STA506 | CALSS# 50 SIDDIQ SYED

UB#50291566

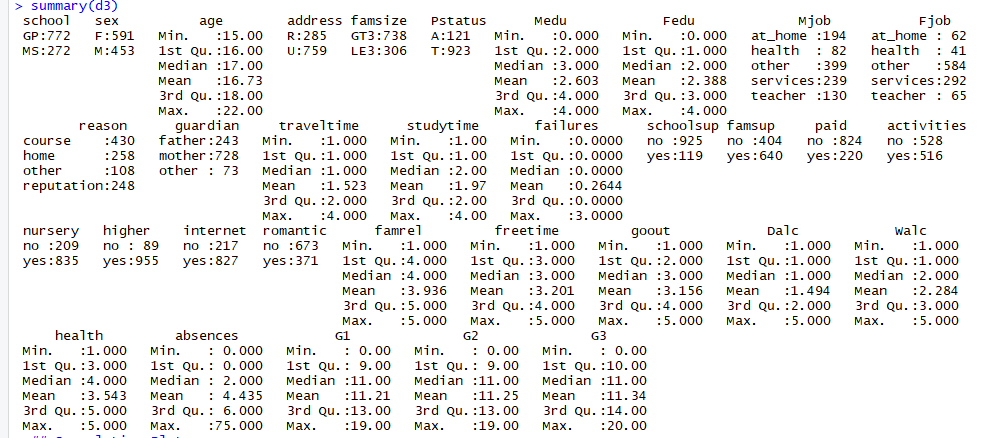
Homework-1

Question:1

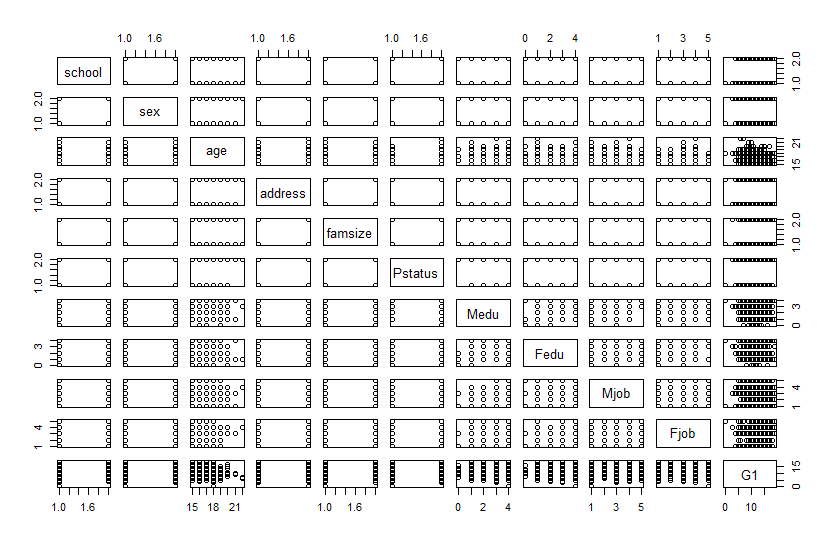
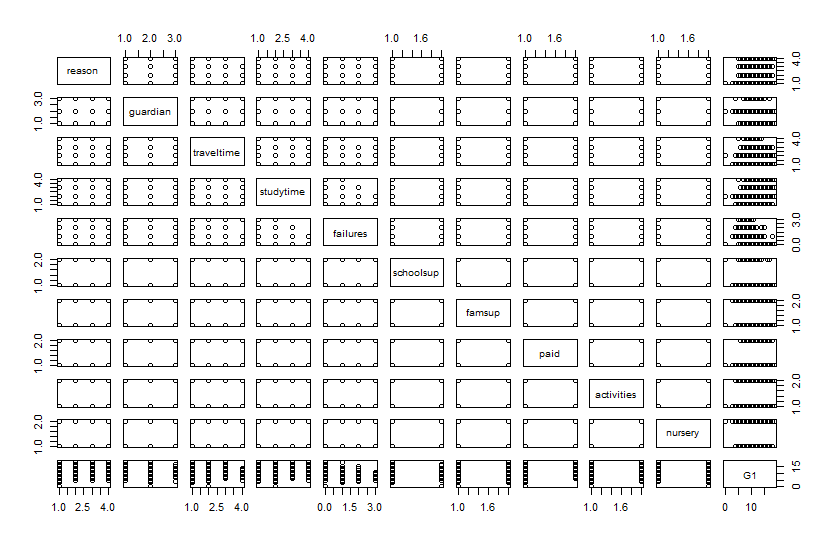
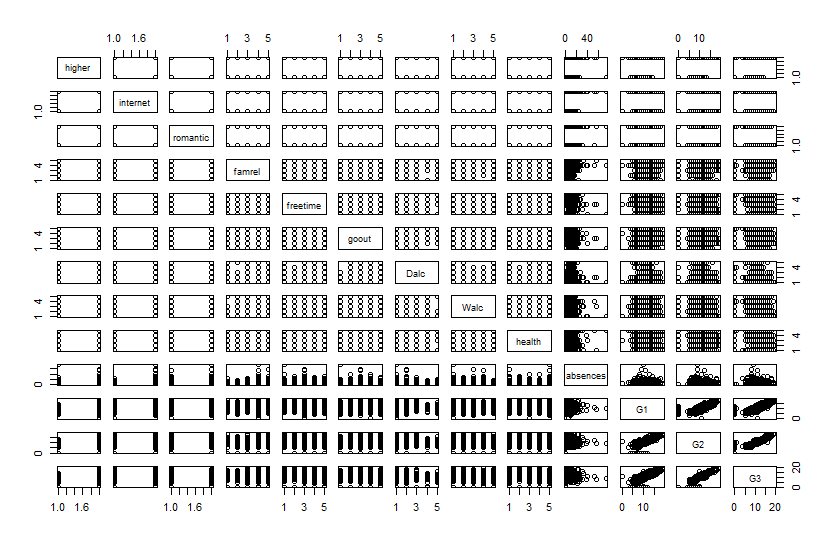
Data given is for two schools GP and MS. For the two given datasets. To analyse on both the given datasets consider rbind for the two datasets with the query below.

