9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

Quantitative: "mpg", "displacement", "weight", "acceleration", "horsepower", "cylinders"

Qualitative:, "year", "origin", "name"  
(b) What is the range of each quantitative predictor? You can answer this using the range() function.

mpg 9-46.6

cylinders 3-8

displacement 68-455

horsepower 46-230

weight 1613-5140

acceleration 8-24.8

(c) What is the mean and standard deviation of each quantitative predictor?

mpg mean: 23.44592 standard deviation: 7.805007

cylinders mean: 5.471939 standard deviation: 1.705783

displacement mean: 194.412 standard deviation: 104.644

horsepower mean: 104.4694 standard deviation: 38.49116

weight mean: 2977.584 standard deviation: 849.4026

acceleration mean: 15.54133 standard deviation: 2.758864

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

mpg mean: 27.74457 standard deviation: 11.97794

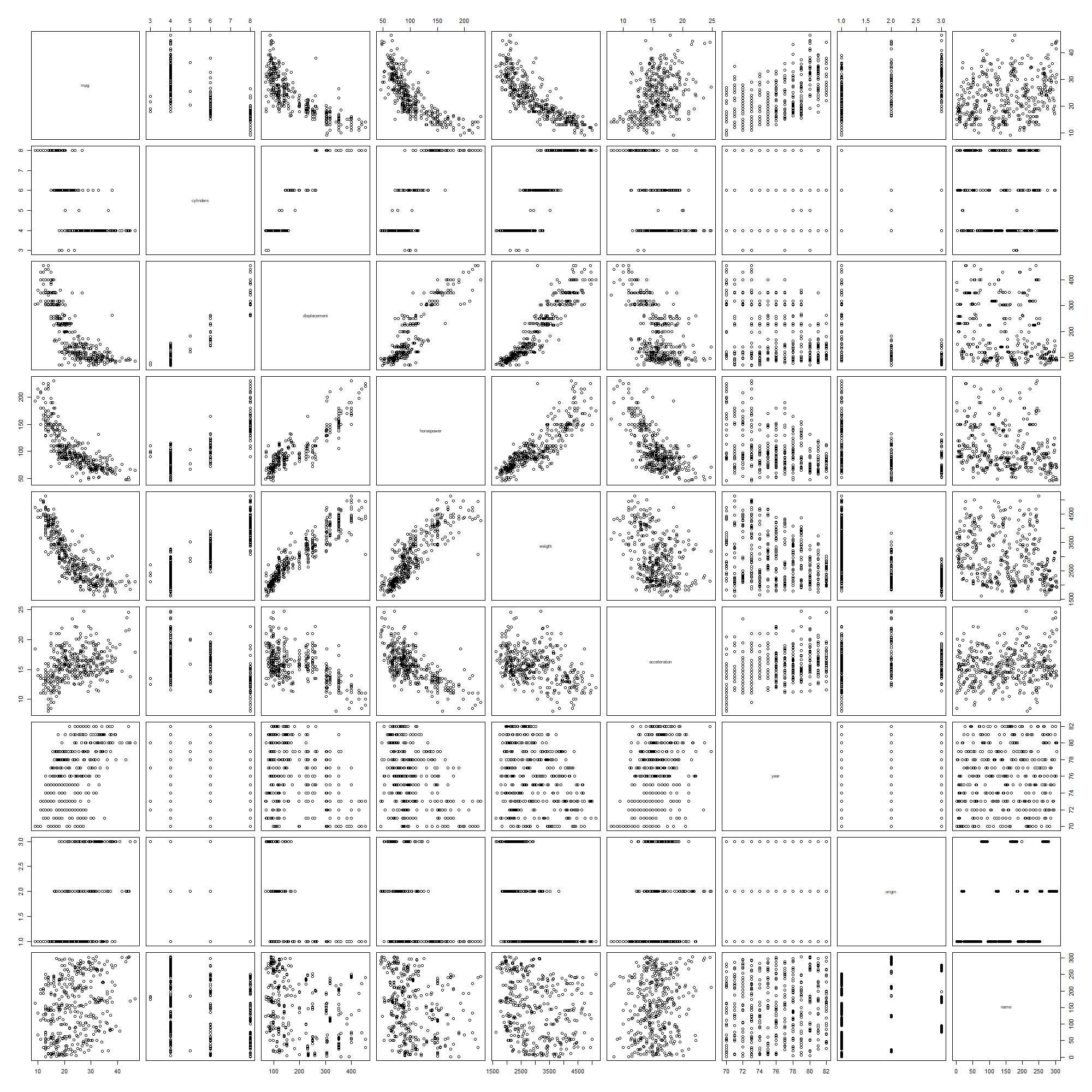
cylinders mean: 3 standard deviation: 0

displacement mean: 264.4615 standard deviation: 148.4484

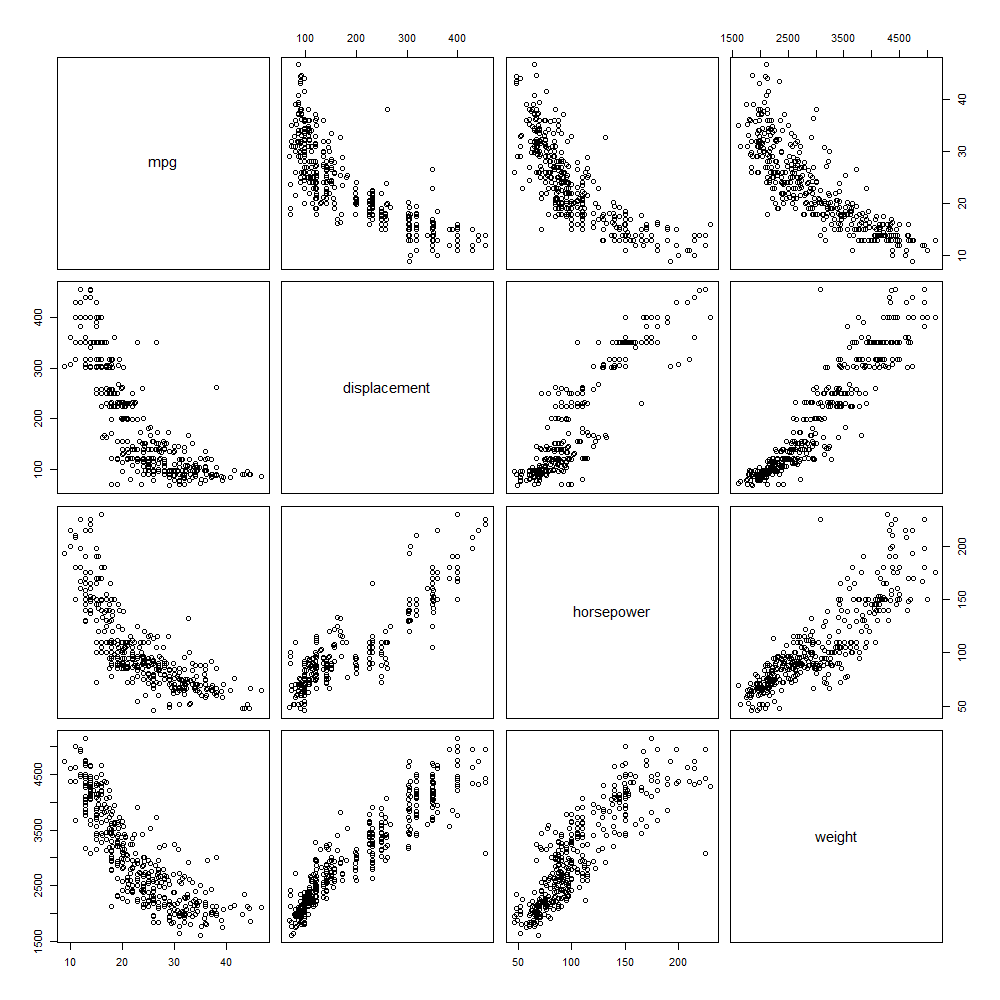
horsepower mean: 126.1205 standard deviation: 64.87632

weight mean: 3415.398 standard deviation: 1269.239

acceleration mean: 16.7186 standard deviation: 4.812594

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.  


There may be a relationship between the following predictors, mpg, displacement, horsepower, weight. The graph below is the scatter plots matrix about those predictors.

