Chase Moreno

@2078503

MATH 4323

Homework 1

1. Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide the sample size and the number of variables for each scenario.

   (a) We are considering launching a new product and wish to know whether it will be a success or failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   It is classification and prediction because we are wanting to find out whether a given product is success or failure. Also, predicting whether it's a success or a failure of the product. The size is 20 products and the number of variables for is 13.

   (b) We are trying to figure out the factors potentially leading to cancer. For this, we collect data on 200 patients that either had or didn't have cancer. In particular, we record their body measurements (weight, height etc), heart rate, blood sugar level, family history of disease - measuring twelve variables. On top of that, we conduct a survey on exercise, eating and drinking habits, adding another three variables.

   It is classification and inference because we are trying to figure it out on each patient that either had or didn't have cancer from the outcome. Most interested in inference because we identify the variables that enable us to determine the outcome of cancer. The size is 200 patients, and the number of variables is 12, but could be more since it mentioned "etc" from body measurements.

   (c) We dig into UH student database and obtain data on 1500 students. We are interested in figuring out the factors affecting the final year GPA of a student depending on the data from the entrance exams, high school and their first year at UH. We extract students' SAT scores, high school GPA, first year GPA at UH, major, age and other five variables.

   It is regression and inference because we are figuring out the factors affecting the final year GPA of a students and not by predicting. The size is 1500 students and the number of variables is 10.

(d) We would like to predict the outcome of a college football game (not the exact score, but simply who wins, team A or team B?) depending on various factors. Assuming that it is mid-season, we obtain the data on all 300 games that have already been played and study such variables as: yards gained/allowed, touchdowns scored/allowed, turnovers committed/forced, whether the game is at home or away, whether it rains or not (all-in-all eight variables).

It is classification and prediction because we are predicting the outcome of a college football game either team A or team B wins. Collecting all data from all 300 games and based on the 12 variables then we are predicting which teams is most likely to win. The size is 300 games, and the number of variables is 8, because it mentioned "(all-in-all eight variables)."

2. (a) What type of statistical learning do all data examples in Problem 1 correspond to - supervised or unsupervised?

Part b and d are supervised. Part a and c are unsupervised.

(b) We know that general model formula is

$$Y = f(X) + \epsilon, \qquad (1)$$

and we try to estimate true f with ˆf. For each of examples (a),(b),(c) from Problem 1, proceed to answer the following:

i. Can our estimate ˆf be treated as a black box?

ii. Why/Why not?

Part a and d can estimate ^f be treated as a black box because it is a prediction.

Part b and c can't estimate ^f be treated as a block box because it is an inference.

(c) When estimating Y from equation (1) with Yˆ = ˆf(X), what two errors can we commit? Which one of them can be improved via a better statistical learning technique? Which one of them can't be improved & why?

The two errors that we can commit are reducible and irreducible errors. I believe reducible can be improved via a better statistical learning technique. On the other hand, irreducible error can't be improved because it will affect the accuracy of the predictions and X variables cannot be predicted.

3. Provide three data examples of unsupervised learning task (on your own, and you can't use the ones already mentioned in class, including the intro lecture), with all of them being from different application areas. Formulate what are the subjects (doesn't have to be people) of interest you are trying to group/cluster, and according to what potential predictors/characteristics. Areas may include, but not limited to, medicine, finance,

economics, sociology, education, marketing, journalism, sports, oil industry, meteorology, etc.

1) **Walmart wants to boost their sales. So, they decide to observe the data based on its sold products. They discovered that customers who bought salsa are most likely to buy chips. Also, customers who bought meatballs tend to buy spaghetti noodles.**

2) **Before the NFL Draft, we have 150 college football players are doing the 40-yard dash where players try to run the fastest time in 40 yards. There is a scout who collected every player's time, but without information about who they are. Then this scout separated the players' time individually.**

3) **A high school student used a personal telescope and saw 25 stars. After seeing the stars, this student doesn't know what they are but dividing them by the color of the star.**

4. This exercise relates to the Credit data set, which can be found in the file Credit.csv. It contains information about credit card debt for 10,000 credit card holders.

