TITLE : CUSTOMER DATA ANALYSIS WITH PYTHON

PRESENTED BY : GROUP 3

SUPERVISED BY : MR LAWAL

ANALYTICS TEAM

- KAYODE OLUWATOBI (TEAM LEAD)
- AKINWALE IDAYAT
- OMONIYI JOSHUA

OBJECTIVES

Analysis of customer data in order to visualize purchasing behaviors, and assess the effectiveness of marketing campaigns.

ANALYSIS PROCESS

- Data Import and Cleaning
- Exploratory Data Analysis (EDA)
- Purchasing Behavior Analysis
- Marketing Campaign Effectiveness
- Predictive Analysis

DATA IMPORTATION AND CLEANING

- **Loading libraries :** These Libraries were loaded before attempting the importation of the dataset, This was done because to import a dataset, one of the libraries (IMPORT PANDA as PD) is responsible for the importation of the dataset.
- **Checking Nulls :** This was done to make our analysis and our result be effective. Nulls were checked in the dataset using the **df.isnull().sum().** This function returns the sum of all nulls in each individual columns. THERE WERE NO NULLS IN THE DATASET.
- **Checking for Duplicates :** This is another important aspect of data cleaning, This was done with the function **df.loc[df.duplicated()].** This function returns the duplicated rows contained in the dataset. THERE WERE NO DUPLICATES IN THE DATASET.
- **Outliers Check :** This was done using the boxplot visualizations. Outliers are mostly found in Numerical columns. Outliers were found in most of the numerical columns contained in the dataset and they were not removed but they were capped instead. Capping outliers helps to mitigate the impact of extreme values that can skew the results or affect the performance of models. This was done using the **INTERQUARTILE RANGE,** two bounds were created “LOWER BOUND” and the “UPPER BOUND”, Any value that fell below and above the lower and the upper bounds were considered outliers. These outliers were replaced with the Lower bound value and the Upper bound value.
-

DATA IMPORTATION AND CLEANING (Cont'd)

- **Renaming columns :** Some columns were renamed for easy readability.
- **Creating the Age column :** This was done by creating a variable for the current year and the YEAR_OF_BIRTH column was deducted from the current year (2024).
- **Removing extra spaces :** This was effected in both the columns and the rows. The INCOME column had extra space before it. The **TRIM()** function gets rid of extra spaces in a dataset.
- **Changing Datatypes :** This was done on the INCOME column, the column being a numerical column was discovered to be in the object data type. The Income column datatype was changed to the INT datatype.
- **Creating Year column :** The Dt_customer which was one of the columns that was renamed was renamed as Cust_reg_date, The column was converted to the date_time datatype before year column and the month name was extracted from the column using the **dt.year** and **dt.strftime('%B')** respectively.

EXPLORATORY DATA ANALYSIS

Genrating Descriptive statistics to understand the distribution data

	ID	Year_Birth	Income	Num_kids_home	Num_Teens_home	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	Num_Discount_purchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Age	Year	
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000
mean	5588.353339	1968.867329	52247.251354	0.441787	0.505415	49.012635	303.272789	21.569043	151.268389	32.189079	21.950812	39.497518	2.216155	4.067690	2.632671	5.800993	5.319043	55.132671	2013.028430	
std	3249.376275	11.770856	25173.076661	0.536896	0.544181	28.948352	331.811951	26.923702	179.016404	40.411139	27.539044	39.874746	1.536238	2.645994	2.736675	3.250785	2.425359	11.770856	0.685618	
min	0.000000	1932.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	28.000000	2012.000000	
25%	2814.750000	1959.000000	35303.000000	0.000000	0.000000	24.000000	24.000000	2.000000	16.000000	3.000000	1.000000	9.000000	1.000000	2.000000	0.000000	3.000000	3.000000	47.000000	2013.000000	
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	174.500000	8.000000	68.000000	12.000000	8.000000	24.500000	2.000000	4.000000	2.000000	5.000000	6.000000	54.000000	2013.000000	
75%	8421.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	505.000000	33.000000	232.250000	50.000000	33.000000	56.000000	3.000000	6.000000	4.000000	8.000000	7.000000	65.000000	2013.000000	
max	11191.000000	1996.000000	666666.000000																	

This was done by applying the **describe()** function. The resulting table contained as shown contained summary statistical calculation of each Numerical columns in the dataset. This function doesn't work on categorical columns.

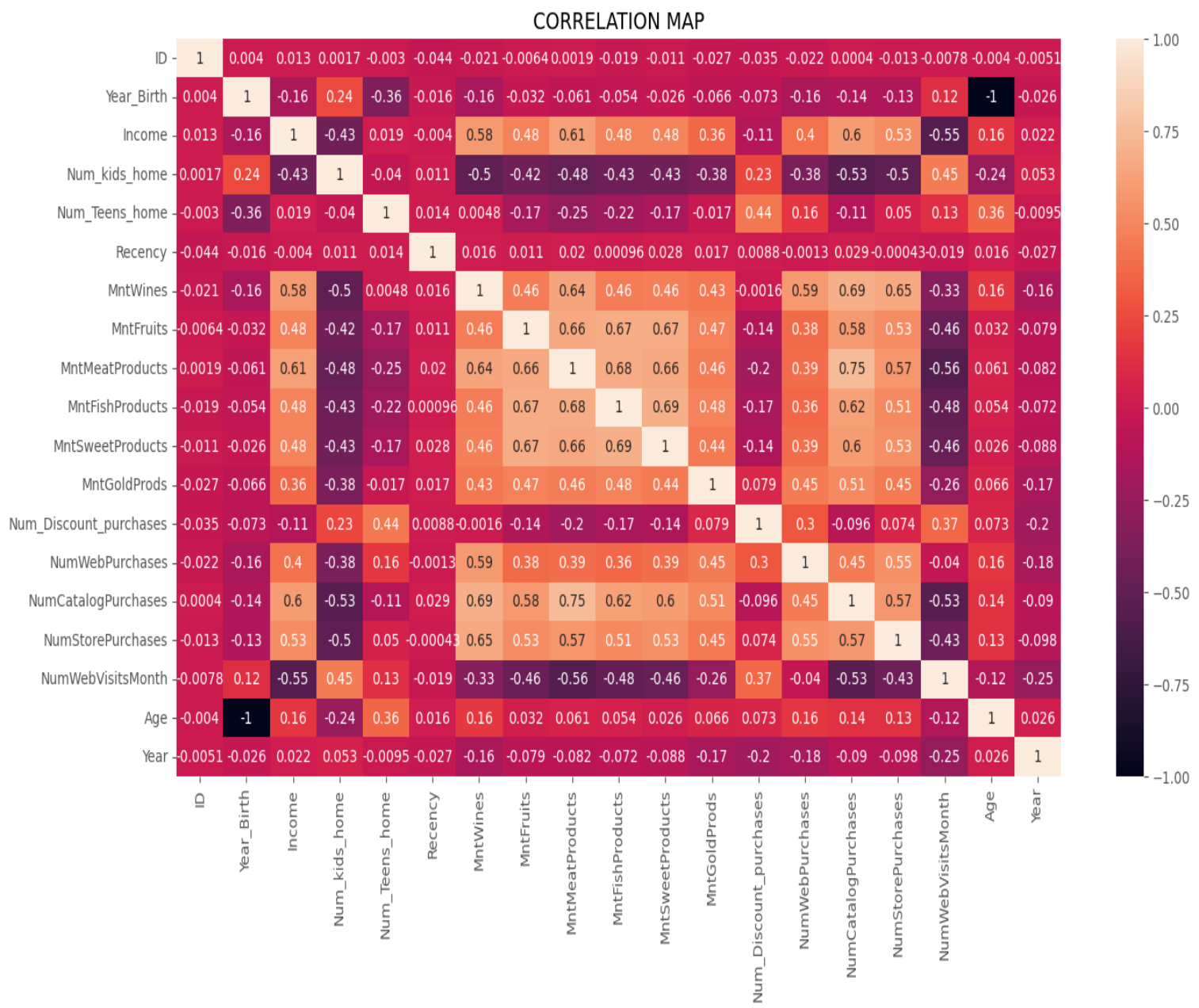
EXPLORATORY DATA ANALYSIS (Cont'd)

