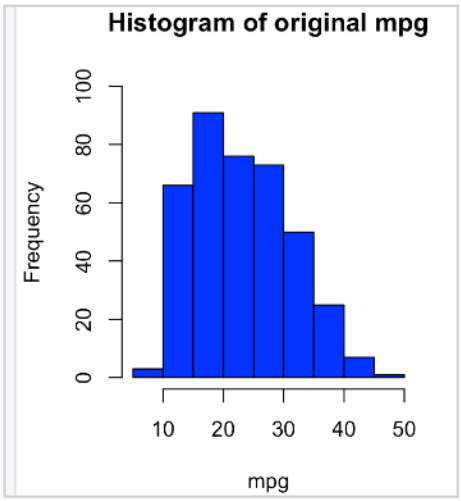
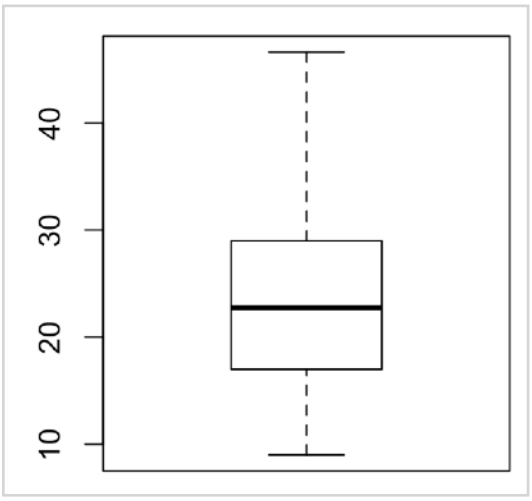
chiehjul
50341634

Data Mining I Homework 1 30 points

Directions: Submit all source codes with write up.

- 1) (10 points) Consider the Auto dataset in the ISLR package. Suppose that you are getting this data in order to build a predictive model for mpg (miles per gallon). Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed. Pre-process this data and justify your choices in your write up. Submit the cleaned dataset as an *.RData file.





Ans: First of all, I checked if there is any missing data in mpg, and there is no NULL data. Then I plot the histogram and boxplot to visualize the distribution of the dataset and removed the outlier data which is < 10 or > 45.

- 2) (10 points) Perform a multiple regression on the dataset you pre-processed in question one. The response variable is mpg. Use the lm() function in R.

  a) Which predictors appear to have a significant relationship to the response.
  Ans: According to the multiple linear regression function, only variable “year” (0.746) and “origin” (1.371) have positive coefficients, so I believe that these variable have significant relationship to the mpg.

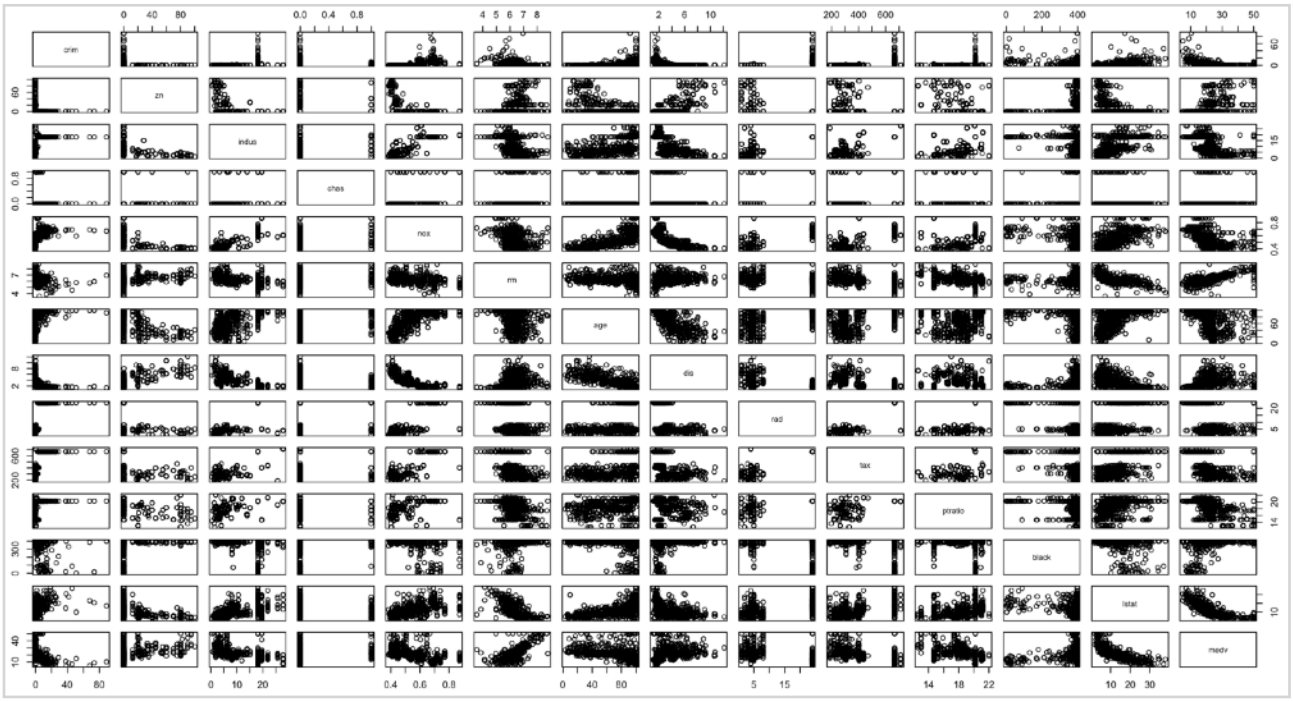
  b) What does the coefficient variable for “year” suggest?

  Ans: 0.746

  c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?
  Ans: The most significant interactions with mpg is cylinders:origin, the regression

coefficient for cylinders:origin is 3857.55760, then are year:origin(233.07716), horsepower:origin(197.80306), cylinders:acceleration:origin(146.27264).

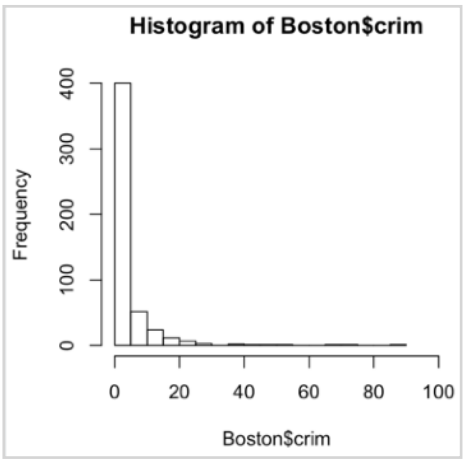
- 3) (10 points) ISLR textbook exercise 2.10 modified: This exercise concerns the boston housing data in the MASS library (>library(MASS) >data(Boston)).
  - a) Make pairwise scatterplots of the predictors, and describe your findings. pairs(Boston)

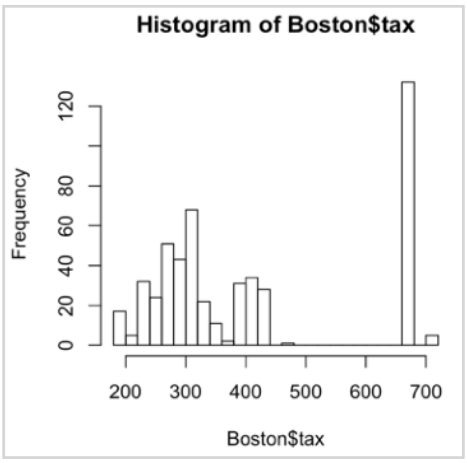


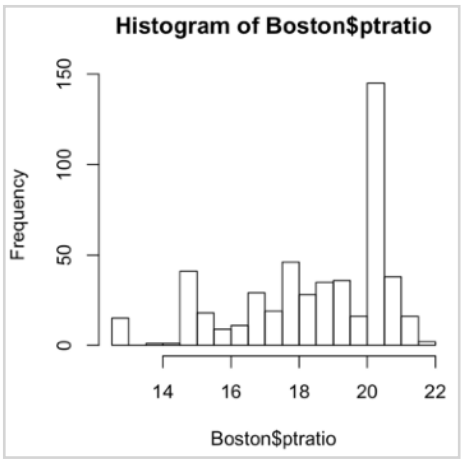
Ans: There are too many predictors that make it hard to find the correlation at a glance, but we can still find that there are negative correlations between “Istat” and “medv” , “nox” and “dis”.

  - b) Are any of the predictors associated with per capita crime rate?
    Ans: According to the correlation coefficients, there is still association between capita and else predictors. The top positive correlation to the per capita crime rate is the index of accessibility to radial highways (0.626), and the top negative one is the median value of owner-occupied homes in \$1000s (-0.388).

- 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.
  Ans:







  - High crime rate: According to the histogram and quantile table, there is 10% of the suburbs have the crime rate that higher than 10.75%, and only 5% of the suburbs have the crime rate that higher than 15.79%.

  - High tax rate: It is clear that the range of higher tax is about 650-750 in the histogram, after calculating, there is about 27% of the suburbs have the high tax (>650).

  - High pupil-teacher ratios: According to the histogram and quantile table, there is 25% of the suburbs have the pupil-teacher ratios that higher than 20.20%.

- 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.
  Ans: There is 64 suburbs have more than seven rooms per dwelling, and 13 more than eight rooms per dwelling. There is only 0.02% of the suburbs have more than eight rooms per dwelling, so it can be classified as the outlier.