Report

Statistical Data Mining I: Homework 1

Sai Lone (MPS)
Person Number: 50139692

---

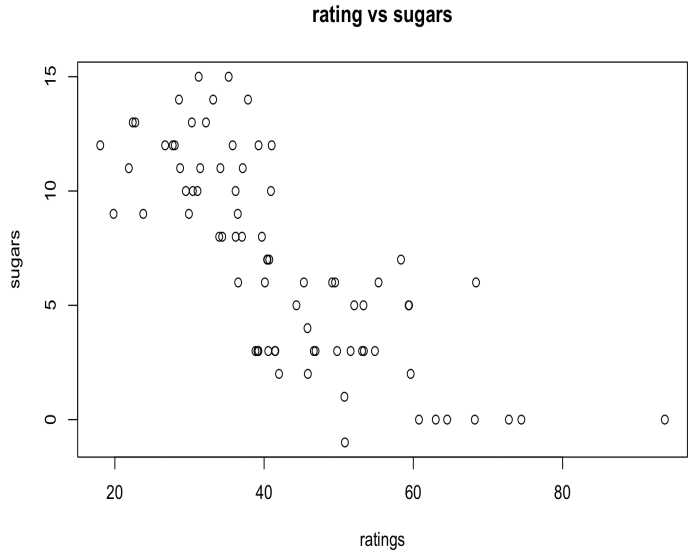
1. Pre-Processing and Data Analysis of the Cereal Dataset

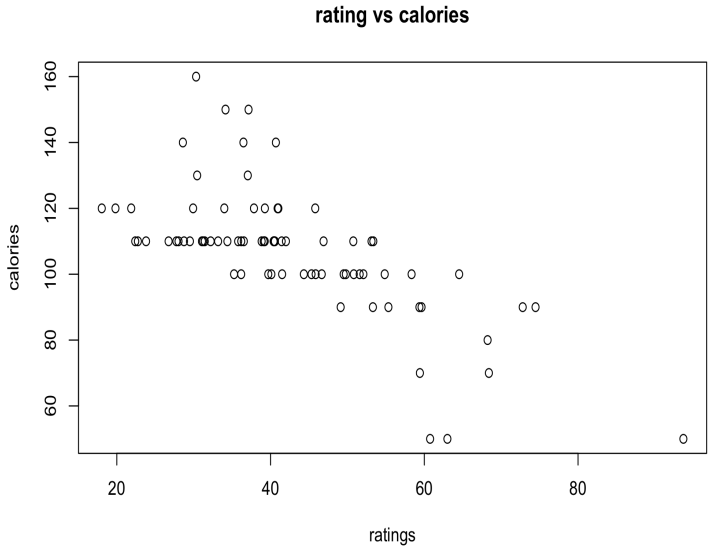
From the cereal dataset, we can see that it contains 77 rows and 16 columns. After investigating the data in the 77x16 data frame, notice that “mfr”, “type”, and “shelf” are classification attributes for the cereals so we can transform them into a factor type and view the summary details after that. This will allow for better manipulation of this type.

```
> cereal$mfr <- factor(cereal$mfr)
> cereal$type <- factor(cereal$type)
> cereal$shelf <- factor(cereal$shelf)
> summary(cereal)
                     name    mfr    type      calories        protein          fat
 100% Bran              : 1   A: 1   C:74   Min.   : 50.0   Min.   :1.000   Min.   :0.000
 100% Natural Bran      : 1   G:22   H: 3   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000
 All-Bran               : 1   K:23          Median :110.0   Median :3.000   Median :1.000
 All-Bran with Extra Fiber: 1   N: 6          Mean   :106.9   Mean   :2.545   Mean   :1.013
 Almond Delight         : 1   P: 9          3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000
 Apple Cinnamon Cheerios  : 1   Q: 8          Max.   :160.0   Max.   :6.000   Max.   :5.000
 (Other)                :71   R: 8
     sodium          fiber          carbo          sugars          potass          vitamins
 Min.   :  0.0   Min.   : 0.000   Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00
 1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00
 Median :180.0   Median : 2.000   Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00
 Mean   :159.7   Mean   : 2.152   Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25
 3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
 Max.   :320.0   Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00

 shelf      weight          cups           rating
 1:20   Min.   :0.50   Min.   :0.250   Min.   :18.04
 2:21   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
 3:36   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
```

Then we can plot the whole data set of cereal to get a feel for the different ranges of data in a boxplot and scatterplot. From the scatterplot, we can notice a slite relationship between a few of the columns with the rating column, specifically calories and sugars. After visualizing these relationships with a scatterplot where the x-axis representing the rating data and the y-axis representing calories and sugar, we can clearly see the direct correlation between the columns.





