Kyle Bermudez

Math 437 Final Report

For this case study, we were to analyze data from Netflix and Cereal. The Netflix data set was a huge data set that had a lot of unnecessary variables for what the problems are asking us, so throughout the problems for Netflix, I handpicked variables of interest to perform statistical inference techniques. The Cereal data set on the other hand, is a much smaller data set that displays the different types of cereals, their nutritional facts, what manufacturer it came from, consumer report ratings, and where they are displayed on shelves of one through three (one being the first row from the floor, two being the second row from the floor, and three being the third row from the floor). My opinion on the questions that were being asked for this problem seems to be a big exploratory data analysis question, which was very fun to me simply because I like seeing the relationships of certain things that I will go over in this report. But first, let's talk about Netflix.

For problems *a* through *c*, I have come across obstacles that I can't seem to overcome. This was because I did not know how to handle the variables genre and box office. I had issues with genre because it is a categorical variable with more than two outcomes, and box office seemed to be too big (double) whenever I try to implement any sort of regression with it. The keyword for problem *a*, among the movies only, also tripped me a bit as well because I have created a binary dummy variable for Series.or.Movie, but it just added more intricacies that I did not know how to address, thus for problems *a* through *c*, I have omitted them and moved on to part *d*. For part *d*, I would say logistic regression would be the most reliable for classifying series vs movies based on past projects, but it seems to show that LDA produced a better classification rate for classifying series versus movies than logistic regression and QDA as seen in Figure 1,2,

and 3 in the appendix. For the ease of the reader, logistic regression, LDA, and QDA produced a classification rate of about 77%, 79%, and 61% respectively.

For part *e*, we were to build a prediction model for hidden gem scores and report our findings. For this, I chose to perform lasso and ridge regression to produce my prediction model. To determine which predictors to use for our prediction model, I used primarily numerical predictors and a categorical variable series or movies for this prediction model. The lasso and ridge regression models came out about to be the same with an MSE of 3.463 for lasso regression, and an MSE of 3.461 for ridge regression; as seen in Figure 4 and 5. Although Ridge regression produced a more accurate model, it would be difficult to interpret how each of the predictors impact the response, but in terms of model prediction accuracy, then Ridge regression would be what we want since it is the scope of our objective.

For part *f*, we were to find out among the movies, what predictors influence box office more, so I used best subset selection and hand picked the variables that would be pertinent to this problem. It showed us that there would be five variables, which are: IMDb score, Rotten Tomatoes Score, Awards Nominated For, IMDb Votes, and Hidden Gem Score; as shown in Figure 6. Problems that could arise from only hand-picked variables that I see fit is the lack of variable selection variety and one important predictor that I left out which is genre. I think genre would be a beneficial predictor to have, but I had some problems trying to implement genre as one of my variables to choose from. I also had troubles implementing the whole data set's variables into my best subset selection because it would simply be too big. It turned out to be 18.2 GB of data to store and filter through for the best subset of variables to determine box office.

For part *g*, we were to build a predictive model for box office. In an ideal world, I would want to use what variables that were selected from part *g* to help me make a lasso regression model for prediction; but I also came across more issues. Like what I did in part *f* I hand-picked my variables for my prediction model, but when I tried to build the parts for lasso regression, I had an error in which I had one multinomial or binomial class that had 1 or 0 observations. I was unsure how to troubleshoot this issue, so I will describe how I would have done this problem if I sought help from a colleague. If I had help from a colleague or a professor, I would address which variable produced that error and even possibly ask how to implement all the categorical variables that I would think would be helpful for making a prediction model of box office. Once I have established that, I would proceed to use lasso regression and cross-validate my model with ridge regression to compare my results and determine which model had better accuracy.

For part *h*, we were to cluster the data and explore each cluster. For this approach, I tried to use hierarchical clustering so that I can explore each cluster in detail with the dendrogram, but I came across a small error that prevented me from doing so. Although it still produced a dendrogram, it was nearly impossible to interpret, but the error that it gave me was "dist(): NAs introduced by coercion". I interpreted this error message as that I could not have any NAs in any cell of the data set, but I thought I addressed that when I made all cells to be 0. Nevertheless, I was able to produce a dendrogram as seen in Figure 7, but as I stated previously, I was unable to interpret the cluster besides the fact that it has three clusters that have great distances between each of them.

