INTRODUCTION TO DATA SCIENCE

.

Unit – I: Introduction

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues. Data Collection Strategies, Data Categorization: NOIR Topology.

Unit – II: Data Pre-Processing & Exploratory Data Analysis

Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics –ANOVA.

Unit – III: Model Development

Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and DecisionMaking.

Unit – IV: Model Evaluation

Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using GridSearch.

REFERENCES:

1. JojoMoolayil, “Smarter Decisions : The Intersection of IoT and Data Science”, PACKT, 2016.
2. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly,2015.
3. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”, EMC 2013
4. Raj, Pethuru, “Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGIGlobal.

	Introduction to Data Science
Sl. No	**UNIT-1 Topics**
1	Introduction to Data Science
2	A brief history of Data Science **Page 1**
3	Data science role and Skill tracks **Page 5 1.2**
4	Problem type **Page 15 1.3.2**
5	List of potential data science careers **Page 21 1.5**
6	Life Cycle of Data Science (Stages of Data Science Project) https://www.knowledgehut.com/blog/data-science/what-is-data-science-life-cycle or https://www.javatpoint.com/life-cycle-phases-of-data-analytics
7	Application of data Science in various Field https://www.geeksforgeeks.org/major-applications-of-data-science/ https://www.edureka.co/blog/data-science-applications/
8	Data Security Issues https://www.imperva.com/learn/data-security/data-security/
9	Data collection strategies https://www.simplilearn.com/what-is-data-collection-article
10	Data Categororization: NOIR https://byjus.com/maths/types-of-data-in-statistics/ https://cse.iitkgp.ac.in/~dsamanta/courses/da/resources/tutorials/PS02%20Data%20Categorization.pdf

1. Introduction to Data Science

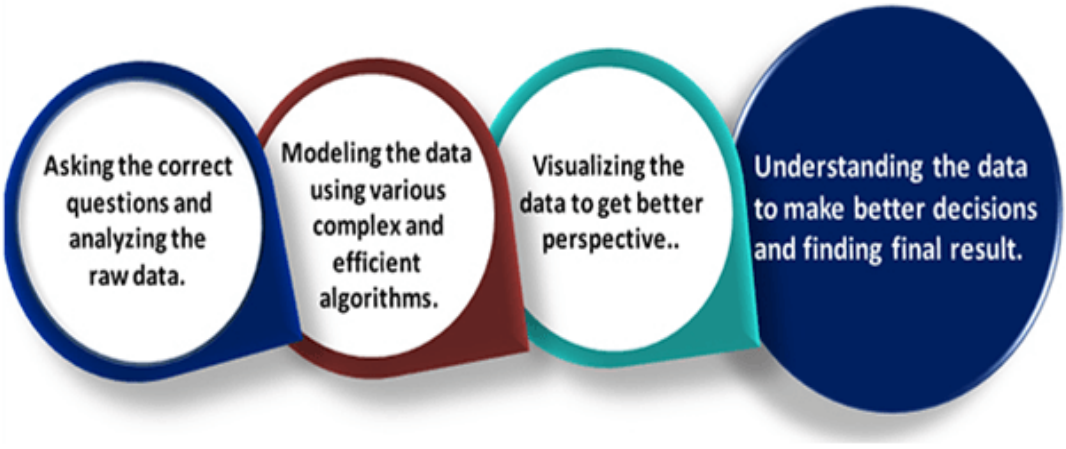
What is Data Science?

Data Science can be explained as the **entire process of gathering actionable insights from raw data** that involves various concepts that include statistical analysis, data analysis, machine learning algorithms, data modeling, preprocessing of data, etc.

Let's consider an example. A case study which also went to become a hollywood feature film "Moneyball". In the movie, they have shown how an underdog team went on to compete at the highest level of the baseball tournament by analyzing the statistical data points of each player and quantifying their performances to win the game. It can be aligned with how data science actually works.

How does Data Science Work?

- ✓ Asking the correct questions and analyzing the raw data.
- ✓ Modeling the data using various complex and efficient algorithms.
- ✓ Visualizing the data to get a better perspective.
- ✓ Understanding the data to make better decisions and finding the final result.



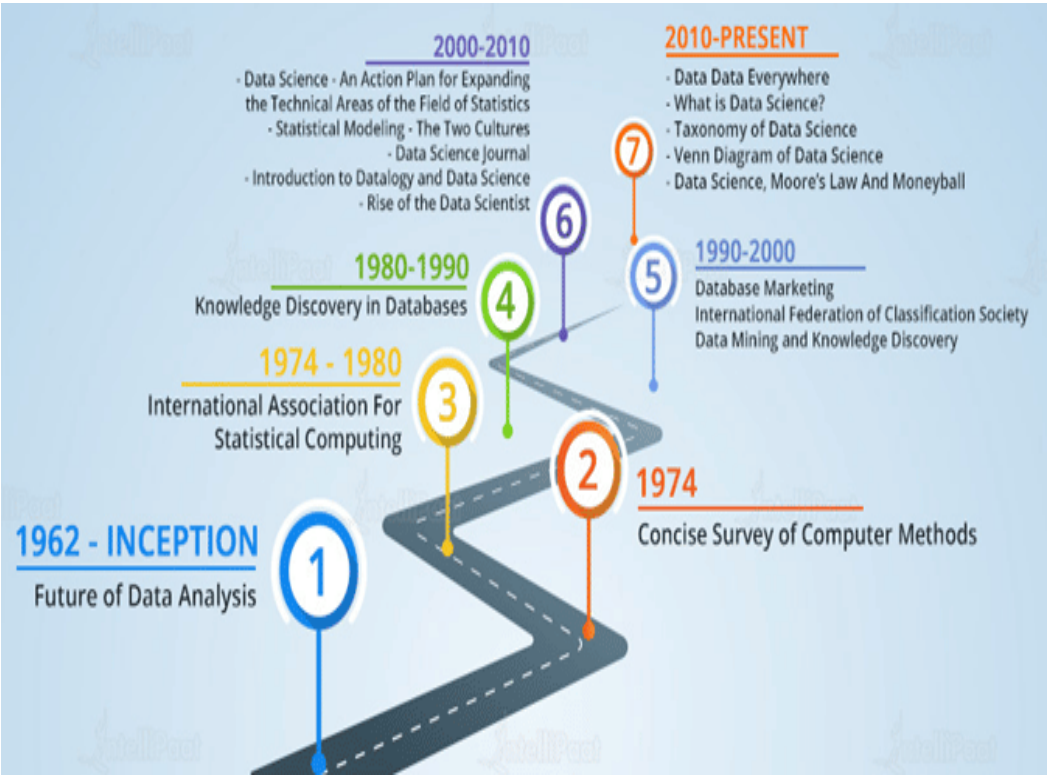
Example:
Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Need for Data Science:
In today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

- ✓ With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
- ✓ Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
- ✓ Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
- ✓ Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

2. A brief history of Data Science



1. 1962 – Inception

a. Future of Data Analysis – In 1962, John W Tukey wrote the "Future of Data Analysis" where he first mentioned the importance of data analysis with respect to science rather than mathematics.

2. 1974

a. Concise Survey of Computer Methods – In 1974, Peter Naur published the "Concise Survey of Computer methods that surveys the contemporary methods of data processing in various applications.

3. 1974 – 1980

a. International Association For Statistical Computing – In 1997, The committee was formed whose sole purpose is to link traditional statistical methodology with modern computer technology to extract useful information and knowledge from the data.

4. 1980-1990

a. Knowledge Discovery in Databases – In 1989, Gregory Piatetsky-Shapiro chaired the Knowledge Discovery in Databases that later went on to become the annual conference on knowledge discovery and data mining.

5. 1990-2000

a. Database Marketing – In 1994, BusinessWeek published a cover story that explains how big organizations are using the customer data to predict the likelihood of a customer buying a specific product or not. Kind of like how targeted ads work in the modern era for social media campaigns.

b. International Federation of Classification Society – For the first time in 1996, the term "Data Science" was used in a conference held in Japan.

6. 2000-2010

a. Data Science – An Action Plan for Expanding the Technical Areas of the Field of Statistics – In 2001, William S Cleveland published the action plan, that majorly focused on major areas of the technical work in the field of statistics and coined the term Data Science.

b. Statistical Modeling – The Two Cultures – In 2001, Leo Breiman wrote "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown".

c. Data Science Journal – April 2002 saw the launch of a journal that focused on management of data and databases in science and technology.

7. 2010-Present

a. Data Everywhere – In February 2010, Kenneth Cukier wrote a special report for The Economist that said a new professional has arrived – a data scientist. Who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data.

b. What is Data Science? – In June 2010, Mike Loukides described data science as combining entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution.

3. Data science role and Skill tracks

Data science has three main skill tracks: engineering, analysis, and modeling.

1. Engineering

Data engineering is the foundation that makes everything else possible. It mainly involves in building the data pipeline infrastructure. It involves the software and the hardware used to store the data and perform data ETL (i.e., extract, transform, and load) process. As cloud service development, it becomes the new norm to store and compute data on the cloud.

(a) Data environment

Designing and setting up the entire environment to support data science workflow is the prerequisite for data science projects. It may include setting up storage in the cloud, Kafka platform, Hadoop and Spark cluster, etc.

(b) Data management

Automated data collection is a common task that includes parsing the logs (depending on the stage of the company and the type of industry you are in), web scraping, API queries, and interrogating data streams. Determine and construct data schema to support analytical and modeling needs. Use tools, processes, guidelines to ensure data is correct, standardized, and documented.

(c) Production

It involves the whole pipeline from data access, preprocessing, modeling to final deployment. It is necessary to make the system work smoothly with all existing software stacks. So, it requires to monitor the system through some robust measures, such as rigorous error handling, fault tolerance, and graceful degradation to make sure the system is running smoothly and the users are happy.

2. Analysis

Analysis turns raw information into insights in a fast and often exploratory way.

(a) Domain knowledge

Domain knowledge is the understanding of the organization or industry where you apply data science.

Some questions about the context are:

- What are the business questions?
- How to translate a business need to a data problem?

Domain knowledge helps you to deliver the results in an audience-friendly way with the right solution to the right problem.

(b) Exploratory analysis

This type of analysis is about exploration and discovery. It often involves different ways to slice and aggregate data.

(c) Storytelling

Storytelling with data is critical to deliver insights and drive better decision making. It usually requires data summarization, aggregation, and visualization. A business-friendly report or an interactive dashboard is the typical outcome of the analysis.

3. Modeling

Modeling is a process that dives deeper into the data to discover the pattern.

(a) Supervised learning

Supervised learning happens in the presence of a supervisor just like learning performed by a small child with the help of his teacher. As a child is trained to recognize fruits, colors, numbers under the supervision of a teacher this method is supervised learning. In this method, every step of the child is checked by the teacher and the child learns from the output that he has to produce.

(b) Unsupervised learning

Unsupervised learning happens without the help of a supervisor just like a fish learns to swim by itself. It is an independent learning process. In this model, as there is no output mapped with the input, the target values are unknown/ unlabeled. The system needs to learn by itself from the data input to it and detect the hidden patterns.

(c) Customized model development

A data scientist may need to develop new models to accommodate the features of the problem at hand.

Here is a list of questions that can help you decide the type of technique to use:

- Is your data labeled?
- Is your data easy to collect?

4. Some common skills

Data Preprocessing:

Data preprocessing is the process of converting raw data into clean data that is proper to use.

(a) Data preprocessing for data engineer

A data lake is a storage repository that stores a vast amount of raw data in its native format, including XML, JSON, CSV, Parquet, etc. It is a data cesspool rather than a data lake. The data engineer's job is to get a clean schema out of the data lake by transforming and formatting the data.

(b) Data preprocessing for data analyst and scientist

A data analyst collects and stores data on sales numbers, market research, logistics, linguistics, or other behaviors. They bring technical expertise to ensure the quality and accuracy of that data, then process, design, and present it in ways to help people, businesses, and organizations make better decisions.
Data Scientist:
A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

4. Problem type

1. Description

The primary analytic problem is to summarize and explore a data set with descriptive statistics (mean, standard deviation, and so forth) and visualization methods. Questions of this kind are:

- What is the annual income distribution?
- What are the mean active days of different accounts?

2. Comparison

The first common modeling problem is to compare different groups.
Here are some examples:

- Are males more inclined to buy our products than females?
- Are there any differences in customer satisfaction in different business districts?

The commonly used statistical tests are chi-square test, t-test, and ANOVA.

3. Clustering

Clustering is a widespread problem, and it can answer questions like:

- How many reasonable customer segments are there based on historical purchase patterns?

Clustering is an unsupervised learning mechanism. Unsupervised learning happens without the help of a supervisor just like a fish learns to swim by itself. It is an independent learning process. In this model, as there is no output mapped with the input, the target values are unknown/ unlabeled. The system needs to learn by itself from the data input to it and detect the hidden patterns.

What Is Unlabeled Dataset?

A dataset with unknown output values for all the input values is called an unlabeled dataset. **For Example,** while buying products online, if butter is put in the cart, then it suggests buying bread, cheese, etc. The unsupervised model looks at the data points and predicts the other attributes that are associated with the product.
The unsupervised learning algorithms include Clustering and Association Algorithms such as:
Apriori, K-means clustering and other association rule mining algorithms.
Clustering Algorithm: The methods of finding the similarities between data items such as the same shape, size, color, price, etc. and grouping them to form a cluster is cluster analysis.
Association Rule Mining: In this type of mining, it finds out the most frequently occurring itemsets or associations between elements. Associations such as "products often purchased together", etc.

4. Classification

Here are some example questions:

- Will this customer likely to buy our product?
- Is it spam email or not?

Classification is a supervised learning mechanism. Supervised learning happens in the presence of a supervisor just like learning performed by a small child with the help of his teacher. As a child is trained to recognize fruits, colors, numbers under the supervision of a teacher this method is supervised learning.
In this method, every step of the child is checked by the teacher and the child learns from the output that he has to produce.

What Is a Labeled Dataset?

The dataset with outputs known for a given input is called a Labeled Dataset. **For example,** an image of fruit along with the fruit name is known. So when a new image of fruit is shown, it compares with the training set to predict the answer.

Supervised learning is a fast learning mechanism with high accuracy. The supervised learning problems include regression and classification problems.

Some of the supervised learning algorithms are:

Decision Trees, K-Nearest Neighbor, Linear Regression, and Neural Networks.

Classification: In these types of problems, we predict the response as specific classes, such as "yes" or "no". When only 2 classes are present, then it is called a Binary Classification. For more than 2 class values, it is called a Multi-class Classification. The predicted response values are discrete values.

For example, Is it the image of the sun or the moon? The classification algorithm separates the data into classes.

Difference Between Supervised Vs Unsupervised Learning

Supervised	Unsupervised
In supervised learning algorithms, the output for the given input is known.	In unsupervised learning algorithms, the output for the given input is unknown.
The algorithms learn from labeled set of data. This data helps in evaluating the accuracy on training data.	The algorithm is provided with unlabeled data where it tries to find patterns and associations in between the data items.
It is a Predictive Modeling technique which predicts the future outcomes accurately.	It is a Descriptive Modeling technique which explains the real relationship between the elements and history of the elements.
It includes classification and regression algorithms.	It includes clustering and association rules learning algorithms.
Some algorithms of supervised learning are Linear Regression, Naïve Bayes, and Neural Networks.	Some algorithms for unsupervised learning are k- means clustering, Apriori, etc.
It is more accurate than unsupervised learning as input data and corresponding output is well known, and the machine only needs to give predictions.	It has less accuracy as the input data is unlabeled. Thus the machine has to first understand and label the data and then give predictions.

5. Regression

Here are some example questions:

- What will be the temperature tomorrow?
- How much inventory should we have?

Regression: Regression problems predict the response as continuous values such as predicting a value that ranges from -infinity to infinity. It may take many values. **For example,** the linear regression algorithm that is applied, predicts the cost of the house based on many parameters such as location, nearby airport, size of the house, etc.

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.

Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It **predicts continuous/real values such as** temperature, age, salary, price, **etc.**

6. Optimization

Optimization is another common type of problems in data science to find an optimal solution by tuning a few tune-able variables with other non-controllable environmental variables. It is an expansion of comparison problem and can solve problems such as:

- What is the best route to deliver the packages?

5. List of potential data science careers

Data infrastructure engineer

Designing, building, and running the data infrastructure to support the Video organizations growing data needs.

Go, Python, AWS/Google Cloud/Azure, logstash, Kafka, and Hadoop

Data Engineer:

A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition, and verification.

Skill required: Data engineer must have depth knowledge of **SQL, MongoDB, Cassandra, HBase, Apache Spark, Hive, MapReduce**, with language knowledge of **Python, C/C++, Java, Perl**, etc.

spark/scala, python, SQL, AWS/Google Cloud/Azure, Data modeling

BI engineer

Design, implement, and maintain systems used to collect and analyze business intelligence data. They create dashboards, databases, and other platforms that allow for efficient collection and evaluation of BI data.

Skill required: Tableau/looker/Mode, etc., data visualization, SQL, Python

Data Analyst:

Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends, and so on. At the end of the day, he comes up with visualization and reporting for analyzing the data for decision making and problem-solving process.

Skill required: For becoming a data analyst, you must get a good background in **mathematics, business intelligence, data mining**, and basic knowledge of **statistics**. You should also be familiar with some computer languages and tools such as **MATLAB, Python, SQL, Hive, Pig, Excel, SAS, R, JS, Spark**, etc.

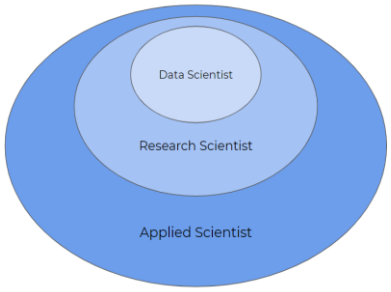
Data Scientist:

A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

Skill required: To become a data scientist, one should have technical language skills such as **R, SAS, SQL, Python, Hive, Pig, Apache spark, MATLAB**. Data scientists must have an understanding of Statistics, Mathematics, visualization, and communication skills.

Research scientist

- ✓ Building research proposals
- ✓ Creating and conducting experiments
- ✓ Analysing results of the experiments
- ✓ Working with other researchers to use and develop end product
- ✓ Applying for grants to continue research



Skill required: R/Python, advanced statistics, experimental design, ML, research background, publications, conference contributions, algorithms

Applied scientist

An applied scientist is more interested in real-life applications.

An applied scientist does scientific research with a focus on applying the results of their studies to solving real-world problems. They use the scientific method to develop research questions and then conduct studies that lead to practical solutions.
Applied Scientists at Amazon, for example, focus on projects to enhance Amazon's customer experience like Amazon's Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Audio Signal Processing, text-to-speech (TTS)
Skill required: ML algorithm design, often with an expectation of fundamental software engineering skills

Machine Learning Engineer

Their tasks involve researching, building, and designing the artificial intelligence responsible for machine learning and maintaining and improving existing artificial intelligence systems.
Machine learning engineers design and create the AI algorithms capable of learning and making predictions that define machine learning (ML). An ML engineer typically works as part of a larger data science team and will communicate with data scientists, administrators, data analysts, data engineers and data architects. They should also have an understanding of various algorithms, problem-solving analytical skill, probability, and statistics.
Skill required: More advanced software engineering skillset, algorithms, machine learning algorithm design, system design

6. Life Cycle of Data Science (Stages of Data Science Project)

The Lifecycle of Data Science

The major steps in the life cycle of Data Science project are as follows:

1. Problem identification

Domain experts and Data Scientists are the key persons in the problem identification of problem. Domain expert has in depth knowledge of the application domain and exactly what is the problem to be solved. Data Scientist understands the domain and help in identification of problem and possible solutions to the problems.

2. Business Understanding

Understanding what customer exactly wants from the business perspective is nothing but Business Understanding. Whether customer wish to do predictions or want to improve sales or minimise the loss or optimise any particular process etc forms the business goals. During business understanding two important steps are followed:

- **KPI (Key Performance Indicator)**

For any data science project, key performance indicators define the performance or success of the project. There is a need to be an agreement between the customer and data science project team on Business related indicators and related data science project goals. Depending on the business need the business indicators are devised and then accordingly the data science project team decides the goals and indicators. To better understand this let us see an example. Suppose the business need is to optimise the overall spending of the company, then the data science goal will be to use the existing resources to manage double the clients. Defining the Key performance Indicators is very crucial for any data science projects as the cost of the solutions will be different for different goals.

- **SLA (Service Level Agreement)**

Once the performance indicators are set then finalizing the service level agreement is important. As per the business goals the service level agreement terms are decided. For example, for any airline reservation system simultaneous processing of say 1000 users is required. Then the product must satisfy this service requirement is the part of service level agreement.
Once the performance indicators are agreed and service level agreement is completed then the project proceeds to the next important step.

3. Collecting Data

Data Collection is the important step as it forms the important base to achieve targeted business goals.

The basic data collection can be done using the surveys. Generally, the data collected through surveys provide important insights. Much of the data is collected from the various processes followed in the enterprise. At various steps the data is recorded in various software systems used in the enterprise which is important to understand the process followed from the product development to deployment and delivery. The historical data available through archives is also important to better understand the business. Transactional data also plays a vital role as it is collected on a daily basis. Many statistical methods are applied to the data to extract the important information related to business. In data science project the major role is played by data and so proper data collection methods are important.

4. Pre-processing data

Large data is collected from archives, daily transactions and intermediate records. The data is available in various formats and in various forms. Some data may be available in hard copy formats also. The data is scattered at various places on various servers. All these data are extracted and converted into single format and then processed. As data warehouse is constructed where the Extract, Transform and Loading (ETL) process or operations are carried out. A data architect role is important in this stage who decides the structure of data warehouse and perform the steps of ETL operations.

5. Analyzing data

Now that the data is available and ready in the format required then next important step is to understand the data in depth. This understanding comes from analysis of data using various statistical tools available. A data engineer plays a vital role in analysis of data. This step is also called as Exploratory Data Analysis (EDA). Here the data is examined by formulating the various statistical functions and dependent and independent variables or features are identified. Careful analysis of data revels which data or features are important and what is the spread of data. Various plots are utilized to visualize the data for better understanding. The tools like Tableau, PowerBI etc are famous for performing Exploratory Data Analysis and Visualization. Knowledge of Data Science with Python and R is important for performing EDA on any type of data.

6. Data Modelling

Data modelling is the important next step once the data is analysed and visualized. The important components are retained in the dataset and thus data is further refined. Now the important is to decide how to model the data? What tasks are suitable for modelling? The tasks, like classification or regression, which is suitable is dependent upon what business value is required. In these tasks also many ways of modelling are available. The Machine Learning engineer applies various algorithms to the data and generates the output. While modelling the data many a times the models are first tested on dummy data similar to actual data.

7. Model Evaluation/ Monitoring

As there are various ways to model the data so it is important to decide which one is effective. For that model evaluation and monitoring phase is very crucial and important. The model is now tested with actual data. The data may be very few and in that case the output is monitored for improvement. There may be changes in data while model is being evaluated or tested and the output will drastically change depending on changes in data. So, while evaluating the model following two phases are important:

8. Model Training

Once the task and the model are finalised and data drift analysis modelling is finalized then the important step is to train the model. The training can be done is phases where the important parameters can be further fine tuned to get the required accurate output. The model is exposed to the actual data in production phase and output is monitored.

9. Model Deployment

Once the model is trained with the actual data and parameters are fine tuned then model is deployed. Now the model is exposed to real time data flowing into the system and output is generated. The model can be deployed as web service or as an embedded application in edge or mobile application. This is very important step as now model is exposed to real world.

10. Driving insights and generating BI reports

After model deployment in real world, next step is to find out how model is behaving in real world scenario. The model is used to get the insights which aid in strategic decisions related to business. The business goals are bound to these insights. Various reports are generated to see how business is driving. These reports help in finding out if key process indicators are achieved or not.

11. Taking a decision based on insight

For data science to make wonders, every step indicated above has to be done very carefully and accurately. When the steps are followed properly then the reports generated in above step helps in taking key decisions for the organization. The insights generated helps in taking strategic decisions like for example the organization can predict that there will be need of raw material in advance. The data science can be of great help in taking many important decisions related to business growth and better revenue generation.

7. Application of data Science in various Field

1. In 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. In Transport

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 and how to handle different situations while driving etc.

3. In Finance

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. In 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. In Health Care

Data Science is used for: Detecting Tumor, Drug discoveries, Medical Image Analysis, Virtual Medical Bots, Genetics and Genomics, Predictive Modeling 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. 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.

8. 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.

9. 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. In Delivery Logistics

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

12. 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.

13. Augmented Reality:

Data Science and Virtual Reality do have a relationship, considering a VR headset contains computing knowledge, algorithms and data to provide you with the best viewing experience. A very small step towards this is the high-trending game of Pokemon GO.

8. Data Security Issues

What is Data Security?

Data security is the process of protecting corporate data and preventing data loss through unauthorized access. This includes protecting your data from attacks that can encrypt or destroy data, such as ransomware, as well as attacks that can modify or corrupt your data. Data security also ensures data is available to anyone in the organization who has access to it.

Why is Data Security important?

Data is a valuable asset that generates, acquires, saves, and exchanges for any company. Protecting it from internal or external corruption and illegal access protects a company from financial loss, reputational harm, consumer trust degradation, and brand erosion.

Main elements of Data Security

- ✓ **Confidentiality:** Ensures that only authorized users, with appropriate credentials, have access to data.
- ✓ **Integrity:** Ensures that all data is accurate, trustworthy, and not prone to unjustified changes.
- ✓ **Availability:** Ensures that data is accessible and available for ongoing business needs in a timely and secure manner.

Data Privacy:

There are two main aspects to enforcing data privacy:

Access control—ensuring that anyone who tries to access the data is authenticated to confirm their identity, and authorized to access only the data they are allowed to access.

Data protection— ensuring that even if unauthorized parties manage to access the data, they cannot view it or cause damage to it. Data protection methods ensure encryption, which prevents anyone from viewing data if they do not have a private encryption key, and data loss prevention mechanisms which prevent users from transferring sensitive data outside the organization.

The primary difference is that data privacy mainly focuses on keeping data confidential, while data security mainly focuses on protecting from malicious activity.

Differentiate between Data Privacy and Data Security

Data Privacy	Data Security
Data Privacy is all about the reflection of what data is important and why.	Data Security is all about the reflection of how those policies got enforced.
Data privacy sets about proper usage, collection, retention, deletion, and storage of data.	Data security sets the policies, methods, and means to secure personal data.
It offers to block websites, internet browsers, cable companies, and internet service providers from tracking your information and your browser history.	It offers to protect you from other people accessing your personal information and other data.
Data privacy tools include browser extensions and add-on, password managers, private browsers and email services, encrypted messaging, private search engines, web proxies, file encryption software, and ad and tracker blockers.	Data Security tools involve with identity and access management, data loss prevention, anti-malware, anti-virus, event management and data masking software.
For e.g. The European Union's General Data Protection Regulation is a type of international standard for protecting the privacy of EU citizens.	For e.g. The Payment Card Industry Data Security Standard is a set of rules which protect the sensitive payment card information and cardholder data.

Data Security Risks

1. Accidental Exposure

A large percentage of data breaches are not the result of a malicious attack but are caused by negligent or accidental exposure of sensitive data. It is common for an organization's employees to share, grant access to, lose, or mishandle valuable data, either by accident or because they are not aware of security policies.

This major problem can be addressed by employee training, but also by other measures, such as data loss prevention (DLP) technology and improved access controls.
A data breach or data leak is the release of sensitive, confidential or protected data to an untrusted environment.

2. Phishing and Other Social Engineering Attacks

Social engineering attacks are a primary vector used by attackers to access sensitive data. They involve manipulating or tricking individuals into providing private information or access to privileged accounts.
Phishing is a common form of social engineering. It involves messages that appear to be from a trusted source, but in fact are sent by an attacker. When victims comply, for example by providing private information or clicking a malicious link, attackers can compromise their device or gain access to a corporate network.

3. Insider Threats

Insider threats are employees who inadvertently or intentionally threaten the security of an organization's data. There are three types of insider threats:
Non-malicious insider—these are users that can cause harm accidentally, via negligence, or because they are unaware of security procedures.
Malicious insider—these are users who actively attempt to steal data or cause harm to the organization for personal gain.
Compromised insider—these are users who are not aware that their accounts or credentials were compromised by an external attacker. The attacker can then perform malicious activity, pretending to be a legitimate user.

4. Ransomware

Ransomware is a major threat to data in companies of all sizes. Ransomware is malware that infects corporate devices and encrypts data, making it useless without the decryption key. Attackers display a ransom message asking for payment to release the key, but in many cases, even paying the ransom is ineffective and the data is lost.
If an organization does not maintain regular backups, or if the ransomware manages to infect the backup servers, there may be no way to recover.

5. Data Loss in the Cloud

Many organizations are moving data to the cloud to facilitate easier sharing and collaboration. However, when data moves to the cloud, it is more difficult to control and prevent data loss. Users access data from personal devices and over unsecured networks. It is all too easy to share a file with unauthorized parties, either accidentally or maliciously.

6. SQL Injection

SQL injection (SQLi) is a common technique used by attackers to gain illicit access to databases, steal data, and perform unwanted operations. It works by adding malicious code to a seemingly innocent database query.
SQL injection manipulates SQL code by adding special characters to a user input that change the context of the query. The database expects to process a user input, but instead starts processing malicious code that advances the attacker's goals. SQL injection can expose customer data, intellectual property, or give attackers administrative access to a database, which can have severe consequences.
SQL injection vulnerabilities are typically the result of insecure coding practices. It is relatively easy to prevent SQL injection if coders use secure mechanisms for accepting user inputs, which are available in all modern database systems.

