Black Friday Project

Radhika Kalani, Dilpreet Singh, Sophia Tsilerides

Introduction

Black Friday is one of the most important retail and spending events in the United States. Businesses and retailers traditionally use Black Friday to make enough sales to put them in "the black" for the year and has become increasingly important for brick and mortar stores (physical stores) because of the presence of online retailers like Amazon taking most of their customers by offering competitively low prices every day.

Using the about 550,000 observations of Black Friday consumer purchase data from a retail store we found on Kaggle.com, we can analyze customer purchase behavior against different products using customer demographics. This can be used by management in big box retail chains to better identify customers who spend a lot of money during Black Friday and target their marketing efforts more precisely. We will be using multiple regression, trees, and clustering to predict purchase amounts of categories for customers of various ages, occupations, gender, and other predictors the dataset provided us with.

Data Structure

In the Black Friday dataset, there are a total of 537,577 records and 12 variables.

There are five qualitative predictors of Purchase and the rest are quantitative predictors. The five qualitative predictors are "Product_ID" (Product ID), "Gender" (Sex of User), "Age" (Age in bins), "City_Category" (Category of the City (A,B,C)) and "Stay_In_Current_City_Years" (Number of years stay in current city).

The quantitative predictors include User_ID (User_ID), Occupation (Occupation), Marital_Status (Marital Status), Product_Category_1 (Clothing), Product_Category_2 (Electronics), Product_Category_3 (Home Goods) and Purchase (Purchase amount in dollars). A customer can make multiple purchases in any of the three product categories and if a customer makes no purchases in any product category, the data reports a null value. Additionally, there were a lot of null values for Product_Category_2 and Product_Category_3. To not lose the purchasing power of these variables, the null values were changed into the respective variable's mean value.

Because occupation and marital status are distinct categories, we have converted them into categorical variables instead of continuous for further analysis.

Raw Results

The minimum and maximum mean and median of each quantitative predictor can be found in Tables 1 to Tables 6 in the Appendix.

Summary statistics based off of Table 1 shows that there are 3,623 unique values for Product_ID and in Table 2 shows that 25% of the data points were Females and 75% were Males. Therefore, the sample data may better represent male gender or it may indicate that there are more males in the population than females.

In Table 3 we can see that ages 26-35 and ages 36-45 constitute the majority of the population sample with 40% and 20% respectively. Table 4 displays that 42% of the population belongs to City_Category B, 31% to City C and the 27% left pertains to City A.

Table 5 presents the number of years stay in current city and we can observe that the majority of the sample population (35%) stays one year, following with 19% of the population staying two years and 46% constitutes the rest of the years.

Observations

Box Plots

In regards to the box plots, they all seem to be very even throughout the different categories as shown in Figures 1 to 5. In Figure 2, Purchases by Gender, the third quartile for males is about 1000 purchases greater than females. In Figure 4, Purchases by City, City C seems to have a slightly higher median than City A and B, which suggests the residents of City C possess a unique factor that influence them to spend more money.

Histograms

In regards to histograms, each one is unique. For size of purchases histogram in Figure 5, the graph has a right skew, with most of the purchase amounts between $4,000-$10,000.

For the frequency of different occupations in our dataset, as shown in Figure 6, the graph curves downward, with 25,000 purchases being made by those with occupations between 0-3, while those with occupations between 3-20 account for the rest 30,000 purchases. Although the actual occupation that relates to each number was not disclosed, because of the high concentration of purchases for occupations 0-3, it can be inferred that these are higher paying jobs.

None of the product category histograms in Figures 7, 8, and 9 seem to have any noticeable pattern. Figure 10 shows the frequency of marital status of customers with the bar less than 0.5 representing singles and the bar greater than 0.5 representing married couples. There was a higher frequency of single individuals, roughly 10,000 more, as opposed to married individuals.

Scatterplots

Since most of our data is categorical, and if it was not we transformed it to be categorical, therefor are no scatterplots of significance to show.

Key Takeaways From EDA

Amongst the approximately 55,000 purchases made in this study, the average amount spent was $9,334.

