Context

“I’m a data analyst, and the Chief Marketing Officer has told me that previous marketing campaigns have not been as effective as they were expected to be. I need to analyze the data set to understand this problem and propose data-driven solutions.”- The context was paraphrased and adjusted from Here.

(Clarification: the above context was quoted and adjusted from this Kaggle Data set. I want to credit the context, idea, and many inspirations of this project back to this Kaggle dataset provider. Thank you for sharing this amazing project idea and background information related to it!)

Dataset Overview

The dataset for this project is provided by Dr. Omar Romero-Hernandez. It is licensed as CC0: Public Domain, which states, “You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission” You can also see the license status and download this dataset on this Kaggle page.

Analysis Process

1. Assessing and Cleaning the data

2. Exploratory data analysis

3. Performing Statistical Analysis

4. Data Visualization and Further Analysis

5. Forming Data-Driven Solutions

6. Give an 8-Minute Presentation to Chief Marketing Officer in the company

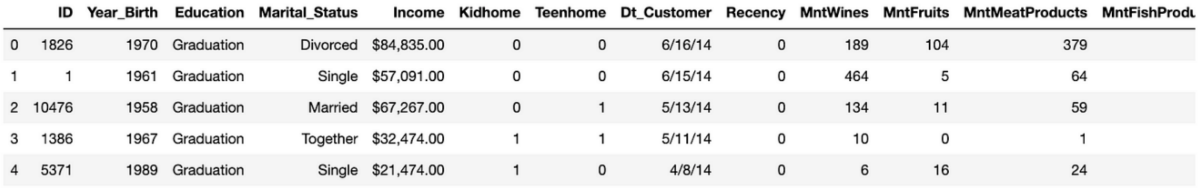
Note: This article is not meant to explain every line of code but the most important part of each analysis step. Therefore, you may find some parts that are just descriptions of the results. If you are interested in the code itself, please check here.

Step 1: Assessing and Cleaning the data

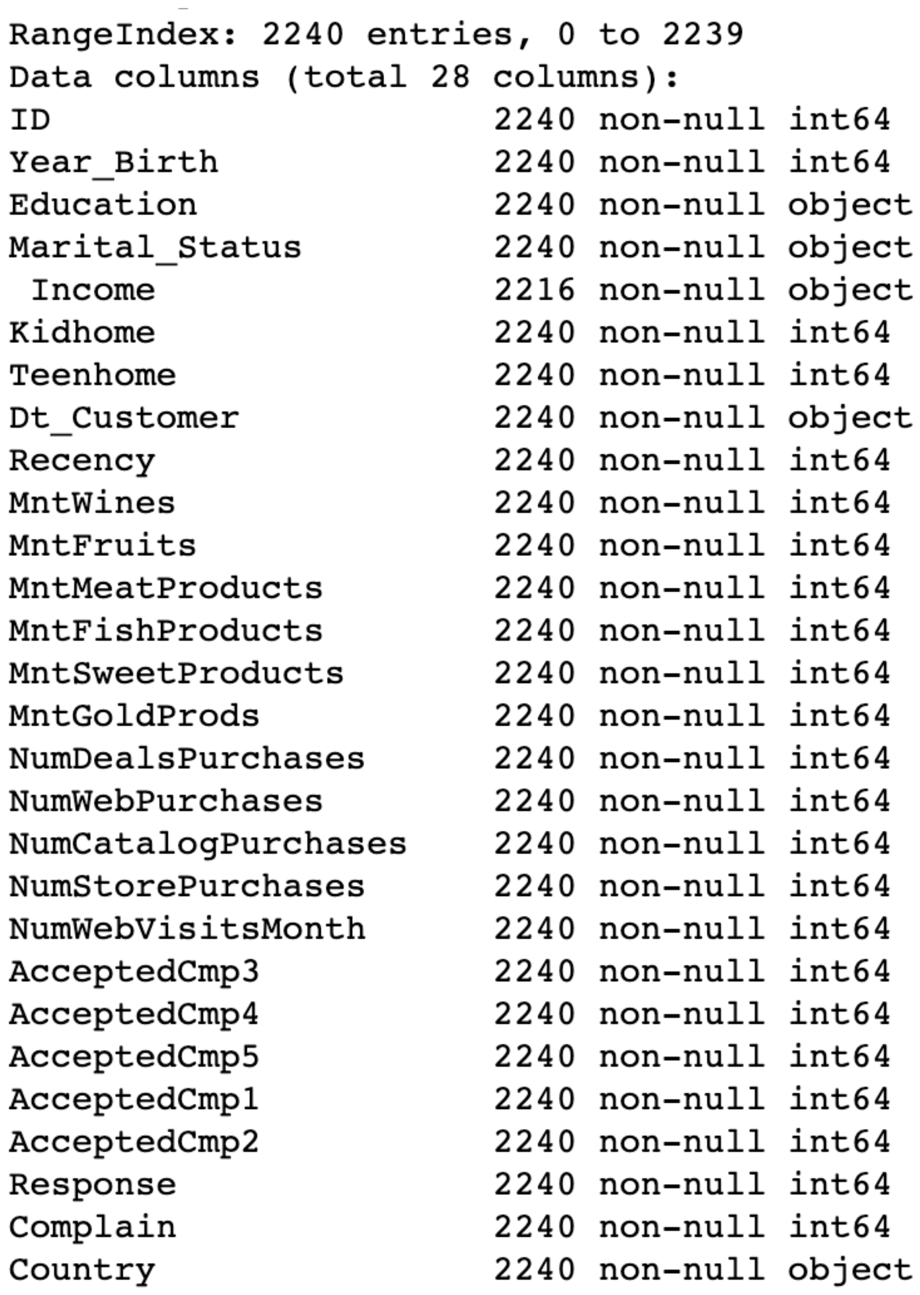
Let’s first look at the feature Information:

- 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
- Tennhome: 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
- MntSweetProducts: Amount spent on sweets in the last 2 years
- MntGoldProds: Amount spent on gold in the last 2 years
- NumDealsPurchase: Number of purchases made with a discount
- NumWebPurchase: Number of purchases made through the company’s website
- NumCatalogPurchase: Number of purchases made using a catalog
- NumStorePurchase: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company’s website in the last month
- 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
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 2 if customer accepted the offer in the 1st campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
- Complain: 1 if a customer complained in the last 2 years, 0 otherwise
- Country: Customer’s location

This dataset has 28 columns, 2240 rows, and 0 duplicated rows.



