Marketing Data Analytics

Author: Jarren Javier

Golden Gate University

August 18, 2021

Introduction

The goal for any good marketing campaign is to have a strategic effort by the company to raise awareness of a product being sold or to capture the attention of a customer. As a marketing analyst, it is one's job to analyze the data on customers and find out why a marketing campaign may or may not be successful. Through the use of data analytics, an analyst can better understand the problem and present a data-driven solution to the correct stakeholders.

The data set analyzed in this report is a fictional dataset taken from the Advanced Data Analytics course in the Master of Science Business Analytics program at Hult International Business School. The data set can be accessed on Kaggle's website: eCommerce Marketing Analytics. The entire analysis was run using Python and stored in a Jupyter Notebook. The methodology of analysis taken was to first use descriptive statistics to gain a better understanding of the company's customer base. Next a predictive linear regression model was produced to help predict the number of store purchases for the company. Lastly, more descriptive analytics was performed to find out what campaign was most successful and what the average customer looked like for that campaign.

Data Set Description

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on *gold* products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Figure 1 - Feature Description Table

Figure 1 - Feature Description Table

The dataset analyzed contains 2217 customers and 29 variables. All variables, with their accompanying description can be seen in Figure 1. Within the dataset, there are five different marketing campaigns that were offered to customers and either were rejected or accepted, denoted by 1's and 0's. The company offers five different products: fish, meat, gold, wine and sweets. A total amount spent within the last two years is provided for each customer. Data on the number of purchases for Store/Web/Deals/Catalog are provided as well. Two of the more important variables are Recency of last purchase and Income. Only birth year was provided and no data indicates present date of the data set. Categorical variables included are education level achieved, marital status, and country.

Data Cleaning

To prepare the data for analysis, several steps of data cleaning occurred before it was ready. There were 24 null values for income and it was decided for these to be dropped. Next totals for purchases, total amount spent, and total dependent columns were created. Aggregation of the number of catalog/web/store purchases was calculated to get the total purchases. Number of deals was excluded as it can be double counted. Total amount spent is an aggregation of the amount spent for each product and dependents is the sum of kids and teens at home. For marital status, answers including 'Yolo', 'Alone', and 'Absurd' were combined with the 'Single' category as they all indicate marital status as single. Education was the last variable that was changed. Under education, 'Graduation' and '2n Cycle' were converted to the 'Bachelor' category. 'Basic' was also converted to 'High School'.

Exploratory Data Analysis

Customer Segmentation

The first step in analyzing a marketing dataset is to perform an exploratory analysis and get a better understanding of who the customers are. The customer segmentation as seen in Figure 2 shows the different categorical variables (Marital Status, Education, and Country) in the data set. It was found that the marital status for customers within the dataset is mostly married and together in a relationship. Customers have lower counts of divorce, single, and widow. Customers' education levels are mostly bachelor and there is a surprising amount of customers with a PhD (481). Over half of the customers base is from Spain at 1093 and South/Central America is second with 337 customers. Mexico notably only has 3 customers, indicating company products are not popular or available to customers from Mexico.

