

Module 01 Lab 01

Joseph Annand

2023-10-29

Question 8

Part A

```
library(ISLR2)
college <- read.csv("College.csv", na.strings = "?", stringsAsFactors = T)
```

Part B

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

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

Part C

Subpart i

```
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
```

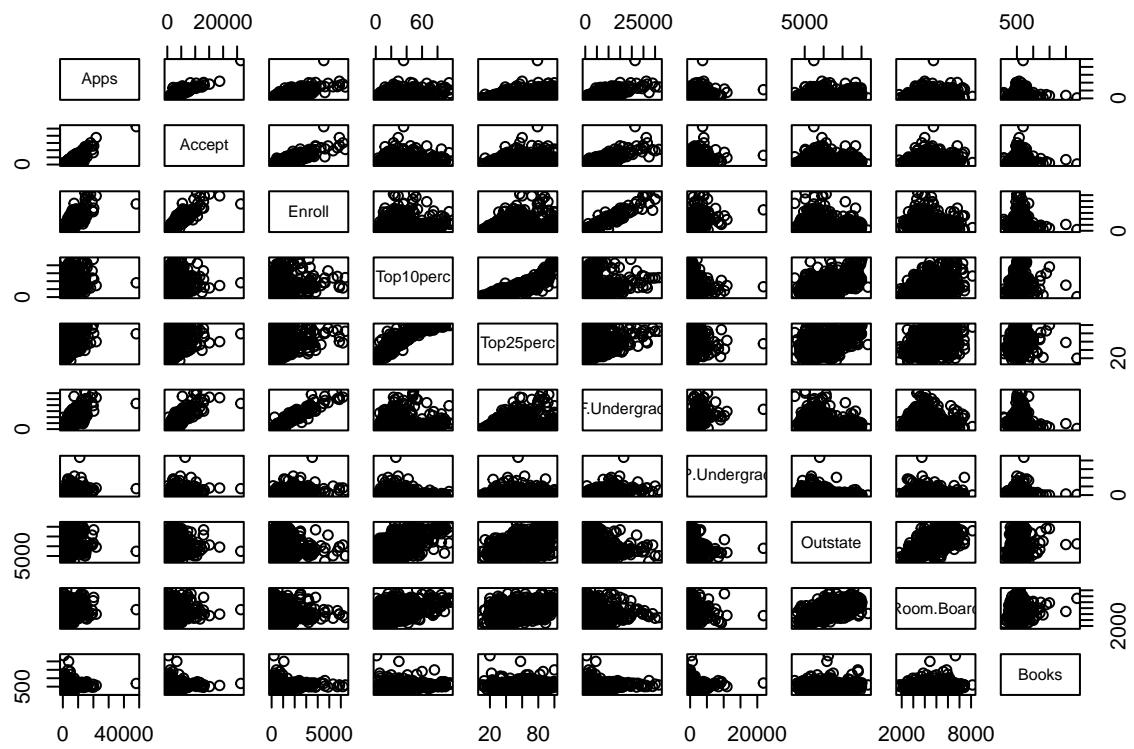
```

##      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

```

Subpart ii

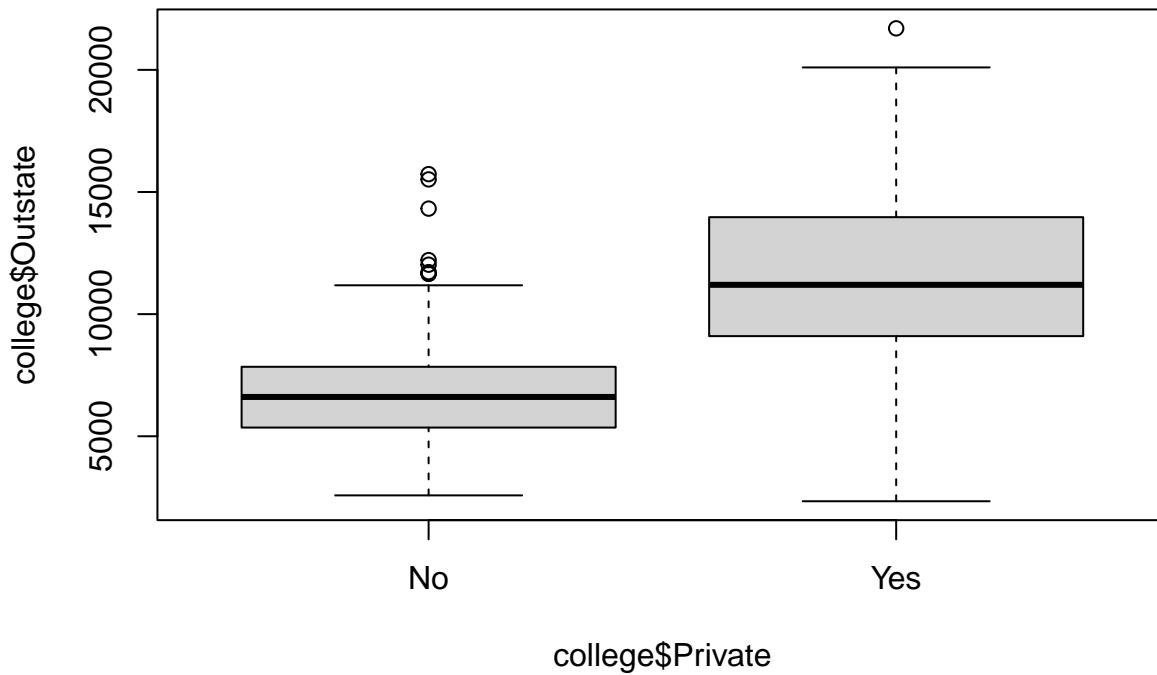
```
pairs(college[, 2:11])
```



Subpart iii

```
boxplot(college$Outstate ~ college$Private, main = "Boxplot of Out-of-State Tuition by Private status")
```

