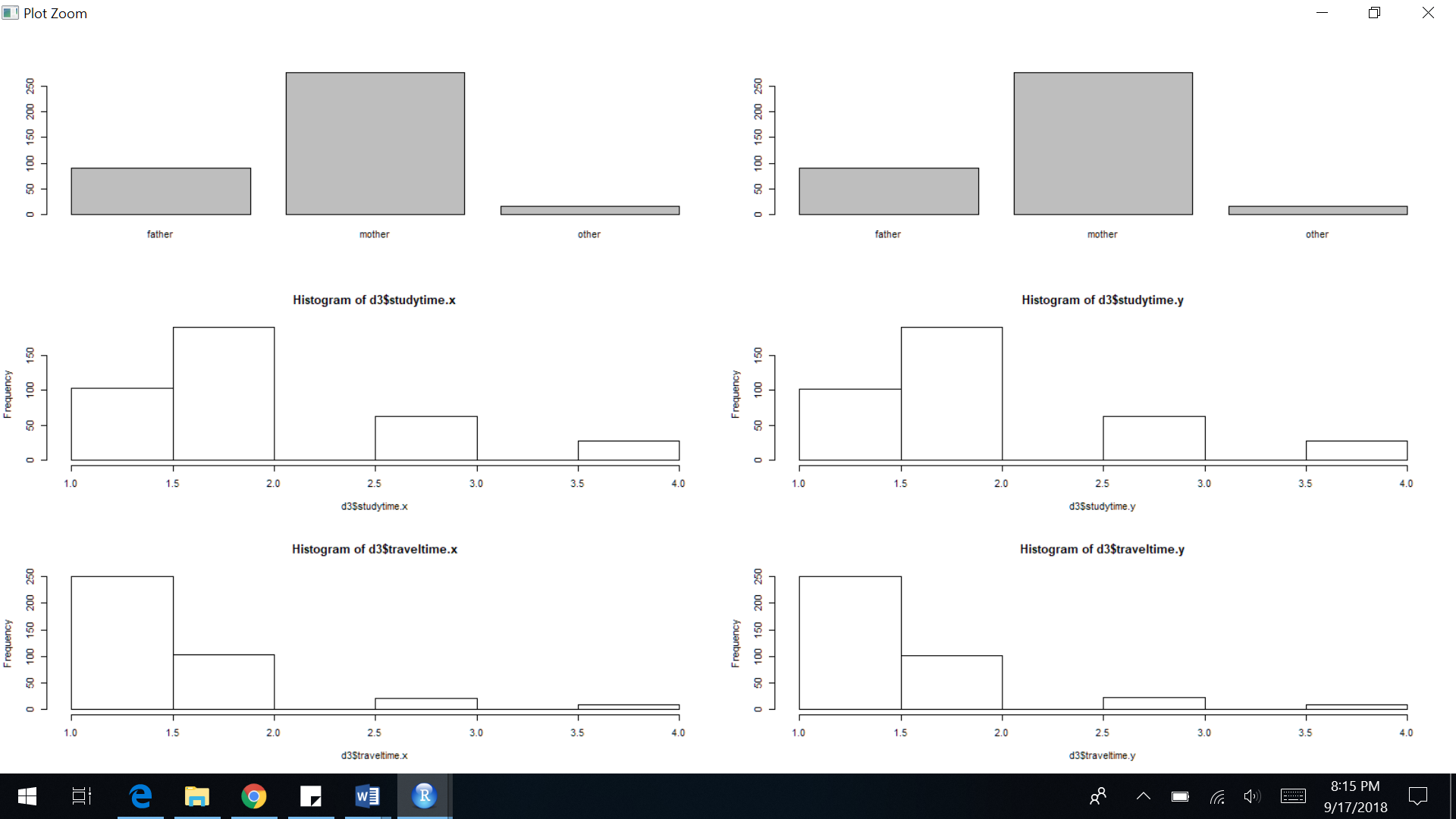
**EAS 506 Statistical Data Mining Jaideep Reddy Kommera**

