Breakfast Cereals: Data Analysis and Discussion

Gabriel Jones

Mercer University

CSC 485: Datascience Applications

Professor Martin Zhao

May 3, 2019

Abstract

This data analysis report explores the various relationships and ideas that can be concluded from using statistical methods and packages from the R programming language, concerning a dataset of seventy-seven different types of cereals. The most notable of findings is that healthier cereals are *generally* disposed to receiving better ratings. Cereals high in sugar and calories score very low compared to healthier options. The manufacturer Kellogg's was found to have the most cereals above the median calorie count. By using the linear regression modeling capabilities of R, it is possible to predict the rating a cereal will get based on nutritional factors, will little error. It would be very useful for manufacturers to be able to receive higher ratings on their cereals while at the same time offering healthier options for consumers.

Introduction

This data analysis report explores the various relationships and ideas that can be concluded from using statistical methods and packages from the R programming language, concerning a dataset of seventy-seven different types of cereals.

The data was found on the Mercer Blackhawk server, and contains twelve different data fields, as follows:

QUANTITATIVE DATA	NOMINAL DATA
NAME	Calories
MANUFACTURER	Protein
TYPE	Fat
	Sodium
	Fiber
	Carbohydrates
	Sugars
	Potassium
	Vitamins
	Shelf
	Weight
	Cups
	Rating*

There are seventy-four cold cereals and three hot cereals.

According to a FoodDive report[1] in 2017, nine out of ten consumers eat cereal for breakfast. Unhealthy dietary habits are some of the most discussed problems concerning teens and children. Therefore, it is important to know the relationships between things like calories, sugar, vitamins, and how these factors compare to the rating of a cereal. After reading this report, it will be clear where the ratings are coming from and how they relate to the nutritional value of the cereal.

Asking the right questions determines the effectiveness of the data analysis. Some of the questions (concerning cereal data) that will be answered in this report are:

- Is there a relation between sugars, calories, carbs, and fat?
- How are calories and potassium distributed?
- Which manufacturers produce cereal with highest calories?
- How does rating compare to calorie count?
- Which nutrients are essential for a good rating for a cereal?
- Is there a relation between manufacturer and rating?

[1] https://www.fooddive.com/news/9-out-of-10-consumers-eat-cereal-for-breakfast-but-just-under-half-like-it/507552/

- Is there a relation between shelf number and rating?
- Can we use machine learning models to predict the rating of a cereal based on its nutritional values?

Methods

R is a robust platform for data manipulation that comes with many tools to visualize, summarize, and compute data.

The tools that were utilized in this analysis include:

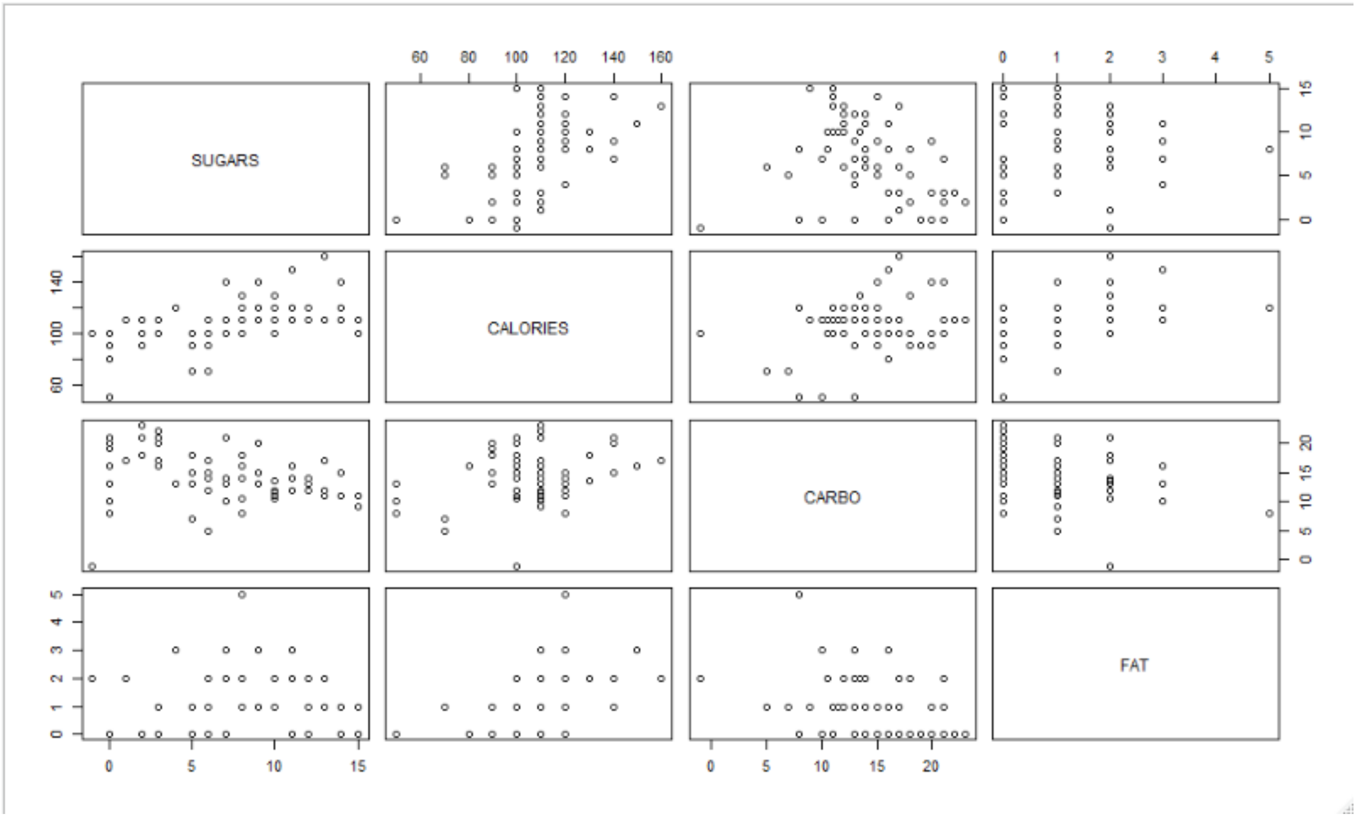
Tool	*Purpose*	*R function*
Scatter-plot Matrix	Creates a correlation matrix plot between parameters in a dataframe.	pairs()
Pie Chart	Creates a pie chart where data is divided into “slices” to illustrate proportions.	
Histogram	Creates a histogram, which is a plot to show the frequency distribution of a variable in a dataframe.	hist()
Bar Chart	Creates a bar chart that represents categorical data with bars with heights proportional to the value they represent.	barplot()
Box-and-whisker Plot	Creates a convenient illustration of the quartiles of a dataset, which is helpful for understanding the spread.	boxplot()
Scatterplot	Displays values for two variables of data on a Cartesian plane, which is helpful for understanding relationships between variables.	plot()
3D Scatterplot	A type of scatterplot that shows the relationship between three variables.	scatterplot3d()
Linear Regression Model	Fits a linear equation to the relationship between two variables. Very helpful in making predictions about future data.	lm()

Results

**Assume set.seed(2048)*

By applying the tools listed in the previous section to the questions that the analysis aimed to answer, the following results were reached.

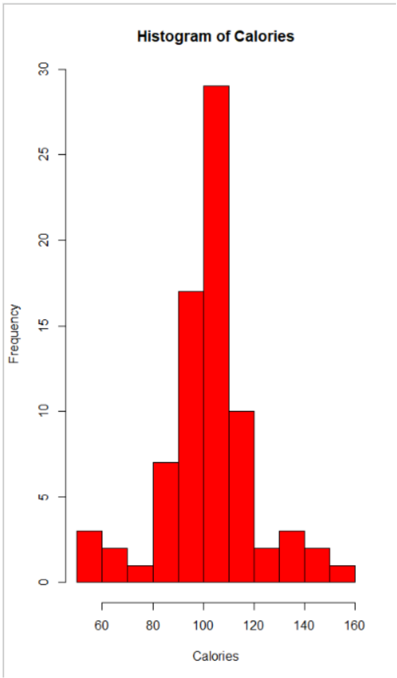
- Is there a relation between sugars, calories, carbs, and fat?

