Statistics Basics | Assignment

Assignment Code: DS-AG-005

Total Marks: 200

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

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

Answer : Statistics is a key field that helps us make sense of data through collection, analysis, and presentation. It plays an important role in many areas, from business to healthcare, by guiding decision-making and drawing conclusions. This process is made easier with the help of two main branches of statistics: descriptive and inferential.

Aspect	Descriptive statistics	Inferential Statistics
Definition	Gives information about raw data; describes and summarizes the data in some manner.	Makes inferences about the population using data drawn from the population.
Purpose	Helps in organizing, analyzing, and presenting data in a meaningful way.	Allows comparison, testing of hypotheses, and making predictions.
Use Case	Used to describe a situation or the characteristics of a dataset.	Used to explain the probability of occurrence of an event and draw conclusions.
Scope	Explains already known data; limited to the dataset (sample or small population).	Attempts to generalize and reach conclusions about the whole population.
Techniques/Examples	Mean, Median, Mode, Range, Variance, Standard Deviation, Charts, Tables, Histograms, etc.	Confidence Intervals, Hypothesis Testing, Regression Models, ANOVA, p-values, etc.
Result Type	Limited to presenting and analyzing the data at hand.	Provides predictions and conclusions that go beyond the observed data.
Application	Used for describing trends and organizing data for presentation.	Used for predicting trends, testing hypotheses, and generalizing results to population.

Example of Descriptive statistics- A teacher calculates the average score of her 40 students in a math exam.

Mean = 68, Standard Deviation = 12.This only tells us about these 40 students; no generalization is made about all students in the school.

Example of Inferential statistics- Suppose the same teacher selects a random sample of 40 students from the school and finds their average math score = 68.

By applying inferential statistics, she estimates the population mean score of all students in the school, and tests whether the average differs significantly from 65 at a 5% significance level.

Question 2: What is sampling in statistics? Explain the differences between random and stratified sampling.
Answer :
Sampling in statistics : Sampling is the process of selecting a subset of individuals, items, or observations from a larger group known as the population. Since studying the entire population is often time-consuming, costly, and impractical, sampling allows researchers to draw reliable conclusions about the population by analyzing a smaller, manageable group. For example, instead of collecting exam scores from 10,000 university students, a researcher may select 500 students as a sample to estimate the average score of the entire university.

Difference between Random and Stratified Sampling:

Random Sampling: One of the most common methods of sampling is random sampling, where every member of the population has an equal chance of being selected. This method is simple, unbiased, and works best when the population is homogeneous. For instance, if a teacher wants to study the performance of students in a class, she may randomly select 10 students using a lottery method or a random number generator. However, random sampling may sometimes lead to underrepresentation of certain subgroups in the population.

Stratified Sampling: To overcome Random Sampling limitations, stratified sampling is used. In this method, the population is first divided into subgroups or strata based on specific characteristics such as gender, age, income level, or education. Then, random samples are drawn from each stratum, usually in proportion to their size in the population. This ensures that all important groups are adequately represented. For example, in a study of university students, the researcher may divide the students into undergraduates and postgraduates, or into male and female categories, and then take a proportional random sample from each group.

Notable Key Differences: The main difference between random and stratified sampling lies in their approach. Random sampling treats the entire population as one group and selects individuals purely by chance, whereas stratified sampling ensures representation by dividing the population into meaningful subgroups before selecting the sample. Random sampling is simpler and easier to implement, but stratified sampling provides more accurate and representative results when the population is heterogeneous.

In conclusion, both methods are widely used in statistics, but the choice between random and stratified sampling depends on the nature of the population being studied. Random sampling works well for uniform populations, while stratified sampling is more suitable when the population has distinct categories that need to be represented fairly.

Question 3: Define mean, median, and mode. Explain why these measures of central tendency are important.
Answer : In statistics, measures of central tendency are numerical values that describe the center or average of a dataset. They give a single representative value around which the data is distributed. The three most common measures are mean, median, and mode.

Mean : The mean (or average) is the sum of all data values divided by the number of observations.

Formula: mean = sum of observations / Total number of observations

Example: For the data 2, 4, 6, 8, the mean = (2+4+6+8)/4 = 20/4 = 5.

Note : Mean may not be fruitful in case of extreme values (outliers).

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

Example: For the data 3, 7, 9, 12, 15, the median = 9. For data 2, 4, 6, 8, the median = (4+6)/2 = 5.

Note : The median is not affected by extreme values and is a better measure when data is skewed.

