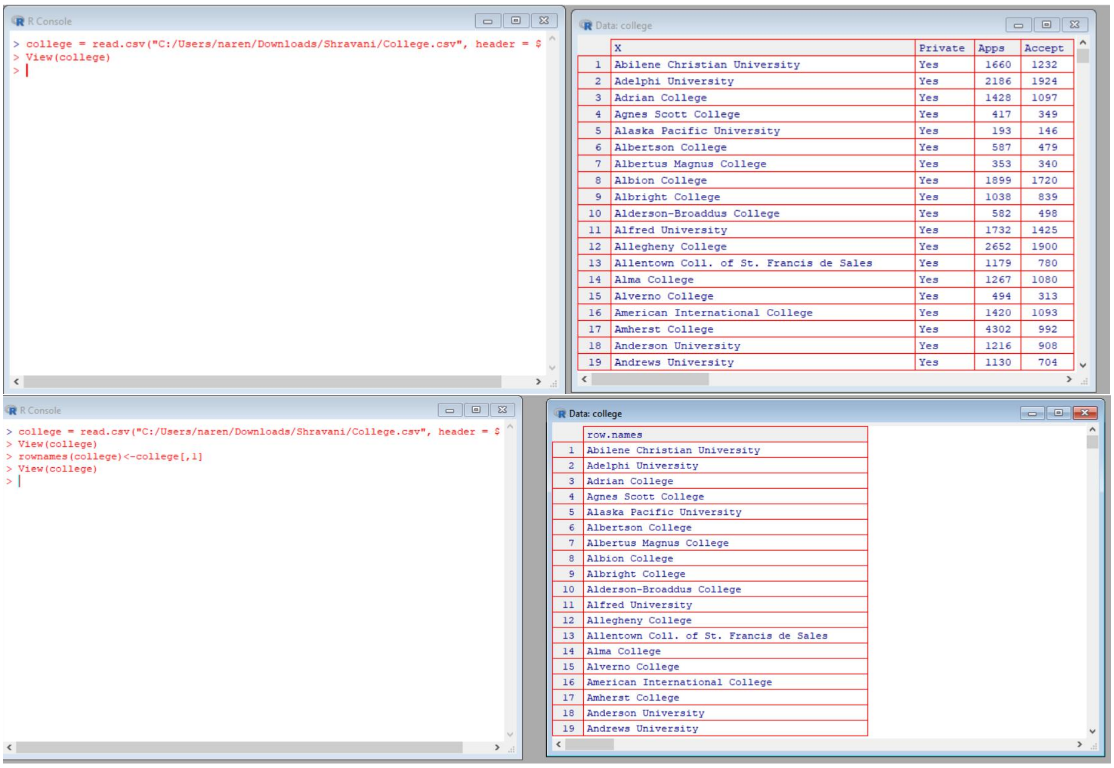
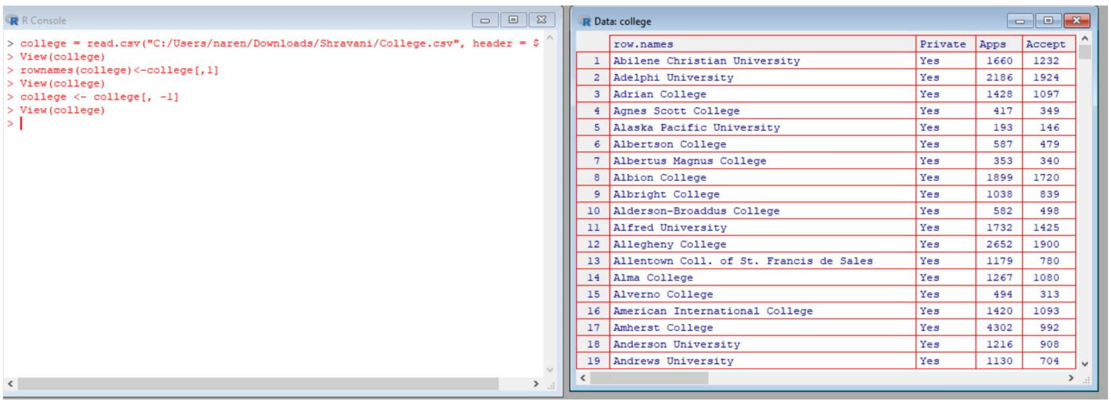
Introduction to Statistical Learning – Lab#1
Student Name: Shravani Nalla
Student ID: 12576204

1. ISLR 2.4 Applied Problem 8

A) **Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.**



B) **Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:**



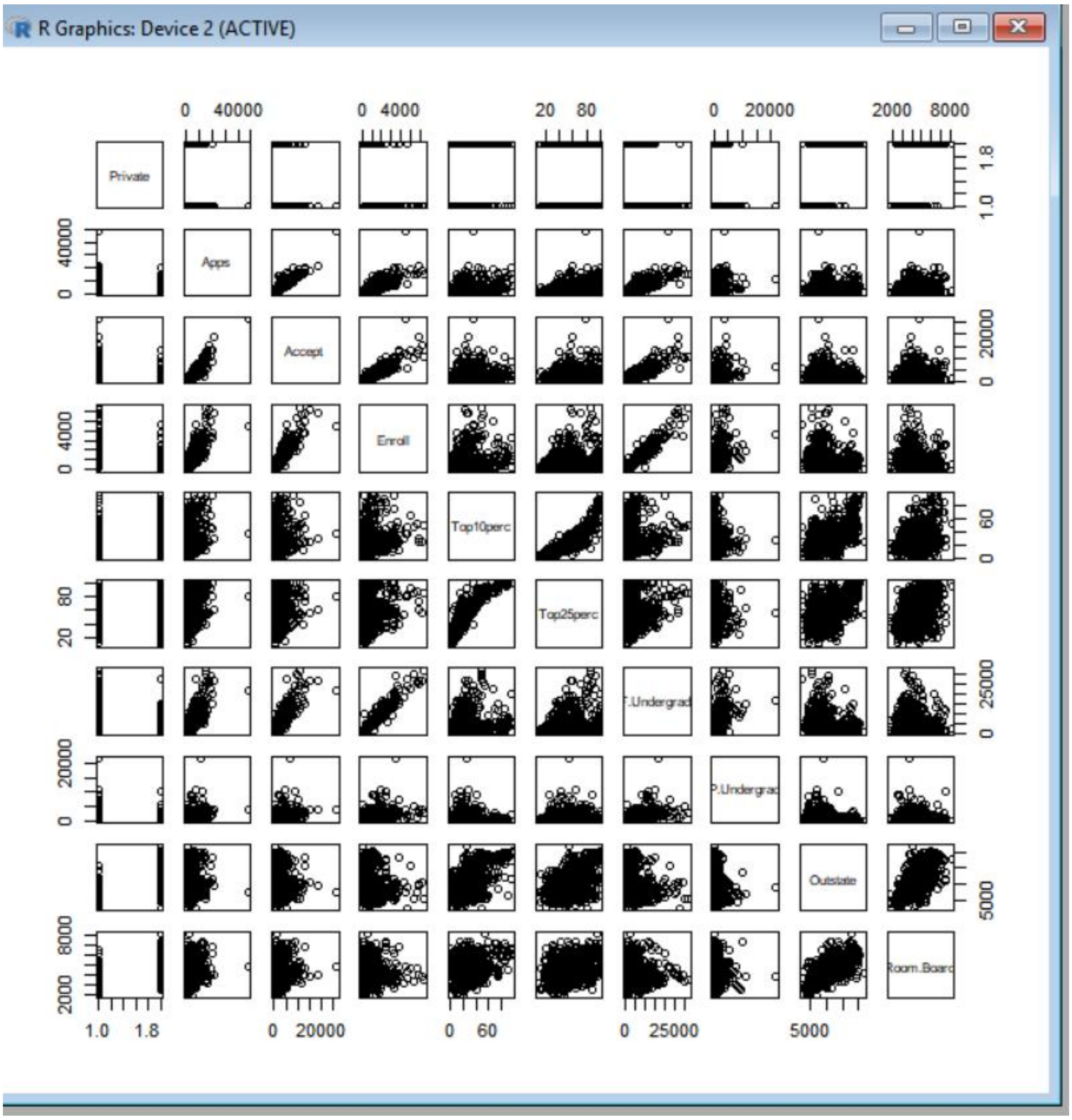
C)

i. **Use the summary() function to produce a numerical summary of the variables in the data set.**

```
> View(college)
> summary(college)
   Private              Apps           Accept          Enroll    
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35  
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242  
 Mode  :character   Median : 1558   Median : 1110   Median : 434  
                    Mean   : 3002   Mean   : 2019   Mean   : 780  
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902  
                    Max.   :48094   Max.   :26330   Max.   :6392  
   Top10perc       Top25perc      F.Undergrad     P.Undergrad     
 Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0  
 1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0  
 Median :23.00   Median : 54.0   Median : 1707   Median :  353.0  
 Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3  
 3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0  
 Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0  
    Outstate       Room.Board       Books           Personal   
 Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250  
 1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850  
 Median : 9990   Median :4200   Median : 500.0   Median :1200  
 Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341  
 3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700  
 Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800  
      PhD            Terminal       S.F.Ratio      perc.alumni   
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00  
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00  
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00  
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74  
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00  
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00  
     Expend        Grad.Rate     
 Min.   : 3186   Min.   : 10.00  
 1st Qu.: 6751   1st Qu.: 53.00  
 Median : 8377   Median : 65.00  
 Mean   : 9660   Mean   : 65.46  
 3rd Qu.:10830   3rd Qu.: 78.00  
 Max.   :56233   Max.   :118.00  
```

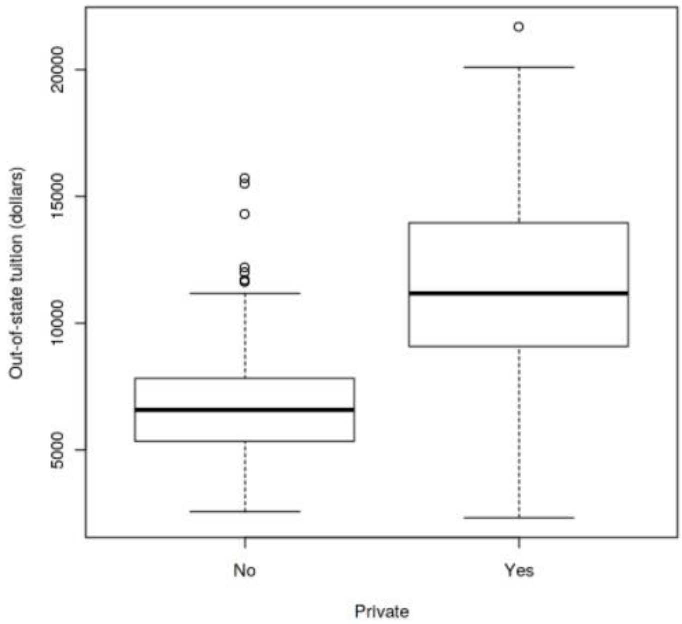
ii. **Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[, 1:10].**

```
> View(college[,1])
> college[,1] = as.numeric(factor(college[,1]))
> pairs(college[,1:10])
> 
```



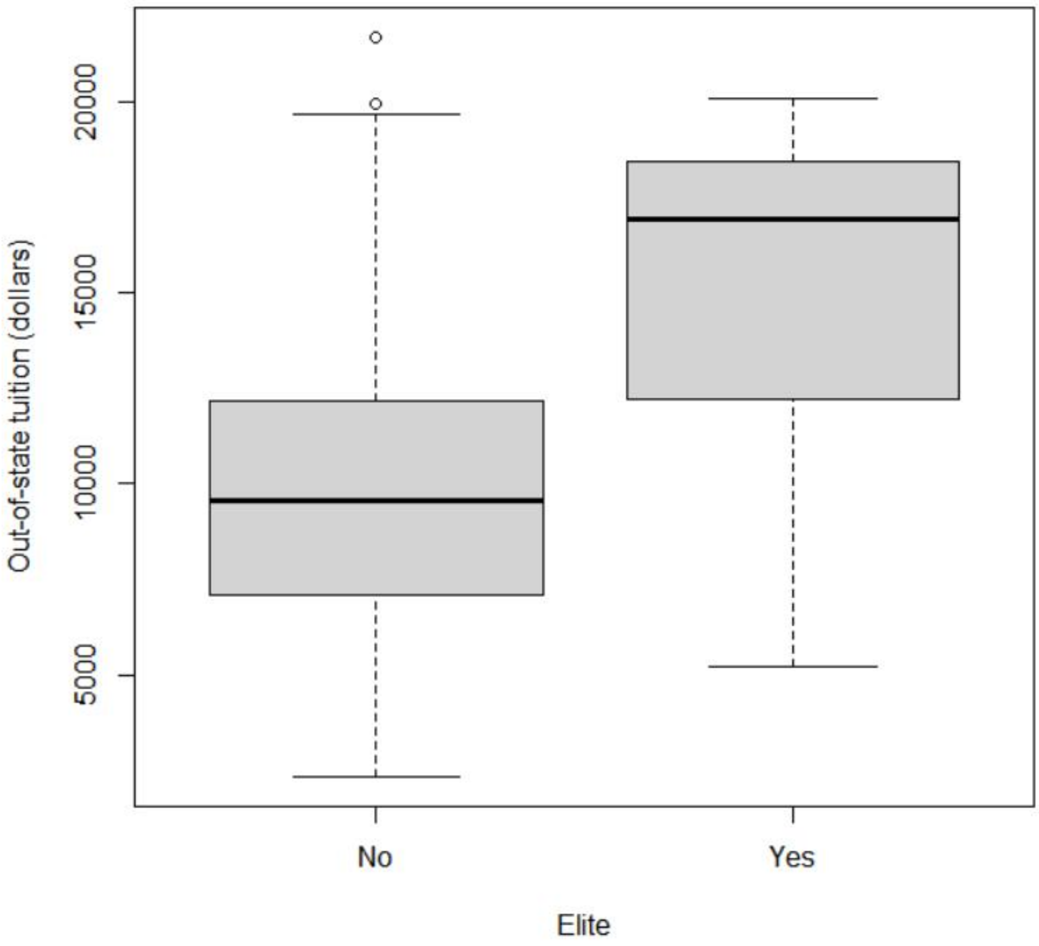
iii. **Use the plot() function to produce side-by-side boxplots of Outstate versus Private.**

```
plot(college$Private, college$Outstate, xlab = "Private", ylab = "Out-of-state tuition (dollars)")
```



