Team 17 Final Report

Group Members: Ziyang Guo, Sean Lee, Vakula Mallapally, Bill Szeto, Tianhua Zhu

Introduction

Increasing sales and continued growth has always been the primary objective for any business. Specifically, the grocery retail industry has seen an influx in the availability and popularity of multi-channel buying experiences and ever-increasing competitors flooding the market. These have been a major factor in the ability of grocery retailers to maintain a consistent level of business. Online shopping, curb-side service, grocery delivery service are just a few of the new modalities. Online retailers such as Amazon and traditional brick and mortar retailers like Walmart and Target joining the grocery business have only added to the business challenge. This project aims to analyze a set of publicly available marketing data in hopes of uncovering customer spending and purchasing habits and any correlation they may have with successive promotional campaigns and availability of various shopping methods.

Objectives

- Understand the relationship amongst the variables, especially how they influence the response variables (i.e., purchase decisions and campaign performance)
- Predict who would be most responsive and valuable to this company, which will in turn maximize the company's profit in future marketing campaigns

Research Questions

- **Consumer value:** Which demographic group is most valuable for the company to target?
- **Promotion effectiveness:** Given the most valuable consumers' behavior, what marketing channel and message should the company focus on?

Description of Dataset

Screenshot of Dataset and Link to Source

- https://www.kaggle.com/jackdaoud/marketing-data/version/1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
ID	Year_Birth	Education	Marital_Stat	Income	Kidhome	Teenhome	Dt_Custome	Recency	MntWines	MntFruits	MntMeatPr	MntFishPro	MntSweetPr	MntGoldPr	NumDealsP	NumWebPu	NumCatalog	NumStoreP	NumWebVi	AcceptedCm	AcceptedCm	AcceptedCm	AcceptedCm	AcceptedCm	Response	Complain	Country
1826	1970	Graduation	Divorced	$84,835.00	0	0	6/16/14	0	189	104	379	111	189	218	1	4	4	6	1	0	0	0	0	0	1	0	SP
1	1961	Graduation	Single	$57,091.00	0	0	6/15/14	0	464	5	64	7	0	37	1	7	3	7	5	0	0	0	0	1	1	0	CA
10476	1958	Graduation	Married	$67,267.00	0	1	5/13/14	0	134	11	59	15	2	30	1	3	2	5	2	0	0	0	0	0	0	0	US
1386	1967	Graduation	Together	$32,474.00	1	1	5/11/14	0	10	0	1	0	0	0	1	1	0	2	7	0	0	0	0	0	0	0	AUS
5371	1989	Graduation	Single	$21,474.00	1	0	4/8/14	0	6	16	24	11	0	34	2	3	1	2	7	1	0	0	0	0	1	0	SP
7348	1958	PhD	Single	$71,691.00	0	0	3/17/14	0	336	130	411	240	32	43	1	4	7	5	2	0	0	0	0	0	1	0	SP
4073	1954	2n Cycle	Married	$63,564.00	0	0	1/29/14	0	769	80	252	15	34	65	1	10	10	7	6	1	0	0	0	0	1	0	GER
1991	1967	Graduation	Together	$44,931.00	0	1	1/18/14	0	78	0	11	0	0	7	1	2	1	3	5	0	0	0	0	0	0	0	SP
4047	1954	PhD	Married	$65,324.00	0	1	1/11/14	0	384	0	102	21	32	5	3	6	2	9	4	0	0	0	0	0	0	0	US
9477	1954	PhD	Married	$65,324.00	0	1	1/11/14	0	384	0	102	21	32	5	3	6	2	9	4	0	0	0	0	0	0	0	IND
2079	1947	2n Cycle	Married	$81,044.00	0	0	12/27/13	0	450	26	535	73	98	26	1	5	6	10	1	0	0	0	0	0	0	0	US
5642	1979	Master	Together	$62,499.00	1	0	12/9/13	0	140	4	61	0	13	4	2	3	1	6	4	0	0	0	0	0	0	0	SP
10530	1959	PhD	Widow	$67,786.00	0	0	12/7/13	0	431	82	441	80	20	102	1	3	6	6	1	0	0	0	0	0	1	0	IND
2964	1981	Graduation	Married	$26,872.00	0	0	10/16/13	0	3	10	8	3	16	32	1	1	1	2	6	0	0	0	0	0	0	0	CA
10311	1969	Graduation	Married	$4,428.00	0	1	10/5/13	0	16	4	12	2	4	321	0	25	0	0	1	0	0	0	0	0	0	0	SP
837	1977	Graduation	Married	$54,809.00	1	1	9/11/13	0	63	6	57	13	13	22	4	2	1	5	4	0	0	0	0	0	0	0	SP
10521	1977	Graduation	Married	$54,809.00	1	1	9/11/13	0	63	6	57	13	13	22	4	2	1	5	4	0	0	0	0	0	1	0	SP
10175	1958	PhD	Divorced	$32,173.00	0	1	8/1/13	0	18	0	2	0	0	2	1	1	0	3	4	0	0	0	0	0	0	0	SP

- In our current dataset, the only definitive variables are the customer age and location. Our plan is to find a related dataset with information on countries and the average household income within that country. We can then compare the spending habits of the customers from dataset 1 compared to the country's mean income.

The dataset we focused on is from Kaggle. With 2240 rows and 28 columns, it contains consumer demographic, product purchase history from the last 2 years, as well as historical response to marketing campaigns. Leaving out the uninformative variables, such as consumer ID, we describe the variables that would potentially contribute to our analysis as follows:

Consumer Profile Information	
Country	The country consumer is located in
DtCustomer	date of customer's enrollment with the company (when they became a customer)
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
Year_Birth	Customer's birth year (from which we can calculate age)
Product Purchase History	
MntFishProducts, MntMeatProducts, MntFruits, MntSweetProducts, MntWines, MntGoldProds	amount spent on this category of products (Fish, Meat, Fruit, Sweet, Wine, Gold) in the last 2 years
Recency	number of days since the last purchase
Past Response to Marketing	
AcceptedCmp1,2,3,4,5	Binary variable for whether or not customer accepted the offer in the given campaign
NumDealsPurchases, NumCatalogPurchases, NumStorePurchases, NumWebPurchases	number of purchases made with the given channels (Catalog, Store, Web, with a Deal)
Response	Binary variable for whether or not customer accepted the offer in the last campaign
Other Behavioral Information	
NumWebVisitsMonth	number of visits to company's web site in the last month
Complain	Binary variable for whether or not customer complained in the last 2 years; 1 if complained, 0 otherwise

Data Cleaning

To prepare our dataset for analysis, the following data cleaning steps were performed:

- ID: Verified values were all unique
- Education: "2n Cycle" was changed to "Masters" as the data is based off of the European Higher Education Area (EHEA); 230 data points affected
- Education: "Master" was changed to "Masters" for consistency in responses; 370 data points affected
- Education: "Graduated" is assumed to represent the completion of a Bachelor's degree; 11 data points affected
- Marital Status: "Absurd" and "YOLO" were removed; 4 data points affected
- Marital Status: "Alone" was changed to "Single" as they are interpreted to be the same; 3 data points affected
- Income: Blank values were interpreted as a lack of input stream and therefore changed to $0.00; 24 data points affected

- Income: There was a single income value that was much larger than all the other values, $666,666.00. We believe this to be an outlier, but have chosen to leave the data point in for preliminary analysis and remove it if it is determined to skew results.
- Birth Year: There are three values – 1893, 1899, 1900 – that may also be outliers as they would be much too old for a 2013/14 registration. They may possibly skew the results, but at this point, we allowed these data points to also remain in the dataset.

