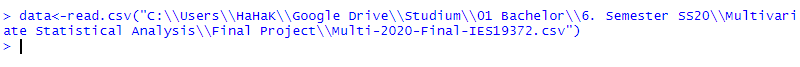
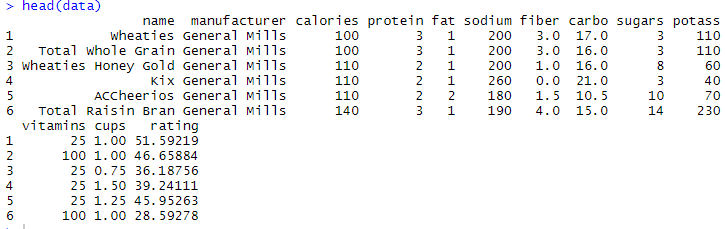
Final Project  
Benedikt Willecke  
IES19372

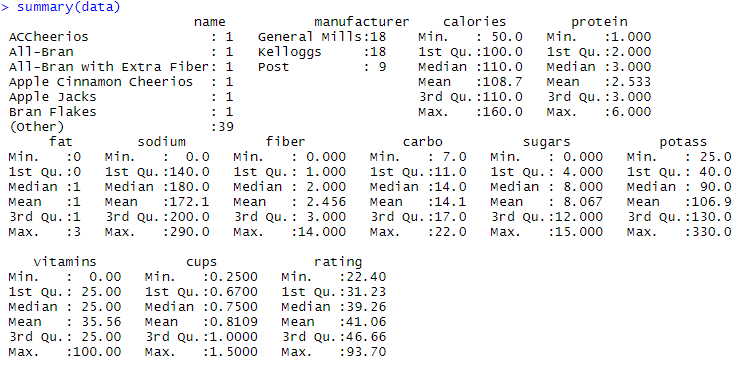
1. Exploratory Data Analysis

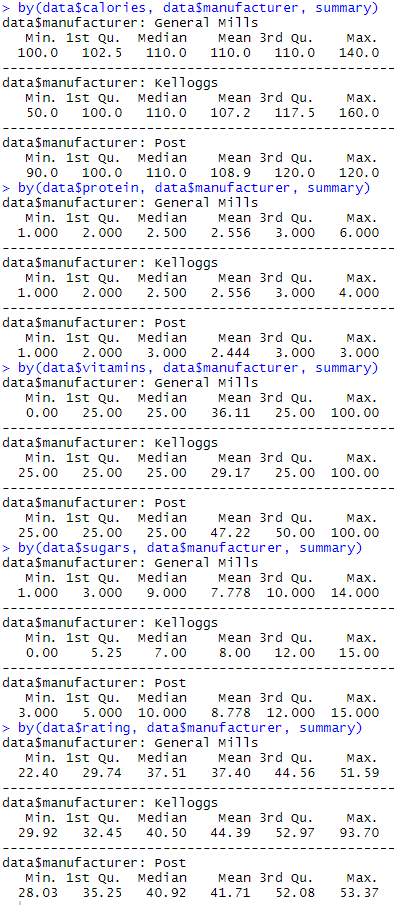
The data includes metrics of different American cereal brands. First I load the data.  
  
Next let’s take a general look at the data. We can see we have the cereal with the respective manufacturer. Furthermore, we have metrics such as calories as well as a rating.  


In total we have 45 datapoints with 13 variables.

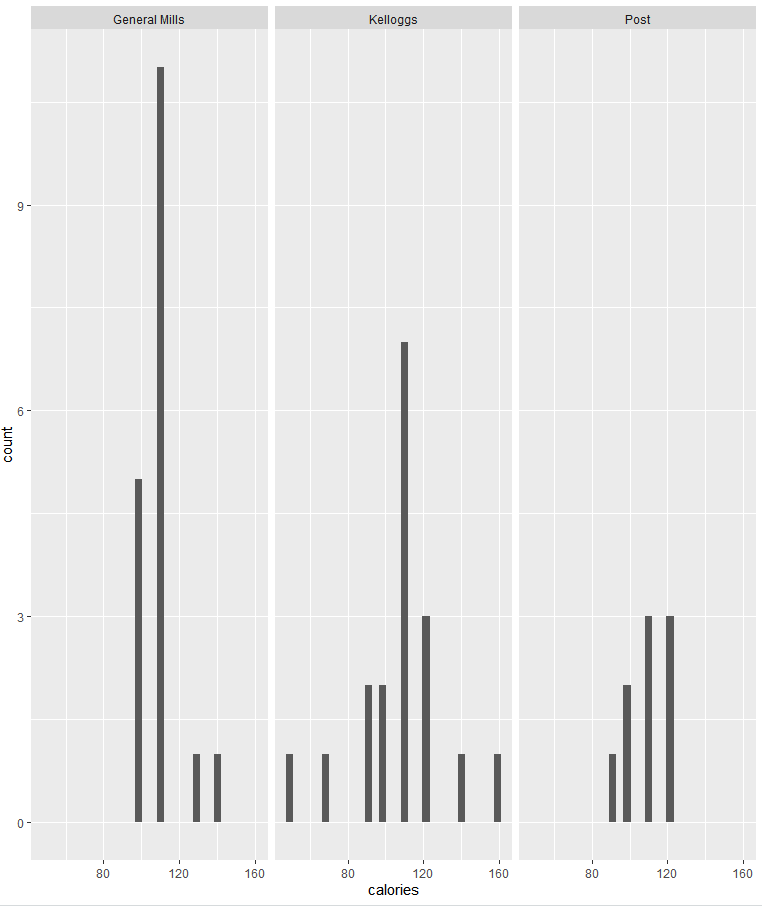


We can see that one datapoints equals one product from one of three different manufacturers. I also would like to point out that all the metrics vary quite a lot. As an example if we look at calories the mean/median are both around 110. But it varies from 50-160. This might be due to different sizes. The rating also varies a lot from 22.4 (bad) to 93.7 (good). On average (41.06) the rating is below the midpoint of 50.0.

