Statistics
for
Data Science and Data Analysis

As it tells the health of Data.

Why Statistics is Important?

- Statistics enables us to make **informed decisions** and **intelligent judgments** despite the presence of **uncertainty and variation.**
- The discipline offers **powerful methods** for gaining insights across a wide range of fields, such as **business, medicine, agriculture, social sciences, and engineering.**
- Various aspects to study it in context of DATA are:

➢**Data Interpretation**:

✓ Statistics helps analysts make sense of complex data by providing methods to summarize, analyze, and visualize patterns.

✓ Without statistical methods, raw data would be difficult to interpret meaningfully.

- **Decision-Making**:

- ✓ Businesses rely on data-driven decisions to optimize processes, reduce risks, and improve outcomes.

- ✓ Statistical techniques like hypothesis testing, confidence intervals, and regression analysis enable decision-makers to make informed, evidence-based decisions rather than relying on intuition.

- **Predictive Analytics**:

- ✓ Statistical models, such as linear regression, time series analysis, and machine learning algorithms, are essential for predicting future trends and behaviors.

- ✓ These forecasts are critical for planning and resource allocation.

- **Data Quality and Reliability**:

- ✓ Statistics provides tools to assess the quality of data, identify outliers, and ensure the reliability of findings.

- ✓ It helps in determining whether the results are statistically significant or just due to random chance.

- **Optimization and Performance Improvement**:

✓Statistical techniques like A/B testing, experimental design, and optimization models help businesses find the best strategies, products, or marketing tactics.

✓This leads to improved performance, cost reductions, and better customer satisfaction.

- **Risk Management**:

✓Statistical tools are used to measure and quantify risks in various business scenarios, enabling businesses to plan for uncertainties and mitigate potential losses.

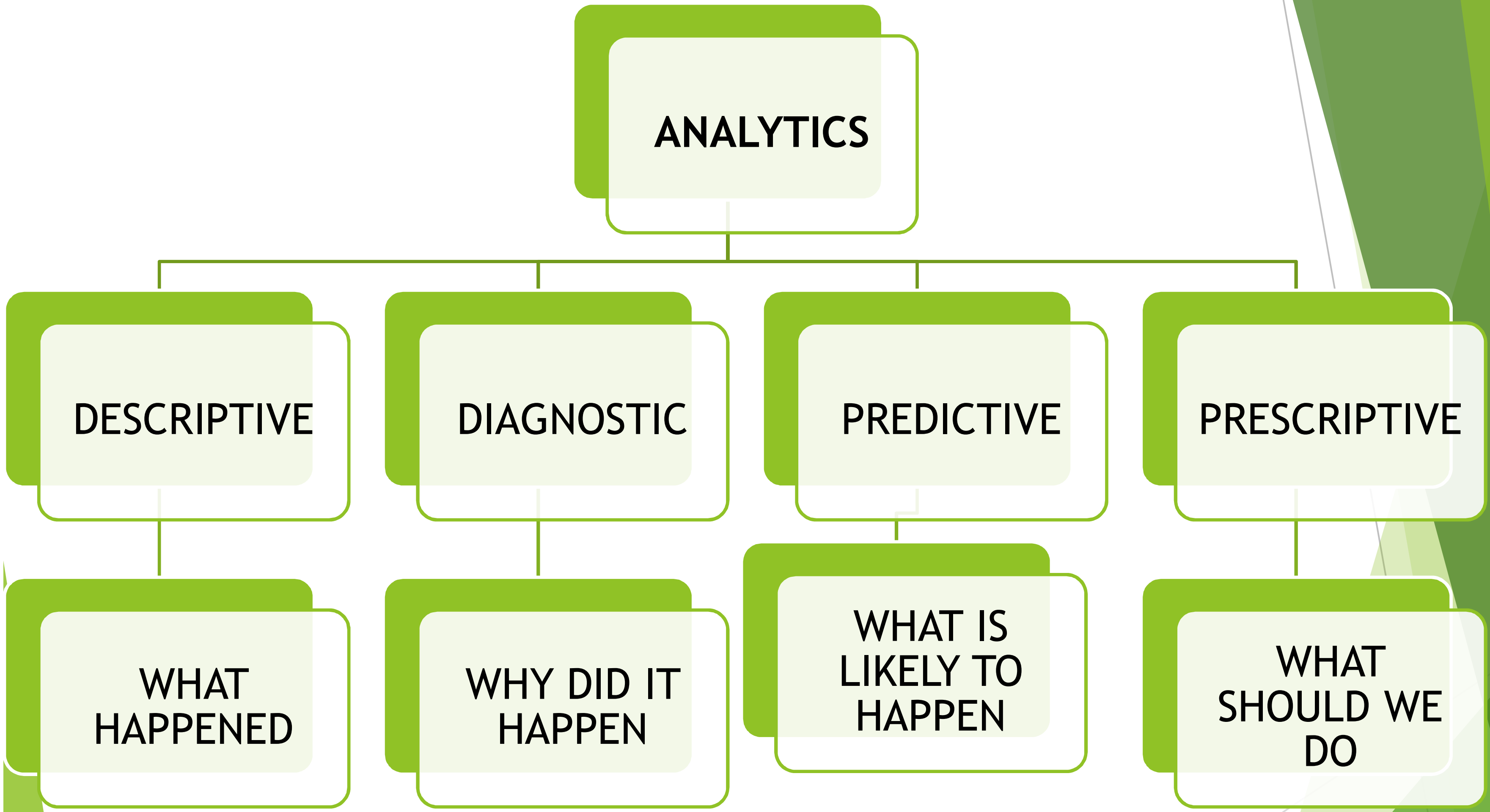
- **Sampling and Inference**:

✓In situations where it is impractical to collect data from the entire population, statistics provides techniques like sampling and estimation to make accurate inferences from smaller datasets, helping businesses draw conclusions about larger populations.

- ➢ **Continuous Improvement**:

- ✓ Statistics is integral in tracking business performance over time.

- ✓ It allows businesses to monitor key performance indicators (KPIs), identify trends, and implement continuous improvements based on data-driven insights.



ANALYTICS

DESCRIPTIVE

WHAT HAPPENED

DIAGNOSTIC

WHY DID IT HAPPEN

PREDICTIVE

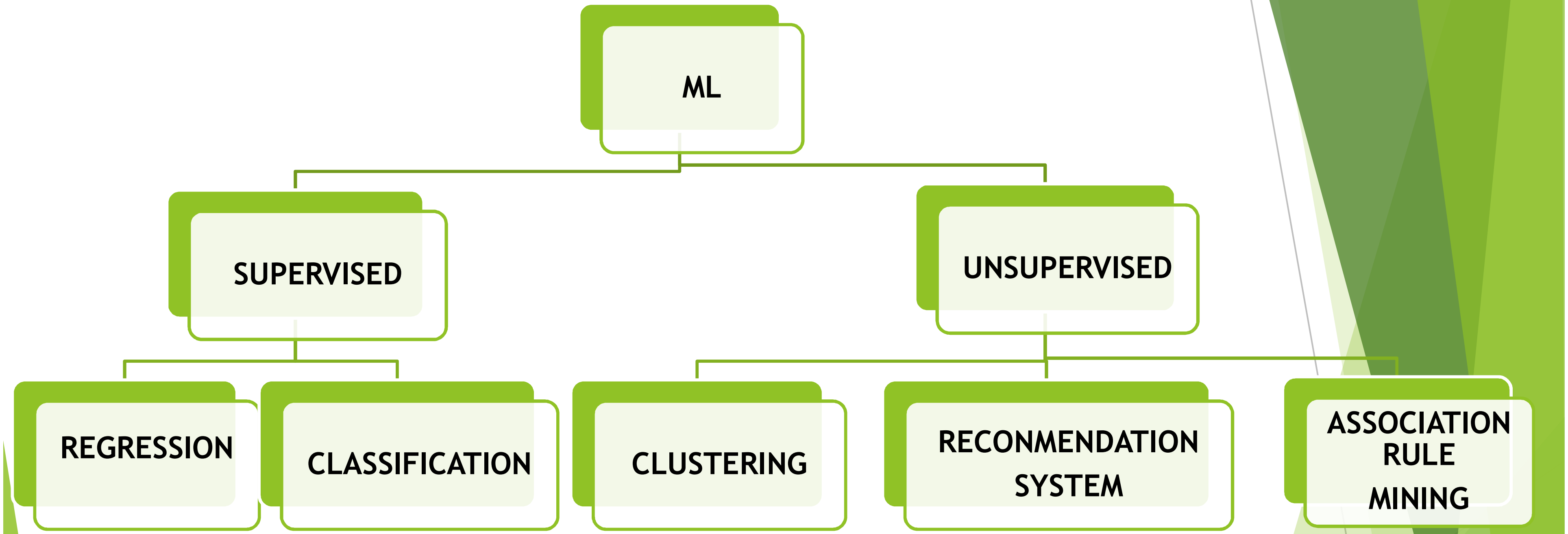
WHAT IS LIKELY TO HAPPEN

PRESCRIPTIVE

WHAT SHOULD WE DO

Difference between Statistician and Data Analytics

- **Stats:** Draft the model for the problem and ask for data.
- **DA:** Start with data and find the solution.
- **Key Points:**
- **Statistician approach**: Formulate a problem and then gather data.
- **DA approach:** Analyze existing data to uncover insights.
- **The importance of asking the right questions in DA vs. collecting the right data in statistics.**



ML

SUPERVISED

UNSUPERVISED

REGRESSION

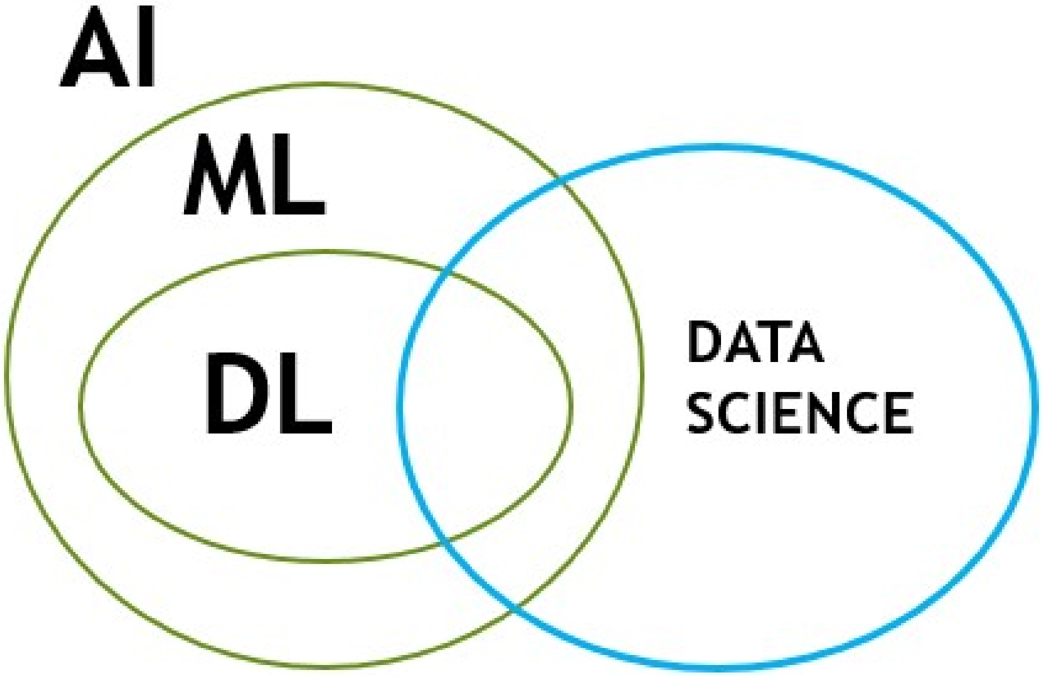
CLASSIFICATION

CLUSTERING

RECONMENDATION SYSTEM

ASSOCIATION RULE MINING

AI : Where AI is system or machine that THINK as well as ACT like human and rationally.



- **ML** is field of computer science/AI that uses statistics techniques to give computer/machine/system the ability to “learn” with data without being explicitly programmed. **Arthur Samuel, 1959.**
- A computer program is said to learn from **experience E** with respect to some **task T** and some **performance measure P**, if its performance on T, as measured by P, improves with experience E. (**Tom Mitchell, 1997)**.
- Where experience is same as Training of model.

Roles and Responsibilities

- **Data Analysts** focus on historical data analysis, reporting, and providing insights to improve decision-making.
- **Business Analysts** focus on identifying business needs, streamlining processes, and aligning business and technical teams.
- **Data Scientists** apply advanced algorithms and machine learning techniques to predict future trends and uncover hidden patterns in data.

Data Analyst

- **Key Focus:** Data Analysts work on understanding patterns and trends from data to help organizations make informed decisions.
- **Responsibilities:**
- Collect, clean, and organize data for analysis.
- Use statistical tools (like Excel, SQL, Python, or R) to analyze datasets.
- Generate reports and dashboards to provide insights into data trends (using tools like Tableau, Power BI).
- Provide actionable insights based on historical data.
- Collaborate with other teams to understand their data needs and provide necessary insights.

Business Analyst

- **Key Focus:** Business Analysts focus on understanding business problems and identifying solutions by analyzing business processes, requirements, and data.
- **Responsibilities:**
- Gather business requirements by working closely with stakeholders (management, clients, etc.).
- Analyze business processes to identify areas of improvement.
- Conduct cost-benefit analysis and recommend solutions.
- Liaise between IT, business teams, and management to align project objectives.
- Translate business needs into technical requirements for developers.
- Prepare reports and presentations to communicate findings to stakeholders.

Data Scientist

- **Key Focus:** Data Scientists use advanced algorithms, machine learning models, and statistical methods to make predictions and uncover patterns in large, complex datasets.
- **Responsibilities:**
- Perform exploratory data analysis (EDA) to understand complex datasets.
- Develop predictive models using machine learning techniques.
- Build algorithms for pattern detection, classification, and prediction.
- Work with unstructured data (text, images, video, etc.).
- Communicate insights and predictions to business leaders for strategic decision-making.
- Deploy models into production environments and monitor performance.
- Collaborate with data engineers and software developers to implement and scale models.

Scenario: Decline in Sales in the Last Quarter

- **Data Analyst Reaction:**
  - **Focus: Exploring historical sales data for patterns and insights.**
  - **Steps:**
  - **Collect Data:** The Data Analyst gathers data on sales from the last few quarters, looking at total sales, product categories, and regions.
  - **Clean and Analyze Data:** They clean the dataset (removing duplicates, handling missing values), then perform a trend analysis to see when the drop started and in which regions or products.
  - **Generate Reports:** They generate visual reports (e.g., in Excel or Power BI) showing sales performance over time, broken down by product category, region, and customer segments.
- **Outcome:**
  - **The Data Analyst presents a report showing that sales have declined mostly in the East region for Product A and gives possible reasons like seasonal variations or customer churn.**

Business Analyst Reaction:

- **Focus: Understanding the business impact and identifying root causes for the decline.**
- **Steps:**
  - **Talk to Stakeholders:** The Business Analyst communicates with the sales and marketing teams to understand any potential non-data-related reasons for the decline (e.g., new competitors, change in product pricing, ineffective promotions).
  - **Review Business Processes:** They analyze business processes such as sales strategy, pricing models, or marketing campaigns to see if any changes could have contributed to the drop.
  - **Conduct a Gap Analysis:** They perform a gap analysis to understand the difference between expected and actual sales, identifying weaknesses in business strategies.
- **Outcome:**
  - **The Business Analyst identifies that a recent price increase for Product A might have reduced demand in the East region, and suggests revising the pricing strategy or offering promotional discounts.**

Data Scientist Reaction:

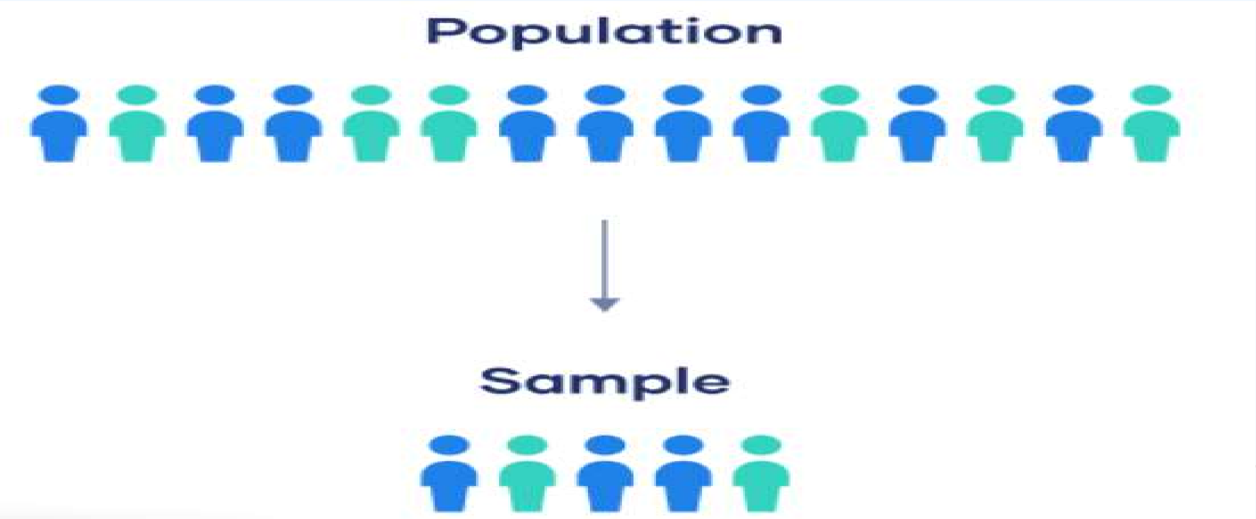
- **Focus: Predicting future sales trends and uncovering complex patterns.**
  - **Steps:**
  - **Data Gathering:** The Data Scientist pulls in not only sales data but also external data such as customer reviews, competitor data, and economic indicators (e.g., inflation, unemployment).
  - **Exploratory Data Analysis (EDA):** They run complex analyses, such as customer segmentation, and create models to identify which customers are most likely to stop purchasing (churn analysis).
  - **Develop Predictive Models:** The Data Scientist builds a machine learning model that predicts future sales based on different factors, such as customer sentiment, product pricing, and market trends.
  - **Outcome:**
  - **The Data Scientist identifies that the decline in sales is likely to continue for Product A unless changes are made. Their model shows that competitor discounts and customer dissatisfaction (via review sentiment analysis) are key reasons for the drop.**

Let us start with some elementary information:

1. Population
2. Sample
3. Variable and its type
4. Frequency distribution
5. Relative Frequency

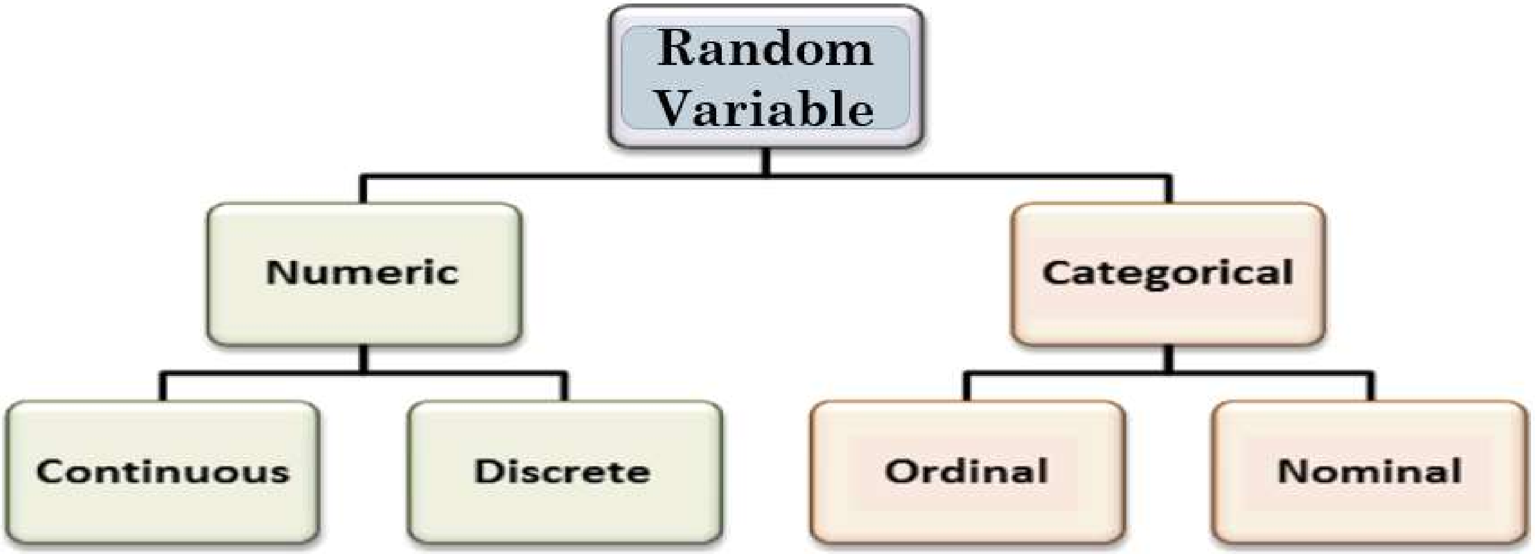
Population vs. Sample

- **Population:** Complete set of individuals, objects, or observations of interest.
- **Sample:** Subset of the population used to make inferences about the population.



Random Variable:

- In **statistics**, random variables are fundamental components that represent the characteristics or properties of the data being studied.
- By definition for a random variable characteristics or quantity changes with time and space and are not predictable. Eg: Blood



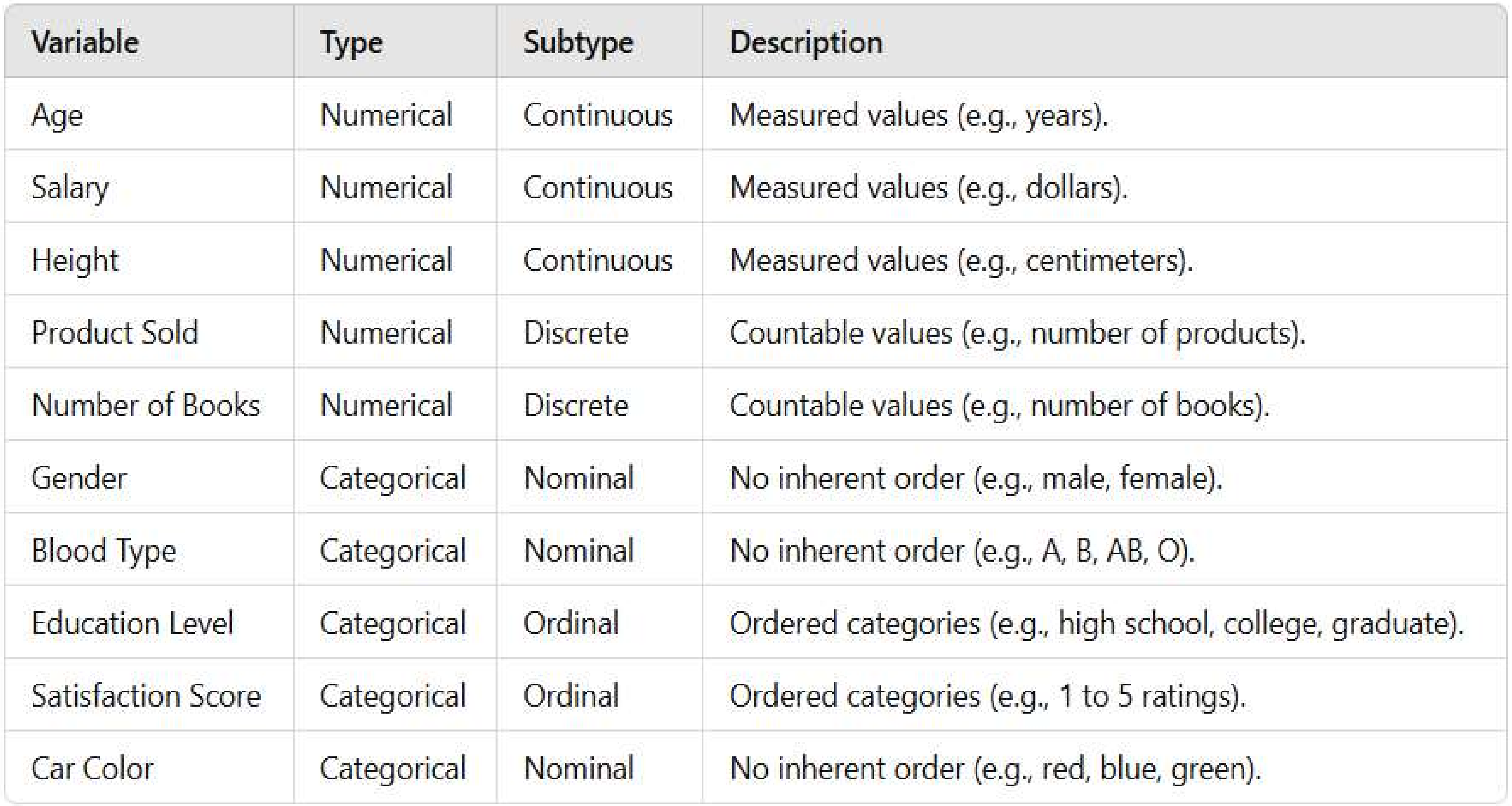
Numerical Variables: Quantitative

- **Continuous Variables:**
  - They can take any numeric value within a range.
  - **Examples:**
    - **Weight of products (e.g., 1.5 kg, 2.34 kg),**
    - **temperature readings.**
- **Discrete Variables :** Countable values, often integers.
  - **Example:**
  - **Number of employees in a company**,
  - **Number of products sold.**

Categorical Variables: Qualitative

- Data is divided into categories or groups.
- **For example**, gender can be categorized as Male or Female, and favorite colors are chosen from Red, Blue, Green, etc.
- **Nominal:** These are categories without a natural order. **Eg:**
  - **Types of Color**: There's no ranking between colors (Red, Blue, Green),
  - **Type of Gender**: There's no ranking between Male or Female.
- **Ordinal:** These categories have an order. **Eg:**
  - **Education Level:** (High School < Bachelor < Master)
  - **Customer Feedback** (Poor < Average < Good < Excellent)

Example:



Variable	Type	Subtype	Description
Age	Numerical	Continuous	Measured values (e.g., years).
Salary	Numerical	Continuous	Measured values (e.g., dollars).
Height	Numerical	Continuous	Measured values (e.g., centimeters).
Product Sold	Numerical	Discrete	Countable values (e.g., number of products).
Number of Books	Numerical	Discrete	Countable values (e.g., number of books).
Gender	Categorical	Nominal	No inherent order (e.g., male, female).
Blood Type	Categorical	Nominal	No inherent order (e.g., A, B, AB, O).
Education Level	Categorical	Ordinal	Ordered categories (e.g., high school, college, graduate).
Satisfaction Score	Categorical	Ordinal	Ordered categories (e.g., 1 to 5 ratings).
Car Color	Categorical	Nominal	No inherent order (e.g., red, blue, green).

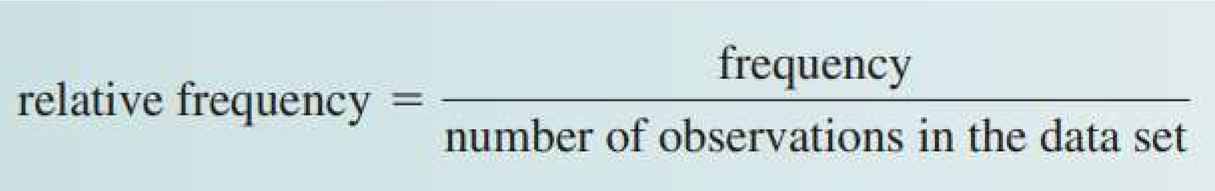
Frequency Distribution

- When the data set is categorical, a common way to present the data is in the form of a table, called a **frequency distribution.**

- A frequency distribution for categorical data is a table that displays the possible categories along with the associated frequencies and/or relative frequencies.

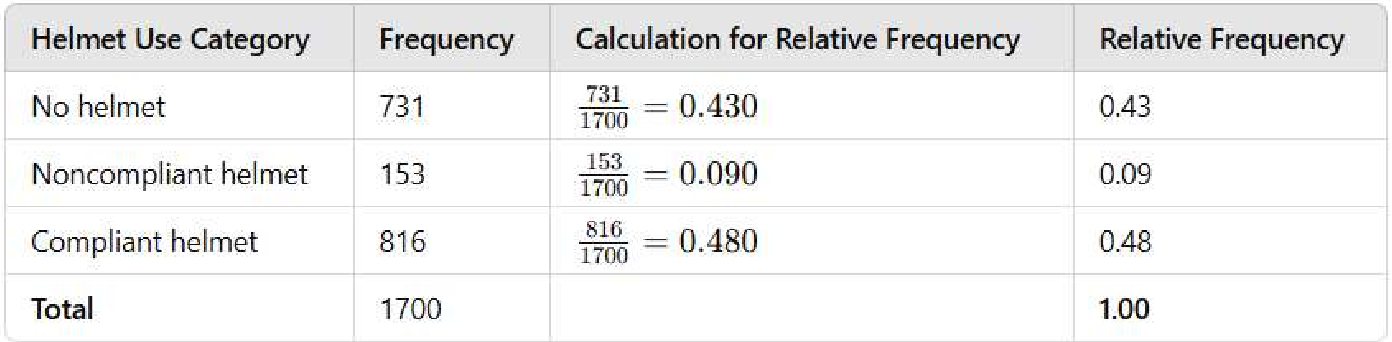
- The frequency for a particular category is the number of times the category appears in the data set.

- The **relative frequency** for a particular category is the fraction or proportion of the observations resulting in the category.

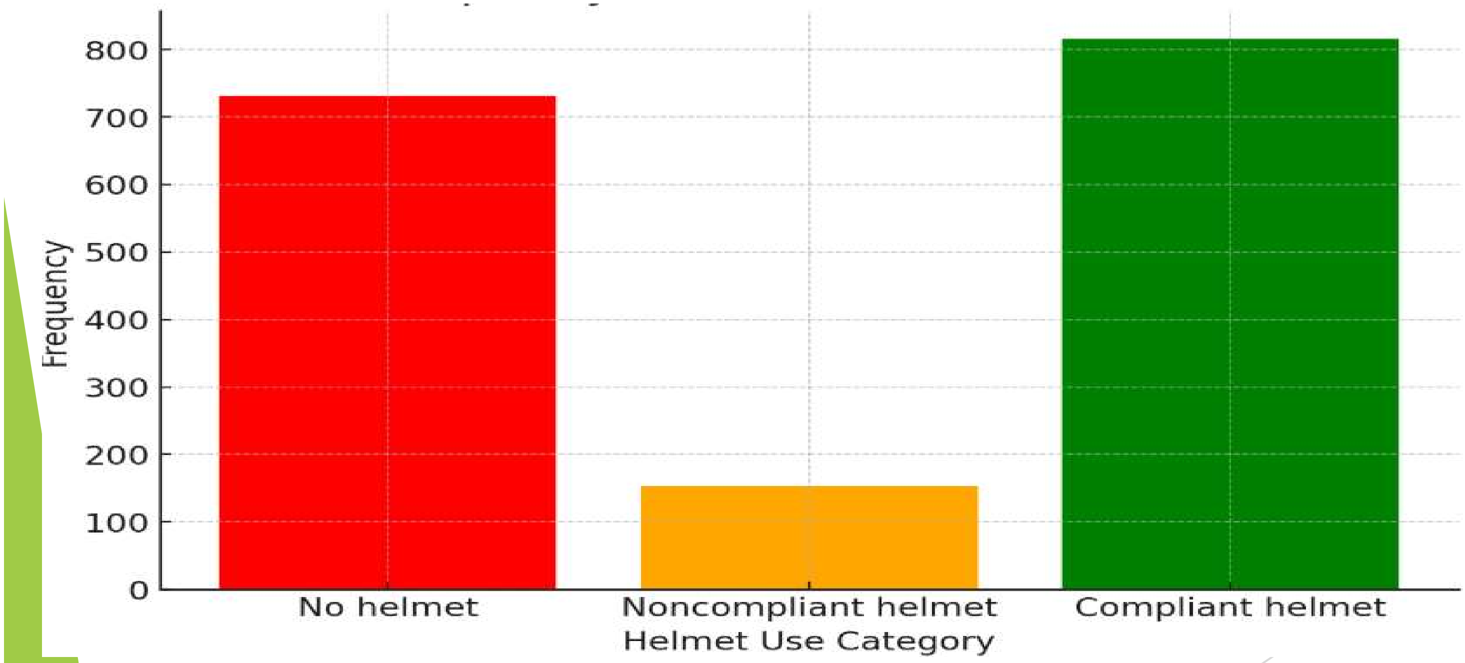


$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

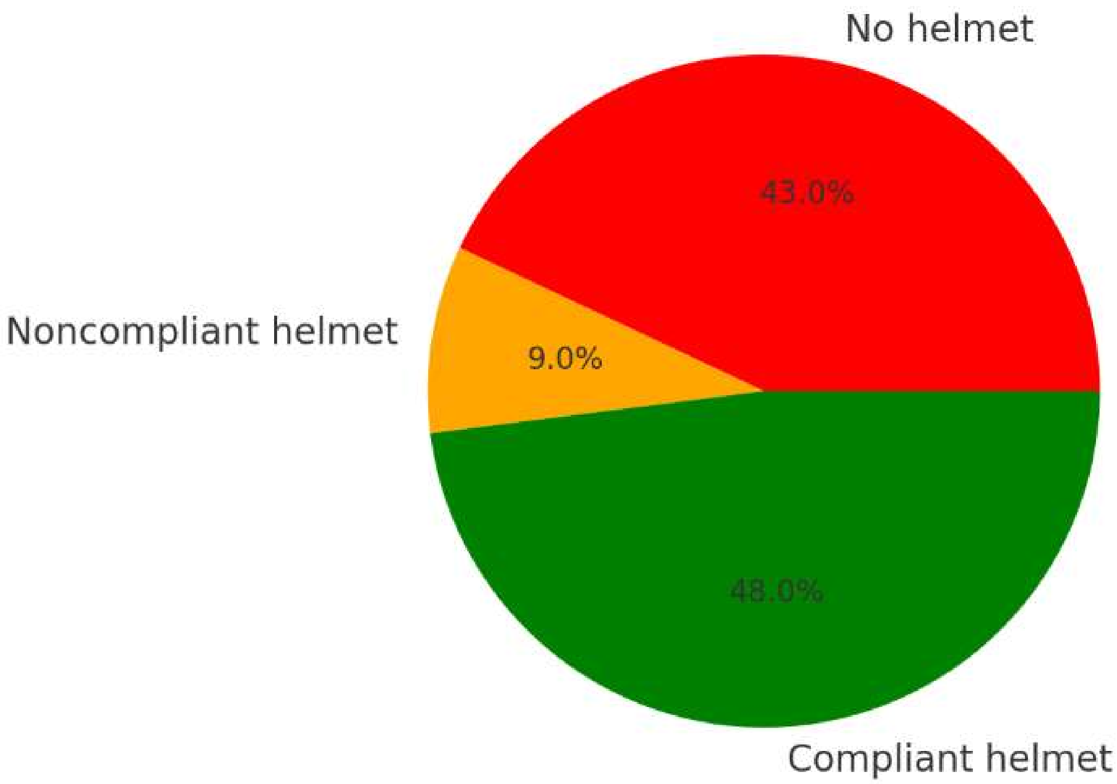
Frequency Distribution of Helmet use



Helmet Use Category	Frequency	Calculation for Relative Frequency	Relative Frequency
No helmet	731	$\frac{731}{1700} = 0.430$	0.43
Noncompliant helmet	153	$\frac{153}{1700} = 0.090$	0.09
Compliant helmet	816	$\frac{816}{1700} = 0.480$	0.48
Total	1700		**1.00**



Relative Frequency Distribution of Helmet use

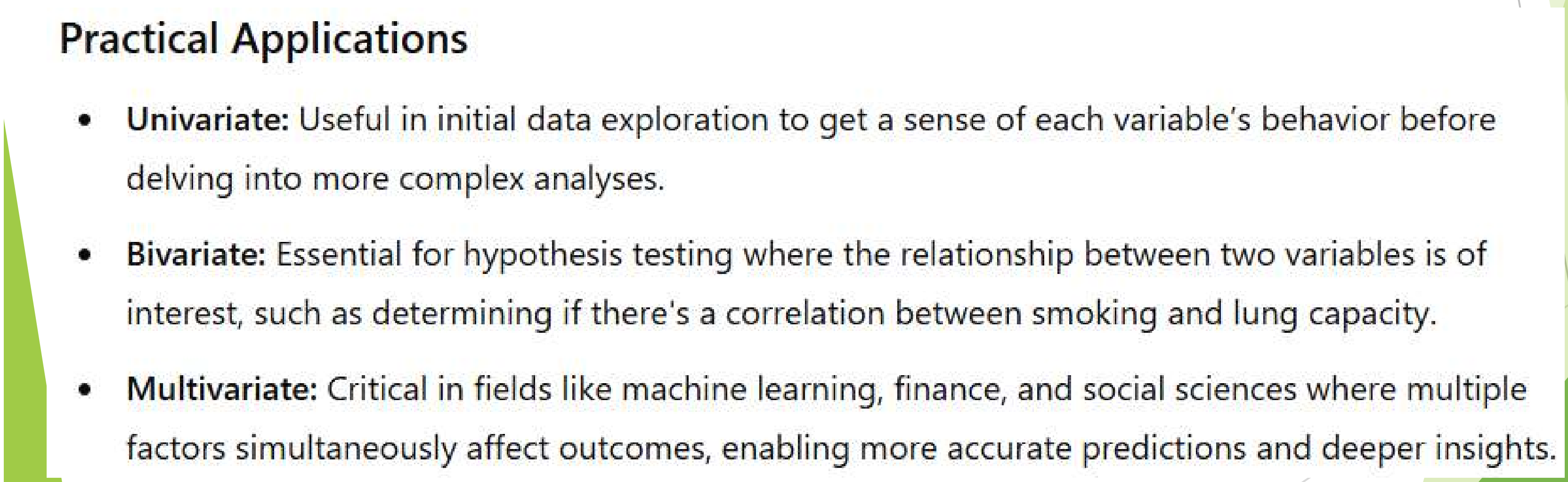


Let us begin with Descriptive Statistics….

Descriptive Statistics

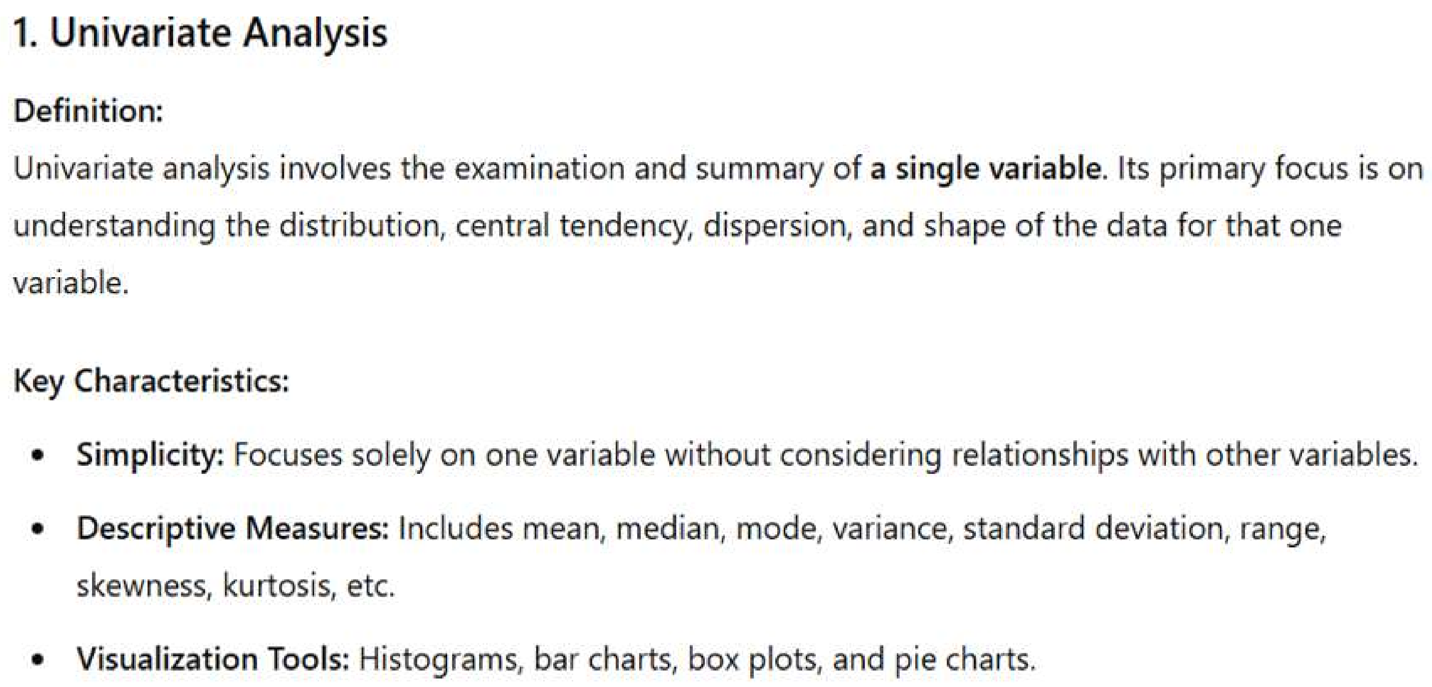
- Descriptive statistics is a branch of statistics that deals with summarizing and describing the features of a dataset.
- Its primary purpose is to provide a concise and meaningful summary of the main characteristics of the data.

- In the context of **descriptive statistics**, the terms **univariate**, **bivariate**, and **multivariate** refer to the number of variables being analyzed and the complexity of the relationships being examined within a dataset.
- Understanding these concepts is fundamental for effectively summarizing and interpreting data.



Practical Applications

- **Univariate:** Useful in initial data exploration to get a sense of each variable's behavior before delving into more complex analyses.
- **Bivariate:** Essential for hypothesis testing where the relationship between two variables is of interest, such as determining if there's a correlation between smoking and lung capacity.
- **Multivariate:** Critical in fields like machine learning, finance, and social sciences where multiple factors simultaneously affect outcomes, enabling more accurate predictions and deeper insights.



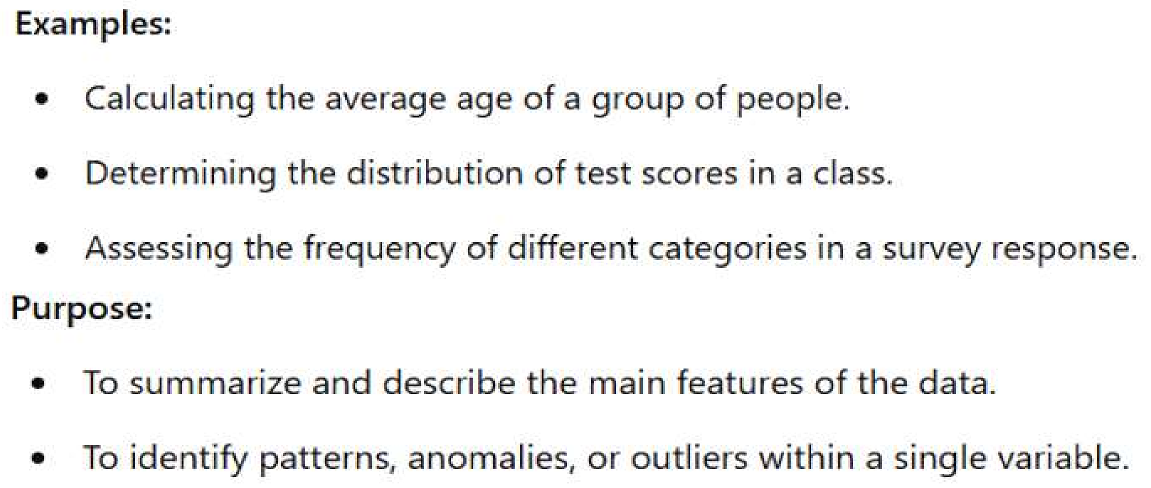
1. Univariate Analysis

Definition:

Univariate analysis involves the examination and summary of **a single variable.** Its primary focus is on understanding the distribution, central tendency, dispersion, and shape of the data for that one variable.

Key Characteristics:

- **Simplicity:** Focuses solely on one variable without considering relationships with other variables.
- **Descriptive Measures:** Includes mean, median, mode, variance, standard deviation, range, skewness, kurtosis, etc.
- **Visualization Tools:** Histograms, bar charts, box plots, and pie charts.

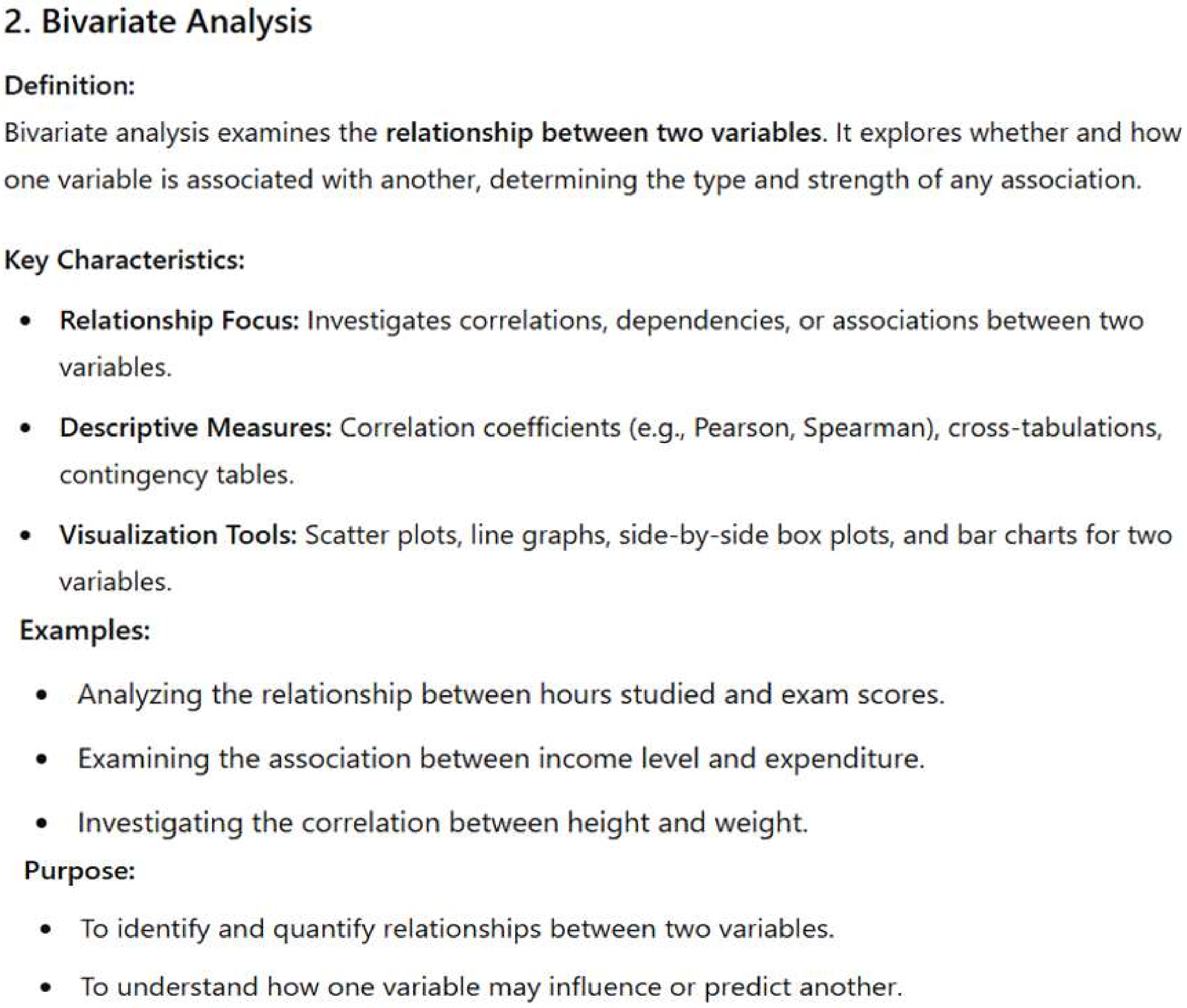


Examples:

- Calculating the average age of a group of people.
- Determining the distribution of test scores in a class.
- Assessing the frequency of different categories in a survey response.

Purpose:

- To summarize and describe the main features of the data.
- To identify patterns, anomalies, or outliers within a single variable.



2. Bivariate Analysis

Definition:

Bivariate analysis examines the **relationship between two variables.** It explores whether and how one variable is associated with another, determining the type and strength of any association.

Key Characteristics:

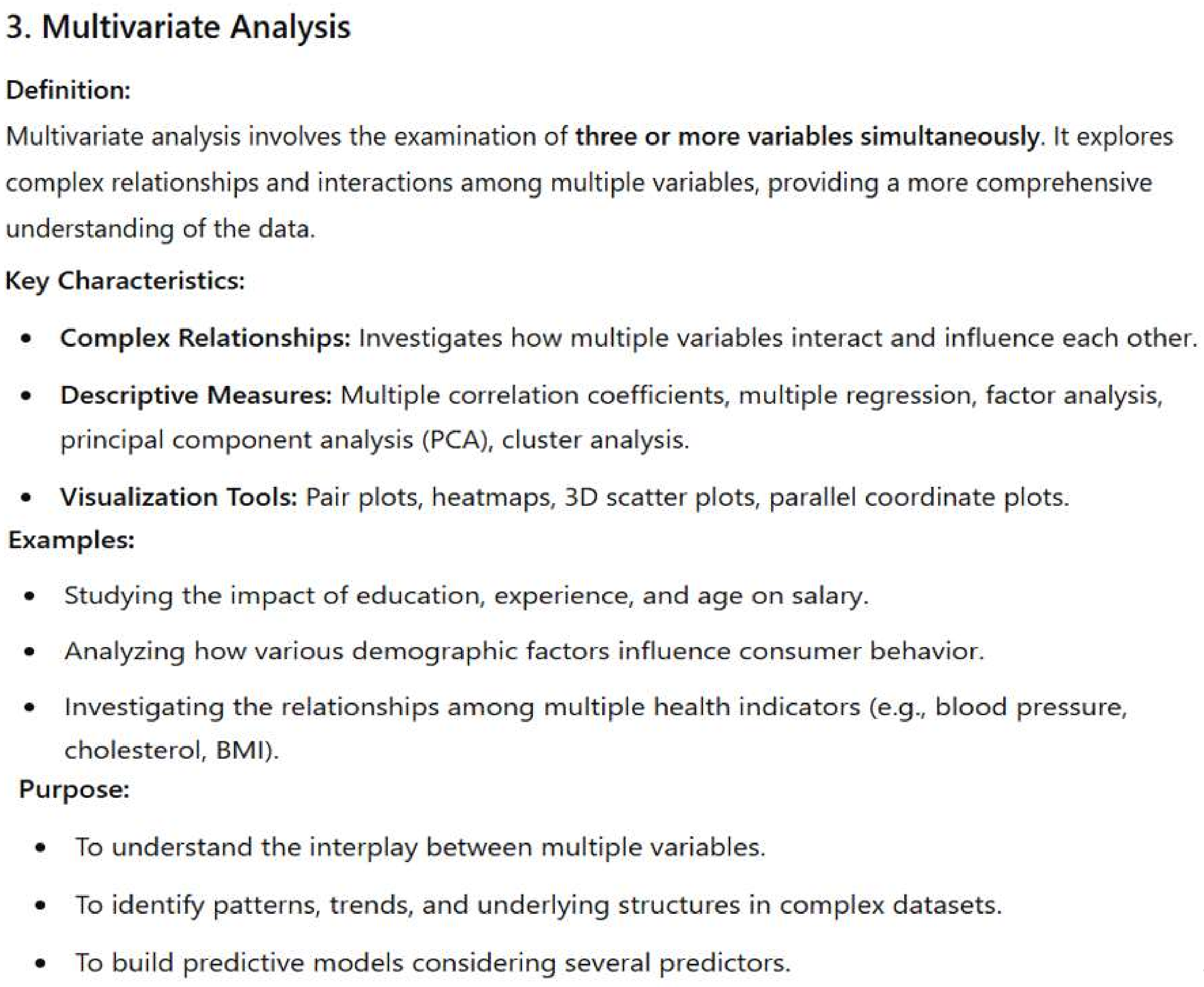
- **Relationship Focus:** Investigates correlations, dependencies, or associations between two variables.
- **Descriptive Measures:** Correlation coefficients (e.g., Pearson, Spearman), cross-tabulations, contingency tables.
- **Visualization Tools:** Scatter plots, line graphs, side-by-side box plots, and bar charts for two variables.

Examples:

- Analyzing the relationship between hours studied and exam scores.
- Examining the association between income level and expenditure.
- Investigating the correlation between height and weight.

Purpose:

- To identify and quantify relationships between two variables.
- To understand how one variable may influence or predict another.



3. Multivariate Analysis

Definition:

Multivariate analysis involves the examination of **three or more variables simultaneously**. It explores complex relationships and interactions among multiple variables, providing a more comprehensive understanding of the data.

Key Characteristics:

- **Complex Relationships:** Investigates how multiple variables interact and influence each other.
- **Descriptive Measures:** Multiple correlation coefficients, multiple regression, factor analysis, principal component analysis (PCA), cluster analysis.
- **Visualization Tools:** Pair plots, heatmaps, 3D scatter plots, parallel coordinate plots.

Examples:

- Studying the impact of education, experience, and age on salary.
- Analyzing how various demographic factors influence consumer behavior.
- Investigating the relationships among multiple health indicators (e.g., blood pressure, cholesterol, BMI).

Purpose:

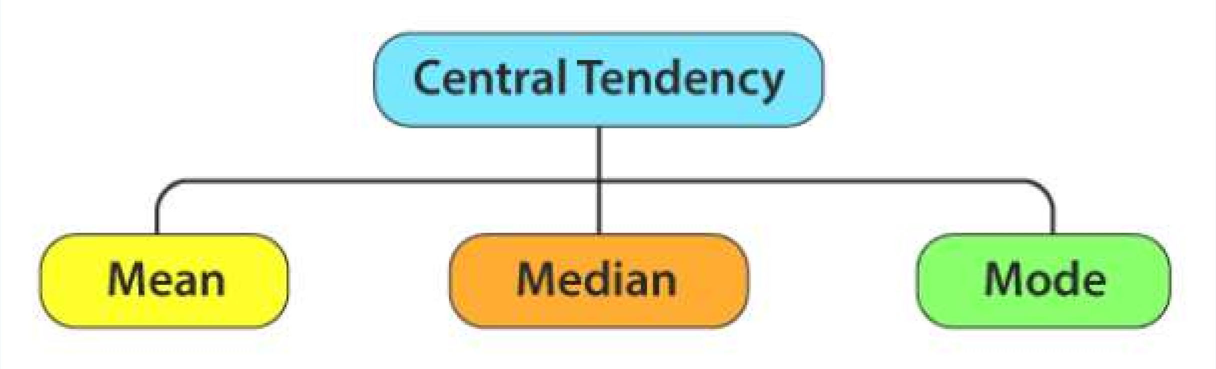
- To understand the interplay between multiple variables.
- To identify patterns, trends, and underlying structures in complex datasets.
- To build predictive models considering several predictors.

- **In Descriptive statistics,** we will use following approach.
- **For Univariate analysis:**
  - We will talk about that number which represent the entire data set or we called it **as Measures of Central Tendency** this includes:
  - Mean,
  - Median,
  - Mode

- **Once we get central value now we will try to understand the spread of data:**
- Different aspects of Spread are:
- **Range**
- **Quartiles**
- **Standard deviation**
- Later we will talk about nature of distribution, using concept of **skewness and kurtosis.**
- **For Bivariate analysis we will talk about covariance and coefficient of Correlation.**

Measures of Central Tendency

- **Definition:** Indicates the central value in a dataset.
  - They help to summarize the distribution of data and provide insights into the central or average value around which the data points tend to cluster.



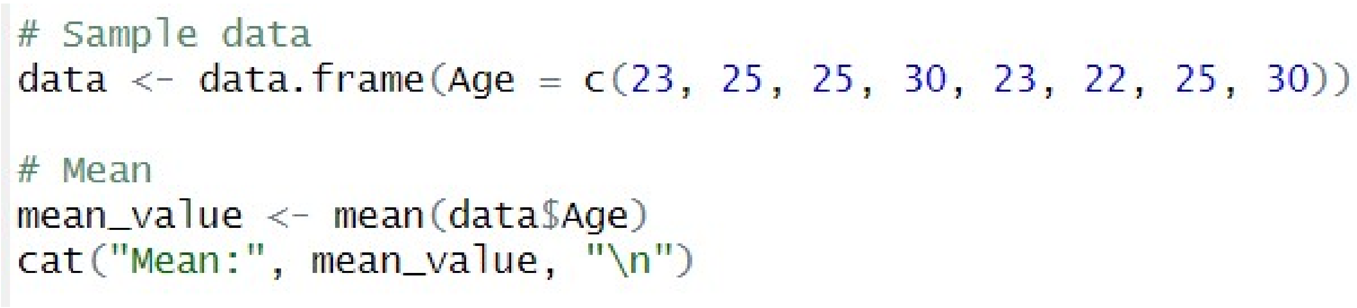
Mean:

- The mean is the average value of a dataset.
- It is calculated by summing all values in the dataset and then dividing by the number of values.
- ***Average age of a man in a group***
- ***Age of average person in a group***
- **Formula:**
- Mean = Sum of all values/Number of values

$$\sum_{i}^{n} \frac{sum\ of\ all\ values}{Number\ of\ values}$$

- The mean is sensitive to extreme values or outliers in the dataset.

Code in R



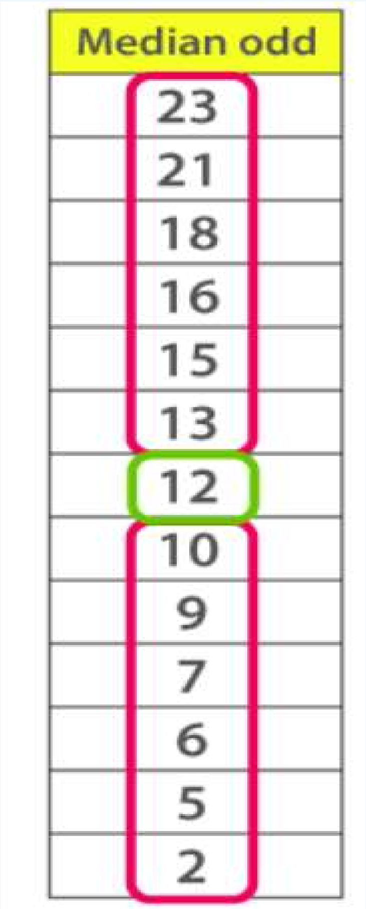
```
# Sample data
data <- data.frame(Age = c(23, 25, 25, 30, 23, 22, 25, 30))

# Mean
mean_value <- mean(data$Age)
cat("Mean:", mean_value, "\n")
```

- cat(): This function is used to concatenate and print objects.
- It is a simple way to display output on the console without including the quotes and with more control over formatting.

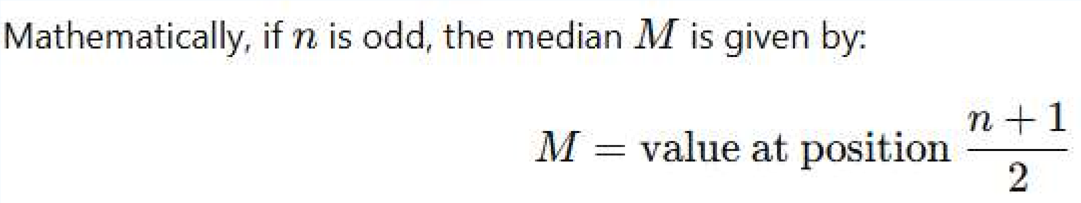
- \n: This is the newline character, which tells R to move the cursor to the next line.
- It is commonly used within cat() to separate different outputs or to create cleaner formatting in the printed text.

Median:

- The median is the middle value of a dataset when it is arranged in ascending or descending order.
- It divides the dataset into two equal halves.
- If the dataset has an odd number of values, the median is the middle value.



Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2



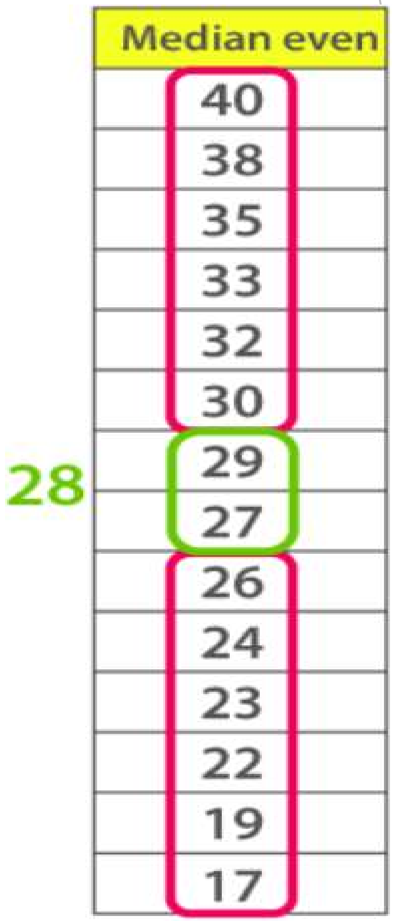
Mathematically, if n is odd, the median M is given by:

$$M = \text{value at position } \frac{n+1}{2}$$

Median:

$$M = \frac{\text{value at } \frac{n}{2} \text{ position} + \text{value at } \left(\frac{n}{2}+1\right) \text{ position}}{2}$$

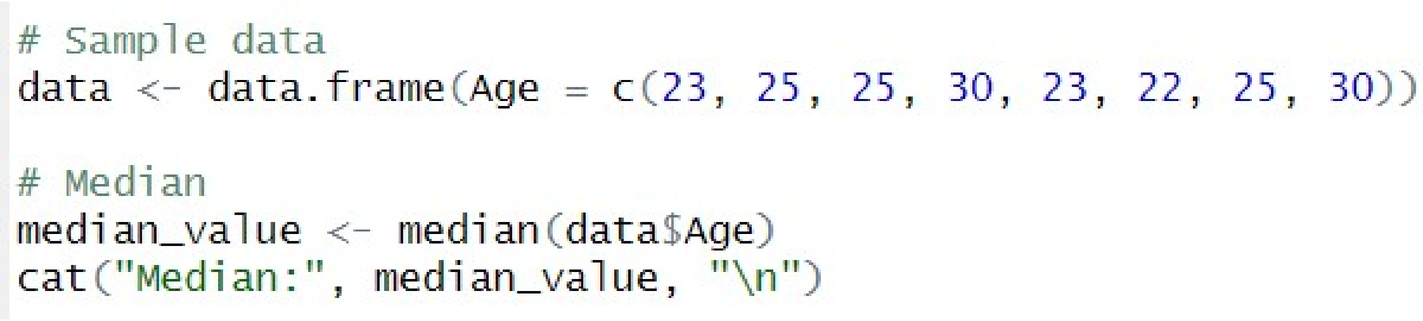
- If the dataset has an even number of values, the median is the average of the two middle values.
- **The median is less affected by outliers compared to the mean and is a robust measure of central tendency.**



Median even
40
38
35
33
32
30
29
27
26
24
23
22
19
17

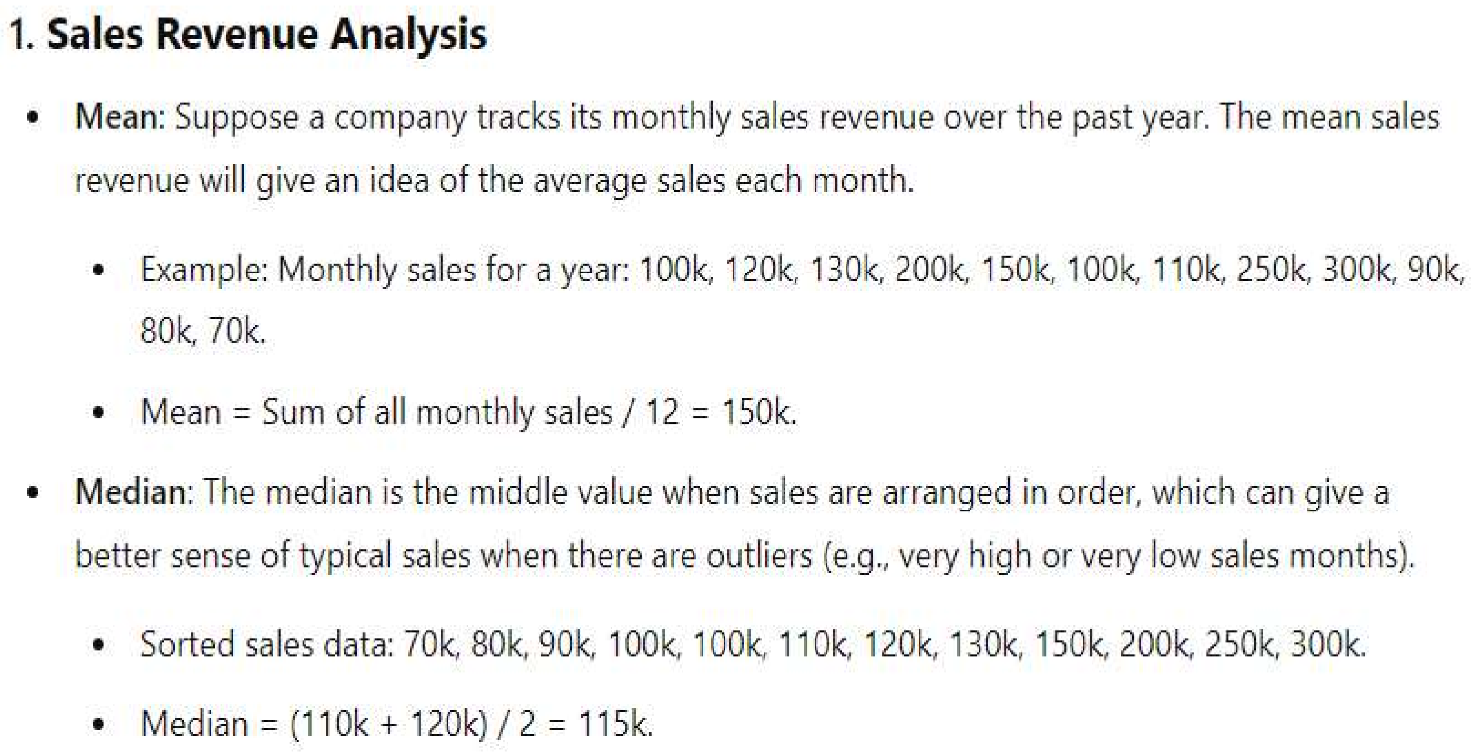
Age of average person in a group

Code in R



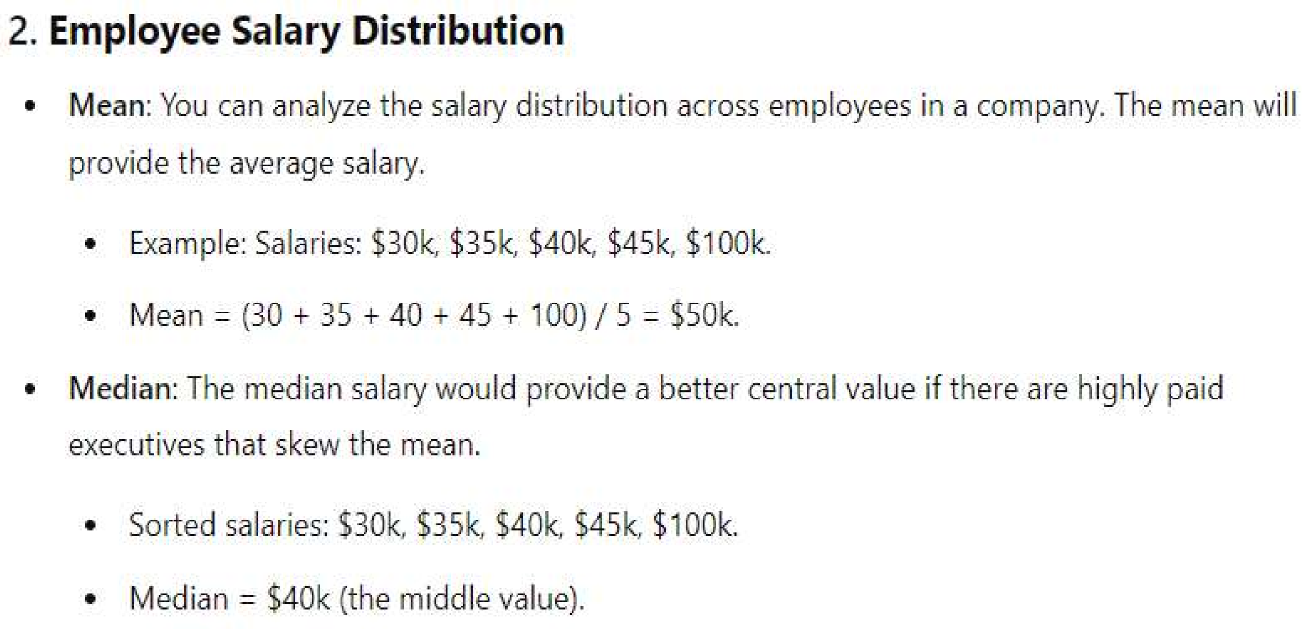
```r
# Sample data
data <- data.frame(Age = c(23, 25, 25, 30, 23, 22, 25, 30))

# Median
median_value <- median(data$Age)
cat("Median:", median_value, "\n")
```

Practical Examples



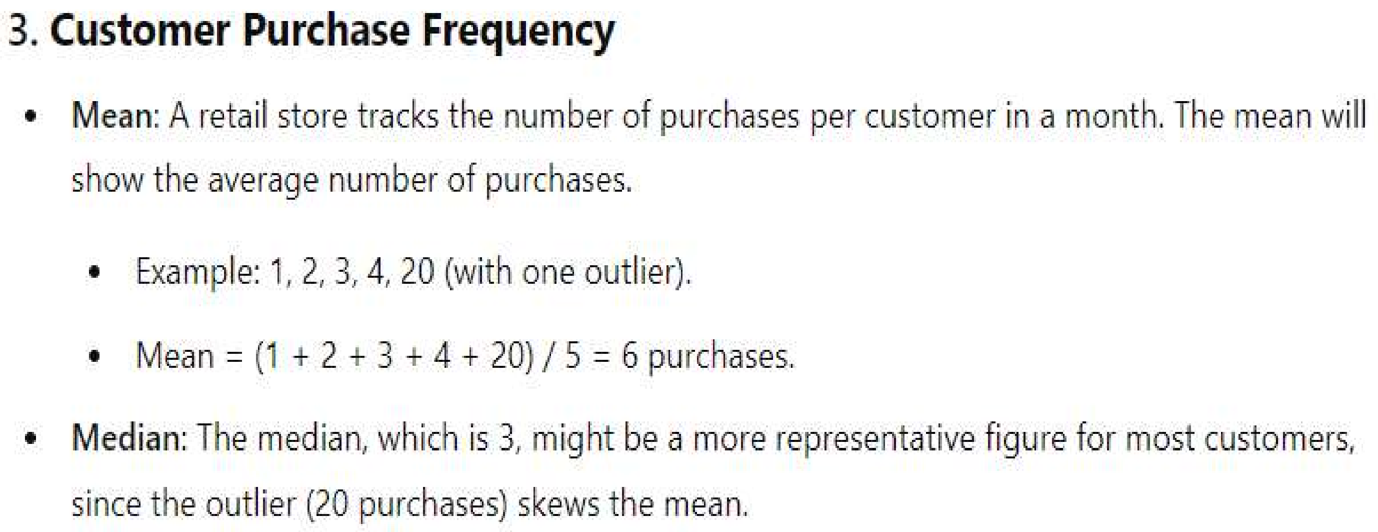
1. Sales Revenue Analysis

- **Mean:** Suppose a company tracks its monthly sales revenue over the past year. The mean sales revenue will give an idea of the average sales each month.
  - Example: Monthly sales for a year: 100k, 120k, 130k, 200k, 150k, 100k, 110k, 250k, 300k, 90k, 80k, 70k.
  - Mean = Sum of all monthly sales / 12 = 150k.
- **Median**: The median is the middle value when sales are arranged in order, which can give a better sense of typical sales when there are outliers (e.g., very high or very low sales months).
  - Sorted sales data: 70k, 80k, 90k, 100k, 100k, 110k, 120k, 130k, 150k, 200k, 250k, 300k.
  - Median = (110k + 120k) / 2 = 115k.



2. Employee Salary Distribution

- **Mean**: You can analyze the salary distribution across employees in a company. The mean will provide the average salary.
  - Example: Salaries: $30k, $35k, $40k, $45k, $100k.
  - Mean = (30 + 35 + 40 + 45 + 100) / 5 = $50k.
- **Median**: The median salary would provide a better central value if there are highly paid executives that skew the mean.
  - Sorted salaries: $30k, $35k, $40k, $45k, $100k.
  - Median = $40k (the middle value).

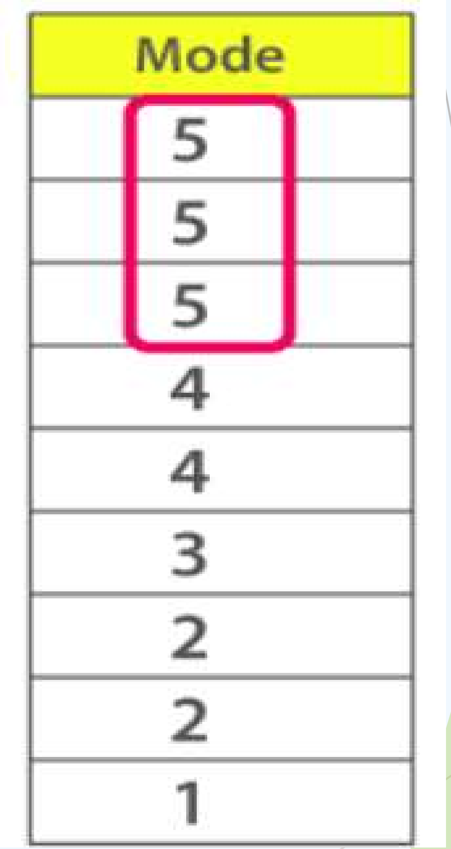


3. Customer Purchase Frequency

- **Mean**: A retail store tracks the number of purchases per customer in a month. The mean will show the average number of purchases.
  - Example: 1, 2, 3, 4, 20 (with one outlier).
  - Mean = (1 + 2 + 3 + 4 + 20) / 5 = 6 purchases.
- **Median**: The median, which is 3, might be a more representative figure for most customers, since the outlier (20 purchases) skews the mean.

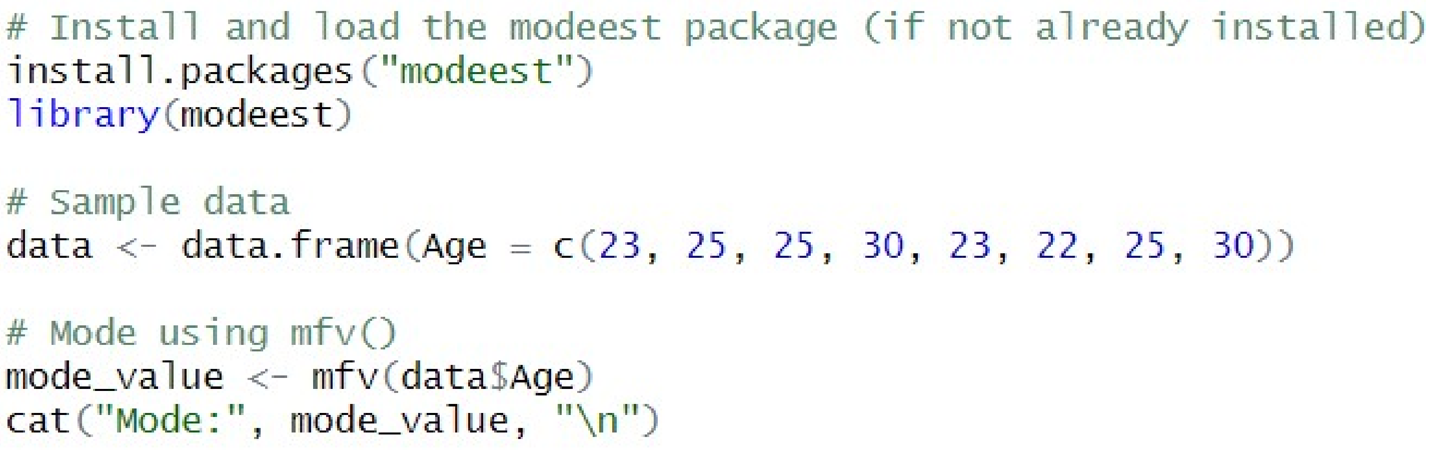
Mode:

- The **mode** is the value(s) that appear most frequently in a dataset.
- **For Unimodal Data (Single Mode)**: If there is only one value that appears more frequently than others, that value is the mode.
- **For Multimodal Data (Multiple Modes)**: If two or more values appear with the same highest frequency, each of these values is considered a mode.
- **For No Mode**: If all values appear with the same frequency (usually 1 for each), the dataset has no mode.
- **The mode is useful for categorical or discrete data but can also be calculated for continuous data.**

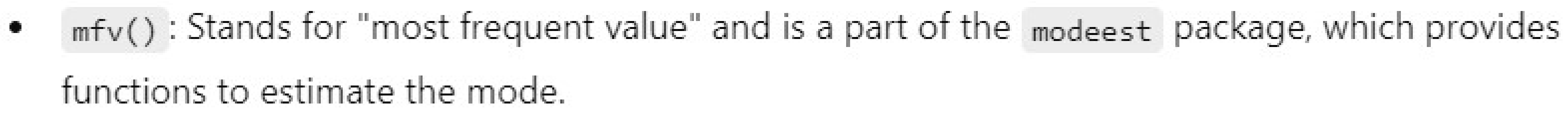


Mode
5
5
5
4
4
3
2
2
1

Code in R

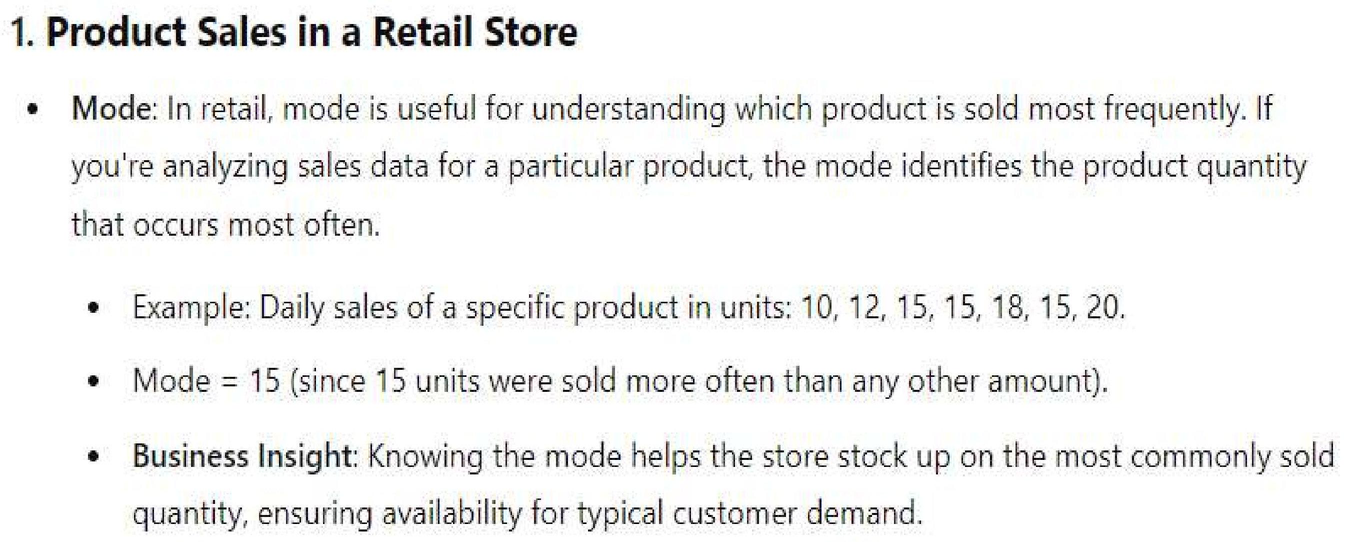


```
# Install and load the modeest package (if not already installed)
install.packages("modeest")
library(modeest)

# Sample data
data <- data.frame(Age = c(23, 25, 25, 30, 23, 22, 25, 30))

# Mode using mfv()
mode_value <- mfv(data$Age)
cat("Mode:", mode_value, "\n")
```



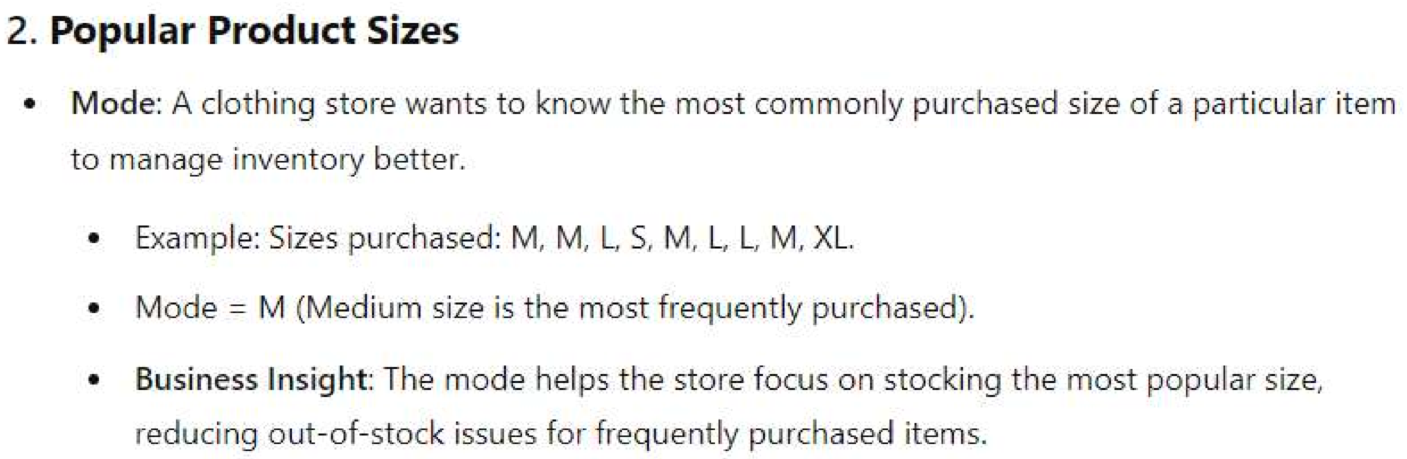
- `mfv()` : Stands for "most frequent value" and is a part of the `modeest` package, which provides functions to estimate the mode.

Practical Examples



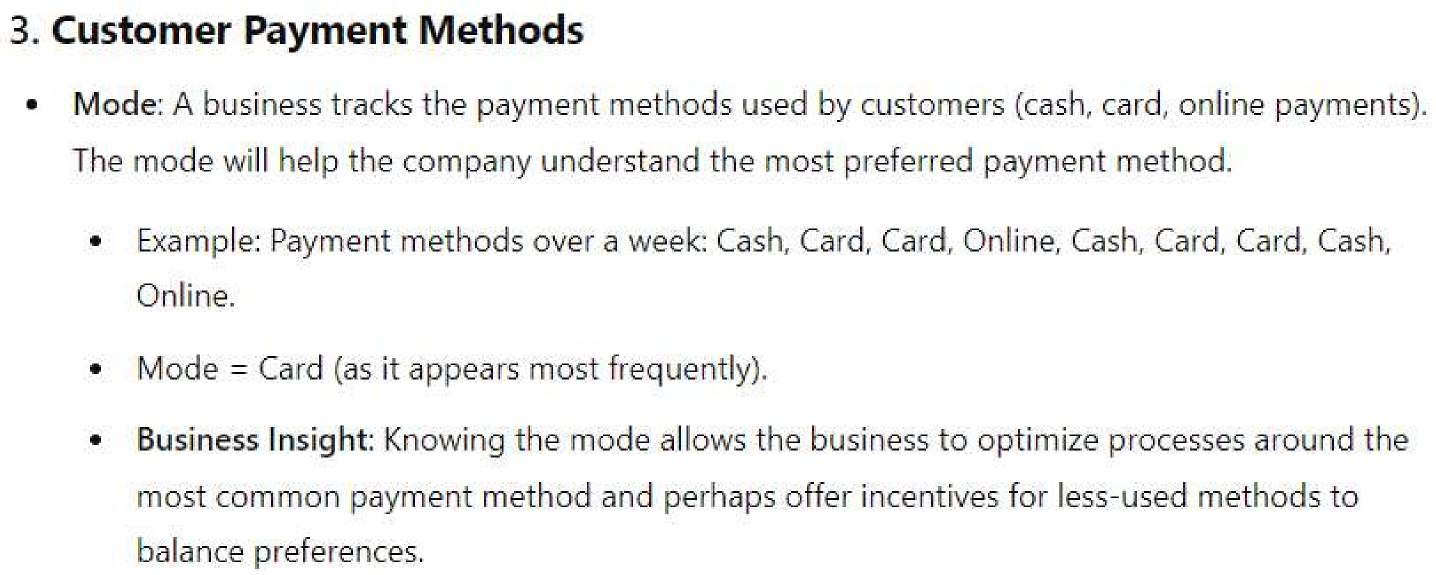
1. Product Sales in a Retail Store

- **Mode**: In retail, mode is useful for understanding which product is sold most frequently. If you're analyzing sales data for a particular product, the mode identifies the product quantity that occurs most often.
  - Example: Daily sales of a specific product in units: 10, 12, 15, 15, 18, 15, 20.
  - Mode = 15 (since 15 units were sold more often than any other amount).
  - **Business Insight**: Knowing the mode helps the store stock up on the most commonly sold quantity, ensuring availability for typical customer demand.



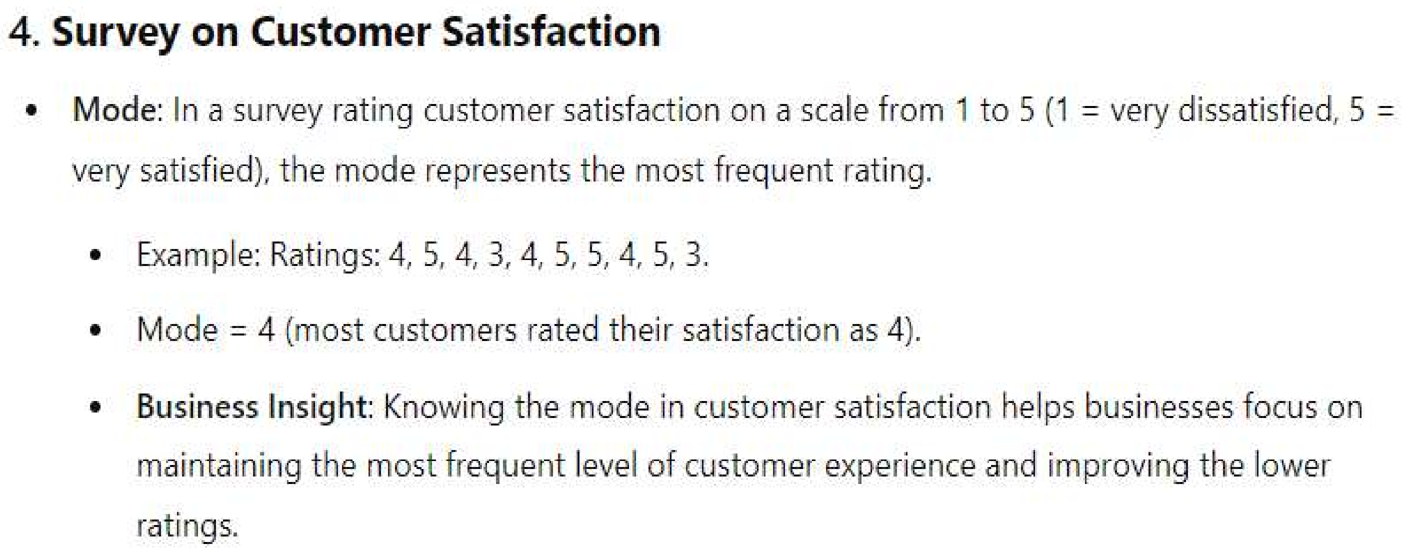
2. Popular Product Sizes

- **Mode**: A clothing store wants to know the most commonly purchased size of a particular item to manage inventory better.
  - Example: Sizes purchased: M, M, L, S, M, L, L, M, XL.
  - Mode = M (Medium size is the most frequently purchased).
  - **Business Insight**: The mode helps the store focus on stocking the most popular size, reducing out-of-stock issues for frequently purchased items.



3. Customer Payment Methods

- **Mode**: A business tracks the payment methods used by customers (cash, card, online payments). The mode will help the company understand the most preferred payment method.
  - Example: Payment methods over a week: Cash, Card, Card, Online, Cash, Card, Card, Cash, Online.
  - Mode = Card (as it appears most frequently).
  - **Business Insight**: Knowing the mode allows the business to optimize processes around the most common payment method and perhaps offer incentives for less-used methods to balance preferences.

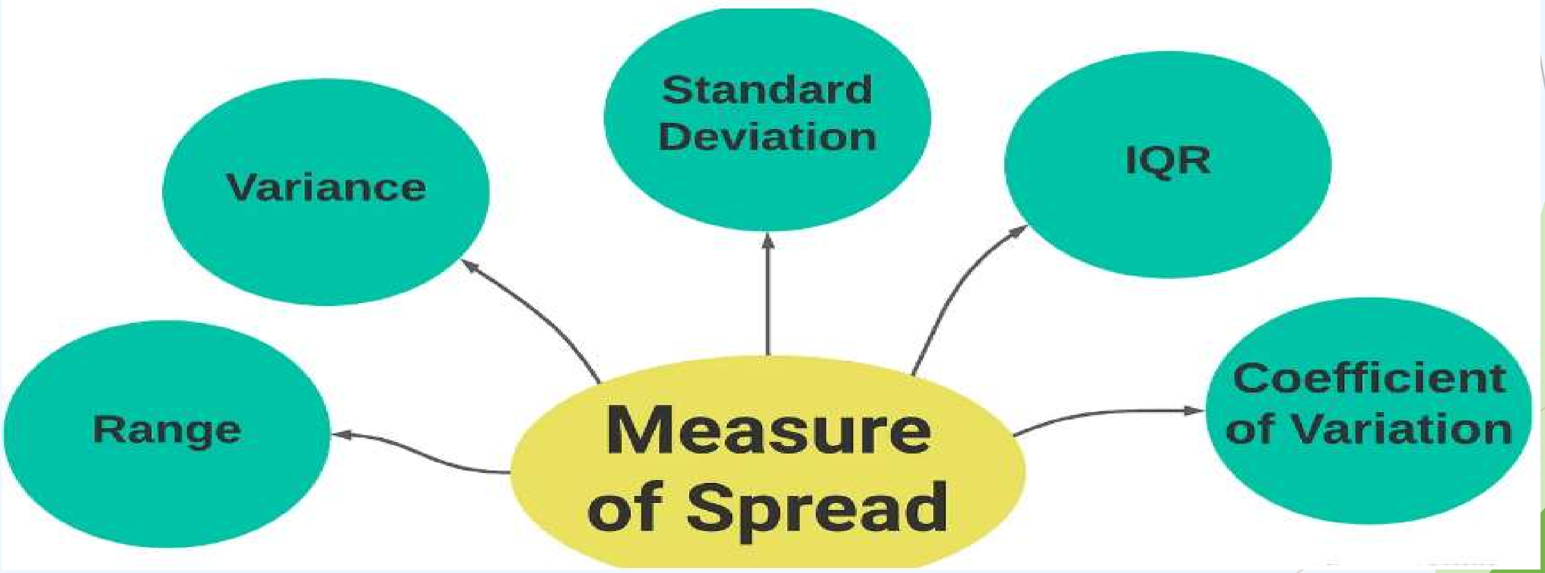


4. Survey on Customer Satisfaction

- **Mode**: In a survey rating customer satisfaction on a scale from 1 to 5 (1 = very dissatisfied, 5 = very satisfied), the mode represents the most frequent rating.
  - Example: Ratings: 4, 5, 4, 3, 4, 5, 5, 4, 5, 3.
  - Mode = 4 (most customers rated their satisfaction as 4).
  - **Business Insight**: Knowing the mode in customer satisfaction helps businesses focus on maintaining the most frequent level of customer experience and improving the lower ratings.

Measures of Spread

- Measures of spread (or dispersion) in descriptive statistics describe how data values are distributed and how they vary from the central tendency (mean or median).



Range:

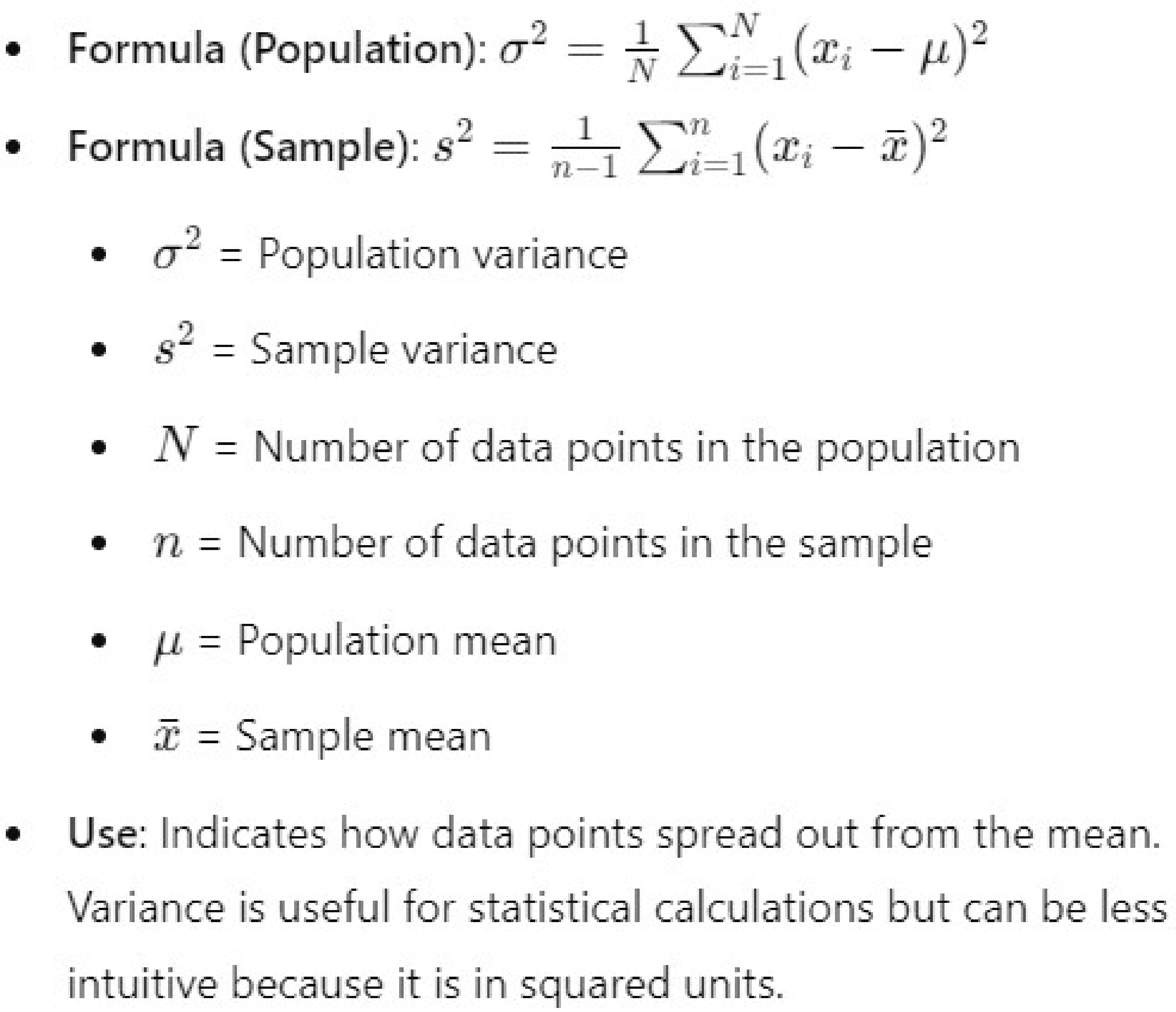
- **Definition**: The difference between the maximum and minimum values in a dataset.
- **Formula**: Range=Maximum−Minimum
- **Use**: Provides a simple measure of total spread.
- However, it is sensitive to outliers.

Example Scenario:

- In a company, if the salaries of employees are $40,000, $45,000, $50,000, $55,000, and $60,000,
- The range is $60,000 - $40,000 = $20,000.
- This tells you the spread between the lowest and highest salaries.

Variance :

- **Definition:** Measures the average squared deviation of each data point from the mean.
- It quantifies the spread of data points around the mean and how much they differ from the average value.
- Variance is influenced by outliers because it squares the deviations from the mean.
- It is less popular due its nature as it is less interpretable.



- **Formula (Population):** $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$
- **Formula (Sample):** $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
  - σ^2 = Population variance
  - s^2 = Sample variance
  - N = Number of data points in the population
  - n = Number of data points in the sample
  - μ = Population mean
  - $\bar{x}$ = Sample mean
- **Use**: Indicates how data points spread out from the mean. Variance is useful for statistical calculations but can be less intuitive because it is in squared units.

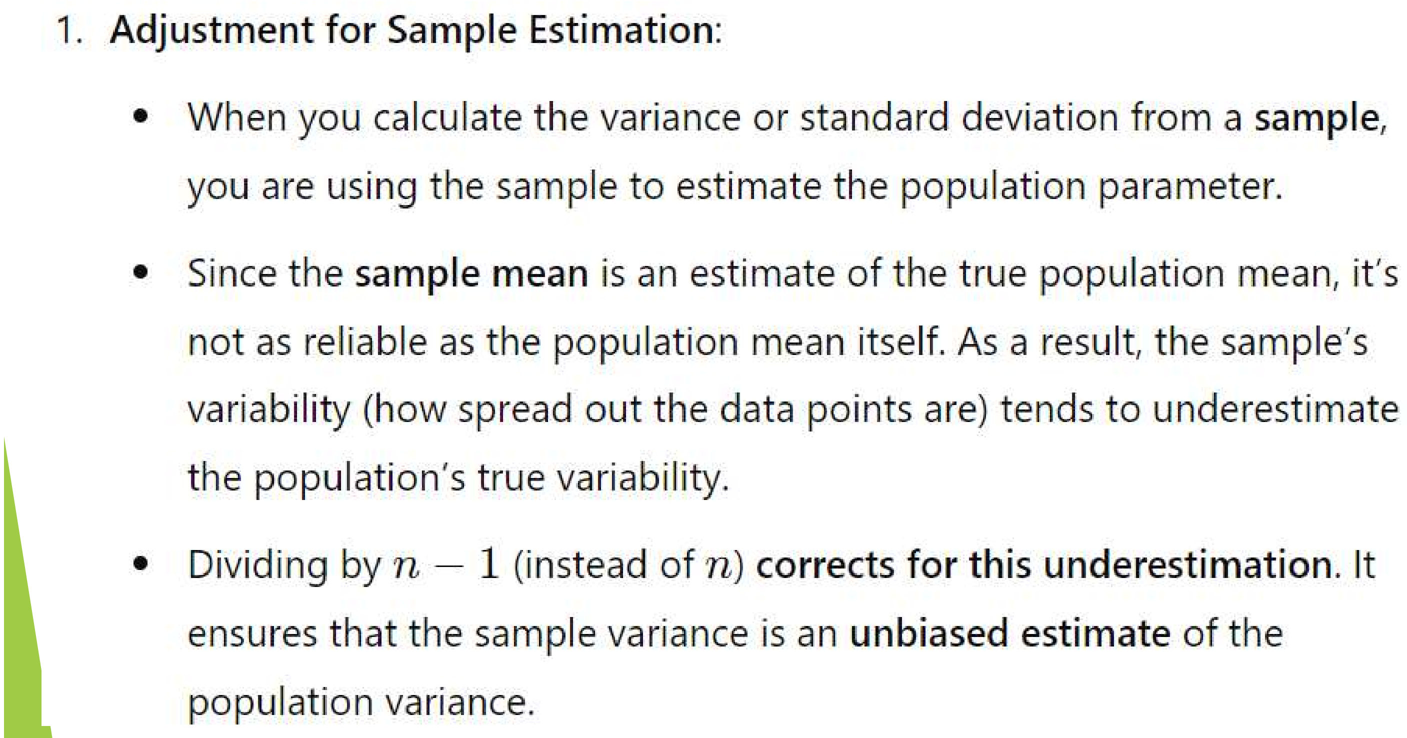
Mean to Standard Deviation?

- To understand this let us move step by step.
- After getting the data : first we find the one number which representing the entire family: **“mean”.**
- But due variability in data now we are interested in knowing this variability that move wrt mean.
- So we calculate the deviation of each variable wrt mean: **(xi– x¯)**
- But this may have problem as positive and negative values wrt mean will neutralize each other.
- So we go for squaring of these deviations and named **variance.** (xi−x¯)2
- Further to standardize this value with entire data set we are dividing this, from **degree of freedom** of my dataset which depends upon number of observation (n).
- This results in the equation as shown below:

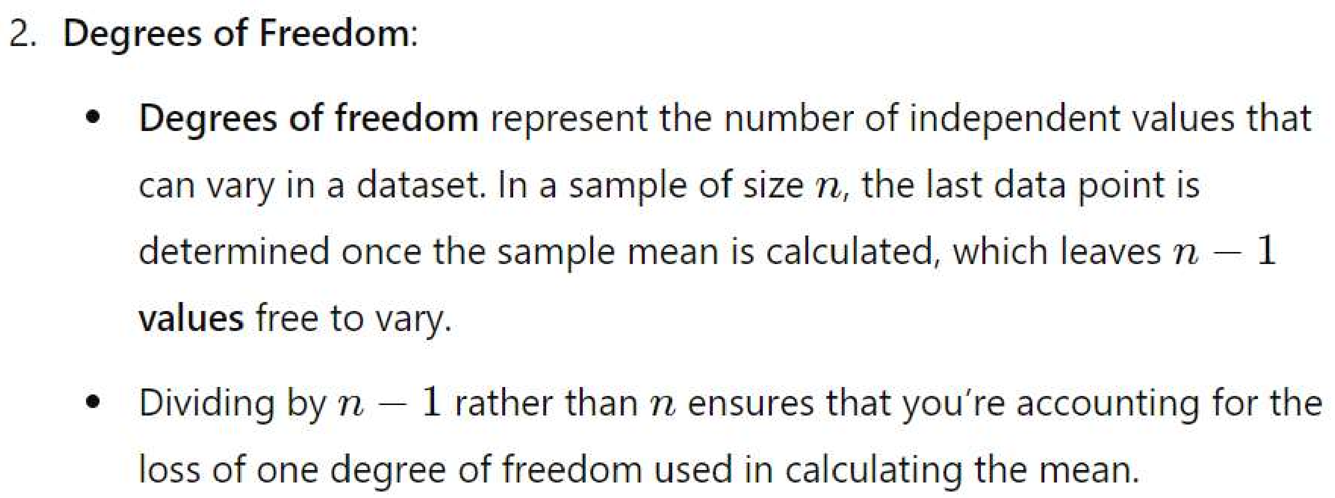
$$\text{(Population)}: \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\text{(Sample)}: s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- So here n or n-1 is different from n used in calculating mean, from definition point of view.
- Here it refers to degree of freedom and as we are interested in variability we must take into account this.

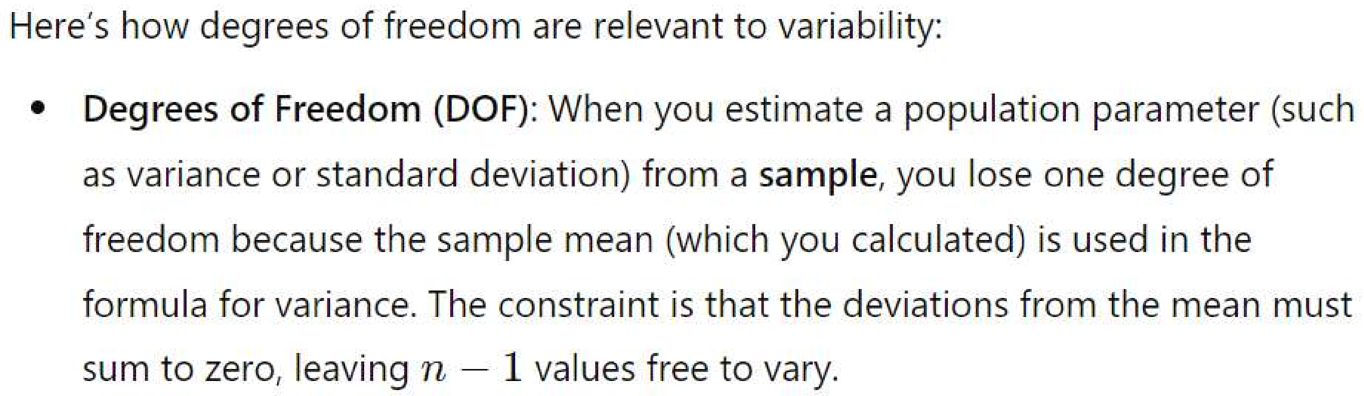


1. **Adjustment for Sample Estimation**:
   - When you calculate the variance or standard deviation from a **sample**, you are using the sample to estimate the population parameter.
   - Since the **sample mean** is an estimate of the true population mean, it's not as reliable as the population mean itself. As a result, the sample's variability (how spread out the data points are) tends to underestimate the population's true variability.
   - Dividing by $n-1$ (instead of n) **corrects for this underestimation**. It ensures that the sample variance is an **unbiased estimate** of the population variance.



2. **Degrees of Freedom**:

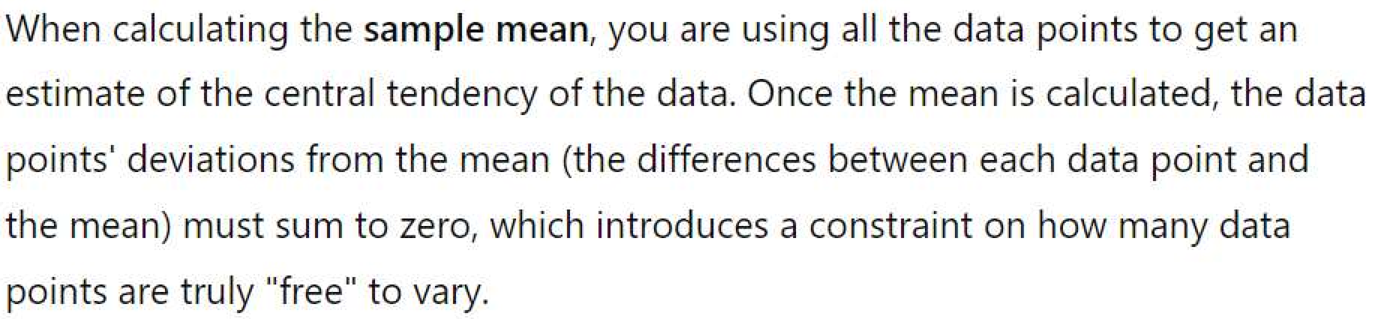
   - **Degrees of freedom** represent the number of independent values that can vary in a dataset. In a sample of size n, the last data point is determined once the sample mean is calculated, which leaves $n - 1$ **values** free to vary.
   - Dividing by $n - 1$ rather than n ensures that you're accounting for the loss of one degree of freedom used in calculating the mean.



Here's how degrees of freedom are relevant to variability:

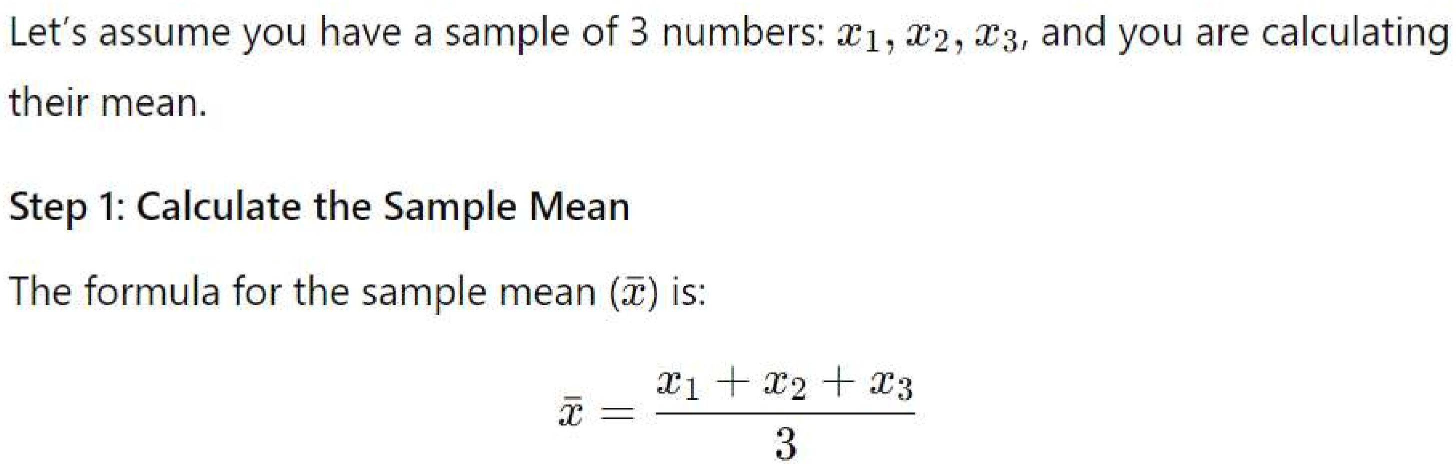
- **Degrees of Freedom (DOF)**: When you estimate a population parameter (such as variance or standard deviation) from a **sample**, you lose one degree of freedom because the sample mean (which you calculated) is used in the formula for variance. The constraint is that the deviations from the mean must sum to zero, leaving $n - 1$ values free to vary.

Degrees of Freedom and Sample Mean



When calculating the **sample mean**, you are using all the data points to get an estimate of the central tendency of the data. Once the mean is calculated, the data points' deviations from the mean (the differences between each data point and the mean) must sum to zero, which introduces a constraint on how many data points are truly "free" to vary.

Example to Understand the Concept:

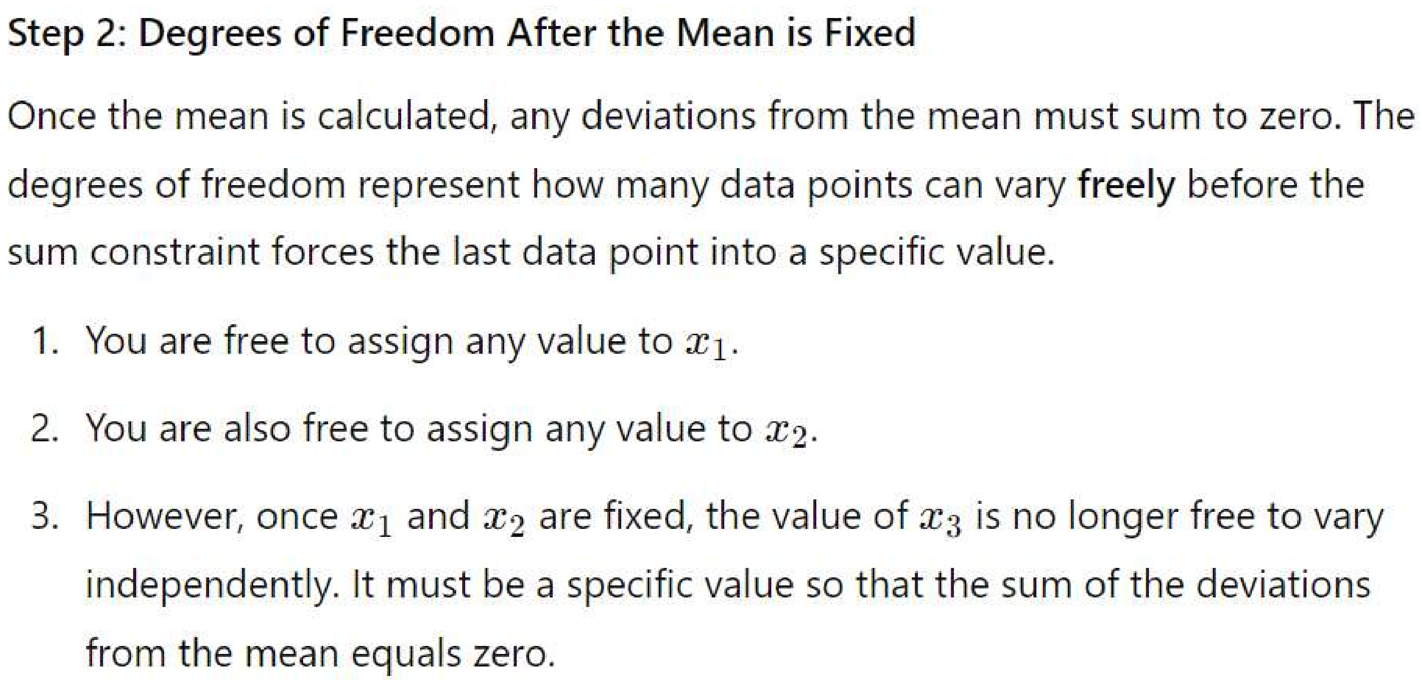


Let's assume you have a sample of 3 numbers: x_1, x_2, x_3, and you are calculating their mean.

Step 1: Calculate the Sample Mean

The formula for the sample mean ($\bar{x}$) is:

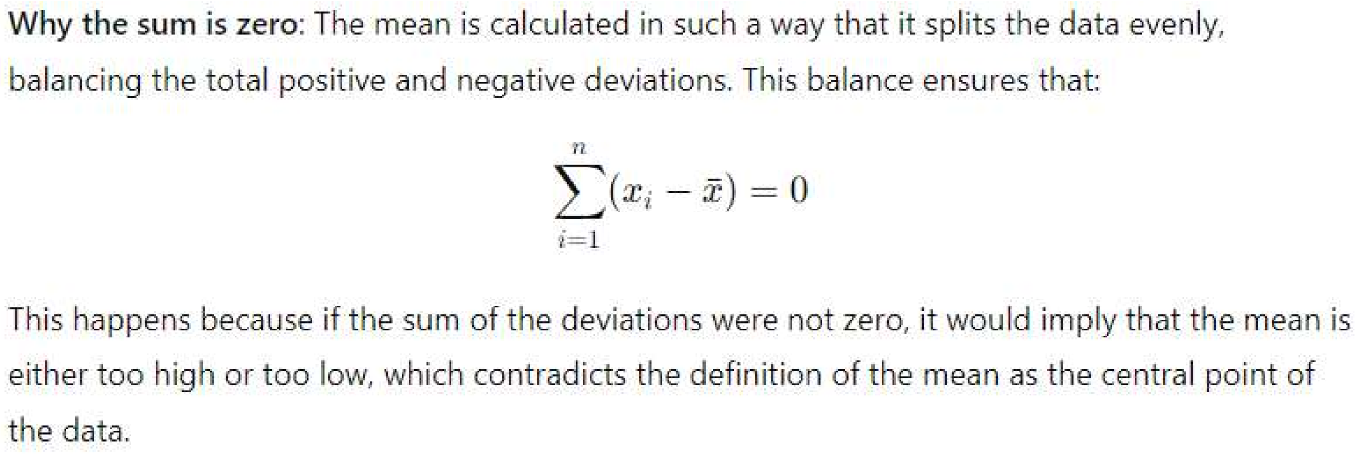
$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$



Step 2: Degrees of Freedom After the Mean is Fixed

Once the mean is calculated, any deviations from the mean must sum to zero. The degrees of freedom represent how many data points can vary **freely** before the sum constraint forces the last data point into a specific value.

1. You are free to assign any value to x_1.
2. You are also free to assign any value to x_2.
3. However, once x_1 and x_2 are fixed, the value of x_3 is no longer free to vary independently. It must be a specific value so that the sum of the deviations from the mean equals zero.

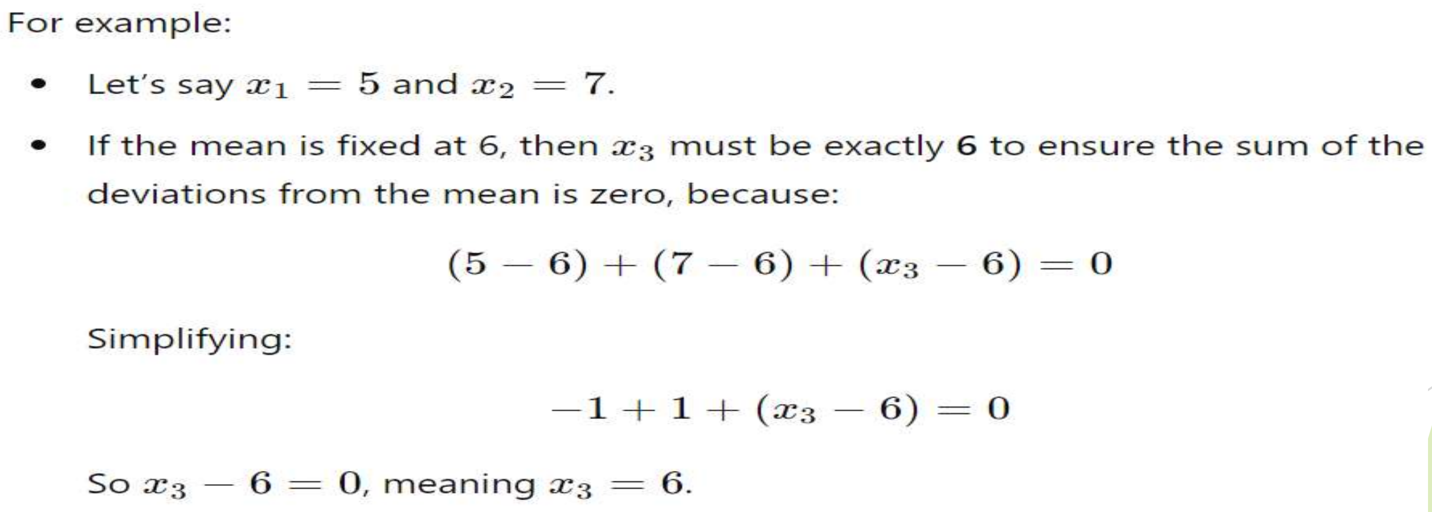


Why the sum is zero: The mean is calculated in such a way that it splits the data evenly, balancing the total positive and negative deviations. This balance ensures that:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

This happens because if the sum of the deviations were not zero, it would imply that the mean is either too high or too low, which contradicts the definition of the mean as the central point of the data.

- So, You can consider it a **fundamental property** of the mean that the sum of the deviations of the data points from the mean is always zero.
- This property reflects the fact that the mean is the **central balancing point** of the data. It ensures that the total amount of "positive" deviation (from points above the mean) is exactly balanced by the total "negative" deviation (from points below the mean).



For example:

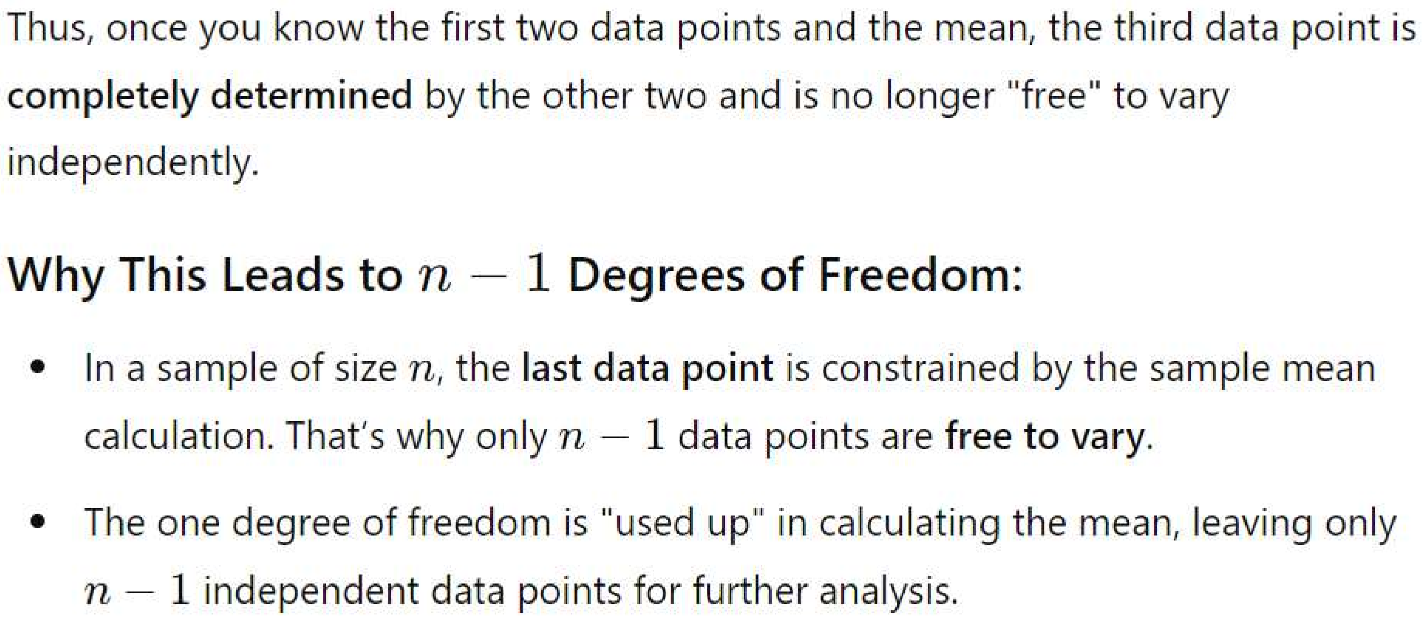
- Let's say $x_1 = 5$ and $x_2 = 7$.
- If the mean is fixed at 6, then x_3 must be exactly **6** to ensure the sum of the deviations from the mean is zero, because:

$$(5-6)+(7-6)+(x_3-6)=0$$

Simplifying:

$$-1+1+(x_3-6)=0$$

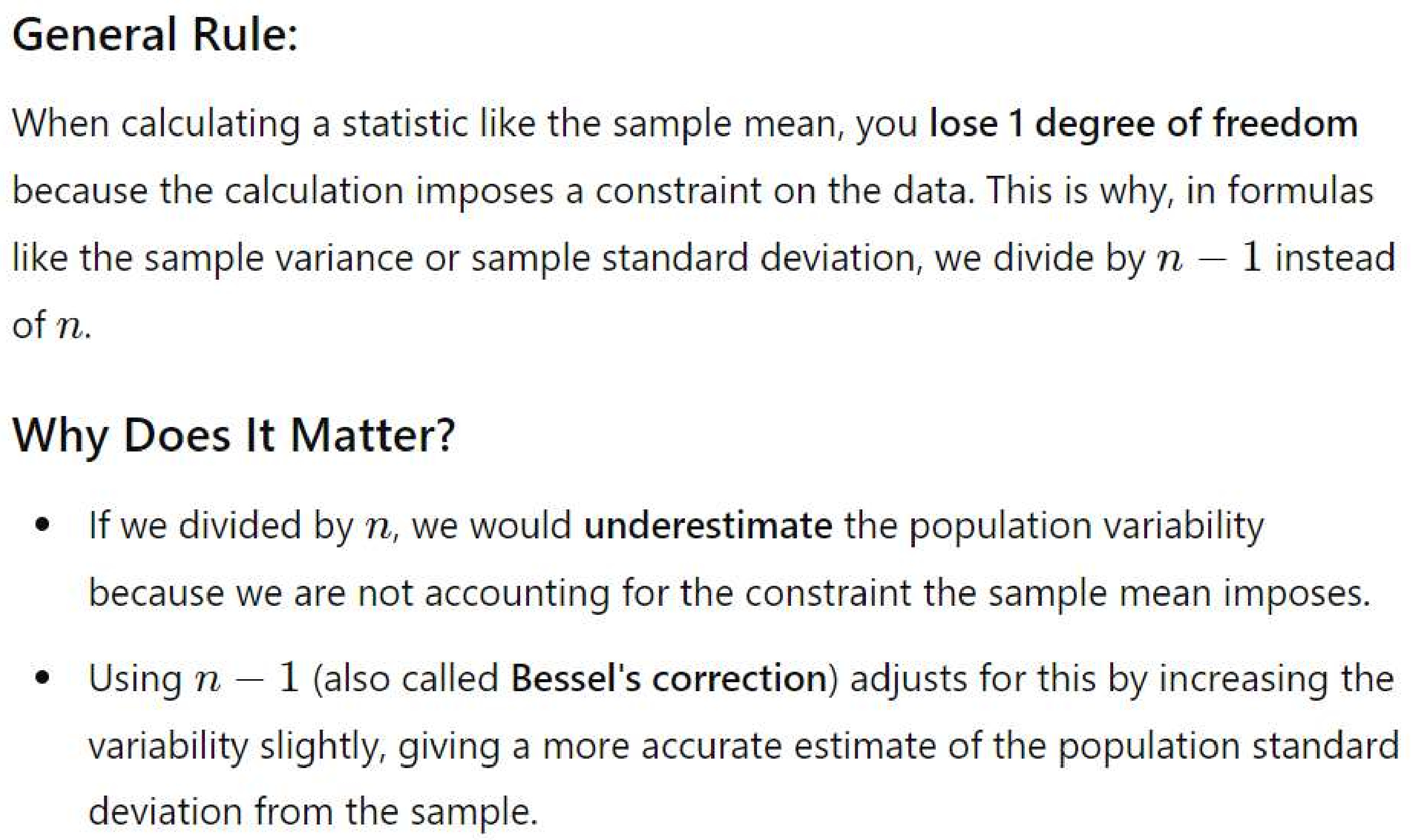
So $x_3 - 6 = 0$, meaning $x_3 = 6$.



Thus, once you know the first two data points and the mean, the third data point is **completely determined** by the other two and is no longer "free" to vary independently.

Why This Leads to $n-1$ Degrees of Freedom:

- In a sample of size n, the **last data point** is constrained by the sample mean calculation. That's why only $n-1$ data points are **free to vary**.
- The one degree of freedom is "used up" in calculating the mean, leaving only $n-1$ independent data points for further analysis.