Feature Engineering

We also created the following variables to facilitate the upcoming analysis:

- Age: Since we don't know the date the data was taken, we used the difference between each customer's year of registration and their year of birth to represent age.
- NoChild: Binary variable for whether or not the customer has dependants
- total_Purchases: Aggregate amount spent across all product categories
- TotalAcceptedCmp: Total number of campaigns the customer accepted
- AnyAcceptedCmp: Binary variable for whether or not the customer accepted any campaign

To further expand our dataset, we have looked into incorporating a second dataset involving various criteria such as daily incomes across various countries, and average household spending on groceries across countries. The potential attributes to combine datasets would be income and country. However, as our Marketing Analytics dataset uses non-standardized abbreviations to represent countries, it would be difficult to combine data due to ambiguity. For example, we would not know a country value of "SA" would indicate "Saudi Arabia" or "South Africa". It would not affect our progress since we plan on using them only for information enrichment in the end, rather than the core modeling process.

Approach and Methodology

Given the two research questions listed above, we used both descriptive and predictive analyses to provide insights into each of them.

For us to identify the most profitable demographic group, we first built a correlation matrix to find relationships between the numeric variables.The correlation matrix revealed which variable has the most influence on purchases and which category of items is most popular amongst that demographic. A positive value would indicate that the demographic variable has a proportional relationship to the customer purchases output, while a negative correlation means that the two are inversely proportional. For categorical variables, we used crosstabs and bar charts to explore their interactions with the response variables we are interested in.

Then, we used cluster analysis to identify the consumer group that has the highest potential to spend more. Regression models were employed to show which demographic factors have the highest influence on the amount spent. By interpreting the coefficient of each predictor, we were able to discover what types of customers the company should focus on acquiring, and what customer behavior they should strive to encourage.

Once we have identified the most influential demographic variables and most bought store items, we turned to historical campaign data. A customer can purchase products from either a catalog, website, or in store. Using classification models to estimate likelihood of purchase and conversion, the company can then prioritize the more profitable channels and allocate more resources there.

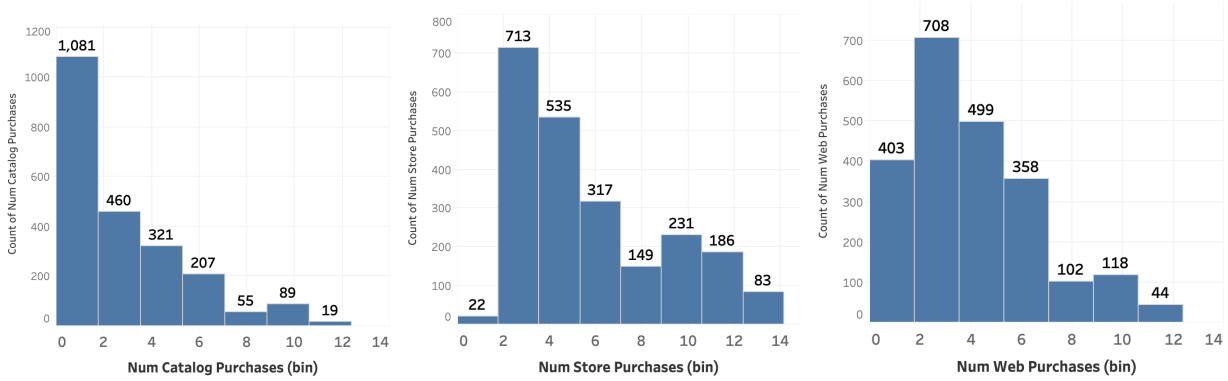
Descriptive Analysis Results

After data cleaning, we were left with 2236 data points. Some summary statistics revealed an overall picture of the company's customer base:

- **Demographics:** A typical customer is middle-aged (36-54 years old), with a Bachelor's degree or higher, earning an average of $51K yearly; the majority of them are not married and don't have children.
- **Recency:** They do not tend to buy from the company frequently; on average, their last purchase was made 51 days before.

- **Purchases:** Overall, store seems to be the most successful channel, inducing the least amount of non-purchases, while catalog performed the worst.

Figure 1. Distribution of the Number of Purchases via Catalog, Store, and Web

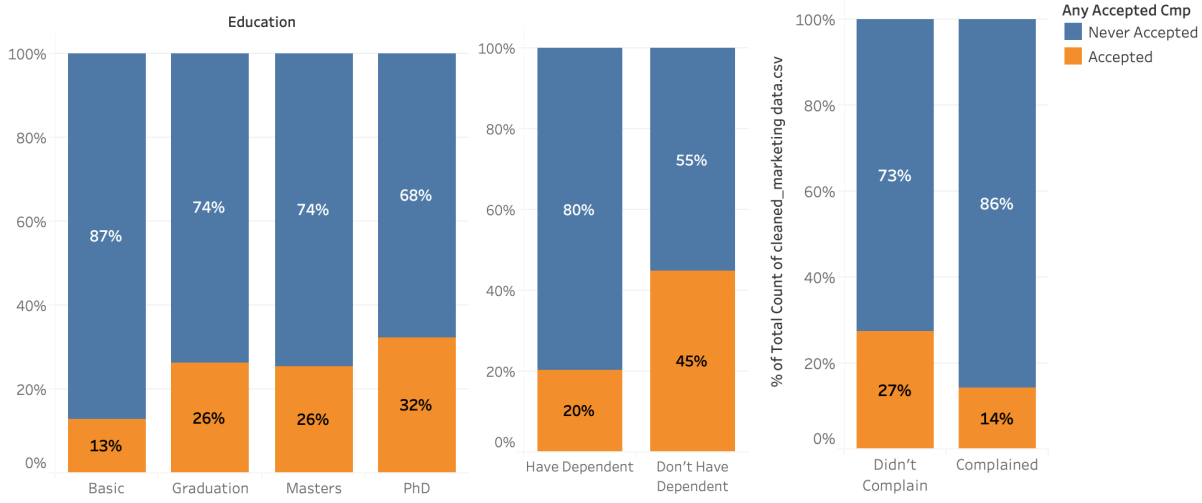


- **Campaign Response:** Of all customers, 27.15% have accepted at least one campaign; the latest campaign with an acceptance rate of 14.85% is the most successful one; among the ones who accepted at least one campaign, it takes an average of 3.59 campaigns to convince them to act.