Moving on to the case study for cereal, I will discuss each point that is made under specific questions in order. It is said that to have a healthy and balanced diet, an adult should consume no more than 30% of their calories in the form of fat, they need about 50g or 63g of

protein daily depending on if they are male or female, eat about 20-35 grams of fiber, and they should provide for the remainder of their caloric intake with complex carbohydrates. First question is regarding if I agree with the consumer reports grading of the cereal. Without any sort of deep statistical inference and just going off the nutritional facts that are presented in the data file, I would have to agree with the consumer grading reports. I agree with the reports of the cereal because if I use All Bran with Extra Fiber as a benchmark for all the other cereals, then we can see why it has the highest rated cereal. It is the highest rated cereal because it is a healthy cereal that fits the needs of a healthy and balanced diet that I have stated previously. All Bran with Extra Fiber has the lowest calories (50 calories/serving), some protein (4g), no fat(0g), low sodium(140mg), high fiber(14g), moderate carbohydrates(8g), zero sugars(0g), a great source of potassium(330mg), and two cups of it equates to one serving, so it is very nutritious and a filling cereal to have for breakfast than let's say Cap'n'Crunch which has higher calories(120 calories/serving), low protein(1g), low fat(2g), high sodium(210mg), zero fiber(0g), high carbs(12g), high sugars(12g), low potassium(35mg), and two cups of it equates to 1.5 servings. This is just going off the numbers alone in comparing the best rated cereal and the lowest rated cereal that are presented from the data file.

Moving on to question 2, this question asks how I would formulate a healthy cereal and is it consistent with the consumer reports. I formulated what makes a healthy cereal by using best subset selection to determine which variables had the most impact on ratings. I have come to the conclusion that there are nine variables that impacted this, which are: calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, and vitamins; as seen in Figure 8. I also used the which.max/min function to determine which would be the highest and lowest rated cereal from the data set, and it concluded that All Bran with Extra Fiber would be our highest, and

Cap'n'Crunch would be our lowest; as seen in Figure 9. I would say my formulation is consisent with the consumer reports because All Bran with Extra Fiber has the best numbers in terms of what predictors impacted ratings, and Cap'n'Crunch is like a foil to All Bran with Extra Fiber. The only thing Cap'n'Crunch has going for it that puts it slight above All Bran with Extra Fiber is that it has higher carbohydrates, but it is still missing a lot of other nutritionally impactful variables like calories, protein, fat, sodium, fiber, sugars, potassium, and vitamins in which it has the worst values for all of those predictors.

In terms of which brands tend to make a healthier cereal, I have decided to use boxplots to illustrate which manfacturer makes the healthier cereal. I made box plots of all the impactful predictors that was given to me from my best subset selection.