Common Data Security Solutions and Techniques

There are several technologies and practices that can improve data security.

1. Data Discovery and Classification

Modern IT environments store data on servers, endpoints, and cloud systems. Visibility over data flows is an important first step in understanding what data is at risk of being stolen or misused. To properly

protect your data, you need to know the type of data, where it is, and what it is used for. Data discovery and classification tools can help.

Data detection is the basis for knowing what data you have. Data classification allows you to create scalable security solutions, by identifying which data is sensitive and needs to be secured. Data detection and classification solutions enable tagging files on endpoints, file servers, and cloud storage systems, letting you visualize data across the enterprise, to apply the appropriate security policies.

2. Data Masking

Data masking lets you create a synthetic version of your organizational data, which you can use for software testing, training, and other purposes that don't require the real data. The goal is to protect data while providing a functional alternative when needed.

Data masking retains the data type, but changes the values. Data can be modified in a number of ways, including encryption, character shuffling and character or word substitution. Whichever method you choose, you must change the values in a way that cannot be reverse-engineered.

3. Identity Access Management

Identity and Access Management (IAM) is a business process, strategy, and technical framework that enables organizations to manage digital identities. IAM solutions allow IT administrators to control user access to sensitive information within an organization.

Systems used for IAM include single sign-on systems, two-factor authentication, multi-factor authentication, and privileged access management. These technologies enable the organization to securely store identity and profile data, and support governance, ensuring that the appropriate access policies are applied to each part of the infrastructure.

4. Data Encryption

Data encryption is a method of converting data from a readable format (plaintext) to an unreadable encoded format (ciphertext). Only after decrypting the encrypted data using the decryption key, the data can be read or processed.

In public-key cryptography techniques, there is no need to share the decryption key – the sender and recipient each have their own key, which are combined to perform the encryption operation. This is inherently more secure.

Data encryption can prevent hackers from accessing sensitive information. It is essential for most security strategies and is explicitly required by many compliance standards.

5. Data Loss Prevention (DLP)

To prevent data loss, organizations can use a number of safeguards, including backing up data to another location. Physical redundancy can help protect data from natural disasters, outages, or attacks on local servers. Redundancy can be performed within a local data center, or by replicating data to a remote site or cloud environment.

Beyond basic measures like backup, DLP software solutions can help protect organizational data. DLP software automatically analyzes content to identify sensitive data, enabling central control and enforcement of data protection policies, and alerting in real-time when it detects anomalous use of sensitive data, for example, large quantities of data copied outside the corporate network.

6. Governance, Risk, and Compliance (GRC)

GRC is a methodology that can help improve data security and compliance:

Governance creates controls and policies enforced throughout an organization to ensure compliance and data protection.

Risk involves assessing potential cybersecurity threats and ensuring the organization is prepared for them.

Compliance ensures organizational practices are in line with regulatory and industry standards when processing, accessing, and using data.

7. Password Hygiene

One of the simplest best practices for data security is ensuring users have unique, strong passwords. Without central management and enforcement, many users will use easily guessable passwords or use the

same password for many different services. Password spraying and other brute force attacks can easily compromise accounts with weak passwords.

A simple measure is enforcing longer passwords and asking users to change passwords frequently. However, these measures are not enough, and organizations should consider multi-factor authentication (MFA) solutions that require users to identify themselves with a token or device they own, or via biometric means.

Another complementary solution is an enterprise password manager that stores employee passwords in encrypted form, reducing the burden of remembering passwords for multiple corporate systems, and making it easier to use stronger passwords. However, the password manager itself becomes a security vulnerability for the organization.

8. Authentication and Authorization

Organizations must put in place strong authentication methods, such as OAuth 2.0 for web-based systems. It is highly recommended to enforce multi-factor authentication when any user, whether internal or external, requests sensitive or personal data.

In addition, organizations must have a clear authorization framework in place, which ensures that each user has exactly the access rights they need to perform a function or consume a service, and no more. Periodic reviews and automated tools should be used to clean up permissions and remove authorization for users who no longer need them.

Authentication verifies the identity of a user or service, and authorization determines their access rights.

Comparing these processes to a real-world example, when you go through security in an airport, you show your ID to authenticate your identity. Then, when you arrive at the gate, you present your boarding pass to the flight attendant, so they can authorize you to board your flight and allow access to the plane.

9. Data Security Audits

The organization should perform security audits at least every few months. This identifies gaps and vulnerabilities across the organizations' security posture. It is a good idea to perform the audit via a third-party expert, for example in a penetration testing model. However, it is also possible to perform a security audit in house. Most importantly, when the audit exposes security issues, the organization must devote time and resources to address and remediate them.

10. Anti-Malware, Antivirus, and Endpoint Protection

Malware is the most common vector of modern cyberattacks, so organizations must ensure that endpoints like employee workstations, mobile devices, servers, and cloud systems, have appropriate protection. Endpoint protection platforms (EPP) take a more comprehensive approach to endpoint security. They combine antivirus with a machine-learning-based analysis of anomalous behavior on the device, which can help detect unknown attacks. Most platforms also provide endpoint detection and response (EDR) capabilities, which help security teams identify breaches on endpoints as they happen, investigate them, and respond by locking down and reimaging affected endpoints.

11. Zero Trust

Zero trust is a security model introduced by Forrester analyst John Kindervag, which has been adopted by the US government, several technical standards bodies, and many of the world's largest technology companies. The basic principle of zero trust is that no entity on a network should be trusted, regardless of whether it is outside or inside the network perimeter.

Zero trust has a special focus on data security, because data is the primary asset attackers are interested in. A zero trust architecture aims to protect data against insider and outside threats by continuously verifying all access attempts, and denying access by default.

Database Security

Database security involves protecting database management systems such as Oracle, SQL Server, or MySQL, from unauthorized use and malicious cyberattacks. The main elements protected by database security are:

- ✓ The database management system (DBMS).

- ✓ Data stored in the database.
- ✓ Applications associated with the DBMS.
- ✓ The physical or virtual database server and any underlying hardware.
- ✓ Any computing and network infrastructure used to access the database.

A database security strategy involves tools, processes, and methodologies to securely configure and maintain security inside a database environment and protect databases from intrusion, misuse, and damage.

Big Data Security

Big data security involves practices and tools used to protect large datasets and data analysis processes. Big data commonly takes the form of financial logs, healthcare data, data lakes, archives, and business intelligence datasets.

Big data security aims to prevent accidental and intentional breaches, leaks, losses, and exfiltration of large amounts of data. Let's review popular big data services and see the main strategies for securing them.

AWS Big Data

AWS offers analytics solutions for big data implementations. There are various services AWS offers to automate data analysis, manipulate datasets, and derive insights, including Amazon Simple Storage Service (S3), Amazon Kinesis, Amazon Elastic Map/Reduce (EMR), and Amazon Glue.

AWS big data security best practices include:

- ✓ **Access policy options**—use access policy options to manage access to your S3 resources.
- ✓ **Data encryption policy**—use Amazon S3 and AWS KMS for encryption management.
- ✓ **Manage data with object tagging**—categorize and manage S3 data assets using tags, and apply tags indicating sensitive data that requires special security measures.

Azure Big Data

Microsoft Azure cloud offers big data and analytics services that can process a high volume of structured and unstructured data. The platform offers elastic storage using Azure storage services, real-time analytics, database services, as well as machine learning and data engineering solutions.

Azure big data security best practices include:

- ✓ Monitor as many processes as possible.
- ✓ Leverage Azure Monitor and Log Analytics to gain visibility over data flows.
- ✓ Define and enforce a security and privacy policy.
- ✓ Leverage Azure services for backup, restore, and disaster recovery.

Google Cloud Big Data

The Google Cloud Platform offers multiple services that support big data storage and analysis. BigQuery is a high-performance SQL-compatible engine, which can perform analysis on large data volumes in seconds. Additional services include Dataflow, Dataproc, and Data Fusion.

Google Cloud big data security best practices include:

- ✓ Define BigQuery access controls according to the least privilege principle.
- ✓ Use policy tags or type-based classification to identify sensitive data.

Snowflake

Snowflake is a cloud data warehouse for enterprises, built for high performance big data analytics. The architecture of Snowflake physically separates compute and storage, while integrating them logically. Snowflake offers full relational database support and can work with structured and semi-structured data.

Snowflake security best practices include:

- ✓ Leverage key pair authentication and rotation to improve client authentication security.
- ✓ Enable multi-factor authentication.

Elasticsearch

Elasticsearch is an open-source full-text search and analytics engine that is highly scalable, allowing search and analytics on big data in real-time.

- ✓ Use strong passwords to protect access to search clusters

- ✓ Encrypt all communications using SSL/TLS SSL (Secure Socket Layer) and TLS (Transport Layer Security)
- ✓ Use IP (Internet Protocol) filtering for client access
- ✓ Turn on auditing and monitor logs on an regular basis

Securing Data in Enterprise Applications

Enterprise applications power mission critical operations in organizations of all sizes. Enterprise application security aims to protect enterprise applications from external attacks, abuse of authority, and data theft.

Email Security

Email security is the process of ensuring the availability, integrity, and reliability of email communications by protecting them from cyber threats.

Technical standards bodies have recommended email security protocols including SSL/TLS, Sender Policy Framework (SPF), and DomainKeys Identified Mail (DKIM). These protocols are implemented by email clients and servers, including Microsoft Exchange and Google G Suite, to ensure secure delivery of emails. A secure email gateway helps organizations and individuals protect their email from a variety of threats, in addition to implementing security protocols.

ERP Security

Enterprise Resource Planning (ERP) is software designed to manage and integrate the functions of core business processes such as finance, human resources, supply chain, and inventory management into one system. ERP systems store highly sensitive information and are, by definition, a mission critical system.

ERP security is a broad set of measures designed to protect an ERP system from unauthorized access and ensure the accessibility and integrity of system data. The Information Systems Audit and Control Association (ISACA) recommends regularly performing security assessments of ERP systems, including software vulnerabilities, misconfigurations, separation of duties (SoD) conflicts, and compliance with vendor security recommendations.

DAM Security

Digital Asset Management (DAM) is a technology platform and business process for organizing, storing, and acquiring rich media and managing digital rights and licenses. Rich media assets include photos, music, videos, animations, podcasts, and other multimedia content. Data stored in DAM systems is sensitive because it often represents company IP, and is used in critical processes like sales, marketing, and delivery of media to viewers and web visitors.

Security best practices for DAM include:

- ✓ Implement the principle of least privilege.
- ✓ Use multi-factor authentication to control access by third parties.
- ✓ Regularly review automation scripts, limit privileges of commands used, and control the automation process through logging and alerting.

CRM Security

Customer Relationship Management (CRM) is a combination of practices, strategies, and technologies that businesses use to manage and analyze customer interactions and data throughout the customer lifecycle. CRM data is highly sensitive because it can expose an organization's most valuable asset—customer relationships.

Security best practices for CRM include:

- ✓ Perform period IT risk assessment audits for CRM systems.
- ✓ Perform CRM activity monitoring to identify unusual or suspicious usage.
- ✓ Encourage CRM administrators to follow security best practices.
- ✓ Educate CRM users on security best practices.

9. Data collection strategies

Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Why is Data Collection important?

- **The trustworthiness of The Research –** A critical purpose behind data collection via quantitative or qualitative techniques is to guarantee that the research question's honesty is kept up without a doubt.
- **Diminish the probability of blunders or errors –** The right utilization of suitable data collection strategies decreases the probability of blunders during different research processes.
- **Effective and accurate decision making –** To limit the danger of blunders or errors in decision making, it is significant that precise data is gathered, so the specialists do not settle on clueless choices.
- **Save Cost and Time –** Data collection plays a significant role in saving time and money that can otherwise be squandered without more profound comprehension of the point or topic.
- **Empowers a new idea or change –** To demonstrate the requirement for an adjustment or new change, it is critical to collect data and information as proof to help these cases.

Depending on the type of data, the data collection method is divided into two categories namely,

- Primary Data Collection methods
- Secondary Data Collection methods

Primary Data Collection Methods

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

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

Observation Method

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

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

Interview Method

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

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple, Should follow a logical sequence
- Provide adequate space for answers, Avoid technical terms

Projective Technique

Projective data gathering is an indirect interview, used when potential respondents know why they're being asked questions and hesitate to answer. For instance, someone may be reluctant to answer questions about their phone service if a cell phone carrier representative poses the questions. With projective data gathering, the interviewees get an incomplete question, and they must fill in the rest, using their opinions, feelings, and attitudes.

Delphi Technique

The Oracle at Delphi, according to Greek mythology, was the high priestess of Apollo's temple, who gave advice, prophecies, and counsel. In the realm of data collection, researchers use the Delphi technique by gathering information from a panel of experts. Each expert answers questions in their field of specialty, and the replies are consolidated into a single opinion.

Focus Groups

Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.

Schedules

This method is similar to the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

Secondary Data Collection Methods

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

Published data are available in various resources including

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

Unpublished data includes

- Diaries, Letters, Unpublished biographies, etc.

Data Collection Tools

Now that we've explained the various techniques, let's narrow our focus even further by looking at some specific tools. For example, we mentioned interviews as a technique, but we can further break that down into different interview types (or "tools").

1. Word Association

The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.

2. Sentence Completion

Researchers use sentence completion to understand what kind of ideas the respondent has. This tool involves giving an incomplete sentence and seeing how the interviewee finishes it.

3. Role-Playing

Respondents are presented with an imaginary situation and asked how they would act or react if it was real.

4. In-Person Surveys

The researcher asks questions in person.

5. Online/Web Surveys

These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

6. Mobile Surveys

These surveys take advantage of the increasing proliferation of mobile technology. Mobile collection surveys rely on mobile devices like tablets or smartphones to conduct surveys via SMS or mobile apps.

7. Phone Surveys

No researcher can call thousands of people at once, so they need a third party to handle the chore. However, many people have call screening and won't answer.

8. Observation

Sometimes, the simplest method is the best. Researchers who make direct observations collect data quickly and easily, with little intrusion or third-party bias. Naturally, it's only effective in small-scale situations.

9. Photography and video: Photographs and videos show still or moving images. Photographs can be used on their own, but are more often accompanied by written captions, providing additional information. Videos are often accompanied by a commentary.

10. Focus group discussions: Focus group discussions (FGDs) are facilitated discussions, held with a small group of people who have specialist knowledge or interest in a particular topic. They are used to find out the perceptions and attitudes of a defined group of people. FGDs are typically carried out with around 6-12 people, and are based around a short list of guiding questions, designed to probe for in-depth information.

11. Case studies and stories of change: A case study is not a data collection tool in itself. It is a descriptive piece of work that can provide in-depth information on a topic. It is often based on information acquired through one or more of the other tools described in this paper, such as interviews or observation. Case studies are usually written, but can also be presented as photographs, films or videos. Case studies often focus on people (individuals, households, communities). But they can also focus on any other unit of analysis such as locations, organisations, policies or the environment. Stories of change are similar to case studies.

12. Surveys and questionnaires: These are designed to collect and record information from many people, groups or organisations in a consistent way. A questionnaire is a form containing questions. It may be a printed form or one designed to be filled in online. Questionnaires may be administered in many different ways. A survey, by contrast, is normally a large, formal exercise. It typically consists of three different aspects: an approved sampling method designed to ensure the survey is representative of a wider population; a standard questionnaire that ensures information is collected and recorded consistently; and a set of analysis methods that allow results and findings to be generated.

What are Common Challenges in Data Collection?

There are some prevalent challenges faced while collecting data, let us explore a few of them to understand them better and avoid them.

1. Data Quality Issues

The main threat to the broad and successful application of machine learning is poor data quality. Data quality must be your top priority if you want to make technologies like machine learning work for you. Let's talk about some of the most prevalent data quality problems in this blog article and how to fix them.

2. Inconsistent Data

When working with various data sources, it's conceivable that the same information will have discrepancies between sources. The differences could be in formats, units, or occasionally spellings. The introduction of inconsistent data might also occur during firm mergers or relocations. Inconsistencies in data have a tendency to accumulate and reduce the value of data if they are not continually resolved. Organizations that have heavily focused on data consistency do so because they only want reliable data to support their analytics.

3. Data Downtime

Data is the driving force behind the decisions and operations of data-driven businesses. However, there may be brief periods when their data is unreliable or not prepared. Customer complaints and subpar analytical outcomes are only two ways that this data unavailability can have a significant impact on businesses. A data engineer spends about 80% of their time updating, maintaining, and guaranteeing the integrity of the data pipeline. In order to ask the next business question, there is a high marginal cost due to the lengthy operational lead time from data capture to insight.

Schema modifications and migration problems are just two examples of the causes of data downtime. Data pipelines can be difficult due to their size and complexity. Data downtime must be continuously monitored, and it must be reduced through automation.

4. Ambiguous Data

Even with thorough oversight, some errors can still occur in massive databases or data lakes. For data streaming at a fast speed, the issue becomes more overwhelming. Spelling mistakes can go unnoticed, formatting difficulties can occur, and column heads might be deceptive. This unclear data might cause a number of problems for reporting and analytics.

5. Duplicate Data

Streaming data, local databases, and cloud data lakes are just a few of the sources of data that modern enterprises must contend with. They might also have application and system silos. These sources are likely to duplicate and overlap each other quite a bit. For instance, duplicate contact information has a substantial impact on customer experience. If certain prospects are ignored while others are engaged repeatedly, marketing campaigns suffer. The likelihood of biased analytical outcomes increases when duplicate data are present. It can also result in ML models with biased training data.

6. Too Much Data

While we emphasize data-driven analytics and its advantages, a data quality problem with excessive data exists. There is a risk of getting lost in an abundance of data when searching for information pertinent to your analytical efforts. Data scientists, data analysts, and business users devote 80% of their work to finding and organizing the appropriate data. With an increase in data volume, other problems with data quality become more serious, particularly when dealing with streaming data and big files or databases.

7. Inaccurate Data

For highly regulated businesses like healthcare, data accuracy is crucial. Given the current experience, it is more important than ever to increase the data quality for COVID-19 and later pandemics. Inaccurate information does not provide you with a true picture of the situation and cannot be used to plan the best course of action. Personalized customer experiences and marketing strategies underperform if your customer data is inaccurate.

8. Hidden Data

The majority of businesses only utilize a portion of their data, with the remainder sometimes being lost in data silos or discarded in data graveyards. For instance, the customer service team might not receive client data from sales, missing an opportunity to build more precise and comprehensive customer profiles. Missing out on possibilities to develop novel products, enhance services, and streamline procedures is caused by hidden data.

9. Finding Relevant Data

Finding relevant data is not so easy. There are several factors that we need to consider while trying to find relevant data, which include -

- Relevant Domain, Relevant demographics
- Relevant Time period and so many more factors that we need to consider while trying to find relevant data.

Data that is not relevant to our study in any of the factors render it obsolete and we cannot effectively proceed with its analysis. This could lead to incomplete research or analysis, re-collecting data again and again, or shutting down the study.

10. Deciding the Data to Collect

Determining what data to collect is one of the most important factors while collecting data and should be one of the first factors while collecting data. We must choose the subjects the data will cover, the sources we will be used to gather it, and the quantity of information we will require. Our responses to these queries will depend on our aims, or what we expect to achieve utilizing your data. As an illustration, we may choose to gather information on the categories of articles that website visitors between the ages of 20 and 50 most frequently access. We can also decide to compile data on the typical age of all the clients who made a purchase from your business over the previous month.

What are the Key Steps in the Data Collection Process?

1. Decide What Data You Want to Gather

The first thing that we need to do is decide what information we want to gather. We must choose the subjects the data will cover, the sources we will use to gather it, and the quantity of information that we would require. For instance, we may choose to gather information on the categories of products that an average e-commerce website visitor between the ages of 30 and 45 most frequently searches for.

2. Establish a Deadline for Data Collection

The process of creating a strategy for data collection can now begin. We should set a deadline for our data collection at the outset of our planning phase. Some forms of data we might want to continuously collect. We might want to build up a technique for tracking transactional data and website visitor statistics over the long term, for instance. However, we will track the data throughout a certain time frame if we are tracking it for a particular campaign. In these situations, we will have a schedule for when we will begin and finish gathering data.

3. Select a Data Collection Approach

We will select the data collection technique that will serve as the foundation of our data gathering plan at this stage. We must take into account the type of information that we wish to gather, the time period during which we will receive it, and the other factors we decide on to choose the best gathering strategy.

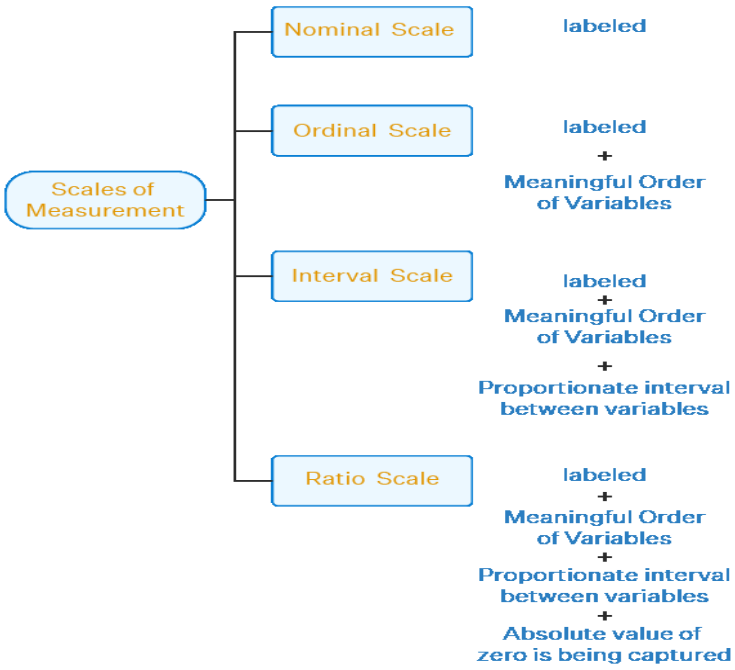
4. Gather Information

Once our plan is complete, we can put our data collection plan into action and begin gathering data. In our DMP, we can store and arrange our data. We need to be careful to follow our plan and keep an eye on how it's doing. Especially if we are collecting data regularly, setting up a timetable for when we will be checking in on how our data gathering is going may be helpful. As circumstances alter and we learn new details, we might need to amend our plan.

5. Examine the Information and Apply Your Findings

It's time to examine our data and arrange our findings after we have gathered all of our information. The analysis stage is essential because it transforms unprocessed data into insightful knowledge that can be applied to better our marketing plans, goods, and business judgments. The analytics tools included in our DMP can be used to assist with this phase. We can put the discoveries to use to enhance our business once we have discovered the patterns and insights in our data.

10. Data Categororization: NOIR



Variable	Data type	Description	Examples
Categorical	Nominal	Named categories with no implied order	Blood groups, breed, gender, neuter status
	Ordinal	Ordered categories where the differences between categories are not necessarily equal	Scoring systems, cancer staging, onset of disease (peracute, acute, chronic)
Continuous	Interval	Equal distances between values but the zero point is arbitrary	IQ, ordinal data with equal-appearing categories
	Ratio	Above as for interval and a meaningful zero; data usually obtained by measurement	Weight, age, temperature, blood pressure

Nominal Scale of Measurement

A nominal scale of measurement is used for qualitative data. It does not give any numerical meaning to the data. Using the nominal scale of measurement, the data can be classified but cannot be added, subtracted, multiplied, or divided. It can cover a wide variety of qualitative data. Some of the situations where nominal measurement scale can be used are given below:

- ✓ Study to find the country of birth of people in a town
- ✓ In collecting data on the eye color of people
- ✓ Classifying people into categories like male/female, working-class population/unemployed, vaccinated/unvaccinated people, etc.

Some of the properties of the nominal scale of measurement are given below:

- ✓ It can categorize variables but does not put them in any order.
- ✓ It does not show any numerical value.
- ✓ It is used for qualitative data.

Binary scale

A nominal variable with exactly two mutually exclusive categories that have no logical order is known as binary variable. Examples: Switch: {ON, OFF}, Attendance: {True, False}, Entry: {Yes, No}

A Binary variable is a special case of a nominal variable that takes only two possible values.

Symmetric and Asymmetric Binary Scale

Different binary variables may have unequal importance.

If two choices of a binary variable have equal importance, then it is called symmetric binary variable.

Example: Gender = {male, female} // usually of equal probability.

If the two choices of a binary variable have unequal importance, it is called asymmetric binary variable.

Example: Food preference = {V, NV}

Ordinal Scale of Measurement

The ordinal scale of measurement groups the data into order or rank. It contains the property of nominal scale as well, which is to classify data variables into specific labels. And in addition to that, it organizes data into groups though it does not have any numerical value. For example, the study of people's satisfaction with a company's product on a scale of #1 - Very happy, #2 - satisfactory, #3 - neutral, #4 - unhappy, and #5 - extremely dissatisfied. This is an example of an ordinal scale of measurement. This measurement scale can be used for the following purposes:

- ✓ Ranks of players in a race.
- ✓ Data collection on variables such as hottest to coldest, richest to poorest, etc.
- ✓ Data on people's satisfaction with any product, person, or government.

Ordered nominal data are known as ordinal data and the variable that generates it is called ordinal variable.

Example: Shirt size = {S, M, L, XL, XXL}

Some of the properties of the ordinal measurement scale are listed below:

- ✓ It displays the order or rating of the variables.
- ✓ It does not give any numerical value to the data. So, it is also used for qualitative data as similar to nominal measurement scale.
- ✓ It contains variables that can be placed in order like heaviest to lightest, ranks of players or students, etc.

Interval Scale of Measurement

The interval scale of measurement includes those values that can be measured in a specific interval, for example, time, temperature, etc. It shows the order of variables with a meaning proportion or difference between them. For example, on a temperature scale, the difference between 20 °C and 30 °C is the same as the difference between 50°C ad 60°C. It is an example of an interval measurement scale. On the other hand, the difference between the scores of the first two rankers in a race and the two runner-ups will be different, which is an example of an ordinal scale.

Some of the properties of the interval scale of measurement are listed below:

- ✓ It includes the properties of both nominal and ordinal scales.
- ✓ It shows meaningful divisions between variables.
- ✓ The difference between the variables can be presented in numerical terms.
- ✓ It includes variables that can be added or subtracted from each other.
- ✓ It gives a meaning to 'Zero" which was not possible in the above two scales. For example, zero degrees of temperature.

Ratio Scale of Measurement

The ratio scale is the most comprehensive scale among others. It includes the properties of all the above three scales of measurement. The unique feature of the ratio scale of measurement is that it considers the absolute value of zero, which was not the case in the interval scale. When we measure the height of the people, 0 inches or 0 cm means that the person does not exist. On the interval scale, there are values possible on both sides of 0, for example, temperature could be negative as well. While the ratio scale does not include negative numbers because of its feature of showing absolute zero. An example of the ratio measurement scale is determining the weight of people from the following options: less than 20 kgs, 20 - 40 kgs, 40 - 60 kgs, 60 - 80 kgs, and more than 80 kgs.

Some of the properties of the ratio scale of measurement are listed below:

- ✓ It is used for quantitative data.
- ✓ It shows the absolute value of zero which means if the value is 0, it's nothing.
- ✓ The variables can be added, subtracted, multiplied, or divided. In addition to these, calculation of mean, median, and mode is also possible with this scale.
- ✓ it doesn't include negative numbers because of the feature of true zero value.

Types of Data:

Qualitative or Categorical Data: Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Quantitative or Numerical Data: Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

Discrete Data

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

Example: Number of students in the class

Continuous Data

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

Unit 2

Descriptive Statistics

Statistics is the science, or a branch of mathematics, that involves collecting, classifying, analyzing, interpreting, and presenting numerical facts and data.

The study of numerical and graphical ways to describe and display your data is called descriptive statistics. It describes the data and helps us understand the features of the data by summarizing the given sample set or population of data.

Types of descriptive statistics

There are 3 main types of descriptive statistics:

[1] The distribution concerns the frequency of each value.

[2] The central tendency concerns the averages of the values.

[3] The variability or dispersion concerns how spread out the values are.

Distribution (also called Frequency Distribution)

Datasets consist of a distribution of scores or values. Statisticians use graphs and tables to summarize the frequency of every possible value of a variable, rendered in percentages or numbers.

Measures of central tendency estimate a dataset's average or center, finding the result using three methods: mean, mode, and median.

Variability (also called Dispersion)

The measure of variability gives the statistician an idea of how spread out the responses are. The spread has three aspects — range, standard deviation, and variance.

Research example

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

- ✓ Go to a library
- ✓ Watch a movie at a theater
- ✓ Visit a national park

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

Measures of central tendency

Advantages and disadvantages of mean, median and mode.

Mean is the most commonly used measures of central tendency. It represents the average of the given collection of data. Median is the middle value among the observed set of values and is calculated by arranging the values in ascending order or in descending order and then choosing the middle value.

The most frequent number occurring in the data set is known as the mode.

Data	Advantages	Disadvantages
Mean	Takes account of all values to calculate the average.	A very small or very large value can affect the mean.
Median	The median is not affected by very large or very small values.	Since the median is an average of position, therefore arranging the data in ascending or descending order of magnitude is time-consuming in the case of a large number of observations.
Mode	The only averages that can be used if the data set is not in numbers.	There can be more than one mode, and there can also be no mode which means the mode is not always representative of the data.

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

4

Measures of Central Tendency

In the previous chapters, data collection and presentation of data were discussed. Even after the data have been classified and tabulated one often finds too much details for many uses that may be made of the information available. We, therefore, frequently need further analysis of the tabulated data. One of the powerful tools of analysis is to calculate a single *average value* that represents the entire mass of data. The word average is very commonly used in day-to-day conversation. For example, we often talk of average work, average income, average age of employees, etc. An 'average' thus is a single value which is considered as the most representative or typical value for a given set of data. Such a value is neither the smallest nor the largest value, but is a number whose value is somewhere in the middle of the group. For this reason an average is frequently referred to as a measure of central tendency or central value. Measures of central tendency show the tendency of some central value around which data tends to cluster.

OBJECTIVES OF AVERAGING

There are two main objectives of the study of averages :

(*i*) *To get one single value that describes the characteristics of the entire data.* Measures of central value, by condensing the mass of data in one single value, enable us to get an idea of the entire data. Thus one value can represent thousands, lakhs and even millions of values. For example, it is impossible to remember the individual incomes of millions of earning people of India and even if one could do it there is hardly any use. But if the average income is obtained, we get one single value that represents the entire population. Such a figure would throw light on the standard of living of an average Indian.

(*ii*) *To facilitate comparison.* Measures of central value, by reducing the mass of data in one single figure, enable comparisons to be made. Comparison can be made either at a point of time or over a period of time. For example, the figure of average sales for December may be compared with the sales figures of previous months or with the sales figure of another competitive firm.

CHARACTERISTICS OF A GOOD AVERAGE

Since an average is a single value representing a group of values, it is desirable that such a value satisfies the following properties :

(*i*) *It should be easy to understand.* Since statistical methods are designed to simplify complexity it is desirable that an average be such that can be readily understood, its use is bound to be very limited.

(*ii*) *It should be simple to compute.* Not only an average should be easy to understand but also should be simple to compute so that it can be used widely. However, though ease of computation is desirable, it should not be sought at the expense of other advantages, *i.e.*, if in the interest of greater accuracy, use of a more difficult average is desirable one should prefer that.

(*iii*) *It should be based on all the observations.* The average should depend upon each and every observation so that if any of the observation is dropped average itself is altered.

(*iv*) *It should be rigidly defined.* An average should be properly defined so that it has one and only one interpretation. It should preferably be defined by an algebraic formula so that if different people compute the average from the same figures they all get the same answer (barring arithmetical mistakes).

(*v*) *It should be capable of further algebraic treatment.* We should prefer to have an average that could be used for further statistical computations. For example, if we are given separately the figures of average income and number of employees of two of more companies we should be able to compute the combined average.

(*vi*) *It should have sampling stability.* We should prefer to get a value which has what the statisticians call 'sampling stability'. This means that if we pick 10 different group of college students, and compute the average of each group, we should expect to get approximately the same values. It does not mean, however, that there can be no difference in the value of different samples. There may be some difference but those averages in which this difference, technically called 'sampling fluctuation,' is less are considered better than those in which this difference is more.

(*vii*) *It should not be unduly affected by the presence of extreme values.* Although each and every observation should influence the value of the average, none of the observations should influence it unduly. If one or two very small or very large observations unduly affect the average, *i.e.*, either increase its value or reduce its value, the average cannot be really typical of the entire set of data. In other words, extremes may distort the average and reduce its usefulness.

The following are the important measures of central tendency which are generally used in business :

A. Arithmetic mean, B. Median, C. Mode, D. Geometric mean, and E. Harmonic mean

A. ARITHMETIC MEAN

The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what the statisticians call the arithmetic mean. Its value is obtained by adding together all the observations and by dividing this total by the number of observations.

Calculation of Arithmetic Mean—Ungrouped Data

For ungrouped data, arithmetic mean may be computed by applying any of the following methods :

(*i*) Direct method,

(*ii*) Short-cut method.

(*i*) Direct Method : The arithmetic mean, often simply referred to as mean, is the total of the values of a set of observations divided by their total number of observations. Thus, if $X_1, X_2, \ldots\ldots, X_N$ represent the values of N items or observations, the arithmetic mean denoted by $\bar{X}$ is defined as :

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

If the subscripts are dropped, the formula becomes :

$$\bar{X} = \frac{\Sigma X}{N}$$

It may be pointed out that in keeping with standard statistical practice, the symbol $\bar{X}$ will represent throughout this text the arithmetic mean of a set of observations.

Illustration 1. The monthly income (in rupees) of 10 employees working in a firm is as follows :

4487 4493 4502 4446 4475 4492 4572 4516 4468 4489

Find the average monthly income.

Solution. Let income be denoted by X.

$\Sigma X = 4487+4493+4502+4446+4475+4492+4572+4516+4468+4489 = 44,940$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{44940}{10} = 4494$$

Hence the average monthly income is Rs. 4,494.

***(ii)* Short-cut Method.** The arithmetic mean* can also be calculated by taking deviations from any arbitrary point in which case the formula shall be :

$$\bar{X} = A + \frac{\Sigma d}{N}$$

where $d = (X - A)$

and A = Arbitrary point (or assumed mean).

It should be noted that any value can be taken as arbitrary point and the answer would be the same as obtained by the direct method.

Illustration 2. Calculate average monthly income by the short-cut method from data of Illustration 1 taking deviations from 4460 as the arbitrary point.

Solution.

CALCULATION OF AVERAGE INCOME

X (Rs.)	$(X-4460)$ d
4487	+27
4493	+33
4502	+42
4446	−14
4475	+15
4492	+32
4572	+112
4516	+56
4468	+ 8
4489	+29
	$\Sigma d = +340$

$$\bar{X} = A + \frac{\Sigma d}{N} = 4460 + \frac{340}{10} = 4460 + 34 = \text{Rs. } 4494.$$

One may find that the short-cut method takes more time as compared to direct method. However, this is true only for ungrouped data. In case of grouped data, considerable saving in time is possible by adopting the short-cut method.

Calculation of Arithmetic Mean—Grouped Data

For grouped data, arithmetic mean may be computed by applying any of the following methods :

(*i*) Direct method,

(*ii*) Short-cut method.

---

*This formula is derived as follows :

Let $d = X - A$ or $X = A + d$

Taking summation of both sides and dividing by N, we get

$$\frac{\Sigma X}{N} = \frac{\Sigma A}{N} + \frac{\Sigma d}{N} \quad \text{or} \quad \bar{X} = A + \frac{\Sigma d}{N}$$

The population mean is denoted by μ (μ is the Greek letter mu) and the sample mean by $\bar{X}$.

(*i*) **Direct Method.** When direct method is used

$$\bar{X} = \frac{\sum fX}{N}$$

where X = mid-point of various classes.
f = the frequency of each class.
N = the total frequency.

Note. For computing mean in the case of grouped data the mid-points of the various classes are taken as representative of that particular class. The reason is that when the data are grouped, the exact frequency with which each value of the variable occurs in the distribution in unknown. We only know the limits within which a certain number of frequencies occur. For example, when we say that the number of persons within the income group 4000–4500 is 50 we cannot say as to how many persons out of 50 are getting 4001, 4002, 4003, etc. We, therefore, make an assumption while calculating arithmetic mean that the frequencies within each class are distributed uniformly or evenly over the range of the class-interval, *i.e.*, there will be as many observations below the mid-point as above it. Unless such an assumption is made, the value of mean cannot be computed.

Illustration 3. The following are the figures of profits earned by 1,400 companies during 2003-04.

Profits (Rs. Lakhs)	*No. of Companies*	*Profits (Rs. Lakhs)*	*No. of Companies*
200–400	500	1,000–1,200	100
400–600	300	1,200–1,400	80
600–800	280	1,400–1,600	20
800–1000	120		

Calculate the average profits for all the companies.

Solution.

CALCULATION OF AVERAGE PROFITS

Profits (Rs. Lakhs)	*Mid-points X*	*No. of Companies f*	*fX*
200–400	300	500	1,50,000
400–600	500	300	1,50,000
600–800	700	280	1,96,000
800–1,000	900	120	1,08,000
1,000–1,200	1100	100	1,10,000
1,200–1,400	1300	80	1,04,000
1,400–1,600	1500	20	30,000
		$N = 1,400$	$\Sigma fX = 8,48,000$

$$\bar{X} = \frac{\sum fX}{N} = \frac{8,48,000}{1,400} = 605.71$$

Thus, the average profit is Rs. 605.71 lakhs.

• $\bar{X} = \frac{\sum fX}{N}$ (direct method)

Now $d = \frac{X - A}{i}$, $\therefore \quad X = A + id$

Substituting the value of X in the direct method, we get

$$\bar{X} = \frac{\sum f(A+id)}{N} = \frac{\sum fA}{N} + \frac{\sum fd}{N} \times i$$

Hence $$\bar{X} = A + \frac{\sum fd}{N} \times i \qquad (\because \Sigma f = N)$$

(*ii*) **Short-cut Method*.** When short-cut method is used, the following formula is applied :

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

where $$d = \frac{X - A}{i}$$

and i = size of the equal class interval.

Illustration 4. Calculate the average profit by the short-cut method from the data of Illustration 3.

Solution. **CALCULATION OF AVERAGE PROFITS**

Profits (Rs. Lakhs)	*Mid points X*	*f*	*(X–900)/200 d*	*fd*
200–400	300	500	– 3	– 1,500
400–600	500	300	– 2	– 600
600–800	700	280	– 1	– 280
800–1000	**900**	120	0	0
1000–1200	1100	100	+ 1	+ 100
1200–1400	1300	80	+ 2	+ 160
1400–1600	1500	20	+ 3	+ 60
		N = 1,400		Σfd = – 2,060

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 900 - \frac{2,060}{1,400} \times 200 = 900 - 294.29 = \text{Rs. } 605.71$$

Hence the average profit is Rs. 605.71 lakhs.

Correcting Incorrect Values

It sometimes happens that due to an oversight or mistake in copying certain wrong values are taken while calculating the mean. The problem is how to find out the correct mean. The process is very simple. From ΣX deduct wrong observations and add correct observations and then divide the correct ΣX by the number of observations and the result so obtained will give the value of the correct mean.

Illustration 5. The average weekly for a group of 25 persons working in a factory was calculated to be Rs. 378.40. It was later discovered that one figure was misread as 160 instead of the correct value Rs. 200. Calculate average wage.

Solution. $\Sigma X = N\bar{X} = 25 \times 378.4 = 9460$

Less: Incorrect figure	160
	9300
Add: Correct figure	200
Total	9500

$\therefore$ *Correct* ΣX = 9500

Hence correct average = $\frac{9500}{25}$ = 380.

(*b*) The mean of 200 observations was 50. Later on, it was discovered that two observations were wrongly read as 92 and 8 instead of 192 and 88. Find out the correct mean.

Solution.

Here $\bar{X}$ = 50 and N = 200

$\Sigma X = 200 \times 50 = 10,000$

Less Incorrect observations	100
	9,900

Add Correct observation correct total = $\frac{280}{10,180}$

Correct mean = $\frac{10,180}{200}$ = 50.9

Mathematical Properties of Arithmetic Mean

The important mathematical properties of arithmetic mean are :

1. The algebraic sum of the deviations of all the observations from arithmetic mean is always zero, *i.e.*, $\Sigma (X - \bar{X}) = 0$. This shall be clear from the following example :

X	$(X - \bar{X})$
10	− 20
20	− 10
30	0
40	+ 10
50	+ 20
$\Sigma X = 150$	$\Sigma (X - \bar{X}) = 0$

Here $\bar{X} = \frac{\Sigma X}{N} = \frac{150}{5} = 30$. When the sum of the deviations from the actual mean, *i.e.*, 30, is taken it comes out to be zero. It is because of this property that the mean is characterised as a *point of balance, i.e.,* the sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

2. The sum of the squared deviations of all the observations from arithmetic mean is minimum, that is, less than the squared deviations of all the observations from any other value than the mean. The following example would clarify the point :

X	$(X - \bar{X})$	$(X - \bar{X})^2$
2	− 2	4
3	− 1	1
4	0	0
5	+ 1	1
6	+ 2	4
$\Sigma X = 20$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 10$

The sum of the squared deviations is equal to 10 in the above case. If the deviations are taken from any other value, the sum of the squared deviations would be greater than 10. This is known as the least square property of the arithmetic mean and becomes the basis for defining the concept of standard deviation.

3. If we have the arithmetic mean and number of observations of two or more than two related groups, we can compute combined average of these groups by applying the following formula :

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$\bar{X}_{12}$ = Combined mean of the two groups.

$\bar{X}_1$ = Arithmetic mean of the first group.

$\bar{X}_2$ = Arithmetic mean of the second group.

N_1 = Number of observations in the first group.

N_2 = Number of observations in the second group.

The following example will illustrate the application of the above formula :

Illustration 6(*a*). There are two branches of a company employing 100 and 80 employees respectively. If arithmetic means of the monthly salaries paid by two branches are Rs. 4570 and Rs. 6750 respectively, find the arithmetic mean of the salaries of the employees of the company as a whole.

Solution. We should compute the combined mean. The formula is

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Given $N_1 = 100,\ \bar{X}_1 = 4570,\ N_2 = 80,\ \bar{X}_2 = 6750$

$$\therefore \quad \bar{X}_{12} = \frac{(100\times 4570)+(80\times 6750)}{100+80} = \frac{997000}{180} = 5538.89$$

If we have to find out the combined mean of three related groups, the above formula can be extended as follows :

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

Illustration 6. (*b*) The mean of marks in Statistics of 100 students of a class was 72. The mean of marks of boys was 75, while their number was 70. Find out the mean marks of girls in the class. *(MBA, Osmania Univ, 2006)*

Solution. We are given $N = 100,\ \bar{X}_{12} = 72,\ \bar{X}_1$, *i.e.*, mean marks of boys = 75, N_1 = number of boys = 70. We have to find out the mean marks of girls, *i.e.*, $\bar{X}_2$.

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$72 = \frac{70(75)+30\bar{X}_2}{70+30}$$

$$7200 = 5250+30\,\bar{X}_2$$

$$\bar{X}_2 = \frac{1950}{30} = 65$$

Hence mean marks of girls in the class = 65.

Illustration 6. (*c*) The mean age of a combined group of men and women is 30 years. If mean age of the group of men is 32 and that of the group of women is 25, find out the percentage of men and women in the group.

Solution. Let N_1 represent percentage of men and N_2 percentage of women so that $N_1+N_2 = 100$.

We are given $\bar{X}_{12} = 30$

$\bar{X}_1 = 32$ (mean age of group of men)

$\bar{X}_2 = 25$ (mean age of group of women)

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$30 = \frac{N_1(32) + N_2(25)}{100}$$

$$3000 = 32N_1+(100-N_1)25$$

$$32N_1 + 2500 - 25N_1 = 3000 \quad \text{or} \quad N_1 = 71.43$$

$$N_2 = 100 - 71.43 = 28.57$$

Hence the percentage of men and women is respectively 71.43 and 28.57.

Merits and Limitations of Arithmetic Mean

The arithmetic mean is the most popular average in practice. It is due to the fact that it possesses first six out of seven characteristics of a good average and no other average possesses such a large number of characteristics.

However, arithmetic mean is unduly affected by the presence of extreme values. Also in open-end frequency distribution, it is difficult to compute mean without making assumption regarding the size of the class-interval of the open-end classes. The arithmetic mean is usually neither the most commonly occurring value nor the middle value in a distribution and in extremely asymmetrical distribution, it is not a good measure of central tendency.

Weighted Arithmetic Mean

One of the limitations of the arithmetic mean discussed above is that it gives equal importance to all the observations. But there are cases where the relative importance of the different observations is not the same. When this is so, we compute weighted arithmetic mean. The terms 'weight' stands for the relative importance of the different observations. The formula for computing weighted arithmetic mean is :

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$$

where $\bar{X}_w$ represents the weighted arithmetic mean

X = The variable.

W = Weights attached to the variable X.

An important problem that arises while using weighted mean is regarding selection of weights. Weights may be either actual or arbitrary, *i.e.,* estimated. Needless to say, if actual-weights are available, nothing like this. However, in the absence of actual-weights, arbitrary or imaginary weights may be used. The use of arbitrary weights may lead to some error, but this is better than no weights at all. In practice, it is found that if weights are intelligently assigned keeping the phenomena in view, the error involved will be so small that it can be easily overlooked.

Weighted mean is specially useful in problems relating to the construction of index numbers and standardised birth and death rates.

Illustration 7. A contractor employs three types of workers—male, female and children. To a male worker he pays Rs. 200 per day, to a female worker Rs. 150 per day and to a child worker Rs. 100 per day. What is the average wage per day paid by the contractor?

Solution. The average wage is not the simple arithmetic mean, *i.e.,* $\frac{200+150+100}{3}$ = Rs. 150 per day. If we assume that the number of male, female and child workers is the same, this answer would be correct. For example, if we take 10 workers in each case then the average wage would be

$$\bar{X} = \frac{(10\times 200)+(10\times 150)+(10\times 100)}{10+10+10} = \frac{2000+1500+1000}{30} = \frac{4500}{30} = \text{Rs. } 150$$

However, the number of male, female and child workers employed is generally different. If we know how many workers of each type are employed by the contractor in question, nothing like this. However, in the absence of this we take assumed weights. Let us assume that the number of male, female and child workers employed is 20, 15 and 5, respectively. The average wage would be the weighted mean calculated as follows :

Wage per day (Rs.) X	*No. of workers* W	WX
200	20	4000
150	15	2250
100	5	500
	$\Sigma W = 40$	$\Sigma WX = 6750$

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W} = \frac{6750}{40} = 168.75$$

Hence the average wage per day paid by the contractor is Rs. 168.75 to all types of workers.

B. MEDIAN

The median is the measure of central tendency which appears in the "middle" of an ordered sequence of values. That is, half of the observations in a set of data are lower than it and half of the observations are greater than it.

As distinct from the arithmetic mean which is calculated from the *value of every* observation in the series, the median is what is called a *positional* average. The term 'position' refers to the

place of a value in a series. The place of the median in a series is such that an equal number of observations lie on either side of it. For example, if the income of five persons is Rs. 7000, 7200, 7500, 7600, 7800, then the median income would be Rs. 7500. Changing any or both of the first two values with any other numbers with value of 7500 or less and/or changing any of the last two values to any other values with values of 7500 and more, would not affect the value of the median which would remain 7500. In contrast, in case of arithmetic mean the change in value of single observation would cause the value of the mean to be changed. Median is thus the central value of the distribution or the value that divides the distribution into two equal parts. If there are even number of observations in a series, there is no actual value exactly in the middle of the series and as such the median is indeterminate. In this case, the median is arbitrarily taken to be halfway between the two middle observations. For example, if there are 10 observations in a series, the median position is 5.5, that is the median value is halfway between the value of the observations that are 5th and 6th in order of magnitude. Thus when N is odd, the median is in an actual value with the remainder of the series in two equal parts on either side of it. If N is even, then the median is a derived figure, *i.e.*, half the sum of two values.

Calculation of Median—Ungrouped Data

Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)

Apply the formula : Median = Size of $\frac{N+1}{2}$th observation.

Illustration 8. From the following data of wages of 7 workers, compute the median wage :

Wages (in Rs.) 4600 4650 4580 4690 4660 4606 4640

Solution :

CALCULATION OF MEDIAN

S. No.	*Wages arranged in ascending order*
1	4580
2	4600
3	4606
4	4640
5	4650
6	4660
7	4690

Median = Size of $\frac{N+1}{2}$th observation = $\frac{7+1}{2}$ = 4th observation.

Value of 4th observation is 4640. Hence median wages = Rs. 4640.

In the above illustration, the number of observations was odd and, therefore, it was possible to determine the value of 4th observation. When the number of observations is even, for example, if in the above case the number of observations are 8 the median would be the value of $\frac{8+1}{2}$ = 4.5th observation. For finding out the value of 4.5th observation, we shall take the average of 4th and 5th observations. Hence the median shall be

$$\frac{4640+4650}{2} = 4645$$

Calculation of Median—Grouped Data

Determine the particular class in which the value of median lies. Use $\frac{N}{2}$ to locate the median class and not $\frac{N+1}{2}$ because in the use of grouped data it is $N/2$ which divides the area of the curve into two equal parts.

Apply the following formula for determining the exact value of median :

$$\text{Median} = L + \frac{N/2 - p.c.f.}{f} \times i$$

L = Lower limit of median class, *i.e.*, the class in which the middle observation in the distribution lies.

$p.c.f.$ = Preceding cumulative frequency to the median class.

f = Frequency of the median class.

i = The class-interval of the median class.

Illustration 9. (*a*) 1,500 workers are working in an industrial establishment. Their age is classified as follows :

Age (yrs.)	*No. of workers*	*Age (yrs.)*	*No. of workers*
18–22	120	38–42	184
22–26	125	42–46	162
26–30	280	46–50	86
30–34	260	50–54	75
34–38	155	54–58	53

Calculate the median age.

Solution :

CALCULATION OF MEDIAN AGE

Age group	*f*	*c.f.*
18–22	120	120
22–26	125	245
26–30	280	525
30–34	260	785
34–38	155	940
38–42	184	1,124
42–46	162	1,286
46–50	86	1,372
50–54	75	1,447
54–58	53	1,500

Median = Size of $\frac{N}{2}$ th observation = $\frac{1,500}{2}$ = 750th observation.

Hence median lies in the class 30–34.

$$\text{Median} = L + \frac{N/2 - p.c.f}{f} \times i = 30 + \frac{750 - 525}{260} \times 4 = 30 + 3.46 = 33.46$$

Hence the median age of the workers is 33.46 years.

(*b*) Calculate the median from the following data pertaining to the profits (in crore Rs.) of 125 companies :

Profits (Rs. crore)	*No. of companies*
less than 10	4
less than 20	16
less than 30	40
less than 40	76
less than 50	96
less than 60	112
less than 70	120
less than 80	125

(MBA, MD Univ., 2000)

Solution :

CALCULATION OF MEDIAN

Profits (Rs. Crore)	*No. of companies (f)*	*c.f.*
0 — 10	4	4
10 — 20	12	16
20 — 30	24	40
30 — 40	36	76
40 — 50	20	96
50 — 60	16	112
60 — 70	8	120
70 — 80	5	125

Med. = Size of $\frac{N}{2}$ th observation = $\frac{125}{2}$ = 62.5th observation.

Median lies in the class 30—40.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$$L = 30, N/2 = 62.5, p.c.f. = 40, f = 36, i = 10$$

$$\therefore \quad \text{Med.} = 30 + \frac{62.5 - 40}{36} \times 10 = 30 + 6.25 = 36.25.$$

Hence 50% of the companies have profits upto Rs. 36.25 crores and the remaining 50% of the companies have profits more than Rs. 36.25 crores.

Merits and Limitations of Median

The median is superior to arithmetic mean in certain respects. For example, it is especially useful in case of open-end distribution and also it is not influenced by the presence of extreme values. In fact when extreme values are present in the data, the median is a more satisfactory measure of central tendency than the mean.

The sum of the deviations of observations from median (ignoring signs) is *minimum*. In other words, the absolute deviation of observations from the median is less than from any other value in the distribution. For example, the median of items 4, 6, 8, 10 and 12 is 8. The deviations from 8 ignoring signs are 4, 2, 0, 2, 4 and the total is 12. This total will be smaller than the one obtained if deviations are taken from any other value. Thus, if deviations are taken from 7, the deviations ignoring signs would be 3, 1, 1, 3, 5 and the total is 13. In an estimation situation, if one is interested in minimising the absolute amount of error and the sign of the error is not particularly important, then the median is preferable to arithmetic mean.

However, since median is a positional average, its value is not determined by each and every observation. Also median is not capable of algebraic treatment. For example, median cannot be used for determining the combined median of two or more groups. Also the median is less reliable average than the mean for estimation purposes since it is more affected by sampling fluctuations. Furthermore, the median tends to be rather unstable value if the number of observations is small.

Related Positional Measures or Quantities

Besides median, there are other measures which divide a series into a equal number of parts. Important amongst these are quartiles, deciles and percentiles. These quartiles, deciles and percentiles are all special cases of *quantities*. Quartiles are those values of the variate which divide the total frequency into four equal parts, deciles divide the total frequency in 10 equal parts and the percentiles divide the total frequency in 100 equal parts. Just as one point divides a series into two parts, three points would divide it into four parts, 9 points into 10 parts and 99 points into 100 parts consequently there are only 3 quartiles, 9 deciles and 99 percentiles for a series. The quartiles are denoted by symbol Q, deciles by D and percentiles by P. The subscripts 1, 2, 3, etc., beneath Q, D and P would refer to the particular value that we want to compute. Thus Q_1 would denote first quartile Q_2 second quartile, Q_3 third quartile, D_1 first decile, D_8 eighth decile, P_1 first percentile, etc.

Graphically any set of these partition values serves to divide the area of the frequency curve or histogram into equal parts. If vertical lines are drawn at each quartile, for example, the area of the histogram will be divided by these lines into four equal parts. The 9 deciles divide the area of the histogram or frequency curve into 10 equal parts and the 99 percentiles divide the area into 100 equal parts.

In economics and business, quartiles are more widely used than deciles and percentiles. The quartiles are the points on the X-scale that divide the distribution into four equal parts. Obviously there are three quartiles, the second coinciding with the median. More precisely stated, the lower quartile, Q_1 is that point on the X-scale such that one-fourth of the total frequency is less than Q_1 and three-fourths is greater than Q_1. The upper quartile, Q_3, is that point on the X-scale such that three-fourths of the total frequency is below Q_3 and one-fourth is above it.

The deciles and percentiles are important in psychological and educational statistics concerning grades, rates, scores and ranks ; they are of use in economics and business in personnel department, productivity ratings and other situations.

Computation of Quartiles, Deciles, Percentiles, etc.

The procedure for computing quartiles, deciles, etc., is the same as for median.

For grouped data, the following formulae are used for quartiles, deciles and percentiles ;

$$Q_j = L + \frac{\frac{jN}{4} - p.c.f.}{f} \times i \qquad \text{for } j = 1, 2, 3$$

$$D_k = L + \frac{\frac{kN}{10} - p.c.f.}{f} \times i \qquad \text{for } k = 1, 2, \ldots, 9$$

$$P_l = L + \frac{\frac{lN}{100} - p.c.f.}{f} \times i \qquad \text{for } l = 1, 2, \ldots, 99$$

where the symbols have their usual meanings and interpretation.

Illustration 10. The profits earned by 100 companies during 2009-10 are given below :

Profits (Rs. lakhs)	*No. of companies*	*Profits (Rs. lakhs)*	*No. of companies*
20 — 30	4	60 — 70	15
30 — 40	8	70 — 80	10
40 — 50	18	80 — 90	8
50 — 60	30	90 — 100	7

Calculate Q_1, median, D_4 and P_{80} and interpret the values. *(MBA, D.U., 2001)*

Solution.

CALCULATION OF Q_1, Q_2, D_4 AND P_{80}

Profits (Rs. lakhs)	f	*c.f.*
20 – 30	4	4
30 – 40	8	12
40 – 50	18	30
50 – 60	30	60
60 – 70	15	75
70 – 80	10	85
80 – 90	8	93
90 – 100	7	100

Q_1 = Size of $N/4$th observation = $\frac{100}{4}$ = 25th observation.

Hence Q_1 lies in the class 40 — 50.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 40 + \frac{25-12}{18} \times 10 = 40 + 7.22 = 47.22$$

25 per cent of the companies earn an annual profit of Rs. 47.22 lakhs or less.

Median or Q_2 = Size of $\frac{2N}{4}$th observation = $\frac{200}{4}$ = 50th observation. Q_2 lies in the class 50—60.

$$Q_2 = L + \frac{2N/4 - p.c.f.}{f} \times i = 50 + \frac{50-30}{30} \times 10 = 50 + 6.67 = 56.67$$

50 per cent of the companies earn an annual profit of Rs. 56.67 lakhs or less.

D_4 = Size of $\frac{4N}{10}$th observation = 40th observation

D_4 lies in the class 50—60.

$$D_4 = L + \frac{4N/10 - p.c.f.}{f} \times i = 50 + \frac{40-30}{30} \times 10 = 50 + 3.33 = 53.33.$$

Thus 40 per cent of the companies earn an annual profit of Rs. 53.33 lakhs or less.

P_{80} = Size of $\frac{80N}{100}$th observation = $\frac{80 \times 100}{100}$ = 80th observation

P_{80} lies in the class 70—80.

$$P_{80} = L + \frac{80N/100 - p.c.f.}{f} \times i = 70 + \frac{80-75}{10} \times 10 = 70 + 5 = 75$$

This means that 80 per cent of the companies earn an annual profit of Rs. 75 lakhs or less and 20 per cent of the companies earn an annual profit of more than Rs. 75 lakhs.

Determination of Median, Quartiles, etc., Graphically

Median can be determined graphically by applying any of the following two methods :

1. Draw two ogives — one by 'less than' method and other by 'more than' method. From the point where both these curves intersect each other, draw a perpendicular on the X-axis. The point where this perpendicular touches the x-axis gives us the value of median.

2. Draw only one ogive by 'less than' method. Take the variable on the X-axis and frequency on the Y-axis. Determine the median value by the formula : median = Size of $\frac{N}{2}$th item. Locate this value on the Y-axis and from it draw a perpendicular on the cumulative frequency curve. From the point where it meets the ogive draw another perpendicular on the X-axis and the point where it meets the X-axis is the median.

The other partition values like quartiles, deciles and percentiles can also be determined graphically.

Illustration 11. Using the data of illustration 10, determine graphically the values of Q_1, Q_2, D_{40} and P_{80}.

Solution. Draw the ogive by the 'less than' method as shown in the graph.

To determine different quartiles, horizontal lines (broken) are drawn from the cumulative frequency values. For example, if we want to determine the value of median, a horizontal line can be drawn from the cumulative frequency value of 0.50 to the less than curve and then extending a vertical line to the horizontal axis. In a similar manner other values can be determined as shown in the graph. Therefore, $Q_1 = 47.22$, $Q_2 = 56.67$, $D_{40} = 53.33$ and $P_{80} = 75$. This may be noted down here that these graphical values are same as obtained by the formulae.

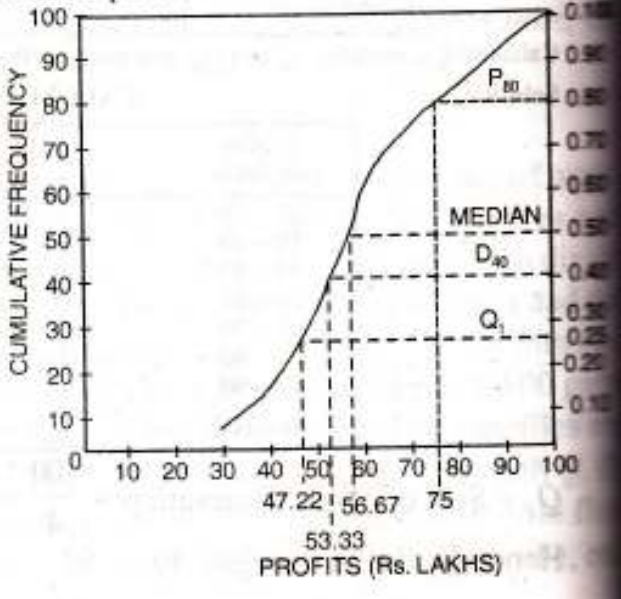


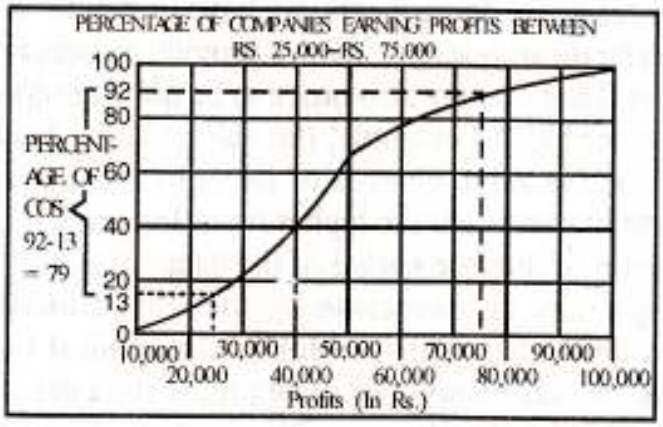
Illustration 12. You are given the net profits earned by some companies. Estimate graphically the percentage of companies getting profits between Rs. 25,000 and Rs. 75,000.

Profits (in Rs.)	*No. of companies*	*Profits (in Rs.)*	*No. of companies*
10,000—20,000	15	60,000—70,000	22
20,000—30,000	35	70,000—80,000	12
30,000—40,000	47	80,000—90,000	11
40,000—50,000	68	90,000—1,00,000	8
50,000—60,000	32		

Solution. Finding percentage from the given data :

Profits less than	*No. of companies*	*Percentage*
Rs. 20,000	15	6.0
" 30,000	50	20.0
" 40,000	97	38.8
" 50,000	165	66.0
" 60,000	197	78.8
" 70,000	219	87.6
" 80,000	231	92.4
" 90,000	242	96.8
" 1,00,000	250	100.0

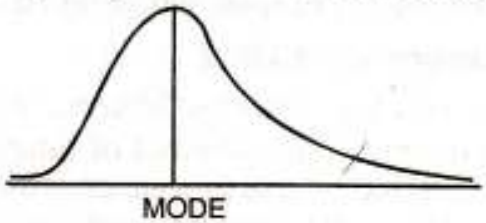
Plotting the data on the graph paper :



The graph shows clearly that the percentage of companies earning profits less than Rs. 75,000 is 92 and the percentage of companies earning profits less than Rs. 25,000 is 13. Thus the percentage of companies making profits between Rs. 25,000 and Rs. 75,000 is (92–13) = 79.

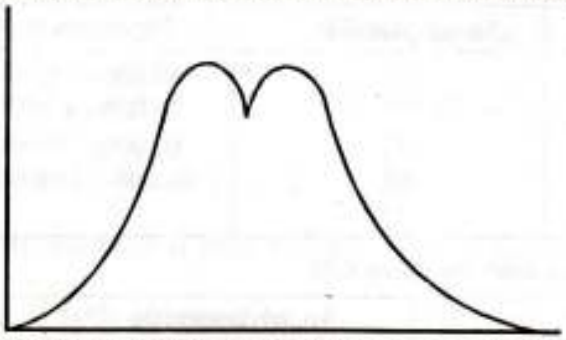
C. MODE

Mode is defined as that value which occurs the maximum number of times, *i.e.*, having the maximum frequency. For example, if we take the values of six different observations as 5, 8, 10, 8, 5, 8, mode will be 8 as it has occurred maximum number of times, *i.e.*, 3 times. Graphically, it is the value on the *X*-axis below the peak, or highest point, of the frequency curve as can be seen from the following diagram.



This interpretation of the statistical mode is analogous to that of the fashion mode. A person dressing with current styles is "in the mode". But a current fashion can be a poor description of what most persons are wearing because of the variety of styles worn by the general public. In statistics, the mode only tells us which single value occurs most often; it may, therefore, represent a majority of the total population.

It is possible that a distribution may be bimodal. This happens when there may be two or more values of equal or nearly equal occurrence as can be seen from the following diagram :



The presence of more than one mode has a special significance in statistical analysis, for it indicates potential trouble. It is usually dangerous to compare bimodal populations or to draw conclusions about them because they usually arise when there is some non-homogeneous factor present in the population.

If the collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused by the taking of too small a sample; the difficulty can be remedied by increasing the sample size. In instances where a distribution is bimodal and nothing can be done to change it, the mode is obviously eliminated as a measure of central tendency.

There are many situations in which arithmetic mean and median fail to reveal the true characteristic of data. For example, when we talk of most common wage, most common income, most common height, most common size of shoe or ready-made garments, we have in mind mode and not the arithmetic mean or median discussed earlier. The mean does not always provide an accurate reflection of the data due to the presence of extreme values. Median may also prove to be quite unrepresentative of the data owing to an uneven distribution of the series. For example, the values in the lower half of a distribution range from, say, Rs. 10 to 100, while the same number of items in the upper half of the series range from Rs. 100 to Rs. 6,000 with most of them near the higher limit. In such a distribution the median value Rs. 100 will provide little indication of the true nature of the data.

Both these shortcomings may be overcome by the use of mode. Mode refers to that value which occurs most frequently in a distribution. Mode is the easiest to compute since it is the value corresponding to the maximum frequency. For example, if the data is :

Size of shoes	:	5	6	7	8	9	10	11
No. of persons	:	10	20	25	40	22	15	6

the modal size is '8' since it appears maximum number of times in the data.

Calculation of Mode

Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially it involves fitting mathematically some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode. These methods have been discussed for ungrouped and grouped data.

Calculation of Mode—Ungrouped Data

For determining mode count, the number of observations the various values repeat themselves, and the value which occurs the maximum number of times is the modal value.

Illustration 13. The following figures relate to the preferences with regard to size of screen (in inches) of T.V. sets of 30 persons selected at random from a locality. Find the modal size of the T.V. screen.

12	20	12	24	29
20	12	20	29	24
24	20	12	20	24
29	24	24	20	24
24	20	24	24	12
24	20	29	24	24

Solution. CALCULATION OF MODAL SIZE

Size in inches	*Tally*	*Frequency*
12	𝍸	5
20	𝍸 |||	8
24	𝍸 𝍸 |||	13
29	||||	4
	Total	30

Since size 24 occurs the maximum number of times, therefore, the modal size of T.V. screen is 24 inches.

Calculation of Mode—Grouped Data

In the case of grouped data, the following formula is used for calculating mode :

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \qquad ...(i)$$

where L = Lower limit of the modal class.

Δ_1 = The difference between the frequency of the modal class and the frequency of the pre-modal class, *i.e.*, preceding class.

Δ_2 = The difference between the frequency of the modal class and the frequency of the post-modal class, *i.e.*, succeeding class.

i = The size of the modal class.

Another form of this formula is :

$$Mo = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \qquad ...(ii)$$

where L = Lower limit of the modal class.

f_1 = Frequency of the modal class.

f_0 = Frequency of the class preceding the modal class.

f_2 = Frequency of the class succeeding the modal class.

While applying the above formula for calculating mode, it is necessary to see that the class intervals are *uniform* throughout. If they are unequal, they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

A distribution having only one mode is called *unimodal.* If it contains more than one mode, it is called *bimodal* or *multimodal.* In the latter case, the value of mode cannot be determined by the above formula and hence mode is *ill-defined* when there is more than one value of mode.

Where mode is ill-defined, its value may be ascertained by the following approximate formula* based upon the relationship between mean, median and mode.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean} \qquad ...(iii)$$

Illustration 14. The following data relate to the sales of 100 companies :

Sales (Rs. lakhs)	*No. of companies*	*Sales (Rs. lakhs)*	*No. of companies*
Below 60	12	66 — 68	10
60—62	18	68—70	3
62—64	25	70—72	2
64—66	30		

Calculate the value of modal sales.

*See page 99.

Solution. Since the maximum frequency 30 is in the class 64—66, therefore, 64—66 is the modal class.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 64,\ \Delta_1 = (30 - 25) = 5,\ \Delta_2 = (30 - 10) = 20,\ i = 2$$

$$\text{Mode} = 64 + \frac{5}{5 + 20} \times 2 = 64 + \frac{10}{25} = 64.4$$

Hence modal sales are Rs. 64.4 lakhs.

Locating Mode Graphically

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are :

1. Draw a histogram of the given data.
2. Draw two lines diagonally on the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.
3. Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis (horizontal scale) which gives us modal value.

Illustration 15. The daily profits in rupees of 100 shops are given as follows :

Profits (in Rs. lakhs)	No. of shops	Profits (in Rs. lakhs)	No. of shops
0—100	12	300—400	20
100—200	18	400—500	17
200—300	27	500—600	6

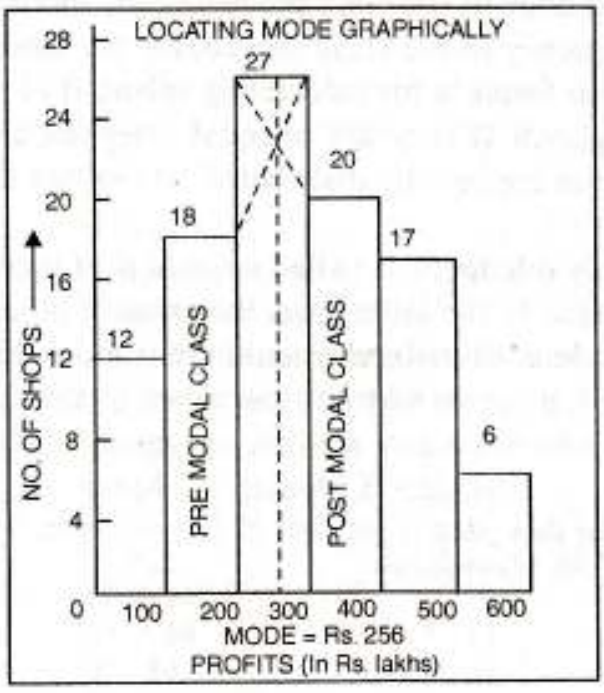
Draw the histogram and thence find the modal value. Check this value by direct calculation.

Solution.

Direct calculation :

Mode lies in the class 200—300.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 200 + \frac{9}{9 + 7} \times 100 = 256.25$$



From the above diagram, the modal value is also 256. Hence by both the methods we get the same value of mode.

Mode can also be determined from frequency polygon in which case perpendicular is drawn on the base from the apex of the polygon and the point where it meets the base gives the modal value.

However, graphic method of determining mode can be used only where there is one class containing the highest frequency. If two or more classes have the same highest frequency, mode cannot be determined graphically.

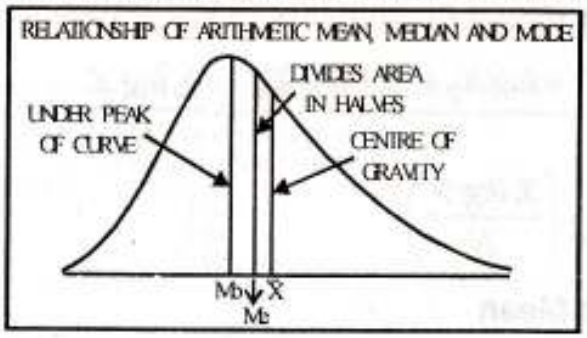
Merits and Limitations of Mode

Like median, the mode is not affected by extreme values and its value can be obtained in open-end distributions without ascertaining the class limits. Mode can be easily used to describe qualitative phenomenon. For example, when we want to compare the consumer preferences for different types of products, say, soap, toothpastes, etc., or different media of advertising, we should compare the modal preferences. In such distributions where there is an outstanding large frequency, mode happens to be meaningful as an average.

However, mode is not a rigidly defined measure as there are several formulae for calculating the mode, all of which usually give somewhat different answers. Also the value of mode cannot always be computed, such as, in case of bimodal distributions.

Relationship among Mean, Median and Mode

A distribution in which the values of mean, median and mode coincide is known as *symmetrical* distribution. Conversely stated, when the values of mean, median and mode are not equal, the distribution is known as *asymmetrical* or *skewed*. In moderately skewed or asymmetrical distributions, a very important relationship exists among mean, median and mode. In such distributions, the distance between the mean and the median is approximately one-thrid of the distance between the mean and mode as will be clear from the following diagram :



Karl Pearson has expressed this approximate relationship as follows :

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$$

or $$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}$$

and $$\text{Median} = \frac{2\ \text{Mean} + \text{Mode}}{3}$$

If we know any of the two values out of the three, we can compute the third from these relationships. The following example will illustrate this point :

Illustration 16. In a moderately asymmetrical distribution the Mode and Mean are 32.1 and 35.4 respectively. Calculate the Median.

Solution. Mode = 3 Median –2 Mean

Mode = 32.1, Mean = 35.4

Substituting the values

32.1 = 3 Median –2 (35.4) or 3 Med = 102.9 or Med. = 34.3.

2. GEOMETRIC MEAN

In business and economic problems, very often we are faced with questions pertaining to percentage rates of change over time. Neither the mean, the median nor mode is the appropriate average to use in these instances. For example, consider the following figures of sale of a company :

Year :	2007	2008	2009	2010
Sales (million tonnes) :	20.2	22.5	23.9	28.0

Suppose we want to find out the average percentage rate of change per year in sales. To answer this question we must specify what we mean by the 'Average percentage rate of change per year'. The most generally useful interpretation of this term is the constant percentage rate of change which if applied each year would take us from the first to the last figure. Hence in the above illustration we would be interested in that constant yearly percentage rate of change which would be required to move from 20.2 million tonnes of sales in 2007 to 28.0 million tonnes in 2010. None of the previously discussed averages provides the correct answer to this question. The correct answer can be obtained through the use of the geometric mean or, what amounts to the same thing, through the use of the familiar compound interest formula. In the discussion which follows, the geometric mean is defined, and the relationship between this average and compound interest calculations is indicated.

Geometric mean is defined as the Nth root of the product of N observations of a given data. If there are two observations, we take the square root; if there are three observations, the cube root; and so on, symbolically.

$$G.M. = \sqrt[N]{X_1 \times X_2 \times X_3 \times ..... \times X_N}$$

where $X_1, X_2, X_3 ..., X_N$, refer to the various observations of the data.

When the number of observations is three or more the task of multiplying the number and of extracting the root becomes quite difficult. To simplify calculations logarithms are used. Geometric mean is then calculated as follows :

$$\log \text{G.M.} = \frac{\log X_1 + \log X_2 + ... + \log X_N}{N} = \frac{\Sigma \log X}{N}$$

$$\therefore \quad \text{G.M.} = \text{antilog} \left(\frac{\Sigma \log N}{N} \right).$$

Calculation of Geometric Mean

In ungrouped data, geometric mean is calculated with the help of the following formula :

$$\text{G.M.} = \text{A.L.} \left(\frac{\Sigma \log X}{N} \right)$$

In grouped data, for calculating geometric mean first we will find the midpoints and then apply the following formula :

$$\text{G.M.} = \text{A.L.} \left(\frac{\Sigma f \log X}{N} \right)$$

where X = midpoint.

Compound Interest Formula

The compound interest formula is expressed as follows :

$$P_n = P_0 (1 + r)^n$$

where P_n = amount accumulated at the end of n periods,

P_0 = Original principal,

r = Rate of interest expressed as a decimal, and

n = Number of compound periods.

It follows from the above formula that :

$$r = \sqrt[n]{\frac{P_n}{P_0}} - 1$$

If interest is compounded at different rates in each time period, and if these successive rates are denoted by $r_1, r_2, \ldots, r_n$, then the amount accumulated at the end of n periods with an original principal of P_0 is

$$P_n = P_0 (1 + r_1)(1 + r_2) \ldots\ldots (1 + r_n).$$

Applications of Geometric Mean

Geometric mean is specially useful in the following cases :

1. The geometric mean is used to find the average per cent increase in sales, production, population or other economic or business data. For example, from 2008 to 2010 prices increased by 5%, 10% and 15% respectively. The average annual increase is not 11% as given by the arithmetic average but 10.9% as obtained by the geometric mean. This average is also useful in measuring the growth of population, because population increases in geometric progression.

2. Geometric mean is theoretically considered to be the best average in the construction of index number.* It makes index numbers satisfy the time reversal test and gives equal weights to equal ratio of change.

3. It is an average which is most suitable when large weights have to be given to small values of observations and small weights to large values of observations, situations which we usually come across in social and economic fields.

The following examples illustrate the use of geometric mean.

Illustration 17. Compared to the previous year the overhead expenses went up by 32% in 2008; they increased by 4% in the next year and by 50% in the following year. Calculate the average rate of increase in the overhead expenses over the three years.

Solution. In average ratios and percentages, geometric mean is more appropriate. Applying geometric mean here :

% Rise	*Expenses at the end of the year taking preceding year as 100*	*log X*
32	132	2.1206
40	140	2.1461
50	150	2.1761
		$\Sigma \log X = 6.4428$

$$\text{G.M.} = \text{A.L.}\left(\frac{\Sigma \log X}{N}\right) = \text{A.L.}\left(\frac{6.4428}{3}\right) = \text{A.L. } 2.1476 = 140.5.$$

Average rate of increase in overhead expenses

$= 140.5 - 100 = 40.5\%$.

Illustration 18. The annual rates of growth of output of a factory in 5 years are 5.0, 7.5, 2.5, 5.0 and 10.0 respectively. What is the compound rate of growth of output per annum for the period ?

Solution.

CALCULATING COMPOUND RATE OF GROWTH

Annual rate of growth	*Output relatives at the end of the year*	*log X*
5.0	105.0	2.0212
7.5	107.5	2.0314
2.5	102.5	2.0107
5.0	105.0	2.0212
10.0	110.0	2.0414
		$\Sigma \log X = 10.1259$

$$\text{G.M.} = \text{A.L.}\left(\frac{\Sigma \log X}{N}\right) = \text{A.L.}\left(\frac{10.1259}{5}\right) = \text{A.L. } 2.0252 = 105.9.$$

The compound rate of growth of output per annum for the period is $105.9 - 100 = 5.9\%$.

*Please refer to chapter on Index Numbers.

Illustration 19. A piece of property was purchased for Rs. 2,00,000 and sold 10 years later for Rs. 23,26,000. What is the average annual rate of return on the original investment ?

Solution. $2,00,000\ X^{10} = 3,26,000$

$$X^{10} = \frac{3,26,000}{2,00,000} = 1.63$$

$$\text{Log } X = \frac{\text{Log } 1.63}{10} = \frac{0.2122}{10} = 0.0212$$

X = A.L. (0.0212) = 1.05 or 105%.

Hence the investment yielded a mean rate return of 105 – 100 = 5 per cent over the 10-year period.

Combined Geometric Mean

Just as we have talked of combined arithmetic mean, in a similar manner we can also talk of combined geometric mean. If the geometric mean of N observations is 6 and these N observations are divided into two sets first containing N_1 and second containing N_2 observations having G_1 and G_2 as the respective geometric means, then

$$\text{Log } G = \frac{N_1 \text{ Log } G_1 + N_2 \text{ Log } G_2}{N_1 + N_2}$$

Thus if the geometric mean of 5 observations is 20 and of another 10 observations is 35.28, the combined geometric mean shall be

$$\text{Log } G = \frac{5 \text{ Log } 20 + 10 \text{ Log } 35.28}{5 + 10} = \frac{(5 \times 1.3010) + (10 \times 1.5475)}{15}$$

$$= \frac{6.505 + 15.475}{15} = \frac{21.98}{15} = 1.465$$

$$\therefore \qquad G = \text{A.L. } 1.465 = 29.17.$$

Illustration 20. Three groups of observations contain 8, 7, and 5 observations. Their geometric means are 8.52, 10.12 and 7.75 respectively. Find the geometric mean of the 20 observations in the single group formed by pooling the three groups.

Solution.

$$\text{Log } G = \frac{N_1 \text{ Log } G_1 + N_2 \text{ Log } G_2 + N_3 \text{ Log } G_3}{N_1 + N_2 + N_3}$$

$$= \frac{8 \text{ Log } 8.52 + 7 \text{ Log } 10.12 + 5 \text{ Log } 7.75}{8 + 7 + 5}$$

$$= \frac{(8 \times .9304) + (7 \times 1.0052) + (5 \times .8893)}{20}$$

$$= \frac{7.4432 + 7.0364 + 4.4465}{20} = \frac{18.9261}{20} = 0.9463$$

$$G = \text{A.L. } 0.9463 = 8.837.$$

Hence the combined geometric mean of the 20 observations taken together is 8.837.

Merits and Limitations of Geometric Mean

Geometric mean is highly useful in averaging ratios and percentages and in determining rates of increase and decrease. It is also capable of algebraic manipulation. For example, if the geometric mean of two or more series and their numbers of observations are known, a combined geometric mean can easily be calculated.

However, compared to arithmetic mean, this average is more difficult to compute and interpret. Also geometric mean cannot be computed when there are both negative and positive values in a series or more observations are having zero value.

E. HARMONIC MEAN

The harmonic mean is based on the reciprocal of the numbers averaged. It is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observation. Thus by definition

$$\text{H.M.} = \frac{N}{\left(\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+...+\frac{1}{X_N}\right)}$$

When the number of observations is large, the computation of harmonic mean in the above manner becomes tedious. To simplify calculations, we obtain reciprocals of the various observations and apply the following formulae :

For ungrouped data, $\text{H.M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)}$

For grouped data, $\text{H.M.} = \frac{N}{\Sigma\left(f \times \frac{1}{X}\right)}$ or $\frac{N}{\Sigma\left(\frac{f}{X}\right)}$*

Illustration 21. (*i*) Calculate harmonic mean of numbers 10, 20 25, 40, 50. (*ii*) Calculate harmonic mean from the following frequency distribution :

X :	0—10	10—20	20—30	30—40	40—50
f :	8	15	20	4	3

Solution. (*i*)

CALCULATION OF HARMONIC MEAN

X	$1/X$
10	0.100
20	0.050
25	0.040
40	0.025
50	0.020
	$\Sigma\ 1/X = 0.235$

$$\text{H.M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)} = \frac{5}{0.235} = 21.28$$

(*ii*)

CALCULATION OF HARMONIC MEAN

Variable	X	f	$f \times 1/X$
0–10	5	8	1.600
10–20	15	15	1.000
20–30	25	20	0.800
30–40	35	4	0.114
40–50	45	3	0.067
		$N = 50$	$\Sigma\left(f \times \frac{1}{X}\right) = 3.581$

$$\text{H.M.} = \frac{N}{\Sigma\left(f \times \frac{1}{X}\right)} = \frac{50}{3.581} = 13.96.$$

*There is no need to first calculate $1/X$ and then multiply it by f. We can directly obtain f/X to simplify calculation.

Measures of Deviation:

Mean Deviation: In statistics, deviation means the difference between the observed and expected values of a variable. In simple words, the deviation is the distance from the centre point. The centre point can be median, mean, or mode. Similarly, the mean deviation definition in statistics or the mean absolute deviation is used to compute how far the values fall from the middle of the data set.

Mean Deviation, also known as Mean Absolute Deviation, is the average Deviation of a Data point from the Data set's Mean, median, or Mode. The term "Mean Deviation" is abbreviated as MAD.

Mean Deviation Formula

	Ungrouped Data	Grouped Data
About Mean	$\frac{\sum_{i=1}^{n} \lvert x_i - \text{Mean} \rvert}{n}$	$\frac{\sum_{i=1}^{n} f_i \lvert x_i - \text{Mean} \rvert}{\sum_{i=1}^{n} f_i}$

Example 1: Calculate the Mean Deviation about the median using the Data given below: (Ungrouped Data)

Test Marks of 9 students are as follows: 86, 25, 87, 65, 58, 45, 12, 71, 35 respectively.

Solution 1) First we have to arrange them into ascending order, i.e., 12, 25, 35, 45, 58, 65, 71, 86, 87.

Then we have to find out the median so,

median = Value of the $\frac{(N+1)^{th}}{2}$ term

Value of the $\frac{(9+1)^{th}}{2}$ term = 58

X	|X-M|
12	46
25	33
35	23
45	13
58	0
65	7
71	13
86	28
87	29
N=9	$\sum \lvert X - M \rvert = 192$

MD=MD= $\frac{\sum \lvert X - M \rvert}{N}$

$= \frac{192}{9}$

$= 21.33$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Example 2:

Example : Find the mean deviation about the mean for {12, 20, 32, 16, 5}

Solution: The data is ungrouped, thus mean = (12 + 20 + 32 + 16 + 5) / 5 = 17

x	$\lvert x - \bar{x} \rvert$
12	5
20	3
32	15
16	1
5	12
Total	36

Using the formula, $\frac{\sum_1^5 \lvert x_i - \mu \rvert}{5}$ = 36 / 5 = 7.2

Answer: Mean deviation about mean = 7.2

Example 3: Find the mean deviation for the following data set. (Grouped Data)

Class Interval	Classmark/midpoint (x_i)	Frequeny (f_i)	$x_i f_i$
0–2	1	4	4
2–4	3	3	9
4–6	5	5	25
6–8	7	7	49
		$\sum f_i = 19$	$\sum x_i f_i = 87$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

(Mean) $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = 4.57$ when $N = \sum_{i=1}^{n} f_i$

The mean deviation $(x_d) = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - \bar{x}| = \frac{38.15}{19} = 2.007$

Example 4: Find the mean deviation about the median for the following data.

Solution:

Class	f	cf	x_i	$\|x - \bar{x}\|$	$f_i\|x - \bar{x}\|$
15 - 25	12	12	20	12.5	150
25 - 35	6	18	30	2.5	15
35 - 45	9	27	40	7.5	67.5
45 - 55	4	31	50	17.5	70
55 - 65	2	33	60	27.5	55
Total	N = 33				357.5

N/2 = 33/2 = 16.5

The cf value that is nearest to 16.5 but higher than it is 18.

Thus, median class is 25 - 35.

l = 25, h = 10, f = 6, cf = 12, $\frac{\Sigma_1^5 f_i}{2} = 16.5$

Substituting these values in the formula,

$M = l + \frac{\frac{\Sigma_1^5 f_i}{2} - cf}{f} \times h = 32.5$

Mean deviation about median = $\frac{\Sigma_1^5 f_i|x_i - M|}{\Sigma_1^5 f_i}$ = 357.5 / 33 = 10.83

Answer: Mean deviation about median = 10.83

***The coefficient of mean deviation is calculated by dividing mean deviation by the average.**

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Standard Deviation and Variance

Standard deviation is the positive square root of the variance. Standard deviation is the degree of dispersion or the scatter of the data points relative to its mean, in descriptive statistics. It tells how the values are spread across the data sample and it is the measure of the variation of the data points from the mean. Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. The basic difference between variance and the standard deviation is in their units. The standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}}$

Population Variance - All the members of a group are known as the population. When we want to find how each data point in a given population varies or is spread out then we use the population variance. It is used to give the squared distance of each data point from the population mean.

Sample Variance - If the size of the population is too large then it is difficult to take each data point into consideration. In such a case, a select number of data points are picked up from the population to form the sample that can describe the entire group. Thus, the sample variance can be defined as the average of the squared distances from the mean. The variance is always calculated with respect to the sample mean.

Standard Deviation of Ungrouped Data

Example: If a die is rolled, then find the variance and standard deviation of the possibilities.

Solution: When a die is rolled, the possible number of outcomes is 6. So the sample space, n = 6 and the data set = { 1;2;3;4;5;6}.

To find the variance, first, we need to calculate the mean of the data set.

