid malformation of data.

3. Pie chart

It is a presentation of data visualization in the circular form or circular chart. It is one of the most popular forms of data visualization, it can only be used when a smart portion of data add up to a whole. For example, 40 % of the marks are considered to pass in an exam, which could be displayed in the pie chart as it is indicating to 40 % out of the total 100 % of the marks. We can convert the percentage to proportions or proportions to the percentage for this aim, additionally, circle charts cannot be used to show an increase or decrease on their own. In case, if a pie chart could be used to present the data over time, there is a need to make a new chart for each time period and every measurement and display them together for comparison.

4. Quantagrams

The repeated pictogram or icon representation to show quantity is termed as Quantagrams, such that A very common example to show the multi-character quantities using Quantagrams is the number of people. You must have seen Quantagrams as classic male and female icons at the doors of the restroom. This technique is suitable for small numbers, small percentages or proportions. If we talk about pictograms, they are so simple and feel sound or reductive if they get used for any severe issues or a large quantity. It would appear as minimized if a severe issue is represented with simple sorted icons. We can opt Typography if we need to visualize data for large statistics.

5. Typography

It is limited to certain cases where it can be accepted as the best solution provider, it is not restricted to provide an old text-only solution, instead, it is intelligently used to achieve a successful and effective piece of content. The data would be fit for typography if it is large or greater than 100, never be a percentage of a whole or increase or decrease in percentage, and can’t be compared to another number. In order to improve typography visualization, it can be combined with a pictogram or icon that gives the viewer a clear visual picture with the context of the subject matter of data and numbers.

4.Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.

5.Area charts. This visualization method is a variation of a line chart; it displays multiple values in a time series -- or a sequence of data collected at consecutive, equally spaced points in time.

6. Scatter plots. This technique displays the relationship between two variables. A scatter plot takes the form of an x- and y-axis with dots to represent data points.

7. Treemaps. This method shows hierarchical data in a nested format. The size of the rectangles used for each category is proportional to its percentage of the whole. Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.

8. Population pyramids. This technique uses a stacked bar graph to display the complex social narrative of a population. It is best used when trying to display the distribution of a population.

Examples of data visualization

Common data visualization use cases

Common use cases for data visualization include the following:

Sales and marketing. Research from the media agency Magna predicts that half of all global advertising dollars will be spent online by 2020. As a result, marketing teams must pay close attention to their sources of web traffic and how their web properties generate revenue. Data visualization makes it easy to see traffic trends over time as a result of marketing efforts.

Politics. A common use of data visualization in politics is a geographic map that displays the party each state or district voted for.

Healthcare. Healthcare professionals frequently use choropleth maps to visualize important health data. A choropleth map displays divided geographical areas or regions that are assigned a certain color in relation to a numeric variable. Choropleth maps allow professionals to see how a variable, such as the mortality rate of heart disease, changes across specific territories.

Scientists. Scientific visualization, sometimes referred to in shorthand as SciVis, allows scientists and researchers to gain greater insight from their experimental data than ever before.

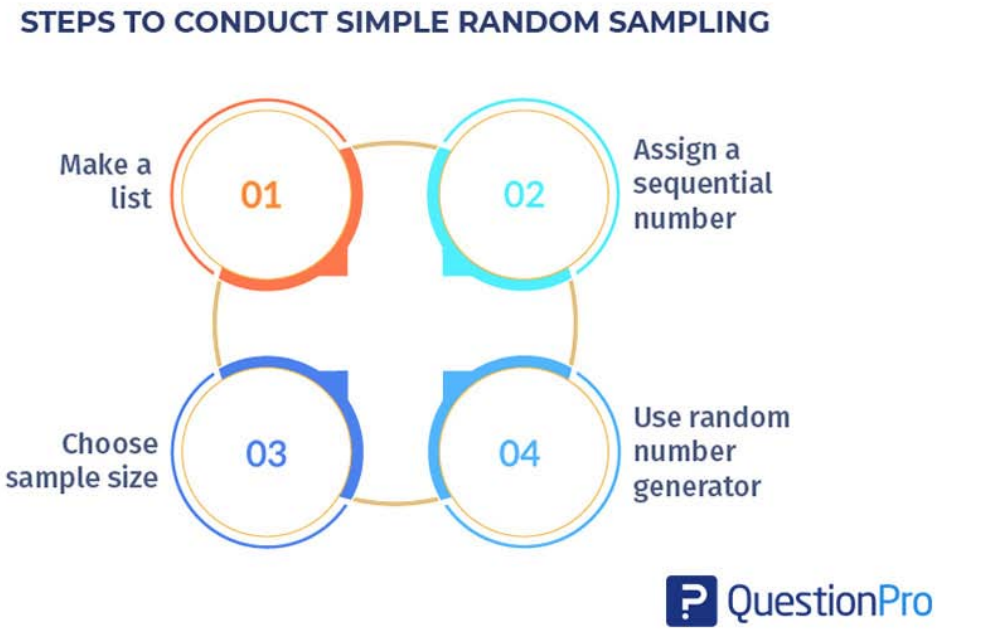
Finance. Finance professionals must track the performance of their investment decisions when choosing to buy or sell an asset. Candlestick charts are used as trading tools and help finance professionals analyze price movements over time, displaying important information, such as

securities, derivatives, currencies, stocks, bonds and commodities. By analyzing how the price has changed over time, data analysts and finance professionals can detect trends.

Logistics. Shipping companies can use visualization tools to determine the best global shipping routes.

Data scientists and researchers. Visualizations built by data scientists are typically for the scientist's own use, or for presenting the information to a select audience. The visual representations are built using visualization libraries of the chosen programming languages and tools. Data scientists and researchers frequently use open source programming languages -- such as Python -- or proprietary tools designed for complex data analysis. The data visualization performed by these data scientists and researchers helps them understand data sets and identify patterns and trends that would have otherwise gone unnoticed.

3. Explain the steps to conduct simple random sampling.



Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample. Here the selection of items entirely depends on luck or probability, and therefore this sampling technique is also sometimes known as a method of chances.

Simple random sampling is a fundamental sampling method and can easily be a component of a more complex sampling method. The main attribute of this sampling method is that every sample has the same probability of being chosen.

Researchers follow these methods to select a simple random sample:

1. They prepare a list of all the population members initially, and then each member is marked with a specific number (for example, there are nth members, then they will be numbered from 1 to N).
2. From this population, researchers choose random samples using two ways: random number tables and random number generator software. Researchers prefer a random number generator software, as no human interference is necessary to generate samples.

Two approaches aim to minimize any biases in the process of simple random sampling:

- **Method of lottery**

Using the lottery method is one of the oldest ways and is a mechanical example of random sampling. In this method, the researcher gives each member of the population a number. Researchers draw numbers from the box randomly to choose samples.

- **Use of random numbers**

The use of random numbers is an alternative method that also involves numbering the population. The use of a number table similar to the one below can help with this sampling technique.

Random Number Table

1	69	24	40	68	29	39	95	60	30
97	23	70	59	79	4	47	19	38	20
13	44	5	71	12	99	78	34	9	96
34	55	83	21	72	3	37	85	61	2
22	80	18	82	54	32	84	16	46	88
7	43	6	48	11	92	63	53	86	28
56	90	36	91	64	45	15	73	10	87
49	65	50	14	51	33	89	52	74	57
98	17	100	58	5	8	77	25	62	31
27	76	66	81	26	93	41	94	67	42

Simple random sampling formula

Consider a hospital has 1000 staff members, and they need to allocate a night shift to 100 members. All their names will be put in a bucket to be randomly selected. Since each person has

an equal chance of being selected, and since we know the population size (N) and sample size (n), the calculation can be as follows:

$$P=1- N-1/N.N-2/N-1....N-n/N-(n-1)$$

$$\text{Cancelling}=1-N-n/N$$

$$=n/N$$

$$=100/1000$$

$$=10\%$$

4. **Discuss simple random sampling with example. Also explain the advantages of simple random sampling**

Follow these steps to extract a simple random sample of 100 employees out of 500.

1. **Make a list** of all the employees working in the organization. (as mentioned above there are 500 employees in the organization, the record must contain 500 names).
2. **Assign a sequential number** to each employee (1,2,3…n). This is your sampling frame (the list from which you draw your simple random sample).
3. **Figure out what your sample size is going to be**. (In this case, the sample size is 100).
4. **Use a random number generator** to select the sample, using your sampling frame (population size) from Step 2 and your sample size from Step 3. For example, if your sample size is 100 and your population is 500, generate 100 random numbers between 1 and 500.

Advantages of simple random sampling

1. It is a fair method of sampling, and if applied appropriately, it helps to reduce any bias involved compared to any other sampling method involved.
2. Since it involves a large sample frame, it is usually easy to pick a smaller sample size from the existing larger population.
3. The person conducting the research doesn't need to have prior knowledge of the data he/ she is collecting. One can ask a question to gather the researcher need not be a subject expert.
4. This sampling method is a fundamental method of collecting the data. You don't need any technical knowledge. You only require essential listening and recording skills.
5. Since the population size is vast in this type of sampling method, there is no restriction on the sample size that the researcher needs to create. From a larger population, you can get a small sample quite quickly.

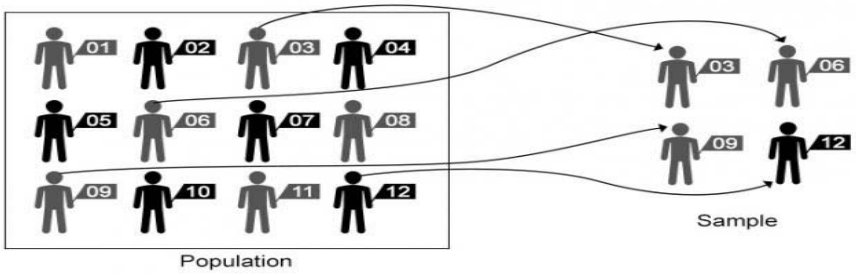
6. The data collected through this sampling method is well informed; more the samples better is the quality of the data.

6.Describe various sample methods in detail.

Sampling is an important strategy of handling large data. If data files are too big to be run by software or just too large to work with, the number of items in the data file can be sampled to provide a new data file that seeks to accurately represent the population from which it comes. In sampling data, there are three components that should be recognized: a population, a sample, and a sample element (the items that make up the sample). A firm's collection of customer service performance documents for one year could be designated as a population of customer service performance for that year. From that population, a sample of a lesser number of sample elements (the individual customer service documents) can be drawn to reduce the effort of working with the larger data. Several sampling methods can be used to arrive at a representative sample. Some of these sampling methods are presented in Table

Sampling Method	Description	Application	Application Notes
Simple Random	Allows each sample element in a population to have an equal chance of selection.	Selecting customers based on their percentage of occurrence as a member of a particular race.	Sample size must be sufficient to avoid sampling bias.
Systematic Random (or Period)	Selects sample elements from a population based on a fixed number in an interval.	Selecting every fifth person leaving an airport to interview.	Assumes the sample elements order in the interval is presented in a random fashion; otherwise, it can result in sampling bias.
Stratified Random	Stage 1: Divide a population into groups (called strata); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people in each of three different economic strata.	Strata must be representative of the population, or it can result in sampling bias.
Cluster Random	Stage 1: Group sample elements geographically (called clusters); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people from voting districts.	Cluster must be representative of the population, or it can result in sampling bias.
Quota	Based on a fixed quota or number of sample elements.	Selecting the first 200 people who enter a store.	• Mainly used to save time and money. • Sample size must be sufficient to avoid sampling bias.
Judgment	Selects sample elements based on expert judgment.	Selecting candidates for an interview with a special offer based on their appearance.	Prone to bias without defined criteria for selection because of dependency on interviewer experience.