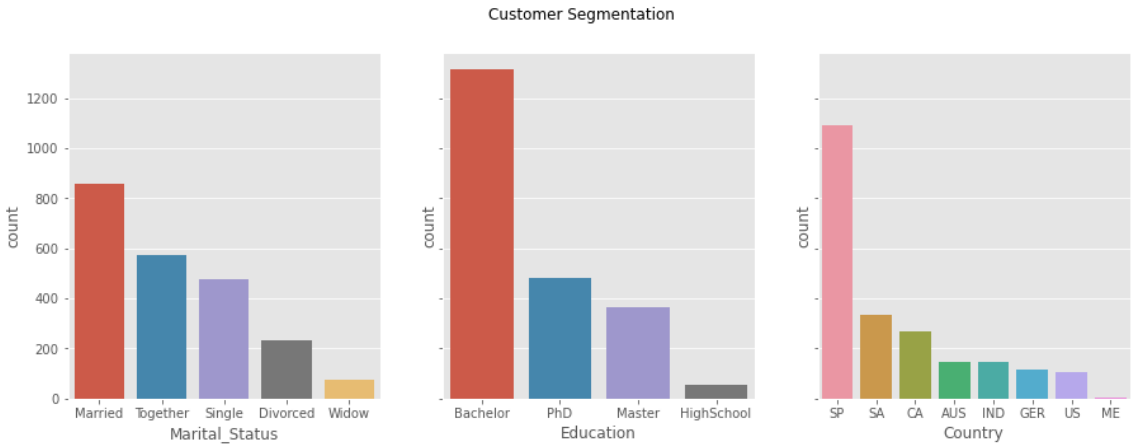


Figure 2 - Categorical Variable Segmentation

Birth year for customers is provided and Figure 3 shows that there is a highest cluster of customers born within 1970-1980. Birth year has extreme outliers that are born before 1900, making them over 110 years old. These can be removed from the dataset. There is an extreme outlier with an income of $666,666 that skews the distribution of income. Once removed the max income value is $162,397. This brought the average income down a couple hundred, $51,969

instead of $52,247. The Recency of purchase distribution has 50% of customers making their most recent purchase between 24 and 79 days and an average of 49 days. Ideally, the company would like to see the average time between purchases decrease and have more frequent customers.

