Project Part 1

Aaradhana “Anya” Bharill

December 8, 2017

## Breakfast Cereal Project Part 1

library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1 v purrr 0.2.4  
## v tibble 1.3.4 v dplyr 0.7.4  
## v tidyr 0.7.2 v stringr 1.2.0  
## v readr 1.1.1 v forcats 0.2.0

## -- Conflicts ------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

cereal <- read\_csv("C:/Users/bharila/Desktop/Winter 2017/MA485-Heyman/Project/HealthyBreakfast.csv")

## Parsed with column specification:  
## cols(  
## name = col\_character(),  
## mfr = col\_character(),  
## type = col\_character(),  
## calories = col\_integer(),  
## protein = col\_integer(),  
## fat = col\_integer(),  
## sodium = col\_integer(),  
## fiber = col\_double(),  
## carbo = col\_double(),  
## sugars = col\_integer(),  
## potass = col\_integer(),  
## vitamins = col\_integer(),  
## shelf = col\_integer(),  
## weight = col\_double(),  
## cups = col\_double(),  
## rating = col\_double()  
## )

I am working on the breakfast cereal dataset. It has 77 different breakfast cereals (i.e. 77 observations), with 15 different variables observed for each observation.

While the site containing the data doesn’t provide any information about how it was collected (except for the variable meanings), all the manufacturers listed in the manufacturer category variable are American companies so clearly American company breakfast cereal data was collected.

The data for all the variables (except for rating) in this dataset, can be collected from the cereal box itself. The fact that there is a shelf variable (counting the shelf number from the floor) means that this data was probably collected from a supermarket. While the shelf variables looks numerical it is actually categorical. The vitamin variable also looks numerical but is categorical as the codebook says that it can have only three values: 0, 25, or 100.

It is unclear where the data for rating was collected. This could be from customer feedback obtained in that supermarket or from somewhere online.

dim(cereal) #gives the dimensions (rows, columns) of the dataset

## [1] 77 16

cereal$vitamins <- as.factor(cereal$vitamins)  
cereal$shelf <- as.factor(cereal$shelf)

First a summary of the cereal dataset was obtained to get a generaloverview of the data. The carbo, sugars, and potass variables have a minimum value of -1 meaning that there is missing data in tha table. On further looking at the table, there are three observations with missing data. All three cereals are from different manufacturers but 2 are of the type, hot and 1 is of the type, cold. There are only 3 hot cereals in the entire dataset.

summary(cereal)

## name mfr type calories   
## Length:77 Length:77 Length:77 Min. : 50.0   
## Class :character Class :character Class :character 1st Qu.:100.0   
## Mode :character Mode :character Mode :character Median :110.0   
## Mean :106.9   
## 3rd Qu.:110.0   
## Max. :160.0   
## protein fat sodium fiber   
## Min. :1.000 Min. :0.000 Min. : 0.0 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:0.000 1st Qu.:130.0 1st Qu.: 1.000   
## Median :3.000 Median :1.000 Median :180.0 Median : 2.000   
## Mean :2.545 Mean :1.013 Mean :159.7 Mean : 2.152   
## 3rd Qu.:3.000 3rd Qu.:2.000 3rd Qu.:210.0 3rd Qu.: 3.000   
## Max. :6.000 Max. :5.000 Max. :320.0 Max. :14.000   
## carbo sugars potass vitamins shelf   
## Min. :-1.0 Min. :-1.000 Min. : -1.00 0 : 8 1:20   
## 1st Qu.:12.0 1st Qu.: 3.000 1st Qu.: 40.00 25 :63 2:21   
## Median :14.0 Median : 7.000 Median : 90.00 100: 6 3:36   
## Mean :14.6 Mean : 6.922 Mean : 96.08   
## 3rd Qu.:17.0 3rd Qu.:11.000 3rd Qu.:120.00   
## Max. :23.0 Max. :15.000 Max. :330.00   
## weight cups rating   
## Min. :0.50 Min. :0.250 Min. :18.04   
## 1st Qu.:1.00 1st Qu.:0.670 1st Qu.:33.17   
## Median :1.00 Median :0.750 Median :40.40   
## Mean :1.03 Mean :0.821 Mean :42.67   
## 3rd Qu.:1.00 3rd Qu.:1.000 3rd Qu.:50.83   
## Max. :1.50 Max. :1.500 Max. :93.70

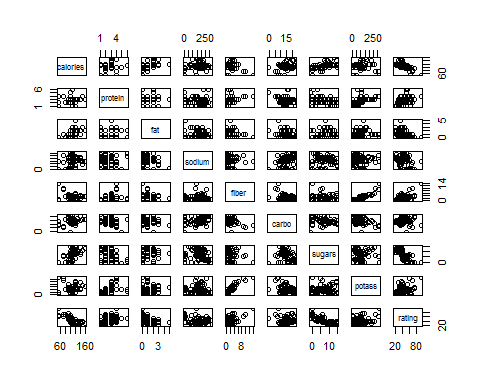
Apart from looking at the summary, I also noticed that the two companies in the dataset manufacture many different cereals whereas the others manufacture a much lower proportion.

w = table(cereal$mfr)  
t = as.data.frame(w)  
t

## Var1 Freq  
## 1 A 1  
## 2 G 22  
## 3 K 23  
## 4 N 6  
## 5 P 9  
## 6 Q 8  
## 7 R 8

Finally I plotted pairwise plots of the main nutrients and rating for each observation. I left out the shelf, vitamin and weight variables because I thought that the vitamin varible was too vague, the shelf variable didn’t have much to do with nutrition and weight is like a sum of all the other nutrients in a way.

pairs(~calories+protein+fat+sodium+fiber+carbo+sugars+potass+rating, data=cereal)