7.Discuss Systematic Random Sampling in detail.



Systematic Random Sampling

Steps in selecting a systematic random sample:

- Calculate the sampling interval (the number of households in the population divided by the number of households needed for the sample)
- Select a random start between 1 and sampling interval
- Repeatedly add sampling interval to select subsequent households

Example of systematic random sampling of 10 households from a list of 40 households

- We first calculate the sampling interval by dividing the total number of households in the population (40) by the number we want in the sample (10). In this case, the sampling is 4. We then select a number between 1 and the sampling interval from the random number table (in this case 3). Household #3 is the first household. We then count down the list starting with household #3 and select each 4^{th} household. For example, the second selected household is 3 + 4, or #7. Note that when you reach the end of the list, you should have selected your desired number of households. If you have not, you have counted wrong or miscalculated the sampling interval. You should go back and start over.

  This is what your final selection should look like:

Households

1	Iwais M	21 Awalum E
2	Teiaba C	22 Douna B
3	Illarau G	23 Guba V
4	Aaron L	24 Posou P
5	Oksen V	25 Huape H
6	Pinpin B	26 Makaen J
7	Illabong P	27 Veia D
8	Ngatia T	28 Hainapa B
9	Gunure N	29 Narakine G
10	Agua C	30 Haung E
11	Illogina W	31 Tito III
12	Aulakua W	32 Pasi N
13	Wasoraba C	33 Kapua K
14	Zonggonau M	34 Tulia R
15	Tobena III	35 Koddy F
16	Mabong F	36 Tulia E
17	Yaman H	37 Tripp K
18	Bagita J	38 Bowen B
19	Bariia III	39 Temple V
20	Harekin J	40 Bowen B

Random number table

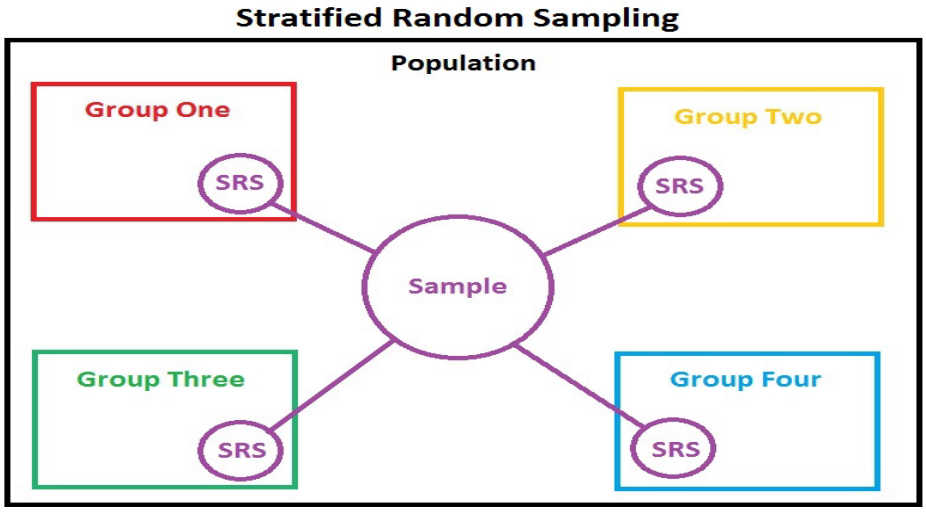
3647	2352	6959	1937	2554	6804	9098	4316
4318	2346	7276	1880	7136	9603	0163	3152
7000	2865	8357	4475	9804	0042	1106	7949
2932	9958	9582	2235	1140	1164	7841	1688
4097	8995	5030	1785	5420	0125	4953	1332
5540	6278	1584	4392	3258	1374	1617	7427
3320	8788	7658	9615	9862	7960	8140	6807
8077	2065	2560	2091	8921	0970	3134	8441

Selected households

1	Marau G
2	Mabong P
3	Mogin W
4	Tobena M
5	Baria M
6	Guba V
7	Veia D
8	Tito M
9	Koddy F
10	Temple V

7. Explain stratified random sampling.

Stratified random sampling is a type of probability sampling using which a research organization can branch off the entire population into multiple non-overlapping, homogeneous groups (strata) and randomly choose final members from the various strata for research which reduces cost and improves efficiency. Members in each of these groups should be distinct so that every member of all groups get equal opportunity to be selected using simple probability. This sampling method is also called "random quota sampling".



Age, socioeconomic divisions, nationality, religion, educational achievements and other such classifications fall under stratified random sampling.

Let's consider a situation where a research team is seeking opinions about religion amongst various age groups. Instead of collecting feedback from 326,044,985 U.S citizens, random samples of around 10000 can be selected for research. These 10000 citizens can be divided into strata according to age,i.e, groups of 18-29, 30-39, 40-49, 50-59, and 60 and above. Each stratum will have distinct members and number of members.

8 Steps to select a stratified random sample:

1. Define the target audience.
2. Recognize the stratification variable or variables and figure out the number of strata to be used. These stratification variables should be in line with the objective of the research. Every additional information decides the stratification variables. For instance, if the objective of research to understand all the subgroups, the variables will be related to the subgroups and all the information regarding these subgroups will impact the variables. Ideally, no more than 4-6 stratification variables and no more than 6 strata should be used in a sample because an increase in stratification variables will increase the chances of some variables canceling out the impact of other variables.
3. Use an already existent sampling frame or create a frame that's inclusive of all the information of the stratification variable for all the elements in the target audience.
4. Make changes after evaluating the sampling frame on the basis of lack of coverage, over-coverage, or grouping.
5. Considering the entire population, each stratum should be unique and should cover each and every member of the population. Within the stratum, the differences should be minimum

whereas each stratum should be extremely different from one another. Each element of the population should belong to just one stratum.

6. Assign a random, unique number to each element.
7. Figure out the size of each stratum according to your requirement. The numerical distribution amongst all the elements in all the strata will determine the type of sampling to be implemented. It can either be proportional or disproportional stratified sampling.
8. The researcher can then select random elements from each stratum to form the sample. Minimum one element must be chosen from each stratum so that there's representation from every stratum but if two elements from each stratum are selected, to easily calculate the error margins of the calculation of collected data.

9. Discuss the types of stratified random sampling.

Types of Stratified Random Sampling:

a. Proportionate Stratified Random Sampling:

In this approach, each stratum sample size is directly proportional to the population size of the entire population of strata. That means each strata sample has the same sampling fraction.

Proportionate Stratified Random Sampling Formula: $n_h = (N_h / N) * n$

$\mathbf{n_h}$= Sample size for h^{th} stratum

$\mathbf{N_h}$= Population size for h^{th} stratum

N = Size of entire population

n = Size of entire sample

If you have 4 strata with 500, 1000, 1500, 2000 respective sizes and the research organization selects ½ as sampling fraction. A researcher has to then select 250, 500, 750, 1000 members from the respective stratum.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/2	1/2	1/2
Final Sampling Size Results	250	500	750	1000

Irrespective of the sample size of the population, the sampling fraction will remain uniform across all the strata.

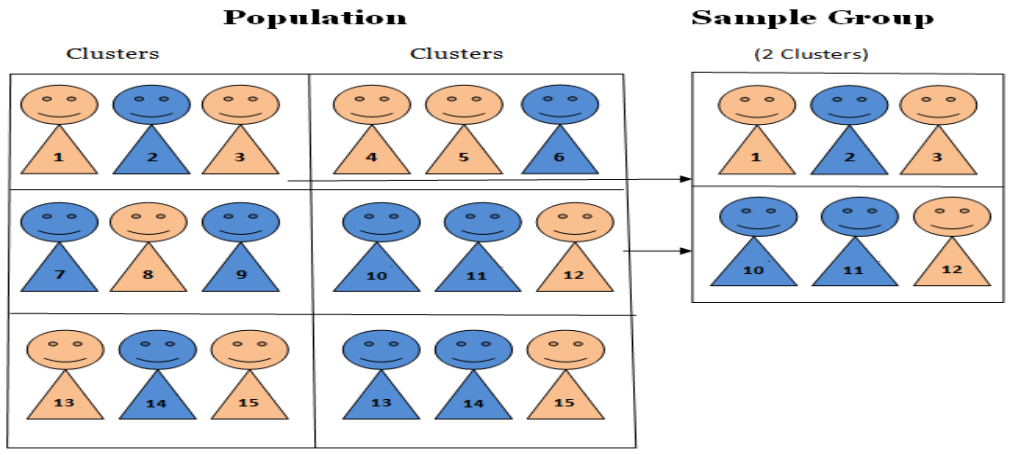
b. Disproportionate Stratified Random Sampling:

Sampling fraction is the primary differentiating factor between the proportionate and disproportionate stratified random sampling. In disproportionate sampling, each stratum will have a different sampling fraction. The success of this sampling method depends on the researcher's precision at fraction allocation. If the allotted fractions aren't accurate, the results may be biased due to the overrepresented or underrepresented strata.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/3	1/4	1/5
Final Sampling Size Results	250	333	375	400

10.Explain cluster sampling in detail .

Cluster sampling is a probability sampling technique where researchers divide the population into multiple groups (clusters) for research. Researchers then select random groups with a simple random or systematic random sampling technique for data collection and data analysis.



Steps to conduct cluster sampling

1. **Sample:** Decide the target audience and also the sample size.
2. **Create and evaluate sampling frames:** Create a sampling frame by using either an existing framework or creating a new one for the target audience. Evaluate frameworks based

on coverage and clustering and make adjustments accordingly. These groups will be varied, considering the population, which can be exclusive and comprehensive. Members of a sample are selected individually.

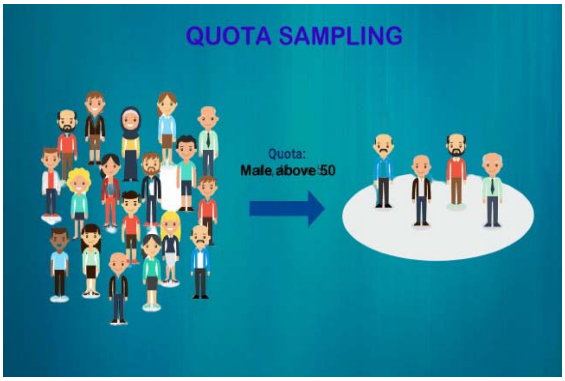
3. **Determine groups:** Determine the number of groups by including the same average members in each group. Make sure each of these groups are distinct from one another.
4. **Select clusters:** Choose clusters by applying a random selection.
5. **Create sub-types:** It is bifurcated into two-stage and multi-stage subtypes based on the number of steps followed by researchers to form clusters.

Applications of cluster sampling

This sampling technique is used in an area or geographical cluster sampling for market research. A broad geographic area can be expensive to survey in comparison to surveys that are sent to clusters that are divided based on region. The sample numbers have to be increased to achieve accurate results, but the cost savings involved make this process of rising clusters attainable.

11. Discuss quota sampling in detail.

The sample obtained from a quota sampling method contains similar proportions of observations as the whole population with some known traits or characteristics. In quota sampling, the researcher selects from his/her judgment or some fixed quota. In other words, the sample observations are to be chosen based on some pre-specified virtues. Then the total sample contains the same distribution of characteristics that were assumed to be found in the population of concern".



