Week 6 Assignment: K-Means & HCA

Scott Schirkofsky  
MSDS 680 Machine Learning

Regis University

K-Means & HCA

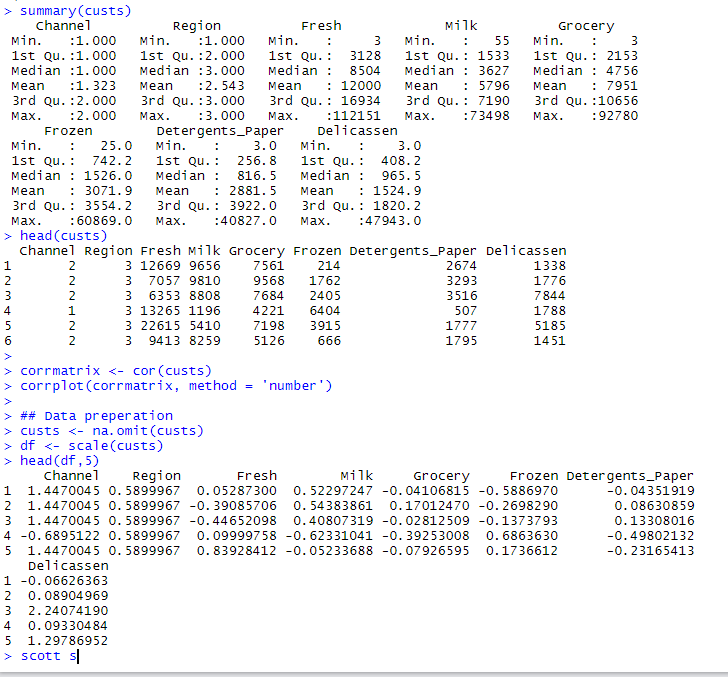
**Introduction**

This assignment in k-means and HCA will utilize the wholesale customer dataset from UCI.

From the given data set: <https://archive.ics.uci.edu/ml/datasets/Wholesale+customers>

**Collecting Data**





**Transforming Data**

It should be noted that the first time I went through this exercise I did not normalize the data, nor did I remove the Channel and Region columns. I kept it this way the first time through and then make the adjustment after question 5 including the performing the clustering methods again. In the output below the channel and region columns still exist.

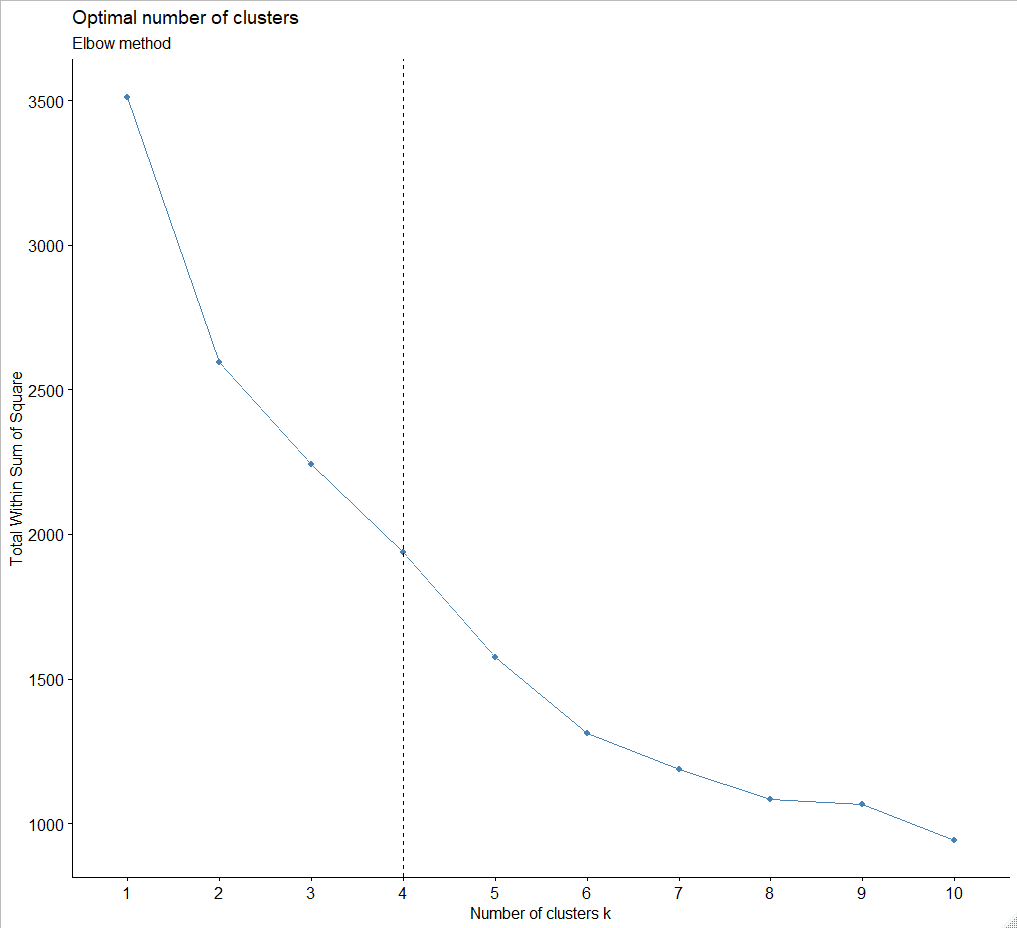


**Part 1: Use K-means to cluster the given data set.**

1. **Find the optimal number of clusters with methods such as Elbow method, silhouette method, and so on.**

