Statistics Basics| Assignment

1. **What is the difference between descriptive statistics and inferential statistics? Explain with examples.**

- **Descriptive Statistics:**
  - **Summarizes and describes data collected from a sample or population.**
  - **Uses measures such as mean, median, mode, standard deviation, variance, and visualizations like histograms.**
  - **Example:**
    If you survey 100 students about their test scores and find the average score is 75, that's descriptive. It only describes the data you have.
- **Inferential Statistics:**
  - **Makes predictions or inferences about a larger population based on a sample of data.**
  - **Uses hypothesis testing, confidence intervals, regression analysis.**
  - **Example:**
    If you use the 100 students' scores to predict the average score of all students in the school, that's inferential.

Question 2: What is sampling in statistics? Explain the differences between random and stratified sampling.

Answer:

- **Sampling**

    - **The process of selecting a subset of individuals from a population to estimate characteristics of the entire population.**
    - **Used because analyzing an entire population is often impractical or costly.**
- **Random Sampling**
    - **Every individual in the population has an equal chance of being selected.**
    - **Example: Choosing 100 customers randomly from a database of 10,000.**
- **Stratified Sampling**
    - **The population is divided into subgroups (strata) based on specific characteristics (e.g., age, income) and random samples are taken from each group proportionally.**
    - **Example: From a company with 60% male and 40% female employees, a sample of 100 would include 60 males and 40 females, selected randomly from their respective groups.**

Question 3: Define mean, median, and mode. Explain why these measures of central tendency are important.

Answer:

- **Mean: Arithmetic average of all data points.**
  Example: [10, 20, 30] → Mean = (10+20+30) / 3 = 20
- **Median: Middle value when data is sorted in ascending or descending order.**
  Example: [10, 20, 30, 40, 50] → Median = 30

- **Mode: Value that occurs most frequently.**
  Example: [5, 10, 10, 20] → Mode = 10

- **Importance:**

  - **Summarize large datasets with a single value.**

  - **Help understand data trends and patterns.**

  - **Useful for comparing datasets and identifying outliers.**

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?

Answer:

- **Skewness**

  - **Measures asymmetry of the distribution of data.**

  - **Positive Skew: Tail of distribution extends to the right; mean > median.**
    Example: Income data where few people earn extremely high salaries.

- **Kurtosis**

  - **Measures how heavy or light the tails of a distribution are compared to a normal distribution.**

  - **High kurtosis → More outliers and sharp peak; Low kurtosis → Flatter curve with fewer outliers.**

- **Positive Skew: Indicates majority of values are concentrated on the lower end with a few high values pulling the mean upward.**

Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

Answer-

```python
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]


import statistics as stats


mean= stats.mean(numbers)

median= stats.median(numbers)

mode= stats.mode(numbers)


print(f'Mean is {mean}')

print(f'Median is {median}')

print(f'Mode is {mode}')
```

Output :

```
Mean is 19.6

Median is 19

Mode is 12
```

:

```python
#Question 6: Covariance and Correlation


import numpy as np
```

```
list_x = [10, 20, 30, 40, 50]

list_y = [15, 25, 35, 45, 60]


cov_matrix = np.cov(list_x, list_y, bias=False)

covariance = cov_matrix[0][1]


correlation = np.corrcoef(list_x, list_y)[0][1]


print(f"Covariance: {covariance}")

print(f"Correlation Coefficient: {correlation}")
```

```
[27]
                                   0s
Covariance: 275.0

Correlation Coefficient: 0.995893206467704
```

Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

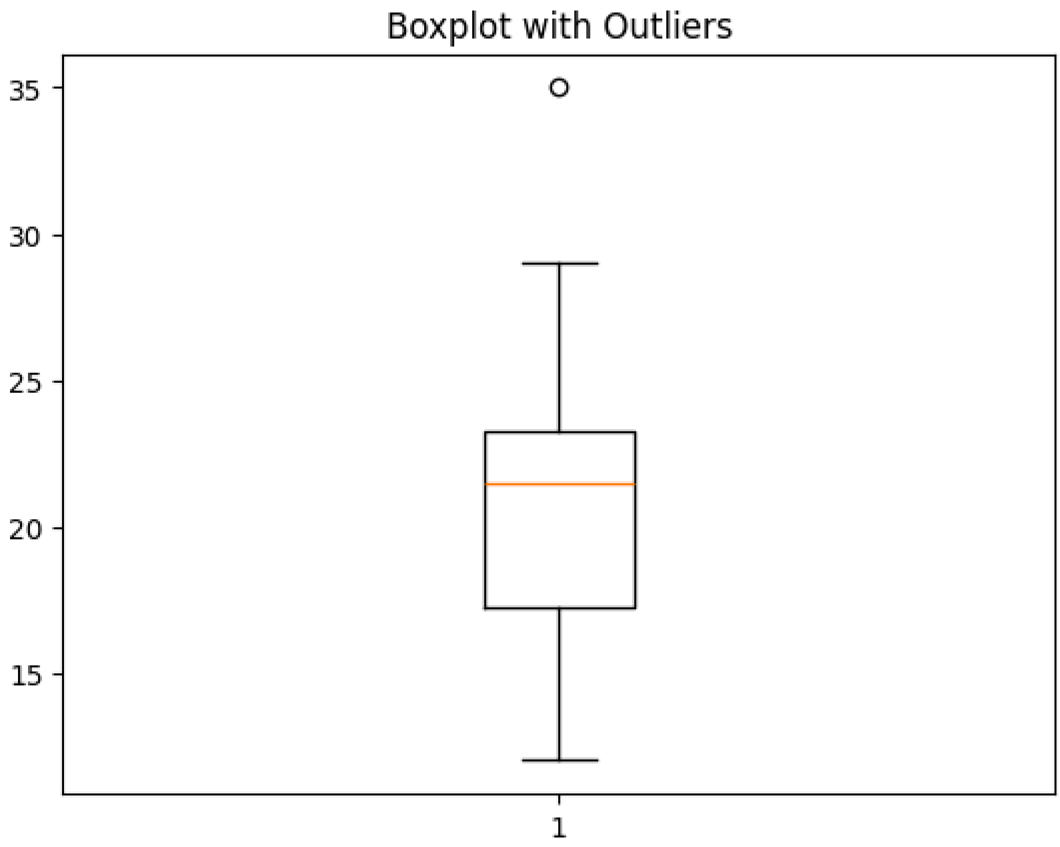
```
import matplotlib.pyplot as plt
```

```
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]


plt.boxplot(data)

plt.title("Boxplot with Outliers")

plt.show()
```



