**DATA**

Data refers to raw facts, information, or values that are collected, stored, and processed. It can take various forms, such as numbers, text, images, audio, and more. Data becomes meaningful when it is organized, analysed, and interpreted.

Example:

Consider a simple spreadsheet containing information about students in a class. The spreadsheet might have columns such as "Student ID," "Name," "Age," and "Grade." Each row represents a different student, and the cells within each row contain specific pieces of information.

| **Student ID** | **Name** | **Age** | **Grade** |
| --- | --- | --- | --- |
| 001 | John Doe | 18 | A |
| 002 | Jane Smith | 19 | B |
| 003 | Bob Brown | 17 | C |

**Structured Data:**

Data that is highly organized and follows a predefined model or schema.

Characteristics:

* Organized into tables with rows and columns.
* Each data element is defined and has a specific data type.
* Examples include relational databases, spreadsheets, and CSV files.

Example: -

| **ID** | **Name** | **Age** | **Grade** |
| --- | --- | --- | --- |
| 001 | John Doe | 18 | A |
| 002 | Jane Smith | 19 | B |
| 003 | Bob Brown | 17 | C |

**Unstructured Data:**

Data that does not have a predefined data model or is not easily organized into a structured format.

Example:

A paragraph of text, an image file, or a video recording without a specific format.

**Semi-Structured Data:**

Data that has some structure but does not fit neatly into a relational database or a strict schema.

Examples include XML (eXtensible Markup Language) and JSON (JavaScript Object Notation)

{

  "id": 1,

  "name": "John Doe",

  "age": 25,

  "grades": {

    "math": "A",

    "science": "B",

    "history": "A"

  }

}

**Numeric Data:**

* Discrete: Consists of separate values. For example, the number of students in a class (1, 2, 3, ...).
* Continuous: Represents a range of values. For example, the height or weight of a person.

**Text Data (Categorical):**

* Also known as qualitative or nominal data.
* Represents categories or labels. For example, color, gender, or types of fruits.

**Variable**

**Quantitative Variable:**

* Represents quantities or measurable amounts.
* Associated with continuous or discrete variables.
* Example: Income, Age

**Discrete Variable:**

* Represents countable, distinct values.
* Often whole numbers.
* Example: Number of siblings, Number of cars in a parking lot.

**Continuous Variable:**

* Represents measurable quantities that can take any value within a range.
* Can have an infinite number of possible values.
* Example: Height, Weight, Temperature.

**Qualitative Variable:**

* Represents qualities or characteristics.
* Often associated with categorical variables.
* Example: Hair color, Marital status.

**Nominal Variable:**

* A type of categorical variable with categories that do not have a specific order.
* Example: Colours, Types of fruit.

**Ordinal Variable:**

* A type of categorical variable with categories that have a meaningful order or ranking.
* Example: Education level (e.g., High School, Bachelor's, Master's).

**Population:**

The entire group that is the subject of a study and about which the researcher wants to draw conclusions.

Characteristics:

* Represents the complete set of individuals, items, or data points that meet certain criteria.
* Often denoted by the symbol "N."

Example:

If a researcher is interested in studying the average income of all households in a particular city, the population would consist of every household in that city.

**Sample:**

A subset of the population that is selected for the actual study.

Characteristics:

* Typically, more manageable and practical to study than the entire population.
* Findings from the sample are then generalized to make inferences about the population.
* Often denoted by the symbol "n."

Example:

In the study mentioned earlier, it might be impractical and time-consuming to collect data from every household in the city. Instead, the researcher might select a sample of, say, 500 households, gather data from them, and use the findings to make inferences about the average income of all households in the city.

**Some key reasons for using sampling**

**Necessity: -**

When dealing with a large population, it may be impractical or impossible to study every individual or element. Sampling allows researchers to study a representative subset, making it more manageable.

**Practicality: -**

Conducting research on an entire population may be time-consuming, require extensive resources, and be logistically challenging. Sampling allows researchers to gather meaningful insights while efficiently using resources.

**Cost-Effective: -**

Conducting a study on an entire population can be costly. Sampling is often more cost-effective as it reduces the expenses associated with data collection, analysis, and other logistical aspects.

**Management: -**

Managing and coordinating a study involving the entire population may be challenging. Sampling facilitates easier management of the research process.

**Data Cleansing (Data Cleaning):**

**Tasks Involved:**

- Handling missing values.

- Correcting inaccuracies and errors.

- Standardizing formats.

- Removing duplicates.

- Dealing with outliers.

**Missing Data**

Missing data refers to the absence of values or information for certain observations or variables in a dataset.

**Causes of Missing Data:**

**Non-response:**

Participants or entities may choose not to provide certain information when responding to surveys or questionnaires.

**Data Entry Errors:**

Mistakes during data entry or recording can lead to missing values.

**Systematic Issues:**

Issues with data collection systems or instruments can result in missing data.

**Incomplete Surveys:**

Some participants might not complete all parts of a survey, leading to missing responses.

**Measurement Issues:**

Certain variables may be challenging to measure, resulting in missing data for those variables.

**Dealing with Missing Data:**

**Accepting Missing Data:**

In some cases, researchers may choose to accept missing data if the proportion of missing values is small and the messiness is deemed random or inconsequential.

**Deleting Missing Data**

In this approach, observations with any missing values are excluded from the analysis.

**Imputation of Missing Data:**

* + Imputation involves estimating or predicting missing values based on observed data.
  + Mean/Median Imputation: Replace missing values with the mean or median of the observed values for that variable.

**Outlier**

An outlier is an observation or data point that significantly differs from the rest of the data in a dataset. Outliers can skew statistical analyses and distort the interpretation of results, making it important to identify and understand them.

**Key Characteristics of Outliers:**

**Deviation from the Norm:**

Outliers exhibit values that deviate substantially from the typical or expected range of values in a dataset.

**Impact on Statistical Measures:**

Outliers can heavily influence summary statistics such as the mean and standard deviation, leading to misleading representations of central tendency and dispersion.

**Identification:**

Outliers are often identified through statistical methods or visual inspection of graphs; such as box plots or scatter plots.

**Causes of Outliers:**

Outliers can arise from measurement errors, data entry mistakes, natural variability, or genuine extreme observations in the population.

**Ways to Identify Outliers:**

**Visual Inspection:**

Plotting the data using graphs like box plots, scatter plots, or histograms can reveal observations that stand out from the majority.

**Statistical Methods:**

* Z-Score: Identifying data points with z-scores beyond a certain threshold (e.g., |z| > 3) as potential outliers.

*Z = (x-µ)/σ*

* Interquartile Range (IQR): Using the IQR to identify observations outside a defined range.

*IQR = Q3 - Q1*

*LF = Q1 - (1.5\*IQR)*

*UF = Q3 - (1.5\*IQR)*

*Ex:*

*l = [2,5,7,10,12,15,18,20,22,30]*

*Q1 = 7*

*Q3 = 20*

*IQR = 20 - 7 => 13*

*LF = 7-(1.5\*13) => -12.5*

*UF = 20-(1.5\*13) => 39.5*

**Dealing with Outliers:**

**Retaining Outliers:**

* In some cases, it may be appropriate to retain outliers, especially if they represent genuine extreme values in the data.
* Retaining outliers allows for an inclusive analysis, considering the full range of variability in the dataset.

**Removing Outliers:**

* + Removing outliers involves excluding extreme values from the dataset before analysis.
  + Common methods include using statistical criteria (e.g., Z-scores, IQR) to identify and exclude observations beyond a certain threshold.
  + Reduces the impact of extreme values on summary statistics and model results
  + Loss of information: Excluding outliers may discard meaningful data points.

**Transformation:**

* + Transformation involves applying mathematical functions to the data to modify its distribution and reduce the impact of outliers.
  + Common transformations include logarithmic, square root, or Cube root transformations.

**Statistics**

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, presenting, and organizing data. It helps in making informed decisions and drawing conclusions from numerical information.

For example, imagine you want to study the average height of students in a class. You collect the heights of each student and calculate the mean (average) height. This mean height represents a central tendency, providing a summary statistic that gives insight into the overall height distribution within the class.

**Types of Statistics**

**Descriptive Statistics:**

- Descriptive statistics are used to summarize, organize, and describe the main features of a dataset.

