Descriptive Stats

By Anshum Banga

www.linkedin.com/in/anshumbanga

Contents

Introduction

Statistics is a branch of mathematics that deals with collecting, organizing, analyzing, interpreting, and presenting data. It provides tools and techniques to make sense of data, enabling informed decision-making in various fields.

Descriptive Statistics

Descriptive statistics focus on summarizing and organizing data to provide a clear overview. It does not involve making predictions or inferences about a larger population; instead, it focuses on what the data shows.

Key Features:

- **Purpose:** To describe and summarize data.
- **Techniques Used:** Measures of central tendency, dispersion, and visualizations.
- **Examples:**
    - Mean, median, and mode (measures of central tendency).
    - Range, variance, and standard deviation (measures of variability).
    - Charts like histograms, bar graphs, or pie charts.

Inferential Statistics

Inferential statistics use data from a sample to make predictions, inferences, or generalizations about a larger population. This involves estimating parameters and testing hypotheses.

Key Features:

- **Purpose:** To make conclusions about a population based on sample data.
- **Techniques Used:** Hypothesis testing, confidence intervals, regression analysis.
- **Examples:**
    - Estimating the average height of all adults in a city based on a sample.
    - Determining whether a new drug is effective by analysing test results from a group of participants.

Differences Between Descriptive and Inferential Statistics

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	To summarize and describe data.	To draw conclusions about a population.
Data	Focuses on the data available.	Focuses on generalizing to the population.
Examples	Mean, median, standard deviation.	Hypothesis testing, confidence intervals.
Visualization	Charts, graphs, and tables.	Not focused on visualizations but predictions.

Measures of Central Tendency

Measures of central tendency are statistical tools used to identify the center or typical value of a dataset. They summarize the data with a single value that represents the entire distribution. The three primary measures are **mean**, **median**, and **mode**.

In Other words,

Measures of central tendency are ways to find the "middle" or "typical" value in a set of numbers. They give us an idea of where most of the data is centred. The three main ways to measure this are **mean**, **median**, and **mode**.

1. Mean (Arithmetic Average)

The mean is the sum of all values divided by the total number of values.

The mean is the sum of all values divided by the total number of values.

Formula:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

Example:

For the dataset: 5, 10, 15

$$\text{Mean} = \frac{5 + 10 + 15}{3} = 10$$

Use Case:

- Useful for numerical data and when all values contribute equally.
- **Limitation**: Affected by outliers (extremely high or low values).

2. Median (Middle Value)

The median is the middle value when the data is arranged in ascending order.

- If the number of values is odd: The middle value is the median.
- If the number of values is even: The median is the average of the two middle values.

Example:

For the dataset: 5, 10, 15 (Odd count):

$$\text{Median} = 10$$

For the dataset: 5, 10, 15, 20 (Even count):

$$\text{Median} = \frac{10 + 15}{2} = 12.5$$

Use Case:

- Useful for skewed data or when there are outliers, as it is not affected by extreme values.

3. Mode

The mode is the value that occurs most frequently in the dataset.

- A dataset can have one mode (**unimodal**), two modes (**bimodal**), or more (**multimodal**).
- If no value repeats, there is no mode.

Example:

For the dataset: 5, 10, 10, 15

Mode=10

Use Case:

- **Useful for categorical data or when identifying the most common value.**

Key Differences

Measure	Best For	Sensitivity to Outliers
Mean	**Symmetrical datasets**	**High**
Median	**Skewed datasets**	**Low**
Mode	**Categorical and repetitive data**	**None**

Use Cases

1. Mean (Average)

Question: What is the average score of students in a class?

- **Scenario**: A teacher wants to understand how students performed in an exam.
- **How to Use**: Add up all the scores and divide by the number of students.
- **Why Preferred**:
    - It gives an overall idea of the performance of the entire class.
    - Works well when data is evenly distributed without extreme values.

Example Calculation:
Scores: 70, 80, 90, 85, 75

$$\text{Mean} = \frac{70 + 80 + 90 + 85 + 75}{5} = 80$$

The average score is **80**.

2. Median (Middle Value)

Question: What is the middle salary of employees in a company?

- **Scenario**: A company wants to report the "typical" salary of its employees, but the CEO's extremely high salary might distort the average.
- **How to Use**: Arrange salaries in order and find the middle value.
- **Why Preferred**:
    - Median is not affected by outliers (e.g., a CEO’s very high salary).
    - It provides a better sense of the "typical" salary when data is skewed.

Example Calculation:
Salaries: 25,000, 30,000, 35,000, 40,000, 5,00,000

Median=35,000

The median shows the "typical" salary, while the mean would be much higher due to the CEO's salary.

3. Mode (Most Common Value)

Question: What is the most preferred shoe size sold in a store?

- **Scenario**: A shoe store wants to stock the most popular sizes to meet customer demand.
- **How to Use**: Identify the size that appears most frequently in sales data.
- **Why Preferred**:
    - Mode helps find the most common preference in a dataset.
    - Useful for categorical or repetitive data.

Example Calculation:
Sizes sold: 8, 9, 9, 10, 8, 9, 7

Mode=9

Shoe size **9** is the most preferred, so stocking more of this size is a good business decision.

