Statistical Learning: Chapter 2

Phuong Dong Le

```
library(ggplot2)
library(tidyverse)
library(gridExtra)
library(kableExtra)
library(GGally)
library(table1)
```

Exercise 8.

This relates to the `College` data set, containing a number of variables for 777 different universities and colleges in the US.

(a). Use the `read.csv()` to read the data into `R`, call the loaded data `college`.

```
## set the working directory:
setwd(dir = "~/Desktop/Statistical Learning/dataset")

### read the dataset college
college = read.csv(file = "college.csv")
```

(b). Look at the data using the `fix()` function. Try the following commands:

```
##rownames(college) = college[,1]
##fix(college)


### college = college[,-1]
### fix(college)
```

(c)

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

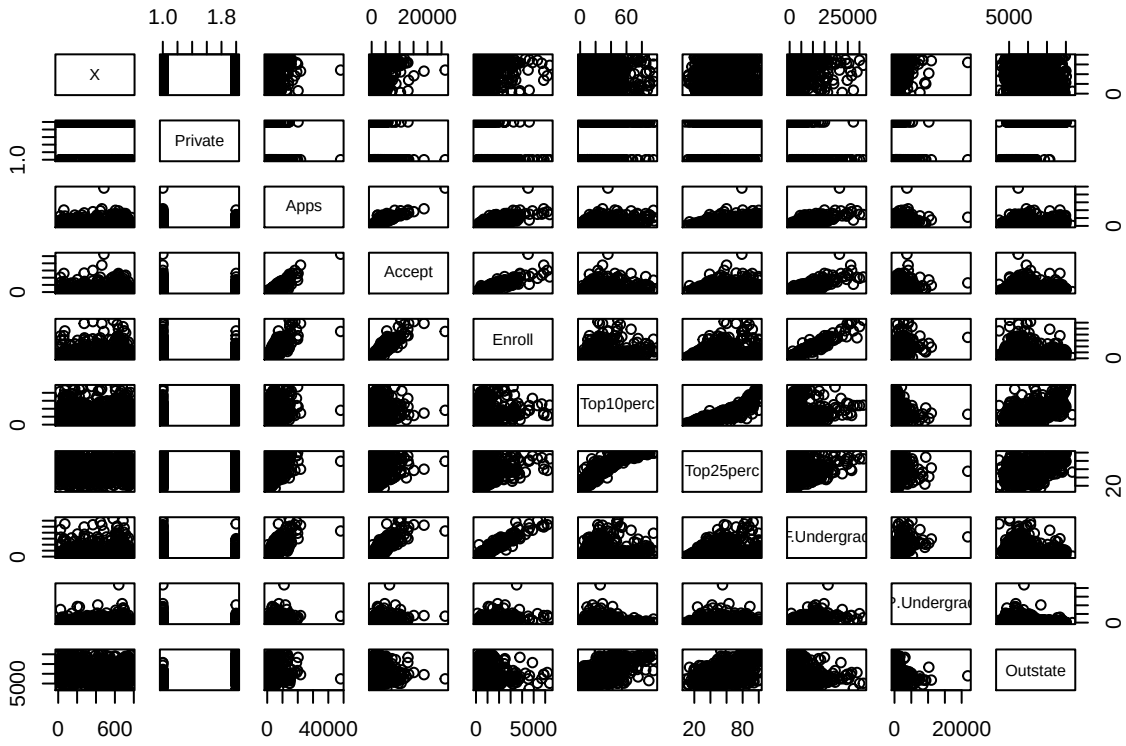
```
TableSummary = summary(college)
print(TableSummary)
```

```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University           :  1   Yes:565   1st Qu.:  776
##  Adrian College               :  1             Median : 1558
##  Agnes Scott College          :  1             Mean   : 3002
##  Alaska Pacific University    :  1             3rd Qu.: 3624
##  Albertson College            :  1             Max.   :48094
##  (Other)                      :771
##      Accept          Enroll       Top10perc       Top25perc
##  Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
```

```
##
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##
##      Books           Personal         PhD            Terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
##
```

