PW SKILLS

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 Statistics:

- Whenever we describe the whole data it is called descriptive statistics. It consists of organizing and summarizing the whole data/population. - Talks about always complete data/population.

- Whenever we want exact information/statistics to make any business decision , we will use descriptive stat.

- Types of descriptive stat : measure of central tendency, measure of dispersion,measure of symmetry and skewness, covariance and correlation. - In eda using the describe function[pandas] you can see what the mean of data is,variance tells about spreads, using box plot or distribution you see outliers are present or not / data is left or right skewed so you can apply transformation. - Example : average height of the student in 10th class, average rainfall in this year,strike rate of player

Inferential Statistics:

- Inferential stats help us make predictions or decisions about larger populations based on sample data.

- Inferential stat talk about the conclusion that is made for the population made from using samples.

- When population is large and we have no time and resources then use - Types of inferential stat : testing of hypothesis , confidence interval , regression - Use in Eda:

1. **hypothesis testing you check differences or relationships in data are statistically significant [example : Is the average income of two cities really different?]**

2. Using confidence intervals Estimate population parameters with a margin of error. [Predict average customer age with 95% confidence]
3. Identify which features truly impact the target variable.

- Example : avg weight and height of population in india (you take sample from state and calculate avg,it's true for all population)

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

Answer:

Sampling statistics is the process of selecting a subset (sample) from a larger group (population) in order to draw conclusions or make inferences about the entire population. Since it is often impractical / impossible to collect data from every member/element of a population, sampling provides a manageable way to gather information.

Simple Random Sampling :

- **Every member in a population(n) has an equal chance of being selected in the sample. [equal chance for all members]**
- **Equal probability of every person being selected for a sample.**
- **Not equal representation across all groups.**
- **Use case : homogenous population**
- **Example : class of 1000 students and randomly pick 150 names using a random number generator it is called simple random sampling.**
- **Pros:**
    1. **easy to implement**
    2. **minimize selection bias**
- **Cons:**
    1. **May not accurately represent subgroups within the population.**
    2. **Possibility of sample and members not being part of sample from a certain group.**
       Example : among selecting 10 cr people in Maharashtra while taking a sample using a simple random sample from a smaller city and least populated area may be not selected for sampling.

Stratified Sampling:

- **The population is divided into distinct subgroups or strata (state,region,colour,gender,age,income) and then a random sample is taken from each stratum.[equal chance for within each stratum]**
- **Equal probability of every person being selected for a sample is only possible in proportional stratified random sampling, not in all types of stratified sampling. - Give equal chance of representation.**
- **Use case: Heterogeneous population with distinct groups**
- **Example : In class it has 60% girls and 40% boys , you want a sample of 50 students, you might select 30 girls and 20 boys randomly from each group. - Pros:**
    1. **Ensures representation of all key subgroups.**
    2. **More precise estimates compared to simple random sampling.**
- **Cons:**
    1. **Requires detailed population information.**
    2. **More complex to design and implement.**

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

Answer:

Mean:

- **The mean is the sum of all values divided by total number of values.**
  Data = 1,2,3,4,5,6
  Mean = 1+2+3+4+5+6/6=3.5
  Population mean = mu = summation x_i / N
  Sample mean = x_bar = summation x_i / n
- **It gives numerical summary of data overall level.**
- **Gives general avg**
- **Mean is arithmetic midpoint of data.**
- **Mean is required for calculating var,standard deviation, testing,normal distribution etc**
- **The mean shows where most value tends to cluster**
- **The mean is often used to compare different groups or categories. - Whenever outliers are present in the dataset,the mean is not a good representative. [affected by outlier]**
- **Example :**
  1. **If the average income in a dataset is 50,000, that gives you a rough idea of what people in that dataset typically earn.**
  2. **Mean test scores of two different classes.**

Median:

- **The median is the middle value in a sorted list of numbers.**
- **Median is physical midpoint of data**
- **Median of even number:**
  Data: 6,3,1,2,4,5
  Sort data : 1,2,3,4,5,6
  Median = 3+4/2=3.5

Median of odd number:

Data = 1 2 3
Median = 2

- **The median isn't affected by outliers.**
- **Using median input missing value when data is skewed.**
- **In the box plot [Q2= is median] seeing median you can understand either data is left skew or right skew.**
- **Example : In a hospital dataset, we might look at the median wait time for patients in the emergency room. If most patients wait 20–30 minutes but a few wait 5 hours, the mean will be misleading .The median tells us what most patients actually experience — a crucial insight for improving service quality.**

Mode:

- **The mode is the value that appears most frequently in a dataset. - It identifies the most common category or popular item.**
- **Data : 7 4 3 2 4 4 3 mode=4**
- **One mode : unimodel , two mode: bimodel, three mode: multimodal**
  1. **Bimodal :**
     - **In bimodal data,we report both values as modes. This tells us there are two dominant clusters or preferences.**
  2. **Multimodal:**
     - **Multiple modes often suggest multiple distinct subgroups. Or reveal pattern in the data**
     - **You can split the dataset based on these modes for targeted analysis.**
     - **Example : Age groups in a customer base — teens, mid-30s, retirees.**

- **mode useful for categorical,ordinal and discrete numerical data - Help to detect dominant patterns, common behaviour or common choice - Impute missing value of categorical column using mode.**
- **Example:**
  1. **In an e-commerce platform, We analyzed users' purchase history to recommend products. For each user, we identified the most purchased frequent category - like sportswear using mode. This helped personalize recommendations by showing trending items from that category. For eg, if a users mode category is sportswear, the system prioritizes new arrivals or bestsellers in that segment.**
  2. **On a video streaming platform, we used mode to find the most watched genre per user like comedy. This became the anchor for personalized recommendations. Even if the user watched multiple genres, the mode gave a clear signal of preference, which improved click-through rates and engagement.**

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

Answer:

Skewness and kurtosis tell us about the shape of data.

Skewness:

1. Skewness quantifies how asymmetric a distribution is around its mean. - Skewness = 1/n summation [(x_i - mu) / sigma] ^ 3

- Pearson 2n coefficient = 3 (mean - median) / sd

2. You measure symmetry using skewness :

- no skewness means symmetric data

- Positive skew/ Right skew : tail on the right side

Mean > median and q3- q2 > q1-q2

- Negative skew/ Left Skew : tail on the left side

Median > mean and q2-q1 > q3-q2

3. Helps decide whether to use mean or median

4. guide data transformation [eg: lognormal, box cox,square root etc] 5. you can seen skewness using box plot,dist plot

6. Machine learning models require symmetric data to build models. 7. Example :

- distribution of wealth in India.

- Imagine monitoring website traffic for an e-commerce platform. On most days, traffic is steady — say, 5,000 to 10,000 visitors. But during a flash sale or influencer shoutout, traffic suddenly jumps to 100,000 visitors in one hour. This creates a right skew distribution The mean traffic is pulled up by the spike, but the median remains closer to normal traffic levels. In this case, skewness helps us detect anomalies and plan for load

balancing or server scaling.

8. Using skewness we can build many systems like App usage during cricket finals or festival launches, News site traffic during breaking events etc.

Kurtosis :

1. Kurtosis measures the tailedness or peakedness of distribution. how heavy or light the tails are compared to a normal distribution.

2. It helps detect outliers and distribution shape, which are critical in

model performance and robustness.

3. Mesokurtic == 3 [equal to normal distribution]

4. Leptokurtic [>3] = Heavy tails, sharp peak (more outliers).

5. Platykurtic [<3] = Light tails, flat peak (fewer outliers).

6. Use in Eda:

- **identified outlier prone distribution**
- **Influences model selection and robustness check - help in risk analysis**
- **help in feature engineering: decide whether to transform or normalize features.**

7. Example :

- **While building a regression model to predict customer lifetime value, we checked the kurtosis of the target variable. It was leptokurtic, indicating heavy tails and potential outliers. This insight helped us choose a robust loss function like Huber loss instead of MSE, which is sensitive to outliers.**
- **In a deep learning pipeline for fraud detection, We analyzed the kurtosis of transaction amounts. High kurtosis revealed rare but extreme fraudulent transactions. This guided us to use stratified sampling and custom evaluation metrics that emphasize tail performance, ensuring the model doesn't ignore rare but critical cases.**

What Does a Positive Skew Imply?

Positive skew means data has a long right tail. Most values are low,but a few high outliers pull the mean upward.This is common in income data, internet traffic spikes, or product sales during a viral campaign.here we can use different types of transformation like log or square root transformation to normalize the distribution.