Comparison Table

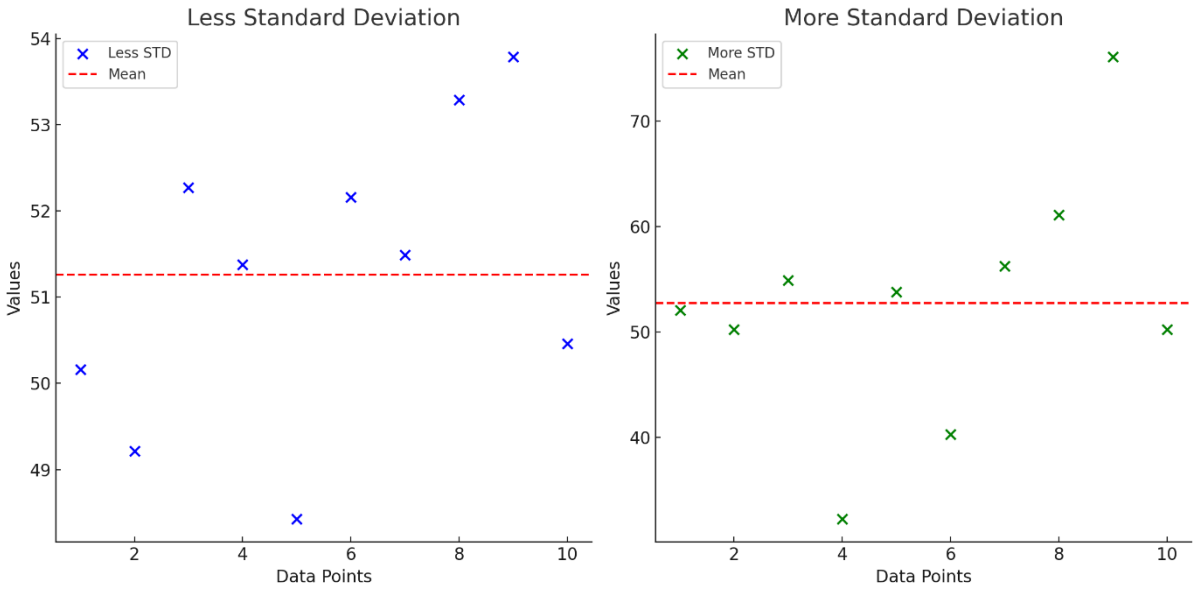
Measure	Scenario	Why Preferred
Mean	Average test scores, average temperatures.	Best for balanced, evenly distributed data.
Median	Middle salary, house prices.	Best for skewed data or data with outliers.
Mode	Most popular product, common shoe size.	Best for identifying trends or preferences.

Measure of Dispersion

A **measure of dispersion** tells us how spread out or scattered the data values are in a dataset. While measures of central tendency (like mean, median, and mode) give us an idea of the center of the data, dispersion shows how much the data varies around that center.

Why is Measure of Dispersion Important?

- Helps us understand **variability** in the data.
- Identifies whether the data points are **closely packed** or **widely spread**.
- Provides insights into **data consistency** (e.g., how consistent students' scores are in an exam).



Key Measures of Dispersion

1. Range

The difference between the largest and smallest values in the dataset.

- **Formula:**

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Example:
For data: 5, 10, 15, 20

Range=20−5=15

Use Case: Simple to calculate but doesn't show how data is distributed.

2. Variance

Measures the average squared difference between each data point and the mean.

- **Formula:**

$$\text{Variance} = \frac{\sum(x_i - \mu)^2}{n}$$

Where x_i is each data value, μ is the mean, and n is the number of values.

3. Standard Deviation (SD)

The square root of the variance. It tells us how much the data points deviate from the mean, in the same units as the data.

- **Formula:**

$$\text{SD} = \sqrt{\text{Variance}}$$

- **Example:** If variance is 16:

$$\text{SD} = \sqrt{16} = 4$$

Use Case: Commonly used in fields like finance and science to measure data consistency.

4. Interquartile Range (IQR)

The range of the middle 50% of the data, calculated as:

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

Where Q1 is the first quartile (25th percentile), and Q3 is the third quartile (75th percentile).

- **Use Case**: Useful for datasets with outliers, as it focuses on the central data

Summary

Measure	Best For	Limitations
Range	Quick overview of spread.	Sensitive to outliers.
Variance	Detailed variability analysis.	Hard to interpret due to squared units.
Standard Deviation	Consistency of data (same units).	Sensitive to outliers.
IQR	Skewed data and outliers.	Ignores extreme values.

Covariance

Covariance is a measure of the relationship between two variables and how they change together. It shows whether an increase in one variable corresponds to an increase or decrease in another variable.

Formula for Covariance:

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

Where:

- X_i, Y_i: Data points of variables X and Y.
- $\bar{X}, \bar{Y}$: Mean of X and Y.
- n: Number of data points.

Key Points:

1. **Positive Covariance**:
   If one variable increase and the other tends to increase too, the covariance is positive.
   Example: Height and weight.
2. **Negative Covariance**:
   If one variable increase while the other tends to decrease, the covariance is negative.
   Example: Time spent studying and errors in a test.
3. **Zero Covariance**:
   If there's no consistent pattern between the variables, the covariance is zero.
   Example: Height and favourite colour.

Limitations of Covariance:

- Covariance only tells the direction (positive/negative) of the relationship, not the strength.
- It is not scaled, so the value depends on the units of the variables.

Pearson's Correlation Coefficient

Pearson's correlation coefficient, often denoted by ***r***, is a standardized measure that shows both the **strength** and **direction** of the relationship between two variables.

Formula for Pearson's Correlation Coefficient:

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

Where:

- $\text{Cov}(X,Y)$: Covariance of X and Y.
- σ_X, σ_Y: Standard deviations of X and Y.

Key Points:

Values of r:

r = +1 : Perfect Correlation

r = -1 : Perfect Negative Correlation

r = 0 : No Correlation

Strength of Correlation:

Strong: ***r*** >0.7 or ***r*** <−0.7

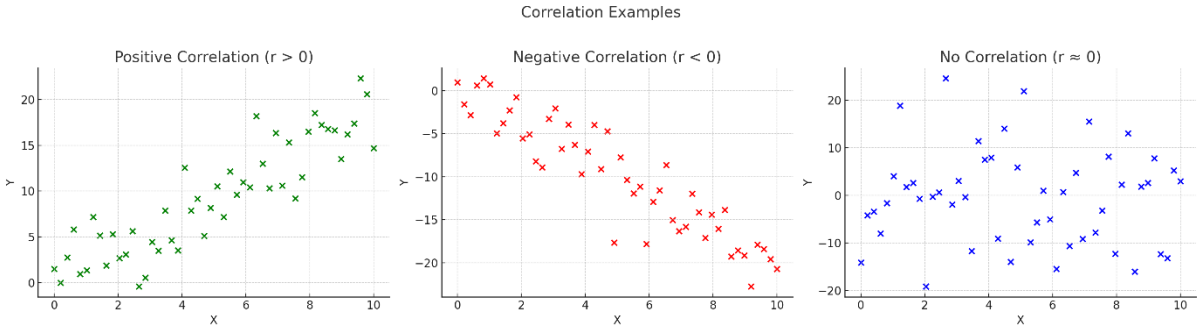
Moderate: 0.3< ***r*** ≤0.7 or −0.7≤ ***r*** <−0.3

Weak: 0.3≤ ***r*** ≤0.3

Direction of Correlation:

- Positive (***+r***): Variables move in the same direction.
- Negative (***−r***): Variables move in opposite directions.