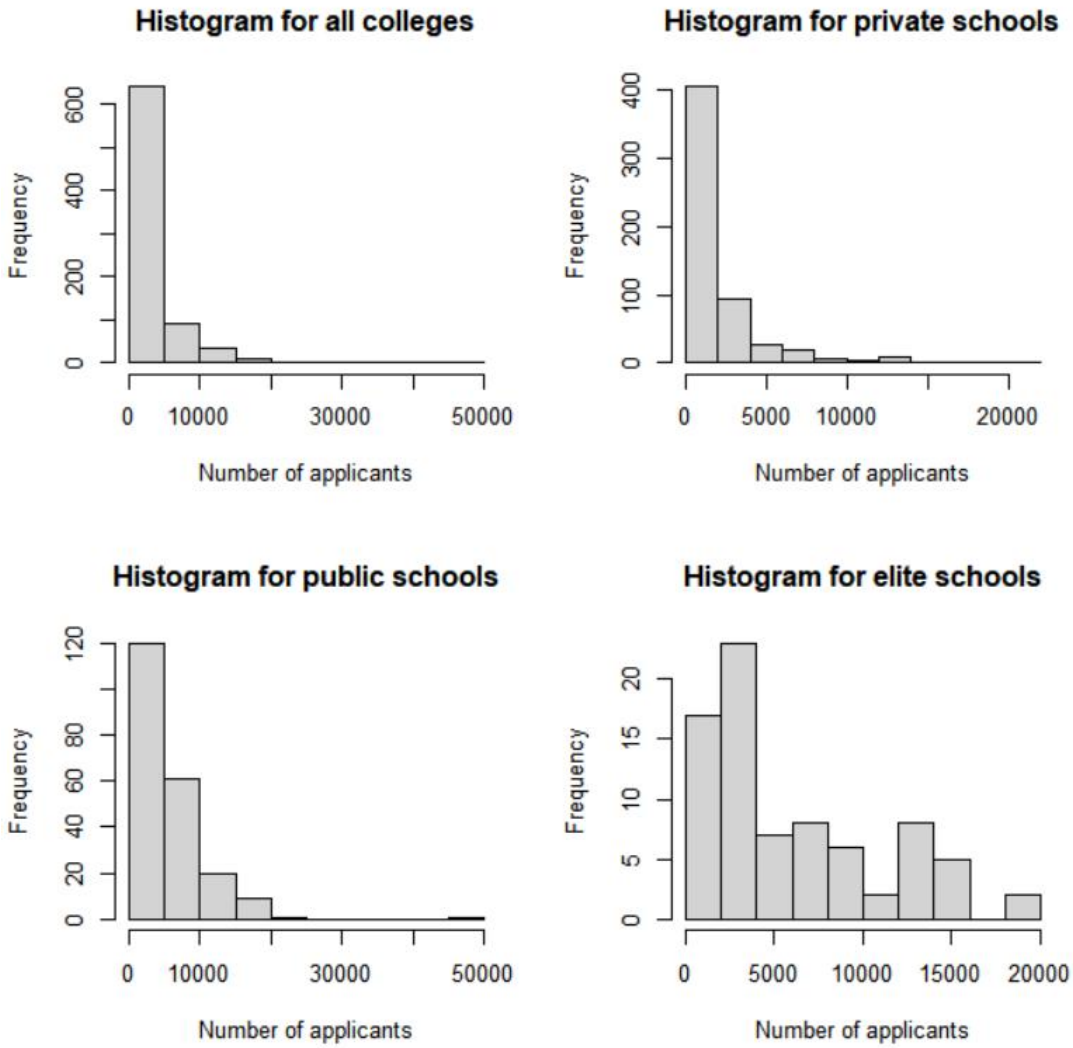
iv. **Create a new qualitative variable, called Elite, by *binning* the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.**

```
> Elite = rep("No", nrow(college))
> Elite[college$Top10per > 50] = "Yes"
> Elite = as.factor(Elite)
> college = data.frame(college, Elite)
> summary(college$Elite)
 No Yes
699  78
> plot(college$Elite, college$Outstate, xlab = "Elite", ylab = "Out-of-state tuition (dollars)")
```

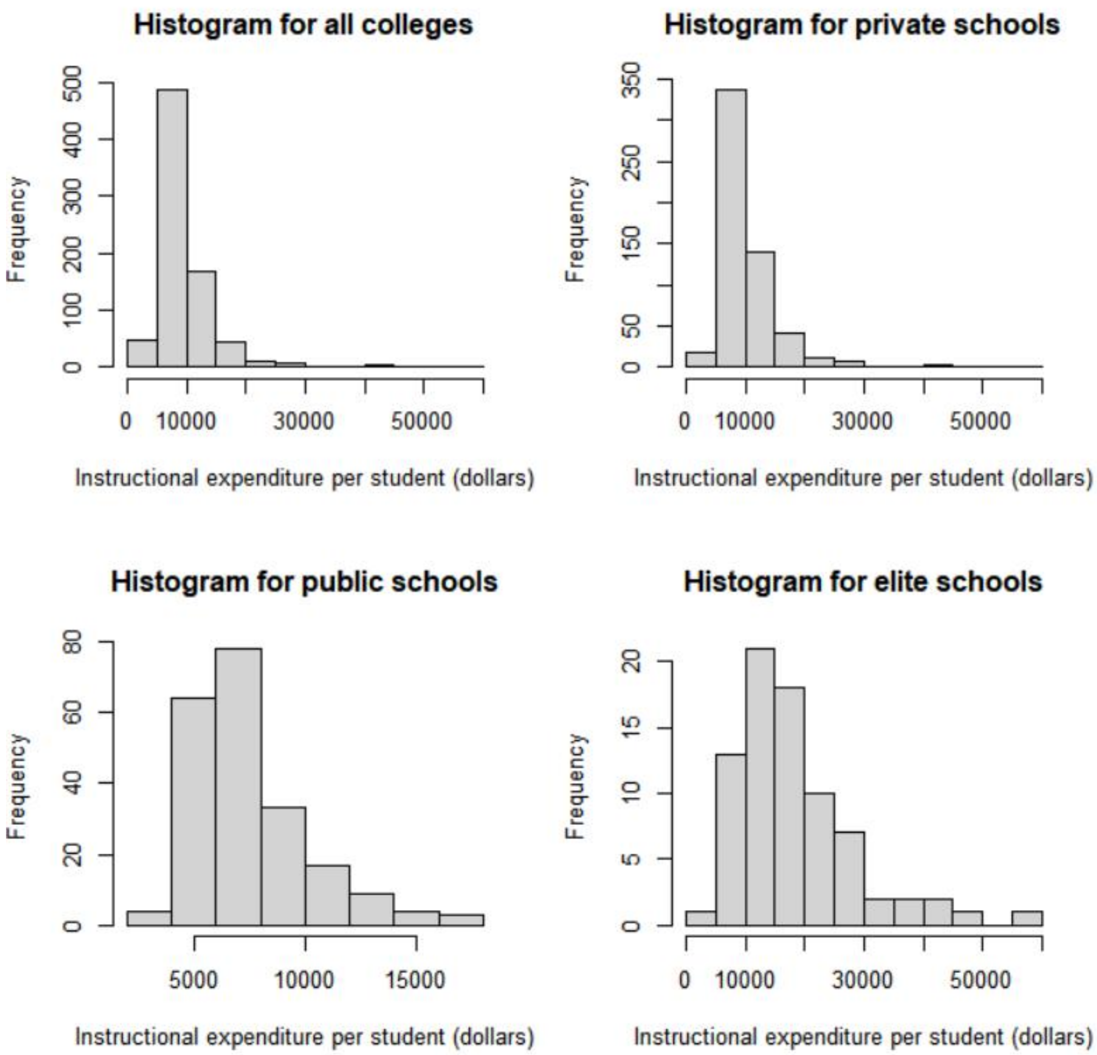


v. **Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.**

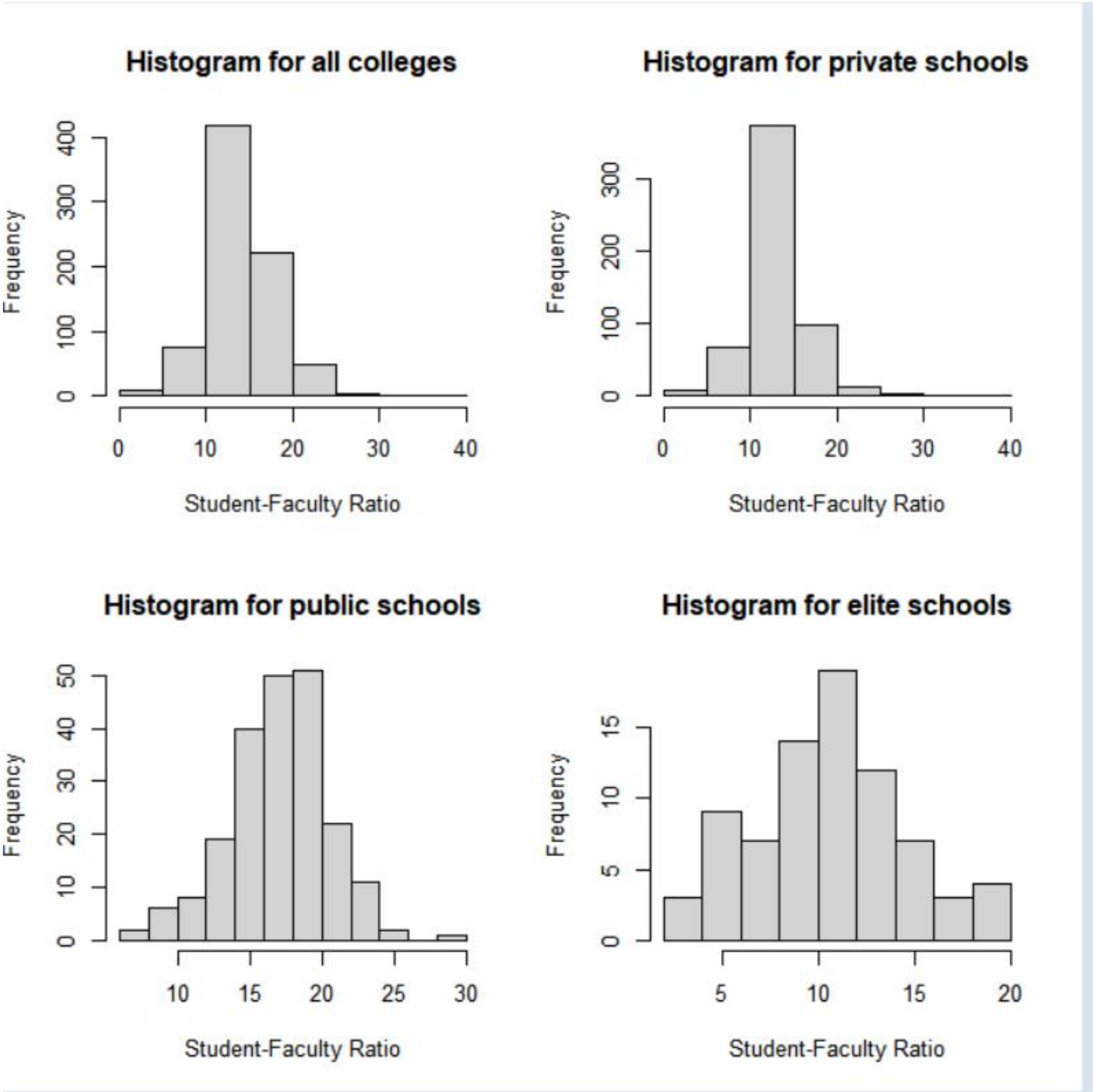
```
> par(mfrow = c(2, 2))
> hist(college$Apps, xlab = "Number of applicants", main = "Histogram for all colleges")
> hist(college$Apps[college$Private == "2"], xlab = "Number of applicants", main = "Histogram for private schools")
> hist(college$Apps[college$Private == "1"], xlab = "Number of applicants", main = "Histogram for public schools")
> hist(college$Apps[college$Elite == "Yes"], xlab = "Number of applicants", main = "Histogram for elite schools")
```




```
> par(mfrow = c(2, 2))
> hist(college$Expend, xlab = "Instructional expenditure per student (dollars)", main = "Histogram for all colleges")
> hist(college$Expend[college$Private == "2"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for private schools")
> hist(college$Expend[college$Private == "1"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for public schools")
> hist(college$Expend[college$Elite == "Yes"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for elite schools")
```

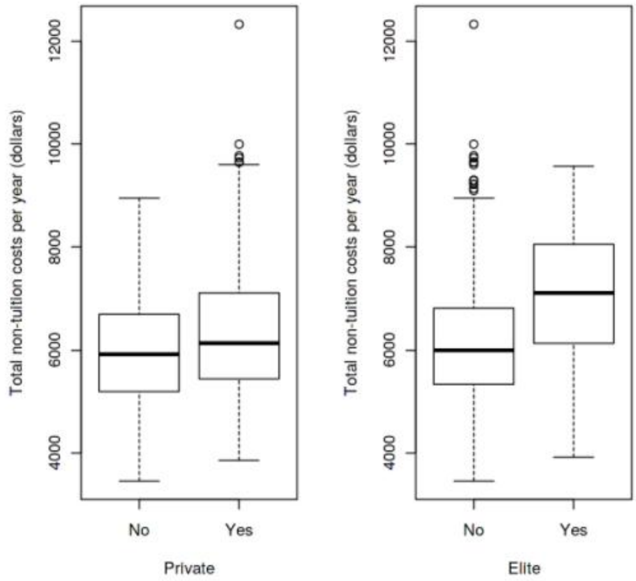



```
> par(mfrow = c(2, 2))
> hist(college$S.F.Ratio, xlab = "Student-Faculty Ratio", main = "Histogram for all colleges")
> hist(college$S.F.Ratio[college$Private == "2"], xlab = "Student-Faculty Ratio", main = "Histogram for private schools")
> hist(college$S.F.Ratio[college$Private == "1"], xlab = "Student-Faculty Ratio", main = "Histogram for public schools")
> hist(college$S.F.Ratio[college$Elite == "Yes"], xlab = "Student-Faculty Ratio", main = "Histogram for elite schools")
```



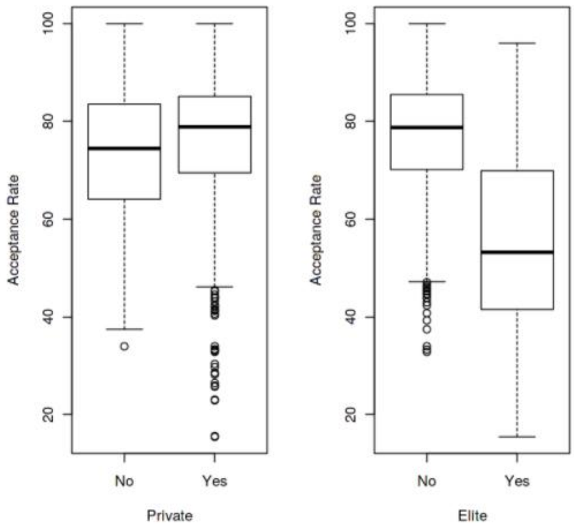
vi. **Continue exploring the data, and provide a brief summary of what you discover.**

```
> NonTuitionCosts = college$Room.Board + college$Books + college$Personal
> college = data.frame(college, NonTuitionCosts)
> par(mfrow = c(1, 2))
> plot(college$Private, college$NonTuitionCosts, xlab = "Private", ylab = "Total non-tuition costs per year (dollars)")
> plot(college$Elite, college$NonTuitionCosts, xlab = "Elite", ylab = "Total non-tuition costs per year (dollars)")
```



Based on the above box plots, it looks like that, aside from some outlier schools with very high costs, there isn't a wide gap for the median non-tution costs between private schools and public schools. The box plots do show, though, that there is a distinct difference in median non-tuition costs between elite and non-elite schools, with elite schools having higher costs.

