Module IV

CHAPTER 4

Introduction to DS

University Prescribed Syllabus w.e.f Academic Year 2021-2022

Introduction and Evolution of Data Science, Data Science Vs. Business Analytics Vs. Big Data, Data Analytics, Lifecycle, Roles in Data Science Projects.

Self-Learning Topics : Applications and Case Studies of Data Science in various Industries

Teaching Hours – 04 **Approximate Weightage of Marks in Exam. – 5 Marks**

4.1 INTRODUCTION AND EVOLUTION OF DATA SCIENCE

4.1.1 Introduction to Data Science

- Data science is the mining procedure in which patterns and useful information is extracted from huge amount of raw data. Here raw data can be structured as well as unstructured.
- This is a field that encompasses a wide range of fields, and the basics of data science include mathematics, inference, computer science, forecasting statistics, the development of learning algorithms, and new technologies to gain understanding from big data.
- In Data science there are many steps and every step is equally important. We should always follow the proper steps to benefit from data science. Every step has its own important and it makes impact on the model. Let us understand these steps one by one.
- **Problem Statement:** Every work should start with motivation, Data science is no exception though. We should clearly and precisely state or declare the problem. Problem formulation is very important. Working of the data science model depends on the statement. Due to this most of the data scientist considers this as the main and much important step. So we should be sure about the problem statement and how well can it add value to business or any other organization.
- **Data Collection:** Once we declare the problem statement, the next step is data collection for model. In the data collection step data is acquired or extracted and then this data is given as input to the system. We must do good research, find all that we need. Data can be unstructured or structured. Data might be in any form like videos, spreadsheets, coded forms, etc. We should collect all these kinds of sources.
- **Data Cleaning:** Once we have formulated our motive and also collected data, the next step to do is cleaning. In Data cleaning step missing, redundant, unnecessary and duplicate data is eliminated from collection.
- There are various tools to do so. One can choose any of them. Various scientist have their opinion on which to choose. For the statistical part, R is preferred over Python as it contains more packages. Python is preferred in some cases as it is fast, easily accessible and we can perform the same things as we can in R with the help of various packages.
- **Data Analysis and Exploration:** In this step structure of the data is analyzed, hidden patterns are found and behavior is studied. Effects of one variable over others are visualized and then concluded. To explore the data various graphs formed with the help of libraries using any programming language are used. In R, ggplot is used whereas in Python matplotlib is used.
- **Data Modelling:** Once analysis is done with the help of data visualization, start building a hypothesis model such that it may yield a good prediction in future. In this step, we must choose a good algorithm that best fit to our model. There are various types of algorithms from regression to classification, Support vector machines, clustering, etc. Model is trained with the train data and then tested with test data. There are various methods to do so. One of them is the K-fold method where whole data is divided into two parts, one is Training data and the other is testing data.
- **Optimization and Deployment:** Model is built by following each and every step. Optimization is used to decide how well our model is performing? We test our data and find how well it is performing by checking its accuracy. In simple words, we check the efficiency of the data model and thus try to optimize it for better accurate prediction.
- Deployment means launch of model and let the people outside there to benefit from that. We can also obtain feedback from organizations and people to know their need and then to work more on model.

4.1.2 Evolution of Data Science

- Statistics, and the use of statistical models, are deeply rooted within the field of Data Science. Data Science started with statistics, and has evolved to include concepts/practices such as Artificial Intelligence, Machine Learning, and the Internet of Things, to name a few.
- As more and more data has become available, first by way of recorded shopping behaviors and trends, businesses have been collecting and storing it in ever

greater amounts. With growth of the Internet, the Internet of Things, and the exponential growth of data volumes available to enterprises, there has been a flood of new information or Big Data.

- Once the doors were opened by businesses seeking to increase profits and drive better decision making, the use of Big Data started being applied to other fields, such as medicine, engineering, and social sciences.
- A functional Data Scientist, as opposed to a general statistician, has a good understanding of software architecture and understands multiple programming languages. The Data Scientist defines the problem, identifies the key sources of information, and designs the framework for collecting and screening the needed data. Software is typically responsible for collecting, processing, and modeling the data. They use the principles of Data Science, and all the related sub-fields and practices encompassed within Data Science, to gain deeper insight into the data assets under review.
- There are many different dates and timelines that can be used to trace the slow growth of Data Science and its current impact on the Data Management industry, some of the more significant ones are outlined below.
- In 1962, John Tukey wrote about a shift in the world of statistics, saying, "… as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…I have come to feel that my central interest is in data analysis…" Tukey is referring to the merging of statistics and computers, at a time when statistical results were presented in hours, rather than the days or weeks it would take if done by hand.
- In 1974, Peter Naur authored the *Concise Survey of Computer Methods*, using the term "Data Science," repeatedly.
- Naur presented his own convoluted definition of the new concept: "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."
- In 1977, The IASC, also known as the International Association for Statistical Computing was formed. The first phrase of their mission statement reads, "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."
- In 1977, Tukey wrote a second paper, titled *Exploratory Data Analysis*, arguing the importance of using data in selecting "which" hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand.
- In 1989, the Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organized its first workshop.
- In 1994, Business Week ran the cover story, *Database Marketing*, revealing the ominous news companies had started gathering large amounts of personal information, with plans to start strange new marketing campaigns. The flood of data was, at best, confusing to company managers, who were trying to decide what to do with so much disconnected information.
- In 1999, Jacob Zahavi pointed out the need for new tools to handle the massive amounts of information available to businesses, in Mining Data for Nuggets of Knowledge.
- He wrote: "Scalability is a huge issue in data mining… Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements… Special data mining tools may have to be developed to address web-site decisions."
- In 2001, Software-as-a-Service (SaaS) was created. This was the pre-cursor to using Cloud-based applications.
- In 2001, William S. Cleveland laid out plans for training Data Scientists to meet the needs of the future. He presented an action plan titled, *Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics*. It described how to increase the technical experience and range of data analysts and specified six areas of study for university departments. It promoted developing specific resources for research in each of the six areas. His plan also applies to government and corporate research.

Module 4

- In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the *Data Science Journal*, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.
- In 2006, Hadoop 0.1.0, an open-source, non-relational database, was released. Hadoop was based on Nutch, another open-source database.
- In 2008, the title, "Data Scientist" became a buzzword, and eventually a part of the language. DJ Patil and Jeff Hammerbacher, of LinkedIn and Facebook, are given credit for initiating its use as a buzzword.
- In 2009, the term NoSQL was reintroduced (a variation had been used since 1998) by Johan Oskarsson, when he organized a discussion on "open-source, non-relational databases".
- In 2011, job listings for Data Scientists increased by 15,000%. There was also an increase in seminars and conferences devoted specifically to Data Science and Big Data. Data Science had proven itself to be a source of profits and had become a part of corporate culture.
- In 2011, James Dixon, CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses. Dixon stated the difference between a Data Warehouse and a Data Lake is that the Data Warehouse pre-categorizes the data at the point of entry, wasting time and energy, while a Data Lake accepts the information using a non-relational database (NoSQL) and does not categorize the data, but simply stores it.
- In 2013, IBM shared statistics showing 90% of the data in the world had been created within the last two years.
- In 2015, using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.
- In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for Artificial Intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.
- In the past ten years, Data Science has quietly grown to include businesses and organizations world-wide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of Big Data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed.
- Data Science has become an important part of business and academic research. Technically, this includes machine translation, robotics, speech recognition, the digital economy, and search engines. In terms of research areas, Data Science has expanded to include the biological sciences, health care, medical informatics, the humanities, and social sciences. Data Science now influences economics, governments, and business and finance.
- One result of the Data Science revolution has been a gradual shift to writing more and more conservative programming. It has been discovered Data Scientists can put too much time and energy into developing unnecessarily complex algorithms, when simpler ones work more effectively. As a consequence, dramatic "innovative" changes happen less and less often. Many Data Scientists now think wholesale revisions are simply too risky, and instead try to break ideas into smaller parts. Each part gets tested, and is then cautiously phased into the data flow.
- Though this play-it-safe philosophy may save companies time and money, and avoid major gaffes, they risk focusing on very narrow constraints, and avoid pursuing true breakthroughs.
- Scott Huffman, of Google, said: "One thing we spend a lot of time talking about is how we can guard against incrementalism when bigger changes are needed. It's tough, because these testing tools can really motivate the engineering team, but they also can wind up giving them huge incentives to try only small changes. We do want those little improvements, but we also want the jumps outside the box."

4.2 DATA SCIENCE VS. BUSINESS ANALYTICS VS. BIG DATA

4.2.1 Data Science vs. Business Analytics

- **Data Science:** Data Science is the complex study of the huge amount of data in an organization or company's repository. In data science study is done on various parameters like the origin of data, content of data, and the usefulness of this data for the growth of the company in the future.
- In any organization the data is always present in two types: Structured or unstructured. This data is studied to extract useful and important information about business or market patterns.
- This data will help the business as compared to the other competitors since they've increased their effectiveness by recognizing patterns in the data. Raw data is converted into critical business matters by the Data scientists.
- These scientists are specialized in algorithmic coding along with concepts like data mining, machine learning, and statistics.
- There are various applications that can benefit from Data Science like healthcare sector, fraud detection sector, internet search, airlines, and so on.

Business Analytics

- Business analytics and data science is slightly similar as both of them involve analyzing data. In Business analytics, we take it a step further and focus on the steps to be taken to positively affect the business after analyzing the data.
- Due to this, we can say, Business Analytics is the study of data in a way that we are able to make decisions for the business in the long run.
- Business analytics goal is to collect data from various business models and interpret it to solve a business goal or target.
- It is usually used to improve the company's overall performance in the market by strictly making business-focused decisions.

Sr. No.	Data Science	Business Analytics
1.	Data Science is the study of complex data using algorithms to find a pattern.	Business Analytics is the analysis of data to find business insights using statistics.
2.	In data science algorithms and pure code is used.	In Business Analytics statistical analysis and business is used.
3.	There are two types of data- structured and unstructured	Here usually data is taken from a business model (structured)
4.	This is relatively a new concept	Has been around since the 19th century
5.	It is the superset of business analytics	It is a part of data science
6.	Very vague and gives generic results	Gives business specific results
7.	Cost of investing is high	Cost of investing is low
8.	Can be used to enhance Machine Learning and Artificial Intelligence in the future	Can be used for Tax Analytics and Cognitive Analysis in the future

4.2.2 Data Science vs. Big Data

Data Science

- Data Science is a study which includes and involves working with a large amount of data and uses it for building predictive, prescriptive and prescriptive analytical models. In Data Science data is digged, captured, analyzed and utilized. It is a convergence of Data and computing. It is a mixing of the field of Computer Science, Business and Statistics together.
- **Big Data:** Big data has huge, large or voluminous data, information or the applicable statistics acquired by the large organizations and ventures. Many software and data storage created and prepared as it is difficult to compute the big data manually.
- It is used to discover patterns and trends and make decisions related to human behavior and interaction technology.

Module 4

Sr. No.	Data Science	Big Data
1.	Data Science is an area.	Big Data is a technique to collect, maintain and process the huge information.
2.	It is about collection, processing, analyzing and utilizing of data into various operations. It is more conceptual.	It is about extracting the vital and valuable information from huge amount of the data.
3.	It is a field of study just like the Computer Science, Applied Statistics or Applied Mathematics.	It is a technique of tracking and discovering of trends of complex data sets.
4.	The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
5.	Tools mostly used in Big Data includes Hadoop, Spark, Flink, etc.	Tools mainly used in Data Science includes SAS, R, Python, etc
6.	It is a super set of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics and many more techniques.	It is a sub set of Data Science as mining activities which is in a pipeline of the Data science.
7.	It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
8.	It broadly focuses on the science of the data.	It is more involved with the processes of handling voluminous data.

4.3 DATA ANALYTICS

4.3.1 Introduction to Data Analytics

In Data Analytics raw data is analysed to find trends and answer questions. Data Analytics includes many techniques with many different goals. The data analytics process comprises of some elements that can help a variety of initiatives. Combination of these elements leads to a successful data analytics initiative that will provide a clear picture of where you are, where you have been and where you should go.

- Mostly, this process starts with *descriptive analytics*. In this process historical trends in the data are described. The goal of Descriptive analytics is to answer the question "what happened?" This generally involves computing traditional indicators such as Return on Investment (ROI). For each industry, there will be different indicators. Descriptive analytics does not make predictions or directly inform decisions. In data analytics focus is on summarizing data in a meaningful and descriptive way.
- The next necessary element of data analytics is *advanced analytics*. This element of data science takes advantage of advanced tools to extract data, make predictions and discover trends. These tools include machine learning as well as classical statistics. Machine learning methods such as sentiment analysis, neural networks, natural language processing and more enable advanced analytics. This information gives new insight from data. Advanced analytics addresses "what if" questions.
- The accessibility of machine learning methods, enormous data sets, and inexpensive computing power has enabled the use of these methods in many industries. The collection of big data sets is instrumental in enabling these methods. Big data analytics is used in businesses to draw meaningful conclusions from complex and varied data sources, which has been made possible by advances in parallel processing and inexpensive computational power.

4.3.2 Types of Data Analytics

Data analytics is a wide field. There are four main types of data analytics: descriptive, diagnostic, predictive and prescriptive analytics. In the data analysis process each type has a different aim and a different place. These are also the main data analytics applications in business.

- Descriptive analytics is used to answer questions about what happened. These methods use large datasets to extract the outcomes to stakeholders. In this methods Key Performance Indicators (KPIs,) are generated to

...ack successes or failures. In many industries Return ...e Investment (ROI) metric is used. In specific ...dustries to track performance specialized metrics are ...veloped. This process requires the collection of ...plicable data, processing of the data, data analysis ...d data visualization. This process provides necessary ...sight into past performance.