Types of Quota Sampling

The quota sampling is classified into two different types, such as:

- Controlled Quota Sampling
- Uncontrolled Quota Sampling

Controlled Quota Sampling:

If the sampling imposes restrictions on the researcher's/Statisticians choice of sample, then it is known as controlled quota sampling. In this method, the researcher can be able to select the limited samples.

Uncontrolled Quota Sampling:

If the sampling does not impose any restrictions on the researcher's/Statisticians choice of sample, then it is known as uncontrolled quota sampling. In this process, the researcher can select the samples of their interest.

Quota Sampling Process

The steps for the process of quota sampling are listed below:

Step 1: Firstly, the population should be divided into mutually exclusive subgroups

Step 2: Next is to determine the proportions of the subgroups made in the previous step, since the same proportion would be used during the sampling process

Step 3: Now, the researcher has to choose individual observations or items from the subgroups. Also, the proportions found previously are to be taken into consideration

Step 4: The last and final step is to ensure that the obtained sample should be representative of the whole population and allow the researcher to study characteristics and traits.

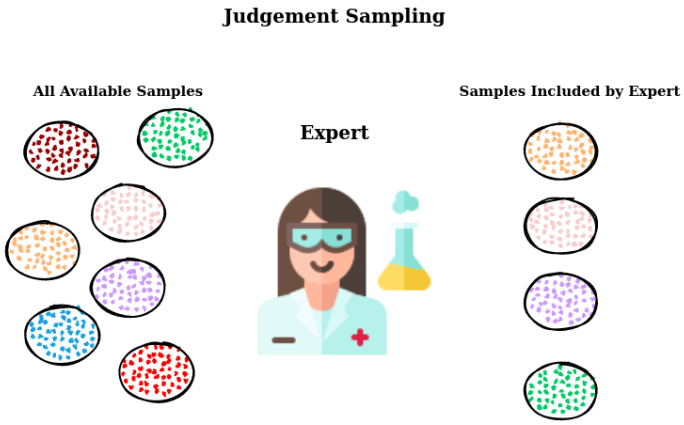
Uses of Quota Sampling

Some of the essential uses of quota sampling are as follows:

- In quota sampling, the judgment of sampling is done on the basis of convenience or a fixed quota. It is used when the decision of the researcher is final for choosing the samples
- The quota sampling is used because it allows the researchers to select a subgroup as a sample that is of great interest to study
- This method is used when the researcher wants to determine the relationship between the subgroups
- The quota sampling method should be selected when there is a limited time

12.Explain Judgment Sampling in detail.

This type of sampling, also known as purposive sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research. **Judgmental sampling is a non-probability sampling technique where the researcher selects units to be sampled based on their knowledge and professional judgment.** Purposive sampling is used in cases where the specialty of an authority can select a more representative sample that can bring more accurate results than by using other probability sampling techniques. The process involves nothing but purposely handpicking individuals from the population based on the authority's or the researcher's knowledge and judgment.



Example of Judgmental Sampling

In a study wherein a researcher wants to know what it takes to graduate summa cum laude in college, the only people who can give the researcher first hand advise are the individuals who graduated summa cum laude. With this very specific and very limited pool of individuals that can be considered as a subject, the researcher must use judgmental sampling.

When to Use Judgmental Sampling

Judgmental sampling design is usually used when a limited number of individuals possess the trait of interest. It is the only viable sampling technique in obtaining information from a very specific group of people. It is also possible to use judgmental sampling if the researcher knows a reliable professional or authority that he thinks is capable of assembling a representative sample.

Setbacks of Judgmental Sampling

The two main weaknesses of authoritative sampling are with the authority and in the sampling process; both of which pertains to the reliability and the bias that accompanies the sampling technique. Unfortunately, there is usually no way to evaluate the reliability of the expert or the authority. The best way to avoid sampling error brought by the expert is to choose the best and most experienced authority in the field of interest. When it comes to the sampling process, it is usually biased since no randomization was used in obtaining the sample. It is also worth noting that the members of the population did not have equal chances of being selected. The consequence of this is the misrepresentation of the entire population which will then limit generalizations of the results of the study.

13.Explain **Linear Programming Problem/Model Formulation Procedure in detail.**

Stepwise Procedure

The hardest part of figuring out any word problem or any real-world problem is always the first step. This stepwise procedure is a strategy for handling any kind of LP model. Big or small, it handles them all by breaking a complex process into small, achievable steps:

Determine the type of problem—A problem has to be either maximization or minimization. If the problem only mentions making profit or sales, it is most likely a maximization problem. If the problem only mentions cost, it most likely is a minimization problem. What if a problem includes sales and cost information? Then subtract the cost from the sales and derive profit.

Maximizing profit both maximizes sales and minimizes cost. The values that can be used to determine the type of problem are called the contribution coefficients.

Define the decision variables—Step 1 determined the type of problem by finding profit or cost contribution coefficients. The number of profit or cost contribution coefficients determines the number of decision variables because these contribution coefficients are attached to the respective decision variables in the objective function. There are two things to remember in defining decision variables: (1) Make clear what the decision variables are determining; (2) State any "time horizon" the problem is requiring. In the Ford Motor Company example

Suppose one wants to decide how many automobiles a Ford Motor Company plant should produce in a week. The plant is capable of producing only two types of automobiles: Mustangs and Thunderbirds. So the decision variables in this LP model will be as follows: X1 = number of Mustangs to produce per week X2 = number of Thunderbirds to produce per week The plant would not produce automobiles unless it could make some profit from the endeavor. Suppose it could make $1,000 on each Mustang and $3,500 on each Thunderbird. These values (1,000 and 3,500) represent the per unit of automobile profit contribution to what will be the total profit (Z) and are the contributions coefficients c1 and c2 in the model. The resulting objective function for this problem would be this: Maximize: Z = c1 X1 + c2 X2 (generalized form) Maximize: Z = 1000 X1 + 3500 X2 (applied form) If this objective function had no constraints to limit the size of the decision variables, they could be set at positive infinity to make as much profit as possible. Unfortunately, in the real world, there are always constraints to limit the optimization effort.

X1 = number of Mustangs to produce per week This definition makes clear that the "number of Mustangs" will be produced. The definition also includes the time horizon of one week. An example of what is not acceptable in the definition of a decision variable is this: X1 = Mustangs

3. Formulate the objective function—Because the contribution coefficients, the type of problem in Step 1, and the decision variables in Step 2 have been identified, all that is left is to combine these into the form of an objective function.

4. Formulate the constraints— (1) Right-hand-side strategy: Look at the problem for a sentence or a column in a table that lists the available resources that the model

needs to achieve. These are the right-hand-side "b" parameters. Create a column vector (a column of numbers) that will represent the "b" values in the model. Then go back and read the problem again to find the technology coefficients to finish the left-handside of the constraint. (2) Left-hand-side strategy: In problems with tabled values, look to see if they are technology coefficients. Take the technology coefficients and align them by row or column to form

the left-hand-side of the constraints. Then go back and read the problem again to find the right-hand-side values.

State the nonnegativity and given requirements —They are designed for beginners but will prep anyone in developing LP models.

PART-C

LP Problem/Model Formulation Practice: Butcher Problem

Problem Statement: Consider the problem of a butcher mixing the day's supply of meatloaf. The butcher has two grades of meatloaf: Grade 1 and Grade 2. The butcher needs to know how many trays of each kind of meatloaf should be made. The butcher may make whole trays or any fractional number of trays. The butcher's profit is increased by $36 for each tray of Grade 1 that is mixed, and by $34 for each tray of Grade 2. If there were no constraints, the butcher would want to make both kinds of meatloaf to maximize profit. Unfortunately, the butcher has constraints that must be considered.

• **Constraint 1—**The butcher cannot sell more than six trays of meatloaf per day.

• **Constraint 2—**Only nine hours of mixing time are available for the butcher and staff. It takes two hours to mix a tray of Grade 1 and one hour to mix a tray of Grade 2.

• **Constraint 3—**The butcher has only 16 feet of shelf space for meatloaf. Each tray of Grade 1 requires 2 feet of shelf space. Each tray of Grade 2 requires 3 feet of shelf space.

• **Formulation—**This problem clearly labels the constraints to make things easy. Remember to use the five-step formulation procedure to reduce a problem to easier and smaller steps to create the model.

1. Determine the type of problem—This problem only mentions profit, so it has to be a maximization problem. The two sales maximizing contribution coefficients ($36 and $34) in this model determine the type of problem.

2. Define the decision variables—The problem says, "The butcher needs to know how many trays of each kind of meatloaf should be made." That is one hint. An easier one is in Step 1. The two contribution coefficients mean two decision variables. Because $36 is the amount of profit on a tray of Grade 1 meatloaf, the related decision variable has to be the "number of trays of Grade 1 meatloaf to make or mix per day." Note the time horizon in the sentence, "Consider the problem of a butcher mixing the day's supply of meatloaf." The resulting two decision variables can be defined as follows: X1 = number of trays of Grade 1 meatloaf to make (or mix) per day X2 = number of trays of Grade 2 meatloaf to make (or mix) per day

3. Formulate the objective function—The formulation of the objective function follows easily from Steps 1 and 2. It is: Maximize: Z = 36 X1 + 34 X2

4. Formulate the constraints—Take one constraint at a time. In reading the sentence for Constraint 1 ("The butcher cannot sell more than six trays of meatloaf per day"), six is a parameter in the constraint. It has to be either a technology coefficient or a right-hand-side value. If it is a technology coefficient, it has to be directly related to an individual decision variable. If it is a right-hand-side value, it must be a total available resource. "Six" is a selling limitation on total trays, not individual trays. So it is a right-hand-side or b value. Because it is also a total maximum selling limitation, the direction of the inequality will be less than or equal to. What about the left-hand-side of this constraint? Well, what is the sum of all the trays of meatloaf? That can be expressed as the sum of both decision variables, resulting in the first constraint of the model that

follows: X1 + X2 ≤ 6 (selling) It is recommended that beginning LP modelers label their constraints with a word or two so that the modeler can remember that the particular limitation as been included as a constraint. It will also be helpful to others wanting to understand the model if the constraints are labeled with understandable terms. In Constraint 2, the sentences are, "Only nine hours of mixing time are available for the butcher and staff. It takes two hours to mix a tray of Grade 1 and one hour to mix a tray of Grade 2." In the first sentence, "nine" is a total available mixing time limitation. So it is a righthand-side value that represents a total maximum amount of this mixing resource resulting in a less than or equal to expression. In the second sentence "two" is attached to the Grade 1 decision variable, and "one" is attached to the Grade 2 decision variable. So these two parameters are technology coefficients. The resulting constraint follows: 2X1 + X2 ≥ 9 (mixing time) In Constraint 3, the sentences are, "The butcher has only 16 feet of shelf space for meatloaf. Each tray of Grade 1 requires 2 feet of shelf space, and each tray of Grade 2requires 3 feet of shelf space." In the first sentence, "16" is the total available shelf space limitation. So it is a right-hand-side value that represents a total maximum amount of this shelf space resource resulting in a less than or equal to expression. In the second sentence, the "2" is attached to the Grade 1 decision variable, and the "3" is attached to the Grade 2 decision variable. So these two parameters are technology coefficients. The resulting constraint is: 2X1 + 3X2 ≤16 (shelf space) **5. State the nonnegativity and given requirements—**Because the model has only two decision variables and the problem specifically allows fractional values, all that is needed is to state the same nonnegative requirements as the basic generalized model presented in Section B.3.4 as here: and X1, X2 ≥ 0