After further investigation of the data, we can notice some key outliers specifically the negative values in sugars. These outliers could have the wrong impact on our model so elimination of the outliers is necessary. We can also create a new data frame for just the rating, sugar and calories. This will allow us to build a smaller model when we perform regression.

2. Multiple Regression

a. Which predictors appear to have a significant relationship to the response.
   - Based on the summary of the large model regression we can see the significant relationship with all the columns with the exception of "shelf", "weight", and "cups".

```
> large_model <- lm(rating ~. - name - mfr, data=cereal)
> summary(large_model)

Call:
lm(formula = rating ~ . - name - mfr, data = cereal)

Residuals:
      Min        1Q    Median        3Q       Max
-5.246e-07 -2.573e-07  4.610e-08  2.242e-07  5.663e-07

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  3.677e-07  1.494e+08   <2e-16 ***
typeH       -3.917e-08  2.478e-07 -1.580e-01    0.875
calories    -2.227e-01  5.750e-09 -3.873e+07   <2e-16 ***
protein      3.273e+00  5.149e-08  6.357e+07   <2e-16 ***
fat         -1.691e+00  6.388e-08 -2.648e+07   <2e-16 ***
sodium      -5.449e-02  5.179e-10 -1.052e+08   <2e-16 ***
fiber        3.443e+00  4.434e-08  7.765e+07   <2e-16 ***
carbo        1.092e+00  1.956e-08  5.584e+07   <2e-16 ***
sugars      -7.249e-01  2.066e-08 -3.509e+07   <2e-16 ***
potass      -3.399e-02  1.486e-09 -2.288e+07   <2e-16 ***
vitamins    -5.121e-02  1.953e-09 -2.622e+07   <2e-16 ***
shelf       -3.700e-08  5.327e-08 -6.950e-01    0.490
weight      -4.010e-07  5.554e-07 -7.220e-01    0.473
cups         1.430e-07  1.965e-07  7.280e-01    0.470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.068e-07 on 63 degrees of freedom
Multiple R-squared:       1,   Adjusted R-squared:       1
F-statistic: 1.226e+16 on 13 and 63 DF,  p-value: < 2.2e-16
```

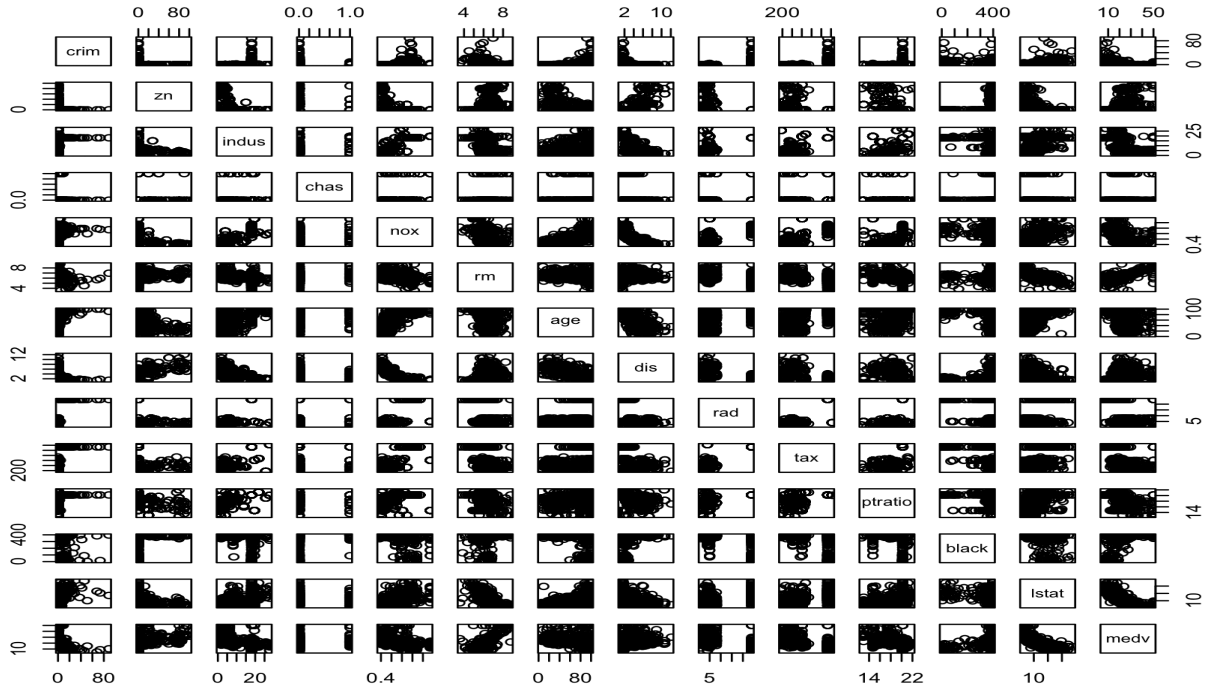
   - Based on the summary of the small model regression we can see the significant relationship in **sugars** because of its high pt values.

b. What does the coefficient variable for "sugar" suggest?
   - The coefficient variable for "sugar" suggest the higher the sugar level in cereal the lowing the rating.

c. Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?
   - After performing the regression using * and : on the variables for sugars and calories there seems no change in the summary details.

