- **Q1. What is statistics, and why is it important ?**
- Ans. Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data.
- It's important because it allows us to:

🔧 **Informed Decision Making**

- Statistics provides methods and tools to analyze data enabling us to perform better and evidenced based results in various fields such as business, healthcare, and public policy.

🔧 **Data Interpretation**:

- It helps in interpreting complex data sets, making it easier to understand trends, patterns, and relationships within the data. Which is further used in machine learning model to provide better and accurate results.

🔧 **Problem Solving**

- Statistics equips individuals with the skills to solve real-world problems by applying appropriate statistical techniques to analyze and interpret data.

🔧 **Predictive Analysis**

- Using the inferential statistics we can make prediction about future of events and the results associated with it and based on these results we can plan our future events accordingly. It is also used in forecasting of events.

🔧 **Quality Control**

- In industries, statistical methods are used for quality control and improvement processes, ensuring products and services meet certain standards and specifications.

🔧 **Research and Innovation**

- Statistics is fundamental in scientific research for designing experiments, analyzing data, and validating hypotheses, driving innovation and new discoveries.

🔧 **Resource Optimization**

- It helps in the efficient allocation and optimization of resources by analyzing data related to usage, demand, and other factors.

🔧 **Risk Assessment**

- Statistics is used to evaluate and manage risks in various sectors, such as finance, insurance, and public health, by quantifying uncertainties and modeling potential outcomes.

🔧 **Social and Economic Understanding:**

- It aids in understanding social and economic phenomena by analyzing data on demographics, economic indicators, and social behaviors, contributing to better policy-making and societal development.

🔧 **Career Opportunities**

- Proficiency in statistics opens up diverse career opportunities in fields such as data science, market research, finance, healthcare, and many more that rely on data-driven insights.

Q2.What are the two main types of statistics ?

Ans.🚀 There are 2 types of statistics:

- Descriptive Statistics
- Inferential Statistics
- 🧩 **Descriptive Statistics**:- Descriptive statistics involve organizing, summarizing, and presenting data in a meaningful way, allowing us to gain insights into the data's main characteristics.
- 🧩 **Inferential Statistics** :- Inferential statistics involve drawing conclusions and making predictions about a population based on a sample of data.

Q3.What are descriptive statistics ?

- Ans:- 🐍 Statistics is the foundation of data science. Descriptive statistics are simple tools that help us understand and summarize data.

🛠️ Types of Descriptive Statistics:-

- There are three categories for standard classification of descriptive statistics methods, each serving different purposes in summarizing and describing data.

1.Measures of Central Tendency – These describe the center of a dataset:

- Mean: The average of all values.
- Median: The middle value when data are ordered.
- Mode: The most frequently occurring value.

2.Measures of Dispersion (Variability) – These describe the spread of the data:

- Range: Difference between the highest and lowest values.
- Variance: The average of the squared differences from the mean.
- Standard Deviation: The square root of the variance.
- Interquartile Range (IQR): The range of the middle 50% of the data.

3.Measures of Position – These indicate the relative standing of data points:

- Percentiles: Indicate the value below which a given percentage of observations fall.
- Quartiles: Divide data into four equal parts.

4.Frequency Distributions – These show how often each value occurs:

- Tables (like frequency tables)
- Graphs (like histograms, pie charts, or bar charts)

Q4. What is inferential statistics ?

- Ans:- 🐍 Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data.

🛠️ Types of Inferential Statistics

- Hypothesis Testing
    - Z Test
    - T Test:
    - F Test
    - Confidence Interval:

Q5.What is sampling in statistics ?

- Ans:- In statistics, sampling is the process of selecting a subset of individuals or items (called a sample) from a larger group (called a population) in order to make inferences or draw conclusions about the whole population.

Q6.What are the different types of sampling methods ?

- Ans:- 📦 Types of Sampling Method

🔧 The two different types of sampling methods are:

- Probability Sampling
- Non-probability Sampling
- Every member of the population has a known, non-zero chance of being selected.

1.Simple Random Sampling

- Every member has an equal chance of being selected.
- Example: Drawing names from a hat.

2.Systematic Sampling

- Select every k-th member from a list after a random start.
- Example: Picking every 10th customer entering a store.

3.Stratified Sampling

- Divide the population into subgroups (strata) based on shared characteristics, then randomly sample from each stratum.
- Example: Sampling students by grade level.

4.Cluster Sampling

- Divide the population into clusters (often geographically), randomly select entire clusters, and sample all or some units within them.
- Example: Selecting random schools, then surveying all students in those schools.

5.Multistage Sampling

- Combines several sampling methods (e.g., cluster sampling followed by stratified or random sampling within clusters).
- Example: Randomly selecting states → then cities → then households.

🔸 **Non-Probability Sampling**

- Not all members have a known or equal chance of being included.

1.Convenience Sampling

- Sample those easiest to reach or access.
- Example: Surveying people at a nearby mall.

2.Judgmental or Purposive Sampling

- Researcher selects individuals based on their knowledge or purpose of the study.
- Example: Choosing experts for an expert panel.

3.Snowball Sampling

- Existing participants refer others, often used for hard-to-reach populations.
- Example: Studying underground music artists or rare disease patients.

4.Quota Sampling

- Ensure representation of certain characteristics, but selection is non-random within groups.
- Example: Selecting a set number of males and females without randomization.

Q7.What is the difference between random and non-random sampling ?

- Ans:- 🔹 **Random Sampling:**

📝 Definition: Every member of the population has an equal chance of being selected.

🧰 Key Features:

- Selection is unbiased and based on chance.
- Uses methods like lottery systems or random number generators.
- Tends to produce representative samples, which helps generalize findings to the broader population.

🌐 Examples:

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling (if the start is random)

🔹 **Non-Random Sampling:**

📝 Definition: Not every member has a known or equal chance of being selected.

🧰 Key Features:

- Selection is often based on convenience, judgment, or quotas.
- Introduces the risk of bias.
- May not accurately represent the larger population.

🌐 Examples:

- Convenience Sampling
- Judgmental (Purposive) Sampling
- Snowball Sampling
- Quota Sampling

Q8.Define and give examples of qualitative and quantitative data ?

- Ans:- Definitions:-

📍 Qualitative Data

Qualitative data refers to non-numerical information that describes characteristics or qualities. It is often used to capture subjective attributes such as colors, labels, opinions, or descriptions.

- Key features: Descriptive, categorical, not measured in numbers.

- 🛣️ Examples:

🧠 Example Description:-

- Hair color of students in a class
- Customer satisfaction feedback
- Types of cuisine in a food court

📍 Quantitative Data

Quantitative data refers to numerical information that can be measured or counted. It is used to express quantities, amounts, or ranges.

- Key features: Measurable, countable, statistical.
- 🛣️ Examples:
- 🧠 Example Description:-
- Number of students in a class
- Height of students
- Temperature readings throughout the day

Q9. What are the different types of data in statistics ?

- Ans:- In statistics, data is typically categorized into different types based on its nature and level of measurement.

📚 **1. Qualitative (Categorical) Data**

- 💡 a. Nominal Data

  Is a type of data that consists of categories or names that cannot be ordered or ranked. Nominal data is often used to categorize observations into groups, and the groups are not comparable.

- 🔄 Examples:-
- Gender (Male or female),
- Race (White, Black, Asian),
- Religion (Hinuduism, Christianity, Islam, Judaism),
- Blood type (A, B, AB, O), etc.
- 💡 b. Ordinal Data

Is a type of data that consists of categories that can be ordered or ranked. However, the distance between categories is not necessarily equal.

- 🔄 Examples:-
- Education level (Elementary, Middle, High School, College),
- Job position (Manager, Supervisor, Employee), etc.