Graphs



Positive Correlation (r > 0): As ***X*** increases, ***Y*** also increases. The points form an upward trend.

Negative Correlation (r < 0): As ***X*** increases, ***Y*** decreases. The points form a downward trend.

No Correlation (r ≈ 0): There is no apparent relationship between ***X*** and ***Y***. The points are scattered randomly.

What is Percentile?

A **percentile** indicates the value below which a given percentage of data falls in a dataset.
Example: If you are at the 90th percentile in a test, it means 90% of the students scored below you.

Difference Between Percentile and Percentage

Feature	Percentile	Percentage
Definition	Value below which a percentage of data lies.	Ratio expressed as a fraction of 100.
Purpose	Compares an individual to the dataset.	Measures part of a total.
Example	90th percentile means 90% of values are below.	90% means 90 out of 100 marks.

What are Quartiles?

Quartiles divide a dataset into four equal parts:

1. **Q1 (First Quartile)**: The 25th percentile (25% of the data is below this value).
2. **Q2 (Second Quartile or Median)**: The 50th percentile (middle value of the dataset).
3. **Q3 (Third Quartile)**: The 75th percentile (75% of the data is below this value).
4. **Q4 (Fourth Quartile)**: The highest value (100th percentile)

Interquartile Range (IQR)

Definition: Measures the spread of the middle 50% of the data.

$$\mathbf{IQR = Q3 - Q1}$$

Lower Fence and Upper Fence

The **lower fence** and **upper fence** are terms used in statistics, particularly in the context of identifying **outliers** in a dataset. They are part of the **interquartile range (IQR)** method for outlier detection.

Formulas:

- **Lower Fence** = Q1 - (1.5 × IQR)
- **Upper Fence** = Q3 + (1.5 × IQR)

Where:

- **Q1** is the first quartile (25th percentile).
- **Q3** is the third quartile (75th percentile).
- **IQR** (Interquartile Range) = Q3 - Q1

Purpose:

- The **lower fence** defines the lower limit below which data points are considered outliers.
- The **upper fence** defines the upper limit above which data points are considered outliers.

Interpretation:

- **Any data point** below the lower fence or above the upper fence is considered an **outlier**.
- Data points that fall between the lower and upper fences are considered **normal** values within the range of the distribution.

Boxplot

A **boxplot** (also known as a **box-and-whisker plot**) is a graphical representation of the distribution of a dataset. It provides a summary of data through its five-number summary:

1. **Minimum**: The smallest data point (excluding outliers).
2. **First Quartile (Q1)**: The median of the lower half of the dataset (25th percentile).
3. **Median (Q2)**: The middle value of the dataset (50th percentile).
4. **Third Quartile (Q3)**: The median of the upper half of the dataset (75th percentile).
5. **Maximum**: The largest data point (excluding outliers).

Question:

The weekly sales of a product over 12 weeks are:
50, 55, 53, 48, 60, 47, 62, 65, 45, 58, 150, 52.
Create a boxplot, calculate the five-number summary, and identify any outliers.

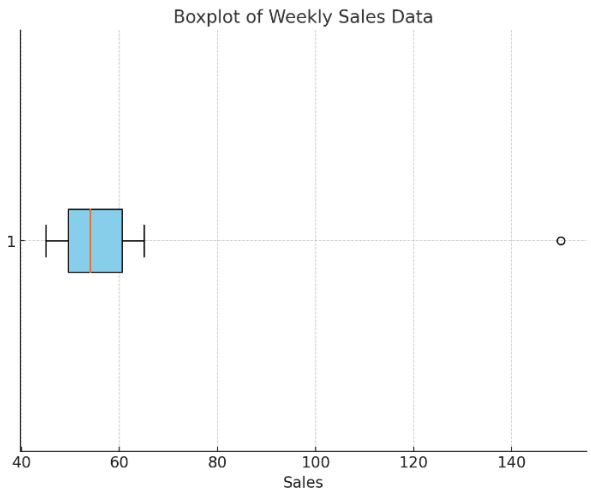
Five-Number Summary:

- **Q1 (25th percentile)**: 49.5
- **Q3 (75th percentile)**: 60.5
- **Interquartile Range (IQR)**: **Q3-Q1 =** 60.5 – 49.5 = 11
- **Lower Fence =** **Q1-1.5*IQR =** 49.5−16.5= 33.0
- **Upper Fence=** **Q3+1.5*IQR =** 60.5+16.5= 77.0

Outliers:

Any data point below 33.0 or above 77.0 is considered an outlier.

- **Outlier**: 150



Shapes of Graphs

Gaussian Curve

A **Gaussian Curve**, also known as the **Normal Distribution** or **Bell Curve**, is a statistical representation of a dataset where the data is symmetrically distributed around the mean. It is one of the most important probability distributions in statistics and is widely used in various fields like data science, physics, and social sciences.

Key Characteristics of a Gaussian Curve:

Shape:

- Bell-shaped and symmetric around the mean.
- The highest point (the peak) represents the mean, median, and mode, which are all equal in a perfect Gaussian distribution.

Mean, Median, and Mode:

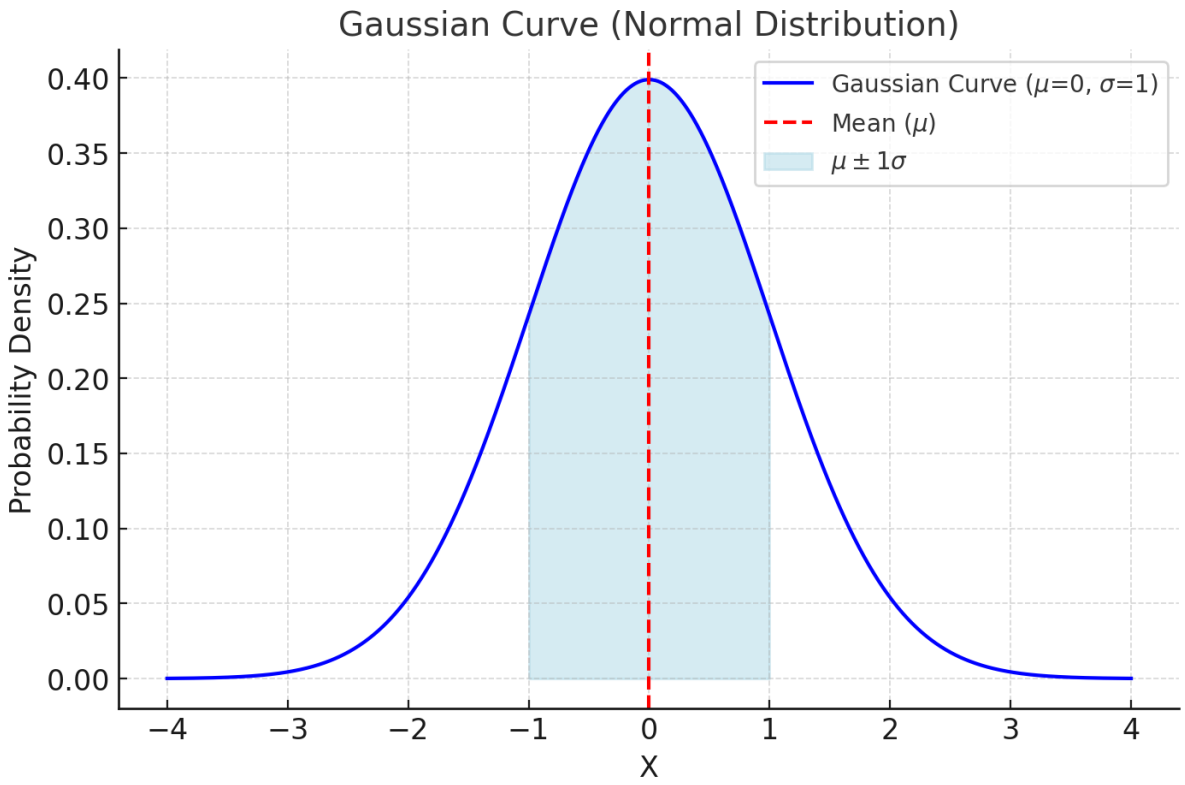
- All are located at the center of the curve.
- The curve is symmetric, so these measures of central tendency coincide.

Standard Deviation (σ):

- Controls the width of the curve.
- A smaller σ results in a narrow and steep curve, while a larger σ makes the curve wider and flatter.

Empirical Rule:

- Approximately **68%** of the data lies within ±1σ of the mean.
- About **95%** of the data lies within ±2σ of the mean.
- Nearly **99.7%** of the data lies within ±3σ of the mean.



Importance of the Gaussian Curve

1. **Central Limit Theorem**:
    - The Gaussian curve plays a central role in the **Central Limit Theorem**, which states that the distribution of the sum (or average) of a large number of independent, identically distributed variables tends to be normal, regardless of the original distribution.
2. **Real-World Applications**:
    - **Natural Phenomena**: Heights, weights, blood pressure, and IQ scores often follow a Gaussian distribution.
    - **Error Distribution**: Measurement errors and noise in systems are usually Gaussian.
    - **Machine Learning**: Used in algorithms like Naive Bayes, and in feature scaling methods.

Left Skewed Distribution (Negatively Skewed)

A **left-skewed distribution**, also known as a **negatively skewed distribution**, is a type of distribution where the **tail extends to the left**, or toward the lower values on a number line.

Characteristics of a Left-Skewed Distribution:

Tail:

- The left tail (low-value end) is longer than the right tail.
- Indicates that there are some smaller outliers pulling the mean downward.

Order of Mean, Median, Mode:

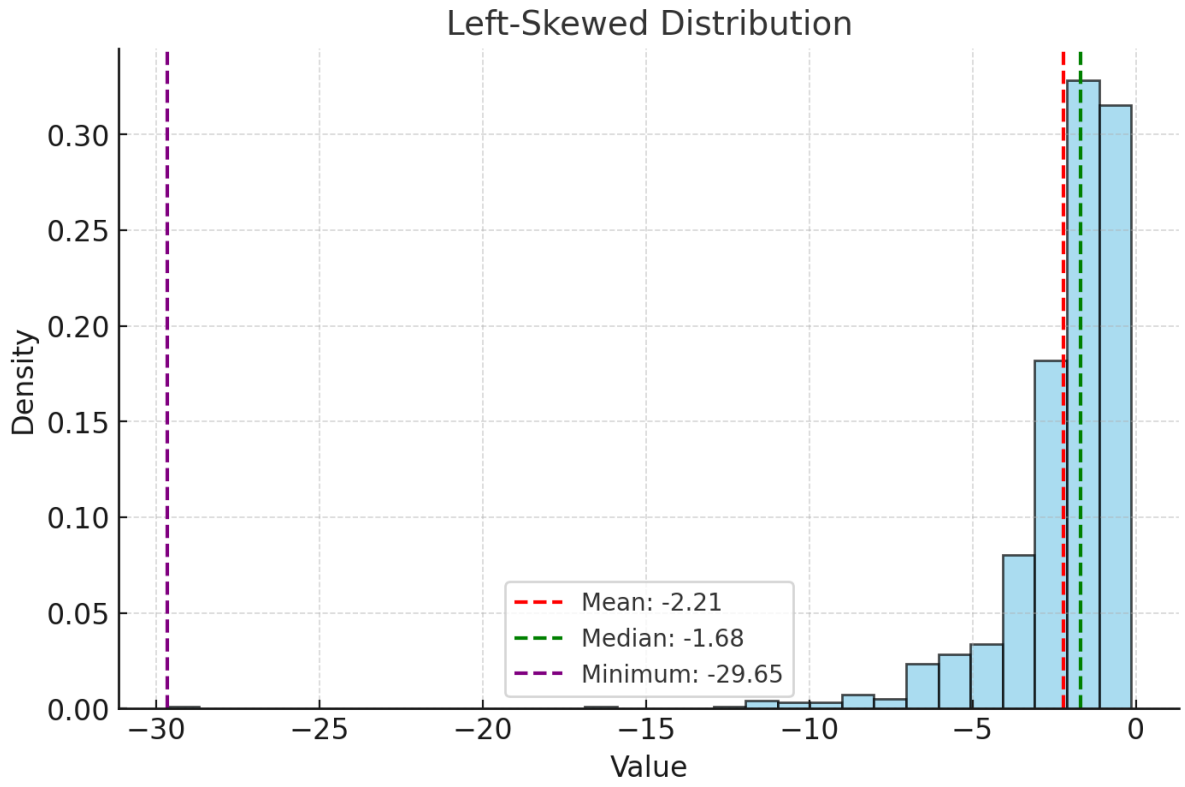
- **Mean** < **Median** < **Mode**.
- The mean is most affected by the small outliers and is shifted to the left.