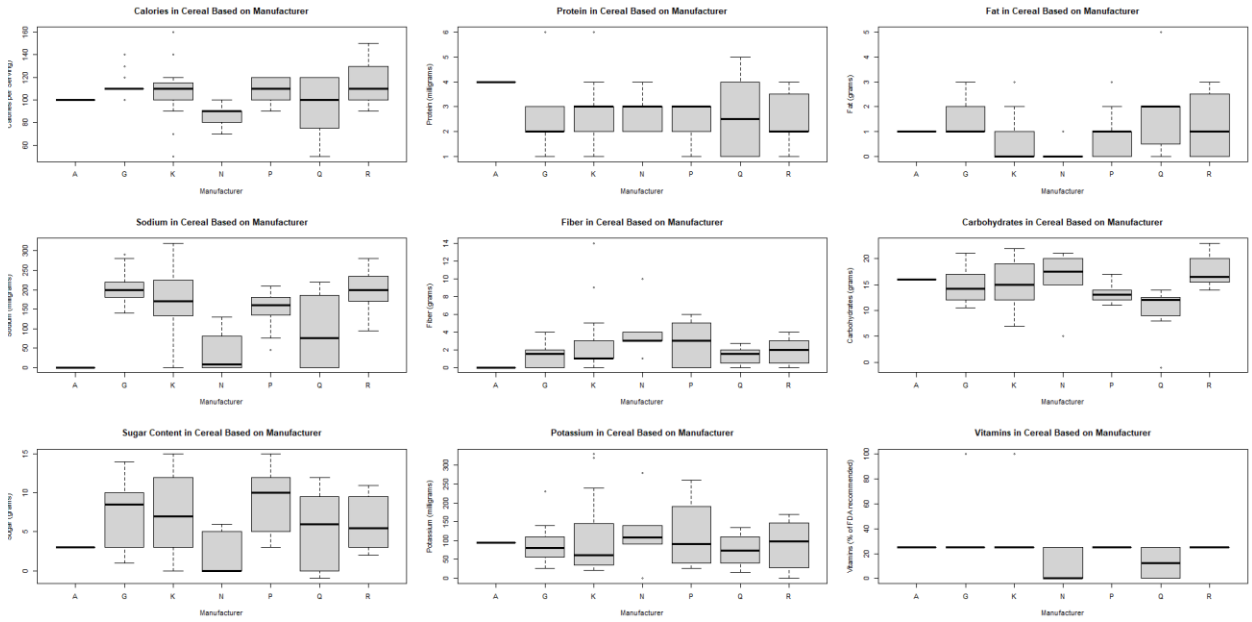

Figure 10: Boxplots of all impactful predictors from best subset selection with manufacturers as our x-axis

From Figure 10, we want to find the average of each predictor and pinpoint who the manufacturer is that made the lowest calorie cereals on average, thus, I will present my findings accordingly. In terms of calories, the lower calories the better, so on average Nabisco produces the lower calorie cereals. For protein, the higher the better, so on average there seems to be a three-way tie of Kellogg's, Nabisco, and Post. An honorable mention for protein would be American Home Food Products which technically has the highest average for protein, but it only had one cereal from them in the data set, so I deemed it as an outlier. For fat, the lower the better, so on average Nabisco and Kellogg's are tied in that category. For sodium, the lower the better, so on average Nabisco produces the lowest sodium cereals. For fiber, the higher the better, so on average Nabisco and Post are tied for highest fiber cereal. For carbohydrates, the higher the better, so on average Nabisco produces the highest carb cereals. For sugar, the lower the better, so on average Nabisco produces the lowest sugar cereals. For potassium, the higher the better, so on average Nabisco produces the highest potassium cereals. Lastly, for vitamins, the higher the better, so on average American Home Food Products, General Mills, Kellogg's, Post, and Ralston Purina are all tied for vitamins. Upon comparing these most impactful nutrition variables, it seems to show that Nabisco tends to make healthier cereals. I don't think there are contradictory points, but it is an interesting point to note that the highest rated cereal in the consumer ratings report is All Bran with Extra Fiber, which is not a Nabisco product, but it is a Kellogg's product. The highest rated cereal for Nabisco is Shredded Wheat'n'Bran, which has a score of 74.47. What this means is that Nabisco as a company is known to make the healthier cereals on average, but Kellogg's produces a variety of types of cereals that range from different scores, but it just so happens that Kellogg's has produced the healthiest cereal in the market according to consumer reports; as shown in Figure 11.

