Data Analysis Report:
Ideal customers Personality Analysis

Alex Wang - 18tw7@queensu.ca

Course Project

STAT 362

R for Data Science

January 20, 2022

Contents

1 Introduction

Customer Personality Analysis is a detailed and comprehensive analysis of a company's ideal customers from Kaggle.

We are curious about the characteristics of the customers who are most likely to buy a certain kind of product and the customer segment which is most likely to have a more significant impact on sales volume from this dataset.

Due to the variety of customers' personalities, customer personality analysis could help companies better understand their customers' shopping preferences to maximize the value of customers to the business. Simultaneously, it will lead the company to find the breakthrough point for sales volume. Furthermore, instead of wasting much workforce and financial resources to market a new product to every customer in the company's database, the company can focus on product promotion for target customers.

For this dataset, many other people also analyzed using R. For example, someone first used univariate analysis to analyze the nine attributes and then bivariate analysis to analyze the relationship between customers' shopping preferences and characteristics. From his analysis, he concluded that companies need to improve the performance of their campaigns since nearly 79% of their customers did not accept it. Moreover, There is also a high correlation between accepting the offers and the income.

Other people use RFM analysis to predict future purchases and monetary value(lifetime or dropout rate of a customer) for the customers based on the time period assigned. They focused on 3 main parts of customers' transactions: recency, frequency and purchase amount. First, they formed the segments and understood their purchasing habits. Secondly, use BG-NBD Model to make a conditional expected number of transactions in the next period in order to predict who needs more attention from the store.

2 Description of The Dataset

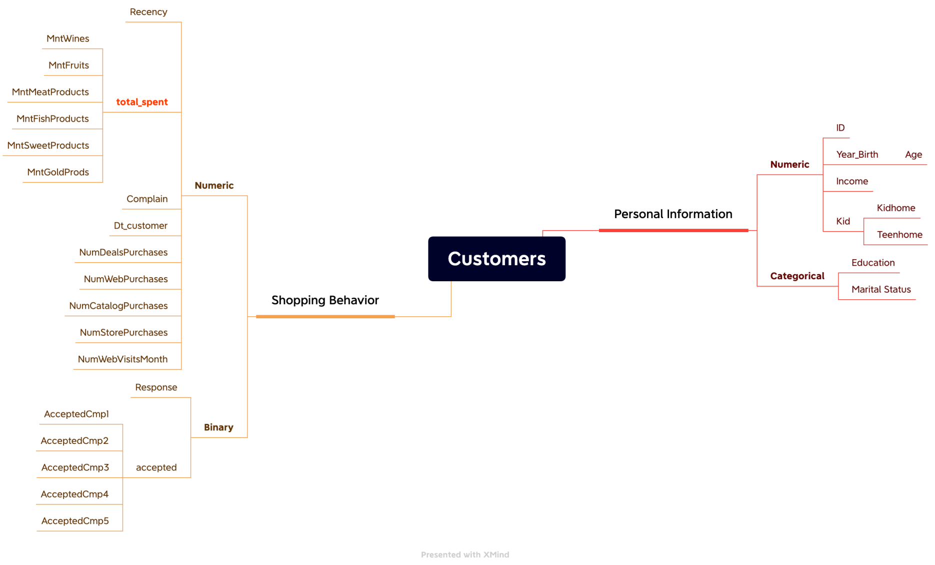

Figure 1: Variables

2.1 Variables

1. ***Customer's basic information***:

Variable Name	Interpretation	Min	Max	Median	Mean
ID	Customer's unique identifier	0	11191	5458	5588
Year_Birth	Customer's birth year	1893	1996	1970	1969
Income($)	Customer's yearly household income	1730	666666	51382	52247
Kidhome	Number of children in customer's household	0	2	0	0.44
Teenhome	Number of teenagers customer's household	0	2	0	0.51

Table 1: Summary of Numeric Variables from Customer's basic information

Variable Name	Interpretation	Level
Education	Customer's education level	0 - 2n Cycle; 1 - Basic; 2 - Graduation; 3 - Master; 4 - PhD
Martial Status	Customer's marital status	0 - Absurd; 1 - Alone; 2 - Divorced; 3 - Married; 4 - Single; 5 - Together

Table 2: Summary of Categorical Variables from Customer's basic information

2. ***Shopping Behavior***:

Variable Name	Interpretation	Min	Max	Median	Mean
Recency	Number of days since customer's last purchase	0	99	49	49.01
MntWines ($)	Amount spent on wine in the last two years	0	1493	174.5	305.1
MntFruits ($)	Amount spent on fruits in the last two years	0	199	8	26.36
MntMeatProducts ($)	Amount spent on meat in the last two years	0	1725	68	167
MntFishProducts ($)	Amount spent on fish in the last two years	0	259	12	37.64
MntSweetProducts ($)	Amount spent on sweets in the last two years	0	262	8	27.03
MntGoldProds ($)	Amount spent on gold in the last two years	0	321	24.50	43.97
NumDealsPurchases	Number of purchases made with a discount	0	15	2	2.32
NumWebPurchases	Number of purchases made through the company's website	0	27	4	4.09
NumCatalogPurchases	Number of purchases made using a catalogue	0	28	2	2.67
NumStorePurchases	Number of purchases made directly in stores	0	13	5	5.8
NumWebVisitsMonth:	Number of visits to the company's website in the last month	0	20	6	5.32
Dt Customer (second)	Date of customer's enrollment with the company	1.34E+09	1.40E+09	1.37E+09	1.37E+09

Table 3: Summary of Numeric Variables from Shopping Behavior

Variable Name	1	0
AcceptedCmp1	customer accepted the offer in the first campaign	otherwise
AcceptedCmp2	customer accepted the offer in the second campaign	otherwise
AcceptedCmp3	customer accepted the offer in the third campaign	otherwise
AcceptedCmp4	customer accepted the offer in the fourth campaign	otherwise
AcceptedCmp5	customer accepted the offer in the fifth campaign	otherwise
Response	customer accepted the offer in the last campaign	otherwise
Complain	customer complained in the last two years	otherwise

Table 4: Summary of Binary Variables from Shopping Behavior

3. ***Other***:

(a) **Numeric**:
Z_CostContact: only contains an integer 3.
Z_Revenue: only contains an integer 11.

4. ***Variable Combinations***:

   We create four continuous variables to help us do data analysis according to the original variables.