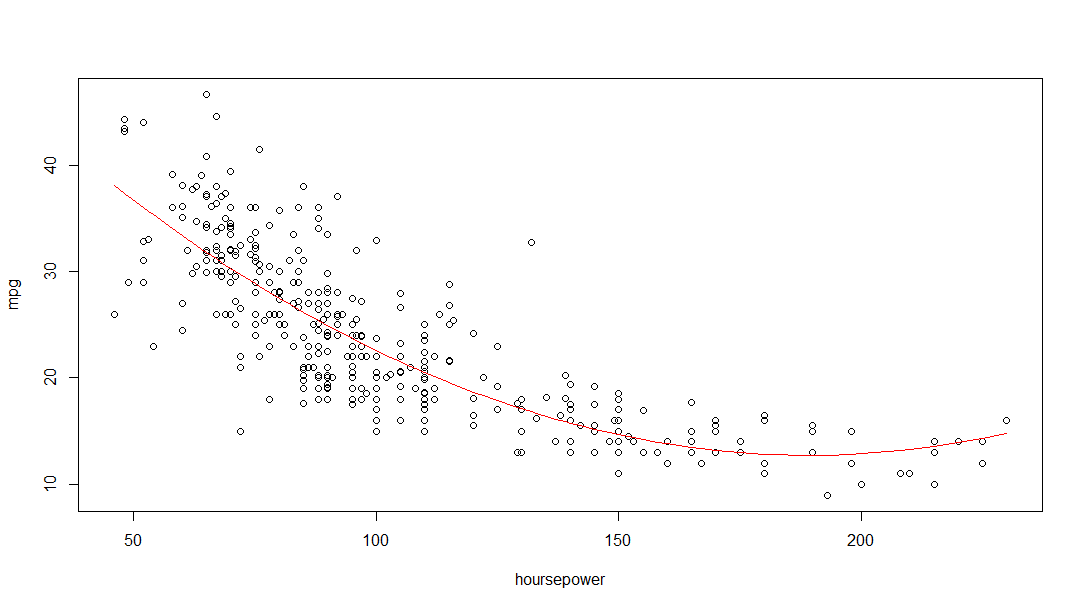


(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer

Using the Polynomial Regression to predict mpg by hoursepower, we can get the graph below.

I also tried the multipule regression by mpg~horsepower+displacement+weight, the function will be

Which says the more horusepower, displacement and weight , the mpg will be lower.

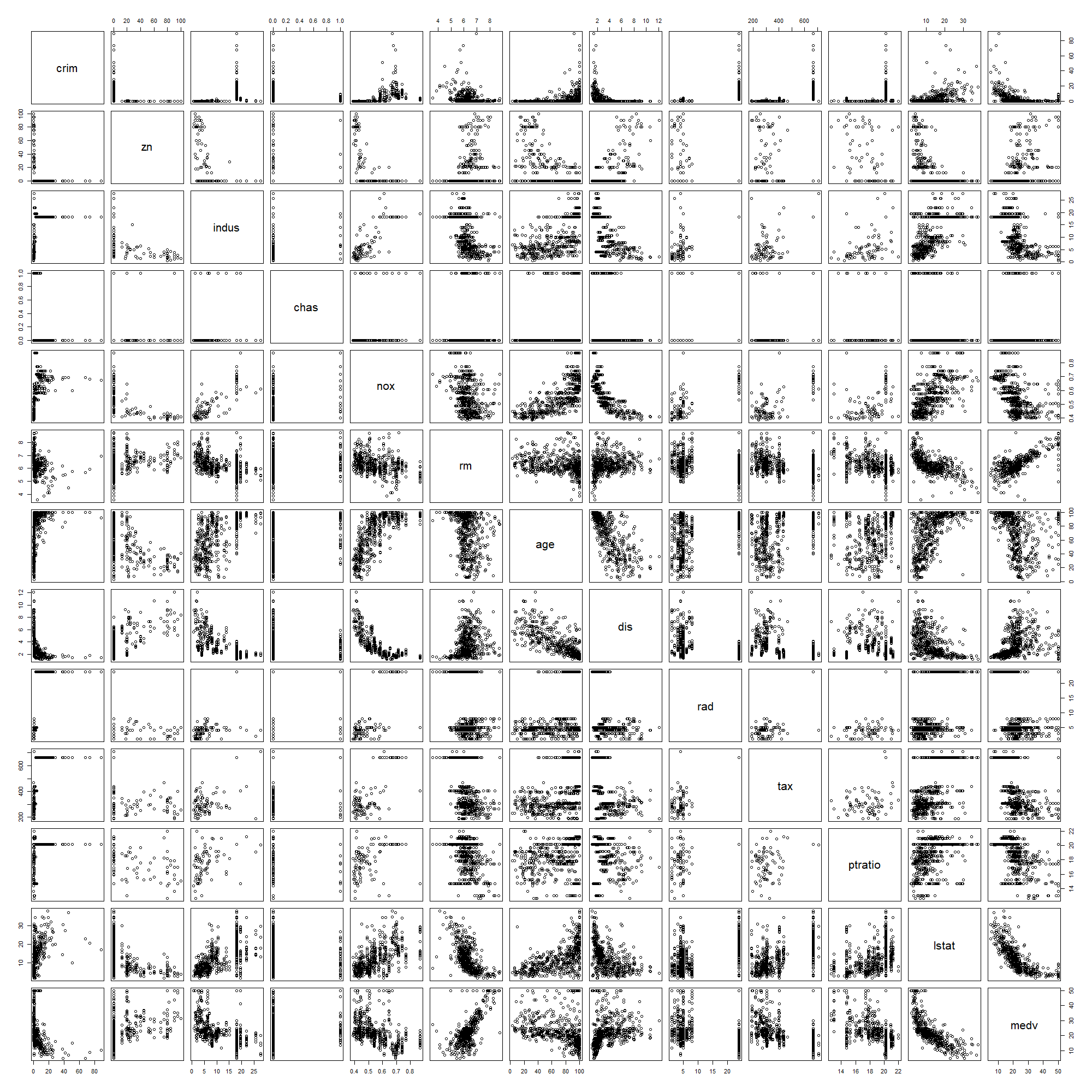


10. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?  
  
There are 506 rows and 13 cols, each column represent one kind of variable and the row represent one sample (census tracts)

The meaning of each col is shown in following table:

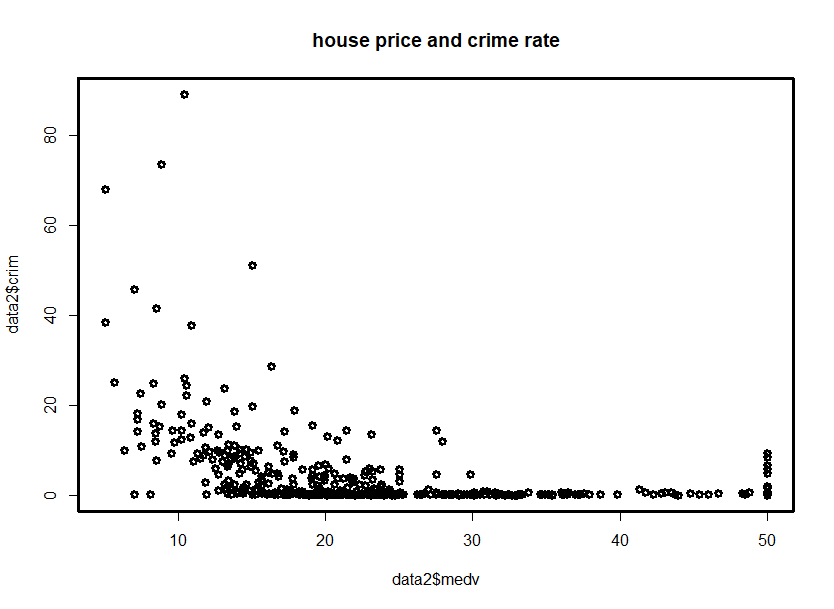
|  |  |  |  |
| --- | --- | --- | --- |
| crim | per capita crime rate by town. | dis | weighted mean of distances to five Boston employment centres. |
| zn | proportion of residential land zoned for lots over 25,000 sq.ft. | rad | index of accessibility to radial highways. |
| indus | proportion of non-retail business acres per town. | tax | full-value property-tax rate per $10,000. |
| Chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). | ptratio | pupil-teacher ratio by town. |
| nox | Nitrogen oxides concentration (parts per 10 million). | lstat | lower status of the population (percent). |
| rm | average number of rooms per dwelling. | medv | median value of owner-occupied homes in $1000s. |
| age | proportion of owner-occupied units built prior to 1940. |  |  |

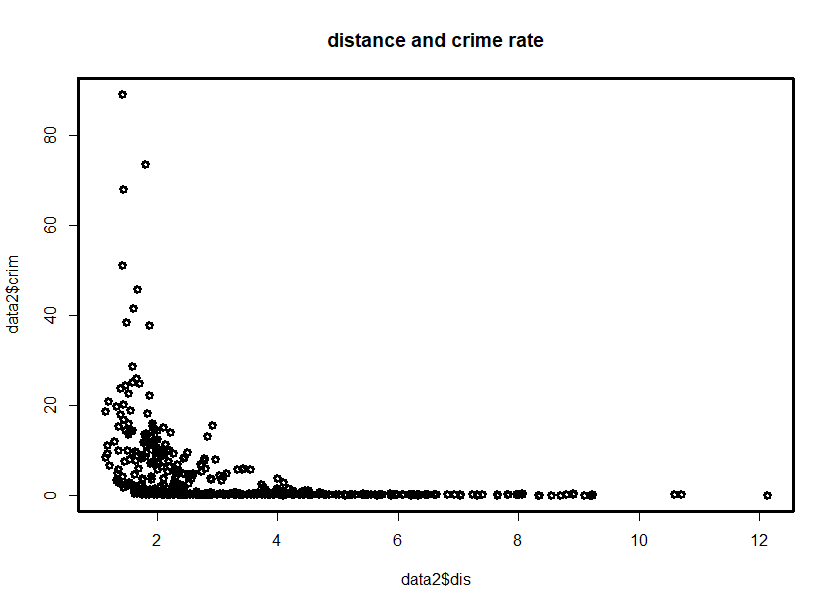
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings. 

There may be positive relationship between “rm” and “medv”, and negative relationship between “nox” and “dis”.

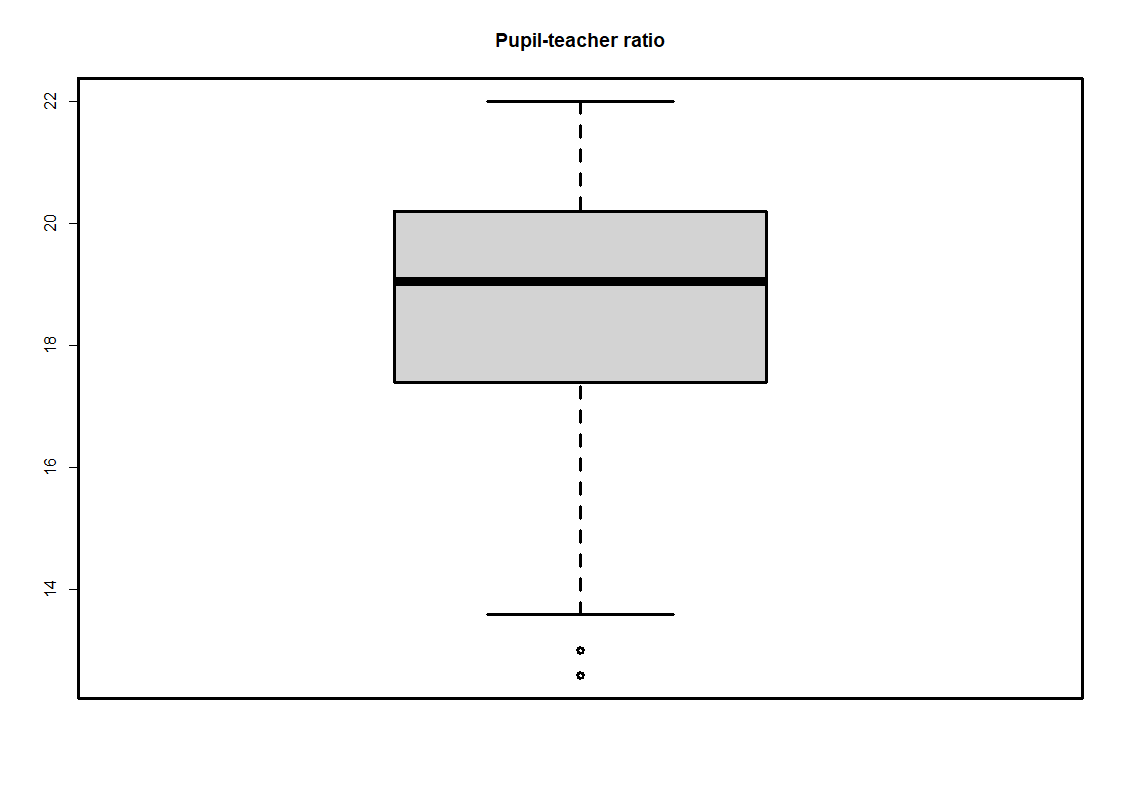
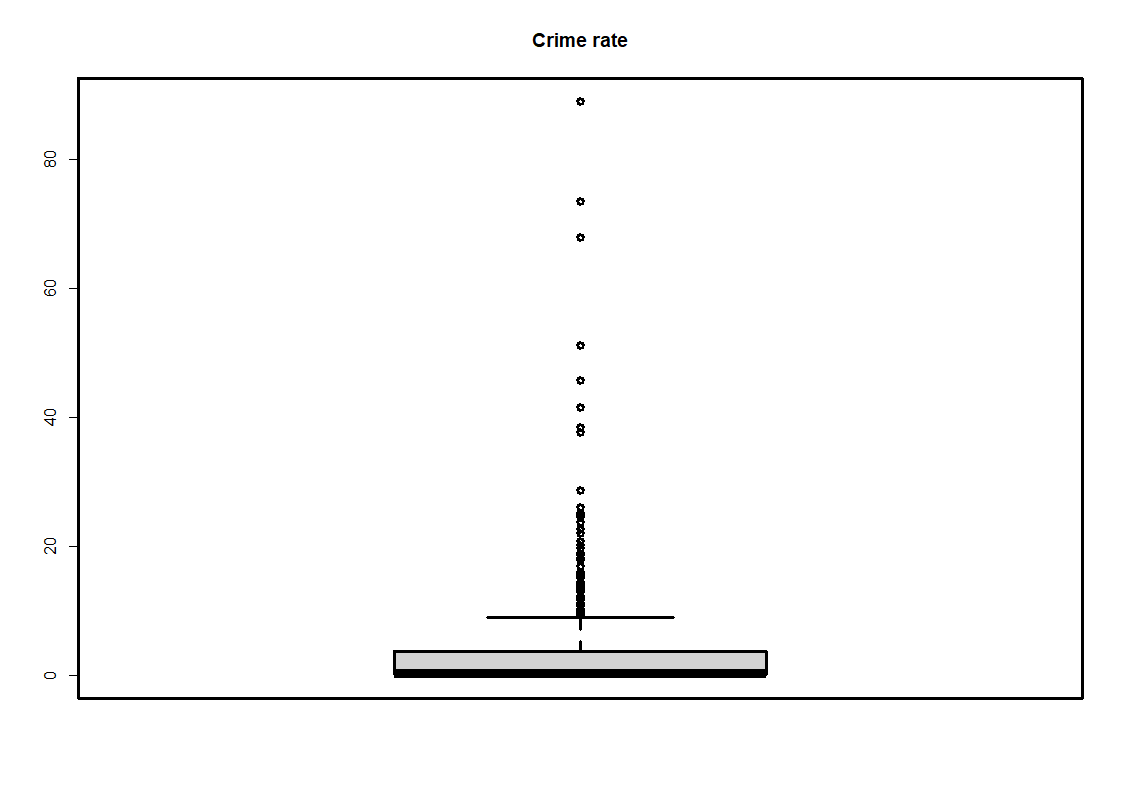
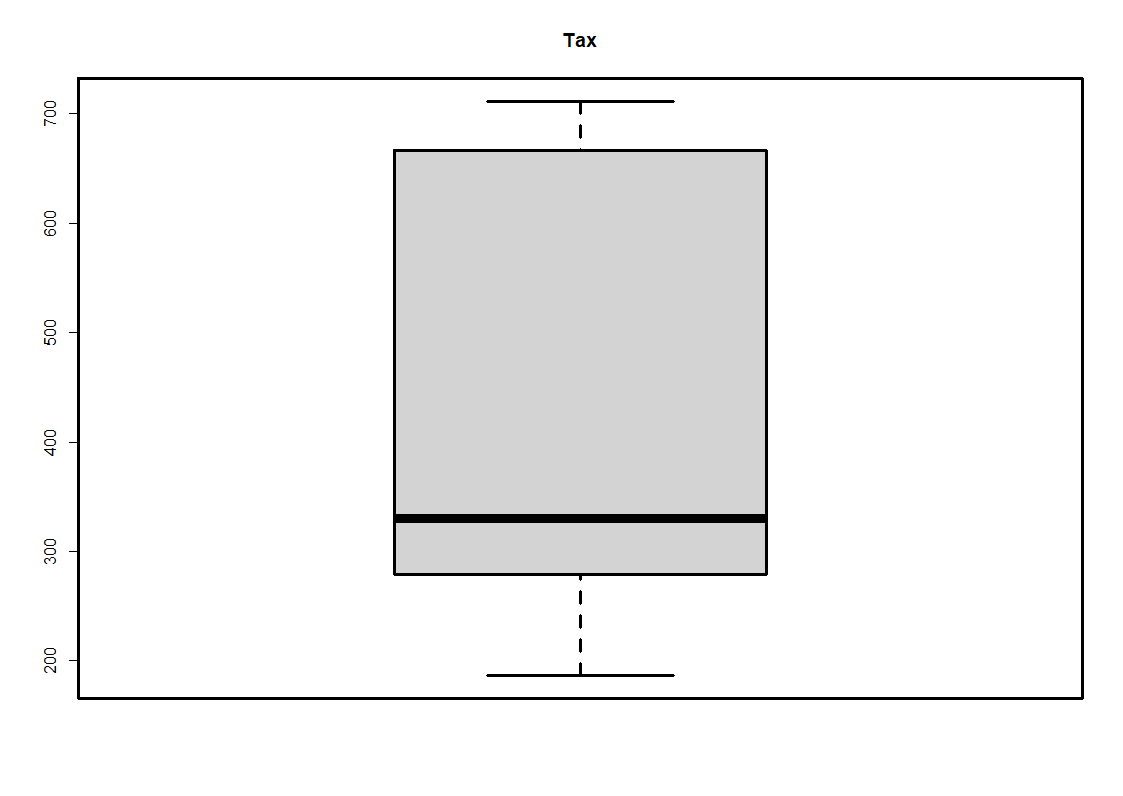
I also find that the census tracts which have high proportion of residential land zoned for lots over 25,000 sq.ft. (“zn”) have the low crime rate than other census tracts.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.  
 As mentioned in (b), the crime rate is low when “zn” value is not equals to zero.

 Other relationship I found was been highlighted by blue square. The distance and the “medv”(house price) have negative relationship with the crime rate, I guess it’s because the countryside are usually have the low crime rate than the cities(“dis” var.), and the place having the high house price is also have the low crime rate, because the people who can afford to live there might have better financial condition.



(d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



Drawing the boxplot of three variable we can see there is many outlier in crime rate, but only few or even no outlier in the other two variable. Although there is many outlier in crime rate variable, but most of census tracts’ crime rate is close to zero(small range).

(e) How many of the census tracts in this data set bound the Charles river?

35 of the census tracts bound the river  
  
(f) What is the median pupil-teacher ratio among the towns in this data set?

The median of pupil-teacher ratio is 19.05

(g) Which census tract of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

The census tracts index 399 and 406 have lowest medv.

一張含有 桌 的圖片

自動產生的描述

From the summary above we can find that these two census tracts have extremely high crime rate and relatively low room number, etc. Lots of variable in no.399 and 406 are the same or close, So it might say the census tract of Boston which has lowest median value will have same property as 399.  
  
(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

There is 64 census tracts have average more than 7 rooms per dwelling, but only 13 census tracts have average more than 8 rooms per dwelling.

By the summary in following page, we can find the census tracts have high average room, will also have high house price(“mdev”) and higher age, but lower in non-retail business acres (“indus” )

一張含有 桌 的圖片

自動產生的描述