❖ Python Coding

Beginner Level:

1. **Basic Data Manipulation:**
   - Write a Python function that takes a list of numbers and returns a list with each number squared.

```
def square_numbers(numbers):

    # Use a list comprehension to square each number in the input list

    return [x ** 2 for x in numbers]

# Example usage:

numbers = [1, 2, 3, 4, 5]

squared_numbers = square_numbers(numbers)

print(squared_numbers)
```

Explanation:

- The function `square_numbers` accepts a list of numbers as an argument (`numbers`).
- Inside the function, a list comprehension is used to iterate over each number in the input list and square it (`x ** 2`).
- The function then returns a new list containing the squared values.

This code will take the list `[1, 2, 3, 4, 5]` and return the squared values `[1, 4, 9, 16, 25]`.

   - Given a list of strings, return the longest string in the list.
   - Write a function that counts the occurrences of each element in a list and returns a dictionary with the count of each item.

```
def count_occurrences(lst):
    count_dict = {}
    for item in lst:
        if item in count_dict:
            count_dict[item] += 1  # Increment the count if item is already in the dictionary
        else:
            count_dict[item] = 1  # Initialize count to 1 if item is not in the dictionary
    return count_dict

      # Example usage:
      numbers = [1, 2, 2, 3, 3, 3, 4, 4, 4, 4]
      occurrences = count_occurrences(numbers)
      print(occurrences)
```

Explanation:

- The function `count_occurrences` takes a list (`lst`) as input.
- It initializes an empty dictionary (`count_dict`).

- The function then loops through each element in the list:
    - If the element is already in the dictionary, it increments its count.
    - If the element is not in the dictionary, it adds the element to the dictionary with an initial count of 1.
- Finally, the function returns the dictionary with the counts of each element.

This code will take the list `[1, 2, 2, 3, 3, 3, 4, 4, 4, 4]` and return a dictionary showing how many times each element occurs.

Pandas Basics:

- Load a CSV file using Pandas and display the first 5 rows of the dataset.
- Write a function that reads a CSV file and returns the total number of missing values in the dataset.
- Write a Python script that filters rows in a Data Frame where a column `age` is greater than 30.

2. **Data Filtering:**
    - Given a DataFrame with columns `Name`, `Age`, and `City`, filter rows where `Age` is greater than 25 and `City` is "New York".
    - Write a function that finds all unique values in a column of a Data Frame.

Intermediate Level:

4. **Aggregating Data:**
    - Write a function that computes the mean, median, and mode of a numerical column in a DataFrame.
    - Write a Python script that groups data by a specific column and computes the sum, average, and count of another numerical column.
5. **Data Merging:**
    - You have two DataFrames. One contains `Product_ID` and `Product_Name`, and the other contains `Product_ID` and `Sales`. Write a script to merge them on the `Product_ID` column.
    - Write a function that joins two DataFrames using a left join on a common column.
6. **Data Transformation:**
    - Write a Python function to create a new column in a DataFrame that contains the square of an existing column's values.
    - Write a function that normalizes a numerical column (scales the data to the range [0,1]).

Advanced Level:

7. **Data Visualization:**
    - Using Matplotlib, plot a histogram of a numerical column in a DataFrame.

```
import pandas as pd
import matplotlib.pyplot as plt

# Sample DataFrame with a numerical column
data = {'values': [10, 15, 10, 30, 25, 30, 10, 20, 25, 30, 35, 40, 30, 25, 20]}
df = pd.DataFrame(data)

# Plotting a histogram of the 'values' column
```

```
plt.hist(df['values'], bins=5, edgecolor='black')  # You can adjust the number of bins as needed
plt.title('Histogram of Values')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

Explanation:

- **`df['values']`**: This refers to the numerical column (`values`) in the DataFrame `df` that you want to plot.
- **`plt.hist()`**: This function is used to plot the histogram. You can specify:
  - The column to plot (`df['values']`).
  - `bins`: The number of bins you want to divide your data into (e.g., `bins=5`).
  - `edgecolor`: The color of the bin edges.
- **`plt.title()`, `plt.xlabel()`, `plt.ylabel()`**: These functions add a title and labels to the X and Y axes.

  - Create a scatter plot using Seaborn that shows the relationship between two numerical columns in a DataFrame.
  - Write a Python script to create a box plot for visualizing the distribution of a numerical column in a DataFrame.

8. **Handling Missing Data:**
   - Write a Python function to handle missing data in a DataFrame by either removing rows with missing values or filling them with the column mean.