Concentration of Data:

- Most data points are clustered on the right (higher values).

Visual Appearance:

- The peak of the distribution is on the right, with a tail extending to the left.



Tail: The longer tail is on the left, representing the smaller values.

Mean and Median:

- The **mean** (red dashed line) is smaller than the **median** (green dashed line).
- This is due to the influence of the smaller values (outliers) pulling the mean downward.

Data Concentration:

- Most data points are concentrated on the right (higher values).

Real-Life Examples of Left-Skewed Distributions

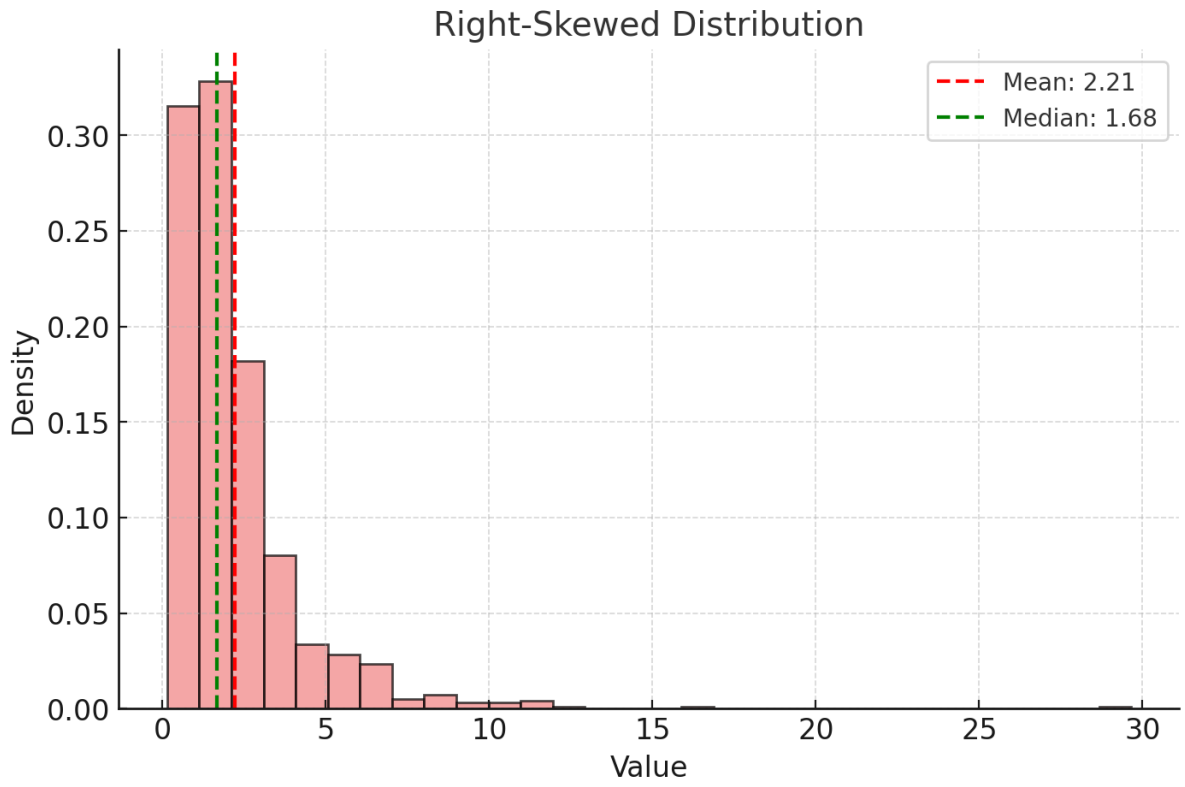
- **Age at Retirement**: Most people retire around a specific age, but a few may retire much earlier, creating a left-skewed pattern.
- **Lifespan of a population**: If the majority of individuals live long lives, but a few die at a younger age, the distribution can be left-skewed.
- **Exam Scores in a Simple Test**: If most students score high marks, but a few perform poorly

Right Skewed Graph (Positively Skewed)

A **right-skewed distribution**, also known as a **positively skewed distribution**, is a type of distribution where the **tail extends to the right**, or toward the higher values on a number line.

Characteristics of a Right-Skewed Distribution

1. **Tail**:
    - The right tail (high-value end) is longer than the left tail.
    - Indicates that there are a few larger outliers pulling the mean upward.
2. **Order of Mean, Median, Mode**:
    - **Mode** < **Median** < **Mean**.
    - The mean is most affected by the larger outliers and is shifted to the right.
3. **Concentration of Data**:
    - Most data points are clustered on the left (lower values).
4. **Visual Appearance**:
    - The peak of the distribution is on the left, with a tail extending to the right.



Tail: The longer tail extends to the right, representing the higher values (outliers).

Mean and Median:

- The **mean** (red dashed line) is greater than the **median** (green dashed line).
- This is because the larger values pull the mean upward.

Data Concentration:

- Most of the data points are clustered on the left, near lower values.

Real-Life Examples of Right-Skewed Distributions

- **Income Distribution**: Most people earn below the average income, with a few high earners creating a right skew.
- **House Prices**: Majority of houses fall into a typical price range, but luxury properties create a right tail.
- **Hospital Stay Duration**: Most stays are short, but some patients stay for extended periods.

Central Limit Theorem

The **Central Limit Theorem (CLT)** is a fundamental principle in statistics that states:

When you take sufficiently large random samples from a population with any shape of distribution, the **sampling distribution of the sample mean** will be approximately **normal** (Gaussian), regardless of the population's original distribution.