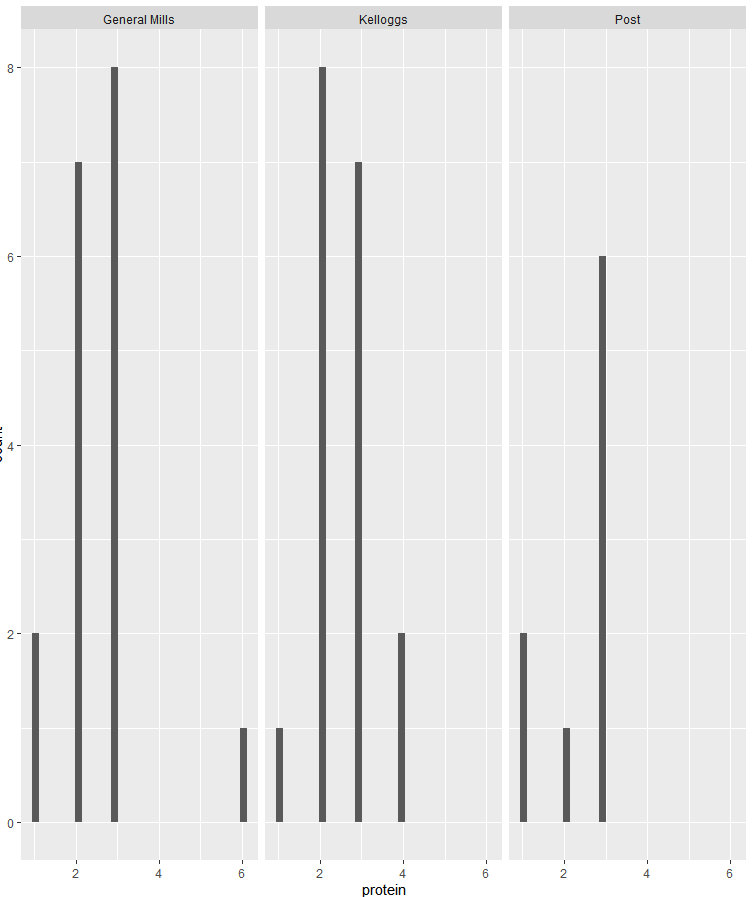


As an example, we can also see that there are differences between the manufacturers. Although the mean/median are quite similar the minimum & maximum values vary. Similar observations can be made about protein, vitamins and sugars. The biggest different is in the ratings though. All values are almost the same except the maximum rating in Kelloggs. Kelloggs seems to have maybe only one product that stands out of the crowd.  


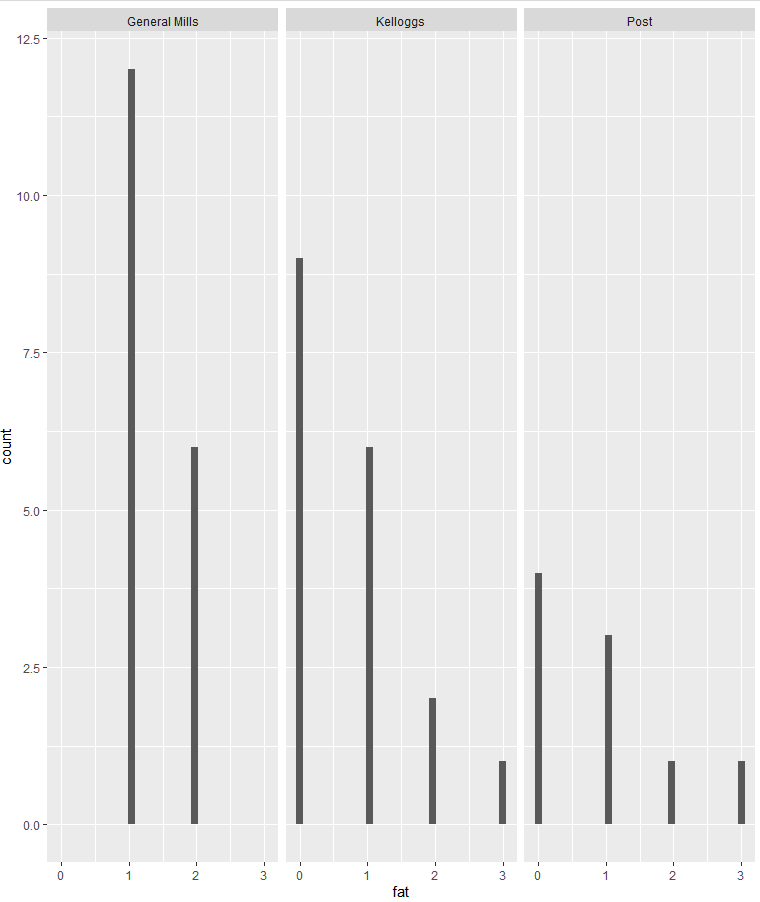
By looking at histograms of the variables by manufacturer we can see how they are distributed in a different way and we can see the differences in the manufacturers. We can see that the calories are very differently distributed. Especially Kellogs is very broadly distributed whereas Post covers only a small portion of the calories. General Mills uses the same calorie amount very often. As for the protein, considering we only have a limited amount of datapoints they seem fairly similarly distributed. When we look at the fat amount General mills has about the same frequency of 1&2 as Kellogs. And Post even has a few zero fat cereal. The sodium distribution is similar like the calorie distribution. In terms of fiber all manufacturers generally have low fiber cereal. Only Kelloggs offers a few high-fiber ones. As for carbo all manufacturers are similar, just Kelloggs has a few outliers on the high end. As for sugar all manufacturers are broadly distributed across the board with higher frequencies at around 5-10 range. General Mills again seems to have standardized amounts and thus high frequency at specific values. Potass I broadly distributed with General Mills products being a little bit on the low end side. As for vitamins it tends to have a lot of products at either 25 or 100. Kelloggs seems to have low and mid cup sizes, Post is very broadly distributed and Kelloggs seems to have mid to high cup sizes. As for the ratings all ratings seems to be equally distributed between 20-60 with Kellogs having a stark outlier at around 90.  