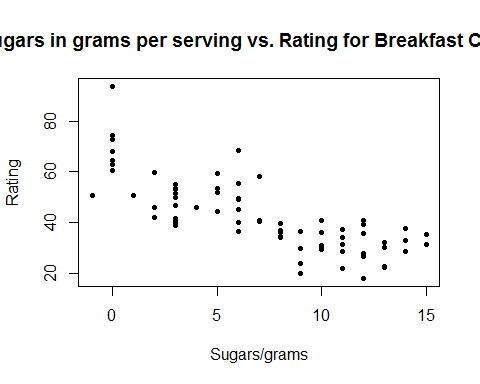


pairs.width = "90%"

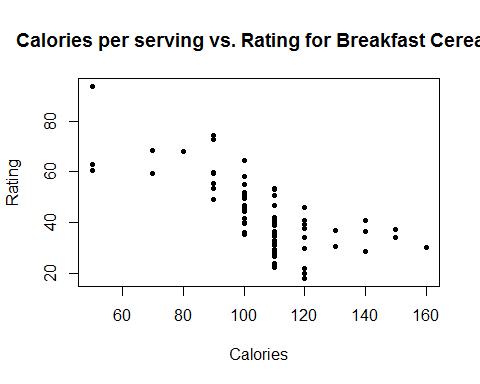
While there are a lot of things that can be noticed in all these pairwise plots (like the clear relationship between fiber and potassium), I think the most important relationships are those between some of the nutrients and the rating since this could give insight into how all of the nutrient content of the breakfast cereal, and some other factors like manufacturer, affect the rating of the cereal. Companies can use this to then make their cereal such that the cereal would get the highest rating as the customers would like it the most, and the company can improve its profits. SO rating is my response variable.

Some of the most evident relationships between the nutrients and rating is in the calories and sugar. While rating seem to have relationships with fat and protein they are also clumped together and it is not very clear.

plot(cereal$sugars, cereal$rating, pch=20, main="Sugars in grams per serving vs. Rating for Breakfast Cereals",  
 xlab="Sugars/grams", ylab="Rating")

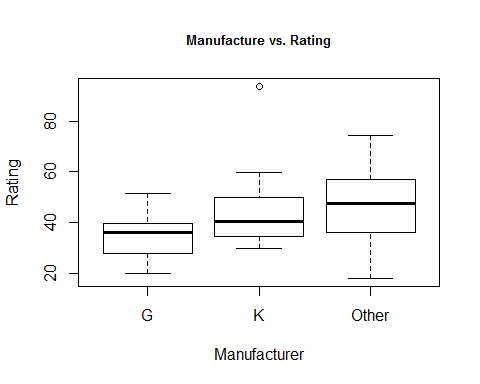


plot(cereal$calories, cereal$rating, pch=20, main="Calories per serving vs. Rating for Breakfast Cereals",  
 xlab="Calories", ylab="Rating")

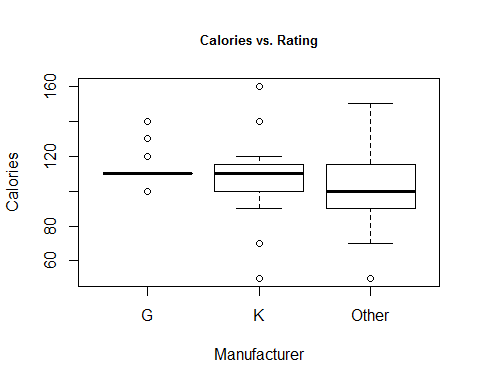
 It sounds reasonable that the number of calories and sugar would affect the rating of the cereal substantially because people who check the nutrient information before eating it are probably trying to minmize calories and sugar. THe data seems ot show a weak negative linear relationship between calories and rating and sugars and rating, supporting this point of view.

Finally, it seems reasonable to compare the ratings by manufacturers since manufacturers. All the small manufacturers are clumped in the small group whereas since there are two big ones which have the most number of observationsin the dataset as mentioned earlier.

cereal$mfr[cereal$mfr=="A"] <- "Other"  
cereal$mfr[cereal$mfr=="N"] <- "Other"  
cereal$mfr[cereal$mfr=="P"] <- "Other"  
cereal$mfr[cereal$mfr=="Q"] <- "Other"  
cereal$mfr[cereal$mfr=="R"] <- "Other"  
  
boxplot(cereal$rating~cereal$mfr, main="Manufacture vs. Rating", ylab="Rating", xlab="Manufacturer", cex.main=0.8)

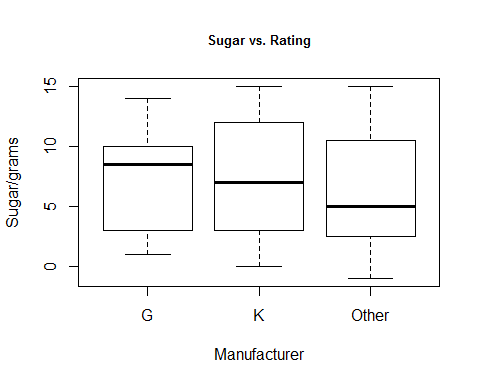


boxplot(cereal$calories~cereal$mfr, main="Calories vs. Rating", xlab="Manufacturer", ylab="Calories", cex.main=0.8)



It seems like the manufacturers with higher ratings have lesser number of calories and that seems reasonable considering our previous evaluations.

boxplot(cereal$sugars~cereal$mfr, main="Sugar vs. Rating", xlab="Manufacturer", ylab="Sugar/grams", cex.main=0.8)

 It seems like the manufacturers with higher ratings have lesser sugar and that seems reasonable considering our previous evaluations.

I think my linear model will contain calories and sugar as the predictors and rating as the response variable.