Individual Assessment Coversheet

To be attached to the front of the assessment.

Campus: East London

Faculty: BSc IT Robotics

Module Code: ITSCA2-12

Group: National

Lecturer's Name: Prosper Sotenga

Student Full Name: Musongela Nyembo

Student Number: EDUV4843866

Indicate	Yes	No
Plagiarism report attached	X	

Declaration:

I declare that this assessment is my own original work except for source material explicitly acknowledged. I also declare that this assessment or any other of my original work related to it has not been previously, or is not being simultaneously, submitted for this or any other course. I am aware of the AI policy and acknowledge that I have not used any AI technology to generate or manipulate data, other than as permitted by the assessment instructions. I also declare that I am aware of the Institution's policy and regulations on honesty in academic work as set out in the Conditions of Enrolment, and of the disciplinary guidelines applicable to breaches of such policy and regulations.

Signature	**Date**
MB.N	June 13 2025

Lecturer's Comments:

Marks Awarded:	**%**

Signature	**Date**

Section A

Question 1

Task Requirements:

Write a Python script that performs the following operations:

1. Data Loading and Initial Display

a. Load the petrolprices.csv file into a Pandas dataframe.(1 Mark)

b. Display the first few rows to confirm successful import.(1 Mark)

```
    1 Load and display the petrol data? or other stuff.
    2 Set the month and year as index, petrol and dizel differnce.
    3 separate data from 2023 and 2024
    4 Visualise data
    5 What insights can be drawn from the price difference trends across the two years?
    0 to exit
    : 1
    Year;Month;Petrol;Diesel
0   2023;January;22.95;23.87
1  2023;February;21.54;22.67
2     2023;March;20.35;21.86
3     2023;April;23.46;24.33
4       2023;May;24.35;25.75
```

2. Data Preprocessing

a. Set the Year and Month columns as the index of the dataframe.(3 Marks)

b. Create a new column called 'Price Difference' that stores the difference between diesel and petrol prices for each month. (5 Marks)

```
    1 Load and display the petrol data? or other stuff.
    2 Set the month and year as index, petrol and dizel differnce.
    3 separate data from 2023 and 2024
    4 Visualise data
    5 What insights can be drawn from the price difference trends across the two years?
    0 to exit
    : 2
              Petrol  Diesel  Price Difference
Year Month
2023 January    22.95   23.87              0.92
     February   21.54   22.67              1.13
     March      20.35   21.86              1.51
     April      23.46   24.33              0.87
     May        24.35   25.75              1.40
```

3. Year-wise Data Segmentation

Create two new dataframes:

a. One containing only data for the year 2023(3 Marks)

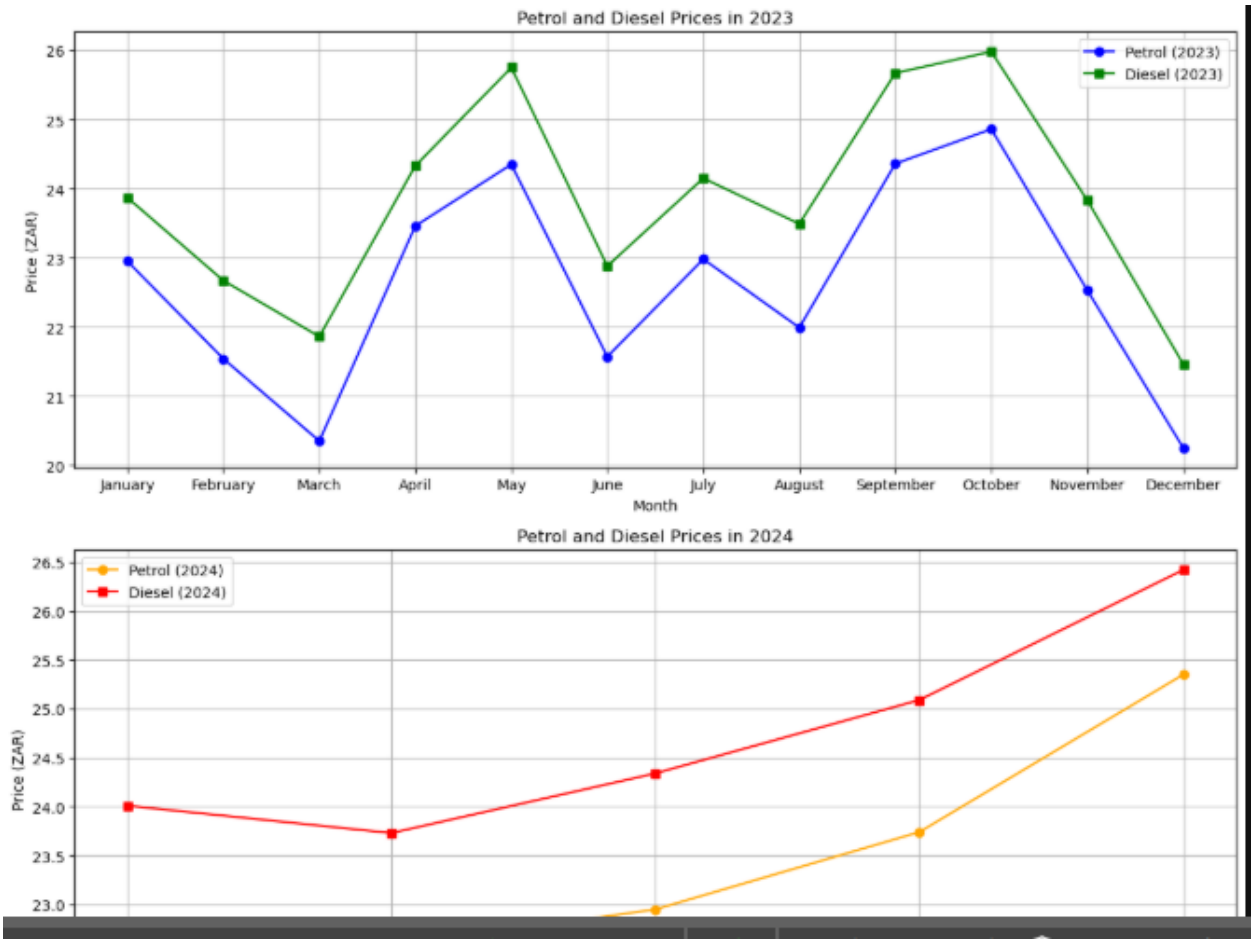
b. One containing only data for the year 2024(3

```
    1 Load and display the petrol data? or other stuff.
    2 Set the month and year as index, petrol and dizel differnce.
    3 separate data from 2023 and 2024
    4 Visualise data
    5 What insights can be drawn from the price difference trends across the two years?
    0 to exit
    : 3
2023 Data:
          Petrol  Diesel
Month
January    22.95   23.87
February   21.54   22.67
March      20.35   21.86
April      23.46   24.33
May        24.35   25.75

2024 Data:
          Petrol  Diesel
Month
January    22.37   24.01
February   22.57   23.73
March      22.95   24.34
April      23.74   25.09
May        25.35   26.42
```