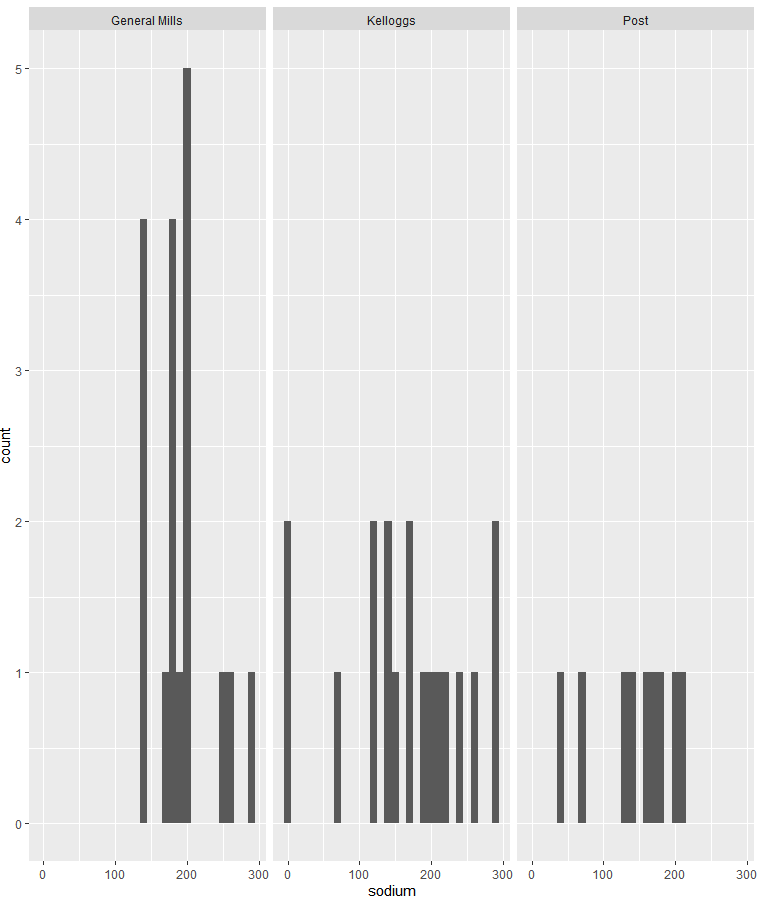





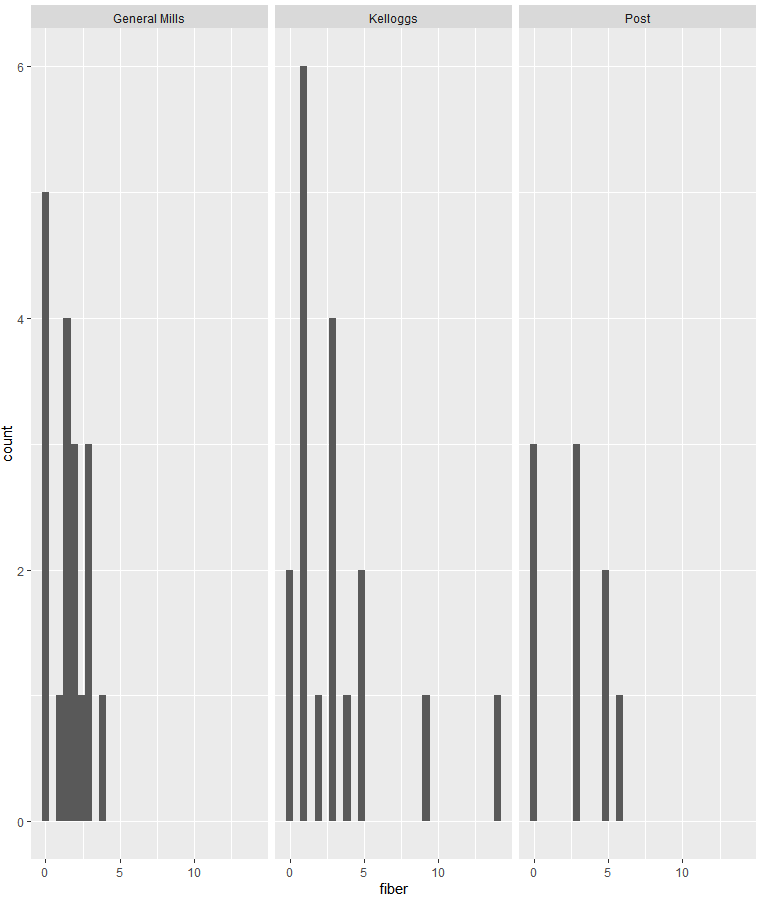


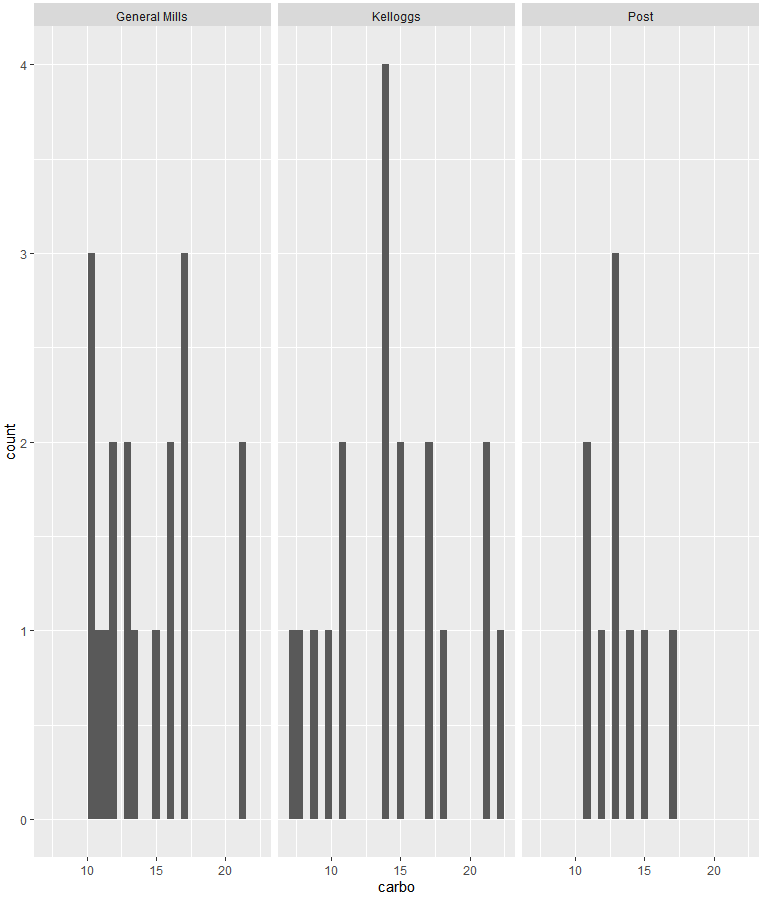


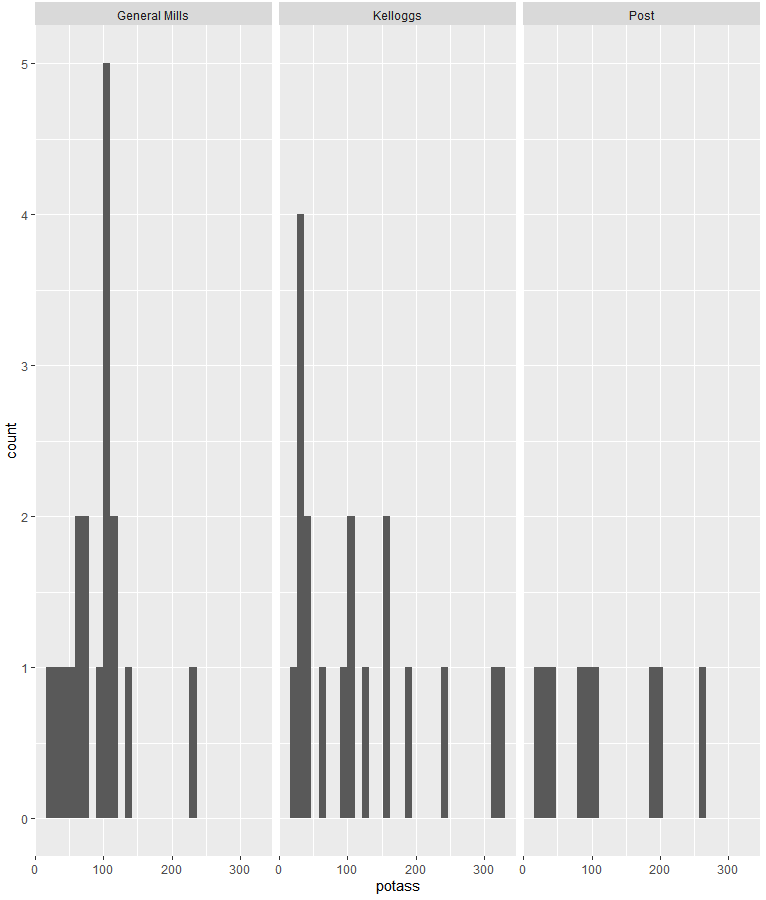
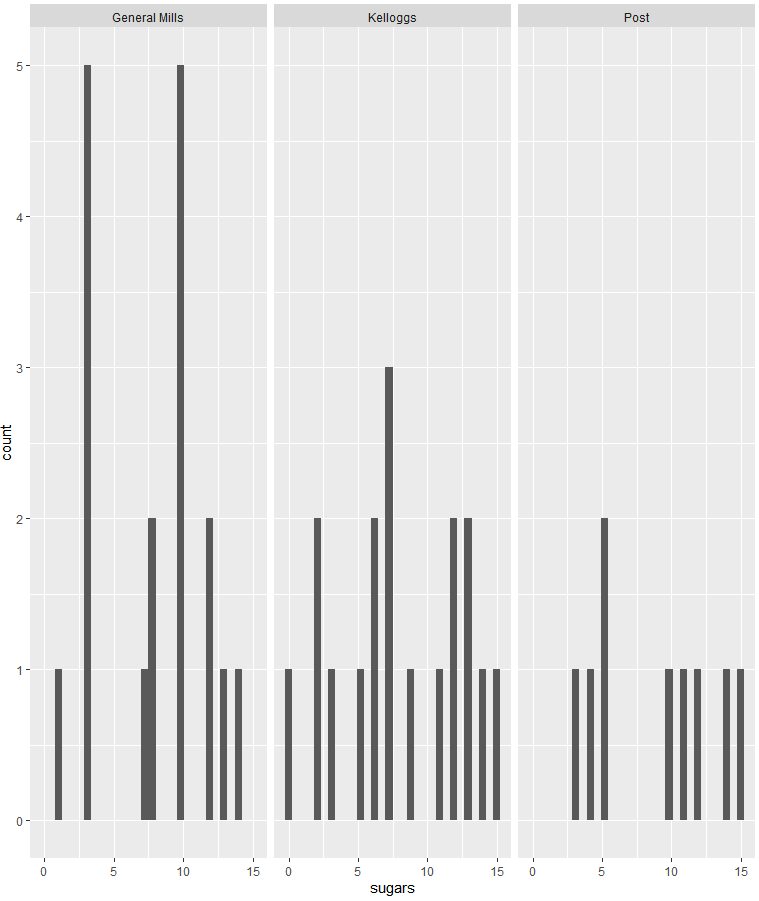