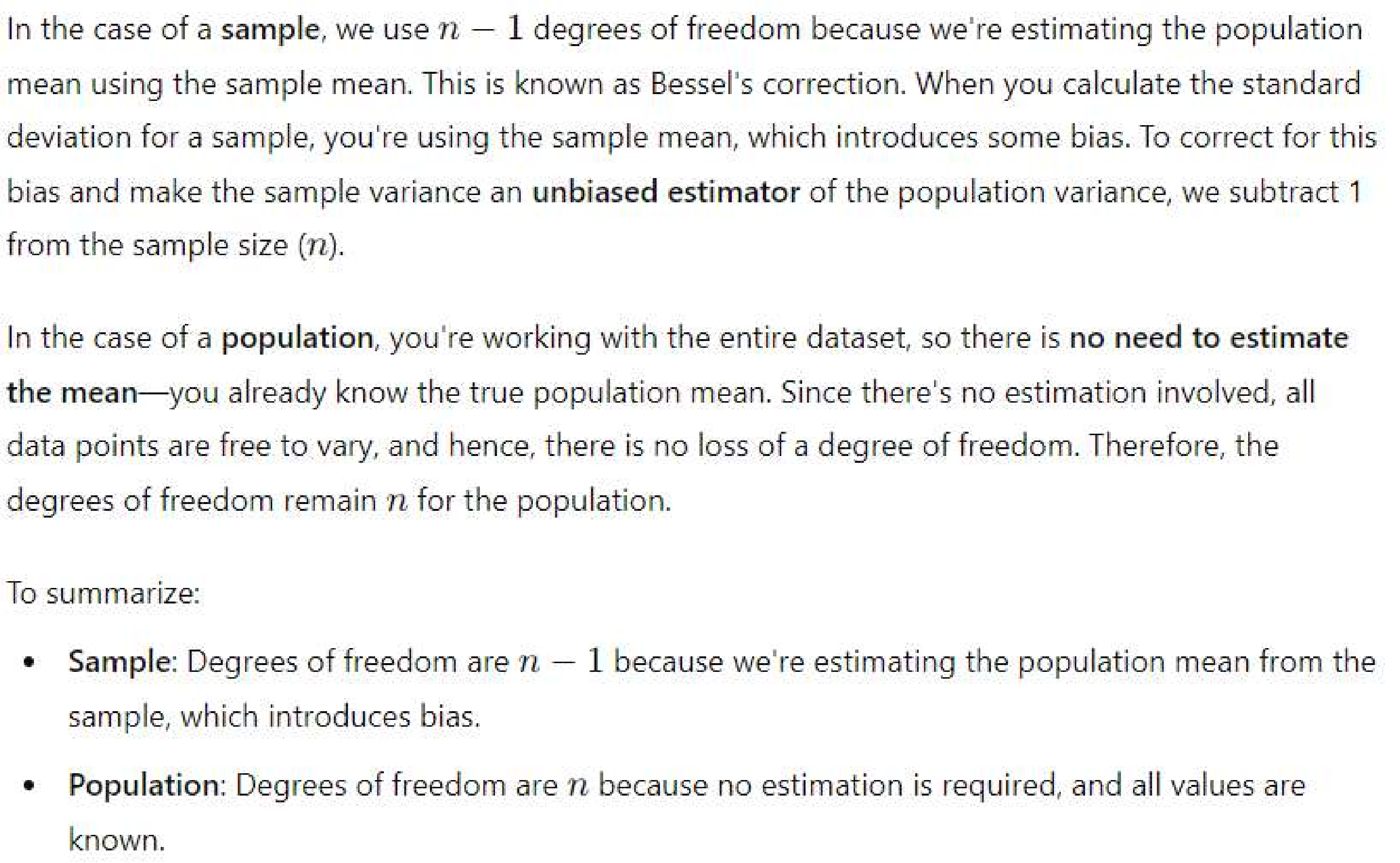
General Rule:

When calculating a statistic like the sample mean, you **lose 1 degree of freedom** because the calculation imposes a constraint on the data. This is why, in formulas like the sample variance or sample standard deviation, we divide by $n - 1$ instead of n.

Why Does It Matter?

- If we divided by n, we would **underestimate** the population variability because we are not accounting for the constraint the sample mean imposes.
- Using $n - 1$ (also called **Bessel's correction**) adjusts for this by increasing the variability slightly, giving a more accurate estimate of the population standard deviation from the sample.

Why not using (n-1) with population?



In the case of a **sample**, we use $n-1$ degrees of freedom because we're estimating the population mean using the sample mean. This is known as Bessel's correction. When you calculate the standard deviation for a sample, you're using the sample mean, which introduces some bias. To correct for this bias and make the sample variance an **unbiased estimator** of the population variance, we subtract 1 from the sample size (n).

In the case of a **population**, you're working with the entire dataset, so there is **no need to estimate the mean**—you already know the true population mean. Since there's no estimation involved, all data points are free to vary, and hence, there is no loss of a degree of freedom. Therefore, the degrees of freedom remain n for the population.

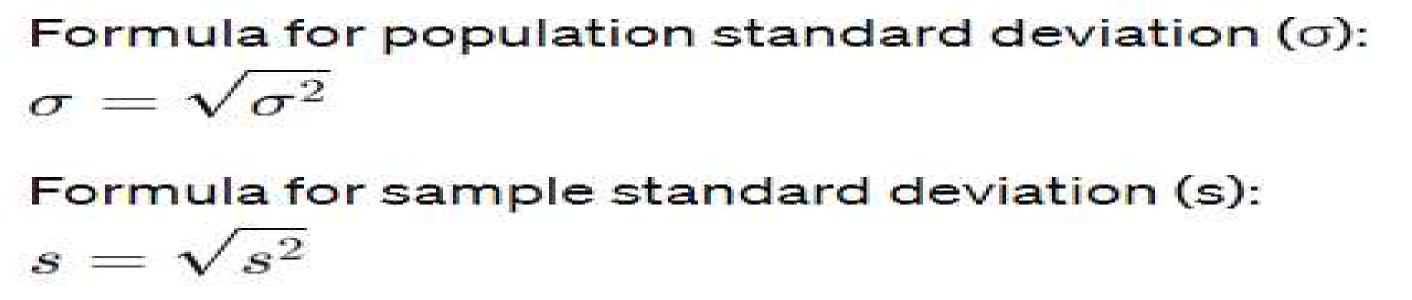
To summarize:

- **Sample**: Degrees of freedom are $n-1$ because we're estimating the population mean from the sample, which introduces bias.
- **Population**: Degrees of freedom are n because no estimation is required, and all values are known.

Standard Deviation:

- Now there another issue as my data points are linear but variance is squared to over come this problem we are taking square root of variance called **standard deviation** and is expressed in the same units as the original data.

- It measures the average distance of data points from the mean and provides a more interpretable measure of spread compared to variance.



Formula for population standard deviation (σ):

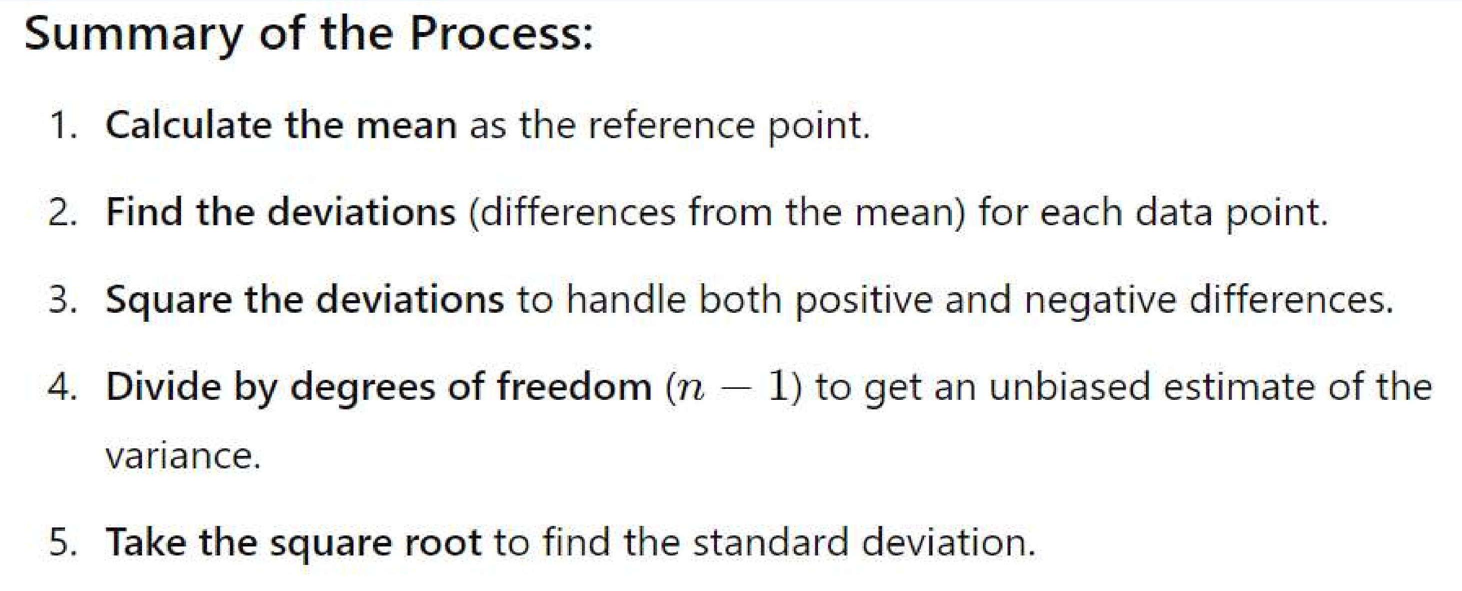
$$\sigma = \sqrt{\sigma^2}$$

Formula for sample standard deviation (s):

$$s = \sqrt{s^2}$$

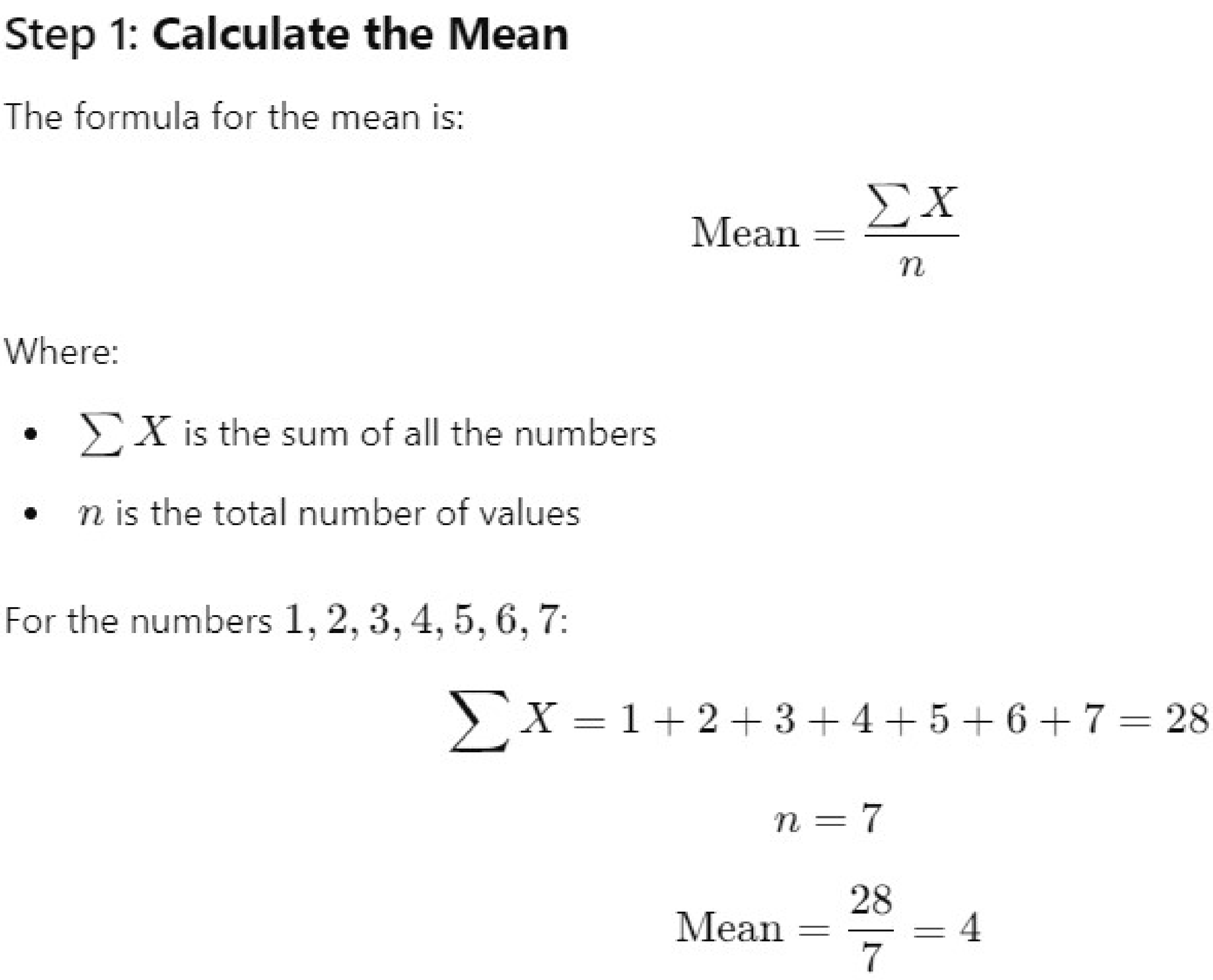
- **Example Scenario:** If the standard deviation of the employee salaries is $7,500, this indicates that on average, individual salaries deviate by $7,500 from the mean salary.
- This is more interpretable than variance because it is in the same units as the data.

Ques: Calculate the mean and standard deviation for the following sets of numbers: 1 2 3 4 5 6 7.



Summary of the Process:

1. **Calculate the mean** as the reference point.
2. **Find the deviations** (differences from the mean) for each data point.
3. **Square the deviations** to handle both positive and negative differences.
4. **Divide by degrees of freedom** $(n - 1)$ to get an unbiased estimate of the variance.
5. **Take the square root** to find the standard deviation.



Step 1: **Calculate the Mean**

The formula for the mean is:

$$\text{Mean} = \frac{\sum X}{n}$$

Where:

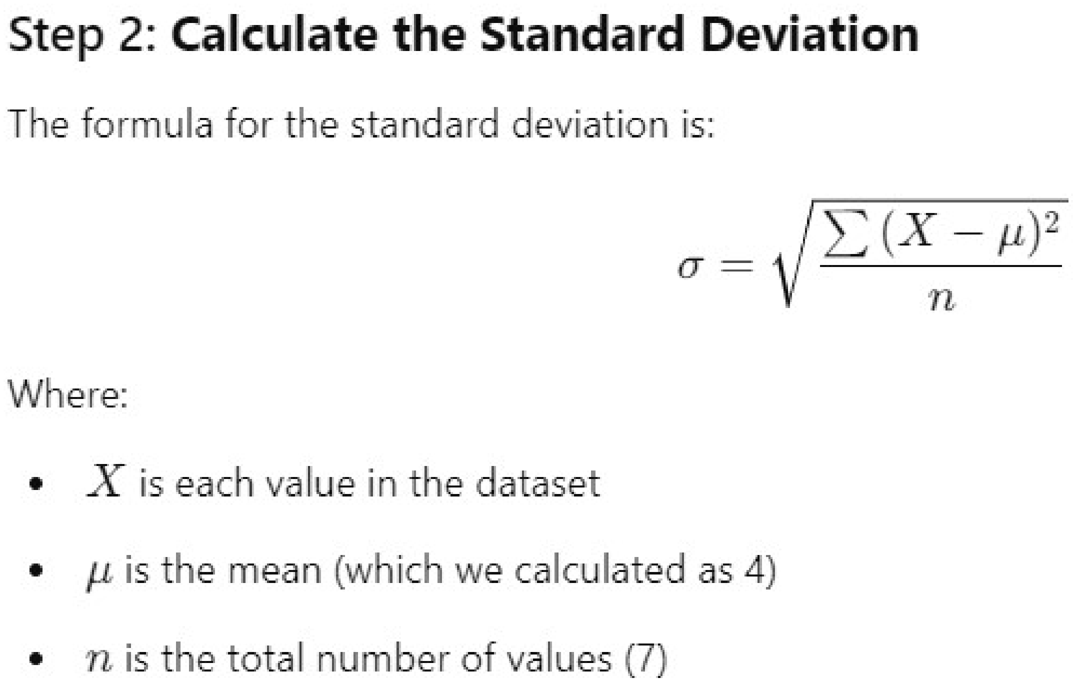
- $\sum X$ is the sum of all the numbers
- n is the total number of values

For the numbers $1, 2, 3, 4, 5, 6, 7$:

$$\sum X = 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$$

$$n = 7$$

$$\text{Mean} = \frac{28}{7} = 4$$



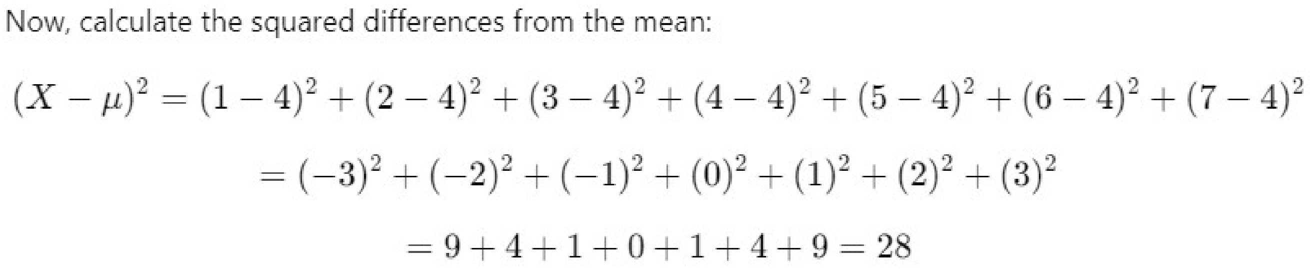
Step 2: **Calculate the Standard Deviation**

The formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{n}}$$

Where:

- X is each value in the dataset
- μ is the mean (which we calculated as 4)
- n is the total number of values (7)

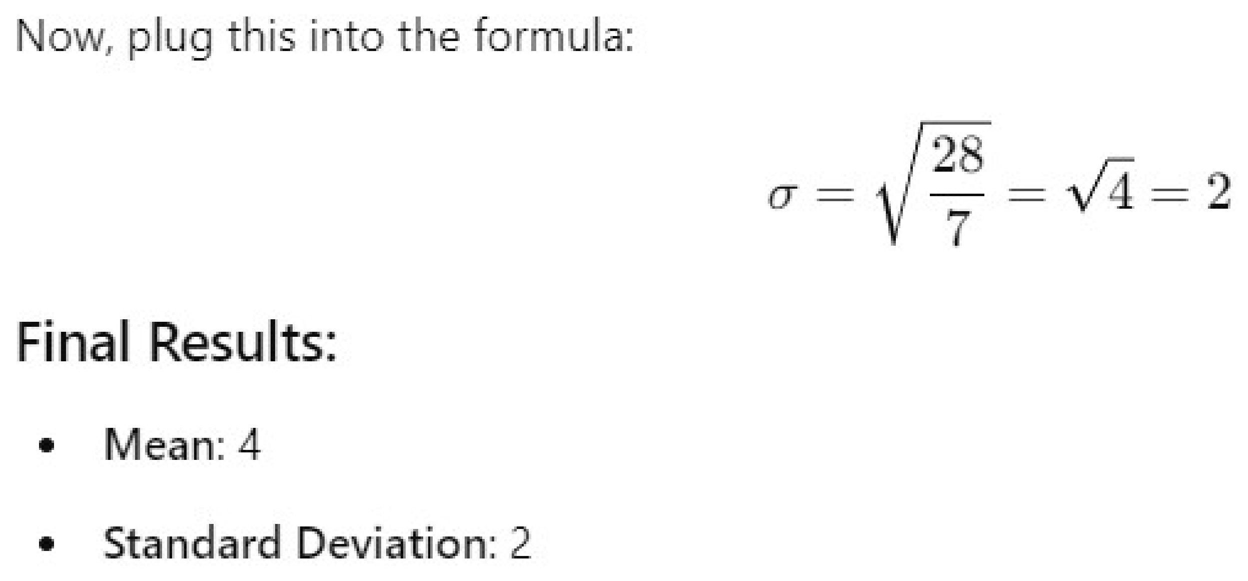


Now, calculate the squared differences from the mean:

$$(X-\mu)^2 = (1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2$$

$$= (-3)^2 + (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 + (3)^2$$

$$= 9 + 4 + 1 + 0 + 1 + 4 + 9 = 28$$



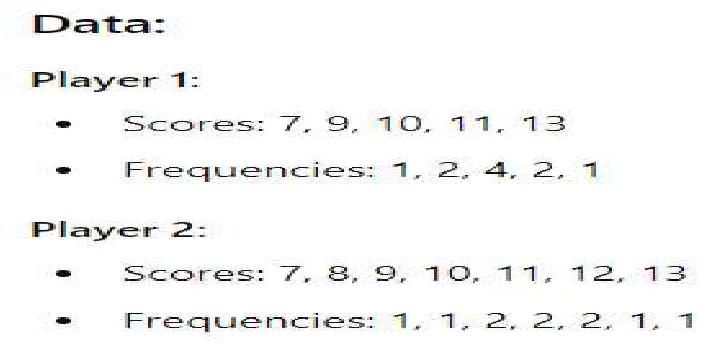
Now, plug this into the formula:

$$\sigma = \sqrt{\frac{28}{7}} = \sqrt{4} = 2$$

Final Results:

- Mean: 4
- Standard Deviation: 2

- **Ques:** The mean for two player is 10. Your job is to play like you're the coach, and work out the standard deviation for each player.
- Which player is the most reliable one for your team?
- (a). Score 7 9 10 11 13, Frequency 1 2 4 2 1,
- (b). Score 7 8 9 10 11 12 13 Frequency 1 1 2 2 2 1 1.
- To determine the reliability of each player, we need to calculate the standard deviation of their scores.
- ***The player with the lower standard deviation is considered more reliable.***



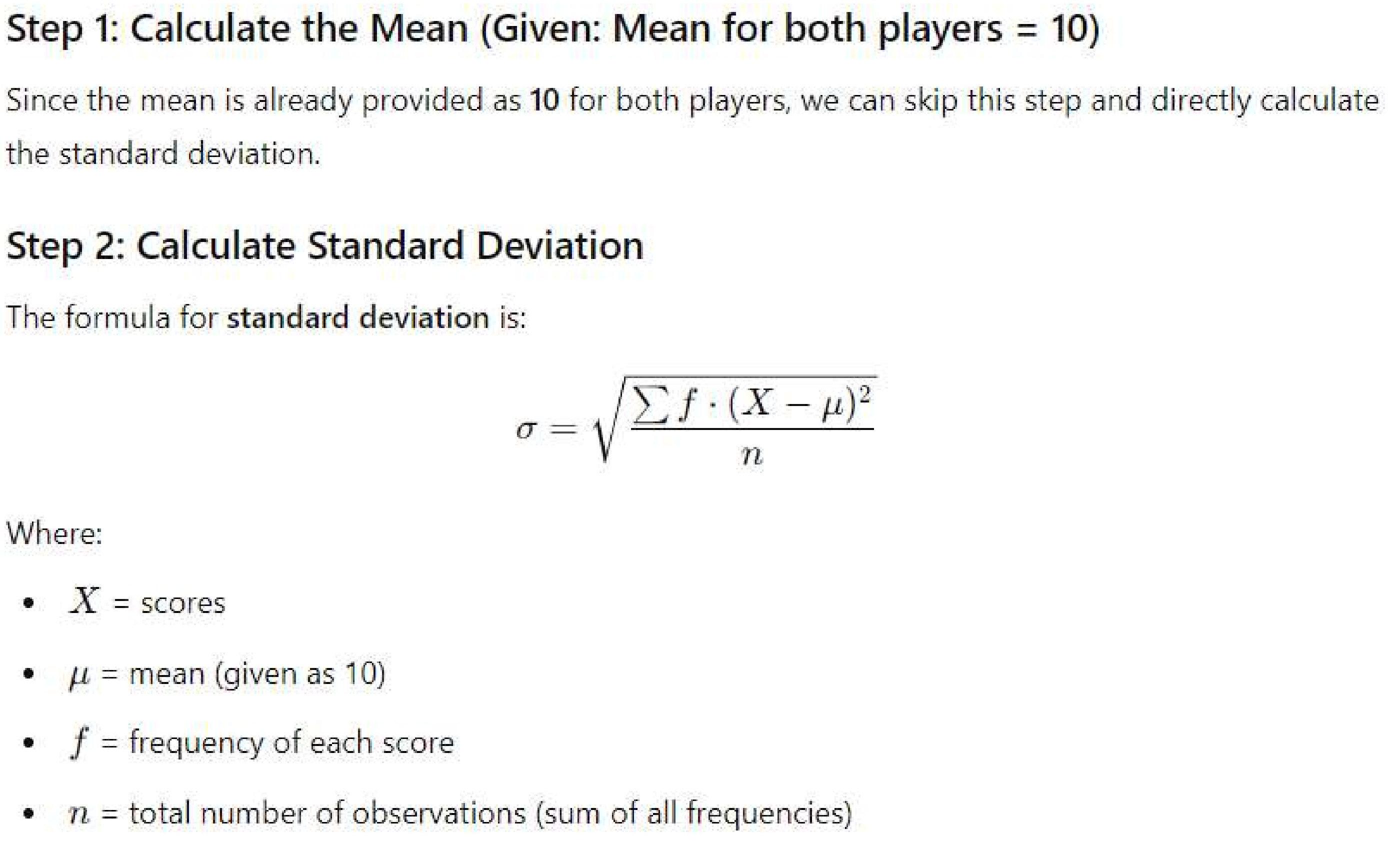
Data:

Player 1:

- Scores: 7, 9, 10, 11, 13
- Frequencies: 1, 2, 4, 2, 1

Player 2:

- Scores: 7, 8, 9, 10, 11, 12, 13
- Frequencies: 1, 1, 2, 2, 2, 1, 1



Step 1: Calculate the Mean (Given: Mean for both players = 10)

Since the mean is already provided as **10** for both players, we can skip this step and directly calculate the standard deviation.

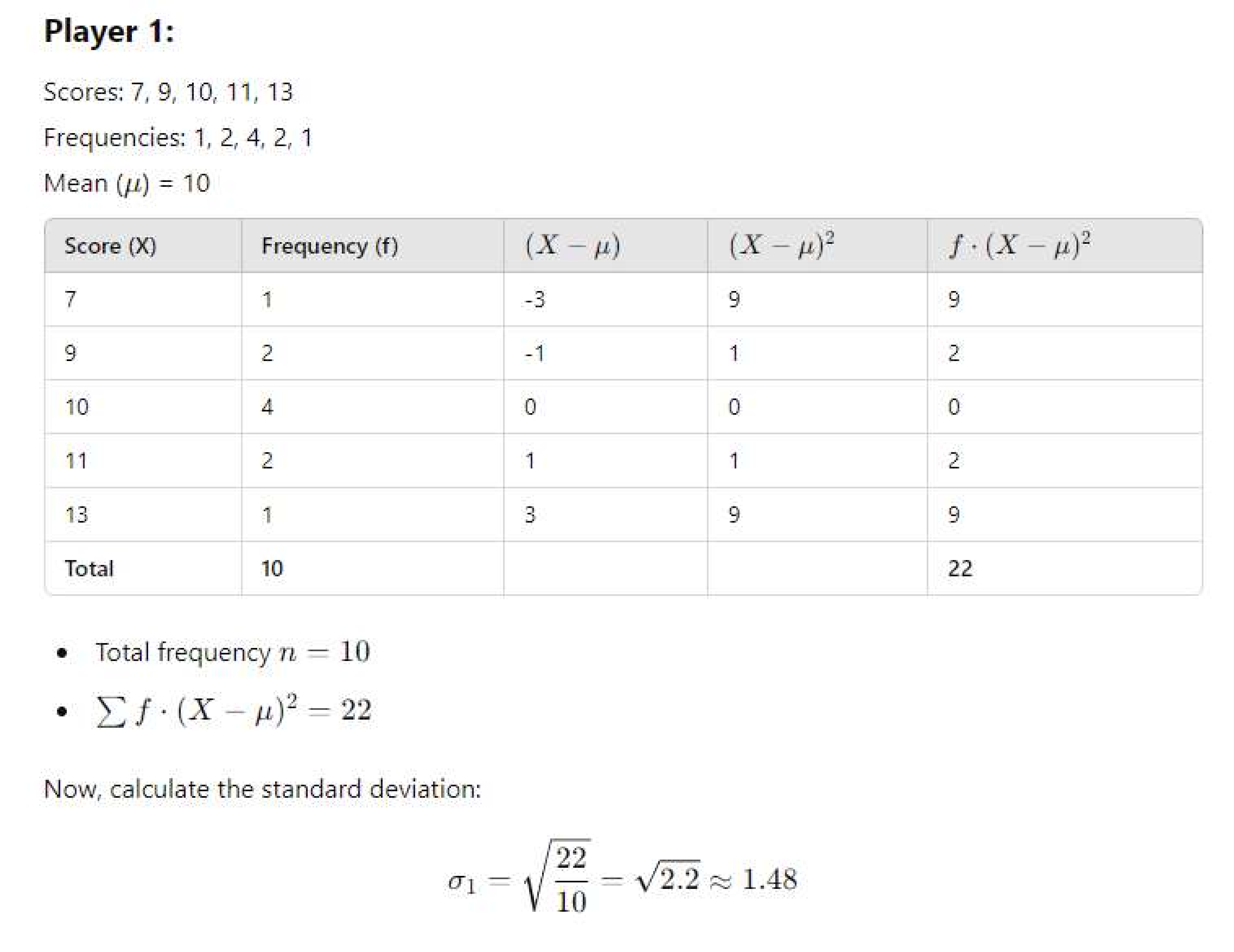
Step 2: Calculate Standard Deviation

The formula for **standard deviation** is:

$$\sigma = \sqrt{\frac{\sum f \cdot (X - \mu)^2}{n}}$$

Where:

- X = scores
- μ = mean (given as 10)
- f = frequency of each score
- n = total number of observations (sum of all frequencies)



Player 1:

Scores: 7, 9, 10, 11, 13

Frequencies: 1, 2, 4, 2, 1

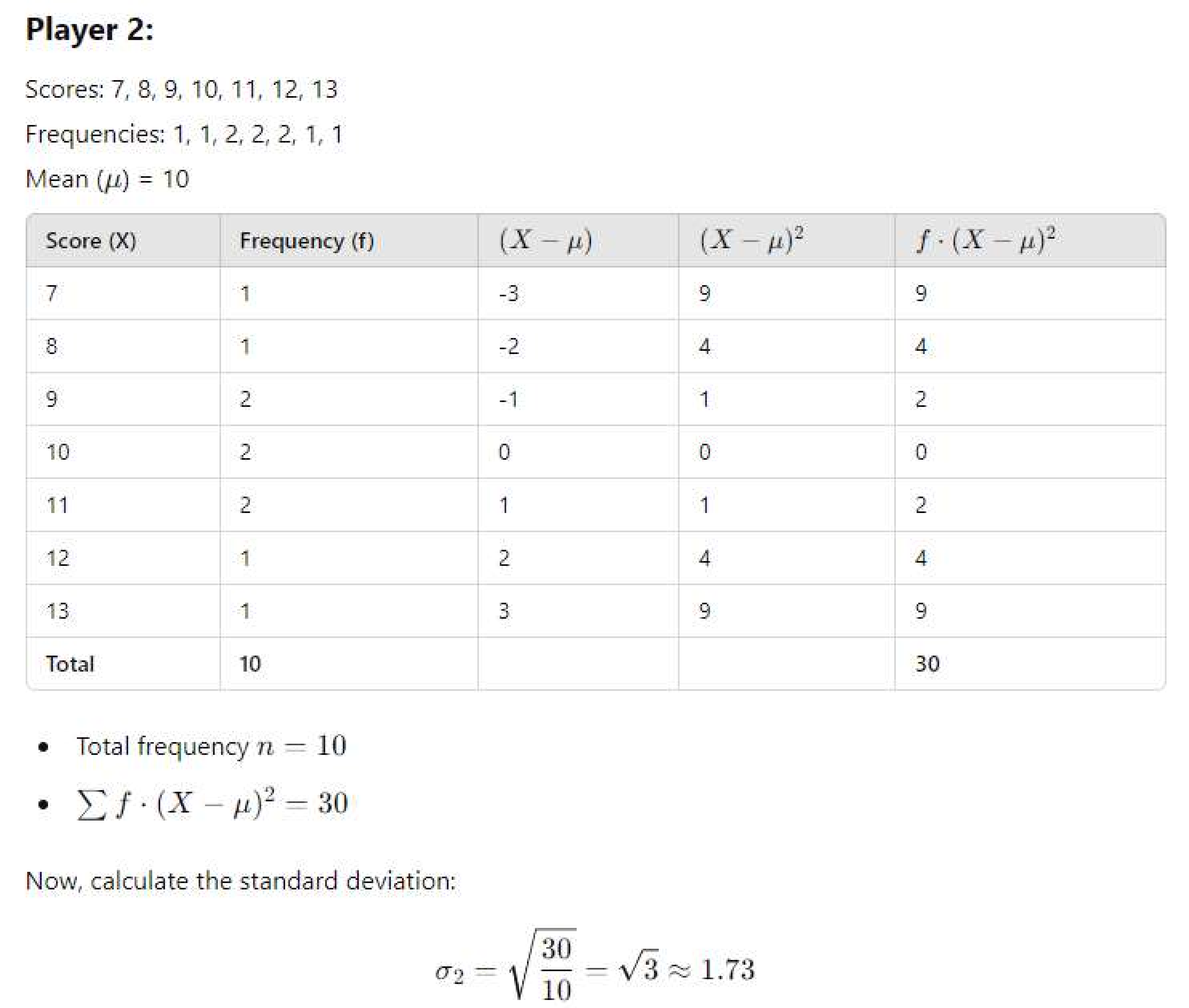
Mean (μ) = 10

Score (X)	Frequency (f)	$(X - \mu)$	$(X - \mu)^2$	$f \cdot (X - \mu)^2$
7	1	-3	9	9
9	2	-1	1	2
10	4	0	0	0
11	2	1	1	2
13	1	3	9	9
Total	**10**			**22**

- Total frequency $n = 10$
- $\sum f \cdot (X - \mu)^2 = 22$

Now, calculate the standard deviation:

$$\sigma_1 = \sqrt{\frac{22}{10}} = \sqrt{2.2} \approx 1.48$$



Player 2:

Scores: 7, 8, 9, 10, 11, 12, 13

Frequencies: 1, 1, 2, 2, 2, 1, 1

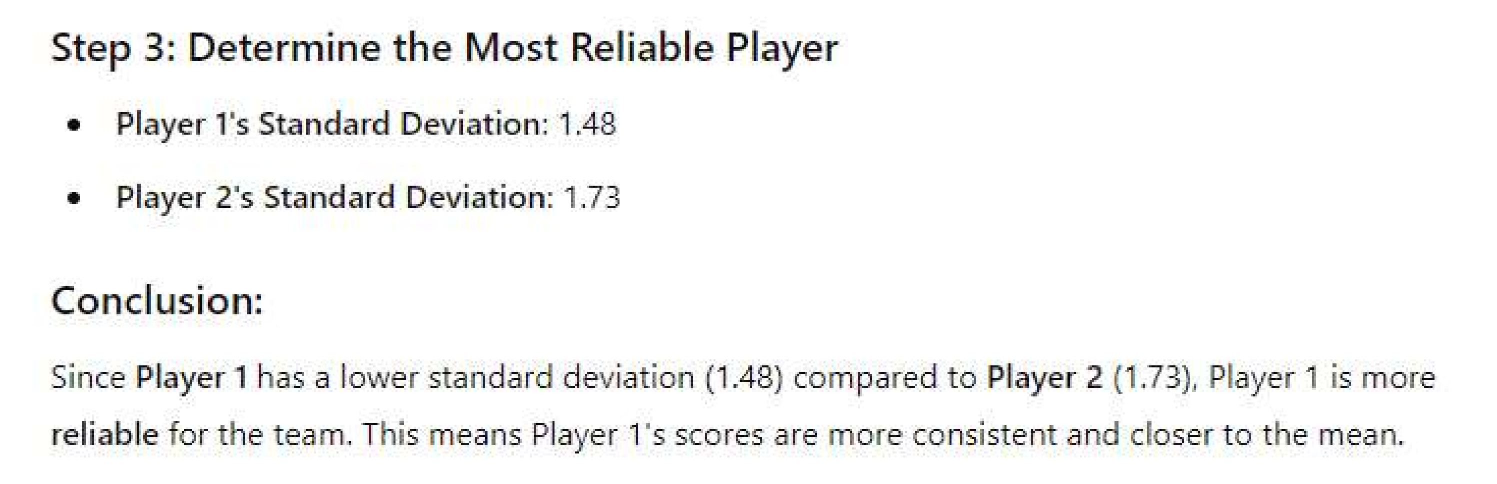
Mean (μ) = 10

Score (X)	Frequency (f)	$(X - \mu)$	$(X - \mu)^2$	$f \cdot (X - \mu)^2$
7	1	-3	9	9
8	1	-2	4	4
9	2	-1	1	2
10	2	0	0	0
11	2	1	1	2
12	1	2	4	4
13	1	3	9	9
Total	10			30

- Total frequency $n = 10$
- $\sum f \cdot (X - \mu)^2 = 30$

Now, calculate the standard deviation:

$$\sigma_2 = \sqrt{\frac{30}{10}} = \sqrt{3} \approx 1.73$$



Step 3: Determine the Most Reliable Player

- **Player 1's Standard Deviation:** 1.48
- **Player 2's Standard Deviation:** 1.73

Conclusion:

Since **Player 1** has a lower standard deviation (1.48) compared to **Player 2** (1.73), Player 1 is more **reliable** for the team. This means Player 1's scores are more consistent and closer to the mean.

Data points arranged in increasing / ascending order

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

First step: Find mean (µ) 31

Second step: find deviation of each point x from mean (x - µ)

Third step: square each deviation (x-µ)^2

Fourth step: Add all squared deviations ∑(x - µ)^2

Fifth step: divide it by n

Sixth step: Find square root of it

x	(x - µ)	(x-µ)^2
10		
11		
12		
25		
25		
27		
31		
33		
34		
34		
35		
36		
43		
50		
59		

∑(x - µ)^2 = 2642

Variance **∑(x - µ)^2 / n =** 176.13

Standard deviation **sqrt(∑(x - µ)^2 / n) =** 13.27

HR Analytics Example: Employee Salaries

- In **HR analytics**, you might analyze the **salary distribution** of employees in a company.
- **Scenario:**
  - You’re working for a company that wants to understand the **salary dispersion** in different departments to ensure fair compensation practices. You collect salary data for employees in the marketing department and calculate the **mean salary**.

Standard Deviation and Variance in Action

- **Mean salary**: Let’s say the average salary in the marketing department is $60,000.
- **Variance**: The variance shows how much the salaries deviate from the average salary. If the variance is large, it indicates that some employees are paid significantly more or less than the mean, suggesting **inequities** in pay.
- **Standard deviation**: If the **standard deviation** is $5,000, it tells you that, on average, employees' salaries deviate from the mean by $5,000.
- A low standard deviation means salaries are fairly uniform, while a high standard deviation could indicate discrepancies that need further investigation, such as pay gaps between senior and junior employees, gender, or different job roles.

Interpretation:

- If the standard deviation is **low** (say $2,000), this would indicate that most employees have salaries close to the mean of $60,000, implying pay equality.
- On the other hand, if the standard deviation is **high** (say $10,000), this indicates a wide salary range, which may require further analysis of compensation practices to identify potential inequalities.

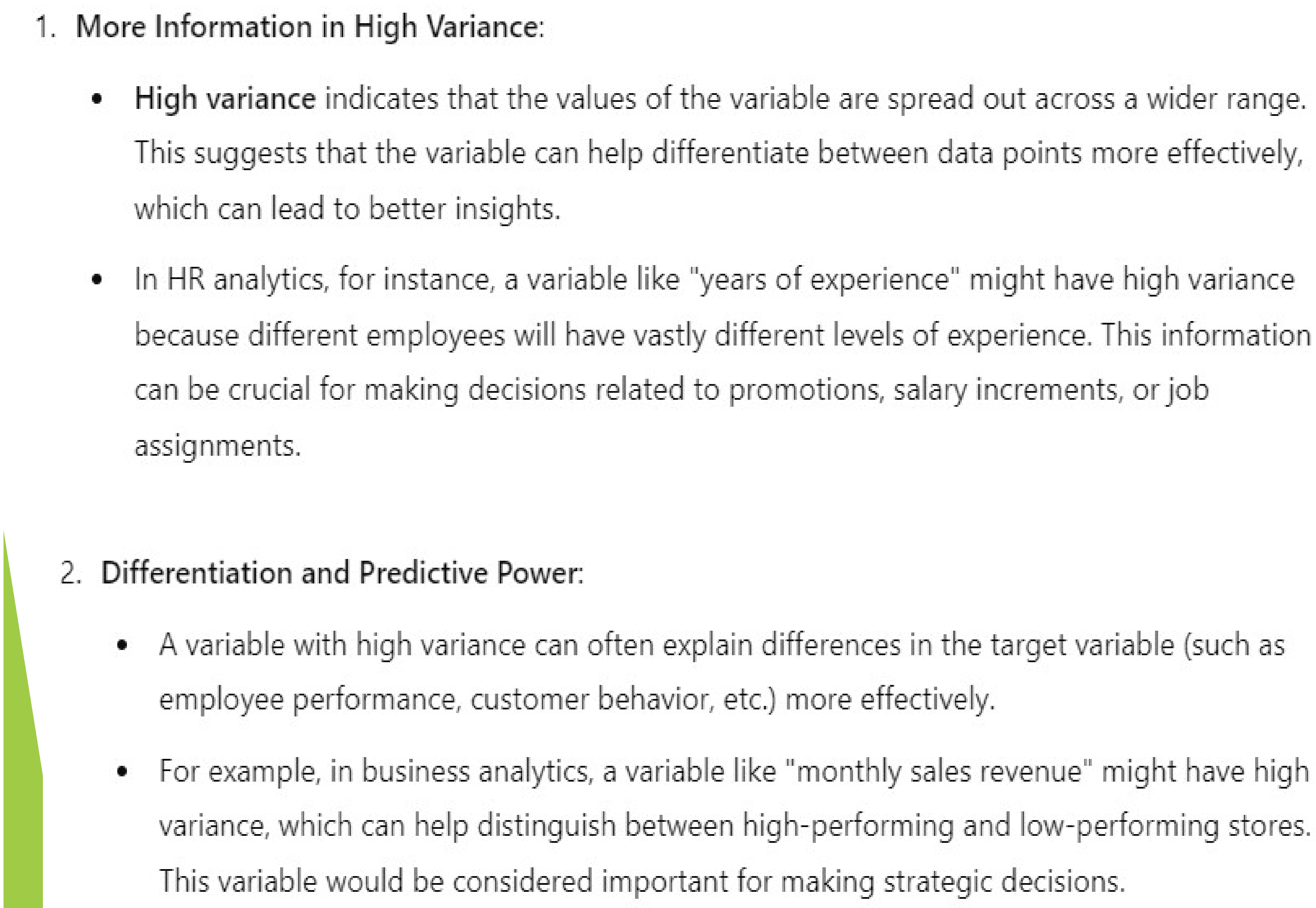
Business Analytics Example: Sales Revenue

- In **business analytics**, you might analyze the **monthly sales revenue** across different stores.
- **Scenario:**
- You are analyzing the **monthly sales revenue** of 10 stores in a retail chain over the past year to understand the consistency of performance.
- **Standard Deviation and Variance in Action:**
- **Mean monthly sales**: Suppose the average monthly sales across stores is $100,000.
- **Variance**: If the variance in monthly sales is **high**, it shows that some stores are performing much better or worse than others, indicating **inconsistent store performance**.
- **Standard deviation**: If the standard deviation is $20,000, it means that, on average, the monthly sales of a store deviate from the mean by $20,000.

Interpretation:

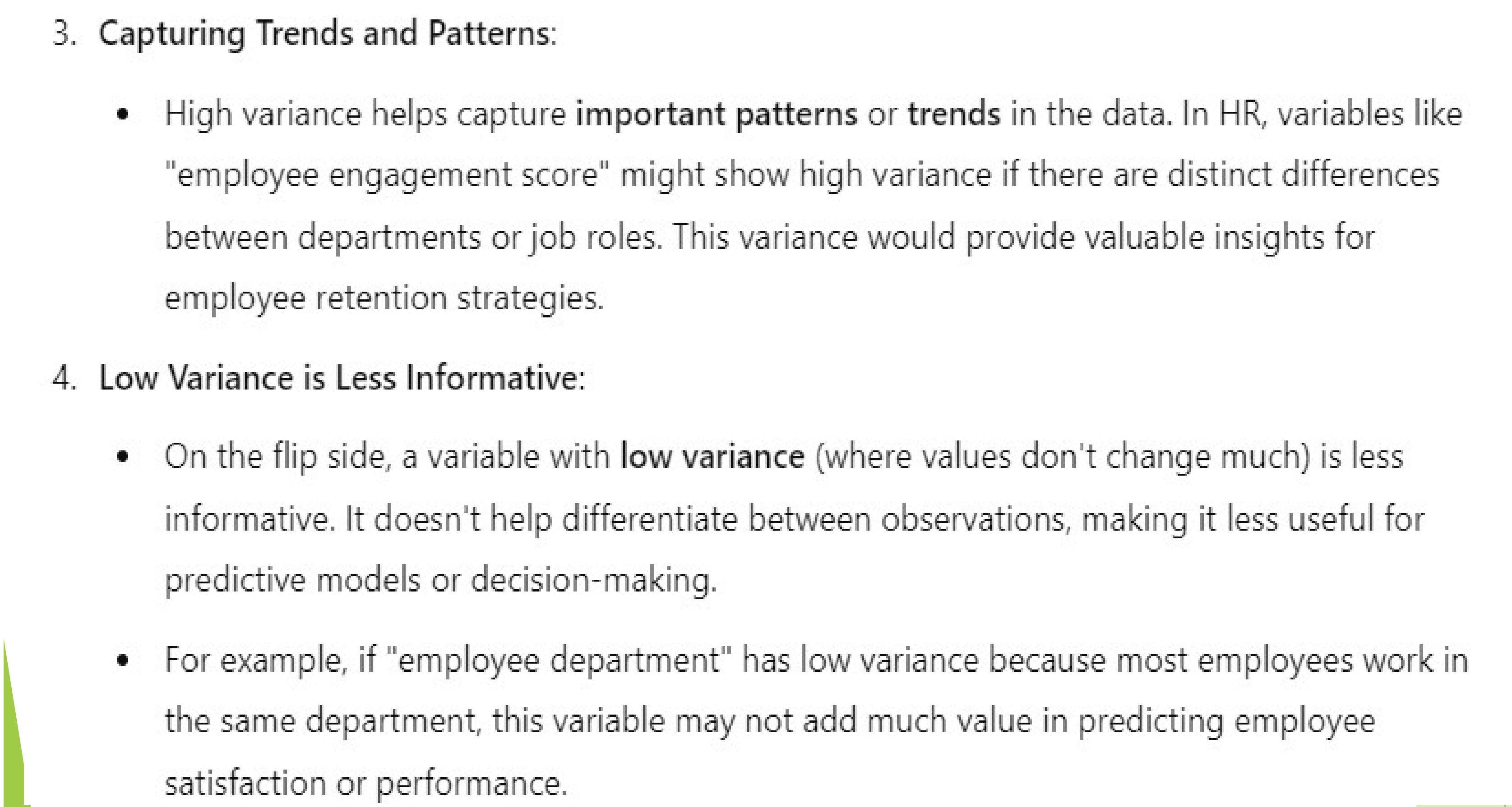
- If the standard deviation is **low** (e.g., $5,000), it means that most stores are generating sales close to the average of $100,000, implying **stable and consistent performance.**
- However, if the standard deviation is **high** (e.g., $30,000), it suggests significant variation in sales between stores.
- This might lead you to investigate the causes—such as differences in marketing strategies, location advantages, or store management practices—and take corrective actions to improve consistency.

Importance of High Variance:

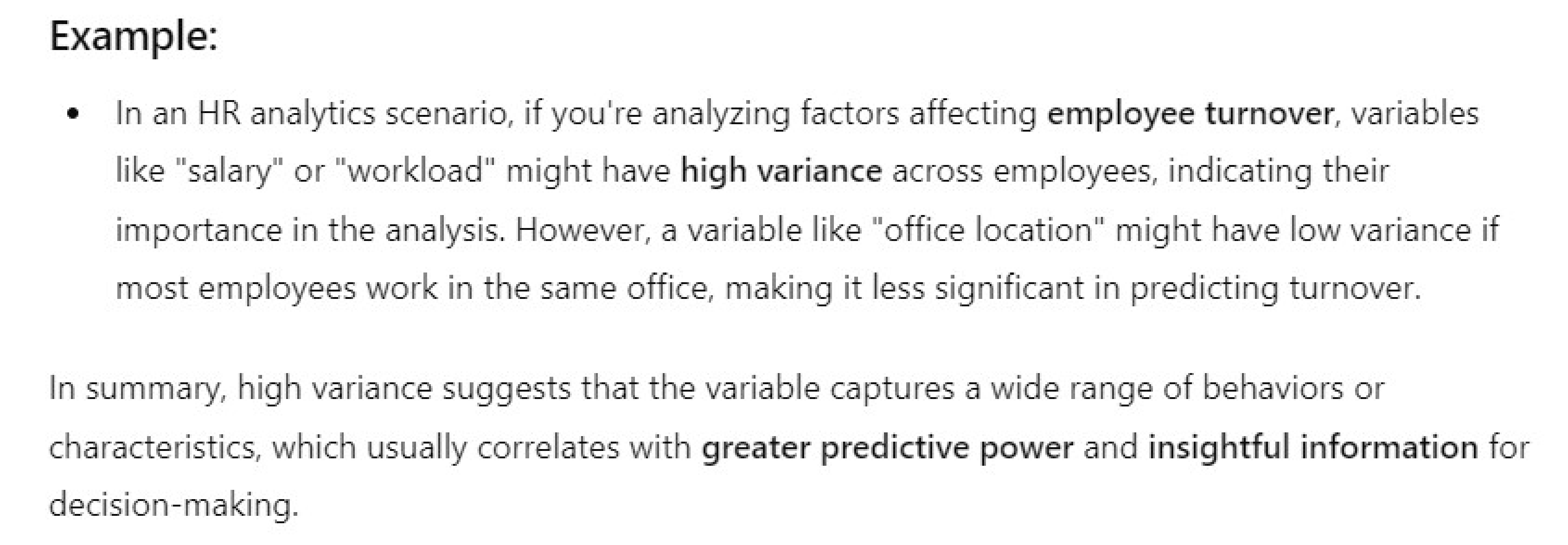


1. **More Information in High Variance**:
   - **High variance** indicates that the values of the variable are spread out across a wider range. This suggests that the variable can help differentiate between data points more effectively, which can lead to better insights.
   - In HR analytics, for instance, a variable like "years of experience" might have high variance because different employees will have vastly different levels of experience. This information can be crucial for making decisions related to promotions, salary increments, or job assignments.

2. **Differentiation and Predictive Power**:
   - A variable with high variance can often explain differences in the target variable (such as employee performance, customer behavior, etc.) more effectively.
   - For example, in business analytics, a variable like "monthly sales revenue" might have high variance, which can help distinguish between high-performing and low-performing stores. This variable would be considered important for making strategic decisions.



3. **Capturing Trends and Patterns:**
   - High variance helps capture **important patterns** or **trends** in the data. In HR, variables like "employee engagement score" might show high variance if there are distinct differences between departments or job roles. This variance would provide valuable insights for employee retention strategies.
4. **Low Variance is Less Informative:**
   - On the flip side, a variable with **low variance** (where values don't change much) is less informative. It doesn't help differentiate between observations, making it less useful for predictive models or decision-making.
   - For example, if "employee department" has low variance because most employees work in the same department, this variable may not add much value in predicting employee satisfaction or performance.



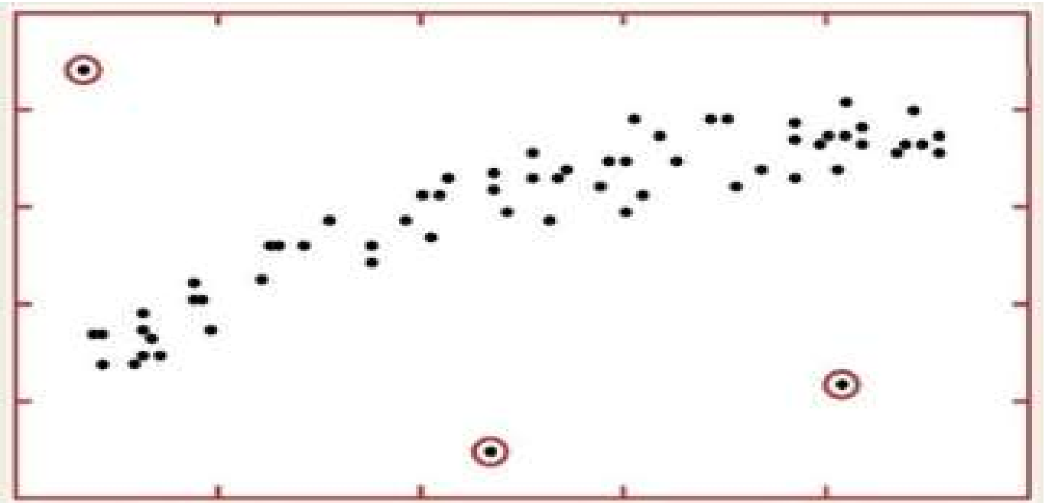
Example:

- In an HR analytics scenario, if you're analyzing factors affecting **employee turnover**, variables like "salary" or "workload" might have **high variance** across employees, indicating their importance in the analysis. However, a variable like "office location" might have low variance if most employees work in the same office, making it less significant in predicting turnover.

In summary, high variance suggests that the variable captures a wide range of behaviors or characteristics, which usually correlates with **greater predictive power** and **insightful information** for decision-making.

Outliers: Understanding Their Impact on Data

- An outlier is an observation that lies at an abnormal distance from other values in a dataset.
- It can be significantly higher or lower than the majority of data points.
- Whether the presence of outliers is good or harmful depends on the context and the type of analysis being conducted.



Types of Outliers

- **Univariate outliers**: These are outliers in a single variable.
- For example, an extremely high salary in a dataset of employee salaries.
- **Multivariate outliers**: These occur when the combination of two or more variables creates an unusual pattern.
- For instance, a person having a very high salary but working very few hours could be considered a multivariate outlier.

Causes of Outliers

- **Measurement errors**: Errors during data entry or recording, like typing mistakes or faulty sensors.
- **Sampling errors**: When the data doesn’t properly represent the population, some unusual values may arise.
- **Genuine outliers**: Sometimes, outliers represent real, rare phenomena in the dataset.
- For example, a company’s top-performing employee with significantly higher sales.

Is the Presence of Outliers Good or Harmful?

- **Harmful Impacts of Outliers**
- **Distortion of statistical measures:** Outliers can significantly distort summary statistics like the **mean** and **standard deviation.** This can mislead conclusions.
- **Example:** Suppose a dataset of salaries has most values between $50,000 and $80,000, but one employee earns $500,000. The **mean** salary would increase dramatically, giving a false impression of the average pay.
- **In this case**, the outlier may need to be removed to provide a better understanding of the central tendency.

- **Misleading visualizations**: Outliers can affect the scale of charts (like histograms, box plots) and make it difficult to visualize the distribution of data.
- **Example:** In a bar chart of store revenues, one store earning 10 times more than others will stretch the graph, making differences between the other stores appear insignificant.
- **Decreased model performance**: Many machine learning algorithms (like linear regression) are sensitive to outliers. Outliers can skew the model’s predictions and reduce accuracy.
- **Example:** In a house price prediction model, one very expensive house could distort the model’s understanding of the relationship between house features and price, leading to poor predictions for other houses.

When Outliers Can Be Beneficial

- **Identification of valuable insights**: Outliers may represent significant or unusual phenomena that deserve attention, such as fraud detection, innovation, or exceptional performance.
  - Example: In fraud detection, a very high transaction amount compared to usual transactions can indicate suspicious activity that needs to be investigated.
- **Highlighting variability**: In some cases, outliers reveal genuine variability in the data, especially in industries where extreme performance is possible.
  - Example: In sports analytics, an athlete's exceptional performance (e.g., breaking a world record) is an outlier, but it's an important one that reflects the athlete's capabilities.

- **Discovering new trends**: Outliers may indicate emerging trends or shifts in data patterns that require attention.
    - **Example:** A sudden spike in a company’s online sales data could be an outlier, but it may also indicate the beginning of a new trend or market demand that needs further exploration.

How to Handle Outliers?

- **Investigate the Cause**
  - First, determine if the outlier is due to a **data entry error**, **measurement error**, or if it’s a genuine observation.
    - If it’s a mistake, correct or remove it.
    - If it’s genuine, decide whether to retain it or remove it based on the impact on your analysis.
- **Use Robust Statistical Measures**
  - Instead of relying on sensitive measures like the mean, use **robust statistics** such as the **median** or **interquartile range (IQR)**,which are less affected by outliers.
    - Example: If a salary dataset has extreme outliers, the **median salary** gives a better representation of central tendency than the mean.

- **Transform or Cap Outliers**
  - You can apply transformations (e.g., **log transformations**) to reduce the influence of outliers.
  - Alternatively, **capping** or **trimming** outliers can limit their impact on your analysis.
    - **Example:** If a small percentage of house prices are extremely high, you might cap them at a reasonable threshold to prevent them from skewing the analysis.

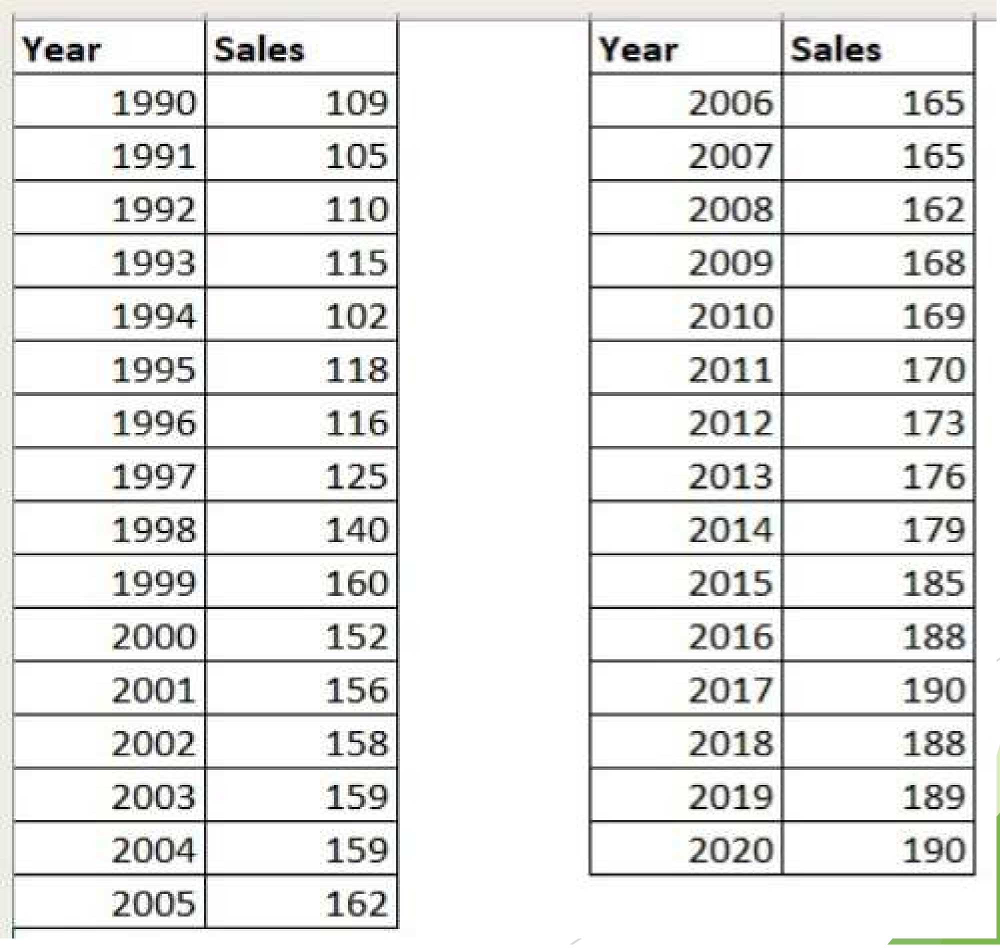
Data Visualization Basics

- What is **data Visualization**?
- **Data visualization** is the representation of data or information in a graph, chart, or other visual format.
- It communicates relationships of the data with images.
- This is important because it allows trends and patterns to be more easily seen.

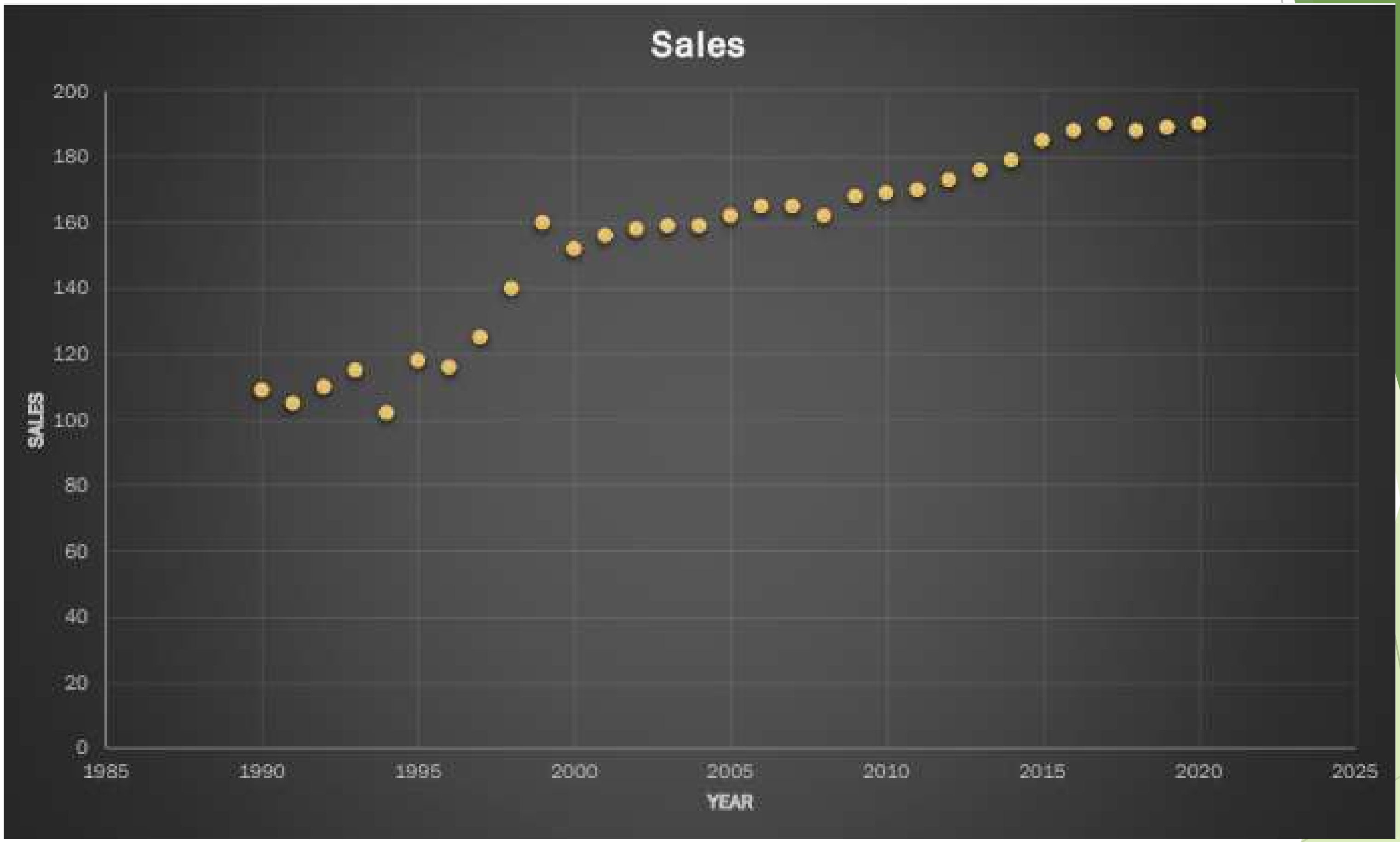
What is its importance?