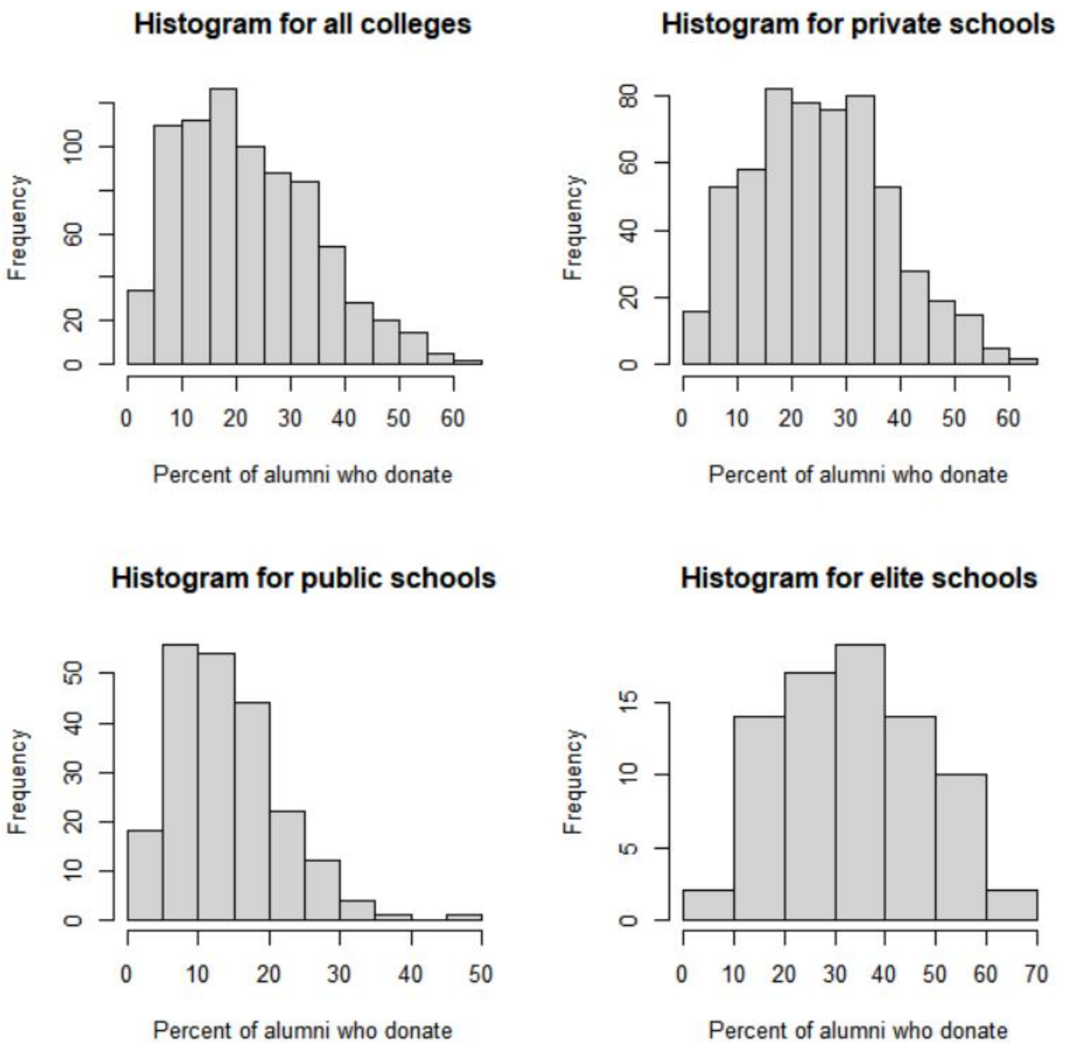
```
> AcceptPerc = college$Accept / college$Apps * 100
> college = data.frame(college, AcceptPerc)
> par(mfrow = c(1, 2))
> plot(college$Private, college$AcceptPerc, xlab = "Private", ylab = "Acceptance Rate")
> plot(college$Elite, college$AcceptPerc, xlab = "Elite", ylab = "Acceptance Rate")
```



```
> summary(college$AcceptPerc[college$Private == "1"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  33.97   64.12   74.43   72.65   83.42  100.00 
> summary(college$AcceptPerc[college$Private == "2"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.45   69.49   78.86   75.46   85.10  100.00 
> summary(college$AcceptPerc[college$Elite == "Yes"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.45   41.53   53.30   54.34   69.59   96.05 
> summary(college$AcceptPerc[college$Elite == "No"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  32.83   70.13   78.81   76.96   85.48  100.00 
```

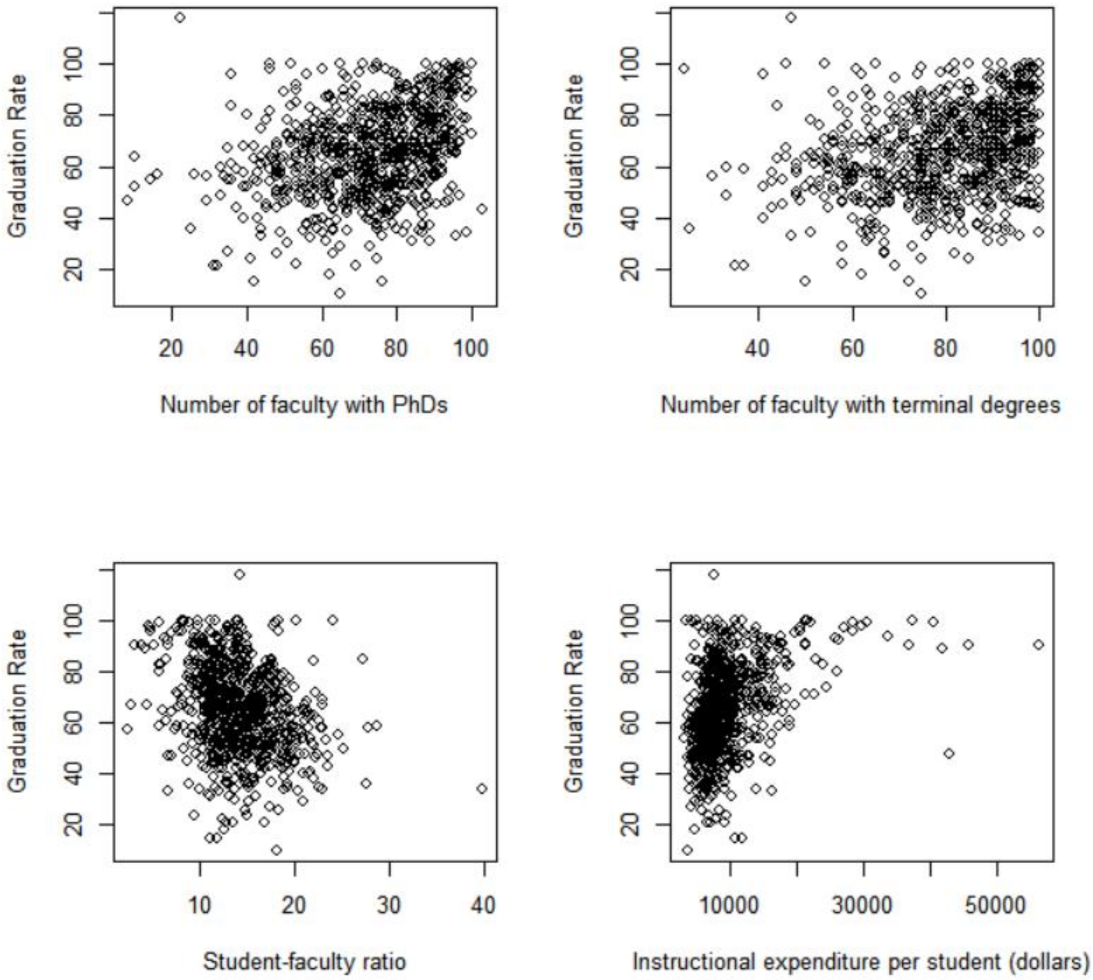
The boxplots show that while the median acceptance rates for both private and public schools are pretty close at around 75-80%, private schools have a much wider range of acceptance rates (going down to a minimum of 15.45%). When we distinguish between elite and non-elite schools, elite schools have a much lower median acceptance rate compared to non-elite ones.

```
> par(mfrow = c(2, 2))
> hist(college$perc.alumni, xlab = "Percent of alumni who donate", main = "Histogram for all colleges")
> hist(college$perc.alumni[college$Private == "2"], xlab = "Percent of alumni who donate", main = "Histogram for private schools")
> hist(college$perc.alumni[college$Private == "1"], xlab = "Percent of alumni who donate", main = "Histogram for public schools")
> hist(college$perc.alumni[college$Elite == "Yes"], xlab = "Percent of alumni who donate", main = "Histogram for elite schools")
```



Based on the above histograms, private schools and elite schools tend to have a higher percent of alumni who donate

```
> par(mfrow = c(2, 2))
> plot(college$PhD, college$Grad.Rate, xlab = "Number of faculty with PhDs", ylab = "Graduation Rate")
> plot(college$Terminal, college$Grad.Rate, xlab = "Number of faculty with terminal degrees", ylab = "Graduation Rate")
> plot(college$S.F.Ratio, college$Grad.Rate, xlab = "Student-faculty ratio", ylab = "Graduation Rate")
> plot(college$Expend, college$Grad.Rate, xlab = "Instructional expenditure per student (dollars)", ylab = "Graduation Rate")
```



The above scatterplots explore some of the factors which might be related to student graduation rates. From the upper-left plot, it appears there is a weak positive relationship between the number of faculty with PhDs and graduation rates. The upper-right plot appears to indicate that there isn't relationship between the number of faculty with terminal degrees and graduation rates. The bottom-left plot indicates that as student-faculty ratios increase, graduation rates generally tend to decrease. Lastly, the bottom-right plot seems to show that there is a definite positive relationship between instructional expenditure per student and graduation rates, with higher expenditures corresponding to higher graduation rates.

2. ISLR 2.4 Applied Problem 9

A) **Which of the predictors are quantitative, and which are qualitative?**

```
> Auto = read.csv("C:/Users/naren/Downloads/Shravani/Auto.csv", header = TRUE, na.strings = "?")
> Auto = na.omit(Auto)
> dim(Auto)
[1] 392   9
> head(Auto)
  mpg cylinders displacement horsepower weight acceleration year origin                      name
1  18         8          307        130   3504         12.0   70      1 chevrolet chevelle malibu
2  15         8          350        165   3693         11.5   70      1         buick skylark 320
3  18         8          318        150   3436         11.0   70      1        plymouth satellite
4  16         8          304        150   3433         12.0   70      1             amc rebel sst
5  17         8          302        140   3449         10.5   70      1               ford torino
6  15         8          429        198   4341         10.0   70      1          ford galaxie 500
> str(Auto)
'data.frame':	392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...
 - attr(*, "na.action")= 'omit' Named int [1:5] 33 127 331 337 355
  ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

The quantitative variables are mpg, displacement, horsepower, weight, and acceleration. Depending on the context, we may want to treat cylinders and year as quantitative predictors or qualitative ones. Lastly, origin and name are qualitative predictors. origin is a quantitative encoding of a car's country of origin, where 1 being American, 2 being European, and 3 being Japanese.

B) **What is the *range* of each quantitative predictor? You can answer this using the range() function.**

```
> 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
> summary(Auto[ , -c(4,9)])
      mpg          cylinders      displacement       weight      acceleration        year           origin     
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000  
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000  
 Median :22.75   Median :4.000   Median :151.0   Median :2804   Median :15.50   Median :76.00   Median :1.000  
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577  
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000  
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000  
```

C) **What is the mean and standard deviation of each quantitative predictor?**

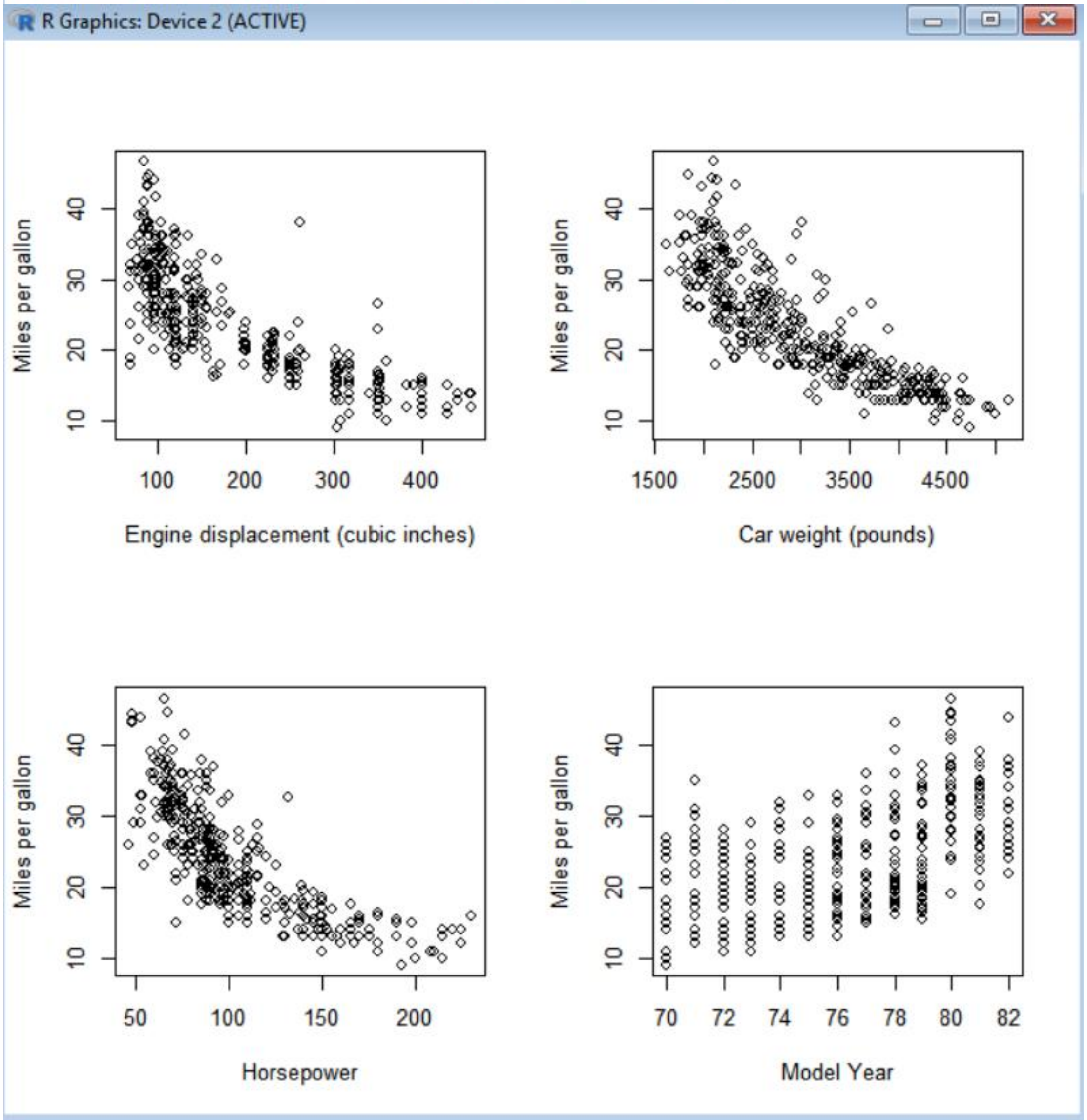
```
> sapply(Auto[ , -c(0,9)], mean)
         mpg    cylinders displacement   horsepower       weight acceleration         year       origin 
   23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327    75.979592     1.576531 
> sapply(Auto[ , -c(0,9)], sd)
         mpg    cylinders displacement   horsepower       weight acceleration         year       origin 
   7.8050075    1.7057832  104.6440039   38.4911599  849.4025600    2.7588641    3.6837365    0.8055182 
```

D) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
> fit <- Auto[-c(10:85),-c(0,9)]
> sapply(fit, range)
      mpg cylinders displacement horsepower weight acceleration year origin
[1,] 11.0         3           68         46   1649          8.5   70      1
[2,] 46.6         8          455        230   4997         24.8   82      3
> sapply(fit, mean)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899    77.145570     1.601266
> sapply(fit, sd)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
   7.867283     1.654179    99.678367    35.708853   811.300208     2.693721     3.106217     0.819910
```

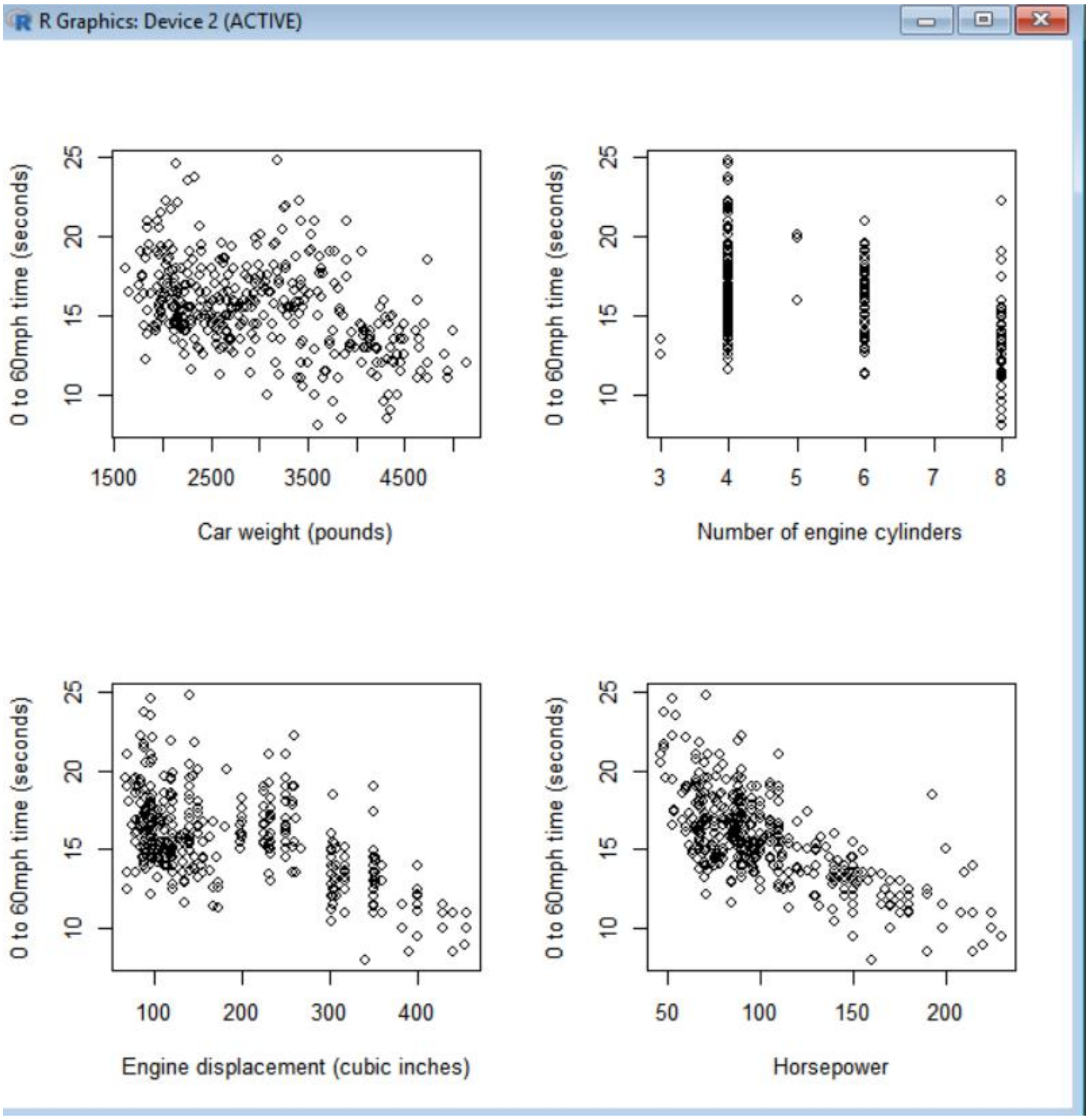
E) **Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.**

```
> par(mfrow = c(2, 2))
> plot(Auto$displacement, Auto$mpg, xlab = "Engine displacement (cubic inches)", ylab = "Miles per gallon")
> plot(Auto$horsepower, Auto$mpg, xlab = "Horsepower", ylab = "Miles per gallon")
> plot(Auto$weight, Auto$mpg, xlab = "Car weight (pounds)", ylab = "Miles per gallon")
> plot(Auto$year, Auto$mpg, xlab = "Model Year", ylab = "Miles per gallon")
```



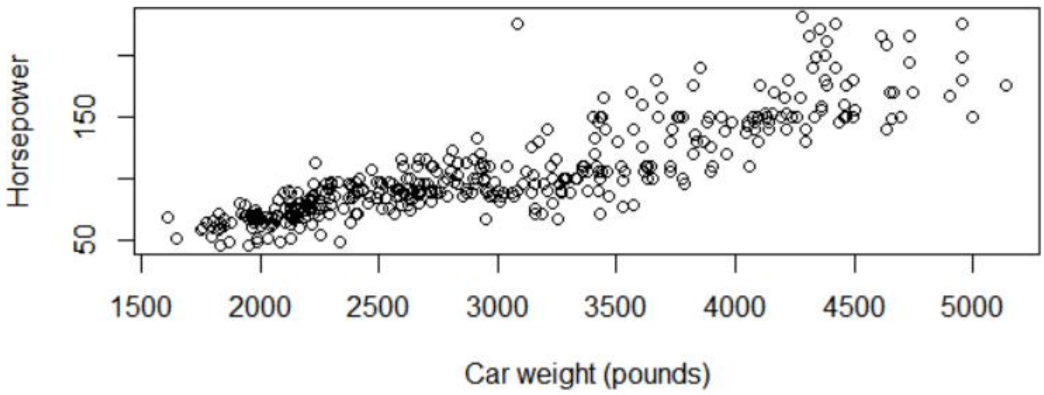
There are still some weak relationships, such as max engine displacement, car weight, and horsepower generally decreasing from 1970 to 1982. From a historical perspective, these changes could be in response to the 1973 and 1979 oil crises, in which spikes in oil prices pushed auto manufacturers to take measures to improve the efficiency of their cars.

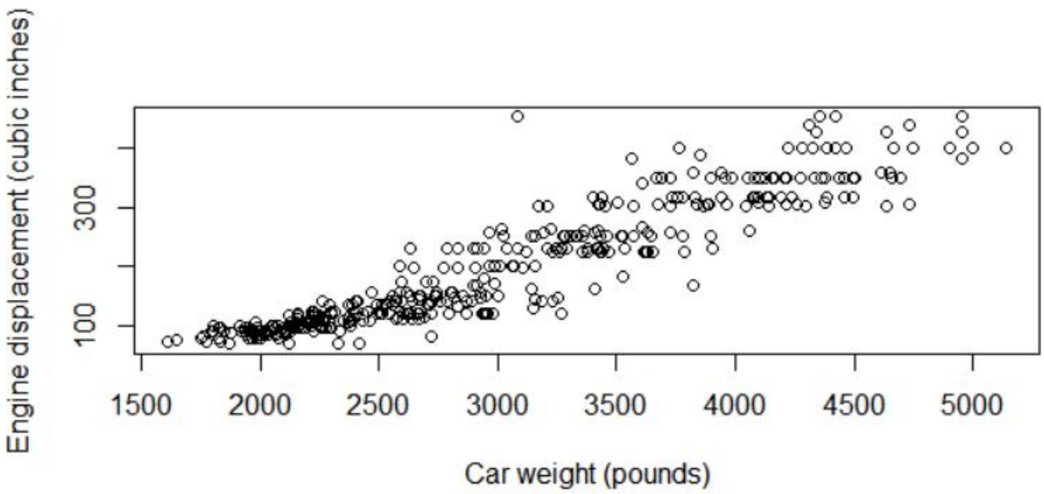
```
> par(mfrow = c(2, 2))
> plot(Auto$weight, Auto$acceleration, xlab = "Car weight (pounds)", ylab = "0 to 60mph time (seconds)")
> plot(Auto$cylinders, Auto$acceleration, xlab = "Number of engine cylinders", ylab = "0 to 60mph time (seconds)")
> plot(Auto$displacement, Auto$acceleration, xlab = "Engine displacement (cubic inches)", ylab = "0 to 60mph time (seconds)")
> plot(Auto$horsepower, Auto$acceleration, xlab = "Horsepower", ylab = "0 to 60mph time (seconds)")
```



Next, I explored the relationship between the number of seconds it takes a car to accelerate from 0 to 60 miles per hour and a number of different factors. As expected, the 0-to-60 time clearly decreases with increased engine displacement and increased horsepower. There is also a weak relationship that as the number of engine cylinders increases the 0-to-60 time tends to decrease. While it may seem counter-intuitive at first, the 0-to-60 time also tends to decrease with car weight. This makes more sense in the context of the two scatterplots below, which shows that the higher weight is correlated with higher horsepower and higher engine displacement.

```
> par(mfrow = c(2, 1))
> plot(Auto$weight, Auto$horsepower, xlab = "Car weight (pounds)", ylab = "Horsepower")
> plot(Auto$weight, Auto$displacement, xlab = "Car weight (pounds)", ylab = "Engine displacement (cubic inches)")
```

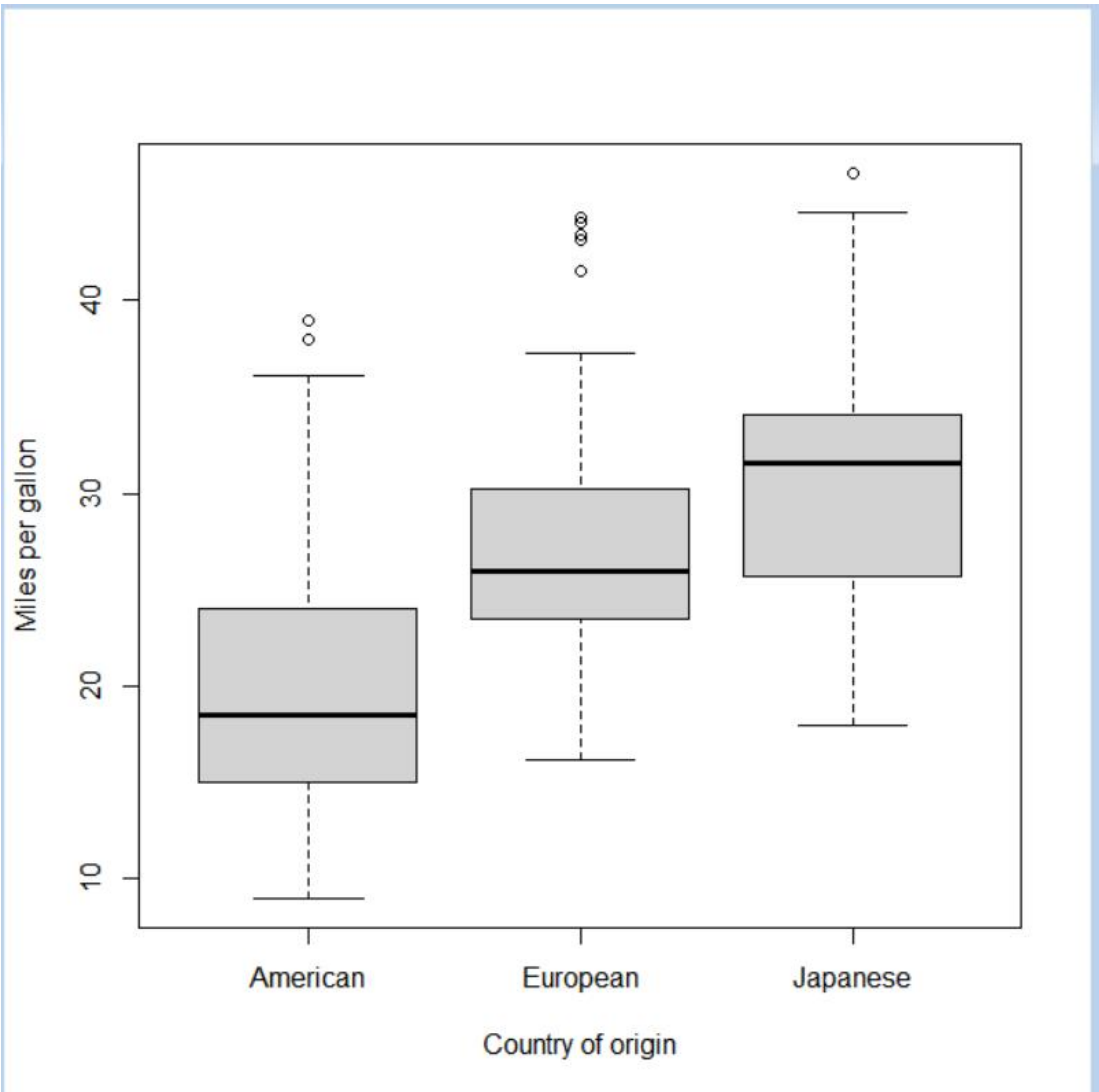




F) Suppose we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Based on the scatter plots I made in part 5 which relate miles per gallon to the predictors engine displacement, horsepower, car weight, and model year, it seems as if the first three factors would be most helpful in predicting mpg, with model year still potentially being helpful but less so. There are clear relationships that increasing engine displacement/horsepower/car weight results in decreased fuel efficiency. There is also a weak relationship that fuel efficiency generally increased going from 1970 to 1982.

```
> Auto$origin[Auto$origin == 1] = "American"
> Auto$origin[Auto$origin == 2] = "European"
> Auto$origin[Auto$origin == 3] = "Japanese"
> par(mfrow = c(1,1))
> plot(Auto$origin, Auto$mpg, xlab = "Country of origin", ylab = "Miles per gallon")
```




```
> cor(Auto$weight, Auto$displacement)
[1] 0.9329944
> cor(Auto$weight, Auto$horsepower)
[1] 0.8645377
> cor(Auto$displacement, Auto$horsepower)
[1] 0.897257
```

3. ISLR 2.4 Applied Problem 10

A) **To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library. How many rows are in this data set? How many columns? What do the rows and columns represent?**

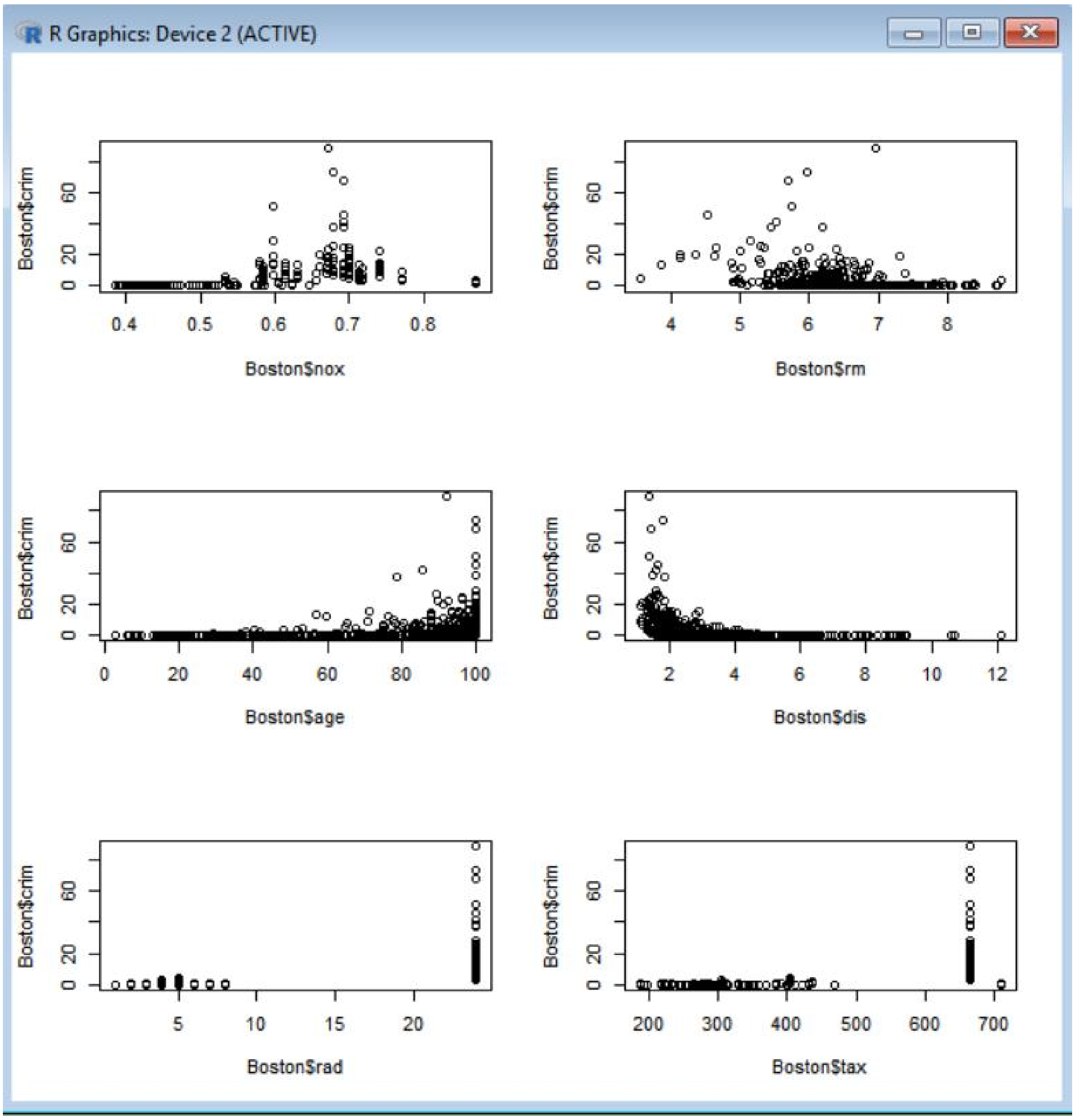
```
> library(MASS)
> Boston$chas <- as.factor(Boston$chas)
> nrow(Boston)
[1] 506
> ncol(Boston)
[1] 14
> |
```

The corrected Boston data set has 506 rows and 14 columns. Each row represents a particular tract of land within the city of Boston. The dataset has the following columns.

1. MEDV: Median value of owner-occupied housing in $1000 for the tract
2. CRIM: Per capita crime rate for the tract
3. ZN: Percent of residential land zoned for lots over 25000 square feet per town (constant for all tracts within the same town)
4. INDUS: Percent of non-retail business acres per town (constant for all tracts within the same town)
5. CHAS: Dummy variable to indicate whether or not the tract borders the Charles River (1 = Borders Charles River, 0 = Otherwise
6. NOX: Nitric oxides concentration (in parts per 10 million) per town (constant for all tracts within the same town)
7. RM: Average number of rooms per dwelling in the tract
8. AGE: Percent of owner-occupied units in the tract built prior to 1940
9. DIS: Weighted distance from the tract to five Boston employment centers
10. RAD: Index of accessibility to radial highways per town (constant for all tracts within the same town)
11. TAX: Full-value property tax rate per $10000 per town (constant for all tracts within the same town)
12. PTRATIO: Pupil-teacher ratio per town (constant for all tracts within the same town)
13. B: 1000(B−0.63)21000(B−0.63)2, where BB is the proportion of black residents in the tract
14. LSTAT: Percent of tract population designated as lower status

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

```
> par(mfrow = c(3,2))
> plot(Boston$nox, Boston$crim)
> plot(Boston$rm, Boston$crim)
> plot(Boston$age, Boston$crim)
> plot(Boston$dis, Boston$crim)
> plot(Boston$rad, Boston$crim)
> plot(Boston$tax, Boston$crim)
> |
```

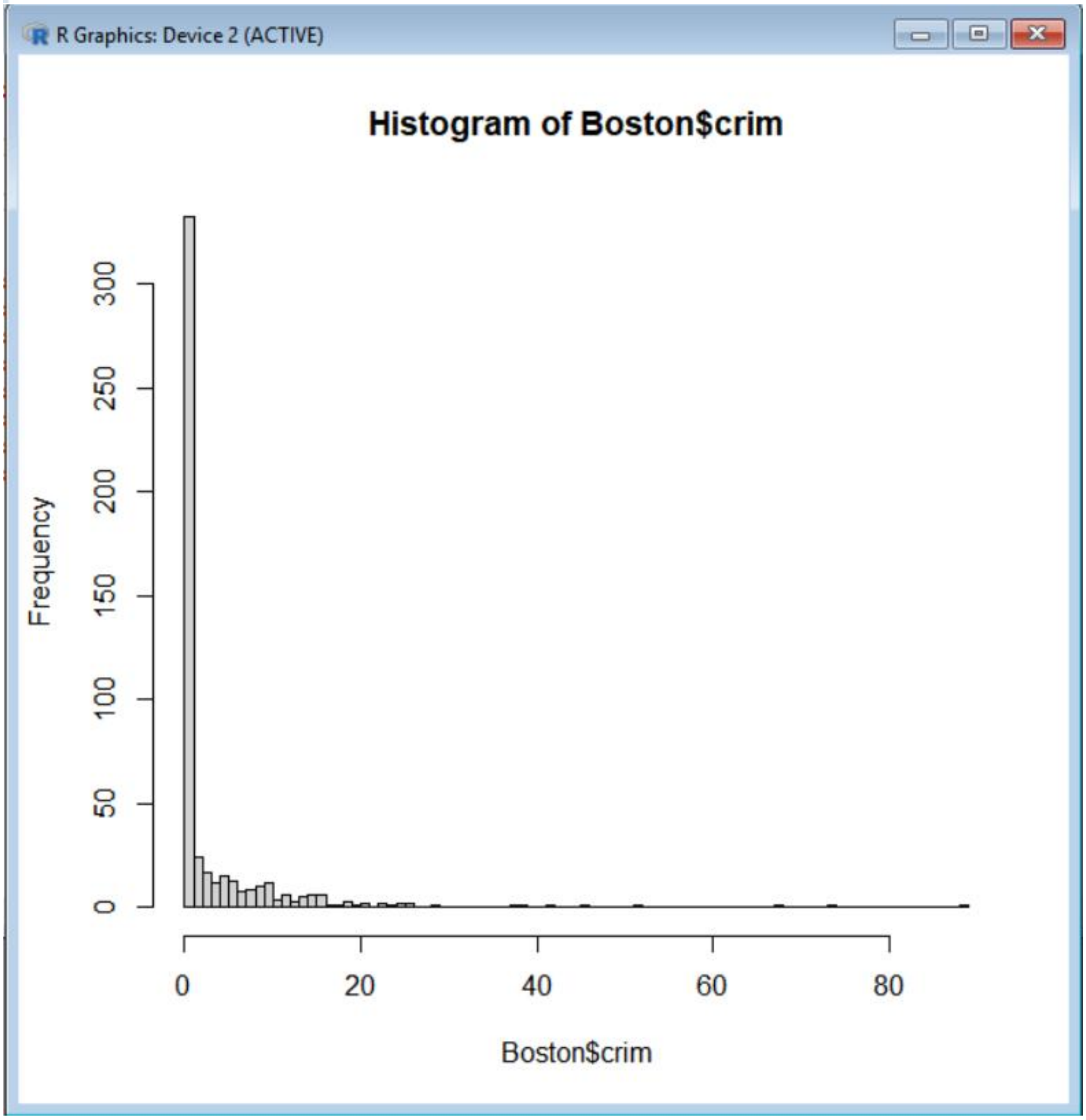


The first two scatter plots in this next group explore factors that might relate to the concentration of nitric oxides. While there isn't a strong relationship, it appears that tracts with higher median home value also weakly tend to have lower concentrations of nitric oxides. There is a much clearer relationship with the percentage of non-retail business acres -- tracts with a higher proportion of non-retail business acres tend to have higher concentrations of nitric oxides. The bottom two plots look at some more factors which might be related to the median home value of a tract.

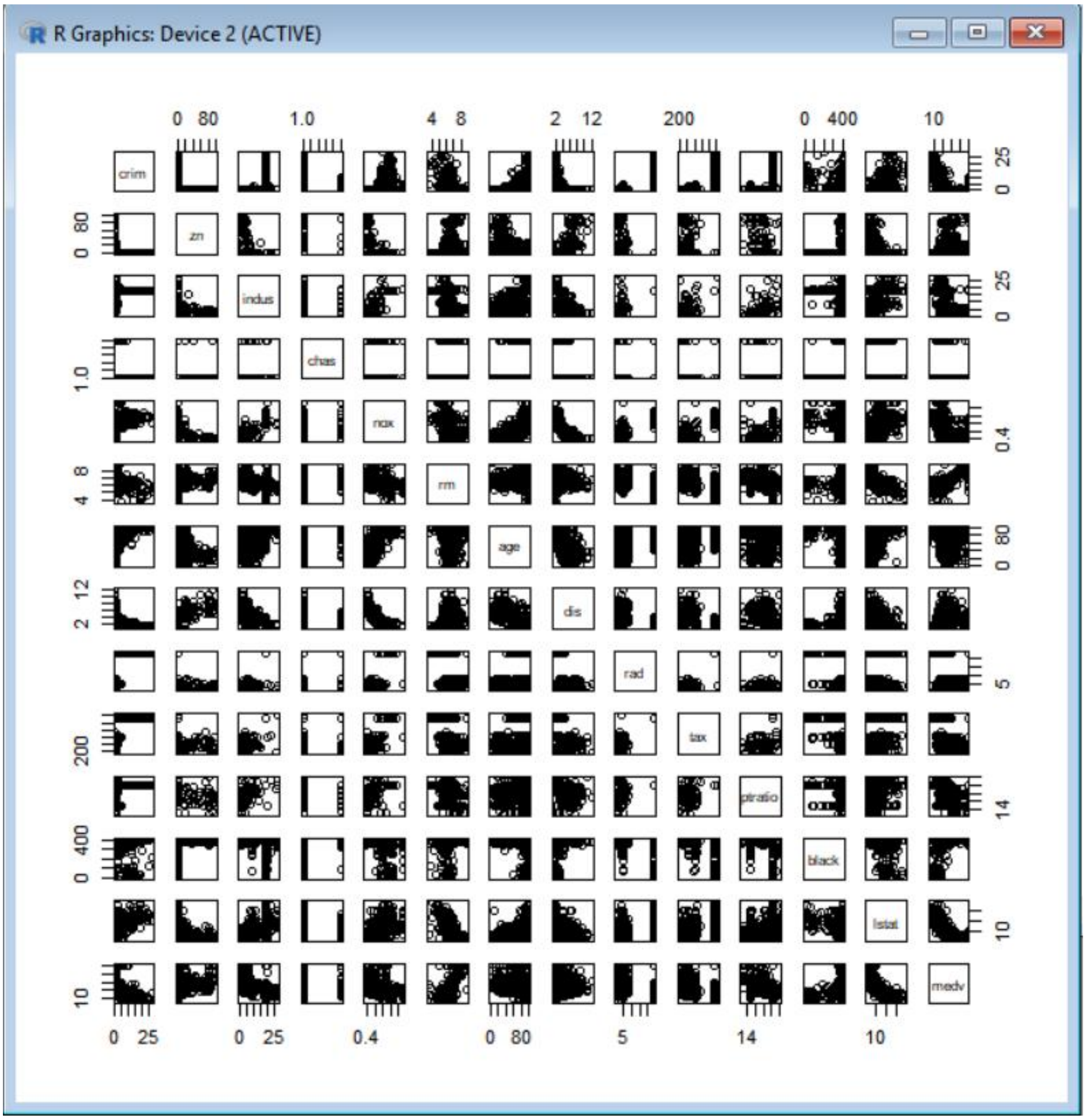
The bottom-left plot seems to indicate that there is a relationship between the value of B and CMEDV, where B increases as CMEDV increases. If I am interpreting this correctly, this means that tracts with high median home values have a very low (close to 0%) proportion of Black residents, while tracts with low median home values have a much higher proportion (close to 63%). The bottom-right plot appears to indicate that there is also a relationship between proximity to Boston employment centers and median home value, with home values generally increasing as one gets further away from the employment centers.

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

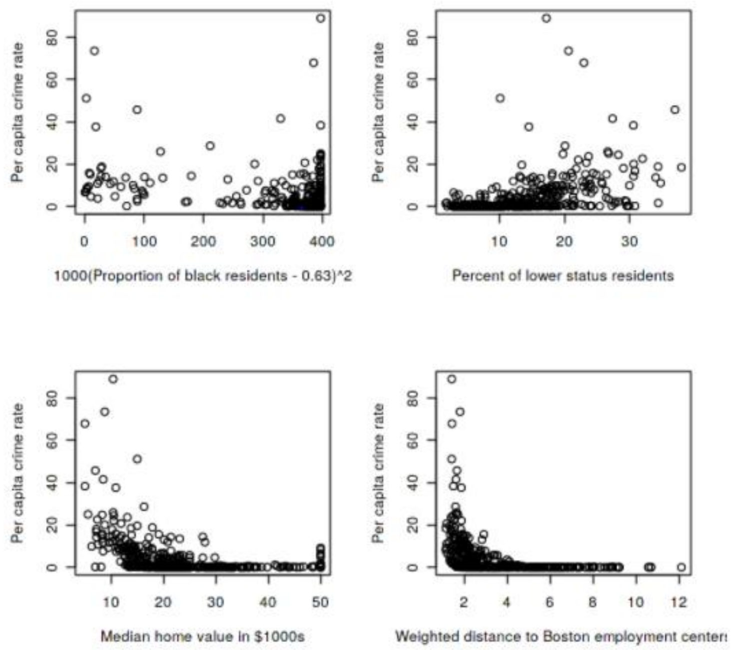
```
> hist(Boston$crim, breaks = 75)
>
```




```
> pairs(Boston[Boston$crim < 30, ])
>
```



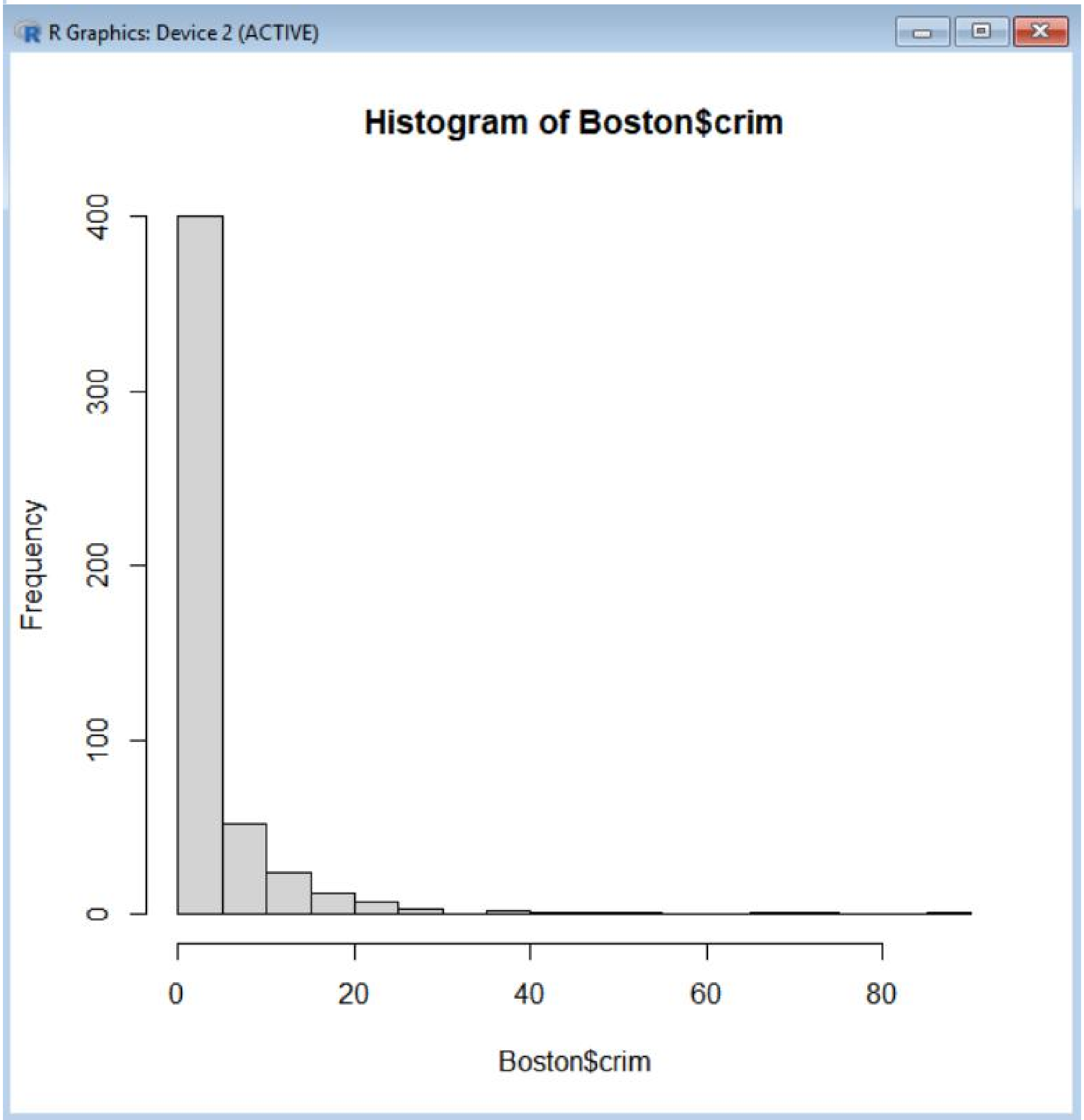

```
> par(mfrow = c(2, 2))
> plot(Boston$b, Boston$crim, xlab = "1000(Proportion of black residents - 0.63)^2", ylab = "Per capita crime rate")
> plot(Boston$lstat, Boston$crim, xlab = "Percent of lower status residents", ylab = "Per capita crime rate")
> plot(Boston$medv, Boston$crim, xlab = "Median home value in $1000s", ylab = "Per capita crime rate")
> |
```



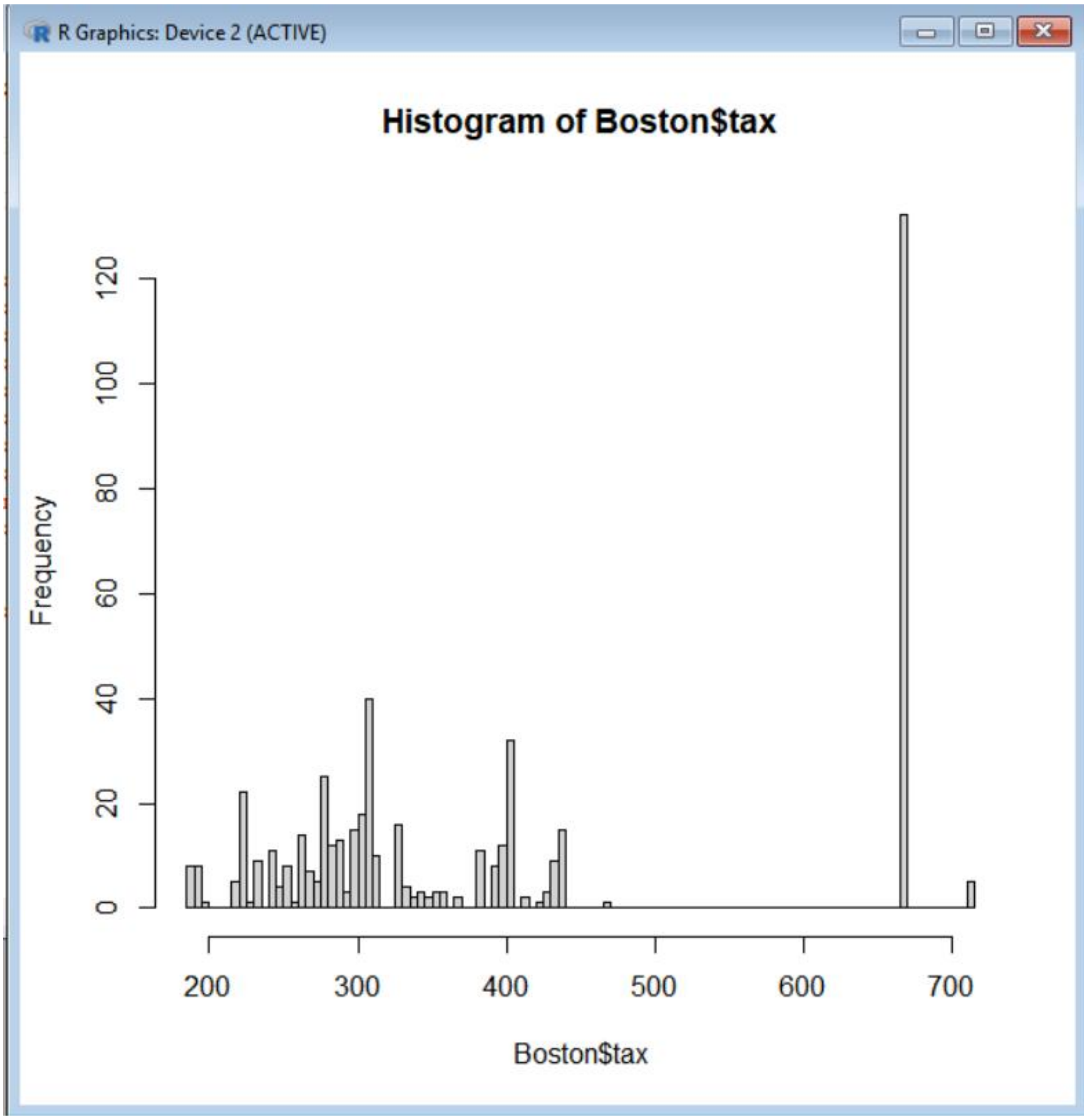
Based on the above four scatter plots, it appears that there are pretty clear relationships between crime rate and median home value, percent of lower status residents, and proximity to Boston employment centers. Tracts with lower home values tend to have higher crime rates, as do tracts which are closer to Boston employment centers. In addition, tracts with higher proportion of lower status residents tend to have higher crime rates. I was also curious if there would be a relationship between crime rate and B, which serves as some kind of measurement for the proportion of Black residents. Based on the scatter plot between those two variables, there doesn't appear to be a clear relationship.

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.**

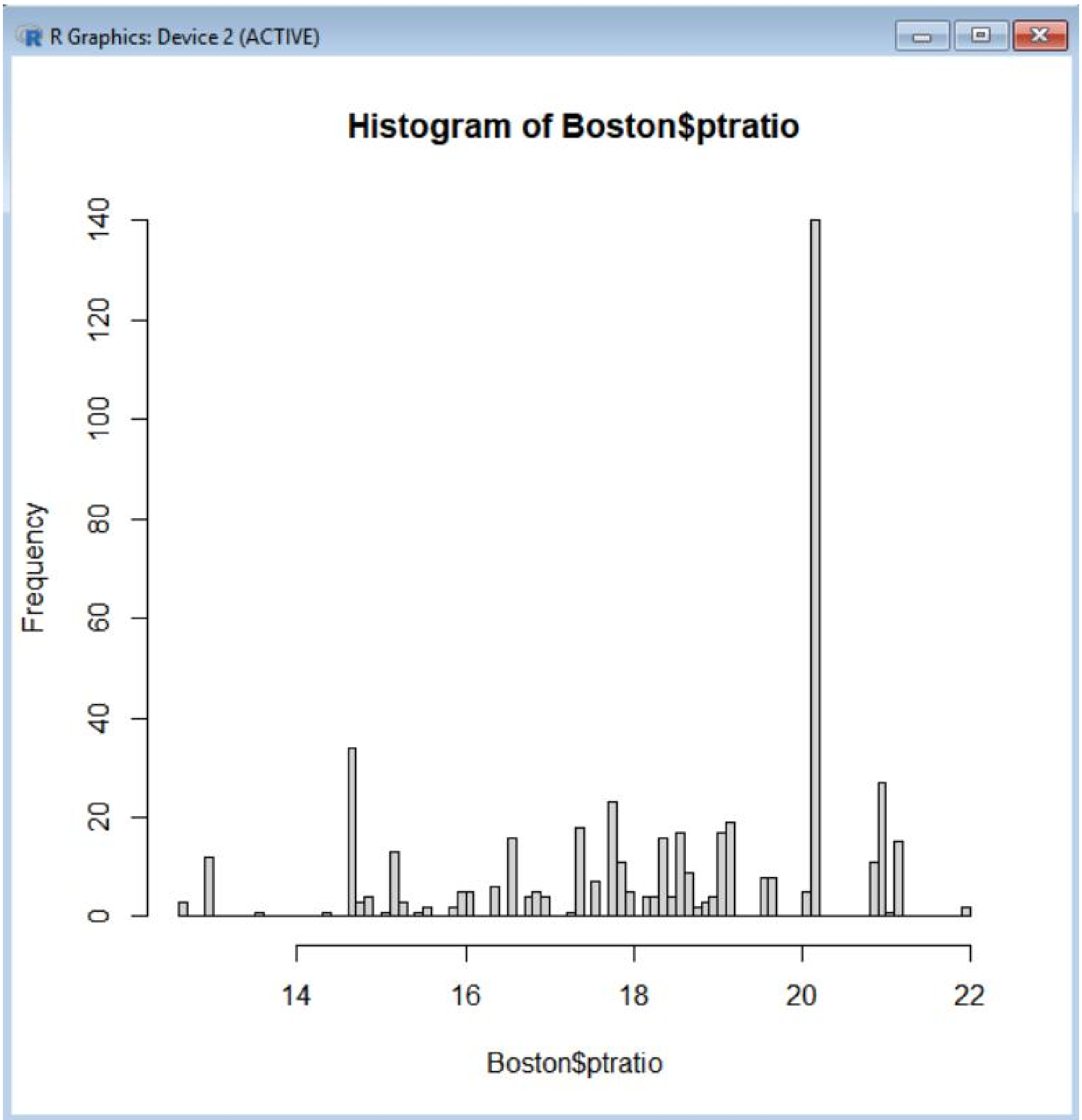
```
> hist(Boston$crim, breaks = 25)
>
```




```
> nrow(Boston[Boston$crim >30, ])
[1] 8
> hist(Boston$tax , breaks = 75)
>
```




```
> nrow(Boston[Boston$tax == 666, ])
[1] 132
> hist(Boston$ptratio , breaks = 75)
> |
```




```
> nrow(Boston[Boston$ptratio > 20, ])
[1] 201
```

Based on the histograms and the numerical summary, there do appear to be tracts within Boston which have particularly high crime rates, tax rates, or pupil-teacher ratios. The minimum crime rate is 0.00632, while the maximum is 88.97620, with a median of 0.25651. The minimum tax rate is $187 per \$10000, while the maximum is $711, with a median of \$330. The minimum pupil-teacher ratio is 12.60 pupils per teacher, while the maximum is 22, with a median of 19.05. Given the median value, the maximum pupil-teacher ratio in the data set isn't outrageously high, since about half of the tracts have a ratio of 19 or more.

E) **How many of the suburbs in this data set bound the Charles river?**

```
> nrow(Boston[Boston$chas ==1,])
[1] 35
```

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

```
> summary(Boston$ptratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.60   17.40   19.05   18.46   20.20   22.00
> median(Boston$ptratio)
[1] 19.05
```

G) **Which suburb of Boston has the 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.**

```
> row.names(Boston[min(Boston$medv), ])
[1] "5"
> range(Boston$tax)
[1] 187 711
> Boston[min(Boston$medv), ]$tax
[1] 222
> |
```

The tracts have relatively high values for B, though one tract has a maximum value while the other, with a value of 384.97, is in between the first and second quartiles. Lastly, the tracts have a high proportion of lower status residents (values of 30.59 and 22.98), putting them in the top quartile of the data.

In summary, these two tracts with the lowest median value of owner-occupied homes have predictors generally at the extreme ends of their respective ranges.

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.**

```
> nrow(Boston[Boston$rm >7, ])
[1] 64
> nrow(Boston[Boston$rm >8, ])
[1] 13
```

From the numerical summary, one thing that stands out is that the tracts which average at least eight rooms per dwelling have low crime rates, low concentrations of nitric oxides, low proportions of Black residents (high values of B), and low proportions of lower status residents compared to the overall data set.

4. ISLR 3.7 Applied Problem 8

A) **Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:**

1. **Is there a relationship between the predictor and the response.**
2. **How strong is the relationship between the predictor and the response?**
3. **Is the relationship between the predictor and the response positive or negative.**
4. **What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?**

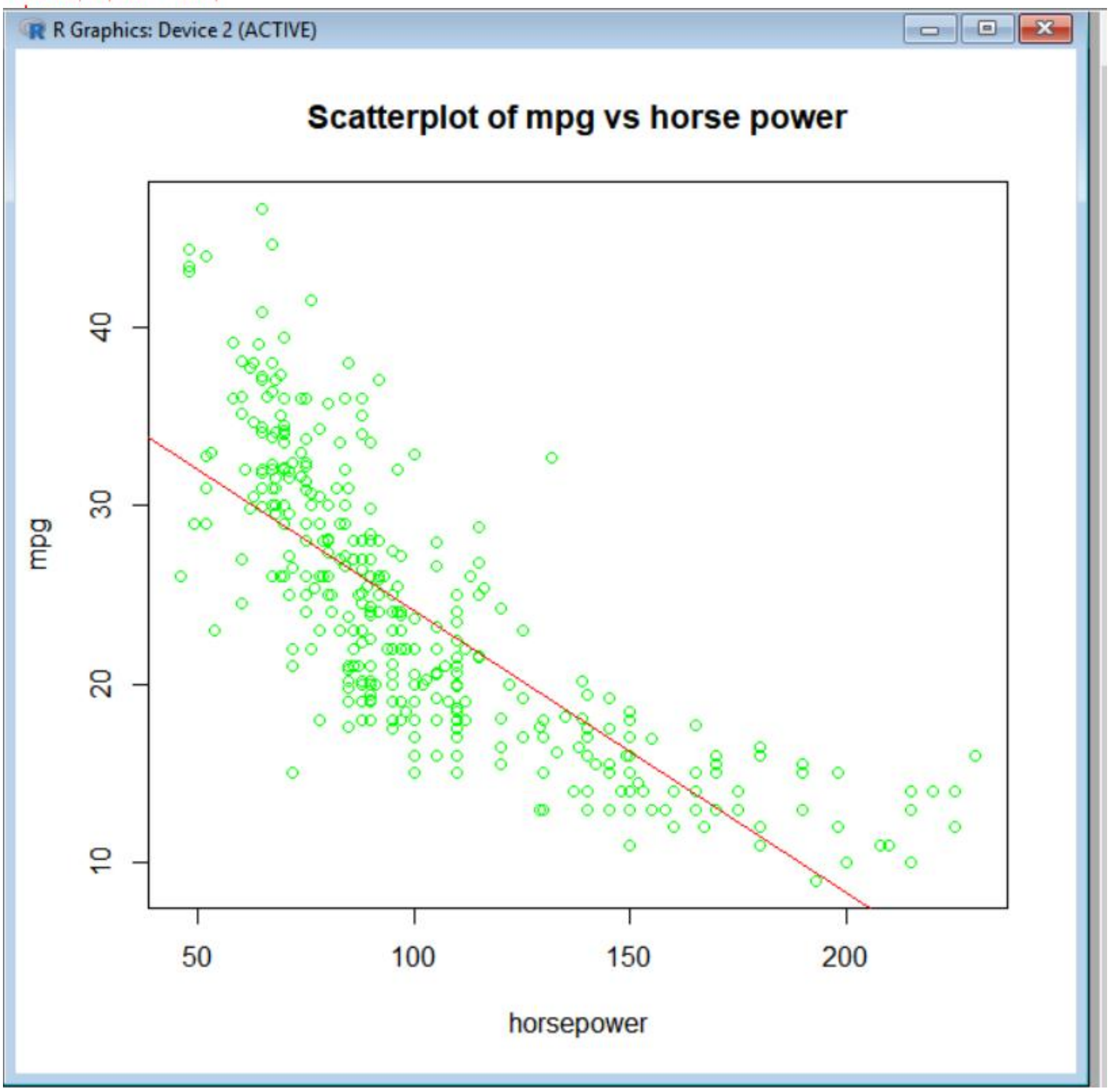
```
> library(ISLR)
> data(Auto)
> fit<-lm(mpg ~ horsepower, data = Auto)
> summary(fir)

Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

> |
```

Simple linear regression gives a model Y^=39.935861−0.157845X1Y^=39.935861−0.157845X1 between the predictor horsepower and the response mpg. A p-value of essentially zero for β^1=−0.157845β^1=−0.157845 gives very strong evidence that there is a relationship between mpg and horsepowerSince R2=0.6059R2=0.6059, approximately 60.6% of the variability in mpg is explained by a linear regression onto horsepower. This is a modest relationship between the predictor and the response, since as discussed in the chapter we can improve our R2R2 value to 0.688 by including a quadratic term. The value of β^1β^1 itself indicates that in the model each increase of 1 horsepower results on average in a decrease of 0.157845 miles per gallon. In other words, in this model there is a negative relationship between the predictor and the response.

```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

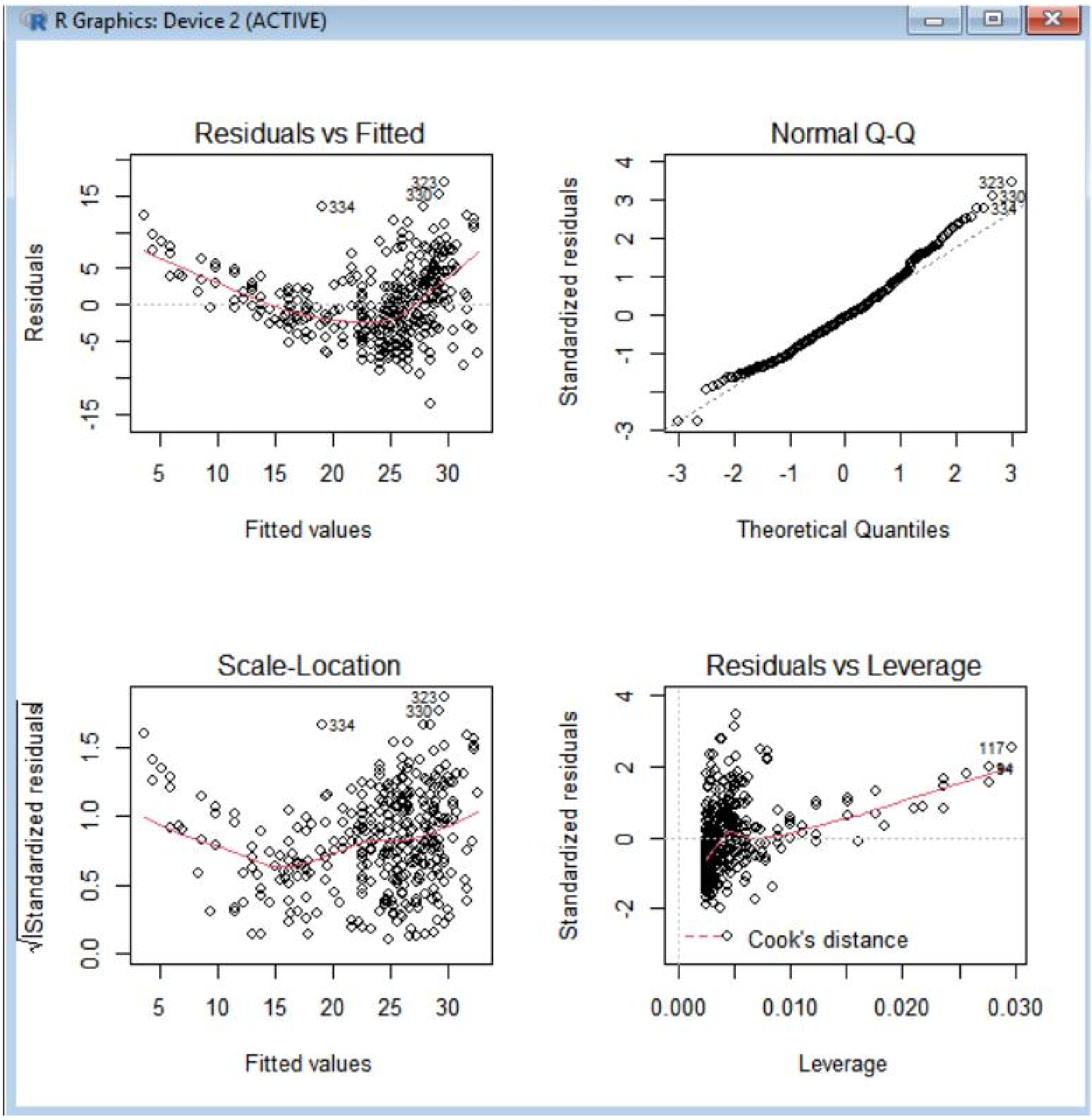
B) **Plot the response and the predictor. Use the abline() function to display the least squares regression line.**

```
> plot(Auto$horsepower, Auto$mpg , main = "Scatterplot of mpg vs horse power", xlab = "horsepower", ylab = "mpg" , col = "green")
There were 50 or more warnings (use warnings() to see the first 50)
> abline(fir, col = "red")
```



C) **Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**

```
> par(mfrow = c(2,2))
> plot(fit)
```

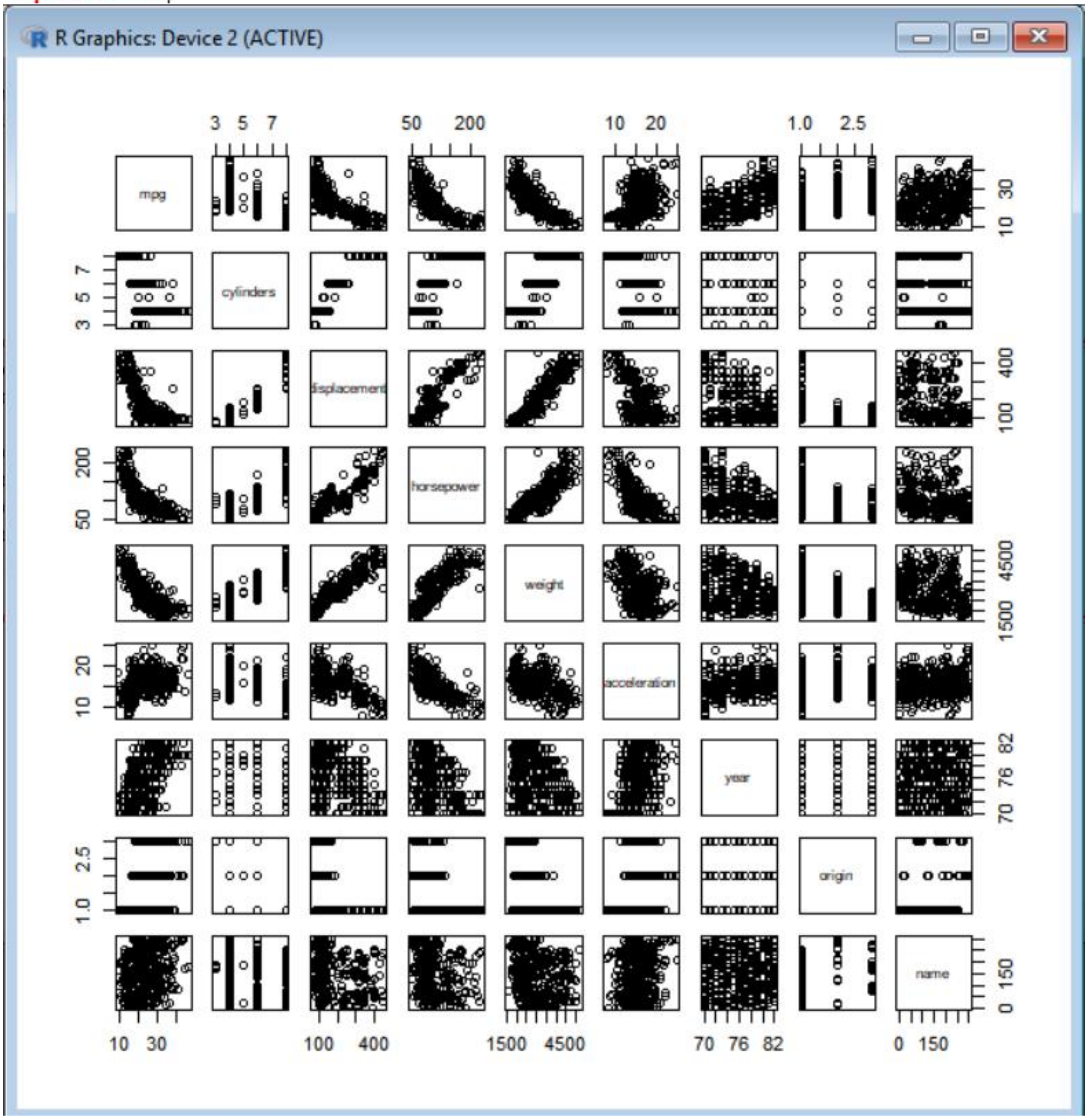


Looking at the Residuals vs. Fitted plot, there is a clear U-shape to the residuals, which is a strong indicator of non-linearity in the data. This, when combined with an inspection of the plot in Part 2, tells us that the simple linear regression model is not a good fit. In addition, when looking at the Residuals vs. Leverage plot, there are some high leverage points (remember that after dropping the rows with null values, there are 392 observations in the data set, giving an average leverage value of 2/392≈0.00512/392≈0.0051) which also have high standardized residual values (greater than 2), which is also of concern for the simple linear regression model. There are also a number of observations with a standardized residual value of 3 or more, which is evidence to suggest that they would be possibile outliers if we didn't already have the suspicion that the data is non-linear.

5. ISLR 3.7 Applied Problem 9

A) **Produce a scatterplot matrix which includes all of the variables in the data set.**

```
> pairs(Auto)
```



B) **Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.**

```
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"       "acceleration" "year"         "origin"       "name"
> |
> cor(Auto[1:8])
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

C) **Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:**

1. **Is there a relationship between the predictors and the response?**
2. **Which predictors appear to have a statistically significant relationship to the response?**
3. **What does the coefficient for the year variable suggest?**

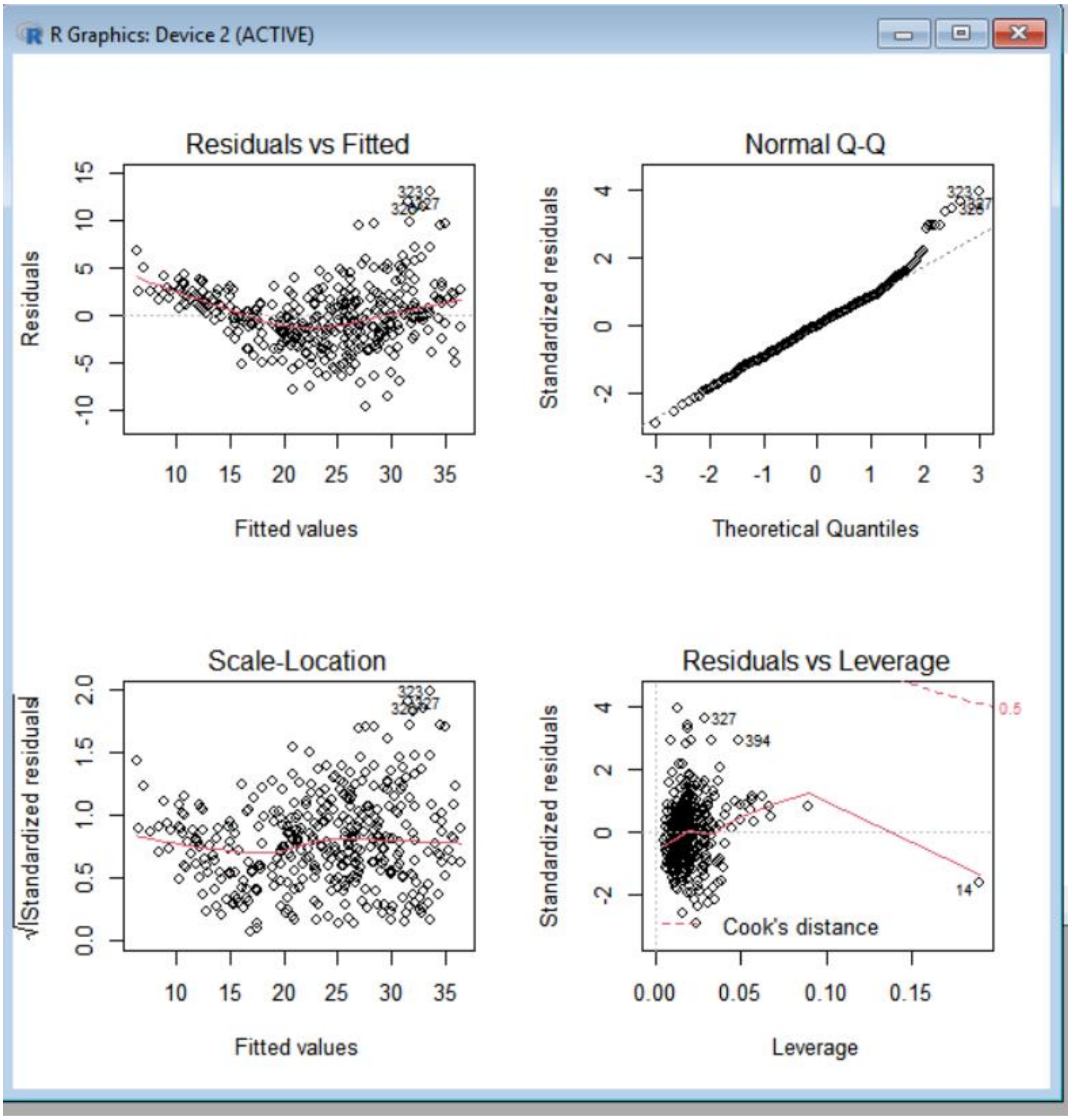
```
> fit2 <- lm(mpg ~ . - name, data = Auto)
> summary(fit2)

Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Since the F-statistic is 224.5, giving a p-value of essentially zero for the null hypothesis $H0:\beta j=0$ for all $jH0:\beta j=0$ for all j, there is strong evidence to believe that there is a relationship between the predictors and the response. The predictors that appear to have a statistically significant relationship to the response mpg are displacement with a p-value of 0.001863, and weight, year, originEuropean, and originJapanese with p-values of essentially zero. The coefficients for cylinders, horsepower, and acceleration have p-values which are not small enough to provide evidence of a statistically significant relationship to the response mpg. The coefficient of 0.777 for the year variable suggests that when we fix the number of engine cylinders, engine displacement, horsepower, weight, acceleration, and country of origin, fuel efficiency increases on average by about 0.777 miles per gallon each year. In other words, the model suggests that we would expect cars from 1971 to be more fuel efficient by 0.777 miles per gallon on average compared to equivalent cars from 1970. Also of interest are the coefficients for originEuropean and originJapanese, which suggest that compared to equivalent cars from the United States, we would expect European cars to be more fuel efficient by 2.630 miles per gallon on average, and Japanese cars to be more fuel efficient by 2.853 miles per gallon on average. Lastly, the $R2R2$ value of

0.8242 indicates that about 82% of the variation in mpg is explained by this least squares regression model.

D) **Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

```
> par(mfrow = c(2,2))
> plot(fit2)
```



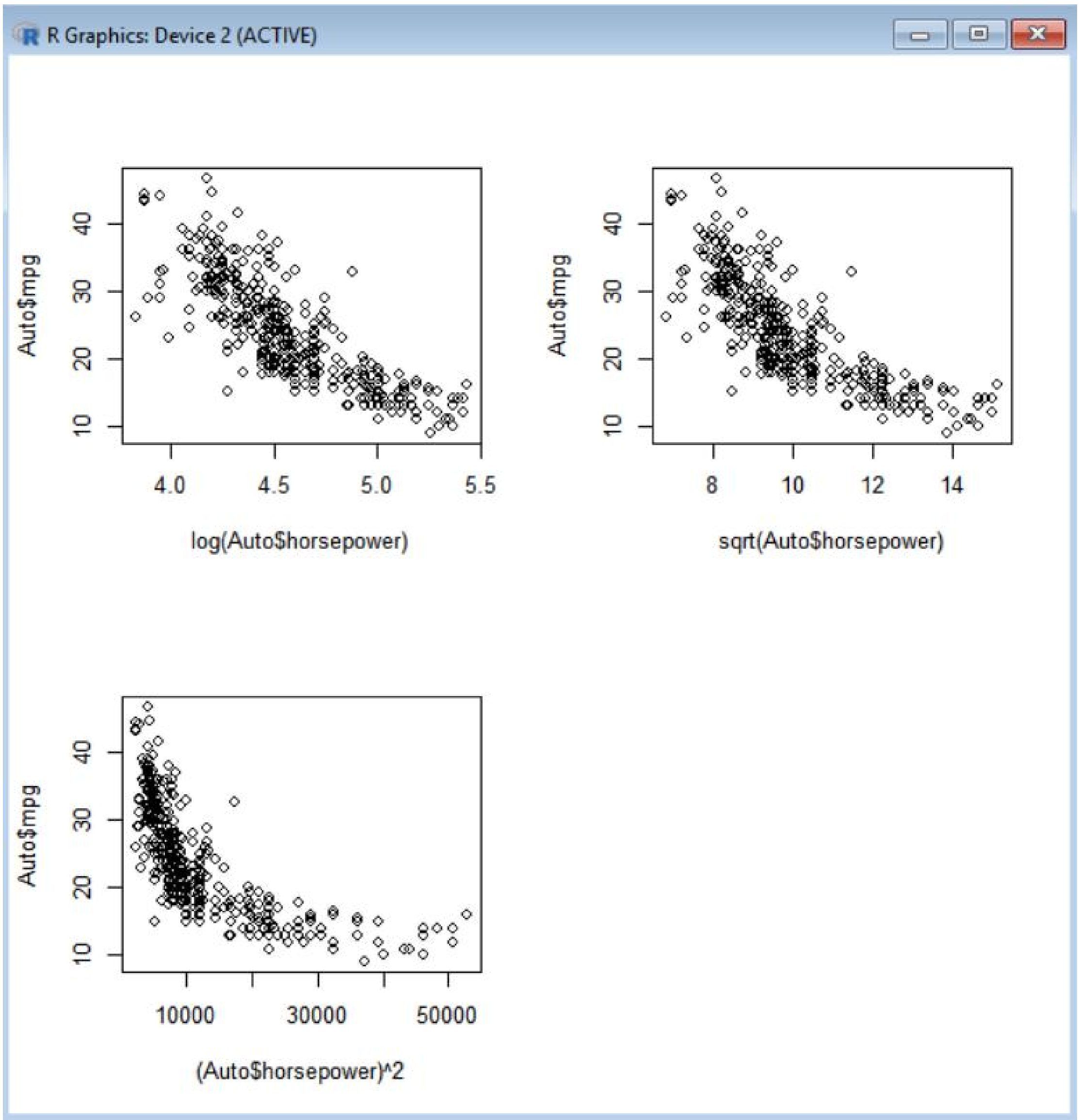
Looking at the Residuals vs. Fitted plot, there appears to be moderate U-shape, which indicates that there might be non-linearity in the data. In addition, when looking at the Residuals vs. Leverage plot we can observe a few things. First, there are a number of observations with standardized residual values with absolute value greater than or equal to 3. Those are likely outliers. This is confirmed by looking at the Scale-Location plot, which has |Standardized residual√|Standardized residual| as the yy-axis. Points with |Standardized residual√≥1.732|Standardized residual|≥1.732 have |Standardized residual|≥3|Standardized residual|≥3, which again means that they are likely outliers. Going back the the Residuals vs. Leverage plot, we also see that there are a couple points with unusually high leverage. Again remember that after dropping the rows with null values, there are 392 observations in the data set, giving an average leverage value of 9/392≈0.0239/392≈0.023. There is one point with a leverage value of about 0.10, which is almost 5 times greater than the average. There is another point with a leverage of about 0.20, which is almost 10 times greater than the average.

E) **Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**

```
> fit3 <- lm(mpg ~ cylinders*displacement +displacement *weight , data = Auto[ , 1:8])
> summary(fit3)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = Auto[, 1:8])

Residuals:
     Min       1Q   Median       3Q      Max
-13.2934  -2.5184  -0.3476   1.8399  17.7723

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322
displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

F) **Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.**

```
> par(mfrow = c(2,2))
> plot(log(Auto$horsepower), Auto$mpg)
> plot(sqrt(Auto$horsepower), Auto$mpg)
> plot((Auto$horsepower)^2, Auto$mpg)
> |
```



While the transformation did bump up the R2R2 value very slightly, it didn't really do anything to help with the residuals. This is probably due to the fact that two cars with the same 0 to 60 mile per hour time could be quite different in other ways that would affect fuel economy, such has differences in engine efficiency. For the remainder of the problem, let's turn our attention the the relationship between engine displacement and fuel efficiency. From the scatterplot, it is pretty clear that there is a nonlinear relationship between the two quantities. Let's start off by comparing a linear model to one that also includes the quadratic term.

First, we notice that none of the terms above order 2 (i.e. the cubic, quartic, and quintic terms) have statistically significant p-values. In addition, the adjusted R2R2 value has dropped slightly from 0.6872 in the quadratic model to 0.6861. Lastly, p-value from the anova() function is 0.65, which means that there is not sufficient evidence to reject the null hypothesis that the quintic model is a better fit than the quadratic one. These pieces of evidence suggest that including terms beyond order 2 does not improve the mode

6. ISLR 3.7 Applied Problem 10

A) **Fit a multiple regression model to predict Sales using Price, Urban, and US.**

```
> data(Carseats)
> fit3 <- lm(Sales ~ Price + Urban + US, data = Carseats)
> summary(fit3)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

B) **Provide an interpretation of each coefficient in the model. Be careful -- some of the variables in the model are qualitative!**

The coefficient of -0.054459 for Price means that, for a given location (i.e. fixed values of Urban and US), increasing the price of a car seat by $1 results in a decrease of sales by approximately 54.46 units, on average, in the model. The coefficient of -0.021916 for Urban Yes means that, for a given carseat price point and value of US, the model predicts urban areas to have approximately 22 fewer carseat sales on average compared to non-urban areas. The coefficient of 1.200573 for USYes means that, for a given carseat price point and value of Urban, the model predicts that stores in the United States have 1201 more carseat sales on average than stores outside the United States.

C) **Write out the model in equation form, being careful to handle the qualitative variables properly**

The model has the following equation.

Y^=13.043−0.054X1−0.022X2+1.200X3Y^=13.043−0.054X1−0.022X2+1.200X3

Here, y^y^ is the estimated carseat sales, in thousands of car seats; x1jx1j is the price of the carseat at the jth store, in dollars; and x2jx2j and x3jx3j are dummy variables to represent whether or not the jjth store at is located in an urban area and in the United States, respectively. More concretely, x2jx2j and x3jx3j use the following coding scheme.

x2jx3j={1,0,if the jth store is in an urban locationif the jth store is not in an urban location={1,0,if the jth store is in the United Statesif the jth store is not in the United States

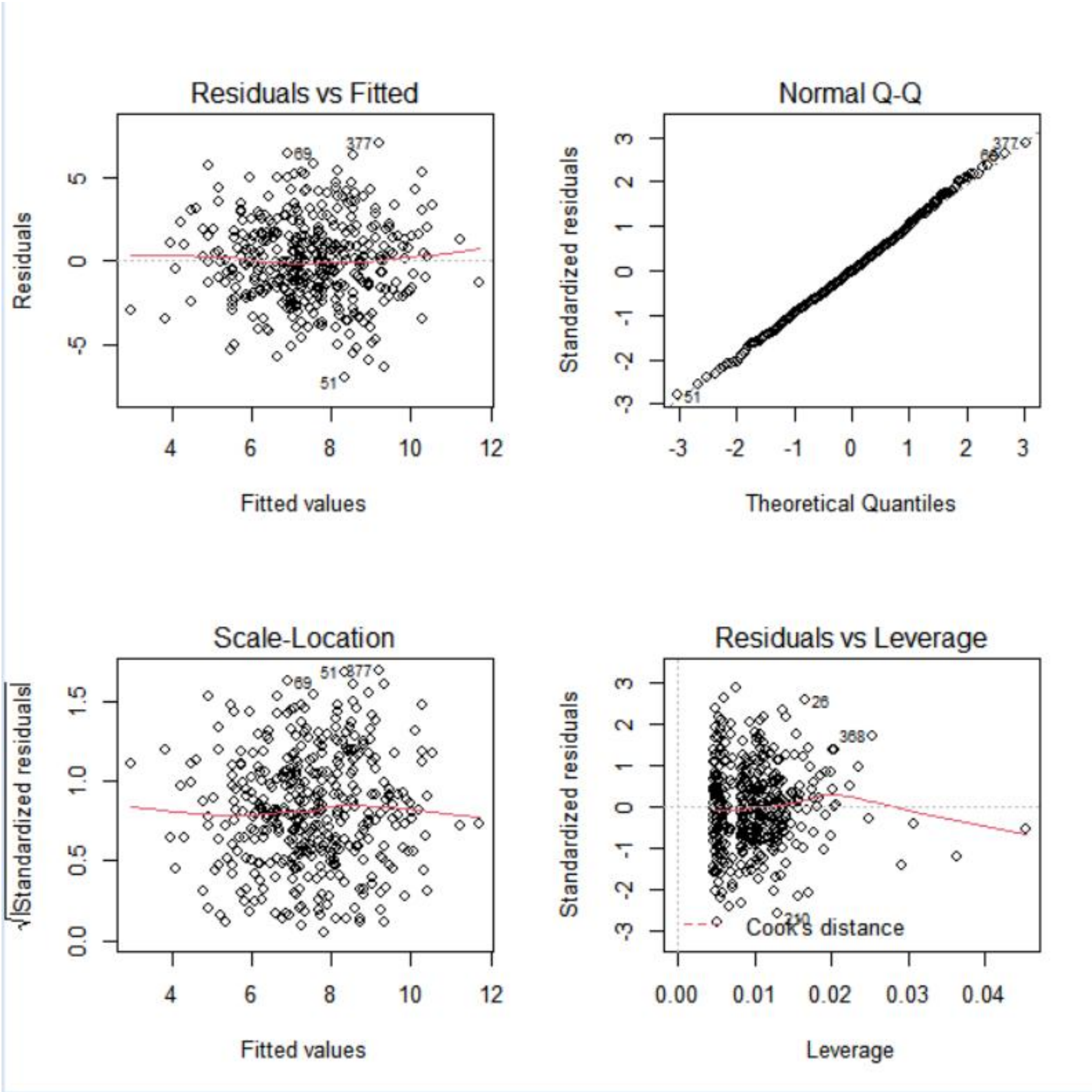
D) **For which of the predictors can you reject the null hypothesis** H0:βj=0H0:βj=0**?**
The p-values for the intercept, Price, and USYes are all essentially zero, which provides strong evidence to reject the null hypothesis H0:βj=0H0:βj=0 for those predictors. The p-value for UrbanYes, however, is 0.936, so there is no evidence to reject the null hypothesis that it has a non-zero coefficient in the true relationship between the predictors and Sales.

E) **On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.**

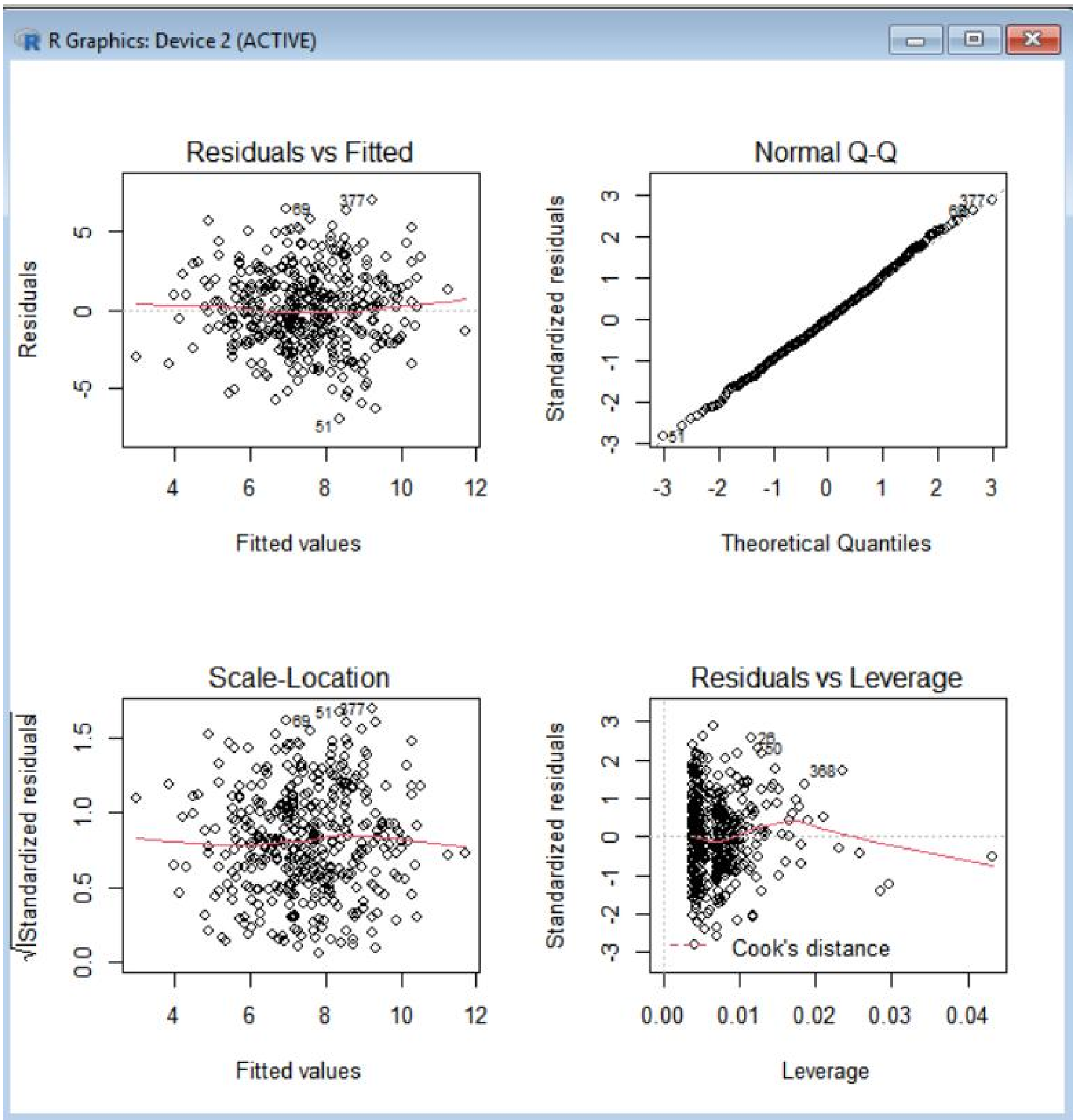
```
> fit4 <- lm(Sales ~ Price + US, data = Carseats)
> summary(fit4)

Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354 
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

How well do the models in Part 1 and Part 5 fit the data?

```
> par(mfrow = c(2, 2))
> plot(fit3)
```




```
> par(mfrow = c(2,2))
> plot(fit4)
```



The models in Part 1 and Part 5 both fit the data about equally well, with identical R2R2 values of 0.2393. In addition, when comparing the diagnostic plots between the two models, there isn't any discernible visual differences that would strongly indicate that one model is a better fit than the other.

G)**Using the model from Part 5, obtain 95% confidence intervals for the coefficient(s).**

```
> confint(fit4)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

H) **Is there evidence of outliers or high leverage observations in the model from Part 5**

When we look at the residuals vs. leverage plot for the model from Part 5 that I generated in Part 6, we see that there are a number of observations with standardized residuals close to 3 in absolute value. Those observations are possible outliers. We can also see in the same plot that there are number of high leverage points with leverage values greatly exceeding the average leverage of 3/400=0.00753/400=0.0075, though those high leverage observations are not likely outliers, as they have studentized residual values with absolute value less than 2.