	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProdu
0	1826	1970	Graduation	Divorced	$84,835.00	0	0	6/16/14	0	189	104	379	
1	1	1961	Graduation	Single	$57,091.00	0	0	6/15/14	0	464	5	64	
2	10476	1958	Graduation	Married	$67,267.00	0	1	5/13/14	0	134	11	59	
3	1386	1967	Graduation	Together	$32,474.00	1	1	5/11/14	0	10	0	1	
4	5371	1989	Graduation	Single	$21,474.00	1	0	4/8/14	0	6	16	24	



```
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 28 columns):
ID                      2240 non-null int64
Year_Birth              2240 non-null int64
Education               2240 non-null object
Marital_Status          2240 non-null object
 Income                 2216 non-null object
Kidhome                 2240 non-null int64
Teenhome                2240 non-null int64
Dt_Customer             2240 non-null object
Recency                 2240 non-null int64
MntWines                2240 non-null int64
MntFruits               2240 non-null int64
MntMeatProducts         2240 non-null int64
MntFishProducts         2240 non-null int64
MntSweetProducts        2240 non-null int64
MntGoldProds            2240 non-null int64
NumDealsPurchases       2240 non-null int64
NumWebPurchases         2240 non-null int64
NumCatalogPurchases     2240 non-null int64
NumStorePurchases       2240 non-null int64
NumWebVisitsMonth       2240 non-null int64
AcceptedCmp3            2240 non-null int64
AcceptedCmp4            2240 non-null int64
AcceptedCmp5            2240 non-null int64
AcceptedCmp1            2240 non-null int64
AcceptedCmp2            2240 non-null int64
Response                2240 non-null int64
Complain                2240 non-null int64
Country                 2240 non-null object
```

After assessing the data, I found that several issues:

1. There is a space in front of the income’s column name

2. There are dollar signs is the values of Income column

3. The “Income” column has 23 missing values

4. Income’s type is string

5. Dt_Customer’s type is string

Since data cleaning is not the main part of this project, let's move forward to the next step. (You can find the codes on cleaning these issues Here)

Step 2: Exploratory Data Analysis

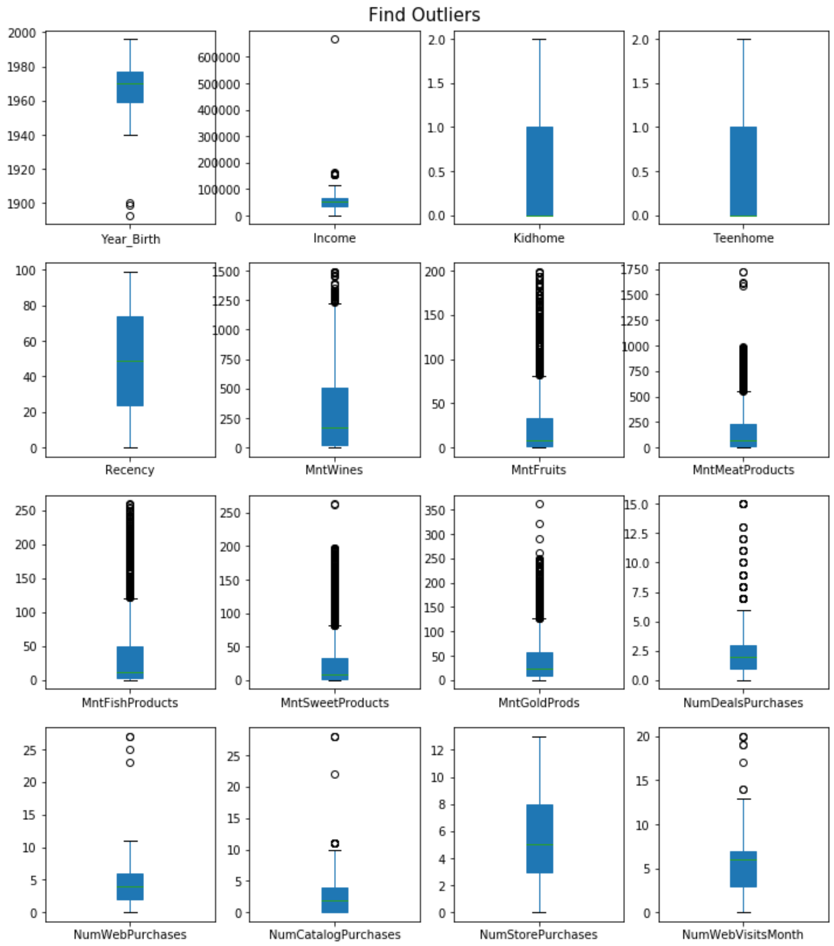
In this dataset's Kaggle page, there are some EDA directions that the data publisher suggested following, and I decided to choose the following three questions to explore:

- Are there any outliers? How will you wrangle/handle them?
- Are there any useful variables that you can engineer with the given data?
- Do you notice any patterns or anomalies in the data? Can you plot them?

Now let's look at the questions one by one.

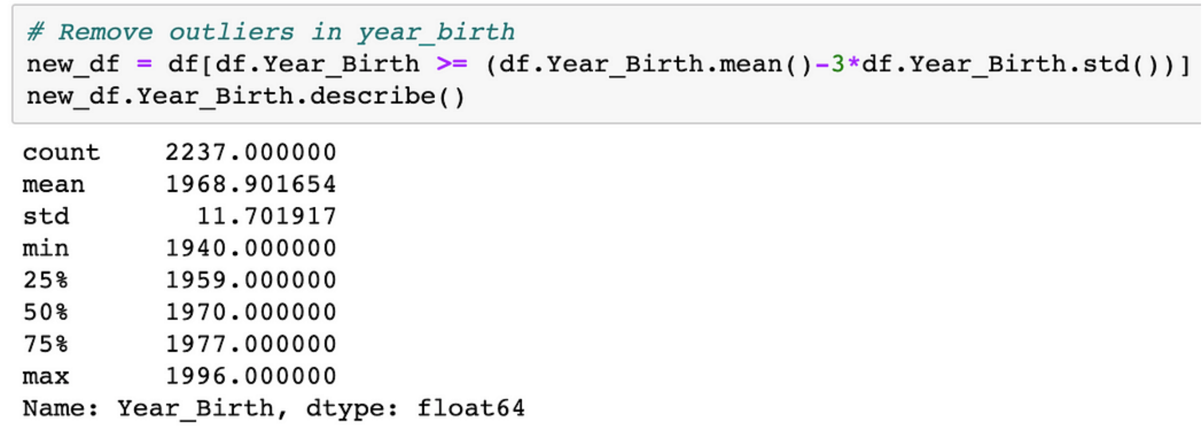
1. Are there any outliers? How will you wrangle/handle them?

I used the boxplots to visualize all the numerical features, and it will show the 5 numbers of the data: the lowest number that is not an outlier, Q1(25th percentile), Q2(50th percentile), Q3(75th percentile), and the highest number that is not an outlier.



