SET 1

1. Scenario: Your HR department has provided you with an employee dataset containing columns like EmployeeID, Department, Salary, and Hire Date. Utilize Pandas data frames to perform the following tasks:

Question:

- Determine the highest and lowest salaries in each department.

- Calculate the average salary in the company.

- Identify employees who were hired in a specific year different events.

2. A clinic wants to know the most common health issues among their patients. They have a list of all the health issues that their patients have been diagnosed with in the past year, along with the number of patients who have been diagnosed with each issue.

Question: Write a program that will calculate the frequency distribution of health issues and print out the most common health issue using the following dataset:

DISEASE_NAME	DIAGNOSED_PATIENTS
Hypertension	250
Asthma	180
Depression	160
Arthritis	140
Migraine	80

3. **Scenario**: A retail store wants to know if there is a correlation between the number of customers they get and the amount of money they spend on promotions. They have data on the number of customers they had each month for the past year, as well as the amount of money spent on promotions each month.

Question: Write a program that will calculate the correlation coefficient between customers and promotional spending, and create a scatter plot of the data.

4. Python program to conduct bivariate analysis and construct a linear regression model for predicting house prices based on selected features, such as house size, using the dataset provided by a real estate company? Additionally, how can I ensure the accuracy and reliability of the model through performance evaluation?

DATASET: 'House Size (sqft)': [1000, 1500, 1200, 1800, 1350],

'Price (USD)': [200000, 300000, 240000, 350000, 280000]

5. Scenario: In a medical study, you have collected data on patients' recovery times after a procedure. Calculate the 10th, 50th, and 90th percentiles to understand the distribution of recovery times.

SET 2

1. Scenario: You are working on a data analysis project that involves analyzing the weekly sales and customer traffic data for a retail store. You have a dataset containing the weekly sales and customer traffic values for each week of a year. Your task is to develop a Python program that generates line plots and scatter plots to visualize the sales and customer traffic data.

- Question:

Develop a Python program to create a line plot of the weekly sales data.

Develop a Python program to create a scatter plot of the weekly customer traffic data.

2. **Scenario:** You are a data analyst working for a company that manufactures electronics. You have been tasked with analyzing the sales data for the past month. The data is stored in a NumPy array.

Question: How would you find the average revenue from all the products sold in the past month? Assume a 4x4 matrix with each row representing the sales for a different product.

3. You are a data scientist working in a pharmaceutical company. Your team wants to classify drug compounds based on their chemical properties to streamline research and development efforts. Your task is to perform hierarchical clustering to group drug compounds into clusters based on their molecular features. Write Python code to load the chemical data, preprocess it, apply hierarchical clustering, and visualize the hierarchical structure using dendrograms.

4.**Scenario:** You have collected data on the ages of customers in a retail store. Write a Python program to calculate and display the 25th percentile of customer ages.

5. You are working with an e-commerce company that has collected data on the purchase amounts made by customers over the past month. The dataset includes the purchase amounts (in dollars) for each transaction. Utilize measures of central tendency to answer the following questions:

- Calculate the mean (average) purchase amount to understand the typical spending behavior of customers.
- Identify the mode of the purchase amounts to find the most frequently occurring purchase amount, helping the company understand popular spending levels

SET 3

1.Scenario: You are managing inventory for a bookstore and need to calculate the total cost of a customer's purchase, including discounts and taxes. You have lists item_prices and quantities where each element corresponds to the price and quantity of an item purchased. The discount rate and tax rate are given as percentages.

Question: Use arithmetic operations to calculate the total cost of a customer's purchase, considering the discounts and taxes based on the item prices, quantities, discount rate, and tax rate.

2. You are a data scientist working in a healthcare organization. Your team wants to predict patient readmission risks based on medical history and treatment details. Your task is to build a classification and regression trees (CART) model for readmission prediction. Write Python code to load the patient data, preprocess it, split it into training and testing sets, train a CART model, and visualize the decision tree for interpretation.

3. Scenario: You are a financial analyst working with an investment portfolio. You need to calculate the total value of investments after applying fees and taxes. You have the following data:

- List of investment values in USD.
- Fee rate as a percentage of the total investment value.
- Tax rate as a percentage applied to the net investment value after deducting fees.

Question: Use arithmetic operations to calculate the net value of investments after deducting fees and applying taxes, given the investment values, fee rate, and tax rate?

4. Scenario: You are a data analyst working for a car manufacturing company. As part of your analysis, you have a dataset containing information about the fuel efficiency of different car models. The dataset is stored in a NumPy array named fuel_efficiency, where each element represents the fuel efficiency (in miles per gallon) of a specific car model. Your task is to calculate the average fuel efficiency and determine the percentage improvement in fuel efficiency between two car models.

Question: How would you use NumPy arrays and arithmetic operations to calculate the average fuel efficiency and determine the percentage improvement in fuel efficiency between two car models?

5. Scenario: You are working with a dataset representing the daily sales of a product over the past month. Calculate the variance of the daily sales to understand how much the sales figures deviate from the mean

SET 4

1. You are working as a financial analyst for a university. The university administration wants to analyze the monthly expenses for different departments to better understand their spending patterns and make informed budgeting decisions. You have collected the following data for four departments (Science, Arts, Engineering, and Business) over four months. The dataset is represented as follows, where each row corresponds to a different month and each column represents a department:

- **Month 1**: [10000, 12000, 11000, 9000]
- **Month 2**: [15000, 14000, 13000, 16000]
- **Month 3**: [9000, 9500, 10000, 11000]
- **Month 4**: [12000, 11000, 12500, 13000]

Question: Using NumPy functions, how would you calculate both the variance and covariance matrix of the monthly expenses for the different departments?

2. You are a data scientist working in an e-commerce company. Your team wants to classify customer reviews as positive or negative to analyze sentiment and improve customer service. Your task is to build a support vector machine (SVM) classifier for sentiment analysis. Write Python code to load the review data, preprocess it, split it into training and testing sets, train an SVM classifier, and evaluate its performance using metrics such as accuracy and F1-score.

3. A company wants to know if there is a correlation between the number of sales they make and the amount of advertising they spend. They have data on the number of sales they made each month for the past year, as well as the amount of advertising they spent each month. Write a program that will calculate the correlation coefficient between sales and advertising, and create a scatter plot of the data.

4. You are a teacher who wants to keep track of your students' exam scores for different subjects. You have collected the following data:

- Each student is identified by a unique student ID.
- For each student, you have their scores in three subjects: Math, Science, and English.
- Write a program that creates a DataFrame that displays each student's scores with the student IDs as index labels as follows

	Math	Science	English
Student1	85	92	78
Student2	90	88	85
Student3	75	95	80

5. Scenario: You are investigating a dataset representing the daily temperatures in a city. Calculate the variance and identify potential outliers that may indicate unusual weather conditions.

SET 5

1. calculate the variance of monthly rainfall in a region's dataset and identify potential outliers that may indicate unusual weather patterns using NumPy, given the following data?

DATASET: 20, 22, 24, 26, 28, 30, 32, 34, 36, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70

2. You are a data scientist working in a financial services firm. Your team wants to predict credit card fraud based on transaction patterns and customer information. Your task is to build a logistic regression model to predict fraudulent transactions. Write Python code to load the transaction data, preprocess it, split it into training and testing sets, train a logistic regression model, and handle class imbalance using techniques such as oversampling or undersampling.

3. Scenario: You are a financial analyst at a credit card company analyzing transaction data to understand customer spending patterns. The dataset includes transaction amounts (in dollars) for each cardholder over the past month. Utilize measures of central tendency to answer the following questions:

Calculate the mean (average) transaction amount to assess the typical spending behavior of cardholders.

Identify the mode of the transaction amounts to determine the most frequently occurring spending level, aiding in understanding common transaction sizes.

4. Scenario: You are a cashier at a grocery store and need to calculate the total cost of a customer's purchase, including applicable discounts and taxes. You have the item prices and quantities in separate lists, and the discount and tax rates are given as percentages. Your task is to calculate the total cost for the customer.

Question: Use arithmetic operations to calculate the total cost of a customer's purchase, including discounts and taxes, given the item prices, quantities, discount rate, and tax rate?

5. Write a python program will take in a dataset containing daily temperature readings for each city over a year and perform the following tasks:

- Calculate the mean temperature for each city.
- Calculate the standard deviation of temperature for each city.
- Determine the city with the highest temperature range (difference between the highest and lowest temperatures).
- Find the city with the most consistent temperature (the lowest standard deviation).

SET 6

1. A music streaming service wants to analyze the popularity of different music genres listened to by users over the past month. They have a dataset containing the genres of songs played and the number of times each genre has been played. Write a Python program that calculates the frequency distribution of music genres and prints out the most played genre.

2. You are a data scientist working in a mobile app development company. Your team wants to segment users based on their app usage behavior to tailor personalized features and notifications. Your task is to perform k-means clustering to segment users into distinct groups. Write Python code to load the app usage data, preprocess it, apply k-means clustering, and visualize the clusters using scatter plots.

3. Scenario: A research institute conducted a study on the relationship between diet and health outcomes among a group of 18 randomly selected participants. They collected data on participants' ages and cholesterol levels with the following results.

Question:Calculate the mean, median, and standard deviation of ages and cholesterol levels using Pandas.

Create boxplots to visualize the distribution of ages and cholesterol levels.

Generate a scatter plot and a quantile-quantile (q-q) plot based on these two variables to explore their relationship and distribution characteristics.

4. Scenario: You are a data analyst at a retail chain tasked with analyzing customer transaction data to optimize pricing strategies. The dataset includes purchase amounts (in dollars) from customer transactions over the past month.

Questions:Calculate the mean (average) purchase amount to determine the typical transaction size and average spending behavior of customers.

Identify the mode of the purchase amounts to pinpoint the most frequently occurring transaction amount, aiding in understanding popular spending levels among customers.

5. Scenario: You are a data scientist working for a company that sells products online. You have been tasked with analyzing the sales data for the past month. The data is stored in a Pandas data frame.

Question: How would you find the top 5 products that have been sold the most in the past month?

SET 7

1. Dataset containing daily temperature data for a specific city. The temperature vector T, which stands for the dataset, has n observations that represent the daily temperatures over a specific time period.

a) Determine the daily temperature changes, represented by the vector D.

b) Use box plots to visualize the distribution of the daily temperature changes and to identify and quantify any outliers based on the interquartile range (IQR) method.

c) Discuss the potential impact of outliers on weather forecasting and analysis.

Consider factors such as prediction accuracy, climate trend analysis, and the reliability of statistical measures in the presence of outliers.

2. You are a data scientist working in a retail analytics firm. Your team wants to analyze customer purchase behavior to optimize product placement and promotions. Your task is to apply market basket analysis to identify which products are frequently bought together by customers. Write Python code to load the transaction data, preprocess it, mine frequent itemsets, and generate association rules to understand customer purchasing patterns.

3. Scenario: You are a climate scientist analyzing climate data from multiple weather stations across different regions. The dataset contains daily temperature readings for each station over the past year.

Calculate the average daily temperature for each weather station to understand regional climate patterns.

Compute the standard deviation of daily temperatures for each station to assess temperature variability.

Identify the weather station with the widest temperature range (difference between the highest and lowest temperatures) to study climate extremes.

Determine the weather station with the most stable temperatures, indicated by the lowest standard deviation, to assess consistency in climate conditions

4. Scenario: You are a financial analyst reviewing the daily returns of an investment portfolio over the past month. Your task is to calculate the variance of the daily returns to assess the volatility and risk associated with the portfolio.

5. Scenario: You are a data scientist working for a company that sells products online. You have been tasked with analyzing the sales data for the past month. The data is stored in a Pandas data frame.

Question: How would you find the top 5 products that have been sold the most in the past month?

SET 8

1. A health analyst wants to investigate whether the average recovery time of two different treatment methods for a specific illness is significantly different. For Treatment A, a random sample of 100 patients was collected, and the mean recovery time was found to be 6.2 days with a standard deviation of 1.5 days. For Treatment B, a random sample of 120 patients was collected, and the mean recovery time was found to be 5.8 days with a standard deviation of 1.2 days. Using a 95% confidence level, test the hypothesis that the mean recovery times of Treatment A and Treatment B are significantly different.

2. You are a data scientist working in a fintech startup. Your team wants to predict loan default risks based on borrower characteristics such as credit score, income, and loan amount. Your task is to build a logistic regression model to predict loan default probabilities. Write Python code to load the loan dataset, preprocess it, split it into training and testing sets, train a logistic regression model, and evaluate its performance on the test set using metrics such as accuracy, precision, recall, and F1-score.

3. Scenario: A public health organization wants to investigate the relationship between air pollution levels and respiratory diseases in a city. They have collected data on the annual average concentration of air pollutants and the number of respiratory disease cases reported each year.

Task: Write a program to calculate the correlation coefficient between air pollution levels and respiratory disease cases, and create a scatter plot of the data using the following mock dataset:

4. You work as a data analyst for a popular online streaming service that offers a variety of content, including movies, TV shows, and documentaries. Your company has collected viewership data over the past year and wants to analyze and visualize this data to gain insights into viewing trends, content performance, and user preferences. To understand which content categories are most popular, create line, scatter, and bar plots that display the distribution of viewership across different content categories. Each plot should represent a category, and the height of the bar should indicate the total viewership count for that category.

Question: Using Python, how would you create line, scatter, and bar plots to visualize the distribution of viewership across different content categories?

5. Scenario: You are a data scientist working for a company that sells products online. You have been tasked with analyzing the sales data for the past month. The data is stored in a Pandas data frame.

Question: How would you find the top 5 products that have been sold the most in the past month?

SET 9

1. A school wants to evaluate the performance of its students based on their scores in a recent exam. The exam scores (out of 100) for a sample of students are as follows: [65, 70, 72, 75, 80, 85, 90, 95, 100]. Calculate the 25th and 75th percentiles of the exam scores.

2. You are a data scientist working in an e-commerce platform. Your team wants to enhance product recommendation strategies by clustering customers based on their browsing and purchasing patterns. Your task is to build a clustering model, such as K-means, to segment customers into distinct groups. Write Python code to load the customer interaction data, preprocess it, apply K-means clustering for customer segmentation, and evaluate the model's performance using metrics such as silhouette score.

3. You are working on a data visualization project for a fitness tracker company. The company has collected data on the number of steps taken

by users each month. Your task is to develop a Python program that generates line plots and bar plots to visualize the monthly step count data.

Question:How would you develop a Python program to create a line plot of the monthly step count data? How would you develop a Python program to create a bar plot of the monthly step count data?

4. Scenario: You are working on a project that involves analyzing the sales performance of a company over the past four quarters. The quarterly sales data is stored in a NumPy array named sales_data, where each element represents the sales amount for a specific quarter. Your task is to calculate the total sales for the year and determine the percentage increase in sales from the first quarter to the fourth quarter.

Question: Using NumPy arrays and arithmetic operations calculate the total sales for the year and determine the percentage increase in sales from the first quarter to the fourth quarter?

5. Scenario: You are a manager at a retail store and need to calculate the total revenue from a recent promotion, including discounts and taxes. You have the sales prices and quantities sold for each item, and the discount and tax rates are provided as percentages. Your task is to compute the total revenue generated after applying discounts and taxes.

Question: Use arithmetic operations to calculate the total revenue from the promotion, including discounts and taxes, given the sales prices, quantities sold, discount rate, and tax rate?

SET 10

1. What is the average satisfaction score and 95% confidence interval for various vacation destinations, and which destinations receive the highest average scores? Additionally, how does the sentiment of travel reviews vary across different destinations?

2. You are a data scientist working in a retail company. Your team wants to analyze customer purchase data to identify patterns and trends in buying behavior. Your task is to preprocess the customer dataset, handle missing values, and perform Principal Component Analysis (PCA) to visualize the high-dimensional data in a lower-dimensional space. Write Python code to load the customer data, preprocess it, apply PCA, and visualize the data points in a scatter plot.

3. Scenario: You are working on a data visualization project and need to create basic plots using Matplotlib. You have a dataset containing the monthly sales data for a company, including the month and corresponding sales values. Your task is to develop a Python program that generates line plots and bar plots to visualize the sales data.

Question: How would you develop a Python program to create a line plot of the monthly sales data?How would you develop a Python program to create a bar plot of the monthly sales data?

4. Scenario: You are working on a project that involves analyzing a dataset containing information about houses in a neighborhood. The dataset is stored in a CSV file, and you have imported it into a NumPy array named house_data. Each row of the array represents a house, and the columns contain various features such as the number of bedrooms, square footage, and sale price.

Question: Using NumPy arrays and operations, how would you find the average sale price of houses with more than four bedrooms in the neighborhood?

5. Scenario: You are working on a project that involves analyzing student performance data for a class of 10 students. The data is stored in a NumPy array named student_scores, where each row represents a student and each column represents a different subject. The subjects are arranged in the following order: Math, Science, English, and History. Your task is to calculate the average score for each subject and identify the subject with the highest average score.

Question: How would you use NumPy arrays to calculate the average score for each subject and determine the subject with the highest average score? Assume 4x4 matrix that stores marks of each student in given order.