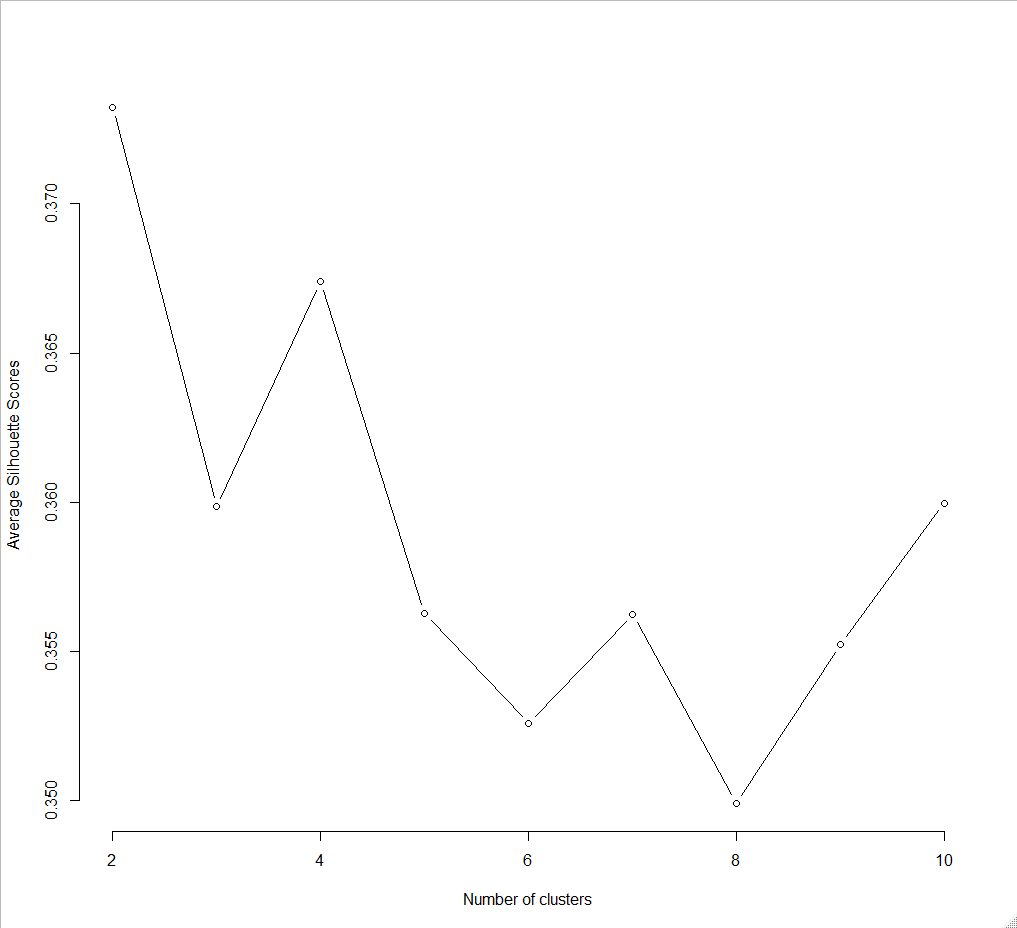
**Elbow:**

Here the elbow method shows the optimal number of clusters to be 4. I will experiment with 4 clusters further on question 4.



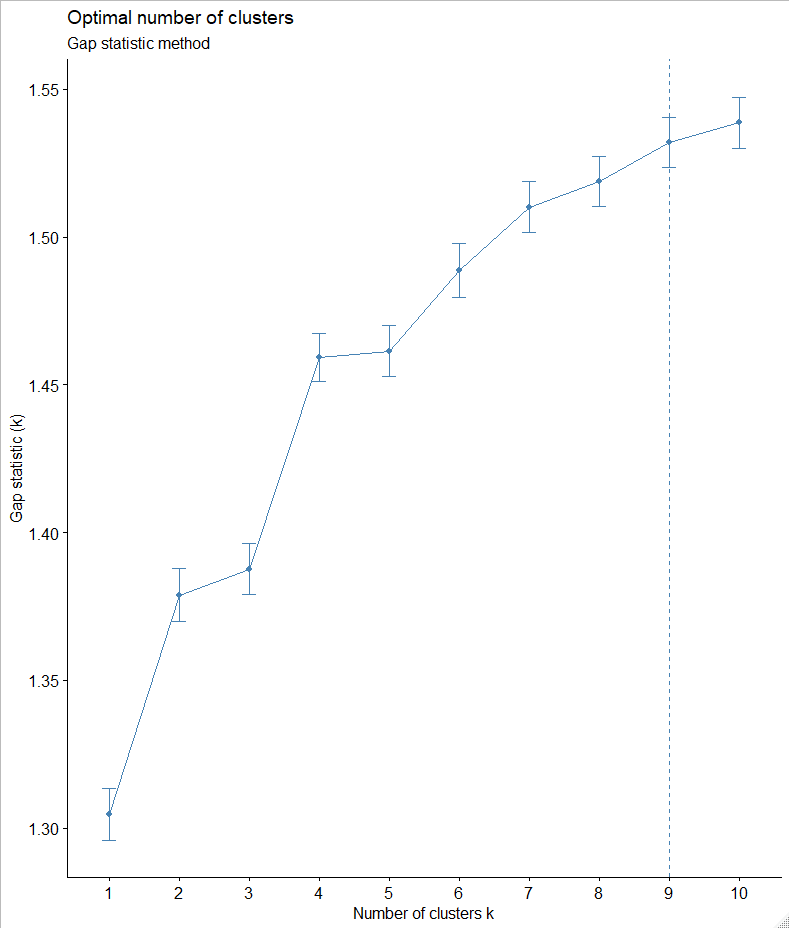
**Silhouette:**

Here the silhouette method shows the optimal number of clusters to be 2. I will experiment with 2 clusters further on question 4.



**Gap Statistic:**

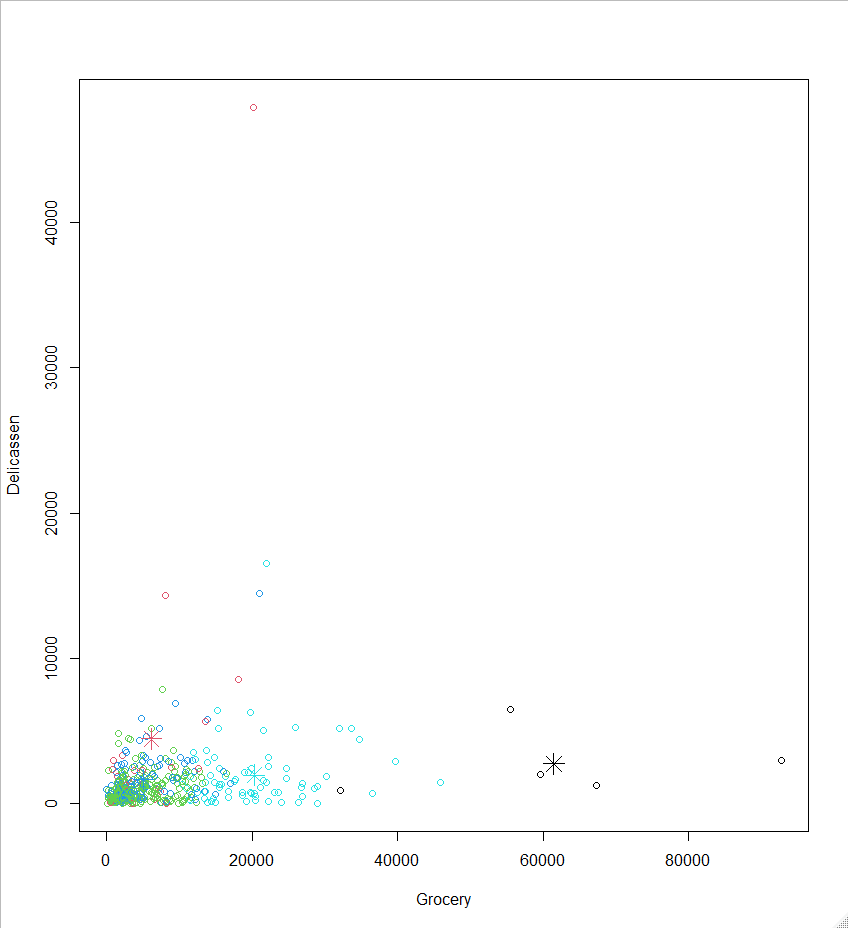
Here the gap statistic method shows the optimal number of clusters to be 9. I will experiment with 9 clusters further on question 4.



1. **Pick two variables such as fresh and frozen, plot the clusters and mark the centers of those clusters. You may try to vary these two variables to get a more readable plot.**

I used Grocery and Delicassen to plot. It’s difficult to see the centroids but they are there for each cluster.