(a) Use RStudio's drop-down menu (Environment → Import Dataset → From Text (base) ...) to read the data into R. Make sure the Heading is set to Y es. Call the loaded data Credit.

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

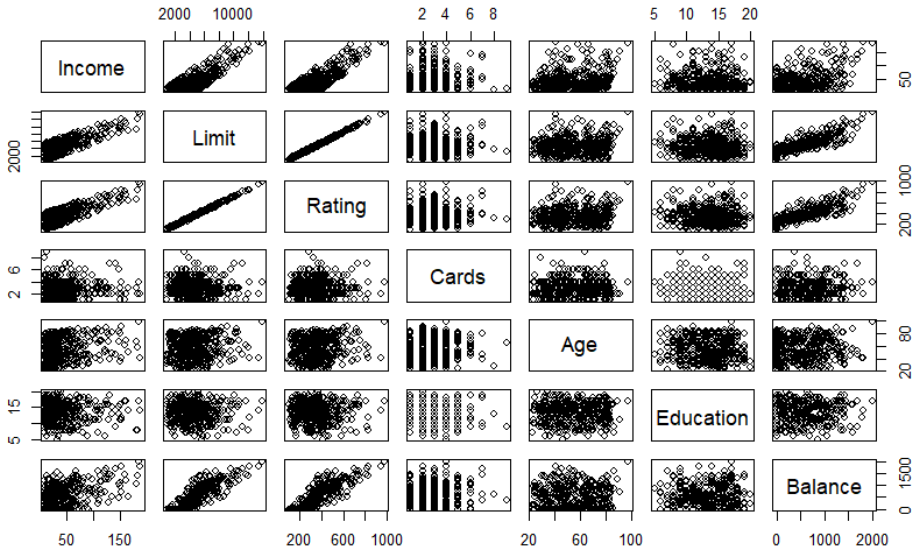
```
> summary(Credit)
     Income           Limit           Rating          Cards            Age
 Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00
 1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75
 Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00
 Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67
 3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00
 Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00
   Education         Own               Student            Married             Region
 Min.   : 5.00   Length:400         Length:400         Length:400         Length:400
 1st Qu.:11.00   Class :character   Class :character   Class :character   Class :character
 Median :14.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :13.45
 3rd Qu.:16.00
 Max.   :20.00
    Balance
 Min.   :   0.00
 1st Qu.:  68.75
 Median : 459.50
 Mean   : 520.01
 3rd Qu.: 863.00
 Max.   :1999.00
>
```

ii. Which columns contain numerical values? Which columns contain categorical values?

Income, Limit, Rating, Cards, Age, Education, and Balance are columns contains numerical values. Own, Student, Married, and Region are columns contains categorical values.

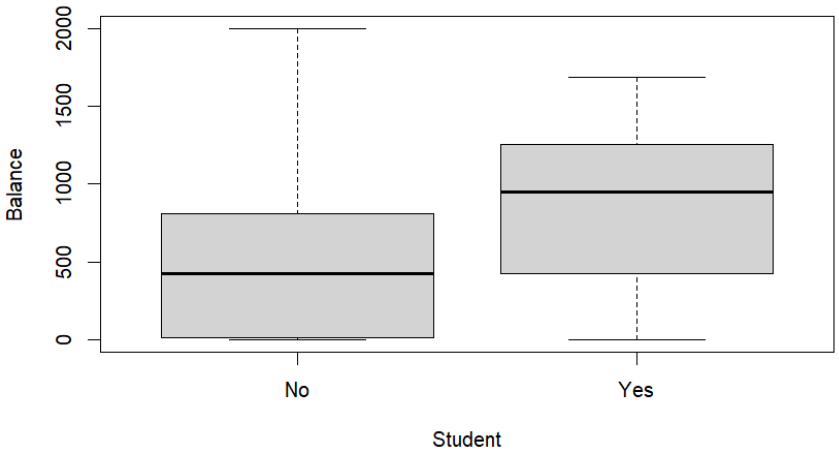
iii. Use the pairs() function to produce a scatterplot matrix of the quantitative variables in the dataset. Note that the pairs() function requires the input as numeric values, so you have to think about how to select the numerical columns from the dataset.

```
> pairs(~ Income + Limit + Rating + Cards + Age + Education +
Balance, Credit)
```



