**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.

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. **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.

**SET 2**

1. Scenario: You are a data analyst working for an e-commerce company. The company has provided you with a dataset containing information about customer orders. The dataset includes columns such as OrderID, CustomerID, ProductID, Quantity, and TotalPrice. Your task is to use Pandas data frames to analyze and derive insights from the dataset.

* How would you load the dataset into a Pandas data frame?
* Create a new column called 'OrderDate' and extract the date information from the OrderID.
* Filter the data to show only orders placed by a specific customer (choose a CustomerID).
* Group the data by CustomerID and calculate the total amount spent by each customer.
* Visualize the distribution of TotalPrice using a histogram.

2. A weather station wants to know the most common types of weather in their area. They have a list of all the weather conditions that have occurred in the past year, along with the number of times each weather condition has occurred. Write a program that will calculate the frequency distribution of weather conditions and print out the most common weather type.

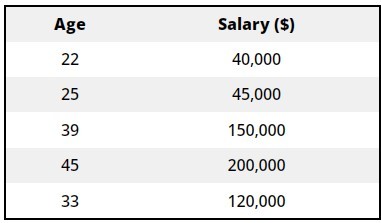
3. Suppose you are working as a data scientist for a medical research organization. Your team has collected data on patients with a certain medical condition and their treatment outcomes. The dataset includes various features such as age, gender, blood pressure, cholesterol levels, and whether the patient responded positively ("Good") or negatively ("Bad") to the treatment. The organization wants to use this model to identify potential candidates who are likely to respond positively to the treatment and improve their medical approach. Build a classification model using the KNN algorithm to predict the treatment outcome ("Good" or "Bad") for new patients based on their features. Evaluate the model's performance using accuracy, precision, recall, and F1-score. Make predictions on the test set and display the results.

4. Imagine you are an analyst for a popular online shopping website. Your task is to analyze customer reviews and provide insights on the average rating and customer satisfaction level for a specific product category. You will use the pandas library to calculate confidence intervals to estimate the true population mean rating. You have been provided with following dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **product\_title** | **product\_category** | **star\_rating** | **review\_headline** | **review\_date** |
| Pineapple slicer | Apparel | 4 | Really good | "2013-01-14" |
| Levis Jeans Pant | Apparel | 5 | Perfect Dress | "2014-04-22" |
| Wallet | Apparel | 5 | Love it | "2015-07-28" |
| Salwar | Apparel | 5 | Awesome | "2015-06-12" |

**SET 3**

1. Consider the age and salary of Individuals in a town. Find the correlation matrix, covariance, plot the on dataset and visualize giving an overview of relationships among data on sample dataset.



2. **Scenario**: You are a data analyst 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:** Develop a code in python to find the frequency distribution of the ages of the customers who have made a purchase in the past month.

3.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

4. 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 4**

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

* Determine the highest and lowest salaries in each department.
* Calculate the average tenure of employees in the company.
* Identify employees who joined before a specific date.

2. 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*

3 **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?

4. You work as a data analyst for a large e-commerce company that sells a variety of products online. Your company has collected sales data over the past year and wants to analyze and visualize this data to gain insights into sales trends, product performance, and customer behavior. To understand which product categories are most popular, create line, scatter and bar plot that displays the distribution of sales across different product categories. Each plot has to represents a category, and the height of the bar indicates the total sales amount for that category.

**SET -5**

**1. 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?

**2.** You work for a real estate agency and have been given a dataset containing information about properties for sale. The dataset is stored in a Pandas DataFrame named property\_data. The DataFrame has columns for property ID, location, number of bedrooms, area in square feet, and listing price. Your task is to analyze the data and answer specific questions about the properties.

Question: Using Pandas DataFrame operations, how would you find the following information from the property\_data DataFrame:

The average listing price of properties in each location.

The number of properties with more than four bedrooms.

The property with the largest area.

**3. Scenario:** You are a data analyst working for a social media platform. As part of your analysis, you have a dataset containing user interaction data, including the number of likes received by each post. Your task is to develop a Python program that calculates the frequency distribution of likes among the posts.