Then, using bivariate analysis to explore the relationship between factors and response (whether the customer accepted any campaign before), we found that:

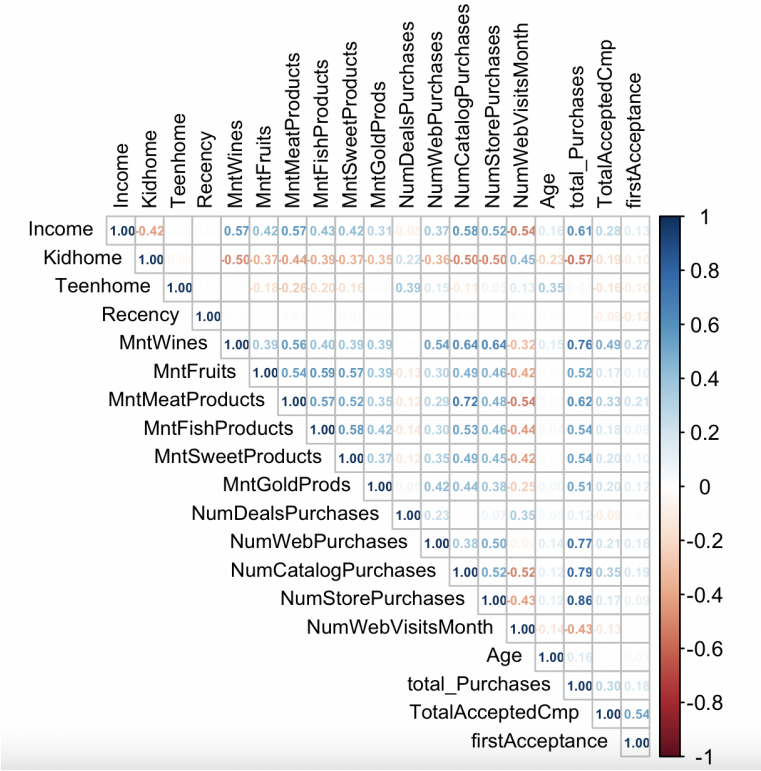
- There's little relationship between marital status and campaign response.
- Customers who are more highly educated don't have kids/ Understandably, the same customers did not complain and are more responsive to the company's marketing campaigns overall.

Figure 2. Campaign Response by Education, Dependent, and Camplain History



It is also interesting to look at the interaction amongst all numeric variables. So we built a correlation matrix as shown in Figure 3.

Figure 3. Correlation Matrix



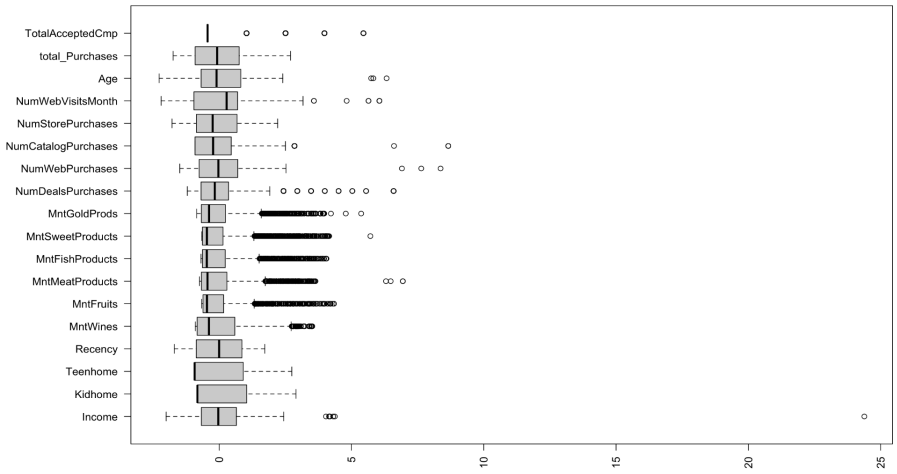
Since total_Purchases was created as the sum of catalog, store, and web purchases, the correlation among them are ignored in this analysis. Among the rest, the most highly correlated pairs are as follows:

total_Purchases	MntWines	0.76
NumCatalogPurchases	MntMeatProducts	0.72
NumStorePurchases	MntWines	0.64
NumCatalogPurchases	MntWines	0.64
total_Purchases	Income	0.61
total_Purchases	Kidhome	-0.57

In summary, wine consumption is positively correlated with both purchase amount and number of accepted campaigns; consumers tend to purchase meat through catalog and wine using both store and catalog; more income is associated with more purchases; Kidhome and NumWebVisitsMonth seem to be suppressing factors on all other variables, except for NumDealsPurchases. We will further examine these findings in the modeling stage, and interpret them in more depth in the final report.

When examining the distribution of each of the variables, we also noted that there are potential outliers in the data set, as shown in the boxplot overview in Figure 4. We chose to keep them in this stage since there is no indication that they are necessarily external errors rather than natural occurrences of extreme values, but we will pay special attention to their influence on the overall findings during modeling.

Figure 4. Boxplots of Numeric Variables After Standardization



Boxplots - Total Purchases by Education/Country/Marital Status

After the data is standardized like naming convention, removing outliers, and categorization, we made the boxplot for Total Purchase. We take three dimensions into consideration which are Education, Country and Marital Status. Following boxplots figures are generated. Note that the red point in each figure is the mean of the Total purchase. It can be seen that the mean is close or similar to the median for every box plot, which means the data is pretty much symmetric after standardization..