📚 2.Quantitative (Numerical) Data

- 🚨 a. Discrete Data

Discrete data type is a type of data in statistics that only uses Discrete Values or Single Values. These data types have values that can be easily counted as whole numbers.

- 🔄 Example
- Height of Students in a class
- Marks of the students in a class test
- Weight of different members of a family, etc.
- 🚨 b. Continuous Data

Continuous data is the type of quantitative data that represents the data in a continuous range. The variable in the data set can have any value within the range of the data set.

- 🔄 Examples
- Temperature Range
- Salary range of Workers in a Factory, etc.

Q10.Explain nominal, ordinal, interval, and ratio levels of measurement ?

- 🧩 1. Nominal Level :- A nominal scale is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects.
- 1️⃣ Characteristics
- A nominal scale variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
- It is qualitative. The numbers are used here to identify the objects.
- The numbers don’t define the object characteristics. The only permissible aspect of numbers in the nominal scale is “counting.”

🧪 Example:

- Gender (male, female, nonbinary)

- Hair color (black, brown, blonde)
- Types of fruit (apple, banana, orange)
- 🧩2. Ordinal Level:- The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them.
- 1️⃣ Characteristics:
- The ordinal scale shows the relative ranking of the variables
- It identifies and describes the magnitude of a variable
- Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables
- The interval properties are not known
- The surveyors can quickly analyse the degree of agreement concerning the identified order of variables

  🧪 Example:
- Ranking of school students – 1st, 2nd, 3rd, etc.
- Ratings in restaurants
- Evaluating the frequency of occurrences
    - Very often
    - Often
    - Not often
    - Not at all
- Assessing the degree of agreement
    - Totally agree
    - Agree
    - Neutral
    - Disagree
    - Totally disagree
- 🧩3. Interval Level
- The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful.
- 1️⃣ Characteristics
- The interval scale is quantitative as it can quantify the difference between the values
- It allows calculating the mean and median of the variables

- To understand the difference between the variables, you can subtract the values between the variables
- The interval scale is the preferred scale in Statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.
- 🧪 Example:
- Likert Scale
- Net Promoter Score (NPS)
- Bipolar Matrix Table
- 🧩4.Ratio Level

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale.

- 1️⃣ Characteristics
- Ratio scale has a feature of absolute zero
- It doesn't have negative numbers, because of its zero-point feature
- It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.
- Ratio scale has unique and useful properties. One such feature is that it allows unit conversions like kilogram – calories, gram – calories, etc.

  🧪 Example:

  - What is your weight in Kgs?
- Less than 55 kgs
- 55 – 75 kgs
- 76 – 85 kgs
- 86 – 95 kgs

Q11.What is the measure of central tendency ?

- Ans:- The measure of central tendency is a statistical concept that identifies a single value as representative of an entire dataset. Its purpose is to describe the center point or typical value of a dataset.

✅ Measures of Central Tendency

- The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.

Q12.Define mean, median, and mode ?

1.Mean (Arithmetic Average):

- The sum of all values divided by the number of values.
- Example: For the numbers 2, 4, 6, the mean is (2 + 4 + 6) / 3 = 4
    - Geometric Mean
    - Harmonic Mean
    - Weighted Mean.

2.Median:

- The middle value when the numbers are arranged in order.
- If the number of observations is odd, it's the middle number.
- If even, it's the average of the two middle numbers.
- Example: For 3, 5, 7, the median is 5. For 2, 4, 6, 8, the median is (4 + 6)/2 = 5.

3.Mode:

- The value that occurs most frequently in a dataset.
- A dataset may have one mode (unimodal), more than one mode (bimodal, multimodal), or no mode.
- Example: In 1, 2, 2, 3, 4, the mode is 2

Q13.What is the significance of the measure of central tendency ?

1.Simplifies Data Interpretation:

- It reduces complex data sets into a single value, giving a quick sense of what a “typical” data point looks like.

2.Comparison Across Groups:

- It allows comparison between different data sets. For example, comparing average test scores between two classes.

3.Supports Decision-Making:

- In business, economics, health, and other fields, knowing the average (like average income, average wait time, etc.) informs policies and decisions.

4.Foundation for Further Analysis:

- Measures like the mean, median, and mode are often prerequisites for other statistical analyses such as standard deviation, regression, or hypothesis testing.

Q14.What is variance, and how is it calculated ?

- Ans:- Variance is a number that tells us how spread out the values in a data set are from the mean (average). It shows whether the numbers are close to the average or far away from it.
- 🛠️ How to Calculate Variance ?
- Step 1: Calculate the mean of the observation using the formula (Mean = Sum of Observations/Number of Observations)
- Step 2: Calculate the squared differences of the data values from the mean. (Data Value - Mean)2
- Step 3: Calculate the average of the squared differences of the given values, which is called the variance of the data set. (Variance = Sum of Squared Differences / Number of Observations)

Double-click (or enter) to edit

Double-click (or enter) to edit

Q15.What is standard deviation, and why is it important ?

- Ans:- Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a set of values. In simple terms, it tells you how spread out the numbers in a data set are around the mean (average).
- Standard Deviation Formula
- Standard Deviation $\sigma = \sqrt{(\sum(x_i - \mu)^2 / N)}$

🎯 Where:

- *x*i = each data point
- μ = mean of the data
- N = number of data points
- σ = standard deviation

🗂️ **Why is it important?**

- 🔁 **1.Measures variability**: It shows how much the data varies from the average.

- 🔁 **2.Interpreting risk**: In finance, a high standard deviation means more risk or volatility in investment returns.
- 🔁 **3.Quality control**: In manufacturing, it helps detect how consistent a process is.
- 🔁 **4.Statistical inference**: Many statistical techniques (like confidence intervals and hypothesis tests) use standard deviation to assess reliability and significance.
- 🔁 **5.Normal distribution context: In a normal distribution**:
- ~68% of data lies within 1 standard deviation of the mean
- ~95% within 2 standard deviations
- ~99.7% within 3 standard deviations

Q16.Define and explain the term range in statistics ?

- Ans:- In statistics, a range refers to the difference between the highest and lowest values in a dataset. It provides a simple measure of the spread or dispersion of the data.
- Range = Maximum Value − Minimum Value
- 🔁 Explanation:
- The maximum value is the largest number in the dataset.
- The minimum value is the smallest number in the dataset.
- The range is simply the difference between these two values.

✅Example:

5, 8, 10, 12, 15

- Maximum = 15
- Minimum = 5
- Range = 15 − 5 = 10

So, the range is 10.

Q17.What is the difference between variance and standard deviation ?

- Ans:- The variance and standard deviation are both measures of how spread out the values in a data set are, but they differ in how they express that spread:
- 📦 1.Variance
- Definition: Variance is the average of the squared differences from the mean.
- 🧩 Formula (for a population):

$$\sigma^2 = \Sigma(x_i - \mu)^2 / N$$

- 📦 2.Standard Deviation
- Definition: Standard deviation is the square root of the variance. It provides a measure of spread in the same units as the original data.
- Formula:

$$\sigma = \sqrt{(\sum(x_i - \mu)^2 / N)}$$

- Units: Same as the original data (e.g., meters if the data is in meters).
- 📥 Key Difference
- Interpretability: Standard deviation is generally more interpretable because it's in the same unit as the data.
- Mathematics: Variance is useful in mathematical computations, especially in statistical theory and inferential statistics, due to its properties (like additivity for independent variables).

Q18 What is skewness in a dataset ?

- Ans:- The skewness in statistics is a measure of asymmetry or the deviation of a given random variable's distribution from a symmetric distribution.
- 📄 Skewness in statistics can be divided into two categories. They are:
- Positive Skewness
- Negative Skewness

Q19.What does it mean if a dataset is positively or negatively skewed ?