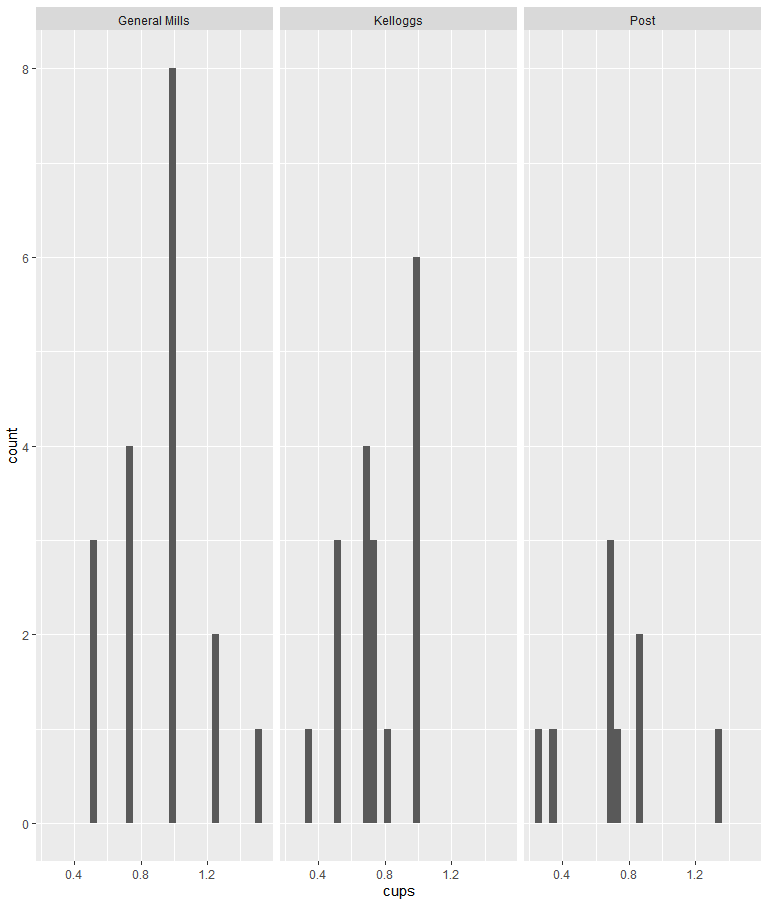
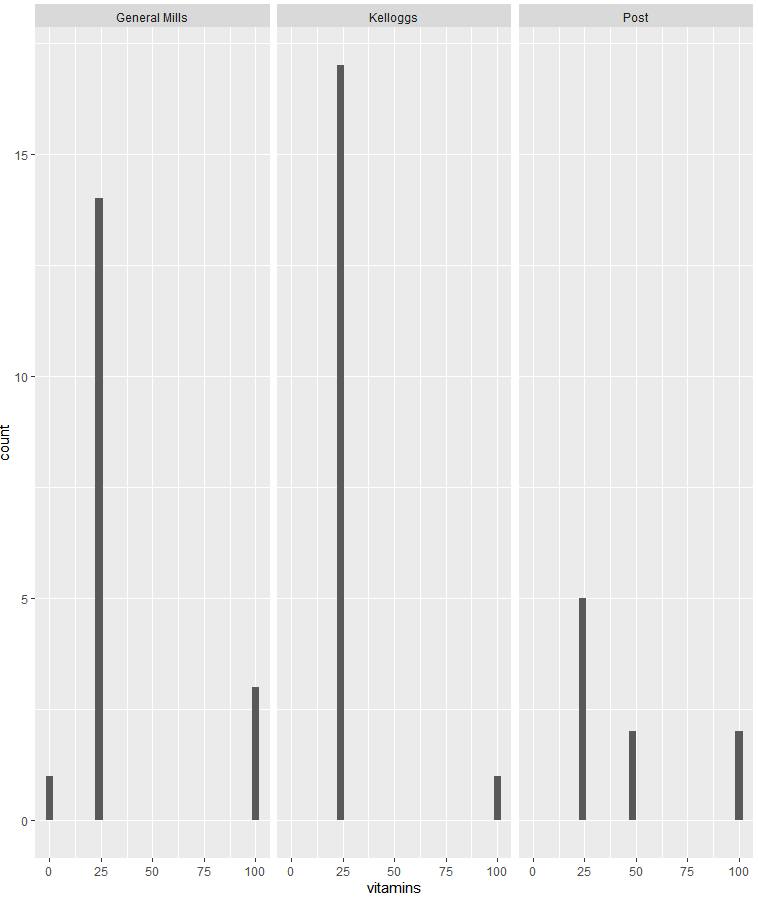