- Measures of central tendency (mean, median, mode)

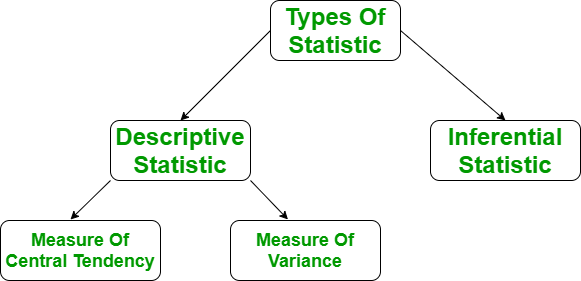
- Measures of dispersion (range, variance, standard deviation)

- Frequency distributions and histograms

- Percentiles and quartiles

**Inferential Statistics:**

inferential statistics involve making inferences or predictions about a population based on a sample of data. It extends findings from a sample to the entire population.



**Understanding the Descriptive Statistics**

In layman’s terms, descriptive statistics generally means describing the data with the help of some representative methods like charts, tables, Excel files, etc. The data is described in such a way that it can express some meaningful information that can also be used to find some future trends.

**There are two types of Descriptive Statistics:**

* 1. The measure of central tendency
  2. Measure of variability

**Measure of Central Tendency**

The measure of central tendency is a single value that attempts to describe the whole set of data. There are three main features of central tendency:

* + Mean
  + Median
  + Mode

**Mean**

It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

The mean () function returns the mean or average of the data passed in its arguments. If the passed argument is empty, Statistics Error is raised.

Mean = Sum/N

Ex:

L = [15,3,12,0,24,3]

mean = (15+3+12+0+24+3)/6 =>9.5

**Median**

It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd, then the centre element is the median and if it is even then the median would be the average of two central elements. it first sorts the data i=and then performs the median operation

Ex:

L = [15,3,12,0,24,3]

median = (12+0)/2 => 6

**Mode**

It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

Ex:

L = [15,3,12,0,24,3]

median = 3

**Measure of Variability**

Till now, we have studied the measure of central tendency but this alone is not sufficient to describe the data. To overcome this, we need the measure of variability. The measure of variability is known as the spread of data or how well our data is distributed. The most common variability measures are:

* + Range
  + Variance
  + Standard deviation

**Range**

The difference between the largest and smallest data point in our data set is known as the range. The range is directly proportional to the spread of data which means the bigger the range, the more the spread of data and vice versa.

Range = Largest data value – smallest data value

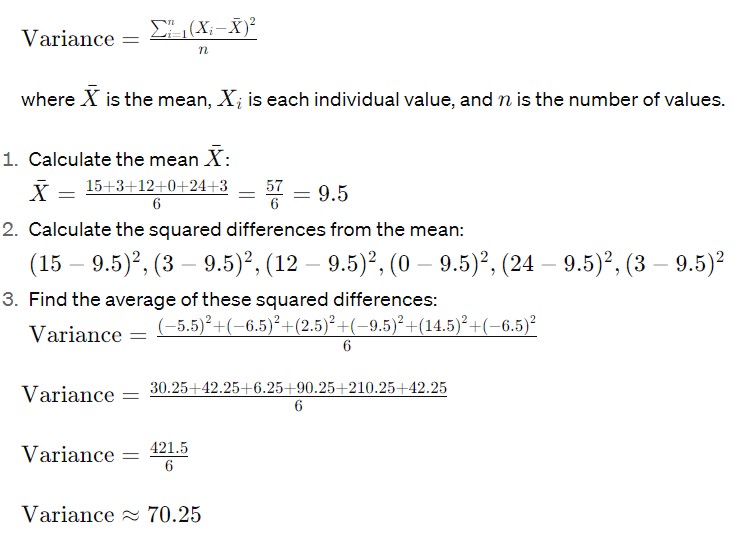
Ex:

L = [15,3,12,0,24,3]

Range=max(L)−min(L)=24−0=24

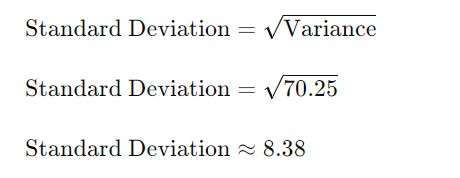
**Variance**

It is defined as an average squared deviation from the mean. It is calculated by finding the difference between every data point and the average which is also known as the mean, squaring them, adding all of them, and then dividing by the number of data points present in our data set.



**Standard Deviation**

It is defined as the square root of the variance. It is calculated by finding the Mean, then subtracting each number from the Mean which is also known as the average, and squaring the result. Adding all the values and then dividing by the no of terms followed by the square root.

****

**Distribution**

a distribution refers to the pattern or shape that the values in a dataset take. It describes how the different values are spread or distributed across the range of possible values. Understanding the distribution of a dataset is fundamental for statistical analysis

**Probability Distribution:**

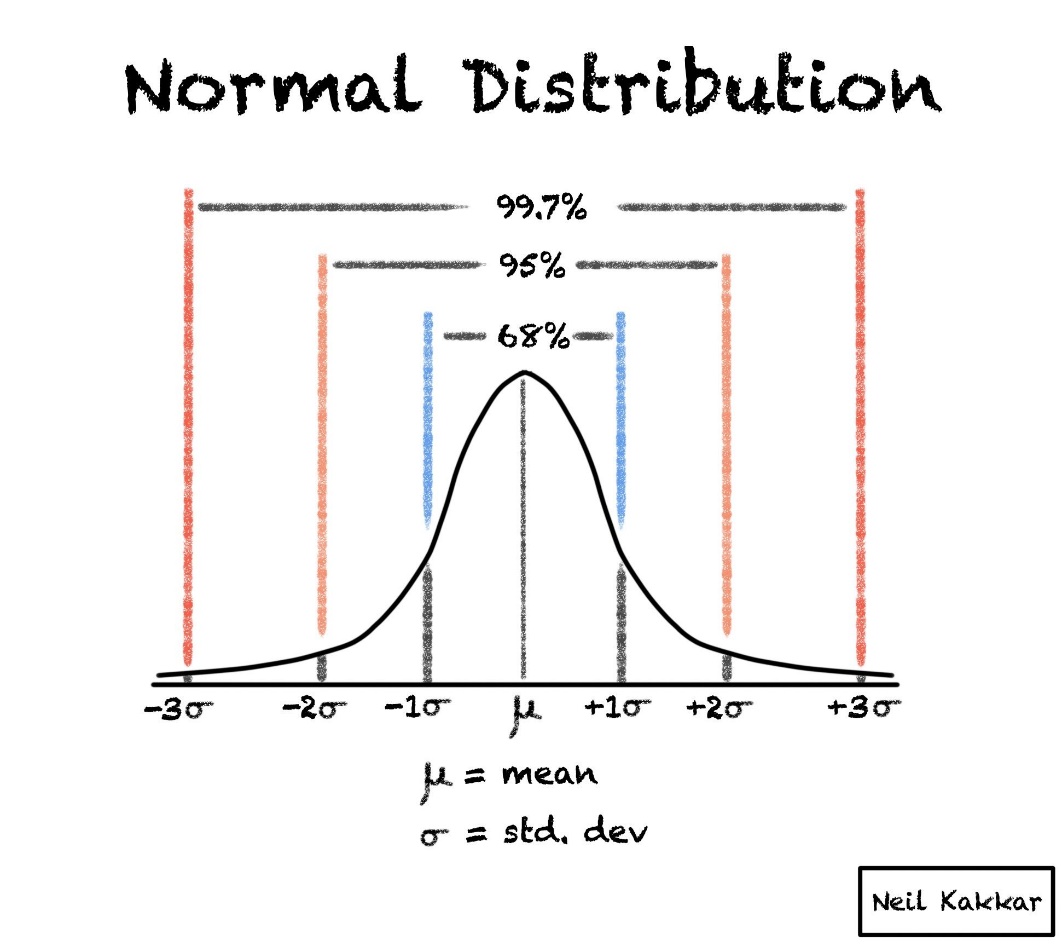
A probability distribution is a mathematical function that describes the likelihood of obtaining the possible values that a random variable can take. It associates each possible outcome with a probability, indicating the likelihood of that outcome occurring.

Ex: -

Consider the probability distribution of the number of heads (successes) when flipping a fair coin five times.

**Normal Distribution:**

* + The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric and bell-shaped.
  + The normal distribution is symmetric around its mean (average), which is located at the centre of the distribution. This means that the probabilities of observing values on the left and right sides of the mean are equal.
  + The normal distribution is completely described by two parameters: the mean (μ) and the standard deviation (σ). The mean determines the centre of the distribution, and the standard deviation controls the spread or variability of the distribution.
  + Approximately 68% of the data falls within one standard deviation of the mean (±μ±σ).
  + Approximately 95% falls within two standard deviations of the mean (±2μ±2σ).
  + Approximately 99.7% falls within three standard deviations of the mean (±3μ±3σ).