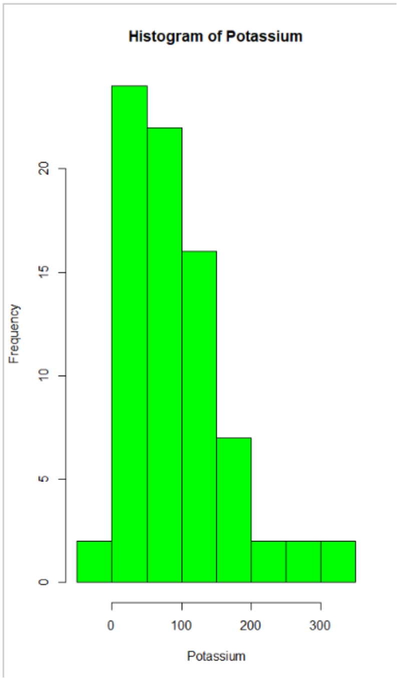


(Refer to script 1)

From the scatter-plot matrix shown above, there is a sharp positive correlation between the number of calories and the number of sugars a cereal has. Also, as the number of sugars rises, the carbohydrate level experiences a decrease, which is peculiar.

- How are calories and potassium distributed?

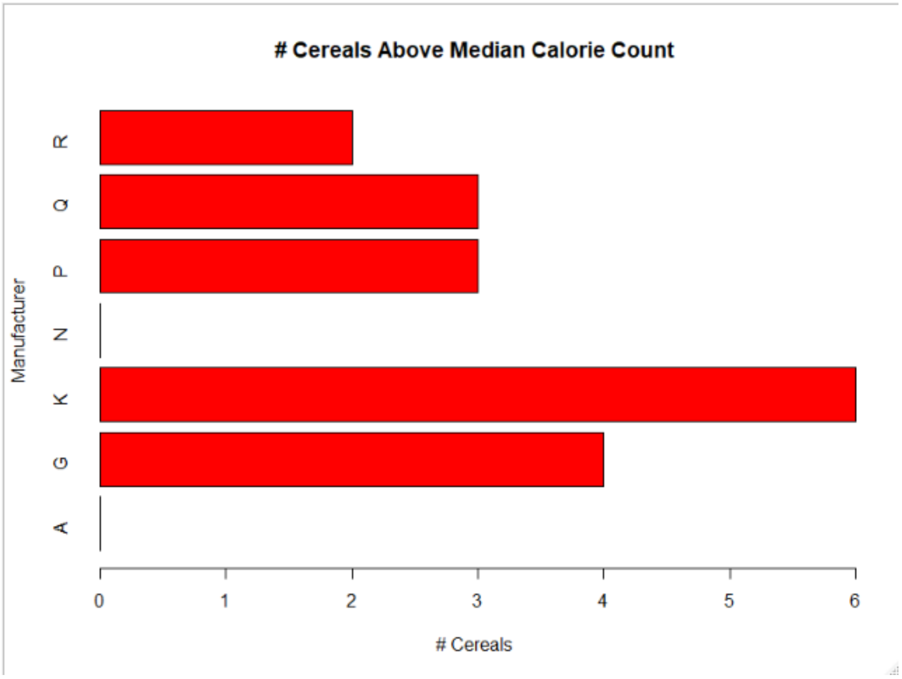




(Refer to script 2)

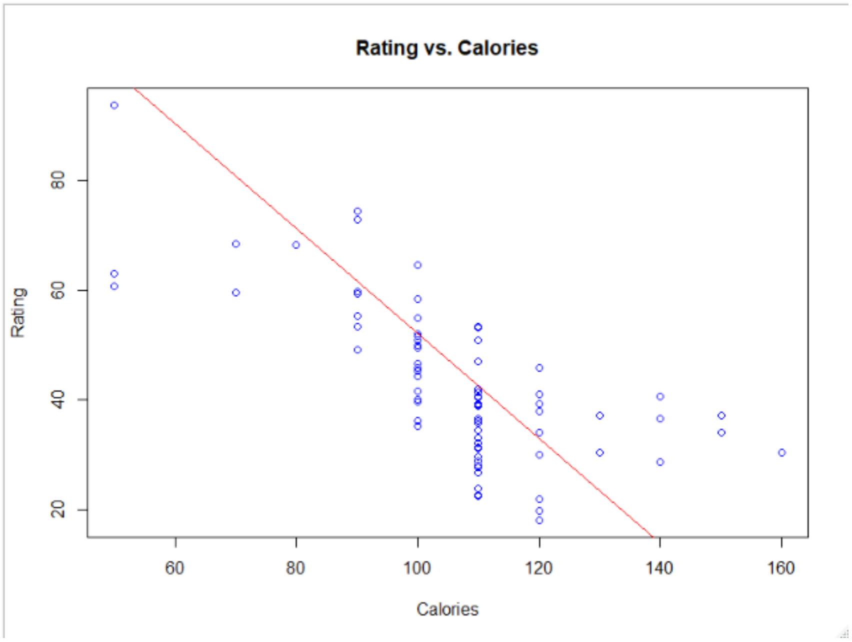
From these histograms, we can see that there is less variance in the number of calories that cereals have compared to the variance of potassium levels. Also, the potassium frequencies are right-skewed, so most cereals do not have high levels of potassium.

