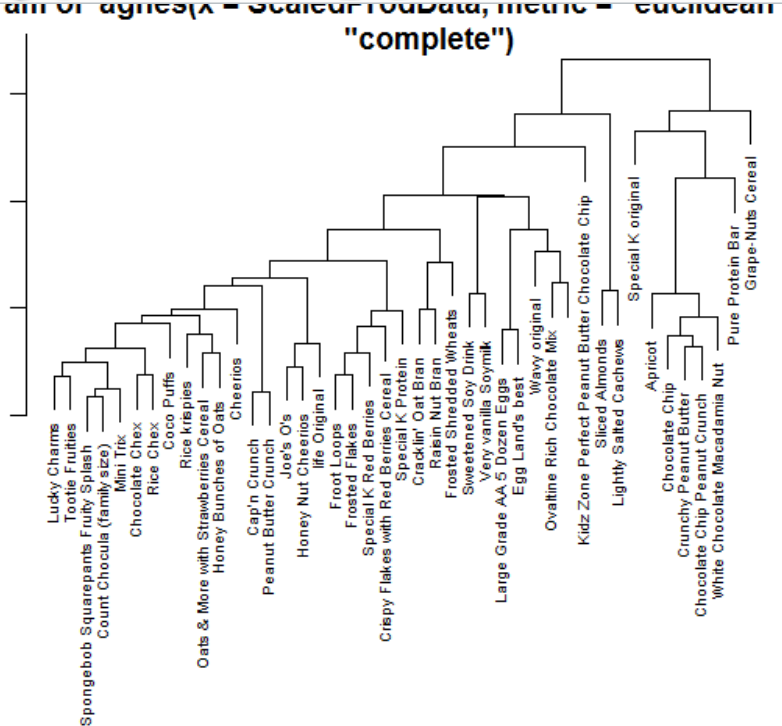
**Data Clustering-Food Products**

-Vaishnavi Ayyadurai

Product.csv is preprocessed using the following methods:

* Prod Id is made NULL
* Subset of product data that has less than 25 NA’s are pulled.
* Duplicate Records are removed and NA’s are made ‘0’.
* Product names are set as rownames.
* Distance matrix is calculated.

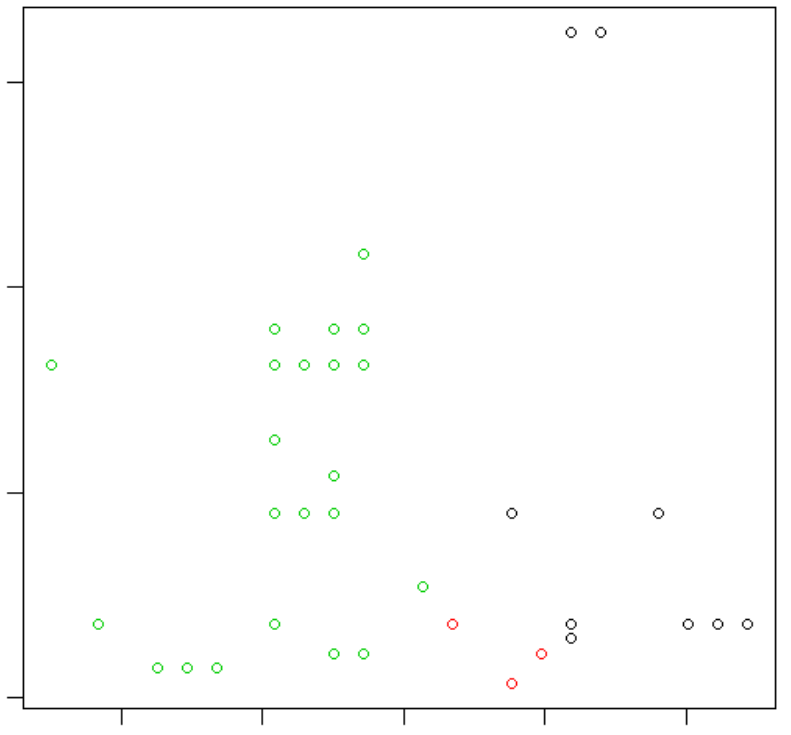
1. **HIERARCHIAL**-**AGNES:**



* Cereals, oats are grouped into one category, while chocolate products are grouped together overall.
* On drilling down the dendrogram, berrie cereals and protein bars are grouped in one category and while products with milk and eggs and rice are grouped in other. This seems to be a pretty decent segmentation of data.

