K-Means Clustering for Customer Segmentation of a Winery

| Yuyan Chen | 186800650 |
| --- | --- |
| Evan Friedlan | 180922230 |
| Steven Ma | 207685580 |
| Cameron Peters | 180482780 |

December 15, 2021

BU425 - Salar Ghamat Table of Contents

| Business Problem: Segmenting Customers | 3 |
| --- | --- |
| Description of Data Set | 3 |
| Proposed Model: K-means clustering | 4 |
| Model Limitations | 5 |
| Formatting and Cleaning Data | 6 |
| Results of Clustering | 6 |
| Results Visualization | 7 |
| Recommendation | 8 |
| Segment 1 | 9 |
| Segment 2 | 10 |
| Conclusion | 11 |

*Business Problem: Segmenting Customers*

When organizations pursue a marketing campaign, it is necessary to segment customers in order to develop the correct campaigns most likely to reach those customers. The business that will be the topic of discussion is a winery experiencing these challenges. In order to ensure efficient targeting of promotions for new and existing products, this company must segment their customers with the given data. This opportunity revolves around more efficient marketing to increase sales. If the company is able to present customers with attractive sales or products to those consumers most likely to buy them, the result would be increased customer traffic and re-purchases. Another opportunity exists to ensure the organization tailors new products to the needs and wants identified by the different customer segments. Failing to segment customers results in wasted effort and capital directed towards customers who are unwilling to purchase the offering. This can in turn have more negative impacts on the organization both financially and internally.

*Description of Data Set*

The data provided allows for the segmenting of customers. From the data, customers who are very similar in their purchasing habits will be placed into the same segment. To begin, the data set provides a customer identifier (ID number) for each customer who purchases products from the organization. The demographic data, such as customer birth year, education levels, marital status, yearly household income, number of children, and number of teenagers in the household will allow us to separate customers with a greater understanding of their personal life. For example, young individuals who may be in school may not have the willingness to purchase more expensive items like married couples with higher income. The rest of the features: date the customer joined the company, the number of days since last purchase, and amount spent on wine represent purchasing habits which will further identify customers and place those more similar to one another in the same segment.

The following data for the company has been accessed on Kaggle, which is a trusted source for legitimate data. Although the database is very thorough, there are missing values in the income column. In order to solve this issue, the missing values were removed so that a model could be run. Instead of imputing, we removed missing and bad values because there were so few, and the removals did not significantly take away from the set.

To ensure the model runs efficiently, features like education, marital status, date of enrollment with the company were temporarily removed. This is because K-means clustering does not work well with categorical variables, therefore the clustering was based only on numeric columns. Lastly, to make the dataset fit with R-Studio, the csv file was read with separator “/t” so that the set would be converted to a database with rows and columns (it would not otherwise appear this way when automatically inputting it into R).

*Proposed Model: K-means clustering* TheK-means algorithm was chosen for customer segmentation. This is an unsupervised learning task, because customer data has not been labeled. Since customers need to be categorized into k groups, K-means is the most best method of clustering for the job.  
After choosing the range for k, the steps were as follows:

**Step 1**. Iterate through the range of integer k, the value “k” represents the number of clusters  
**Step 2**. Randomly select k distinct centroid (new data points as cluster initialization)  
**Step 3**. Measure the Euclidean distance between each point and the centroid  
**Step 4**. Assign the each point to the nearest cluster  
**Step 5**. Calculate the mean of each cluster as new centroid  
**Step 6**. Repeat step 3–5 with the new center of cluster until the algorithm converges or maximum numbers of iterations are reached  
**Step 7**. Calculate the variance of each cluster. K-means algorithm aims to choose centroids that minimise the inertia, or within-cluster sum-of-squares criterion. Since K-means clustering can’t “see” the best clustering, the only option is to keep track of these clusters, and their total variance, and do the whole thing over again with different starting points.  
**Step 8**. Repeat step 2-7 until get the lowest sum of variance