- Which manufacturers produce cereal with highest calories?



(Refer to script 3)
From this bar graph, it becomes known that Kellogg's brand cereals *generally* have the most cereals where the calorie count is above the median.

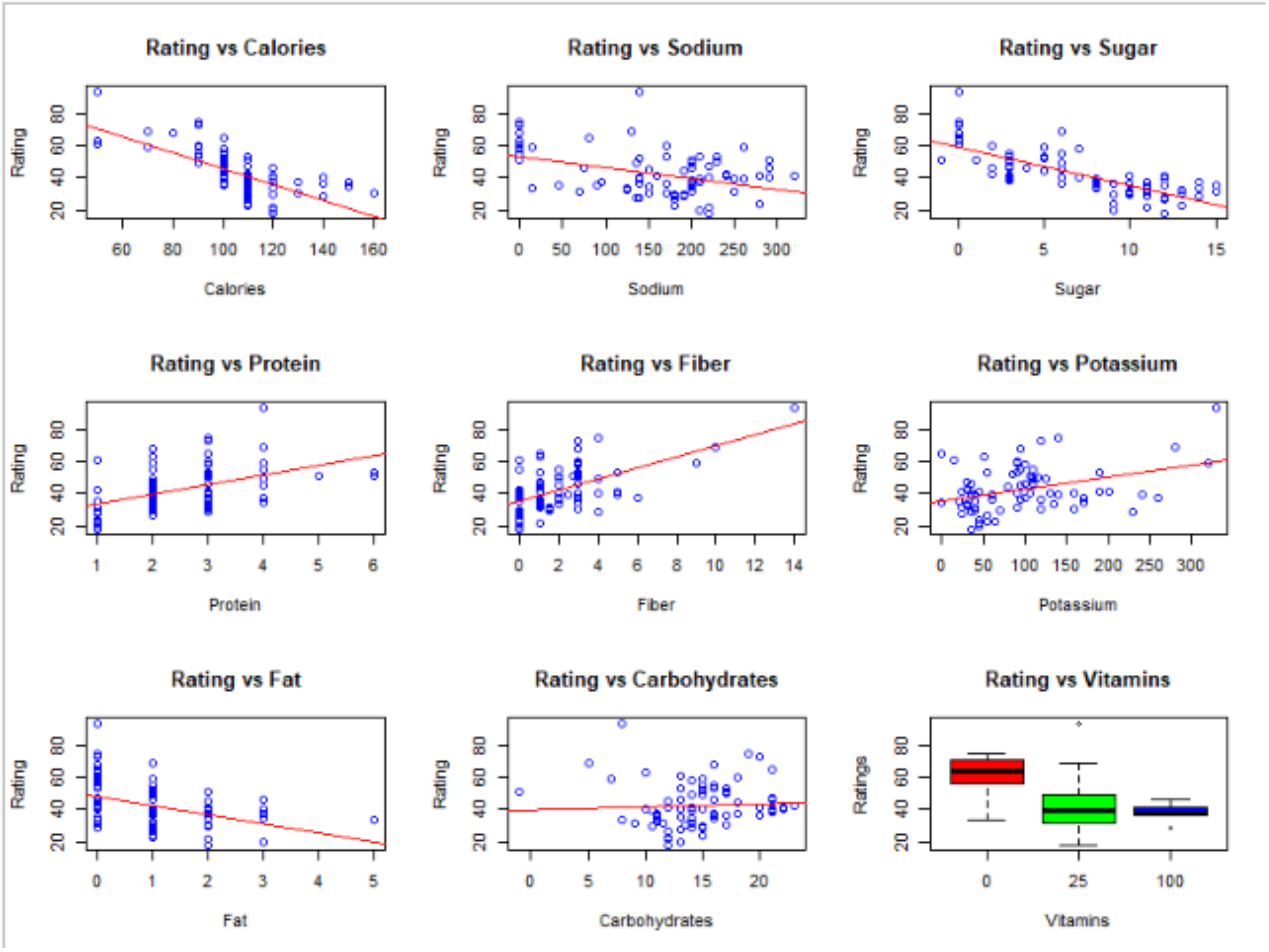
- How does rating compare to calorie count?



