Lab 1: Exploratory Data Analysis

This Bitter Earth

Alice Chang

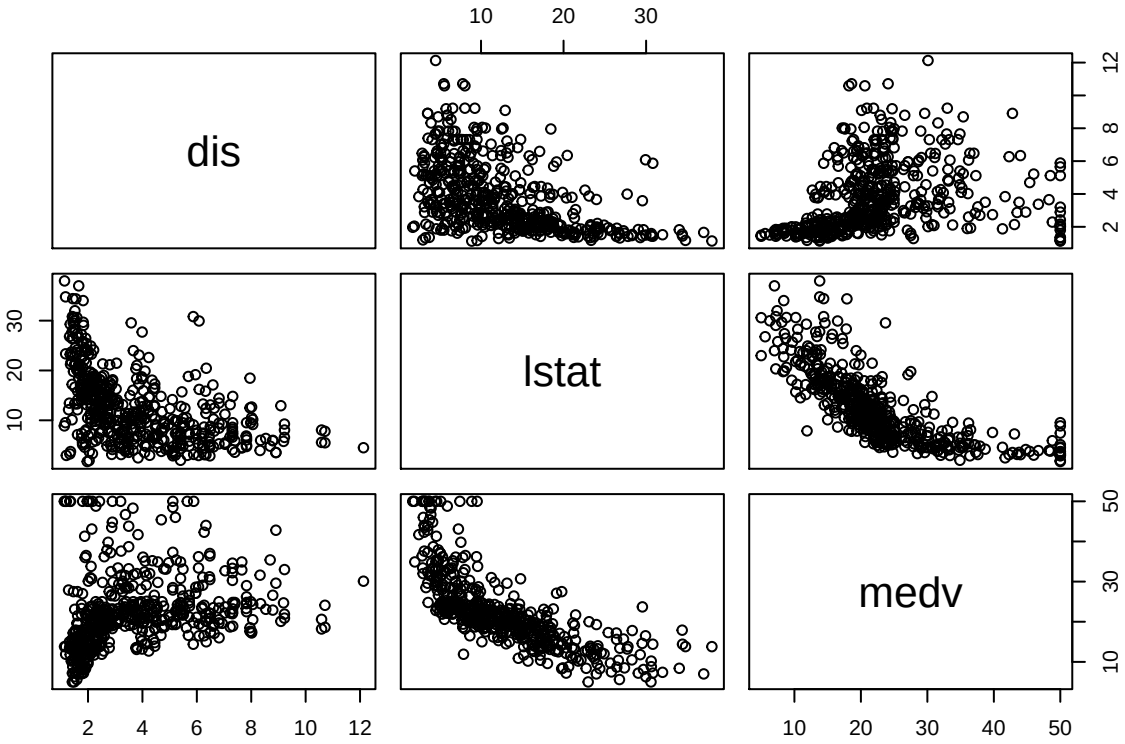
Exercise 1

```
library(MASS)
data(Boston)
```

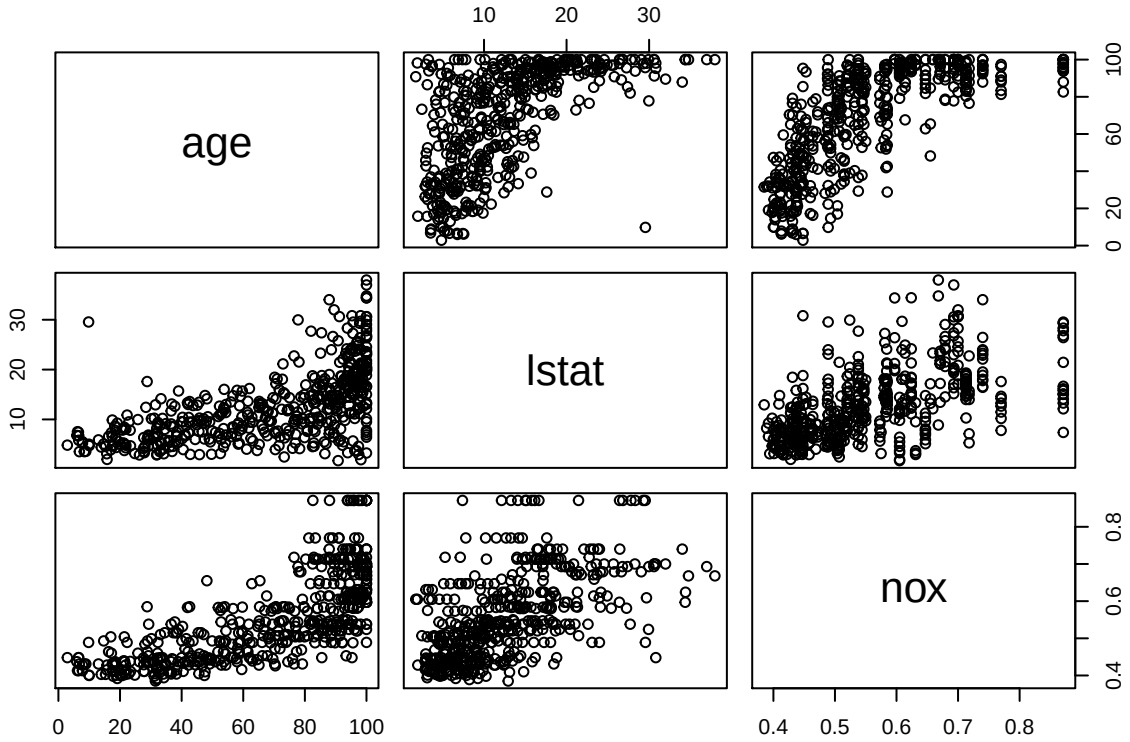
There are 506 rows and 14 columns. The rows represent the number of cases/observations in the dataset. The cases are each of the suburbs in Boston that are included in this dataset. The columns represent the variables in this dataset measuring different attributes of each suburb in the Boston housing area. If we were to build a model assessing the relationship between different features of the suburb ond housing values, there would be 13 independent variables (x) and one target dependent variable (y). The independent variables include charactersitics of the town's businesses/education/employment/crime, the racial/economic composition of the population, as well as features of individual dwellings. The dependent variable (medv) would measure the median housing value in each suburb.

