Week 6 Assignment – K-Means and HCA

Zach Adair

Regis University

Abstract

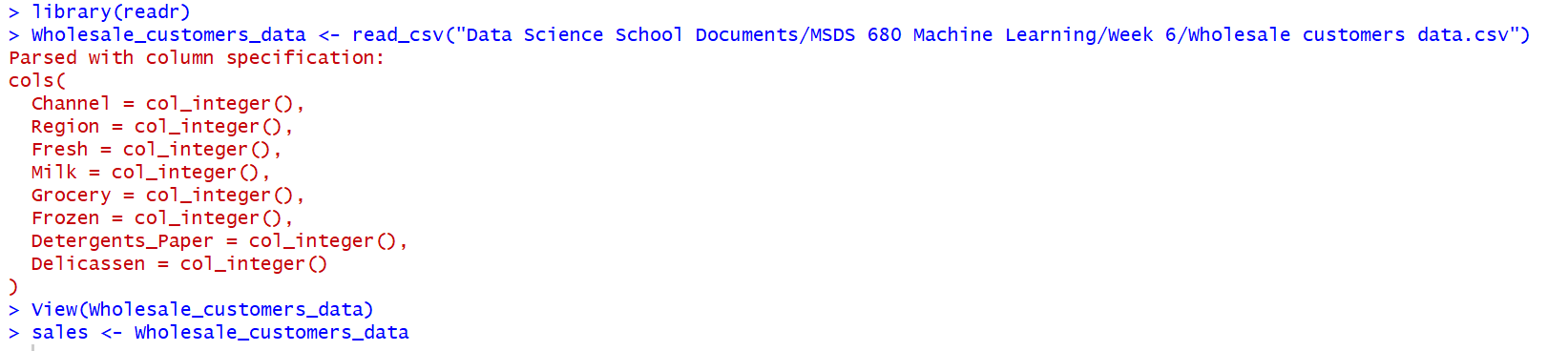
For week 6, the assignment is focused around the use of K-means and HCA for cluster analysis. The data for this assignment is data from a wholesale store and their customer data. There are also several questions to answer within this analysis and I will answer those from the different sections of this report. K-means and HCA are a form of unsupervised learning which revolve around clustering data based on similarities, the difference between the two is that k-means are set up for a specific, k number, of clusters while HCA determines the number of clusters while going through the analysis and doesn’t preconceive its number.

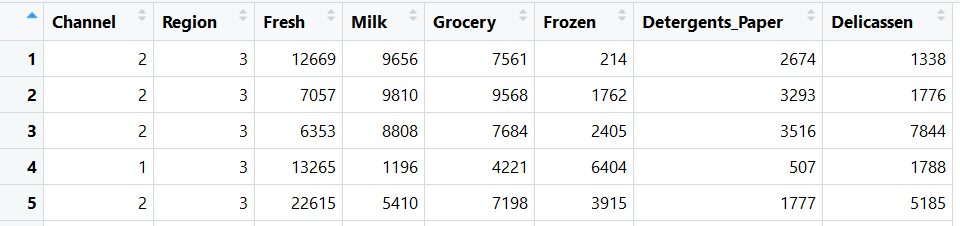
Week 6 Assignment – K-means and HCA

The body of your paper uses a half-inch first line indent and is double-spaced. APA style provides for up to five heading levels, shown in the paragraphs that follow. Note that the word Introduction should not be used as an initial heading, as it’s assumed that your paper begins with an introduction.

# Load the data

The first step in any analysis is to load the data. So I will load the wholesale data into RStudio and begin my analysis.



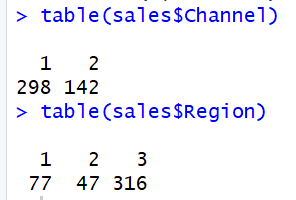


## Explore the Data

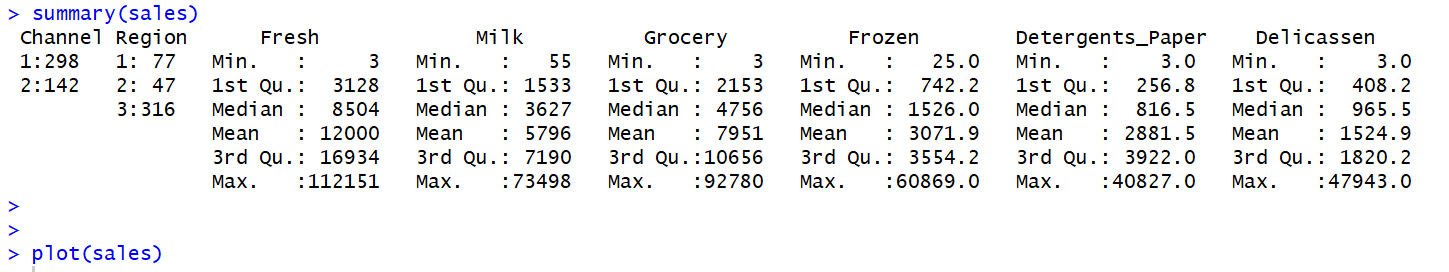
Now that the d ata is loaded into RStudio, I’m going to explore the data and prepare it for a K-Means and HCA analysis.

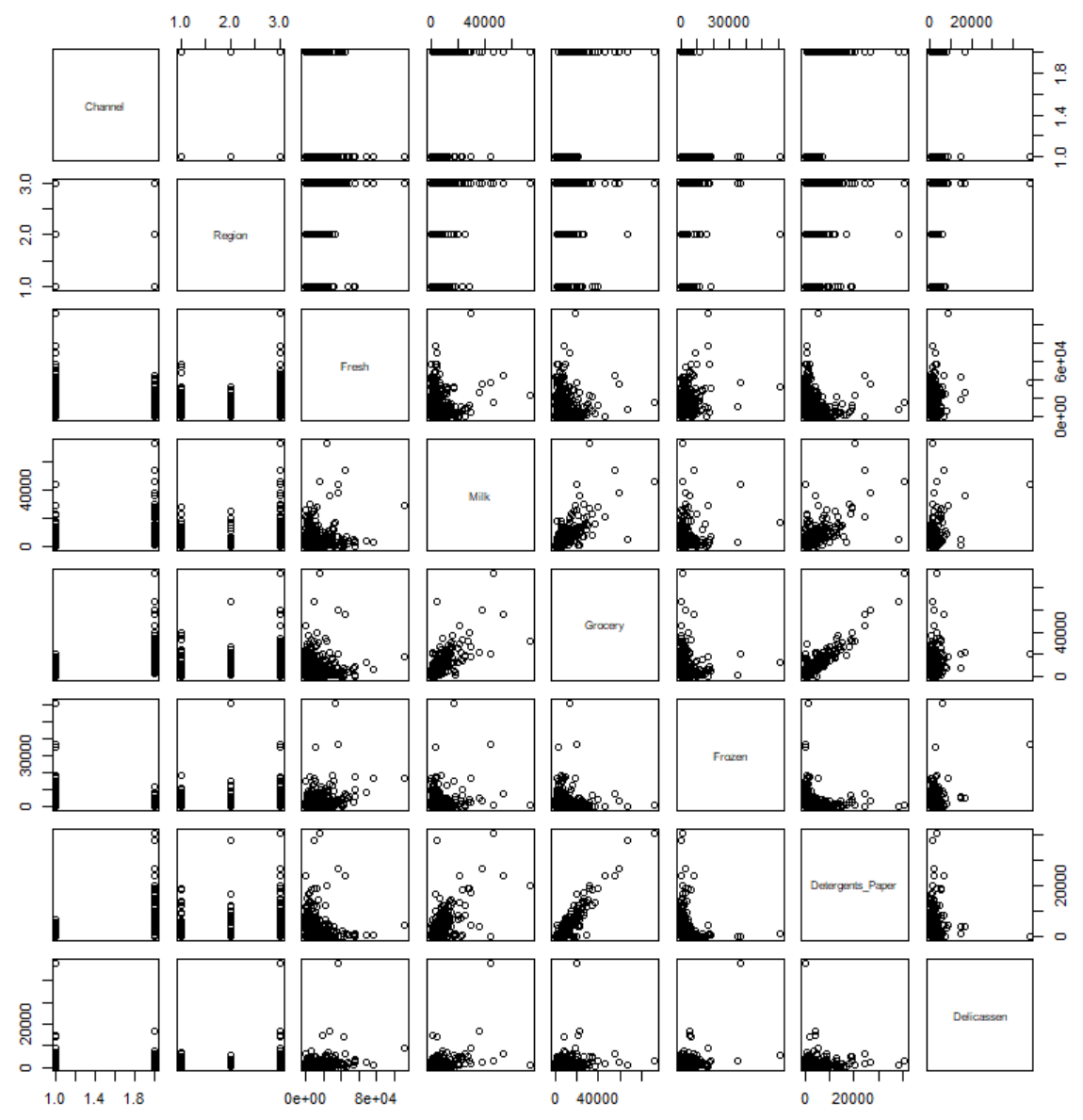


I wanted to set Channel and Region to factors because they both are more factor based fields and I can also now run a table on each to see their distributions.



Now I can see the distributions of both, 1 is skewed favorably for channel and 3 is skewed favorably for the region field.



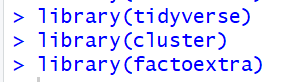


Then I also need to check for any missing values.

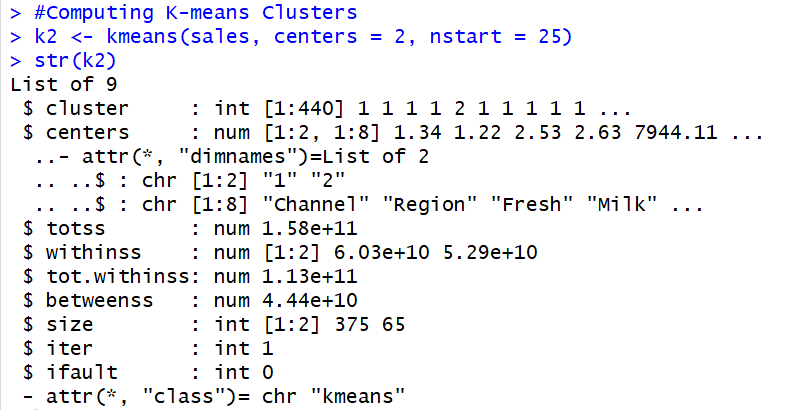


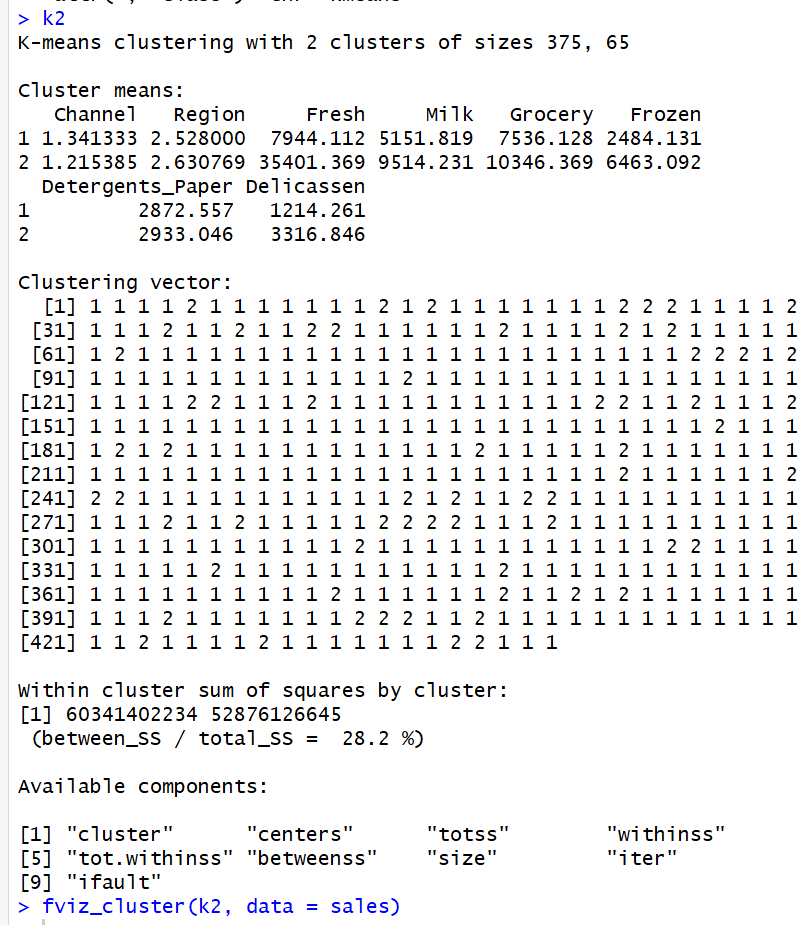
### **K Means Analysis.**

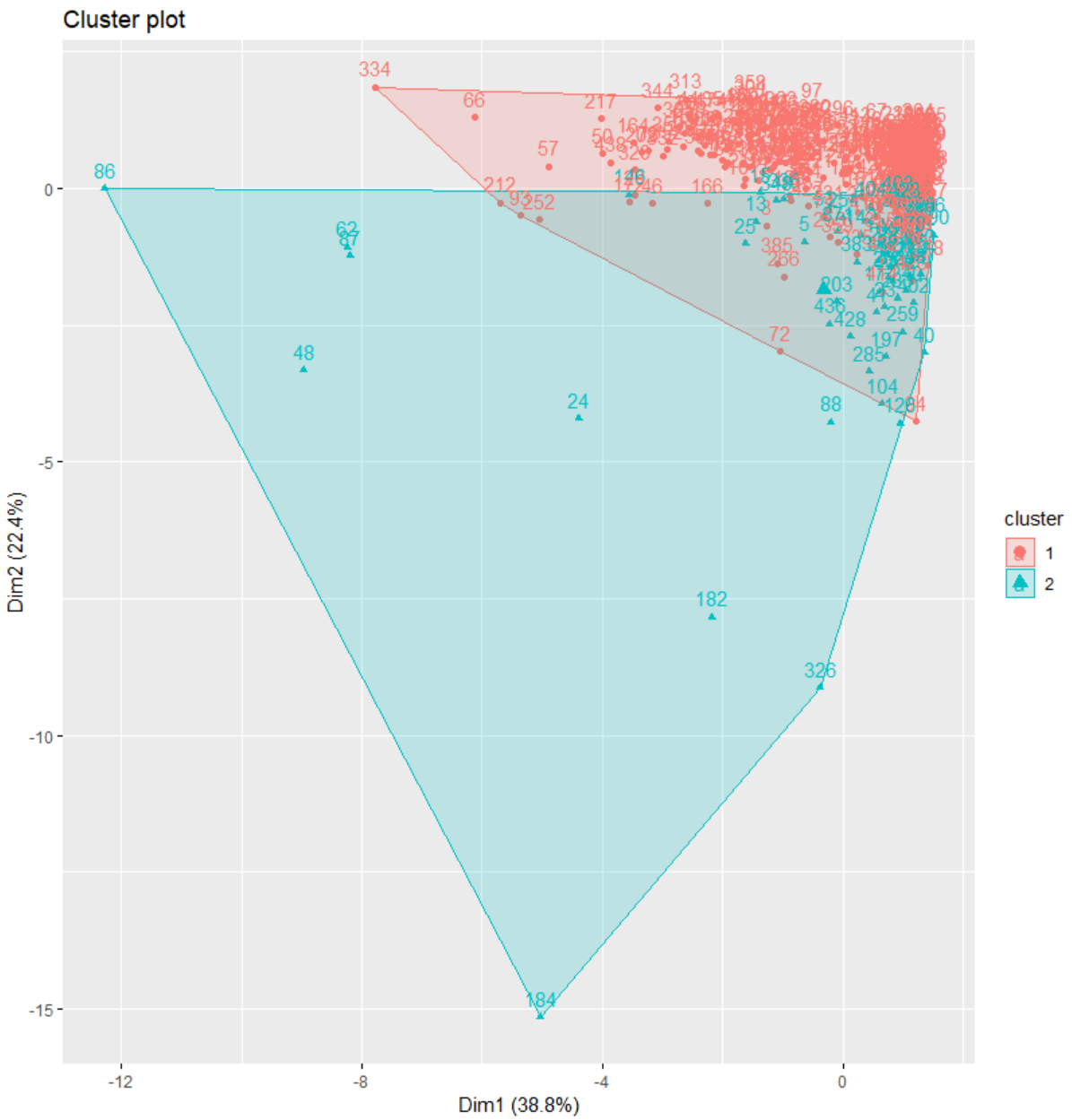
I will start by bringing in several RStudio libraries, they will help me with running my K-means algorithm on my wholesale data. I will start by adding some RStudio libraries to use for this process.



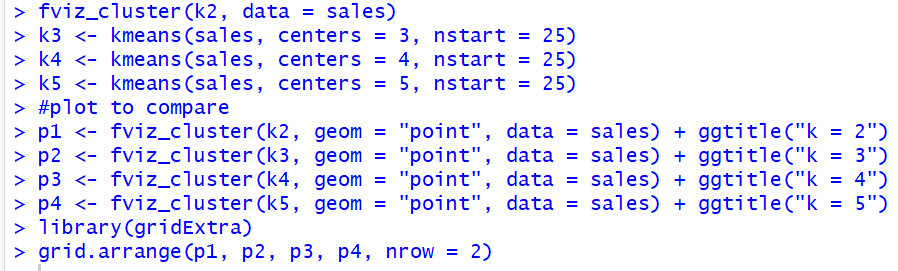
Now that these libraries are in I can begin my k-means analysis. I will start by creating two clusters and charting them to get a look at the data distributed and how the algorithm clusters the data together.

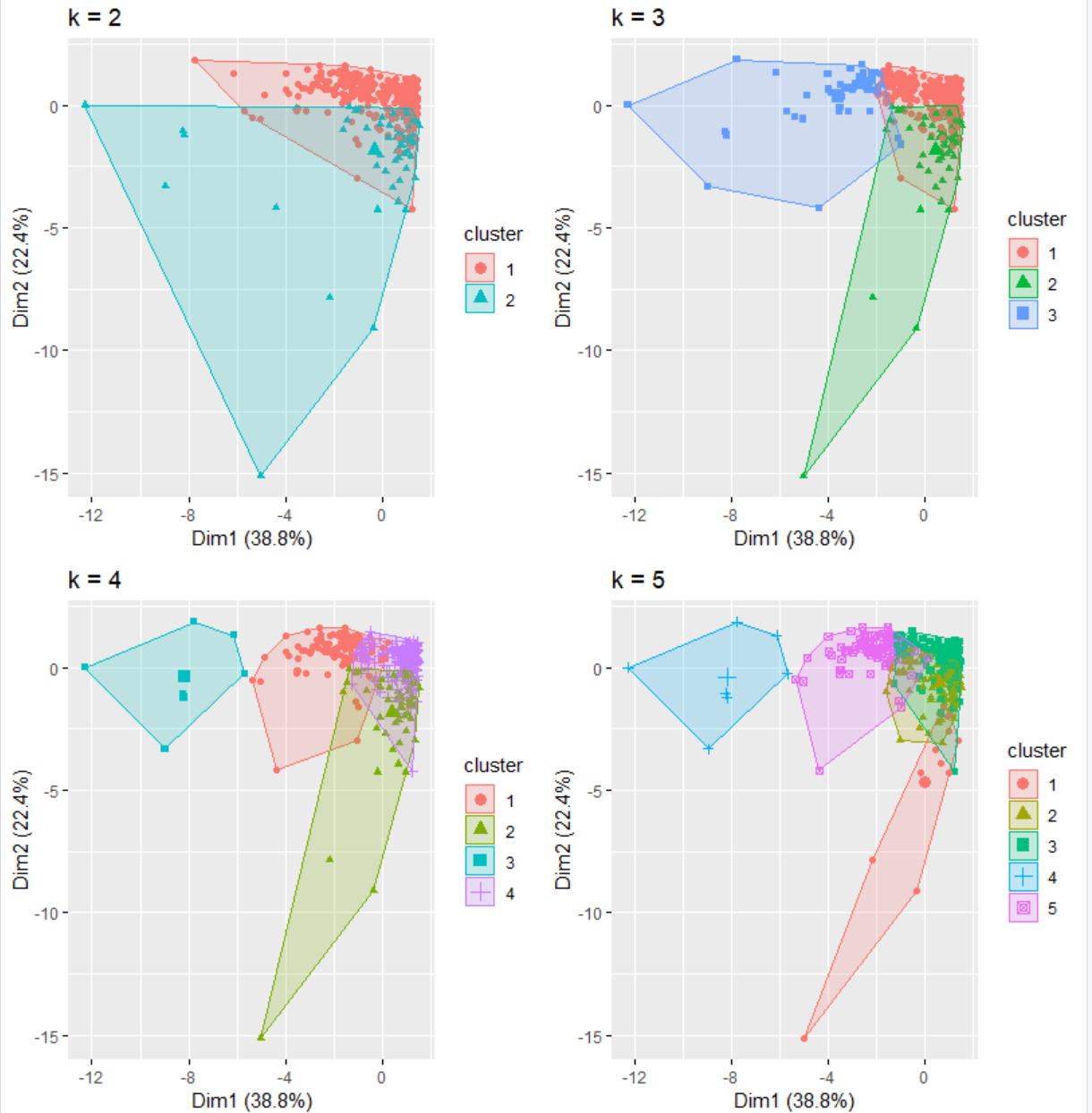




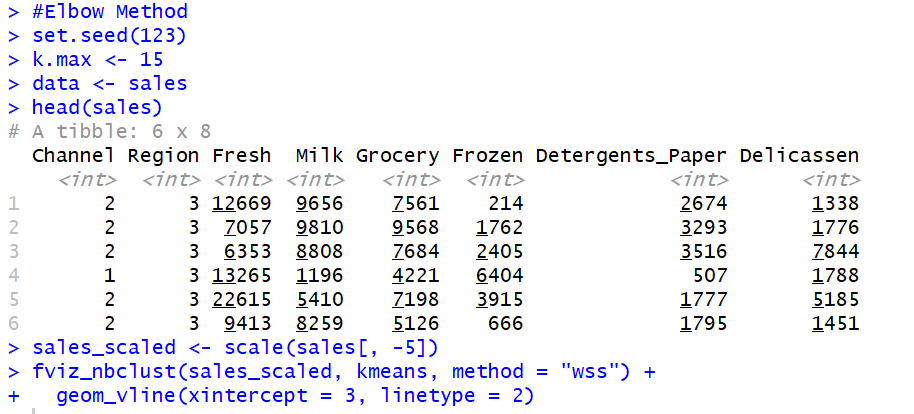


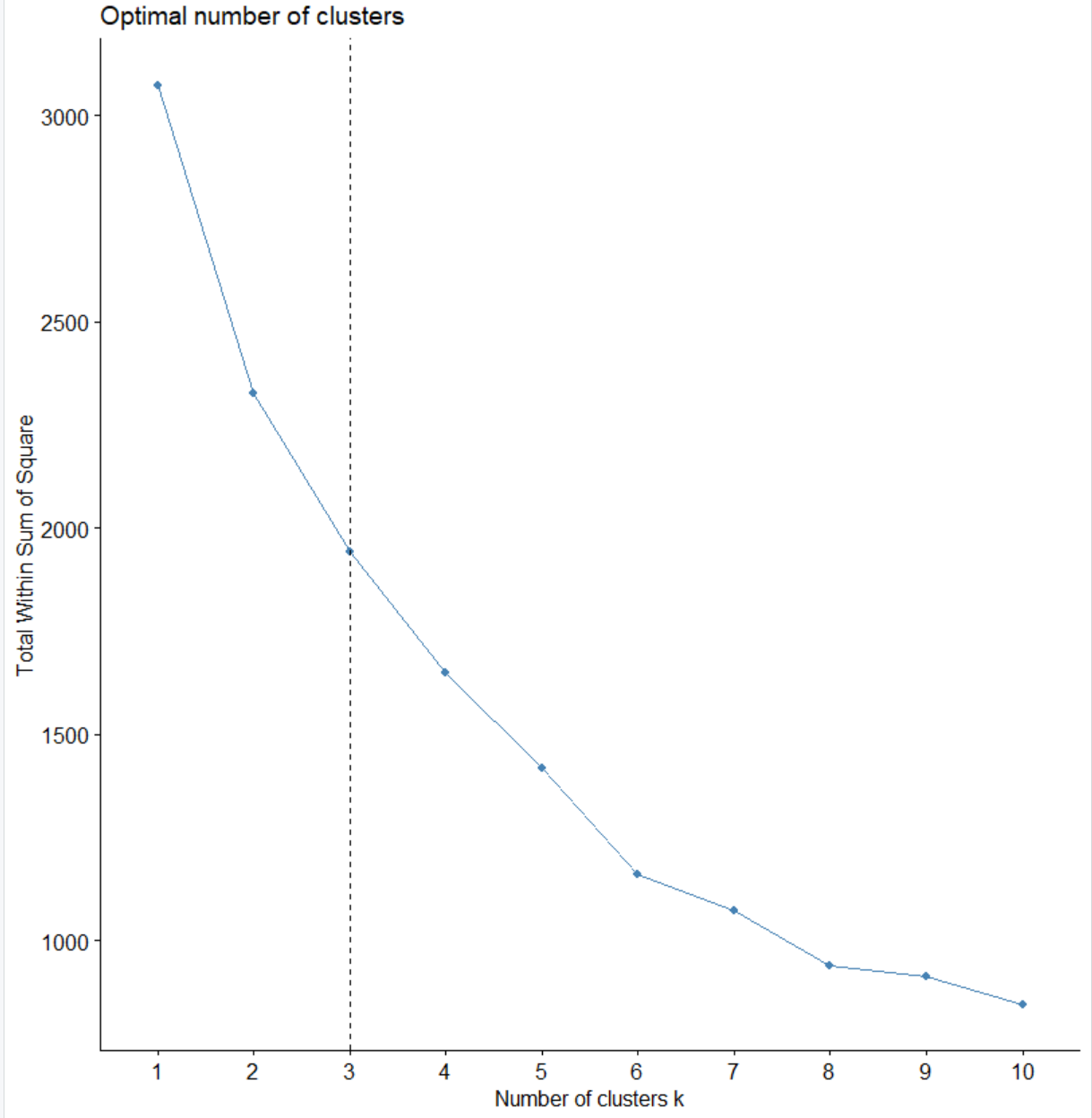
From the k-means result from two clusters we see how each variable stacked up within the separate clusters, and how the data charts within the range of the clusters. It appears a majority of data points plot around the same space but the first cluster is much smaller than the second which has a much wider spread. Now I’m going to compare when I run the k-means process and get 3, 4 and 5 clusters. I want to look at all these plots simultaneously so I will also be adding the library gridExtra before I plot my k-means.





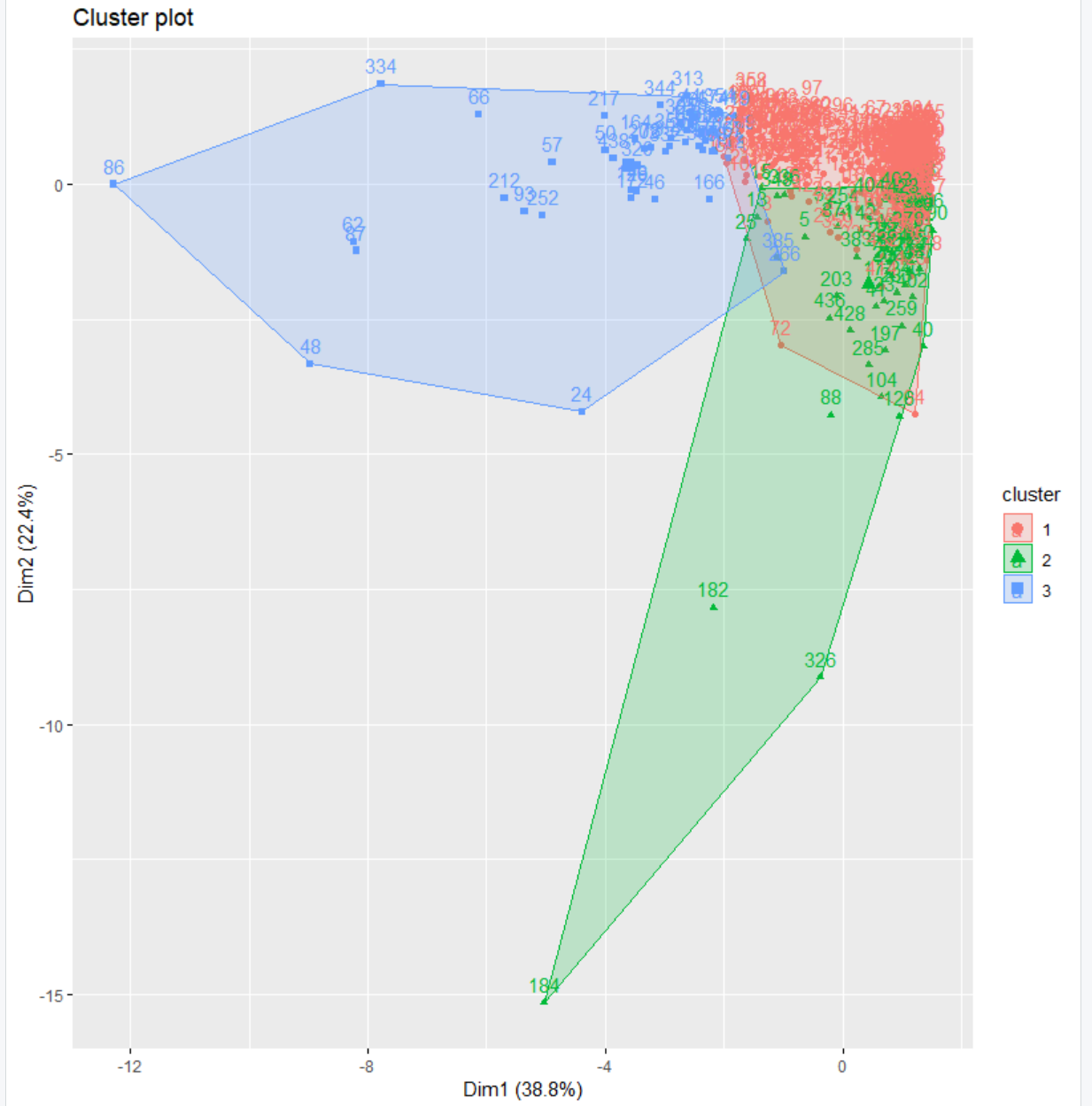
The result of the of the data appear to get a bit more difficult to distinguish between one cluster to another as more clusters are added. For k 5 there seems to be a real overlap that makes it hard when k = 5. I need to find the ideal number of clusters for this analysis and to do this I will use the elbow method and silhouette method.





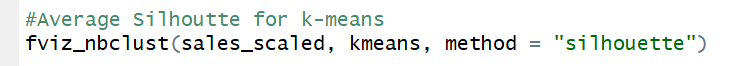
From the elbow method it is apparent that the ideal number of clusters is 3. So the plot that is ideal for showing these clusters is at k = 3 which is here below.

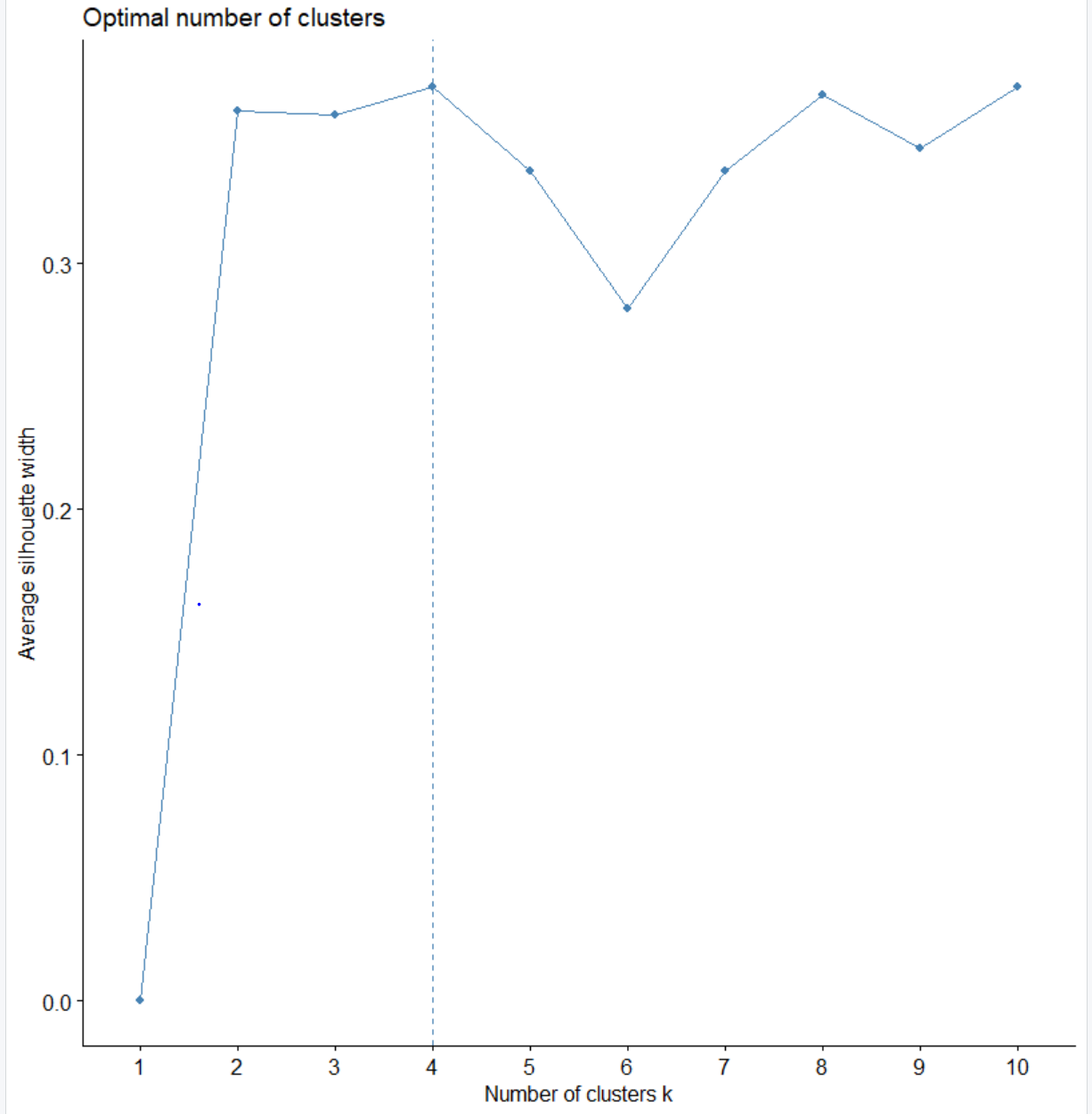




As we can see, even though the clusters share some of the same space, they are distinguishable based on the volume of data points within their individual clusters and we can see where they really grouped together.

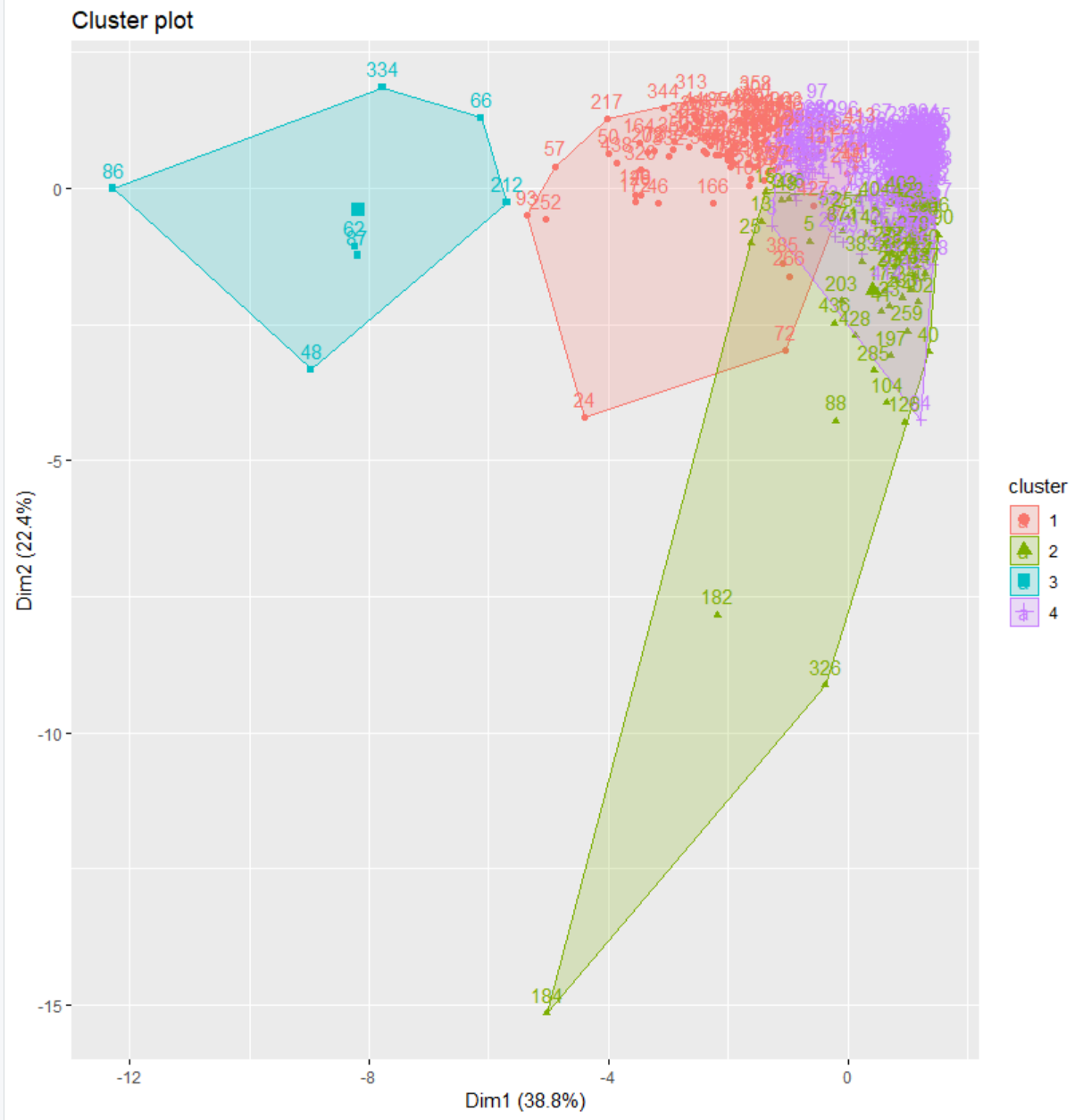
Now using the silhouette method:





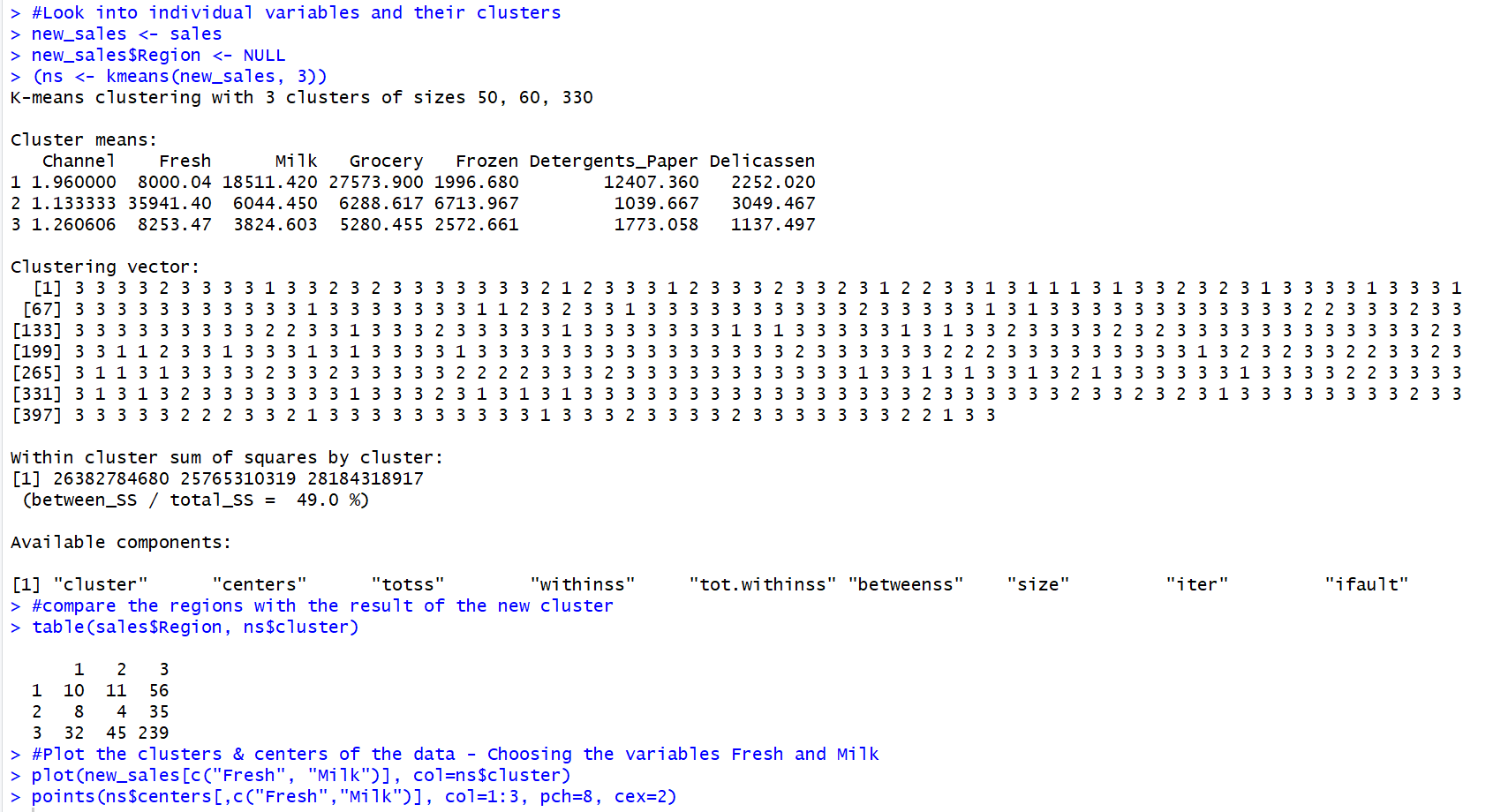
The optimal number of clusters from the silhouette method is 4 which is one more than from the elbow method. The plot for this is looks like this with four clusters:

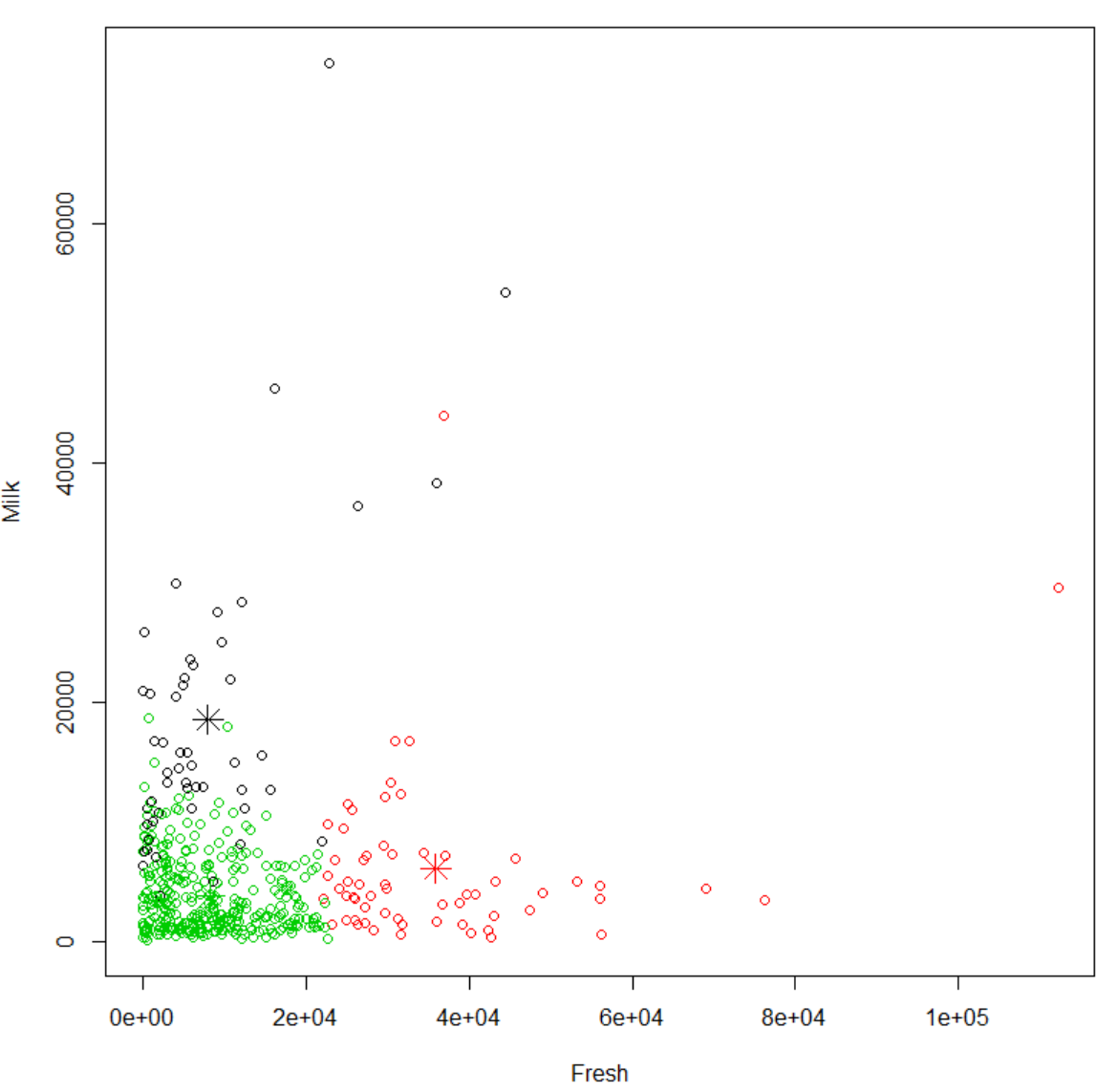




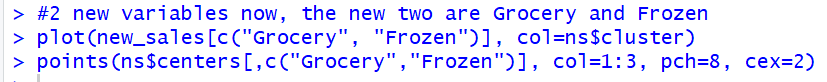
This is the look of the four clusters plotted out. It appears the clusters are all centered around the top right of x-y plane but cluster three is off in its own area away from the rest of the clusters.

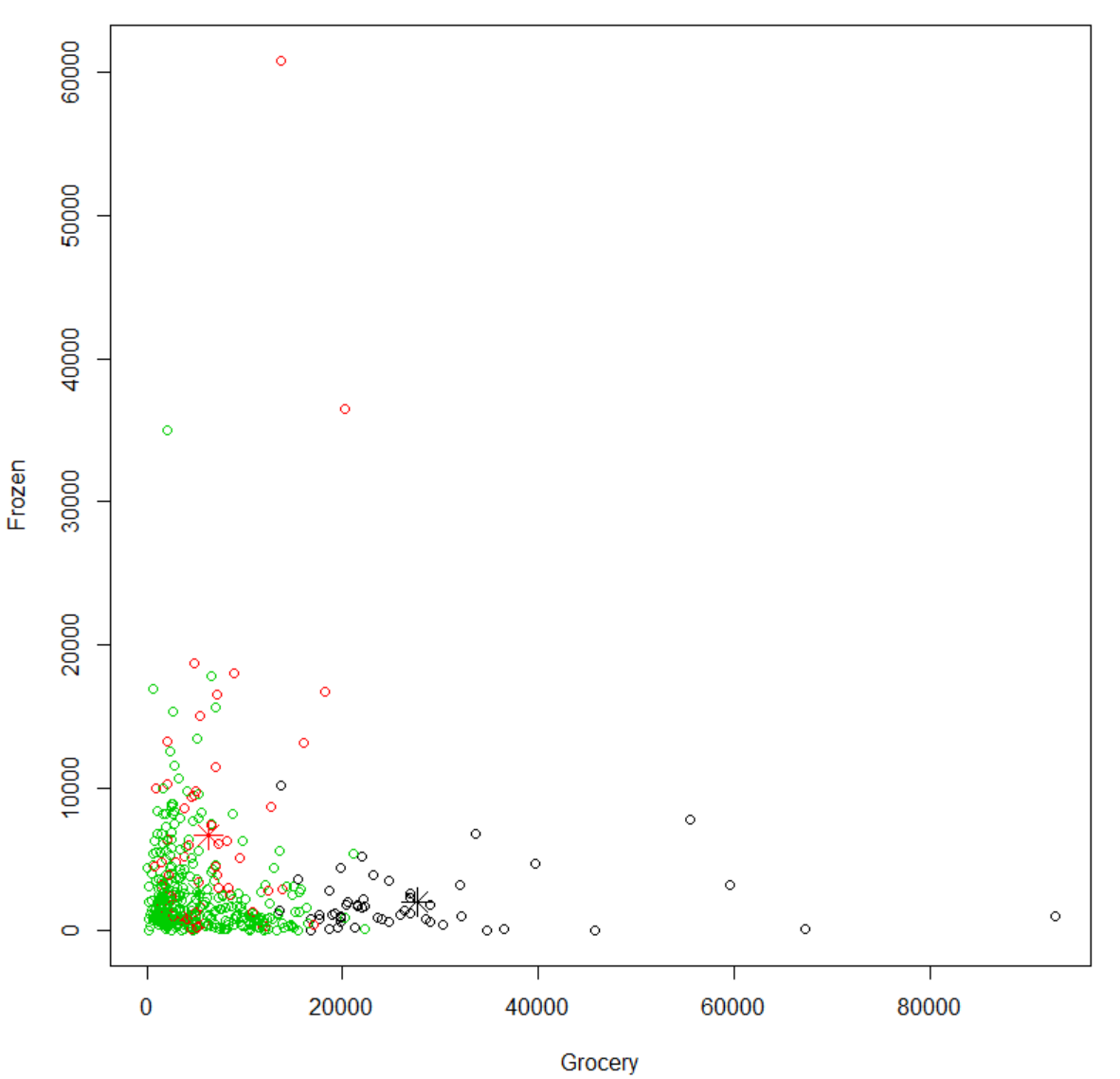
I’m now going to choose two variables, plot and mark the centers of the clusters to see what those centers are based on individual fields. The individual fields I will choose are fresh and milk and compare them across the three different regions the data has.





The result of looking at the three regions and then the two variables of fresh and milk are shown above. The data is clearly marked and color coded so we can see the difference in the regions and the star symbol shows where the center lies for the two variables. It would appear that for milk the center is around (1e+04, 19,000) and though the scale for fresh is extremely large numbers the center for fresh is at just under (4e + 04, 5000).

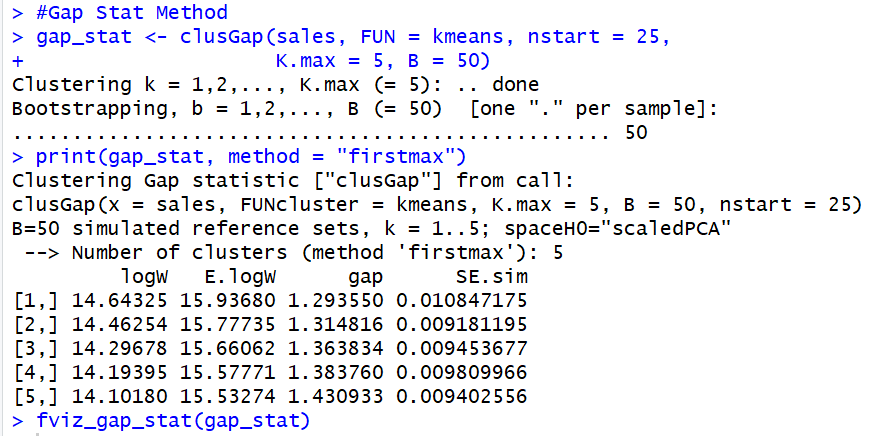


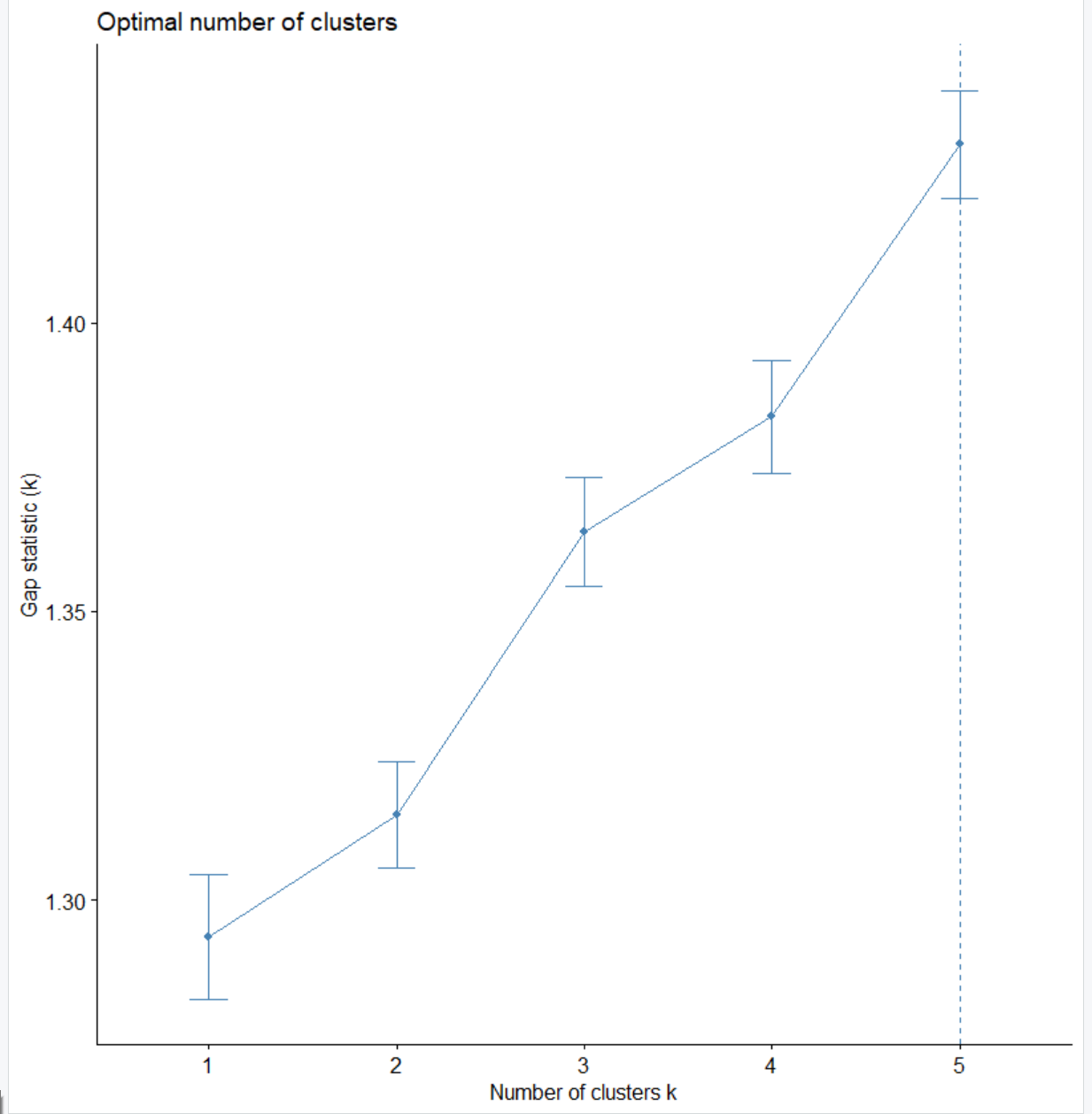


The result of the two new variables is different from the first two variables plotted. With Frozen and Grocery, frozen ended up centering fairly low on the scale, around (< 10,000,5000) on the scale, while grocery ended up being around (30,000, 2,000).

Finally, we will judge the distances of the clusters using the gap statistic method.

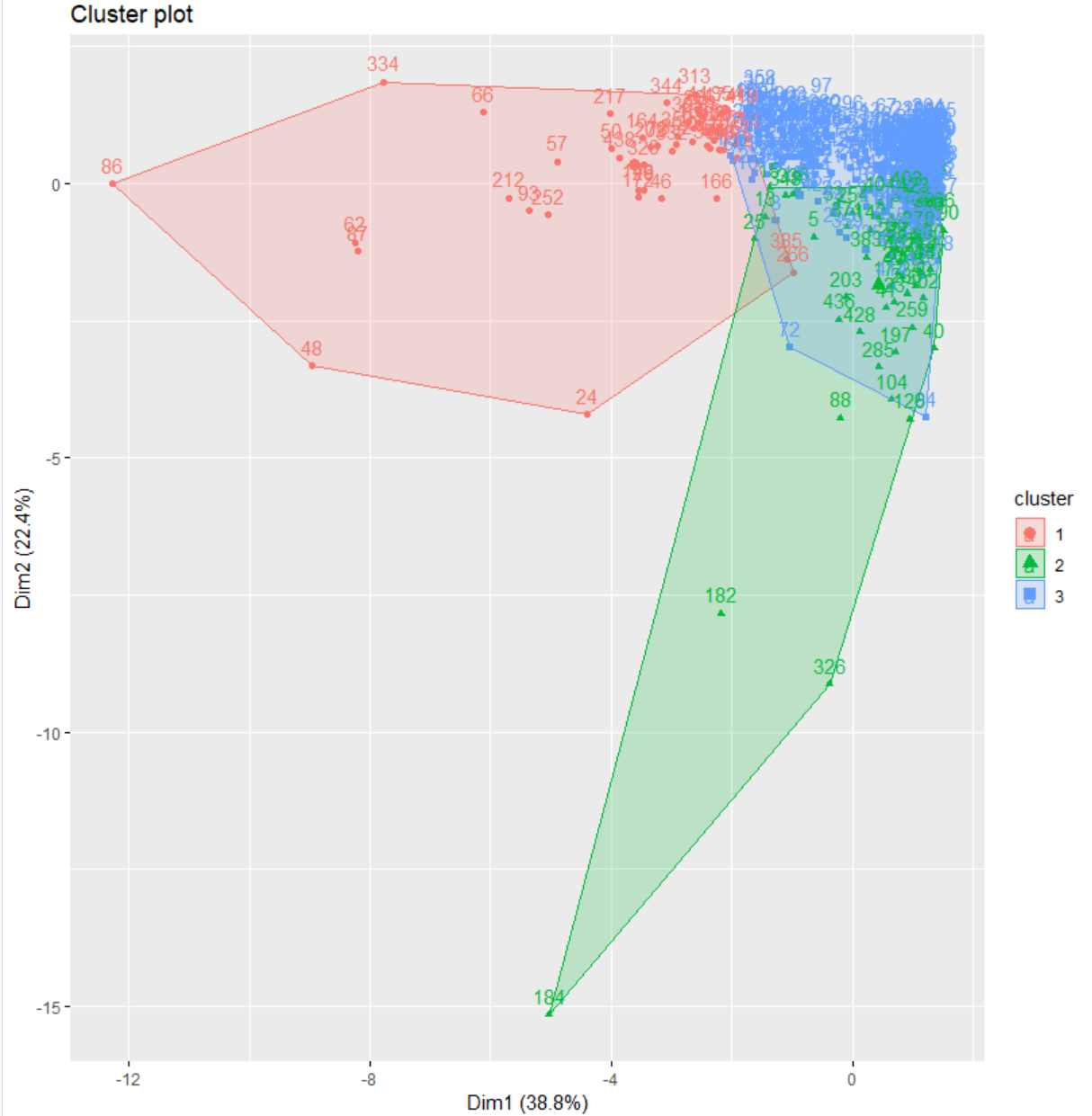
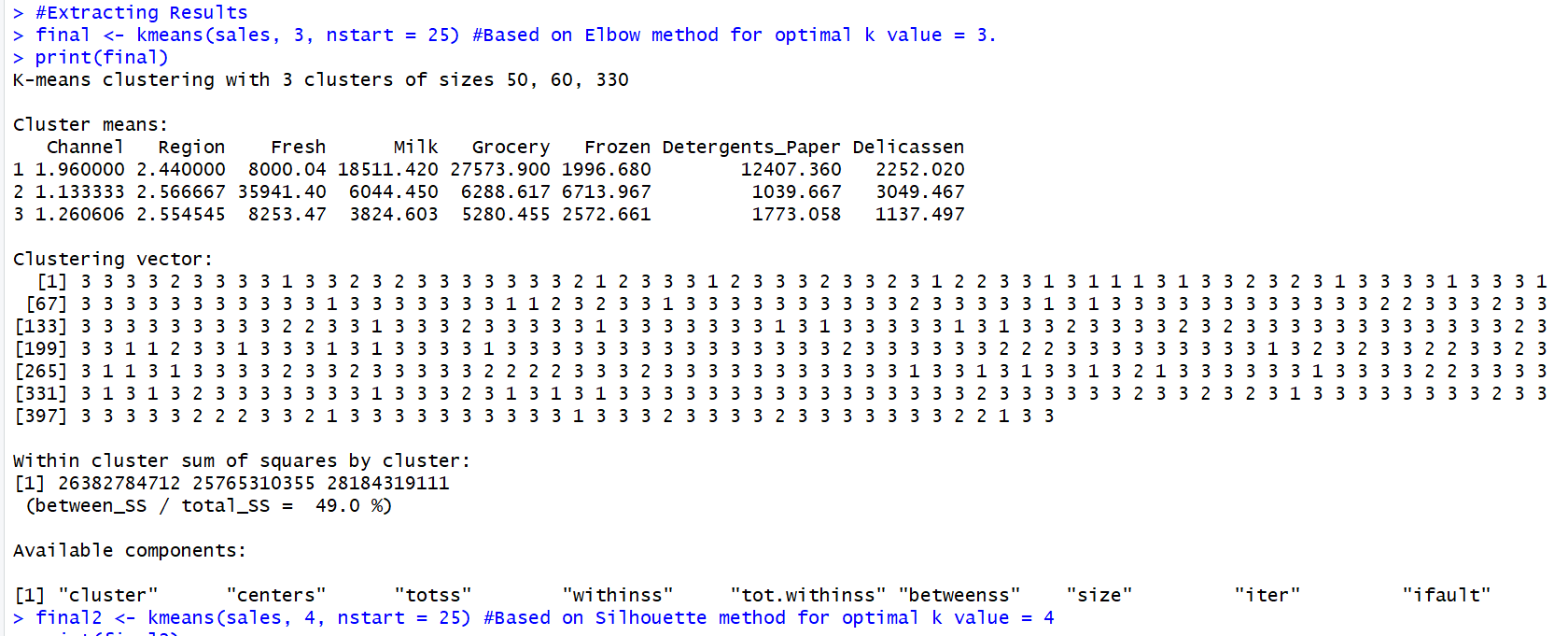
Start by constructing the Gap Stat method, setting it to the sales data that I have and setting K = to a max of 5.



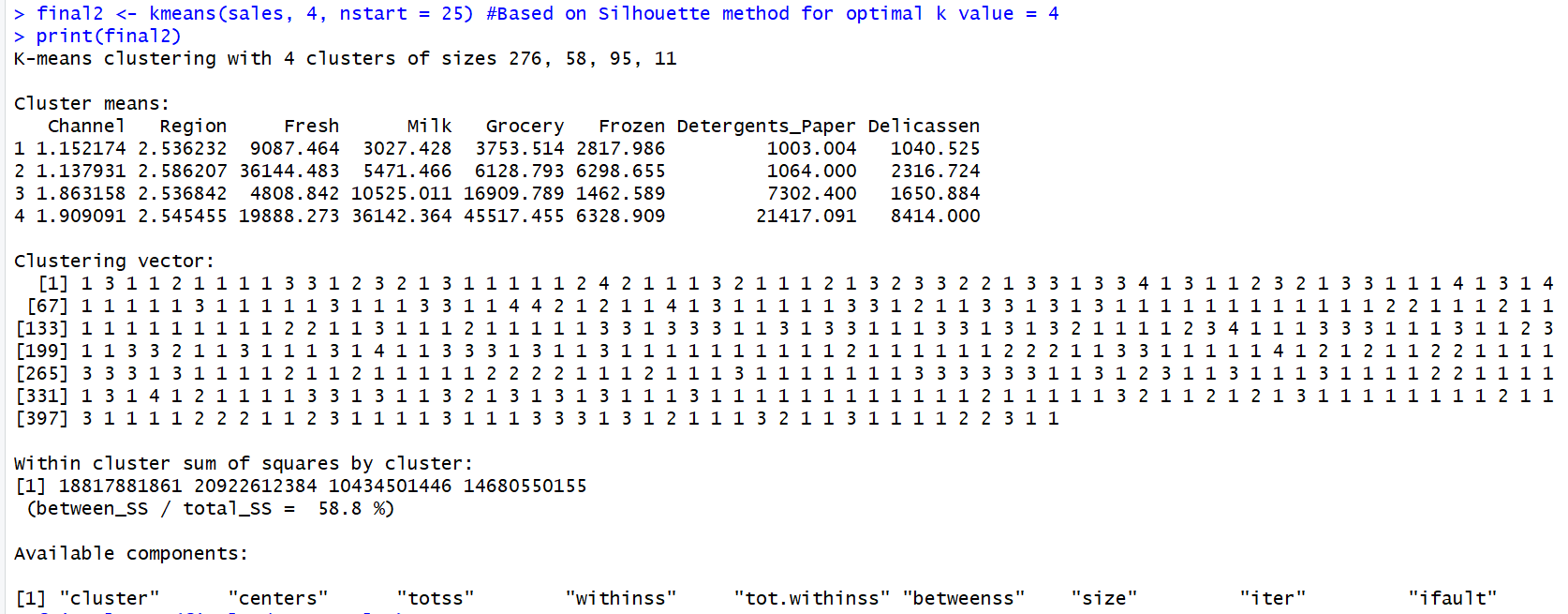


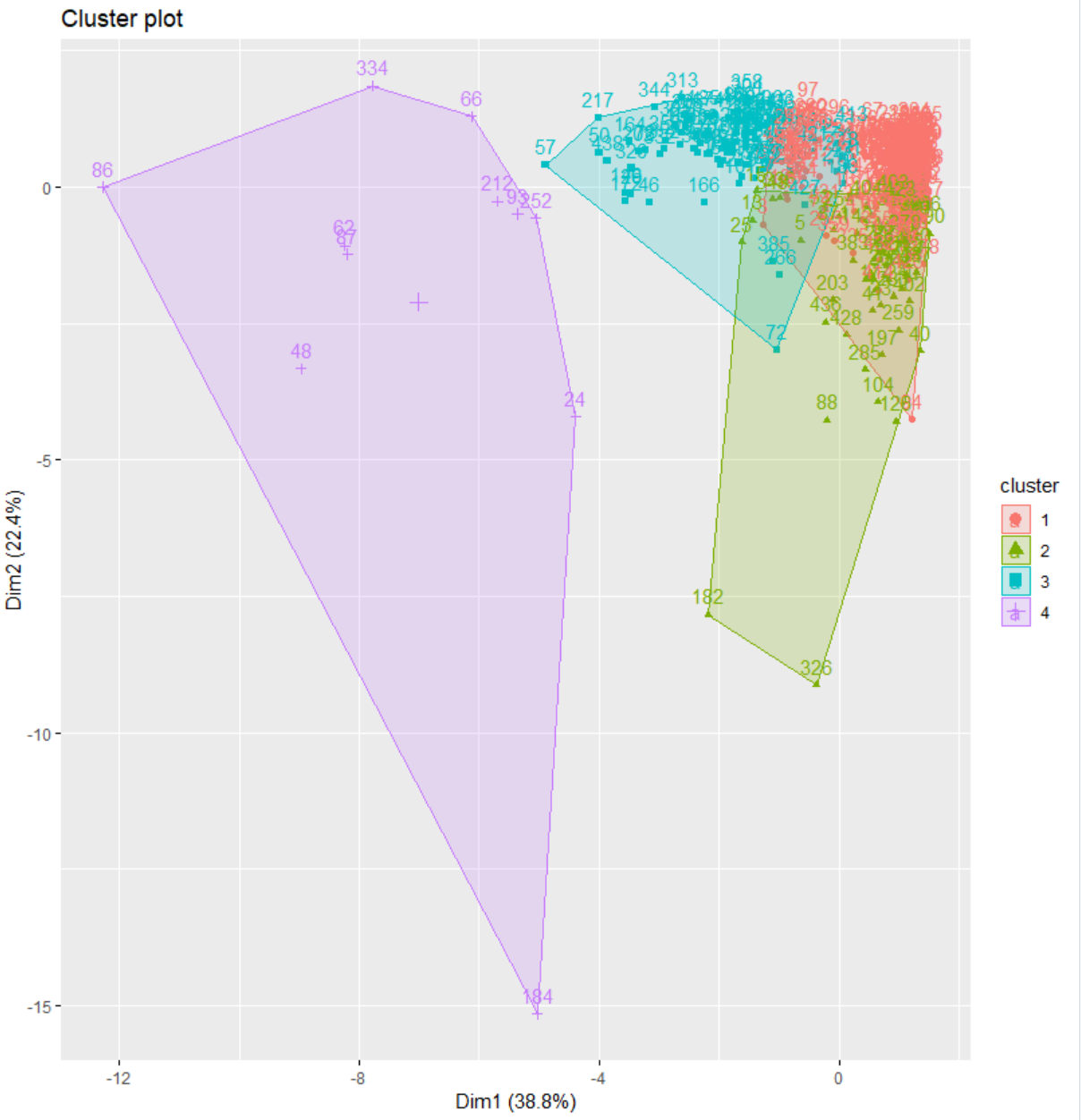
From our result it appears the gaps really aren’t as big as they may appear. From gap 1 to gap 5 there is less than a 0.15 difference. This is misconstrued in the RStudio graph because of the y-axis scale not starting at 0. But it is easy to see from looking at the results of the clusGap command. The closest clusters are clusters 1 and 2, which is about 0.02 distance and then also clusters 3 and 4 are a small distance of a little less than 0.02. These cluster pairings that run closer together signal how alike these clusters are to one another. The fact that all 5 of the clusters looked at in this method are pretty close shows how close all clusters are in an analysis when k = 5, but you can also look at the charts from when k = 3 or 4 and you can get the same result that a lot of the results are similar to one another because the clusters are so close to each other.

The last step of this k-means analysis will be the analysis of extracting the data and setting it according to the optimal k values from the elbow method and silhouette method.



From the result of the plot, the 3 clusters are very close together meaning there is a lot of similarity between them. We will now run the same process but again at k = 4 which should yield a similar result in the clusters.





This result of the silhouette valued k of 4 yielded different results, where cluster 4 is not overlapping any of the first three clusters. But this could also be the case because of the wide discrepancy of data points cluster 4 has compared to the others. Where the smallest of the first three is cluster 2 with 58 and cluster 1 is at a seemingly giant amount at 276 data points.

#### ***Questions to answer on K-means***

##### Find the optimal number of clusters with methods such as the Elbow method, silhouette method.

The optimal number of clusters changed when going from the elbow method to the silhouette method. The elbow method output 3 clusters, while the silhouette method output 4 in total.

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

This was done above at the end of my k-means analysis. I started with looking at fresh and milk and then I switched and decided to look at grocery and frozen.

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

When I originally started plotting data before bringing in any methods to find the ideal number of clusters, I found when I set k = 5 it was hard to find one of the clusters, but once I ran the elbow and silhouette methods that cleared up how many clusters that should be used and it made it easy to read the clusters when they were plotted.

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

The distance between the clusters is shown in the gap statistic method where the distance between the clusters is measured. The findings are also concluded upon the completion of that analysis.

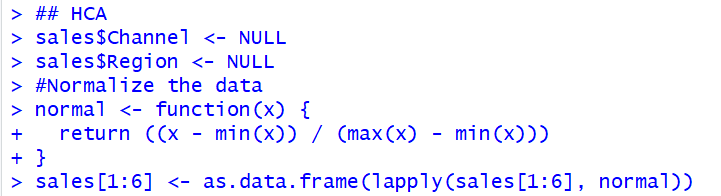
1. Do you have any concerns on the experiments?

I don’t have any concerns with the experiment, but I would like to be able to see what would happen when the process is allowed to pick the number of clusters itself instead of me controlling it or calculating the number of clusters with a different method.

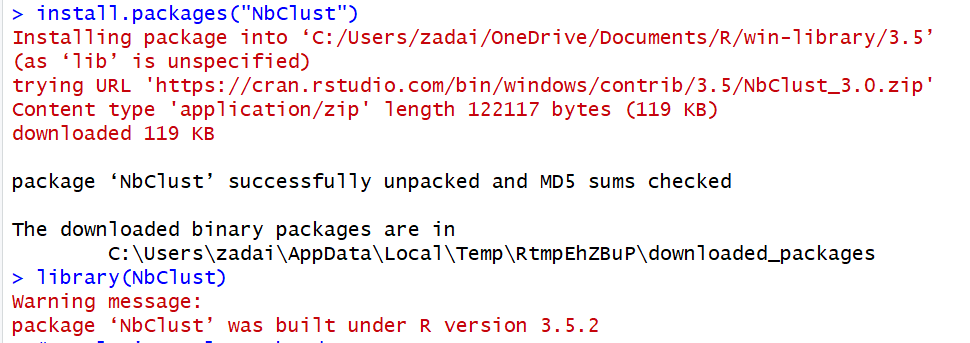
##### ***HCA Analysis***

Now that I have finished the K-means analysis I will now work with the sales data and do hierarchical clustering as well as create a dendrogram of the data.

I’ll start by setting both the variables Channel and Region to NULL to help normalize the data.

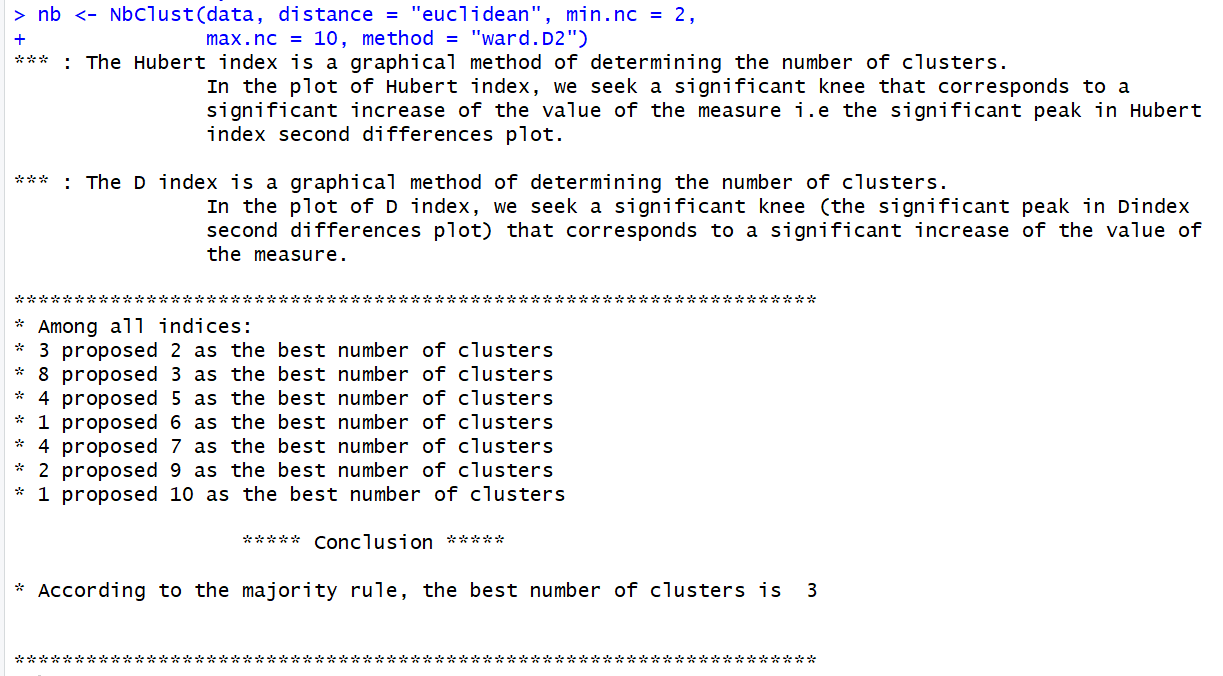


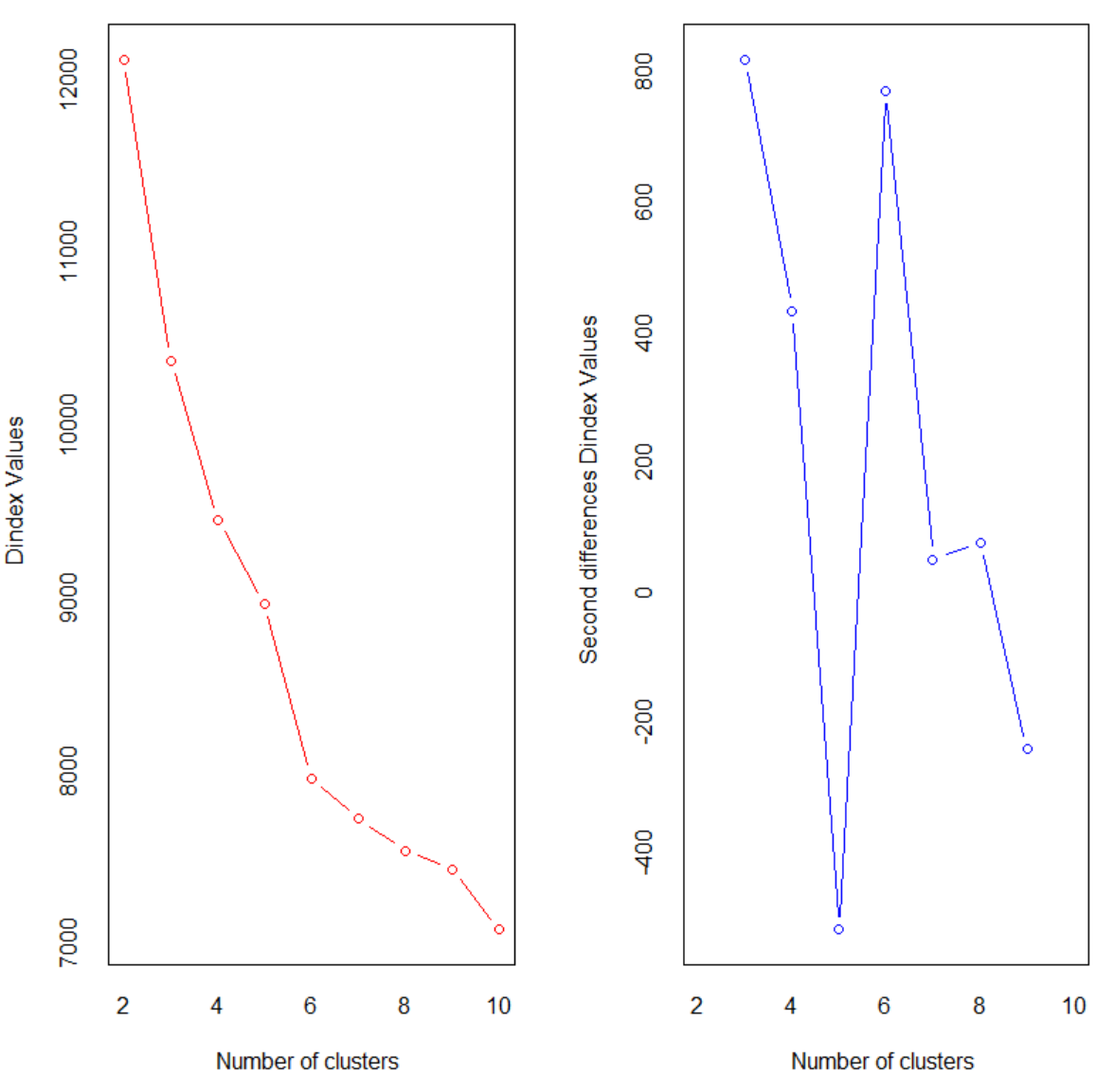
Next, I’ll bring in the NbClust library.



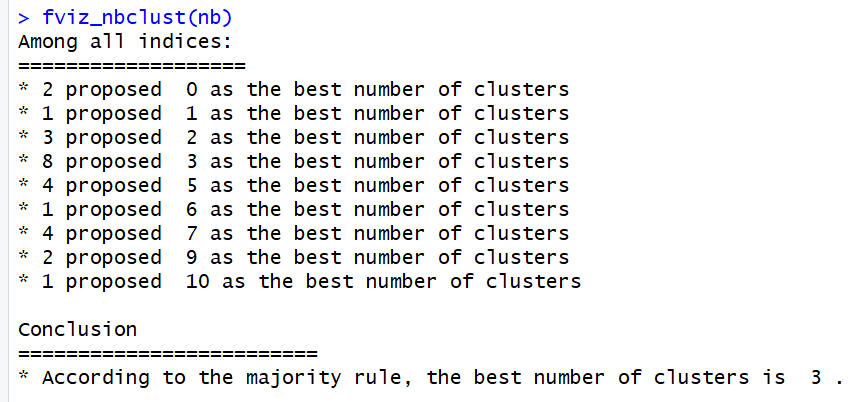
Now that I have my data right and the libraries I need, I can start the analysis.

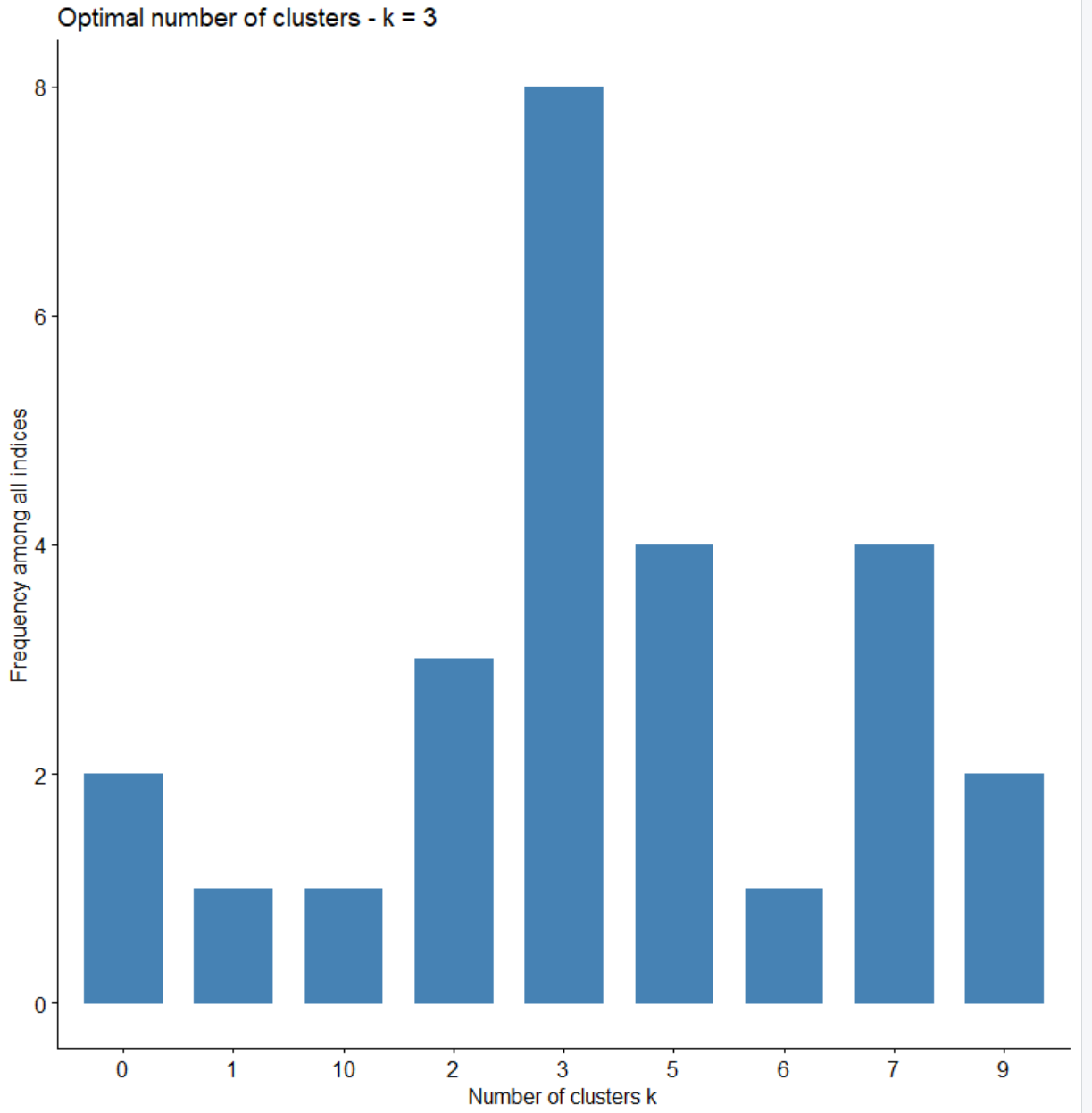
First start with calculating the proper number of clusters.



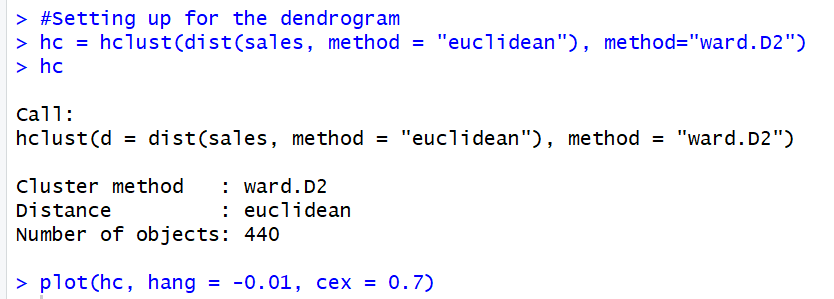


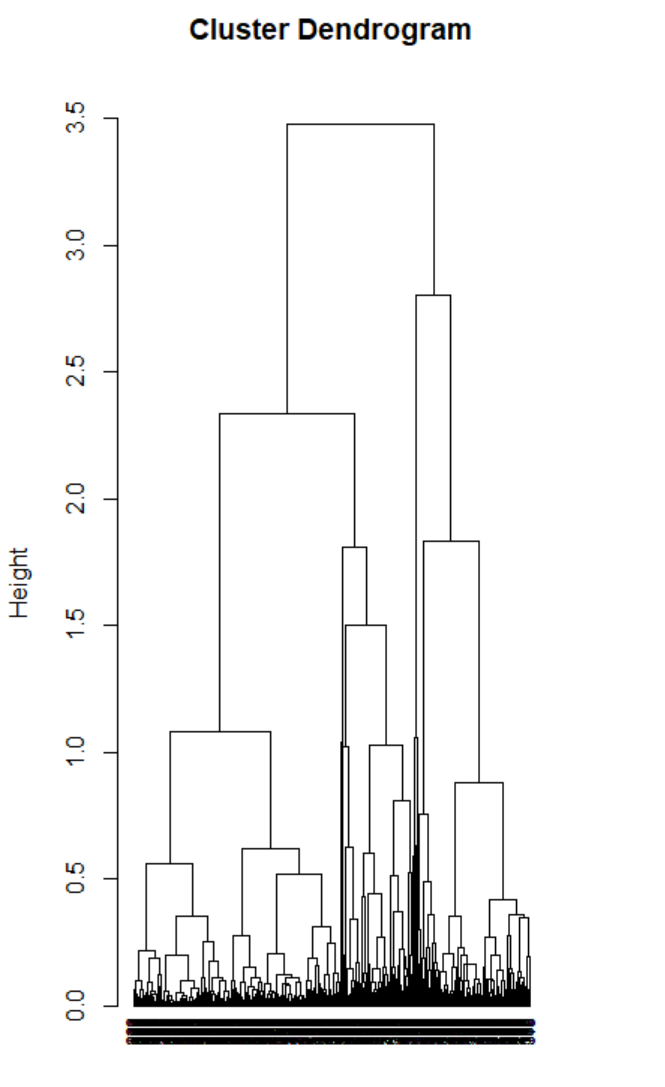
It was distinguished that the best number of clusters appears to be 3.



