|  |  |
| --- | --- |
| Unsupervised Learning project  Cincinnati Zoo | Abstract  An implementation of Clustering methods and Association rules  Deepa Das, Om Prakash  Data Mining for BI |

Contents

[Data Introduction: 1](#_Toc6318716)

[Data Reading and Preparation: 2](#_Toc6318717)

[Exploratory Data Analysis: 3](#_Toc6318718)

[Modeling Techniques: 7](#_Toc6318719)

[Clustering: 7](#_Toc6318720)

[K-means Clustering: 7](#_Toc6318721)

[K means with k=2: 12](#_Toc6318722)

[Determining optimal number of clusters: 15](#_Toc6318723)

[Hierarchical Clustering: 18](#_Toc6318724)

[Association Rules: 22](#_Toc6318725)

[Data Exploration: 22](#_Toc6318726)

[Apriori Model: 26](#_Toc6318727)

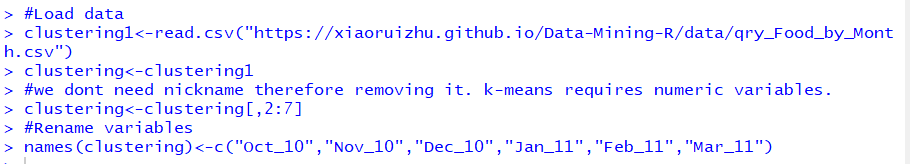
[Improving model performance: 28](#_Toc6318728)

[Conclusion: 33](#_Toc6318729)

# Data Introduction:

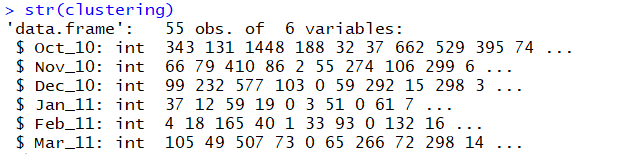
The Cincinnati Zoo was founded in 1873 and officially opened in 1875. It is the second oldest in the nation after Pennsylvania Zoo and serves over a million visitors each day. The goal of this analysis is to identify useful and/or hidden information in the data collected by the zoo and to study the buying and/or visiting behavior of zoo members. The data comes from 2 files - one with aggregated sales data of food items across the months Oct’10, Nov’10, Dec’10, Jan’11, Feb’12, and Mar’12 and another one with individual transaction data of customer purchase behavior across the same months and food items.

# Data Reading and Preparation:

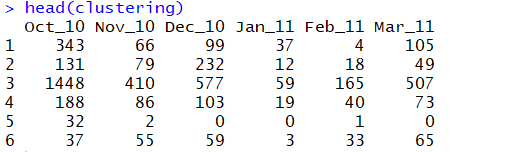


We don’t need nickname therefore we removed this column as k-means requires numeric variables and renamed the remaining columns for ease.

Let’s check the class of each variables.



A sample data looks like below:



There are 55 rows and 6 columns in this data frame as below:



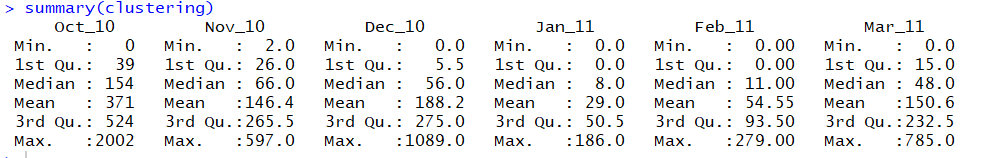
To check if there are any missing values in the data, we use the below function which shows there are no missing values.



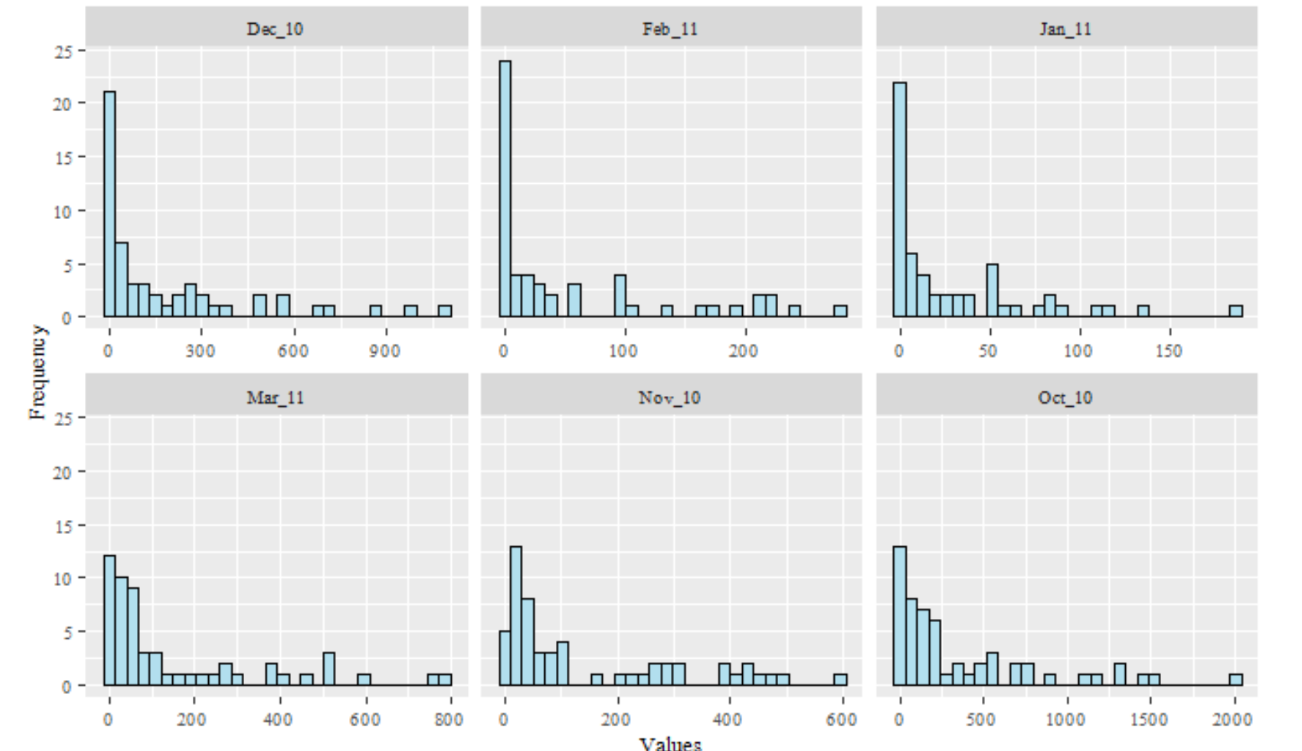
# Exploratory Data Analysis:

Now, we try to investigate each of the variables using summary and different plots.

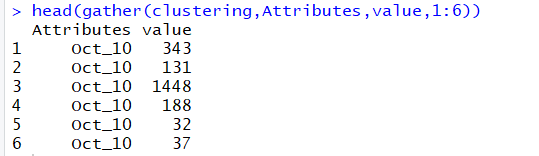
The summary of the dataset is below:



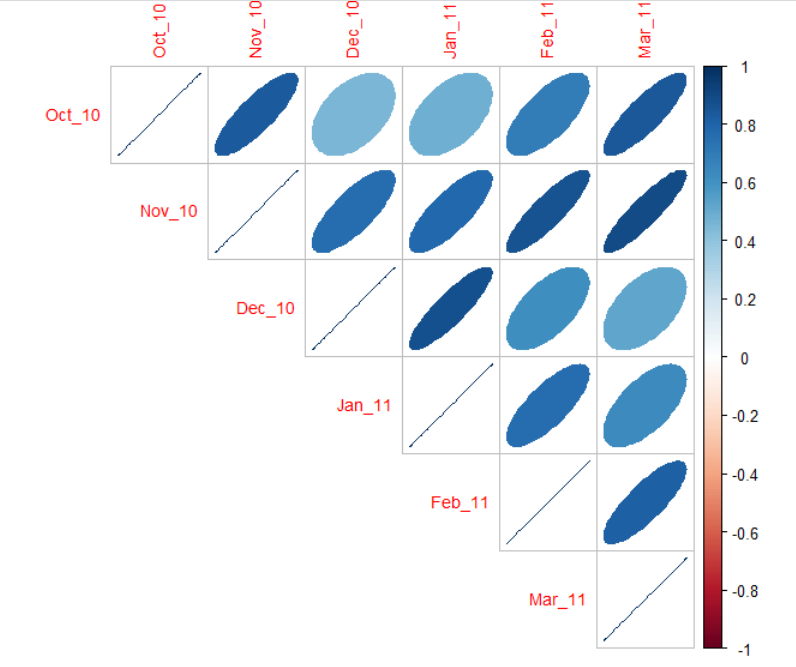
We plot histograms for each variables and combine them into a single plot to do a comparative study of these variables.



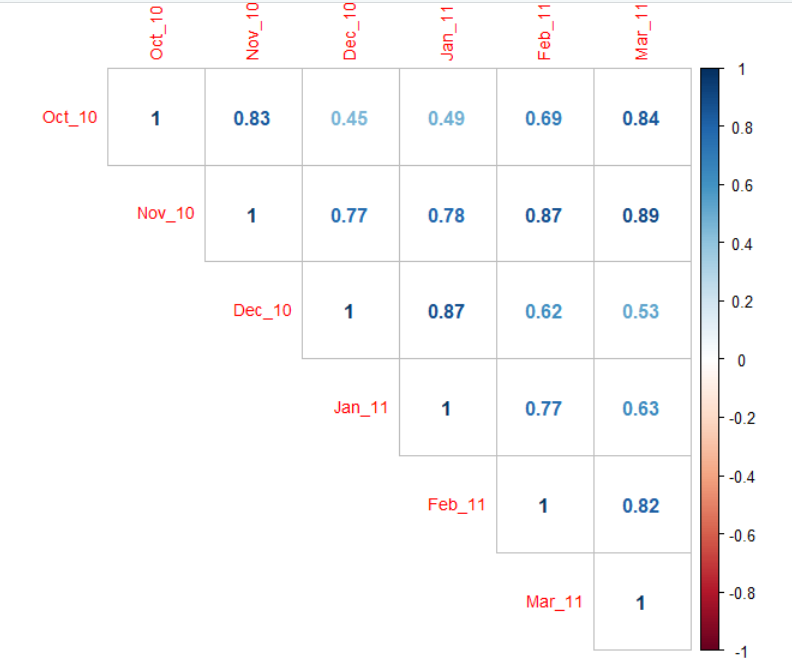
Let’s check the attributes and the values of the dataframe.