Exercise 2

```
pairs(~dis + lstat + medv, Boston)
```




```
pairs(~age + lstat + nox, Boston)
```



a) There isa strong negative curvilinear relationship between the perentage of the population that is "lower status" andthe median housing value of owner-occupied homes in a suburb (lstat/medv). In general, as the proportionof the population that is "lower status" increases, the housing value of owner-occupied homes in a suburb descreases.

There is a moderately strong negative curvilinear relationship between a suburb's distance from Boston employment centers and the percentage of the population that is "lower status" (lstat/dis). In general, as the "lower status" population of a suburb increases, its distance from employment centers decreases.

There is a very weak positive curvilinear relationship between the median housing value of owner-occupied homes in a suburb and its distance from employment centers (dis/medv). To speak very generally, as a suburb's distance from employment centers increases, the median housing value of homes increases as well.

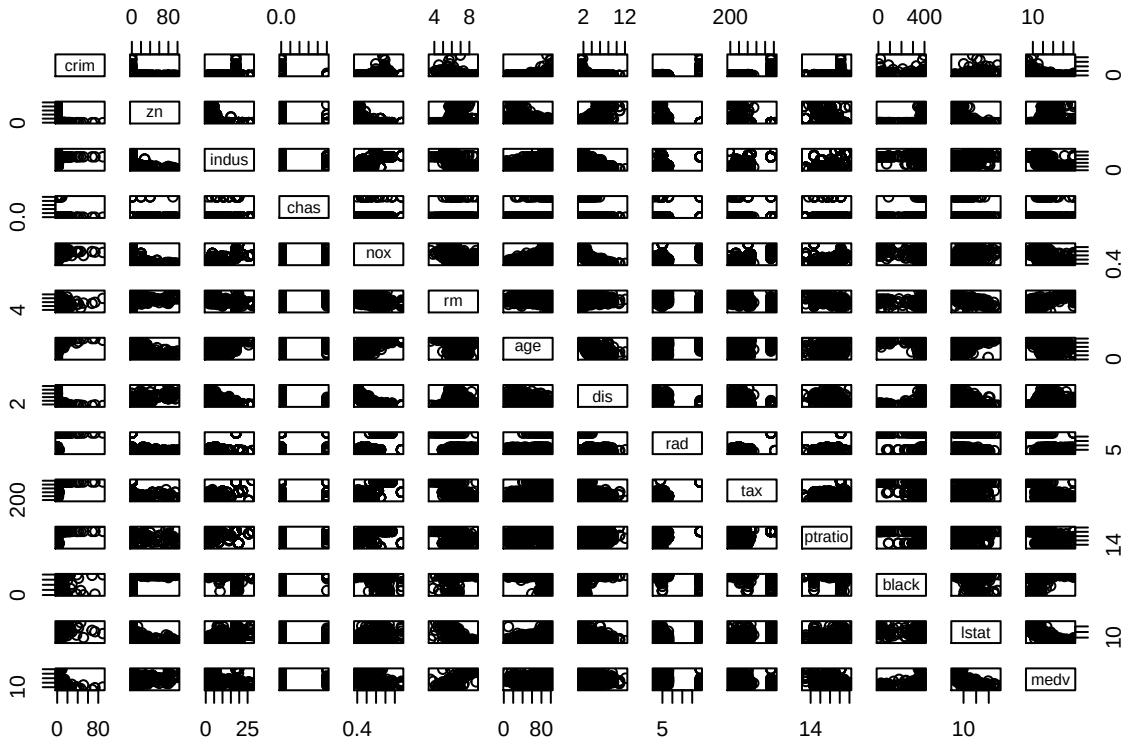
b) There is a weak positive linear relationship between the percentage of the population that is "lower status" and the concentration of nitrogen oxides in a suburb (nox/lstat). In general, as the population of "lower status" residents increases, the concentration of nitrogen oxides in the town increases as well.

There is a moderately strong positive curvilinear relationship between the proportion of older owner-occupied units in a suburb and the concentration of nitrogen oxides in a town (age/nox). In general, as the proportion of owner-occupied units built prior to 1940 increases, the concentration of nitrogen oxides in a suburb increases as well.

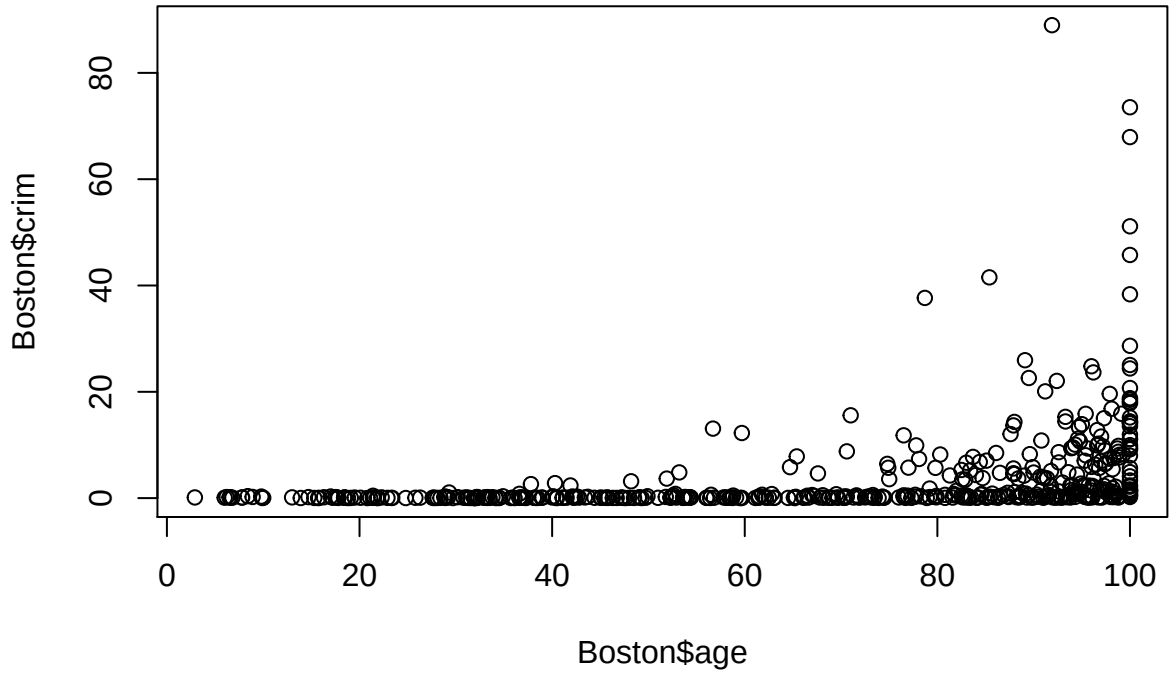
There is a weak positive curvilinear relationship between the proportion of older owner-occupied units and the percentage of lower-status residents in a suburb (age/lstat). In general, as the percentage of lower-status residents in a suburb increases, the proportion of older owner-occupied units in a suburb increases as well.