“One Picture speaks better than one thousand words”

- As in this example:



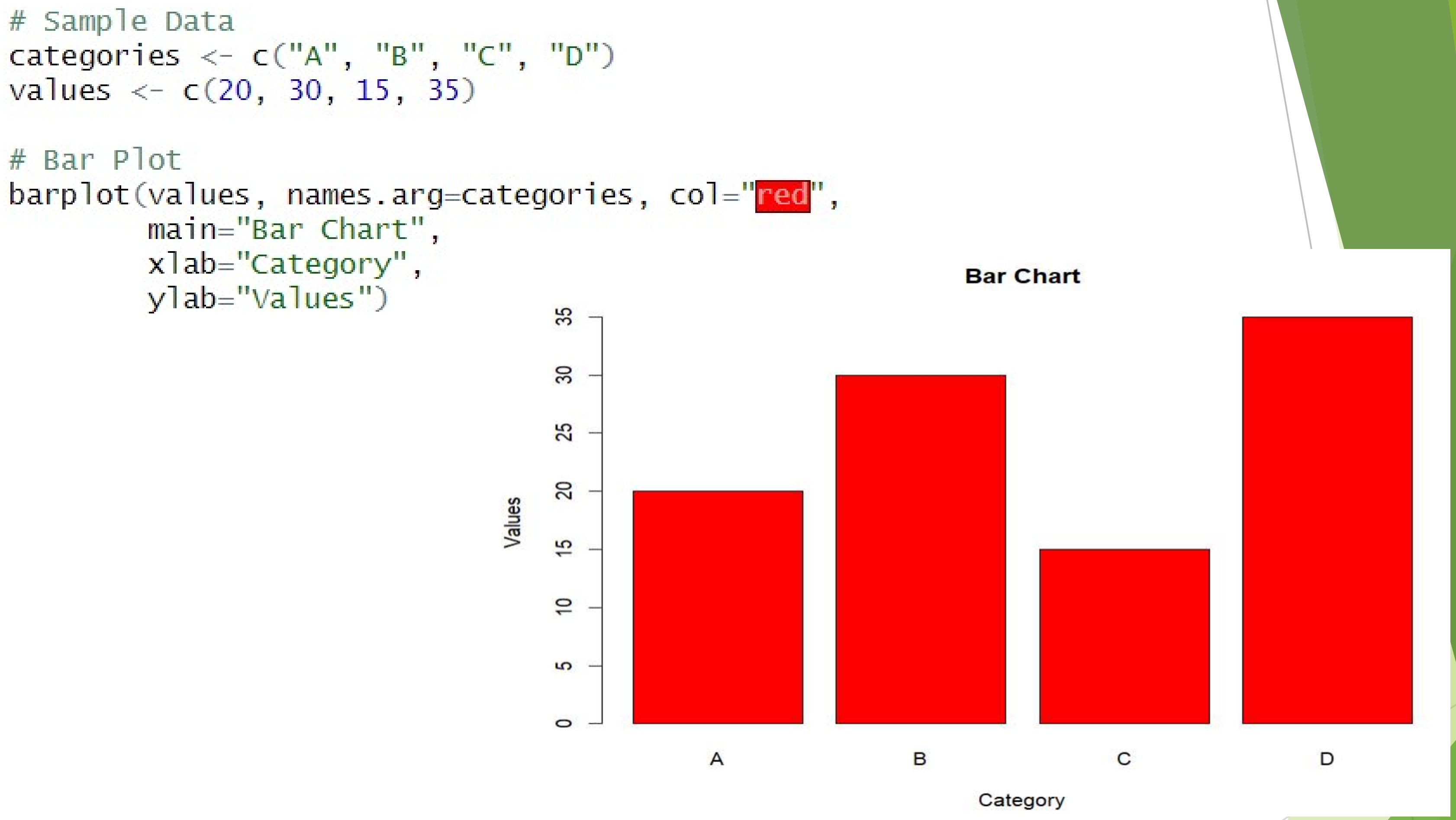
Year	Sales
1990	109
1991	105
1992	110
1993	115
1994	102
1995	118
1996	116
1997	125
1998	140
1999	160
2000	152
2001	156
2002	158
2003	159
2004	159
2005	162
2006	165
2007	165
2008	162
2009	168
2010	169
2011	170
2012	173
2013	176
2014	179
2015	185
2016	188
2017	190
2018	188
2019	189
2020	190


Sales
SALES
200
180
160
140
120
100
80
60
40
20
0
1985
1990
1995
2000
2005
2010
2015
2020
2025
YEAR

Different Types of Visual Format

- **Bar Chart**: Categorical data distribution.
- **Histogram**: Distribution of a continuous variable.
- **Pie Chart**: Proportion of categories.
- **Box Plot**: Summarizes the distribution of a continuous variable, highlighting outliers.
- **Scatter Plot**: Relationship between two variables.
- **Line Chart**: Trends over time.
- **Density Plot**: Estimated distribution of a continuous variable.
- **Correlation Matrix Plot**: Shows correlations between multiple variables.

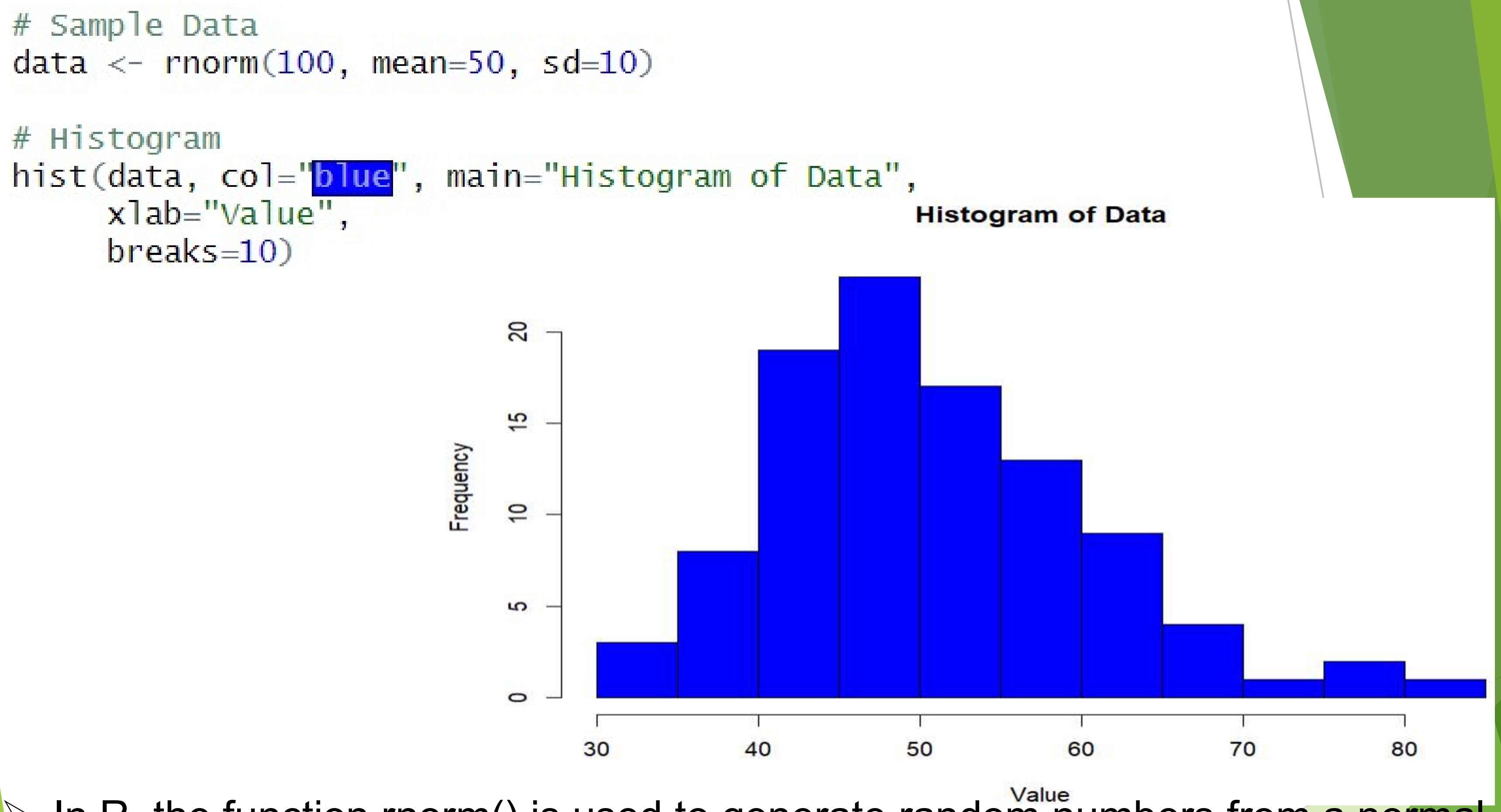
Bar Plot using R



```
# Sample Data
categories <- c("A", "B", "C", "D")
values <- c(20, 30, 15, 35)

# Bar Plot
barplot(values, names.arg=categories, col="red",
        main="Bar Chart",
        xlab="Category",
        ylab="Values")
```

- ➢ In R, the symbol <- is the assignment operator.
- ➢ It is used to assign values to variables.
- ➢ It works similarly to the equals sign = in many other programming languages, but is preferred in R for assignment in most cases.

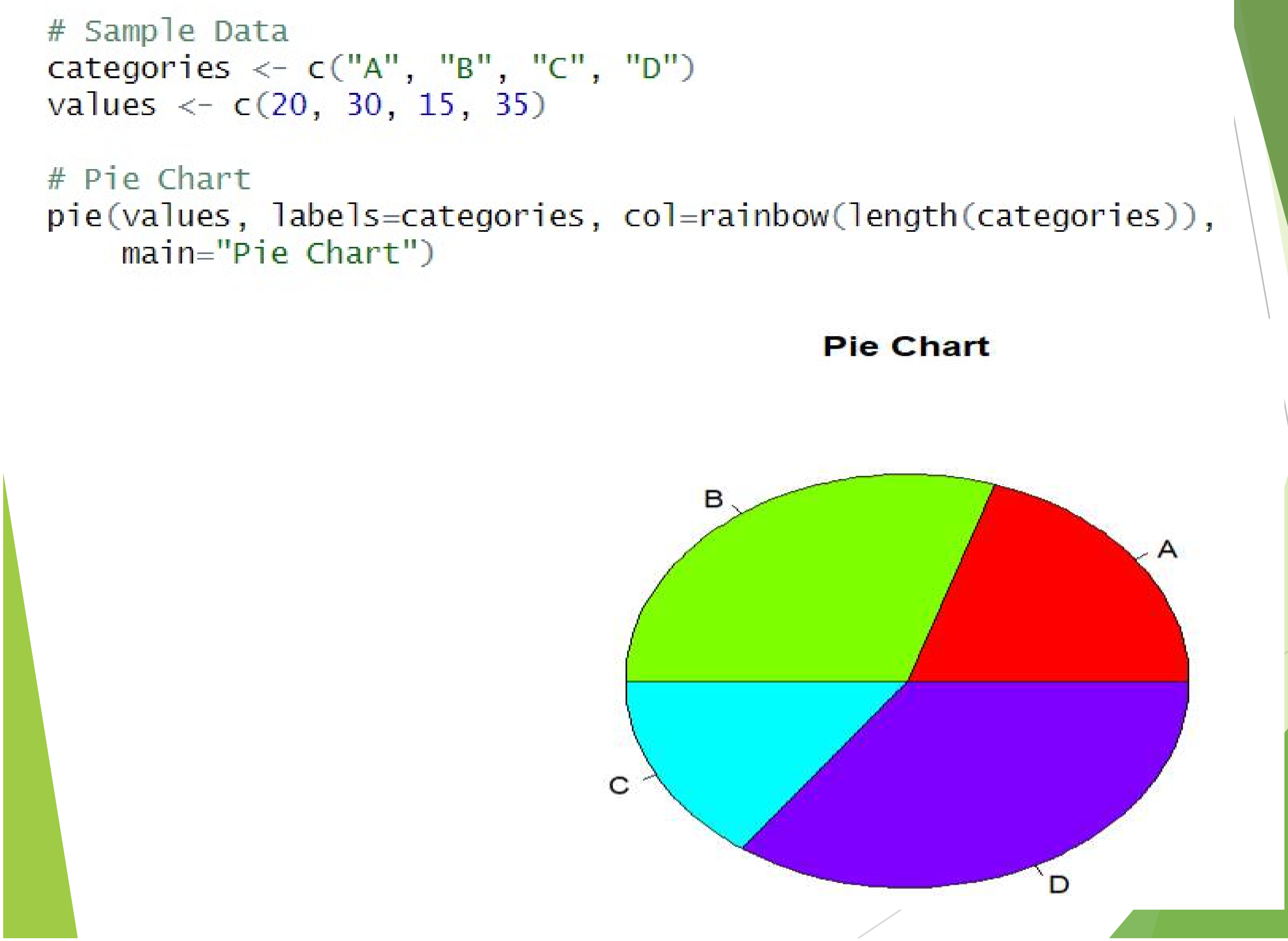
Histogram using R



```
# Sample Data
data <- rnorm(100, mean=50, sd=10)

# Histogram
hist(data, col="blue", main="Histogram of Data",
     xlab="Value",
     breaks=10)
```

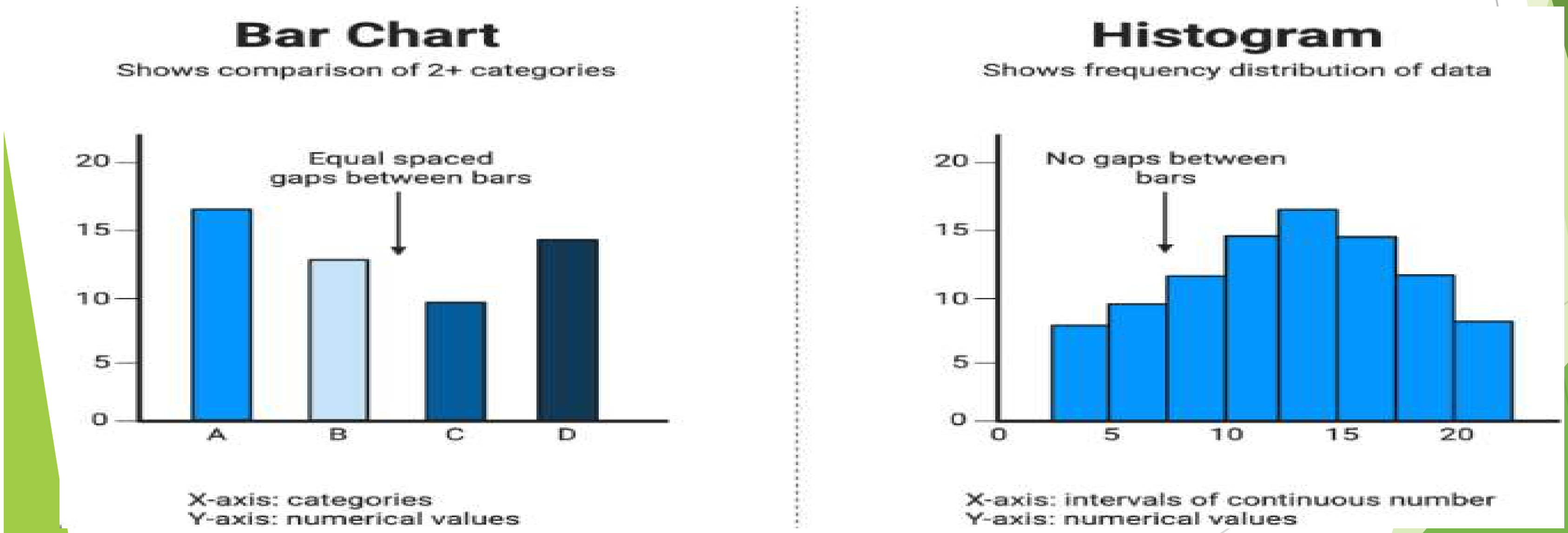
- ➢ In R, the function rnorm() is used to generate random numbers from a normal distribution (also known as the Gaussian distribution).
- ➢ The "r" in rnorm() stands for "random," and "norm" refers to the normal distribution.
- ➢ Breaks tells the bin size.

Pie Chart using R



```
# Sample Data
categories <- c("A", "B", "C", "D")
values <- c(20, 30, 15, 35)

# Pie Chart
pie(values, labels=categories, col=rainbow(length(categories)),
    main="Pie Chart")
```

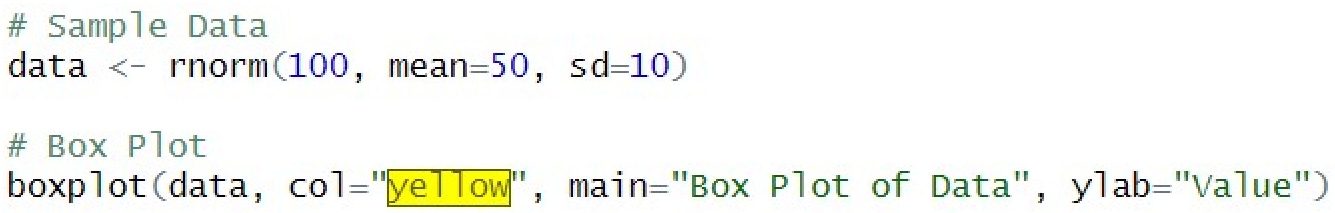
- **Histograms:** A graphical representation of the frequency distribution of numerical data, showing the distribution of values across intervals (bins).
- **Bar Charts:** Similar to histograms but used for categorical data, showing the frequency of each category.



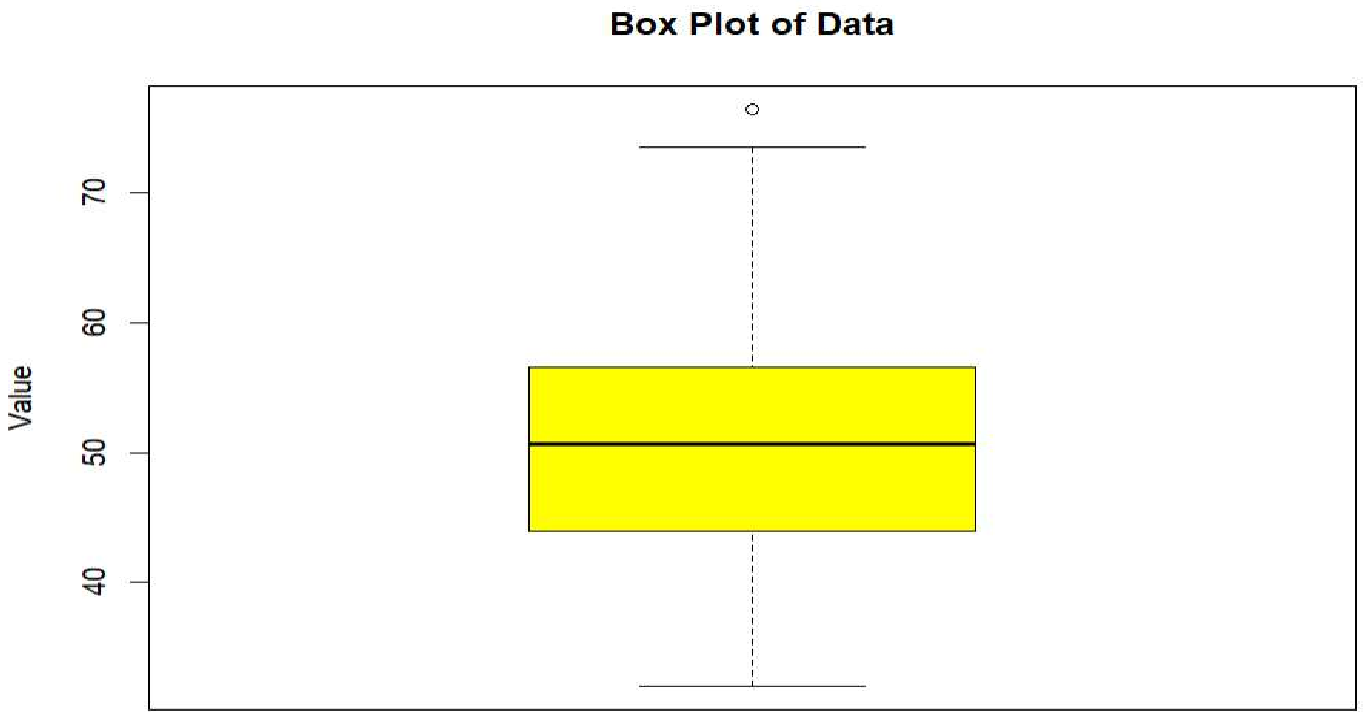
Histograms vs Bar Charts: Key Differences

- **Histograms** are used for numerical data, while **bar charts** are used for categorical data.
- **Histograms** show the distribution of values across intervals, while **bar charts** show the frequency of categories.
- **Histograms** bars typically touch each other to represent a continuous range of values, while **bar chart** bars are separated to show distinct categories.

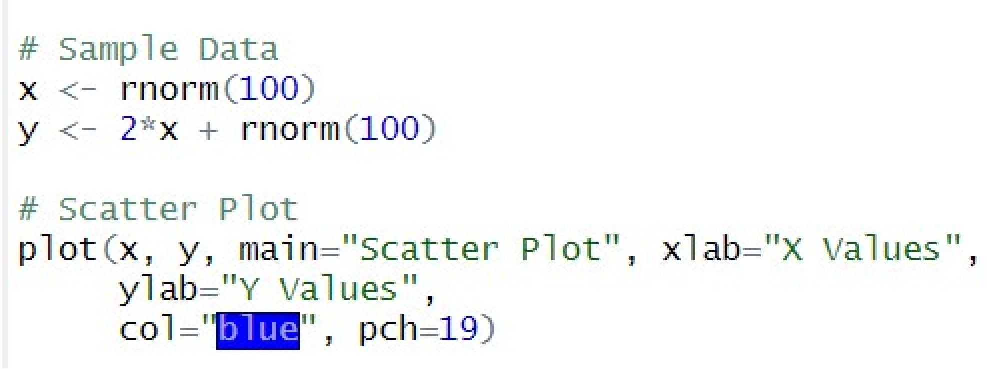
Box Plot



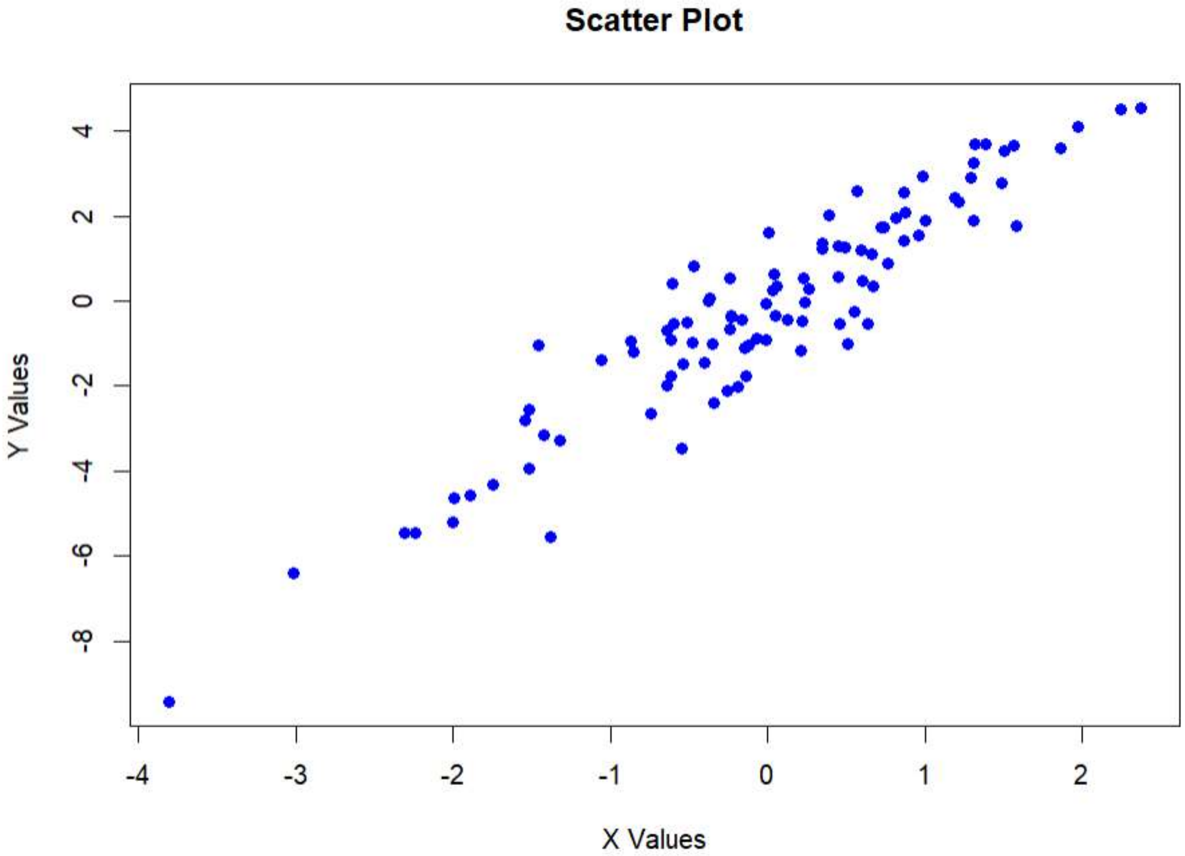
```
# Sample Data
data <- rnorm(100, mean=50, sd=10)

# Box Plot
boxplot(data, col="yellow", main="Box Plot of Data", ylab="Value")
```

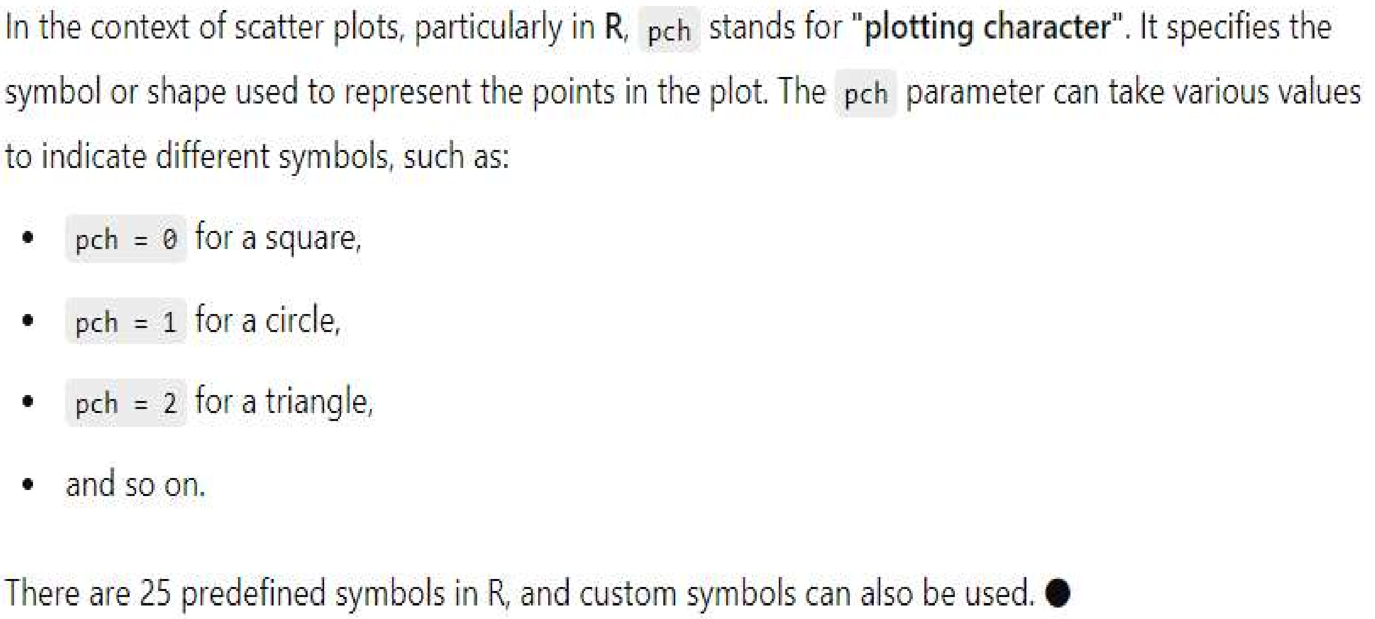


Scatter Plot



```
# Sample Data
x <- rnorm(100)
y <- 2*x + rnorm(100)

# Scatter Plot
plot(x, y, main="Scatter Plot", xlab="X Values",
     ylab="Y Values",
     col="blue", pch=19)
```

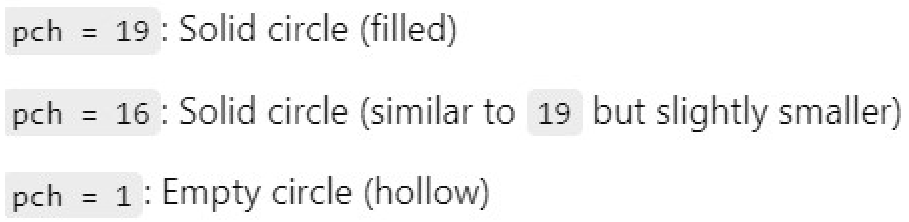




In the context of scatter plots, particularly in **R**, `pch` stands for "**plotting character**". It specifies the symbol or shape used to represent the points in the plot. The `pch` parameter can take various values to indicate different symbols, such as:

- `pch = 0` for a square,
- `pch = 1` for a circle,
- `pch = 2` for a triangle,
- and so on.

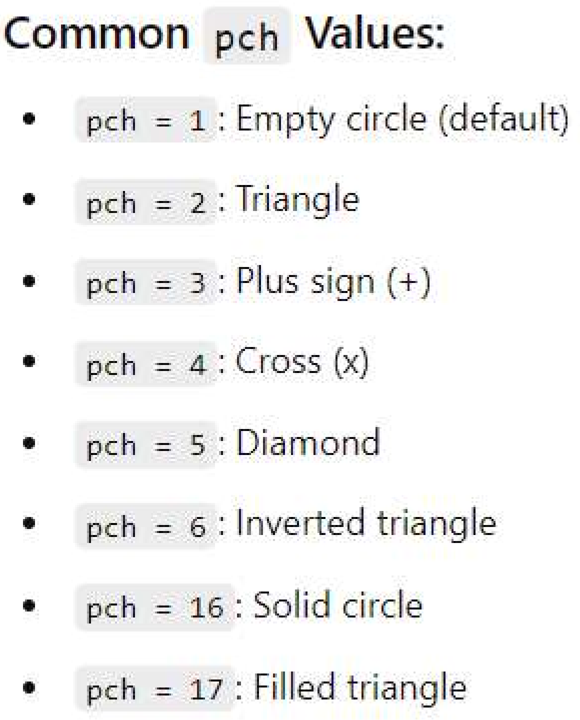
There are 25 predefined symbols in R, and custom symbols can also be used. ●



`pch = 19`: Solid circle (filled)

`pch = 16`: Solid circle (similar to `19` but slightly smaller)

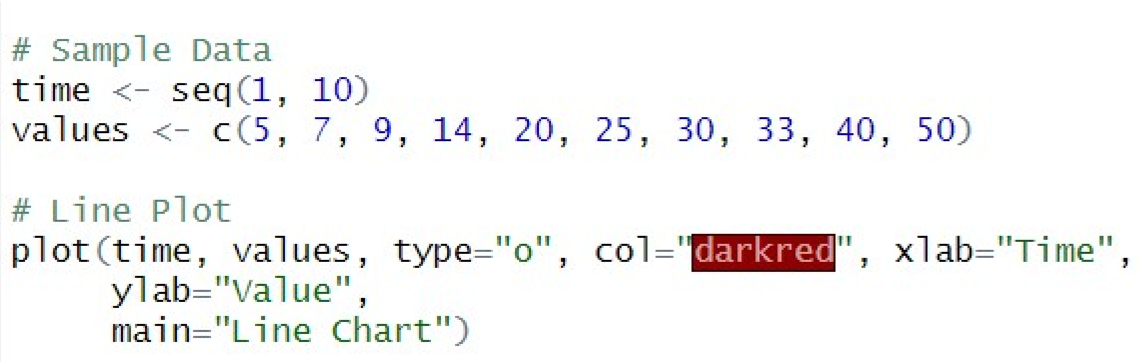
`pch = 1`: Empty circle (hollow)



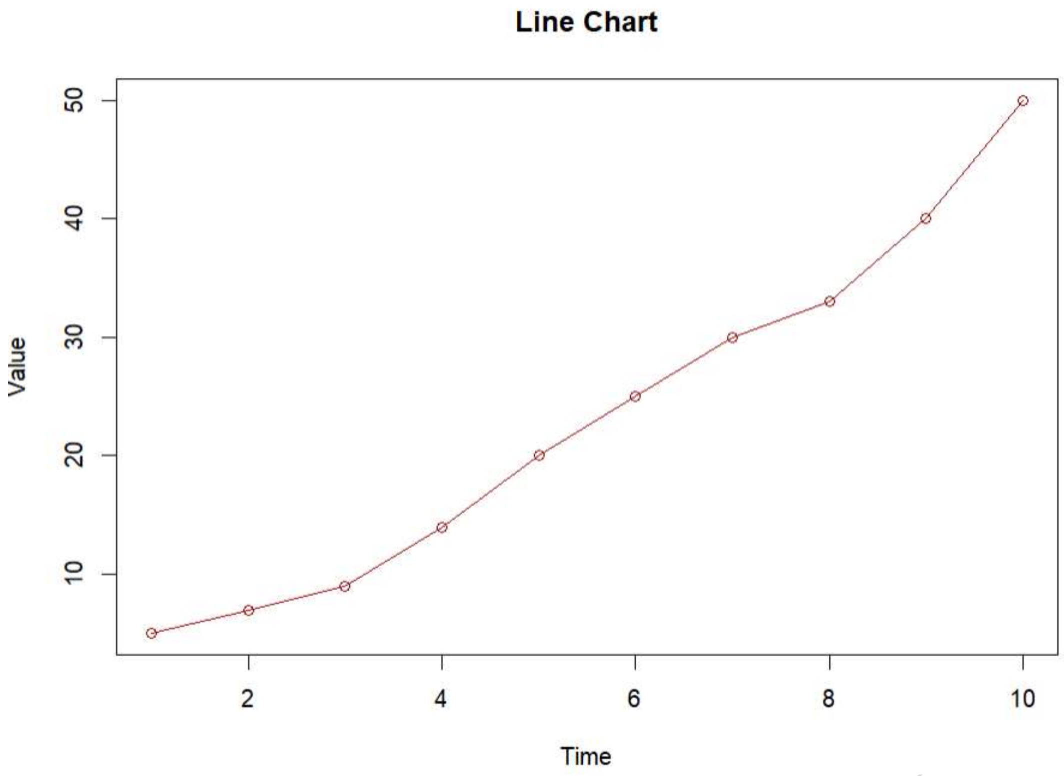
Common `pch` Values:

- `pch = 1`: Empty circle (default)
- `pch = 2`: Triangle
- `pch = 3`: Plus sign (+)
- `pch = 4`: Cross (x)
- `pch = 5`: Diamond
- `pch = 6`: Inverted triangle
- `pch = 16`: Solid circle
- `pch = 17`: Filled triangle

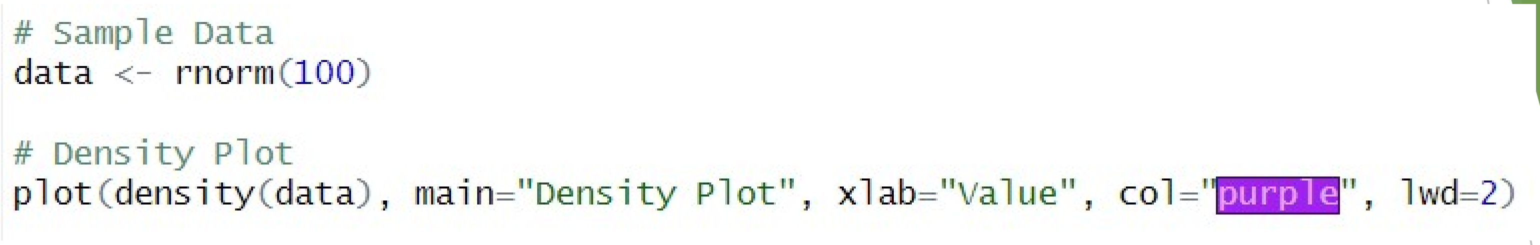
Line Chart



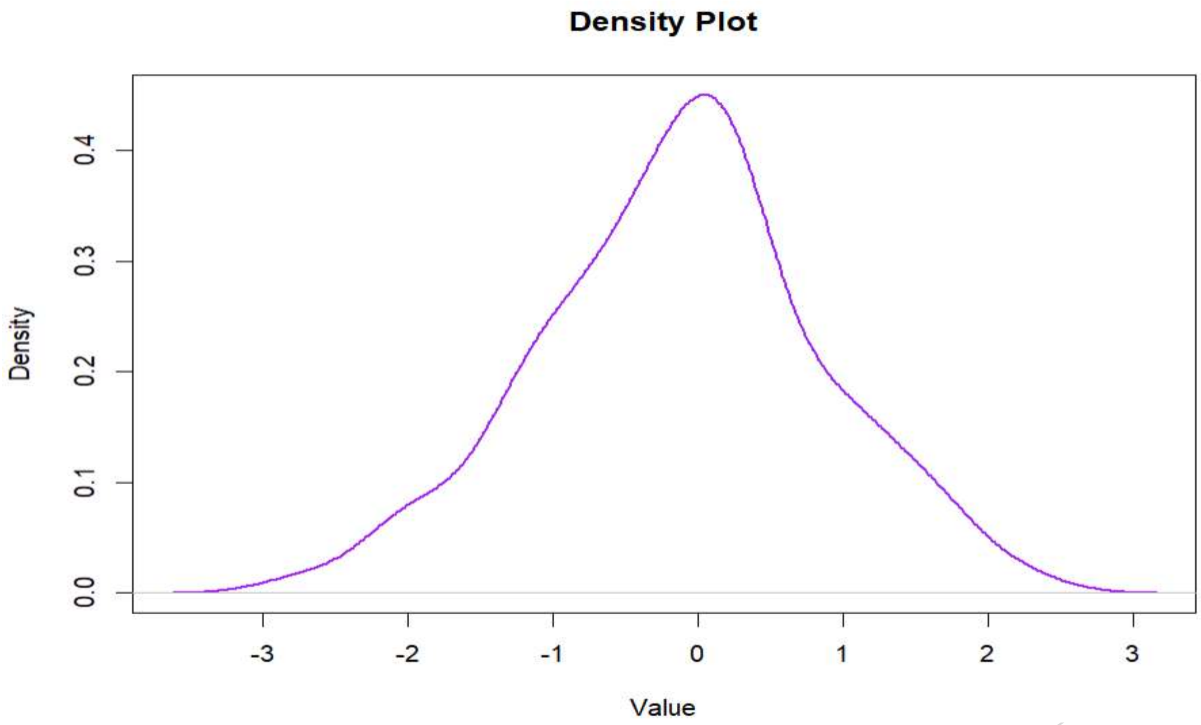
```
# Sample Data
time <- seq(1, 10)
values <- c(5, 7, 9, 14, 20, 25, 30, 33, 40, 50)

# Line Plot
plot(time, values, type="o", col="darkred", xlab="Time",
     ylab="Value",
     main="Line Chart")
```



Density Plot

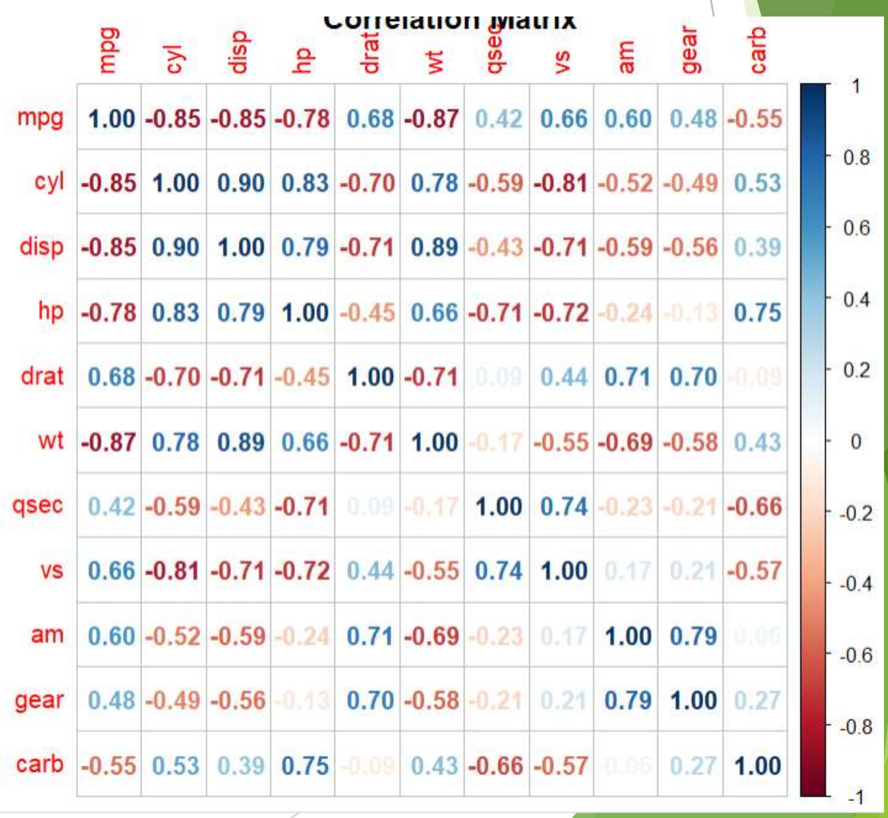


```
# Sample Data
data <- rnorm(100)

# Density Plot
plot(density(data), main="Density Plot", xlab="Value", col="purple", lwd=2)
```



Correlation Matrix Plot

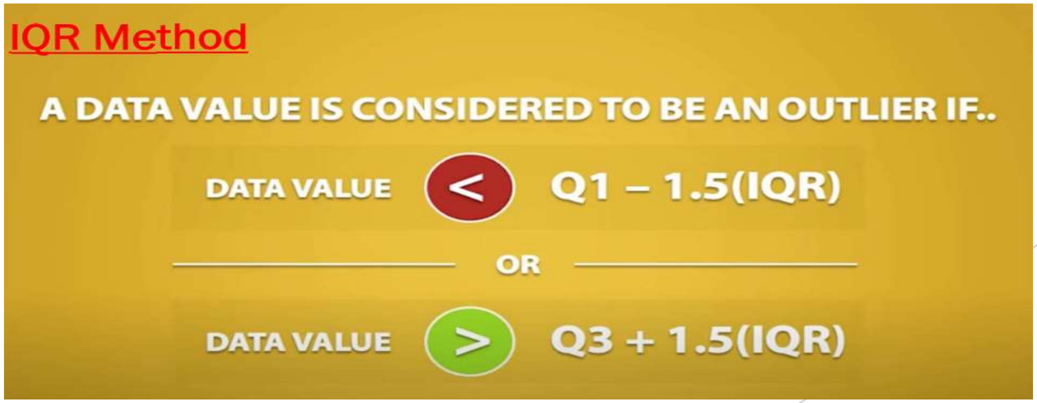
```
# Install and load the necessary package
install.packages("corrplot")
library(corrplot)

# Sample Data
data <- mtcars

# Correlation Matrix
cor_matrix <- cor(data)

# Correlation Plot
corrplot(cor_matrix, method="number",
         main="Correlation Matrix")
```



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

Use Outlier Detection Techniques

- Techniques like **Z-score** or **IQR method**, can help identify outliers systematically.
- **Z-score:** Measures how far a data point is from the mean in terms of standard deviations. A Z-score greater than 3 or less than -3 is often considered an outlier.
- **IQR method:** Values that lie beyond 1.5 times the IQR above the third quartile or below the first quartile are flagged as outliers.



IQR Method

A DATA VALUE IS CONSIDERED TO BE AN OUTLIER IF..

DATA VALUE < Q1 – 1.5(IQR)

OR

DATA VALUE > Q3 + 1.5(IQR)

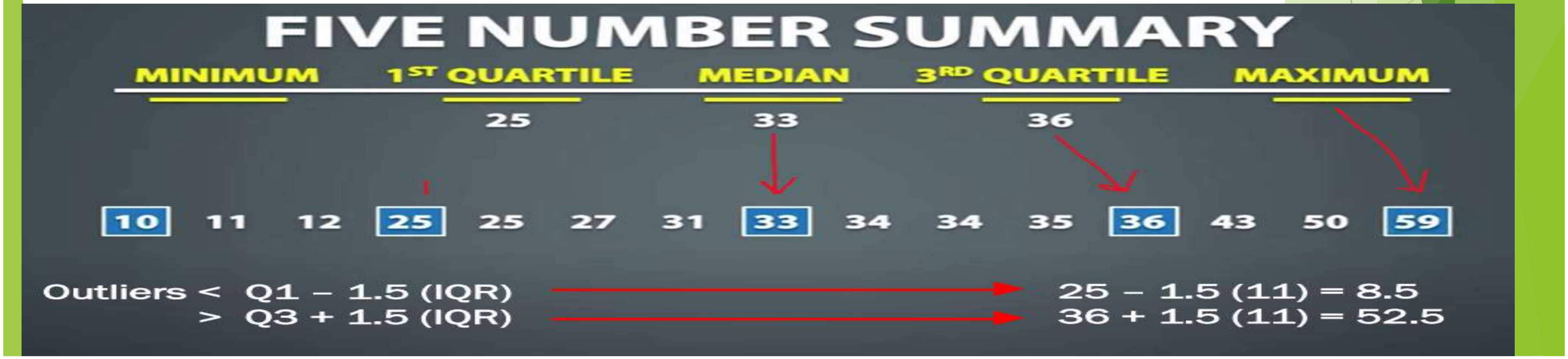
Example:

- **HR Analytics Example: Employee Performance**
  - Imagine you're analyzing the **performance scores** of employees, which range from 1 to 100.
  - Most employees have scores between 40 and 70, but one employee has a score of 95.
  - You may initially consider this a **positive outlier**—an exceptional performer.
  - **Scenario 1**: If you’re running a training program and want to find struggling employees, the outlier may not be useful, and you could consider **excluding** this value from certain analyses.
  - **Scenario 2**: However, if you’re analyzing potential future leaders or high performers, the outlier might represent someone who deserves **special attention** for promotion or leadership training.

- **Business Analytics Example: Sales Revenue**
- You are analyzing monthly **store sales**. Nine out of ten stores have sales between $10,000 and $15,000, but one store reports $100,000.
- This is an **outlier**, and it can affect your conclusions about overall store performance.
  - **Scenario 1**: If your goal is to understand typical store performance, this outlier might skew your analysis, and you might consider **removing** it for better insights.
  - **Scenario 2**: If your goal is to investigate **best practices**, this outlier could indicate a high-performing store whose strategies should be replicated across other stores.
  - **Typical performance** refers to the **average or most common range of performance** within a dataset.

FIVE POINT SUMMARY

- While discussing Box plot we come across the term FIVE point summary. That includes:
- **This divides data in to 4 parts.**
- **1. Minimum**
- **2. 25% = Q1 = 25% of data**
- **3. 50% = Q2 = Median = 50% of data**
- **4. 75% = Q3 = 75% of data**
- **5. Maximum**



FIVE NUMBER SUMMARY

MINIMUM	1ST QUARTILE	MEDIAN	3RD QUARTILE	MAXIMUM
	25	33	36	

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

Outliers < Q1 – 1.5 (IQR) → 25 – 1.5 (11) = 8.5
> Q3 + 1.5 (IQR) → 36 + 1.5 (11) = 52.5

Interquartile Range (IQR):

- **Definition:** The range within which the central 50% of data values lie, between the 1st quartile (Q1) and the 3rd quartile (Q3).

- **Steps to find IQR**

- Order the data from least to greatest
- Find the median
- The left side of median is lower half and right side of the data set is upper half.
- Calculate the median of both the lower and upper half of the data named as Q1 andQ3.

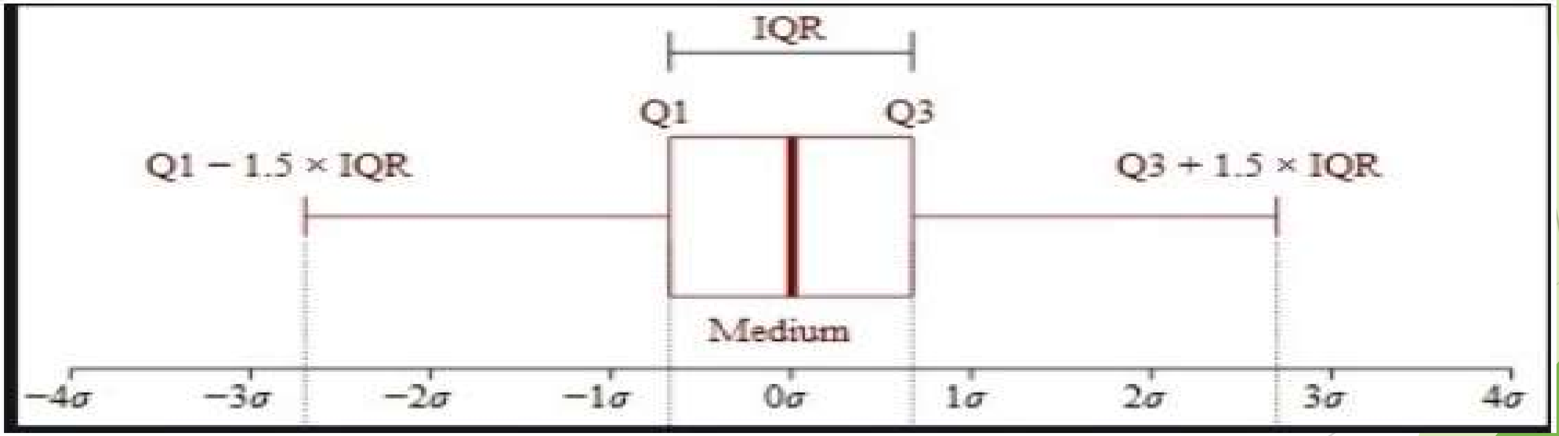
- The IQR is the difference between the upper and lower half.

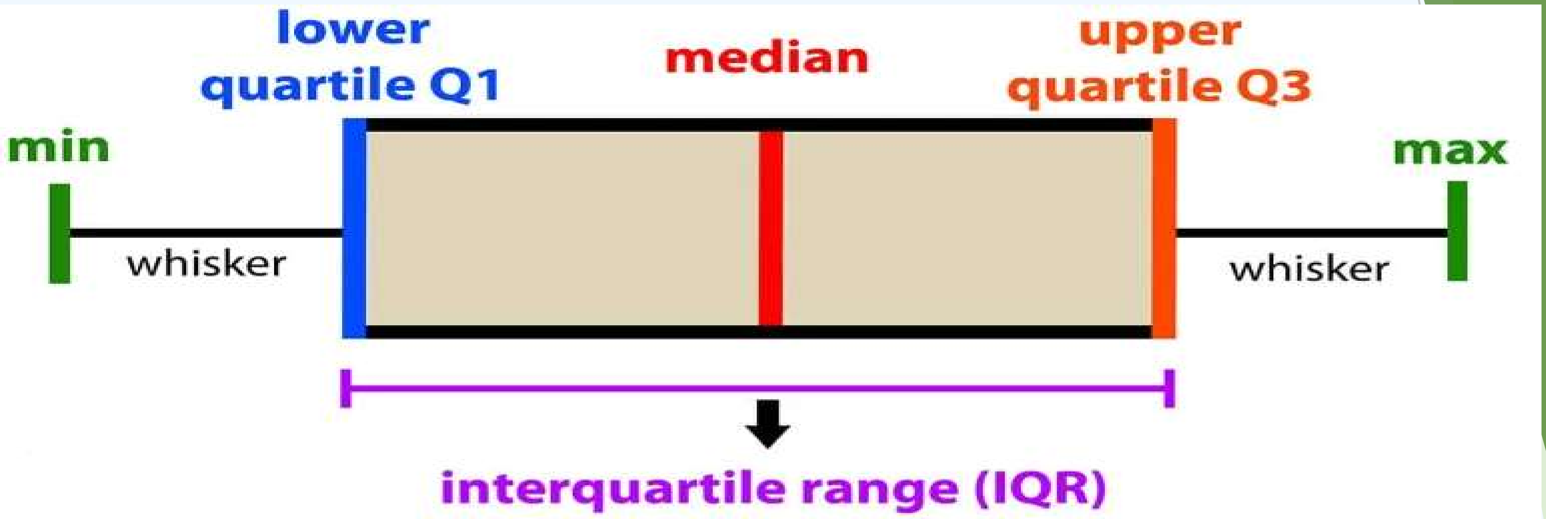
Formula:

$$\text{IQR} = Q3 \quad - Q1$$

- **Use:** Measures the spread of the middle 50% of data, reducing the influence of outliers and giving a more robust measure of spread.

- Generally a **box plot**, also known as a box and whisker plot, is a graphical representation of the five-number summary of a dataset including Q1,Q2,Q3,Max,Min.
- This helps to visualize the central tendency, spread, and skewness of the data.

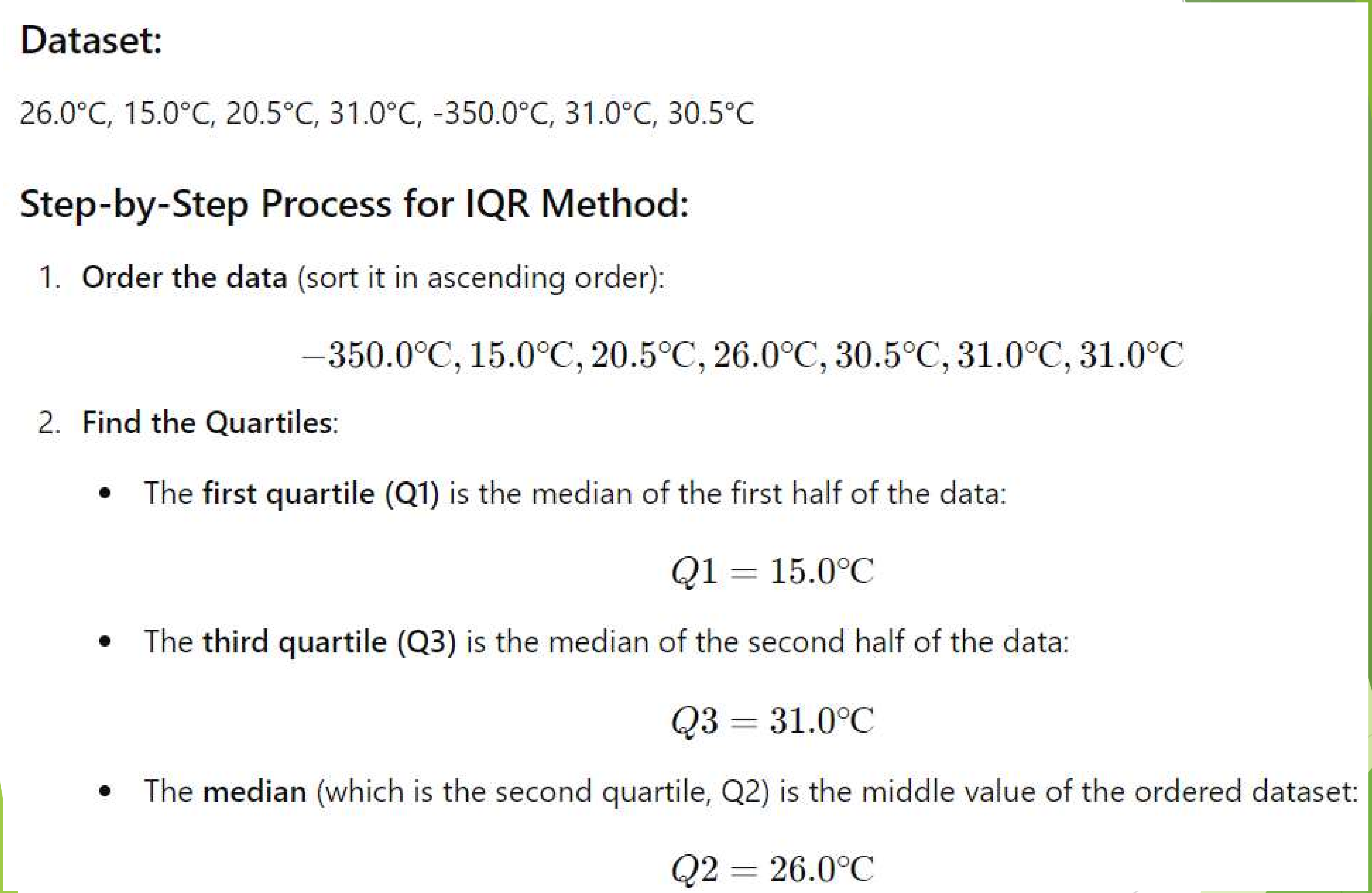




- **Example Scenario:** For a dataset of employee salaries, if Q1 is $45,000 and Q3 is $55,000,
- the IQR is $55,000 - $45,000 = $10,000.
- This shows that the central 50% of salaries fall within a $10,000 range.

- **Ques.** Can you identify the outliers from the below dataset, using the IQR method?

- **DATA** : 26.0 ℃ , 15.0 ℃ , 20.5 ℃ , 31 ℃ , -350.0 ℃ , 31.0 ℃ , 30.5 ℃
- n = 7



Dataset:

26.0°C, 15.0°C, 20.5°C, 31.0°C, -350.0°C, 31.0°C, 30.5°C

Step-by-Step Process for IQR Method:

1. **Order the data** (sort it in ascending order):

$$-350.0°\text{C}, 15.0°\text{C}, 20.5°\text{C}, 26.0°\text{C}, 30.5°\text{C}, 31.0°\text{C}, 31.0°\text{C}$$

2. **Find the Quartiles**:

- The **first quartile (Q1)** is the median of the first half of the data:

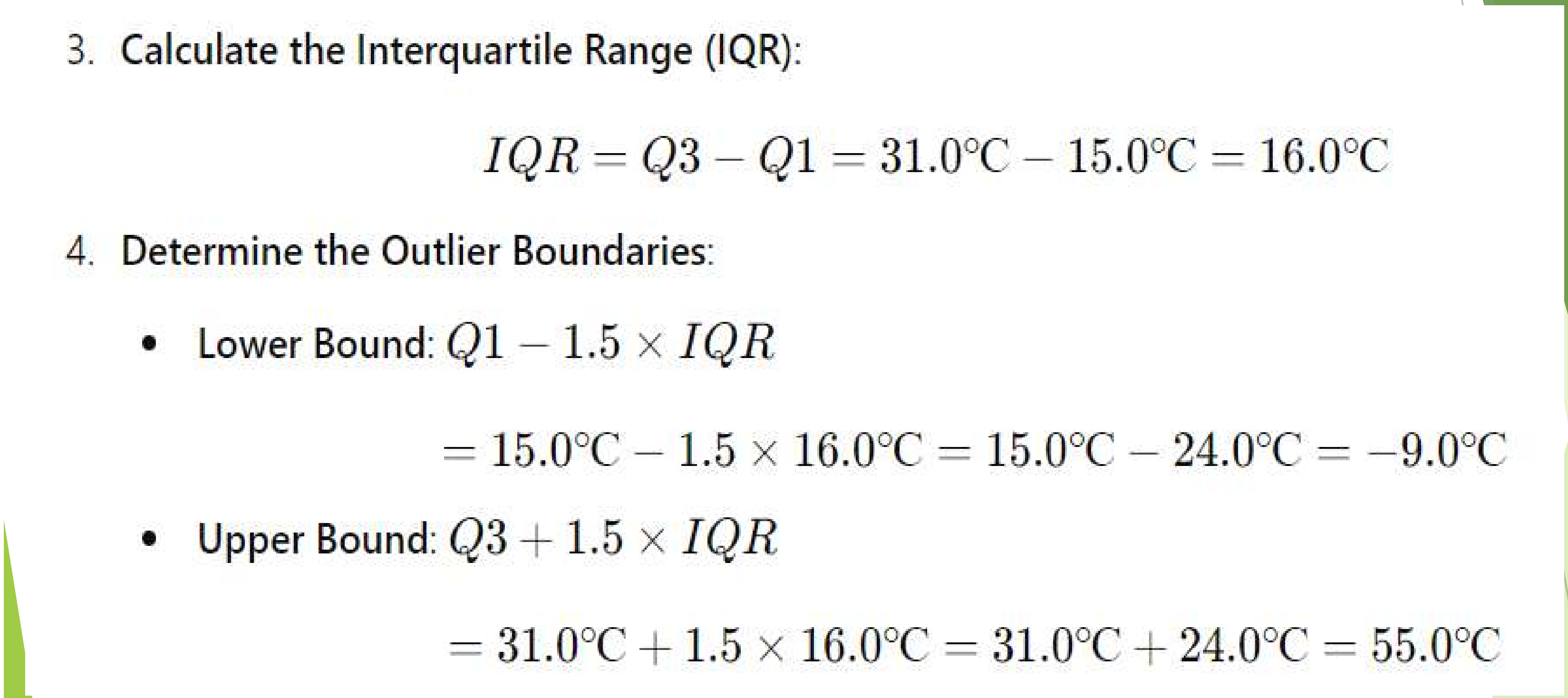
$$Q1 = 15.0°\text{C}$$

- The **third quartile (Q3)** is the median of the second half of the data:

$$Q3 = 31.0°\text{C}$$

- The **median** (which is the second quartile, Q2) is the middle value of the ordered dataset:

$$Q2 = 26.0°\text{C}$$



3. Calculate the Interquartile Range (IQR):

$$IQR = Q3 - Q1 = 31.0°\text{C} - 15.0°\text{C} = 16.0°\text{C}$$

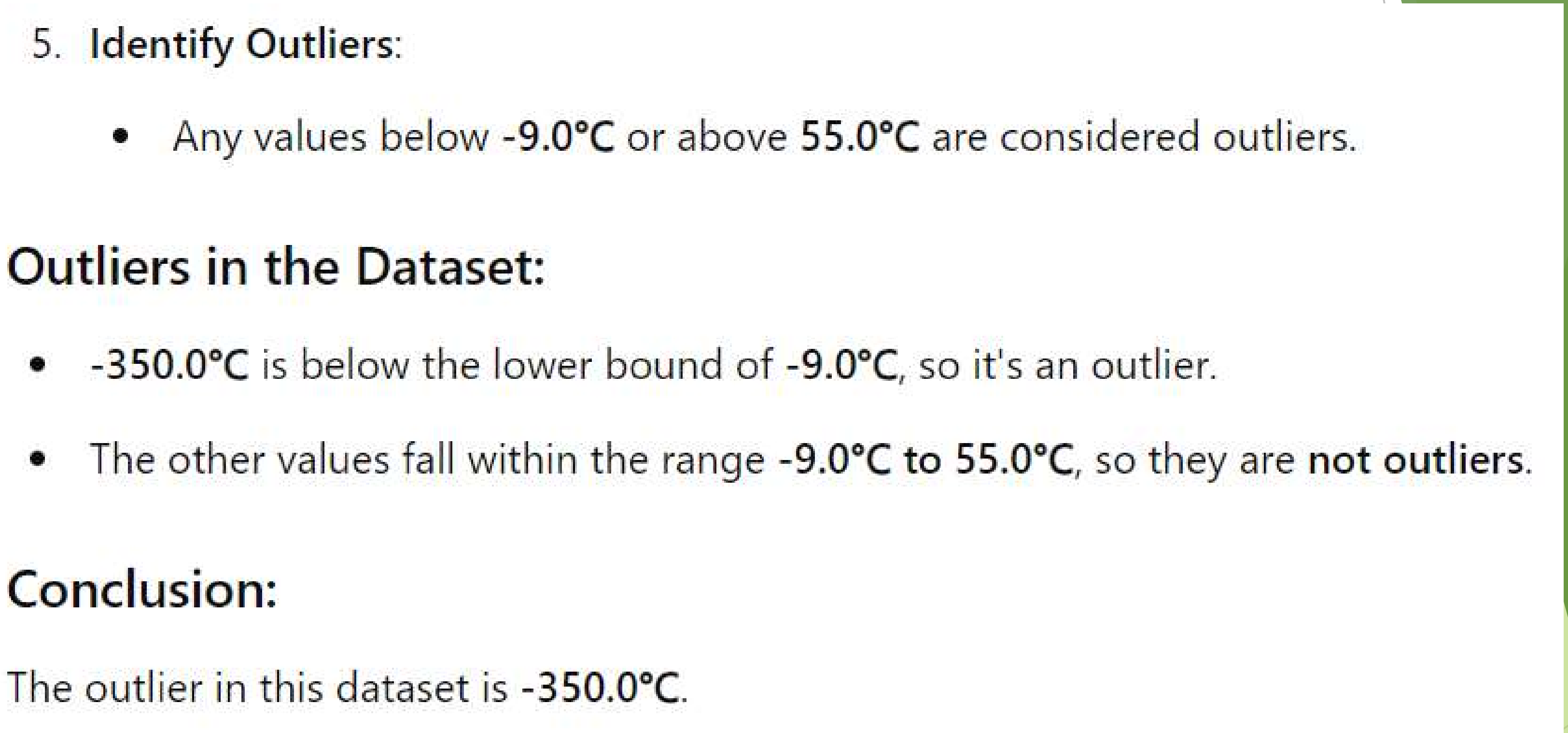
4. Determine the Outlier Boundaries:

- Lower Bound: $Q1 - 1.5 \times IQR$

$$= 15.0°\text{C} - 1.5 \times 16.0°\text{C} = 15.0°\text{C} - 24.0°\text{C} = -9.0°\text{C}$$

- Upper Bound: $Q3 + 1.5 \times IQR$

$$= 31.0°\text{C} + 1.5 \times 16.0°\text{C} = 31.0°\text{C} + 24.0°\text{C} = 55.0°\text{C}$$



5. **Identify Outliers**:
   - Any values below **-9.0°C** or above **55.0°C** are considered outliers.

