BANA 7047 – Prof. Yan YuIndividual Case II

Last Name: Ramakrishnan  
First Name: Siva  
M#: M10669365  
Signature: Siva Ramakrishnan

Executive Summary for Problem 1:

To perform the clustering analysis in IRIS dataset the following steps where followed

1. Dropped the species column form the dataset
2. Sampled the data by taking 90% of the observations of the dataset using M-number as seed
3. Used the K-means method present in the fpc package to perform the k-means clustering
4. Inspected the K-means clustered outputs of the considering different cluster numbers.
5. Estimated the ideal cluster size using within group sum of squares, Average Silhouette Coefficient and Dunn index
6. Looked at the Hierarchical clustering with various cluster sizes

Ideal Cluster size for K-means as per various methods are : Prediction strength method: 2 , Sum of squares method: 4, Average Silhouette Coefficient: 2, Dunn’s Index: 2

**Conclusion**: Overall the best guess would be that the data be divided into two clusters.

Below is the plot of the data considering number of clusters as 2 ( K-means on the left and Hierarchical clustering on the right). We observe that there is a good separation between the two clusters

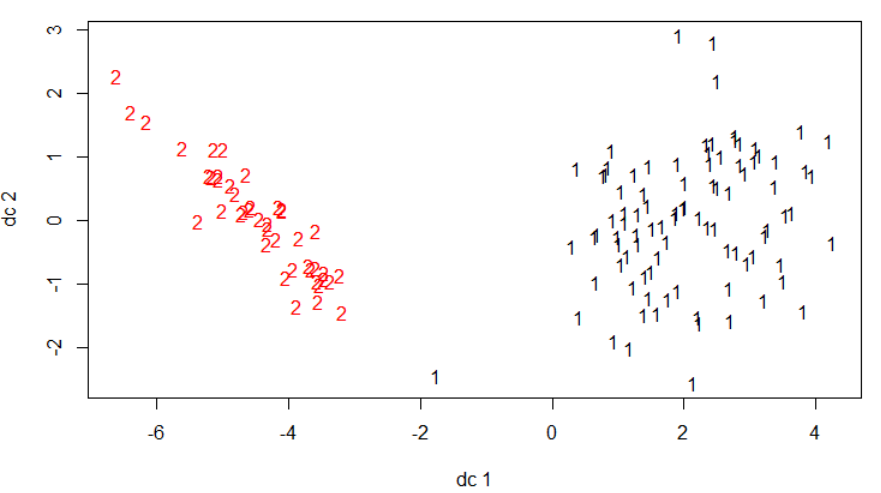
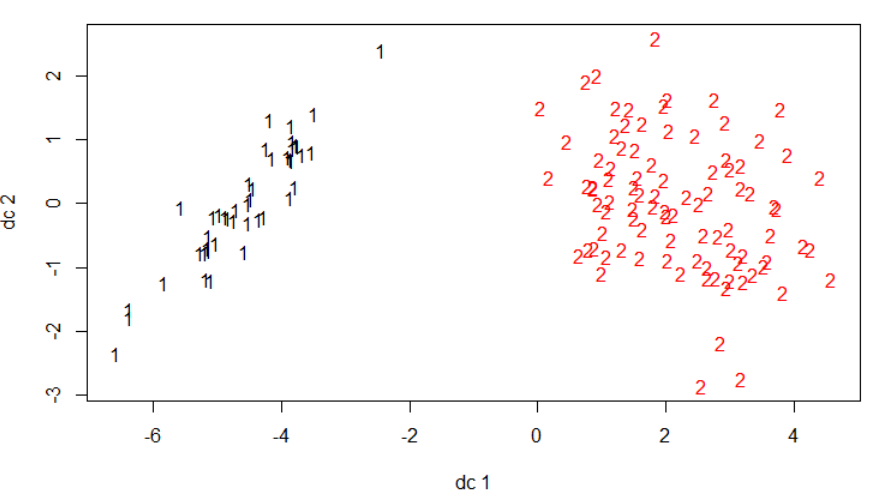


Figure :k-means and hierarchical clusters

The aggregate means for the various attributes of the scaled dataset in the two groups in case of K-means and hierarchical clustering is given below

Table : K-means clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
| 1 | -1.0112814 | 0.8376126 | -1.321308 | -1.26538 |
| 2 | 0.4889712 | -0.4049995 | 0.638874 | 0.611831 |

Table : Hierarchial Clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
| 1 | 0.4658072 | -0.4193923 | 0.6165243 | 0.592155 |
| 2 | -0.9966108 | 0.8973045 | -1.3190752 | -1.26694 |

**Inferences**: Both hierarchical and K-means clusters have around the same mean values within the clusters with just the cluster number being different( 1 in hierarchical is same as 2 in k-means). Inferences wrt the k-means cluster are listed below:

* The cluster 1 seems to have lower Sepal length, petal length and Petal width when compared to cluster 2
* Flowers in cluster 2 have lower Sepal width than cluster 1

Executive Summary for Problem 2:

The objective of this problem is to perform clustering analysis and find the association rules in the Cincinnati Zoo dataset.

Clustering analysis was performed on the quantity sold of 55 different categories of food across 6 months (from oct-10 to mar-11). Both K-means and hierarchical clustering the data was performed. To estimate the number of clusters for k-means within sum of squares, prediction strength, silhouette width and Dunn-index were used.

Estimate of 6 clusters from k-means was arrived based on the above analysis.