Many columns have outliers, but most of them seem like natural outliers that came from the population. In contrast, the outliers in Year_birth seem like entry errors since it’s impossible that people born before 1900 still alive. Therefore, I will remove the outliers in Year_birth.

Outliers mean they are below or above 3 standard deviations from the mean.



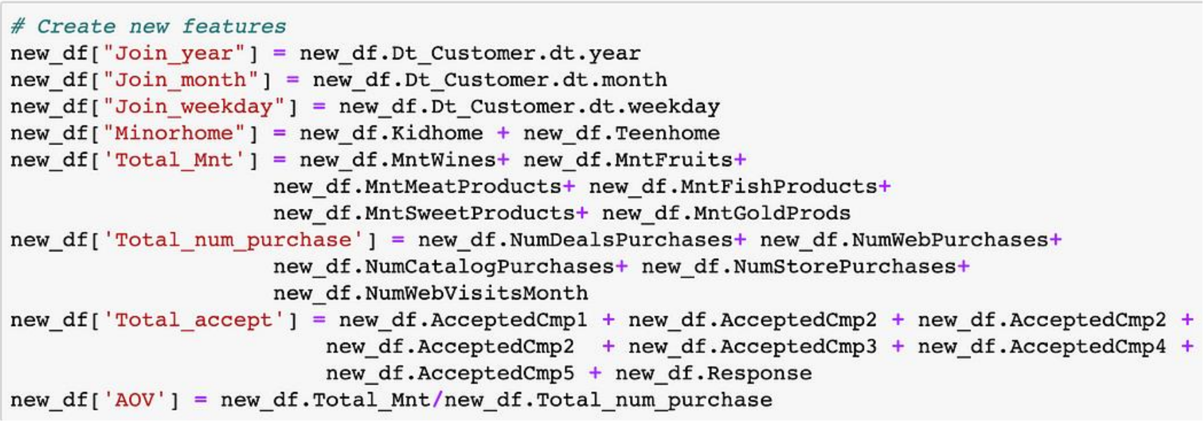
```
# Remove outliers in year_birth
new_df = df[df.Year_Birth >= (df.Year_Birth.mean()-3*df.Year_Birth.std())]
new_df.Year_Birth.describe()
```

```
count    2237.000000
mean     1968.901654
std        11.701917
min      1940.000000
25%      1959.000000
50%      1970.000000
75%      1977.000000
max      1996.000000
Name: Year_Birth, dtype: float64
```

2. Are there any useful variables that you can engineer with the given data?

After assessing the dataset, I list the new features that I think can be useful for the last analysis. For example, if we know the average month and the day of the week that the average person became a customer, then when we build a campaign on that day or in that month, it might help boost more first-time customers.

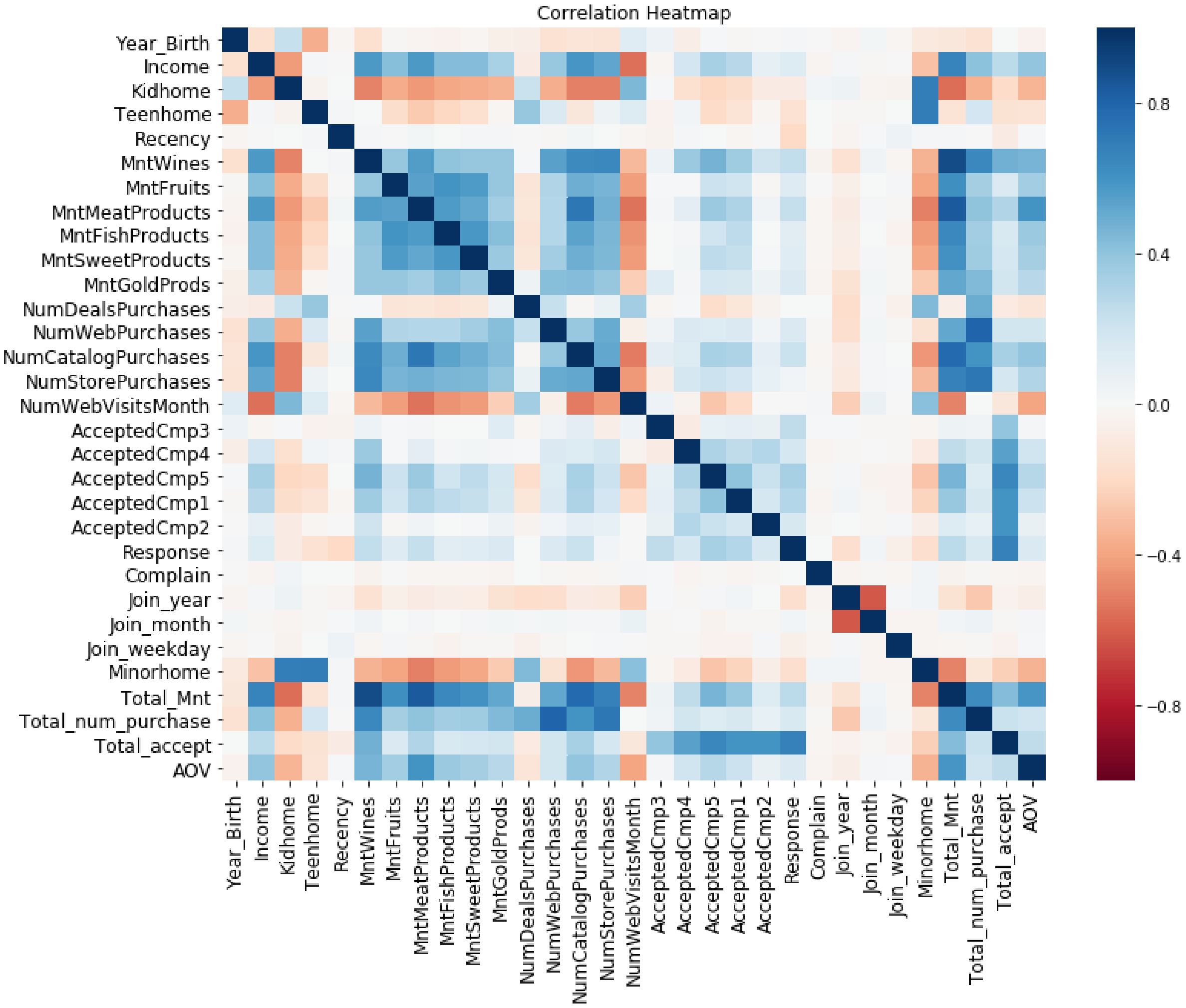
- Join_month: The month that person became a customer, which can be engineered from “Dt_Customer”
- Join_weekday: The day of the week that person became a customer, which can be engineered from “Dt_Customer”
- Minorhome: The total number of minors in their family, which can be acquired by summing up by Kidhome and Teenhome.
- Total_Mnt: Total amount spent in the last two years, which can be acquired by summing up all the “Mnt”-related columns
- Total_num_purchase: Total number of purchases in the last two years, which can be acquired by summing up all the “Num”-related columns
- Total_accept: Total amount a customer accepted the offer in all the marketing campaigns, which can be acquired by summing up all the “Accepted”-related columns and the “Response” column
- “AOV”: AOV stands for the average order volume of each customer, which can be engineered by dividing Total_Mnt by Total_num_purchase



```
# Create new features
new_df["Join_year"] = new_df.Dt_Customer.dt.year
new_df["Join_month"] = new_df.Dt_Customer.dt.month
new_df["Join_weekday"] = new_df.Dt_Customer.dt.weekday
new_df["Minorhome"] = new_df.Kidhome + new_df.Teenhome
new_df['Total_Mnt'] = new_df.MntWines+ new_df.MntFruits+
                    new_df.MntMeatProducts+ new_df.MntFishProducts+
                    new_df.MntSweetProducts+ new_df.MntGoldProds
new_df['Total_num_purchase'] = new_df.NumDealsPurchases+ new_df.NumWebPurchases+
                    new_df.NumCatalogPurchases+ new_df.NumStorePurchases+
                    new_df.NumWebVisitsMonth
new_df['Total_accept'] = new_df.AcceptedCmp1 + new_df.AcceptedCmp2 + new_df.AcceptedCmp2 +
                        new_df.AcceptedCmp2  + new_df.AcceptedCmp3 + new_df.AcceptedCmp4 +
                        new_df.AcceptedCmp5 + new_df.Response
new_df['AOV'] = new_df.Total_Mnt/new_df.Total_num_purchase
```