4. ISL textbook exercise 2.10

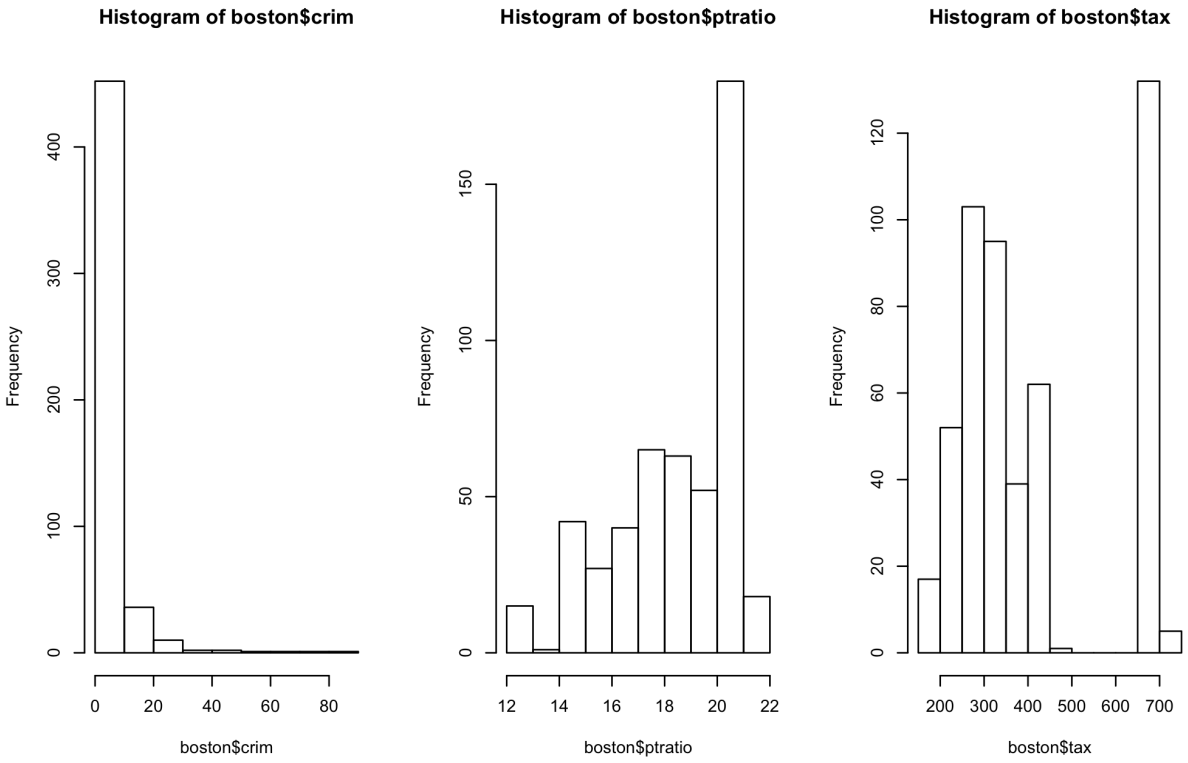
a. Make pairwise scatterplots of the predictors, and describe your findings.



- The scatterplots of all the predictors show that relationships between the variables in the Boston dataset do exist.

b. Are any of the predictors associated with the per capita crime rate?
   - After using the cor() function in r along with building a regression model for per capital crime rate we can see a strong association in dis, rad, and medv with crim.

c. Do any of the suburbs of Boston appear to have particularly high crime rates?
   Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
   - Yes, some of the suburbs of Boston appear to have high crime rates, tax rates, and pupil-teacher ratios. This can be visualized with a histogram plot of each predictor as seen below.



d. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

- **64** suburbs with more than 7 rooms per dwelling and **13** suburbs with more than 7 rooms per dwelling.

```
> dim(subset(boston, rm>7))
[1] 64 14
> dim(subset(boston, rm>8))
[1] 13 14
>
```

- After viewing summary details for both the subsets of suburbs that average more than seven rooms per dwelling versus eight rooms per dwelling, we can see that the max value of crim decreased from 19.60910 to 3.47428.