Stats 101C HW 1

Anna Piskun

10/14/2020

Problem 1: Exercise 5

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

   One advantage of a flexible approach for regression or classification is that it allows us to simplify complex data sets and better fit non-linear models as well as decrease bias. By having a very flexible model we can more easily fit complicated data and trends. However, a disadvantage of a very flexible approach is that we run the risk of overfitting the training data due to having a greater number of parameters which as a result may capture excess noise not allowing us to extrapolate our findings and directly increasing variance. When faced with strange data that follows very strange, complicated, non-linear trends it is preferred to use a more flexible approach. Additionally, a more flexible approach may be preffered when we want to increase the predictability of our model. On the other had, a less flexible approach may be preferred when attempting to simplify data patterns to make it easier to visualize and understand. Likewise, if the data follows a linear trend it is preferred to use a less flexible approach.

Problem 2: Exercise 6

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a para- metric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

   Using a parametric statistical learning approach, we first make an assumption about the form of function f(x). This leads us to have a fixed set of parameters, that we fit the training data set to. Non-parametric statistical models may have the potential to have an infinite number of parameters, meaning that the complexity of a non-parametric model grows with the training data. Thus, we end up choosing a function that closely fits the data (or in other words we fit a function to the data). Some advantages of a parametric approach are that it allows us to fit any sort of data to a functional form (linear regression, logistic, etc.) as well as allow us to be able to interpret the results more easily. Additionally, parametric models do not require as large of a training data set in comparison to a non-parametric model and can create meaningful findings even if overall fit is not completely perfect. However, some disadvantages of the parametric statistical learning approach is that with too many parameters, the model may over-fit the training data preventing us from being able to extrapolate the model's findings and leading to large variance and potentially high bias.

Problem 3: Exercise 10

10. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

```r
library(MASS)
```

Now the data set is contained in the object Boston.

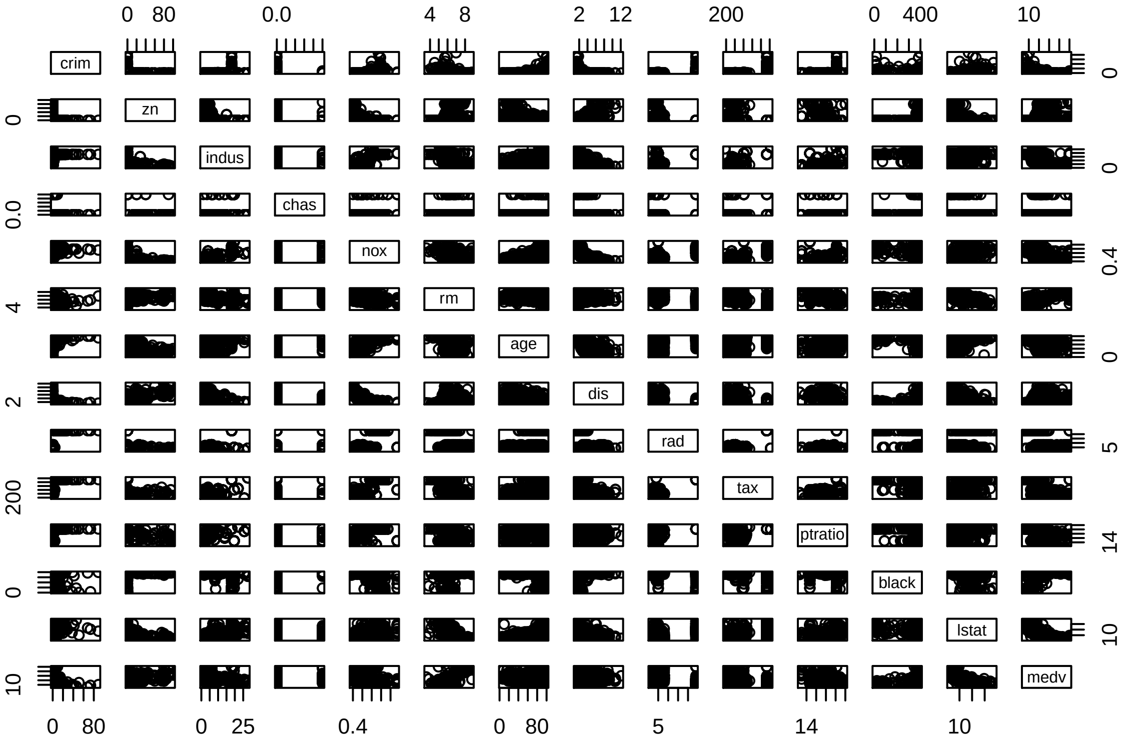
Read about the data set:

How many rows are in this data set? How many columns? What do the rows and columns represent?

There are 506 rows and 14 columns. The columns represent predictors used (crime, proportion of non-retail business, pupil-teacher ratio, etc.) while the rows represent the individual suburbs included in this data set.

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

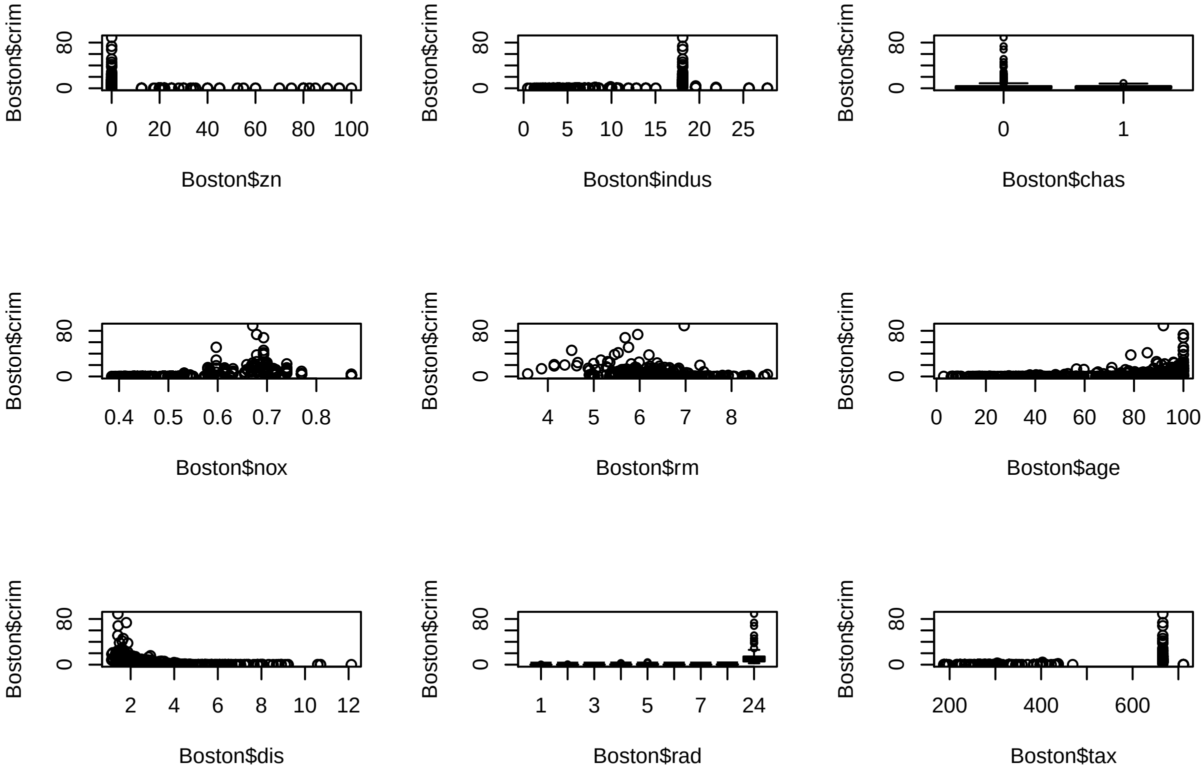
```
pairs(Boston)
```



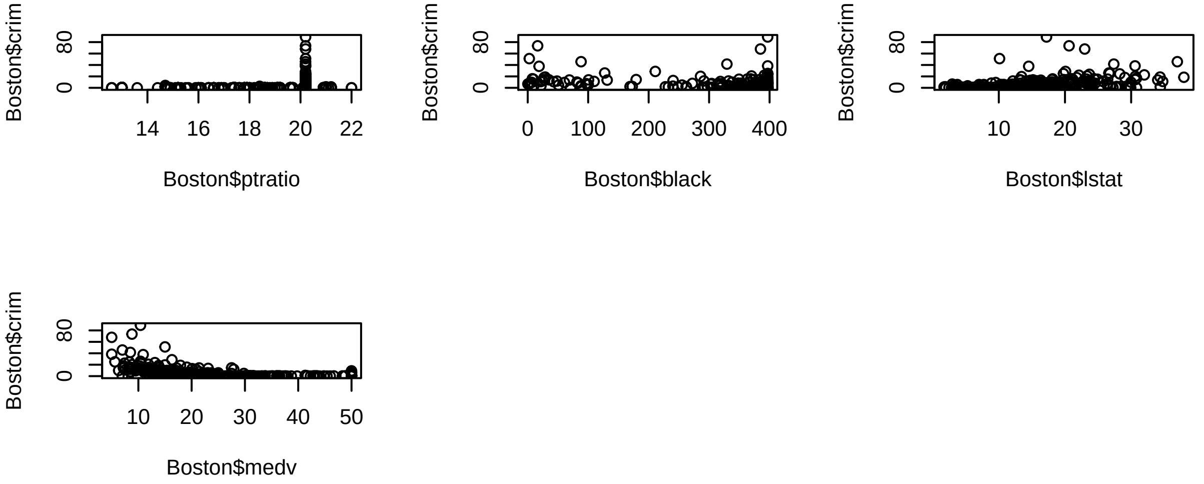
From the pairwise scatterplots of all the predictors we are able to see a few associations. For example, in the most bottom left hand corner we can clearly see that there is a potential relationship between per capita crime rate by town and the median value of owner occupied homes. It looks as though the lower the crime rate is, the higher the median value is of a home. This makes sense given the fact that most people see living in a safe community as desirable. Another clear relationship exists between crime and age, where the higher the proportion of owner-occupied units built prior to 1940 is then the higher the crime rate is. Another potential relationship exists between lstat and medv, where the higher percentage of people of lower status is associated with a lower median home value. Despite seeing some initial relationships, we must conduct further testing to see if these predictors are significant.

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

```
par(mfrow = c(3,3))
plot(Boston$zn, Boston$crim)
plot(Boston$indus, Boston$crim)
boxplot((Boston$crim~Boston$chas))
plot(Boston$nox, Boston$crim)
plot(Boston$rm, Boston$crim)
plot(Boston$age, Boston$crim)
plot(Boston$dis, Boston$crim)
boxplot(Boston$crim~Boston$rad)
plot(Boston$tax, Boston$crim)
```



```r
plot(Boston$ptratio, Boston$crim)
plot(Boston$black, Boston$crim)
plot(Boston$lstat, Boston$crim)
plot(Boston$medv, Boston$crim)
```



From the above plots we can see that there may be a relationship between age and crime, dis and crime, lstat and crime, and medv and crime. The higher the proportion of owner-occupied units built prior to 1940 the higher the crime rate is. However, a longer mean distance from employment centers is associated with a lower crime rate. There is a slight upward trend between lstat and crim indicating that the higher the proportion of "lower status" people then the higher the crime rate is. Lastly, as to be expected, higher crime rates are associated with lower median value homes.

(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.

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

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

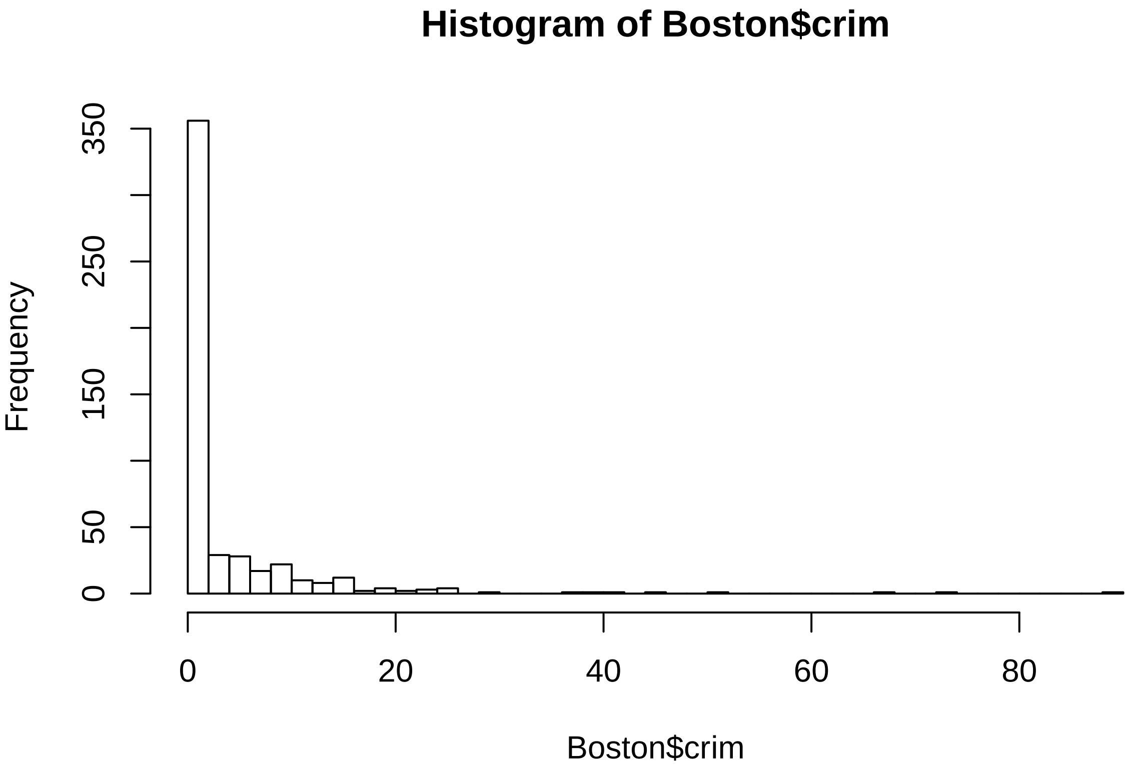
```r
range(Boston$tax)
```

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

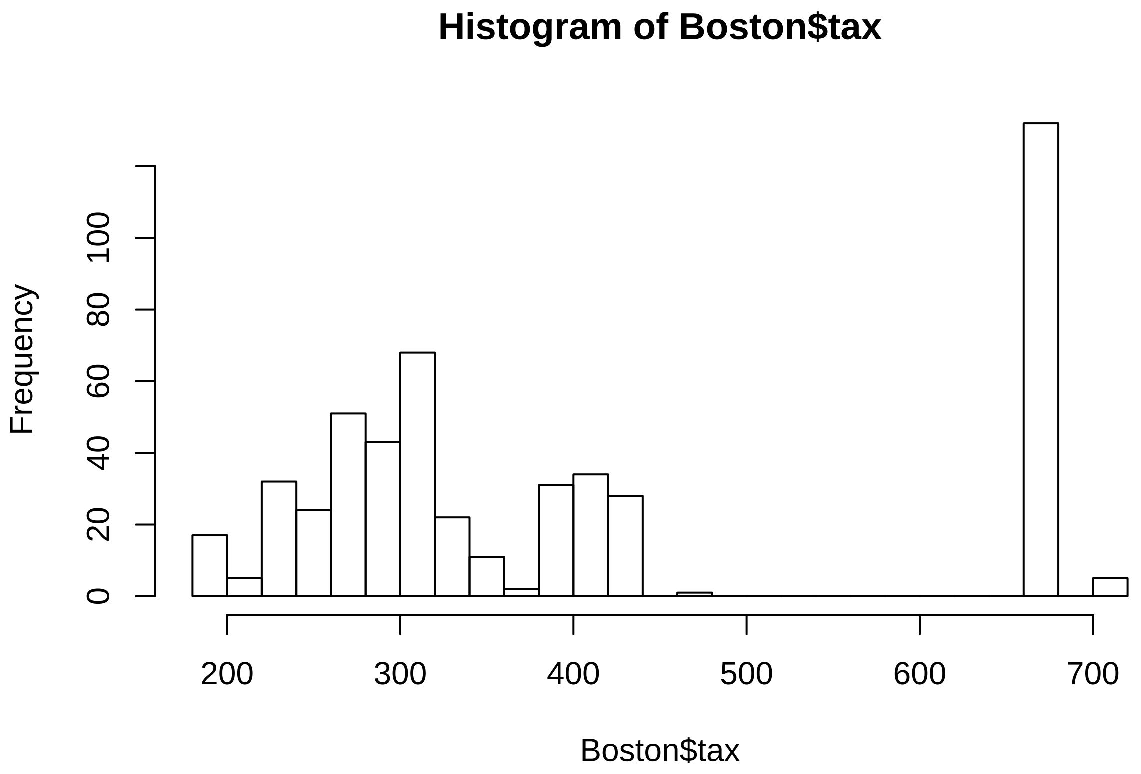
```r
range(Boston$ptratio)
```

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

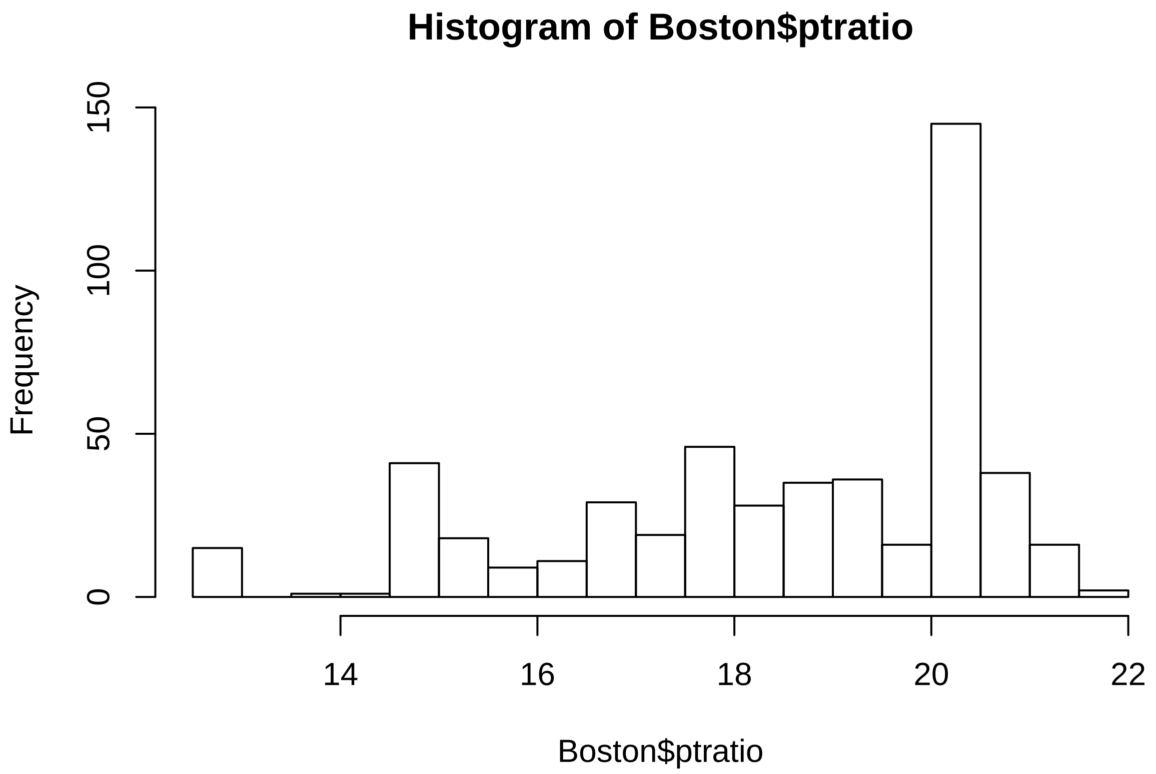
```r
hist(Boston$crim, breaks = 35)
```



```r
hist(Boston$tax, breaks = 20)
```



```
hist(Boston$ptratio, breaks = 20)
```



While most of the Boston suburbs had very low crime rates, a few suburbs stuck out as having extraordinarily high crime rates reaching above 60%. The range of crime rates was 0.00632 to 88.97620. There was also a high frequency of suburbs with high tax rates, but this could potentially be explained another predictor such as proportion of residential land zoned or the proportion of non-retail business acres per town. The range of property tax rate per $10,000 ranged from 187 to 711. Lastly, the majority of Boston suburbs had a pupil to teacher ratio of about 20. While the histogram is left skewed, most suburbs clustered around this ratio. The range for this predictor is from 12.6 to 22.

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

```
sum(Boston$chas != 0)
```

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

35 suburbs are bound by the Charles River.

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

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

19.05 is the median pupil teacher ratio among the towns in this data set.

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

```
n <- order(Boston$medv)

ordered <-  Boston[which(Boston$medv == 5), ]

summary(ordered)
```

```
##       crim             zn        indus          chas        nox       
##  Min.   :38.35   Min.   :0   Min.   :18.1   Min.   :0   Min.   :0.693  
##  1st Qu.:45.74   1st Qu.:0   1st Qu.:18.1   1st Qu.:0   1st Qu.:0.693  
##  Median :53.14   Median :0   Median :18.1   Median :0   Median :0.693  
##  Mean   :53.14   Mean   :0   Mean   :18.1   Mean   :0   Mean   :0.693  
##  3rd Qu.:60.53   3rd Qu.:0   3rd Qu.:18.1   3rd Qu.:0   3rd Qu.:0.693  
##  Max.   :67.92   Max.   :0   Max.   :18.1   Max.   :0   Max.   :0.693  
##        rm             age           dis             rad           tax     
##  Min.   :5.453   Min.   :100   Min.   :1.425   Min.   :24   Min.   :666  
##  1st Qu.:5.511   1st Qu.:100   1st Qu.:1.441   1st Qu.:24   1st Qu.:666  
##  Median :5.568   Median :100   Median :1.458   Median :24   Median :666  
##  Mean   :5.568   Mean   :100   Mean   :1.458   Mean   :24   Mean   :666  
##  3rd Qu.:5.625   3rd Qu.:100   3rd Qu.:1.474   3rd Qu.:24   3rd Qu.:666  
##  Max.   :5.683   Max.   :100   Max.   :1.490   Max.   :24   Max.   :666  
##     ptratio         black           lstat            medv  
##  Min.   :20.2   Min.   :385.0   Min.   :22.98   Min.   :5  
##  1st Qu.:20.2   1st Qu.:388.0   1st Qu.:24.88   1st Qu.:5  
##  Median :20.2   Median :390.9   Median :26.79   Median :5  
##  Mean   :20.2   Mean   :390.9   Mean   :26.79   Mean   :5  
##  3rd Qu.:20.2   3rd Qu.:393.9   3rd Qu.:28.69   3rd Qu.:5  
##  Max.   :20.2   Max.   :396.9   Max.   :30.59   Max.   :5
```

```
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
```

Suburb 399 and 406 have the lowest median values of owner-occupied homes. Looking at the other predictors for Suburb 399 we find that it has a crime rate of 38.35%, no zoning for residential land, a proportion of 18.10 non-retail business acres per town, is not bound by the Charles river, 0.6930 nitrogen oxide concentration, an average of 5.453 rooms per dwelling, 100% of owner occupied units built prior to 1940, is only 1.4896 miles from 5 Boston employment centers, an index of accessibility to radial highways of 24, tax rate of 666, pupil to student ratio of 20.2, a proportion of 396.90 African Americans by town, poverty rate of 30.59%, and median home value of $5,000. Compared with the other suburbs, combined Suburb 399 and Suburb 406 have some of the highest crime rates, lowest residential zoning, high proportion of non-retail business acres, some of the highest levels of nitrogen oxide concentration, lower amount of rooms per dwelling, highest proportion of owner occupied units, are some of the closest suburbs to employment centers, have incredibly high accessibility to highways, a high tax rate and pupil to student ratio, and lastly both these suburbs have one of the highest rates of poverty. All of these factors make Suburb 399 and 406 not desirable and helps to explain their low home value.

(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.

```r
new <- Boston[which(Boston$rm > 8), ]
summary(new)
```

```
##       crim               zn            indus             chas
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
##       nox               rm             age             dis
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv     
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0  
```

```r
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  
```

In this data set, 64 suburbs average more than seven rooms per dwelling, and 13 average more than eight rooms per dwelling. The suburbs that average more than 8 rooms per dwelling on average have lower crime rates, higher residential zoning, lower proportion of non-retail business acres per town, slightly lower nitrogen oxide concentration, higher proportion of owner-occupied units, shorter distance to employment centers, lower index of accessibility, lower tax rates, smaller pupil to student ratio, much lower percentage of poverty, and significantly higher median values of homes. Although a subset of only 13 suburbs, from our above analysis these suburbs may be considered more desirable and therefore be of more value.