Table : Means across the clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Oct..10 | Nov..10 | Dec..10 | Jan..11 | Feb..11 | Mar..11 |
| 1 | 486.4 | 79.4 | 54 | 12.4 | 20 | 78.2 |
| 2 | 1342.125 | 462.75 | 522.375 | 81.125 | 182.25 | 548.375 |
| 3 | 9 | 260.5 | 1028.5 | 109 | 27.5 | 24 |
| 4 | 86.8 | 24.68 | 15.2 | 2.6 | 5.24 | 21.24 |
| 5 | 587.1429 | 289.5714 | 323.5714 | 59 | 127.1429 | 313 |
| 6 | 117.375 | 98.375 | 149.875 | 23.5 | 45.75 | 91.625 |

The categories of items in each cluster was also found (please refer to the detailed results section for the table of items in each cluster)

Following inferences can be made from the analysis of clusters:

1. Cluster 1 which consisted items popular during October like Cheese, Chicken Nugget basket chips
2. Cluster 2 items are usually the most purchased ones across all months. It contained items like Bottled water, Snacks..
3. Cluster 3 had only hot chocolate which seems to be popular during December.
4. Cluster 4 Consisted of several items items which aren’t purchased frequently like Iced tea, Salad etc
5. Cluster 5 items also seem to be popular across all months though not as much as cluster 2
6. Cluster 6 looks like it is the 3 most frequently purchased category across all months

For deriving the association rules, the dataset consisting of 19076 observations detailing the purchases made by individuals across 55 food categories was analyzed.

After the initial EDA, the association rules where derived using the arules packages. The best rules were chosen among them using support, Confidence and lift.

Following inferences were drawn based on the top association rules

1. Cheese coney / Side of Cheese had the strongest association with Hot Dog’s indicating that they should be bundled together.

2. Chicken Tenders had strong association with French fries.

3. Ice Cream Cones had strong association with Toppings and bottled water.

### Problem 1 Detailed analysis:

Plot of K-Means with various number of clusters is shown below:

1. There is good separation between the cluster in case of size 2
2. For cluster size 3 there is no proper separation for group 1
3. The separation between the groups decreases when there the cluster size is greater than 3

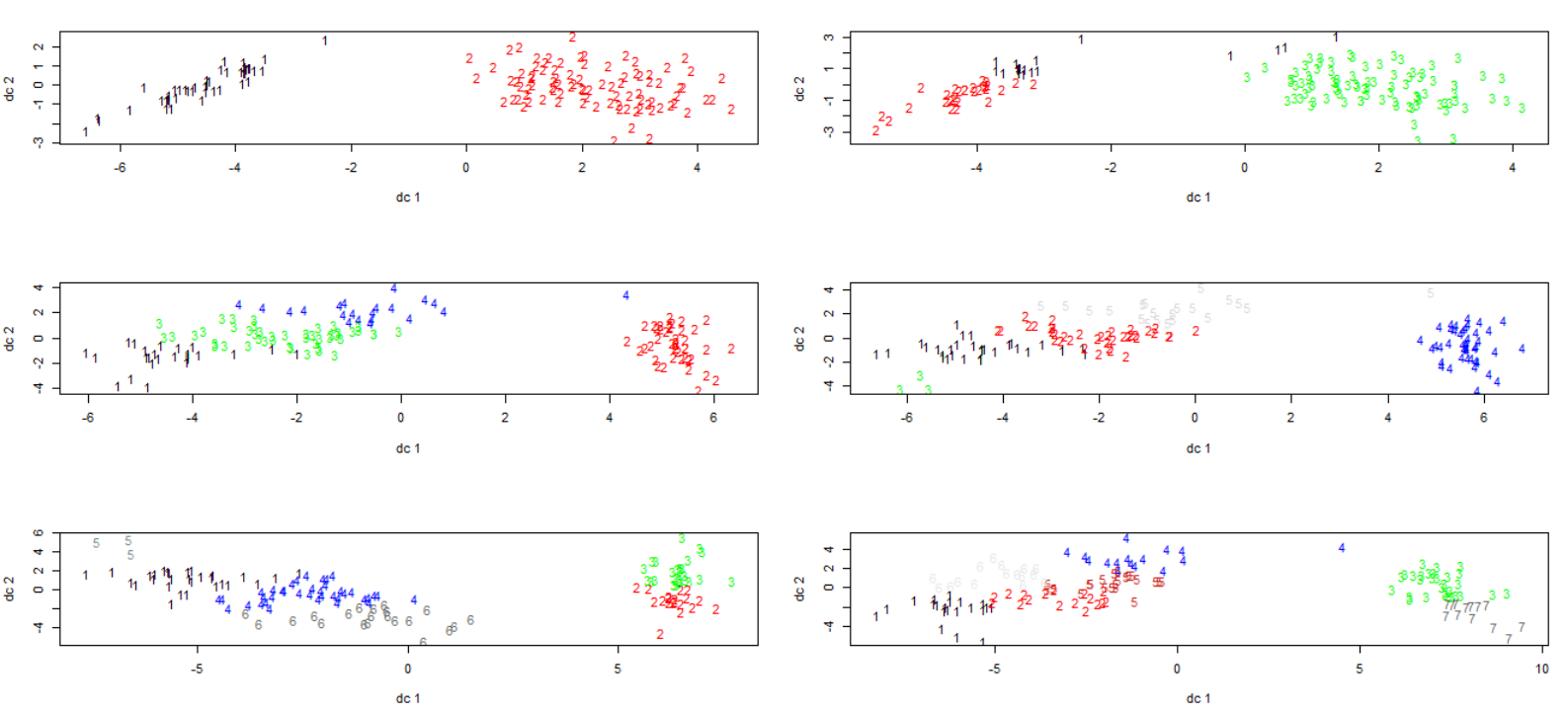


Figure : k-means clusters

**Estimating Number of clusters using Within Group Sum of squares:**

Below graph plots the within group sum of squares vs the number of clusters. The elbow point seem to be at 2

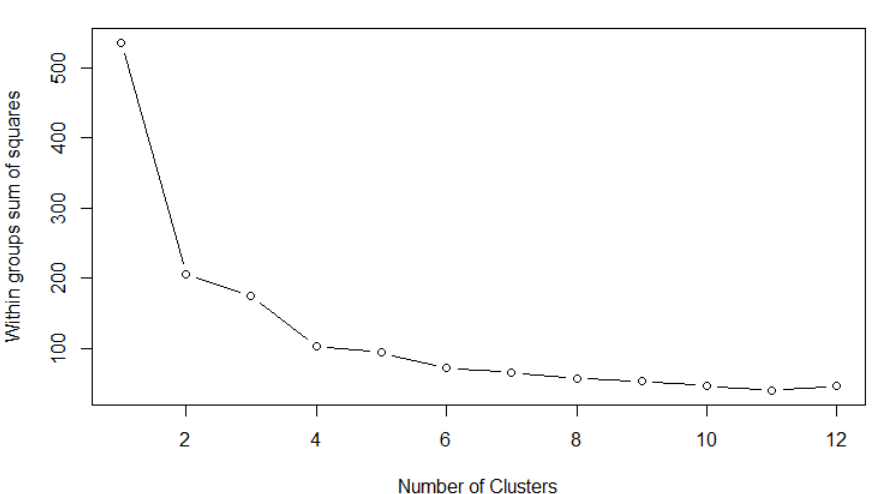


Figure Within group sum of squares

**Estimating Number of clusters using Prediction Strength:**

Shown below is the R output for Prediction strengths for various cluster sizes. Based on this output 2 seems to be the choice for the number of clusters.

Prediction strength

Clustering method: kmeans

Maximum number of clusters: 15

Resampled data sets: 10

Mean pred.str. for numbers of clusters: 1 1 0.6546604 0.5396884 0.4859692 0.4262525 0.3927876 0.3428078 0.283311 0.2788079 0.2223377 0.1409163 0.1672203 0.1018506 0.09392857

Cutoff value: 0.8

Largest number of clusters better than cutoff: 2

**Estimating Number of clusters using Average Silhouette Coefficient:**

The silhouette coefficient measures similarity of observations within the own cluster when compared to other clusters. It’s value ranges from -1 to 1 with a high indicating that the within cluster similarity is high.

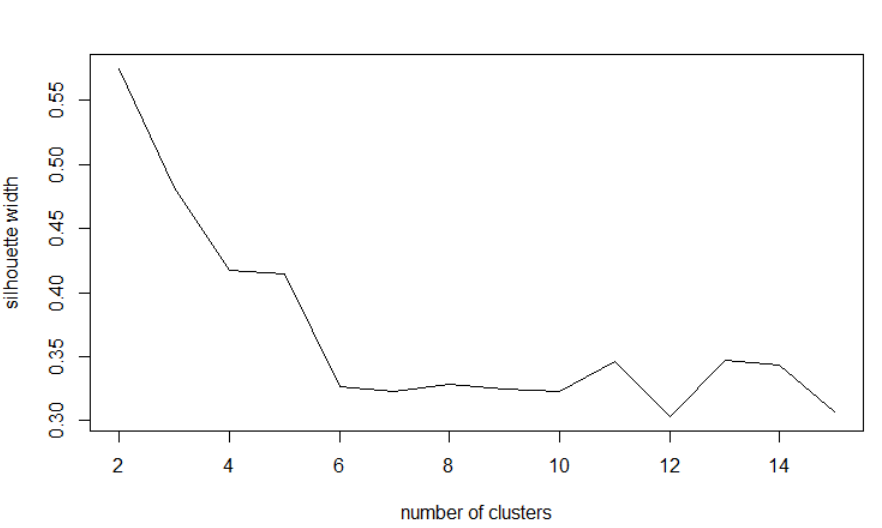


Figure : Silhouette coefficient

As per the above plot 2 seems to have the highest coefficient value.

**Estimating Number of clusters using Dunn’s Index:**

Dunn’s index s another metric in evaluating the clusters. It’s value ranges from 0 to infinity with a higher value being better.

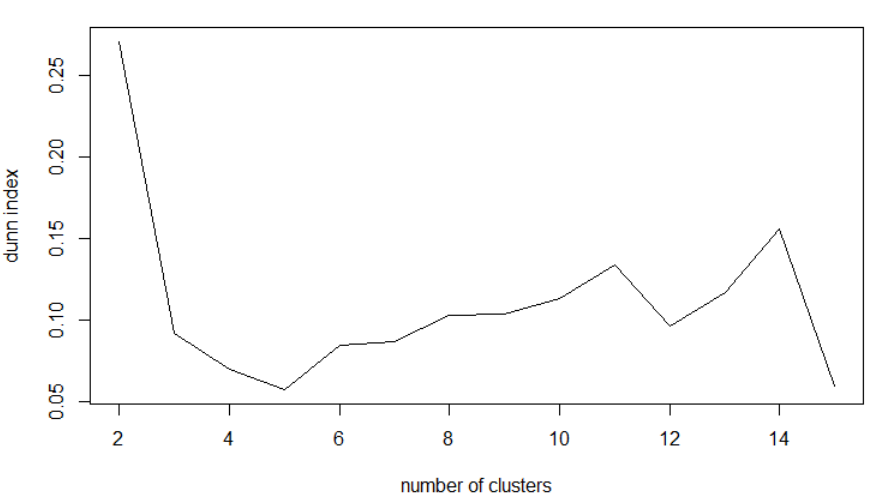


Figure : Dunn's Index

Based on the above plot number of clusters as 2 gives the highest index value.

**Hierarchical Clustering**

Hierarchical clustering cluster the data using bottom-up approach and cluster based on closest distance.

Below plot shows the clusters with number of cluster from 2 to 7.

On visual inspection, Number of clusters as 2 seems to be a good choice as there is good separation between the clusters.

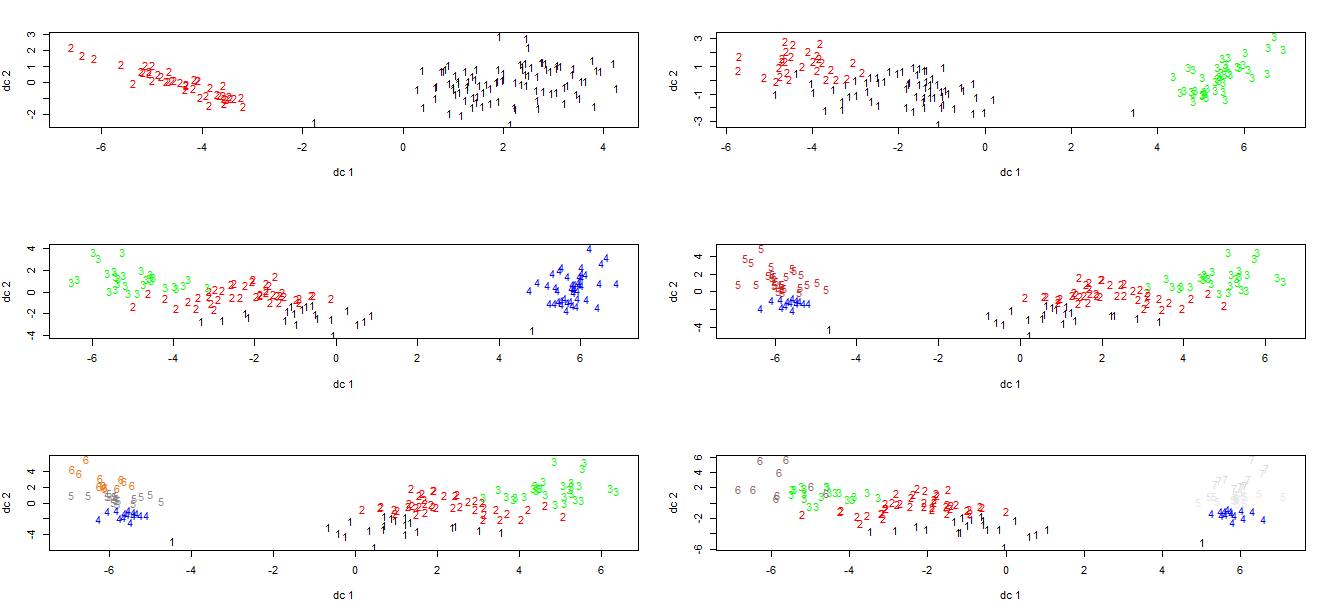


Figure : Hierarchical clusters

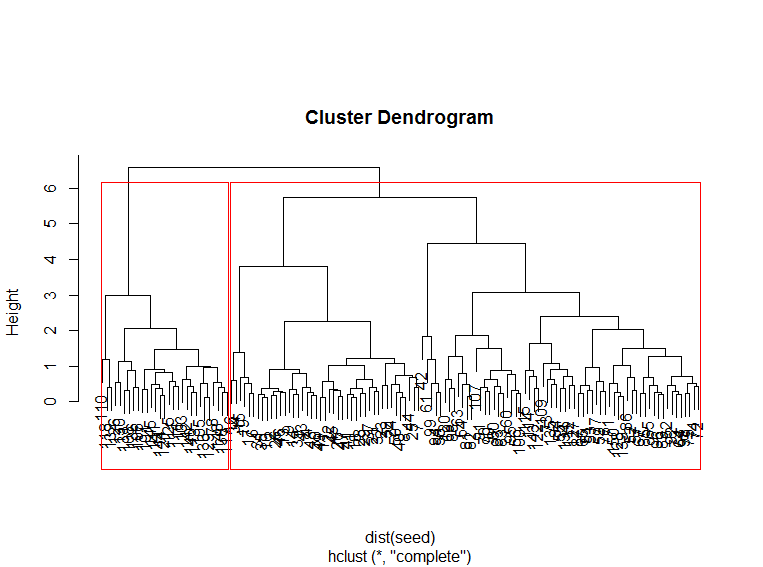


Figure : Hierarchical cluster Dendrogram

### Problem 2 Detailed Analysis

**Exploratory Data analysis**

Initial EDA across the datasets was performed.

Purchases across various months was looked at and it was found that most purchases happed during October with January being the least.

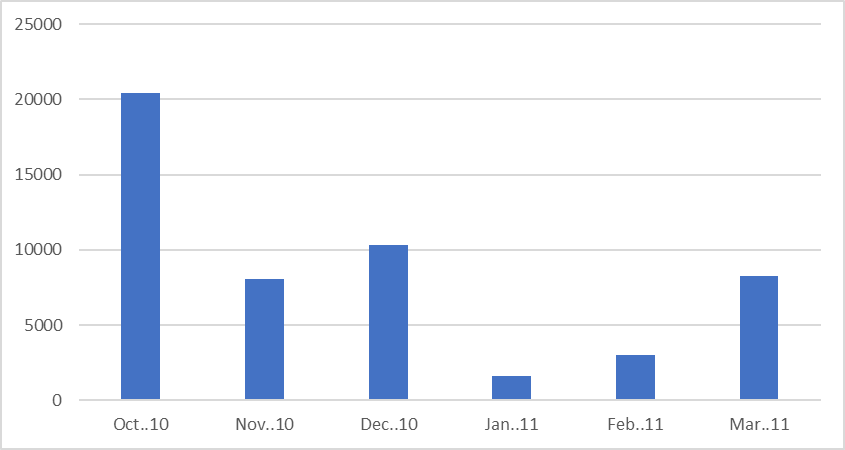


Figure : Sum of items Vs month

The Most frequently sold items in the Transactions dataset was looked at and it looks like Bottle of water and slice of cheese are the most popular items.

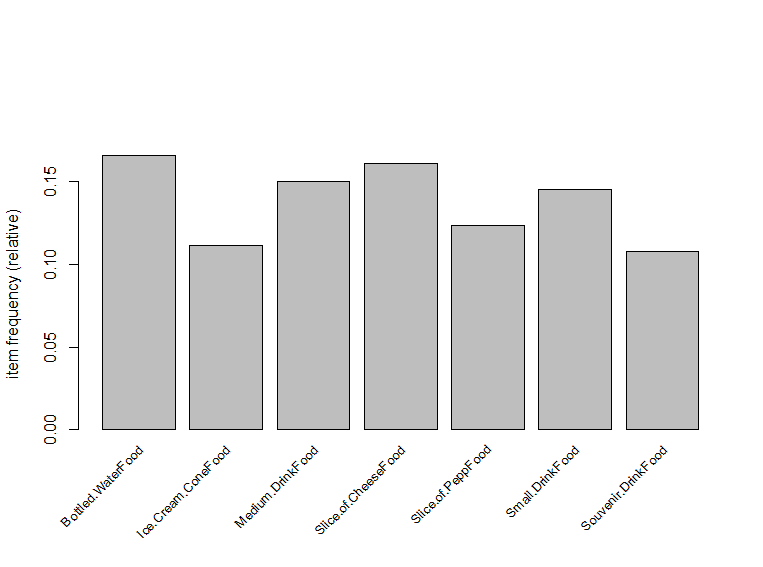


Figure : Top items sold

**Clustering Analysis**

Plot of K-Means with various cluster sizes is shown below:

Based on the visual analysis of the plots it looks like 4,5,6 seem to be good number for cluster sizes.

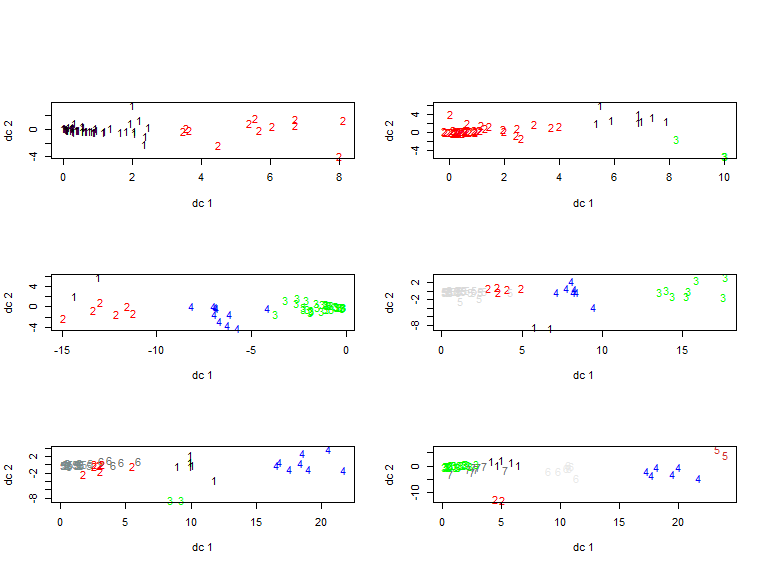


Figure : K-means Clusters

**Estimation of number of clusters using within group sum of squares**

The elbow point in the below plot is at 6.

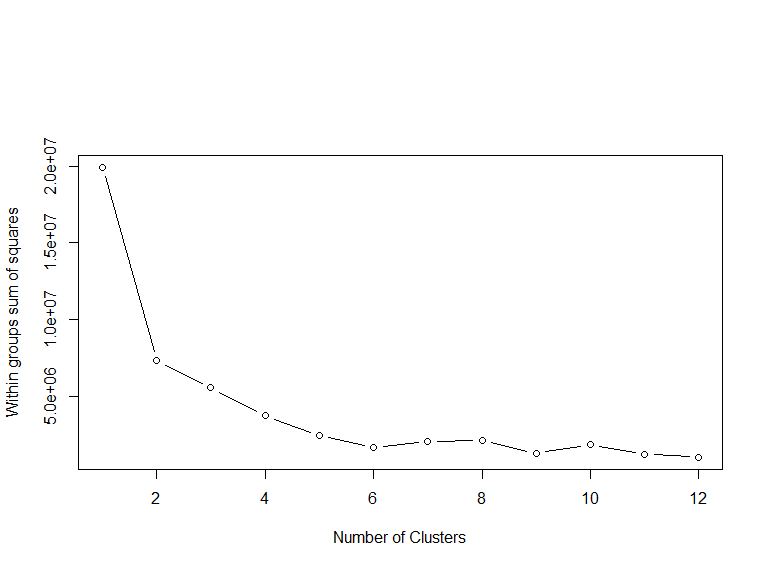


Figure :Within group Sum of Squares

**Estimation of number of clusters using Prediction strengths**

This method gives the result as 1 cluster. None of the other numbers cross the cutoff threshold.

Prediction strength

Clustering method: kmeans

Maximum number of clusters: 15

Resampled data sets: 10

Mean pred.str. for numbers of clusters: 1 0.7036531 0.5503645 0.4395635 0.5389894 0.4205795 0.3870165 0.2801587 0.1415476 0.1363095 0.06666667 0.07166667 0.055 0.04642857 0.2083333

Cutoff value: 0.8

Largest number of clusters better than cutoff: 1

**Estimating Number of clusters using Average Silhouette Coefficient:**

The number of cluster from this plot seems to be 2 although 6 also seems to be a good number with a local maximum occurring at that point.

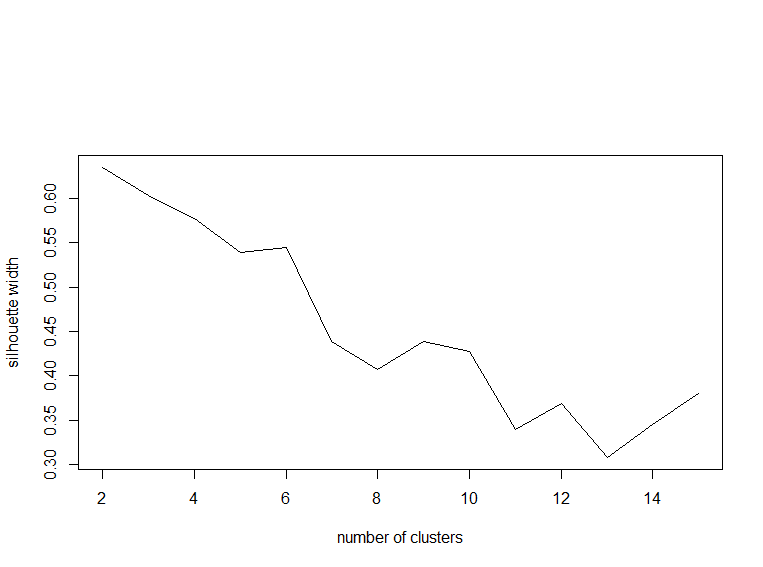


Figure : Silhouette width

**Estimating Number of clusters using Dunn’s Index:**

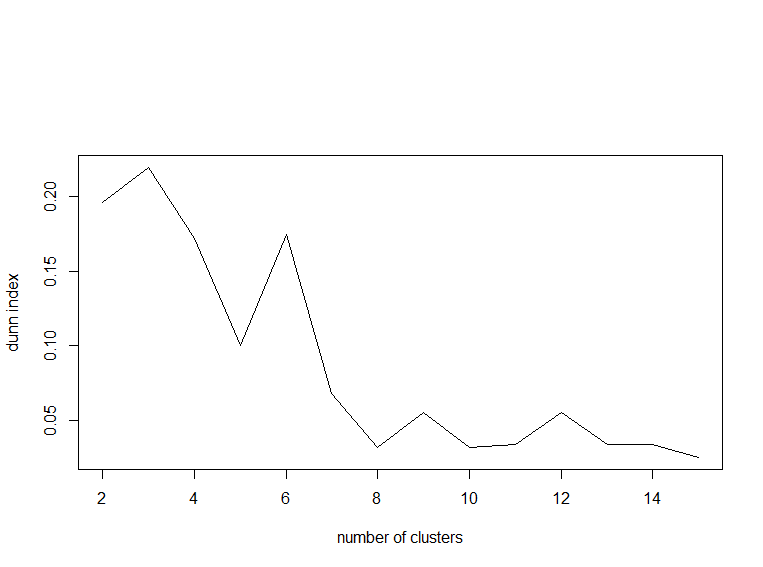


Figure : Dunn's Index

Based on Dunn’s index it looks like 3 and 6 are good choices for number of clusters.

Overall based on the analysis of all the methods number of clusters as 6 seems to be a good value.

The mean values of all the attributes in the 6 clusters are shown below.

Table : Means across the clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group.1 | Oct..10 | Nov..10 | Dec..10 | Jan..11 | Feb..11 | Mar..11 |
| 1 | 486.4 | 79.4 | 54 | 12.4 | 20 | 78.2 |
| 2 | 1342.125 | 462.75 | 522.375 | 81.125 | 182.25 | 548.375 |
| 3 | 9 | 260.5 | 1028.5 | 109 | 27.5 | 24 |
| 4 | 86.8 | 24.68 | 15.2 | 2.6 | 5.24 | 21.24 |
| 5 | 587.1429 | 289.5714 | 323.5714 | 59 | 127.1429 | 313 |
| 6 | 117.375 | 98.375 | 149.875 | 23.5 | 45.75 | 91.625 |

The food categories present in each of the 6 clusters is shown below:

Table : Items in the clusters

|  |  |
| --- | --- |
| Cluster 1 | Cheese Chicken Nugget Basket Chips Coney |
| Icee |
|  |  |
| Cluster 2 | Bottled Water IceCreamCone Medium Drink Small Drink |
| Snack Souvenir Drink Whole/Slice Cheese Whole/Slice Pepp |
|  |  |
| Cluster 3 | Hot Chocolate Hot Chocolate Souvenir |
|  |  |
| Cluster 4 | Capri Sun Cheese Fries Basket Chili Cheese Sandwich |
| Diet Dr Pepper Diet Pepsi Dr Pepper |
| Fish Basket Float Iced Tea |
| Land Shark Mountain Dew Nacho |
| Pepsi Pink Lemonade Red River Hog |
| Salad Shinerboch Siberian Chill |
| Sierra Mist Small/Large Fry Soup |
| Special Stella Veggie Burger Basket |
| Walking 3-Way |
|  |  |
| Cluster 5 | Cheeseburger Basket Chicken Tender Basket French Fries Basket Gatorade |
| Hot Dog Basket Krazy Kritter Soft Pretzel |
|  |  |
| Cluster 6 | Alchohol Burger |
| Coffee/HotTea Funnel Cake Topping |
| Gourmet Cup Grilled Chicken Sandwich Basket |
| Milk Sandwich Basket |

**Hierarchical Clustering**

Below plot shows the clusters with number of cluster from 2 to 7.

On visual inspection, Number of clusters as 5 seems to be a good choice as there is good separation between the clusters.

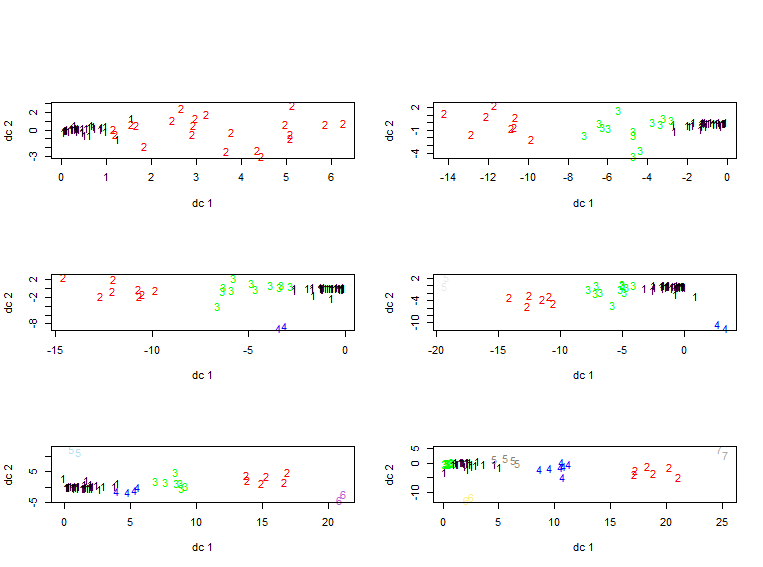


Figure : Hierarchical clusters

Cluster Dendrogram: On Visual analysis of the dendrogram 5 seems to be good number for number of clusters.

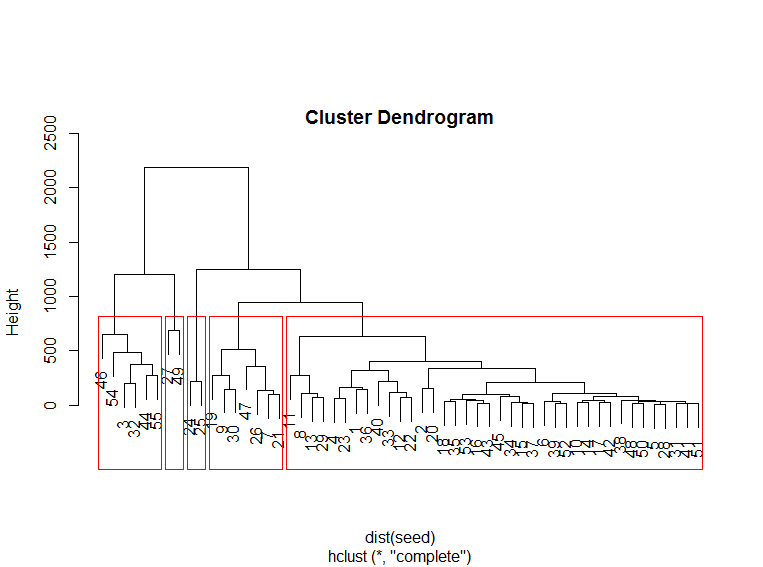


Figure : Dendrogram

Overall number of clusters as 5 was considered in case of hierarchical cluster.

Following table shows us the means of the attributes across the clusters :

Table : Means across the clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group.1 | Oct..10 | Nov..10 | Dec..10 | Jan..11 | Feb..11 | Mar..11 |
| 1 | 101.5294 | 43.23529 | 49.35294 | 8.529412 | 14.73529 | 40.26471 |
| 2 | 1202.5 | 474.5 | 636.33333 | 99.166667 | 206.16667 | 518.83333 |
| 3 | 563.5455 | 214.36364 | 221.45455 | 39.818182 | 89.63636 | 225.18182 |
| 4 | 9 | 260.5 | 1028.5 | 109 | 27.5 | 24 |
| 5 | 1761 | 427.5 | 180.5 | 27 | 110.5 | 637 |

**Association Rules:**

The Apriori alogorithm was used to determine the top association rules. A support of .4% and confidence of 80% gives us 12 association rules.

The 12 association rules obtained are

Table : Top Association Rules

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lhs |  | rhs | support | confidence | lift |
| [1] | {Cheese.ConeyFood, |  |  |  |  |  |
|  | Side.of.CheeseFood} | => | {Hot.DogFood} | 0.004351017 | 0.9325843 | 21.828193 |
| [2] | {Side.of.CheeseFood} | => | {Hot.DogFood} | 0.006290627 | 0.9230769 | 21.605663 |
| [3] | {Chicken.TendersFood, |  |  |  |  |  |
|  | Krazy.KritterFood} | => | {French.Fries.BasketFood} | 0.005661564 | 0.9557522 | 9.791584 |
| [4] | {CheeseburgerFood, |  |  |  |  |  |
|  | Krazy.KritterFood} | => | {French.Fries.BasketFood} | 0.005451877 | 0.8813559 | 9.029402 |
| [5] | {CheeseburgerFood, |  |  |  |  |  |
|  | Medium.DrinkFood} | => | {French.Fries.BasketFood} | 0.005189767 | 0.8761062 | 8.975619 |
| [6] | {Bottled.WaterFood, |  |  |  |  |  |
|  | ToppingFood} | => | {Ice.Cream.ConeFood} | 0.004036486 | 1 | 8.964286 |
| [7] | {ToppingFood} | => | {Ice.Cream.ConeFood} | 0.028569931 | 0.9981685 | 8.947868 |
| [8] | {Chicken.TendersFood, |  |  |  |  |  |
|  | Slice.of.CheeseFood} | => | {French.Fries.BasketFood} | 0.005399455 | 0.8728814 | 8.94258 |
| [9] | {CheeseburgerFood, |  |  |  |  |  |
|  | Slice.of.CheeseFood} | => | {French.Fries.BasketFood} | 0.005242189 | 0.8695652 | 8.908607 |
| [10] | {CheeseburgerFood, |  |  |  |  |  |
|  | Small.DrinkFood} | => | {French.Fries.BasketFood} | 0.004141329 | 0.8315789 | 8.519441 |
| [11] | {Chicken.TendersFood, |  |  |  |  |  |
|  | Small.DrinkFood} | => | {French.Fries.BasketFood} | 0.004822814 | 0.8214286 | 8.415452 |
| [12] | {Chicken.TendersFood, |  |  |  |  |  |
|  | Medium.DrinkFood} | => | {French.Fries.BasketFood} | 0.004141329 | 0.814433 | 8.343783 |

The lift value can be interpreted as the importance of the rule.

A lift value >1 indicates that the LHS and RHS appear more often together. This means that occurrence of LHS means there is a higher probability of RHS also occurring.

Following inferences were drawn based on the top association rules

1. Cheese coney / Side of Cheese had the strongest association with Hot Dog’s indicating that they should be bundled together.
2. Chicken Tenders had strong association with French fries.
3. Ice Cream Cones had strong association with Toppings and bottled water.

The scatter plot of the 12 rules is shown below:

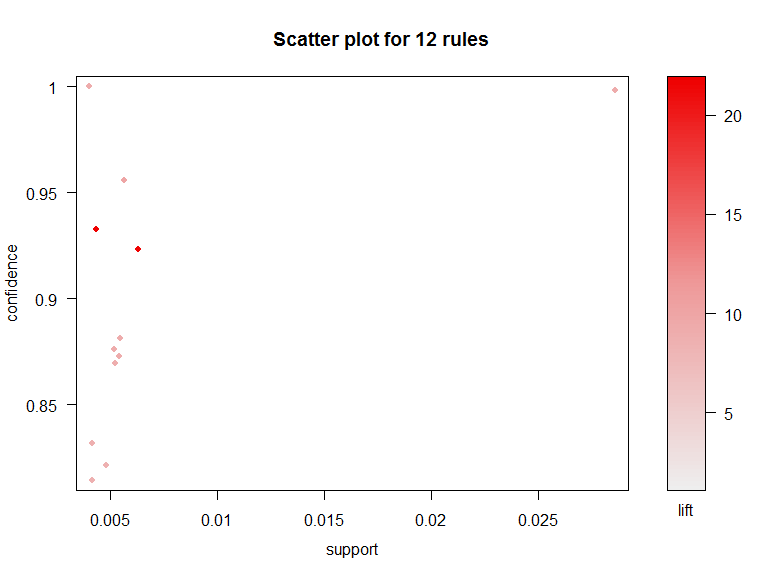


Figure : Scatter plot of the top rules

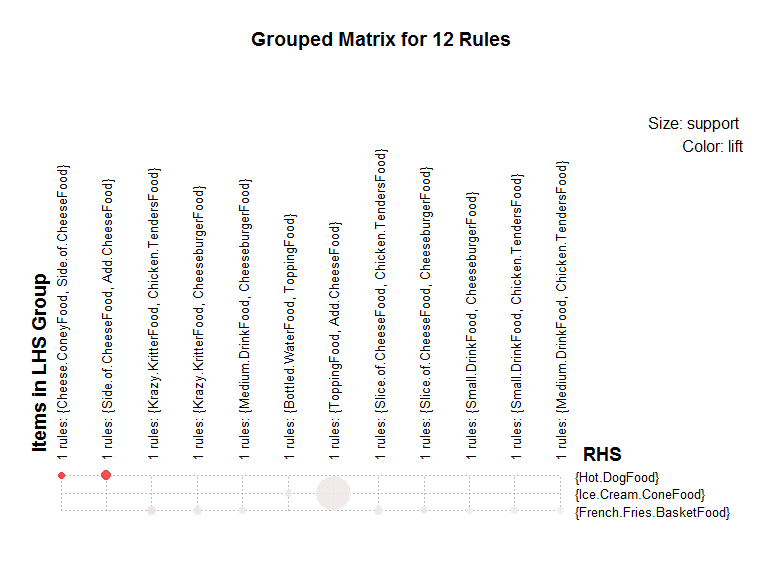


Figure : Grouped Matrix

plot of the LHS vs RHS is shown above.

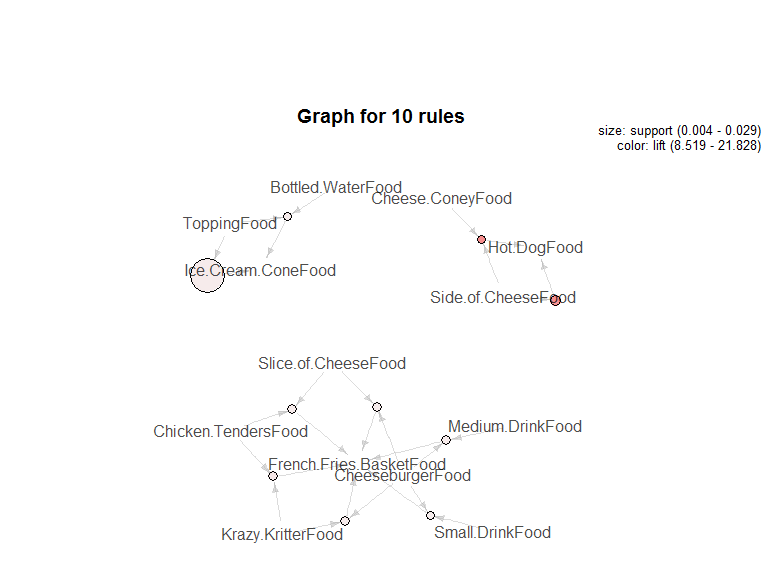


Figure : Graph of the top rules

Graph of the top 10 rules with the vertices as the items in the rules is shown above.