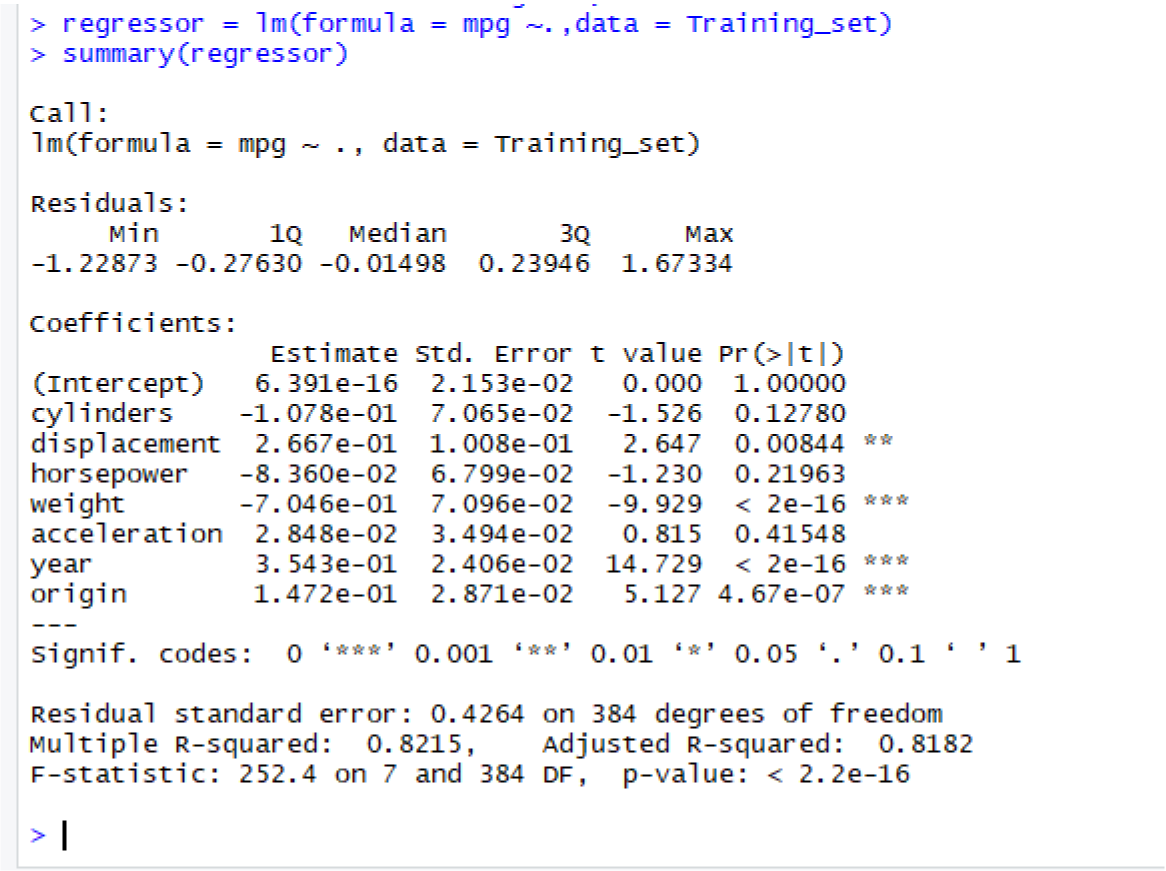
Data Mining I

Homework 1

1. Loading the dataset :code is in the r file
   For cleaning the dataset the 1st step is handling **<u>the missing data</u>…..**since there is no missing data in the dataset no need for doing the step
   For handling the missing data you can either remove the column or handle it by averaging or taking the mean of the column
2. **<u>Finding out the patterns in the data</u>** this can be done by plotting the different kind of graphs and by getting insights out of it as explained in the code
3. **Feature selection** and extraction: in this step the features that are useful are selected based on the insight we got in the above step .
4. **Splitting into training and test sets**: nearly 80% of the data is used for training the model and 20% is used to test the model after predicting and to check the accuracy
5. Feature scaling: this is the most important step in the process of preprocessing the data
   The data is scaled so that one of the feature do not dominate the other in terms of size. Generally scaled around 0 with standard deviation of 1

2.Question Answer

The lower the P values is the More significant impact the dependent variable makes on the independent variable



```
> regressor = lm(formula = mpg ~.,data = Training_set)
> summary(regressor)

Call:
lm(formula = mpg ~ ., data = Training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.22873 -0.27630 -0.01498  0.23946  1.67334 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.391e-16  2.153e-02   0.000  1.00000    
cylinders    -1.078e-01  7.065e-02  -1.526  0.12780    
displacement  2.667e-01  1.008e-01   2.647  0.00844 ** 
horsepower   -8.360e-02  6.799e-02  -1.230  0.21963    
weight       -7.046e-01  7.096e-02  -9.929  < 2e-16 ***
acceleration  2.848e-02  3.494e-02   0.815  0.41548    
year          3.543e-01  2.406e-02  14.729  < 2e-16 ***
origin        1.472e-01  2.871e-02   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4264 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16

> |
```

Summary of The model

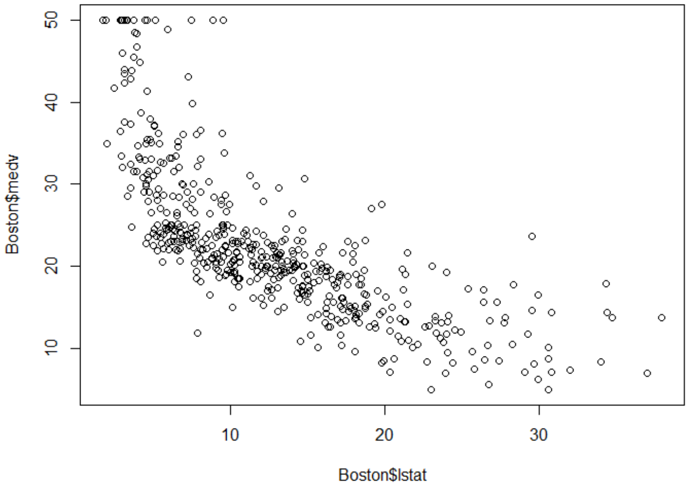
- Generally, it is considered that if the P-value is less than 5% then the variables are considered as More significant.

a) the variables Year, origin, horsepower, are appeared to be most significant and displacement is also significant when compared to the other variables.

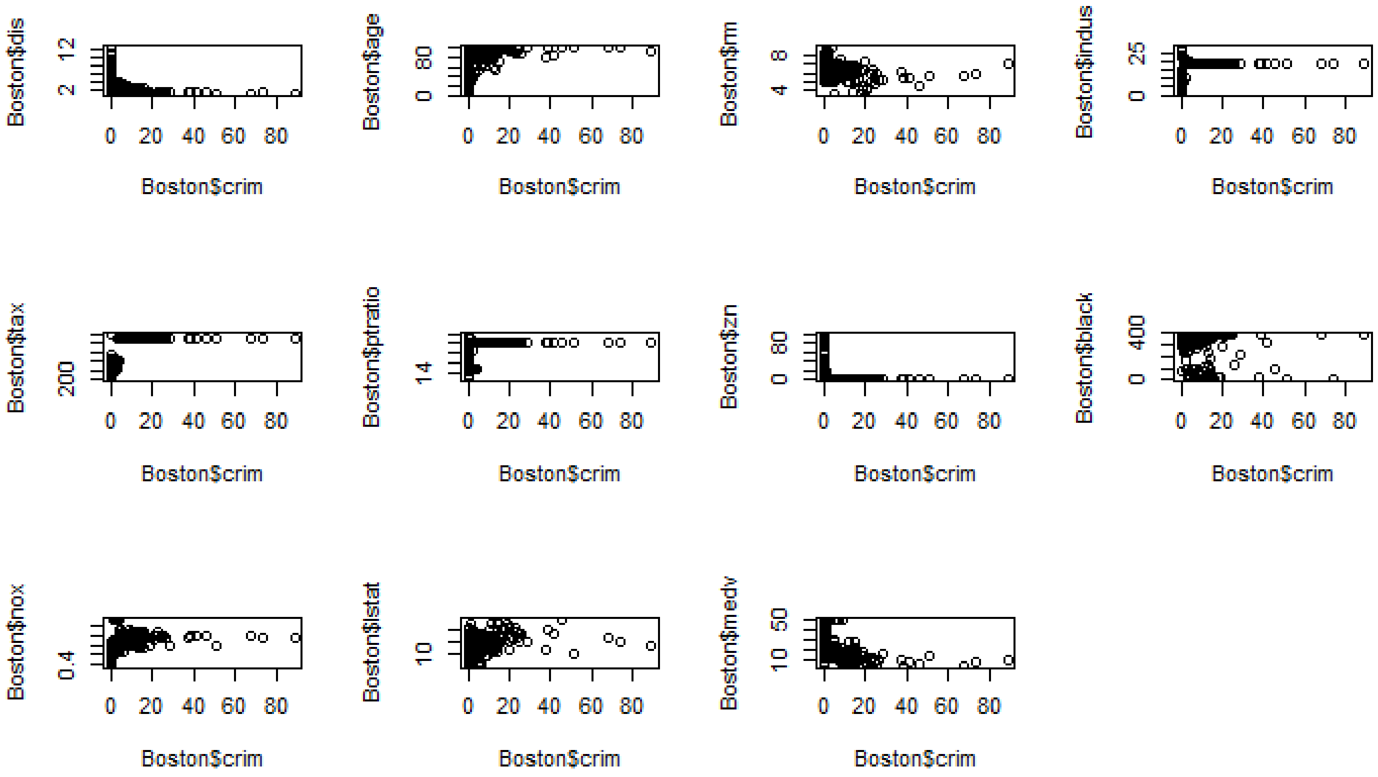
b) the coefficient variable suggests that it is one of the most significant variables in predicting the MPG variable, and if year increases then mpg also tend to increase.

c) *** = most significant 0-0.001 ** = Significant 0.001 – 0.01
So, we can say that the variables origin, year, weight are the most significant variables.

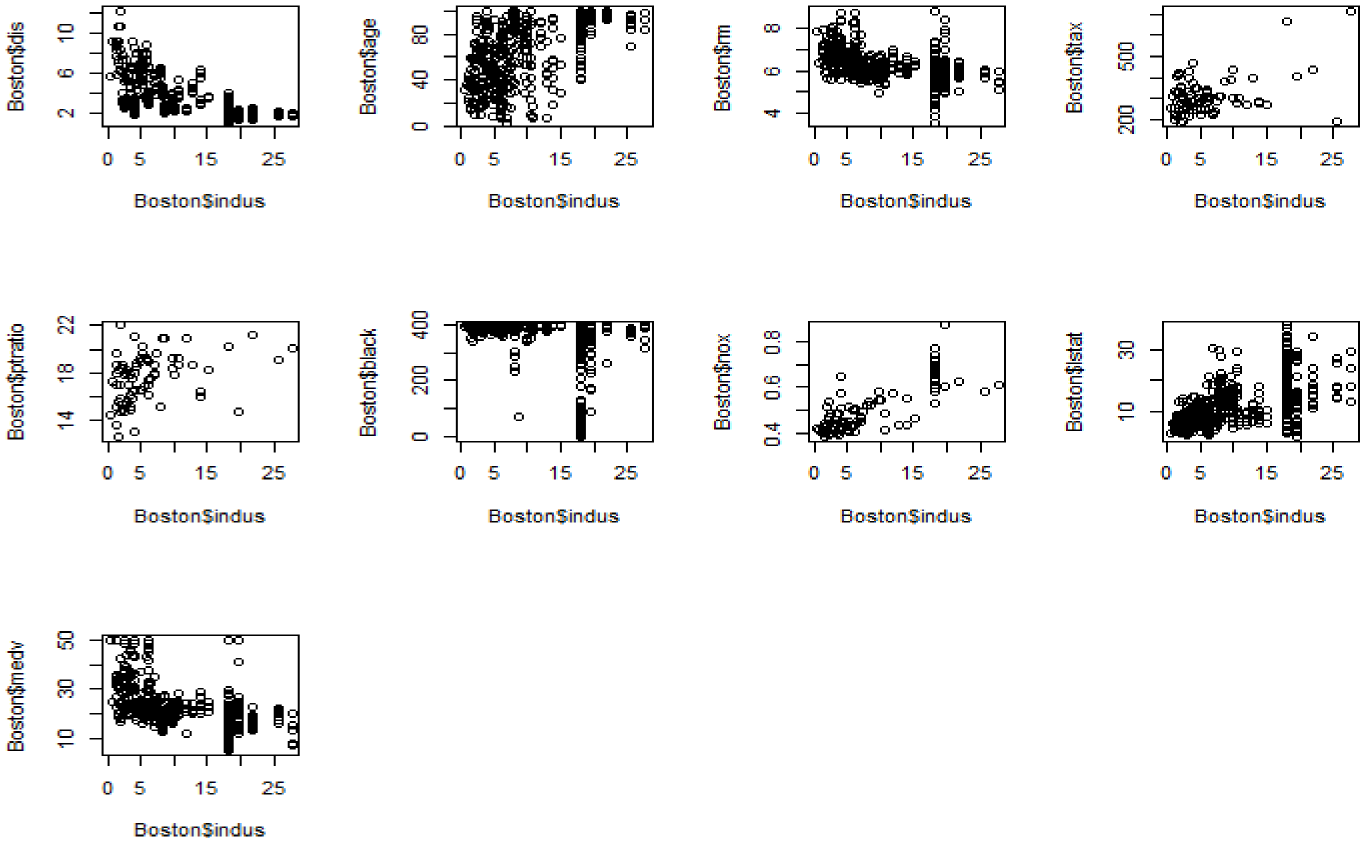
3rd Question Answer



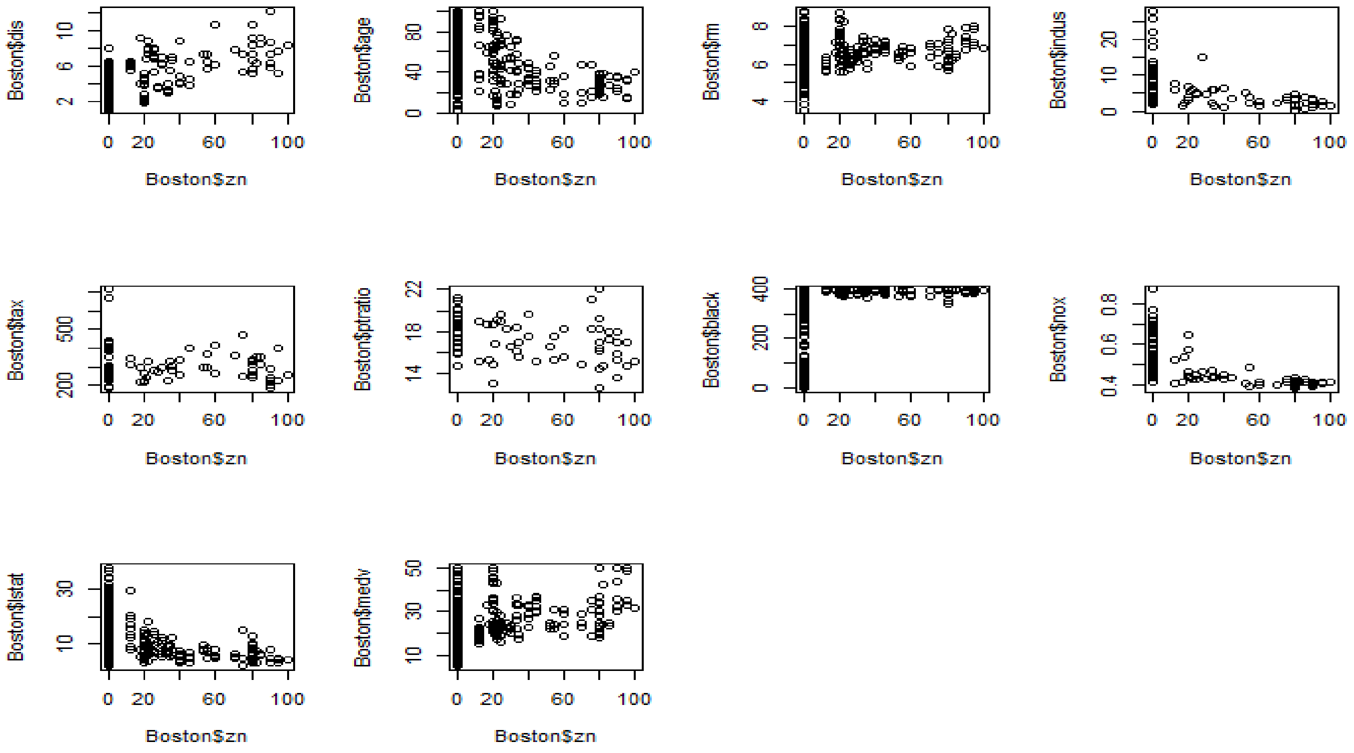
a)crim vs all



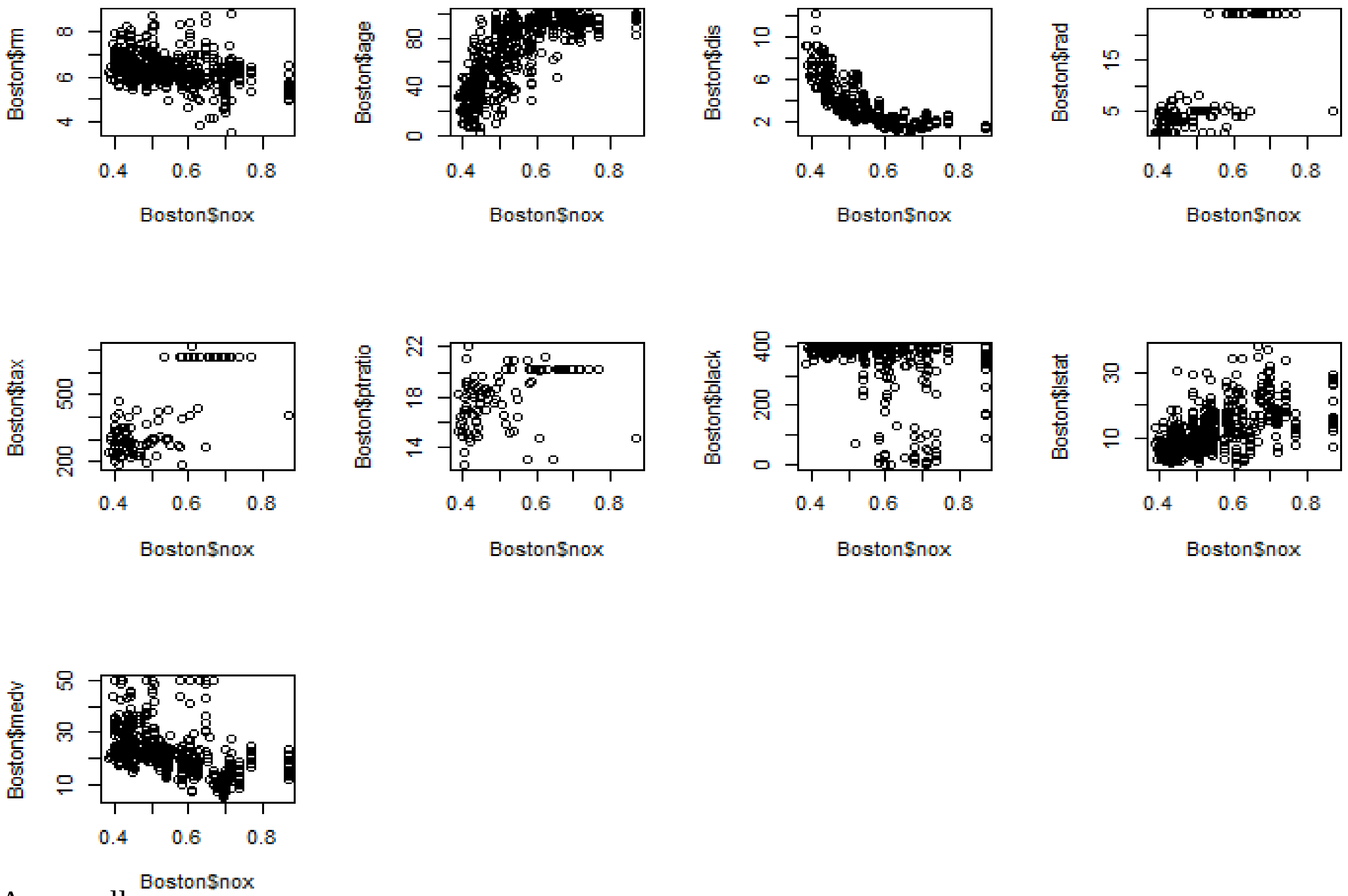
Indus vs all



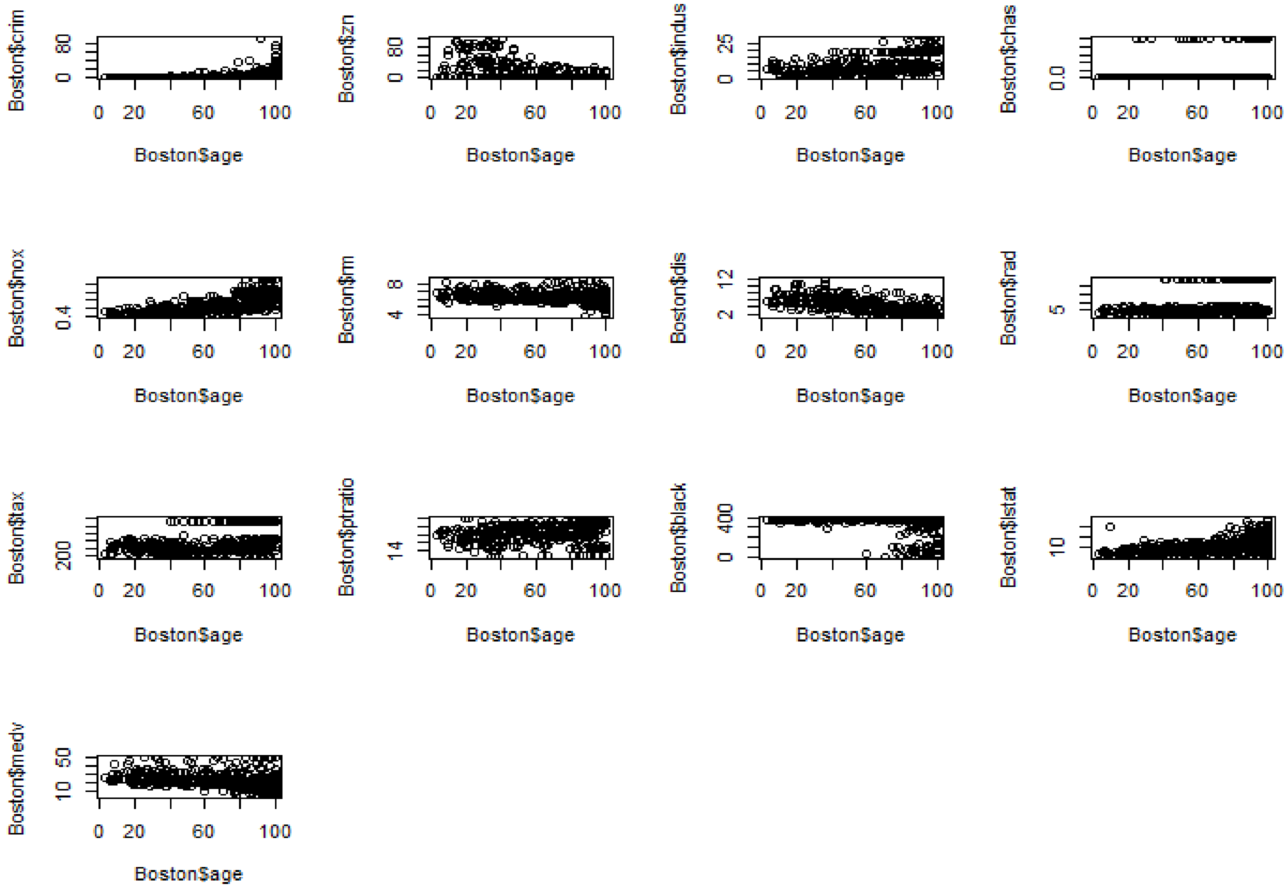
Zn vs all



nox vs all

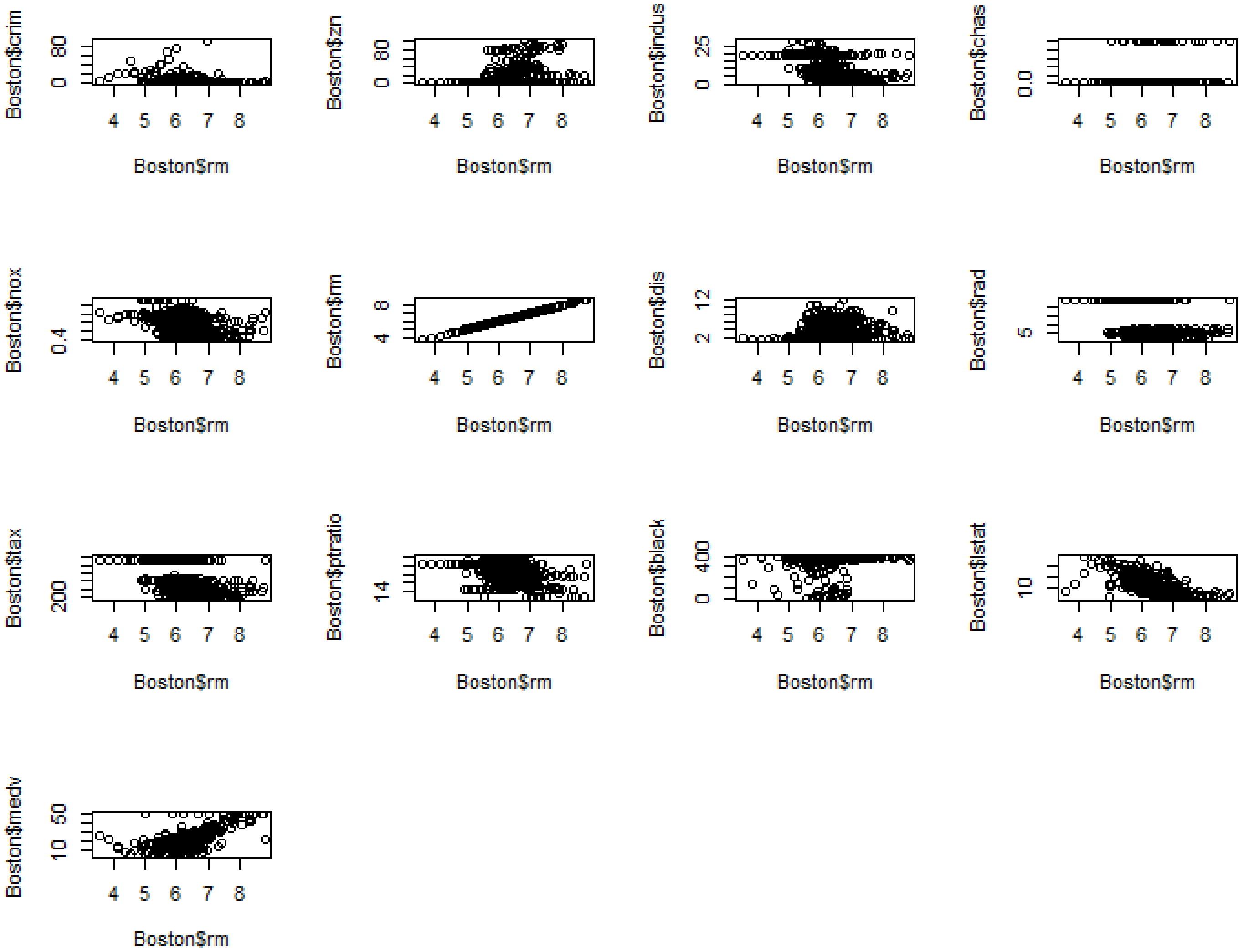


Age vs all



d) the suburbs average more than seven rooms per dwelling are around more than 50%

e) the suburbs average more than seen rooms per dwelling


Boston$crim
Boston$rm
Boston$zn
Boston$rm
Boston$indus
Boston$rm
Boston$chas
Boston$rm
Boston$nox
Boston$rm
Boston$rm
Boston$rm
Boston$dis
Boston$rm
Boston$rad
Boston$rm
Boston$tax
Boston$rm
Boston$ptratio
Boston$rm
Boston$black
Boston$rm
Boston$lstat
Boston$rm
Boston$medv
Boston$rm