UNIT I

❖ **Data Analysis:**

It is the technique of observing, transforming, cleaning, and modeling raw facts and figures with the purpose of developing beneficial information and acquiring profitable conclusions.

The process of studying the data to find out the answers to how and why things happened in the past. Usually, the result of data analysis is the final dataset, i.e a pattern, or a detailed report that you can further use for Data Analytics.

- Data is omnipresent in today's world, spanning various sources like spreadsheets, social media, and feedback platforms.

- Data is generated rapidly in the modern information age.

- Correct data analysis can be a company's most valuable asset.

- Effective data analysis is essential for business growth and personal success.

- When growth stagnates, analyzing past mistakes and devising new plans is necessary.

- For continuously growing businesses, ongoing analysis of data and processes is crucial.

- Data analysis is a fundamental step in achieving business and personal objectives

❖ **Data Analytics :**

Data analytics is the process of examining, cleaning, transforming, and interpreting data to extract valuable insights, patterns, and trends. It involves the use of various tools, techniques, and algorithms to make informed decisions, optimize processes, and gain a competitive edge.

Use of Data Analytics

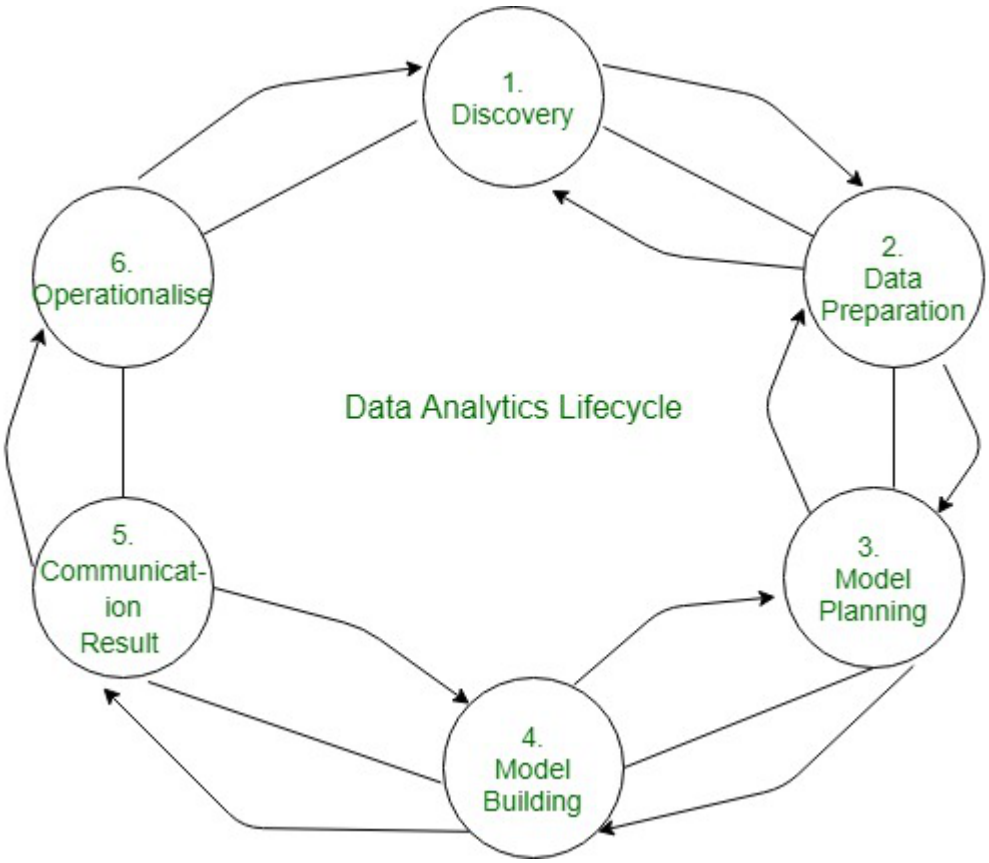
There are some key domains and strategic planning techniques in which the Data ANlaytics has played a very important role:

1. Improved Decision-Making – If we will have supporting data in favor of a decision that then we will be able to implement them with even more success probability. For example, if a certain decision or plan has to lead to better outcomes then there will be no doubt in implementing them again.
2. Better Customer Service – Churn modeling is the best example of this in which we try to predict or identify what leads to customer churn and change those things accordingly so, that the attrition of the customers is as low as possible which is a most important factor in any organization.
3. Efficient Operations – Data Analytics can help us understand what is the demand of the situation and what should be done to get better results then we will be able to streamline our processes which in turn will lead to efficient operations.
4. Effective Marketing – Market segmentation techniques have been implemented to target this important factor only in which we are supposed to find the marketing

techniques which will help us increase our sales and leads to effective marketing strategies.

Life Cycle of Data Analytics

The Data analytics lifecycle was designed to address Big Data problems and data science projects. The process is repeated to show the real projects. To address the specific demands for conducting analysis on Big Data, the step-by-step methodology is required to plan the various tasks associated with the acquisition, processing, analysis, and recycling of data.



Phase 1: Discovery -

The data science team is trained and researches the issue.

Create context and gain understanding.

Learn about the data sources that are needed and accessible to the project.

The team comes up with an initial hypothesis, which can be later confirmed with evidence.

Phase 2: Data Preparation -

Methods to investigate the possibilities of pre-processing, analysing, and preparing data before analysis and modelling.

It is required to have an analytic sandbox. The team performs, loads, and transforms to bring information to the data sandbox.

Data preparation tasks can be repeated and not in a predetermined sequence.

Some of the tools used commonly for this process include - Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning -

The team studies data to discover the connections between variables. Later, it selects the most significant variables as well as the most effective models.

In this phase, the data science teams create data sets that can be used for training for testing, production, and training goals.

The team builds and implements models based on the work completed in the modelling planning phase.

Some of the tools used commonly for this stage are MATLAB and STASTICA.

Phase 4: Model Building -

The team creates datasets for training, testing as well as production use.