Outliers in the Dataset:

- **-350.0°C** is below the lower bound of **-9.0°C**, so it's an outlier.
- The other values fall within the range **-9.0°C to 55.0°C**, so they are **not outliers**.

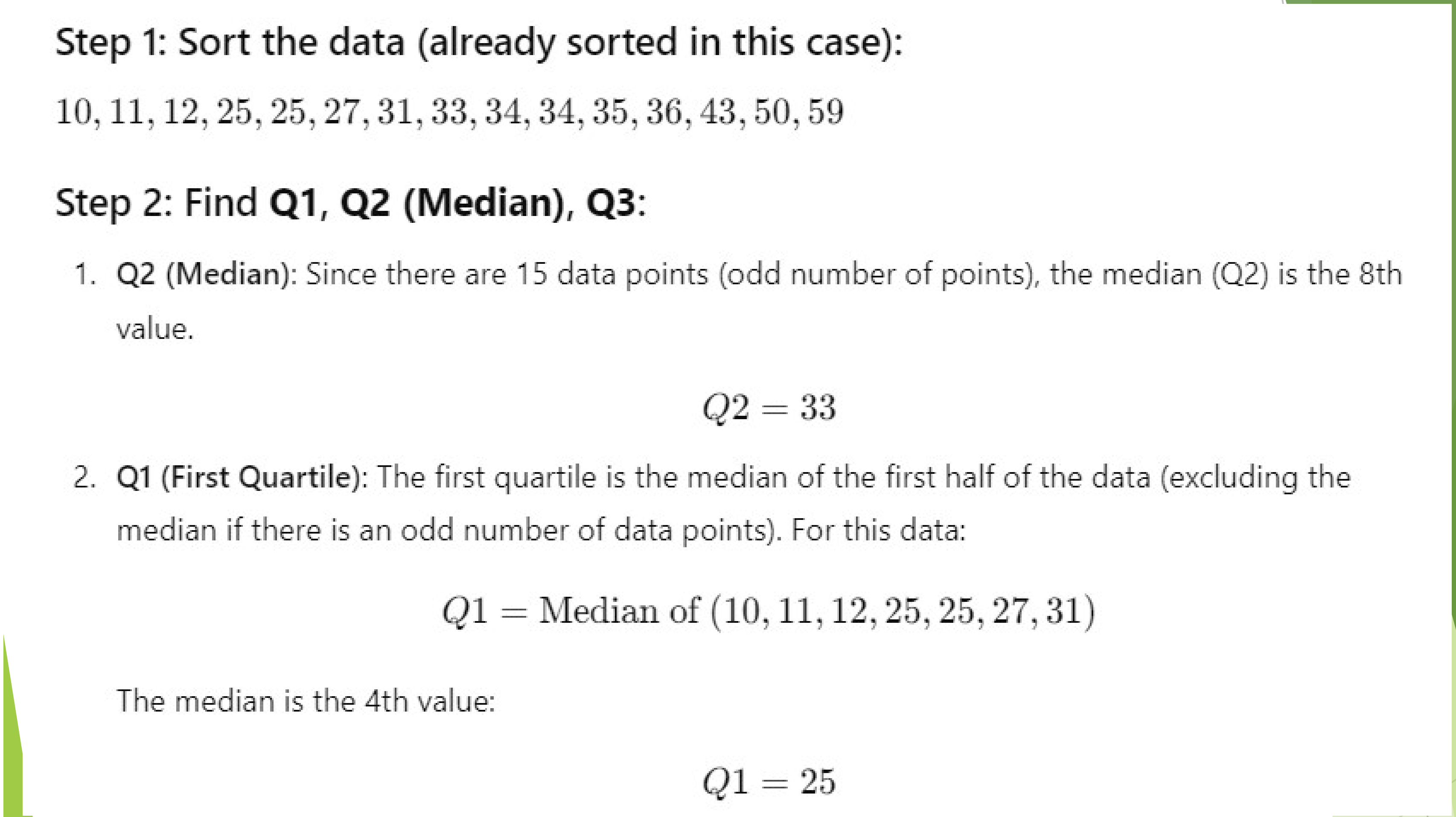
Conclusion:

The outlier in this dataset is **-350.0°C**.

- **Ques : For these number find q1,q2, q3, iqr, max, mini, and draw box plot.**

Data = 10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

- n =15



Step 1: Sort the data (already sorted in this case):

$$10, 11, 12, 25, 25, 27, 31, 33, 34, 34, 35, 36, 43, 50, 59$$

Step 2: Find **Q1, Q2 (Median), Q3**:

1. **Q2 (Median)**: Since there are 15 data points (odd number of points), the median (Q2) is the 8th value.

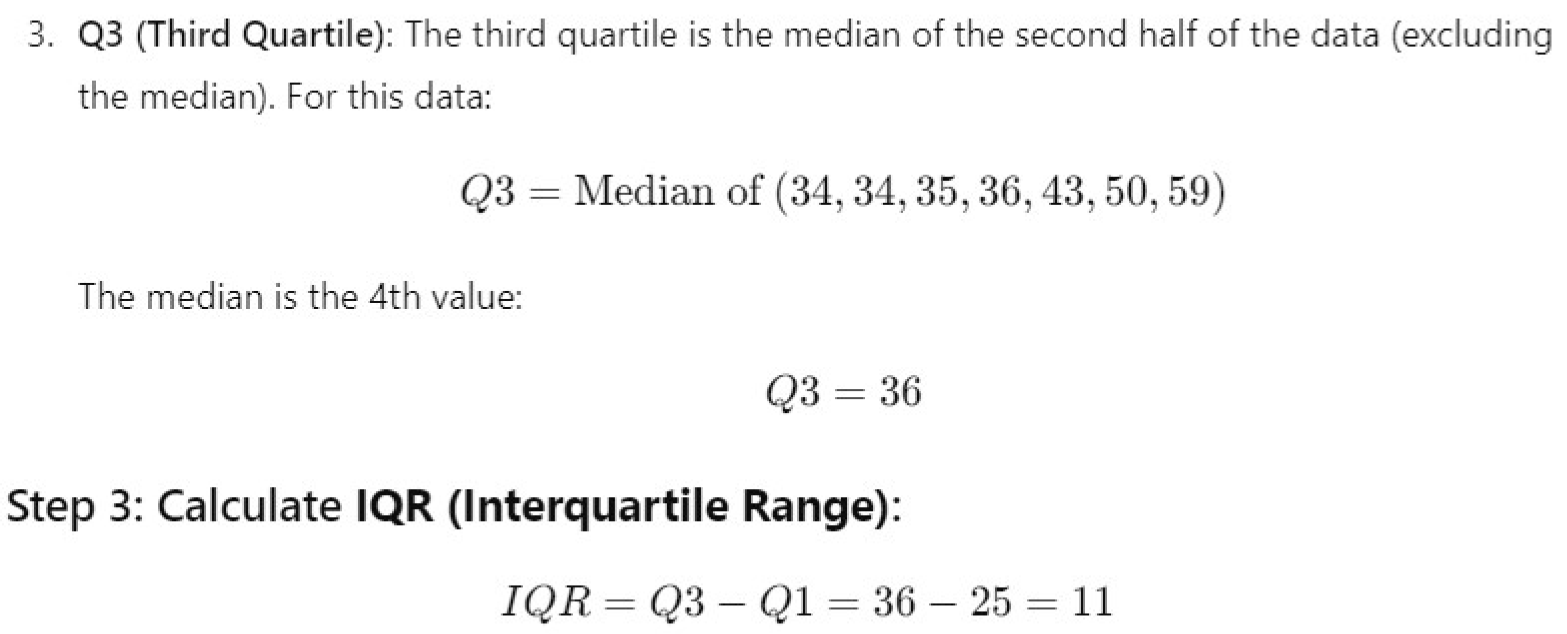
$$Q2 = 33$$

2. **Q1 (First Quartile)**: The first quartile is the median of the first half of the data (excluding the median if there is an odd number of data points). For this data:

$$Q1 = \text{Median of } (10, 11, 12, 25, 25, 27, 31)$$

The median is the 4th value:

$$Q1 = 25$$



3. **Q3 (Third Quartile)**: The third quartile is the median of the second half of the data (excluding the median). For this data:

$$Q3 = \text{Median of } (34, 34, 35, 36, 43, 50, 59)$$

The median is the 4th value:

$$Q3 = 36$$

Step 3: Calculate **IQR (Interquartile Range)**:

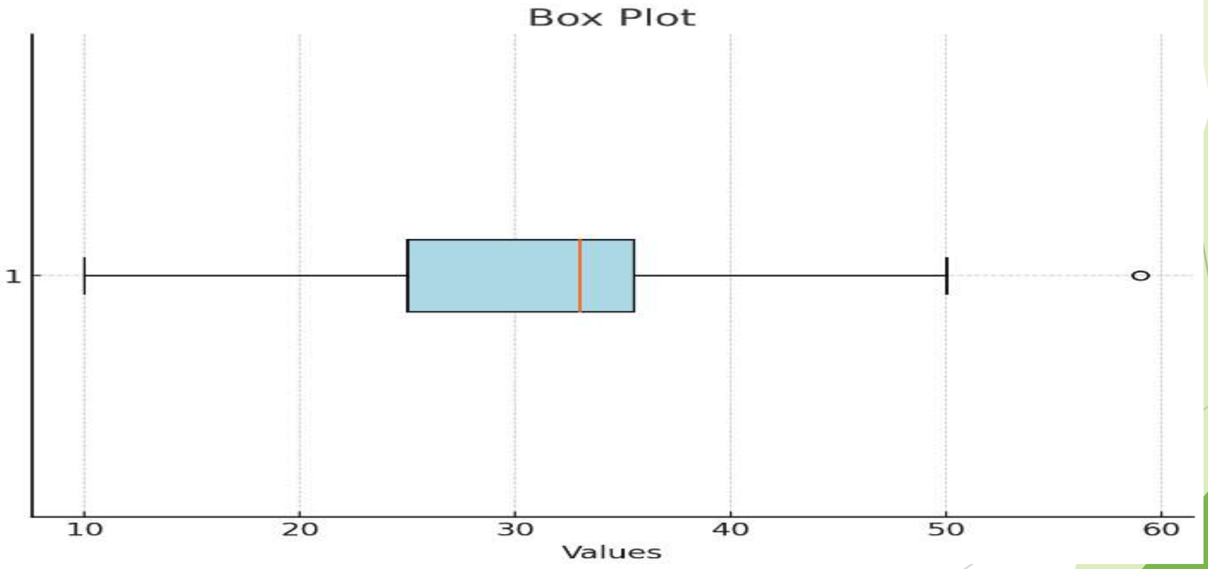
$$IQR = Q3 - Q1 = 36 - 25 = 11$$

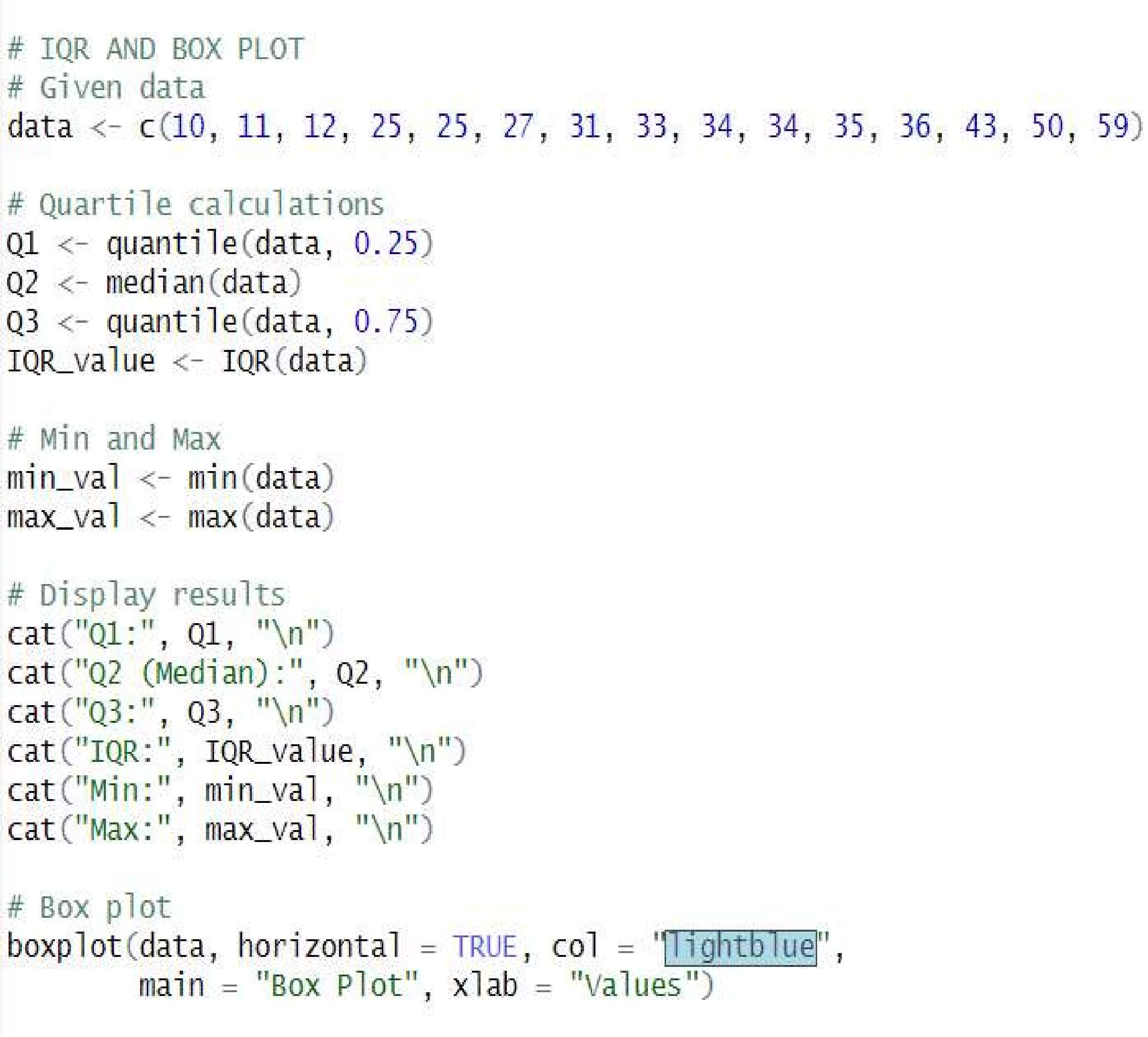
Step 4: Find **Minimum** and **Maximum** values:

- **Minimum** value in the data: 10
- **Maximum** value in the data: 59

Summary:

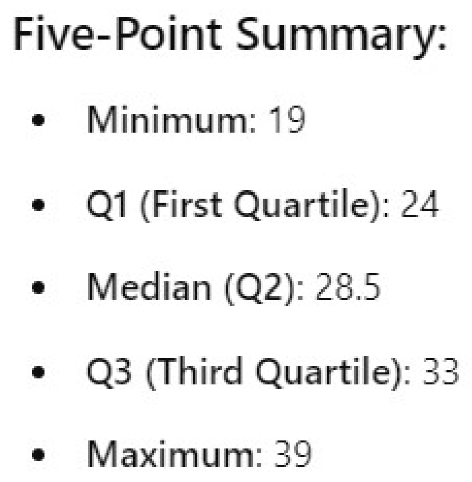
- Q1 = 25
- Q2 (Median) = 33
- Q3 = 36
- IQR = 11
- Minimum = 10
- Maximum = 59





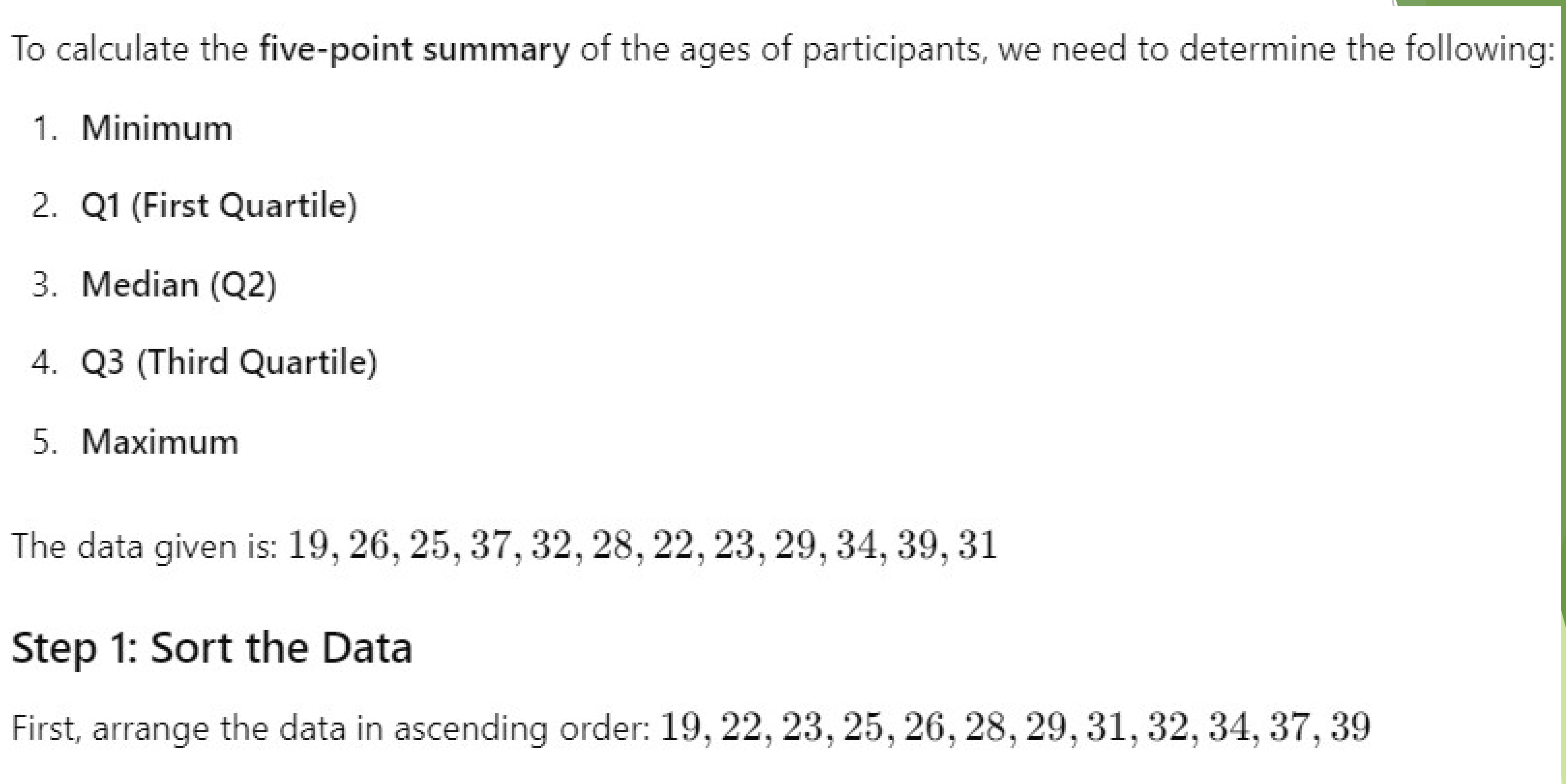
```r
# IQR AND BOX PLOT
# Given data
data <- c(10, 11, 12, 25, 25, 27, 31, 33, 34, 34, 35, 36, 43, 50, 59)

# Quartile calculations
Q1 <- quantile(data, 0.25)
Q2 <- median(data)
Q3 <- quantile(data, 0.75)
IQR_value <- IQR(data)

# Min and Max
min_val <- min(data)
max_val <- max(data)

# Display results
cat("Q1:", Q1, "\n")
cat("Q2 (Median):", Q2, "\n")
cat("Q3:", Q3, "\n")
cat("IQR:", IQR_value, "\n")
cat("Min:", min_val, "\n")
cat("Max:", max_val, "\n")

# Box plot
boxplot(data, horizontal = TRUE, col = "lightblue",
        main = "Box Plot", xlab = "Values")
```

- Ques: Given is the ages of people registered for a webinar, calculate the five point summary of the ages of the participants?
- Data : 19,26,25,37,32,28,22,23,29,34,39,31



Five-Point Summary:

- Minimum: 19
- Q1 (First Quartile): 24
- Median (Q2): 28.5
- Q3 (Third Quartile): 33
- Maximum: 39



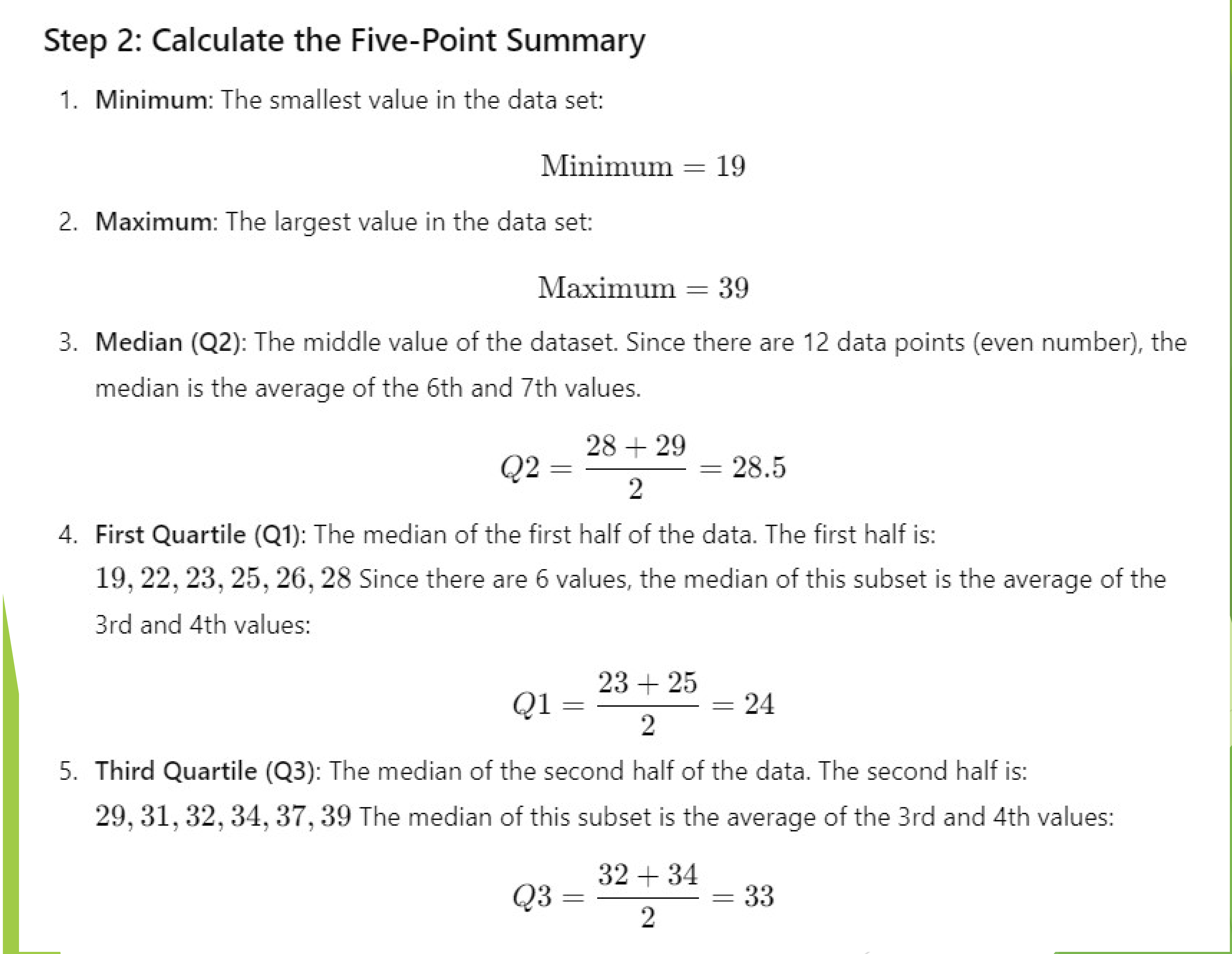
To calculate the **five-point summary** of the ages of participants, we need to determine the following:

1. **Minimum**
2. **Q1 (First Quartile)**
3. **Median (Q2)**
4. **Q3 (Third Quartile)**
5. **Maximum**

The data given is: $19, 26, 25, 37, 32, 28, 22, 23, 29, 34, 39, 31$

Step 1: Sort the Data

First, arrange the data in ascending order: $19, 22, 23, 25, 26, 28, 29, 31, 32, 34, 37, 39$



Step 2: Calculate the Five-Point Summary

1. **Minimum**: The smallest value in the data set:

$$\text{Minimum} = 19$$

2. **Maximum**: The largest value in the data set:

$$\text{Maximum} = 39$$

3. **Median (Q2)**: The middle value of the dataset. Since there are 12 data points (even number), the median is the average of the 6th and 7th values.

$$Q2 = \frac{28 + 29}{2} = 28.5$$

4. **First Quartile (Q1)**: The median of the first half of the data. The first half is: $19, 22, 23, 25, 26, 28$ Since there are 6 values, the median of this subset is the average of the 3rd and 4th values:

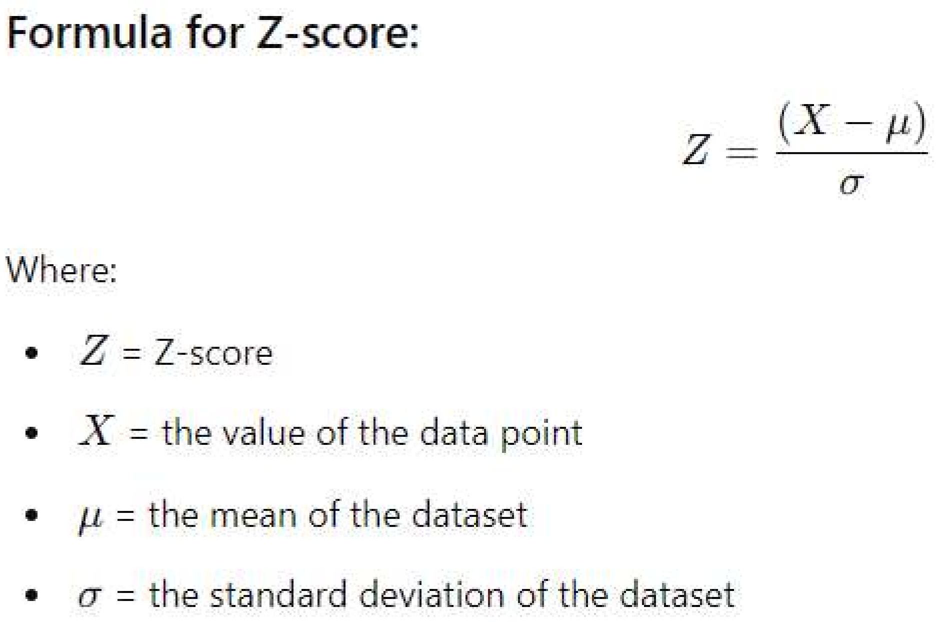
$$Q1 = \frac{23 + 25}{2} = 24$$

5. **Third Quartile (Q3)**: The median of the second half of the data. The second half is: $29, 31, 32, 34, 37, 39$ The median of this subset is the average of the 3rd and 4th values:

$$Q3 = \frac{32 + 34}{2} = 33$$

Z-score to find outliers

- The **Z-score** is a statistical measure that tells you how many standard deviations a data point is from the mean of a dataset.
- It's often used to identify **outliers** because a Z-score that is very high or very low indicates that the data point is far from the mean, potentially classifying it as an outlier.



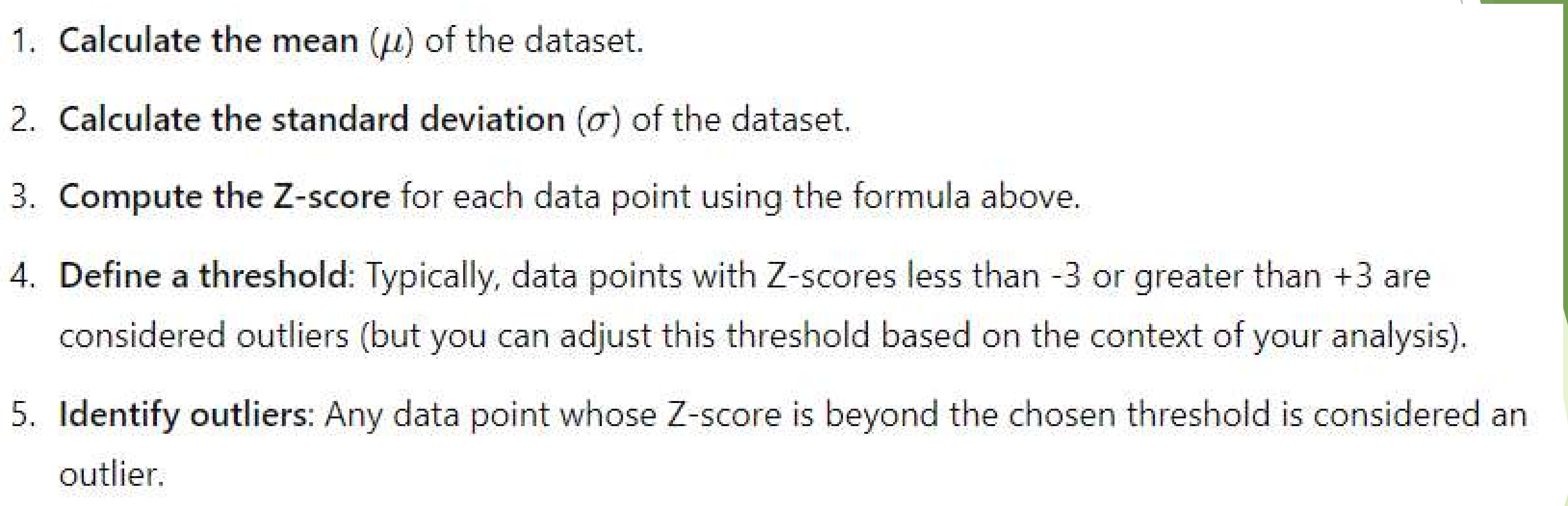
Formula for Z-score:

$$Z = \frac{(X - \mu)}{\sigma}$$

Where:

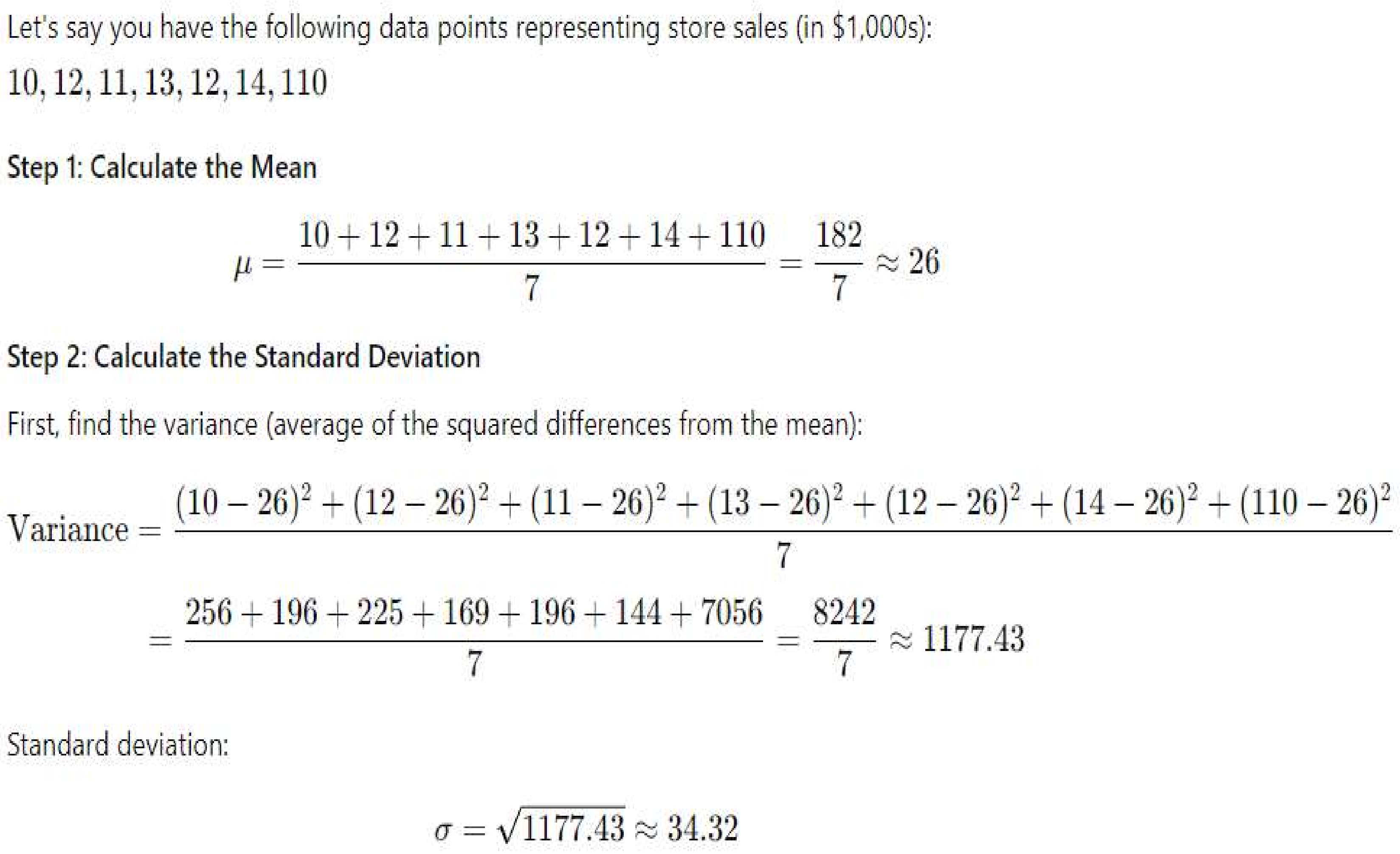
- Z = Z-score
- X = the value of the data point
- μ = the mean of the dataset
- σ = the standard deviation of the dataset

Steps to Find Outliers Using Z-score:



1. **Calculate the mean** (μ) of the dataset.
2. **Calculate the standard deviation** (σ) of the dataset.
3. **Compute the Z-score** for each data point using the formula above.
4. **Define a threshold**: Typically, data points with Z-scores less than -3 or greater than +3 are considered outliers (but you can adjust this threshold based on the context of your analysis).
5. **Identify outliers**: Any data point whose Z-score is beyond the chosen threshold is considered an outlier.

Example:



Let's say you have the following data points representing store sales (in $1,000s):

$10, 12, 11, 13, 12, 14, 110$

Step 1: Calculate the Mean

$$\mu = \frac{10 + 12 + 11 + 13 + 12 + 14 + 110}{7} = \frac{182}{7} \approx 26$$

Step 2: Calculate the Standard Deviation

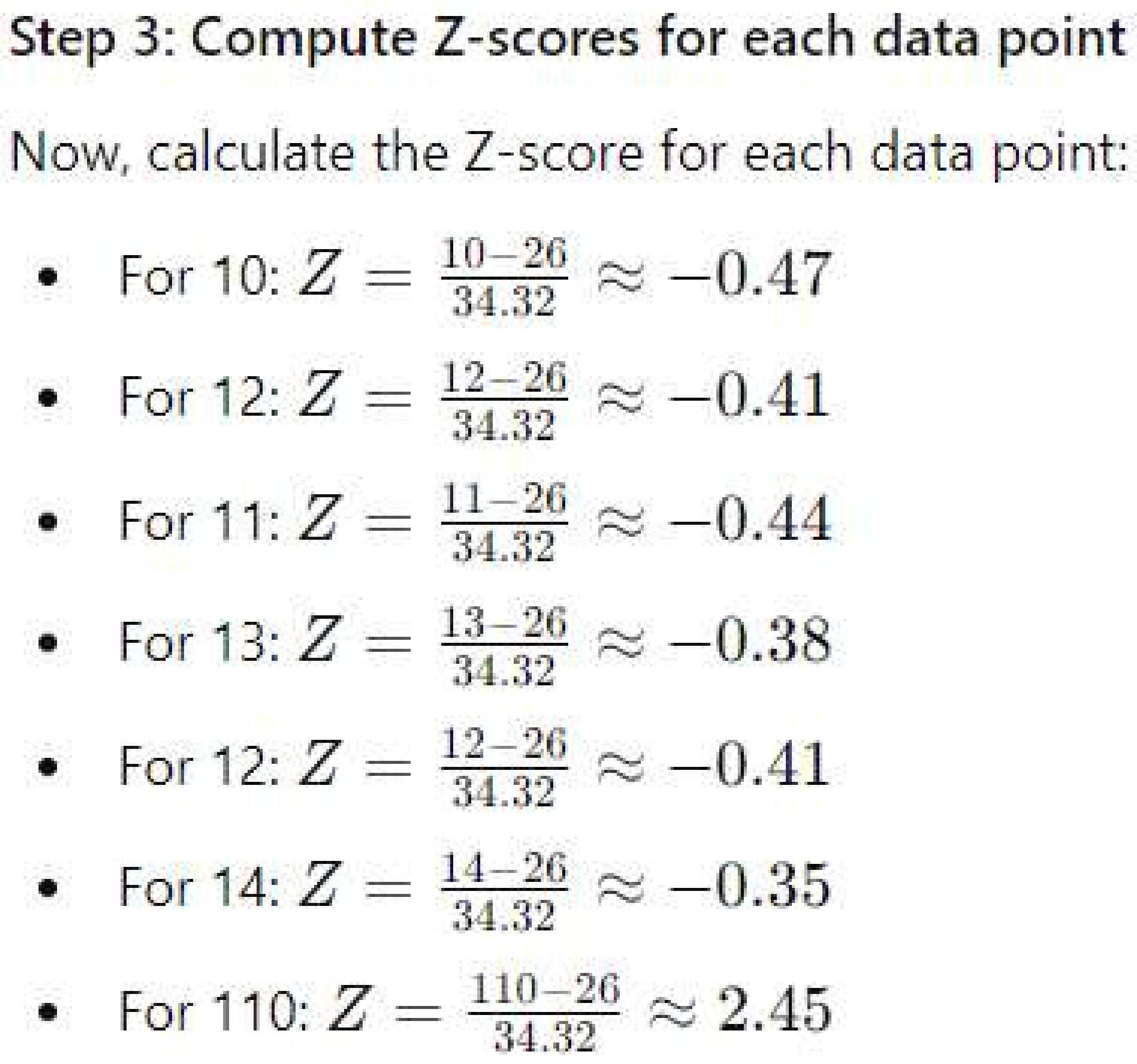
First, find the variance (average of the squared differences from the mean):

$$\text{Variance} = \frac{(10-26)^2 + (12-26)^2 + (11-26)^2 + (13-26)^2 + (12-26)^2 + (14-26)^2 + (110-26)^2}{7}$$

$$= \frac{256 + 196 + 225 + 169 + 196 + 144 + 7056}{7} = \frac{8242}{7} \approx 1177.43$$

Standard deviation:

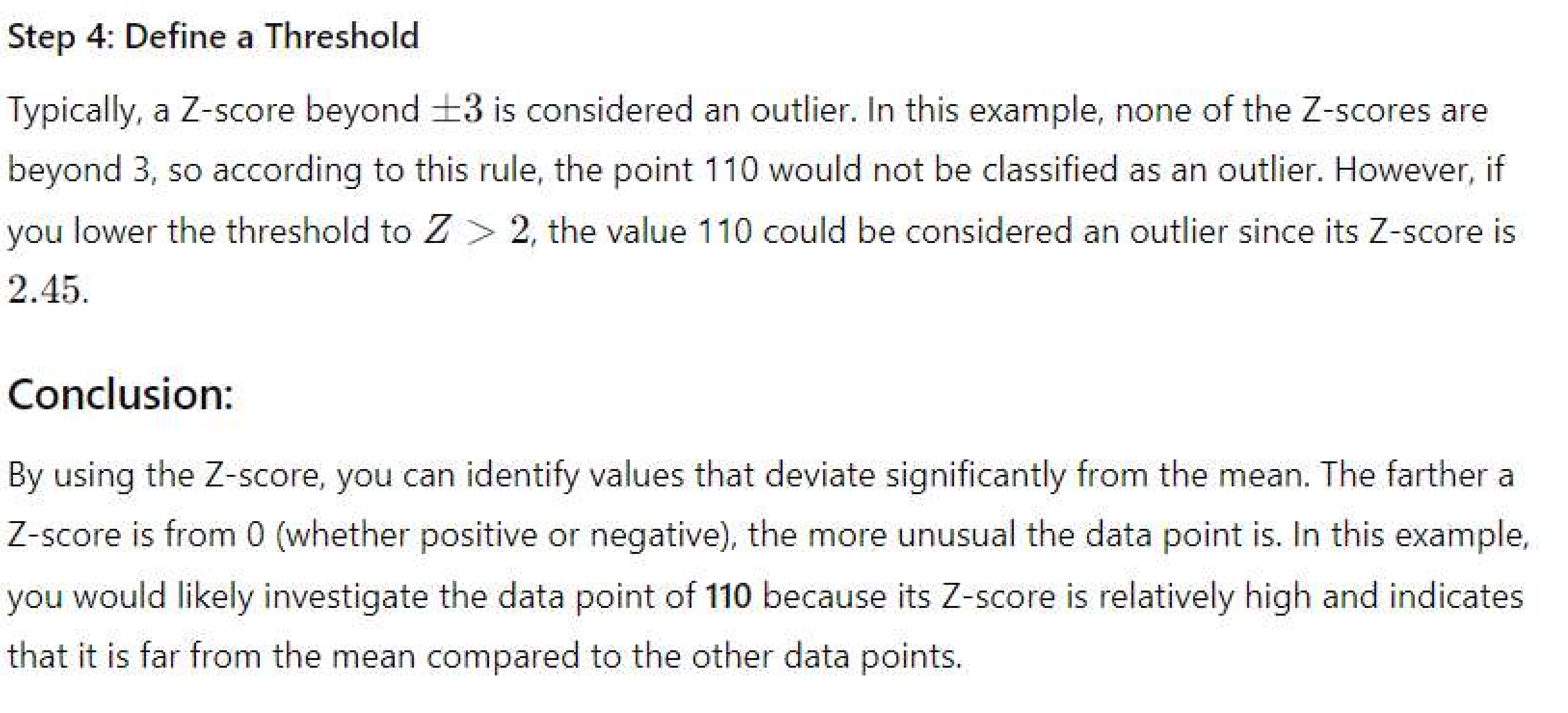
$$\sigma = \sqrt{1177.43} \approx 34.32$$



Step 3: Compute Z-scores for each data point

Now, calculate the Z-score for each data point:

- For 10: $Z = \frac{10-26}{34.32} \approx -0.47$
- For 12: $Z = \frac{12-26}{34.32} \approx -0.41$
- For 11: $Z = \frac{11-26}{34.32} \approx -0.44$
- For 13: $Z = \frac{13-26}{34.32} \approx -0.38$
- For 12: $Z = \frac{12-26}{34.32} \approx -0.41$
- For 14: $Z = \frac{14-26}{34.32} \approx -0.35$
- For 110: $Z = \frac{110-26}{34.32} \approx 2.45$



Step 4: Define a Threshold

Typically, a Z-score beyond ± 3 is considered an outlier. In this example, none of the Z-scores are beyond 3, so according to this rule, the point 110 would not be classified as an outlier. However, if you lower the threshold to $Z > 2$, the value 110 could be considered an outlier since its Z-score is 2.45.

Conclusion:

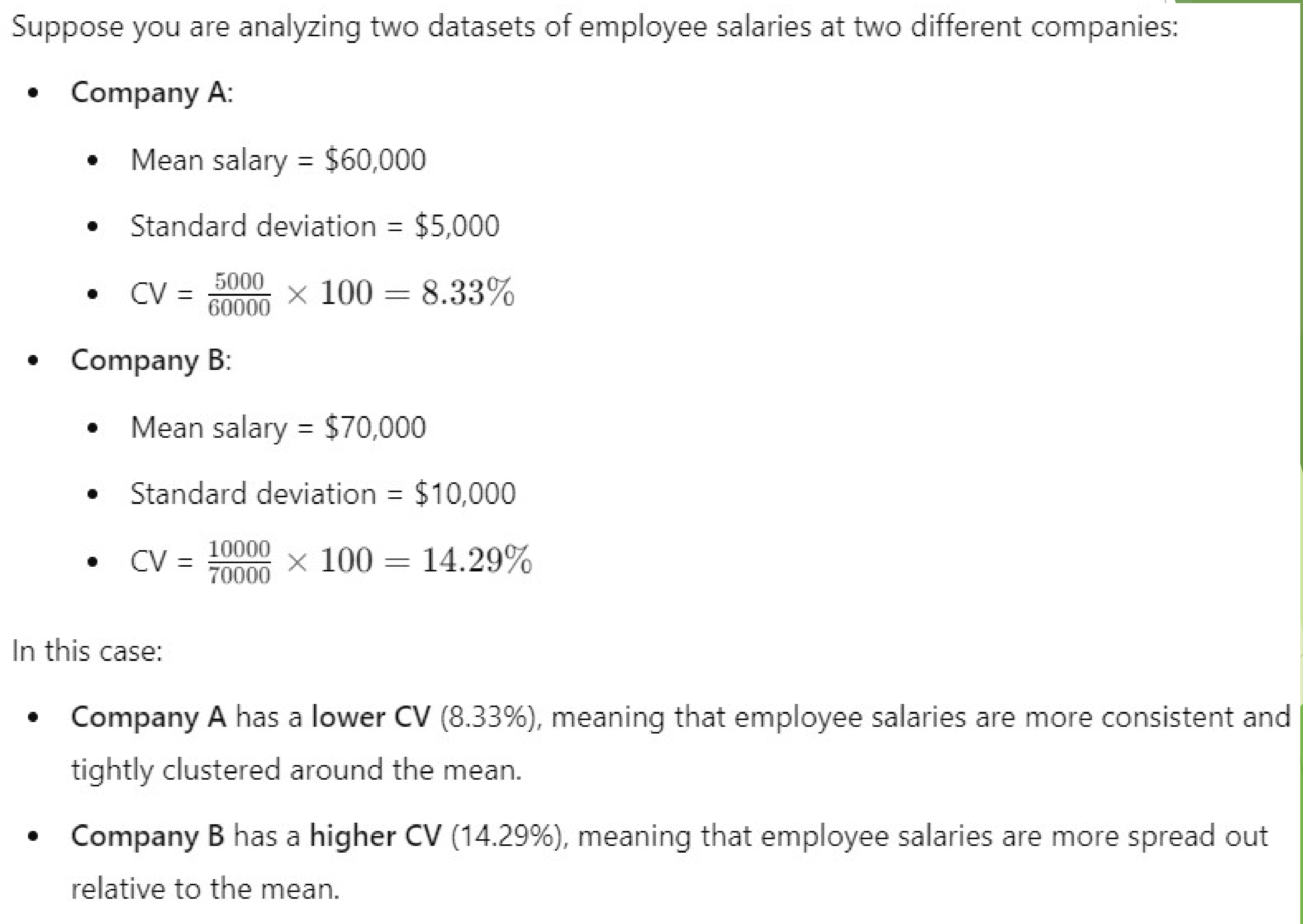
By using the Z-score, you can identify values that deviate significantly from the mean. The farther a Z-score is from 0 (whether positive or negative), the more unusual the data point is. In this example, you would likely investigate the data point of **110** because its Z-score is relatively high and indicates that it is far from the mean compared to the other data points.

Coefficient of Variation (CV)

- **Purpose:** Like variance, standard deviation, it also measures the relative variability in a data set with respect to **mean**.
- The coefficient of variation (CV) quantifies the variation of a dataset relative to the mean and expresses this variation as a percentage.
- **Formula:** CV = $(\sigma/\mu)\times 100$ as %

- **Use Case**: The CV is often used in fields like finance to assess the risk (volatility) of investments relative to their expected return.
- **Example**: If a stock's return has a CV of 10%, it means the standard deviation is 10% of the mean return.

Example:



Suppose you are analyzing two datasets of employee salaries at two different companies:

- **Company A**:
  - Mean salary = $60,000
  - Standard deviation = $5,000
  - CV = $\frac{5000}{60000} \times 100 = 8.33\%$
- **Company B**:
  - Mean salary = $70,000
  - Standard deviation = $10,000
  - CV = $\frac{10000}{70000} \times 100 = 14.29\%$

In this case:

- **Company A** has a **lower CV** (8.33%), meaning that employee salaries are more consistent and tightly clustered around the mean.
- **Company B** has a **higher CV** (14.29%), meaning that employee salaries are more spread out relative to the mean.

Applications:

1.Comparing datasets with different units or scales:

- Since the CV normalizes the standard deviation in terms of the mean, it allows you to compare the relative variability of two datasets, even if they have different units or means.

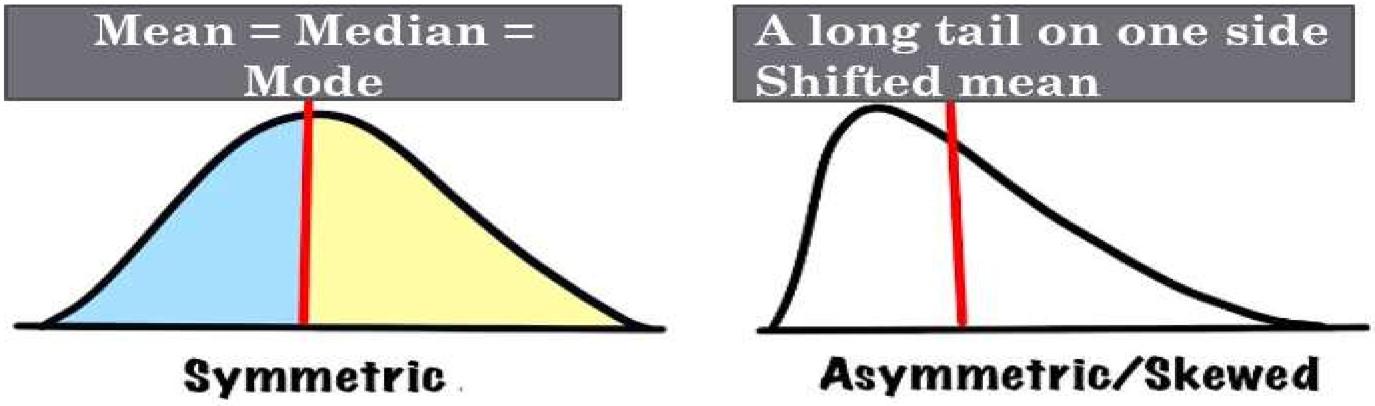
2. Assessing risk: In finance, for example, the CV is used to compare the risk of different investments relative to their expected return.

Key Point:

•**CV** expresses the **relative variability** of the data as a **percentage of the mean**, making it easier to interpret how dispersed the data is around the mean, regardless of the absolute scale or units of the data.

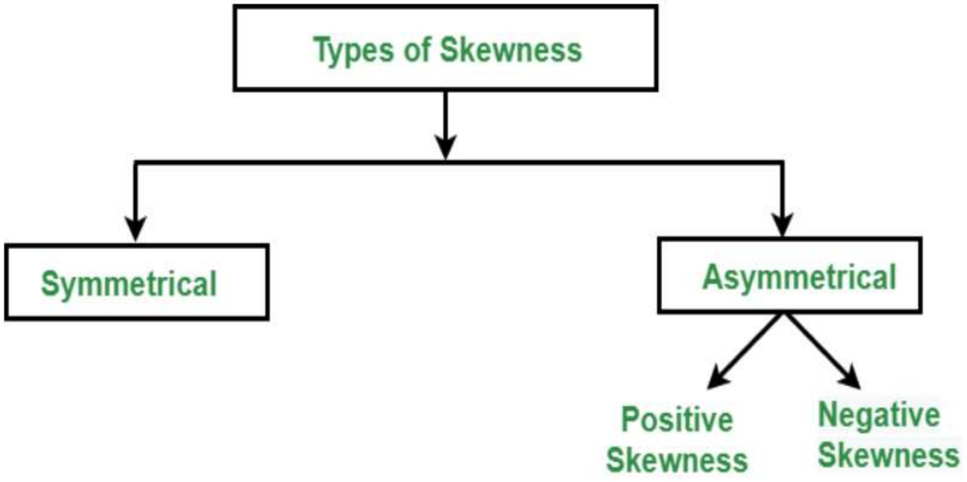
Concept of Symmetry:

- After understanding the **spread** and **center** of the data, the next step is to understand the **shape** of the distribution, which leads to skewness.
- Remember how outliers can stretch or skew our data? When we have a few extreme values pulling the data in one direction, the shape of the distribution becomes skewed.
- **Symmetry** means the **mean ≈ median**, and there’s no skew.

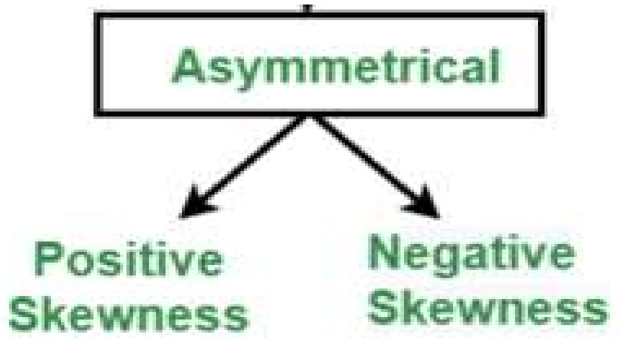


Skewness

- **Definition:** "Skewness measures whether data points are **symmetrically distributed** around the mean, or if they're pulled in one direction more than the other."



Types of Skewness:



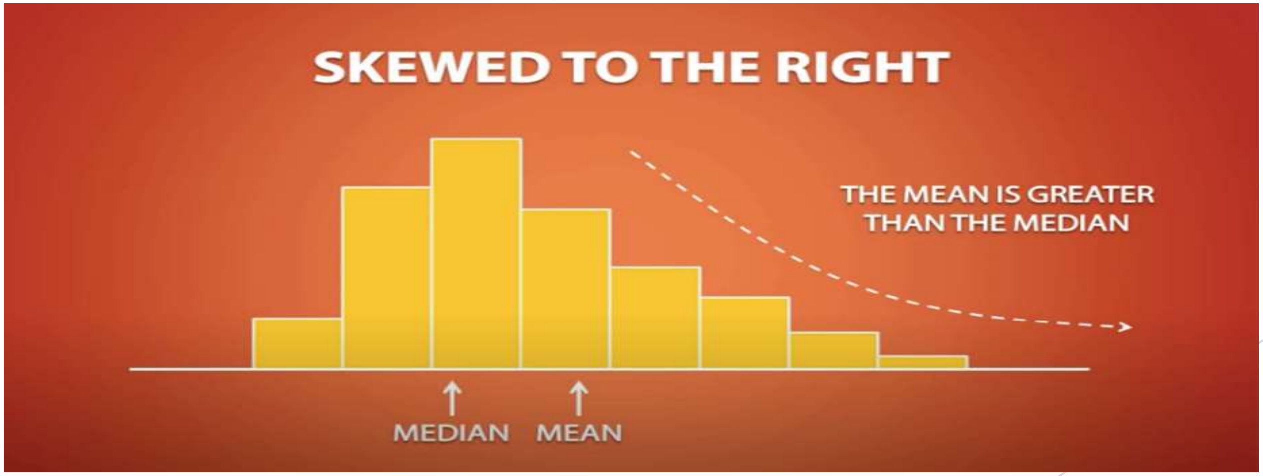
Asymmetrical

Positive Skewness

Negative Skewness

- **Right-Skewed (Positive Skew):** Most data points are concentrated on the left, but there are outliers on the right. Here, the **mean** is greater than the **median**.
- **Left-Skewed (Negative Skew):** Most data points are concentrated on the right, but outliers drag the data to the left. Here, the **mean** is less than the **median**.

- **Right-Skewed Distribution (Positive Skewness):**
- In a **right-skewed** distribution, most of the data points are concentrated on the left side, with a tail extending to the right.
- These distributions are common when a limit prevents lower values but allows for extreme high values.
- Positive skewness (> 0) means the distribution is skewed to the right, with a longer tail on the right side.



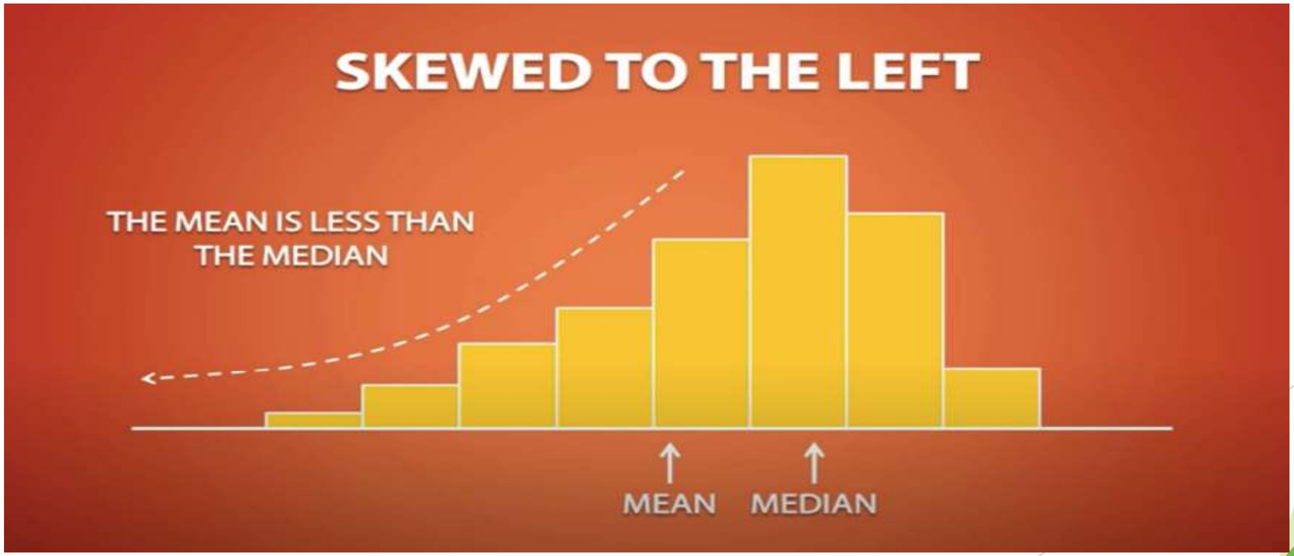
- **Right Skewed:**
- **Example 1: Income Distribution**
- In most countries, the **income distribution** is right-skewed. A majority of people earn below or around the average income, but a small percentage of the population (wealthy individuals) earn extremely high incomes, creating a long tail on the right.
- **Inevitability**: Wealth inequality tends to persist in societies, with a small fraction of people earning significantly more than the majority, leading to a naturally skewed income distribution.

- **Example 2: Housing Prices**
- **Real estate prices** are often right-skewed. Most homes in an area may be priced within a certain range, but there are luxury homes that sell for much higher prices, creating a long right tail.

- **Inevitability**: Differences in neighborhood, home features, and demand drive extreme prices for high-end homes, making the skewness in housing prices inevitable.

- **Example 3: Waiting Times**
- **Customer service waiting times** often exhibit right-skewed behavior. Many people may be served quickly, but a small portion experiences much longer waiting times due to delays or inefficiencies.
- **Inevitability**: Operational inefficiencies, peak demand times, or unforeseen issues result in longer wait times for some customers, making the right tail inevitable.

- **Left-Skewed Distribution (Negative Skewness):**
- In a **left-skewed** distribution, most of the data points are concentrated on the right side, with a tail extending to the left.
- This often occurs when a natural limit prevents high values but allows for extremely low values.
- Negative skewness (< 0) means the distribution is skewed to the left, with a longer tail on the left side.



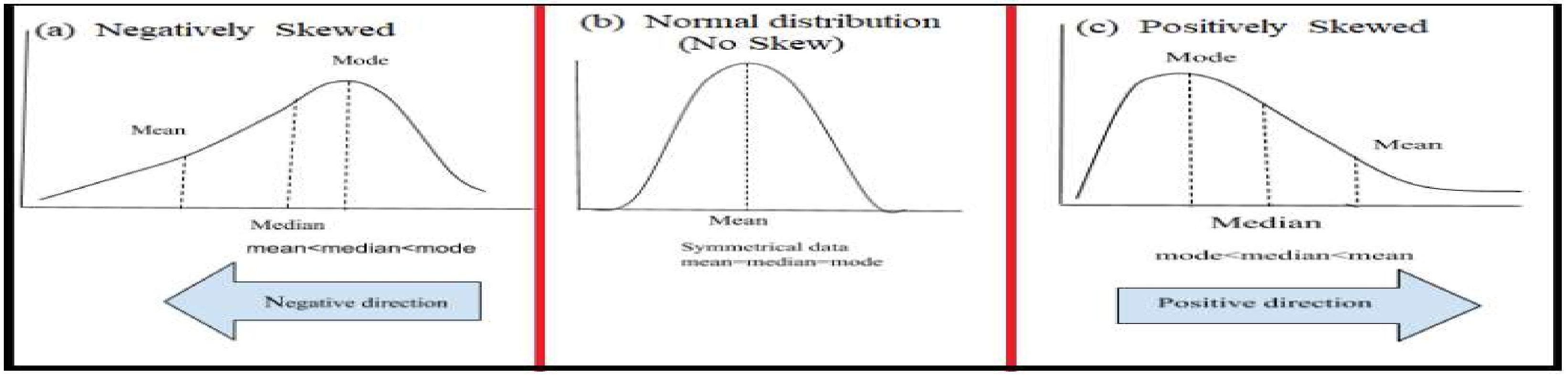
- **Left Skewed:**
- **Example 1: Age at Retirement**
- **Retirement ages** are often left-skewed. Most people retire around a certain age (e.g., 60-65), but a few people retire much earlier, creating a left tail.
- **Inevitability:** The nature of work, policies, and personal preferences ensure that most retirements occur within a narrow window, but early retirement due to financial independence or health issues creates a skewed distribution.

