BANA 7047

DATA MINING II

INDIVIDUAL CASE II

Last Name: SURINENI

First Name: VENKATA GOPALA SRIHARSHA

M10574454

Signature: S V G SRIHARSHA

Executive Summary

1. Iris Data
   a) Goal and Background:

   The following analysis is carried out on the IRIS dataset available in default library of R. IRIS data set contains 150 observations of 5 variables each defining the dimensions of the petals and sepals of different flowers and the species they belong to respectively. The goal of the analysis is to cluster the dataset using K-means and Hierarchical clustering and comparing their performance. K-means clustering starts with "K" random observations as centroids of clusters and then assigns all observations to the nearest centroid and the new centroids are updated. Whereas, Hierarchical clustering starts with all the observations as different clusters and nearby points are clustered. K-means algorithm is computationally less intensive but highly sensitive to number of clusters and initial set of centroids.

   b) Approach:

   The approach here is to cluster 90% random sample of the dataset using K-means and hierarchical clustering algorithms and analyze their performance for different number of clusters. Arrive at number of clusters which minimizes intra-cluster variance.

   c) Major Findings:

   - Using K means clustering, the optimum number of clusters using different methods like : i) Minimizing within group sum of squares error is 3 and ii) Using prediction method is 2
   - The dendrogram for hierarchical clustering is plotted for 3 levels.

2. Cincinnati Zoo Data:
   a) Goal and Background:

   The first phase of this part of analysis is on the "qry_Food_by_Month" data set consisting of number of purchases of 55 food items for the months of July 2010 through March 2011, sold in Cincinnati Zoo. Goal of this analysis is to try and find groups of items which are similar in terms of purchases made using K-means and Hierarchical clustering algorithms and analyze the results.

   The second phase of the analysis is on the "food_4_association" dataset containing 19076 transactions of sets of food items purchased along with each other in each transaction at Cincinnati Zoo. Goal of the analysis is to try and find association rules describing the purchasing nature of zoo visitors to recommend improvements in business strategy based on the findings.

   b) Approach:

   The first phase of analysis starts with exploratory analysis of the data and then proceeds to build clusters using K-means and Hierarchical clustering algorithms. Minimum intra-cluster variance is used to arrive at number of clusters for K-means clustering algorithm. Clusters from both the methods are analyzed to arrive at key findings.

   The second phase is association analysis of food_4_association dataset, which starts starts with inspection of the nature of transactions and uses apriori algorithm to choose major association rules based on support, confidence and lift ratio of the association rules.

   c) Major Findings:

      i) Cluster Analysis of Food items purchase trends:
         - Hierarchical clustering resulted in more coherent results interpretable with similar purchase trends
         - Sales most of the items were low during Jan and Feb

      ii) Market Basket Analysis of Food items sold per transaction
         - There are 8 major rules with confidence (reliability) greater than 0.8 and lift (efficiency) greater than 20
         - These rules suggest the items, "Hot Dog Food", "Side of Cheese Food", "Small Drink Food", "Cheese Coney Food" and "Medium Drink Food" are sold together

Clustering on Iris Dataset:

At first 90% random sample of the iris data is scaled. Scaling helps in standardizing the data and reducing the effect of a single variable with higher influence. When clustering is done with k=3, the following division of clusters is seen in each cluster.

1	2	3
49	39	47

The clusters are plotted as follows:

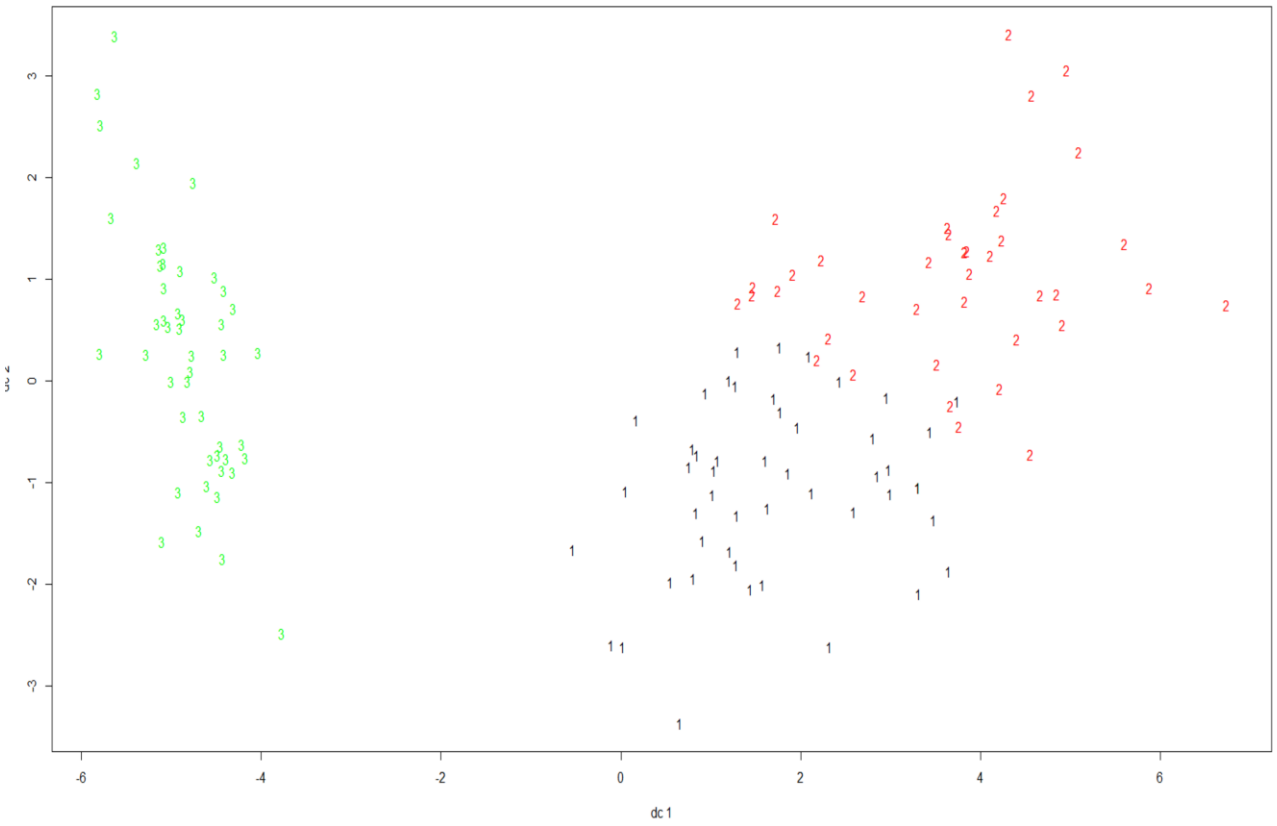

Fig 1: Plot of clusters – K-means clustering

When each cluster is analyzed following statistics are obtained:

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width

1  -0.01733468 -0.87460796    0.3732437   0.3215175
2   1.18772258  0.09653719    1.0437784   1.0556731
3  -0.96748470  0.83171999   -1.2552403  -1.2111832
```

Main drawback of K-means clustering is that the number of clusters must be specified before hand. To know the best number of clusters for the data, we plot a grid and determine the number of clusters. The Sum of squared error in each cluster must be the least in each cluster Using this point, the right number of clusters is identified to be equal to 4.

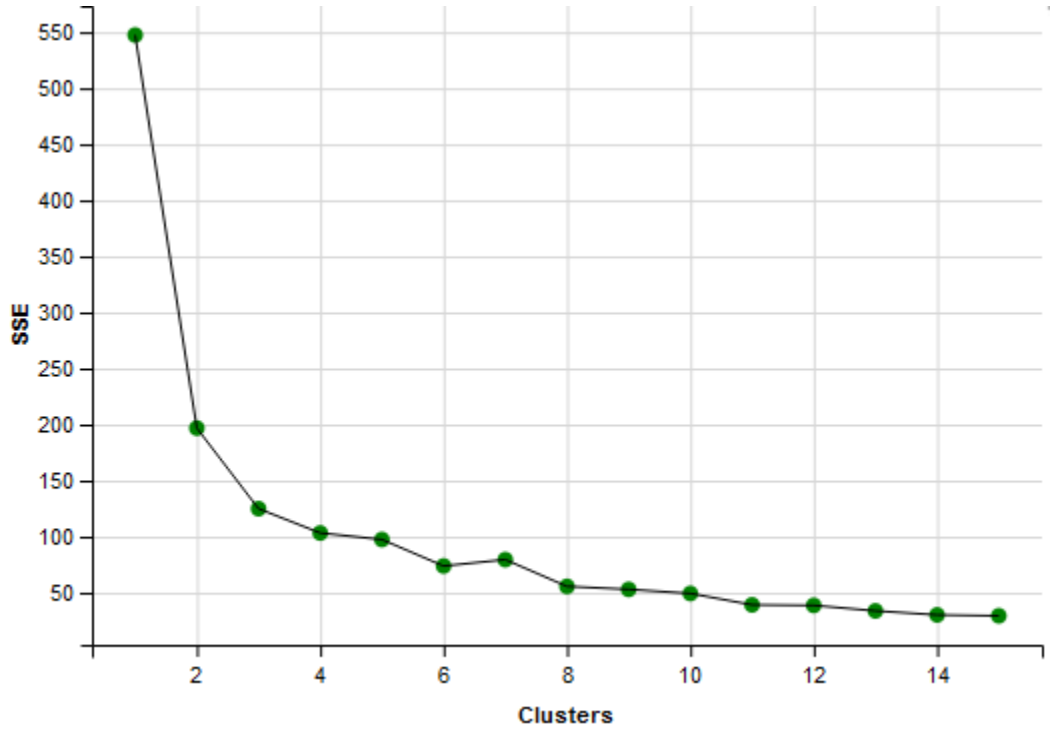

Fig 2: Plot showing intra-cluster variation vs. number of clusters

Plotting the clusters:

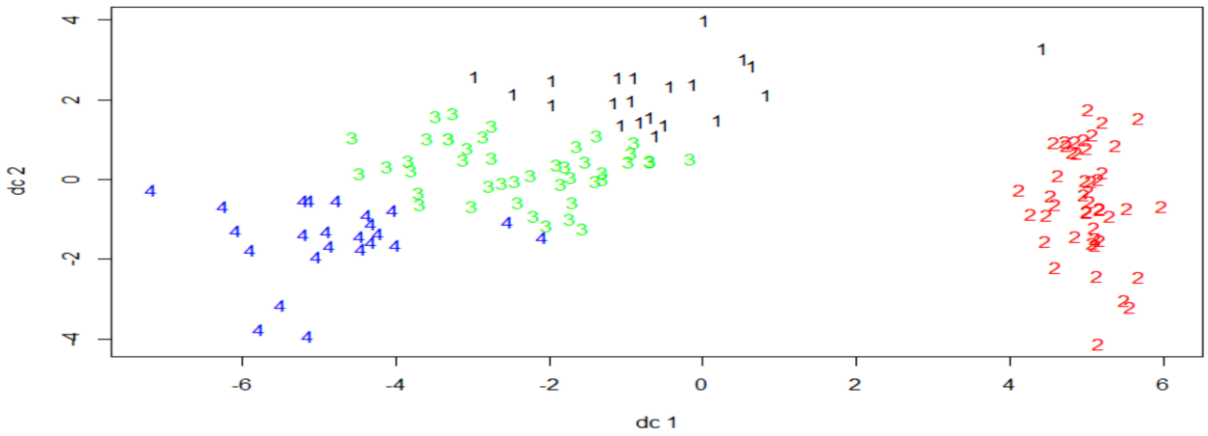

Fig 3: Plot of 4 clusters

Silhouette helps in understanding how well each object lies in a cluster.

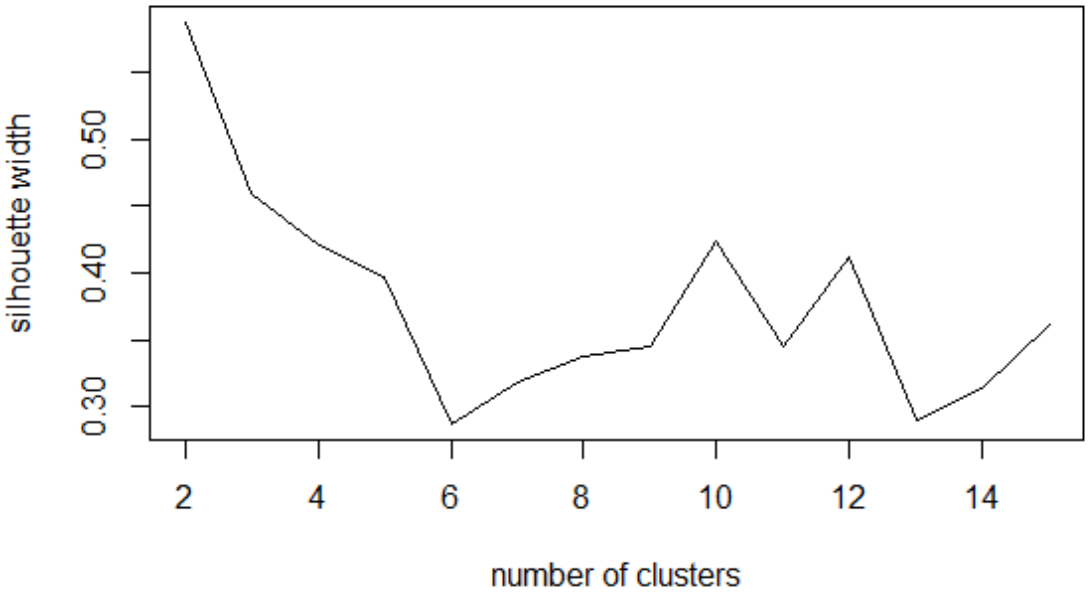

Fig 4: Silhoutte width plot

For each point p, first find the average distance between p and all other points in the same cluster (this is a measure of cohesion, call it A). Then find the average distance between p and all points in the nearest cluster (this is a measure of separation from the closest other cluster, call it B). The silhouette

