Free Analytics Environment R Assignment 2

Clustering analysis report

Contents

Overview

In this report we will be performing clustering analysis using "wholesale" dataset. The dataset contains annual spending's of clients in different categories of goods. The aim is to partition the data into groups such that each partition conveys a unique characteristic of that group. The output of our analysis will group clients into different clusters with each cluster representing a unique quality.

Exploratory Data Analysis

First we search our data for null and incomplete values. For this dataset no incomplete values exist. Let's take a look at the variables more closely.

Channel & Region

The variable's channel and region are categorical variables. Channel takes on the values 1 and 2 while Region contains three unique values 1,2 and 3. 298 observations belong to Channel 1 while 142 belong to Channel 2. 77 observations belong to Region 1, 47 to Region 2 and 316 to Region 3.

Fresh

The values for this variable range from 3 to 112,151. The mean is 12,000 and the median is at 8,504. The mean is higher than the median which indicates that the distribution is right skewed. Let's take a look at the histogram to get a better understanding of the distribution.

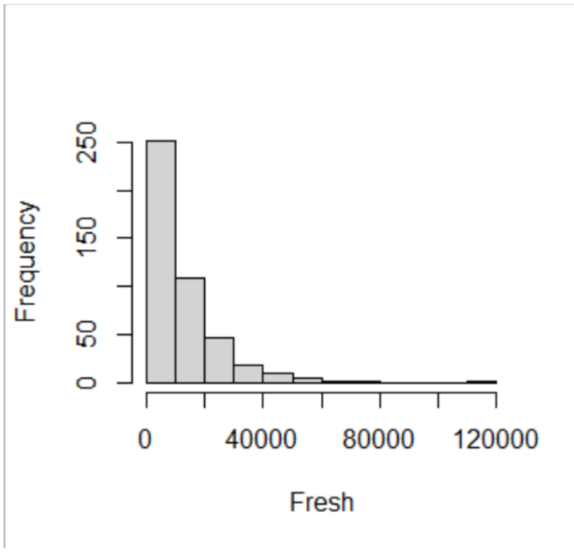


Figure 1

It is visible that the distribution is right skewed. There also seems to be some potential outliers present. We can make use of box plot to get more insightful information on the distribution.

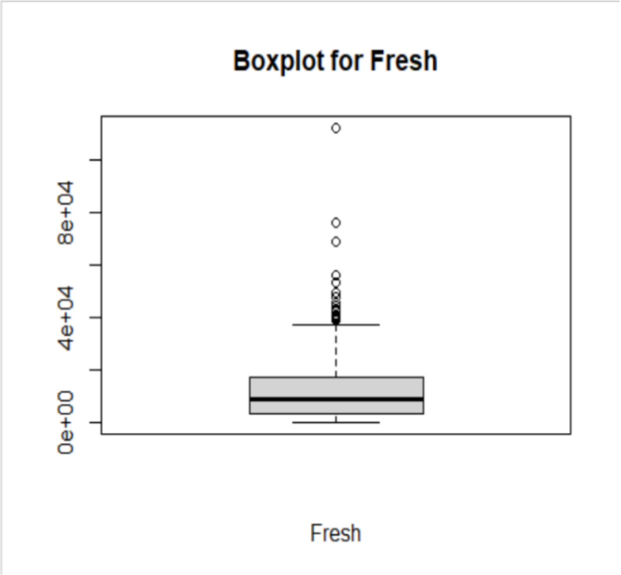


Figure 2

Figure 2 shows that a lot of potential outliers exist in the variable Fresh.

Milk

Moving on we look at the spending that clients have made on milk. The value ranges from 55 to 73,498. The median is at 3,627 while the mean is 5,796. Lets take a look at the distribution through a histogram.

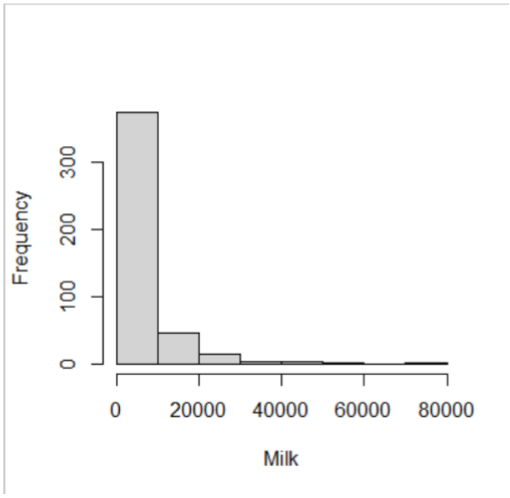


Figure 3

The distribution is right skewed as expected with some potential outliers again. Lets take a look at the box plot now to better understand them.

