Econ 573: Problem Set 1 - Part II

Exercise 9

a.

```
college = read.csv("C:/Users/mateo/OneDrive - University of North Carolina at Chapel Hill/Courses/Spring 2025/Econ 573/College.csv") #Read data into R
```

b.

```
View(college)
#view data
rownames(college) <- college[, 1] #creates a column with the names of each university recorded
View(college)
college <- college[, -1] #eliminates first column
View(college)
```

c.

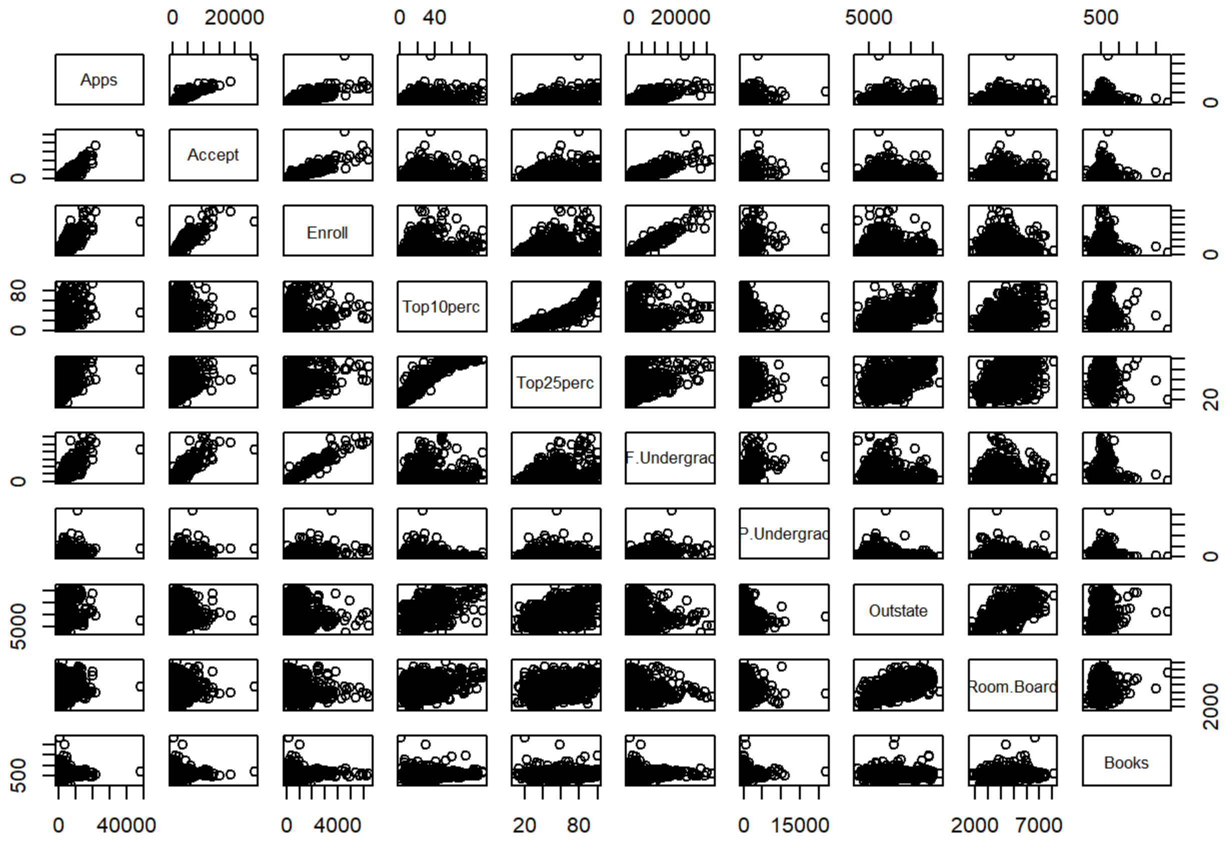
i.

```
summary(college) #produces a numerical summary of the variables in the data set
```

```
##    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
```

