1. This problem is about predicting the model for Student Performance from UCI machine Learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance). And to investigate the data using exploratory data analysis with different plots, eliminate outliers, transform variable and eliminate variable.

Knowing my Data:

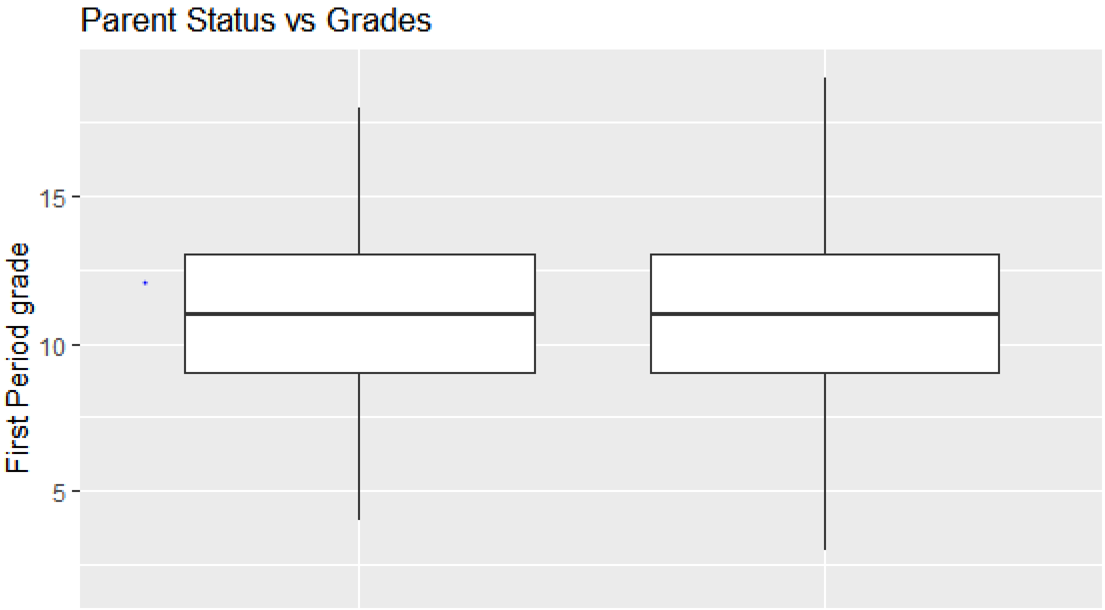
Dataset has 1044 rows and 33 variables most of the variables are categorical and First period grades for which we need to predict model.

Sub-setting the Data:

Based upon my intuition and some domain knowledge, I find some of the Data are irrelevant to my predictive model. So, I will pick few columns which I believe are the most important variables in the dataset. I am eliminating these columns, 11,12,17,18,19,20,23,24,25,27,28,32,33 from the Student Data set. Now the dimension will be 19 variables and 1044 columns.

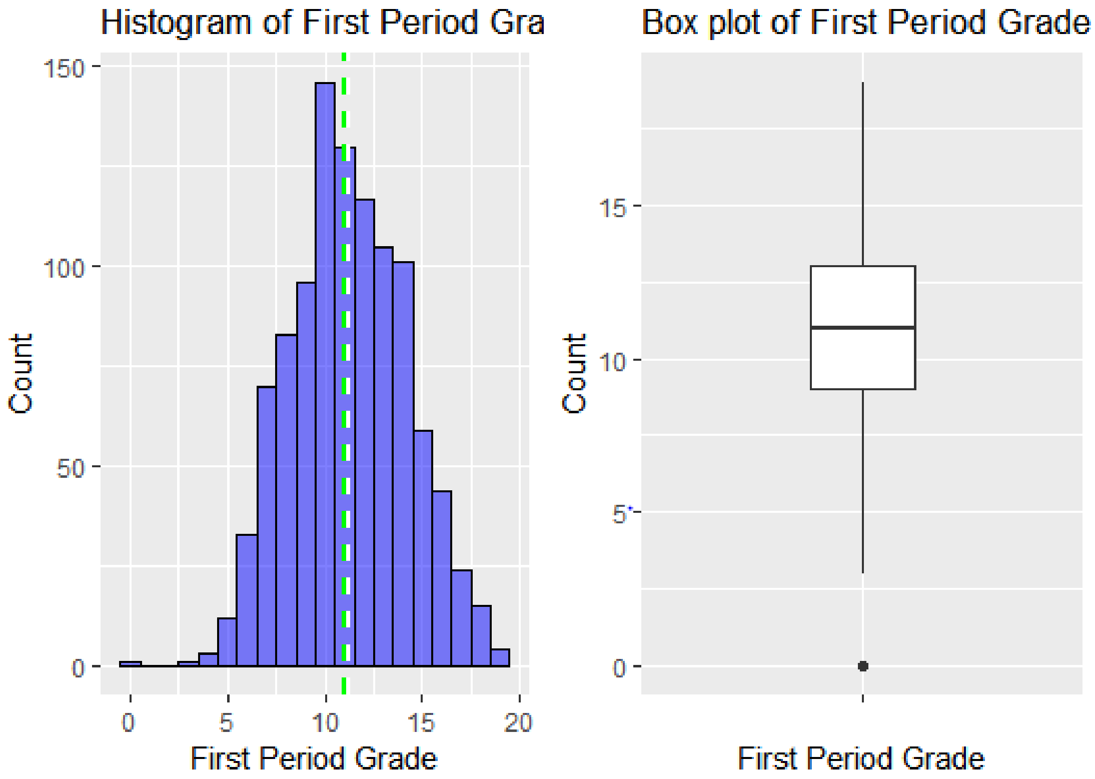
Analysing Data:

Based upon mean value between pstatus and first period grade what I found out was it pstatus does not have any effect on my predictor variable G1.since mean value for both cases are same .So, I eliminating pstatus from Data set.



Creating histogram for First Period Grade:

Histograms help us to understand how the data or a variable is distributed in the data set .I like to inspect the First Period Grade from my Dataset .So, I like to plot the histogram using ggplot2 library. From this plot we can infer that G1 is distributed the most between values 9 and 15.Green and yellow dashed line shows mean and median value.



From my understanding the histogram was left skewed mainly because of one person with zero grades. So I consider this an outlier and eliminated the corresponding row and this can also be shown with the help of the box plot too which shows clearly there was an outlier in the distribution of histogram. The summary of the student data is given by

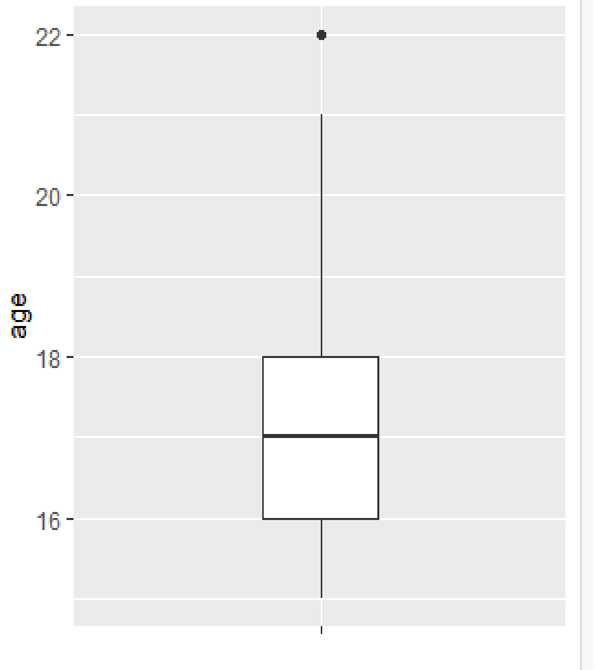
```
school   sex          age        address Pstatus      Medu            Fedu
Mjob           Fjob
 GP:772   F:591   Min.   :15.00   R:285   A:121   Min.   :0.000   Min.
:0.000   at_home :194   at_home : 62
 MS:272   M:453   1st Qu.:16.00   U:759   T:923   1st Qu.:2.000   1st
Qu.:1.000   health  : 82   health  : 41
                  Median :17.00                   Median :3.000   Median
:2.000   other   :399   other   :584
                  Mean   :16.73                   Mean   :2.603   Mean
:2.388   services:239   services:292
                  3rd Qu.:18.00                   3rd Qu.:4.000   3rd
Qu.:3.000   teacher :130   teacher : 65
                  Max.   :22.00                   Max.   :4.000   Max.
:4.000
   traveltime      studytime       failures      schoolsup higher    internet
goout           health
 Min.   :1.000   Min.   :1.00   Min.   :0.0000   no :925   no : 89   no :217
Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.0000   yes:119   yes:955   yes:827
1st Qu.:2.000   1st Qu.:3.000
```

```
 Median :1.000   Median :2.00   Median :0.0000  
Median :3.000   Median :4.000  
 Mean   :1.523   Mean   :1.97   Mean   :0.2644  
Mean   :3.156   Mean   :3.543  
 3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:0.0000  
3rd Qu.:4.000   3rd Qu.:5.000  
 Max.   :4.000   Max.   :4.00   Max.   :3.0000  
Max.   :5.000   Max.   :5.000  
    absences            G1       
 Min.   : 0.000   Min.   : 0.00  
 1st Qu.: 0.000   1st Qu.: 9.00  
 Median : 2.000   Median :11.00  
 Mean   : 4.435   Mean   :11.21  
 3rd Qu.: 6.000   3rd Qu.:13.00  
 Max.   :75.000   Max.   :19.00  
```

With the summary, we shall go ahead and plot the variables in pairs to understand how each variable is affecting the other.

Eliminating other outliers:

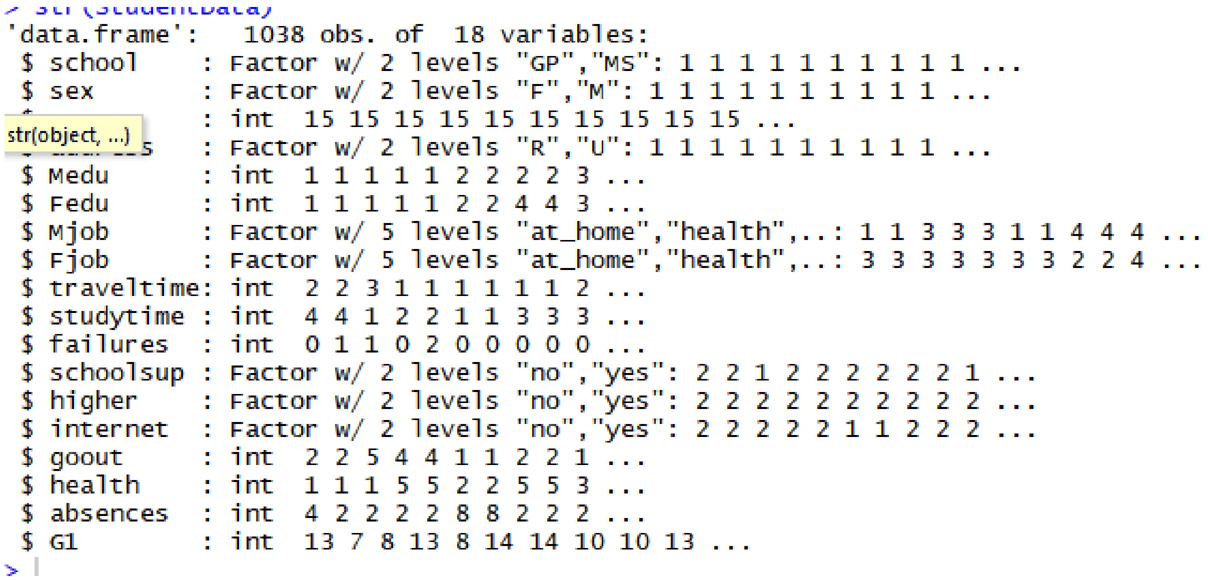
Now plotting the box plot for the age column I found out that there is outliers in age column which can be clearly seen from the box plot.