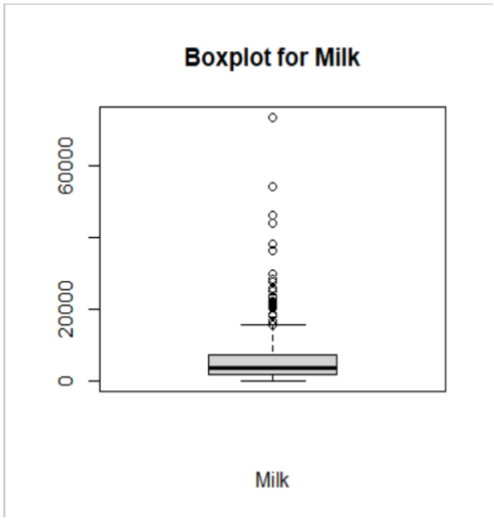


Figure 4

The data is centered around the median value, mostly above the median. The distribution is strongly right skewed with a lot of potential outliers.

Grocery

The spending that clients made on groceries ranges from 3 to 92,780. The mean of the distribution is at 7,951 while the median is 4,756. Just as the other two variables this distribution also seems to be right skewed. Lets take a look at the histogram for this variable.

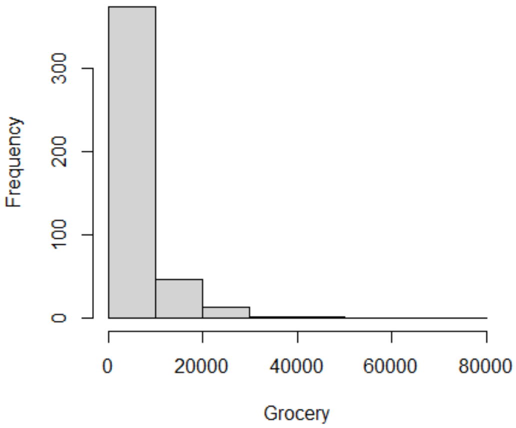


Figure 5

It can be seen that distribution is unimodal and concentrated between the mean and the median. It can be seen that most frequently clients spend between 0 to 10,000 on groceries. The box plot will shed more light on the outliers and their distance from the center of the distribution.

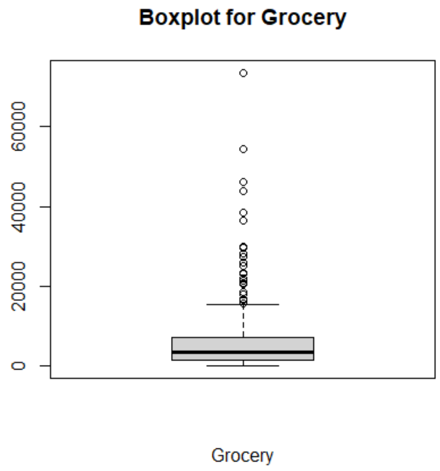


Figure 6

Frozen

The spending on Frozen category ranges from 25 to 60,869 with a mean value of 3,071 and median 1,526. Again the mean is higher than the median which means that the distribution is concentrated between the median and the mean. The following histogram depicts that :

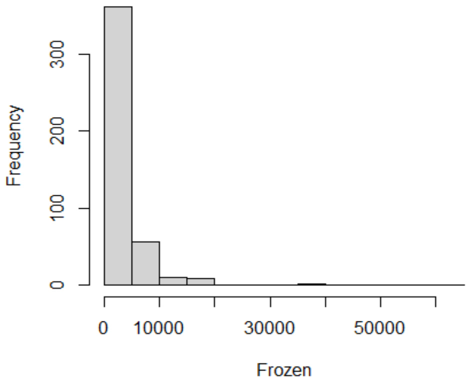


Figure 7

We can skip the box plot for this variable as it will be quite similar to the previous variables and we have extracted the useful information regarding potential outliers from the histogram.

Detergents_Paper

The distribution for detergents paper ranges from 3 to 40,827. The mean is at 2,881 while the median is lower compared to the mean at 816. The distribution we visualized looks similar to the previous distributions :

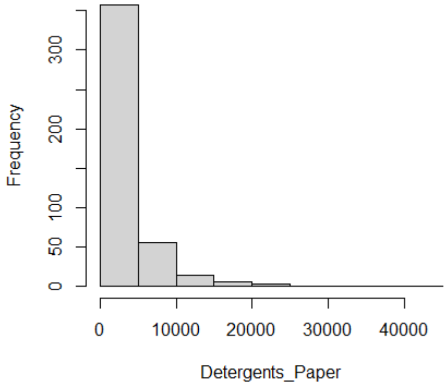


Figure 8

Delicassen

For delicassen the distribution ranges from 3 to 47, 943. The mean is 1,525 and median is at 965.5. The distribution is visualized again using a histogram:

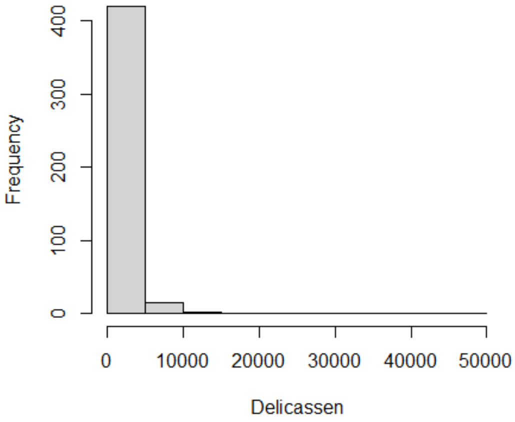


Figure 9

Looking at the histogram the distribution is heavily concentrated between 0 and 5000. It can be seen the maximum value is far away from the center of the distribution.

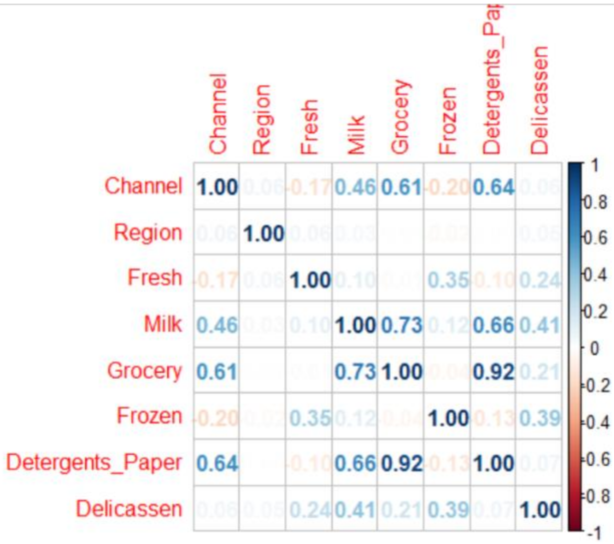
Now that we have an idea about the individual variables we can point out some similarities. All the non categorical variables have distributions with higher mean values compared to the median. It shows that only a handful of clients spend astounding amounts of money since the number of outliers is low but the values of these outliers are quite high. The range of the distributions are quite wide as depicted by the box plots. We can now take a look at the total spending by clients in each category :

Categories of goods	Total spending
Fresh	5,280,131
Milk	2,550,357
Grocery	3,498,562
Frozen	1,351,650
Detergents_Paper	1,267,857
Delicassen	670,943

It can be seen that the clients spent most money on the Fresh category while the least amount of money was spent on Delicassen.

Correlation analysis

Since our dataset is entirely numeric we can run correlation analysis on the entire dataset. We plot the correlation using corrplot library in R and the result is as follows:



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Pa	Delicassen
Channel	1.00	0.06	-0.17	0.46	0.61	-0.20	0.64	0.06
Region	0.06	1.00	0.06	0.03		0.02		0.05
Fresh	-0.17	0.06	1.00	0.10	0.01	0.35	-0.10	0.24
Milk	0.46	0.03	0.10	1.00	0.73	0.12	0.66	0.41
Grocery	0.61		0.01	0.73	1.00	-0.04	0.92	0.21
Frozen	-0.20	0.02	0.35	0.12	-0.04	1.00	-0.13	0.39
Detergents_Paper	0.64		-0.10	0.66	0.92	-0.13	1.00	0.07
Delicassen	0.06	0.05	0.24	0.41	0.21	0.39	0.07	1.00

Figure 10

Figure 12 shows correlation between each variable and it is color coded to show strong positive correlation with a dark blue color while it shows strong negative correlation in dark red. The colors fade to white as the correlation approaches 0 indicating no correlation.

We start the correlation analysis with variables with the highest correlation. Detergents_Paper and Grocery are highly positively correlated with a correlation value of 0.92. Which indicates that as the value of Grocery increases the value of Detergents_Paper increases as well in the positive direction and vice versa. We plot a scatter plot to in order to depict this.