iv. Use the plot() function to produce side-by-side boxplots of Balance versus Student.

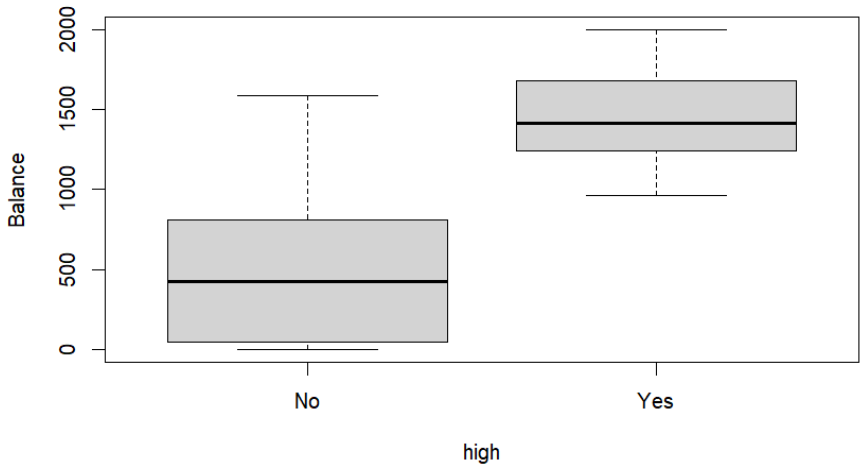
```
> boxplot(Balance ~ Student, Credit)
```



v. Create a new qualitative variable, called high, by binning the Rating variable. We are going to divide the card holders into two groups based on whether their credit ratings exceed 680. > Credit$high=ifelse(Credit$Rating>680,"Yes","No") Use the table() function to see how many card holders in this dataset have high credit ratings. Now use the plot() function to produce side-by-side boxplots of

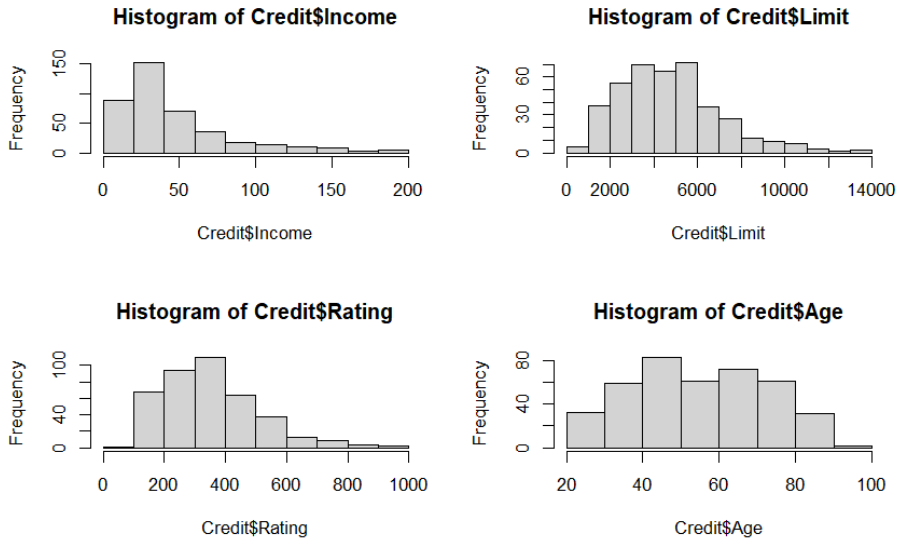
Balance versus high. What would you comment on the findings based on the boxplot?

```
> Credit$high = ifelse(Credit$Rating > 680, "Yes", "No")
> table(Credit$high)

 No Yes
382  18

> boxplot(Balance ~ high, Credit)
```



vi. Use the hist() function to produce some histograms with differing numbers of bins (e.g. 5, 10, 15, 20) for a few of the quantitative variables (e.g. Income and Age). You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
> par(mfrow = c(2,2))
> hist(Credit$Income)
> hist(Credit$Limit)
> hist(Credit$Rating)
> hist(Credit$Age)
```



5. This exercise involves the Boston data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

> library(MASS)

Now the data set is contained in the object Boston .

> Boston

Read about the data set:

> ?Boston

How many rows are in this data set? How many columns? What do ”lstat”, ”ptratio”,”chas”, and ”medv” represent?

There are 506 rows and 14 columns.

lstat represent lower status of the population (percent).

ptratio represent pupil-teacher ratio by town.

chas represent Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

medv represent median value of owner-occupied homes in $1000.

(b) Of what type are most of the predictors - quantitative or qualitative?

Quantitative

(c) What is the range of each predictor? You can answer this either by applying the range() function to each predictor, or by using summary() function on the whole data set and extracting the range from there. Please provide a table with ranges for all predictors:

```
> sapply(Boston, range)
         crim  zn indus chas   nox    rm   age     dis rad tax ptratio  black lstat medv
[1,]  0.00632   0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6   0.32  1.73    5
[2,] 88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0 396.90 37.97   50
```

[1,] = Min

[2,] = Max

(d) What is the mean and standard deviation of each quantitative predictor? Provide the answer in the form of a table:

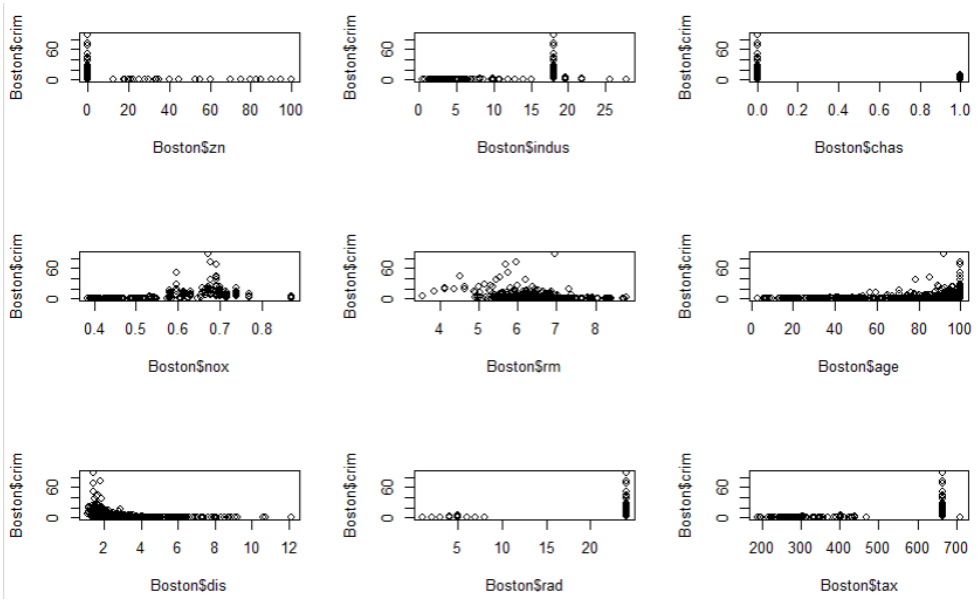
```
> sapply(Boston, mean)
       crim          zn       indus        chas         nox          rm         age
 3.61352356 11.36363636 11.13677866  0.06916996  0.55469506  6.28463439 68.57490119
        dis         rad         tax     ptratio       black       lstat        medv
 3.79504269  9.54940711 408.23715415 18.45553360 356.67403162 12.65306324 22.53280632