Which can be found from the function which I had used from the code to remove outliers if any from the data set? So it will remove corresponding points from the data set.

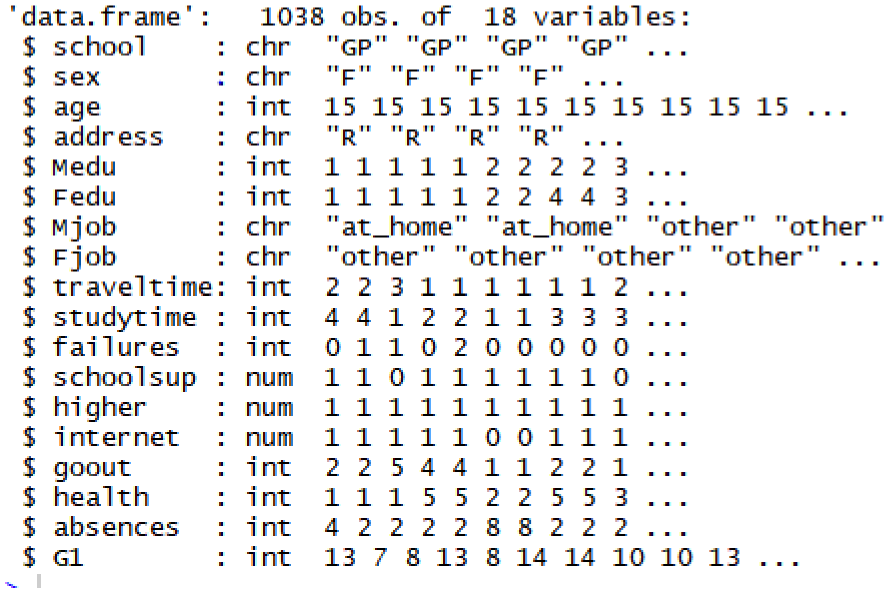
Variable transformation:

Most of the data frame consider factors as string so it is easier to convert into binary value.Since it is easier to work with 0’s and 1’s instead of “YES” and “NO” I had used function to transform factors into binary values.



```
> str(StudentData)
'data.frame':   1038 obs. of  18 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 str(object, ...)  : int  15 15 15 15 15 15 15 15 15 15 ...
           s : Factor w/ 2 levels "R","U": 1 1 1 1 1 1 1 1 1 1 ...
 $ Medu      : int  1 1 1 1 1 2 2 2 2 3 ...
 $ Fedu      : int  1 1 1 1 1 2 2 4 4 3 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 3 3 3 1 1 4 4 4 ...
 $ Fjob      : Factor w/ 5 levels "at_home","health",..: 3 3 3 3 3 3 3 2 2 4 ...
 $ traveltime: int  2 2 3 1 1 1 1 1 1 2 ...
 $ studytime : int  4 4 1 2 2 1 1 3 3 3 ...
 $ failures  : int  0 1 1 0 2 0 0 0 0 0 ...
 $ schoolsup : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 2 2 1 ...
 $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 1 1 2 2 2 ...
 $ goout     : int  2 2 5 4 4 1 1 2 2 1 ...
 $ health    : int  1 1 1 5 5 2 2 5 5 3 ...
 $ absences  : int  4 2 2 2 2 8 8 2 2 2 ...
 $ G1        : int  13 7 8 13 8 14 14 10 10 13 ...
>
```

Data set before transformation

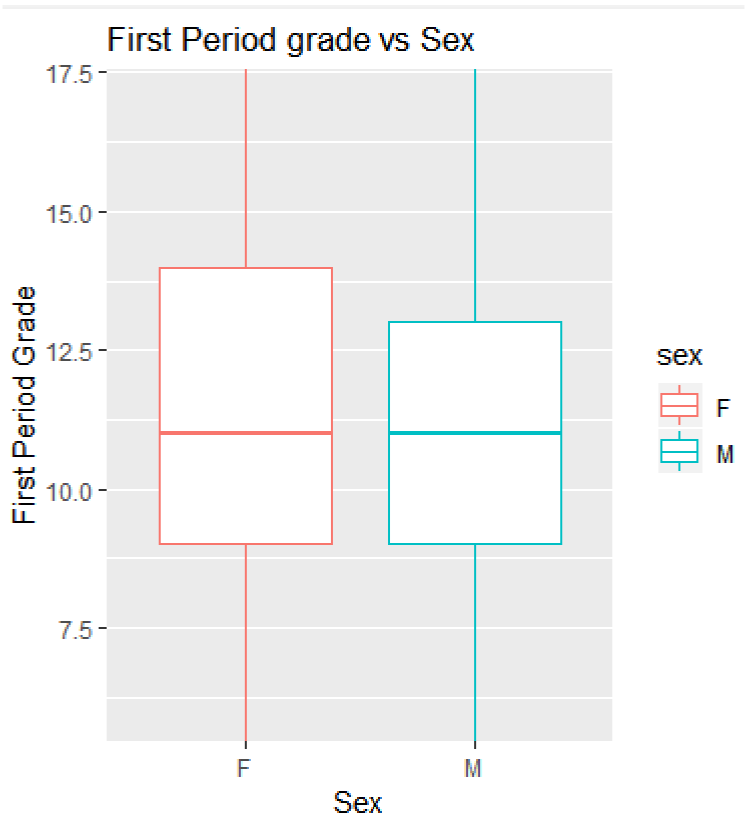


```
'data.frame':   1038 obs. of  18 variables:
 $ school    : chr  "GP" "GP" "GP" "GP" ...
 $ sex       : chr  "F" "F" "F" "F" ...
 $ age       : int  15 15 15 15 15 15 15 15 15 15 ...
 $ address   : chr  "R" "R" "R" "R" ...
 $ Medu      : int  1 1 1 1 1 2 2 2 2 3 ...
 $ Fedu      : int  1 1 1 1 1 2 2 4 4 3 ...
 $ Mjob      : chr  "at_home" "at_home" "other" "other"
 $ Fjob      : chr  "other" "other" "other" "other" ...
 $ traveltime: int  2 2 3 1 1 1 1 1 1 2 ...
 $ studytime : int  4 4 1 2 2 1 1 3 3 3 ...
 $ failures  : int  0 1 1 0 2 0 0 0 0 0 ...
 $ schoolsup : num  1 1 0 1 1 1 1 1 1 0 ...
 $ higher    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ internet  : num  1 1 1 1 1 0 0 1 1 1 ...
 $ goout     : int  2 2 5 4 4 1 1 2 2 1 ...
 $ health    : int  1 1 1 5 5 2 2 5 5 3 ...
 $ absences  : int  4 2 2 2 2 8 8 2 2 2 ...
 $ G1        : int  13 7 8 13 8 14 14 10 10 13 ...
```

Data set after transformation

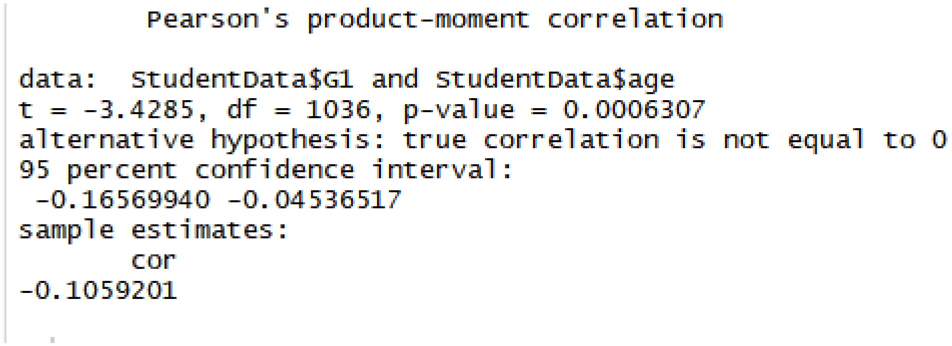
<u>Relation between First Period Grades and Sex</u>

By using aggregating it found that sex ratio has small effect on First period variable I decided to keep the variable for prediction



Relation between First Period Grades and age:

By finding the correlation between G1 and age it was clear that age has no effect on the G1.So I had eliminated the age variable from Data set



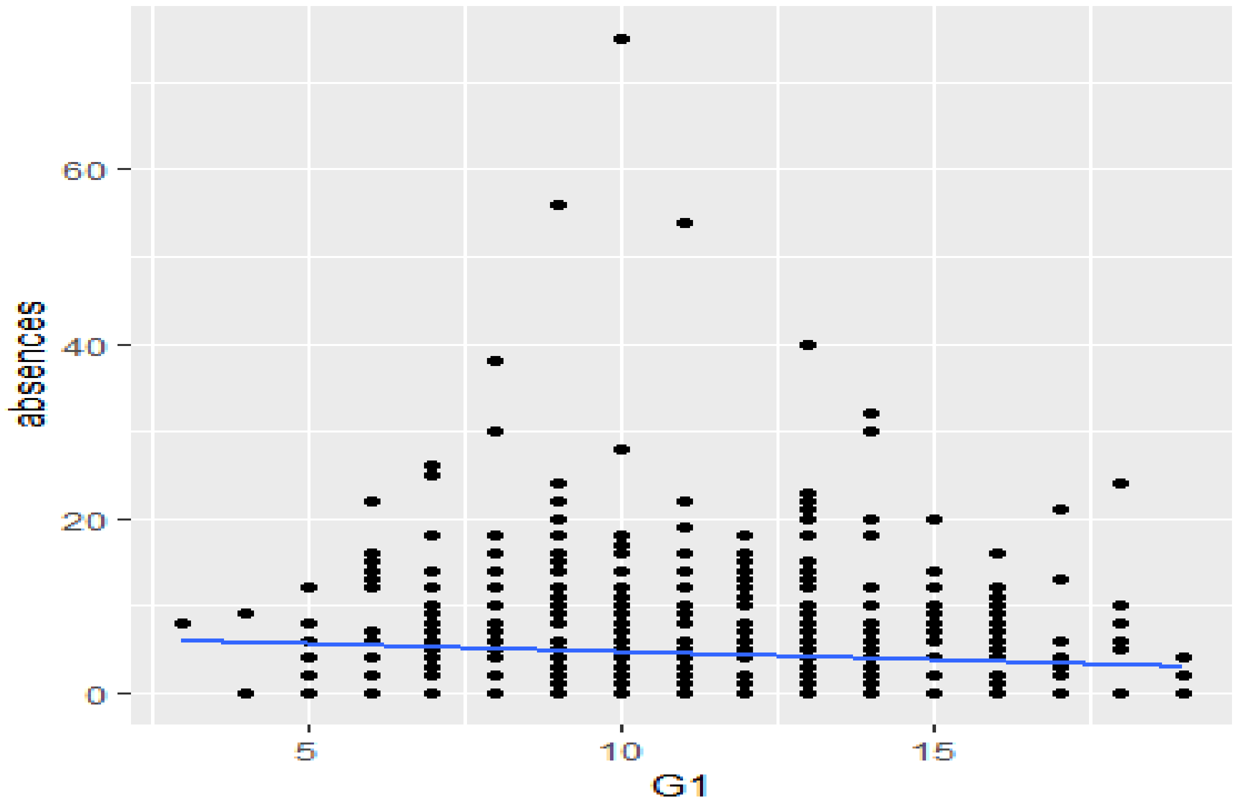
```
	Pearson's product-moment correlation

data:  StudentData$G1 and StudentData$age
t = -3.4285, df = 1036, p-value = 0.0006307
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.16569940 -0.04536517
sample estimates:
       cor 
-0.1059201 
```