Mode :The mode is the value that occurs most frequently in the dataset. A dataset may have no mode, one mode (unimodal), or more than one mode (bimodal or multimodal).

Example: For the data 2, 3, 4, 4, 5, 6, the mode = 4.

Note : The mode is particularly useful for categorical data (e.g., finding the most popular product or color choice).

Importance of measuring Central Tendencies :

- They summarize data into a single value, making large datasets easier to interpret.
- They help in comparison between different groups or datasets.
- They serve as the foundation for further statistical analysis such as variance, standard deviation, and hypothesis testing.
- They provide insights for decision-making in fields like business, economics, education, and health sciences.
- Each measure has its own strength: mean is widely used, median is robust to outliers, and mode is suitable for qualitative/categorical data

---

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?
Answer: The shape of a data distribution is essential in addition to knowing its central tendency and dispersion. Two important measures that describe the shape of a distribution are skewness and kurtosis.

Skewness : Skewness measures the asymmetry of a data distribution around its mean.

- Types of Skewness:
    - ➔ Symmetrical Distribution (Skewness = 0): Data is evenly distributed on both sides of the mean (e.g., normal distribution).
    - ➔ Positive Skew (Right-Skewed): Tail on the right side is longer; mean > median > mode.
    - ➔ Negative Skew (Left-Skewed): Tail on the left side is longer; mean < median < mode.

Example: Income distribution in a country is usually positively skewed because a small number of very high incomes pull the mean to the right.

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

- ➢ Types of Kurtosis:
    - ○ Mesokurtic (Kurtosis = 3): Normal distribution; moderate tails and peak.
    - ○ Leptokurtic (Kurtosis > 3): Distribution has a sharper peak and fatter tails → more extreme values.
    - ○ Platykurtic (Kurtosis < 3): Distribution is flatter with thinner tails → fewer extreme values.

Example: Stock market returns often show leptokurtic behavior because of occasional extreme fluctuations.

Positive Skew Imply About Data : A positive skew means the distribution has a longer right tail. This implies that:

- Most of the data values are concentrated on the lower side (left side) of the distribution.
- A few unusually large values (outliers) pull the mean to the right.
- Results in - Mean>Median>Mode

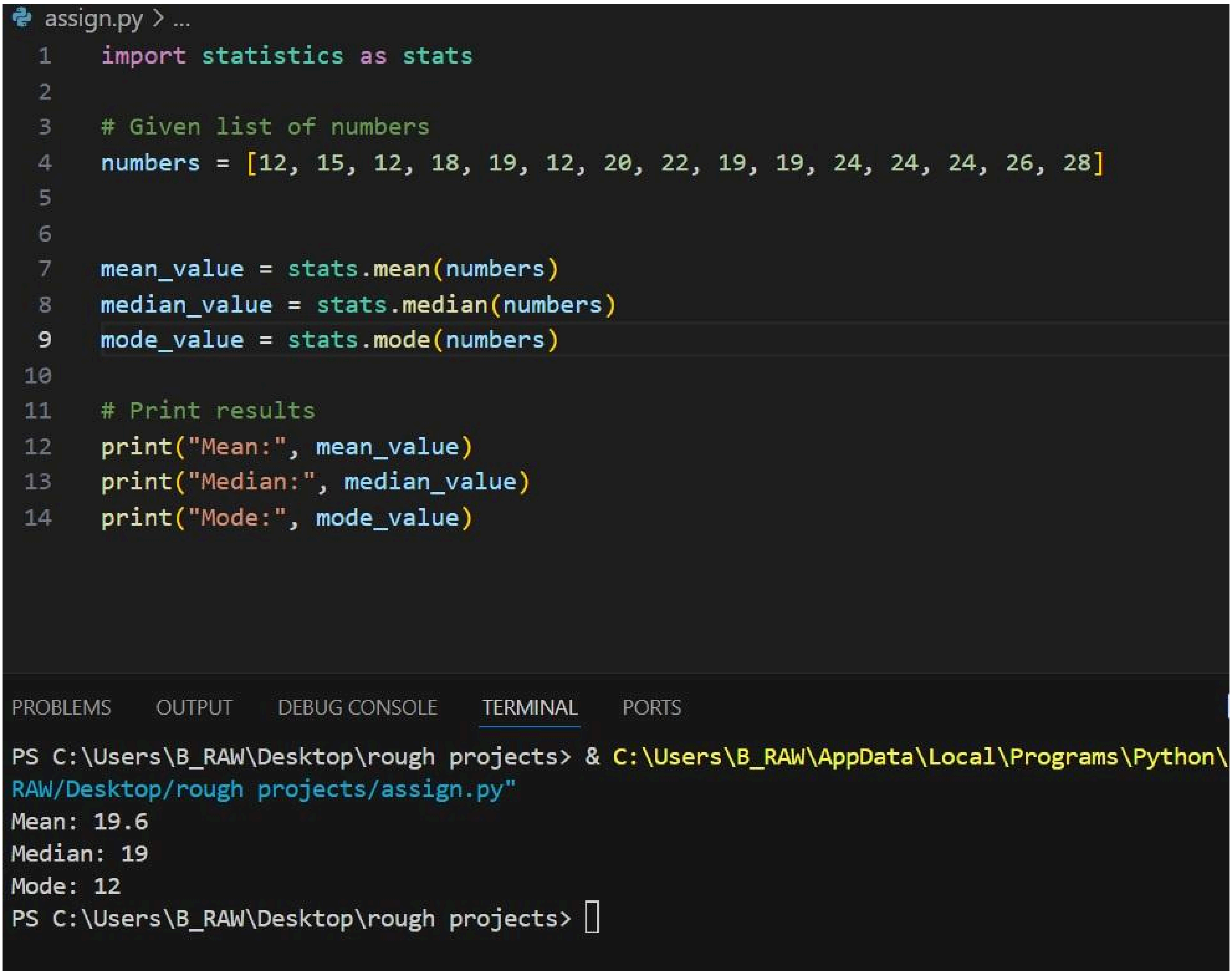
Example: Personal income, house prices, and waiting times are typically positively skewed.

