STAT 1361 - Homework 1

Gordon Lu

2/4/2021

Exercise 1:

1a) A flexible statistical learning method would perform better, since the sample size would be large enough to fit more parameters. Additionally, a small number of predictors would limit the variance of the model.

1b) A flexible statistical learning method would perform worse, since the introduction of more parameters and a smaller sample size would increase the chance of overfitting. The model would try to fit to the small number of observations, and almost fit the data "too well".

1c) A flexible statistical learning method would perform better, as with more degrees of freedom, we would obtain a better fit.

1d) A flexible statistical learning method would perform worse, since the model would fit to the noice in the error terms, and variance would increase. Additionally, the chance of overfitting increases.

Exercise 2:

2a)

This is a `regression` problem, since the response is continuous. This problem concerns `inference`, we want to know which factors impact CEO salary, rather than how much does CEO salary increase/decrease.

Here, `n` is the top 500 firms in the US. On the other hand, The parameters, `p` are number of employees, industry, and the CEO salary.

2b)

Since the response is categorical, in particular, a binary response, this is a `classification` problem, we want to know whether launching a new product will be a success or a failure, rather than what factors influence whether a new product will be a success or failure. This problem is one concerning `prediction`.

Here, `n` is the 20 similar products that were previously launched. The parameters, `p` are the price charged for the product, marketing budget, competition price, and the ten other variables.

2c)

This is a `regression` problem, since the response is continuous. This problem concerns `prediction`, we want to know how USD/Euro exchange rate is related to weekly changes in the world stock markets, rather than knowing what factors influence the exchange rate.

Here, `n` is the 52 weeks of data collected for all of 2012. The parameters, `p` are the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Exercise 5:

An advantage of a very flexible model would be that they are highly data-driven. In the case of fitting a non-linear models, a flexible model will typically do a better job. Flexible models tend to have less bias, and higher variance.

One disadvantage with a very flexible model is that for regression and classsifcation, which require estimating a large number of parameters, the variance increases, and as a result the model may overfit.

A more flexible model would be preferred to a less flexible model if we are interested in prediction, rather than interpretation.

On the other hand, a less flexible model would be preferred if we are interested in inference and interpretation of the results.

Problem 3:

ISLR Conceptual Exercise 2 asks you (among other things) to determine whether each scenario is a classification or regression problem. For each scenario, now suppose we wanted to do the opposite. That is, if it was a classification problem, suppose we wanted to instead treat it as a regression problem and vice versa. What would need to change about the response and the way it's measured? In other words, think about how the descriptions could be restated in order to change the type of problem (regression or classification) being discussed.

3a)

In 2a), to go from a `regression` to `classification`, we can change the continuous response, to a binary response. In particular, we can consider encoding a 0 to indicate that the CEO salary did not change significantly, and a 1 to indicate that the CEO salary changed significantly. In this sense, our problem would be "Determining what factors influence whether a CEO's salary will change."

3b)

In 2b), to go from a `classification` to `regression`, we can change the categorical response, to a continuous response. In particular, we can consider the revenue that the new product brings in. In this sense, our problem would be "Predicting how successful launching a new product will be."

3c)

In 2c), to go from a `regression` to `classification`, we can change the continuous response, to a binary response. In particular, we can consider encoding a 0 to indicate that the model predicts there will be no % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets, a 1 to indicate that the model predicts there will be an increase in the % change in the USD/Euro exchange rate, and a -1 to indicate that the model predicts there will be a decrease in the % change in the USD/Euro exchange rate. In this sense, our problem would be "Predicting how the USD/Euro exchange rate will change."

Exercise 8:

8a)

```
college <- read.csv(file = 'C:/Users/gordo/Desktop/College.csv')
college
```

8b)

```
fix(college)
rownames(college) <- college[,1]


college <- college[,-1]
fix(college)
college
```

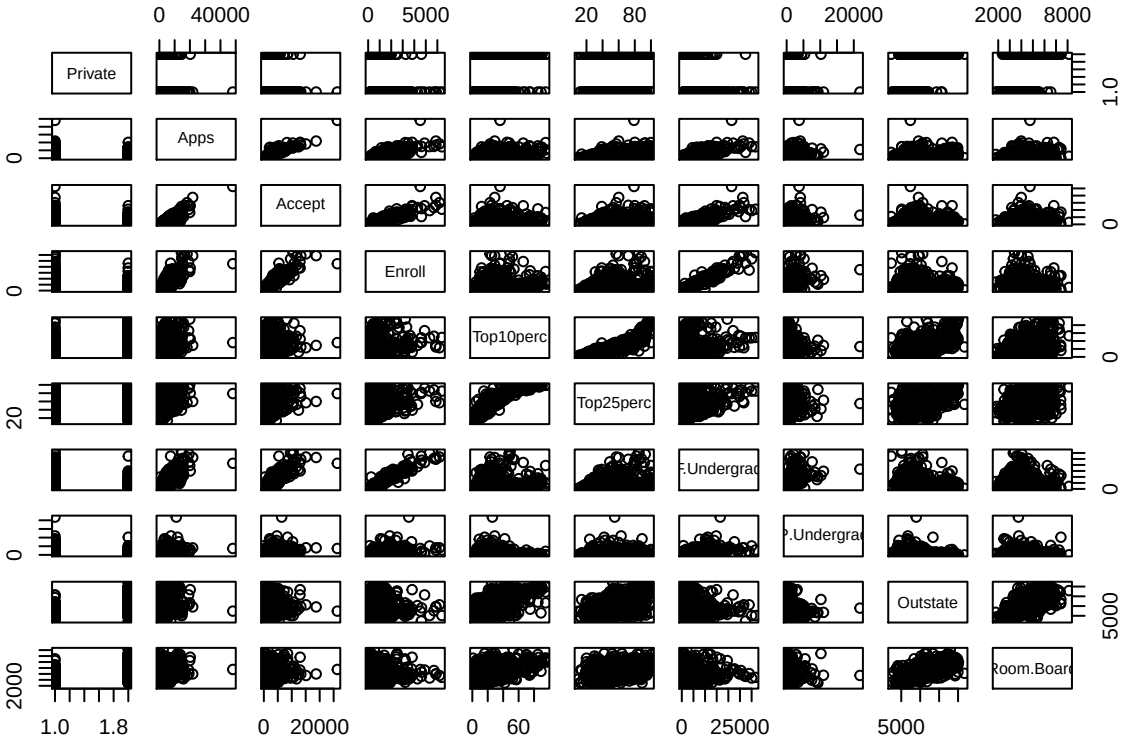
8c)

```
summary(college)
```

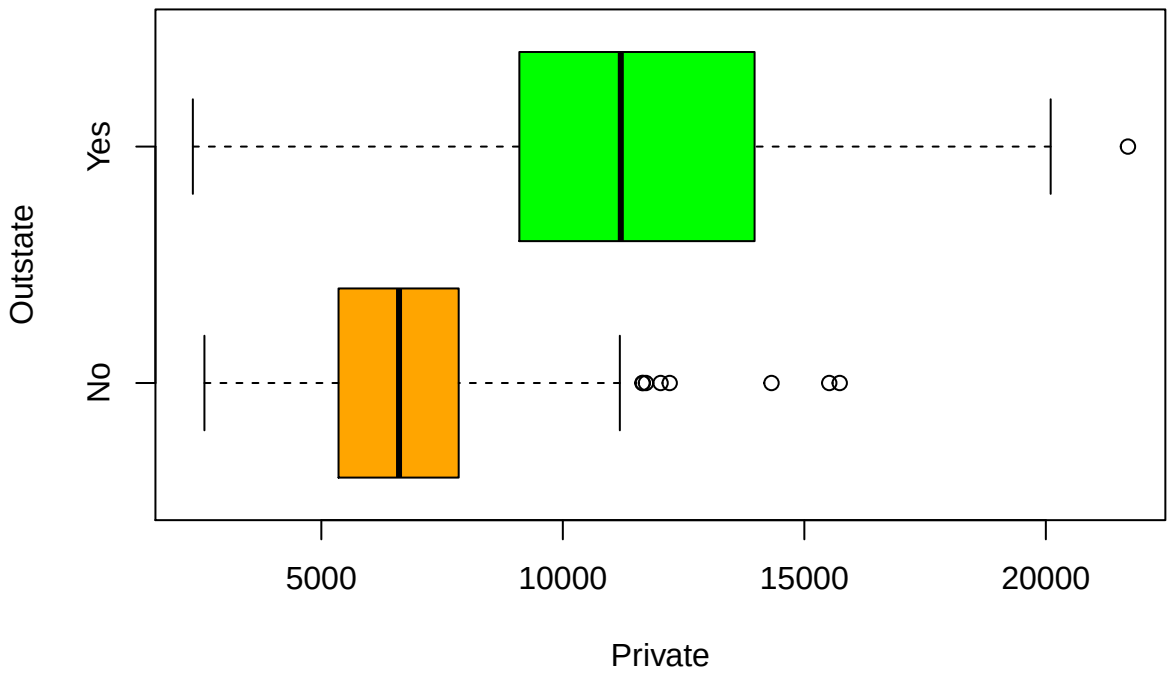
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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
```

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




```
boxplot(college$Outstate~college$Private, horizontal=TRUE, xlab = 'Private',
        ylab = 'Outstate', col = c("orange", "green"))
```



8d)

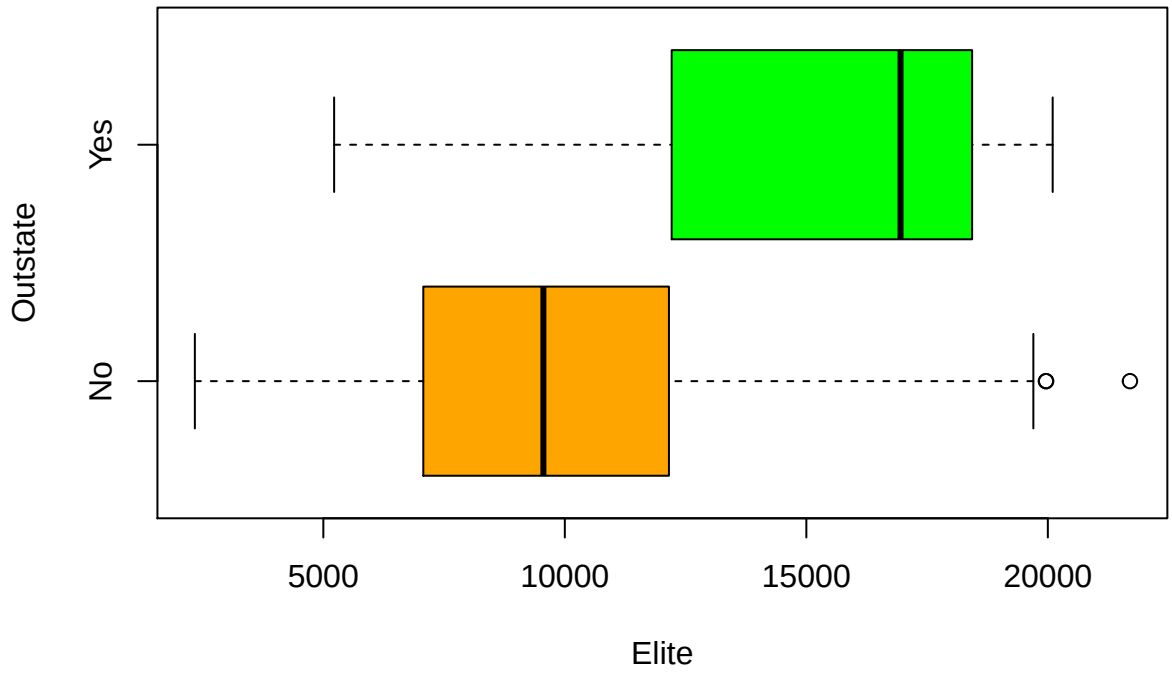
```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] = "Yes"
Elite <- as.factor(Elite)
college = data.frame(college, Elite)

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
```

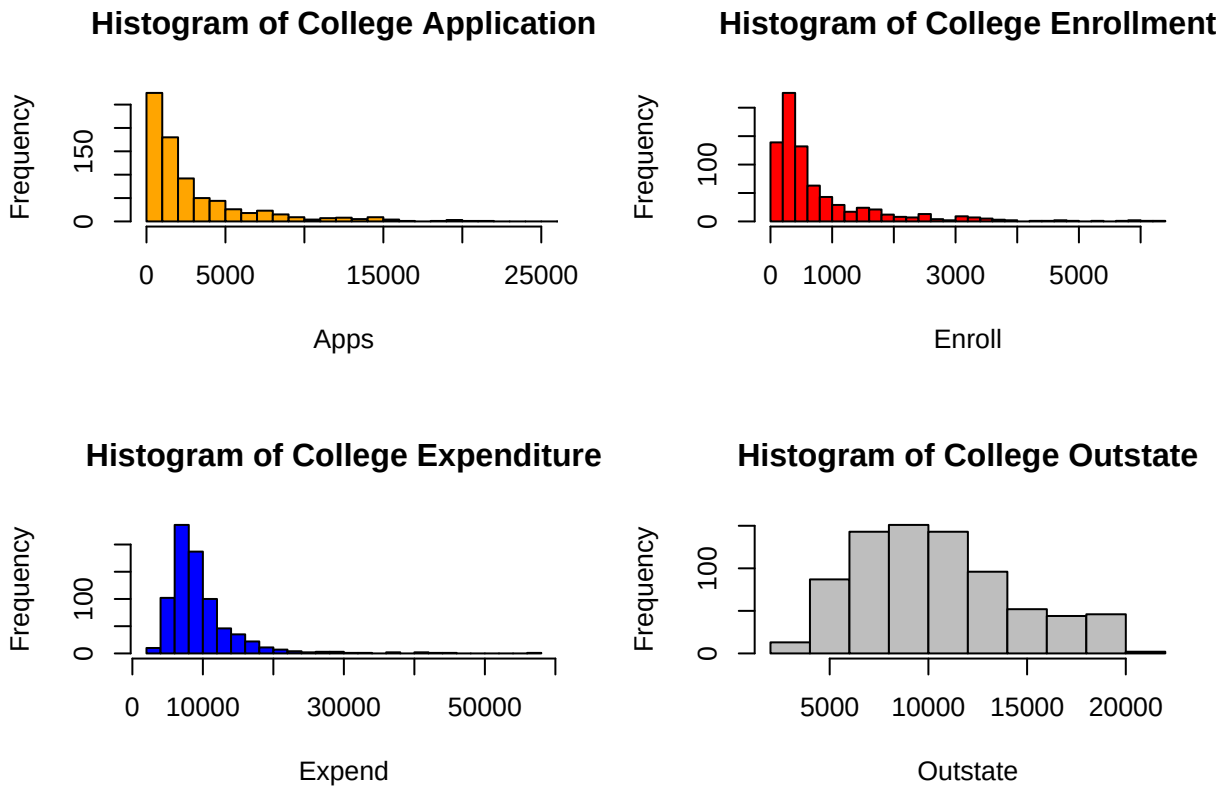
```
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   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
```

```
boxplot(college$Outstate~college$Elite, horizontal=TRUE, xlab = 'Elite',
        ylab = 'Outstate', col = c("orange", "green"))
```



8e)

```
par(mfrow=c(2,2))
hist(college$Apps, breaks=60, xlim=c(0,25000), main="Histogram of College Application",
     xlab = 'Apps', col = 'orange')
hist(college$Enroll, breaks=30, main="Histogram of College Enrollment",
     xlab = 'Enroll', col = 'red')
hist(college$Expend, breaks=30, main="Histogram of College Expenditure",
     xlab = 'Expend', col = 'blue')
hist(college$Outstate, main="Histogram of College Outstate",
     xlab = 'Outstate', col = 'gray')
```



8f)

```
summary(lm(college$Expen ~ college$Outstate), data = college) #print out result
```

```
##
## Call:
## lm(formula = college$Expen ~ college$Outstate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
##  -6131  -2022   -637   1027  39273
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      542.86969  385.92915   1.407     0.16
## college$Outstate   0.87325    0.03449  25.315   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 775 degrees of freedom
## Multiple R-squared:  0.4526, Adjusted R-squared:  0.4519
## F-statistic: 640.9 on 1 and 775 DF,  p-value: < 2.2e-16
```

```
summary(lm(college$Expen ~ college$Apps), data = college) #print out result
```

```
##
## Call:
## lm(formula = college$Expen ~ college$Apps)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14980  -2828   -997   1211  44656
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8608.8542   229.1441  37.570  < 2e-16 ***
## college$Apps    0.3503     0.0468   7.483 1.97e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5046 on 775 degrees of freedom
## Multiple R-squared:  0.06739,    Adjusted R-squared:  0.06618
## F-statistic:    56 on 1 and 775 DF,  p-value: 1.971e-13
```

The p-value is significantly less than 0.01. Thus, we can conclude that the there is a significant relationship between college expenditure and whether the student is out of state. Perhaps out of state students have to fork more money up for college than in-state students.

The p-value is significantly less than 0.01. Thus, we can conclude that the there is a significant relationship between college expenditure and whether the number of college applications. Perhaps the number of applications is correlated to the number of accepted students, resulting in an increase in cost for tuition.

Exercise 9:

9a)

```
auto <- read.csv(file = 'C:/Users/gordo/Desktop/Auto.csv')
auto
```

The quantitative predictors are: `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, and `year`.

The qualitative predictors are: `origin`, and `name`.

9b)

```
range(auto$mpg)
```

```
## [1]  9.0 46.6
```

```
range(auto$cylinders)
```

```
## [1] 3 8
```

```
range(auto$displacement)
```

```
## [1]  68 455
```

```
range(auto$horsepower)
```

```
## [1]  46 230
```

```
range(auto$weight)
```

```
## [1] 1613 5140
```

```
range(auto$acceleration)
```

```
## [1]  8.0 24.8
```

```
range(auto$year)
```

```
## [1] 70 82
```

9c)

```
sapply(auto[,1:7], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
sapply(auto[,1:7], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

9d)

```r
new_mat <- auto[,-(8:9)]   # drop origin, name
new_mat <- new_mat[-(10:85),]  # drop rows
sapply(new_mat, range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
```

```r
sapply(new_mat, mean)
```

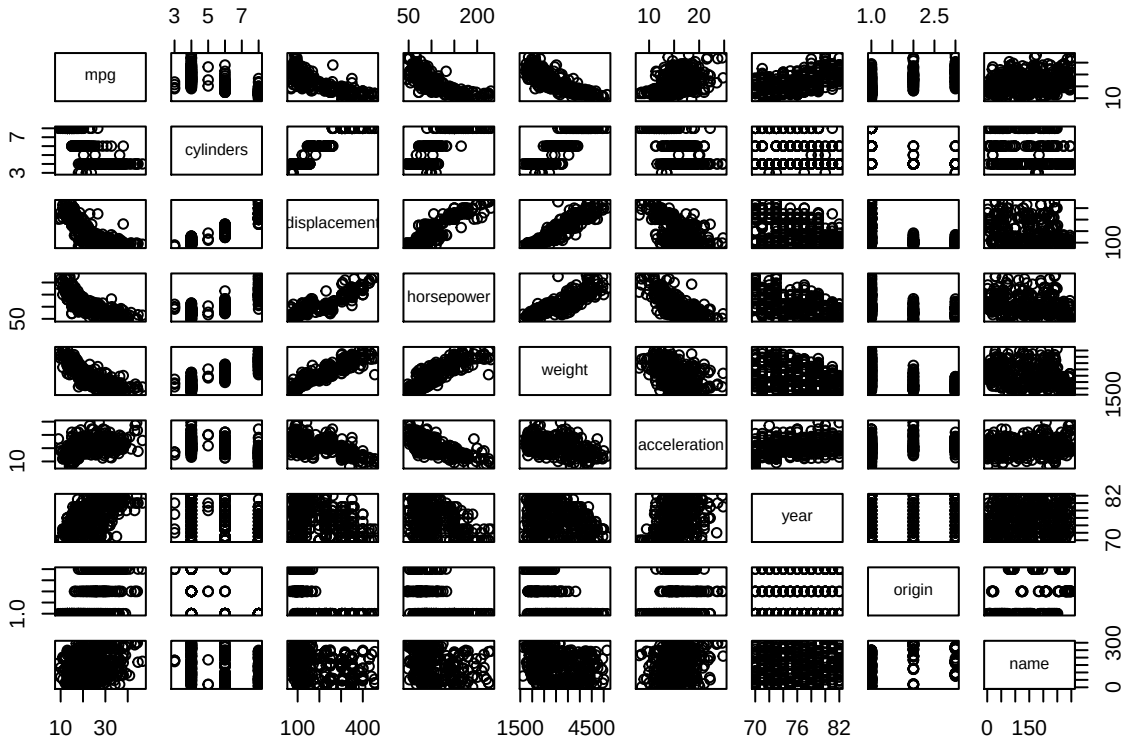
```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
##         year
##    77.145570
```

```r
sapply(new_mat, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

9e)

```r
pairs(auto)
```




```
cor(auto$mpg, auto$cylinders)
```

```
## [1] -0.7776175
```

```
cor(auto$mpg, auto$displacement)
```

```
## [1] -0.8051269
```

```
cor(auto$mpg, auto$horsepower)
```

```
## [1] -0.7784268
```

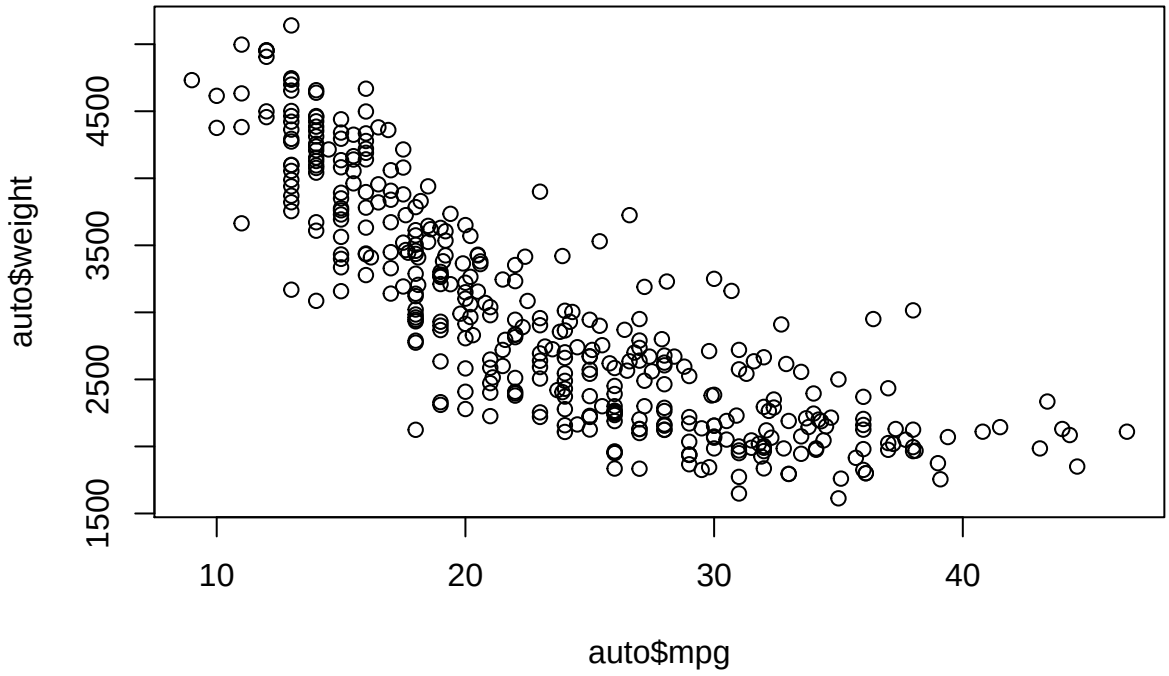
```
cor(auto$mpg, auto$year)
```

```
## [1] 0.580541
```

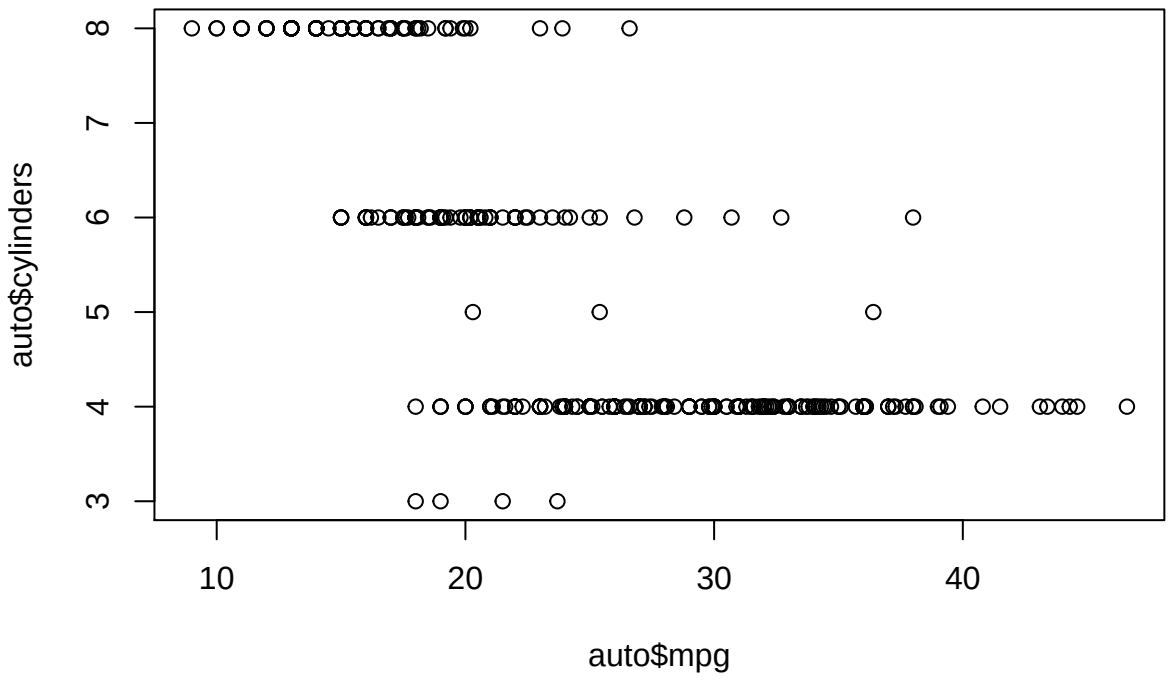
```
cor(auto$horsepower, auto$weight)
```

```
## [1] 0.8645377
```

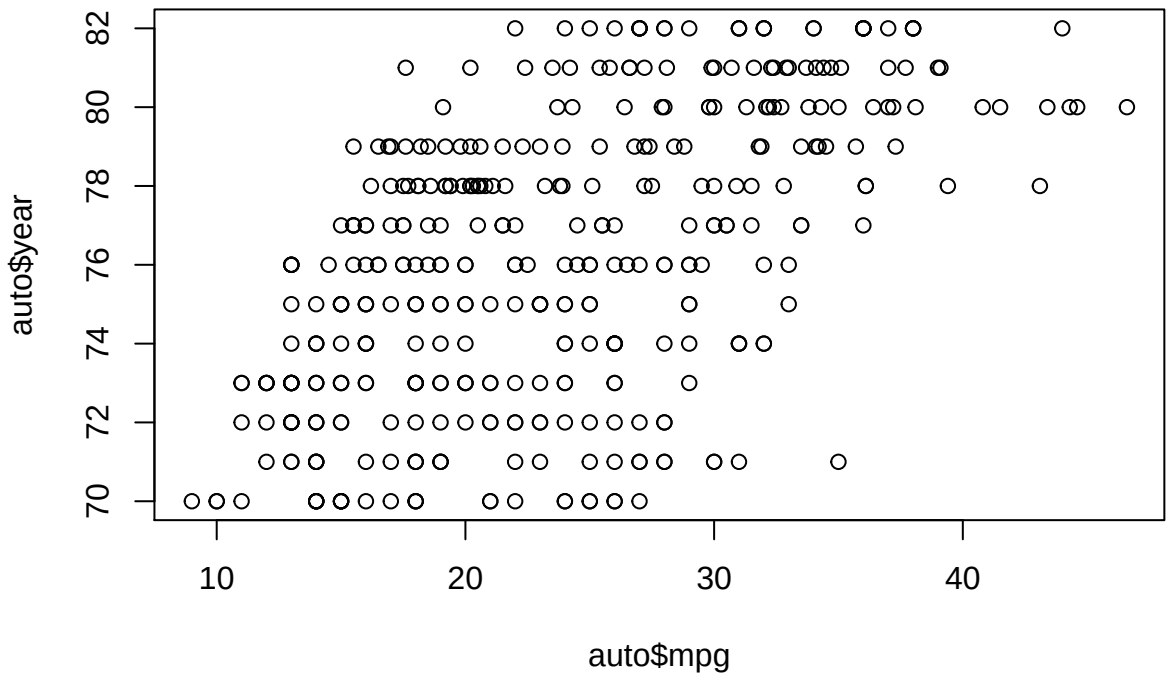
```r
plot(auto$mpg, auto$weight)
```




```r
plot(auto$mpg, auto$cylinders)
```




```
plot(auto$mpg, auto$year)
```



From the plots, we can see that `mpg` is negatively correlated with`cylinders`, `displacement`, `horsepower`, and `weight`.

We can also say that `horsepower` is negatively correlated with `weight`, and typically `mpg` increases with newer models.

9f)

Yes, the plots indicate that there are relationships between `mpg` and other variables in the Auto data set. We can also tell by looking at the respective correlations between `mpg` and other variables.

Exercise 10:

10a)

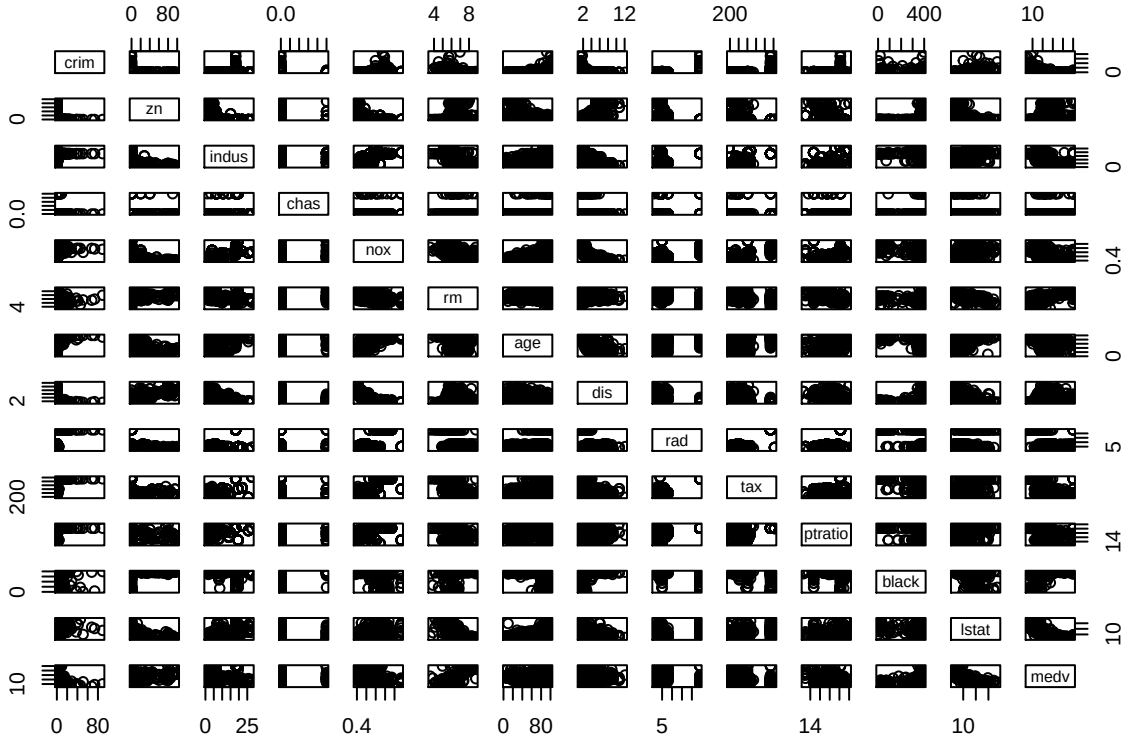
```
library(MASS)
dim(Boston)
```

```
## [1] 506  14
```

There are 506 rows and 14 columns. The rows represent an observation for a housing suburb in Boston. The columns represent the features.

10b)

```
pairs(Boston)
```



`crim` seems to be correlated with: `age`, `dis`, `rad`, `tax`, `ptratio`.

`zn` seems to be correlated with: `indus`, `nox`, `age`, `lstat`.

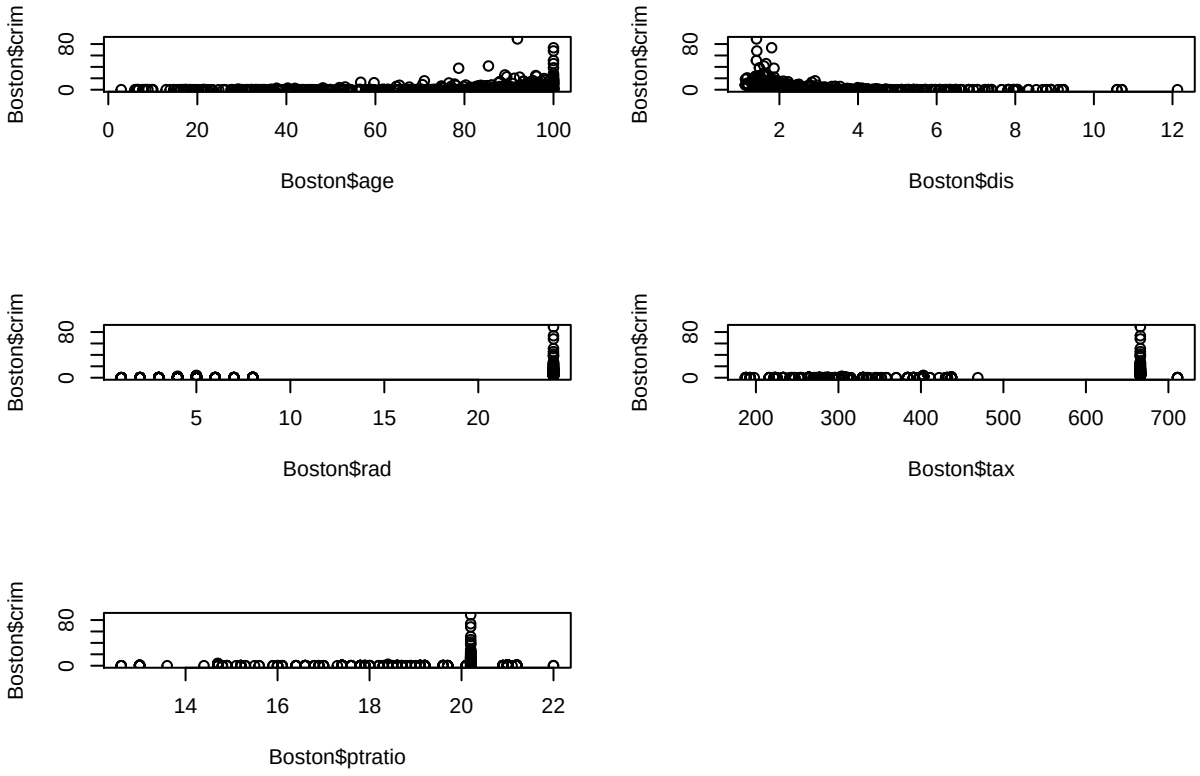
`indus` seems to be correlated with: `age`, `dis`.

`nox` seems to be correlated with: `lstat`.

`lstat` seems to be correlated with: `medv`.

10c)

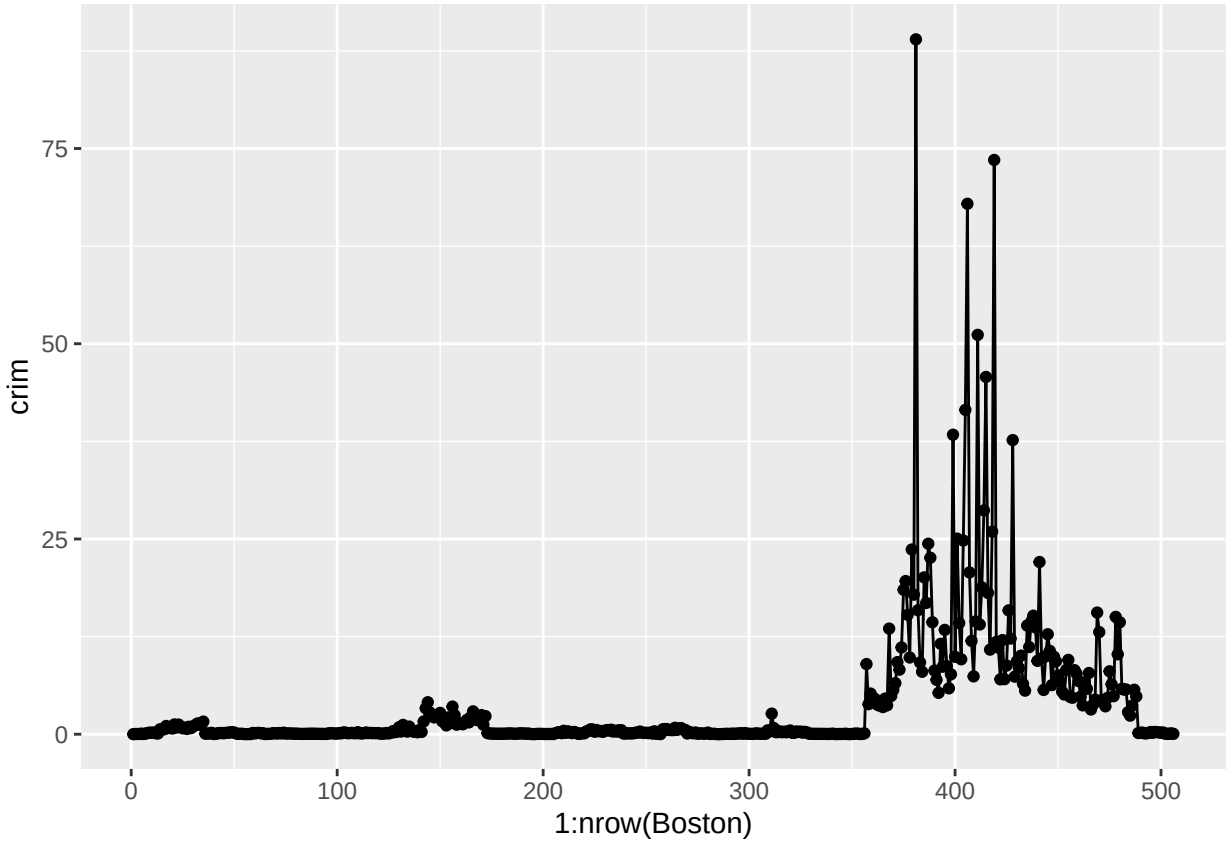
```
par(mfrow=c(3,2))
plot(Boston$age, Boston$crim)
plot(Boston$dis, Boston$crim)
plot(Boston$rad, Boston$crim)
plot(Boston$tax, Boston$crim)
plot(Boston$ptratio, Boston$crim)
```



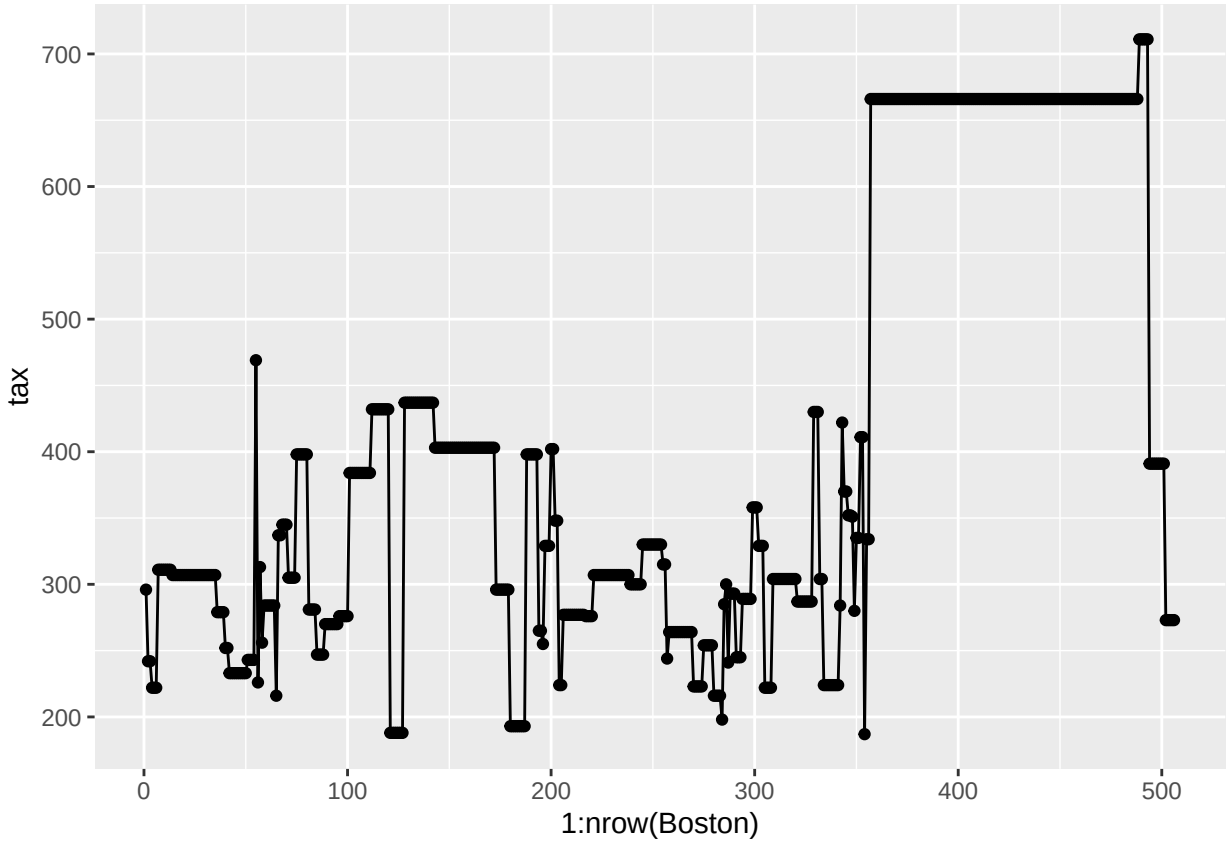
Seems that with older homes, there tends to be more crime. With homes closer to the work-area, there are also more crimes. The higher the index of accessibility seems, the higher the crime seems to be. The higher the tax rate, the more crime there is. Also the higher the pupil:teacher ratio, the more crime there seems to be.

10d)

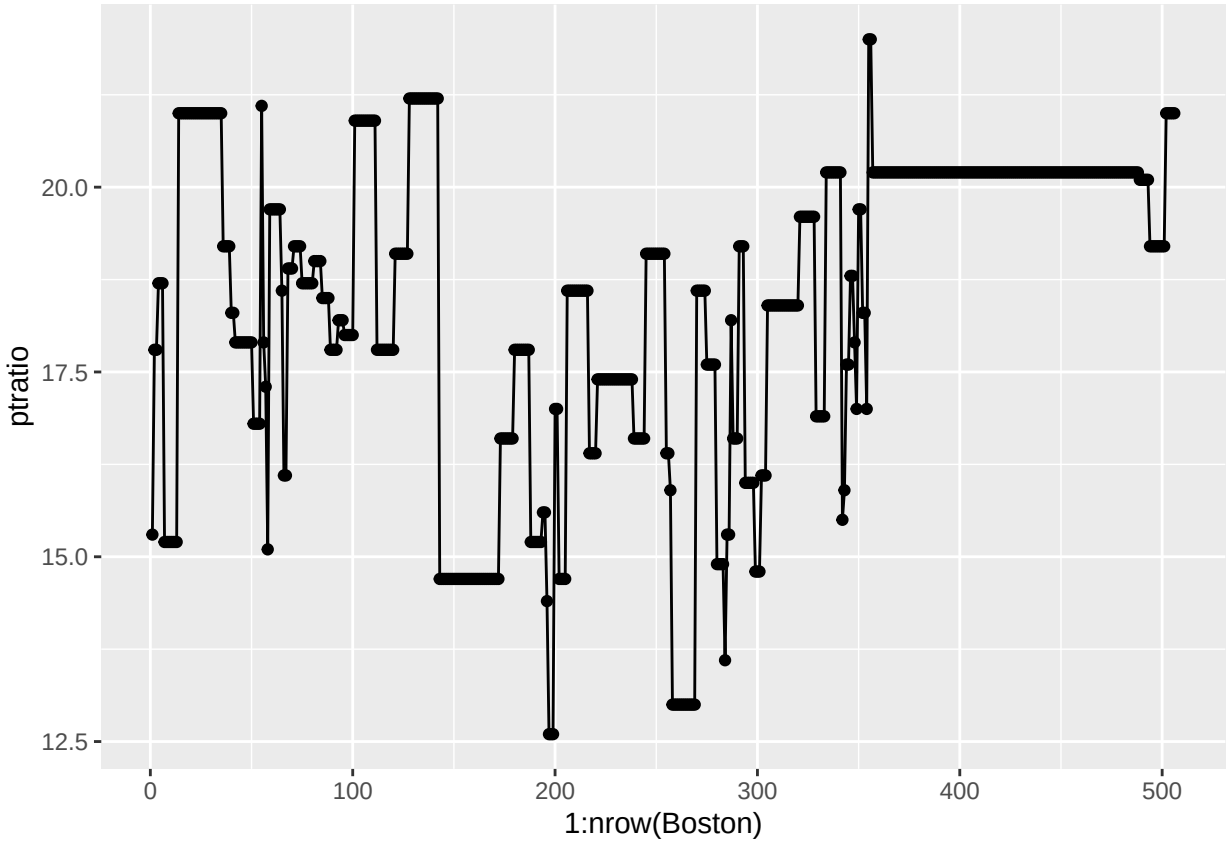
```
library(ggplot2)
g <- ggplot(Boston, aes(x=1:nrow(Boston), y=crim))
g + geom_point() + geom_line()
```




```
g <- ggplot(Boston, aes(x=1:nrow(Boston), y=tax))
g + geom_point() + geom_line()
```




```
g <- ggplot(Boston, aes(x=1:nrow(Boston), y=ptratio))
g + geom_point() + geom_line()
```



Looking at the plots, there are definitely outliers for `crim` and `tax`, but with the `ptratio` variable, there is no clear outlier. I would say that for crime rates and tax rates, there are definitely suburbs with outliers, and are unusualy, but with the Pupil-teacher tatios, there are no apparent outliers, but potentially some influential points.

10e)

```
nrow(subset(Boston, chas == 1))
```

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

There are `35` suburbs that bound the Charles river.

10f)

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

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

The median pupil-teacher ratio among towns in Boston is: `19.05`.

```
(min_median <- Boston[Boston$medv == min(Boston$medv),])
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
```

```
sapply(Boston, quantile)
```

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

`age` and `rad` are at max.

`crim`, `indus`, `nox`, `tax`, `ptratio`, `lstat` at or above the 3rd quartile.

`zn`, `rm`, `dis` are at min.

10h)

```
nrow(subset(Boston, rm > 7))
```

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

```
nrow(subset(Boston, rm > 8))
```

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

```
summary(subset(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
```

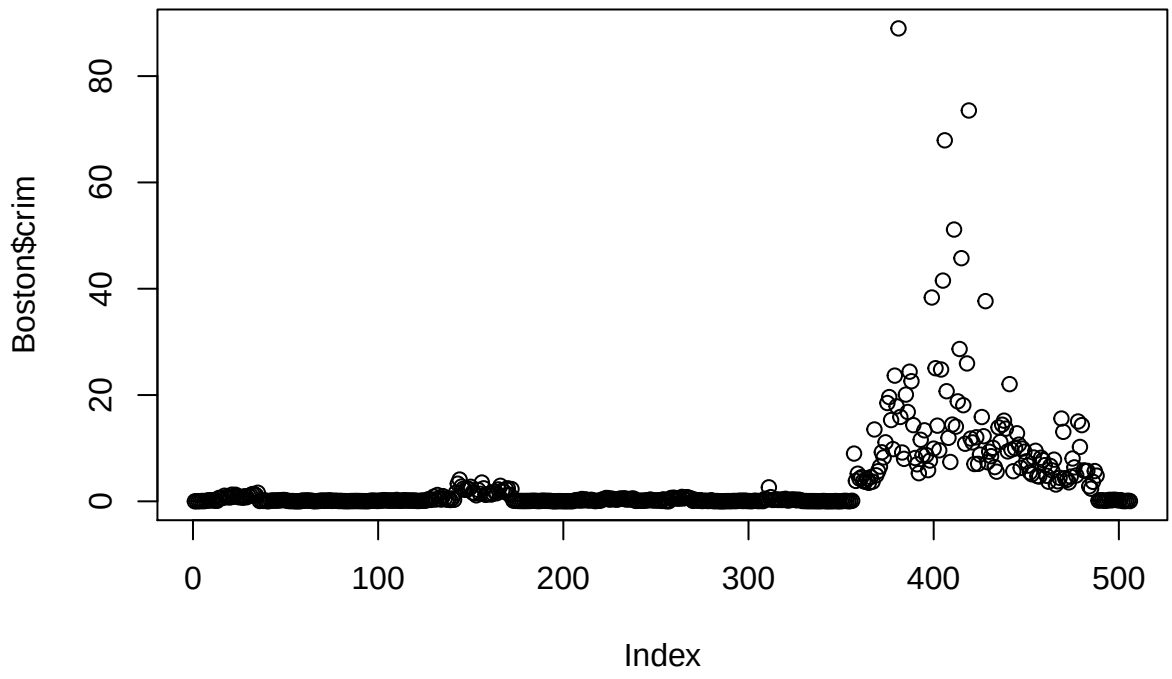
```
summary(Boston)
```

```
##       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
```

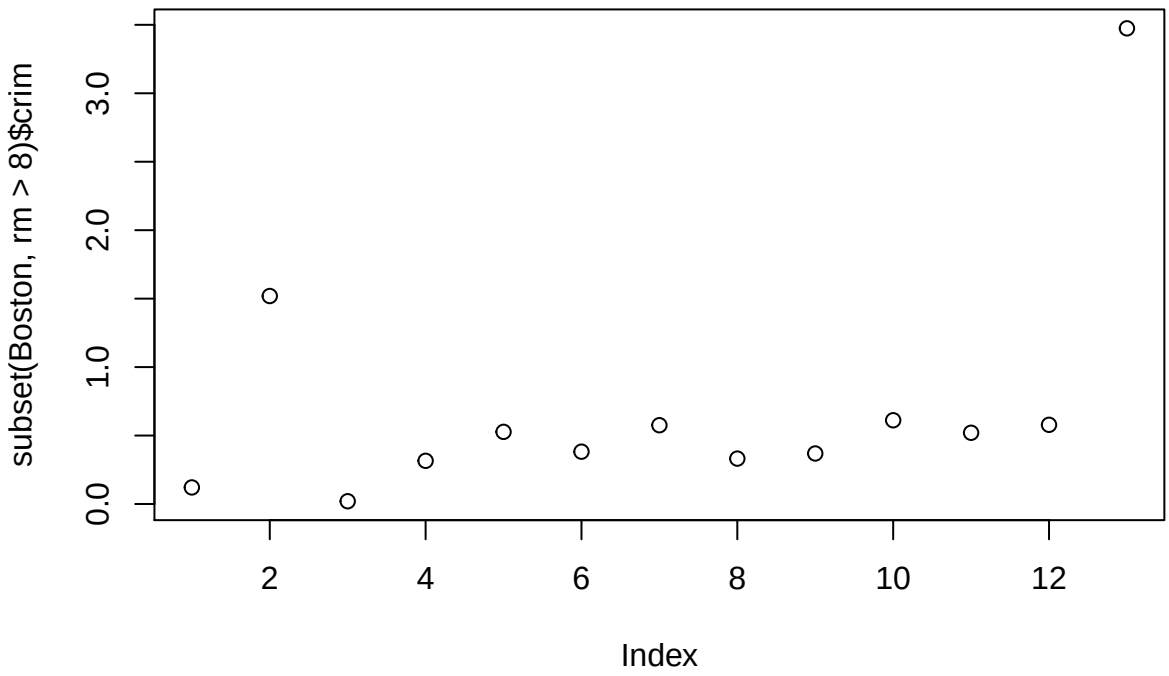
```
rbind(sapply(Boston[Boston$rm > 8,], mean), sapply(Boston, median))
```

```
##           crim       zn    indus      chas       nox       rm      age      dis
## [1,] 0.7187954 13.61538 7.078462 0.1538462 0.5392385 8.348538 71.53846 3.430192
## [2,] 0.2565100  0.00000 9.690000 0.0000000 0.5380000 6.208500 77.50000 3.207450
##           rad      tax  ptratio    black lstat medv
## [1,] 7.461538 325.0769 16.36154 385.2108  4.31 44.2
## [2,] 5.000000 330.0000 19.05000 391.4400 11.36 21.2
```

```
plot(Boston$crim)
```




```
plot(subset(Boston, rm > 8)$crim)
```



For the suburbs that average more than eight rooms per dwelling, there is a lower crime rate, and a lower lstat. This is simply by looking at the range, and the respective scatter plots. Also there is a much lower `medv` value than the overall data set, and a much higher `lstat` value.