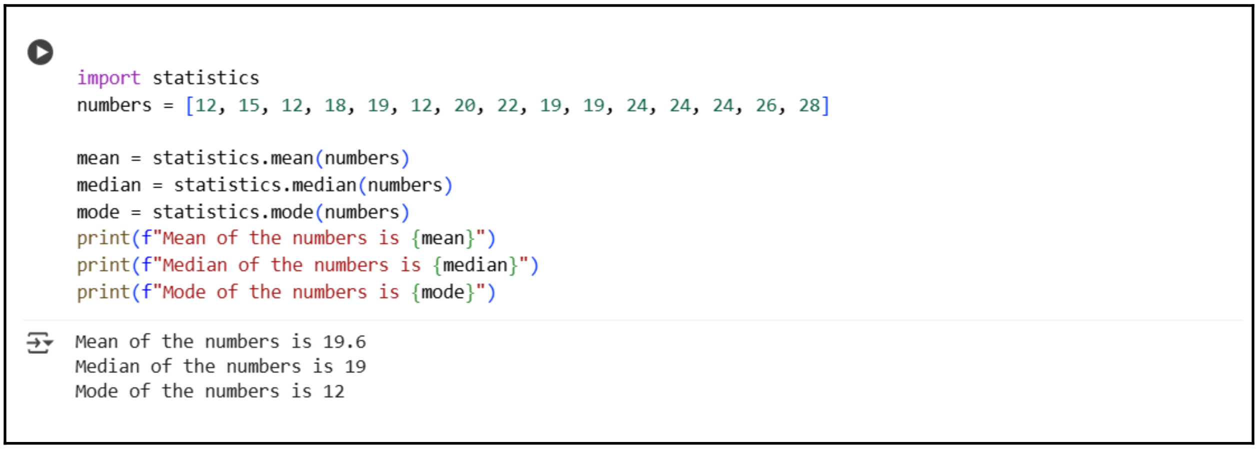
Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

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:

Paste your code and output inside the box below:



```
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(f"Mean of the numbers is {mean}")
print(f"Median of the numbers is {median}")
print(f"Mode of the numbers is {mode}")
```

```
Mean of the numbers is 19.6
Median of the numbers is 19
Mode of the numbers is 12
```

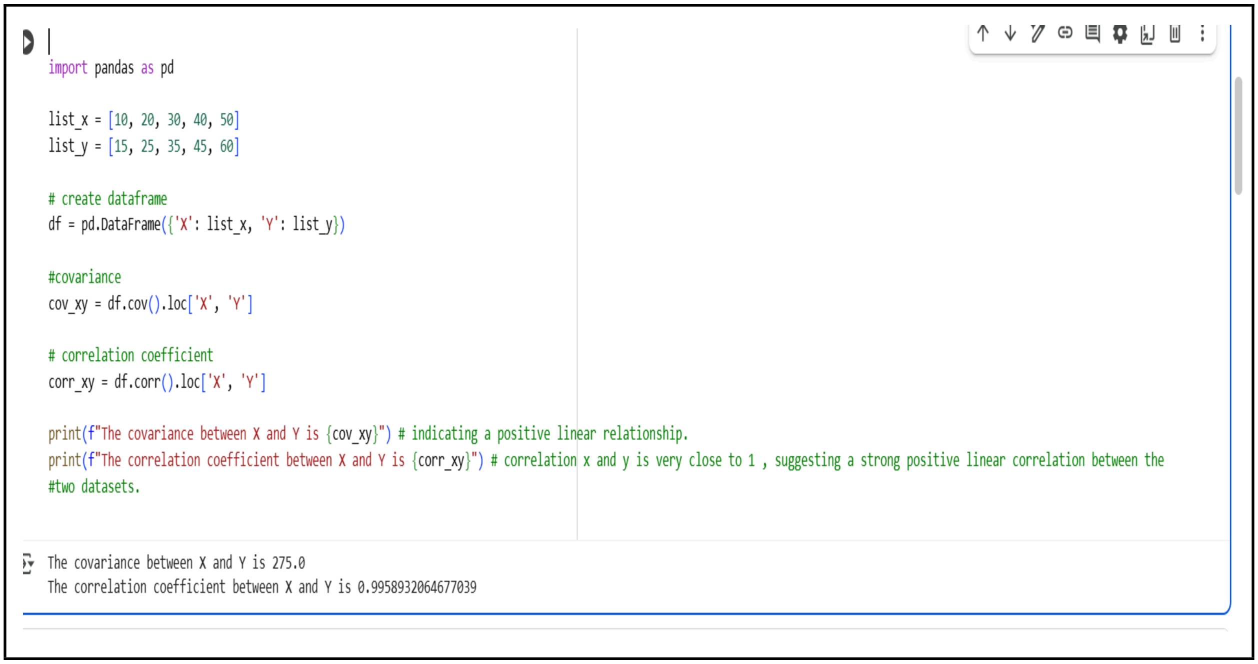
Question 6: Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

