Gouthamrajan Nadarajan

**Homework 1**

**Problem 1:**

I looked at several different plots to explore the data in the cereal dataset. I first started off looking at histograms and visualize the distribution for specific variables. This is shown in the figure 1 below where I plot the raw data in histograms for the numerical predictors.

A picture containing text, receipt

Description automatically generated

Figure 1: Histogram of raw data of nutritional data

I then noticed that there were negative values among the predictors that skewed the results. Therefore, I eliminated the rows with the negative values and replotted the results in Figure 2.

A picture containing text, receipt

Description automatically generated

Figure 2: Histogram of nutritional data with outliers removed

**Problem 2:**

a)

Using rating as the response variable and the rest of the numerical variables as predictors, I trained an multiple regression model and using the coefficient values for each predictor, I found that the variables carbohydrates, fiber, protein, and fat were most correlated with rating as they had a coefficient magnitude closest to 1.

Fiber had a coefficient value of 3.443.

Carbohydrates had a coefficient value of 1.092

Protein had a coefficient value of 3.273

Fat had a coefficient value of -1.691

b)

The coefficient for the sugar predictor was found to be -7.25e-01. This tells us that the sugar predictor is negatively correlated with the response variable. It is also not as strongly correlated with the response variable as the other coefficients discussed in part (a).

c)

To look at the effect of interactions when using linear models, I looked at the carbohydrate and the protein predictor.

Using the \* method, we obtained a coefficient of -0.5296

Using the : method, I obtained a coefficient of 0.25829

This tells me that multiplying these variables does not yield a better predictor that is better than each of the individual predictors combined.

**Problem 3:**

a)

I first graphed some scatter plots below which I had initially hypothesized would show some sort of relation. These are shown below.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 3: List of Scatterplots for housing dataset

b)

Distance from employment centers was associated with crime rate. The suburbs that were the closest to the employment centers had the most crime and the suburbs further away had significantly less.

c)

Suburbs that have the highest crime rate include 381,406, and 419. The range of the crime feature was (0-90). The standard deviation and mean for this feature was 8.60 and 3.61 respectively

Suburbs that have the highest taxes are 489, 490, 491, 492, and 493. The range of the tax rate was between (150-750). The standard deviation and mean for this feature was 168.53 and 408.23 respectively

Suburbs that have the highest pupil-teacher ratio are 355 and 356. The range of the pupil-teacher ratio was between (12-22). The standard deviation and mean for this feature was 2.16 and 18.45 respectively

d)

I found that there were 64 suburbs average more than seven rooms per dwellings and 13 suburbs average more than eight rooms per dwelling.

Interestingly, I also found that the suburbs that average more than seven rooms per dwelling had much lower tax and much lower crime than the overall means of these features as described in 3c. The mean tax and crime statistic was 312.2344 and 0.9791089 respectively. The suburbs that average more than eight rooms per dwelling also follow a similar trend where the tax and crime statistic was 325.0769 and 0.7187954 respectively.