*Limitations of Chosen Model* Since the k-means model performs clustering based on a distance-based measure, the model does not handle categorical variables well. If categorical data is also nominal, then values need to be assigned to each category to represent similarities. However, if an ordering for categorical variables does not make sense, then they must be excluded from the algorithm. Additionally, the optimal k value is unknown; it is difficult to decide which integer to start with without background information. It is also possible that the optimal solution might not be found if the data set is very large and the algorithm does not converge. Whenever the dataset is updated, which includes deleting or changing existing data points and adding new data points, the whole process needs to be repeated again. There might be some abnormal data entries (which may be mistakes in recording) such as outliers that can affect the accuracy of the result — these data points must be addressed to ensure the validity of our model.

*Formatting and Cleaning the Data* From the original data, several steps were taken to clean the data to make it usable for the K-means clustering algorithm. Of the 29 variables/columns in the data, only Income had missing values, with 24 observations with NA values for Income. An initial attempt was to perform imputation on the data, however the resulting regression was a poor fit to the data, and so the rows with missing data were simply removed from the data frame for the analysis. This was considered a reasonable approach as the original data set contains 2240 observations; removing 24 observations should not have a significant impact on the results of the clustering model.  
 Since k-means uses a distance metric to find optimal clusters, only numeric data should be included in the clustering algorithm. Therefore, columns including Education, Martial\_Status, Kidhome, Teenhome, DT\_Customer, and all other binary zero-one columns were removed from the resulting data set. Additionally, the ID column was also dropped for the resulting analysis. Although ID is a numeric variable, it does not provide any information for clustering, and can impact the resulting clusters.  
 Finally, the k-means algorithm requires scaled data as the magnitude of certain variables will impact the results of distance measures and the resulting centroids (clusters). A common approach to scaling is normalization (subtracting by the mean, dividing by the standard deviation) and is the method used in scaling the data. However, other methods such as range scaling (subtracting by the min, dividing by the range) are also frequently used.

*Results of the Clustering Output* Using the scaled data, the k-means model was fit to the data, with k ranging from 2 to 10 clusters. The number of random starts was chosen to be 5 — although this choice was arbitrary, by increasing the number of random starts, the model can find better clusters which reduce the within cluster sum of squares. Using the fviz\_nbclust function, 2 clusters resulted in the highest average silhouette width (See Exhibit 1 for Silhouette score). The elbow plot (See Exhibit 1) also exhibits a significant decrease in total within sum of squares for 2 clusters.

Using 2 as the optimal number of clusters, the k-means model was fit to the data with 10 random starts. The R squared value for the 2 clusters is 0.3209, meaning that approximately 32.09 percent of total variability in the data is explained by the resulting centroids in the k-means model. Although the R squared value is quite low, this can be attributed to the fact that only 2 clusters are used to describe the entire data set — the R squared values will increase as the number of clusters increases. Based on the silhouette plot (see Exhibit 2) the average silhouette width is 0.33. Additionally, two cluster plots are used to help visualize the resulting clusters. It appears that the cluster represented by circles in both cluster plots is much more concentrated and centralized compared to the other cluster.

Since clustering requires scaled variables, the resulting centroids can be difficult to interpret and describe numerically. Since the k-means algorithm assigns each observation to a given cluster, one way to describe the two clusters is to create two new data sets with observations in each data frame separated by cluster. By doing so, numerical and graphical summaries can be given to describe and contrast each group. From a marketing perspective, these two groups form distinctive customer segments which can be targeted based on their unique preferences.

*Results Visualization*

With the results of the K-means clustering complete, analysis can be conducted on these two clusters to gain valuable insights from the data. The best way to do so is to make the results digestible and simple to examine, and by creating charts to visualize the data, some interesting patterns can be identified. With the data that had been separated into two sets by cluster, the mean of each variable was calculated and used to create a bar chart (Exhibit 4). The chart displays the average customer of the two clusters, creating a tangible customer prototype with which to target from a marketing perspective. The two clusters were quite distinct, with cluster 1 (C1) presenting the image of a lower-income customer who buys much less of each product category than cluster 2 (C2). The average income of C1 was approximately $39,000, whereas C2 nearly doubled that at just under $73,000. These measures are reflected in spending on each kind of product, with C2 spending 3 – 10 times C1’s amount in the various categories. For example, C1 spends $37 on meat products while C2 spends $365, on average.   
 There was also a notable difference in product preferences between the clusters. As shown in the pie charts of spending habits (Exhibit 6), C1 spends proportionally more on wines than C2. Instead of wine, C2 diverts their money to a significantly higher percentage of meat products and increases sweet products, fish products, and fruits. This could be due to C2’s higher income, allowing for additional products to be purchased alongside the wines. C1 may be shopping at the winery specifically for their flagship product, wine, whereas C2 can afford to make purchases of some accompanying items like meats, fish, or fruits. Surprisingly, aside from wine, the one area that C1 spends a higher proportion on is gold products.  
 Coming back to the bar plot, the shopping behaviours of the two segments is also of note. In terms of number of purchases, C2 has a higher value for all three of web, catalogue, and store. This is in line with the higher spending ability characteristic of this cluster, however the number of purchases made using a deal, or discount, is actually higher for C1. This, along with the higher number of web visits, demonstrates that C1 likes to find discounts and will only make a purchase if the price is right, which would make sense given their lower income.

*Recommendation*

Based on the results of the k-means clustering algorithm, the 2 clusters should form the basis of the customer segments for this company. Therefore, any potential marketing strategies should be focused on targeting these two groups. Since data used in this analysis is only related to purchasing behaviour, segmentation for these two groups are based on a behavioural segmentation basis only. In order to create effective marketing tactics, a thorough understanding of the unique buying behaviours for each segment is required.

*Segment 1* Segment 1 is a lower-income customer segment, with a mean annual income of approximately $39,000. This factors heavily into their spending habits, as they buy far less of every product type than Segment 2 (See Exhibit 4). Proportionally, however, Segment 1’s spending leads to some interesting conclusions. They seem to focus on a few products, specifically wine, meat products, and gold, and stick mainly to those areas. In fact, they spend even more on gold than the richer Segment 2, by percentage (Exhibit 6).  
 Demonstrated by their higher number of deals purchases and web visits per month, Segment 1 can be characterized as bargain hunters. They will repeatedly return to the website, even more than Segment 2 who purchase far more and in larger orders (Exhibit 4). The typical Segment 1 customer seems to make orders a few times per year, with just a few items in each. This is a potential area for improvement, as it would be beneficial to increase the frequency and size of Segment 1’s orders.

With their propensity to search for discounts, our client would be wise to offer Segment 1 a promotional campaign to boost sales. In 2014, the time of this case, digitization is becoming ever more important, so building an online presence is paramount. Segment 1 was found to most commonly do their shopping in-store, with web purchases coming in a close second (Exhibit 5). Segment 2 also shows store purchases to be their preferred method of shopping, so we can guide Segment 1 to the site to grow our online customer base without risking the main drivers of in-store profits. Another supporting factor is that Segment 1 already visits the website more often than Segment 2, so it makes sense to tap into this aspect of their behavior.

We recommend introducing one or multiple deals on website orders, which can include discount deals like 10% off, or perhaps free shipping on orders over a qualifying amount (ex. $50, $100, etc.). These promotions could be sent as coupon codes to Segment 1’s email, or home address by traditional mail. This would increase website traffic, as the deals would be exclusive to online orders, and influence customers to spend more to get a better deal. Given Segment 1’s tendency to spend on a few types of products, management might prefer to focus their efforts on boosting lacking departments, namely fruits and sweet products. A discount could be offered on these categories, which would result in more evenly distributed sales figures. They might instead wish to further increase the current top sellers, but additional data like profit margins and seasonality, for example, would be needed to make that decision.

*Segment 2*

Segment 2 is considered to be the store's most profitable customer segment. With a higher mean annual income of 72,500 dollars, customers in this segment are frequent purchasers: they consistently buy more of every product category compared to Segment 1, contributing to a majority of the store’s annual sales. This can be explained by higher disposable income, which allows for higher dollar amount purchases. Segment 2 also spends less on deal purchases, but not by a significant amount. The spending habits of this segment are similar to Segment 1; these customers tend to spend the majority of their store purchases on wine and meat products. On average, a typical customer in this segment spent 603 dollars in the past 2 years on wine, and spent on average 365 dollars on meat products. Comparatively, only 50 to 80 dollars were spent on fruits, fish products, sweet products, and gold products respectively. Finally, Segment 2 spends a roughly equal amount of time on in-store, online, and catalogue purchases.

Based on the description of Segment 2, various marketing strategies can be used to target this particular customer group. For instance, a customer loyalty program can be implemented which rewards shoppers that make frequent purchases using a rewards-based point system. By doing so, this company can further stimulate sales from their most valuable customer segment, while building brand loyalty and goodwill.

Another core focus would be increasing the promotional efforts through in-store, online, and catalogue retail channels to better reach these customers. In conjunction, a marketing campaign will also be deployed, which is oriented towards the buying habits of this segment. Since customers in this segment like to purchase wine and meat in larger amounts, two different approaches can be taken: the company can either decide to further increase the sales of their meat and wine products, or push the sales of their lesser bought categories, such as fruits, fish, sweets, and gold. For both scenarios, a possible strategy is to promote their premium brands to encourage sales, with the assumption that customers in this segment will prefer more expensive brands because of their high disposable income.

*Conclusion*

The essence of a successful marketing campaign is the target market that is being pursued. This company is now presented with 2 segments that they can focus on. The first segment is characterized by lower spending power and can be targeted via promotions focused on discount deals and savings programs. The second segment has considerably more disposable income and can be targeted with customer loyalty programs and product-oriented promotions. Our client can follow our suggested strategies, or use these segments to create other specific, customized marketing campaigns that will improve their financial performance and boost other key indicators that are of interest to the organization.

Appendix

Exhibit 1: Elbow Plot and Silhouette Scores by Cluster

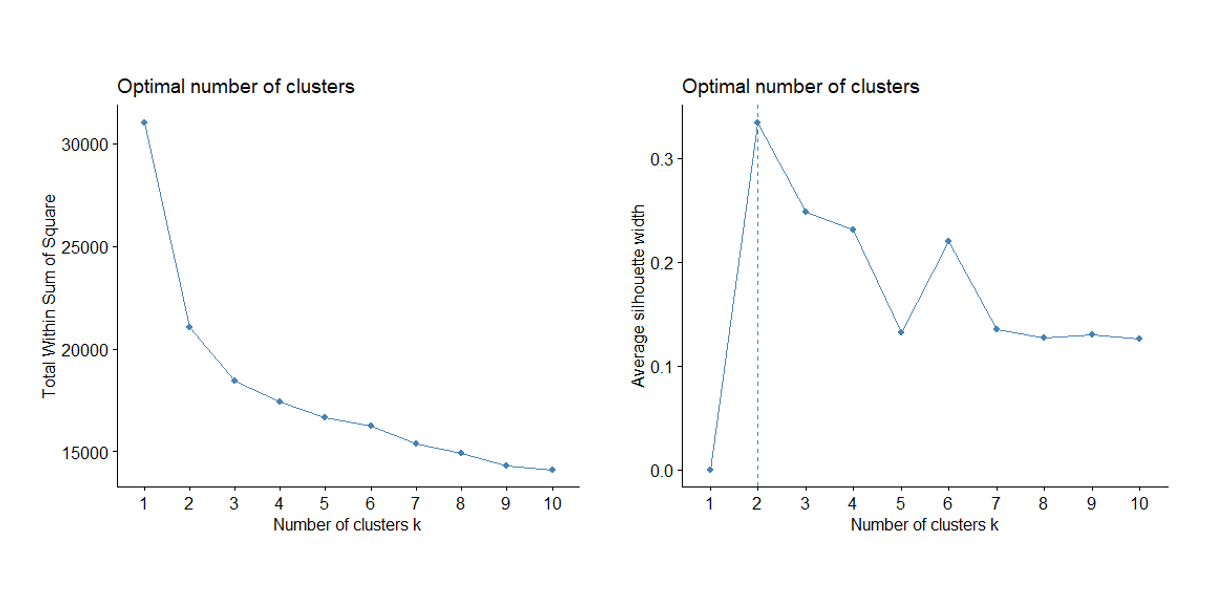


Exhibit 2: Silhouette Plot

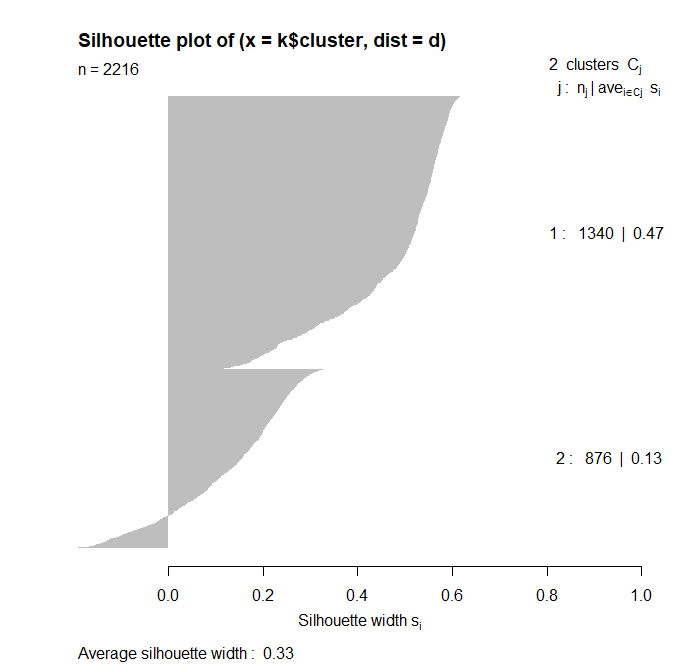


Exhibit 3: Clusterplots of K-means

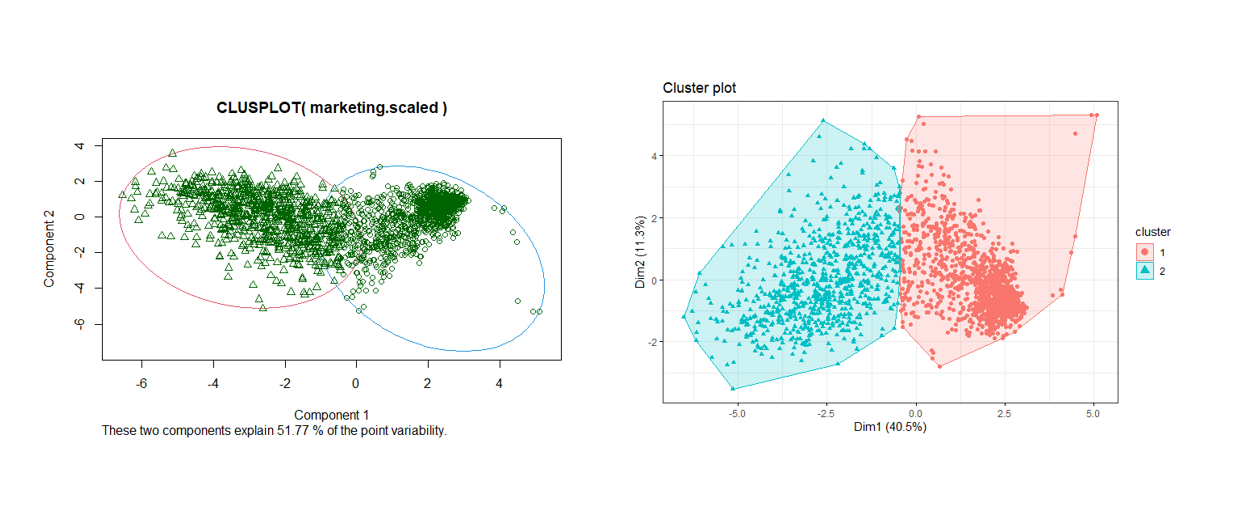


Exhibit 4: Barplot of Column means by Cluster

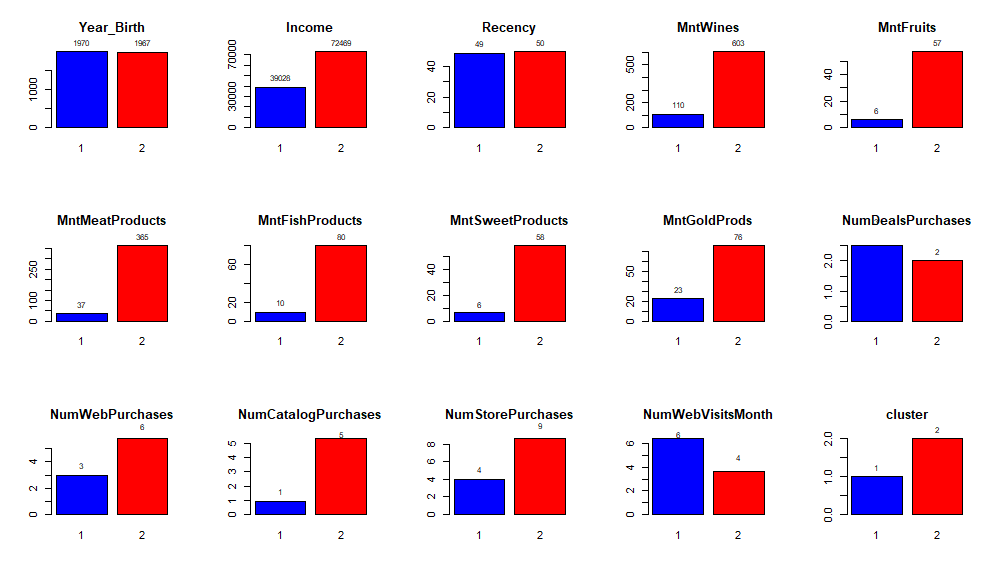


Exhibit 5: Purchasing Channels

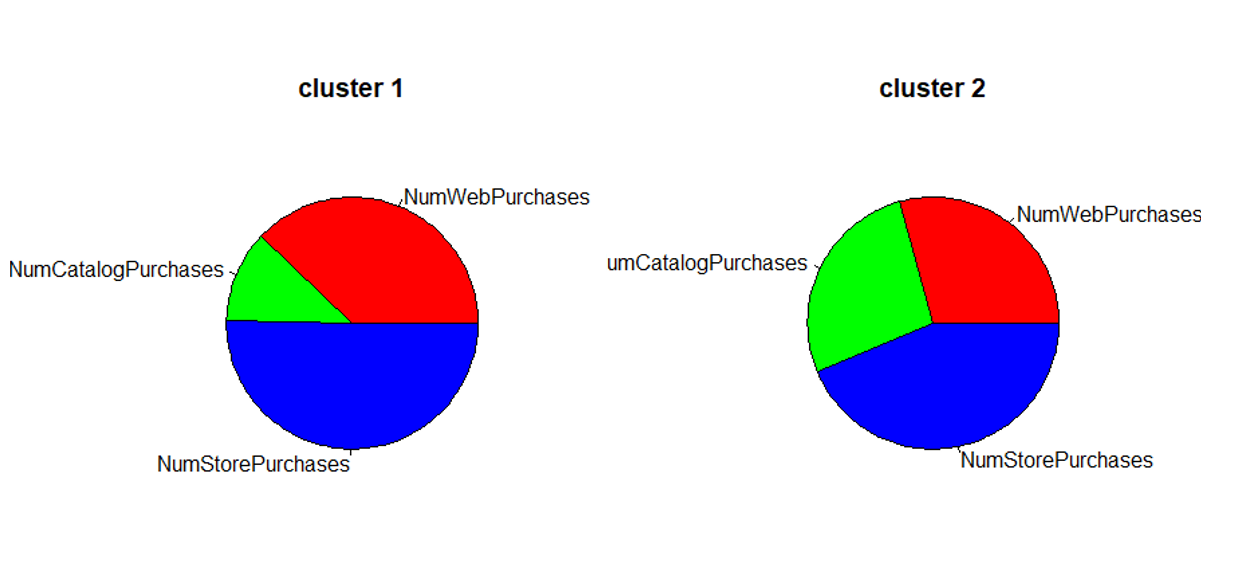
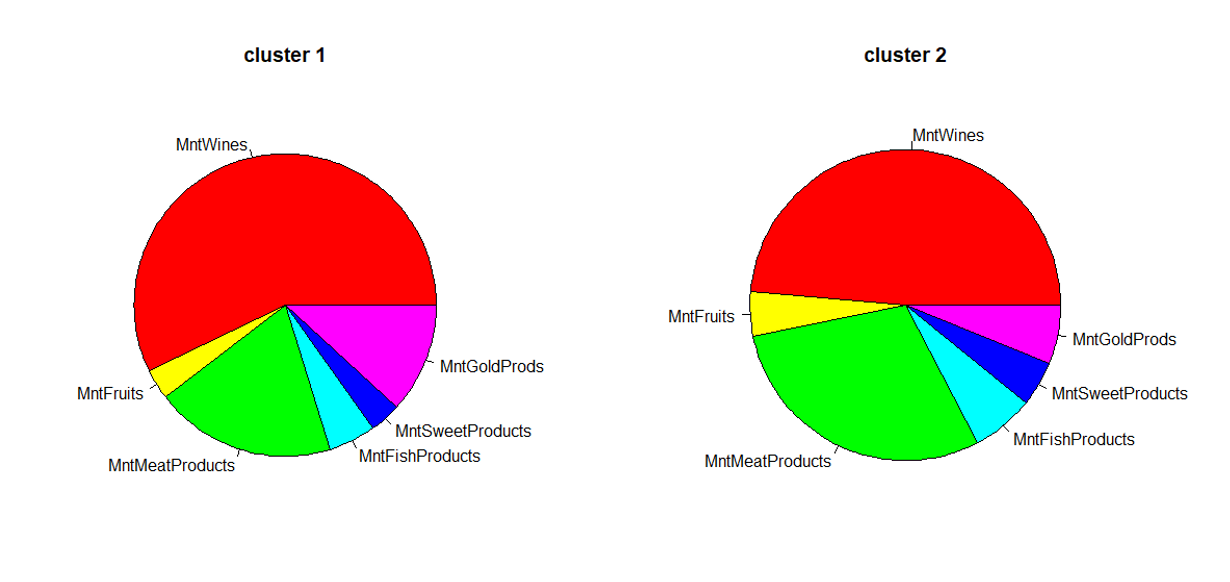


Exhibit 6: Sales by Product Category



Works Cited

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

https://medium.com/data-folks-indonesia/step-by-step-to-understanding-k-means-clustering-and-implementation-with-sklearn-b55803f519d6