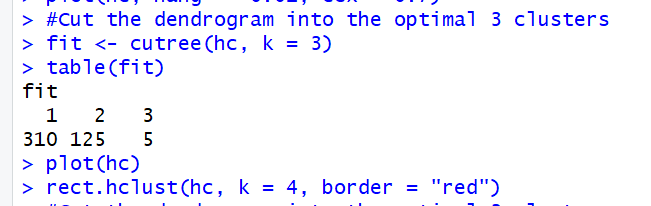


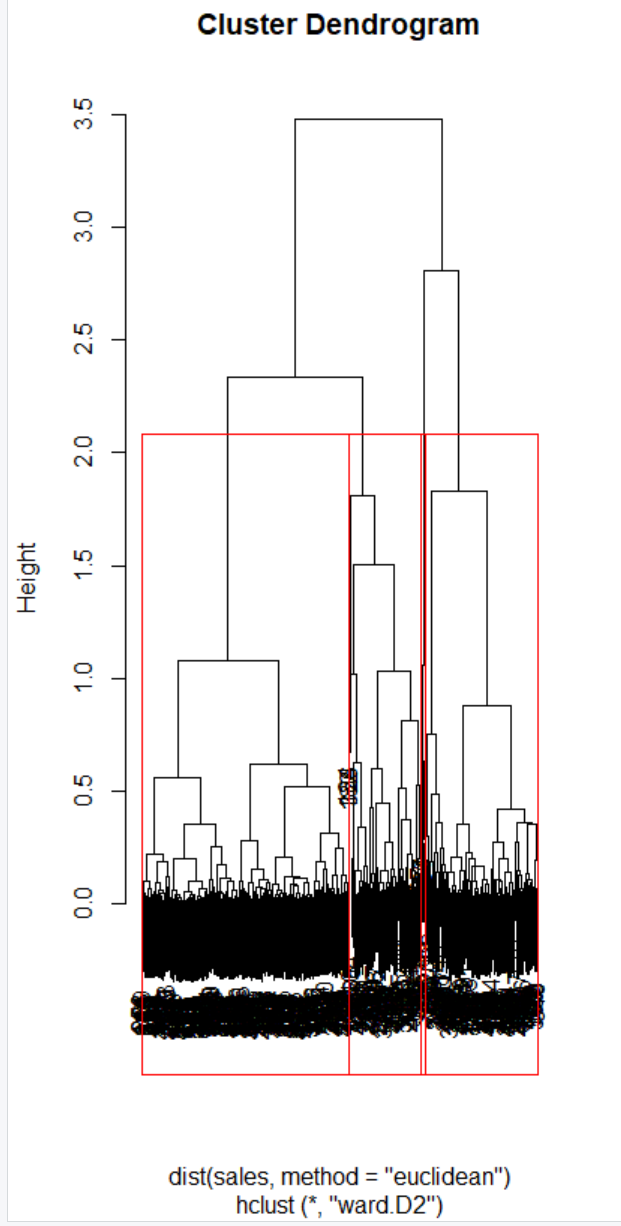
Now it’s time for the dendrogram.





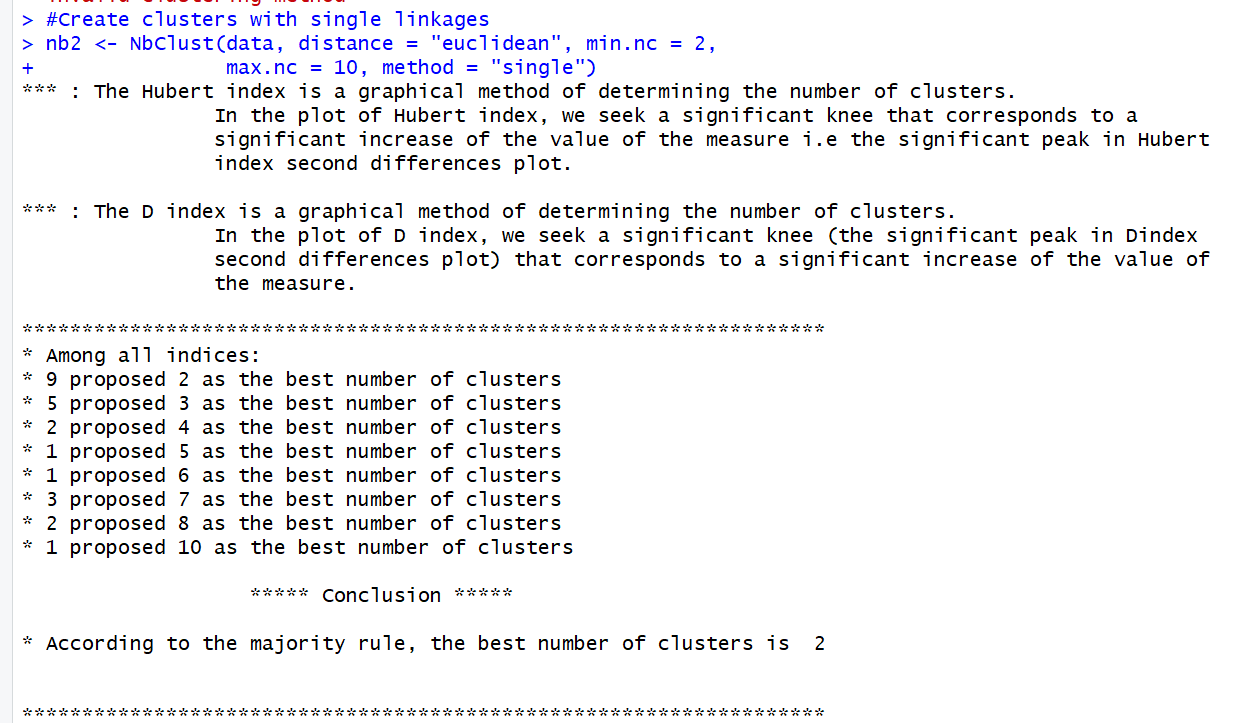
With the dendrogram now set up we now need to cut it into our optimal 3 clusters.

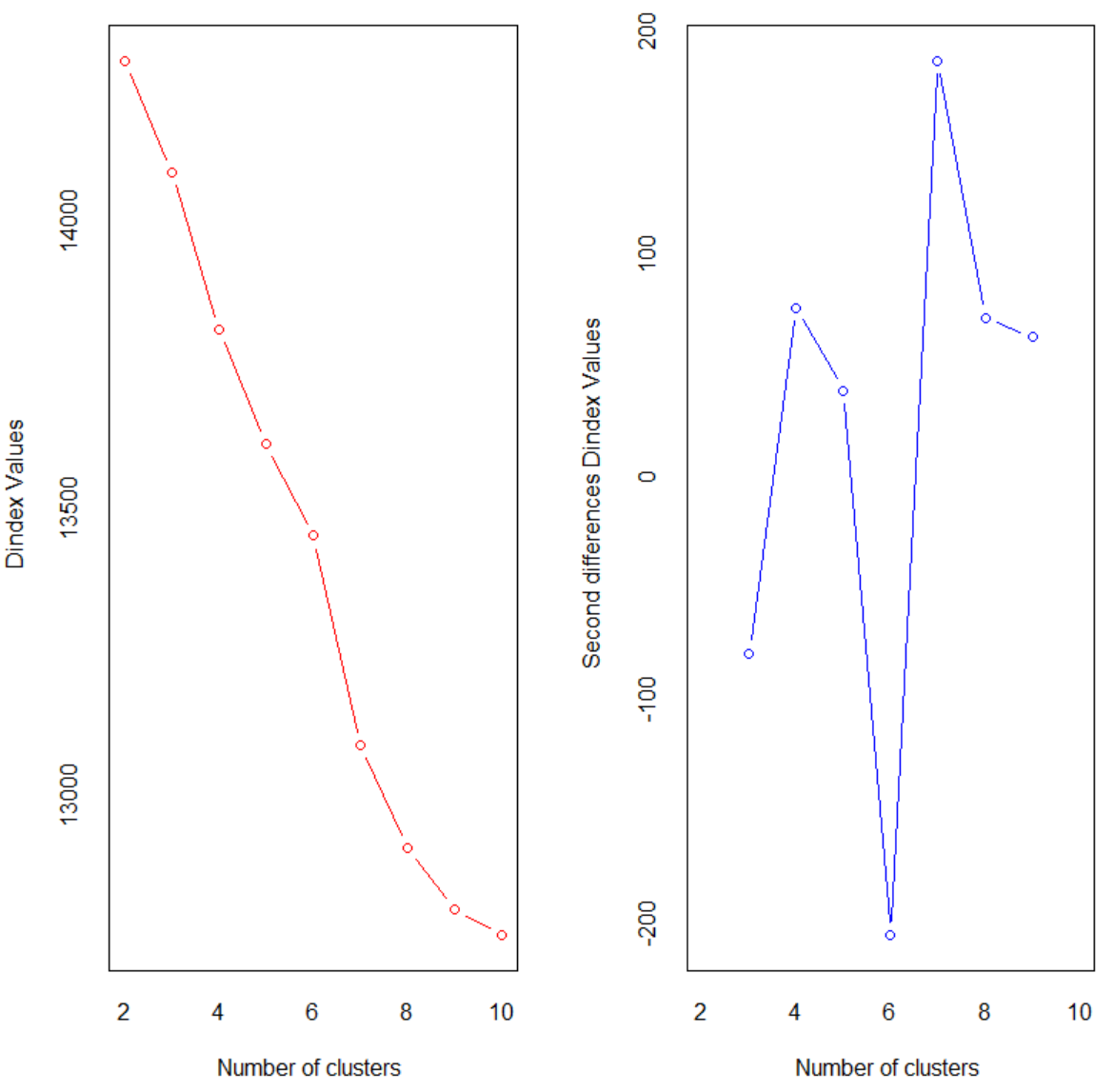




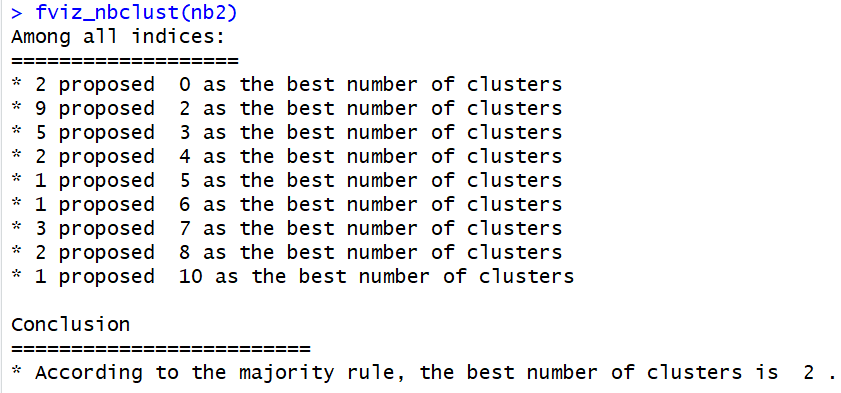
As we can see above, the clusters are split up and plotted upon the dendrogram.

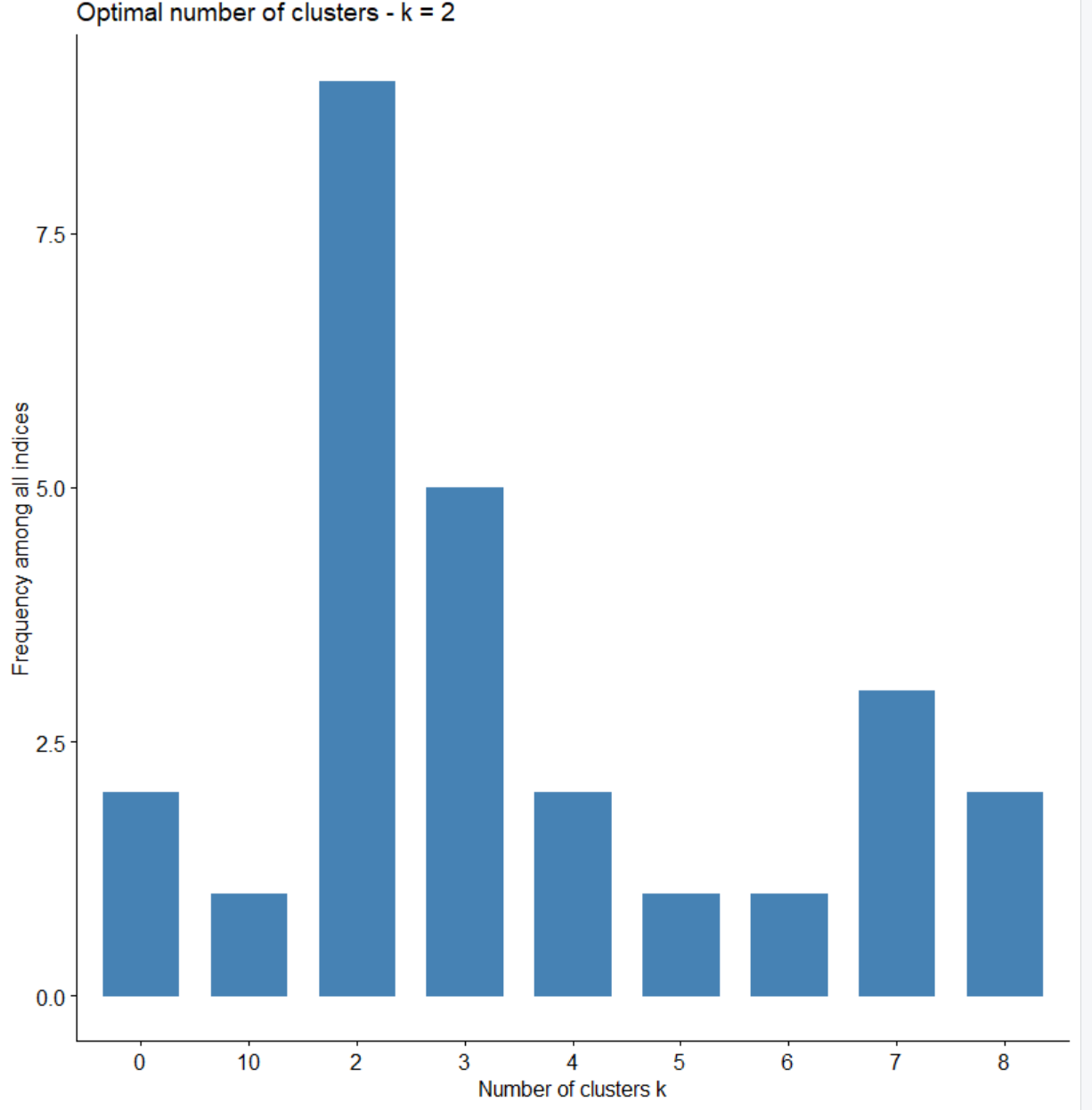
Now I’m going to redo the process, this time for single linkage of the clusters.

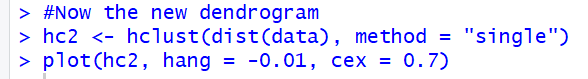


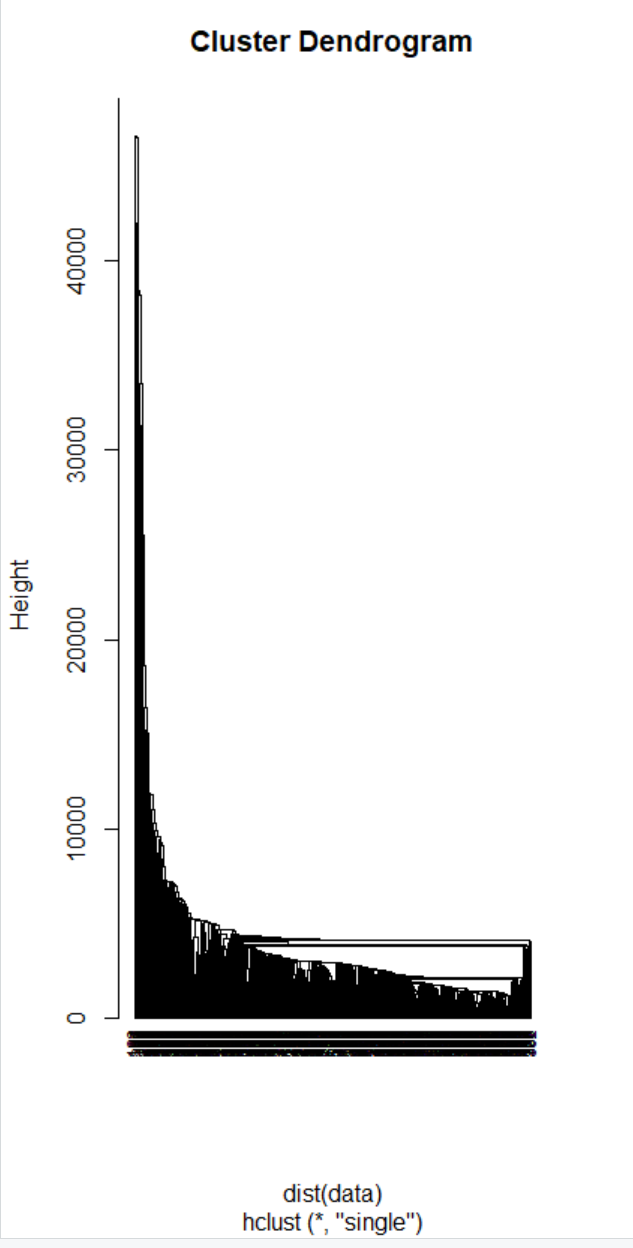


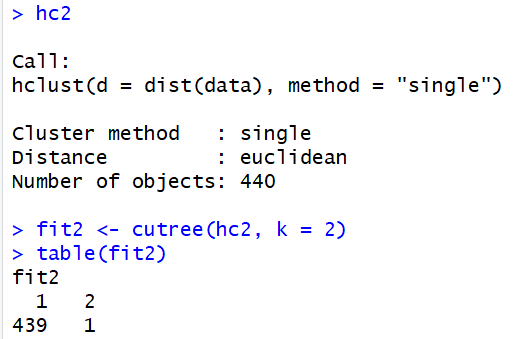
This time the optimal number of clusters is 2.



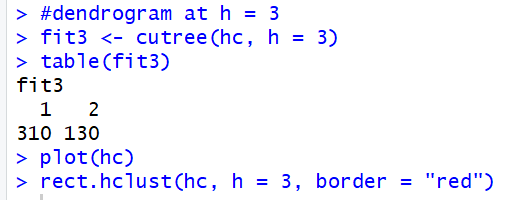


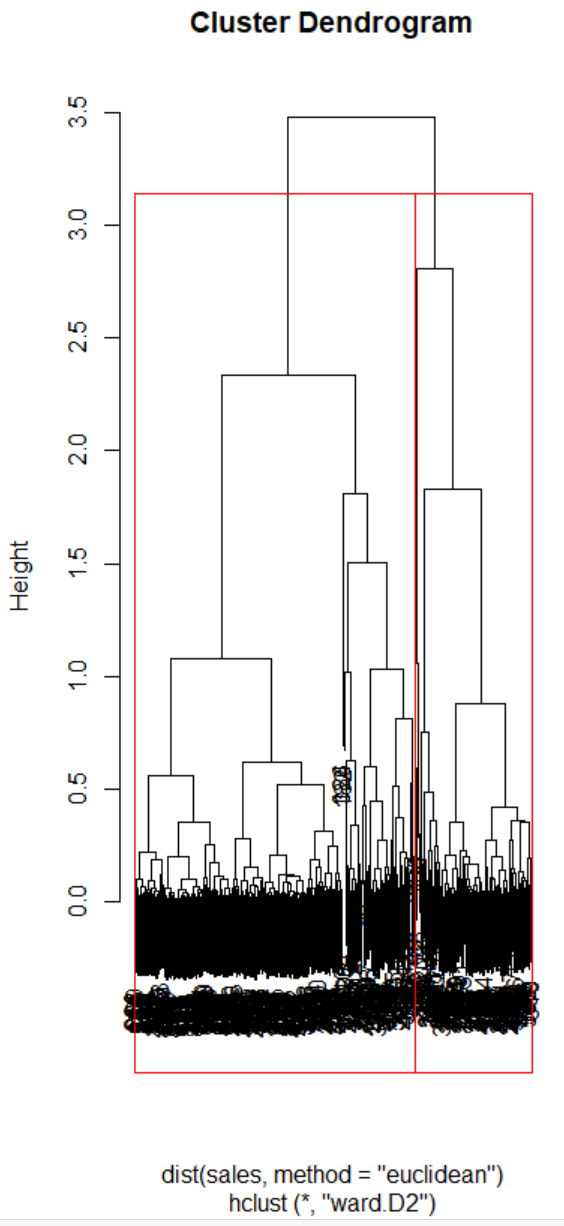




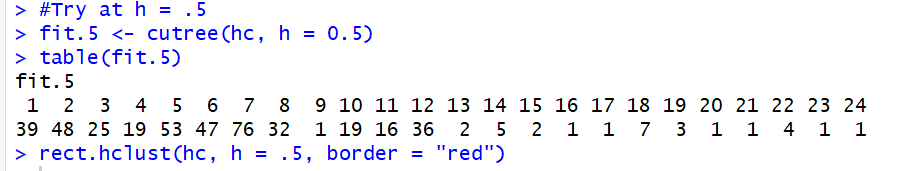


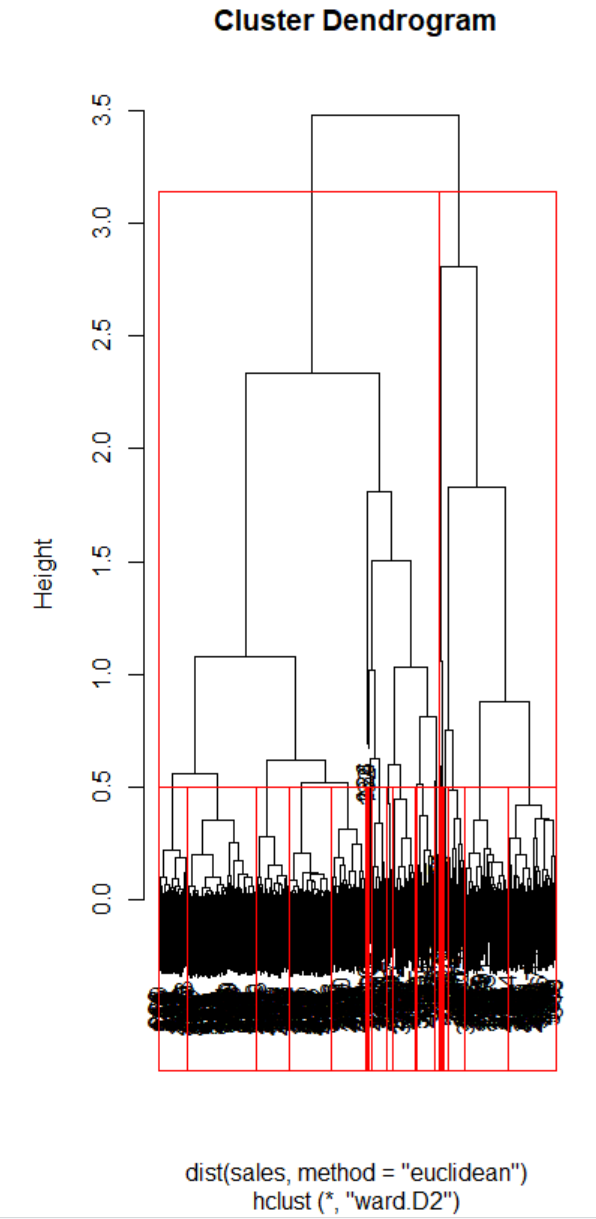
It appears with the single linkage approach there wasn’t any meaningful clustering taking place, so I will move back to the original dendrogram.



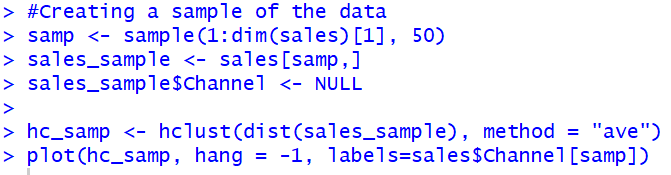


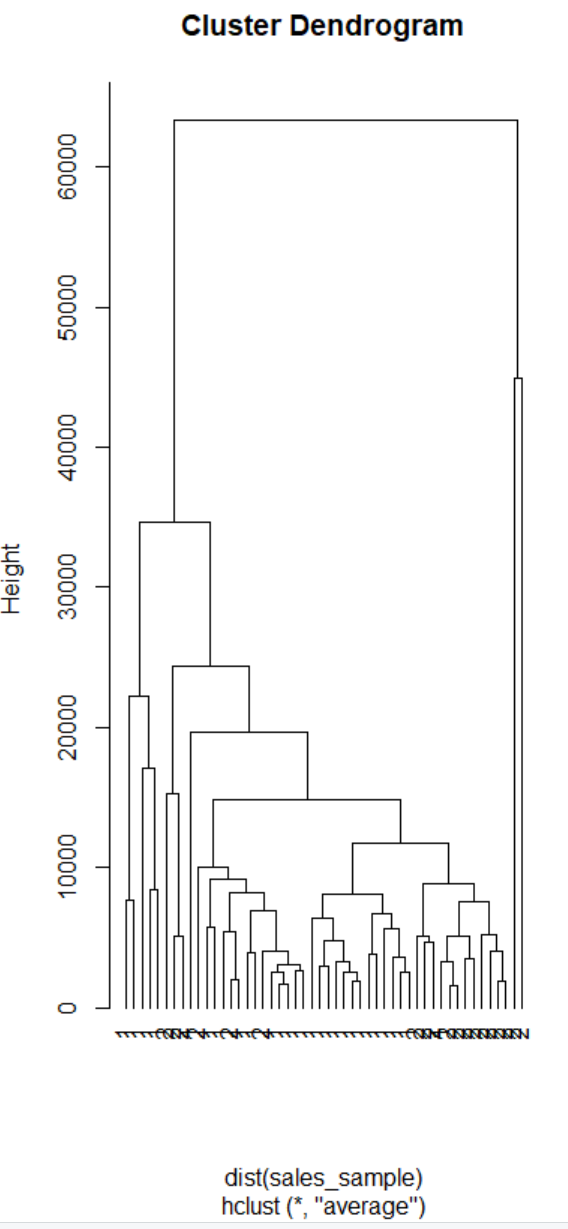
Now trying at h = 0.5



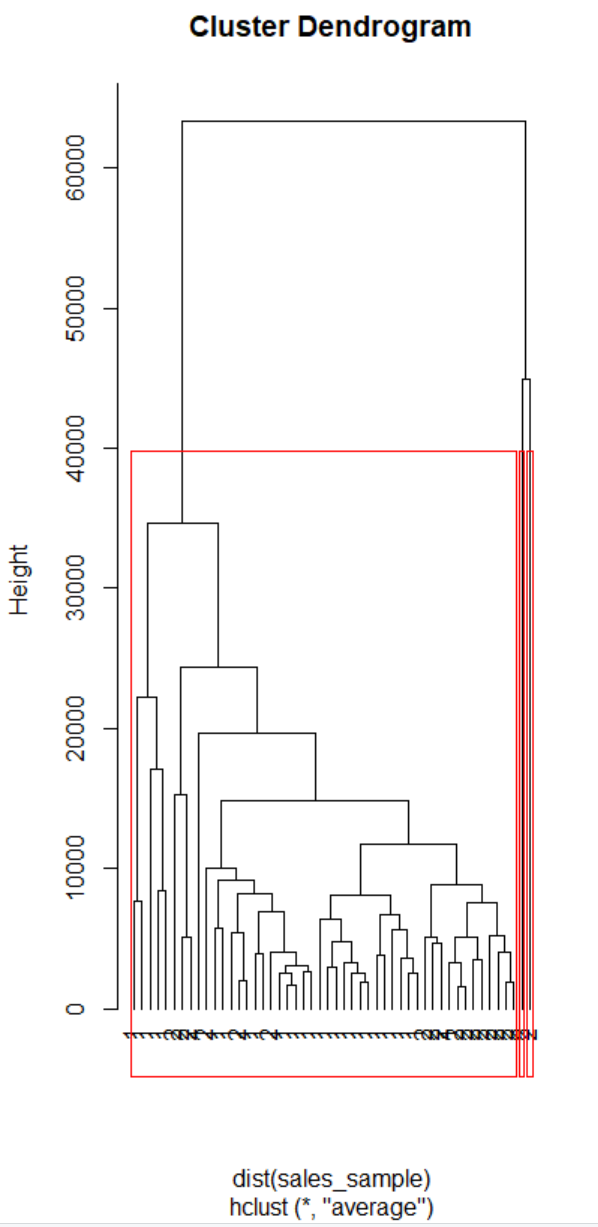


No matter how the dendrogram is cut it appears to have too many observations, so from here what I can do is try and create a sample of the data and hopefully that should help with reasonably cutting this data.

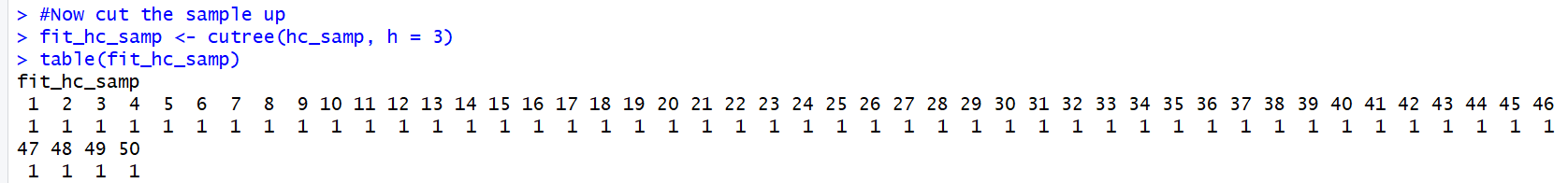








The dendrogram is much easier to get meaningful information from when the sample size is cut down from 440 to 50. The only problem is, it appears each of the sample entries are to channel 1, and not any other channels.



So unfortunately there isn’t a big difference in the clusters, but the dendrogram is much easier to read so that is a benefit of being able to take a sample of the dendrogram like I just did.

**Questions for HCA.**

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

This was done above, my optimal number of clusters was 3 and using NbClust I was able to find that number

1. Draw a dendrogram

I have several drawn above.

1. Cut at different heights of the dendrogram. Draw it. Discuss and summarize your findings.

I cut the dendrogram several times with the full data and once when I took just a 50 entry sample. The results are discussed above.

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

The concern I have for the analysis primarily comes from the k-means, because I was unable to create a reasonable spread for clusters, it was always heavily skewed so that only one cluster had seemingly 99% or more of the entries in the data, no matter if I took a sample of the data or took the full dataset.

**Summary**

For this assignment I took the data from a wholesales grocery store and ran an K-means and HCA unsupervised learning algorithms upon the data to check out the different clusters these methods would sort and put the data together. This was a great learning experience because I was able to see the data being used in two different methods and be able to find out which one I would prefer more depending on the data I have. From this data, I liked using the k-means more, but that is just because the HCA clusters were so skewed it was difficult to get any meaningful result from it. If they were less skewed, with either the full data set or the sample, I feel like I would have had more actionable results to pull from that.