d3 <- do.call(rbind,list(d1,d2))

For the obtained dataset lets check the summary and analyse what it actually consists of.

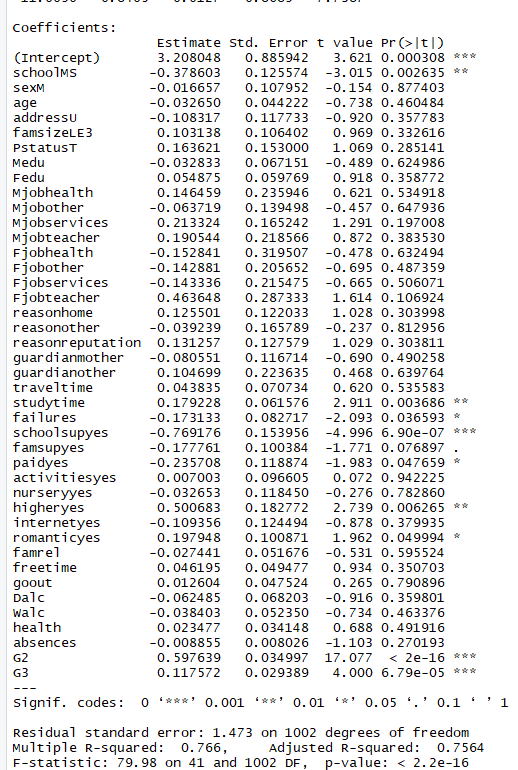


From the above summary the First Grade value is G1 varying from 0 to 19. Corresponding summary of the variables can be seen above. From which outliers looks like present for failures and absences.

Also find the scatter plots for the 33 variables given below with respect to G1. 

It can be infer that from the above plots G1 can be linearly dependent on absences and G2 and G2.

For further more analysis lets check the correlation and p values for the dataset with respect to G1.



The p value close to 0 can be inferred as the relatable values from linear regression methods. Upon which we can say that G1 can be dependent on School, address, pstatus, fedu, mjob, fjob, reason, studytime, failures, schoolsup, famsup, paid, higher, internet, romantic, freetime, dalc, absences, G2 and G3.

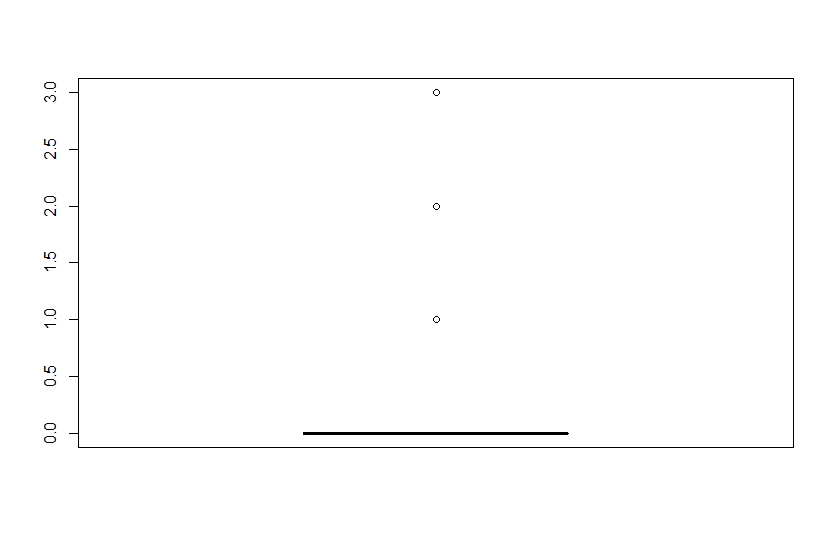
From summary of comparison from given field with respect to G1

1. Urban living students to perform better than rural area students. Summary -1 : Rural and Summary – 2: Urban

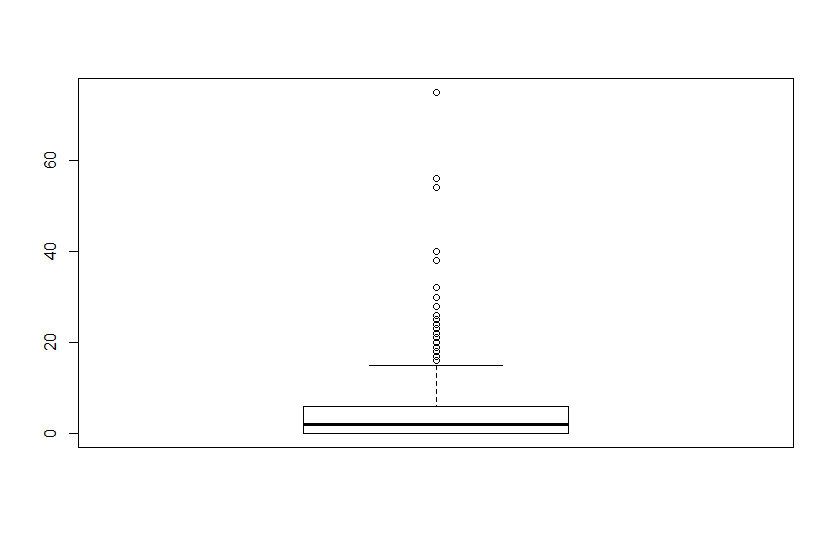
|  |
| --- |
| Min. 1st Qu. Median Mean 3rd Qu. Max.  3.00 8.00 10.00 10.66 13.00 19.00  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.00 9.00 11.00 11.42 14.00 19.00 |
|  |
| 1. Family size less than 3 tend to perform slightly better than greater than 3 family size.   summary(d3\_GT3$G1)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 9.0 11.0 11.1 13.0 19.0  summary(d3\_LE3$G1)  Min. 1st Qu. Median Mean 3rd Qu. Max.  3.00 9.25 11.00 11.48 14.00 19.00   |  | | --- | |  | |

Similar comparisons can be made for the G1 with respect to fields.

Also below is the boxplot for the absences and failures.



For absences

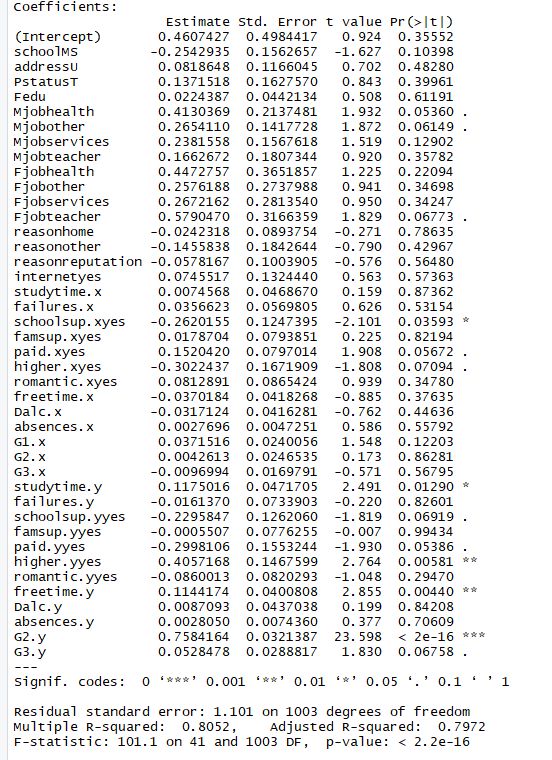


For absences

We can observe few outliners for the above two cases and these can associate with first grade score in an anomalous ways.

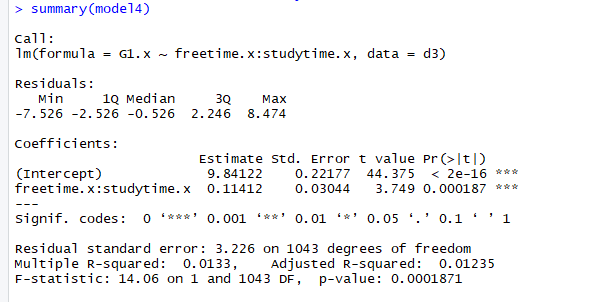
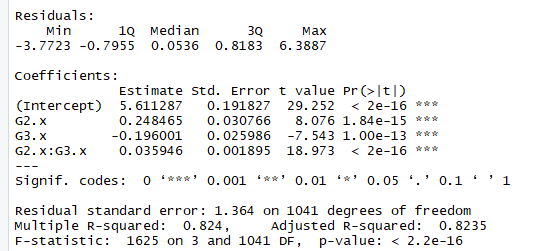
Question-2:

1. Using the lm function below summary for the lm models are obtained for G1.x and G2.y.



From which we can say that the predictors Mjob, Fjob, schoolsup, paid, higher, studytime and freetime will play a major role deciding G1 grades.

1. To achieve good grades in first year a student has to give more efforts to study by having greater study times. Also parents of the student need to be well educated. Apart from that extra educational support a student should have like library and friends to help while studying. Student need to have a good higher education. Student need to be funded well and during his free time quality work needs to be done to improve on his strengths and work to better his weaknesses.
2. Below are the two interaction models for which the significance can be explained as below.



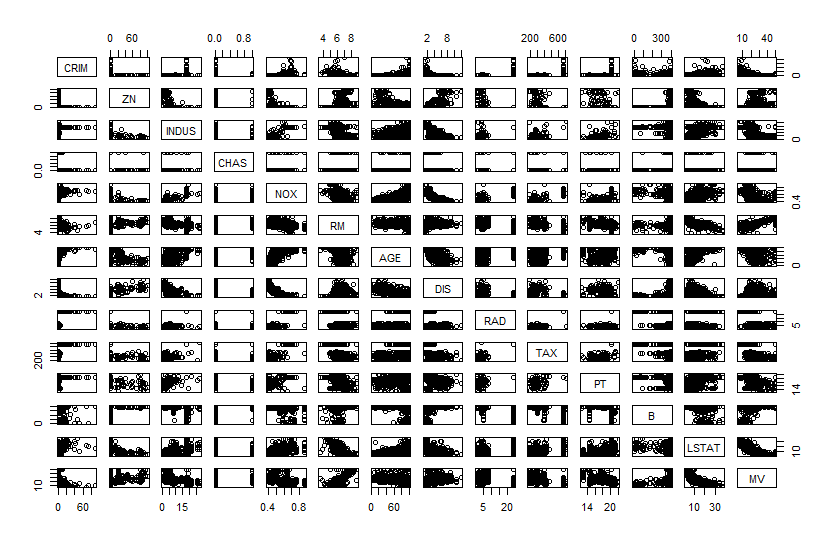
The \* act the product of the two predictors with respect to G1.x and also can be defined as the independent predictors G2.x and G3.x with respect to the G1.x.

Also from the plot it is evident that G2.x\*G3.x is pretty significant for interpreting G1.x.

Also “:” results in predictor 1:predictor 2 and for the above lm function the freetime.x:studytime.x is significant value for interpreting G1.x.

Question 3:

1. After loading the required libraries and Boston dataset, scatter plot has been obtained for all the 14 variables in the given dataset.



Scatter Plot for Boston dataset with given 14 variables.

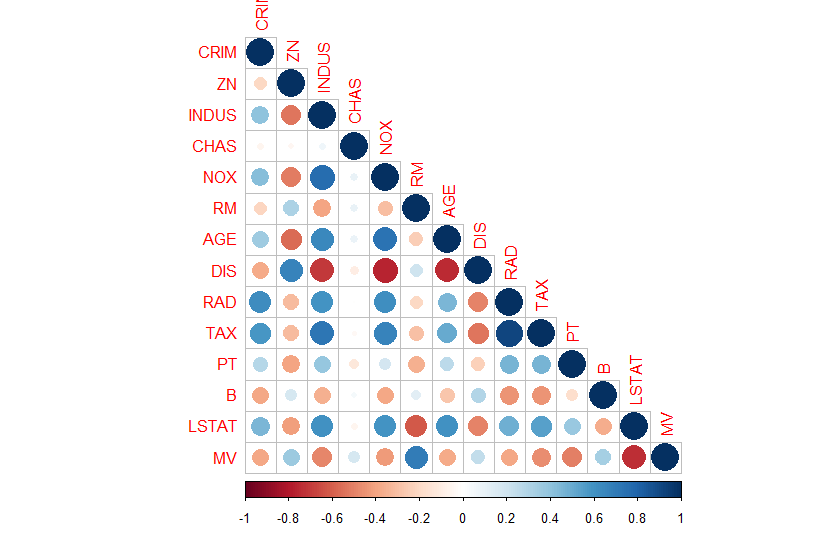
The pairs that have linear dependencies with positive or increasing trends are given below.

(AGE, CRIM), (RM,ZN), (DIS,ZN), (AGE,INDUS), (MV,ZN), (INDUS,NOX), (AGE,NOX), (PT,NOX), (LSTAT,NOX), (MV,RM), (B,DIS), (B,MV), (INDUS,LSTAT), (AGE,LSTAT), (PT,LSTAT).

Interaction with variable CHAS is mostly unrelatable to any of the other variable for the given data.

Similar cases with most of the interactions of the variables with RAD and TAX variables which have non linear interactions.

1. From the below correlation plot

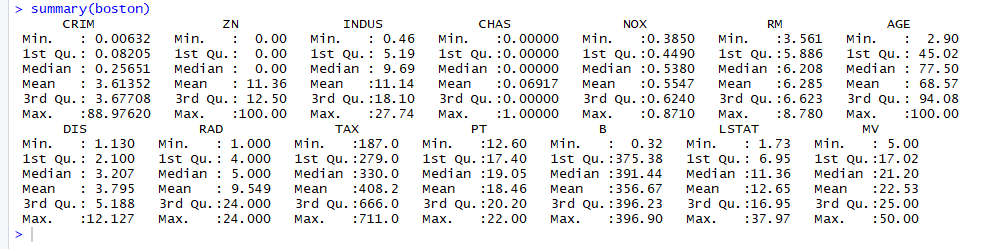


We can infer that CRIM is positively correlated with RAD and TAX which means that with increase in Index of accessibility to raidal highways i.e. houses close to highways are more prone to crime rate.

Also for TAX which is full value property tax rate per 10000$ i.e. as the cost of the house is more and which will imply to chances of frequent crime to be happening.

CRIM also has association with other predictors too.

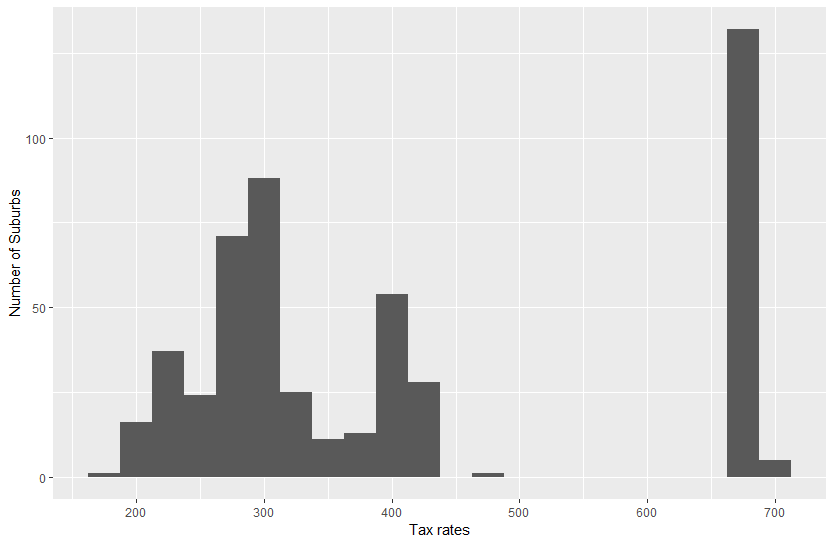
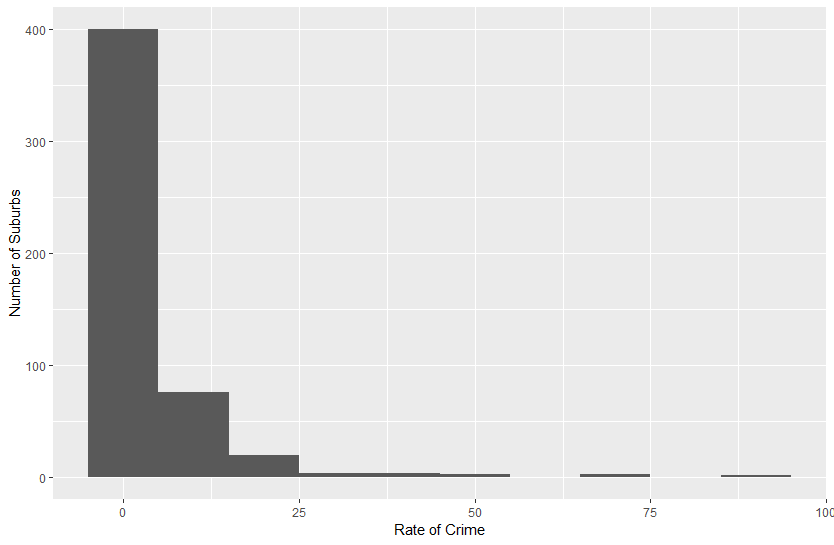
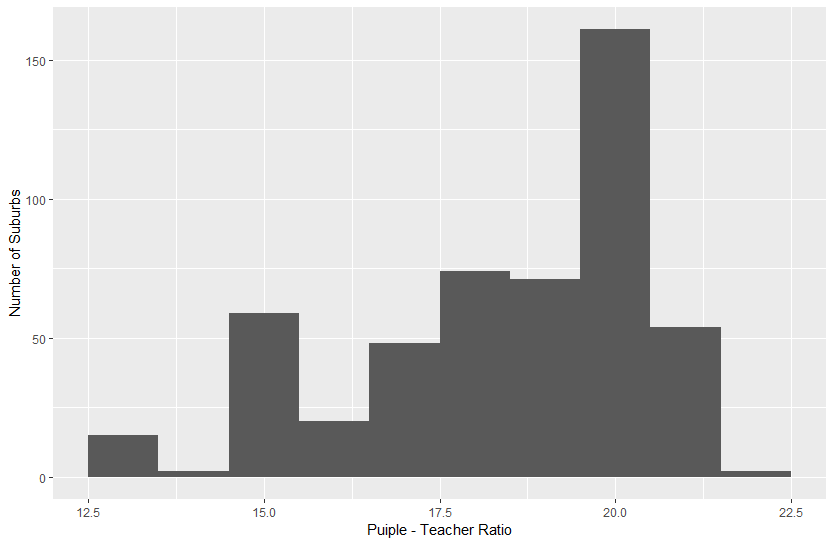
1. Below is the summary for the data



Selecting the subset for the CRIM data above 20 we get 3.6% of the 506 given housing data.

Selecting the subset for the CRIM data above 5 we get 21% of the 506 localities given.

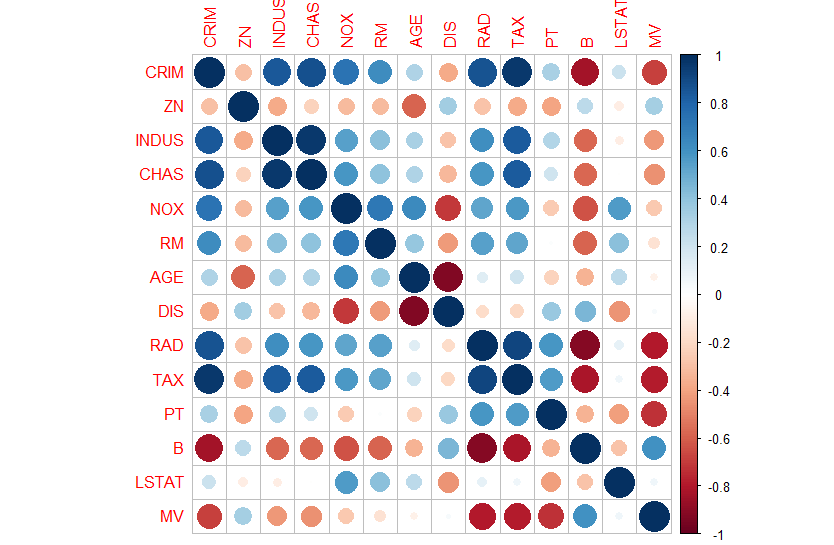
Also the TAX summary looks overrated as few of the housing are paying tax over 500$ and can have greater impact considering. Calculating the percentages we get 27% of the housing pay tax over 500$.

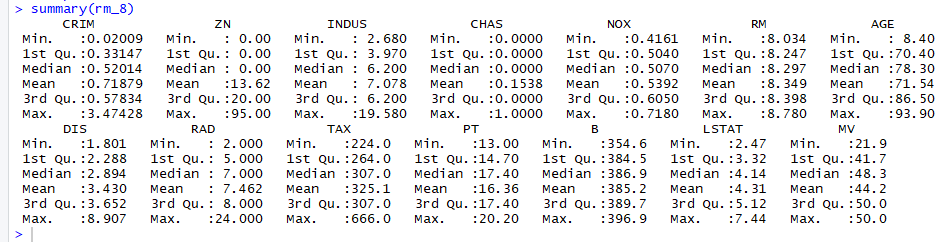


Above the plots shows the counts of the houses belonging to particular ranges for crime rates, taxes and pupil teacher ratio.

1. Calculating the number of rooms per dwelling having more than 7 and 8 rooms.

We get 64 above 7 and 13 above 8.

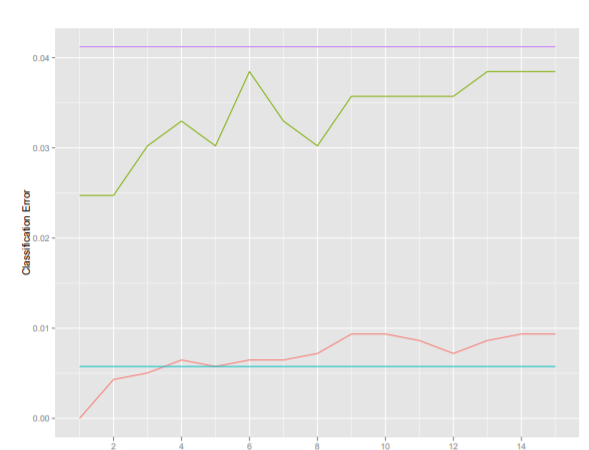




Above is the correlation and summary for the data having RM above 8.

Question 4:

The performance of the KNN classification and Linear regression can be compared with the obtained plot of the given code.



For all the values of K this can be checked and both training and test can be obtained from the R code given.