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:



```
import pandas as pd

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

# create dataframe
df = pd.DataFrame({'X': list_x, 'Y': list_y})

#covariance
cov_xy = df.cov().loc['X', 'Y']

# correlation coefficient
corr_xy = df.corr().loc['X', 'Y']

print(f"The covariance between X and Y is {cov_xy}") # indicating a positive linear relationship.
print(f"The correlation coefficient between X and Y is {corr_xy}") # correlation x and y is very close to 1 , suggesting a strong positive linear correlation between the
#two datasets.
```

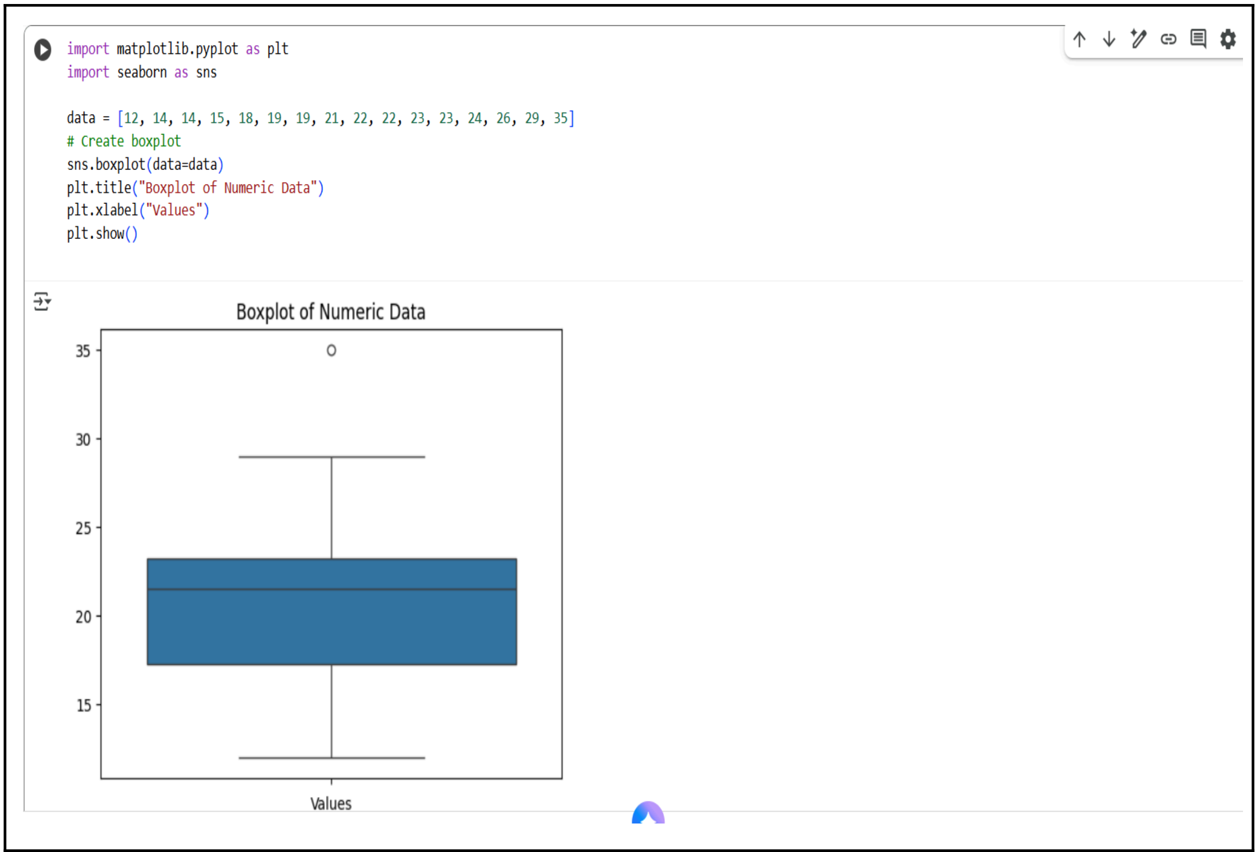
```
The covariance between X and Y is 275.0
The correlation coefficient between X and Y is 0.9958932064677039
```

