Exploratory Data Analysis

Graded Project Week 2

Part - I (25 points)

- Numpy
- Pandas
- Visualization

Part - II (25 points)

- Encoding categorical data
- Scaling and Normalization
- Imputing Missing Values
- Handling outlier values
- pandas-profiling library

Part-I (25 Points)

Dataset: Online_sales.csv
Domain: Retail
Objective: Analyzing sales data to understand sales trend

1. Import necessary libraries and read the provided dataset (online_sales.csv). (1 point)
2. Check top 5 and random 5 samples of the dataframe. (1 point)

3. Check info of the dataframe and write your observations. Comment on data types and shape of the dataset. (1 points)
4. Check for null values and report the percentage of null values of each column. And drop the rows having null values in it. (1 points)
5. Check statistical summary of the dataset. (1 point)
6. Drop the instances having quantity less than zero. (1 point)
7. Check unique values of the country and report the name of the country that has the highest number of instances. (2 points)
8. Create a new column with the name as ‘sales’ having total sales. The total sales is defined as Quantity*UnitPrice. (3 points)
9. Report the top 5 countries in terms of sales. (2 points)
    a. Consider the size of sales.
    b. Consider the mean value of sales.
10. Report the top 5 products which bring the highest sales. Use StockCode for product information. (2 points)
11. Convert the ‘InvoiceDate’ into a date format and report the month on which the maximum sales occur? (5 points)
12. Check statistical summary of the sales and use an appropriate plot to display the distribution of sales and write your inferences. (2 points)
13. Submit a business report including your findings and interpretations of the above project. Please refer to the do’s and don’t document for more information. (3 points)

Part- II (25 Points)

Marketing Data Analysis

Domain:

Marketing

Objective:

To extract actionable insights that will enable growth in the market

Data Description: maketing_data.csv

The dataset can be found here

Feature Details:

ID: Customer's unique identifier

Year_Birth: Customer's birth year

Education: Customer's education level

Marital_Status: Customer's marital status

Income: Customer's yearly household income

Kidhome: Number of children in customer's household

Teenhome: Number of teenagers in customer's household

Dt_Customer: Date of customer's enrollment with the company

Recency: Number of days since customer's last purchase

MntWines: Amount spent on wine in the last 2 years\

MntFruits: Amount spent on fruits in the last 2 years

MntMeatProducts: Amount spent on meat in the last 2 years

MntFishProducts: Amount spent on fish in the last 2 years

and so on..

The complete feature details can be found in the above mentioned link.

Tasks to be performed:

1. Import necessary libraries. (1 point)

2. Load the file and display the first 5 and last 5 instances. (1 point)

3. Check the shape of the data (number of rows and column). (1 point)

4. Generate pandas profiling report of the original data. (2 points)

5. Check the dtype of values in column 'Income'. (1 point)

6. Convert the values in the 'Income' column to numeric format. (2 points)

7. Check the distribution of the income column. (1 point)

8. Check the presence of outliers in the feature 'Income'. (1 point)

9. Encode categorical features to numerical. (3 points)

- Convert the column **'Education'** from categorical to numerical format.
  Map them as Basic=1, Graduation=2, Master=3, PhD=4, 2n Cycle=5
- Check the number of unique values in the column "Country"

Since the column **Country** and **Marital Status** is **Nominal**

- So we will one-hot encode these variables.

10. Convert the values in column **'Dt_Customer'** to datetime. (2 points)

- After converting the values to datetime, convert it to numerical values.

11. Check the number of null values present in each column. (1 point)

12. Handle null values using the below given approaches. (3 points)

- **1st Approach:** Since the number of instances having null values is too less, we can drop the null instances. And drop the null instances and save it in a new DataFrame df2
- **2nd Approach:** Fill the null instances with median value and save it in new dataframe df3
  We are not using mean as the column contains some extreme values
- **3rd Approach:** Use sklearn's KNNImputer to impute the data, and save it in dataframe df4

13. Visualize the outliers using a scatter plot. (2 points)

14. Handle the outlier values in the column **Income. (3 points)**

- **1st Approach:** Drop the instances where income is greater than 1,50,000, save it in df2
- **2nd Approach:** Drop the instances which have outlier values using the IQR, save it in df3
- **3rd Approach:** Cap the instances to max or min value using the IQR, save it in df4

15. Scale the data in column 'Income' to have mean=0 and standard deviation = 1. (1 points)