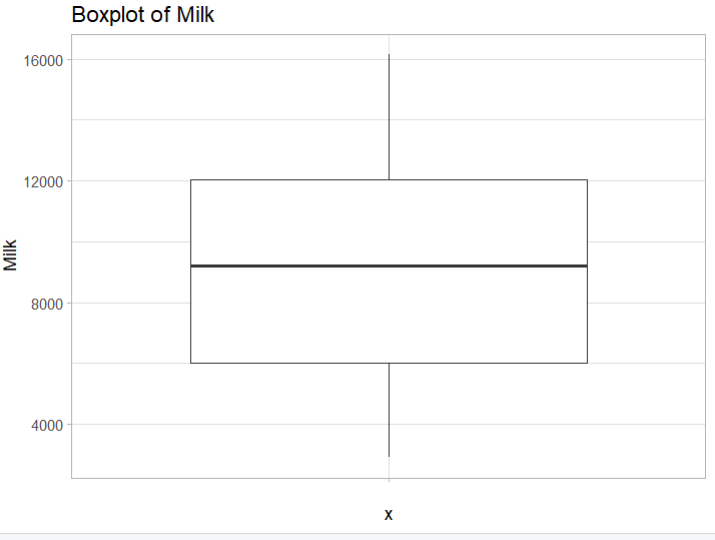
**1. Data Transformation and Descriptive Analysis**

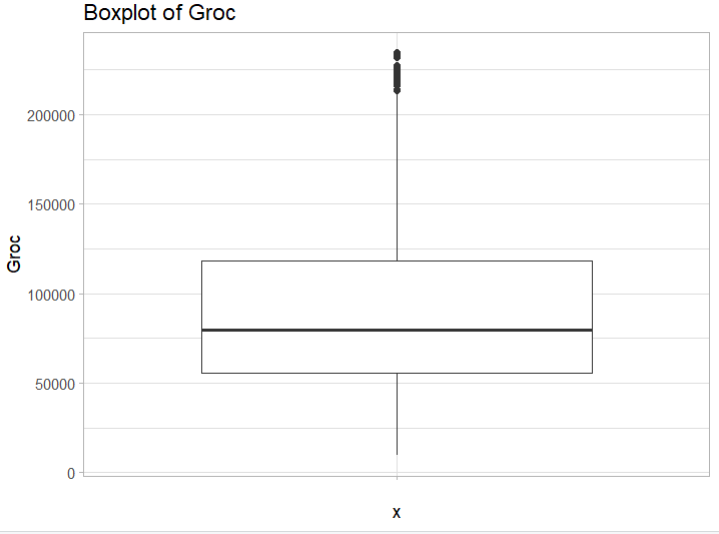
**Question 1.1: Rename all variables with your initials appended (just as was done in assignment 1)**

I have renamed all the variables by appending them with my initials NM.

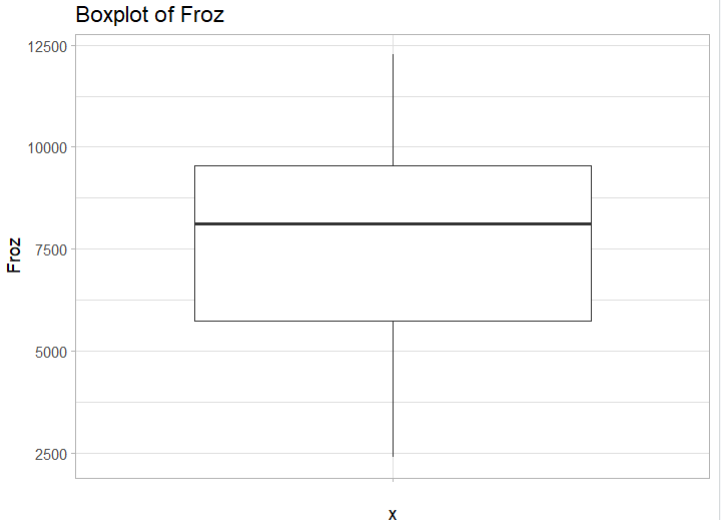
**2)** **Create graphical summaries of the data (as demonstrated in class: boxplots or histograms) and comment on any observations you make.**

****

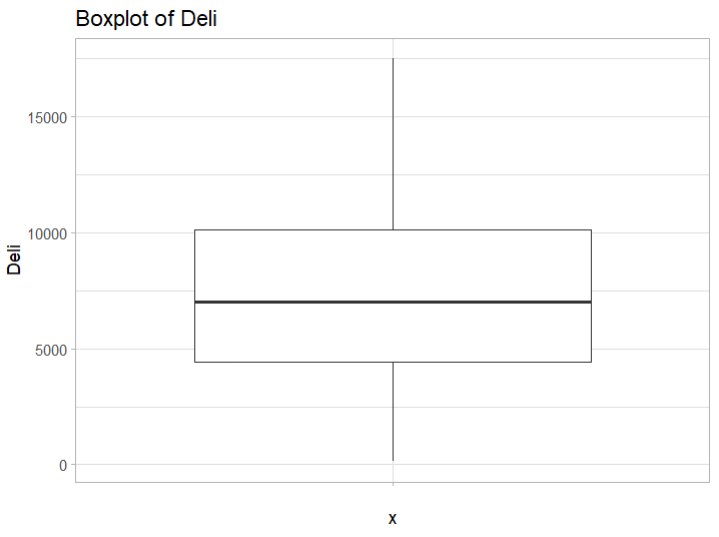
This is the Boxplot of annual spending on Milk Products as we can see the spending on Milk products is pretty even and the median lies at a good centre point that is approximately 9000. The upper and lower bounds of the Milk products aren’t that high or low either so we can infer that this is pretty normal

****

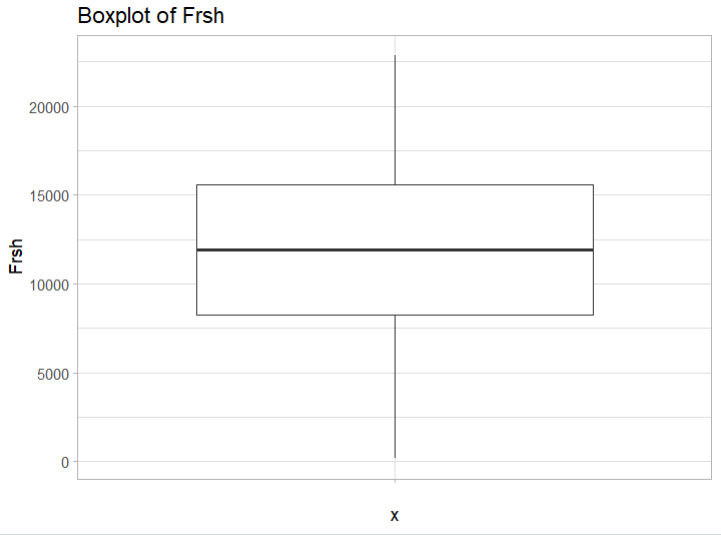
This is the Boxplot for the annual spending on Grocery and as we can see there are many outliers in this Boxplot but this can be explained by the fact that many people tend to buy different groceries and there are multiple items a person can choose to buy or not to buy. On the other hand, we can see that the box is pretty low when compared to the Milk and this is pretty self-explanatory as spends on grocery in the common public tend to be lower than most people and this is evident by the lower median.

****

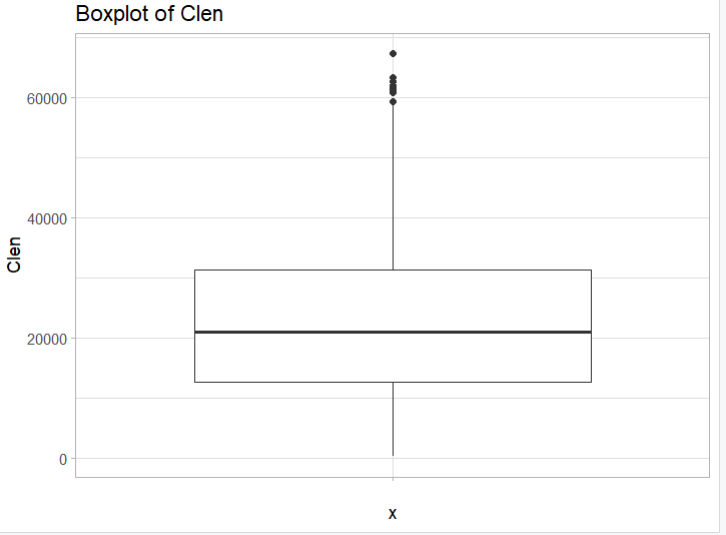
This is the Boxplot of frozen products and from this we can infer that the box is pretty high and the annual spendings are very low. According to me this can mean that people tend to spend low for the frozen products and since frozen products tend to have a longer shelf life.

****

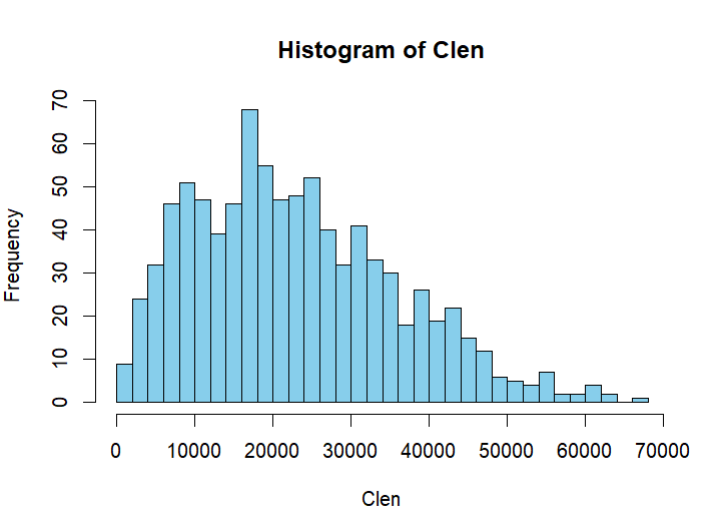
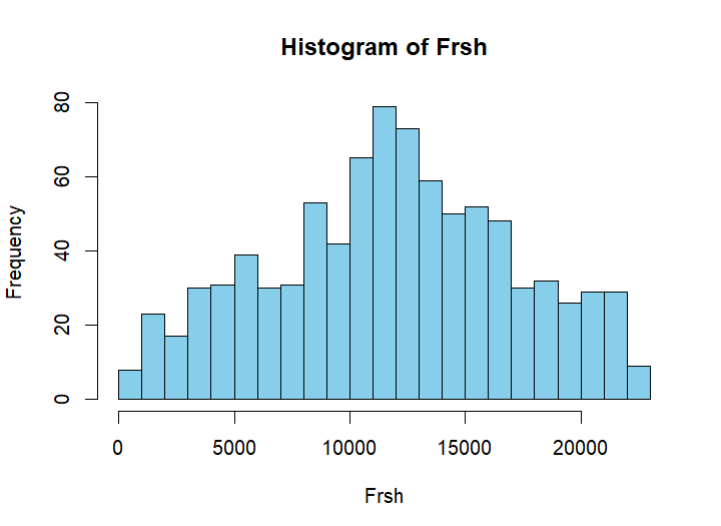
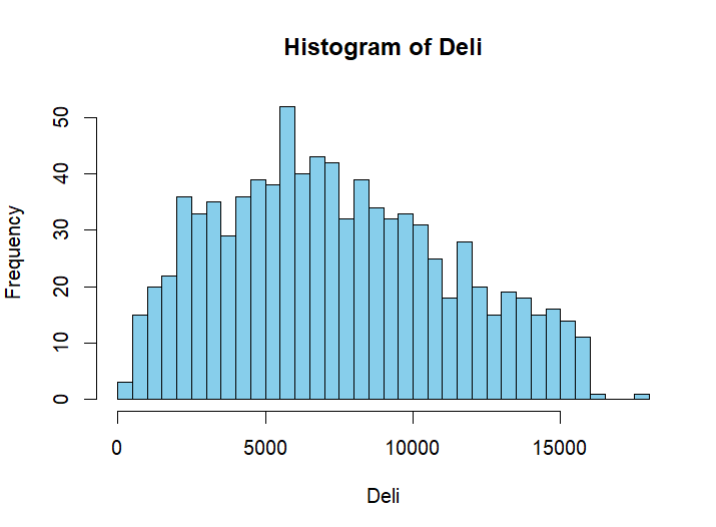
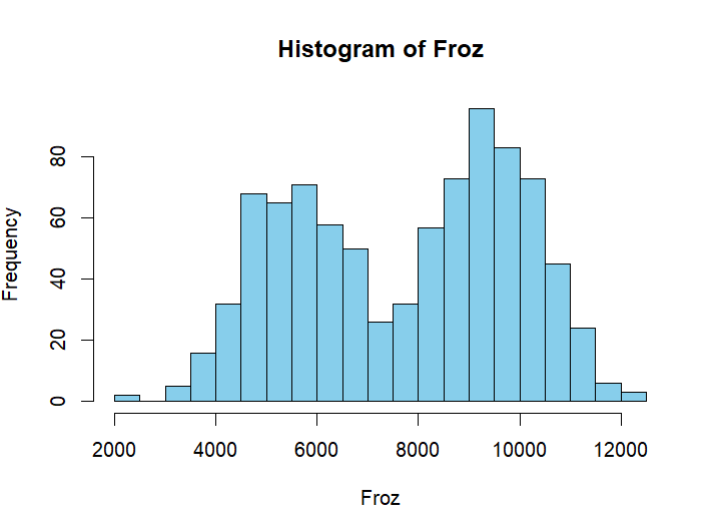
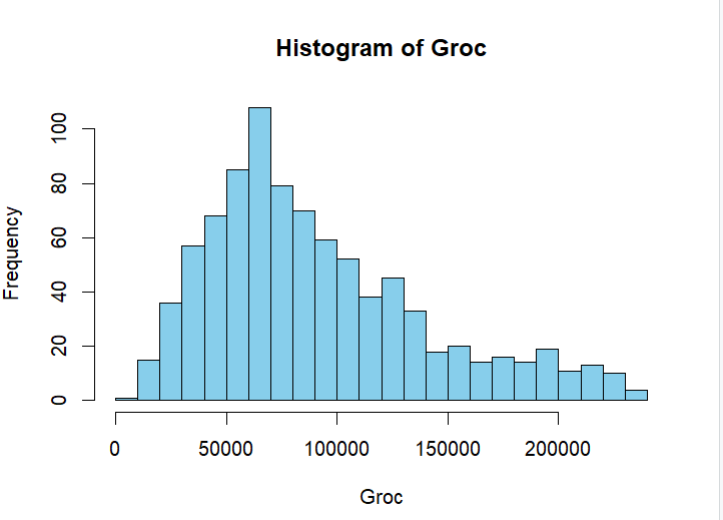
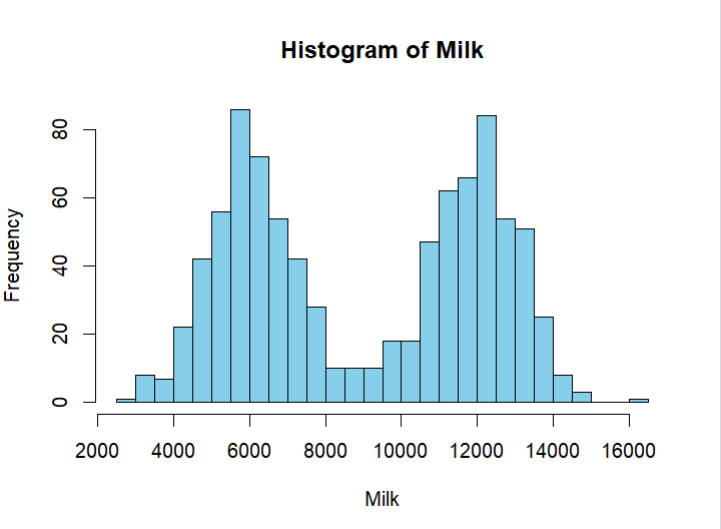
This is the Boxplot for Delicatessen and the annual spending on these products. The boxplots dictates that the people spending on Deli items range from 0 to more than 15000. This can be explained by the fact that many people don’t like to eat Deli products and that there are many expensive cuts of meat and more Deli products.

****

This is the Boxplot for annual spending on Fresh products. We can see from the equal spending and the boxplot that the median is much more centred compared to other box plots. Moreover, there are no outliers in this boxplot.

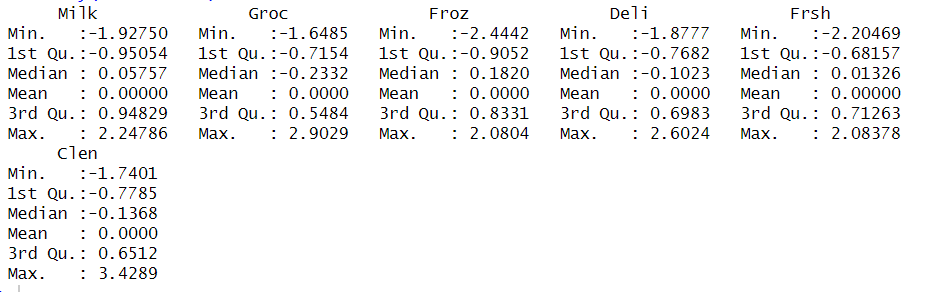
****

This Boxplots shows us the annual spendings on detergents and paper products. There are many visible outliers in this boxplot. As many people tend to buy more or have more family members which can lead to them buying more products compared to others. But the median line is pretty low when compared to the upper bounds and the outliers which means that the average person tends to spend pretty low on these products.



These are the Histogram for different variables in our dataset. We can see that the data is mostly normally distributed. Some might be having a left tail or a right tail or could be a bit skewed but they are distributed normally. But, the Histogram of Milk and Frozen ingredients has two peaks instead of one and hence they are Bimodal Distribution meaning that they have two peaks from two underlying groups. From the frozen group we can see that different groups of people tend to spend drastically differently. This can suggest that one group tends to spend more on frozen products compared to the other. This can also be seen in the Histogram from Milk as well. Here, as there are multiple price ranges for milk with multiple percentage so that could be a way to explain why these two histograms are bimodal.

**3) Standardize all of the variables using either of the two functions demonstrated in class. Describe why you chose the method you did.**

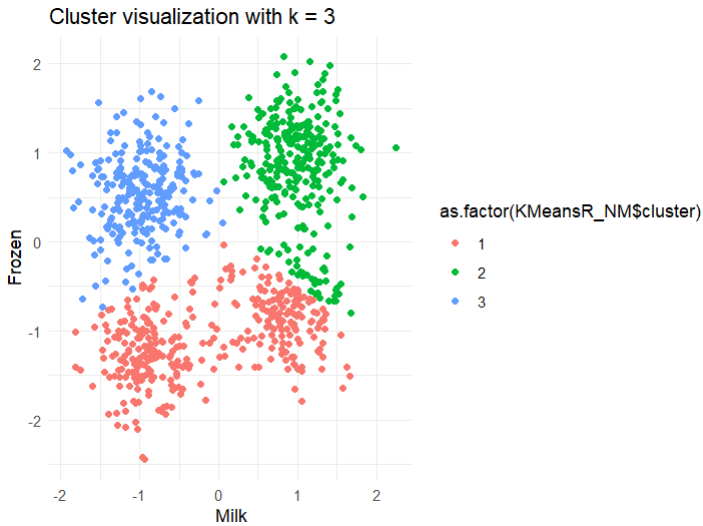
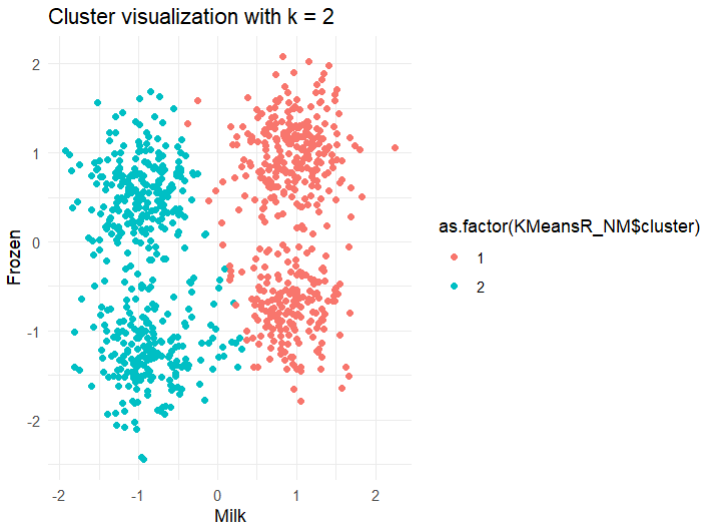
****

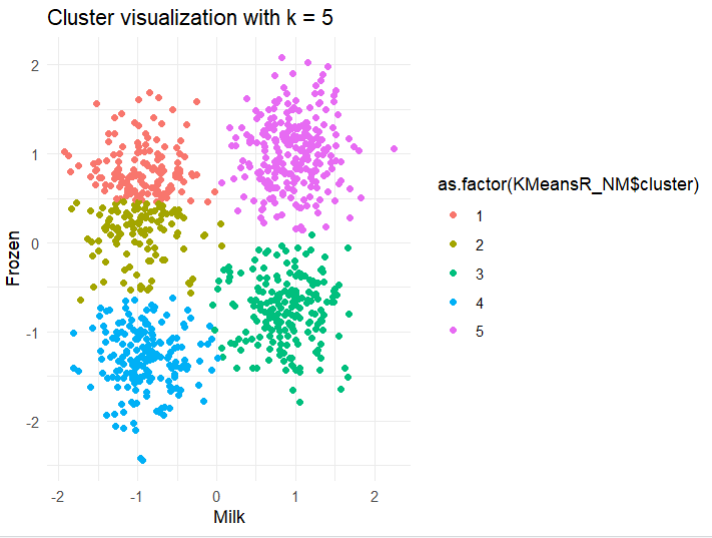
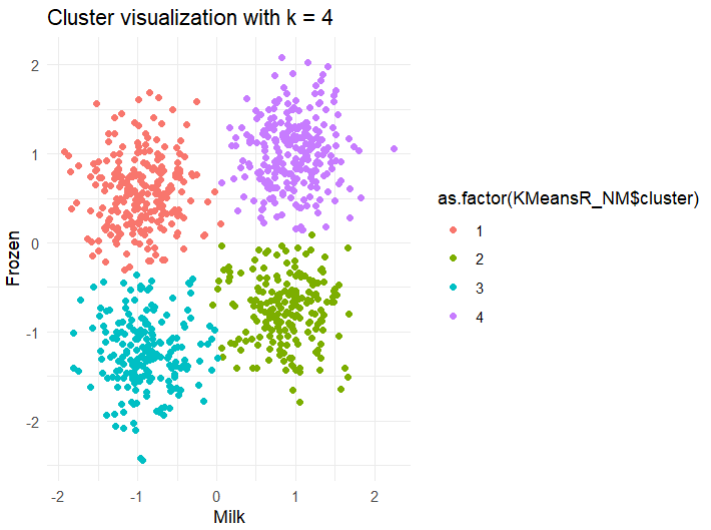
Here I have standardized the data and we can see that by the mean value which has become 0 for all the variables. I have used this by using the scale method. We use this scale method because all of our data is continuous. As the data is numerical, we use the scale function to standardize our data. We standardize the data so that the data doesn’t have any biases and that we can handle outliers in a better manner. Moreover, standardizing the data makes it easier to process and analyse the data.

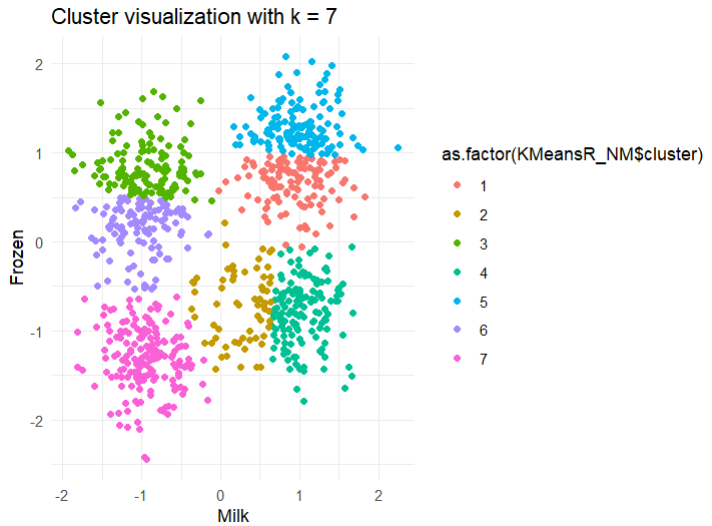
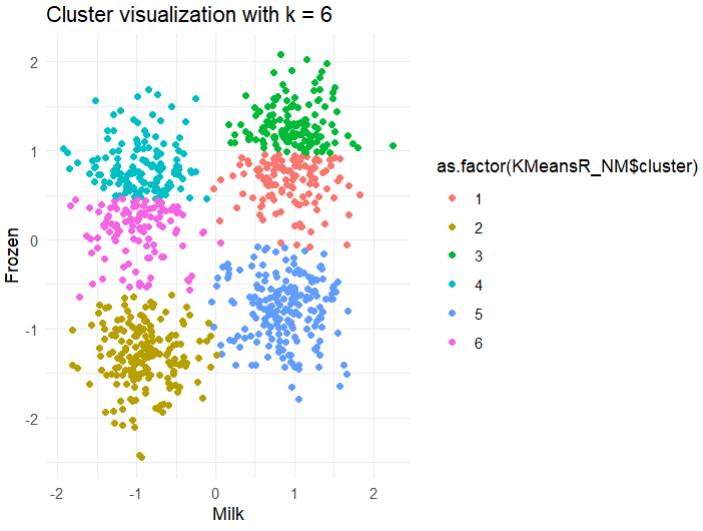
**2. Clustering**

**Using the K-Means procedure as demonstrated in class, create clusters with k=2,3,4,5,6,7. You will be using only two variables as your centroids (Milk and Froz).**

1. **Create segmentation/cluster schemes for k=2,3,4,5,6,7.**



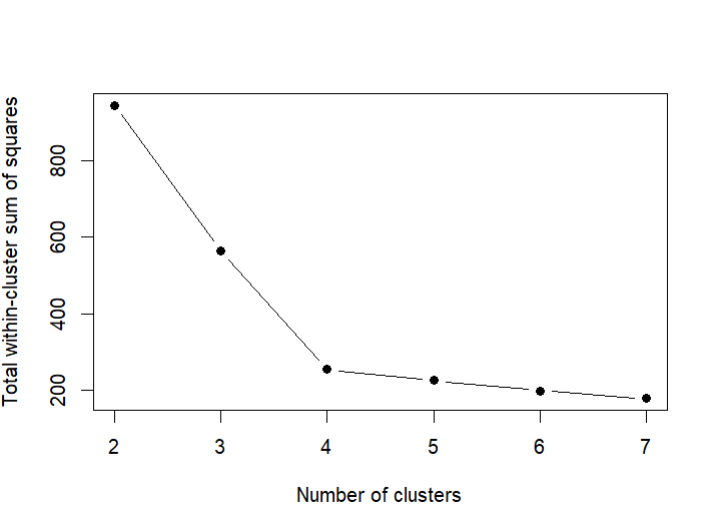




Clusters for k=2,3,4,5,6,7

We can clearly make it out overfitting and underfitting from these diagrams Let’s take k=7 for example. Here we can see that the algorithm takes particular shapes and everything to properly fit the data according to the model and its training. Which is not good for evaluation and hints at overfitting. Similarly on the other end of the spectrum, when we take k=2 or k=3 we can see underfitting taking place. Both of these are not good for the model.