**Homework 1 Class no. 28**

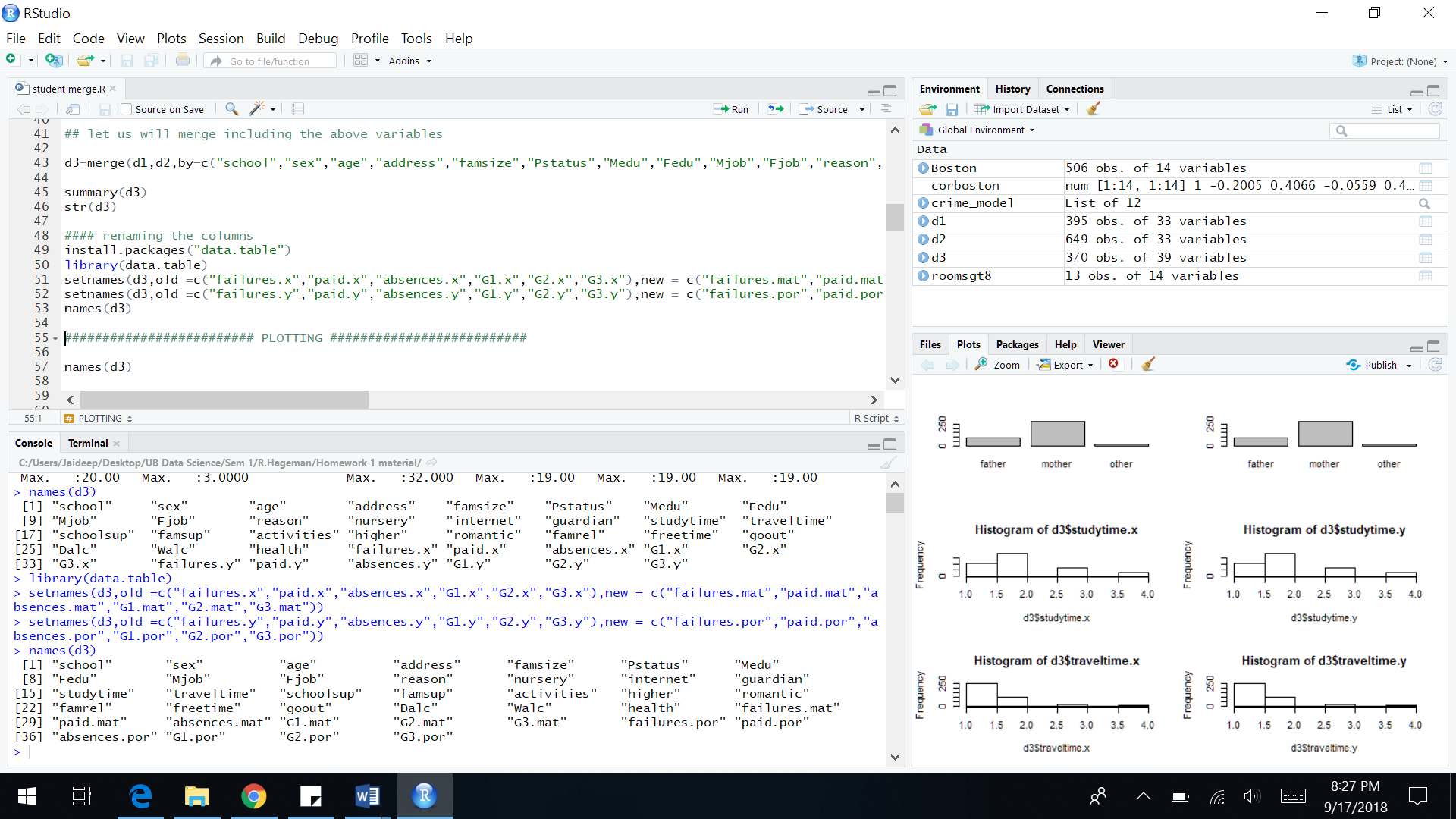
**Question 1:**

Introduction: The given dataset is about Student Performance in two different subjects, Math and Portuguese. There are students who are present in both datasets and therefore both the datasets are merged based on common variables. The following work is done on the merged dataset.

Pre-processing: The initial merged dataset has x and y components for different variables. On further inspection of the variables both the components are similar. Therefore, one of the columns can be dropped.



In the image above, we can notice that the x and y components of the variables are same. Therefore, the datasets are merged again including all the similar components. The grades have been renamed based on the subject they belong to. The names and summary of the final dataset is as below