Mean, $\bar{x} = (1+2+3+4+5+6)/6 = 3.5$

We can put the value of data and mean in the formula to get;

$$\sigma^2 = \Sigma (xi - \bar{x})^2/n$$

$$\sigma^2 = ⅙ (6.25+2.25+0.25+0.25+2.25+6.25)$$

$$\sigma^2 = 2.917$$

Answer: Therefore the variance is $\sigma^2 = 2.917$, and standard deviation, $\sigma = \sqrt{2.917} = 1.708$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Example: There are 39 plants in the garden. A few plants were selected randomly and their heights in cm were recorded as follows: 51, 38, 79, 46, and 57. Calculate the standard deviation of their heights. N=5.

$$\text{Mean } (\bar{x}) = (51+38+79+46+57)/5 = 54.2$$

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma(x_i - x)^2}{N-1}}$$

$$= \sqrt{\frac{(51-54.2)^2+(38-54.2)^2+(79-54.2)^2+(46-54.2)^2+(57-54.2)^2}{4}}$$

$$= 15.5$$

Standard Deviation of Grouped Data

Class (1)	Frequency (f) (2)	Mid value (x) (3)	$f \cdot x$ (4) = (2) × (3)	$f \cdot x^2 = (f \cdot x) \times (x)$ (5) = (4) × (3)
2-4	3	3	9	27
4-6	4	5	20	100
6-8	2	7	14	98
8-10	1	9	9	81
---	---	---	---	---
--	$n = 10$	--	$\sum f \cdot x = 52$	$\sum f \cdot x^2 = 306$

$$\text{Population Variance } \sigma^2 = \frac{\sum f \cdot x^2 - \frac{(\sum f \cdot x)^2}{n}}{n}$$

$$= \frac{306 - \frac{(52)^2}{10}}{10} = 3.56$$

$$\text{Sample Variance } S^2 = \frac{\sum f \cdot x^2 - \frac{(\sum f \cdot x)^2}{n}}{n-1}$$

$$= \frac{306 - \frac{(52)^2}{10}}{9} = 3.9556$$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Coefficient of Variation

Coefficient of variation is a type of relative measure of dispersion. It is expressed as the ratio of the standard deviation to the mean. A measure of dispersion is a quantity that is used to gauge the extent of variability of data.

	Coefficient of Variation	Standard Deviation
Population	$\frac{\sigma}{\mu} \times 100$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Sample	$\frac{s}{\mu} \times 100$	$s = \sqrt{\frac{\sum(x_i - \mu)^2}{N - 1}}$

Example 1: Find the population coefficient of variation of the given data set (320, 540, 480, 540, 420, 240)

Solution: Mean = (320 + 540 + 480 + 540 + 420 + 240) / 6 = 423.33

Standard Deviation =

$$\sqrt{\frac{(320-423.33)^2+(540-423.33)^2+(480-423.33)^2+(540-423.33)^2+(420-423.33)^2+(240-423.33)^2}{6}}$$

Standard Deviation = 111.6

Coefficient of Variation = (Standard Deviation / mean) * 100 = (111.6 / 423.33) * 100 = 26.36%

Answer: Coefficient of variation = 26.36%

Example 2: If the coefficient of variation is given as 20.75 and the mean is 22.6 then find the standard deviation.

Solution: Coefficient of Variation = (Standard Deviation / mean) * 100

20.75 = (SD / 22.6) * 100

SD = 4.69

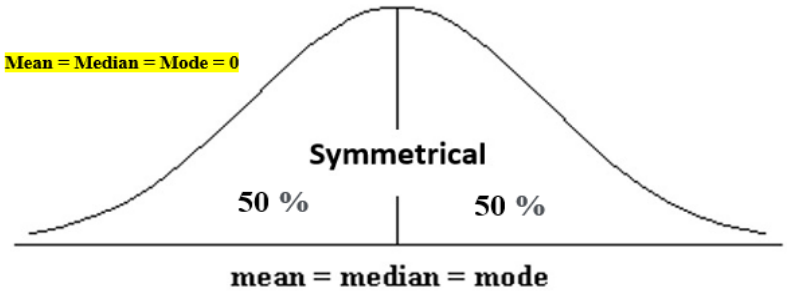
Answer: Standard deviation = 4.69

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Shape of data: Skewness and Kurtosis

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half. For example, the normal distribution is a symmetric distribution with no skew. The tails are exactly the same. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

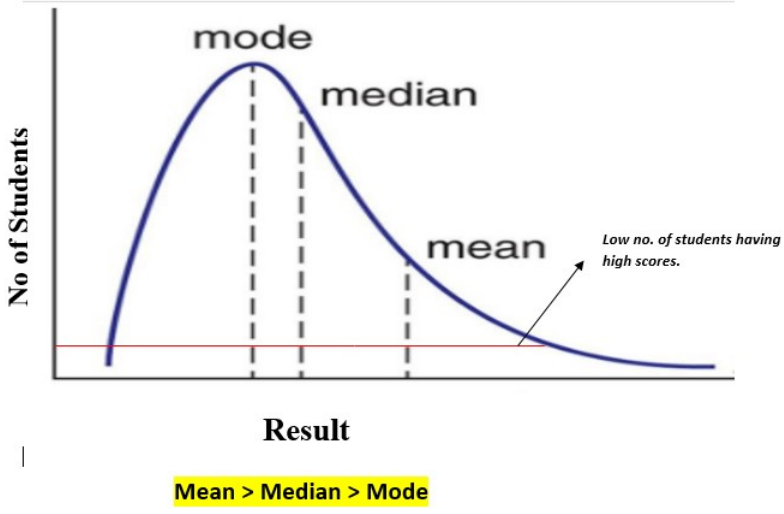


When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.

Types of skewness

1. Positive skewed or right-skewed

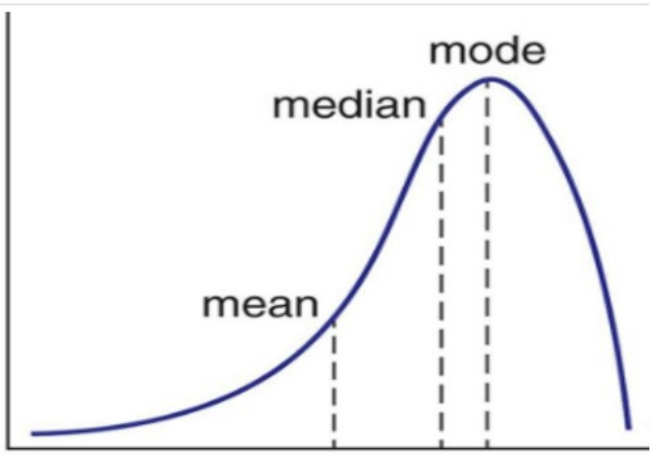
In statistics, a positively skewed distribution is a sort of distribution where, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other, with positively skewed data, the measures are dispersing, which means Positively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

2. Negative skewed or left-skewed

A negatively skewed distribution is the straight reverse of a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side. In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



Mode > Median > mean

Calculate the skewness coefficient of the sample

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of -1 to +1. That accurately the range of the correlation values.

Pearson's first coefficient of skewness is helping if the data present high mode. But, if the data have low mode or various modes, Pearson's first coefficient is not preferred, and Pearson's second coefficient may be superior, as it does not rely on the mode.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.

If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

Solution. CALCULATION OF COEFFICIENT OF SKEWNESS

Profits (Rs. lakhs)	*m.p. X*	*f*	*(X–170)/20 d*	*fd*	*fd²*
100–120	110	17	–3	–51	153
120–140	130	53	–2	–106	212
140–160	150	199	–1	–199	199
160–180	170	194	0	0	0
180–200	190	327	+1	+327	327
200–220	210	208	+2	+416	832
220–240	230	2	+3	+6	18
		$N = 1{,}000$		$\Sigma fd = 393$	$\Sigma fd^2 = 1{,}741$

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Calculation of Mean : $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 170 + \frac{393}{1000} \times 20 = 170 + 7.86 = 177.86$

Calculation of Mode : By inspection mode lies in the class 180–200.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 180 + \frac{133}{133+119} \times 20 = 180 + 10.56 = 190.56$$

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1741}{1000} - \left(\frac{393}{1000}\right)^2} \times 20$$

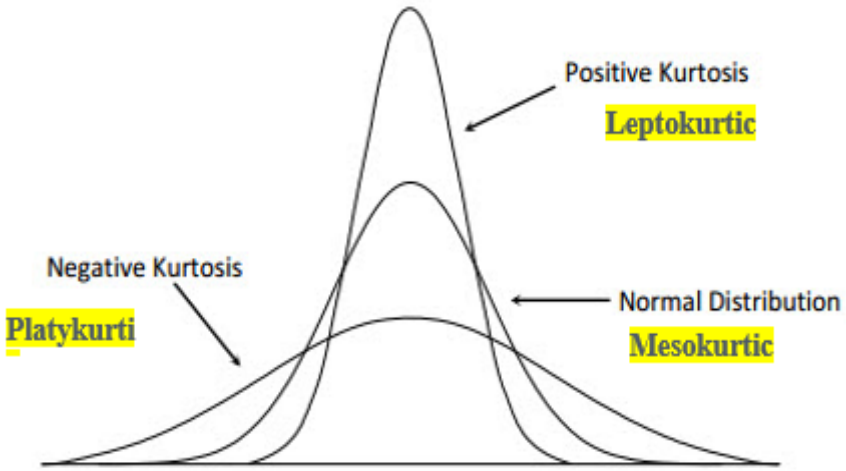
$$= \sqrt{1.74 - 0.15} \times 20 = 1.26 \times 20 = 25.2$$

$$Sk_P = \frac{177.86 - 190.56}{25.2} = -0.504$$

The mode is greater than the mean by an amount equal to about 50.4 per cent of the value of standard deviation. It is a case of moderate negatively skewed distribution.

Kurtosis

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

Excess kurtosis = Kurt – 3

Types of excess kurtosis

1. Leptokurtic (kurtosis > 3)

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

2. platykurtic (kurtosis < 3)

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

3. Mesokurtic (kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near to 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution).

Leptokurtic distribution (kurtosis more than normal distribution).

Mesokurtic distribution (kurtosis same as the normal distribution).

Platykurtic distribution (kurtosis less than normal distribution).

Calculate Population Skewness and Population Kurtosis from the following grouped data

Example-1

Class (1)	Mid value (x) (2)	f (3)	$f \cdot x$ (4) = (2) × (3)	$(x - \bar{x})$ (5)	$f \cdot (x - \bar{x})^2$ (6) = (3) × (5)	$f \cdot (x - \bar{x})^3$ (7) = (5) × (6)	$f \cdot (x - \bar{x})^4$ (8) = (5) × (7)
2 - 4	3	3	9	-2.2	14.52	-31.944	70.2768
4 - 6	5	4	20	-0.2	0.16	-0.032	0.0064
6 - 8	7	2	14	1.8	6.48	11.664	20.9952
8 - 10	9	1	9	3.8	14.44	54.872	208.5136
---	---	---	---	---	---	---	---
--	--	$n = 10$	$\sum f \cdot x = 52$	`--	= 35.6	= 34.56	= 299.792

Population Standard deviation $\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$

$= \sqrt{\frac{35.6}{10}} = 1.8868$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

$$\text{Population Skewness} = \frac{\sum f(x-\bar{x})^3}{n \cdot S^3}$$

$$= \frac{34.56}{10 \cdot (1.8868)^3}$$

$$= \frac{34.56}{10 \cdot 6.717}$$

$$= 0.5145$$

$$\text{Population Kurtosis} = \frac{\sum f(x-\bar{x})^4}{n \cdot S^4}$$

$$= \frac{299.792}{10 \cdot (1.8868)^4}$$

$$= \frac{299.792}{10 \cdot 12.6736}$$

$$= 2.3655$$

Karl Pearson Coefficient of Correlation

The study of Karl Pearson Coefficient is an inevitable part of Statistics. Statistics is majorly dependent on Karl Pearson Coefficient Correlation method. The Karl Pearson coefficient is defined as a linear correlation that falls in the numeric range of -1 to +1.

This is a quantitative method that offers the numeric value to form the intensity of the linear relationship between the X and Y variable. But is it really useful for any economic calculation? Let, us find and delve into this topic to get more detailed information on the subject matter – Karl Pearson Coefficient of Correlation.

What do You mean by Correlation Coefficient?

Before delving into details about Karl Pearson Coefficient of Correlation, it is vital to brush up on fundamental concepts about correlation and its coefficient in general.

The correlation coefficient can be defined as a measure of the relationship between two quantitative or qualitative variables, i.e., X and Y. It serves as a statistical tool that helps to analyze and in turn, measure the degree of the linear relationship between the variables.

For example, a change in the monthly income (X) of a person leads to a change in their monthly expenditure (Y). With the help of correlation, you can measure the degree up to which such a change can impact the other variables.

Types of Correlation Coefficient

Depending on the direction of the relationship between variables, correlation can be of three types, namely –

Positive Correlation (0 to +1)

In this case, the direction of change between X and Y is the same. For instance, an increase in the duration of a workout leads to an increase in the number of calories one burns.

Negative Correlation (0 to -1)

Here, the direction of change between X and Y variables is opposite. For example, when the price of a commodity increases its demand decreases.

Zero Correlation (0)

There is no relationship between the variables in this case. For instance, an increase in height has no impact on one's intelligence.

What is Karl Pearson's Coefficient of Correlation?

This method is also known as the Product Moment Correlation Coefficient and was developed by Karl Pearson. It is one of the three most potent and extensively used methods to measure the level of correlation, besides the Scatter Diagram and Spearman's Rank Correlation.

The Karl Pearson correlation coefficient method is quantitative and offers numerical value to establish the intensity of the linear relationship between X and Y. Such a coefficient correlation is represented as 'r'.

The Karl Pearson Coefficient of Correlation formula is expressed as:

Actual Mean Method Which is Expressed as

Actual Mean Method Which is Expressed as -

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sqrt{(Y - \bar{Y})^2}}}$$

Where, $\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

The Coefficient of Correlation can also be Calculated as (without taking deviation from mean)

$$= \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Assumed Mean Method Which is Expressed as

Assume Mean Method

d_x = X - A

d_y = Y - A

$$r = \frac{N\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{N\sum d_x^2 - (\sum d_x)^2} - \sqrt{N\sum d_y^2 - (\sum d_y)^2}}$$

In this Karl Pearson Correlation formula,

- dx = x-series' deviation from assumed mean, wherein (X - A)
- dy = Y-series' deviation from assumed mean = (Y - A)
- Σdx.dy implies summation of multiple dx and dy.

Illustration Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value :

Roll No.	:	1	2	3	4	5
Marks in Accountancy	:	48	35	17	23	47
Marks in Statistics	:	45	20	40	25	45

Direct Method/ Actual Mean Method

Solution. Let marks in accountancy be denoted by X and that in statistics by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

X	$(X-34)$ x	x^2	Y	$(Y-35)$ y	y^2	xy
48	+14	196	45	+10	100	+140
35	+1	1	20	−15	225	−15
17	−17	289	40	+5	25	−85
23	−11	121	25	−10	100	+110
47	+13	169	45	+10	100	+130
$\Sigma X = 170$	$\Sigma x = 0$	$\Sigma x^2 = 776$	$\Sigma Y = 175$	$\Sigma y = 0$	$\Sigma y^2 = 550$	$\Sigma xy = 280$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{280}{\sqrt{776 \times 550}} = \frac{280}{653.3} = +0.429$$

It is a moderate case of positive correlation between marks in accountancy and statistics.

Marks obtained by 5 students in algebra and trigonometry as given below: Calculate Karl Pearson Coefficient of Correlation without taking deviation from mean.

x	y	x^2	y^2	xy
16	11	256	121	176
15	18	225	324	270
12	10	144	100	120
10	20	100	400	200
8	17	64	289	136
$\sum x = 61$	$\sum y = 76$	$\sum x^2 = 789$	$\sum y^2 = 1234$	$\sum xy = 902$

r= -0.424

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Solution. Let age and sick days be represented by variable X and Y respectively. Assumed Mean Method

CALCULATION OF CORRELATION COEFFICIENT

Age X	$(X-43)$ d_x	d_x^2	Sick days Y	$(Y-14)$ d_y	d_y^2	$d_x d_y$
20	−23	529	11	−3	9	+69
30	−13	169	12	−2	4	+26
32	−11	121	10	−4	16	+44
35	−8	64	13	−1	1	+8
40	−3	9	14	0	0	0
46	+3	9	16	+2	4	+6
52	+9	81	15	+1	1	+9
55	+12	144	17	+3	9	+36
58	+15	225	18	+4	16	+60
62	+19	361	19	+5	25	+95
$\Sigma X = 430$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 1712$	$\Sigma Y = 145$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 85$	$\Sigma d_x d_y = 353$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10\times 353 - (0)(-5)}{\sqrt{10\times 1712 - (0)^2}\sqrt{10\times 85 - (-5)^2}}$$

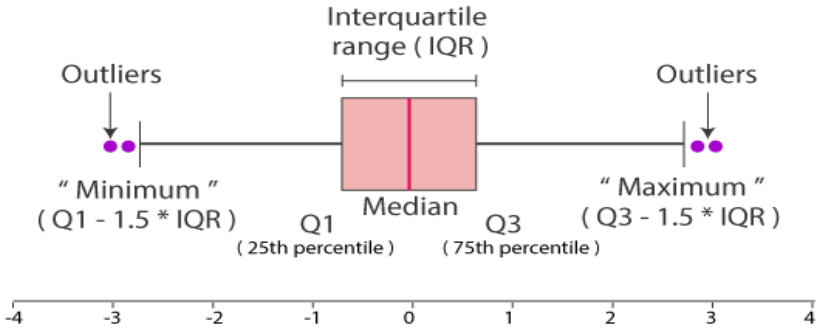
$$= \frac{3530}{\sqrt{17120}\sqrt{825}} = \frac{3530}{130.85\times 28.72} = 0.939$$

Thus, there is a very high degree of positive correlation between age and sick days taken. Hence, we can conclude that the age of an employee increases, he is liable to be sick more often than others.

Box Plot

When we display the data distribution in a standardized way using 5 summary – minimum, Q1 (First Quartile), median, Q3(third Quartile), and maximum, it is called a Box plot. It is also termed as box and whisker plot. It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

Parts of Box Plots



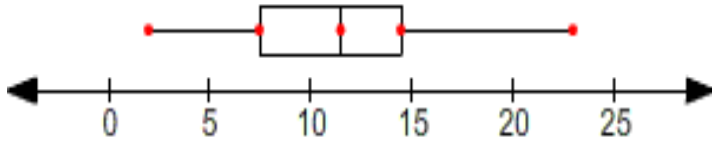
Different parts of boxplot

Minimum: The minimum value in the given dataset

First Quartile (Q1): The first quartile is the median of the lower half of the data set.

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Median: The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.
Third Quartile (Q3): The third quartile is the median of the upper half of the data.
Maximum: The maximum value in the given dataset.
Apart from these five terms, the other terms used in the box plot are:
Interquartile Range (IQR): The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1
Outlier: The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.
(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).
Positively Skewed: If the distance from the median to the maximum is greater than the distance from the median to the minimum, then the box plot is positively skewed.
Negatively Skewed: If the distance from the median to minimum is greater than the distance from the median to the maximum, then the box plot is negatively skewed.
Symmetric: The box plot is said to be symmetric if the median is equidistant from the maximum and minimum values.
The median (Q2) divides the data set into two parts, the upper set and the lower set. The **lower quartile** (Q1) is the median of the lower half, and the **upper quartile** (Q3) is the median of the upper half.
Example -1:
Find Q1 , Q2 , and Q3 for the following data set, and draw a box-and-whisker plot.
{2,6,7,8,8,11,12,13,14,15,22,23}
There are 12 data points. The middle two are 11 and 12. So the median, Q2 , is 11.5 .
The "lower half" of the data set is the set {2,6,7,8,8,11}. The median here is 7.5 So Q1=7.5.
The "upper half" of the data set is the set {12,13,14,15,22,23}. The median here is 14.5. So Q3=14.5.
A box-and-whisker plot displays the values Q1 , Q2 , and Q3 , along with the extreme values of the data set (2 and 23 , in this case):



A box & whisker plot shows a "box" with left edge at Q1 , right edge at Q3 , the "middle" of the box at Q2 (the median) and the maximum and minimum as "whiskers".
Example -2: Let us take a sample data to understand how to create a box plot.
Here are the runs scored by a cricket team in a league of 12 matches.

```
Ascending Order -
100,110,110,110,120,120,130,140,140,150,170,220


Median (Q2) = (120+130)/2 = 125   ; Since there were even values
```

To find the First Quartile we take the first six values and find their median.

```
Q1 = (110+110)/2 = 110
```

For the Third Quartile, we take the next six and find their median.

```
Q3 = (140+150)/2 = 145
```

Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.

IQR = Q3-Q1 = 145-110 = **35**

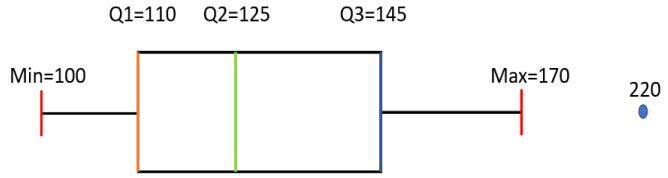
We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

Lower Limit = Q1-1.5*IQR = 110-1.5*35 = **57.5**
Upper Limit = Q3+1.5*IQR = 145+1.5*35 = **197.5**

So the minimum and maximum between the range [57.5,197.5] for our given data are –

Minimum = 100
Maximum = 170

The outliers which are outside the range are **Outliers= 220**



Example-3: Calculate Box and Whisker Plots from the following grouped data

Solution: n=25

x	Frequency	CF
0	1	1
1	5	6
2	10	16
3	6	22
4	3	25

Minimum value = 0

Maximum value = 4

First quartile Q_1 :

Here, $n = 25$

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{ value of the observation}$$

$$= \left(\frac{26}{4}\right)^{th} \text{ value of the observation}$$

$$= (6.5)^{th} \text{ value of the observation}$$

$$= 2$$

Median Q_2 :

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} \text{ value of the observation}$$

$$= \left(\frac{2 \cdot 26}{4}\right)^{th} \text{ value of the observation}$$

$$= (13)^{th} \text{ value of the observation}$$

$$= 2$$

Third quartile Q_3 :

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \text{ value of the observation}$$

$$= \left(\frac{3 \cdot 26}{4}\right)^{th} \text{ value of the observation}$$

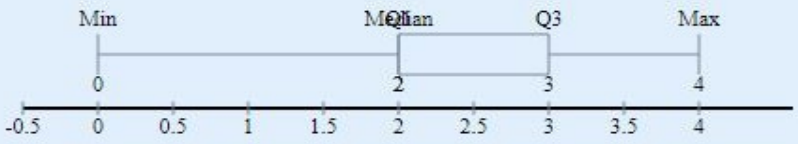
$$= (19.5)^{th} \text{ value of the observation}$$

$$= 3$$

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Thus Five number summary is

1. Minimum value = 0
2. First quartile $Q_1 = 2$
3. Median $Q_2 = 2$
4. Third quartile $Q_3 = 3$
5. Maximum value = 4



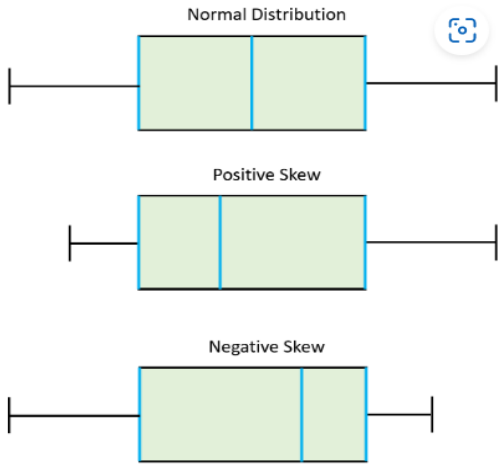
Uses of a Box Plot

Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).

a) If the Median is at the **center** of the Box and the **whiskers** are almost **same on both the ends** then the data is **Normally Distributed.**

b) If the Median lies **closer to the First Quartile** and if the **whisker at end is shorter** (as in the above example) then it has a **Positive Skew (**

c) If the Median lies **closer to the Third Quartile** and if the **whisker at upper end is shorter** then it has a **Negative Skew (Left Skew).**



Pivot Tables:

A pivot table is a powerful data summarization tool that can automatically sort, count, and sum up data stored in tables and display the summarized data. Pivot tables are useful to quickly create crosstabs (a process or function that combines and/or summarizes data from one or more sources into a concise format for analysis or reporting) to display the joint distribution of two or more variables.

Typically, with a pivot table the user sets up and changes the data summary's structure by dragging and dropping fields graphically. This "rotation" or pivoting of the summary table gives the concept its name.

Three key reasons for organizing data into a pivot table are:

- ✓ To summarize the data contained in a lengthy list into a compact format.
- ✓ To find relationships within the data those are otherwise hard to see because of the amount of detail.
- ✓ To organize the data into a format that's easy to read.

HeatMap

Heatmaps visualize the data in a 2-dimensional format in the form of colored maps. The color maps use hue, saturation, or luminance to achieve color variation to display various details. This color variation gives visual cues to the readers about the magnitude of numeric values. HeatMaps is about replacing numbers with colors because the human brain understands visuals better than numbers, text, or any written data.

A heatmap (or heat map) is a graphical representation of numerical data, where individual data points contained in the data set are represented using different colors. The key benefit of heatmaps is that they simplify complex numerical data into visualizations that can be understood at a glance. For example, on website heatmaps 'hot' colours depict high user engagement, while 'cold' colours depict low engagement.

Uses of HeatMap

1. Business Analytics: A heat map is used as a visual business analytics tool. A heat map gives quick visual cues about the current results, performance, and scope for improvements. Heatmaps can analyze the existing data and find areas of intensity that might reflect where most customers reside, areas of risk of market saturation, or cold sites and sites that need a boost.

2. Website: Heatmaps are used in websites to visualize data of visitors' behavior. This visualization helps business owners and marketers to identify the best & worst-performing sections of a webpage.

3. Exploratory Data Analysis: EDA is a task performed by data scientists to get familiar with the data. All the initial studies are done to understand the data are known as EDA. EDA is done to summarize their main features, often with visual methods, which includes Heatmaps.

4. Molecular Biology: Heat maps are used to study disparity and similarity patterns in DNA, RNA, etc.

5. Marketing and Sales: The heatmap's capability to detect warm and cold spots is used to improve marketing response rates by targeted marketing. Heatmaps allow the detection of areas that respond to campaigns, under-served markets, customer residence, and high sale trends, which helps optimize product lineups, capitalize on sales, create targeted customer segments, and assess regional demographics.

Types of HeatMaps

Typically, there are two types of Heatmaps:

Grid Heatmap: The magnitudes of values shown through colors are laid out into a matrix of rows and columns, mostly by a density-based function. Below are the types of Grid Heatmaps.

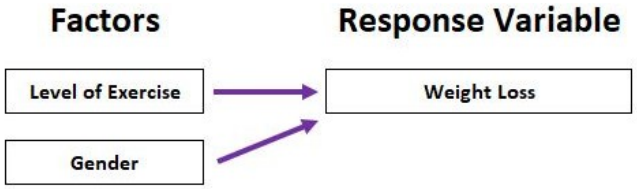
Clustered Heatmap: The goal of Clustered Heatmap is to build associations between both the data points and their features. This type of heatmap implements clustering as part of the process of grouping similar features. Clustered Heatmaps are widely used in biological sciences for studying gene similarities across individuals. **Correlogram:** A correlogram replaces each of the variables on the two axes with numeric variables in the dataset. Each square depicts the relationship between the two intersecting variables, which helps to build descriptive or predictive statistical models.

Spatial Heatmap: Each square in a Heatmap is assigned a color representation according to the nearby cells' value. The location of color is according to the magnitude of the value in that particular space. These Heatmaps are data-driven "paint by numbers" canvas overlaid on top of an image. The cells with higher values than other cells are given a hot color, while cells with lower values are assigned a cold color.

ANOVA, which stands for Analysis of Variance, is a statistical test used to analyze the difference between the means of more than two groups. A one-way ANOVA uses one independent variable, while a two-way ANOVA uses two independent variables.

A One-Way ANOVA is used to determine how one factor impacts a response variable. For example, we might want to know if three different studying techniques lead to different mean exam scores. To see if there is a statistically significant difference in mean exam scores, we can conduct a one-way ANOVA.

A **Two-Way ANOVA** is used to determine how two factors impact a response variable, and to determine whether or not there is an interaction between the two factors on the response variable. For example, we might want to know how gender and how different levels of exercise impact average weight loss. We would conduct a two-way ANOVA to find out.



ANOVA Real Life Example #1

A large scale farm is interested in understanding which of three different fertilizers leads to the highest crop yield. They sprinkle each fertilizer on ten different fields and measure the total yield at the end of the growing season. To understand whether there is a statistically significant difference in the mean yield that results from these three fertilizers, researchers can conduct a one-way ANOVA, using "type of fertilizer" as the factor and "crop yield" as the response.

ANOVA Real Life Example #2

An example to understand this can be prescribing medicines. Suppose, there is a group of patients who are suffering from fever. They are being given three different medicines that have the same functionality i.e. to cure fever. To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.

ANOVA is used in a wide variety of real-life situations, but the most common include:

Retail: Store are often interested in understanding whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales. This is the exact type of analysis that ANOVA is built for.

Medical: Researchers are often interested in whether or not different medications affect patients differently, which is why they often use one-way or two-way ANOVA's in these situations.

Environmental Sciences: Researchers are often interested in understanding how different levels of factors affect plants and wildlife. Because of the nature of these types of analyses, ANOVA's are often used.

What is ANOVA Test

ANOVA test, in its simplest form, is used to check whether the means of three or more populations are equal or not. The ANOVA test applies when there are more than two independent groups. The goal of the ANOVA test is to check for variability within the groups as well as the variability among the groups. The ANOVA test statistic is given by the f test.

ANOVA Test Definition

ANOVA test can be defined as a type of test used in hypothesis testing to compare whether the means of two or more groups are equal or not. This test is used to check if the null hypothesis can be rejected or not

depending upon the statistical significance exhibited by the parameters. The decision is made by comparing the ANOVA test statistic with the critical value.

The steps to perform the one way ANOVA test are given below:

Step 1: Calculate the mean for each group.

Step 2: Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.

Step 3: Calculate the SSB/ SSC.

Step 4: Calculate the between groups degrees of freedom.

Step 5: Calculate the SSE.

Step 6: Calculate the degrees of freedom of errors.

Step 7: Determine the MSB and the MSE.

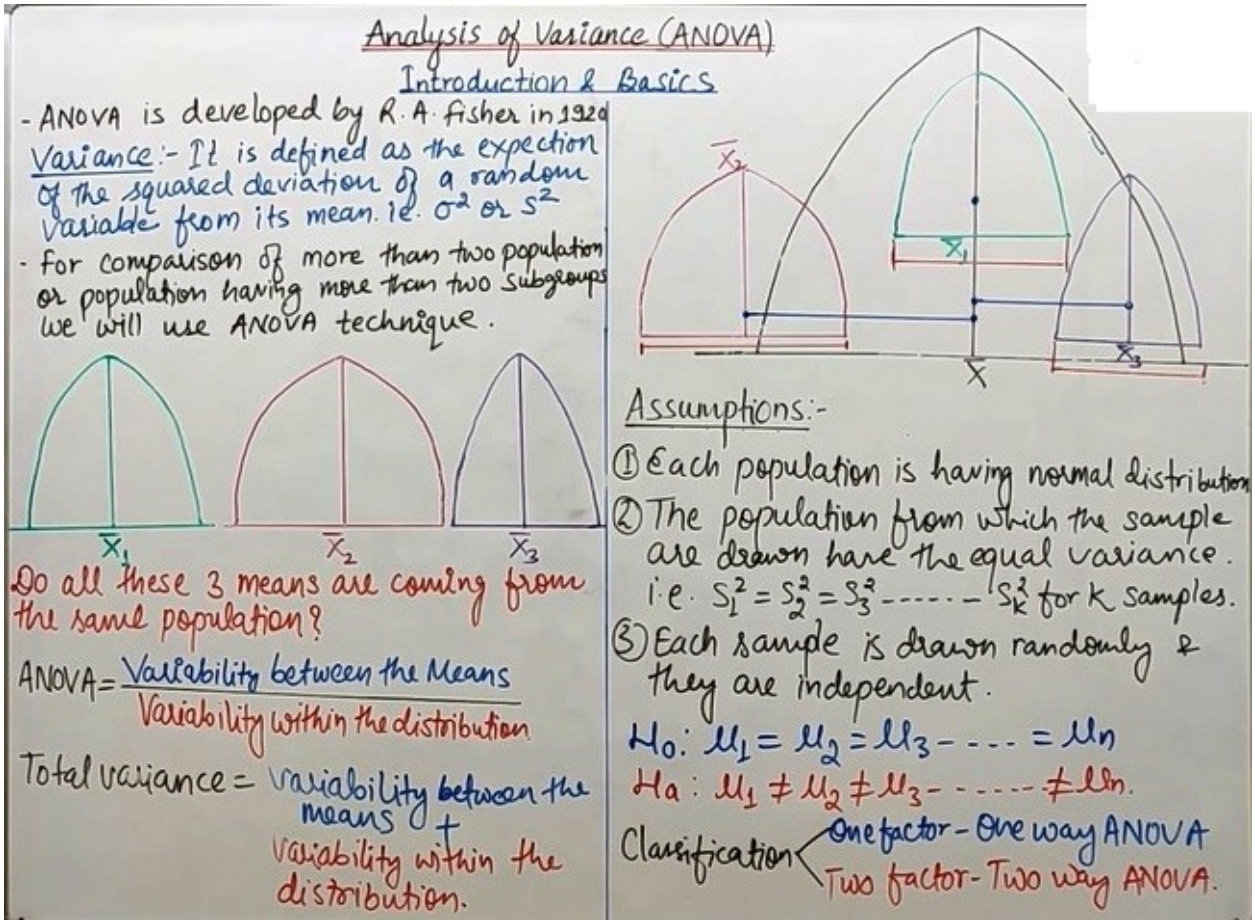
Step 8: Find the f test statistic.

Step 9: Using the f table for the specified level of significance, find the critical value. This is given by F(, df1. df2).

Step 10: If f > F then rejects the null hypothesis.

Assumptions for ANOVA

- ✓ Each group sample is drawn from a normally distributed population
- ✓ All populations have a common variance
- ✓ All samples are drawn independently of each other
- ✓ Within each sample, the observations are sampled randomly and independently of each other



Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

ANOVA also uses a Null hypothesis and an Alternate hypothesis. The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don't have any significant difference. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:

$$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_L \qquad \textit{Null hypothesis}$$

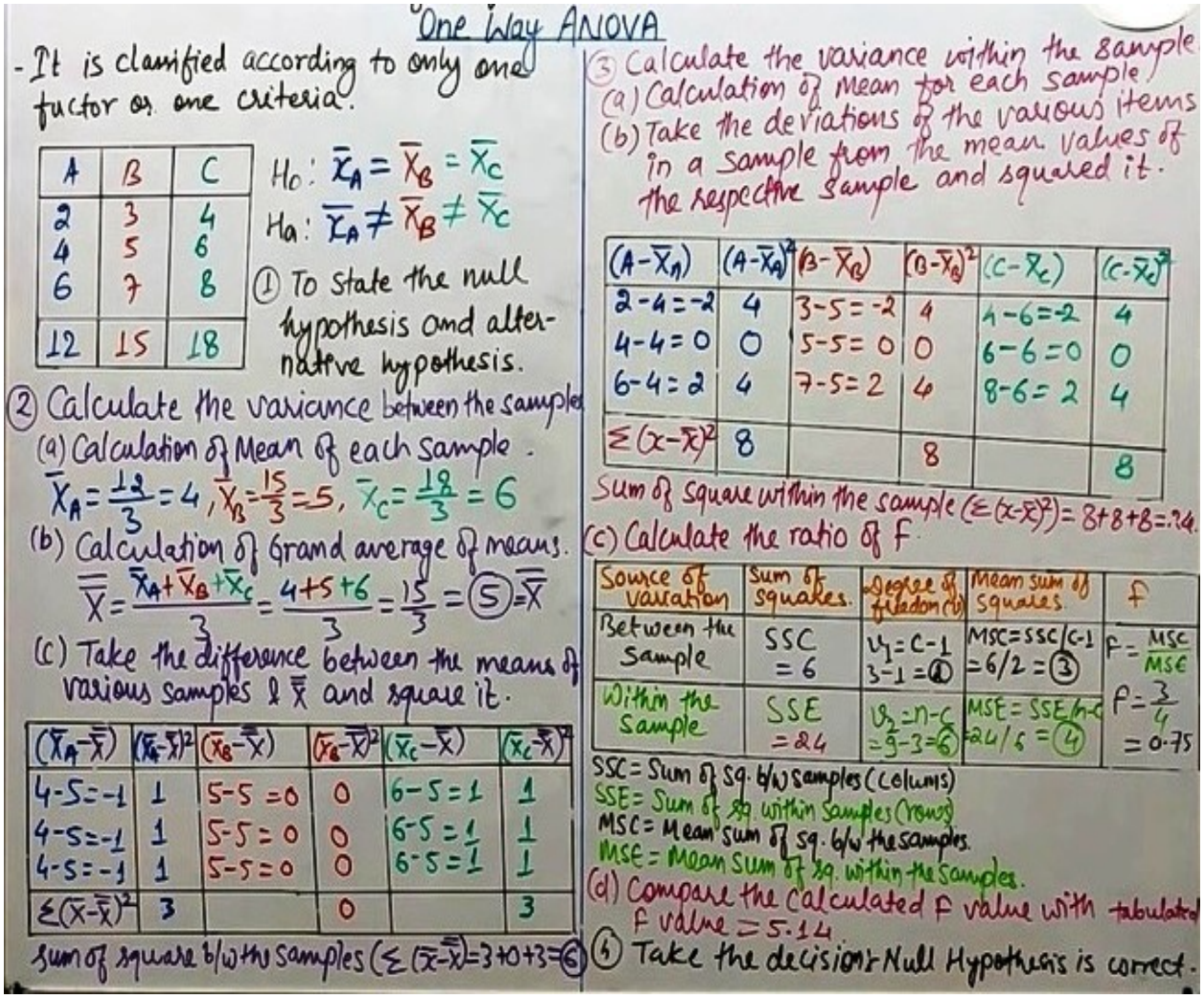
$$H_1: \ \mu_l \neq \mu_m \qquad \textit{Alternate hypothesis}$$

Sums of Squares: In statistics, the sum of squares is defined as a statistical technique that is used in regression analysis to determine the dispersion of data points. In the ANOVA test, it is used while computing the value of F. As the sum of squares tells you about the deviation from the mean, it is also known as variation.

Degrees of Freedom: Degrees of Freedom refer to the maximum numbers of logically independent values that have the freedom to vary in a data set.

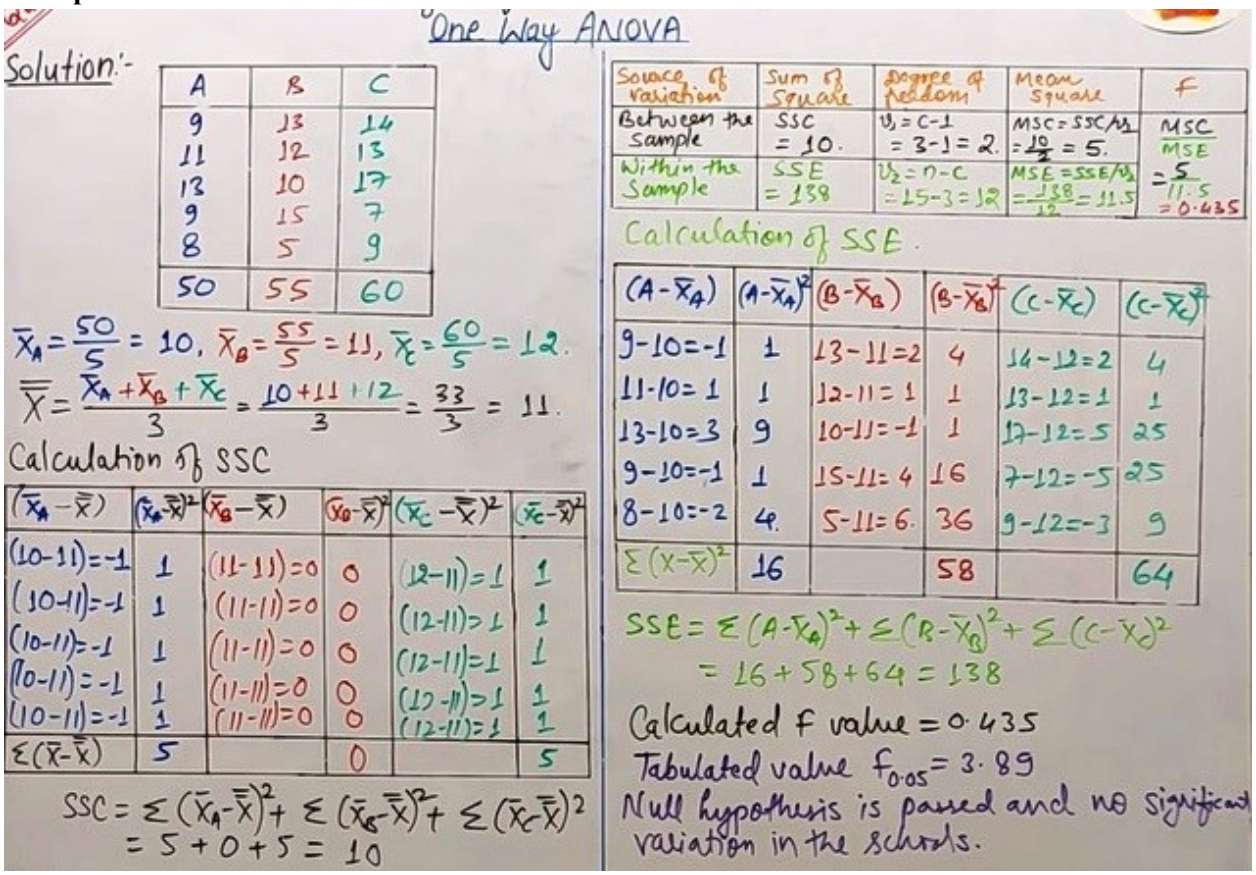
The Mean Squared Error: It tells us about the average error in a data set. To find the mean squared error, we just divide the sum of squares by the degrees of freedom.

Example-1



Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Example-2:



One Way ANOVA

Solution:-

A	B	C
9	13	14
11	12	13
13	10	17
9	15	7
8	5	9
50	55	60

$\bar{X}_A = \frac{50}{5} = 10$, $\bar{X}_B = \frac{55}{5} = 11$, $\bar{X}_C = \frac{60}{5} = 12$.

$\bar{\bar{X}} = \frac{\bar{X}_A + \bar{X}_B + \bar{X}_C}{3} = \frac{10+11+12}{3} = \frac{33}{3} = 11$.

Calculation of SSC

$(\bar{X}_A - \bar{\bar{X}})$	$(\bar{X}_A - \bar{\bar{X}})^2$	$(\bar{X}_B - \bar{\bar{X}})$	$(\bar{X}_B - \bar{\bar{X}})^2$	$(\bar{X}_C - \bar{\bar{X}})^2$	$(\bar{X}_C - \bar{\bar{X}})^2$
(10-11)=-1	1	(11-11)=0	0	(12-11)=1	1
(10-11)=-1	1	(11-11)=0	0	(12-11)=1	1
(10-11)=-1	1	(11-11)=0	0	(12-11)=1	1
(10-11)=-1	1	(11-11)=0	0	(12-11)=1	1
(10-11)=-1	1	(11-11)=0	0	(12-11)=1	1
$\Sigma(\bar{X}-\bar{\bar{X}})$	5		0		5

$SSC = \Sigma(\bar{X}_A - \bar{\bar{X}})^2 + \Sigma(\bar{X}_B - \bar{\bar{X}})^2 + \Sigma(\bar{X}_C - \bar{\bar{X}})^2$
$= 5 + 0 + 5 = 10$

Source of variation	Sum of Square	Degree of freedom	Mean Square	F
Between the sample	SSC = 10.	$v_1 = c-1 = 3-1 = 2$.	MSC = SSC/v_1 = 10/2 = 5.	MSC/MSE = 5/11.5 = 0.435
Within the Sample	SSE = 138	$v_2 = n-c = 15-3 = 12$	MSE = SSE/v_2 = 138/12 = 11.5	

Calculation of SSE.

$(A-\bar{X}_A)$	$(A-\bar{X}_A)^2$	$(B-\bar{X}_B)$	$(B-\bar{X}_B)^2$	$(C-\bar{X}_C)$	$(C-\bar{X}_C)^2$
9-10=-1	1	13-11=2	4	14-12=2	4
11-10=1	1	12-11=1	1	13-12=1	1
13-10=3	9	10-11=-1	1	17-12=5	25
9-10=-1	1	15-11=4	16	7-12=-5	25
8-10=-2	4	5-11=6	36	9-12=-3	9
$\Sigma(X-\bar{X})^2$	16		58		64

$SSE = \Sigma(A-\bar{X}_A)^2 + \Sigma(B-\bar{X}_B)^2 + \Sigma(C-\bar{X}_C)^2$
$= 16 + 58 + 64 = 138$

Calculated f value = 0.435

Tabulated value $f_{0.05} = 3.89$

Null hypothesis is passed and no significant variation in the schools.

F-Table

Critical Values of the F-Distribution: $\alpha = 0.05$

Denom.	Numerator Degrees of Freedom									
d.f.	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Data Pre-processing

Data preprocessing is the process of converting raw data into clean data that is proper for modeling. A model fails for various reasons. Data preprocessing can significantly impact model results, such as imputing missing value and handling with outliers. Data preprocessing is done to improve the quality of data in data warehouse.

- ✓ Increases efficiency
- ✓ Ease of data mining process
- ✓ Removes noisy data, inconsistent data and incomplete data

Data mining and data warehousing are very powerful and popular techniques for analyzing and storing data, respectively. Data warehousing is all about compiling and organizing data in a common database, while data mining refers to the process of extracting important data from the databases. With the definition, we can conclude that the data mining process is dependent on the data warehouse for identifying patterns in data and draw relevant conclusions. The process of data mining involves the use of statistical models and algorithms to find hidden patterns in the data.

Major tasks of Data preprocessing-

Data Cleaning: It is a process to clean the data in such a way that data can be easily integrated.

Data Integration: It is a process to integrate/combine all the data.

Data Reduction: It is a process to reduce the large data into smaller once in such a way that data can be easily transformed further.

Data Transformation: It is a process to transform the data into a reliable shape.

Data Discretization: It converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy.

Data Cleaning

After you load the data, the first thing is to check how many variables are there, the type of variables, the distributions, and data errors. It cleans the data by filling in the missing values, smoothing noisy data, resolving the inconsistency and removing the outliers.

Things to pay attention are:

- ✓ There are some missing values.
- ✓ There are outliers for store expenses (store_exp). The maximum value is 50000. Who would spend $50000 a year buying clothes? Is it an imputation error?
- ✓ There is a negative value (-500) in store_exp which is not logical.
- ✓ Someone is 300 years old.
- ✓ Enter phone number in wrong format.

How can we clean Data:

Data validation: apply some constraints to make sure you have valid and consistent data. Data validation is the process of ensuring data has undergone data cleansing to ensure they have, that is, that they are both correct and useful.

Data screening: Data screening is a method which applies to remove all error from data and make it correct for statistical analysis

De-duplication: Delete the duplicate data.

String matching method: Identify the close matches between your data and valid value

Approaches in Data Cleaning

1. Missing values
2. Noisy Data

1. Missing Values:

It is defined as the value or data that are not stored for some variable in the given dataset

How can you go about filling in the missing values for this attribute? Let's look at the following methods.

- ✓ Ignore the data row: This method is suggested for records where maximum amount of data is missing, rendering the record meaningless. This method is usually avoided where only less attribute values are missing. If all the rows with missing values are ignored i.e. removed, it will result in poor performance.
- ✓ Fill the missing values manually: This is a very time consuming method and hence infeasible for almost all scenarios.
- ✓ Use a global constant to fill in for missing values: A global constant like "NA" or 0 can be used to fill all the missing data. This method is used when missing values are difficult to be predicted.
- ✓ Use attribute mean or median: Mean or median of the attribute is used to fill the missing value.
- ✓ Use forward fill or backward fill method: In this, either the previous value or the next value is used to fill the missing value. A mean of the previous and succession values may also be used.
- ✓ Use the most probable value to fill in the missing value: (Decision tree, Regression method)

Dirty data	Examples
Incomplete data	salary=" "
Inconsistent data	Age ="5 years", Birthday ="06/06/1990", Current Year ="2017"
Noisy data	Salary = "-5000", Name = "123"
Intentional error	Sometimes applications a lot auto value to attribute. e.g some application put gender value as male by default. gender="male"

2. Noisy Data: Noise is a random error or variance in a measured variable.

Approaches for Noisy Data

1. Binning 2. Regression 3.Clustering

1. Binning Methods for Data Smoothing

The binning method can be used for smoothing the data.

Mostly data is full of noise. Data smoothing is a data pre-processing technique using a different kind of algorithm to remove the noise from the data set. This allows important patterns to stand out.

Unsorted data for price in dollars

Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

First of all, sort the data

After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Smoothing the data by equal frequency bins

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Smoothing by bin means

For Bin 1:

(8+ 9 + 15 +16 / 4) = 12

(4 indicating the total values like 8, 9 , 15, 16)

Bin 1 = 12, 12, 12, 12

For Bin 2:

(21 + 21 + 24 + 26 / 4) = 23

Bin 2 = 23, 23, 23, 23

For Bin 3:

(27 + 30 + 30 + 34 / 4) = 30
Bin 3 = 30, 30, 30, 30

Smoothing by bin boundaries

Bin 1: 8, 8, 8, 15
Bin 2: 21, 21, 25, 25
Bin 3: 26, 26, 26, 34

How to smooth data by bin boundaries?

You need to pick the minimum and maximum value. Put the minimum on the left side and maximum on the right side.
Now, what will happen to the middle values?
Middle values in bin boundaries move to its closest neighbor value with less distance.
Unsorted data for price in dollars:
Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34
First of all, sort the data
After sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34
Smoothing the data by equal frequency bins
Bin 1: 8, 9, 15, 16
Bin 2: 21, 21, 24, 26,
Bin 3: 27, 30, 30, 34
Smooth data after bin Boundary
Before bin Boundary: Bin 1: 8, 9, 15, 16
Here, 1 is the minimum value and 16 is the maximum value.9 is near to 8, so 9 will be treated as 8. 15 is more near to 16 and farther away from 8. So, 15 will be treated as 16.
After bin Boundary: Bin 1: 8, 8, 16, 16
Before bin Boundary: Bin 2: 21, 21, 24, 26,
After bin Boundary: Bin 2: 21, 21, 26, 26,
Before bin Boundary: Bin 3: 27, 30, 30, 34
After bin Boundary: Bin 3: 27, 27, 27, 34

2. Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

3. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Data Integration

Data integration is the process of merging data from several disparate sources. While performing data integration, you must work on data redundancy, inconsistency, duplicity, etc.
Data integration is important because it gives a uniform view of scattered data while also maintaining data accuracy.
Data Integration Approaches
There are mainly two types of approaches for data integration. These are as follows:
Tight Coupling: It is the process of using ETL (Extraction, Transformation, and Loading) to combine data from various sources into a single physical location.
Loose Coupling: Facts with loose coupling are most effectively kept in the actual source databases. This approach provides an interface that gets a query from the user, changes it into a format that the supply database may understand, and then sends the query to the source databases without delay to obtain the result.

Data Integration Techniques

There are various data integration techniques in data mining. Some of them are as follows:

Manual Integration: This method avoids using automation during data integration. The data analyst collects, cleans, and integrates the data to produce meaningful information. This strategy is suitable for a mini organization with a limited data set. It is a time-consuming operation.
Middleware Integration: The middleware software is used to take data from many sources, normalize it, and store it in the resulting data set.
Application-based integration: It is using software applications to extract, transform, and load data from disparate sources. This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding.
Uniform Access Integration: This method combines data from a more disparate source. However, the data's position is not altered in this scenario; the data stays in its original location. This technique merely generates a unified view of the integrated data. The integrated data does not need to be stored separately because the end-user only sees the integrated view.

Data Transformation:

Data transformation changes the format, structure, or values of the data and converts them into clean, usable data.

Data Transformation Techniques

1. Data Smoothing

Data smoothing is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

2. Attribute Construction

In the attribute construction method, the new attributes consult the existing attributes to construct a new data set that eases data mining. New attributes are created and applied to assist the mining process from the given attributes. This simplifies the original data and makes the mining more efficient.

For example, suppose we have a data set referring to measurements of different plots, i.e., we may have the height and width of each plot. So here, we can construct a new attribute 'area' from attributes 'height' and 'weight'. This also helps understand the relations among the attributes in a data set.

3. Data Aggregation

Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used.

For example, we have a data set of sales reports of an enterprise that has quarterly sales of each year. We can aggregate the data to get the enterprise's annual sales report.

4. Data Normalization: Normalizing the data refers to scaling the data values to a much smaller range such as [-1, 1] or [0.0, 1.0]. There are different methods to normalize the data, as discussed below. Consider that we have a numeric attribute A and we have n number of observed values for attribute A that are V1, V_2, V_3, ….V_n.

Min-max normalization: This method implements a linear transformation on the original data. Let us consider that we have min_A and max_A as the minimum and maximum value observed for attribute A and $V_i$is the value for attribute A that has to be normalized. The min-max normalization would map V_i to the V'_i in a new smaller range [new_min_A, new_max_A].

The formula for min-max normalization is given below:

$$v'_i = \frac{v_i - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new\_min_A$$

For example, we have $1200 and $9800 as the minimum, and maximum value for the attribute income,

and [0.0, 1.0] is the range in which we have to map a value of $73,600.
The value $73,600 would be transformed using min-max normalization as follows:

$$\frac{73600-1200}{9800-1200}(1.0-0.0)+0.0=0.716$$

Z-score normalization: This method normalizes the value for attribute A using the ***mean*** and ***standard deviation***. The following formula is used for Z-score normalization:

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}$$

Here $\bar{A}$ and σ_A are the mean and standard deviation for attribute A, respectively.
For example, we have a mean and standard deviation for attribute A as $54,000 and $16,000. And we have to normalize the value $73,600 using z-score normalization.

$$\frac{73600-5400}{1600}=1.225$$

Decimal Scaling: This method normalizes the value of attribute A by moving the decimal point in the value. This movement of a decimal point depends on the maximum absolute value of A. The formula for the decimal scaling is given below:

$$v'_i = \frac{v_i}{10^j}$$

Here j is the smallest integer such that $\max(|v'_i|)<1$

Salary bonus	Formula	CGPA Normalized after Decimal scaling
400	400 / 1000	0.4
310	310 / 1000	0.31

We will check the maximum value of our attribute "**salary bonus**". Here maximum value is 400 so we can convert it into a decimal by dividing it by 1000. Why 1000? 400 contain three digits and we so we can put three zeros after 1. So, it looks like 1000.

5. Data Discretization

This is a process of converting continuous data into a set of data intervals. Continuous attribute values are substituted by small interval labels. This makes the data easier to study and analyze. If a data mining task handles a continuous attribute, then its discrete values can be replaced by constant quality attributes. This improves the efficiency of the task.
For example, the values for the age attribute can be replaced by the interval labels such as (0-10, 11-20…) or (kid, youth, adult, senior).

6. Data Generalization

It converts low-level data attributes to high-level data attributes using concept hierarchy. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data. Data generalization can be divided into two approaches:
For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old).

Data Reduction

Data reduction techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Techniques of Data Reduction

1. Dimensionality reduction is the process in which we reduced the number of unwanted variables, attributes, and. Dimensionality reduction is a very important stage of data pre-processing. Dimensionality reduction is considered a significant task in data mining applications. For example, let's start with an example. Suppose you have a dataset with a lot of dimensions (features or columns in your database).

RollNo	Name	Mobile Number	Mobile Network
T4Tutorials1	Sameed	+92 302 XX XXX XX	Mobilink
T4Tutorials1	Ali	+92 333 XX XXX XX	Ufone

Figure 1: Before Dimension reduction

If we know Mobile Number, then we can know the Mobile Network. So we need to reduce the one dimension.

RollNo	Name	Mobile Number
T4Tutorials1	Sameed	+92 302 XX XXX XX
T4Tutorials1	Ali	+92 333 XX XXX XX

Figure 2: After Dimension reduction

In this example, we can see that if we know the mobile number, then we can know the mobile network or sim provider. So, we reduce a dimension of mobile network. When we reduce the dimensions, then you can reduce those dimensions of attributes of data by combining the dimensions in such a way that it will not lose significant characteristics of the original dataset that is going to be ready for data mining.

2. Numerosity Reduction:

Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation. There are two techniques for numerosity reduction- Parametric and Non-Parametric methods.

Parametric Methods –

For parametric methods, data is represented using some model. The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data. Regression and Log-Linear methods are used for creating such models.

Non-Parametric Methods –

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

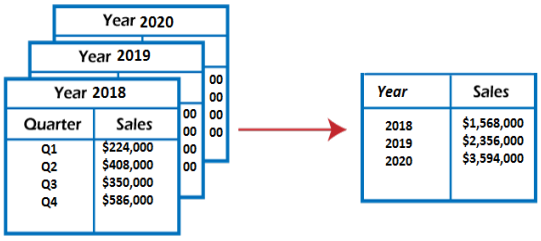
Histograms: Histogram is the data representation in terms of frequency.

Clustering: Clustering divides the data into groups/clusters.

Sampling: Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset).

3. Data Cube Aggregation

This technique is used to aggregate data in a simpler form. Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction.



Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

The data cube aggregation is a multidimensional aggregation that eases multidimensional analysis. The data cube present precomputed and summarized data which eases the data mining into fast access.

4. Data Compression

Data compression employs modification, encoding, or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form.

Data that can be restored successfully from its compressed form is called Lossless compression. In contrast, the opposite where it is not possible to restore the original form from the compressed form is Lossy compression. Dimensionality and numerosity reduction method are also used for data compression.

This technique reduces the size of the files using different encoding mechanisms, such as Huffman Encoding and run-length Encoding. We can divide it into two types based on their compression techniques.

Lossless Compression: Encoding techniques (Run Length Encoding) allow a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.

Lossy Compression: In lossy-data compression, the decompressed data may differ from the original data but are useful enough to retrieve information from them. For example, the JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image. Methods such as the Discrete Wavelet transform technique PCA (principal component analysis) are examples of this compression.

5. Discretization & Concept Hierarchy Operation:

Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals.

Concept Hierarchies:

It reduces the data size by collecting and then replacing the low-level concepts (such as 43 for age) to high-level concepts (categorical variables such as middle age or Senior).

Data Discretization

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

Table before Discretization

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level.

Some Famous techniques of data discretization:

Histogram analysis: Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

Binning: Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

Cluster Analysis

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

Data discretization using decision tree analysis

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

Data discretization using correlation analysis

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

Data discretization and concept hierarchy generation

The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance. For example, in computer science, there are different types of hierarchical systems. A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model. There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping.

Let's understand this concept hierarchy for the dimension location with the help of an example.

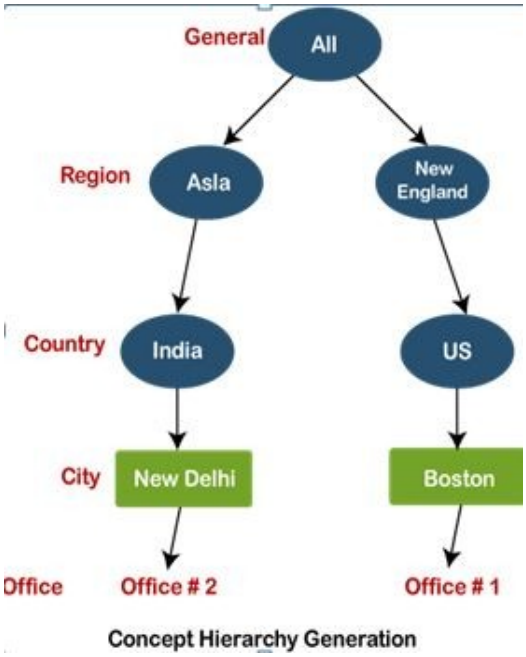
A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

Top-down mapping

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

Bottom-up mapping

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.



Concept Hierarchy Generation

Notes Prepared by **Chandrakanta Mahanty**, *Assistant Professor*
Department of CSE, 8093488380/8249119544 chandra.mahanty@giet.edu

Dimensionality Reduction :-

Ex:-

Roll NO Name Mobile Number

Date - 20/10/22 UNIT-3 Model Development

Regression :-

→ Regression analysis is a ~~stat~~ statistical Method to model the relationship between a dependent and independent variable with one or more independent variables.

→ Regression analysis reveals average relationship between two variables and this makes possible estimation / Prediction.

→ Regression used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

→ Ex:- Prediction of rain using temp. and other factor

Determining Market trends

Prediction of road accidents due to rash driving.

Simple Linear Regression :-

It is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable.

↳ The key point of Simple Linear Regression is that the dependent variable must be a continuous/ real value.

↳ ex:- Model relationship between income and expenditure.

Forecasting new observation.

Least Square Method

$$\Sigma (Y - Y_c)^2$$

Regression equation of y on x. $\boxed{Y = a + bx}$

formula:- $\boxed{\begin{aligned} \Sigma y &= Na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}}$

N = no. of observation.

Example-1

(in) (X) over	(d) (Y) run	xy	x^2
1	2	2	1
2	5	10	4
3	8	24	9
4	12	48	16
5	14	70	25
6	18	108	36
$\Sigma x = 21$	$\Sigma y = 59$	$\Sigma xy = 262$	$\Sigma x^2 = 91$

$\Sigma y = na + b\Sigma x$

$\Rightarrow 59 = 6 \times a + b \times 21$

$\Rightarrow 59 = 6a + 21b$ — (1)

Multiply 7 on eq (1)

$7(6a + 21b) = 7 \times 59$

$\Rightarrow 42a + 147b = 413$ — (3)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Rightarrow 262 = a \times 21 + b \times 91$

$\Rightarrow 262 = 21a + 91b$ — (2)

multiply 2 on eqn (2)

$2(21a + 91b) = 2 \times 262$

$\Rightarrow 42a + 182b = 524$ — (4)

Subtract eqn (3) from eqn (4), we get

$42a + 182b - 42a - 147b = 524 - 413$

$\Rightarrow 35b = 111$

$\Rightarrow b = \frac{111}{35}$

$\Rightarrow \boxed{b = 3.17}$

$y = a + bx$

$\Rightarrow \boxed{y = -1.26 + 3.17x}$

put this on eqn (1)

$6a + 21b = 59$

$\Rightarrow 6 \times a + 21 \times 3.17 = 59$

$\Rightarrow 6a + 66.57 = 59$

$\Rightarrow 6a = -7.57$

$\Rightarrow a = -\frac{7.57}{6} = -1.26.$

Q.

(x) over	(y) Runs	xy	y^2
1	2	2	4
2	5	10	25
3	8	24	64
4	12	48	144
5	14	90	196
6	18	108	324
21	59	282	757

$$\Sigma x = na + b\Sigma y$$

$$\Rightarrow 21 = 6a + 59b \quad \text{---} (1)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$\Rightarrow 282 = 59a + 757b \quad \text{---} (2)$$

$$59(6a + 59b) = 21 \times 59$$

$$\Rightarrow 354a + 3481b = 1239 \quad \text{---} (3)$$

$$6(59a + 757b) = 6 \times 282$$

$$\Rightarrow 354a + 4542b = 1692 \quad \text{---} (4)$$

$35a + 4547b - 34.95 - 3481b = 1572 - 1239$

$\Rightarrow 1066b = 333$

$\Rightarrow b = \frac{333}{1066}$

$\Rightarrow \boxed{b = 0.3}$

$6a + 59b = 21$

$\Rightarrow 6a + 59 \times 0.31 = 21$

$\Rightarrow 6a + 18.29 = 21$

$\Rightarrow 6a = 2.71$

$\Rightarrow a = \frac{2.71}{6}$

$\Rightarrow \boxed{a = 0.45}$

$X = a + by$

$\Rightarrow X = 0.41 + 0.31y$

$X = 0.41 + 0.31 \times 36$

$\Rightarrow X = 0.417$

$\Rightarrow \boxed{X = 11.57}$

Utility of Regression :-

1) Nature of Relationship
2) Estimation of Relationship.
3) Prediction
4) Useful in economic and Business Research.

Multiple Regression :-

→ It considering the values of the available multiple independent variables and predicting the value of one dependent variables.

→ The variables considered for the model should be

$\Sigma y =$

Date - 27/10/22

Q) ~~Structure~~

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$\Sigma y x_1 = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Sigma y x_2 = a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2.$$

} Multiple Regression

Stu-Name	Marks	Liveclass	Book
A	8	3	4
B	9	4	5
C	7	3	3
D	10	5	5
E	6	2	3

Name	Y	x_1	x_2	$Y x x_1$	$Y x x_2$	$x_1 x x_2$	Y^2	x_1^2	x_2^2
A	8	3	4	24	32	12	64	9	16
B	9	4	5	36	45	20	81	16	25
C	7	3	3	21	21	9	49	9	9
D	10	5	5	50	50	25	100	25	25
E	6	2	3	12	18	6	36	4	9
	$\Sigma=40$	17		$\Sigma=143$	$\Sigma=166$	$\Sigma=72$	$\Sigma=63$		$\Sigma=84$

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$40 = 5a + 17b_1 + 20b_2 \quad \text{——} (1)$$

$$\Sigma y x_1 = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Rightarrow 143 = 17a + 63b_1 + 72b_2 \quad \text{——} (2)$$

$$\Sigma y x_2 = a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

$$\Rightarrow 166 = 20a + 72b_1 + 84b_2 \quad \text{——} (3)$$

$a = 2.52$

$b_1 = 0.833$

$b_2 = 0.66$

$$\boxed{Y = 2.52 + 0.833 x_1 + 0.66 x_2}$$

Regression Error analysis :-

a)

X	Y
6	7
2	9
8	5
4	6
10	8

y on x : $y_c = a + bx$

x on y : $x_c = a + by$

X	Y	X_c	Y_c	$X - X_c$	$(X-X_c)^2$	$Y - Y_c$	$(Y-Y_c)^2$
6	7	6	7	0	0	0	0
2	9	7.6	4.8	-5.6	31.36	4.2	17.64
8	5	7.2	6.7	0.8	0.64	-1.7	2.89
4	6	6.6	7.3	-2.6	6.76	-1.3	1.69
10	8	5.4	6.4	4.6	21.16	1.6	2.56
					59.92		24.78

Reg eqn y on x $\quad y = 7.9 - 0.15x$

eq eqn x on y $\quad x = 10.2 - 0.6y$

$y_c = 7.9 - 0.15x$

$\Rightarrow y_c = 7.9 - 0.15 \times 6$

$\Rightarrow y_c = 7$

$$S_{ny} = \sqrt{\frac{\Sigma (x - x_c)^2}{N}}$$

$= \sqrt{11.984}$

$= 3.46$

$$S_{yx} = \sqrt{\frac{\Sigma (y - y_c)^2}{n}}$$

$$= \sqrt{\frac{24.78}{5}}$$

$$= \sqrt{4.946}$$

$$= 2.223$$

Date - 28/10/22

Matrix formulation for Multiple Regression Model :-

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

Q) find the coefficient of Regression in matrix form from the data below

y	9	10	13	14	16
x_1	1	3	4	6	7
x_2	10	14	15	18	20

$\beta_0 = 1$

$$y = \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T y = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix}$$

$$= \begin{bmatrix} 1+1+1+1+1 & 1+3+4+6+7 & 10+14+15+18+20 \\ 1+3+4+6+7 & 1+9+16+36+49 & 10+42+60+108+140 \\ 10+14+15+18+20 & 10+42+60+108+140 & 100+196+225+324+400 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 21 & 77 \\ 21 & 111 & 360 \\ 77 & 360 & 1245 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}$$

$$= \begin{bmatrix} 9+10+13+14+16 \\ 9+30+52+84+112 \\ 90+140+195+252+320 \end{bmatrix}$$

$$= \begin{bmatrix} 162 \\ 287 \\ 997 \end{bmatrix}$$

$$A^{-1} = \frac{1}{|A|} \cdot \text{Adj } A$$

$$\det(X^T X) = \begin{bmatrix} 5 & 21 & 77 \\ 21 & 111 & 360 \\ 77 & 360 & 1245 \end{bmatrix}$$

$$= 5\left[111 \times 1245 + 360 \times 360\right] - 21\left[\right.$$

$$= 5\left(138375 + 129600\right) - 21\left(26145 + 27720\right) + 77\left(75600 - 79541\right)$$

~~714375 = 138195 + 1240239~~

$|X| = \det(X^T X)$

adj of $(X^T X)$ ⇒ co efficient of $(X^T X)$

$$\begin{bmatrix} \begin{bmatrix} 111 & 360 \\ 360 & 1245 \end{bmatrix} - \begin{bmatrix} 21 & 360 \\ 77 & 1245 \end{bmatrix} + \begin{bmatrix} 21 & 111 \\ 77 & 360 \end{bmatrix} \\ \begin{bmatrix} 111 & 360 \\ 360 & 1245 \end{bmatrix} + \begin{bmatrix} 21 & 360 \\ 77 & 1245 \end{bmatrix} - \begin{bmatrix} 21 & 111 \\ 77 & 360 \end{bmatrix} \\ \begin{bmatrix} 111 & 360 \\ 360 & 1248 \end{bmatrix} + \begin{bmatrix} 21 & 360 \\ 77 & 360 \end{bmatrix} - \begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} (138195 - 129600) - (26145 - 27720) + (7560 - 8547 \\ -(138195 - 129600) + (26145 - 27720) - (7560 - 854 \end{bmatrix}$$

$$= \begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 298 & -183 \\ -987 & -183 & 114 \end{bmatrix} = \text{adj } A$$

$$B = (X^T X)^{-1} X^T Y = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 296 & -183 \\ -987 & -185 & 114 \end{bmatrix} * \begin{bmatrix} 63 \\ 287 \\ 947 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 532890 + 452025 - 984039 \\ 97650 + 84952 - 182451 \\ -61194 - 52521 + 113658 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 876 \\ 151 \\ -57 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 17.17 \\ 2.96 \\ -1.11 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\boxed{y_i = 17.17 + 2.96 x_1 - 1.11 x_2}$$

Date- 29/10/22

Model Evalution and Selection :-

- Metrics for performance Evalution
- Methods for performance Evalution
- Methods for Model comparison
- Model selection.

Confusion Matrix :-

A table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known

Ex

Actual class	Predicted class: yes	Predicted class: no	
yes	40 TP	20 FN	60 (P)
no	10 FP	90 TN	100 (N)
			T = P + N = 160

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Date- 31/10/22

True positive (TP)

↳ Positive tupples that were correctly level by the classified.

↳ negative tupples that were correctly leveled by the classifien.

False positive (FP)

↳ negative tupples that were incorrectly level by the positive.

↳ Positive tupples that were mislebeled as negative.

Performance matrices :—

1. Accuracy :-

↳ It is also called as Recognition Rate.

↳ Percentage of test set tupples that are correctly classified.

↳ $\frac{TP + TN}{P + N}$

2. Error Rate :-

↳ 1 - Accuracy

↳ It is also called as Misclassification Rate

↳ $\frac{FP+FN}{P+N} = \frac{10+20}{160} = 0.19$

3. Calculate sensitivity :—

↳ It also refers to the TP recognition rate.

↳ The proportion of positive tuppple that are currently identified.

↳ sensitivity = $\frac{TP}{P}$

↳ specificity :—

↳ It refers to the TN recognition rate.

↳ The proportion of negative tupples that are currently identified.

↳ specificity = $\frac{TN}{N}$

5. Precision :—

↳ It is called measure of exactness. It means what % of tupples is truely positive.

↳ It lies between 0 to 1.

↳ precision = $\frac{TP}{TP+fp}$

6. Recall :-

↳ It is a measure of completeness.

↳ what % of positive tupples are level as Positive.

↳ $\frac{TP}{TP + FN}$

7. F-Score / F-Measure :-

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{1.056}{1.46} = 0.72$$

↳ It is the harmonic mean precesion & Recall.

↳ It gives same weightage to precesion & Recall.

↳ It is also called as F_1 score.

Date - 05/11/22 Measures of insample Evalution :-

Loss function

A function that calculate loss of 1 Data point is known as Loss function.

↳ $(y_i - \hat{y}_i)^2$

y_i = Actual value

$\hat{y}_i$ = Predicted value.

cost function

A function that calculate Loss of entire data is known as Cost function.

↳ Mean sq. error = $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

n = sample size.

Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Lower the MAE

higher the accuracy

Arithmatic avg of absolute Errors.

Mean Bias Error

$$MBE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

A positive Bias Means the error from the data is over estimated

A negative bias means the error is underesti.

Relative Absolute Error

$$RAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}$$

where, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \rightarrow$ Mean of actual values.

Result - 0 - 1

$0 \rightarrow$ good model.

Mean absolute percentage error

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| * 100\%$$

Root Mean square error

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

RMSE = 0 → Perfect Model

Relative squared error :-

$$RSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ → Mean of actual value.

Relative Root Mean squared error :-

$$RRMSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\hat{y}_i)^2}}$$

Excellent when RRMSE < 10%.
good 10% to 20%.
Fair 20 to 30%
Poor > 30%.

Root Mean squared Logarithmic Error :-

$$RMSLE = \sqrt{\log(y_i + 1) - \log(\hat{y}_i + 1)^2}$$

Coefficient of Determination (R squared) :-

- It is used to analyze how differences one variable can be explained by a difference in a second variable.
- Its Ranges from 0 to 1 (0% to 100%)
- If $R^2 = 0$, then dependent variable can't be predicted from the independent variable.
- If $R^2 = 1$, then the dependent variable can be predicted from the independent variable.
- If R^2 = between 0 and 1, then the dependent variable can be predicted.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Formula-2

$$R^2 = 1 - (RSS/TSS)$$

RSS ↓ Residual sum square
TSS ↓ Total sum squa.

Date - 07/11/22

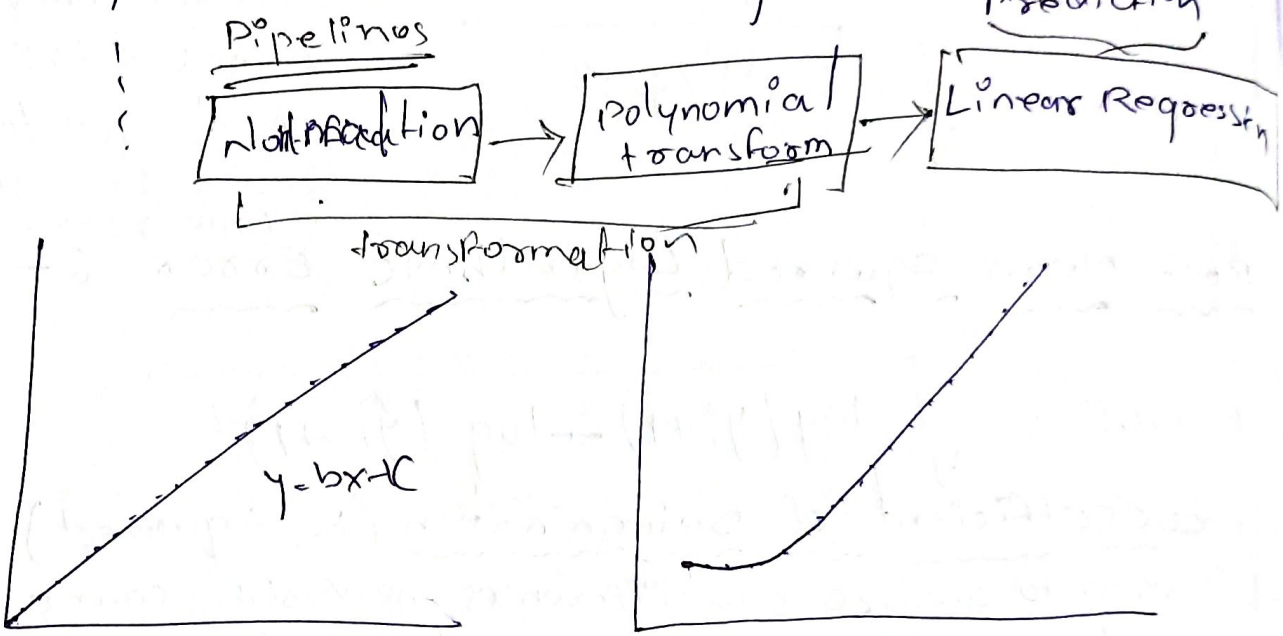
Polynomial Regression :-

↳ Polynomial Regression is a Regression algorithm that models a relationship bet'n a dependent(y) and independent variable(x) as n^{th} degree polynomial

if y = constant - 0-degree.

$y = bx^1 + c$ - 1-degree

$y = ax^2 + bx + c$ - 2-degree polynomial



$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 \ldots b_n$

↳ Polynomial Regression fits a non linear relationship.

→ Here the data points are arranged non-linear fashion.

If we apply a linear model or simple RG in a linear data set, then it provides us a good result but if we apply the same model without any modification on a non linear data set, then it will produce a drastic output. due to which loss function will increase, the error rate will be high and accuracy will be decreased. so for such cases we need the polynomial regression model.

Example :-

x	y	$y_i x_i$		x^2	x^3	x^4
3	9.5	7.5	22.5	9	27	
4	3.2	12.8	51.2	16	64	
5	3.8	19	95	25	125	
6	6.5	39	234	36	216	
7	19.5	80.5	563.5	49	343	
25	22	158.8	966.5	135	775	4659

$$\Sigma y_i = n a_0 + a_1 \Sigma x_i + a_2 (\Sigma x_i^2)$$

$$\Sigma y_i x_i = a_0 (\Sigma x_i) + a_1 (\Sigma x_i^2) + a_2 (\Sigma x_i^3)$$

$$\Sigma y_i x_i^2 = a_0 (x_i^2) + a_1 (\Sigma x_i^3) + a_2 (\Sigma x_i^4)$$

$$\Rightarrow 22 = 5 a_0 + 25 a_1 + 135 a_2$$

$$\Rightarrow 158.5 = 25 a_0 + 135 a_1 + 775 a_2$$

$$\Rightarrow 966.2 = 135 a_0 + 775 a_1 + 4659 a_2$$

$a_0 = 12.42$

$a_1 = -5.51$

$a_2 = 0.76$

$y = a_0 + a_1 x + a_2 x^2 - - - a_n x_i^2$

$y = 12.42 - 5.51 x + 0.76 x^2$

Date- 10/11/23

Residual plots for regression model validation ?

A Residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value.

↳ Residual (ϵ) = $y - \hat{y}$

Residual plot

A typical residual plot has the residual values on the y-axis and the independent variable on the X-axis.

Residual plot Analysis :-

The most important assumption of a linear regress. model is that the errors are independent and normally distributed.

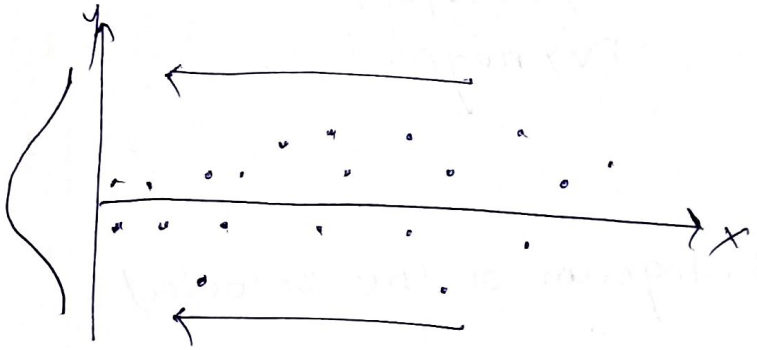
Response = Deterministic + Stochastic

Characteristics of Good Residual plots :-

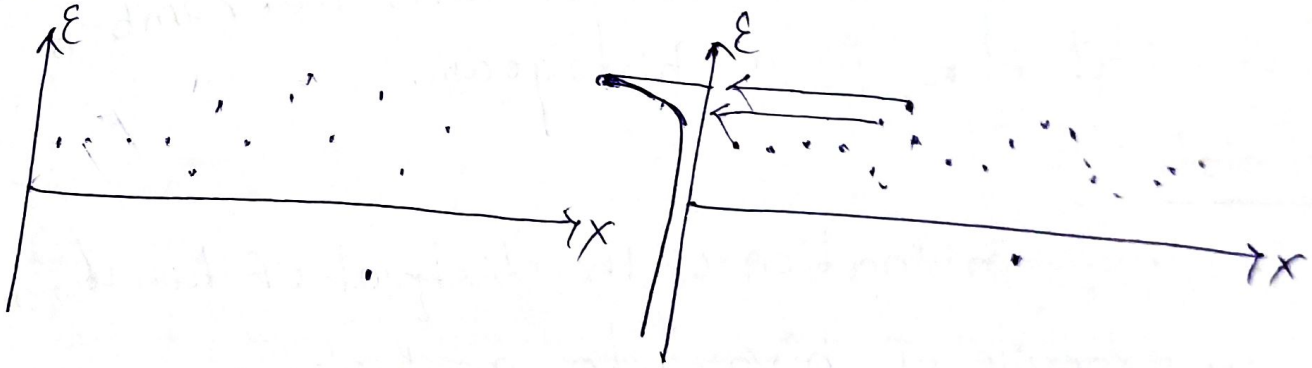
↳ It has a high density of point close to the origin and a low density of point away from the origin.

↳ It is symmetoic about the origin.

Good Residual plot



↳ Residual errors are approximately distributed in the same manner.



Bad Residual plot

Distribution plot :-

~~These plot~~

↳ seaborn is a python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

↳ These plot help us to visualize the distribution of data. we can use these plots to understand the mean, median, range, variance, deviation, etc of the data.

↳ It is of 4 types of plots :- i) joinplot
ii) distplot
iii) Pairsplot
iv) rugplot

a) Distplot

↳ Dist plot gives us the histogram of the selected continuous variable.

↳ It is an example of a univariate analysis.

↳ We can change the number of bins i.e. number of vertical bars in a histogram.

b) Join plot

→ It is the combination of the distplot of two vari

→ It is an example of bivariate analysis.

Pair plot :-

↳ It takes all the numerical attributes of the data and plot pairwise scatter plot for two different variables and histograms from the same variables.

Rugplot :-

↳ It draw a dash mark instead of a uniform distribution as in distplot.

↳ It is an example of a univariate analysis.

~~Pipelines~~ :-

Prediction and Decision making :-

↳ Do the predicted values make sense

↳ Visualization

↳ numerical measures for evaluation.

↳ comparing Models

Date-12/11/22

↳ Mean square Error for a multiple Linear Regression model will be smaller than the Mean square Error for a simple Linear Regression model, since the errors of the data will decrease when more variables are included in the model.

↳ Polynomial regression will also have a smaller Mean Square Error than the linear regular regression.

UNIT-4 — Model Evaluation — Date-14/11

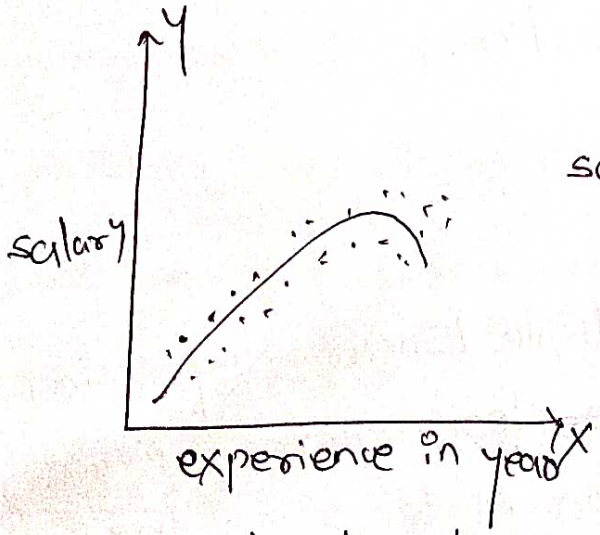
Overfitting :-

↳ Overfitting refers to a model that models the training data too well.

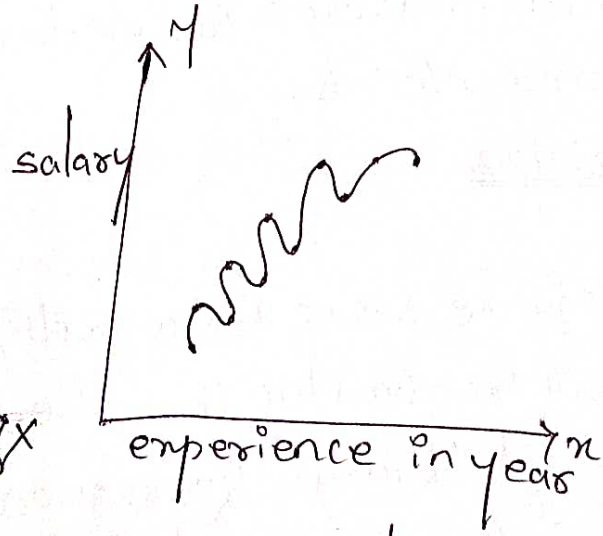
↳ It does not make accurate prediction on testing data.

↳ When a model get trained with so much data, it starts learning from the noise and inaccurat data entries in data set.

↳ Overfitting happens when a models learns the detail and noise in the training dataset to extend that it negatively impact the performance of the model.



optimal fit



overfit

* high Training Acc.
* Low Testing Acc.

→

Reasons for overfitting :-

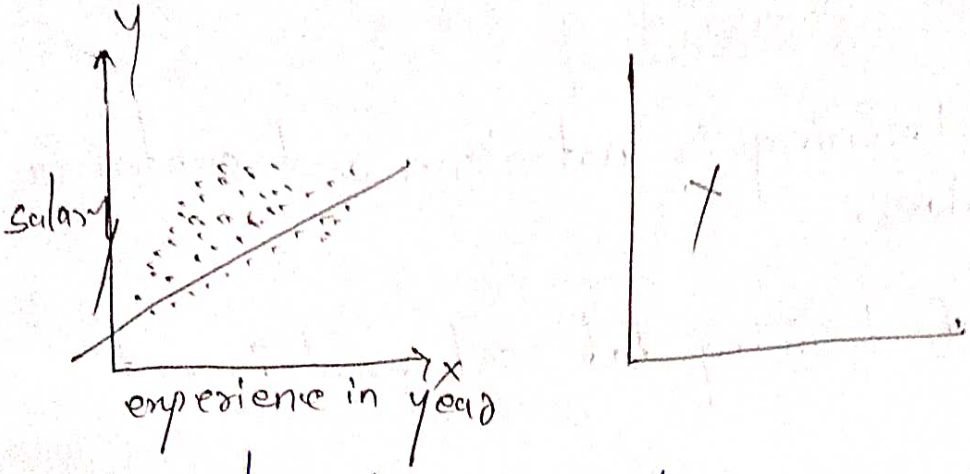
- Data used training is not clean and contains garbage values.
- The size of training data set used is not enough.
- The model is too complex.
- The model has higher variance and low Bias.

Prevention of overfitting :-

- Increase Training data.
- Reduce model complexicity.
- Ridge Regularization
- Bias - Variance Toadeoff.
- Early stopping during training phase. (Have an eye over the loss over the training Period, as soon as loss begins to increase stop training).

Underfitting :-

- Underfitting happens when the model does not learn enough from the data.
- It performs well on training data but performs Poorly on testing data.
- It destory the accuracy of the model.



It can't capture the underlying trend of data.

Reasons for underfitting :-

↳ Higher bias and Low variance.

↳ The size of the training data set used is not enough

↳ The model is too simple.

↳ Training data is not cleaned and contain noise.

Prevention of underfitting :-

↳ Increase Model complexity.

↳ Increase the no. of features.

↳ Remove noise from the data.

↳ Increase the duration of training.

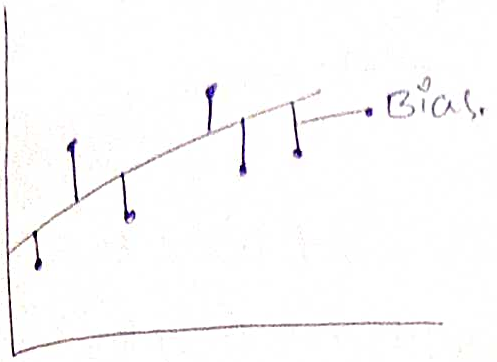
↳ Bias - Variance Tradeoff.

Date

Bias - Variance Tradeoff :-

Bias :-

Bias is the difference bet^n the average prediction of our model and the correct value which we are trying to predict.

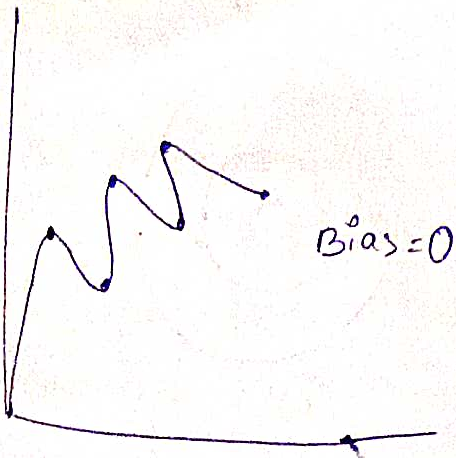


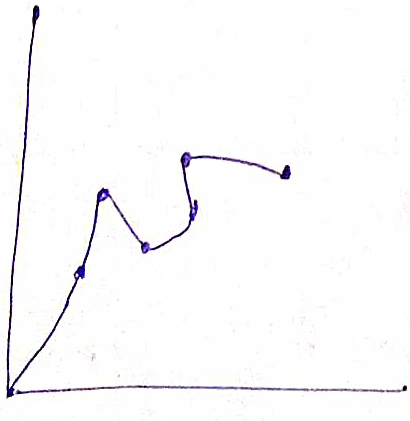
→ Bias is something

→ high Bias Means the model has not learned well on the training data. This further leads to a high error on the training and test data as the model has become oversimplified from not learning anything about the features, data points etc.

Variance :-

→ Variance is the amount that the estimate of the target function will change if different training data was used.

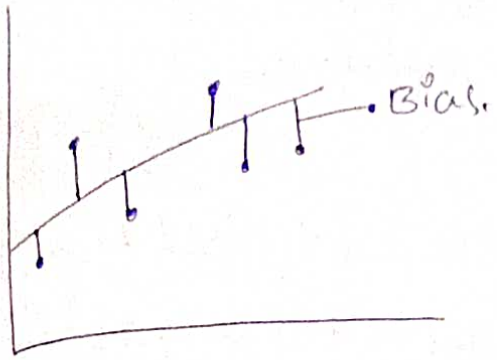




Bias - Variance Tradeoff :-

Bias :-

Bias is the difference bet^n the average prediction of our model and the correct value which we are trying to predict.

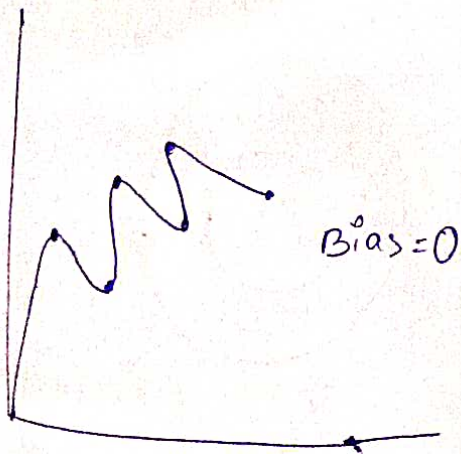


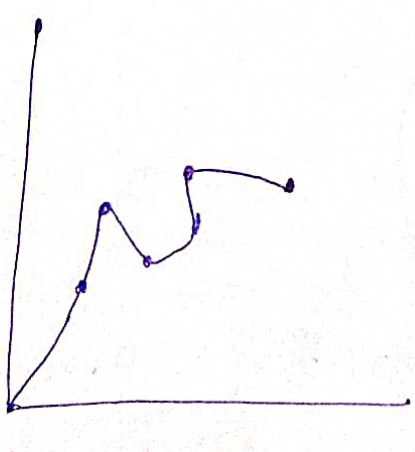
→ Bias is something

→ high Bias Means the model has not learned well on the training data. This further leads to a high error on the training and test data as the model has become oversimplified from not learning anything about the features, data points etc.

Variance :-

→ Variance is the amount that the estimate of the target function will change if different training data was used.



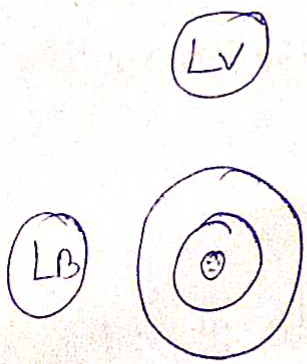


↳ high variance means the model has learned well on the training data, however it will not be able to generalize well on unseen or test data. Therefore, this will lead to having a high error rate on test data and causing overfitting.

↳ The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model.

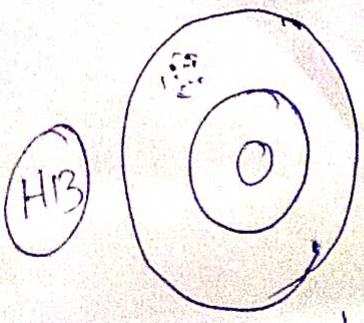
Trade-off

If a model is way too simple - this could lead to high bias and low variance. If a model is too complex with a number of parameters - this could lead to high variance and low bias. Therefore, our aim here is to find the perfect point in which overfitting or underfitting does not occur.



(1)

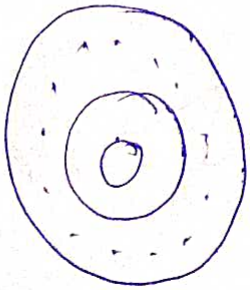
After Lockdown



(3) underfitting

(2)

During Lockdown

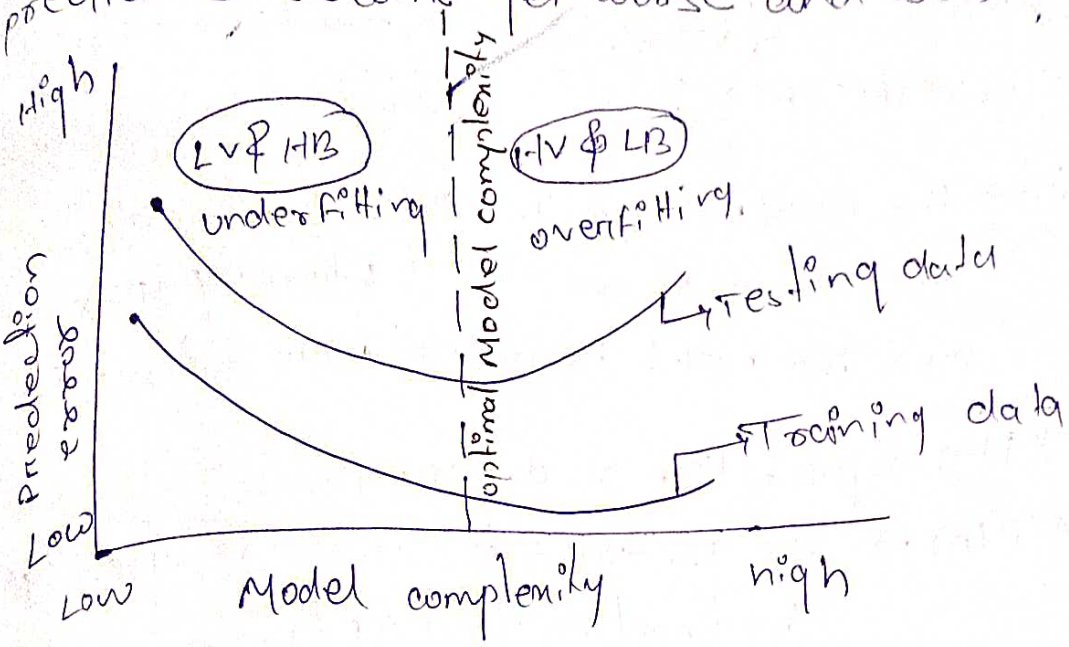


(4)

overfitting.

In the above diagram, centre of the target is a model that perfectly predicts correct values.

As we move away from the bulls-eye our predictions become get worse and worse.



[Bias-Variance Tradeoff]

Date - 19/11/22

Generalization Error :-

- It is called as out-of-sample error.
- It is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data.
- It can be minimized by avoiding overfitting in the learning algorithm.
- The performance of a machine learning algorithm is visualized by plots that show values of estimate of the generalization error through the learning process.

Mathematical notation:

D → the training dataset

x → our sample

y → the real values of the outcome variable.

y_D → the observed values of the outcome variable in dataset D.

$f(x;D)$ → the fitted values of the outcome variable, i.e. the output of our model when input $= x$, and the model is learned from dataset D

E_D → the expectation on dataset D

$Error(f;D)$ → the generalization error of model f trained on dataset D.

Generalization error

$$Error(f;D) = E[(f(x;D) - y_D)^2]$$

where

$$\overline{f(x)} = E_D[f(x;D)]$$

bias could be denoted as

$$bias^2(x) = (\overline{f(x)} - y_D)^2$$

Variance be

$$Var(x) = E_D[(f(x;D) - \overline{f(x)})^2]$$

Also noise

$$\varepsilon^2 = E_D[(y_D - y)^2]$$

Same?

$$E_{train} = \frac{1}{n}\sum_{i=1}^{n} \text{error}(f_D(x_i), y_i)$$

$$E_{gen} = \int \text{error}(f_D(x), y)\, P(y, x)\, dx$$

x: all possible 20×20 black/white bitmaps

y: {0, 1, ..., 9}

Testing Error:-

$$E_{test} = \frac{1}{w}\sum_{i=1}^{n} \text{error}(f_D(x_i), y_i)$$

Date - 21/11/22

CROSS Validation :- out-of-sample

↳ CROSS Validation is a technique.

↳ It is a re-sampling procedure used to evaluate mechine learning models and access how the model will perform for an independent test data set.

↳ It is used for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

Types of cross validation technique :-

1) Validation set approach

↳ Here training or test data split into 50% of the data set.

↳ Here the model may miss out to capture important information of the data set.

↳ It also tends to give the underfitted model.

↳ Data is to split into training data and validation data.

* used for model development

* used for validating performance of the same Model

2) Leave p out of cross validation (LPOCV)

↳ In this approach the p data set are left out of the training data.

↳ It means if there total nth data point in the original input data set, then n-p points will be used for the training data set. and the p data set used on validation set.

↳ This complete process is repeted to all the sample and the average error is calculated to know the effective race of the Model.

Disadvantage :-

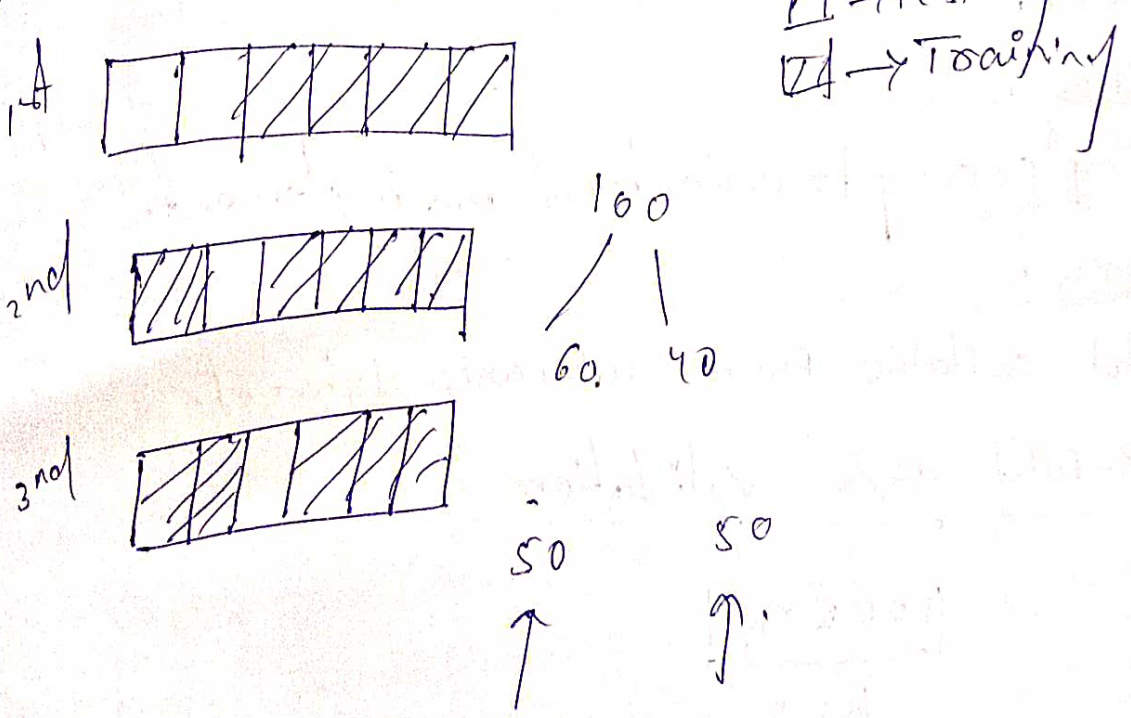
↳ It can be computational difficult for large p.

Leave one out of cross validation

↳ For a data set having 'n' rows, first rows is selected for validation and the rest n-1 rows are used for training.

↳ For next iteration, second row is selected for validation and the rest will be for training the Model.

↳ This process repeat for each data point.



□ → Testing

▨ → Training

Pros

↳ It is simple and easy to understand and implement.

Cons :-

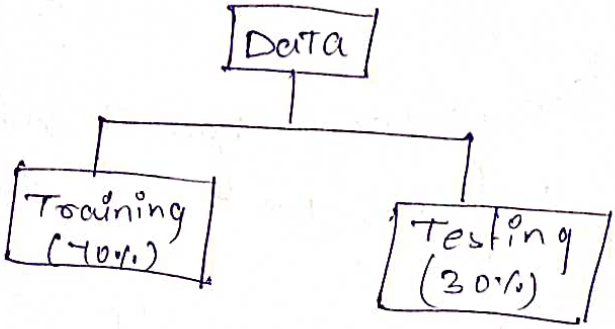
The process is executed for n times, hence execution time is high.

↳ The Model may lead to low bias.

Date - 24/11/22

4) Holdout Cross Validation :-

↳ It randomly splits the data set into train and test data.



↳ The split of training data is more than test data.

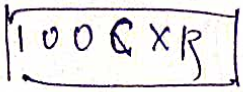
Pros :-

simple, easy to understand and implement.

cons :-

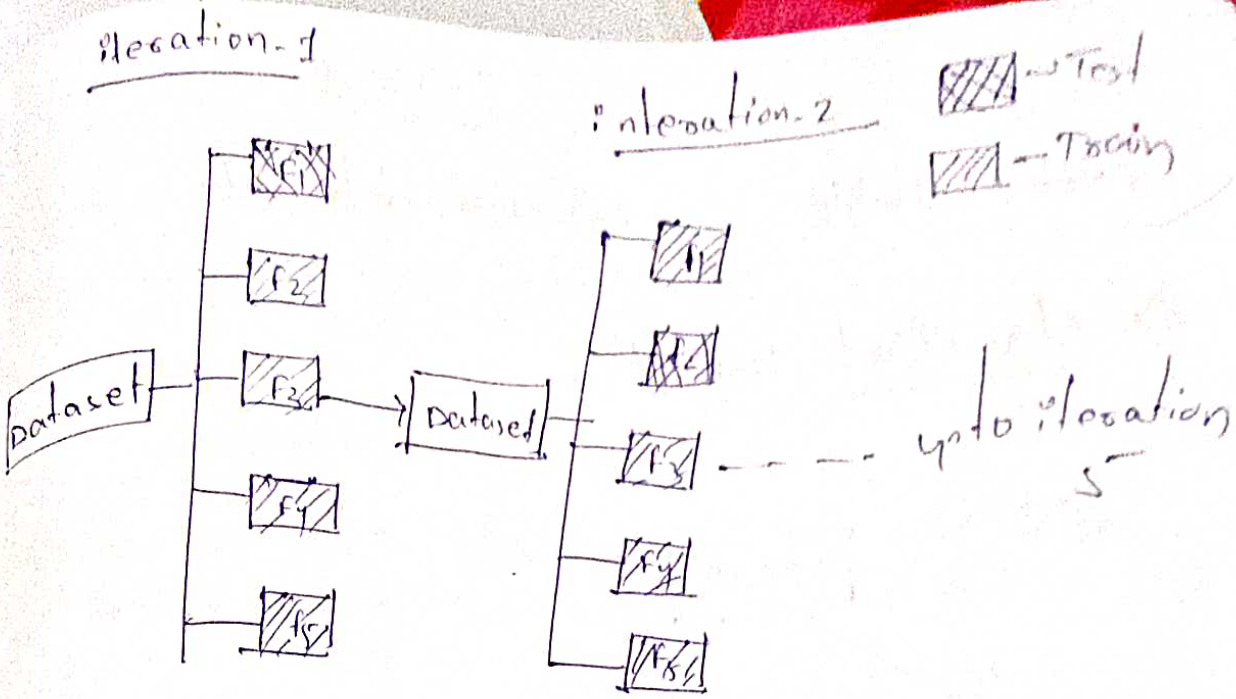
not suitable for an imbalance data set.

5) K-fold Cross Validation



$k = 5$

each fold = 100/5 = 20 images.



↳ Here the original data set is equally partion into k-folds/subparts.

↳ For each iteration one group is selected as validation data and the remaining (k-1) groups are selected as training.

↳ The process is repeated for k times until each group is treated as validation and remaining as training data.

↳ The final accuracy of the model as computed by taking the mean accuracy of the k-models validation data.

↳ $acc_{CV} = \sum_{i=1}^{k} \frac{acc_i}{k}$

PROS

↳ The model has low bias, low time complexity.

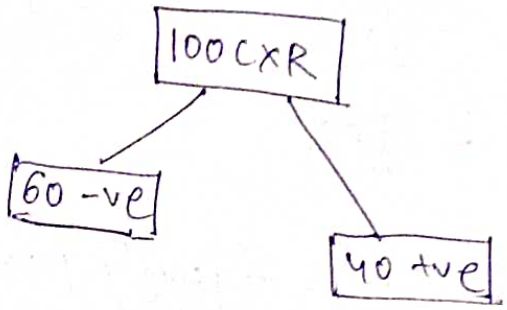
↳ The entire data set is ~~used~~ utilised for both training and testing.

Cons :-

↳ Not suitable for imbalance data set.

6) Stratified K-Fold

Key



Fold 1 | 15 (-) | 10 (+) |

each fold = $\frac{100}{4}$ = 25 images.

Fold 2 | 15 | 10 |

Fold 3 | 15 | 10 |

Fold 4 | 15 | 10 |

↳ It is a process of rearranging the data to ensure that each fold is a good representative of the complete data set.

Pros :-

↳ It works well for an imbalance data set.

Cons :-

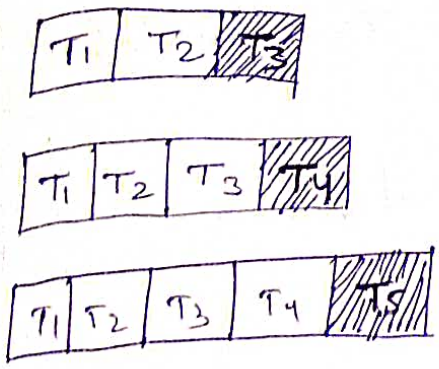
Not suitable for time series data set.

7) Time Series Cross Validation

↳ Here we split the data into train and validation is according to time.

↳ It is also called as forward chaining / Rolling cross validation.

↳ For a particular iteration, the next instance of trained data can be treated as validation data.



Date - 25/11/22

8) Nested cross-validation

In the case of k-fold and stratified k-fold cross-validation, we get a poor estimate of the error in training and test data. Hyperparameter tuning is done separately in the earlier methods. When cross-validation is used simultaneously for tuning the hyperparameters and generalizing the error estimate, nested cross-validation is required.

Model selection :-

It is the process of choosing the best suited model for a particular problem. Selecting a model depends on various factor such as the dataset, task, nature of the model.

Two factors to be considered :-

1) Logical Reason to select a model
2) Comparing the performance of the models.

↳ Models can be selected based on

i) Type of data available :

a. image & videos - CNN
b. Text data or speech data - RNN
c. Numerical data - SVM, Logistic Regression, Decision tree

2) Based on the task we need to carry out :

a. classification tasks - SVM, Logistic Regression, Decision tree
b. Regression tasks - Linear Regression, Random forest, Polynomial Regression, etc.
c. clustering tasks - k-means clustering, Hierarchical clustering.

Linear Regression :-

Advantages :-

1) very simple to implement
2) performs well on data with linear relationship

Disadvantages :-

1) Not suitable for data having non-linear relationship
2) underfitting issue
3) sensitive to outliers

* salary prediction
Monthly sales prediction

Logistic Regression

Advantages :-

1) Easy to implement
2) Performs well on data with linear relationship
3) Less prone to over-fitting for low dimensional dataset.

Disadvantages :-

1) High dimensions dataset causes over-fitting
2) Difficult to capture complex relationships in a dataset
3) Sensitive to outliers
4) Needs a larger dataset

Logistic

Decision Tree :-

Advantages :-

1) can be used for both classification & Regressing
2) Easy to interpret
3) No need for normalization or Scaling.
4) Not sensitive to outliers.

Disadvantages :-

1) Overfitting issue
2) Small changes in the data alter the tree structure causing instability.
3) Training time is relatively higher.

Ex got several recommendation based on what we are purchasing.

Date

Regularization :-

→ It is a technique to prevent the model from overfitting by adding extra information to it.

→ It is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

→ In this technique, we reduce the magnitude of the features by keeping the same number of features.

→ Technique of Regularization.

i) Ridge Regularization

ii) Lasso Regularization.

Predection by useing Ridge Regularization :-

Ridge Regularization :-

→ It is one of the types of ~~Ridge~~ linear regression in which a small amount of bias is introduced so that we can get better long-time predictions.

→ It is a regularization technique, which is used to reduce the complexity of the model. It is called as L2 regularization.

→ In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual

↳ Ridge Regression is to advantage is to avoid overfitting.

↳ Ridge regression works by attempting at increasing the bias to improve variance (generalization capability).

↳ This works by changing the slope of the line.

↳ The model performance might be little poor on the training set but it will perform consistently well on both the training and testing datasets.

↳ slope has been reduced with ridge regression penalty and therefore the model becomes less sensitive to changes in the independent variable (# years of experience).

Least Square Regression :-

Min (sum of the squared residuals)

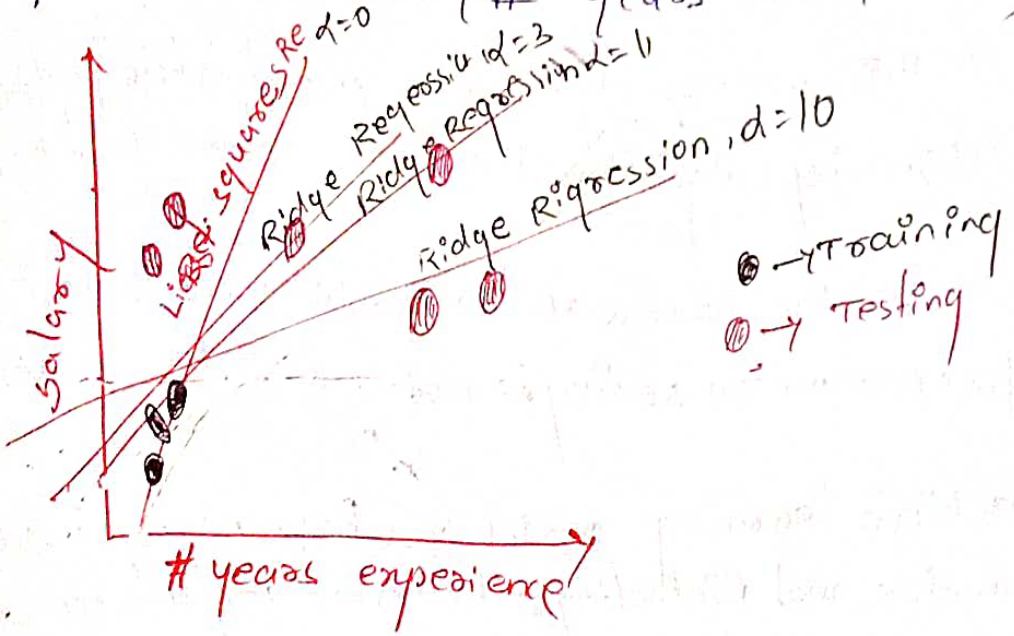
Ridge Regression :-

Min (sum of squared residuals + (α * slope^2))

↑ Penalty term

As Alpha increases, the slope of the regression line is reduced and becomes more horizontal.

As Alpha increases, the model becomes less sensitive to the variations of the independent variable (# years of experience)



Date - 28/11/22

Types of parameter :-

Parameter

Model parameter

- These are the parameter of the model that can be determined by training with training data. These can be considered as internal Parameters.
- weight, Bias
- depend on dataset
- $Y = W * X + b$

Hyperparameter

- Hyperparameters are parameters whose values control the learning process. These are adjustable parameters used to obtain an optimal model. External parameters.
- does not depend on dataset.
-

Hyperparameter tuning :-

↳ It refers to the process of choosing the optimum set of hyperparameter for a Machine Learning Model. This process is also called Hyperparameter optimizal.

↳ It is the process of finding parameter values of a learning algorithm that produce the best Model.

↳ It Makes the process of determining the best Hyperparameter setting is easier and less tedious.

↳ Machine learning model can have Many hyperparameter and finding the best combination of parameters can be prepared as such problem.

↳ We can search the best value by trial and error Method.

Hyperparameter tuning Methods :-

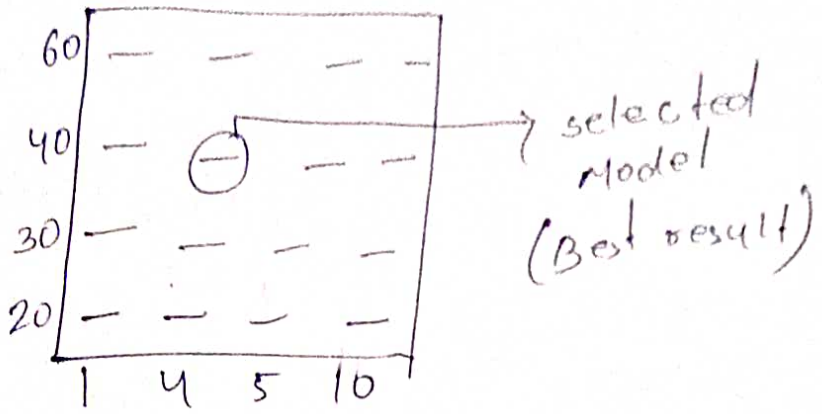
1 Grid Search :-

↳ It searches for best set of hyperparameters from a grid of hyperparameter values.

↳ It brute force all combination.

↳ Here, we simply build a Model for each possible combination of all of the hyperparameter values,

evaluating each model and selecting architecture which produce best result.



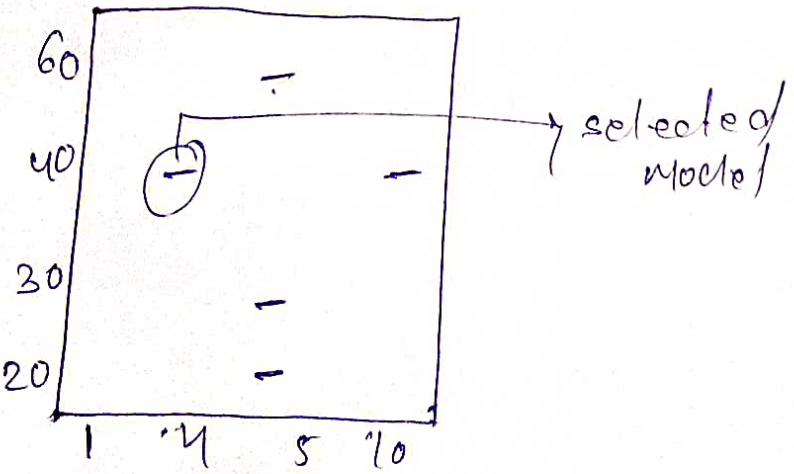
Disadvantages :-

Grid search computationally very expensive.

2) Random Search :-

↳ Instead of searching the entire grid, random search only evalutes a random sample of points on the grid.

↳ It is not chepper than grid search.



Disadvantage :-

It is very difficult to predict which model is giving the best results.

Advantages :-

It required less time than grid search.