MTH 765P Mini project

Venkata Sri Naga Sailusha Peddibhotla

1 Introduction

Data Analysis is process of collecting data, analysing it and perform modelling to predict outputs with the information collected. These are not fully understood by the businesses, who only need how the data can be useful to improve their strategy. This requires creativity while doing data analysis to enable the reader to understand the analysis to help provide insight. Businesses collect information from their consumers to better understand their behaviour while buying to discover patterns, trends and relationships that are not visible directly. EDA is an important task in helping identify and understand the outliers, relationship of features within the data using graphs and plots, https://www.ibm.com/uk-en/cloud/learn/exploratory-data-analysis.EDA.

The analysis is performed on a dataset to derive some conclusions from the data by making visualisations between different features using libraries such as Matplotlib and seaborn. This is accomplished using following steps: Data collection, cleaning, visualisation and sharing the results with client.

In this report a sample data set of customer profiles was taken, and EDA was performed to understand the trends observed in the data. The next section discusses the process of obtaining the data, followed by description of the dataset. This data was initially cleaned and then uni-variate analysis was completed using graphical and non-graphical methods. This was followed by bi-variate analysis. Finally, conclusions on the insights from the data were presented.

2 Obtaining /Acquiring the data

2.1 About the dataset

The dataset has been taken from Kaggle (https://www.kaggle.com/jackdaoud/marketing-data). This is a public repository containing many data sets published for machine learning projects by analysts and data scientists. The dataset chosen had, customers profiles enrolled in a company, their spending behaviour on different products, impact of advertising campaigns on customer spending and the channel where the products were sold (online or store). The

downloaded dataset was in a .csv format. It had 2,240 customers with 28 columns with the customer information. The details of the information provided is discussed in the next section.

3 Description

The dataset was about customer data collected by a company on how they purchase products, which channels they prefer for purchases, relationships between the features like income and children on spending. Table 1 shows a list of the information present in the dataset. The table defines the information present in each column of the dataset.

Table 1: Customer details enrolled in a company

Column_name	Data
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
MntWines	Amount spent on wine in the last 2 years
MntMeatProducts:	Amount spent on meat in the last 2 years
MntFruits:	Amount spent on fruits in the last 2 years
MntFishProducts:	Amount spent on fish in the last 2 years
MntSweetProducts:	Amount spent on sweets in the last 2 years
MntGoldProds:	Amount spent on gold in the last 2 years
NumDealsPurchases	Number of purchases made with a discount
NumCatalogPurchases:	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth:	Number of visits to company's web site in the last month
AcceptedCmp1:	1 if customer accepted the offer in the 1st campaign, 0 otherwise (Target variable)
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise (Target variable)
AcceptedCmp3:	1 if customer accepted the offer in the 3rd campaign, 0 otherwise (Target variable)
AcceptedCmp4:	1 if customer accepted the offer in the 4th campaign, 0 otherwise (Target variable)
AcceptedCmp5:	1 if customer accepted the offer in the 5th campaign, 0 otherwise (Target variable)
Response	1 if customer accepted the offer in the last campaign, 0 otherwise (Target variable)
Complain:	1 if customer complained in the last 2 years, 0 otherwise
Recency	Number of days since customer's last purchase
Country	Customer's location

4 Analysis

4.1 Defining the problem

Different questions can be explored to improve the business using the dataset. The expectations of the business is normally used to define the problem. The questions which were considered

as key parameters to improve sales and conclusions were derived after performing the analysis. The goal was to answer questions such as: the most purchased product, best performing channels for sales, the successful advertising campaigns and factors effecting products purchased.

4.2 Data cleaning and exploration

4.2.1 Importing Libraries:

Initially all the necessary libraries required for analysis were imported, this consists of methods required for writing the code namely Pandas and Matplotlib libraries. **Pandas** library is a data analysis tool used for data manipulation. **Matplotlib** is a comprehensive library for creating static, animated, and interactive visualizations in Python. Seaborn was used as the Python data visualization library based on matplotlib. This provided a high-level interface for drawing attractive and informative statistical graphics.

4.2.2 Importing dataset and identification of variables and data types:

Variables were of two types Numerical and Categorical. The numerical data can be further classified as discrete, continuous, nominal and ordinal. The dataset was imported, and the first few data points were used to identify the variable types.

The data was imported using the 'read_csv' command and assigned to the variable 'marketing'. Figure 1 shows a few rows of dataset. There were 3 categorical variables (education, marital_status and country) while the remaining 25 variables were of type numerical. The '.dtypes' method was used to identify the data type of the variables in the dataset. After executing the command 'marketing.dtypes'. It was observed there were 5 variables of object type and other variables were of integer type. This can be useful while making transformations and visualisations.