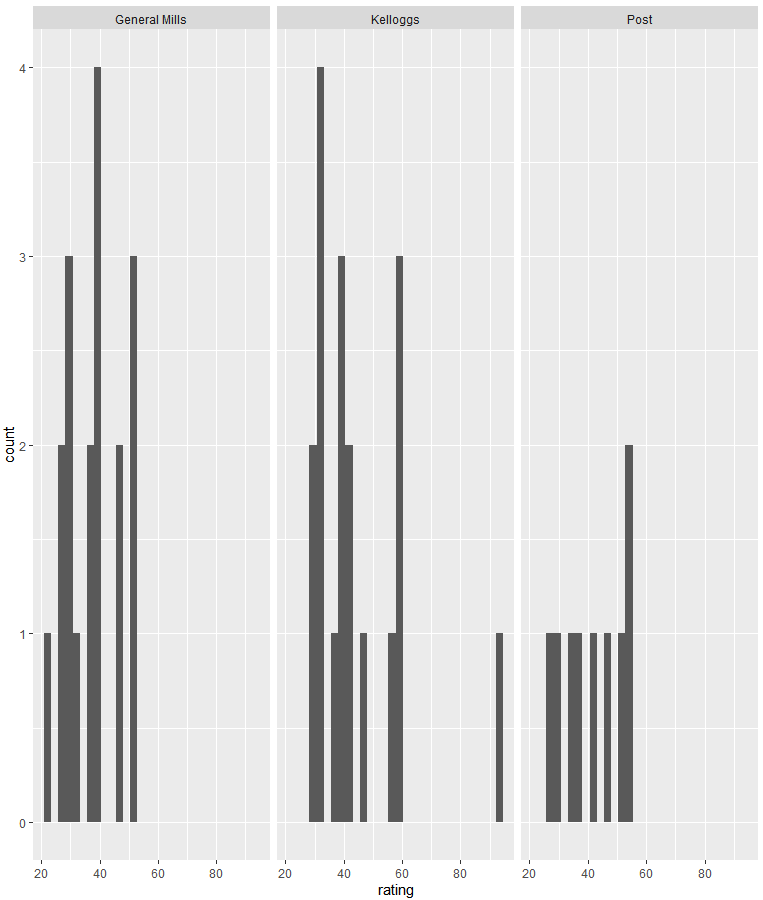






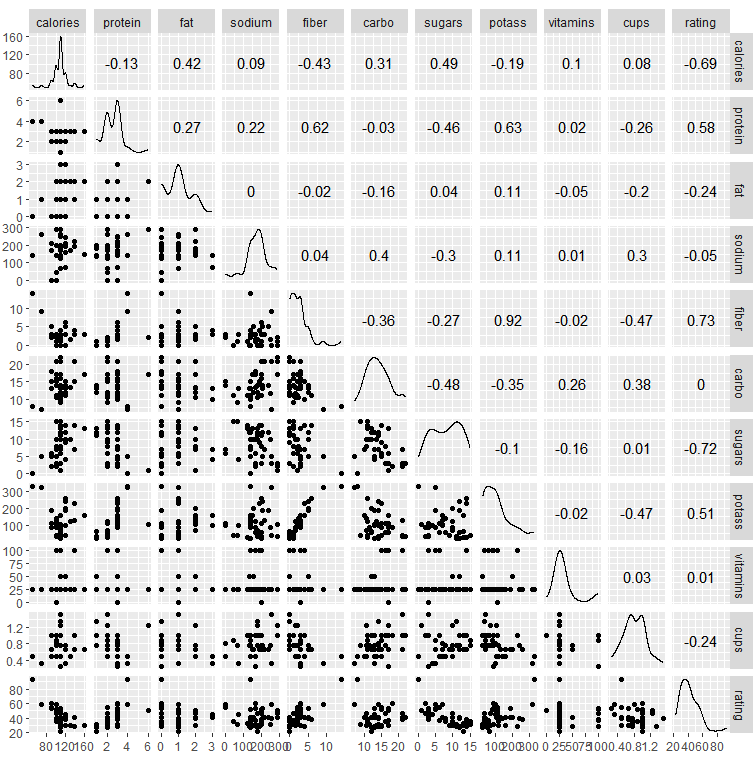




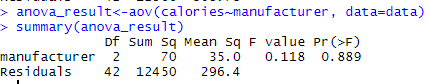


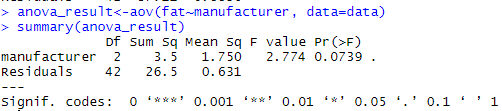
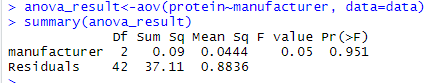
Next we can look at the scatter matrix. We can make several observations: fat&calories, fiber&protein, potass&protein and rating and fiber seem to be positively correlated with potass&fiber being very strongly correlated. Fiber&calories, rating&calories, sugars&protein, cups&protein, rating&fat, carbo&fiber, sugars&fiber, sugars&carbo, potass&carbo, rating&sugars, cups&potass, cating&cups seem to be negatively correlated. From this we can conclude that high rated products seem to have low calories, high protein, low fat, high fiber, low sugar, high potass and low cups. This could mean that the rating reflects the healthiness of the product.

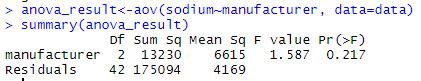


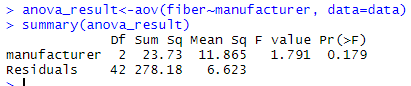


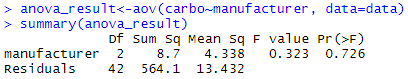
2. Test whether the mean of each variables between the manufacturers are the same or not.

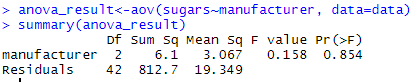
Here I test whether the mean of each variable is different between the manufacturers. With ANOVA I can deduct that most means seem to be the same. Only for fat and cups we cannot reject H0 with a 10% significance level. Using MANOVA it seems that the means of all variables put together seem to not be different. (From my understanding MANOVA was not asked for, I just wanted to include it anyways out of my own curiosity).  


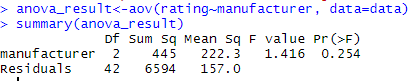
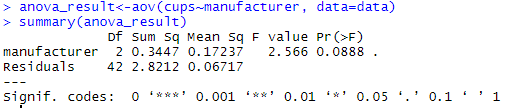
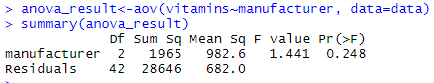
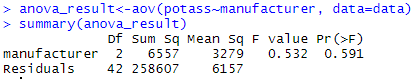


