HOMEWORK 1

STATISTICAL DATA MINING

EAS 506

NAME : SHUBHAM SHARMA

PERSON # : 50290293

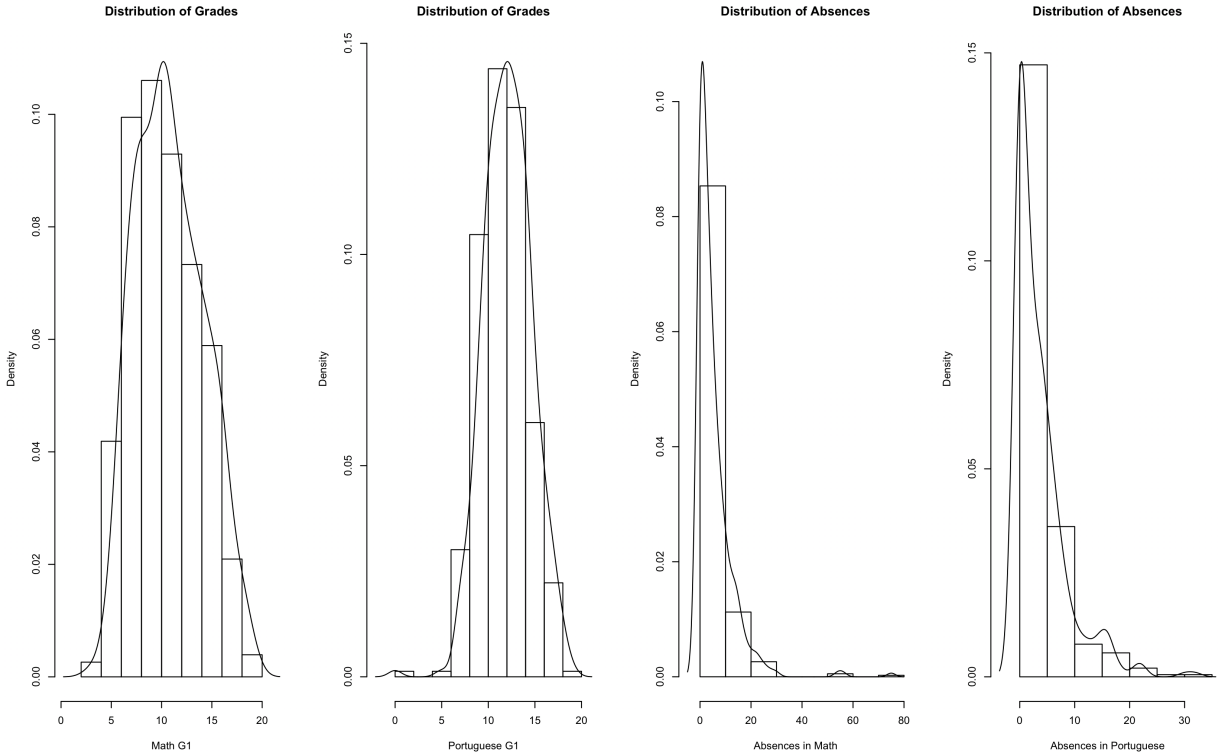
CLASS # : 44

PROFESSOR : DR. RACHAEL H. BLAIR

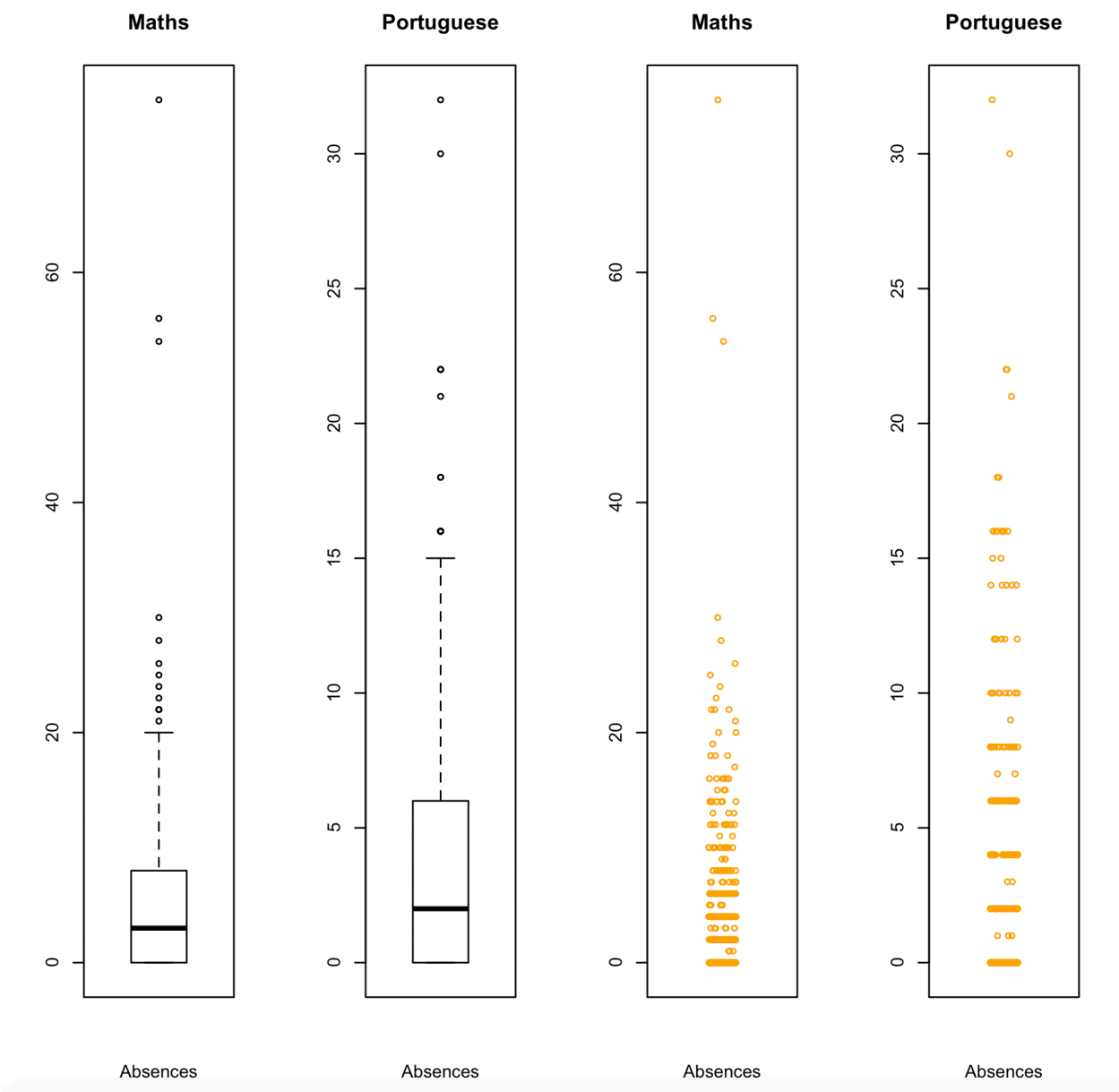
*1) Consider the Student Performance Data Set on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance). Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Pre-process this data and justify your choices (elimination of outliers, elimination of variables, variable transformations, etc.) in your write up. Submit the cleaned dataset as an *.RData file.*

The merged dataset has 382 rows and 53 columns. The given data set was found to be complete and have zero missing values. After checking the summary of the data we get some insights which can be used further in our analysis. We find that more students are from Gabriel Pereira(342), the sex ratio is almost equal and the median of the ages is 17 years. Most of the students are from urban setting (301) with access to internet (324) and a majority relies on family educational support. The median of absences is 3 and the average first period grade for the group is 10.86 for Maths while the median of absences is 2 and the average first period grade for the group is 12.11 for Portuguese.

Following is the density distribution for first period grades and absences in Maths and Portuguese.

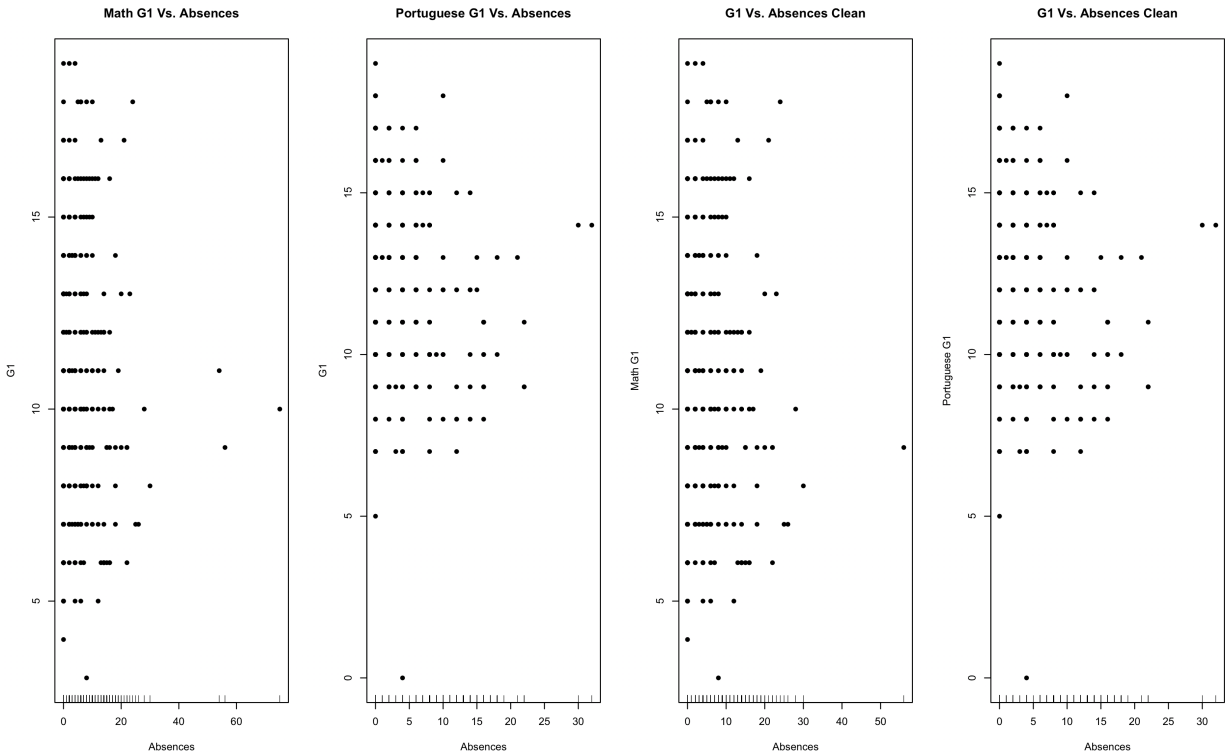


We can see that the grades are almost a normal distribution although the grades for Math are slightly skewed to the left. For binning, Sturges formula has been used. It can be inferred that students tend to perform better in Portuguese than maths and have less absence in the Portuguese classes. Although we observe that some data points are fairly large for Absences in Math and Portuguese which can be considered as potential outliers. We can confirm this in the following graph:

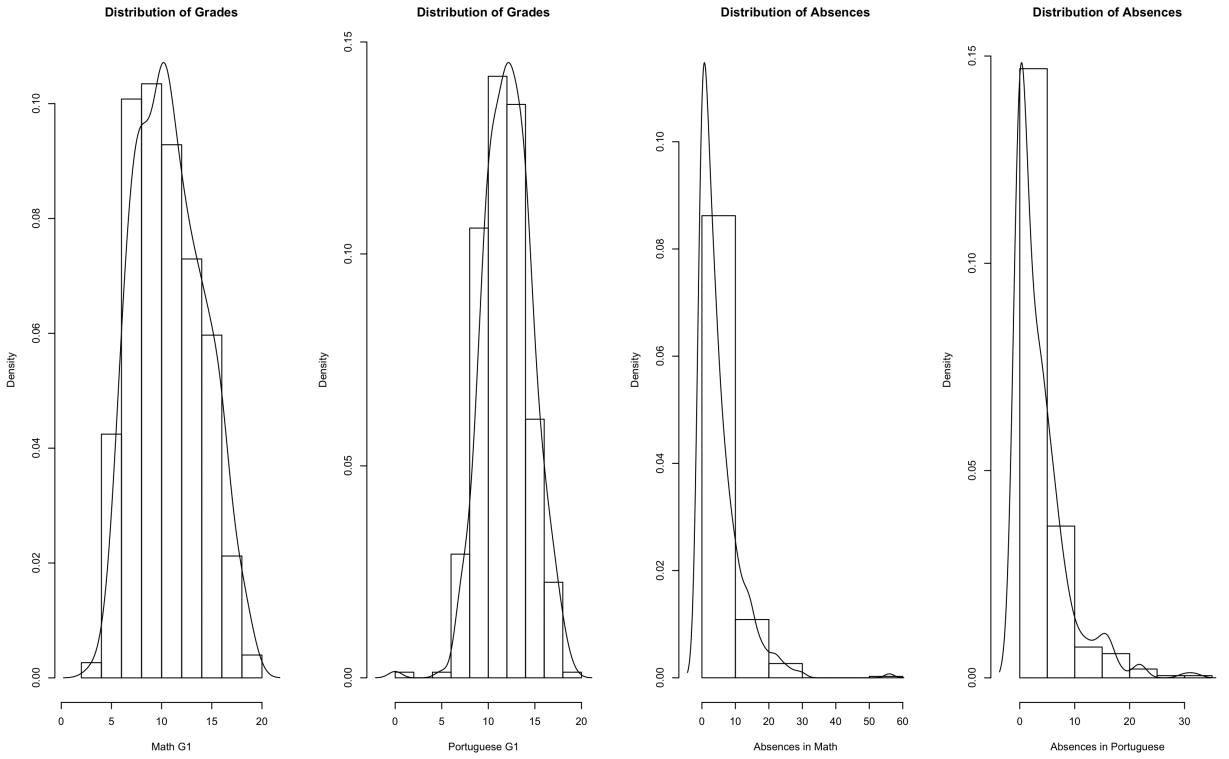


The points towards the top of the graph are potential outliers which can be eliminated to make the data cleaner. It is done using a for loop and taking the assumption of whatever data lies outside of 3.5 times of the standard deviation from the mean.

After outlier removal the distribution looks as follows:

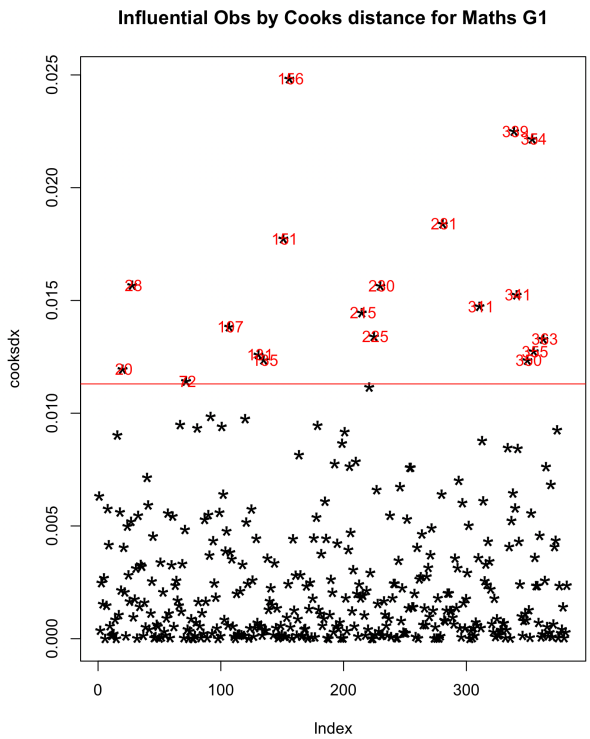


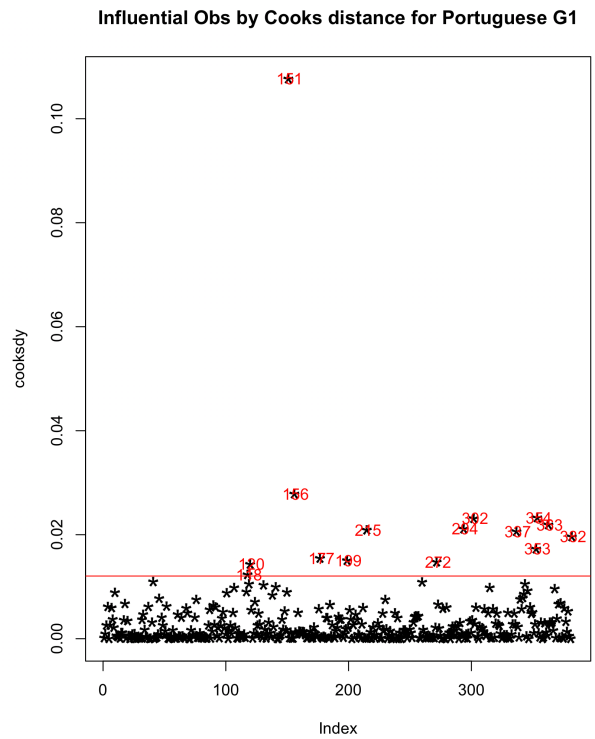
The relation shows that for the majority part of the data set the grades decreases as the absences increase although there is an interesting observation here which shows that upto a certain value of absences there in no inverse effect on grades. This is more clearly depicted in the cleaner data.

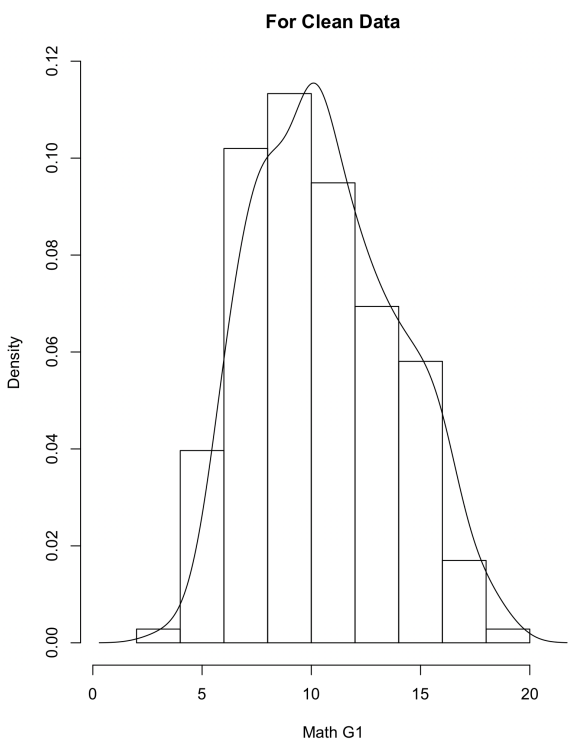


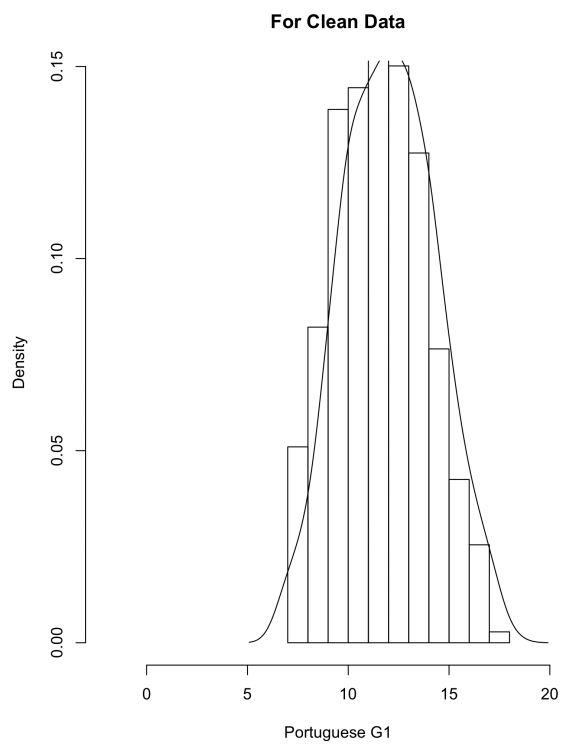
We can compare the Absences vs G1 grade trend for original as well as clean data(graph above).

Outlier Removal Strategy 2: Cooks Distance[1]



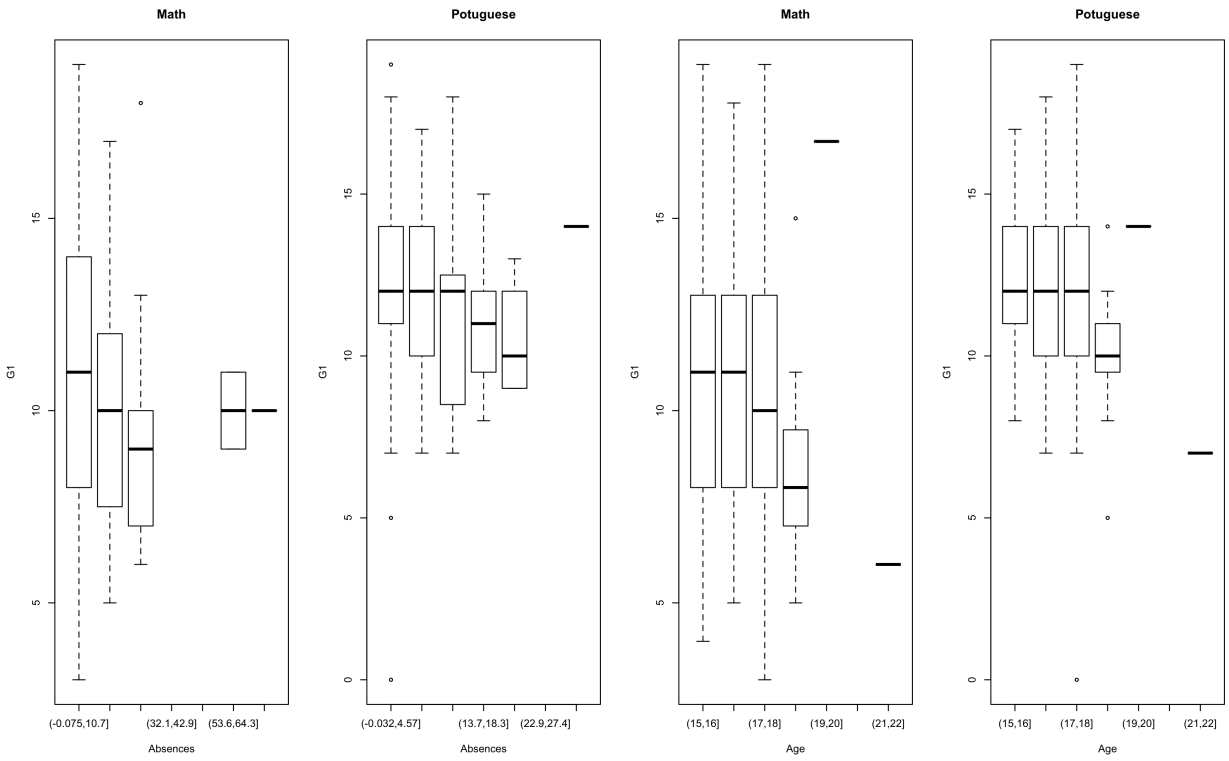




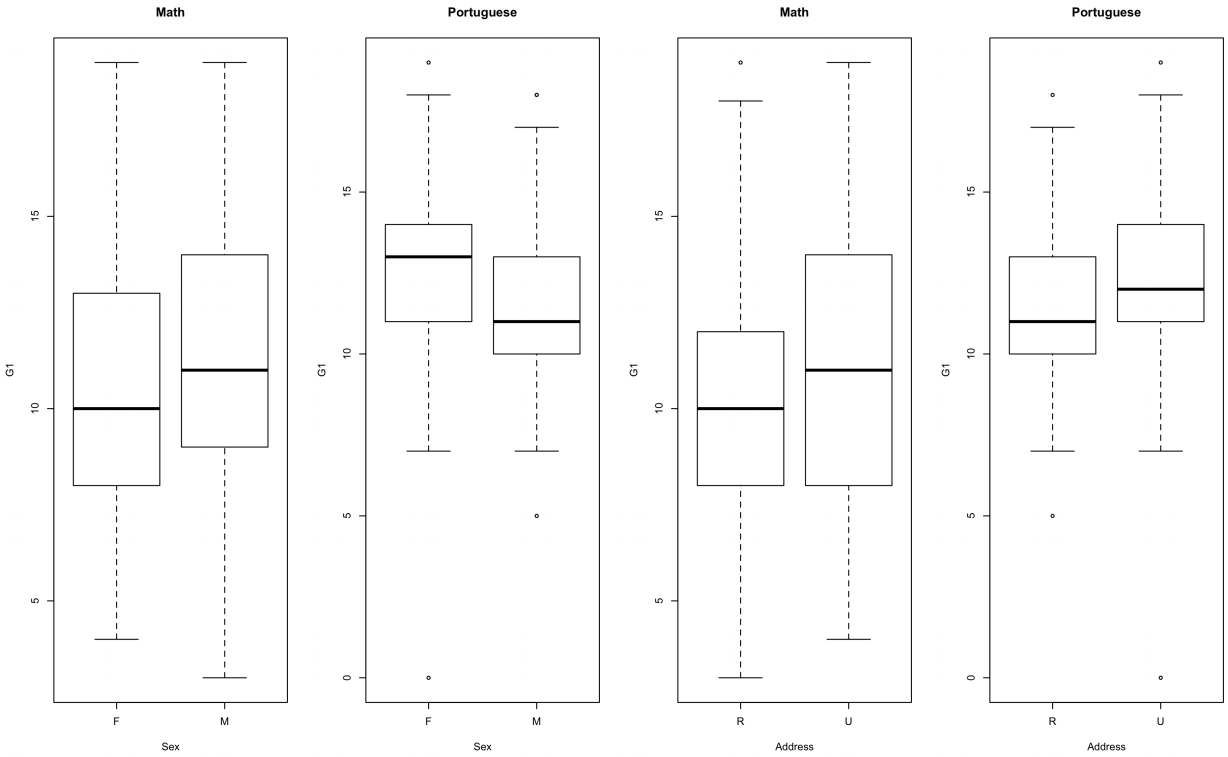


We eliminate the influential rows and check the accuracy of the prediction in question 2. The extreme observation from this computation for the given model comes from row 230 for Maths G1 and from row 151 for Portuguese G1. Above is the distribution of grades in Math and Portuguese after the data has been cleaned using the cook's distance.

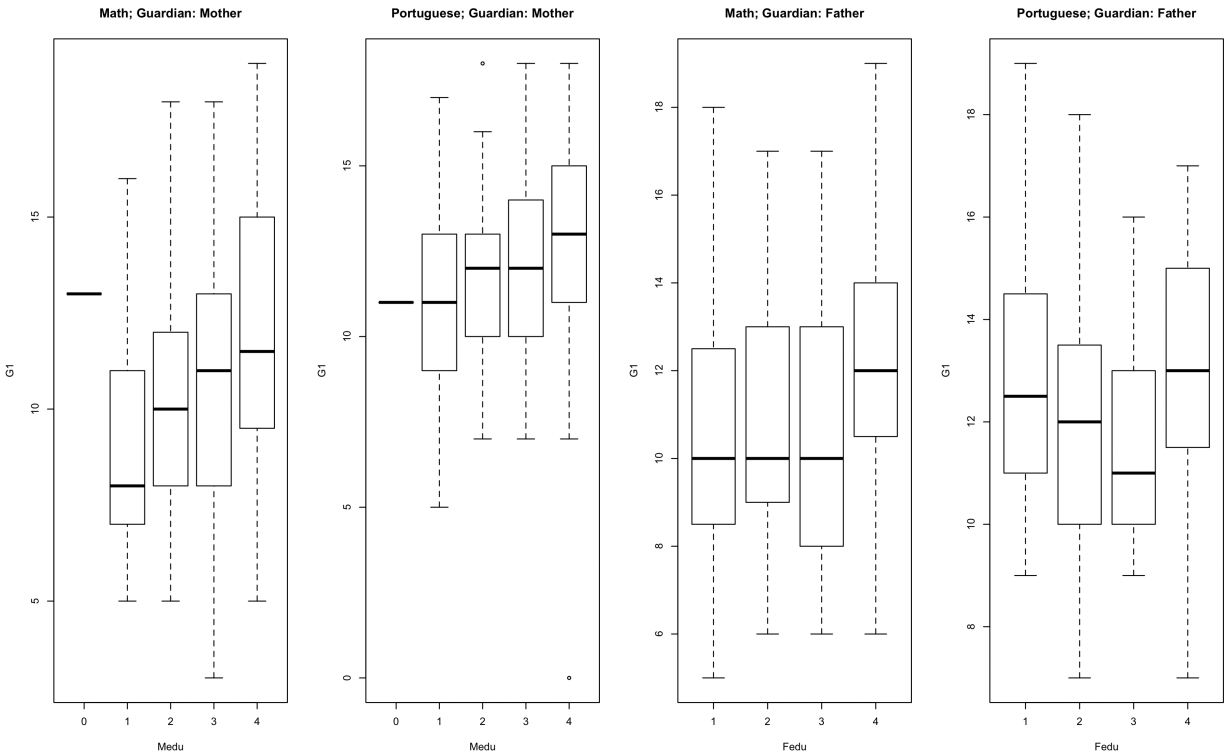
In the following graph the trend of box plots depict the inverse effect of absences on grades for Portuguese as well as Maths. We also observe an inverse relationship between Age and Grades which is more prominent in Maths.



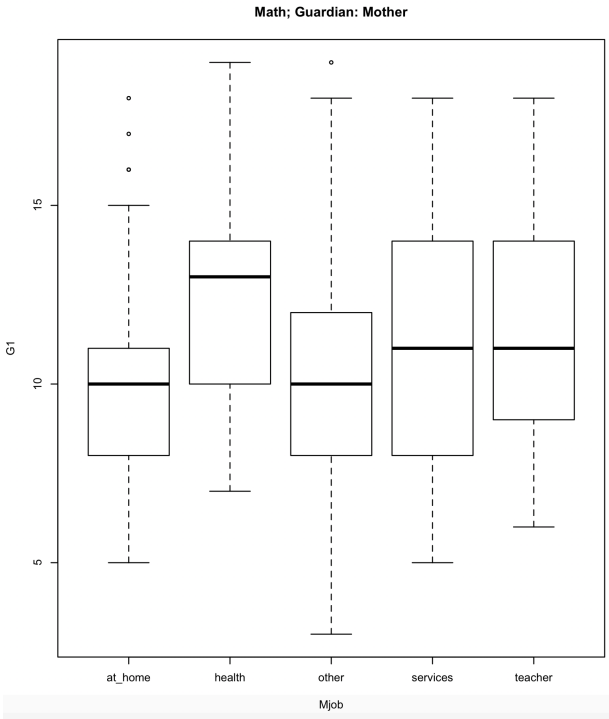
One-hot encoding[2] was done to categorical data: sex, age and address. Box-plots were plotted for these data sets. It can be deduced that females tend to perform poorer than boys in Maths while better than boys in Portuguese. Also, students in rural settings are doing poorly for both subjects. One interesting observation is majority of the students are above the median value of the grade for Portuguese which shows better chances of success in the subject for a student compared to Maths.

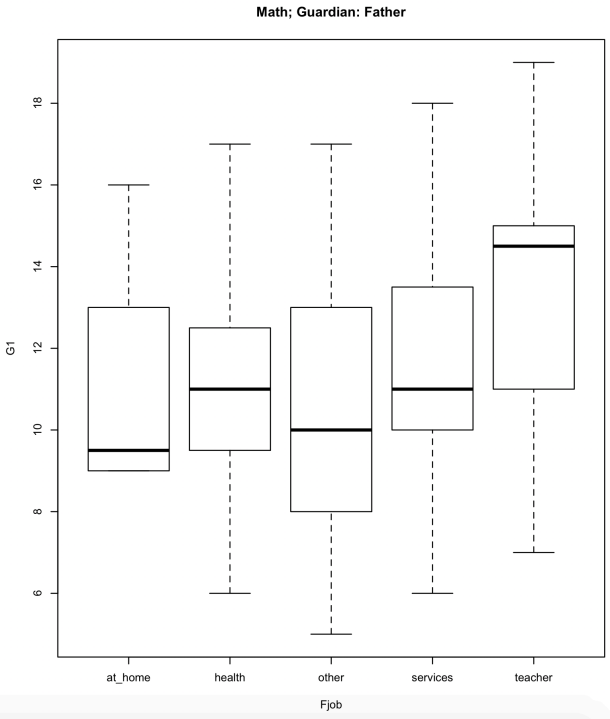


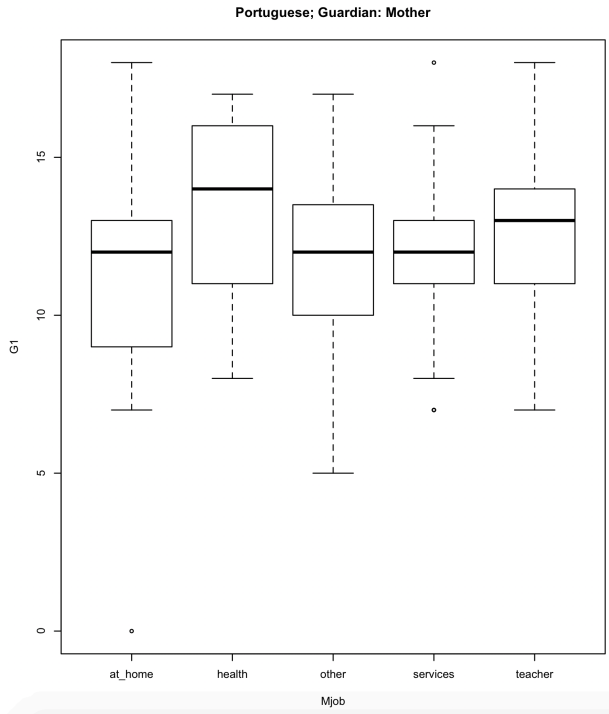
In the following graph we observe a direct relationship between the performance of students and the education level of mother when the guardian for the student is the mother. Although, when the guardian is father the students performance in Portuguese decreases as the education level of the father increases. Also, we see better performance of students when mother is the guardian.

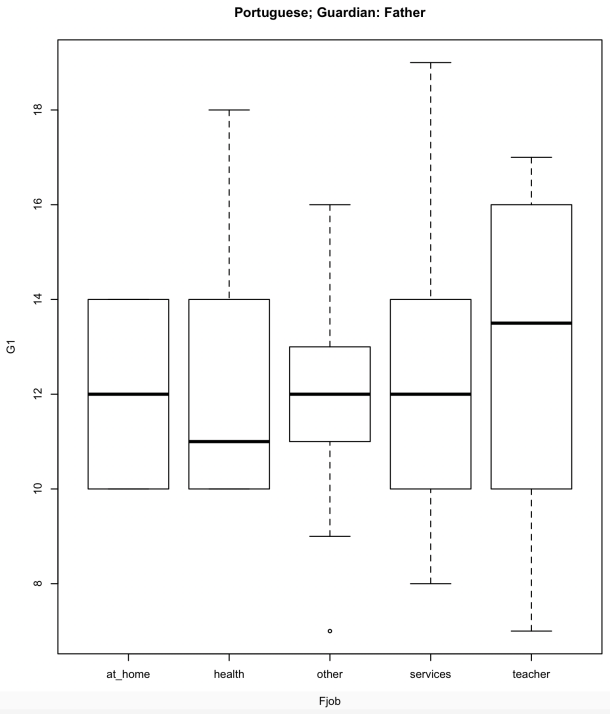


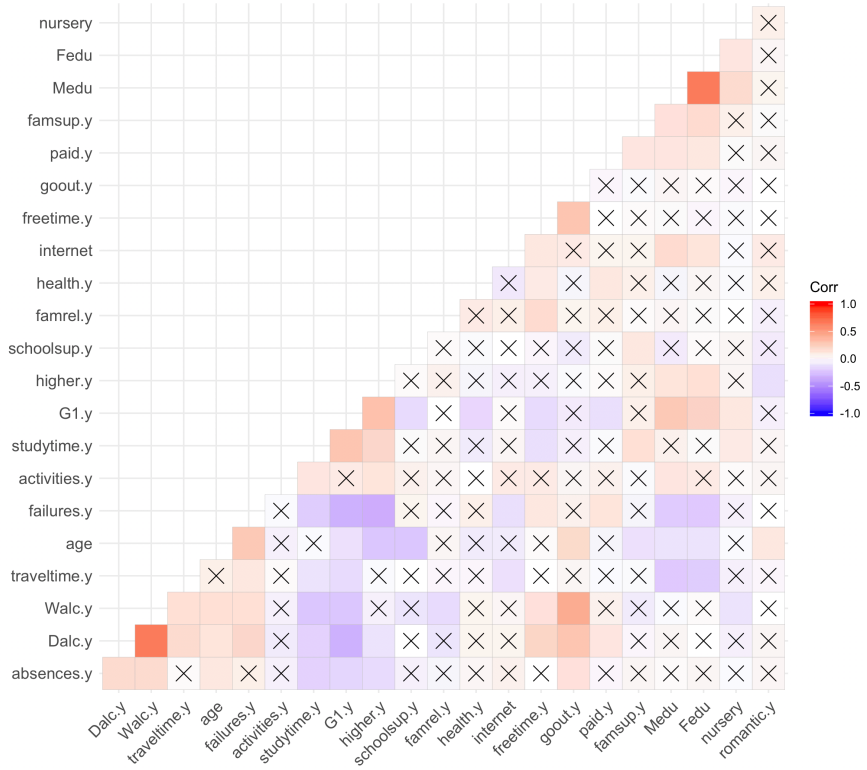
From the below plots we infer that students tend to perform better in cases when father is a teacher and/or mother is in health services.

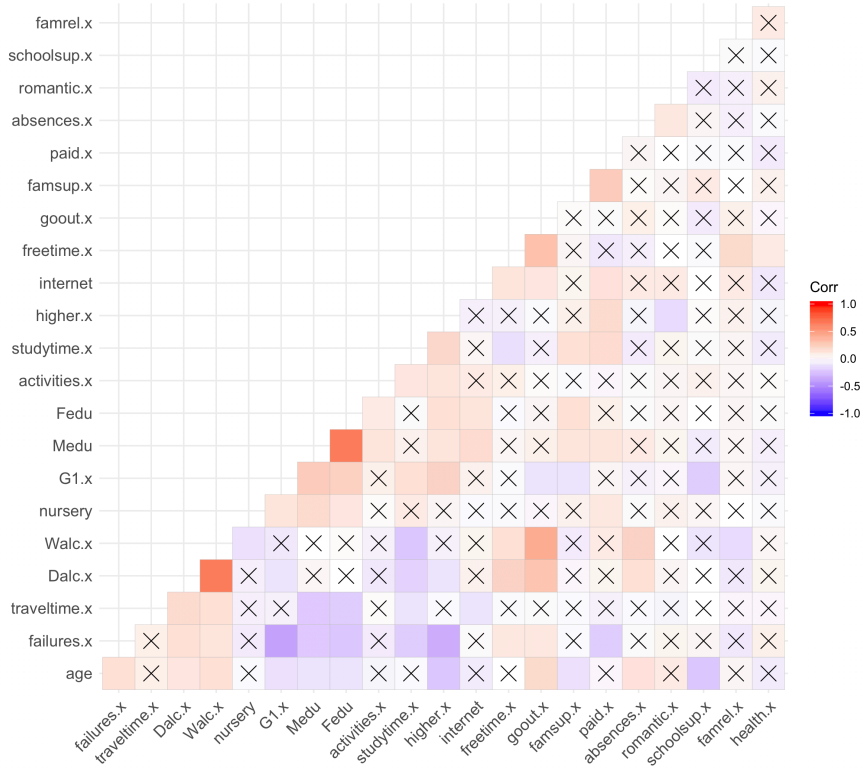




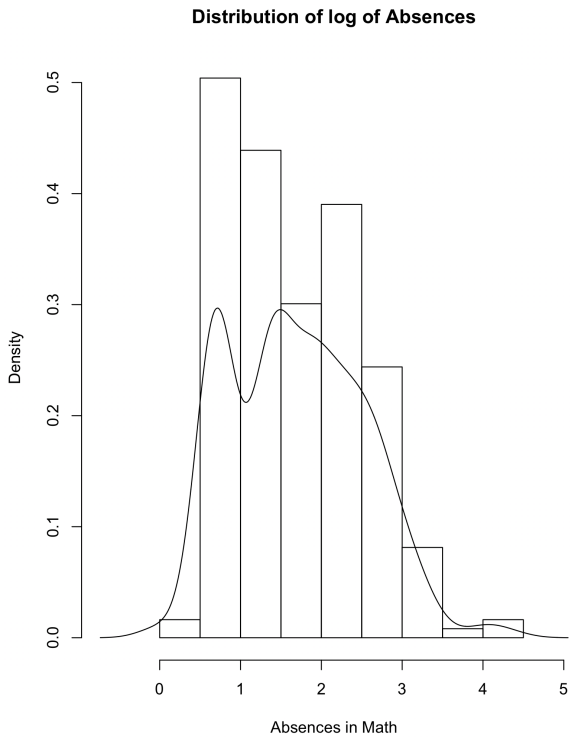


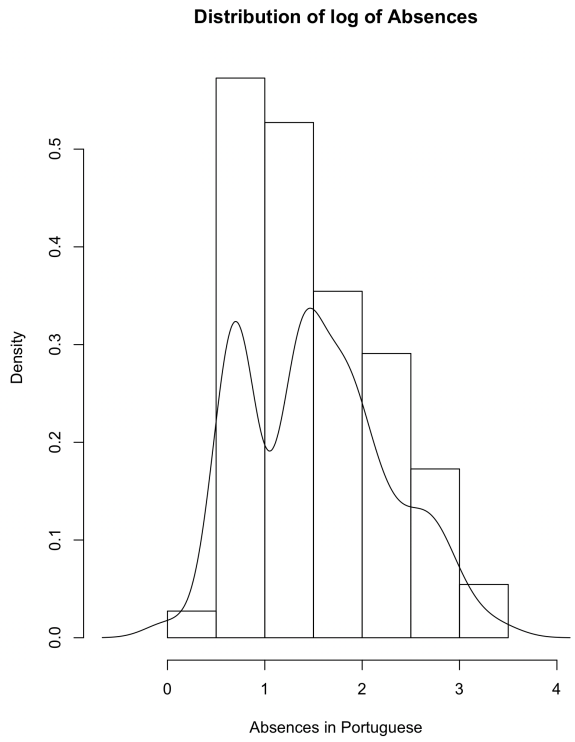






The above Correlation Plots show the effects of various variables on the grades of Maths(G1.x) and Portuguese(G1.y). On observing we find that the grades have a direct relationship with nursery, Mother's and Father's education, study time, goal for higher education and an inverse relationship with going out, age, past failures(strong relationship) and alcohol consumption.





We transform the variables absences.x and absences.y logarithmically. After taking the log values of absences.x and absences.y we observe a better distribution of grades with log of absences values.

2) Perform a multiple regression on the dataset you pre-processed in question one. The response are the first period grades. Use the lm() function in R.
a) Which predictors appear to have a significant relationship to the response.
b) What suggestions would you make to a first-year student trying to achieve good grades.
*c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?*

The multiple regression analysis between G1.x (left) and G1.y (right) and the rest of the variables(predictors) gives the following respons

Math

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.1244 -1.4472 -0.0484  1.5519  5.7671

Coefficients: (6 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.739940   3.520238   1.631 0.104013
age           -0.379541   0.132842  -2.857 0.004568 **
Medu           0.049754   0.181006   0.275 0.783596
Fedu           0.197361   0.160379   1.231 0.219421
traveltime.x  -0.702863   2.912504  -0.241 0.809464
studytime.x    1.360790   1.195555   1.138 0.255924
failures.x    -1.165285   0.236137  -4.935 1.32e-06 ***
famrel.x       0.891076   1.485940   0.600 0.549169
freetime.x     2.084402   2.119289   0.984 0.326120
goout.x       -1.070292   0.581364  -1.841 0.066590 .
Dalc.x        -2.640038   1.757355  -1.502 0.134056
Walc.x         0.485182   1.154819   0.420 0.674681
health.x      -0.257995   0.840939  -0.307 0.759209
absences.x     0.149800   0.146704   1.021 0.308013
traveltime.y   0.703734   2.924468   0.241 0.809998
studytime.y   -1.045093   1.195509  -0.874 0.382704
failures.y     1.248284   0.365598   3.414 0.000726 ***
famrel.y      -0.771101   1.485687  -0.519 0.604122
freetime.y    -1.890139   2.107350  -0.897 0.370463
goout.y        0.775986   0.571775   1.357 0.175733
Dalc.y         2.730891   1.757680   1.554 0.121292
Walc.y        -0.532929   1.138741  -0.468 0.640119
health.y       0.308050   0.841672   0.366 0.714619
absences.y    -0.164325   0.152118  -1.080 0.280883
G1.y           0.655399   0.067713   9.679  < 2e-16 ***
school         0.597165   0.514020   1.162 0.246241
sex            1.280738   0.299497   4.276 2.54e-05 ***
address       -0.006316   0.360917  -0.017 0.986049
famsize       -0.012042   0.298375  -0.040 0.967833
Pstatus       -0.007792   0.456182  -0.017 0.986383
Mjob          -0.075341   0.127377  -0.591 0.554636
Fjob           0.211190   0.174683   1.209 0.227599
reason         0.091955   0.112071   0.821 0.412566
nursery        0.488673   0.336970   1.450 0.148028
internet       0.194802   0.385033   0.506 0.613266
guardian.x    -0.031141   0.274928  -0.113 0.909890
schoolsup.x   -0.990500   0.420875  -2.353 0.019234 *
famsup.x      -0.901009   0.291885  -3.087 0.002208 **
paid.x        -0.025639   0.287783  -0.089 0.929067
activities.x  -0.090186   0.271570  -0.332 0.740047
higher.x       0.228651   0.728761   0.314 0.753922
romantic.x     0.001844   0.281581   0.007 0.994779
guardian.y           NA         NA      NA       NA
schoolsup.y          NA         NA      NA       NA
famsup.y             NA         NA      NA       NA
paid.y        -0.215875   0.553760  -0.390 0.696930
activities.y         NA         NA      NA       NA
higher.y             NA         NA      NA       NA
romantic.y           NA         NA      NA       NA
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.32 on 306 degrees of freedom
Multiple R-squared:  0.5146,    Adjusted R-squared:  0.4479
F-statistic: 7.723 on 42 and 306 DF,  p-value: < 2.2e-16
```

Portuguese

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1146 -1.0584 -0.0061  1.0669  3.6895

Coefficients: (6 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.460204   2.585428   2.499  0.01299 *
age            0.203503   0.098747   2.061  0.04016 *
Medu           0.224565   0.133108   1.687  0.09261 .
Fedu           0.002093   0.118765   0.018  0.98595
traveltime.x   2.531433   2.146799   1.179  0.23925
studytime.x   -1.062538   0.882937  -1.203  0.22975
failures.x     0.112496   0.181128   0.621  0.53501
famrel.x       0.204773   1.098246   0.186  0.85221
freetime.x    -1.776377   1.564699  -1.135  0.25715
goout.x        0.319262   0.431440   0.740  0.45987
Dalc.x         3.569434   1.286859   2.774  0.00588 **
Walc.x        -0.386585   0.853025  -0.453  0.65073
health.x      -0.059387   0.621288  -0.096  0.92391
absences.x    -0.069126   0.108483  -0.637  0.52447
G1.x           0.357636   0.036950   9.679  < 2e-16 ***
traveltime.y  -2.612177   2.155342  -1.212  0.22646
studytime.y    1.166139   0.881708   1.323  0.18696
failures.y    -0.553334   0.273340  -2.024  0.04380 *
famrel.y      -0.338250   1.097790  -0.308  0.75820
freetime.y     1.624366   1.555976   1.044  0.29733
goout.y       -0.280766   0.423335  -0.663  0.50769
Dalc.y        -3.744301   1.285815  -2.912  0.00386 **
Walc.y         0.346832   0.841255   0.412  0.68042
health.y      -0.026757   0.621877  -0.043  0.96571
absences.y    -0.167905   0.112173  -1.497  0.13547
school        -0.959896   0.376566  -2.549  0.01129 *
sex           -1.000398   0.220456  -4.538 8.17e-06 ***
address        0.396729   0.265643   1.493  0.13635
famsize        0.343435   0.219534   1.564  0.11876
Pstatus       -0.114917   0.336918  -0.341  0.73328
Mjob          -0.005802   0.094146  -0.062  0.95090
Fjob           0.193277   0.128873   1.500  0.13471
reason         0.075891   0.082764   0.917  0.35989
nursery       -0.469965   0.248325  -1.893  0.05936 .
internet      -0.341980   0.283871  -1.205  0.22925
guardian.x    -0.211795   0.202733  -1.045  0.29699
schoolsup.x   -0.640137   0.311560  -2.055  0.04076 *
famsup.x       0.402939   0.217732   1.851  0.06519 .
paid.x        -0.041954   0.212574  -0.197  0.84368
activities.x   0.196479   0.200330   0.981  0.32748
higher.x       1.285255   0.533385   2.410  0.01656 *
romantic.x    -0.060774   0.207975  -0.292  0.77032
guardian.y           NA         NA      NA       NA
schoolsup.y          NA         NA      NA       NA
famsup.y             NA         NA      NA       NA
paid.y        -0.310068   0.408780  -0.759  0.44872
activities.y         NA         NA      NA       NA
higher.y             NA         NA      NA       NA
romantic.y           NA         NA      NA       NA
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.714 on 306 degrees of freedom
Multiple R-squared:  0.5348,    Adjusted R-squared:  0.4709
F-statistic: 8.375 on 42 and 306 DF,  p-value: < 2.2e-16
```

It shows that 6 predictor variables: guardian.y, schoolsup.y, famsup.y, activities.y, higher.y and romantic.y show singularity which means they show perfect collinearity with one of the other variables present and are thus redundant. Thus we remove these variables in the next iteration.

We can see the F-statistic value to be fairly greater than 1 here and since we have 48 variables we can say that this provides the evidence against the null hypothesis and it can be stated that there exists some relationship between the response variables and the predictors.

Also we can observe that pvalues for G1.x are low for age, failures.x, goout.x, failures.y, G1.y, sex, schoolsup.x and famsup.x. For G1.y it is age, Medu, Dalc.x, G1.x, failures.y, Dalc.y, school, sex, schoolsup.x, famsup.x and higher.x. Thus, regressing G1.x and G1.y on the above predictors we get:

Math

```
Call:
lm(formula = G1.x ~ ., data = dcfin1x)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1622 -1.5738 -0.0466  1.5392  5.8376 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.74309    2.56411   2.630 0.008949 ** 
age         -0.38323    0.12288  -3.119 0.001978 ** 
Fedu         0.20255    0.12434   1.629 0.104277    
studytime.x  1.47978    0.82677   1.790 0.074409 .  
failures.x  -1.19388    0.22244  -5.367 1.52e-07 ***
failures.y   1.24117    0.33993   3.651 0.000304 ***
freetime.x   2.33198    0.70465   3.309 0.001039 ** 
freetime.y  -2.12060    0.70905  -2.991 0.002995 ** 
Dalc.x      -2.84624    1.12703  -2.525 0.012029 *  
absences.x   0.11058    0.13374   0.827 0.408946    
studytime.y -1.15921    0.82943  -1.398 0.163184    
goout.y      0.42136    0.32489   1.297 0.195580    
goout.x     -0.70853    0.32584  -2.174 0.030388 *  
Dalc.y       2.88395    1.12957   2.553 0.011130 *  
absences.y  -0.14938    0.14246  -1.049 0.295164    
G1.y         0.65445    0.06277  10.426  < 2e-16 ***
school       0.51029    0.47108   1.083 0.279508    
sex          1.27119    0.27518   4.619 5.55e-06 ***
Fjob         0.19379    0.16164   1.199 0.231457    
reason       0.09674    0.10442   0.927 0.354855    
nursery      0.47860    0.31775   1.506 0.132973    
schoolsup.x -0.99067    0.39768  -2.491 0.013230 *  
famsup.x    -0.89522    0.26735  -3.348 0.000908 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.261 on 326 degrees of freedom
Multiple R-squared:  0.5089,	Adjusted R-squared:  0.4757 
F-statistic: 15.35 on 22 and 326 DF,  p-value: < 2.2e-16
```

Portuguese

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.0566 -1.0702 -0.0499  1.0431  3.7739 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.73584    2.35724   2.433 0.015510 *  
age           0.18783    0.09431   1.992 0.047267 *  
Medu          0.22413    0.10072   2.225 0.026759 *  
traveltime.x  1.55256    1.14709   1.353 0.176860    
traveltime.y -1.62308    1.14914  -1.412 0.158794    
studytime.x  -1.12457    0.77616  -1.449 0.148350    
freetime.x   -1.19586    0.68725  -1.740 0.082811 .  
failures.y   -0.45527    0.22772  -1.999 0.046428 *  
freetime.y    1.03024    0.68566   1.503 0.133942    
Dalc.x        3.25425    0.98611   3.300 0.001076 ** 
studytime.y   1.23275    0.77682   1.587 0.113518    
goout.x       0.02886    0.09068   0.318 0.750527    
Dalc.y       -3.46192    0.98715  -3.507 0.000518 ***
absences.y   -0.19533    0.09863  -1.980 0.048510 *  
G1.x          0.34919    0.03448  10.129  < 2e-16 ***
school       -0.92370    0.36082  -2.560 0.010927 *  
sex          -1.05101    0.21095  -4.982 1.03e-06 ***
address       0.41959    0.25655   1.635 0.102928    
famsize       0.33936    0.21012   1.615 0.107281    
Fjob          0.20008    0.12179   1.643 0.101414    
reason        0.08920    0.07842   1.137 0.256183    
nursery      -0.46081    0.24169  -1.907 0.057470 .  
internet     -0.33827    0.27007  -1.253 0.211296    
guardian.x   -0.19663    0.19485  -1.009 0.313690    
schoolsup.x  -0.65106    0.29716  -2.191 0.029179 *  
famsup.x      0.35265    0.20373   1.731 0.084416 .  
activities.x  0.16714    0.19262   0.868 0.386206    
higher.x      1.23544    0.50191   2.461 0.014363 *  
paid.y       -0.34540    0.38964  -0.886 0.376036    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.69 on 320 degrees of freedom
Multiple R-squared:  0.5269,	Adjusted R-squared:  0.4855 
F-statistic: 12.73 on 28 and 320 DF,  p-value: < 2.2e-16
```

Thus, we get a linear model for our response variables G1.x and G1.y. Adding interaction between study time and gooier

Adding Interactions:

We add interactions between studytime.x & goout.x , absences.x & goout.x , studytime.x & absences.x , freetime.x & studytime.x , Dalc.x & studytime.x , (studytime.x, absences.x & goout.x), (studytime.x, freetime.x & goout.x), (freetime.x, absences.x & goout.x), studytime.y & goout.y , absences.y & goout.y , studytime.y & absences.y , freetime.y & studytime.y , Dalc.y & studytime.y , (studytime.y, absences.y & goout.y), (studytime.y, freetime.y & goout.y) and (freetime.y, absences.y & goout.y).

Although we don't find significant p-values for the above interactions but the Multiple R-squared parameter value and Adjusted R-squared value is higher for the model with interactions between (freetime.y, absences.y & goout.y) for G1.y: 0.5277,0.4848) and studytime.x & goout.x for G1.x(0.5121, 0.4775) as shown in the below figure.

```
Call:
lm(formula = G1.x ~ . + goout.x:studytime.x, data = dcfin1x)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5478 -1.5638 -0.0693  1.5375  6.2426 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          5.46575    2.70529   2.020 0.044162 *  
age                 -0.37409    0.12282  -3.046 0.002511 ** 
Fedu                 0.21035    0.12424   1.693 0.091390 .  
studytime.x          2.03918    0.91006   2.241 0.025720 *  
failures.x          -1.21931    0.22273  -5.474 8.80e-08 ***
failures.y           1.22023    0.33965   3.593 0.000378 ***
freetime.x           2.40424    0.70517   3.409 0.000733 ***
freetime.y          -2.18110    0.70904  -3.076 0.002275 ** 
Dalc.x              -2.87450    1.12525  -2.555 0.011088 *  
absences.x           0.11242    0.13351   0.842 0.400413    
studytime.y         -1.13668    0.82814  -1.373 0.170832    
goout.y              0.47739    0.32660   1.462 0.144786    
goout.x             -0.38307    0.39443  -0.971 0.332165    
Dalc.y               2.88155    1.12762   2.555 0.011060 *  
absences.y          -0.14766    0.14222  -1.038 0.299935    
G1.y                 0.65268    0.06267  10.414  < 2e-16 ***
school               0.49727    0.47035   1.057 0.291194    
sex                  1.27405    0.27472   4.638 5.11e-06 ***
Fjob                 0.19476    0.16137   1.207 0.228337    
reason               0.09302    0.10427   0.892 0.372971    
nursery              0.48237    0.31721   1.521 0.129316    
schoolsup.x         -1.01472    0.39733  -2.554 0.011110 *  
famsup.x            -0.86658    0.26761  -3.238 0.001327 ** 
studytime.x:goout.x -0.19831    0.13593  -1.459 0.145555    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.257 on 325 degrees of freedom
Multiple R-squared:  0.5121,	Adjusted R-squared:  0.4775 
F-statistic: 14.83 on 23 and 325 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = G1.y ~ . + goout.y:freetime.y:absences.y, data = dcfin1y)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0676 -1.0594 -0.0527  0.9968  3.7233 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      5.45326    2.38850   2.283 0.023080 *  
age                              0.18453    0.09448   1.953 0.051668 .  
Medu                             0.22681    0.10085   2.249 0.025202 *  
traveltime.x                     1.70120    1.16472   1.461 0.145107    
traveltime.y                    -1.77159    1.16671  -1.518 0.129889    
studytime.x                     -1.08727    0.77826  -1.397 0.163371    
freetime.x                      -1.25251    0.69182  -1.810 0.071166 .  
failures.y                      -0.45462    0.22788  -1.995 0.046893 *  
freetime.y                       1.13305    0.69958   1.620 0.106303    
Dalc.x                           3.36402    0.99749   3.372 0.000837 ***
studytime.y                      1.19638    0.77885   1.536 0.125505    
goout.x                          0.07513    0.10960   0.686 0.493497    
Dalc.y                          -3.56576    0.99740  -3.575 0.000404 ***
absences.y                      -0.05938    0.20576  -0.289 0.773064    
G1.x                             0.34990    0.03451  10.138  < 2e-16 ***
school                          -0.88166    0.36536  -2.413 0.016379 *  
sex                             -1.04663    0.21117  -4.956 1.17e-06 ***
address                          0.42822    0.25698   1.666 0.096627 .  
famsize                          0.33573    0.21032   1.596 0.111409    
Fjob                             0.20189    0.12190   1.656 0.098669 .  
reason                           0.09230    0.07858   1.175 0.241000    
nursery                         -0.46101    0.24186  -1.906 0.057530 .  
internet                        -0.33569    0.27028  -1.242 0.215140    
guardian.x                      -0.19036    0.19516  -0.975 0.330101    
schoolsup.x                     -0.64810    0.29739  -2.179 0.030041 *  
famsup.x                         0.34851    0.20394   1.709 0.088441 .  
activities.x                     0.16489    0.19277   0.855 0.392981    
higher.x                         1.21145    0.50326   2.407 0.016642 *  
paid.y                          -0.36758    0.39101  -0.940 0.347888    
freetime.y:absences.y:goout.y -0.01217    0.01616  -0.753 0.452005    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.691 on 319 degrees of freedom
Multiple R-squared:  0.5277,	Adjusted R-squared:  0.4848 
F-statistic: 12.29 on 29 and 319 DF,  p-value: < 2.2e-16
```

This shows that the model is able to fit about 51% of the data in the DataSet for G1.x. We can also see that the predictors age, studytime.x , failures.x, failures.y, freetime.x, freetime.y, Dalc.x, Dalc.y, sex, schoolsup.x and famsup.x are the predictors having significant association with the response G1.x while the predictors Medu, failures.y, Dalc.x, Dalc.y, school, sex, schoolsup.x and higher.x have significant association with the response G1.y. The coefficient variable for failures shows very strong negative relation (-1.21: failure.x with G1.x) with the predictor which shows

students with past failures have more chance to perform poorly and thus need to improve on other parameters to get good grades.

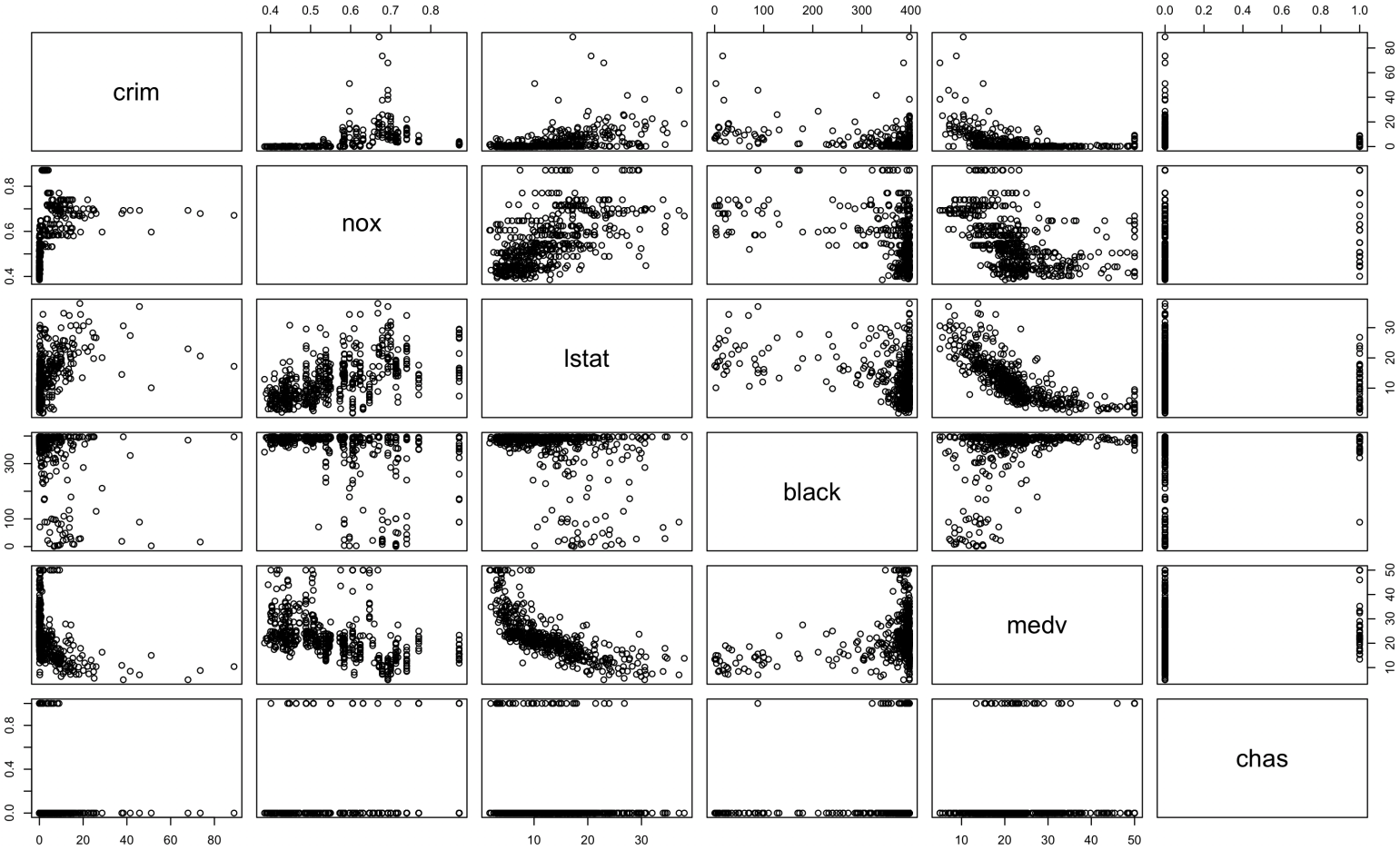
The chances of good grades increases in the following cases:

1. More time devoted to study
2. Reducing the Alcohol consumption or eliminating it
3. School support and Family support also helps in improving grades

3) ISL textbook exercise 2.10 modified: This exercise concerns the boston housing data in the MASS library (>library(MASS) >data(Boston)).

a) Make pairwise scatterplots of the predictors, and describe your findings.
b) Are any of the predictors associated with per capita crime rate?
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 seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Following are the pairwise scatter plots between the predictors:



The Boston data set contains the following predictors:

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per \$10,000.

ptratio: pupil-teacher ratio by town.

black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in \$1000s.

The above scatter plot shows the following relations between the predictors:

a. crim & lstat : The towns with less people belonging to lower status have a high density of crime rate.

b. crim & black : There are more instances of crime per capita in towns with more number of black people

c. crim & medv : The number of crime instances per capita is high in towns where the median value of homes is low.

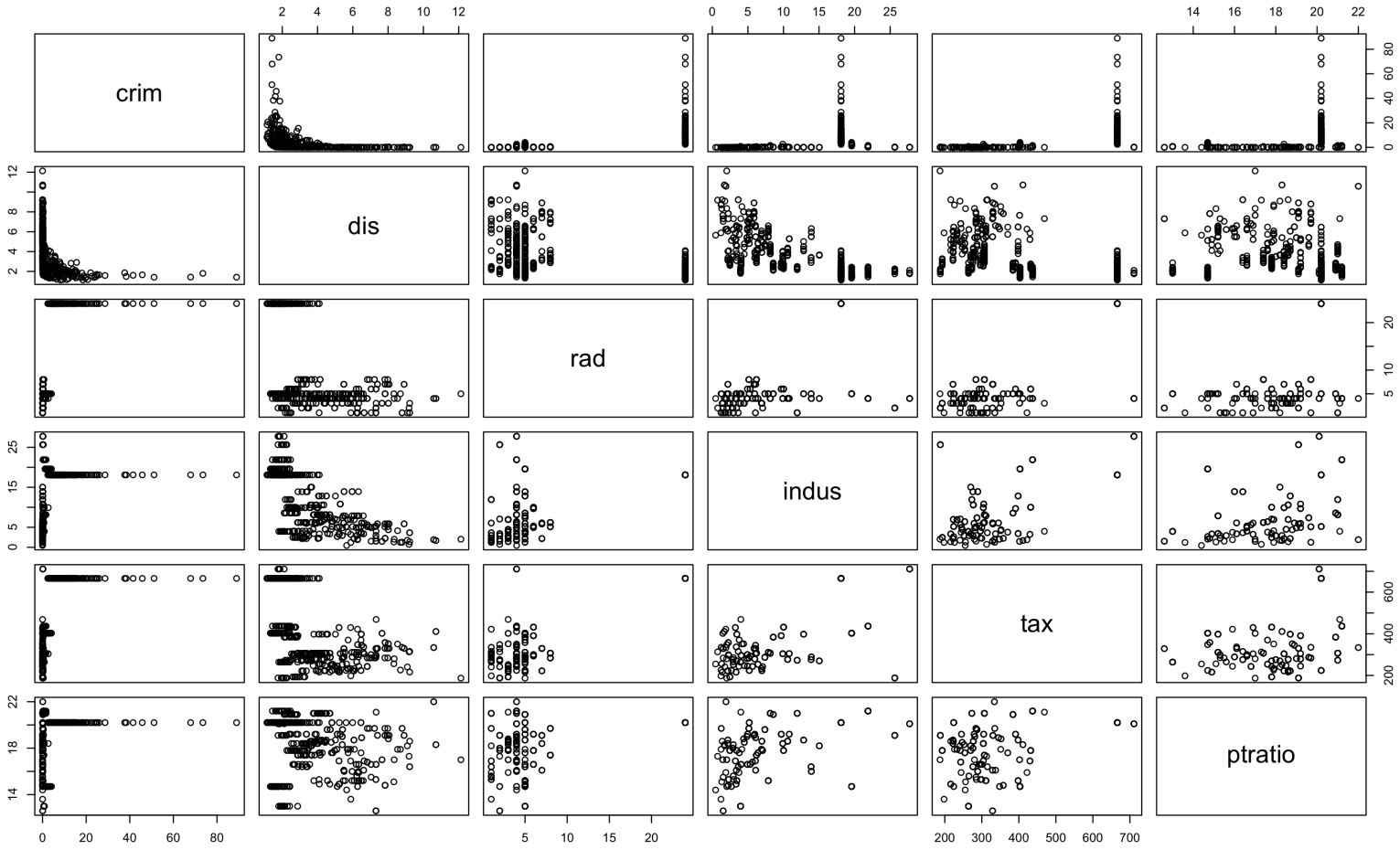
d. crim & chas : The crimes would be more in a town where Charles river is not surrounding the town

e. nox & lstat : A weak inference can be formed that nitrogen oxides concentration distribution is denser where there are less people in the lower status.

f. nox & medv : The towns where the owner occupied homes have a lower median value have higher concentration of Nitrogen Oxide.

g. nox & chas : If a town is not bounded by Charles River, there would be more Nitrogen Oxide concentration.

h. lstat & medv : The towns with more people in lower status have low median value of owner occupied homes.



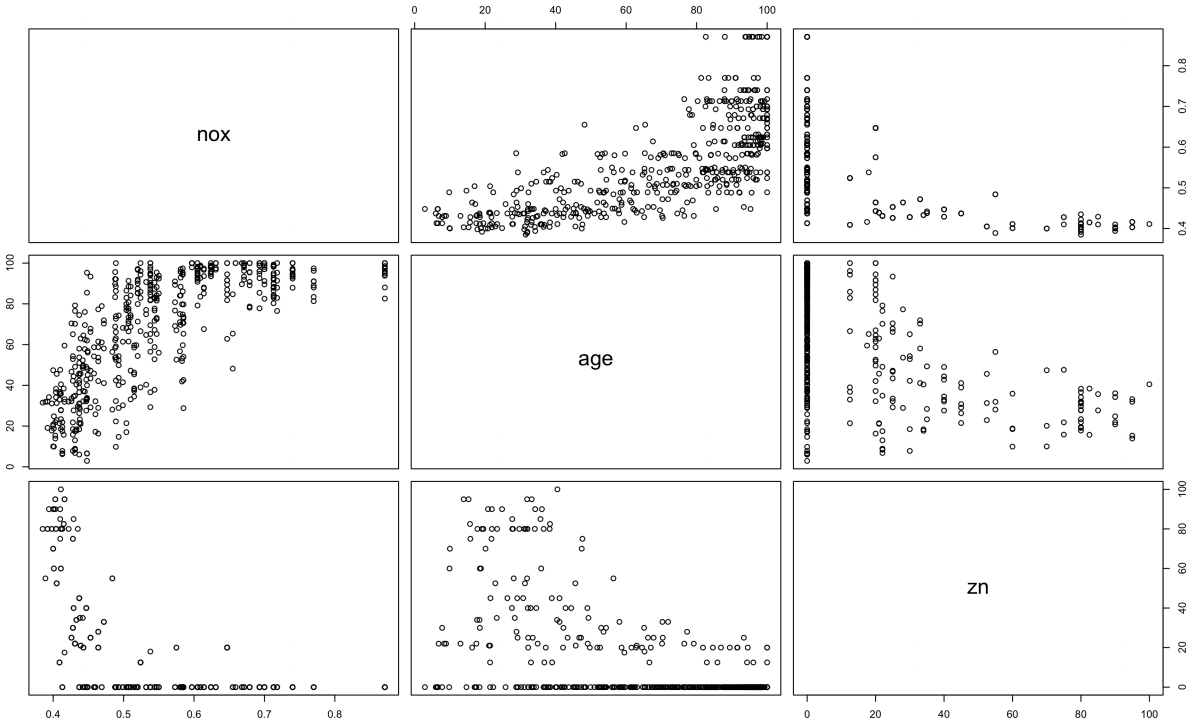
For the above pairwise scatterplot following can be deduced:

a. crim & dis : Closeness to five Boston Employment centers increases the density of crime in the towns.

b. crim & rad : If a town has a better index of accessibility to radial highways, the density of crimes in those towns tend to be higher.

c. crim & tax : Towns with higher taxes have a high density of crime.

d. crim & pratio : If the pupil teacher ratio is on the higher side, the density of crime is also high for a particular town.

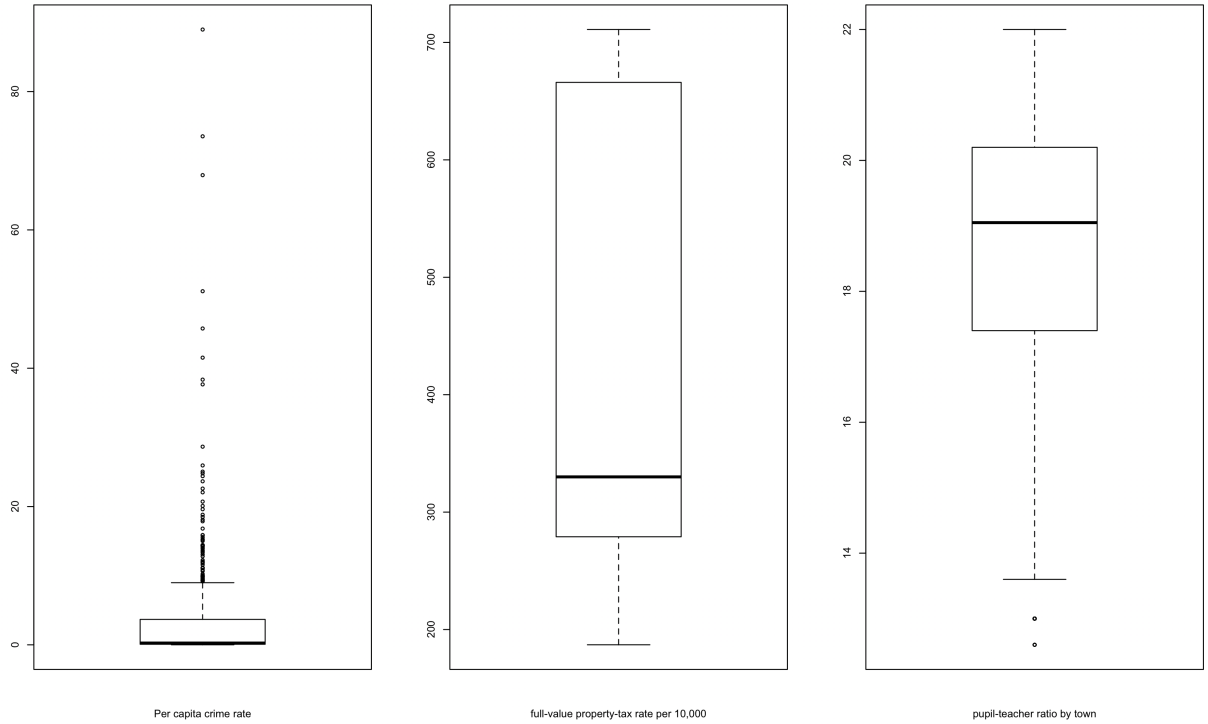


For the above pairwise scatterplot following can be deduced:

a. nox & age : We observe a positive relationship here and as the proportion of owner-occupied units built prior to 1940 increases: the nitrogen oxides concentration (parts per 10 million) also increases.

b. nox & zn : The concentration of Nitrogen oxide will be more where the proportion of residential land zoned over 25000 sq. ft. will be less

We can see that the following predictors have associations with the per capita crime rate in some form: **lstat, black, medv, tax, dis, rad, ptratio**



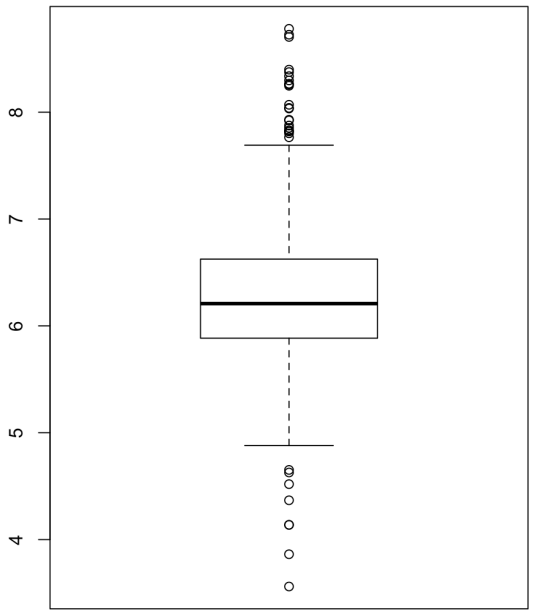
We can see in the above graph that there are some suburbs which have particularly high crime rates, tax rates and Pupil-teacher ratio.

For crime rates there are many suburbs for which the values are above the upper whisker which means there are large number of suburbs which have crime rates more than the median + 1.5*IQR. The crim value range from 0.0063 to 88.976 which means the per capita crime rate span over an interval of 88.9698.

For tax we can see there are many values which are above the median value which also indicates high tax rates for many suburbs. The tax rate span over an interval of 524 from 187 % to 524%.

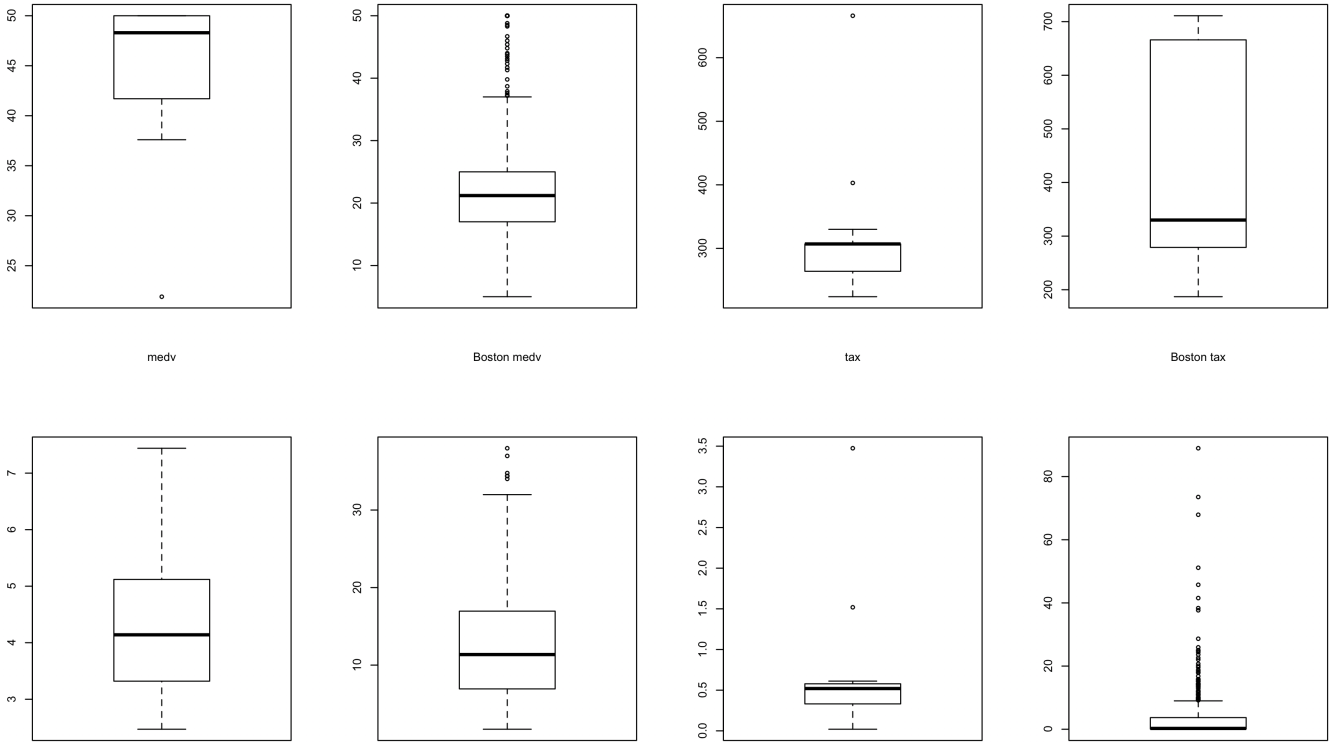
For ptratio the distribution shows that the box plot is towards the higher side and there are some suburbs which have high ptratio which lie on the upper whisker of the box plot. The ptratio range from 12.6 to 22.0.

64 suburbs average more than seven rooms per dwelling while **13 suburbs** average more than eight rooms per dwelling. For the suburbs that average more than 8 rooms per dwelling, the mean of medv(44.2) is significantly higher than for the full data set(22.53). Thus, for the suburbs which have more than 8 rooms per dwelling on an average have a higher mean value of the median value of homes. Similarly, the mean per capita crime rate (0.7188) is also lower than for the full data set(3.6135): relatively safer to live. The mean tax rate is 325.077 for this group which is



average number of rooms per dwelling

lower than the mean tax rate for the full data set(408.237). In the suburbs having more than 8 rooms per dwelling on an average there are less percentage of people (mean = 4.31) who belong to lower status of the populaion as compared to the full data set.



4) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k. The zipcode data is available in the ElemStatLearn package. Note that you do not have to divide the data into test and training as it is done for you.
> ls("package:ElemStatLearn") >?zip.test
>?zip.train

For Linear Regression:

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.59281 -0.09679 -0.00057  0.09365  0.70122 
```

```
Residual standard error: 0.1745 on 1132 degrees of freedom
Multiple R-squared:  0.9005,    Adjusted R-squared:  0.878 
F-statistic: 40.01 on 256 and 1132 DF,  p-value: < 2.2e-16
```

After performing linear regression, we can fit the linear model to get the predicted value. Thus we calculate the Mean Squared error = Mean of (actual value - predicted value)^2

Now, we can check the accuracy of the model by checking the prediction compared to the actual value. For how many actual values of 2 and 3 has the model actually predicted 2 and 3.

For K- nearest neighbor classification, we compute the Mean Squared error for different values of K for both data sets. The errors increase as the KNN values increase.

On comparing we find that the performance of KNN is better than the linear regression model.

SOURCES:

1. https://www.analytics-link.com/single-post/2017/08/25/HOW-TO-R---One-Hot-Encoding

2. http://r-statistics.co/Outlier-Treatment-With-R.html

3. https://stackoverflow.com/questions/47922184/convert-categorical-variables-to-numeric-in-r/47922653#47922653 - code to convert categorical values to numerical values

4. http://jamesmarquezportfolio.com/correlation_matrices_in_r.html - to plot correlation plot