**Cereal Data Analysis**

**Prema Krishna Gadde (z1834025@students.niu.edu)**

**Northern Illinois University**

**OMIS 665**

**Professor: Ying Wang**

**December 5, 2018**

**INTRODUCTION**

The data is taken from the cereal section found in a grocery store, this dataset consists name of each cereal, location of the cereal box which could be on Top, middle or bottom shelf aisle section of the store. This dataset also has different types of cereals along with their nutritional breakdowns which include: Calories, Fat, Carbohydrates, Fiber, Sugar and Protein. In the dataset manufacture, type and the shelf information of the cereal are of coded categorical or the nominal type, while the data related to the vitamins content in each cereal are in numerical or ration type.

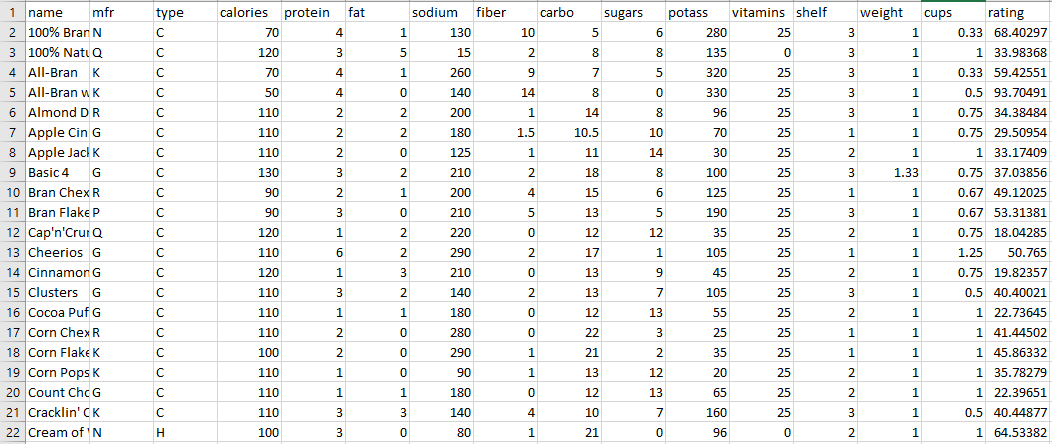
The reason why we chose this dataset is, most of the Americans start off their day with a bowl of cereal and usually the customers in a grocery store spend an ample amount of time at the cereal aisle to determine which cereal is healthy. Also having a decent amount of sample data which is 77 cereal types, made us think it would be a legitimate source to analyze the data to determine which cereals are healthy to consume at the start of the day. It’s just not the kids who enjoy cereal but also the adults enjoy it quite equally and it is important to know what we are eating.

Our goal here is to analyze which company is manufacturing the best cereals and what are the similarities among the various brands. Apart from this, we would be implementing different types of Analysis methods to determine the what affects the ratings provided by the customers to the manufacturers. We would also determine the factors that affect the calorie content in a cereal. We are also determining the brands which have similarities with respect to the calories in the cereals and ratings given to the cereals. The dataset present is taken from Kaggle website.

**VARIABLE DESCRIPTION**

|  |  |
| --- | --- |
| Field | Description |
| Name | Name of the Cereal |
| Mfr | Manufacturer of the Cereal |
| A | American Home Food Products |
| G | General Mills |
| K | Kellogg’s |
| N | Nabisco |
| P | Post |
| Q | Quaker |
| R | Ralston Purina |
| Type | Cold, Hot |
| Calories | Calories per Serving |
| Protein | Grams of Protein |
| Fat | Grams of Fat |
| Sodium | Milligrams of Sodium |
| Fiber | Grams of dietary fiber |
| Carbo | Grams of complex carbohydrates |
| Sugars | Grams of Sugars |
| Shelf | Display Shelf |
| Weight | Weight in Ounces per serving |
| Cups | Number of cups per serving |
| Ratings | Rating of the Cereals |

**DATASET OVERVIEW**



**ANALYSIS METHODS AND TOOLS**

|  |  |
| --- | --- |
| K-Means Clustering | R Studio |
| Hierarchical Clustering | R Studio |
| Linear Regression | SAS Studio, R Studio |

**EXPLANATORY ANALYSIS**

**Distribution of manufacturers**

**R code:**

install.packages("plotrix")

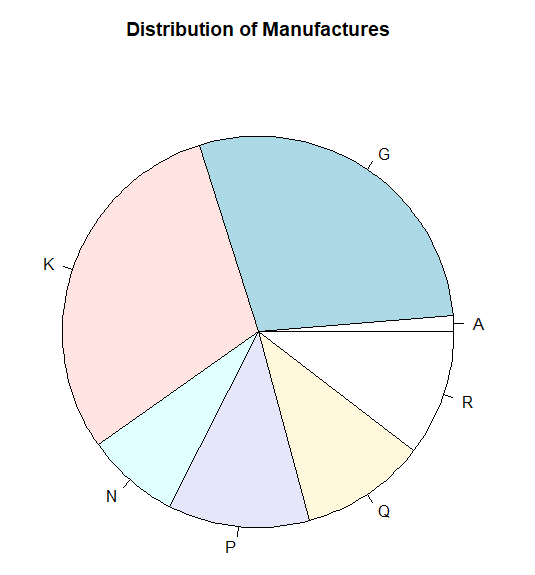
library(plotrix)

cereal<-read.csv ("cereal.csv")

mfr=cereal$mfr

mfr.freq = table(mfr)

pie(mfr.freq,main="Distribution of Manufactures", cex = 1)



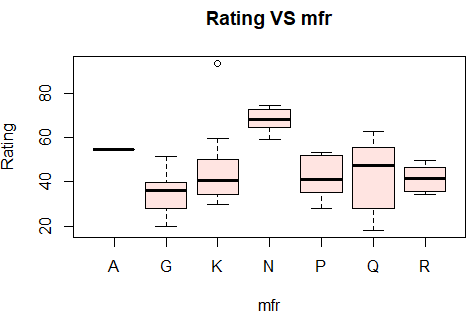
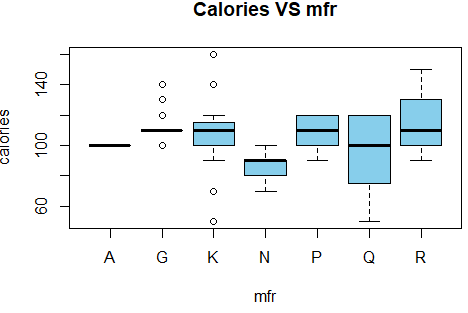
From the above graph we can say that Kellogg’s manufactures a greater number of cereals and it is followed by General mills, American home food have least number of cereals in the market.

**Rating vs calories for manufactures**

**R code:**

boxplot(cereal$rating ~ cereal$mfr,xlab="mfr",ylab="Rating",main="Rating VS mfr",col="mistyrose")

boxplot(cereal$calories ~ cereal$mfr,xlab="mfr",ylab="calories",main="Calories VS mfr",col="skyblue")

From the above figures we can say that if the ratings are higher for the ones that have less calories, Nabisco company manufactures cereals with least number of calories and has the highest rating.

**Manufacturers vs type**

**R code:**

install.packages("ggplot2")

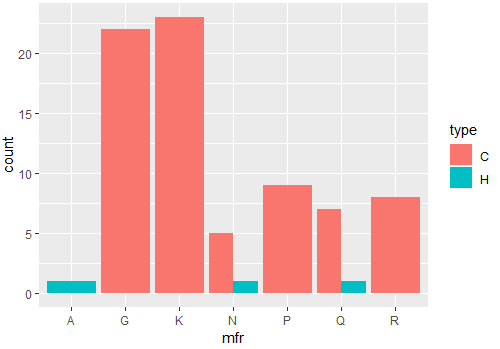
library(ggplot2)

mfr <- cereal$mfr

type <- cereal$type

df <- data.frame(mfr,type)

ggplot(df, aes(mfr, ..count..)) + geom\_bar(aes(fill = type), position = "dodge")



From the above figure we can see that very few companies manufacture hot cereals, business must focus on coming up with new variety of hot cereal if they want to be unique.

**Correlation Matrix**

A correlation matrix can help us determine the degree of correlation between the variables. It also shows the direction of relationship between two variables. Correlation is between -1 to +1 and if the value is 0 then there is no correlation between the variables.

**R code:**

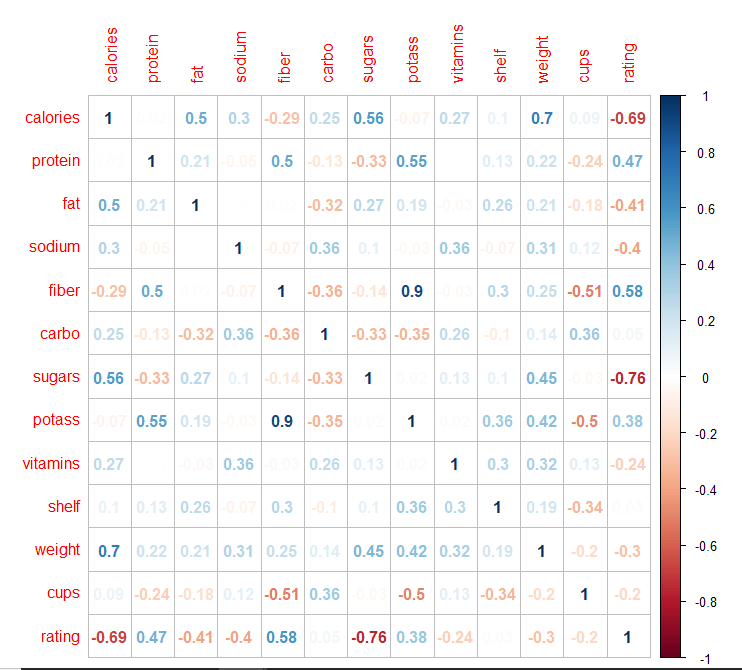
install.packages("corrplot")

library(corrplot)

cerealcor<-read.csv ("cerealcorelation.csv")

M <- cor(cerealcor)

corrplot(M, method = "circle")



**Interpretation:**

* From the above figure let’s look at some significant correlations, we can say the there is a strong negative correlation between rating and sugars, rating and calories.
* A strong positive correlation between potassium and fiber, rating and fiber.
* As per our research, we can say that a cup and weight ratio may mean different things to different companies.