The team is also evaluating whether its current tools are sufficient to run the models or if they require an even more robust environment to run models.

Tools that are free or open-source or free tools Rand PL/R, Octave, WEKA.

Commercial tools - MATLAB, STASTICA.

Phase 5: Communication Results -

Following the execution of the model, team members will need to evaluate the outcomes of the model to establish criteria for the success or failure of the model.

The team is considering how best to present findings and outcomes to the various members of the team and other stakeholders while taking into consideration cautionary tales and assumptions.

The team should determine the most important findings, quantify their value to the business and create a narrative to present findings and summarize them to all stakeholders.

Phase 6: Operationalize -

The team distributes the benefits of the project to a wider audience. It sets up a pilot project that will deploy the work in a controlled manner prior to expanding the project to the entire enterprise of users.

This technique allows the team to gain insight into the performance and constraints related to the model within a production setting at a small scale and then make necessary adjustments before full deployment.

The team produces the last reports, presentations, and codes.

Open source or free tools such as WEKA, SQL, MADlib, and Octave.

- **Difference between Data Analytics and Data Analysis :**

S.No.	Data Analytics	Data Analysis
1.	It is described as a traditional form or generic form of analytics.	It is described as a particularized form of analytics.
2.	It includes several stages like the collection of data and then the inspection of business data is done.	To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data.
3.	It supports decision making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.
4.	It uses various tools to process data such as Tableau, Python, Excel, etc.	It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc.
5.	Descriptive analysis cannot be performed on this.	A Descriptive analysis can be performed on this.
6.	One can find anonymous relations with the help of this.	One cannot find anonymous relations with the help of this.
7.	It does not deal with inferential analysis.	It supports inferential analysis.

- **Data science:**

Data science is a multidisciplinary field that combines various techniques, algorithms, processes, and systems to extract insights and knowledge from structured and unstructured data. It involves the integration of aspects from

computer science, mathematics, statistics, domain knowledge, and data engineering to analyze and interpret complex data sets.

Advantages of data science:

Improved decision-making: Data science can help organizations make better decisions by providing insights and predictions based on data analysis.

Cost-effective: With the right tools and techniques, data science can help organizations reduce costs by identifying areas of inefficiency and optimizing processes.

Innovation: Data science can be used to identify new opportunities for innovation and to develop new products and services.

Competitive advantage: Organizations that use data science effectively can gain a competitive advantage by making better decisions, improving efficiency, and identifying new opportunities.

Personalization: Data science can help organizations personalize their products or services to better meet the needs of individual customers.

Disadvantages of data science:

Data quality: The accuracy and quality of the data used in data science can have a significant impact on the results obtained.

Privacy concerns: The collection and use of data can raise privacy concerns, particularly if the data is personal or sensitive.

Complexity: Data science can be a complex and technical field that requires specialized skills and expertise.

Bias: Data science algorithms can be biased if the data used to train them is biased, which can lead to inaccurate results.

Interpretation: Interpreting data science results can be challenging, particularly for non-technical stakeholders who may not understand the underlying assumptions and methods used.

- **Buzzwords of Data Science**
  - Big Data: Refers to extremely large and complex datasets that cannot be effectively managed or analyzed using traditional data processing tools. Big data technologies like Hadoop and Spark are used to process and analyze such datasets.
  - Machine Learning: A subset of artificial intelligence (AI) that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed.
  - Artificial Intelligence (AI): The broader field of computer science focused on creating systems or machines that can perform tasks that typically require human intelligence, including problem-solving and decision-making.

- Data Mining: The process of discovering patterns, correlations, and information from large datasets, often used for knowledge discovery and making predictions.

- Predictive Analytics: Using historical data and statistical algorithms to predict future outcomes or trends, helping organizations make proactive decisions.

- Deep Learning: A subfield of machine learning that involves neural networks with multiple layers (deep neural networks) for more complex and sophisticated data analysis, often used in image and speech recognition.

- Natural Language Processing (NLP): The branch of AI that focuses on enabling computers to understand, interpret, and generate human language, used in chatbots, language translation, and sentiment analysis.

- Data Visualization: The graphical representation of data to make complex information more understandable and accessible, often using charts, graphs, and dashboards.

- Feature Engineering: The process of selecting, transforming, or creating new features (variables) from data to improve the performance of machine learning models.

- Data Wrangling: The process of cleaning, transforming, and preparing data for analysis, which often involves tasks like handling missing values and outliers.

- A/B Testing: A statistical method used to compare two versions of a webpage, product, or marketing campaign to determine which one performs better.

- Unstructured Data: Data that lacks a predefined structure, such as text, images, audio, and video, requiring specialized techniques for analysis.

- Structured Data: Data that is organized in a specific format with defined fields and relationships, making it easy to query and analyze.

- Supervised Learning: A machine learning approach where the model is trained on labeled data, meaning it learns from both input data and corresponding target outcomes.

- Unsupervised Learning: A machine learning approach where the model is trained on unlabeled data to discover patterns and structures within the data without predefined target outcomes.

- Ensemble Learning: A machine learning technique that combines multiple models (e.g., decision trees) to improve predictive performance and reduce overfitting.