Diagnostic analytics is used to answer questions about why things happened. These methods supplement more basic descriptive analytics. They take the outcomes from descriptive analytics and dig deeper to find the cause. The performance indicators are further investigated to discover why they got better or worse. This generally occurs in three steps :

- Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.
- Data that is related to these anomalies is collected.
- Statistical techniques are used to find relationships and trends that explain these anomalies.

- Predictive analytics is used to answer questions about what will happen in the future. These methods use historical data to identify trends and determine if they are likely to recur. Predictive analytical tools provide important insight into what may happen in the future and its methods include a variety of statistical and machine learning methods, such as: regression, neural networks and decision trees.
- Prescriptive analytics is used to answer questions about what should be done. By using insights from predictive analytics, data-driven decisions can be made. This allows businesses to make informed decisions in the face of uncertainty. Prescriptive analytics methods rely on machine learning approaches that can find patterns from large datasets. By analyzing past decisions and events, the likelihood of different outcomes can be estimated.
- These types of data analytics provide the insight that businesses need to make effective and efficient decisions. Used in combination they provide a well-rounded understanding of a company's needs and opportunities.

4.3.3 Role of Data Analytics

- Data analysts exist at the crossings of information technology, statistics and business. These fields are combined to help businesses and organizations succeed. The main goal of a data analyst is to improve efficiency and performance with the help of patterns discovery in data.
- The profile of a data analyst involves working with data throughout the data analysis pipeline. This means working with data in various ways. The main steps in the data analytics process are data mining, data management, statistical analysis, and data presentation. The importance and balance of these steps depend on the data being used and the aim of the analysis.
- Data mining is a necessary process for many data analytics tasks. In this data is extracted from unstructured data sources. These may include raw sensor data, written text, or large complex databases. The main steps in this process are to extract, transform, and load data (often called ETL.) These steps convert raw data into a useful and manageable format. This prepares data for storage and analysis. Data mining is mostly the most time-intensive step in the data analysis pipeline.
- Data warehousing is another main aspect of a data analyst's job. Data warehousing comprises of designing and implementing databases that allow easy access to the results of data mining. This step generally involves creating and managing SQL databases.
- Statistical analysis is used to generate insights from data. Both machine learning and statistics methods are used to analyze data. Big data is used to generate statistical models that finds trends in data. These models can then be applied to new data to make predictions and inform decision making. Statistical programming languages like Python or R are necessary to this process. In addition, open source libraries and packages such as TensorFlow enable advanced analysis.
- The last step in most data analytics processes is data presentation. In this step insights are shared with stakeholders. Data visualization is the most important tool in data presentation.

Module 4

- Compelling visualizations can help tell the story in the data which may help executives and managers understand the importance of these insights.

4.3.4 Importance of Data Analytics

- The applications of data analytics are wide. In many different industries, analyzing big data can optimize efficiency. Improving performance enables businesses to succeed in an increasingly competitive world.
- One of the earliest adopters is the financial sector. Data analytics has an important role in the banking and finance industries, used to predict market trends and assess risk. Credit scores are an example of data analytics that affects everyone. These scores use many data points to determine lending risk. Data analytics is also used to detect and prevent fraud to improve efficiency and reduce risk for financial institutions.
- The use of data analytics goes beyond maximizing profits and ROI, however. Data analytics can provide critical information for healthcare (health informatics), crime prevention, and environmental protection. These applications of data analytics use these techniques to improve our world.
- Though statistics and data analysis have always been used in scientific research, advanced analytic techniques and big data allow for many new insights. These techniques can find trends in complex systems. Researchers are currently using machine learning to protect wildlife.
- The use of data analytics in healthcare is already widespread. Predicting patient outcomes, efficiently allocating funding and improving diagnostic techniques are just a few examples of how data analytics is revolutionizing healthcare.
- The pharmaceutical industry is also being revolutionized by machine learning. Drug discovery is a complex task with many variables. Machine learning can greatly improve drug discovery. Pharmaceutical companies also use data analytics to understand the market for drugs and predict their sales.
- The internet of things (IoT) is a field that is used alongside machine learning. These devices provide a great opportunity for data analytics. IoT devices often contain many sensors that collect meaningful data points for their operation.
- Devices like the Nest thermostat track movement and temperature to regulate heating and cooling. Smart devices like this can use data to learn from and predict your behavior. This will provide advance home automation that can adapt to the way you live.
- The applications of data analytics are seemingly endless. More and more data is being collected every day — this presents new opportunities to apply data analytics to more parts of business, science and everyday life.

4.4 LIFECYCLE OF DATA SCIENCE

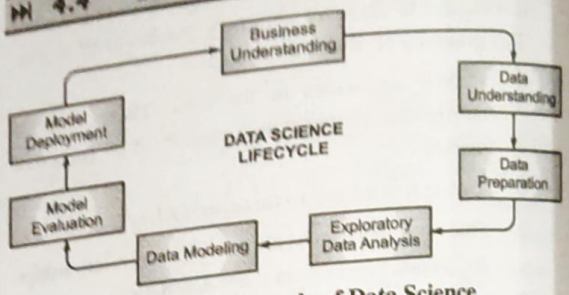


Fig. 4.4.1 : Life cycle of Data Science

1. Business Understanding

- The complete cycle go around the enterprise goal. To resolve a specific problem we should first properly formulate it. It is extraordinarily necessary to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis.
- After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective.
- You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

2. Data Understanding

- After enterprise understanding, the next step is data understanding. This step includes a collection of series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used

for this commercial enterprise problem, and different information.

- This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

3. Preparation of Data

- Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them.
- Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and features.
- Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

4. Exploratory Data Analysis:

- This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model.
- Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps.
- Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

5. Data Modelling

- Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output.
- This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem.
- After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyper parameters of every model to obtain the preferred performance.
- We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6. Model Evaluation

- Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality.
- If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved.
- Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric.
- We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7. Model Deployment

- The model after a rigorous assessment is at the end deployed in the preferred structure and

Module 4

channel. This is the last step in the data science life cycle.

- Each step in the data science life cycle defined above must be labored upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste.
- For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work.
- If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

4.5 ROLES IN DATA SCIENCE PROJECTS

There are certain main roles that are required for the complete and fulfilled functioning of the data science team to execute projects on analytics successfully. There are seven main roles.

Each plays an important role in developing a successful analytics project. There is no hard and fast rule for considering the listed seven roles, they can be used fewer or more depending on the scope of the project, skills of the participants, and organizational structure.

Main Roles for a Data analytics project

1. Business User

- The business user understands the main area of the project and is also basically benefited from the results.
- This user gives advice and consults the team working on the project about the value of the results obtained and how the operations on the outputs are done.
- The business manager, line manager, or deep subject matter expert in the project mains fulfils this role.

2. Project Sponsor

- The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
- He generally provides the funds and measures the degree of value from the final output of the team working on the project.
- This person introduces the prime concern and brooms the desired output.

3. Project Manager

- This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

4. Business Intelligence Analyst

- Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.
- This person generally creates fascia and reports and knows about the data feeds and sources.

5. Database Administrator (DBA)

- DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
- His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

6. Data Engineer

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
- The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

Data Scientist

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
- He ensures overall analytical objectives are met.
- Data scientists outline and apply analytical methods and proceed towards the data available for the concerned project.

4.6 APPLICATIONS AND CASE STUDIES OF DATA SCIENCE IN VARIOUS INDUSTRIES

- Data Science is the deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc. It uses various tools and techniques to extract meaningful data from raw data. Data Science is also known as the Future of Artificial Intelligence.
- For Example, Ram loves books to read but every time when he wants to buy some books he was always confused that which book he should buy as there are plenty of choices in front of him. Here Data Science Technique will be useful. When he opens Amazon he will get product recommendations on the basis of his previous data. When he chooses one of them he also gets a recommendation to buy these books with this one as this set is mostly bought. So all Recommendation of Products and Showing set of books purchased collectively is one of the examples of Data Science.

Applications of Data Science

1. Search Engines

- The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.
- For Example, when we search something suppose "Data Structure and algorithm courses" then at that time on the Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is done using Data Science, and we get the topmost visited Web Links.

2. Transport

- There is always a need for people to reach their destinations on time and data science and analytics can be used by transportation providers, both public and private, to increase the chances of successful journeys. For instance, Transport for London uses statistical data to map customer journeys, manage unexpected circumstances, and provide people with personalized transport details.
- Public transport officials also use predictive analytics to keep things functioning smoothly. In 2017, Americans took **10.1 billion** public transit trips. The substantial data generated from these trips can allow data scientists to analyze this data to ensure that all obstacles are properly dealt with.
- Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.
- For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. Finance

- The banking industry is generally not looked at as being one that uses technology a lot. However, this is slowly changing as bankers are beginning to increasingly use technology to drive their decision-making.

Module 4

channel. This is the last step in the data science life cycle.

- Each step in the data science life cycle defined above must be labored upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste.
- For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work.
- If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

4.5 ROLES IN DATA SCIENCE PROJECTS

There are certain main roles that are required for the complete and fulfilled functioning of the data science team to execute projects on analytics successfully. There are seven main roles.

Each plays an important role in developing a successful analytics project. There is no hard and fast rule for considering the listed seven roles, they can be used fewer or more depending on the scope of the project, skills of the participants, and organizational structure.

Main Roles for a Data analytics project

1. Business User

- The business user understands the main area of the project and is also basically benefited from the results.
- This user gives advice and consults the team working on the project about the value of the results obtained and how the operations on the outputs are done.
- The business manager, line manager, or deep subject matter expert in the project mains fulfils this role.

2. Project Sponsor

- The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
- He generally provides the funds and measures the degree of value from the final output of the team working on the project.
- This person introduces the prime concern and brooms the desired output.

3. Project Manager

- This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

4. Business Intelligence Analyst

- Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.
- This person generally creates fascia and reports and knows about the data feeds and sources.

5. Database Administrator (DBA)

- DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
- His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

6. Data Engineer

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
- The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

5. Data Scientist

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
- He ensures overall analytical objectives are met.
- Data scientists outline and apply analytical methods and proceed towards the data available for the concerned project.

4.6 APPLICATIONS AND CASE STUDIES OF DATA SCIENCE IN VARIOUS INDUSTRIES

- Data Science is the deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc. It uses various tools and techniques to extract meaningful data from raw data. Data Science is also known as the Future of Artificial Intelligence.
- For Example, Ram loves books to read but every time when he wants to buy some books he was always confused that which book he should buy as there are plenty of choices in front of him. Here Data Science Technique will be useful. When he opens Amazon he will get product recommendations on the basis of his previous data. When he chooses one of them he also gets a recommendation to buy these books with this one as this set is mostly bought. So all Recommendation of Products and Showing set of books purchased collectively is one of the examples of Data Science.

☞ Applications of Data Science

1. Search Engines

- The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.
- For Example, when we search something suppose "Data Structure and algorithm courses" then at that time on the Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is done using Data Science, and we get the topmost visited Web Links.

2. Transport

- There is always a need for people to reach their destinations on time and data science and analytics can be used by transportation providers, both public and private, to increase the chances of successful journeys. For instance, Transport for London uses statistical data to map customer journeys, manage unexpected circumstances, and provide people with personalized transport details.
- Public transport officials also use predictive analytics to keep things functioning smoothly. In 2017, Americans took **10.1 billion** public transit trips. The substantial data generated from these trips can allow data scientists to analyze this data to ensure that all obstacles are properly dealt with.
- Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.
- For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. Finance

- The banking industry is generally not looked at as being one that uses technology a lot. However, this is slowly changing as bankers are beginning to increasingly use technology to drive their decision-making.

Module 4

- For instance, the Bank of America uses natural language processing and predictive analytics to create a virtual assistant called Erica to help customers view information on upcoming bills or view transaction histories.
- Erica, the virtual assistant, is also trained to get smarter with every transaction. Bank of America representatives say that the assistant will eventually study their customers' banking habits and suggest relevant financial advice at appropriate times.
- Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.
- For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. E-Commerce

- E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.
- For Example, when we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

5. Health Care

- The medical industry is using big data and analytics in a big way to improve health in a variety of ways. For instance, the use of wearable trackers to provide important information to physicians who can make use of the data to provide better care to their patients. Wearable trackers also provide information like whether the patient is taking his/her medication and following the right treatment plan.
- Data compiled over time provide physicians with comprehensive information on patients' wellbeing and provide much more actionable data than just short in-person visits.
- Big data and analytics can also help hospital managers improve care and reduce waiting times. Medical data is a great example of how providers can look at large amounts of data to find patterns and prescribe appropriate courses of action.
- In the Healthcare Industry data science act as a boon. Data Science is used for:
  - Detecting Tumour.
  - Drug discoveries.
  - Medical Image Analysis.
  - Virtual Medical Bots.
  - Genetics and Genomics.
  - Predictive Modelling for Diagnosis etc.

6. Image Recognition

- Currently, Data Science is also used in Image Recognition. For Example, when we upload our image with our friend on Facebook, Facebook gives suggestions tagging who is in the picture.
- This is done with the help of machine learning and Data Science. When an Image is recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

- Targeting Recommendation is the most important application of Data Science.
- Whatever the user searches on the Internet, he/she will see numerous posts everywhere.
- This can be explained properly with an example. Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy

offline. Data Science helps those companies who are paying for Advertisements for their mobile.

- So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

7. Airline Routing Planning

- With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays.
- It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

8. Data Science in Gaming

- In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance.
- There are many games like Chess, EA Sports, etc. will use Data Science concepts.

10. Medicine and Drug Development

- The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life.
- Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors.
- The algorithms based on data science will forecast how this will react to the human body without lab experiments.

11. Autocomplete

- AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line.
- In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.
- Data science has been effective in tackling many real-world problems and is being increasingly adopted across industries to power more intelligent and better-informed decision-making.
- With the increased use of computers for day-to-day business and personal operations, there is a demand for intelligent machines, can learn human behavior and work patterns. This brings Data science and big data analytics to the forefront.