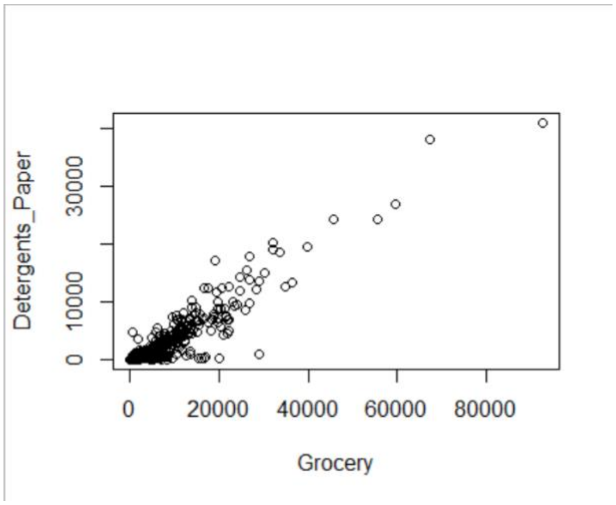


Figure 11

The next highest correlation value is 0.73, between Milk and Groceries. When we plot the two variables now it is visible that the relation between these variables is not as positively correlated.

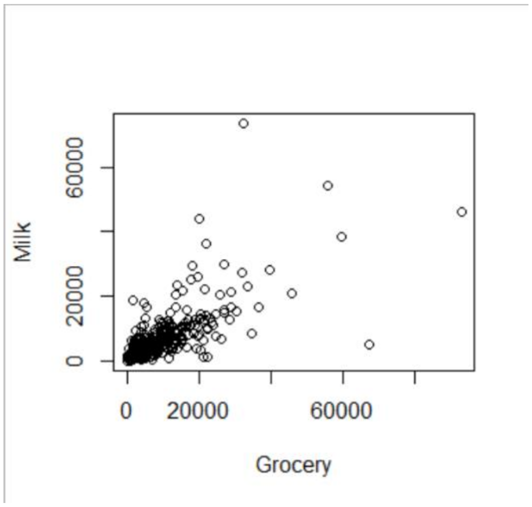


Figure 12

Moving on the next highest correlation values are not significantly high enough to depict any sort of a relationship between the variables. We conclude our correlation analysis with the fact that only Groceries and Detergent Paper are linearly correlated.

Normalization

Normalization is an important data preprocessing step in the context of clustering. As we saw in the previous section the values the non-categorical variables range from very small to very large values. This range of values can disrupt the quality of the clusters that are produced according to our algorithm. Therefore normalization is used to standardize all attributes of the dataset and eliminate redundant data and ensure good cluster qualities. Since the range of values is shrinked hence the efficiency of the algorithm also increases.

For our case we will be performing Min-Max normalization which performs a linear transformation on each column of the data. For each feature in a specific column the minimum value of the feature gets transformed into a 0 while the maximum value gets transformed into a 1. Every other value in between gets transferred into a decimal between 0 and 1 depending on how close the actual value is to the minimum and to the maximum.

We perform Min-Max normalization on all non-categorical variables since they have a similar scale. ”Channel” and “Region” do not require normalization and will not be a part of our clustering analysis since they are categorical variables with 3 unique values. Since K-means algorithm operates on least sum of squares of the Euclidian distances the distance calculation between the categorical variables are meaningless.

Choosing the number of clusters

Choosing the right number of clusters is a crucial step for efficient outputs from our analysis. We will be using the K means algorithm in order to group our data into clusters. Selecting optimal number of clusters is important since selecting few clusters may lead to many long distances of features to the centroid while selecting too many clusters leads to an increased number of centroids. Large number of clusters is not preferable to identify simple similarities to interpret.

There are a number of ways to identify the optimal number of clusters. Let's review and perform these methods on our dataset.

Elbow method

The elbow method is visually used to determine the suitable number of clusters. It makes use of the within clusters sum of squares (WCSS). WCSS is the measure of variability of the observations within each cluster. The more the clusters the smaller the within sum of squares, hence WSS is monotonically decreasing. Let visualize the elbow method for our dataset for up to 10 clusters.

