Assignment Code: DS-AG-005

Statistics Basics| **Assignment**

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.

Total Marks: 200

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

Answer:

Descriptive Stats : It Deals with summarizing, organizing and describing data.
Eg: Suppose you collected exam scores of 50 students in your class :

- The average score (mean) is 72.
- The highest score is 95, and the lowest is 40.

Inferential Statistics : It Deals with making predictions or generalization about a population based on sample data.
Eg: Imagine you want to know the average height of all students in your college.

- It's impossible (or very hard) to measure every single student.
- So, you randomly select 50 students and measure their heights.
- From this sample, you find the average height = 165 cm.

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

Answer:

Sampling is a method used in statistics to select a smaller group (sample) from a larger group (population) so that we can study the sample and draw conclusions about the entire population.
Random Sampling : In random sampling every individual has equal chance of being selected in the sample.

Stratified Sample : In Stratified Sampling population is divided into groups based on certain traits and then a random sample is taken from each strata

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

Answer:

Mean : It is the average of the entire sample.
Example : s = {1,2,3,4}
Mean = (1+2+3+4)4 = 2.4
Median : It is the middle most value of the sample.
Example : s = {1,2,3,4,5}
Median = 3
Mode : It is the most frequently occuring value in the sample.
Example : s = {1,2,1,1,2,1,3,4,5,4,3}
Mode = 1

The measure of central tendency is important because it gives us a single value that represents the entire dataset. Instead of looking at all the numbers, we can use the mean, median, or mode to quickly understand the "typical" or "central" value.

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

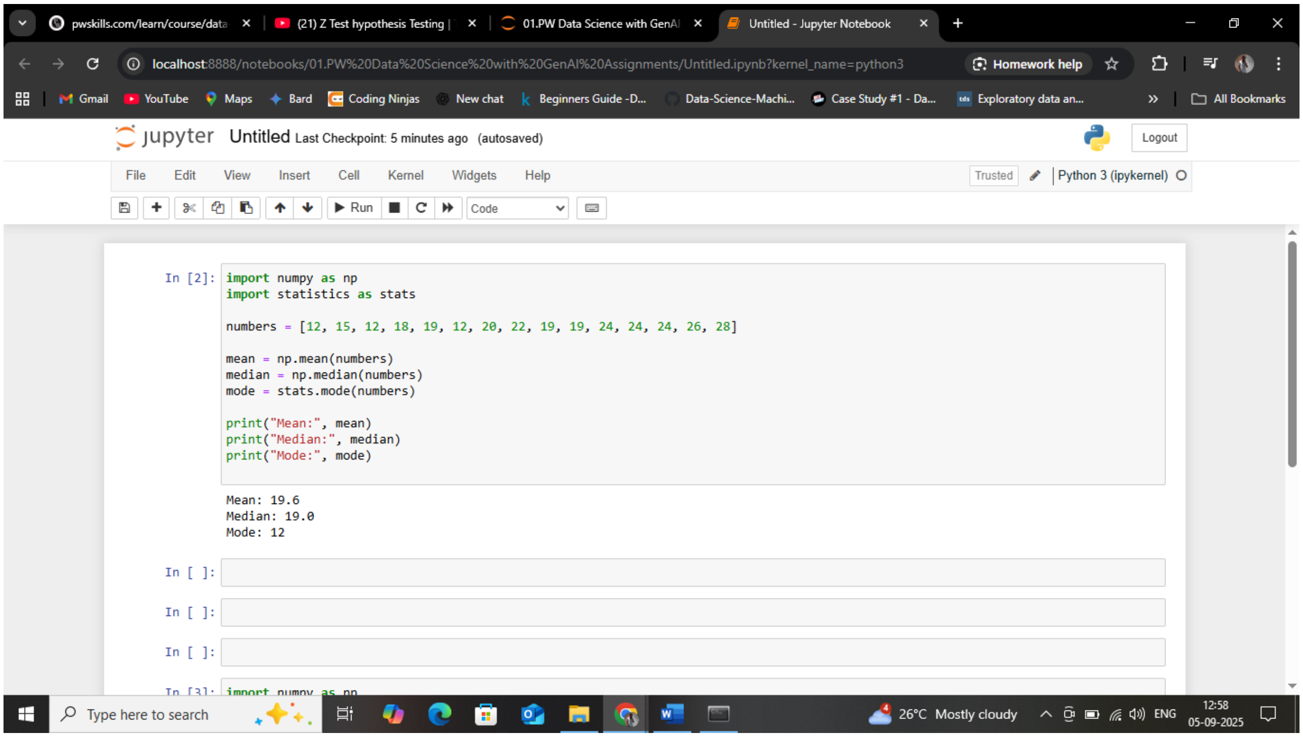
Answer: Skewness tells us about the symmetricity of the distribution of data.