- Ans:- If a dataset is positively skewed or negatively skewed, it refers to the asymmetry in the distribution of the data values.
- 🔹 Positively Skewed (Right-Skewed)
- Tail direction: The right tail (higher values) is longer.
- Distribution shape: Most data values are concentrated on the left (lower values), with a few extreme values on the right.
- Mean > Median > Mode: The mean is pulled in the direction of the tail (to the right).

Example: Income data — most people earn modest amounts, but a few earn extremely high salaries.

📉 Skew goes toward the positive (right) direction.

🔹 Negatively Skewed (Left-Skewed)

- Tail direction: The left tail (lower values) is longer.
- Distribution shape: Most data values are concentrated on the right (higher values), with a few extreme values on the left.
- Mean < Median < Mode: The mean is pulled in the direction of the tail (to the left).

Example: Test scores where most students score high, but a few score very low.

📉 Skew goes toward the negative (left) direction.

Q20.Define and explain kurtosis ?

- Ans:- Kurtosis is a statistical measure that describes the shape of a distribution's tails in relation to its overall shape.
- 🔍 Definition:

Kurtosis measures the "tailedness" of the probability distribution of a real-valued random variable.

📊 Types of Kurtosis:

1.Mesokurtic (Kurtosis = 3)

- This is the kurtosis of a normal distribution.
- Tails are moderate; neither too fat nor too thin.

2.Leptokurtic (Kurtosis > 3):

- Heavy tails and a sharp peak.
- More data in the tails and central peak → more outliers.
- Example: t-distribution with small degrees of freedom.

3.Platykurtic (Kurtosis < 3):

- Light tails and a flatter peak.
- Fewer outliers.
- Example: Uniform distribution.

🧮 **Formula (Excess Kurtosis):**

To make comparison easier, we often use excess kurtosis, which subtracts 3 from the regular kurtosis so the normal distribution has a kurtosis of 0.

```
Excess Kurtosis = μ4 / σ4 - 3
```

Q21.What is the purpose of covariance ?

- Ans:- Covariance is a statistical measure used to assess the direction of the linear relationship between two random variables.
- 🧰 Purpose of Covariance:

▶️ 1.Measure Direction of Relationship:

- Positive covariance: As one variable increases, the other tends to increase.
- Negative covariance: As one variable increases, the other tends to decrease.
- Zero covariance: No linear relationship.

▶️ 2.Foundation for Correlation:

- Covariance is a building block for correlation, which standardizes the relationship so it's not influenced by scale.
- Correlation = Covariance divided by the product of standard deviations.

▶️ 3.Used in Portfolio Theory (Finance):

- Helps in portfolio diversification. Covariance between asset returns tells whether they tend to rise or fall together.
- Low or negative covariance between assets can reduce overall risk.

▶️ 4.Statistical Modeling and Machine Learning:

- Covariance matrices are used in algorithms like Principal Component Analysis (PCA) and Gaussian processes.
- In multivariate statistics, covariance helps model relationships between variables.

Example:

Suppose you track temperature and ice cream sales. If higher temperatures usually lead to more ice cream being sold, their covariance will be positive.

Q2.What does correlation measure in statistics ?

- Ans:- In statistics, Correlation studies and measures the direction and extent of relationship among variables, so the correlation measures co-variation, not causation.

1.Range: Correlation values range from -1 to +1.

- +1: Perfect positive linear relationship (as one variable increases, the other increases).
- -1: Perfect negative linear relationship (as one variable increases, the other decreases).
- 0: No linear relationship between the variables.

2.Common Measure: The most common type is the Pearson correlation coefficient (denoted as r), which assesses how well the relationship between two variables can be described by a straight line.

Q23.What is the difference between covariance and correlation ?

- Ans:- Covariance and correlation are both statistical measures used to describe the relationship between two variables, but they differ in scale, interpretation, and how they are computed.

1.Covariance

- **Definition**: Covariance measures the direction of the linear relationship between two variables.
- **Range:**
- Can be any real number (positive, negative, or zero).
- **Interpretation:**
- Positive covariance: As one variable increases, the other tends to increase.
- Negative covariance: As one increases, the other tends to decrease.
- Zero covariance: No linear relationship.
- **2.Correlation**
- **Definition**: Correlation is a scaled version of covariance that measures both the direction and strength of a linear relationship.
- **Range:**
- Always between -1 and +1.
- Interpretation
- r=1: Perfect positive linear relationship.
- r=−1: Perfect negative linear relationship.
- r=0: No linear relationship.

Q24.What are some real-world applications of statistics ?

- Ans:- Statistics has a vast range of real-world applications across virtually every field.

1.🧪Healthcare and Medicine

- Clinical Trials: Testing the effectiveness of new drugs and treatments.
- Epidemiology: Tracking disease outbreaks and public health trends (e.g., COVID-19 modeling).
- Medical Diagnostics: Developing predictive models for disease based on patient data.

2.🧪Business and Economics

- Market Research: Understanding consumer behavior through surveys and product feedback.
- Quality Control: Monitoring production processes to maintain product quality (Six Sigma).
- Financial Forecasting: Predicting stock market trends, investment risks, and economic indicators.

3.🧪Government and Public Policy

- Census Data Analysis: Allocating funding, drawing electoral districts, planning infrastructure.
- Policy Evaluation: Measuring the impact of programs (e.g., education reform, tax policy).
- Crime Statistics: Understanding crime patterns and allocating law enforcement resources.

🧪4.Education

- Standardized Testing: Designing and analyzing exams like the SAT or GRE.
- Performance Metrics: Assessing school and teacher effectiveness.
- Learning Analytics: Customizing education using student data.

🧪6.Technology and Artificial Intelligence

- Machine Learning: Training models using statistical techniques.
- Data Mining: Finding patterns in large datasets.
- A/B Testing: Optimizing software features or website layouts based on user behavior.

Practical Questions

```
# 1. How do you calculate the mean, median, and mode of a dataset ?
dataset = [1, 2, 3, 4, 5, 5, 6, 7, 8, 9]
import statistics
mean = statistics.mean(dataset)
median = statistics.median(dataset)
mode = statistics.mode(dataset)
print(f"Mean: {mean}")
```

```
print(f"Median: {median}")
print(f"Mode: {mode}")
```

```
Mean: 5
Median: 5.0
Mode: 5
```

```
# 2.Write a Python program to compute the variance and standard deviation of a datas

import math

def calculate_variance(data):
    if len(data) == 0:
        return 0
    mean = sum(data) / len(data)
    squared_diff = [(x - mean) ** 2 for x in data]
    variance = sum(squared_diff) / len(data)  # Use (len(data)-1) for sample varianc
    return variance

def calculate_standard_deviation(data):
    variance = calculate_variance(data)
    standard_deviation = math.sqrt(variance)
    return standard_deviation

# Example usage
dataset = [10, 12, 23, 23, 16, 23, 21, 16]
variance = calculate_variance(dataset)
std_deviation = calculate_standard_deviation(dataset)

print("Dataset:", dataset)
print("Variance:", variance)
print("Standard Deviation:", std_deviation)
```

```
Dataset: [10, 12, 23, 23, 16, 23, 21, 16]
Variance: 24.0
Standard Deviation: 4.898979485566356
```

```
# 3.  Create a dataset and classify it into nominal, ordinal, interval, and ratio ty
dataset = {
    "nominal": ["red", "blue", "green"],
    "ordinal": ["low", "medium", "high"],
    "interval": [1, 2, 3, 4, 5],
    "ratio": [10, 20, 30, 40]
}
print("Dataset Classification:")
for key, value in dataset.items():
    print(f"{key.capitalize()}: {value}")
```

```
Dataset Classification:
Nominal: ['red', 'blue', 'green']
Ordinal: ['low', 'medium', 'high']
```

```
    Interval: [1, 2, 3, 4, 5]
    Ratio: [10, 20, 30, 40]
```

