Final Project

Big Data Analysis and Visualization in Python

Submission: (5 pts)

You need to submit the python code and the screenshots of the execution results for every question. Please submit a word file and a python file.

The use of Jupyter notebook is a MUST.

Problem Context: (45 pts)

Customer behavior refers to an individual's buying habits, including social trends, frequency patterns, and background factors influencing their decision to buy something.

Businesses study customer behavior to understand their target audience and create more-enticing products and service offers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

Customers are first segmented into buyer personas based on their common characteristics. Then, each group is observed at the stages on your customer journey map to analyze how the personas interact with your company. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

Goal: Shape the data to create different customer segments based on their respective persona.

Data : You are provided with customers.csv file. The data has many columns organized based on their scope as follows:

People

- 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
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Part I: Data Overview

1. Use the pandas library to read the data file and to create the data frame.
2. Display the first **5** rows and the last **3** rows of your data.
3. Show quick statistics of your data
4. Show the data type of each column
5. Show how many columns and rows in your data.
6. Show the list of columns in your data frame.
7. Show the number of duplicated rows in your data.

Part II: Data Preparation and Cleaning

8. Rename the columns belonging to **Products** by removing Mnt from them.
    a. You should have: **'Wines','Fruits','Meat','Fish','Sweet','Gold'**
9. Rename the columnsbelonging to **Place** by removing Num and Purchases from them.
    a. You should have: **'Web ','Catalog', 'Store'**
10. Keep only **NumDealsPurchases** column in the **Promotion** section
11. Delete the column **NumWebVisitsMonth**
12. The attribute **Year_Birth** is not helpful to segment your data. Instead create a new derived column named **Age**.
13. Create a new column named **Spending** that holds the total amount spent on all products categories.
14. Change the **Dt_Customer** type from object to **datetime**. Use the format YYYY-MM-DD.
15. Show the unique values of the column `data` **Marital_Status**.
16. It is better to work with lower categorical values. Hence, we'll classify customers in two segments for their martial status: change the column **Marital status** as follows:
    a. Change the values: Divorced, Single, Absurd and Widow to **'Alone'**
    b. Change Married and Together to **'Not Alone'**

17. Show the unique values of the column **Education**.

18. We want to segment the customers in 2 groups based on their education level: change the column **Education** as follows:
    a. Change the values: 'Basic' and '2n Cycle' to **'Undergraduate'**,
    b. All other values to **'Postgraduate'**
19. Have a look at column **Income**. Investigate the presence of outliers? Delete them accordingly.
20. Show quick statistics for the column **Income**.
21. Show the number of missing values for each column.
22. Fill the missing values with the average.
23. Create a new column named **Children** to hold the total number of children for every customer.
24. Create a new column **Has_Complaint** as follows:
    a. **'Has complaint'** if the customer complained in the last 2 years,
    b. Otherwise **'No complaint'**
25. Show the first 5 rows of customers with **'No complaint'**

Part III: Data Visualization & Analytics

26. Let's analyse the profile of customers based on their background factors. To do so, we will use our cleaned data frame (in part II) to create a new one with the columns: **Age, Spending, Marital_Status, Education, Income, Children, Dt_Customer and Has_Complaint.**
27. Create histogram to show the number of children per customers in your data. What do you see?
28. Create a histogram and a Boxplot to show the Income of customers. What do you find?
29. Create a histogram and a Boxplot to show the Spending of customers. What do you find?
30. Create a histogram and a boxplot to analyse the customers based on their Age.
31. Create a histogram to analyse the Education level of the company's customers. What do you find?

**The column Education is defined as Object and has 2 values (Undergraduate, Postgraduate). You need to shape the values as numeric to be able to plot the histogram.