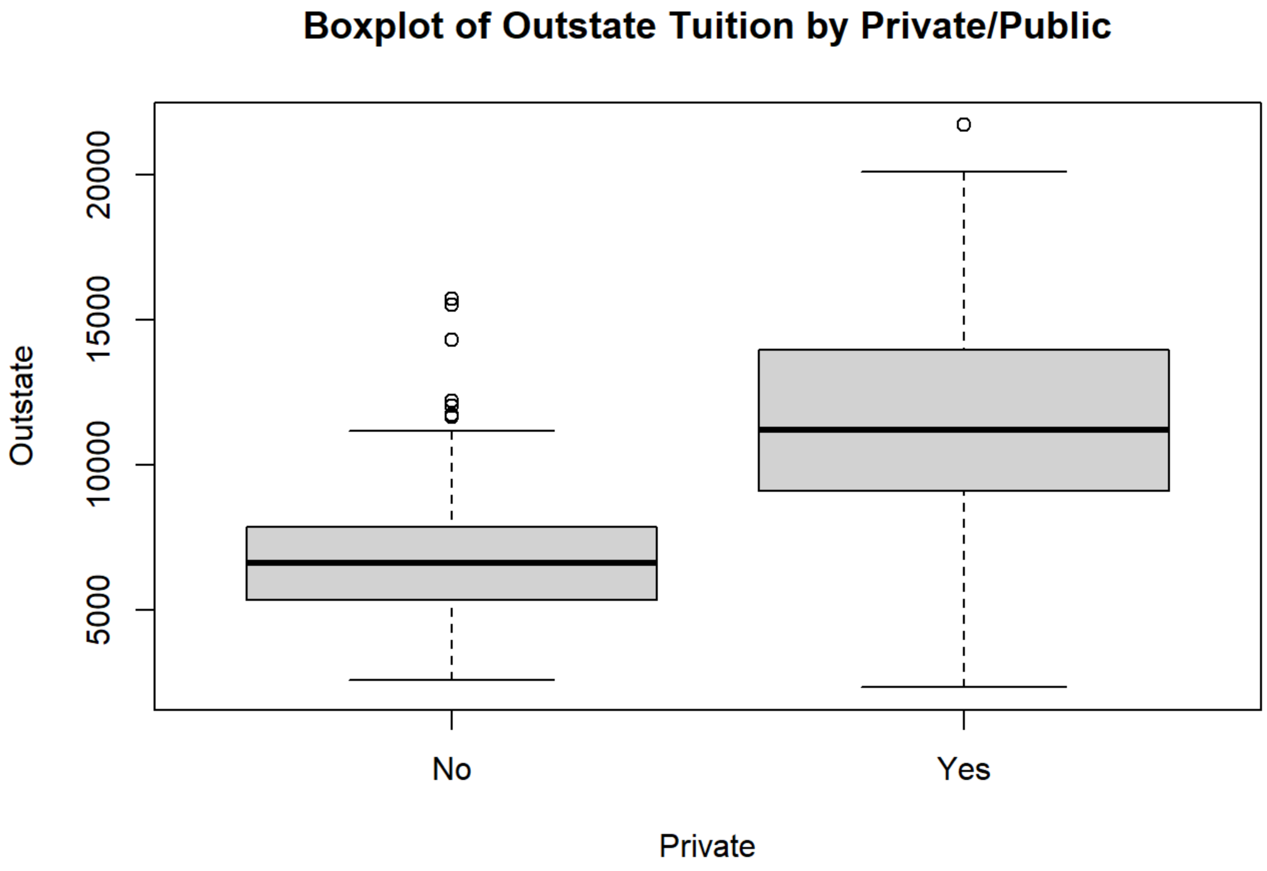
ii.

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



iii.

```
attach(college) #makes variables from college available by name.
Private <- as.factor(Private) #the Private column will be treated as a categorical variable with
two levels: "Yes and "No"
plot(Outstate ~ Private, #Plots outsate as a function of Private
     xlab = "Private",
     ylab= "Outstate",
     main = "Boxplot of Outstate Tuition by Private/Public")
```



iv.

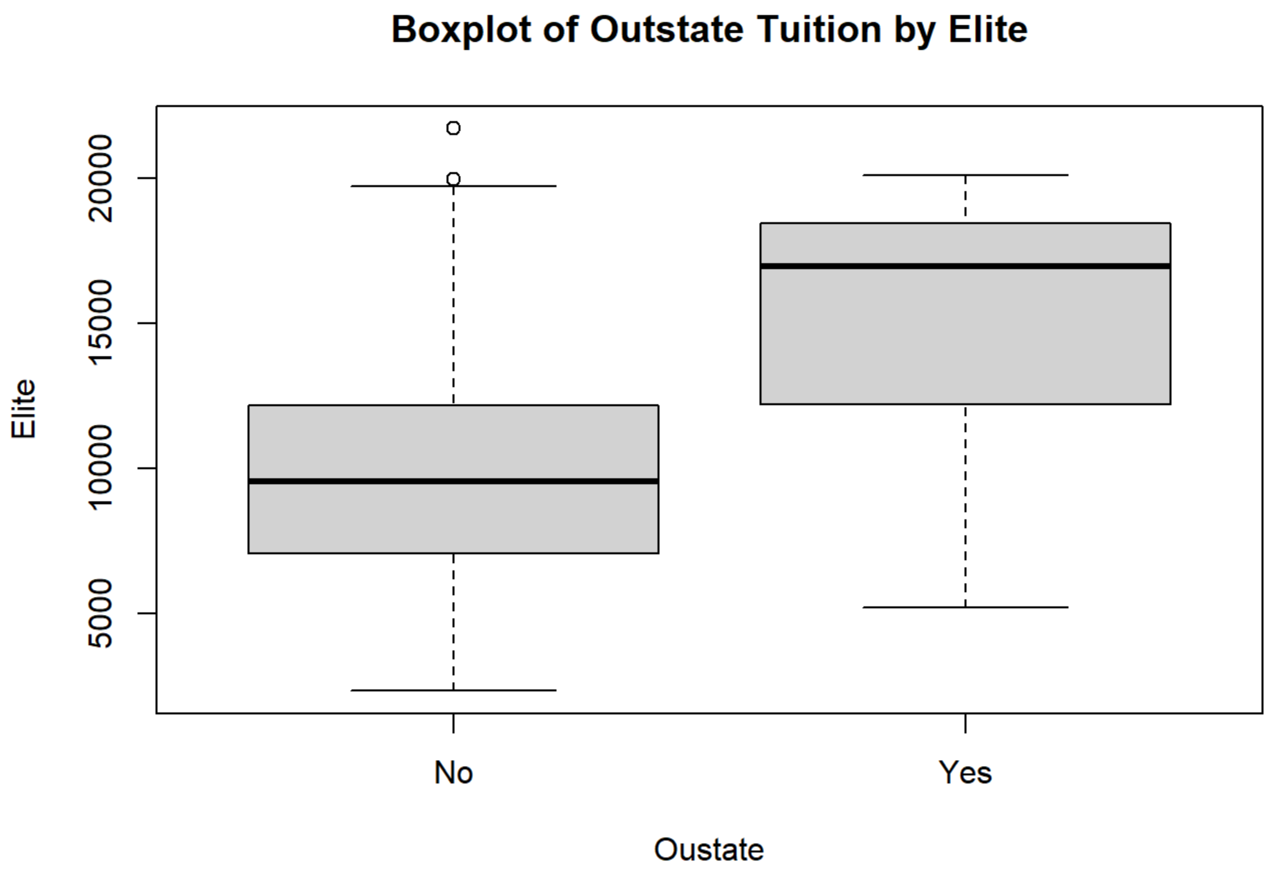
```
Elite <- rep("No", nrow(college)) #creates a vector of the same length as the college data set with the value "No"
Elite[college$Top10perc > 50] <- "Yes" #Check for rows where the percentage of top students is greater than 50%. For those rows, the value in the Elite vectors is changed to "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite) #converts vector into a factor, such that R will treate it as a categoricla grouping variable with two levels: "NO" and "Yes"
summary(Elite)
```