Question 8: You are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales.

Explain how you would use covariance and correlation to explore this relationship:

- Covariance
  - **Measures direction of relationship.**
  - **Positive covariance → Sales increase as ad spend increases.**
  - **Negative covariance → Sales decrease as ad spend increases.**
- Correlation
  - **Measures strength and direction of relationship (range −1 to +1).**
  - **+1 = Perfect positive relation; 0 = No relation; −1 = Perfect negative relation.**

```
advertising_spend = [200, 250, 300, 400, 500]

daily_sales = [2200, 2450, 2750, 3200, 4000]



correlation = np.corrcoef(advertising_spend, daily_sales)[0][1]

print(f"Correlation: {correlation}")
```

Output:

```
Correlation: 0.9935824101653329
```

Question 9: Your team has collected customer satisfaction survey data on a scale of 1-10 and wants to understand its distribution before launching a new product.

```
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
```

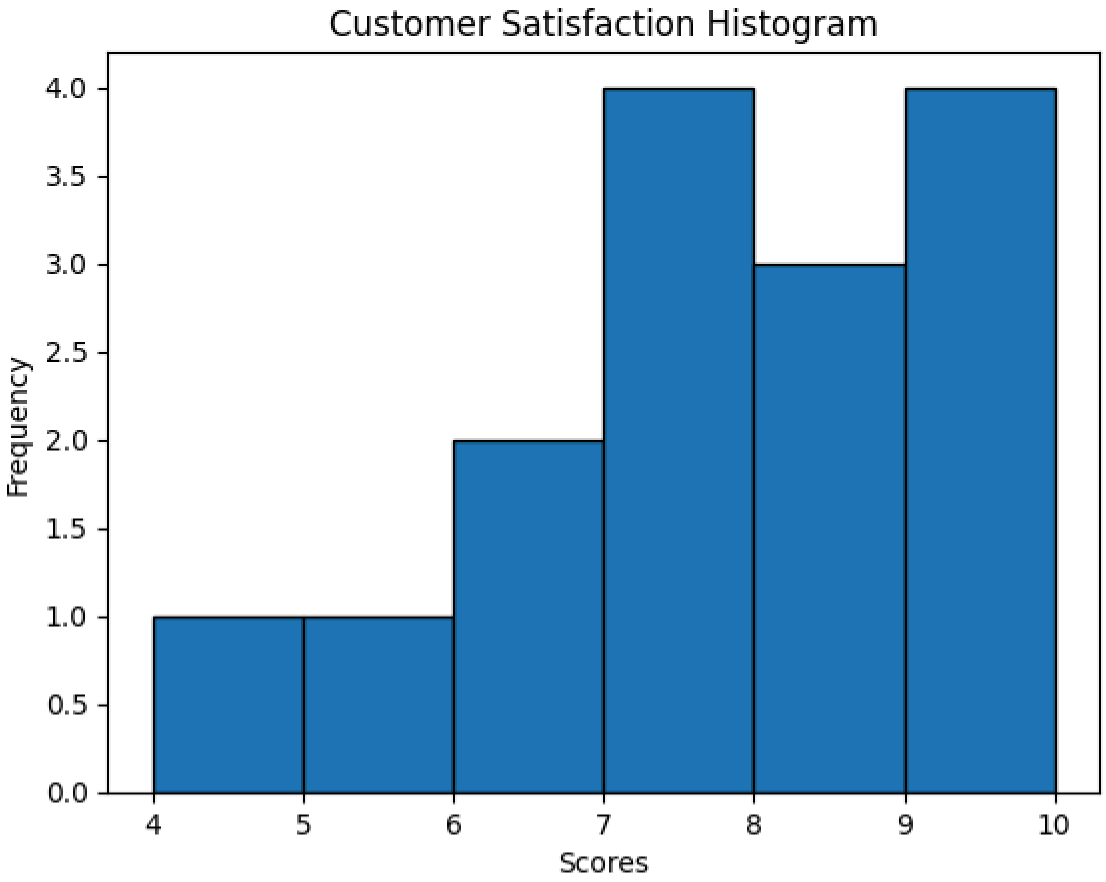
```
plt.hist(survey_scores, bins=6, edgecolor='black')

plt.title("Customer Satisfaction Histogram")

plt.xlabel("Scores")

plt.ylabel("Frequency")

plt.show()
```


Customer Satisfaction Histogram
Frequency
Scores