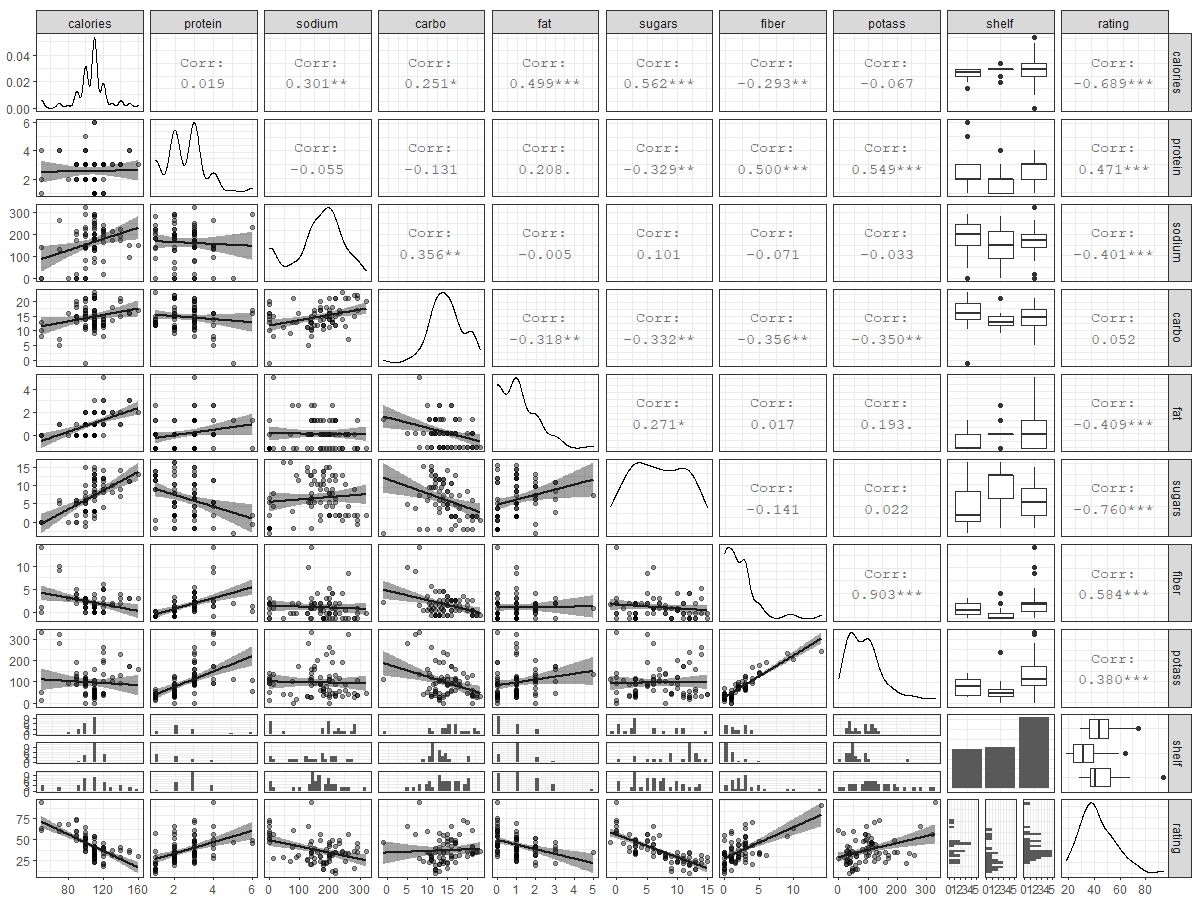
**First do full model lm**

**Then compare to full model cp etc stepwise**

**Plot(modle) look at residuals see if need transformation QQ plot**

**Look at studentized residuals**

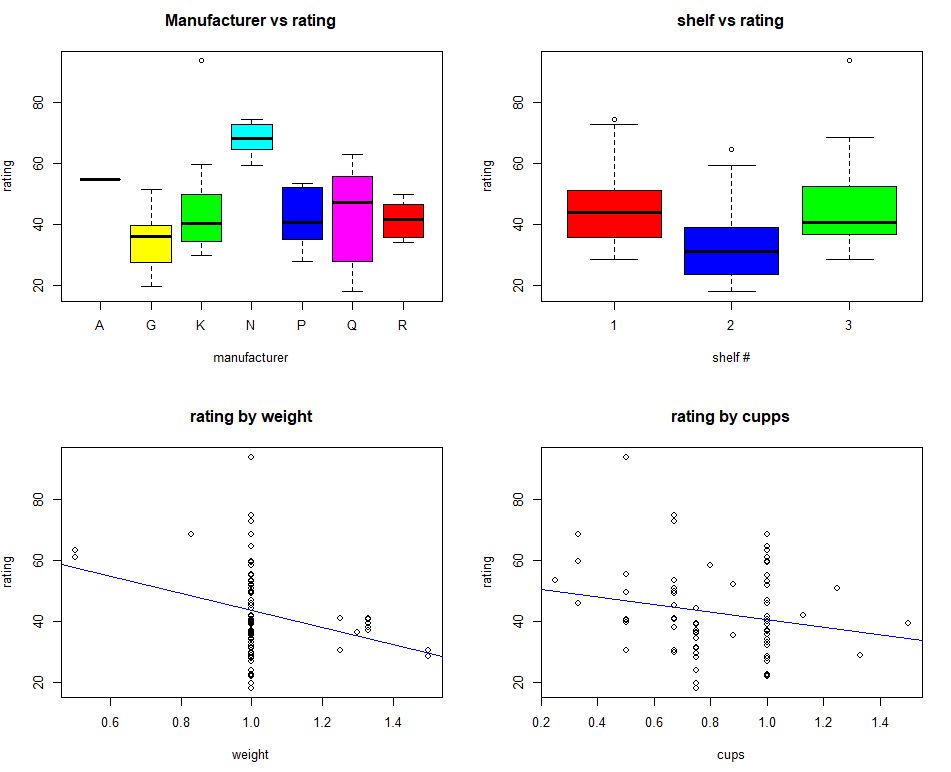
****

**Here we can see a few things:**

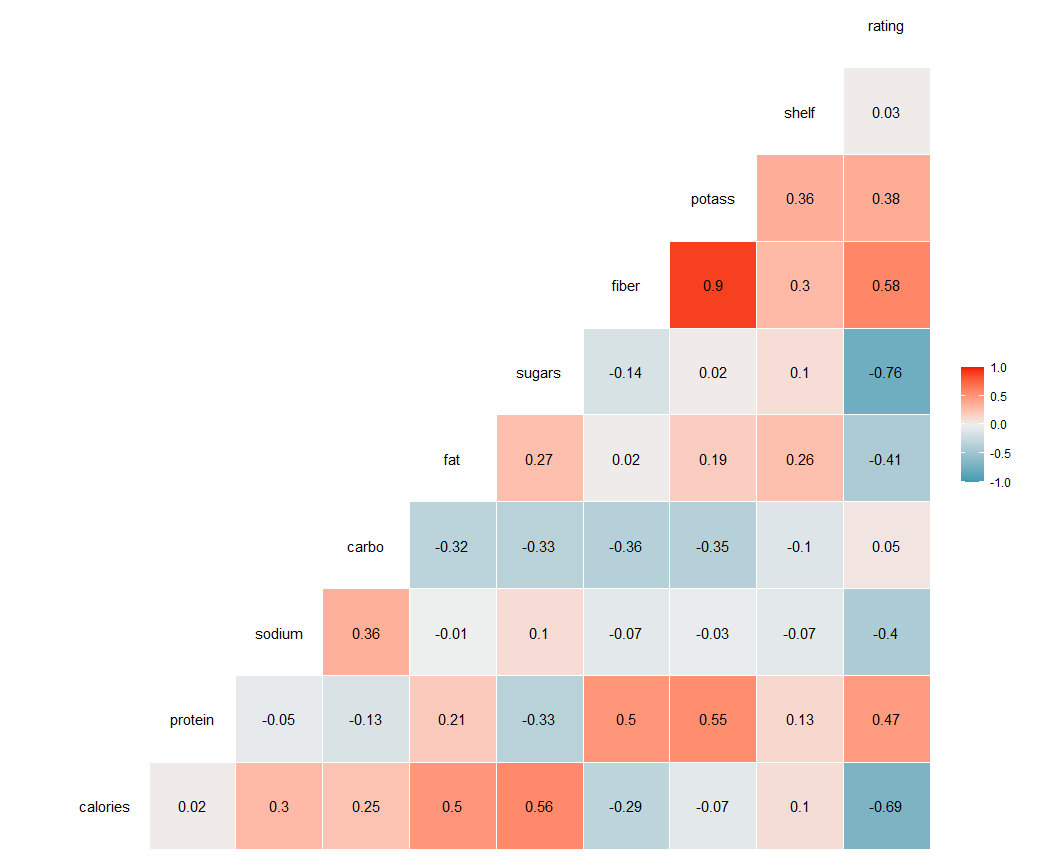
1. **The higher the rating the lower the calories and highest mean is for shelf 2**
2. **The higher the rating the higher the protein and highest mean is for shelf 3**
3. **The higher the rating the lower the sodium and highest mean is for shelf 1**
4. **Rating and carbohydrate almost have a flat relationship. and highest mean is for shelf 1**
5. **The higher the rating the lower the fat content and highest mean is for shelf 3**
6. **The higher the rating the lower the sugar content and highest mean is for shelf 2**
7. **The higher the rating the higher the fiber content and highest mean is for shelf 3**
8. **The higher the rating the higher the potassium content and highest mean is for shelf 3**

**Rating vs Cup, Weight, Manufacturer and Shelf**

**From the blow plots, rating seems to be dependent on manufacturers. Nabisco has the highest manufacturer rating and General mills the lowest. Cereals on shelf 2 has the overall worst rating and there is a negative weak linear relationship when comparing weight vs rating and rating by cups.**



**Here we can see in the below correlation plot that there aren’t many variables that could be questioned to have multicollinearity except potassium and calories. We will look further into their relationship later into the exploration.**

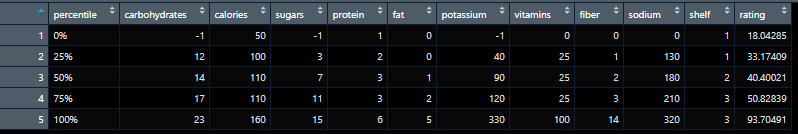
****

**Percentile number for each predictor**

**Qualifications of healthy Cereal**

1. **Fat: No more than 30% of calories from fat**
   1. **1g fat is 9 calories**
2. **Protein: 50g women and 63g daily men**
   1. **12 – 20 % of protein: (**[**Balancing Carbs, Protein, and Fat | Kaiser Permanente Washington**](https://wa.kaiserpermanente.org/healthAndWellness/?item=%2Fcommon%2FhealthAndWellness%2Fconditions%2Fdiabetes%2FfoodBalancing.html#:~:text=In%20a%20healthy%20diet%2C%20about,growth%2C%20maintenance%2C%20and%20energy.)**)**
3. **Carbohydrates and proteins contain 4 calories per gram**
   1. **Recommended 325g a day (**[**Carbohydrates: How carbs fit into a healthy diet - Mayo Clinic**](https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/carbohydrates/art-20045705#:~:text=The%20Dietary%20Guidelines%20for%20Americans,grams%20of%20carbohydrates%20a%20day.)**)**
   2. 45-65% of daily recommended calories **(**[**Carbohydrates: How carbs fit into a healthy diet - Mayo Clinic**](https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/carbohydrates/art-20045705#:~:text=The%20Dietary%20Guidelines%20for%20Americans,grams%20of%20carbohydrates%20a%20day.)**)**
4. **Fiber =** 25-35 grams of fiber daily
5. **Sodium =** Consume less than 2,300 mg a day **([Boiling down the dietary guidelines - Mayo Clinic](https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/dietary-guidelines/art-20045584))**
6. **Sugar =** less than 10% of daily calories from sugar **(**[**Boiling down the dietary guidelines - Mayo Clinic**](https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/dietary-guidelines/art-20045584)**)**
   1. Has 4g calories per gram sugar
7. **Potassium =** 3500-4700 mg daily **([Potassium - Health Professional Fact Sheet (nih.gov)](https://ods.od.nih.gov/factsheets/Potassium-HealthProfessional/))**
8. **Vitamins =** at least 1000 mg per day

**Based on findings let’s find the brands in the healthy quantiles for each variable**

****

**Percent of calories from carbs, fats, proteins and sugars**

****

1. **Carbs:** The plot above shows that almost all cereals stay within the 45-65% of calories from carbs the max being 57.7% just not the 0 percentile.
2. **Fat:** This also shows all brands in regards to fat only 28% of daily fat at 9 calories per g of fat (5x9 = 45/160 = 28), which is less than 30% of calories from fat.
3. **Protein:** The highest protein is 6 grams and around 3g for most cereals, also coming in at a max of 15 percent of calories which falls in the range of 12-20% of calories from.
4. **Sugar:** sugar all brands are over 10% for the sugar content.
5. **Potassium:** Based on the above information the maximum potassium is 330 and recommended daily is 3500-4700 so it doesn’t seem to be a good source.
6. **Vitamins:** The daily recommended amount is 1000mg and max here is 100 the more the better but its not a good source.
7. **Sodium:** Doesn’t hit the max suggested of 2,300mg per day so it is ok but the less the better
8. **Fiber:** 25-35 grams of fiber daily the max is 14 which seems to be a sufficient source of fiber at about 50% of daily recommended value

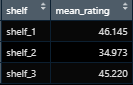
**Scientific Goals**

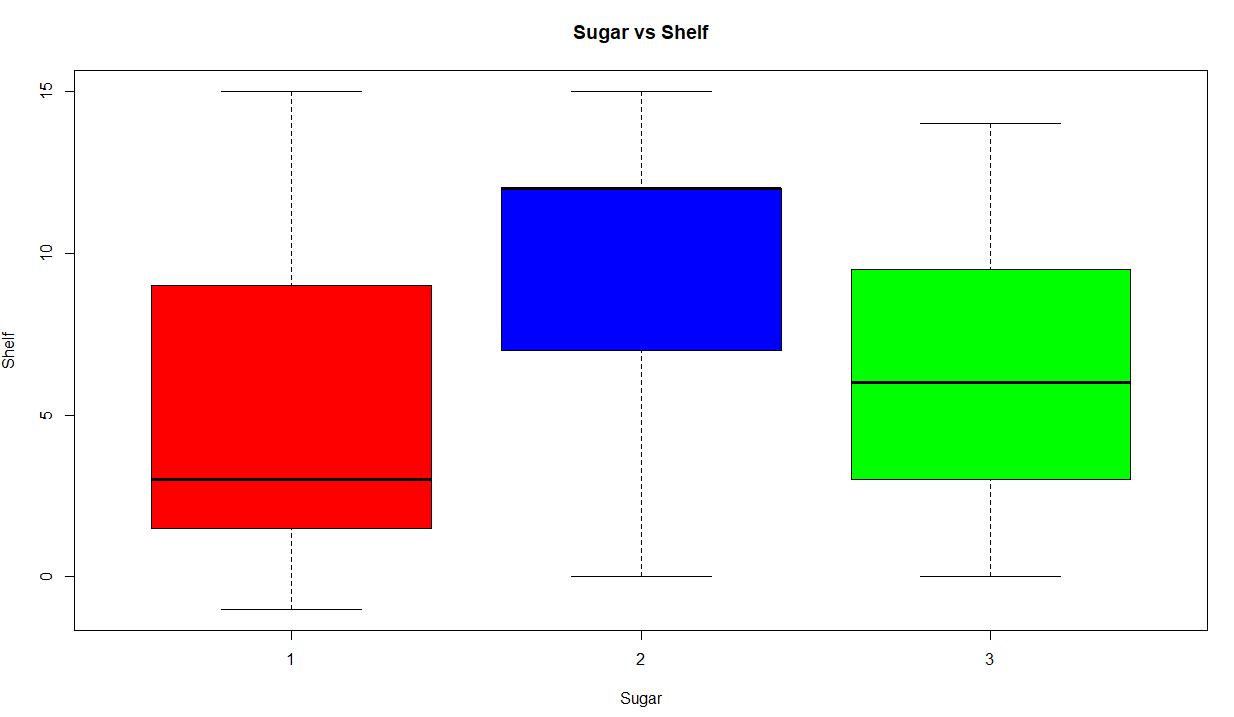
1. **How would you formulate a healthy cereal? I would follow these guidelines for a healthy cereal.**

* Carbs to stay in range I would say a cereal brand would have to be between 12 and 23g of carbs
* Fats to stay in range a healthy cereal should be 5g of fats or less
* Proteins to stay in range the range of 11-19g of proteins to stan in the 12-20% of calories range
* Sugars should be less than a gram to hit the suggest range of less than 10%
* All vitamins and fiber and potassium should be high and sodium low

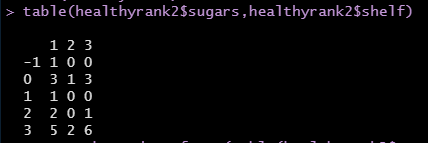
1. **Based on your formulation, can you say whether healthier cereals tend to be displayed on the upper or lower shelves?**
   1. Because carbs, fats and proteins are all pretty much in a healthy range for all brands and vitamins and cereal is not a sufficient source of vitamins or potassium. I would say that how much sugar a cereal has is an indicator of how healthy it is. Based on the plot below the second shelf is the shelf that holds the cereals with the highest sugar contents. And also see that the brands with best ratings are on shelf 1.

**Mean Rating By Shelf #**

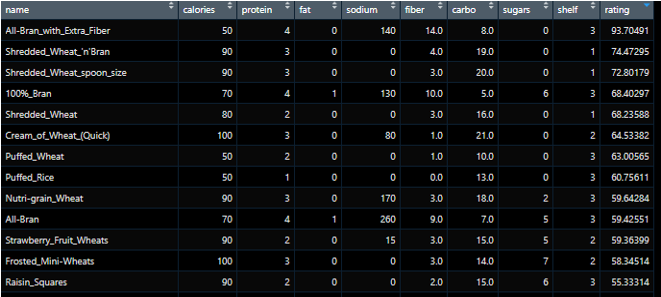




* 1. Based on that we can see in the chart below shelves being column header and row names being grams of sugar. We can see cereals at the 0 percentile occur mostly on shelf 1 and 3. But if you add -1 gram of sugar and 1 gram you can see that shelf 1 is where the healthy cereals are displayed.



* Below we can see some of the top brands with the highest ranking (They seem to fit the structure of low sodium, low sugar, low to medium fat, high to mid carb and high to mid protein.



**Simple Linear Regression Modeling**

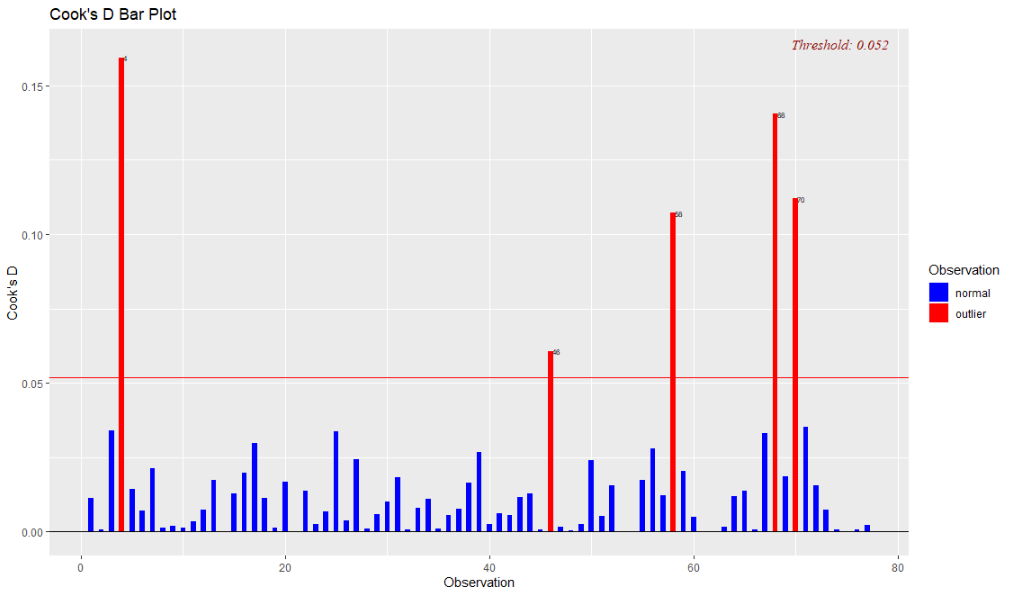
To under stand the relationships between each individual predictor variable in more depth we isolate them in a single regression model and create a table below. This analysis shows us the model, R squared values, p-values and correlation of each predictor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor Variables** | **Correlation** | **P-Value** | **R2** | **Model** |
| Calories | -0.689 | 4.1403e-12\*\*\* | 0.475 | 95.788 – 0.497 |
| Carbohydrates | 0.052 | 0.653 | 0.002 | 40.171 + 0.170 |
| Protein | 0.470 | 1.5663e-05\*\*\* | 0.221 | 27.29 + 6.04 |
| Fats | -0.409 | 0.0002 \*\*\* | 0.167 | 48.45 – 5.71 |
| Sugars | -0.759 | 1.1534e-15\*\*\* | 0.577 | 59.3 – 2.4 |
| Fiber | 0.584 | 2.4453e-08\*\*\* | 0.341 | 35.26 + 3.44 |
| Sodium | -0.401 | 0.0003\*\* | 0.161 | 53.4027 – 0.0672 |
| Vitamins | -0.241 | 0.0350\* | 0.058 | 46.938 – 0.151 |
| Potassium | 0.380 | 0.0006\*\*\* | 0.144 | 35.4682 + 0.074 |

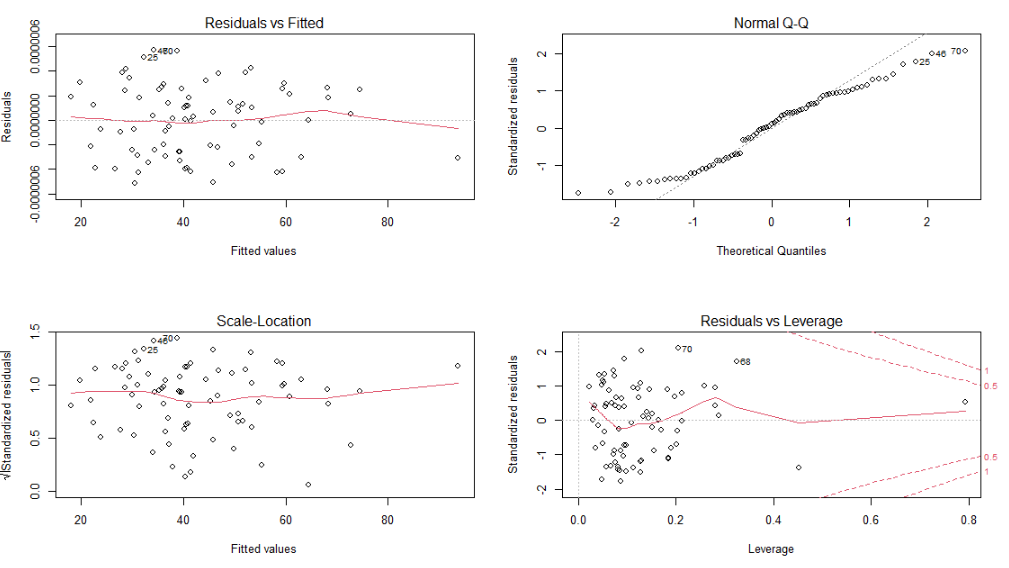
The above simple linear regression models tell us that correlation between rating and sugars is the strongest. This model also gives us the largest R2 and smallest p-values. We can reject that the regression coefficients are zero for, sugars, protein, calories, fiber, sodium, vitamins, fats and potassium. The variable carbohydrates has a p-value of .653 indicating that we fail to reject the null hypothesis that the carbohydrate coefficient is not significantly different from 0. We will take a look into seeing if that variable is worth taking out of the full model later.

**Multiple Linear Regression Modeling**

Now we will fit multiple predictor variables into the full model in this case will be: Rating ~ Calories + Carbohydrates + Protein + Fats + Sugars + Fiber + Sodium + Vitamins + Potassium. We will examine the assumptions of the model first by doing some diagnostics in below plots.

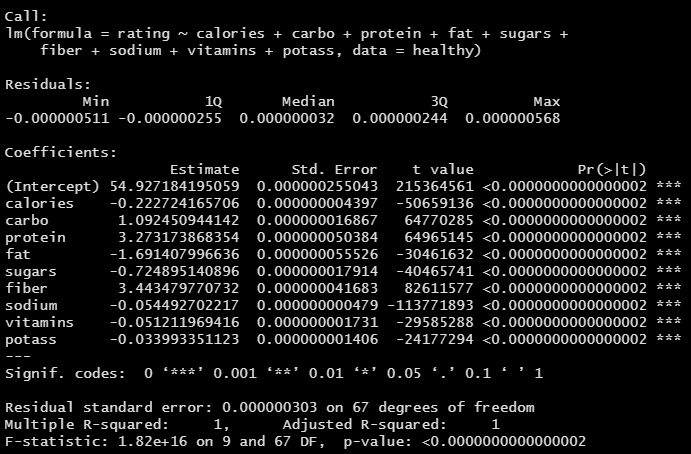


In the cook’s distance plot above we can see observation 70,68,58,46 and 4 are outliers but observation 4 is the most influential to the model.



Based on the plot above the QQ plot shows the assumption of normality looks like it might not be a normal distribution but it also looks like it is kind of holding up a linear shape. There doesn’t look like any curvilinear trends in residuals as well.

**Full Model Summary**



Here in the full model summary above we can see that all of the p-values are significant. This is the case even though individually the predictor carbo (carbohydrates) was insignificant, we will keep this variable when added it adds some important information to the model. We can also see that the R2 value is 1 which means that the model full explains the variability in the model. This means we can drop any further investigation to any transformation such as log transformation because the model is fully explained. We can conclude that the model is as follows:

**Full Model**

Rating = 54.93 – 0.22 \* Calories + 1.09 \* Carbohydrates + 3.27 \* Protein – 1.69 \* Fat -0.72 \* Sugars + 3.44 \* Fiber – 0.054 \* Sodium – 0.051 \* Vitamins – 0.034 \* Potassium

1. **Do you agree with Consumer Reports grading of the cereals?**
   1. Consumer report is a unknown formula based on the nutritional content. Thus, we examine the relationships between the different nutritional variables in this dataset to understand the significance of these relationships in their secret formula.

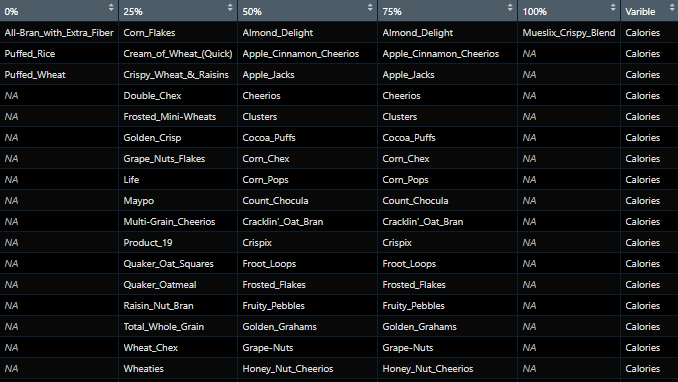
**Answer:** I do agree with Consumer Reports Grading of Cereals because the regression model fitted was an R2 of 1 explaining all of the variance in the data.I think they did a pretty accurate job on describing the ranking of cereals.

**4. Clearly, the prices of the cereals are not reported in this data. Would it be the case that healthier cereals are the more expensive ones?**

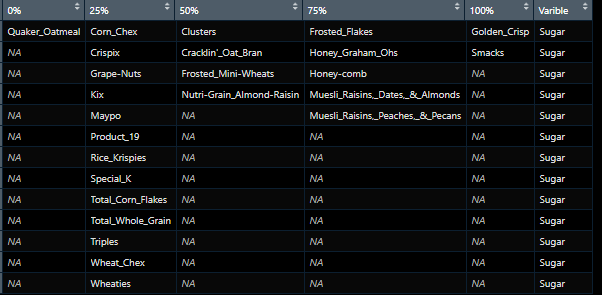
**What brands are in what percentile for carbohydrates**

****

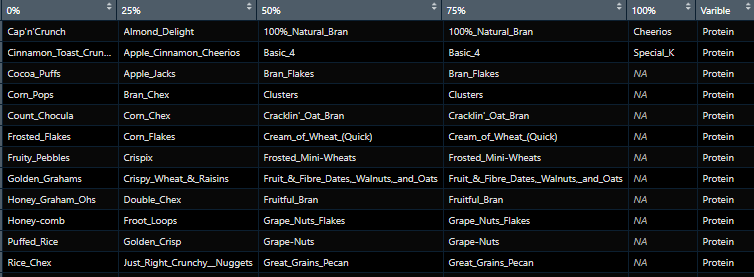
**What brands are in what percentile for calories**

****

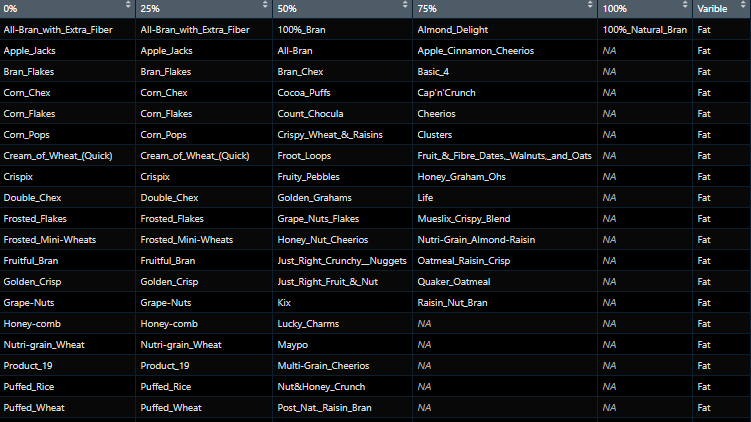
**What brands are in what percentile for sugar**

****

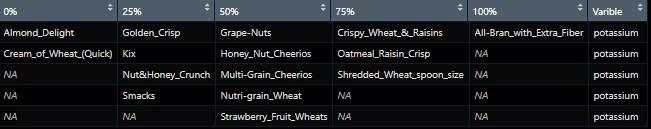
**What brands are in what percentile for protein**

****

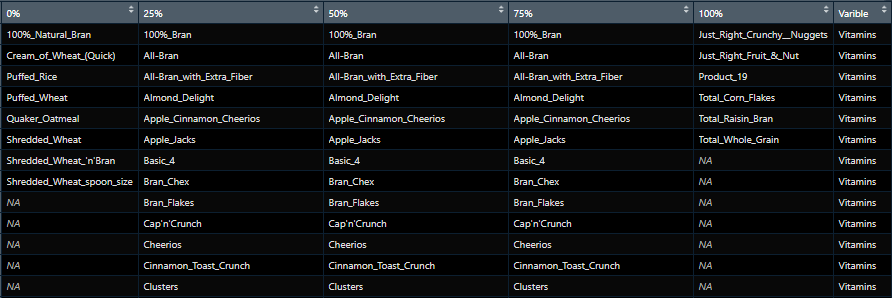
**What brands are in what percentile for fat**

****

**What brands are in what percentile for potassium**

****

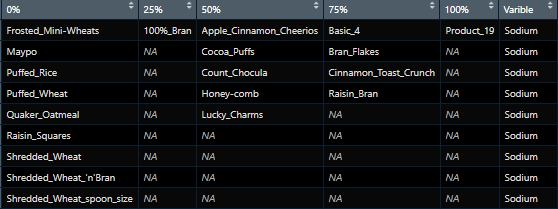
**What brands are in what percentile for vitamins**

****

**What brands are in what percentile for fiber**

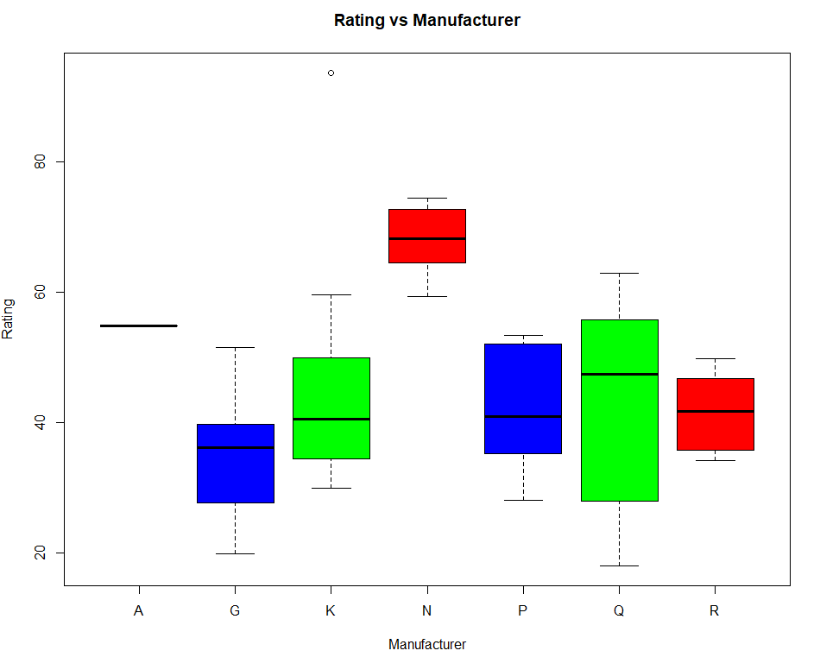
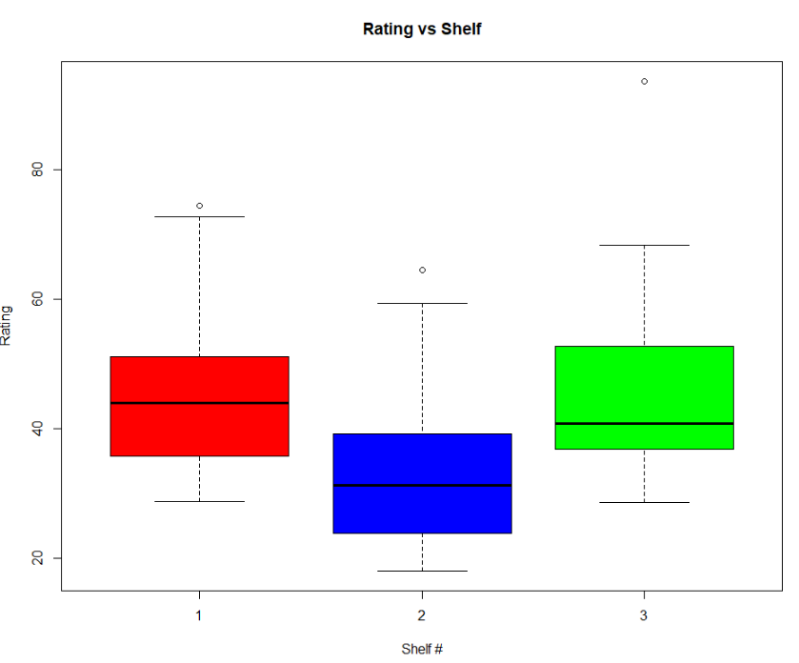
****

**What brands are in what percentile for sodium**

****

**What brands are in what percentile for ratings**

****

****