```
import pandas as pd

def handle_missing_data(df, method='drop'):
    """
    Handles missing data in a DataFrame by either removing rows with missing values
    or filling them with the column mean.

    Parameters:
    df (pd.DataFrame): The input DataFrame with missing data.
    method (str): The method to handle missing data ('drop' or 'fill').
                  'drop' will remove rows with missing values.
                  'fill' will fill missing values with the column mean.

    Returns:
    pd.DataFrame: The DataFrame with missing data handled.
    """

    if method == 'drop':
        # Remove rows with any missing values
        return df.dropna()

    elif method == 'fill':
        # Fill missing values with the column mean
        return df.fillna(df.mean())

    else:
        raise ValueError("Method must be either 'drop' or 'fill'")

# Example usage:
data = {'A': [1, 2, 3, None, 5], 'B': [None, 2, 3, 4, 5], 'C': [1, None, None, 4, 5]}
df = pd.DataFrame(data)

print("Original DataFrame:")
print(df)

# Handle missing data by filling with column mean
df_filled = handle_missing_data(df, method='fill')
print("\nDataFrame after filling missing values with column mean:")
print(df_filled)

# Handle missing data by dropping rows with missing values
df_dropped = handle_missing_data(df, method='drop')
print("\nDataFrame after dropping rows with missing values:")
print(df_dropped)
```

Explanation:

- **`handle_missing_data(df, method='drop')`**:
    - The function accepts two parameters:
        - `df`: The input DataFrame with missing data.
        - `method`: A string that determines how to handle missing data. It can be `'drop'` to remove rows with missing values or `'fill'` to replace missing values with the column mean.
- **If `method` is `'drop'`**: The function removes rows containing any missing values using `df.dropna()`.
- **If `method` is `'fill'`**: The function fills missing values with the mean of the respective column using `df.fillna(df.mean())`.
- **`df.mean()`**: This calculates the mean for each column, which is then used to fill the missing values.
- • **Column Mean Filling**: When filling missing values with the column mean, the `df.mean()` function computes the mean for each column ignoring `NaN` values.
- • **Customization**: If you want to fill with another statistic (like median or mode), you can easily modify the function to use `df.median()` or `df.mode()`.

    - Given a DataFrame with missing values, write a script to predict and fill missing values using linear regression.

9. **Advanced Data Aggregation:**
    - Write a Python function that calculates the rolling average of a numerical column in a DataFrame with a window size of 5.
    - Write a function that creates a pivot table from a DataFrame, summarizing data by a specific categorical column.
10. **Time Series Analysis:**
    - Using a time series dataset, write a function that converts a date column into a pandas `Datetime` object and sets it as the index.
    - Write a Python script to plot the trend of a time series data over time using Matplotlib or Seaborn.

❖ Here are some basic theory questions related to data analytics that might be helpful for your semester exams:

1. What is Data Analytics?

- Explain the concept of data analytics. How does it differ from data analysis?
- What are the main types of data analytics (Descriptive, Diagnostic, Predictive, and Prescriptive)? Provide examples of each.

2. What is the importance of Data Cleaning?

- What is data cleaning, and why is it important in the data analytics process?
- What are some common techniques used in data cleaning (e.g., handling missing values, removing duplicates, etc.)?

3. What are the types of data in Data Analytics?

- Explain the difference between structured and unstructured data.
- What is semi-structured data? Give examples.
- What are the different data types in a database (integer, float, string, date, etc.)?

4. What is Exploratory Data Analysis (EDA)?

- What is the purpose of Exploratory Data Analysis (EDA)?
