Statistical Data Mining Project 1

Question 1

The first problem provided us with cereal data in order to build a predictive model for nutritional rating i.e. response variable. This was done by pre-processing the data and using exploratory data analysis.

From reading the dataset, we can see there were predictors that need to be categorical, such as vitamins (either 0, 25, 100), cups and shelf. These factors are described as numbers and thus they were converted to factors. Other predictors such as protein, fat and sodium are measured in grams and are quantitative, thus were changed to numeric. To clean the data, missing values in the columns was first checked for, which there was none. When reading the data, negative values were identified in the predictors/columns potassium, carbohydrate, and sugars. Variance for potassium, carbohydrate, and sugars was calculated with values being 14.8, 15.0, and 18.9, respectively. These were roughly low variances, which indicates the data points were roughly close to the mean and were a suitable rough approximation. Thus, the negative values in these columns where imputed with the mean of the column.

Nutritional rating is determined by the nutritional value of the food, which is a well-balanced ratio measure of the vital nutrients carbohydrates, fat, protein, fibers, minerals (sodium and potassium), and vitamins in food. Thus, these predictors are assumed to affect the cereal rating and were used in the exploratory data analysis. As we can see in figure 1, the density plot of the response variable i.e. rating of cereal is slightly close to normality (Fig.1), with an average cereal rating of roughly 40.

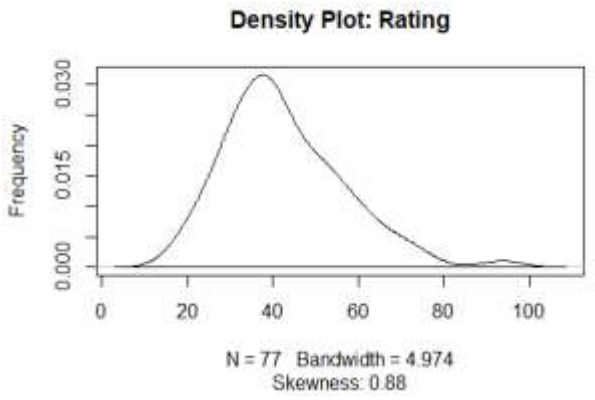


Figure 1: *Density plot of the rating of cereals with skewness of 0.88, cereals with skewness of 0.88, indicating the data is positively skewed*

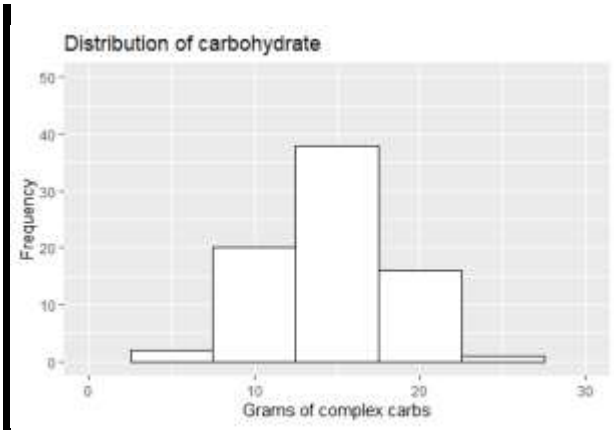


Figure 2: *Histogram plot of carbohydrate in the cereals, showing close it close to normal distribution, with skewness of 0.57.*

Moreover, the cereal data was cleaned for the presence outliers. Boxplots were used to determine the presence of outliers since having outliers in the predictor can drastically affect our prediction and the slope of the line of best fit. Outliers were defined as the data points that

lie outside 1.5 * IQR (Inter Quartile Range- which is the difference between 75th and 25th quartiles), these were identified using boxplot.stats in R. Histogram plots and density function were used to determine the distribution of the predictors in the cereal data set.

From figure 2, we can see the carbohydrate predictor has close to a normal distribution with a skewness close to 0.5. Furthermore, carbohydrate, vitamins, fat, and sodium did not have any outliers which were computed using boxplot.stats, thus no further pre-processing was done. However, there were outliers present in potassium, protien and dieraty fiber.

From figure 3 we can see the predictor poassiusm has a distrubtion that is positively skewed with a value of 1.37 and has 5 outliers. From figure 4 we can see the predictor dietary fiber has a distrubtion that is positively skewed with a value of 2.34 and has 3 outliers. From figure 5 we can see the predictor protein has a distrubtion that is moderately positively skewed with a value of 0.72 and has 3 outliers.

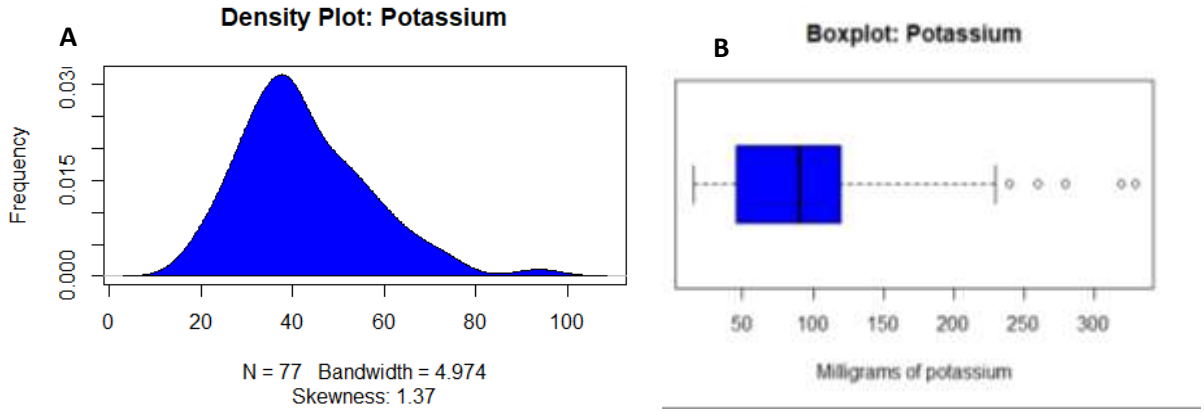


Figure 3: *A) Density plot of potassium, with skewness of 1.37, showing it is positively skewed. B) Boxplot of potassium, with 5 outliers (280, 320, 330, 260, and 240).*

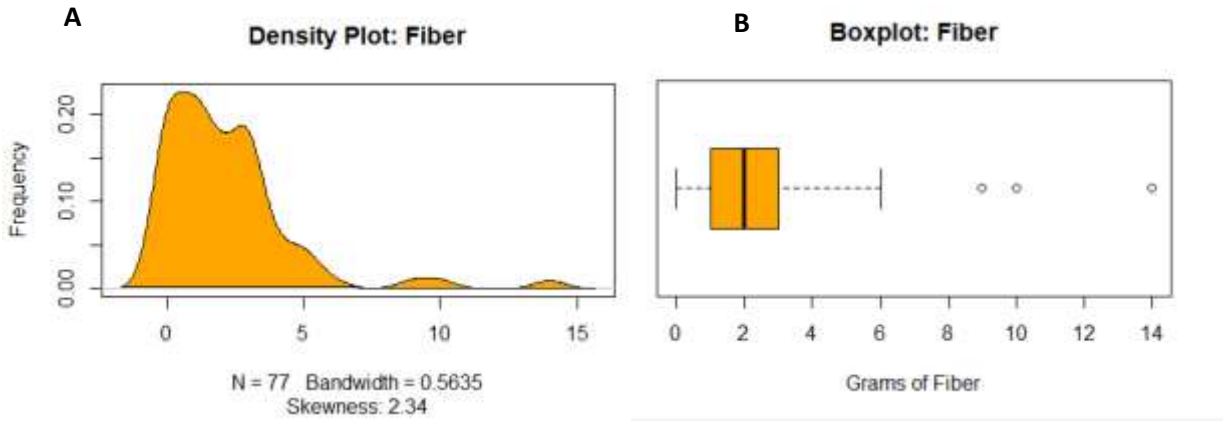


Figure 4: *A) Density plot of fiber, with skewness of 2.34, indicating it is positively skewed. B) Boxplot of fiber, with 3 outliers (9, 10, and 14)*

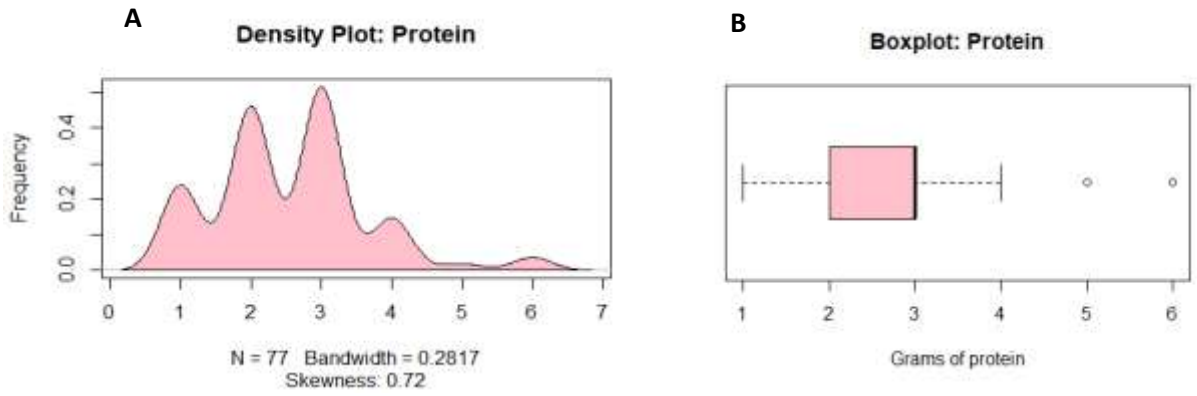


Figure 5: *A) Density plot of protein, with skewness of 0.72, indicating it is moderately positively skewed B) Boxplot of fiber, with 3 outliers (5, 6, and 6) found.*

To preprocess and replace the outliers with presentative values, the capping method was used. This was done by replacing the data points outside the lower limit with the value of 5th percentile and those above the upper limit with the value of 95th percentile[1]. This was a better method because we have a few number of data points (77 obs.) and each observation is as important as others, so eliminating them is not reasonable. After replacing the outliers density plots for each of the predictors was plotted to show the distribution as shown in figure 6, with their skewness closer to 0 compared to previously. Furthermore, boxplots were plotted as shown in figure 7 after the replacement of the outliers.

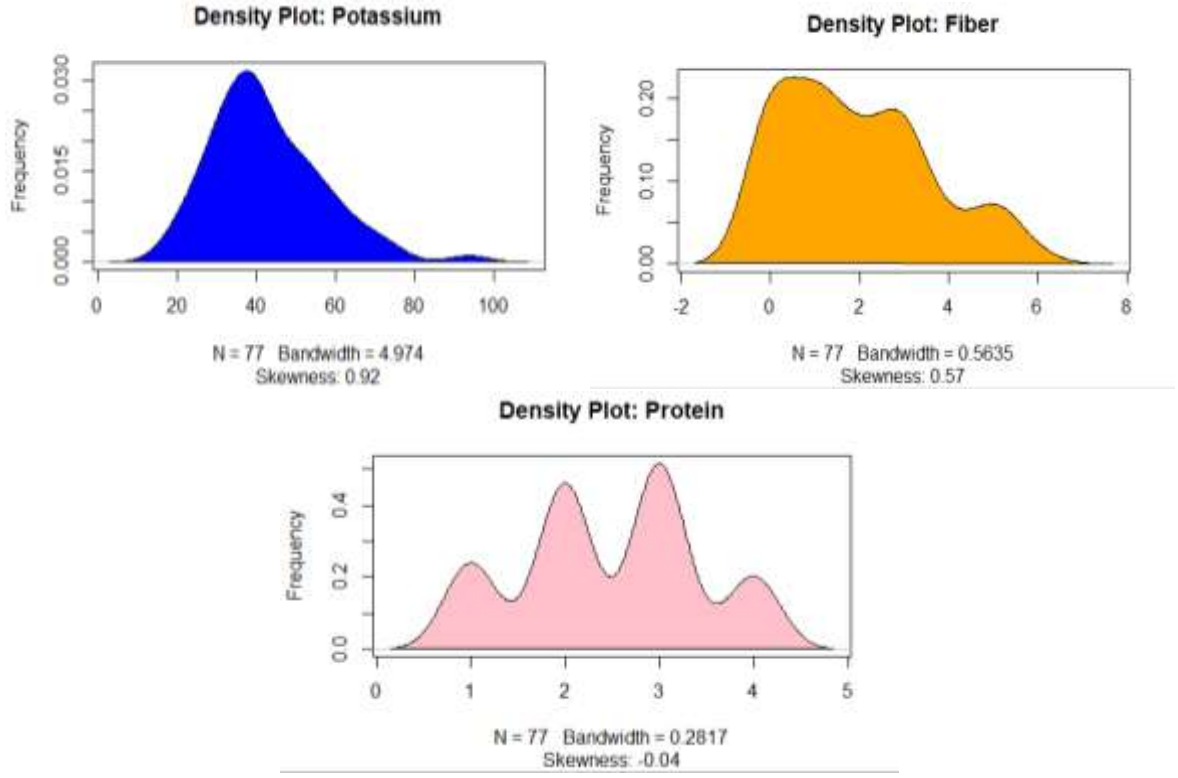


Figure 6: *A) Density plot of potassium after using the capping method, with a skewness closer to 0. B) Density plot of Fiber after using the capping method, with skewness closer to 0. C) Density plot of protein after using the capping method with a normal distribution.*

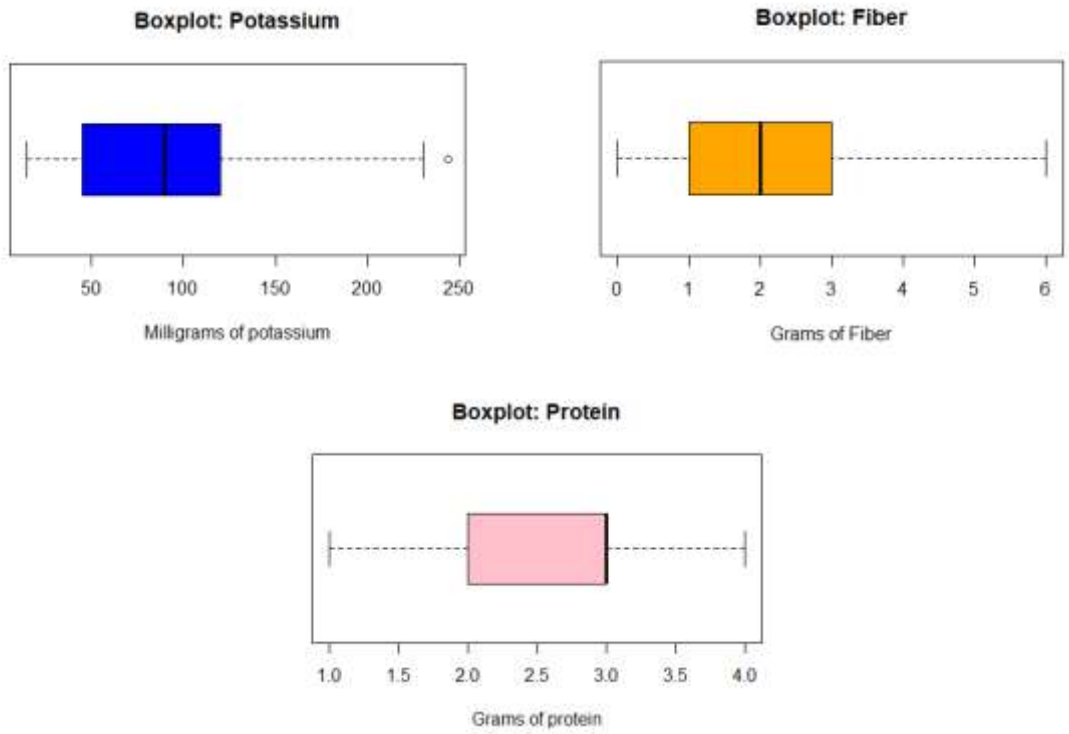


Figure 7: *Boxplots of predictors after pre-processing of outliers. A) Boxplot of potassium after using the capping method. B) Boxplot of Fiber after using the capping method. C)Boxplot of protein after using the capping method.*

After preproccessing the cereal data, an exploratory anayslsis was peroformed. Some scatter plots where plotted to determine the relationship the relationship between the response varibale and the predictor. From figure 8 and figure 9 we can see there is a negative correlation of both sugar and sodium with cereal rating. From figure 10 and 11, we can see a postive correlation of both potassium dietary fiber with cereal rating. Transformation as an exploratory analysis was perfromed using log of the calories and the ceral rating as shown in figure 12.

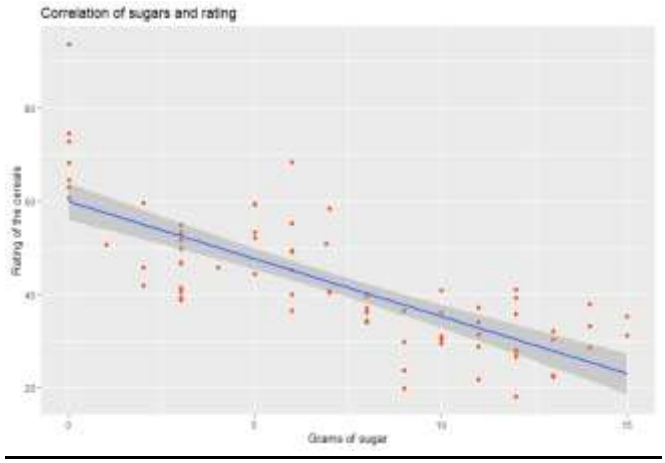


Figure 8: *Scatterplot of the relationship between sugar and the cereal rating*

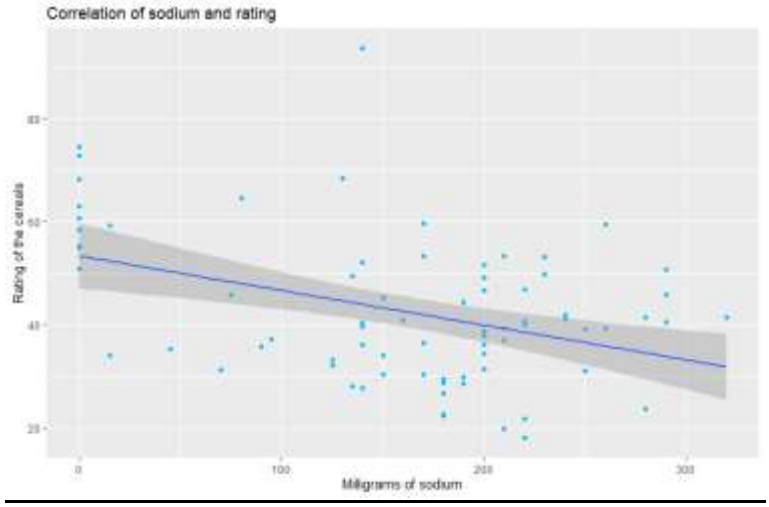
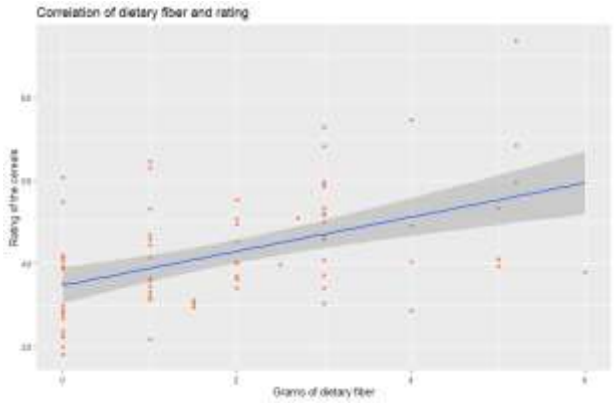


Figure 9: *Scatterplot between sodium and cereal rating*



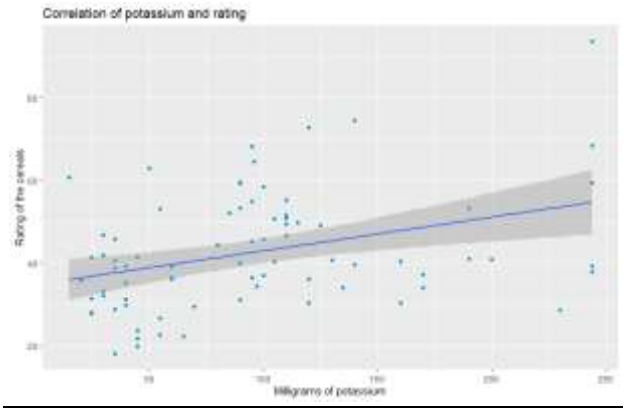


Figure 10: *Scatterplot of dietary fiber &cereal rating* ***Figure 11:*** *Scatterplot of potassium & cereal rating*

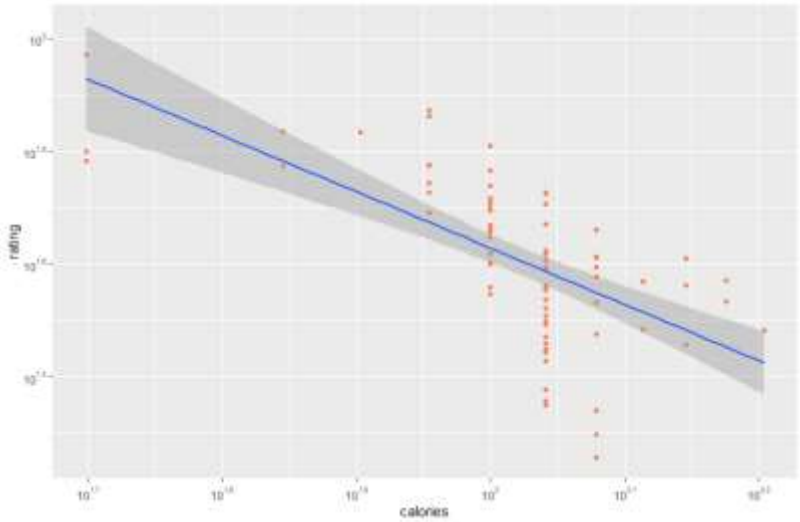


Figure 12: *Scatterplot of log of the dietary fiber and log of cereal rating*

Question 2

In this problem we had to perform a multiple regression on the dataset which was preprocessed in the first problem, with the response variable being the rating and we had to see the significant relationship of the predictors with the response and examine interactions with predictors.

First, the lm() function in R was used to create the multiple linear regression for the cleaned cereal data set, with cereal rating as the response variable and all the other predictors except for the name column, as it is qualitative. The summary of fitted model was created as shown in Table 1.

Table 1: Summary statistics of the multiple linear regression model of cleaned cereal data

```
Call:
lm(formula = rating ~ . - name, data = cereal_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2760 -1.4755 -0.2448  0.9722 12.9693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.353947   3.991602  16.874  < 2e-16 ***
mfr          0.182113   0.220539   0.826   0.4121
type        -3.752174   2.132174  -1.760   0.0834 .
calories    -0.155883   0.063168  -2.468   0.0164 *
protein      3.504107   0.630424   5.558 6.12e-07 ***
fat         -3.620524   0.651718  -5.555 6.19e-07 ***
sodium      -0.045234   0.005327  -8.491 5.59e-12 ***
fiber        1.342397   0.547065   2.454   0.0170 *
carbo       -0.182532   0.281216  -0.649   0.5187
sugars      -1.771557   0.266329  -6.652 8.58e-09 ***
potass       0.014792   0.016165   0.915   0.3637
vitamins    -1.909290   1.132924  -1.685   0.0970 .
shelf       -0.337822   0.492018  -0.687   0.4949
weight      12.111471   5.601346   2.162   0.0345 *
cups         0.096422   0.174335   0.553   0.5822
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.796 on 62 degrees of freedom
Multiple R-squared:  0.9677,    Adjusted R-squared:  0.9604
F-statistic: 132.6 on 14 and 62 DF,  p-value: < 2.2e-16
```

a) Table 1 was used to determine the predictors which appear to have a significant relationship to the response. The p-value of each predictor under the coefficients was used to check for statistical significance.
   The three significance stars in the last column indicate p<0.001, which means the predictor is very significant (as the more the stars, the more significant the predictor). This suggest calories, protein, fat, sodium, sugar are more significant predictors than others (we don’t interpret the intercept β0).
   The Null hypothesis for the linear regression is the coefficient βj of the predictor=0. Since the βj of protein, fat, sodium, sugar is not equal to 0 (3.5,-3.6,-0.04, -1.77,

respectively), we reject the null hypothesis and therefore βj is significant and the predictor is corresponds to is important.
The predictors with one star are weight and calories are significant as $p<0.05$. The other predicators do not have statistically significant relationship with cereal rating.

b) The coefficient variable for "sugar" which is $\beta 1 = -1.77$is not equal to 0, therefore we reject the null hypothesis and βj of sugar is significant and sugar is an important predictor i.e. there is a relationship between sugar and cereal rating. The regression coefficient β for sugar is -1.77. This means other predictors are fixed the cereal rating decreases by 1.77, with every gram increase in sugar i.e. the cereal rating is greater with less sugar (y = b0 + b1*x1).

c) To fit the model with interactions and try to predict cereal rating from all the different variables, we first see how correlated the variables are, which are statistically significant as shown in table 1 above. This was done using a chart correlation as shown figure 13, with the distribution of the predictors on the diagonal. The top diagonal with the correlation coefficients and p-values as stars and the lower diagonal has scatterplots.

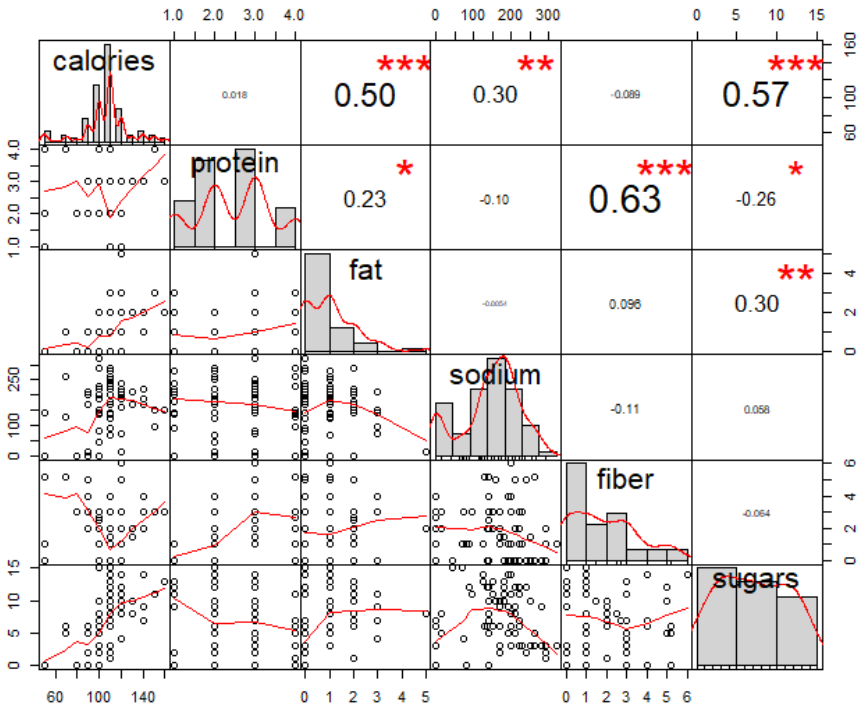


***Figure 13**: Correlation chart for statistically significant predictors for rating*

From figure 13 above, we can see that protein and fiber are correlated and protein is more significant, whereas fiber is less given the p-values in table 1. If we take out protein we see

more significance in fiber as they are correlated and the model is focusing on one and ignoring the other. The same happens with fat and calories, as well as sugars and calories, as they are correlated as shown in figure 13. The calories is less significant than either fat or protein as seen in table 1 above. We have to take out one of them to fit the model and reduce the number of predictors to make it simpler.