(Refer to script 4)

This scatterplot (which includes a line of regression) shows that as the number of calories goes up, the rating generally goes down. This is good to know because high-calorie cereals should not be receiving inflated ratings.

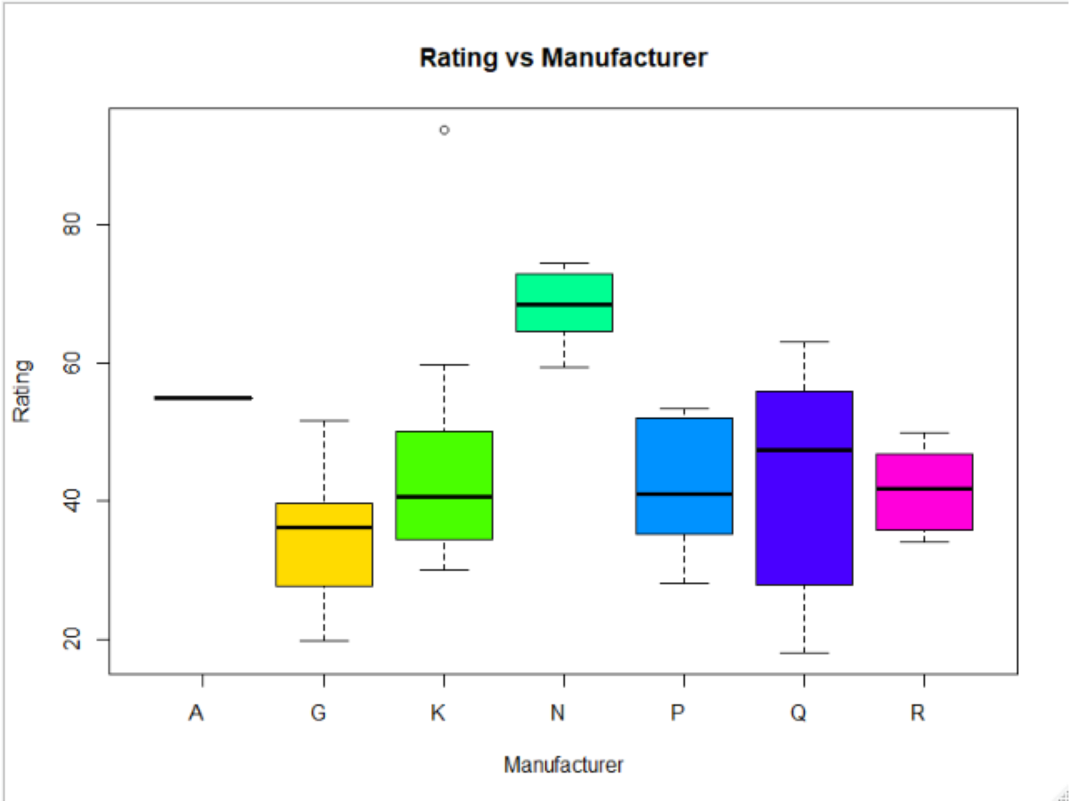
- Which nutrients are essential for a good rating for a cereal?



(Refer to script 5)

This collection of scatterplots is extremely useful in determining how certain nutritional factors determine the end rating of a cereal. We can see that there is a positive correlation between rating and protein, fiber, and potassium. On the other hand, there is a negative correlation between rating and calories, sodium, sugar, and fat. Healthy cereals get better ratings!