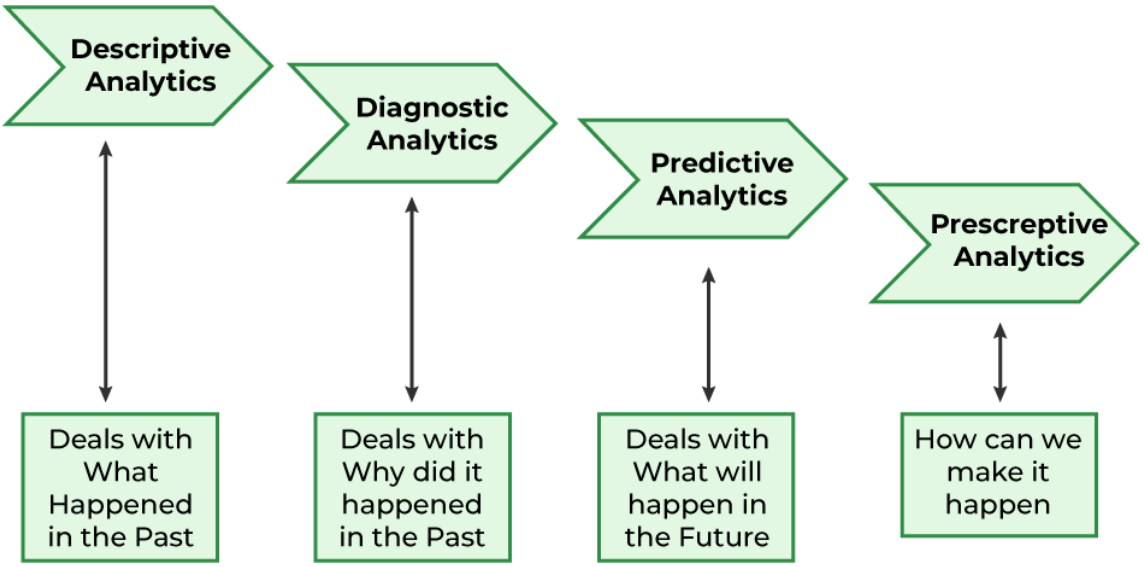
- Regression Analysis: A statistical method used to analyze the relationship between a dependent variable and one or more independent variables, often used for predictive modeling.

- Classification: A type of machine learning task where the goal is to assign data points to predefined categories or classes, such as spam vs. non-spam emails.

- Clustering: A machine learning task that involves grouping similar data points together based on their characteristics, often used for customer segmentation.

- Neural Networks: Computational models inspired by the human brain, used in deep learning for tasks like image recognition and natural language processing.

❖ **Types of Data Analytics**



There are four major types of data analytics:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)
4. Diagnostic analytics

Predictive Analytics

Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event. Techniques that are used for predictive analytics are:

Linear Regression

Time Series Analysis and Forecasting

Data Mining

Basic Corner Stones of Predictive Analytics

Predictive modeling

Decision Analysis and optimization

Transaction profiling

Descriptive Analytics

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups. Unlike a predictive model that focuses on predicting the behavior of a single customer, Descriptive analytics identifies many different relationships between customer and product.

Common examples of Descriptive analytics are company reports that provide historic reviews like:

Data Queries

Reports

Descriptive Statistics

Data dashboard

Prescriptive Analytics

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen. Further, Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.

For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.

Diagnostic Analytics

In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

For example, companies go for this analysis because it gives a great insight into a problem, and they also keep detailed information about their disposal otherwise data collection may turn out individual for every problem and it will be very time-consuming. Common techniques used for Diagnostic Analytics are:

Data discovery

Data mining

Correlations

- **Statistics:**

Statistics is a branch of mathematics and a field of study that involves the collection, analysis, interpretation, presentation, and organization of data. Its primary purpose is to make sense of data, extract meaningful information, and draw conclusions or make predictions based on that data. Statistics is widely used in various disciplines, including science, business, economics, social sciences, and more. Here are some key aspects of statistics:

Data Collection: Statistics begins with the collection of data, which can be gathered through surveys, experiments, observations, or other methods. Data can be in the form of numbers, measurements, or categories.

Data Analysis: Once data is collected, statistical techniques are applied to analyze and summarize it. This includes calculating measures of central tendency (e.g., mean, median, mode), measures of dispersion (e.g., variance, standard deviation), and exploring relationships between variables.

Data Interpretation: Statistical analysis helps in interpreting data by identifying patterns, trends, and associations. It allows researchers to draw conclusions or make inferences based on the data.

Inferential Statistics: This branch of statistics is concerned with making predictions or drawing conclusions about a population based on a sample of data. Techniques like hypothesis testing, confidence intervals, and regression analysis are used for inference.

Descriptive Statistics: Descriptive statistics involves summarizing and presenting data in a meaningful way. Common tools include tables, charts, histograms, and summary statistics.

Basics of Statistics

The basics of statistics comprise of :

1. Measure of Central tendency (Mean, Median, Mode)
2. Measure of Variance, Range, Quartile

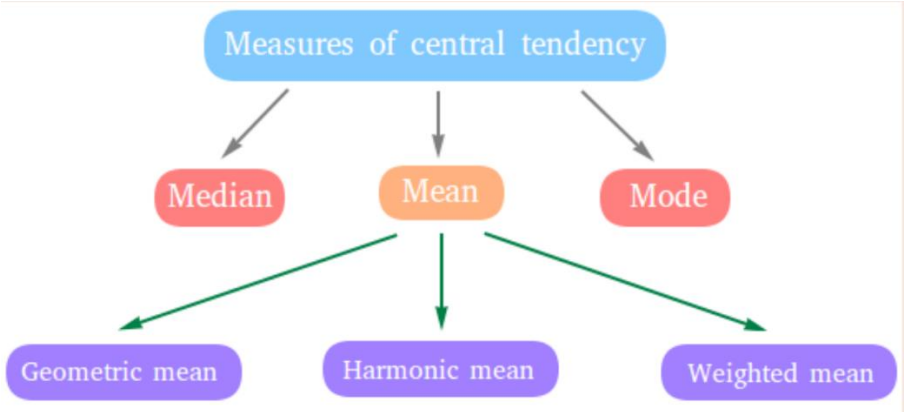
Mean is the average of the observations.

Median is the middle or the central value of the observations when arranged in order.

The mode is the most frequently occurring observation in the data set.

Variance is defined as the measure of how far a set of data are dispersed from its mean value and is calculated by the average of the squared differences from the mean. Standard Deviation is the square root of variance.