- What are some common techniques used in EDA?
- Explain the role of visualization in EDA.

5. What are the steps involved in a Data Analytics Project?

- List and describe the key steps involved in a data analytics project (Data Collection, Data Cleaning, Data Analysis, Data Interpretation, etc.).
- Why is it necessary to define the problem statement before starting a data analytics project?

6. What is Descriptive Analytics?

- Explain descriptive analytics and provide examples of its applications.
- How does descriptive analysis summarize data, and what types of tools are used in this type of analysis?

7. What is Predictive Analytics?

- What is predictive analytics, and how does it help in decision-making?
- How are machine learning models used in predictive analytics?

8. What is the Role of Statistics in Data Analytics?

- How do statistical methods contribute to data analytics?
- What are the key statistical concepts used in data analytics (e.g., mean, median, standard deviation, correlation, hypothesis testing)?

9. What are the Different Types of Data Visualizations?

- What are the various types of data visualizations, and when should each be used (e.g., bar charts, histograms, scatter plots, pie charts, etc.)?
- How do different visualizations help in understanding the underlying patterns in the data?

10. What is the concept of Big Data?

- Define Big Data and explain its 3 Vs (Volume, Velocity, Variety).
- How does Big Data differ from traditional data analytics in terms of processing and storage?

11. What is Machine Learning and its Role in Data Analytics?

- Define machine learning and explain its relevance in the context of data analytics.
- What are the different types of machine learning (supervised, unsupervised, reinforcement learning)? Provide examples of each.

12. What is Data Mining?

- What is the difference between data mining and data analytics?
- Explain some common data mining techniques, such as classification, clustering, and association rule mining.

13. What is Hypothesis Testing in Data Analytics?

- What is hypothesis testing, and why is it important in data analysis?
- Describe the steps in hypothesis testing (formulating hypotheses, choosing the test, calculating the test statistic, etc.).

14. What is Correlation and Causation in Data Analytics?

- Explain the difference between correlation and causation.
- How can you interpret a correlation coefficient, and what does it indicate about the relationship between two variables?

15. What is Data Warehousing?

- What is data warehousing, and why is it important for businesses?
- How does data warehousing support data analytics in organizations?

16. What are the challenges of Data Analytics?

- What are some common challenges faced during the data analytics process (e.g., data quality, data security, handling large volumes of data, etc.)?
- How can these challenges be addressed?

17. What is Data Privacy and Security in Data Analytics?

- Why is data privacy and security critical in data analytics projects?
- What are some common data protection laws and practices (e.g., GDPR, data encryption, etc.)?

18. What is the Role of a Data Analyst in an Organization?

- What are the key responsibilities of a data analyst in an organization?
- How does a data analyst support decision-making through data analytics?

19. What is Data Integration?

- What is data integration, and why is it important for data analysis?
- What are the common techniques used in integrating data from multiple sources?

20. What are the differences between Data Analytics and Business Intelligence?

- How does data analytics differ from business intelligence (BI)?
- What role does BI play in making data-driven decisions, and how do analytics complement it?

❖ Mean, Median, Mode, and Standard Deviation Theory Questions

1. **Define the following terms:**
   - Mean
   - Median
   - Mode
   - Standard Deviation
   - How are each of these measures of central tendency and dispersion useful in analyzing data?