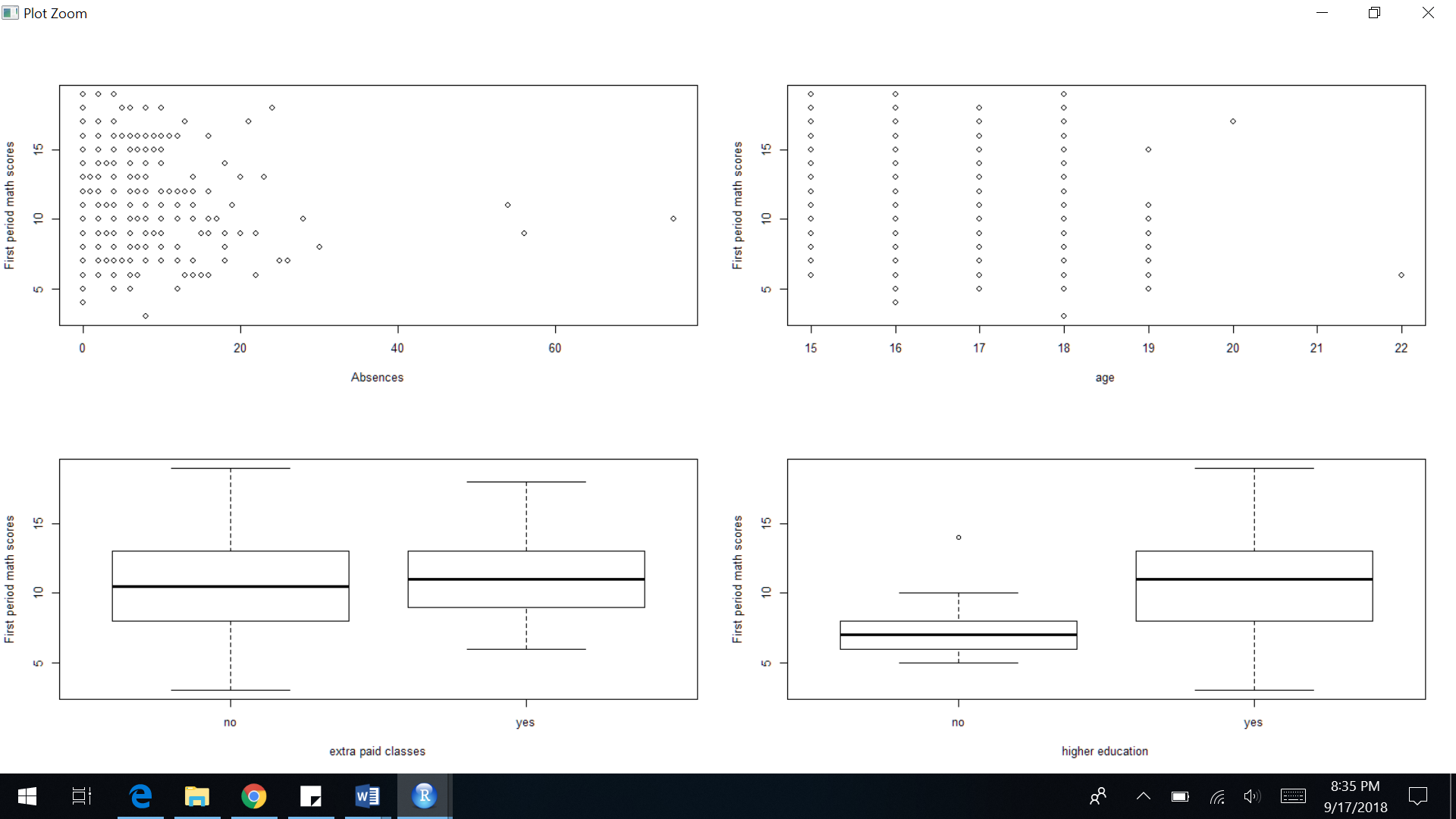


The dataset did not have any missing or NA values.



Plotting:

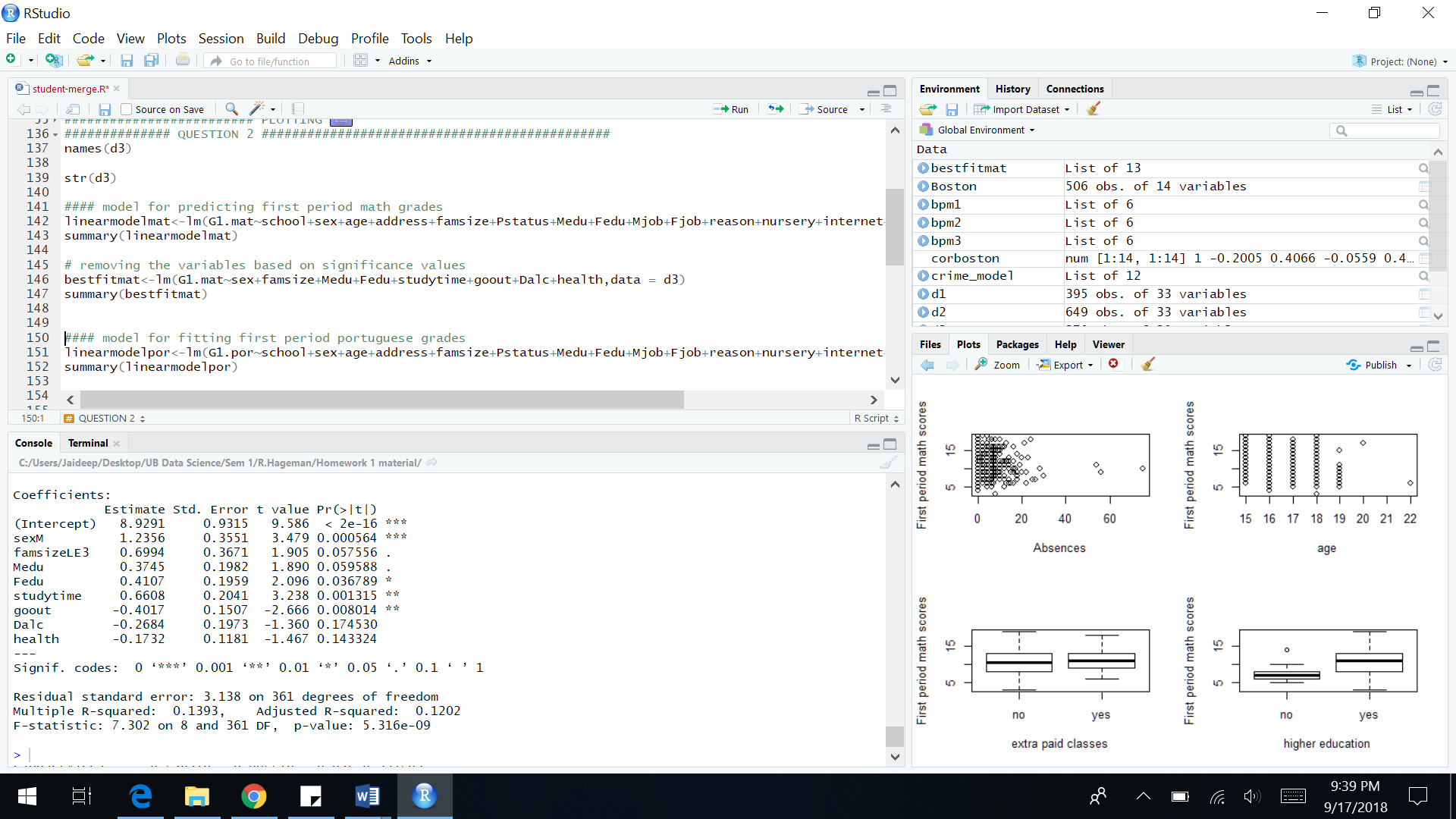
Plotting first period math grades against different variables to see the interaction



**Question 2:**

Performing Linear Regression on the merged dataset. The first period grades are the response variables. Two separate regressions have been performed on Math grades and Portuguese grades.

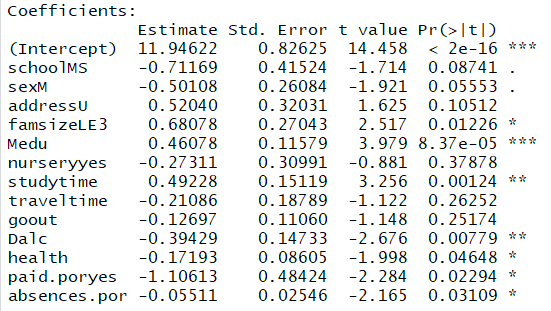
Response Variable: First Period Math Grade



From the above summary we can see that the first period math grade is dependent on gender, study time, father’s education and frequency of going out with friends.

Gender, study time and father’s education affect positively (i.e. higher the better) but going out with friends affects negatively.

Response Variable: First Period Portuguese Grade



From the above table we can see that the factors affecting first period Portuguese grades are Mother’s education, study time, Daily alcohol consumption, health, paid for extra classes and absences.