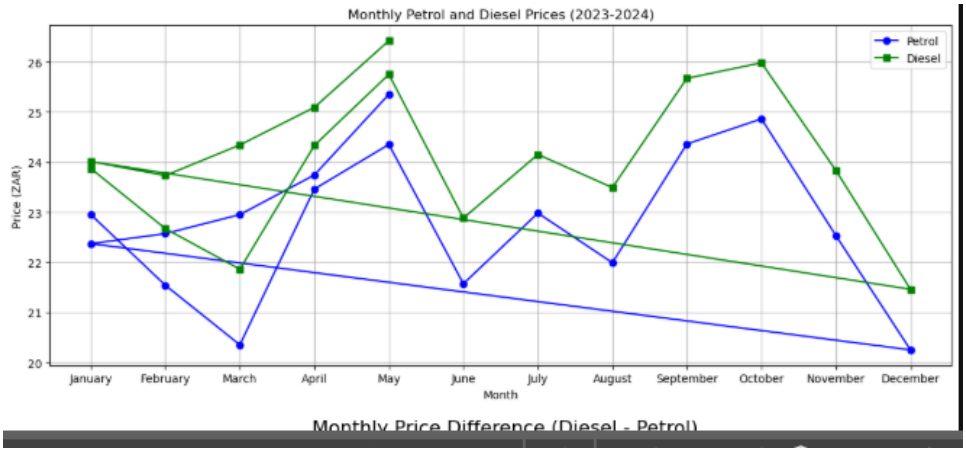
4. Data Visualisation

a. Plot individual line charts showing the petrol and diesel prices for each year (2023 and 2024) separately.(2 Marks)



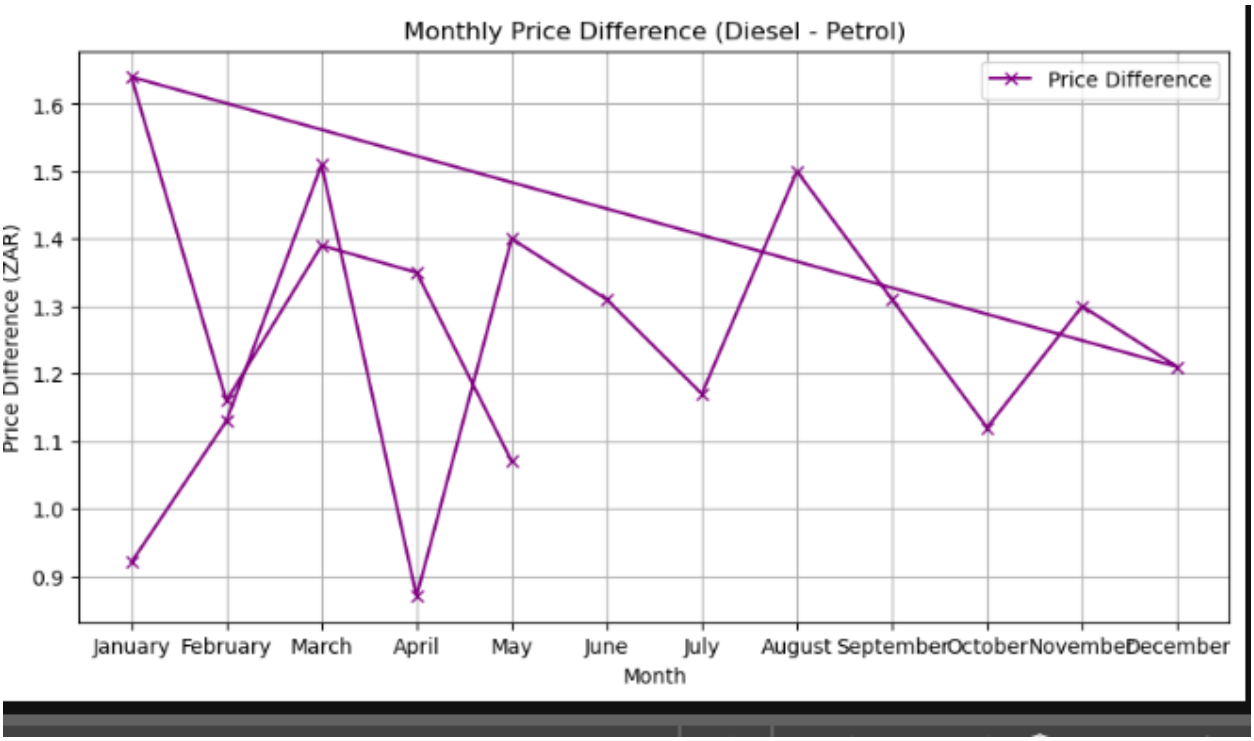
b. Plot a combined line chart that shows the monthly trend in both petrol and diesel prices