3. Do you notice any patterns or anomalies in the data? Can you plot them?

We can use a heatmap to see the correlations between each variable. When it gets bluer, they are more positively correlated, and when it gets redder, they are more negatively correlated.



Findings:

Patterns:

1. High-Income People
— tend to spend more and purchase more.
— tend to visit the company's website less frequently than other people.
— tend to has few numbers of purchases made with a discount

2. People having kids at home
— tend to spend less and purchase less.
— tend to has a high number of purchases made with a discount

3. People who purchased with high average order volume
— tend to buy more wines and meat products
— tend to make a high number of purchases made using a catalog
— tend not to visit the company's website.

Anomalies:

1. Intuitively, I'd think the more complaints a customer has, the less they may spend on our store, but the number of complaints in the last two years has almost no correlation with the total amount spent in the last two years. => After further investigating the data, I found that it is because we only have 20 customers who complained in the last two years, but we have 2200 customers in total. So, because of the imbalanced ratio, they don't correlate. The customer service department in the company has done a wonderful job in the last two years.

Step 3: Performing Statistical Analysis

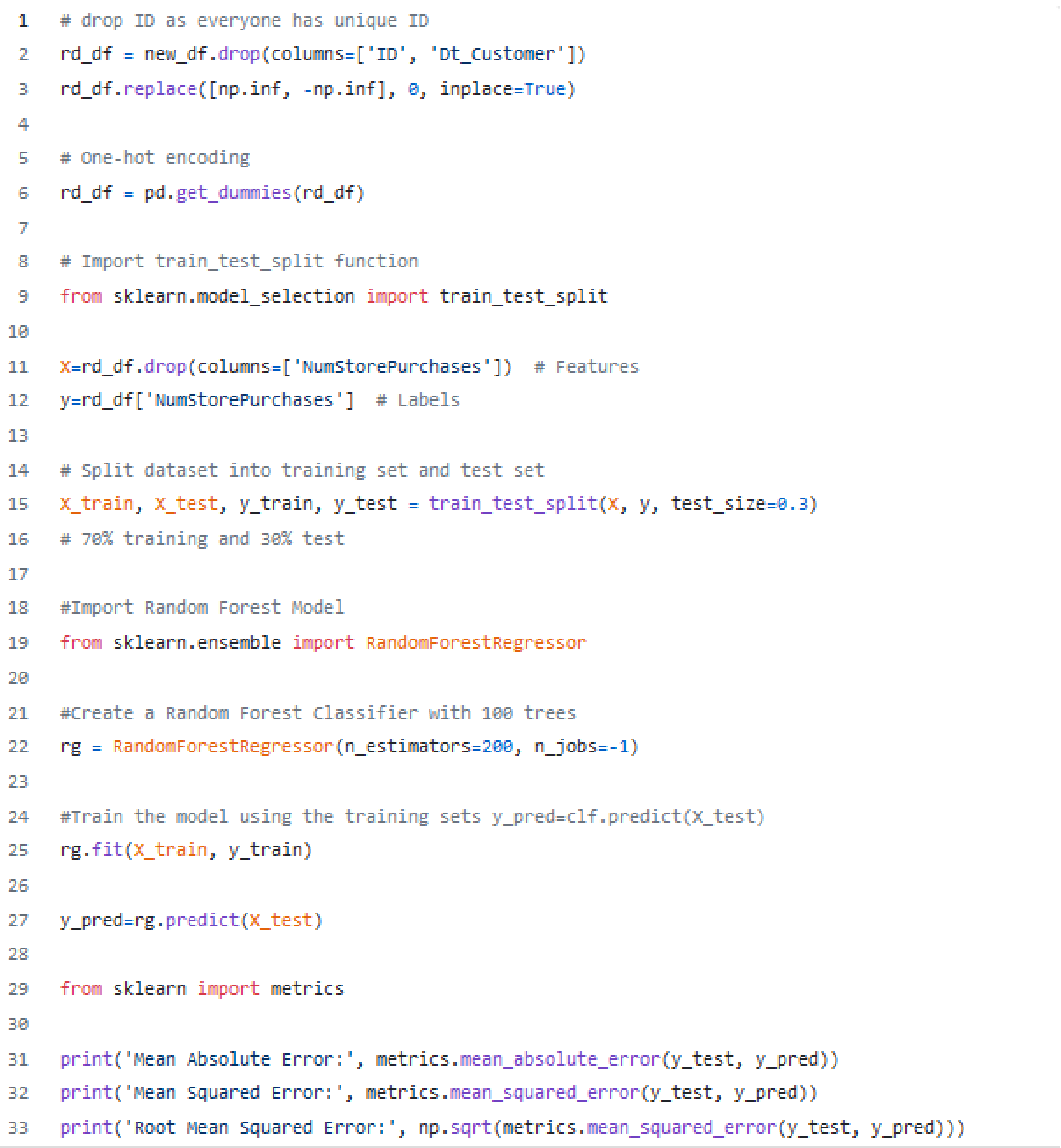
In this dataset's Kaggle page, there are some statistical analysis questions that the data publisher suggested answering, and I decided to choose the following three questions to explore:

- What factors are significantly related to the number of store purchases?
- Your supervisor insists that people who buy gold are more conservative. Therefore, people who spent an above-average amount on gold in the last 2 years would have more in-store purchases. Justify or refute this statement using an appropriate statistical test
- Fish has Omega 3 fatty acids, which are good for the brain. Accordingly, do "Married Ph.D. candidates" have a significant relation with the amount spent on fish?