- **Example 2: Time to Complete a Task in a Test**
- **Completion times for tasks or exams** are often left-skewed. Most participants complete the task within the allowed time, but a small number of fast individuals finish much earlier.
- **Inevitability**: While most individuals work at a similar pace, variations in skill level or focus result in some participants finishing exceptionally quickly.

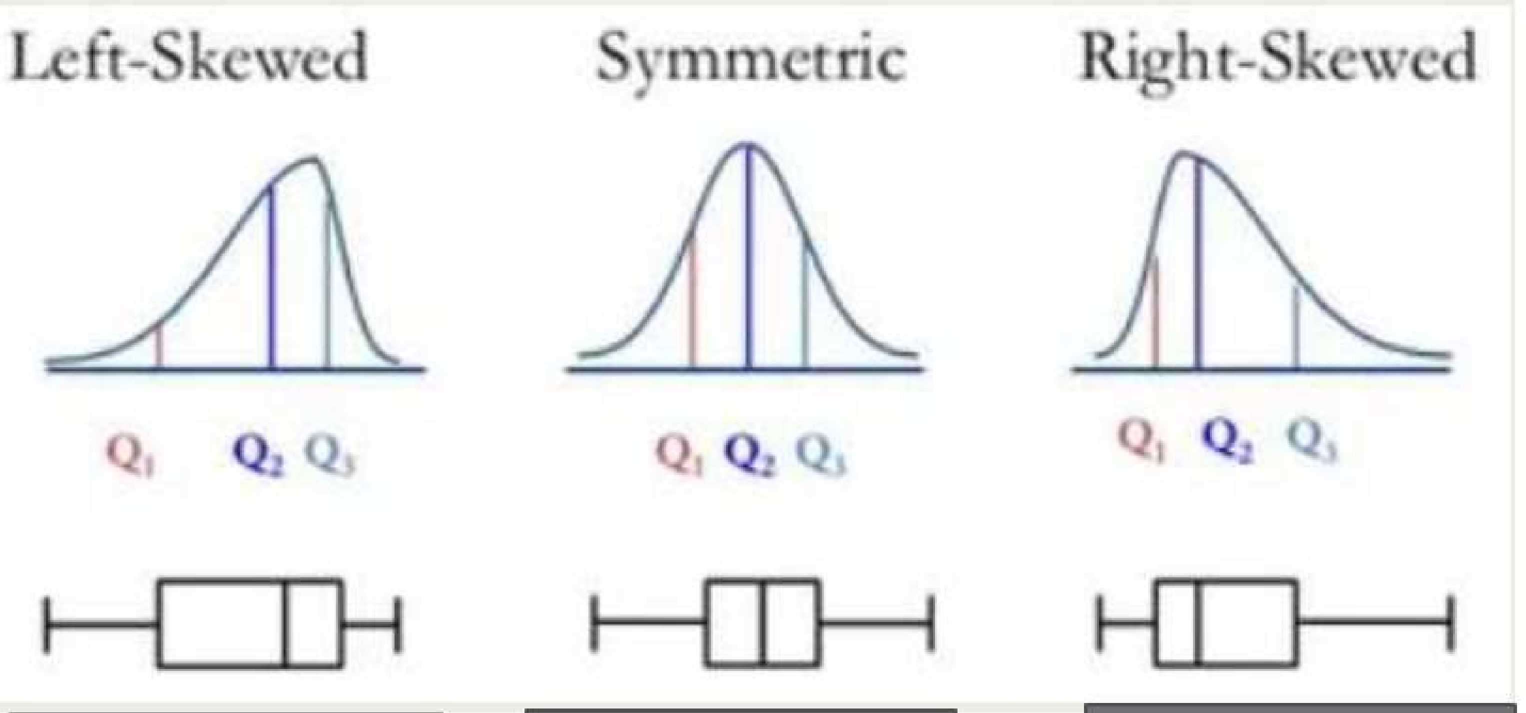
- **Example 3: Loan Repayment Times**
- The **repayment period of loans** can be left-skewed. Many borrowers will repay the loan close to the scheduled time, but a few may repay much earlier.
- **Inevitability:** Financial situations differ, and some borrowers may be able to repay large amounts quickly, leading to early repayments and skewness.

- **In Business Analytics**: Show how **sales data** might be skewed if a few products generate significantly higher sales compared to the majority.
- **In HR Analytics**: Discuss **employee tenure** or **salary distributions**. If a company has a few long-tenured employees or very high-salary individuals, the data will likely be skewed.

- In HR, the **median salary** might be more useful than the mean salary in understanding typical pay, especially if a few employees earn exceptionally high salaries.



- A symmetrical distribution has a skewness of 0, indicating that the data is evenly distributed around the mean.
- If the skewness is between -0.5 and 0.5, the data are **fairly symmetrical.**
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are **moderately skewed.**
- If the skewness is less than -1 or greater than 1, the data are **highly skewed.**

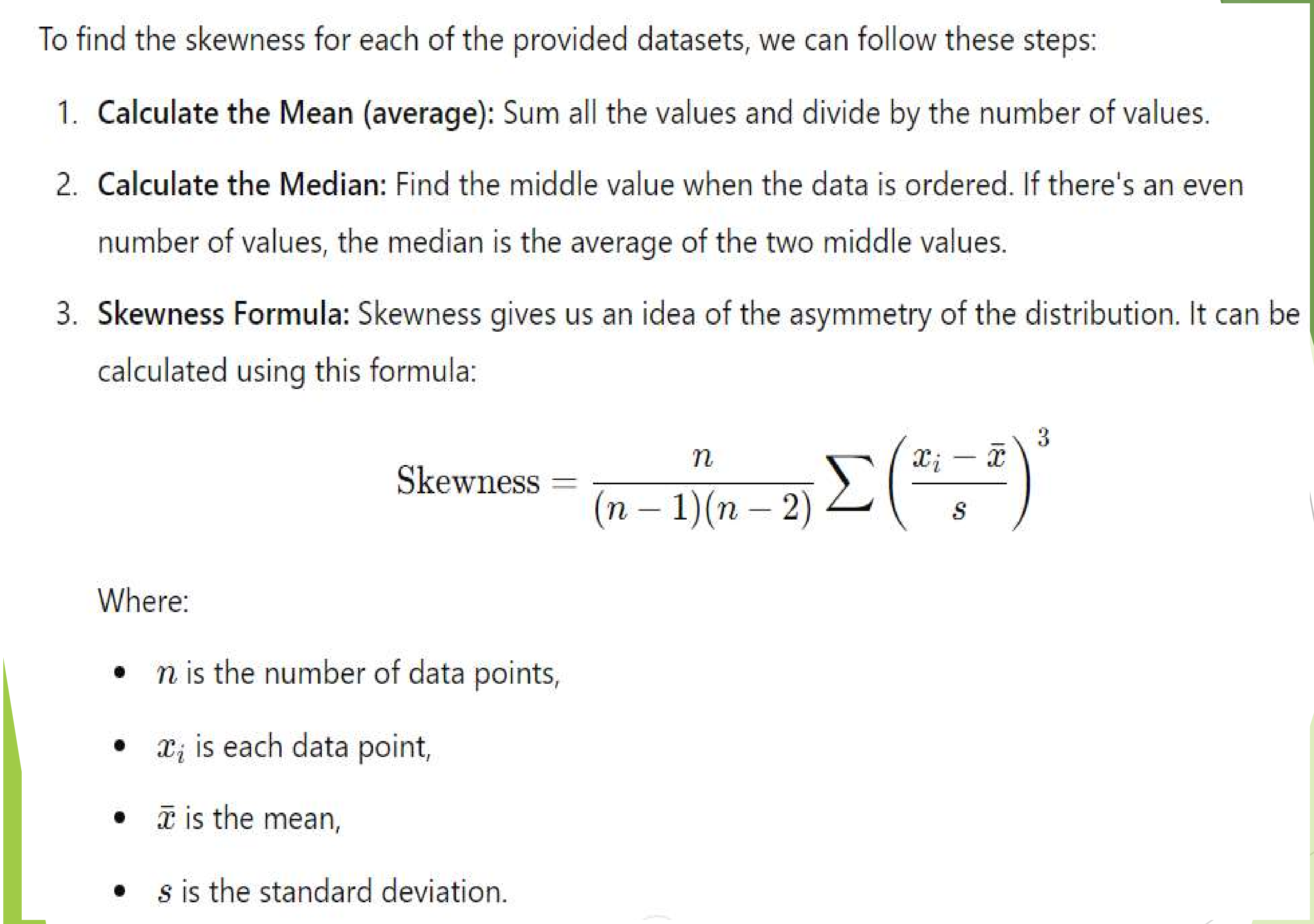


Q2 –Q1 > Q3- Q2

Q2 –Q1 = Q3- Q2

Q2 –Q1 < Q3- Q2

- Ques 1: 4,5,6,6,6,7,7,7,7,7,7,7,8,8,8,9,10.
- Ques 2: 4,5,6,6,6,7,7,7,7,8.
- Ques 3: 6,7,7,7,7,8,8,8,9,10
- Find: Mean, median, mode, RS/LS/ZS.



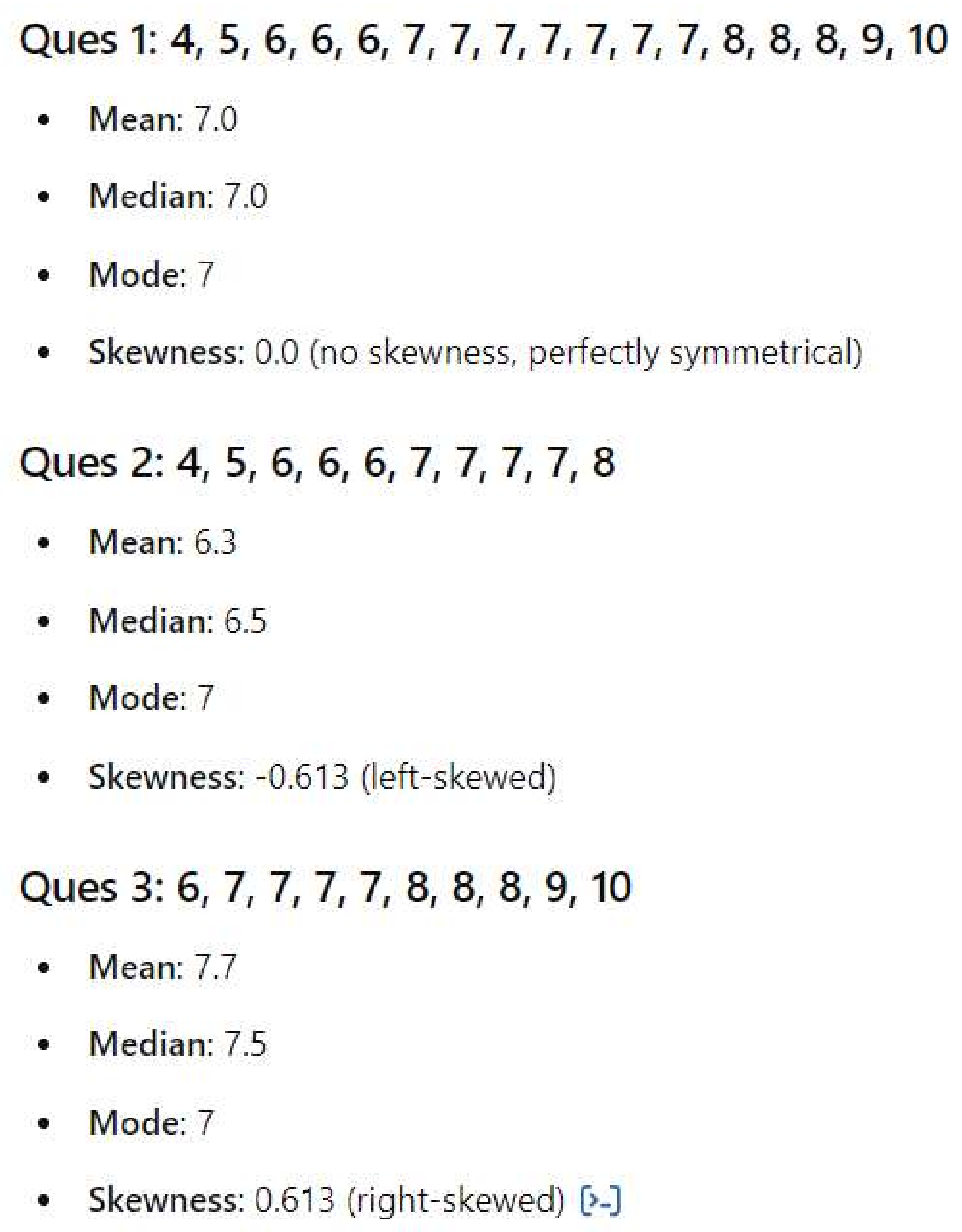
To find the skewness for each of the provided datasets, we can follow these steps:

1. **Calculate the Mean (average):** Sum all the values and divide by the number of values.
2. **Calculate the Median:** Find the middle value when the data is ordered. If there's an even number of values, the median is the average of the two middle values.
3. **Skewness Formula:** Skewness gives us an idea of the asymmetry of the distribution. It can be calculated using this formula:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Where:

- n is the number of data points,
- x_i is each data point,
- $\bar{x}$ is the mean,
- s is the standard deviation.



Ques 1: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

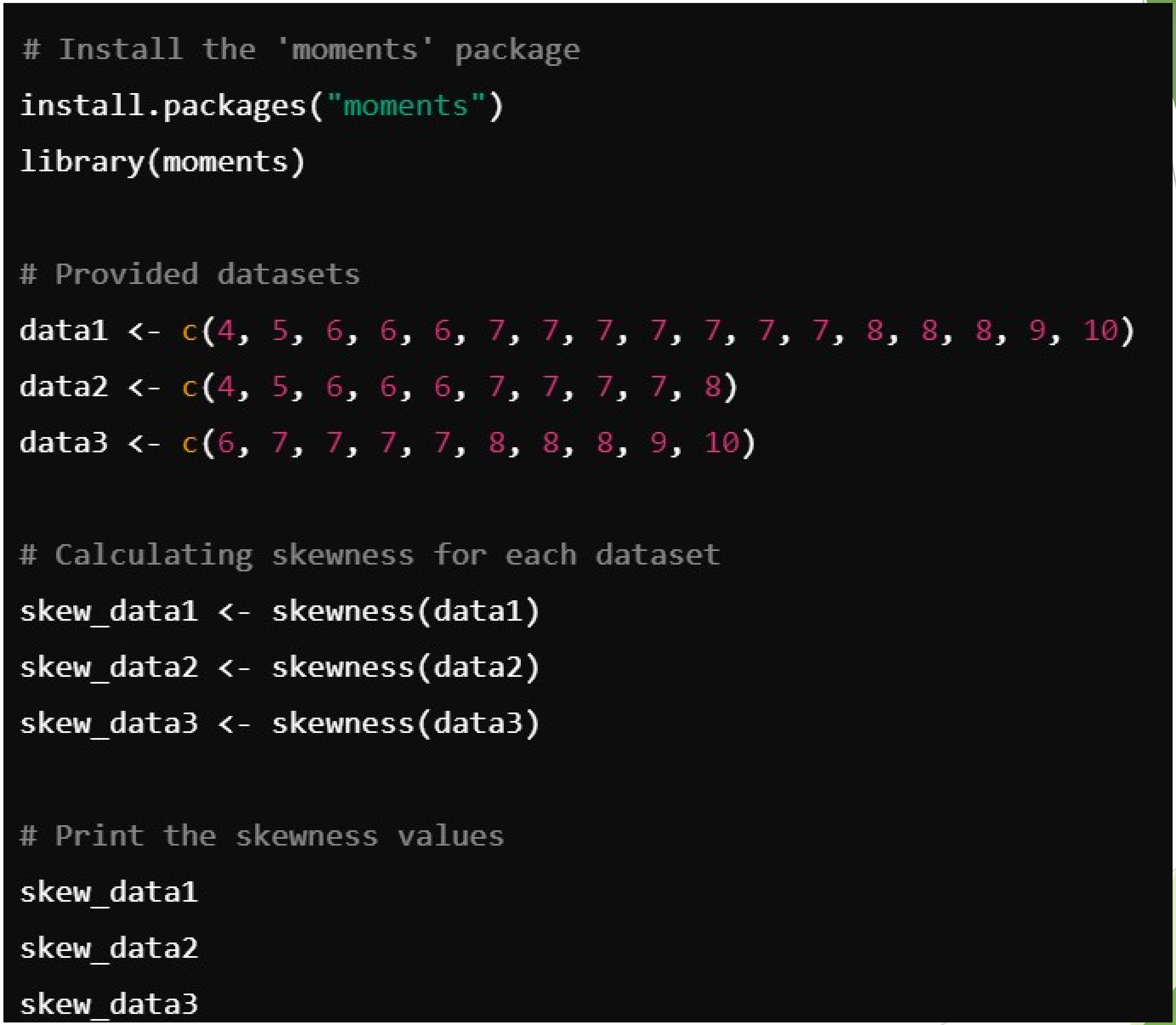
- **Mean**: 7.0
- **Median**: 7.0
- **Mode**: 7
- **Skewness**: 0.0 (no skewness, perfectly symmetrical)

Ques 2: 4, 5, 6, 6, 6, 7, 7, 7, 7, 8

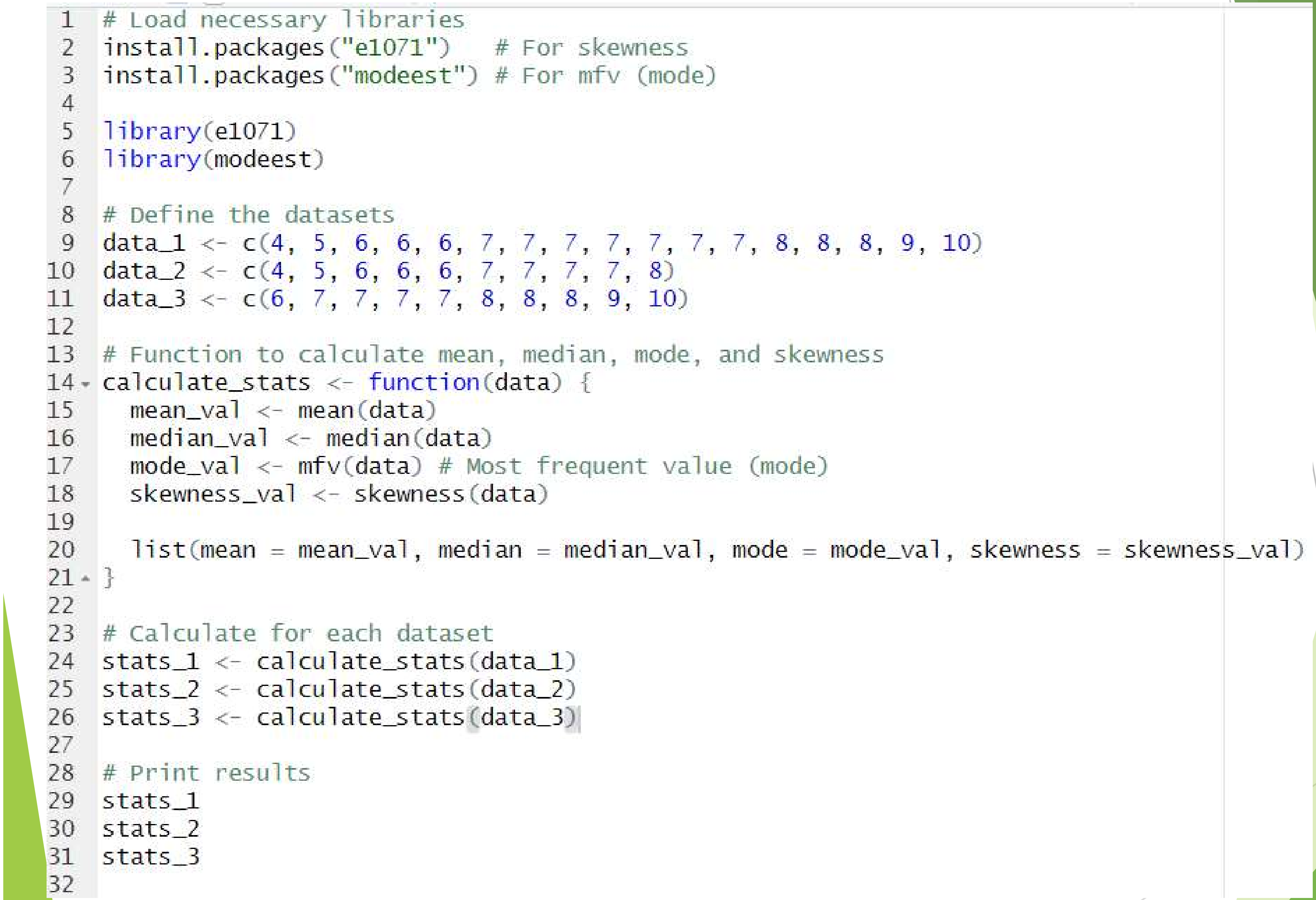
- **Mean**: 6.3
- **Median**: 6.5
- **Mode**: 7
- **Skewness**: -0.613 (left-skewed)

Ques 3: 6, 7, 7, 7, 7, 8, 8, 8, 9, 10

- **Mean**: 7.7
- **Median**: 7.5
- **Mode**: 7
- **Skewness**: 0.613 (right-skewed)



```r
# Install the 'moments' package
install.packages("moments")
library(moments)

# Provided datasets
data1 <- c(4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10)
data2 <- c(4, 5, 6, 6, 6, 7, 7, 7, 7, 8)
data3 <- c(6, 7, 7, 7, 7, 8, 8, 8, 9, 10)

# Calculating skewness for each dataset
skew_data1 <- skewness(data1)
skew_data2 <- skewness(data2)
skew_data3 <- skewness(data3)

# Print the skewness values
skew_data1
skew_data2
skew_data3
```



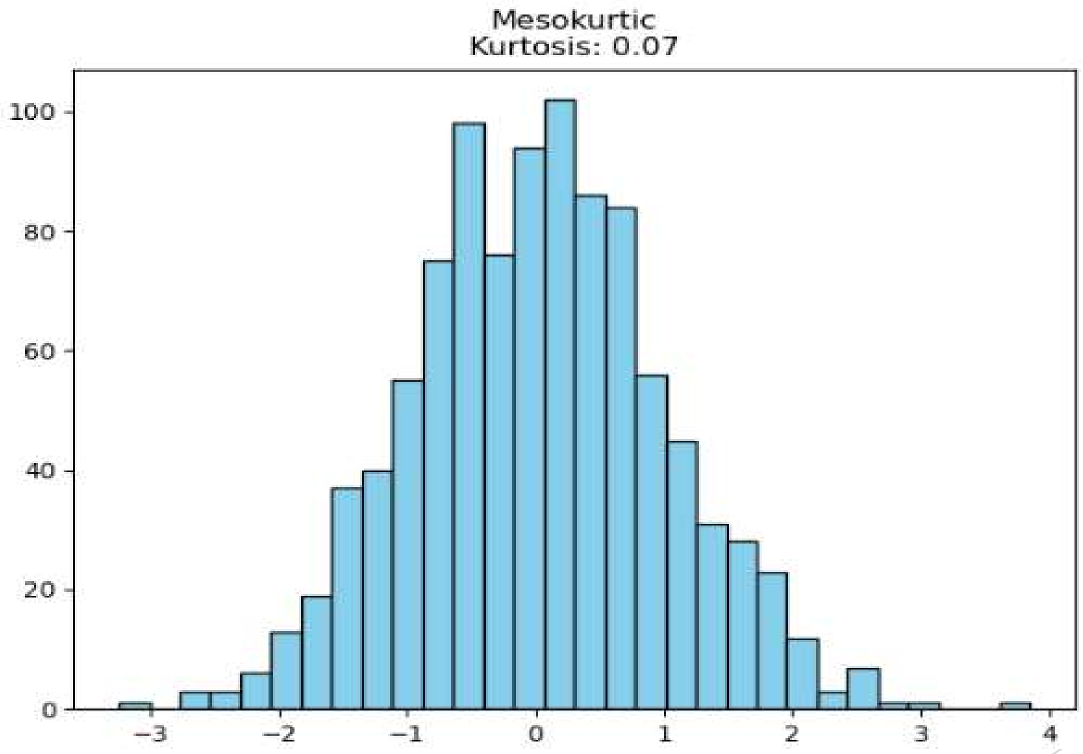
```r
# Load necessary libraries
install.packages("e1071")   # For skewness
install.packages("modeest") # For mfv (mode)

library(e1071)
library(modeest)

# Define the datasets
data_1 <- c(4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10)
data_2 <- c(4, 5, 6, 6, 6, 7, 7, 7, 7, 8)
data_3 <- c(6, 7, 7, 7, 7, 8, 8, 8, 9, 10)

# Function to calculate mean, median, mode, and skewness
calculate_stats <- function(data) {
  mean_val <- mean(data)
  median_val <- median(data)
  mode_val <- mfv(data) # Most frequent value (mode)
  skewness_val <- skewness(data)

  list(mean = mean_val, median = median_val, mode = mode_val, skewness = skewness_val)
}

# Calculate for each dataset
stats_1 <- calculate_stats(data_1)
stats_2 <- calculate_stats(data_2)
stats_3 <- calculate_stats(data_3)

# Print results
stats_1
stats_2
stats_3

```

Kurtosis

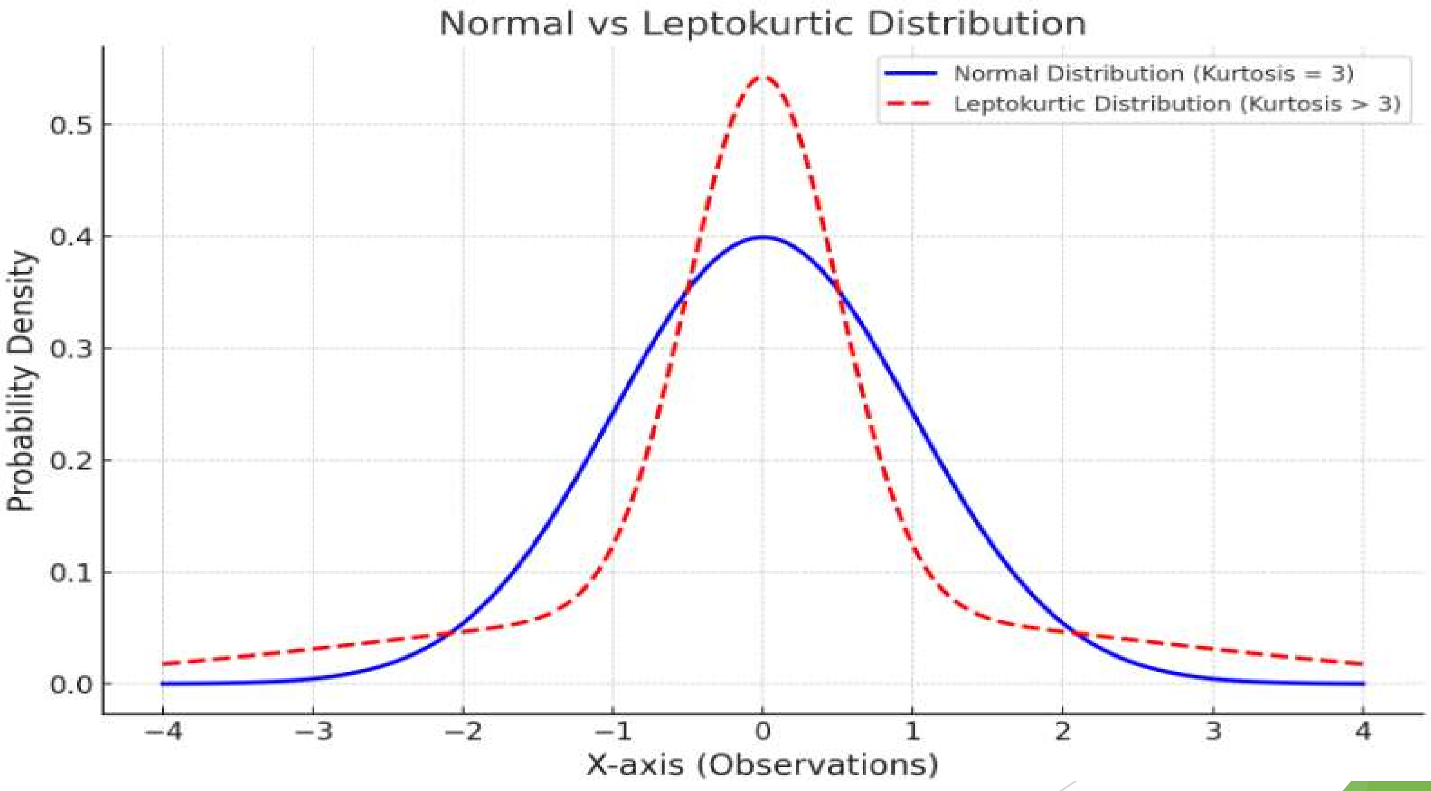
- **Definition**: Kurtosis measures whether the data has **heavy tails** or **light tails** compared to a normal distribution.
- It shows whether extreme values (outliers) occur more or less frequently.
- **Example**: "Kurtosis answers the question: Is the data mostly concentrated in the middle, or are there extreme values in the tails?"

- **Mesokurtic**: Kurtosis ≈ 3 (Normal distribution)
- **Leptokurtic**: Kurtosis > 3 (Heavy tails)
- **Platykurtic**: Kurtosis < 3 (Light tails)

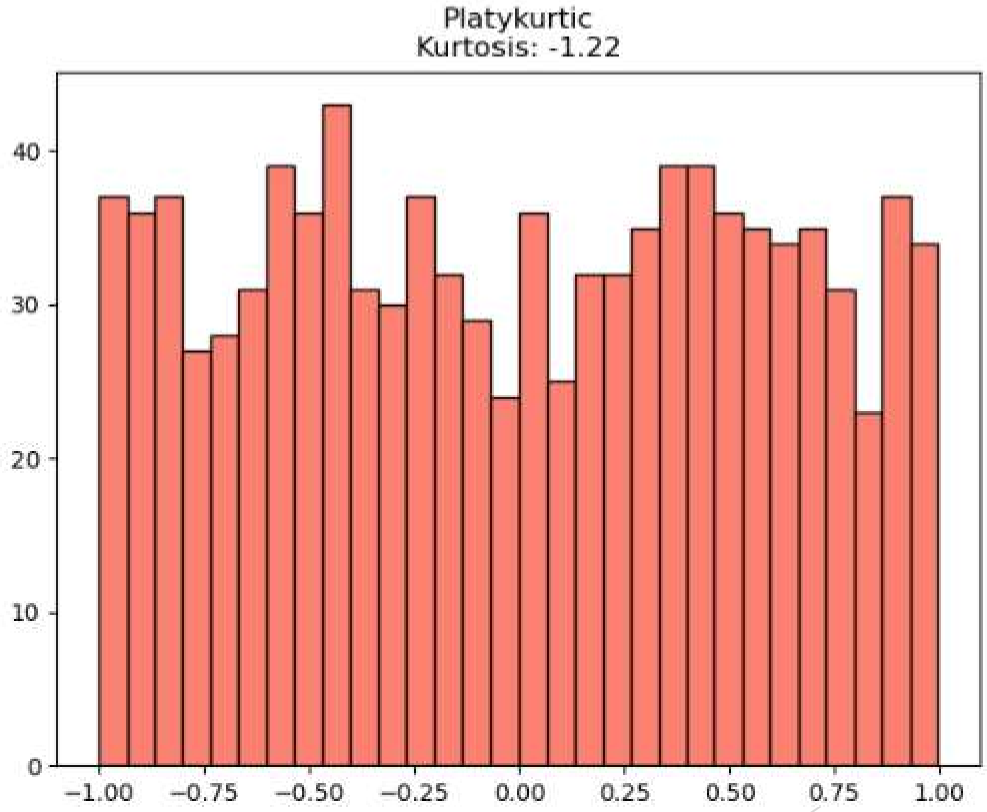
- **Mesokurtic (Normal Distribution):**This is the reference point: a distribution with **normal kurtosis** (kurtosis = 3).
  - **Real-world example**: Height distribution among a large population is often close to mesokurtic, where most people’s heights are near the average.



- **Leptokurtic (Heavy Tails):Leptokurtic distributions** have **sharp peaks** and **fat tails**, meaning there are **more outliers.** The kurtosis value is **greater than 3.**
  - **Real-world example:** Income data may show a leptokurtic distribution, where most people have an average income, but there are a few extremely high earners (outliers).



- **Platykurtic (Light Tails):Platykurtic distributions** have **flatter peaks** and **thin tails**, meaning there are **fewer outliers.** The kurtosis value is **less than 3.**
- **Real-world example:** Test scores where most students perform similarly, without much deviation from the average, would likely have platykurtic distribution.



Understanding Kurtosis Beyond the Formula

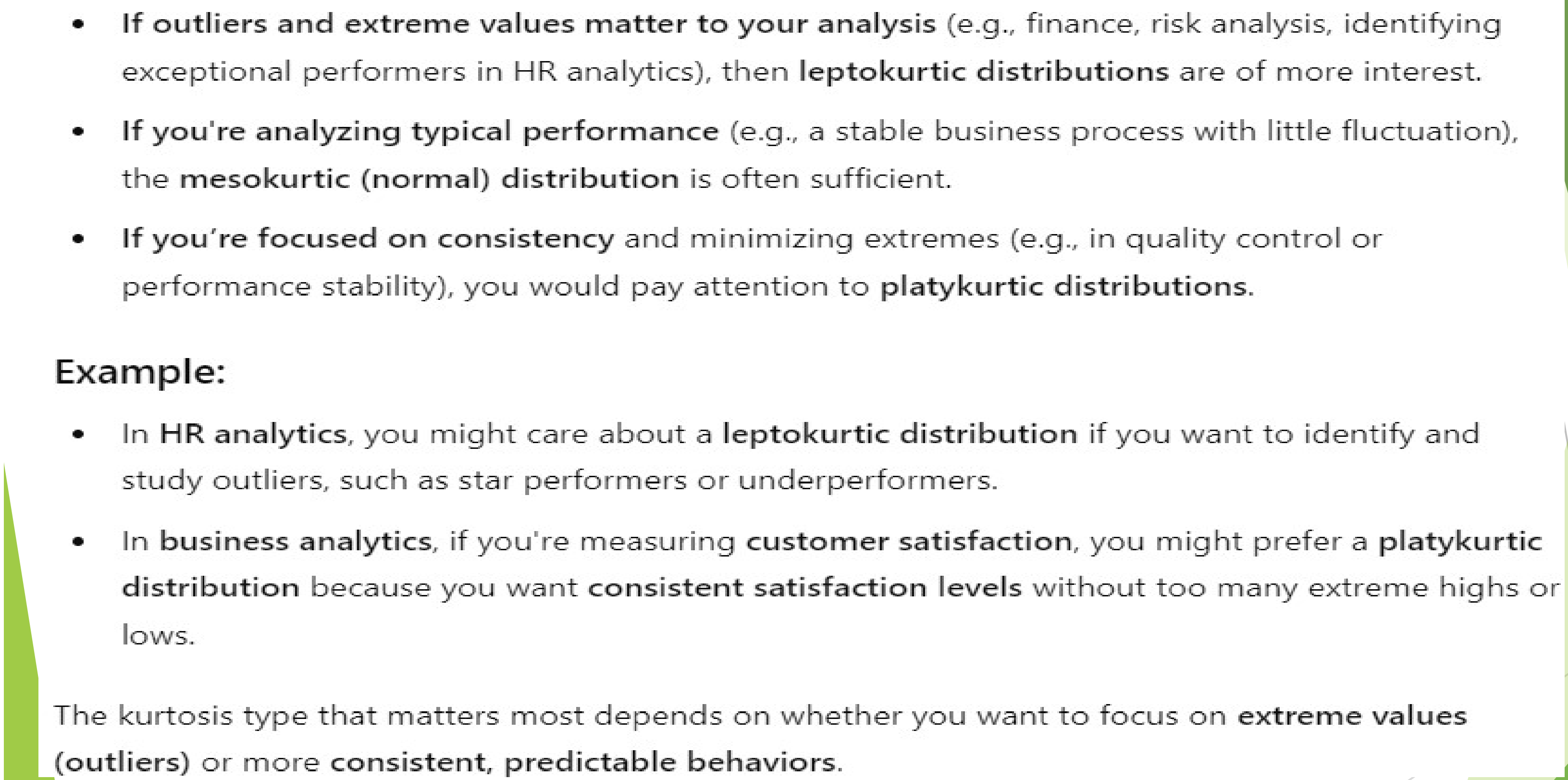
- While kurtosis is mathematically defined using the fourth moment, its practical interpretation revolves around the propensity of the distribution to produce outliers:
- **High Kurtosis (Leptokurtic)**: More data in the tails and a sharper peak. This implies a higher likelihood of extreme values (outliers).
- **Low Kurtosis (Platykurtic)**: Less data in the tails and a flatter peak. This suggests fewer extreme values.

- Let’s look at some examples to illustrate how kurtosis relates to data observations:
- **Normal Distribution (Kurtosis = 3)**
  - **Tail Behavior**: Approximately 0.3% of data lies beyond ±3 standard deviations from the mean.
  - **Interpretation**: Baseline for tail comparisons.
- **Leptokurtic Distribution (Kurtosis > 3)**
  - **Example**: Kurtosis = 5
  - **Tail Behavior**: More than 0.3% of data lies beyond ±3 standard deviations.
  - **Interpretation**: Higher chance of extreme values (outliers). For instance, instead of 0.3%, you might observe 0.5% or more in the tails.
- **Platykurtic Distribution (Kurtosis < 3)**
  - **Example**: Kurtosis = 2
  - **Tail Behavior**: Less than 0.3% of data lies beyond ±3 standard deviations.
  - **Interpretation**: Fewer extreme values. Data points are more uniformly spread around the mean.

Example Scenario

- **Scenario:** You have 1,000 observations of daily returns for a stock.
- **Calculated Kurtosis:** 4.5 (Excess Kurtosis = 1.5)
- **Interpretation:**
- **Leptokurtic:** Higher probability of extreme returns (both positive and negative) compared to a normal distribution.
- **Practical Implications:**
- **Risk Assessment:** The stock may experience more significant gains or losses than expected under normal assumptions.
- **Action:** Use risk management strategies that account for higher volatility and potential extreme movements.

Which Kurtosis Should You Focus On?



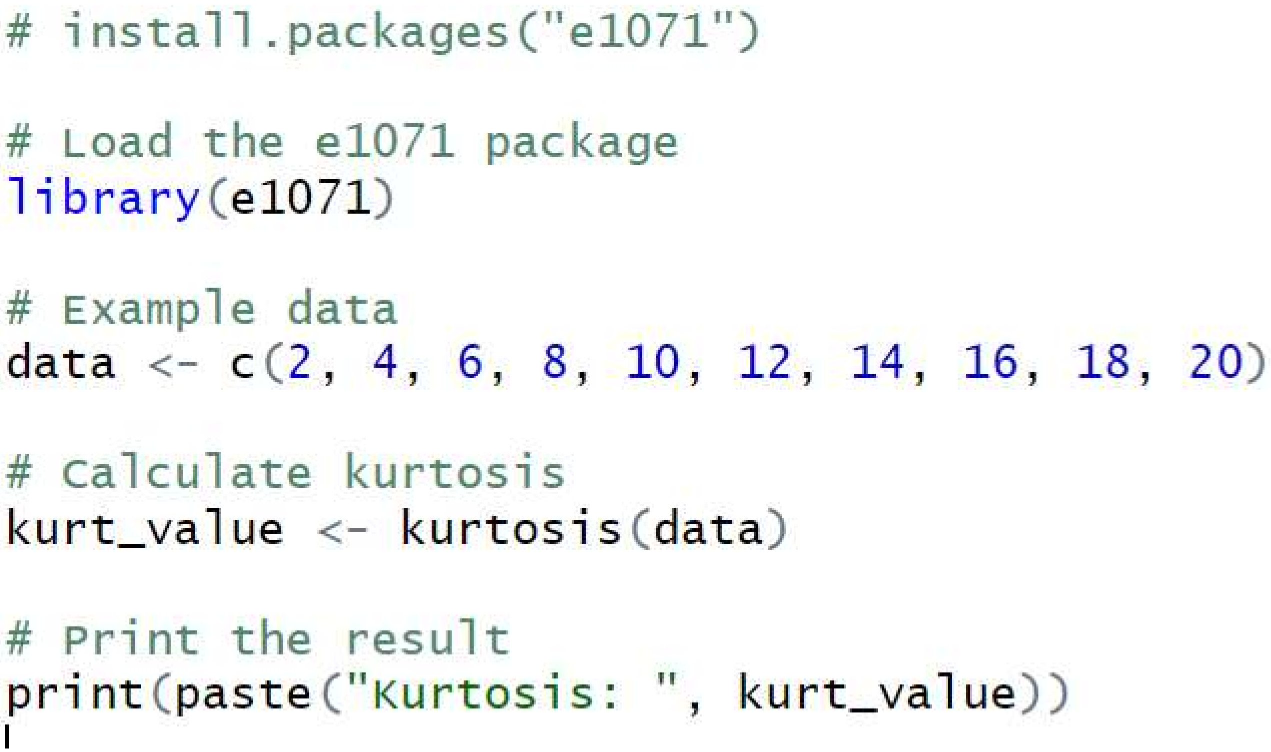
- **If outliers and extreme values matter to your analysis** (e.g., finance, risk analysis, identifying exceptional performers in HR analytics), then **leptokurtic distributions** are of more interest.
- **If you're analyzing typical performance** (e.g., a stable business process with little fluctuation), the **mesokurtic (normal) distribution** is often sufficient.
- **If you're focused on consistency** and minimizing extremes (e.g., in quality control or performance stability), you would pay attention to **platykurtic distributions.**

Example:

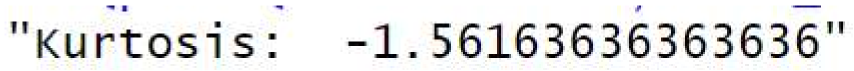
- In **HR analytics**, you might care about a **leptokurtic distribution** if you want to identify and study outliers, such as star performers or underperformers.
- In **business analytics**, if you're measuring **customer satisfaction**, you might prefer a **platykurtic distribution** because you want **consistent satisfaction levels** without too many extreme highs or lows.

The kurtosis type that matters most depends on whether you want to focus on **extreme values (outliers)** or more **consistent, predictable behaviors**.

Code in R



```
# install.packages("e1071")

# Load the e1071 package
library(e1071)

# Example data
data <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)

# Calculate kurtosis
kurt_value <- kurtosis(data)

# Print the result
print(paste("Kurtosis: ", kurt_value))
```



```
 "Kurtosis:  -1.56163636363636"
```

Review of Key Journey till now

- **So till now we talked about :**

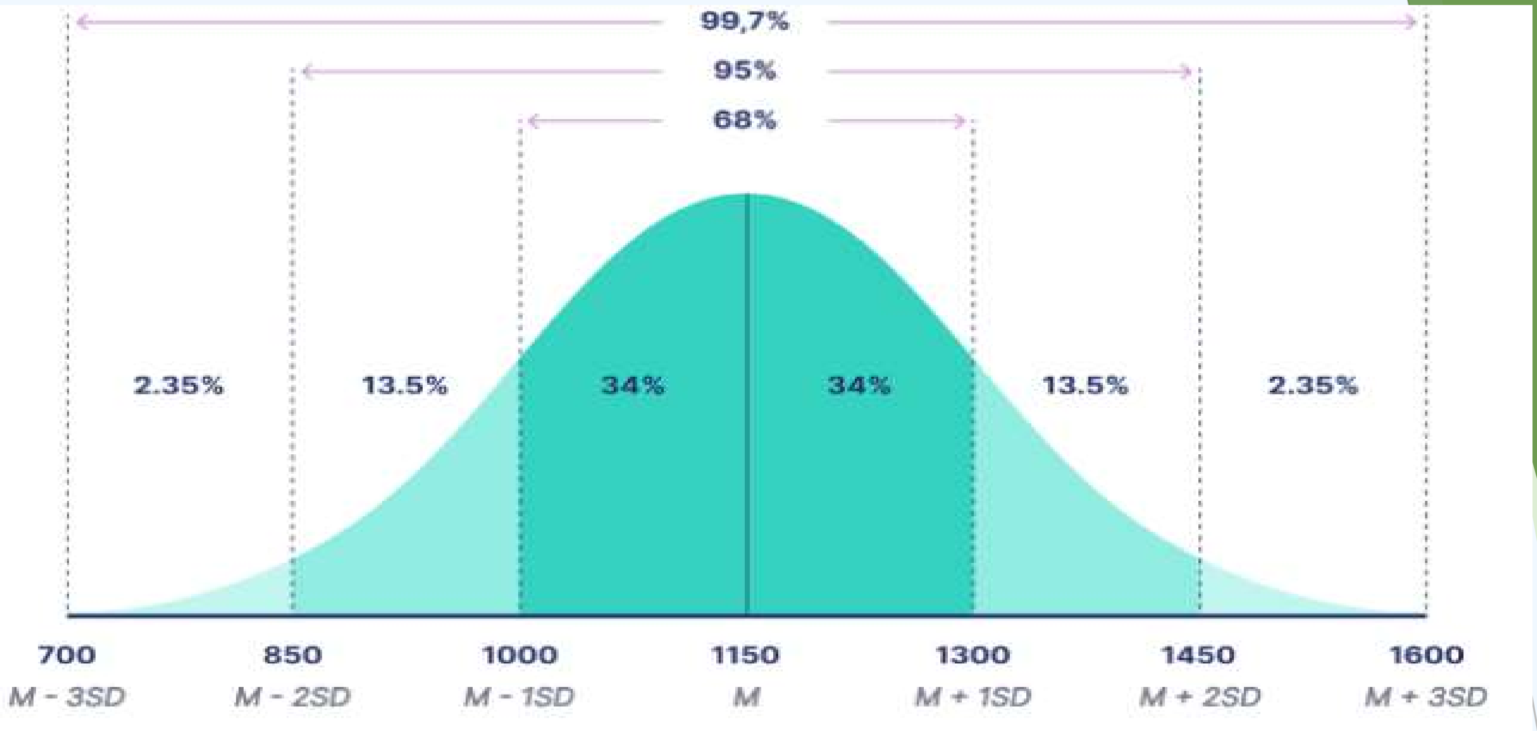
- **Mean:** The average of a dataset.
- **Variance:** Measures the spread of the data around the mean.
- **Skewness:** Indicates the asymmetry of the data distribution.
- **Kurtosis:** Reflects the "tailedness" of the distribution.
- **Outliers:** Data points that are significantly different from others.

Introducing the Normal Distribution

- **Definition:**
- The **normal distribution** is a continuous probability distribution characterized by its bell-shaped curve, which is symmetric around the mean.
- **Key Properties:**
- **Symmetry:** The left and right sides of the distribution are mirror images.
- **Mean, Median, Mode:** All three are equal and located at the center of the distribution.
- **Defined by Mean (μ) and Variance (σ²):** These parameters determine the center and spread of the distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

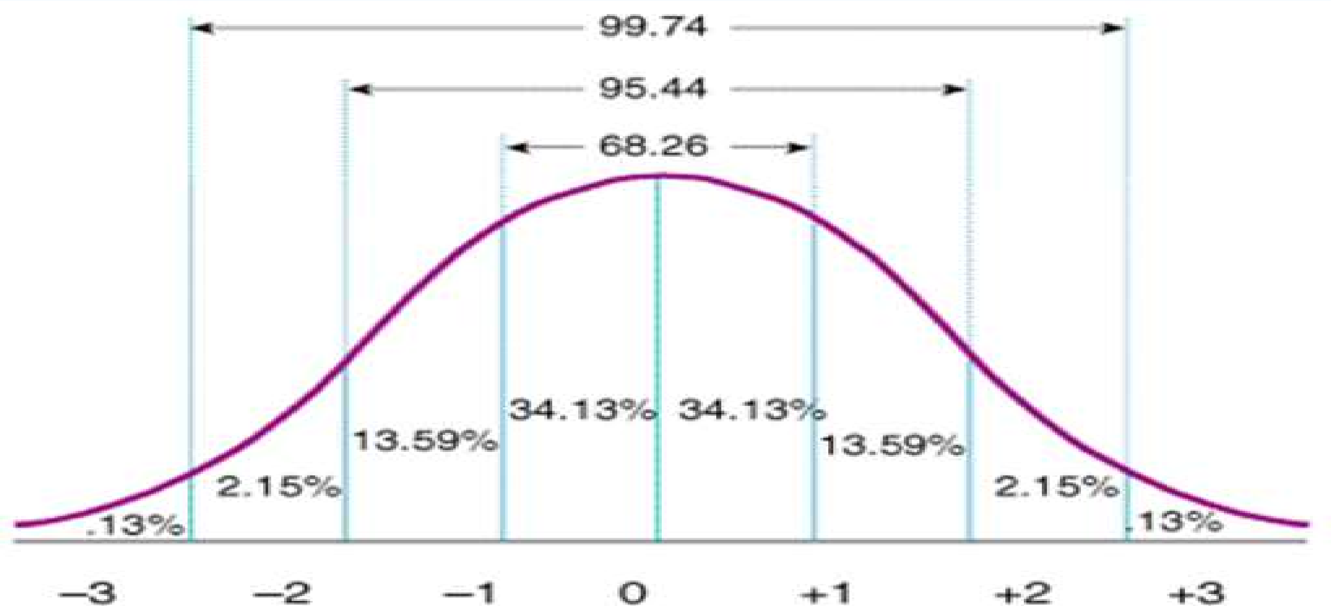
- In analytics it is assumed that data is normally distributed for analysis due its occurrence in nature and real world phenomena.
- Because it is assumed that for normal distribution, 99.7% of all the values will fall with in 3* S.D. of the mean on either side on curve.
- It is also known as Gaussian Distribution or Bell curve.
- It talks about symmetric around its mean and has a single peak at the centre of the distribution.



99,7%
95%
68%
2.35%
13.5%
34%
34%
13.5%
2.35%
700
M - 3SD
850
M - 2SD
1000
M - 1SD
1150
M
1300
M + 1SD
1450
M + 2SD
1600
M + 3SD

Introducing the Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution.
- It has the following characteristics:
  - **Mean (μ)**: 0
  - **Standard Deviation (σ)**: 1



Importance:

- **Simplification:** Allows for easier calculations and the use of standard tables (Z-tables).
- **Z-Scores:** Any normal distribution can be converted to the standard normal distribution using Z-scores.
- It is also called Z distribution.
- Z-score formula:

$$Z = (x - \mu)/\sigma$$

- The SND is also a probability distribution, so the area under the curve between two points tells the probability of variables taking on a range of values.
- The total area under the curve is 1 or 100%.
- This process, called standardization or normalization, converts any value from a normal distribution to the standard normal distribution by subtracting the mean and dividing by the standard deviation.
- The standardization allows for meaningful comparisons across different datasets with different scales and units.

- The **Z-score** is primarily a tool used in **inferential statistics**, but it also has applications in **descriptive statistics**.
- **In Descriptive Statistics:**
  - **Z-score** is used to describe the position of a single data point relative to the mean of a dataset.

- **In Inferential Statistics:**
  - **Z-score** is crucial for hypothesis testing and confidence interval estimation.
  - It tells us how far the mean of sample from population mean.

- **Interpretation:** A Z score tells you how far a particular value is from the mean of the distribution in terms of standard deviations.
  - A Z score of 0 indicates that the value is at the mean.
  - A Z score of 1 indicates that the value is one standard deviation above the mean.
  - A Z score of -1 indicates that the value is one standard deviation below the mean.
  - Typically, values with z-scores beyond a certain threshold (e.g., z-score > 3 or z-score < -3) are considered outliers.
  - Z scores tell you how many standard deviations from the mean each value lies.

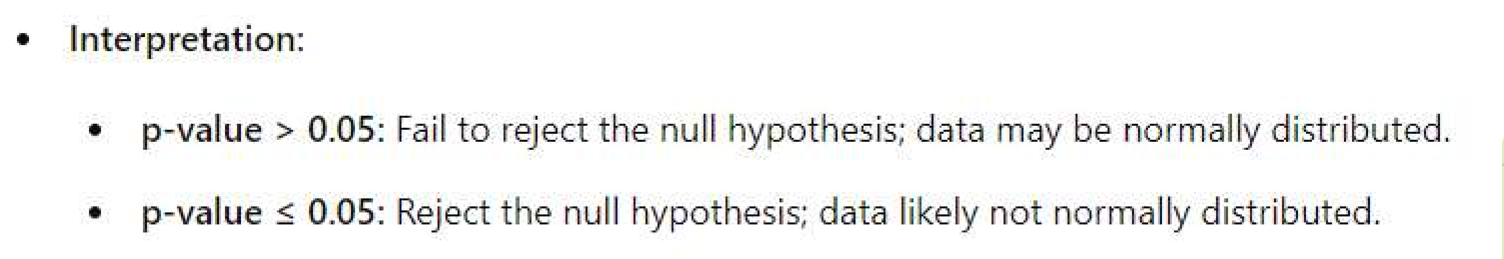
Connecting the Dots: From Descriptive Statistics to Distribution

- **Descriptive Statistics (Mean, Variance, Skewness, Kurtosis):**
  - These measures describe the key features of any dataset, including its central tendency, spread, asymmetry, and tail behavior.
- **Normal Distribution:**
  - A specific type of distribution where skewness is 0 and kurtosis is 3.
  - Defined completely by its mean and variance.
  - Serves as a foundational model in statistics due to the Central Limit Theorem.
- **Standard Normal Distribution:**
  - A normalized form of the normal distribution.
  - Facilitates the calculation of probabilities and the comparison of different datasets.
- **Outliers:**
  - In the context of a normal distribution, outliers are rare (extremely high or low values).
  - Understanding the normal distribution helps in identifying and interpreting outliers.

Practical Applications and Examples

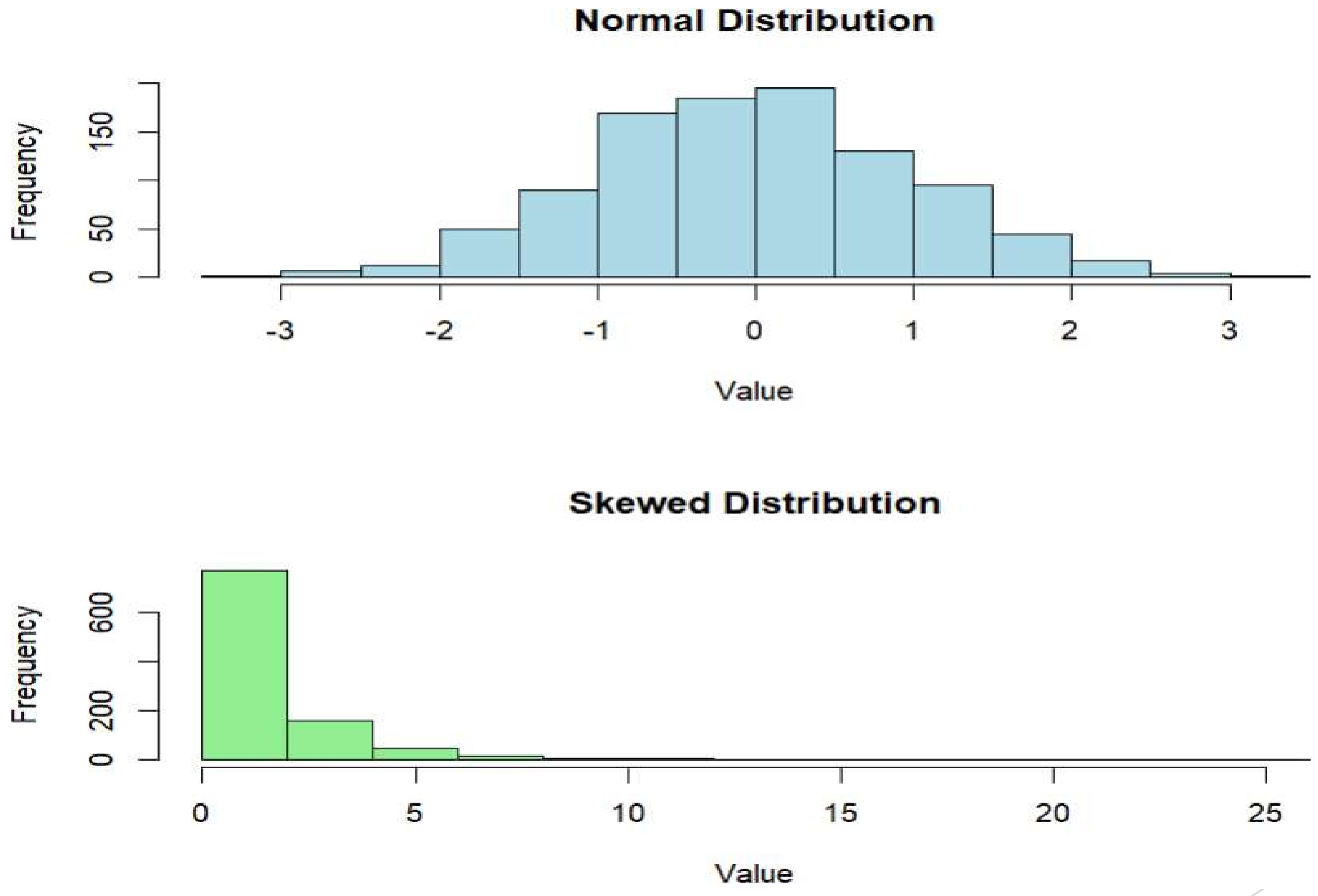
- **Example 1: Checking Normality of Data**
- *Using Shapiro-Wilk Test in R:*
- # Shapiro-Wilk test for normality
- shapiro_test <- shapiro.test(data)
- print(shapiro_test)

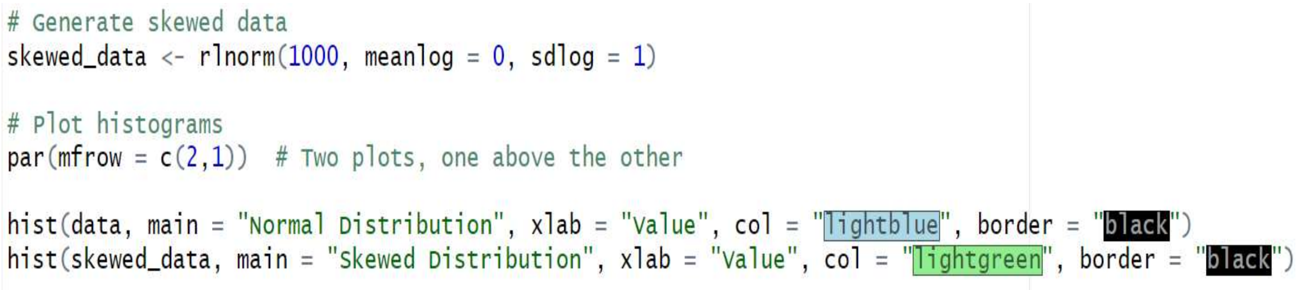
Interpretation: If p-value > 0.05, data is likely normally distributed.



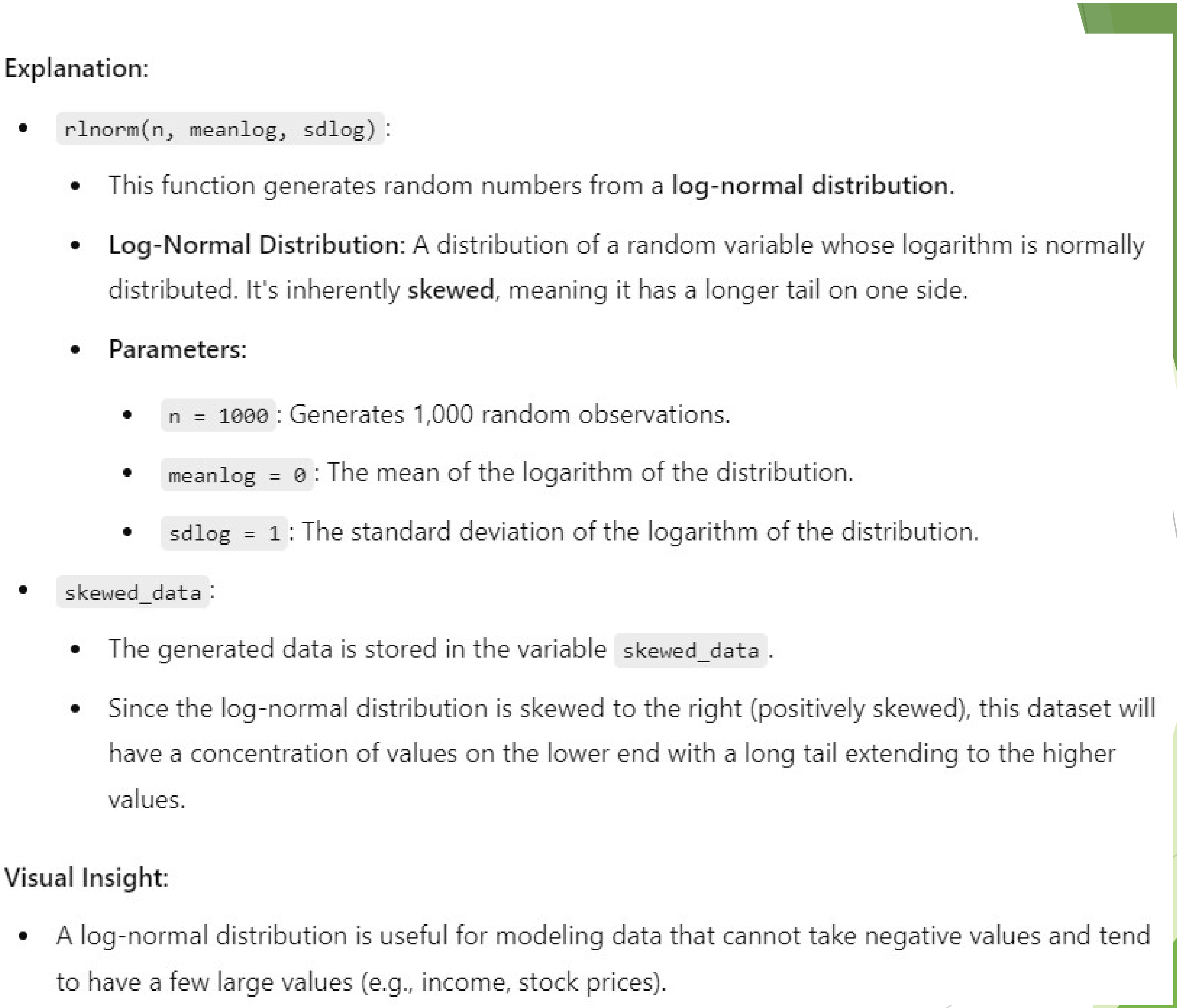
- Interpretation:
  - **p-value > 0.05:** Fail to reject the null hypothesis; data may be normally distributed.
  - **p-value ≤ 0.05:** Reject the null hypothesis; data likely not normally distributed.