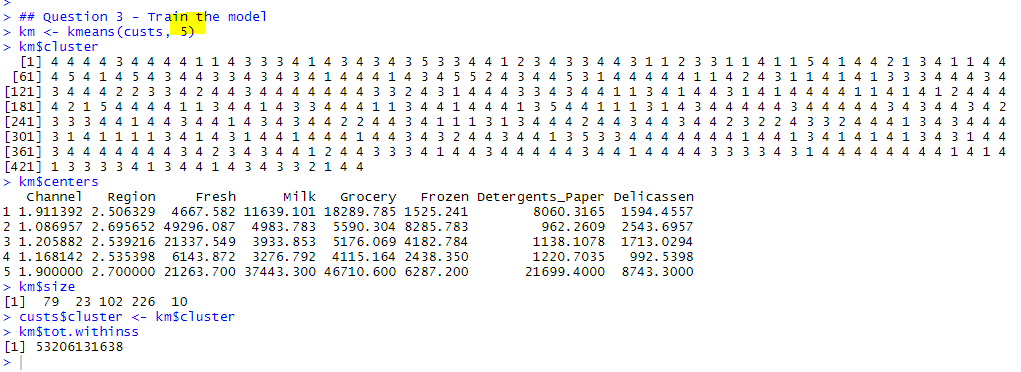
1. **Is there any cluster that is difficult to detect? Discuss.**

The clusters containing the objects with the lowest values, and therefore the lowest value centroids, seem to be the hardest to detect. In my plot the green cluster is very difficult to detect while the black cluster is quite easy.

The other concern I have is that the three methods all showed vastly different

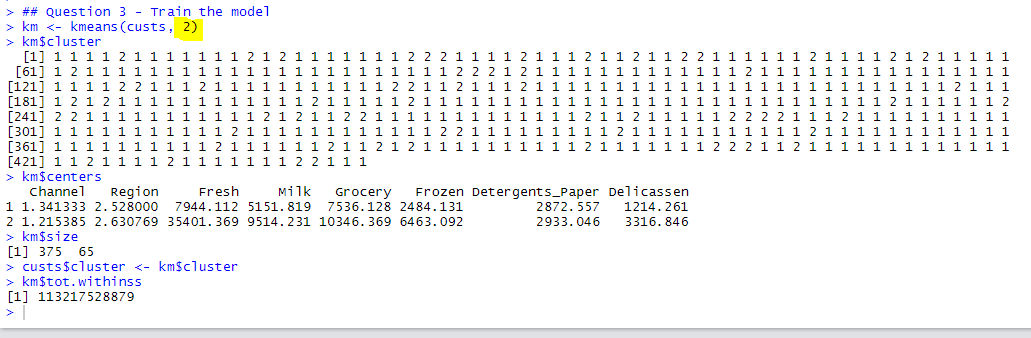
1. **Interpret the meaning of each cluster and conclude your findings.**

The 5 clusters are very different in size and density with one containing 244 objects and another with only 5. There is a large variation in cluster centers from 1.13 to 51254.60. The sum of squares is 53,206,131,638.

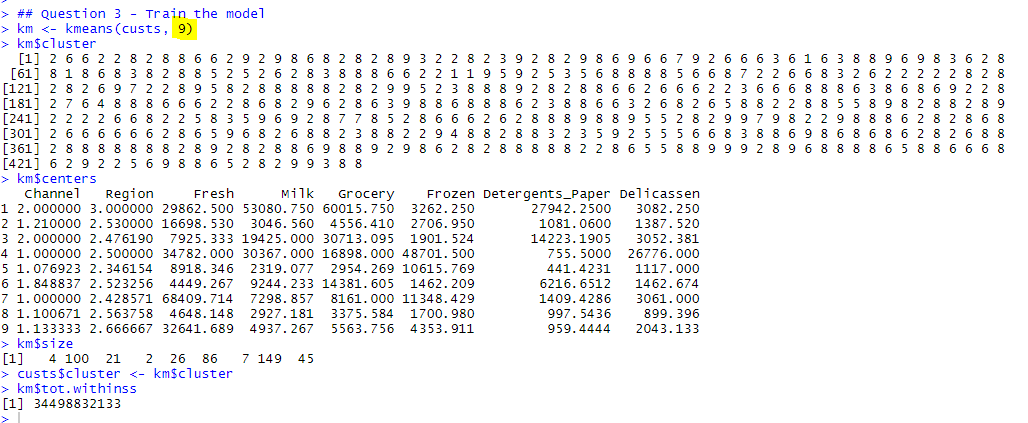


I performed k-means again, this time with 2 clusters suggested by the silhouette method. This provides a value of

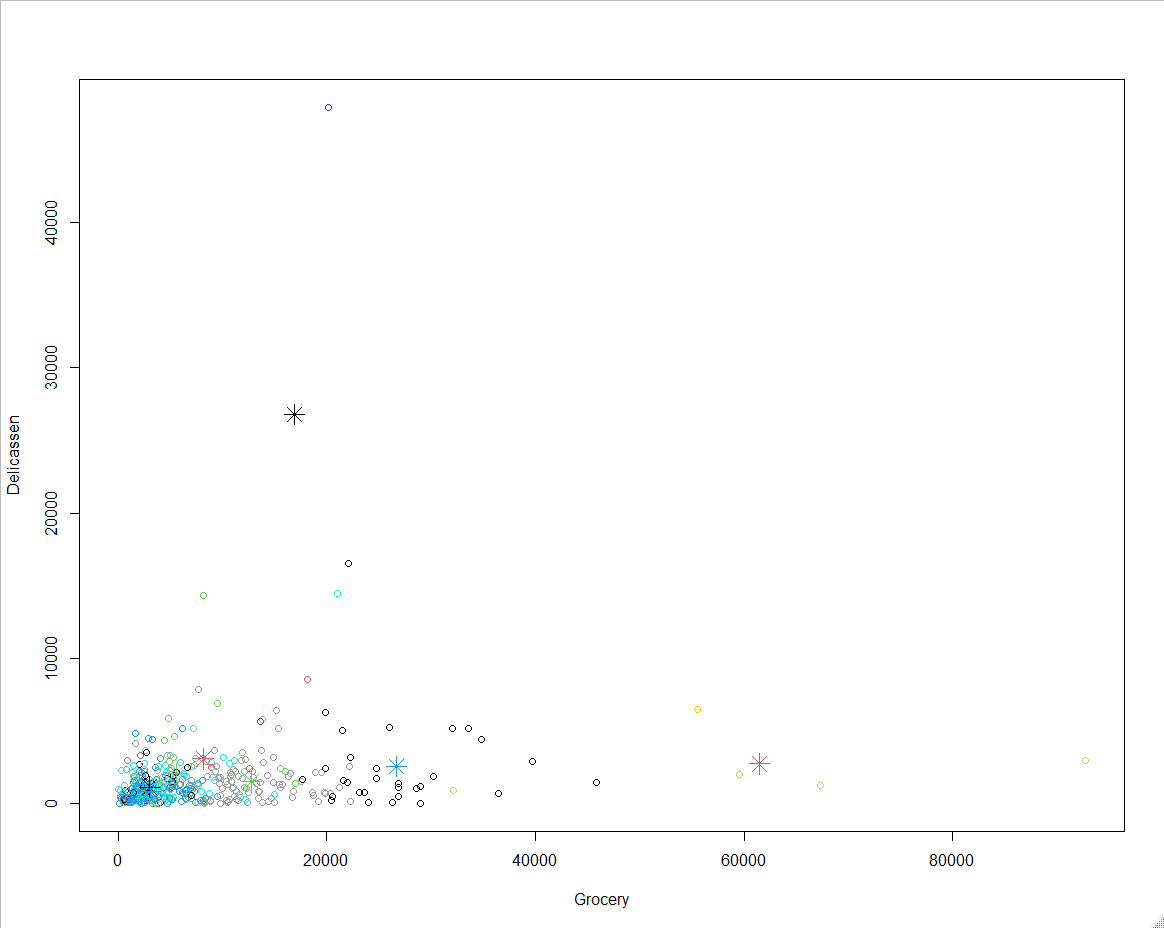
113,217,528,879 for sum of squares, nearly twice the value returned when using 5 clusters in the example above.