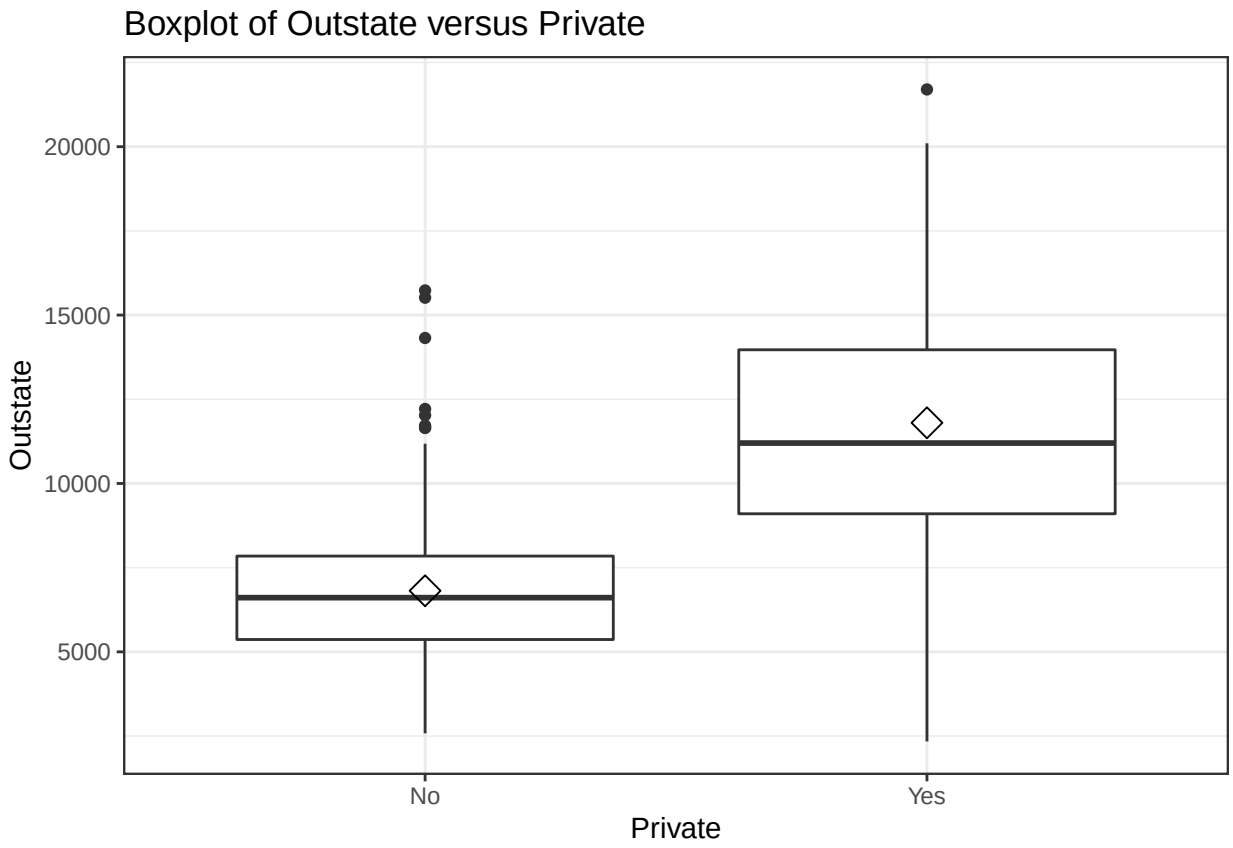
(ii). Use the `pair()` function to produce the scatterplot matrix of first ten columns or variables of the data

```
PairData = college[,1:10]
pairs(x = PairData)
```



(iii). Make a boxplot of side-by-side of Outstate versus Private

```
Boxplot1 = ggplot(data = college, mapping = aes(x = Private, y = Outstate)) +
  geom_boxplot() +
  ggtitle(label = "Boxplot of Outstate versus Private") +
  stat_summary(fun.y = mean, geom = "point", shape = 23, size = 4 ) +
  xlab(label = "Private") +
  ylab(label = "Outstate") +
  theme_bw()

print(Boxplot1)
```