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

```
attach(college)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     Elite, Private
## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, Books, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
```

```
plot(Outstate ~ Elite,
     xlab = "Oustate",
     ylab = "Elite",
     main = "Boxplot of Outstate Tuition by Elite")
```



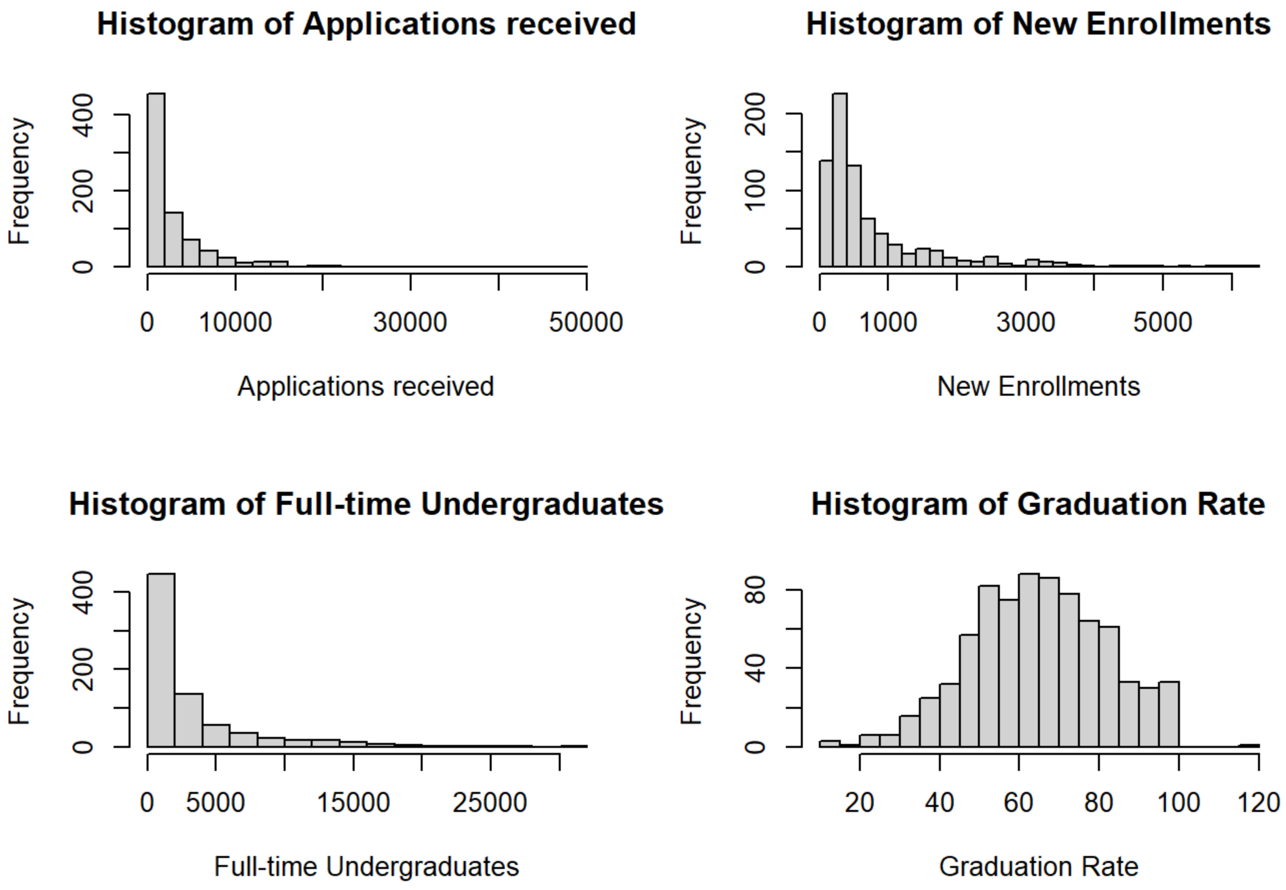
v.

```
par(mfrow = c(2, 2)) #Divides the plotting area into 2x2 grid
attach(college)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     Elite, Private
## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, Books, Elite, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
```

```
## The following objects are masked from college (pos = 4):
##
##     Accept, Apps, Books, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
```

```
hist(Apps, main = "Histogram of Applications received", xlab = "Applications received", breaks =
20)
hist(Enroll, main = "Histogram of New Enrollments", xlab = "New Enrollments", breaks = 30)
hist(F.Undergrad, main = "Histogram of Full-time Undergraduates", xlab = "Full-time Undergraduat
es", breaks = 15)
hist(Grad.Rate, main = "Histogram of Graduation Rate", xlab = "Graduation Rate", breaks = 25)
```