Age and the length of residency in a city seem to be irrelevant to the amount purchased, as the different statistical measures are all fairly close throughout the numbers. In regards to gender, more males seemed to have spent a higher amount as opposed to females. Amongst the three cities, City C edged out over the others with a higher median and quartile range.

There is a noticeable correlation between gender and purchase amount, as well as the city category and purchase amount.

We will take this summary into consideration in further analysis.

Linear and Multiple Regression

When trying to create models, we realized that there were a lot of empty values in the database, mostly within Product_Category_1, Product_Cateogry_2, and Product_Cateogry_3. Since this would cause a problem when making models and comparing MSE, we decided to replace all null values with a 0's. We also tested out our models to see if MSE changed if we replaced all null values with the respective variable's mean value.

Thus, the first model we created included all variables except Age and Stay_In_Current_City_Years since we discovered through EDA these two predictors have no affect on the amount purchased. After checking the summary for the first model, User_ID and Marital_Status both revealed to be insignifcant to the model.

The second model we created included only significant predictors. Based on the summary, all of these variables remained significant. After plotting the predicted values and the residual values, the residual graph seemed quite streaky. Thus, we knew we had to try other types of models, not just linear, since this was not a linear relationship.

In the third model, we decided to take the reciprocals of the values in Product_Category_1, Product_Category_2, and Product_Category_3. We tried taking the log and squares for the remaining variables as well. The residual plot was less streaky and seems to have multiple negativly-sloped lines.

In the fourth model, we decide to multiply the different product categories with each other and add that with the product of Occupation and City_Category. In the fifth model, we got rid of the Occupation variable and just added the product of the product categories to Gender and City_Category.

To select the better of the two models (model 4 and 5), we develop the models further by using a training data set with 1% of the original observations and test the models' performance using the remaining 99% of the data. We only trained 1% of the data set because this such a large data set and processing MSE for a large dataset was taking a long time on R Studio. Using Mean Squared Error (MSE) to access the models' performance, Model 4 is the better model becase it has a smaller MSE.

We also decided to see if our MSE or answer would change if we didn't replace the null values with the mean, but instead replaced them with 0's. We trained our model the exact same way and used 1% of the original observations to train the data set. According to this method, MSE, AIC, and BIC were all smaller for the model with 0 values and Model 5 is a better model because it has a smaller MSE and AIC.

While creating and testing different models, everytime we would take out Product_Category_1 from the model, the MSE would increase heavily. When we would take out other significant variables, MSE would increase only a little. Based on this simple fact, we can determine that Product_Category_1 has the most siginficant impact on Purchase power.

Decision Trees

After applying the cv.tree() function to our tree fitted for the training data of a random 1% (5,375 rows) of the original observation in the Black Friday dataset, with Purchase as the response and all other variables as predictors, the number of nodes corresponding to the lowest error rate was 10 nodes. This means that the optimal tree size to make predictions for Purchase is 10 terminal nodes as this amount of terminal nodes produces the least Mean Square Error *(*Figure 11*)*.

The pruned tree corresponding to the optimal tree size is shown in Figure 12 in the Appendix. This tree has nine internal nodes and ten terminal nodes or leaf nodes. The MSE for the pruned tree is 9,538,890. Additionally, a random forest analysis of 500 trees and 4 variables tried at each split was conducted. The MSE for the random forest analysis was 10,203,551, thus showing the pruned tree was a better model due to lower MSE.

For the random forest analysis, the importance of each variable was also measured, which can be seen in Figure 13. From these graphs, it is clear that the variable *"Product_Category_1"* has the greatest importance, which can be also seen in the pruned tree as it is the only variable present in the tree.

Optimal Model

Based on our multiple regression and trees/random forests analysis, we conclude that the trees method created the optimal value since the lowest MSE between all the models was for the Pruned Tree (9,538,890) and the variable with the greatest importance is "*Product_Category_1."*

K-Means Clustering

To provide the best data for our audience's targeting marketing efforts, we have performed a clustering analysis on our dataset. Unlike our previous models that make predictions, we want to see if there are any similarities amongst our observations so that we can divide them into subgroups and try to predict future purchase behaviors. The proposed variables we begin with are gender, age, occupation, and city. In other words, we want to see amongst these variables, is there any pattern for determining how many clothing products purchases a customer makes. There are too many observations to perform hierarchical clustering on our machines, resulting in a memory exhaust error, so we use K-means which is better for large databases. However, the centers for K-means must be continuous, so we converted all our variables to continuous variables for the purpose of this clustering.