Boxplot of Out-of-State Tuition by Private status



Subpart iv

There are 78 “elite” universities.

```
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

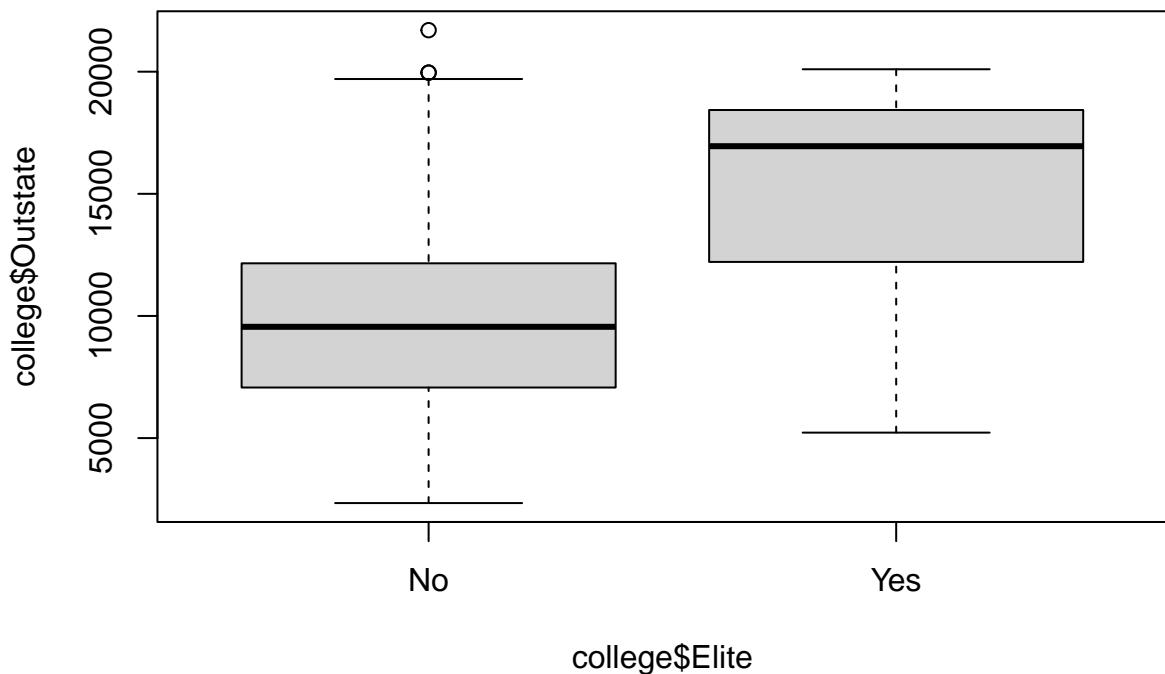
```

##      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, main = "Boxplot of Out-of-State Tuition by Elite status")
```

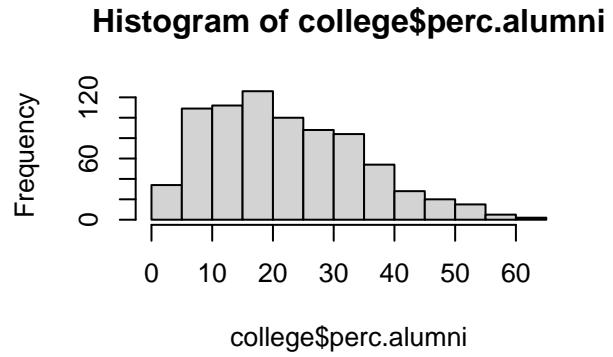
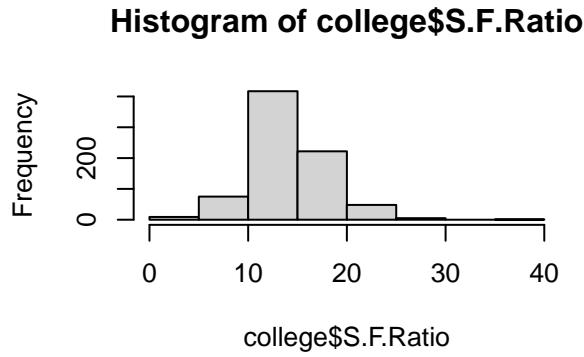
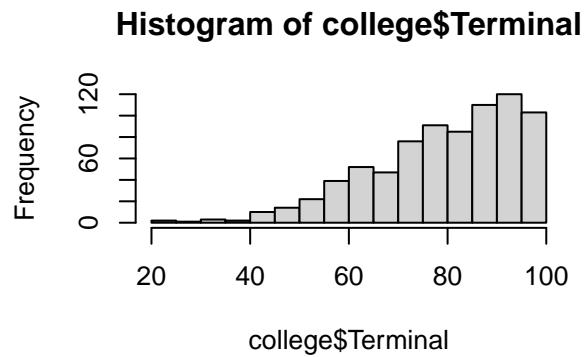
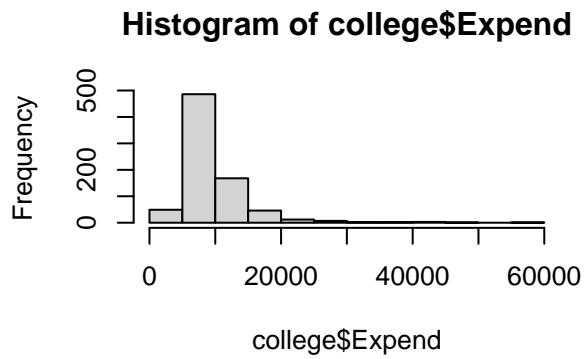
Boxplot of Out-of-State Tuition by Elite status



Subpart v

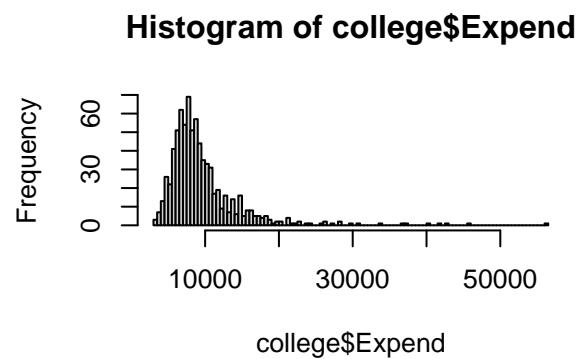
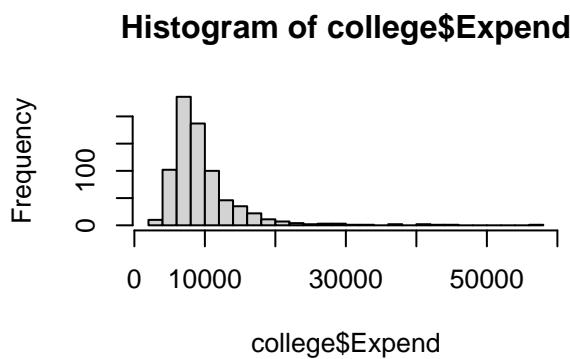
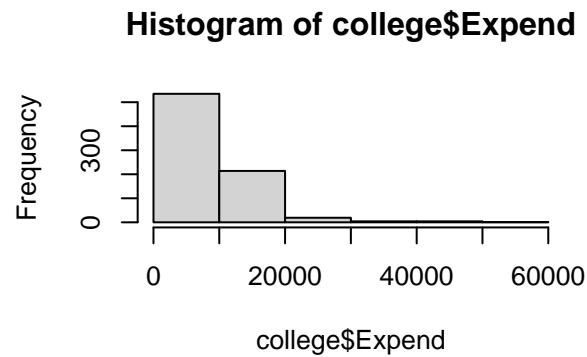
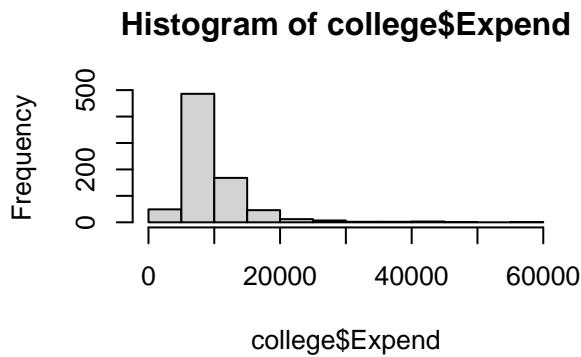
```
par(mfrow = c(2,2))