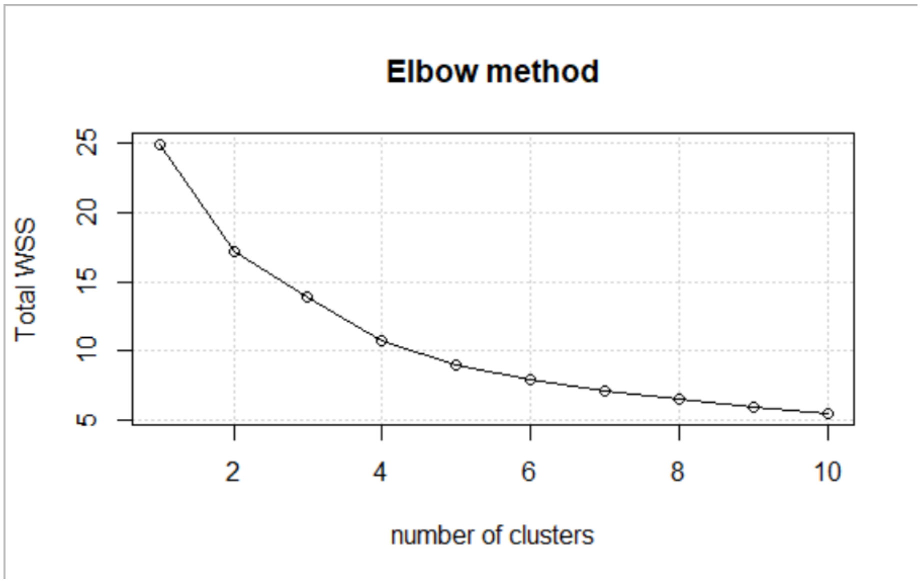


Figure 13

In order to determine the number of clusters we notice the change in WSS when an additional cluster is added from a considerable improvement to a much smaller improvement.

Looking at the elbow method line plot the optimal number of clusters is not easily determinable. It can be either 2, 3 or 4. The shift from considerable change to a smaller improvement is most prominent at 2 clusters. Let's take a look at the other methods to resolve our confusion.

Silhouette method

The silhouette method plots the number of clusters against a silhouette coefficient which is between -1 and 1. A score closer to 1 means that the data points are very compact within the cluster and far from other clusters. The visualization of the silhouette method on our dataset is as follows:

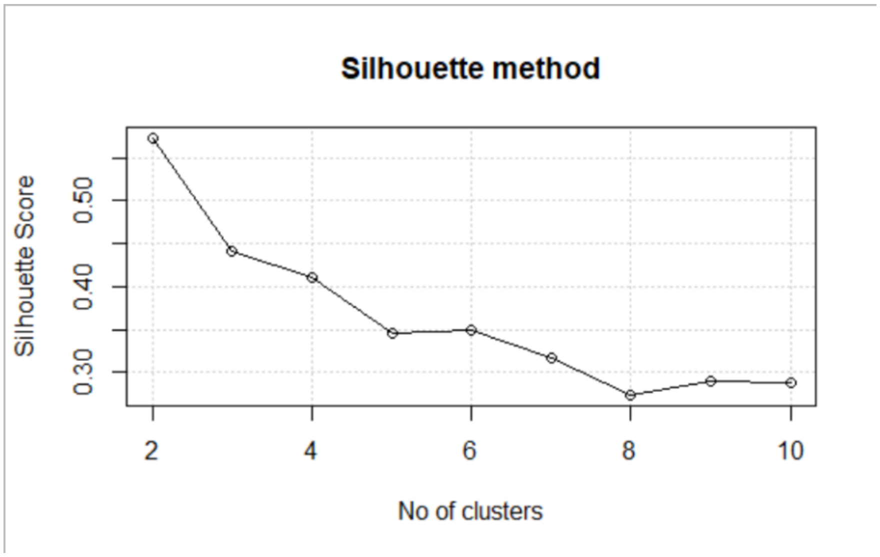


Figure 14

The figure shows 2 to be the optimal number of clusters according to the Silhouette method.

Calinski-Harabasz Index

The Calinski-Harabasz Index measures the ratio between the Between-Group-Sum-of-Squares (BGSS) and Within-Group-Sum-of-Squares (WGSS). In general the higher the Calinski-Harabasz score the better the performance. If on the plot of Calinski-Harabasz value against number of clusters there is a prominent peak or an abrupt elbow we can choose it. It can be seen that the graph has a peak at 2 clusters in the following figure:

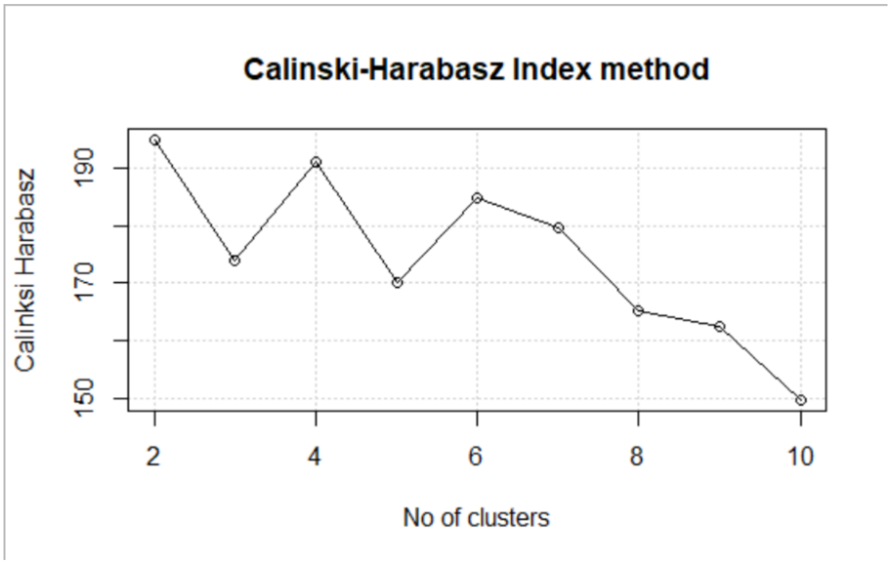


Figure 15

Gap Statistics

Gap statistics uses the idea of comparison between the actual data and reference uniformly distributed data. For this report we plot the simplified gap statistic against number of clusters.

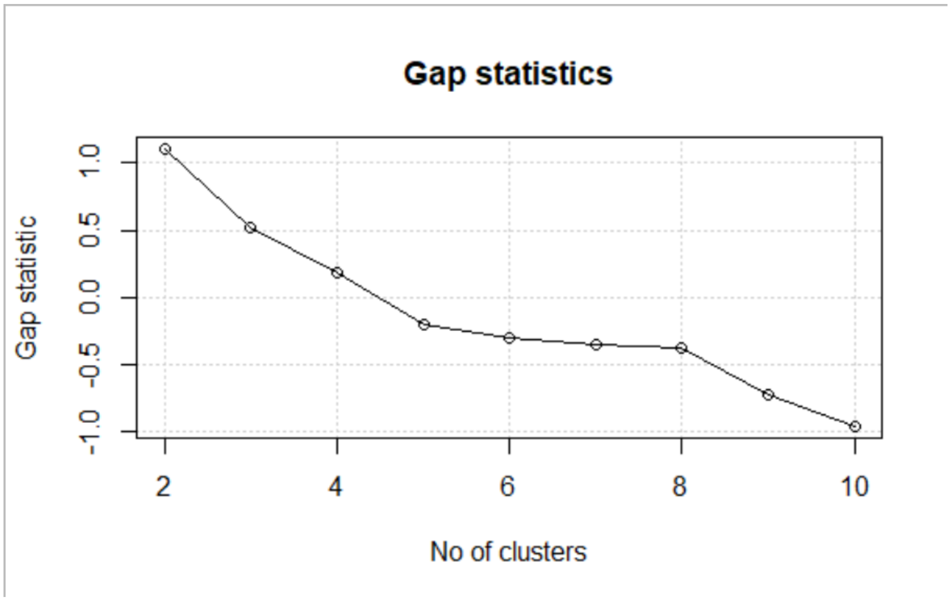


Figure 16

The cluster value with the highest gap statistic is 2, which indicates that 2 is the optimal number of clusters for our data set according to gap statistics.

After examing all the methods the joint consensus for the number of clusters is 2. Now that we have the optimal number of clusters we move on to partition our dataset into 2 clusters using a clustering algorithm.

Clustering

K-means

We will be using K-means algorithm which is an unsupervised machine learning algorithm used to partition the data into a given number of clusters. We start by defining k number of clusters, the algorithm then initializes the centroids of the k clusters at random. Then it assigns each observation to the closest centroid by calculating least squared Euclidean distance between centroids and data points. After the assignment step the mean is updated for the new clusters. Both the assignment and update step are repeated until convergence or until maximum iteration is reached.

When running the k-means algorithm in R we set the Nstart parameter to 25. The Nstart parameter makes the algorithm go through multiple configurations and chooses the best one. The configuration refers to variation between Euclidian distance between the centroid which is randomly selected initially, and the points in a cluster.

Selection of the initial centroids is an important step as a wrong estimate can lead to misleading clusters. Therefore setting the Nstart value is important in order to prevent our model from ending up in a local minimum situation.

After running the algorithm on our dataset we look at some information on the means of each variable grouped by cluster.

Variable	**Mean Cluster 1**	**Mean Cluster 2**
Fresh	0.108	0.996
Milk	0.056	0.269
Grocery	0.060	0.304
Frozen	0.050	0.048
Detergents_Paper	0.041	0.324
Delicassen	0.027	0.070

We notice that variables “Grocery”, “Detergents_Paper” and “Milk” have higher differences in the means between the two clusters. We can assume that these variables are responsible for this grouping. Let’s take a look at a few scatter plots involving these variables color coded by the clusters.