We begin by removing Product Category 1, the clothing category, from our dataset. If we had a set y variable, we could have used this saved category to check the extent to which these classes agree with our results later. However, we will just use them to make a Confusion Matrix and see if our data can be clustered.

There are 18 product categories for clothing and we would like to separate them into 4 subgroups. In other words, we would like customers to fall into one of four purchasing patterns for targeting marketing efforts.

We first begin by setting seed to 1 on RStudio 5.0 Macintosh, running the function 1 time. We then tried a variety of initial cluster assignments and nstarts to see how they affect the clustering results in terms of between-cluster sum of squares and within-cluster sum of squares. We also tried standardizing our data. The result of both outputs are shown in Figures 14 and 15.

Doing K-means many times with many different assignments shows how that initial random assignments impact the final clustering result. The best iteration of K-means clustering had a global optimal of 21,839,520 for between-cluster sum of squares and 2,435,312 within-cluster sum of squares.

When we plot the within-cluster sum of squares against K, we can try to find the optimal value of K. As shown in Figure 16, between 4 clusters and 6 clusters, the within-cluster sum of squares change is

not significant. Therefore, performing K-means using our selected K=4 is satisfactory, however, our confusion matrix showed clusters were well distributed and there was not much we could interpret. Therefore, we increased our cluster size to 6.

When we increased our K to 6, with a seed of 1 and nstart of 1, our Confusion Matrix had better results. Also, the within-cluster sum of squares was lower as shown in Figure 17.

In Figure 18, we take a closer look at some of the clustering results using the Confusion Matrix which has been conditionally formatted for legibility. Consumers that bought 11 Product Category 1 items are best clustered in Cluster 1, 5 or 4 where 40%, 27% and 14% of the results fall. However, Consumers that bought 9 items have 35% of its records in Cluster 1 and 35% in Cluster 6, so it doesn't matter which cluster we assign it to. Lastly, consumers that bought 17 items had 40% of its records clustered in Cluster 4 and less than 15% of the rest of the records in Cluster 2, 3 and 6. So customers of this cluster demographic that bought 17 items are very dissimilar to customers that only bought 9 items.

Similar results were obtained when we clustered using just gender and age as our x variables.

Given by our results, we can see there was no product category that had more than 40% of its records in one cluster. The clusters are not great, even with the optimal K, because of the actual data collection method itself.

Conclusions and Implications

We recall that the data was collected on Black Friday, an exceptional shopping day for consumers. Therefore, the data set does not have reflect the regular purchase behavior of customers. Sex and gender, which usually distinguish purchase patterns, do not influence purchase choices on Black Friday as they do on other days. Instead, the purchase behavior is extremely random and influenced by factors that are not included in our dataset like who would be at actual end user of the product or if the item is a gift, or if the purchase amount is made on a whim influence by the price. The variance of the clusters reflects this. If we tried to produce Association Rules from this dataset, they would be meaningless because they would not be able to be applied in future instances.

Therefore, given this dataset, supervised algorithms that target big buyers based on purchase amount are more useful.

Appendix

Summary Statistics - Categorical Data

Table 1. Counts/ Proportions for Product_ID

	Product_ID
P00265242	1858
P00110742	1591
P00025442	1586
P00112142	1539
P00057642	1430
P00184942	1424
(Other)	528149

Number of Observations = 537,577

Table 2. Counts/ Proportions for Gender

	Gender
F	132197 (0.25)
M	405380 (0.75)

Number of Observations = 537,577

Table 3. Counts/ Age

	Age
0-17	14707 (0.03)
18-25	97634 (0.18)
26-35	214690 (0.40)
36-45	107499 (0.20)
46-50	44526 (0.08)
51-55	37618 (0.07)
55+	20903 (0.03)

Number of Observations = 537,577

Table 4. Counts/ Proportions for City_Category

	City_Category
A	144,638 (0.27)
B	226,493 (0.42)
C	166,446 (0.31)

Number of Observations = 537,577

Table 5. Counts/ Proportions for Stay_In_Current_City_Years

	Stay_In_Current_City_Years
0	72725 (0.14)
1	189192 (0.35)
2	99459 (0.19)
3	93312 (0.17)
4+	82889 (0.15)

Number of Observations = 537,577

Summary Statistics – Continuous Data

Table 6a. Summary Statistics of User_ID, Occupation, Marital_Status

	User_ID	Occupation	Marital_Status
Min	1000001	0	0
Median	1003031	7	0
Mean	1002992	8.083	0.4088
Max	1006040	20	1

Number of Observations = 537,577

Table 6b. Summary Statistics of Product Category 1, 2, and 3 and Purchase

	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
Min	1	2	3	185
Median	5	9	14	8062
Mean	5.296	9.84	12.7	9334
Max	18	18	18	23961

Number of Observations = 537,577

Figure 1. Stay in Current City vs Purchase in Box Plot; width adjusted to proportion percentage

Figure 2. Purchases by Gender

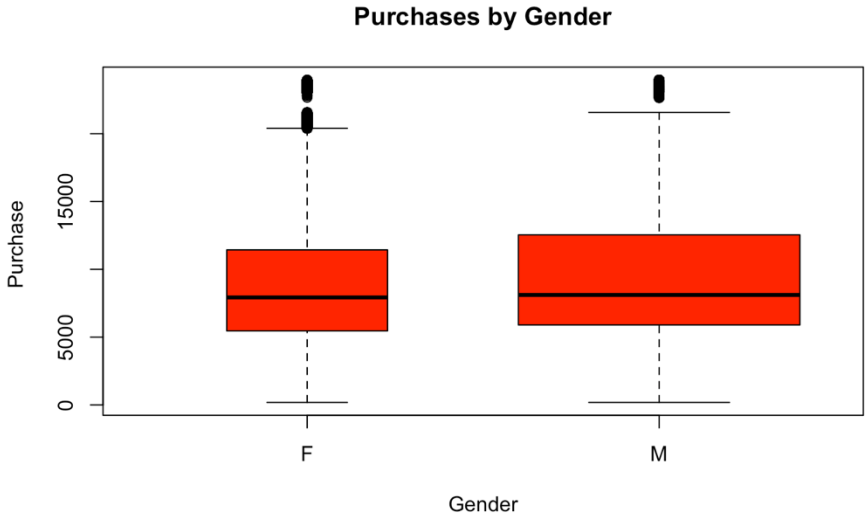

Figure 3. Purchases by Age

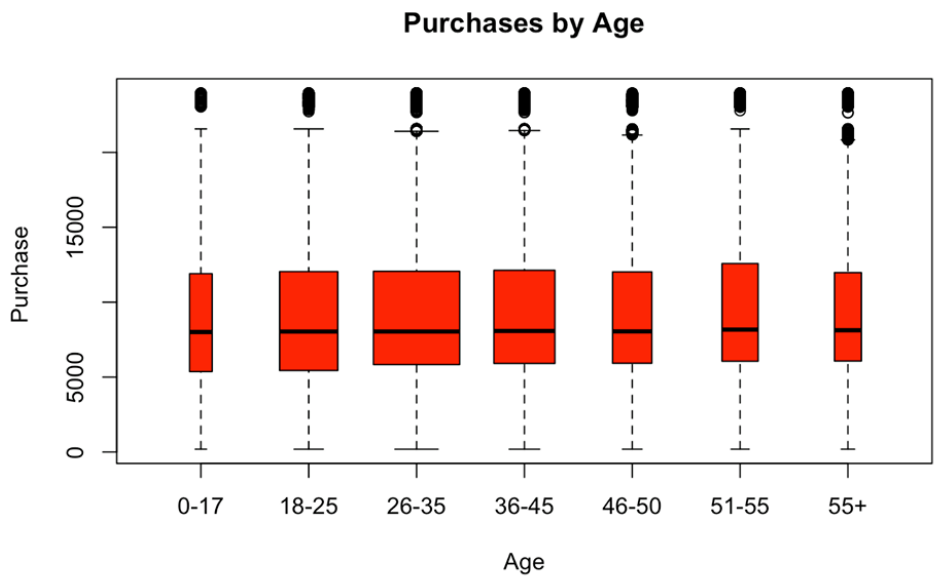

Figure 4. Purchases by City

Figure 5. Frequency of Purchase Size presented in Histogram

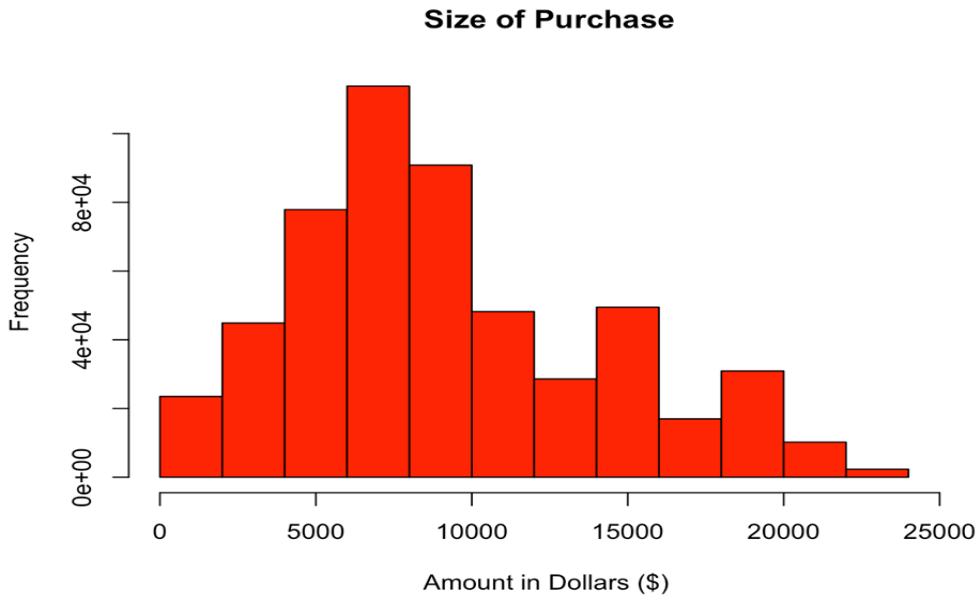

Figure 6. Frequency of Occupation of Purchasers presented in Histogram

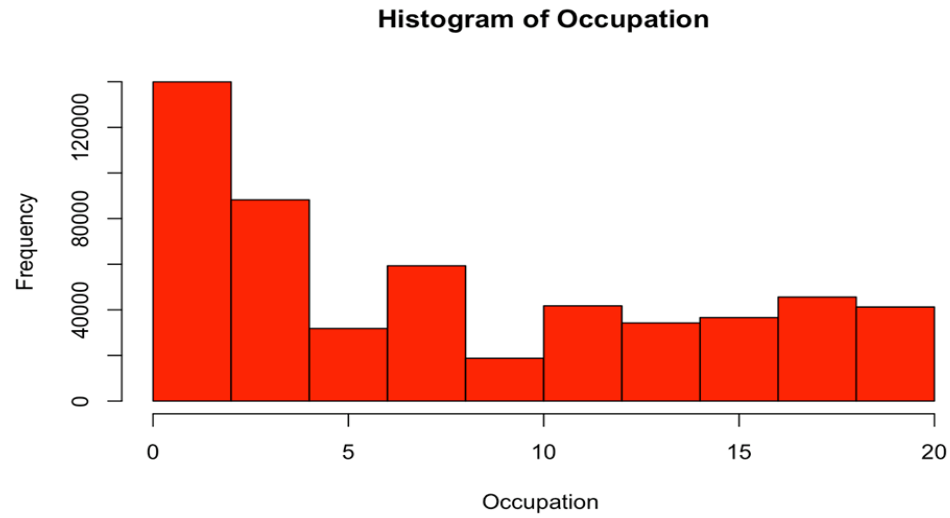

Figure 7. Frequency of Product Category 1