hist(college$Expend)
hist(college$Terminal)
hist(college$S.F.Ratio)
hist(college$perc.alumni)
```



```
par(mfrow = c(2,2))

hist(college$Expend)
hist(college$Expend, breaks = 5)
hist(college$Expend, breaks = 20)
hist(college$Expend, breaks = 100)
```

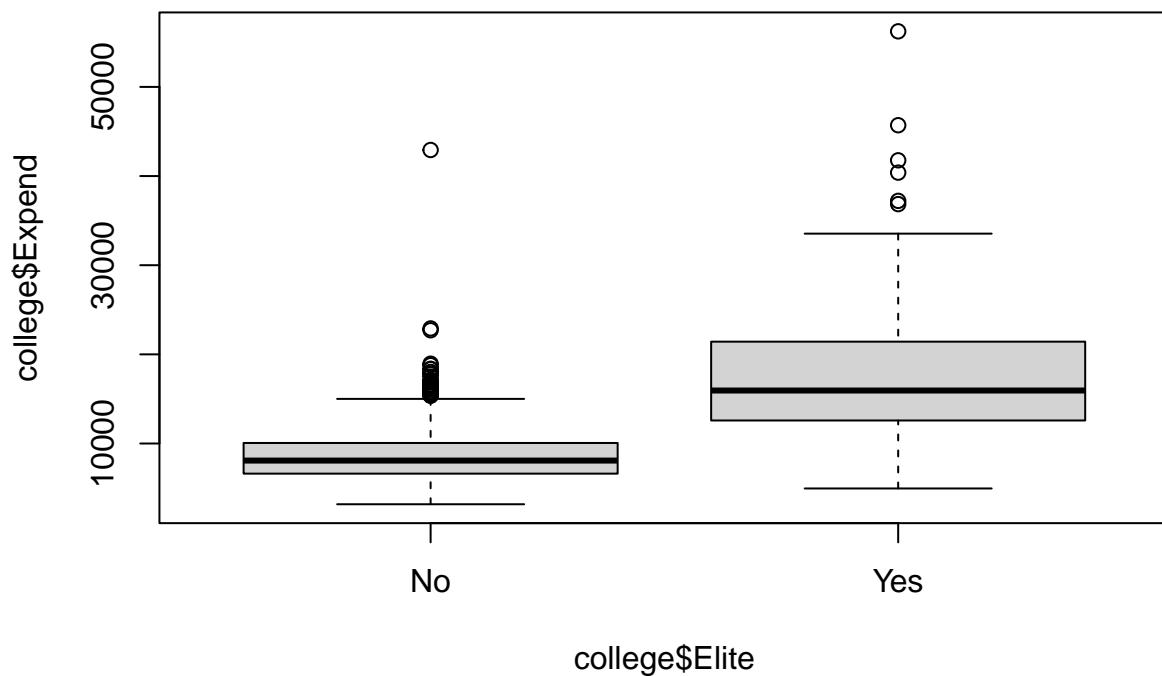


Subpart vi

Elite universities spend more on instruction per student than non-Elite schools. Elite schools tend to have smaller student-to-faculty ratios and have lower overall acceptance rates.

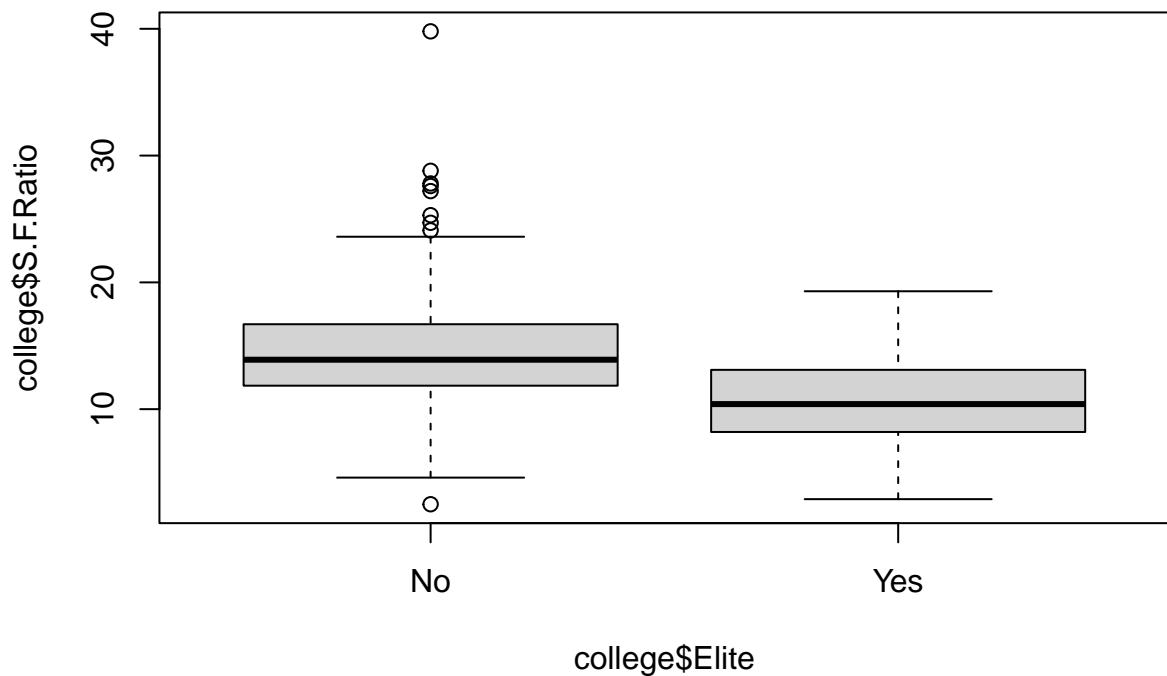
```
boxplot(college$Expend ~ college$Elite, main = "Boxplot of Instructional Expenditure by Elite status")
```

Boxplot of Instructional Expenditure by Elite status

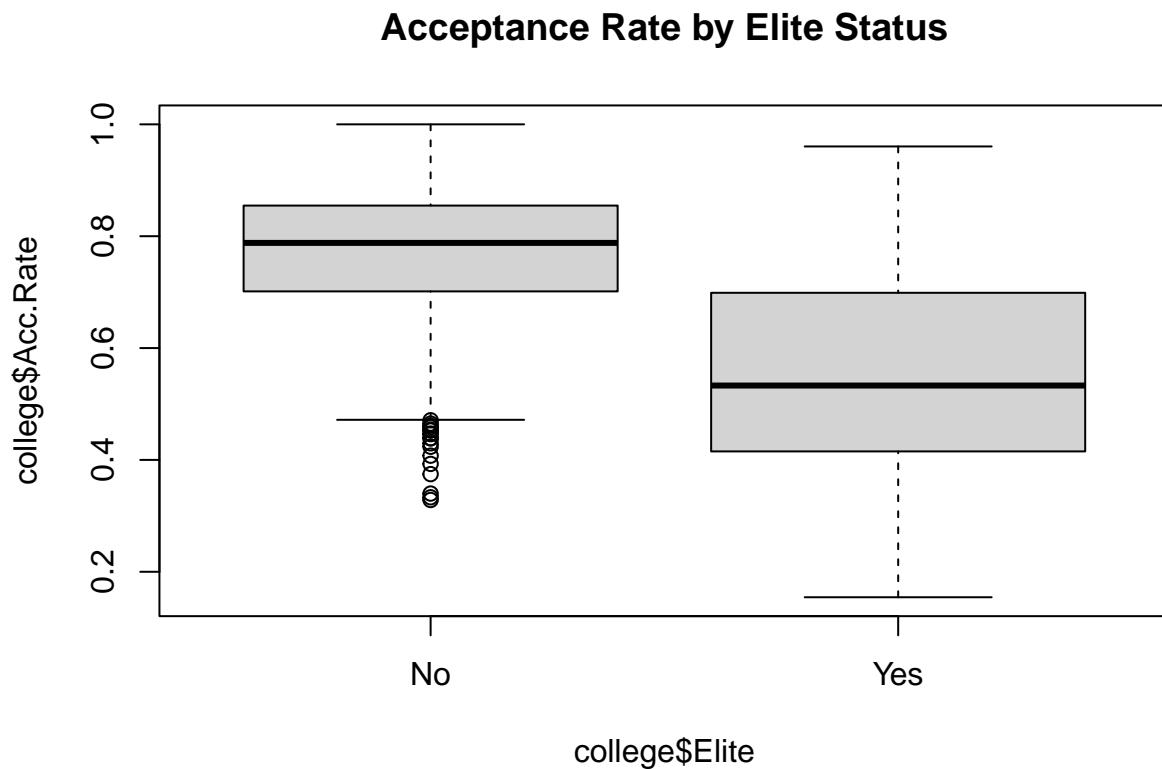


```
boxplot(college$S.F.Ratio ~ college$Elite, main = "Boxplot of Student/Faculty Ratio Expenditure by Elite")
```

Boxplot of Student/Faculty Ratio Expenditure by Elite status



```
college$Acc.Rate <- college$Accept / college$Apps  
boxplot(college$Acc.Rate ~ college$Elite, main = "Acceptance Rate by Elite Status")
```



Question 9

Part A

mpg, displacement, horsepower, weight, and acceleration are all quantitative predictors. Year, origin, name, and cylinders are qualitative predictors.