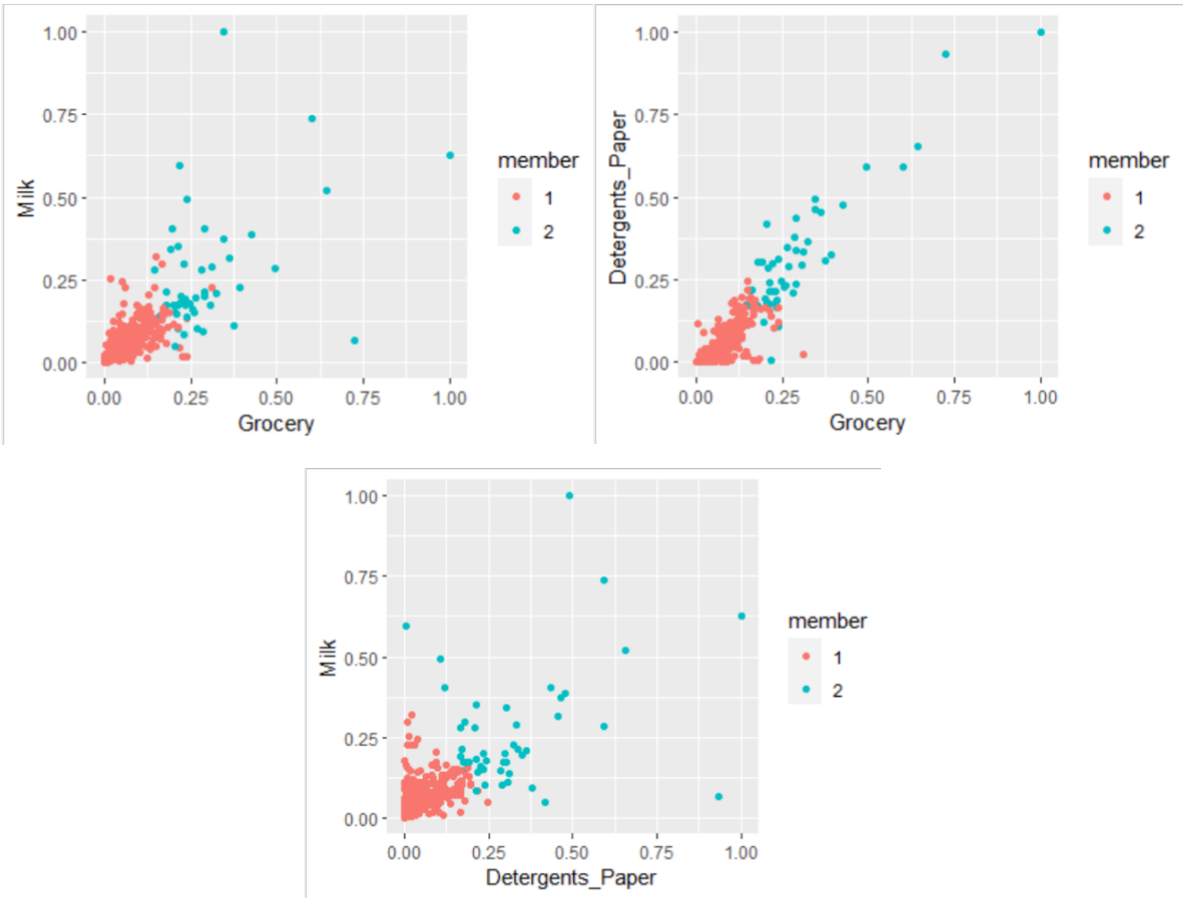


Figure 17

We can investigate the variables individually with respect to each cluster using the initial non-normalized data with box plots.

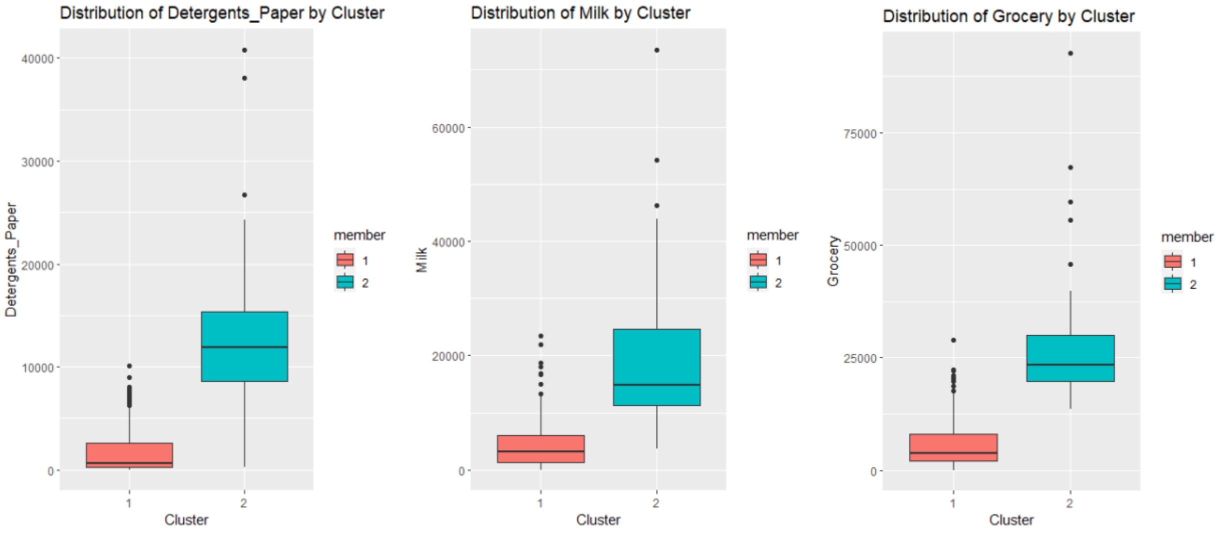


Figure 18

The box plot shows there is overlapping between the clusters which means that the clusters have high intra class similarities. The total number of observations was 440. After clustering 394 observations were assigned to cluster 1 while 46 observations are assigned to cluster 2.

Conclusion

We started our analysis with a diverse set of variables scaling from very small to very larger values. After finding the correct number of clusters and performing the K-means algorithm on our dataset we were able to get some useful insights.

The output of our analysis shows two different groups of clients based on their spending. One group spends lower amounts of money annually while the other group spends a lot more. The differences in spending are more visible for milk, grocery and detergent papers. For these products the super market chain can devise separate supply chain schemes in order to fulfill their client’s needs more efficiently. They can create different costing models for clients who spend more annually in order to sustain high paying clients in the long run. The clustering also shows that the number of clients that spend lower amounts of money annually are relatively high in number. The company can manage their budget accordingly to reach the larger set of customers with targeted advertisements. Marketing schemes can also be targeted according to region and channel depending on the partition of each cluster within a Region or Channel with the help of bar graphs.

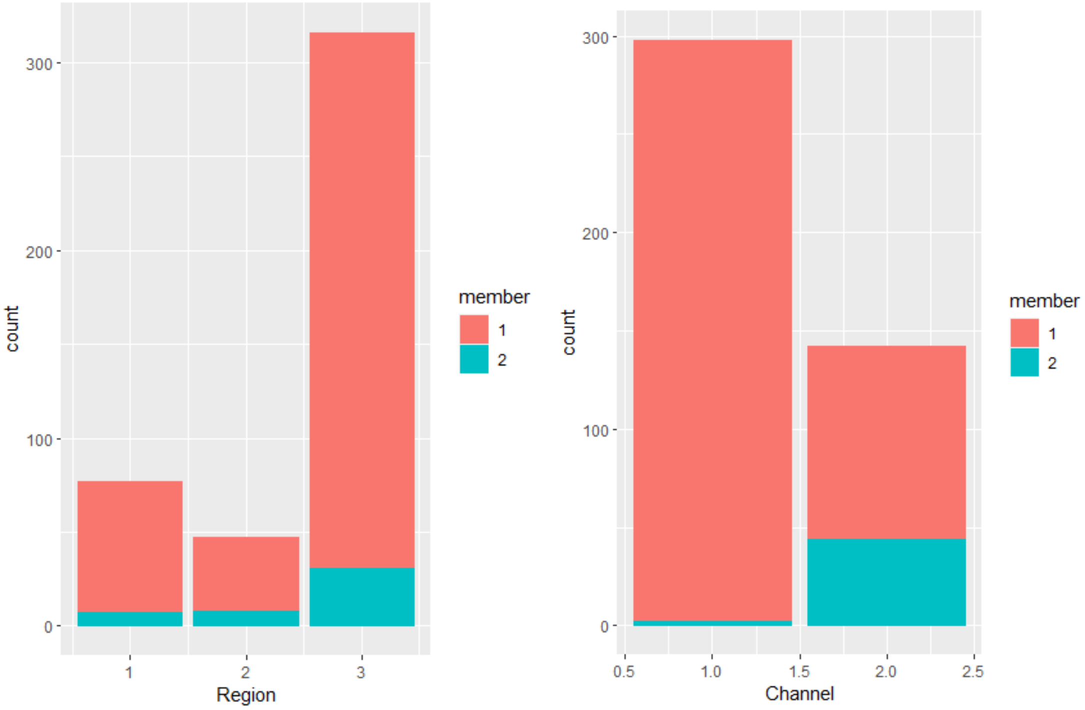


Figure 19

The number of clients belonging to cluster 2 are higher in Channel 1 as compared to Channel 2. Therefore the marketing team can focus their marketing schemes for the high paying clients on channel 2. They can combine this information with the information gathered from the Region bar graph and target the separate marketing scheme designed for high paying clients to clients from Region 3 and Channel 2.