Next, we will plot a correlation plot using ggplot and corrplot function as shown below:



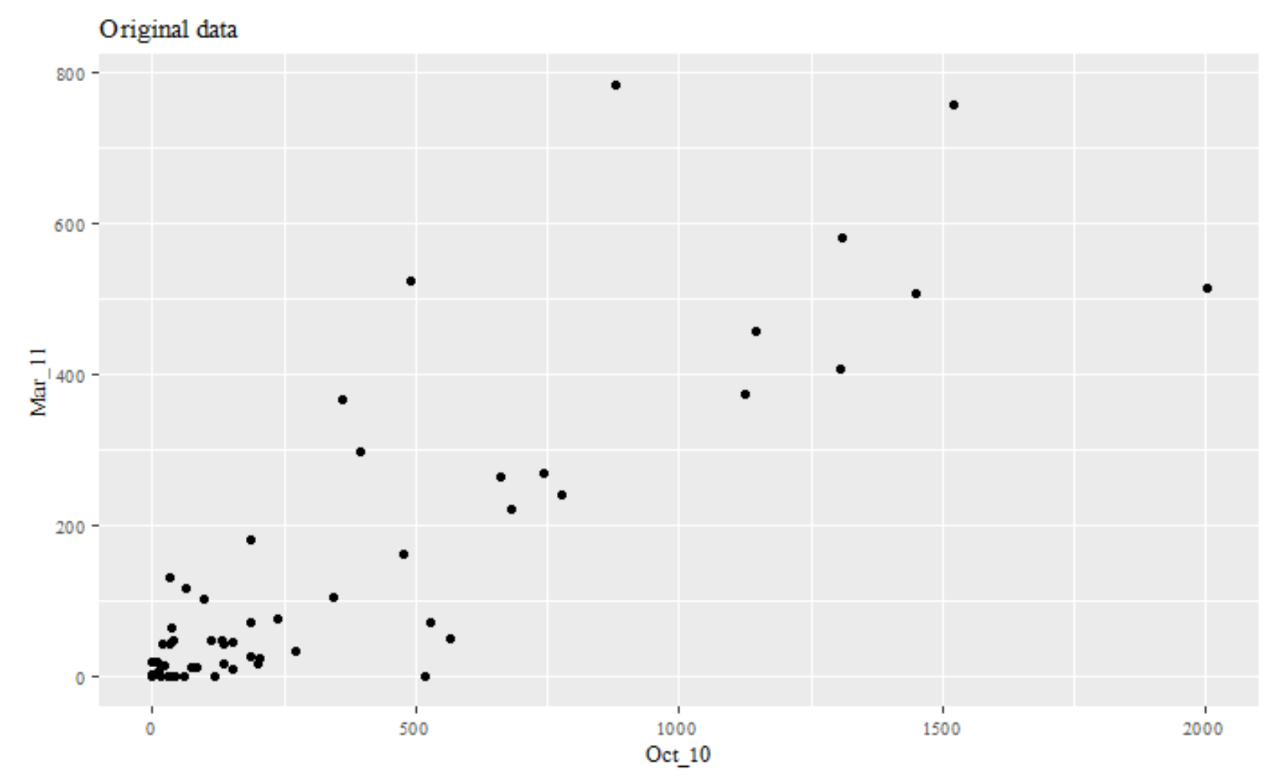
We see that transactions in Oct 10 has high correlation with Nov 10 and March 11. Feb 11 has high correlation with March 11. Dec 10 has high correlation with Jan 11. Nov 10 has high correlation with Feb 11 and March 11.

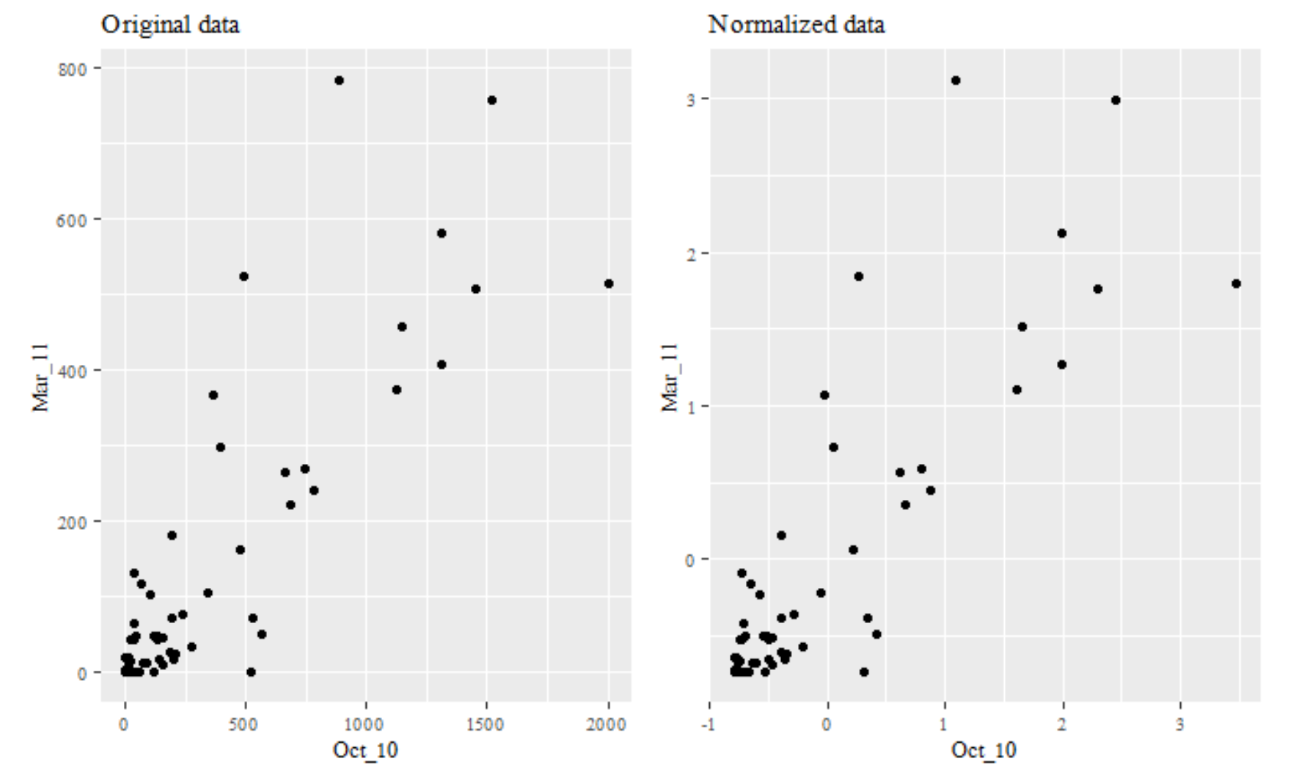


We now normalize our data using scale function and use this new scaled data to perform our modeling.



Let’s plot a scatter plot of the original data and normalized data.





# Modeling Techniques:

# Clustering:

Clustering is a data exploratory technique used for discovering groups or pattern in a dataset. There are two standard clustering strategies: k-means and hierarchical clustering.

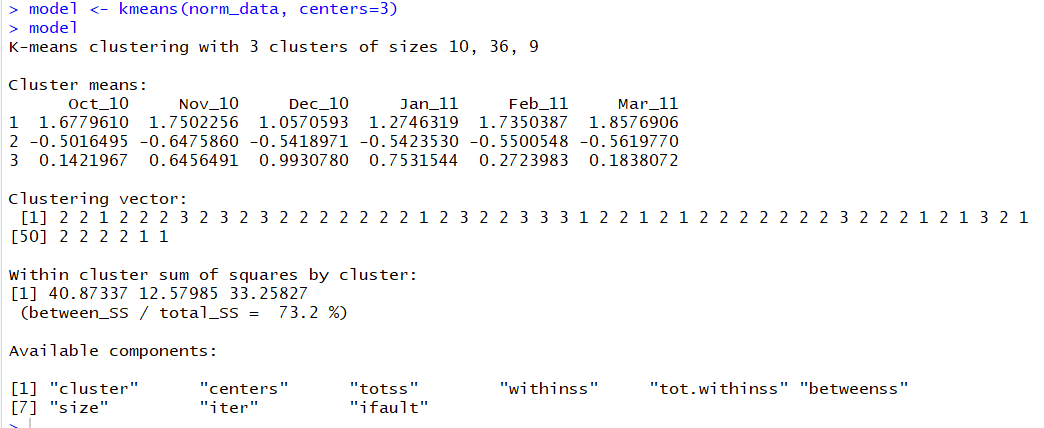
# K-means Clustering:

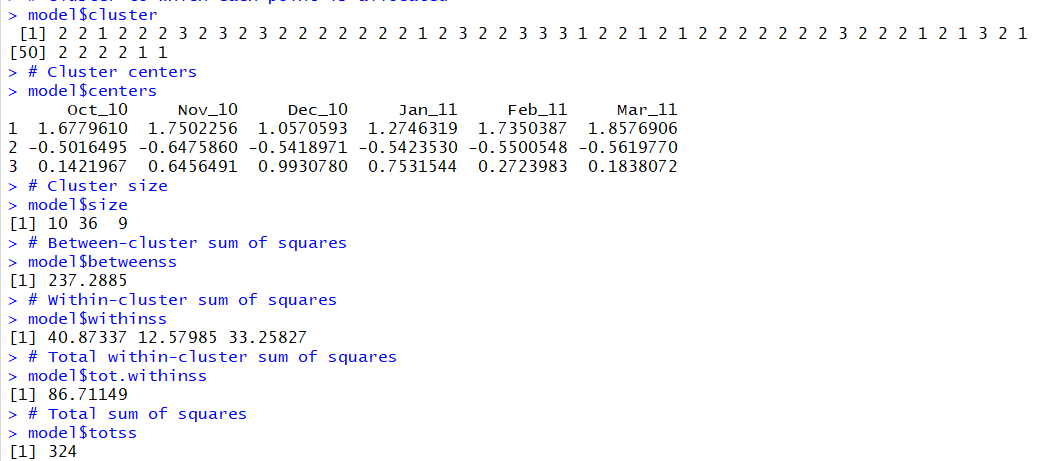
K-means clustering is most widely used partitioning method for splitting a dataset into a set of k groups (i.e. clusters). In this method we need to specify the number of optimal clusters to be generated from the data. There are several methods to determine the optimal number of clusters like a dissimilarity measure of Within Cluster Sum of Squares (WSS), Silhoutte coefficient, Dunn’s Index etc .

Within Cluster Sum of Squares can be calculated by:

First identify k clusters, it can be random - Identify the significant clusters and this process is iterative. If the distance between the observation and its closest cluster center is greater than the distance between the others closest cluster centers (Cluster 1, Cluster 2 …),then the observation will replace the cluster center depending on which one is closer to the observation. Each observation is allocated to the closest cluster, and the distance between an observation and a cluster is calculated from the Euclidean distance between the observation and the cluster center. The sum of these distances is called the WSS.

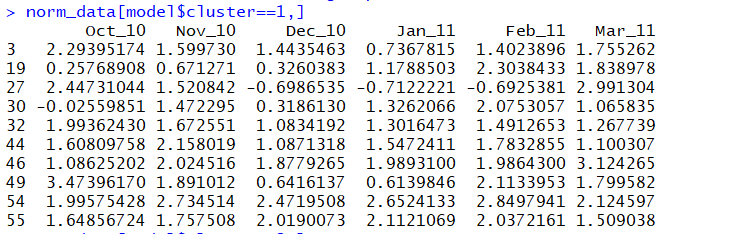
We build our model using k-means function and specifying 3 as number of clusters to be obtained and check the summary of this model .

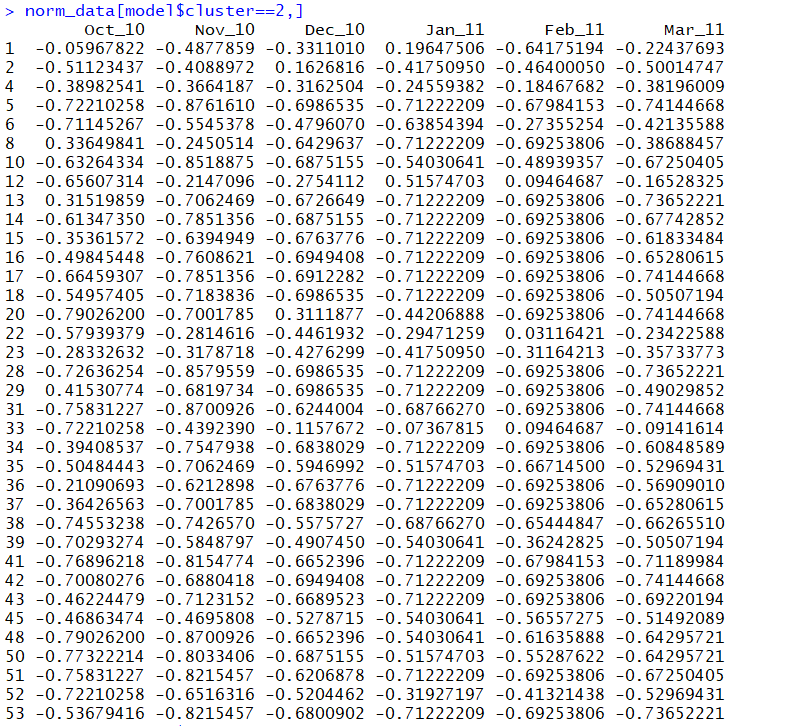


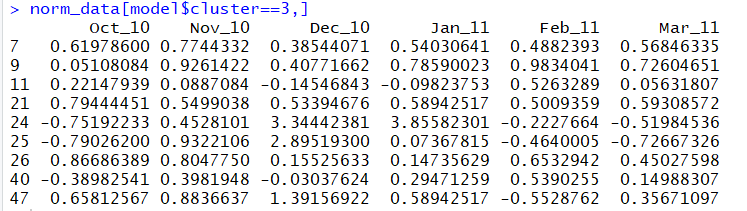


In the above output, we can see the cluster assignment of each data point, the centres allocated for each of the clusters, the number of data points assigned to each cluster are 10, 36 and 9. The between sunm of squares of the clusters is 237.28. The within sum of squares of the three clusters are 40.87, 12.57 and 33.25. The total within cluster sum of squares is 86.71 and total sum of squares is 324.

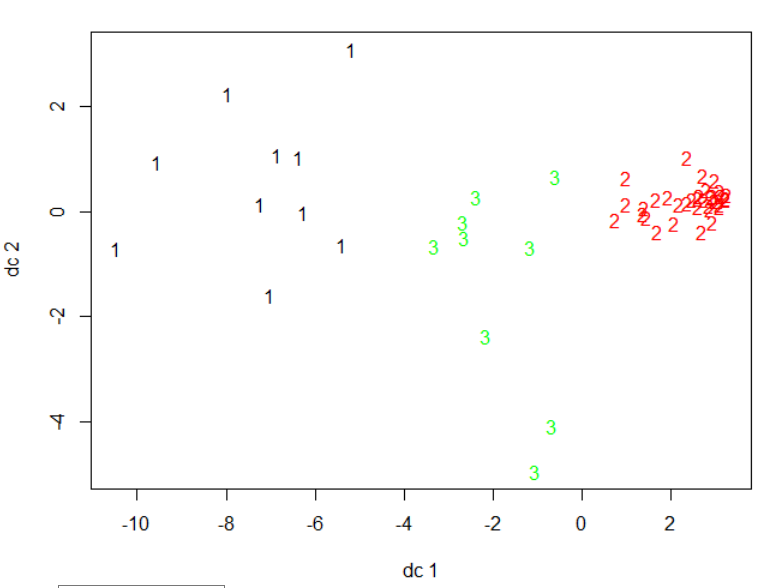
We next check the items assigned to each cluster as below:



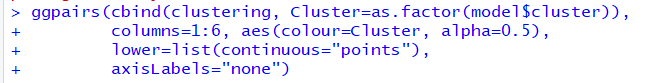


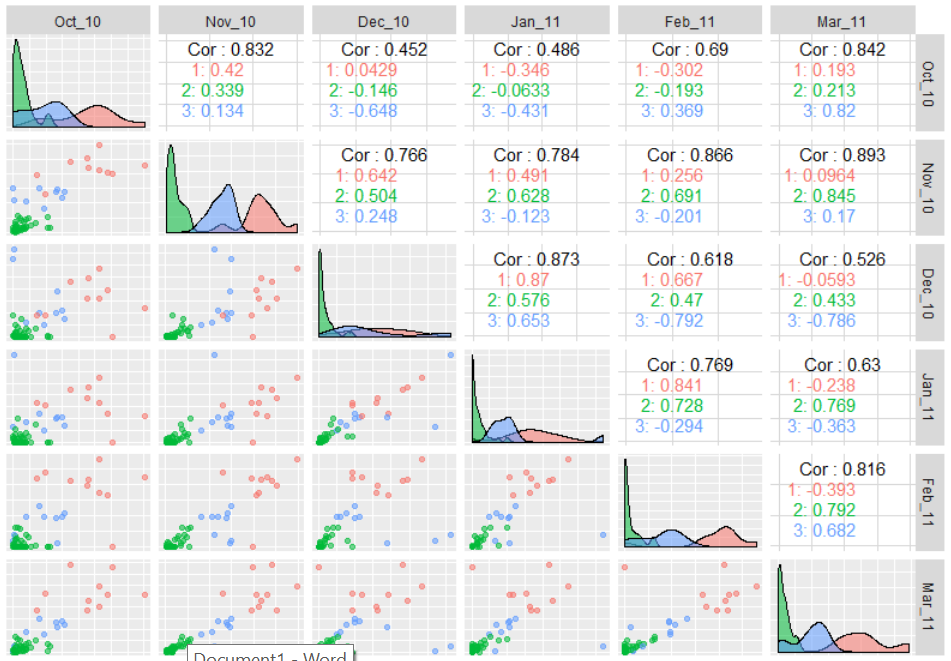


We will plot these items where we can visually see the clusters assigned to these members as below:

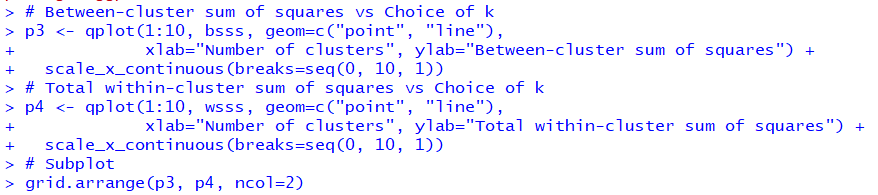


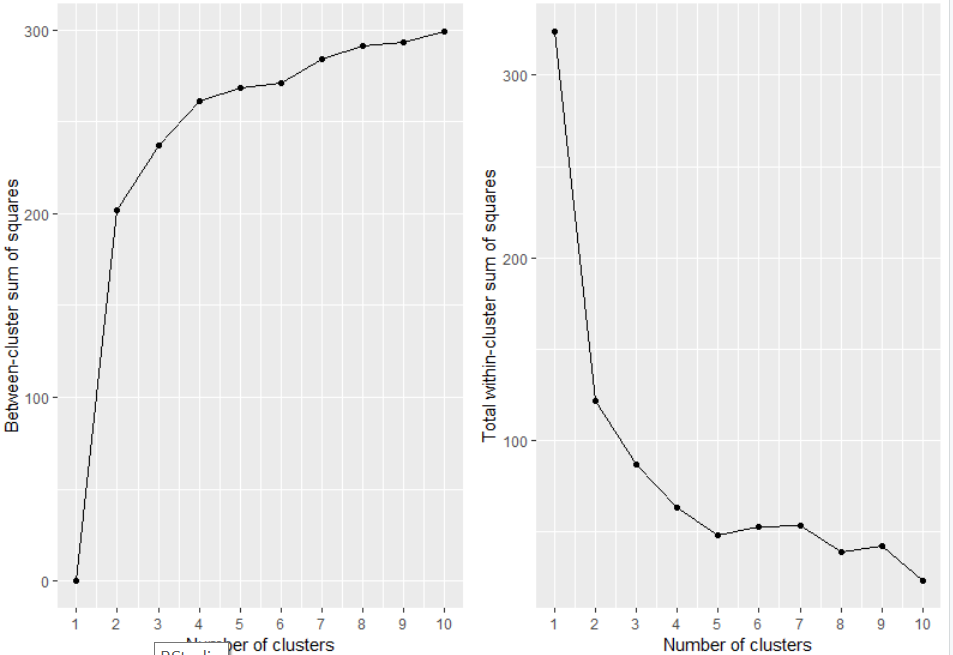
We will plot different plots and combine them to see the clusters and it’s behavior visually.





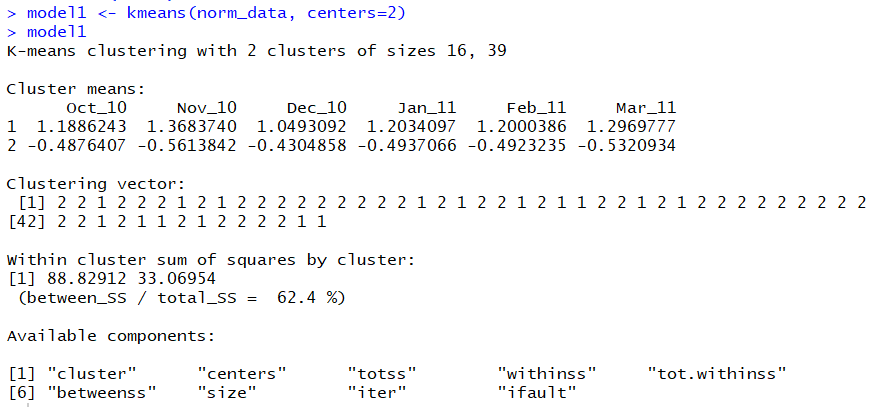
We will plot the between sum of squares and total within cluster sum of squares with the number of clusters to have an idea about the number of clusters which may be used.

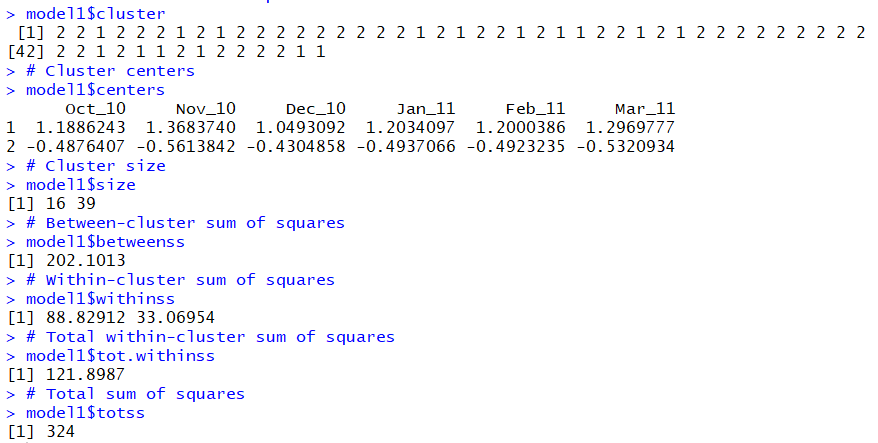




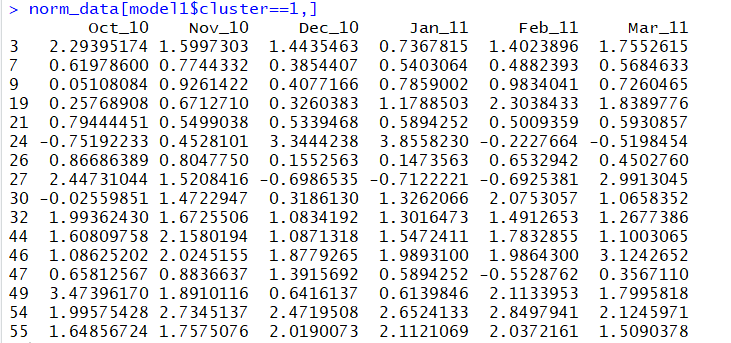
In the above plots, we see that the sum of squares values are very steep at k=2 hence we will evaluate using k=2 clusters and see the results.

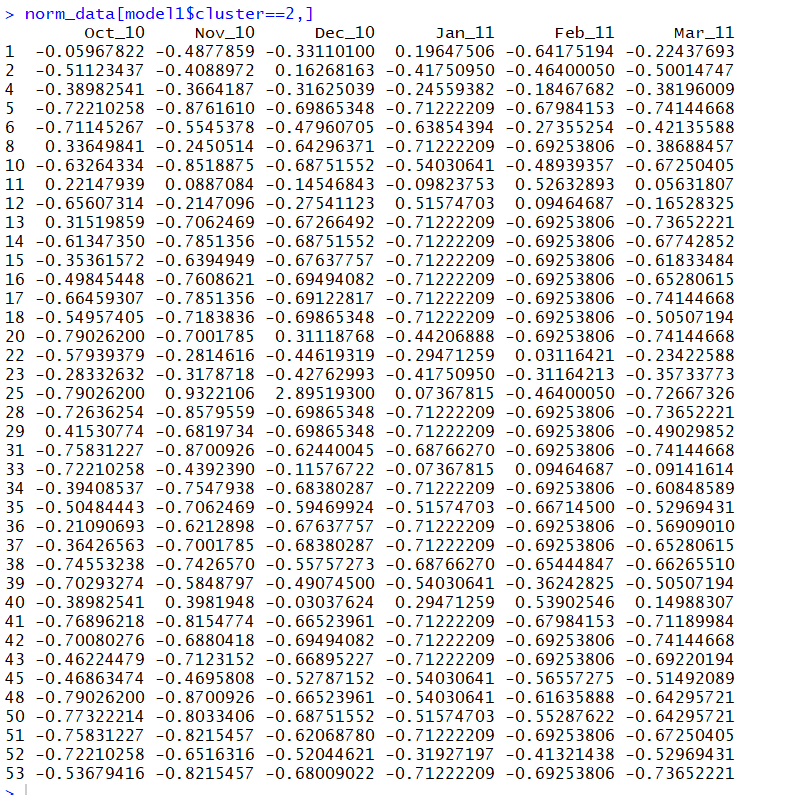
## K means with k=2:

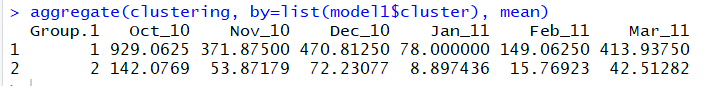




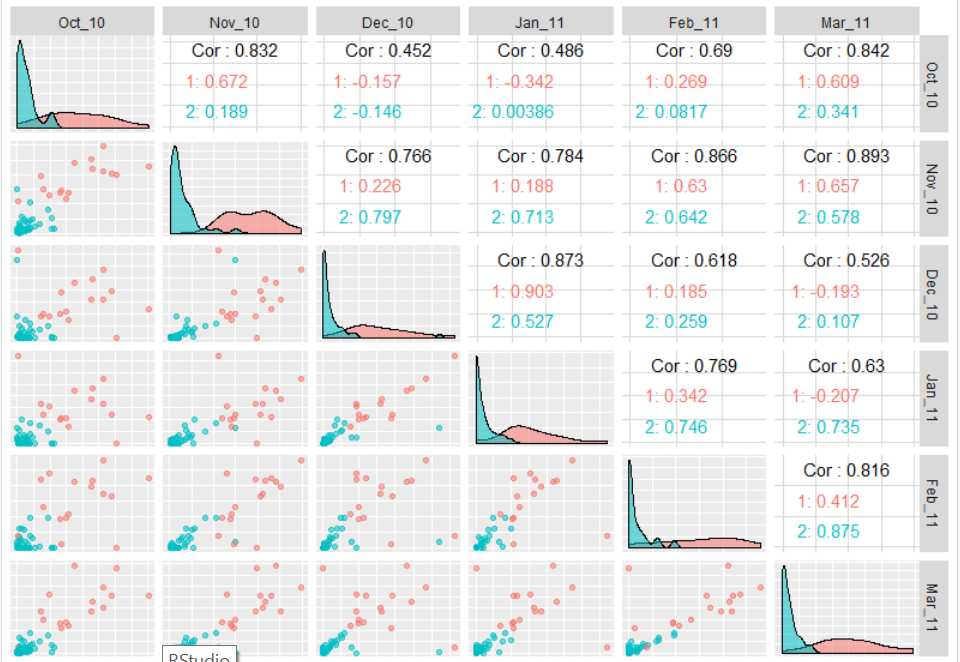
**Items in each cluster:**







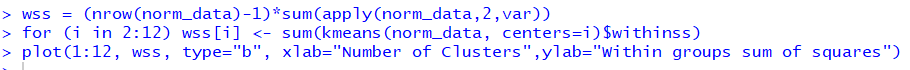
Let’s now plot the clusters using different plots combination for k=2.

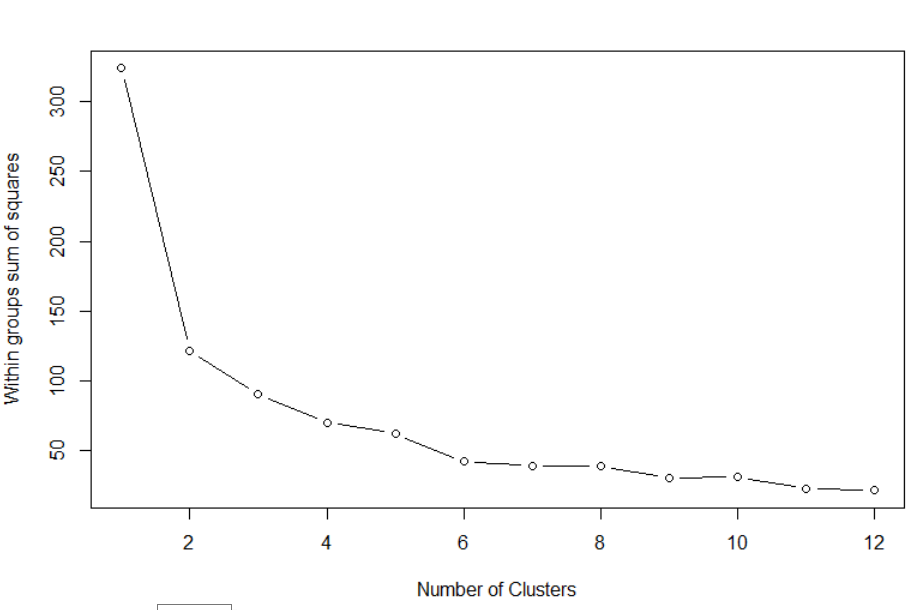


## Determining optimal number of clusters:

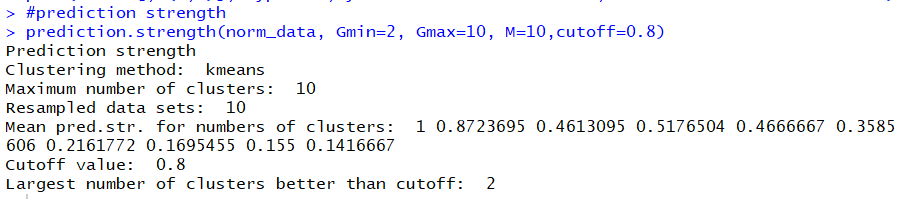
Here, we will implement 3 methods to determine optimal number of clusters.

1. Group Sum of squares:

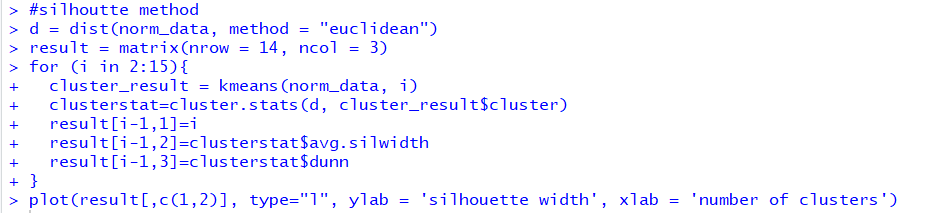


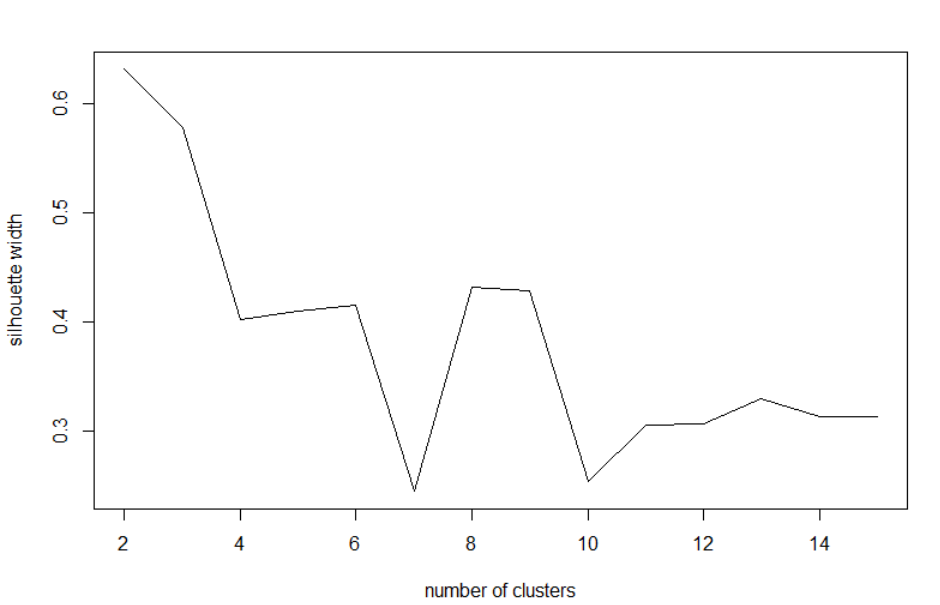


There is a steep decrease at 2 and after that from 5 to 6. Therefore, based on this graph we would prefer to have 2 clusters.



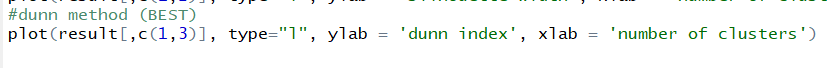
1. Silhouette Method:



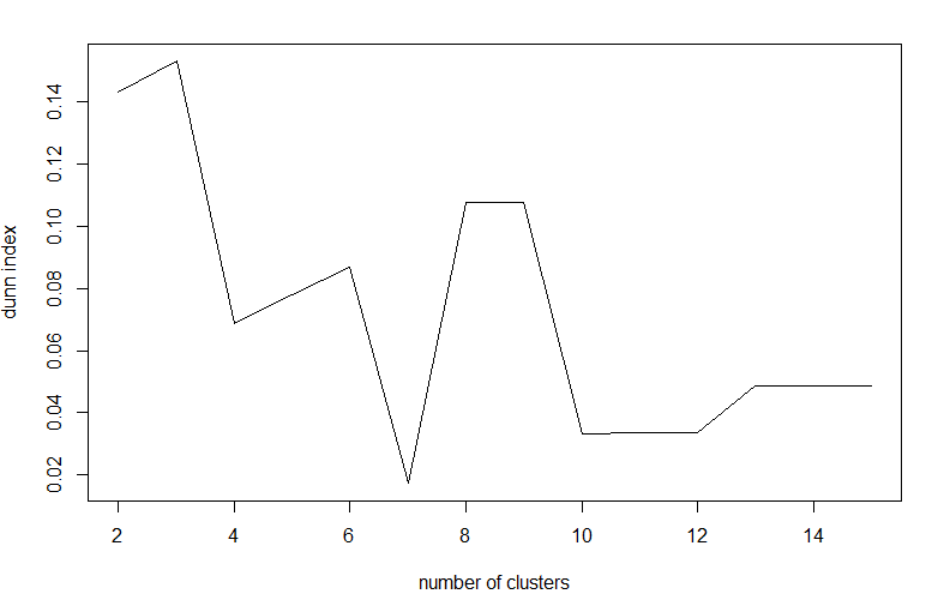


In the Silhouette method, we can see in the above graph that 2 has a very high silhouette index. We can take this value as this was also verified by the previous elbow plot.

1. Dunn Index Method:



In this method, we will plot number of clusters vs their Dunn index value to determine the optimal number of clusters.



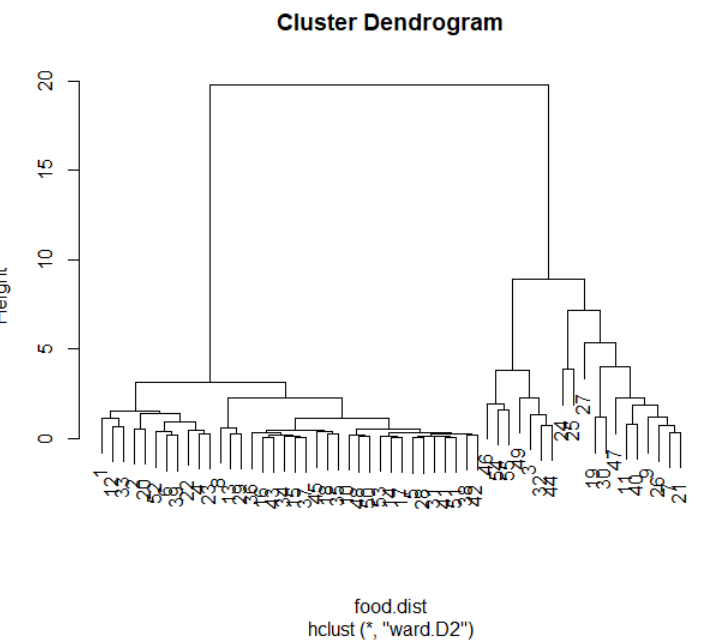
In the above plot, we see that the Dunn index increases from 2 to 3.

# Hierarchical Clustering:

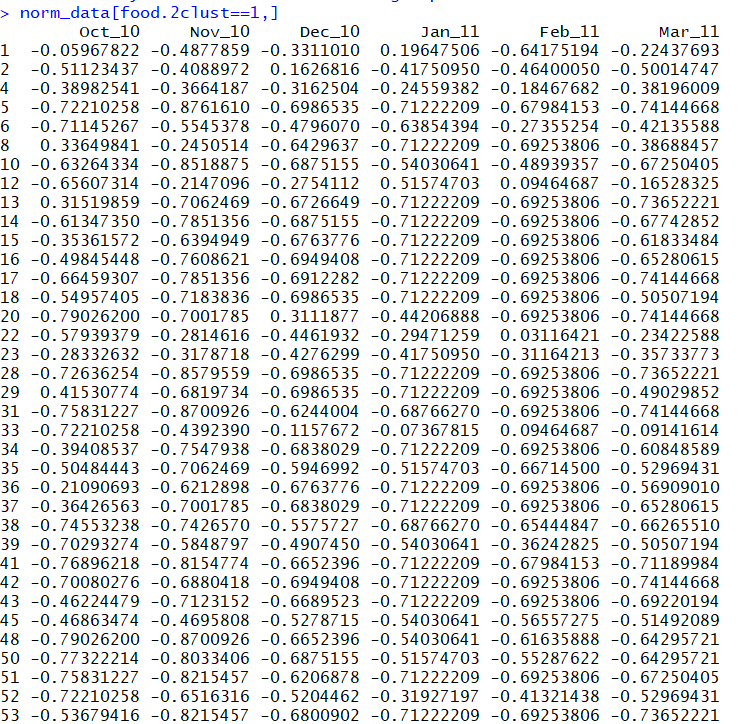
Hierarchical clustering is an alternative approach to k-means clustering for identifying subgroups in the dataset. It does not require to pre-specify the number of clusters to be generated. Dendrogram is a tree-based representation of the observations. It uses pairwise distance matrix between observations as clustering criteria. Hierarchical clustering can be divided into two main types: agglomerative and divisive.

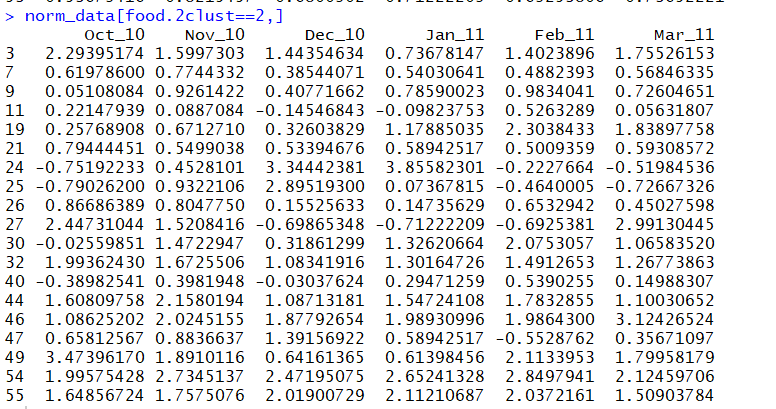
Agglomerative clustering: It’s also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. Each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root). The result is a tree which can be plotted as a dendrogram.

We will plot the dendogram using hclust function and see the visualization below. In this plot, we can see that the dissimilarity measure i.e the euclidean distance difference is large when we cut the cluster at k=2. Thus, we choose k= 2 as the optimal cluster which was also verified by the previous methods.

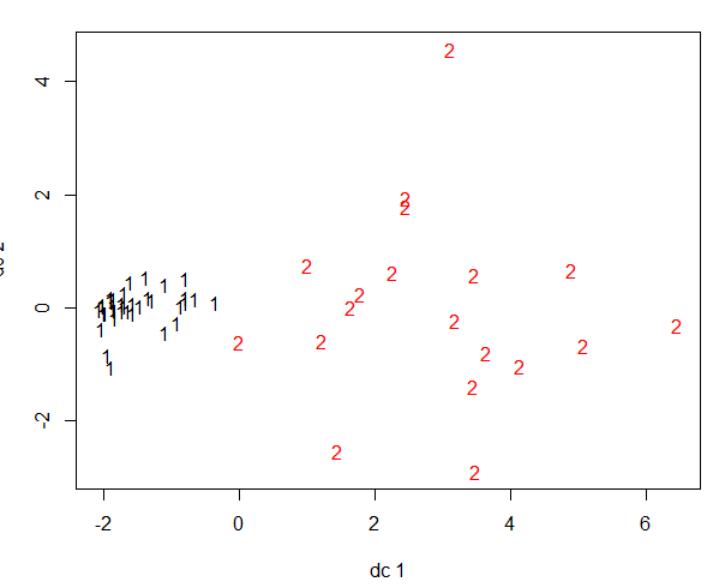


We will check the members assigned to each clusters as below.





**We will plot the centroid plot using plotcluster function as below.**



# Association Rules:

Association rule mining is a data mining technique where we can find different co-occurring associations among a collection of items. It is sometimes referred to as “Market Basket Analysis”, since that was the original application area of association mining. The goal is to use this algorithm in our Cincinnati zoo data to find similar associations of items that should be sold together and hence may be placed together as compared to a random sampling possibility. Association rules analysis on the food data was done using the Arules package in R.

Let’s also define the parameters used to find these associations.

1. Support: It says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.
2. Confidence: It says how likely the item Y is purchased when item X is purchased, expressed as {X -> Y}.
3. Lift: It says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

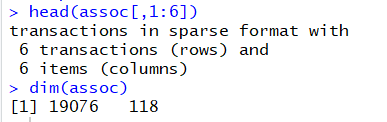
|  |
| --- |
|  |

## Data Exploration:

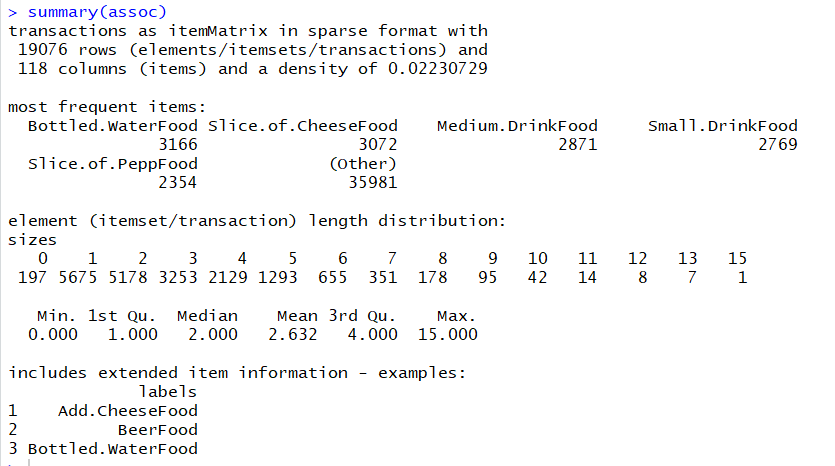
The below output shows the information about the transcational data for the products bought together which we will use to find the association rules.



We don’t need transaction id for modeling, hence we remove this variable for modeling purpose.



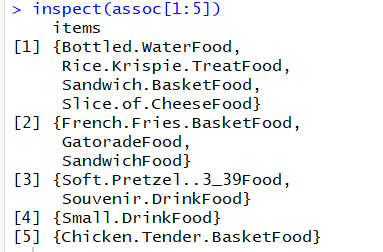
Our data has 19076 rows and 118 columns.



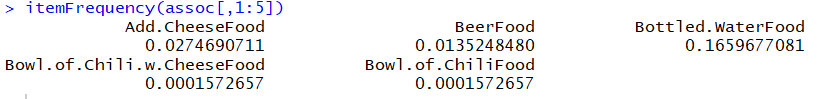
The density value of 0.02230729 (2.2 percent) refers to the proportion of nonzero matrix cells. Since there are 19076 \* 118 = 2250968 positions in the matrix, we can calculate that a total of 2250968 \* 0.02230729 = 50212.996 items was purchased.

The most frequent items show: items that were most commonly found in the transactional data. Since, 3166 / 118= 26.83, we can determine that bottled water appeared in 26.8 percent of the transactions. A total of 5178 transactions contained only a single item, while 1 transaction had 15 items. The first quartile and median purchase sizes are 1 and 2 items, respectively, implying that 25 percent of the transactions contained 1 or less items. The mean of 2.632 items per transaction took place.

Let’s look at the center of sparse matrix:



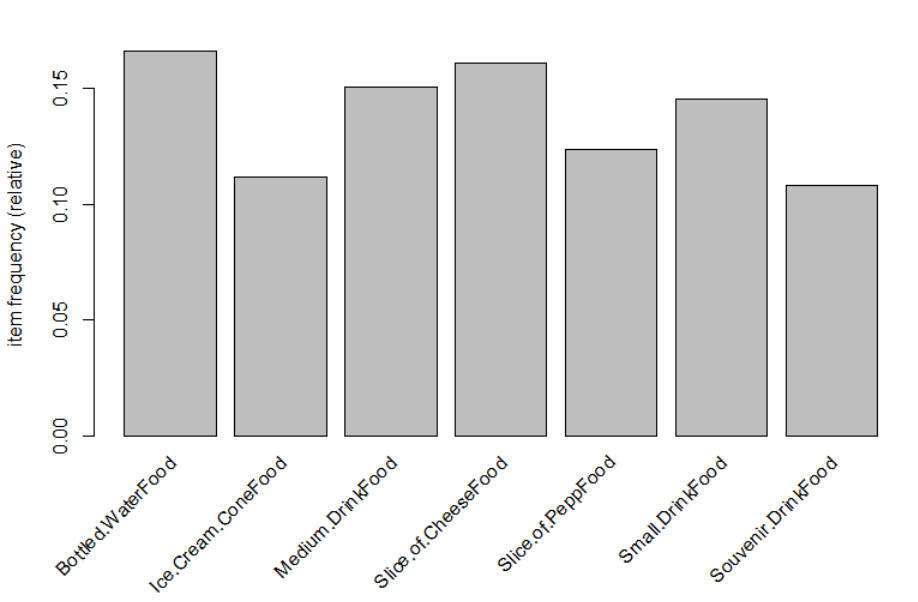
We can see the proportion of transactions that contain the item as below.