**2.b)**

Suggestions for a first-year student based on model findings:

* Improve study time
* Reduce going out with friends often
* Reduce Daily alcohol consumption
* Maintain proper health

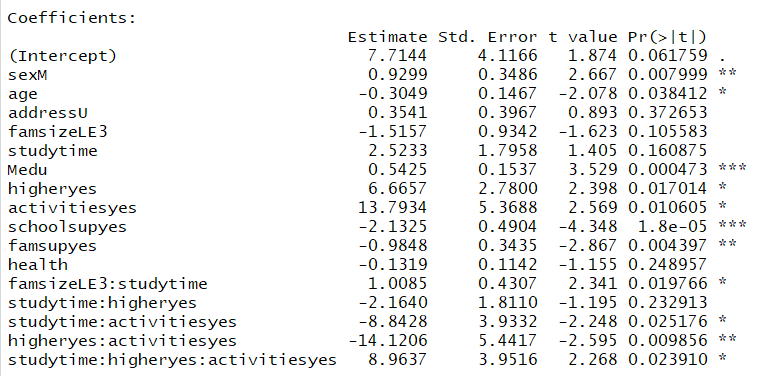
**2.c)**

Interaction between variables

Formula:



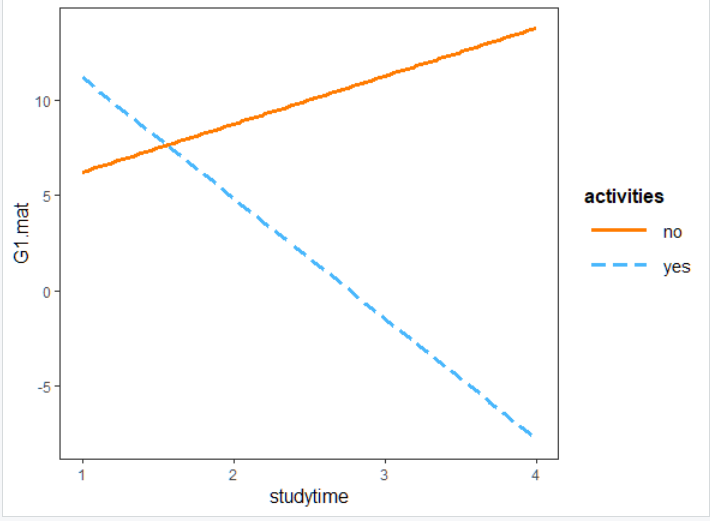
Summary:

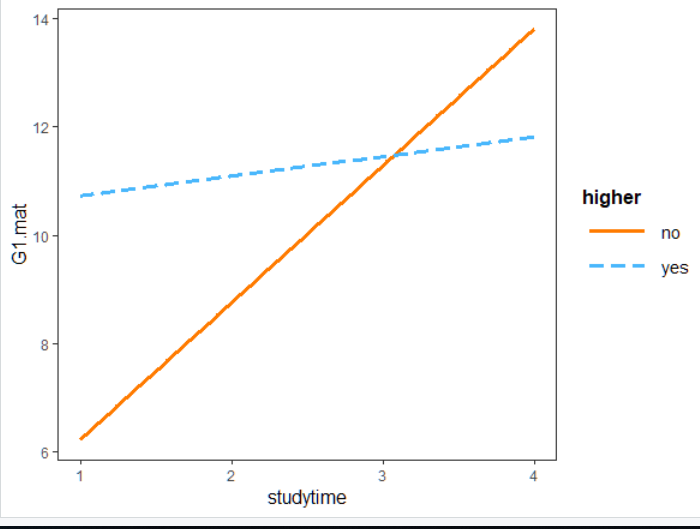


From the above table we can notice significant interactions such as

* Extracurricular activities and higher education
* Family size and study time
* Study time and extracurricular activities

**Plotting interaction between significant variables:**



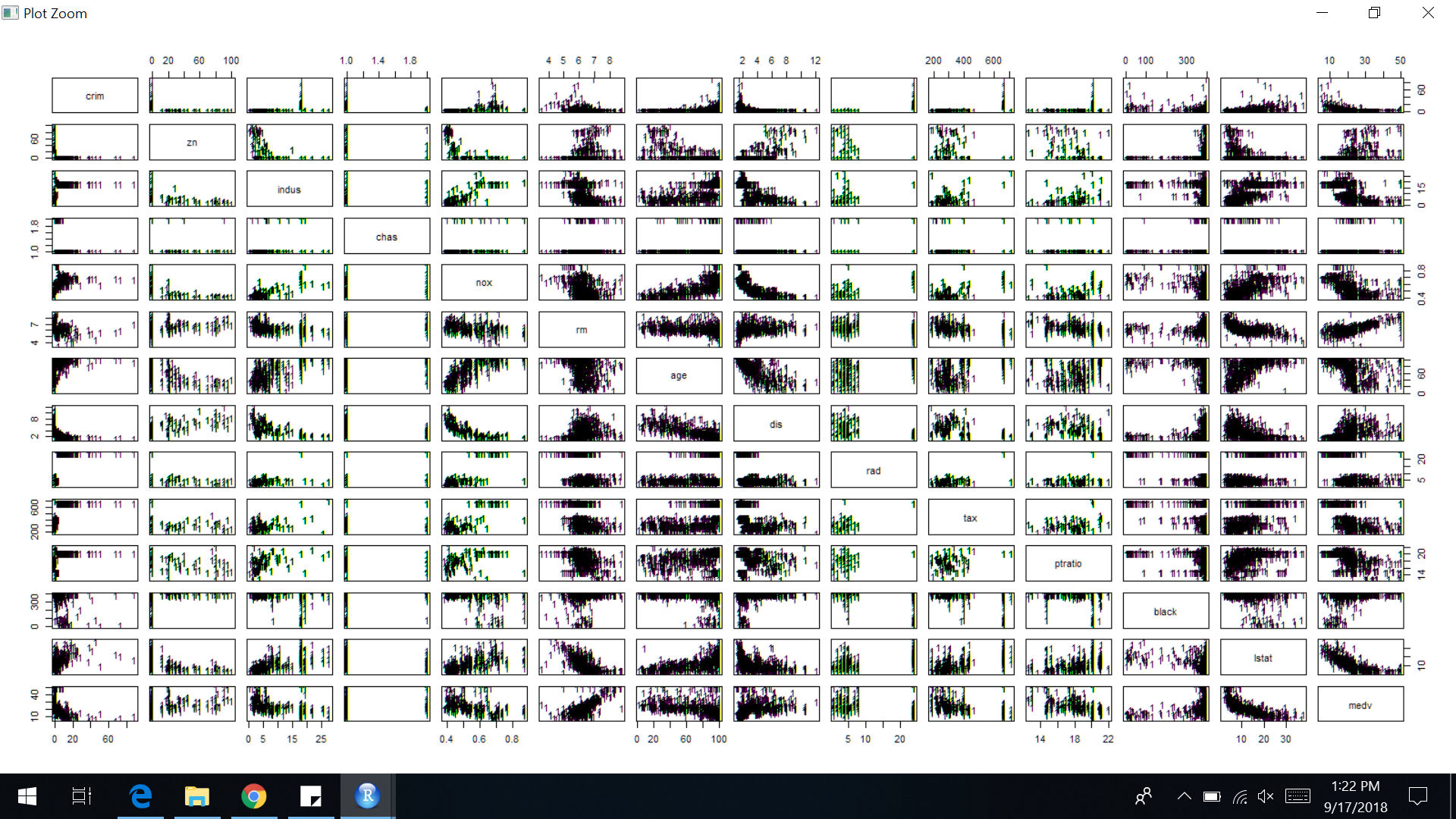


**Question 3:**

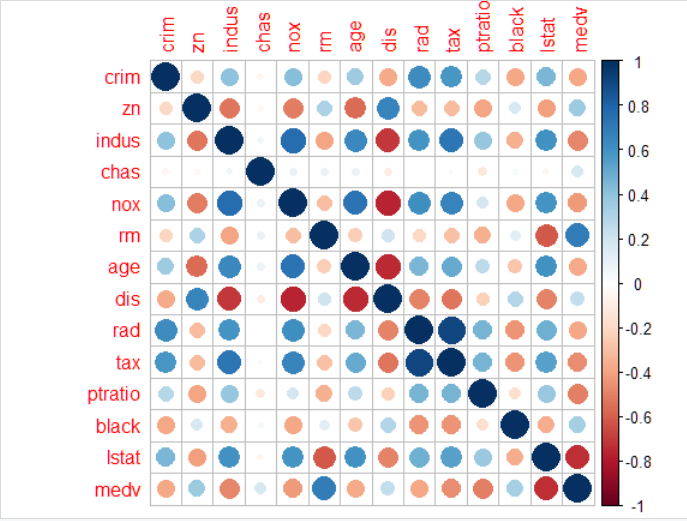
Introduction: The Boston dataset has the housing values in the suburbs of Boston from the 1970’s. The dataset has 506 rows and 14 columns.

**3.a)**

Pairwise Scatterplot:



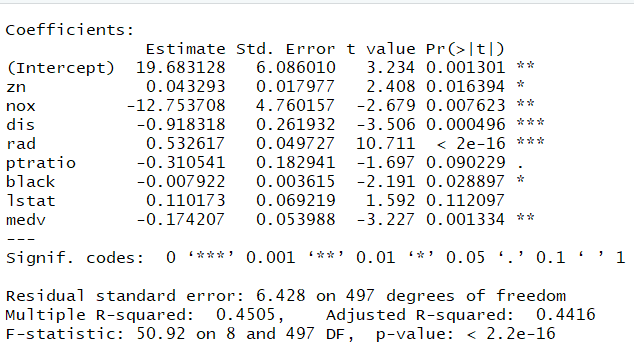
Correlation Matrix:



From these plots we can say that the predictors are correlated

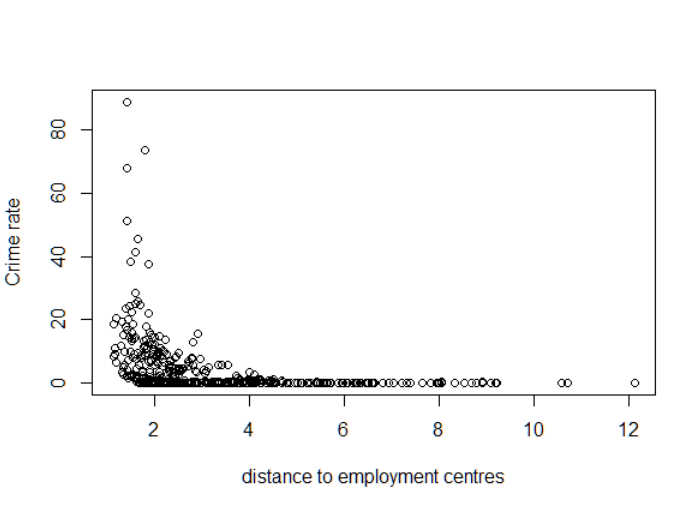
**3.b)**

We can use a linear model to find significant variables that are associated with crime rate

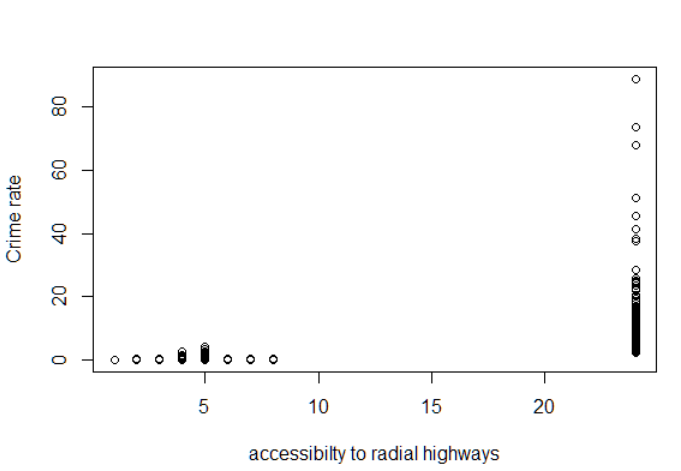


Accessibility to radial highways, distances to employment centers, median value of houses and nitrous oxide concentration are significant variables that affect the per capita crime rate.

Plots of the significant variables with crime rate:



Crime rates are higher in areas closer to the Boston employment centers

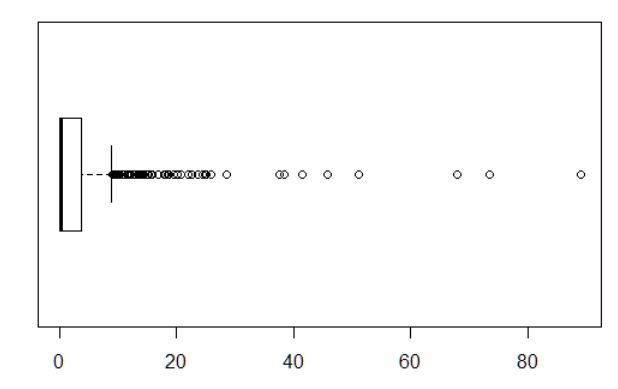




Crime rates are higher where the housing price is lower

**3.c)**

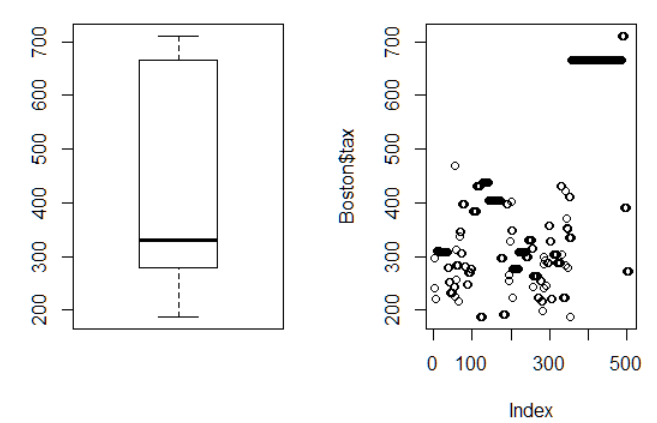
Based on the boxplot suburbs with crime rate greater than 35 are considered