12. Retail

- Retailers need to correctly anticipate what their customers want and then provide those things. If they don't do this, they will likely be left behind the competition.
- Big data and analytics provide retailers the insights they need to keep their customers happy and returning to their stores.
- One study by IBM said that **62%** of retail respondents claimed that insights provided by analytics and information provided them with competitive advantages.
- There are many ways retailers can use big data and analytics to keep their shoppers coming back for more.
- For instance, retailers can use big data and analytics to create hyper-personal and relevant shopping experiences that make their customers highly satisfied and more prone to making purchase decisions.

13. Construction

- It is no surprise that construction companies are beginning to embrace data science and analytics in a big way.

- Construction companies track everything from the average time needed to complete tasks to materials-based expenses and everything in between.
- Big data is now being used in a big way in the construction industry to drive better decision-making.

14. Communications, Media, and Entertainment

- Consumers now expect rich media in different formats as and when they want it on a variety of devices. Collecting, analyzing, and utilizing these consumer insights is now a challenge that data science is stepping in to tackle.
- Data science is being used to leverage social media and mobile content and understand real-time, media content usage patterns.
- With data science techniques, companies can better create content for different target audiences, measure content performance, and recommend on-demand content.
- For example, Spotify, the on-demand music streaming service, uses Hadoop big data analytics to collect and analyze data from its millions of users to provide better music recommendations to individual users.

15. Education

- One challenge in the education industry where data science and analytics can help is to incorporate data from different vendors and sources and use them on platforms not designed for varying data.
- For example, the University of Tasmania with over 26,000 students has developed a learning and management system that can track when a student logs into the system, the overall progress of the student, and how much time is spent on different pages, among other things.
- Big data can also be used to measure teachers' effectiveness by fine-tuning teachers' performance by measuring against subject matter, student numbers, student aspirations, student demographics, and many other variables.

16. Manufacturing and Natural Resources

- The increasing demand and supply of natural resources, such as oil, minerals, gas, metals, agricultural products, etc. has led to the generation of huge amounts of data that is complex, difficult to handle, and a prime candidate for big data analytics. The manufacturing industry also generates huge amounts of data that has so far gone untapped.
- Big data allows decision-making to be supported by predictive analytics in the natural resources industry. Large amounts of geospatial data, text, temporal data and graphical data can be analyzed using data science to ingest and integrate these large datasets. Big data also has a role to play in reservoir characterization and seismic interpretation, among others.

17. Government

- Big data has many applications in the public services field. Places where big data is/can be used include in financial market analysis, health-related research, environmental protection, energy exploration, and fraud detection.
- One specific example is the use of big data analytics by the Social Security Administration (SSA) to analyze large numbers of social disability claims that come in as unstructured data.
- Analytics is being used to rapidly process medical information and detect fraudulent or suspicious claims. Another example is the use of data science techniques by the Food and Drug Administration (FDA) to identify and analyze patterns related to food-related diseases and illnesses.

18. Energy and Utilities

- The energy and utilities industry generates and will continue to generate huge amounts of data that can be analyzed using big data analytics. For instance, nowadays, smart readers allow data to be collected every 15 minutes or so as compared to how it was previously when it was once a day.
- This data can be used to better study the consumption of utilities, which in turn allows for

better control of utility use and improved customer feedback. The use of big data by utility companies also allows for improved asset and workforce management and is useful for identifying and correcting errors as soon as possible.

19. Outsourcing Industry

- The value of the global data science and analytics outsourcing market was US$ 2.49 Bn in 2018 and is expected to grow to **USD 19.36 Bn** by 2027 at a CAGR of **25.8%**. Factors driving this growth are shortage of skilled resources and high adoption by diverse industries.
- Outsourcing companies are not far when it comes to Data Science Services. They are making use of data science to automate back-office processes, keep prices in check, and shorten the turnaround time.
- Flatworld Solutions is one such company using Artificial Intelligence (AI) and Machine Learning (ML) to automate the back-end processes for clients to automatically classify and index documents, process PDF files, name and classify files, automatically discover documents, use image annotation for inventory management, and more.
- The term 'Data Science' was first coined in 2001 and it took less than two decades for it to become the phenomenon it is today. Finance was the first industry to understand data science advantages when no one could and used it to sift through and analyze large amounts of data and help companies reduce losses.
- Today, Data Science is a force to reckon with and almost all industries are trying to leverage it's potential, and this number will only continue to increase as data science technology becomes more reliable and cost-effective.
- However, to capitalize on data science opportunities, you will need to understand industry-specific challenges, understand data characteristics of each industry, and match market needs with custom capabilities and solutions.

4.7 SAMPLE QUESTIONS AND ANSWERS

Q. 1 What is Data Science? Write a note on evolution of Data Science. *(Ans. : Refer section 4.1)* **(10 Marks)**

Q. 2 Differentiate between Data Science and Business Analytics: *(Ans.: Refer Example 4.2.1)* **(10 Marks)**

Q. 3 Differentiate between Data Science and Big Data *(Ans.: Refer section4.2.2)* **(10 Marks)**

Q. 4 Write a short note on Data Analytics. *(Ans.: Refer section 4.3)* **(10 Marks)**

Q. 5 Explain the steps in Data Science Process. *(Ans.: Refer Example 4.4)* **(10 Marks)**

Q. 6 Explain the different roles in Data Science Projects *(Ans.: Refer section4.5)* **(10 Marks)**

Q. 7 Explain the applications of Data Science. *(Ans. :Refer section 4.6)* **(10 Marks)**

Module 4

Chapter Ends...

❑❑❑

Module V

CHAPTER 5

Exploratory Data Analysis

University Prescribed Syllabus w.e.f Academic Year 2021-2022

Introduction to exploratory data analysis, Typical data formats. Types of EDA, Graphical/Non graphical Methods. Univariate/multivariate methods Correlation and covariance, Degree of freedom Statistical Methods for Evaluation including ANOVA.

Self-Learning Topics : Implementation of graphical EDA methods.

Teaching Hours – 04 **Approximate Weightage of Marks in Exam. – 10 Marks**

Some key benefits of EDA include :

1. Spotting missing and incorrect data :
2. Understanding the underlying structure of data
3. Testing hypothesis and checking assumptions
4. Calculating the most important variables
5. Creating the most efficient model
6. Determining error margins
7. Identifying the most appropriate statistical tools to help us

1. Spotting missing and incorrect data :

As part of the **data cleaning process, an** initial data analysis (IDA) can help us spot any structural issues with our dataset.

2. Understanding the underlying structure of data

Properly mapping data ensures that you maintain **high data quality** when transferring it from its source to your database, data warehouse etc.

Understanding data structure means one can avoid mistakes entering in.

3. Testing hypothesis and checking assumptions

Before analysing data, any assumptions or hypothesis should be scrutinised. EDA may not give all the details but it helps us to spot if we are inferring the right outcomes based on our understanding of the data.

4. Calculating the most important variables

Before carrying out data analysis, the relationship between the variables must be identified.

For example, which independent variables which dependent variables ? This way one call extract the most useful information.

5. Creating the most efficient model

To carry out perfect analysis, extraneous information is to be removed. It is because needless information may obscure key insights. Here, EDA helps to identify information that you can extract.

6. Determining error margins

EDA is also capable of determining which data may lead to unavoidable errors in later analysis. One can avoid accepting false conclusions in such cases, if one knows beforehand which data will impact the results.

7. Identifying the most appropriate statistical tools to help us

The most important use of EDA is that it helps us to determine which techniques and statistical models will help us to get what we need, from the dataset. EDA will guide us to carry out either a predictive analysis or a sentiment analysis. One can learn different types of data analysis using EDA.

5.2.4 The Underlying Principles of EDA

(i) EDA considers what to look for, how to look for it, and, how to interpret what we discover.

(ii) In short, EDA is more of an **Attitude** than it is a step-by-step process.

(iii) Avoid making assumptions about the rules that we think, the data will adhere to

(iv) Exploring data with an open mind reveals its underlying nature in toto.

(v) Exploratory data analysis is **a qualitative investigation, and not a quantitative one.**

Thus EDA looks at a dataset's inherent qualities with an inquisitive mindset.

5.2.5 Five-number Summary

The five-number summary is a set of five descriptive statistics. They make a useful starting point for any explorative data analysis. The five number summary often a concise overview of how different observations in the dataset are distributed.

The five number summary includes the fixed most common sample percentiles :

1. **The sample minimum** (the smallest observations)
2. **The lower quartile** (the median of the lower half of data).
3. The median (the average/middle value)
4. The upper quartile (the median of the upper half of the data)
5. The sample maximum (the largest observation).

These can be used to determine the **interquartile range,** which is the middle 50% of the dataset.

This helps to describe the overall spread of the data; This allows us to identify any outliers.

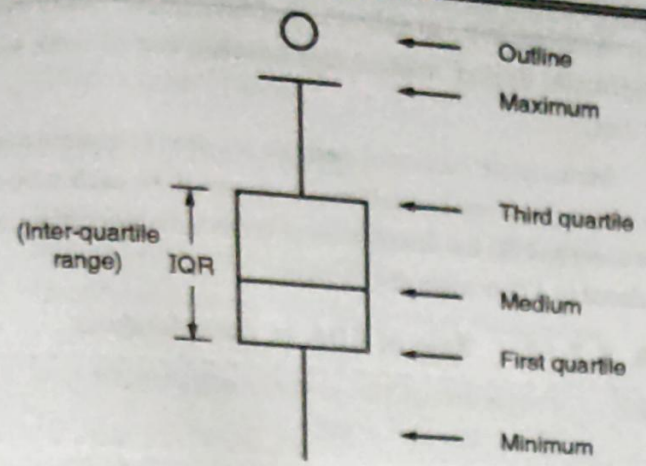


Fig. 5.2.2

The five number summary is a foundational part of data exploration because the five-number summary can be used to determine a great number of additional attributes about a given dataset.

5.2.6 Exploratory Data Analysis Tools

GQ. What are different Exploratory Data Analysis Tools ?

Specific statistical functions and techniques one can perform with EDA tools include:

1. Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
2. Univariate visualization of each field in the raw dataset, with summary statistics.
3. Bivariate visualizations and summary statistics that allow one to assess the relationship between each variable in the dataset and the target variable one is looking at.
4. Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
5. K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
6. Predictive models, such as linear regression, use statistics and data to predict outcomes.

5.2.7 Advantages of EDA

GQ. What are different advantages of EDA ?

1. **Improve understanding** of variables by extracting averages, mean, minimum and maximum values, etc.
2. **Discover errors**, outliers and missing values in the data.
3. **Identify patterns** by visualising data in graphs such as scatter plots, histograms.

Thus, the main goal is to understand data better and use tools effectively to gain valuable insights or draw conclusions.

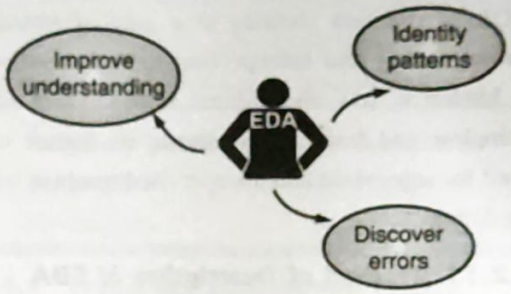


Fig. 5.2.3

5.2.8 Process of EDA

1. Getting **maximum insights** from a data set.
2. Uncovering underlying structure.
3. Extracting important variables from the dataset.
4. Detecting **outliers and anomalies** (if any), Testing underlying assumptions.
5. Determining the optimal factor settings.

5.2.9 Performing EDA

Step by step approach to perform EDA:

Step 1 : Use resources like blogs, Massive Open Online Courses (MOOCs) for setting EDA.

Step 2 : To become familiar with various data visualisation techniques, **e.g. charts, plots**.

Step 3 : To demonstrate some of the steps with python, code snippet.

Step 4 : In statistics, EDA is an approach of analysing data sets to summarise their main characteristics, often using **statistical graphics** and other data visualisation methods. EDA is the process of

investigating the dataset to discover patterns and anomalies outliers, and form hypothesis based on our understanding of the data-set. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.

5.2.10 EDA in Machine Learning

- EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques we are considering for data analysis are appropriate.
- EDA in machine learning is a way of **visualizing, summarizing and interpreting** the **information** that is hidden in row and column format. Once EDA is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling.

5.2.11 Method of Description of EDA

EDA code is supposed to perform the following steps :

1. Review data
2. Check total, number of entries and column types.
3. Check any null values.
4. Check duplicate entries.
5. Plot distribution of numeric data (Univariate and pairwise joint distribution).

5.2.12 EDA Plots

EDA is heavily based on **graphical techniques**. One can use graphical techniques to identify most important properties of EDA.

EDA plots are : Bar Graph, Histogram, Pie-charts, frequency-polygon.

5.2.13 Four Primary Types of EDA

Four primary types of EDA are :

1. Univariate, Non-graphical
2. Univariate, Graphical
3. Multivariate, Graphical
4. Multivariate, Non-graphical

Multivariate data arises from more than one variable.

Multivariate graphical : Multivariate data uses graphics to display relationships between two or more sets of data.

Multivariate statistical methods are used to analyse data in which (1) several variables are observed for each subject (or case) and (2) the distribution of those variables cannot be reduced to a univariate distribution.

5.2.14 Role of EDA in Data Analysis

GQ. What is Role of EDA in Data Analysis ? OR What is box plot graph.

- Exploratory graphs serve mostly the same functions on graphs. They help us find patterns in data and make us understand its properties.
- They suggest **modelling strategies** and help debug analyses. **Python and R-language** are the two most commonly used data science tools to create an EDA. EDA can be done using python **for identifying the missing value in a data set.**
- **Box-plot graph :** divides the data into sections that each contain approximately **25% of data** in that set. Box plots are useful as they provide a visual summary of the data enabling the researchers to quickly identify **mean values, the dispersion of the data set, and signs of skewness.**
- Exploratory data visualization are the types of visualizations that we assemble when we do not have a clue about what information lies within our data.