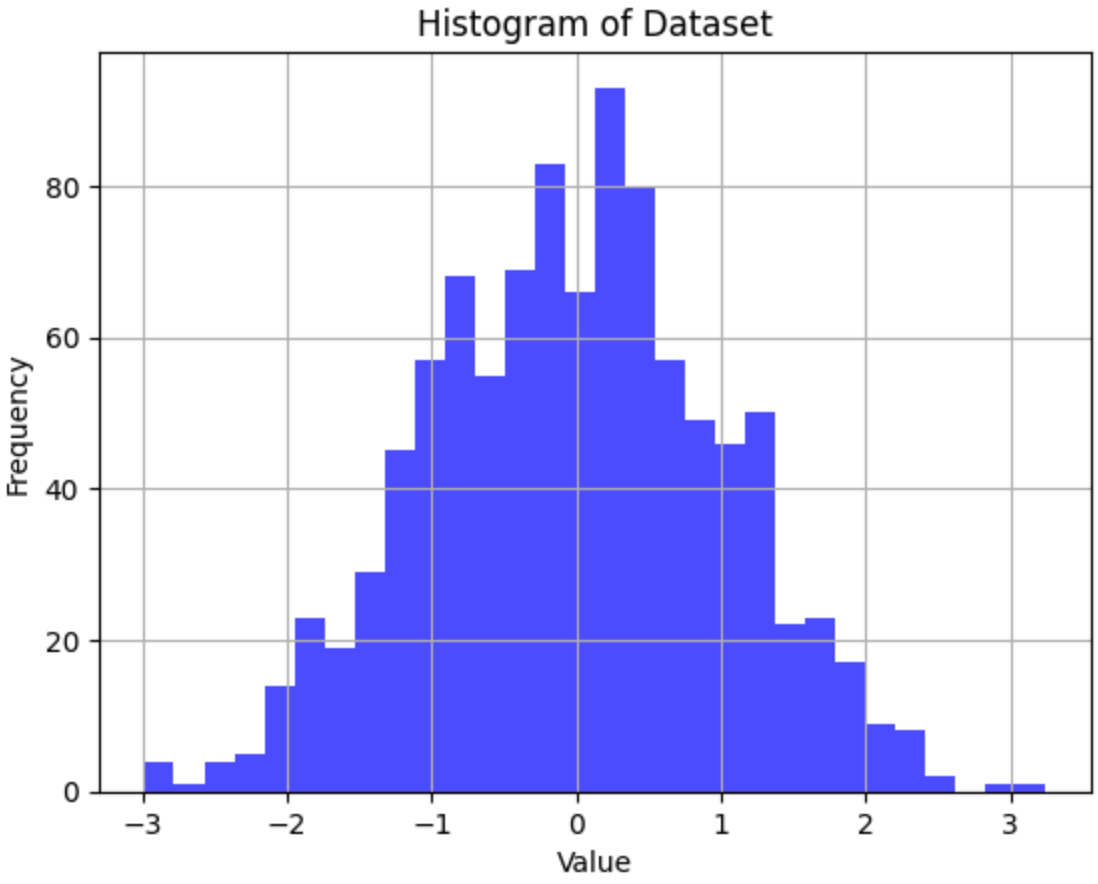
```
# 4. Implement sampling techniques like random sampling and stratified sampling.
import random
def random_sampling(data, sample_size):
    return random.sample(data, sample_size)
def stratified_sampling(data, strata, sample_size):
    stratified_samples = []
    for stratum in strata:
        stratum_data = [item for item in data if item['stratum'] == stratum]
        stratified_samples.extend(random.sample(stratum_data, sample_size))
    return stratified_samples
data = [
    {'value': 1, 'stratum': 'A'},
    {'value': 2, 'stratum': 'A'},
    {'value': 3, 'stratum': 'B'},
    {'value': 4, 'stratum': 'B'},
    {'value': 5, 'stratum': 'C'},
    {'value': 6, 'stratum': 'C'},
]
sample_size = 1
random_sample = random_sampling(data, sample_size)
strata = ['A', 'B', 'C']
stratified_sample = stratified_sampling(data, strata, sample_size)
print(f"Random Sample: {random_sample}")
print(f"Stratified Sample: {stratified_sample}")
```

```
Random Sample: [{'value': 3, 'stratum': 'B'}]
Stratified Sample: [{'value': 1, 'stratum': 'A'}, {'value': 3, 'stratum': 'B'},
```

```
# 5.Write a Python function to calculate the range of a dataset ?
def calculate_range(data):
    return max(data) - min(data)
data = [1, 2, 3, 4, 5]
data_range = calculate_range(data)
print(f"Range of dataset: {data_range}")
```

```
Range of dataset: 4
```

```
# 6. Create a dataset and plot its histogram to visualize skewness.
import matplotlib.pyplot as plt
import numpy as np
data = np.random.normal(0, 1, 1000)
plt.hist(data, bins=30, alpha=0.7, color='blue')
plt.title('Histogram of Dataset')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

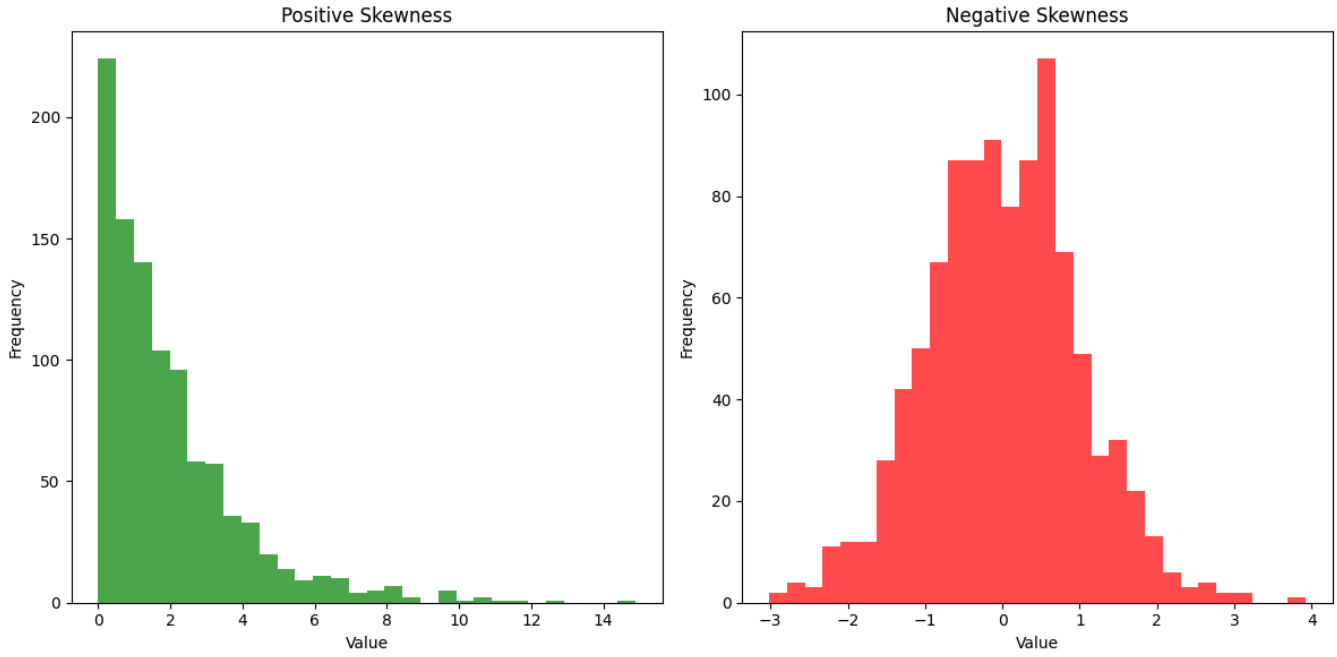



```
# 7.Calculate skewness and kurtosis of a dataset using Python libraries ?

# ✅ Using scipy.stats and pandas

import pandas as pd
from scipy.stats import skew, kurtosis

# Example dataset
data = [10, 12, 23, 23, 16, 23, 21, 16]

# Convert to pandas Series
series = pd.Series(data)

# Calculate skewness
data_skewness = skew(series)
```

```
# Calculate kurtosis (Fisher’s definition by default; normal ==> 0)
data_kurtosis = kurtosis(series)

print(f"Skewness: {data_skewness}")
print(f"Kurtosis: {data_kurtosis}")
```

```
Skewness: -0.363596133694378
Kurtosis: -1.3984375
```

```
# 8. Generate a dataset and demonstrate positive and negative skewness.
data_positive_skew = np.random.exponential(scale=2, size=1000)
data_negative_skew = np.random.normal(loc=0, scale=1, size=1000)
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.hist(data_positive_skew, bins=30, alpha=0.7, color='green')
plt.title('Positive Skewness')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.subplot(1, 2, 2)
plt.hist(data_negative_skew, bins=30, alpha=0.7, color='red')
plt.title('Negative Skewness')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```




```
# 9.Write a Python script to calculate covariance between two datasets ?

# ✅ Version 1: Pure Python

def calculate_covariance(x, y):
    if len(x) != len(y):
        raise ValueError("Datasets x and y must have the same length.")

    n = len(x)
    mean_x = sum(x) / n
    mean_y = sum(y) / n

    covariance = sum((x[i] - mean_x) * (y[i] - mean_y) for i in range(n)) / n
    return covariance

# Example usage
x = [2.1, 2.5, 4.0, 3.6]
y = [8, 10, 12, 14]

cov = calculate_covariance(x, y)
```

```
print("Covariance:", cov)

# ✅ Version 2: Using NumPy

import numpy as np

# Example datasets
x = np.array([2.1, 2.5, 4.0, 3.6])
y = np.array([8, 10, 12, 14])

# Use np.cov with bias=True to match the pure Python result
cov_matrix = np.cov(x, y, bias=True)
cov = cov_matrix[0][1]

print("Covariance (NumPy):", cov)
```

```
Covariance: 1.5
Covariance (NumPy): 1.5
```

```
# 9.Write a Python script to calculate covariance between two datasets ?

import numpy as np

def calculate_covariance(x, y):
    if len(x) != len(y):
        raise ValueError("Datasets x and y must have the same number of elements.")

    mean_x = np.mean(x)
    mean_y = np.mean(y)

    # Calculate covariance
    covariance = np.sum((x - mean_x) * (y - mean_y)) / (len(x) - 1)
    return covariance

# Example usage
x = np.array([2, 4, 6, 8])
y = np.array([1, 3, 5, 7])

cov = calculate_covariance(x, y)
print("Covariance:", cov)
```

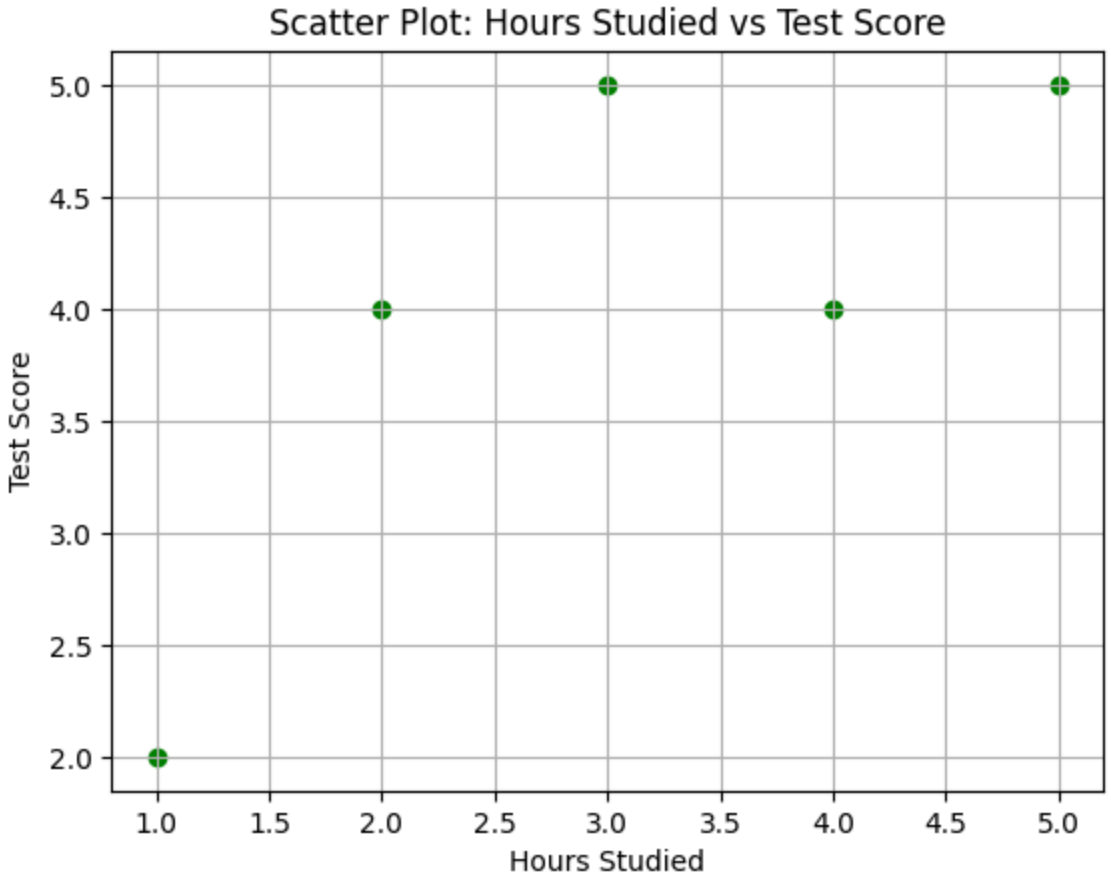
```
Covariance: 6.666666666666667
```

```
# 10.Write a Python script to calculate the correlation coefficient between two data

import numpy as np

# Sample datasets
x = [10, 20, 30, 40, 50]
```

```
y = [15, 25, 35, 45, 55]

# Manual calculation of Pearson correlation coefficient
def calculate_correlation(x, y):
    if len(x) != len(y):
        raise ValueError("Datasets x and y must be of the same length.")

    n = len(x)
    mean_x = sum(x) / n
    mean_y = sum(y) / n

    numerator = sum((xi - mean_x) * (yi - mean_y) for xi, yi in zip(x, y))
    denominator = (sum((xi - mean_x)**2 for xi in x) * sum((yi - mean_y)**2 for yi i

    if denominator == 0:
        raise ValueError("Denominator is zero. Correlation coefficient is undefined.

    return numerator / denominator

# Using NumPy for verification
def correlation_with_numpy(x, y):
    return np.corrcoef(x, y)[0, 1]

# Calculate and print results
manual_corr = calculate_correlation(x, y)
numpy_corr = correlation_with_numpy(x, y)

print(f"Manual Pearson Correlation Coefficient: {manual_corr:.4f}")
print(f"NumPy Pearson Correlation Coefficient: {numpy_corr:.4f}")
```

```
Manual Pearson Correlation Coefficient: 1.0000
NumPy Pearson Correlation Coefficient: 1.0000
```