from January 2023 to December 2024.(3 Marks)



c. Plot a separate line chart showing the monthly 'Price Difference'.

(2 Marks)



Additional Questions

5. Diesel has been more expensive on average, the closest they have ever was in April 2023 and the furthest was January 2024

Question 2

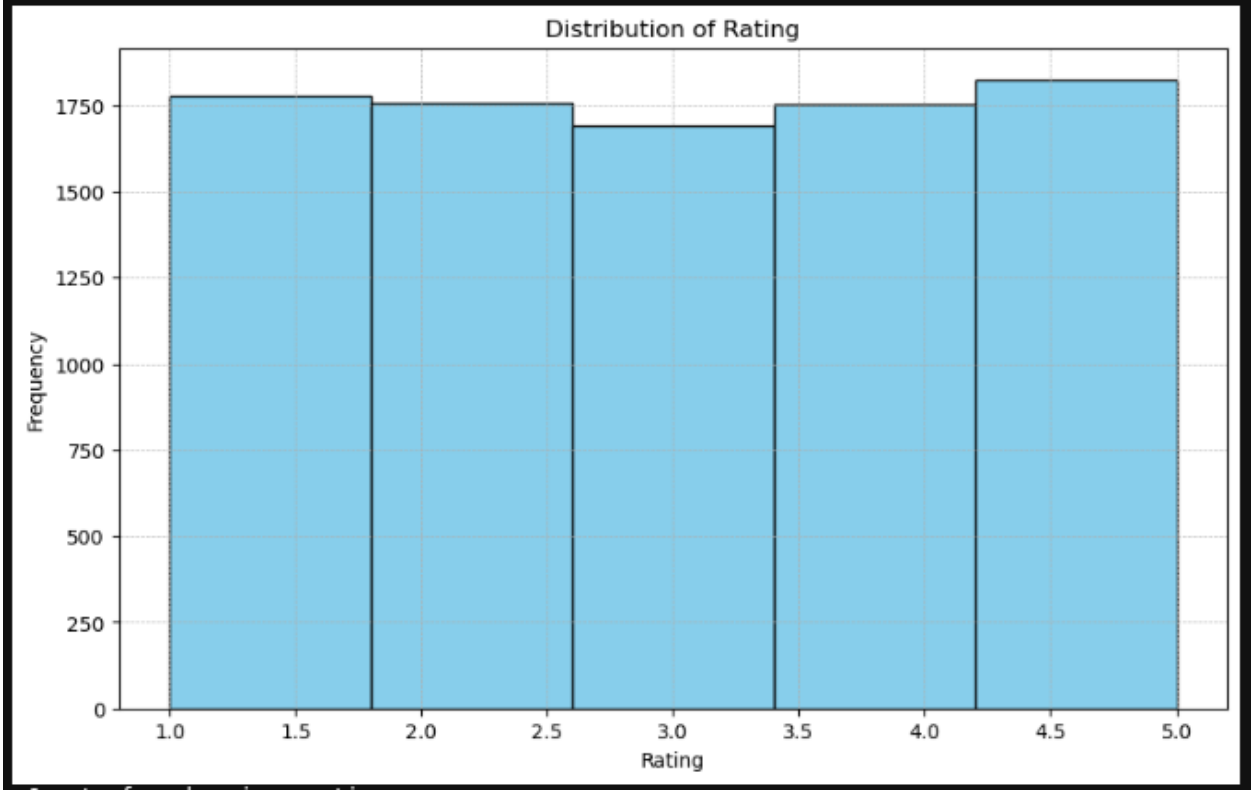
1. Data Loading
    a. Load the dataset series_movies_descriptions.csv into a Pandas dataframe.
    b. Display the first few rows of the dataframe to confirm successful import. (3 Marks)

```
    1 Load and display the series_movies data?.
    2 Visualize the distribution of rating.
    3 Sentiment Analysis
    4 Classification of Sentiment
    5 Visualisation of Sentiment
    0 to exit
    : 1
                 title                                        description  \
0   Dick Johnson Is Dead  As her father nears the end of his life, filmm...
1          Blood & Water  After crossing paths at a party, a Cape Town t...
2              Ganglands  To protect his family from a powerful drug lor...
3  Jailbirds New Orleans  Feuds, flirtations and toilet talk go down amo...
4           Kota Factory  In a city of coaching centers known to train I...

   Rating
0       2
1       2
2       4
3       1
4       5
```