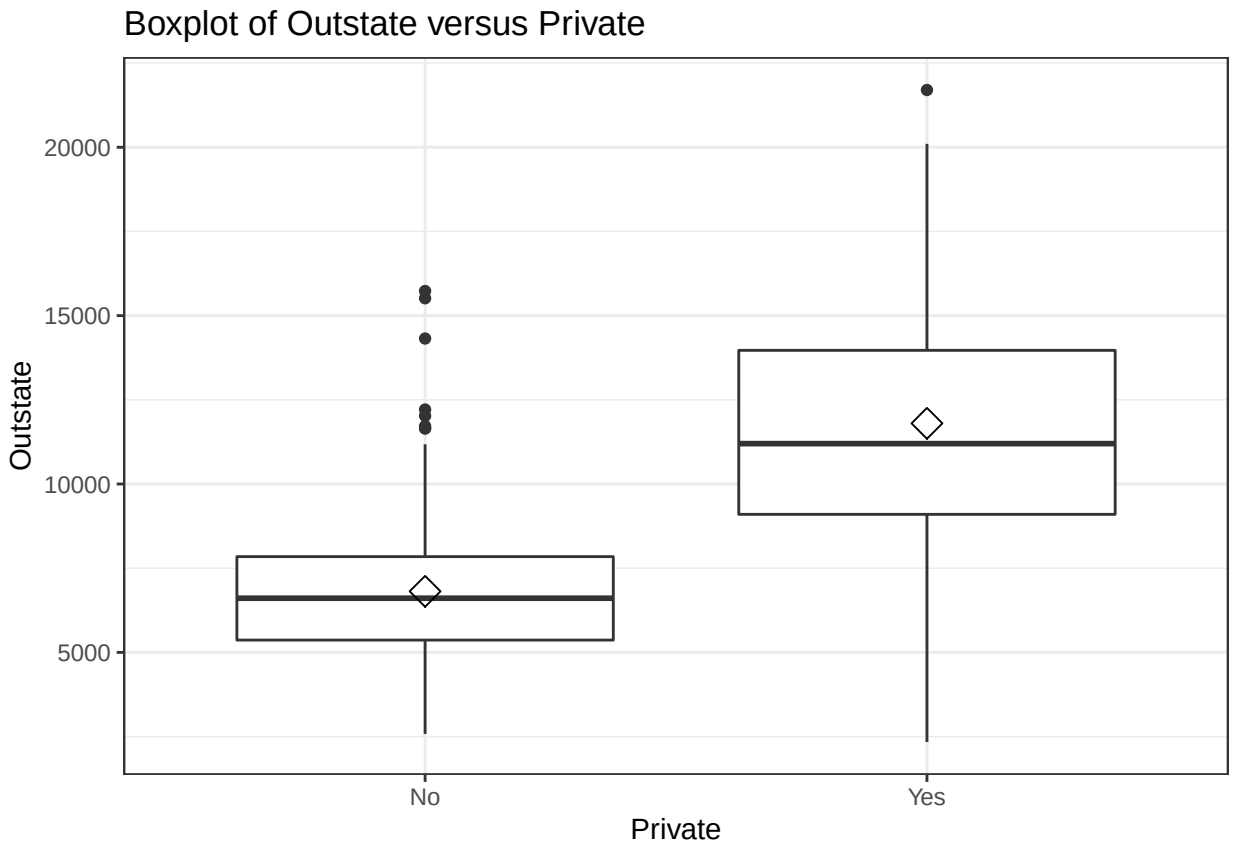
(iv). Create a qualitative variable `Elite`: binning the `Top10perc` variable: divide the universities into 2 groups based on prop. of students comming from the top 10% of high school exceeds 50%

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

### bind the dataframe
college = data.frame(college, Elite)

summary(college)
```

```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University           :  1   Yes:565   1st Qu.:  776
##  Adrian College               :  1             Median : 1558
```

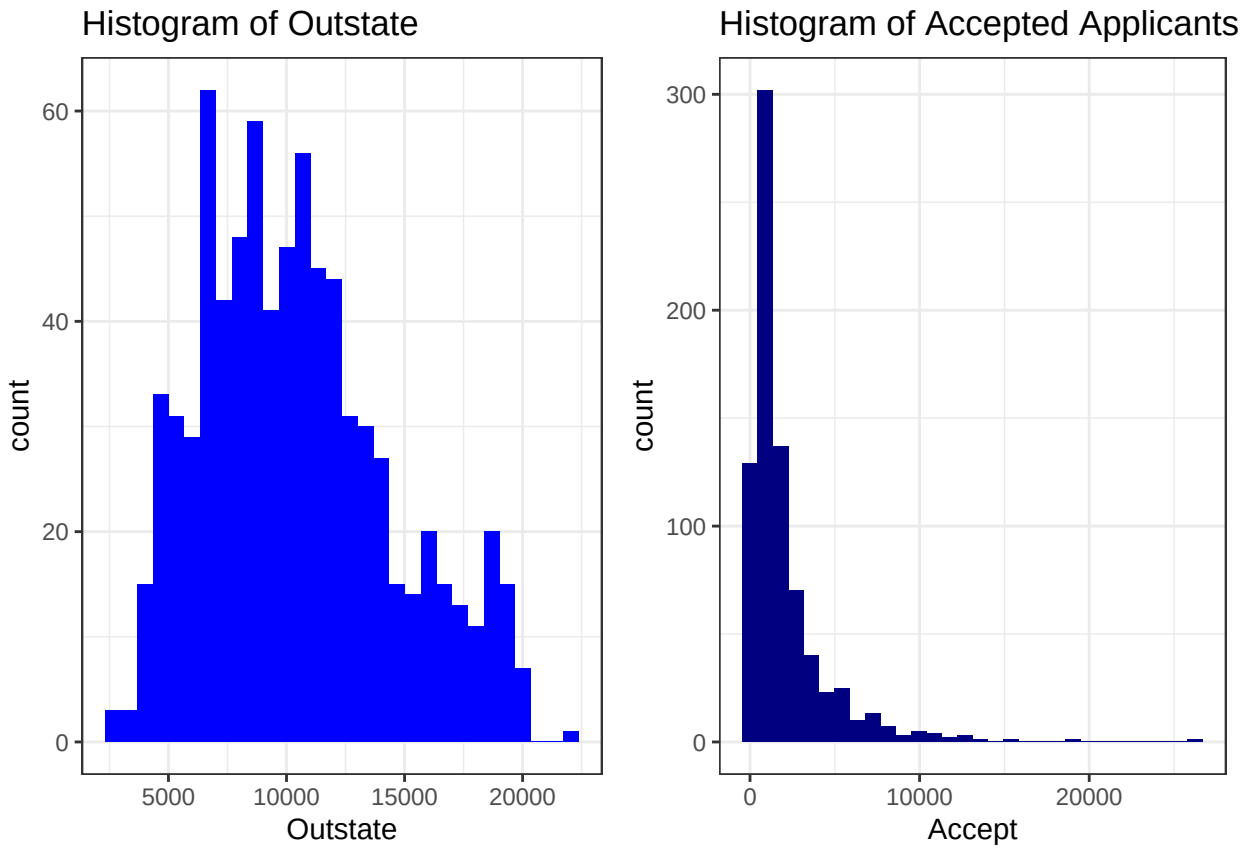
```
##  Agnes Scott College         :  1             Mean   : 3002
##  Alaska Pacific University   :  1             3rd Qu.: 3624
##  Albertson College           :  1             Max.   :48094
##  (Other)                     :771
##      Accept          Enroll       Top10perc       Top25perc
##  Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##
##      Books           Personal         PhD            Terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
##
##  Elite
##  No :699
##  Yes: 78
##
##
##
##
##
```