Question 5: 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.)

Answer:



```python
import statistics as stats

# Given list of numbers
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]


mean_value = stats.mean(numbers)
median_value = stats.median(numbers)
mode_value = stats.mode(numbers)

# Print results
print("Mean:", mean_value)
print("Median:", median_value)
print("Mode:", mode_value)
```

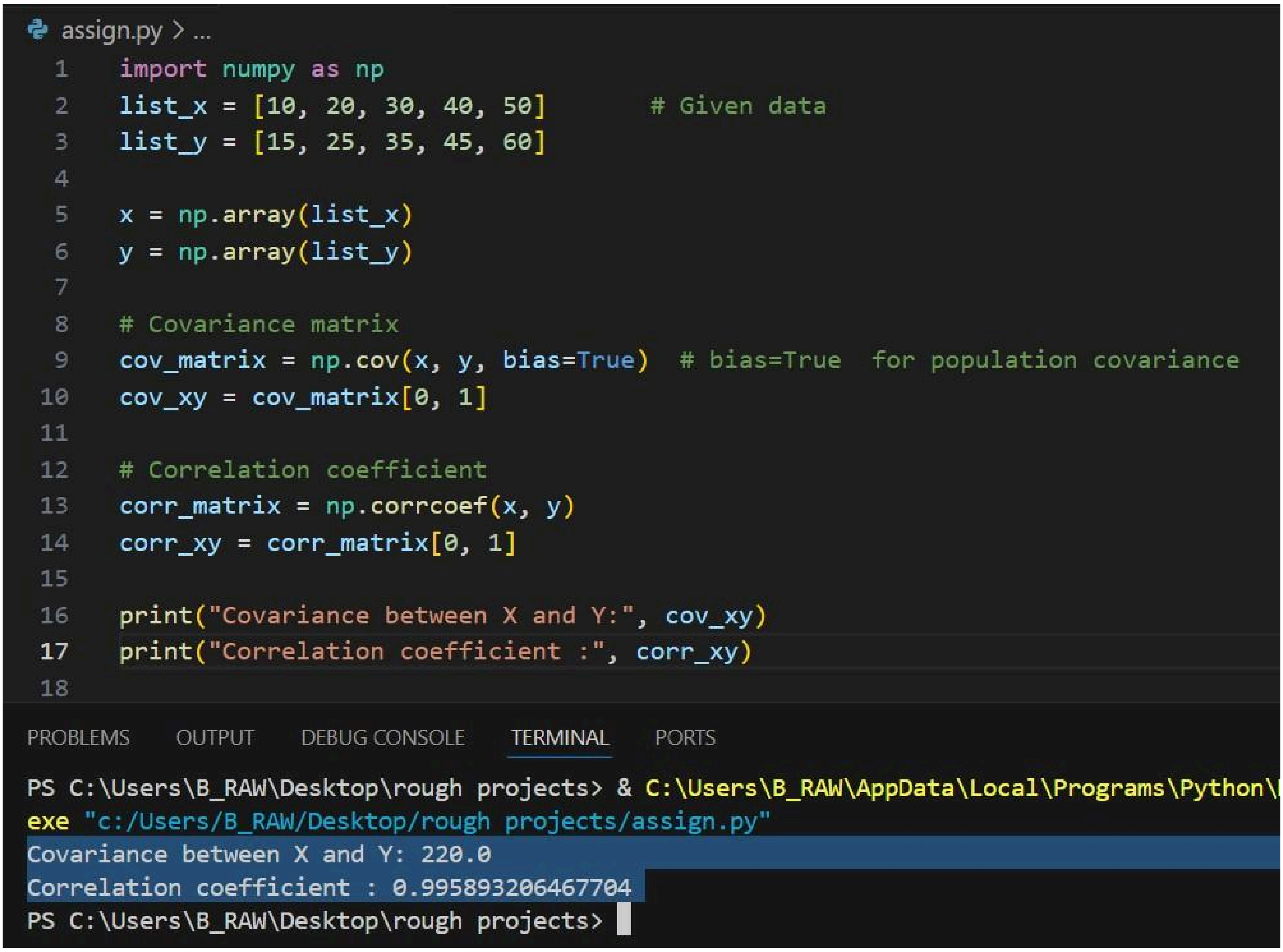
```
PS C:\Users\B_RAW\Desktop\rough projects> & C:\Users\B_RAW\AppData\Local\Programs\Python\
RAW/Desktop/rough projects/assign.py"
Mean: 19.6
Median: 19
Mode: 12
PS C:\Users\B_RAW\Desktop\rough projects> 
```