Variable Name	Composition	Min	Max	Median	Mean
Age	2022 - Year_Birth	26	100	52	53.18
total_spent ($)	MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds	5	2525	396.50	607.10
Kid	Kidhome + Teenhome	0	3	1	0.94
Accepeted	AcceptedCmp1 + AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5	0	4	0	0.30

Table 5: Summary of New Created Continuous Variables

2.2 Data Visualization & Summary Statistics

1. ***Customer Properties***:

   As Figure 2 shows, the first numeric variable in the dataset is the year of birth. Use 2022 minus their birth year to get the result of their age. It can be found that it is an approximately normal distribution, and the mean is around 53(53.18) years old. The second numeric variable is the customers' yearly household income. A few outliers beyond 15 hundred thousand can be found in the histogram. Those outliers are removed in the following analysis. The Third numeric variable is the number of children the customers had. In the raw data, it splits into teenagers and kids. We sum them up and create a new variable called Kid. There is another numeric variable which is the customers' ID. Since it is just random numbers, it is unnecessary to make data visualization.

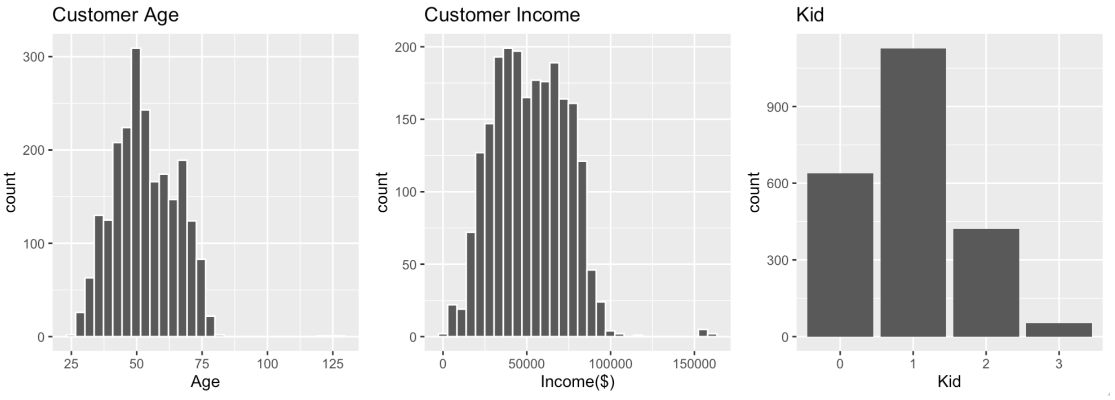

Figure 2: Customer's age, income and kid

2. ***Customer Behavior***:

   In Customer behaviour, There are six numeric variables: money spent on Wines, Fruits, Meat, Fish, sweets, and gold. We sum them up and create a numeric variable named total spend. As the Figure 3 illustrates below, most customers are not used to spending too much on these products overall.

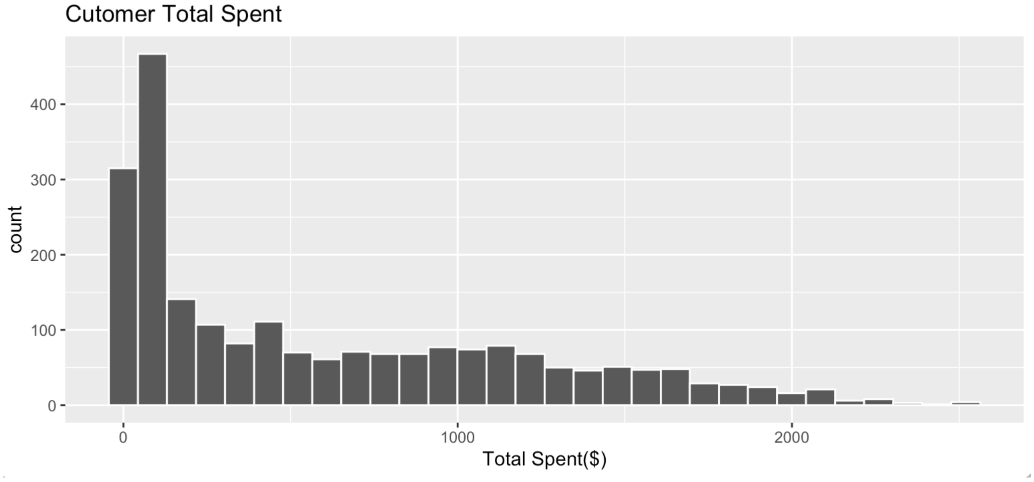

Figure 3: Total spent histogram

3. ***Total Spent & Independent Variables***:

   It can be seen from the scatter plot in Figure 4 and calculated the correlation(age:0.11, income:0.67) that income has a stronger linear relationship with the total spending than age.

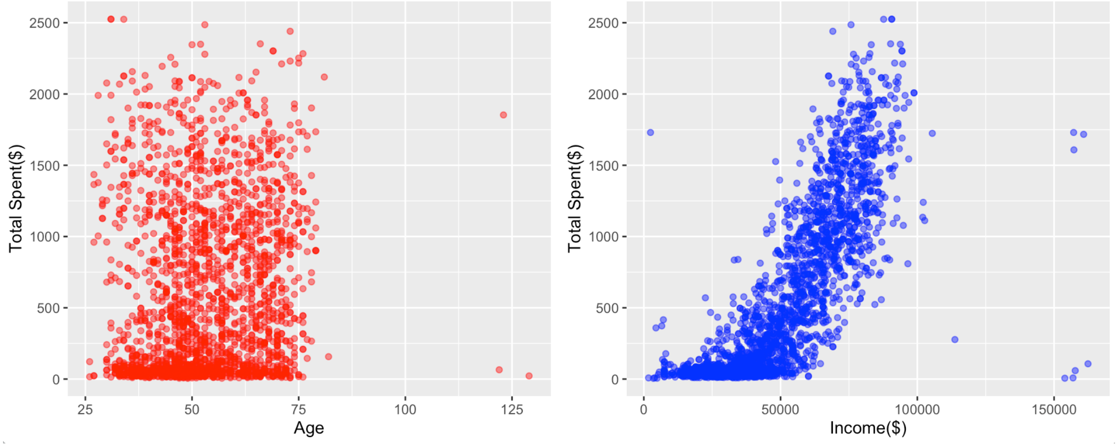

Figure 4: Total spent/age & income scatter plot

As the Figure 5 shows, moreover, people without children spend more than the others, and people who only receive a primary education spend the least. Furthermore, relatively, people who do not have a family and are alone prefer to spend less in the store than the other groups.