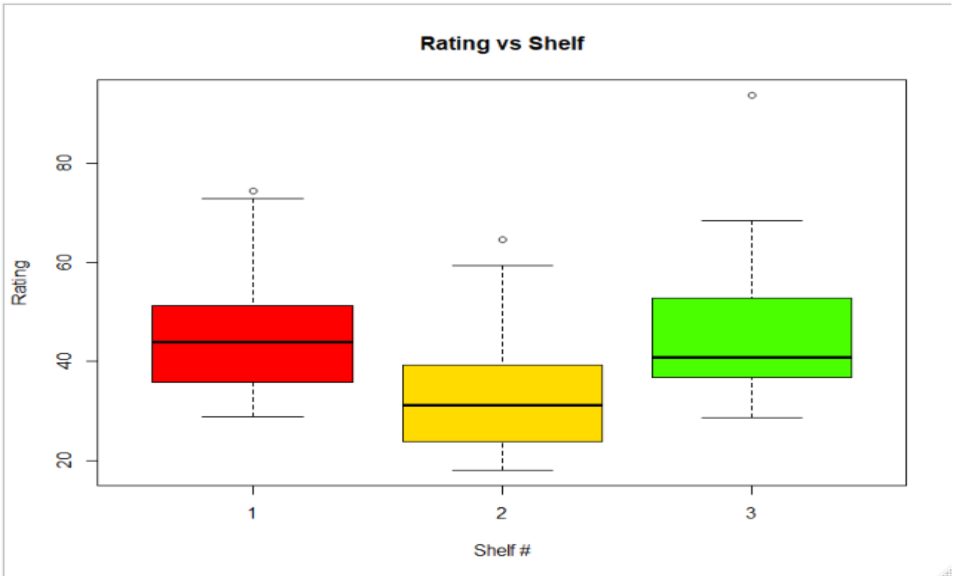
- Is there a relation between manufacturer and rating?



(Refer to script 6)

Almost all of the manufactures align similarly when it comes to rating, aside from the "N" manufacturer. This suggests that this manufacturer is probably focusing on creating healthier cereals which results in less low ratings and more high ratings.

- Is there a relation between shelf number and rating?



(Refer to script 7)
Judging from the box-and-whisker plot, a definite relationship cannot be established between rating and shelf number. **Although, it does appear that shelves one and three will generally contain the highest-rated (and likely healthiest) cereals.**

- Can we use machine learning models to predict the rating of a cereal based on its nutritional values?

(Refer to script 8)

Targeting rating with reference to sugars, proteins, and calories, using the linear regression model we can accurately predict a cereal's rating just be knowing the nutritional data.

The linear regression model returns the following equation:

$$76.73 - 1.15 * SUGARS - 0.35 * CALORIES + 4.62 * PROTEIN$$

By removing records from the dataset and testing their nutritional values against the equation, an average error percentage of 7.5% was calculated when testing using training data on an 80/20 split.

(Refer to script 9 for model evaluation)

Conclusions

- There is a positive correlation between calories and sugars in cereal.
- Most cereals do not have relatively high potassium values.
- Kellogg's offers the most cereals out of any manufacturer that are above the median calorie count (110).
- The more calories that a cereal has, the less likely it is to receive a high rating.
- Manufacturers that want to bring in high ratings should create cereals that are high in fiber, protein, and potassium and avoid creating cereals with high calorie counts or lots of sugar or fat.
- Cereals with high ratings are more likely to be placed on the first or third shelf, because that is generally where the consumers' eyes gravitate.
- Using a linear regression model can allow for accurate predictions of future cereal with less than ten percent error on average.
    - For instance, a cereal that has thirteen grams of sugars, one-hundred and ten calories, and two grams of protein is projected to receive a rating of 32.03.

Appendix

Script 1

```
pairs(~SUGARS+CALORIES+CARBO+FAT,data=cereals_dat)
```

Script 2

```
calorie_histogram <- hist(cereals_dat$CALORIES, breaks=10, col="red",
xlab = "Calories", main = "Histogram of Calories")
potassium_histogram <- hist(cereals_dat$POTASS, breaks = 10, col =
"green", xlab = "Potassium", main = "Histogram of Potassium")
```

Script 3

```
cereals_dat_calorie_select <- subset(cereals_dat, cereals_dat[4]>110)
#110 is median
bar_labels <- table(cereals_dat_calorie_select$MANUF)
barplot(bar_labels, col = "red", horiz = TRUE, main = "# Cereals Above
Median Calorie Count", ylab = "Manufacturer", xlab = "# Cereals")
```

Script 4

```
plot(cereals_dat$CALORIES, cereals_dat$RATING, xlab = "Calories", ylab
= "Rating", main = "Rating vs. Calories", col = "blue")
abline(lm(cereals_dat$CALORIES~cereals_dat$RATING), col = "red")
```

