

Data mining (Itm6285)

ASSIGNMENT 3



March 10, 2017

ANANYA AUDDY

NET ID: PY7552

Objective: To identify food product clusters based on nutritional content

Dataset: The Product2.csv file consists of data on food products sold in the US from openfoodfacts.org



Before we start our analysis, we need to get familiarized with the data set.

Let’s explore before creating any clustering model - The dataset has 1463 unique products and 48 features. We create a subset of data – considering all those data having less than 26 NA’s.

The subset has 60 rows. We clean the dataset by removing the duplicate products and replacing the missing values with 0.

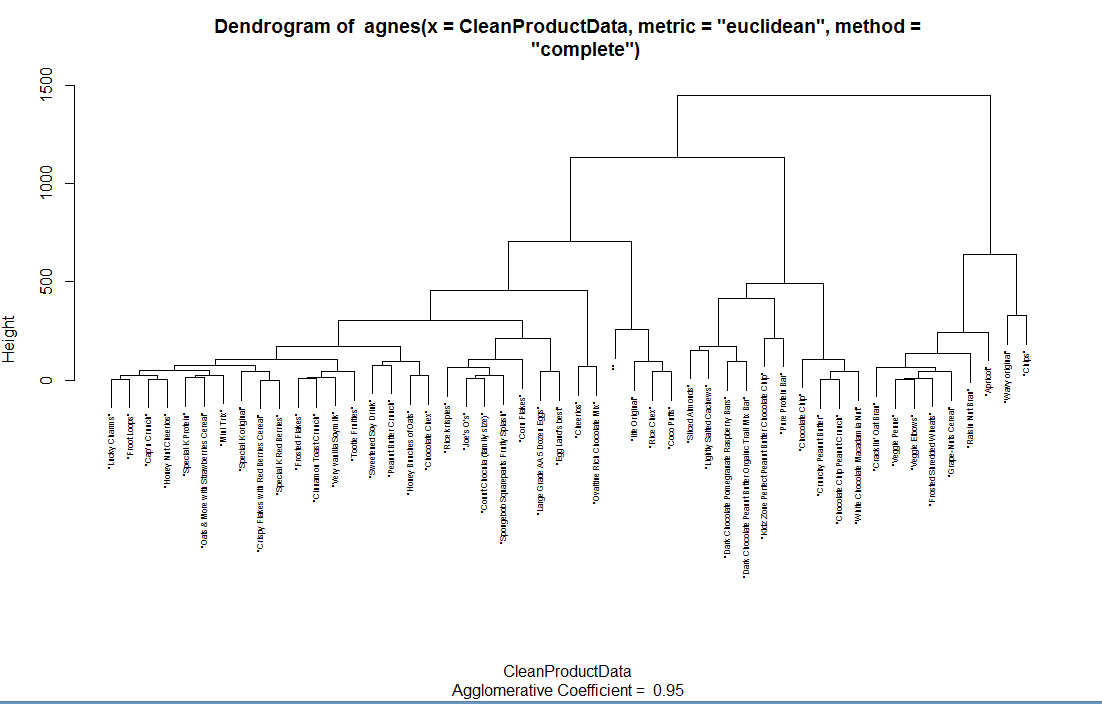
We also set the row names of the data frame as the product names.

After cleaning we are left with 50 products. We also scaled the data,

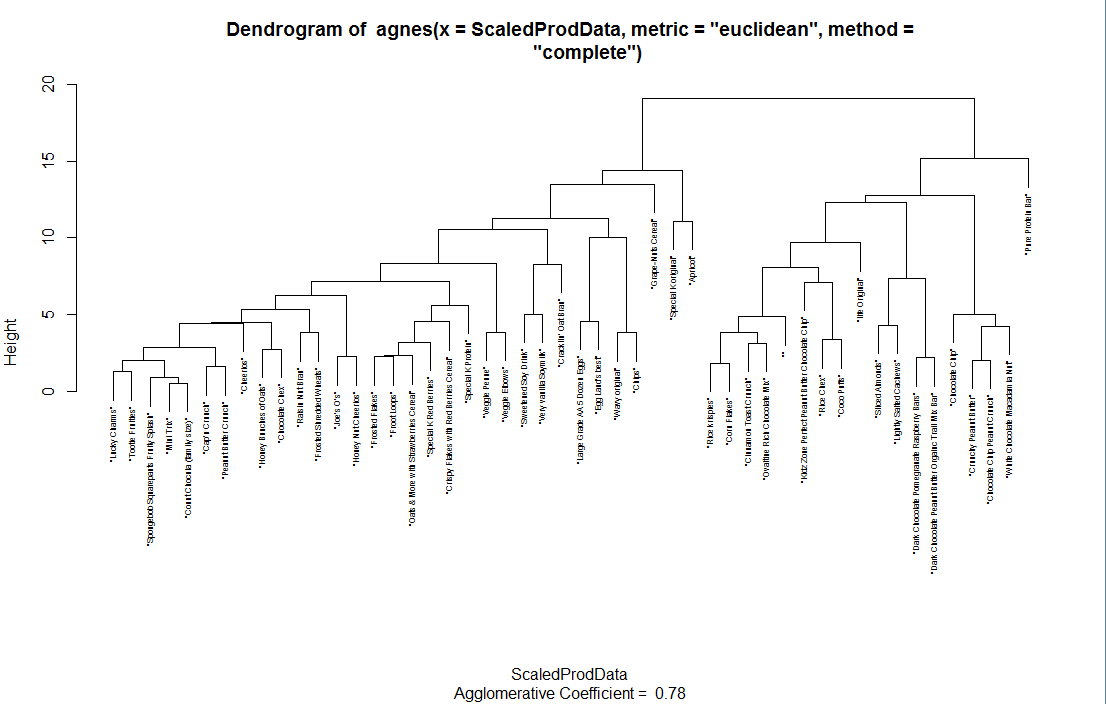
The clustering model would be using the features to build clusters to achieve internal cohesion and external separation.