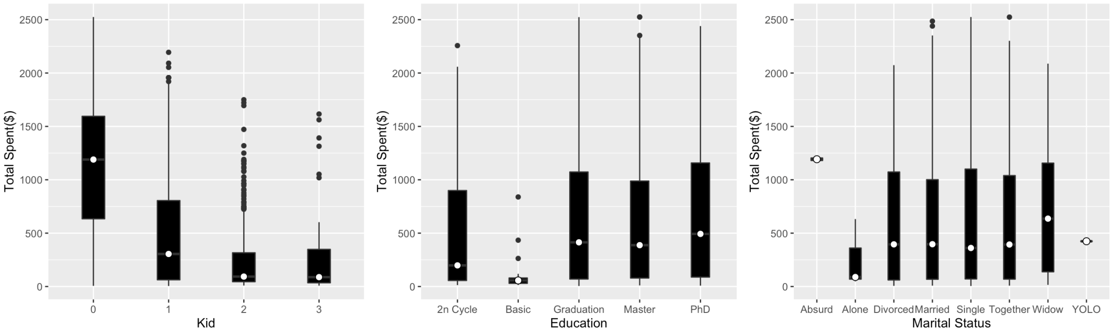

Figure 5: Total spent histogram

3 Method & Results

3.1 K-means Clustering

1. ***Introduction & Direction***

   The first research method uses the total spent as the response variable and personal information and the ways of shopping as the independent variable. The goal is to find customer groups that have a more significant impact on sales. We first tried Hierarchical Clustering and K-means clustering, and we found that the former has enormous complexity, which can result in long computation times compared with K-means. It is complicated to determine the correct number of clusters by the dendrogram for our dataset.

   So we choose to perform k-means clustering with the total spent as the response variable. The reason why we choose this method is that the objective of K-means clustering is straightforward. It can help group similar data points and discover underlying patterns of those customers. Meanwhile, it is very suitable for this dataset since it can tell the relationships between total spent and other information like their income, age, education, and so on in different groups.

2. ***Results***

   As the Figure 6 shows that K-means Clustering divides the customers into four groups, and they all show their information in this plot. For example, group 3 has the most significant total spend, and then we will find which variable will significantly influence the total spent.

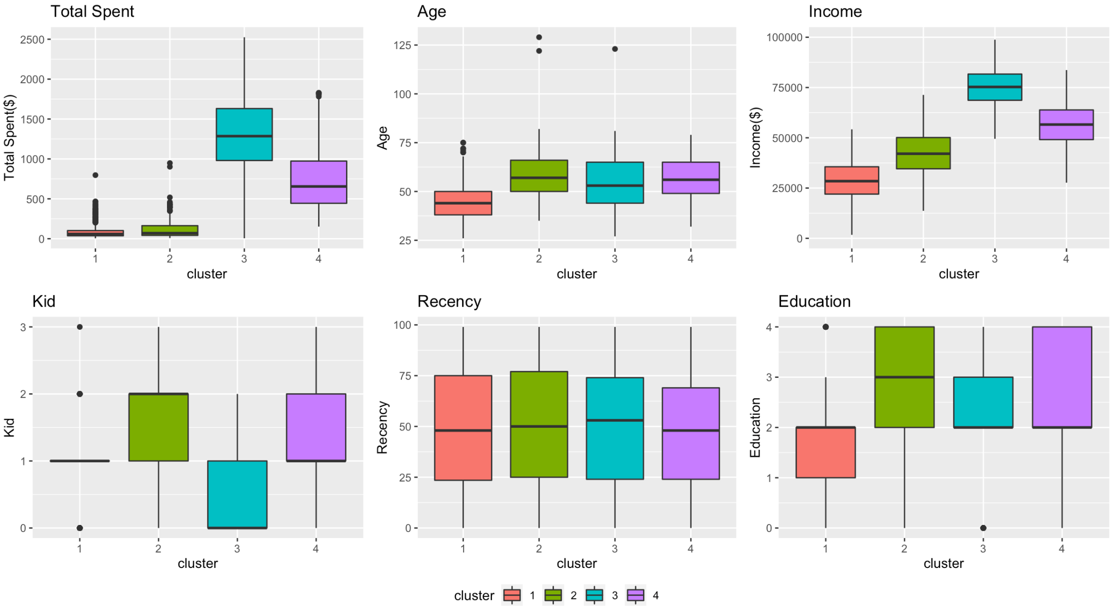

Figure 6: K-means Clustering Result 1

Figure 7 shows the result of relationships between total spent and their ways of purchasing. One of the observations is the cluster that spent the most but spent the least on discounted items.

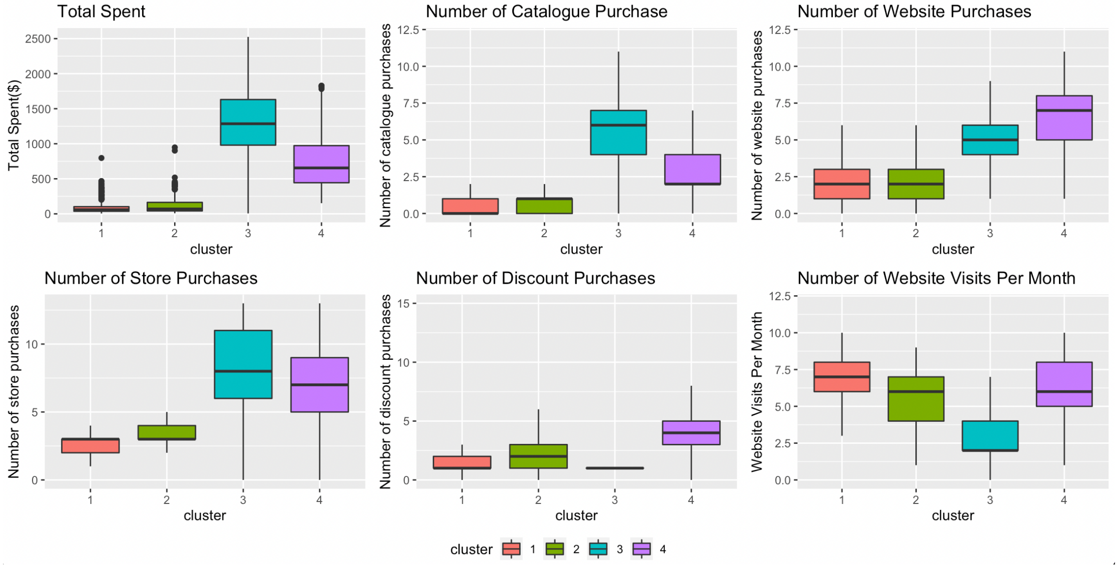

Figure 7: K-means Clustering Result 2

Figure 8 is a summary of each group of customers. Generally, we found our ideal customers are group 3 and group 4, and they have higher income, spending a lot on website and catalogue purchases. Meanwhile, group 4 also has the most discount purchases.

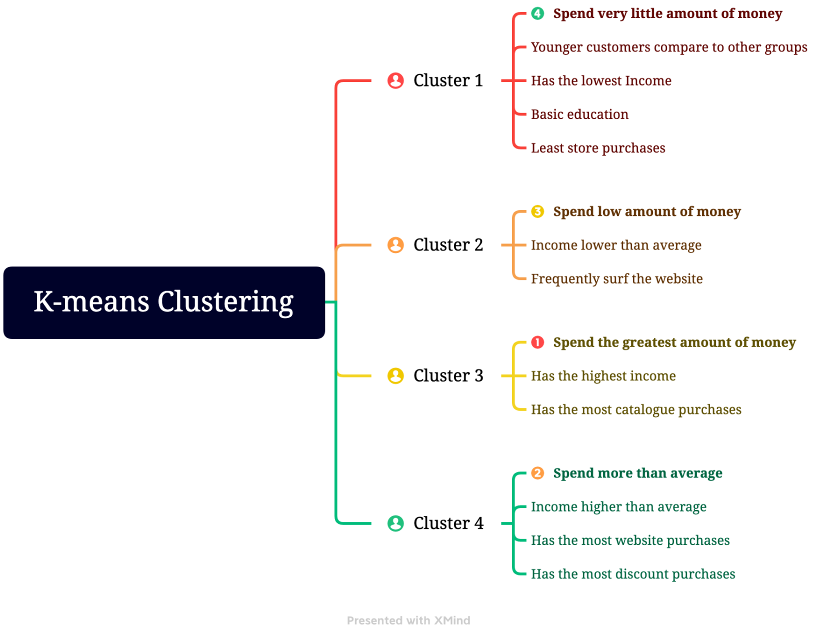
K-means Clustering
Cluster 1
4 Spend very little amount of money
Younger customers compare to other groups
Has the lowest Income
Basic education
Least store purchases
Cluster 2
3 Spend low amount of money
Income lower than average
Frequently surf the website
Cluster 3
1 Spend the greatest amount of money
Has the highest income
Has the most catalogue purchases
Cluster 4
2 Spend more than average
Income higher than average
Has the most website purchases
Has the most discount purchases
Presented with XMind

Figure 8: Four Groups of Clusters

3. ***Implications***

   (a) Cluster1: Even though they are younger and visit the web most frequently, they rarely buy goods. This will remind us that we should improve the customer experience on the web and be able to predict their favorite items through big data.

   (b) Cluster2: The total consumption of this group is somewhat higher than that of the second group, which also proves that the total consumption is relatively proportional to their wages and education. They are also the most frequent web surfers. The same implication as Cluster1 shows.

   (c) Cluster3: They are the most loyal customers, the highest in terms of total spending, and at the same time have the highest salaries of all customer groups. They spend the most in catalogue, while hardly ever buying discounted goods. It is worth studying for their backgrounds and experience, and these will help improving the experience of other customers.

   (d) Cluster4: Although their total spending is not as much as the third group, they prefer to shop online and buy discounted items, which is more in line with the popular spending style nowadays. We can enhance our connection with such customers by learning their spending habits to place ads on the web and predict their future spending habits.

4. ***Potential Problems***

   In our preliminary k-means clustering, we found that a minimal number of clusters appeared all the time, and we found that it was the only 21 customers who had generated a complaint.

If we insist on adding complaints to clustering, there will be a waste of one of the clusters. As the saying goes: One bad apple spoils the whole bunch.

5. ***Cross-validation***

   For this k-means clustering, we did not apply the cross-validation because one of the answers is found from jmb. He says there is usually no clear definition of error in unsupervised learning, such as clustering. Due to this, cross-validation cannot be used for this purpose(jmb, 2016).

3.2 Multiple Linear Regression

1. ***Introduction & Direction***

   The second research direction uses the amount of money spent on different kinds of products as the response variable and personal information and the ways of shopping as the independent variable. Our goal is to determine which aspect is the most important to the customers when selecting products.

   As a start, linear regression is the most straightforward but informative model to help us get a brief idea about the models. First, we formulate some hypotheses and perform the hypothesis testings. We fit six multiple linear regressions with different responses and 13 variables.

   Response:

   There are six different responses.

   Y1: Amount spent on wine in last 2 years - Wines

   Y2: Amount spent on fruits in last 2 years - Fruits

   Y3: Amount spent on meat in last 2 years - Meat Products

   Y4: Amount spent on fish in last 2 years - Fish Products

   Y5: Amount spent on sweets in last 2 years - Sweet Products

   Y6: Amount spent on gold in last 2 years - Gold Products

   Predictor:

   Predictors are the same for the six models.

   X: Education, Marital_Status, Income, Dt_Customer, Recency, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, Complain, Age, Kid

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{13} x_{i13} + \varepsilon \qquad \varepsilon \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \\
\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{13} x_{i13} + \varepsilon \\
H_0 &: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{13} = 0 \\
H_1 &: \text{at least one of } \beta_1, \beta_2, \beta_3, \cdots, \beta_{13} \neq 0
\end{aligned}
$$

2. ***Results***

   Table 6 depicts the adjusted R^2, F-statistics, P-value, and Correlation in six models. We need to consider whether each model has a good overall performance and is statistically significant or not.

	Wines	Fruits	Meat Products	Fish Products	Sweet Products	Gold Products
Adjusted R^2	0.65	0.36	0.63	0.41	0.38	0.30
F-statistics	160	98.44	285.1	120.2	106.5	74.31
P-value	2.2E-16	2.2E-16	2.2E-16	2.2E-16	2.2E-16	2.2E-16
Correlation	0.75	0.58	0.77	0.62	0.58	0.54

Table 6: Summary of Multiple Linear Regression Models

For the responses: the amount spent on Wine and Meat Products in the last two years, the adjusted R^2 is about 0.6, suggesting the two models fit the data reasonably well.

For the left four responses: the amount spent on Fruits, Fish Products, Sweet Products, and Gold Products in the last two years, the adjusted R^2 is relatively low (around 0.3), but the model can still be useful.

Noticed that the P-value in each model is lower than 0.05, which indicates that we should reject the null hypothesis and the six models are all statistically significant from 0.

Since the response is continuous, we cannot use “accuracy” to measure the performance of our prediction. One possible measure is to use correlation. The higher the correlation, the better the prediction. From the table, we can see that each model has a relatively high correlation. Therefore, the prediction is quite good.

In the model evaluation, we do not use cross-validation in model performance.

3. ***Conclusion***

All six models have a good performance and higher correlation and are statistically significant.

3.3 Decision Tree

1. ***Introduction & Direction***

Since multiple linear regression provides a very intuitive model structure as they assume a monotonic linear relationship between the predictor variables and the response, we decided to use the decision tree method because when features interact, it could display more details.

We choose regression trees as our decision tree because our responses are all continuous. First, we fit the regression trees and calculate the MSE in the test data for six responses respectively. Then we use cross-validation to determine the optimal choice of size, which is the number of terminal nodes, creating the pruned trees and calculating the MSE in the test data.

2. ***Results***

Table 7 shows the difference between mean square error and terminal nodes. For the responses: Amount spent on Wine in the last two years, we find that the original terminal nodes are the same as the optimal size, so there is no difference with and without pruning the tree.

Then we look at another two responses: Amount spent on Meat Products and Fish Products in last two years. The test error and the optimal node in the pruned tree are both smaller than the tree without pruning. They achieve the same accuracy as the model with larger

	Wines	Fruits	Meat Products	Fish Products	Sweet Products	Gold Products
MSE Before Prune	42030.38	996.589	20472.15	1818.463	1126.875	1812.777
MSE After Prune	42030.38	1003.492	18852.38	1816.439	1129.661	1941.879
Difference between two MSE	0	-6.9027	1619.77	2.024	-2.786	-129.102
Original Size	9	7	8	11	5	6
Optimal Size	9	4	7	6	4	3
Difference	0	3	1	5	1	3

Table 7: MSE and Size from Decision Tree

numbers, and the performance of the pruned tree in test data is similar to the tree without pruning.

For the last three responses: Amount spent on Fruits, Sweet Products and Gold Products in last two years. The test error in the pruning tree is bigger than the one without pruning, but pruning trees make the models simpler and easier to interpret.

Sometimes we want to both improve model accuracy and reduce the tree size. However, for the three responses we mentioned before, decision trees may not both satisfy the two conditions.

3. ***Implication and Insight***

Decision trees do have some advantages. Trees require very little pre-processing and outliers typically do not bias the results as much since the binary partitioning simply looks for a single location to do a split within the distribution of each feature.

However, individual decision trees generally do not often achieve state-of-the-art predictive accuracy. They typically lack predictive performance compared to more complex algorithms like neural networks and MARS.

Sometimes we want to both improve model accuracy and reduce the tree size. However, for the three responses we mentioned before, decision trees may not both satisfy the two conditions.

We tried to split the dataset into three parts: training data (50%), validation data (25%), and testing data (25%). First, use training data to do cross-validation and choose the optimal terminal nodes. Then, use validation data to evaluate the model performance. Last, use testing data to refit the model, create the pruned trees and calculate the mean square error in the test data. In this case, we may determine the “best fit” tree in each model to have both a smaller size and mean square error.

Another improvement is that we can use the greedy algorithm, making an optimal local choice at each mode, to decrease the test error and make a better prediction on each model.

3.4 Ridge Regression

1. ***Introduction & Direction***

   Through the property of ridge regression, it is clear to know that this method can significantly reduce overfitting, which may cause failure to find unique solutions. We apply ridge regression by using the amount of money spent on each kind of product as a response and customers' personal information and the number of purchases made by different ways of shopping as predictors. We use half of the data as a training dataset and perform ridge regression with cross-validation. Since we have a large number of observations, we use the default option as k equals 10.

2. ***Results***

	Wines	Fruits	Meat Products	Fish Products	Sweet Products	Gold Products
Linear Regression	53022.67	1046.47	21669.64	1809.19	1123.65	1869.47
Ridge Regression	43541.63	1017.79	20302.77	1764.93	1049.95	1836.13
Random Forest	4846.10	165.50	2588.31	274.94	181.03	317.17

Table 8: Summary of MSE from Linear Regression, Ridge Regression and Random Forest

   From Table 8, we compared with the MSE between linear regression and ridge regression, it is obvious that there is a decreasing pattern in each kind of product.

3. ***Conclusion***

   By fitting ridge regression, the model has a better performance than linear regression. However, the MSE is still preserved at a high level, and we would like to get a relatively small number by applying other methods.

3.5 Random Forest

1. ***Introduction & Direction***

   We compared the MSE of random forest, linear regression, and ridge regression in predicting the amount of money spent on each kind of product using customers' personal information and the number of purchases made by different ways of shopping. As a result, random forest generates the smallest MSE. We tried models with 1000 trees, and the number of variables tried at each split (mtry) was valued at p/3 because we have a regression type of random forest. The node size was set to default, and the importance of predictors should be assessed. The cross-validation performance was used to tune the model, and used a separate test of 50% of the valid data to evaluate its accuracy.

   The random forest method could provide an intuitive plot to illustrate the importance of each variable affecting the amount of money spent on each kind of product. These plots could facilitate our analysis of sales for different products. At the same time, the method of random forest maintains the smallest MSE value, which implies it has the best performance among the random forest, linear regression, and ridge regression.

2. ***Results***

   Through the six variable importance plots, we can summarize the following table.

	Wines	Fruits	Meat Products	Fish Products	Sweet Products	Gold Products	Average Importance Level
Income	2	1	1	1	1	2	1.33
NumCatalogPurchases	1	2	2	2	2	1	1.67
NumStorePurchases	3	4	5	3	3	5	3.83
Dt_Customer	6	4	7	4	4	3	4.67
NumWebPurchase	4	8	1	9	7	4	5.50
NumWebVisitsMonth	5	7	4	6	5	9	6
Age	7	5	9	5	8	6	6.67
Recency	9	6	10	7	6	7	7.50
Kid	10	9	3	8	10	12	8.67
Education	8	10	12	10	9	8	9.50
NumDealsPurchases	11	11	8	11	11	10	10.33
Marital_Status	12	12	11	12	12	11	11.67
Complain	13	13	13	13	13	13	13

Table 9: Summary of Variables Importance Level

From Table 9, we can clearly see that income is the most important factor influencing the sales of each kind of product. Meanwhile, catalog shopping and in-store shopping are the two most important purchase methods for each category's sales. Then, the length of time consumers spends as members can also seriously affect their consumption.

For almost every category of product, the presence of complaints and marital status have a limited impact on consumption.

It is worth noting that the number of children in a family has a strong correlation with the consumption of meat products. What's more, differences in age groups also lead to changes in sales of fish and fruit.

3. ***Conclusion***

   The MSE value of the random forest is the smallest, which means that its performance is better than the other models, and we can predict the customers' consumption tendency more quickly through the results obtained by this method. Although this method is very effective in providing a ranking of the important factors that affect sales for different types of products, there is no access to generate a range that shows in detail how much each factor would affect sales.

4 Conclusion

With total spent and individual category consumption as different prediction targets and customers' personal information and purchase methods as predictors, our models and their results can help companies find breakthroughs in sales.

Based on the combination of all the above methods, we can provide the following suggestions to the company. First, from the customers' information, we can conclude that the high-income group is the main customer of this company; thus, the products can be selected according to their preferences. Then, from the purchase method, the catalog and in-store shopping methods are dominant, so for future sales breakthroughs, we can increase our investment in online shopping methods.

However, if the dataset provider could increase the clarity of the description of several variables, our analysis could be more detailed. For example, the unclear relationship among variables ‘AcceptedCmp1-5’ (customer accepted the offer in the 1st-5th campaign) and ‘Response’ (customer accepted the offer in the last campaign). Throughout checking the value of the dataset, we can find that the fifth campaign is not the last campaign, but the data provider didn’t give the total number of campaigns and acceptance of the rest of the campaign. Moreover, for the variables ‘Z CostContact’ and ‘Z Revenue’, we did not understand their use.

Each of the analysis methods we use has its own limitations. First, since the number of populations in K-means clustering is determined by the Elbow Method, changing the number of populations will give us very different results. Then, compared to the random forest, both linear regression and ridge regression have higher MSE. Furthermore, if we do not use set.seed function, we will generate different results by decision tree each time. The result will not be very representative. Finally, the random forest cannot predict the specific interval to be influenced by the change of each individual factor.

5 Reference

Data Source: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

jmb. (2016, November 18). Cross-validation for K-means clustering in R. Cross Validated. Retrieved April 13, 2022, from https://stats.stackexchange.com/questions/200254/cross-validation-for-k-means-clustering-in-r

Palani, G. (n.d.). Customer Behavior Analysis with R. Kaggle. Retrieved April 13, 2022, from https://www.kaggle.com/code/gomathimaliga/customer-behavior-analysis-with-r

Boehmke, B., & Greenwell, B. (2020, February 01). Hands-On Machine Learning with R. Retrieved from https://bradleyboehmke.github.io/HOML/

6 Appendix

Figure 9-14 are decision tree in the setting of set.seed(362).

The information in Table 9 is provided by the Variables Importance Plots from Figure 15-20 , which are obtained from random forest.

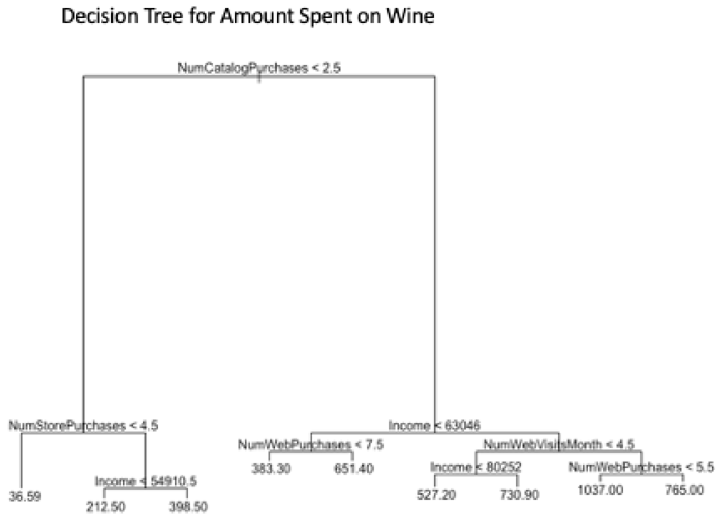

Figure 9: Decision Tree for Amount Spent on Wine

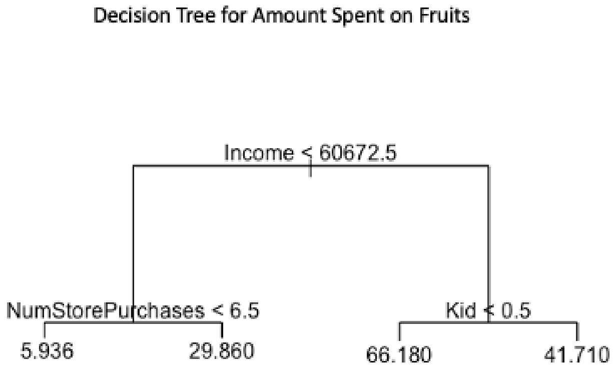

Figure 10: Decision Tree for Amount Spent on Fruits

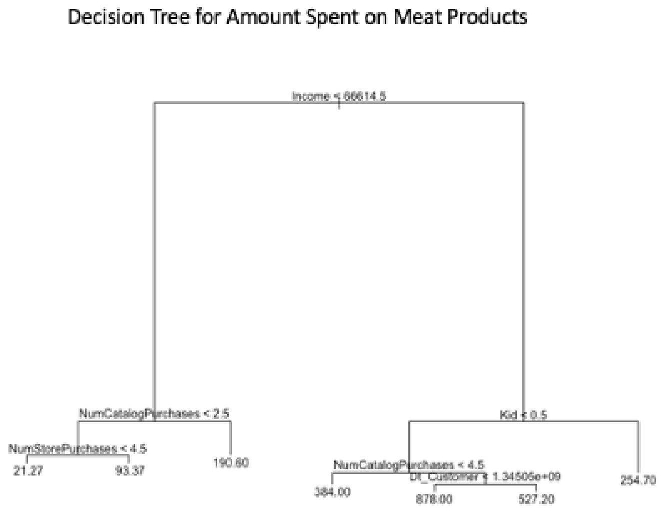
Decision Tree for Amount Spent on Meat Products
Income < 66614.5
NumCatalogPurchases < 2.5
NumStorePurchases < 4.5
21.27
93.37
190.60
Kid < 0.5
NumCatalogPurchases < 4.5
Dt_Customer < 1.34505e+09
384.00
878.00
527.20
254.70