**Standard normal distribution**

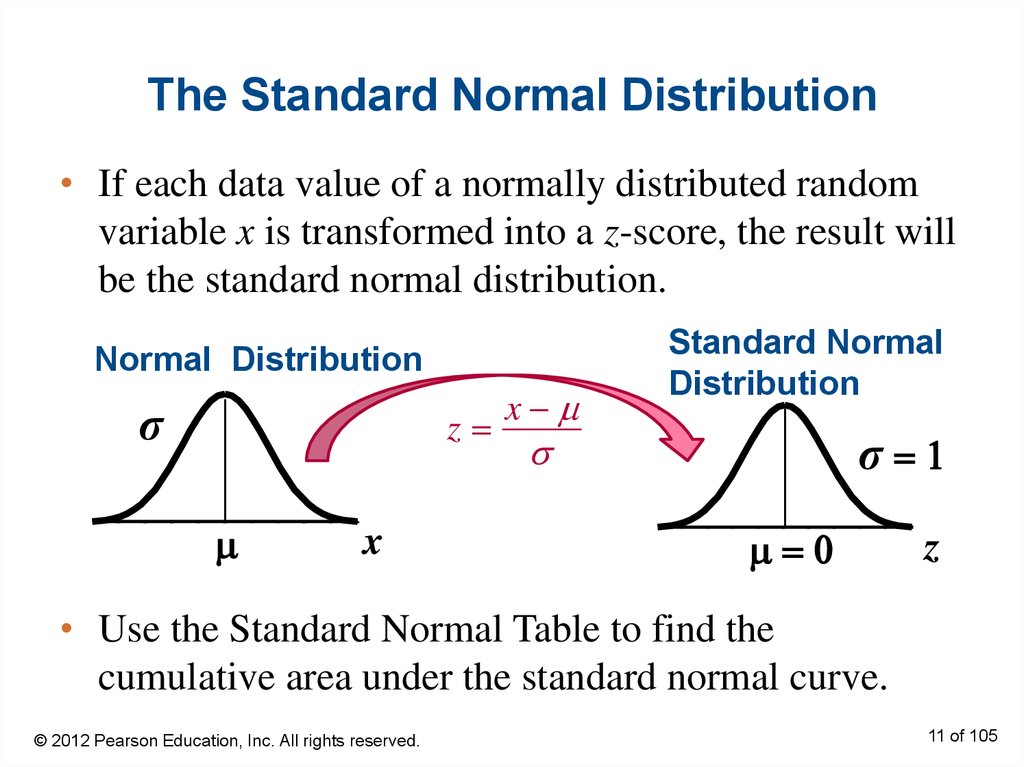
The standard normal distribution, also known as the z-distribution or the standard Gaussian distribution, is a special case of the normal distribution where the mean (μ) is 0 and the standard deviation (σ) is 1

Properties of the standard normal distribution include:

- Mean (μ): 0

- Standard Deviation (σ): 1

- Symmetry: Like the general normal distribution, the standard normal distribution is symmetric around 0.



**Probability density function**

The Probability Density Function (PDF) is a concept used in probability theory and statistics to describe the likelihood of a continuous random variable falling within a particular range of values. The PDF is associated with continuous probability distributions

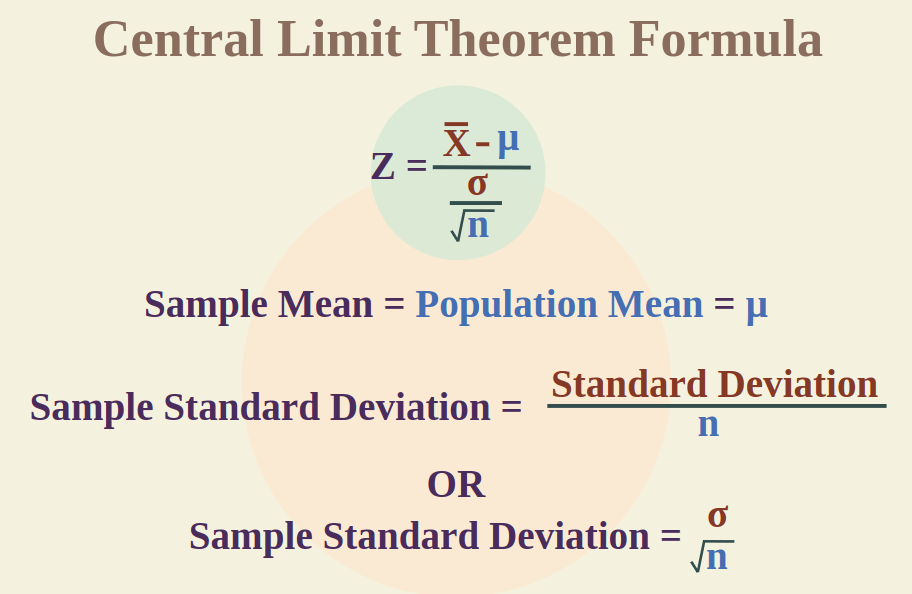
Probability within an interval: The probability that X falls within a specific interval [a,b] is given by the integral of the PDF over that interval.

**Inferential Statistics**

**Central Limit Theorem**

Central Limit Theorem explains that the sample distribution of the sample mean resembles the normal distribution irrespective of the fact that whether the variables themselves are distributed normally or not. Central Limit Theorem is often called CLT in abbreviated form.

Central Limit Theorem states that when large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.



Ex: -

Imagine you have a school with a large number of students. The heights of individual students may not follow a perfect normal distribution. Some students might be shorter, some taller, and the distribution of heights may not look like a bell curve.

Now, let's apply the Central Limit Theorem:

Original Distribution:

The individual heights of students may have a varied distribution; it might not follow a normal pattern.

Sample Means:

Now, randomly select groups of students (samples) from the school and calculate the average height in each group. For example, you might take a sample of 30 students and find the average height of those 30.

Distribution of Sample Means:

According to the Central Limit Theorem, as you collect more and more of these average heights from different samples, the distribution of these sample means will start to resemble a bell curve, even if the individual heights themselves don't follow a normal distribution.

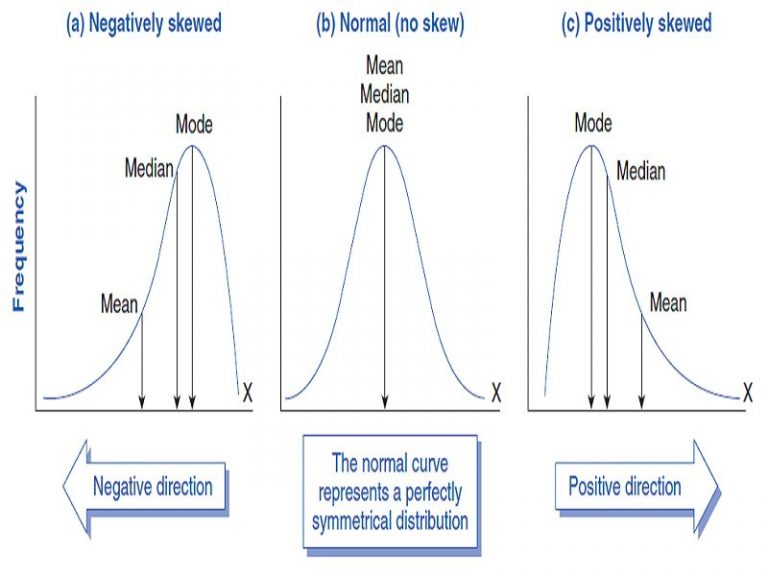
**Assumptions of Statistical test**

**Skew**

skewness is a measure of the asymmetry of a probability distribution. It quantifies the extent and direction of skew (departure from horizontal symmetry) in a dataset.

There are three main types of skewness:

1. **Positive skewness (right-skewed):** The right tail (larger values) of the distribution is longer or fatter than the left tail. In other words, the majority of the data points are concentrated on the left side, and the right side has some extreme values.
2. **Negative skewness (left-skewed):** The left tail (smaller values) of the distribution is longer or fatter than the right tail. This indicates that the majority of the data points are concentrated on the right side, and the left side has some extreme values.
3. **Zero skewness:** A perfectly symmetric distribution has a skewness of zero. This means that the right and left sides of the distribution are mirror images of each other.



Skewness is often measured using different formulas, with one common method being the Pearson's skewness coefficient. It is calculated as:

Skewness=3(*x*ˉ−*μ*)/σ​

where:

* ˉ*x*ˉ is the sample mean,
* *μ* is the population mean,
* *σ* is the standard deviation.