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

Exercise 9

```
auto = read.csv("C:/Users/mateo/OneDrive - University of North Carolina at Chapel Hill/Courses/S
pring 2025/Econ 573/Auto.csv")
View(auto)
```

a.

```
summary(auto)
```

```
##       mpg          cylinders      displacement    horsepower       
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Length:397        
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   Class :character  
##  Median :23.00   Median :4.000   Median :146.0   Mode  :character  
##  Mean   :23.52   Mean   :5.458   Mean   :193.5                     
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0                     
##  Max.   :46.60   Max.   :8.000   Max.   :455.0                     
##      weight      acceleration        year           origin     
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000  
##  1st Qu.:2223   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000  
##  Median :2800   Median :15.50   Median :76.00   Median :1.000  
##  Mean   :2970   Mean   :15.56   Mean   :75.99   Mean   :1.574  
##  3rd Qu.:3609   3rd Qu.:17.10   3rd Qu.:79.00   3rd Qu.:2.000  
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000  
##      name          
##  Length:397        
##  Class :character  
##  Mode  :character  
##                    
##                    
## 
```

```
str(auto)
```

```
## 'data.frame':    397 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  : chr  "130" "165" "150" "150" ...
##  $ 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" "a
mc rebel sst" ...
```

Based on the outputs we can conclude that “miles per gallon”, “cylinders”, “displacement”, “weight”, and “acceleration” are all quantitative variables. While “name”, “horsepower”, “year”, and “origin” are qualitative variables.

b.

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

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

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

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

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

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

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

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

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

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

c.

```
mean(auto$mpg)
```

```
## [1] 23.51587
```

```
sd(auto$mpg)
```

```
## [1] 7.825804
```

```
mean(auto$cylinders)
```

```
## [1] 5.458438
```

```
sd(auto$cylinders)
```

```
## [1] 1.701577
```

```
mean(auto$displacement)
```

```
## [1] 193.5327
```

```
sd(auto$displacement)
```

```
## [1] 104.3796
```

```
mean(auto$weight)
```

```
## [1] 2970.262
```

```
sd(auto$weight)
```

```
## [1] 847.9041
```

```
mean(auto$acceleration)
```

```
## [1] 15.55567
```

```
sd(auto$acceleration)
```

```
## [1] 2.749995
```

d.

```
auto_sub <- auto[-(10:85), ]
range(auto_sub$mpg)
```

```
## [1] 11.0 46.6
```

```
mean(auto_sub$mpg)
```

```
## [1] 24.43863
```

```
sd(auto_sub$mpg)
```

```
## [1] 7.908184
```

```
range(auto_sub$cylinders)
```

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

```
mean(auto_sub$cylinders)
```

```
## [1] 5.370717
```

```
sd(auto_sub$cylinders)
```

```
## [1] 1.653486
```

```
range(auto_sub$displacement)
```

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

```
mean(auto_sub$displacement)
```

```
## [1] 187.0498
```

```
sd(auto_sub$displacement)
```

```
## [1] 99.63539
```

```
range(auto_sub$weight)
```

```
## [1] 1649 4997
```

```
mean(auto_sub$weight)
```

```
## [1] 2933.963
```

```
sd(auto_sub$weight)
```

```
## [1] 810.6429
```

```
range(auto_sub$horsepower)
```

```
## [1] "?"  "98"
```

```
mean(auto_sub$horsepower)
```

```
## Warning in mean.default(auto_sub$horsepower): argument is not numeric or
## logical: returning NA
```

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

```
sd(auto_sub$horsepower)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introducidos por coerción
```

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

```
range(auto_sub$acceleration)
```

```
## [1]  8.5 24.8
```

```
mean(auto_sub$acceleration)
```

```
## [1] 15.72305
```

```
sd(auto_sub$acceleration)
```

```
## [1] 2.680514
```

```
range(auto_sub$year)
```

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

```
mean(auto_sub$year)
```

```
## [1] 77.15265
```

```
sd(auto_sub$year)
```

```
## [1] 3.11123
```

```
range(auto_sub$origin)
```

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

```
mean(auto_sub$origin)
```

```
## [1] 1.598131
```

```
sd(auto_sub$origin)
```

```
## [1] 0.8161627
```

```
range(auto_sub$name)
```

```
## [1] "amc ambassador brougham" "vw rabbit custom"
```

```
mean(auto_sub$name)
```

```
## Warning in mean.default(auto_sub$name): argument is not numeric or logical:
## returning NA
```