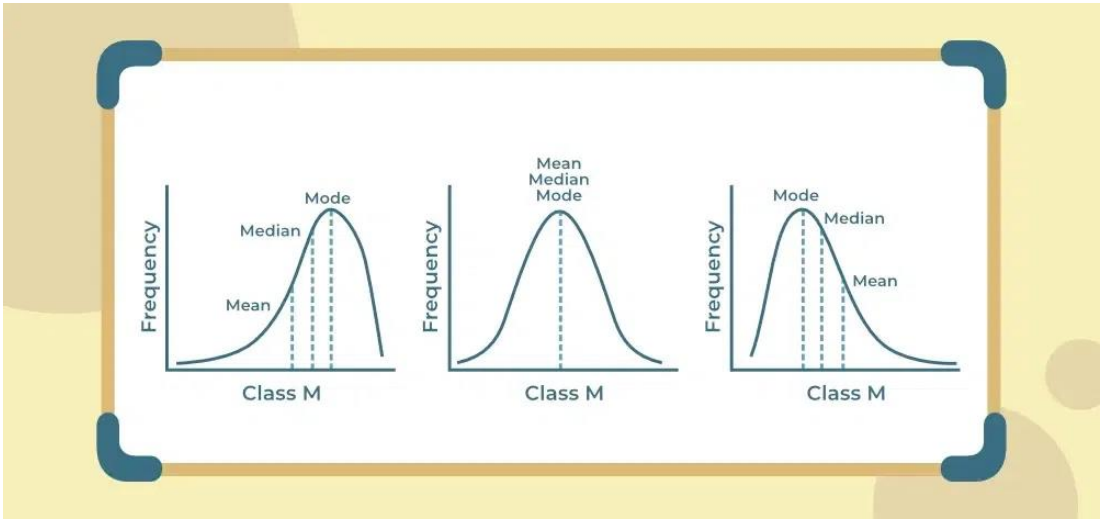
1. **Measure of central tendency**



Some of the most commonly used measures of central tendency are:

1) Mean
2) Median

3) Mode



1)Mean:

Mean in general terms is used for the arithmetic mean of the data, but other than the arithmetic mean there are geometric mean and harmonic mean as well that are calculated using different formulas.

Mean for Ungrouped Data

Arithmetic mean ($\bar{X}$) is defined as the sum of the individual observations (X_i) divided by the total number of observations N. In other words, the mean is given by the sum of all observations divided by the total number of observations.

$$\bar{x} = \frac{\sum x_i}{N}$$

OR

Mean = Sum of all Observations ÷ Total number of Observations

Example:

If there are 5 observations, which are 27, 11, 17, 19, and 21 then the mean ($\bar{X}$) is given by

$\bar{X}$ *= (27 + 11 + 17 + 19 + 21) ÷ 5*

= 95 ÷ 5

= 19

Mean for Grouped Data

Mean ($\bar{X}$)) is defined for the grouped data as the sum of the product of observations (X_i)) and their corresponding frequencies (f_i) divided by the sum of all the frequencies (f_i).

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Example: Calculate the mean height for the following data using the direct method.

Height (in inches)	60 – 62	62 – 64	64 – 66	66 – 68	68 – 70	70 – 72
						2

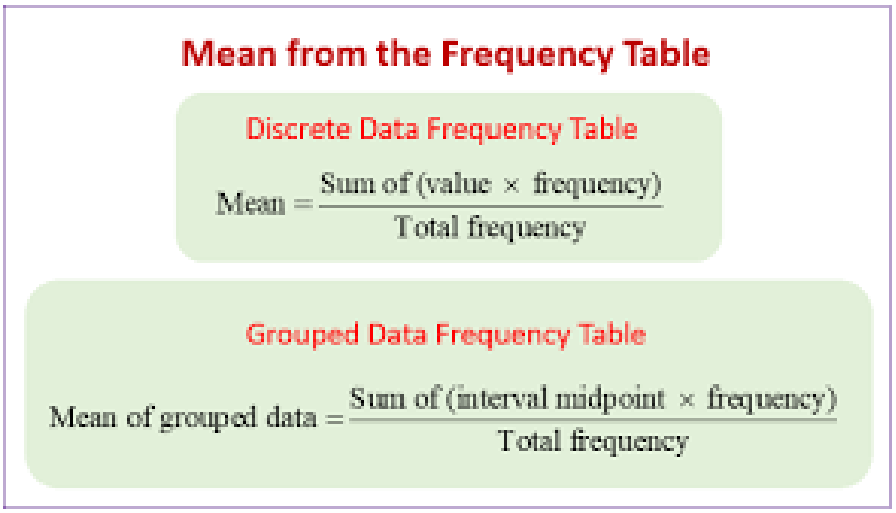
Solution:

As,

Height (in inches)	Frequency(f_i)	Midpoint (x_i)	$f_i \times x_i$
60 – 62	3	61	183
62 – 64	6	63	378
64 – 66	9	65	585
66 – 68	12	67	804
68 – 70	8	69	552
70 – 72	2	71	142
	$\sum f_i = 40$		**$\sum f_i x_i = 2644$**

⇒ Mean = 2644/40 = 66.1

Thus, mean height is 66.1 inches.



Types of Mean

Mean can be classified into three different class groups which are

1) Arithmetic Mean
2) Geometric Mean
3) Harmonic Mean

Arithmetic Mean: The formula for Arithmetic Mean is given by

$$\bar{x} = \frac{\sum x_i}{N}$$

Where,

- ***x1, x2, x3, . . ., xn*** *are the observations, and*
- ***N*** *is the number of observations.*

Geometric Mean: The formula for Geometric Mean is given by

$$\text{Geometric mean} = \sqrt[n]{x1 * x2 * x3 \ldots\ldots * xn}$$

Where,

- ***x1, x2, x3, . . ., xn*** *are the observations, and*
- ***n*** *is the number of observations.*

Harmonic Mean: The formula for Harmonic Mean is given by

$$\text{Harmonic mean} = \frac{n}{\left(\frac{1}{(x1)} + \frac{1}{(x2)} + \frac{1}{(x3)} + \ldots\ldots + \frac{1}{(xn)}\right)}$$

Where,

- ***x1, x2, . . ., xn*** *are the observations, and*
- ***n*** *is the number of observations.*

Disadvantage of Mean as Measure of Central Tendency

Although Mean is the most general way to calculate the central tendency of a dataset however it can not give the correct idea always, especially when there is a large gap between the datasets.