I could make a linear model and make a statement with respect to the roles that continuous variables have such as sugars or calories play on ratings by taking the summary of the linear model from Figure 12. The key thing to look at here in the summary is the p-values. From Figure 12, it shows that both sugars and calories are statistically significant variables to tell the story of ratings, but it seems that sugars are the more statistically significant predictor than calories since it has the lower p-value, but both are still solid predictors from our summary. This all means that having lower calories and sugars will improve consumer ratings. When it comes to making a statement for the categorical variables of the effects that manufacturer and shelf have on ratings, I decided to display a box plot because it is easier to tell a story than a generalized linear model. Thus, from Figure 11 and Figure 13, we can see that Nabisco tends to produce the higher rated cereals according to consumer reports, but Kellogg's has one cereal that has the highest consumer report rating in the market according to this data set. From figure 13, we can see that higher rated cereals according to consumer reports tend to be on the first shelf, while the lower rated cereals, which tend to be the unhealthier ones, happen to be on the second shelf

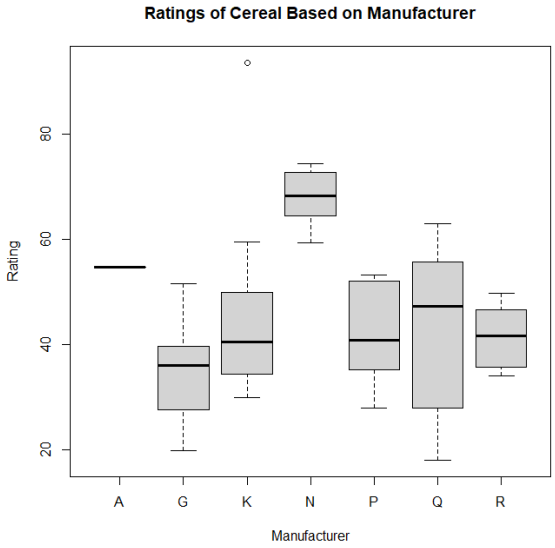

Figure 11: Boxplot of Ratings based on Manufacturer

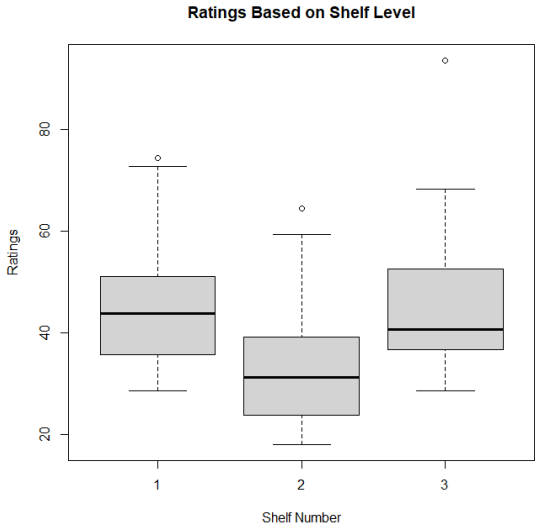

Figure 13: Ratings of Cereal Based on Shelf Level

Addressing the association/relationship between continuous variables other than ratings with each other and their associations with the categorical variables has already been established before when talking about Figure 10. Now I will address the association/relationship between the continuous variables and shelf level, which can be seen below in Figure 14.