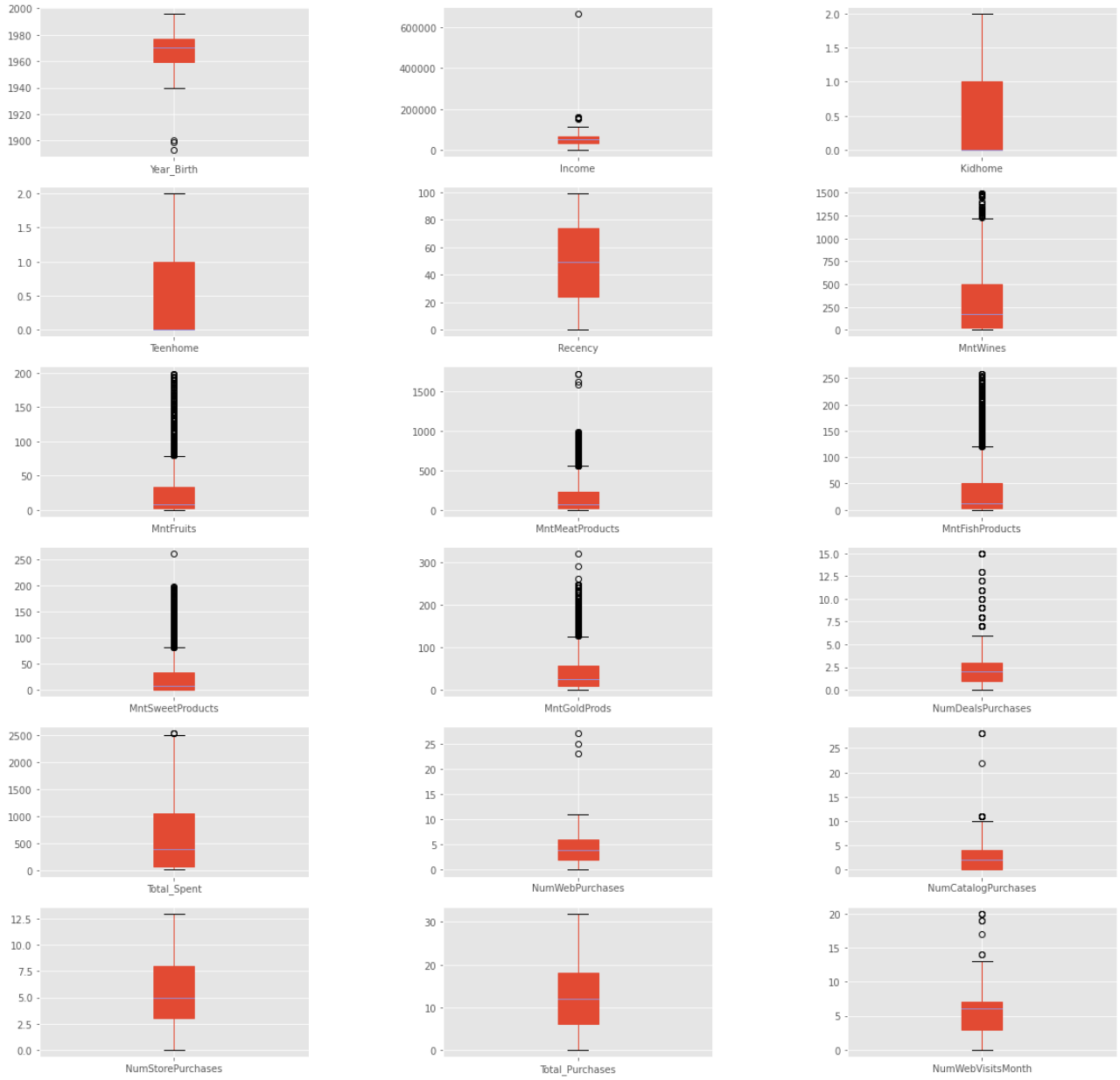


Figure 3 - Numerical Variable Distribution Boxplots

Figure 4 breaks down the sales by each individual product. Wine has the highest amount of sales totaling $676,083 and meat in second with $370,063. With this knowledge it would be

beneficial to further explore what are the most popular products by country. Further analysis in Figure 5 shows that for all countries, wine and meat are the top selling products. There is no distinct product popularity difference between the countries in the dataset.

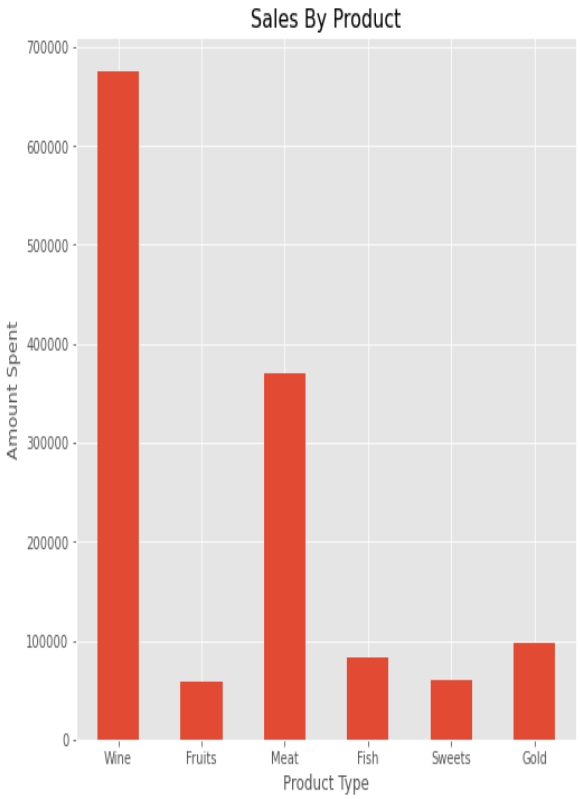
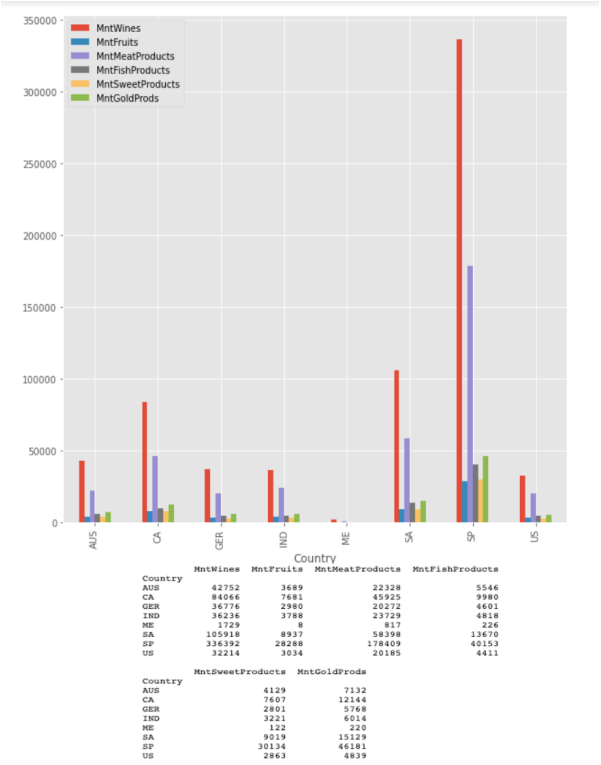


Figure 4 - Sales by Product



Country	MntWines	MntFruits	MntMeatProducts	MntFishProducts
AUS	42752	3689	22328	5546
CA	84066	7681	45925	9980
GER	36776	2980	20272	4601
IND	36236	3788	23729	4818
ME	1729	8	817	226
SA	105918	8937	58398	13670
SP	336392	28288	178409	40153
US	32214	3034	20185	4411

Country	MntSweetProducts	MntGoldProds
AUS	4129	7132
CA	7607	12144
GER	2801	5768
IND	3221	6014
ME	122	220
SA	9019	15129
SP	30134	46181
US	2863	4839

Figure 5 - Product Sales by Country

Linear Regression Analysis Number of Store Purchases

Correlation Analysis

The next goal was to use a regression model to try and predict the number of store purchases. By predicting the number of store purchases, the company can improve their inventory management. Improving inventory management can help reduce waste on perishable products offered by the company as well as increase the amount of cash flow available. Upon running a correlation analysis for the number of store purchases, the positive and negative

relationships with the independent variables was determined. Total Purchases (0.86), TotalSpent(0.68), MntWines(0.64), Income (0.53), and NumCatalogPurchases(0.52), NumWebPurchases(0.52) were the top six variables positively correlated with Number of Store Purchases. A positive correlation indicates as these variables increase, so does the number of store purchases. The highest negative correlation was found in the variables KidsHome(-0.50), Number of web visits per month (-0.43), and dependents(-0.32). The data indicates customers with more dependents are less likely to make in store purchases.