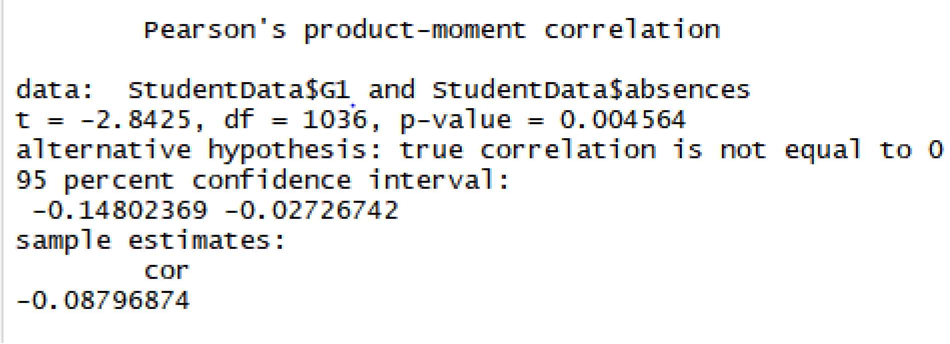
Relation between First Period Grades and father and mother education:

By analysing between aggregate of mean of both father and mother education with respect to grade I had found out that father education does not have any effect on G1 and mother education has some effect on G1.I had decided to eliminate father education from the data set.

Relation between First Period Grades and absences:

Since First Period Grade and absences have no effect on each other which was evident from the scatter linear regression line. So I had decided to eliminate absences column from the data set. And also the correlation factor is very low .

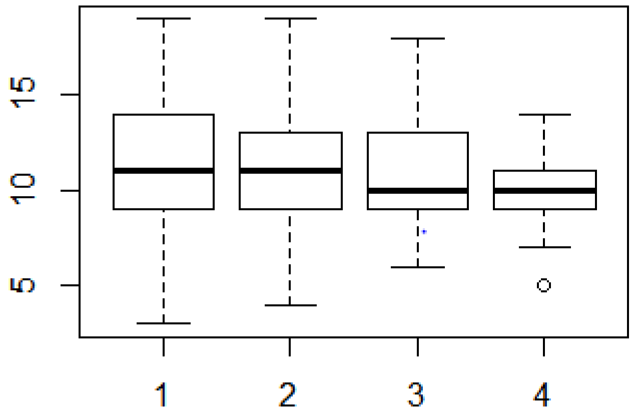




```
        Pearson's product-moment correlation

data:  StudentData$G1 and StudentData$absences
t = -2.8425, df = 1036, p-value = 0.004564
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14802369 -0.02726742
sample estimates:
        cor
-0.08796874
```

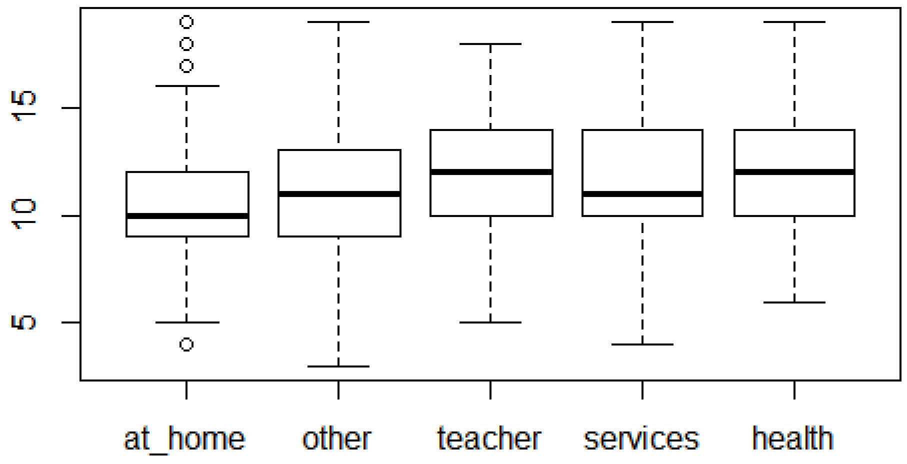
Relation between travel time and Grades:

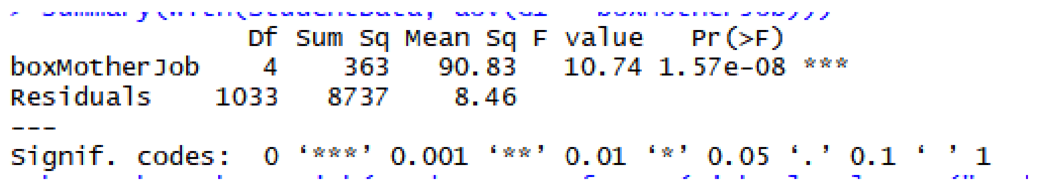
It is evident from the box plot that the travel time and corresponding values non-linear and correlation factor is low I decide to eliminate travel time column from the dataset



Relation between Mothers Job and Grades:

It is clear from the box plot that the mother who is working in health sector or working as a teacher has a significant effect on First period Grade. It can be shown by both plot and by finding the value of the correlation

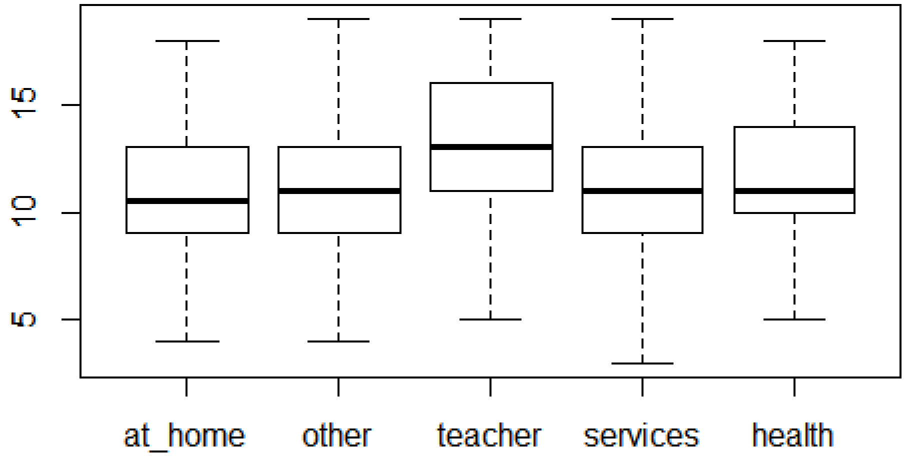


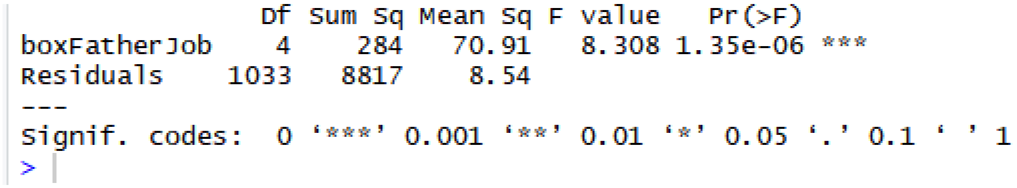


```
               Df Sum Sq Mean Sq F value   Pr(>F)    
boxMotherJob    4    363   90.83   10.74 1.57e-08 ***
Residuals    1033   8737    8.46                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Relation between Father Job and Grades:

It is clear from the box plot that the father who is working as a teacher has a significant effect on First period Grade. It can be shown by both plot and by finding the value of the correlation between them





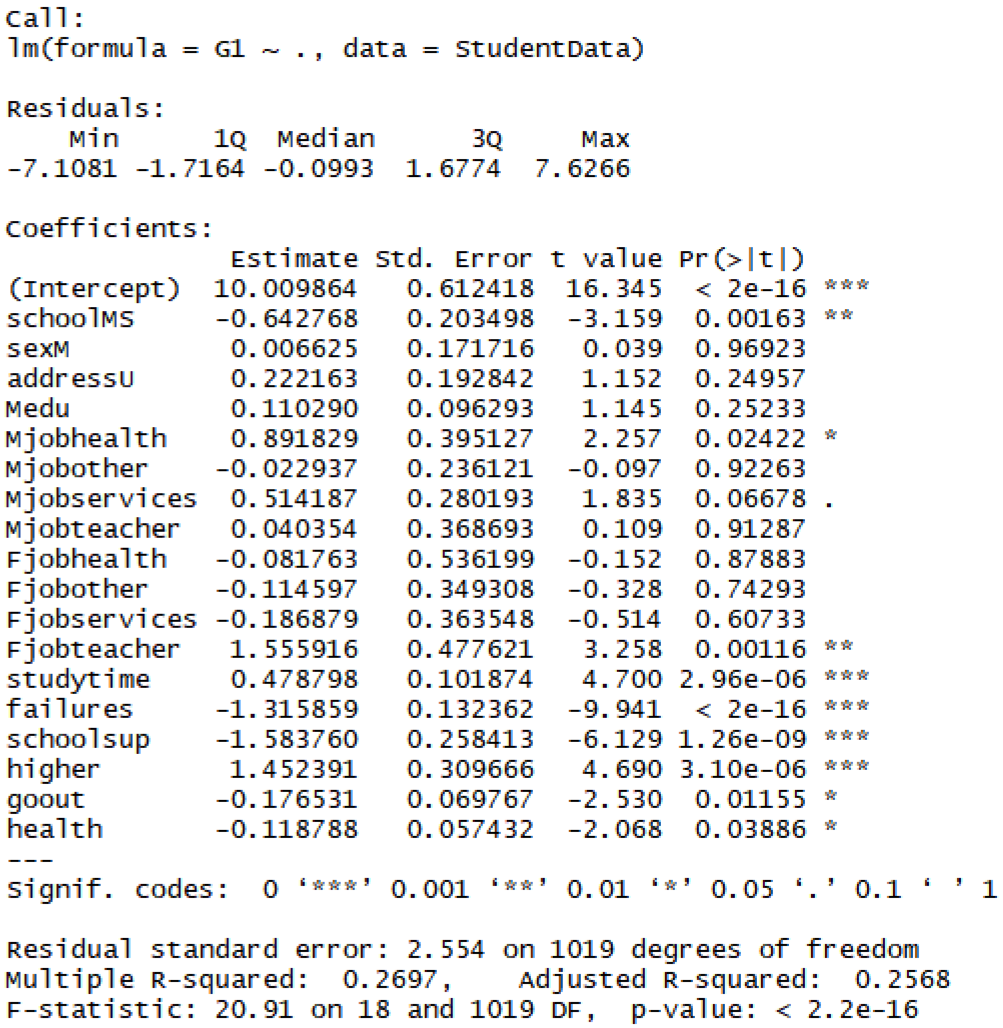
```
             Df Sum Sq Mean Sq F value   Pr(>F)    
boxFatherJob    4    284   70.91   8.308 1.35e-06 ***
Residuals    1033   8817    8.54                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