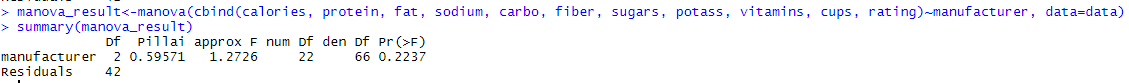






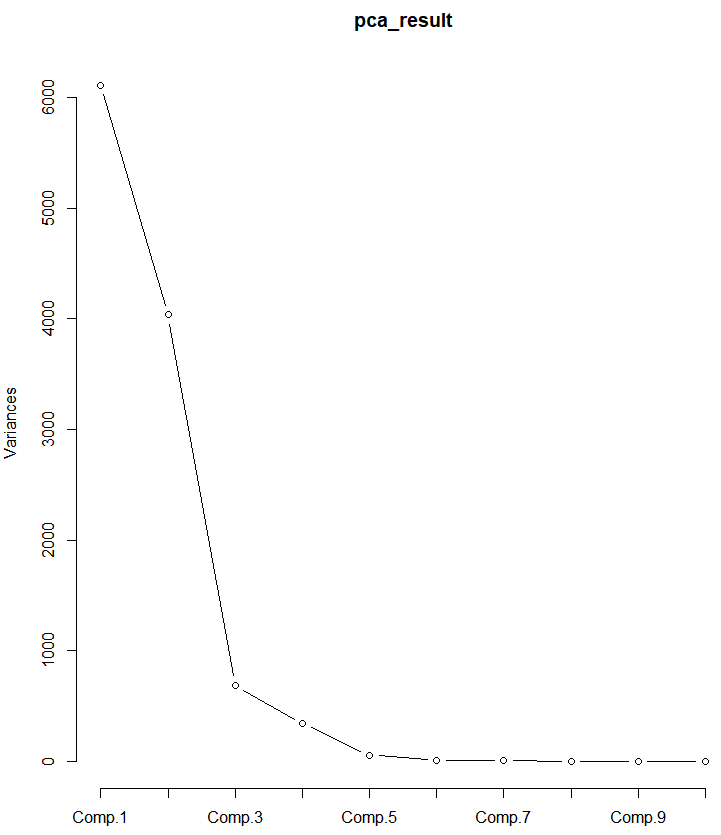
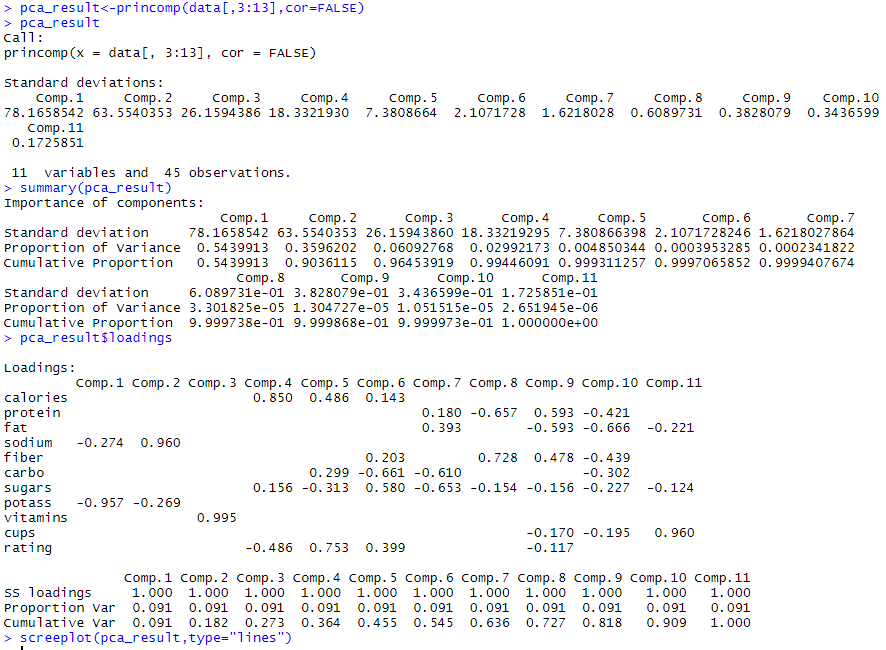






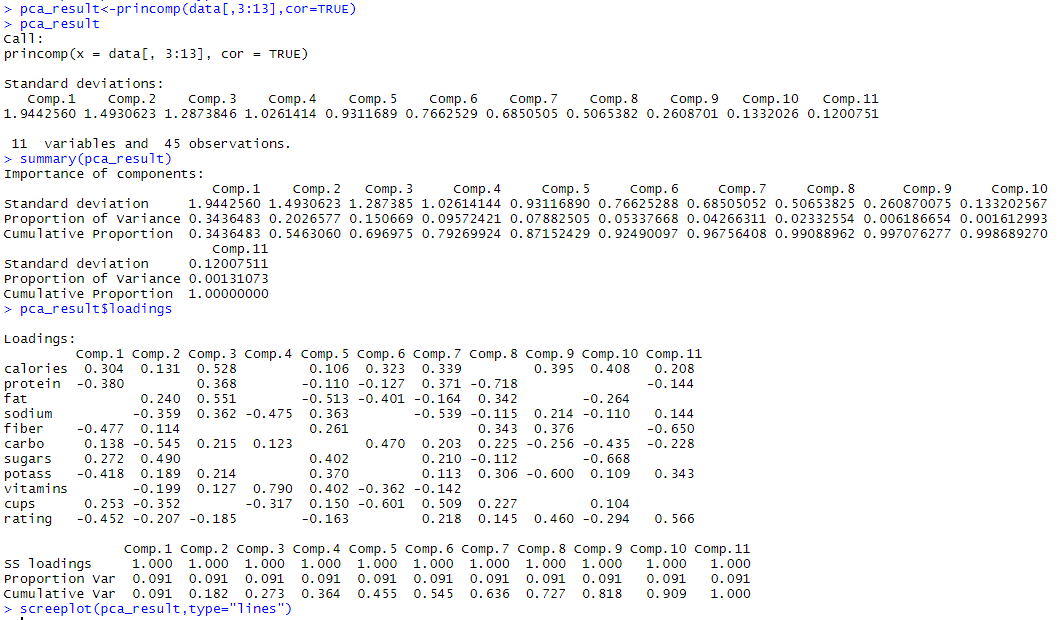
3. PCA

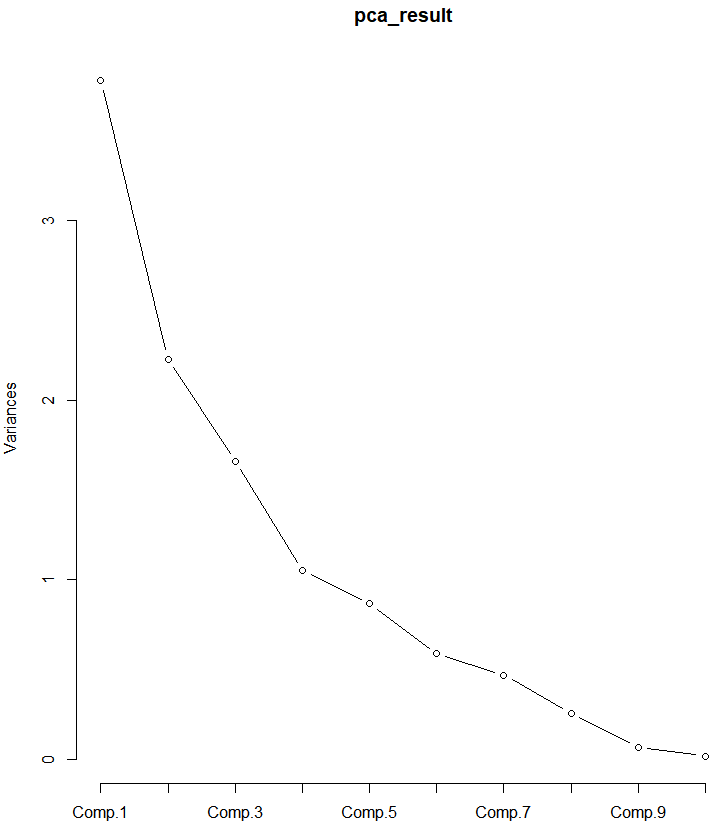
a) Here we can see that the first two components describe about 90% of the data. This provides enough accuracy to describe the data. For 99% precision 4 components could be used but 90% is already so high that m=2 should be chosen.



b)

Using the correlation matrix we can see that in order to get the previous 90% accuracy we would need 6 components here. Instead, in this case, we can use m=4 components to get roughly 80% accuracy. This is still higher than the standard minimum of 70%.