Now let's look at the questions one by one.

1. What factors are significantly related to the number of store purchases?

We can use the random forest to predict store purchases and then utilize the model's feature importance score to rank the factors.



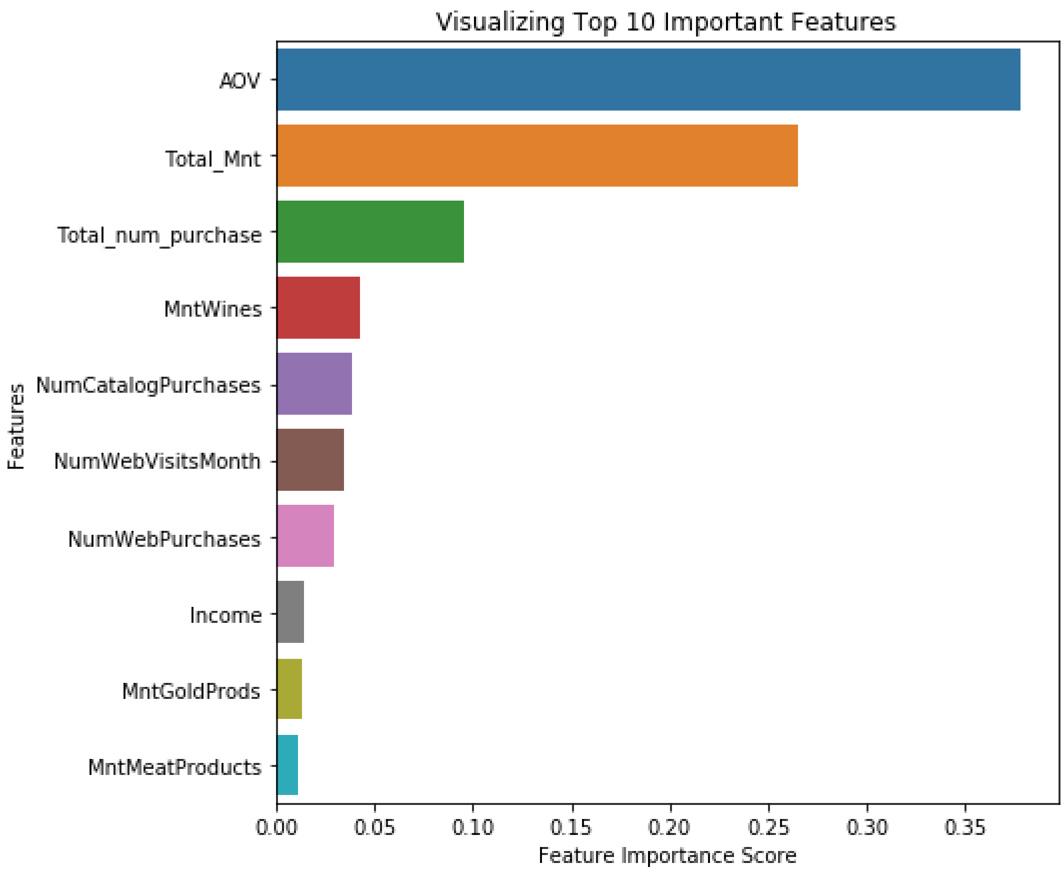
```python
# drop ID as everyone has unique ID
rd_df = new_df.drop(columns=['ID', 'Dt_Customer'])
rd_df.replace([np.inf, -np.inf], 0, inplace=True)

# One-hot encoding
rd_df = pd.get_dummies(rd_df)

# Import train_test_split function
from sklearn.model_selection import train_test_split

X=rd_df.drop(columns=['NumStorePurchases'])  # Features
y=rd_df['NumStorePurchases']  # Labels

# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
# 70% training and 30% test

#Import Random Forest Model
from sklearn.ensemble import RandomForestRegressor

#Create a Random Forest Classifier with 100 trees
rg = RandomForestRegressor(n_estimators=200, n_jobs=-1)

#Train the model using the training sets y_pred=clf.predict(X_test)
rg.fit(X_train, y_train)

y_pred=rg.predict(X_test)

from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

Result:

```
Mean Absolute Error: 0.78703125
Mean Squared Error: 1.4546007812500001
Root Mean Squared Error: 1.2060683153329252
```

The range of NumStorePurchases is 13, and the Root Mean Squared Error is only 1.2(less than 10% of the range), which means it is a reliable model.

Now, let’s use random forest’s feature importance score to see which factors most contribute to the NumStorePurchase.

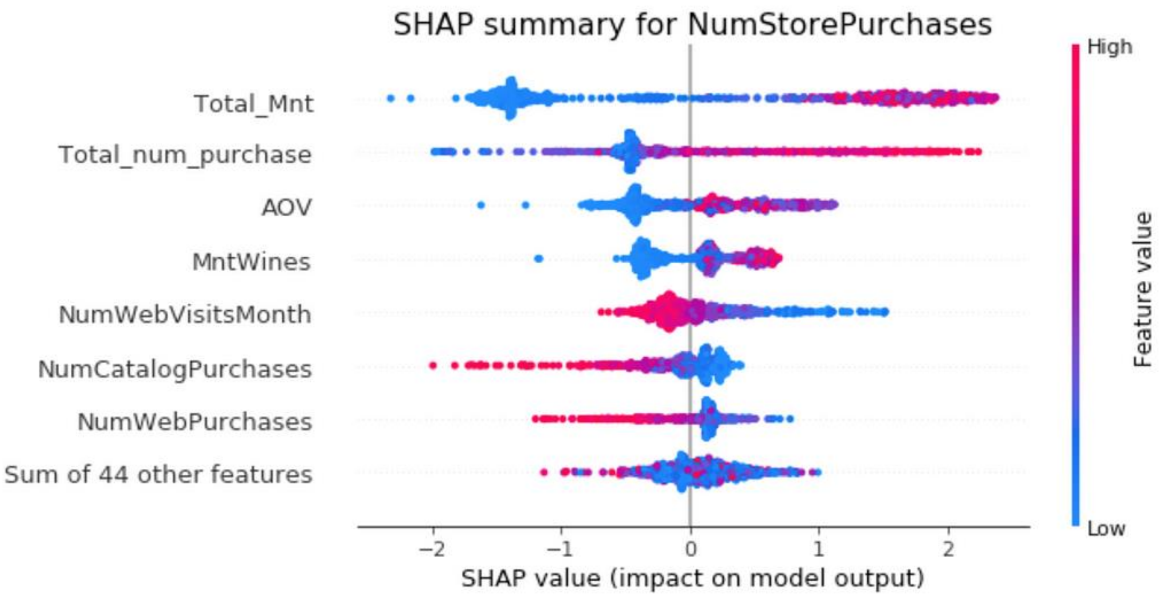


We can now see that the top 7 factors are

```
1. Average order volume
2. Total amount spent in the last two years
3. Total number of purchases in the last two years
4. Amount spent on wine in the last 2 years
5. Number of purchases made using a catalog
6. Number of visits to company's web site in the last month
7. Total number of purchases through website in the last two years
```

However, we can’t tell whether each factor is positively or negatively correlated to the number of store purchases. We can use SHAP to explain it.

There is a famous article by Samuele Mazzanti explaining what SHAP is. Please check out here.



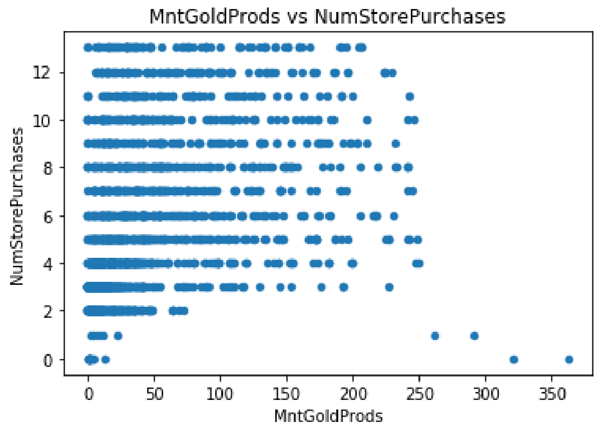
Finding:

1. The number of store purchases increases with the higher total amount spent(Total_Mnt), higher total purchase amount(Total_num_purchase), higher AOV, and higher amount of wines purchases(MntWines).
2. The number of store purchases decreases with more website visits(NumWebVisitsMonth), a higher number of purchases through the catalog(NumCatalogPurchases), and a higher number of purchases through websites(NumWebPurchases).

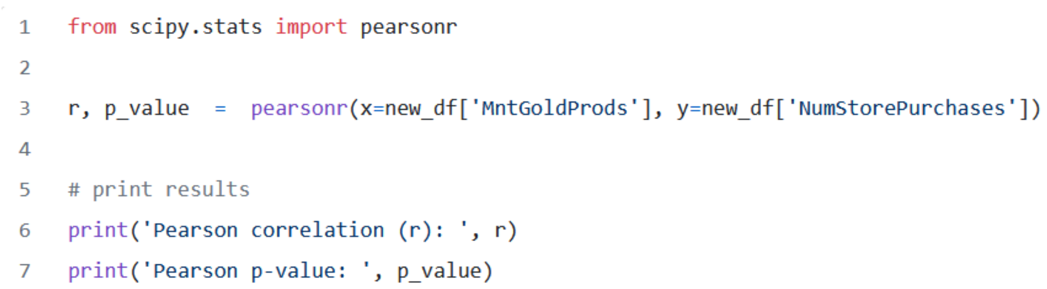
Summary: **People who mostly shop at stores tend to buy more wines, have a higher average order volume, and shop less through the internet or catalog.**

2. Your supervisor insists that people who buy gold are more conservative. Therefore, people who spent an above average amount on gold in the last 2 years would have more in store purchases. Justify or refute this statement using an appropriate statistical test.

To statistically verify this claim, we need to use a correlation test to see if MntGoldProds and NumStorePurchases are positively correlated. First, let’s look at the scatterplot of the two variables.



As we can see, there is a very vague trend that says as MntGoldProds increases, NumStorePurchases also increases. Now, let's look at the correlation test.



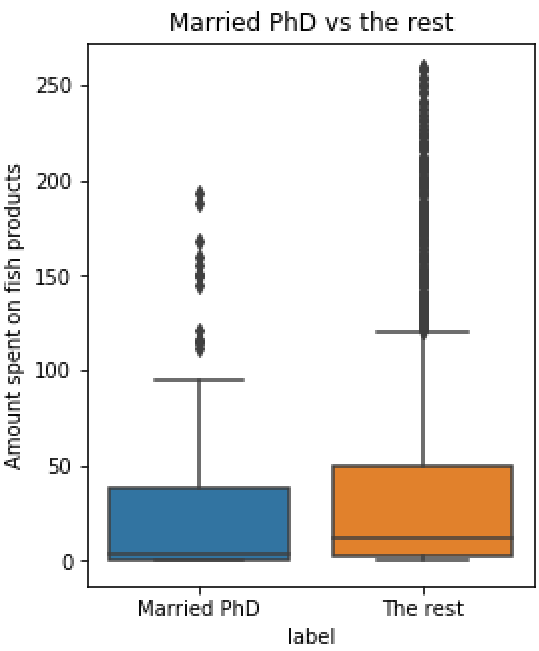
```
from scipy.stats import pearsonr

r, p_value  =  pearsonr(x=new_df['MntGoldProds'], y=new_df['NumStorePurchases'])

# print results
print('Pearson correlation (r): ', r)
print('Pearson p-value: ', p_value)
```

```
Pearson correlation (r):  0.38326418634704296
Pearson p-value:  3.4668974417790955e-79
```

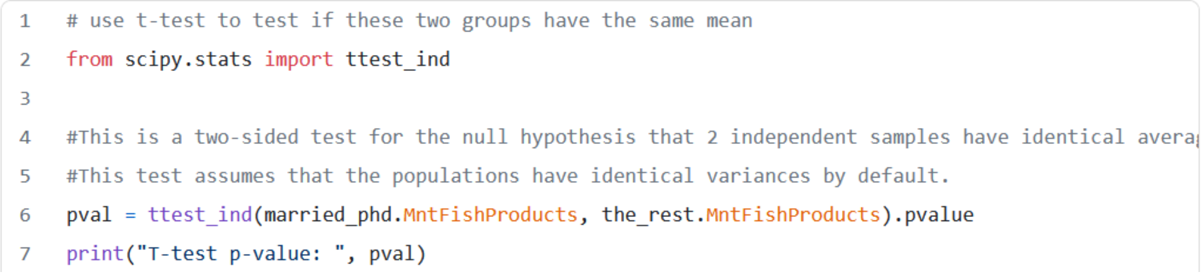
We got a Pearson correlation of 0.38 and a p-value of almost zero, which states that **they are statistically significant and have a positive correlation**. (If the p-value is > 0.05, we will fail to reject the null hypothesis, where they do not correlate.)

3. Fish has Omega 3 fatty acids which are good for the brain. Accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish?

To statistically verify these, I first divide the data into two groups. One is the married Ph.D. group and the rest. And then, we can use a boxplot to visualize these two groups to see if they are different. Lastly, we can use a t-test to test whether their mean is similar.



This plot shows that the rest of the customers spent more on fish products as its 50th percentile is higher than the married Ph.D. group. Now, let’s look at the t-test.



```
# use t-test to test if these two groups have the same mean
from scipy.stats import ttest_ind

#This is a two-sided test for the null hypothesis that 2 independent samples have identical avera
#This test assumes that the populations have identical variances by default.
pval = ttest_ind(married_phd.MntFishProducts, the_rest.MntFishProducts).pvalue
print("T-test p-value: ", pval)
```

```
T-test p-value:  0.005297012242158541
```

Since the p-value is less than 0.05, I concluded that we reject the null hypothesis, meaning that their means are different, but the Married Ph.D.’s mean is lower than the rest, as we can see from the graph.

Step 4: Data Visualization and Further Analysis

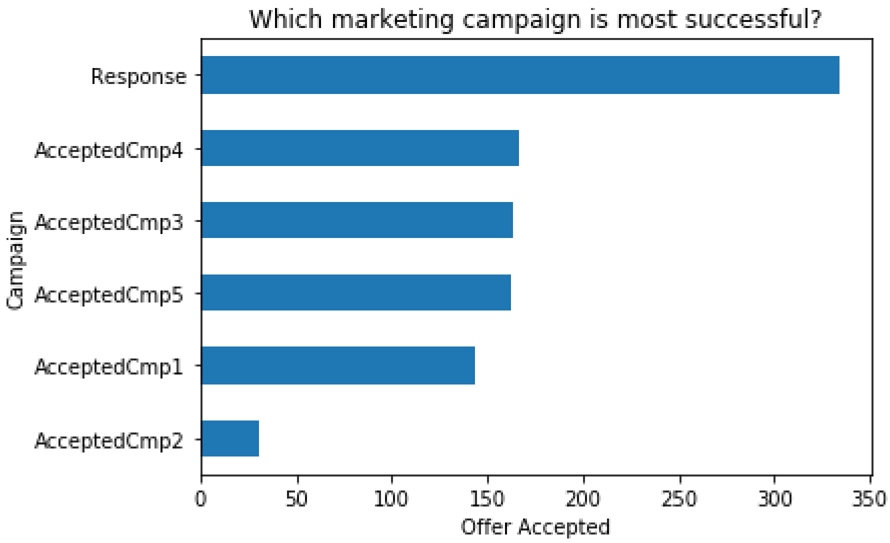
Here are the questions that I’d be exploring using data visualization:

- Which marketing campaign is most successful?
- What does the average customer look like for this company? Which products are performing best?

- Investigate the differences in the customer characteristics and purchases behaviors between the most successful campaign and the rest.

Now let’s look at the questions one by one.

1. Which marketing campaign is most successful?



Response means the last marketing campaign, which is the most successful one. It performed nearly twice as well as the previous campaigns, except campaign 2.

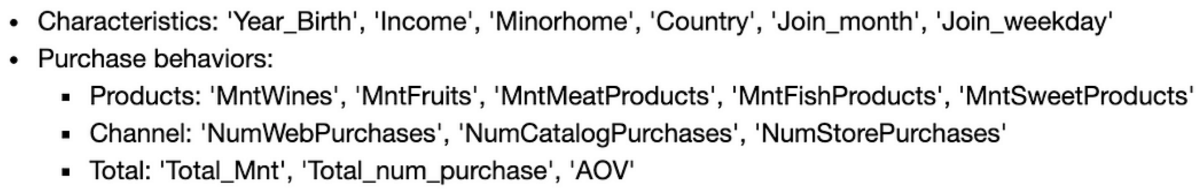
2. What does the average customer look like for this company? Which products are performing best?

After using .mean(), I found that an average customer…

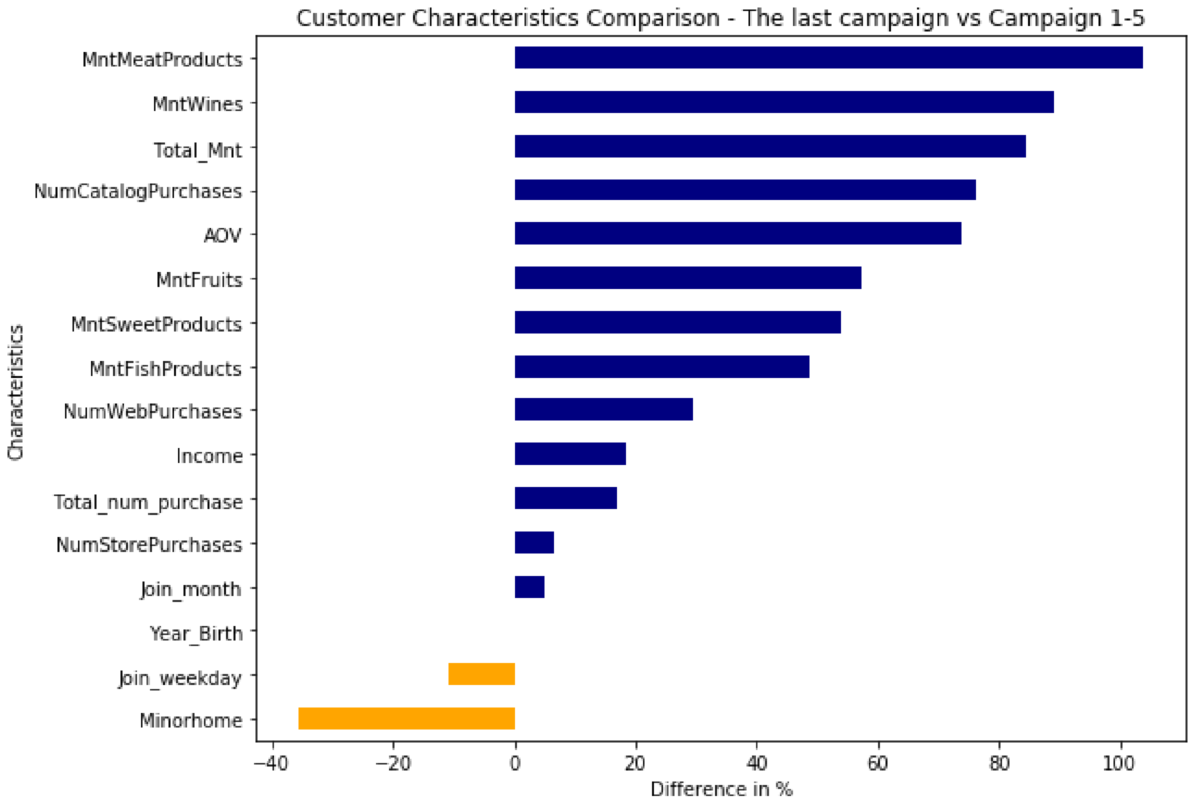
- has an annual income of 52200 dollars
- had purchased 49 days ago
- has an AOV of 26.8 dollars
- has spent 605 dollars
- has purchased 20 times
- became a customer in mid-June
- became a customer on Thursday
- spent most on wines(300 dollars) and then meat products(165 dollars)
- spent least on fruit(26 dollars) and sweet products(27 dollars)

3. Investigate the differences in the customer characteristics and purchases behaviors between the most successful campaign and the rest.

Now that we know the last campaign is the most successful one, we can further investigate the differences in the customer characteristics and purchases behaviors(listed below) between the most successful campaign, the last one, and the rest of the campaigns, campaign 1–5.



- Characteristics: 'Year_Birth', 'Income', 'Minorhome', 'Country', 'Join_month', 'Join_weekday'
- Purchase behaviors:
  - Products: 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts'
  - Channel: 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases'
  - Total: 'Total_Mnt', 'Total_num_purchase', 'AOV'



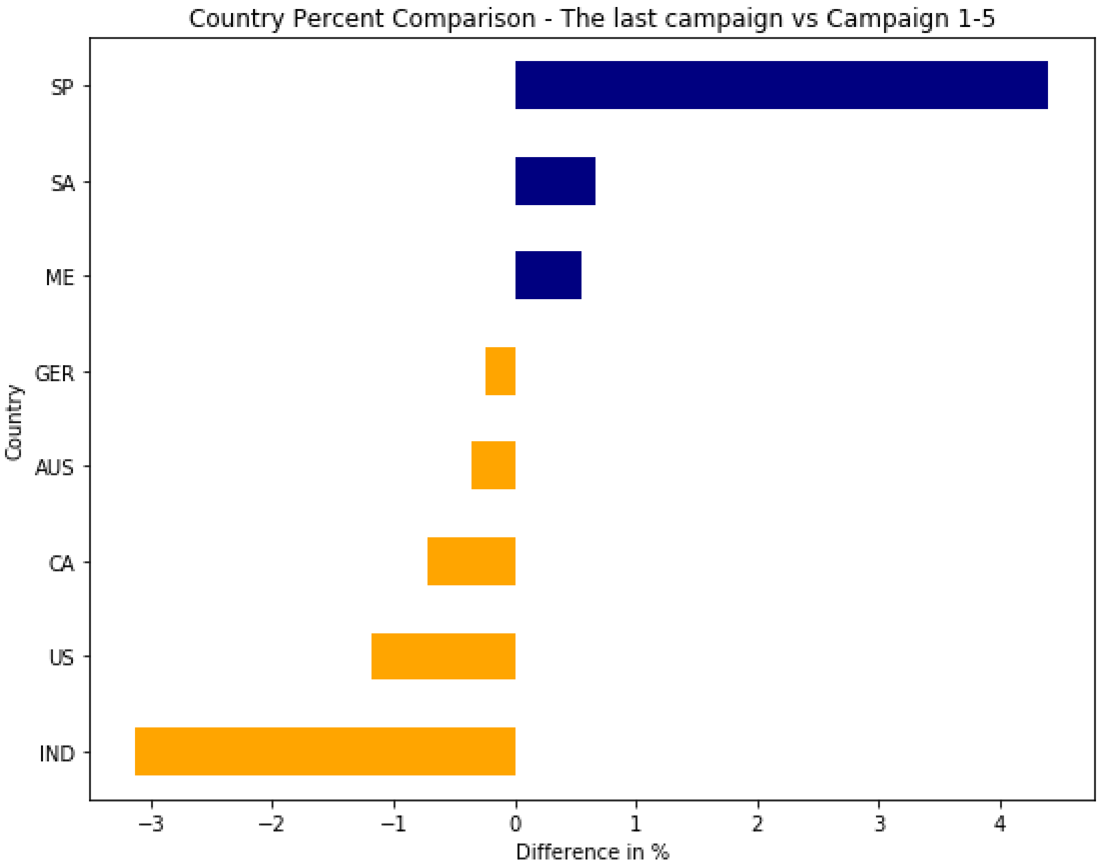
The last campaign attracted more valuable customers in terms of AOV, the total amount spent, and the total number of purchases compared to the customers attracted by the previous campaigns.

In terms of product categories, the customers in the last campaign spent nearly two times more money on meat products and wines compared to the customers in the previous campaigns.

Regarding purchasing channels, the customers in the last campaign purchased more evenly through stores, websites, and catalogs, whereas the customers in the previous campaigns mostly purchased through stores and websites.

The customers in the last campaign earned 20% more salary than the customers in the previous campaigns.

Let’s look at the proportion change of each country from the previous campaigns to the most successful campaign.



Spain has relatively more customers (+4%), and India has fewer customers (-3%) attracted to the last campaign.

Step 5: Forming Data-Driven Solutions

Summaries of insights:

1. The last campaign performed nearly twice as good as the previous campaigns

The last campaign attracted more valuable customers in terms of AOV, the total amount spent, and the total number of purchases, compared to the customers who were attracted by the previous campaigns.

Spain has relatively more customers (+4%) and India has fewer customers (-3%) that were attracted to the last campaign

In terms of product categories, the customers in the last campaign spent nearly two times more money on meat products and wines compared to the customers in the previous campaigns.

In terms of purchasing channels, the customers in the last campaign purchased more evenly through stores, websites, and catalogs, whereas the customers in the previous campaigns mostly purchased through stores and websites.

The customers in the last campaign earned 20% more salary than the customers in the previous campaigns.

2. Most customers purchase through physical stores, where people tend to spend more amount per purchase. The reason might be the customers had more impulsive purchases when they saw other similar products in stores.

3. People having kids at home are less valuable customers as they…

tend to purchase less

tend to has a high number of purchases made with a discount

4. The average customer…

became a customer on Thursdays

became a customer in Mid-June

Actionable Data-Driven Solutions

On Acquisition:

1. Keep using the same marketing techniques in the last campaign, but with a focus on promoting meat products and wines
2. Spend more marketing budget in Spain, and less in India
3. Have a brand discount day on Thursday or a brand discount month in June to attract new customers

On Increasing revenue:

1. Have marketing campaigns to convert customers who shop mostly on a website or catalog to in-store purchasers because most in-store purchases have a high average order volume.
2. Build a loyalty program to make high-income customers loyal as long as possible