RELATIONSHIPS BETWEEN VARIABLES

The table shows the relationships between the columns in the dataset. It describes how the variables interact with each other

	ID	Year_Birth	Income	Num_kids_home	Num_Teens_home	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	Num_Discount_purchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Age	Year
ID	1.000000	0.004009	0.013095	0.001736	-0.003030	-0.044376	-0.021126	-0.006438	0.001924	-0.019028	-0.010899	-0.026843	-0.034962	-0.021974	0.000395	-0.013070	-0.007794	-0.004009	-0.005071
Year_Birth	0.004009	1.000000	-0.163459	0.236879	-0.358145	-0.016386	-0.164102	-0.031723	-0.061008	-0.054108	-0.025874	-0.066470	-0.072939	-0.158566	-0.140501	-0.134109	0.122935	-1.000000	-0.025769
Income	0.013095	-0.163459	1.000000	-0.428669	0.019133	-0.003970	0.582337	0.477692	0.613637	0.478098	0.480148	0.360192	-0.114962	0.403123	0.599316	0.529362	-0.553088	0.163459	0.022451
Num_kids_home	0.001736	0.236879	-0.428669	1.000000	-0.039869	0.011492	-0.501882	-0.421730	-0.483936	-0.431856	-0.429032	-0.380421	0.227367	-0.379850	-0.533530	-0.501349	0.447477	-0.236879	0.052943
Num_Teens_home	-0.003030	-0.358145	0.019133	-0.039869	1.000000	0.013838	0.004833	-0.174525	-0.253139	-0.215464	-0.171441	-0.016961	0.435220	0.163100	-0.112345	0.049737	0.131240	0.358145	-0.009488
Recency	-0.044376	-0.016386	-0.003970	0.011492	0.013838	1.000000	0.015988	0.011264	0.019609	0.000955	0.027680	0.017157	0.008791	-0.001337	0.029065	-0.000434	-0.018564	0.016386	-0.027064
Mnt Wines	-0.021126	-0.164102	0.582337	-0.501882	0.004833	0.015988	1.000000	0.456118	0.640556	0.456703	0.461539	0.434035	-0.001565	0.585316	0.689392	0.648922	-0.325862	0.164102	-0.155884
MntFruits	-0.006438	-0.031723	0.477692	-0.421730	-0.174525	0.011264	0.456118												

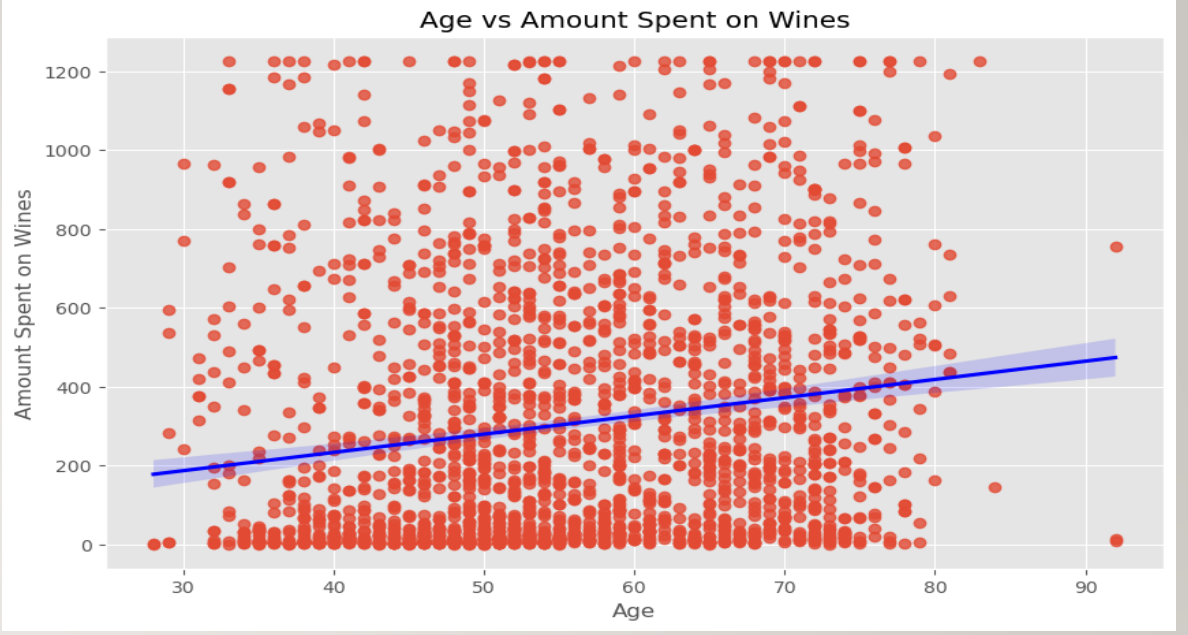
- Correlation heatmap shows the relationship between two or more numerical variables in h
- Correlation values ranges from 0 to 1. 0-0.19 is regarded as **very weak**, 0.2-0.39 as **weak**, 0.40-0.59 as **moderate**, 0.6-0.79 as strong and 0.8-1 as very strong correlation.
- For instance, the correlation between the **MntSweetProducts** and **Income** is 0.48, this value falls in the **Moderately correlated** range.

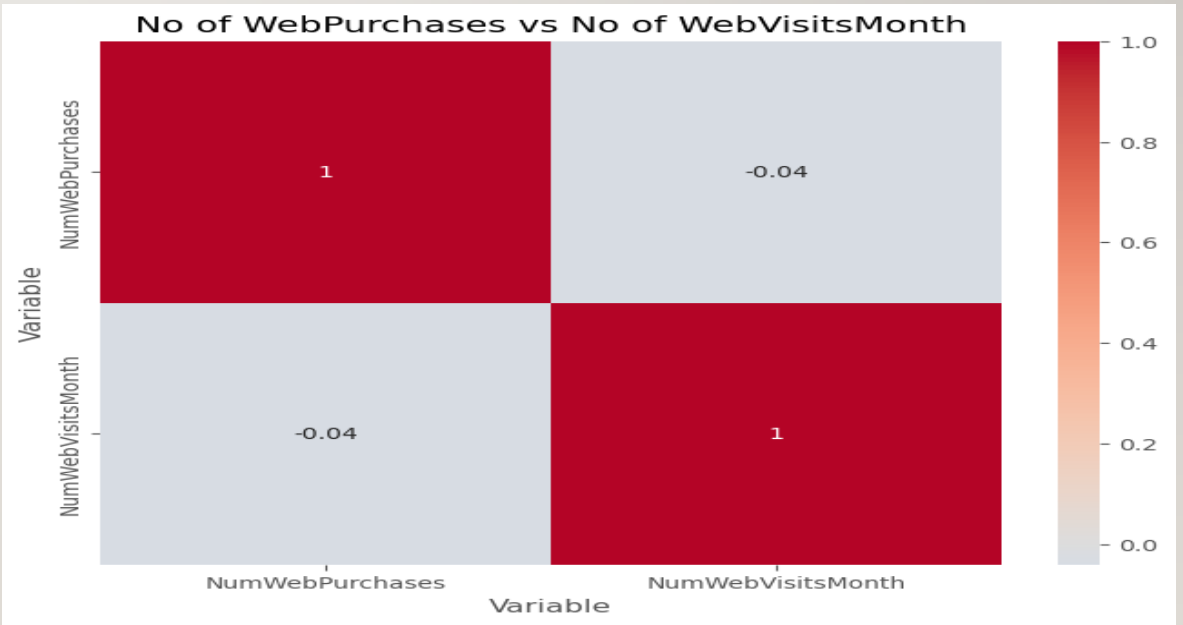


- The First Visualization shows the relationship between the Age and the amount spent on wine.