coefficient for p is defined as th and A divided by the greater of the two (max(A,B)).

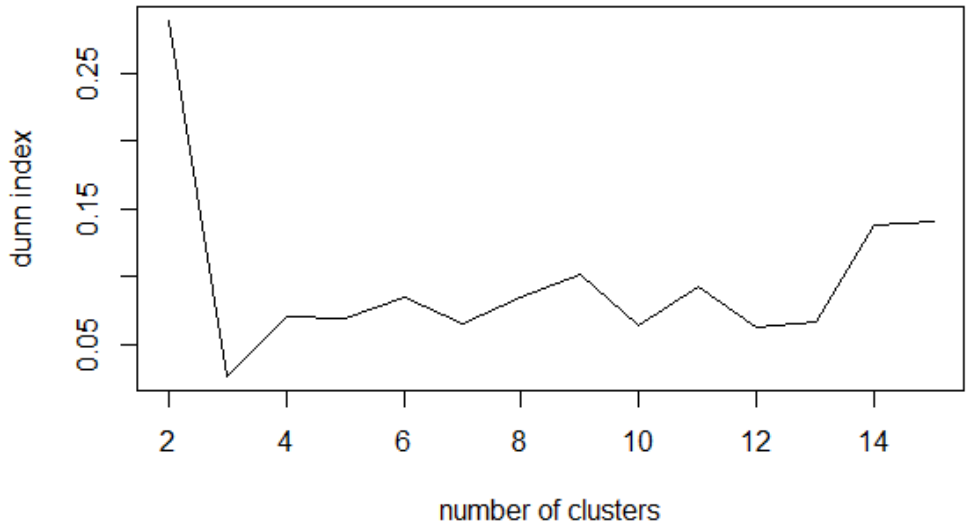

Fig 5: dunn index plot

In the dunn index, a higher dunn index value represents better clustering. i.e the best number of clusters obtained is 2.

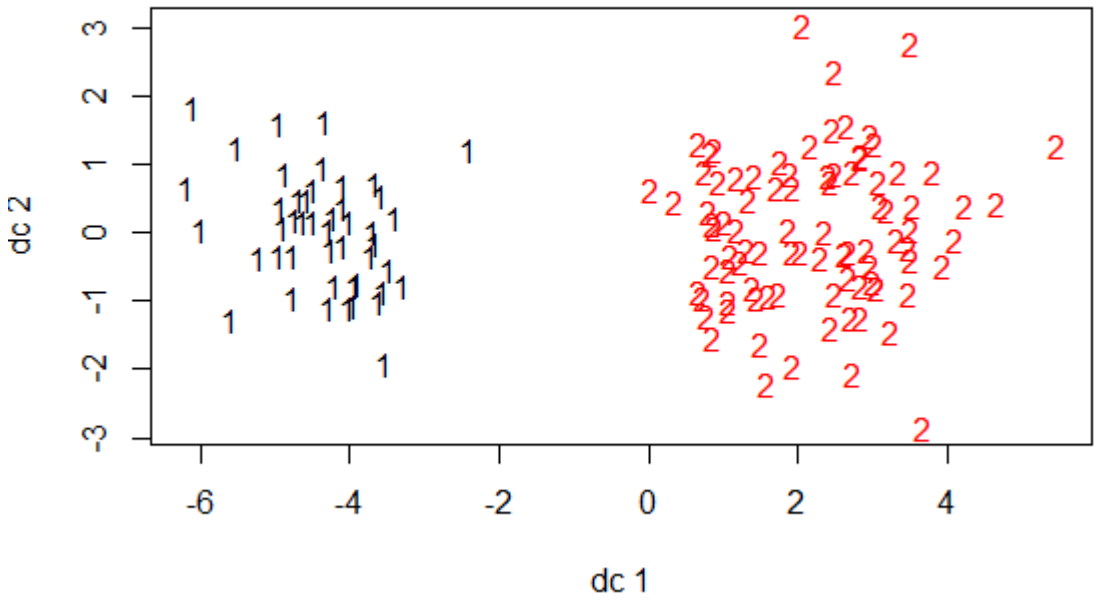

Fig 6

Based on this, count of observations in the two clusters is as follows:

1	2
89	46

Means in each cluster are as follows:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	0.5542	-0.3919	0.68256	0.658312
2	-1.00531	0.851012	-1.3025	-1.25116

Hierarchical clustering:

A distance matrix is formed and clusters are obtained using Ward method. Following Dendrogram is plotted .

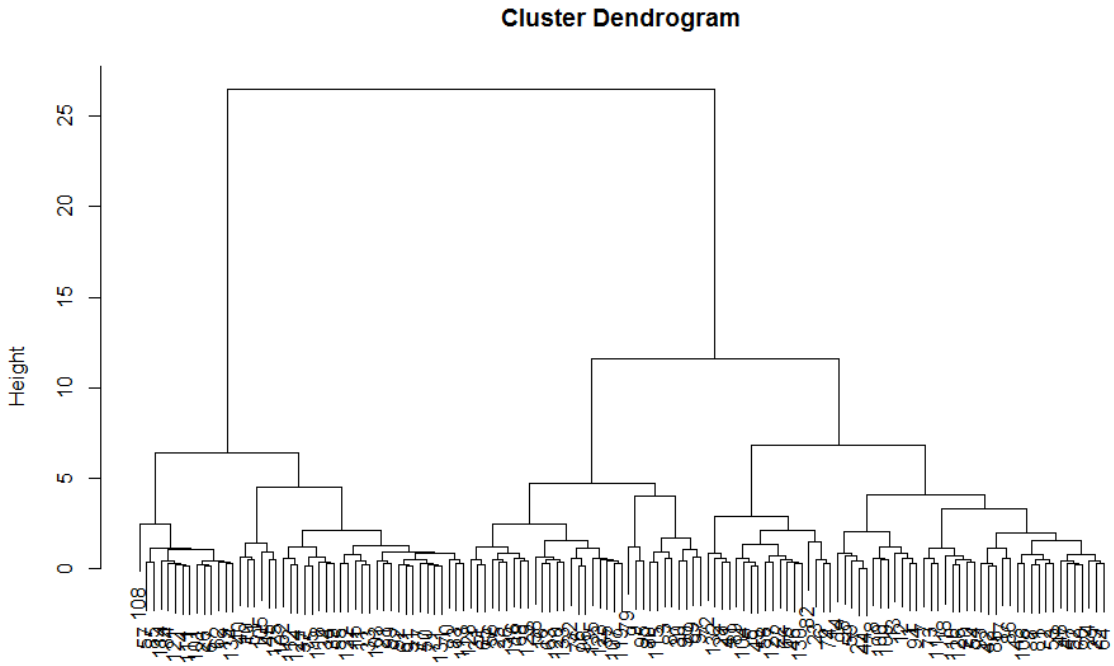

Fig 7

The obtained dendrogram is cut at 3 clusters level and cluster membership is obtained.

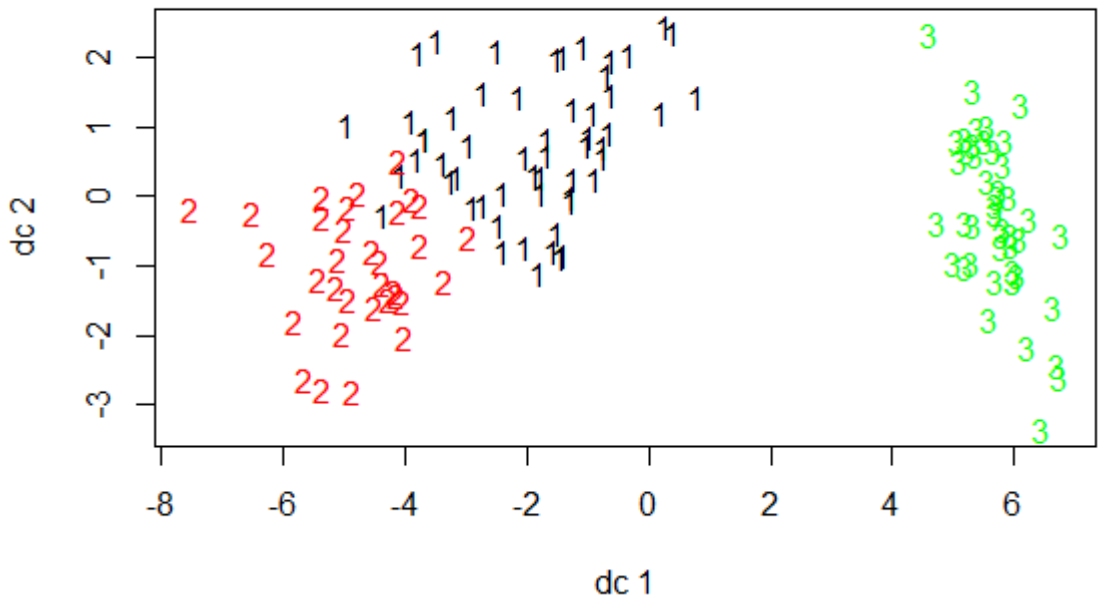

Fig 8

2.1 Cluster Analysis of Cincinnati Zoo Food Sales data:

A glimpse of the structure of the data:

```
##              NickName Oct..10 Nov..10 Dec..10 Jan..11 Feb..11 Mar..11
## 1              Cheese     343      66      99      37       4     105
## 2           Alchohol      131      79     232      12      18      49
## 3       Bottled Water    1448     410     577      59     165     507
## 4              Burger     188      86     103      19      40      73
## 5           Capri Sun      32       2       0       0       1       0
## 6 Cheese Fries Basket      37      55      59       3      33      65
```

K-means clustering:

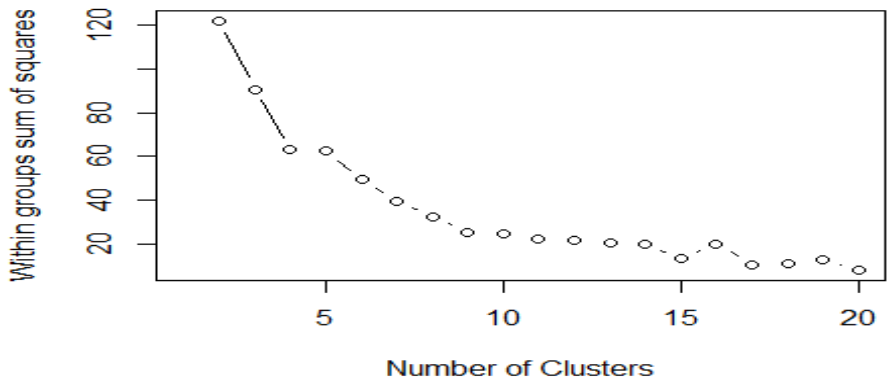

Fig 9: Plot of within cluster variance vs. number of cluster

From the above plot, it is evident that 8 is a better choice of k, the number of clusters.

A glimpse of how these clusters are separated:

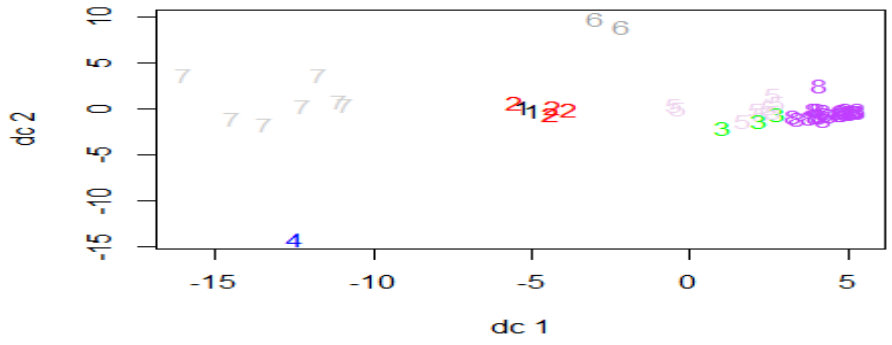

Fig 10: Plot of clusters – K-means Clustering

Here is an example of one of the clusters:

	NickName	Oct..10	Nov..10	Dec..10	Jan..11	Feb..11	Mar..11	cluster
7	Cheeseburger Basket	662	274	292	51	93	266	1
9	Chicken Tender Basket	395	299	298	61	132	298	1
21	Gatorade	744	237	332	53	94	271	1
26	Hot Dog Basket	778	279	230	35	106	242	1
47	Soft Pretzel	680	292	563	53	11	223	1

Mean of the purchases of clusters:

	Oct..10	Nov..10	Dec..10	Jan..11	Feb..11	Mar..11	cluster
1	651.80000	276.20000	343.000000	50.600000	87.200000	260.00000	1
2	538.00000	56.66667	7.333333	0.000000	0.000000	41.33333	2
3	1520.00000	397.00000	0.000000	0.000000	0.000000	758.00000	3
4	83.46154	24.92308	25.076923	2.923077	5.038462	20.42308	4
5	9.00000	260.50000	1028.500000	109.000000	27.500000	24.00000	5
6	195.22222	109.22222	130.555556	26.555556	51.777778	111.11111	6
7	1316.71429	472.14286	597.000000	92.714286	208.285714	518.42857	7
8	425.50000	323.00000	275.000000	80.000000	227.000000	445.50000	8

Hierarchical Clustering:

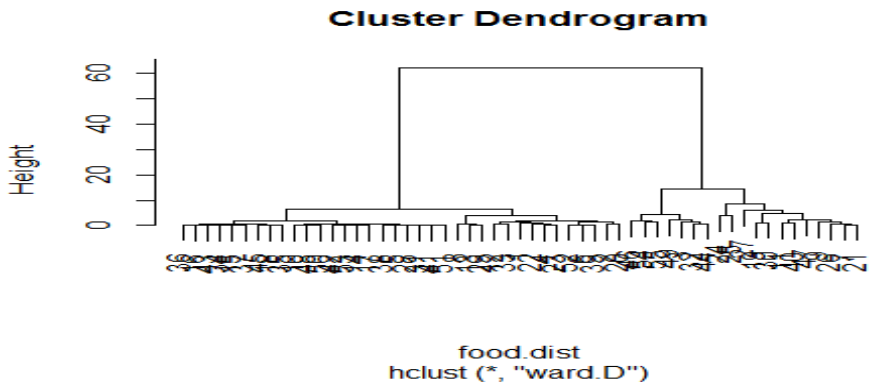

Fig 11: Cluster Dendrogram of Hierarchical clustering

Capping the number of clusters to 8:

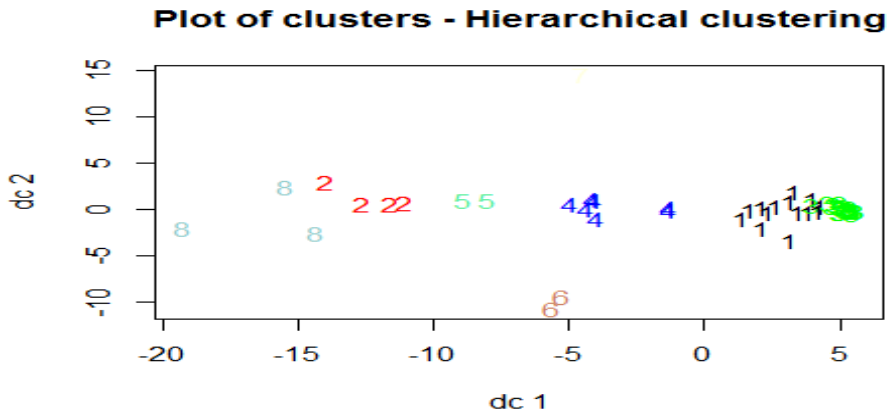

Fig 12: Plot of Clusters – Hierarchical Clustering

Here is a look at one of the clusters:

```
              NickName Oct..10 Nov..10 Dec..10 Jan..11 Feb..11 Mar..11 cluster
19 French Fries Basket      492      257      276       77      236      524       5
30       Krazy Kritter      359      389      274       83      218      367       5
                NickName Oct..10 Nov..10 Dec..10 Jan..11 Feb..11 Mar..11 cluster
7    Cheeseburger Basket      662      274      292       51       93      266       4
9  Chicken Tender Basket      395      299      298       61      132      298       4
11                 Chips      475      161      149       25       96      162       4
21             Gatorade      744      237      332       53       94      271       4
26        Hot Dog Basket      778      279      230       35      106      242       4
40       Sandwich Basket      188      212      180       41       97      181       4
47          Soft Pretzel      680      292      563       53       11      223       4
```

Hierarchical clustering gives more coherent clusters with items in same clusters varying in same manner in terms of monthly purchases.

```
Mean purchases of clusters:
```

```
11
```

```
##      Oct..10   Nov..10     Dec..10    Jan..11    Feb..11   Mar..11 cluster
## 1  201.28571  68.21429   93.071429  15.000000  25.285714  66.92857       1
## 2 1470.75000 448.00000  474.750000  71.750000 188.250000 451.25000       2
## 3   93.63636  21.50000    9.863636   1.772727   2.272727  17.04545       3
## 4  560.28571 250.57143  292.000000  45.571429  89.857143 234.71429       4
## 5  425.50000 323.00000  275.000000  80.000000 227.000000 445.50000       5
## 6    9.00000 260.50000 1028.500000 109.000000  27.500000  24.00000       6
## 7 1520.00000 397.00000    0.000000   0.000000   0.000000 758.00000       7
## 8 1111.33333 504.33333  760.000000 120.666667 235.000000 608.00000       8
```

Market Basket Analysis of Food items purchased in Cincinnati Zoo:

Here is a glimpse of the structure of the data:

```
## transactions as itemMatrix in sparse format with
##  19076 rows (elements/itemsets/transactions) and
##  118 columns (items) and a density of 0.02230729
##
## most frequent items:
##   Bottled.WaterFood Slice.of.CheeseFood    Medium.DrinkFood
##                3166                3072                2871
##     Small.DrinkFood   Slice.of.PeppFood             (Other)
##                2769                2354               35981
##
## element (itemset/transaction) length distribution:
## sizes
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   15
##  197 5675 5178 3253 2129 1293  655  351  178   95   42   14    8    7    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.632   4.000  15.000
##
## includes extended item information - examples:
##              labels
## 1    Add.CheeseFood
## 2          BeerFood
## 3 Bottled.WaterFood
```

There are 197 transactions with zero items. Majority of the transactions contain 1 to 5 items. After removing transactions with zero items as they are of no use in the following analysis, we have 18879 transactions.

Here is a glimpse of the frequency plot of the items with minimum support of 0.05:

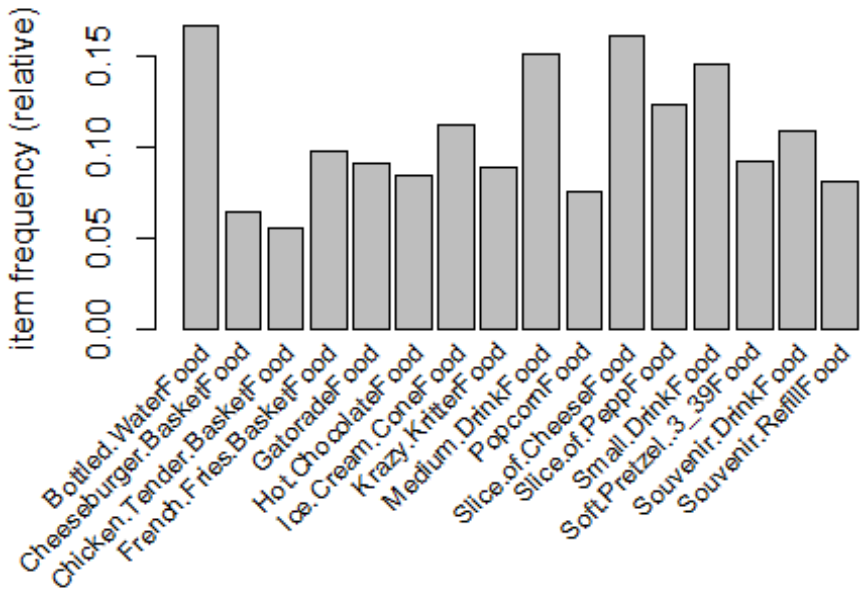

Fig 13

Applying appriori algorithm to find association rules with minimum support of 0.001, confidence of 0.5 and lift ratio of 1.2:

```
## set of 1812 rules
##
## rule length distribution (lhs + rhs):sizes
##    2    3    4
##  217 1150  445
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   3.000   3.126   3.000   4.000
##
## summary of quality measures:
##     support           confidence          lift
##  Min.   :0.001006   Min.   :0.2000   Min.   : 1.193
##  1st Qu.:0.001218   1st Qu.:0.2424   1st Qu.: 1.904
##  Median :0.001589   Median :0.3015   Median : 2.543
##  Mean   :0.002708   Mean   :0.3566   Mean   : 4.192
##  3rd Qu.:0.002595   3rd Qu.:0.3925   3rd Qu.: 4.507
##  Max.   :0.061497   Max.   :1.0000   Max.   :76.722
##
## mining info:
##  data ntransactions support confidence
##   dat         18879   0.001        0.2

##     lhs                             rhs                              support confidence      lift
## [1] {Souvenir.Sierra.MistFood}  => {Chicken.Nugget.BasketFood} 0.001218285  0.4693878 12.569605
## [2] {Souvenir.Dr.PepperFood}    => {Chicken.Nugget.BasketFood} 0.001324223  0.4901961 13.126825
## [3] {Medium.Diet.Dr.PepperFood} => {Chicken.Nugget.BasketFood} 0.001059378  0.4255319 11.395201
## [4] {Soft.Pretzel..3_89Food}    => {Add.CheeseFood}            0.001271254  0.3243243 11.684960
## [5] {Soft.Pretzel..3_89Food}    => {Bottled.WaterFood}         0.001165316  0.2972973  1.772797
## [6] {ChiliFood}                 => {French.Fries.BasketFood}   0.001006409  0.4750000  4.816071
```

There are very few rules with a good value of support. Majority of the rules have support less than 0.03. This can be expected with the high number of possible combinations of items with varied sizes. Majority of the rules have confidence greater than 0.3, very few of them have confidence greater than 0.5 (which is not a good measure of relaibility) and majority of the rules have lift ratio greater than 4 (which is a good thing, rules are useful). Filtering rules with confidence greater than 0.7.

```
## set of 119 rules
##
## rule length distribution (lhs + rhs):sizes
##  2  3  4
##  6 71 42
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   3.000   3.303   4.000   4.000
##
## summary of quality measures:
##     support           confidence          lift
##  Min.   :0.001006   Min.   :0.7037   Min.   : 5.077
##  1st Qu.:0.001218   1st Qu.:0.8146   1st Qu.: 8.510
##  Median :0.001483   Median :0.8889   Median : 8.872
##  Mean   :0.002519   Mean   :0.8844   Mean   :10.102
##  3rd Qu.:0.002384   3rd Qu.:0.9649   3rd Qu.: 9.459
##  Max.   :0.028868   Max.   :1.0000   Max.   :30.175
##
## mining info:
##  data ntransactions support confidence
##   dat         18879   0.001        0.2
```

There are 119 rules with confidence greater than 0.7. Lift ratio of majority of the rules is greater than 7. We lost some rules with better lift ratios than 30 (which is the maximum in this basket of rules). Let's have a look at the rules with lift ratio greater than 20.