Skewness provides valuable information about the shape of a distribution and can help identify potential outliers

**Kurtosis**

Kurtosis is a statistical measure that describes the distribution of data in terms of the tails and the peakedness (or flatness) relative to a normal distribution. It provides information about the shape and thickness of the tails of a probability distribution.

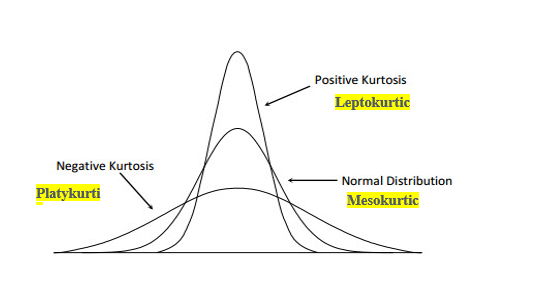
There are three main types of kurtosis:

1. **Mesokurtic (Normal kurtosis):** A distribution with kurtosis equal to 3 (or 0 excess kurtosis) is considered mesokurtic. This means that the distribution has similar tail behavior to a normal distribution.
2. **Leptokurtic (Excess kurtosis > 0):** A distribution with positive excess kurtosis has heavier tails and a more peaked central portion compared to a normal distribution. This indicates the presence of more extreme values.
3. **Platykurtic (Excess kurtosis < 0):** A distribution with negative excess kurtosis has lighter tails and a flatter central portion compared to a normal distribution. This suggests fewer extreme values than a normal distribution.

The excess kurtosis is often used in practice, and it is calculated as:

Excess Kurtosis=Kurtosis−3

Kurtosis is affected by the presence of outliers in a dataset. High kurtosis can indicate the presence of heavy tails and potentially more extreme values, while low kurtosis can suggest lighter tails and a lack of extreme values.



**Correlation**

Correlation is the statistical analysis of the relationship or dependency between two variables. Correlation allows us to study both the strength and direction of the relationship between two sets of variables.

Change in variable Couse change in another variable

**Types**

**+ ve Correlation**

It is a relation b/w two variable where if one variable increases, the other variable also increases & vice-verse

Ex: - No of tree cut down increases, the probability of erosion Increases

**- ve Correlation**

Inverse relationship b/w Variable, where one variable decreases other variable increases.

ex: - Car speed decreases, time taken to reach destination increases.

It always takes on a value between -1 and 1 where:

* + -1 indicates a perfectly negative linear correlation between two variables
  + 0 indicates no linear correlation between two variables
  + 1 indicates a perfectly positive linear correlation between two variables

**Correlation coefficients**

The correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables.

The most commonly used correlation coefficient is Pearson's correlation coefficient, denoted by ***r***.

**Coefficient of determination**

The coefficient of determination, often denoted as *R*2 (pronounced "R-squared"), is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

he *R*2 value ranges between 0 and 1, or it can be expressed as a percentage between 0% and 100%. The interpretation is as follows:

* *R*2=0: The model does not explain any of the variability in the dependent variable.
* *R*2=1: The model explains all of the variability in the dependent variable.
* 0<*R*2<1: The model explains a proportion of the variability in the dependent variable.

it is calculated as the ratio of the explained sum of squares to the total sum of squares. The formula for *R*2 is:

*R2 = Explained Sum of Squares (ESS)/ Total Sum of Squares (TSS)*

**Machine learning**

Machine learning is a field of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform tasks without being explicitly programmed for those tasks. The primary goal of machine learning is to create systems that can learn from data and make predictions or decisions based on that learning.

1. **Supervised Learning:** In supervised learning, the algorithm is trained on a labelled dataset, which means that the input data is paired with corresponding output labels. The algorithm learns the mapping from inputs to outputs, and once trained, it can make predictions on new, unseen data.
2. **Unsupervised Learning:** Unsupervised learning deals with unlabelled data, where the algorithm tries to find patterns, relationships, or structures in the data without explicit guidance on the output. Clustering and dimensionality reduction are common tasks in unsupervised learning.

**Supervised Learning:**

The main goal of supervised learning is to build a predictive model that can accurately generalize to new, unseen data. There are two primary types of supervised learning tasks:

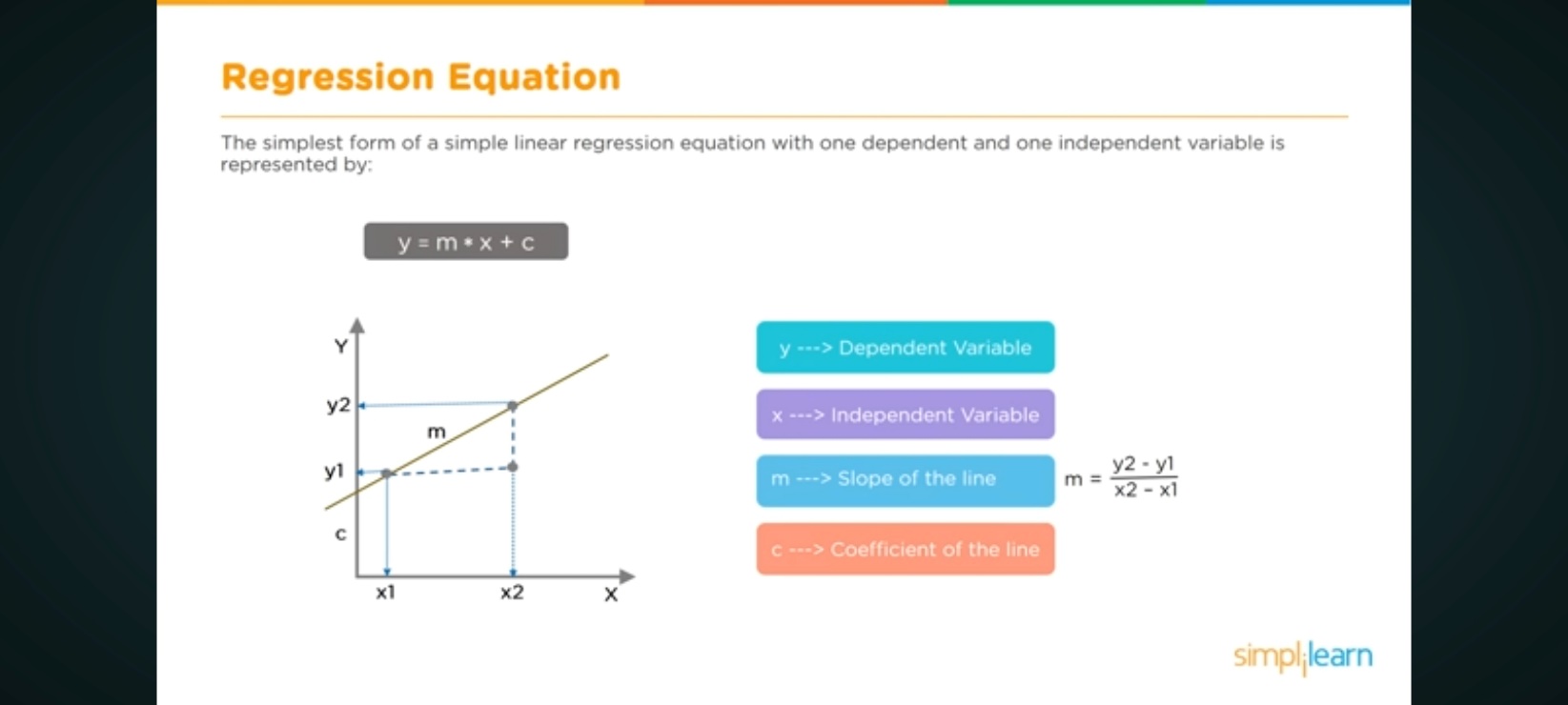
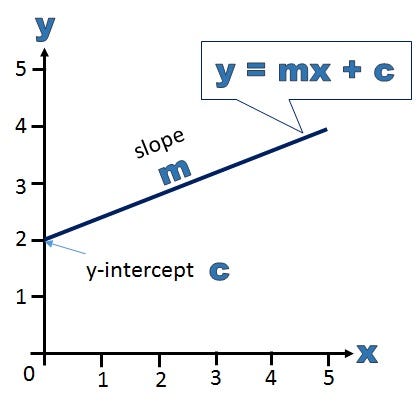
1. **Regression:** In regression tasks, the goal is to predict a continuous output or target variable. Examples include predicting house prices, temperature, or stock prices.
2. **Classification:** In classification tasks, the goal is to predict a categorical output or label. Examples include spam detection, image classification, and sentiment analysis.

**Regression**

It involves modelling the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors or features). The primary goal of regression is to understand and quantify the relationship between variables and to make predictions based on that understanding.

**Simple Linear Regression**

* Simple Linear Regression is a statistical method used to model the relationship between a single independent variable (predictor) and a dependent variable (response).
* The relationship is assumed to be linear, meaning it can be represented by a straight line.
* The purpose of simple linear regression is to find the best-fitting line that minimizes the difference between the predicted and observed values of the dependent variable.



The simple linear regression model is represented by the equation:

y=mx + c

where:

* *y* is the dependent variable.
* *x* is the independent variable.
* *m* is the slope of the line, representing the rate of change in *y* for a one-unit change in *x*.
* *c* is the y-intercept, which is the value of *y* when *x* is 0.

The goal in simple linear regression is to estimate the values of *m* and *c* that minimize the sum of squared differences between the observed values (*y*) and the values predicted by the line (*mx*+*c*).

The formula for the slope (*m*) is given by:

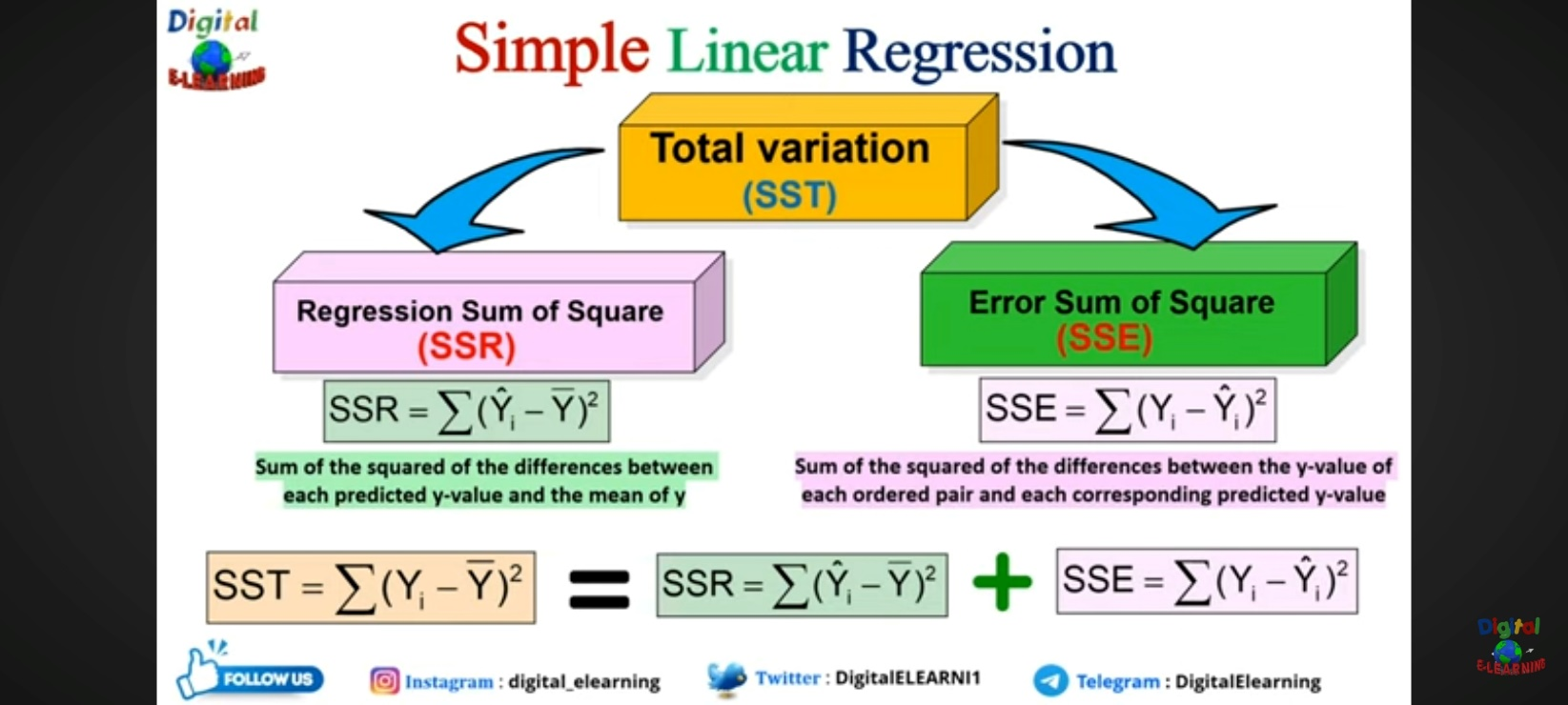
m=n(∑x\*y) −(∑x) (∑y) ​/n(∑x2) −(∑x)2

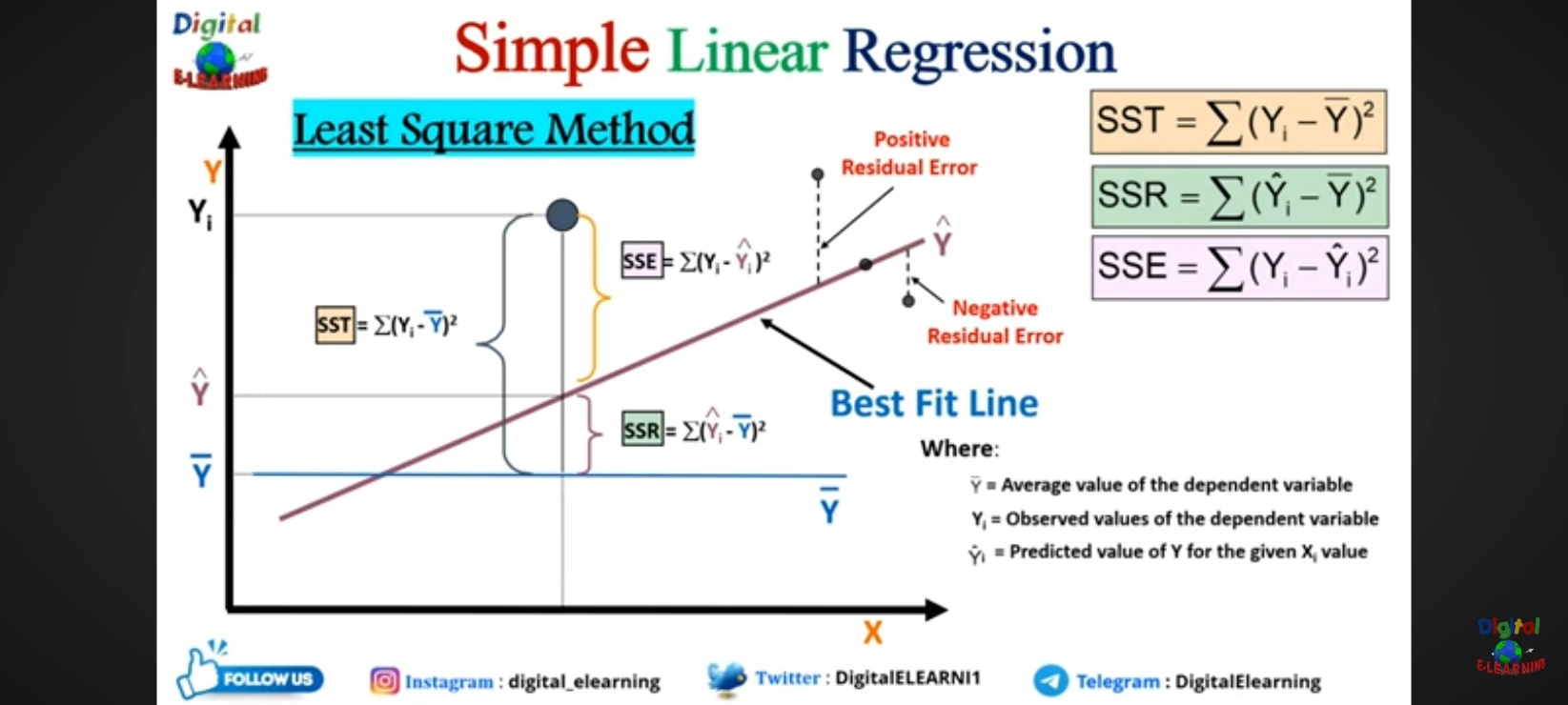
where:

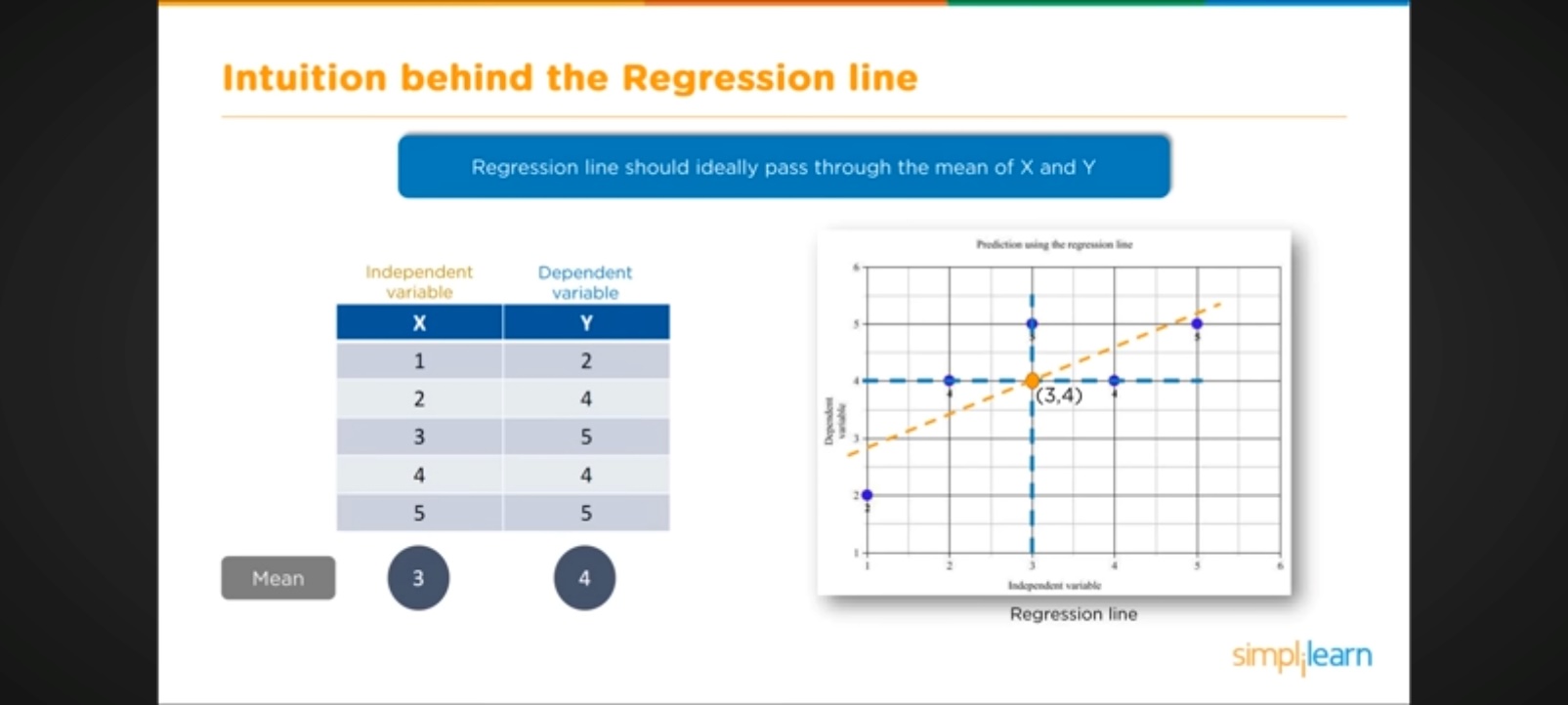
* *n* is the number of data points.
* ∑ denotes summation.
* *x* and *y* are the respective means of the independent and dependent variables.