5.2.15 EDA and Visualization

- Exploratory data analysis is a way to better understand the data which helps in further preprocessing. And data visualization is a key, making exploratory data analysis process streamline and easily analyzing data using wonderful **charts and plots.**
- Data processing and EDA are essential tasks for any data-science projects. We observe that EDA and data processing are distinct terms, but they have many overlapping subtasks and are usually used interchangeably. The data set and original code can be accessed through the **Git-Hub link.**

Consider the data set as :

- There are dress-shoes, hiking boots, sandals etc. Using EDA, one is open to the fact that any number of people might buy any number of different types of shoes.

- Using EDA, one can visualize the data to find the maximum number of customers buying **'1 to 3'** different types of shoes.

5.2.16 Inclusion in EDA

1. Extracting important variables from the data and leaving behind useless variables.
2. Identifying outliers, missing values or human-error.
3. Understanding relationships or lack of relationships between variables.

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set while providing all of the specific items that an analyst would want to extract from a data set, such as a **good-fitting, parsimonious model, a list of outliers**.

EDA is used for seeing what the data can tell us before the modelling task. It is not easy to look at a column of numbers and determine important characteristics of the data. It may be tedious, boring to derive insights by looking at plain numbers.

5.3 UNIVARIATE ANALYSIS (U.A)

Univariate analysis is the simplest form of analysing data 'Uni' means 'one', so in other words, the data has only **one variable.** It does not deal with causes or relationships (unlike regression) and its major purpose is to take data, summarise that data and find patterns in that data.

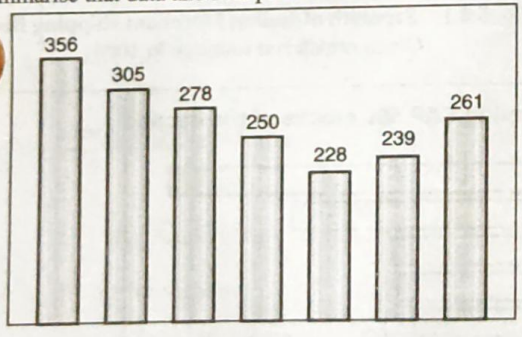


Fig. 5.3.1

5.3.1 Steps to be Followed for U.A.

1. Prepare your dataset.
2. Choose analysis (type) : Descriptive statistics or frequencies.
3. Click statistics and analyse the required data and then click continue.
4. To click the chart.
5. Choose the expected chart and then click continue.
6. Click O.K. and finish the analysis.
7. See and interpret the output.

Univariate analysis means **analysis of one variable or one feature**. Univariate analysis basically tells us how data in each feature is distinguished and also tells us about the central tendencies like **mean, median** and **mode**. U.A. is characterised by or is dependent on only one random variable, a **uni-linear model**. In a dataset, it explores each variable separately.

5.4 METHODS OF UNIVARIATE DISTRIBUTION

GQ. Discuss different methods of Univariate Distribution.

Univariate distribution can be described as :

1. Frequency Distribution	2. Bar-Charts
3. Bar graph	4. Histogram
5. Pie-diagram	6. Frequency polygon

5.4.1 Frequency Distribution (F.D.)

F.D. reflects how often an occurrence has taken place in data. It gives a brief idea of the data and makes it easier to find patterns : Consider an experiment of taking IQ test of 17 students in a Engineering college.

The scores obtained are :

118, 139, 141, 142, 144, 147, 148, 149, 157, 152, 154, 157 ;

The frequency representation is :

118 – 138 – 1 ; 139 – 144 – 7
145 – 149 – 4 ; 152 – 157 – 5

5.4.2 Bar-Charts

Module 5

The bar-chart is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualising data. **For example**, the data relating the sales, profits, production, population etc. For different periods may be presented by bar diagram.

Remark

If there are a large number of items or values of the variable under study, then instead of bar diagram, line diagram may be drawn.

5.4.3 Bar-Graph

A Bar graph (also known as a bar chart or bar diagram) is a visual tool that uses bars to compare data among categories.

A bar graph may be horizontal or vertical. The longer the bar, the greater its value. Bar graphs consist of two axes :

On a vertical graph, the horizontal axis P or X-axis, shows the data categories. The vertical axis is the scale.

☞ Attributes of Bar Graphs

(i) A bar diagram makes it easy to compare sets of data between different groups.

(ii) The graph represents categories on one axis and a discrete value in the other. This represents the relationship between the two axes.

(iii) Bar charts can also big changes in data over time.

☞ Use of a Bar-Graph

(i) Bar graphs are an effective way to compare items between different groups.

(ii) They are effective visual in presentations and reports.

(iii) From the bar-graph, one can recognitions patterns or trends for more easily than at a table of numerical data.

☞ Types of a Bar graph

(1) Vertical bar Graph

(i) The most common type is the vertical bar graph. It is useful when presenting a series of data over time.

One disadvantage is that they don't leave much space at the bottom if long labels are required

(ii) **Horizontal bar Graph**

Here there is plenty of room for long label along the vertical axis.

Best performing S and P 500 stocks of the decade.

☞ Example

The following data relating to the strength of the Indian merchant shipping fleet gives the Gross Registered Tonnage (GRT) as on 31st Dec. For the different years.

Year	1961	1966	1971	1975	1976
GRT in' 000	901	1,792	2,500	4,464	5,115

Represent data by suitable bar-diagram.

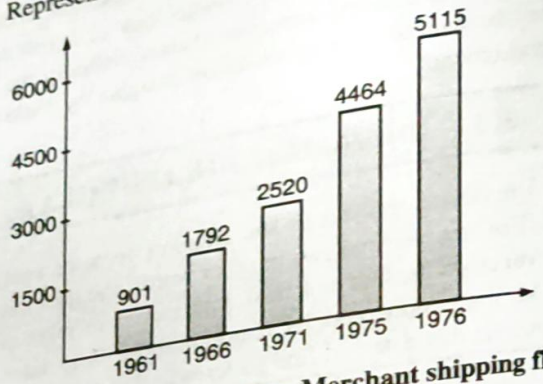


Fig. 5.4.1 : Strength of Indian Merchant shipping fleet (Gross registered tonnage in 1000)

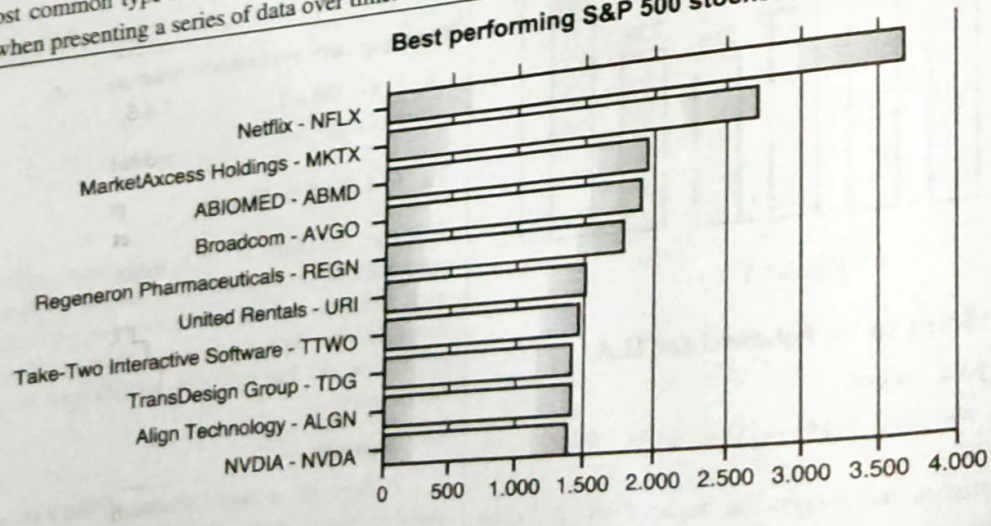


Fig. 5.4.2

5.4.4 Histogram

A histogram is the most commonly used graph to exhibit frequency distribution.

It looks very much like a bar chart, but these are important differences between them.

Histogram is to be used when

1. The data is numerical
2. To note the shape of the data's distribution, especially when whether the output of a process is distributed approximately normal.
3. To check whether a process can satisfy the customer's requirements.
4. To analyse the output from a supplier's process.
5. To observe whether a process change has occurred from one time period to another.
6. To check whether the outputs of two or more processes are different.
7. To communicate the distribution of data quickly and easily to others.

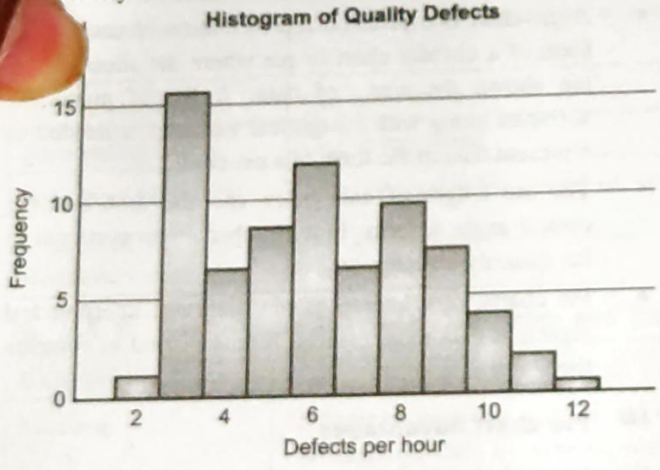


Fig. 5.4.3 : Histogram Example

To create a Histogram

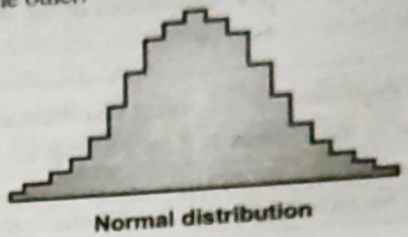
1. Collect at least 50 consecutive data points from a process.
2. Draw X-and Y-axis on graph paper. Mark and label the Y-axis for counting data values. Mark and label the X-axis the values of data.

The spaces between these numbers will be the bars on histogram. There is no space between bars.

Typical Histogram shapes and what they mean

(1) Normal Distribution

It is a bell-shaped curve known as the "normal distribution". In a normal or "typical" distribution, points are as likely to occur on one side of the average as on the other.



Normal distribution

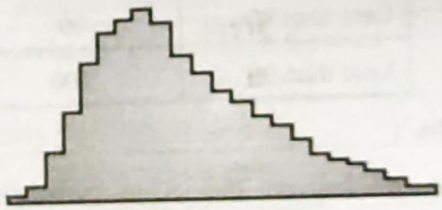
Fig. 5.4.4

(2) Skewed Distribution

The skewed distribution is asymmetrical because a natural limit prevents outcomes on one side.

The distributions peak is off centre toward the limit and a tail stretches away from it.

These distributions are called right or left-skewed according to the direction of the tail.



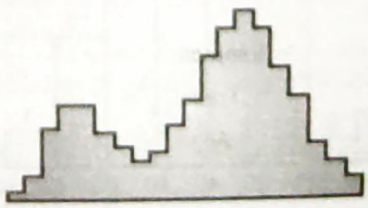
Right-skewed distribution

Fig. 5.4.5

(3) Double peaked or Bimodal

Here the outcomes of two processes with different distributions are combined in one set of data.

The bimodal distribution looks like the back of a two-humped camel.



Bimodal (double-peaked) distribution

Fig. 5.4.6

Module 5

It is one of the most popular and commonly used devices for charting **continuous frequency distribution**.

It consists in erecting a series of adjacent vertical rectangles on the sections of the **horizontal axis (X-axis),** with bases (sections) equal to the width of the corresponding class intervals and heights are so taken that the areas of the rectangles are equal to the frequencies of the corresponding classes.

Ex. 5.4.1 : Represent the adjoining distribution of marks of 100 students in the examination by a histogram.

Marks obtained	No. Of students
Less than 10	4
Less than 20	6
Less than 30	24
Less than 40	46
Less than 50	67
Less than 60	86
Less than 70	96
Less than 80	99
Less than 90	100

☑ **Soln. :**

First we convert the given cumulative frequency distribution into the frequency distribution of marks in each range.

Marks	No. Of students
0 – 10	4
10 – 20	6 – 4 = 2
20 – 30	24 – 6 = 18
30 – 40	46 – 24 = 22
40 – 50	67 – 46 = 21
50 – 60	86 – 67 = 19
60 – 70	96 – 86 = 10
70 – 80	99 – 96 = 3
80 – 90	100 – 99 = 1

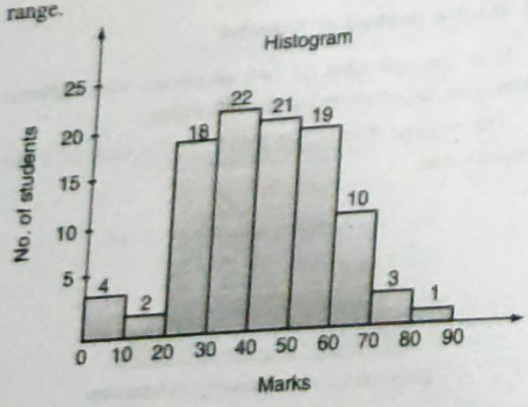


Fig. Ex. 5.4.1 : Marks

5.4.5 Pie-Chart

☞ **What is pie-chart :**

- A pie chart is a type of a change that displays data in a circular graph. If one of the most commonly used graphs to represent data using the attributes of **circle spheres**, and angular data to represent real world information.
- A pie-chart is a pictorial representation of data in the form of a circular chart or pie where the slices of the pie shown the size of data. A list of numerical variables along with categorical variables is needed to represent data in the form of a pie-chart.
- The arc length of each slice and the area and the central angle it forms in a pie chart is proportional to the quantity it represents.
- Pie charts, also known as pie diagrams interpret and represent data more clearly. It is also used to compare the given data.

☞ **Pie-chart advantages**

(i) A straight forward and easy-to-understand illustrations.

(ii) It visually portrays data as a fraction of a whole, and is an important communication tool for even inexperienced audience.