```
auto <- read.csv("Auto.csv", na.strings = "?", stringsAsFactors = T)
auto <- na.omit(auto)

View(auto)
```

Part B

The range of mpg: 9 - 46.6 The range of displacement: 68 - 455 The range of horsepower: 46 - 230 The range of weight: 1613 - 5140 The range of acceleration: 8 to 24.8

```
quantitative_auto = c('mpg', 'displacement', 'horsepower', 'weight', 'acceleration')
for (x in quantitative_auto) {
  print(range(auto[, x]))
}
```

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

```
## [1] 68 455
## [1] 46 230
## [1] 1613 5140
## [1] 8.0 24.8
```

Part C

```
sapply(auto[, quantitative_auto], function(x) c("Mean:" = mean(x, na.rm = T), "Std Dev:" = sd(x)))

##          mpg displacement horsepower      weight acceleration
## Mean:    23.445918      194.412 104.46939 2977.5842     15.541327
## Std Dev: 7.805007      104.644  38.49116 849.4026     2.758864
```

Part D

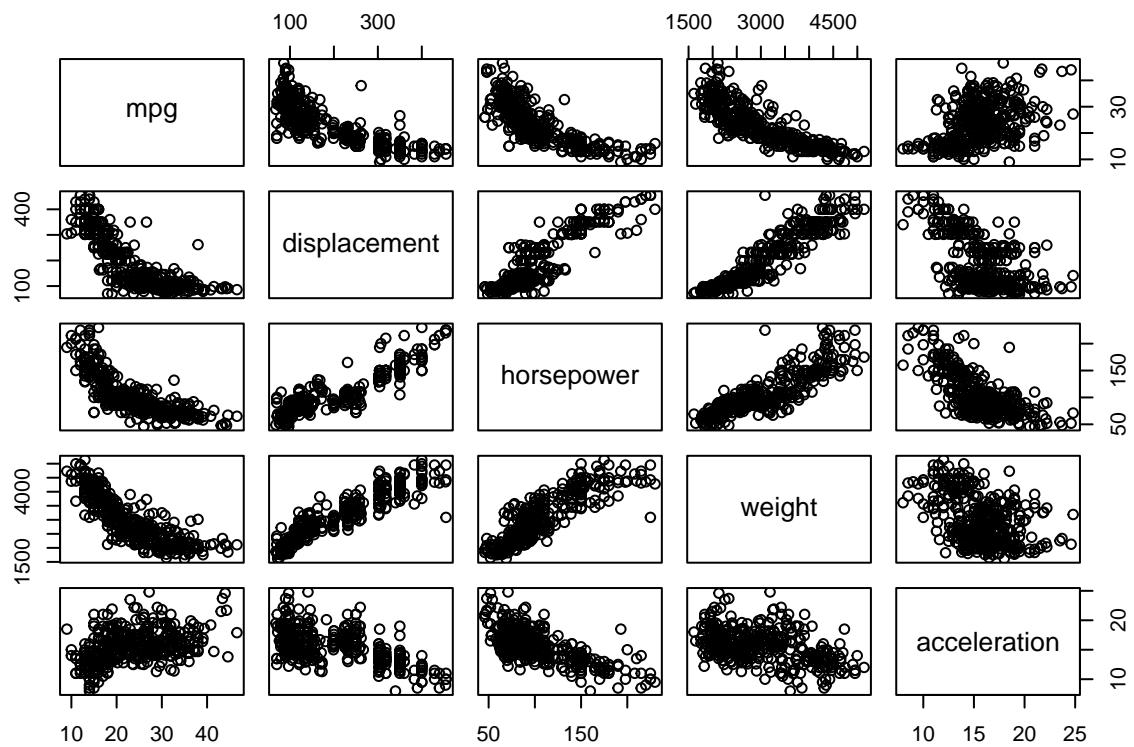
```
rmv_auto <- auto[-c(10:85),]
sapply(rmv_auto[,quantitative_auto], function(x) c("Mean:" = mean(x, na.rm = T),
                                                 "Std Dev:" = sd(x),
                                                 "Range" = range(x)))
```