2. Exploratory Data Analysis
    a. Visualise the distribution of the ratings using an appropriate chart (e.g., histogram or bar chart). (2 Marks)
    b. Display the count of each unique rating to understand how ratings are distributed across the dataset. (2 Marks)

```
4 Classification of Sentiment
5 Visualisation of Sentiment
0 to exit
: 2
```



3. Sentiment Analysis
    a. Use a suitable natural language processing library (e.g., TextBlob, VADER) to compute sentiment polarity scores for each movie or series description.
    b. Add a new column to the dataframe called 'Polarity' to store the sentiment score of each description. (5 Marks)

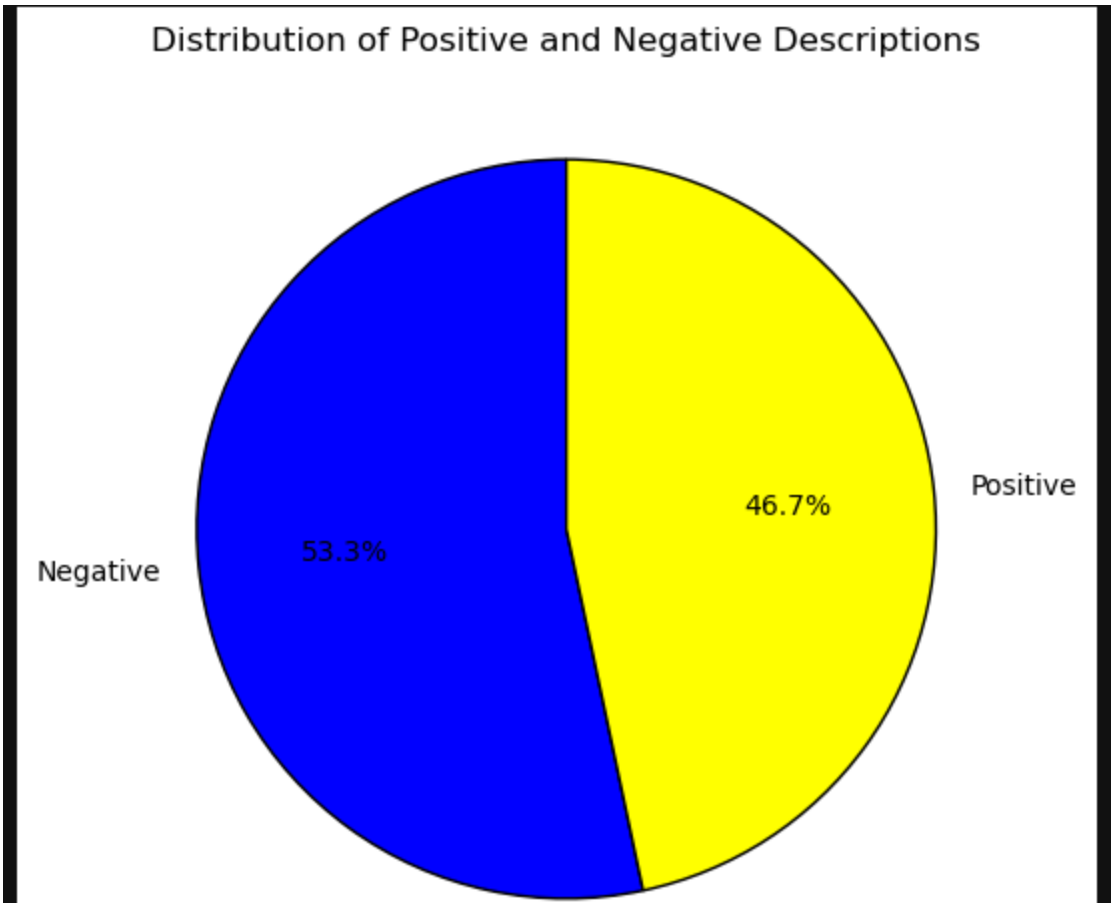
```
    1 Load and display the series_movies data?.
    2 Visualize the distribution of rating.
    3 Sentiment Analysis
    4 Classification of Sentiment
    5 Visualisation of Sentiment
    0 to exit
    : 3
                title                                        description  \
0   Dick Johnson Is Dead  As her father nears the end of his life, filmm...
1          Blood & Water  After crossing paths at a party, a Cape Town t...
2              Ganglands  To protect his family from a powerful drug lor...
3  Jailbirds New Orleans  Feuds, flirtations and toilet talk go down amo...
4           Kota Factory  In a city of coaching centers known to train I...

   Rating  Polarity
0       2   -0.2960
1       2   -0.1531
2       4   -0.7783
3       1    0.2263
4       5    0.7430
```

4. Classification of Sentiment
    a. Based on the polarity value, classify each description as either 'Positive' (if polarity > 0) or 'Negative' (if polarity ≤ 0).
    b. Add this classification as a new column in the dataframe called 'Sentiment Label'. (6 Marks)

```
    1 Load and display the series_movies data?.
    2 Visualize the distribution of rating.
    3 Sentiment Analysis
    4 Classification of Sentiment
    5 Visualisation of Sentiment
    0 to exit
    : 4
                                        description  Polarity Sentiment Label
0  As her father nears the end of his life, filmm...   -0.2960        Negative
1  After crossing paths at a party, a Cape Town t...   -0.1531        Negative
2  To protect his family from a powerful drug lor...   -0.7783        Negative
3  Feuds, flirtations and toilet talk go down amo...    0.2263        Positive
4  In a city of coaching centers known to train I...    0.7430        Positive
```

5. Visualisation a. Plot a chart (e.g., bar chart or pie chart) that shows the distribution of 'Positive' and 'Negative' descriptions. (3 Marks)



Additional Questions

6. Suggest one method to improve the accuracy of the sentiment classification, especially for longer or more nuanced descriptions. (2 Marks)

To improve sentiment classification accuracy in longer or complex descriptions, use pre-trained deep learning models like BERT or RoBERTa.
These models are better at understanding language than older methods like VADER. They have contextual understanding, they can handle complex language and can be adjusted to specific datasets for accuracy

7. What limitations might exist in using rule-based tools like TextBlob for movie or series sentiment analysis?

They cannot pickup context, they would take things like sarcasm more seriously.

Question 3

Instructions

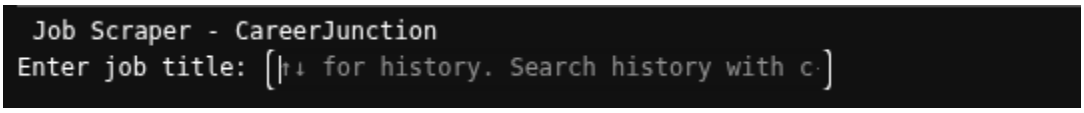
Develop a Python application that automates the retrieval of job postings from the CareerJunction

website (https://www.careerjunction.co.za).

Task Requirements:

1. User Interaction

Prompt the user to input a job title they wish to search for.



2. Web Scraping

Use an appropriate Python library (e.g., requests, BeautifulSoup, Selenium) to perform the web

scraping task.

Extract the following data fields from the first page of search results:

Job Title

Recruiter Name

Salary

Job Position or Type

Job Location

Date Posted (8 Marks)

```
#web data extraction
    # Extract title
    title_elem = (job.select_one('h2 a') or
                  job.select_one('h3 a') or
                  job.select_one('[data-testid="job-title"]') or
                  job.select_one('.job-title') or
                  job.select_one('a[title]'))

    # Extract company
    company_elem = (job.select_one('[data-testid="company-name"]') or
                    job.select_one('.company-name') or
                    job.select_one('.recruiter') or
                    job.select_one('.employer'))

    # Extract location
    location_elem = (job.select_one('[data-testid="location"]') or
                     job.select_one('.location') or
                     job.select_one('.place') or
                     job.select_one('.city'))

    # Extract salary from text
    salary_match = re.search(r'R\s*\d+[\d\s,]*(?:\s*-\s*R?\s*\d+[\d\s,]*)?', text)
    salary = salary_match.group().strip() if salary_match else "Not specified"

    # Extract date
    date_match = re.search(r'\d+\s+(?:day|week|month)s?\s+ago|today|yesterday', tex
    date = date_match.group().strip() if date_match else "Not specified"
```

The data extracted must be stored inside a dataframe as follows:

Title

Recruiter

Salary

Position

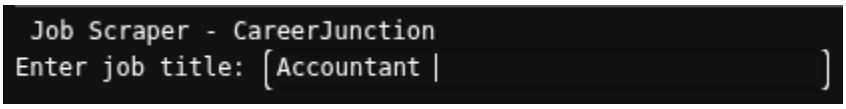
Location

Date Posted (6 Marks)

```
    job_data = {
        'Title': title_text,
        'Recruiter': company_elem.get_text().strip() if company_elem else "Not specified",
        'Salary': salary,
        'Position': title_text,
        'Location': location_elem.get_text().strip() if location_elem else "Not specified",
        'Date Posted': date
    }
```

The dataframe must be saved as a csv. The csv filename must be in the following format:

search_term + ‘job-results.csv’.(2 Marks)



The script works as a python file

Additional Questions

3. What challenges or limitations might arise when scraping dynamic websites such as

CareerJunction?(2 Marks)

Anti bot measures, early I encountered errors that did not allow my to request due to the site suspecting my script as a bit

Sites like CareerJunction can change and do not stay stationary over time some element may change leading to the scrip being useless

4. How would you handle pagination to extract jobs beyond the first page?

I could place a variable the increases on the URL, through a loop.

Section B

Question 4

Task Requirements:

1. Data Preparation

a. Load the dataset into a Pandas dataframe.(1 Mark)

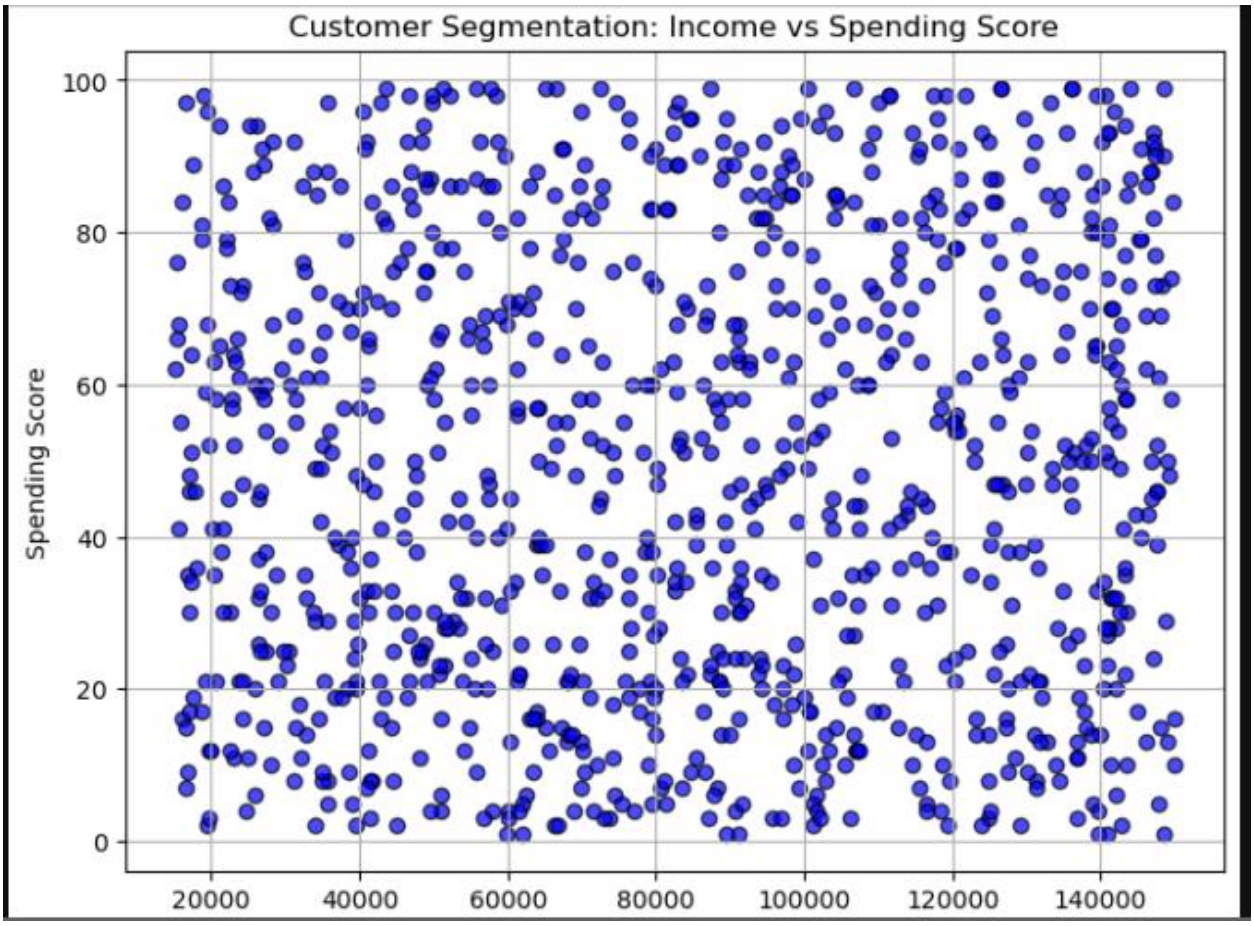
b. Identify and remove any rows containing missing (null) values.(1 Mark)

c. Drop all columns except the 'Income' and 'Spending Score' columns.(1 Mark)

```
def instructions ():
# a. Loads the dataset
    df = pd.read_csv("customer_segmentation.csv", sep=";")
# b. removes rows containing missing values
    df_cleaned = df.dropna()
# c. Drop all columns except 'Income' and 'Spending Score'
    df_selected = df_cleaned[['income', 'spending_score']]
# d. Display a scatter plot to examine patterns
    plt.figure(figsize=(8, 6))
    plt.scatter(df_selected['income'], df_selected['spending_score'], alpha=0.7, c='blue',
    plt.xlabel("Income")
    plt.ylabel("Spending Score")
    plt.title("Customer Segmentation: Income vs Spending Score")
    plt.grid(True)
    plt.show()
    askUser()
```

d. Display a visual scatter plot of the remaining features to examine initial patterns.(2 Marks)