Example 2: Visualizing Normal vs. Skewed Distribution





```
# Generate skewed data
skewed_data <- rlnorm(1000, meanlog = 0, sdlog = 1)

# Plot histograms
par(mfrow = c(2,1))  # Two plots, one above the other

hist(data, main = "Normal Distribution", xlab = "Value", col = "lightblue", border = "black")
hist(skewed_data, main = "Skewed Distribution", xlab = "Value", col = "lightgreen", border = "black")
```

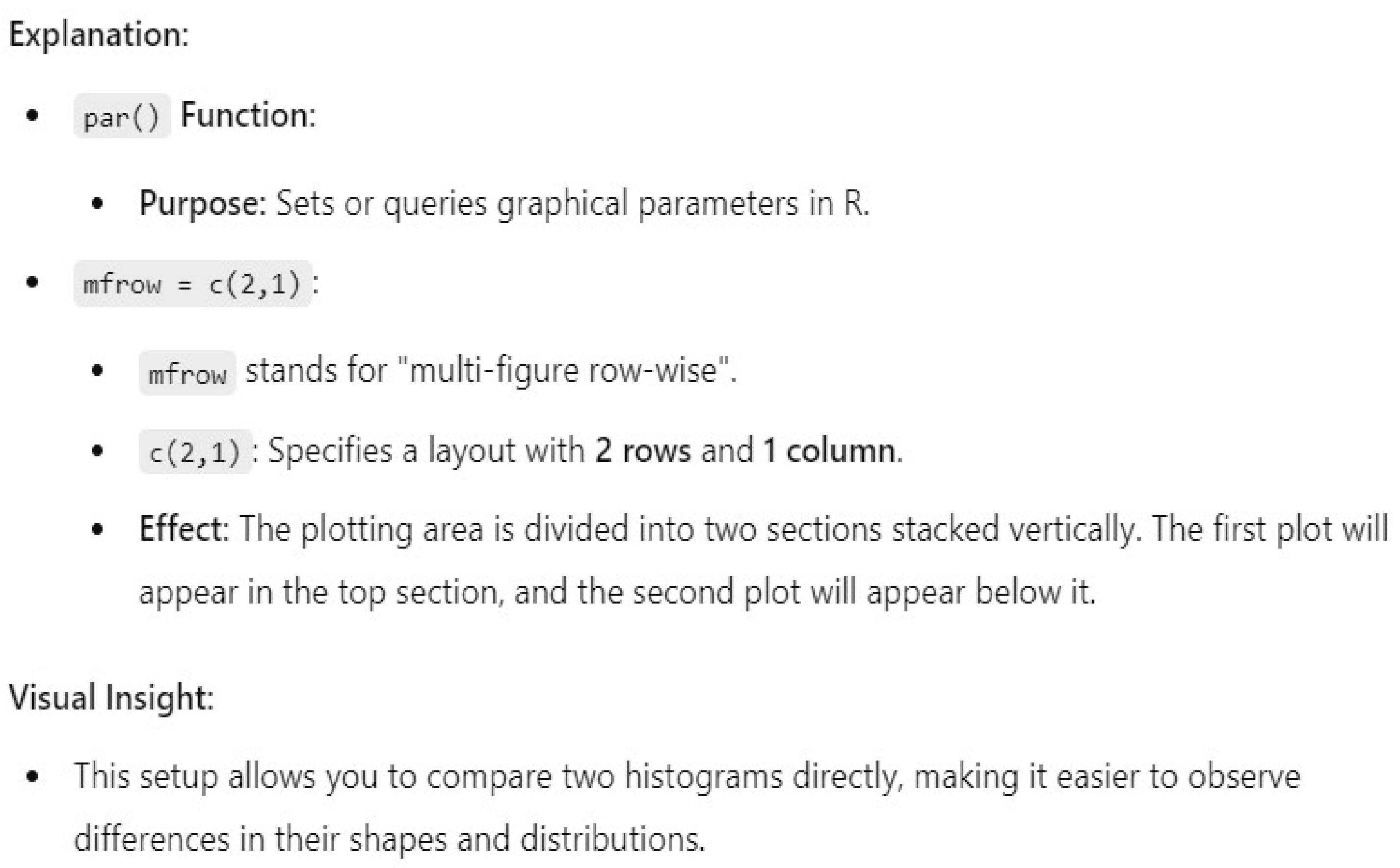


Explanation:

- `rlnorm(n, meanlog, sdlog)`:
  - This function generates random numbers from a **log-normal distribution**.
  - **Log-Normal Distribution:** A distribution of a random variable whose logarithm is normally distributed. It's inherently **skewed**, meaning it has a longer tail on one side.
  - **Parameters:**
    - `n = 1000`: Generates 1,000 random observations.
    - `meanlog = 0`: The mean of the logarithm of the distribution.
    - `sdlog = 1`: The standard deviation of the logarithm of the distribution.
- `skewed_data`:
  - The generated data is stored in the variable `skewed_data`.
  - Since the log-normal distribution is skewed to the right (positively skewed), this dataset will have a concentration of values on the lower end with a long tail extending to the higher values.

Visual Insight:

- A log-normal distribution is useful for modeling data that cannot take negative values and tend to have a few large values (e.g., income, stock prices).

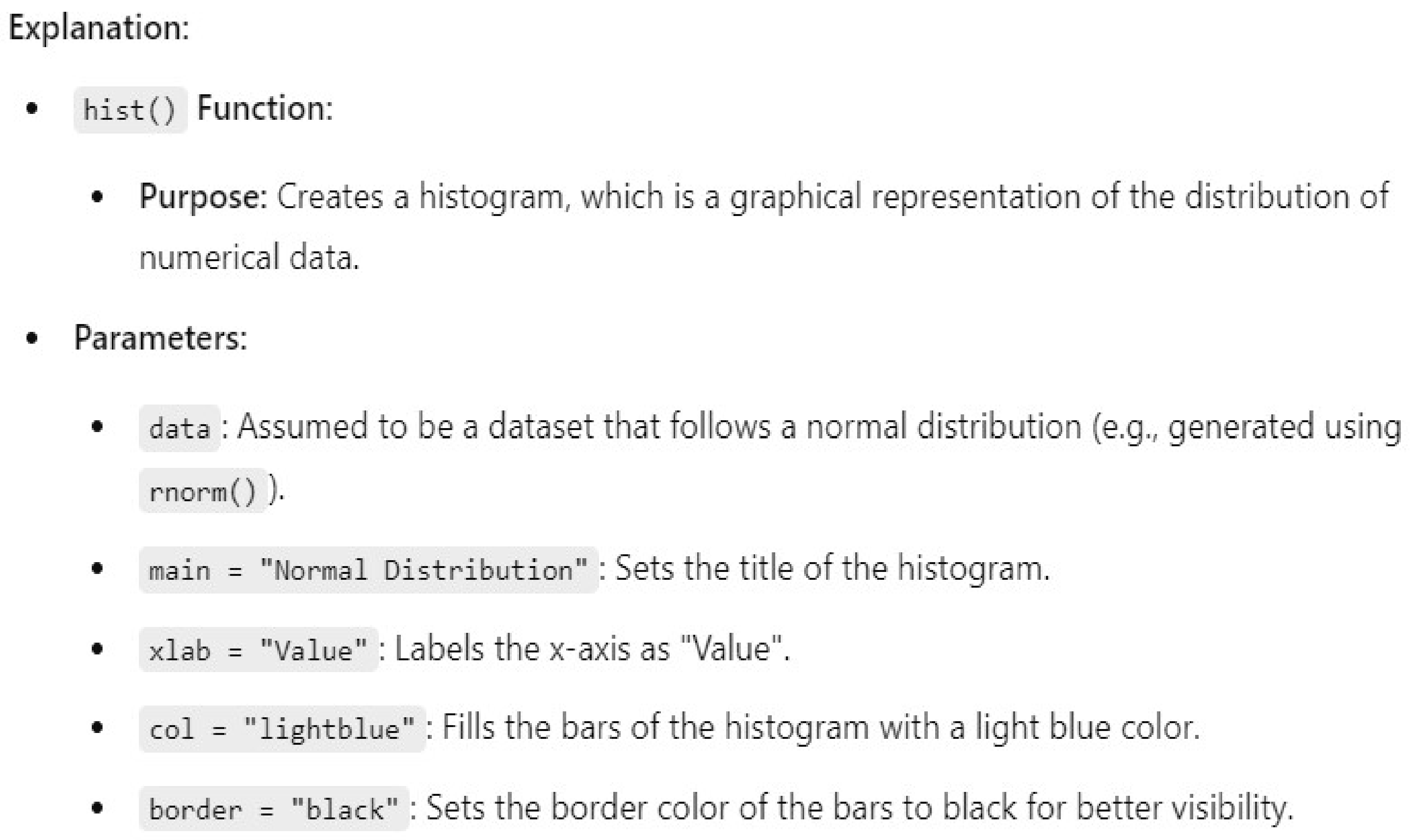


Explanation:

- `par()` **Function:**
  - **Purpose:** Sets or queries graphical parameters in R.
- `mfrow = c(2,1)`:
  - `mfrow` stands for "multi-figure row-wise".
  - `c(2,1)`: Specifies a layout with **2 rows** and **1 column**.
  - **Effect:** The plotting area is divided into two sections stacked vertically. The first plot will appear in the top section, and the second plot will appear below it.

Visual Insight:

- This setup allows you to compare two histograms directly, making it easier to observe differences in their shapes and distributions.



Explanation:

- `hist()` Function:
  - **Purpose:** Creates a histogram, which is a graphical representation of the distribution of numerical data.
- Parameters:
  - `data` : Assumed to be a dataset that follows a normal distribution (e.g., generated using `rnorm()`).
  - `main = "Normal Distribution"` : Sets the title of the histogram.
  - `xlab = "Value"` : Labels the x-axis as "Value".
  - `col = "lightblue"` : Fills the bars of the histogram with a light blue color.
  - `border = "black"` : Sets the border color of the bars to black for better visibility.

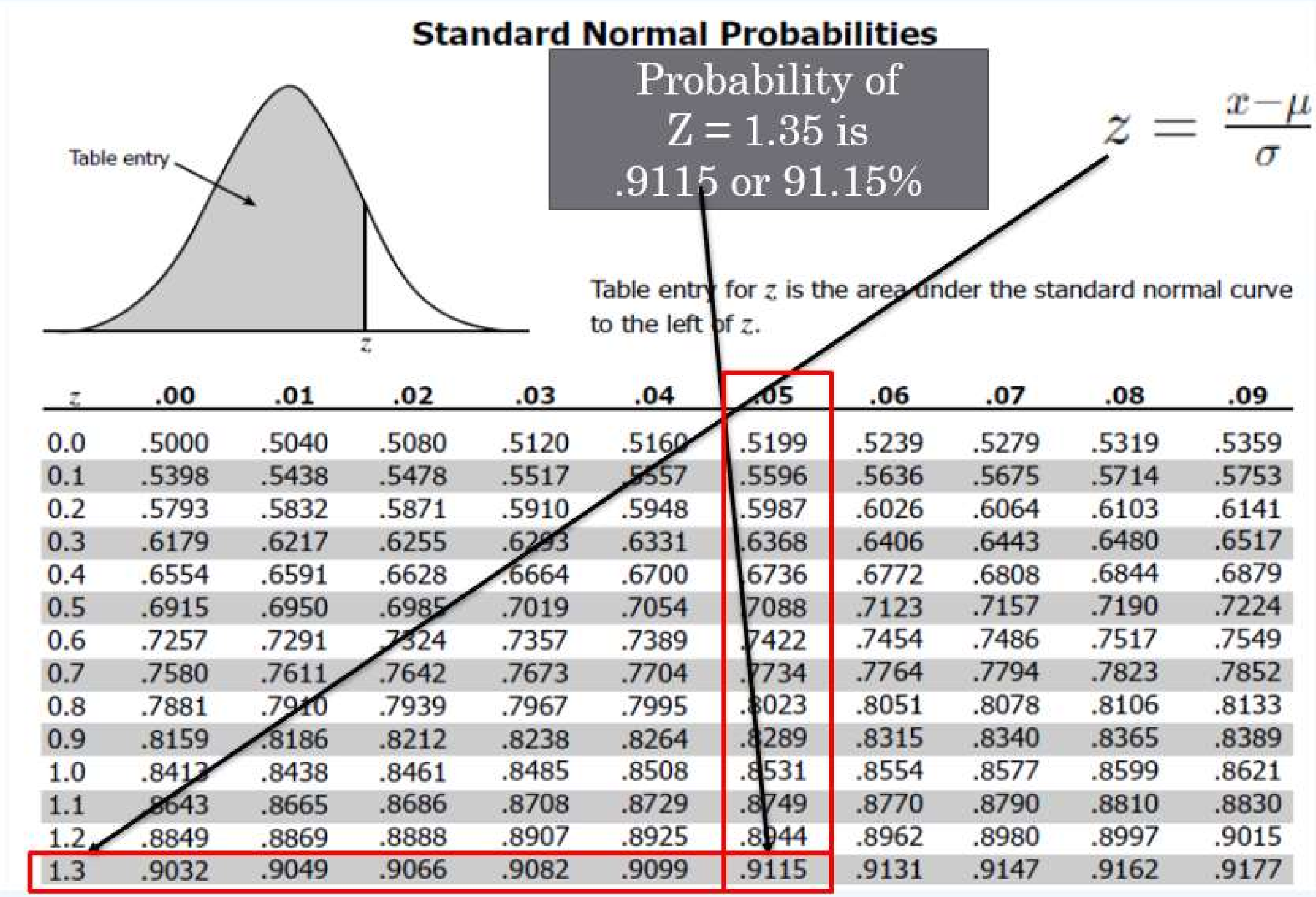


Table entry
z

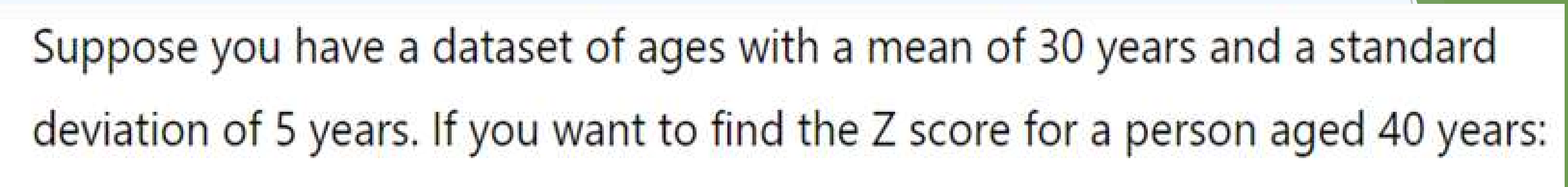
Standard Normal Probabilities

Probability of
Z = 1.35 is
.9115 or 91.15%

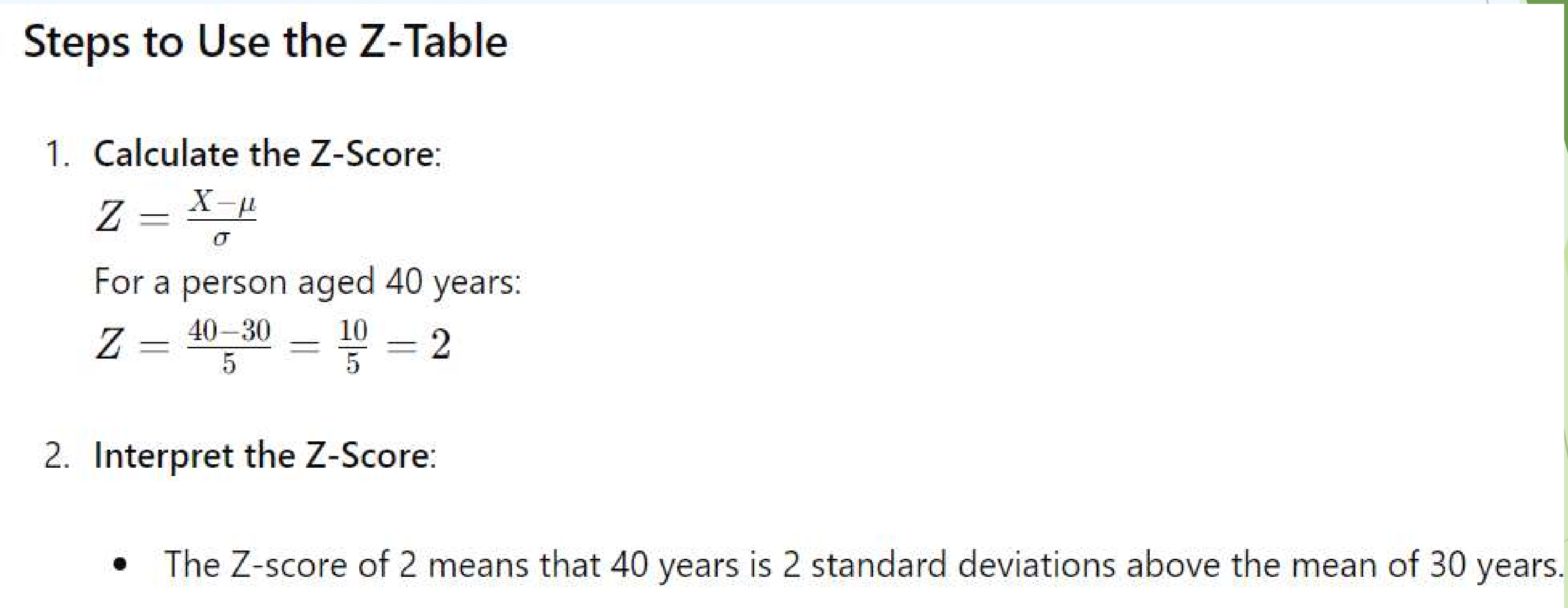
$$z = \frac{x-\mu}{\sigma}$$

Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177



Suppose you have a dataset of ages with a mean of 30 years and a standard deviation of 5 years. If you want to find the Z score for a person aged 40 years:



Steps to Use the Z-Table

1. **Calculate the Z-Score**:

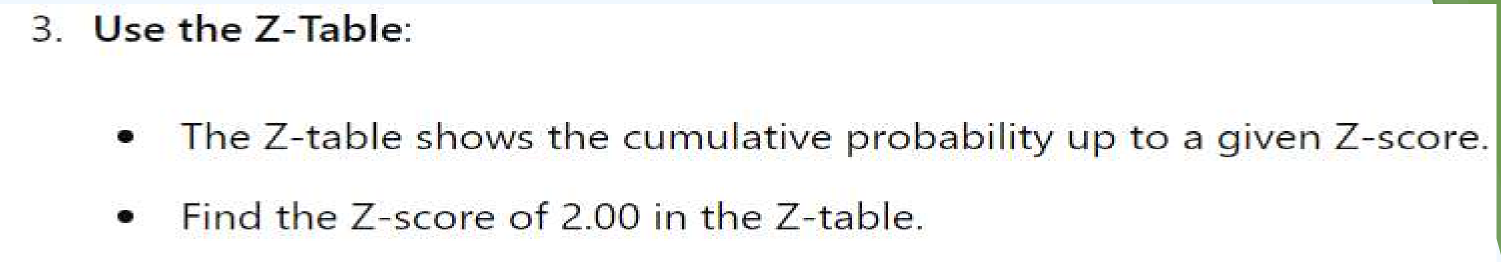
   $Z = \frac{X - \mu}{\sigma}$

   For a person aged 40 years:

   $Z = \frac{40 - 30}{5} = \frac{10}{5} = 2$

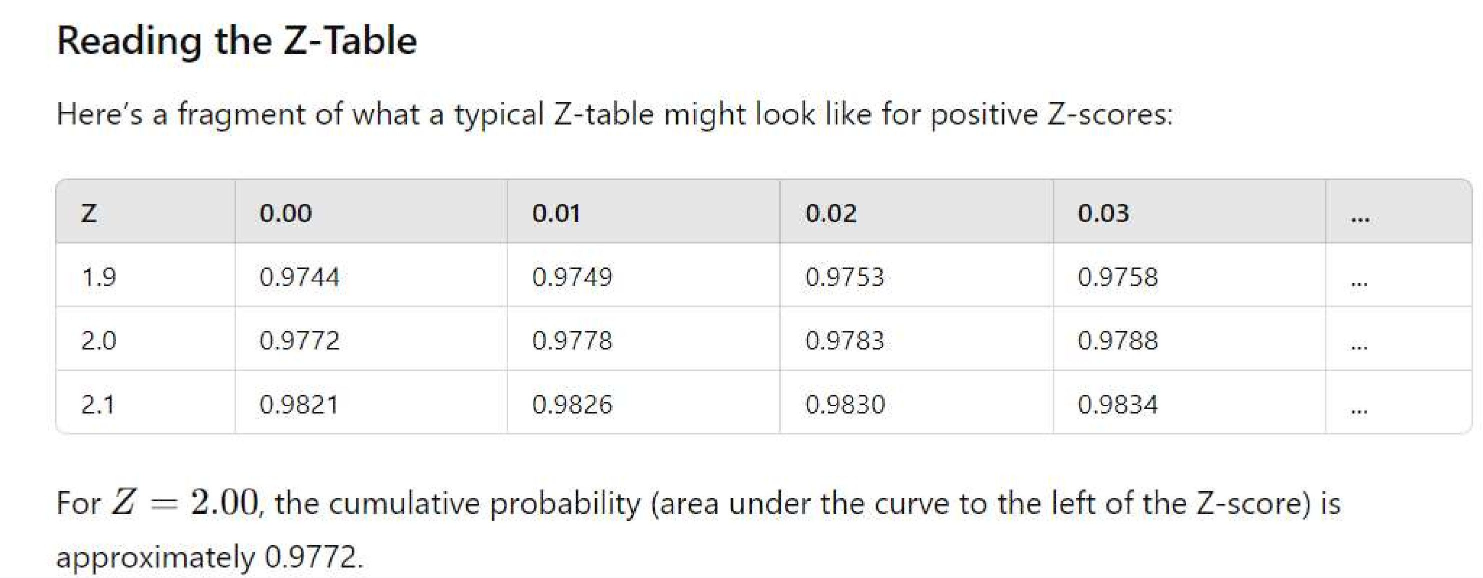
2. **Interpret the Z-Score**:

   - The Z-score of 2 means that 40 years is 2 standard deviations above the mean of 30 years.



3. **Use the Z-Table**:

- The Z-table shows the cumulative probability up to a given Z-score.
- Find the Z-score of 2.00 in the Z-table.

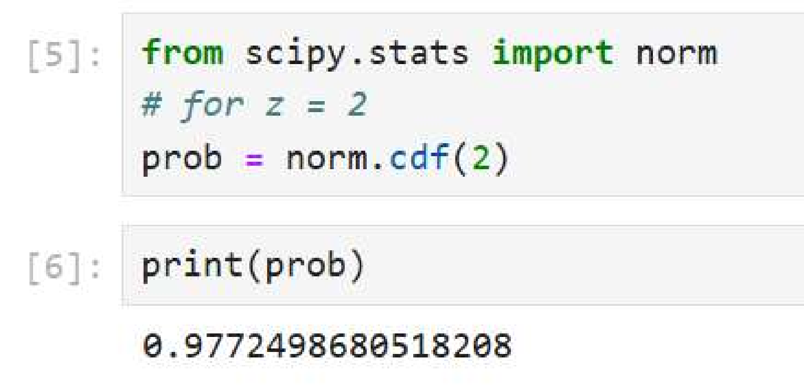


Reading the Z-Table

Here's a fragment of what a typical Z-table might look like for positive Z-scores:

Z	0.00	0.01	0.02	0.03	...
1.9	0.9744	0.9749	0.9753	0.9758	...
2.0	0.9772	0.9778	0.9783	0.9788	...
2.1	0.9821	0.9826	0.9830	0.9834	...

For $Z = 2.00$, the cumulative probability (area under the curve to the left of the Z-score) is approximately 0.9772.



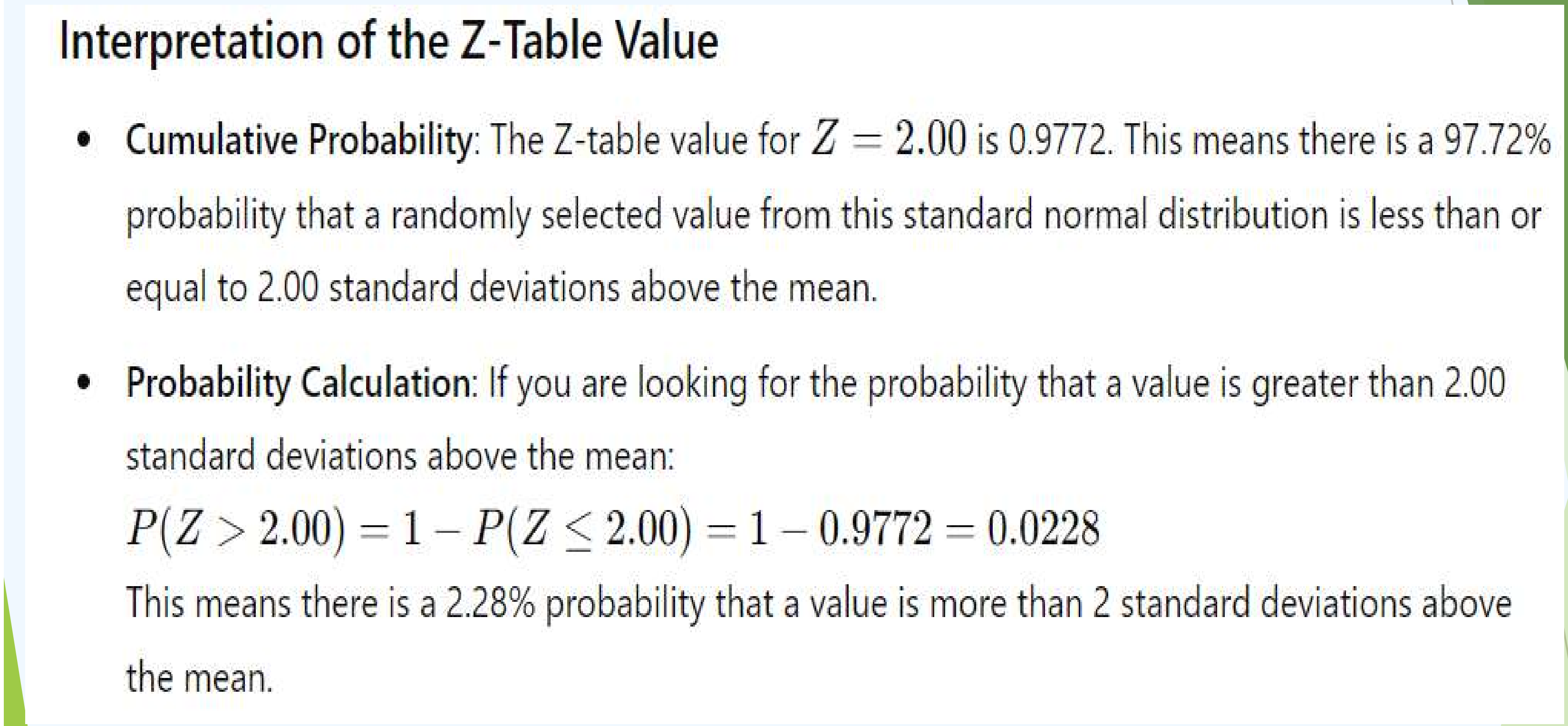
[5]:

```
from scipy.stats import norm
# for z = 2
prob = norm.cdf(2)
```

[6]:

```
print(prob)
```

```
0.9772498680518208
```



Interpretation of the Z-Table Value

- **Cumulative Probability**: The Z-table value for $Z = 2.00$ is 0.9772. This means there is a 97.72% probability that a randomly selected value from this standard normal distribution is less than or equal to 2.00 standard deviations above the mean.
- **Probability Calculation**: If you are looking for the probability that a value is greater than 2.00 standard deviations above the mean:

  $$P(Z > 2.00) = 1 - P(Z \leq 2.00) = 1 - 0.9772 = 0.0228$$

  This means there is a 2.28% probability that a value is more than 2 standard deviations above the mean.

Example: The test scores of students in a class test has a mean of 70 and with a standard deviation of 12. What is the probable percentage of students scored more than 85?

Solution: The z score for the given data is,

$$Z = (85-70)/12 = 1.25$$

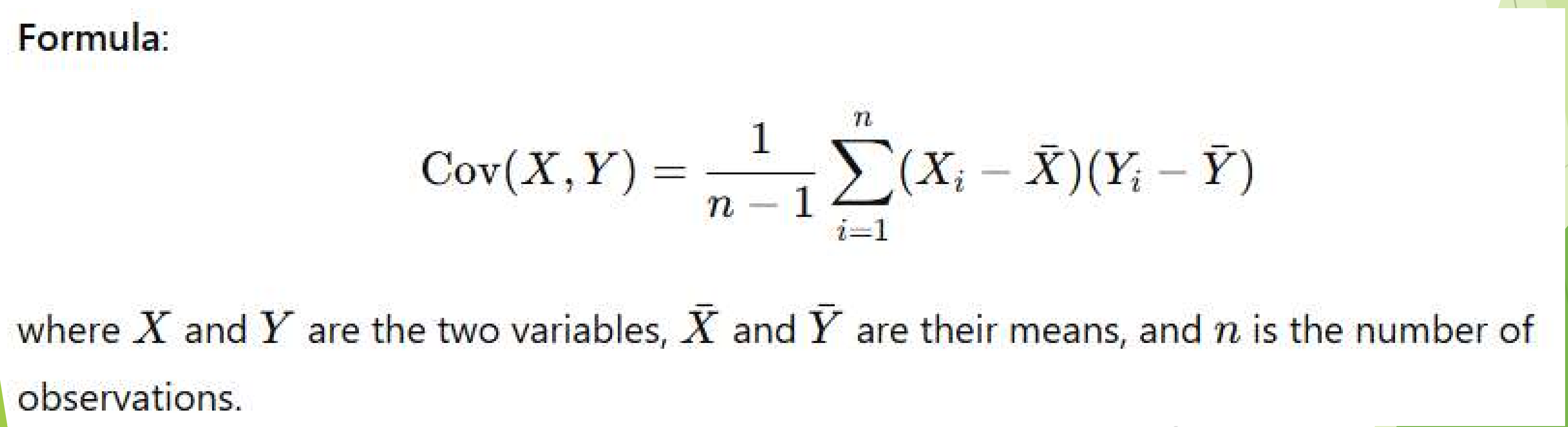
From the z score table, the fraction of the data within this score is 0.8944.

- **This means 89.44 % of the students are within the test scores of 85 and hence the percentage of students who are above the test scores of 85 = (100-89.44)% = 10.56 %.**

Covariance

- **Definition:**
  - Covariance measures the degree to which two variables change together or we can say it talks about direction.

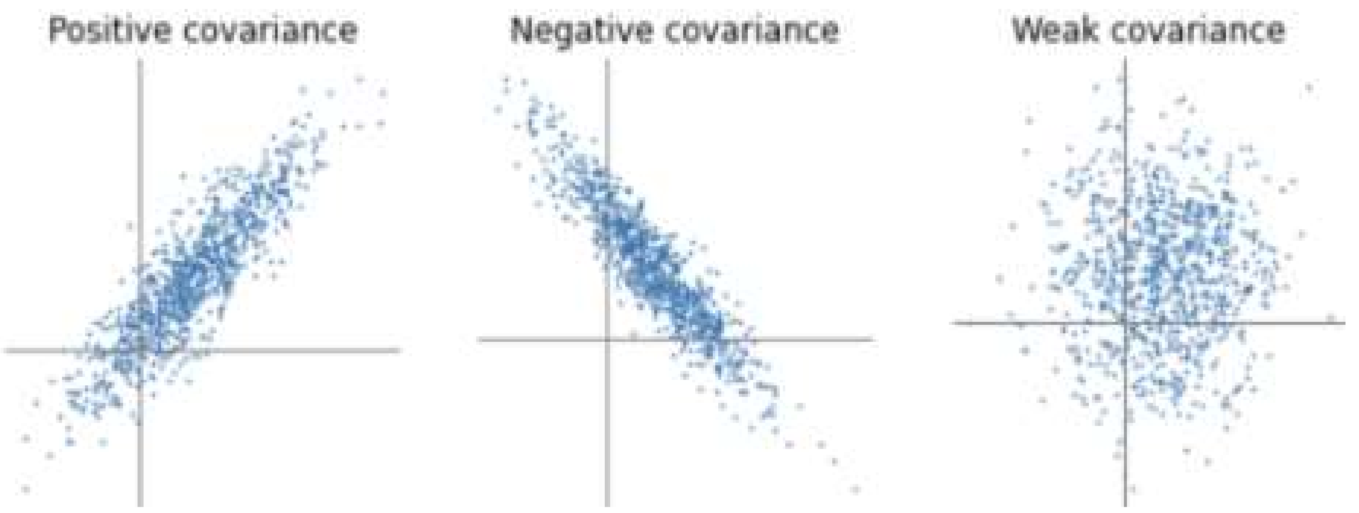
- If both variables tend to increase together, the covariance is positive;
- if one increases while the other decreases, the covariance is negative.

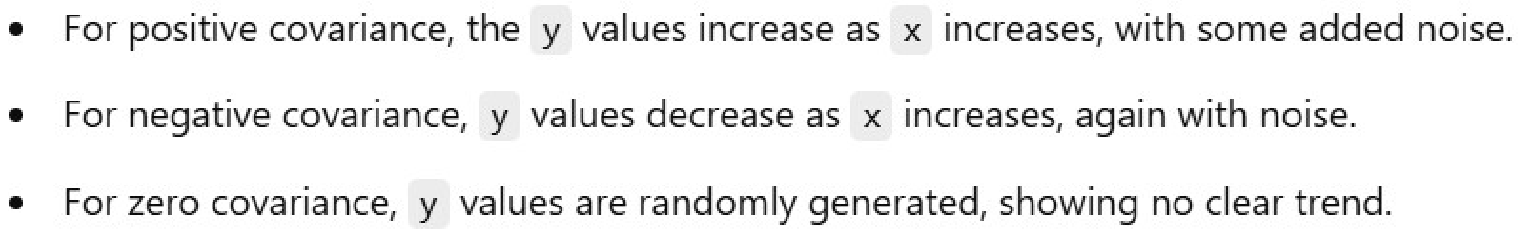


Formula:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

where X and Y are the two variables, $\bar{X}$ and $\bar{Y}$ are their means, and n is the number of observations.

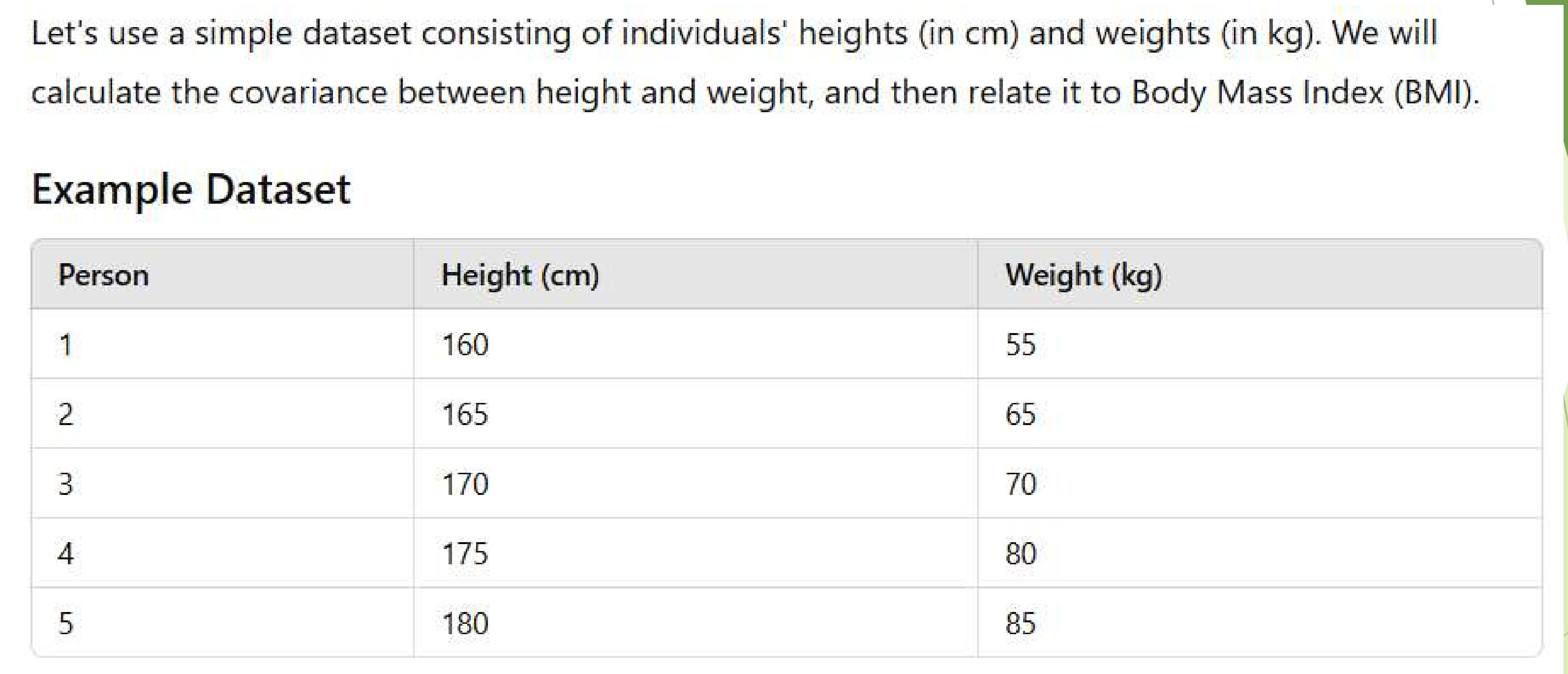




- For positive covariance, the `y` values increase as `x` increases, with some added noise.
- For negative covariance, `y` values decrease as `x` increases, again with noise.
- For zero covariance, `y` values are randomly generated, showing no clear trend.

- Covariance can take any value between −∞ and +∞.
- It’s not standardized, so its magnitude is difficult to interpret.

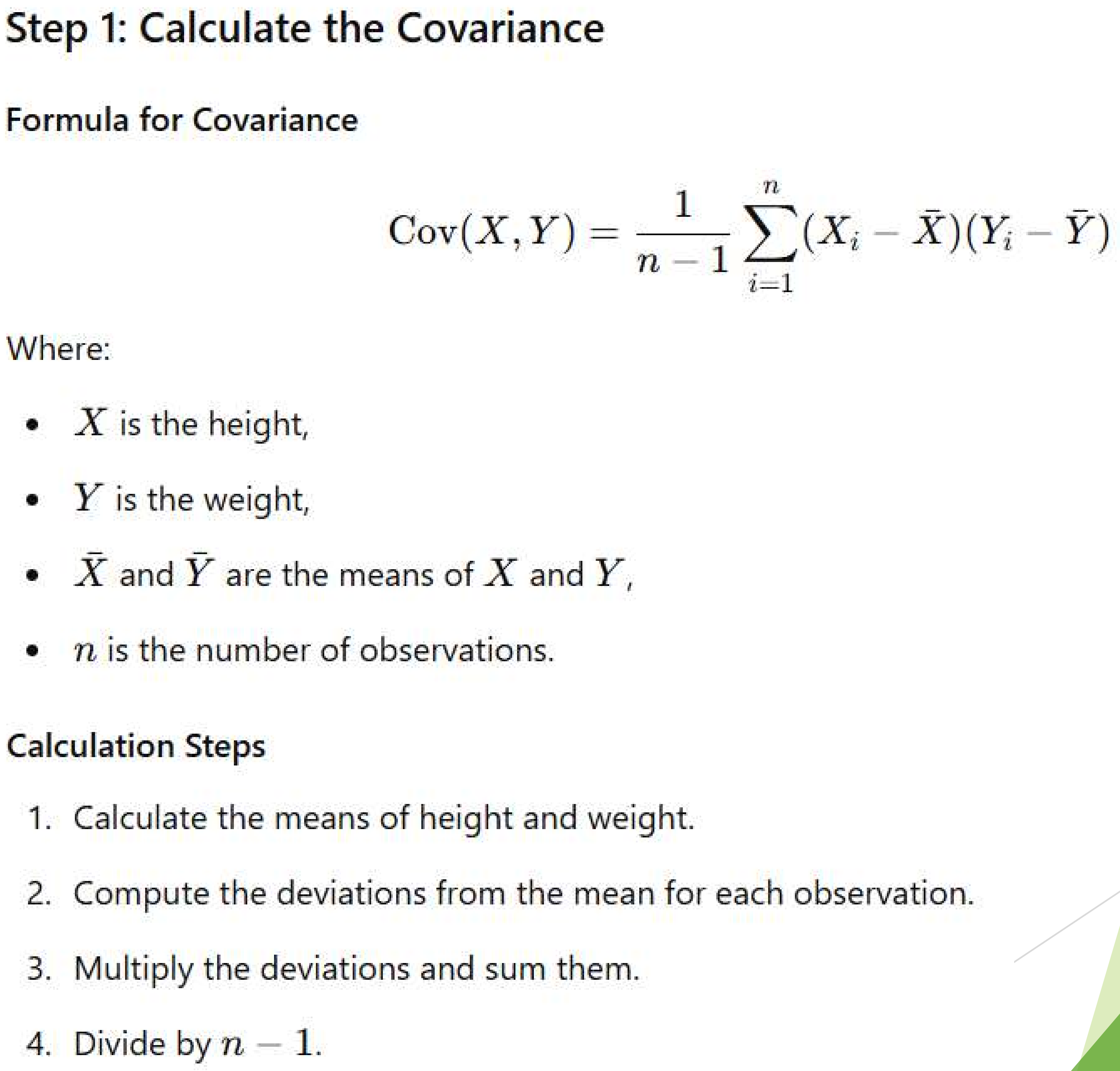
Practical Example: Calculating Covariance with a Simple Dataset



Let's use a simple dataset consisting of individuals' heights (in cm) and weights (in kg). We will calculate the covariance between height and weight, and then relate it to Body Mass Index (BMI).

Example Dataset

Person	Height (cm)	Weight (kg)
1	160	55
2	165	65
3	170	70
4	175	80
5	180	85



Step 1: Calculate the Covariance

Formula for Covariance

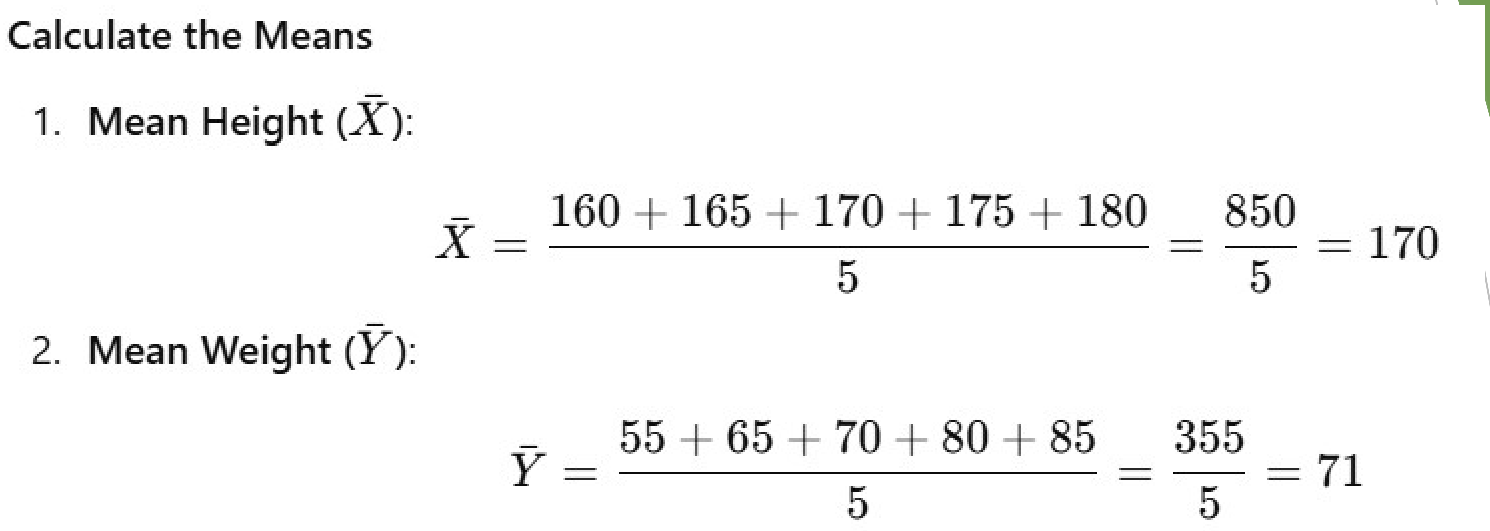
$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Where:

- X is the height,
- Y is the weight,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y,
- n is the number of observations.

Calculation Steps

1. Calculate the means of height and weight.
2. Compute the deviations from the mean for each observation.
3. Multiply the deviations and sum them.
4. Divide by $n - 1$.



Calculate the Means

1. **Mean Height ($\bar{X}$):**

$$\bar{X} = \frac{160 + 165 + 170 + 175 + 180}{5} = \frac{850}{5} = 170$$

2. **Mean Weight ($\bar{Y}$):**

$$\bar{Y} = \frac{55 + 65 + 70 + 80 + 85}{5} = \frac{355}{5} = 71$$

Calculate Deviations from the Mean

- Deviations for Height:
  - $160 - 170 = -10$
  - $165 - 170 = -5$
  - $170 - 170 = 0$
  - $175 - 170 = 5$
  - $180 - 170 = 10$

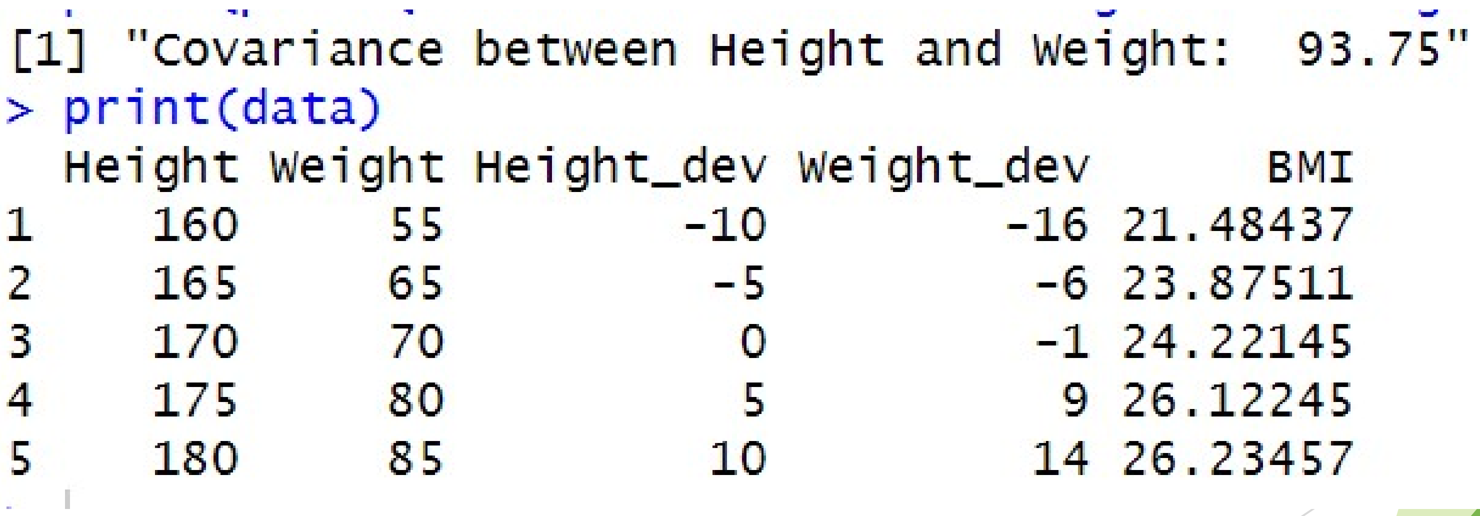
- Deviations for Weight:
  - $55 - 71 = -16$
  - $65 - 71 = -6$
  - $70 - 71 = -1$
  - $80 - 71 = 9$
  - $85 - 71 = 14$

Calculate Covariance

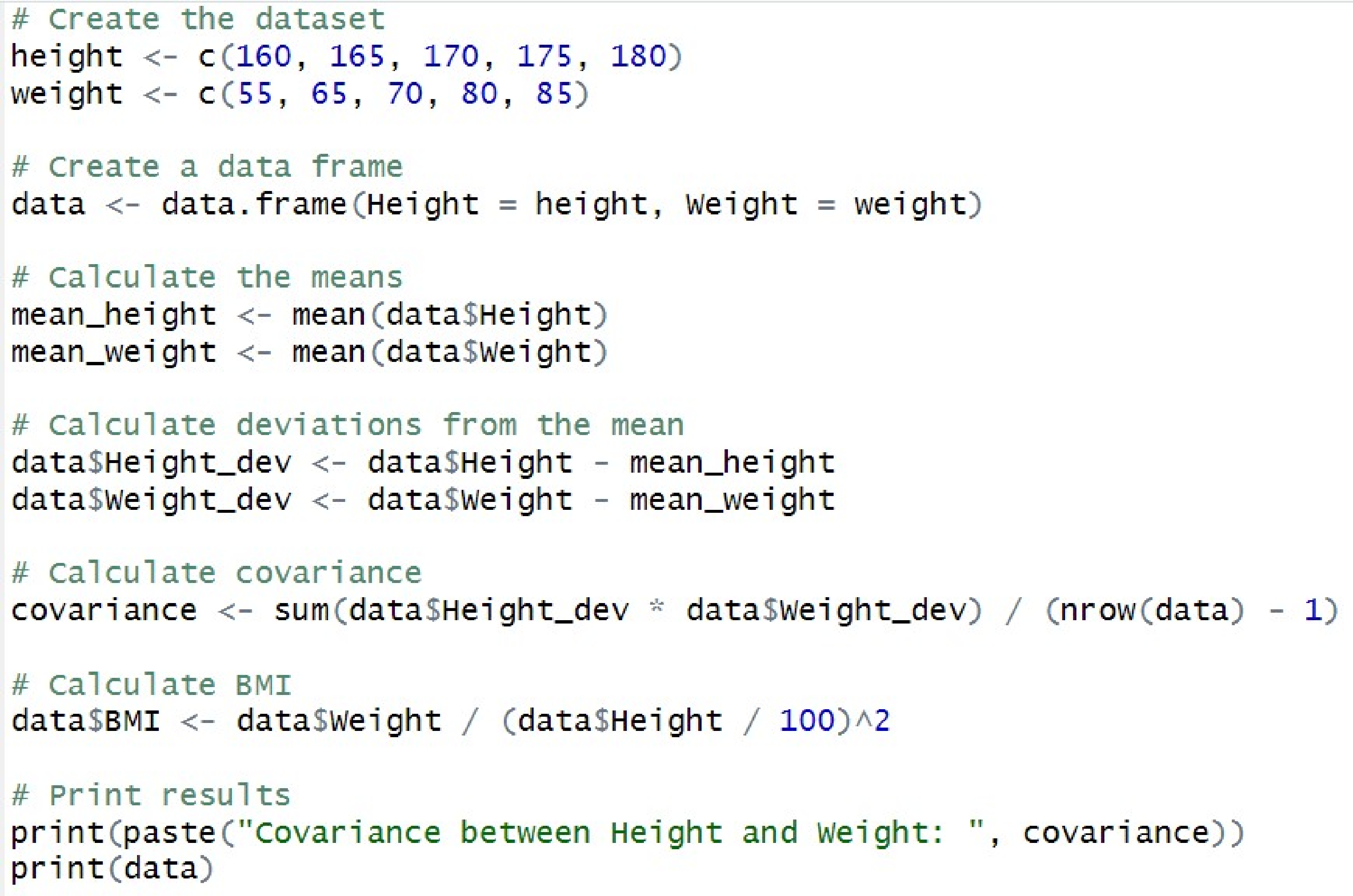
$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\mathrm{Cov}(X,Y) = \frac{1}{4}[(-10)(-16) + (-5)(-6) + (0)(-1) + (5)(9) + (10)(14)]$$

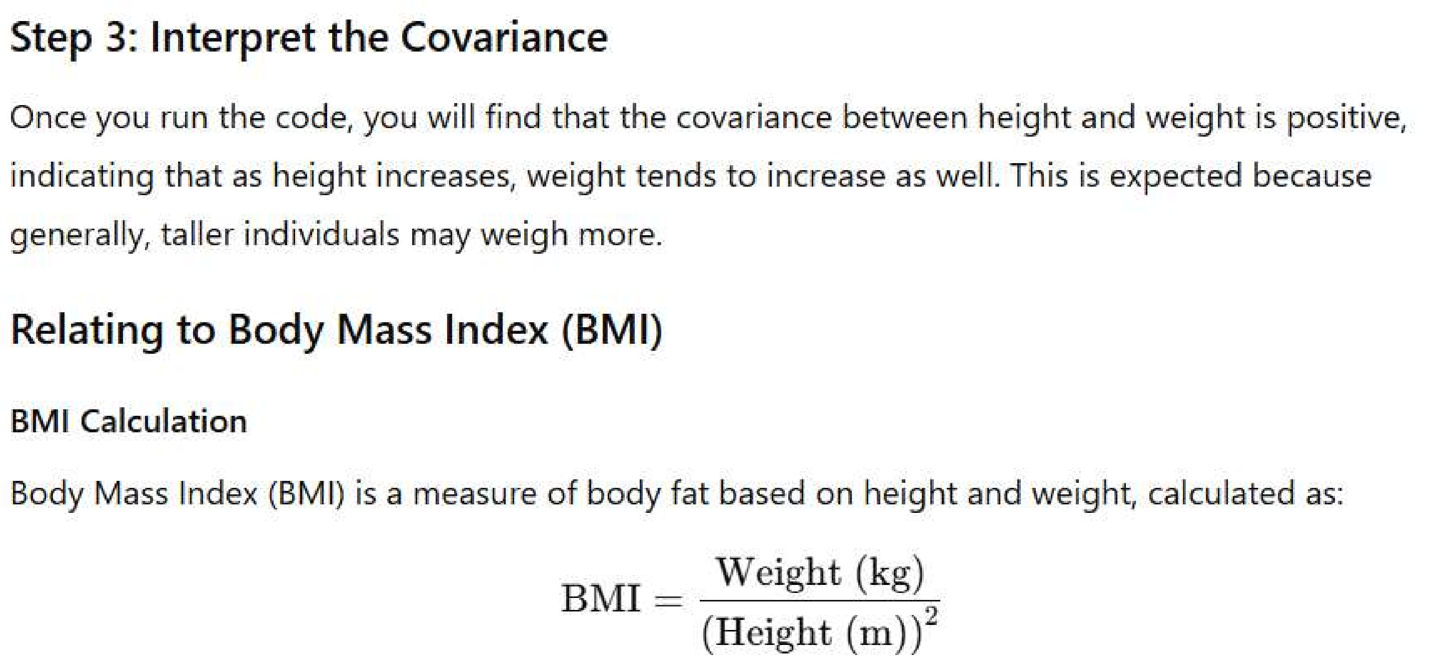
$$= \frac{1}{4}[160 + 30 + 0 + 45 + 140] = \frac{375}{4} = 93.75$$



```
[1] "Covariance between Height and Weight:  93.75"
> print(data)
  Height Weight Height_dev Weight_dev      BMI
1    160     55        -10        -16 21.48437
2    165     65         -5         -6 23.87511
3    170     70          0         -1 24.22145
4    175     80          5          9 26.12245
5    180     85         10         14 26.23457
```



```r
# Create the dataset
height <- c(160, 165, 170, 175, 180)
weight <- c(55, 65, 70, 80, 85)

# Create a data frame
data <- data.frame(Height = height, Weight = weight)

# Calculate the means
mean_height <- mean(data$Height)
mean_weight <- mean(data$Weight)

# Calculate deviations from the mean
data$Height_dev <- data$Height - mean_height
data$Weight_dev <- data$Weight - mean_weight

# Calculate covariance
covariance <- sum(data$Height_dev * data$Weight_dev) / (nrow(data) - 1)

# Calculate BMI
data$BMI <- data$Weight / (data$Height / 100)^2

# Print results
print(paste("Covariance between Height and Weight: ", covariance))
print(data)
```



Step 3: Interpret the Covariance

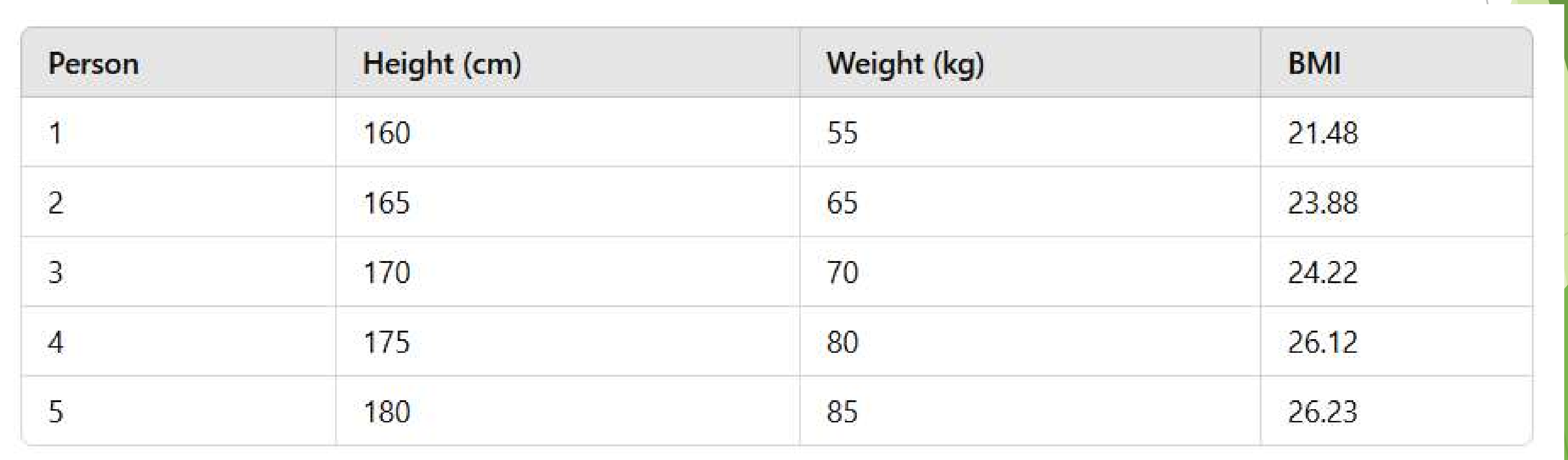
Once you run the code, you will find that the covariance between height and weight is positive, indicating that as height increases, weight tends to increase as well. This is expected because generally, taller individuals may weigh more.

Relating to Body Mass Index (BMI)

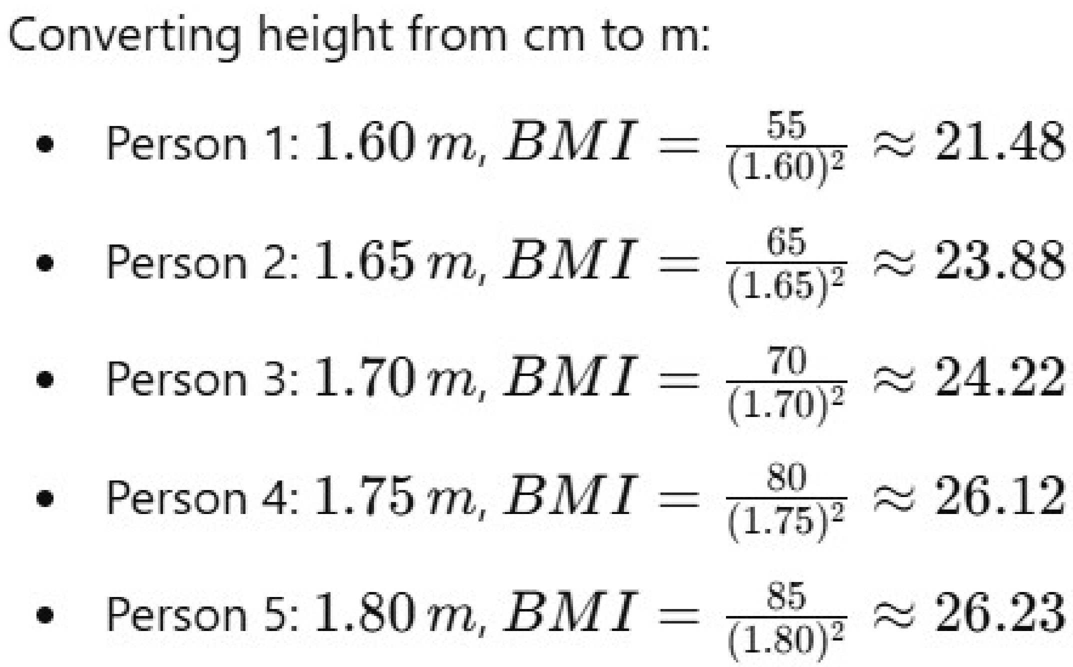
BMI Calculation

Body Mass Index (BMI) is a measure of body fat based on height and weight, calculated as:

$$\text{BMI} = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$



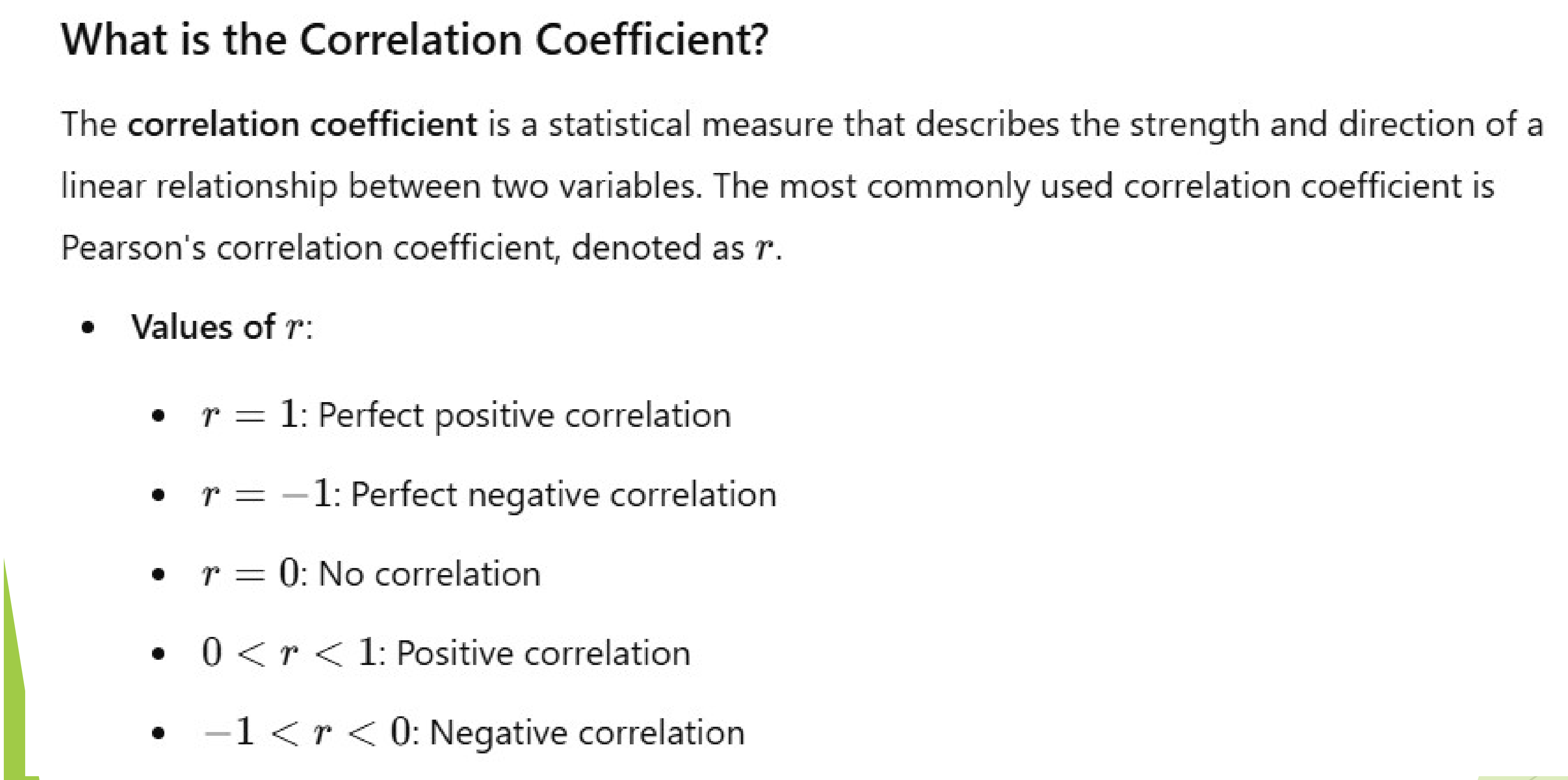
Person	Height (cm)	Weight (kg)	BMI
1	160	55	21.48
2	165	65	23.88
3	170	70	24.22
4	175	80	26.12
5	180	85	26.23



Converting height from cm to m:

- Person 1: $1.60\,m$, $BMI = \frac{55}{(1.60)^2} \approx 21.48$
- Person 2: $1.65\,m$, $BMI = \frac{65}{(1.65)^2} \approx 23.88$
- Person 3: $1.70\,m$, $BMI = \frac{70}{(1.70)^2} \approx 24.22$
- Person 4: $1.75\,m$, $BMI = \frac{80}{(1.75)^2} \approx 26.12$
- Person 5: $1.80\,m$, $BMI = \frac{85}{(1.80)^2} \approx 26.23$

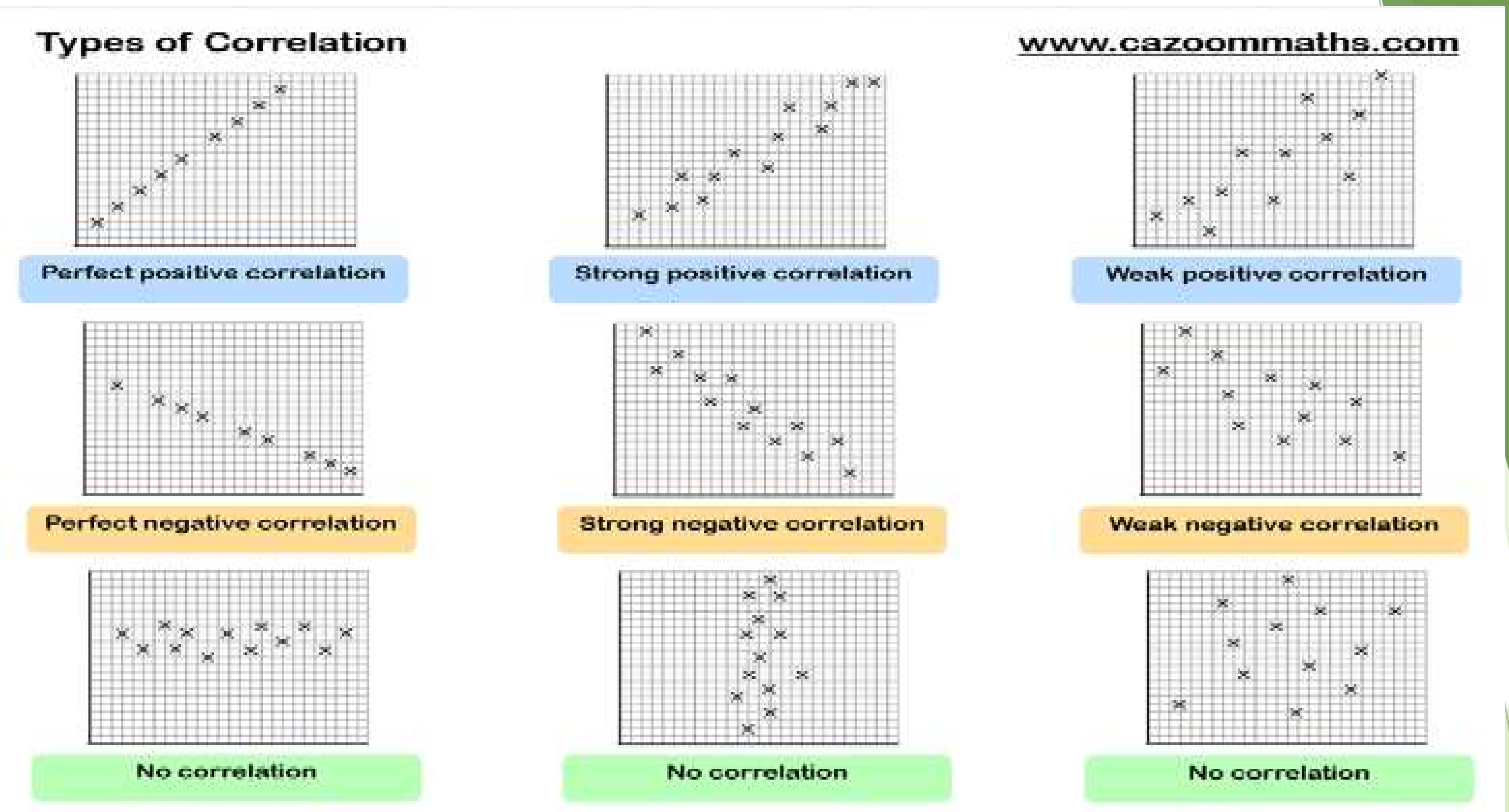
Correlation



What is the Correlation Coefficient?