## Hierarchical Agglomerative Model (Agnes)

Dendogram using unscaled data:

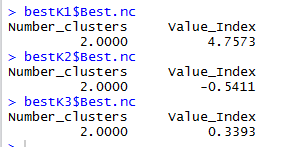


Dendogram using scaled data:

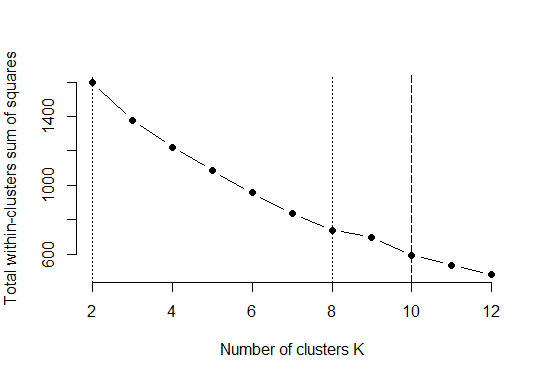


## Optimal Cluster Number & K-Means

Optimal Cluster number is 2. All 3 indices- “kl”, “gap” and “silhouette” suggest that the optimal cluster should be 2. Moreover, from the above dendogram on the scaled data also suggest that the data can be broadly classified into 2 categories.



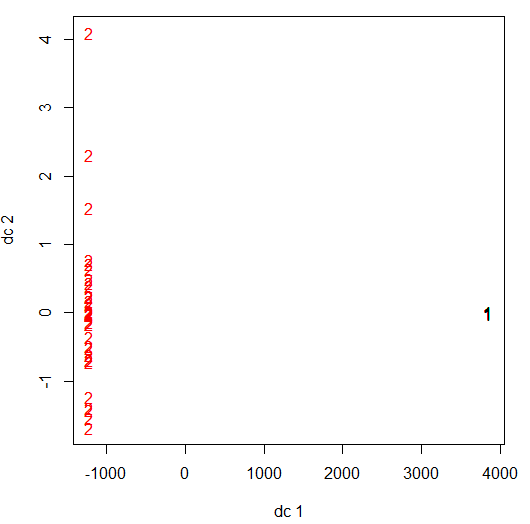
However, from the elbow method, we see the knee bend is at 2,8 and 10. We would be doing K-means clustering with all three and choose the best results.



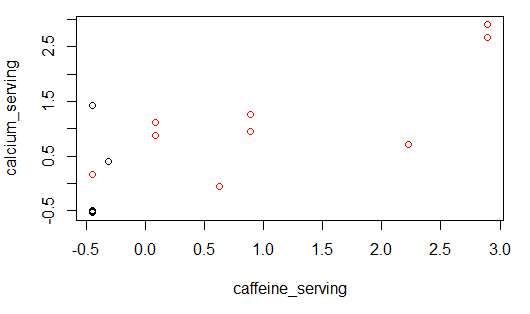
**K-Means Clustering using # of clusters as 2:**

|  |  |
| --- | --- |
| **Cluster2** | **Cluster1** |
| Lucky Charms | Sliced Almonds |
| Rice krispies | Kidz Zone Perfect Peanut Butter Chocolate Chip |
| Frosted Flakes | Lightly Salted Cashews |
| Froot Loops | life Original |
| Cap'n Crunch | Dark Chocolate Pomegranate Raspberry Bars |
| Special K Protein | Rice Chex |
| Sweetened Soy Drink | Pure Protein Bar |
| Peanut Butter Crunch | Dark Chocolate Peanut Butter Organic Trail Mix Bar |
| Oats & More with Strawberries Cereal | Chocolate Chip |
| Joe's O's | Crunchy Peanut Butter |
| Honey Bunches of Oats | Chocolate Chip Peanut Crunch |
| Cracklin' Oat Bran | White Chocolate Macadamia Nut |
| Cinnamon Toast Crunch | Coco Puffs |
| Large Grade AA 5 Dozen Eggs |  |
| Veggie Penne |  |
| Spongebob Squarepants Fruity Splash |  |
| Wavy original |  |
| Cheerios |  |
| Ovaltine Rich Chocolate Mix |  |
| Special K original |  |
| Very vanilla Soymilk |  |
| Raisin Nut Bran |  |
| Crispy Flakes with Red Berries Cereal |  |
| Mini Trix |  |
| Veggie Elbows |  |
| Special K Red Berries |  |
| Honey Nut Cheerios |  |
| Frosted Shredded Wheats |  |
| Grape-Nuts Cereal |  |
| Chips |  |
| Corn Flakes |  |
| Count Chocula (family size) |  |
| Tootie Fruities |  |
| Apricot |  |
| Egg Land's best |  |
| Chocolate Chex |  |
| “” (unknown product) |  |