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

```
sd(auto_sub$name)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introducidos por coerción
```

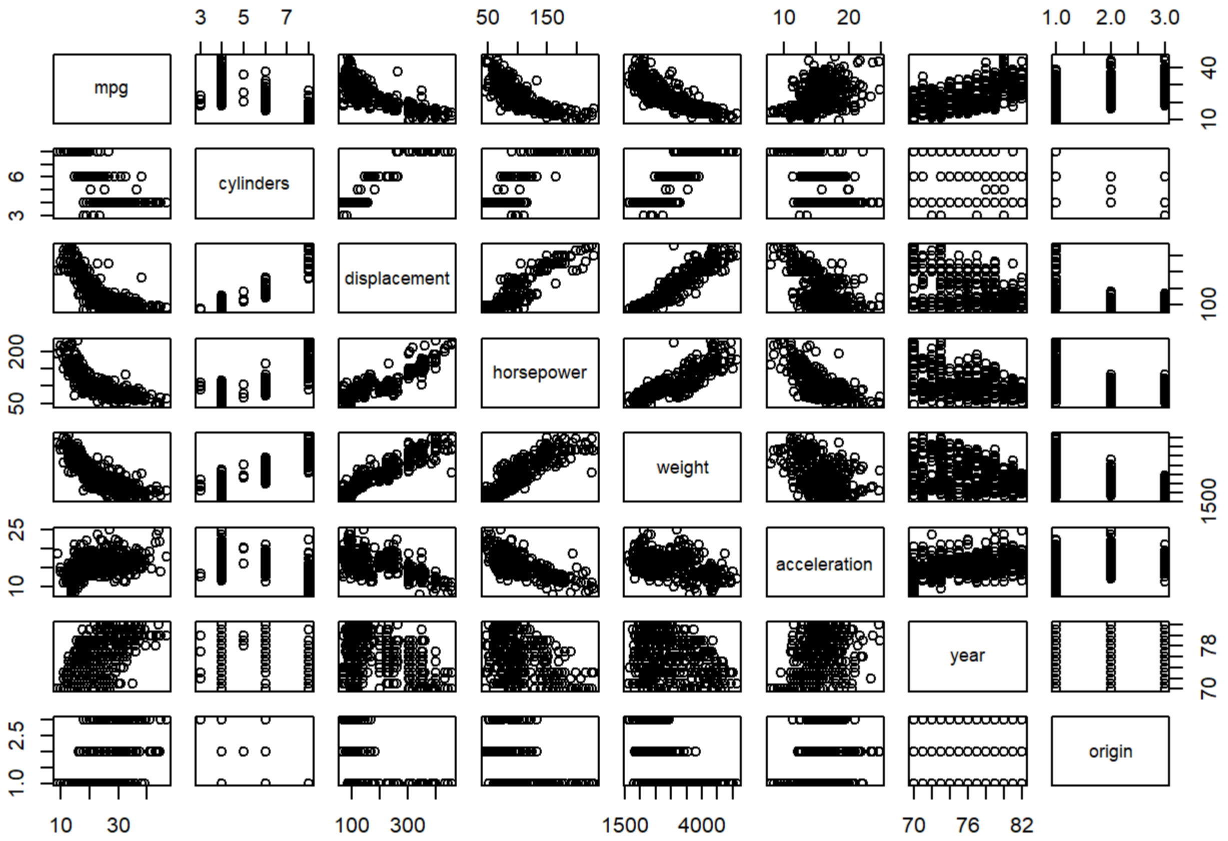
```
## [1] NA
```

e.

```
auto$horsepower <- as.numeric(auto$horsepower)
```

```
## Warning: NAs introducidos por coerción
```

```
pairs(auto[,1:8])
```



“miles per gallon” seems to have an inverse relationship with “displacement”, “horsepower”, and “weight”. However, it seems to have a converse relationship with “acceleration” and “year”.

“displacement” has an inverse relationship with “acceleration”, but a converse relationship with “horsepower” and “weight”.

“horsepower” has an inverse relationship with “acceleration”, but a converse relationship with “weight”.

“acceleration” and “weight” seem to have somewhat of an inverse relationship.

f. Yes. The plotting suggests that “displacement”, “horsepower”, “weight”, “acceleration”, and “year” could be useful when prediction gas mileage. Larger engines with higher horsepower have lower gas mileage as they tend to consumer fuel. Heavier cars also tend to have lower fuel efficiency. Newer cars might have improvements in fuel efficiency.

Exercise 10

```
boston = read.csv("C:/Users/mateo/OneDrive - University of North Carolina at Chapel Hill/Course
s/Spring 2025/Econ 573/Boston.csv")
```