2) Median

The Median of any distribution is that value that divides the distribution into two equal parts such that the number of observations above it is equal to the number of observations below it. Thus, the median is called the central value of any given data either grouped or ungrouped.

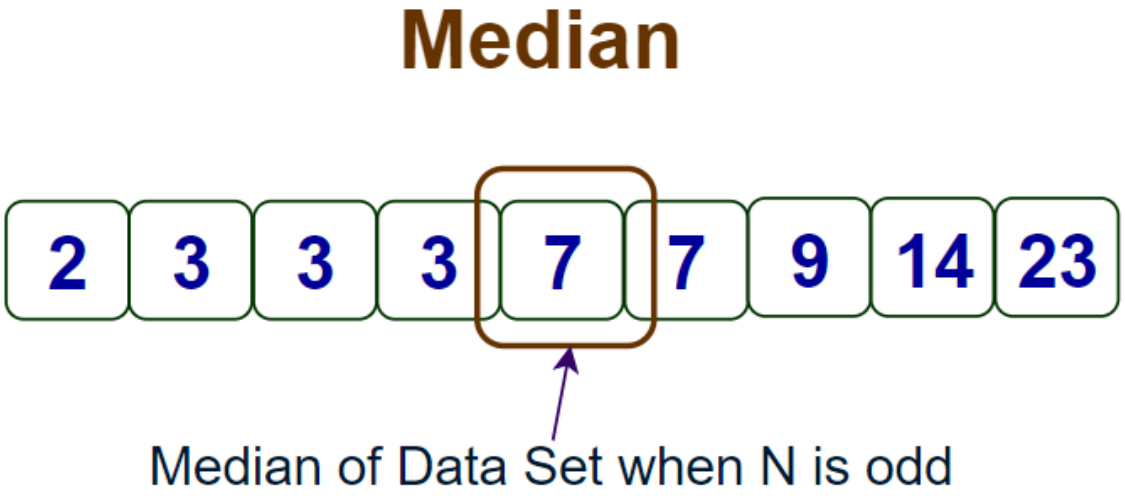
Median of Ungrouped Data

To calculate the Median, the observations must be arranged in ascending or descending order. If the total number of observations is N then there are two cases

Case 1: N is Odd

Median = Value of observation at [(n + 1) ÷ 2]th Position

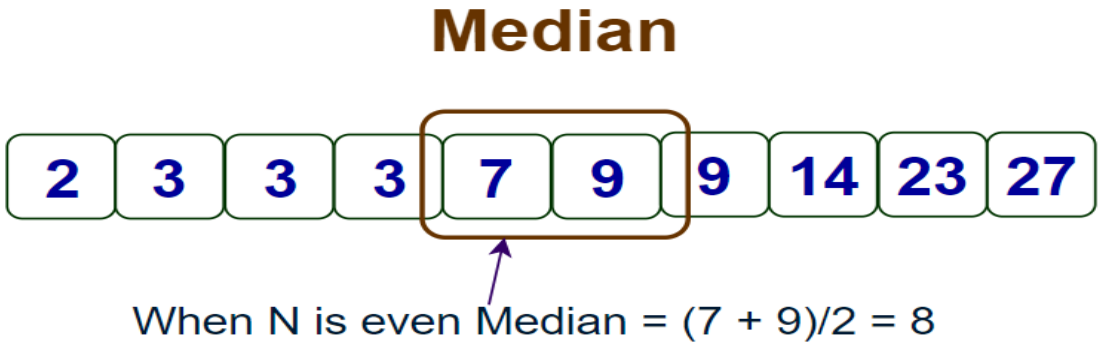
When **N** is odd the median is calculated as shown in the image below.



Case 2: N is Even

Median = Arithmetic mean of Values of observations at (n ÷ 2)th and [(n ÷ 2) + 1]th Position

When **N** is even the median is calculated as shown in the image below.



Example 1: If the observations are 25, 36, 31, 23, 22, 26, 38, 28, 20, 32 then the Median is given by

*Arranging the data in ascending order: 20, 22, 23, 25, **26**, **28**, 31, 32, 36, 38*

N = 10 which is even then

Median = Arithmetic mean of values at (10 ÷ 2)th and [(10 ÷ 2) + 1]th position

⍰ Median = (Value at 5th position + Value at 6th position) ÷ 2

⍰ Median = (26 + 28) ÷ 2

⍰ Median = 27

Example 2: If the observations are 25, 36, 31, 23, 22, 26, 38, 28, 20 then the Median is given by

*Arranging the data in ascending order: 20, 22, 23, 25, **26**, 28, 31, 36, 38*

N = 9 which is odd then

Median = Value at [(9 + 1) ÷ 2]th position

Median = Value at 5th position

Median = 26

$$\text{Median} = \frac{n+1}{2} \text{ th term , n = odd}$$

$$\left\{\frac{n}{2} \text{ th term} + \frac{n}{2} + 1 \text{ th term}\right\} / 2 \text{ , n = even}$$

Median of Grouped Data

The Median of Grouped Data is given as follows:

$$\text{Median} = l + \frac{N/2 - c_f}{f} \times h$$

Where,

- ***l** is the lower limit of median class,*
- ***n** is the total number of observations,*
- ***cf** is the cumulative frequency of the preceding class,*
- ***f** is the frequency of each class, and*
- ***h** is the class size.*

Example: Calculate the median for the following data.

Class	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Frequency	5	10	12	8	5

Solution:

Create the following table for the given data.