Key Points of the Central Limit Theorem

1. **Applies to Sampling Distributions**:
    - The theorem pertains to the **distribution of sample means**, not the original population distribution.
2. **Sample Size Matters**:
    - The approximation to a normal distribution improves as the **sample size (n)** increases.
    - A sample size of n≥30 is often considered sufficient for the CLT to hold in most cases.
3. **Population Mean and Variance**:
    - The mean (μ) of the sampling distribution equals the mean of the population.
    - The standard deviation (σ) of the sampling distribution, called the **standard error**, is given by:

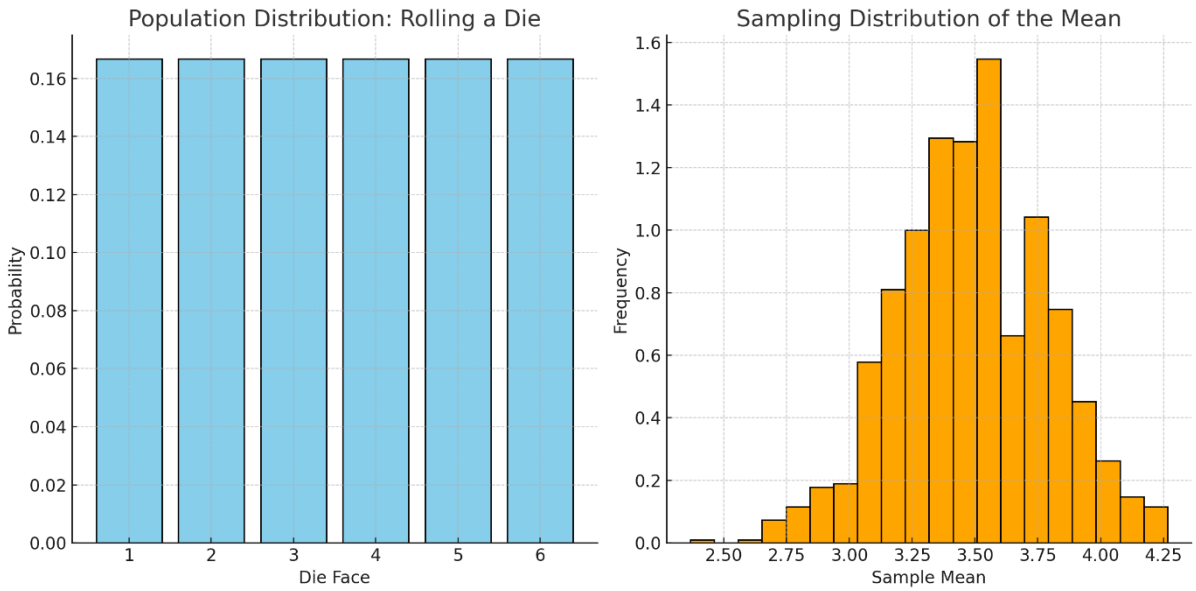
$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation, and n is the sample size.

4. **Shape of Population Distribution**:
    - The CLT holds true regardless of the population's distribution shape (e.g., skewed, uniform, bimodal), as long as the sample size is large enough.

Importance of the Central Limit Theorem

1. **Foundation for Inferential Statistics**:
    - It justifies using the normal distribution to approximate sampling distributions, enabling hypothesis testing and confidence interval construction.
2. **Practical Applications**:
    - Used in quality control, polling, finance, and machine learning, where the population distribution is unknown or not normal.
3. **Simplifies Analysis**:
    - Allows analysts to use the properties of the normal distribution, which is mathematically convenient.



Left Graph: The population distribution represents the uniform probability of rolling a standard 6-sided die. Each die face (1 to 6) has an equal chance (1/6).

Right Graph: The sampling distribution of the sample mean, created by taking 1000 random samples of size 30 from the population, shows a bell-shaped curve. This demonstrates the Central Limit Theorem: regardless of the original population distribution, the distribution of sample means approaches a normal distribution as the sample size increases.

Real-Life Example

Customer Spending in a Supermarket

- **Population**: Individual spending by customers in a supermarket (likely skewed as some spend a lot, while others spend very little).
- **Sampling**: Select random groups of 50 customers and compute the average spending for each group. The distribution of these sample means will be approximately normal.

Quality Control in Manufacturing

- **Population**: The weight of a product (e.g., packets of chips), which may vary slightly due to machine imperfections.
- **Sampling**: Take random samples of 50 packets from each day's production and calculate the mean weight. The sampling distribution of the mean will be normal, helping manufacturers detect production anomalies.