Positive Skew means: The tail on the right side is longer → most data values are on the left, but a few very large values pull the mean to the right.

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]

Answer:



localhost:8888/notebooks/01.PW%20Data%20Science%20with%20GenAI%20Assignments/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: 5 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

In [2]:

```
import numpy as np
import statistics as stats

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

mean = np.mean(numbers)
median = np.median(numbers)
mode = stats.mode(numbers)

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

```
Mean: 19.6
Median: 19.0
Mode: 12
```

In [ ]:

In [ ]:

In [ ]:

In [3]:

```
import numpy as np
```

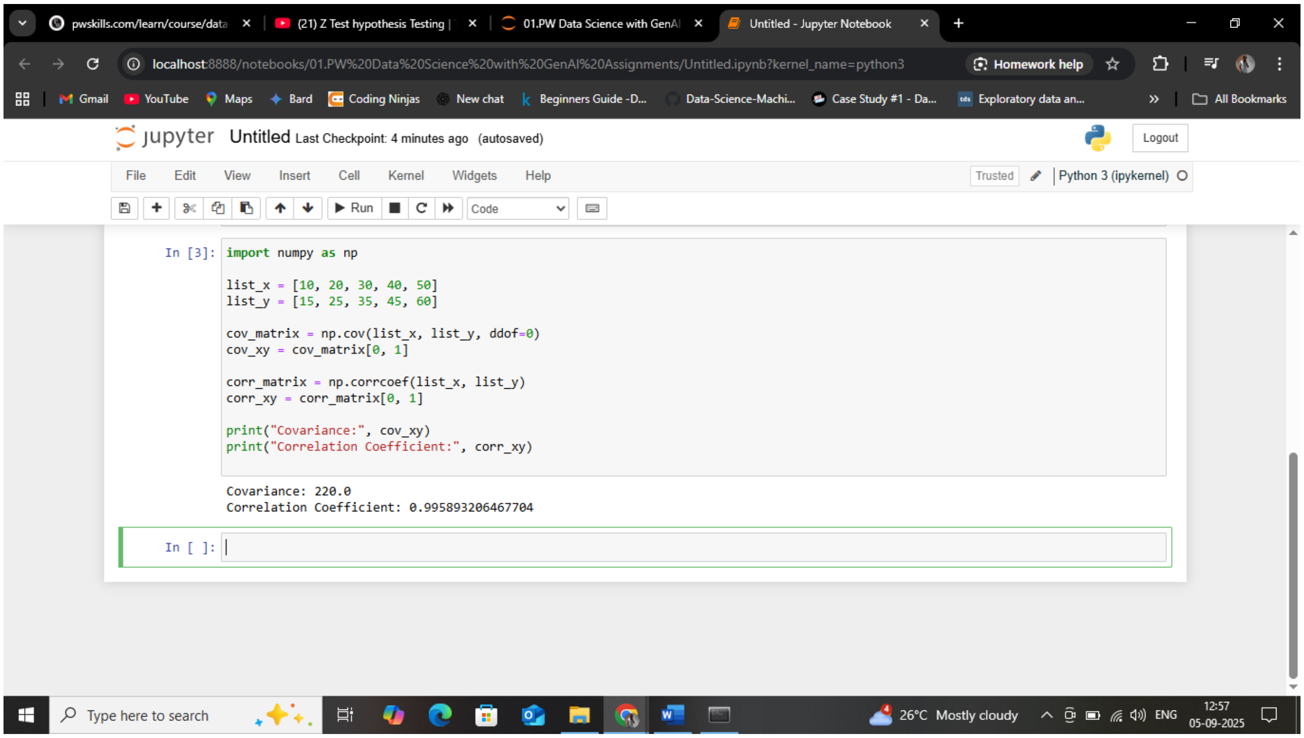
Question 6: Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

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:

Paste your code and output inside the box below:



```
In [3]: 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, ddof=0)
        cov_xy = cov_matrix[0, 1]

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

        print("Covariance:", cov_xy)
        print("Correlation Coefficient:", corr_xy)
```

```
Covariance: 220.0
Correlation Coefficient: 0.995893206467704
```

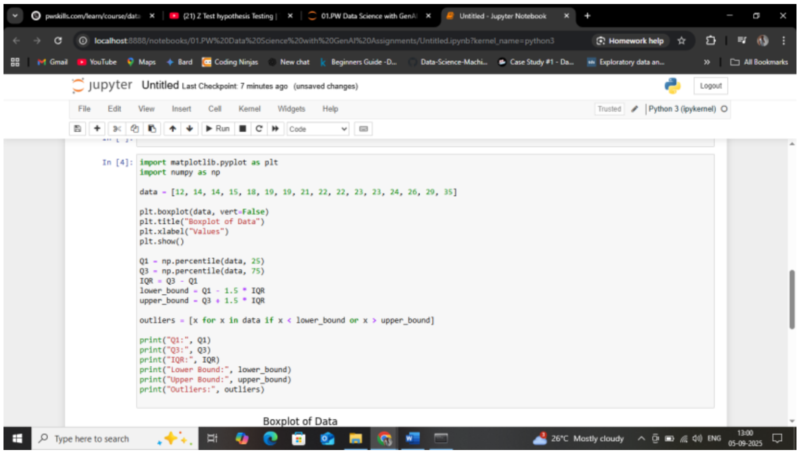
```
In [ ]:
```

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

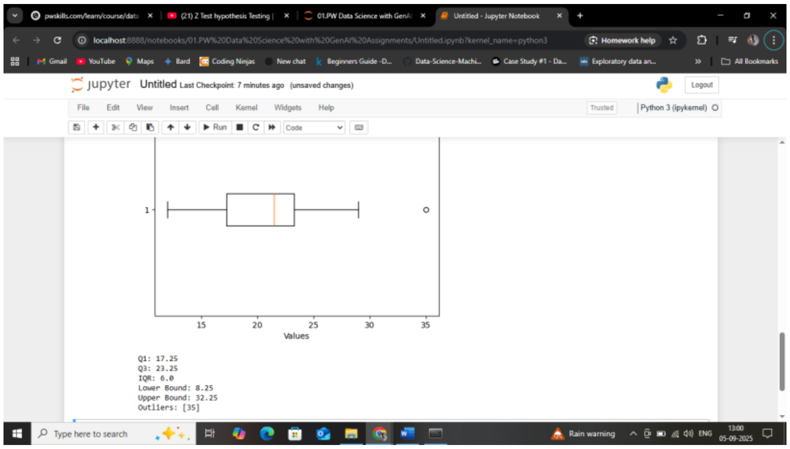
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

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

Answer:



```
In [4]: 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]

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

        Q1 = np.percentile(data, 25)
        Q3 = np.percentile(data, 75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        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]
```

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.

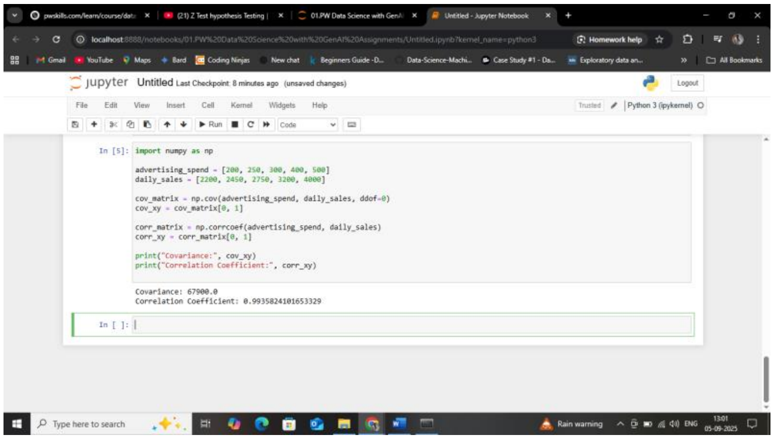
- Explain how you would use covariance and correlation to explore this relationship.
- Write Python code to compute the correlation between the two lists:

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:



```
In [5]: import numpy as np

        advertising_spend = [200, 250, 300, 400, 500]
        daily_sales = [2200, 2450, 2750, 3200, 4000]

        cov_matrix = np.cov(advertising_spend, daily_sales, ddof=0)
        cov_xy = cov_matrix[0, 1]

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

        print("Covariance:", cov_xy)
        print("Correlation Coefficient:", corr_xy)

        Covariance: 67900.0
        Correlation Coefficient: 0.9935824101653329

In [ ]:
```

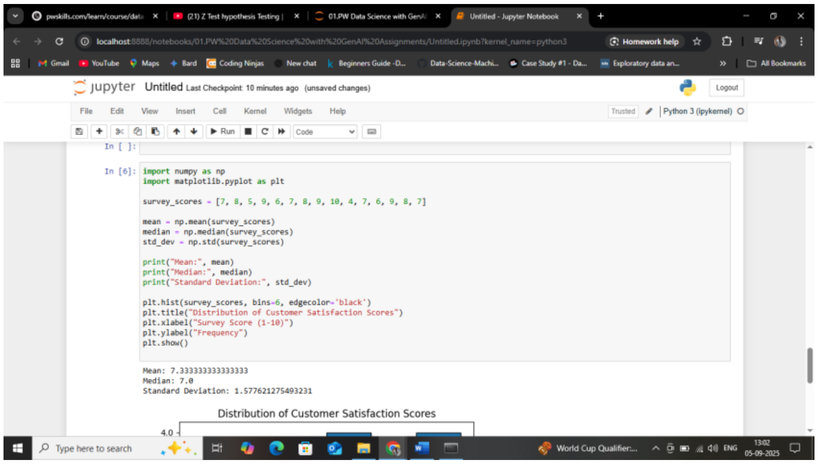
There is a strong positive correlation between advertising spend and daily sales. This means increasing advertising spend is highly likely to increase sales.

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]

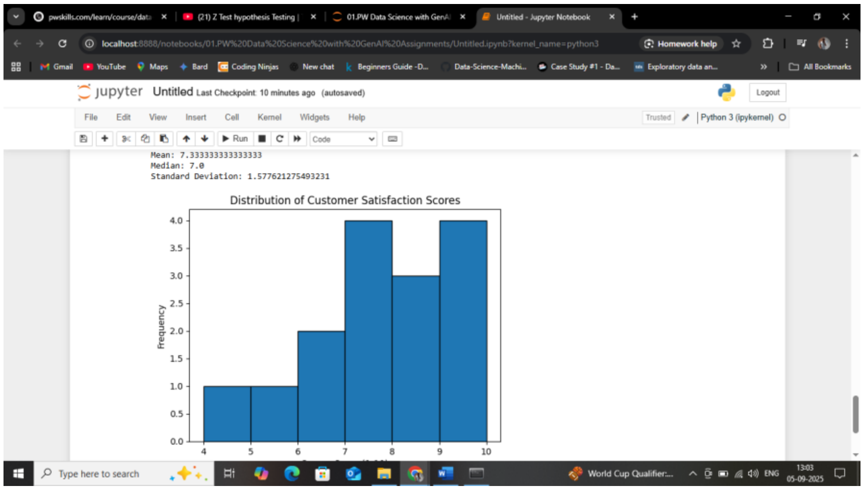
Answer:



```
In [6]: import numpy as np
        import matplotlib.pyplot as plt

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

        mean = np.mean(survey_scores)
        median = np.median(survey_scores)
        std_dev = np.std(survey_scores)

        print("Mean:", mean)
        print("Median:", median)
        print("Standard Deviation:", std_dev)

        plt.hist(survey_scores, bins=6, edgecolor='black')
        plt.title("Distribution of Customer Satisfaction Scores")
        plt.xlabel("Survey Score (1-10)")
        plt.ylabel("Frequency")
        plt.show()
```

```
Mean: 7.333333333333333
Median: 7.0
Standard Deviation: 1.577621275493231
```

jupyter Untitled Last Checkpoint: 10 minutes ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
Logout
Distribution of Customer Satisfaction Scores
4.0



```
Mean: 7.333333333333333
Median: 7.0
Standard Deviation: 1.577621275493231
```

jupyter Untitled Last Checkpoint: 10 minutes ago (autosaved)
Distribution of Customer Satisfaction Scores
Frequency
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
4
5
6
7
8
9
10