Class	Frequency	Cumulative Frequency
10 – 20	5	5
20 – 30	10	15
30 – 40	12	27
40 – 50	8	35
50 – 60	5	40

As n = 40 and n/2 = 20,
Thus, 30 – 40 is the median class.
l = 30, c_f = 15, f = 12, and h = 10

$$\text{Median} = l + \frac{N/2 - c_f}{f} \times h$$

Putting the values in the formula
Median = 30 + (20 – 15)/12) × 10
⇒ Median = 30 + (5/12) × 10
⇒ Median = 30 + 4.17
⇒ Median = 34.17
So, the median value for this data set is 34.17

3)Mode:

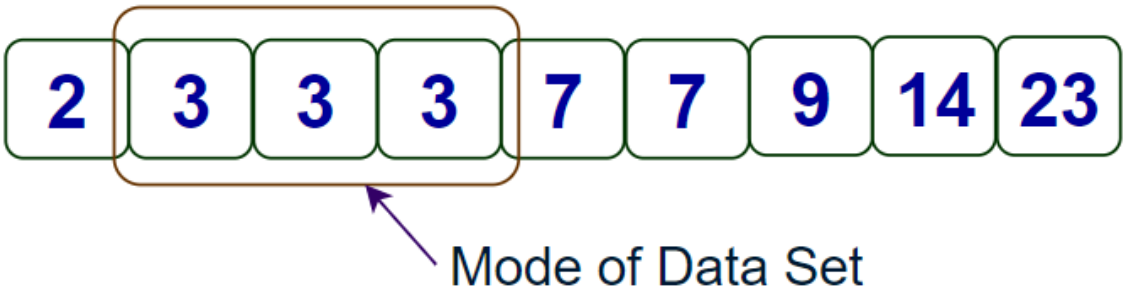
The Mode is the value of that observation which has a maximum frequency corresponding to it. In other, that observation of the data occurs the maximum number of times in a dataset.

Mode of Ungrouped Data

Mode of Ungrouped Data can be simply calculated by observing the observation with the highest frequency. Let's see an example of the calculation of the mode of ungrouped data.

The mode of the data set is the highest frequency term in the data set as shown in the image added below.

Mode



Example: Find the mode of observations 5, 3, 4, 3, 7, 3, 5, 4, 3.
Solution:
Create a table with each observation with its frequency as follows:

xi	5	3	4	7
fi	2	4	2	1

Since 3 has occurred a maximum number of times i.e. 4 times in the given data;
Hence, Mode of the given ungrouped data is 3.

Mode of Grouped Data

The formula to find the mode of the grouped data is:

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] h$$

Where,

- ***l*** *is the lower class limit of modal class,*
- ***h*** *is the class size,*
- ***f1*** *is the frequency of modal class,*
- ***f0*** *is the frequency of class which proceeds the modal class, and*
- ***f2*** *is the frequency of class which succeeds the modal class.*

$$\text{Mode} = l + \left(\frac{f1 - f0}{2f1 - f0 - f2}\right) \times h$$

Class Interval	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	16	10

Solution:

As the class interval with the highest frequency is 40-50, which has a frequency of 16. Thus, 40-50 is the modal class.

*Thus, **l** = 40 , **h** = 10 , **f1** = 16 , **f0** = 12 , **f2** = 10*

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] h$$

Plugging in the values in formula , we get

Mode = 40 + (16 – 12)/(2 × 16 – 12 – 10) × 10

Mode = 40 + (4/10)×10

Mode = 40 + 4

Mode = 44

Therefore, the mode for this set of data is 44.

❖ Empirical Relation Between Measures of Central Tendency

The three central tendencies are related to each other by the empirical formula which is given as follows:

2 × Mean + Mode = 3 × Median

This formula is used to calculate one of the central tendencies when two other central tendencies are given.

- What is the purpose of central tendency?

The primary goal of central tendency is to offer a single value that effectively represents a set of collected data. This value aims to capture the core or typical aspect of the data, providing a concise summary of the overall information.

❖ Difference between Mean, Mode, Median

	Mean	Median	Mode
Meaning	The average number across a group of statistics is known as the mean	The middle number in the collection of data is called the median	The mode is the number that occurs in data collection the most commonly
Type of Average	It is an average determined by math	It is an average of positions	A positional average is known as a mode
Basis	The mean is affected by every single statistic	The median is the midpoint	Mode is a common item
Capability	Mean can be further algebraically processed	The median needs to be more competent	Mode is not also capable

Observation	Mean can only be determined mathematically	Simple observation may be used to determine the median	Additionally, it may be determined via simple observation
Location	In the graph, the mean cannot be found	On the graph, the median may be found	On a graphic, the mode can also be located
Affected by	Extreme numbers impact these	The middle is not significantly affected by extreme numbers	Extreme values also do not have a significant impact on mode
Defined	In each instance, the mean is clearly defined	In every instance, the median is clearly stated	In some instances, the definition of mode could be more straightforward
Usage	It cannot be used when %. 1. The distribution is substantially skewed 2. Open-ended classes are included in the distribution. 3. The average needed is for rates and ratios	When the following conditions exist 1. The data cannot be measured directly 2. Each of the groups in the distribution is open-ended	A lot of the time, the mode could be more straightforward 1. When expressing preferences in issues, the mode is employed

2. Measure of variability:

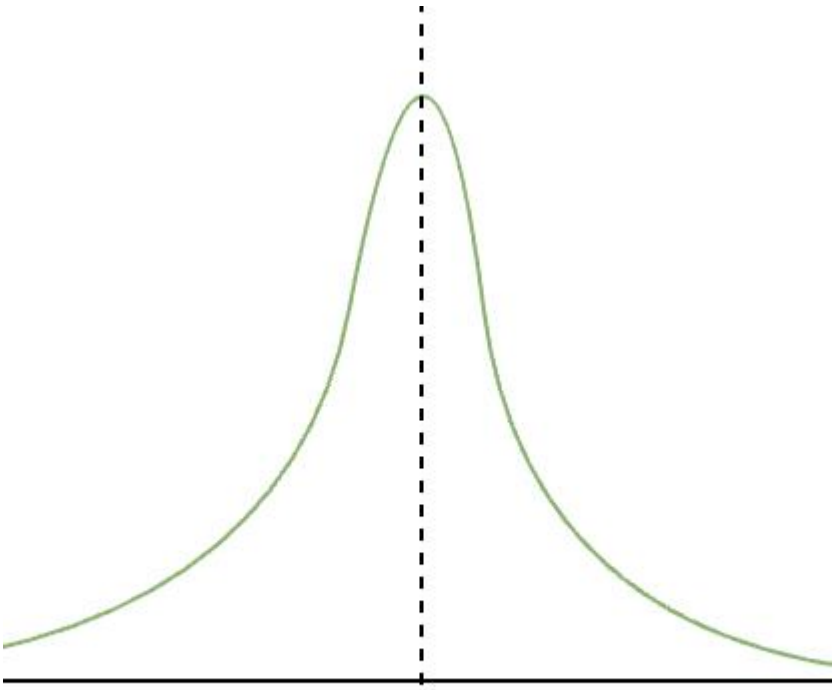
Variability is most commonly measured with the following descriptive statistics:

1) Range: the difference between the highest and lowest values
2) Interquartile range: the range of the middle half of a distribution
3) Standard deviation: average distance from the mean
4) Variance: average of squared distances from the mean

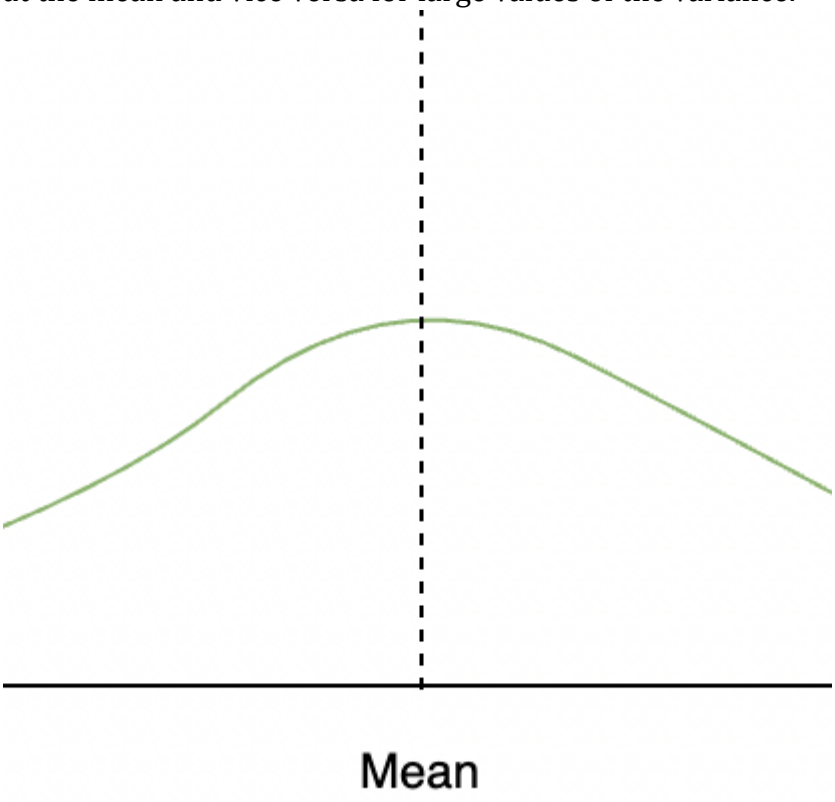
1) Variance:

The variance of the data is given by measuring the distance of the observed values from the mean of the distribution. Here we are not concerned with the sign of the distance of the point, we are more interested in the magnitude. So, we take squares of the distance from the mean. Let's say we have x_1, x_2, x_3 .... x_n as n observations and be the mean.

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 \ldots (x_n - \bar{x})^2 = \sum_0^n (x_i - \bar{x})^2$$



If this sum zero, then each term has to be zero which means that there is no scattering in the data. If it is small, then it means that the data is concentrated at the mean and vice versa for large values of the variance.



But this measure is still dependent on the number of observations in the data. That is if there are lots of observations this value will become large. So, we take the mean of the data,

$$\text{Variance} = \frac{\sum_0^n (x_i - \bar{x})^2}{n}$$

$$\sigma^2 = \frac{\sum_0^n (x_i - \bar{x})^2}{n}$$

2) **Range:**

The range of the data is given as the difference between the maximum and the minimum values of the observations in the data.

For example, let's say we have data on the number of customers walking in the store in a week.

10, 14, 8, 10, 15, 4, 7

Minimum value in data = 7

Maximum Value in the data = 15

Range = Maximum Value in the data – Minimum value in the data

= 15 – 7

= 8

Now we can say that the range of the data is 8. This gives us an idea about the spread of the data but doesn't tell how the data is distributed.

Example1: Ungrouped Data - Range

Given the following dataset: 12, 15, 17, 20, 22, 25, 30, 35, 40, 45, calculate the range.

Formula:

Range = Maximum Value - Minimum Value

1. Find the maximum value in the dataset: Max = 45

2. Find the minimum value in the dataset: Min = 12

3. Calculate the range: Range = Max - Min = 45 - 12 = 33

Example 2: Ungrouped Data - Variance

Using the same dataset as Example 1, calculate the variance.

Formula:

Variance = $\Sigma((x - \mu)^2) / (n - 1)$

Steps:

1. Calculate the mean (average) of the dataset: $\mu = (12 + 15 + 17 + 20 + 22 + 25 + 30 + 35 + 40 + 45) / 10 = 27.2$

2. Calculate the squared difference between each data point and the mean.

- For example, for 12: $(12 - 27.2)^2 = 233.64$

3. Sum all the squared differences: $\Sigma((x - \mu)^2) = 1215.6$

4. Divide the sum by (n - 1), where n is the number of data points: Variance = 1215.6 / (10 - 1) = 135.07 (rounded to two decimal places)

Example 3: Grouped Data - Range

Given the frequency distribution below, calculate the range:

Class Interval	Frequency
10-20	5
20-30	8
30-40	12
40-50	7

Formula:

Range = Maximum Value - Minimum Value

Steps:

1. Find the maximum class interval boundary: Max = 50

2. Find the minimum class interval boundary: Min = 10

3. Calculate the range: Range = Max - Min = 50 - 10 = 40

Example 4: Grouped Data - Variance

Using the same frequency distribution as Problem 3, calculate the variance.

Formula:

Variance = $[\Sigma(fx^2) / n] - [(\Sigma fx / n)^2]$

Steps:

1. Calculate the midpoint (x) for each class interval.

- For 10-20: $x = (10 + 20) / 2 = 15$

- For 20-30: $x = (20 + 30) / 2 = 25$

- For 30-40: $x = (30 + 40) / 2 = 35$

- For 40-50: x = (40 + 50) / 2 = 45

2. Calculate fx (frequency multiplied by midpoint) for each class.

- For 10-20: fx = 5 * 15 = 75

- For 20-30: fx = 8 * 25 = 200

- For 30-40: fx = 12 * 35 = 420

- For 40-50: fx = 7 * 45 = 315

3. Calculate Σfx: Σfx = 75 + 200 + 420 + 315 = 1010

4. Calculate $\Sigma(fx^2)$ (frequency multiplied by squared midpoint) for each class.

- For 10-20: $fx^2 = 5 * (15)^2 = 1125$

- For 20-30: $fx^2 = 8 * (25)^2 = 5000$

- For 30-40: $fx^2 = 12 * (35)^2 = 14700$

- For 40-50: $fx^2 = 7 * (45)^2 = 14175$

5. Calculate $\Sigma(fx^2)$: $\Sigma(fx^2) = 1125 + 5000 + 14700 + 14175 = 35000$

6. Calculate n (total frequency): n = 5 + 8 + 12 + 7 = 32

7. Calculate the variance formula using the values from steps 3, 5, and 6:

- Variance = $[\Sigma(fx^2) / n] - [(\Sigma fx / n)^2]$

- Variance = $[35000 / 32] - [(1010 / 32)^2] = 1093.75 - 320.31 = 773.44$

3)Quartiles

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. They are essential in analyzing and understanding the distribution and variability of data. Quartiles are especially useful for identifying the spread and central tendency of data, as well as detecting outliers.

Formulas for Quartiles:

1. **First Quartile (Q1):** This is the 25th percentile, separating the lowest 25% of the data.
    - **For Ungrouped Data:**

        Q1 = (n + 1) / 4-th value in the ordered dataset, where n is the number of data points.

    - **For Grouped Data:**

        Q1 = L1 + [(N/4 - F1) / f1] * C1

        Where:

        L1 is the lower boundary of the class interval containing Q1.

        N is the total number of data points.

        F1 is the cumulative frequency of the class interval before Q1.

f1 is the frequency of the class interval containing Q1.

C1 is the width of the class interval.

2. **Second Quartile (Q2):** This is the median, dividing the data into two equal halves.
   - **For Ungrouped Data:**

     Q2 = (n + 1) / 2-th value in the ordered dataset, where n is the number of data points.
   - **For Grouped Data:**

     Q2 is calculated the same way as the median.
3. **Third Quartile (Q3):** This is the 75th percentile, separating the lowest 75% of the data.
   - **For Ungrouped Data:**

     Q3 = (3n + 1) / 4-th value in the ordered dataset, where n is the number of data points.
   - **For Grouped Data:**

     Q3 = L3 + [(3N/4 - F3) / f3] * C3

     Where:

     L3 is the lower boundary of the class interval containing Q3.

     N is the total number of data points.

     F3 is the cumulative frequency of the class interval before Q3.

     f3 is the frequency of the class interval containing Q3.

     C3 is the width of the class interval.

Example for Ungrouped Data:

Consider the dataset: 12, 15, 17, 20, 22, 25, 30, 35, 40, 45 (n = 10).

1. Calculate Q2 (the median):
   - Q2 = (10 + 1) / 2 = 5.5-th value, so Q2 = (5th value + 6th value) / 2 = (22 + 25) / 2 = 23.5
2. Calculate Q1:
   - Q1 = (10 + 1) / 4 = 2.75-th value, so Q1 = (2nd value + 3rd value) / 2 = (15 + 17) / 2 = 16
3. Calculate Q3:
   - Q3 = (3 * 10 + 1) / 4 = 7.75-th value, so Q3 = (7th value + 8th value) / 2 = (30 + 35) / 2 = 32.5

Quartile Deviation:

Quartile Deviation, also known as Semi-Interquartile Range (SIQR), measures the spread of data within the interquartile range (Q3 - Q1) and is calculated as:

Quartile Deviation (QD) = (Q3 - Q1) / 2

- It represents half the difference between the third quartile (Q3) and the first quartile (Q1).
- QD is a measure of the middle 50% of the data's variability and is less sensitive to outliers.
- A smaller QD indicates less variability within the interquartile range, while a larger QD suggests more variability.

	Frequency
10-20	5
20-30	8
30-40	12
40-50	7

1. Calculate Q2 (the median) using the same method as for ungrouped data: Q2 = 30 + [(10/2 - 5) / 8] * 10 = 30 + (0.625 * 10) = 36.25.
2. Calculate Q1 and Q3:

   For Q1:

   L1 = 10 (lower boundary of the class interval containing Q1)

   N = 32 (total number of data points)

   F1 = 0 (cumulative frequency before Q1)

   f1 = 5 (frequency of the class interval containing Q1)

   C1 = 10 (width of the class interval)

   Q1 = 10 + [(8/4 - 0) / 5] * 10 = 10 + (2/5) * 10 = 14.

   For Q3:

   L3 = 30 (lower boundary of the class interval containing Q3)

   N = 32 (total number of data points)

   F3 = 5 + 8 (cumulative frequency before Q3)

   f3 = 12 (frequency of the class interval containing Q3)

   C3 = 10 (width of the class interval)

   Q3 = 30 + [(3 * 32/4 - 13) / 12] * 10 = 30 + (16/12) * 10 = 42.5.
3. Calculate Quartile Deviation:

   QD = (Q3 - Q1) / 2 = (42.5 - 14) / 2 = 14.25.