HW2_Gonzalez_Jadyn

Jadyn Gonzalez

9/9/2019

Chapter 2 Questions 8 and 10

Question 8

- a) Lets import and view our data set

```
setwd("~/Documents/MGSC310")
college = read.csv("College.csv")
```

- b) We'll use the head function in R to view the data.

```
head(college, 5)
```

```
##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
## 5    Alaska Pacific University     Yes  193    146     55        16
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
## 5       72      11.9           2  10922        15
```

```
#set rownames to be the first element of each row
rownames(college) = college[,1]
#dop the first column of the set
college = college[,-1]
head(college, 5)
```

```
##                             Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
```

```
##                              Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University        52        2885         537     7440
## Adelphi University                  29        2683        1227    12280
## Adrian College                      50        1036          99    11250
## Agnes Scott College                 89         510          63    12960
## Alaska Pacific University           44         249         869     7560
##                              Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
## Agnes Scott College                5450   450      875  92       97
## Alaska Pacific University          4120   800     1500  76       72
##                              S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
## Agnes Scott College                7.7          37  19016        59
## Alaska Pacific University         11.9           2  10922        15
```

We see that R has given the rownames the value that was in the first column of the original dataset, our new first column values are now 'Private'

- c) Let's plot and run some statistics.

1. A summary of the data

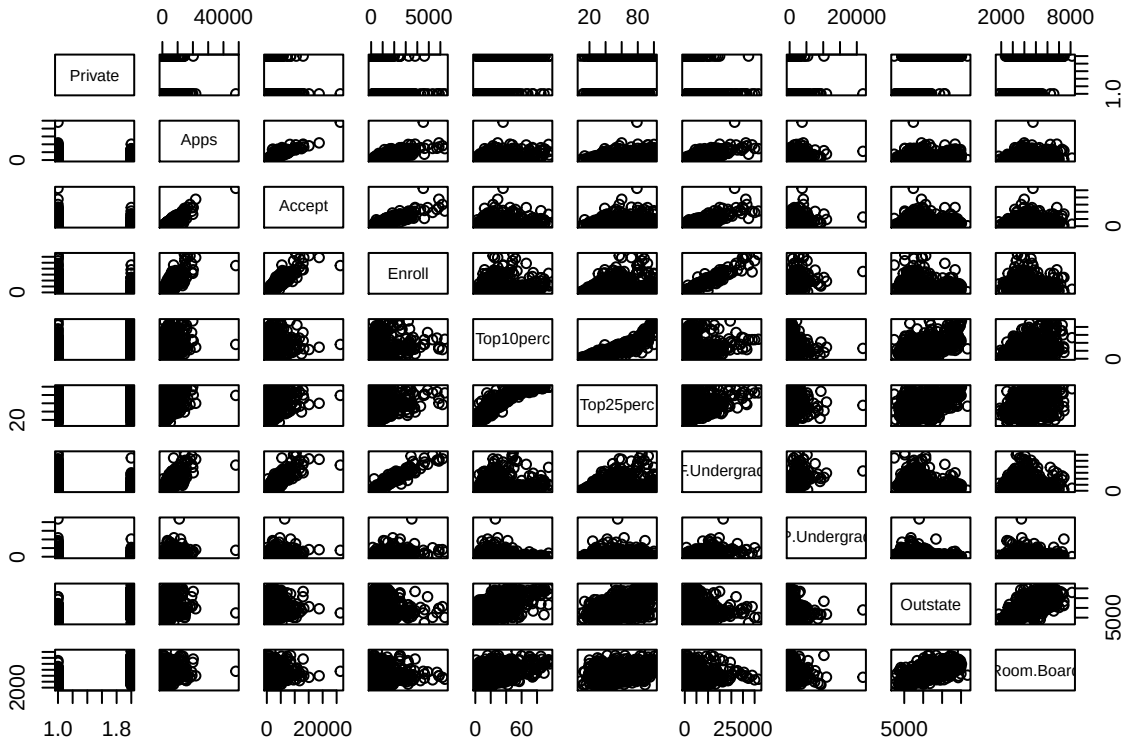
```
summary(college)
```

```
##  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
```

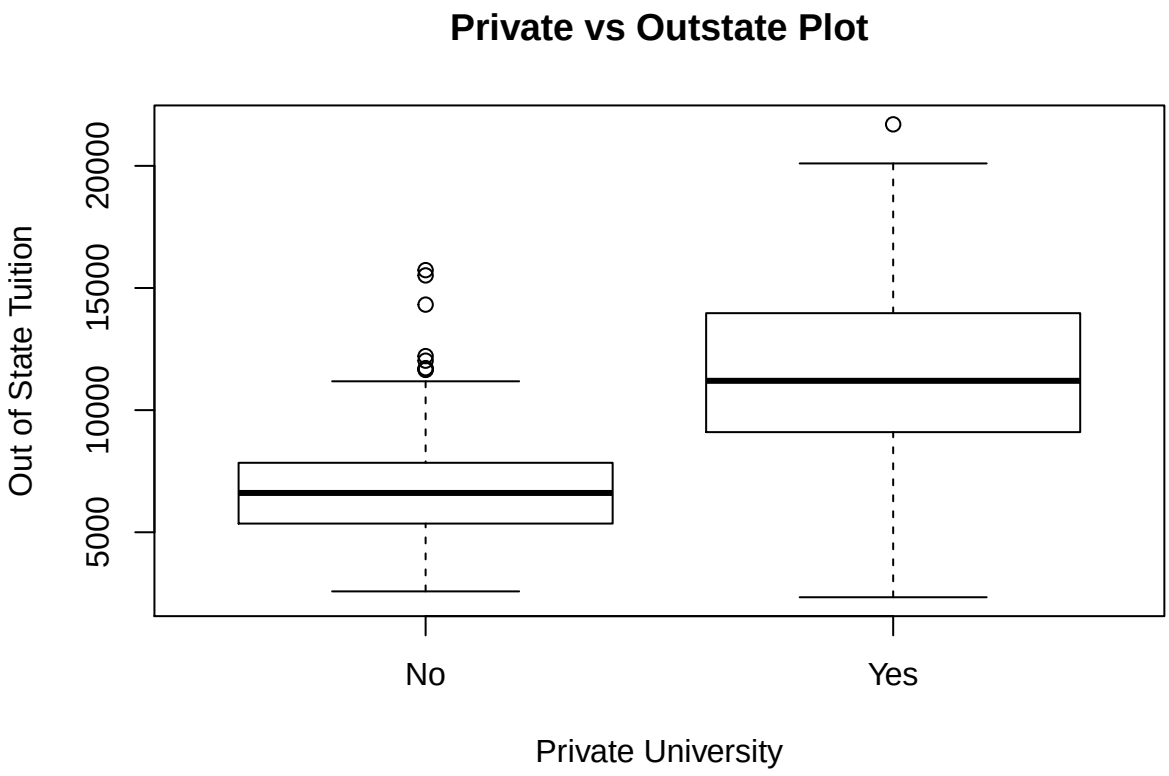
2. A scatterplot matrix of the first 10 attributes

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



3. Side-by-side boxplots of Outstate vs. Private

```
plot(college$Private, college$Outstate, xlab = "Private University", ylab = "Out of State Tuition", mai
```

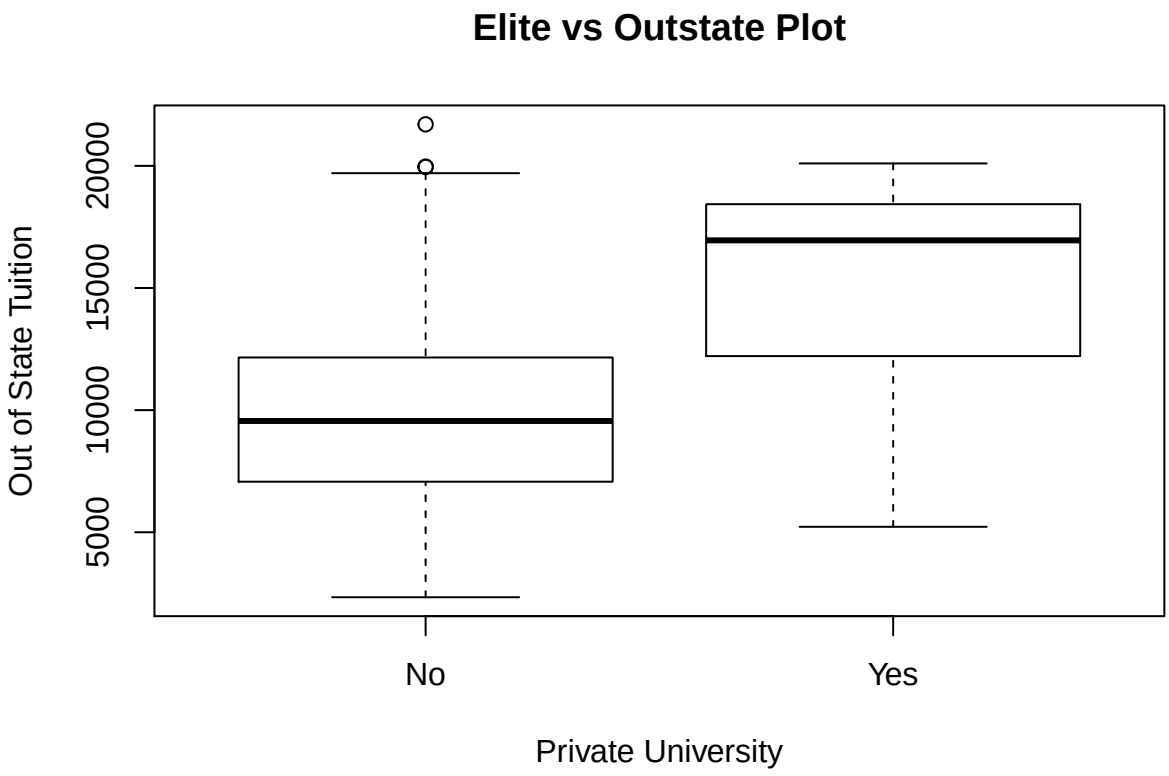


4. Create a new attribute 'Elite' and observe the count of Elite universities, then a plot between Outstate and Elite

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

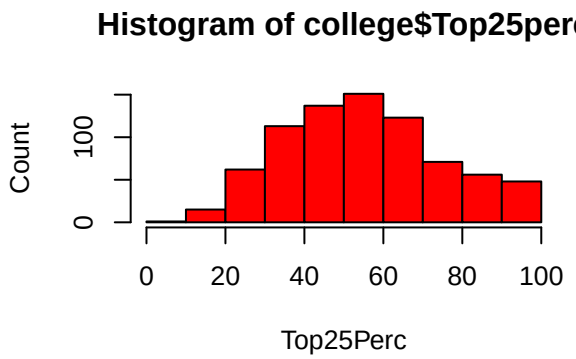
```
##  No Yes
## 699  78
```

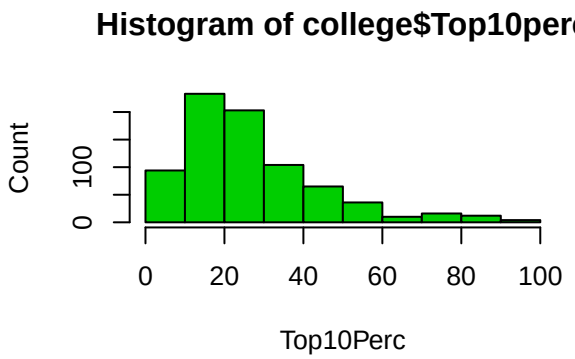
```
plot(college$Elite, college$Outstate, xlab = "Private University", ylab = "Out of State Tuition", main =
```

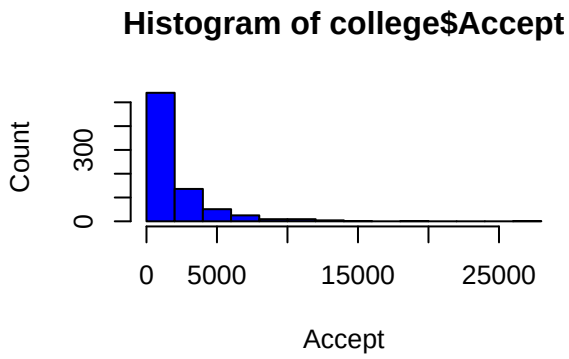


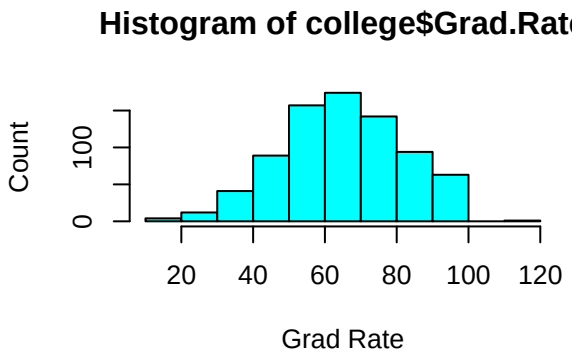
5. Some histograms of a few quantitative attributes

```
par(mfrow = c(2,2))
hist(college$Top25perc, col = 2, xlab = "Top25Perc", ylab = "Count")
hist(college$Top10perc, col = 3, xlab = "Top10Perc", ylab = "Count")
hist(college$Accept, col = 4, xlab = "Accept", ylab = "Count")
hist(college$Grad.Rate, col = 5, xlab = "Grad Rate", ylab = "Count")
```









6. Let's explore a bit more

```
summary(college$Top25perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    41.0    54.0    55.8    69.0   100.0
```

```
summary(college$Top10perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   15.00   23.00   27.56   35.00   96.00
```

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

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

```
summary(college$Grad.Rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   53.00   65.00   65.46   78.00  118.00
```

Here we can see that PhD and Grad.Rate exceed 100%, how is that possible? Let's look into this a little further.

```
#we use 103 and 118 respectivly as that was the output from summary()
anom_phd = college[college$PhD == 103,]
nrow(anom_phd)
```

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

```
anom_phd
```

```
##                                 Private Apps Accept Enroll Top10perc
## Texas A&M University at Galveston      No  529    481    243        22
##                                 Top25perc F.Undergrad P.Undergrad
## Texas A&M University at Galveston        47        1206         134
##                                 Outstate Room.Board Books Personal PhD
## Texas A&M University at Galveston     4860       3122   600      650 103
##                                 Terminal S.F.Ratio perc.alumni Expend
## Texas A&M University at Galveston       88      17.4          16   6415
##                                 Grad.Rate Elite
## Texas A&M University at Galveston        43    No
```

```
anom_grad = college[college$Grad.Rate == 118,]
nrow(anom_grad)
```

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

```
anom_grad
```

```
##                   Private Apps Accept Enroll Top10perc Top25perc
## Cazenovia College     Yes 3847   3433    527         9        35
##                   F.Undergrad P.Undergrad Outstate Room.Board Books
## Cazenovia College        1010          12     9384       4840   600
##                   Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Cazenovia College      500  22       47      14.3          20   7697
##                   Grad.Rate Elite
## Cazenovia College       118    No
```

Ok, so we see PhD for Texas is 103 and Grad.Rate for Cazenovia is 118. This is probably some mistyped data that would have to be cleaned to continue further working with the dataset.

Question 10

- a) Let's load and read about the dataset

```
library(MASS)
head(Boston, 5)
```

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

```
?Boston
```

We can see how many rows and columns 'Boston' has by using the dim() function

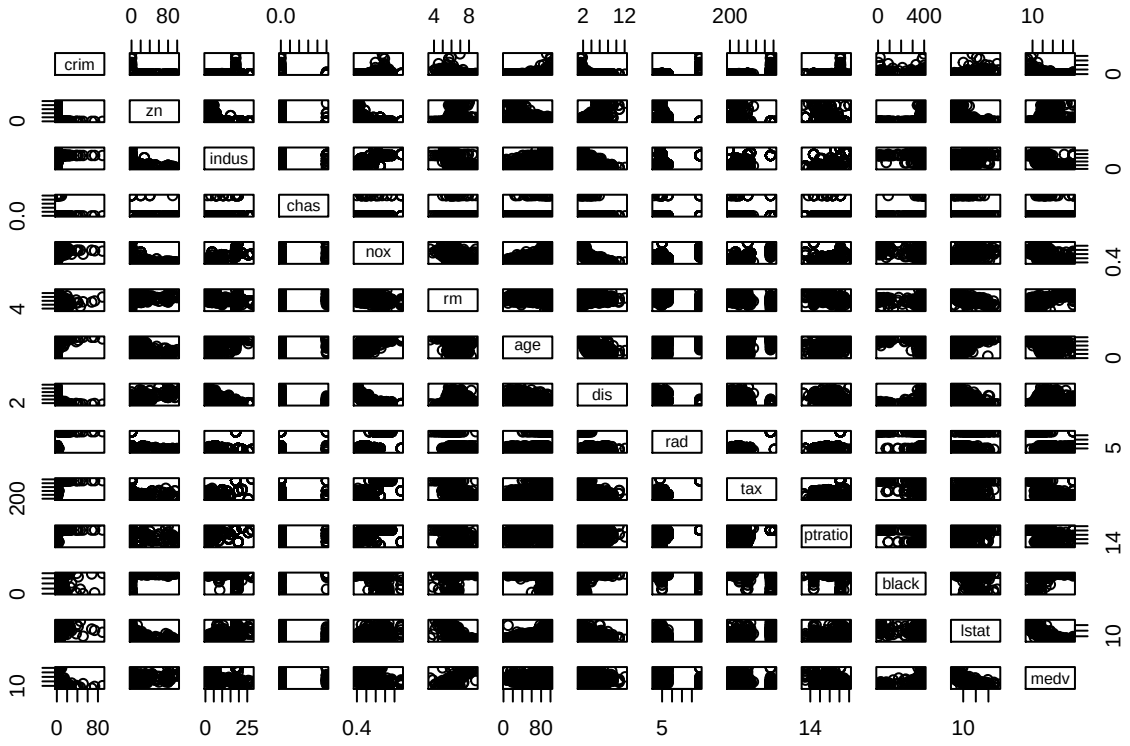
```
dim(Boston)
```

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

We see it has 506 rows and 14 columns, What they represent can be found by reading the output using '?Boston'

- b) Here are some scatterplots af a few predictors

```
pairs(Boston)
```



From this it looks like crim may contain some outliers

- c) Let's look for correlations between crime rate

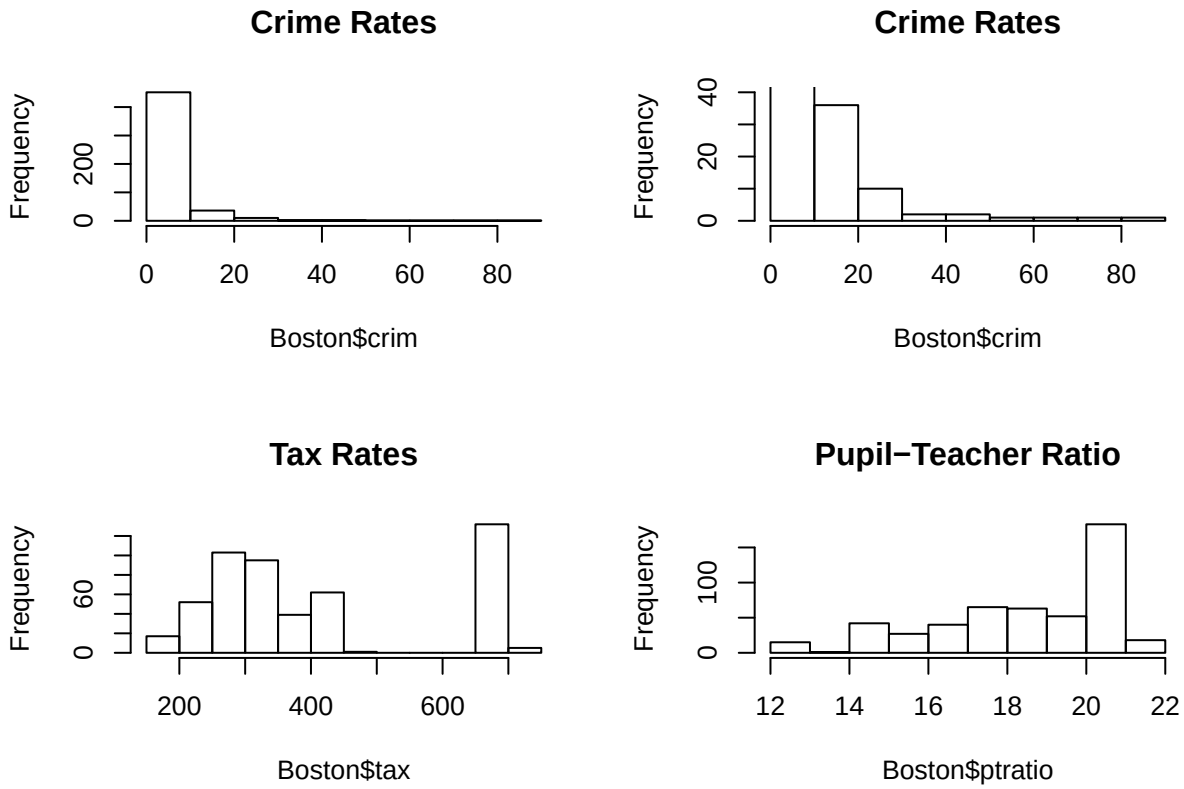
```
cor(Boston, use = "pairwise.complete.obs")
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593
##             ptratio       black      lstat       medv
## crim     0.2899456 -0.38506394  0.4556215 -0.3883046
## zn      -0.3916785  0.17552032 -0.4129946  0.3604453
## indus    0.3832476 -0.35697654  0.6037997 -0.4837252
## chas    -0.1215152  0.04878848 -0.0539293  0.1752602
## nox      0.1889327 -0.38005064  0.5908789 -0.4273208
## rm      -0.3555015  0.12806864 -0.6138083  0.6953599
## age      0.2615150 -0.27353398  0.6023385 -0.3769546
## dis     -0.2324705  0.29151167 -0.4969958  0.2499287
## rad      0.4647412 -0.44441282  0.4886763 -0.3816262
## tax      0.4608530 -0.44180801  0.5439934 -0.4685359
## ptratio  1.0000000 -0.17738330  0.3740443 -0.5077867
## black   -0.1773833  1.00000000 -0.3660869  0.3334608
## lstat    0.3740443 -0.36608690  1.0000000 -0.7376627
## medv    -0.5077867  0.33346082 -0.7376627  1.0000000
```

We can see that there is some correlation between variables such as ptratio, rad, tax, lstat, age, indus, and nox

- d) Let's look for some high crime, tax, and pupil-teacher rates

```
par(mfrow = c(2,2))
hist(Boston$crim, main = "Crime Rates")
hist(Boston$crim, main = "Crime Rates", ylim = c(0,40))
hist(Boston$tax, main = "Tax Rates")
hist(Boston$ptratio, main = "Pupil-Teacher Ratio")
```



It looks like crime is heavily skewed to one side and tax rates has high outliers

- e) Let's count how many suburbs are bound by the river

```
summary(Boston$chas == 1)
```

```
##    Mode   FALSE    TRUE
## logical     471      35
```

- f) Let's find the median pupil-teacher ratio

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

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

- g) Let's see the suburb of the lowest meadian value of homes

```r
which.min(Boston$medv)
```

```
## [1] 399
```

- h) Let's see the number of homes with 7 and 8 rooms

```r
summary(Boston$rm > 8)
```

```
##    Mode   FALSE    TRUE
## logical     493      13
```

```r
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
```

```r
summary(Boston)
```

```
##       crim                zn              indus             chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
```

```
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

It appears that all of these homes have generally lower crime rate in their suburbs