1. **K-MEANS:**

**K(No of Clusters)=3**

***Energy serving and Iron serving*** :

# dist between each point with the cluster

[1] 230026.69 30482.44 463395.18

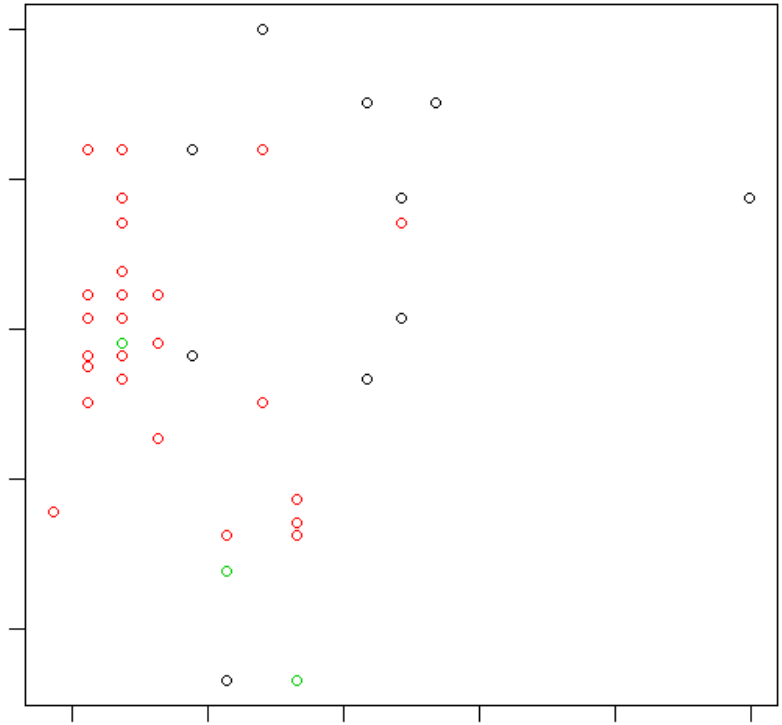
# total distance in the above

[1] 723904.3

## dist between the clusters

[1] 2417408

**For protein serving & Salt Serving:**



**Optimal Number of clusters:**

Index=KL

bestK1$All.index:

2 3 4 5 6 7 8 9 10 11

0.7116 3.5877 0.7177 2.2385 0.5777 1.5122 1.9209 0.3650 1.0305 2.4117

12 13 14 15

2.0322 0.3313 2.8111 0.5796

bestK1$Best.nc:

Number\_clusters Value\_Index

3.0000 3.5877

Index=GAP

bestK2$All.index:

2 3 4 5 6 7 8 9 10

-0.8987 -1.2449 -1.2833 -2.0648 -2.4887 -2.4459 -2.4628 -2.5203 -2.9787

11 12 13 14 15

-2.9450 -2.6982 -2.9238 -2.7331 -2.9600

bestK2$Best.nc:

Number\_clusters Value\_Index

2.0000 -0.8987

Index=SILHOUETTE

bestK3$All.index:

2 3 4 5 6 7 8 9 10 11

0.2440 0.3092 0.3371 0.2788 0.2981 0.3346 0.2360 0.2196 0.3025 0.2832

12 13 14 15

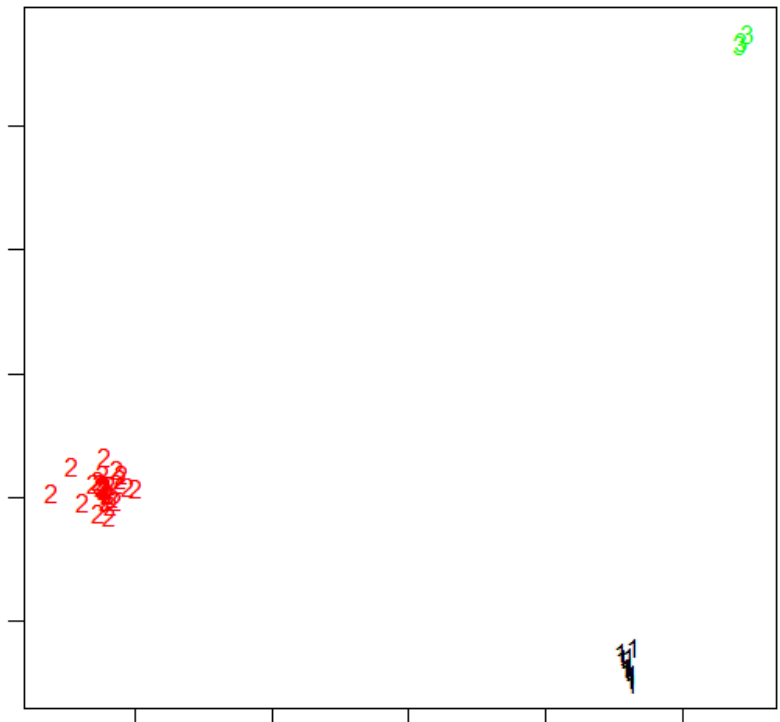
0.2997 0.2925 0.3109 0.2974

bestK3$Best.nc

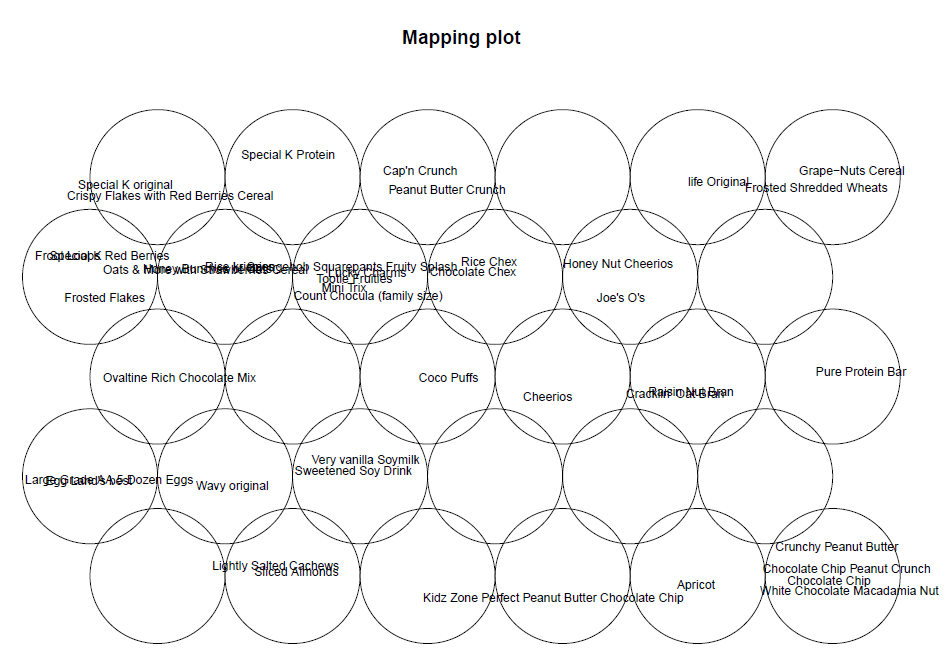
Number\_clusters Value\_Index

3.0000 0.3371

* Optimal Number of clusters that can be used is “***Three”***

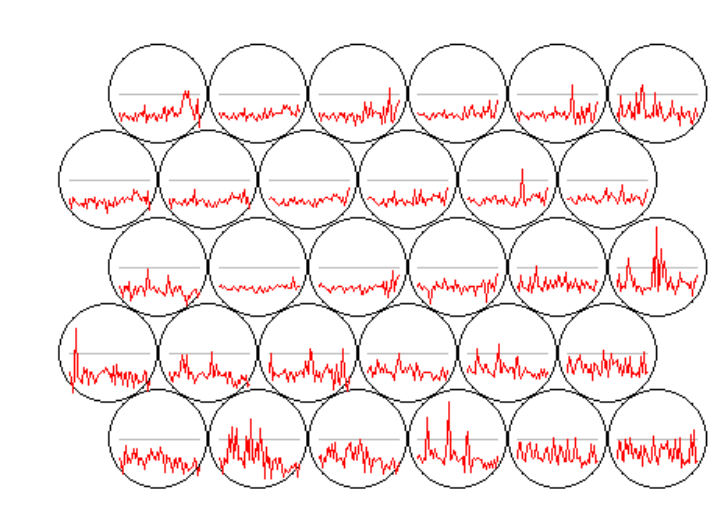


1. **Kohonen SOM:**

Created a Kohonen SOM with 6 Rows and 5 Grids:

The following product types are framed into same grids or closely adjacent:

* Vanilla Soymilk +Sweetened Soy Drinks
* Special K original+Berries Cereal+Special K protein+Flakes+Oats
* Chocolate chip peanut crunch+chocolate chip+White chocolate Nut+
* Kids Zone
* Coco Puffs+savories+Cheerios