Question 6: 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.)

Answer:



```python
import numpy as np
list_x = [10, 20, 30, 40, 50]       # Given data
list_y = [15, 25, 35, 45, 60]

x = np.array(list_x)
y = np.array(list_y)

# Covariance matrix
cov_matrix = np.cov(x, y, bias=True)  # bias=True  for population covariance
cov_xy = cov_matrix[0, 1]

# Correlation coefficient
corr_matrix = np.corrcoef(x, y)
corr_xy = corr_matrix[0, 1]

print("Covariance between X and Y:", cov_xy)
print("Correlation coefficient :", corr_xy)

```

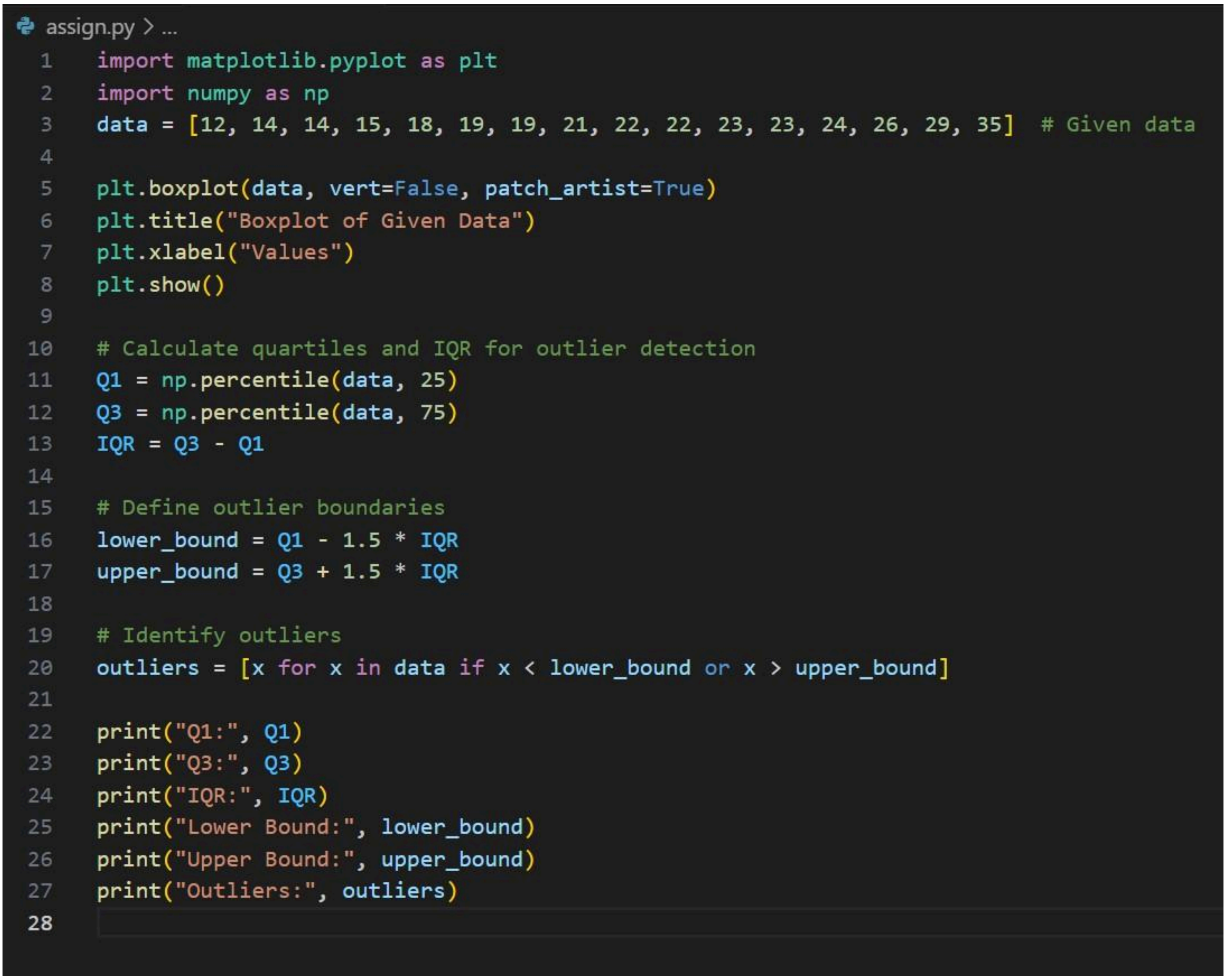
```
PS C:\Users\B_RAW\Desktop\rough projects> & C:\Users\B_RAW\AppData\Local\Programs\Python\
exe "c:/Users/B_RAW/Desktop/rough projects/assign.py"
Covariance between X and Y: 220.0
Correlation coefficient : 0.995893206467704
PS C:\Users\B_RAW\Desktop\rough projects>
```

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

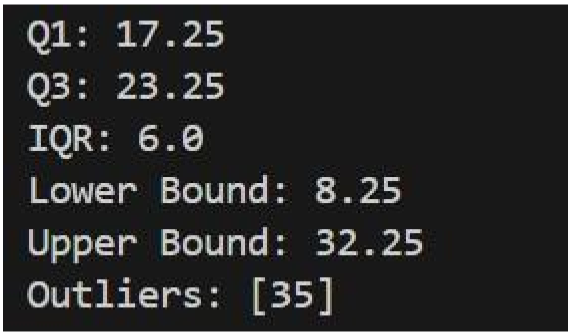
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

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

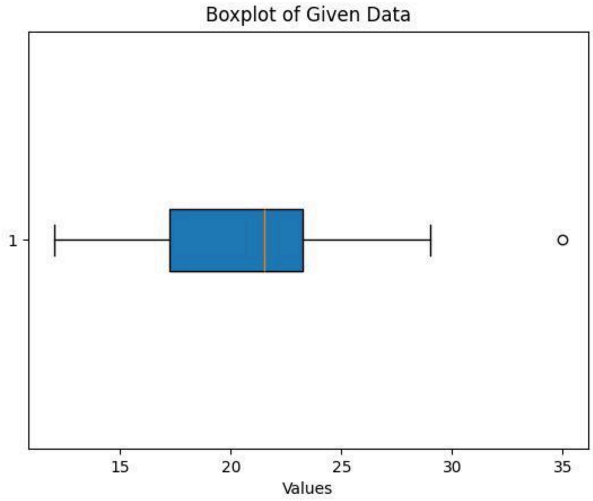
Answer:



```python
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]  # Given data

plt.boxplot(data, vert=False, patch_artist=True)
plt.title("Boxplot of Given Data")
plt.xlabel("Values")
plt.show()

# Calculate quartiles and IQR for outlier detection
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

# Define outlier boundaries
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identify outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]

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

```



```
Q1: 17.25
Q3: 23.25
IQR: 6.0
Lower Bound: 8.25
Upper Bound: 32.25
Outliers: [35]
```



The result we have is a box plot representing the values in the x axis and gives information on lower bound ,upper bound, IQR , Q1, Q2 and the outliers.

From the above box plot it is clearly visible that data is well distributed and there are not much outliers as from above data we can find only 35 is present as extreem value.

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.**
- **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.)

Answer:

As a data analyst, to understand whether advertising spend influences daily sales, I would use:

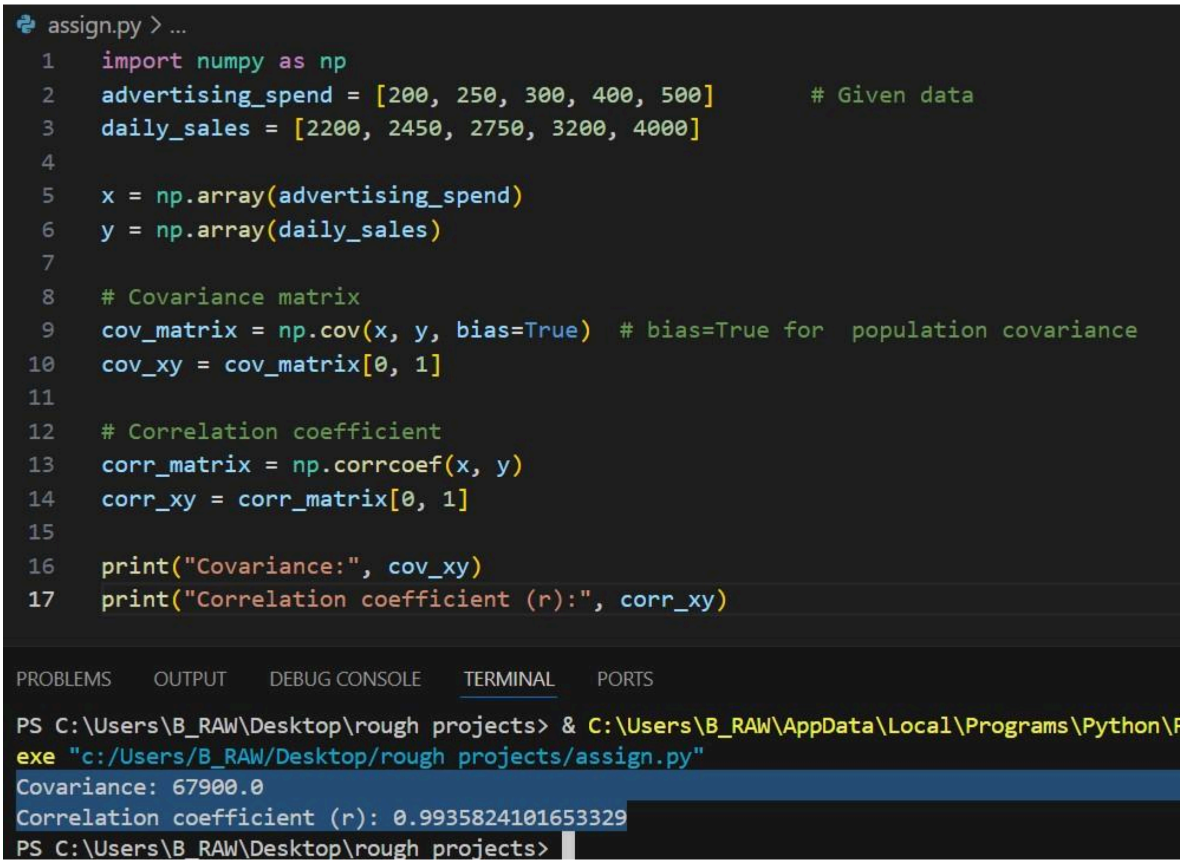
Covariance:

- Measures the direction of the relationship.
- If covariance > 0 means when advertising spend increases, daily sales also increase.
- If covariance < 0 means when advertising spend increases, daily sales decrease.
- Note : Covariance does not tell us about the strength of the relationship, only the direction.

Correlation:

- Standardized measure of the relationship between two variables.
- Range: -1 ≤ r ≤ +1
- r ≈ +1 means strong positive relationship (sales rise with ad spend).
- r ≈ -1 means strong negative relationship (sales fall with ad spend).
- r ≈ 0 means no linear relationship.
- Note : Correlation removes the unit of measurement, making it easier to interpret compared to covariance.

Thus, I would first check covariance to see the direction of the relationship and then compute the correlation coefficient to measure its strength.



```python
import numpy as np
advertising_spend = [200, 250, 300, 400, 500]       # Given data
daily_sales = [2200, 2450, 2750, 3200, 4000]

x = np.array(advertising_spend)
y = np.array(daily_sales)

# Covariance matrix
cov_matrix = np.cov(x, y, bias=True)  # bias=True for  population covariance
cov_xy = cov_matrix[0, 1]

# Correlation coefficient
corr_matrix = np.corrcoef(x, y)
corr_xy = corr_matrix[0, 1]

print("Covariance:", cov_xy)
print("Correlation coefficient (r):", corr_xy)
```

```
PS C:\Users\B_RAW\Desktop\rough projects> & C:\Users\B_RAW\AppData\Local\Programs\Python\P
exe "c:/Users/B_RAW/Desktop/rough projects/assign.py"
Covariance: 67900.0
Correlation coefficient (r): 0.9935824101653329
PS C:\Users\B_RAW\Desktop\rough projects>
```

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.

● 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]

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

Answer : To understand the distribution of customer satisfaction scores (1–10 scale), Let me first describe the meaning of each one :

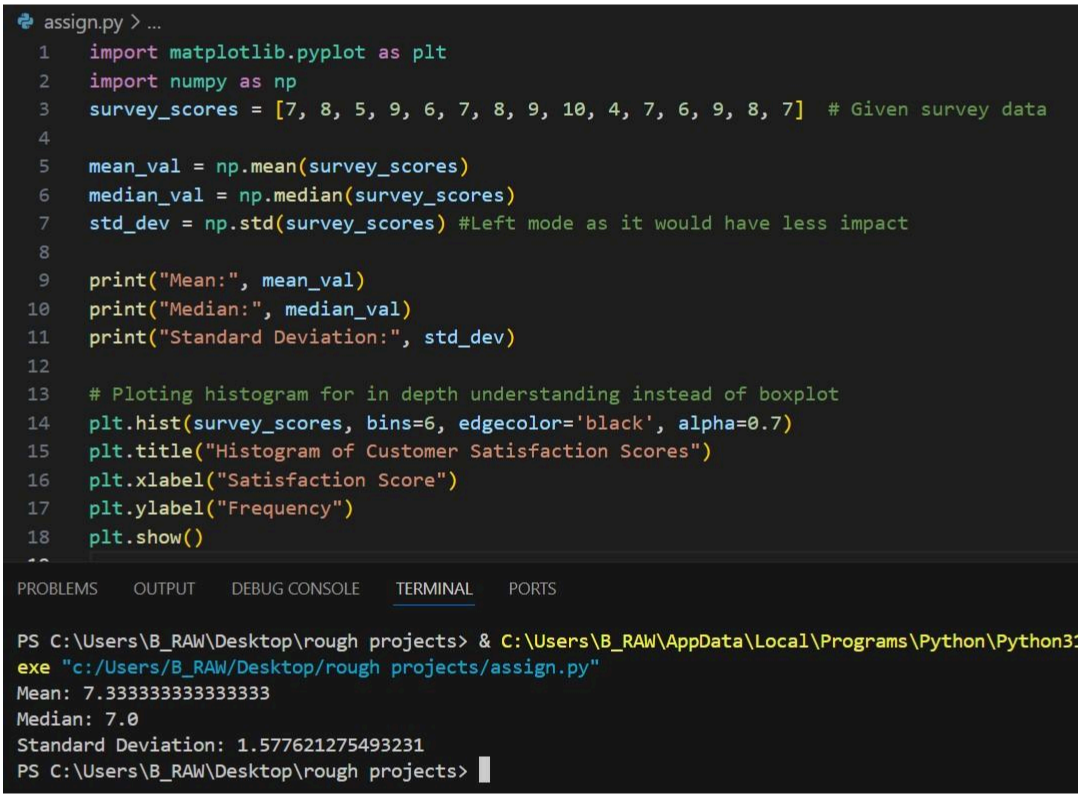
- Mean (Average): Provides the central tendency of satisfaction levels.
- Median: Shows the middle score and is less affected by outliers.
- Mode: Identifies the most common rating given by customers.
- Standard Deviation (SD): Measures the variability of responses; higher SD means customers' opinions are more spread out.

Since outliers is not much and mean can work well with the data set I would use mean and standard deviation more than median and mode to know in depth of customer satisfaction.