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumStorePurchases	NumWebVisitsl
0	1826	1970	Graduation	Divorced	$84,835.00	0	0	6/16/14	0	189	...	6	
1	1	1961	Graduation	Single	$57,091.00	0	0	6/15/14	0	464	...	7	
2	10476	1958	Graduation	Married	$67,267.00	0	1	5/13/14	0	134	...	5	
3	1386	1967	Graduation	Together	$32,474.00	1	1	5/11/14	0	10	...	2	
4	5371	1989	Graduation	Single	$21,474.00	1	0	4/8/14	0	6	...	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	
2235	10142	1976	PhD	Divorced	$66,476.00	0	1	3/7/13	99	372	...	11	
2236	5263	1977	2n Cycle	Married	$31,056.00	1	0	1/22/13	99	5	...	3	
2237	22	1976	Graduation	Divorced	$46,310.00	1	0	12/3/12	99	185	...	5	
2238	528	1978	Graduation	Married	$65,819.00	0	0	11/29/12	99	267	...	10	
2239	4070	1969	PhD	Married	$94,871.00	0	2	9/1/12	99	169	...	4	

2240 rows × 28 columns

Figure 1: Few rows of dataset

4.2.3 Understanding data from the identification of variables

The columns can easily be understood from their names. Columns beginning with 'Mnt' denotes the amount spent by the consumer on the particular type of product. Columns beginning with 'Num' denote a number. These values would not have any unit ($, cm, kg, etc.) associated with them. Columns beginning with 'Accepted' was used to define if a marketing campaign has been successful or not. Five different campaigns were held and a value of 1 denoted success while a value of 0 denoted failure.

4.3 Visualisation

4.3.1 Non-Graphical Univariate Analysis:

Transformation of variables:

Data cleaning included changing the variables by renaming existing variable names to a more meaningful name. The income column was also modified to remove any '$' signs and converted to a float type variable. The spending on all products were combined to visualise total spending of each customer.

Filtering based on Conditions

Datasets were filtered using different conditions using logical operators. The marital status column had meaningless variables (Alone, Absurd and YOLO), these were replaced with 'single' category. Table 2 shows the count of the marital status categories. The campaigns were also all combined into 'TotalCampaignsAcc' to find the relationships between these variables and factors that influence success of campaign.

Table 2: Count of values in marital status category

Marital status	Value
Married	861
Together	575
Single	486
Divorced	230
Widow	77

Finding null values

When dataset was imported many blank columns were imported as null values into the Data Frame, which can later create problems while operating that data frame. Pandas 'isnull()' method was used to check and manage null values in a data frame. There were 24 missing

records in the column '*Income*'. These missing records were set to the median value of income to enable the data to be used for the analysis.

4.3.2 Graphical Univariate Analysis:

Box Plots:

A box plot was used to identify the income outlier in the data set. Figure 2 shows the box plot and the outliers. The minimum income was at $1730, the mean was at $52247, 25% percentile was at $35303, the median was at $51381 and the 75% percentile was at $68522. The maximum income was at $666666, as shown in the plot this was considered an outlier in the dataset.

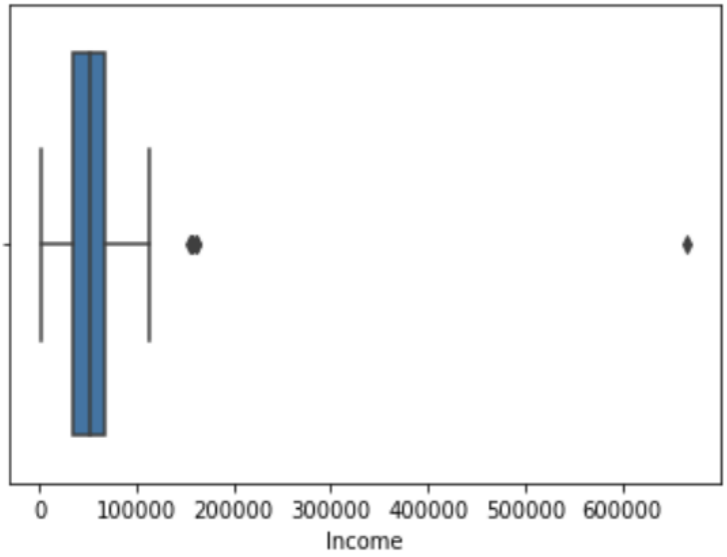

Figure 2: Boxplot of Income with outliers at the upper range

The birth years were plotted to compare customer ages. It was observed that there were few outliers near 1900 and below, shown in Figure 3. These data points were not considered and were removed.