So finally, I will be having 1038 rows by eliminating outliers and 13 variables by removing the variables which does not have significant effect on my response. So I will be using this pre-processed data in the question

2. This question is about predicting the model for the response variable which is First Period Grades. Based upon the value got from the question1

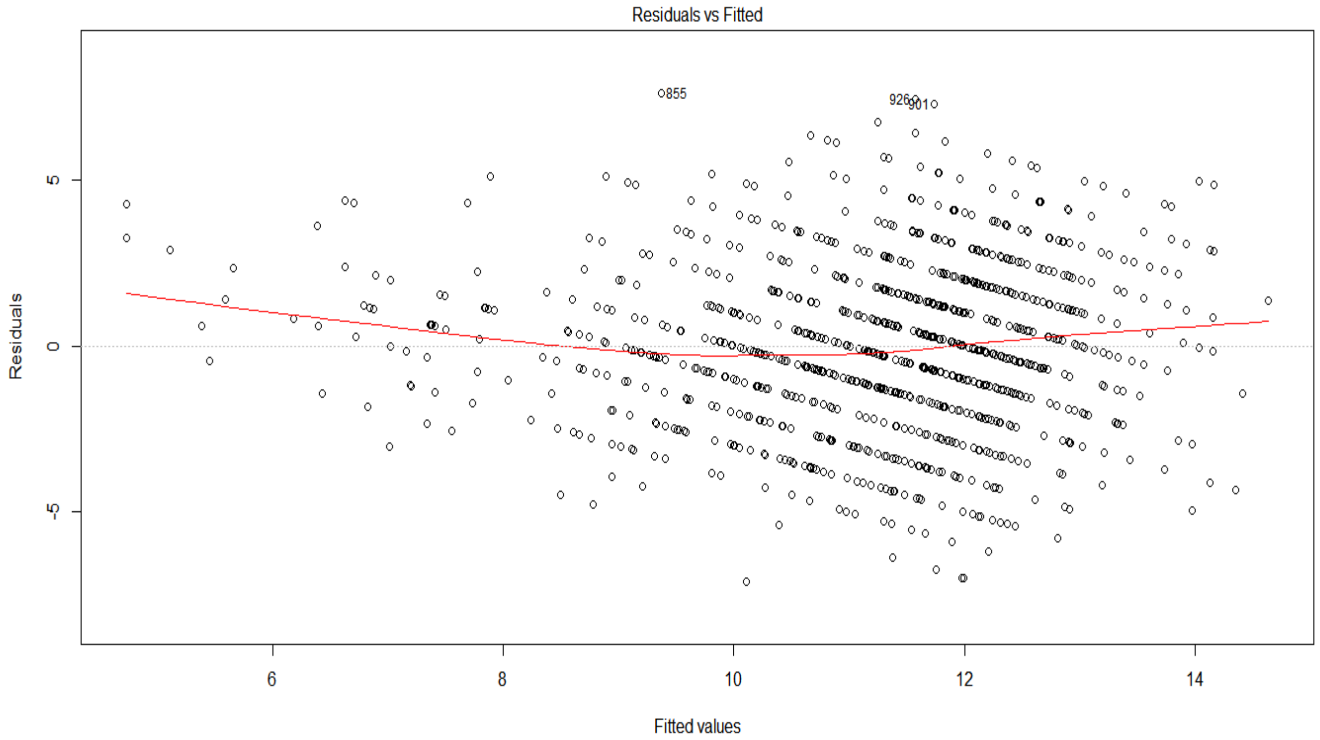
In this I had used the lm() to find the multiple linear regression

<u>Summary of Linear Model:</u>



```
Call:
lm(formula = G1 ~ ., data = StudentData)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1081 -1.7164 -0.0993  1.6774  7.6266 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.009864   0.612418  16.345  < 2e-16 ***
schoolMS     -0.642768   0.203498  -3.159  0.00163 ** 
sexM          0.006625   0.171716   0.039  0.96923    
addressU      0.222163   0.192842   1.152  0.24957    
Medu          0.110290   0.096293   1.145  0.25233    
Mjobhealth    0.891829   0.395127   2.257  0.02422 *  
Mjobother    -0.022937   0.236121  -0.097  0.92263    
Mjobservices  0.514187   0.280193   1.835  0.06678 .  
Mjobteacher   0.040354   0.368693   0.109  0.91287    
Fjobhealth   -0.081763   0.536199  -0.152  0.87883    
Fjobother    -0.114597   0.349308  -0.328  0.74293    
Fjobservices -0.186879   0.363548  -0.514  0.60733    
Fjobteacher   1.555916   0.477621   3.258  0.00116 ** 
studytime     0.478798   0.101874   4.700 2.96e-06 ***
failures     -1.315859   0.132362  -9.941  < 2e-16 ***
schoolsup    -1.583760   0.258413  -6.129 1.26e-09 ***
higher        1.452391   0.309666   4.690 3.10e-06 ***
goout        -0.176531   0.069767  -2.530  0.01155 *  
health       -0.118788   0.057432  -2.068  0.03886 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.554 on 1019 degrees of freedom
Multiple R-squared:  0.2697,	Adjusted R-squared:  0.2568 
F-statistic: 20.91 on 18 and 1019 DF,  p-value: < 2.2e-16
```

The below graph shows the relation between Residuals and Fitted values G1.Where the linear line is fitted towards the correlated variables and linear regression line is drawn and sum of the residuals squares are predicted. The Graph shows the non-linear response and the predictors



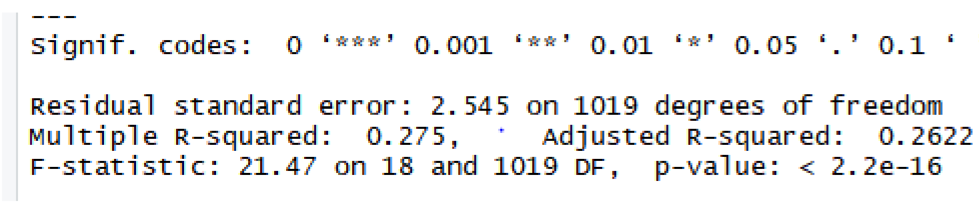
A) It was clear from the summary of the linear regression model that some of the significant predictors are **studytime, failures ,schoolups, higher** from this failures and schoolups are negatively correlated whereas studytime and higher are positively correlated and some of the other predictors for the output variable are Father job as teacher, school as MS, goout, health, Mother job as health among this Mother Job as health and Father Job as Teacher has positively correlated and goout and school as MS are negatively corelated

B) Suggestions I like to give was

- First factor to consider that is School, If you study in school GP other than MS your score can probably can go high compared to students from other schools.
- Second, Study time plays an important role in your grades. If your study time is more then you will get good grades in your exam
- Third, if you have any higher education plans then you will be having an interest towards the subjects and you will like spend more time on your studies, so your grades will go up.

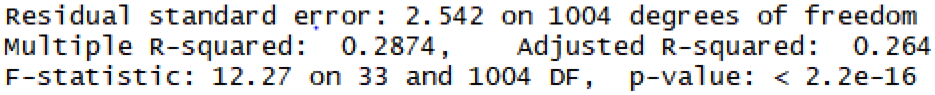
- Fourth ,If you have any past arrears then you need to concentrate more .since your are finding difficult in understanding .You need work on your free time
- Fifth, In case of extra educational support you will be confused with the teaching in regular class and tuition class. And also you need some other activities to get relaxed and concentrate more when you’re studying in regular school. So avoid extra education

C) : and * gives the response based upon the interaction between two predictors
When each output variable interact with higher for both: and * symbols residual standard error had decreased and Multiple R-square and adjusted R-square value had increased



