Chapter 2 Problem 10

Andira Putri

This exercise involves the `Boston` housing data set.

a.) Load in the `Boston` data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

```r
library(MASS)
data=Boston
```

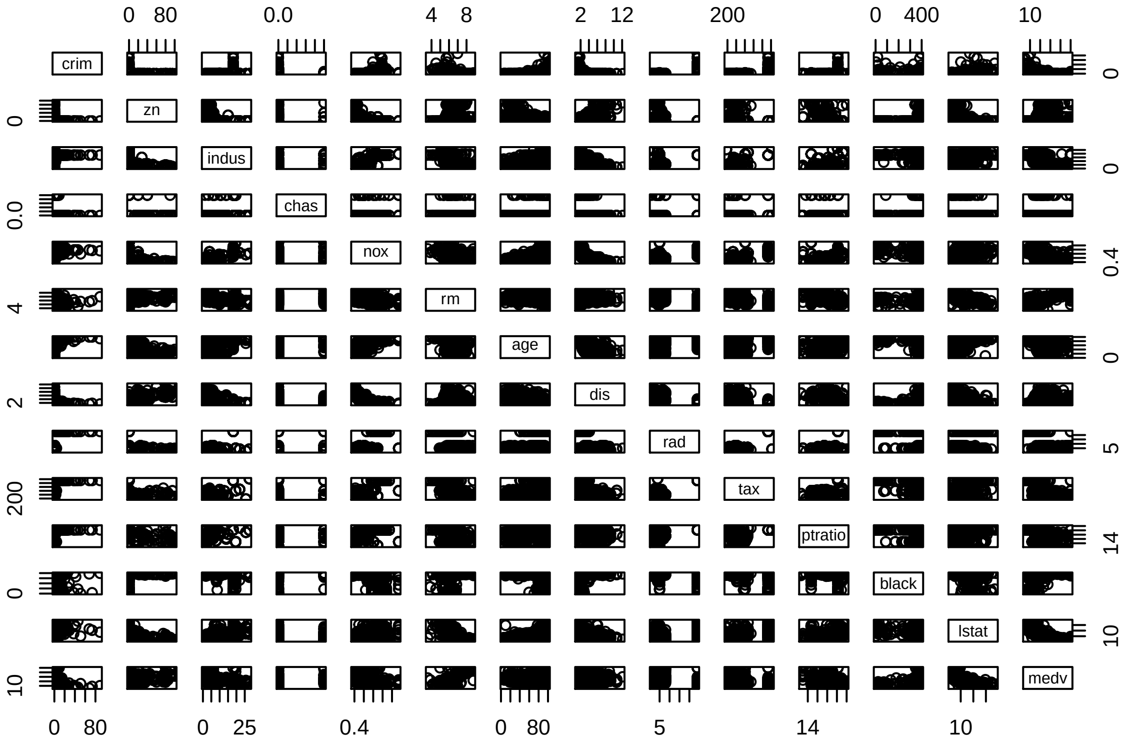
Number of rows: 506 (suburbs of Boston)

Number of columns: 14, representing:

1. crim-per capita crime rate by town.
2. zn-proportion of residential land zoned for lots over 25,000 sq.ft.
3. indus-proportion of non-retail business acres per town.
4. chas-Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. nox-nitrogen oxides concentration (parts per 10 million).
6. rm-average number of rooms per dwelling.
7. age-proportion of owner-occupied units built prior to 1940.
8. dis-weighted mean of distances to five Boston employment centres.
9. rad-index of accessibility to radial highways.
10. tax-full-value property-tax rate per $10,000.
11. ptratio-pupil-teacher ratio by town.
12. black-1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks by town.
13. lstat-lower status of the population (percent).
14. medv-median value of owner-occupied homes in $1000s.

b.) Make some pairwise scatterplots of the predictors. Describe your findings.

```r
pairs(data)
```



c.) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Continuation from part B.... As age of town increases, crime rate increases.

Crime rates are higher when in closer proximity to employment centers.