Exploratory Analysis

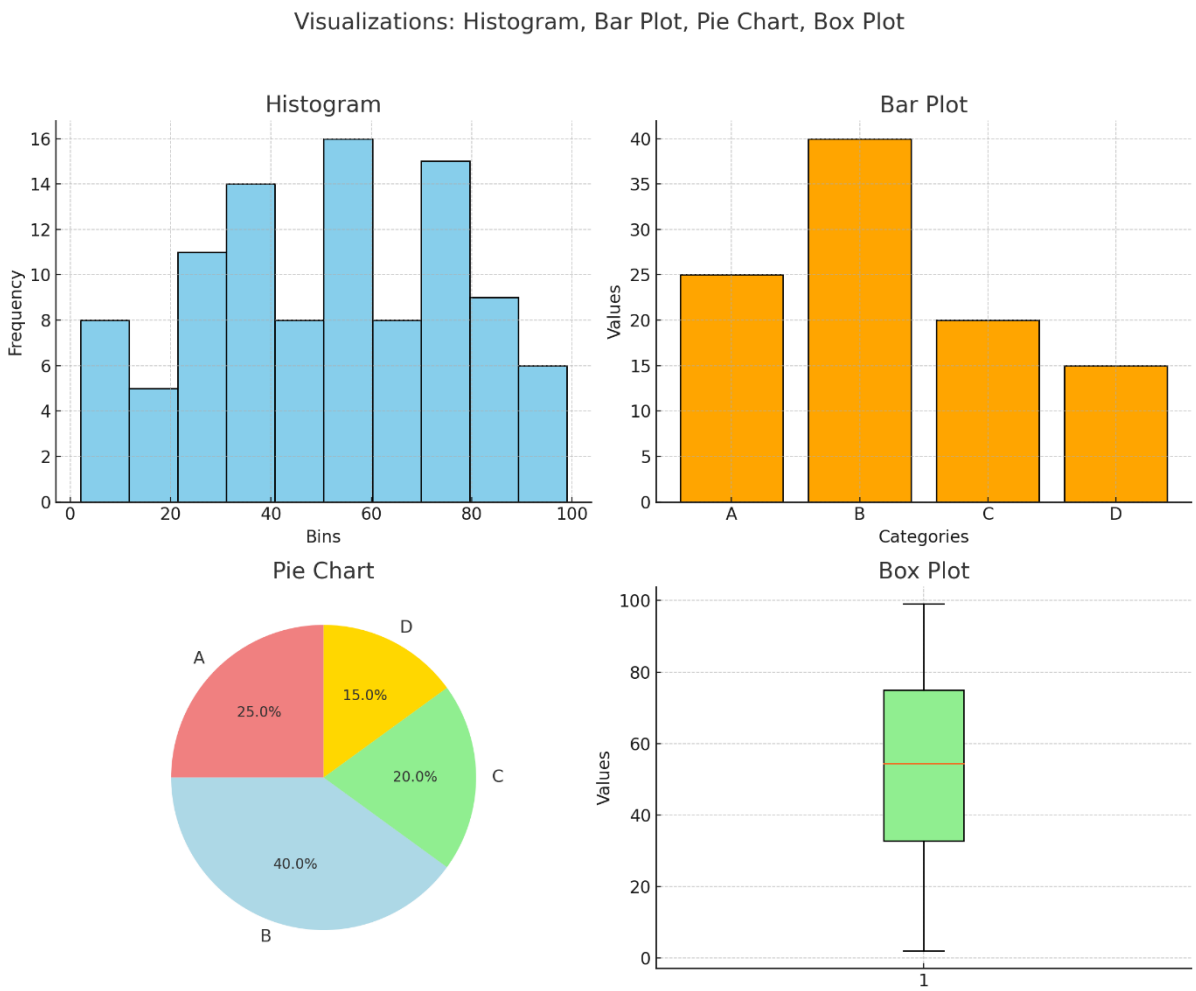
Univariate Analysis

Definition: Analysis of a single variable to summarize and understand its characteristics.

Purpose: Focuses on one variable at a time to describe its distribution, central tendency, and variability.

Techniques:

- **Descriptive Statistics**: Mean, median, mode, variance, standard deviation.
- **Visualization**: Histograms, bar plots, pie charts, box plots.



1. **Histogram** (Top-Left): Shows the frequency distribution of the data values across bins.
2. **Bar Plot** (Top-Right): Displays values associated with different categories (A, B, C, D).
3. **Pie Chart** (Bottom-Left): Represents the proportion of values for each category as a percentage.
4. **Box Plot** (Bottom-Right): Summarizes the distribution of the data, including the median, quartiles, and potential outliers.

Bivariate Analysis

Definition: Analysis of the relationship between two variables to understand their association.

Purpose: Explores how one variable influence or correlates with another.

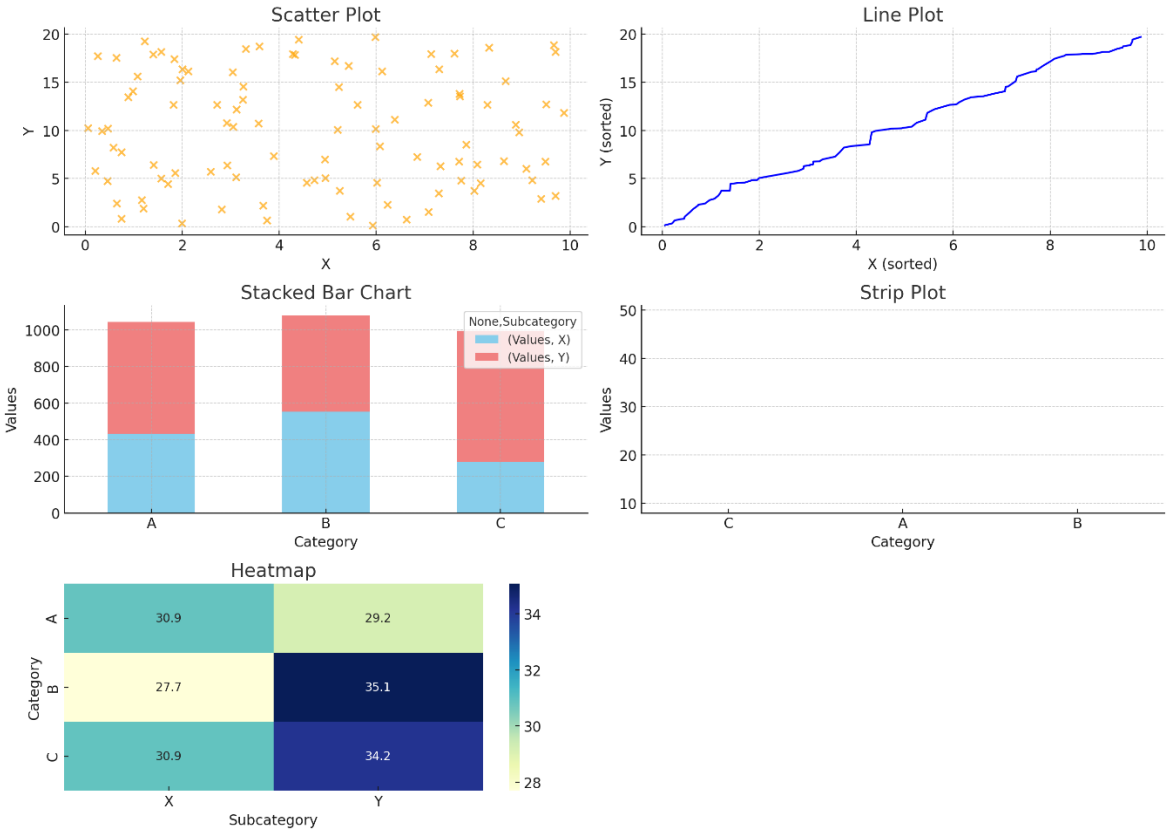
Types:

- **Numerical-Numerical**: Correlation, scatter plots.
- **Categorical-Categorical**: Cross-tabulation, chi-square test.
- **Numerical-Categorical**: Box plots, t-tests, ANOVA.

Techniques:

- Correlation coefficient (e.g., Pearson, Spearman).
- Regression analysis (simple linear regression).
- Visualization: Scatter plots, bar plots, grouped box plots.