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.545 on 1019 degrees of freedom
Multiple R-squared:  0.275,     Adjusted R-squared:  0.2622
F-statistic: 21.47 on 18 and 1019 DF,  p-value: < 2.2e-16
```

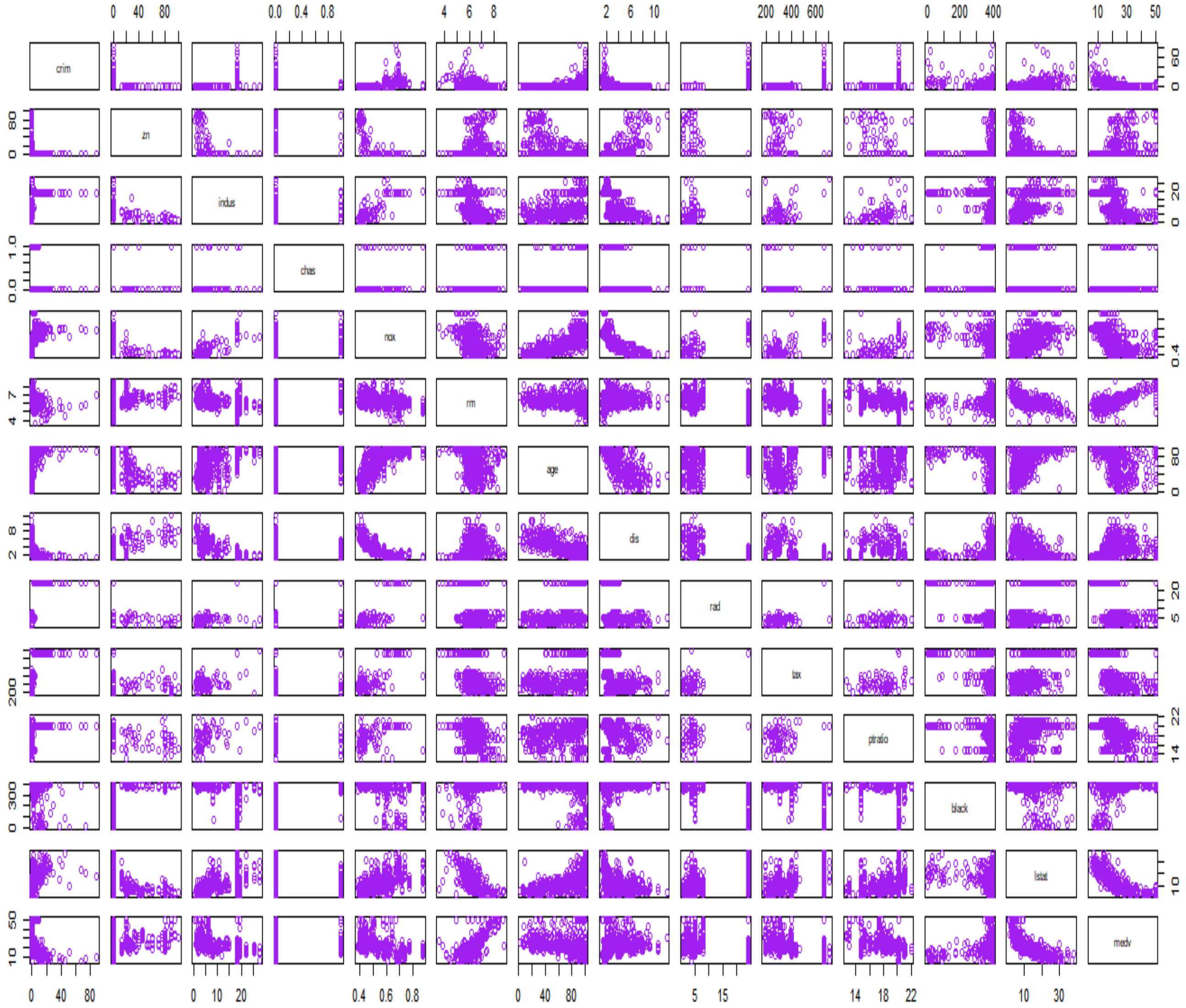
When interaction of other column with higher has increased R-squared and Multiple R-squared when using “:” symbol



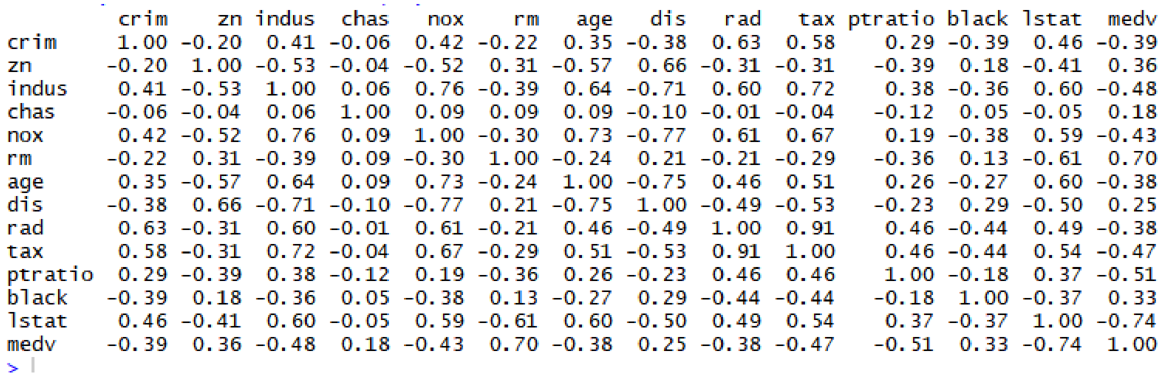
```
Residual standard error: 2.542 on 1004 degrees of freedom
Multiple R-squared:  0.2874,    Adjusted R-squared:  0.264
F-statistic: 12.27 on 33 and 1004 DF,  p-value: < 2.2e-16
```

When interaction of other column with higher has increased R-squared and multiple R-squared when using “*” symbol

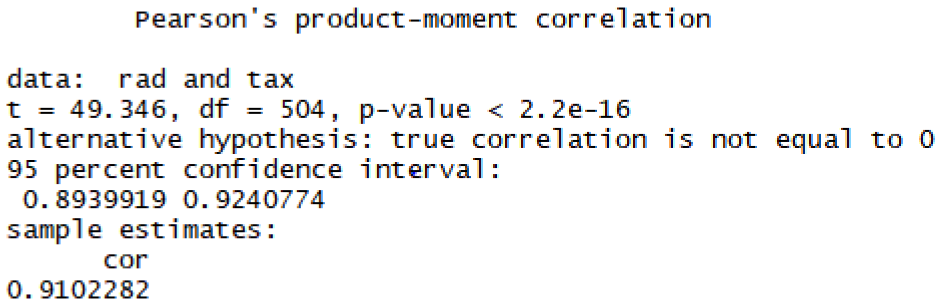
3) This exercise concerns about BOSTON housing data model



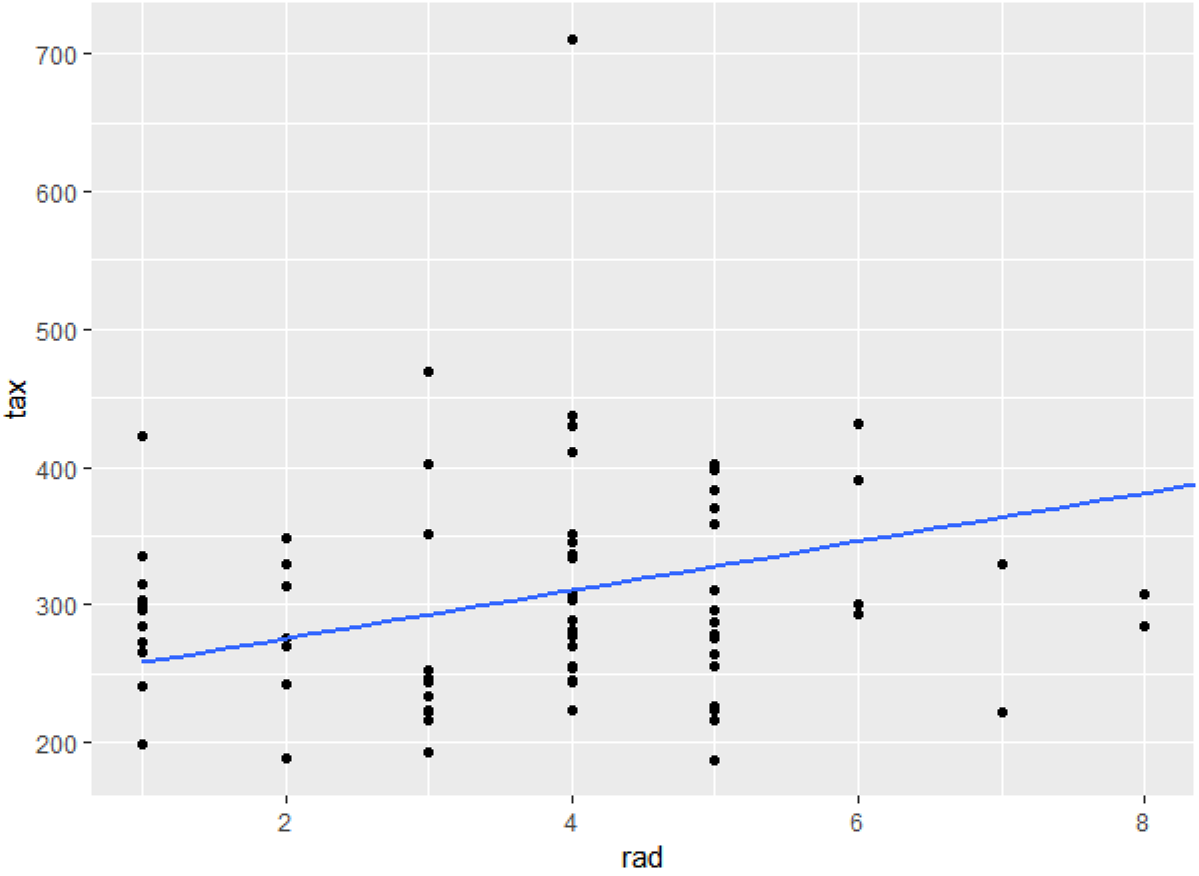
Above plot shows the pair wise plot of each and every variable in BOSTON data set. Not much information can be predicted with above graph. So correlation might be help full


```
         crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio black lstat  medv
crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29 -0.39  0.46 -0.39
zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39  0.18 -0.41  0.36
indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38 -0.36  0.60 -0.48
chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12  0.05 -0.05  0.18
nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19 -0.38  0.59 -0.43
rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36  0.13 -0.61  0.70
age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26 -0.27  0.60 -0.38
dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23  0.29 -0.50  0.25
rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46 -0.44  0.49 -0.38
tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46 -0.44  0.54 -0.47
ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00 -0.18  0.37 -0.51
black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18  1.00 -0.37  0.33
lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37 -0.37  1.00 -0.74
medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51  0.33 -0.74  1.00
> 
```

By observing the above correlation matrix it was found that some of the variable have high correlation between them some of them are rad and tax there is high correlation which is about 0.91

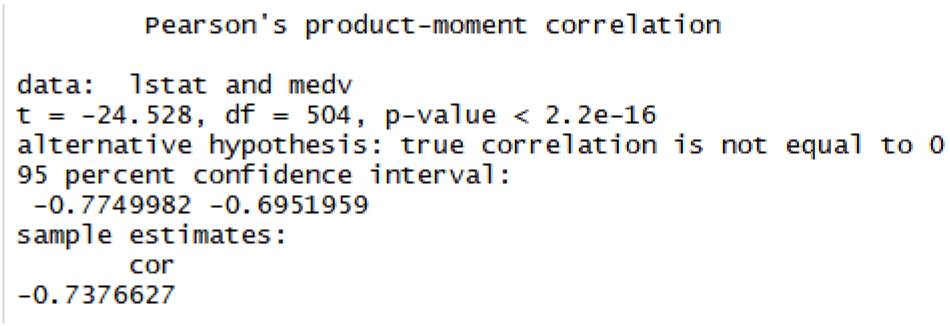


```
	Pearson's product-moment correlation

data:  rad and tax
t = 49.346, df = 504, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8939919 0.9240774
sample estimates:
      cor 
0.9102282 
```

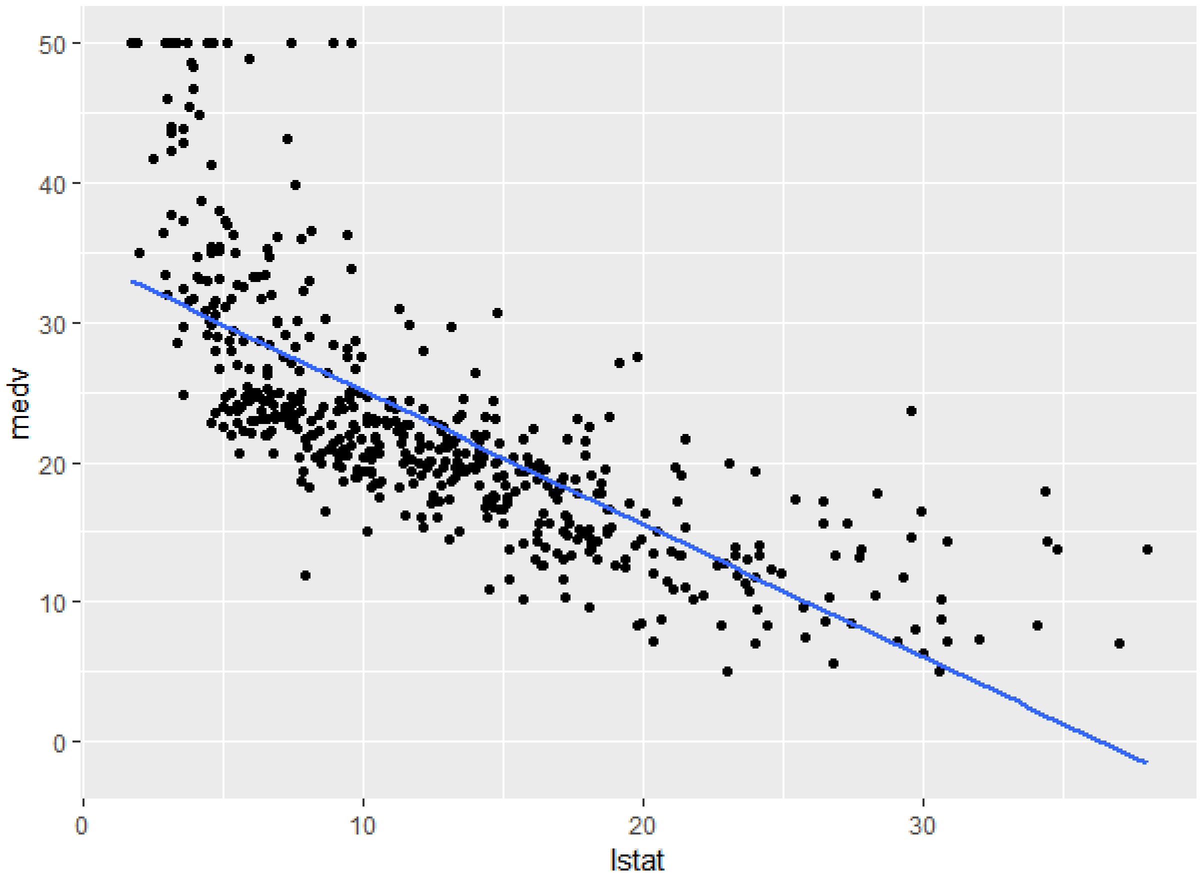


The line is linearly increasing, so it has positive correlation

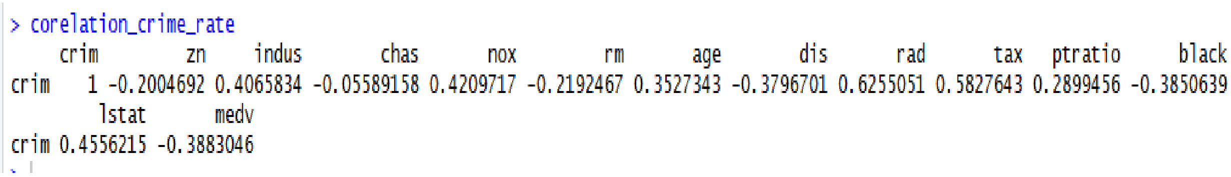
Another pair which has high correlation is given by lstat and medv . They both have the negative correlation.



```
        Pearson's product-moment correlation

data:  lstat and medv
t = -24.528, df = 504, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7749982 -0.6951959
sample estimates:
       cor 
-0.7376627 
```

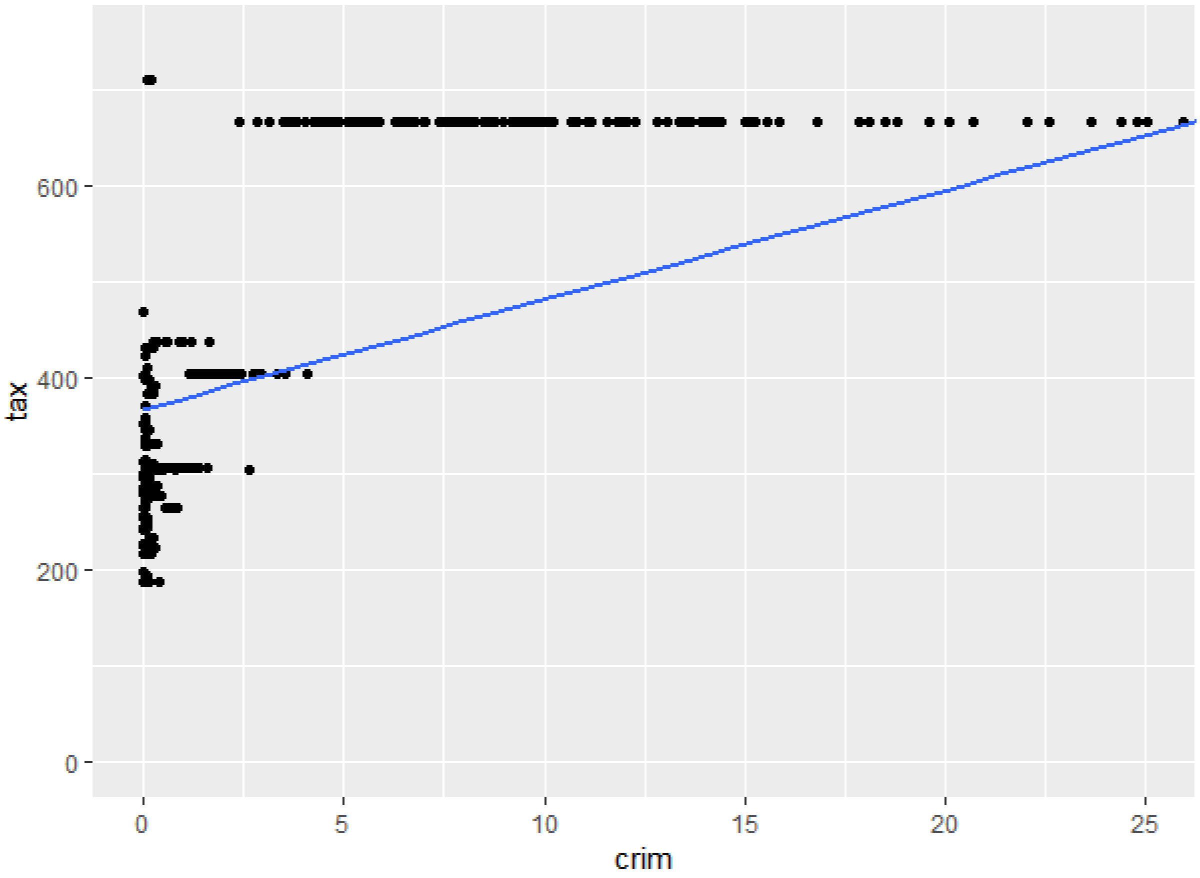


B) Predictors involved in Per Capita Income



```
> corelation_crime_rate
     crim         zn     indus        chas       nox         rm       age        dis       rad       tax   ptratio      black
crim    1 -0.2004692 0.4065834 -0.05589158 0.4209717 -0.2192467 0.3527343 -0.3796701 0.6255051 0.5827643 0.2899456 -0.3850639
         lstat       medv
crim 0.4556215 -0.3883046
```

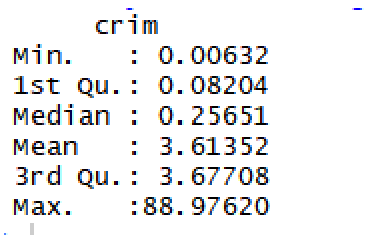
By Analysing the above correlation matrix, It was found out that crimerate and rad and also crimerate and tax ratio has high correlation.



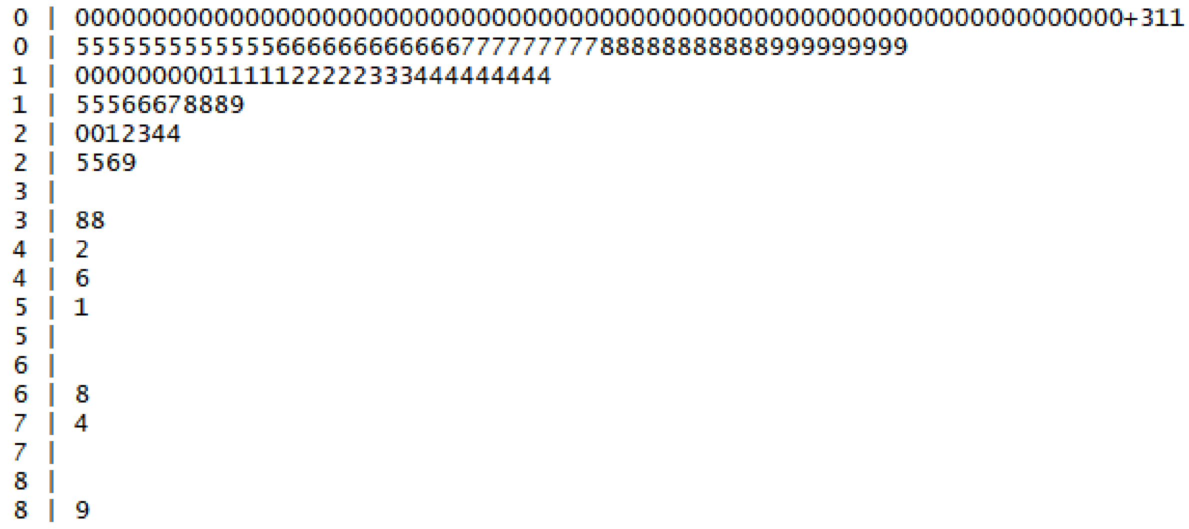
But by seeing scatter plot does not provide that much information regarding the correlation.

C) **High crime rate**

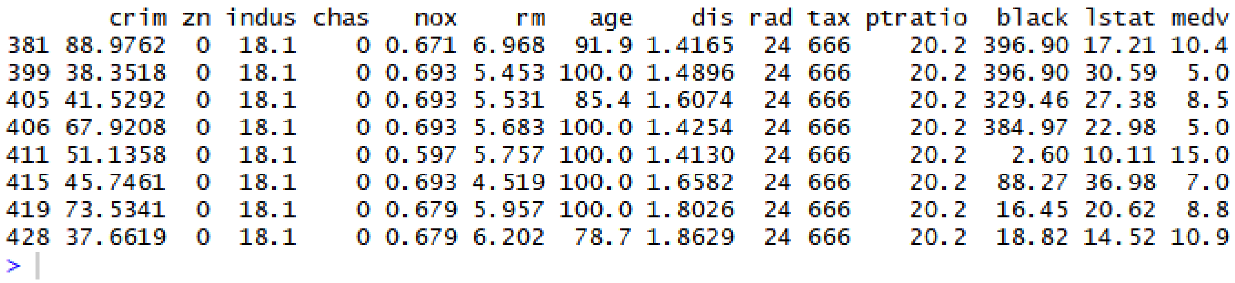
The majority of towns have close to zero crime rate, while some suburbs have a rate as high as 70 or 80. There are many outlier suburbs marked as outliers indicating most suburbs have a very low crime rate, although there are a significant number of towns in the minority with an outlying crime rate on the higher end.