There are some peaks of crime rate in the following:

*Low proportion in residential-zoned lots

*When tract does not bound Charles River

*High access to radial highways

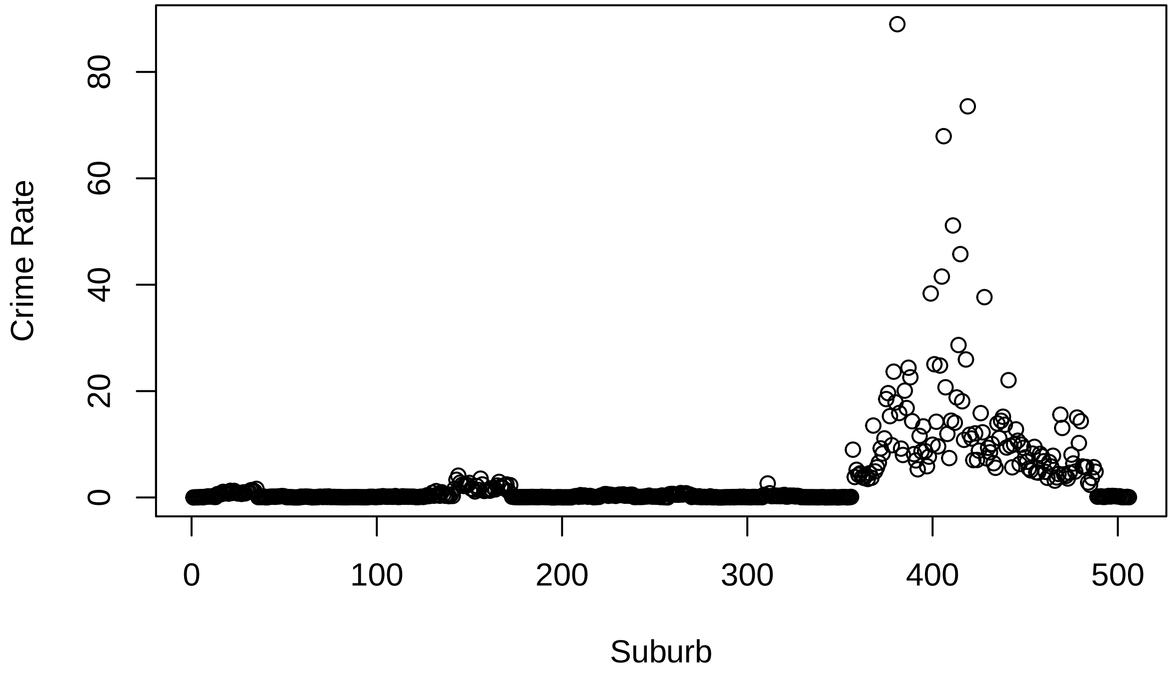
d.) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