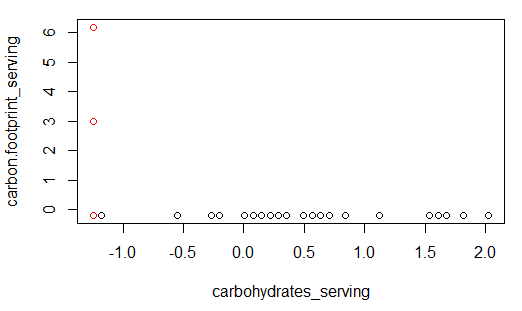
Overall Plot with 2 clusters:



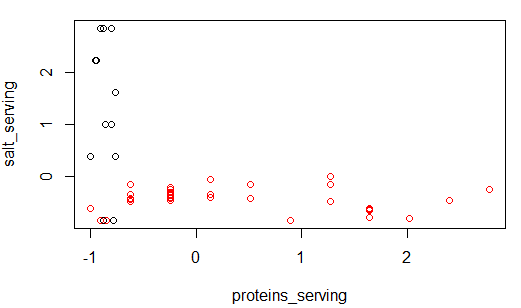
Plotting the product clusters based on its caffeine and calcium serving



Plotting the product clusters based on its carbohydrate and carbon footprint serving



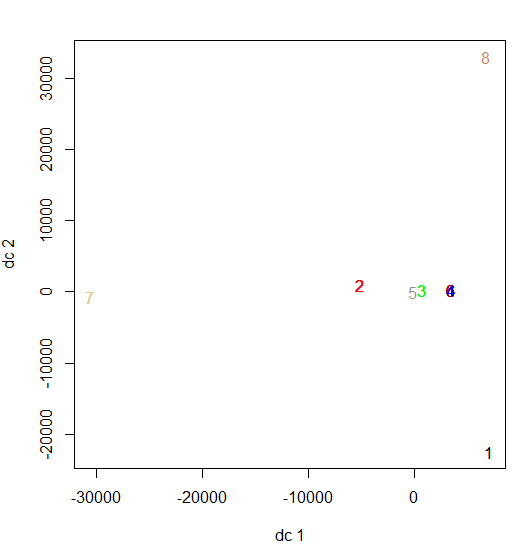
Plotting the product clusters based on its protein and salt serving



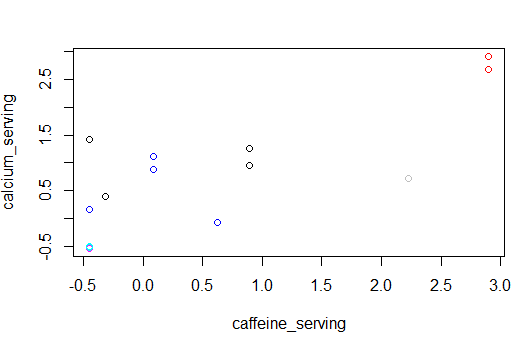
**K-Means Clustering using # of clusters as 8:**

|  |  |
| --- | --- |
| **Clusters** | **Products** |
| Cluster 1 | “” (Unknown) |
| Kidz Zone Perfect Peanut Butter Chocolate Chip |
| life Original |
| Rice Chex |
| Coco Puffs |
| Cluster 2 | Chocolate Chip |
| Crunchy Peanut Butter |
| Chocolate Chip Peanut Crunch |
| White Chocolate Macadamia Nut |
| Cluster 3 | Veggie Penne |
| Veggie Elbows |
| Grape-Nuts Cereal |
| Cluster 4 | Sliced Almonds |
| Lightly Salted Cashews |
| Dark Chocolate Pomegranate Raspberry Bars |
| Dark Chocolate Peanut Butter Organic Trail Mix Bar |
| Cluster 5 | Lucky Charms |
| Rice krispies |
| Frosted Flakes |
| Froot Loops |
| Cap'n Crunch |
| Special K Protein |
| Peanut Butter Crunch |
| Oats & More with Strawberries Cereal |
| Joe's O's |
| Honey Bunches of Oats |
| Cracklin' Oat Bran |
| Cinnamon Toast Crunch |
| Spongebob Squarepants Fruity Splash |
| Cheerios |
| Ovaltine Rich Chocolate Mix |
| Very vanilla Soymilk |
| Raisin Nut Bran |
| Crispy Flakes with Red Berries Cereal |
| Mini Trix |
| Special K Red Berries |
| Honey Nut Cheerios |
| Frosted Shredded Wheats |
| Corn Flakes |
| Count Chocula (family size) |
| Tootie Fruities |
| Chocolate Chex |
| Cluster 6 | Sweetened Soy Drink |
| Large Grade AA 5 Dozen Eggs |
| Wavy original |
| Chips |
| Egg Land's best |
| Cluster 7 | Special K original |
| Apricot |
| Cluster 8 | Pure Protein Bar |