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

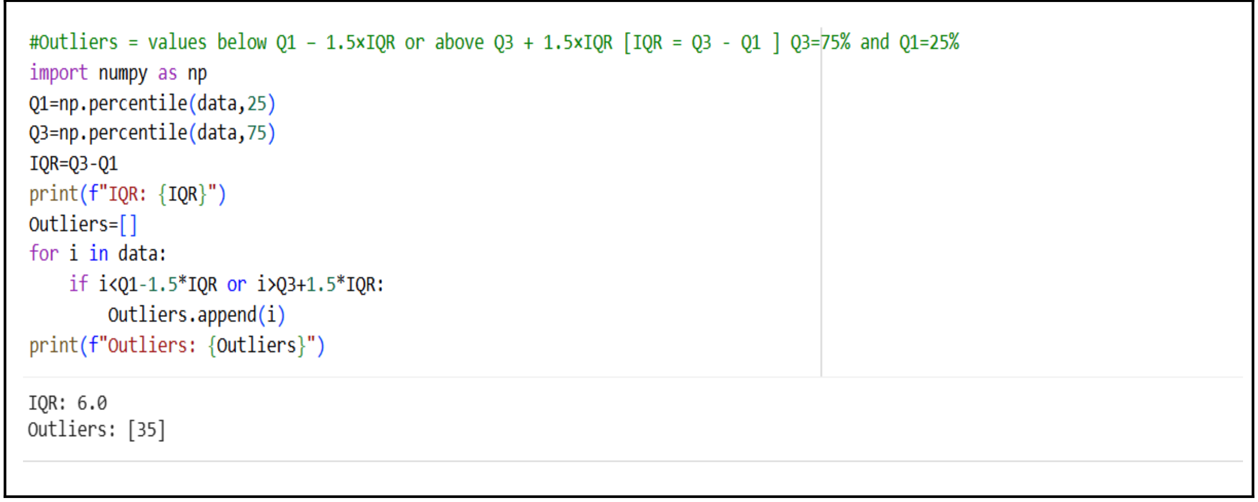
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