Exercise 3

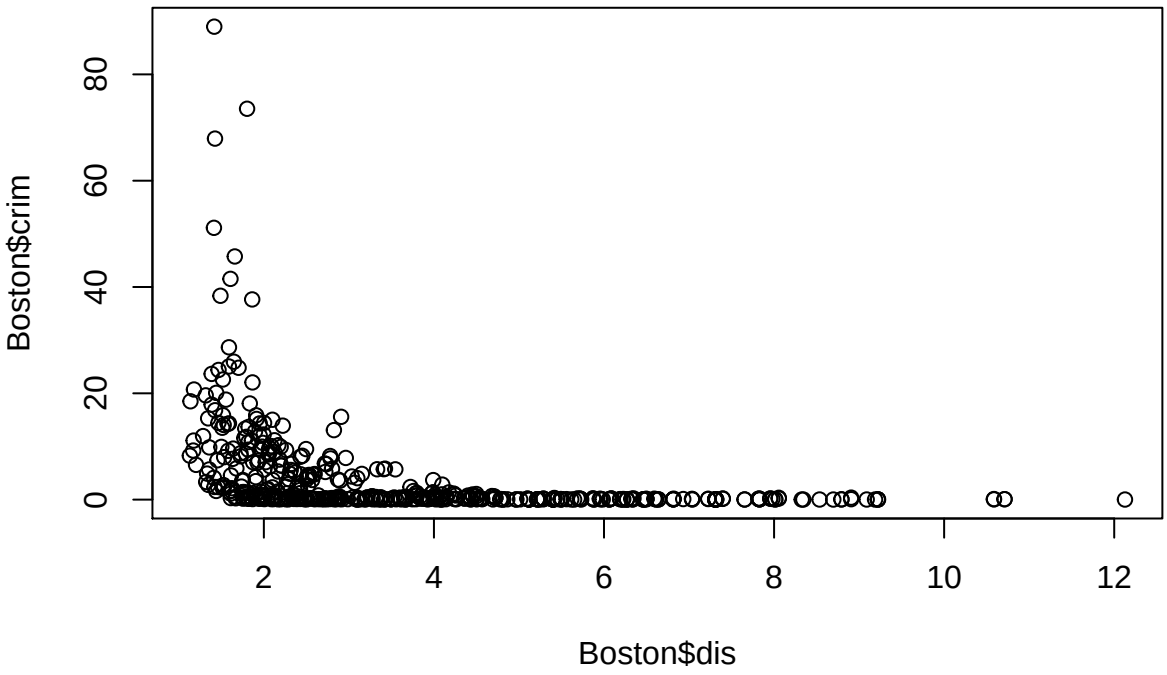
```
pairs(Boston)
```



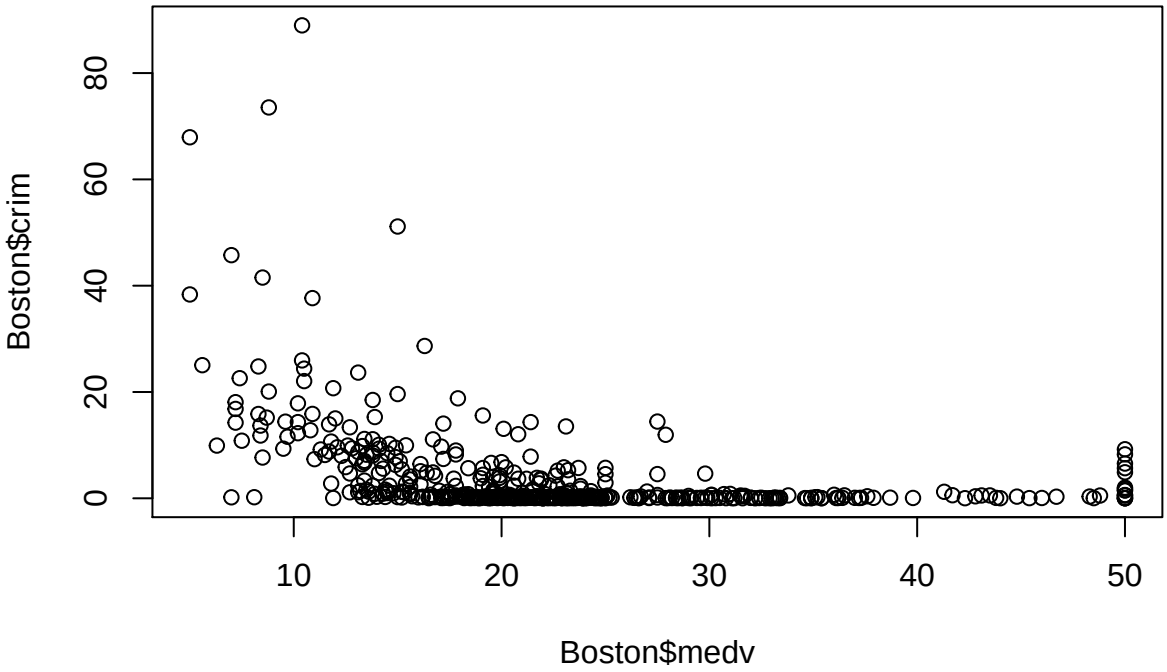

```
plot(Boston$age, Boston$crim)
```




```
plot(Boston$dis, Boston$crim)
```




```
plot(Boston$medv, Boston$crim)
```



The predictors that have the strongest association with per capita crime rate are the proportion of older owner-occupied units (age), the weighted distance of suburbs from employment centers (dis), and the median value of owner-occupied homes (medv).

There is a very weak positive curvilinear relationship between the proportion of older owner-occupied units in a suburb and per capita crime rate. In general, as the proportion of older homes increases, per capita crime rate increases as well.

There is a very weak negative curvilinear relationship between the distance of a suburb from employment centers and per capita crime rate. In general as the distance of a suburb from employment centers in Boston

increases, per capita crime rate decreases.

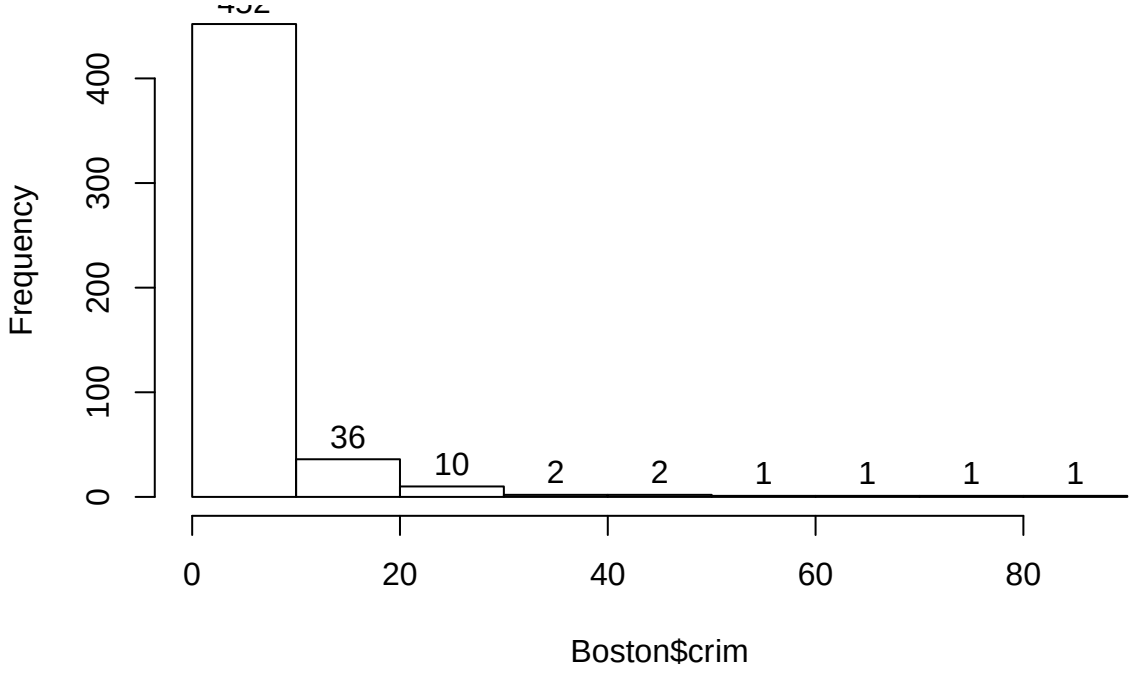
There is a very weak negative curvilinear relationship between median housing value in a suburb and per capita crime rate. In general, as the median value of owner-occupied units in a suburb increases, per capita crime rate decreases.

Exercise 4

```
hc <- hist(Boston$crim)
hc
```

```
## $breaks
##  [1]  0 10 20 30 40 50 60 70 80 90
##
## $counts
## [1] 452  36  10   2   2   1   1   1   1
##
## $density
## [1] 0.0893280632 0.0071146245 0.0019762846 0.0003952569 0.0003952569
## [6] 0.0001976285 0.0001976285 0.0001976285 0.0001976285
##
## $mids
## [1]  5 15 25 35 45 55 65 75 85
##
## $xname
## [1] "Boston$crim"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
text(hc$mids,hc$counts,labels=hc$counts, adj=c(0.5, -0.5))
```



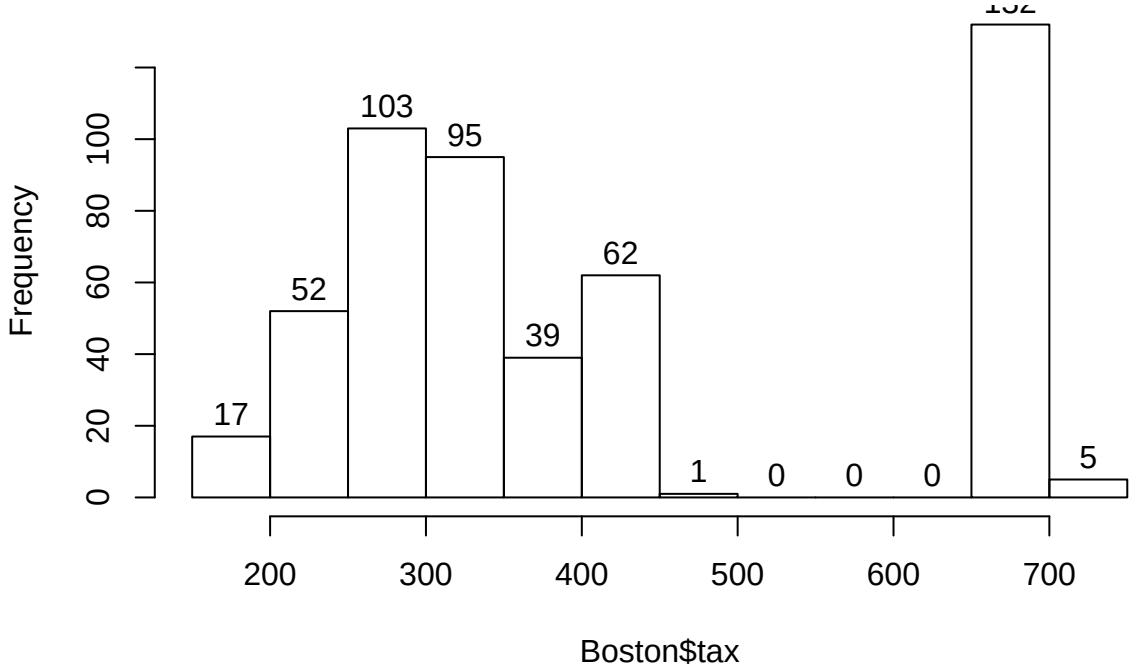

```
range(Boston$crim)
```

```
## [1]  0.00632 88.97620
```

```
ht <- hist(Boston$tax)
ht
```

```
## $breaks
##  [1] 150 200 250 300 350 400 450 500 550 600 650 700 750
##
## $counts
##  [1]  17  52 103  95  39  62   1   0   0   0 132   5
##
## $density
##  [1] 6.719368e-04 2.055336e-03 4.071146e-03 3.754941e-03 1.541502e-03
##  [6] 2.450593e-03 3.952569e-05 0.000000e+00 0.000000e+00 0.000000e+00
## [11] 5.217391e-03 1.976285e-04
##
## $mids
##  [1] 175 225 275 325 375 425 475 525 575 625 675 725
##
## $xname
## [1] "Boston$tax"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
text(ht$mids,ht$counts,labels=ht$counts, adj=c(0.5, -0.5))
```



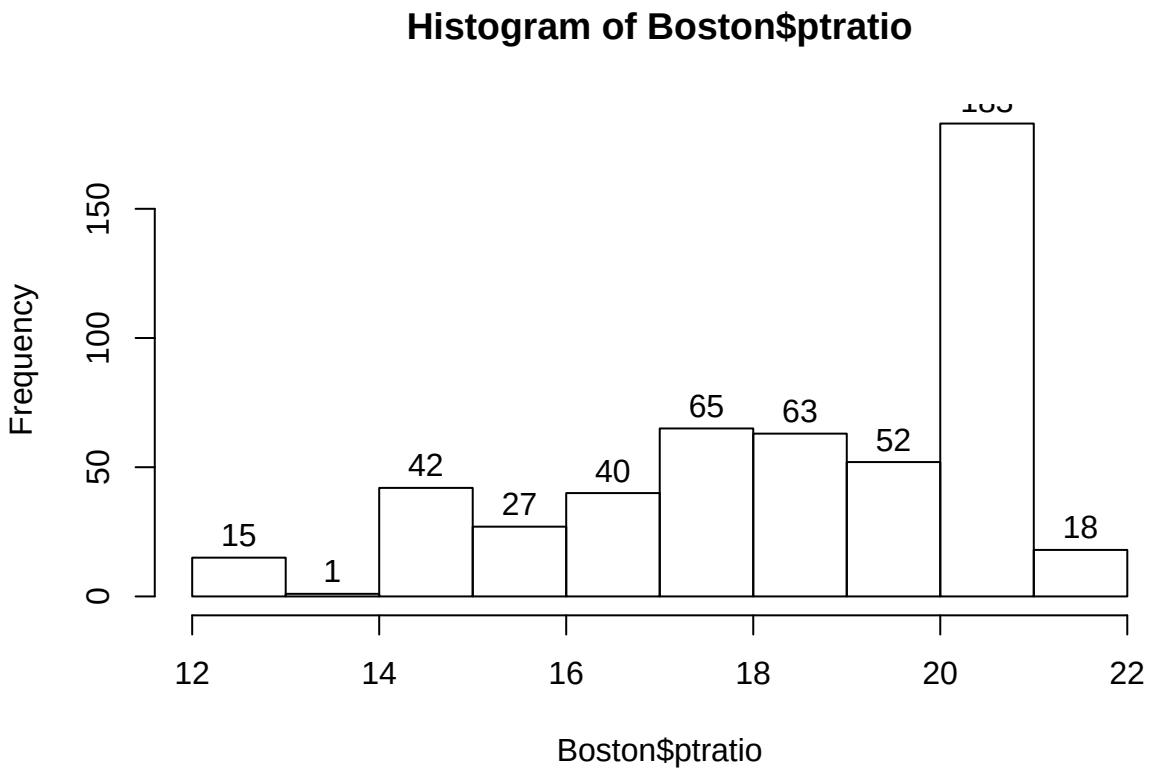

```
range(Boston$tax)
```

```
## [1] 187 711
```

```
hp <- hist(Boston$ptratio)
hp
```

```
## $breaks
##  [1] 12 13 14 15 16 17 18 19 20 21 22
##
## $counts
##  [1]  15   1  42  27  40  65  63  52 183  18
##
## $density
##  [1] 0.029644269 0.001976285 0.083003953 0.053359684 0.079051383
##  [6] 0.128458498 0.124505929 0.102766798 0.361660079 0.035573123
##
## $mids
##  [1] 12.5 13.5 14.5 15.5 16.5 17.5 18.5 19.5 20.5 21.5
##
## $xname
## [1] "Boston$ptratio"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
text(hp$mids,hp$counts,labels=hp$counts, adj=c(0.5, -0.5))
```




```
range(Boston$ptratio)
```

```
## [1] 12.6 22.0
```

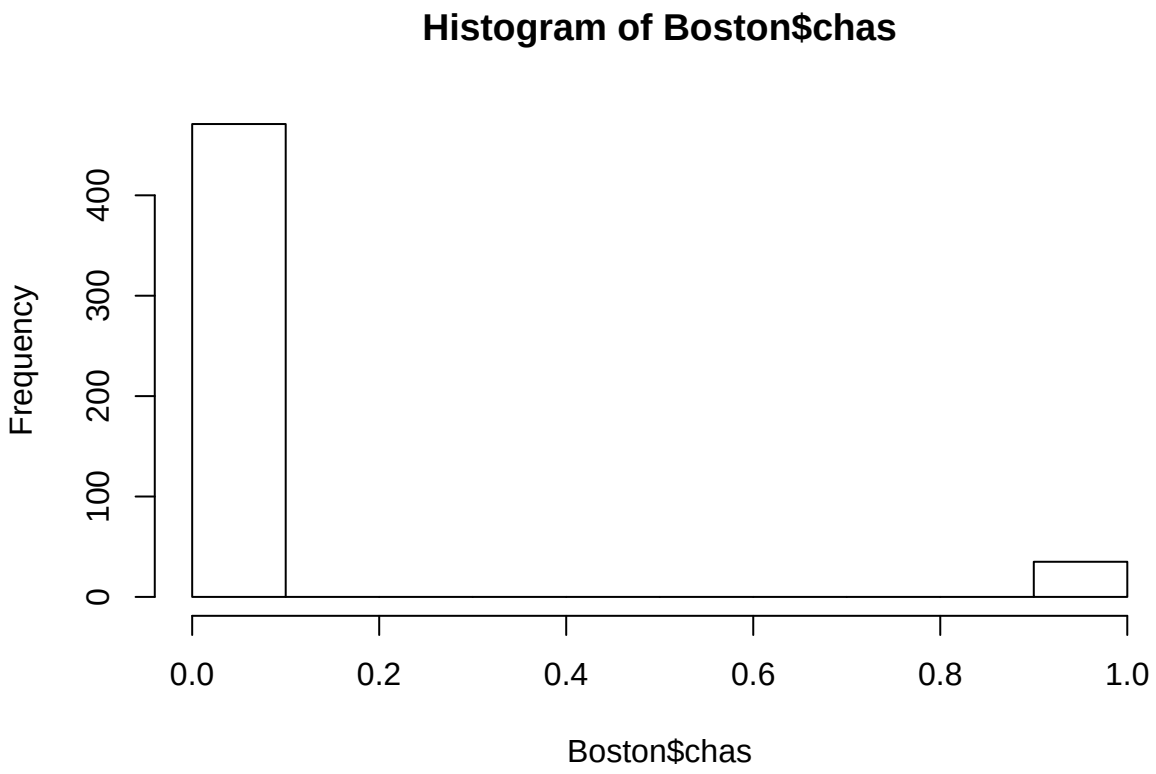
Per capita crime rates in the suburbs range from 0.00632 to 88.9762. The distribution of crime rates are skewed right. A large majority of the suburbs have low crime rates below 20. Only 18 of the suburbs have particularly high crime rates above 20.

Property tax rates in the suburbs range from 187 to 711. The distribution of property tax rates are roughly symmetrical, but there is a peak of 132 suburbs with property tax rates between 650 and 700. Most of the suburbs have property tax rates below 500, but there are 137 suburbs with higher tax rates above 650.

Pupil-teacher ratios among suburbs range from 12.6 to 22. The distribution of pupil-teacher ratios is skewed left. Only 18 suburbs have the highest pupil-teacher ratio of 21, but that is not particularly high.

Exercise 5

```
hist(Boston$chas)
```




```
table(Boston$chas)
```

```
##
##   0   1
## 471  35
```

35 suburbs in the data set bound the Charles River.

Exercise 6

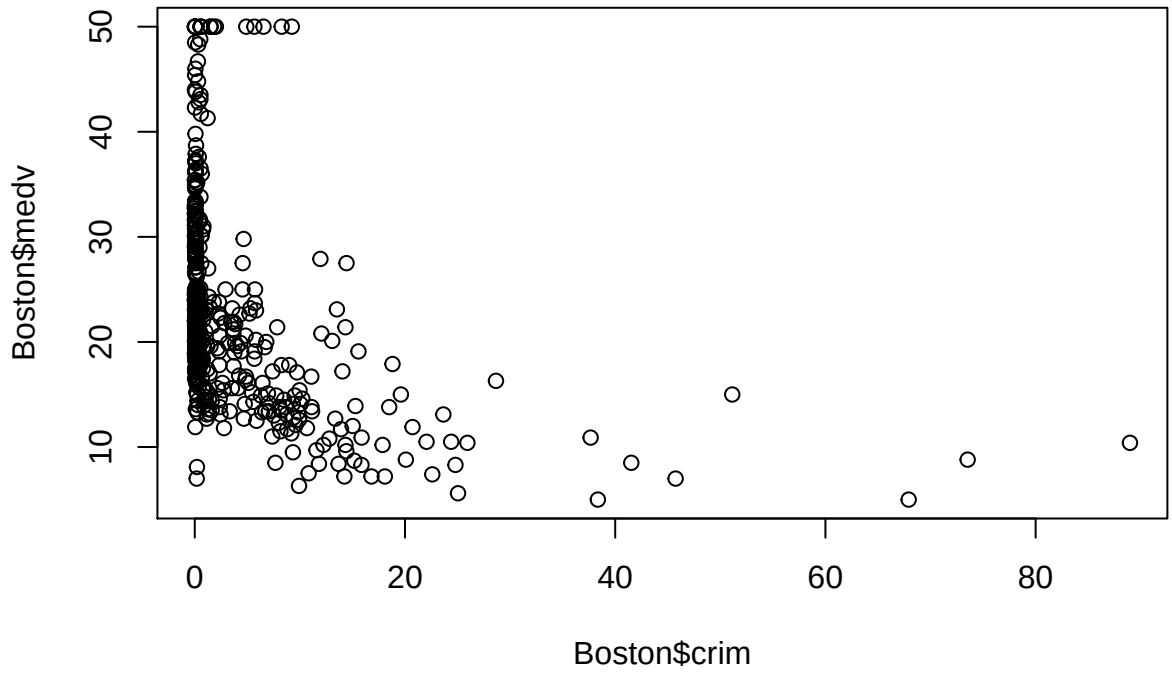
```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

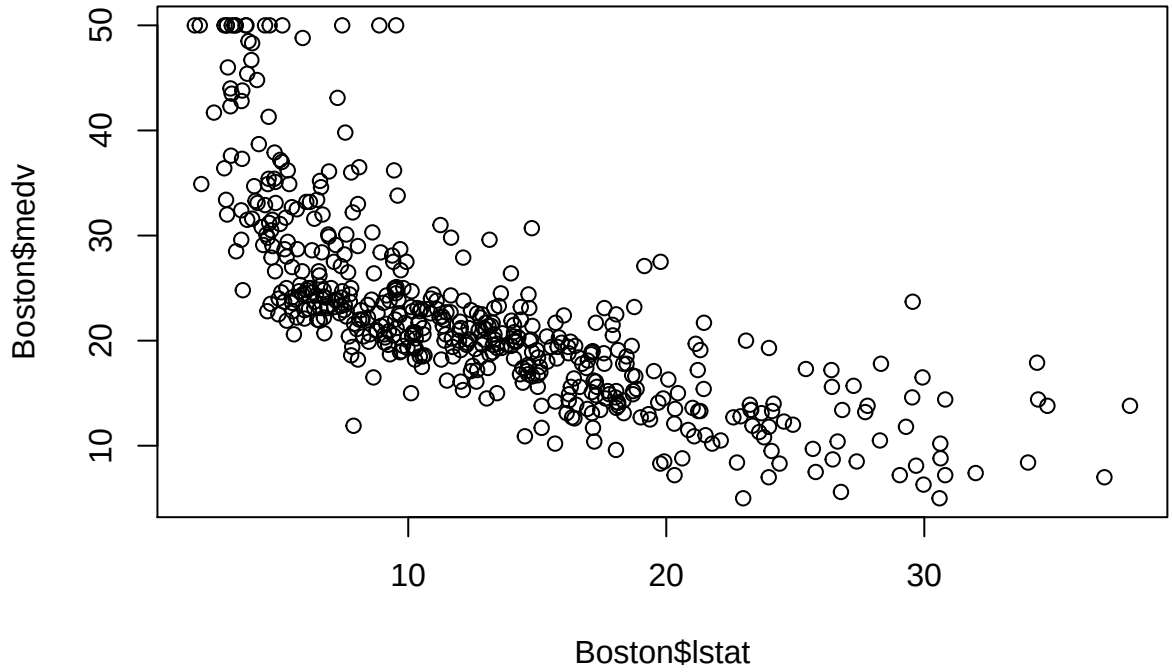
The median pupil-teacher ratio among the towns in the dataset is 19.05

Exercise 7

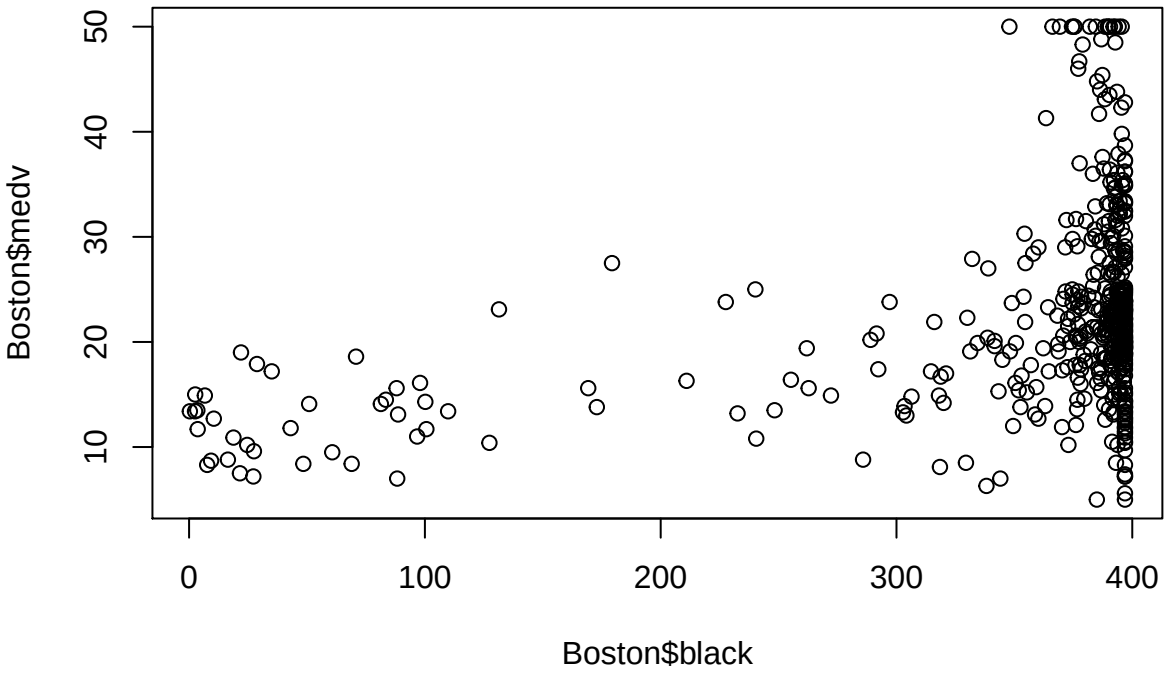
```
plot(Boston$crim, Boston$medv)
```




```
plot(Boston$lstat, Boston$medv)
```




```
plot(Boston$black, Boston$medv)
```



In building a model to predict the average value of a home, my output/response variable would be "medv" which measures the median value of owner-occupied homes. I would first evaluate all of the other 13 variables measuring other attributes of Boston suburbs as inputs/predictors. However, based off of my preliminary exploration of the data set, per capita crime rate (crim), the proportion of black residents in the town population (black), and the percentage of "lower status" residents in the suburb seem to have stronger associations with housing value. Consequently, it might be reasonable to focus on those three variables (crim, black, lstat) as predictors in my model.