Figure 5. Boxplots of Education vs Total Purchases After Standardization

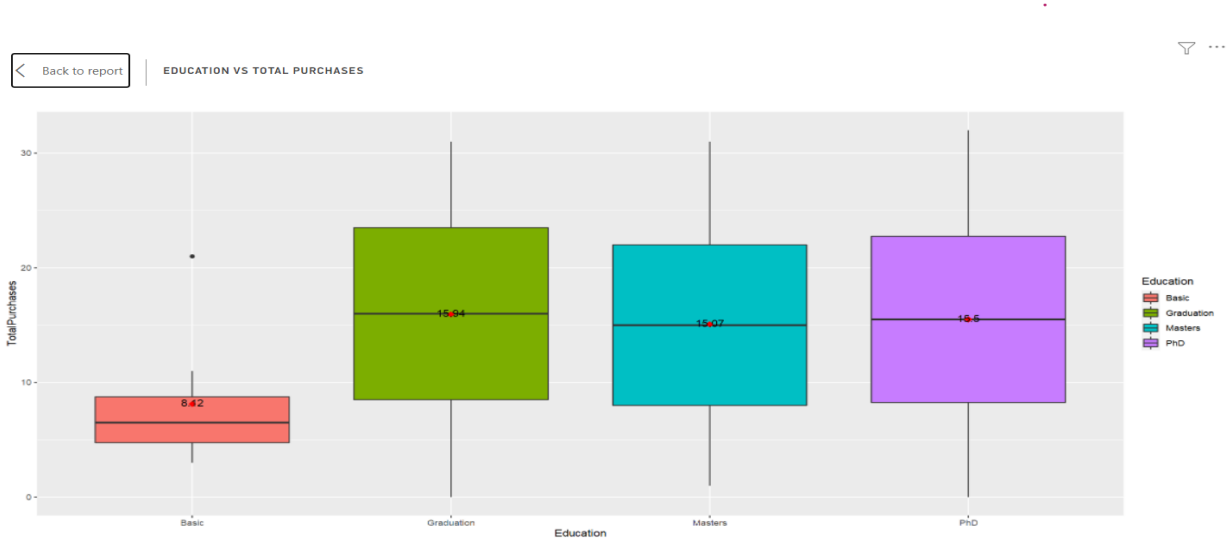


Figure 6. Boxplots of Marital Status vs Total Purchases After Standardization

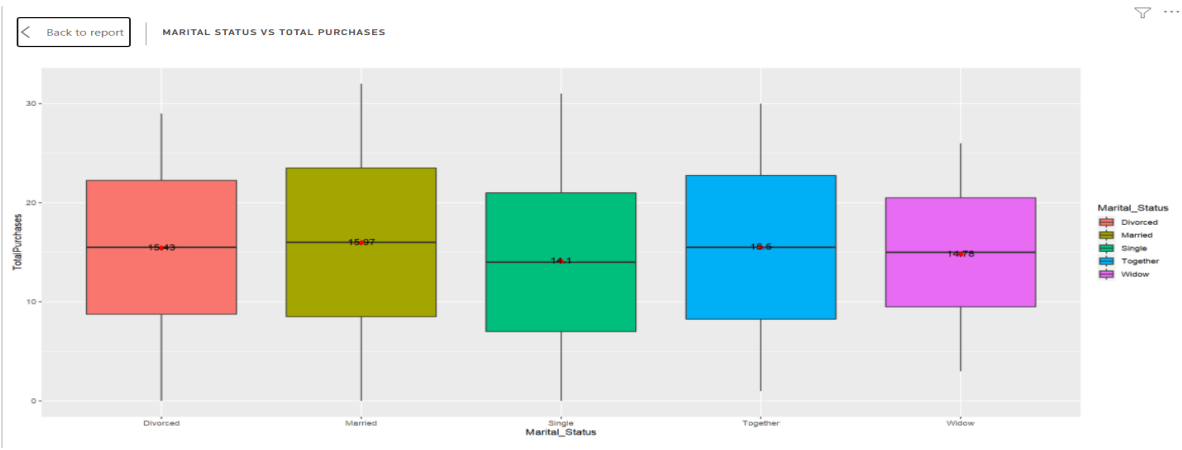
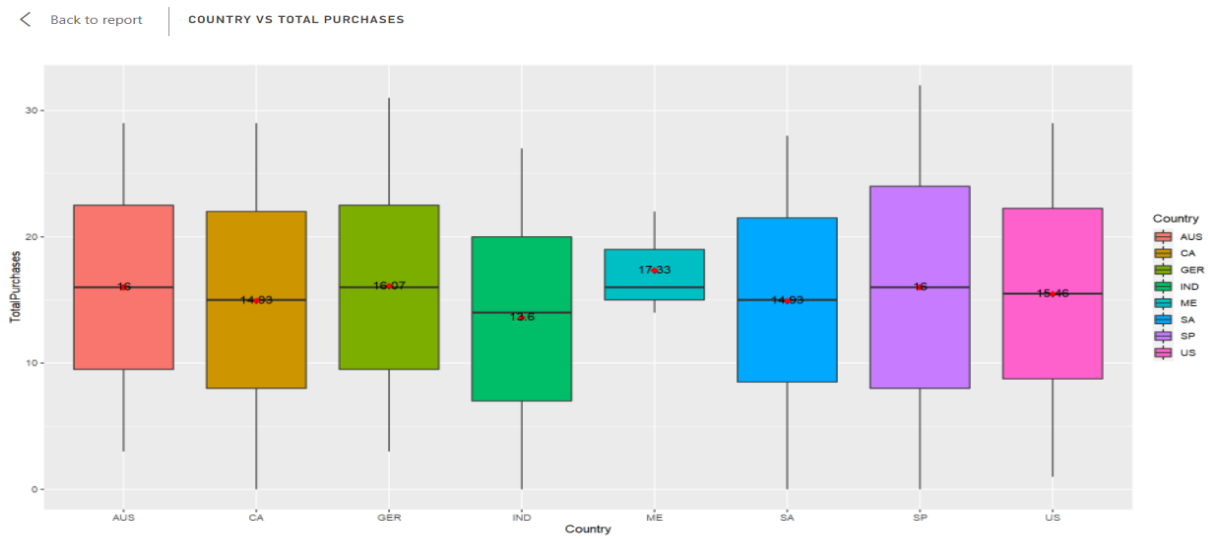


Figure 7. Boxplots of Country vs Total Purchases After Standardization



Results and Interpretation of Results

Regression Analysis

To get a better understanding of how the independent variables (customer demographics, types of store purchases, and campaign acceptance) relate to the dependent variable (total purchases), we ran multiple regression models. This information is important, since identifying the variable weights could indicate which factors are most important in customer spending habits, allowing the business to create targeted marketing campaigns.

Before running the model, we first looked at the scatterplot between each independent variable and the dependent variable. We then decided to transform the six variables related to food consumption and gold purchases since they exhibit a relationship that is clearly non-linear. After natural log transformation, the R2 of each of the trend lines was improved by at least 10%. And the adjusted R2 of the model increased from 0.7724 to 0.8029.To further improve upon the current 28-variable model, we performed variable selection using stepwise regression with both

directions. With the goal to minimize AIC, the procedure left out Marital_Status, Teenhome, and AcceptedCmp2, which were not statistically significant in the original model either.