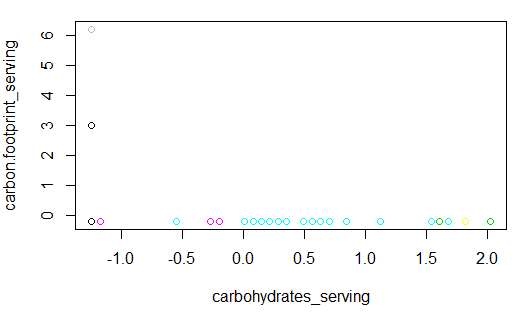
Overall Plot using 8 clusters:



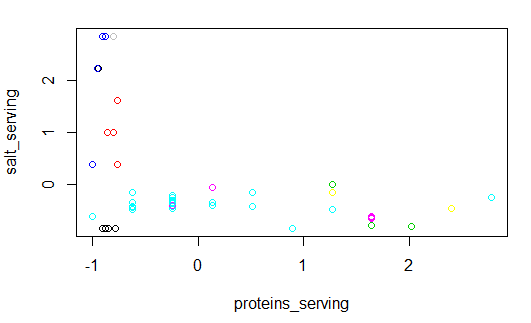
Plotting the product clusters based on its caffeine and calcium serving



Plotting the product clusters based on its carbohydrate and carbon footprint serving



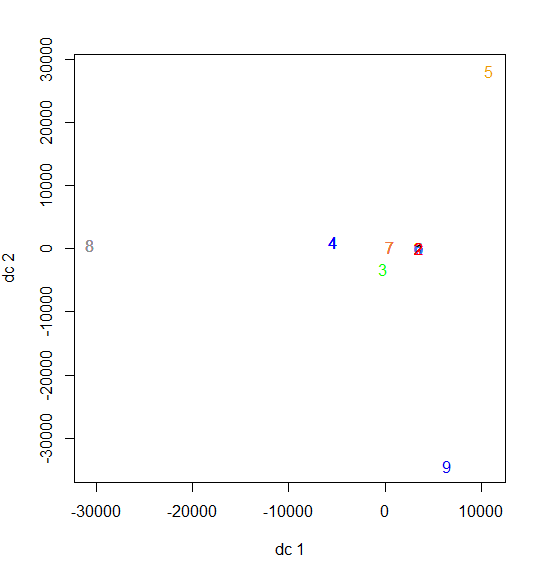
Plotting the product clusters based on its protein and salt serving



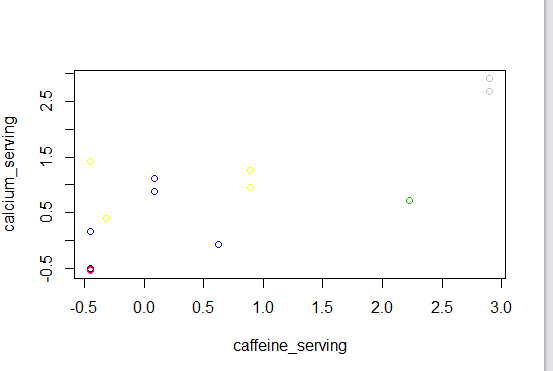
**K-Means Clustering using # of clusters as 10:**

|  |  |
| --- | --- |
| **Clusters** | **Products** |
| Cluster 1 | Sweetened Soy Drink |
| Cracklin' Oat Bran |
| Very vanilla Soymilk |
| Cluster 2 | Lucky Charms |
| Froot Loops |
| Cap'n Crunch |
| Peanut Butter Crunch |
| Joe's O's |
| Honey Bunches of Oats |
| Cinnamon Toast Crunch |
| Spongebob Squarepants Fruity Splash |
| Cheerios |
| Ovaltine Rich Chocolate Mix |
| Raisin Nut Bran |
| Mini Trix |
| Honey Nut Cheerios |
| Frosted Shredded Wheats |
| Count Chocula (family size) |
| Tootie Fruities |
| Chocolate Chex |
| Cluster 3 | Pure Protein Bar |
| Cluster 4 | Sliced Almonds |
| Lightly Salted Cashews |
| Dark Chocolate Pomegranate Raspberry Bars |
| Dark Chocolate Peanut Butter Organic Trail Mix Bar |
| Cluster 5 | Apricot |
| Cluster 6 | Large Grade AA 5 Dozen Eggs |
| Wavy original |
| Chips |
| Egg Land's best |
| Cluster 7 | Kidz Zone Perfect Peanut Butter Chocolate Chip |
| “” (Unknown) |
| life Original |
| Rice Chex |
| Coco Puffs |
| Cluster 8 | Chocolate Chip |
| Crunchy Peanut Butter |
| Chocolate Chip Peanut Crunch |
| White Chocolate Macadamia Nut |
| Cluster 9 | Grape-Nuts Cereal |
| Cluster 10 | Rice krispies |
| Frosted Flakes |
| Special K Protein |
| Oats & More with Strawberries Cereal |
| Veggie Penne |
| Special K original |
| Crispy Flakes with Red Berries Cereal |
| Veggie Elbows |
| Special K Red Berries |
| Corn Flakes |

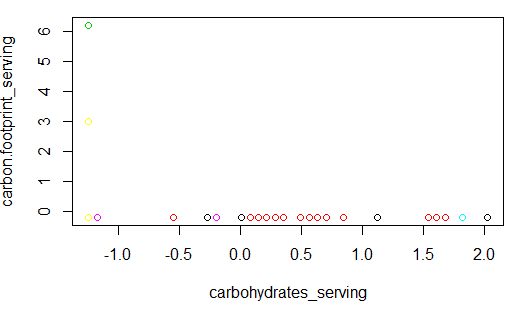
Overall Plot using 10 clusters:



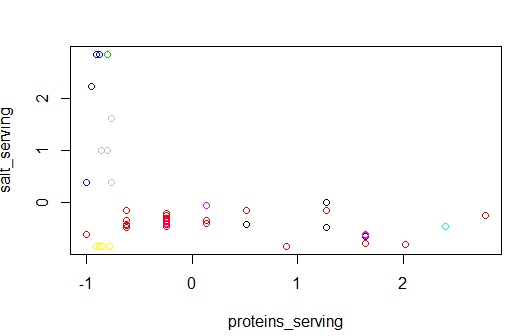
Plotting the product clusters based on its caffeine and calcium serving



Plotting the product clusters based on its carbohydrate and carbon footprint serving



Plotting the product clusters based on its protein and salt serving



Clustering Measures:

|  |  |  |
| --- | --- | --- |
| **Clusters** | **Total Withins** | **Betweens** |
| 2 Clusters | 1597.794 | 656.2062 |
| 8 Clusters | 740.9548 | 1513.045 |
| 10 Clusters | 605.4501 | 1648.55 |

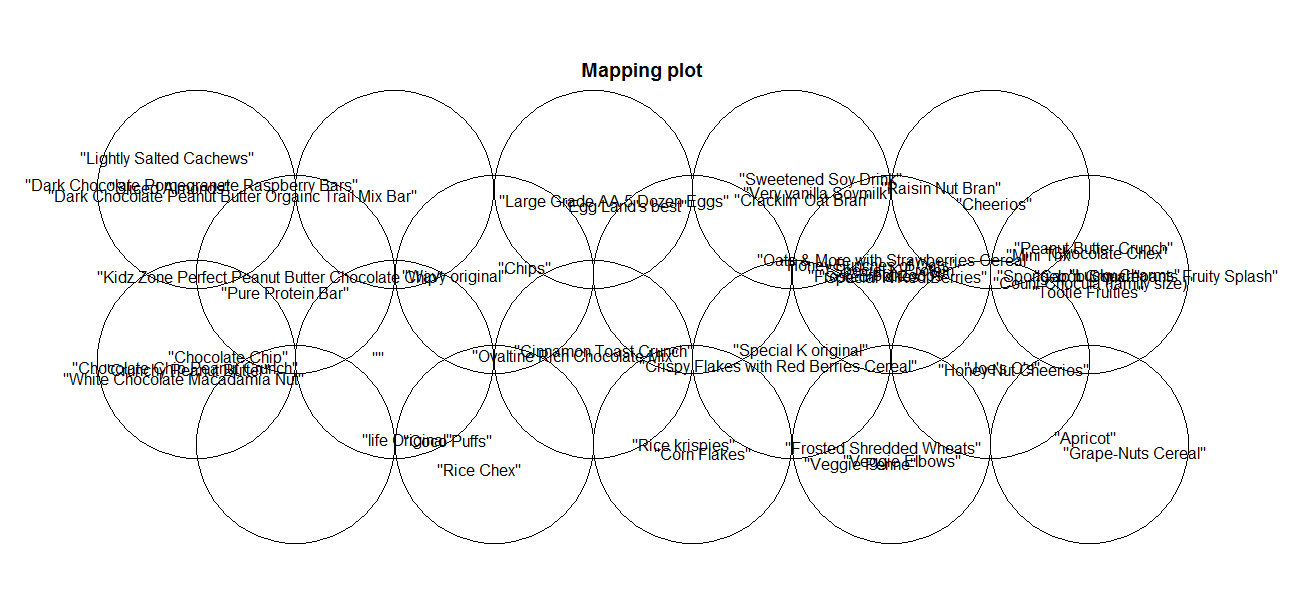
## Optimum Cluster # : 2

The dendogram, nbclust function and elbow method all states 2 as the optimum cluster. Moreover, the overall plot suggests 2 as the best number of clusters- the 2 groups seem to have internal cohesion and external separation.

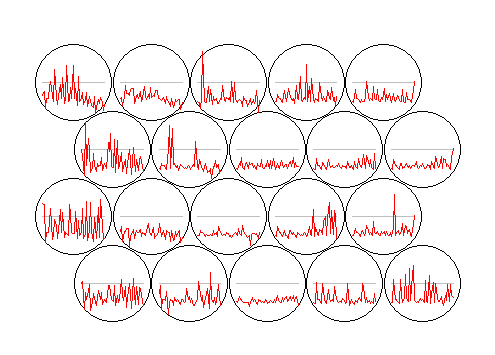
The overall plot of 8 and 10 clusters clearly shows, that the clusters were created forcibly. They seem to overlap each other, thus creating unnecessary groups.