The entire formulation of the butcher problem is again presented here:

Maximize: $Z = 36X_1 + 34X_2$

subject to:

$X_1 + X_2 \leq 6$ (selling)

$2X_1 + X_2 \leq 9$ (mixing time)

$2X_1 + 3X_2 \leq 16$ (shelf space)

and $X_1, X_2 \geq 0$

UNIT-4

PART- A

1. What is predictive analytics?

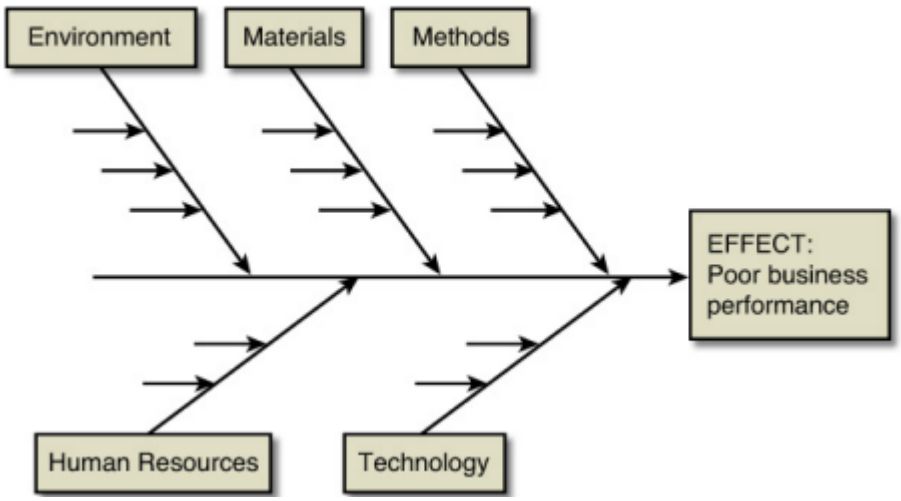
Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities.

2. What is logic driven model?

A logic-driven model is one based on experience, knowledge, and logical relationships of variables and constants connected to the desired business performance outcome situation.

3. What is cause and effect diagram?

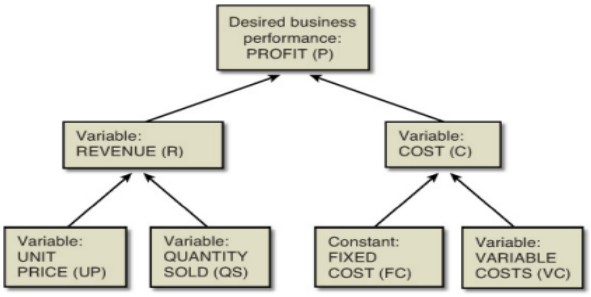
The cause-and-effect diagram is a visual aid diagram that permits a user to hypothesize relationships between potential causes of an outcome.



This diagram lists potential causes in terms of human, technology, policy, and process resources in an effort to establish some basic relationships that impact business performance.

4.What is influence diagram?

According to Evans (2013, pp. 228– 229), influence diagrams can be useful to conceptualize the relationships of variables in the development of models.



5 . list the variables that define profit.

Profit = Revenue− Cost, or Profit = (Unit Price × Quantity Sold) − [(Fixed Cost) + (Variable Cost × Quantity Sold)], or P = (UP × QS) − [FC + (VC × QS)]

6. data-driven models (using data collected from many sources to quantitatively establish model relationships).

7. List data driven techniques.

Data-Driven Models	Possible Applications
Sampling and Estimation	Generate statistical confidence intervals to define limitations and boundaries on future forecasts for other forecasting models.
Regression Analysis	(1) Create a predictive equation useful for forecasting time series forecasts. (2) Weed out predictive variables in forecasting models that add little to predicting values. (3) Generate a trend line for forecasting.
Correlation Analysis	(1) Assess variable relationships. (2) Weed out predictive variables in forecasting models that add little to predicting values.
Probability Distributions	(1) Estimate trend behavior that follows certain types of probability distributions. (2) Conduct statistical tests to confirm significance of variables.
Predictive Modeling and Analysis	Fit linear and nonlinear models to data to use the models for forecasting.
Forecasting Models	Those listed in this table and others such as smoothing models can be used to forecast values.
Simulation	Project future behavior in variables by simulating the past behavior found in probability distributions.

8. what is data mining?

data mining is a discovery-driven software application process that provides insights into business data by finding hidden patterns and relationships in big or small data and inferring rules from them to predict future behavior. These observed patterns and rules guide decision□making.

9. **What Is Clustering in Data Mining?**

Clustering refers to the process of grouping a series of different data points based on their characteristics. By doing so, data miners can seamlessly divide the data into subsets, allowing for more informed decisions in terms of broad demographics (such as consumers or users) and their respective behaviors.

10. What Is Association in Data Mining?

Data miners use association to discover unique or interesting relationships between variables in databases. Association is often employed to help companies determine marketing research and strategy.

11. What Is Data Cleaning in Data Mining?

Data cleaning involves organizing data, eliminating duplicate or corrupted data, and filling in any null values. When this process is complete, the most useful information can be harvested for analysis.

12. What Is Data Visualization in Data Mining?

Data can be presented in visual ways through charts, graphs, maps, diagrams, and more. This is a primary way in which data scientists display their findings.

13. What Is Classification in Data Mining?

In data mining, classification is considered to be a form of clustering — that is, it is useful for extracting comparable points of data for comparative analysis. Classification is also used to designate broad groups within a demographic, target audience, or user base through which businesses can gain stronger insights.

14. What Is Machine Learning in Data Mining?

In data mining, machine learning's applications are vast. Machine learning and data mining fall under the umbrella of data science but aren't interchangeable terms. For instance, computers perform data mining as part of their machine learning functions.

15. What Are Neural Networks in Data Mining?

Neural networks combine many computer processors (similar to the way the brain uses neurons) to process data, make decisions, and learn as a human would — or at least as closely as possible.

Neural networks are used to find associations where connections between words or numbers can be determined. Specifically, neural networks can take large volumes of data and potential variables and explore variable associations to express a beginning variable (referred to as an input layer), through middle layers of interacting variables, and finally to an ending variable (referred to as an output).

16. What Is Outlier Detection in Data Mining?

While other data mining methods seek to identify patterns and trends, outlier detection looks for the unique: the data point or points that differ from the rest or diverge from the overall sample. Outlier detection finds errors, such as data that was input incorrectly or extracted from the wrong sample. Natural data deviations can be instructive as well.

17.What Is Prediction in Data Mining?

Predictive modeling is among the most common uses of data mining and works best with large data sets that represent a broad sample size.

18. What Is Data Warehousing in Data Mining?

Data miners collect data from multiple sources into a common archive before it can be used in business analysis. This process, called data warehousing, typically occurs before the data mining process.

19. List the two functions of Neural network in SPSS.

SPSS has two versions of neural network software functions: Multilayer Perceptron (MLP) and Radial Basis Function (RBF). Both procedures produce a predictive model for one or more dependent variables based on the values of the predictive variables. Both allow a decision maker to develop, train, and use the software to identify particular traits (such as bad loan risks for a bank) based on characteristics from data collected on past customers).

20. What is discriminant analysis?

The analysis generates a regression function whereby values of the independent variables can be incorporated to generate a predicted value for the dependent variable.

21. what is logistic regression?

Logistic regression is **a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set**. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

22. what is hierarchical clustering?

Hierarchical clustering, also known as hierarchical cluster analysis, is **an algorithm that groups similar objects into groups called clusters**. The endpoint is a set of clusters, where each cluster

is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

23. what is agglomerative strategy?
The agglomerative strategy is a bottom-up approach, where one starts with each item in the data and begins to group them.

24. what is divisive stragegy?
The divisive strategy is a top-down approach, where one starts with all the items in one group and divides the group into clusters.

25. what is k-mean clustering?
K-mean clustering is a classification methodology that permits a set of data to be reclassified into K groups, where K can be set as the number of groups desired. The algorithmic process identifies initial candidates for the K groups and then interactively searches other candidates in the data set to be averaged into a mean value that represents a particular K group. The process of selection is based on maximizing the distance from the initial K candidates selected in the initial run through the list. Each run or iteration through the data set allows the software to select further candidates for each group.

Part- B

1. Explain predictive analytics process

1. **Define the requirements.** Understand the business problem you're trying to solve. Is it managing inventory? Reducing fraud? Predicting sales? Generating questions about the problem and listing them in order of importance is a good start. Collaborating with a statistician at this stage can help form metrics for measuring success. A business user or subject matter expert generally takes charge of this first step.

2. **Explore the data.** Here, you'll want to loop in a statistician or data analyst or both. The job is to identify the data that informs the problem you're trying to solve and the goal. Consider the relevancy, suitability, quality and cleanliness of the data.

3. **Develop the model.** A data scientist can help figure out which predictive models are best suited to solving the problem. It's important to experiment with different features, algorithms and processes in order to strike a balance between performance, accuracy and other requirements, such as explainability.

4. **Deploy the model.** Once the model is approved by the data scientist, a data engineer determines how best to retrieve, clean and transform the required raw data to deploy the model at scale and, above all, in a way that makes a meaningful difference -- e.g., integrating a new scoring algorithm into the sales team's workflow.

5. **Validate the results.** Performance of the model can change over time due to shifts in customer preferences or the business climate, or unforeseen events such as a pandemic. Thresholds for updating models vary, requiring the joint expertise of a business user and a data scientist in this step.

2. Discuss the real time applications of predictive analytics in business?

Examples of predictive analytics applications span business functions and industries. Moreover, as the technology becomes more accurate, easier to use and less expensive, the uses and benefits of predictive analytics will increase, reported technology journalist Maria Korolov in her updated look at the top predictive analytic use cases in business.

Here is a sampling of how businesses are applying predictive analytics.

Marketing. The use of predictive analytics in marketing has transformed how companies sell to customers. Technology writer Mary K. Pratt reported in her article on how to drive success in marketing with predictive analytics that the variety of use cases include next best action, lead qualification, proactive churn management, demand forecasting and "data-driven creatives" -- the use of predictive analytics to help decide what media style and form of messaging will resonate best with certain customers.

Supply chain management. The COVID-19 pandemic highlighted an increased need for better statistical models and forecasting in supply chain management. The pandemic forced companies to throw "historical data out the window," research analyst Alexander Wurm told Korolov, and to update their processes with real-time data and third-party information. For example, IoT-

generated real-time data alerts companies to goods that have gone bad or been otherwise damaged, increasing the usefulness of predictive analytics in rapidly changing environments.

Fraud detection. The latest global crime survey from PricewaterhouseCoopers found that fraud rates are at record highs, costing companies worldwide a staggering $42 billion over the past two years. Many companies have small teams of investigators, making predictive technology essential to getting a handle on fraud. Predictive analytics is being used to scan through hundreds of thousands of insurance claims and refer just the ones that are most likely to be fraudulent to investigative teams. It's also being used by retailers to authenticate customers as they log in and monitor them for suspicious behavior as it happens.

Healthcare. As Pratt reported in her compendium of 12 ways predictive analytics provides value in healthcare, its use is both challenging and expected to increase in this field. Predicting the likelihood of patients developing certain medical conditions and forecasting the progression of diseases in patients are big uses, involving data stores from electronic health records, federal repositories, biometric data, claims data and more. Health administration also benefits from predictive analytics, which is used to identify patients at high risk for hospital readmission, optimize resource allocations and manage supply chains, among other applications.