8 suburbs have higher crime rate

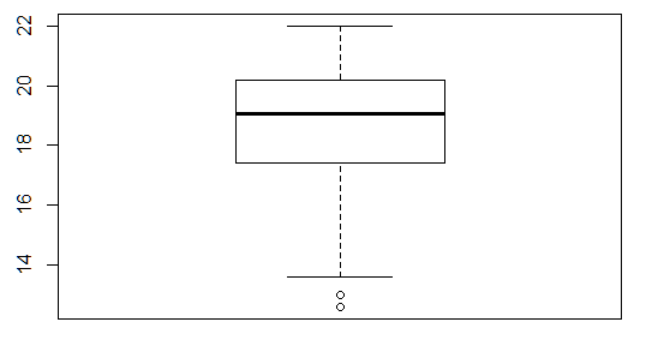
Based on the third quartile the tax rates greater than 666 are considered





5 suburbs have higher tax rates

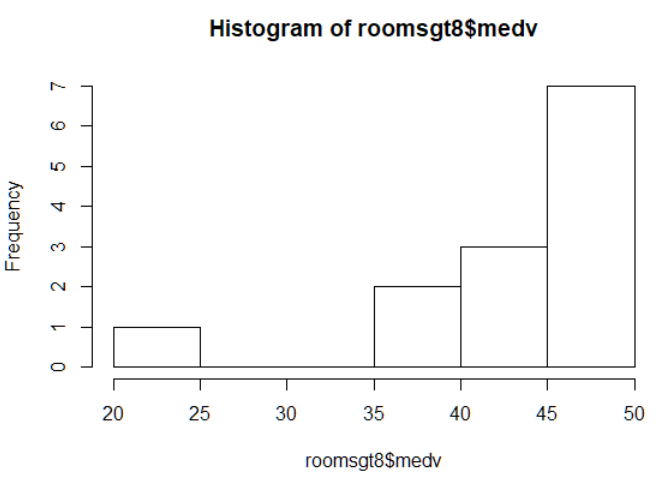
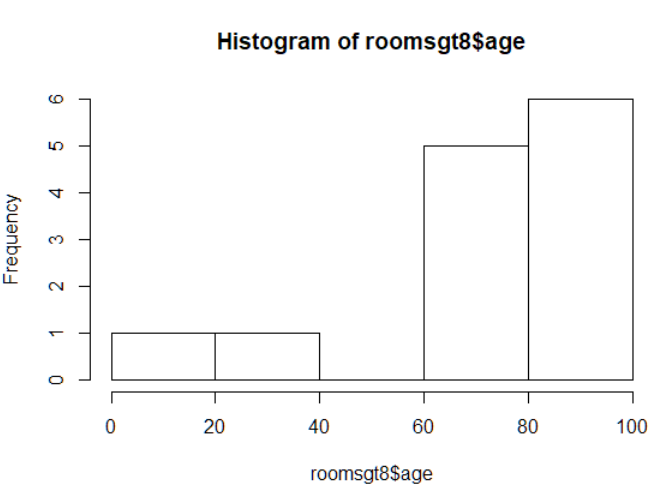
Based on the boxplot pupil-teacher ratio less than 14 is considered



16 suburbs have fewer pupil-student ratios

**3.d)**

* 64 suburbs average more than seven rooms per dwelling
* 13 suburbs average more than eight rooms per dwelling
* most houses that have >8 rooms are built prior to 1940
* most houses that have >8 rooms have higher median value



**Question 4:**

Introduction: Zip code data where we have to only consider 2’s and 3’s

Pre-processing: No NA or missing values. Data is filtered by considering only 2’s and 3’s.

Model Building: A linear model and KNN model for K values of 1,3,5,7,9,11, 13,15 need to be build.

The mis-classification rate for the linear model comes as **0.15166**

Errors for different values of k are:



Plotting:

