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	Inferential Statistics
Focuses on summarizing and describing the main features of a dataset.	Focuses on generalization and predictions about a population based on a sample.
Calculating measures like: Mean, Median & Mode and creating visualization like histograms and charts.	Using techniques like hypothesis testing, confidence intervals, and regression analysis.
To present a clear and concise view of the data.	To draw conclusions about a larger group based on a smaller subset.
Example: creating a bar graph to show the distribution of income levels in an organization.	Example: Estimating the average height of an adult in a city, based on a sample survey.

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

Answer:

In statistics, sampling is the way of selecting a subset of individuals from a larger population to study and draw conclusions about the entire population.

Random Sampling	Stratified Sampling
Every individual in the population has an equal and independent chance of being selected for the sample.	The population is divided into homogeneous subgroups(strata) based on specific characteristics, and then random samples are taken from each strata.
Assumes homogeneity	Only support homogeneity within subgroups.
It is more simple to implement	It is more complex and potentially more time consuming
Generally lower cost	Potentially higher cost due to complexity

---

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

Answer:

Mean: The mean is the arithmetic average of the certain data. Sum of all the numbers by total number of numbers.

Median: The median is the middle value of a dataset when all the data is arranged in ascending order.

Mode: It is the value that occurs most in the given dataset. Having highest number of occurrences in the dataset.

These measures of central tendency are important as:

1. It allows to abstract a large dataset into a single value, which makes it easy to grasp the overall characteristics of the data.
2. It enables the comparison of data with central tendency within any two or more datasets to see how they differ.
3. It is useful in Exploratory data analysis

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

Answer:

Skewness: Skewness measures the degree to which a distribution deviates from symmetry. A symmetric distribution has a skewness of zero.

Kurtosis: It measures the “tailedness” of the distribution’ tails to a normal distribution. It indicates how much data falls into tails.

- Positive skew indicates that the tail of graph is inclining towards the right side of the distribution. The mean is typically greater than the median.

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]

```python
import numpy as np
from scipy import stats
numbers =[12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

print(f"Mean: {np.mean(numbers)}")
print(f"Median: {np.median(numbers)}")
print(f"Mode: {stats.mode(numbers)}")
```

Output:

```
Mean: 19.6
Median: 19.0
Mode: ModeResult(mode=np.int64(12), count=np.int64(3))
```

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]

Answer:

```
import pandas as pd

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

df = pd.DataFrame({'x': list_x, 'y': list_y})
data1 = df.cov()
data2 = df.corr()
print("Covariance Matrix:")
print(data1)
print("\nCorrelation Matrix:")
print(data2)
```

Output:

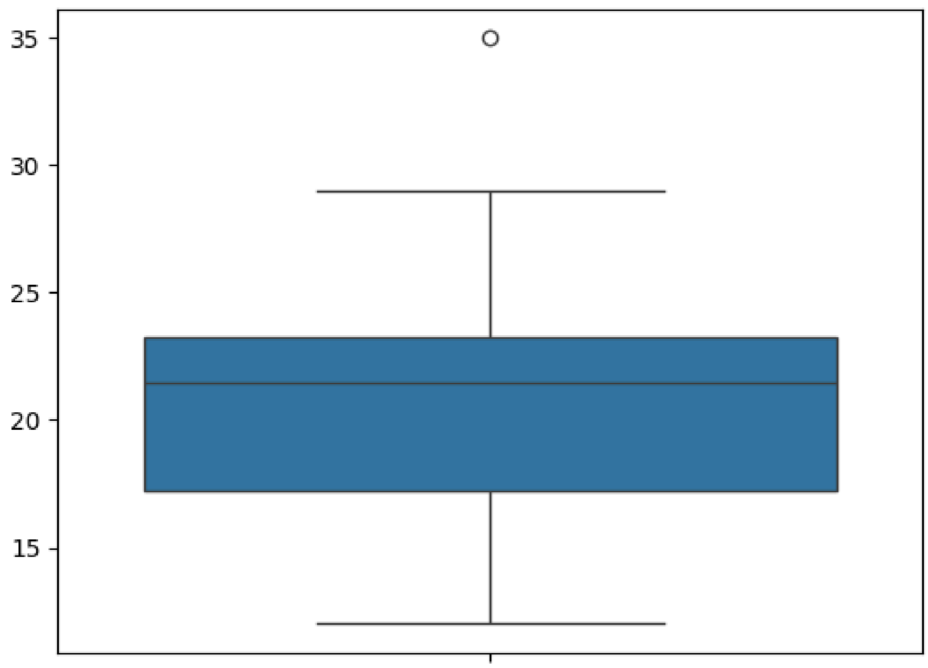
```
Covariance Matrix:
       x      y
x  250.0  275.0
y  275.0  305.0

Correlation Matrix:
          x         y
x  1.000000  0.995893
y  0.995893  1.000000
```

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]
Answer:

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

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29,
35]
sns.boxplot(data=data)
plt.show()

# Calculate Q1 and Q3
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)

# Calculate IQR
iqr = q3 - q1

# Define outlier bounds
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(f"Q1: {q1}")
print(f"Q3: {q3}")
print(f"IQR: {iqr}")
print(f"Lower bound for outliers: {lower_bound}")
print(f"Upper bound for outliers: {upper_bound}")
print(f"Identified outliers: {outliers}")
```

Output:

```
Q1: 17.25
Q3: 23.25
IQR: 6.0
Lower bound for outliers: 8.25
Upper bound for outliers: 32.25
Identified outliers: [35]
```



Explanation:
In the box plot:- 1. The bottom of the box is displaying Q1:17.25
2. The middle line inside the box is median: (between 20 to 25)
3. The top is showing Q3: 23.25
4. Outlier is (35), as it is above the upper whisker.
5. Lower whisker = around 9, upper whisker = 29

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]

Answer:

To know the relationship between advertising spend and daily sales,

1. using covariance will depict the measure of the relationship with respect to each other.On such a basis, if the measure shows positive covariance then it conveys that such will increase in any one of the parameters, the other will

increase simultaneously.Or else the measure will show negative covariance implying that any one of the parameters decreasing shows the other will increase.

2. Using correlation I will explore the strength & linear relationship between two parameters.The correlation coefficient ranges from -1 to +1.
   Likewise, if the correlation coefficient is more close to +1 shows strong linear relationship, which means there is a strong tendency that if advertising spend increases then, daily sales will obviously increase.
   - -1 shows strong negative relationship
   - 0 shows no linear relationship.

```
import pandas as pd

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


data = {'advertising_spend': advertising_spend, 'daily_sales':
daily_sales}
df = pd.DataFrame(data)
print(df)

correlation = df.corr(numeric_only = True)

print("Correlation:")
print(correlation)
```

Output:

```
advertising_spend  daily_sales
0                  200         2200
1                  250         2450
2                  300         2750
3                  400         3200
4                  500         4000
Correlation:
                    advertising_spend  daily_sales
advertising_spend            1.000000     0.993582
daily_sales                  0.993582     1.000000
```

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]

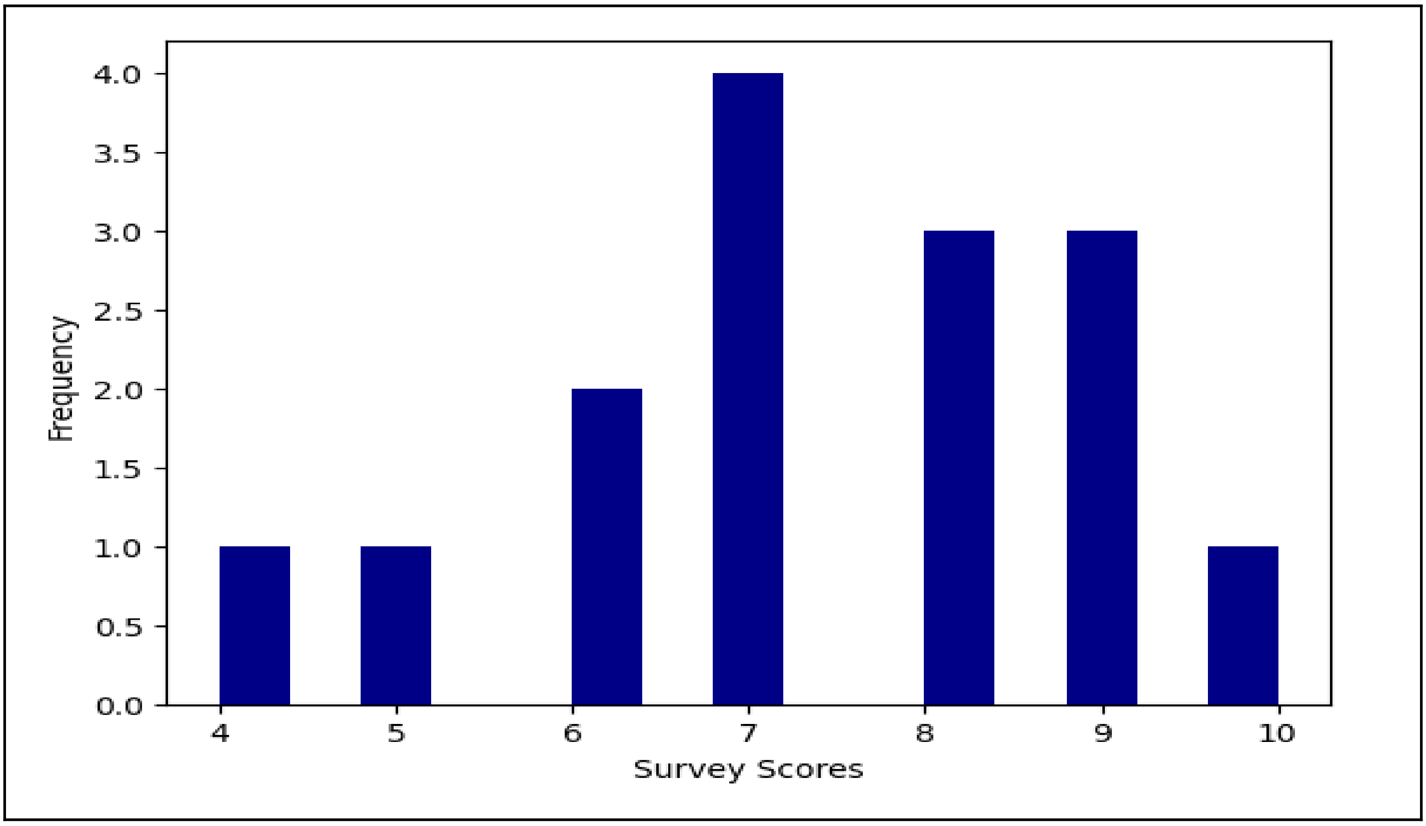
Answer:

In summary statistics:

1. Central Tendency : I will use **mean, median & mode** to analyze how much satisfied customers actually are.
2. Dispersion: In dispersion i will use range, percentile or Quartiles, to know the consistency in customer satisfaction and any Outliers can be found.
3. Skewness: To check the inclination of the happy or unhappy part.
4. Visualizations: to comprehend properly with graphs.

```
import matplotlib.pyplot as plt

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

plt.hist(survey_scores, bins=15, color='darkblue')
plt.xlabel("Survey Scores")
plt.ylabel("Frequency")
plt.show()
```

Output:


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Frequency
4
5
6
7
8
9
10
Survey Scores