AppliedEx

Anuhya B S

2023-01-22

Applied Exercise

8. 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. The variables are

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10% of high school class
- `Top25perc` : New students from top 25% of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

(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 `View()` function. Store the college names as row names and remove the column.

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

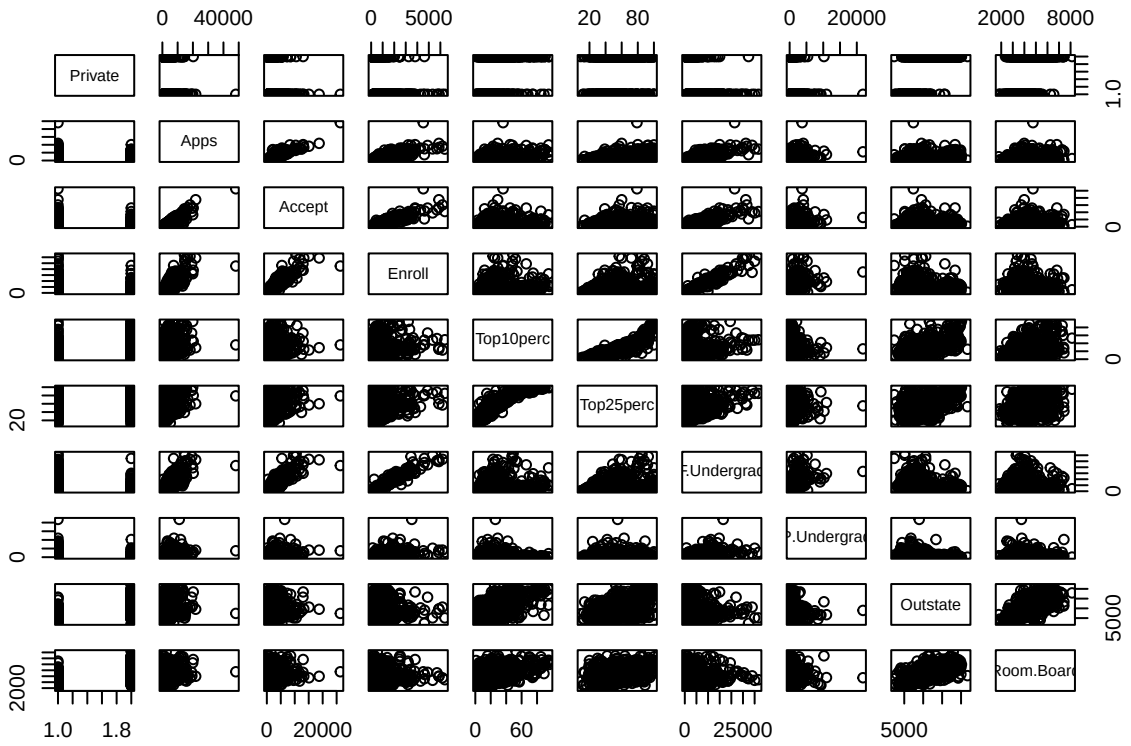
(c) i. Use the `summary()` function to produce a numerical summary of the variables in the data set.

```
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
```

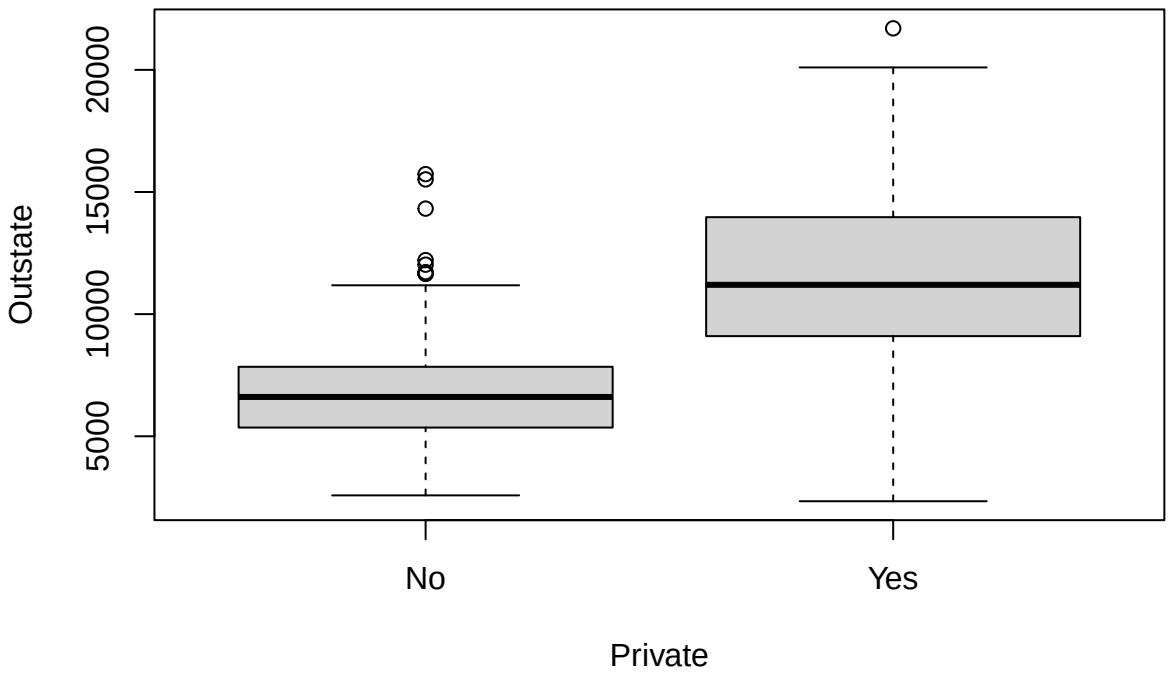
ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
college$Private <- as.factor(college$Private)
pairs(college[,1:10])
```



iii. Use the `plot()`s function to produce side-by-side boxplots of Outstate versus Private.

```
plot(college$Private,college$Outstate,xlab="Private",ylab="Outstate")
```



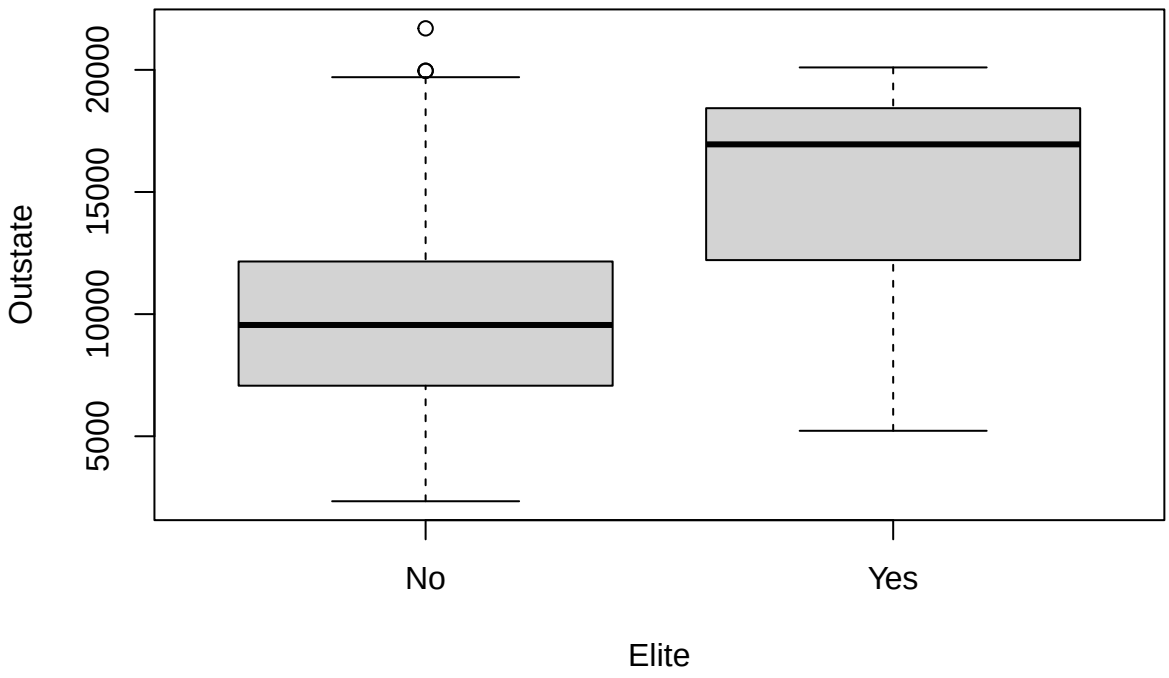
iv. Create a new qualitative variable, called `Elite`, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50 %. Use the `summary()` function to see how many 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 <- data.frame(college,Elite)
```

```
summary(Elite)
```

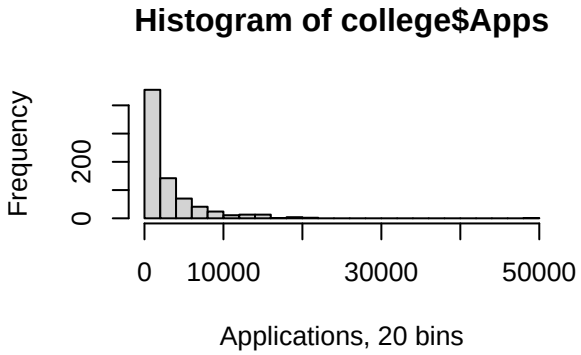
```
##  No Yes
## 699  78
```

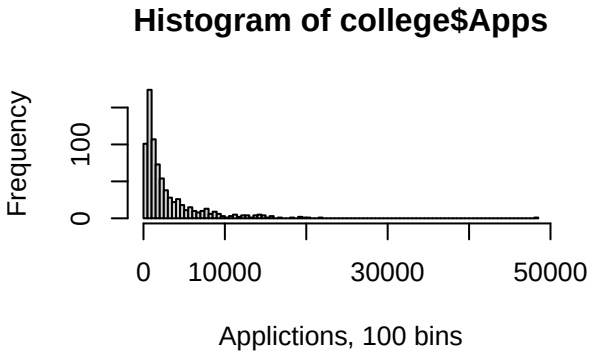
```
plot(college$Elite,college$Outstate,xlab="Elite",ylab="Outstate")
```

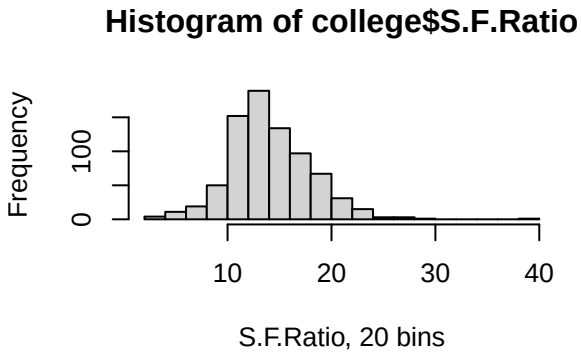


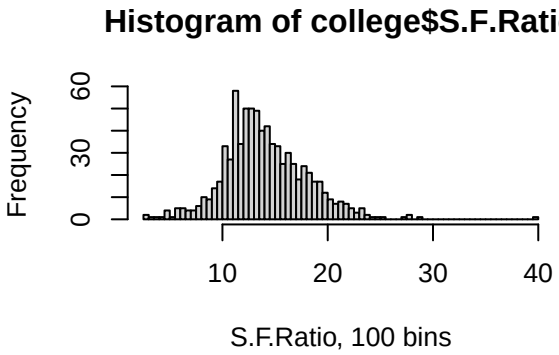
v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables.

```
par(mfrow=c(2,2))
hist(college$Apps, breaks = 20, xlab = "Applications, 20 bins")
hist(college$Apps, breaks = 100, xlab = "Applictions, 100 bins")
hist(college$S.F.Ratio, breaks = 20, xlab = "S.F.Ratio, 20 bins")
hist(college$S.F.Ratio, breaks = 100, xlab = "S.F.Ratio, 100 bins")
```









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

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

```
auto <- read.table("Auto.data", header = T,na.strings = "?", stringsAsFactors = T)
auto <- na.omit(auto)
head(auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

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

Quantitative: - mpg - displacement - horsepower - weight - acceleration

Qualitative: - cylinders - year - origin - name

(b) What is the range of each quantitative predictor?

```
quant = subset(auto,select=c('mpg','displacement','horsepower','weight','acceleration'))
t(sapply(quant, range))
```

```
##              [,1]   [,2]
## mpg             9   46.6
## displacement   68  455.0
## horsepower     46  230.0
## weight       1613 5140.0
## acceleration    8   24.8
```

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

```
t(sapply(quant, mean))
```

```
##           mpg displacement horsepower   weight acceleration
## [1,] 23.44592      194.412   104.4694 2977.584     15.54133
```

```
t(sapply(quant, sd))
```

```
##           mpg displacement horsepower   weight acceleration
## [1,] 7.805007      104.644   38.49116 849.4026     2.758864
```

(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?

```
new_df =quant[-c(10:85),]
apply(new_df,2,range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,] 11.0           68         46   1649          8.5
## [2,] 46.6          455        230   4997         24.8
```

```
apply(new_df,2,mean)
```

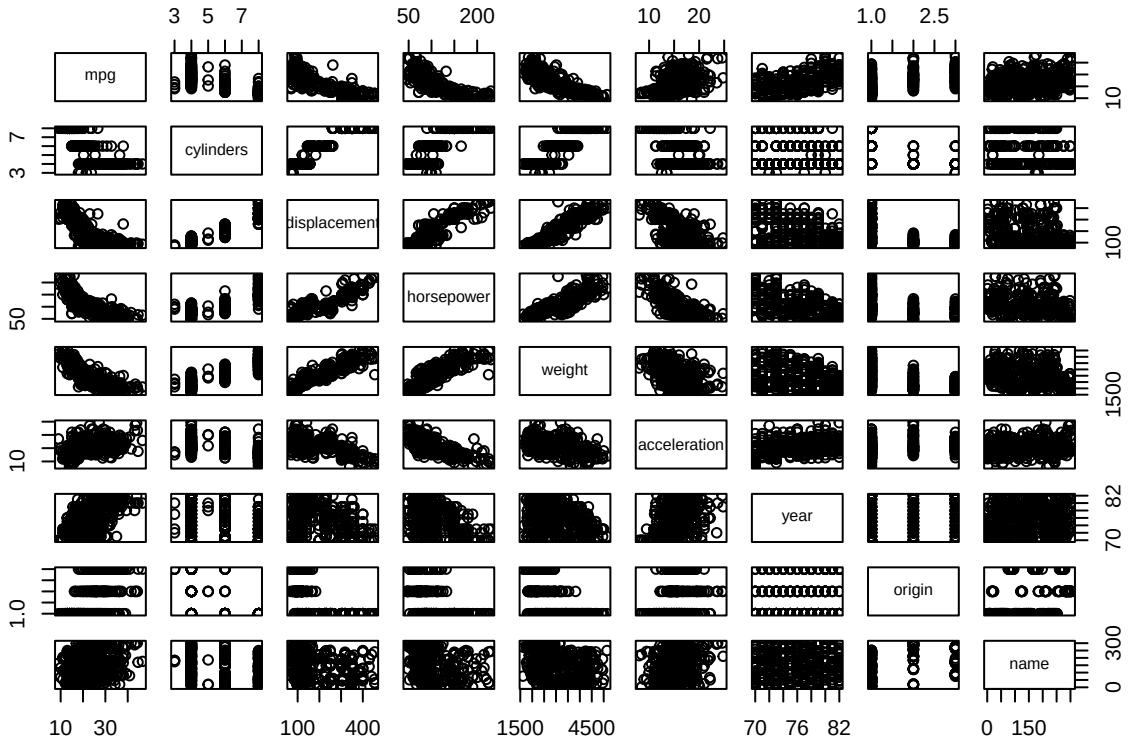
```
##          mpg displacement   horsepower       weight acceleration
##     24.40443    187.24051    100.72152   2935.97152     15.72690
```

```
apply(new_df,2,sd)
```

```
##          mpg displacement   horsepower       weight acceleration
##     7.867283    99.678367    35.708853   811.300208     2.693721
```

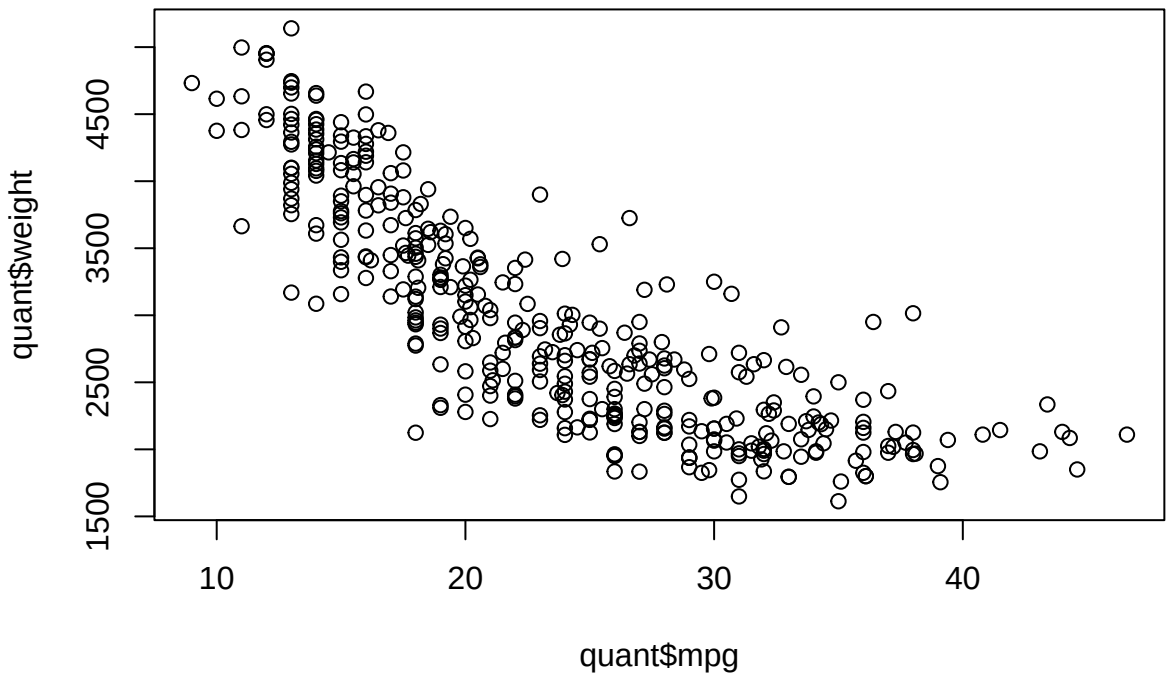
(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
pairs(auto)
```

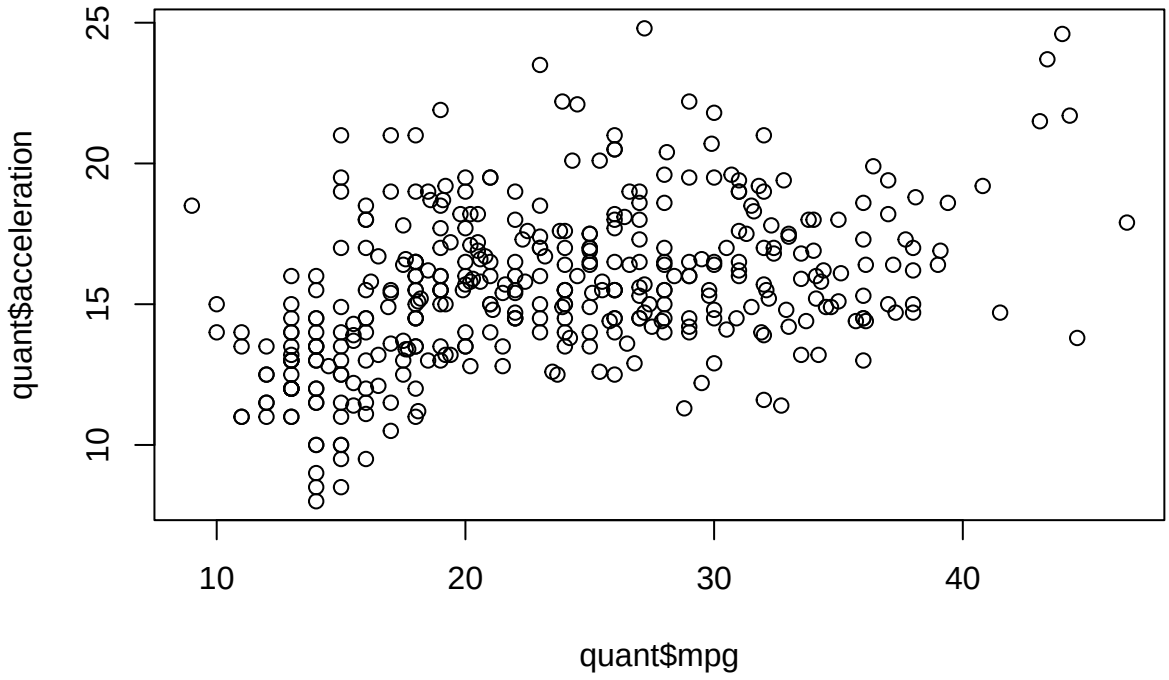


(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

```
plot(quant$mpg,quant$weight)
```




```
plot(quant$mpg,quant$acceleration)
```



10. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library. How many rows are in this data set? How many columns? What do the rows and columns represent?

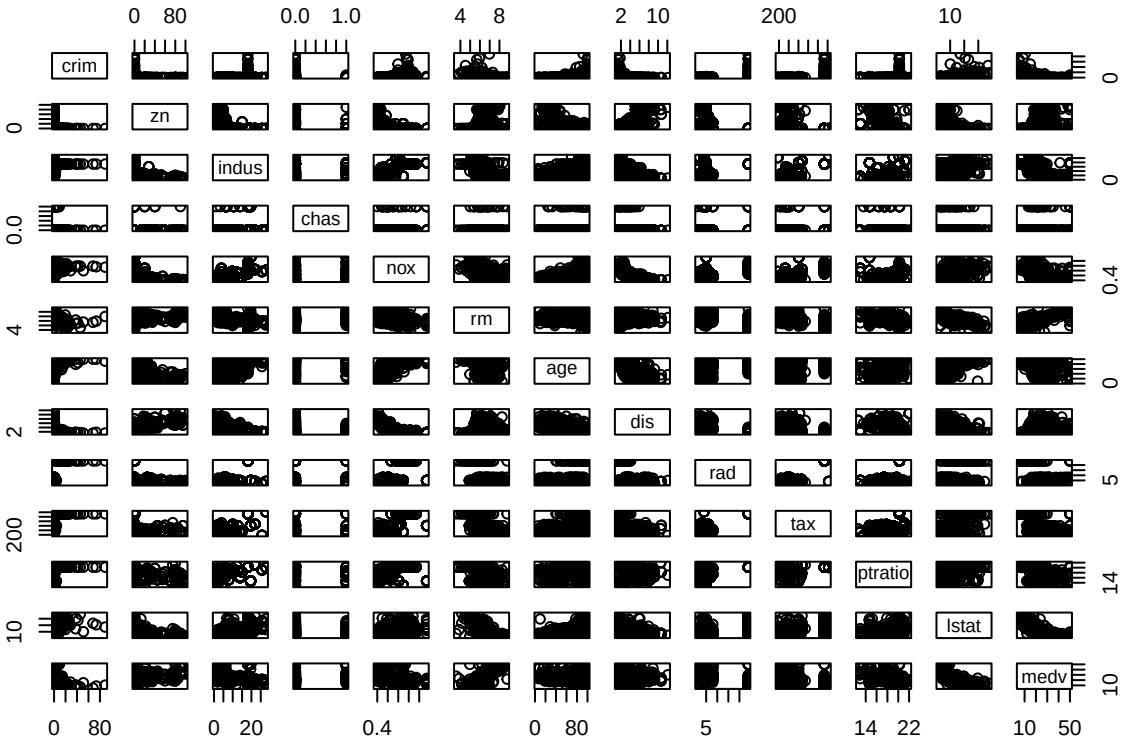
```
library(ISLR2)
dim(Boston)
```

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

A data frame with 506 rows and 13 variables (columns). The data set contains housing values in 506 suburbs of Boston. Rows represent data for the 506 suburbs of Boston. Columns represent the housing values for these suburbs.

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
pairs(Boston)
```




```
round(cor(Boston),2)
```

```
##           crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
## crim      1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
## zn       -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
## indus     0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
## chas     -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
## nox       0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
## rm       -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
## age       0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
## dis      -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
## rad       0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
## tax       0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
## ptratio   0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
## lstat     0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
## medv     -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
##          lstat  medv
## crim      0.46 -0.39
## zn       -0.41  0.36
## indus     0.60 -0.48
## chas     -0.05  0.18
## nox       0.59 -0.43
## rm       -0.61  0.70
## age       0.60 -0.38
## dis      -0.50  0.25
## rad       0.49 -0.38
```

```
## tax      0.54 -0.47
## ptratio  0.37 -0.51
## lstat    1.00 -0.74
## medv    -0.74  1.00
```

On a first glance, correlations between some of the predictors seem obvious. For example: 1. between the nitrogen oxides concentration and the proportion of owner-occupied units prior to 1940 2. between nitrogen oxides concentration and proportion of non-retail business acres per town 3. between nitrogen oxides concentration and weighted mean of distances to five Boston employment centres 4. weighted mean of distances to five Boston employment centres and the proportion of owner-occupied units prior to 1940

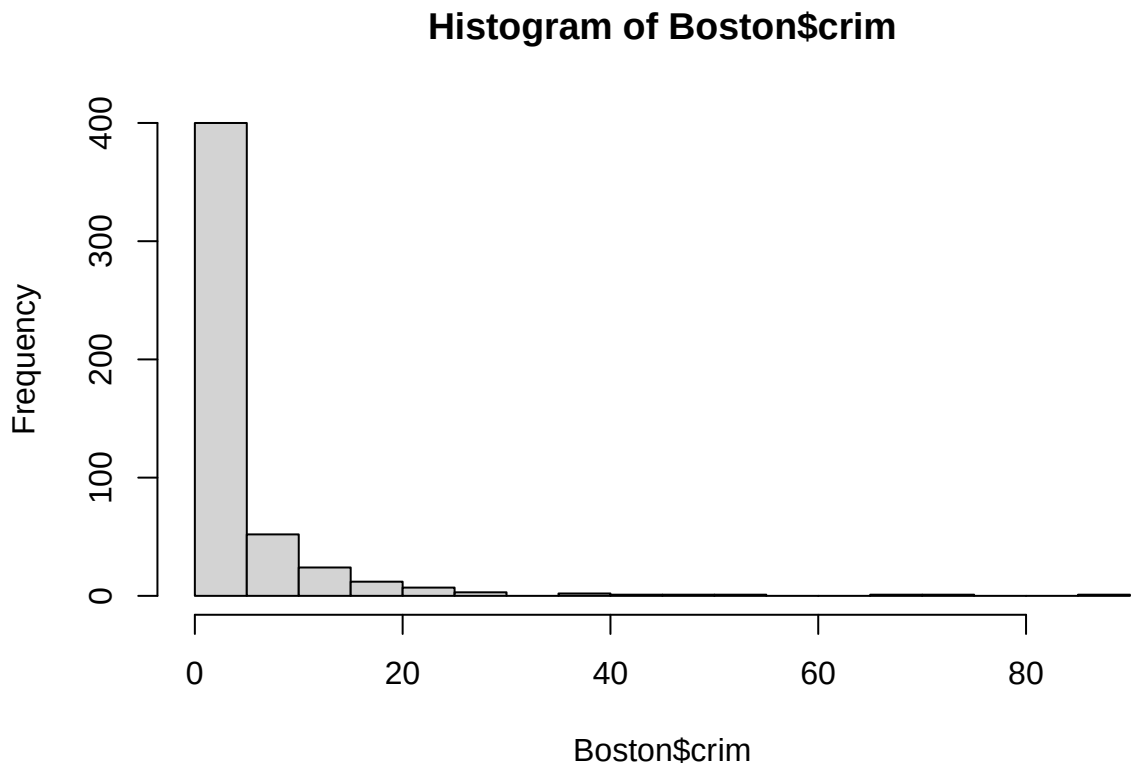
(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

There don't seem to be any very strong linear associations between the predictors and the per capita crime rate variable. However, based on the correlations it can be said that the variables rad(index of accessibility to radial highways), tax (full-value property-tax rate per $10,000), lstat (lower status of the population (percent)), nox(nitrogen oxides concentration (parts per 10 million)) and indus(proportion of non-retail business acres per town) are more associated with per capita crime rate.

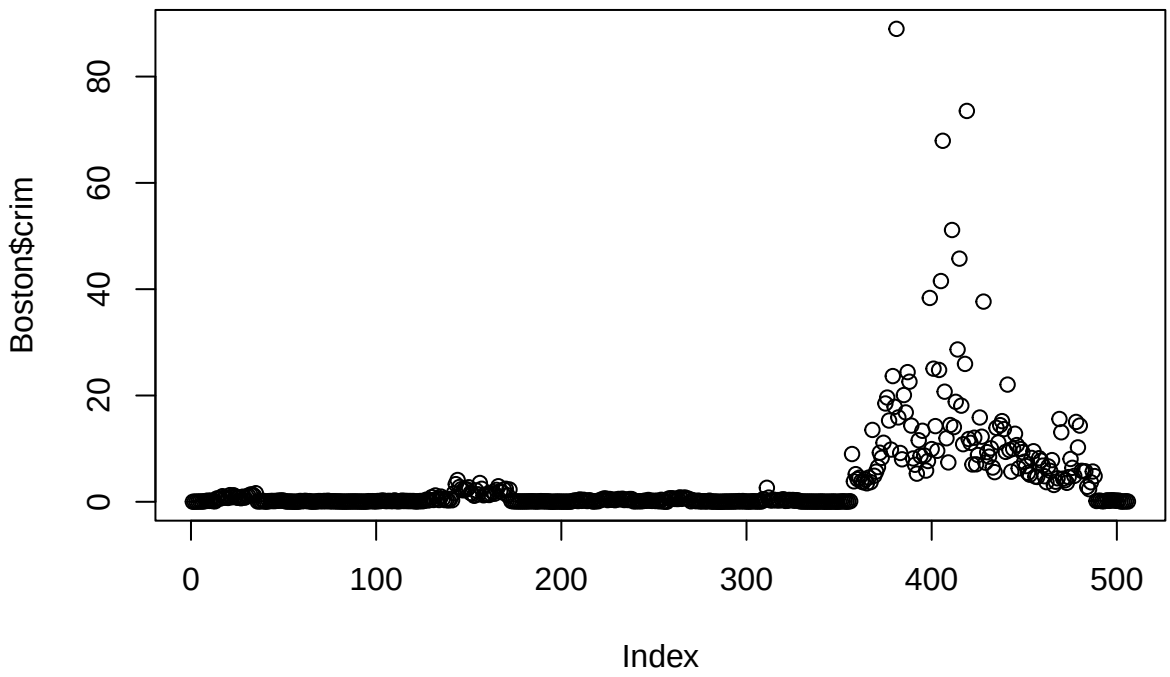
It can also be noted that all the data points seem to be right skewed for the predictors and per capita crime rate.

(d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
hist(Boston$crim,breaks = 20)
```

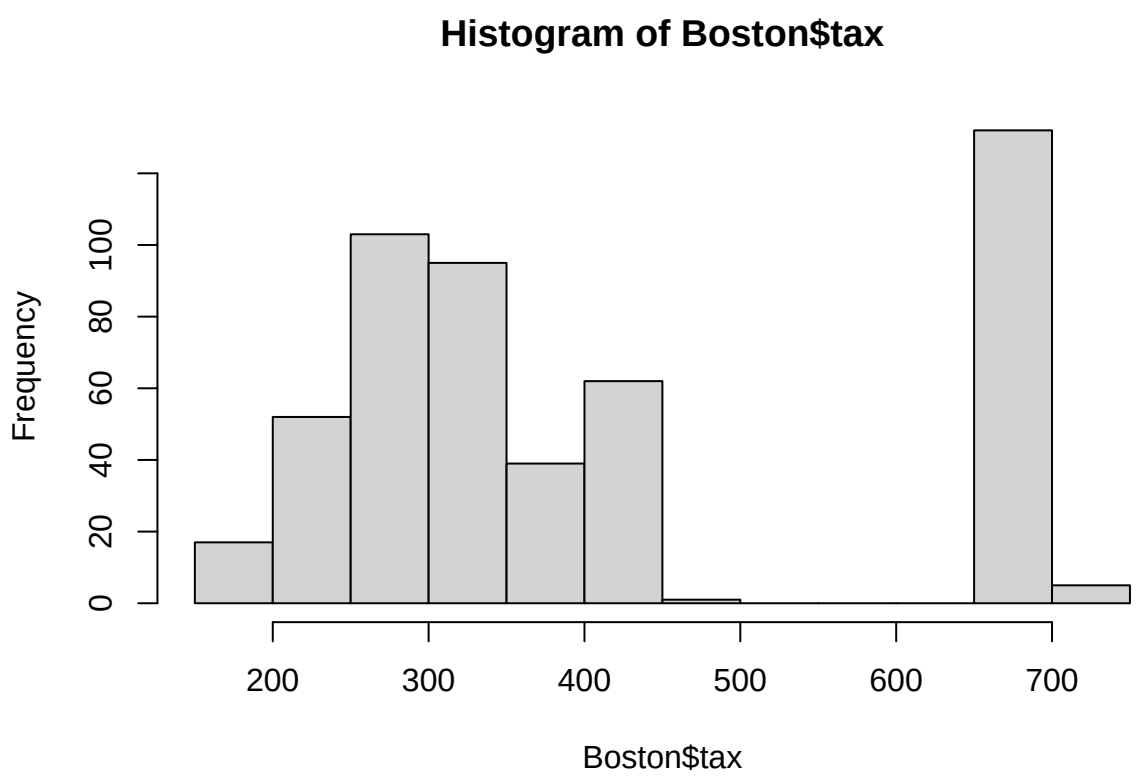


```
plot(Boston$crim)
```

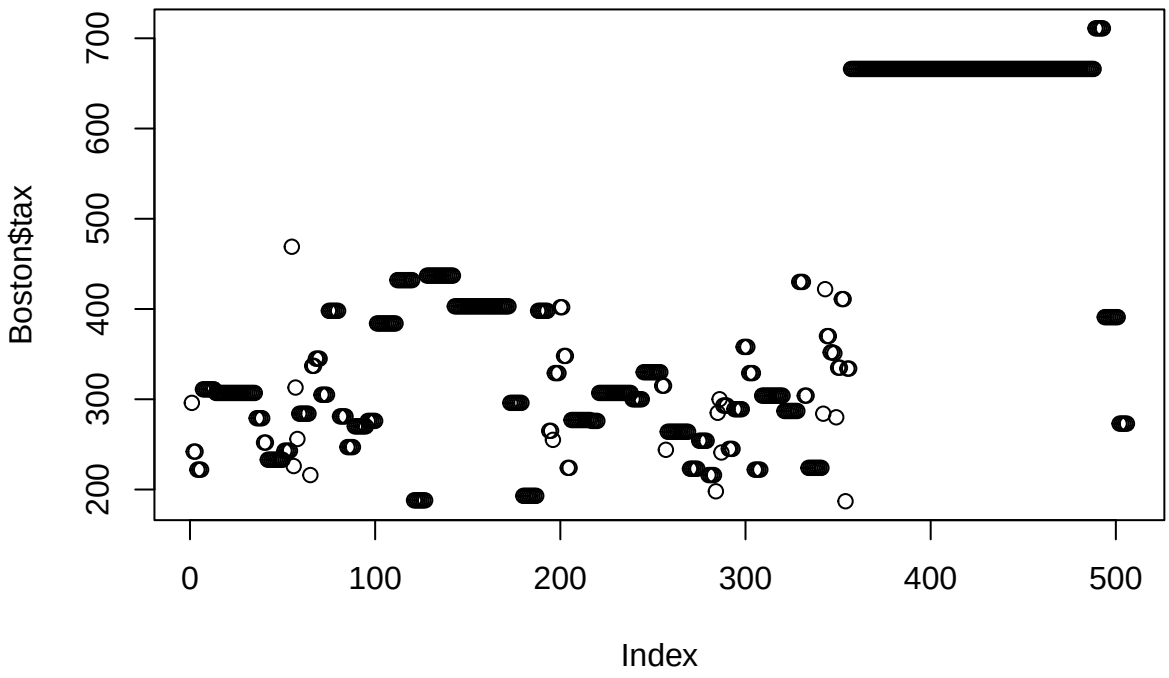


As it can be seen from the above plots, there are some suburbs (especially the ones between indexes 350-500) that have higher crime rate but most suburbs have low per capita crime rates.

```
hist(Boston$tax)
```

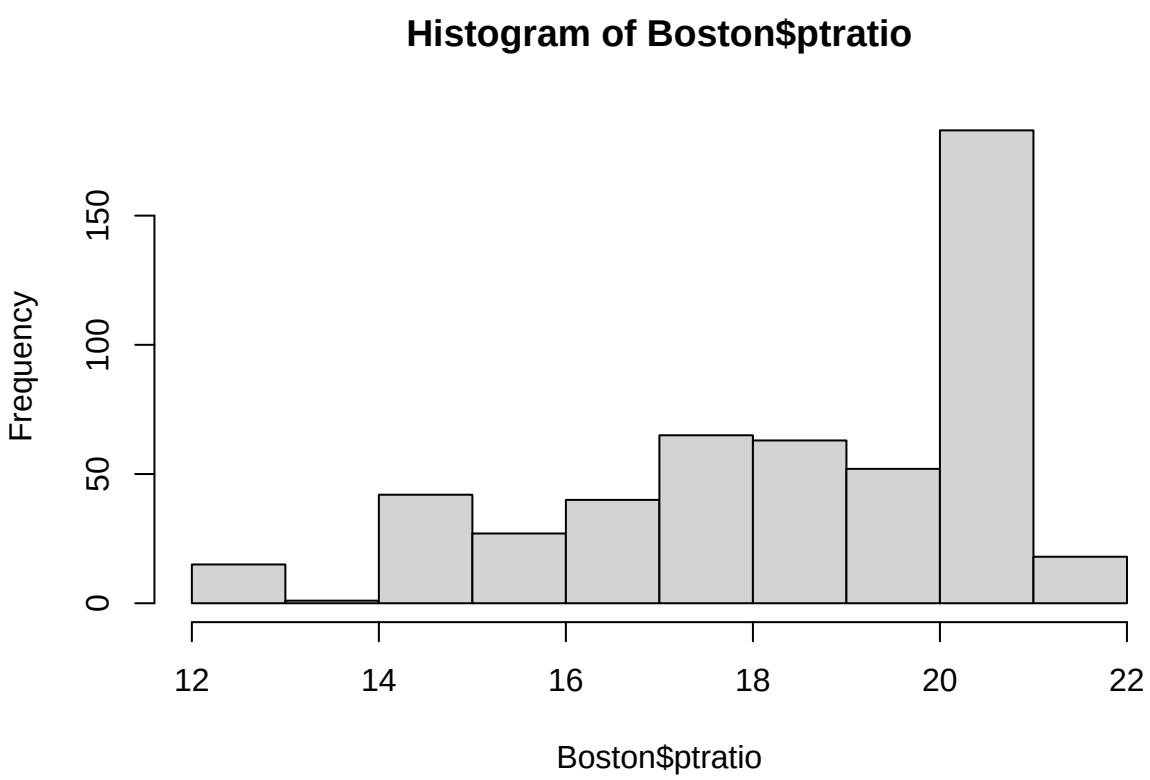



```
plot(Boston$tax)
```

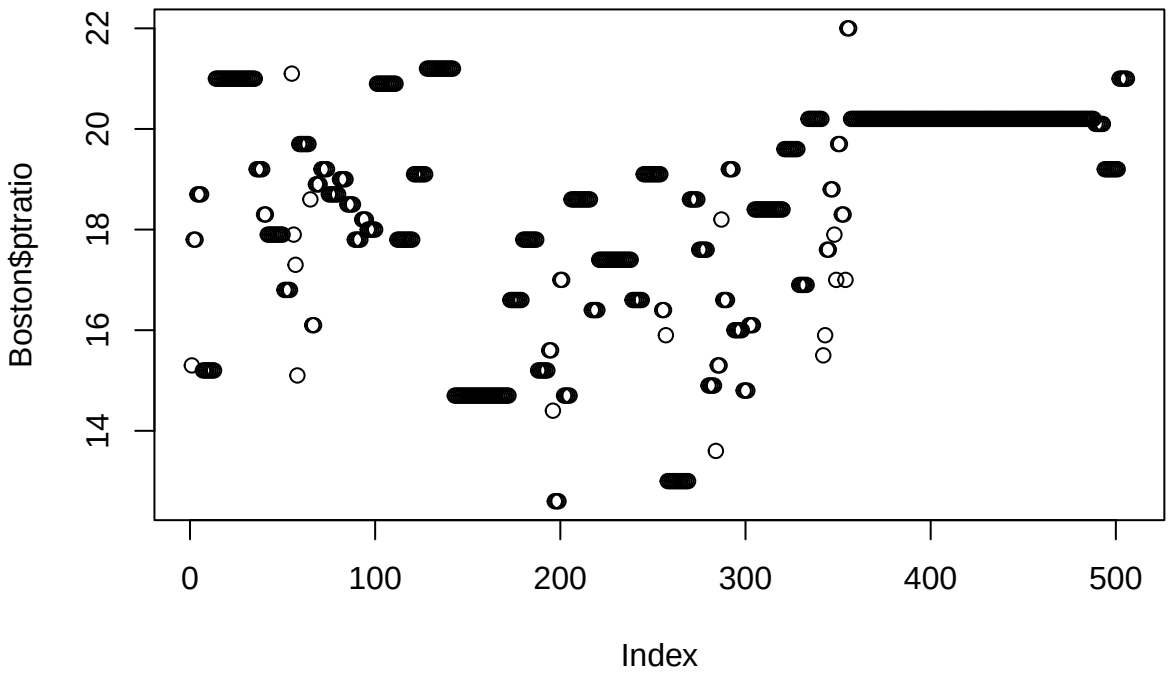


Yes, there are some suburbs that have particularly high rates as can be seen from the plots above.

```
hist(Boston$ptratio)
```




```
plot(Boston$ptratio)
```



The pupil teacher ratio is particulary high (between 20-21) for some suburbs as can be seen from the plots above.

The ranges of each of these predictors is as follows: 1. Crime : 0.00632-88.97620 2. Tax: 187-711 3. Pupil: Student Ratio : 12.6-22.0

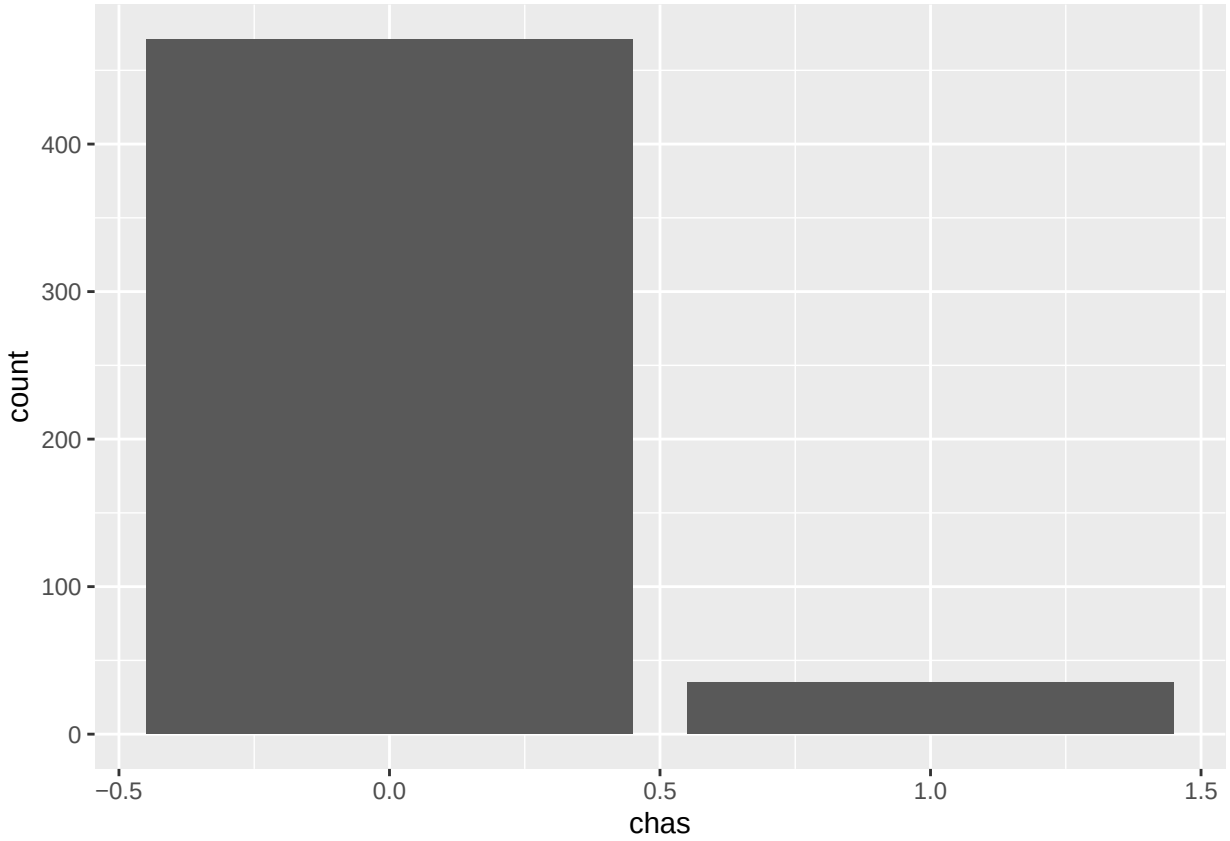
The crime rate on an avergae is lesser for all suburbs and the max value in the range of crime rate is particularly influenced by a single data point.

The max value of the range of tax rate is influenced by multiple suburbs in the dataset. It can also be seen that there is a huge gap in the histogram which means that most suburbs have tax rate less than 500 and then greater than 650.

The max value in the range of pupil-teacher ratio is due to a single suburb but there sames seem to be an uncommonly high number of suburbs that have high ptratio.

(e) How many of the census tracts in this data set bound the Charles river?

```
library(ggplot2)
ggplot(Boston, aes(x = chas)) +
  geom_bar()
```




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

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

There are 35 suburbs that bound the Charles river.

(f) What is the median pupil-teacher ratio among the towns in this data set?

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

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

(g) Which census tract of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
range(Boston$medv)
```

```
## [1]  5 50
```

```
low_medv = subset(Boston,medv==5)
nrow(low_medv)
```

```
## [1] 2
```

```
low_medv
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 22.98    5
```

```
apply(low_medv,2,range)
```

```
##         crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
## [1,] 38.3518  0  18.1    0 0.693 5.453 100 1.4254  24 666    20.2 22.98    5
## [2,] 67.9208  0  18.1    0 0.693 5.683 100 1.4896  24 666    20.2 30.59    5
```

```
apply(Boston,2,range)
```

```
##          crim  zn indus chas   nox    rm   age     dis rad tax ptratio lstat
## [1,]  0.00632   0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6  1.73
## [2,] 88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0 37.97
##      medv
## [1,]    5
## [2,]   50
```

There are 2 suburbs that have the lowest median value of owner-occupied homes (5).

(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

```
nrow(subset(Boston,rm>7))
```

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

```
nrow(subset(Boston,rm>8))
```

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

There are 64 suburbs that average more than 7 rooms per dwelling and 13 suburbs that average 8 rooms per dwelling.

Most suburbs that average more than 8 room per dwelling are older in age, relatively closer to radial highways, have lower crime rates and are not really bound the Charles river.