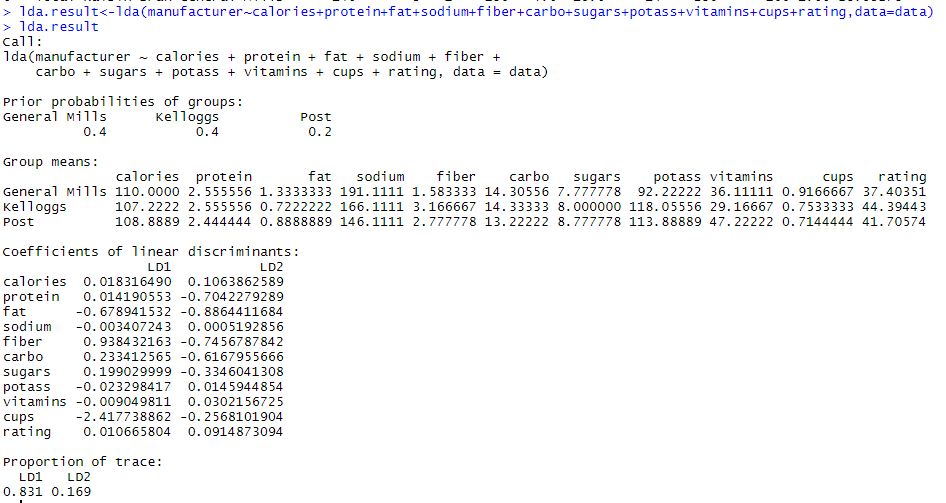


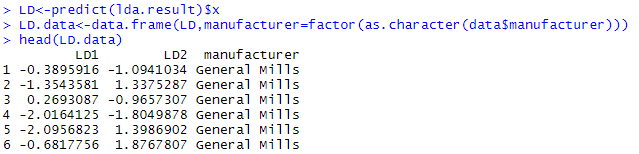


c)  
In conclusion we can say that using the variance-covariance matrix we can get more accuracy (90% vs 80%) with less components (2 instead of 4). We can see at the loadings that when we use the correlation matrix that a lot of variables are considered. Because of the usage of negative and positive values we can see that this model looks at the difference between the variables. On the other hand when using the variance- covariance matrix it simplifies the model a lot by only using sodium and potass. Although very simple (2 vs 11 variables) it explains the data much better.

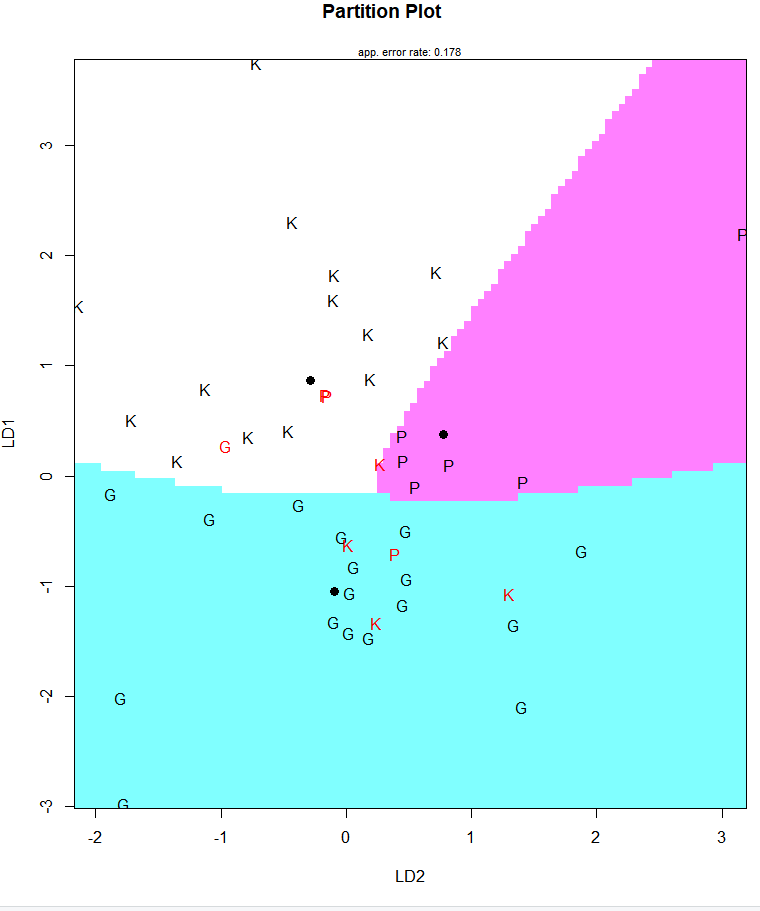
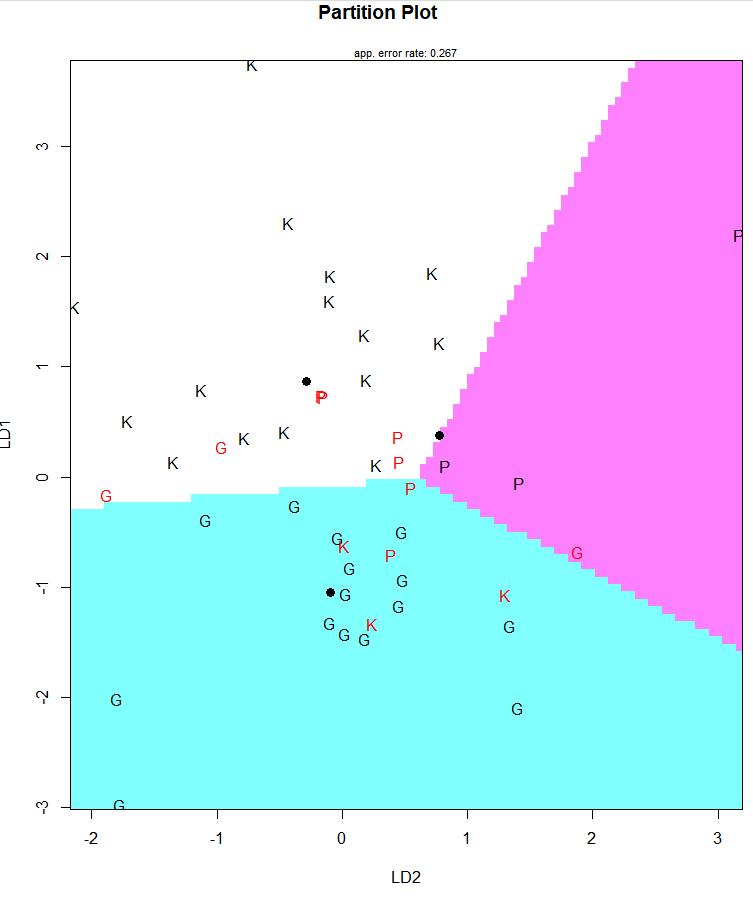
4.Discriminant Analysis

a) & b)

With LDA we can find the linear discriminants as described below.   


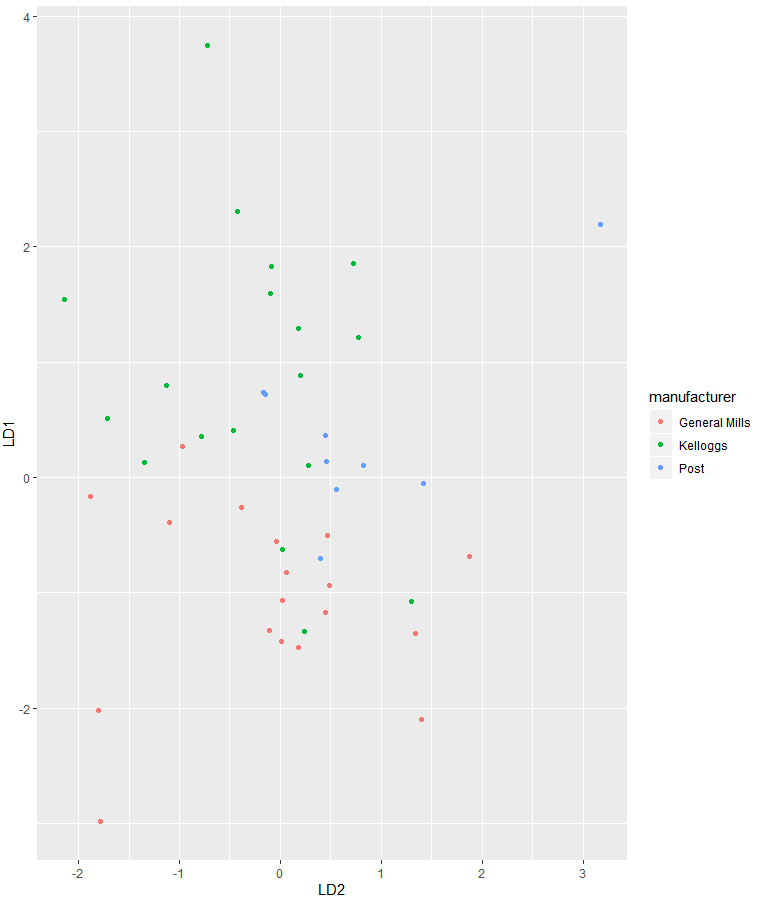
Next we can make a prediction.  
  
