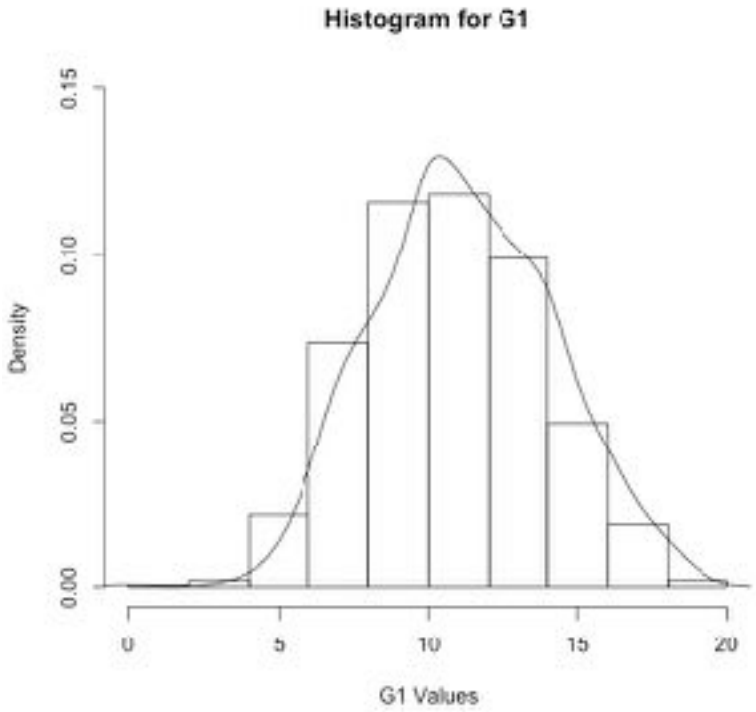
September 14, 2018

Homework 1

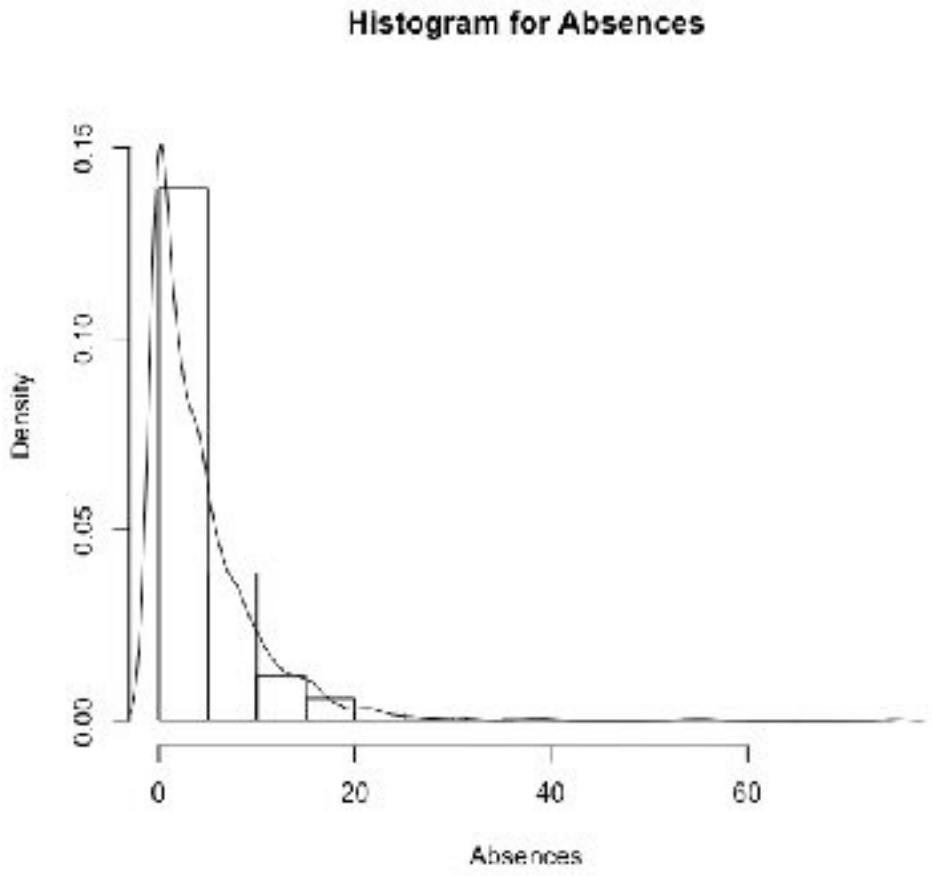
Q1) Consider the Student Performance Data Set on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance). Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory 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.

Solution 1: The student dataset we are dealing with has 33 independent variables and the target variable, G1 (First Period Grades), holds numeric values from 0 to 20.

Univariate Analysis:

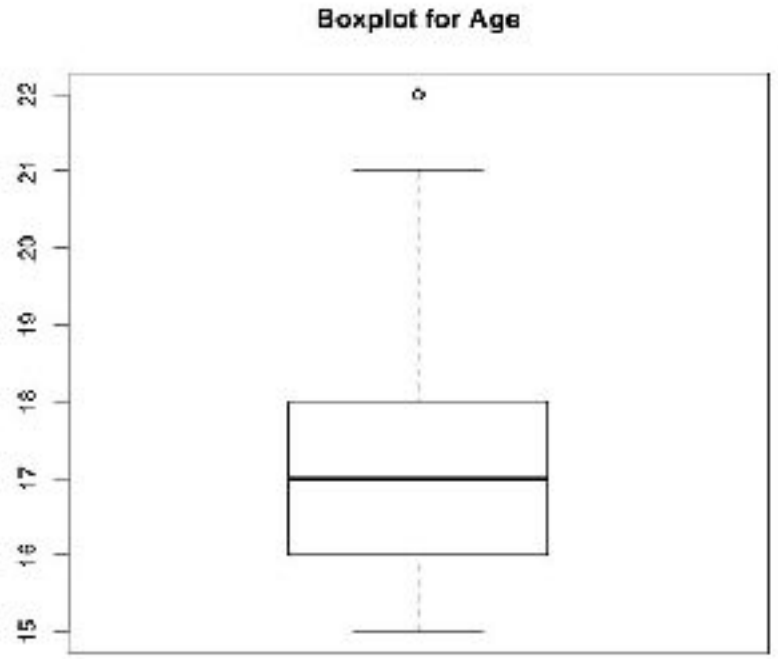


The histogram was plotted to understand the distribution of the target variable. The graph shows normal distribution with slight skewness to the left. I tried correcting the skewness by taking square root of the variable, however, this aggravated the skewness even further.



This graph indicates the number of absences by students. This looks like an exponentially decreasing curve. From the graph we can infer that large number of students have taken lesser number of absences.

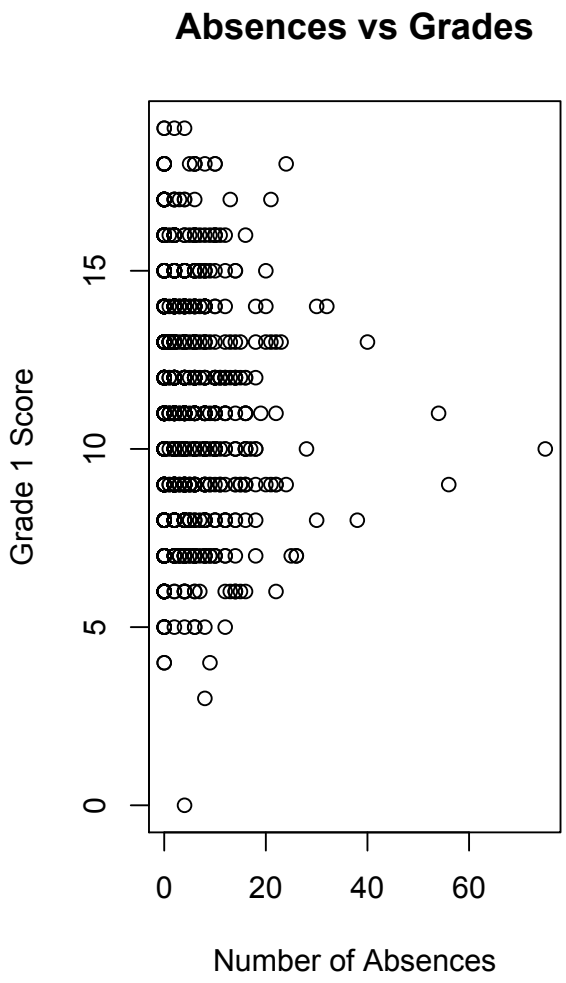
Next, we move on to analyze the independent variables and their relationship with the target.

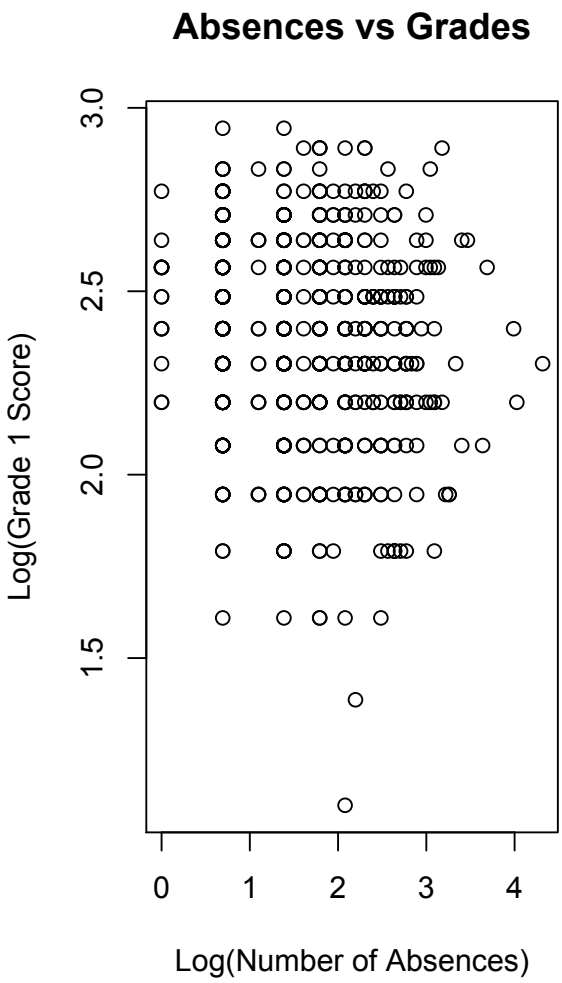


Here, looking at the Age parameter, we can see that the data is distributed between 15 and 21. The median is at 17. The first and third quartile are at 16 and 18 respectively. There are very few viz. two entries for the age 22. Since the point for age 22 doesn't contribute much to the distribution, I have considered them outliers, and removed the entries.

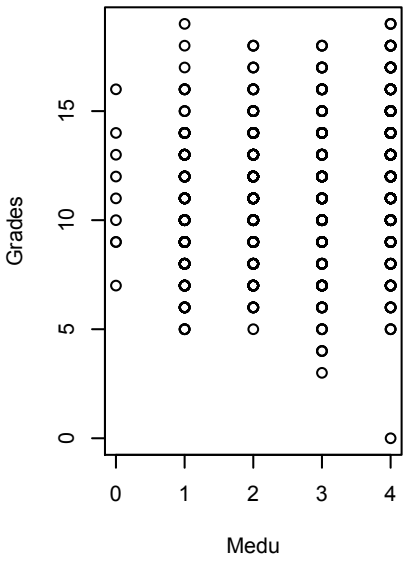
Bivariate Analysis:

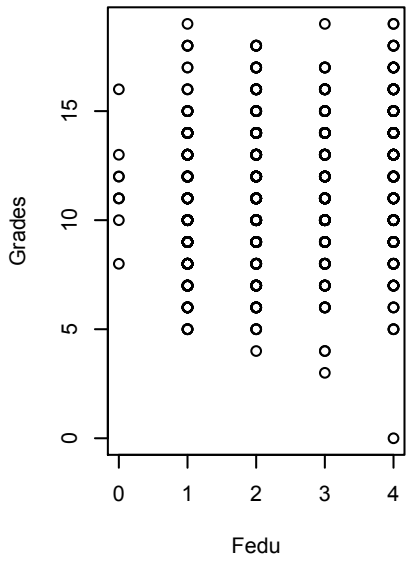
Motivation to do the bivariate analysis was identifying those predictors that directly impact the target.

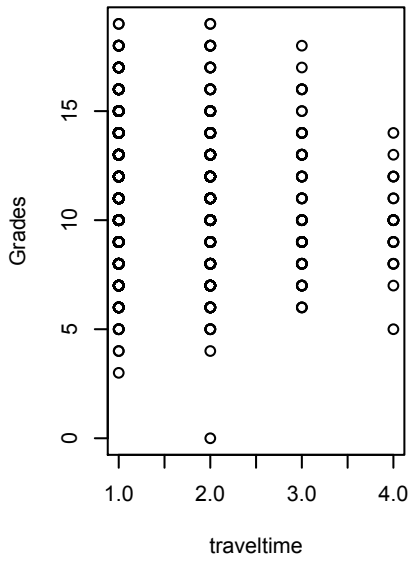


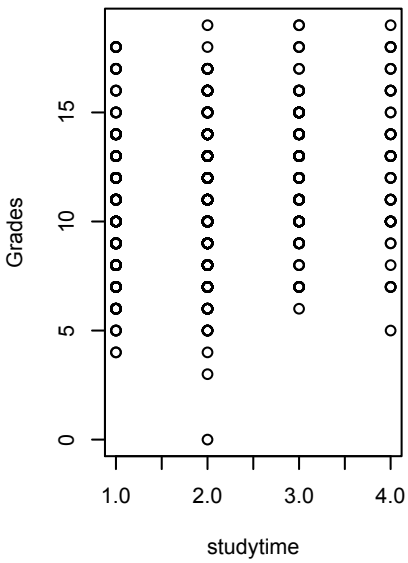


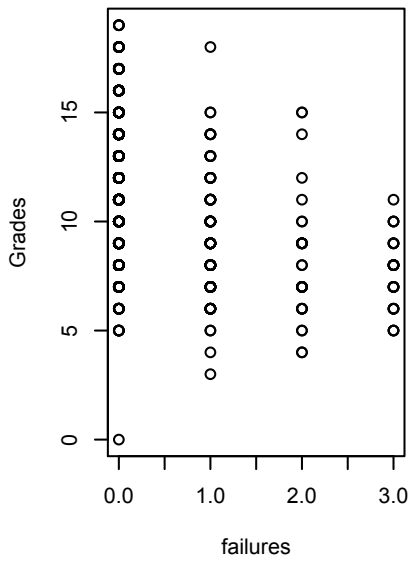
This graphs tries to plot a relationship between the absences and grades. The graph on the left shows data-points concentrated in the region between 0-20 and 7-15 on the X-axis and Y-axis respectively. I chose to perform log transformation on the variables to spread them apart, but it doesn't show a direct relationship or a pattern between absences and grades.

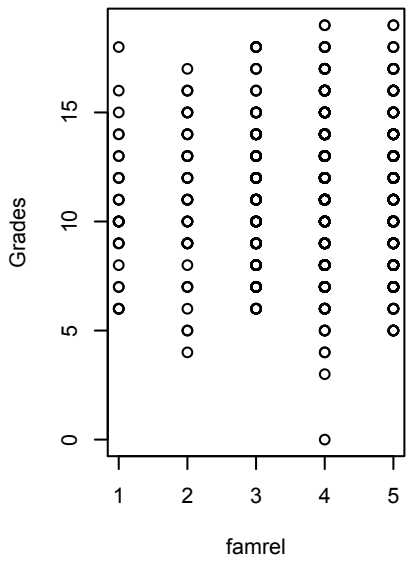

Grades
0
5
10
15
0
1
2
3
4
Medu

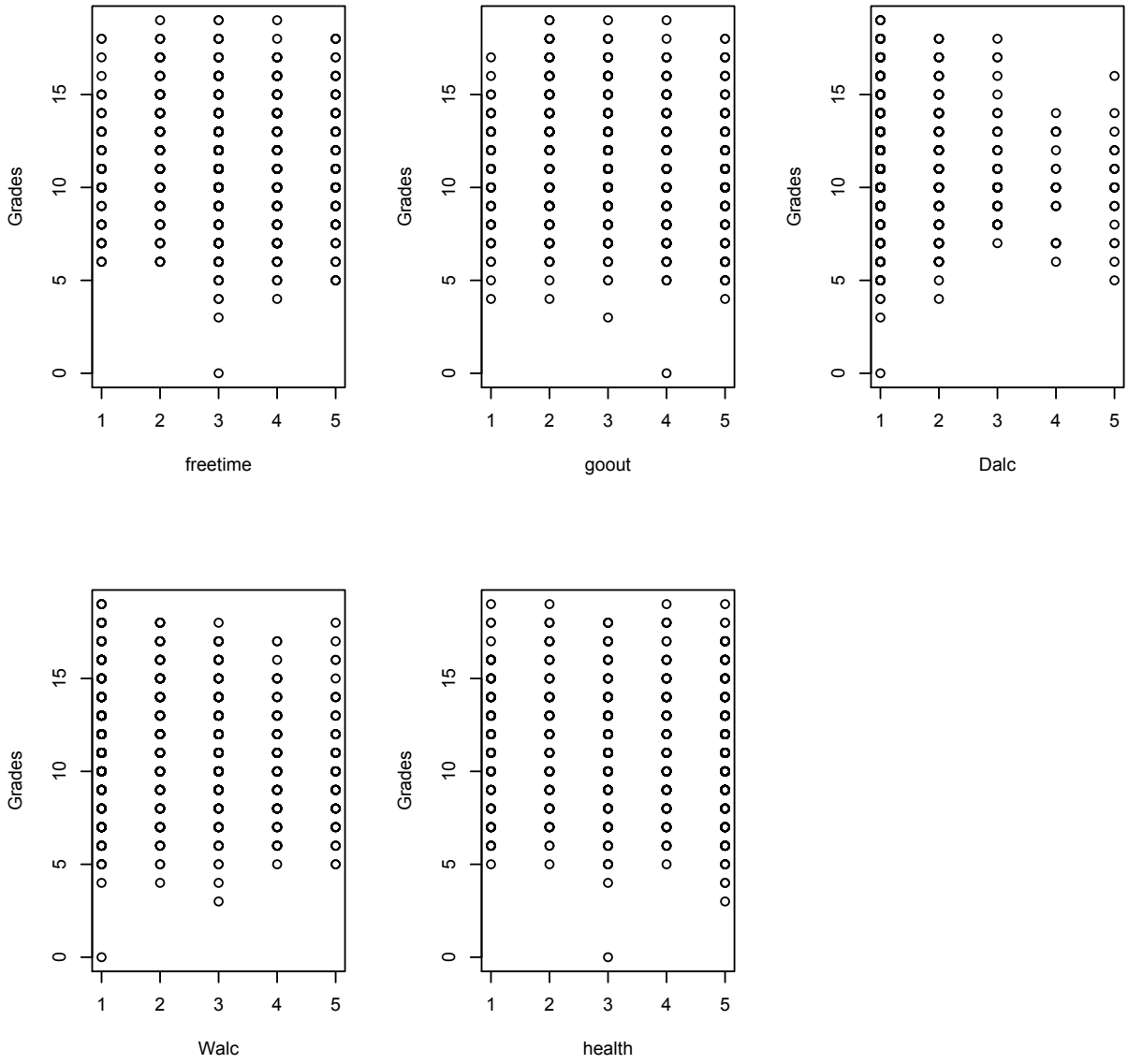

Grades
0
5
10
15
0
1
2
3
4
Fedu


Grades
0
5
10
15
1.0
2.0
3.0
4.0
traveltime

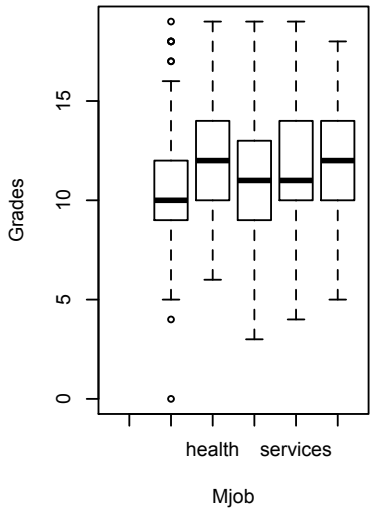

Grades
0
5
10
15
1.0
2.0
3.0
4.0
studytime

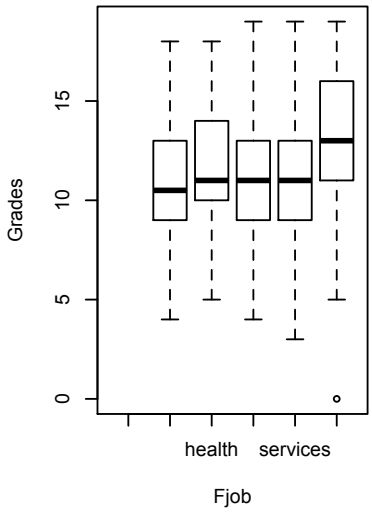

Grades
0
5
10
15
0.0
1.0
2.0
3.0
failures

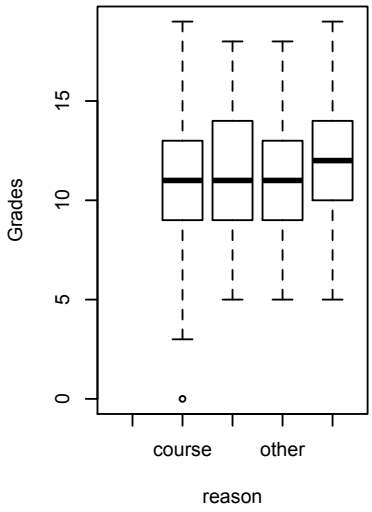

Grades
0
5
10
15
1
2
3
4
5
famrel

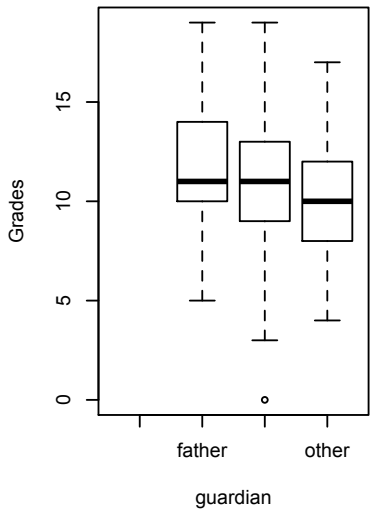


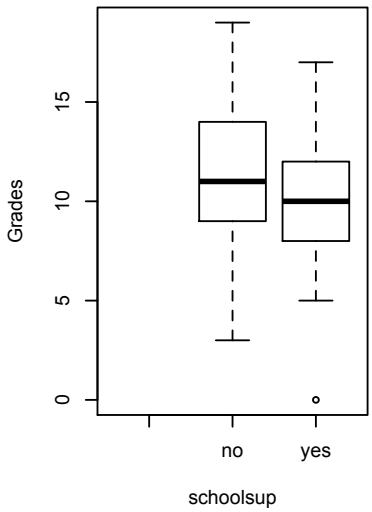
Performed bivariate analysis between the target value and ordinal predictors using scatter plots. There appears to be a relationship between grades and failure. The rest do not appear to hold a relationship.

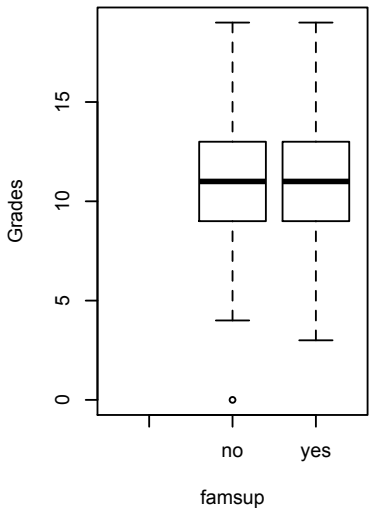

Grades
0
5
10
15
health
services
Mjob

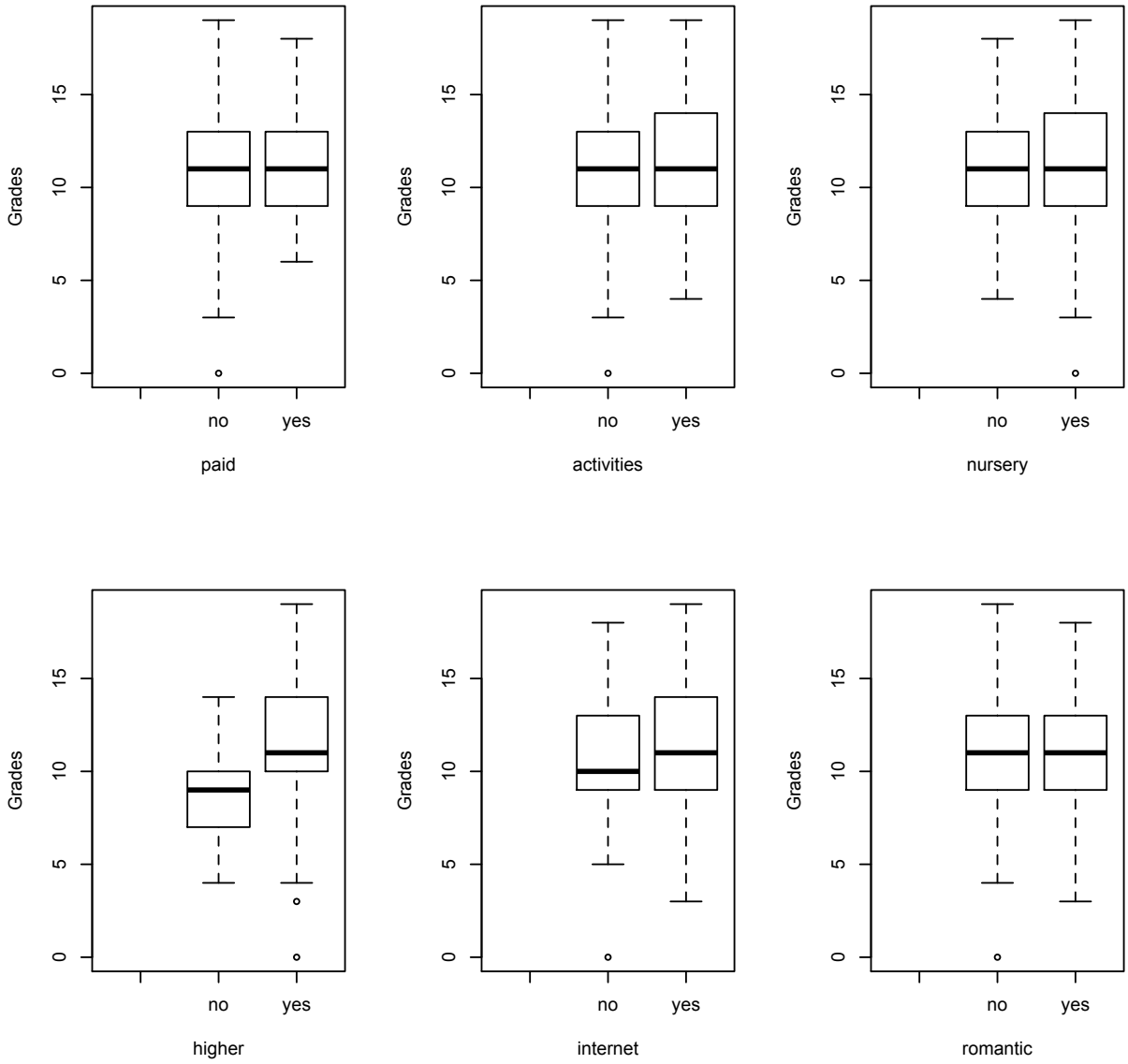

Grades
0
5
10
15
health
services
Fjob


Grades
0
5
10
15
course
other
reason


Grades
0
5
10
15
father
other
guardian


Grades
0
5
10
15
no
yes
schoolsup


Grades
0
5
10
15
no
yes
famsup



Performed bivariate analysis between target and categorical predictor values. I do not observe any obvious relationship from these graphs. I have retained the entry with grade value 0 because it doesn't seem like an outlier, because it may actually be a valid score due to factors not so obvious.

We can now go ahead and model the data to see if how significantly the independent variables affect the target value.

Q2) 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?

Solution 2:

Performed multiple linear regression on the pre-processed data.

A) The predictors studytime, failures, schoolsup, famsup, paid, higher, Fjob, Mjob, famsize, school, goout, health and absences appear to have a significant relationship to the response.

B) As ‘studytime’ and ‘schoolsup' are highly significant predictors in our model, I would suggest that he/she tries to maintain an average of 2.5hours of weekly study time, and if possible, to acquire additional educational support from the school.

C) I used the * symbol between *Mjob* and *higher* because their combination boost the model performance. This resulted in a highly significant interaction *Mjobhealth*, a nd nominally significant interactions *higheryes* and *Mjobteacher:higheryes*. I used : symbol between *studytime* and *schoolsup* which gave a nominally significant interaction *studytime:schoolsupyes*.

```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               10.79601    0.66706  16.184  < 2e-16 ***
studytime                  0.59708    0.10594   5.636 2.25e-08 ***
failures                  -1.27117    0.12985  -9.790  < 2e-16 ***
schoolsupyes              -0.04300    0.70447  -0.061 0.951339
famsupyes                 -0.31487    0.16865  -1.867 0.062191 .
paidyes                   -0.69533    0.20040  -3.470 0.000543 ***
health                    -0.12593    0.05657  -2.226 0.026226 *
Fjobhealth                 0.05592    0.53661   0.104 0.917021
Fjobother                 -0.16291    0.34697  -0.470 0.638800
Fjobservices              -0.31023    0.36214  -0.857 0.391843
Fjobteacher                1.39697    0.47420   2.946 0.003293 **
famsizeLE3                 0.26000    0.17629   1.475 0.140558
schoolMS                  -0.85527    0.19517  -4.382 1.30e-05 ***
goout                     -0.17710    0.06953  -2.547 0.011005 *
absences                  -0.02362    0.01314  -1.797 0.072648 .
Mjobhealth                 1.31024    0.36024   3.637 0.000290 ***
Mjobother                 -0.22262    0.61112  -0.364 0.715723
Mjobservices               0.11129    0.82239   0.135 0.892380
Mjobteacher               -3.30239    1.89341  -1.744 0.081435 .
higheryes                  1.12047    0.49210   2.277 0.023000 *
studytime:schoolsupyes    -0.75782    0.30909  -2.452 0.014381 *
Mjobhealth:higheryes          NA         NA      NA       NA
Mjobother:higheryes        0.42416    0.65550   0.647 0.517727
Mjobservices:higheryes     0.76856    0.86100   0.893 0.372265
Mjobteacher:higheryes      3.93215    1.90376   2.065 0.039132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.541 on 1018 degrees of freedom
Multiple R-squared:  0.2883,    Adjusted R-squared:  0.2722
```

Q3) 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.

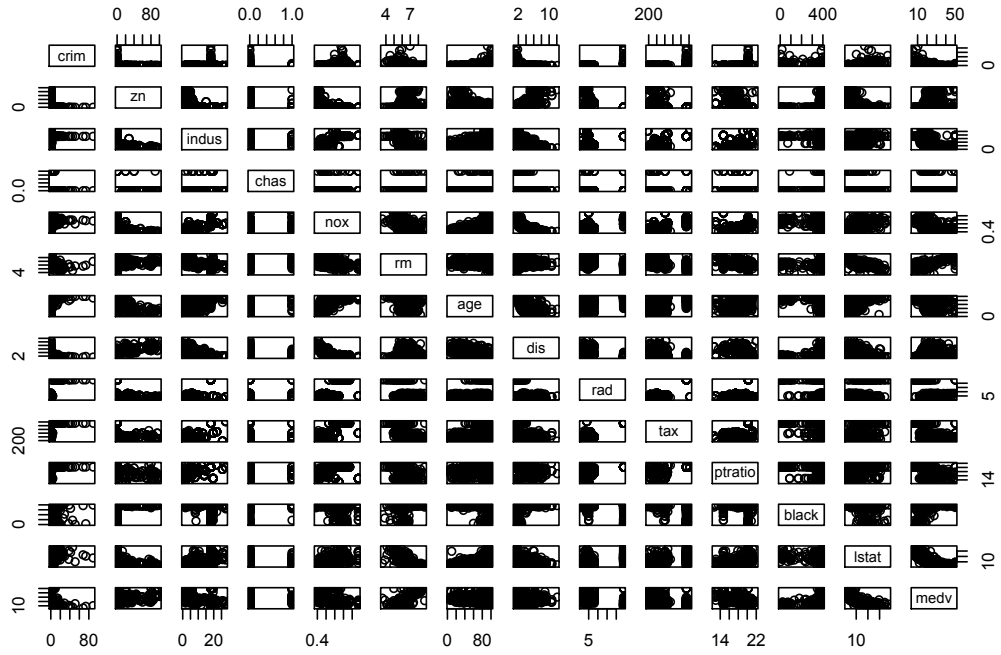
b) Are any of the predictors associated with per capita crime rate?

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.

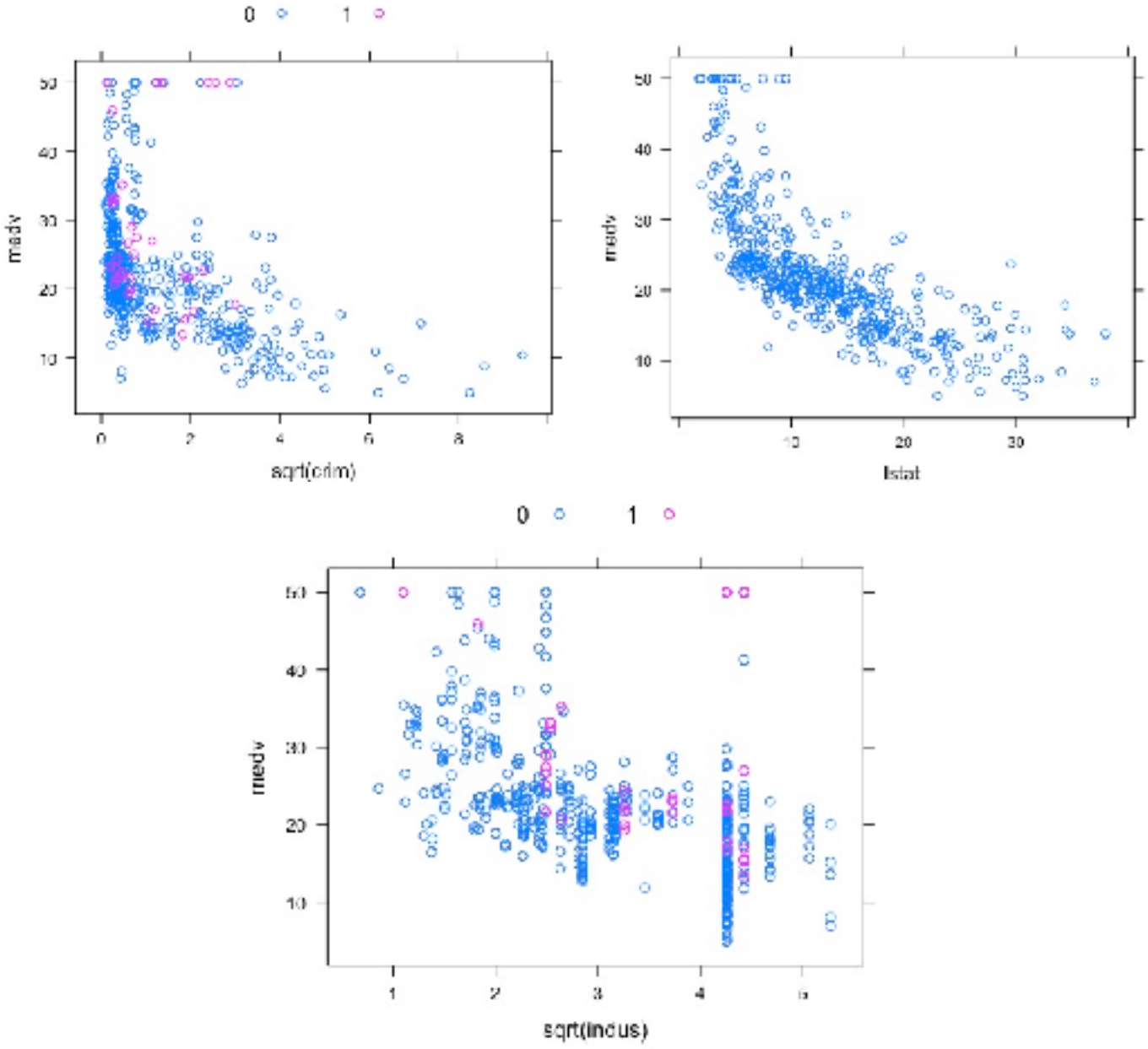
d) In this data set, how many of the suburbs average more than seen rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Solution 3:

Plotted a correlation plot for the Boston dataset as the initial step of EDA.

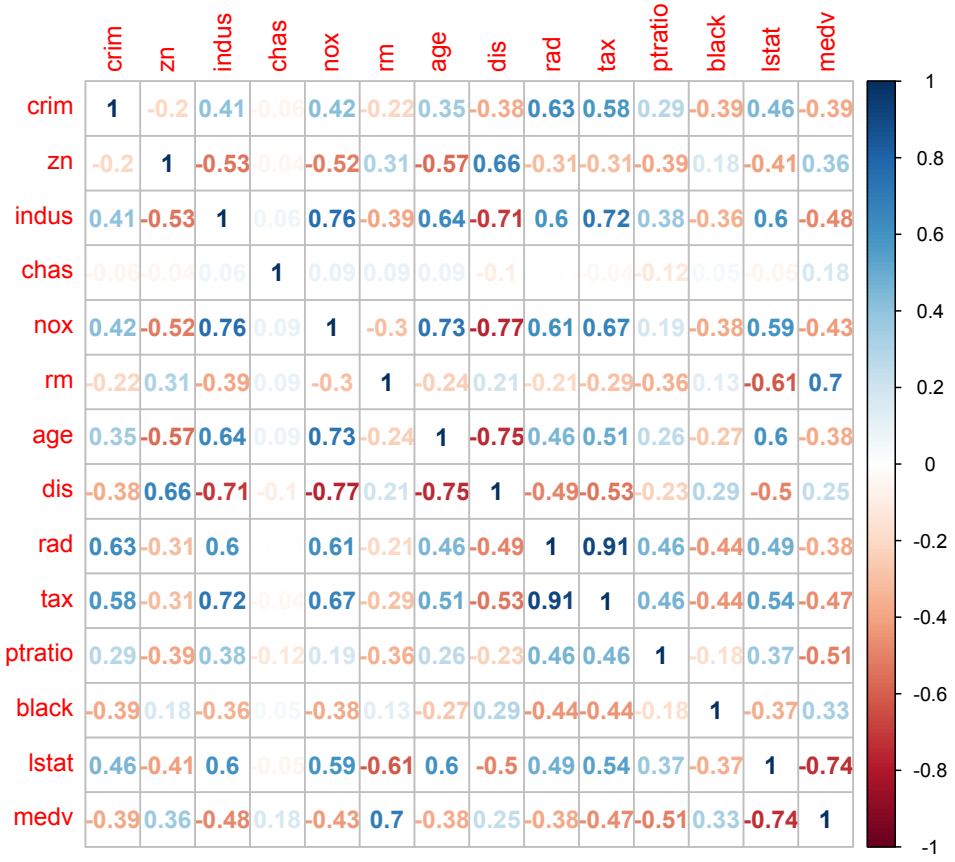


A)From the above correlation matrix plot, crim, lstat and indus show a strong relation with the target, medv. Performed bivariate analysis for these predictors to understand the underlying pattern.



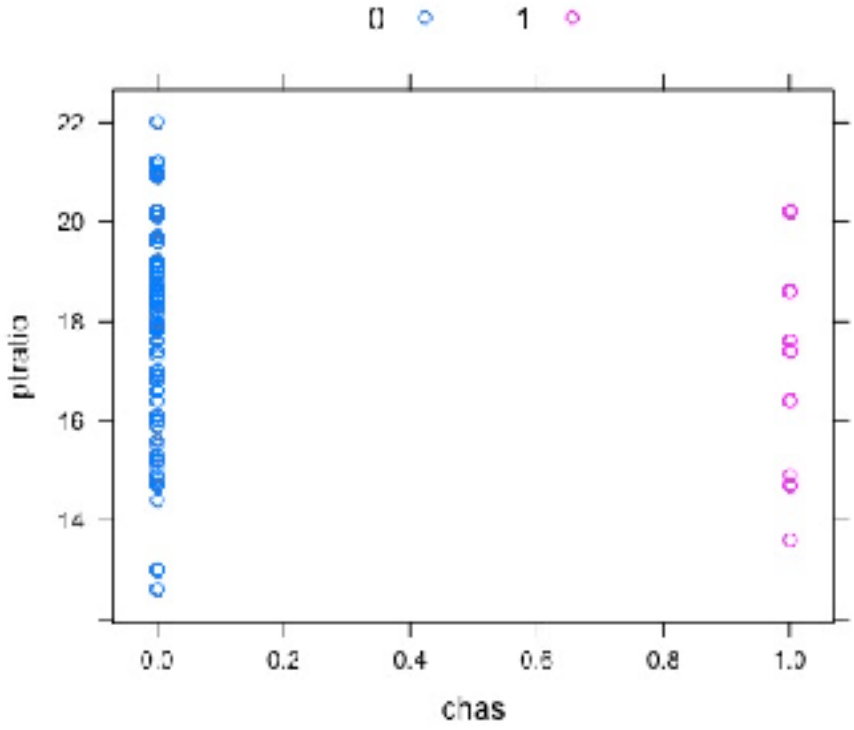
From these graphs, excluding a few outliers, we can see a clear pattern of decline in the median value of homes with an increase in the independent predictors crim, lstat and indus.

B) To observe whether a predictor is associated with the per capita crime rate (crim), we plot a heat map.

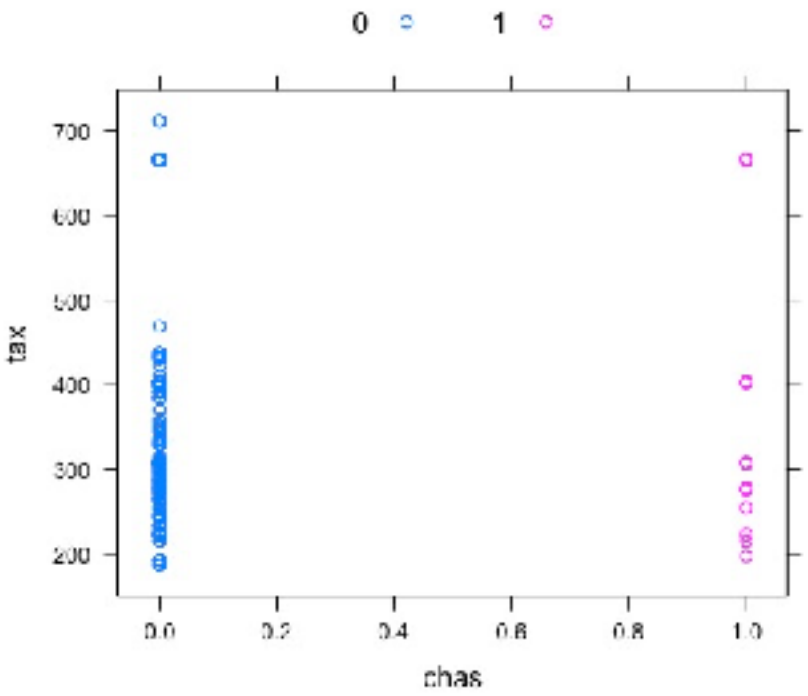


The first column of the heat map is observed to understand the relationship between crime and other predictors. Here we see that crim is more correlated to rad (index of accessibility to radial highways) with 63% and tax (full-value property-tax rate per \ $10,000) with 58%, compared to other values.

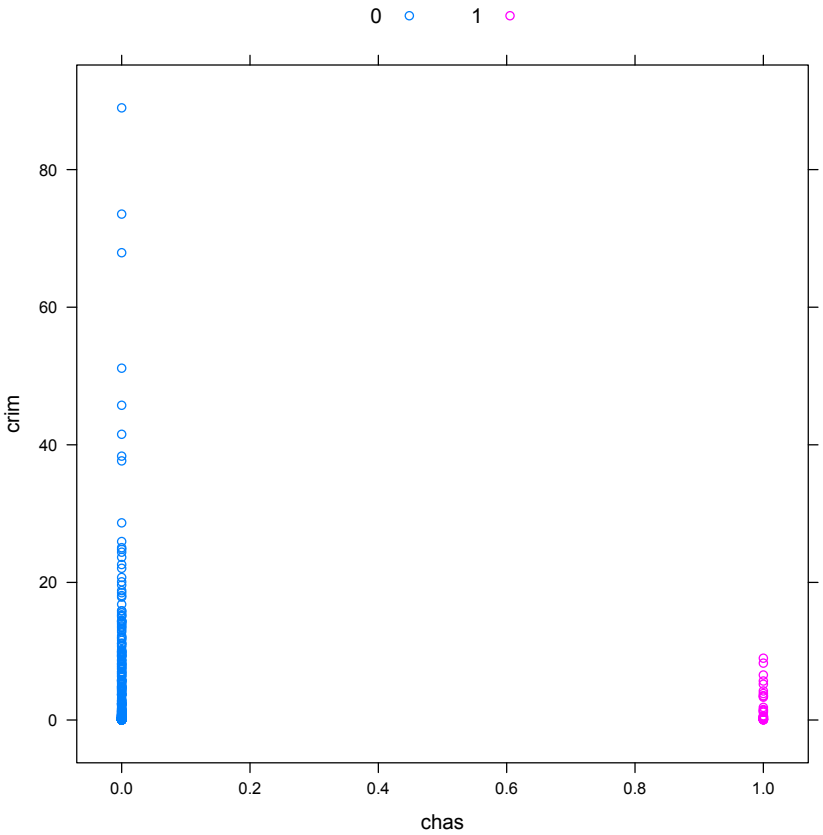
C) To check whether a suburb had high pupil-teacher ratio (ptratio), tax rates (tax) and crime rates (crim), split the data based on the variable chas. Suburbs that are bound by the river are marked with pink and those that are not bound by the river are blue.



The graph above shows that suburbs that are blue have higher pupil-teacher ration (max. 22, median 19) as compared to the others (max. 20, median 17).



From the above graph it is seen that the blue suburbs are the ones with higher tax rates as compared to the pink group, with a maximum of 711/$10,000 as against 666/$10,000.



The graph above shows that the suburbs marked with blue have a much higher per capita crime rate (max. 88, median 24) while those of the pink group have a much less-er per capita crime rate (max. 8, median 0.4).

D) There are 64 entries that average more than 7 rooms per dwelling while 13 entries that average 8 rooms per dwelling.

```
> summary(subset_rm8)
      crim                zn             indus            chas              nox              rm             age       
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000   Min.   :0.4161   Min.   :8.034   Min.   : 8.40  
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000   1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40  
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000   Median :0.5070   Median :8.297   Median :78.30  
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538   Mean   :0.5392   Mean   :8.349   Mean   :71.54  
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000   3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:85.50  
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000   Max.   :0.7180   Max.   :8.780   Max.   :93.90  
      dis              rad              tax           ptratio          black            lstat            medv      
 Min.   :1.801   Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6   Min.   :2.47   Min.   :21.9  
 1st Qu.:2.288   1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5   1st Qu.:3.32   1st Qu.:41.7  
 Median :2.894   Median : 7.000   Median :307.0   Median :17.40   Median :386.9   Median :4.14   Median :48.3  
 Mean   :3.430   Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2   Mean   :4.31   Mean   :44.2  
 3rd Qu.:3.652   3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7   3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :8.907   Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9   Max.   :7.44   Max.   :50.0  
>
```

We use the above summary to infer about the dwellings that average 8 rooms that it is bound by the river Charles, has maximum crime rate of 3/capita, the median indus is 6 with a maximum of 19, tax/$10,000 is maximum of 666 with a median on 307. Also, the maximum lstat is 7, while medv is 50.

Q4) 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.

Solution 3:

Train Error:

```
> knnt__1
[1] 0
> knnt__3
[1] 0.005039597
> knnt__5
[1] 0.005759539
> knnt__7
[1] 0.006479482
> knnt__9
[1] 0.009359251
> knnt__11
[1] 0.008639309
> knnt__13
[1] 0.008639309
> knnt__15
[1] 0.009359251
>
```

Test Error:

```
> knn_1
[1] 0.4967603
> knn_3
[1] 0.49964
> knn_5
[1] 0.49964
> knn_7
[1] 0.5025198
> knn_9
[1] 0.5025198
> knn_11
[1] 0.5025198
> knn_13
[1] 0.5017999
> knn_15
[1] 0.5017999
```