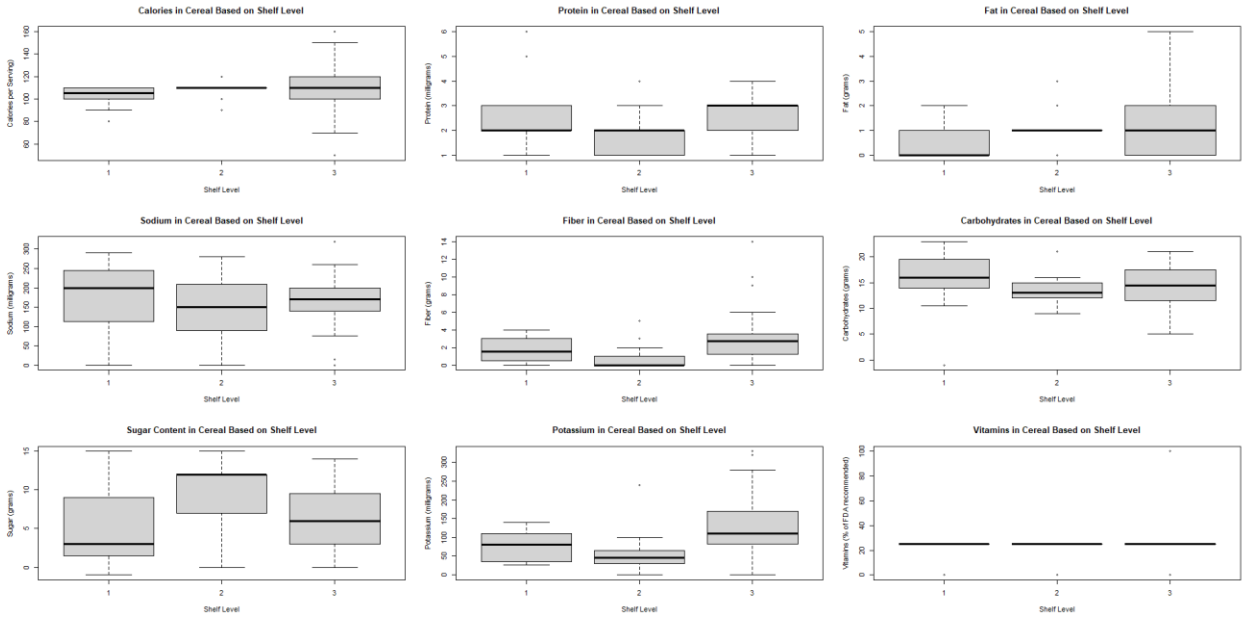

Figure 14: Boxplots of all impactful predictors from best subset selection with shelf level as our x-axis

Therefore, I will now address my findings. The higher calorie cereals on average tend to be on the second shelf. The lower protein cereals on average tend to be on the first and second shelf. The highest fat cereals tend to be on the second shelf. The higher sodium cereals on average tend to be on the first shelf. The least fiber cereals on average tend to be on the second shelf. The least carbohydrate cereals tend to be on the second shelf. The highest sugar cereals on average tend to be on the second shelf. The least potassium cereal on average tends to be on the second shelf.

Lastly, shelf one, two, and three have the same average when it comes to vitamins. From Figure 10 and Figure 14, we can deduce that Nabisco tends to make the healthier cereals, and shelf three tends to have the more nutritional cereals according to the consumer reports. Another thing we can deduce is that shelf two tends to have the least nutritional cereals, and Kellogg's and Post tend to make the least nutritional cereals, which is ironic because Kellogg's has the highest rated cereal according to consumer reports, but they tend to make the least nutritional cereals.

In terms of being able to cluster this data set, we can cluster this data. Clustering this data produces three clusters which I can denote below in Figure 15 and Figure 16.