Figure 11: Decision Tree for Amount Spent on Meat Products

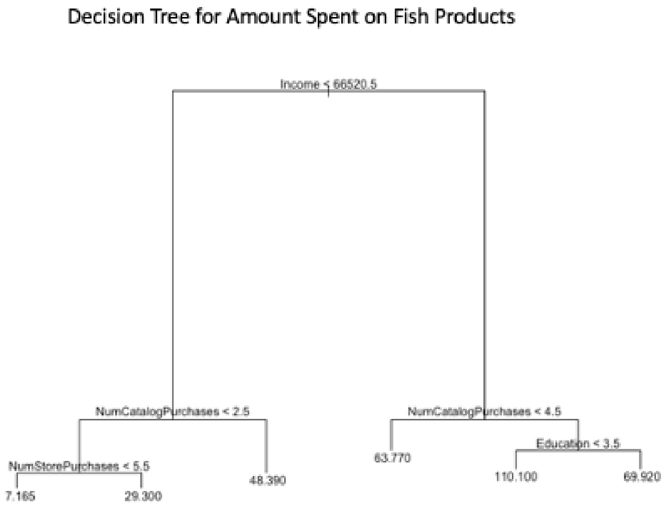
Decision Tree for Amount Spent on Fish Products
Income < 66520.5
NumCatalogPurchases < 2.5
NumStorePurchases < 5.5
7.165
29.300
48.390
NumCatalogPurchases < 4.5
63.770
Education < 3.5
110.100
69.920

Figure 12: Decision Tree for Amount Spent on Fish Products

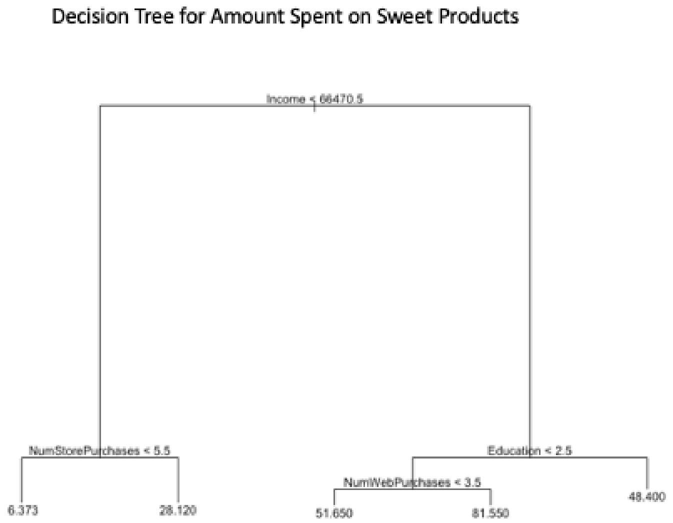

Figure 13: Decision Tree for Amount Spent on Sweet Products

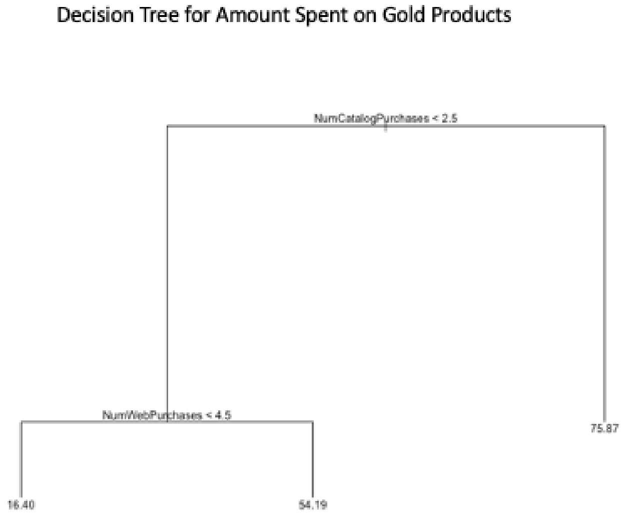
Decision Tree for Amount Spent on Gold Products
NumCatalogPurchases < 2.5
NumWebPurchases < 4.5
75.87
16.40
54.19

Figure 14: Decision Tree for Amount Spent on Gold Products

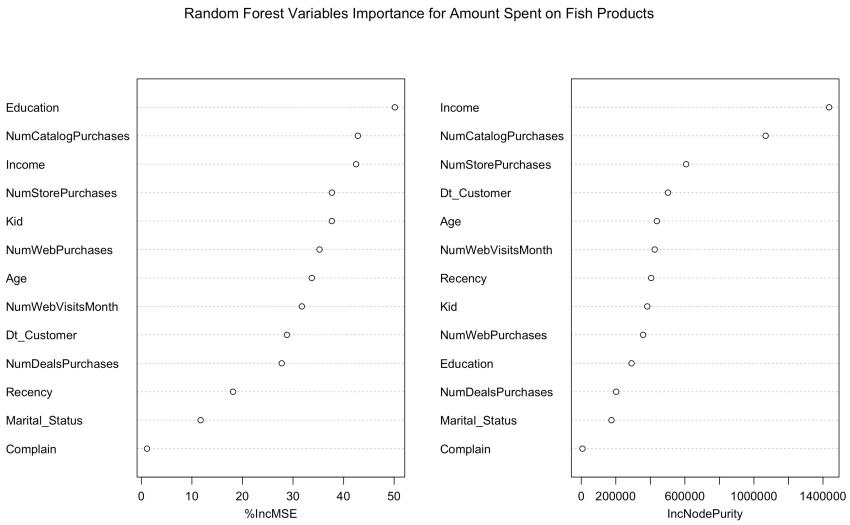

Figure 15: Random Forest for Amount Spent on Wine

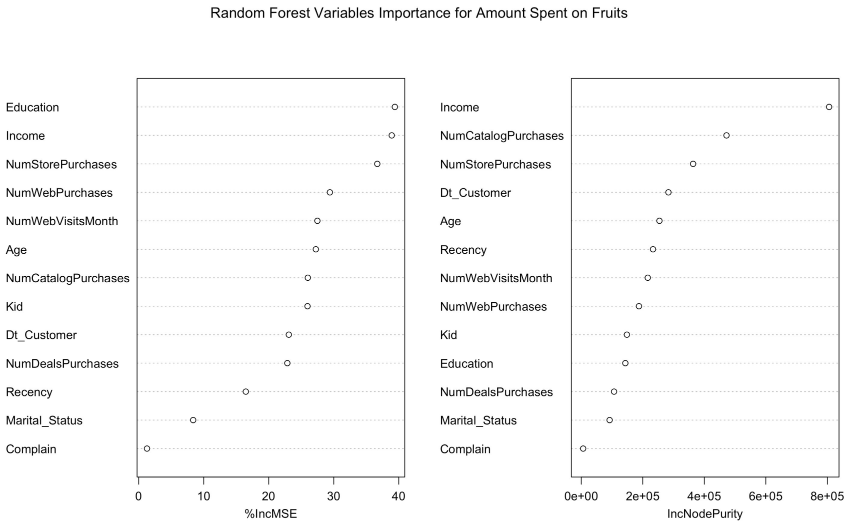

Figure 16: Random Forest for Amount Spent on Wine

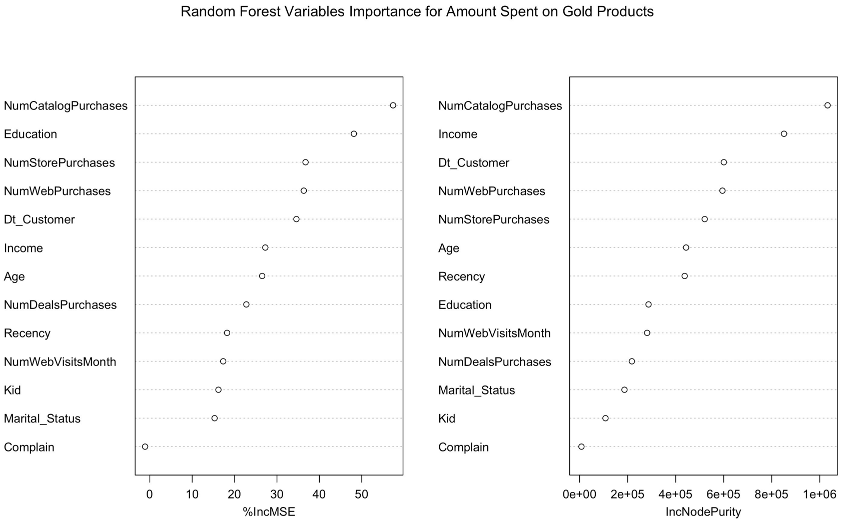

Figure 17: Random Forest for Amount Spent on Wine

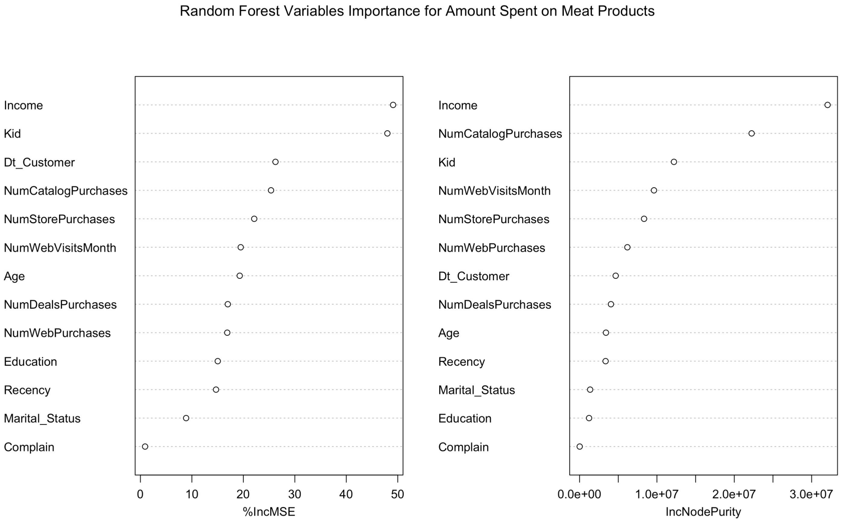

Figure 18: Random Forest for Amount Spent on Wine

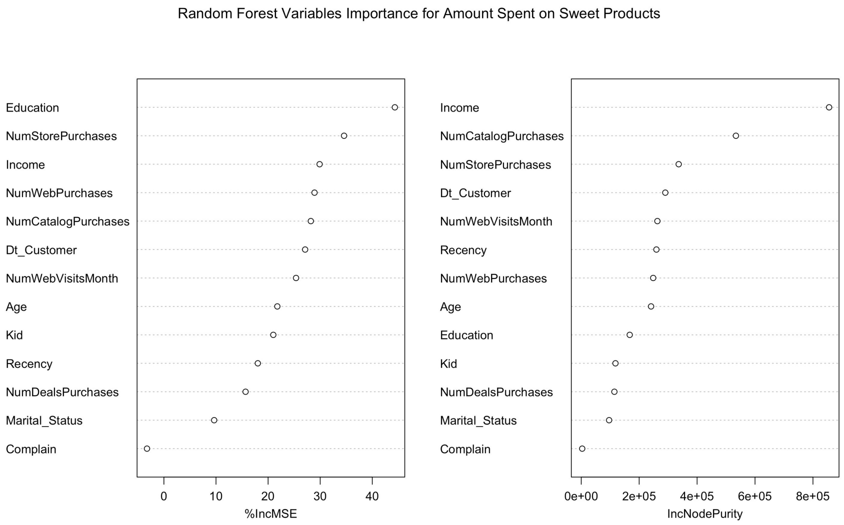

Figure 19: Random Forest for Amount Spent on Wine

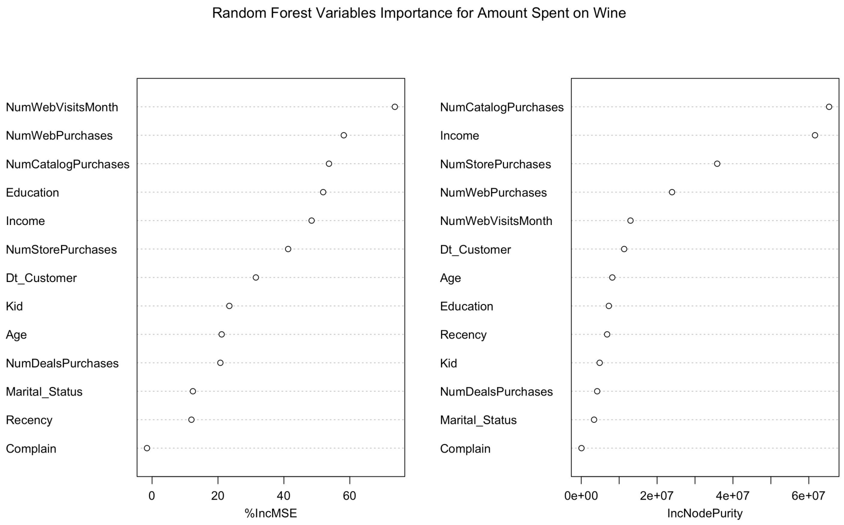

Figure 20: Random Forest for Amount Spent on Wine