```
Boxplot2 = ggplot(data = college, mapping = aes(x = Private, y = Outstate)) +
  geom_boxplot() +
  ggtitle(label = "Boxplot of Outstate versus Private") +
  stat_summary(fun.y = mean, geom = "point", shape = 23, size = 4 ) +
  xlab(label = "Private") +
  ylab(label = "Outstate") +
  theme_bw()

print(Boxplot2)
```

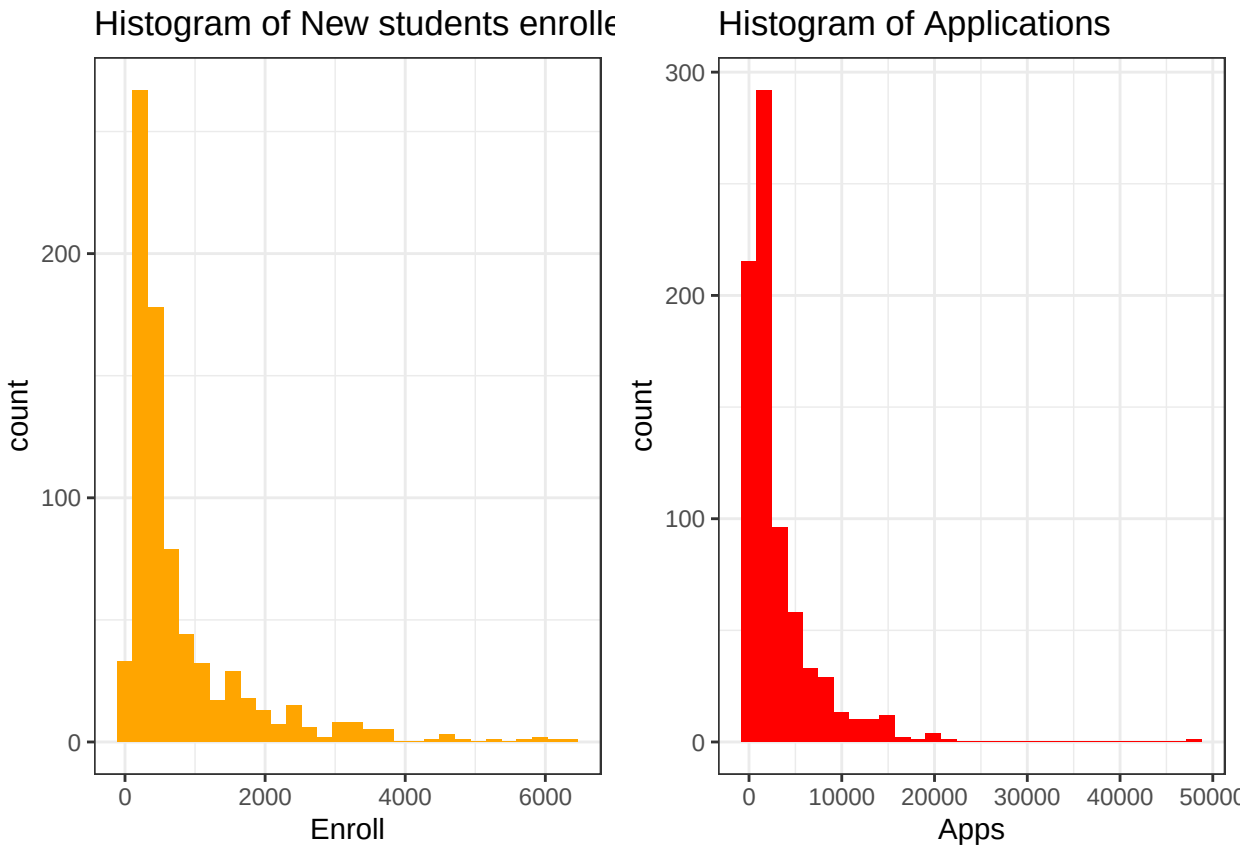


(v), Make histograms for a few quantitative variables

