DSA0404 - Fundamentals of Data Science - Lab Questions

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. **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:

1. The total number of orders made by each customer.
2. The average order quantity for each product.
3. The earliest and latest order dates in the dataset.

3**. Scenario:** You are working for a retail company that has multiple stores across different regions. The company wants to analyze the sales data of their products to understand the performance across various stores and regions.

Question: You have a DataFrame sales_data with columns 'Product', 'Price', and 'Quantity'. Can you calculate the total sales for each product (Price * Quantity) and add a new column 'Total Sales'?

4. **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:

1. The average listing price of properties in each location.
2. The number of properties with more than four bedrooms.
3. The property with the largest area.

5. **Scenario:** You are working on a project that involves analyzing student performance data for a class of 15 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.

6. Scenario:You are working on a project that involves analyzing the trajectory of a projectile. You have recorded the time intervals and corresponding vertical positions of a projectile over a given time period. Using this data, you aim to calculate the average velocity of the projectile during its flight using NumPy.

Question: You have two NumPy arrays representing time intervals and vertical positions of the projectile. Calculate the average velocity of the projectile over this time interval. The vertical positions denote the height of the projectile at each corresponding time interval.

7. **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?

8. **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 NumPy array.

Question: How would you find the average price of all the products sold in the past month?

9. **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?

10. **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.

11. **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?

12. **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?

13. **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?

14. **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.

15. **Scenario:**You are a researcher working in a medical lab, investigating the effectiveness of a new treatment for a specific disease. You have collected data from a clinical trial with two groups: a control group receiving a placebo, and a treatment group receiving the new drug.Your goal is to analyze the data using hypothesis testing and calculate the p-value to determine if the new treatment has a statistically significant effect compared to the placebo. You will use the matplotlib library to visualize the data and the p-value.

16. **Scenario:** 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.

Question: Your task is to 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.

17. **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.

18**. Scenario:** You work for a weather data analysis company, and your team is responsible for developing a program to calculate and analyze variability in temperature data for different cities.

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

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

19. **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.

20. **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.

21. **Scenario:** You are working on a dataset that contains information about various types of fruits. The dataset includes features such as weight, color, and texture of the fruit. Your task is to build a k-Nearest Neighbors (kNN) classifier to predict the type of fruit based on these features.

Question: Given a dataset with features like 'weight', 'color', and 'texture' of fruits, and their respective 'type' (e.g., apple, orange, banana), how would you implement a k-Nearest Neighbors classifier to predict the type of an unknown fruit based on its 'weight', 'color', and 'texture' features? Additionally, discuss the process of choosing the optimal value of 'k' and handling categorical features (like 'color' or 'type') in a kNN classifier.

22. **Scenario**: You are tasked with implementing a decision tree classifier in Python to predict whether an online shopper will make a purchase on an e-commerce platform. The dataset provided includes attributes such as 'age', 'income', 'browsing_duration', 'device_type', and the target variable 'purchase' (indicating whether a purchase was made or not).

Question: Given the dataset with the mentioned attributes and the 'purchase' label, how would you use Python's scikit-learn library to create a decision tree classifier? Provide code to preprocess categorical variables like 'device_type' for model training and predict whether a new customer, with specific 'age', 'income', 'browsing_duration', and 'device_type', is likely to make a purchase or not.

23. **Scenario:** You work for a financial institution, and your task is to develop a classification model to assess the credit risk associated with loan applicants. The dataset provided contains various attributes such as income, credit score, debt-to-income ratio, employment duration, and the final 'risk' label indicating whether an applicant is high-risk or low-risk for a loan.

Question: Using the Classification and Regression Trees (CART) algorithm in Python, build a predictive model to evaluate the credit risk of loan applicants based on features like 'income',

'credit score', 'debt-to-income ratio', and 'employment duration'? Provide a Python code that preprocesses the data, builds a CART classifier, and predicts the credit risk level for a new loan applicant with specific attribute values

24. A company wants to know the most popular product they sell. They have a list of all the products they have sold in the past year, along with the number of times each product was sold. Write a program that will calculate the frequency distribution of products sold and print out the most popular product.

25. Given a dataset, write a program to perform estimation techniques such as mean estimation, variance estimation, and sampling techniques to infer population characteristics.

26. A weather station wants to know if there is a correlation between the temperature and the amount of rainfall in a city. They have data on the temperature and rainfall each day for the past year in that city. Write a program that will calculate the correlation coefficient between temperature and rainfall, and create a scatter plot of the data.

27. **Scenario**: Suppose you are a data analyst working for a marketing firm. The firm is interested in estimating the average revenue generated from a recent marketing campaign conducted on social media. You've collected a sample of the revenue generated by 100 customers who made purchases after clicking on the ads.

Question: Using Python, how would you calculate the confidence interval for the average revenue from these 100 customers? Provide Python code that computes the confidence interval at a specified confidence level (e.g., 95%) for the mean revenue.

28. Consider a sample car dataset and plot Mutlivariate graphs to show the distribution of data from multiple variables for Multivariate Scatterplot and Scatter Plot Matrix

29. 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 a CSV file named "customer_reviews.csv," which contains customer ratings for products in the chosen category.

30. You are a data scientist working for a company that sells shoes. You are tasked with writing a program that will calculate the frequency distribution of shoe sizes sold in the past year. The data is stored in a file called shoe_sales.csv. The file contains the following columns:

- shoe_size: The size of the shoe sold.
- quantity: The number of shoes sold in that size.

Write a program that will read the data from the file and calculate the frequency distribution of shoe sizes. The program should output the frequency distribution table, as well as a bar chart showing the frequency of each shoe size.