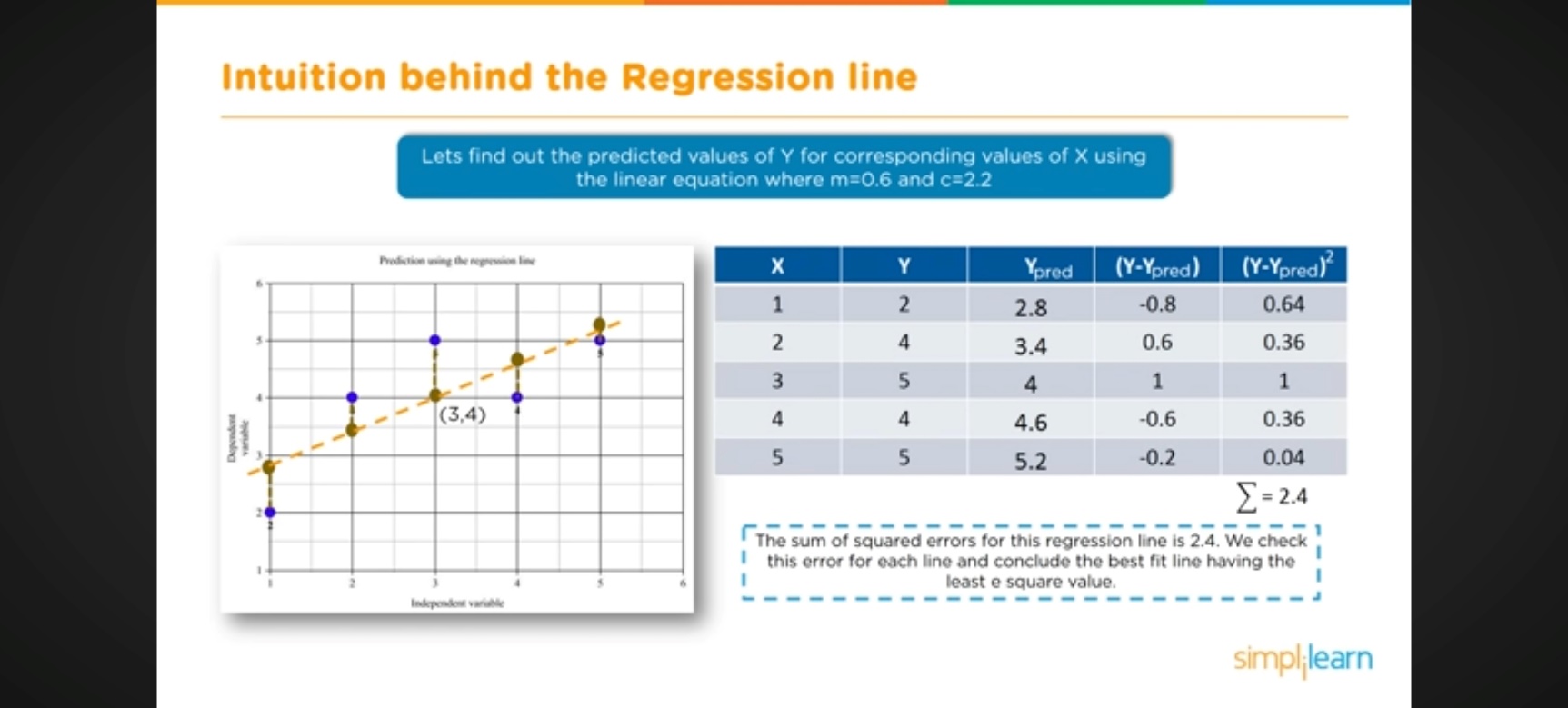
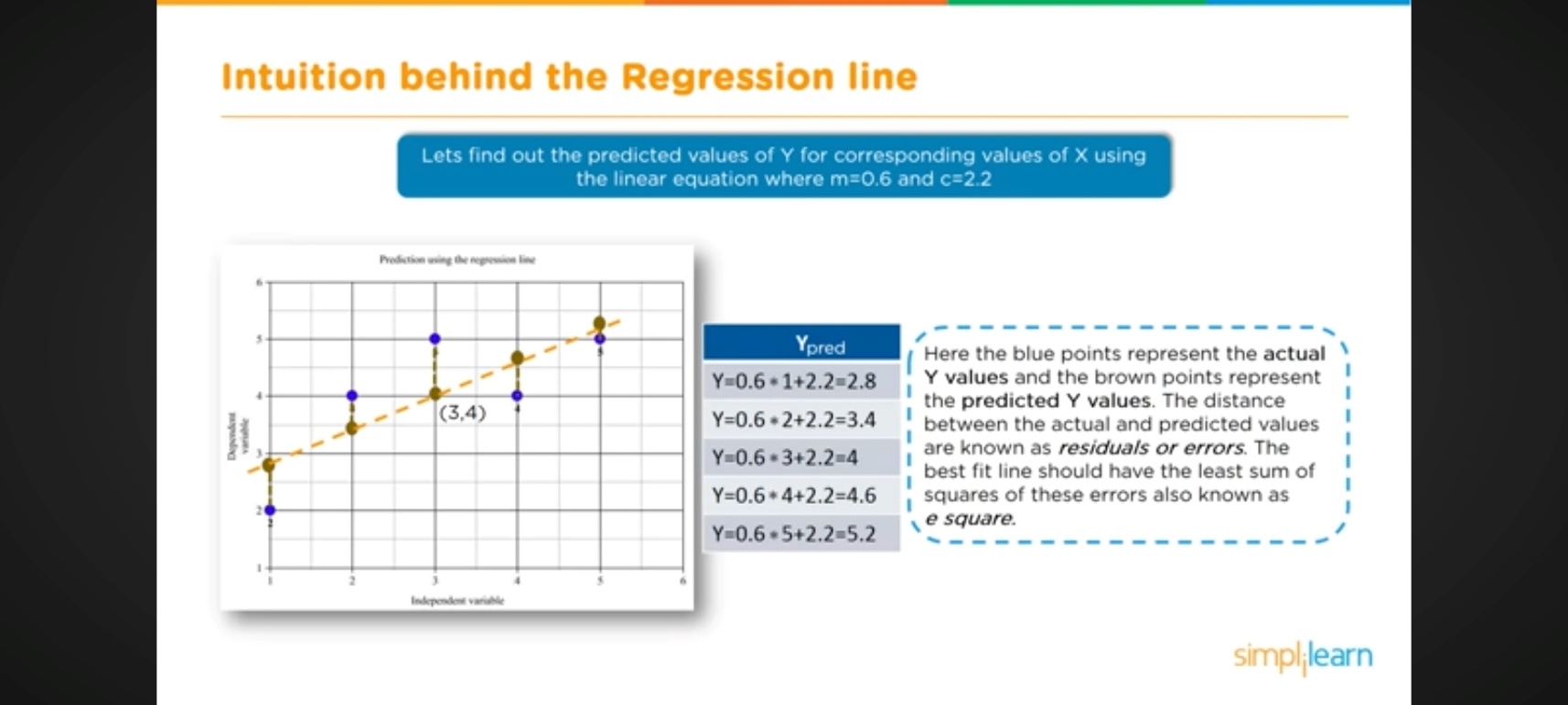
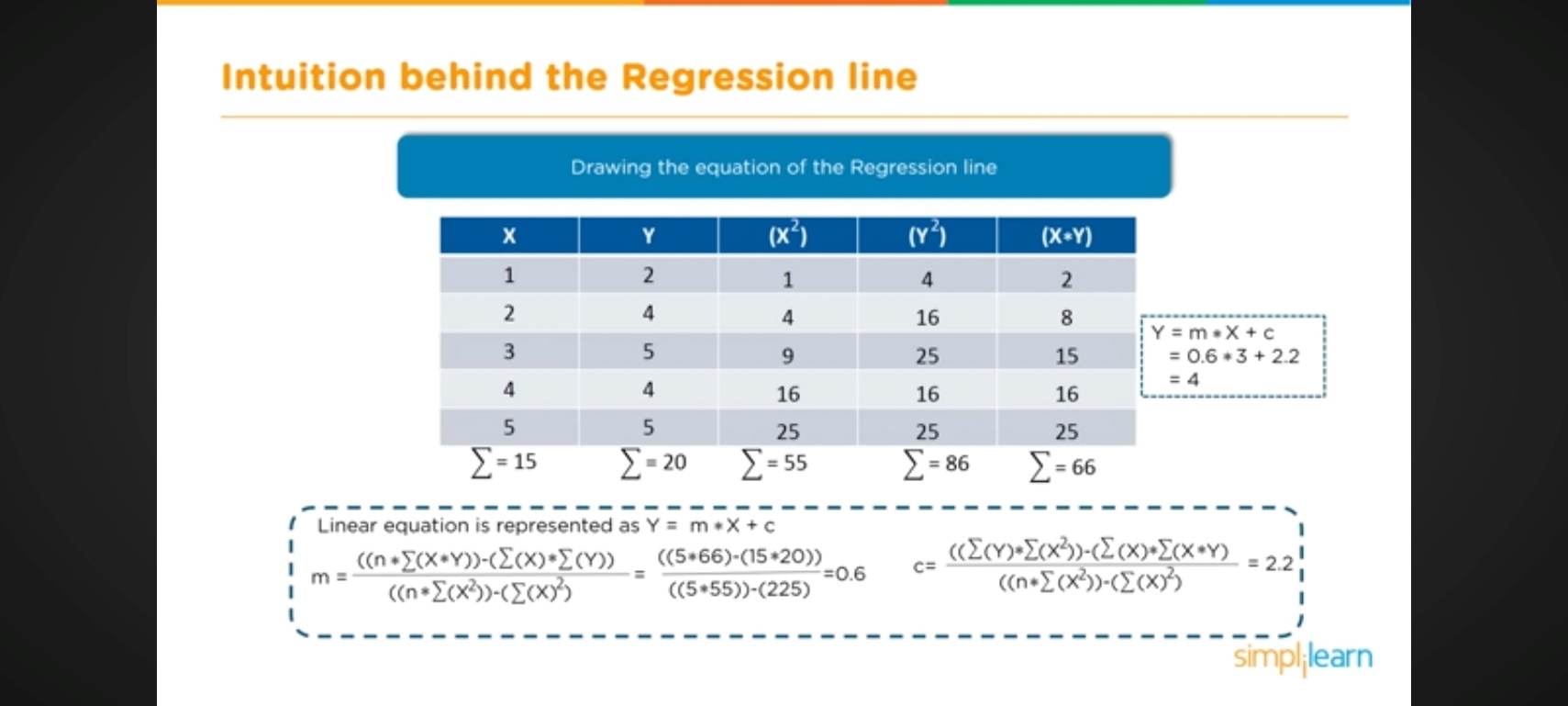
The formula for the y-intercept (*c*) is given by:

C = (∑x) (∑y) ​/ n









**Polynomial Regression**

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modelled as an *nth-degree* polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y.

**What is a Polynomial Regression?**

* There are some relationships that a researcher will hypothesize is curvilinear. Clearly, such types of cases will include a polynomial term.
* Inspection of residuals. If we try to fit a linear model to curved data, a scatter plot of residuals (Y-axis) on the predictor (X-axis) will have patches of many positive residuals in the middle. Hence in such a situation, it is not appropriate.
* An assumption in the usual multiple linear regression analysis is that all the independent variables are independent.

**Why Polynomial Regression?**

Polynomial regression is a type of regression analysis used in statistics and machine learning when the relationship between the independent variable (input) and the dependent variable (output) is not linear. While simple linear regression models the relationship as a straight line, polynomial regression allows for more flexibility by fitting a polynomial equation to the data.

When the relationship between the variables is better represented by a curve rather than a straight line, polynomial regression can capture the non-linear patterns in the data.

When the relationship is non-linear, a polynomial regression model introduces higher-degree polynomial terms.

The general form of a polynomial regression equation of degree n is:

Y = a+b1x+b2x2+b3x3+b4x4 ………. *n x*n+

Where :

* y is the dependent variable.
* x is the independent variable.
* ​ 0,1, ……., *n* are the coefficients of the polynomial terms.
* n is the degree of the polynomial.
* represents the error term.

1. The choice of the polynomial degree (n) is a crucial aspect of polynomial regression. A higher degree allows the model to fit the training data more closely, but it may also lead to overfitting, especially if the degree is too high. Therefore, the degree should be chosen based on the complexity of the underlying relationship in the data.
2. The polynomial regression model is trained to**find the coefficients** that minimize the difference between the predicted values and the actual values in the training data.
3. Once the model is trained, it can be used to make predictions on new, unseen data.

**Overfitting Vs Under-fitting**

While dealing with the polynomial regression one thing that we face is the problem of [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/) this happens because while we increase the order of the polynomial regression to achieve better and better performance model gets overfit on the data and does not perform on the new data points.

**Application of Polynomial Regression**

* The growth rate of tissues.
* Progression of disease epidemics
* Distribution of carbon isotopes in lake sediments

**Disadvantages of using Polynomial Regression**

* These are too sensitive to outliers.
* The presence of one or two [outliers](https://www.geeksforgeeks.org/machine-learning-outlier/) in the data can seriously affect the results of nonlinear analysis.
* In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

**Multivariate Regression**

Multivariate regression refers to the statistical technique that establishes a relationship between multiple data variables. It estimates a linear equation that facilitates the analysis of multiple dependent or outcome variables depending on one or more predictor variables at different points in time.

**Assumptions**

The validity and reliability of the multivariate regression findings depend upon the following four assumptions:

1. **Linearity**: The correlation between the predictor and outcome variables is linear.
2. **Independence**: The observations are autonomous of each other, i.e., the value of the other independent variable should not influence the value of the independent variables.
3. **Homoscedasticity**: The [**variance**](https://www.wallstreetmojo.com/variance/) of the errors (residuals) is even across all levels of the explanatory variables. This ensures that the spread of residuals is the same for all predicted values.
4. **Normality**: The residuals (differences between observed and predicted values) should be normally distributed, ensuring that statistical inferences about [**regression**](https://www.wallstreetmojo.com/regression/) coefficients are valid.

**Formula**

The multivariate regression equation is represented as follows:

***Y = β0 + β1X1+ βkXk + residual***

* Y represents the dependent variable.
* β0​ is the intercept.
* β1​,β2​,…,βk​ are the coefficients for the respective independent variables X1​,X2​,…,Xk​.
* Residual represents the error term

**Advantages And Disadvantages**

The researchers need to consider the following pros and cons before conducting the multivariate regression analysis:

**Advantages**

Some of the benefits of this model are discussed below:

* **Better Comprehends Relationships**: Unlike simple linear regression, which considers only one predictor, multivariate regression can account for interactions and interdependencies among various predictors, capturing complex relationships between these variables.
* **Reliable Predictions**: By including multiple predictors, the model might provide more accurate estimations than simple regression models, leading to a better fit for the data.
* **Correlation, Strength, and Direction**: Multivariate regression can help identify which explanatory variables significantly influence the dependent variable, establishing a correlation and quantifying the direction and strength of these correlations.

**Disadvantages**

The various limitations of this regression technique are as follows:

* **Difficult to Interpret**: Multivariate regression can be challenging to interpret, especially for individuals unfamiliar with [**statistical analyses**](https://www.wallstreetmojo.com/statistical-analysis/), due to multiple predictors.
* **Complex Calculations**: Since this model incorporates multiple variables, its computation involves complex mathematical calculations.
* **Extensive Data Requirement**: Multivariate regression requires a larger [**sample size**](https://www.wallstreetmojo.com/sample-size-formula/) than simple regression. Small sample sizes can result in unreliable parameter estimates and low statistical power.
* **Overfitting**: It occurs when the model fits the training data too closely, capturing noise rather than the underlying pattern.

**Multivariate Regression vs Multiple Regression**

Multivariate regression and [**multiple regression**](https://www.wallstreetmojo.com/multiple-regression-formula/) are terms that are often used interchangeably, but they are two different methods, as discussed below:

| **Basis** | **Multivariate Regression** | **Multiple Regression** |
| --- | --- | --- |
| Meaning | A regression analysis method involving multiple variables, both dependent and independent variables | A statistical measure that facilitates the analysis of the association between a dependent variable and two or more independent variables |
| Purpose | It determines how a set of predictors impacts multiple related outcome variables. | It is used to understand how multiple predictors influence a single outcome variable. |
| Number of Dependent Variables | Several | Single |
| Outcome | Provides insights into how a set of predictors influences multiple related outcomes simultaneously, revealing patterns among the dependent variables | Identifies individual effects of predictors on the dependent variable while controlling for other predictors |

**Decision Tree**

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

**Decision Tree Terminologies**

Some of the common Terminologies used in Decision Trees are as follows:

* **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
* Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
* **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
* **Splitting:**The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
* **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
* **Parent Node:** The node that divides into one or more child nodes.
* **Child Node:**The nodes that emerge when a parent node is split.
* **Impurity**: A measurement of the target variable’s homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The **Gini index** and **entropy** are two commonly used impurity measurements in decision trees for classifications task
* **Variance**: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. **Mean squared error, Mean Absolute Error, friedman\_mse, or Half Poisson deviance** are used to measure the variance for the regression tasks in the decision tree.
* **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
* **Pruning**: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.

**Entropy:**

Entropy is the measure of the degree of randomness or uncertainty in the dataset. In the case of classifications, It measures the randomness based on the distribution of class labels in the dataset.

**Important points related to Entropy:**

1. The entropy is 0 when the dataset is completely homogeneous, meaning that each instance belongs to the same class. It is the lowest entropy indicating no uncertainty in the dataset sample.
2. when the dataset is equally divided between multiple classes, the entropy is at its maximum value. Therefore, entropy is highest when the distribution of class labels is even, indicating maximum uncertainty in the dataset sample.
3. Entropy is used to evaluate the quality of a split. The goal of entropy is to select the attribute that minimizes the entropy of the resulting subsets, by splitting the dataset into more homogeneous subsets with respect to the class labels.
4. The highest information gain attribute is chosen as the splitting criterion (i.e., the reduction in entropy after splitting on that attribute), and the process is repeated recursively to build the decision tree.

**Gini Impurity or index:**

Gini Impurity is a score that evaluates how accurate a split is among the classified groups. The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible. Gini Index is the evaluation metric we shall use to evaluate our Decision Tree Model.

**Information Gain:**

Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

**Advantages of the Decision Tree:**

1. It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
2. It can be very useful for solving decision-related problems.
3. It helps to think about all the possible outcomes for a problem.
4. There is less requirement of data cleaning compared to other algorithms.

**Disadvantages of the Decision Tree:**

1. The decision tree contains lots of layers, which makes it complex.
2. It may have an overfitting issue, which can be resolved using the Random Forest algorithm.
3. For more class labels, the computational complexity of the decision tree may increase.

**K-means Clustering**

[Unsupervised Machine Learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/)is the process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to operate on that data without supervision. Without any previous data training, the machine’s job in this case is to organize unsorted data according to parallels, patterns, and variations.

K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster. After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster. In the analysis we assume that number of cluster is given in advanced and we have to put points in one of the group.

In some cases, K is not clearly defined, and we have to think about the optimal number of K. K Means clustering performs best data is well separated. When data points overlapped this clustering is not suitable. K Means is faster as compare to other clustering technique. It provides strong coupling between the data points. K Means cluster do not provide clear information regarding the quality of clusters. Different initial assignment of cluster centroid may lead to different clusters. Also, K Means algorithm is sensitive to noise. It maymhave stuck in local minima.

**objective**

The goal of [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) is to divide the population or[set](https://www.geeksforgeeks.org/set-in-cpp-stl/) of data points into a number of groups so that the data points within each group are more[comparable](https://www.geeksforgeeks.org/comparable-vs-comparator-in-java/) to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

The algorithm works as follows:

1. First, we randomly initialize k points, called means or cluster centroids.
2. We categorize each item to its closest mean, and we update the mean’s coordinates, which are the averages of the items categorized in that cluster so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

**Advantages of K-Means Clustering:**

1. **Simplicity and Ease of Implementation**:
   * K-Means is relatively straightforward to understand and implement.
   * It’s a great starting point for beginners exploring clustering algorithms.
2. **Scalability**:
   * K-Means can handle large datasets efficiently.
   * Its computational complexity scales linearly with the number of data points.
3. **Convergence Guarantee**:
   * K-Means guarantees convergence to a local minimum.
   * The algorithm iteratively refines cluster centroids until they stabilize.
4. **Warm-Start Capability**:
   * You can initialize K-Means with pre-defined centroid positions (warm-start).
   * This helps speed up convergence, especially when dealing with multiple runs.
5. **Adaptability to New Data**:
   * K-Means can easily incorporate new examples without retraining the entire model.
   * When new data points arrive, you can update the centroids efficiently.
6. **Generalization to Different Cluster Shapes and Sizes**:
   * K-Means can handle clusters of various shapes (e.g., elliptical clusters) and sizes.
   * By modifying K-Means, you can adapt it to naturally imbalanced clusters.

**Disadvantages of K-Means Clustering:**

1. **Manual Selection of K**:
   * Choosing the optimal number of clusters (K) is challenging.
   * You need to decide K beforehand, which can be subjective.
2. **Sensitivity to Initial Values**:
   * K-Means is sensitive to the initial placement of centroids.
   * Running K-Means multiple times with different initializations can mitigate this issue.
3. **Handling Varying Cluster Sizes and Densities**:
   * K-Means struggles when clusters have different sizes and densities.
   * Generalizing K-Means can address this limitation.
4. **Clustering Outliers**:
   * Outliers can significantly impact K-Means.
   * Centroids may be dragged by outliers or form separate clusters.
5. **Dimensionality Scaling**:
   * As the number of dimensions increases, K-Means becomes less effective.
   * Consider dimensionality reduction techniques like PCA or spectral clustering.

Euclidean distance between two points is the **length of the straight line segment** that directly connects those points.