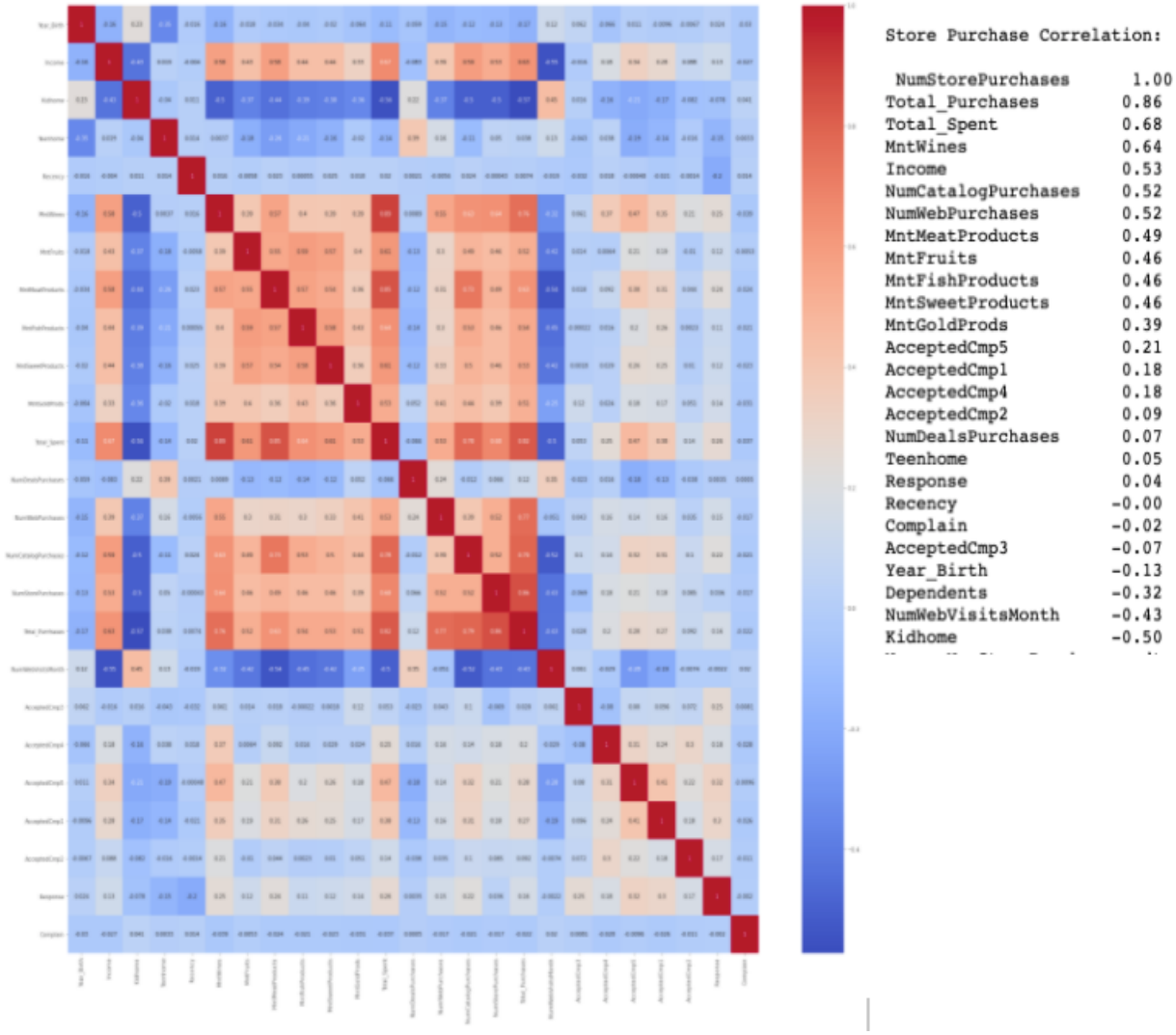


Figure 6 - Correlation Test and Heatmap

Model Data Preparation

The first step in preparing the data was to create dummy variables for the categorical variables (Country, Marital_Status, and Education). Creation of the dummy variable allows to run a single regression equation to represent the multiple groups as opposed to creating multiple equations. The target variable, number of store purchases, was stored in a Y variable for the test and the independent variables stored in the X variable. Once stored in the proper variables, the model was fitted by splitting the data into training and testing sets. Now that the model is properly fit, it can run for testing. The Ordinary Least Squares function in Python's Statsmodel is used for the regression analysis.

Model 1 Results

Using the Ordinary Least Squares model, we are able to predict the values 100% of the time. Model 1 had a R squared value of 1.00 showing a perfect fit and the adjusted R-squared value is the 1.00 as well. The p-values in figure 7A show that for Teenhome, Kidhome, Income, Recent, Dependents, MntFruits, Total_Purchases, NumWebPurchases, TotalSpent, MntGoldProds, MntSweetProducts, MntFishProducts, MntMeatProducts, MntWines, and NumCatalogPurchases are all less than 0.05, indicating statistical significance for predicting number of store purchases. No other model was created because of the perfect fit of model 1.

	coef	std err	t	P>|t|	[0.025	0.975]
Year_Birth	2.082e-17	7.76e-17	0.268	0.789	-1.31e-16	1.73e-16
Income	-4.28e-19	4.85e-20	-8.826	0.000	-5.23e-19	-3.33e-19
Kidhome	-1.51e-14	1.45e-15	-10.409	0.000	-1.79e-14	-1.23e-14
Teenhome	1.497e-14	1.36e-15	11.041	0.000	1.23e-14	1.76e-14
Dependents	-5.988e-15	1.01e-15	-5.927	0.000	-7.97e-15	-4.01e-15
Recency	-1.631e-16	2.84e-17	-5.736	0.000	-2.19e-16	-1.07e-16
MntWines	4.608e-15	6.08e-18	758.329	0.000	4.6e-15	4.62e-15
MntFruits	3.58e-15	2.7e-17	132.943	0.000	3.54e-15	3.64e-15
MntMeatProducts	4.65e-15	8.36e-18	556.413	0.000	4.63e-15	4.67e-15
MntFishProducts	4.856e-15	2.15e-17	226.038	0.000	4.81e-15	4.9e-15
MntSweetProducts	5.382e-15	2.58e-17	208.879	0.000	5.33e-15	5.43e-15
MntGoldProds	4.644e-15	1.81e-17	256.264	0.000	4.61e-15	4.68e-15
Total_Spent	-4.629e-15	5.12e-18	-904.024	0.000	-4.64e-15	-4.62e-15
NumDealsPurchases	8.431e-16	5.45e-16	1.546	0.122	-2.27e-16	1.91e-15
NumWebPurchases	-1.0000	6.19e-16	-1.62e+15	0.000	-1.000	-1.000
NumCatalogPurchases	-1.0000	5.97e-16	-1.68e+15	0.000	-1.000	-1.000
Total_Purchases	1.0000	3.86e-16	2.59e+15	0.000	1.000	1.000
NumWebVisitsMonth	-7.702e-16	5.14e-16	-1.499	0.134	-1.78e-15	2.38e-16
Marital_Status_Divorced	-6.882e-16	5.32e-14	-0.017	0.987	-1.06e-13	1.03e-13
Marital_Status_Married	-6.217e-15	5.34e-14	-0.116	0.907	-1.11e-13	9.84e-14
Marital_Status_Single	0	5.35e-14	0	1.000	-1.05e-13	1.05e-13
Marital_Status_Together	2.665e-15	5.33e-14	0.050	0.960	-1.02e-13	1.07e-13
Marital_Status_Widow	-2.665e-15	5.28e-14	-0.050	0.960	-1.06e-13	1.01e-13
Education_Bachelor	3.553e-15	6.65e-14	0.053	0.957	-1.27e-13	1.34e-13
Education_HighSchool	0	6.65e-14	0	1.000	-1.31e-13	1.31e-13
Education_Master	7.994e-15	6.63e-14	0.120	0.904	-1.22e-13	1.38e-13
Education_PhD	3.553e-15	6.63e-14	0.054	0.957	-1.27e-13	1.34e-13
Country_AUS	-1.332e-15	3.33e-14	-0.040	0.968	-6.68e-14	6.4e-14
Country_CA	-7.994e-15	3.33e-14	-0.240	0.810	-7.33e-14	5.73e-14
Country_GER	6.883e-16	3.34e-14	0.021	0.979	-6.46e-14	6.64e-14
Country_IND	-6.661e-16	3.35e-14	-0.020	0.984	-6.63e-14	6.5e-14
Country_ME	-6.883e-16	3.93e-14	-0.023	0.982	-7.79e-14	7.61e-14
Country_SA	4.441e-15	3.33e-14	0.133	0.894	-6.09e-14	6.98e-14
Country_SP	3.553e-15	3.33e-14	0.107	0.915	-6.17e-14	6.88e-14
Country_US	3.109e-15	3.34e-14	0.093	0.926	-6.23e-14	6.85e-14

Figure 7A - Model 1 Summary Statistics

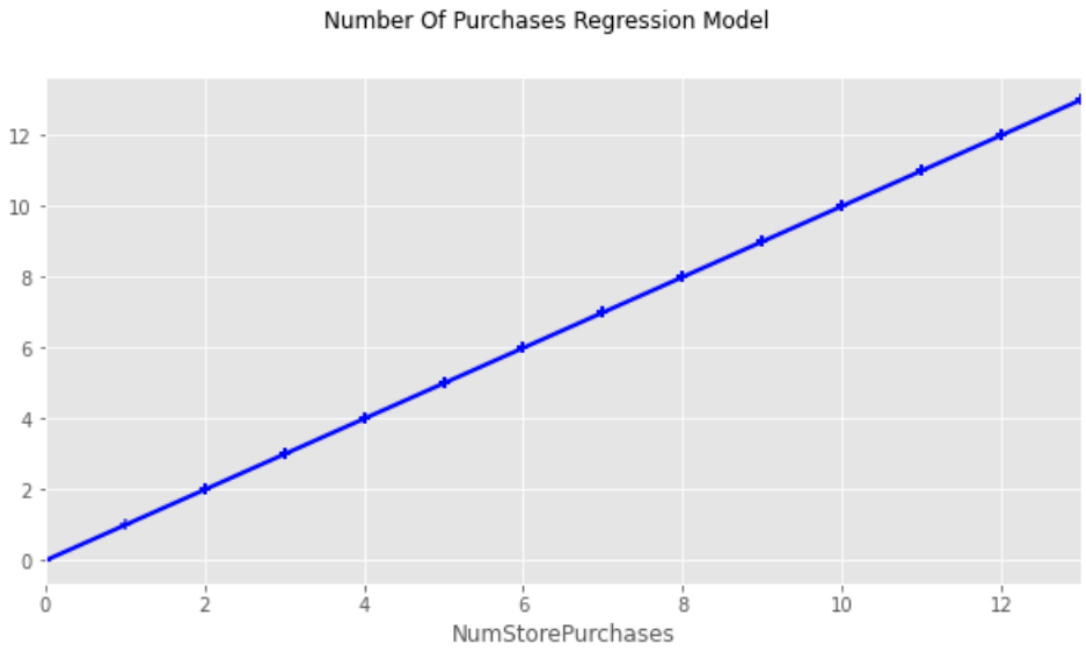


Figure 7B - Number of Purchases Model 1 Regression Plot

Campaign Analysis

For the final part of the analysis, I wanted to figure out what campaign performed best based on response rates for each campaign and what the average customer looked like for that campaign. There are several instances where multiple campaigns were accepted and a response was given by customers. For example campaign 1 and campaign 2 were accepted and received a response. Due to this fact, it was impossible to tell what campaign was responded to the most. In order to distinguish what campaign holds significance, only columns where one campaign was accepted was kept for analysis. With the campaigns properly adjusted, plotting the total count in Figure 8 for each campaign found that Campaign 3 was accepted the most by far totaling 129.

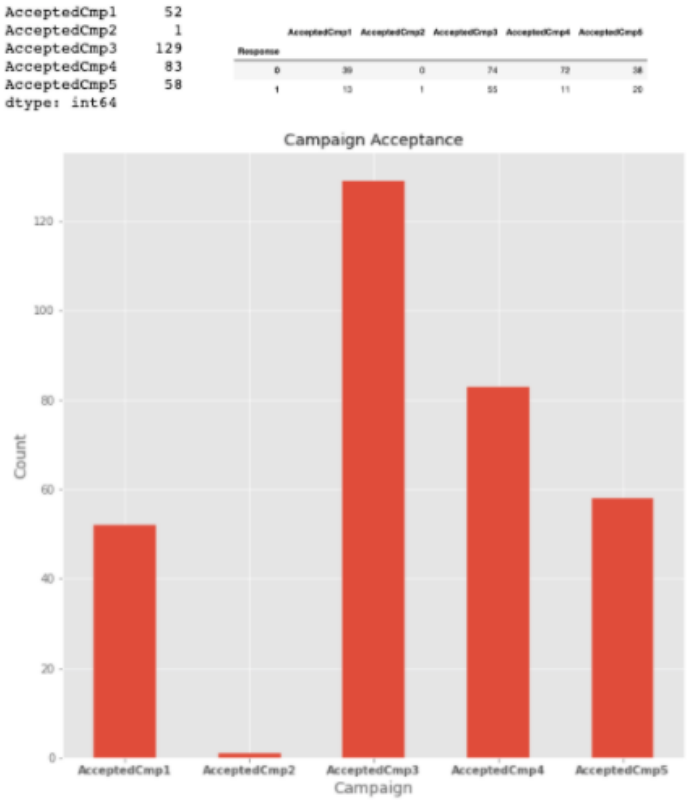


Figure 8 - Campaign Acceptance Bar Chart

55 out of the 129 accepted customers responded yes to Campaign 3.This brought the reponsonse rate to 42.64% as seen in Figure 9. Campaign 2 only had 1 response from a customer where it was the only campaign accepted. There is not enough data to analyze Campaign 2 thoroughly. Campaign 1 (52 acceptants) and Campaign 5 (58 acceptants) were significantly lower in acceptance. Campaign 4 had the lowest response rate at 13.25%. This indicates Campaign 4 was the least liked by customers and should not be used any further. Campaign 3 looks to be the most successful by far.

```
Campaign 1 Reponse Rate: 25.0
Campaign 2 Reponse Rate: 100.0
Campaign 3 Reponse Rate: 42.63565891472868
Campaign 4 Reponse Rate: 13.253012048192772
Campaign 5 Reponse Rate: 34.48275862068966
```

Figure 9 - Campaign Response Rate

Analysis of Campaign 3

The final goal was to answer the question, "What does the average customer who responded to Campaign 3 look like?". Figure 10 shows that the average customer has an income of $42,664, was born in the year 1969, recently purchased within 33 days, has 1 dependent, and has spent $532 in the last two years. Figure 11 segments the customer for Campaign 3 for marital status, education, and country. Customer demographic for Campaign 3 looks to be most popular with married/single, bachelor/PhD, and Spain/South America customers. With this information, the company can better target their campaigns to this particular demographic.

	Year_Birth	Income	Kidhome	Dependents	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
count	55.00	55.00	55.00	55.00	55.00	55.00	55.00	55.00	55.00	55.00
mean	1969.85	42664.02	0.56	0.91	0.35	33.09	282.49	23.42	129.98	26.07

Total_Spent	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	Total_Purchases	NumWebVisitsMonth
55.00	55.00	55.00	55.00	55.00	55.00	55.00
532.85	2.58	4.51	3.20	4.18	11.89	6.76

Figure 10 - Averages for Variables

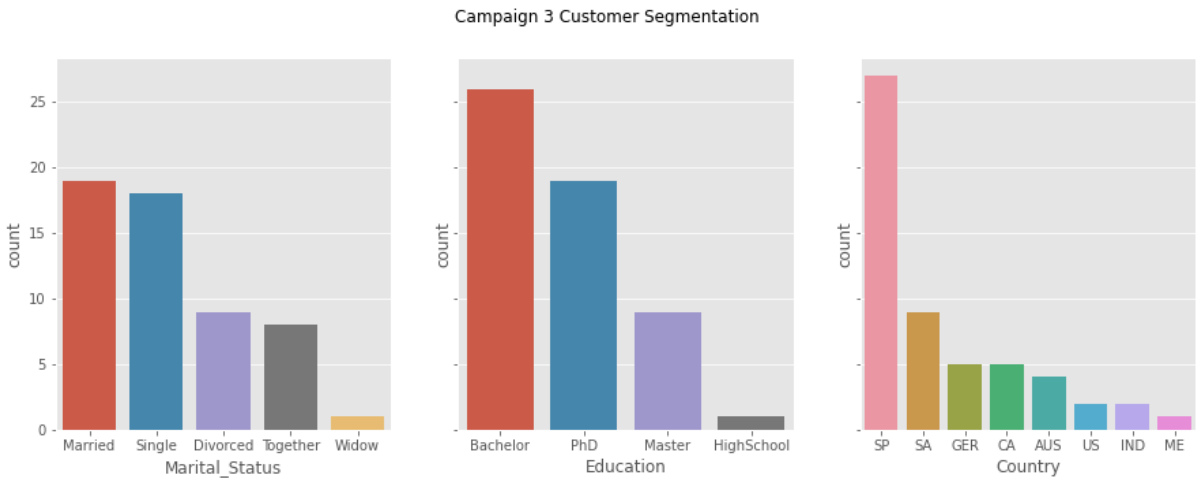


Figure 11 - Campaign 3 Customer Segmentation

Conclusion

The analysis performed gives a lot of data driven insight the company can use to better run the business. The exploratory analysis really highlighted that wine and meat are the best selling products and are performing the same across all countries. Next it found that the average recency of purchase was really high at 49 days and actions should be taken to get more frequent customers. The results show Spain holds most of the company's customer base and brings in most of the revenue. The accuracy of the linear regression model for the number of store purchases would be particularly helpful for the company in deciding how much stock to carry in stores. If the company can more accurately predict the demand of customers, revenue can be increased with better inventory management and the company will have a better cash flow. Lastly, analysis of Campaign 3 will help provide the company with a better understanding of their customer base for future campaigns. With better marketing Campaigns, the company can reach the right customers, push the more popular products, and continue to grow their market in other countries.

References

Crockett, J. (2021, February 2). Marketing analytics eda task [Final]. Kaggle. https://www.kaggle.com/jennifercrockett/marketing-analytics-eda-task-final/notebook.

Daoud, J. (2020, December 19). Marketing analytics. Kaggle. https://www.kaggle.com/jackdaoud/marketing-data.