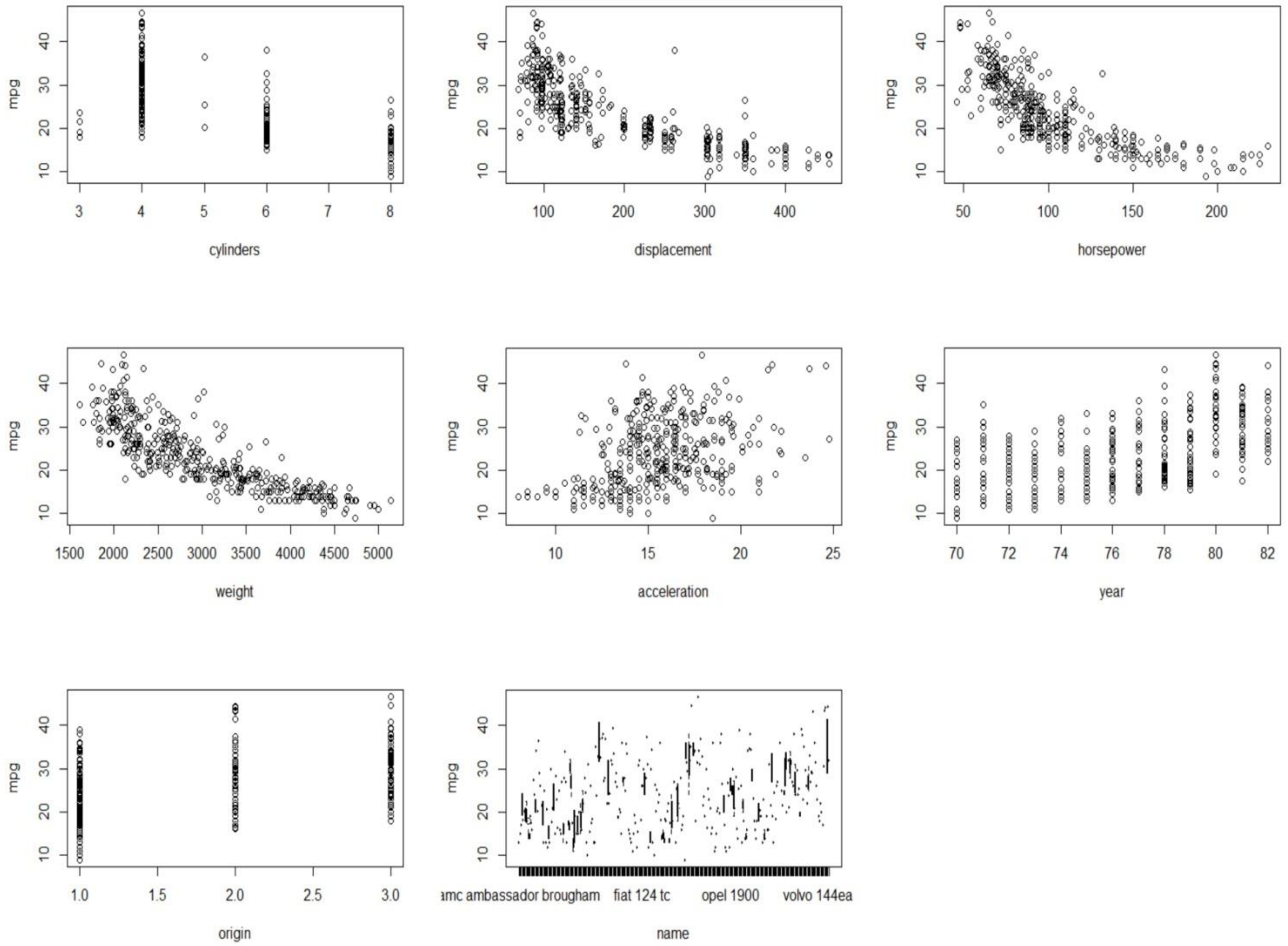
Statistical Data Mining I

Homework 1

Student Number:71

Problem 1:

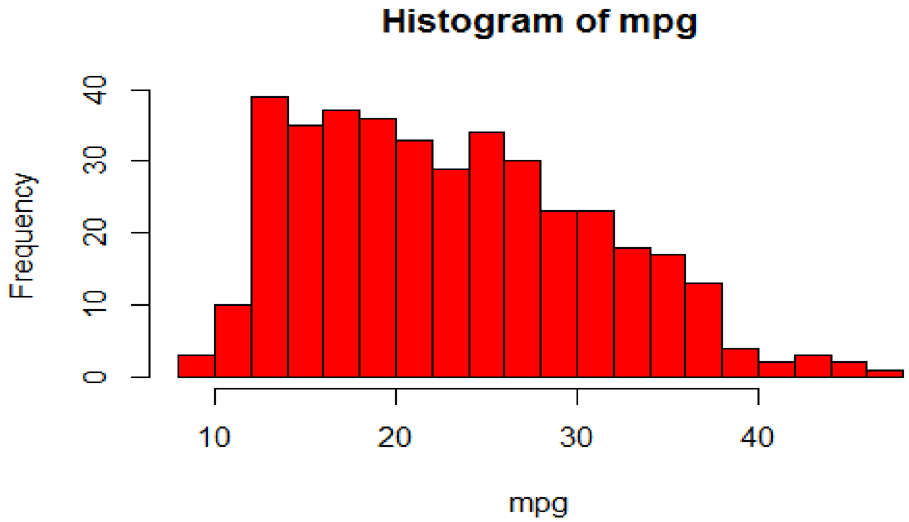


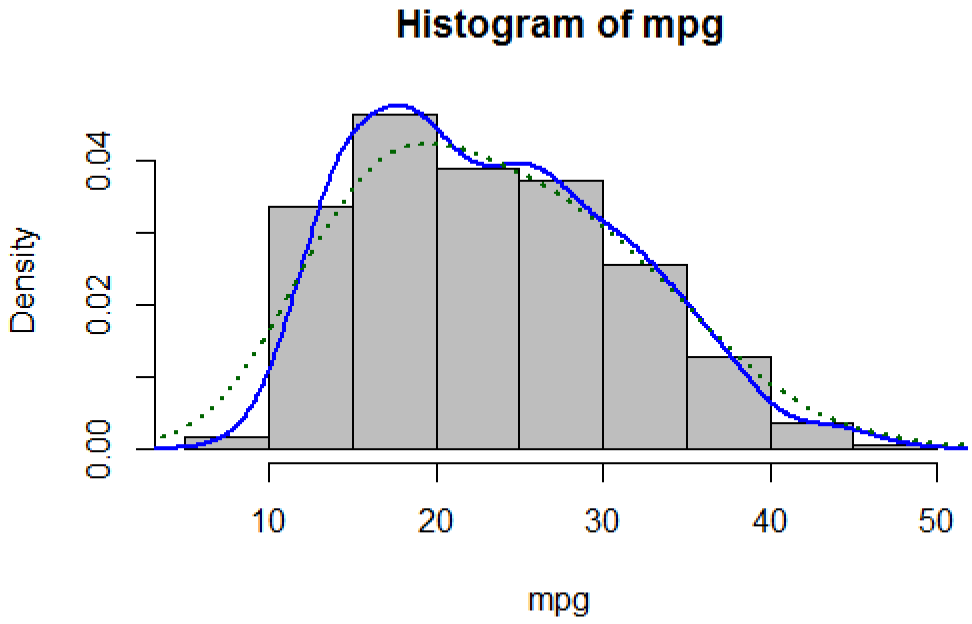
i) We seem to get more mileage per gallon on a 4 cylinder vehicle than the others.

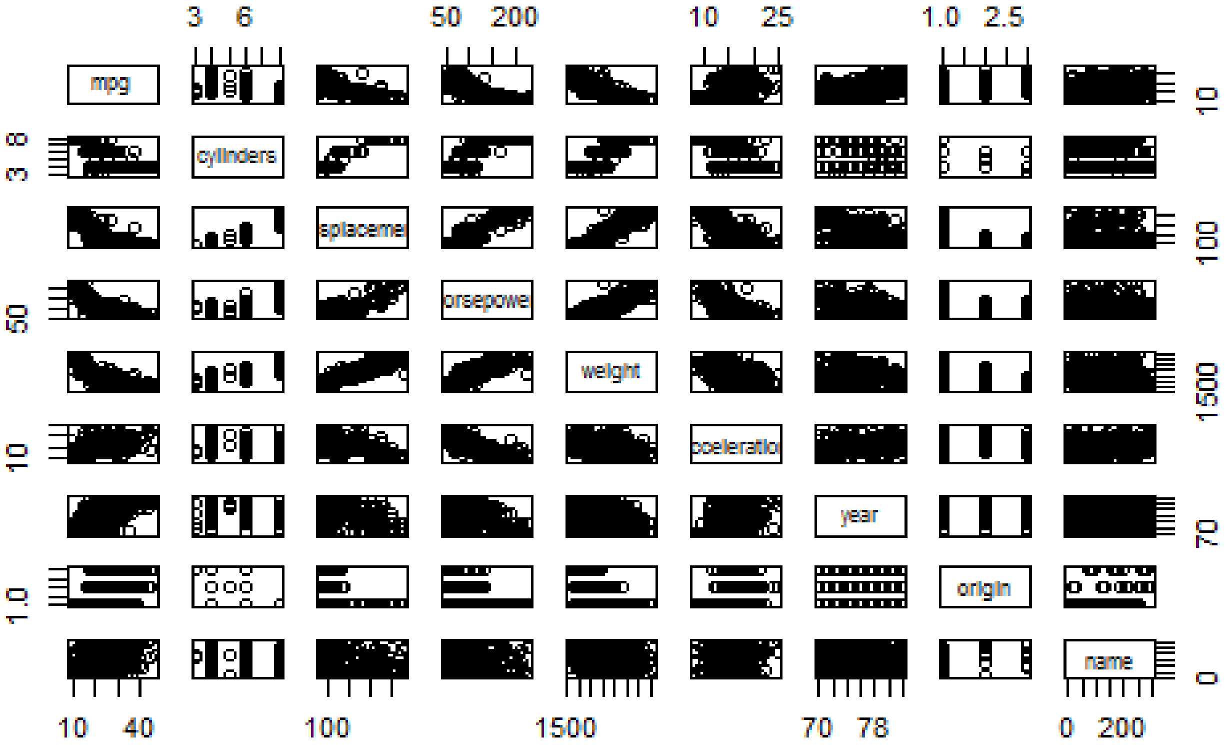
ii) Weight, displacement and horsepower seem to have an inverse effect with mpg. We see an overall increase in mpg over the years. Almost doubled in one decade.

iii) Japanese cars have higher mpg than US or European cars.

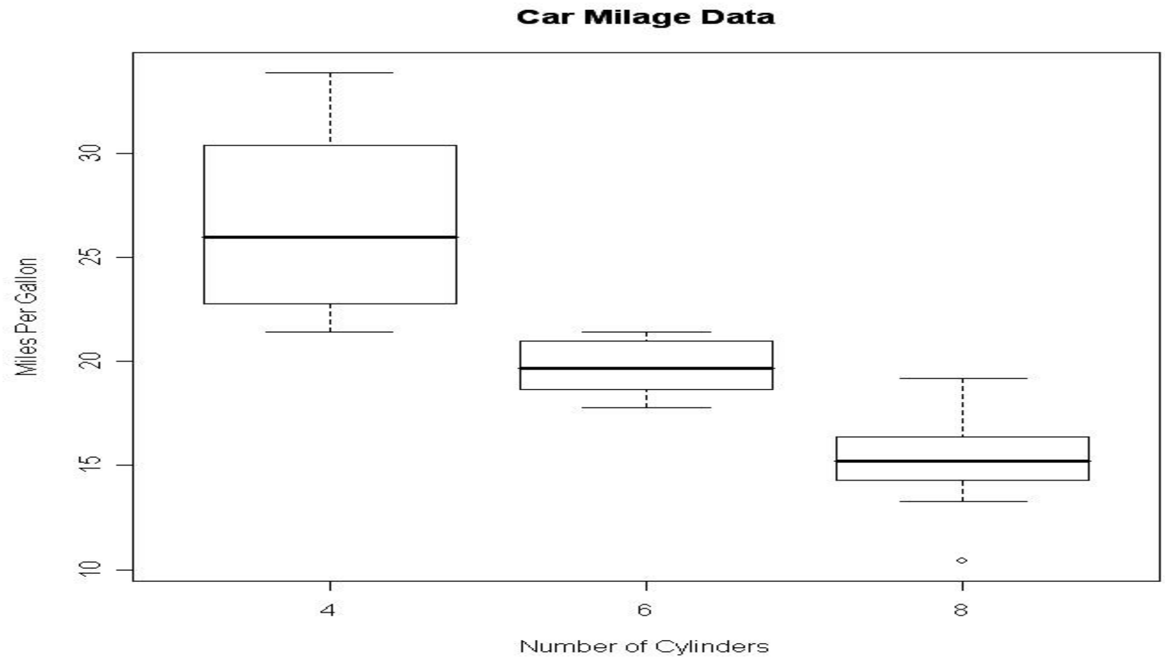
Exploratory Data Analysis:



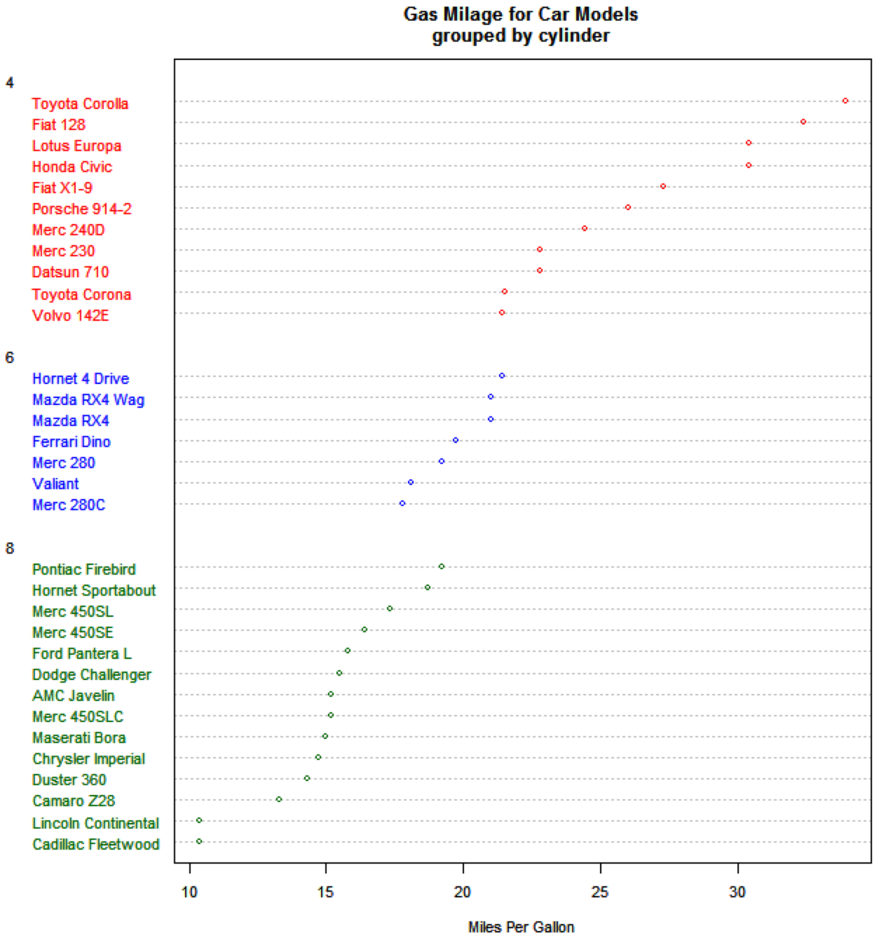



3
6
50
200
10
25
1.0
2.5
mpg
cylinders
splaceme
orsepowe
weight
cceleratio
year
origin
name
3 8
50
10
1.0
10
100
1500
70
0
10
40
100
1500
70
78
0
200

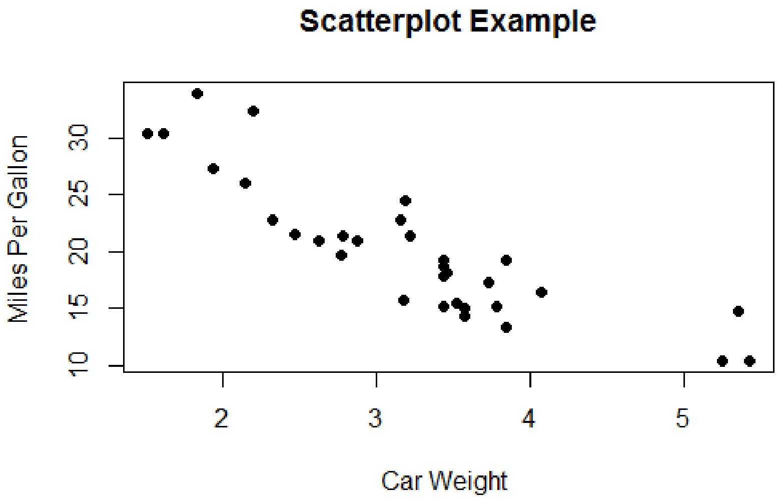
B)Box Plot


Car Milage Data
Miles Per Gallon
10
15
20
25
30
4
6
8
Number of Cylinders

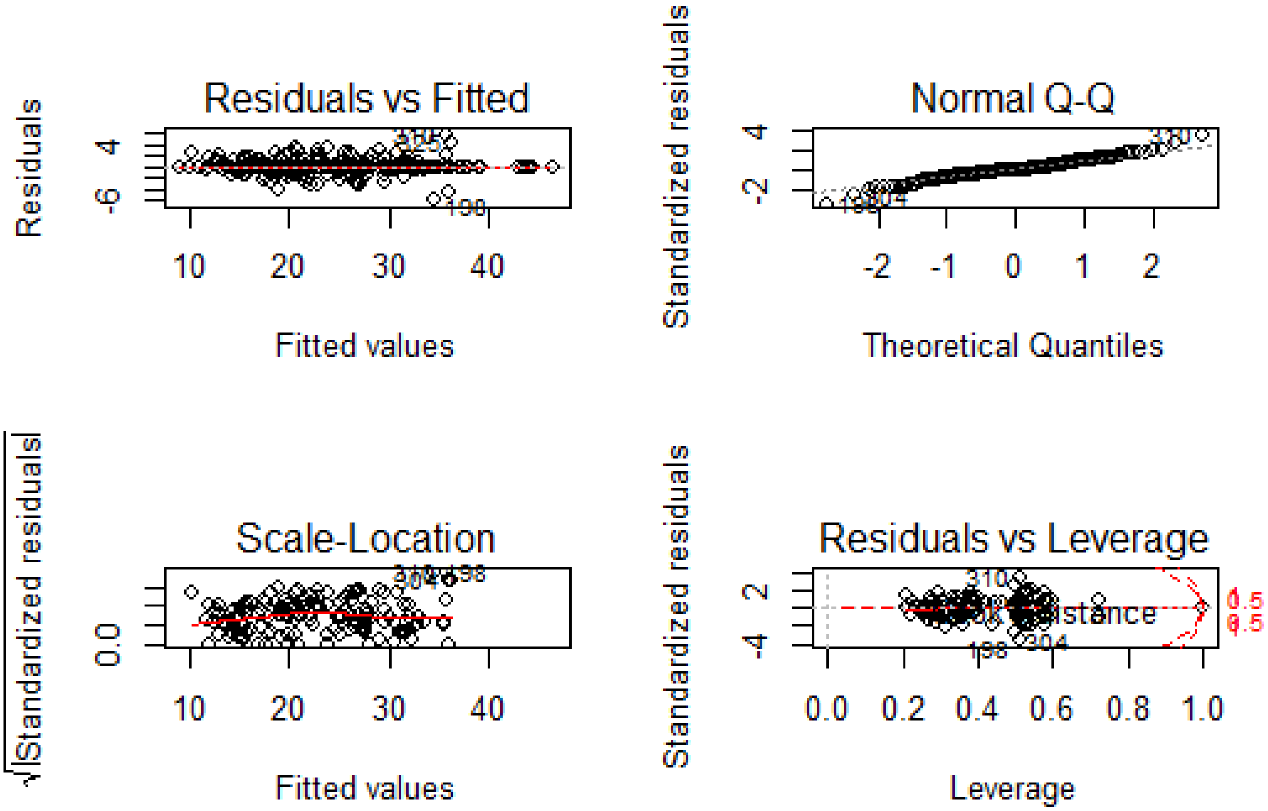
C)Dot Plot



4) Scatter Plot:



Problem 2:



A) Indeed multiple predictors appear to be statistically significant. Displacement, weight, year, and origin are included there.

B) The year coefficient suggests a positive correlation between increasing year, and increasing mpg, in other words, mpg gets better with newer cars. This makes sense, as advances are made and engines become more efficient.

C) From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not..

Problem 3:

i) Linear regression is does not help us here, partly, because the pixels for different samples do not align properly.

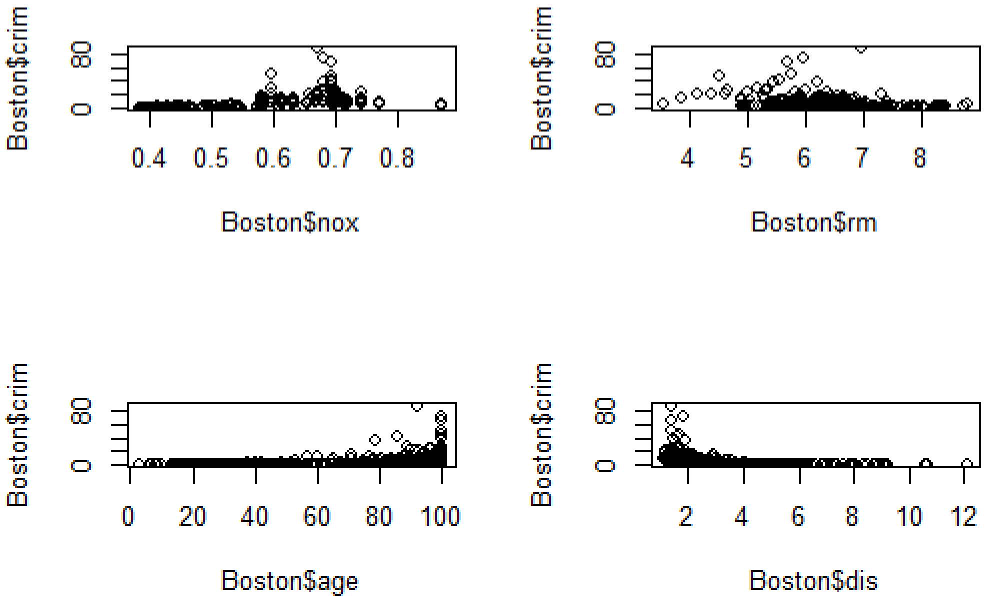
ii) The linear regression does better on the test data than on the training data.

iii) Nearest neighbor results are quite reasonable. The training error results are reduced by the fact that there is one direct hit.

iv) The amount of error increases as the number of neighbors is increased.

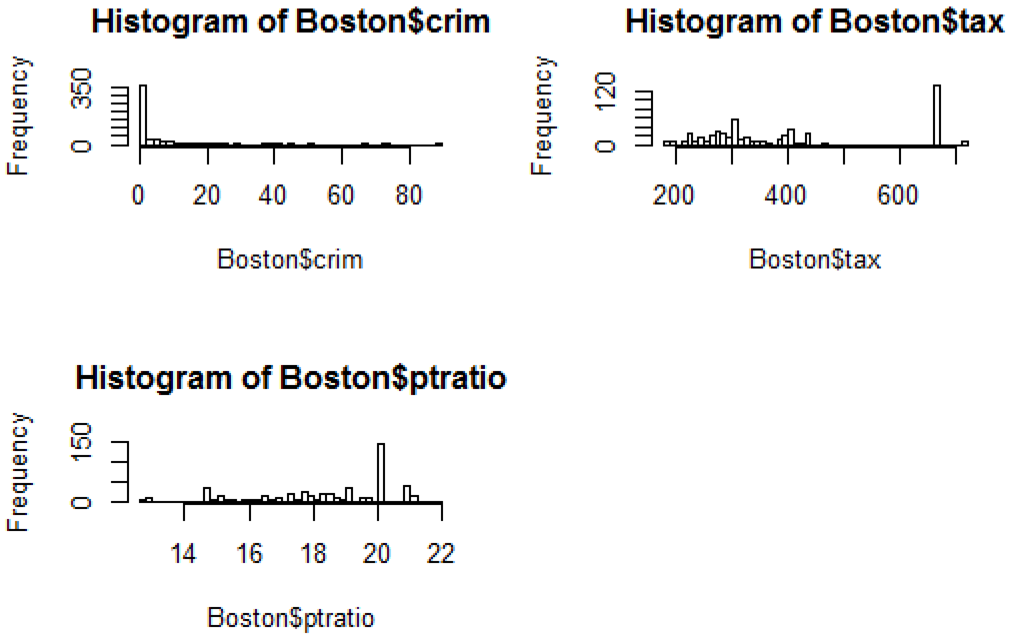
Problem 4:

a) Pairwise Scatterplots



b) There may be a relationship between crim and nox, rm, age, dis, lstat and medv.

c)



d) 64 suburbs average more than seven rooms per dwelling.
13 suburbs average more than eight rooms per dwelling.