```
hist1 = ggplot(data = college, mapping = aes(x = Outstate)) +
  geom_histogram(fill = "blue") +
  ggtitle(label = "Histogram of Outstate") +
  theme_bw()

hist2 = ggplot(data = college, mapping = aes(x = Accept)) +
  geom_histogram(fill = "navy") +
  ggtitle(label = "Histogram of Accepted Applicants") +
  theme_bw()

hist3 = ggplot(data = college, mapping = aes(x = Enroll)) +
  geom_histogram(fill = "orange") +
  ggtitle(label = "Histogram of New students enrolled") +
  theme_bw()

hist4 = ggplot(data = college, mapping = aes(x = Apps)) +
  geom_histogram(fill = "red") +
  ggtitle(label = "Histogram of Applications") +
  theme_bw()


grid.arrange(hist1, hist2, ncol = 2)
```




```
grid.arrange(hist3, hist4, ncol = 2)
```



Exercise 9:

(a). Find the predictors: quantitative, qualitative

```
## set the working directory:
setwd(dir = "~/Desktop/Statistical Learning/dataset")

### read the datafile:

auto = read.csv(file = "Auto.csv")
### print out the structure:
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  : Factor w/ 94 levels "?","100","102",..: 17 35 29 29 24 42 47 46 48 40 ...
##  $ 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        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

- The quantitative predictors: `mpg`, `displacement`, `weight`, `acceleration`.
- The qualitative predictors: `cylinders`, `horsepower`, `year`, `origin`, `name`.

(b). Find the range of each quantitative predictor

```
## range of mpg:

range_mpg = range(auto$mpg)
range_displacement = range(auto$displacement)
range_weight = range(auto$weight)
range_acc = range(auto$acceleration)


print(list(
  range_mpg = range_mpg,
  range_displacement = range_displacement,
  range_weight = range_weight,
  range_acc = range_acc
))
```

```
## $range_mpg
## [1]  9.0 46.6
##
## $range_displacement
## [1]  68 455
##
## $range_weight
## [1] 1613 5140
##
## $range_acc
## [1]  8.0 24.8
```

(c). Find the mean and standard deviation of each quantitative predictor

```
op_auto  = auto[,c("mpg", "displacement", "weight", "acceleration")]

print(list(
  mean_mpg = mean(op_auto$mpg),
  mean_dis = mean(op_auto$displacement),
  mean_weight = mean(op_auto$weight),
  mean_acc = mean(op_auto$acceleration)
))
```

```
## $mean_mpg
## [1] 23.51587
##
## $mean_dis
## [1] 193.5327
##
## $mean_weight
## [1] 2970.262
##
## $mean_acc
## [1] 15.55567
```

```
print(list(
  sd_mpg = sd(op_auto$mpg),
  sd_dis = sd(op_auto$displacement),
  sd_weight = sd(op_auto$weight),
  sd_acc = sd(op_auto$acceleration)
```

```
))
```

```
## $sd_mpg
## [1] 7.825804
##
## $sd_dis
## [1] 104.3796
##
## $sd_weight
## [1] 847.9041
##
## $sd_acc
## [1] 2.749995
```

(d). Remove the 10th through 85th obsevation. Find the range, mean, and std. of each predictor in the subset of the data that remains

```
sub_auto = auto[-c(10:85),]

range_mpg = range(sub_auto$mpg)
range_displacement = range(sub_auto$displacement)
range_weight = range(sub_auto$weight)
range_acc = range(sub_auto$acceleration)


print(list(
  range_mpg = range_mpg,
  range_displacement = range_displacement,
  range_weight = range_weight,
  range_acc = range_acc
))
```

```
## $range_mpg
## [1] 11.0 46.6
##
## $range_displacement
## [1]  68 455
##
## $range_weight
## [1] 1649 4997
##
## $range_acc
## [1]  8.5 24.8
```

```
op_auto  = sub_auto[,c("mpg", "displacement", "weight", "acceleration")]

print(list(
  mean_mpg = mean(op_auto$mpg),
  mean_dis = mean(op_auto$displacement),
  mean_weight = mean(op_auto$weight),
  mean_acc = mean(op_auto$acceleration)
))
```

```
## $mean_mpg
## [1] 24.43863
##
```