2. **Explain the differences between mean, median, and mode.**
   - Under what circumstances would you use each measure of central tendency?
   - What effect do outliers have on the mean, median, and mode?
3. **What is the importance of standard deviation in data analysis?**
   - How is standard deviation related to the spread or variability of data?
   - What does a high standard deviation indicate about a dataset? What does a low standard deviation indicate?
4. **What is the relationship between mean, median, and mode in a normal distribution?**
   - How do the mean, median, and mode compare in a perfectly symmetrical, normal distribution?
5. **What are the steps to calculate the mean, median, and mode for a given dataset?**
   - Provide a brief explanation of how to calculate each of these statistics.
6. **How does the presence of outliers affect the mean, median, and mode?**
   - Give examples of how outliers can skew the results when calculating the mean but have less impact on the median or mode.
7. **How do you calculate the standard deviation of a sample vs. a population?**
   - What is the difference in the formulas for calculating the standard deviation of a sample and a population?
   - Why is there a difference, and why is it important?
8. **What is the coefficient of variation, and how is it related to standard deviation?**
   - Explain the concept of the coefficient of variation and how it helps in comparing the spread of two different datasets.

❖ Mean, Median, Mode, and Standard Deviation Application-Based Questions

1. **Calculate the Mean, Median, Mode, and Standard Deviation:** Given the dataset:

   $$\{10,12,12,13,15,16,18,20,22,22,25\}$$

   - Find the mean, median, mode, and standard deviation.
2. **Understanding the Effect of Outliers:** Consider the following dataset:

$$\{2,4,6,8,100\}$$

   - Calculate the mean, median, and standard deviation.
   - How does the outlier (100) affect the mean, median, and standard deviation?

3. **Comparison of Two Datasets:** Dataset 1: {10, 12, 13, 15, 17}
   Dataset 2: {2, 8, 15, 18, 25}
   - Calculate the mean, median, and standard deviation for both datasets.
   - Which dataset has higher variability, and why?
4. **Application of Standard Deviation in Business:** In a company, the monthly sales revenue (in thousands) for 5 months is:

   $$\{50,60,55,45,70\}$$

   Calculate the mean and standard deviation of the sales revenue.

   - Based on the standard deviation, what can you infer about the consistency of the company's sales?
5. **Identifying Skewness in Data:** Given the following dataset:

   $$\{1,2,3,3,4,5,6,8,9,10\}$$

   - Calculate the mean, median, and mode.
   - Does the dataset show any skewness based on the relationship between the mean and median?
6. **Impact of Changing Data Values:** Dataset: {5, 7, 9, 10, 15}
   - Calculate the mean, median, and standard deviation.
   - If the value 15 is replaced with 30, how does it affect the mean, median, and standard deviation?
7. **Comparison of Two Distributions:** Dataset 1: {2, 4, 6, 8, 10}
   Dataset 2: {1, 5, 7, 9, 11}
   - Calculate the mean and standard deviation for both datasets.
   - Which dataset has more variability?
8. **Median and Mode in Skewed Data:** Given the dataset of monthly salaries (in thousands) for 7 employees:

   $$\{25,30,35,40,50,75,80\}$$

   - Find the median and mode.
   - If the data is skewed, how does this affect the choice of the best measure of central tendency?

Example Problem with Full Calculation

Problem:

Given the data set: $5, 8, 12, 14$, find the **sample standard deviation**.

Solution:

1. **Find the Mean:**

$$\bar{x} = \frac{5 + 8 + 12 + 14}{4} = \frac{39}{4} = 9.75$$

2. **Find the Squared Deviations:**

$$(5 - 9.75)^2 = 22.5625$$

$$(8 - 9.75)^2 = 3.0625$$

$$(12 - 9.75)^2 = 5.0625$$

$$(14 - 9.75)^2 = 18.0625$$

3. **Sum of Squared Deviations:**

$$22.5625 + 3.0625 + 5.0625 + 18.0625 = 48.75$$

4. **Sample Variance:**

$$s^2 = \frac{48.75}{4 - 1} = \frac{48.75}{3} = 16.25$$

5. **Sample Standard Deviation:**

$$s = \sqrt{16.25} \approx 4.03$$