For visualization, I would use:

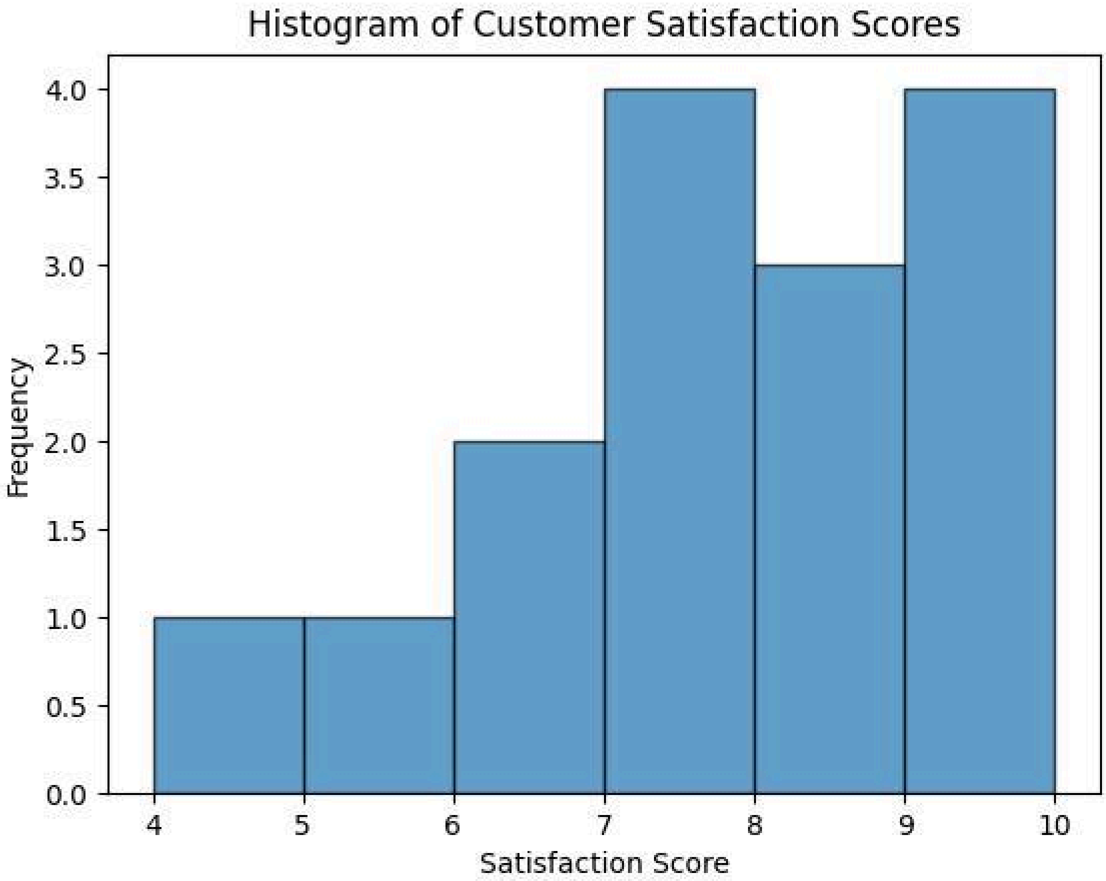
- Histogram: To show the frequency distribution of survey scores.
- Boxplot: To detect spread and outliers.

Together, these statistics and visualizations provide insights into whether customers are generally satisfied (scores skewed high) or dissatisfied (scores skewed low), and whether opinions are consistent or highly varied.



```python
import matplotlib.pyplot as plt
import numpy as np
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]  # Given survey data

mean_val = np.mean(survey_scores)
median_val = np.median(survey_scores)
std_dev = np.std(survey_scores) #Left mode as it would have less impact

print("Mean:", mean_val)
print("Median:", median_val)
print("Standard Deviation:", std_dev)

# Ploting histogram for in depth understanding instead of boxplot
plt.hist(survey_scores, bins=6, edgecolor='black', alpha=0.7)
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Satisfaction Score")
plt.ylabel("Frequency")
plt.show()
```

```
PS C:\Users\B_RAW\Desktop\rough projects> & C:\Users\B_RAW\AppData\Local\Programs\Python\Python3
exe "c:/Users/B_RAW/Desktop/rough projects/assign.py"
Mean: 7.333333333333333
Median: 7.0
Standard Deviation: 1.577621275493231
PS C:\Users\B_RAW\Desktop\rough projects>
```



The histogram above clearly reflects that customers like the product with a satisfactory range mostly between 7 to 10 and the new product can be launched with ease as per this sample survey.