Statistics Basics| Assignment

Q1. What is the difference between descriptive statistics and inferential statistics? Explain with examples.

A1. Descriptive statistics deals with collecting, organizing, summarizing, and presenting data in a meaningful way. It only describes the data that is already available and does not make any predictions or conclusions beyond it. Ex calculating the average marks of students in a class.

Inferential statistics is used to draw conclusions, make predictions, or generalize results about a population based on a sample of data. It helps in decision-making under uncertainty. Ex estimating the average income of a city based on a sample survey.

Q2. What is sampling in statistics? Explain the differences between random and stratified sampling.

A2. Sampling is the process of selecting a small group (sample) from a large group (population) to study and analyze, so that conclusions can be drawn about the entire population. Sampling is used because studying the whole population is often time-consuming, costly, and impractical.

Random Sampling : In random sampling, every member of the population has an equal chance of being selected. The selection is done purely by chance, without any bias. Ex : selecting 100 students randomly from a list of all students using a lottery.

Stratified Sampling : In stratified sampling, the population is first divided into subgroups (called strata) based on certain characteristics such as age, gender, department, or income. Then, samples are randomly selected from each subgroup. Ex : dividing students into Science, Commerce, and Arts streams, and then selecting students randomly from each stream.

Q3. Define mean, median, and mode. Explain why these measures of central tendency are important.

A3. Mean : The mean is the average value of a dataset. It is calculated by adding all the values and dividing the sum by the total number of observations.

Median : The median is the middle value of a dataset when the data is arranged in ascending or descending order. If the number of observations is odd, the median is the middle value. If the number of observations is even, the median is the average of the two middle values.

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

Measures of central tendency are important because they simplify data analysis, support decision-making, and provide a clear picture of the data distribution.

Q4. Explain skewness and kurtosis. What does a positive skew imply about the data?

A4. Skewness : Skewness measures the degree of asymmetry of a data distribution around its mean. If the distribution is symmetrical, skewness is zero. If the distribution is stretched more on one side, it is said to be skewed.

Kurtosis : Kurtosis measures the peakedness or flatness of a data distribution compared to a normal distribution.

A positive skew means that the right tail of the distribution is longer than the left tail.Most data values are concentrated on the lower side.

Q5. Implement a Python program to compute the mean, median, and mode of a given list of numbers.

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

(Include your Python code and output in the code box below.)

A5.

[6]
0s

```
import statistics

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

mean = statistics.mean(numbers)
median = statistics.median(numbers)
mode = statistics.mode(numbers)

print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
```

```
Mean: 19.6
Median: 19
Mode: 12
```

Q6. Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

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

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

(Include your Python code and output in the code box below.)

A6.

[7]
0s

```
import statistics

list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

covariance = statistics.covariance(list_x, list_y)
correlation = statistics.correlation(list_x, list_y)

print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

```
Covariance: 275.0
Correlation Coefficient: 0.9958932064677039
```

Q7. 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]

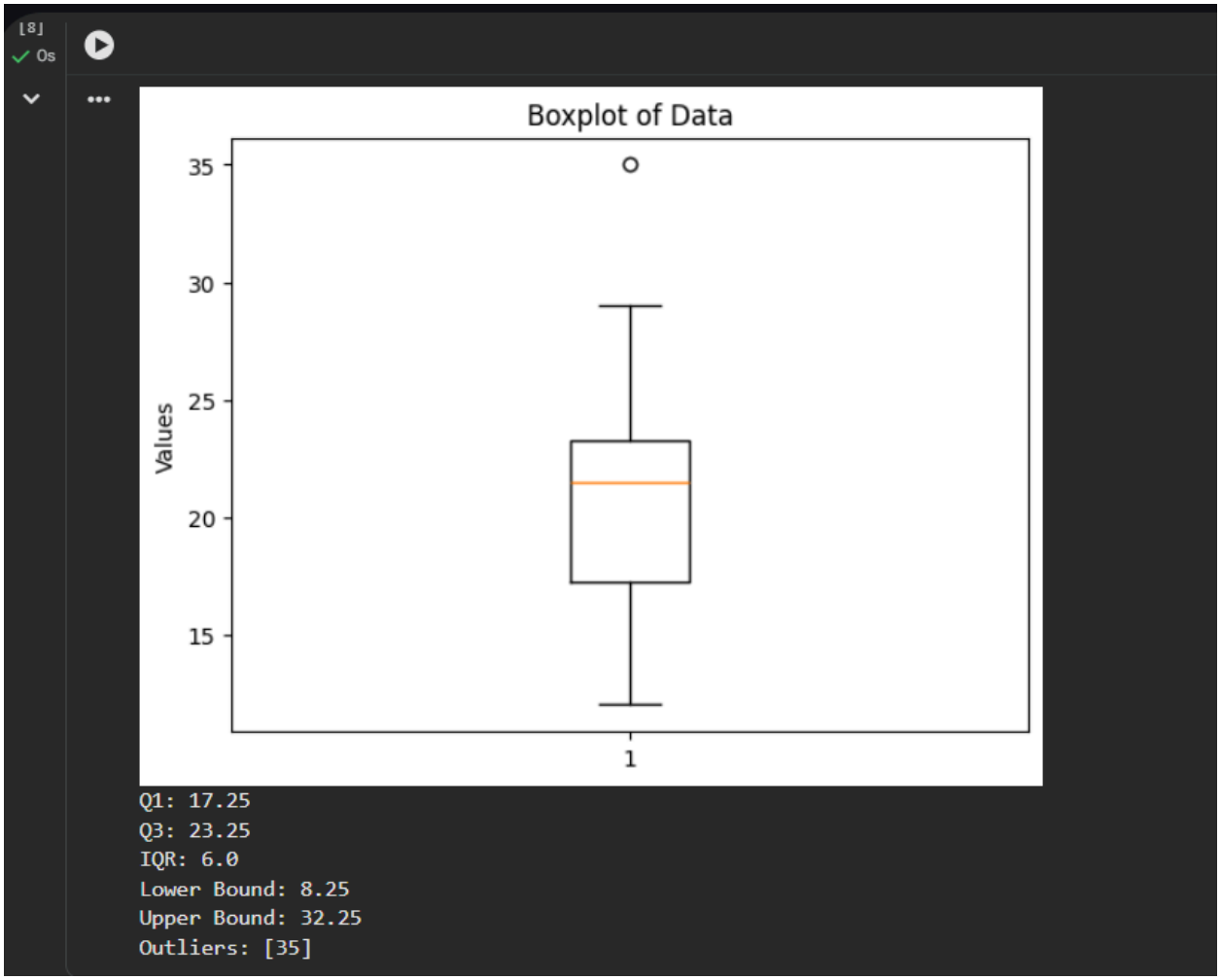
(Include your Python code and output in the code box below.)

A7.

[8]
✓ 0s

```
import matplotlib.pyplot as plt
import numpy as np

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

arr = np.array(data)

Q1 = np.percentile(arr, 25)
Q3 = np.percentile(arr, 75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = arr[(arr < lower_bound) | (arr > upper_bound)]

plt.boxplot(arr)
plt.title("Boxplot of Data")
plt.ylabel("Values")
plt.show()

print("Q1:", Q1)
print("Q3:", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers.tolist())
```

Output :



Q8. 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.

● Write Python code to compute the correlation between the two lists:

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

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

(Include your Python code and output in the code box below.)

A8. I would use covariance and correlation as follows:

Covariance helps identify whether advertising spend and sales move together or in opposite directions.

Positive covariance : both increase or decrease together.

Negative covariance : one increases while the other decreases

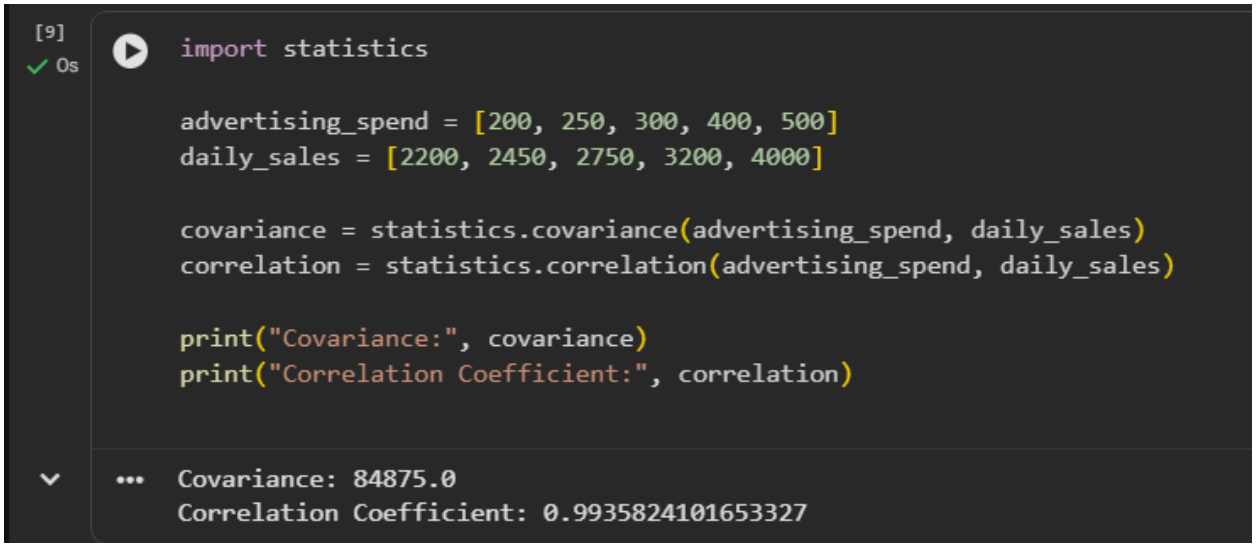
Correlation measures both the strength and direction of the relationship on a standardized scale from -1 to +1.

+1 : perfect positive relationship.

0 : no relationship.

-1 : perfect negative relationship.

In this case, correlation helps the marketing team understand how strongly advertising spend impacts daily sales.



Q9. 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.

● Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you'd use.

● Write Python code to create a histogram using Matplotlib for the survey data:

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

A9. I would use the following summary statistics and visualizations:

Summary Statistics:

Mean: Shows the average customer satisfaction level.

Median: Indicates the central value and is useful if data is skewed.

Standard Deviation: Measures how spread out the satisfaction scores are. A high value indicates varied opinions, while a low value shows consistent feedback.

Visualizations

Histogram: Helps visualize the distribution of scores and identify patterns such as concentration, skewness, or gaps.

These measures together provide a clear understanding of customer sentiment and readiness for a product launch.

code :

```
[10]
✓ 0s
```

```python
import matplotlib.pyplot as plt
import statistics

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

mean_score = statistics.mean(survey_scores)
std_dev = statistics.stdev(survey_scores)

print("Mean Score:", mean_score)
print("Standard Deviation:", std_dev)

plt.hist(survey_scores, bins=7)
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Survey Scores")
plt.ylabel("Frequency")
plt.show()
```

Output :

```
Mean Score: 7.333333333333333
Standard Deviation: 1.632993161855452
```