(iii) It allows the viewer to do an immediate analysis or quickly comprehend details.

(iv) One can manipulate data in the pie-chart to highlight points one wants to make.

(v) Pie-charts are pleasing, therefore great for gaining the attention of the viewers.

Disadvantages of the pie-chart

(i) When there are many data points in a pie-chart, it loses its effectiveness.

(ii) It there are many pieces of data, they can become confusing and difficult to read

(iii) Since the chart reflects one data set, you will need a serves of pie-charts to compare different settings.

(iv) It is not easy to compare data slices because the reader has to account for angles and compare non-adjacent pieces.

(v) Where there is negative data, a pie-chart is not a good choice.

5.4.6 Pie-diagram

Steps of construction of pie-diagram

(1) Express each of the component values as a percentage of the respective total.

(2) Since the angle at the centre of a circle is **360°** and each component part is to be expressed proportionately in degrees. Since 1 percent of the total value is equal to $\frac{360}{100} = 3.6°$; the percentage of the common value from step 1. Can be converted to degrees by multiplying each of them by 3.6.

(3) Draw a circle of appropriate radius and different sectors representing various components should be distinguished from one another by using different shades, dotting, colours etc.

5.4.7 Comparison Table (Bar Chart V/s Histogram)

Comparison term	Bar chart	Histogram chart
Usage	To compare different categories of data	To display the frequency of occurrences.
Indicates	Discrete values	Non-discrete values
Data	Categorical data	Quantitative data
rdering bars	Each data point is rendered as a separate bar	The data points are grouped and rendered based on the bin value
between bars	Can have space	No space
nting bass	Can be reordered	Cannot be reordered
label placement	Axis labels can be placed on or between the ticks	Axis labels are placed on the ticks
Required values	x and y	Only y

5.4.8 Comparison Between Histogram and Bar Graph

Basis for comparison	Histogram	Bar Graph
Meaning	Histograms refers to a graphical representation, that displays data by way of bars to show the frequency of numerical data	Bar graph is a pictorial representation of data that used bars to compare different categories of data
Indicates	Distribution of non-discrete variables	Comparison of discrete variables
Orients	Quantitative data	Categorical data
Spaces	Bars touch each other, hence there are no spaces between bars	Bars do not touch each other, hence there are spaces
Elements	Elements are grouped together, so that they are considered as ranges.	Elements are taken as individual entities
Can bars be reordered	No	Yes
Width of bars	Need not be same	same

Module 5

Ex. 5.4.2 : The following data represent expenditure by a state government for the year 1997-98. Draw a pie-diagram.

Items	Agriculture and rural development	Industries and urban development	Health and education	Miscellaneous
Proposed expediting in million Rs.	4,200	1,500	1,000	500

☑ **Soln. :**

Calculation for pie-chart

Items (1)	Proposed expenditure (in million Rs.) (2)	Angle at the centre (3) $\frac{2}{7200} \times 360°$
Agriculture and rural development	4,200	$\frac{4200}{7200} \times 360° = 210°$
Industries and urban development	1,500	$\frac{1500}{7200} \times 360° = 75°$
Health and education	1,000	$\frac{1000}{7200} \times 360° = 50°$
Miscellaneous	500	$\frac{500}{7200} \times 360° = 25°$
Total	7,200	360°

Pie-diagram representing proposed expenditure by state-government on different items for 97-98.

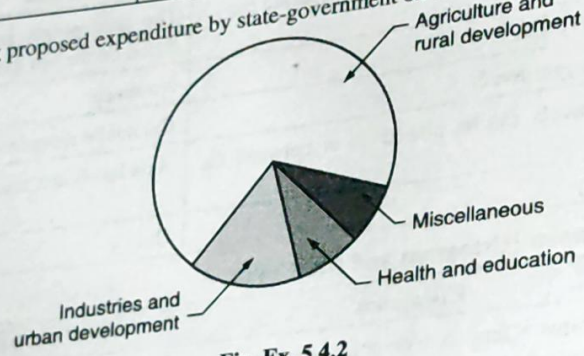


Fig. Ex. 5.4.2

5.4.9 Frequency Polygon

Frequency polygons are a graphical representation of data distribution that helps in understanding the data through a specific shape. Frequency polygons are very similar to histograms but are useful while comparing two or more data.

☞ **Definition**

Frequency polygon is defined as a form of a graph that interprets information or data that is widely used in statistics. This visual form of data representation helps in depicting the shape and trend of the data in an organised and systematic manner. Frequency polygons through the shape of the graph depict the number of occurrence of class intervals.

While a histogram is a graph with rectangular bars without spaces, a frequency polygon graph is a line graph that represents cumulative frequency distribution data.

Steps to construct frequency polygons

The curve in a frequency polygon is drawn on X-axis and Y-axis. X-axis represents the value in a dataset and Y-axis shows the number of occurrences of each category.

- **Step 1** : Mark the class intervals for each class on X-axis while we plot the curve on Y-axis.
- **Step2** : Calculate the midpoint of each of the class interval which is the class mark.
- **Step 3** : Mark the class-marks on X-axis
- **Step 4** : Plot the frequency according to each class mark
- **Step 5** : Once the points are marked, join them with a line segment similar to a line graph. The curve that is obtained by this line segment is the frequency polygon.

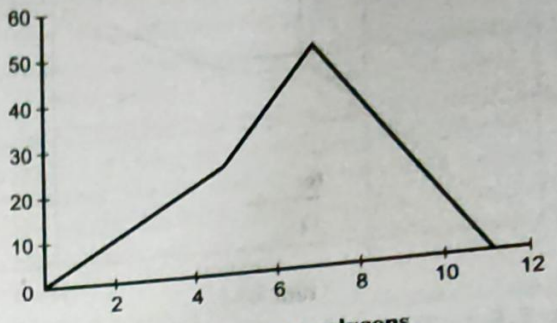


Fig. 5.4.7 frequency polygons

Ex. 5.4.3 : The following data show the number of accidents sustained by 313 drivers of a public utility company over a period of 5 years. Draw the frequency polygon.

No. of accidents	0	1	2	3	4	5	6	7	8	9	10	11
No. of drivers	80	44	68	41	25	20	13	7	5	4	3	2

Soln. :

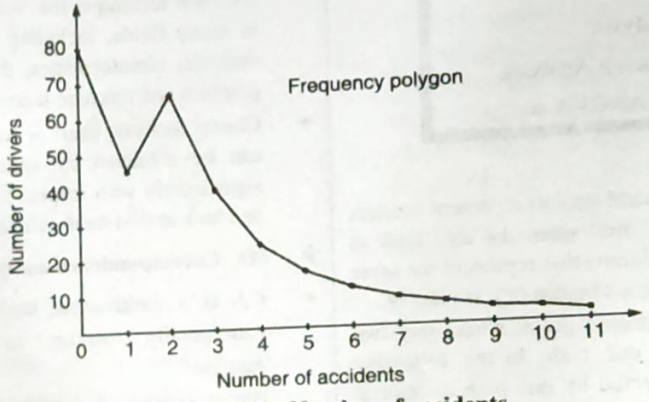


Fig. Ex. 5.4.3 : Number of accidents

5.5 MULTIVARIATE ANALYSIS (MANOVA)

- Multivariate data consists of individual measurements that are acquired on a function of **more than two variables**, e.g., kinetic measurements at many wavelengths and a function of temperature, or as a function of initial concentrations, of the reacting solutions.
- Multivariate analysis is used 'to study more complex sets of data than what univariate analysis methods can handle.
- This type of analysis is almost always performed with software or by graphical methods.
- Multivariate analysis looks at two or more variable and explores the relationship between them.
- Initially, it is better to carry out univariate analysis on each of the variables before doing explorative data analysis.

Module 5

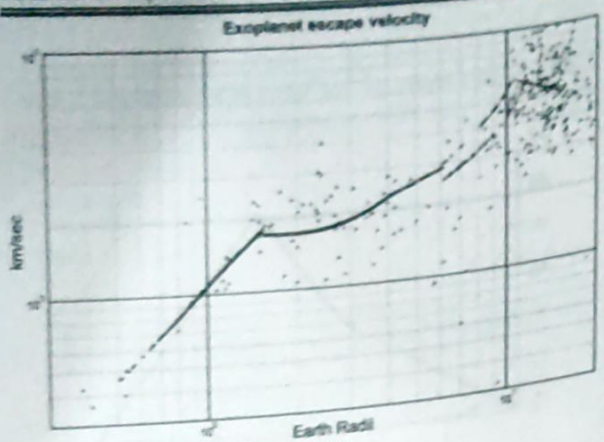


Fig. 5.5.1

5.5.1 Types of Multivariate Analysis (MANOVA)

GQ. Discuss different types of Multivariate Analysis (MANOVA)

(1) Additive Tree,
(2) Canonical Correlation Analysis,
(3) Cluster Analysis,
(4) Correspondence Analysis,
(5) Multiple Correspondence Analysis.
(6) Factor Algorithm, MANOVA etc.

(1) Additive Tree

- An additive tree is a 'general way to represent clusters of data in a graph'. It is used when the data table is composed of rows and columns that represent the same units; the measure must be a distance or a similarity.
- A "tree" is a finite, connected graph where any two nodes are connected by one path. In the following diagram, node B is connected by one path to node E and node E is connected by one path to node F.
- The additive tree is a similar technique to 'cluster analysis'. Both techniques have the "leaves" of the tree representing units. Where the additive tree differs is that the distance is graphically represented by the distance of those units on the tree.

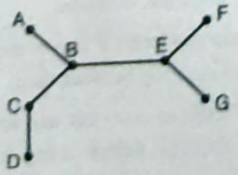


Fig. 5.5.2

Software that can create additive trees

The Matlab command is [add tree.m] and displaying them is [displaytree.m]. One can find more matlab commands.

(2) Canonical Correlation Analysis

- Canonical correlation analysis is used to identify and measure the associations among two sets of variables.
- Canonical correlation is appropriate in the same situations where multiple regression is appropriate.
- Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set and that best explain the variability both within and between sets.

(3) Cluster Analysis

- 'Cluster analysis' or 'clustering' is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).
- It is a main task of exploratory data analysis and a common technique for 'statistical data analysis', used in many fields, including pattern recognition, image analysis, bioinformatics, data compression, computer graphics and machine learning.
- Cluster analysis itself is not one specific algorithm. It can be achieved by various algorithms that differ significantly with respect to what constitutes a cluster and how to find them efficiently.

(4) Correspondence analysis (CA)

- CA is a multivariate statistical technique which is conceptually similar to 'Principal Component Analysis'.
  But it applies to categorical rather than continuous data.
- It provides a means of displaying or summarising a set of data in two-dimensional graphical form. Its aim is to display in a biplot any structure hidden in the multivariate setting of the data table.
- As such it is a technique from the field of multivariate ordination.
- Since the variant of CA described here can be applied either with a focus on the rows or an the columns.
- It is also called as simple (symmetric) correspondence analysis.

(5) Multiple Correspondence Analysis : (MCA)

- Multiple correspondence analysis (MCA) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set.
- It does this by representing data as points in a low dimensional Euclidean space.
- The procedure is thus appears to be the counterpart of principal component analysis for categorical data.
- MCA can be viewed as an extension of simple correspondence analysis (CA) in that it is applicable to a large set of categorical variables.

(6) Factor Algorithm

- Unsolved problem in computer science :
- Can integer factorisation be solved in polynomial time on a classical computer?
- In number theory, integer factorisation is the decomposition of a composite number into a product of smaller integers.

For Example :

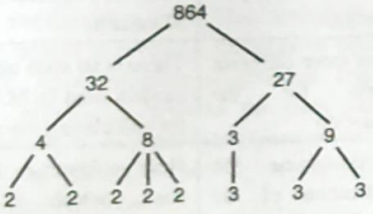


Fig. 5.5.3

5.5.2 Applications of M.A.

1. Multivariate analysis data can be used to process information in a meaningful fashion. These methods can afford **hidden data structures**. On the one hand the elements of measurements often do not contribute to the relevant property and on the other hand hidden phenomena are recorded.
2. The terms multivariable analysis and multivariate analysis are often used interchangeably in **medical and health sciences research.** However, multivariate analysis refers to the analysis of multiple outcomes whereas multivariable analysis deals with only one outcome each time.

5.6 ANOVA

- ANOVA means 'Analysis of Variance' the overall statistical test averages acts as ANOVA. An ANOVA tests the relationship between a categorical and a numeric variable by testing the difference between two or more means. This test produces a **p-value** to determine whether the relation is significant or not.
- Anova of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyse the differences among means.
- Anova is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalises the t-test beyond two means.

5.6.1 Assumptions

(1) ANOVA assumes that the data is normally distributed.

(2) ANOVA also assumes homogeneity of variance, which means that the variance among the groups should be approximately equal.

(3) ANOVA also assumes that the observations are independent of each other.

5.6.2 Steps to perform ANOVA

Step I : Calculate all the means i.e., calculate the sum of squares within groups (SSW).

Then add them (SSW) for all the groups. Hence we have the sum symbol twice in the formula :

Assuming that there are three groups : then SST can be found as :

[Group I – mean – total mean]^2 + [Group II – mean – total mean]^2 + [Group III – mean – total mean]^2

i.e. $SST = SS_{Total} = \sum_{j=1}^{k} \left(\bar{X}j - \bar{X}^2\right)$.

Step 2 : Set up the null and alternate hypothesis :

The null hypothesis assumes that there is no variance data in different groups; i.e. the means are all same.

Module 5

The alternate hypothesis states that the means are different :

i.e. H_0 : (null hypothesis) : $\mu_1 = \mu_2 = \mu_3$

H_1 : (alternate hypothesis)

$\mu_1 \neq \mu_2 \neq \mu_3$

- **Step 3 :** Write down (after calculating) the sum of squares
- **Step 4 :** Calculate the degrees of freedom (DFT) d.f. = (n – 1) :

Where n is the total of all the data sets combined.