**3) Create the WSS plots as demonstrated in class and select a suitable k value based on the “elbow”.**

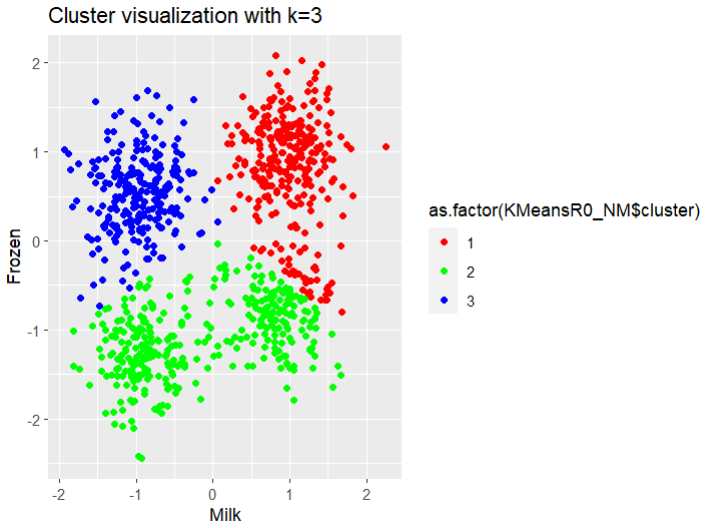
****

Based on this WSS plot we can clearly see at cluster value k=4 there is a sharp change in the graph and the graph changes direction drastically. Hence, based on the explanation of Elbow method we can select the optimal number of clusters as 4.

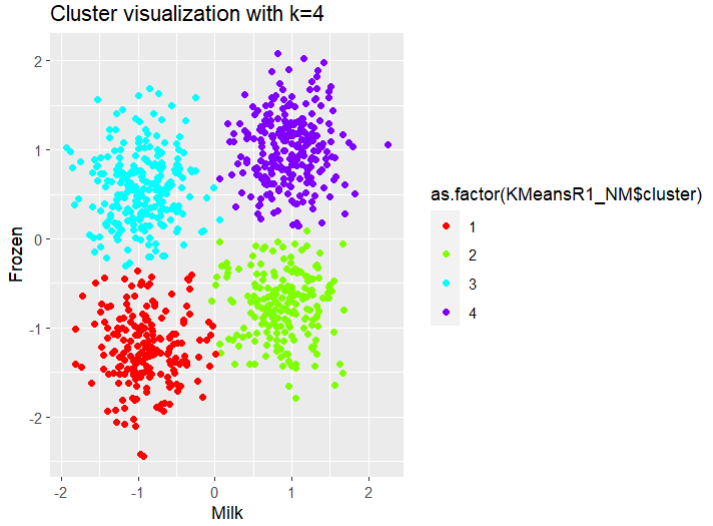
**3. Evaluation of Clusters**

**1) Based on the “k” chosen above, create a scatter plot showing the clusters and colour-coded datapoints for each of “k-1”, “k”, “k+1”. For example, if you think the “elbow” is at k=5 create the charts for k=4, k=5 and k=6**

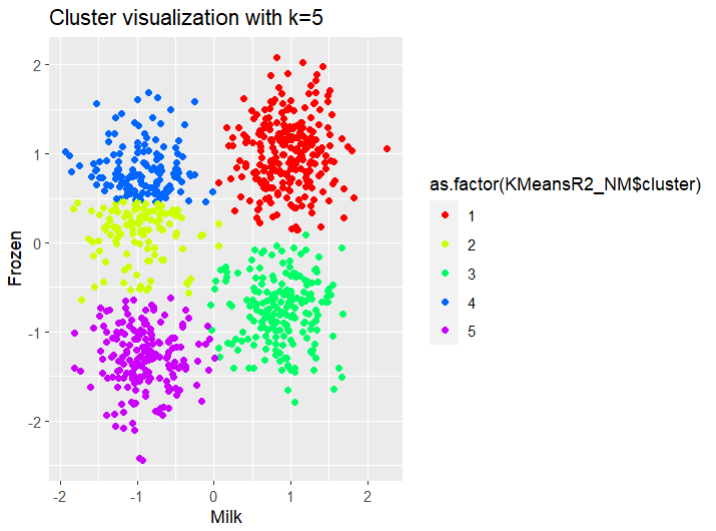
With the help of elbow method and the inflexion point I have decided to go with k=4 as our optimal number of clusters.  
Here is k-means clustering for k=3,4 and 5 as asked in the question.