Figure 8. Before Transformation

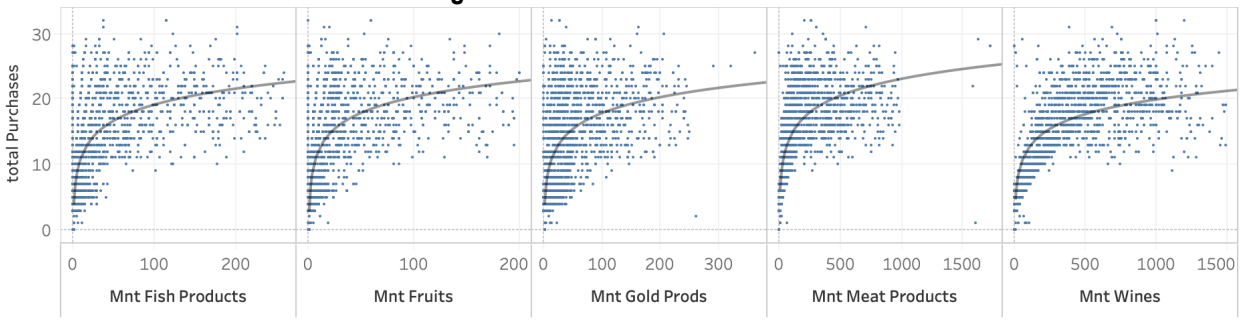
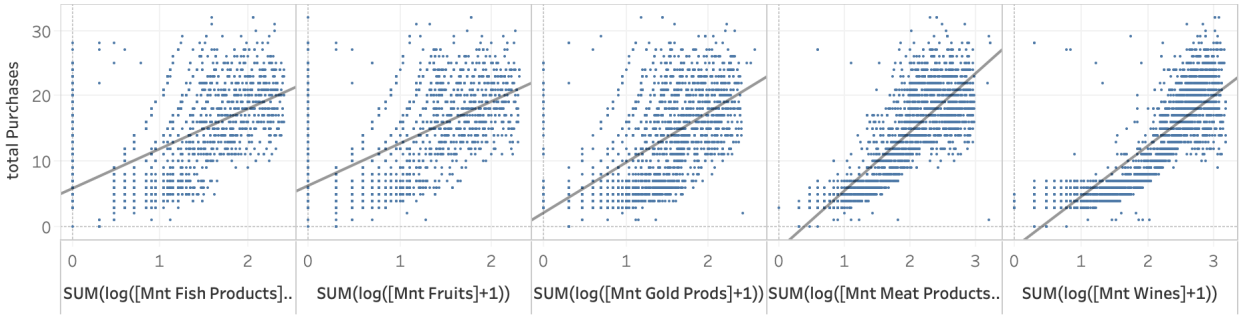