Script 5

```
par(mfcol=c(3,3))
#RATING vs CALORIES
plot(cereals_dat$RATING~CALORIES,
     data = cereals_dat,
     xlab="Calories",
     ylab="Rating",
     main="Rating vs Calories",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$CALORIES), col="red")

#RATING vs PROTEIN
plot(cereals_dat$RATING~PROTEIN,
     data = cereals_dat,
     xlab="Protein",
     ylab="Rating",
     main="Rating vs Protein",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$PROTEIN), col="red")
```

```
#RATING vs FAT
plot(cereals_dat$RATING~FAT,
     data = cereals_dat,
     xlab="Fat",
     ylab="Rating",
     main="Rating vs Fat",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$FAT), col="red")

#RATING vs SODIUM
plot(cereals_dat$RATING~SODIUM,
     data = cereals_dat,
     xlab="Sodium",
     ylab="Rating",
     main="Rating vs Sodium",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$SODIUM), col="red")

#RATING vs FIBER

plot(cereals_dat$RATING~FIBER,
     data = cereals_dat,
     xlab="Fiber",
     ylab="Rating",
     main="Rating vs Fiber",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$FIBER), col="red")

#RATING vs CARBO

plot(cereals_dat$RATING~CARBO,
     data = cereals_dat,
     xlab="Carbohydrates",
     ylab="Rating",
     main="Rating vs Carbohydrates",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$CARBO), col="red")

#RATING vs SUGARS

plot(cereals_dat$RATING~SUGARS,
     data = cereals_dat,
     xlab="Sugar",
     ylab="Rating",
     main="Rating vs Sugar",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$SUGARS), col="red")

#RATING vs POTASS

plot(cereals_dat$RATING~POTASS,
     data = cereals_dat,
     xlab="Potassium",
     ylab="Rating",
     main="Rating vs Potassium",
     col="blue")

abline(lm(cereals_dat$RATING~cereals_dat$POTASS), col="red")


#RATING vs VITAMINS

boxplot(cereals_dat$RATING~VITAMINS,
        data = cereals_dat,
        xlab="Vitamins",
        ylab="Ratings",
        main="Rating vs Vitamins",
        col=c("red","green","blue"))
```

Script 6

```
boxplot(cereals_dat$RATING~cereals_dat$MANUF, data = cereals_dat, xlab
= "Manufacturer", ylab = "Rating", main = "Rating vs Manufacturer",
col = rainbow(7))
```

Script 7

```
boxplot(cereals_dat$RATING~cereals_dat$SHELF, data = cereals_dat, xlab
= "Shelf #", ylab = "Rating", main = "Rating vs Shelf", col =
rainbow(7))
```

Script 8

```
fit <- lm(RATING ~ SUGARS + CALORIES + PROTEIN, data = cereals_dat)
summary(fit)

predict(fit, data.frame(SUGARS = 13, CALORIES = 110, PROTEIN = 2))
```

Script 9

```
trainingRowIndex <- sample(1:nrow(cereals), 0.8*nrow(cereals))
trainingData <- cereals[trainingRowIndex, ]
testData <- cereals[-trainingRowIndex, ]
```

```
pred <- predict(fit, testData)
actual_preds <- data.frame(cbind(actuals=testData$RATING,
predicteds=pred))
correlation_accuracy <- cor(actual_preds)

#returns the table below
ID   Actual     Predicted
1    68.40297   63.49662
2    33.98368   38.8779
10   53.31381   52.95716
11   18.04285   25.02857
15   22.73645   27.41311
17   45.86332   48.26507
18   35.78279   28.56738
21   64.53382   55.18973
33   52.0769    49.41835
36   21.87129   26.18285
37   31.07222   40.10816
46   34.13976   29.41475
47   30.31335   18.95128
65   74.47295   58.72854
66   72.80179   58.72854
77   36.18756   37.8006

#accuracy
           actuals   predicted
 actuals           1 0.9250653
 predicted 0.9250653         1
```

Using the following equation, we can calculate the average of the minimum and maximum accuracy:

$$MinMaxAccuracy = mean\left(\frac{min\left(actuals, predicteds\right)}{max\left(actuals, predicteds\right)}\right)$$

The equation returns 0.8467601, which represents a fair error ratio.

