HW1

Jingshi

1/25/2017

Question 4

a)

Example 1: Suppose a doctor want to find out whether or not a patient has diabetes (this is the response). The doctor collects several information from the patient such as blood glucose, weight, height, and etc. (these variables are the predictors). Then the doctor decides whether or not the patient has diabetes based on the information he collected. This is an example of classification. The goal is prediction.

Example 2: Suppose a medical institute wants to investigate which variables indicate diabetes (whether or not a patient has diabetes is the response). The institute collects data from a sample of random patients of both diabetes and non-diabetes. Several variables of each patient are collected such as blood glucose, weight, height, gender and etc. By statistical analysis, the institute decide which variables are necessary or significant in order to predict if a patient has diabetes (these variables are the predictors). This is an example of classification. The goal is inference.

Example 3: Suppose we are given Capital Bikeshare data from the second quarter of 2016. Each row of the data set is a bike that is out for service. Columns in the data set are variables such as duration of the trip, bike identification number, date, and etc. (these variables are the predictors). We would like to predict whether or not a bike is in repair shop instead of used by customers based on the given variables. So the response is whether or not a bike is in repair shop. This is an example of classification. The goal is prediction.

b)

Example 1: Suppose we want to predict a student's GRE score based on his or her GPA and hours spent for preparing GRE test that we know. We also know there is a function that takes his or her GPA and hours spent for studying GRE as input and output GRE score. We plug in this function and calculate the student score. This is an example of regression. The response is a student's GRE score. The predictors are his or her GPA and hours spent for preparing GRE test. The goal is prediction.

Example 2: Suppose we want to generate a function to predict a student's GRE score based on two variables- his or her GPA and hours spent for preparing GRE test. We collect a sample of random students who have taken GRE test along with their GPA and hours spent for preparing GRE test. Then, we fit a model with multiple linear regression to see if the variables are significant. This is an example of regression. The response is a student's GRE score. The predictors are his or her GPA and hours spent for preparing GRE test. The goal is inference.

Example 3: Suppose we want to find out whether or not hours of studying per week can predict a student's GPA at Georgetown University. We collected a sample of random students along with their hours of studying per week and GPA at Georgetown University. Then, we fit a model with linear regression to see if hours of studying per week is a significant predictor of GPA. This is an example of regression. The response is a student's GPA at Georgetown University. The predictors are his or her hours spent for studying per week. The goal is inference.

c)

Example 1: Suppose we want to cluster tweets on twitter based on their sentiments of six so-called "basic" emotions (i.e. anger, disgust, happiness, sadness, surprise, and fear). Each tweet is assigned a score for each of the emotion which is the percentage of words appear in the emotional category. Then, we conduct a K-means cluster analysis to determine if the tweets fell into emotion patterns that could be used for grouping purposes.

Example 2: Suppose we want to group students at Georgetown University based on their hobbies, major, years of study. In this case, cluster analyses might be useful.

Example 3: Suppose we want to group species based on species' characteristics such as bipedalism or reptile, amphibian or not, and etc. In this case, cluster analyses might be useful.

Question 7

a)

$$d1 = \sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = 3$$
$$d2 = \sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = 2$$
$$d3 = \sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = 3.162278$$
$$d4 = \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = 2.236068$$
$$d5 = \sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = 1.414214$$
$$d6 = \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = 1.732051$$

b)

Let x0 be the point that X1=X2=X3=0. When k = 1, the 1 point that closest to x0 is observation number 5 which is green. Pr(Y =Green|X=x0) =1. Thus, the largest probability is green. So the predicton is green.

c)

Let x0 be the point that X1=X2=X3=0. When k = 3, the 3 points that closest to x0 are observatino number 2, 5, and 6. There are two reds and one green. So, Pr(Y =Green|X=x0) =1/3, Pr(Y =Red|X=x0) =2/3. Thus, the largest probability is red. So the prediction is red.

d)

As K increase, the boundary tends to be linear, thus, the best value of K is small in this problem (when the Bayes decision boundary is highly non- linear).

Question 8

a)

```r
#install.packages('ISLR')
library('ISLR')
data(College)
setwd("/Users/jingshisun/Desktop/Spring 2017 Courses/ANLY 512/HW1")
college <- read.csv("College.csv")
```

b)

```r
rownames(college)=college[,1]
fix(college)
college=college[,-1]
fix(college)
```

c)

i)

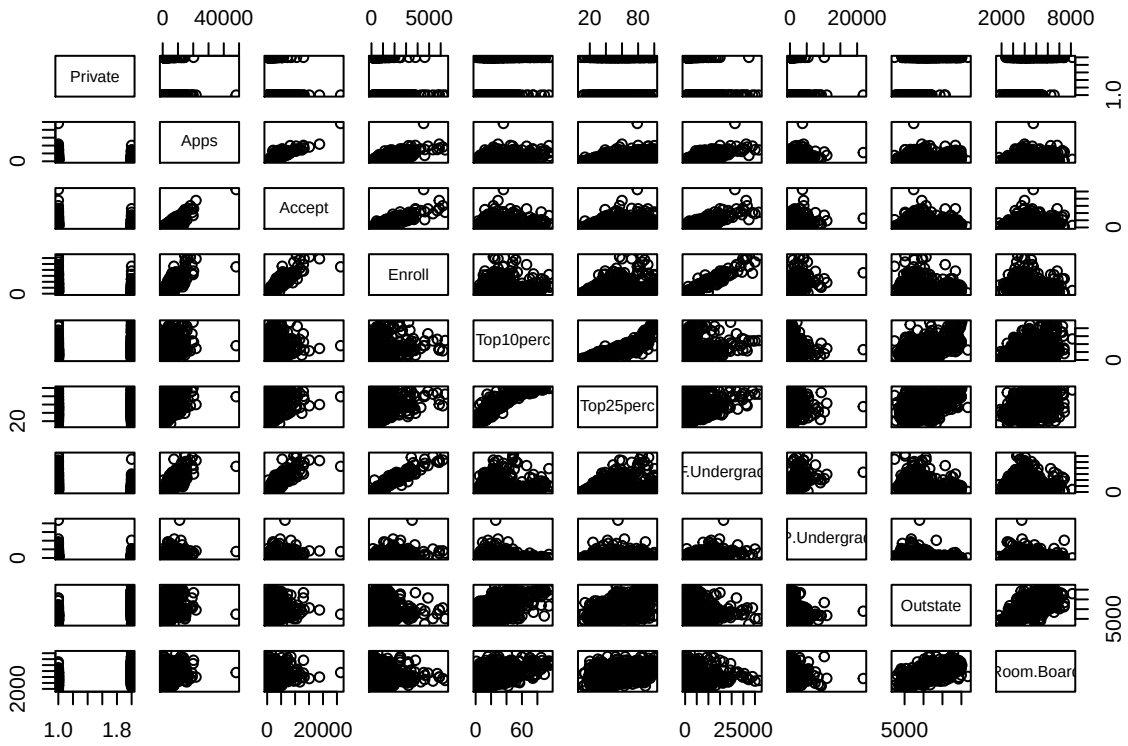
```r
summary(college)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
```

```
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

ii)

```
pairs(college[,1:10])
```



iii)

```
plot(college$Private, college$Outstate, xlab = "Whether or not it is a private university",
     ylab ="Out-of-state Tuition", main = "Side-by-side Boxplots of Outstate Versus Private")
```

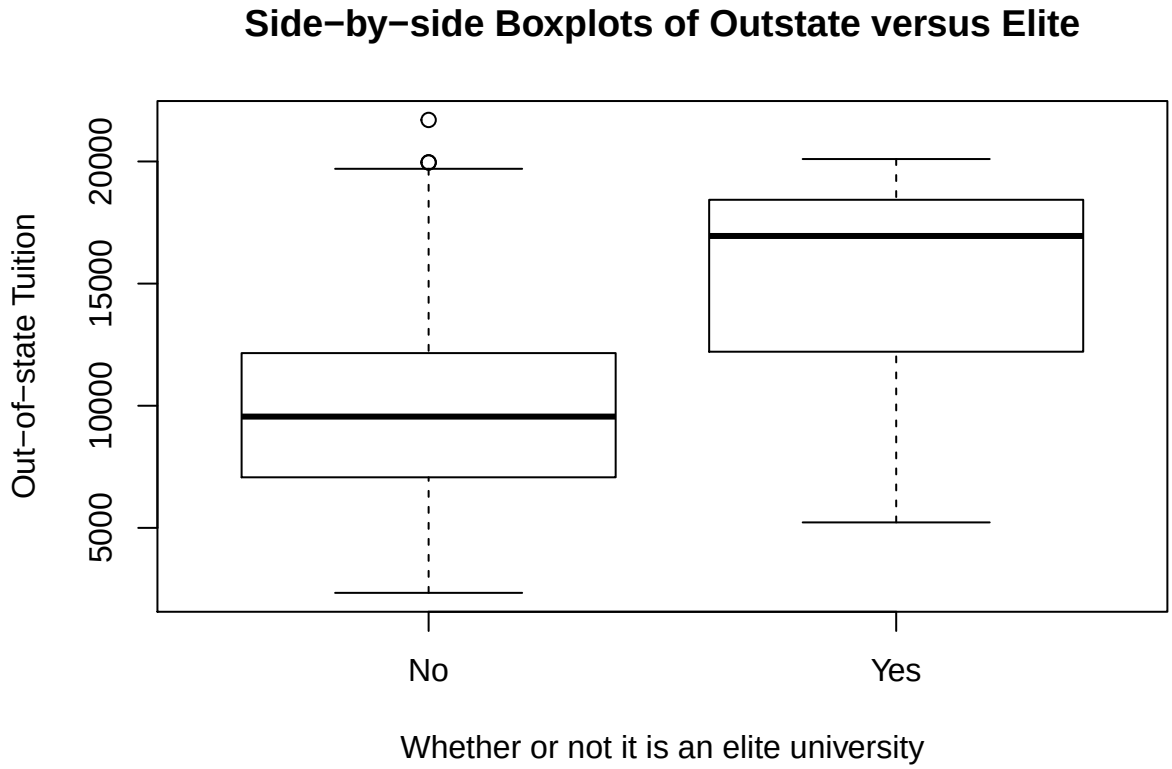
iv)

```
Elite=rep("No",nrow(college))
Elite[college$Top10perc >50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)
summary(college)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
```

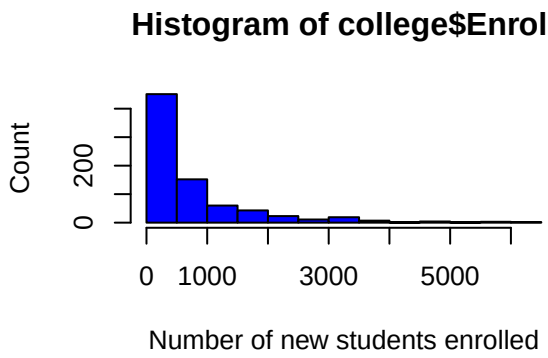
```
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate      Elite
##  Min.   : 10.00   No :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

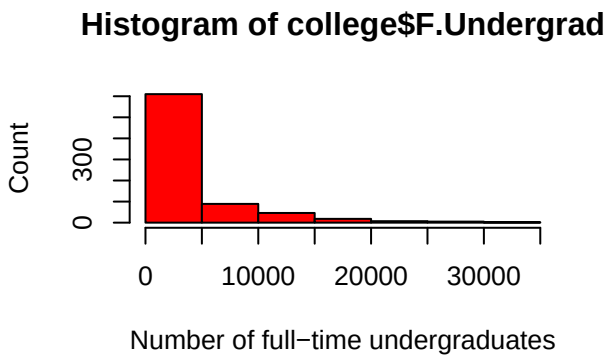
```r
plot(college$Elite,college$Outstate, xlab="Whether or not it is an elite university",
     ylab="Out-of-state Tuition", main="Side-by-side Boxplots of Outstate versus Elite")
```

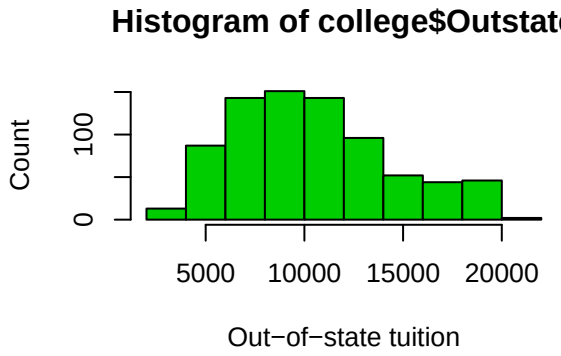


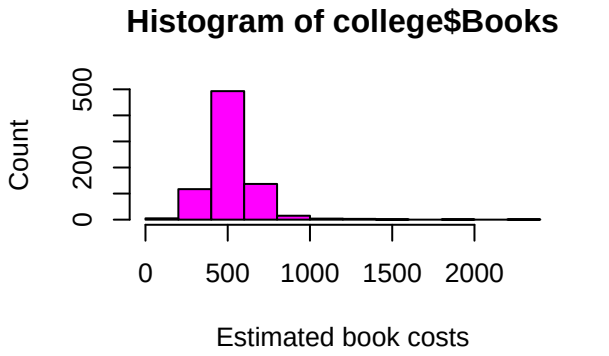
v)

```r
par(mfrow = c(2,2))
hist(college$Enroll, col = 12, xlab = "Number of new students enrolled", ylab = "Count")
hist(college$F.Undergrad, col = 10, xlab = "Number of full-time undergraduates", ylab = "Count")
hist(college$Outstate, col = 3, xlab = "Out-of-state tuition", ylab = "Count")
hist(college$Books, col = 6, xlab="Estimated book costs", ylab = "Count")
```









vi)

```
summary(college$Enroll) # summary of number of new students enrolled
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      35     242     434     780     902    6392
```

According to the summary and histogram, most (more than 3/4) of the colleges have new students entrolled under 1000. Although, there is a university has 6392 new students enrolled, the median is 434.

```
summary(college$F.Undergrad) # summary of number of full-time undergraduates
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     139     992    1707    3700    4005   31640
```

According to the summary and histogram, most (more than 3/4) of the colleges have number of full-time undergraduates less than 5000. Although, thre is a university has 31640 full-time undergraduates, the median is 1707.

```
summary(college$Outstate) #summary of out-of-state tuition
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2340    7320    9990   10440   12920   21700
```

Most of colleges' out of state tuitions are between around 5000 USD and 15000 USD. The college with the highest tuition (21700 USD) is about 10 times of the lowerst (2340 USD). The median is 9990 USD.

```
summary(college$Books) #summary of estimated book costs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    96.0   470.0   500.0   549.4   600.0  2340.0
```

The median estimated book costs is about 500 USD which is close to the mean (549.4 USD). As the histogram shown, the distribution of estimated book costs is approximated symmetrical. The college with the highes estimated book cost (2340 USD, could be considered as an outlier) is about 24 time the college with the lowest estimated book cost (96 USD).

Q10

a)

```
library(MASS)
head(Boston)
```

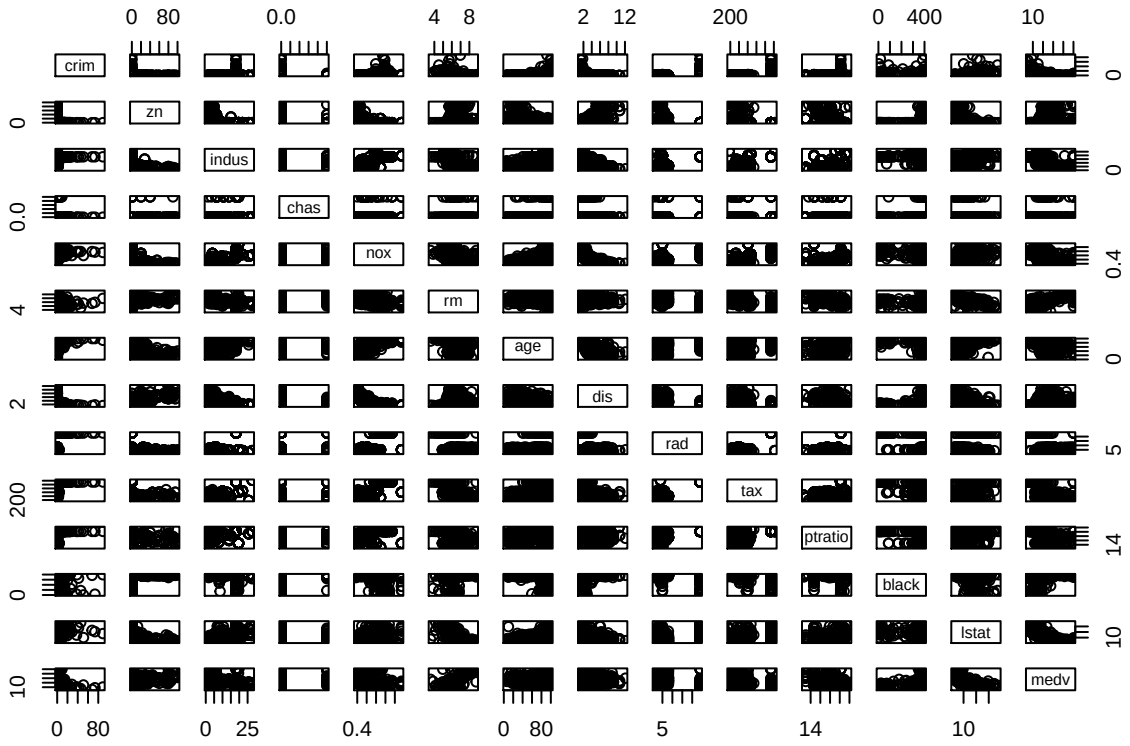
```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

```
?Boston
```

There are 506 rows and 14 columns in the Boston data set. Each row represents an observation of a suburb in Boston. Each column represents one attribute (such as per capita crime rate by town,average number of rooms per dwelling, pupil-teacher ratio by town and etc.) of a town in Boston.

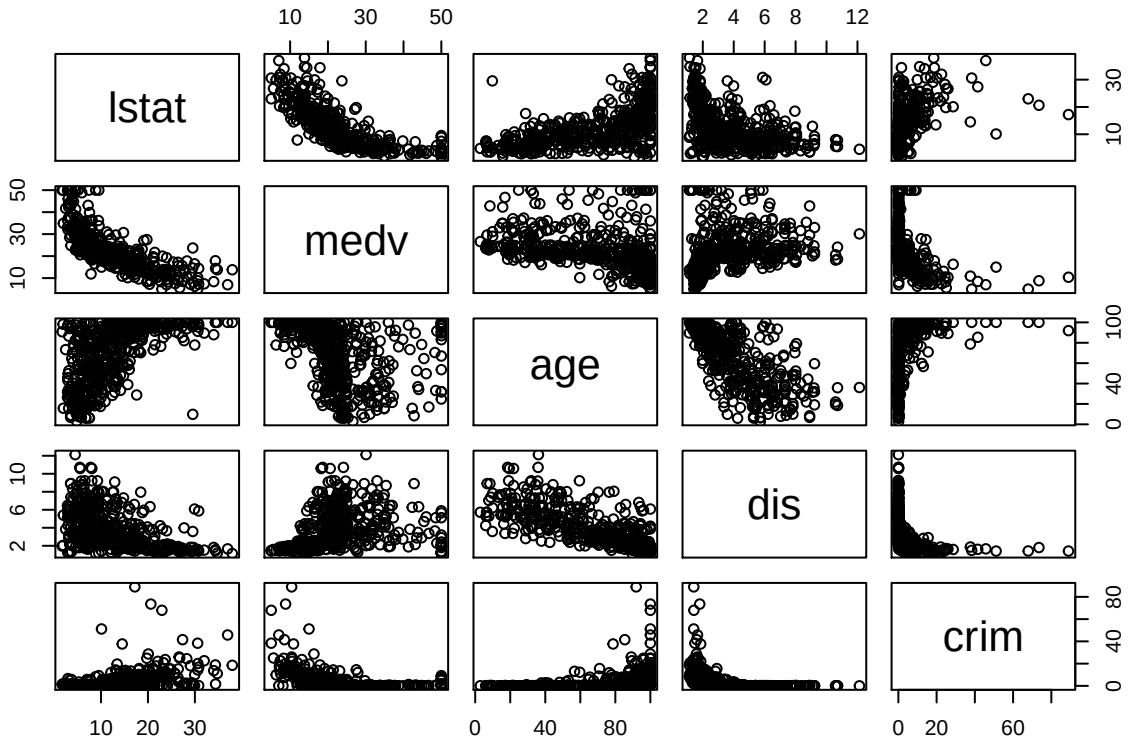
b)

```
pairs(Boston) #pairwise scatterplots of all predictors.
```



At first glance, the plots are so small if I do pairwise scatterplots of every predictors. Let's just focus a few predictors.

```
# make pairwise scatterplots of lstat, medv, age, dis and crim.
pairs(~lstat+medv+age+dis+crim, data=Boston )
```

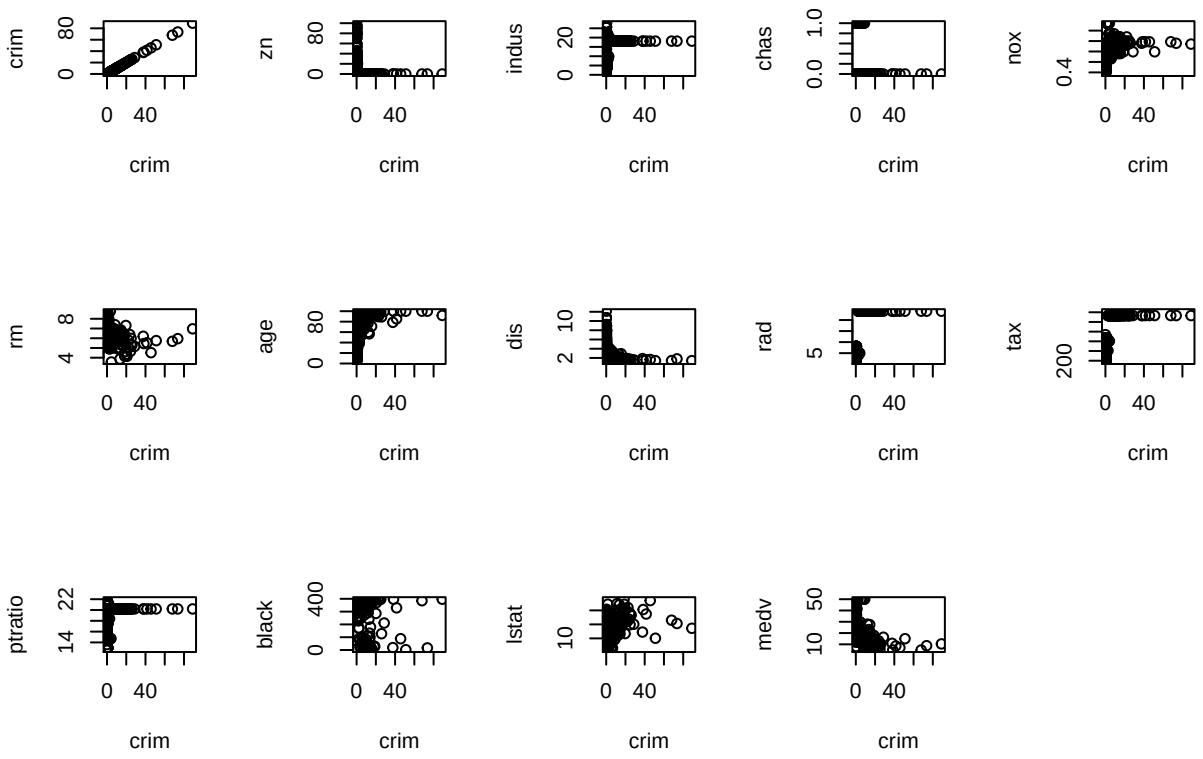


According to the scatter plots, there is a strong negative curved relationship between lstat (lower status of the population (percent)) and medv (median value of owner-occupied homes in $1000s). There is also a moderate

negative linear relationship between age (proportion of owner-occupied units built prior to 1940) and dis (weighted mean of distances to five Boston employment centres). There is also a moderate positive curved relationship between age (proportion of owner-occupied units built prior to 1940) and crim (per capita crime rate by town). Additionally, there is a strong negative curved relationship between dis (weighted mean of distances to five Boston employment centres) and crim (per capita crime rate by town).

c)

```
# divide the print window into 15 (3 x 5) regions.
par(mfrow = c(3,5))
# for each predictors, add a scatter plot of the predictor and crim
# into the window.
for(variable in names(Boston)){
  plot(Boston$crim, Boston[, variable], xlab="crim", ylab = variable)
}
```



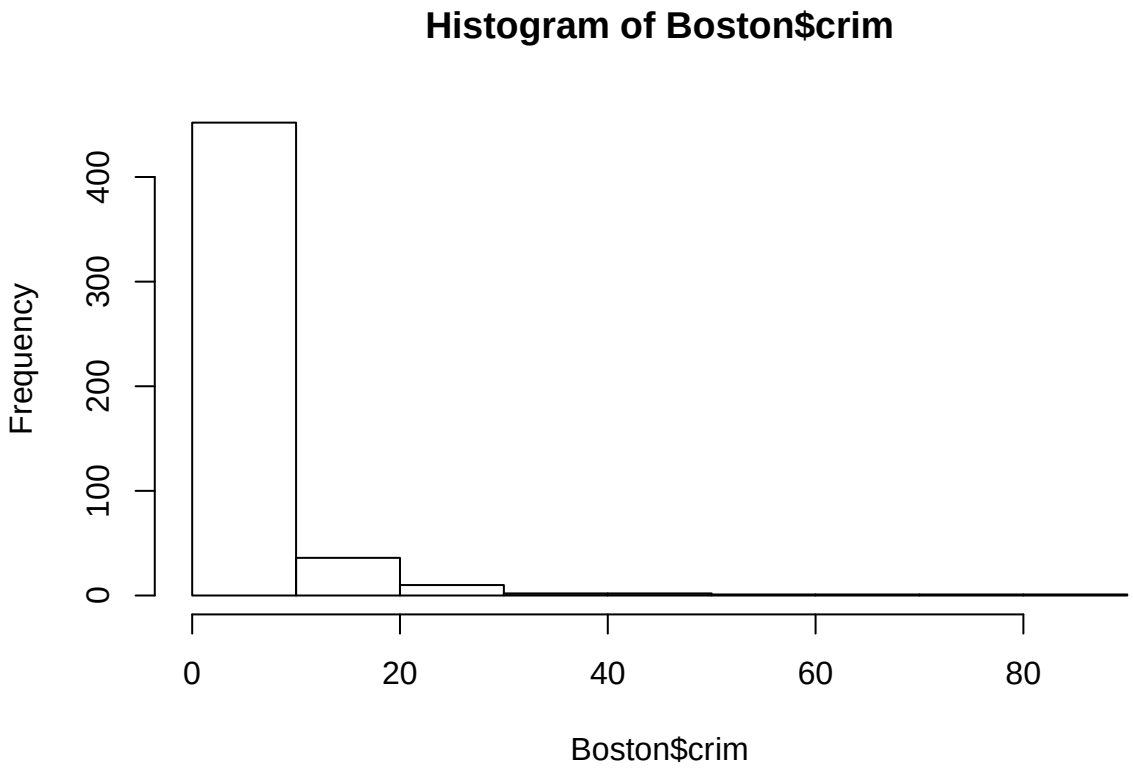
As the output shown, there is a moderate positive curved relationship between crim (per capita crime rate by town) and age (proportion of owner-occupied units built prior to 1940). There is a strong negative curved relationship between crim (per capita crime rate by town) and dis (weighted mean of distances to five Boston employment centres). Moreover, there is a moderate negative curved relationship between crim (per capita crime rate by town) and medv (median value of owner-occupied homes in $1000s).

d)

```
# set the print window back into 1 region.
par(mfrow = c(1,1))
summary(Boston$crim)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00632   0.08204   0.25650   3.61400   3.67700 88.98000
```

```
hist(Boston$crim)
```




```
#count number of rows/suburbs that have crime rate above 20.
nrow(Boston[Boston$crim > 20, ])
```

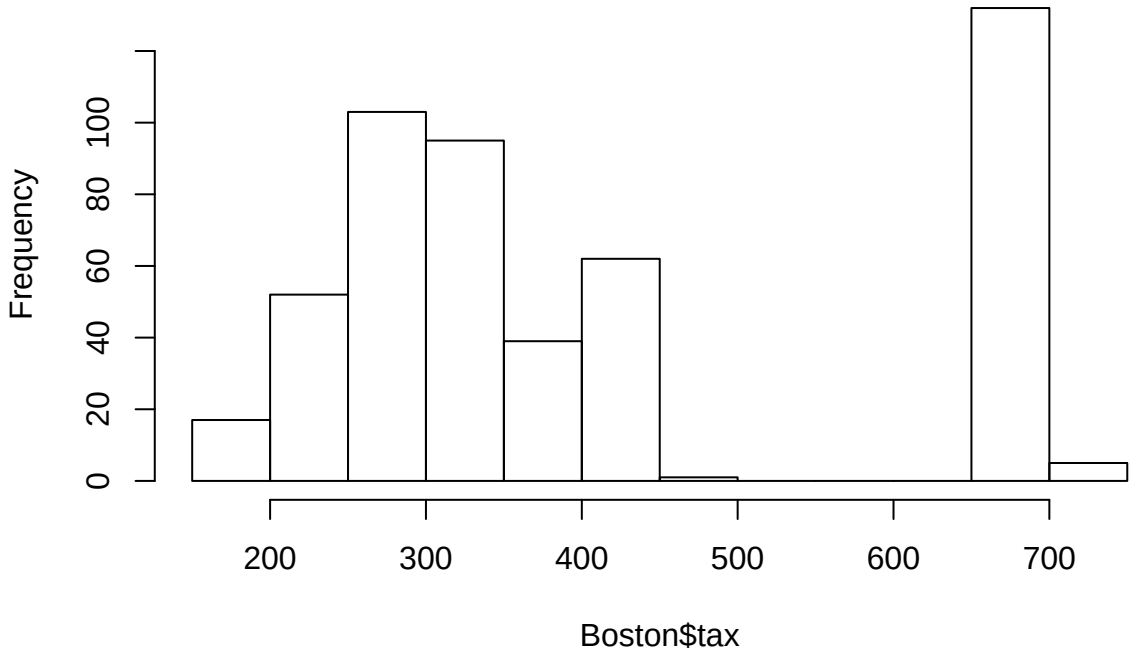
```
## [1] 18
```

As you can see from the output, most of the suburbs have crime rate under 20. However, there are some extreme cases. There are 18 suburbs have crime rate above 20 and one has a extremely high crime ratio of 88.98. The range is between 0.00632 and 88.98.

```
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
hist(Boston$tax)
```




```
#count number of rows/suburbs that have tax rate above 700.
nrow(Boston[Boston$tax > 700, ])
```

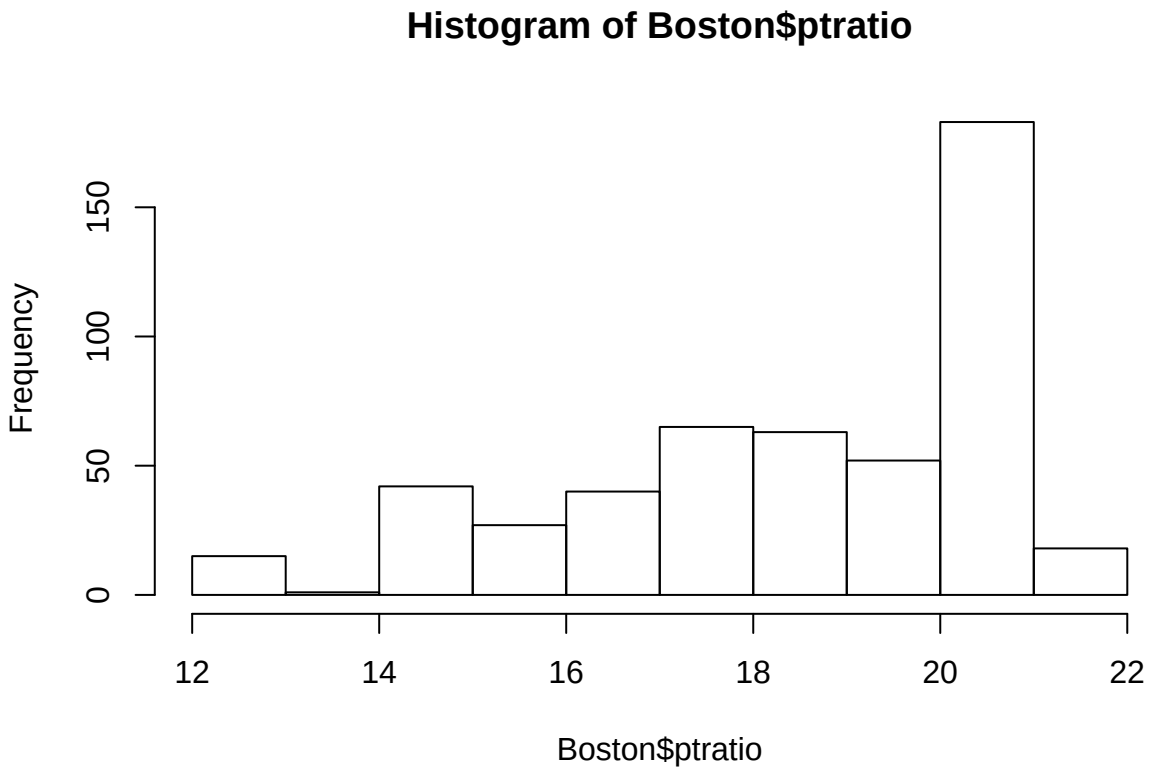
```
## [1] 5
```

The range of tax rate is betwwen 187 and 711. There seems to be two clusters. The lower cluster ranges between 187 and 500. The higher cluster ranges above 600. There are 5 particularly high cases (higher than 700).

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

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

```
hist(Boston$ptratio)
```




```
#count number of rows/suburbs that have pupil-teacher ratio above 21.
nrow(Boston[Boston$ptratio > 21, ])
```

```
## [1] 18
```

The pupil-teacher ratios range from 12.6 to 22. As you can see from the histogram, most of them are between 14 and 21. There are 18 particularly high cases (higher than 21) of pupil-teacher ratio.

e)

```
#count number of rows/suburbs that bound the Charles river.
nrow(Boston[Boston$chas == 1, ])
```

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

f)

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

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

According to the summary, median pupil-teacher ratio among the towns in this data set is 19.05.

g)

```
# Find suburbs of Boston that has lowest
# median value of owner-occupied homes.
Boston[Boston$medv == min(Boston$medv),]
```

```
##        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    20.2 396.90
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97
##     lstat medv
## 399 30.59    5
## 406 22.98    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
```

According to the output, there are 2 surburbs that have lowest median value of owner-occupied homes. They are observation number 399 and 406. As I noticed, both two suburbs have high crime rates (above 30). One is 38.3518, the other is 67.9208 which is extremely high. Both of the suburbs have 0 zn (proportion of residential land zoned for lots over 25,000 sq.ft.) which is lower than the mean value of zn (11.36%).

Both of the suburbs are not bounded by Charles River. Both suburbs have higher than median nitrogen oxides concentration. Both of the suburbs have 100 percent of owner-occupied units build prior to 1940. Both suburbs have low weighted mean of distances to five Boston employment centres. Both suburbs have the highest index of accessibility to radial highways. Both suburbs have higher than median pupil-teacher ratio. Suburbs number 399 have the highest proportion of blacks.

What's more, both suburbs higher than median (and mean) lower status of the population. Besides these,

the rest of the predictors range between 1st and 3rd quarter of the total.

h)

```
#Count the number of rows/suburbs average more than 7 rooms per dwelling
nrow(Boston[Boston$rm > 7, ])
```

```
## [1] 64
```

```
#Count the number of rows/suburbs average more than 8 rooms per dwelling
nrow(Boston[Boston$rm > 8, ])
```

```
## [1] 13
```

As you can see from the output, 64 suburbs average more than seven rooms per dwelling, 13 suburbs average more than seven rooms per dwelling.

```
summary(Boston) # summary of all suburbs
```

```
##       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
```

```
# summary of suburbs average more than 8 rooms per dwelling.
summary(Boston[Boston$rm > 8, ])
```

```
##       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
```

From the above output, we can see that the suburbs that average more than eight rooms per dwelling have lower average crime rates, higher average median value of owner-occupied homes in $1000s, lower percent in lower status of the population, lower tax rates, and have a higerer proportion that are bounded by Charles River than the average of the suburbs.

The Shiny Question

a)

Simulation	Residual SS	Highest Order Coefficient
1	63.88	-4.5
2	104.03	-4.5
3	77.96	-3.5
4	80.13	-3.5
5	124.26	-4.3
6	92.68	-4.1
7	73.55	-4.7
8	79.8	-3.5
9	77.03	-3.9
10	79.96	-4.2

```
1/10*(63.88+104.03+77.96+80.13+124.26+92.68+73.55+79.8+77.03+79.96) # The average of residual SS
```

```
## [1] 85.328
```

```
#The approximate range of the highest order coefficient
range(c(-4.5,-4.5,-3.5,-3.5,-4.3,-4.1,-4.7,-3.5,-3.9,-4.2))
```

```
## [1] -4.7 -3.5
```

According to the output, the average residual SS is 85.328 and the approximate range of the highest order coefficient is from -4.7 to -3.5.

b)

Simulation	Residual SS	Coefficient #2
1	44.2	14.9
2	39.58	-31.5
3	28.79	-7.4
4	32.17	0.2
5	24.7	-26
6	38.94	12.6
7	23.95	-16.5
8	26.13	17.5
9	40.4	-14.9
10	35.46	-26

```
1/10*(44.2+39.58+28.79+32.17+24.7+38.94+23.95+26.13+40.4+35.46) # The average of residual SS
```

```
## [1] 33.432
```

```
#The approximate range of the coefficient #2
range(c(14.9,-31.5,-7.4,0.2,-26,12.6,-16.5,17.5,-14.9,-26))
```

```
## [1] -31.5  17.5
```

According to the output, the average residual SS is 33.432 and the approximate range of the coefficient #2 is from -31.5 to 17.5.

c)

After many times of simulations, it becomes clear to me that the coefficient has the largest variation is coefficient #6. Thus, I will record 10 simulations with residual ss and coefficient #6 in below.

Simulation	Residual SS	Coefficient #6
1	4.53	3.1×10^5
2	8.5	8.5×10^5
3	15.77	-1.75×10^5
4	4.9	1.3×10^6
5	14.19	8×10^5
6	2.09	-3.6×10^5
7	6.74	-7.8×10^5
8	7.14	-5.2×10^5
9	0.06	-9×10^5
10	2.63	6.5×10^5

```
1/10*(4.53+8.5+15.77+4.9+14.19+2.09+6.74+7.14+0.06+2.63) # The average of residual SS
```

```
## [1] 6.655
```

```
#The approximate range of the coefficient #6
range(c(3.1*10^5,8.5*10^5,-1.75*10^5,1.3*10^6,8*10^5,-3.6*10^5,-7.8*10^5,-5.2*10^5,-9*10^5,6.5*10^5))
```

```
## [1] -900000 1300000
```

According to the output, the average residual SS is 6.655 and the approximate range of the coefficient #6 is from -9×10^5 to 1.3×10^6.

e)

As the model complexity increase, the variance decrease (as you can see from previous questions, the residual SS decrease when model complexity increase from 1 to 4, then to 16), but the bias increase (as you can see from previous questions, the range of coefficient with the largest range increase dramatically).

f)

When model complexity = 2, I typically obtain a curve which is most similar to the unknown curve that is to be estimated, because the curve looks very close to a quadratic equation.