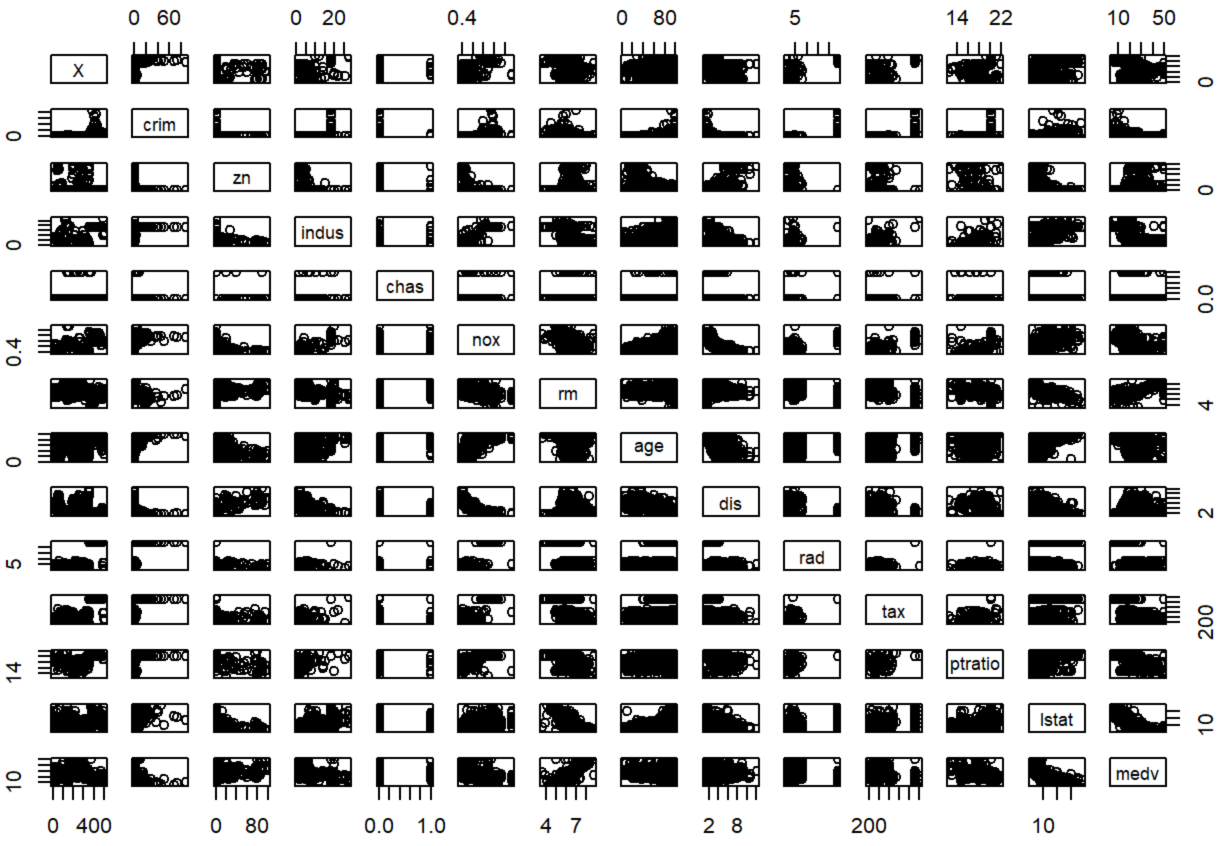
a.

```
dim(boston)
```

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

b.

```
pairs(boston)
```



Both the proportion of non-retail business acres per town and the nitrogen oxides concentration seem inversely related to the weighted mean of distances to five Boston employment centres. However, the later might not be relevant.

c.

```
cor(boston$crim, boston)
```

```
##              X crim         zn     indus        chas       nox         rm
## [1,] 0.4074072    1 -0.2004692 0.4065834 -0.05589158 0.4209717 -0.2192467
##            age        dis       rad       tax   ptratio     lstat       medv
## [1,] 0.3527343 -0.3796701 0.6255051 0.5827643 0.2899456 0.4556215 -0.3883046
```

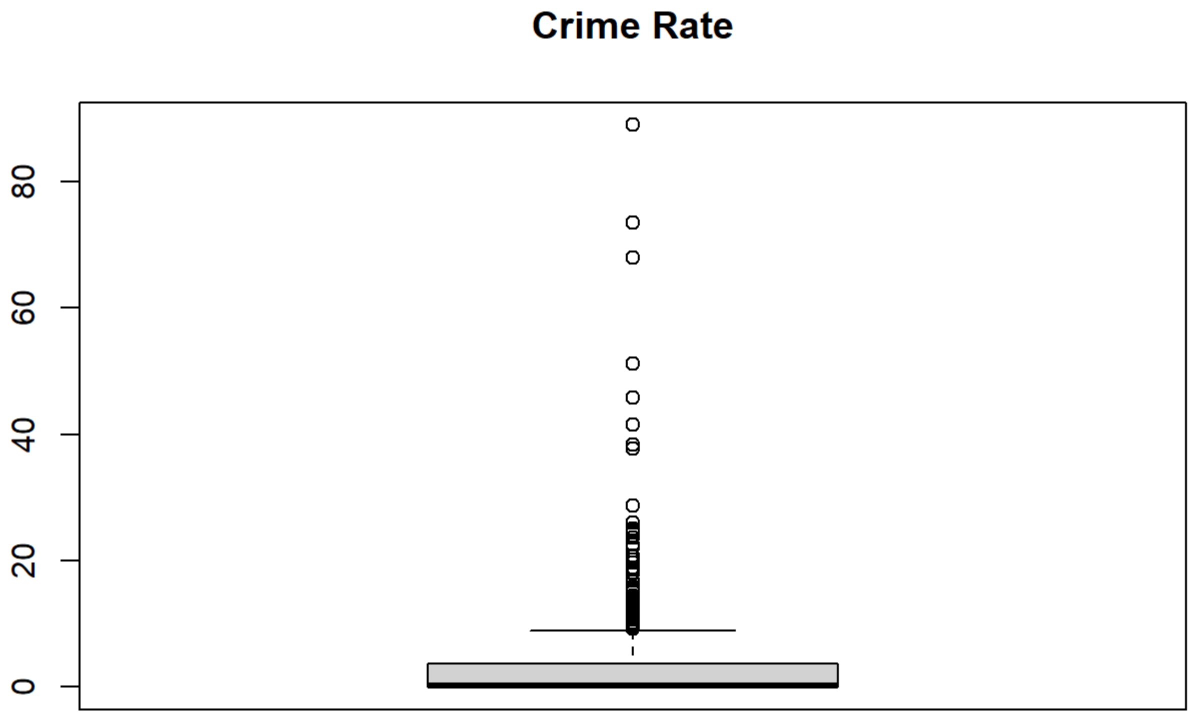
The per capital crime rate seems to have a relevant positive correlation with property tax values(tax), the lower status of the population(lstat), and the nitrogen oxides concentration(nox), and the proportion of non-retail business acres per town(indus). While it also appears to have a weaker negative correlation with the mean of distances to five Boston employment centres (dis) and the median value of owner-occupied homes (medv) that might turn out to be relevant.