```
##          mpg displacement horsepower      weight acceleration
## Mean:    24.404430      187.24051 100.72152 2935.9715     15.726899
## Std Dev: 7.867283      99.67837  35.70885 811.3002     2.693721
## Range1  11.000000      68.00000  46.00000 1649.0000     8.500000
## Range2  46.600000      455.00000 230.00000 4997.0000     24.800000
```

Part E

Scatter plots show that most predictors are not strongly related to each other; however, the relationship between mpg and the other predictors appears to be strong based on the graphs.

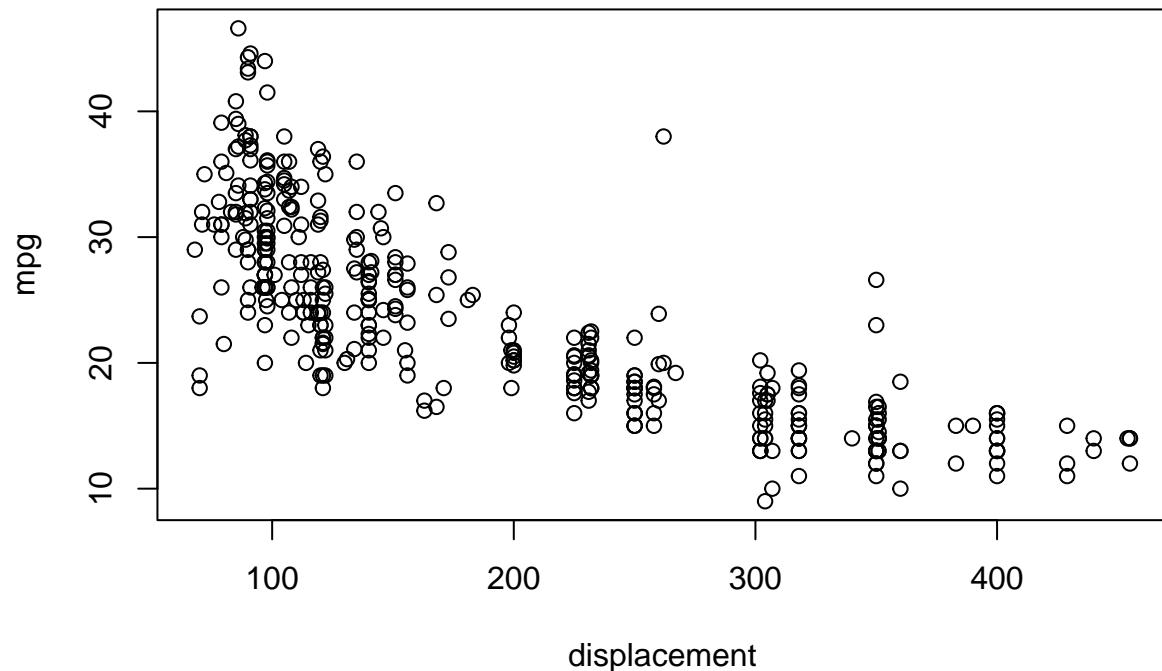
```
pairs(auto[,quantitative_auto])
```



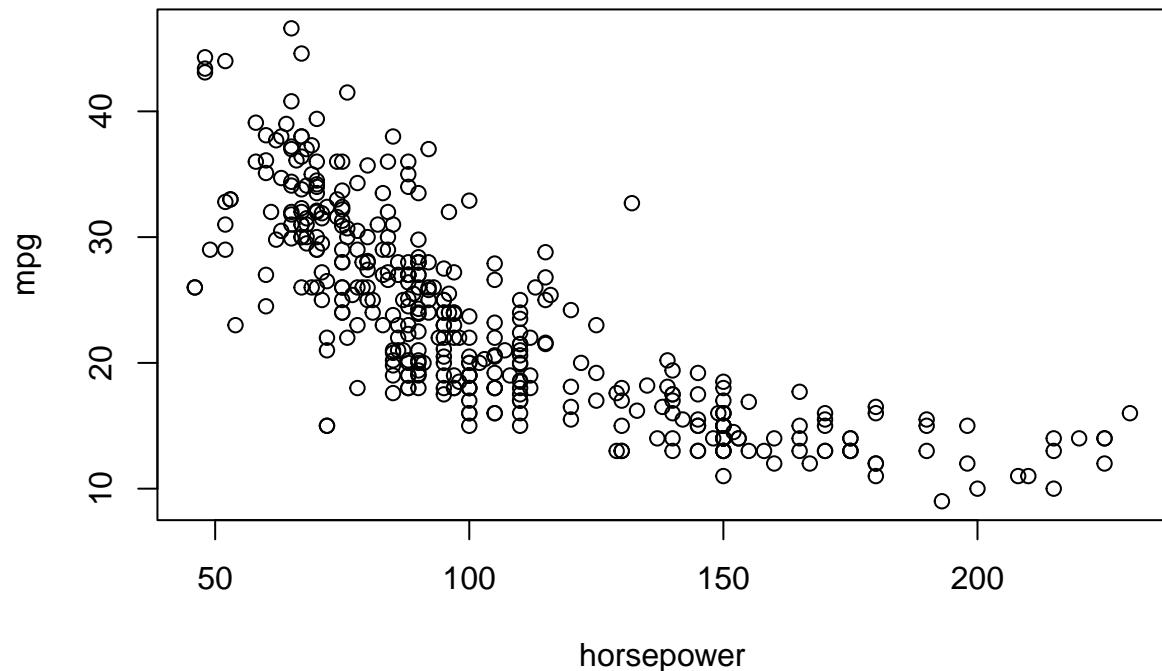
Part F

The plots as well as correlation coefficients show that displacement, horsepower, and weight could be useful for predicting mpg.

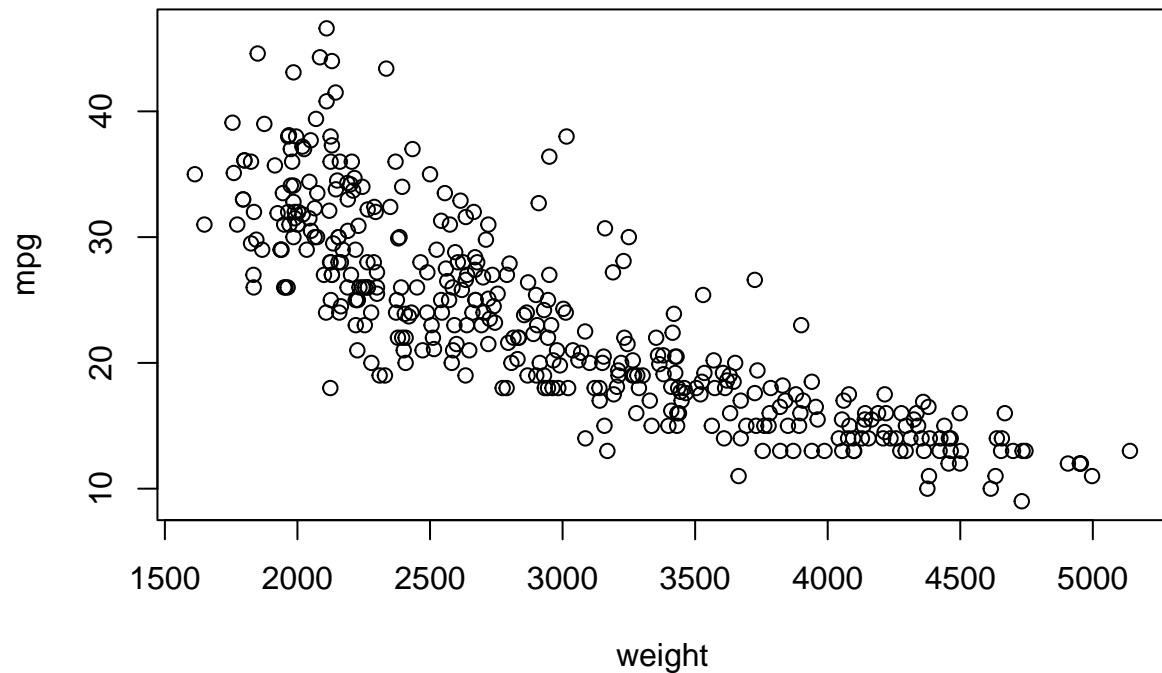
```
attach(auto)
plot(displacement, mpg)
```



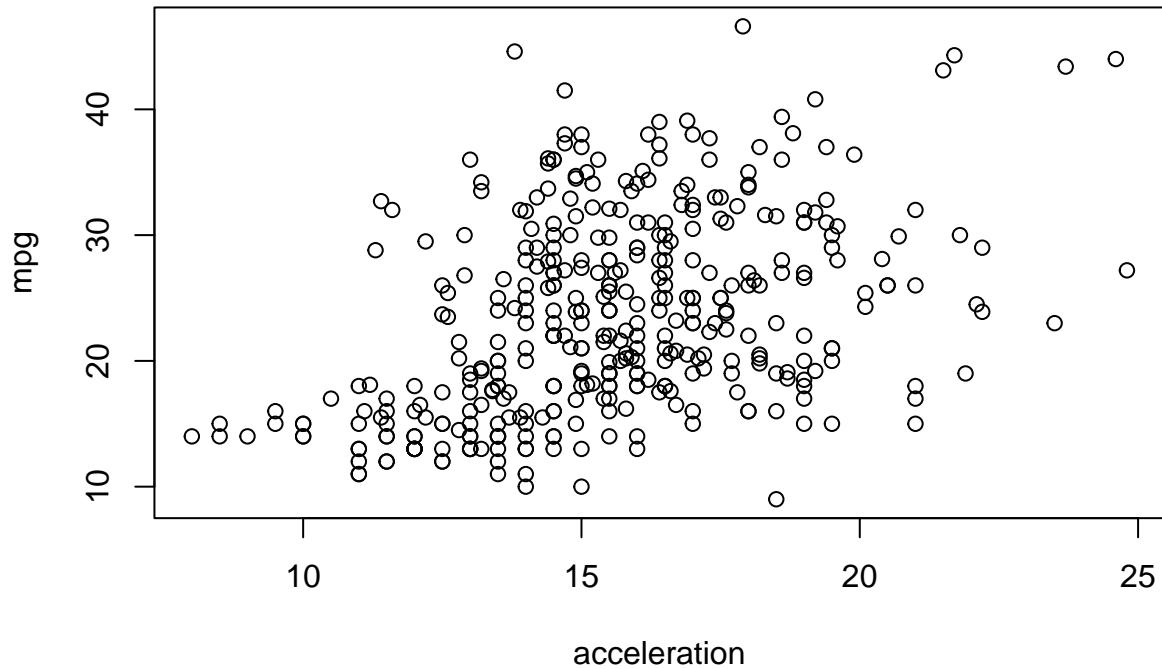
```
plot(horsepower, mpg)
```



```
plot(weight, mpg)
```



```
plot(acceleration, mpg)
```



```

for (predictor in c('displacement', 'horsepower', 'weight', 'acceleration')) {
  print(paste("Correlation coeff between mpg and ", predictor, " is ", cor(auto$mpg, auto[, predictor])))
}

## [1] "Correlation coeff between mpg and displacement is 0.423328536902787"
## [1] "Correlation coeff between mpg and horsepower is 0.423328536902787"
## [1] "Correlation coeff between mpg and weight is 0.423328536902787"
## [1] "Correlation coeff between mpg and acceleration is 0.423328536902787"

```

Question 10

Part A

There are 506 rows or observations in the Boston data set and 13 columns. Each observation is a suburb of Boston and each column is some geographic statistic describing an aspect of the town.

```
dim(Boston)
```

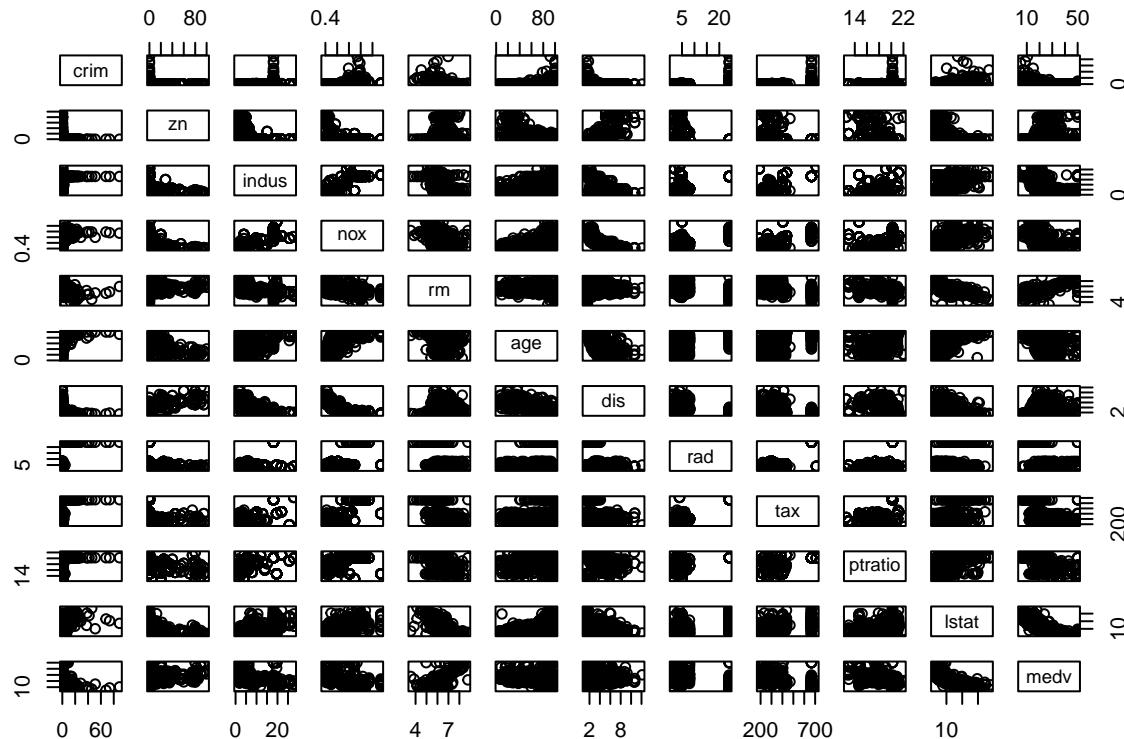
```
## [1] 506 13
```