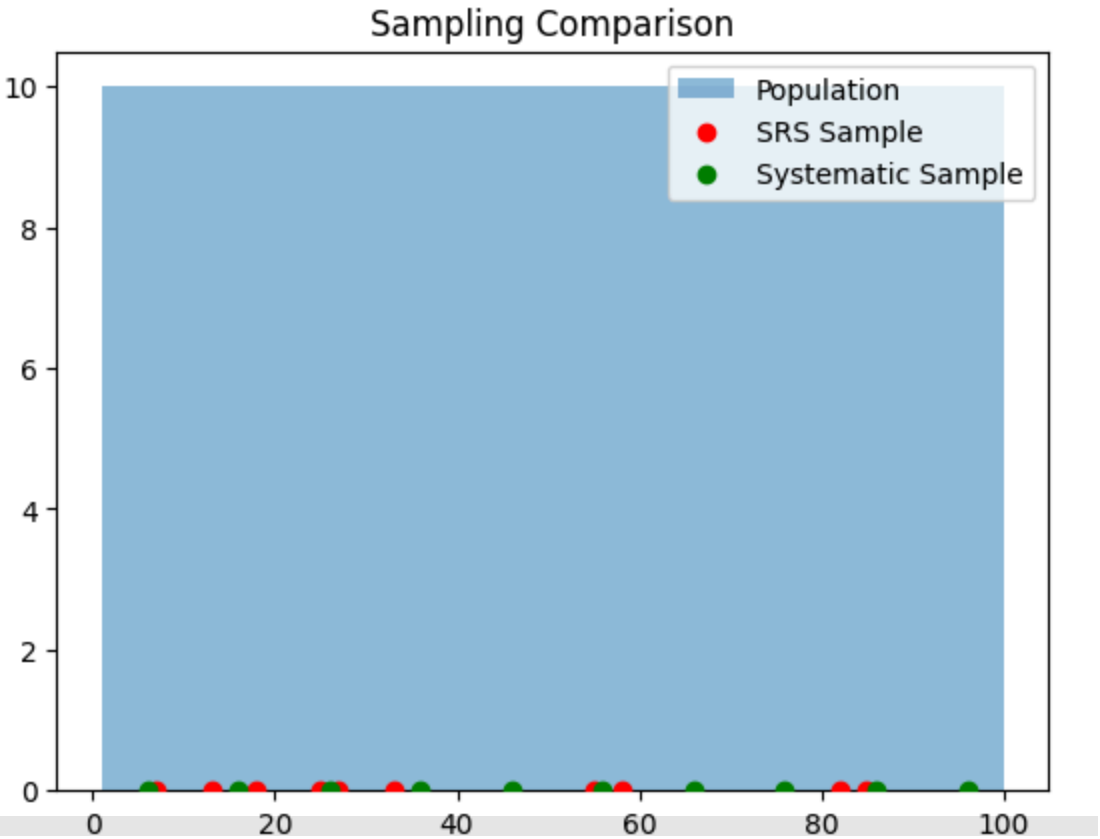
```
# 11.Create a scatter plot to visualize the relationship between two variables ?

import matplotlib.pyplot as plt

# Example data
x = [1, 2, 3, 4, 5]  # Independent variable
y = [2, 4, 5, 4, 5]  # Dependent variable

# Create scatter plot
plt.scatter(x, y, color='green', marker='o')
plt.title('Scatter Plot: Hours Studied vs Test Score')
plt.xlabel('Hours Studied')
plt.ylabel('Test Score')
plt.grid(True)
plt.show()
```




```
# Q12. Implement and compare simple random sampling and systematic sampling ?
import numpy as np
import matplotlib.pyplot as plt

# Generate a population
population = np.arange(1, 101)  # Population from 1 to 100
population_mean = np.mean(population)

sample_size = 10

# Simple Random Sampling
srs_sample = np.random.choice(population, size=sample_size, replace=False)
srs_mean = np.mean(srs_sample)

# Systematic Sampling
k = len(population) // sample_size  # Sampling interval
start = np.random.randint(0, k)     # Random start point
systematic_sample = population[start::k][:sample_size]
systematic_mean = np.mean(systematic_sample)

# Output results
print(f"Population Mean: {population_mean}")
print(f"SRS Sample: {srs_sample}, Mean: {srs_mean}")
print(f"Systematic Sample: {systematic_sample}, Mean: {systematic_mean}")
```

```
# Optional: Plotting
plt.hist(population, bins=10, alpha=0.5, label='Population')
plt.scatter(srs_sample, [0]*sample_size, color='red', label='SRS Sample', zorder=5)
plt.scatter(systematic_sample, [0]*sample_size, color='green', label='Systematic Sam
plt.legend()
plt.title("Sampling Comparison")
plt.show()
```

```
Population Mean: 50.5
SRS Sample: [25 85  7 55 58 27 18 82 33 13], Mean: 40.3
Systematic Sample: [ 6 16 26 36 46 56 66 76 86 96], Mean: 51.0
```



```
# 13. Calculate the mean, median, and mode of grouped data.
def grouped_data_statistics(data):
    total_frequency = sum(data.values())
    mean = sum(value * frequency for value, frequency in data.items()) / total_frequ
    median = None
    cumulative_frequency = 0
    for value, frequency in sorted(data.items()):
        cumulative_frequency += frequency
        if cumulative_frequency >= total_frequency / 2:
            median = value
            break
    mode = max(data, key=data.get)
    return mean, median, mode
grouped_data = {1: 2, 2: 3, 3: 5, 4: 4}
mean, median, mode = grouped_data_statistics(grouped_data)
print(f"Grouped Data - Mean: {mean}, Median: {median}, Mode: {mode}")
```

```
Grouped Data - Mean: 2.7857142857142856, Median: 3, Mode: 3
```

```
# 14.Simulate data using Python and calculate its central tendency and dispersion.

import numpy as np
import scipy.stats as stats
import pandas as pd

# 1. Simulate data: 1000 data points from a normal distribution
np.random.seed(42)  # for reproducibility
data = np.random.normal(loc=50, scale=10, size=1000)

# 2. Central Tendency
mean = np.mean(data)
median = np.median(data)
mode = stats.mode(data, keepdims=True).mode[0]

# 3. Dispersion
variance = np.var(data)
std_dev = np.std(data)
data_range = np.max(data) - np.min(data)
iqr = stats.iqr(data)

# 4. Display results
summary = pd.DataFrame({
    "Measure": ["Mean", "Median", "Mode", "Variance", "Std Dev", "Range", "IQR"],
    "Value": [mean, median, mode, variance, std_dev, data_range, iqr]
})

print(summary)
```

```
     Measure      Value
0       Mean  50.193321
1     Median  50.253006
```