```
##     lhs                       rhs                    support confidence     lift
## [1] {Side.of.CheeseFood}  => {Hot.DogFood}      0.006356269  0.9230769 21.38254
## [2] {Cheese.ConeyFood,
##      Side.of.CheeseFood}  => {Hot.DogFood}      0.004396419  0.9325843 21.60277
## [3] {Side.of.CheeseFood,
##      Small.DrinkFood}     => {Cheese.ConeyFood} 0.001536098  0.8055556 30.17477
## [4] {Side.of.CheeseFood,
##      Souvenir.RefillFood} => {Hot.DogFood}      0.001059378  0.9523810 22.06135
## [5] {Side.of.CheeseFood,
##      Small.DrinkFood}     => {Hot.DogFood}      0.001853912  0.9722222 22.52096
## [6] {Medium.DrinkFood,
##      Side.of.CheeseFood}  => {Hot.DogFood}      0.002065787  0.9285714 21.50982
## [7] {Cheese.ConeyFood,
##      Side.of.CheeseFood,
##      Small.DrinkFood}     => {Hot.DogFood}      0.001483129  0.9655172 22.36564
## [8] {Cheese.ConeyFood,
##      Medium.DrinkFood,
##      Side.of.CheeseFood}  => {Hot.DogFood}      0.001430160  0.9642857 22.33712
```

There are 8 rules with lift ratio greater than 20 and confidence greater than 0.8. Graphical visualization of these 8 major rules:

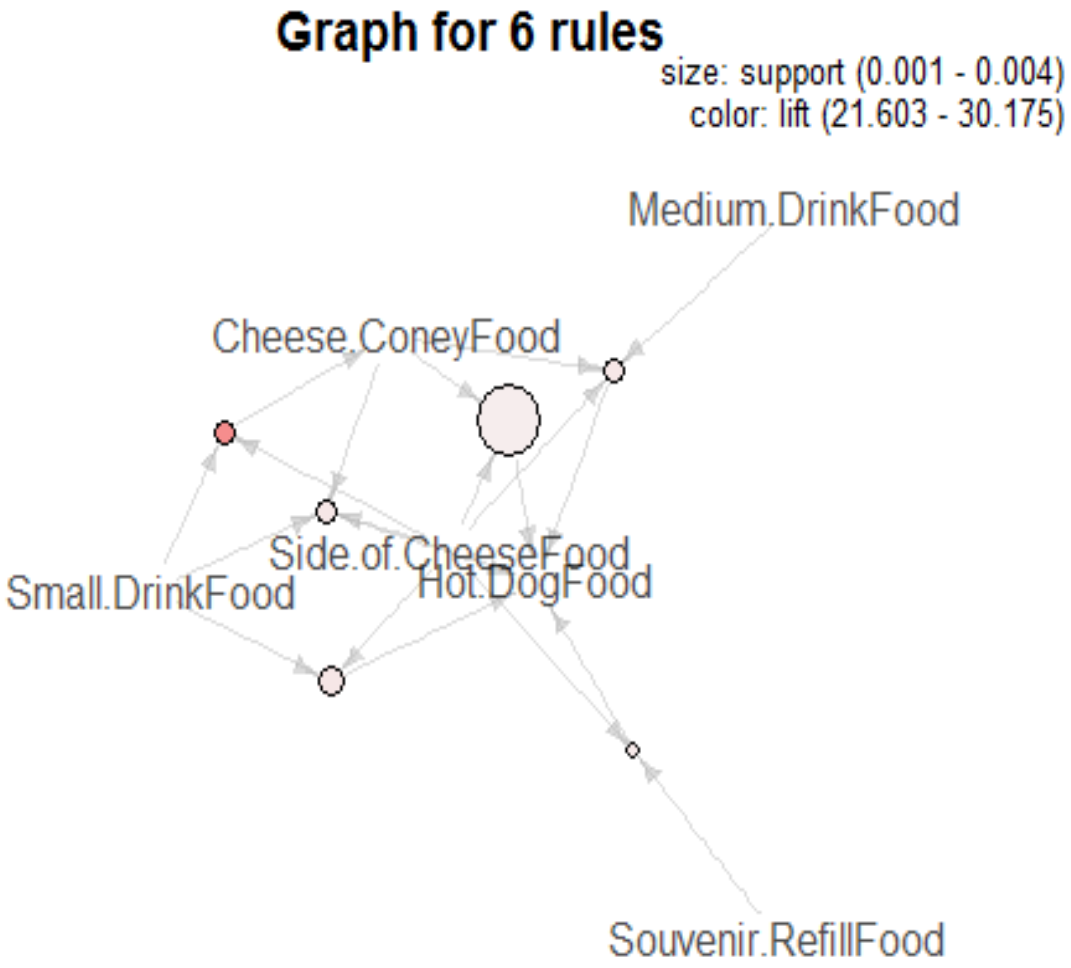

Fig 14: Plot of 6 major rules

These 6 rules are reliable (confidence approaching 1).

Rule with maximum lift ratio (76) is as follows:

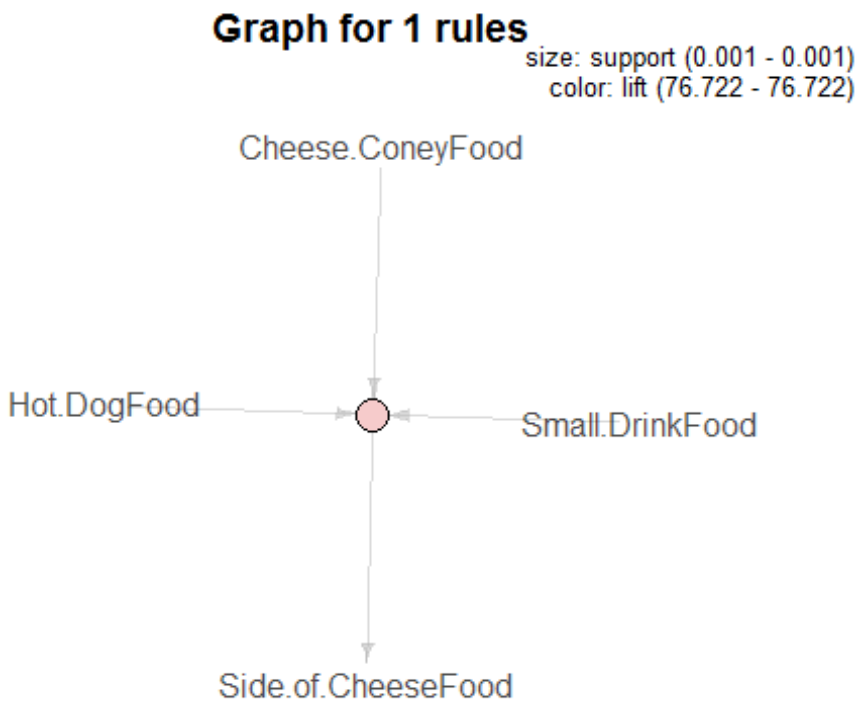

Fig 15: Plot of rule with highest lift ratio

Almost all the major rules involve one or the other items of: "Hot Dog Food", "Side of Cheese Food", "Small Drink Food", "Cheese Coney Food" and "Medium Drink Food".