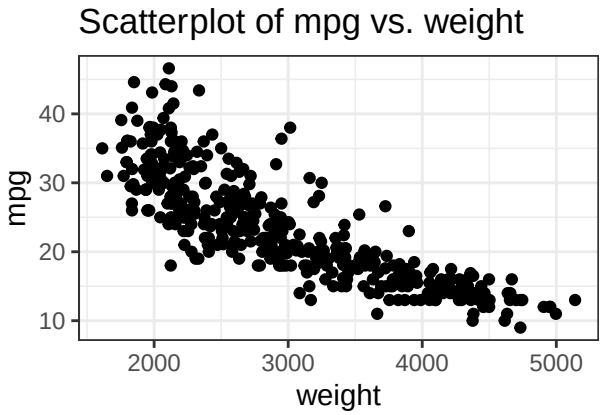
```
## $mean_dis
## [1] 187.0498
##
## $mean_weight
## [1] 2933.963
##
## $mean_acc
## [1] 15.72305
```

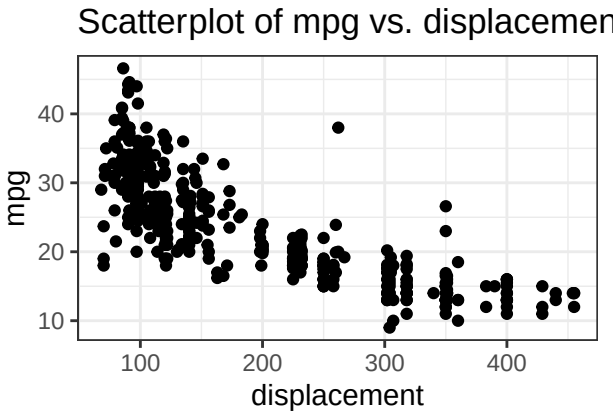
```
print(list(
  sd_mpg = sd(op_auto$mpg),
  sd_dis = sd(op_auto$displacement),
  sd_weight = sd(op_auto$weight),
  sd_acc = sd(op_auto$acceleration)
))
```

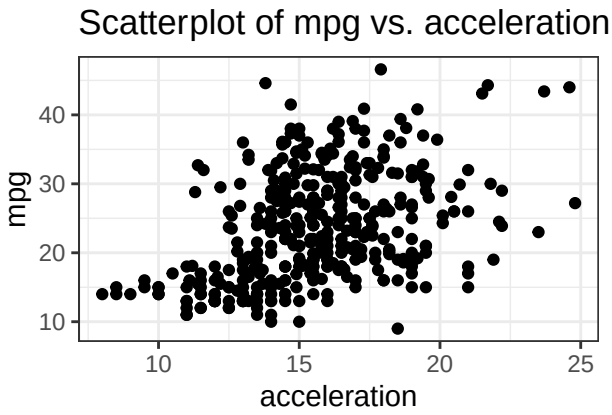
```
## $sd_mpg
## [1] 7.908184
##
## $sd_dis
## [1] 99.63539
##
## $sd_weight
## [1] 810.6429
##
## $sd_acc
## [1] 2.680514
```

(e). Make some plots highlighting the relationships among the predictors using the full data set

```
plot1  = ggplot(data = auto, aes(x = weight, y = mpg)) +
  geom_point() +
  ggtitle(label = "Scatterplot of mpg vs. weight") + theme_bw()

plot2  = ggplot(data = auto, aes(x = displacement, y = mpg)) +
  geom_point() +
  ggtitle(label = "Scatterplot of mpg vs. displacement") + theme_bw()

plot3 =  ggplot(data = auto, aes(x = acceleration, y = mpg)) +
  geom_point() +
  ggtitle(label = "Scatterplot of mpg vs. acceleration") + theme_bw()

grid.arrange(plot1,plot2,plot3, ncol = 2)
```







(f). We want to predict gas mileage (mpg) on the basis of other variables

- Based on the scatterplots above, we can predict `mpg` using the variables: `weight` and `displacement`. There is a linear negative trend between these variables.

Exercise 10:

(a). This exercise involves the `Boston` housing data set

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Read about the dataset:
## ?Boston