d.

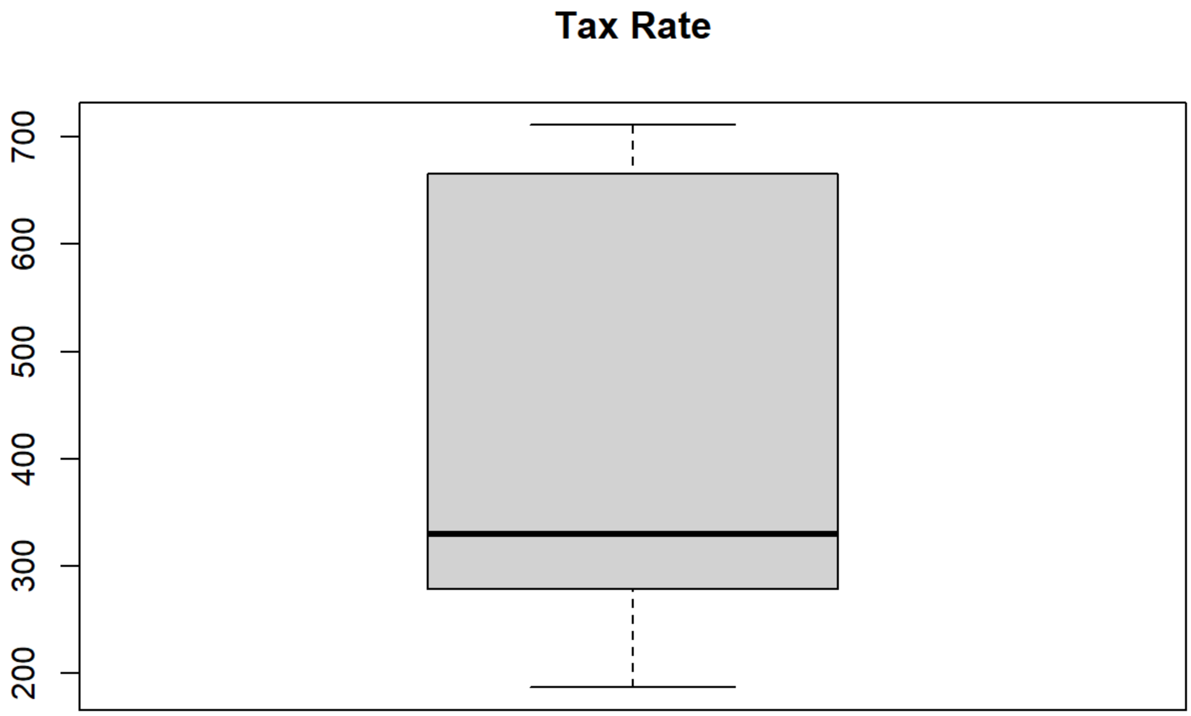
```
summary(boston)
```

```
##        X              crim                zn             indus      
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46  
##  1st Qu.:127.2   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19  
##  Median :253.5   81 Median :  0.00   Median : 9.69  
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14  
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10  
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74  
##       chas              nox               rm             age        
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
##  1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
##  Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
##  Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
##  3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
##       dis              rad              tax           ptratio     
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60  
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40  
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05  
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46  
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20  
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00  
##      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
```