**Question:** Develop a Python program to calculate the frequency distribution of likes among the posts?

4. **Question:** You are a data scientist working for an e-commerce company. The marketing team has conducted an A/B test to evaluate the effectiveness of two different website designs (A and B) in terms of conversion rate. They randomly divided the website visitors into two groups, with one group experiencing design A and the other experiencing design B. After a week of data collection, you now have the conversion rate data for both groups. You want to determine whether there is a statistically significant difference in the mean conversion rates between the two website designs.

**Question:**

"Based on the data collected from the A/B test, is there a statistically significant difference in the mean conversion rates between website design A and website design B?"

**SET – 6**

**1. 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

**2.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?

**3.Scenario:** You are a data scientist working for a company that sells products online. You have been tasked with creating a simple plot to show the sales of a product over time.

**Question:**

a) Write code to create a simple line plot in Python using Matplotlib to predict sales happened in a month?

b). Write code to create a scatter plot in Python using Matplotlib to predict sales happened in a month?

c) Develop a Python program to create a bar plot of the monthly sales data.

**4. Question:** Evaluation Metrics for Model Performance

You have trained a machine learning model on a dataset, and now you want to evaluate its performance using various metrics.

Write a Python program that loads a dataset and trained model from scikit-learn. The program should ask the user to input the names of the features and the target variable they want to use for evaluation. The program should then calculate and display common evaluation metrics such as accuracy, precision, recall, and F1-score for the model's predictions on the test data.

**SET - 7**

1. **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?

2. **Scenario:** You are working on a text analysis project and need to determine the frequency

distribution of words in a given text document. You have a text document named "sample\_text.txt"

containing a paragraph of text. Your task is to develop a Python program that reads the text

document, processes the text, and generates a frequency distribution of the words.

**Question:** How would you develop a Python program to calculate the frequency distribution of

words in a text document?

3. A weather station wants to know the most common types of weather in their area. They have a list of all the weather conditions that have occurred in the past year, along with the number of times each weather condition has occurred. Write a program that will calculate the frequency distribution of weather conditions and print out the most common weather type.

4.**Question:** K-Means Clustering for Customer Segmentation

You are working for an e-commerce company and want to segment your customers into distinct groups based on their purchasing behavior. You have collected a dataset of customer data with various shopping-related features.

Write a Python program that allows the user to input the shopping-related features of a new customer. The program should use K-Means clustering from scikit-learn to assign the new customer to one of the existing segments based on the input features.

**SET – 8**

1. **Scenario:** You are a data scientist working for a company that sells products online. You have been tasked with creating a simple plot to show the sales of a product over time.

**Question:**

Write code to create a simple line plot in Python using Matplotlib to predict sales happened in a month?

Write code to create a scatter plot in Python using Matplotlib to predict sales happened in a month?

Develop a Python program to create a bar plot of the monthly sales data.

**2 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?

3 **Scenario:** You are working on a project that involves analyzing customer mar for a product. You have a dataset containing customer reviews, and your task is to develop a Python program that calculates the frequency distribution of words in the reviews.

**Question:** Develop a Python program to calculate the frequency distribution of words in the customer reviews dataset?

4. **Scenario:** You work as a data scientist for a marketing agency, and one of your clients is a large e-commerce company. The company wants to understand the purchasing behavior of its customers and segment them into different groups based on their buying patterns. The e-commerce company has provided you with transaction data, including customer IDs, the total amount spent in each transaction, and the number of items purchased.

**Question:** Build a clustering model using the K-Means algorithm to group customers based on their spending and purchase behavior and visualize the clusters using scatter plots or other appropriate visualizations to gain insights into customer distribution and distinguish different segments.

**SET – 9**

**1.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?

**2.Scenario:** You work for a real estate agency and have been given a dataset containing information about properties for sale. The dataset is stored in a Pandas DataFrame named property\_data. The DataFrame has columns for property ID, location, number of bedrooms, area in square feet, and listing price. Your task is to analyze the data and answer specific questions about the properties.

**Question:** Using Pandas DataFrame operations, how would you find the following information from the property\_data DataFrame:

The average listing price of properties in each location.

The number of properties with more than four bedrooms.

The property with the largest area.

3.**Scenario:** You are working on a project that involves analyzing customer reviews for a product. You have a dataset containing customer reviews, and your task is to develop a Python program that calculates the frequency distribution of words in the reviews.

**Question:** Develop a Python program to calculate the frequency distribution of words in the customer reviews dataset?

4. **Scenario:** You work as a data scientist for a retail company that operates multiple stores. The company is interested in segmenting its customers based on their purchasing behavior to better understand their preferences and tailor marketing strategies accordingly. To achieve this, your team has collected transaction data from different stores, which includes customer IDs, the total amount spent in each transaction, and the frequency of visits.

**Question:** Your task is to build a clustering model using the K-Means algorithm to group customers into distinct segments based on their spending patterns.

**SET – 10**

**1.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?

2. a) Write a code to plot a graph and implement the important functions using mat plot lib.

b) Write a code to print number of days between the months using NumPy library

3.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).

4. **Scenario:** You work as a data scientist for a real estate company. The company has collected data on various houses, including features such as the size of the house, number of bedrooms, location, and other relevant attributes. The marketing team wants to build a predictive model to estimate the price of houses based on their features. They believe that linear regression modeling can be an effective approach for this task.

**Question:** Your task is writing a Python program to perform bivariate analysis and build a linear regression model to predict house prices based on a selected feature (e.g., house size) from the dataset. Additionally, you need to evaluate the model's performance to ensure its accuracy and reliability.

**SET – 11**

1. **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:**

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

b) How would you develop a Python program to create a bar plot of the monthly sales data?

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

* + - Determine the highest and lowest salaries in each department.
    - Calculate the average tenure of employees in the company.
    - Identify employees who joined before a specific date.

**3.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

4. **Question**: Classification and Regression Trees (CART) for Car Price Prediction

You are working for a car dealership, and you want to predict the price of used cars based on various features such as the car's mileage, age, brand, and engine type. You have collected a dataset of used cars with their respective prices.

Write a Python program that loads the car dataset and allows the user to input the features of a new car they want to sell. The program should use the Classification and Regression Trees (CART) algorithm from scikit-learn to predict the price of the new car based on the input features.

The CART algorithm will create a tree-based model that will split the data into subsets based on the chosen features and their values, leading to a decision path that eventually predicts the price of the car. The program should output the predicted price and display the decision path (the sequence of conditions leading to the prediction) for the new car.

**SET - 12**

**1.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?

2. 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

**3. Scenario:** You are dealing with a dataset containing the monthly expenses of different departments in a company. Use NumPy functions to efficiently calculate both the variance and covariance matrix of these expenses.

4.**Question:** Linear Regression for Housing Price Prediction

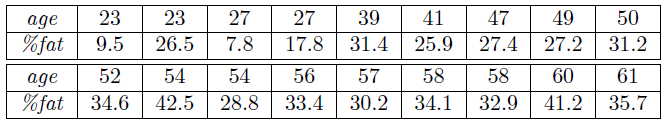
You are a real estate analyst trying to predict housing prices based on various features of the houses, such as area, number of bedrooms, and location. You have collected a dataset of houses with their respective prices.

Write a Python program that allows the user to input the features (area, number of bedrooms, etc.) of a new house. The program should use linear regression from scikit-learn to predict the price of the new house based on the input features.

**SET – 13**

1.. **Senario:** 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.

2. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.



**Question:**

* Calculate the mean, median and standard deviation of age and %fat using Pandas.
* Draw the boxplots for age and %fat.
* Draw a scatter plot and a q-q plot based on these two variables

**3.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.

4.**Question:** Logistic Regression for Customer Churn Prediction

You are working for a telecommunications company, and you want to predict whether a customer will churn (leave the company) based on their usage patterns and demographic data. You have collected a dataset of past customers with their churn status (0 for not churned, 1 for churned) and various features.

Write a Python program that allows the user to input the features (e.g., usage minutes, contract duration) of a new customer. The program should use logistic regression from scikit-learn to predict whether the new customer will churn or not based on the input features.

**SET - 14**

1. **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?

2.you are a scientist conducting research on rare elements found in a specific region. Your goal is to

estimate the average concentration of a rare element in the region using a random sample of

measurements. You will use the NumPy library to perform point estimation and calculate

confidence intervals for the population mean. The rare element concentration data is stored in a CSV

file named "rare\_elements.csv," where each row contains a single measurement of the

concentration.

**Question:**

write a Python program that allows the user to input the sample size, confidence level, and desired level of precision

**3.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?

4. **Scenario:** You are working on a data analysis project that involves analyzing the monthly temperature and rainfall data for a city. You have a dataset containing the monthly temperature and rainfall values for each month of a year. Your task is to develop a Python program that generates line plots and scatter plots to visualize the temperature and rainfall data.

**Question:**

1. Develop a Python program to create a line plot of the monthly temperature data.

2: Develop a Python program to create a scatter plot of the monthly rainfall data.

**SET – 15**

1. You are a data scientist working for a medical research institute. The institute is conducting a study to understand the relationship between smoking habits and the incidence of lung cancer among a group of individuals. As part of your analysis, you are tasked with calculating the correlation coefficient between smoking and lung cancer rates and creating a scatter plot to visualize the data.

Data = {

‘Smoking’: [20, 15, 5, 25, 30, 10, 18, 22, 8, 12],

‘Lung Cancer’: [5, 4, 1, 6, 8, 2, 3, 7, 1, 2]

}

1. Scenario: You are a data analyst working for an e-commerce company. The company has provided you with a dataset containing information about customer orders. The dataset includes columns such as Order ID, Customer ID, Product ID, Quantity, and Total Price. Your task is to use Panda’s data frames to analyze and derive insights from the dataset.

* How would you load the dataset into a Pandas data frame?
* Create a new column called 'Order Date' and extract the date information from the Order ID.
* Filter the data to show only orders placed by a specific customer (choose a Customer ID).
* Group the data by Customer ID and calculate the total amount spent by each customer.
* Visualize the distribution of Total Price using a histogram.

3**.Scenario**: You are a data analyst working for a sports analytics company. The company has

collected data on various soccer players, including their names, ages, positions, number of goals

scored, and weekly salaries. Create dataset on your own and store in a CSV file.

**Question**: Develop a Python program to read the data from the CSV file into a pandas data frame,

to find the top 5 players with the highest number of goals scored and the top 5 players with the

highest salaries. Also calculate the average age of players and display the names of players who are

above the average age and visualize the distribution of players based on their positions using a bar

chart.

1. **Scenario**: You are a data scientist working for an educational institution, and you want to

explore the correlation between students' study time and their exam scores. You have collected data

from a group of students, noting their study time in hours and their corresponding scores in an

exam.

**Question:** Identify any potential correlation between study time and exam scores and explore

various plotting functions to visualize this relationship effectively.

**Set -16**

1.You work as a data analyst for a large e-commerce company that sells a variety of products online. Your company has collected sales data over the past year and wants to analyze and visualize this data to gain insights into sales trends, product performance, and customer behavior. To understand which product categories are most popular, create line, scatter and bar plot that displays the distribution of sales across different product categories. Each plot has to represents a category, and the height of the bar indicates the total sales amount for that category

**2.Scenario:** You are a data analyst working for a marketing research company. Your team has collected a large dataset containing customer feedback from various social media platforms. The dataset consists of thousands of text entries, and your task is to develop a Python program to analyze the frequency distribution of words in this dataset. Your program should be able to perform the following tasks:

* Load the dataset from a CSV file (data.csv) containing a single column named "feedback" with each row representing a customer comment.
* Preprocess the text data by removing punctuation, converting all text to lowercase, and eliminating any stop words (common words like "the," "and," "is," etc. that don't carry significant meaning).
* Calculate the frequency distribution of words in the preprocessed dataset.
* Display the top N most frequent words and their corresponding frequencies, where N is provided as user input.
* Plot a bar graph to visualize the top N most frequent words and their frequencies.