This is clustering for k=3 we can see clear differences in the data but they are not properly separated because we have chosen the k-value as 3. This leads to underfitting of the data by the model.



This is the k-means clustering for k=4 i.e. 4 clusters. Here as per the elbow method this is the optimal number of clusters for this dataset.

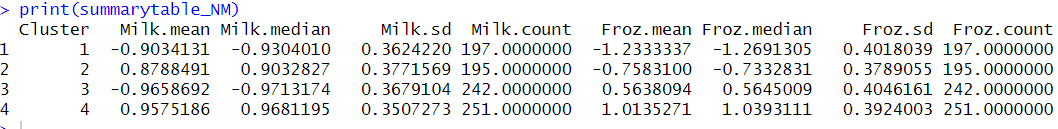


This is the k-means clustering for the k-value of 5. Here we can see that the yellow cluster is forcibly making a section from the blue cluster which is not optimal.

**2) Based on the WSS plot (3.2) and the charts (4.1) choose one set of clusters that best describes the data.**

Based on the Weighted Sum of Squares plot against the number of clusters and the plot of k=4 cluster we can confidently choose k=4 as our optimal amount of cluster between Frozen products and Milk products. Elbow method suggests us the sharpest angle between WSS and the number of clusters should be our optimal number. This leads us to select the number of clusters as 4. Moreover, from the plot of k=4 we can see that the optimal amount of fit is present as there are some errors here and there and the model is not overfitted or underfitted to the extremes and it shows no biases. With the help of the Visual method, we can choose k=4 as there are 4 distinct clusters. These are the reasons why I have decided to go with k=4 as the optimal number for our clusters which describes us the best fit for our data.

**3) Create summary tables for the segmentation/clustering scheme.**



This is our summary table for k=4 number of clusters.

**4) Create suitable descriptive names for each cluster.**

Based on the summary of our cluster we can infer many details and ideas. Some descriptions for those said clusters that jump to mind are these:

**Cluster 1** it is evident that people are buying less amounts of milk and frozen products as the mean and median for both of these products is lower than other clusters. Hence, this cluster can be named as **Low Purchasers for Frozen and Milk Products.**

**Cluster 2** gives the idea of high amount of milk product purchases. Although, the milk product sales have increased the frozen products remains lower than usual. Therefore, this cluster can be described as **High Milk and Low Frozen Products Purchasers.**

In **Cluster 3** we can infer that the Milk product purchasers are lower then the previous 2 clusters. On the flip side, the average mean and median of frozen product buyers is moderately higher. This cluster can be surmised as **Low Milk Product Purchasers and Moderate Frozen Product Purchasers.**

Moving on to **Cluster 4** we can clearly make sense how the average mean and median for both the milk products and the frozen products is much higher than the rest of the clusters. This cluster clearly gives the idea of **High Milk and Frozen Product Purchases.**

**5) Suggest possible uses for this clustering scheme**

Based on the table of summary and the clusters we can make use of this data to leverage better sales and make use of targeted advertisement to the customers. Like for example, we can advertise more costly frozen products to customers in cluster number 4 as they are the group that spend the most in the frozen products category and they can be enticed to splurge more money on more costly options. Similarly, we can advertise sales and buy one get one offers to customers in cluster number 3. Since they are buying a moderate number of frozen products, we can push them to spend a bit more on frozen products by giving them sales and offers. Moreover, we can advertise frozen products that go together well with milk-based products to customers in cluster number two to boost more frozen product sales.

We can also stock products based on these results as people in cluster 4 tend to spend more on the milk and frozen products so we can stock more. Moreover, we can use to segment customers which can lead to higher satisfaction if we provide them with offers and cater to their needs based on their spending habits and their history. We can also use this information to advertise top selling brands and products which are frequently bought together by other customer to boost more sales. Furthermore, we can offer more discounts to people in cluster where milk and frozen sales are low to incentivize them to actually buy the products.

If we take into consideration of other external factors then we can also use these clusters to our advantage. For example, lets take customer from cluster 3 who buys lesser number of milk-based products then frozen products because of some external factors. Then we can solely cater to them with frozen products to suit them better. These are some of the uses that come to mind where we can make use of k-means clustering. Some more uses from the top of my head can be Customer Retention Strategies, Increasing customer values, etc.