And we can plot both LDA and QDA so we can see the decision rules. We can divide the data at about LDA1=0. Doing so gives us almost only Kelloggs at the top and General Mills at the bottom. In order to predict Post as well we need to include the top right part of the plot. To achieve this QDA gives slightly better result.





When looking at a different plot as well which shows the different manufacturers in color. Because Post overlaps a lot with the other two manufacturers, which seem pretty distinct, I would guess Post is harder to predict than the others.

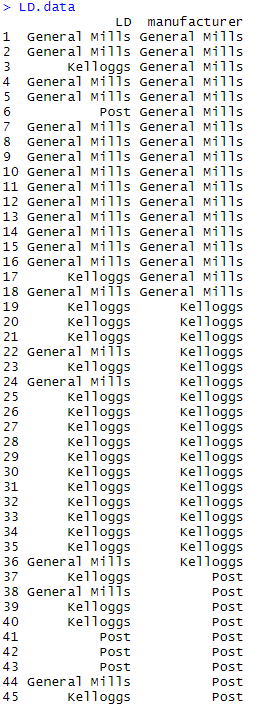




c)

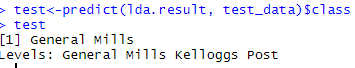
I can confirm this when I try to predict the manufacturer and compare the truth vs the prediction. We can see that General Mills got misclassified 3 times, Kelloggs 3 times and Post 6 times which is double the amount at much less datapoints.





d)

My prediction is that this is General Mills.

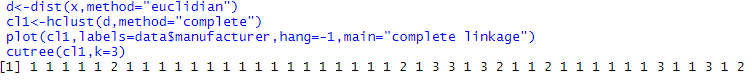


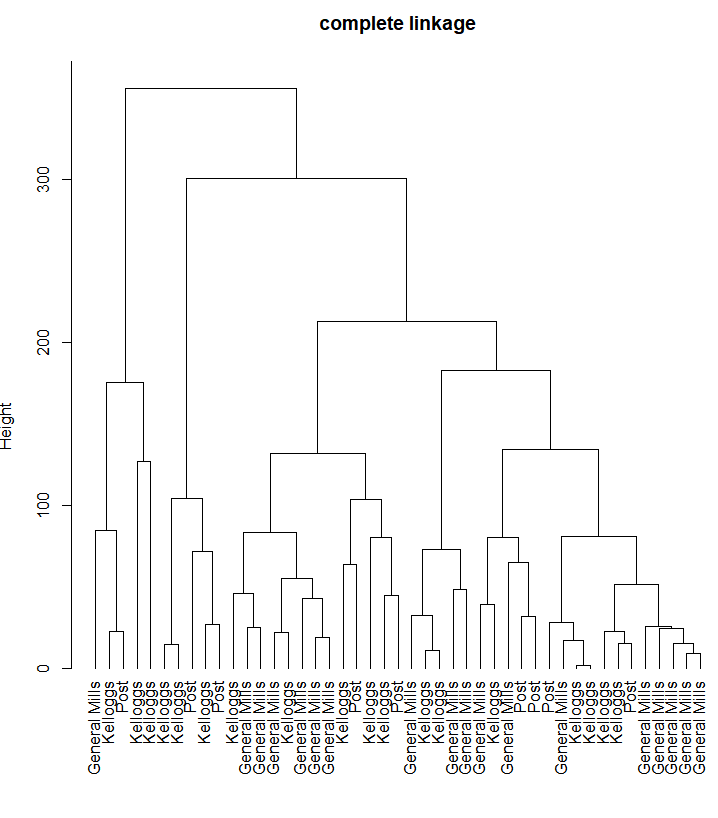
5. Clustering

a)

Using complete linkage with Euclidian distance we can get the following dendogram with the following cluster.







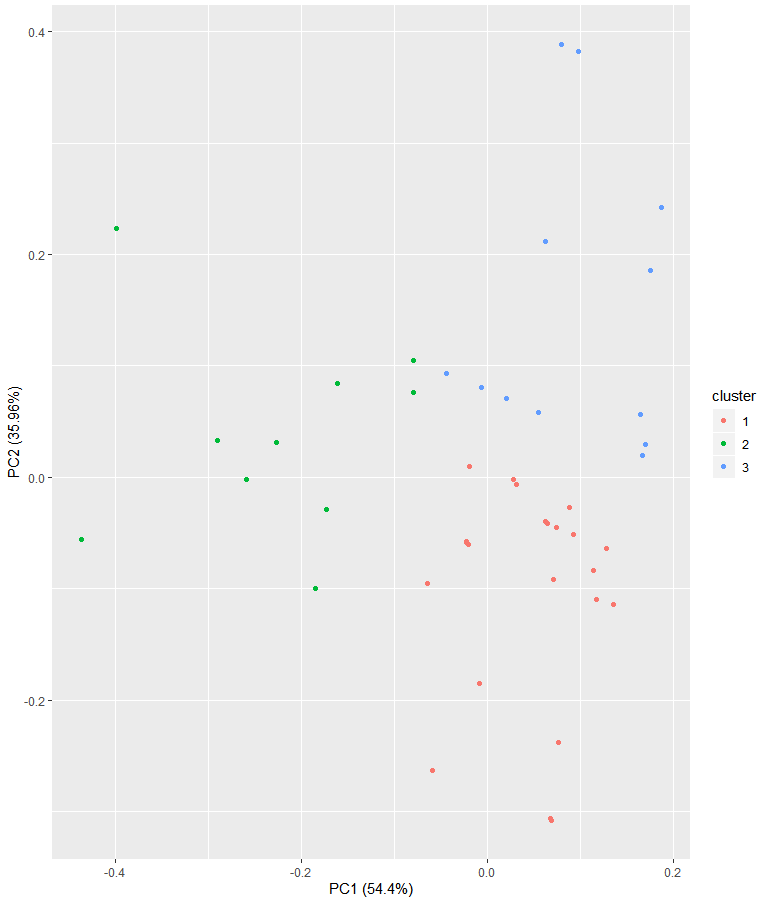
b)

When applying kmeans method we get the following clusters.









c)

When comparing the two results we can see that both clustering methods give similar results. Only 7 datapoints are clustered differently. With complete linkage having a strong bias towards group 1. Note that group 1 in complete linkage equals group 2 in k-means and vice versa.