It tells the proportion of transactions that contain the item. (Also known as Support)

We will visualize the item support by the below plot as known as item frequency plots



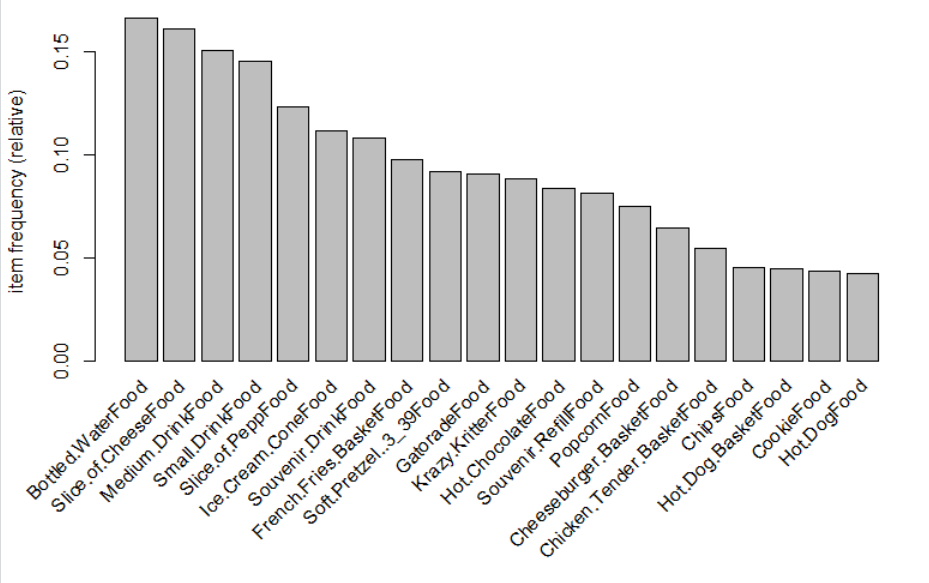


Here, Support means items that occurs 10% of transactions. We see that Bottled food, ice cream, drink food, cheese food etc. appeared in 10% of transactions.

Support = 0.1 means an item must have appeared in at least 0.1 \* 19076 =1907 transactions.

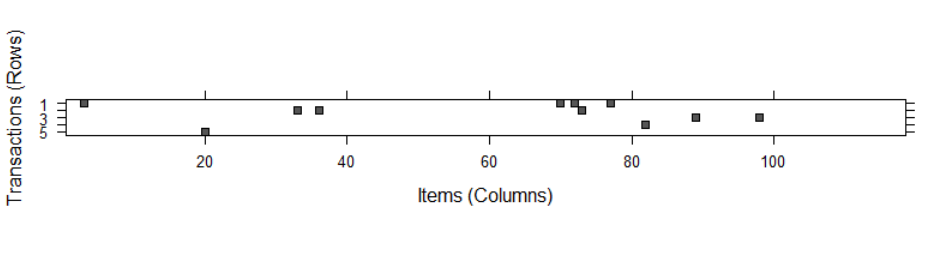
Now, let’s check the top 20 items that are in the transactions.





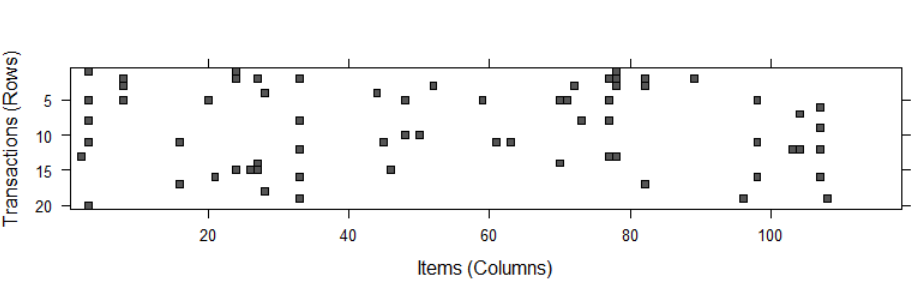
We will plot the sparse matrix by visualizing the transactional data as below.





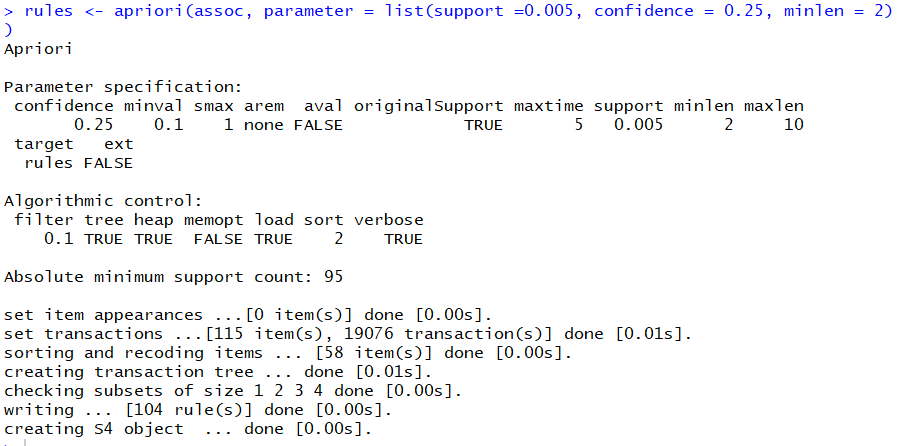
Cells in the matrix are filled with black for transactions (rows) where the item (column) was purchased. We will now plot for random 20 transactions as below:





## Apriori Model:

We will fit our model using apriori function of the arules library and see the summary as below.

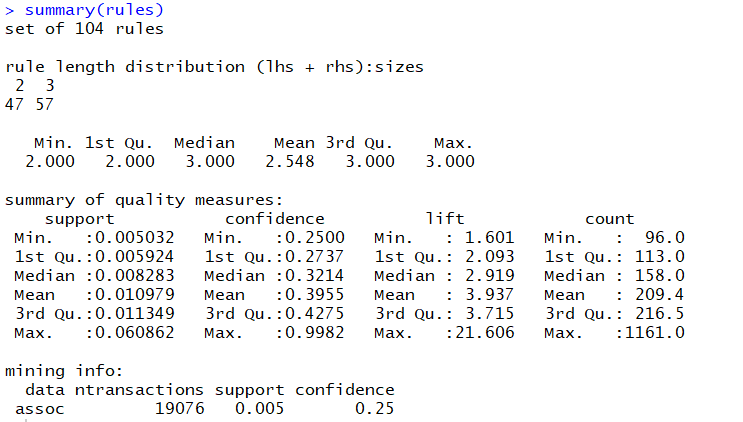


We set minlen = 2 to eliminate rules that contain fewer than two items, confidence threshold of 0.25 means that in order to be included in the results, the rule has to be correct at least 25 percent of the time and support as 0.005.

This model gives us 104 rules.

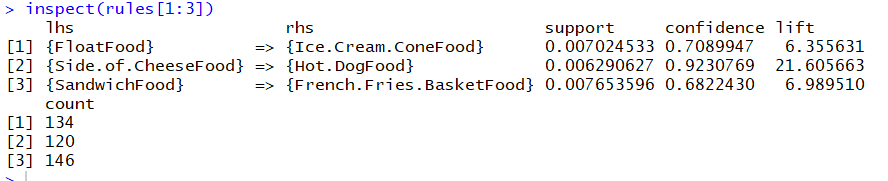


We will check the summary of these rules as below.



We see that 47 rules contains 2 items and 57 rules contain 3 items. Lift of a rule measures how much more likely one item or itemset is purchased relative to its typical rate of purchase, given that we know another item or itemset has been purchased. If lift is greater than one, it implies that the two items are found together more often than one would expect by chance. A large lift value is therefore a strong indicator that a rule is important, and reflects a true connection between the items.

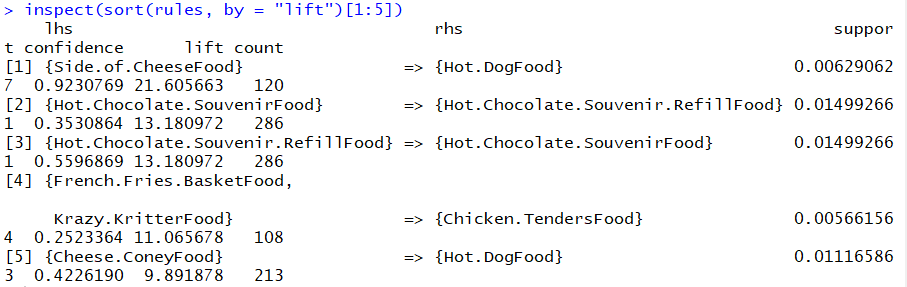
Let’s look at first 3 rules:



1st rule tells us that this rule covers 0.7 percent of the transactions and is correct in 70 percent of purchases involving float food. The lift value tells us how much more likely a customer is to buy whole Ice cream conefood relative to the average customer, given that he or she bought a float food.

## Improving model performance:

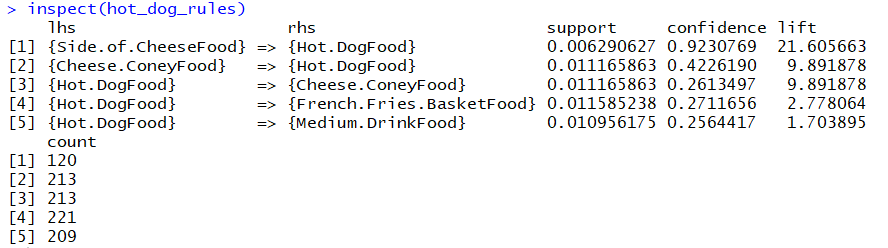
We will now sort the set of association rules as below.