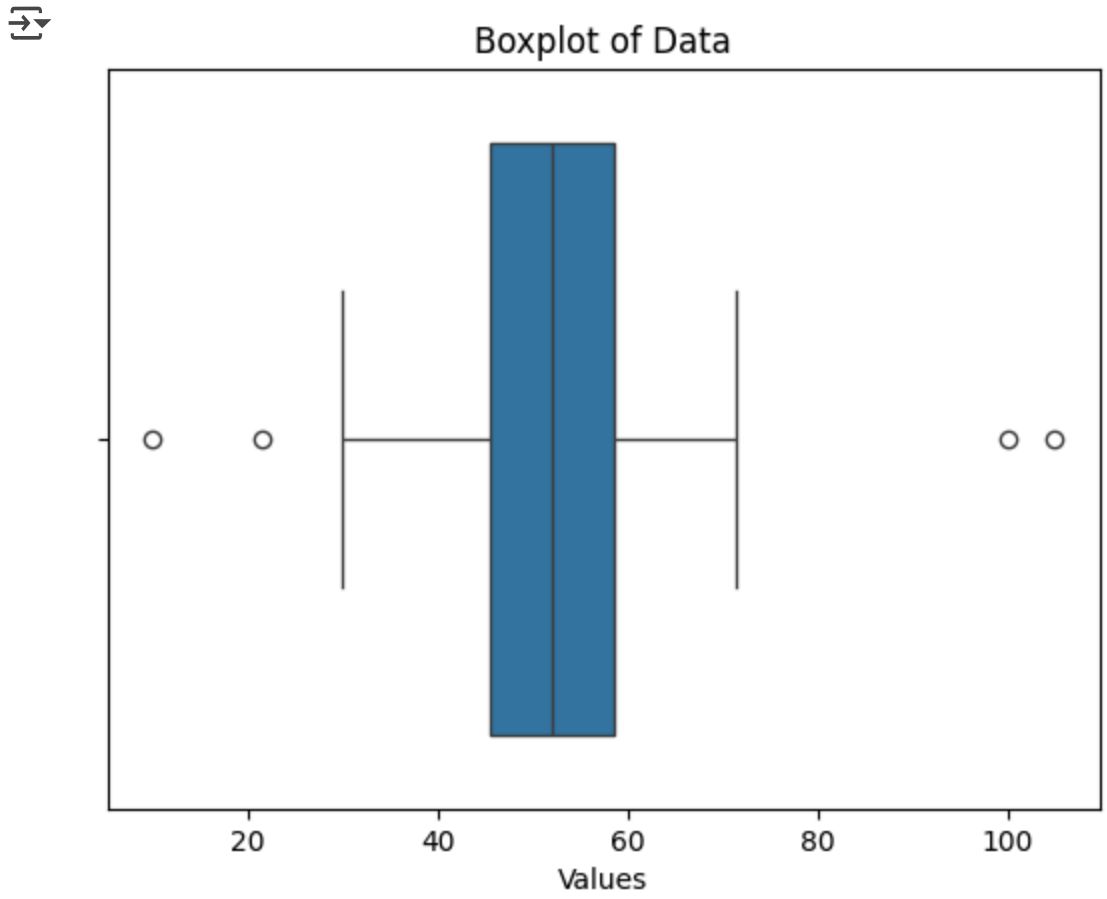
```
2      Mode  17.587327
3  Variance  95.790499
4   Std Dev   9.787262
5     Range  70.939988
6       IQR  12.955342
```

```
# 15.Use NumPy or pandas to summarize a dataset’s descriptive statistics ?

import pandas as pd

# Sample dataset
data = {
    'Age': [25, 32, 47, 51, 62],
    'Salary': [50000, 60000, 75000, 82000, 90000]
}

# Create a DataFrame
df = pd.DataFrame(data)

# Get descriptive statistics
summary = df.describe()
print(summary)

# 🧮 Using NumPy for descriptive statistics

import numpy as np

# Sample data
ages = np.array([25, 32, 47, 51, 62])
salaries = np.array([50000, 60000, 75000, 82000, 90000])

# Compute statistics
print("Age - Mean:", np.mean(ages), " Std:", np.std(ages), " Min:", np.min(ages), "
print("Salary - Mean:", np.mean(salaries), " Std:", np.std(salaries), " Min:", np.mi
```

```
             Age        Salary
count   5.000000      5.000000
mean   43.400000  71400.000000
std    14.876155  16272.676485
min    25.000000  50000.000000
25%    32.000000  60000.000000
50%    47.000000  75000.000000
75%    51.000000  82000.000000
max    62.000000  90000.000000
Age - Mean: 43.4  Std: 13.305637902783918  Min: 25  Max: 62
Salary - Mean: 71400.0  Std: 14554.724318928202  Min: 50000  Max: 90000
```

```
# 16 Plot a boxplot to understand the spread and identify outliers ?

import matplotlib.pyplot as plt
```

```
import seaborn as sns
import numpy as np

# Example data
data = np.random.normal(loc=50, scale=10, size=100)  # Normally distributed data

# Introduce some outliers
data = np.append(data, [10, 100, 105])

# Create boxplot
sns.boxplot(x=data)
plt.title('Boxplot of Data')
plt.xlabel('Values')
plt.show()
```




```
# 17. Calculate the interquartile range (IQR) of a dataset.
def calculate_iqr(data):
    q1 = np.percentile(data, 25)
    q3 = np.percentile(data, 75)
    iqr = q3 - q1
    return iqr
data = [1, 2, 3, 4, 5, 6, 7, 8, 9]
iqr = calculate_iqr(data)
print(f"Interquartile Range (IQR): {iqr}")
```

```
Interquartile Range (IQR): 4.0
```

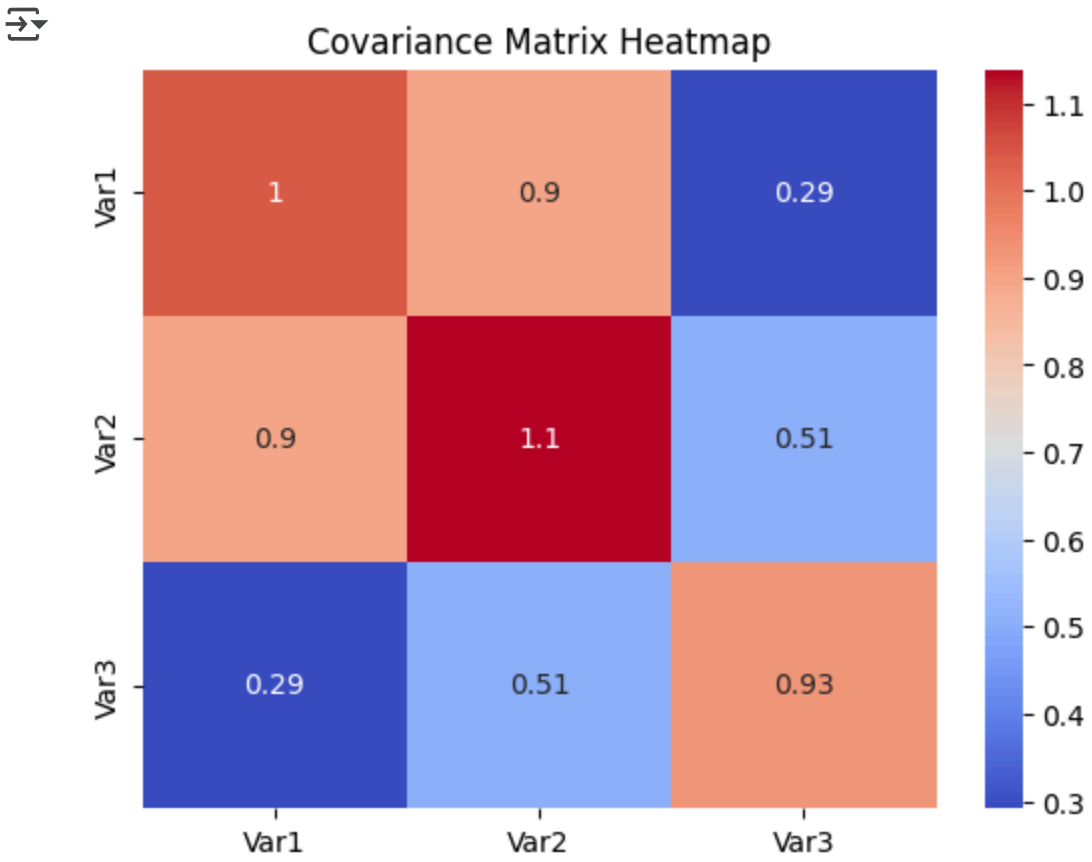
```
# 18. Implement Z-score normalization and explain its significance.
def z_score_normalization(data):
    mean = np.mean(data)
    std_dev = np.std(data)
    return [(x - mean) / std_dev for x in data]
data = [1, 2, 3, 4, 5]
normalized_data = z_score_normalization(data)
print(f"Z-score Normalized Data: {normalized_data}")
```

```
Z-score Normalized Data: [np.float64(-1.414213562373095), np.float64(-0.70710678
```