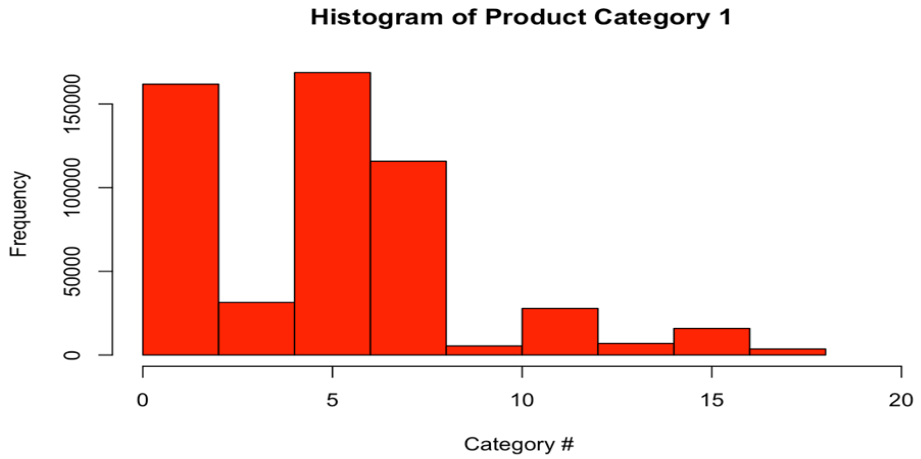

Figure 8. Frequency of Product Category 2

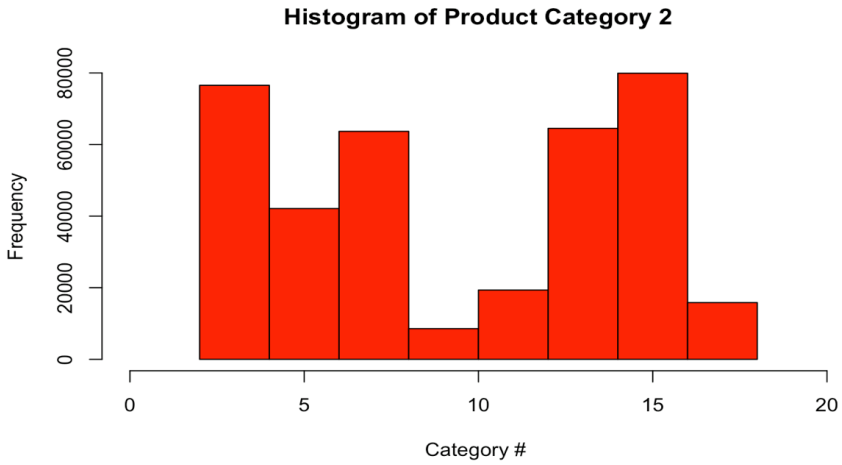

Figure 9. Frequency of Product Category 3

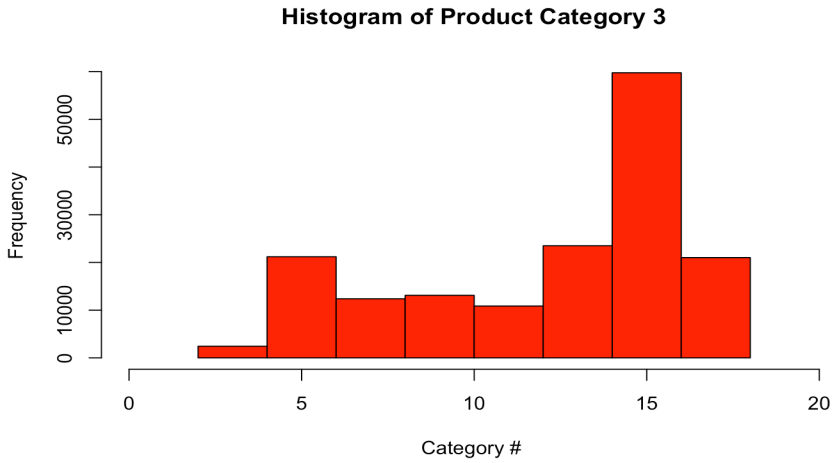

Figure 10. Frequency of Marital Status

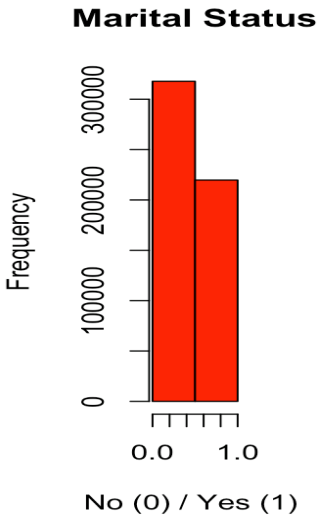

Figure 11. Relationship between Tree Size and Terminal Nodes

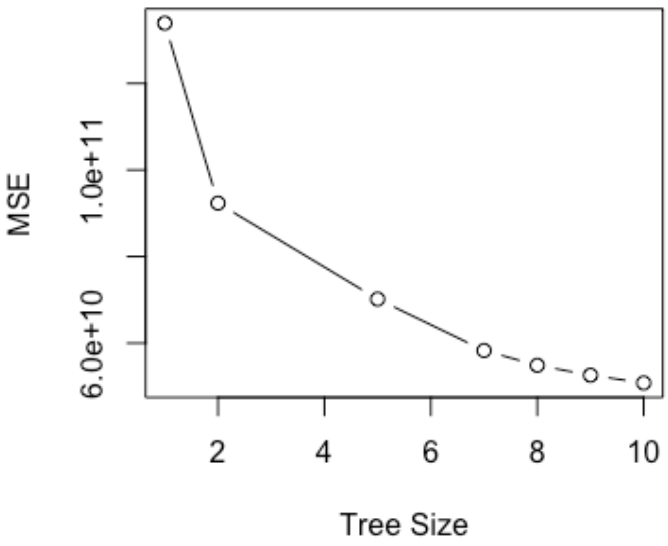

Figure 12. Optimal Pruned Tree with 10 Terminal Nodes

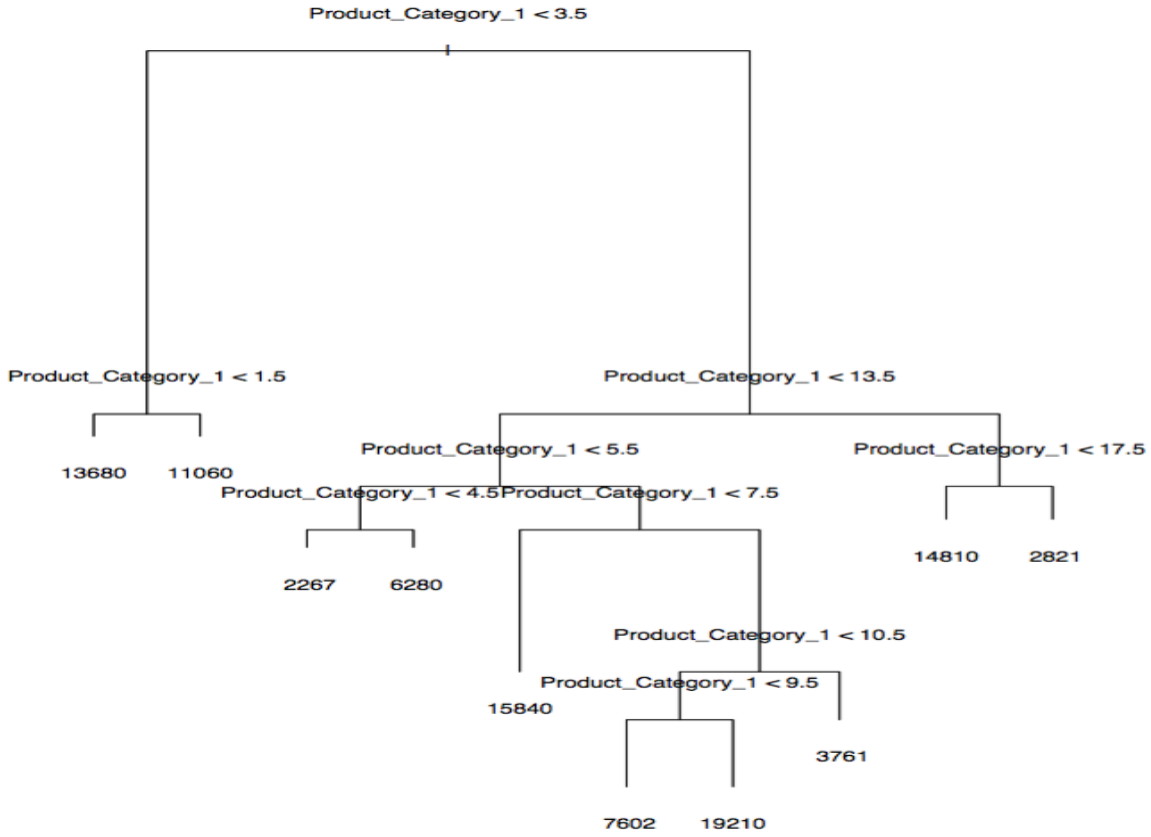

Figure 13. Importance of Variables

Figure 14. Different Variations of K-means in terms of Assignment and nstart without Standardizing