```
      crim
 Min.   : 0.00632
 1st Qu.: 0.08204
 Median : 0.25651
 Mean   : 3.61352
 3rd Qu.: 3.67708
 Max.   :88.97620
```

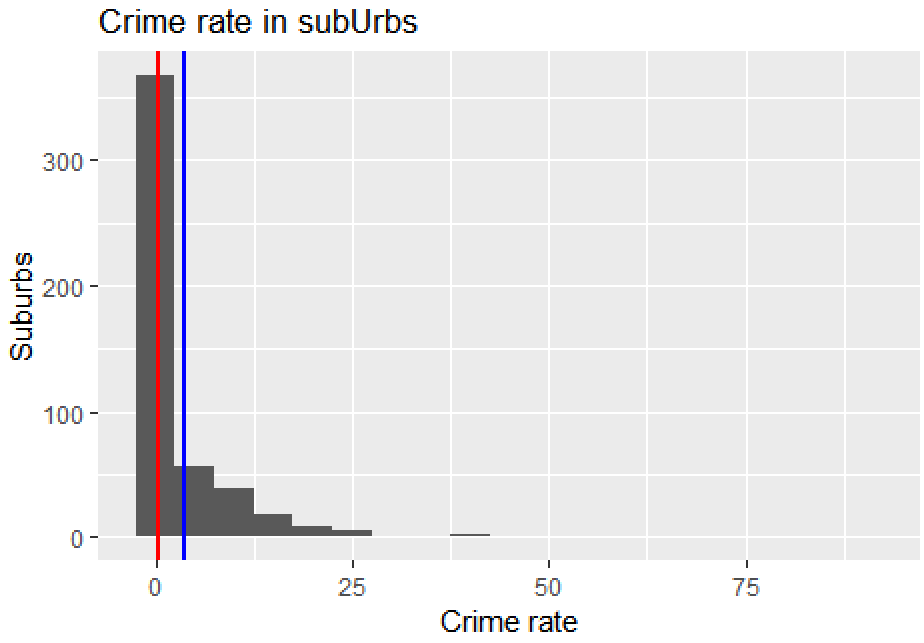


```
0 | 000000000000000000000000000000000000000000000000000000000000000000000+311
0 | 555555555555566666666666677777777788888888888999999999
1 | 0000000001111122222333444444444
1 | 55566678889
2 | 0012344
2 | 5569
3 |
3 | 88
4 | 2
4 | 6
5 | 1
5 |
6 |
6 | 8
7 | 4
7 |
8 |
8 | 9
```



```
       crim zn indus chas   nox    rm   age    dis rad tax ptratio  black lstat medv
381 88.9762  0  18.1    0 0.671 6.968  91.9 1.4165  24 666    20.2 396.90 17.21 10.4
399 38.3518  0  18.1    0 0.693 5.453 100.0 1.4896  24 666    20.2 396.90 30.59  5.0
405 41.5292  0  18.1    0 0.693 5.531  85.4 1.6074  24 666    20.2 329.46 27.38  8.5
406 67.9208  0  18.1    0 0.693 5.683 100.0 1.4254  24 666    20.2 384.97 22.98  5.0
411 51.1358  0  18.1    0 0.597 5.757 100.0 1.4130  24 666    20.2   2.60 10.11 15.0
415 45.7461  0  18.1    0 0.693 4.519 100.0 1.6582  24 666    20.2  88.27 36.98  7.0
419 73.5341  0  18.1    0 0.679 5.957 100.0 1.8026  24 666    20.2  16.45 20.62  8.8
428 37.6619  0  18.1    0 0.679 6.202  78.7 1.8629  24 666    20.2  18.82 14.52 10.9
>
```

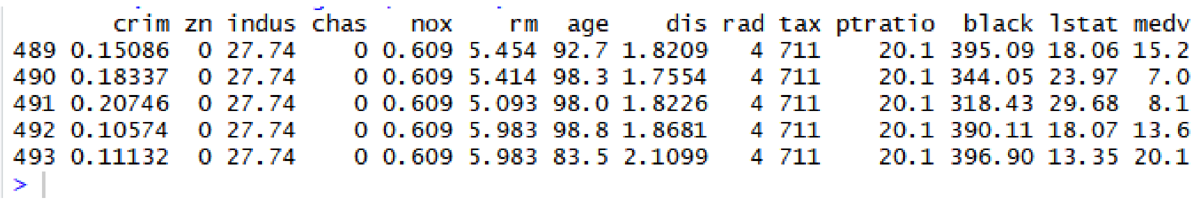
The above matrix shows the data of the suburbs with particularly higher crime rate.



This histogram shows the mean and median values of crime rate and also the outliers who are particularly contributing high crime rate

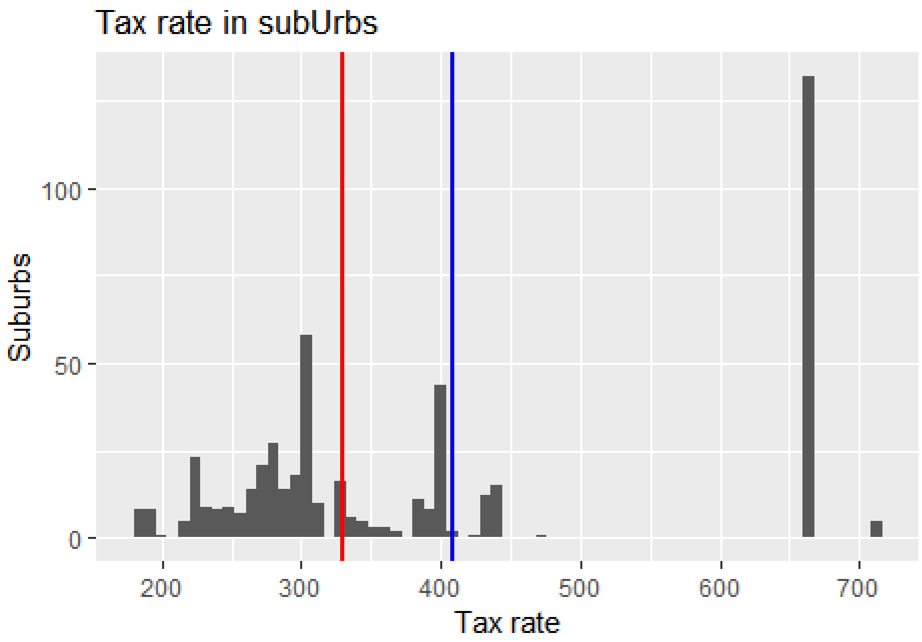
<u>High Tax Rate:</u>

The property tax ranges from approximately 200 to 700, approximately. The median lies around 320 dollars. The mean lies around 410.There were around 5 suburbs which are particularly with higher tax rate.



```
       crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
489 0.15086  0 27.74    0 0.609 5.454 92.7 1.8209   4 711    20.1 395.09 18.06 15.2
490 0.18337  0 27.74    0 0.609 5.414 98.3 1.7554   4 711    20.1 344.05 23.97  7.0
491 0.20746  0 27.74    0 0.609 5.093 98.0 1.8226   4 711    20.1 318.43 29.68  8.1
492 0.10574  0 27.74    0 0.609 5.983 98.8 1.8681   4 711    20.1 390.11 18.07 13.6
493 0.11132  0 27.74    0 0.609 5.983 83.5 2.1099   4 711    20.1 396.90 13.35 20.1
> 
```

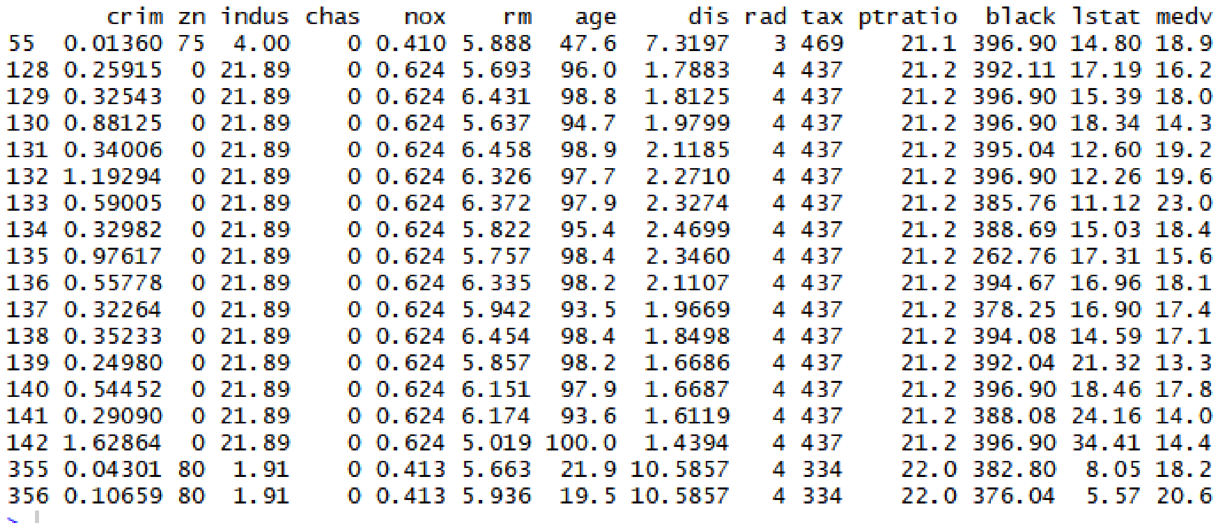
The above matrix shows the suburbs which have higher crime ratio



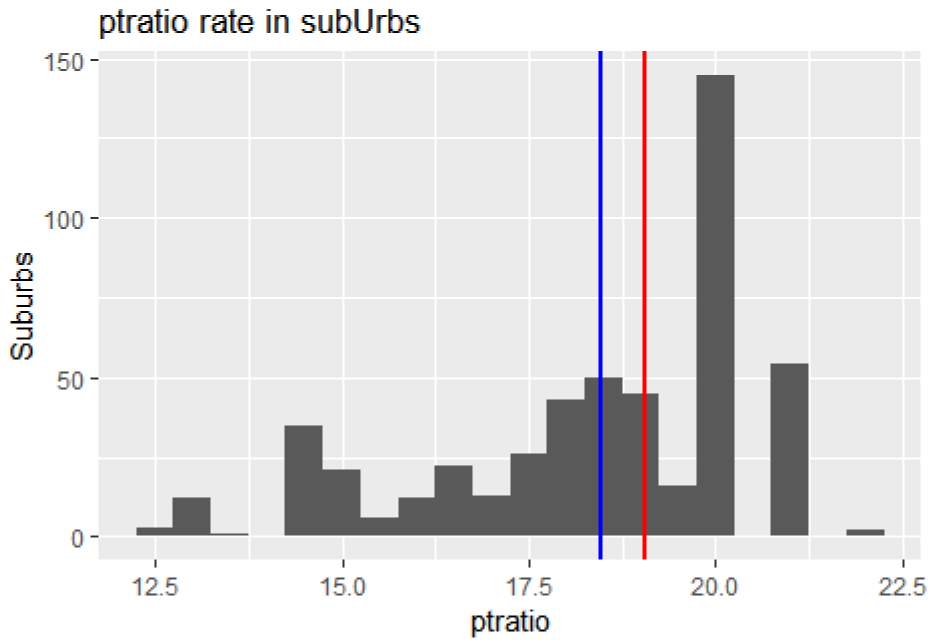
The Above histogram shows the outliers and mean and mode of the tax rate with respect to suburbs

C) **<u>Pupil-Teacher Ratio</u>**

The below matrix shows the suburbs with relatively higher Pupil-Teacher ratio. By seeing the matrix we can see that there were some outliers .There were around 18 suburbs which are greater than outer quartile range .It also increases the mean of the distribution



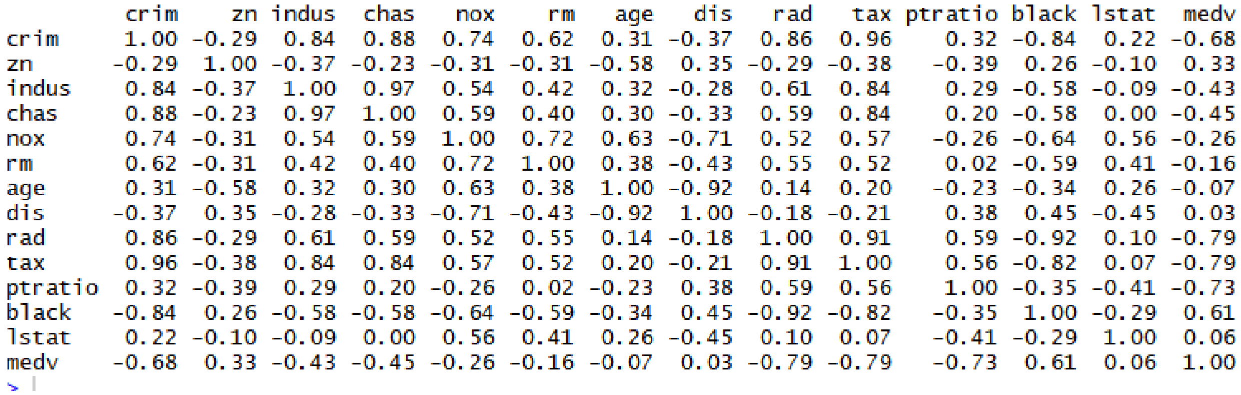
```
       crim zn indus chas   nox    rm   age     dis rad tax ptratio  black lstat medv
55  0.01360 75  4.00    0 0.410 5.888  47.6  7.3197   3 469    21.1 396.90 14.80 18.9
128 0.25915  0 21.89    0 0.624 5.693  96.0  1.7883   4 437    21.2 392.11 17.19 16.2
129 0.32543  0 21.89    0 0.624 6.431  98.8  1.8125   4 437    21.2 396.90 15.39 18.0
130 0.88125  0 21.89    0 0.624 5.637  94.7  1.9799   4 437    21.2 396.90 18.34 14.3
131 0.34006  0 21.89    0 0.624 6.458  98.9  2.1185   4 437    21.2 395.04 12.60 19.2
132 1.19294  0 21.89    0 0.624 6.326  97.7  2.2710   4 437    21.2 396.90 12.26 19.6
133 0.59005  0 21.89    0 0.624 6.372  97.9  2.3274   4 437    21.2 385.76 11.12 23.0
134 0.32982  0 21.89    0 0.624 5.822  95.4  2.4699   4 437    21.2 388.69 15.03 18.4
135 0.97617  0 21.89    0 0.624 5.757  98.4  2.3460   4 437    21.2 262.76 17.31 15.6
136 0.55778  0 21.89    0 0.624 6.335  98.2  2.1107   4 437    21.2 394.67 16.96 18.1
137 0.32264  0 21.89    0 0.624 5.942  93.5  1.9669   4 437    21.2 378.25 16.90 17.4
138 0.35233  0 21.89    0 0.624 6.454  98.4  1.8498   4 437    21.2 394.08 14.59 17.1
139 0.24980  0 21.89    0 0.624 5.857  98.2  1.6686   4 437    21.2 392.04 21.32 13.3
140 0.54452  0 21.89    0 0.624 6.151  97.9  1.6687   4 437    21.2 396.90 18.46 17.8
141 0.29090  0 21.89    0 0.624 6.174  93.6  1.6119   4 437    21.2 388.08 24.16 14.0
142 1.62864  0 21.89    0 0.624 5.019 100.0  1.4394   4 437    21.2 396.90 34.41 14.4
355 0.04301 80  1.91    0 0.413 5.663  21.9 10.5857   4 334    22.0 382.80  8.05 18.2
356 0.10659 80  1.91    0 0.413 5.936  19.5 10.5857   4 334    22.0 376.04  5.57 20.6
```



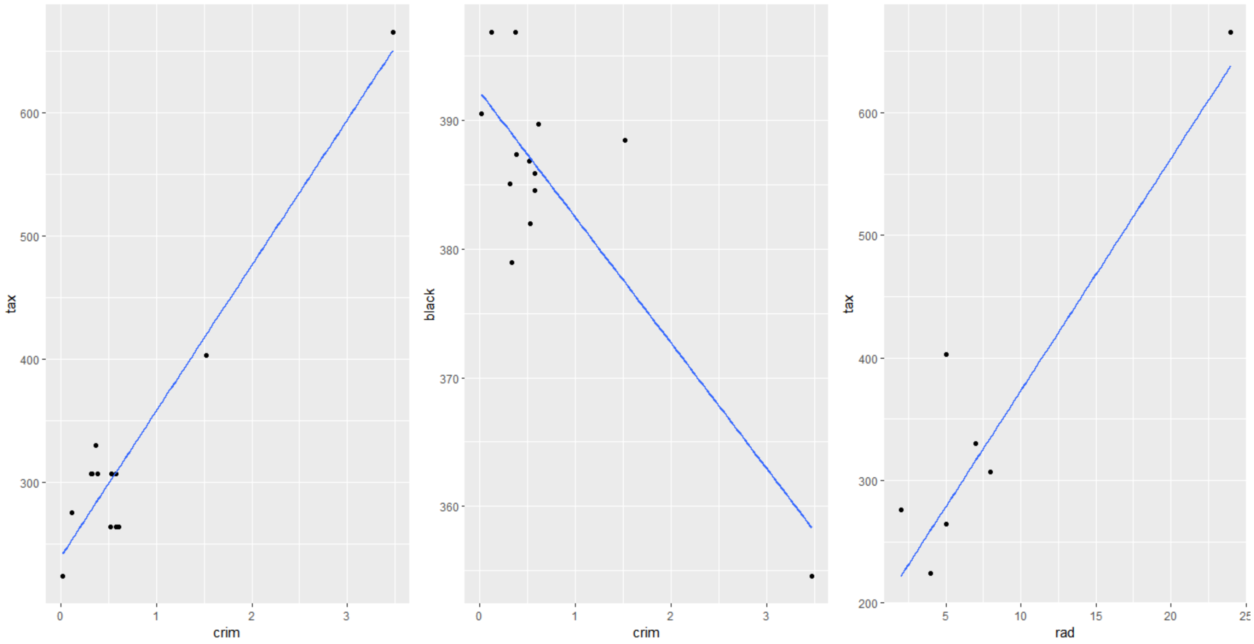
The above histogram shows the mean and median of the suburbs with respect to pupil-teacher ratio. And some of the outliers which significantly contributing to higher Pupil teacher ratio

D) Comment on pairs greater than 8

Below are the correlation coefficients for the average dwelling rate greater than 8. It gives some correlation between different things like crime and tax it has high correlation. When crime rate is high tax rate will also be high (i.e. positively correlated) when average dwelling is greater than 8 and also crime rate and black people also highly correlated so when crime will be minimum when black people are higher (i.e. negatively correlated).similarly correlation coefficient is high for rad and tax (i.e positively correlated) when average of dwelling is greater than 8

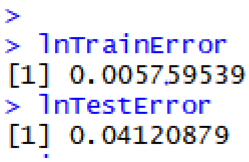


```
        crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio black lstat  medv
crim     1.00 -0.29  0.84  0.88  0.74  0.62  0.31 -0.37  0.86  0.96    0.32 -0.84  0.22 -0.68
zn      -0.29  1.00 -0.37 -0.23 -0.31 -0.31 -0.58  0.35 -0.29 -0.38   -0.39  0.26 -0.10  0.33
indus    0.84 -0.37  1.00  0.97  0.54  0.42  0.32 -0.28  0.61  0.84    0.29 -0.58 -0.09 -0.43
chas     0.88 -0.23  0.97  1.00  0.59  0.40  0.30 -0.33  0.59  0.84    0.20 -0.58  0.00 -0.45
nox      0.74 -0.31  0.54  0.59  1.00  0.72  0.63 -0.71  0.52  0.57   -0.26 -0.64  0.56 -0.26
rm       0.62 -0.31  0.42  0.40  0.72  1.00  0.38 -0.43  0.55  0.52    0.02 -0.59  0.41 -0.16
age      0.31 -0.58  0.32  0.30  0.63  0.38  1.00 -0.92  0.14  0.20   -0.23 -0.34  0.26 -0.07
dis     -0.37  0.35 -0.28 -0.33 -0.71 -0.43 -0.92  1.00 -0.18 -0.21    0.38  0.45 -0.45  0.03
rad      0.86 -0.29  0.61  0.59  0.52  0.55  0.14 -0.18  1.00  0.91    0.59 -0.92  0.10 -0.79
tax      0.96 -0.38  0.84  0.84  0.57  0.52  0.20 -0.21  0.91  1.00    0.56 -0.82  0.07 -0.79
ptratio  0.32 -0.39  0.29  0.20 -0.26  0.02 -0.23  0.38  0.59  0.56    1.00 -0.35 -0.41 -0.73
black   -0.84  0.26 -0.58 -0.58 -0.64 -0.59 -0.34  0.45 -0.92 -0.82   -0.35  1.00 -0.29  0.61
lstat    0.22 -0.10 -0.09  0.00  0.56  0.41  0.26 -0.45  0.10  0.07   -0.41 -0.29  1.00  0.06
medv    -0.68  0.33 -0.43 -0.45 -0.26 -0.16 -0.07  0.03 -0.79 -0.79   -0.73  0.61  0.06  1.00
> |
```



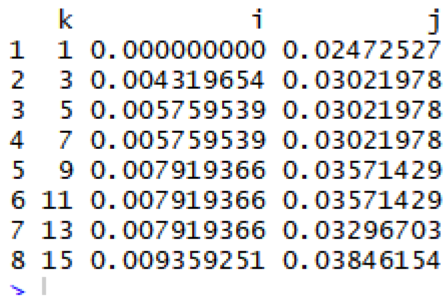
The above scatter plot shows the relationship between **crime rate vs tax** and **crim vs black** and **rad vs tax**. Where two linearly increasing where other one is decreasing linearly.

4.Based upon the value of the test error in the knn process we will selecting the value of k.When k=1 there may cause over fitting of Data .When K value increases then there may be change of over fitting



```
>
> lnTrainError
[1] 0.005759539
> lnTestError
[1] 0.04120879
```

The above picture is for linear regression error



```
   k           i          j
1  1 0.000000000 0.02472527
2  3 0.004319654 0.03021978
3  5 0.005759539 0.03021978
4  7 0.005759539 0.03021978
5  9 0.007919366 0.03571429
6 11 0.007919366 0.03571429
7 13 0.007919366 0.03296703
8 15 0.009359251 0.03846154
> 
```

The above is the picture of knn error

Based upon the value of we will selecting whether to chose linear regression or Knn .In our case we will be selecting knn when K value is less than 3