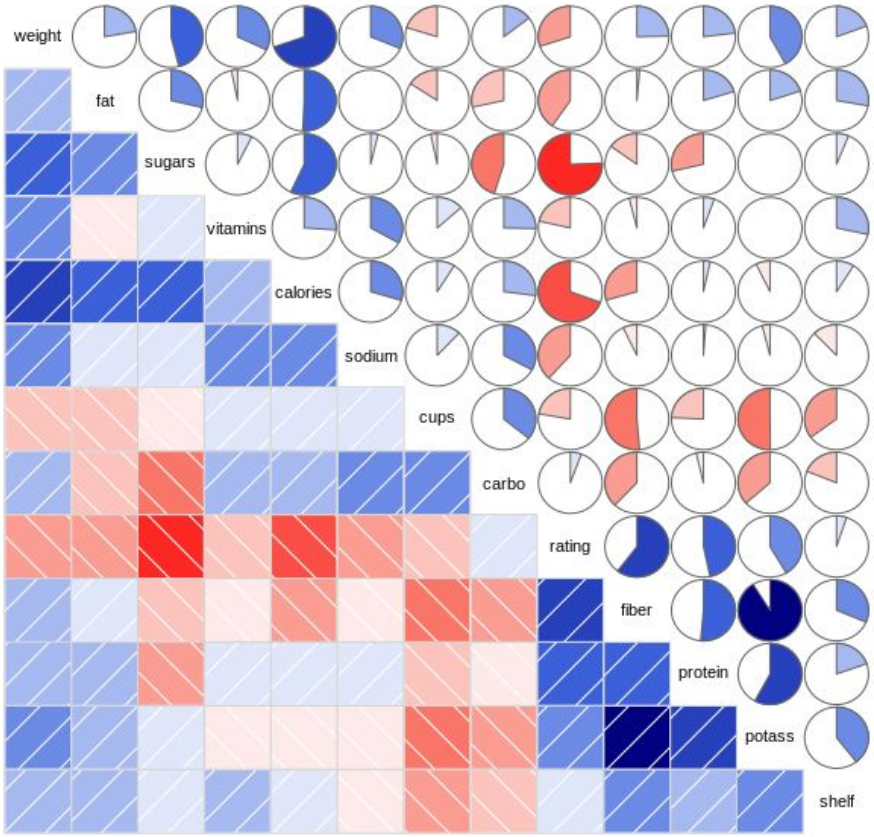
# Jeetendra Gan (jgan2), Number-21
# Problem 1

### Cleaning the data

Carbohydrates, sugar and potassium have negative values in some cases. For instance, there are 1 cereal with -1 carbo value, 2 cereals with -1 potass value, and 1 cereal with -1 sugar value. Quaker Oatmeal is common in two cases. If we drop the values, then we lose only 3 data values. Other option was to set the incorrect values to the mean. Dropping is better than adding some value on the basis of assumption.

### Data exploration

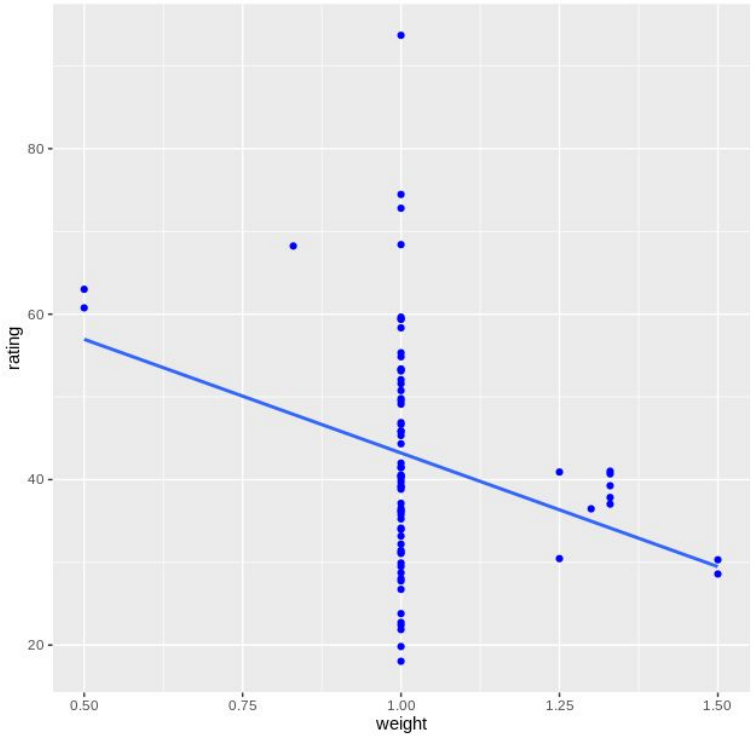
In the correlation matrix shown below, it can be observed that there are a few correlations between predictors and the dependent variable.

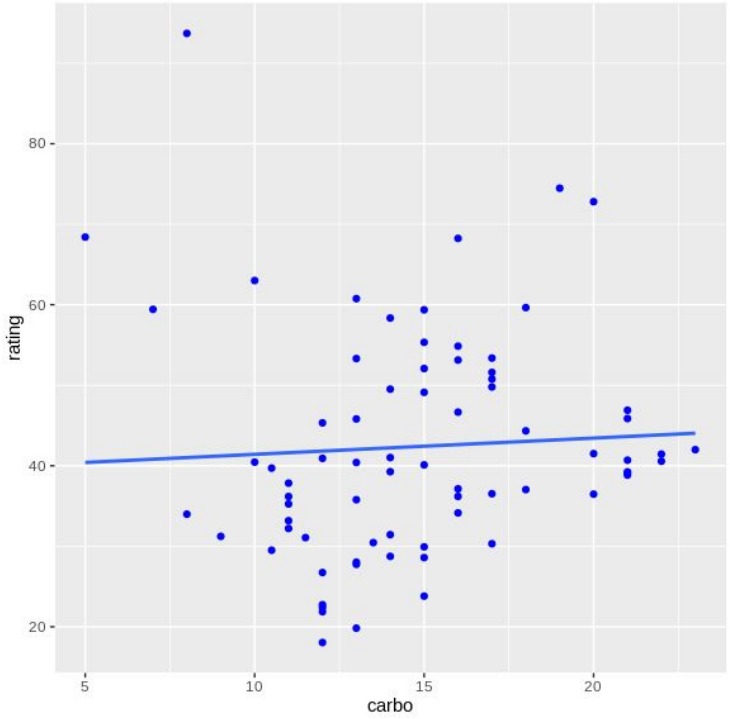



Following are the observations

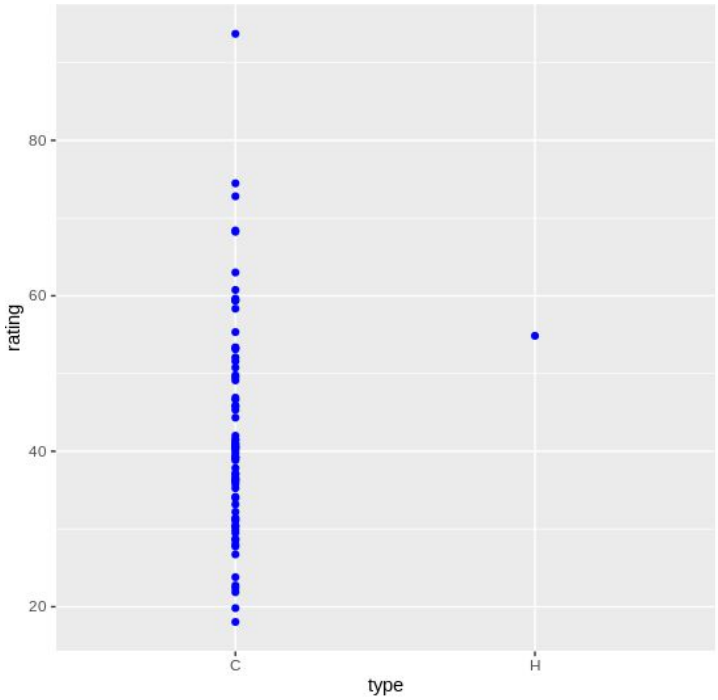
1) Fat, and sugar are positively correlated to calories
2) Calories, Sugar, and sodium are negatively correlated to rating
3) Fiber, protein, and potassium are positively correlated to rating.

Most of the weight is around 1 kgs. Even though there are not enough data points for other weights, it seems that rating is drops with increase in weight.




There is little to no correlation between carbohydrates and rating.




**Another point to note is that there aren't as many data points for type H.**

## Problem 2:

## a) Which predictors appear to have a significant relationship to the response.

Following variables seem to have a strong relationship to rating.

| | Est | Std. Err | T Value | Pr(>\|t\|) |
|---|---|---|---|---|
| **calories** | **-2.227e-01** | **8.329e-09** | **-2.674e+07** | **<2e-16 ***** |
| **protein** | **3.273e+00** | **5.906e-08** | **5.542e+07** | **<2e-16 ***** |
| **fat** | **-1.691e+00** | **9.123e-08** | **-1.854e+07** | **<2e-16 ***** |
| **sodium** | **-5.449e-02** | **6.639e-10** | **-8.208e+07** | **<2e-16 ***** |
| **fiber** | **3.443e+00** | **6.123e-08** | **5.624e+07** | **<2e-16 ***** |
| **carbo** | **1.092e+00** | **4.217e-08** | **2.591e+07** | **<2e-16 ***** |
| **sugars** | **-7.249e-01** | **3.855e-08** | **-1.880e+07** | **<2e-16 ***** |
| **potass** | **-3.399e-02** | **1.946e-09** | **-1.747e+07** | **<2e-16 ***** |
| **vitamins** | **-5.121e-02** | **2.079e-09** | **-2.464e+07** | **<2e-16 ***** |

## b) What does the coefficient variable for "sugar" suggest?

It suggests that sugar is strongly negatively correlated to rating.

## c) Use the * and : symbols to fit models with interactions. Are there any
## interactions that are significant?

The interactions between rating, calories, proteins, fat, sodium, fiber, carbo, sugars, potass, and vitamins are significant.

**Without any interactions:**
Residual standard error: 3.114e-07 on 55 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 8.235e+15 on 18 and 55 DF, p-value: < 2.2e-16

**Interaction 1 : calories:fat + all other features**
Residual standard error: 3.143e-07 on 54 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 7.661e+15 on 19 and 54 DF, p-value: < 2.2e-16

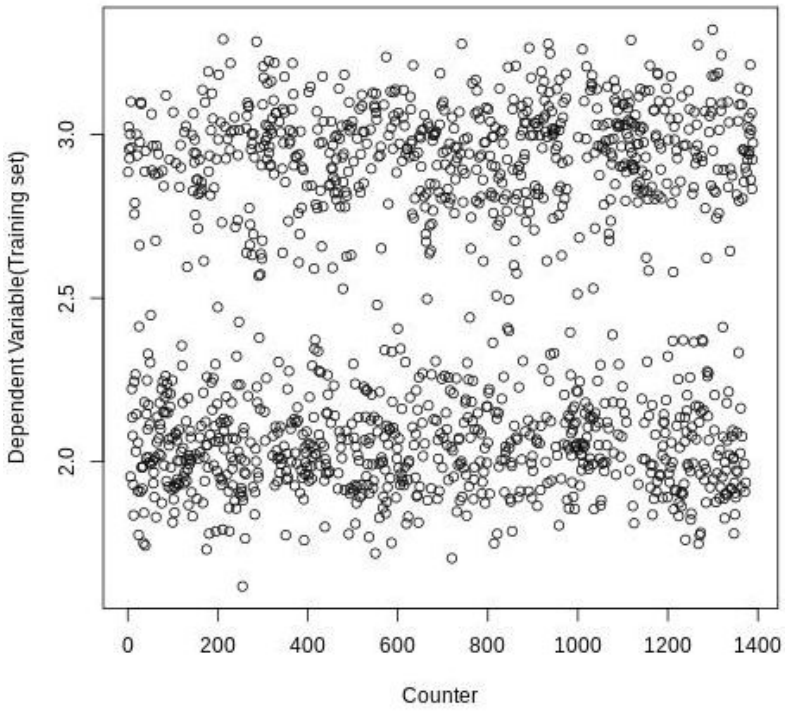
**Interaction 2 : sugars:calories + all other features**
Residual standard error: 3.123e-07 on 54 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 7.758e+15 on 19 and 54 DF, p-value: < 2.2e-16

## Problem 3:

**Classification performance of linear regression and k-nearest**
Multiple regression -
After the multiple regression was used to fit the training data, a classification mechanism was to be devised as the output was quantitative. In order to do so, I passed the training data into the model and produced the output. The output values had the following statistics-

| | |
|---|---|
| **Min =** 1.619 | **Mean=**2.474 |
| **1st Qu. =** 2.020 | **3rd Qu.=**2.953 |
| **Median =** 2.304 | **Max. =** 3.321 |

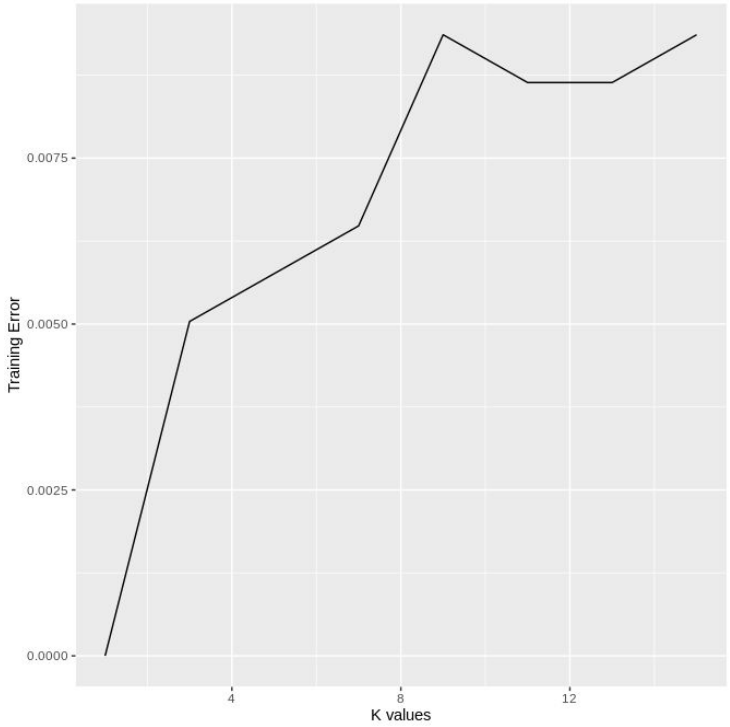
The following is the graph for quantitative predictions made on training data.




After looking at the graph, a value of 2.5 was chosen as a boundary separator. I classified all the points as 3 if quantitative prediction is above or equal to 2.5, else it will be classified as 2. On the training data, accuracy of 0.994 was obtained. On the test data, it was 0.958.

**K-Nearest Neighbours**

Here is a graph that shows errors in the training data for various K-values.




**Training Errors**

The error is almost 0 for k = 1, and increases slowly. The statistics for the training error in K-nearest neighbours is as follows.

| Min | 1st. Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|
| 0.00 | 0.0055 | 0.0075 | 0.0066 | 0.0088 | 0.0093 |

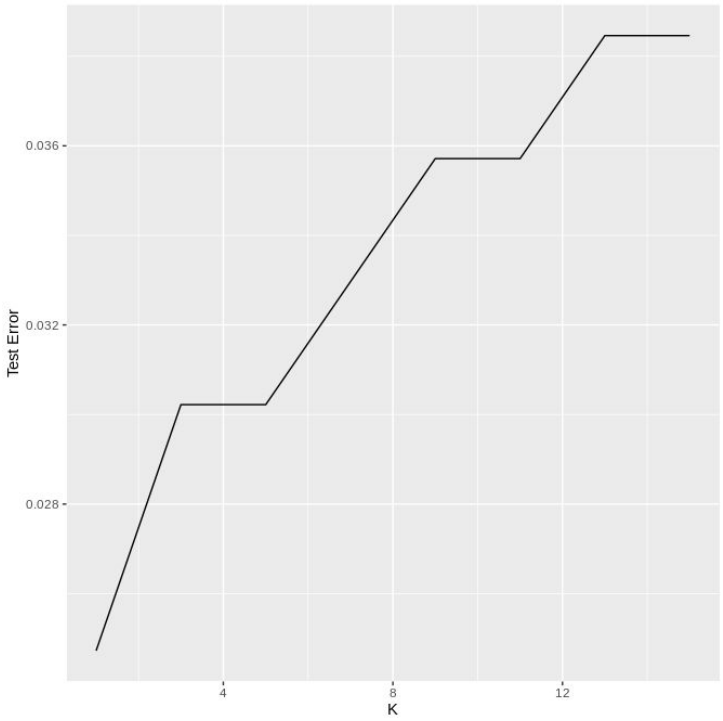
**For training Data**

Here are the statistics for the test data

| Min | 1st. Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|
| 0.024 | 0.0302 | 0.0343 | 0.033 | 0.036 | 0.0384 |

**For test data**

As expected, the test error is more than the training error. Increase in errors with an increase in K-values, is common to both, training and test data sets.
Below is a graph that shows errors for the test set.




**Test errors**

Finally following is the summary of test and training errors for every value of K.

| K | Test Error | Training Error |
|---|---|---|
| 1 | 0.024 | 0 |
| 3 | 0.0302 | 0.005 |
| 5 | 0.0302 | 0.0057 |
| 7 | 0.0329 | 0.0064 |
| 9 | 0.0357 | 0.0093 |
| 11 | 0.0357 | 0.0086 |
| 13 | 0.0384 | 0.0086 |
| 15 | 0.0384 | 0.0093 |

**Classification performance of linear regression and KNN**

The classification errors for multiple regression are as follows

| Test Set | Training Set |
|---|---|
| 0.042 | 0.006 |

According to the two tables shown above, KNN performs better or equal to Linear Regression for K values 1, 3, 5, and 7 on the training set. On the test set, KNN is always better for each of

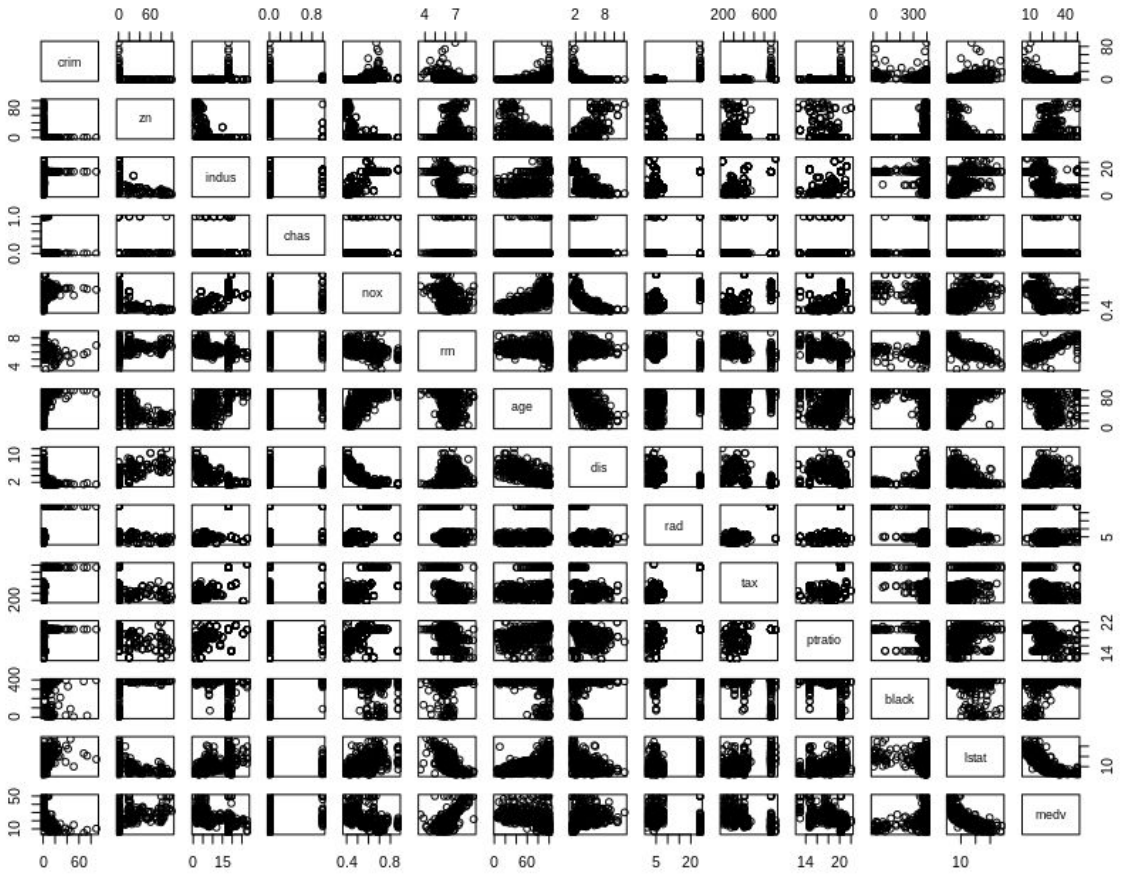
the K values(1, 3, 5, 7, 9, 11, 13, 15). **It is safe to conclude that KNN generalizes better than linear regression for small values of K.**

## Problem 4:

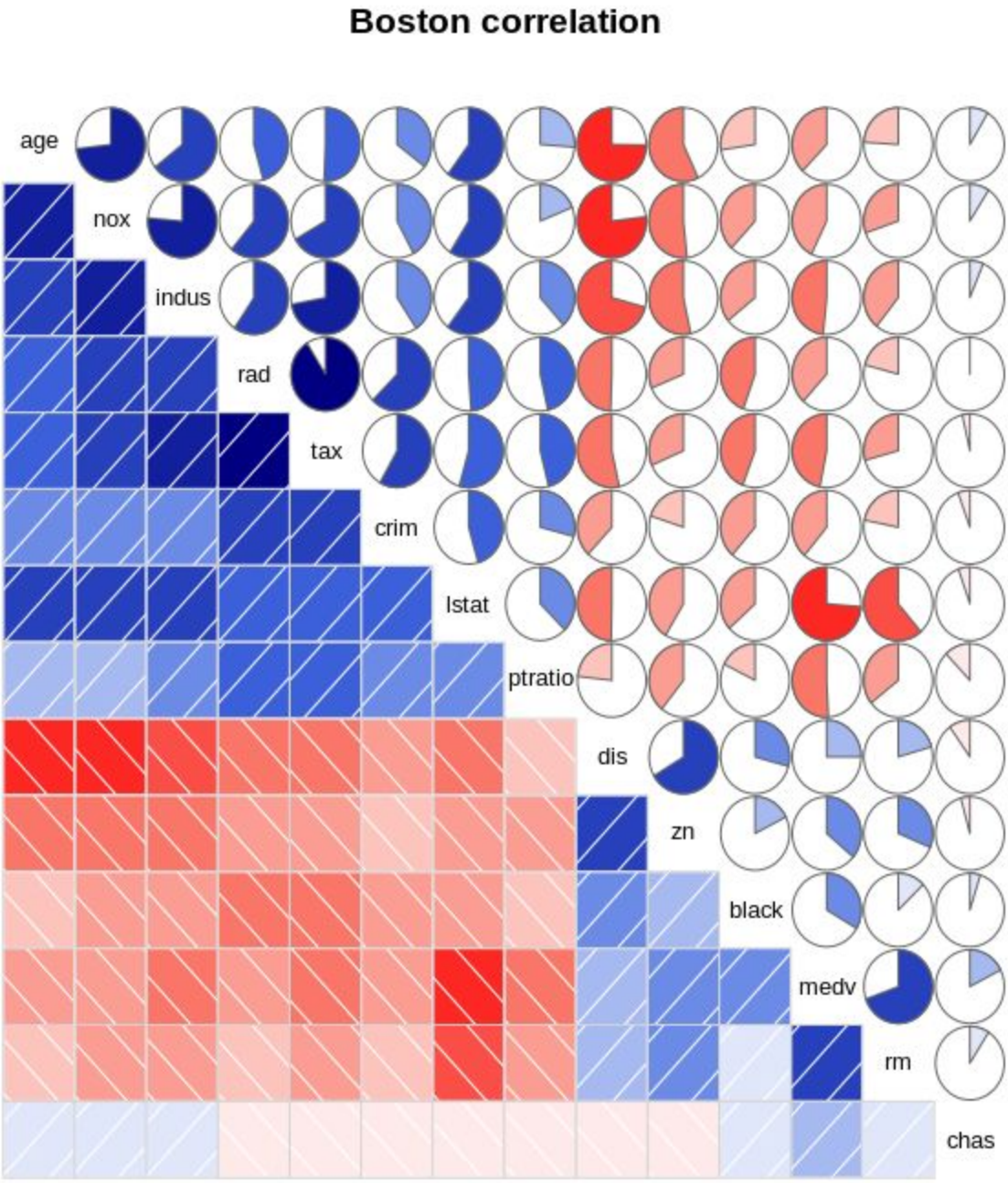
a) **Make pairwise scatterplots of the predictors, and describe your findings.**
b) **Are any of the predictors associated with per capita crime rate?**
    i) After fitting a linear model to the data, it was seen that crime rate does have some association with predictors.
c) **Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**
d) **In this data set, how many of the suburbs average more than seen rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.**

**Problem a)**

Here are the pairwise scatterplots of the predictors.




The graph is not clear enough to give a conclusion on the data, hence a more visual tool (Correlogram) has been used to understand correlations better.




Correlation matrix drawn using Correlogram. Blue sectors show a positive correlation, and red show negative correlation. Darker the circle, stronger is the correlation type.

Strong correlations have been addressed below:

1) Age is positively correlated with Nox - Hard to explain as Nitrogen oxides are majorly produced by cars. That might not have much to do with the old units. But, the correlation still needs to be noted.
2) Age is negatively correlated with distance from the employment centers as the employment centers are likely to be away from old house settlements.

3) Nox is positively correlated to prop. Of non-retail business acers. So, non retail businesses might be contributing to nox value in some way.
4) Nox is negatively correlated to dis. As the distance to employment centers decreases, the number of vehicles increase, and so does the nox value.
5) Indus(Prop. Of non-retail businesses) is positively related to tax. Maybe greater number of non-retail businesses are started to avoid taxes as they increase.
6) Indus is negatively correlated to dist- Maybe because owning property close to employment centers is expensive.
7) rad is positively related to taxes - Closeness to highways may lead to increase in taxes.
8) Crime is also slightly positively correlated to closeness to highways, and lower status population and negatively related to distance from employment units.
9) Lower status population is negatively correlated to median value of owner occupied homes and average rooms per dwelling.
10) Median value of owner occupied home is positively correlated to average rooms per dwelling.

**Sub-Problem b)**

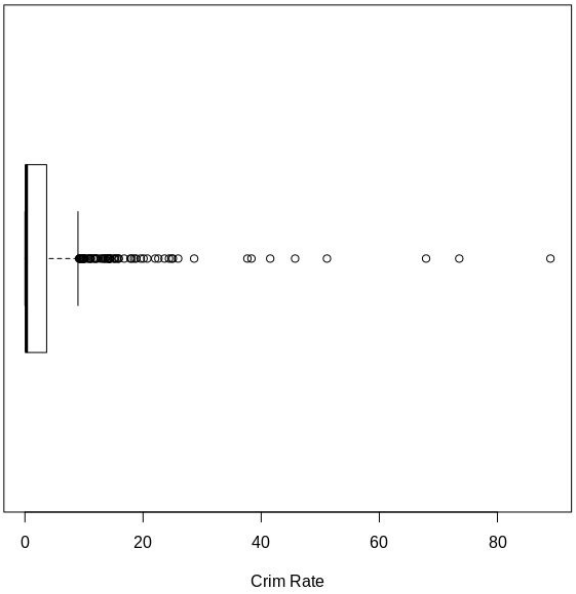
From the correlation matrix image drawn for the complete data, crime is correlated with the following features - rad, tax, and istat, dist, and medv. Here is the pairwise scatterplot. Positive correlation between crim and rad can be explained as closeness/accessibility to highways might lead to crimes. Positive correlation with tax is hard to explain. There is also a positive correlation with lower status population. Crime is negatively correlated with distance to employment centers, i.e. as the distance increases, crime decreases. It is also slightly negatively correlated to median value of owner occupied homes.

**Sub-problem c)**

i) The crime rate is highly skewed (Left). Here are the statistics for the crime rate.

| Min | 1st. Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|
| 0.00632 | 0.08204 | 0.256 | 3.61 | 3.67 | 88.97 |

Here is the box plot for crime.



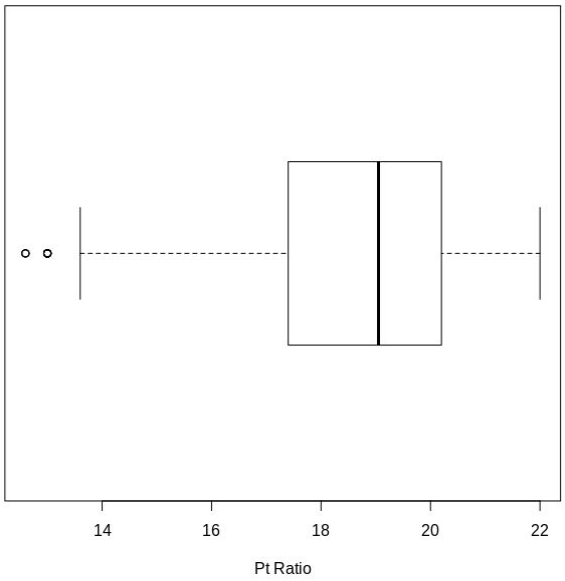

We will use 85th percentile as the dividing rage. So anything above 85% will be classified as high. There are 14 such suburbs. Following is the observation about other predictors. We divide the two subsets of data as sub1(>85%) and sub2(<=85%)

1) The number of non-retail business acers is higher in sub2 than in sub1.
2) The of number of rooms is less in sub2
3) Age of units is way heigher in sub2
4) Distance to employment units is way lesser in sub2
5) Index of accessibility to highways is also higher in sub2
6) Black population is lower in sub2
7) Lower status population is higher in sub2
8) Median value of owner occupied homes is less in sub2

ii) The ptratio is also slightly left skewed. Here are its ranges.

| Min | 1st. Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|
| 12.60 | 17.40 | 19.05 | 18.46 | 20.20 | 22.00 |

Here is its box plot.




In this case we consider 85% as higher. We have around 56 suburbs with a higher PT-ratio.
Again we subdivide data into two parts, sub1(ptratio <= 85%), and sub2(ptratio > 85%)
Here are the observations about other predictors

1) Crime is higher in sub1 than sub2
2) Age of units is higher in sub2
3) Access to radial highways is less for sub 2.
4) Tax are lesser is sub2
5) Lower status population is slightly greater in sub2
6) Median value of owner occupied homes is greater in sub1

iii) The distribution of tax is slightly skewed. Here are its statistics.

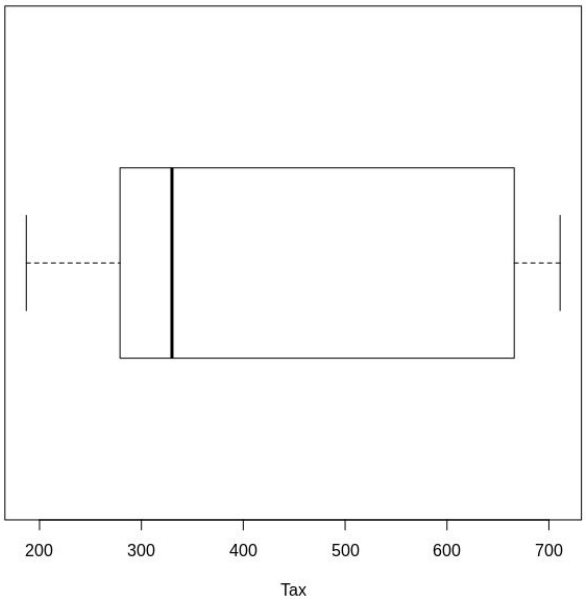
| Min | 1st. Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|
| 187.0 | 279.0 | 330.0 | 408.2 | 666.0 | 711.0 |

Assume sub1 <- tax <= 85%, sub2 <- tax > 85%

Observations about other predictors

1) Distance to employment centers is less in sub2
2) Accessibility to highways is slightly less in sub2
3) Median value of homes is also less in sub2

Here is the box plot for tax.




d) Suburbs with more than 7 rooms per dwelling 64. Suburbs with more than 8 rooms per dwelling 13. The statistics for suburbs with more than 8 rooms per dwelling are more or less the same as the rest of the data set.