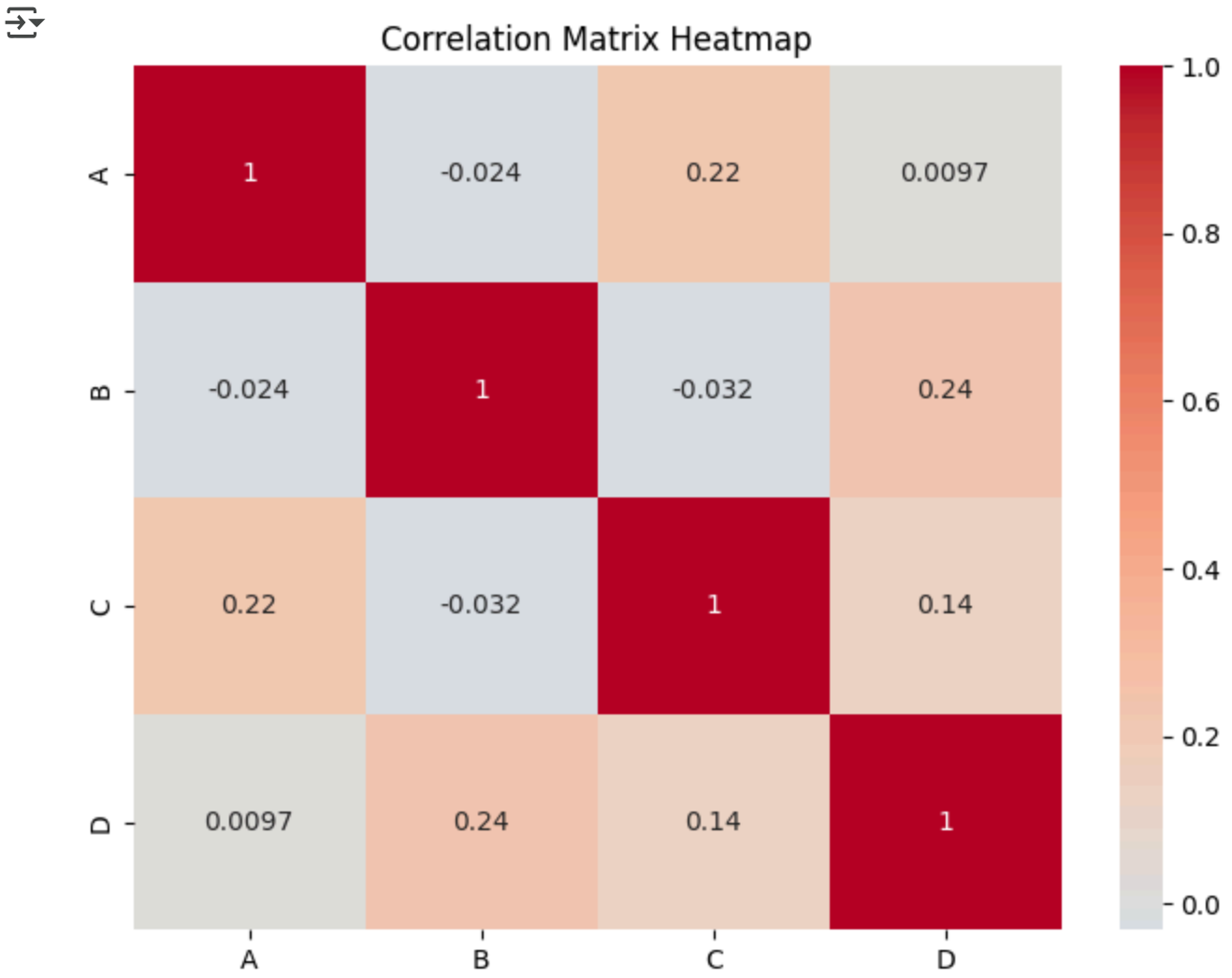
```
# 19. Compare two datasets using their standard deviations.
def compare_standard_deviations(data1, data2):
    std_dev1 = np.std(data1)
    std_dev2 = np.std(data2)
    return std_dev1, std_dev2
data1 = [1, 2, 3, 4, 5]
data2 = [5, 6, 7, 8, 9]
std_dev1, std_dev2 = compare_standard_deviations(data1, data2)
print(f"Standard Deviation of Data1: {std_dev1}")
print(f"Standard Deviation of Data2: {std_dev2}")
```

```
Standard Deviation of Data1: 1.4142135623730951
Standard Deviation of Data2: 1.4142135623730951
```

```
# 20 Write a Python program to visualize covariance using a heatmap ?

import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Generate some sample data (e.g., 3 variables with 100 observations)
np.random.seed(0)
data = np.random.multivariate_normal(
    mean=[0, 0, 0],
    cov=[[1, 0.8, 0.3], [0.8, 1, 0.5], [0.3, 0.5, 1]],
    size=100
)

# Compute the covariance matrix
cov_matrix = np.cov(data, rowvar=False)

# Visualize covariance matrix as heatmap
sns.heatmap(cov_matrix, annot=True, cmap='coolwarm', xticklabels=['Var1', 'Var2', 'V
plt.title('Covariance Matrix Heatmap')
plt.show()
```




```
#21 Use seaborn to create a correlation matrix for a dataset ?

import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Example dataset
data = pd.DataFrame({
    'A': np.random.rand(100),
    'B': np.random.rand(100),
    'C': np.random.rand(100),
    'D': np.random.rand(100),
})

# Calculate correlation matrix
corr_matrix = data.corr()

# Plot heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Matrix Heatmap')
plt.show()
```




```
# 22.Generate a dataset and implement both variance and standard deviation computati

import random
import math

# Generate a dataset of 20 random numbers between 1 and 100
dataset = [random.randint(1, 100) for _ in range(20)]

# Function to calculate variance
def variance(data):
    n = len(data)
    mean = sum(data) / n
    var = sum((x - mean) ** 2 for x in data) / n
    return var

# Function to calculate standard deviation
def standard_deviation(data):
    return math.sqrt(variance(data))

# Display results
print("Dataset:", dataset)
```

```
print("Variance:", variance(dataset))
print("Standard Deviation:", standard_deviation(dataset))
```

```
Dataset: [43, 37, 84, 38, 3, 44, 68, 82, 52, 44, 30, 1, 62, 83, 65, 70, 30, 85,
Variance: 685.6400000000001
Standard Deviation: 26.18472837361503
```

```
# 23. Visualize skewness and kurtosis using Python libraries like matplotlib or seab
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
data = np.random.normal(loc=0, scale=1, size=1000)
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.histplot(data, kde=True)
plt.title('Histogram with KDE')
plt.subplot(1, 2, 2)
sns.boxplot(data=data)
plt.title('Boxplot')
plt.tight_layout()
plt.show()
```


```
#24.Implement the Pearson and Spearman correlation coefficients for a dataset.

import numpy as np
from scipy.stats import rankdata

def pearson_correlation(x, y):
    """
    Compute Pearson correlation coefficient between two arrays.

    Args:
        x (array-like): First variable
        y (array-like): Second variable
```