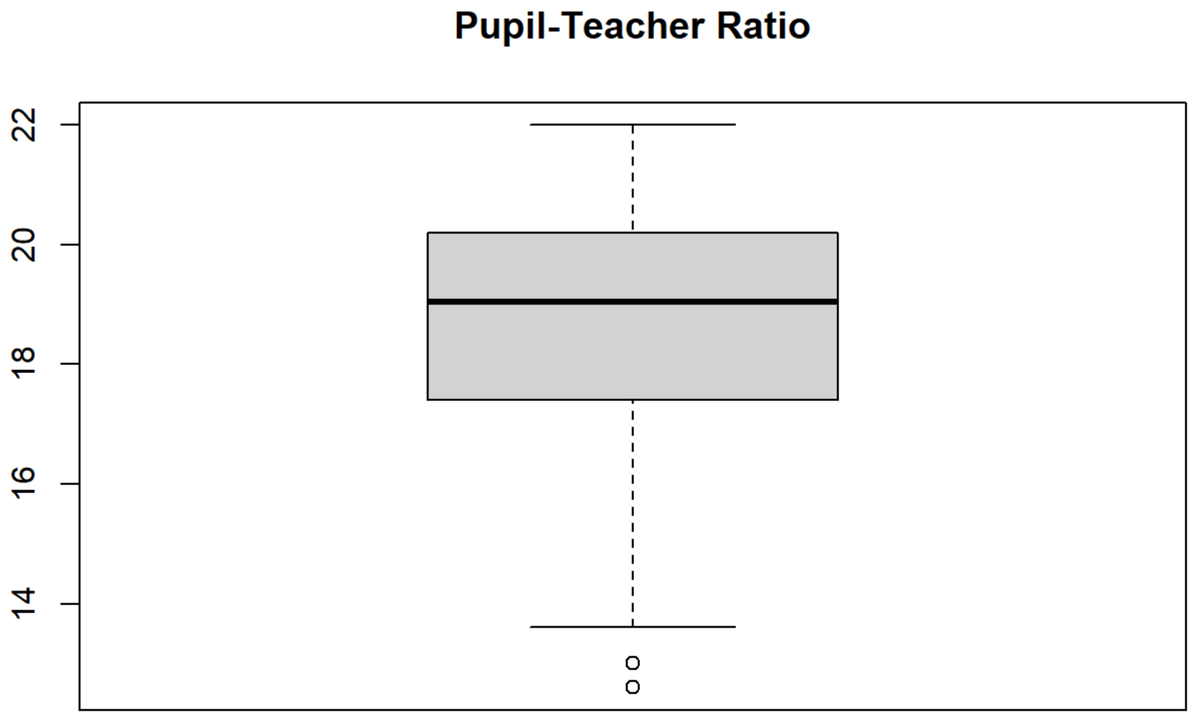
```
boxplot(boston$crim, main="Crime Rate")
```



```
boxplot(boston$tax, main="Tax Rate")
```



```
boxplot(boston$ptratio, main="Pupil-Teacher Ratio")
```



The data for crime rates shows that the majority of the data clusters at the bottom, such that most census tracts have lower crime rates. We observe that the mean is 3.61352 and the maximum is 88.9762. The 3rd quantile representing only represents up to 3.67708. There is a few outliers with particularly high crime rates.

For tax rates, the range is from 187 to 711, with the mean being 408.2, which maintains that the distribution is semi-uniform.

For pupil-teacher ratios tha range is from 12.60 to 22, with the mean being 18.46. It seems as it is somewhat skewed towards higher values.

e.

```
sum(boston$chas == 1)
```

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

35 census tracts in the data set bound the Charles Rivers.

f.

```
median(boston$ptratio)
```

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

The median pupil-teacher ration is 19.05

g.

```
lowest_medv_index <- which.min(boston$medv)
lowest_medv_index <- boston[lowest_medv_index, ]
lowest_medv_index
```

	X	crim	zn	indus	chas	nox	rm	age	dis ▸
	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
399	399	38.3518	0	18.1	0	0.693	5.453	100	1.4896

1 row | 1-10 of 15 columns

```
range(boston$crim)
```

```
## [1]  0.00632 88.97620
```

```
range(boston$zn)
```

```
## [1]   0 100
```

```
range(boston$indus)
```

```
## [1]  0.46 27.74
```

```
range(boston$chas)
```

```
## [1] 0 1
```

```
range(boston$nox)
```

```
## [1] 0.385 0.871
```

```
range(boston$rm)
```

```
## [1] 3.561 8.780
```

```
range(boston$age)
```

```
## [1]   2.9 100.0
```

```
range(boston$dis)
```

```
## [1]  1.1296 12.1265
```

```
range(boston$rad)
```

```
## [1]  1 24
```

```
range(boston$tax)
```

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

```
range(boston$ptratio)
```

```
## [1] 12.6 22.0
```

```
range(boston$lstat)
```

```
## [1]  1.73 37.97
```

```
summary(boston)
```

```
##        X               crim                zn             indus
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46
##  1st Qu.:127.2   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19
##  Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74
##       chas              nox               rm             age
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90
##  1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02
##  Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50
##  Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57
##  3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00
##       dis              rad              tax           ptratio
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
##      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
```

```
summary(boston$lstat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73    6.95   11.36   12.65   16.95   37.97
```

The census track with lowest median value of owner-occupied homes only has 5. In terms of crime rates, the value is way above the mean, at a relatively higher value of 38.3518. However, not as high as the maximum of 88.97620. In terms of zn it’s at the minimum of 0. For indus its at the start of the 3rd quintile, above the mean but below the maximum. It’s not bound by the Charles River. It’s nox value is above the 3rd quintile, but below the maximum. The rm variable is below the mean but above the minimum. It’s at the maximum age for census tracks at 100. The dis is close to the minimum. The rad is above the minimum but below the 1st quintile. Both the tax and the ptratio are at their 3rd quintiles. The lstat is below the 1st quintile.

h.

```
sum(boston$rm >7)
```

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

```
sum(boston$rm >8)
```

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

64 census tracts in the data set average more than seven rooms per dwelling, while only 13 average more than eight rooms per dwelling.