Input = 1



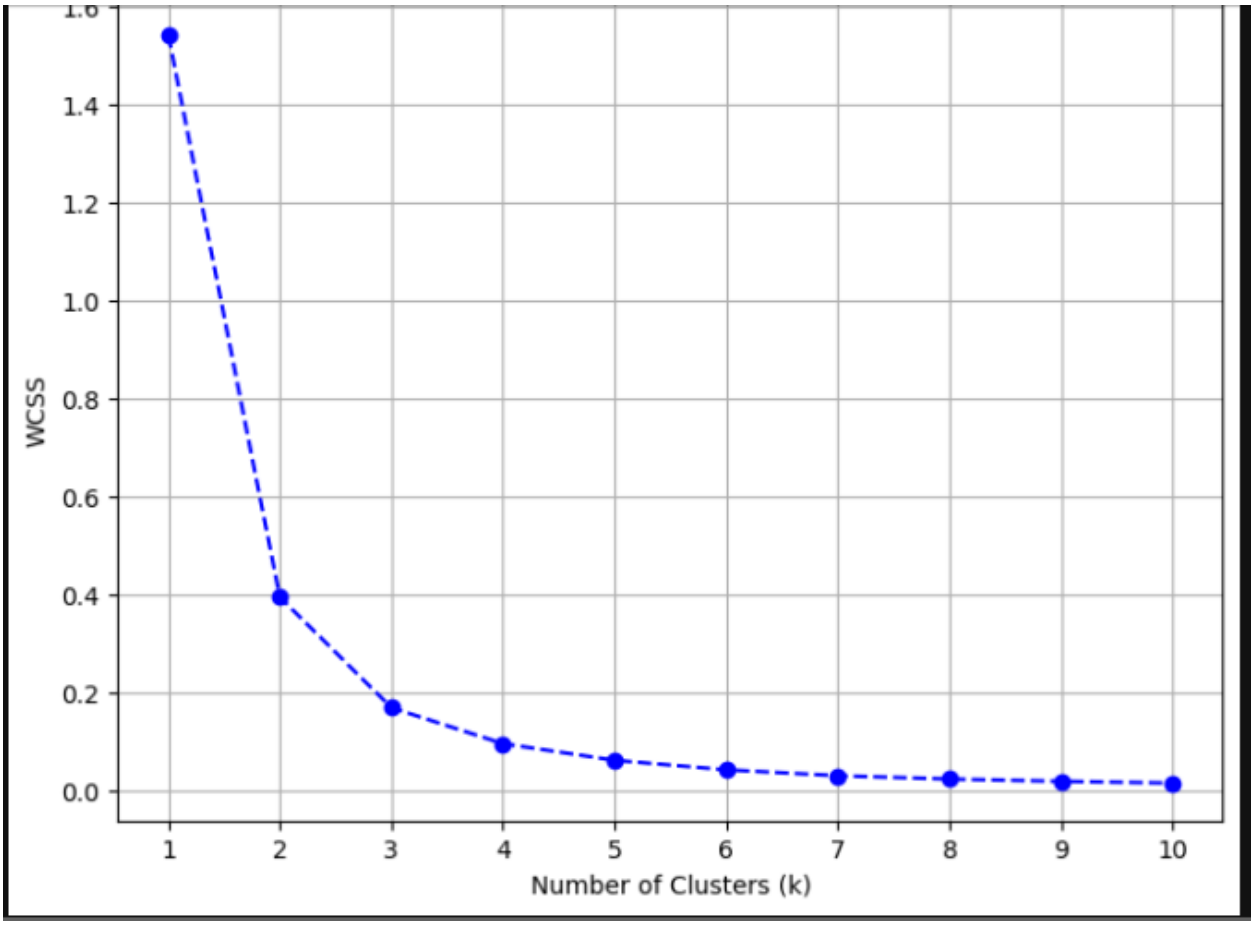
2. Clustering Analysis

a. Apply the elbow method to determine an appropriate number of clusters.

Use KMeans clustering from sklearn.cluster for this step.

Plot the Within-Cluster-Sum-of-Squares (WCSS) for different values of k (e.g., 1 to 10) to

determine the optimal number of clusters.(7 Marks)

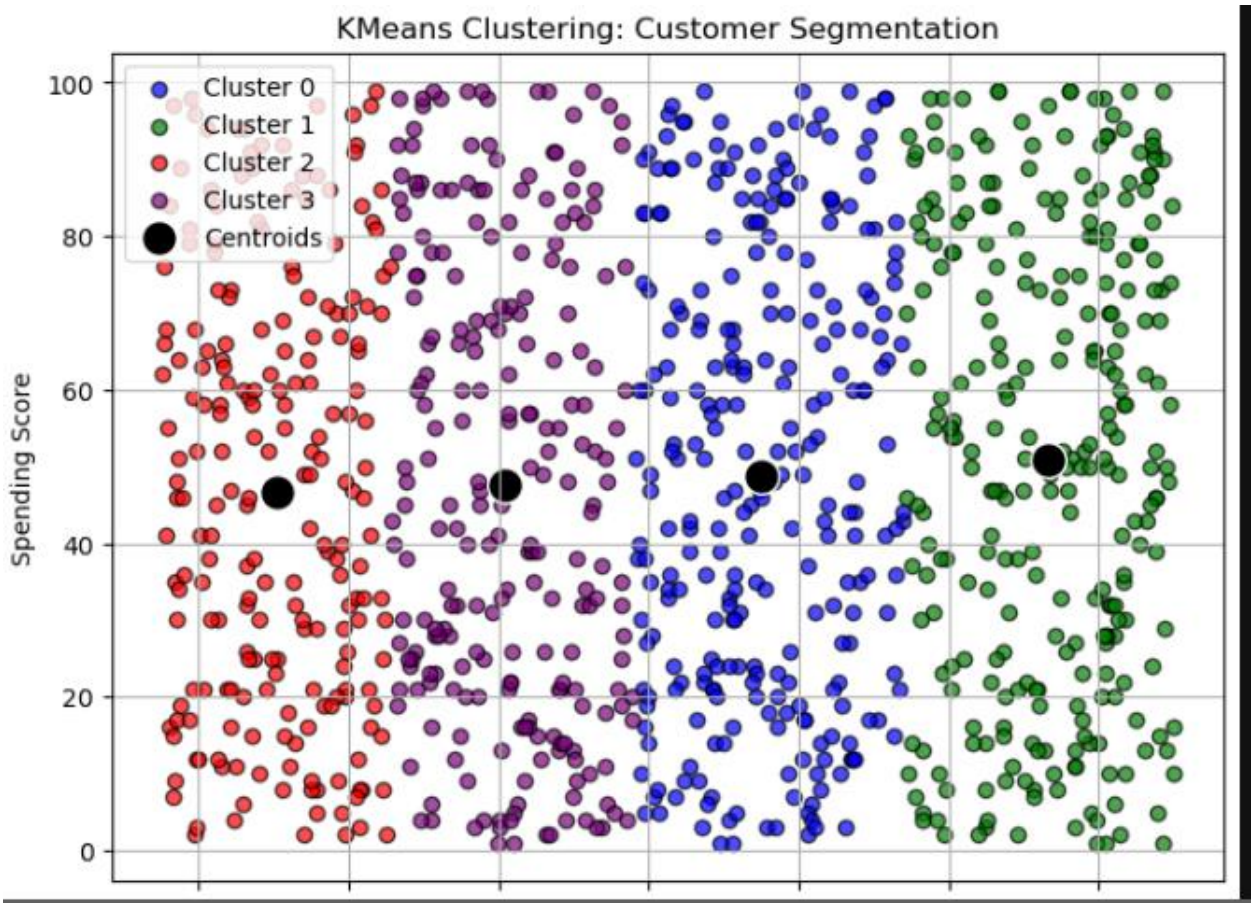


3. Cluster Visualisation

a. Apply KMeans clustering using the optimal number of clusters determined in the previous step.

b. Plot the resulting clusters using different colours for each cluster.(10 Marks)

c. Mark the centroids of the clusters on the plot using a distinct marker or colour.(4 Marks)



4. Interpretation and Summary

a.

Write a short summary (3–5 lines) explaining the number of clusters formed and the apparent characteristics of each cluster.(2 Marks)

We discovered four customer groups based on income and spending habits.
Some earn and spend a lot, while others spend little despite high income.
There's a group that spends a lot even with low income, possibly impulsive buyers.
And finally, one group earns and spends very little — likely more cautious shoppers.

b.

Discuss how this clustering output might assist the retailer in tailoring advertisements for different customer groups.

The clusters help the retailer target ads more effectively. Big spenders can get premium offers, while cautious buyers might prefer savings or deals. Discount ads work well for low earners who spend more, and budget shoppers can get basic, value-focused promotions.

References

Anon. 2024. Python Pandas Tutorial for Data Analysis. Real Python, 10 December 2024. [Online] Available at: https://realpython.com/pandas-python-tutorial/ [Accessed: 2025-06-02].

Anon. 2025. Sentiment Analysis in Python Using TextBlob. Machine Learning Mastery, 20 January 2025. [Online] Available at: https://machinelearningmastery.com/sentiment-analysis-in-python-with-textblob/ [Accessed: 2025-06-04].

Anon. 2025. Web Scraping with Python: BeautifulSoup and Requests. Towards Data Science, 15 February 2025. [Online] Available at: https://towardsdatascience.com/web-scraping-with-python-beautifulsoup-requests-7d7a023bd4f3 [Accessed: 2025-06-06].

Anon. 2025. K-Means Clustering with Python and Scikit-learn. DataCamp, 5 March 2025. [Online] Available at: https://www.datacamp.com/tutorial/k-means-clustering-python [Accessed: 2025-06-07].

Anon. 2025. Data Visualization in Python Using Matplotlib. GeeksforGeeks, 1 April 2025. [Online] Available at: https://www.geeksforgeeks.org/data-visualization-in-python-using-matplotlib/ [Accessed: 2025-06-09].

Anon. 2025. Handling Missing Data in Pandas. Analytics Vidhya, 12 March 2025. [Online] Available at: https://www.analyticsvidhya.com/blog/2025/03/handling-missing-data-in-pandas/ [Accessed: 2025-06-08].

Anon. 2024. How to Plot Line Graphs in Python Using Matplotlib. PythonProgramming.net, 22 November 2024. [Online] Available at: https://pythonprogramming.net/matplotlib-line-graphs/ [Accessed: 2025-06-03].

Anon. 2025. Using the Elbow Method to Find the Optimal Number of Clusters in K-Means. KDnuggets, 18 February 2025. [Online] Available at: https://www.kdnuggets.com/2025/02/elbow-method-k-means-clustering.html [Accessed: 2025-06-10].