Figure 9. After Transformation



Regression Model - Total Purchases

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.890e+00  5.091e-01  -3.712 0.000211 ***
EducationGraduation       -2.163e+00  4.814e-01  -4.493 7.38e-06 ***
EducationMasters          -2.113e+00  4.919e-01  -4.296 1.82e-05 ***
EducationPhD              -1.678e+00  5.196e-01  -3.230 0.001256 **
Income                     3.296e-05  5.393e-06   6.111 1.16e-09 ***
Kidhome                   -1.290e+00  1.681e-01  -7.678 2.40e-14 ***
Recency                   -3.783e-03  2.407e-03  -1.572 0.116116
NumDealsPurchases          3.402e-01  4.115e-02   8.267 2.33e-16 ***
AcceptedCmp31              6.395e-01  2.786e-01   2.295 0.021803 *
AcceptedCmp41              5.920e-01  2.930e-01   2.020 0.043503 *
AcceptedCmp51             -5.759e-01  3.183e-01  -1.810 0.070490 .
AcceptedCmp11              6.852e-01  3.181e-01   2.154 0.031343 *
Response1                 -5.616e-01  2.224e-01  -2.526 0.011622 *
Complain1                  1.036e+00  7.038e-01   1.473 0.140998
log(MntWines + 1)          1.558e+00  8.805e-02  17.698  < 2e-16 ***
log(MntFruits + 1)         1.856e-01  7.086e-02   2.620 0.008859 **
log(MntMeatProducts + 1)   1.125e+00  1.117e-01  10.068  < 2e-16 ***
log(MntFishProducts + 1)   1.777e-01  6.756e-02   2.631 0.008573 **
log(MntSweetProducts + 1)  4.230e-01  6.812e-02   6.210 6.30e-10 ***
log(MntGoldProds + 1)      3.010e-01  7.500e-02   4.014 6.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.2 on 2215 degrees of freedom
Multiple R-squared:  0.8047,    Adjusted R-squared:  0.803
F-statistic: 480.3 on 19 and 2215 DF,  p-value: < 2.2e-16
```

Based on the output above, Education, Income, KidHome, NumDealsPurchases, MntWine, MntMeatProducts, MntSweetProducts, and MntGoldProds were the most statistically significant factors contributing to total purchases. From this data, we can draw a few reasonable conclusions. Running this leaner model gives us an adjusted R2 of 0.803. The model output is shown below. We can observe that higher education and having more kids are associated with lower purchase; an increase of $10K in income would result in 0.3296 increase in purchase on average; accepting 10 more deals is expected to lead to 3.402 more purchases; buying more food and gold have positive effect on purchase as well. Buying wine has the strongest positive effect – 1% increase in wine consumption could result in 1.558 more purchases overall.

For our second regression model, , we turned to each of the three channel purchases. It turned out that these independent variables didn't do as good a job at predicting channel-specific purchases as they did for total purchases, as indicated in a much lower R2. However, since the factor significance and coefficient magnitude may still provide some insights for each channel's marketing strategy, we summarized output in a table in Appendix A.

It is interesting to note that the three channels seem to have a competitive relationship against each other. For example, more store or web purchases lead to lower catalog purchases. Also web purchases are negatively associated with campaign #2. Catalog purchases are positively associated with campaign #1 and #3. Store purchases are positively associated with campaign #2 and #4. Without the actual campaign creatives at hand, we cannot provide more in-depth analysis, but it may be beneficial for the company to follow this path and analyze if there's any significant implications of this association between purchase channels and the promotion strategy.

All of the three channels were negatively associated with homes that have kids. Catalog purchases were negatively associated with teens at home. The number of web visits per month is positively correlated with the

number of web purchases but negatively correlated with the number of catalog and store purchases. Customers who make web purchases are more likely to continue to make more web visits.

The third regression analysis produced interesting results when channel purchases were regressed using the various campaigns. The findings seemed to indicate that the campaigns may have been channel-focused; i.e.certain campaigns may have been drivers for promoting in-store or catalog, combinations of channel purchases. Campaigns 1 and 2 affected Web purchases and store purchases negatively and positively, respectively. The same effects appear for campaign 1 and 3 for Store and Web purchases. With Group #3 representing one half of the company's customer base, a recommendation for the campaigns would be to include all Company purchasing modalities; do not limit campaigns to "online purchases only" or "in-store purchases only". Unfortunately, the actual details of the Company's campaigns were not available with the dataset.

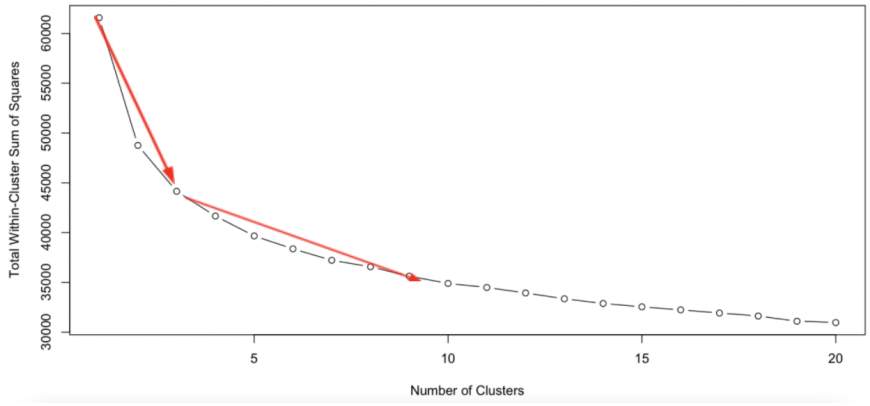
After identifying the significant variables in regression models for total purchases, types of purchases, and campaign acceptance, the business can focus on targeted marketing campaigns for specific types of customers, therefore capitalizing on customers with high profit margins and acquiring customers with the highest profit potential.

Discussion

Cluster Analysis

Since we have a mixture of numeric and categorical variables, we used K-Prototypes algorithm which uses cluster means for numeric variables and modes for factors to segment the customers. From the elbow chart plotting the total within-cluster sum of squares against the number of clusters, we decided to use the three-cluster solution.

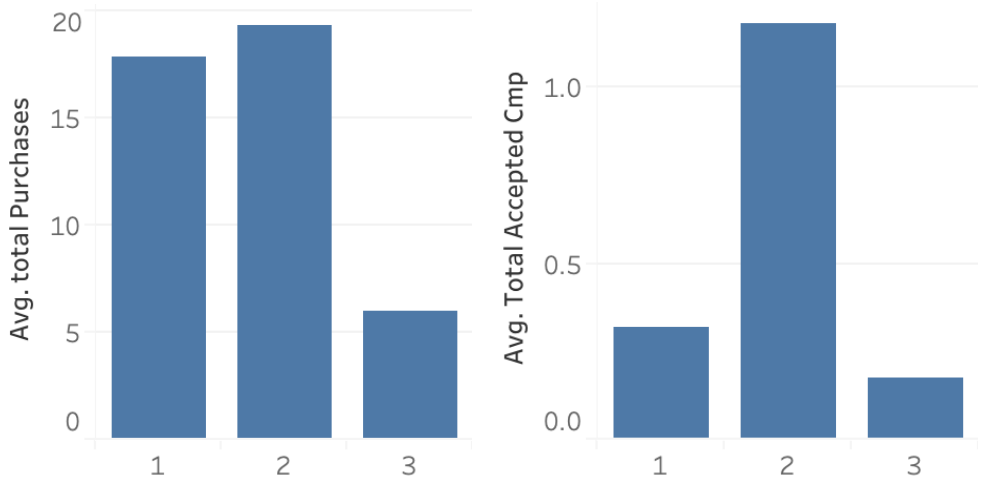
Figure 10. #Clusters vs Sum of Squared Errors



With 20 random initiations, the algorithm converged, and we found 662 data points belonging to cluster #1, 507 belonging to #2, and 1067 to #3. From the summary data of each cluster, with their total purchases and total accepted campaigns shown in the plots below, we can easily see the differences among clusters:

- **Low value:** a typical customer in cluster #3 is married, in their early 40s, has more than one child, earns the least income among the three, sometimes complains, rarely makes purchases with the company, with deals occupying a large proportion of their overall purchases.