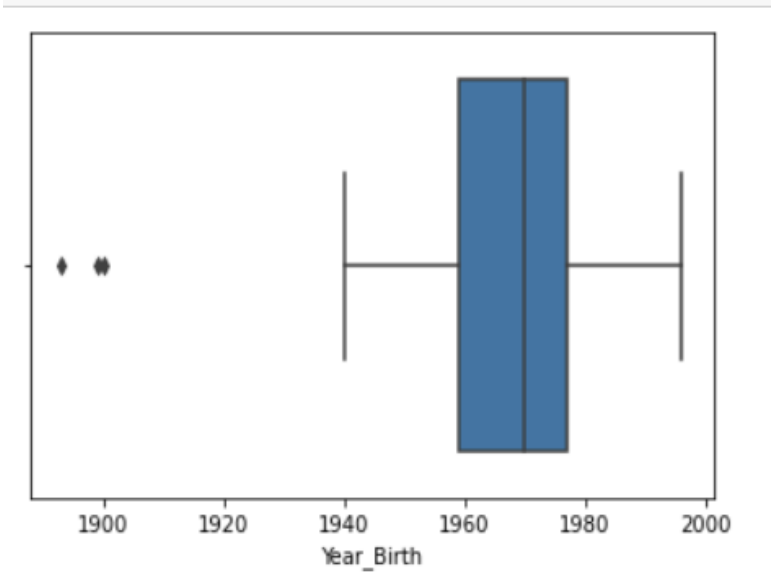

Figure 3: Boxplot of Customer's Date of Birth with outlier near 1900

Plotting Histograms:

The outliers in income was set at $150,000 from boxplot. These datapoints were not contributing to the objective and hence were ignored in this analysis. A histogram plot of the income in shown in Figure 4. The income follows a near normal distribution curve with a flattened peak from $40000 to $70000 approximately. The plot shows a steeper fall above $80000 compared to below $40000 suggesting the majority of the customers have lower incomes.

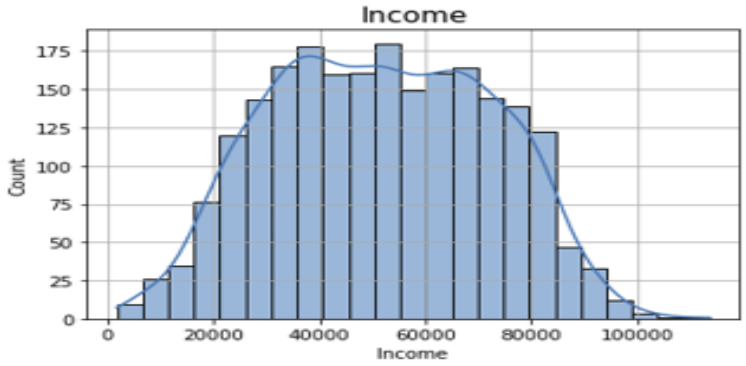

Figure 4: Histogram of Income of the customer from the data set. Outliers above 100000 not shown in the plot

4.3.3 Bi-variate-analysis:

In this case, analysis on how education was impacted the products they purchased were compared. As shown in Figure 5, Wine was most purchased by PhD's as they tended to be older and richer compared to other categories of people and least bought by basic educated people as they may not have money to afford it. Graduates were interested more in purchasing fruits compared to other sectors of people. Meat Products were mostly bought by everyone and it was higher than other products excluding basic educated people. People with 2nd cycle education were interested in fish products which was similar in case of sweets and gold products.

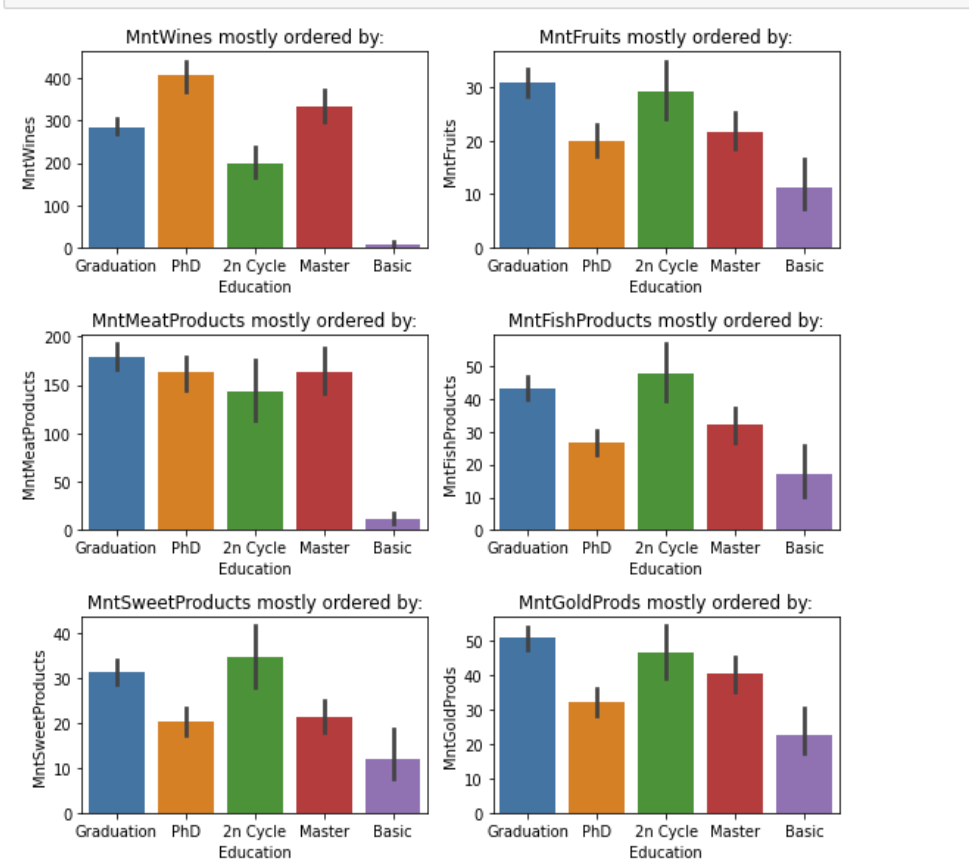

Figure 5:Barplots of products bought by people with different education levels

A second analysis was conducted to understand the patterns in products purchased by parents. In Figure 6, It was observed that most of the products were bought by parents with no kids as they had less expenses to manage and more disposable income to buy items for themselves. Generally as the number of kids increased the amount of items purchased reduced across all product types. The least purchaseds product was fruits, sweets and gold while the most

purchased products were wines and meats with puchases for wines and meat approximately ten times of fruits.

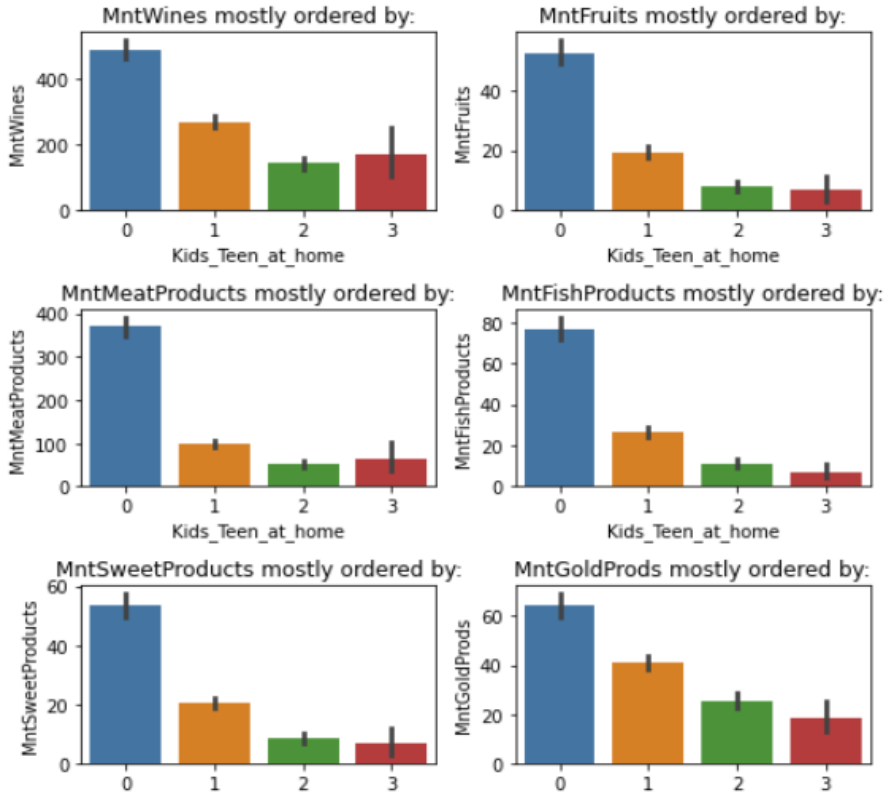

Figure 6:Barplots of parents buying pattern with Kids

Figure 7 shows the spending of an average customer on different product groups. The highest amount was spent wine ($305.00) followed by meat products ($165.30), while the least amount was spent on fruits ($26.3).