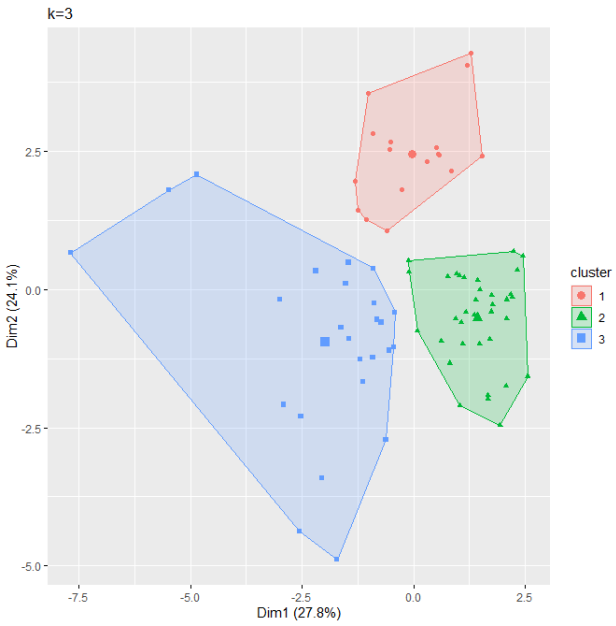

Figure 15: K-means Clustering on Cereal data set

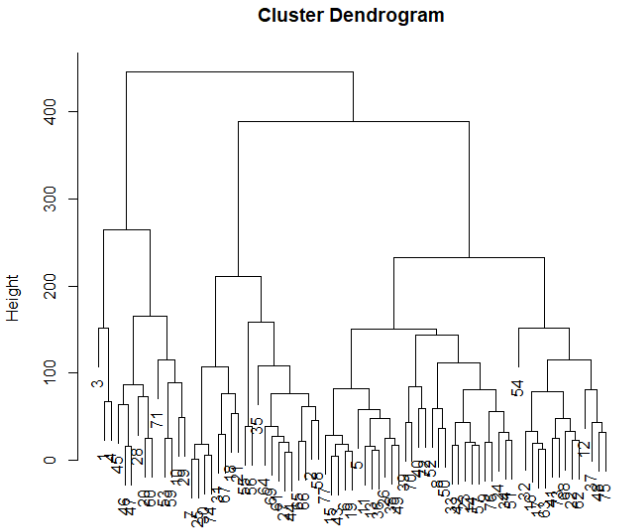

Figure 16: Hierarchical Clustering on Cereal data set

From analyzing the dendrogram, I have made a rough estimation of what each cluster would target which demographic. For cluster 3, I know for a fact that this cluster would target health-conscious individuals because it has the highest rated cereal according to consumer reports. For cluster 2, I think this cluster would target kids because this cluster has Lucky Charms, which would mean this would be the cluster for sugary cereals. Lastly, cluster 1 would target adults just looking for cereal to eat that are not really health-conscious and not a kid either, so these three clusters target the people who would buy those cereals.

I definitely saw outliers in my analysis. For example, if we take a look at Figure 10 above, we can see that there were some cereals that had a high protein level outlier such as Kellogg's and General Mills, which could possibly inflate the manufacturer's overall rating since protein is one of the statistically significant predictors for consumer rating reports. The same could be said about the reverse where Kellogg's would have a couple cereals that have high calories that could adversely affect their consumer ratings. Variables of interest would be the variables that are the best to predict ratings, which are calories, protein, fats, sodium, fiber, carbohydrates, sugar, potassium, and vitamins. Either one of those dots in the boxplots from Figure 10 can positively or negatively affect their ratings. Having higher calories would negatively affect their ratings, but lower calories would increase it. Having high protein would increase their rating but having low protein would not, and the list goes on and on. An obvious example of how outliers can affect ratings is by looking at Figure 11 in the Appendix, we can see that generally, Nabisco produces higher rated cereals, but Kellogg's has one cereal that is the best rated cereal, so that would inflate the ratings of Kellogg's more than it should.

Appendix

```
> glm.pred=rep("Movie",7740)
> glm.pred[glm.probs>.5]="Series"
> table(glm.pred,Series.or.Movie)
         Series.or.Movie
glm.pred Movie Series
  Movie   5631   1651
  Series   309    798
> mean(glm.pred==Series.or.Movie)
```

Figure 1: Logistic Regression Classification Rate calculated by hand from table. Classification rate of .7663

```
> table(lda.class,Series.or.Movie)
         Series.or.Movie
lda.class Movie Series
   Movie  11493   3062
   Series   204    721
> mean(lda.class==Series.or.Movie)
[1] 0.7890181
```

Figure 2: LDA Method for Classification Rate

```
> table(qda.class,Series.or.Movie)
         Series.or.Movie
qda.class Movie Series
   Movie   5826    240
   Series  5871   3543
> mean(qda.class==Series.or.Movie)
[1] 0.6052326
```

Figure 3: QDA Method for Classification Rate

```
> lasso.pred=predict(mod.lasso,newx=test.mat,s=lambda.best)
> mean((data.test[,"Hidden.Gem.Score"]-lasso.pred)^2)
[1] 3.462791
```

Figure 4: MSE of Lasso Regression

Kyle Bermudez

Appendix

```
> ridge.pred=predict(mod.ridge,newx=test.mat,s=lambda.best)
> mean((data.test[,"Hidden.Gem.Score"]-ridge.pred)^2)
[1] 3.461274
```

Figure 5: MSE of Ridge Regression

```
> reg.summary$rsq
[1] 0.1983747 0.2816011 0.3357648 0.3425814 0.3444421 0.3444607
> which.max(reg.summary$adjr2)
[1] 5
> which.min(reg.summary$cp)
[1] 5
> which.min(reg.summary$bic)
[1] 5
> coef(regfit.full2,5)
          (Intercept)            IMDb.Score Rotten.Tomatoes.Score
         2.087611e+02          1.713659e+02          7.571615e+00
 Awards.Nominated.For            IMDb.Votes      Hidden.Gem.Score
        -2.383270e+00          1.018269e-03         -1.715905e+02
```

Figure 6: Best Subset Selection for box office

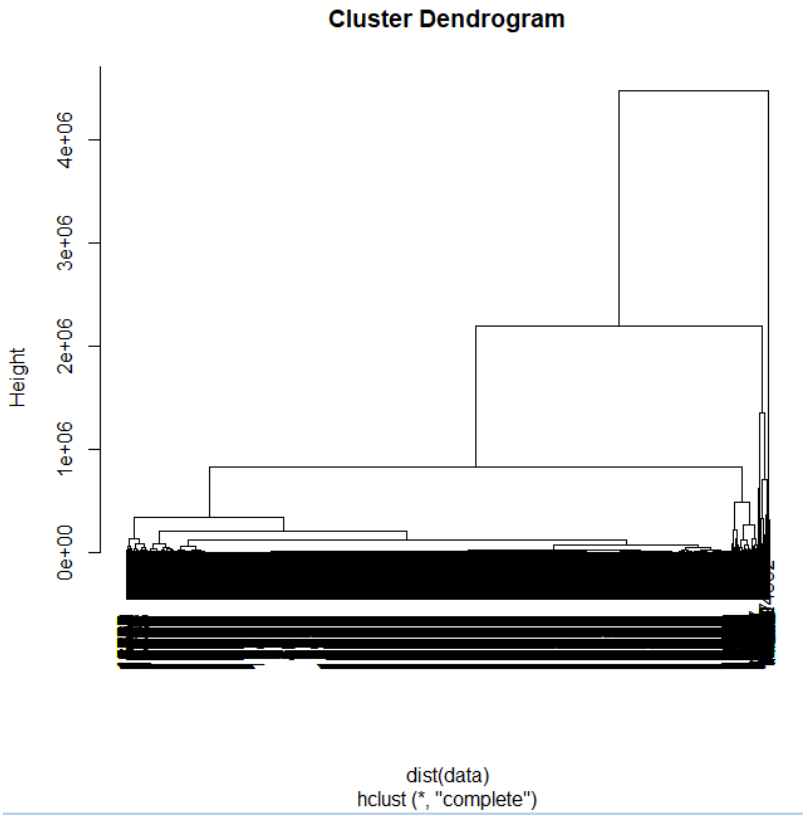

Figure 7: Dendrogram of the data

Appendix

```
> reg.summary$rsq
 [1] 0.5771056 0.8091568 0.8984590 0.9604887 0.9735251 0.9831029 0.9908565
 [8] 0.9964295 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
> which.max(reg.summary$adjr2)
[1] 9
> which.min(reg.summary$cp)
[1] 9
> which.min(reg.summary$bic)
[1] 9
> coef(regfit.full,9)
(Intercept)    calories     protein         fat      sodium       fiber
54.92718420 -0.22272417  3.27317387 -1.69140800 -0.05449270  3.44347977
      carbo      sugars      potass    vitamins
 1.09245094 -0.72489514 -0.03399335 -0.05121197
```

Figure 8: Best Subset Selection with rating as the response for the Cereal Data Set

```
> data2[which.max(rating),"name"]
[1] "All-Bran with Extra Fiber"
> data2[which.min(rating),"name"]
[1] "Cap'n'Crunch"
```

Figure 9: Highest and lowest rated cereal given the best subset selection

Kyle Bermudez

Appendix

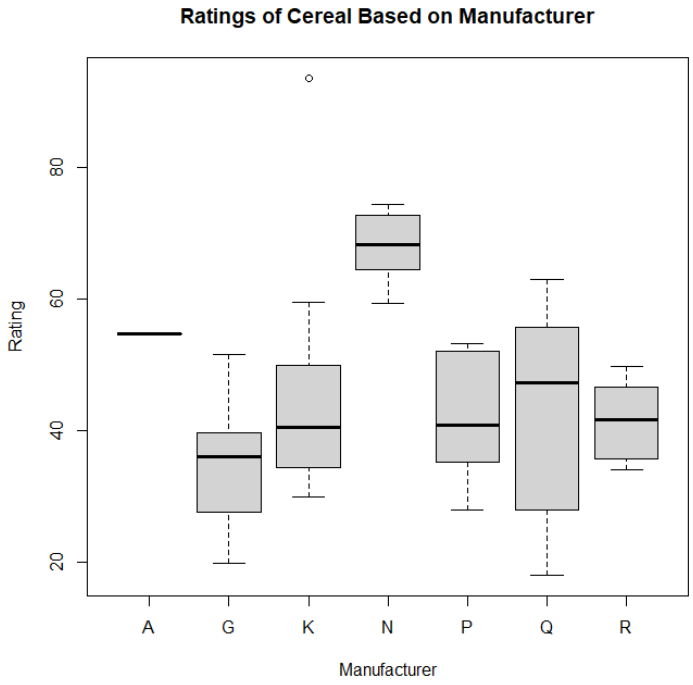

Figure 11: Boxplot of Ratings based on Manufacturer

```
Call:
lm(formula = rating ~ sugars + calories, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-15.643  -6.339  -1.221   4.823  23.413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.11417    5.44513  15.448  < 2e-16 ***
sugars      -1.71939    0.25225  -6.816 2.16e-09 ***
calories    -0.27644    0.05755  -4.804 7.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.083 on 74 degrees of freedom
Multiple R-squared:  0.6776,    Adjusted R-squared:  0.6689
F-statistic: 77.78 on 2 and 74 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of linear model with rating as the response and sugars and calories as the predictors

Appendix

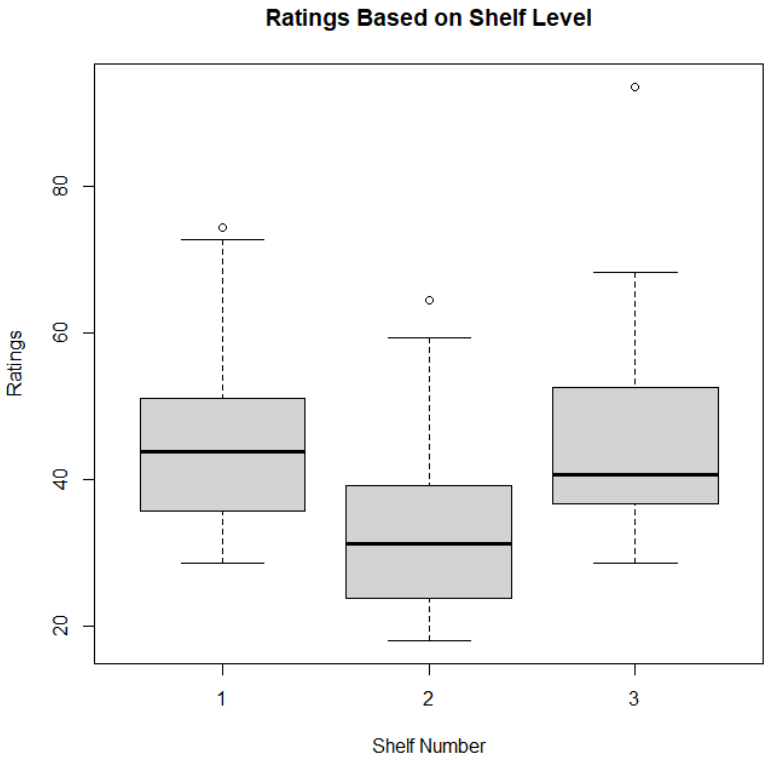

Figure 13: Ratings of Cereal Based on Shelf Level