**LINEAR REGRESSION**

Linear regression is a supervised leaning method used to predict the relationship between the dependent variable and the independent variable and it also tell us how the independent variable coefficient is impacting the dependent variable.

In previous analysis we have learnt that cereal products with less calories have higher ratings, and therefore we wanted to further analyze the data set to discover the effects of some variables on calories. This will help manufacturers of cereals to have better insight on what to add or reduce in their cereal to get better ratings and grocery stores to stock up on the cereals that are going to have more demand.

Please note that for each one of these models, the hypothesis remains the same:

H0 = the model is not useful (The slope is equal to zero that is the independent variable does not influence the dependent variable)

H1 = the model is useful (The slope is not equal to zero that means the independent variable influences the dependent variable)

**Rating vs Sugar**

Here we are using linear regression to see how the amount of sugar in the cereal effects the rating.

**R code:**

cereal<-read.csv ("cereal.csv")

linearMod <- lm(rating ~ sugars, data=cereal)

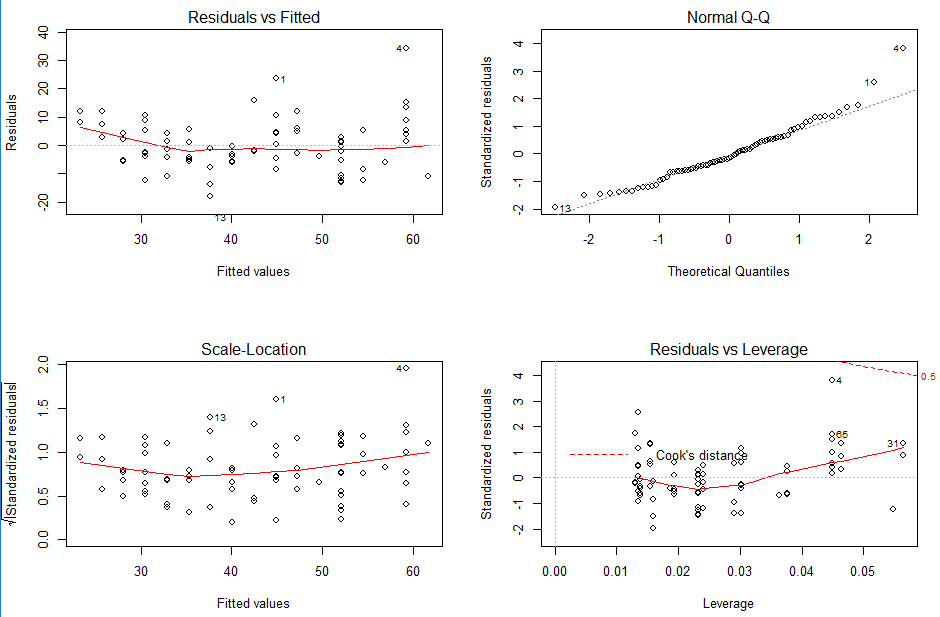
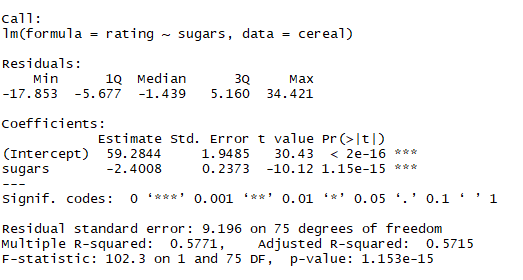
summary(linearMod)

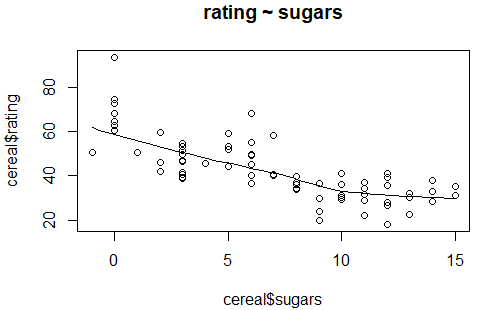
par(mfrow=c(2,2))

plot(linearMod)

scatter.smooth(x=cereal$sugars, y=cereal$rating, main="rating ~ sugars")

**Results:**





**Interpretation:**

* Rating = 59.2844 - 2.4008 sugars.
* With every unit increase in sugar the rating is reducing by 2.4008 units.
* 57.15 % of variance in rating can be explained by sugars we look at adjusted R square

for this.

* P value is less than 0.05, we reject the null hypothesis and say the model is a good fit.
* We can look at the above graphs and say assumptions of linear regression have meet.

**Recommendation:**

The cereal companies must reduce the amount of sugar in the cereal if they want to increase the rating, because having too much of sugar or artificial for breakfast can cause rapid fluctuations in the body sugar levels.

**Rating vs Fiber**

**R code:**

linearMod <- lm(rating ~ fiber, data=cereal)

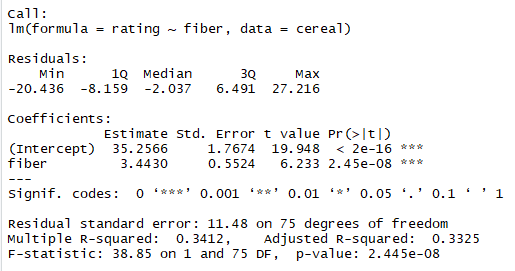
summary(linearMod)

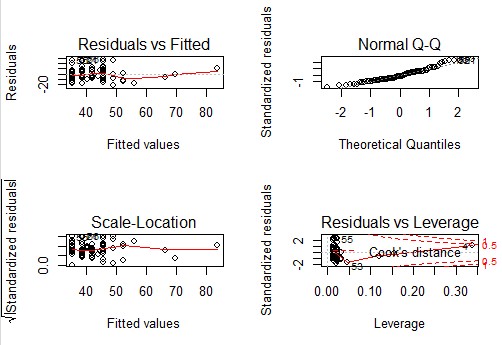
par(mfrow=c(2,2))

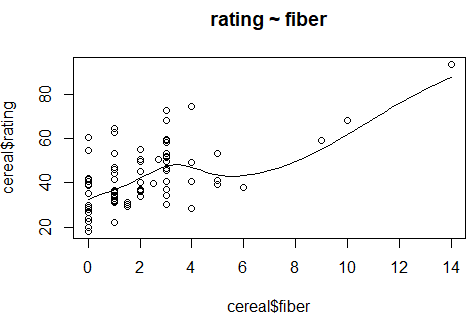
plot(linearMod)

scatter.smooth(x=cereal$fiber, y=cereal$rating, main="rating ~ fiber")

**Results:**







**Interpretation:**

* Rating = 35.2566 + 3.4430 Fiber.
* With every unit increase in fiber the rating is increasing by 3.4430.
* Variance of 33.25 % in rating can be explained by fiber we look at adjusted R square and the model is a moderate fit.
* P value is less than 0.05, we reject the null hypothesis and say model is a good fit.
* We can look at the Q-Q plot and say that the residuals are normally distributed.

**Recommendation:**

The cereal companies must increase the amount of fiber in the cereal if they want to keep up their current rating or want to increase their rating, because studies have shown that having high fiber diet can have related health benefits and can also reduce the blood pressure.

**Rating vs Calories**

**R code:**

**l**inearMod <- lm(rating ~ calories, data=cereal)

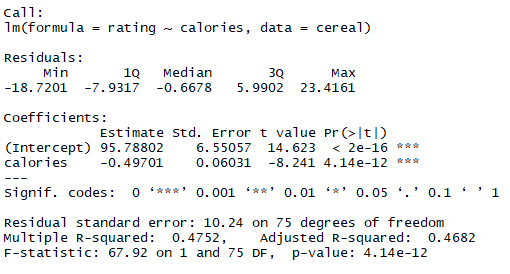
summary(linearMod)

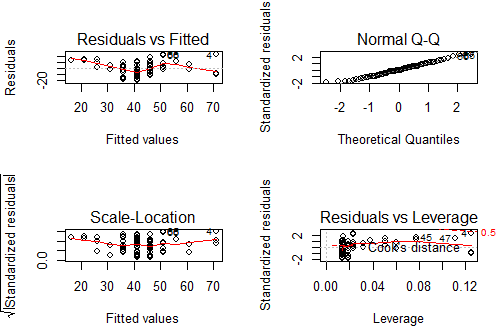
par(mfrow=c(2,2))

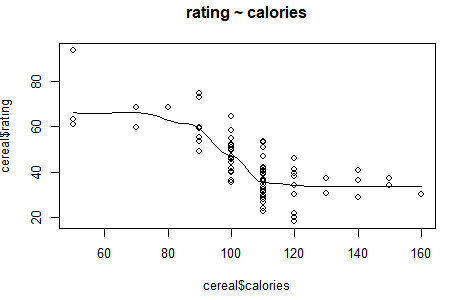
plot(linearMod)

scatter.smooth(x=cereal$calories, y=cereal$rating, main="rating ~ calories")

**Results:**







**Interpretation:**

* Rating = 95.78802 – 0.49701 calories.
* With every unit increase in sugar the rating is reducing by 2.4008 units.
* Variance of 46.82 % in rating can be explained by calories we look at adjusted R square for this and we can also that the model is a moderate fit.
* P value is less than 0.05, we reject the null hypothesis and say the model is a good fit.
* We can look at the Q-Q plot and say that the residuals are normally distributed.

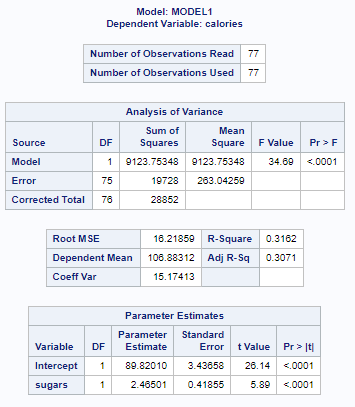
**Recommendation:**

The cereal companies must reduce the number of calories in the cereal if they want to increase the rating, having high amount of sugar and fat can lead to increase in calories in cereal this can also be determined from correlation matrix, a lot of people are becoming health conscious and are trying to avoid bad calories and unhealthy cereals, so when manufacturing cereals attention must be paid to the amount of calories included in the cereal.

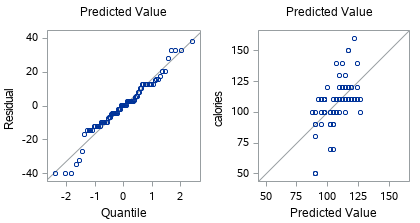
**Calories vs Sugar**

We will like to find out the effect of sugar on calories based on the cereal data set. We will use linear regression as a supervised learning to do this:

**SAS Results:**







**Interpretation:**

* From the above results of our analysis, we can see that the model is significant, because the F – Statistics P value is less than Alpha (0.05) and thus we reject the null.
* Observing the adjusted R-Square we can conclude that 30.71% of the variance in Calories that is explained by Sugar.
* The Linear regression equation here is: Calories = 89.82010 + 2.46501 Sugar.
* With every unit increase in sugar, the calories increase by 2.46502 units.
* The graph shows that the normality of the residuals holds good since the histogram fits ok in the normal curve.

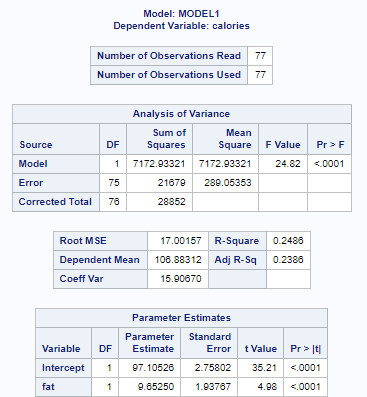
**Recommendation:**

We can see from the above analysis that Sugar increases the number of calories in cereal and as we learnt earlier also, calories causes lower ratings. Therefore, we can conclude that sugar may be one of the ingredients to be reduced if a manufacturer wants lower calories and increase ratings.

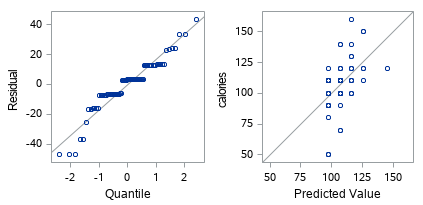
**Calories vs Fat**

We will like to find out the effect of fat on calories based on the cereal data set. We will use linear regression as a supervised learning to do this:

**SAS Results:**







**Interpretation:**

* From the above results of our analysis, we can see that the model is significant, because the F – Statistics P value is less than Alpha (0.05) and thus we reject the null.
* Observing the adjusted R-Square we can conclude that 23.86% of the variance in Calories is explained by Fat.
* The Linear regression equation here is: Calories = 97.10526 + 9.65250 Fat
* For a unit increase in Fat, the calories increase by 9.65250 units.
* The graph shows that the normality of the residuals holds good since the histogram fits ok in the normal curve.

**Recommendation**

We can see from the above analysis that Fat increases the number of calories in cereal and as we learnt earlier also, calories causes lower ratings. Therefore, we can conclude that Fat may be one of the ingredients to be reduced if a manufacturer wants lower calories and having high amount of saturated fat in the morning is not healthy for people.

**CLUSTERING**

Clustering analysis allows us to find any hidden structure in an unlabeled dataset. The items in a cluster are grouped based on the similarities and items within a cluster are more similar to each other as compared to the items in other clusters. To further understand how the cereal brands are similar, an analysis has been performed to understand the groupings.

**Calories and Rating**

We want to know what kind of brands have similarity based on the calories and ratings

we used WSS plot to get the optimal number of clusters

**R code:**

setwd("C:/Users/gprem/Downloads/cereal")

cereal<-read.csv ("cereal.csv")

mydata <- cereal[ ,c("calories","rating")]

wss <- (nrow(mydata)-1)\*sum(apply(mydata,2,var))

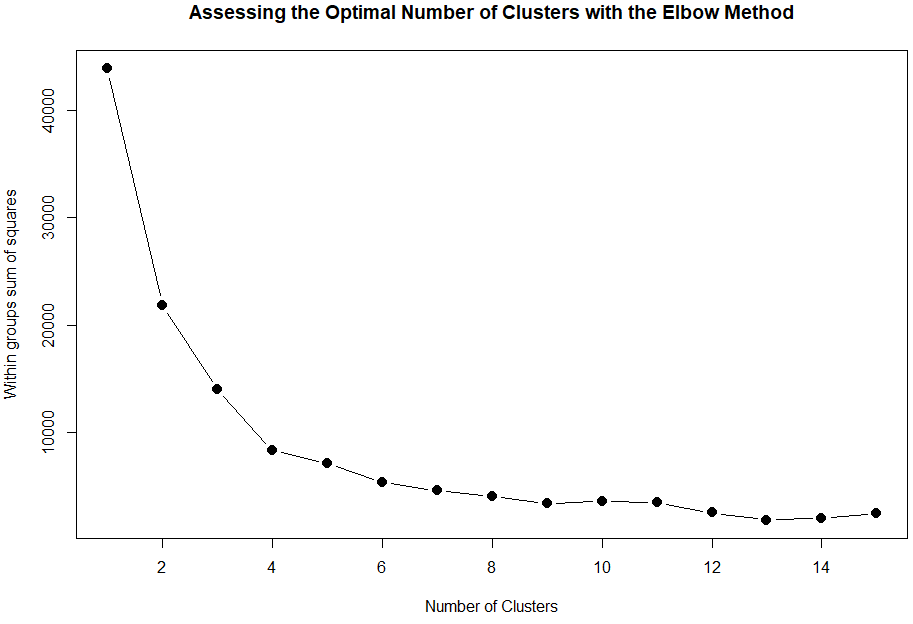
for (i in 2:15) wss[i] <- sum(kmeans(mydata,

centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",

main="Assessing the Optimal Number of Clusters with the Elbow Method",

pch=20, cex=2)

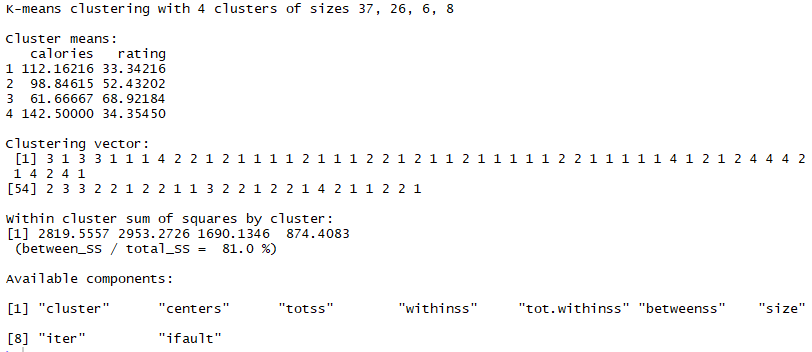


After looking at the WSS plot, the elbow is at 4 and we take the optimal number of clusters to be 4

**R code:**

km2 = kmeans(mydata, 4, nstart=100)

km2



From the above figure we can say that the clusters are divided into sizes of 37, 26, 6, 8.

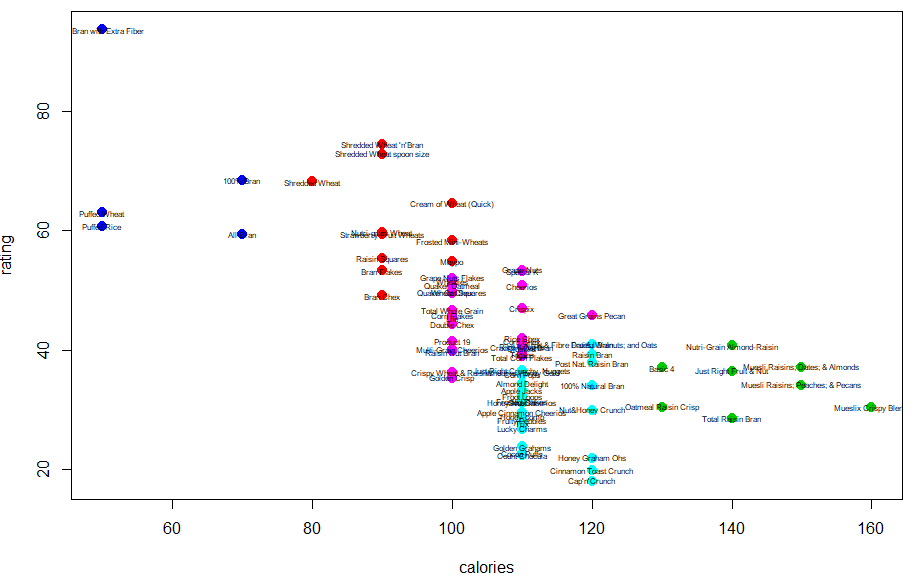
Now let’s plot the graph and see which cereal brand belongs to which cluster.

**R code:**

plot(mydata, col =(km2$cluster +1) , main="K-Means result with 4 clusters", pch=20, cex=2)

text(x=mydata$calories, y=mydata$rating, labels=cereal$name, cex=0.5)

**Results:**



**Interpretation:**

* The blue observations are the ones with the least number of calories and highest ratings.
* The green observations are the brands with the greatest number of calories and the worst rating.
* From our analysis we can say that the cereals that have the least number of calories are the most preferred by people, everyone these days are becoming health conscious and are trying to eat healthy, so it is obvious that people take into consideration about the amount of calories they eat for breakfast, brands like Muesli Crispy Blend and Muesli Raisins; Dates & Almonds should take this into consideration if they want to improve their ratings among the customers and our opinion is the ones in blue are considered best to buy.

**Nutrients**

We wanted to check how the cereals are grouped together based on the nutrients present in them as per the research and with the data available with us we have used protein, fat, carbohydrates and fiber since they are among the top essential nutrients to the body, using these variables we have conducted a cluster analysis. To determine the optimal number of clusters we have done k means clustering below are the results.

**R code:**

cereal<-read.csv ("cerealhclust.csv")

clu <- cereal[,-1]

head(clu)

mydata <- clu

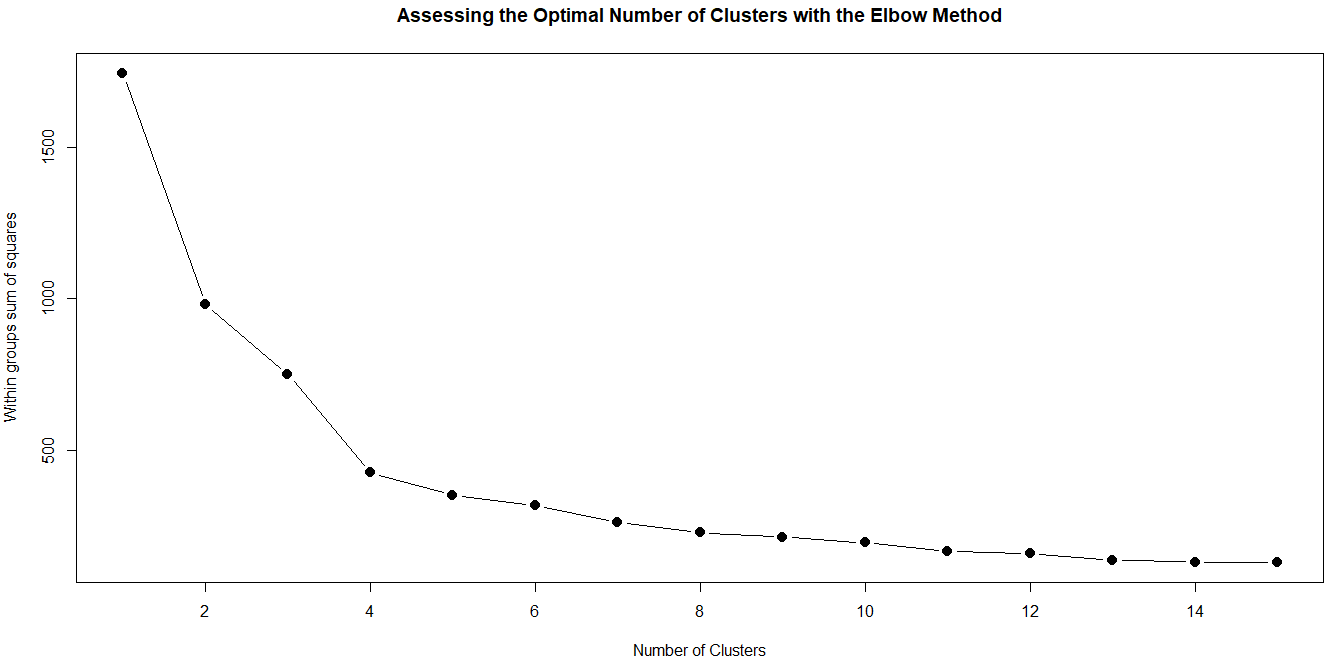
wss <- (nrow(mydata)-1)\*sum(apply(mydata,2,var))

for (i in 2:15) wss[i] <- sum(kmeans(mydata,centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares",main="Assessing the Optimal Number of Clusters with the Elbow Method",

pch=20, cex=2)

**Results:**

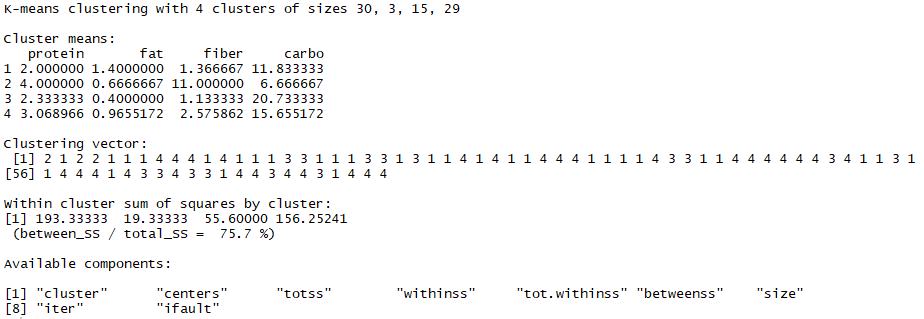


From the above graph we can see that the elbow is at 4 so we have considered 4 to be the optimal number of clusters.

**R code:**

km1 = kmeans(clu, 4, nstart=100)

km1



Here we have also tried to check if there is much of a change in the between\_SS / total\_SS value if we took 5 clusters and the different was not significant. The clusters sizes are 30,3,15,29

Now after getting the optimal number of clusters we are doing the hierarchical clustering to see which cereal belongs to which cluster.

**R code:**

mydata<-read.csv ("cerealhclust.csv")

mydata

row.names(mydata) <- mydata[,1]

d <- dist(mydata,method="euclidean")

d

fit <- hclust(d)

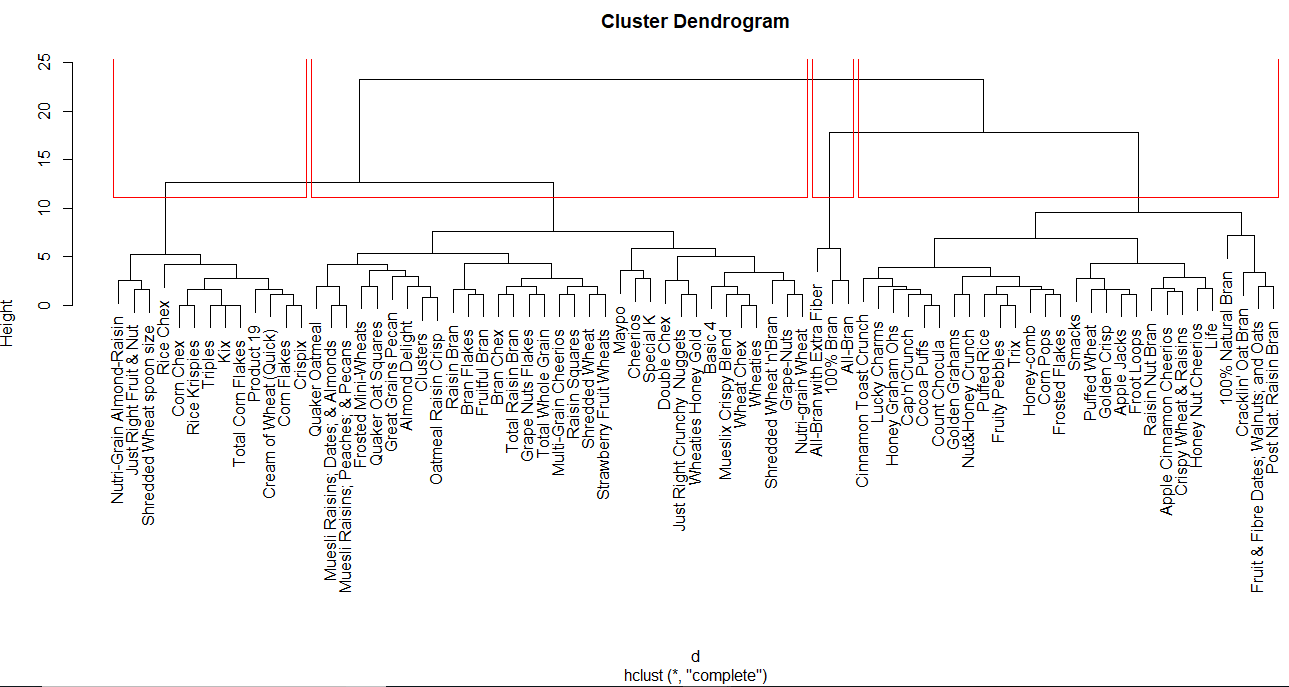
fit

plot(fit)

groups <- cutree(fit, k=4)

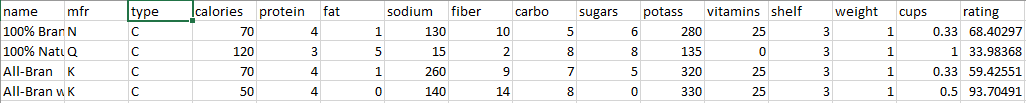
groups

rect.hclust(fit,k=4,border="red")



**Interpretation:**

From the above plot we can see that All bran with extra fiber, 100% bran and all bran all belong to the same cluster. We have also compared these results with the results we obtained from the k-means clustering; the first, third and the fourth observations belong to the same cluster -number 2. Also, when you look at the dataset given, we can see that 100% Bran is the first observation, All-Bran is the third observation and All-Bran with Extra Fiber is the fourth observation.



From the above plot we can say that the observations are grouped together based on the essential nutrients present in them. On further analysis we have investigated cereals based on cluster division and we found that the cereals which have high content of protein, fiber and less carbohydrates are grouped into a single cluster. Another cluster is formed where there is no or very little fiber content and optimal calorie content, whose average was around 12 grams as well as the protein content average is 1 gram. A cluster is formed where the protein content is around 3 grams and the average carbohydrate present is 15 grams.

**CONCLUSION**

The aim of our research was to analyze and understand some relevant information about cereals. From our analysis, we were able to analyze some of the ingredients contained in cereals and how they impact customers rating. We did this because it gives us some insight on the nutrition facts in a cereal that customers regard as important.

Furthermore, the analysis in our report can serve as a reference or source of information for companies that manufacture cereals. Cereal manufacturing companies can now identify the factors that can impact their product. Breakfast being the most important meal of the day, as most of the cereals are consumed in the morning the company should pay attention to the amount of nutrients they are adding into their products.

**RECOMMENDATIONS**

Our recommendation is targeted at cereal manufacturing companies. We can see that customers pay a large amount of attention to the nutrition contents in cereal products therefore we recommend that cereal manufacturing companies try to factor this when they make their products if they want to improve their rating.

**REFERENCES**

<https://www.kaggle.com/crawford/80-cereal>