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

Answer:



```python
import matplotlib.pyplot as plt
import seaborn as sns

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
# Create boxplot
sns.boxplot(data=data)
plt.title("Boxplot of Numeric Data")
plt.xlabel("Values")
plt.show()
```



```python
#Outliers = values below Q1 - 1.5×IQR or above Q3 + 1.5×IQR [IQR = Q3 - Q1 ] Q3=75% and Q1=25%
import numpy as np
Q1=np.percentile(data,25)
Q3=np.percentile(data,75)
IQR=Q3-Q1
print(f"IQR: {IQR}")
Outliers=[]
for i in data:
    if i<Q1-1.5*IQR or i>Q3+1.5*IQR:
        Outliers.append(i)
print(f"Outliers: {Outliers}")
```

```
IQR: 6.0
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.*)

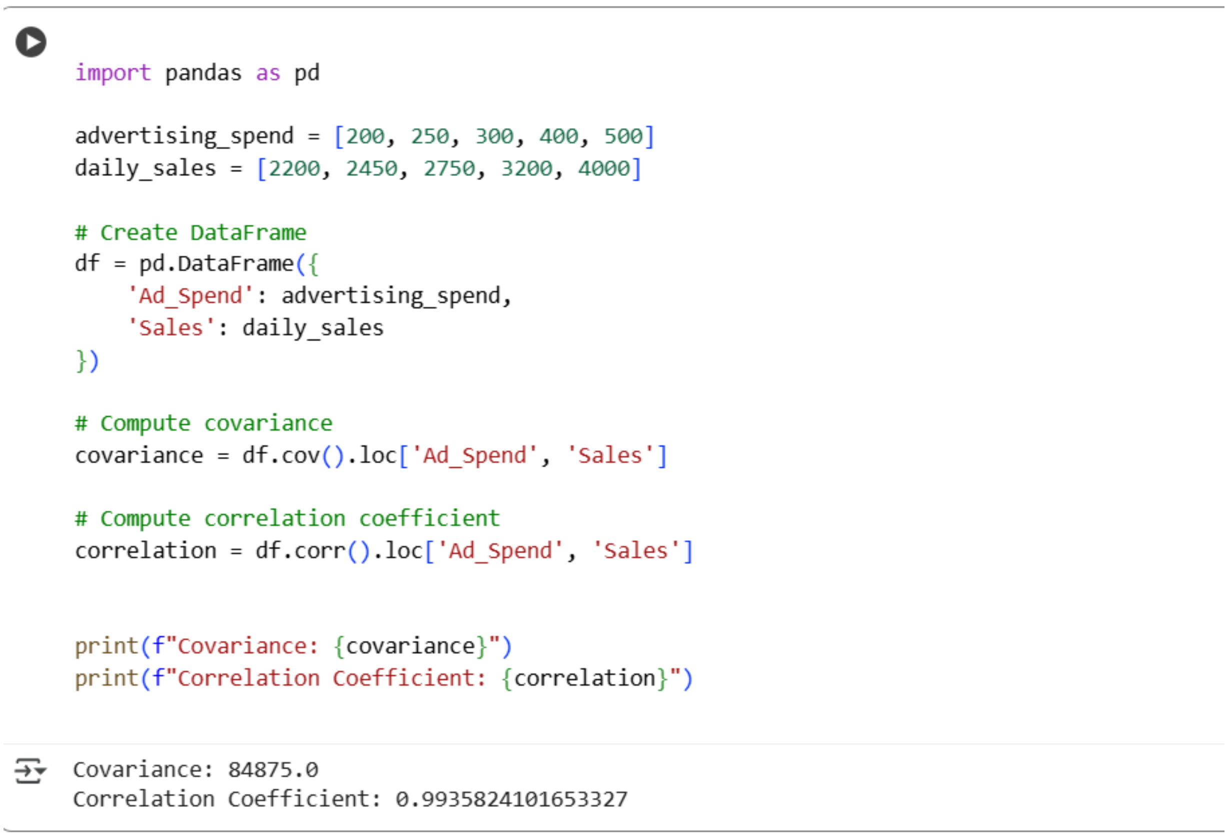
Answer:

Covariance :

1. Measure the direction of the relationship
2. If positive , it means higher ad spend tends to result in higher sales.
3. If negative , it means inverse relationship
4. But it is scale dependent, so not ideal for comparing across datasets.

Correlation :

1. Measures the strength and direction of the linear relationship.
2. Range -1 to 1.
    - 1 means perfect positive correlation
    - 0 means no correlation
    - -1 means perfect negative correlation
3. scale independent,making it ideal for interpretation.



```
import pandas as pd

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

# Create DataFrame
df = pd.DataFrame({
    'Ad_Spend': advertising_spend,
    'Sales': daily_sales
})

# Compute covariance
covariance = df.cov().loc['Ad_Spend', 'Sales']

# Compute correlation coefficient
correlation = df.corr().loc['Ad_Spend', 'Sales']


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

```
Covariance: 84875.0
Correlation Coefficient: 0.9935824101653327
```

Interpretation :

- Covariance positive means ad spend and sales move together
- Correlation will be close to 1 means a strong positive linear relationship
- it shows that increased advertising spend is strongly associated with higher daily sales,which supports marketing investment decision

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.