Table 3: Summary statistics of the interaction between fiber and protein

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.831733   3.576132  18.968  < 2e-16 ***
mfr            0.069713   0.199426   0.350 0.727866
type          -3.173286   1.914567  -1.657 0.102565
calories      -0.139809   0.056702  -2.466 0.016504 *
protein        2.067466   0.667132   3.099 0.002936 **
fat           -3.537142   0.583928  -6.057 9.35e-08 ***
sodium        -0.051056   0.004983 -10.246 6.99e-15 ***
fiber         -1.209288   0.799211  -1.513 0.135418
carbo         -0.133544   0.252099  -0.530 0.598225
sugars        -1.825634   0.238853  -7.643 1.81e-10 ***
potass        -0.007812   0.015518  -0.503 0.616478
vitamins      -1.598729   1.017355  -1.571 0.121250
shelf         -0.466642   0.441717  -1.056 0.294939
weight        13.700149   5.030967   2.723 0.008420 **
cups           0.112340   0.156153   0.719 0.474627
protein:fiber  1.166605   0.288711   4.041 0.000152 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.504 on 61 degrees of freedom
Multiple R-squared:  0.9745,   Adjusted R-squared:  0.9682
F-statistic: 155.4 on 15 and 61 DF,  p-value: < 2.2e-16
```

Table 4: Summary statistics of the interaction between fiber and protein

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.79403    5.02150   5.137 2.24e-06 ***
protein        4.99766    2.32509   2.149   0.0349 *
fiber          2.53722    3.37707   0.751   0.4549
protein:fiber -0.07162    1.12141  -0.064   0.9493
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.01 on 73 degrees of freedom
Multiple R-squared:  0.2976,   Adjusted R-squared:  0.2687
F-statistic: 10.31 on 3 and 73 DF,  p-value: 9.741e-06
```

Table 3, shows the interaction between protein and fiber is significant (with 3 stars). The R^2 value and the F-statistic increased, and the standard errors decrease suggesting a better fit. Table 4 also shows that there is no significance and thus they are independent and there is no relationship between them since it is a full first-order model.

Table 5: Summary statistics of the interaction between calories and fat using ":"

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.159050   3.929548  18.618  < 2e-16 ***
mfr            0.088451   0.201501   0.439 0.662241
type          -0.937367   2.070844  -0.453 0.652407
calories      -0.293901   0.067852  -4.332 5.62e-05 ***
protein        3.514702   0.571672   6.148 6.58e-08 ***
fat          -13.089816   2.564355  -5.105 3.49e-06 ***
sodium        -0.038854   0.005115  -7.596 2.18e-10 ***
fiber          1.681314   0.504050   3.336 0.001452 **
carbo         -0.207847   0.255092  -0.815 0.418361
sugars        -1.691970   0.242415  -6.980 2.51e-09 ***
potass        -0.006182   0.015666  -0.395 0.694508
vitamins      -0.803194   1.067877  -0.752 0.454859
shelf         -0.866345   0.467392  -1.854 0.068638 .
weight        17.650745   5.284852   3.340 0.001434 **
cups          -0.062865   0.163563  -0.384 0.702060
calories:fat   0.087586   0.023081   3.795 0.000342 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.536 on 61 degrees of freedom
Multiple R-squared:  0.9738,   Adjusted R-squared:  0.9674
F-statistic: 151.4 on 15 and 61 DF,  p-value: < 2.2e-16
```

Table 5, shows the interaction between calories and fat is significant (with 3 stars). The R^2 value and the F-statistic increased, and the standard errors decrease suggesting a better fit.

Table 6: Summary statistics of the interaction between calories and sugars using ":"

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    70.433031   4.707920  14.961  < 2e-16 ***
mfr             0.161086   0.220344   0.731  0.46754
type           -3.248968   2.163377  -1.502  0.13831
calories       -0.200936   0.072937  -2.755  0.00773 **
protein         3.587450   0.631638   5.680 4.01e-07 ***
fat            -3.450847   0.663852  -5.198 2.47e-06 ***
sodium         -0.043945   0.005410  -8.122 2.71e-11 ***
fiber           1.463347   0.553837   2.642  0.01045 *
carbo          -0.149936   0.281377  -0.533  0.59606
sugars         -2.317839   0.520066  -4.457 3.63e-05 ***
potass          0.008145   0.016996   0.479  0.63349
vitamins       -1.602899   1.156014  -1.387  0.17062
shelf          -0.487287   0.505131  -0.965  0.33852
weight         12.257527   5.580560   2.196  0.03187 *
cups            0.041665   0.179343   0.232  0.81707
calories:sugars  0.005475   0.004483   1.221  0.22670
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.785 on 61 degrees of freedom
Multiple R-squared:  0.9684,   Adjusted R-squared:  0.9607
F-statistic: 124.8 on 15 and 61 DF,  p-value: < 2.2e-16
```

Table 7: Summary statistics of the interaction of sugars and calories using "*"

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     101.519956   7.564525  13.421  < 2e-16 ***
sugars           -5.006682   1.074284  -4.660 1.39e-05 ***
calories         -0.447940   0.078420  -5.712 2.27e-07 ***
sugars:calories   0.030311   0.009792   3.095  0.00279 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.635 on 73 degrees of freedom
Multiple R-squared:  0.7162,   Adjusted R-squared:  0.7046
F-statistic: 61.41 on 3 and 73 DF,  p-value: < 2.2e-16
```

Table 6, shows the interaction between calories and sugars is not significant. The R^2 value and the F-statistic decreased, and the standard errors increased suggesting a not a good fit. Table 7 shows that there sugars and calories are dependent, as * indicates a full-first order model.

Question 3

In this problem we had to compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data, which is hand written recognition data. We considered only the 2's and 3's digits id's which were followed by the 256 grayscale values. For the knn model we performed it on k = 1,3,5,7,9,11, 13,15. Both the training and the test error for each choice of k was determined, shown in table . This was also performed on linear regression as shown in table.

Table8: k-NN error for both training and testing datasets

k-NN model	Training error	Testing error
1	0.0000	0.0247
3	0.0050	0.0302
5	0.0058	0.0302
7	0.0065	0.0330
9	0.0094	0.0357
11	0.0086	0.0357
13	0.0086	0.0385
15	0.0094	0.0385

Table 9: Linear Regression model error for both training and testing datasets

Training error	Testing error
0.0058	0.0412

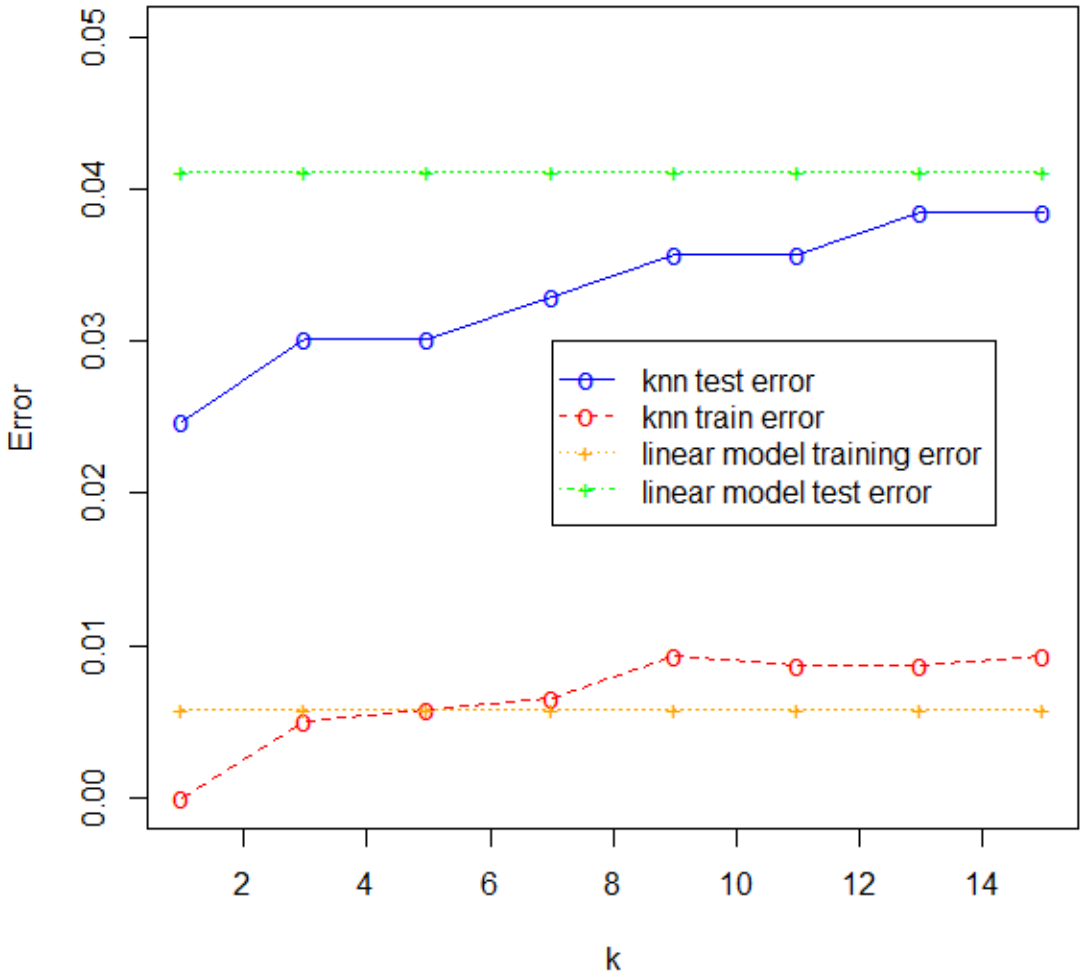


Figure 13: *A plot of the knn and linear regression model with both of their training and testing data*

From the table 8 and 9, we can see that with regards to the testing error, the KNN model has a lower error rate compared to the linear regression model. Additionally, the KNN model performs better than the linear regression model at small values of k. This is due to the curse of dimensionality, which is when the dimensions increase, then problems arise when making good predictions. Therefore, as we increase the number of predictors, the neighbors get further and further away and we need to cover a lot more space. In this data set we have 256 predictors and the observations are spread out and thus we have an under fitting and error increases. Furthermore, from the table for KNN, we can see that as the k increases, the error in both the training and testing datasets increases.

Question 4

In this problem, we had to perform an exploratory anayslsis on the Boston housing data. To view the correlated predictors in the Boston data set, pairwise scatterplots of the predictors was plotted, as shown in figure 14 , as well as, a graph of correlation matrix as shown in figure 15. From these figures there are some high and low correlations between variables.

- From pairwise plot we see there is positive correlation between ox - nitric oxides concentration- and indus - proportion of non-retail business acres per town. This is supported in the correlation matrix, with correlation coefficient of 0.76. This is predictable as nitric oxide is produced from non-retail businesses in their production of goods/services.
- From pairwise plot we see there is positive correlation between ox- nitric oxides concentration and dis- weighted distances to five Boston employment centers. This is supported in the correlation matrix, with correlation coefficient of 0.77. This is expected as more nitric oxides is produced the further the distance to travel to the center (by vehicle).
- From pairwise plot we see there is positive correlation between the tax rate and rad - index of accessibility to radial highways. This is supported in the correlation matrix, with correlation coefficient of 0.91. This is expected neighborhoods that exhibit high accessibility to highways rate also have high tax rate to help the government maintain them.
- From pairwise plot we see there is low correlation between the median value of owner-occupied homes in dollars (medv) and and dis (distance to employment centres). This is supported in the correlation matrix, with correlation coefficient of 0.25. This is predictable as the value of home doesn't affect the distance to center.
- From pairwise plot we see there is negative correlation between the median value of owner-occupied homes in dollars (medv) and lower status of the population (lstat). This is supported in the correlation matrix, with correlation coefficient of -0.74. This is expected as the highest lower status of the population have the lowest median value.
- From pairwise plot we see there is negative correlation between the median value of owner-occupied homes in dollars (medv) and average number of rooms per dwelling (rm). This is supported in the correlation matrix, with correlation coefficient of 0.7. This is predictable as in general the more rooms in a dwelling, the higher the median value.

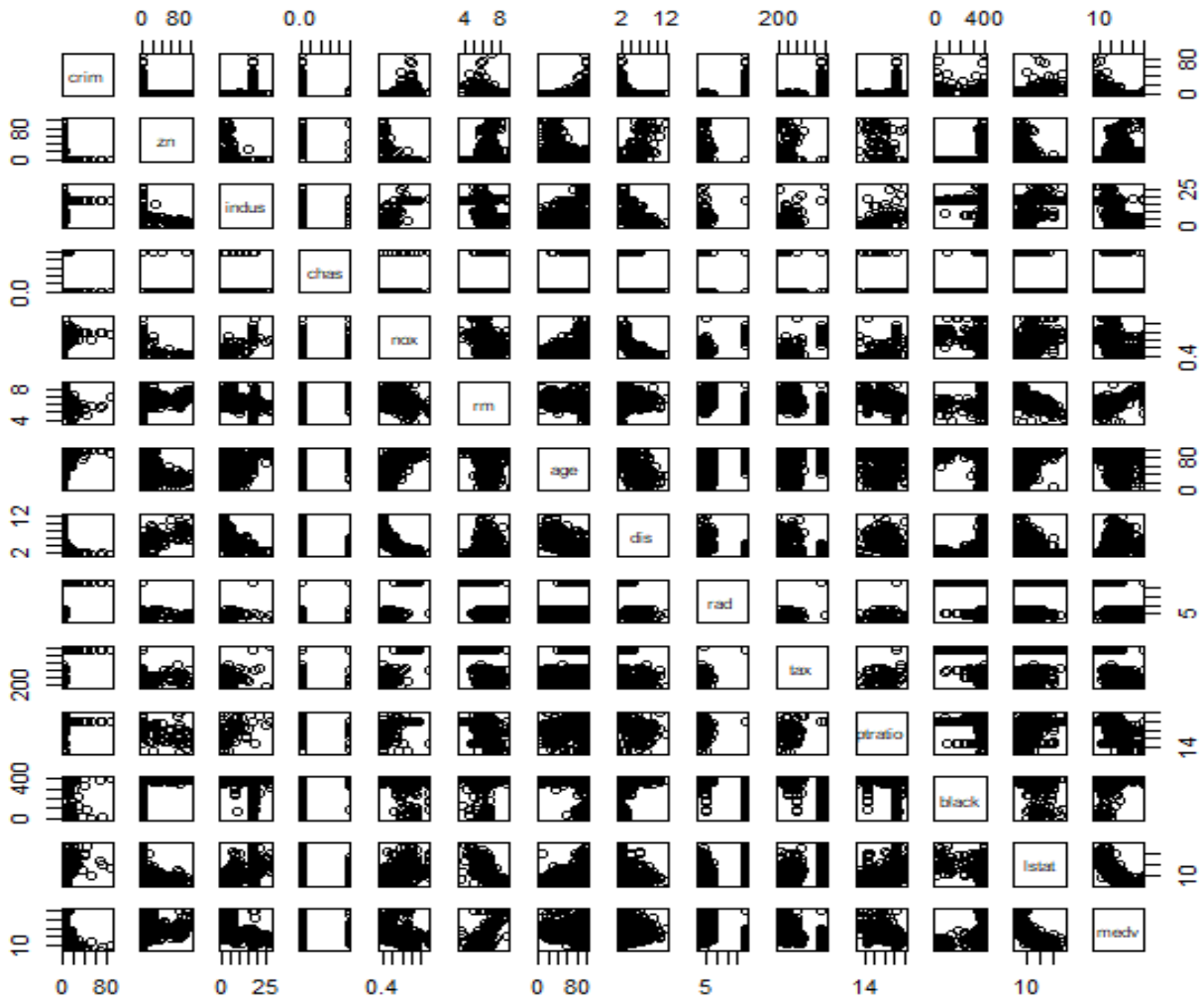


***Figure 14**: Pairwise scatterplots of the predictors of the suburbs of Boston*

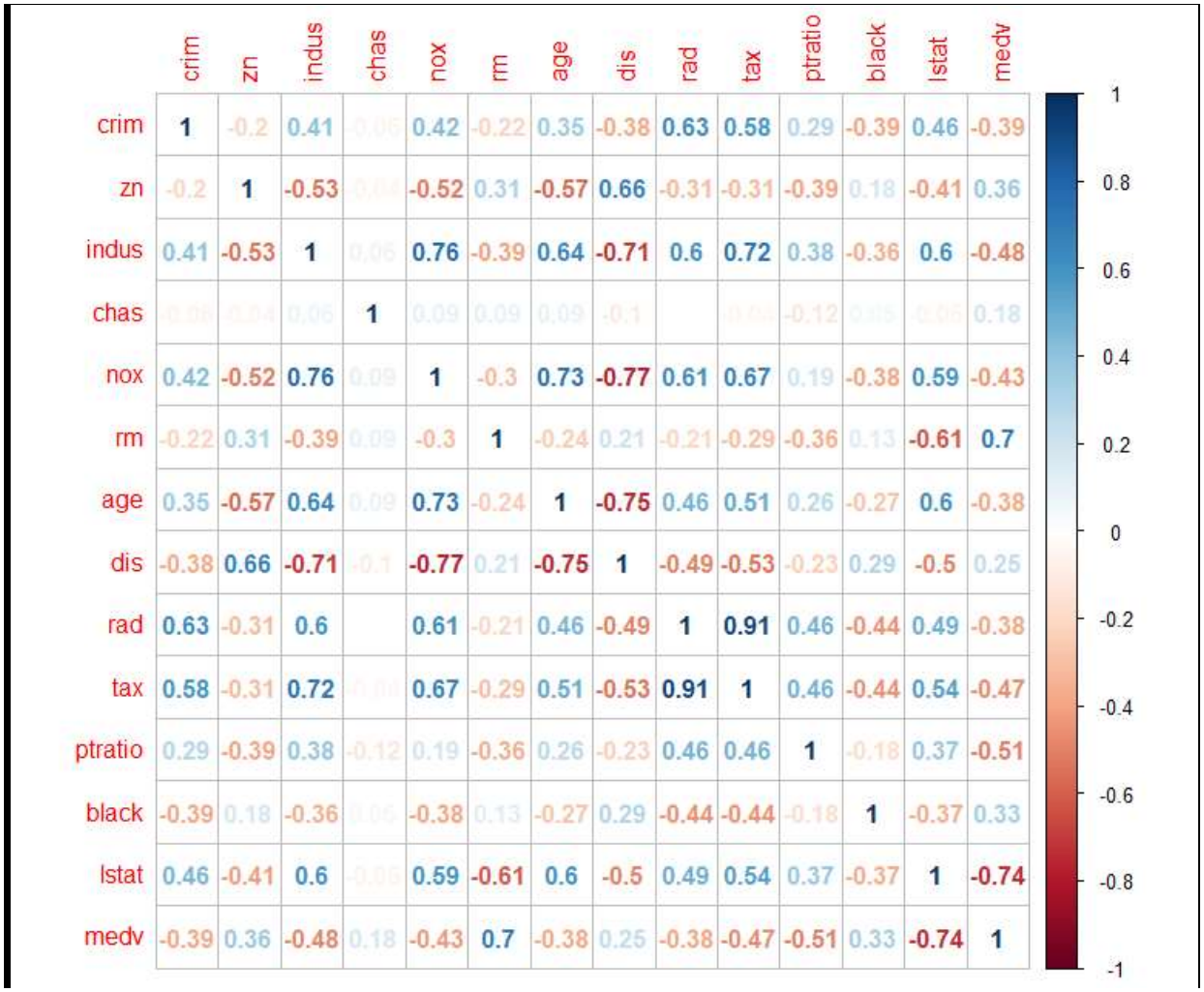


Figure 15: *Graph of correlation matrix for the predictors of the suburbs of Boston*

To determine if any of the predictors are associated with per capita crime rate, the pairwise plot and the graph of the correlation matrix were used. From these plots, we can see that the rad (index of accessibility to radial highways), tax (full-value property-tax rate per $10,000), and lstat (lower status of the population), have a positive correlation with crime rate, with correlation coefficients of 0.63, 0.58, and 0.46 respectively. The rad and tax have strong positive correlation with crime rate as the value is between 0.5 and 1, whereas the lstat has a moderate positive correlation as it is between 0.3 and 0.5.

To find whether the suburbs of Boston appear to have particularly high crime rates, tax rates, pupil-teacher ratios, boxplots where plotted to determine the range of these predictors as shown in figure 16.

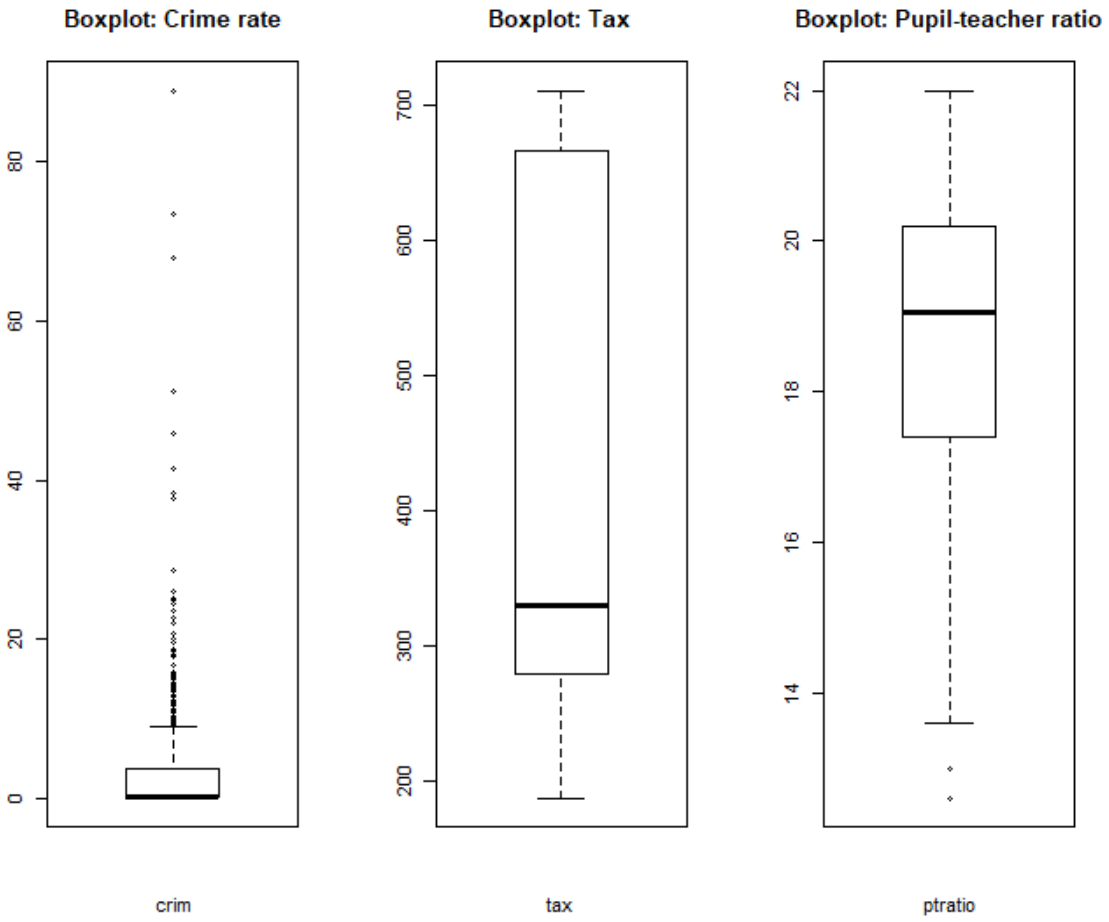


Figure 16: *Boxplots of crime rates, tax rates, pupil-teacher ratios for suburbs of Boston*

- The per capita crime rate by town ranges from 0.0063 to 88.97. From the boxplot we see there are several outlier suburbs observed from 10 to 89 that that lie outside 1.5*IQ R. Since the high crime rate are outliers, this suggest most of the suburbs in Boston hav e a low crime rate.
  The median crime rate of the suburbs is near zero with 0.25 crime rate.
- The tax rate ranges of the suburbs have crime from about 187 to 711 dollars. The median rates is 3200 dollars. From the boxplot we can see there are no outliers.
- The pupil-teacher ratio by town ranges from 12.6 to 22. The median pupil-teacher ratio is 19.5. From the boxplot we see there are some outliers from 12.6 to 13, but these don't lie outside 1.5*IQR.

In this Boston data set, 64 suburbs average more than seen rooms per dwelling and 13 average more than eight rooms per dwelling. Based on the suburbs that average more than eight rooms per dwelling.

Table 10: Summary of the suburbs that average more than eight rooms per dwelling in Boston

crim	zn	indus	chas	nox	rm	age	dis
Min. :0.02009	Min. : 0.00	Min. : 2.680	Min. :0.0000	Min. :0.4161	Min. :8.034	Min. : 8.40	Min. :1.801
1st Qu.:0.33147	1st Qu.: 0.00	1st Qu.: 3.970	1st Qu.:0.0000	1st Qu.:0.5040	1st Qu.:8.247	1st Qu.:70.40	1st Qu.:2.288
Median :0.52014	Median : 0.00	Median : 6.200	Median :0.0000	Median :0.5070	Median :8.297	Median :78.30	Median :2.894
Mean :0.71879	Mean :13.62	Mean : 7.078	Mean :0.1538	Mean :0.5392	Mean :8.349	Mean :71.54	Mean :3.430
3rd Qu.:0.57834	3rd Qu.:20.00	3rd Qu.: 6.200	3rd Qu.:0.0000	3rd Qu.:0.6050	3rd Qu.:8.398	3rd Qu.:86.50	3rd Qu.:3.652
Max. :3.47428	Max. :95.00	Max. :19.580	Max. :1.0000	Max. :0.7180	Max. :8.780	Max. :93.90	Max. :8.907

rad	tax	ptratio	black	lstat	medv
Min. : 2.000	Min. :224.0	Min. :13.00	Min. :354.6	Min. :2.47	Min. :21.9
1st Qu.: 5.000	1st Qu.:264.0	1st Qu.:14.70	1st Qu.:384.5	1st Qu.:3.32	1st Qu.:41.7
Median : 7.000	Median :307.0	Median :17.40	Median :386.9	Median :4.14	Median :48.3
Mean : 7.462	Mean :325.1	Mean :16.36	Mean :385.2	Mean :4.31	Mean :44.2
3rd Qu.: 8.000	3rd Qu.:307.0	3rd Qu.:17.40	3rd Qu.:389.7	3rd Qu.:5.12	3rd Qu.:50.0
Max. :24.000	Max. :666.0	Max. :20.20	Max. :396.9	Max. :7.44	Max. :50.0

Table 11: Summary of the suburbs for the less than or equal to eight rooms per dwelling in Boston

crim	zn	indus	chas	nox	rm	age	dis
Min. : 0.00632	Min. : 0.0	Min. : 0.46	Min. :0.00000	Min. :0.3850	Min. :3.561	Min. : 2.9	Min. : 1.130
1st Qu.: 0.08014	1st Qu.: 0.0	1st Qu.: 5.19	1st Qu.:0.00000	1st Qu.:0.4490	1st Qu.:5.879	1st Qu.: 44.4	1st Qu.: 2.088
Median : 0.24522	Median : 0.0	Median : 9.69	Median :0.00000	Median :0.5380	Median :6.185	Median : 77.3	Median : 3.216
Mean : 3.68986	Mean : 11.3	Mean :11.24	Mean :0.06694	Mean :0.5551	Mean :6.230	Mean : 68.5	Mean : 3.805
3rd Qu.: 3.77498	3rd Qu.: 12.5	3rd Qu.:18.10	3rd Qu.:0.00000	3rd Qu.:0.6240	3rd Qu.:6.575	3rd Qu.: 94.3	3rd Qu.: 5.215
Max. :88.97620	Max. :100.0	Max. :27.74	Max. :1.00000	Max. :0.8710	Max. :7.929	Max. :100.0	Max. :12.127

rad	tax	ptratio	black	lstat	medv
Min. : 1.000	Min. :187.0	Min. :12.60	Min. : 0.32	Min. : 1.73	Min. : 5.00
1st Qu.: 4.000	1st Qu.:280.0	1st Qu.:17.40	1st Qu.:374.71	1st Qu.: 7.34	1st Qu.:16.70
Median : 5.000	Median :334.0	Median :19.10	Median :391.83	Median :11.65	Median :21.00
Mean : 9.604	Mean :410.4	Mean :18.51	Mean :355.92	Mean :12.87	Mean :21.96
3rd Qu.:24.000	3rd Qu.:666.0	3rd Qu.:20.20	3rd Qu.:396.24	3rd Qu.:17.11	3rd Qu.:24.80
Max. :24.000	Max. :711.0	Max. :22.00	Max. :396.90	Max. :37.97	Max. :50.00

Comparing table 10 and table 11, we can see that the suburbs that average more than eight rooms per dwelling have low crime rate, with a mean of 0.71 as opposed to 3.68. However, the proportion of blacks is relatively high in these suburbs with an average of 385 as opposed to an average of 356 and a minimum of 354 rather than 0.32. Furthermore, the suburbs that average more than eight rooms per dwelling have a high median value i.e. are more expensive, with a mean of 44 instead of 22 dollars. However, the property tax rate is low with an average of 325, rather than 408, suggesting this predictor doesn't depend on the more dwellings of a property. There is a lower teach-pupil ratio in areas with more than eight rooms per dwelling with mean of 16.36, instead of 18.51. There is a smaller low status people in more dwelling house with mean of 4.31, rather than 12.87. Furthermore, locations with more than eight rooms per dwelling seem to be far away from radial highways compared to those less than eight.

Reference

[1] R-statistics.co. (2019). *Outlier Treatment With R | Multivariate Outliers*. [online] Available at: http://r-statistics.co/Outlier-Treatment-With-R.html [Accessed 20 Sep. 2019].