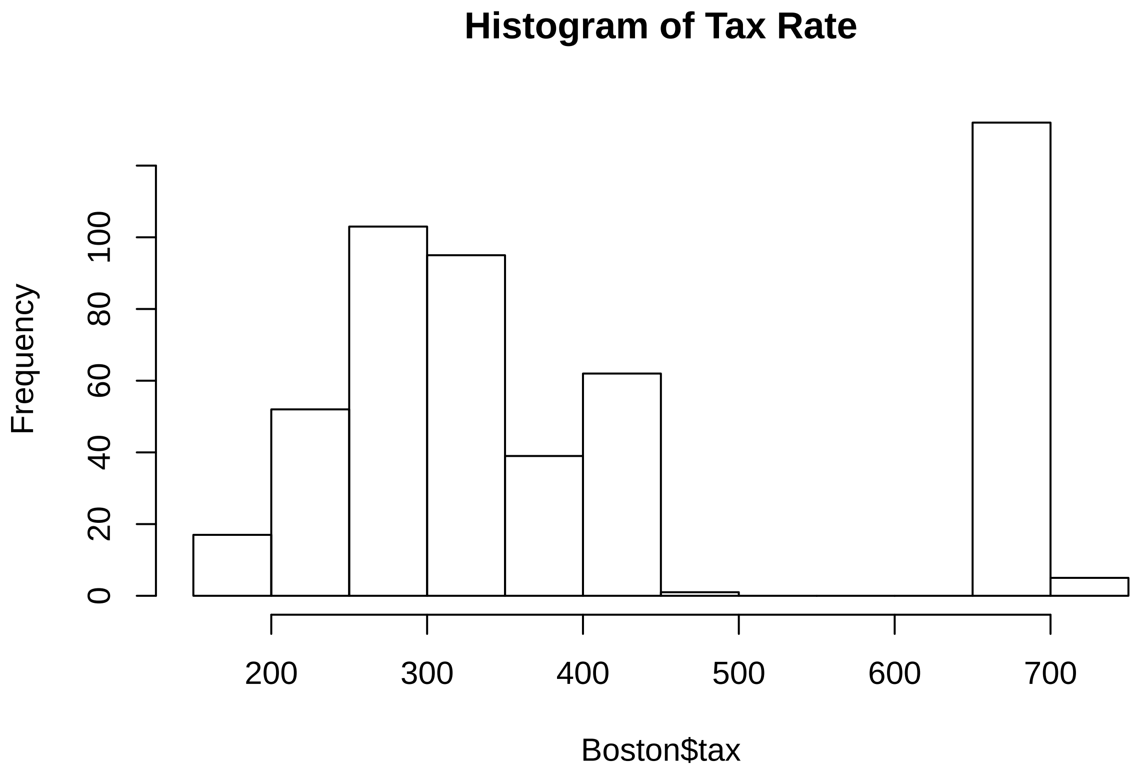
The range of the crime predictor is 88.97%, so the data is very spread out. Given that the maximum crime rate is 88.98% and the mean value is 3.61%, at least one suburb has a significantly higher crime rate than others. Out of curiosity, here is the plot of crime rate with each suburb:

```
plot(Boston$crim,xlab="Suburb",ylab="Crime Rate")
```



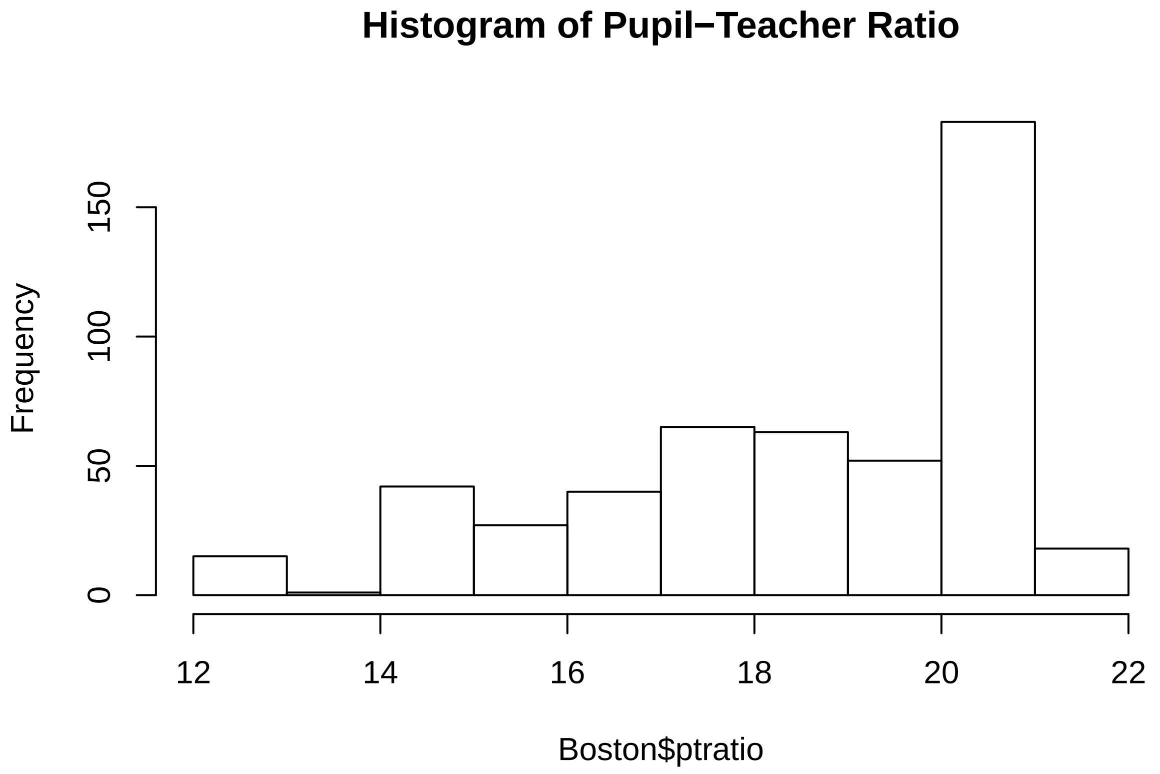
Then, I observed a histogram of Boston property tax rates since a regular plot did not help as much. The two peaks indicate that the tax is bimodal which suggests we have two different groups of suburbs. One group pays a high tax while the other pays a medium tax. There is not a major differentiation like in the crime predictor. The range is 513, so again, the data for tax rates is spread out.

```r
hist(Boston$tax,main="Histogram of Tax Rate")
```



Finally, we observe pupil-teacher ratios with a histogram. There is one obvious peak, though some cities have very low pupil-teacher ratios comparatively. The range is 9.4.

```r
hist(Boston$ptratio,main="Histogram of Pupil-Teacher Ratio")
```



e.) How many of the suburbs in this data set bound the Charles River?

```
count=nrow(Boston[Boston$chas==1,])
count
```

```
## [1] 35
```

f.) What is the median pupil-teacher ratio among the towns in the data set?

19.05 (refer to the summary table in part d).

g.) Which suburb has the lowest median value of owner-occupied houses (medv)? What are the values of the other predictors for that suburb, and how do these values compare to the overall ranges for those predictors? Comment on your findings.

```
#use which.min/which.max to find min/max values
Min=Boston[which.min(Boston$medv),]
Min
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9
##     lstat medv
## 399 30.59    5
```

This returns observation 399. Now, we analyze how suburb 399 compares with the whole data set.

```
sapply(Boston[,1:14], quantile) #whole data set
```

```
##             crim    zn indus chas    nox      rm      age        dis rad tax
## 0%    0.006320   0.0  0.46    0 0.385 3.5610   2.900  1.129600   1 187
## 25%   0.082045   0.0  5.19    0 0.449 5.8855  45.025  2.100175   4 279
## 50%   0.256510   0.0  9.69    0 0.538 6.2085  77.500  3.207450   5 330
## 75%   3.677083  12.5 18.10    0 0.624 6.6235  94.075  5.188425  24 666
## 100% 88.976200 100.0 27.74    1 0.871 8.7800 100.000 12.126500  24 711
##      ptratio    black  lstat   medv
## 0%     12.60   0.3200  1.730  5.000
## 25%    17.40 375.3775  6.950 17.025
## 50%    19.05 391.4400 11.360 21.200
## 75%    20.20 396.2250 16.955 25.000
## 100%   22.00 396.9000 37.970 50.000
```

From observation 399, these predictors are at or above the 75th percentile when compared to the entire Boston data set: crim, indus, nox, age, rad, tax, ptratio, lstat

h.) 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.

```
count7rooms=nrow(Boston[Boston$rm>7,])
count8rooms=nrow(Boston[Boston$rm>81,])
```

There are 64 suburbs with more than 7 rooms per house on average, and there are 13 suburbs with more than 8 rooms per house on average.

```
sapply(Boston[Boston$rm > 8,], mean)
```

```
##        crim          zn       indus        chas         nox          rm
##   0.7187954  13.6153846   7.0784615   0.1538462   0.5392385   8.3485385
##         age         dis         rad         tax     ptratio       black
##  71.5384615   3.4301923   7.4615385 325.0769231  16.3615385 385.2107692
##       lstat        medv
##   4.3100000  44.2000000
```

We compare the above table with the one given by `sapply(Boston[,1:14], quantile)`.

- Crime rate is above the 50th percentile.
- There is a lower pupil-teacher ratio.
- There is a small percentage of people in the lower status.
- The median value of homes is much higher.