Predictive maintenance and monitoring. IoT data is being used in predictive modeling to forecast equipment breakdowns. Manufacturers attach sensors to machinery on the factory floor and to mechatronic products, such as automobiles; the sensor data is then used to forecast when maintenance and repair work should be done in order to prevent problems. Predictive analytics is also used for monitoring oil and gas pipelines, drilling rigs, windmill farms and various other industrial IoT installations. Localized weather forecasts for farmers based partly on data collected from sensor-equipped weather data stations installed in farm fields is another IoT-driven predictive modeling application.

3. Discuss data mining methodologies in detail

Data Mining Techniques

1. Clustering

Clustering is a technique used to represent data visually — such as in graphs that show buying trends or sales demographics for a particular product.

What Is Clustering in Data Mining?

Clustering refers to the process of grouping a series of different data points based on their characteristics. By doing so, data miners can seamlessly divide the data into subsets, allowing for more informed decisions in terms of broad demographics (such as consumers or users) and their respective behaviors.

Methods for Data Clustering

- **Partitioning method:** This involves dividing a data set into a group of specific clusters for evaluation based on the criteria of each individual cluster. In this method, data points belong to just one group or cluster.
- **Hierarchical method:** With the hierarchical method, data points are a single cluster, which are grouped based on similarities. These newly created clusters can then be analyzed separately from each other.
- **Density-based method:** A machine learning method where data points plotted together are further analyzed, but data points by themselves are labeled "noise" and discarded.
- **Grid-based method:** This involves dividing data into cells on a grid, which then can be clustered by individual cells rather than by the entire database. As a result, grid-based clustering hase a fast processing time.
- **Model-based method:** In this method, models are created for each data cluster to locate the best data to fit that particular model.

Examples of Clustering in Business

Clustering helps businesses manage their data more effectively. For example, retailers can use clustering models to determine which customers buy particular products, on which days, and with what frequency. This can help retailers target products and services to customers in a specific demographic or region.

Clustering can help grocery stores group products by a variety of characteristics (brand, size, cost, flavor, etc.) and better understand their sales tendencies. It can also help car insurance companies that want to identify a set of customers who typically have high annual claims in order to price policies more effectively. In addition, banks and financial institutions might use clustering to better understand how customers use in-person versus virtual services to better plan branch hours and staffing.

2. Association

Association rules are used to find correlations, or associations, between points in a data set.

What Is Association in Data Mining?

Data miners use association to discover unique or interesting relationships between variables in databases. Association is often employed to help companies determine marketing research and strategy.

Methods for Data Mining Association

Two primary approaches using association in data mining are the single-dimensional and multi-dimensional methods.

- **Single-dimensional association:** This involves looking for one repeating instance of a data point or attribute. For instance, a retailer might search its database for the instances a particular product was purchased.
- **Multi-dimensional association:** This involves looking for more than one data point in a data set. That same retailer might want to know more information than what a customer purchased — such as their age, method of purchase (cash or credit card), or age.

Examples of Association in Business

The analysis of impromptu shopping behavior is an example of association — that is, retailers notice in data studies that parents shopping for childcare supplies are more likely to purchase specialty food or beverage items for themselves during the same trip. These purchases can be analyzed through statistical association.

Association analysis carries many other uses in business. For retailers, it's particularly helpful in making purchasing suggestions. For example, if a customer buys a smartphone, tablet, or video game device, association analysis can recommend related items like cables, applicable software, and protective cases.

Additionally, association is used by the government to employ census data and plan for public services; it is also used by doctors to diagnose various illnesses and conditions more effectively.

3. Data Cleaning

Data cleaning is the process of preparing data to be mined.

What Is Data Cleaning in Data Mining?

Data cleaning involves organizing data, eliminating duplicate or corrupted data, and filling in any null values. When this process is complete, the most useful information can be harvested for analysis.

Methods for Data Cleaning

- **Verifying the data:** This involves checking that each data point in the data set is in the proper format (e.g, telephone numbers, social security numbers).
- **Converting data types**: This ensures data is uniform across the data set. For instance, numeric variables only contain numbers, while string variables can contain letters, numbers, and characters.
- **Removing irrelevant data:** This clears useless or inapplicable data so full emphasis can be placed on necessary data points.
- **Eliminating duplicate data points:** This helps speed up the mining process by boosting efficiency and reducing errors.
- **Removing errors:** This eliminates typing mistakes, spelling errors, and input errors that could negatively affect analysis outcomes.

- **Completing missing values:** This provides an estimated value for all data and reduces missing values, which can lead to skewed or incorrect results.

Examples of Data Cleaning in Business

According to Experian, 95 percent of businesses say they have been impacted by poor data quality. Working with incorrect data wastes time and resources, increases analysis costs (because models need to be repeated), and often leads to faulty analytics.

Ultimately, no matter how great their models or algorithms are, businesses suffer when their data is incorrect, incomplete, or corrupted.

4. Data Visualization

Data visualization is the translation of data into graphic form to illustrate its meaning to business stakeholders.

What Is Data Visualization in Data Mining?

Data can be presented in visual ways through charts, graphs, maps, diagrams, and more. This is a primary way in which data scientists display their findings.

Methods for Data Visualization

Many methods exist for representing data visually. Here are a few:

- **Comparison charts:** Charts and tables express relationships in the data, such as monthly product sales over a one-year period.
- **Maps:** Data maps are used to visualize data pertaining to specific geographic locations. Through maps, data can be used to show population density and changes; compare populations of neighboring states, counties, and countries; detect how populations are spread over geographic regions; and compare characteristics in one region to those in other regions.
- **Heat maps:** This is a popular visualization technique that represents data through different colors and shading to indicate patterns and ranges in the data. It can be used to track everything from a region's temperature changes to its food and pop culture trends.
- **Density plots:** These visualizations track data over a period of time, creating what can look like a mountain range. Density plots make it easy to represent occurrences of single events over time (e.g., month, year, decade).
- **Histograms:** These are similar to density plots but are represented by bars on a graph instead of a linear form.
- **Network diagrams:** These diagrams show how data points relate to each other by using a series of lines (or links) to connect objects together.
- **Scatter plots:** These graphs represent data point relationships on a two-variable axis. Scatter plots can be used to compare unique variables such as a country's life expectancy or the amount of money spent on healthcare annually.
- **Word clouds:** These graphics are used to highlight specific word or phrase instances appearing in a body of text; the larger the word's size in the cloud, the more frequent its use.

Examples of Data Visualization in Business

Representing data visually is an important skill because it makes data readily understandable to executives, clients, and customers. According to Markets and Markets, the market size for global data visualization tools is expected to nearly double (to $10.2 billion) by 2026.

Companies can make faster, more informed decisions when presented with data that is easy to understand and interpret. Today, this is typically accomplished through effective, visually accessible mediums such as graphs, 3D models, and even augmented reality. As a result, it's a good idea for aspiring data professionals to consider learning such skills through a data science and visualization bootcamp.

5. Classification

Classification is a fundamental technique in data mining and can be applied to nearly every industry. It is a process in which data points from large data sets are assigned to categories based on how they're being used.

What Is Classification in Data Mining?

In data mining, classification is considered to be a form of clustering — that is, it is useful for extracting comparable points of data for comparative analysis. Classification is also used to designate broad groups within a demographic, target audience, or user base through which businesses can gain stronger insights.

Methods for Data Mining Classification

- **Logistic regression:** This algorithm attempts to show the probability of a specific outcome within two possible results. For example, an email service can use logistic regression to predict whether or not an email is spam.
- **Decision trees:** Once data is classified, follow-up questions can be asked, and the results diagrammed into a chart called a decision tree. For example, if a computer company wants to predict the likelihood of laptop purchases, it may ask, *Is the potential buyer a student?* The data is classified into "Yes" and "No" decision trees, with other questions to be asked afterward in a similar fashion.
- **K-nearest neighbors (KNN):** This is an algorithm that tries to identify an unknown object by comparing it to others. For instance, grocery chains might use the K-nearest neighbors algorithm to decide whether to include a sushi or hot meals station in their new store layout based on consumer habits in the local marketplace.
- **Naive Bayes:** Based on the Bayes Theorem of Probability, this algorithm uses historical data to predict whether similar events will occur based on a different set of data.
- **Support Vector Machine (SVM):** This machine learning algorithm is often used to define the line that best divides a data set into two classes. An SVM can help classify images and is used in facial and handwriting recognition software.

Examples of Classification in Business

Financial institutions classify consumers based on many variables to market new loans or project credit card risks. Meanwhile, weather apps classify data to project snowfall totals and other

similar figures. Grocery stores also use classification to group products by the consumers who buy them, helping forecast buying patterns.

6. Machine Learning

Machine learning is the process by which computers use algorithms to learn on their own. An increasingly relevant part of modern technology, machine learning makes computers "smarter" by teaching them how to perform tasks based on the data they have gathered.

What Is Machine Learning in Data Mining?

In data mining, machine learning's applications are vast. Machine learning and data mining fall under the umbrella of data science but aren't interchangeable terms. For instance, computers perform data mining as part of their machine learning functions.

Methods for Machine Learning

- **Supervised learning:** In this method, algorithms train machines to learn using pre-labeled data with correct values, which the machines then classify on their own. It's called supervised because the process trains (or "supervises") computers to classify data and predict outcomes. Supervised machine learning is used in data mining classification.
- **Unsupervised learning:** When computers handle unlabeled data, they engage in unsupervised learning. In this case, the computer classifies the data itself and then looks for patterns on its own. Unsupervised models are used to perform clustering and association.
- **Semi-supervised learning:** Semi-supervised learning uses a combination of labeled and unlabeled data, making it a hybrid of the above models.
- **Reinforcement learning:** This is a more layered process in which computers learn to make decisions based on examining data in a specific environment. For example, a computer might learn to play chess by examining data from thousands of games played online.

Examples of Machine Learning in Business

With machine learning, companies can use computers to quickly identify all sorts of data patterns (in sales, product usage, buying habits, etc.) and develop business plans using those insights. This is a growing need in many industries.

According to a MicroStrategy survey, 18 percent of analytics professionals said machine learning and AI will have the most significant impact on their strategies over the next five years. Learning more advanced topics like machine learning is thus becoming imperative for data scientists.

7. Neural Networks

Computers process large amounts of data much faster than human brains but don't yet have the capacity to apply common sense and imagination in working with the data. Neural networks are one way to help computers reason more like humans.

What Are Neural Networks in Data Mining?

Artificial neural networks attempt to digitally mimic the way the human brain operates. Neural networks combine many computer processors (similar to the way the brain uses neurons) to process data, make decisions, and learn as a human would — or at least as closely as possible.

Neural Network Methods

Neural networks consist of three main layers: input, "hidden," and output. Data enters through the input layer, is processed in the hidden layer, and is resolved in the output layer where any relevant action based on the data is then taken. The hidden layer can consist of many processing layers, depending on the amount of data being used and learning taking place.

Supervised and unsupervised learning also apply to neural networks; neural networks use these types of algorithms to "train" themselves to function in ways similar to the human brain.

Examples of Neural Networks in Business

Neural networks have a wide range of applications. They can help businesses predict consumer buying patterns and focus marketing campaigns on specific demographics. They can also help retailers make accurate sales forecasts and understand how to use dynamic pricing. Furthermore, they help to improve diagnostic and treatment methods in healthcare, improving care and performance.

8. Outlier Detection

Outlier detection is a key component of maintaining safe databases. Companies use it to test for fraudulent transactions, such as abnormal credit card usage that might suggest theft.

What Is Outlier Detection in Data Mining?

While other data mining methods seek to identify patterns and trends, outlier detection looks for the unique: the data point or points that differ from the rest or diverge from the overall sample. Outlier detection finds errors, such as data that was input incorrectly or extracted from the wrong sample. Natural data deviations can be instructive as well.

Methods for Outlier Detection

- **Numeric outlier:** Outliers are detected based on the Interquartile Range, or the middle 50 percent of values. Data points outside that range are considered outliers.
- **Z-score:** The Z-Score denotes how many standard deviations a data point is from the sample's mean. This is also known as extreme value analysis.
- **DBSCAN:** This stands for "density-based spatial clustering of applications with noise" and is a method that defines data as core points, border points, and noise points, which are the outliers.
- **Isolation forest:** This method isolates anomalies in large sets of data (the forest) with an algorithm that searches for those anomalies instead of profiling normal data points.

Examples of Outlier Detection in Business

Almost every business can benefit from understanding anomalies in their production or distribution lines and how to fix them. Retailers can use outlier detection to learn why their

stores witness an odd increase in purchases, such as snow shovels being bought in the summer, and how to respond to such findings.

Generally, outlier detection is employed to enhance logistics, instill a culture of preemptive damage control, and create a smoother environment for customers, users, and other key groups.

9. Prediction

Predictive modeling seeks to turn data into a projection of future action or behavior. These models examine data sets to find patterns and trends, then calculate the probabilities of a future outcome.

What Is Prediction in Data Mining?

Predictive modeling is among the most common uses of data mining and works best with large data sets that represent a broad sample size.

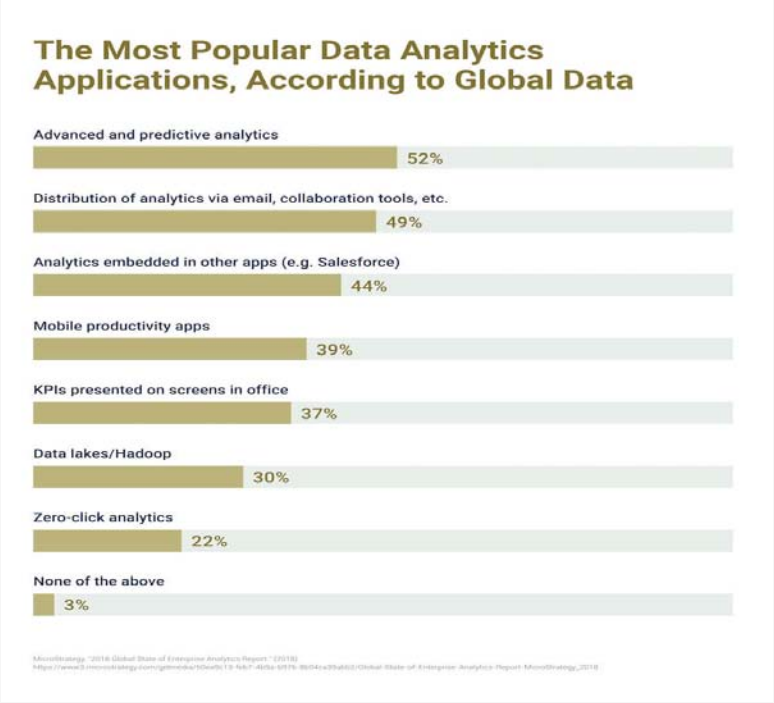
Methods for Prediction

Predictive modeling uses some of the same techniques and terminology as other data mining processes. Here are four examples:

- **Forecast modeling:** This is a common technique in which the computer answers a question (for instance, *How much milk should a store have in stock on Monday?*) by analyzing historical data.
- **Classification modeling:** Classification places data into groups where it can be used to answer direct questions.
- **Cluster modeling:** By clustering data into groups with shared characteristics, a predictive model can be used to study those data sets and make decisions.
- **Time series modeling:** This model analyzes data based on when the data was input. A study of sales trends over a year is an example of time series modeling.

Examples of Prediction in Business

Predictive modeling is a business imperative that impacts nearly every corner of the public and private sectors. According to MicroStrategy, 52 percent of global businesses consider advanced and predictive modeling their top priority in analytics.



Predictive models can be built to determine sales projections and predict consumer buying habits. They help manufacturers forecast distribution needs and determine maintenance schedules. Government agencies use census data to map population trends and project spending needs while baseball teams use predictive models to determine contracts and build rosters.

10. Data Warehousing

Data warehousing is the process by which data is collected and stored before it is evaluated.

What Is Data Warehousing in Data Mining?

Data miners collect data from multiple sources into a common archive before it can be used in business analysis. This process, called data warehousing, typically occurs before the data mining process.

Methods for Data Warehousing

Data goes through a three-stage process known as ETL before being loaded into a data warehouse. ETL stands for extract, transform, and load:

- **Extract:** Data is copied and moved from its source to a warehouse staging area. Data can be structured (names, dates, credit card numbers, etc.) or unstructured (photos, videos, audio files, social media posts).
- **Transform:** In this step, the data is filtered and cleaned — errors are removed and the data is validated. The data is also formatted to fit the warehouse.
- **Load:** In the final step, the transformed data is uploaded to the data warehouse. These steps can be repeated as data is updated.

Examples of Data Warehousing in Business

Data warehouses make working with big data easier — particularly for businesses that deal with large customer bases, sales and billing reports, and resource plans. Through data warehousing, businesses can segment and target customers from vast collections of sales orders, product searches, or loyalty program registrations. They also can store and analyze a wide variety of data points, even social media posts about products and businesses.

Data warehousing also consolidates various data sources into one place, making mining and decision-making more efficient and saving businesses time and money.

UNIT 5
PRESCRITIVE ANALYTICS

Introduction to Prescriptive analytics - Prescriptive Modeling - Non Linear Optimisation - Demonstrating Business Performance Improvement.

PART-A

1.What is prescriptive analytics?

It involves the application of decision science, management science, or operations research methodologies to make best use of allocable resources.

2.List the methodologies of prescriptive analytics.

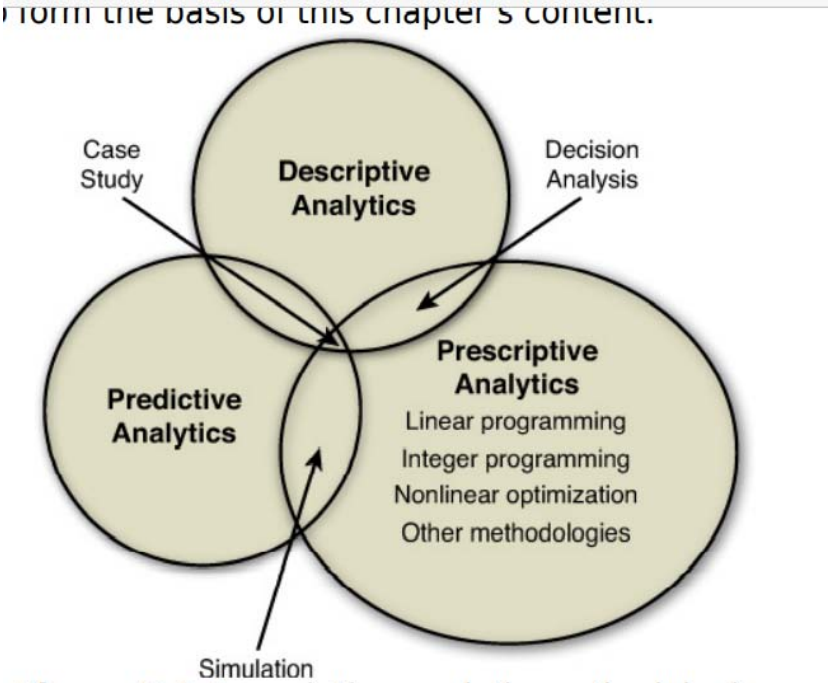


Figure 7.1 Prescriptive analytic methodologies

3.What is linear programming?

linear programming, mathematical modeling technique in which a linear function is maximized or minimized when subjected to various constraints. This technique has been useful for guiding

quantitative decisions in business planning, in industrial engineering, and—to a lesser extent—in the social and physical sciences.

4. list the example of linear programming in real life?

Examples include allocating advertising budgets to differing media, allocating human and technology resources to product production, and optimizing blends of mixing ingredients to minimize costs of food products.

5. What is Integer programming?

Integer programming expresses the optimization of a linear function subject to a set of linear constraints over integer variables. An integer programming (IP) problem is a linear programming (LP) problem in which the decision variables are further constrained to take integer values. Both the objective function and the constraints must be linear. The most commonly used method for solving an IP is the method of branch-and–bound.

6. Write the example of Integer programming in real life?

Examples include allocating stocks to portfolios, allocating personnel to jobs, and allocating types of crops to farm lands.

7. List the application of mixed integer programming?

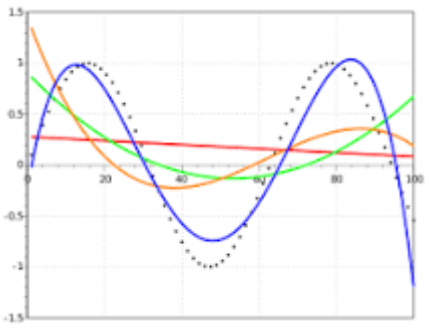
Mixed-integer programming has many applications in industrial productions, including job-shop modelling. One important example happens in agricultural production planning involves determining production yield for several crops that can share resources (e.g. Land, labor, capital, seeds, fertilizer, etc.)

8 What is nonlinear optimization?
In nonlinear optimization, nonlinear constraints. of the decision variable are used. If the possible solution space is bounded by nonlinear constraints then the method used to find possible solution is called non-linear programming (NLP).

9. What is decision analysis?

Decision analysis is a systematic, quantitative, and visual approach to making strategic business decisions. Decision analysis is a formalized approach to making optimal choices under conditions of uncertainty. It allows the user to enter costs, probabilities, and health-related quality of life values among other inputs of interest, and then calculates probabilistically weighted means of these outcome measures.

10. What is meant by curve fitting?



Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints.

11. What is simulation and example?

The definition of a simulation is a model or representative example of something. When you create a computer program that is intended to model flying a plane, this is an example of a simulation.

12. What is regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

13. What is ANOVA?

Analysis of variance, or ANOVA, is **a statistical method that separates observed variance data into different components to use for additional tests**. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

14.What is parametric hypothesis test?

When the statistic being compared is a population parameter based on a probability distribution, the hypothesis test is called a parametric hypothesis test.

15.What is nonparametric hypothesis test?
For hypothesis tests with measures that are not population parameters (based on counts or frequencies), use a nonparametric hypothesis test.

16. What does Ho implies?
H0 implies that there is no significant difference between the mean that we have from our sample (or some other measure of central tendency) and the population mean.

17. What does 9595 percent confidence indicate?
It indicates that a 5 percent chance the population mean is not in the interval. That 5 percent is called the level of significance, because it marks the boundary of where one designs a significant difference between the sample mean and a population mean.

18.What is multiple regression?
Multiple regression is used to develop a model when multiple independent variables might predict a dependent variable more accurately than the one independent variable simple regression model. It is not limited to time-series data but can be used to generate time series forecasts. It is ideal for sorting through possible predictive variables and determining those that should be used and those that should not be used in a forecasting model. The generalized model for a multiple regression model can be presented as such: Yp = a + b1 X1 + b2 X2 +... + bnXn where:
Yp= the forecasted or predicted value of the dependent
variables of trend a = vertical axis intercept value X
i = (for i
= 1,2, ..., n) different independent variables b
i = (for i=1,2,
..., n) the proportional contribution of the related independent variable to the forecast of Yp

PART- B

1. Discuss in detail Linear Programming Problem/Model Formulation procedure

Stepwise Procedure

1. Determine the type of problem—A problem has to be either maximization or minimization. If the problem only mentions making profit or sales, it is most likely a maximization problem. If the problem only mentions cost, it most likely is a minimization problem. What if a problem includes sales and cost information? Then subtract the cost from the sales and derive profit. Maximizing profit both maximizes sales and minimizes cost. The values that can be used to determine the type of problem are called the contribution coefficients.

2. Define the decision variables—Step 1 determined the type of problem by finding profit or cost contribution coefficients. The number of profit or cost contribution coefficients determines the number of decision variables because these contribution coefficients are attached to the respective decision variables in the objective function. There are two things to remember in defining decision variables: (1) Make clear what the decision variables are determining; (2) State any “time horizon” the problem is requiring.

3. Formulate the objective function— The objective function is generally expressed as one of the following:
Maximize: Z = c1 X1 + c2 X2 + ... + cn Xn

or
Minimize: Z = c1 X1 + c2 X2 + ... + cn Xn
where:
Z = an unknown that is not a variable but one that will be solved when the values of the decision variables are determined Xj = decision variables for j = 1, 2, ..., n; which are the unknowns to solve for an optimal value cj = contribution coefficients for j = 1, 2, ..., n; which represent the per-unit contribution to Z for each unit of the decision variable to which they are related

4. Formulate the constraints—

The constraints in an LP model can generally be expressed as the following: subject to:
a11 X1 + a12 X2 + ... + a1n Xn ≤ b1
a21 X1 + a22 X2 + ... + a2n Xn ≥ b2 ...
am1 X1 + am2 X2 + ... + amn Xn = bm
where:
bi = a right-hand-side value for i = 1, 2, ... m; and "m" is
the number of constraints in the model each having a righthand- side value usually representing a total resource
availability or requirement aij = technology coefficients for I = 1, 2, ..., m and j = 1, 2, ... n; which represent the per-unit usage of the related ith right-hand-side value by the related jth decision variable In the constraints, the technology coefficients (aij) are located by row with the first subscript and by column with the second subscript.

5. State the nonnegativity and given requirements

The decision variables in LP models are required to be zero or some positive value. As a formal part of the correct way of formulating an LP model (as is the case in most of the mathematical programming methods), one must add an additional statement in LP model formulations that looks like this: and X1, X2, ... , Xn ≥ 0

2. Explain Necessary Assumptions for Linear Programming Models.

Five basic assumptions must be met for LP to be used in a modeling a situation. These assumptions are also useful in deciding whether LP should be used to model a problem.
Here are the five assumptions:

1. Linearity—All constraints and the objective function must be linear. If one has a nonlinear profit or cost function or a nonlinear constraint, other nonlinear programming methodologies must be used. A number of these are available in the literature, including such techniques as quadratic programming, separable programming, and Kuhn-Tucker conditions.

2. Additivity—All the constraints and the objective function must for any value of the decision variables add up exactly as modeled. That is, one cannot have synergistic impact, where 2 + 2 = 5. Regardless of the size of the decision variable value, the added values of the coefficients must be the simple sum of the products. If they're not, use another methodology.

3. Divisibility—In the LP models presented in this appendix, the nonnegativity and given requirements allow the decision variable values to be real numbers or any fractional value. This means that if a decision variable ended up being 0.5, one-half of the profit or cost of that decision variable is exactly what will be received. Also, if the labor hour usage of that decision variable is two, then 0.5 of two means that exactly one hour of labor will be used. Sometimes fractional

answers are not realistic. In such cases, use something other than LP—perhaps integer programming.

4. Finiteness—This requirement simply means that the values of the decision variables must be finite. If they are not finite, they are infinite and, therefore, unbounded.

5. Certainty and a static time period—All of the a, b, and c parameters of an LP model must be known with certainty. We can help ensure this certainty by stating a time horizon or a static time period when the decision variables are defined. The static time period specifies the period over which the answer and the parameters remain true.

2.Solve LP Problem - Diet Problem

Problem Statement: A diet is to contain at least 10 ounces of nutrient P, 12 ounces of nutrient R, and 20 ounces of nutrient S. These nutrients are acquired from foods A and B. Each pound of A costs four cents and has four ounces of P, three ounces of R, and no S. Each pound of B costs seven cents and has one ounce of P, two ounces of R, and four ounces of S. Desiring minimum cost, how many pounds of each food should be purchased if the stated dietary requirements are to be met?

Formulation, by steps:

1. Determine the type of problem—This problem only mentions costs. Therefore, it must be a minimization problem.

2. Define the decision variables—How many cost values were used in Step 1 to determine the type of problem? Two (four cents and seven cents) are required. So how many decisions variables are needed? Two: If the four cents is the cost per pound of food A, the first decision variable follows: X1 = number of

pounds of food A to purchase Note that there is no time

horizon (day, week, and so on) in this problem. So do not put one in. The second decision variable is as follows: X2 = number of pounds of food B to purchase

3. Formulate the objective function—Note next that "cents" are being used. Some modelers might express the cents as 0.04 and 0.07, and others might express them as integer values. Note here that they are modeled as cents.

Minimize: $Z = 4X1 + 7X2$

4. Formulate the constraints—This problem illustrates how the "right-hand-side strategy" for formulating constraints might be helpful. Note in the first sentence, "A diet is to contain at least 10 ounces of nutrient P, 12 ounces of nutrient R, and 20 ounces of nutrient S," how the total minimum requirements are listed. These values create a column vector (10, 12, and 20), as presented in the right-hand-side values of the

constraints that follow: Note how the technology coefficients for food A (the X1 column) can be found in a single sentence, "Each pound of A costs four cents and has four ounces of P,

three ounces of R, and no S," and food B (the X2 column) can be found in a single sentence, "Each pound of B costs seven cents and has one ounce of P, two ounces of R, and four ounces of S." Because all the constraints had total minimum amounts of nutrients, the resulting expressions are all greater than or equal to.

5. State the nonnegativity and given requirements

Because the model has only two decision variables, all that is needed is to state the same nonnegative requirements as the basic generalized model and $X1, X2 \geq 0$ The entire formulation of the diet problem is again

presented here:

$$\text{Minimize: } Z = 4X_1 + 7X_2$$

$$\begin{aligned} \text{subject to:} \quad 4X_1 + X_2 &\geq 10 \text{ (nutrient P)} \\ 3X_1 + 2X_2 &\geq 12 \text{ (nutrient R)} \\ 4X_2 &\geq 20 \text{ (nutrient S)} \\ \text{and} \quad X_1, X_2 &\geq 0 \end{aligned}$$

3. **Solve LP Problem- Farming Problem**

Problem Statement: The Smith family owns 175 acres of farmland for breeding pigs and sheep. On average, it takes 0.5 acres of land to support either a pig or a sheep. The family can produce up to a total of 7,000 hours of labor for breeding. It takes 15 hours of labor to breed a pig and 20 hours of labor to breed a sheep. Although the family is willing to breed sheep, they do not want to breed more than 200 sheep at a time. Also, pig breeding is limited to 250.It is expected that each pig will contribute $300 profit, whereas each sheep will contribute $350.

Formulation, by steps:

1. Determine the type of problem—The problem only mentions profit, so it has to be a maximization problem.

2. Define the decision variables—The profit coefficients are attached to pigs and sheep, and there is no stated time horizon, so: X1 = number of pigs to breed X2 = number of sheep to breed

3. Formulate the objective function: Maximize: Z = 300X1 + 350X2

4. Formulate the constraints:

$$\begin{aligned} 0.5X_1 + 0.5X_2 &\leq 175 \text{ (acres of land)} \\ 15X_1 + 20X_2 &\leq 7000 \text{ (labor hours)} \\ X_1 &\leq 250 \text{ (maximum pig breeding)} \\ X_2 &\leq 200 \text{ (maximum sheep breeding)} \end{aligned}$$

5. State the nonnegativity and given requirements: and X1, X2 ≥ 0

4. Solve LP Problem Customer Service Problem

Problem Statement: The customer service department of a local department store provides repair services for merchandise sold. During one week, 5 television sets, 12 radios, and 18 electric percolators were returned for repair, representing overload work items. Two repair people are temporarily employed as part-time helpers to deal with the overload work. In a normal 8-hour workday, Person 1 can repair 1 television, 3 radios, and 3 electric percolators. In a normal 8-hour workday, Person 2 repairs 1 television, 2 radios, and 6 electric percolators. Person 1 makes $55 per day, and Person 2 makes $52 per day. The customer service department wants to minimize the total cost of operation, while maintaining good customer relationships. How many days should the two repair people be employed to handle the overload of work during this one week?

Solution

Formulation, by steps:

1. Determine the type of problem—The problem only mentions cost, so it has to be a minimization problem.

2. Define the decision variables—The cost coefficients are attached to Person 1 and Person 2. Now this is a "fuzzy" time horizon problem. Are these people being hired for a week? No! They are hired for some unknown number of days to process a week's overload. So the decision variables in this problem do not need a time horizon other than to say, specifically, they are handling the overload work. It can be written as follows:
X1 = number of days Person 1 should be hired to handle the overload work X2 = number of days Person 2 should be hired to handle the overload work
3. Formulate the objective function: Minimize: Z = 55X1 + 52X2
4. Formulate the constraints:

subject to: $X_1 + X_2 \geq 5$ (TVs)
$3X_1 + 2X_2 \geq 12$ (radios)
$3X_1 + 6X_2 \geq 18$ (percolators)

5. State the nonnegativity and given requirements: and
X1, X2 ≥ 0

5. Solve LP Problem - Model Clarke Special Parts Problem
Problem Statement: The Clarke Special Parts Company manufactures three products: A, B, and C. Three manufacturing centers are necessary for the production process. Product A only passes through Centers 1 and 2; Products B and C must pass through all three manufacturing centers. The time required in each center to produce one unit of each of the three products is noted as follows:

Product	Center 1	Center 2	Center 3
A	3 hours	2 hours	0 hours
B	1	2	2
C	2	1	3

So a unit of Product A takes three hours at Center 1, two hours at Center 2, and zero hours at Center 3. Each center is on a 40-hour week. The time available for production must be decreased by the necessary cleanup time. Center 1 requires four hours of cleanup, Center 2 requires seven hours, and Center 3 requires five hours. It is estimated that the profit contribution is $60 per unit of Product A, $40 per unit of Product B, and $30 per unit of Product C. How many units of each of these special parts should the company produce to obtain the maximum profit?
Formulation, by steps:
1. Determine the type of problem—The problem only mentions profit, so it has to be a maximization problem.
2. Define the decision variables—The profit coefficients are attached to Products A, B, and C and have a weekly stated time horizon, so: X1 = number of units of Product A to produce per week X2 = number of units of Product B to produce per week X3 = number of units of Product C to produce per week
3. Formulate the objective function: Maximize: Z = 60X1 + 40X2 + 30X3
4. Formulate the constraints—This problem illustrates that some arithmetic may be needed to derive model parameters. In this case, the right-hand-side b values need to be adjusted for the cleanup time. In a week, each department starts with 40 hours for production purposes. They

then have to be decreased for the cleanup time, as stated in the problem sentences, "The time available for production must be decreased by the necessary cleanup time. Center 1 requires four hours of cleanup, Center 2 requires seven hours, and Center 3 requires five hours." So, for Center 1 we have 36 hours (40 − 4), for Center 2 we have 33 hours (40− 7), and so on to formulate the right-hand-side values in each constraint. This problem also illustrates the use of the left-hand-side strategy for formulating constraints. Note how the tabled values are the technology coefficients listed by columns in the constraints that follow:

$$\begin{aligned} \text{subject to: } & 3X_1 + X_2 + 2X_3 \leq 36 \text{ (Center 1)} \\ & 2X_1 + 2X_2 + X_3 \leq 33 \text{ (Center 2)} \\ & 2X_2 + 3X_3 \leq 35 \text{ (Center 3)} \end{aligned}$$

5. State the nonnegativity and given requirements: and X1, X2, X3 ≥ 0

6. Solve LP Problem - Federal Division Problem

Problem Statement: The Federal Division has a contract to supply at least 72 engine parts. There are three different production processes for engine parts. The processes require different amounts of skilled labor, unskilled labor, and computer time for machine tools. Any one process is, by itself, capable of producing an engine part.

Process	Hours of Skilled Labor for One Engine Part	Hours of Unskilled Labor for One Engine Part	Computer Time in Minutes for One Engine
1	3	4	1
2	6	4	0
3	0	1	4

In the foreign country where they operate their plant, skilled labor costs $8.00/hour, and no more than 288 hours can be obtained. Unskilled labor costs $3.00/hour, and no more than 324 hours can be obtained. Computer time costs $10/minute, and no more than 196 minutes are available. Recommend a course of action.

Formulation, by steps:

1. Determine the type of problem—The problem only mentions cost, so it has to be a minimization problem.

2. Define the decision variables—This problem is meant to challenge and build skills in identifying decision variables. What is the variable here? In this problem, only one product, an engine part, is produced. So what is the variable? The variable in this problem is the "process" by which engine parts are produced. So the decision variables become this: X1 = number of engine parts to produce by Process 1 X2 = number of engine parts to produce by Process 2

X3 = number of engine parts to produce by Process 3 Like many firms today that have older technology to produce current products, this problem seeks to make the best use of a combination of old and new process technologies to produce the single product called engine parts.

3. Formulate the objective function—Given the definition of the preceding decision variables as processes, identify the correct contribution coefficients. Because they are directly related to the decision variable definitions, they can be defined as follows: c1 = cost of producing one engine

part by Process 1 c2 = cost of producing one engine part by Process 2 c3 = cost of producing one engine part by Process 3 How are these parameters found? Use a little arithmetic to compute them as follows:

Cost of Skilled Labor	Cost of Unskilled Labor	Cost of Computer Time
c_1 = \$8 × 3 hours +	\$3 × 4 hours +	\$10 × 1 minute = \$46
c_2 = \$8 × 6 hours +	\$3 × 4 hours +	\$10 × 0 minutes = \$60
c_3 = \$8 × 0 hours +	\$3 × 1 hour +	\$10 × 4 minutes = \$43

The resulting three parameters can then be put in an objective function as follows: Minimize: Z = 46X1 + 60X2 + 43X3

4. Formulate the constraints:

subject to:

$$X_1 + X_2 + X_3 \geq 72 \text{ (Supply)}$$
$$3X_1 + 6X_2 \leq 288 \text{ (Skilled labor)}$$
$$4X_1 + 4X_2 + X_3 \leq 324 \text{ (Unskilled labor)}$$
$$X_1 + 4X_3 \leq 196 \text{ (Computer time)}$$

4. State the nonnegativity and given requirements: and X1, X2, X3 ≥ 0

6. List the non-parametric test used for taking decisions?

Test Statistic	Application Area
Binomial Test	This test compares the observed frequencies of the two categories of a dichotomous variable to the frequencies that are expected under a binomial distribution with a specified probability parameter (α).
Chi-Square	This test (used often for a goodness-of-fit test) compares the observed and expected frequencies in each category of a distribution to test that all categories contain the same proportion of values or to test that each category contains a user-specified proportion of values.
Kolmogorov-Smirnov (One-Way)	Multiple versions of this test can be applied to differing comparative analyses. The one-way procedure compares the observed cumulative distribution function for a variable with any specified theoretical distribution.
Wilcoxon Signed-Rank	Multiple versions of this test can be applied to differing comparative analyses. This test is applicable when two samples (two populations) are related (not independent). The test is designed to compare some *n* matched pairs of ranked or ordinal data from two populations.
Run	A *run* is a sequence of like observations. This tests whether the order of occurrence of two values of a variable is random. A sample with too many or too few runs suggests that the sample is not random.

Table A.4 Common Nonparametric Statistical Tests and Software Access Information

7. Explain Discrete Probability distribution in detail.

Probability Distribution	Description	Application Examples
Geometric	Similar to the binomial in that there is a two-outcome situation of success r or failure, the p probability of success occurring, and q probability of failure of the event not occurring. Unlike the binomial, one is interested in the first success that will occur out of some unknown number of n trial periods. This distribution is particularly useful when the interest is in determining the probability of some events occurring in a specific number of discrete time periods (for example, in hours, days, minutes).	Examples of the use of this distribution can include determining the probability that a customer will wait two or more minutes in a queue for service; the probability that out of four people surveyed, one will buy a new product; and the probability that one will have to sample at least 20 tax returns before finding the first person who will need to be audited.
Hypergeometric	Similar to the binomial distribution in that it is a two-outcome situation of success r or failure, but with this distribution the trials are dependent. The dependence of the trials occurs because the results are taking place *without replacement*. In using this probability distribution, unlike all other distributions discussed so far, a total number or *population* of elements must be specified, denoted N, from which sampling takes place without replacement. This distribution is based on counting and does not require known probability parameters in its computation.	Examples of the use of this distribution include determining the probability of lot rejection based on lot size and sample size, the probability that a person has been discriminated against based on population size, and determining the proportion of voters that will vote for a specific political party's candidate based on a sample of registered voters.

8.Discuss the Limitations on the Use of Multiple Regression Models in Forecasting Time Series Data

The use of multiple regression models in time series forecasting has limitations. The application of the model to look beyond the ranges of its independent variables (like time) can violate the model's necessary assumptions previously stated for the simple regression model. Moreover, there are numerous other mathematical conditions that make forecasting with multiple independent variables risky at best. There are lag effects between independent variables that can falsely lead researchers to assume they have a fairly accurate model by bloating the correlation coefficients, when in fact the independent variables may only be correlating between themselves, not the dependent variable they seek to forecast. A common statistical test used in regression analysis is called the Durbin-Watson Autocorrelation Test, which tests for the lagged cause-and-effect relationship between the variables. It measures the residual errors when comparing

forecast values with actual values. Ideally, there should be no autocorrelation present. If there is autocorrelation present in the model, the relationship between the variablesin the model is not accurately expressed. The Durbin-Watson d-test computes a statistic d value (similar to a ttest or Z-test) that can range from 0 to 4. The closer the value is to 2, the less residual correlations (less autocorrelation) are assumed. The closer the value is to 0, the stronger the degree of positive correlations of the residuals. The closer the d-test statistic is to 4, the stronger the negative correlation of the residuals. As can be seen in Figure E.6, the Durbin-Watson d-test statistic is 2.837 for the multiple regression model. That suggests a slight negative correlation of residuals, but not much autocorrelation. The Durbin-Watson is just one of many tests that can be run to lend validity to the use of multiple regression.

8. Explain exponential smoothing model in detail.

An exponential smoothing model, like the name implies, smooths raw data to reveal nonlinear behavior. This smoothing is accomplished by mathematically weighting the inaccuracy of a prior forecast in an effort to generate a new forecast. In other words, exponential smoothing models allow the mathematical weighting of a prior inaccurate forecast in an effort to seek to improve and make a more accurate forecast in the future. Because the model is limited to forecasting only one time period into the future, it is viewed as a short-term forecasting model and can be useful in identifying cyclical and seasonal behavior in data. The formula for the simple exponential forecasting model follows: Ft = Ft -1 + α (At −1 − Ft − 1) where: Ft = exponential smoothed forecast value for the t time period Ft − 1 = forecast value for the t − 1 prior time period At − 1= actual value for the t − 1 prior time period α = an alpha weight ranging from 0 to 1 The values of F t − 1 and At − 1 for the first forecast value of F1, (F0 and A0, which are found by F1 − 1 and A1 − 1) are usually assumed values or an average of some prior set of data. The effect of these assumed F0 and A0 values will eventually be averaged out by this forecasting model, so the selection can be arbitrary if the number of forecast values n is significantly large. The number of periods t to use in a model at one time, as well as the value of α in the formula, is experimental and must be determined by selecting the best combination that minimizes forecasting error. This is usually accomplished by trial and error methods, where various values for the parameters are substituted into the model and results simulated, whereby a comparison can be made of accuracy statistics to find the best α to use in the formula or the number of time periods to run before the desired forecast value can be assumed to be generated.

How to Select Models and Parameters for Models

The selection of a model to forecast or a parameter in a model (for example, alpha or an independent variable in multiple regression) can be based on a number of criteria. The type of variation (linear or nonlinear) is one criterion. Other criteria include cost to develop a model, time it takes to develop a forecasting model, and time horizon of the forecast (long-term or short-term). The single, most important criterion for making a final selection of a model or a parameter in a model is forecasting accuracy. Although statistical methods like correlation and t-tests provide some measure of variable relationships and their potential to predict values, in forecasting, actual results are vital. Accuracy statistics can help make this selection decision. The most accurate model will be the one that generates the least forecasting error. Several statistics (MAD, MSE, and MAPE statistics) can be computed for any model once it has been developed.

In this way, differing models can be compared, and parameters can be selected for use. Following are the formulas for these commonly used forecast accuracy statistics. Experienced forecasters often use a simpler statistic called the mean absolute deviation (MAD). The formula for MAD follows:

$$MAD = \frac{\sum |A_t - F_t|}{n}$$

where:
At = actual value in time period t
Ft = forecast value for time period t
n = total number of t time periods that are being summed in the numerator The MAD statistic will be zero if the predictive model used to generate Ft perfectly predicts At.
As the error in forecasting increases, so will the MAD statistic value. When comparing the MADs from different models or forecasts based on differing parameters in a model, the smaller the MAD, the more accurate is the model. A similar statistic that seeks to minimize error in forecasting is the minimizing mean square error (MSE), using the same principles of standard error. Here's the formula for MSE:

$$MSE = \frac{\sum (A_t - F_t)^2}{n}$$

where:
At = actual value in time period t
Ft = forecast value for time period t
n = total number of t time periods that are being summed in the numerator Like the MAD statistic, the smaller the MSE, the more accurate the use of the parameter or model. Another useful error metric is mean absolute percentage error (MAPE). Mean square error and mean absolute error get larger with more observations and need to be compared with other measures of the same type. MAPE has the relative advantage in that it presents error in percentage form, making it possible to learn something about relative error immediately. Here's the formula for MAPE:

$$MAPE = \left(\frac{100}{n}\right) \sum \left| \frac{A_t - F_t}{A_t} \right|$$