Though Clustering into 8 and 10 groups helps in reducing the total sum of squares within further but it creates cluster having single products – which is not ideal.

## Kohonen SOM

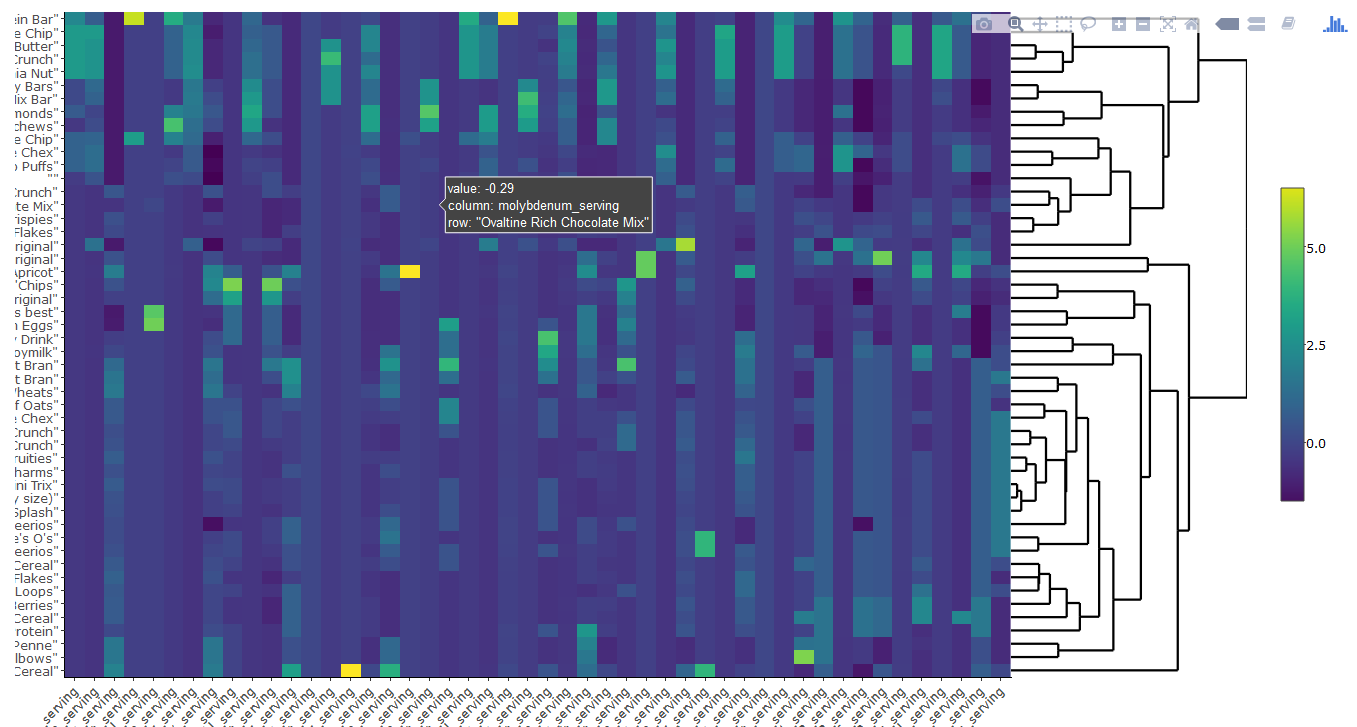


**Code Plot**

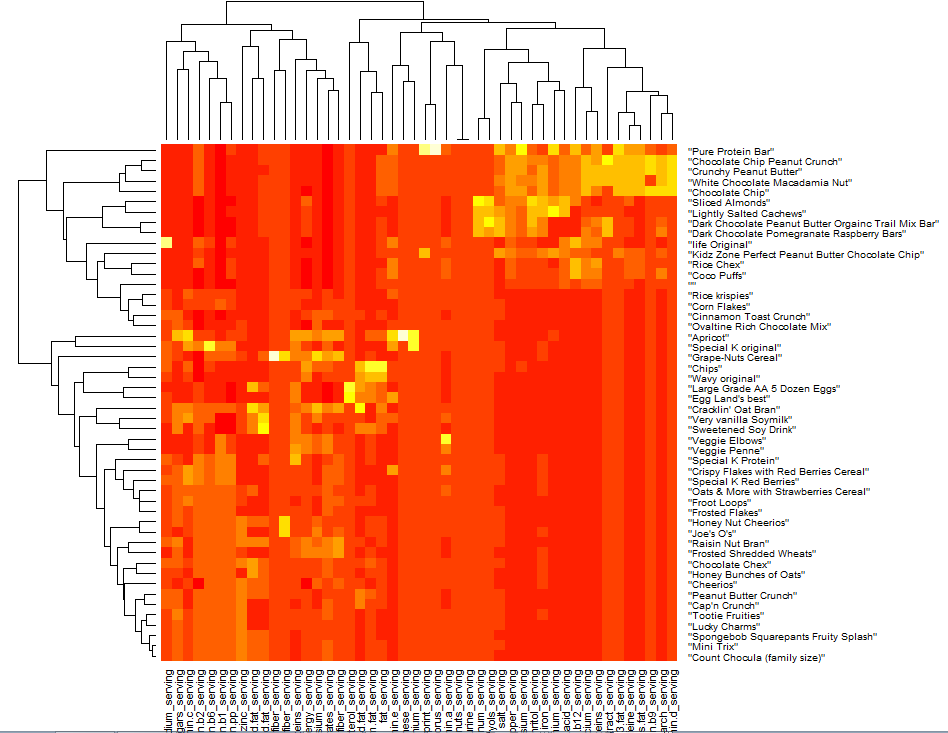


## Heat Map

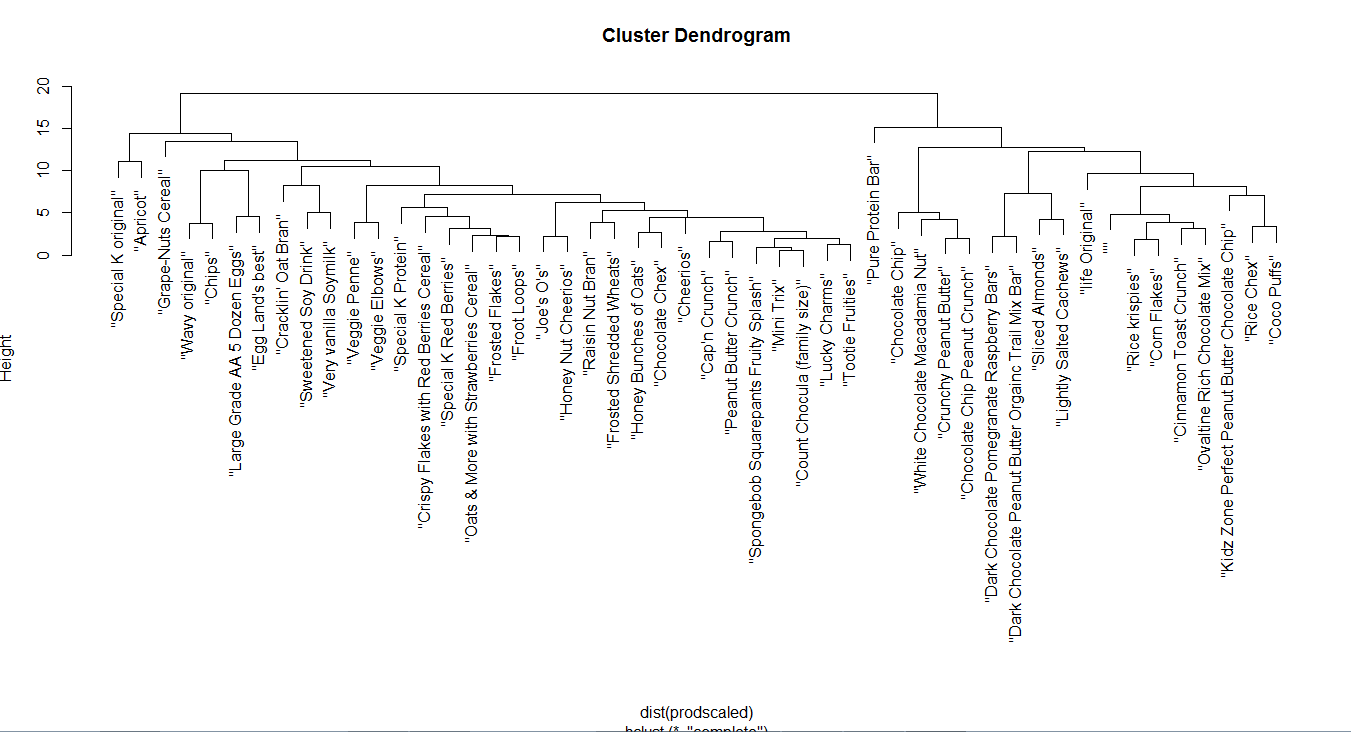
Dynamic Heat Map:



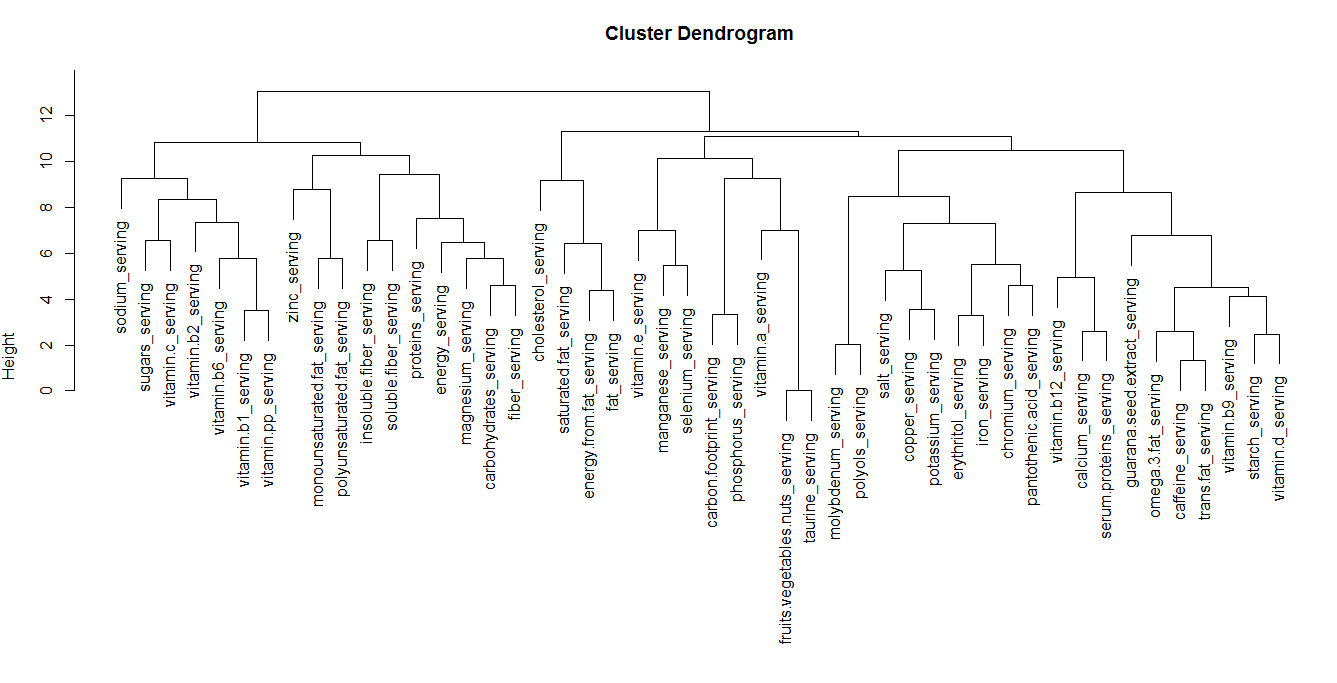
Static Heat Map:



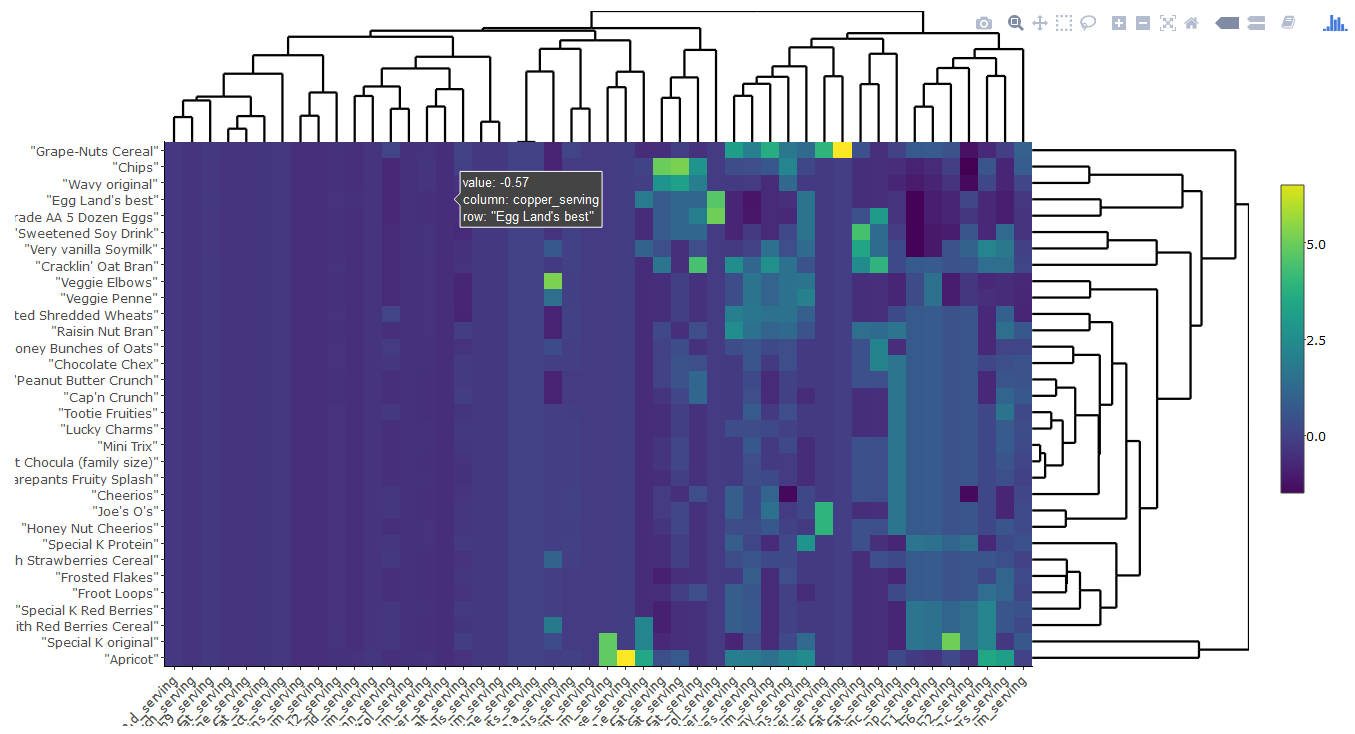
Clustering Rows:



Clustering Columns:



HeatMap for the 1st Cluster:



HeatMap for the 2nd Cluster:

