Chapter - 2

Q8. This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US

(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.

```
college <- read.csv("College.csv")
```

(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.

```
head(college[, 1:5])
```

```
##                              X Private Apps Accept Enroll
## 1 Abilene Christian University     Yes 1660   1232    721
## 2           Adelphi University     Yes 2186   1924    512
## 3               Adrian College     Yes 1428   1097    336
## 4          Agnes Scott College     Yes  417    349    137
## 5    Alaska Pacific University     Yes  193    146     55
## 6            Albertson College     Yes  587    479    158
```

```
rownames <- college[, 1]
college <- college[, -1]
head(college[, 1:5])
```

```
##   Private Apps Accept Enroll Top10perc
## 1     Yes 1660   1232    721        23
## 2     Yes 2186   1924    512        16
## 3     Yes 1428   1097    336        22
## 4     Yes  417    349    137        60
## 5     Yes  193    146     55        16
## 6     Yes  587    479    158        38
```

(c)

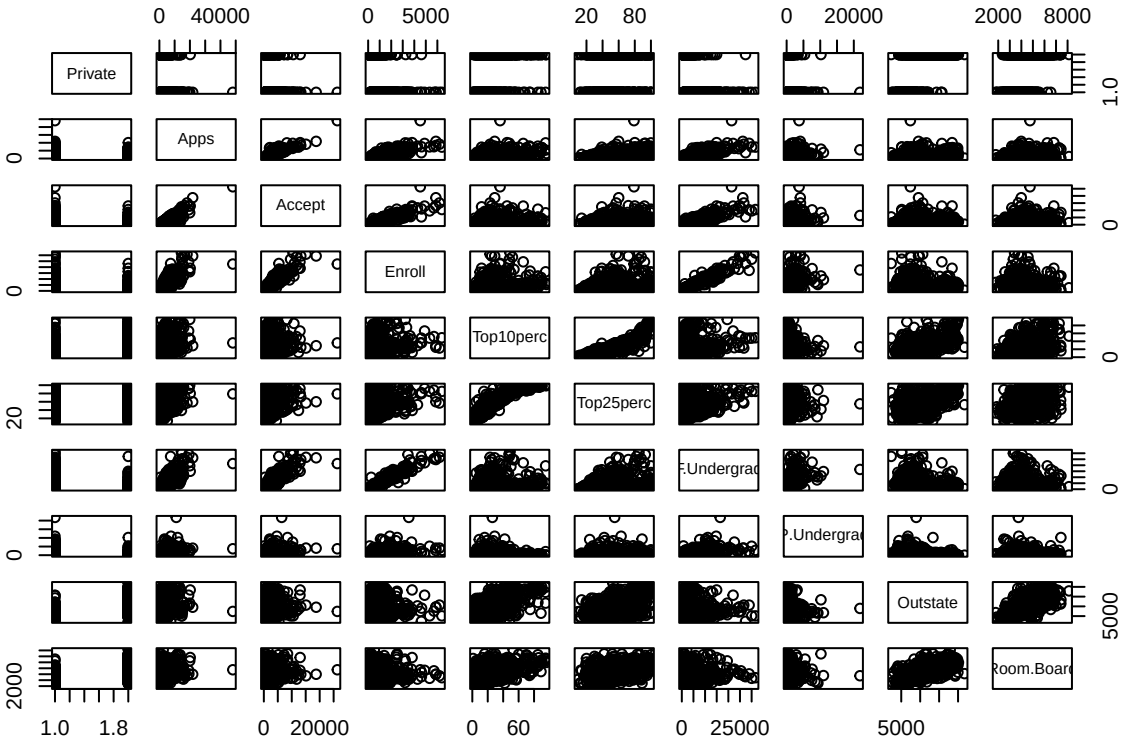
i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
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
```

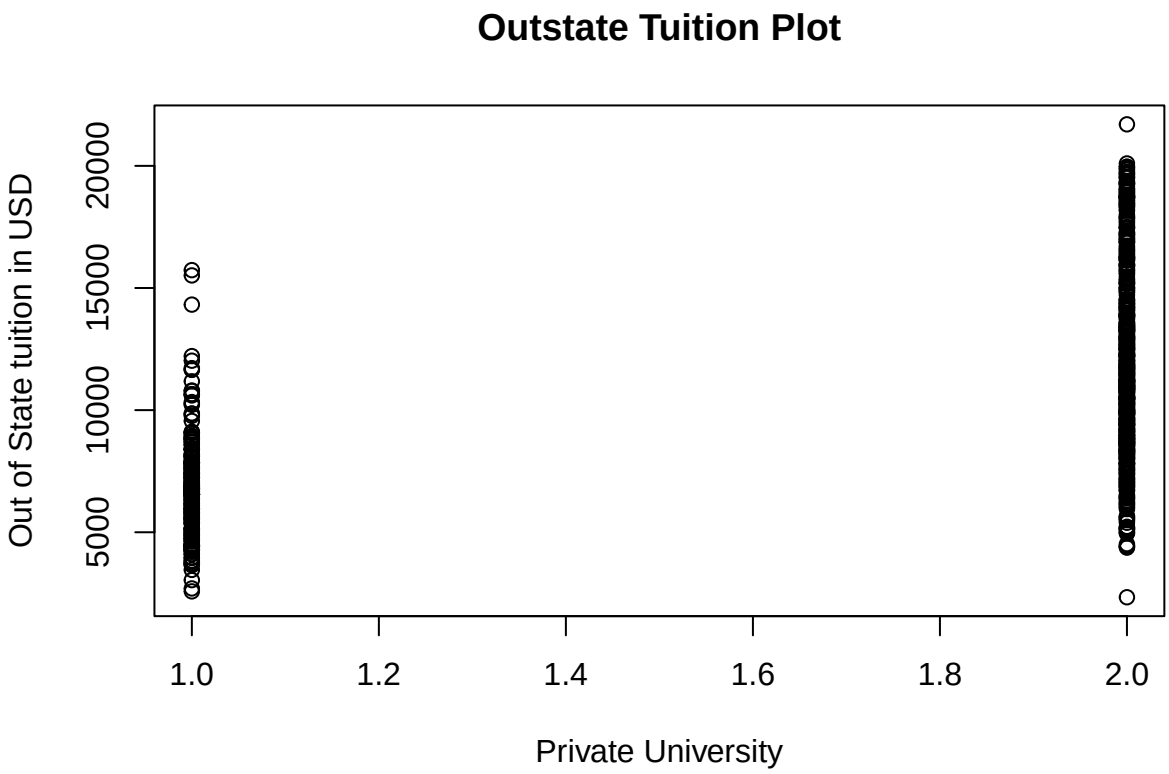
ii.Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.

```
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 University", ylab ="Out of State tuition in USD
```

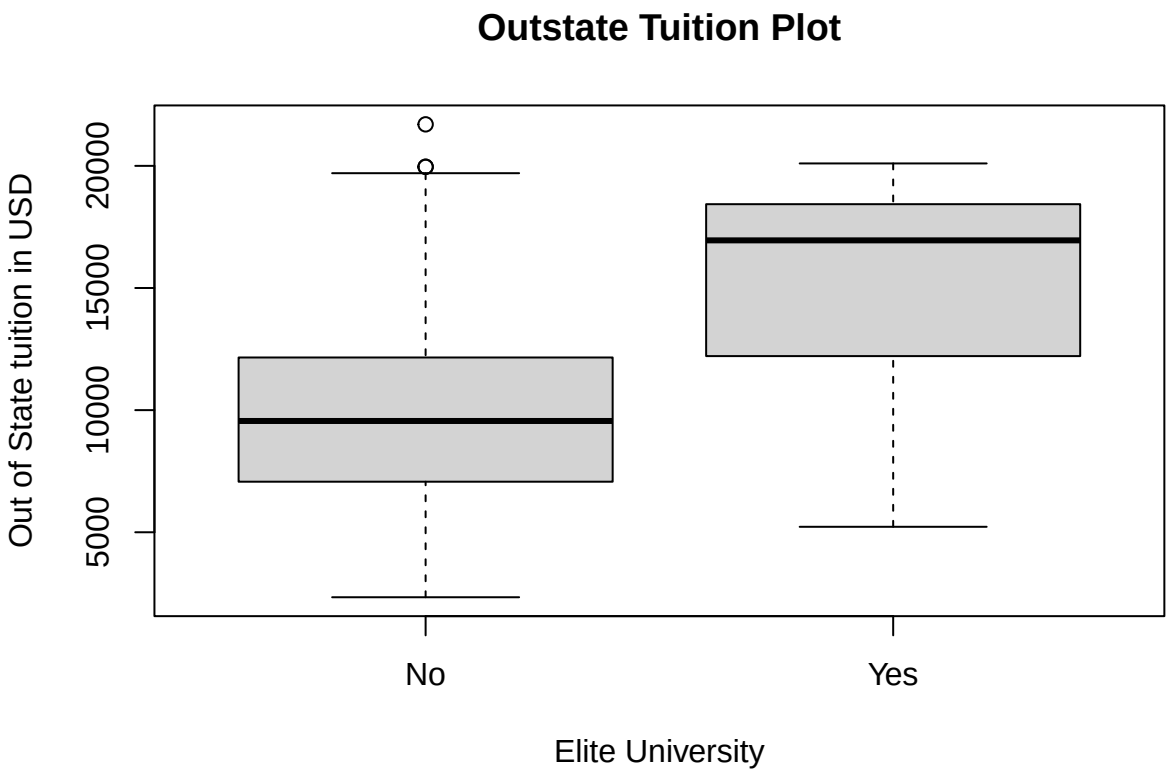


iv. Create a new qualitative variable, called "Elite", by binning the "Top10perc" variable. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of "Outstate" versus "Elite".

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

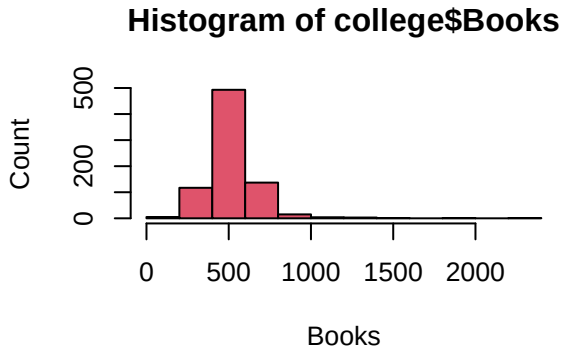
```
##  No Yes
## 699  78
```

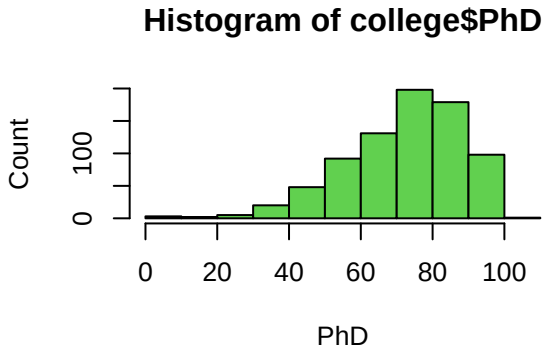
```
plot(college$Elite, college$Outstate, xlab = "Elite University", ylab ="Out of State tuition in USD", ma
```

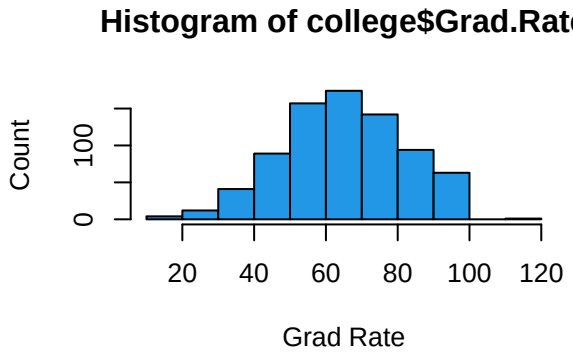


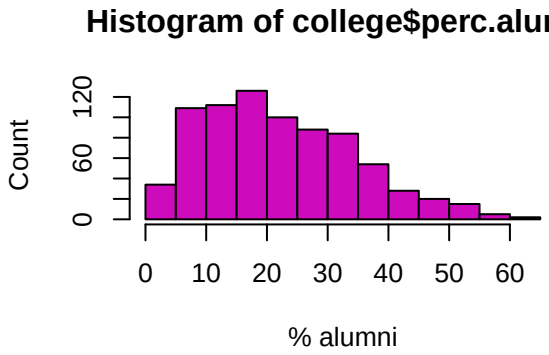
v. Use the hist() function to produce some histograms with numbers of bins for a few of the quantitative variables.

```
par(mfrow = c(2,2))
hist(college$Books, col = 2, xlab = "Books", ylab = "Count")
hist(college$PhD, col = 3, xlab = "PhD", ylab = "Count")
hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count")
hist(college$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count")
```









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

```
summary(college$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

It is a little weird to have universities with 103% of faculty with Phd's, let us see how many universities have this percentage and their names.

```
weird.phd <- college[college$PhD == 103, ]
nrow(weird.phd)
```

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

```
rownames[as.numeric(rownames(weird.phd))]
```

```
## [1] "Texas A&M University at Galveston"
```

Q9. This exercise involves the "Auto" data set studied in the lab. Make sure the missing values have been removed from the data.

a. Which of the predictors are quantitative, and which are qualitative ?

```r
auto <- read.csv("Auto.csv", na.strings = "?")
auto <- na.omit(auto)
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 rebe
##  - attr(*, "na.action")= 'omit' Named int [1:5] 33 127 331 337 355
##   ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

b. What is the range of each quantitative predictor ?

```r
summary(auto[, -c(4, 9)])
```

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

c. What is the mean and standard deviation of each quantitative predictor ?

```r
sapply(auto[, -c(4, 9)], mean)
```

```
##          mpg    cylinders displacement       weight acceleration         year
##    23.445918     5.471939   194.411990  2977.584184    15.541327    75.979592
##       origin
##     1.576531
```

```r
sapply(auto[, -c(4, 9)], sd)
```

```
##          mpg    cylinders displacement       weight acceleration         year
##    7.8050075    1.7057832  104.6440039  849.4025600    2.7588641    3.6837365
##       origin
##    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 ?

```
subset <- auto[-c(10:85), -c(4,9)]
sapply(subset, range)
```

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

```
sapply(subset, mean)
```

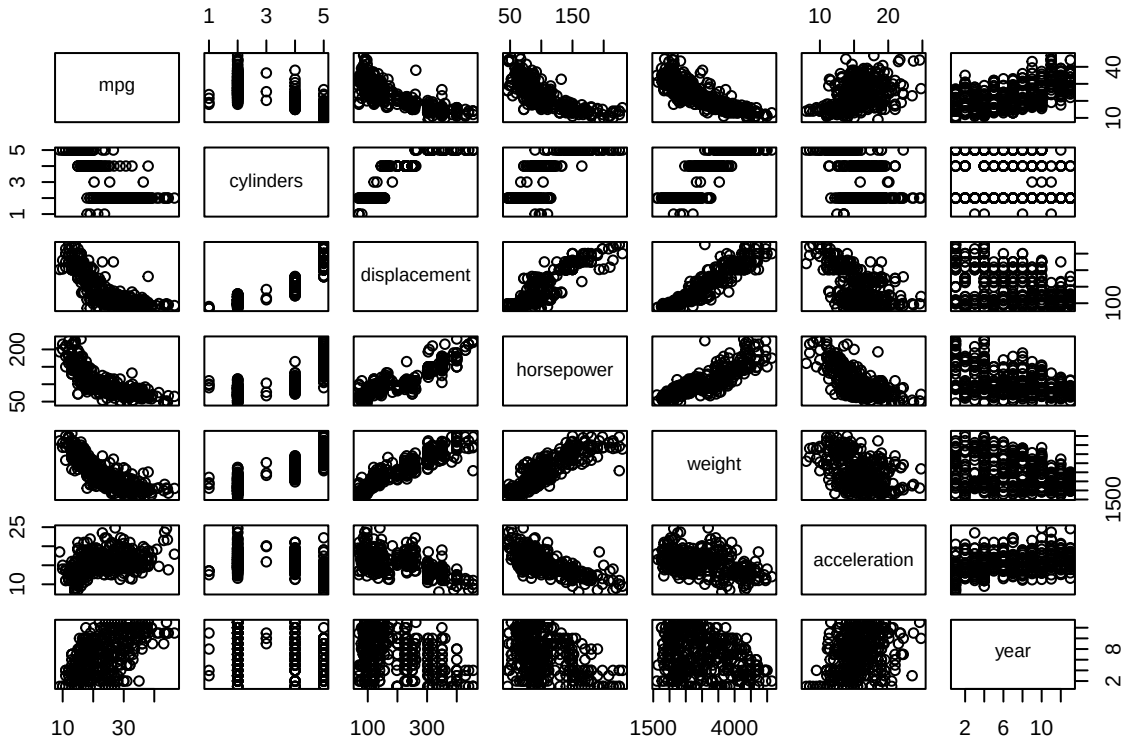
```
##          mpg    cylinders displacement       weight acceleration         year
##    24.404430     5.373418   187.240506  2935.971519    15.726899    77.145570
##       origin
##     1.601266
```

```
sapply(subset, sd)
```

```
##          mpg    cylinders displacement       weight acceleration         year
##     7.867283     1.654179    99.678367   811.300208     2.693721     3.106217
##       origin
##     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.

```
auto$cylinders <- as.factor(auto$cylinders)
auto$year <- as.factor(auto$year)
auto$origin <- as.factor(auto$origin)
pairs(auto[,1:7])
```



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

```
auto$horsepower <- as.numeric(auto$horsepower)
cor(auto$weight, auto$horsepower)
```

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

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

```
## [1] 0.9329944
```

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

```
## [1] 0.897257
```

Q10. This exercise involves the "Boston" housing data set.

a. To begin, load in the "Boston" data set.

```
library(MASS)
Boston$chas <- as.factor(Boston$chas)
nrow(Boston)
```

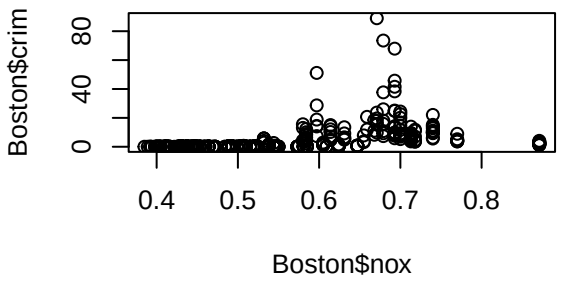
```
## [1] 506
```

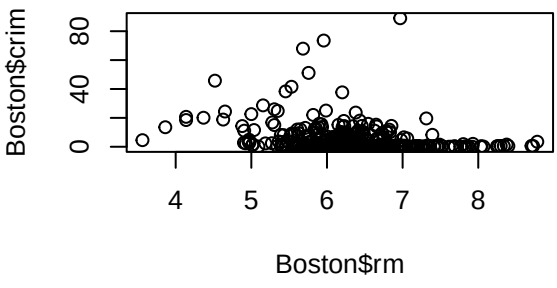
```
ncol(Boston)
```

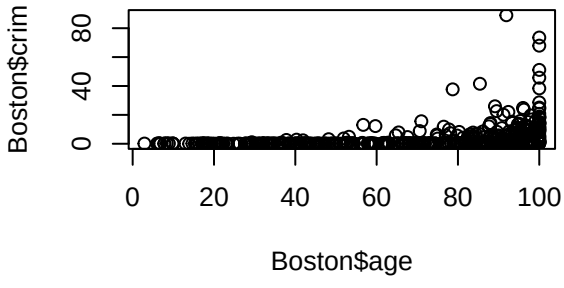
```
## [1] 14
```

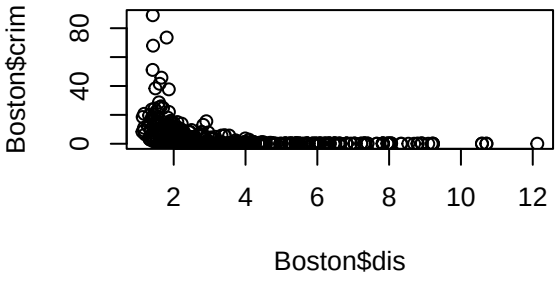
b. Make some pairwise scatterplots of the predictors in this data set.

```
par(mfrow = c(2, 2))
plot(Boston$nox, Boston$crim)
plot(Boston$rm, Boston$crim)
plot(Boston$age, Boston$crim)
plot(Boston$dis, Boston$crim)
```



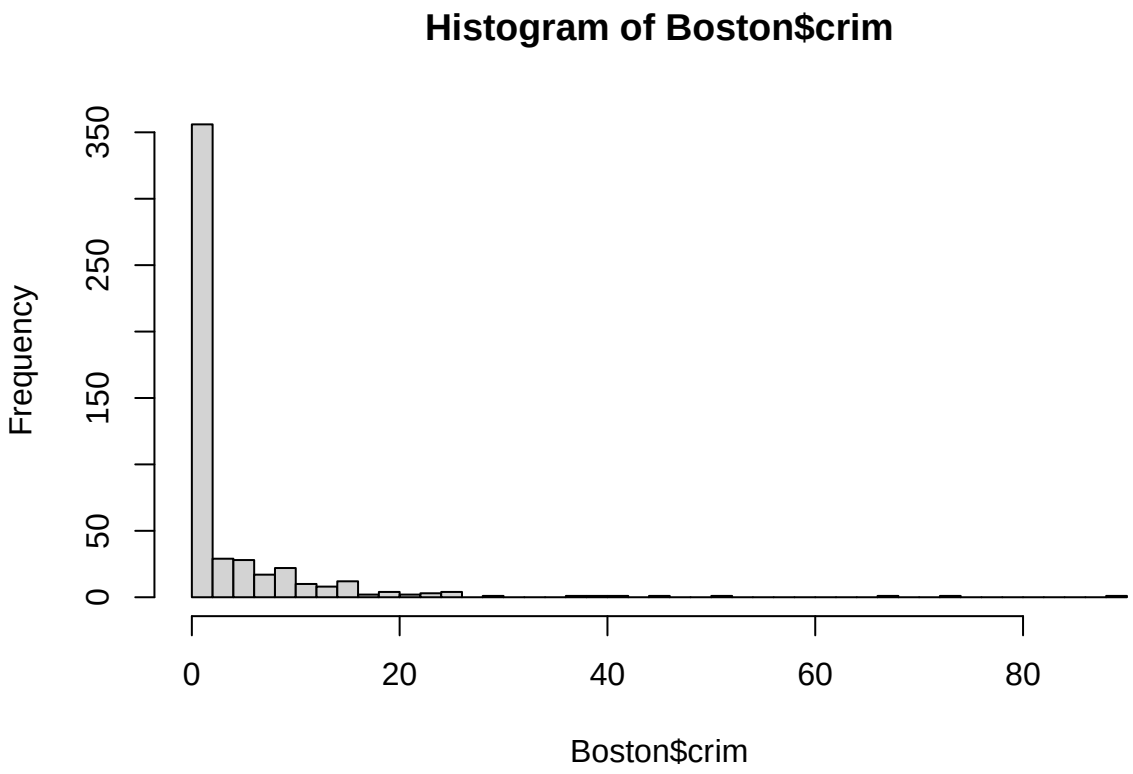






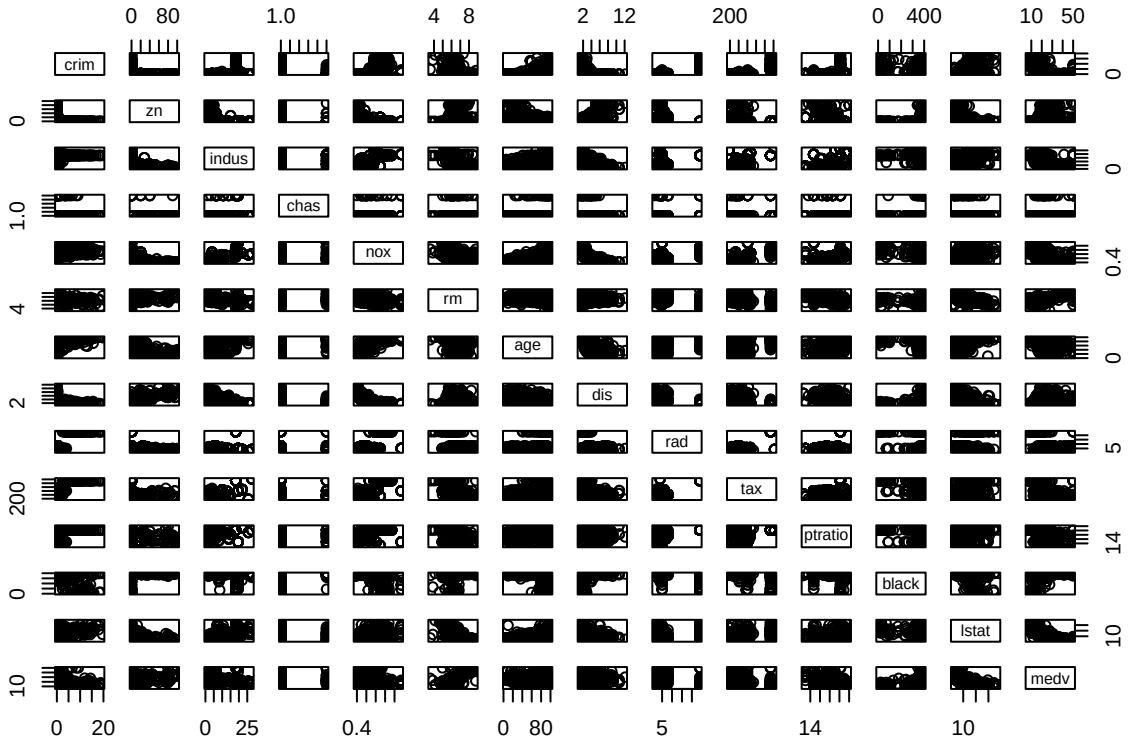
c. Are any of the predictors associated with per capita crime rate ?

```
hist(Boston$crim, breaks = 50)
```



Most suburbs do not have any crime (80% of data falls in crim < 20).

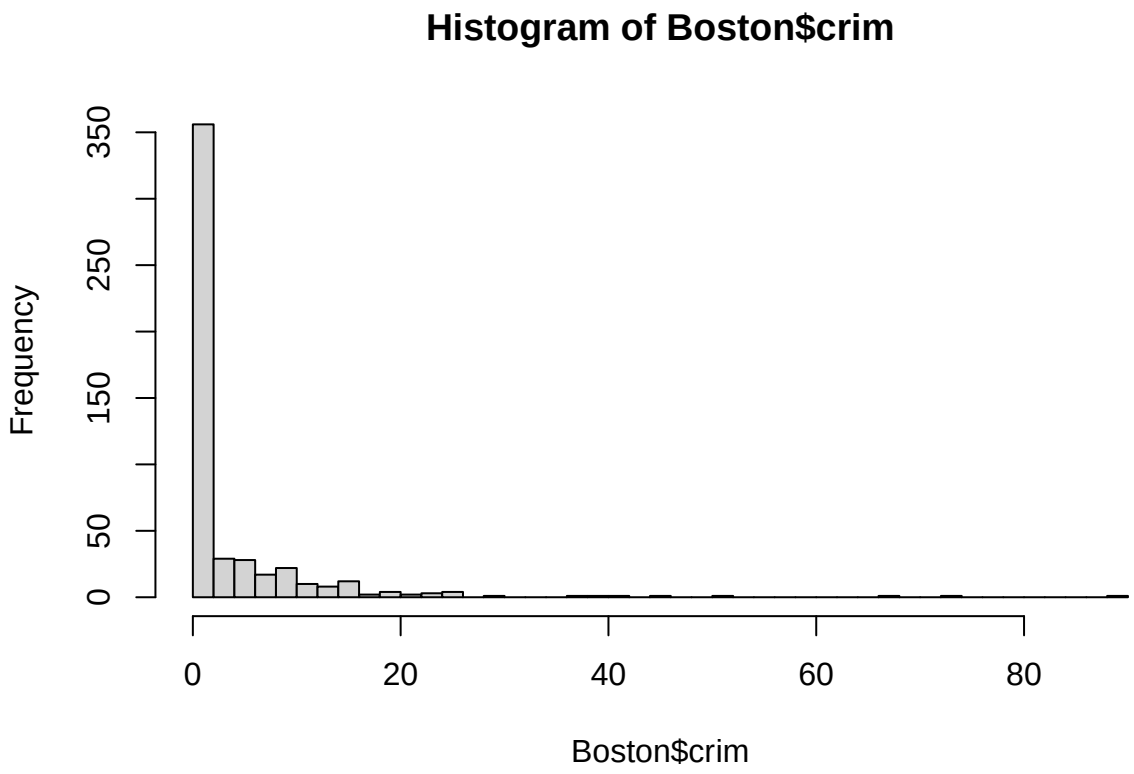
```
pairs(Boston[Boston$crim < 20, ])
```



There may be a relationship between crim and nox, rm, age, dis, lstat and medv.

d. Do any of the suburbs of Boston appear to have particularly high crime rates ? Tax rates ? Pupil-teacher ratios ?

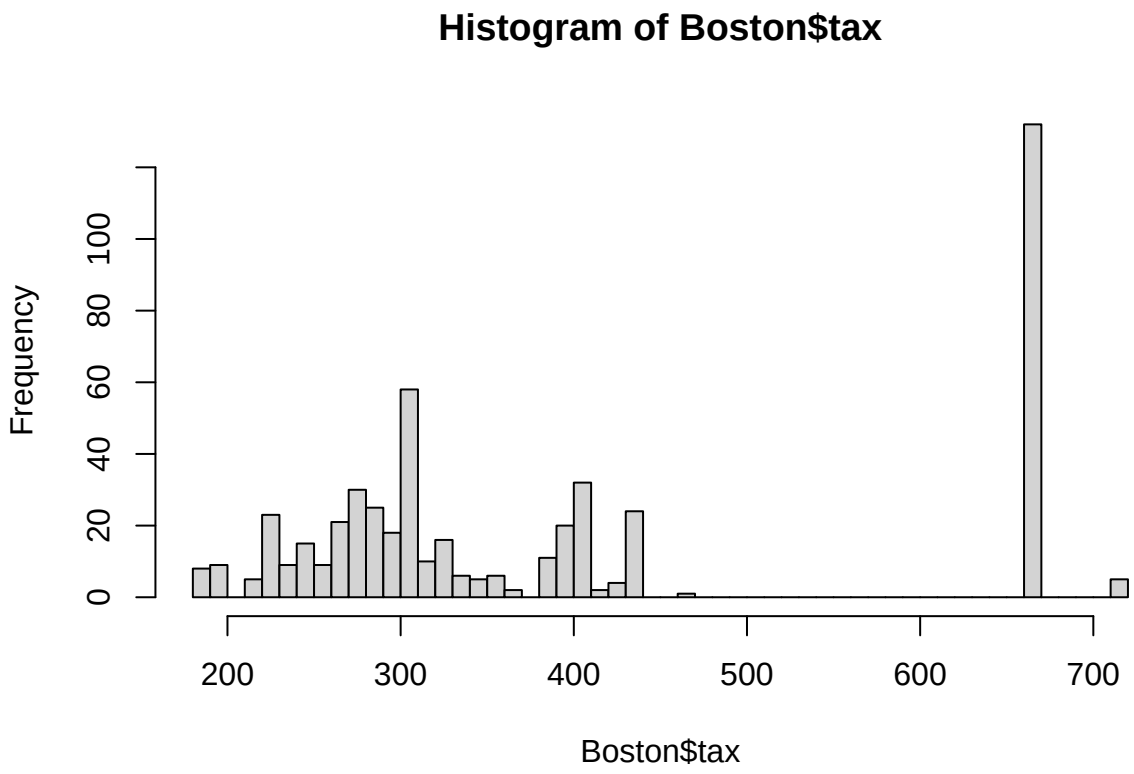
```
hist(Boston$crim, breaks = 50)
```




```
nrow(Boston[Boston$crim > 20, ])
```

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

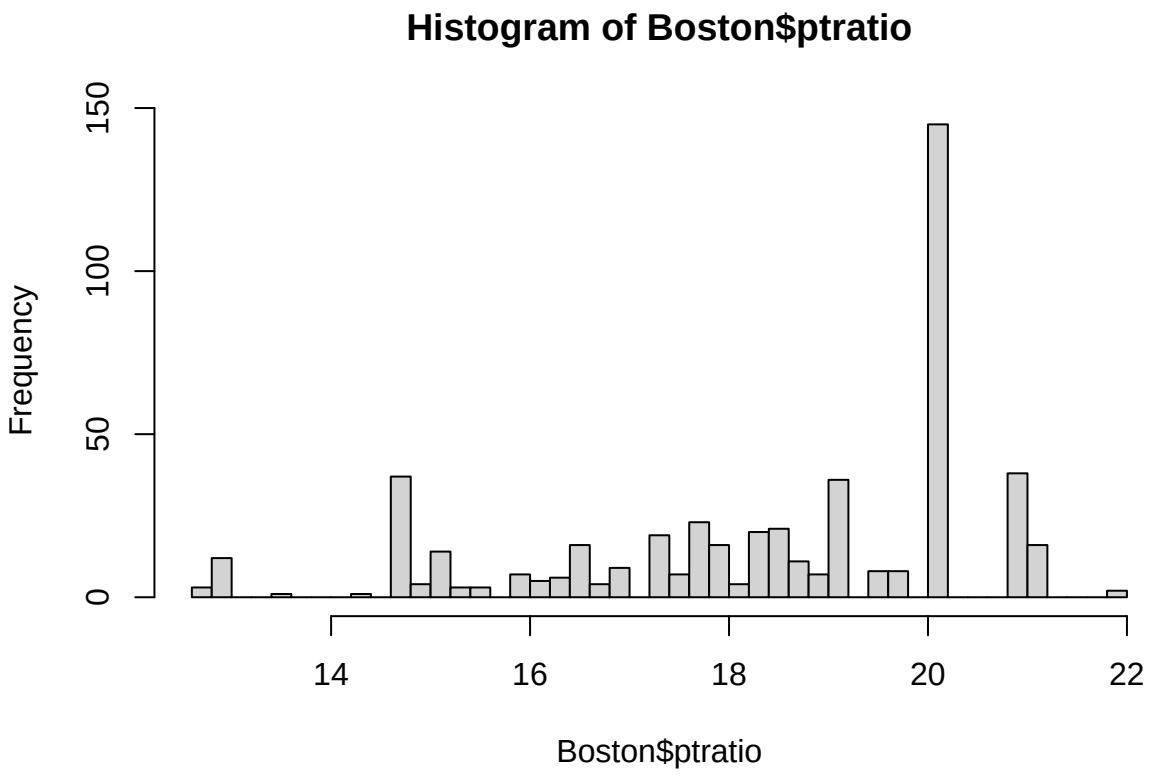
```
hist(Boston$tax, breaks = 50)
```




```
nrow(Boston[Boston$tax == 666, ])
```

```
## [1] 132
```

```
hist(Boston$ptratio, breaks = 50)
```




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

```
## [1] 201
```

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 the towns in this data set ?

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

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

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

```
row.names(Boston[min(Boston$medv), ])
```

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

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

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

```
Boston[min(Boston$medv), ]$tax
```

```
## [1] 222
```

h. In this data set, how many of the suburbs average more than seven rooms per dwelling ? More than eight rooms per dwelling ?

```
nrow(Boston[Boston$rm > 7, ])
```

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

```
nrow(Boston[Boston$rm > 8, ])
```

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