K=4	seed =1, nstart =1	seed = 1, nstart = 20	seed = 10, nstart = 1	seed = 10, nstart = 20
Between-Cluster Sum of Squares	21,839,520	21,839,520	21,433,471	21,839,520
Within-Cluster Sum of Squares	2,435,312	2435,312	2,841,362	2,435,312

Figure 15. Different Variations of K-means in terms of Assignment and nstart with Standardizing

K=4	seed =1, nstart =1	seed = 1, nstart = 20	seed = 10, nstart = 1	seed = 10, nstart = 20
Between-Cluster Sum of Squares	1,107,587	1,109,290	1,073,605	1,108,957
Within-Cluster Sum of Squares	1,042,717	1,041,014	1,076,699	1,041,347

Figure 16. Optimal Value of K

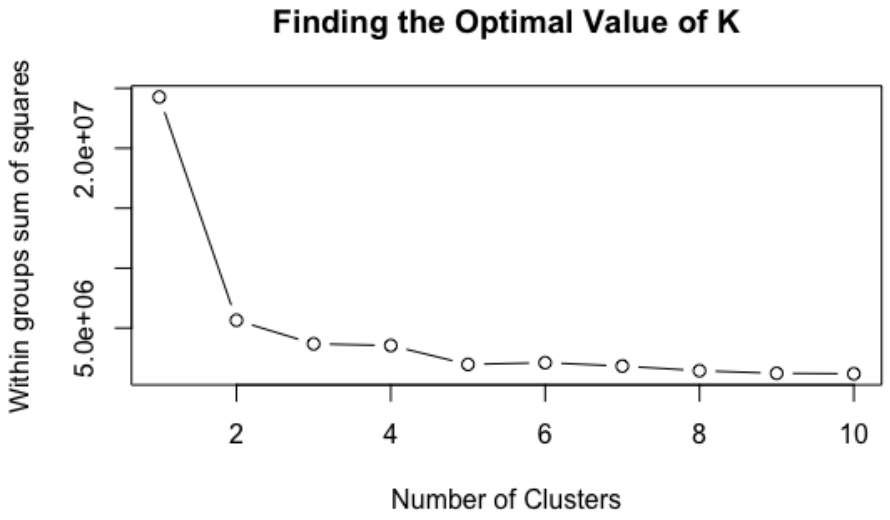

Figure 17. Betweenss and Withinss with different Ks

	K=5	K=6
Between-Cluster Sum of Squares	1,187,769	1,267,465
Within-Cluster Sum of Squares	962,535	882,838.9

Figure 18. Sample 6-means Clustering Confusion Matrix

	Clothing Purchase Items					
Clusters	9 items		11 items		17 items	
	Count	*% of total*	*Count*	*% of total*	*Count*	*% of total*
1	*142*	*35%*	*9546*	*40%*	*146*	*26%*
2	*19*	*5%*	*1627*	*7%*	*17*	*3%*
3	*28*	*7%*	*1798*	*8%*	*10*	*2%*
4	*51*	*13%*	*3239*	*14%*	*227*	*40%*
5	*141*	*35%*	*6516*	*27%*	*133*	*23%*
6	*23*	*6%*	*1234*	*5%*	*34*	*6%*
Totals	404	100%	23960	100%	567	100%