Final Script

```
set.seed(2048)
cereals <- read.delim("~/Desktop/cereals_dat.txt")

#view summary data#
summary(cereals)

#create histograms for POTASS and CALORIES#
calorie_histogram <- hist(cereals_dat$CALORIES, breaks=10, col="red",
xlab = "Calories", main = "Histogram of Calories")
potassium_histogram <- hist(cereals_dat$POTASS, breaks = 10, col =
"green", xlab = "Potassium", main = "Histogram of Potassium")

#create barplot showing manufacturers with high calorie cereals#
cereals_calorie_selection <- subset(cereals, cereals[4]>110)
bar_labels <- table(cereals_calorie_selection$MANUF)
barplot(bar_labels, col = "red", horiz = TRUE, main = "# Cereals Above
Median Calorie Count", ylab = "Manufacturer", xlab = "# Cereals")

#create scatterplot matrix for CARBO, CALORIES, SUGARS#
pairs(~SUGARS+CALORIES+CARBO+FAT,data=cereals)

#create RATING vs CALORIE scatterplot#
plot(cereals$CALORIES, cereals$RATING, xlab = "Calories", ylab =
"Rating", main = "Rating vs. Calories", col = "blue")
#linear regresson line added to plot#
abline(lm(cereals$CALORIES~cereals$RATING), col = "red")


#create multiple scatterplot display#
par(mfcol=c(3,3))
#RATING vs CALORIES
plot(cereals_dat$RATING~CALORIES,
     data = cereals_dat,
     xlab="Calories",
     ylab="Rating",
     main="Rating vs Calories",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$CALORIES), col="red")

#RATING vs PROTEIN
plot(cereals_dat$RATING~PROTEIN,
     data = cereals_dat,
     xlab="Protein",
     ylab="Rating",
     main="Rating vs Protein",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$PROTEIN), col="red")

#RATING vs FAT#
plot(cereals_dat$RATING~FAT,
     data = cereals_dat,
     xlab="Fat",
     ylab="Rating",
     main="Rating vs Fat",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$FAT), col="red")

#RATING vs SODIUM#
plot(cereals_dat$RATING~SODIUM,
     data = cereals_dat,
     xlab="Sodium",
     ylab="Rating",
     main="Rating vs Sodium",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$SODIUM), col="red")

#RATING vs FIBER#

plot(cereals_dat$RATING~FIBER,
     data = cereals_dat,
     xlab="Fiber",
     ylab="Rating",
     main="Rating vs Fiber",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$FIBER), col="red")

#RATING vs CARBO#

plot(cereals_dat$RATING~CARBO,
     data = cereals_dat,
     xlab="Carbohydrates",
     ylab="Rating",
     main="Rating vs Carbohydrates",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$CARBO), col="red")

#RATING vs SUGARS#

plot(cereals_dat$RATING~SUGARS,
     data = cereals_dat,
     xlab="Sugar",
     ylab="Rating",
     main="Rating vs Sugar",
     col="blue")
abline(lm(cereals_dat$RATING~cereals_dat$SUGARS), col="red")

#RATING vs POTASS#

plot(cereals_dat$RATING~POTASS,
     data = cereals_dat,
     xlab="Potassium",
     ylab="Rating",
     main="Rating vs Potassium",
     col="blue")

abline(lm(cereals_dat$RATING~cereals_dat$POTASS), col="red")


#RATING vs VITAMINS#

boxplot(cereals_dat$RATING~VITAMINS,
        data = cereals_dat,
        xlab="Vitamins",
        ylab="Ratings",
        main="Rating vs Vitamins",
        col=c("red","green","blue"))


#create grouped box-and-whisker plots based on MANUF#
boxplot(cereals_dat$RATING~cereals_dat$MANUF, data = cereals_dat, xlab
= "Manufacturer", ylab = "Rating", main = "Rating vs Manufacturer",
col = rainbow(7))

#create grouped box-and-whisker plots based on SHELF#
boxplot(cereals_dat$RATING~cereals_dat$SHELF, data = cereals_dat, xlab
= "Shelf #", ylab = "Rating", main = "Rating vs Shelf", col =
rainbow(7))

#create linear regression model#
fit <- lm(RATING ~ SUGARS + CALORIES + PROTEIN, data = cereals_dat)
summary(fit)

#test case prediction#
predict(fit, data.frame(SUGARS = 13, CALORIES = 110, PROTEIN = 2))
```

```
#testing model with training data#
trainingRowIndex <- sample(1:nrow(cereals), 0.8*nrow(cereals))
trainingData <- cereals[trainingRowIndex, ]
testData <- cereals[-trainingRowIndex, ]
pred <- predict(fit, testData)
actuals_preds <- data.frame(cbind(actuals=testData$RATING,
predicteds=pred))
correlation_accuracy <- cor(actual_preds)
min_max_accuracy <- mean(apply(actuals_preds, 1, min) /
apply(actuals_preds, 1, max))
```