Next, I performed k-means with 9 clusters suggested by Gap Statistic method. This provides a value of 34,498,832,133 for sum of squares, considerably less than 53,206,131,638 returned with 5 clusters.

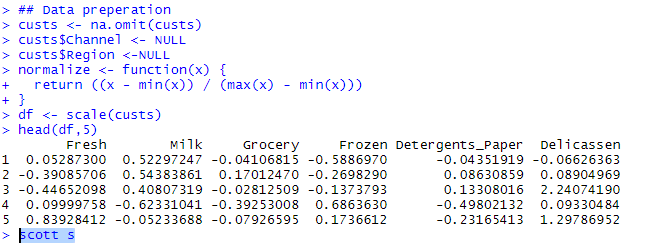


Because 9 clusters returns the lowest sum of squares figure I decided that accept it as the optimum value and plotted the results. The centers are easier to see now but still not great. The outlying values are really throwing the clustering off.

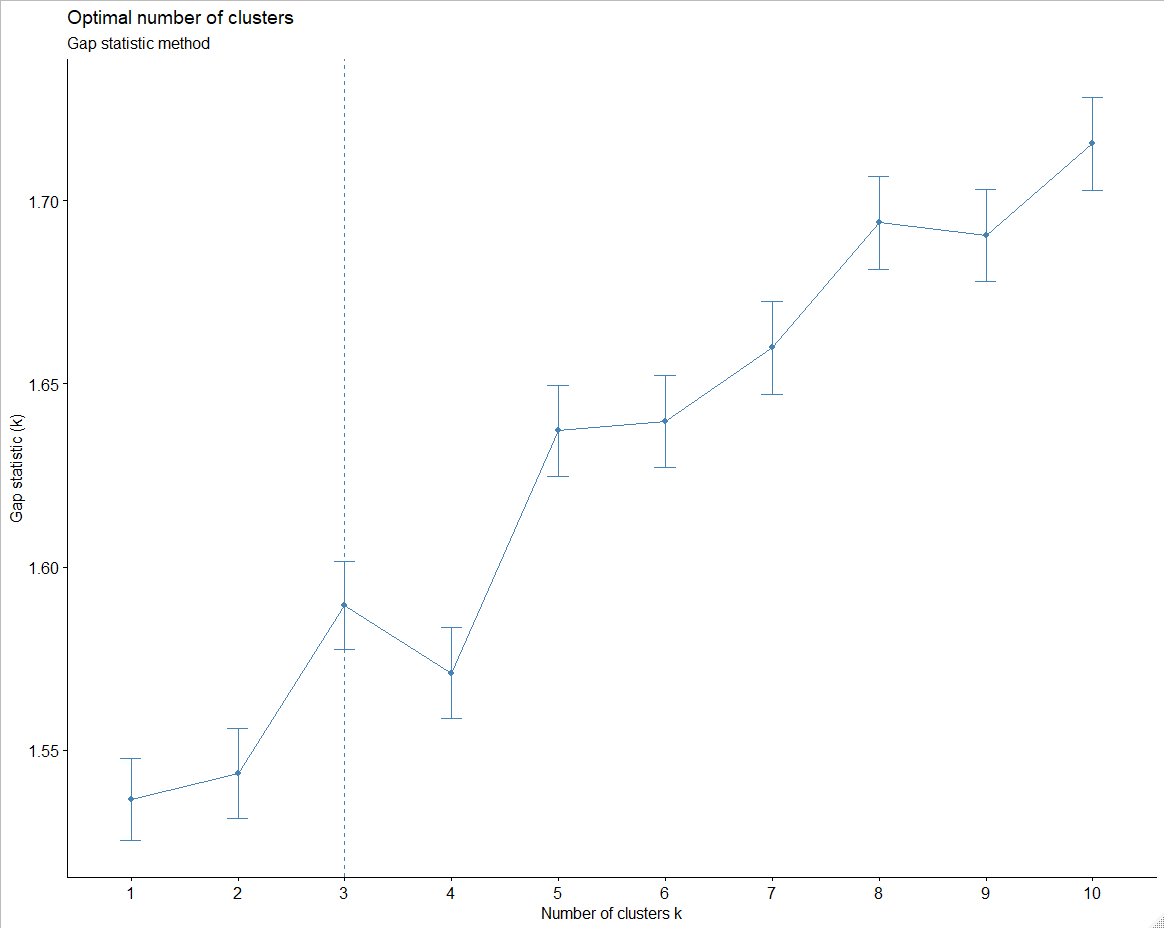
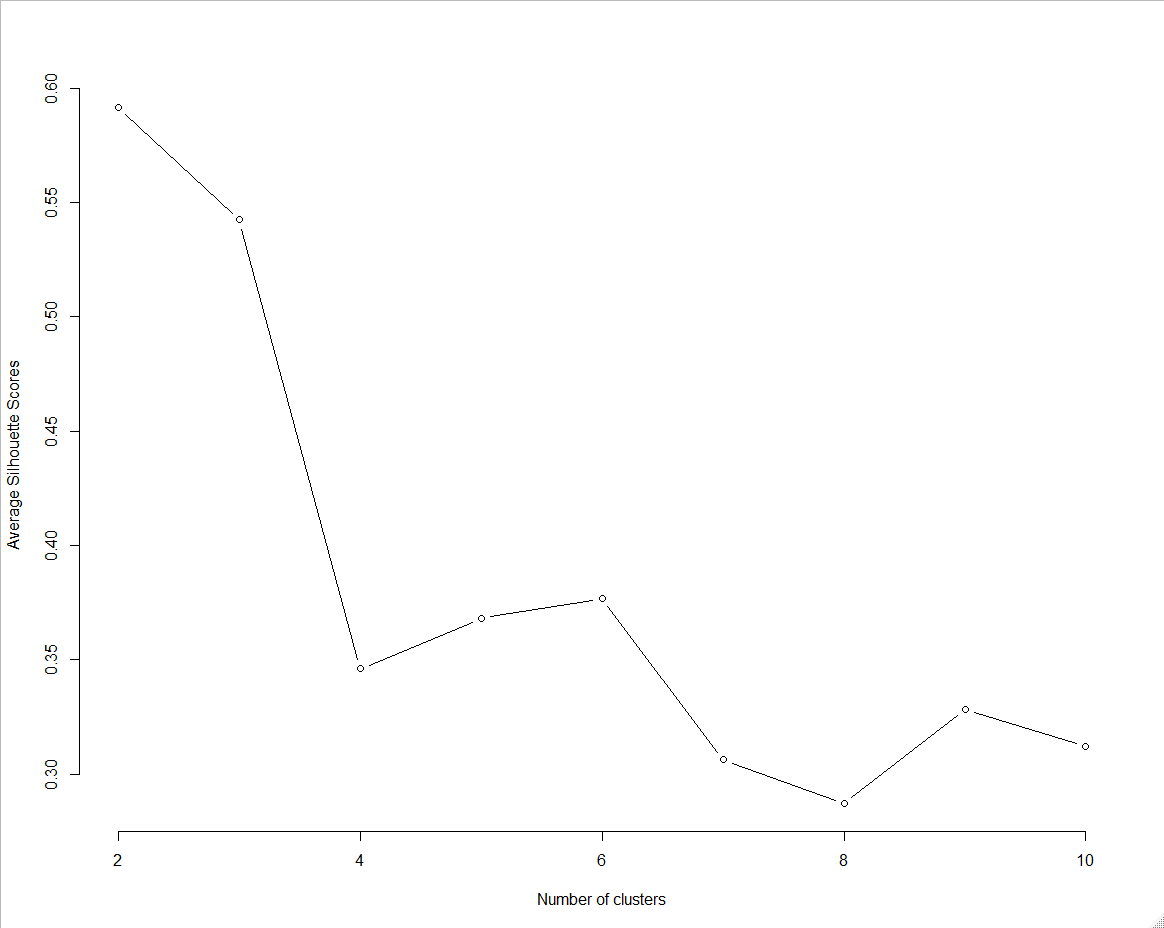
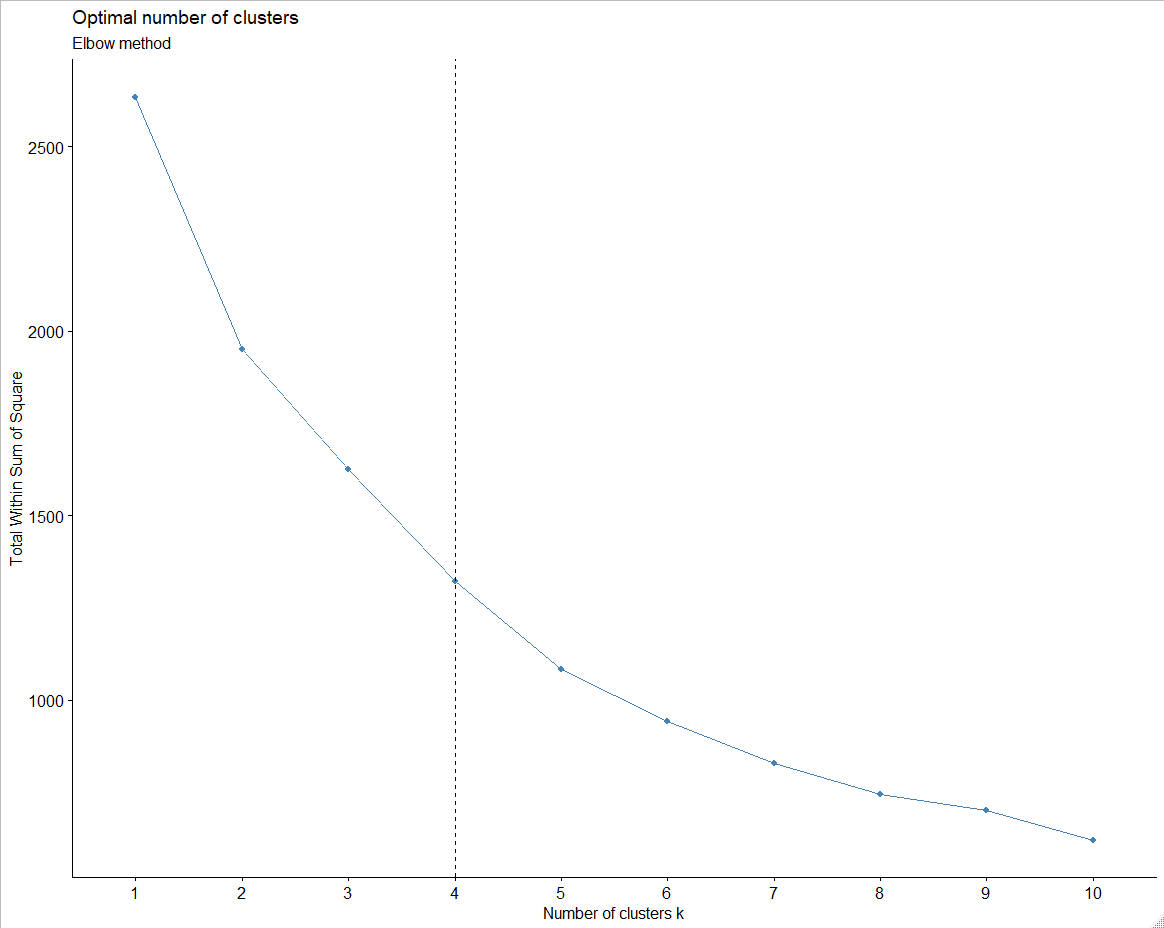


**5) Do you have any concern on the experiments?**

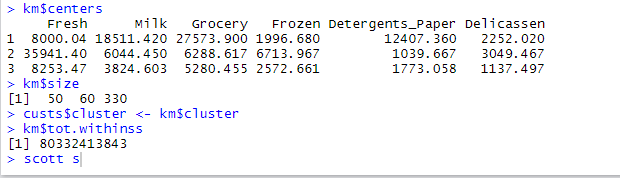
One of the concerns I have is the lack of separation between the clusters. I believe this is due to the large discrepancy of size and density between the clusters. Perhaps standardizing the data to have a mean of zero, normalizing and/or removing the outliers would help with this issue so I decided to give it a try and normalize the data and removing the channel and region columns.



The interesting thing after normalizing and standardizing the data, Gap Statistic (plots below) is now indicating that 3 may be the optimum number of clusters.



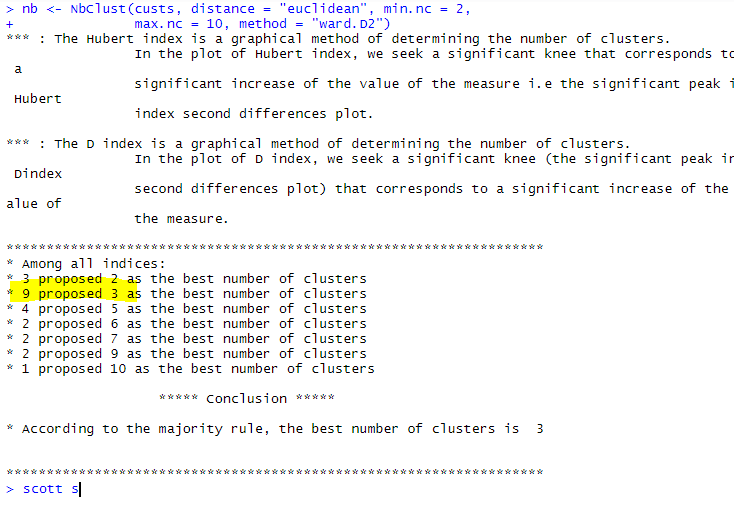
Next, I go ahead and perform Gap Statistic again with 3 clusters and get a sum of squares of 80,332,413,843. I believe that 3 clusters is the right number for k.

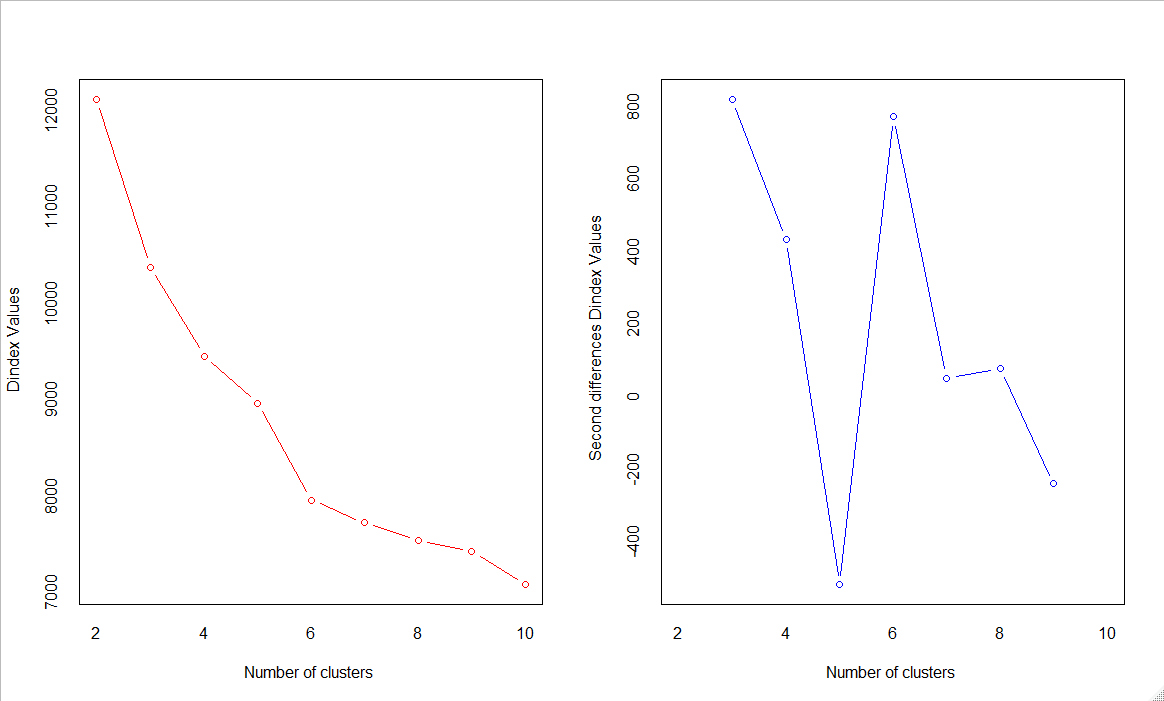


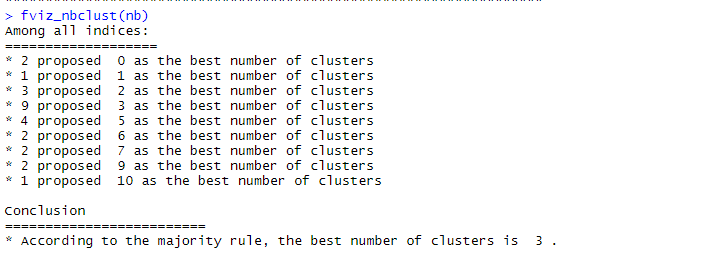
**Part 2: Use HCA to cluster the given data set (Choose TWO linkages of your choice)**

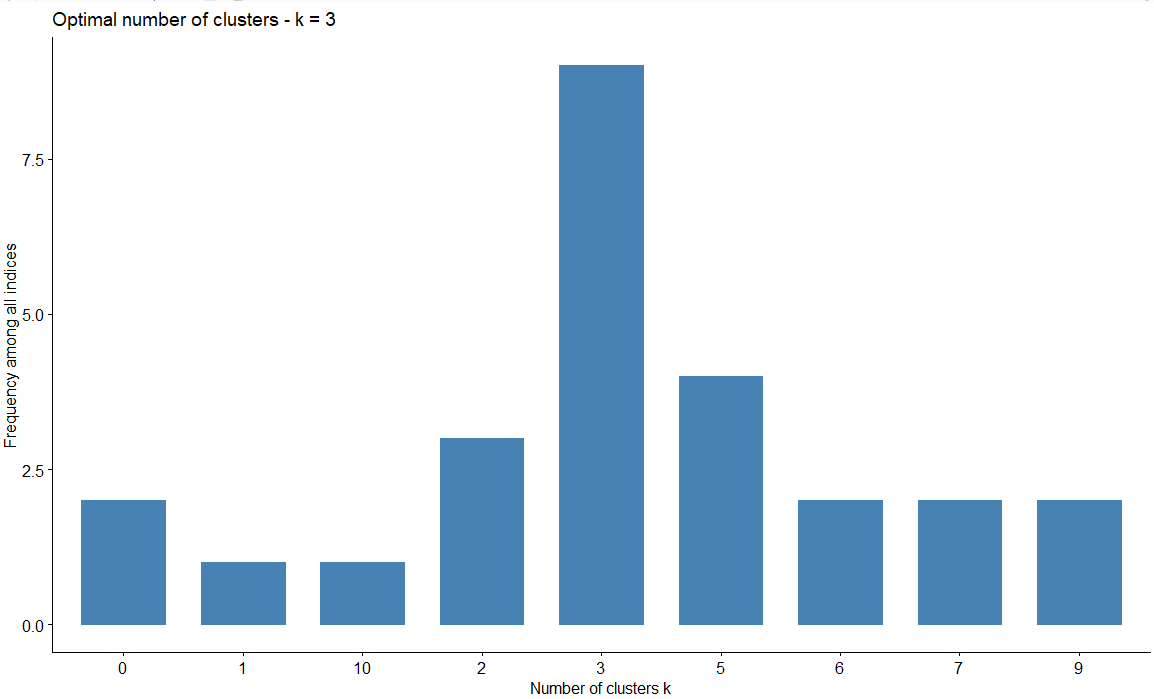
1. **Identify the number of optimal clusters and justify how do you pick the number of clusters.**

Hierarchical clustering also shows that 3 is the optimum number of clusters. I justified this number of clusters by transforming the data, then normalizing and standardize the data and performing multiple clustering methods. Gap Statistic revealed 3 would be the optimum number and HCA is confirming this number.