dim(Boston)
```

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

```
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
head(Boston)
```

```
##      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
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   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
## 6  5.21 28.7
```

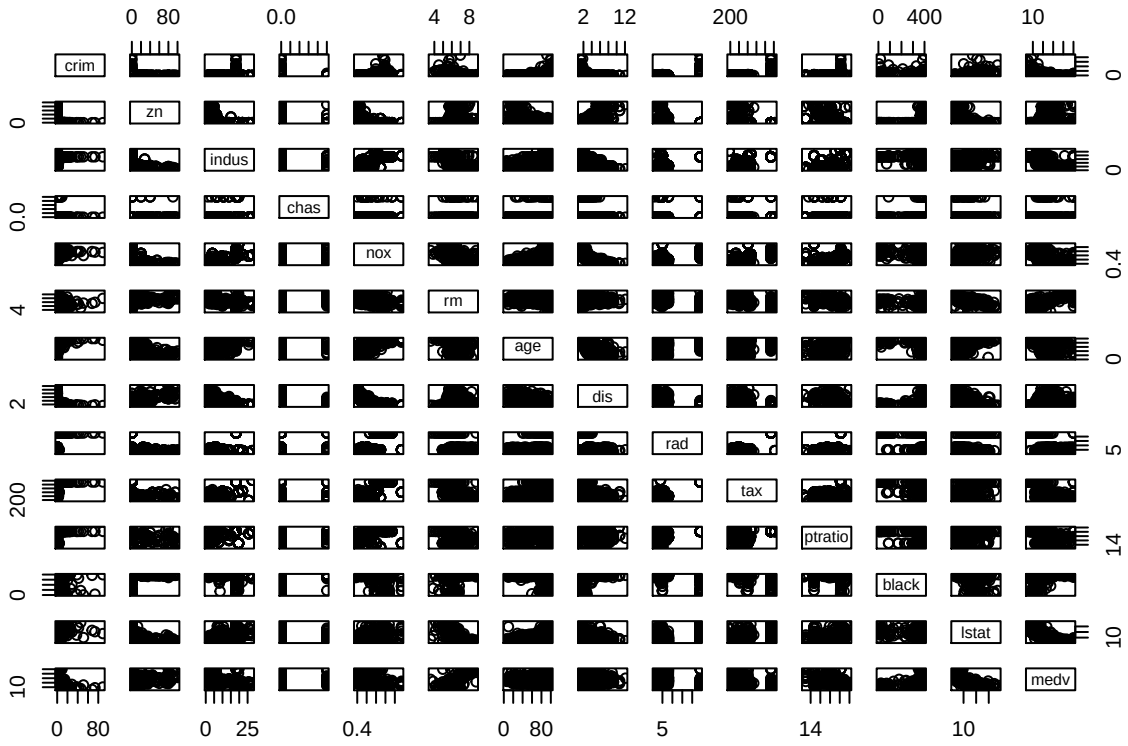
- There are 506 rows and 14 columns in data set. `crim` represent the per capita crime rate by town. `zn` is the prop. of residential land zoned for lots over 25,000 sqtft. `indus` is prop. of non-retail business acres per town. `chas` is Charles River dummy variable. `ncox` is nitrogent oxides concerntration. `rm` is average number of rooms per dwelling. `age` is prop. of owner-occupited units built prior to 1940. `dis` is weighted mean of distances to five Boston emp. centres.`rad` is index of acceibility of radial highways. `tax` is full prop. tax rate. `pratio` is pupil-teacher ratio by town. `black` prop. of blacks by town.`lstat` lower status of pop. percent. `medv` median value of owner occupied homes.

(b). Make some pairwise scatterplots

```
str(Boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
Boston$chas <- as.numeric(Boston$chas)
Boston$rad <- as.numeric(Boston$rad)
pairs(x = Boston)
```



- Not much can be discerned other than the fact that some variables appear to be correlated. A correlation matrix would be more helpful and question-c gives us the opportunity to make one.

(c). Make a correlation matrix to find association with the per capita crime rate

```
round(
  cor(x = Boston$crim, y = Boston[,c("zn", "indus","chas","nox", "rm", "age", "dis", "rad",
                                     "tax", "ptratio","black", "lstat","medv")]),
  digits = 3)
```

```
##        zn indus   chas   nox     rm   age   dis   rad   tax ptratio  black
## [1,] -0.2 0.407 -0.056 0.421 -0.219 0.353 -0.38 0.626 0.583    0.29 -0.385
##      lstat   medv
## [1,] 0.456 -0.388
```

- Based on the correlation coefficients and their corresponding p-values, there is indeed an association between the per capita crime rate (crim) and the other predictors.

Part (d)

```
summary(Boston$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```
summary(Boston$tax)
```

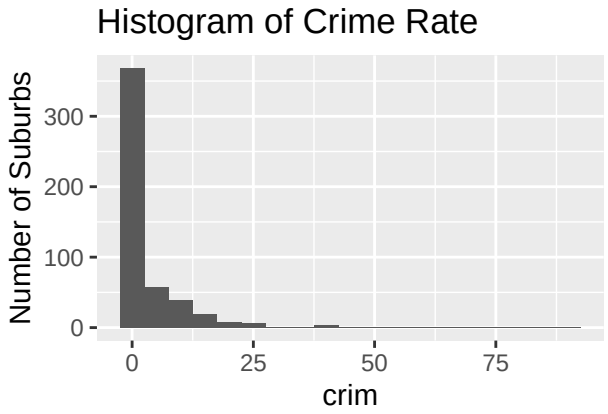
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

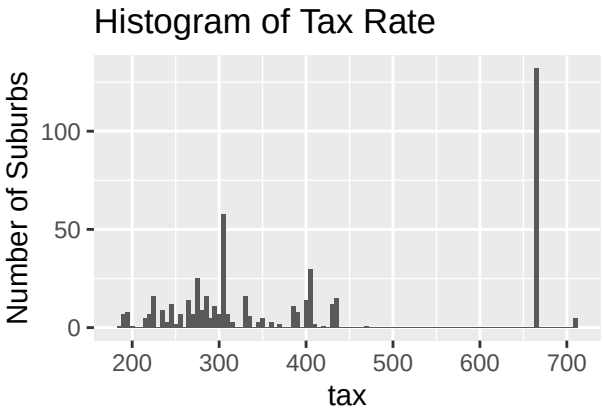
```
summary(Boston$ptratio)
```

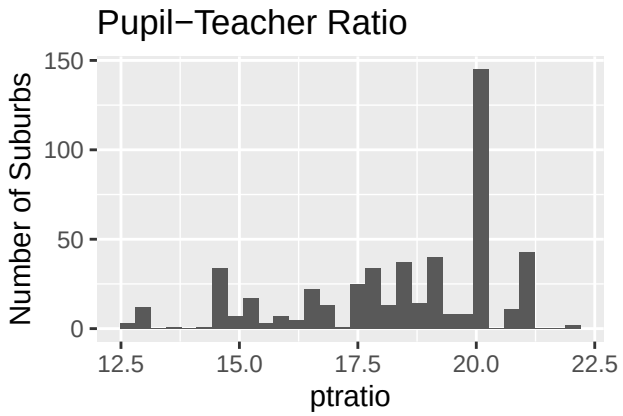
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