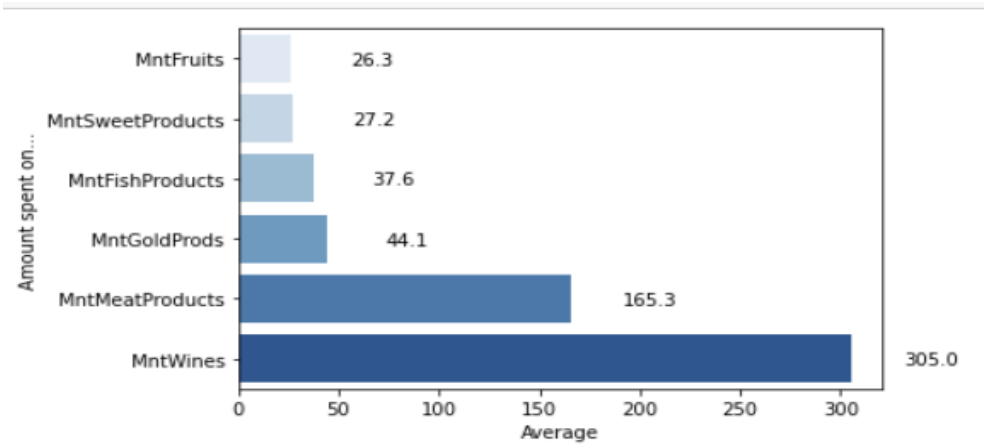

Figure 7:Average Customer money spent on different products

Figure 8 shows the percentage of success rate of advertising campaigns. It was observed that response campaign had the highest success rate (nearly double compared to other campaigns). The least successful campaign was campaign 2 with only 1% success rate.

Figure 8:Advertising campaign success rate

Figure 9 shows the channel of business interaction and purchases completed for an average customer. Most purchases were completed in store (5.8) and followed by website purchases (4.1). The data showed that 77% of web visits led to web purchases. The advertising campaigns amounted to the least amount of purchases.

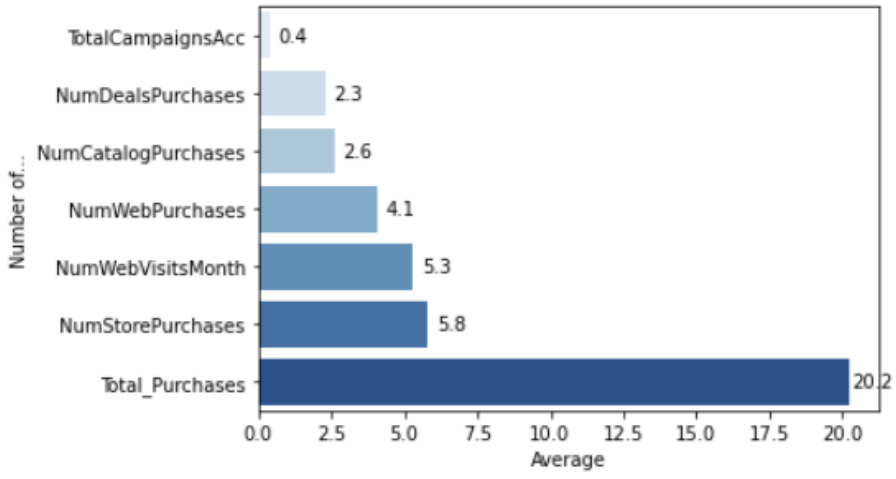

Figure 9:Type of purchases made by customers

5 Conclusion:

The EDA was conducted on a dataset with over 2000 customers to gain insight for better targeting customers. The mean customer income was $51381 with 64% of customer being couples (married or together). The most successful and purchased products were wines and meats with the average customer spending being the highest on these items. It was found that education and parents having kids impacted the type of products purchased. Wine was mostly purchased by PhD holders and parents with no kids purchased the most products. The analysis suggested that the most successful advertising campaign was the most recent campaign (column name: Response). The best performing purchasing channels were web and store purchases (i.e. the average customer made the most purchases via these channels). It was found that currently advertising campaigns produced the least amount of sales. Further improvements are required to improve the advertising campaigns for increased sales. Analysis can be further improved by performing correlation test of each and every variable. While cursory analysis can be conducted on this data further and more detailed insights can be obtained with a larger dataset.