**Question**: Create a Python program that fulfills these requirements and helps your team gain insights from the customer feedback data.

3. You are working for a retail company that maintains a sales dataset with columns like ProductID, SalesDate, UnitsSold, and Revenue. Use Pandas data frames to answer the following:

* Calculate the total revenue for each product.
* Identify the top-selling products based on UnitsSold.
* Find the average revenue per sale.

**4.Scenario:** You work as a data scientist for a marketing agency, and one of your clients is a large e-commerce company. The company wants to understand the purchasing behavior of its customers and segment them into different groups based on their buying patterns. The e-commerce company has provided you with transaction data, including customer IDs, the total amount spent in each transaction, and the number of items purchased.

**Question:** Build a clustering model using the K-Means algorithm to group customers based on their spending and purchase behavior and visualize the clusters using scatter plots or other appropriate visualizations to gain insights into customer distribution and distinguish different segments.

**SET -17**

**1. Scenario:** You are working as a data analyst for an e-commerce company. You have been given a dataset containing information about customer orders, stored in a Pandas DataFrame named order\_data. The DataFrame has columns for customer ID, order date, product name, and order quantity. Your task is to analyze the data and answer specific questions about the orders.

**Question:** Using Pandas DataFrame operations, how would you find the following information from the order\_data DataFrame:

The total number of orders made by each customer.

The average order quantity for each product.

The earliest and latest order dates in the dataset.

**2.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

3.Imagine you are an analyst for a popular online shopping website. Your task is to analyze customer reviews and provide insights on the average rating and customer satisfaction level for a specific product category. You will use the panda’s library to calculate confidence intervals to estimate the true population mean rating. You have been provided with following dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **product\_title** | **product\_category** | **star\_rating** | **review\_headline** | **review\_date** |
| Pineapple slicer | Apparel | 4 | Really good | "2013-01-14" |
| Levis Jeans Pant | Apparel | 5 | Perfect Dress | "2014-04-22" |
| Wallet | Apparel | 5 | Love it | "2015-07-28" |
| Salwar | Apparel | 5 | Awesome | "2015-06-12" |

4. Decision Tree for Iris Flower Classification

You are analyzing the famous Iris flower dataset to classify iris flowers into three species based on

their sepal and petal dimensions. You want to use a Decision Tree classifier to accomplish this task.

Write a Python program that loads the Iris dataset from scikit-learn, and allows the user to input the

sepal length, sepal width, petal length, and petal width of a new flower. The program should then

use the Decision Tree classifier to predict the species of the new flower.

**SET -18**

1.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

2. Scenario: You are a data analyst working for a finance company. Your team is interested in

analyzing the variability of stock prices for a particular company over a certain period. The

company's stock data includes the closing prices for each trading day of the specified period.

Question: Your task is to build a Python program that reads the stock data from a CSV file,

3. calculates the variability of stock prices, and provides insights into the stock's price movements.

A hospital wants to know if there is a correlation between the number of patients who smoke and the number of patients who get lung cancer. They have data on the number of patients who smoke and the number of patients who get lung cancer each year. Write a program that will calculate the correlation coefficient between smoking and lung cancer, and create a scatter plot of the data using following mock dataset.

|  |  |  |
| --- | --- | --- |
| **smoking\_patients** | **lung\_cancer\_patients** | **Year** |
| 200 | 25 | 2015 |
| 220 | 30 | 2016 |
| 240 | 35 | 2017 |
| 260 | 40 | 2018 |
| 300 | 55 | 2019 |

1. **Scenario:** You work as a data scientist for a retail company that operates multiple stores. The company is interested in segmenting its customers based on their purchasing behavior to better understand their preferences and tailor marketing strategies accordingly. To achieve this, your team has collected transaction data from different stores, which includes customer IDs, the total amount spent in each transaction, and the frequency of visits.

**Question:** Your task is to build a clustering model using the K-Means algorithm to group customers into distinct segments based on their spending patterns.