**Extra Credit:**

I have updated the heatmap app in **Github IO**, since shiny IO had connection issue with my R studio. Also, this assignment is published in Github along with its heatmap IO

<https://github.com/Vaishnavi28/DataSegmentation_FoodProducts>

The following method are adopted during heat map generation:

* Heat Map without re-ording columns:
* Cluster Rows
* Transpose rows and Cluster columns
* Heat Map for First Cluster
* Heat Map for Second Cluster

***HEATMAPLY and HEATMAP*** packages are used.

**# create heatmap and don't reorder columns**

heatmaply(prodscaled, Colv=F, scale='none')

heatmap(prodscaled, Colv=F, scale='none')

**# cluster rows**

hc.rows <- hclust(dist(prodscaled))

plot(hc.rows)

**# transpose the matrix and cluster columns**

hc.cols <- hclust(dist(t(prodscaled)))

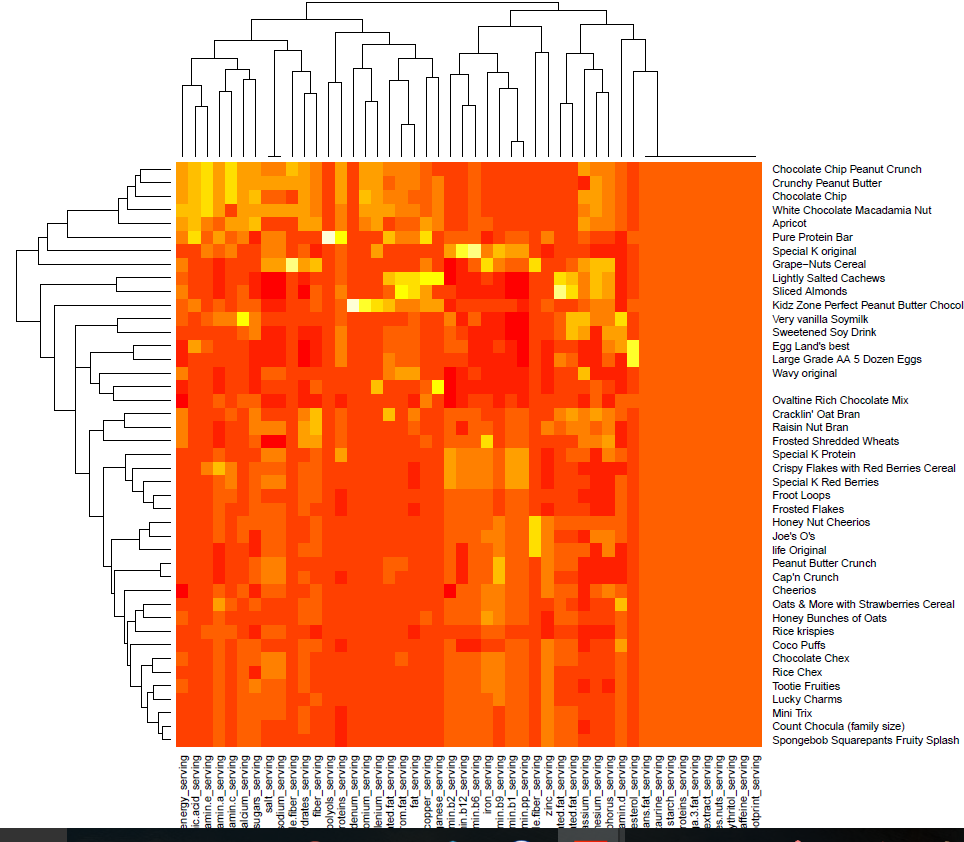
plot(hc.cols)

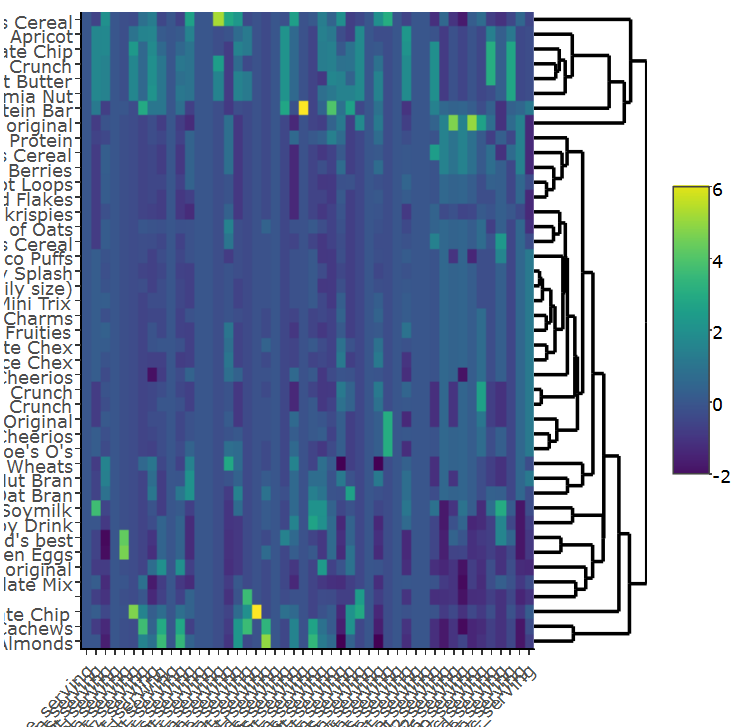
**# draw heatmap for first cluster**

heatmaply(prodscaled[cutree(hc.rows,k=2)==1,], Colv=as.dendrogram(hc.cols), scale='none',margins = c(60, 150))

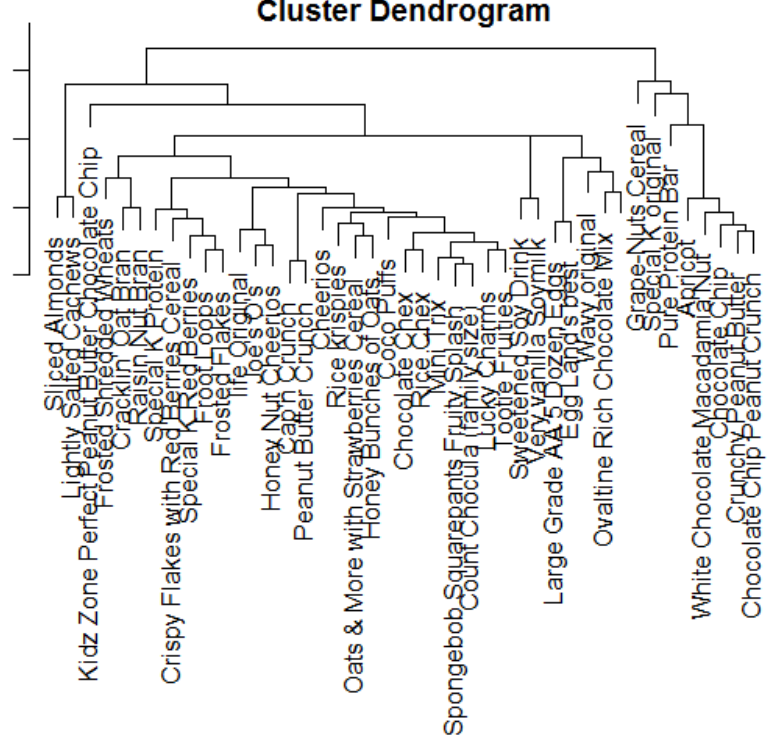
**# draw heatmap for second cluster**

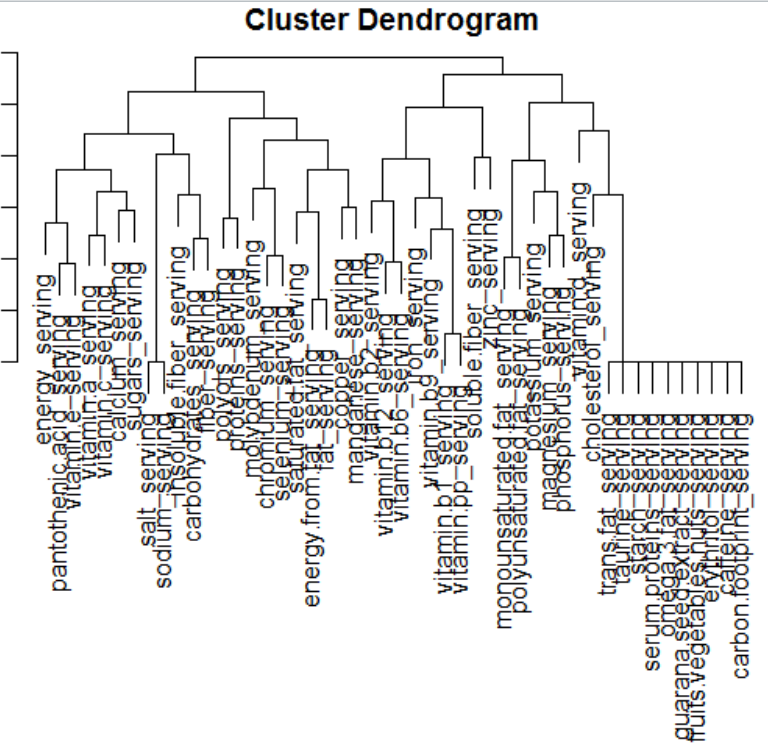
heatmaply(prodscaled[cutree(hc.rows,k=2)==2,], Colv=as.dendrogram(hc.cols), scale='none')

****

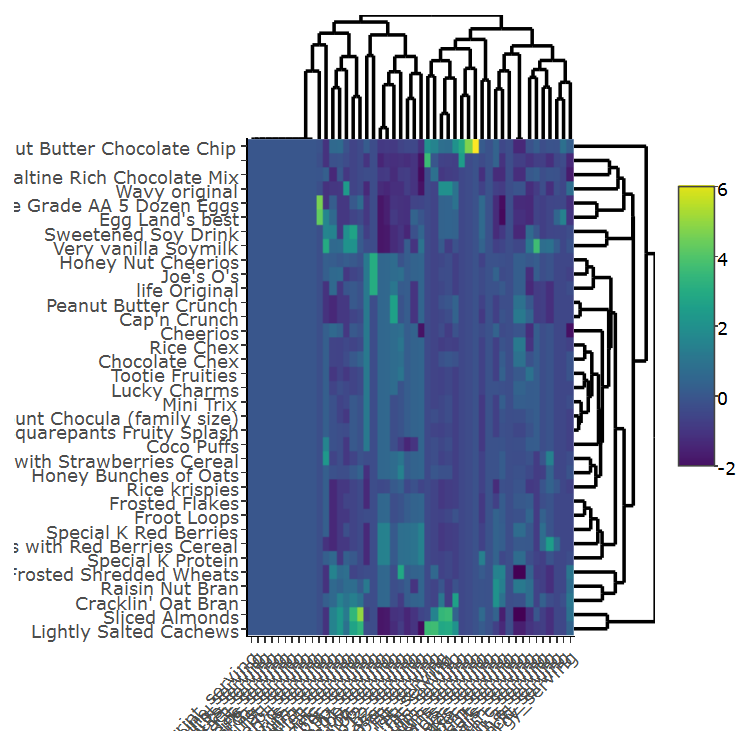
****

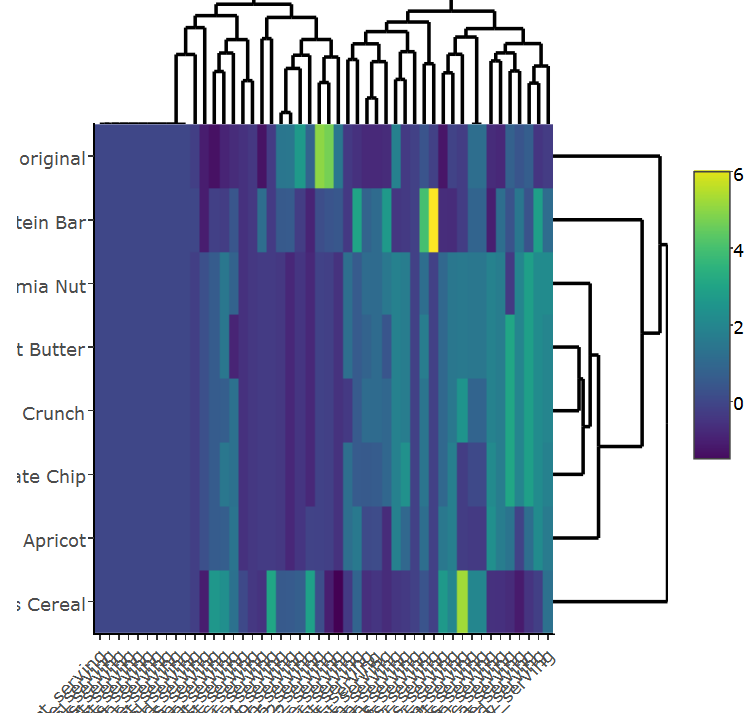
**Cluster Rows:**

****

**Cluster Columns:**

**HeatMap for first Cluster:**

****

**Heat Map for Second Cluster:**