  This was done using the SNS library and the REGPLOT visual to display the relationship. The regression line in blue shows that the amount spent on wine increases with Age.
- Although the correlation coefficient which is 0.16 shows very weak relationship
- The second visualization shows the correlation between the Number of web visits and Number of web purchases.

  The correlation coefficient of this relationship is -0.04. This tells us that they are negatively correlated. The Visualization was done using the SNS library and the HEATMAP visualization.





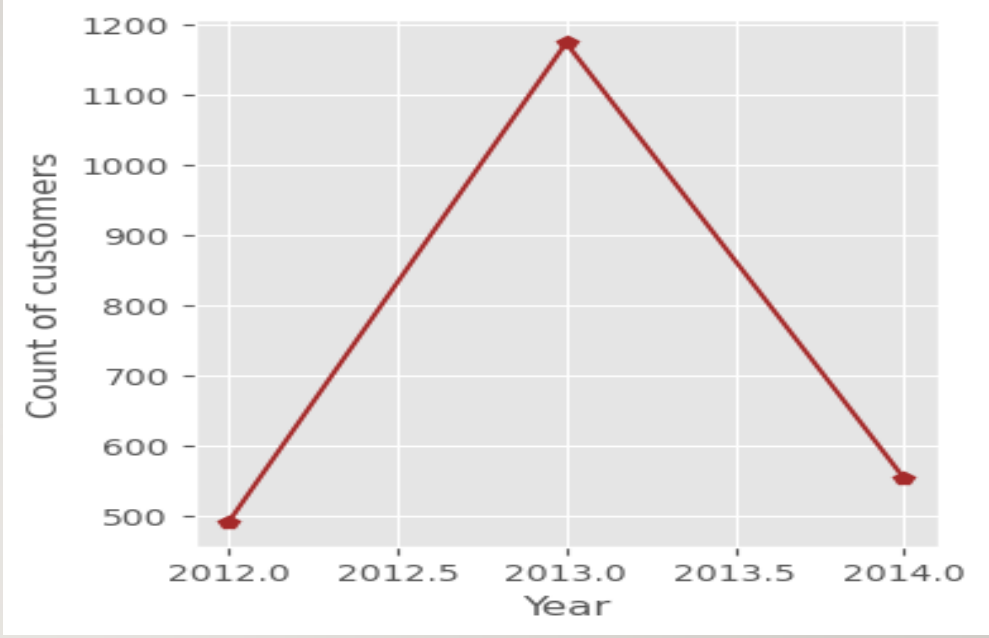
IDENTIFYING TRENDS IN THE DATASET

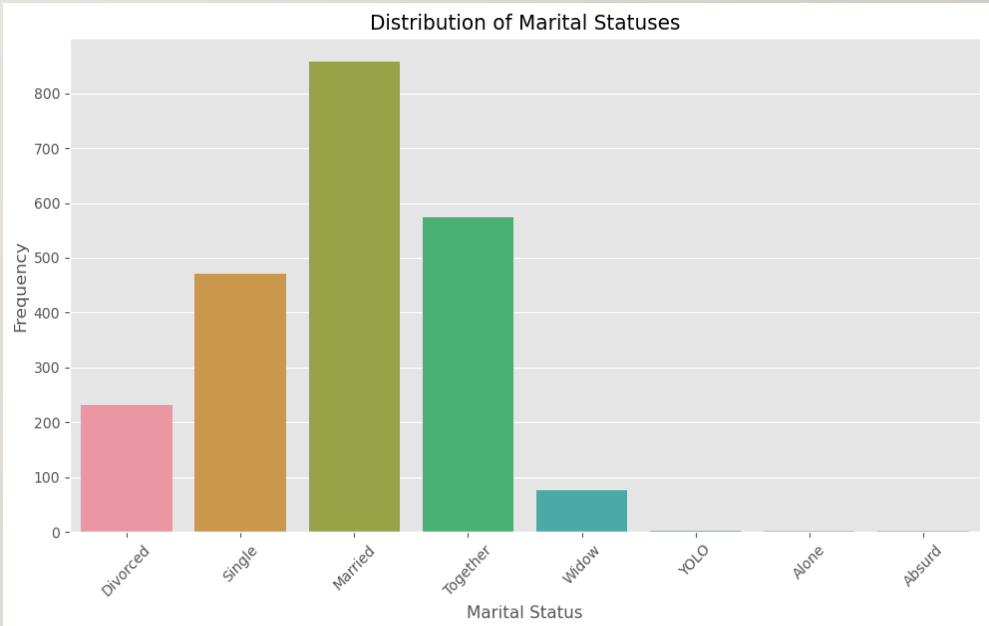
- The first Visualization shows the count of customers over the years. This was done by employing the use of the PLOT and the LINE functions of Visualizations. The diagram displays a spike in the year **2013.**
- According to research, there were some factors behind this spike and some of them includes

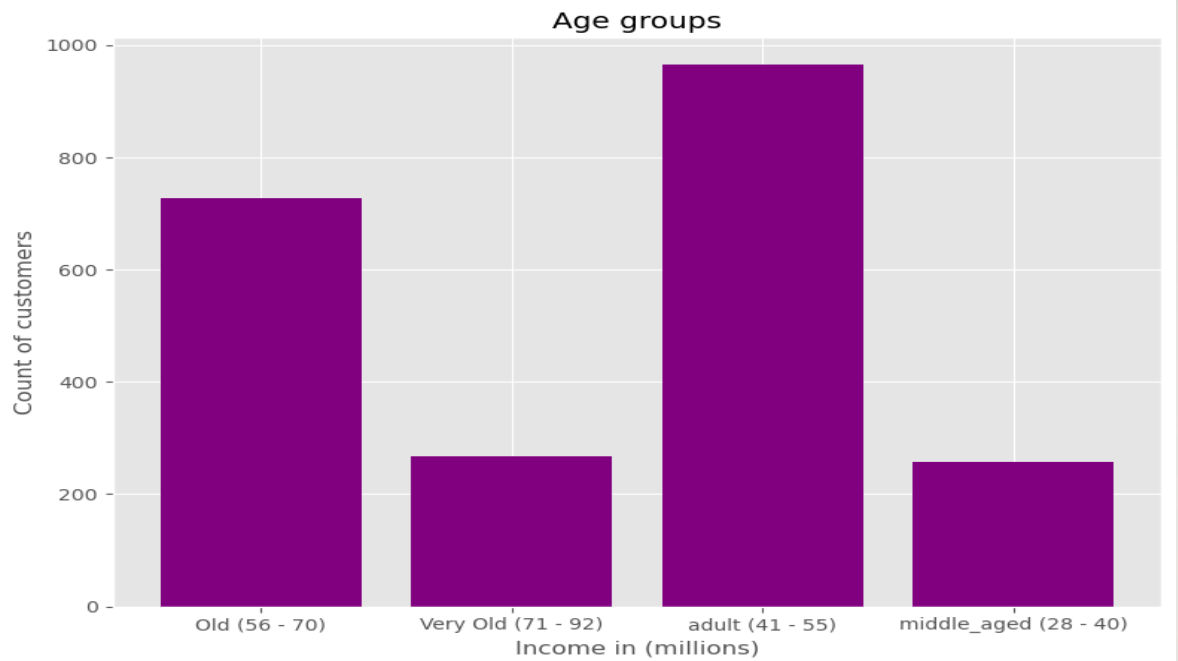
a. **Minimum Wage Increases**: Several states and municipalities increased their minimum wage rates in 2013.

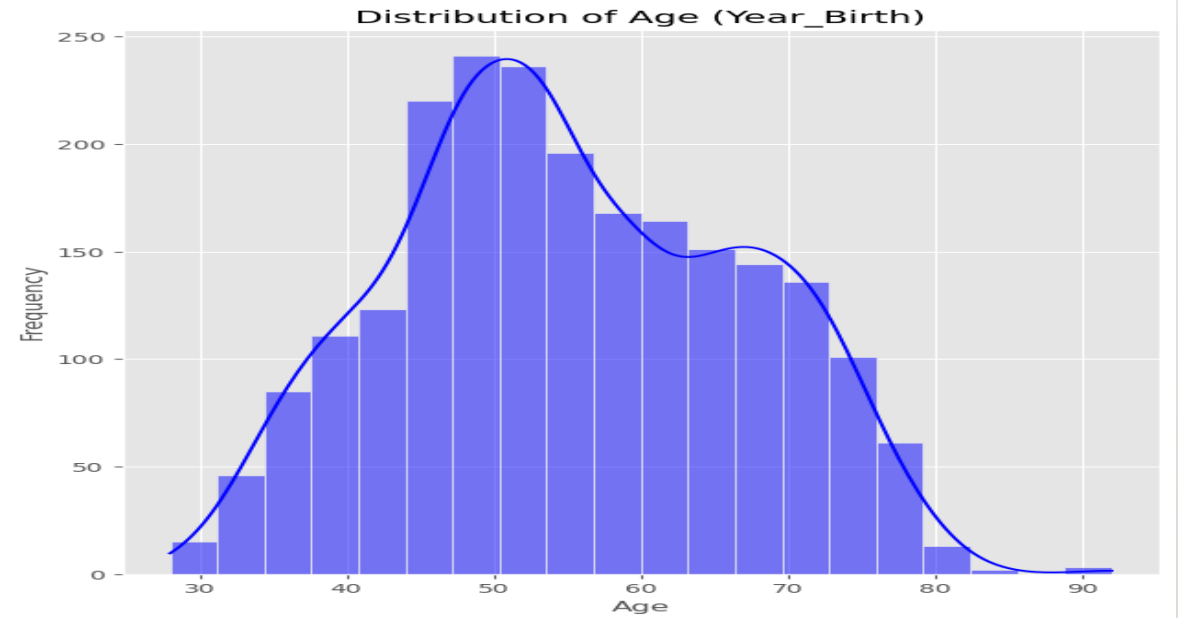
b. **Federal Reserve's Monetary Policy**: The Federal Reserve maintained a policy of low interest rates and continued its quantitative easing program, which aimed to stimulate economic growth by making borrowing cheaper and encouraging investment. This policy helped support the recovery and had indirect effects on employment and income levels.

- The second viz shows the distribution of marital statuses in the dataset, with the Married group been the most dominant of all statuses with over 800 customers. This Visualization was implemented using the SNS.COUNTPLOT visualization function.





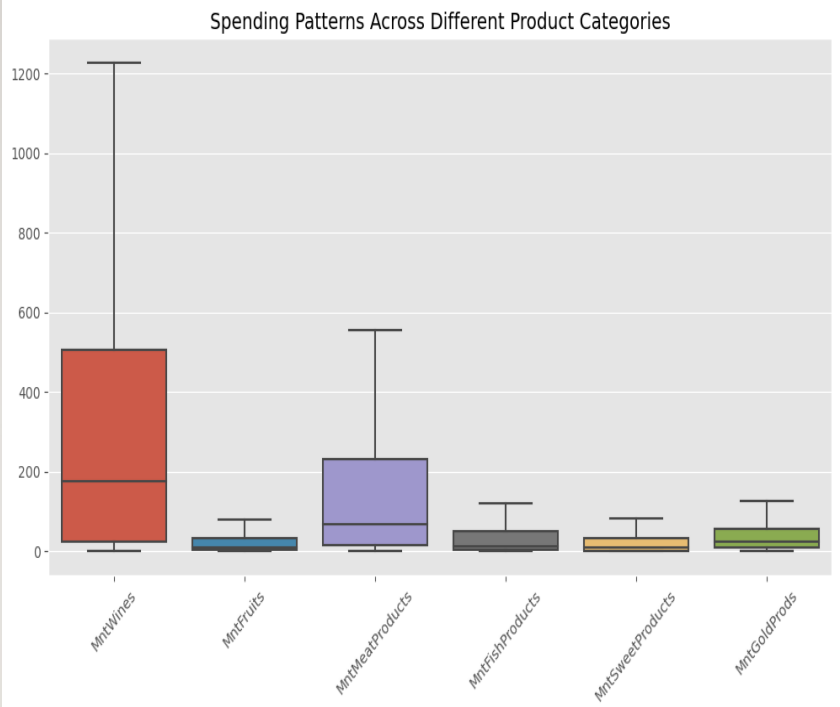


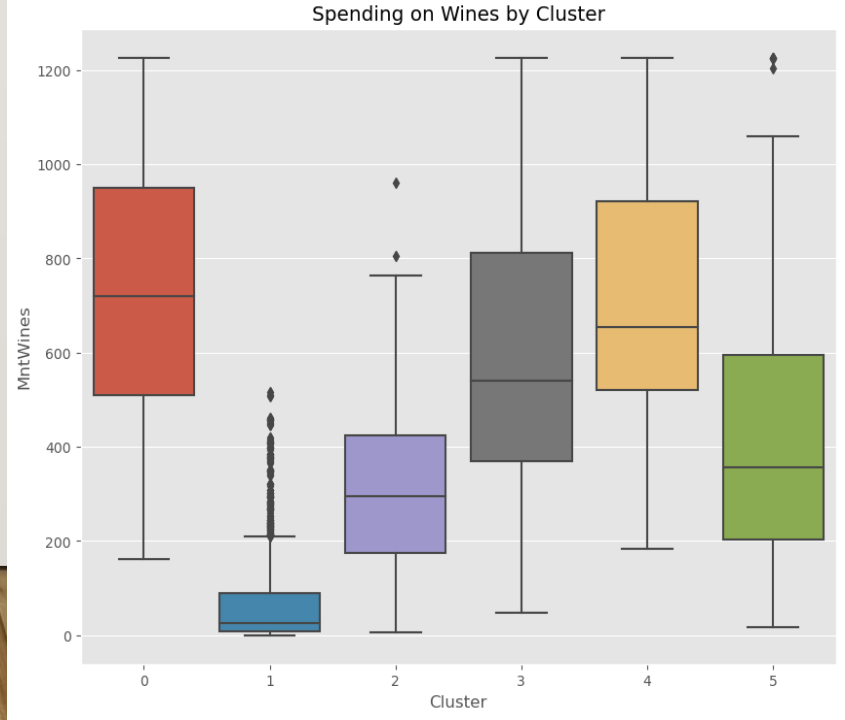


- In the First visualization, The bar graph shows the distribution of the age groups among the customers in the dataset.
- The Age group column was created because the age column has a unique value of 57 which might make our data visualizations messy, so the ages were grouped using a self_defined range for easy usage and readablility.
- Firstly, the MAX() and the MIN() function was used to find the Maximum and the Minimum value for the age respectively.
- The age range **28 – 40** was considered to be MIDDLE_AGED, the **41 – 55** was considered to be ADULT, the **56 – 70** was considered to be OLD and the **71 – 92** was considered to be VERY OLD.
- The second visualization is a Histogram shows the distribution of the age column in the dataset. According to the diagram, it was shown that the 50-60 years group are the dominant age in the dataset. This visualization was done by using the SNS.HISTPLOT

ANALYSIS ON PURCHASING BEHAVIOR

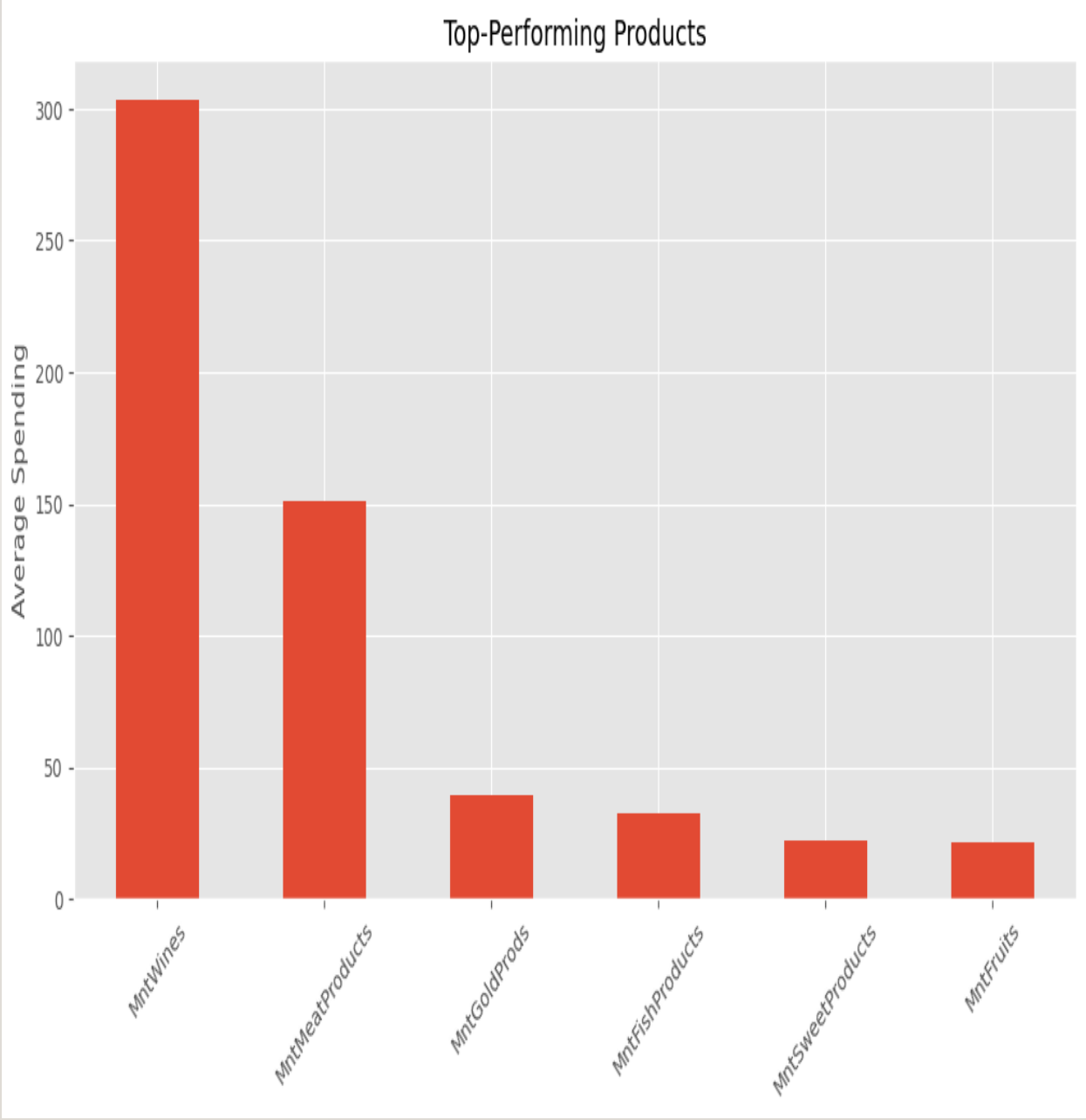
- The first boxplot shows the different product categories. "MntWines" (Money spent on wines) has the highest median spending, as well as the largest range and interquartile range (IQR). This indicates that spending on wines varies significantly among customers, with some spending much more than others.
- "MntMeatProducts" (Money spent on meat products) has the second-highest median spending and a moderate range and IQR. This suggests that meat products are also a significant category for customer spending, but with less variation compared to wines.
- The second boxplot contains Clusters. The distribution of spending within each cluster shows different spending behaviors. Cluster 0's spending is highly dispersed, showing a varied spending on wines.
- Clusters 1, 2, and 5 have more tightly packed spending, showing less variation and indicating a more consistent spending pattern within these clusters.
- Clusters 3 and 4 shows moderate spending with some variation, implying a mix of consistent and varied spending behaviors.
- This clustering technique used in this analysis is the K_MEANS CLUSTERRING TECHNIQUE.





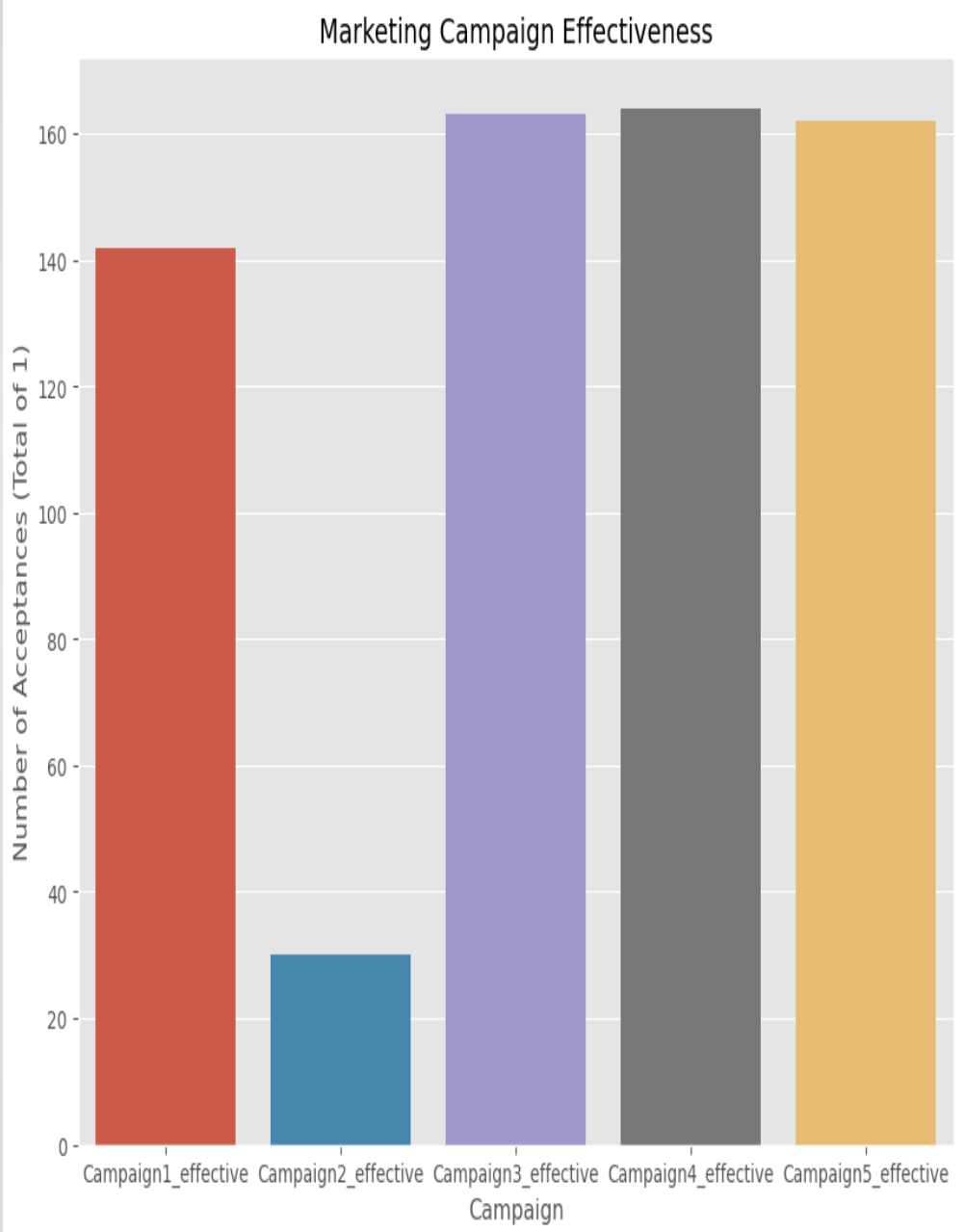
IDENTIFY TOP-PERFORMING PRODUCTS AND CUSTOMER SEGMENTS

- The highest average spending is on wines (MntWines), which significantly outperforms all other product categories.
- This suggests that wines are the most popular and high-demand product among the customers.
- The second highest average spending is on meat products (MntMeatProducts). Although it is less than half of the spending on wines, it indicates a substantial interest in meat products.

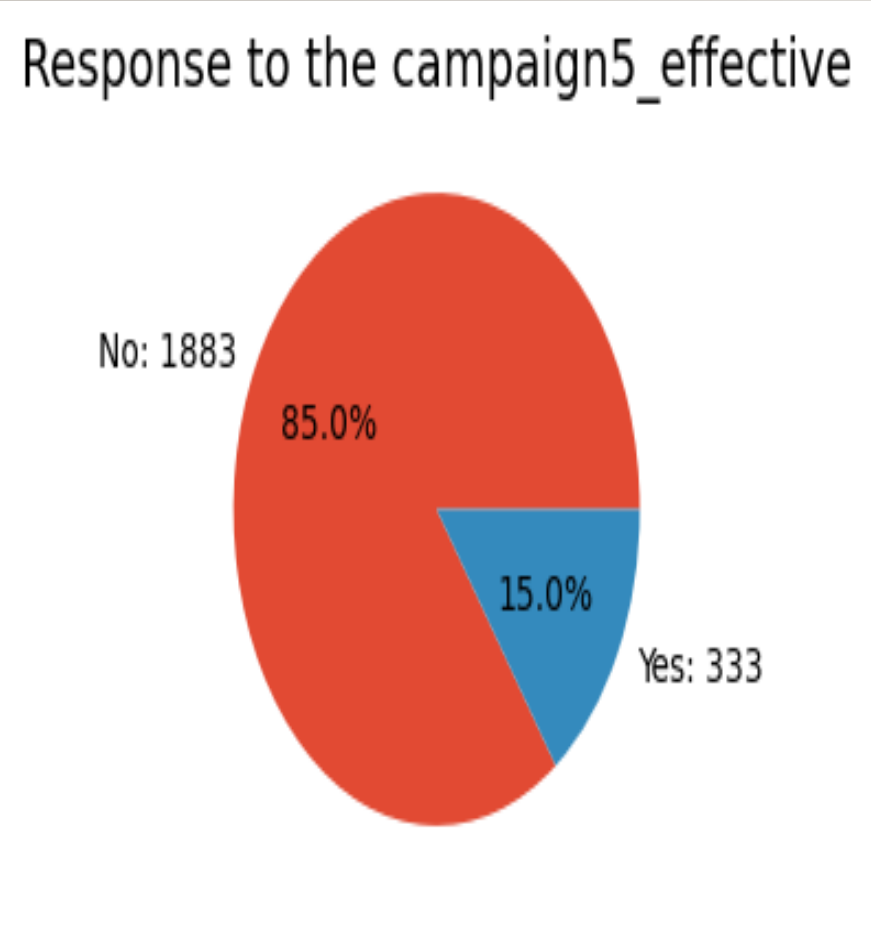


Marketing Campaign Effectiveness

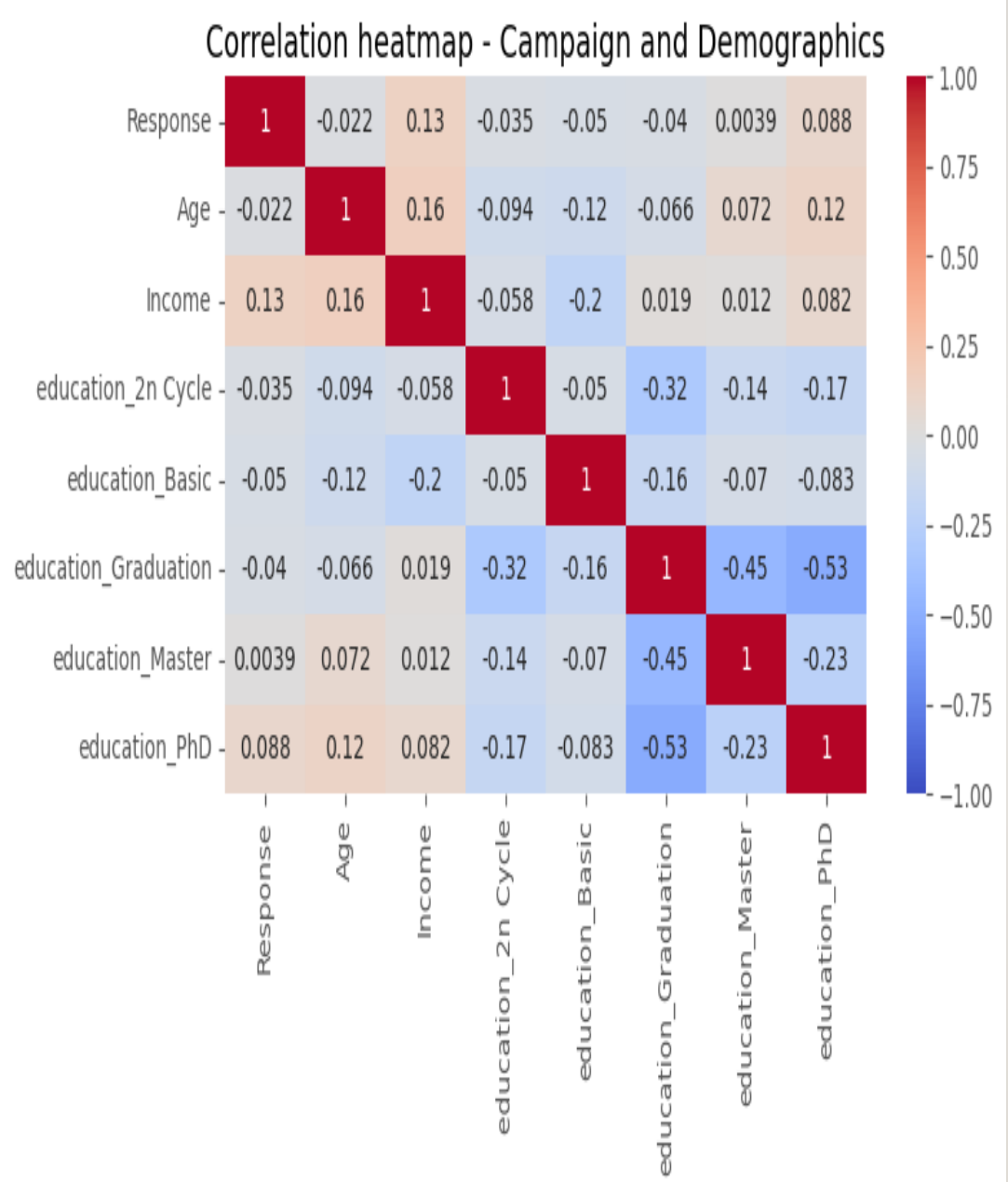
- The relevant campaign columns were selected and assigned to a variable for easy access.
- Contained in these columns are the values “0” and “1” which represents No and Yes respectively.
- The No means the campaign wasn’t effective for that particular customer while the yes which is “1” means the campaign was affected for the customer
- The campaigns were carried out five(5) times
- Since The Campaign columns contains “0” and “1” and they are numeric, The **SUM()** function was used on them to get the total number of customers that accepted the campaigns.
- The SNS.BARPLOT visualization function was used to display the distribution of the total number of acceptances.
- The fourth campaign (Campaign4_effective) was found out to be slightly higher than the third campaign which makes the **CAMPAIGN4_EFFECTIVE** to be the most effective of all the campaigns.



MODEL TO PREDICT CUSTOMERS RESPONSE TO FUTURE CAMPAIGN

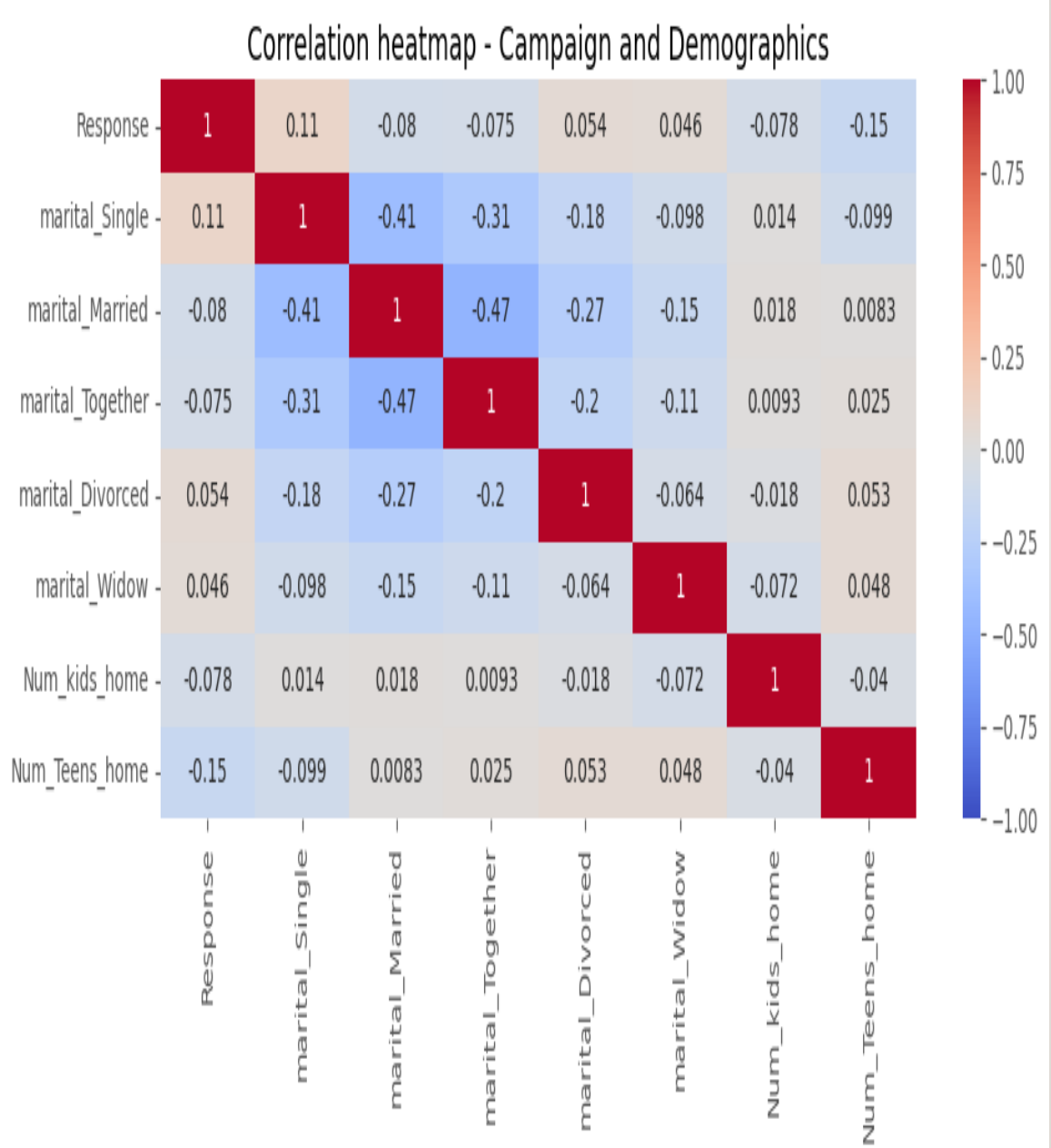


- The Pie chart displays the percentage of the response of customers to the last campaign which is the fifth campaign (Campaign5_effective).
- The chart shows that majority of the customers with **85%** rejected the last campaign.



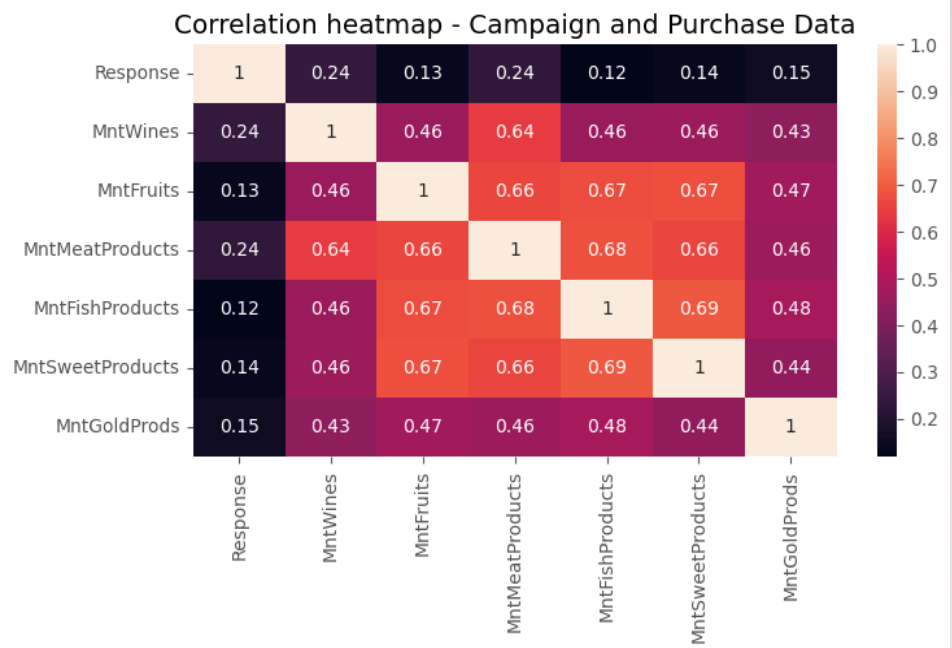
Correlation Analysis of Campaign Response and Demographic Variables

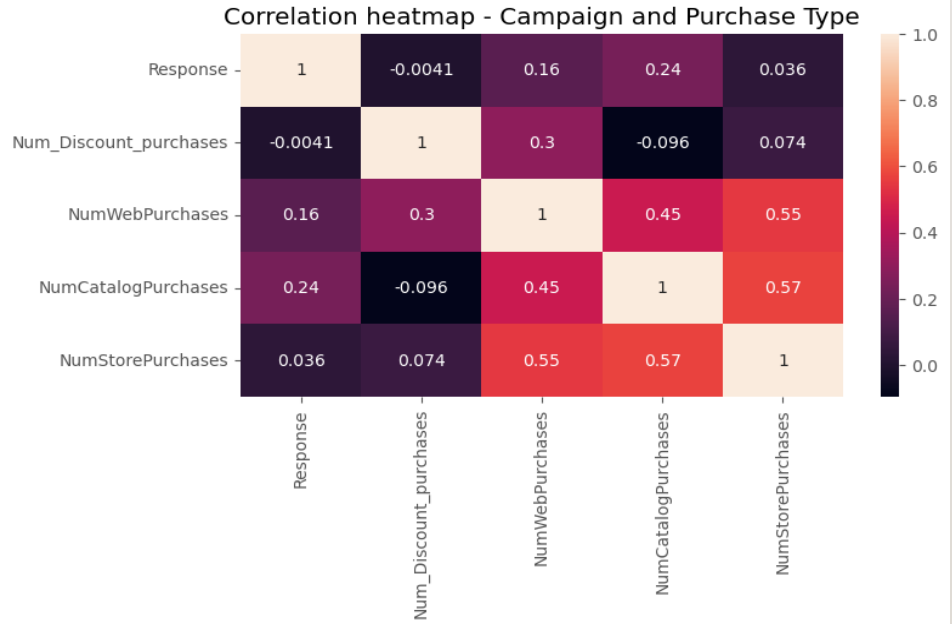
- A correlation heat map was generated to visualize the linear relationships between the response variable and selected demographic variables.
- There is a moderate positive correlation **(0.13)** between income and response rate, suggesting that individuals with higher incomes are slightly more likely to respond to the campaign.
- There is a weak negative correlation between basic education and income **(-0.20),** suggesting that individuals with basic education tend to have lower incomes.
- A slight positive correlation **(0.088)** between PhD holders and response rate indicates a marginally higher response rate among those with a PhD.



Correlation Analysis of Campaign Response and Demographic Variables

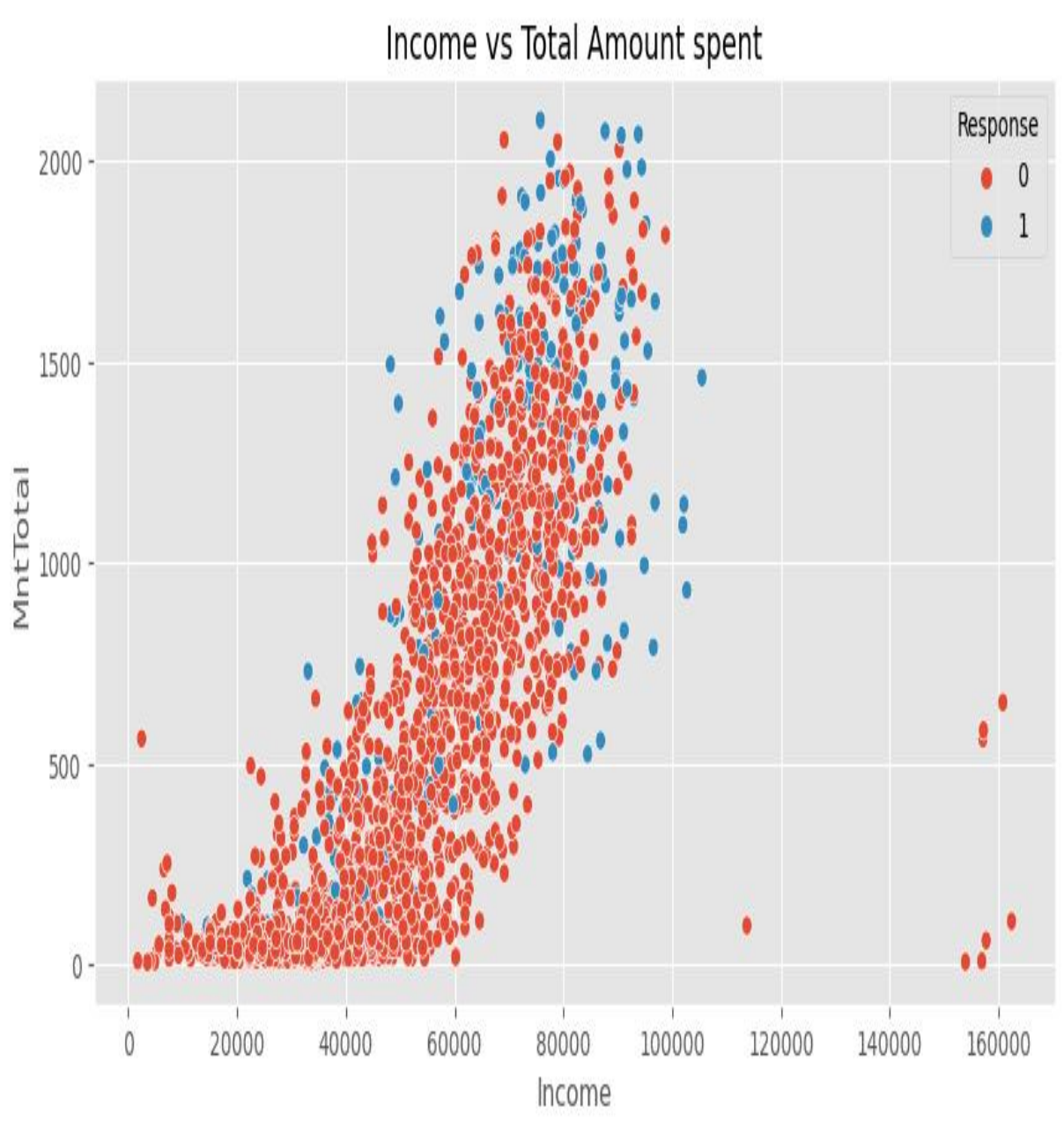
- A weak negative correlation **(-0.04)** between the number of kids and the number of teens at home suggests a slight inverse relationship between these two variables.
- Divorced and widowed individuals also show a slight tendency to respond more to the campaign, with correlations of **0.054** and **0.046,** respectively.
- Single individuals show a slightly higher response rate to the campaign with a correlation of **0.11**.
- Households with more kids tend to have a slightly lower response rate to the campaign, as shown by a weak negative correlation **(-0.078)**.





Correlation Analysis of Campaign Response and Demographic Variables

- A correlation heat map was generated to visualize the linear relationships between the response variable and product spending categories.
- There is a moderate positive correlation between campaign response and spending on wines **(0.24)** and meat products **(0.24),** indicating that customers who spend more on these products are more likely to respond to the campaign.
- Weak positive correlations exist between campaign response and spending on fruits **(0.13),** fish products **(0.12),** sweet products **(0.14),** and gold products **(0.15).** These indicate slight tendencies for higher spending in these categories to be associated with campaign response.
- The second heat map reveals that different types of purchases (web, catalog, and store) are positively correlated with each other, indicating a pattern where customers engage in multiple purchasing channels.
- The response to campaigns is more strongly correlated with catalog and web purchases than with store purchases or discount purchases.



Correlation Analysis of Campaign Response and Total amount spent on categories

- The scatter plot shows the relationship between the Total amount spent on the products by each customers and the Income.
- The Response to the campaign “0” means “No” and the “1” means Yes.
- The positive correlation between income and total spending is evident, suggesting that income is a strong predictor of spending pattern.
- Campaign responses (represented by blue points) are scattered throughout the income and spending ranges, indicating that positive responses to campaigns are not confined to specific income levels.
- Customers (around $40,000 to $80,000) exhibit a higher frequency of positive campaign responses compared to other income levels.

EVALUATING MODEL PERFORMANCES

Logistic Regression Results on Test Data
Accuracy: 0.878195
Precision: 0.838710
Recall: 0.254902

- The logistic regression model demonstrates high accuracy and precision but suffers from low recall. While it performs well overall and is reliable when predicting positive outcomes, its inability to identify a large portion of positive instances suggests room for improvement.
- The low recall indicates that the model struggles to identify a large portion of the true positive cases.

SUMMARY OF FINDINGS

- Wines are the most popularly demanded product among the product categories.
- Customers (around $40,000 to $80,000) exhibit a higher frequency of positive campaign responses compared to other income levels.
- Income is a strong predictor of spending patterns among customers
- The response to campaigns is more strongly correlated with catalog and web purchases than with other purchasing patterns.
- The fourth campaign is the most effective campaign.
- The most dominant age groups are the 50 – 60 years old customers
- The year 2013 experienced a surge in the number of customers.

RECOMMENDATIONS

- Campaign strategies could be devised to target higher-income individuals for better response rates.
- Streamlining campaign strategies to target single, divorced, and widowed individuals may improve response rates.
- Campaign strategies could be tailored to target customers who spend more on wines and meat products to improve response rates.
- The lower frequency of positive campaign responses at the lower income levels suggests that campaign strategies might need to be adjusted for this segment.
- Given the high spending on wines, targeted marketing campaigns, promotions, and loyalty programs for wine buyers could be highly effective.
- Meat products can also be a focus for promotions and bundled offers, potentially pairing them with wines to boost sales further.
- For the less prioritized categories, strategies such as discounts, special offers, or highlighting the unique aspects of these products could help increase customer interest on these categories and thereby improve the amount spent on them.