> sapply(Boston, sd)
       crim          zn       indus        chas         nox          rm         age         dis
 8.6015451  23.3224530   6.8603529   0.2539940   0.1158777   0.7026171  28.1488614   2.1057101
        rad         tax     ptratio       black       lstat        medv
 8.7072594 168.5371161   2.1649455  91.2948644   7.1410615   9.1971041
```

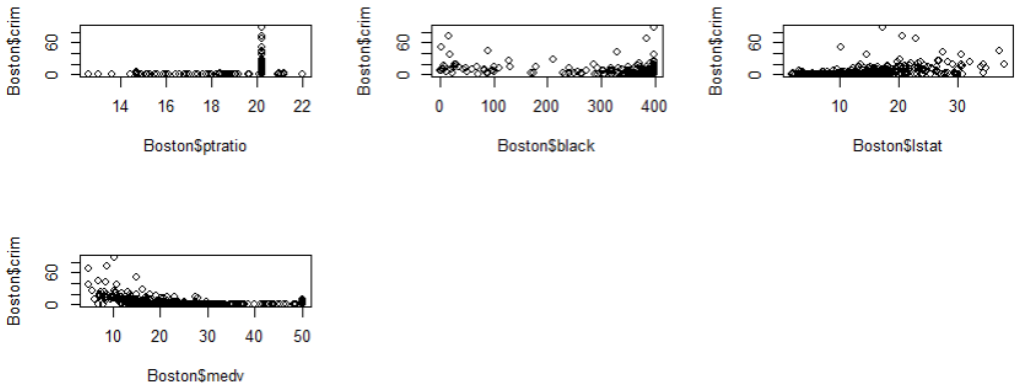
(e) Now remove the 50th through 100th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? Provide the answer in the form of a table:

```
> newSet <- Boston[-c(50:100),]
> sapply(newSet, range)
        crim zn indus chas   nox    rm   age     dis rad tax ptratio  black lstat medv
[1,]  0.00632  0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6   0.32  1.73    5
[2,] 88.97620 95 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0 396.90 37.97   50
> sapply(newSet, mean)
       crim          zn       indus        chas         nox          rm         age
  4.00977073 10.47692308 11.67147253  0.07692308  0.56776615  6.27542857 71.10483516
        dis         rad         tax     ptratio       black       lstat        medv
  3.60324967 10.16043956 420.83516484 18.46483516 352.68008791 13.11892308 22.36571429
> sapply(newSet, sd)
      crim         zn      indus       chas        nox         rm        age        dis
  8.9853408 22.7607025  6.9070525  0.2667627  0.1147901  0.7202964 27.6766735  2.0594589
       rad        tax    ptratio      black      lstat       medv
  8.9604230 172.3011056  2.2547050 95.4248925  7.3187121  9.5255725
```

(f) Investigate graphically (via scatterplots) whether any of the predictors are associated with per capita crime rate (crim)? If so, comment on the relationship.

```
> par(mfrow = c(3,3))
> plot(Boston$zn, Boston$crim)
> plot(Boston$indus, Boston$crim)
> plot(Boston$chas, Boston$crim)
> plot(Boston$nox, Boston$crim)
> plot(Boston$rm, Boston$crim)
> plot(Boston$age, Boston$crim)
> plot(Boston$dis, Boston$crim)
> plot(Boston$rad, Boston$crim)
> plot(Boston$tax, Boston$crim)
> plot(Boston$ptratio, Boston$crim)
> plot(Boston$black, Boston$crim)
> plot(Boston$lstat, Boston$crim)
> plot(Boston$medv, Boston$crim)
```

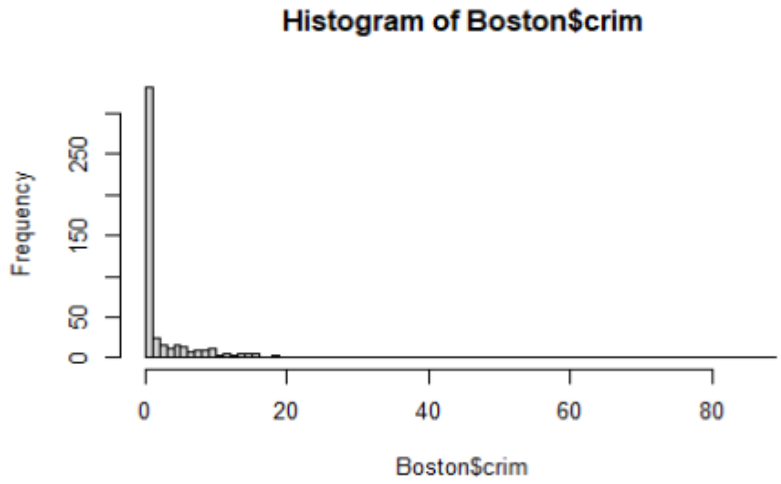




Yes, I see that crime rate increases as age of home increases. I see that the crime rate is high when the weighted mean of distance to five Boston employees' centers is at the lower level of distance. The crime rate increases slowly when the lower status of the population (percent) increases. Also, I see that the crime rate decreases as the median value of owner-occupied homes in $1000s increases.

(g) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? If yes - please comment on those (e.g. which suburbs are those, and what is the highest or lowest the value gets).

```
> hist(Boston$crim, breaks = 75, xlim = c(0,30))
```

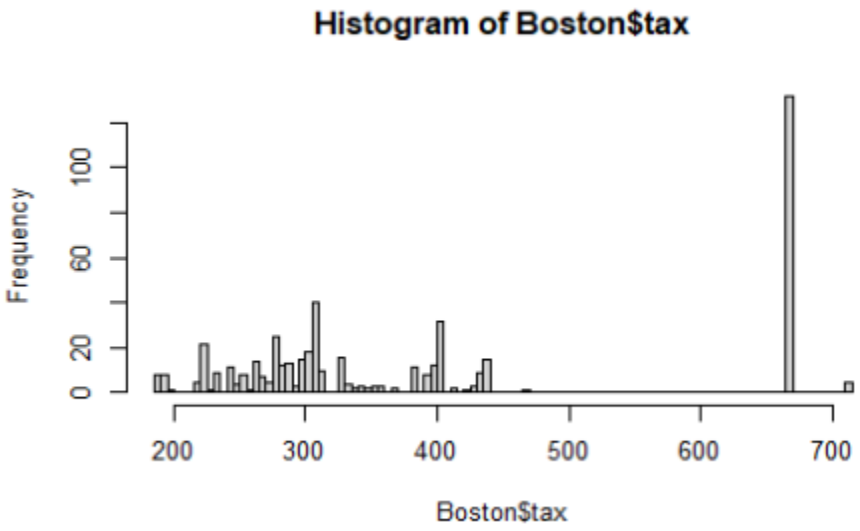


From the histogram of crime rate, we can see that most of the suburbs of Boston have low crime rates. But some of the suburbs of Boston do have high crime rate as crime increases to the right.

```
> range(Boston$crim)
[1]  0.00632 88.97620
```

From calculating the range of crime rate, the lowest crime rate for one of the suburbs of Boston is 0.00632 and the highest crime rate is 88.97620.

```
> hist(Boston$tax, breaks = 75)
```

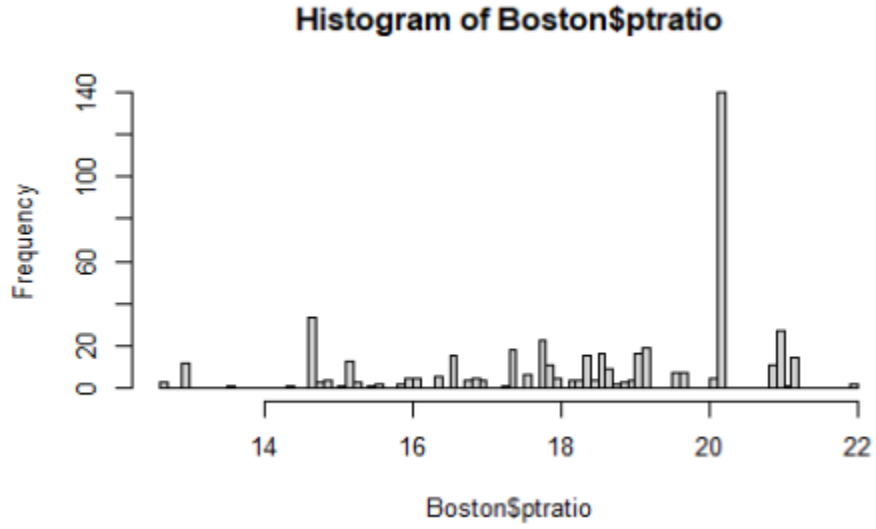


From the histogram of tax rates, there are few many suburbs of Boston between around 200 to less than 500. However, is there a large spike at the far right of the graph.

```
> range(Boston$tax)
[1] 187 711
```

From calculating the range of tax rate, the lowest tax rate for one of the suburbs of Boston is 187 and the highest tax rate is 711.

```
> hist(Boston$ptratio, breaks = 75)
```



From the histogram of pupil-teacher ratio, we can see that there are many activities between the ratio of 14 to 22. There are few spikes and one large spike after the ratio of 20.

```
> range(Boston$ptratio)
```

```
[1] 12.6 22.0
```

From calculating the range of pupil-teacher ratio, the lowest ratio for one of the suburbs of Boston is 12.6 and the highest ratio is 22.0.

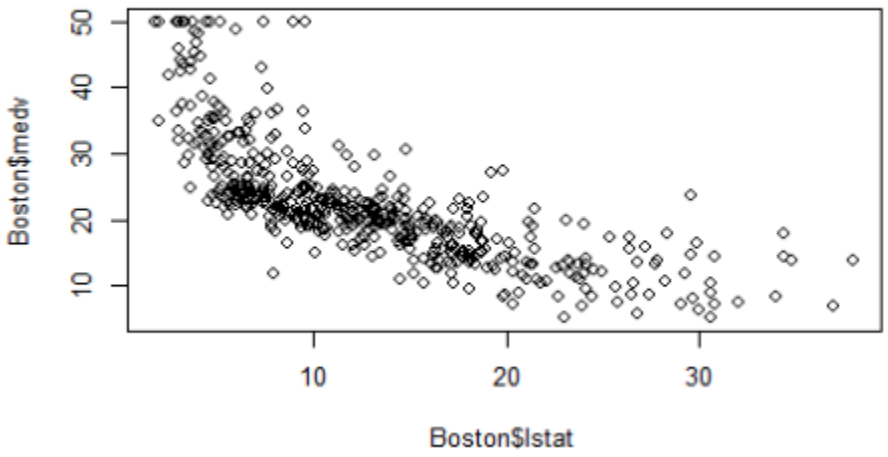
(h) How many of the suburbs in this data set bound the Charles river?

```
> summary(as.factor(Boston$chas))
  0   1
471  35
```

1 = if tract bounds river. So, there are 35 of suburbs in this data set bound the Charles river.

(i) Suppose that we wish to predict median value price of the house (medv) on the basis of the other variables. Do any of the scatter plots (medv vs other predictor) suggest that any of the other variables might be useful in predicting medv? Justify your answer.

```
> plot(Boston$medv~Boston$lstat)
```



Here is the median value of owner-occupied homes in 1000s (medv) vs lower status of the population by percent (lstat) scatterplot. We can see that the medv are high at the low percentage of lstat. However, as lstat increases then the medv decreases.