```
## Histogram:
Hist1 = ggplot(data = Boston, aes(x = crim)) +
  ggtitle(label = "Histogram of Crime Rate") +
  ylab(label = "Number of Suburbs") + geom_histogram(binwidth = 5)

Hist2 = ggplot(data = Boston, aes(x = tax)) +
  ggtitle(label = "Histogram of Tax Rate ") +
  ylab(label = "Number of Suburbs") + geom_histogram(binwidth = 5)

Hist3 = ggplot(data = Boston, aes(x = ptratio)) +
  ggtitle(label = "Pupil-Teacher Ratio ") +
  ylab(label = "Number of Suburbs") + geom_histogram()

grid.arrange(Hist1, Hist2, Hist3, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







- Considering that the median and maximum crime rate values are respectively about 0.26% and 89%, there are indeed some neighborhoods where the crime rate is alarmingly high

```
selection <- subset( Boston, crim > 10)
nrow(selection)/ nrow(Boston)
```

```
## [1] 0.1067194
```

- There is 11% of the neighborhood's have crime rates above 10%

```
selection <- subset( Boston, crim > 50)
nrow(selection)/ nrow(Boston)
```

```
## [1] 0.007905138
```

- There is 0.8% of the neighborhoods have crim rates above 50%.
- Based on the histogram of the Tax rates, they are few neighborhoods where rates are relative higher. The median and average tax amount are $330 and $408.20 (per Full-value property-tax rate per $10,000) respectively.

```
selection <- subset( Boston, tax< 600)
nrow(selection)/ nrow(Boston)
```

```
## [1] 0.729249
```

- 73% of the neighborhood pay under $600

```
selection <- subset( Boston, tax> 600)
nrow(selection)/ nrow(Boston)
```

```
## [1] 0.270751
```

- There is around 27% of the neighborhood pay over $600

(e). The number of surbibs in this data set bound to the Charles River

```
nrow(subset(Boston, chas == 1))
```

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

- There are 35 such surburbs bound to the Charles River.

(f). The median pupil-teacher ratio among the towns

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

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

- The median is 19.05

(g). Find the least median surburb? values of other predictors for that surburb, and compare to the overall ranges for those predictors.

```
selection <- Boston[order(Boston$medv),]
selection[1,]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9
##     lstat medv
## 399 30.59    5
```

- Suburb #399 with a median value of $5000. We can use the following summary information to answer part-2 of this question
- Crime is very high compared to median and average rates of all Boston neighborhoods.
- No residential land zoned for lots over 25,000 sq.ft. This applies to more than half of the neighborhoods in Boston * Proportion of non-retail business acres per town is very high compared to most suburbs.
- This suburd is not one of the suburbs that bound the Charles river.
- Nitrogen oxides concentration (parts per 10 million) is one of the highest.
- Average number of rooms per dwelling is one of the lowest
- Highest proportion of owner proportion of owner-occupied units built prior to 1940.

- One of the lowest weighted mean of distances to five Boston employment centres. * Highest index of accessibility to radial highways.
- One of the highest full-value property-tax rate per $10,000.
- One of the highest pupil-teacher ratio by town * Highest value for 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- One of the highest lower status of the population (percent)
- Lowest median value of owner-occupied homes in $1000s.

Based on the list above, suburb 399 can be classified as one of the least desirable places to live in Boston.

Part (h)

```
rm_over_7 <- subset(Boston, rm>7)
nrow(rm_over_7)
```

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

```
rm("rm_over_7")
```

- There are 64 suburbs with more than 7 rooms per dwelling.

```
rm_over_8 <- subset(Boston, rm>8)
nrow(rm_over_8)
```

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

```
summary(rm_over_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
```

- There are 13 suburbs with more than 7 rooms per dwelling