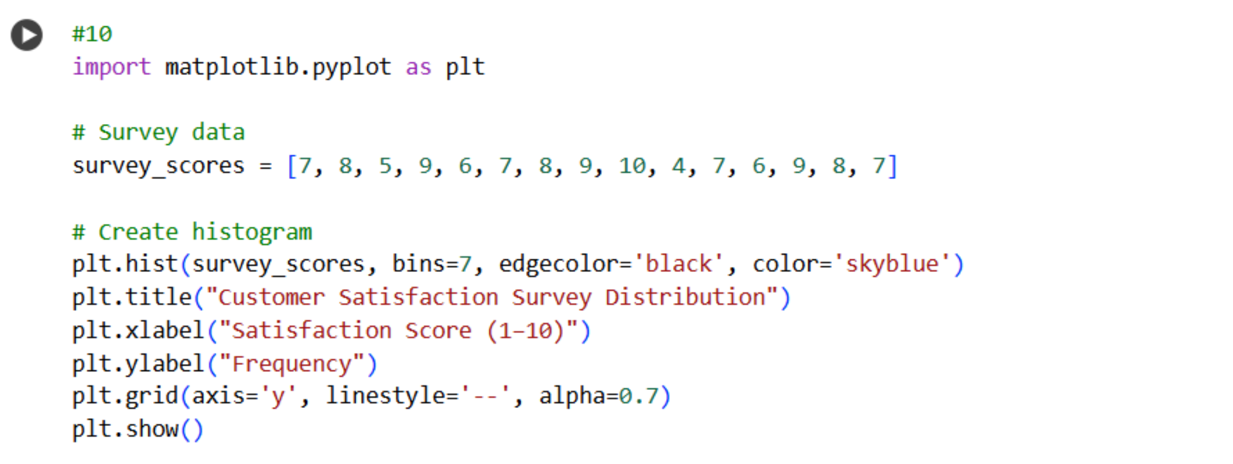
- Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you'd use.
- Write Python code to create a histogram using Matplotlib for the survey data:

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

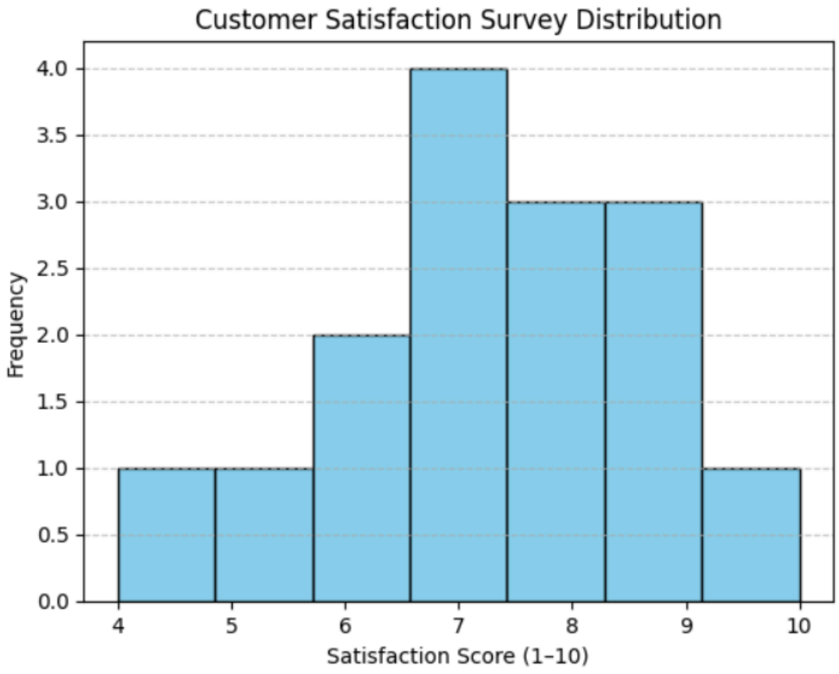
(*Include your Python code and output in the code box below.*)

Answer:

```
Mean : Average satisfaction level
Median: Middle score — robust to outliers
Mode:  Most common score
Standard deviation: Spread of scores — how consistent feedback is
Skewness and kurtosis: Shape and outlier sensitivity

Visualizations:

Histogram: Shows frequency of each score — reveals distribution shape
Boxplot: Highlights spread, median, and outliers
```



```
#10
import matplotlib.pyplot as plt

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

# Create histogram
plt.hist(survey_scores, bins=7, edgecolor='black', color='skyblue')
plt.title("Customer Satisfaction Survey Distribution")
plt.xlabel("Satisfaction Score (1-10)")
plt.ylabel("Frequency")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



Conclusion :

1. The histogram will likely show a peak around 7-9, indicating high satisfaction.

2. No extreme outliers — scores are within expected range.

3. If the distribution is right-skewed, it means most customers are satisfied.

4. If it's left-skewed, it may signal dissatisfaction and require deeper analysis