The **correlation coefficient** is a statistical measure that describes the strength and direction of a linear relationship between two variables. The most commonly used correlation coefficient is Pearson's correlation coefficient, denoted as r.

- **Values of r:**
  - $r = 1$: Perfect positive correlation
  - $r = -1$: Perfect negative correlation
  - $r = 0$: No correlation
  - $0 < r < 1$: Positive correlation
  - $-1 < r < 0$: Negative correlation



Types of Correlation

www.cazoommaths.com

Perfect positive correlation

Strong positive correlation

Weak positive correlation

Perfect negative correlation

Strong negative correlation

Weak negative correlation

No correlation

No correlation

No correlation

Pearson's Correlation Coefficient:

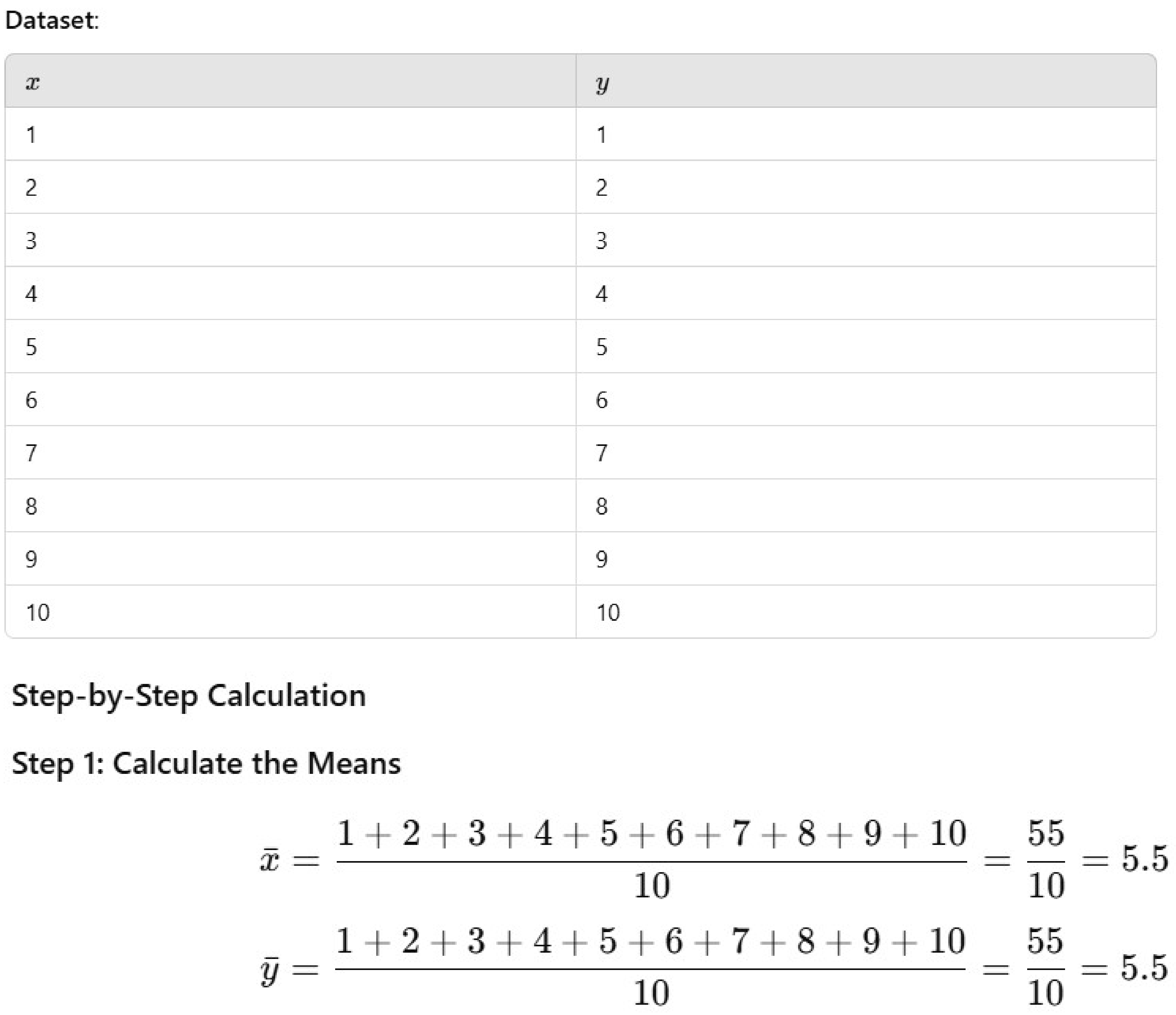
$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y.

Interpretation: Discuss the interpretation of the correlation coefficient:

- **Positive Correlation**: $0 < r \leq 1$ (e.g., as one variable increases, the other also increases)
- **Negative Correlation**: $-1 \leq r < 0$ (e.g., as one variable increases, the other decreases)
- **No Correlation**: $r = 0$ (no linear relationship)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$



Dataset:

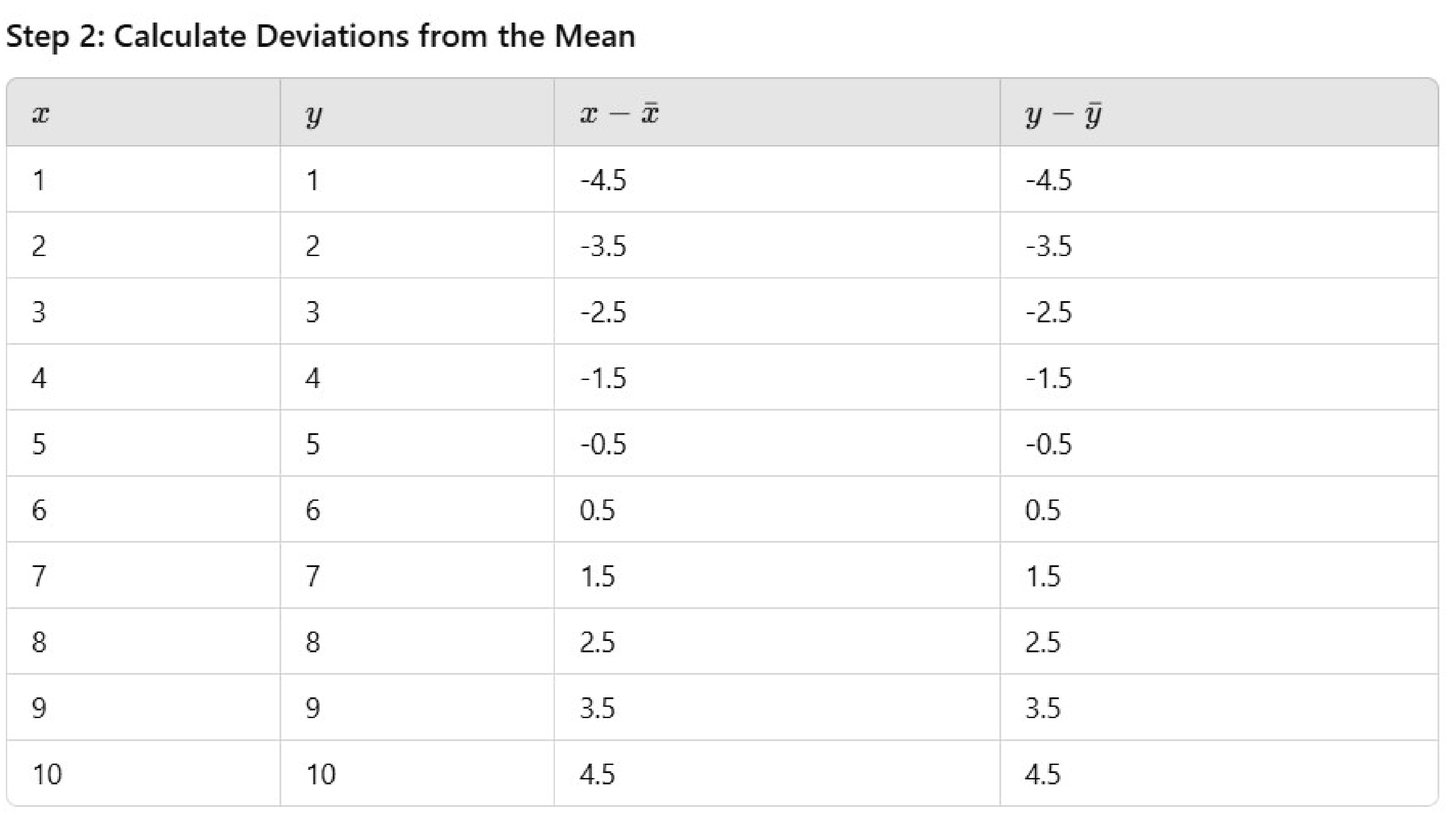
x	y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

Step-by-Step Calculation

Step 1: Calculate the Means

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

$$\bar{y} = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

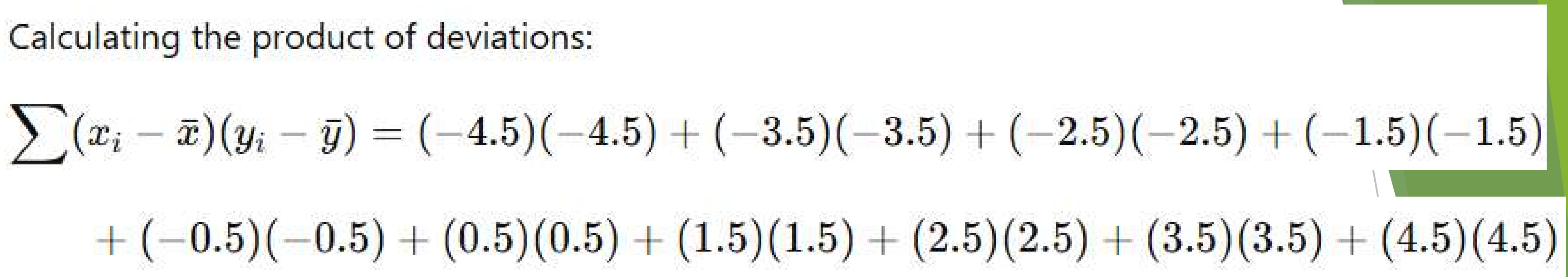


Step 2: Calculate Deviations from the Mean

x	y	$x - \bar{x}$	$y - \bar{y}$
1	1	-4.5	-4.5
2	2	-3.5	-3.5
3	3	-2.5	-2.5
4	4	-1.5	-1.5
5	5	-0.5	-0.5
6	6	0.5	0.5
7	7	1.5	1.5
8	8	2.5	2.5
9	9	3.5	3.5
10	10	4.5	4.5

Step 3: Calculate the Covariance

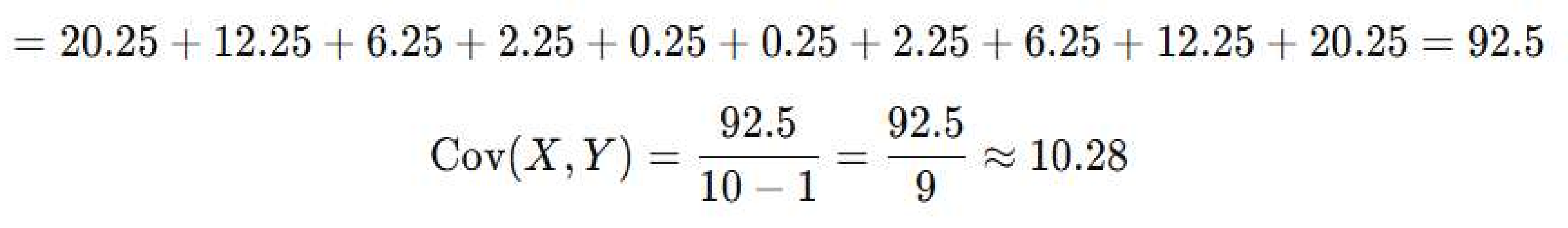
$$\mathrm{Cov}(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



Calculating the product of deviations:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (-4.5)(-4.5) + (-3.5)(-3.5) + (-2.5)(-2.5) + (-1.5)(-1.5)$$

$$+ (-0.5)(-0.5) + (0.5)(0.5) + (1.5)(1.5) + (2.5)(2.5) + (3.5)(3.5) + (4.5)(4.5)$$



$$= 20.25 + 12.25 + 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 + 12.25 + 20.25 = 92.5$$

$$\text{Cov}(X, Y) = \frac{92.5}{10-1} = \frac{92.5}{9} \approx 10.28$$

Step 4: Calculate the Standard Deviations

$$\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} =$$

$$\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} =$$

$$= \sqrt{\frac{20.25 + 12.25 + 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 + 12.25 + 20.25}{9}}$$

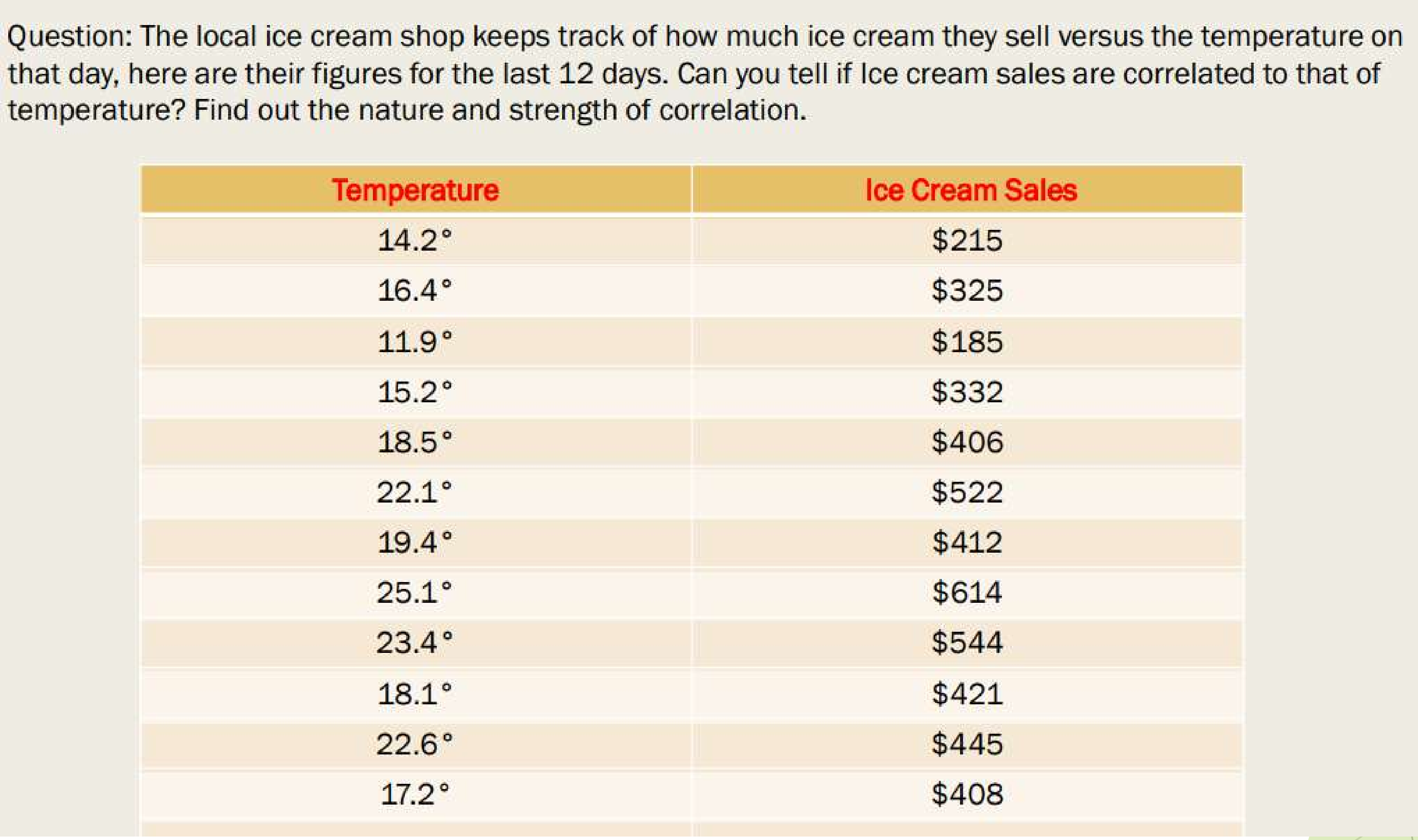
$$= \sqrt{\frac{20.25 + 12.25 + 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 + 12.25 + 20.25}{9}} =$$

$$= \sqrt{\frac{92.5}{9}} \approx 3.22$$

$$= \sqrt{\frac{92.5}{9}} \approx 3.22$$

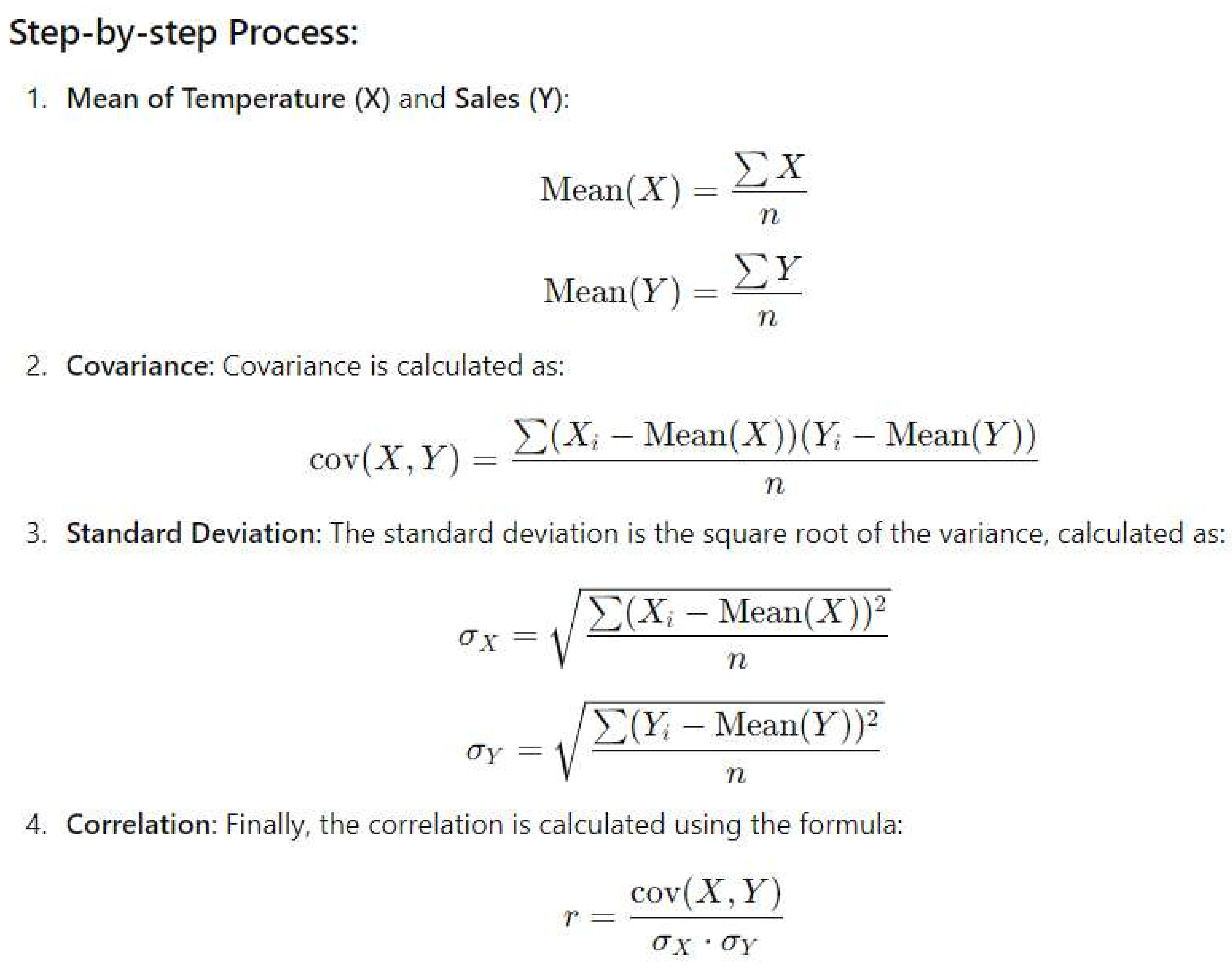
Step 5: Calculate the Correlation Coefficient

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} = \frac{10.28}{3.22 \times 3.22} = \frac{10.28}{10.38} \approx 1$$



Question: The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days. Can you tell if Ice cream sales are correlated to that of temperature? Find out the nature and strength of correlation.

Temperature	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Step-by-step Process:

1. **Mean of Temperature** (X) and **Sales** (Y):

$$\text{Mean}(X) = \frac{\sum X}{n}$$

$$\text{Mean}(Y) = \frac{\sum Y}{n}$$

2. **Covariance**: Covariance is calculated as:

$$\text{cov}(X, Y) = \frac{\sum(X_i - \text{Mean}(X))(Y_i - \text{Mean}(Y))}{n}$$

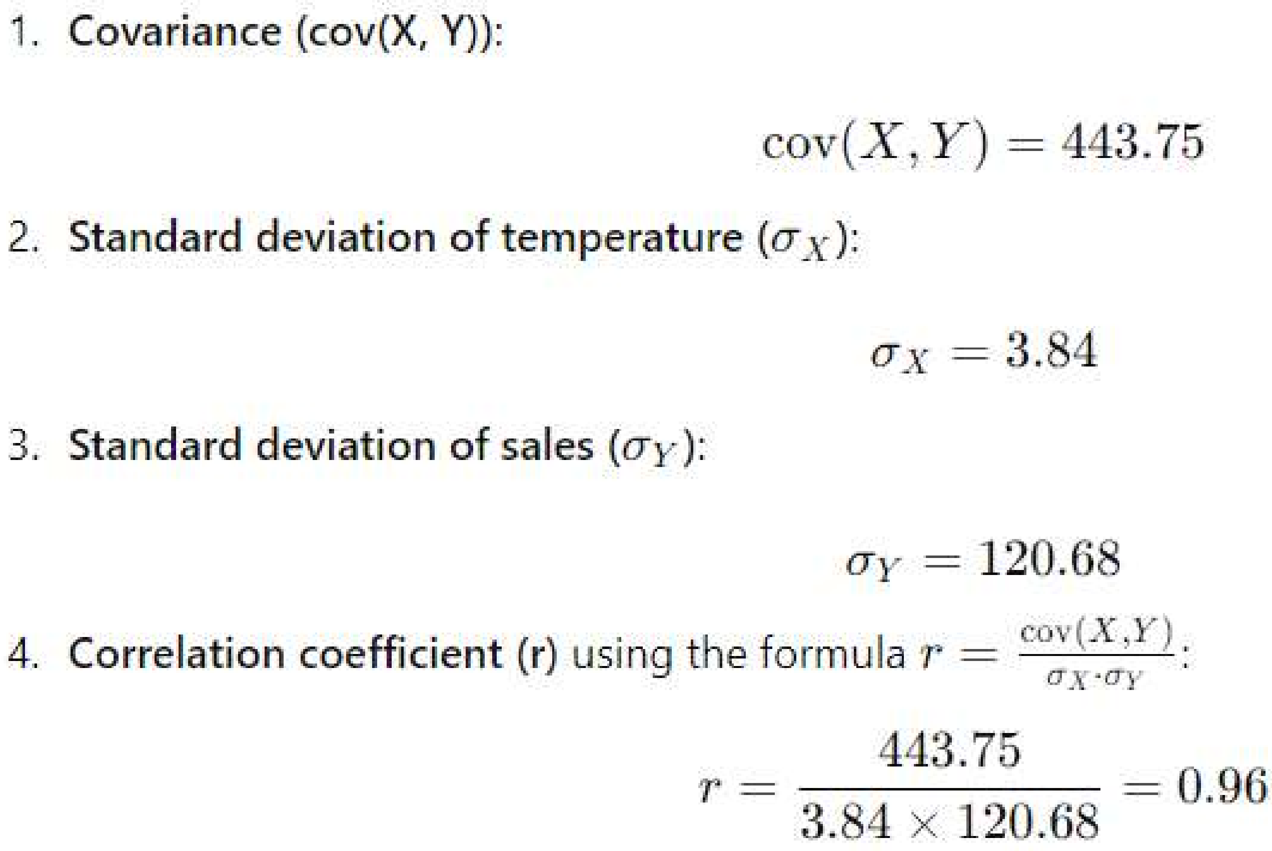
3. **Standard Deviation**: The standard deviation is the square root of the variance, calculated as:

$$\sigma_X = \sqrt{\frac{\sum(X_i - \text{Mean}(X))^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{\sum(Y_i - \text{Mean}(Y))^2}{n}}$$

4. **Correlation**: Finally, the correlation is calculated using the formula:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$



1. **Covariance (cov(X, Y)):**

$$\text{cov}(X, Y) = 443.75$$

2. **Standard deviation of temperature (σ_X):**

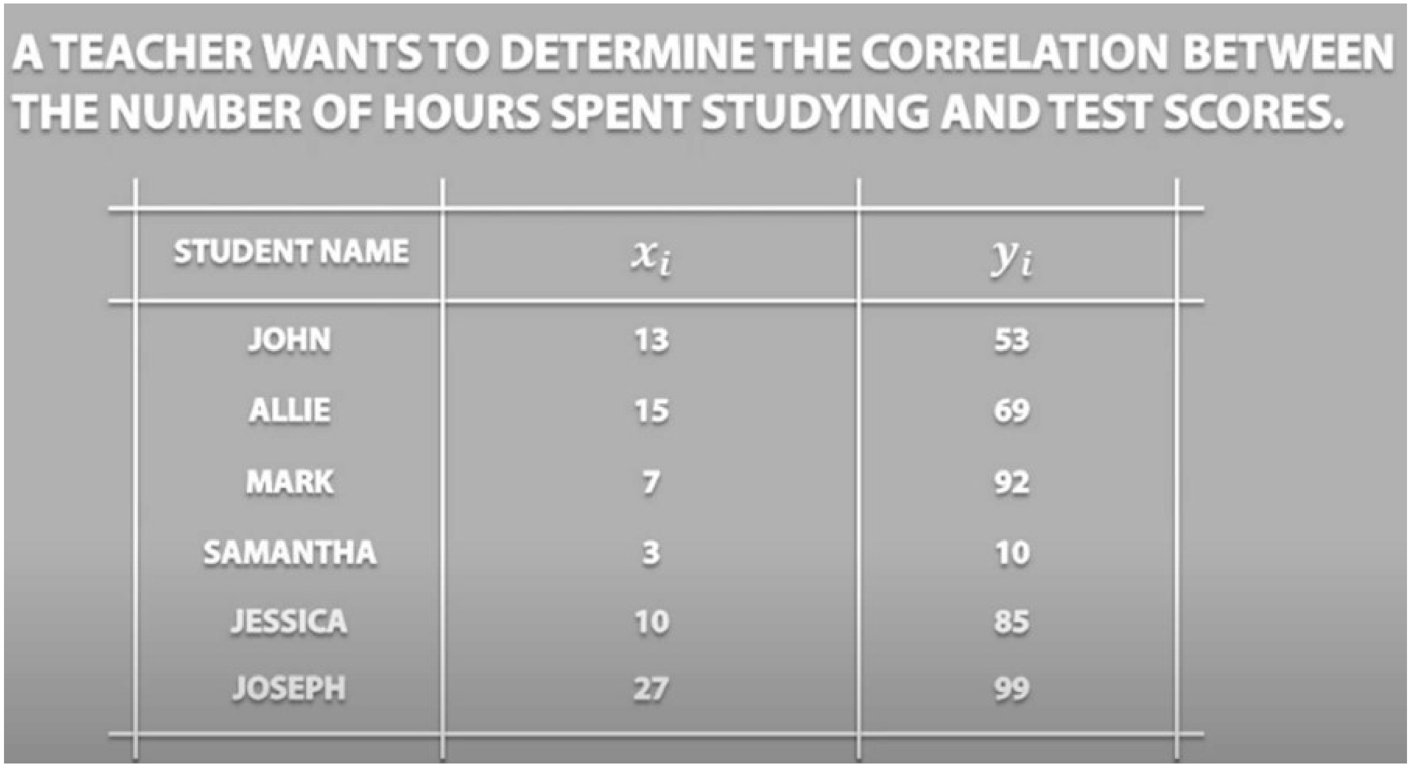
$$\sigma_X = 3.84$$

3. **Standard deviation of sales (σ_Y):**

$$\sigma_Y = 120.68$$

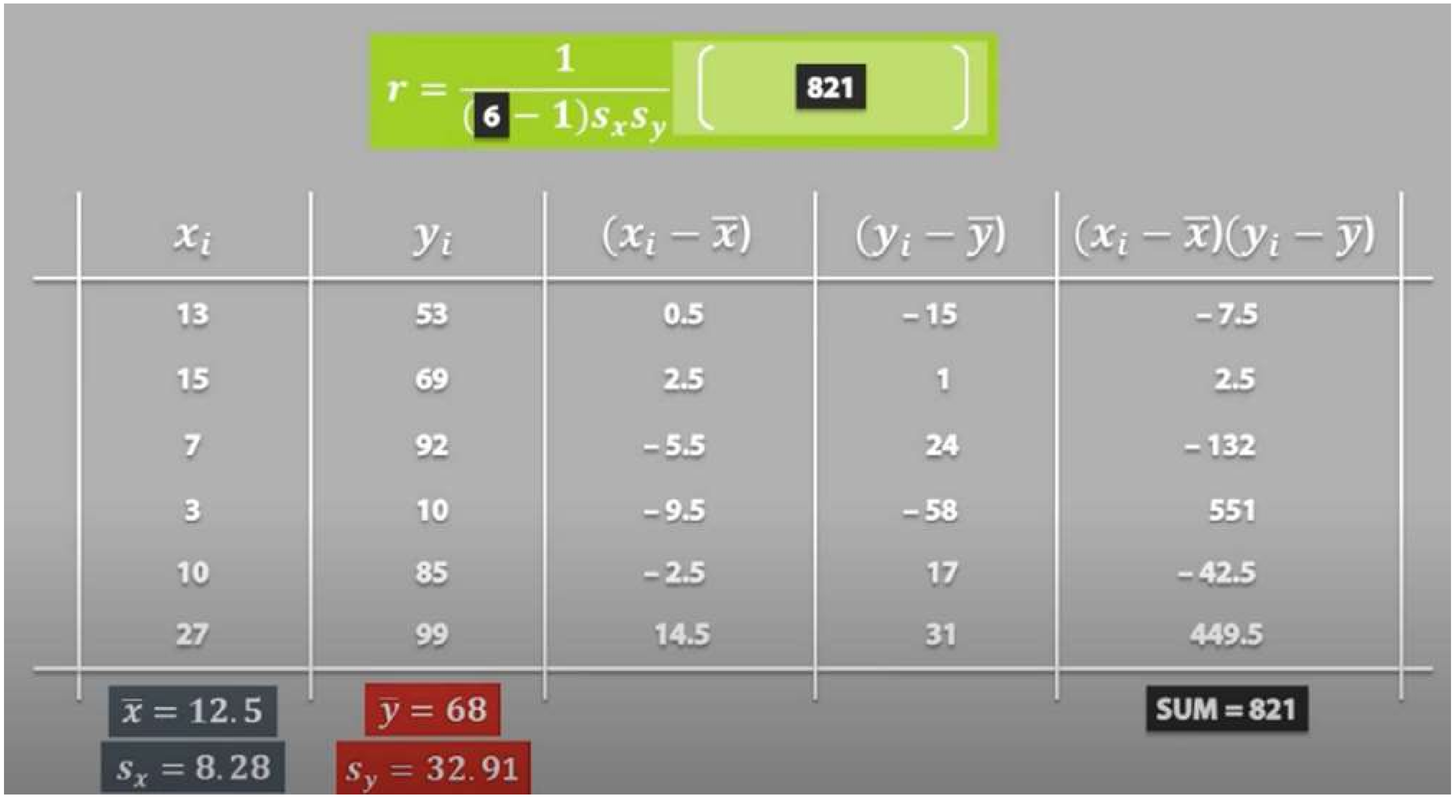
4. **Correlation coefficient (r)** using the formula $r = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$:

$$r = \frac{443.75}{3.84 \times 120.68} = 0.96$$



A TEACHER WANTS TO DETERMINE THE CORRELATION BETWEEN THE NUMBER OF HOURS SPENT STUDYING AND TEST SCORES.

STUDENT NAME	x_i	y_i
JOHN	13	53
ALLIE	15	69
MARK	7	92
SAMANTHA	3	10
JESSICA	10	85
JOSEPH	27	99

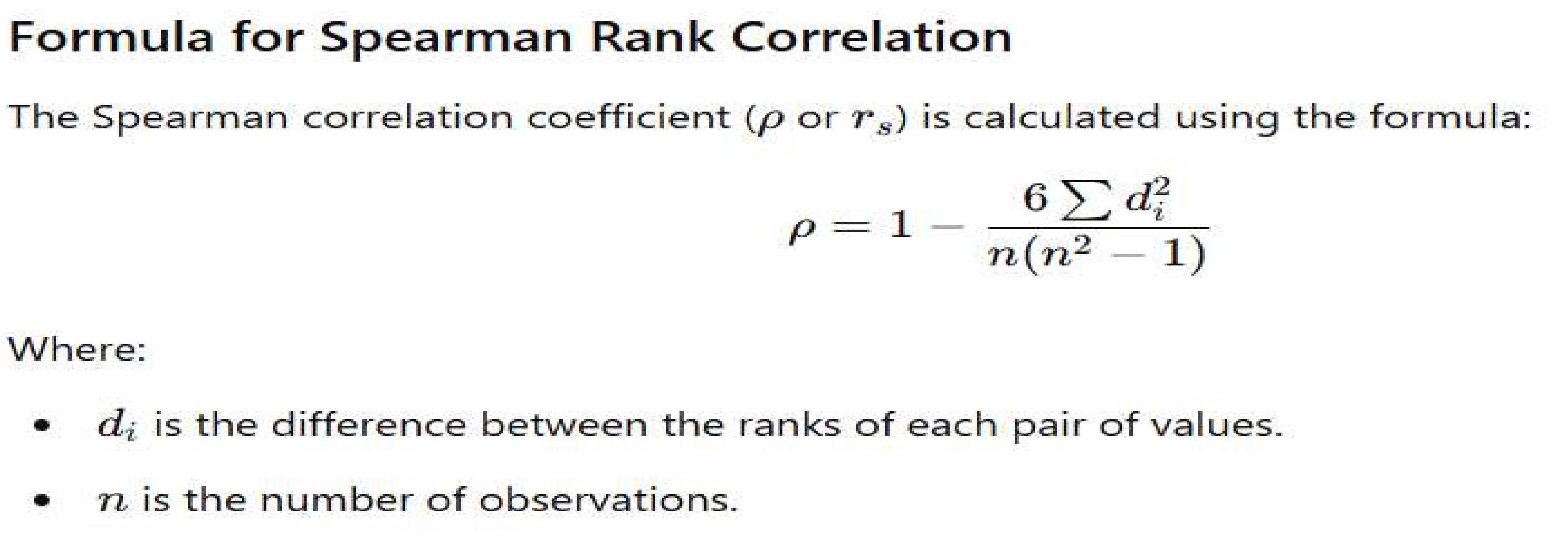


$$r = \frac{1}{(6 - 1)s_x s_y}\left(821 \right)$$

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
13	53	0.5	–15	–7.5
15	69	2.5	1	2.5
7	92	–5.5	24	–132
3	10	–9.5	–58	551
10	85	–2.5	17	–42.5
27	99	14.5	31	449.5
$\overline{x} = 12.5$	$\overline{y} = 68$			SUM = 821
$s_x = 8.28$	$s_y = 32.91$			

Spearman Rank Correlation:

- It is a non-parametric measure of correlation that assesses how well the relationship between two variables can be described using a monotonic function.
- Unlike Pearson’s correlation, which measures linear relationships, Spearman's correlation evaluates the strength and direction of the association between two **ranked variables**.
- It is especially useful when the data do not meet the assumptions required for Pearson's correlation, such as normality.



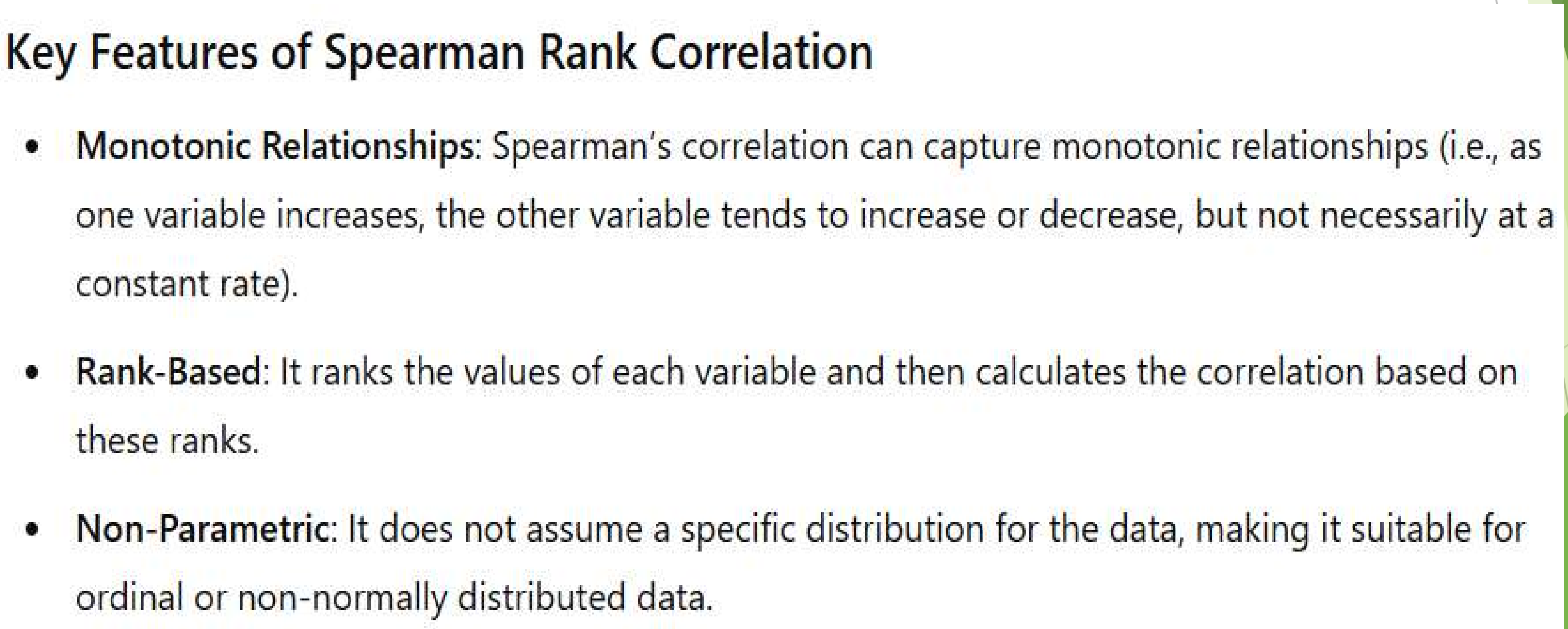
Formula for Spearman Rank Correlation

The Spearman correlation coefficient (ρ or r_s) is calculated using the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

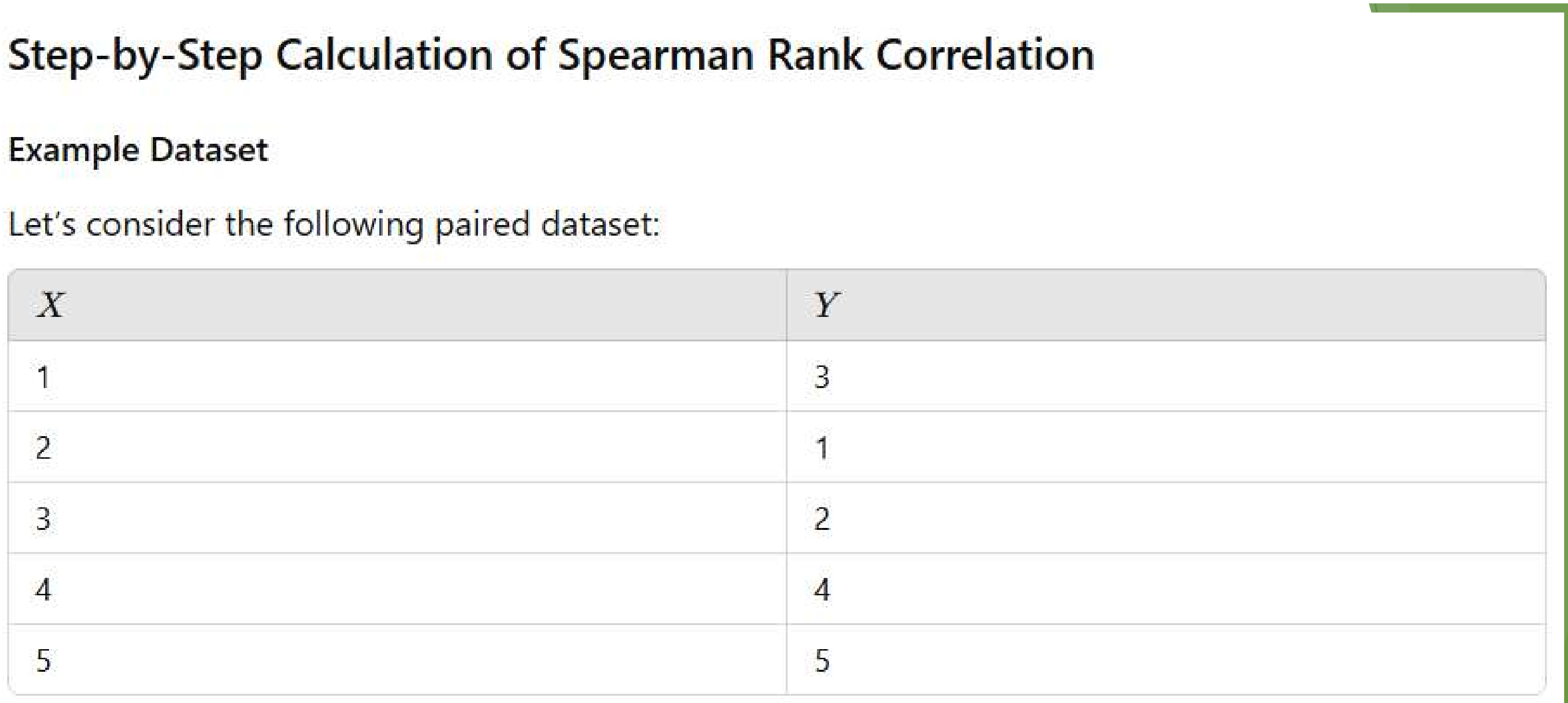
Where:

- d_i is the difference between the ranks of each pair of values.
- n is the number of observations.



Key Features of Spearman Rank Correlation

- **Monotonic Relationships**: Spearman's correlation can capture monotonic relationships (i.e., as one variable increases, the other variable tends to increase or decrease, but not necessarily at a constant rate).
- **Rank-Based**: It ranks the values of each variable and then calculates the correlation based on these ranks.
- **Non-Parametric**: It does not assume a specific distribution for the data, making it suitable for ordinal or non-normally distributed data.

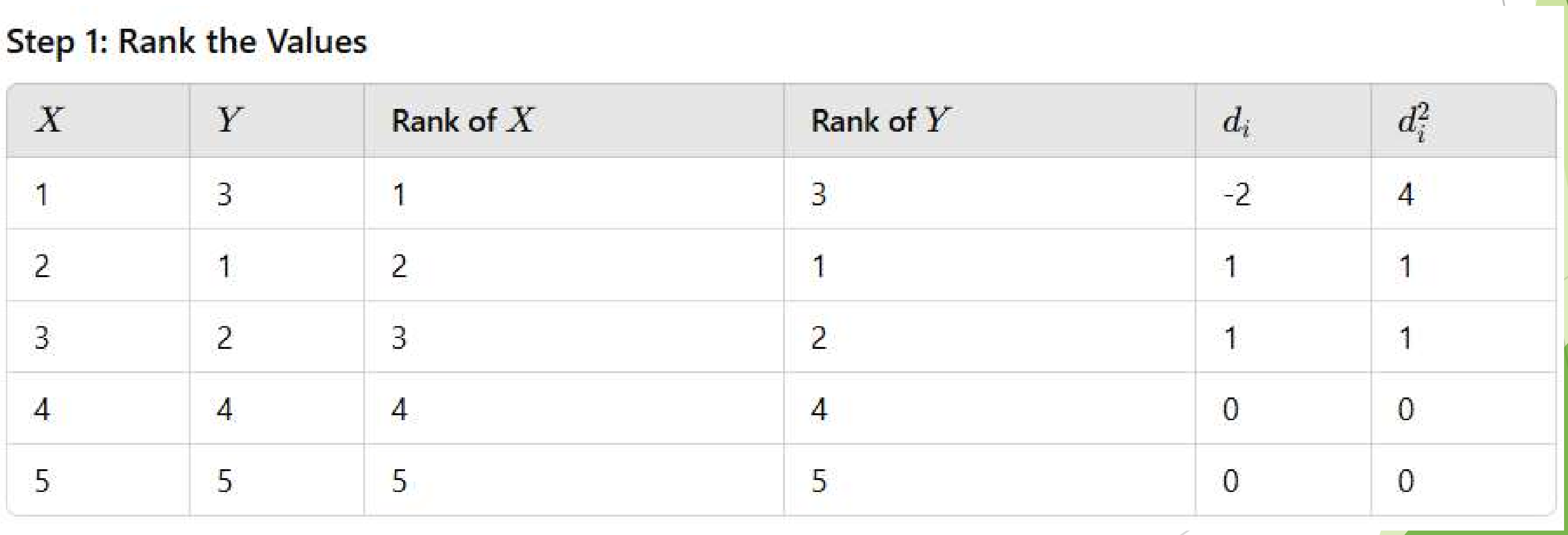


Step-by-Step Calculation of Spearman Rank Correlation

Example Dataset

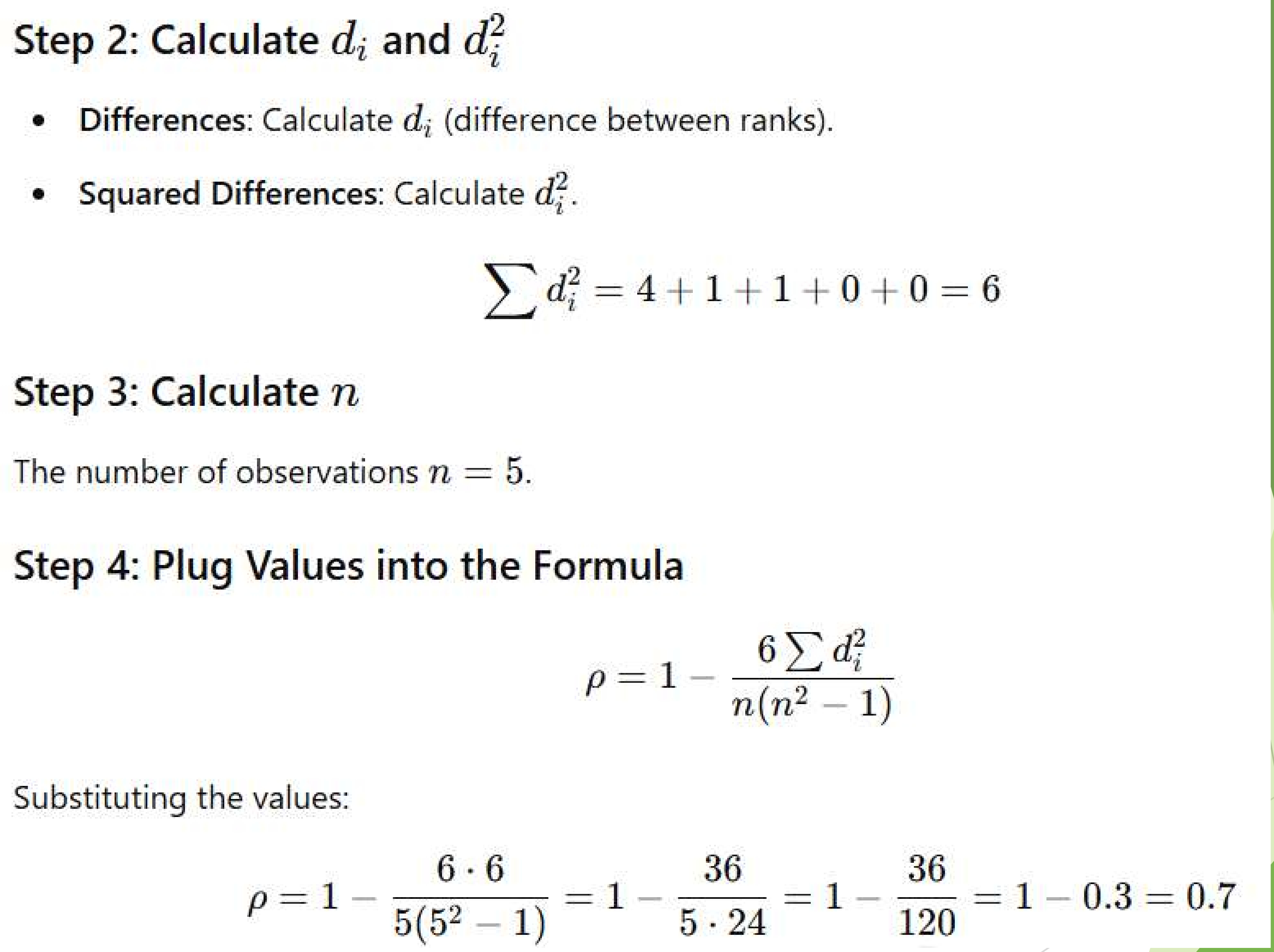
Let's consider the following paired dataset:

X	Y
1	3
2	1
3	2
4	4
5	5



Step 1: Rank the Values

X	Y	Rank of X	Rank of Y	d_i	d_i^2
1	3	1	3	-2	4
2	1	2	1	1	1
3	2	3	2	1	1
4	4	4	4	0	0
5	5	5	5	0	0



Step 2: Calculate d_i and d_i^2

- **Differences**: Calculate d_i (difference between ranks).
- **Squared Differences**: Calculate d_i^2.

$$\sum d_i^2 = 4 + 1 + 1 + 0 + 0 = 6$$

Step 3: Calculate n

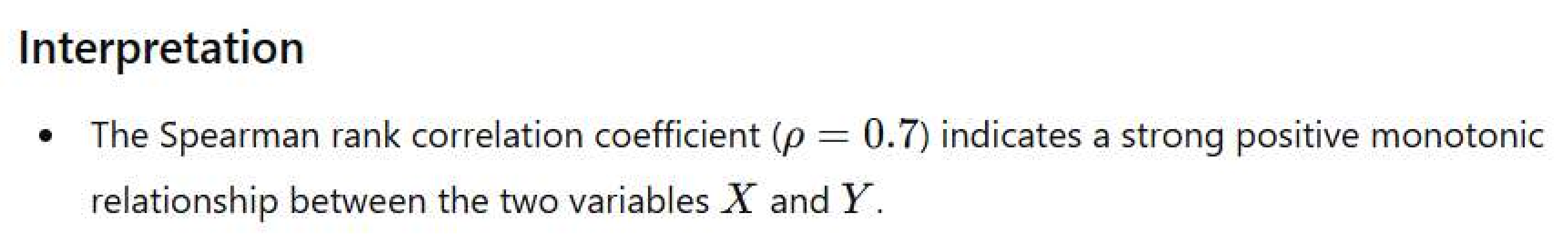
The number of observations $n = 5$.

Step 4: Plug Values into the Formula

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

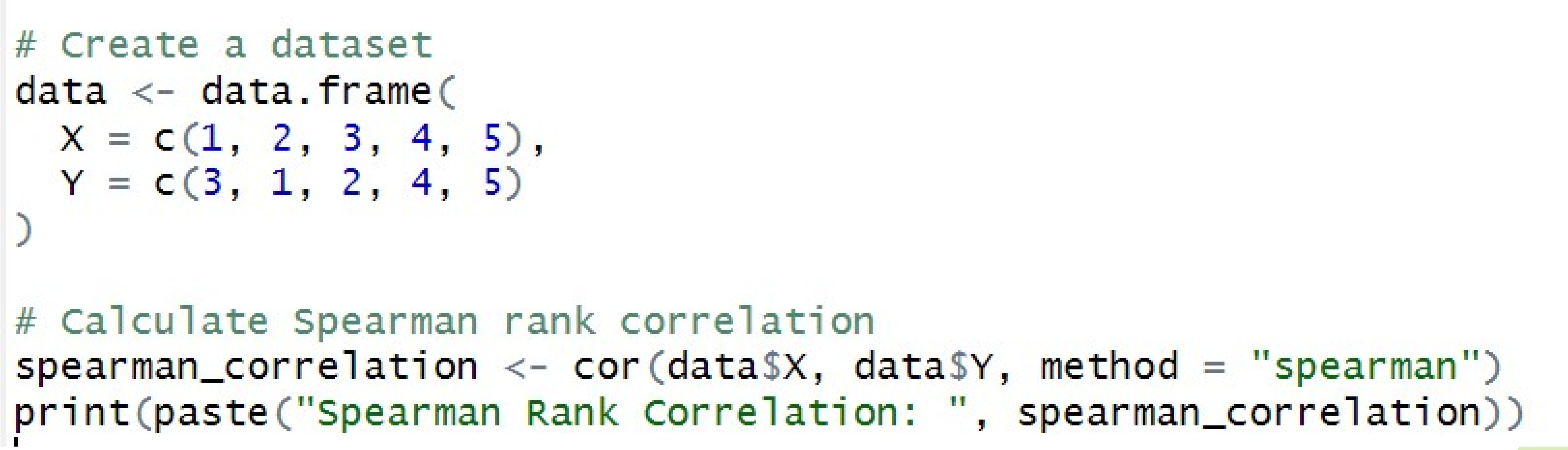
Substituting the values:

$$\rho = 1 - \frac{6 \cdot 6}{5(5^2 - 1)} = 1 - \frac{36}{5 \cdot 24} = 1 - \frac{36}{120} = 1 - 0.3 = 0.7$$

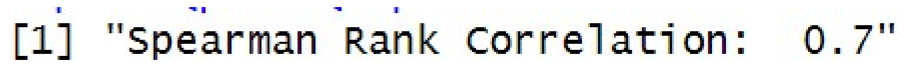


Interpretation

- The Spearman rank correlation coefficient ($\rho = 0.7$) indicates a strong positive monotonic relationship between the two variables X and Y.

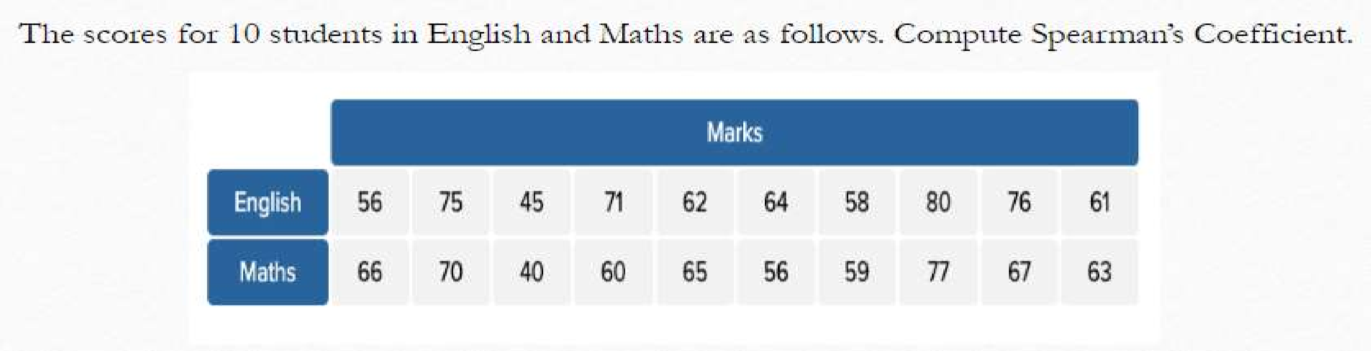


```r
# Create a dataset
data <- data.frame(
  X = c(1, 2, 3, 4, 5),
  Y = c(3, 1, 2, 4, 5)
)

# Calculate Spearman rank correlation
spearman_correlation <- cor(data$X, data$Y, method = "spearman")
print(paste("Spearman Rank Correlation: ", spearman_correlation))
```



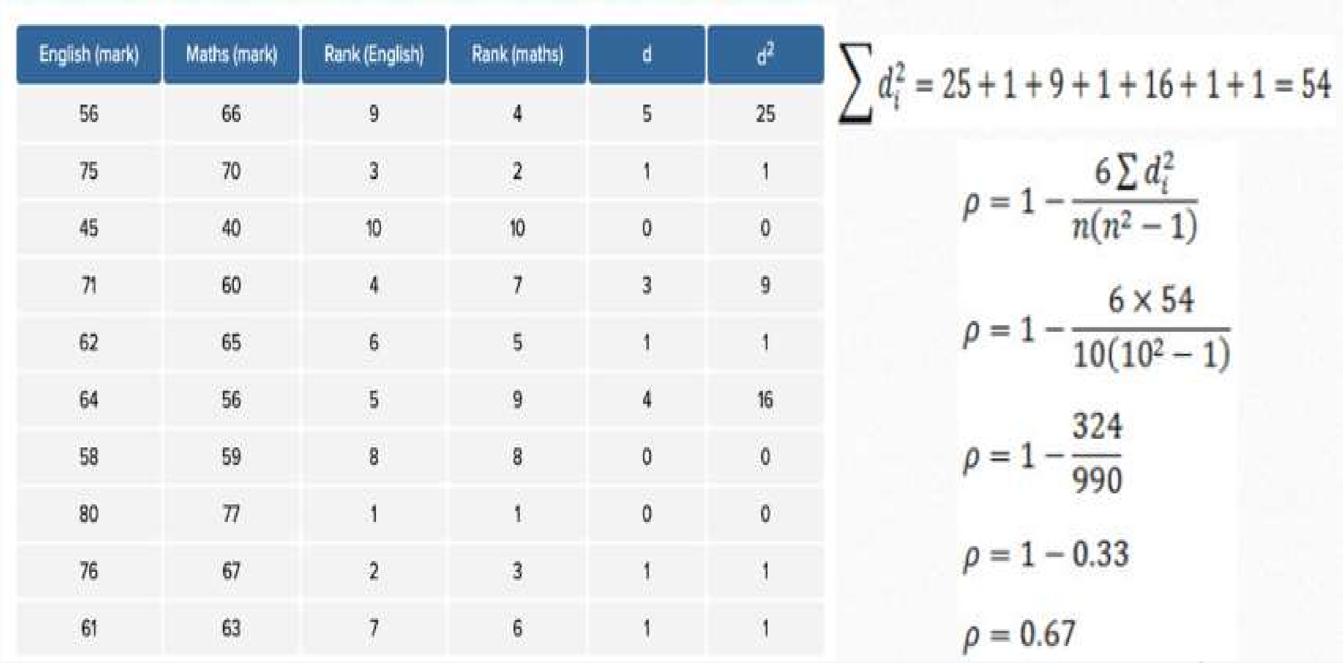
```
[1] "Spearman Rank Correlation:  0.7"
```

Question:



The scores for 10 students in English and Maths are as follows. Compute Spearman's Coefficient.

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63



English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d²
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Here Descriptive part of Stats done……..