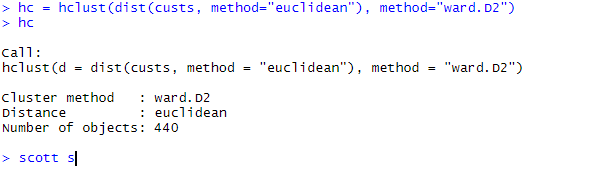


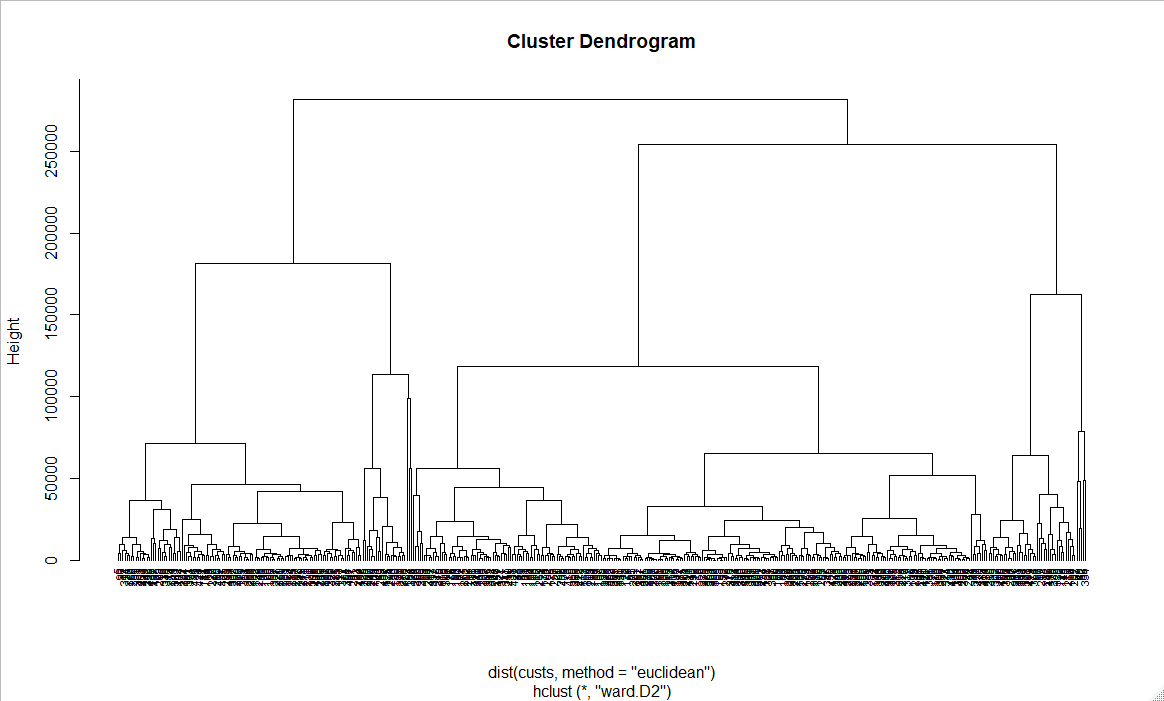


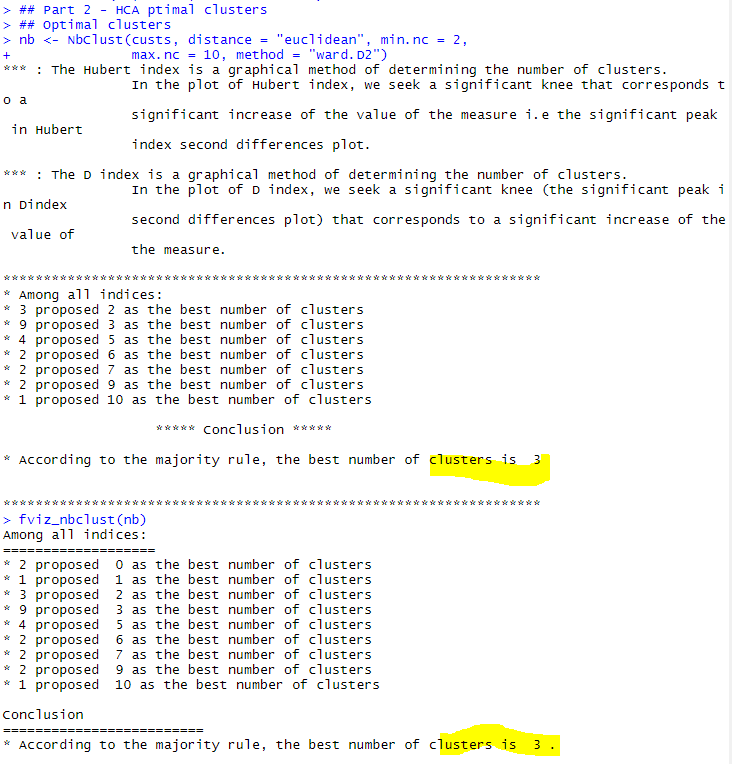
1. **Draw a dendrogram (both linkages)**

**Complete linkage**

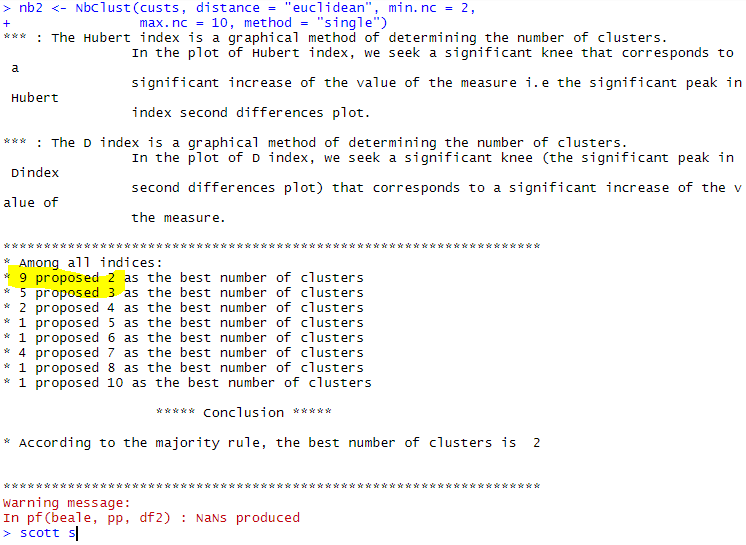
The complete linkage method results in an optimum clusters of 3.

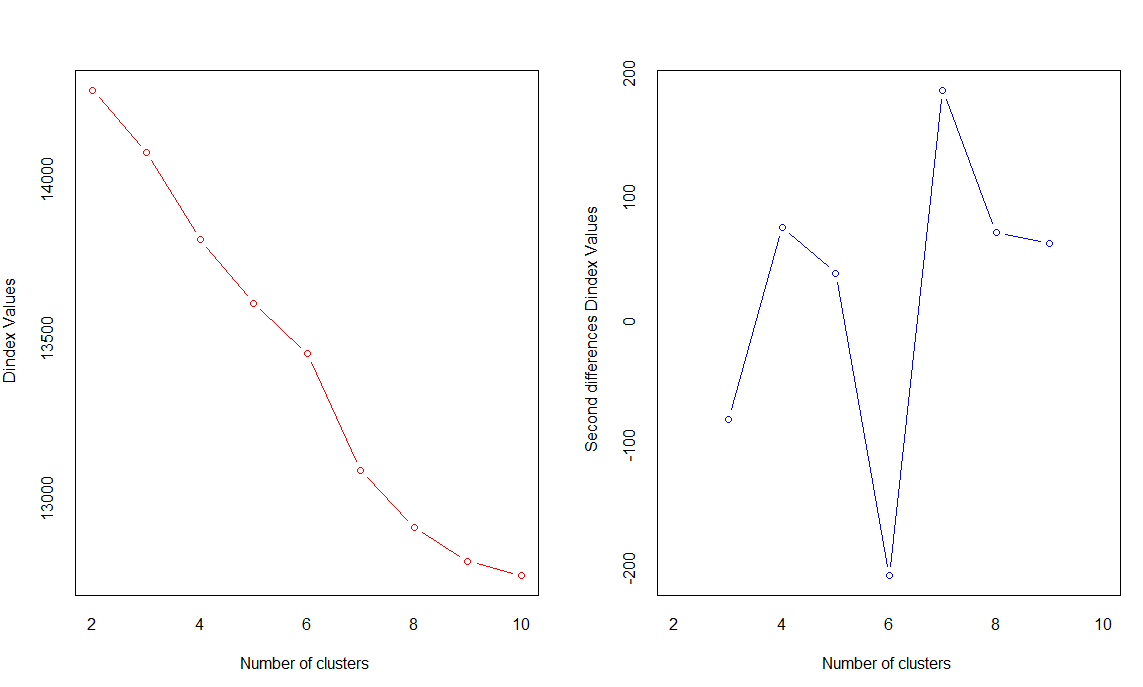


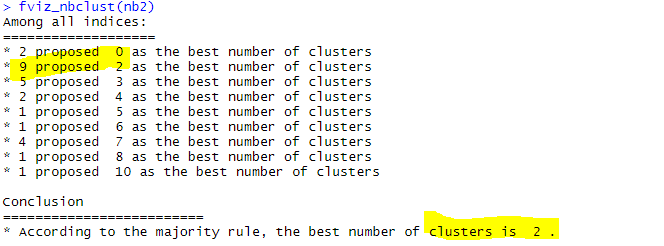


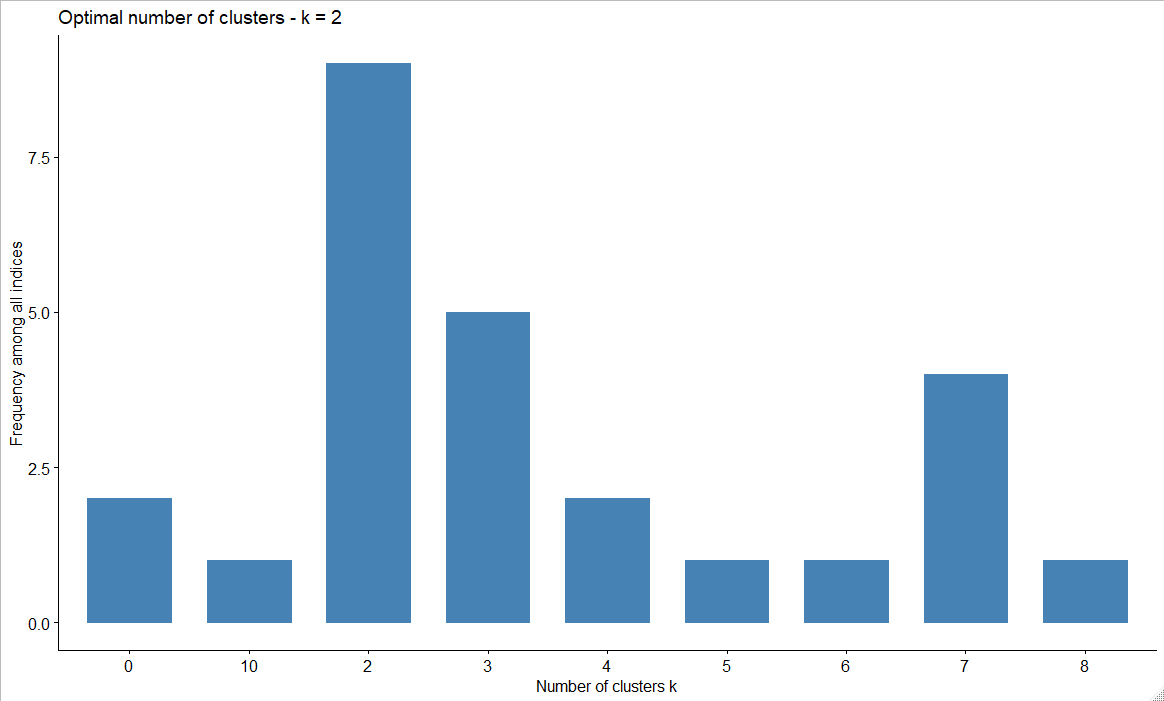


**Single Linkage**Now to create clusters with “single” linkage. Single linkage reveals an optimum cluster size of 2.

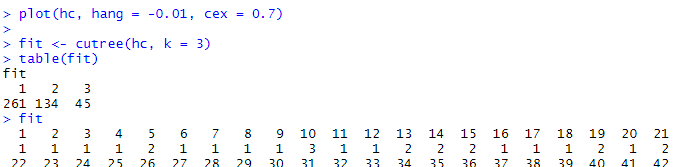


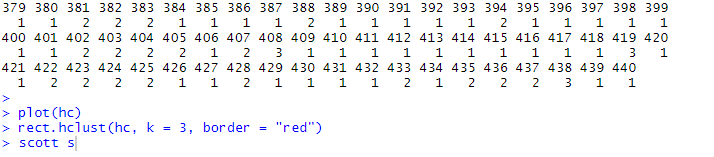
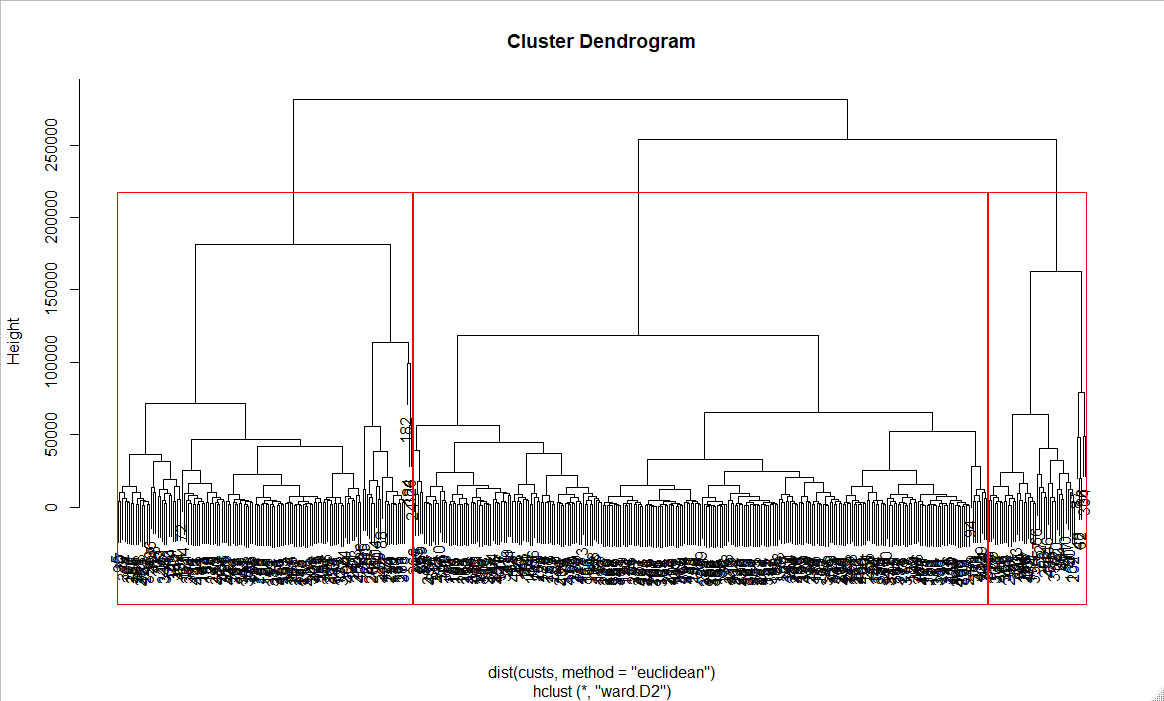






1. **Cut at different heights of the dendrogram. Draw it. Discuss and summarize your findings.**  
   The cut dendrogram shows 261 observations in cluster 1, 134 in cluster 2, and 45 in cluster 3. It’s quite easy to visually see why the tree is cut into the 3 clusters. Those clusters contain the highest levels on the tree, and their widths help fit the lower levels as they fall in line with one of the three clusters. Amazing to see in a dendrogram, just makes it so clear.



1. **State any concerns you might have from the analysis.**

No real concerns that come to mind. The fact that the complete and single linkage methods resulted in different numbers of clusters is not concerning as they were only one number apart. This makes sense to me given that complete linkage is the farthest neighbor approach while single is the nearest neighbor approach. So, the fact that the farthest neighbor approach resulted in 1 additional cluster does not seem highly unusual.