```
View(Boston)
```

Part B

Many of the variables do not show a strong relationship with each other. For example, crime rate and residential land zoning do not appear to be good predictors for any of the other variables. Some other variables like lower status and home value appear have strong correlations with other variables and may be useful in predicting those variables.

```
pairs(Boston[, -4])
```



Part C

Based on graphs from Part B and correlation coefficients, the predictors most closely associated with crime rate are lower status and median home value. Crime is slightly positively correlated with crime rate and crime is slightly negatively correlated with median home value.

```
for (x in c('lstat', 'medv')) {  
  print(paste("Correlation coeff between crim and ", x, " is ", cor(Boston[x], Boston$crim)))  
}  
  
## [1] "Correlation coeff between crim and lstat is 0.455621479447946"  
## [1] "Correlation coeff between crim and medv is -0.388304608586812"
```

Part D

Some census tracts of Boston appear to have very high crime rates and tax rates based on the difference between the median and the maximum values for these variables. These statistics do not suggest that any census tracts have particularly high pupil-teacher ratios.

```
boston <- data.frame(Boston)

for (x in c('crim', 'tax', 'ptratio')) {
  print(range(Boston[, x]))
}

## [1] 0.00632 88.97620
## [1] 187 711
## [1] 12.6 22.0

print(median(boston$crim))

## [1] 0.25651

print(median(boston$tax))

## [1] 330

print(median(boston$ptratio))

## [1] 19.05
```

Part E

There are 35 census tracts that border the Charles river.

```
length(which(boston$chas == 1))

## [1] 35
```

Part F

The median pupil-teacher ratio among towns is 19.05.

```
median(boston$ptratio)

## [1] 19.05
```

Part G

Census tract in index 399 has the lowest median value of owner-occupied homes. It is in the upper range for non-retail business acres, the maximum for owner-occupied homes built before 1940, lower range for weighted mean of distances to five Boston employment centers, upper range for tax rate, upper range for pupil-teacher ratio, upper range for lower status, and the minimum for median home value.

```
min_med <- which.min(boston$medv)
print(min_med)
```

```
## [1] 399

min_med_df <- boston[min_med,]
View(min_med_df)

for (x in colnames(boston)) {
  print(range(boston[, x]))
}

## [1] 0.00632 88.97620
## [1] 0 100
## [1] 0.46 27.74
## [1] 0 1
## [1] 0.385 0.871
## [1] 3.561 8.780
## [1] 2.9 100.0
## [1] 1.1296 12.1265
## [1] 1 24
## [1] 187 711
## [1] 12.6 22.0
## [1] 1.73 37.97
## [1] 5 50
```

Part H

There are 64 census tracts with more than seven rooms per dwelling on average, and there are 13 census tracts with more than eight rooms per dwelling on average. These 13 census tracts have better accessibility to highways on average, lower average percentage of lower status, greater average median home value, and a lower maximum crime rate.

```
length(which(boston$rm > 7))
```

```
## [1] 64

length(which(boston$rm > 8))

## [1] 13
```

```
summary(boston[boston$rm > 8,])
```

```
##      crim            zn            indus            chas
##  Min.   : 0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.00000
##  1st Qu.: 0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.00000
##  Median : 0.52014   Median : 0.00   Median : 6.200   Median :0.00000
##  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.:0.00000
##  Max.   : 3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.00000
##      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          lstat          medv
##  Min.   : 2.000     Min.   :224.0    Min.   :13.00   Min.   :2.47   Min.   :21.9
##  1st Qu.: 5.000     1st Qu.:264.0    1st Qu.:14.70   1st Qu.:3.32   1st Qu.:41.7
##  Median : 7.000     Median :307.0    Median :17.40   Median :4.14   Median :48.3
##  Mean   : 7.462     Mean   :325.1    Mean   :16.36   Mean   :4.31   Mean   :44.2
##  3rd Qu.: 8.000     3rd Qu.:307.0    3rd Qu.:17.40   3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :24.000     Max.   :666.0    Max.   :20.20   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.: 0.08205   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          lstat
##  Min.   : 1.000     Min.   :187.0    Min.   :12.60   Min.   : 1.73
##  1st Qu.: 4.000     1st Qu.:279.0    1st Qu.:17.40   1st Qu.: 6.95
##  Median : 5.000     Median :330.0    Median :19.05   Median :11.36
##  Mean   : 9.549     Mean   :408.2    Mean   :18.46   Mean   :12.65
##  3rd Qu.:24.000     3rd Qu.:666.0    3rd Qu.:20.20   3rd Qu.:16.95
##  Max.   :24.000     Max.   :711.0    Max.   :22.00   Max.   :37.97
##      medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
```

```
## 3rd Qu.:25.00  
## Max. :50.00
```