Calculate the degrees of freedom within groups (DFW)

i.e. $d.f_{witin} = (k - 1)$

Where k is number of groups

Then calculate the degrees between groups (DFB) i.e. d.

$f_{between} = n - k$

- **Step 5 :** calculate the Mean squares within (MSW)

$$M\,S_{witin} = \frac{S\,S_{within}}{d.\,f_{within}}$$

- **Step 6 :** Calculate F – statistic, using the formula

$$F = \frac{M\,S_{between}}{M\,S_{within}}$$

- **Step 7 :** Observe the tabulated value of F (critical value) from the statistical value and compare it with the value that is being calculated.

If the absolute value is greater than the critical value, we reject the null hypothesis and conclude that there is significant difference between the means of the populations.

Otherwise accept the null hypothesis or fail to reject the null hypothesis.

5.7 BIVARIATE ANALYSIS (B.A)

Bivariate analysis is more analytical than univariate analysis. When the dataset contains two variables and researchers aim to undertake **comparisons** between the two data sets then Bivariate analysis is the right type of analysis technique.

Example of Bivariate Analysis

In a survey of classroom, the researcher may be looking for analysis of : the ratio of students who scored above 85% corresponding to their genders. In this case, there are two variables - Gender X (independent variable) and result Y (dependent variable).

5.7.1 ANOVA V/S MANOVA

- ANOVA and MANOVA are basically two different statistical methods that are used to calculate the mean for a given data. The word ANOVA stands for analysis of variant, while the word MANOVA stands for multivariate analysis of variant.
- The ANOVA method used for calculating mean includes only one dependent variable, while the MANOVA method used for calculating mean includes multiple dependent variables.
- It is basically used to determine if there is any difference in the variant groups or if there is more than one dependent variable present. And this is how it is different from ANOVA in one way, which requires only one variable.

Sr. No.	ANOVA	MANOVA
1.	There is only one variable (dependent) for calculating mean.	There are multiple variables for the calculation of the mean.
2.	ANOVA is analysis of variant	It is multivariate analysis of variants
3.	It uses three different models for the calculation	There is no such number of models used in MANOVA for calculating the mean.
4.	To determine the significance of the factor F-test is used.	Here multivariate F-test is used, which is called Wilk's Lambda test.
5.	The comparison of the factor variance to the error variance decides the value of F in the ANOVA.	The factor variance-covariance matrix is compared to the error variance-covariance matrix in order to obtain Wilk's Lambda.

Conclusion

From the discussion so far, it is to be concluded that ANOVA and MANOVA are basically two different statistical methods that are used to calculate the mean for a given data. The word ANOVA stands for analysis of variant, while the word MANOVA stands for multivariate analysis of variant.

The ANOVA method used for calculating mean includes only one dependent variable, while the MANOVA method used for calculating mean includes multiple

dependent variables. It is basically used to determine if there is any difference in the variant groups or if there is more than one dependent variable present. And this is how it is different from ANOVA in one way, which requires only one variable.

ANOVA has three different models that are used in different aspects to calculate the mean. A fixed-effect model is applied when the object is subjected to be having one or even more than one treatment. The random effect model is applied when the treatment that is applied is not fixed before for the subject in the large population. A mixed-effect model is applied when the treatment has both the previous methods, the fixed one and the mixed one too.

Multivariate analysis (MANOVA) extends the capabilities of ANOVA by assessing **multiple dependent variables simultaneously**. ANOVA statistically tests the difference between three or more group means. This statistical procedure tests multiple dependent variables at the same time. For example, if you have 3 different teaching methods and you want to evaluate the average scores for these groups, you can use ANOVA.

But ANOVA has drawbacks. It can assess only one dependent variable at a time. This limitation can be an enormous problem in certain circumstances because it can prevent you from detecting effects that actually exist. MANOVA provides a solution for such studies. This statistical procedure tests multiple dependent variable at the same time. By doing so, MANOVA can offer several advantages over ANOVA.

Manova is used for

- The one-way multivariate analysis of variance (one-way manova) is used to determine whether there are any differences between independent groups on more than one continuous dependent variable.
- In this regard, it differs from a one-way ANOVA, which only measures one dependent variable.
- One can use a one-way MANOVA to understand whether there were differences in students short-term and long-term recall of facts based on three different lengths of lecture (i.e., the two dependent variables are "short-term memory recall" and "long-term memory recall", while the independent variable is "lecture-duration", which has four independent groups : ""30 minutes", "60 minutes", "90 minutes" and "120 minutes").

Remark

- If you have to independent variables rather than one, you can run a 'two way' MANOVA.
- Alternatively, if you have one independent variable and a continuous covariance, you can run a one-way MANOVA.
- In addition, if your independent variable consists of repeated measures, you can use the one-way repeated measures MANOVA.

MANOVA is used instead of ANOVA

The correlation structure between the dependent variables provides additional information to the model which gives MANOVA the following capabilities : 'Greater Statistical Power.' When the dependent variables are correlated, MANOVA can identify effects that are smaller than those that regular ANOVA can find.

Benefits of MANOVA and when to use it

- Multivariate ANOVA (MANOVA) extends the capabilities of analysis of variance (ANOVA) by assessing multiple dependent variables simultaneously.
- ANOVA statistically tests the differences between three or more group means.
- For example, if you have three different teaching methods and you want to evaluate the average scores for these groups, you can use ANOVA.
- However, ANOVA does have a drawback. It can assess only one dependent variable at a time.
- This limitation can be a problem in certain circumstances because it can prevent you from detecting effects that actually exist.
- MANOVA provides a solution for some studies. This statistical procedure tests multiple dependent variables at the same time. By doing so, MANOVA can offer several advantages over ANOVA.

Module 5

5.8 GRAPHICAL FORMAT

Graphic images are stored digitally using a small number of **standardised graphic file formats** including **bit map, TIFF (Tagged image file format), JPEG, PNG (Portable network graphic)**; they can also be stored as raw, unprocessed data. Many graphic are created as vector graphics and then published as raster images.

5.8.1 Common Image File Format

- TIFF, GIF or graphic interchange format files are widely used for **web-graphics**, because they are limited to only **256 colours**, can allow for transparency and can be animated.
- **Most common graphic format :** GIF format (gif file extensional) is one of the two most common file formats for image on the world wide web (www) because it is supported by almost all web browsers.

5.8.2 JPEG File (Joint Photographic Experts Group)

- JPEG file (Joint Photographic Experts Group) is a standard image format for containing **lossy and compressed image data**.
- JPEG file can also contain high-quality image data with compression. In paint shop, **pro-JPEG** is a commonly used format for storing the edited image.

5.8.3 3 Types of Graphic Formats

- There are a number of different types of graphics file formats. Each type stores graphics data in a different way. **Bitmap, vector** and **metafile formats** are most commonly used formats.
- Bitmap, format, also known as bitmap image file is a **vaster graphics image** file format used to store bitmap digital images, independently of the display device.
- Vector graphics, as a form of computer graphics, is the set of mechanisms for creating visual images directly from geometric shapes defined on a Cartesian plane, such as points, lines, curves and polygons. A true vector image can be scaled to no end, with no pixels or distortion.

5.8.4 Metafile Format

- Windows metafile is an image file format originally designed for Microsoft windows in 1990. Essentially, a metafile stores a list of records consisting of **drawing commands, property definitions** and **graphics object** to display an images on screen.
- A metafile contains specifications for another file. It is commonly associated with digital graphics, particularly vector images. However metafiles can contain other formats as well, such as bitmaps or other data.

WMF Header Record
WMF Record
...
WMF Record
WMF End-of-File Record

WMF placeable Header Record
WMF Header Record
WMF Record
...
WMF Record
WMF End-of-file record

Fig. 5.8.1 : Structures of original and placeable Windows metafiles

5.9 CORRELATION ANALYSIS

5.9.1 Correlation Coefficient

- In correlation analysis, the degree of relationship between two variables, say X and Y, is measured by a single number r called a 'correlation coefficient'.
- Here both X and Y are random variables. In regression analysis dependent variable Y is assumed to be a random variable but the regress or (independent) variable X is not a random variable. It may be a mathematical, physical or scientific variable.

Examples of correlation coefficient

(i) Rainfall and crop-yield correlated.

(ii) Two coins being tossed simultaneously, uncorrelated.

5.9.2 Properties of Correlation Coefficient

GQ. Discuss the properties of correlation coefficient.

(i) Correlation coefficient is a pure number, i.e. It has no unit.

(ii) The correlation coefficient r ranges from **– 1 to 1**.

(iii) The correlation between two variables is known as **simple correlation** or correlation of zero order.

(iv) It is not affected by coding (linear transformation) of variables or variate values.

(v) The relation between the correlation coefficient 'r' and the regression coefficient b_{YX} and b_{XY} is

$$r = \sqrt{b_{YX} \cdot b_{XY}}$$

(vi) The sign of r will be the same for b_{YX} or b_{XY}.

(vii) r^2, the **square of correlation coefficient** is referred to as coefficient of determination and $(1 - r^2)$ as coefficient of non-determination.

(viii) If the two variables are independent, the correlation coefficient between them is zero, but the converse is not true.

5.9.3 Types of Correlation

1. Positive or direct
2. Negative or inverse
3. Linear
4. Non linear

By plotting a given set of pairs of random variables, (X_i, Y_i), $i = 1$ to n, as a scatter diagram, the correlation is said to be

1. **Positive or direct :** If Y increases, X also increases.
2. **Negative or inverse :** If Y decreases, X increases.
3. **Linear :** If all the points lie near a straight line.
4. **Non-linear :** If all the points lie on some non-linear curve.

Examples of correlation

1. Age and physical health are negatively correlated.
2. Income and expenditure positively correlated.

5.9.4 Simple Correlation

Correlation between **two variables** is said to be simple correlation.

Multiple correlation

The correlation between **more than two variables** is called as multiple correlation.

If $r = \pm 1$, there is a perfect positive (or negative) correlation. If $r = 0$, there is no linear correlation, but a non-linear correlation may exist.

A high correlation due to a third variable is known as a **spurious correlation**. For example, poverty and crime are highly correlated but the spurious correlation is because of variable illiteracy.

5.9.5 Standard Error of Estimate

Y on X is denoted by $S_{Y,X}$ is defined as,

$$S_{Y,X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}}$$

Where Y_{est} is the estimated or predicted value of Y from the least squares regression line $Y = a_0 + a_1 X$.

Similarly, the standard error of estimate X on Y is.

$$S_{Y,X} = \sqrt{\frac{\sum (X - Xest)^2}{N}}$$

In general, $S_{Y,X} \neq S_{X,Y}$.

Result : We note that,

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{est})^2 + \sum (y_{est} - \bar{Y})^2$$

i.e. Total variation = unexplained variation + explained variation.

It can be shown that

$$r = \pm \sqrt{\frac{\text{Explained variation}}{\text{Unexplained variation}}}$$

i.e., $$r = \pm \sqrt{\frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - Y_{est})^2}}$$

The positive and negative signs correspond to positive and negative correlation respectively.

5.9.6 Properties of r

(i) r lies in the interval $[-1, 1]$, i.e. $-1 \leq r \leq 1$

(ii) r is independent of origin.

(iii) r is independent of scale of measurement unit.

5.9.7 Karl Pearson Product-Moment

We develop the formula for correlation coefficient r:

Let $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$ be the least squares regression lines.

Let $y = Y - \bar{Y}$ and $x = X - \bar{X}$

Then it can be shown that

$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2) - (\sum y^2)}}$$

The $\pm$ sign can be omitted without any loss of generality since y_{est} increases or decreases as x increases or decreases.

Module 5

$$\therefore \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

We introduce covariance S_{XY} of X and Y by

$$S_{XY} = \frac{\sum xy}{N}$$

And s.d. of X and Y by,

$$S_X = \sqrt{\frac{\sum x^2}{N}}, S_Y = \sqrt{\frac{\sum y^2}{N}}$$

$$\therefore \quad r = \frac{S_{XY}}{S_X \cdot S_Y}$$

☞ **Computational formula**

$$r = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{[N(\sum X^2) - (\sum X)^2][N(\sum Y^2) - (\sum Y)^2]}}$$

☞ **Regression lines and the linear correlation coefficient**

The least squares regression line Y on X

$Y = a_0 + a_1 X$ can be written as

$$Y = \left(\frac{\sum xy}{\sum x^2}\right) x$$

Where $y = Y - \bar{Y}$ and $x = X - \bar{X}$

$$\therefore \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

$$\therefore \quad \frac{\sum xy}{\sum x^2} = \frac{r\sqrt{\sum x^2 \cdot \sum y^2}}{\sum x^2} = r\sqrt{\frac{\sum y^2}{\sum x^2}}$$

$$= r\frac{Sy}{Sx}$$

$$\therefore \quad y = a_1 x = \left(\frac{\sum xy}{\sum x^2}\right) x$$

In a similar way; we can have

$$x = r\left(\frac{S_x}{S_y}\right) y$$

Ex. 5.9.1 : Show that the coefficient of correlation r is the geometric mean between the regression coefficients.

☑ **Soln. :**

Let $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$ be the least squares regression lines with a_1 and b_1 as regression coefficients.

We have that, $a_1 = r\frac{S_r}{S_x}$ and $b_1 = r\frac{Sx}{Sy}$

$$\therefore \quad a_1 b_1 = \left(r\frac{Sy}{Sx}\right)\left(r\frac{Sx}{Sy}\right) = r^2$$

$$\therefore \quad r = \sqrt{a_1 b_1}$$

$\therefore$ r is the G.M. between the regression coefficient.

5.9.8 Examples on Correlation Coefficient

Ex. 5.9.2 : (a) Predict the blood pressure (B. P.) of a woman of age 45 years from the following data which shows the ages X and systolic B.P. Y of 12 women. (b) Are the two variable ages X and B.P. Y correlated ?

Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
B.P. (Y)	147	125	160	118	149	128	150	145	115	140	152	155

☑ **Soln. :**

(a) Let the least squares regression equation of Y as X be $Y = a_0 + a_1 X$

Its normal equations are

$$\sum Y = Na_0 + a_1 \sum X$$

$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

From the given data, we have

$N = 12$, $\sum X = 628$, $\sum Y = 1684$, $\sum X^2 = 34416$,

$\sum Y^2 = 238822$, $\sum XY = 89894$

Substituting,

$$1684 = 12\,a_0 + 628\,a_1$$

$$89894 = 628 a_0 + 34416\,a_1$$

Solving, $a_0 = 80.777$, $a_1 = 1.138$

$\therefore$ The prediction equation is

$$Y = 80.777 + 1.138\,X$$

The B.P. of a woman with age X = 45 is

$$Y = 80.777 + 1.138\,(45) = 131.987 \simeq 132$$

(b) We determine correlation coefficient r to find the association between age and B.P.

$$\text{Now, } r = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2]\,[N\sum (Y^2) - (\sum Y)^2]}}$$

$$= \frac{12\,(89894) - (628)\,(1684)}{\sqrt{[(12)\,(34416) - (628)^2]\,[(12)\,(238822) - (1684)^2]}}$$

$$= 0.8961$$

$\therefore$ Age X and B.P. Y are strongly positively correlated.

Ex. 5.9.3 : For the following data determine

(a) Least squares regression line of y on x.

(b) y (3). (c) least squares regression line of x on y.

(d) x (4). (e) S_{yx}

(f) S_{xy} (g) Total variation in y

(h) unexplained variation in y

(i) explained variation in y.

x	6	5	8	8	7	6	10	4	9	7
y	8	7	7	10	5	8	10	6	8	6

☑ **Soln. :**

x	y	x^2	y^2	xy	
6	8	36	64	48	$\sum x = 70$
5	7	25	49	35	$\sum y = 75$
8	7	64	49	56	$\sum x^2 = 520$
8	10	64	100	80	$\sum y^2 = 587$
7	5	49	25	35	$\sum xy = 540$
6	8	36	64	48	$N = 10$
10	10	100	100	100	
4	6	16	36	24	
9	8	81	64	72	
7	6	49	36	42	
70	75	520	587	540	

(a) Let the L.S.R,L. equation of y on x be

$$y = a_0 + a_1 x$$

Normal equations are

$$\sum y = a_0 N + a_1 \sum x$$

$$\sum xy = a_0 \sum x + a_1 \sum x^2$$

$$\therefore 75 = 10\,a_0 + 70\,a_1$$

$$540 = 70\,a_0 + 520\,a_1$$

Solving $a_1 = 0.5,\ a_0 = 4$

$\therefore$ L.S.R.L. of y on x is

$$y = 4 + 0.5\,x \qquad \ldots(1)$$

(b) $y(3) = 4 + 0.5\,(3) = 4 + 1.5 = 5.5$

(c) Let L.S.R.L. of x on y be

$$x = b_0 + b_1 y$$

With normal equations

$$\sum x = Nb_0 + b_1 \sum y$$

$$\sum xy = b_0 \sum y + b_1 \sum y^2$$

$$\therefore 70 = 10\,b_0 + 75 b_1$$

$$540 = 75 b_0 + 587\,b_1$$

Solving, $b_1 = 0.612, \quad b_0 = 2.41$

$\therefore$ L.S.R.L. of x on y is

$$x = 2.41 + 0.612\,y \qquad \ldots(2)$$

(d) $x\,(4) = 2.41 + 0.612\,(4) = 4.858$

(e) From Equation (1) estimated value of $y = Y_{est} = 4 + 0.5\,x$

x	y	y_{est}	$y - y_{est}$	$(y - y_{est})^2$
6	8	7	1	1
5	7	6.5	0.5	0.25
8	7	8	– 1	1
8	10	8	2	4
7	5	7.5	– 2.5	6.25
6	8	7	1	1
10	10	9	1	1
4	6	6	0	0
9	8	8.5	– 0.5	0.25
7	6	7.5	–1.5	2.25
			Total	**17.0**

Module 5

Now, $S_{yx} = \sqrt{\frac{\sum (y - y_{est})^2}{N}} = \sqrt{\frac{17.0}{10}} = \sqrt{1.7} = 1.304$

(f) From (2), $x_{est} = 2.41 + 0.612\,y$

y	x	x_{est}	$x - x_{est}$	$(x - x_{est})^2$
8	6	7.306	–1.306	1.7056
7	5	6.694	–1.694	2.8696
7	8	6.694	1.306	1.7056
10	8	8.53	–0.53	0.2809
5	7	5.47	1.53	2.3409
8	6	7.306	– 1.306	1.7056
10	10	8.53	1.47	2.1609
6	4	6.082	– 2.082	4.3347
8	9	7.306	1.694	2.8696
6	7	6.082	0.918	0.8427
			Total	**20.816**

Now, $S_{xy} = \sqrt{\frac{\sum (x - x_{est})^2}{N}} = 1.443$

(g, h, i) : $\bar{y} = \frac{\sum y}{N} = \frac{75}{10} = 7.5$

Total variation = unexplained variation + explained variation

$$\sum (y - \bar{y})^2 = \sum (y - y_{est})^2 + \sum (y_{est} - \bar{y})^2 \quad ...(3)$$

$(y_{est} - \bar{y})$	0.5	–1	0.5	0.5	0	–0.5	1.5	–1.5	1	0	Total
$(y_{est} - \bar{y})^2$	0.25	1	0.25	0.25	0	0.25	2.25	2.25	1	0	7.50
$(y - \bar{y})$	0.5	–0.5	–0.5	2.5	–2.5	0.5	2.5	–1.5	0.5	–1.5	
$(y - \bar{y})^2$	0.25	0.25	0.25	6.25	6.25	0.25	6.25	2.25	0.25	2.25	24.50

Total Variation = $\sum (y - \bar{y})^2 = 24.50$

Explained variation = $\sum (y_{est} - \bar{y})^2 = 7.50$

∴ From (3), unexplained variation

$= \sum (y - y_{est})^2 = 24.50 - 7.50$

$= 17.0$

Ex. 5.9.4 : Calculate the coefficient of correlation r for the above data in 3 ways.

☑ **Soln. :**

(1) We have $r = \pm \sqrt{\frac{\text{explained variation}}{\text{total variation}}}$

$= \sqrt{\frac{7.50}{24.50}} = 0.553$

(2) Correlation coefficient is the geometric mean between the regression coefficients, i.e.

$r = \sqrt{bxy \,.\, byx}$

Now, from (i), the regression coefficient of y on x is bxy = 0.5

And regression coefficient of x on y is bxy = 0.612

$\therefore \quad r = \sqrt{(0.5)(0.612)} = 0.5532$

(3) By product – moment formula.

$$r = \frac{N\sum xy - (\sum x) \cdot (\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2]\,[N\sum (y^2) - (\sum y)^2]}}$$

$$= \frac{10\,(540) - (70)\,(75)}{\sqrt{[10\,(520) - (70)^2]\,[10\,(587) - (75)^2]}}$$

$$= \frac{150}{\sqrt{73500}} = 0.5513$$

Ex. 5.9.5 : From 10 pairs of observations for x and y the following data is obtained :

$n = 10, \sum x = 66, \sum y = 69, \sum x^2 = 476, \sum y^2 = 521,$

$\sum xy = 485$. It was later found that two pairs of (correct) values.

x	y	Were copied down as	x	y
4	6		2	3
9	8		7	5

Calculate the correct value of the coefficient of correlation

☑ **Soln. :**

To obtain the correct data, we subtract the incorrect data and add the correct data as :

$\sum x = 66 - 2 - 7 + 4 + 9 = 70$

$\sum y = 69 - 3 - 5 + 6 + 8 = 75$

$$\sum x^2 = 476 - 4 - 49 + 16 + 81 = 520$$

$$\sum y^2 = 521 - 9 - 25 + 36 + 64 = 587$$

$$\sum xy = 485 - 6 - 35 + 24 + 72 = 540$$

Now, $$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

$$= \frac{10(540) - (70)(75)}{\sqrt{[10(520) - (70)^2][10(587) - (75)^2]}}$$

$$= 0.5533$$

Exercise

1. Compute r for the data given below

X :	1	2	3	4	5	6
Y :	6	4	3	5	4	2

Hints : $N = 6, \sum X = 21, \sum Y = 24, \sum X^2 = 91, \sum Y^2 = 106, \sum XY = 75$

Ans. : r = – 0.68

2. Determine r for the following data :

X :	50	60	70	90	100
Y :	65	51	40	26	8

Ans. : – 0.99

5.10 RANK CORRELATION

Sometimes data turns out to be non-numeric i.e. it is **qualitative, For example** :

1. **Appearance :** Beautiful, ugly
2. **Efficiency :** Excellent, good, average, bad etc.

- In such cases data is ranked according to that particular character and not according to numeric measurements on them. And hence correlation coefficient cannot be calculated in the usual manner.
- Hence Charles Edward Spearman developed a nonparametric counterpart of correlation coefficient as follows : For a given set of n paired observations (X_i, Y_i), for **i = 1 to n**; ranks 1, 2, ..., n are assigned to the **X-observations** in order of **magnitude** and similarly to the **Y-observations**.
- Then these ranks are substituted for the actual numerical values, and correlation coefficient is calculated, which is called as **'Rank correlation coefficient'** and is given by,

$$r_{rank} = 1 - \left[\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}\right]$$

where d_i = difference between ranks assigned to X_i and Y_i

n = number of pairs of data.

Remarks

(i) r lies between – 1 and 1.

(ii) If there are ties among either X or Y observations, substitute for each of the tied observations, the mean of the ranks that they jointly occupy.

5.10.1 Example On Rank Correlation

Ex. 5.10.1 : Determine rank correlation for the following data which shows the marks obtained in two quizes in mathematics:

Marks in 1st quiz (X) :	6	5	8	8	7	6	10	4	9	7
Marks in 2nd quiz (Y) :	8	7	7	10	5	8	10	6	8	6

Soln. :

Assigning ranks to the data of X, we get

X :	4	5	6	6	7	7	8	8	9	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2	3.5	3.5	5.5	5.5	7.5	7.5	9	10
Similarly Y :	5	6	6	7	7	8	8	8	10	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2.5	2.5	4.5	4.5	7	7	7	9.5	9.5

Data assigned with ranks is

X :	3.5	2	7.5	7.5	5.5	3.5	10	1	9	5.5
Y :	7	4.5	4.5	9.5	1	7	9.5	2.5	7	2.5
D :	– 3.5	–2.5	3	–2	4.5	–3.5	0.5	–1.5	2	3
D^2 :	12.25	6.25	9	4	20.25	12.25	0.25	2.25	4	9

$$\sum D^2 = 79.5$$

$$\text{Rank correlation} = 1 - \left[\frac{6\sum D^2}{n(n^2 - 1)}\right]$$

Module 5

$$= 1 - \left[\frac{6\,(79.5)}{10\,(99)}\right] = 1 - 0.4818$$

$$= 0.5182$$

Exercise

1. Find the rank correlation for the following data :

X :	2	4	5	6	8	11
Y :	18	12	10	8	7	5

Ans. : rank $= 1 - \frac{6\,(70)}{6(35)} = -1$

For Examples

Consider the following paired observations :

X :	–3	–2	–1	0	1	2	3
Y :	9	4	1	0	1	4	9

For this data, $\sum X = 0$, $\sum Y = 0$, $\sum XY = 0$.

$$\therefore r_{XY} = 0$$

but the data clearly reveals that, $Y = X^2$, which means that X and Y are dependent.

5.11 ANALYSIS OF VARIANCE (ANOVA)

Introduction

The analysis of variance is a powerful **statistical tool test** of significance. The term 'Analysis of Variance' was introduced by Prof. R.A. Fisher to deal with problems in the analysis of agronomical data.

5.11.1 Definition of ANOVA

- Analysis of variation is the **'separation of variance ascribable to one group of causes from the variance ascribable to the other group".**
- By this technique the total variation in the sample data is expressed as the sum of its **non-negative components** where each of these components is a measure of the variation due to some specific independent source or factor or cause.
- The ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to assignable factors (causes) with the estimate due to chance factors (causes).

5.11.2 Assumptions for ANOVA Test

ANOVA test is based on **statistic F-test** for variance Ratio. For the validity of the F-test in ANOVA, the following assumptions are made:

(i) The observations are independent,

(ii) Parent population from which observations are taken are additive in nature.

(iii) Various treatment and environmental effects are additive in nature.

5.12 HYPOTHESIS TESTING FOR MORE THAN TWO MEANS (ANOVA)

5.12.1 Rejection Region Method

We carry out the test for the equality of several (k) population means by the Rejection Region method. The various steps are given below.

▶ **Step 1 :** Set up the hypothesis

Null Hypothesis : $H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_k$,

i.e. all the means are equal.

Alternate Hypothesis

H_1 : At least two means are different.

▶ **Step 2 :** Compute the means and standard deviations for each class by the formulae

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \; ;$$

$$\sigma_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right)^2 ; \; i = 1, 2, \ldots, k$$

Also compute the mean $\bar{X}$ of all the data observations in the k-classes by the formula

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} \left(\sum_{j=1}^{n_j} X_{ij} \right) = \frac{\sum_i n_i \bar{X}_i}{\sum_i n_i}$$

▶ **Step 3 :** Obtain the 'between classes' sum of squares (BSS) by the formula

$$BSS = n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + ... + n_k(\bar{X}_k - \bar{\bar{X}})^2$$

- **Step 4 :** Obtain the 'between classes' mean sum of squares (MBSS)

$$MBSS = \frac{\text{Between classes S.S.}}{\text{Degrees of freedom}} = \frac{BSS}{k-1}$$

- **Step 5 :** Obtain the 'within classes' sum of squares (WSS) by the formula

$$WSS = \sum_i \sum_j (X_{ij} - \bar{X}_i) = \sum_{i=1}^{k} n_i S_1^2$$

$$= n_1 s_1^2 + n_2 s_2^2 + ... + n_k s_k^2$$

- **Step 6 :** Obtain the within classes mean sum of squares (MWSS)

$$MWSS = \frac{\text{Within classes S.S.}}{\text{Degrees of freedom}} = \frac{WSS}{n-k}$$

- **Step 7 :** Obtain the test statistic F or Variance Ratio (V.R.)

$$F = \frac{\text{Between clases MSS}}{\text{Within classes MSS}}$$

$$= \frac{\text{step (4)}}{\text{step (6)}} \sim F(k-1, n-k)$$

Which follows F-distribution with
$(\gamma_1 = k-1,\ \gamma_2 = n-k)$ degrees of freedom.

- **Step 8 :** Find the critical value of the test statistic F, for $(k-1, n-k)$ d.f. and at desired level of significance, say α, from the given table.
- **Step 9 :** Write down the conclusion.

5.12.2 Alternative for Computation of Various Sums of Squares

- **Step 1 :** Compute $G = \sum_i \sum_j X_{ij}$ = Grand total of all observations.
- **Step 2 :** Compute Correction factor (C.F.) $= \frac{G^2}{n}$

Where $n = n_1 + n_2 + ... + n_k$,

(1) Total number of observations

- **Step 3 :** Compute Raw sum of squares

$$(RSS) = \sum_i \sum_j X_{ij}^2$$

= sum of squares of all observations

- **Step 4 :** Total S.S.

$$= \Sigma\Sigma (X_{ij} - \bar{\bar{X}})^2 = RSS - CF$$

- **Step 5 :** Compute $T_i = \sum_{j=1}^{n_i} X_{ij}$ = Sum of the all observations in i^{th} class; $(i = 1, 2, ..., k)$
- **Step 6 :** Between classes (or Treatment) S.S.

$$S.S. = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C.F.$$

$$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + ... + \frac{T_k^2}{n_k} - C.F.$$

- **Step 7 :** Within classes or Error S.S.

= Total S.S. – Between classes S.S.

5.13 SOLVED EXAMPELS

Ex. 5.13.1 : A trucking company wishes to test the average life of each of the four brands of tyres. The company uses all the brands on randomly selected trucks. The records showing the lives (thousands of miles) of tyres are as given in the table.

Test the hypothesis that the average life for each brand of tyres is the same. Assume $\alpha = 0.01$.

Table Ex. 5.13.1

Brand 1	Brand 2	Brand 3	Brand 4
20	19	21	15
23	15	19	17
18	17	20	16
17	20	17	18
	16	16	

Soln. :

Here, the factors of variation are brands of tyres.

Set up the hypothesis

Null Hypothesis : $\mu_1 = \mu_2 = \mu_3 = \mu_4$,

i.e. the mean life of the tyres of all the brands is same.

Alternative Hypothesis : At least two means are different.

Module 5

	Brand 1	Brand 2	Brand 3	Brand 4	
	20	19	21	15	
	23	15	19	17	
	18	17	20	16	
	17	20	17	18	
Total $T_i = \sum_j X_{ij}$	$T_1 = 78$	$T_2 = 87$	$T_3 = 93$	$T_4 = 66$	$G = \Sigma\Sigma X_{ij} = 324$
T_i^2	$T_1^2 = 6084$ $n_1 = 4$	$T_2^2 = 7569$ $n_2 = 5$	$T_3^2 = 8649$ $n_3 = 5$	$T_4^2 = 4356$ $n_4 = 4$	

G = Grand total

$= 78 + 87 + 93 + 66 = 324$

$n = n_1 + n_2 + n_3 + n_4$

$= 4 + 5 + 5 + 4 = 18$

$\therefore$ Correction factor (C.F.) $= \frac{G^2}{n} = \frac{(324)^2}{18}$

$= \frac{104976}{18} = 5832$

Raw sum of squares (RSS) $= \sum X_{ij}^2$

$= (400 + 529 + 324 + 289) + (361 + 225 + 289 + 400 + 256)$

$+ (441 + 361 + 400 + 289 + 256) + (225 + 289 + 256 + 324)$

$= 1542 + 1531 + 1747 + 1094 = 5914$

$\therefore$ Total sum of squares = (TSS)

$= RSS - CF = 5914 - 5832 = 82$

Between Samples (brands of tyres)

Sum of squares (BSS) is given by

$$BSS = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - C.F.$$

$$= \frac{6084}{4} + \frac{7569}{5} + \frac{8649}{5} + \frac{4356}{4} - 5832$$

$= (1521 + 1513.8 + 1729.8 + 1089) - 5832 = 21.6$

Within samples (Error) Sum of squares (WSS)

$= TSS - BSS = 82 - 21.6 = 60.4$

ANOVA Table

Sources of Variation	d.f.	Sum of Squares	Mean S.S.	Variance ratio
Between brands of tyres	$4 - 1 = 3$	21.6	$\frac{21.6}{3} = 7.2$	$\frac{7.2}{4.31} = 1.67$
Error	$17 - 3 = 14$	60.4	$\frac{60.4}{14} = 4.31$	-
Total	$18 - 1 = 17$	82	-	-

☞ **Critical Value**

The critical (tabulated) value of F ($\gamma_1 = 3, \gamma_2 = 14$) d.f. at $\alpha = 0.01$ is 5.56 (from table)

Since the calculated value of the test statistic F = 1.67 is less than the critical value, it is not significant, i.e. it does not fall in the rejection region. Hence the null hypothesis H_0 is to be rejected.

☞ **Conclusion**

There is no significant difference between the average lives of the four brands of tyres 1, 2, 3 and 4.

Ex. 5.13.2 : Given the following data, test the hypothesis :

H_0 : All the means are equal

H_1 : At least two means are different

$\bar{X}_1 = 27$, $\sigma_1 = 8$, $n_1 = 4$; $\bar{X}_2 = 25$, $\sigma_2 = 9$, $n_2 = 7$;

$\bar{X}_3 = 28$, $\sigma_3 = 5$, $n_3 = 5$ and $\alpha = 0.05$

☑ **Soln. :**

As we are given the values of the sample means ($\bar{X}$), S.D.^{s} (σ) and sizes (n) =, we can directly compute 'Between classes (Samples) S.S.' and 'Within Classes (Samples) S.S.'

The overall mean $\bar{X}$ is given by

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + n_3 \bar{X}_3}{n_1 + n_2 + n_3} = \frac{4(27) + 7(25) + 5(28)}{4 + 7 + 5}$$

$$= \frac{108 + 175 + 140}{16} = \frac{423}{16} = 26.44$$

Between Samples S.S. (B.S.S.)

$$BSS = n_1 \left(X_1 - \bar{X} \right)^2 + n_2 \left(X_2 - \bar{X} \right)^2 + n_3 \left(X_3 - \bar{X} \right)^2$$

$$= 4(27 - 26.44)^2 + 7(25 - 26.44)^2 + 5(28 - 26.44)^2$$

$$= 4(0.3136) + 7(2.0736) + 5(2.4336)$$

$$= 1.2544 + 14.5152 + 12.1680$$

$$= 27.9376 \doteqdot 27.94$$

☞ **Within samples (Error) S.S. (W.S.S)**

d.f. for between samples S.S.= K – 1 = 3 – 1 = 2

d.f. for within samples (Error) = n – k = (4 + 7 + 5) – 3 = 13

ANOVA Table

Sources of Variation	d.f.	Sum of Squares	Mean S.S.	Variance Ratio (F)
Between Samples	3 – 1 = 2	21.6	$\frac{27.94}{2}$ = 13.97	$\frac{13.97}{72.92} = 0.19$ ~ F (2, 13)
Within Samples	15 – 2 = 13	948	$\frac{948}{13} = 72.92$	–
Total	16 – 1 = 15	–	–	–

The tabulated (Critical) value of F for (2, 13) d.f. at 5% level of significance is 3.80 (from the table). Since the calculated value of F = 0.19 is less than the critical value, it doesn't lie in the rejection region. Hence the null hypothesis H_0 can be accepted.

We conclude that null hypothesis of equality of means may be taken as true.

Ex. 5.13.3 : A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in producing a certain output. Five hourly production figures are observed at random from each machine and the results are given in the table. Use analysis of variance technique and determine whether the machines are significantly different in their mean speeds. Use $\alpha = 0.05$.

Table Ex. 5.13.3

	Machine A_1	Machine A_2	Machine A_3
Observations	25	31	24
	30	39	30
	36	38	28
	38	42	25
	31	35	28

☑ **Soln. :**

Here, the factor of variation is machines (A1, A2, A3).

We set up the hypothesis.

H_0 : Null hypothesis : $\mu_1 = \mu_2 = \mu_3$;

i.e. all machines are equally effective.

H_1 : Alternate Hypothesis

At least two of the means are not equal.

In the usual notation, we have

$n_1 = n_2 = n_3 = 5$; $k = 3$; $n = n_1 + n_2 + n_3 = 15$

Machine	Sample Observation (X_{ij})					Total	
A_1	25	30	36	38	31	$T_1 = 160$	$T_1^2 = 25600$
A_2	31	39	38	42	35	$T_2 = 185$	$T_2^2 = 34225$
A_3	24	30	28	25	28	$T_3 = 135$	$T_3^2 = 18225$
					Total	$G = 480$	$\Sigma T_i^2 = 78050$

Module 5

$$\text{Raw S.S. (RSS)} = \sum_i \sum_j X_{ij}^2$$

$$= 25^2 + 30^2 + \ldots + 25^2 + 28^2$$

$$= 15810$$

$$G = \sum_i \sum_j X_{ij}^2 = 160 + 185 + 135 = 480$$

$$\text{Correction factor (C.F.)} = \frac{G^2}{n} = \frac{(480)^2}{15}$$

$$= \frac{230400}{15} = 15360$$

$$\text{Total S.S. (TSS)} = \text{RSS} - \text{CF}$$

$$= 15810 - 15360 = 450$$

$$\text{Treatment S.S. } (S_T^2) = \frac{\Sigma T_i^2}{5} - \text{C.F.}$$

$$= \frac{78050}{5} - 15360$$

$$= 15610 - 15360 = 250$$

$$\text{Error S.S.} = \text{TSS} - \text{SST}$$

$$= 450 - 250 = 200$$

$$\text{Test statistic : } F = \frac{MSST}{MSSE} \sim F(2,12)$$

Sources of Variation	Sum of Squares	d.f.	MSS = s.s./d.f.	Variance Ratio (F)
Treatment (Machines)	250	3 – 1 = 2	MSST = 250/2 = 125	F = 125/16.67 = 7.4985 ~ F(2,12)
Error	200	14 – 2 = 12	MSSE = 200/12 = 16.67	–
Total	450	15 – 1 = 14	–	–

The critical (tabulated) value of F for (2,12) d.f. and $\alpha = 0.05$ level of significance is 3.89 (from the table).

Since the computed value of test statistic F is significant, so we reject H_0 at 5% level of significance and conclude with 95% confidence that the treatment (machine) A_1, A_2, A_3 differ significantly.

Ex. 5.13.4 : Following are the weekly sale records (in thousand Rs.) of three salesmen A, B and C of a company during 13 sale-cells

Table Ex. 5.13.4

A	300	400	300	500	
B	600	300	300	400	
C	700	300	400	600	500

Test whether the sales of three salesmen are different.

$F_{2/10, 0.05} = 4.10$, $F_{2,13,0.05} = 3.81$

☑ **Soln. :**

Since the variation ratio is independent of change of origin and scale, we change the scale conveniently in the original values (X_{ij}) to $V_{ij} = X_{ij} + 100$. This will not affect the final result.

In the usual notation we have, $n_1 = 4, n_2 = 4, n_3 = 5$

$$N = n_1 + n_2 + n_3 = 13,$$

$$G = \sum\sum V_{ij} = 56$$

Decoded data ($V_{ij} = X_{ij} + 100$)						Total
A	3	4	3	5		$R_1 = 15$
B	6	3	3	4		$R_2 = 16$
C	7	3	4	6	5	$R_3 = 25$
						$G = \sum\sum V_{ij} = 56$

Now, We have

$$\text{Correction factor} = \frac{G^2}{N} = \frac{(56)^2}{13} = 241.23$$

$$\text{Raw sum of squares (RSS)} = \sum\sum Y_{ij}^2$$

$$= 9 + 16 + 9 + 25 + 36 + 9 + 9 + 16 + 49 + 9 + 16 + 35 + 25$$

$$= 264$$

$\therefore$ Total sum of squares TSS

$$= \text{RSS} - \text{CF}$$

$$= 264 - 241.23 = 22.77$$

Sum of squares due to rows (salesmen) is given by

$$\text{SSR (salesman)} = \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} - \text{C.F.}$$

$$= \frac{15^2}{4} + \frac{16^2}{4} + \frac{25^2}{5} - 241.23$$

$$= 26.25 + 64 + 125 - 241.23$$

$$= 245.25 - 241.23 = 4.02$$

$$SSE = TSS - SSR$$

$$= 22.77 - 4.02 = 18.75$$

Degrees of freedom

For Total SS = $N - 1 = 13 - 1 = 12$

$\therefore$ for SSR = $3 - 1 = 2$

For SSE = $12 - 2 = 10$

Sources of Variation	Sum of Squares	d.f.	Mean S.S.	Variance Ratio (F)
Rows (Salesmen)	4.02	2	$\frac{4.02}{2} = 2.01$	$F = \frac{2.01}{1.875} = 1.072$
Error	18.75	10	$\frac{18.75}{10} = 1.876$	–
Total	22.77	12	–	–

The critical value (tabulated) value of F for d.f. $V_1 = 2$ and $V_2 = 10$ and at level of significance $\alpha = 0.05$ is 3.81 (from table)

$\because$ Calculated valve is less than table valve, it is not significant. Hence the null hypothesis H_0 is rejected and we conclude that there is no significant difference in the sales of the three salesmen.

Chapter Ends...

❑❑❑

Module 5