- **Moderate value:** a typical customer in cluster #1 is married, in their late 40s, has 1 child, relatively well-off, sometimes complains, likes to purchase wine and sometimes meat from the company, prefers store purchases most and web second, and likes to take deals.
- **High value:** a typical customer in cluster #2 in their mid-40s, has no child, wealthiest among the three, rarely complains, likes to purchase both wine and meat, prefers store purchases most and catalog second, and takes few deals.

Figure 11. Customer Clusters vs Total Purchases and Total Accepted Campaigns



Based on these cluster demographics, the two that have the highest profit margins are Cluster 1 and 2. It is important for the company to maintain these customers, as they are highly profitable and it is much cheaper to retain customers than to acquire new ones. A typical customer in cluster 1 appears to be more of an average family-centered consumer, who primarily spends money on things they need. They tend to buy more when there is a deal and are fairly well-off, so there is an untapped potential on how much more they could buy from the business. As such, the company's targeted marketing campaigns for cluster 1 should be a loyalty/membership program that focuses on returning customers, offers family discounts, and offers bulk buys. This will entice more customers from cluster 1, since they will be rewarded with deals for both bulk buying and returning as customers, which will further encourage them to keep coming back so on and so forth.

A typical customer in cluster 2 is in their mid 40s with no children and no dependents. They are the wealthiest customers among the three groups. They have the least amount of complaints. They are more likely to make store and catalog purchases and least likely to make purchases with deals. They have the most average amount of purchases, especially in the meat and wine food categories. The customers are the ideal customers in terms of profit. They don't usually don't take deals and make purchases of items at the retail price. The main goals for the company for cluster 2 would be to retain these customers and to increase their spending in the other food categories. If we were to make a recommendation to the company to drive more purchases from this cluster, we would say to showcase the other food categories in the catalog to appeal to these customers since the customers in cluster 2 are visiting the catalog more than the other clusters.

With cluster 3, being roughly the same size as cluster 1 and cluster 2 combined, attention should be focused on increasing new customers from this demographic. A customer from this cluster is older, married with kids, has the lowest income of the three clusters, and rarely buys unless there is a deal. Building awareness and branding are the first steps to conversion. By increasing promotions focusing on goods that align with this group's demographics, such as food necessities and children-related foods and drinks, the company can build additional sales. The cost for a repeat sale is cheaper and building upon those newly acquired customers through loyalty incentives and discounts could prove beneficial.

Logistic Regression

Using binary campaign acceptance as the response variable, we built logistic regression models to find the likelihood a customer may respond to any campaign, as well as specific campaigns 1 through 6. Overall, consumption of wine was proven to be an important factor positively associated with campaign acceptance, as well as number of web visits (positive for all but campaign 5). Interestingly, the number of store purchases and recency tend to have a negative impact on campaign response, whereas customers having no child have a higher chance of accepting a campaign.

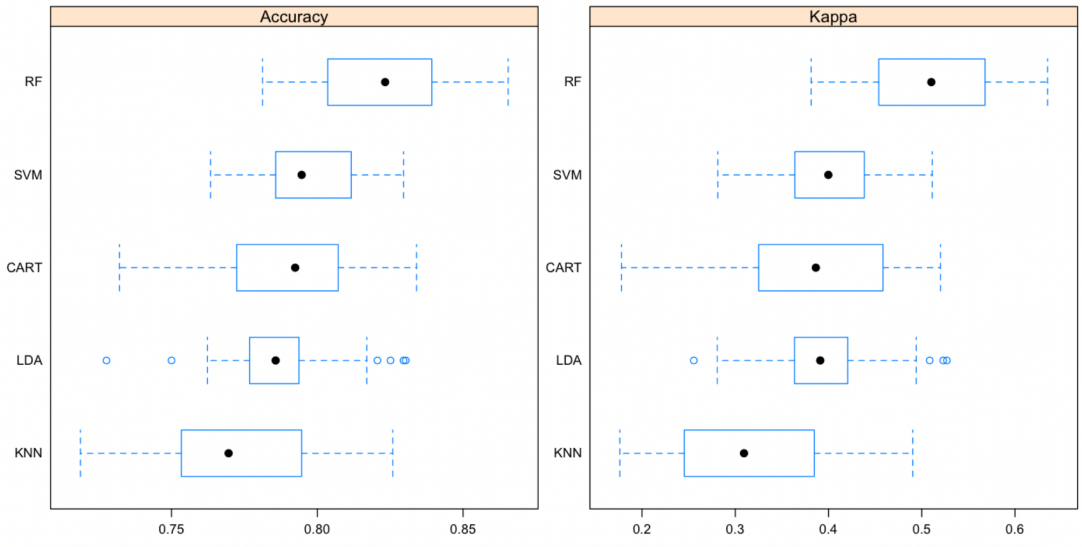
It is important to note that previous campaign acceptance has a consistent positive effect on future campaign acceptance, which represents an opportunity for the company to build up the momentum by engaging with their customers consistently.

Take the model with "any accepted campaign" as the response as an example, customers without a dependent have a probability of 12.9% to accept at least one campaign, 7.2 percentage points higher than customers having dependent(s). However, a customer making 5 purchases in store would have 2 percentage points' drop in acceptance probability compared to a customer who made just 1 purchase in store. A summary of the significant variables in the tested models are attached in Appendix B.

Other Classification Models

Another objective of this project is to provide the company with a model to predict future campaign acceptance. To do that, we used the same predictors and 10-fold cross validation to compare the performance of a set of classification models, including Decision Tree, Random Forest, KNN, Linear Discriminant Analysis, and Support Vector Machine. The results are presented in the box plots below. With a mean accuracy rate of 0.82, the Random Forest model performed the best.

Figure 12. Accuracy and Kappa scores for Classification Models

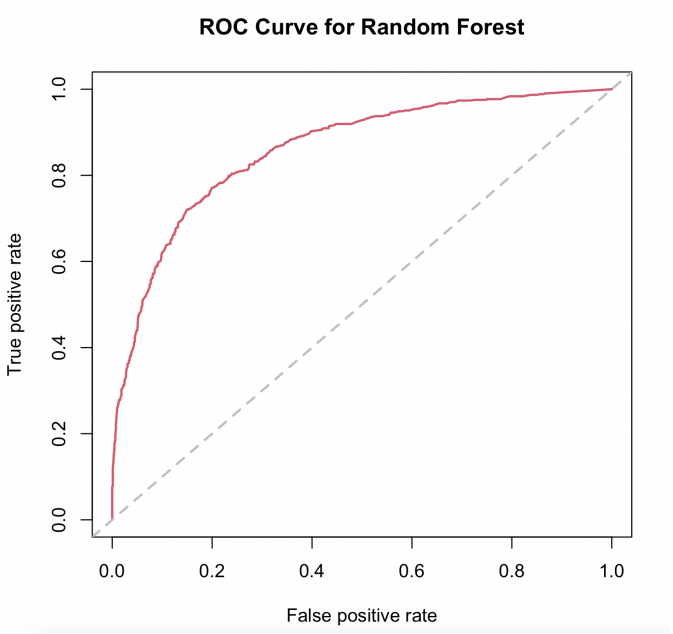


After running the Random Forest model again with 100 trees, we see that although the out-of-bag error rate is 17.58%, a sensitivity of 53.71% is quite low. The company needs to be reminded that the model is better at eliminating the non-converters than picking out the true supporters.

	Predicted Negative	Predicted Positive
Actual Negative	1517	112
Actual Positive	281	326

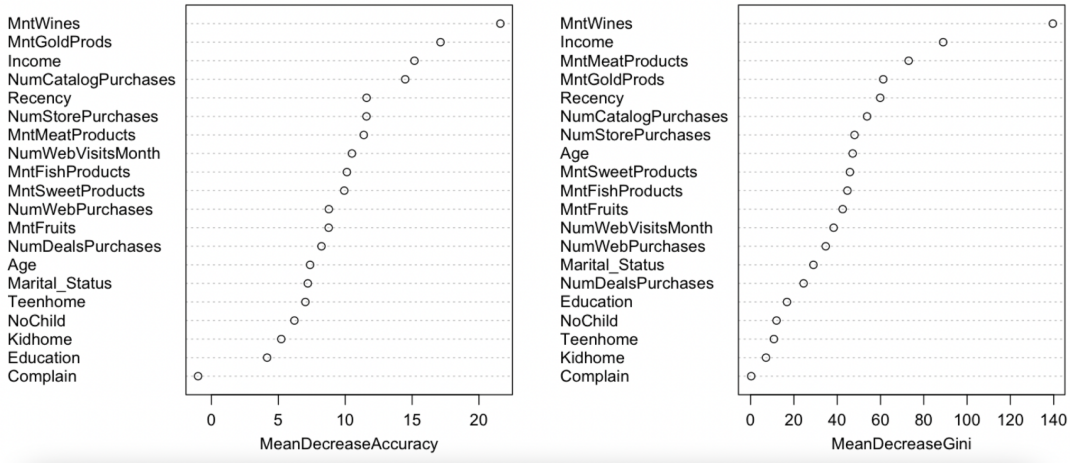
The ROC curve is plotted as below and the area under The ROC curve (AUC) is 0.859471.

Figure 13.



This model also helps confirm the importance of variables. Income, recency, purchase of wine, gold and meat, as well as catalog and store purchases prove to be the most important.

Figure 14. Variable Importance



2nd Dataset Conclusions

In terms of incorporating a second dataset, we attempted to associate average income data (per capita) from World Bank (https://data.worldbank.org/indicator/NY.ADJ.NNTY.PC.CD?end=2014&start=2012). We collected per capita data for the years 2012 through 2014, which coincided with customer registration in our main data. From here, we calculated the average income as a means to combine the two data sets. However, the average incomes calculated from this second set were significantly lower than presented in our main dataset, specifically, for India, Saudi Arabia, and Montenegro. See the table below for the comparison of mean incomes. One possible reasoning for this disparity could be associated with the data sample sizes. Our main dataset comprised 2,240 total records whereas the worlddata sample is significantly larger. Seemingly, the company dataset represents a wealthier sub-population for the three countries.

Country	Code	Mean Income 2012	Mean Income 2013	Mean Income 2014	Mean Income from dataset
Canada	CA	$42,991.14	$43,034.90	$41,604.75	$53,050.62
Australia	AUS	$53,474.85	$53,920.04	$48,738.75	$51,804.29
India	IND	$1,249.08	$1,249.52	$1,368.60	$49,016.41
Montenegro	ME	$5,775.26	$6,348.01	$6,475.53	$57,680.33
Saudi Arabia	SA	$19,659.59	$19,473.33	$19,536.13	$54,830.82
United States	US	$44,727.36	$45,665.51	$47,701.37	$53,218.37

Conclusion

In this project, we have first examined the consumer base of the company. We found a most valuable cluster of consumers: A typical customer belonging to this cluster is single, in their mid-40s, has no child, earning an average of $77K per year, rarely complains, likes to purchase both wine and meat, prefers store purchases most and catalog second, and takes few deals.

Then, we analyzed campaign acceptance. We found that the latest (6^{th}) campaign with an acceptance rate of 14.85% is the most successful one, almost twice as successful as the previous ones. In addition, consistent results showed that customers who don't have kids, enjoy buying wine, prefer to shop via catalog but not store, and visit the website more often are more responsive to the company's marketing campaigns overall.

We have also confirmed that wine consumption has a significant positive impact on both number of purchases and campaign acceptance, while higher education and more dependents are consistently the suppressing factors on revenue and engagement. It is clear that previous campaign acceptance has a consistent positive effect on future campaign acceptance.

Lastly, from our experiments, Random Forest is proven to be the most accurate classification model. It has a high specificity that can help the company eliminate the non-converters.

We conclude with the following business implications based on the findings summarized above:

- Customer acquisition:
    - Improve store front presentation and customer service as stores are the most profitable channel
    - Focus acquisition effort on those who fit the most valuable customer profile stated above

    - Make offers of wine and meat to those who fit the most valuable customer profile to encourage them to buy more
    - Evaluate customer value by predicting cluster assignment or applying regression with purchase volume
- Campaign engagement:
    - Push campaigns digitally to those who are more savvy on the Internet
    - Examine the materials of the most recent campaign and replicate its success
    - Run a widenet campaign and then focus on building up the momentum using follow-up targeted campaigns as previous campaign acceptance influences the future
    - Predict likelihood of campaign acceptance via the trained Random Forest model and focus resources on the potentials

For further analyses, the company can conduct more in-depth regression analysis within each of the clusters. It would also be beneficial for the company to examine the campaign materials – copies and creatives – as well as targeting strategies in order to map our findings to the reason for success of individual campaigns.

Appendix A

Significant Variables in Multiple Regressions
(ranked in the order of the number of times they are significant)

Significant Variables	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
EducationGraduation	-0.7022*	-0.6137*	-0.9740**
EducationMasters	-0.7456*	-0.6424*	-0.8334**
EducationPhD			-0.8594**
Income	9.195e-06*	2.937e-05***	
Kidhome	-0.2981**	-0.7202***	-0.5755***
Teenhome	0.2147*	-0.3532**	
NumDealsPurchases	0.07837**	0.1921***	0.09245**
NumCatalogPurchases	-0.05688*	N/A	-0.09028***
NumWebPurchases	N/A	-0.04741*	
NumStorePurchases		-0.07047***	N/A
NumWebVisitsMonth	0.3317***	-0.1666***	-0.1594***
Recency			-0.003795*
AcceptedCmp11		0.6768***	
AcceptedCmp21	-1.003**		1.259**
AcceptedCmp31		1.100***	-0.4484*
AcceptedCmp41			0.3715*
Response1			-0.8665***
Age			-0.008839*
log(MntWines + 1)	0.7258***		0.8638***
log(MntMeatProducts + 1)		0.8444***	0.3069***
log(MntFishProducts + 1)		0.07406*	0.09265*
log(MntSweetProducts + 1)	0.1759***		0.2264***

log(MntFruits + 1)			0.1751***
log(MntGoldProds + 1)	0.3171***		
Adjusted R-squared	0.5253	0.6345	0.6227

Appendix B

Significant Variables in Logistic Regressions
(ranked in the order of the number of times they are significant)

Significant Variables	Any	Campaign 1	Campaign 2	Campaign 3	Campaign 4	Campaign 5	Response
Baseline	-2.799*** (0.057)	-20.56 (1.18e-09)	-21.50 (4.60e-10)	-1.819* (0.140)	-0.1972 (0.451)	-0.1742* (0.543)	-4.113*** (0.016)
MntWines	0.003212***	0.002584***	0.003141***	0.0007698*	0.003983***	0.004441***	
NumWebVisitsMonth	0.1587***	0.1185**	0.2755**	0.1660***		-0.3858***	0.2338***
NoChild1	0.8923***	2.22***	2.395*				0.6471*
Kidhome	0.4909**	1.258**			-0.8088*	1.489*	
Income		0.00001011**			0.000006514*	0.00001529***	
NumDealsPurchases				-0.1651**		-0.5850**	0.1121*
NumCatalogPurchases	0.07593*	0.1363***		0.2534***			
NumStorePurchases	-0.1311***			-0.1681***			-0.1309***
MntFishProducts		0.004035*			-0.006260***	-0.006808**	
MntGoldProds	0.002627*			0.008238***	-0.006967**		
MntMeatProducts	0.0009035*						0.001998***
NumWebPurchases						0.1217**	0.08752**
Recency	-0.01145***						-0.02822***
Teenhome		0.9262*					-6.393**
EducationP							1.931*

hD[1]							
Marital_StatusMarried[2]							-1.155***
Marital_StatusTogether							-1.171***
Marital_StatusWidow					1.077*		
Age				-0.01911*			
AcceptedCmp11	N/A	N/A	1.851***	0.8126*	1.774***	0.8391**	1.053***
AcceptedCmp21	N/A	N/A	N/A		5.097***		1.288*
AcceptedCmp31	N/A	N/A	N/A	N/A		0.8203*	1.782***
AcceptedCmp41	N/A	N/A	N/A	N/A	N/A	1.579***	1.032***
AcceptedCmp51	N/A	N/A	N/A	N/A	N/A	N/A	1.546***

[1] Base level: “Basic”
[2] Base level: “Divorced”