Commonly used Graphs example:



1. **Scatter Plot**: Shows the relationship between two numerical variables, X and Y.
2. **Line Plot**: Displays a trend line by sorting X and Y values.
3. **Stacked Bar Chart**: Compares values for Subcategories (X and Y) within each Category (A, B, C).
4. **Strip Plot**: Displays individual data points for a numerical variable (Values) across categories (A, B, C).
5. **Heatmap**: Represents the average values for each combination of Category and Subcategory using a colour gradient.

1. Numerical vs. Numerical

These graphs help analyse the relationship or correlation between two numerical variables.

- **Scatter Plot**: Displays points representing two variables. Ideal for showing trends, patterns, and correlations.
- **Line Plot**: Useful if one variable is time or sequential in nature.
- **Joint Plot**: Combines scatter plots and histograms for both variables.
- **Bubble Plot**: Similar to a scatter plot but includes a third variable through the size of the points.

---

2. Categorical vs. Categorical

These graphs help compare or explore relationships between two categorical variables.

- **Stacked Bar Chart**: Compares the proportions of categories across groups.
- **Grouped Bar Chart**: Similar to stacked, but groups bars side by side.
- **Heatmap**: Uses a color-coded matrix to display relationships between categories.

---

3. Numerical vs. Categorical

These graphs help compare a numerical variable across categories.

- **Box Plot**: Shows the distribution of a numerical variable within each category.
- **Violin Plot**: Similar to a box plot but adds a kernel density estimate to show distribution.
- **Strip Plot**: Displays individual data points along with their distribution across categories.
- **Swarm Plot**: Similar to a strip plot but adjusts for overlapping points.
- **Bar Plot with Error Bars**: Compares means or medians across categories with error bars for variance.

Multivariate Analysis

Definition: Analysis of more than two variables simultaneously to understand complex relationships and patterns.

Purpose: Captures interactions and dependencies among multiple variables.

Types:

- **Regression Analysis**: Multiple linear regression, logistic regression.
- **Dimensionality Reduction**: PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis).
- **Clustering**: K-means, hierarchical clustering.
- **Machine Learning**: Classification, regression, and unsupervised learning techniques.

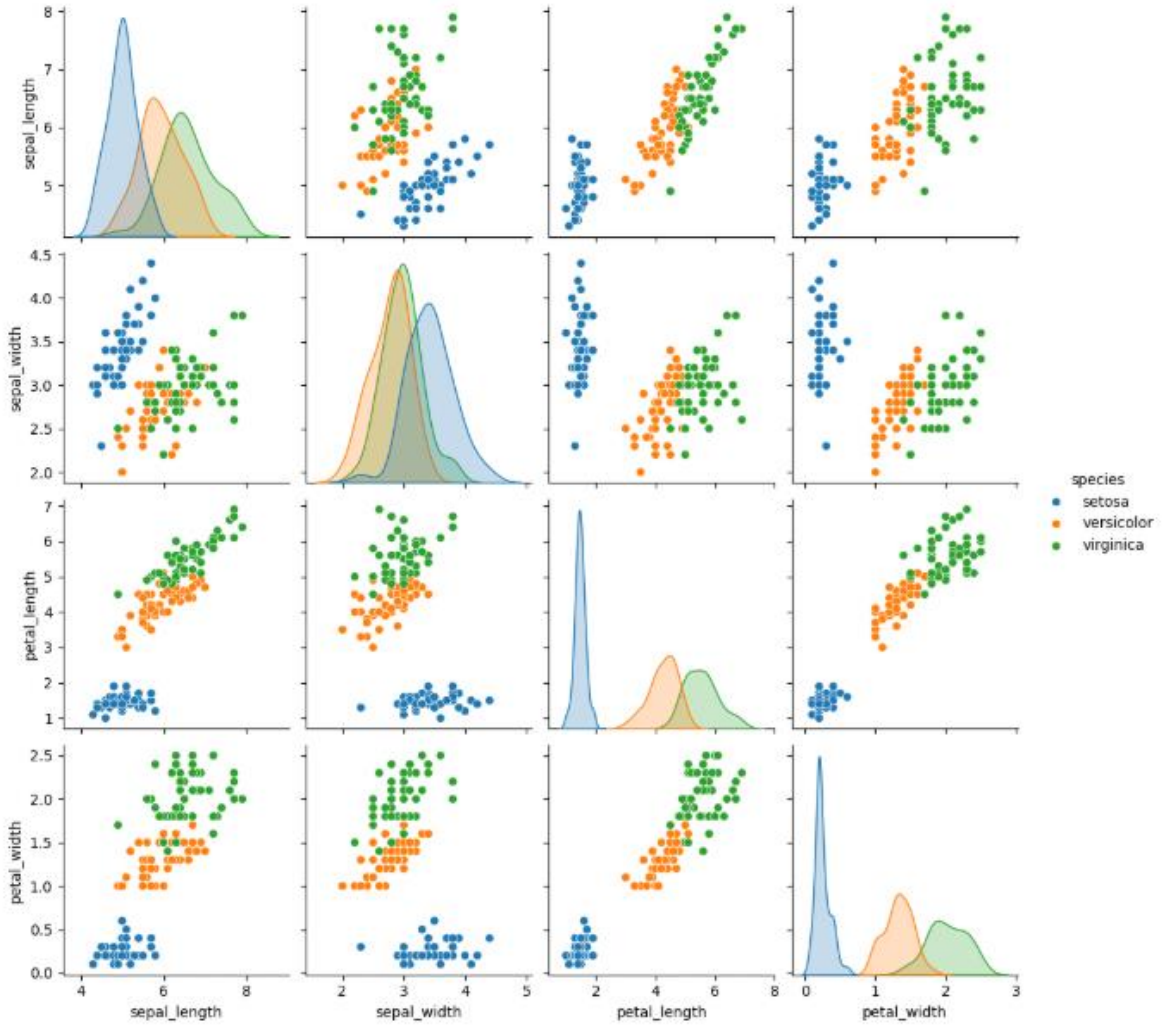
Visualization:

- Pair plots, heatmaps, 3D scatter plots.

Example: Predicting house prices based on **size, location, number of bedrooms, and age of the house**.

- Insights: Understand how all these factors collectively impact house prices.

Commonly used Graphs examples:



- Each subplot shows the scatterplot between two features.
- The diagonal shows the distribution of a single feature.
- Colours represent different species of flowers

Follow for more content:

www.github.com/anshumbanga ; www.linkedin.com/in/anshumbanga