In the top 5 rules, we see that people who buy hot dog food are nearly 21 times more likely to buy side of cheese food than the typical customer



subset() function provides a method to search for subsets of transactions, items, or rules.



We can see that hot dog food is purchased frequently with side of cheese food, cheese coney food, french fries basketfood and medium drinkfood.

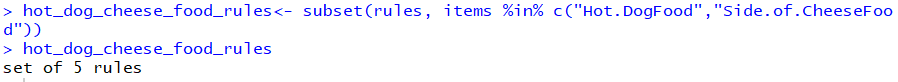
Let’s check the rules containing hot dog food on the left side.



Let’s check the rules containing hot dog food on right hand side.



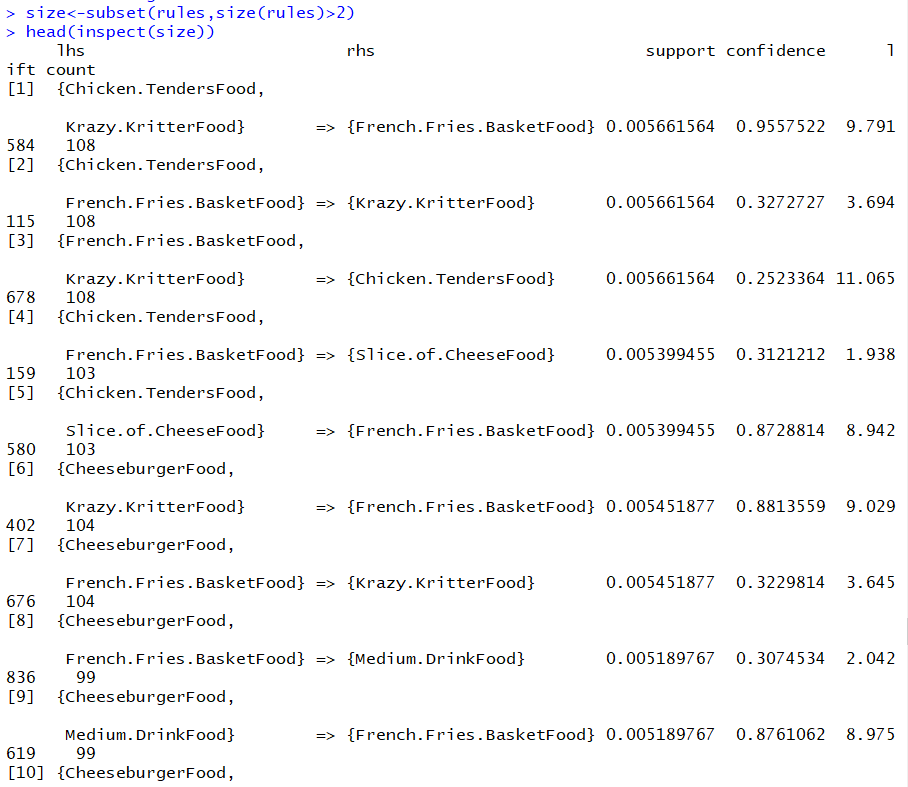
Let’s check the rules matching either hot dog food or side of cheese food.



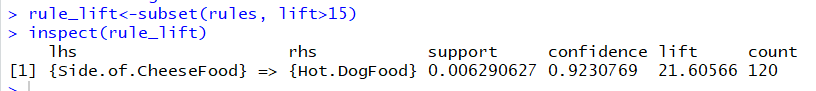
Partial matching allows us to find any item containing Chicken.



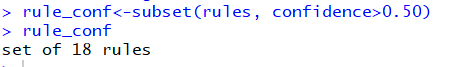
Let’s check rules having size >2.

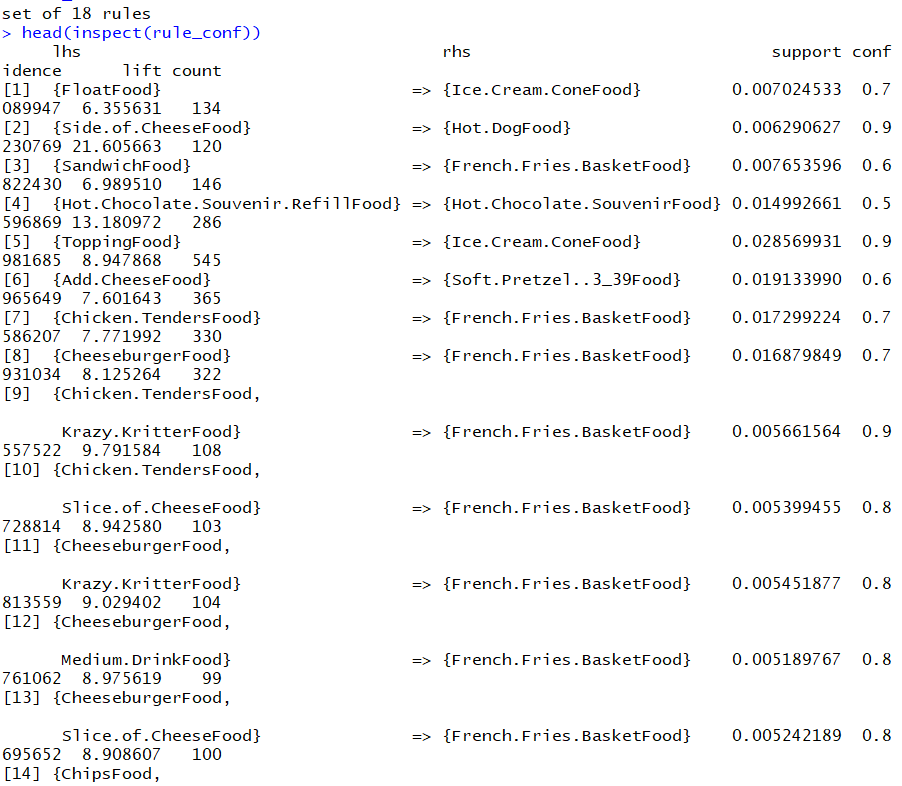


Let’s check rules having life >15.

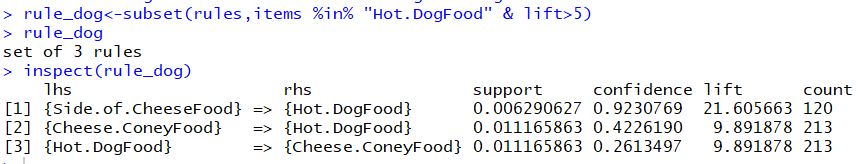


Let’s check rules having confidence>0.50

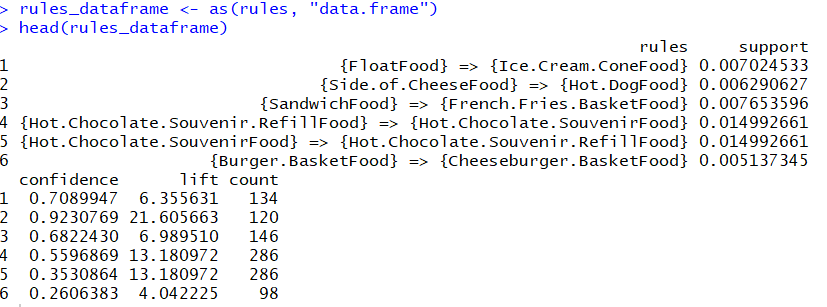




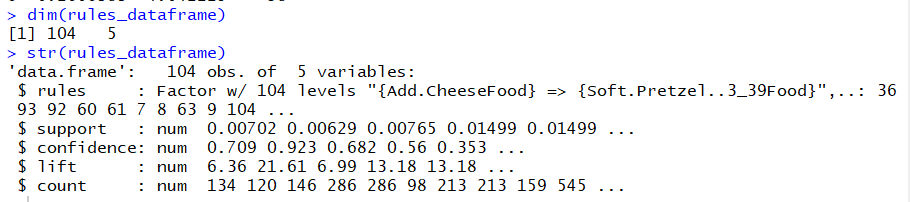
Let’s check the rules having lift>5 and item containing Hot Dog Food.



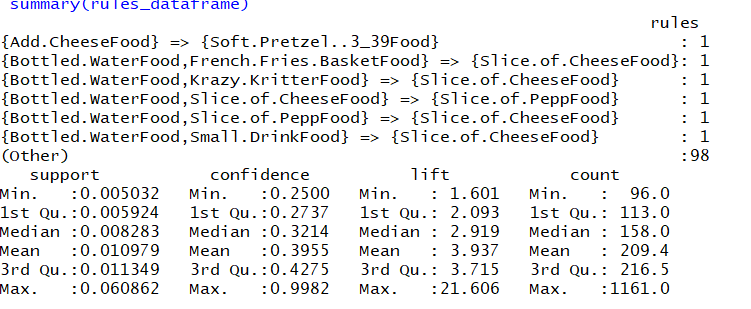
We will now save these rules to a dataframe as below.



We have obtained 104 association rules and their support, confidence, lift and count in this data frame.



Let’s check the summary for the association rules as below.



The final rules was written into a csv file which can be seen below.



# Conclusion:

We have used two unsupervised learning methods to give recommendations of items to be purchased/sold together at the Cincinnati Zoo. In this project we use champion-challenger method to evaluate the performance of these methods. Using clustering, we found the subgroups of the items that were sold at the zoo. Hence based on this result, we could give recommendations which product would fall into which subgroup which will help in increasing the sales and/or keeping the stock of items as required. In the next method, we implemented, association rules for these products to give a better insight on which product goes together with which other product/s. We found 104 association rules which can be implemented as recommendations at the Cincinnati zoo.

Hence based on this reasoning, we conclude that Association Rule will be our Champion and Clustering will be our Challenger as Association Rules was able to give us better insights about the market and the sales of product.