MATH 4322 Homework 1

Dr. Cathy Poliak

Spring 2023

Instructions

1. Due date: January 31, 2023, 11:59 PM
2. Answer the questions fully for full credit.
3. Scan or Type your answers and submit only one file. (If you submit several files only the recent one uploaded will be graded).
4. Preferably save your file as PDF before uploading.
5. Submit in Canvas under Homework 1.
6. These questions are from *An Introduction to Statistical Learning*, second edition by James, et. al., chapter 2.
7. The information in the gray boxes are R code that you can use to answer the questions.

Problem 1

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Answer
Regression
Most interested in prediction
$n = 52$
$p = 4$

b) An online store is determining whether or not a customer will purchase additional items. This online store collected data from 1500 customers and looked at cost of initial purchase, if there was a special offer, type of item purchased, number of times the customer logged into their account, and if they purchased additional items.

Answer
Classification
Most interested in prediction
$n = 1500$
$p = 5$

Problem 2

This is an exercises about bias, variance and MSE.

Suppose we have n independent Bernoulli trails with true success probability p. Consider two estimators of p: $\hat{p}_1 = \hat{p}$ where $\hat{p}$ is the sample proportion of successes and $\hat{p}_2 = \frac{1}{2}$, a fixed constant.

a) Find the expected value and bias of each estimator.

b) Find the variance of each estimator.

c) Find the MSE of each estimator and compare them by plotting against the true p. Use $n = 4$. Comment on the comparison.

Answer

a) $E(\hat{p}_1) = p$ and $E(\hat{p}_2) = \frac{1}{2}$.

$$Bias(\hat{p}_1) = E(\hat{p}) - p = p - p = 0$$

$$Bias(\hat{p}_2) = E\left(\frac{1}{2}\right) - p = \frac{1}{2} - p$$

b) Determining Variances:

$$Var(\hat{p}_1) = Var\left(\frac{X}{n}\right)$$

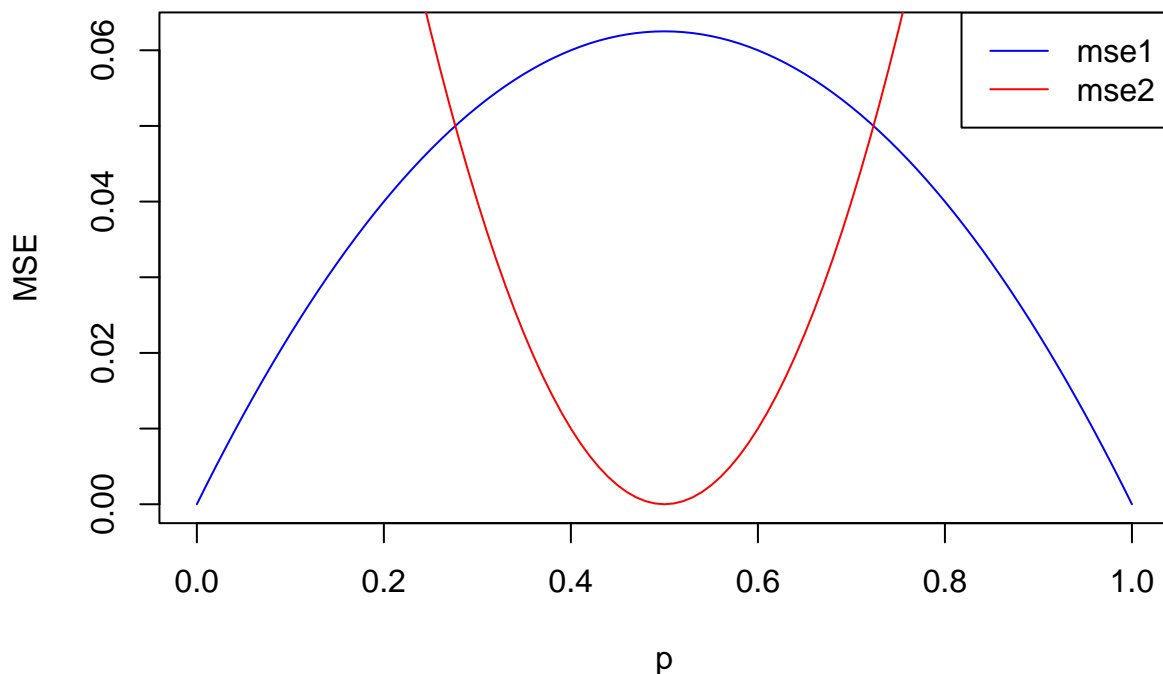$$= \frac{1}{n^2}Var(X)$$

$$= \frac{1}{n^2}n(p)(1-p)$$

$$= \frac{p(1-p)}{n}$$

$$Var\left(\frac{1}{2}\right) = 0$$

c) $MSE = E(\hat{p} - p)^2 = Var(\hat{p}) + [Bias(\hat{p})]^2$

$$MSE(\hat{p}_1) = \frac{p(1-p)}{n}$$

$$MSE(\hat{p}_2) = \left(\frac{1}{2} - p\right)^2$$

```
p = seq(0,1,by = 0.01)
mse1 = 1/4*p*(1 - p)
mse2 = (1/2 - p)^2
plot(p,mse1,type = "l",col = "blue",ylab = "MSE")
lines(p,mse2,col="red")
legend("topright",legend = c("mse1","mse2"), col = c("blue","red"), lty = c(1,1))
```

Notice that the closer p gets to 0.5, the smaller the MSE for $\hat{p} = \frac{1}{2}$.

Problem 3

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Answers

1. Parametric: Make an assumption about the functional form, or shape, of f. Then requires a procedure that uses the training data to fit or train the model.

2. Non-Parametric: do not make explicit assumptions about the functional form of f.

Advantages of parametric approach: It can reduce the problem of estimating f to a smaller set of parameters.
Disadvantage of a parametric approach: The model that we choose may not fit the correct form of f. Whereas the non-parametric approach estimates f that gets a close to the data points as possible thus can get a better estimate of the form of f.

See the *ISLR* textbook pages 22 & 23.

Problem 4

This exercise involves the `Auto` data set in `ISLR` package. Make sure that the missing values have been removed from the data.

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

Answer

Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year
Qualitative: origin, name

(b) What is the range of each quantitative predictor? You can answer this using the `summary()` function.

```
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight
## Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
## 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
## Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
## Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
## 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
## Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##
##   acceleration        year          origin                   name
## Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
## 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
## Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
## Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
## 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
## Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                 (Other)           :365
```

Answer

mpg (9, 46.6), cylinders (3, 8), displacement (68, 455), horsepower (46, 230), weight (1613, 5140), acceleration (8, 24.8), year (70, 82)

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

Answer

```
auto.quant = Auto[,1:7]
colMeans(auto.quant) #Means
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
sqrt(diag(var(auto.quant))) #standard deviation
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

(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?

Answer

```
auto.quant.2 = auto.quant[1:9,]
summary(auto.quant.2) #gives range and mean
```

```
##       mpg          cylinders    displacement     horsepower        weight
## Min.   :14.00   Min.   :8     Min.   :302.0   Min.   :130     Min.   :3433
## 1st Qu.:14.00   1st Qu.:8     1st Qu.:307.0   1st Qu.:150     1st Qu.:3449
## Median :15.00   Median :8     Median :350.0   Median :165     Median :3693
```

4

```
##  Mean   :15.67   Mean   :8    Mean   :373.2   Mean   :177   Mean   :3883
##  3rd Qu.:17.00   3rd Qu.:8    3rd Qu.:440.0   3rd Qu.:215   3rd Qu.:4341
##  Max.   :18.00   Max.   :8    Max.   :455.0   Max.   :225   Max.   :4425
##   acceleration        year
##  Min.   : 8.5   Min.   :70
##  1st Qu.:10.0   1st Qu.:70
##  Median :10.5   Median :70
##  Mean   :10.5   Mean   :70
##  3rd Qu.:11.5   3rd Qu.:70
##  Max.   :12.0   Max.   :70
```

```
sqrt(diag(var(auto.quant.2)))
```

```
##          mpg    cylinders displacement    horsepower       weight acceleration
##     1.658312     0.000000    69.481612     37.446629   458.232474     1.250000
##         year
##     0.000000
```

(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.

Answer

```
auto2 = Auto
auto2$origin = as.factor(auto2$origin)
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
ggpairs(auto.quant)
```
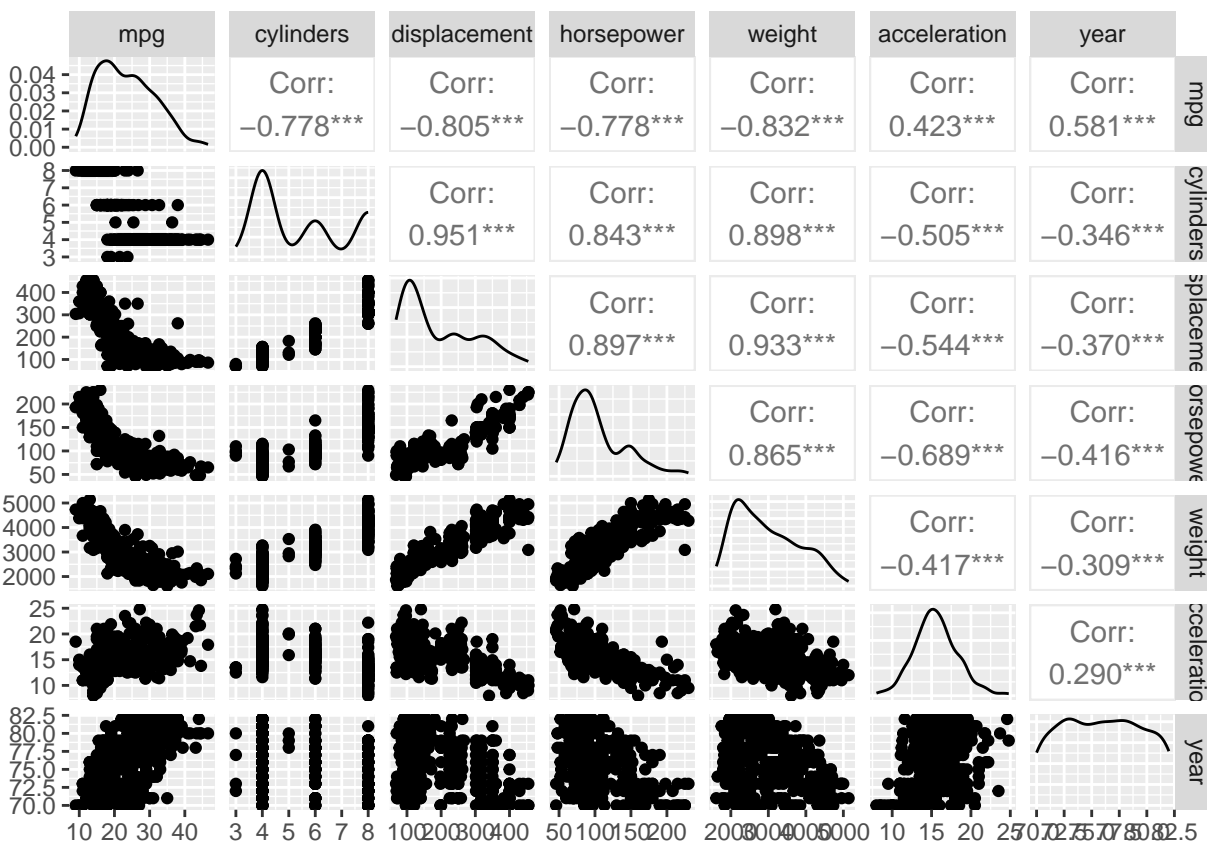
Explanation is subjective.

(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.

Answer: Yes, `mpg` seems to be directly related to `displacement`, `horsepower` and `weight`.

Problem 5

This exercise relates to the `College` data set, which can be found in the file `College.csv` attached to this homework set in Blackboard. 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. You can also import this data set into **RStudio** by using the **Import Dataset → From Text** drop down list in the Environment window.

b) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We will not use this column as a variable but it may be handy to have these names for later. Try the following commands in R:

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

If you are getting an error make sure your data frame is named with a lowercase "c".
Give a brief description of what you see in the data frame.

c) Use the `summary()` function to produce a numerical summary of the variables in the data set. Is there any variables that do not show a numerical summary?

```
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
```

Answer

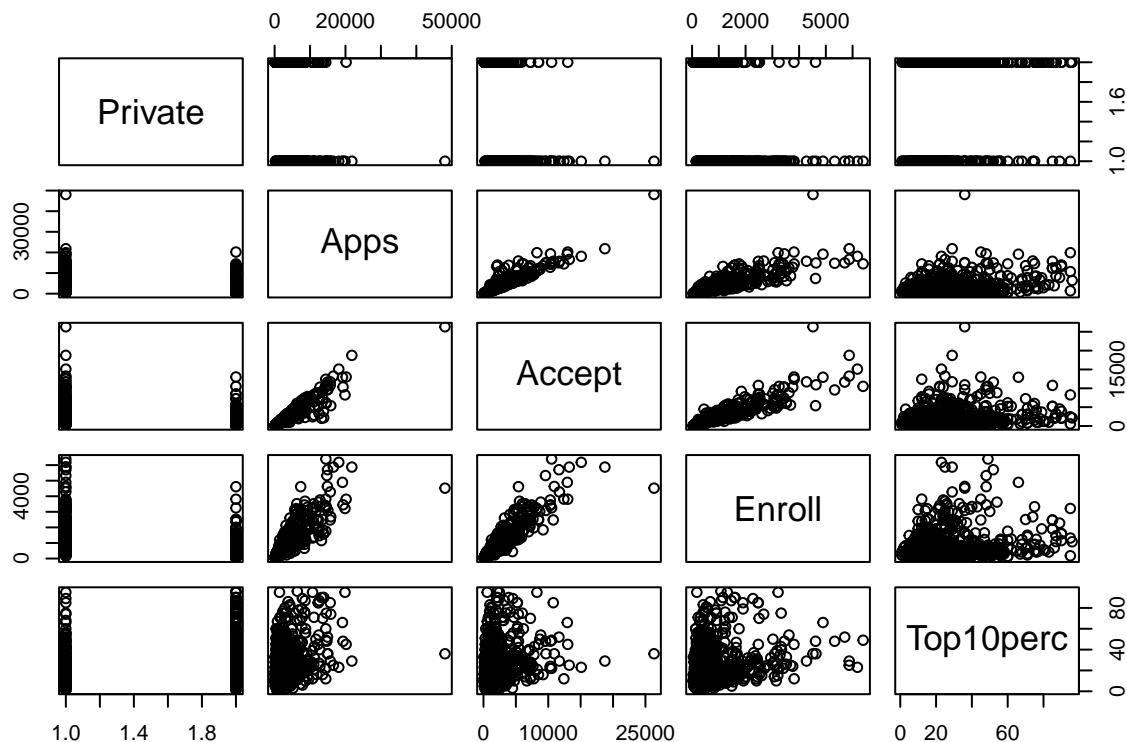The variable `Private` is listed as a character.

Type in the following in `R`:

```
college$Private <- as.factor(college$Private)
```

    d) Use the `pairs()` function to produce a scatterplot matrix of the first five columns or variable of the dataset. Describe any relationships you see in these plots.
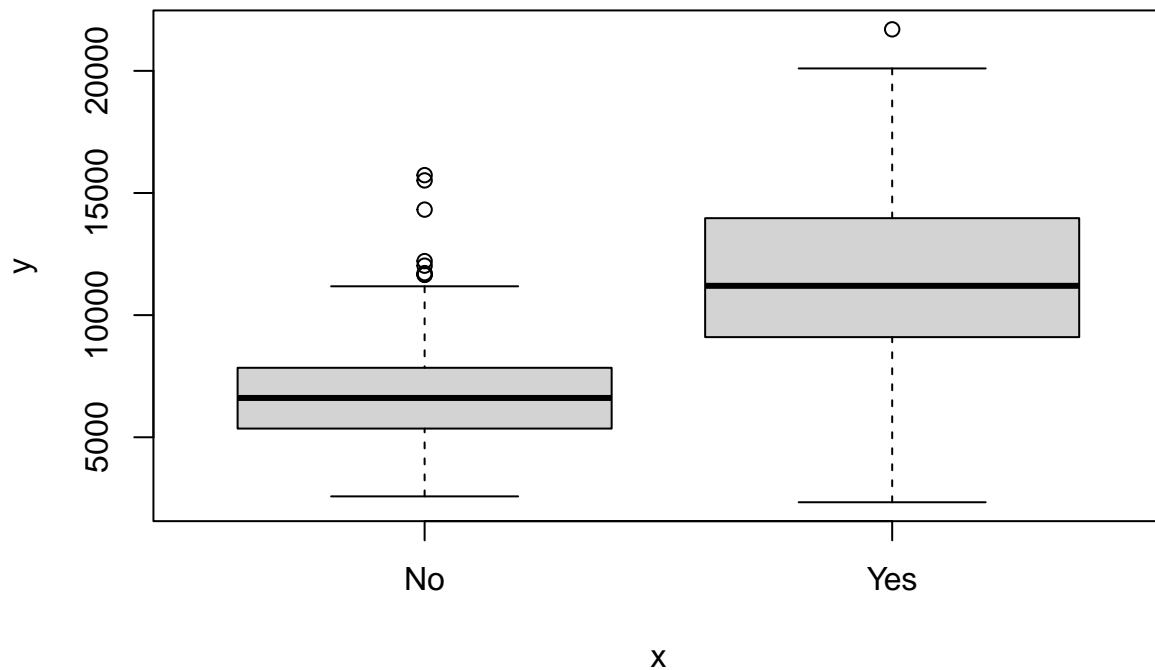
Answer

```
pairs(college[,1:5])
```


There seems to be a linear relationship between number of applications, number accepted, and number enrolled.

    e) Use the `plot()` function to produce a plot of `Outstate` versus `Private`. What type of plot was produced? Give a description of the relationship. *Hint: 'Outstate is in the y-axis.*

Answer

```
plot(college$Private,college$Outstate)
```

This is a box plot. There seems to be a higher out state tution for private universities.

f) 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%. Type in the following in R:

```r
Elite <- rep("No", nrow(college)) #this gives a column of No's for the same number of rows college.
Elite[college$Top10perc > 50] <- "Yes" #changes to Yes if top 10% is greater than 50
Elite <- as.factor(Elite)
college <- data.frame(college,Elite) #adds Elite as a column
```

Use the `summary()` function to see how many elite universities there are.

Answer

```r
summary(Elite)
```

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

There are 78 so called elite schools.

Problem 6

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. You may have to install the ISLR2 library then call for this library.

```
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'
## The following objects are masked from 'package:ISLR':
##
##      Auto, Credit
```

```
data(Boston)
```

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 the rows and columns represent?
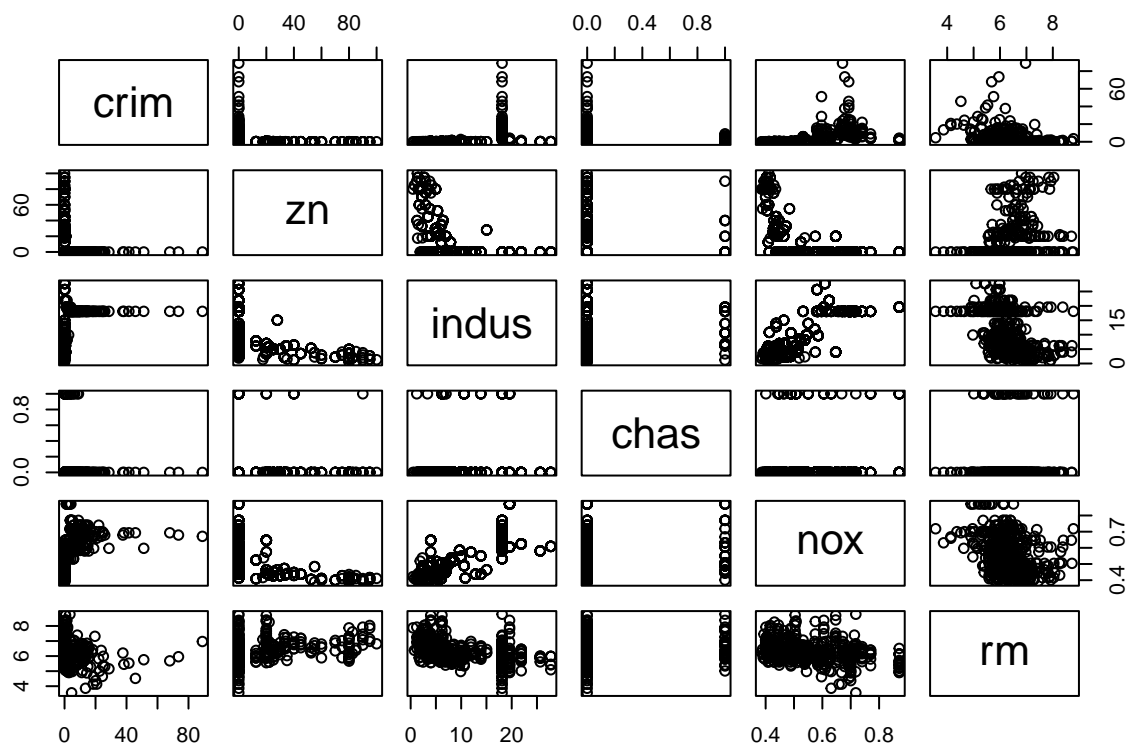
Answer
There are 506 rows, this is the number of observations number of suburbs in Boston and 13 columns, this is the number of variables in the data set.
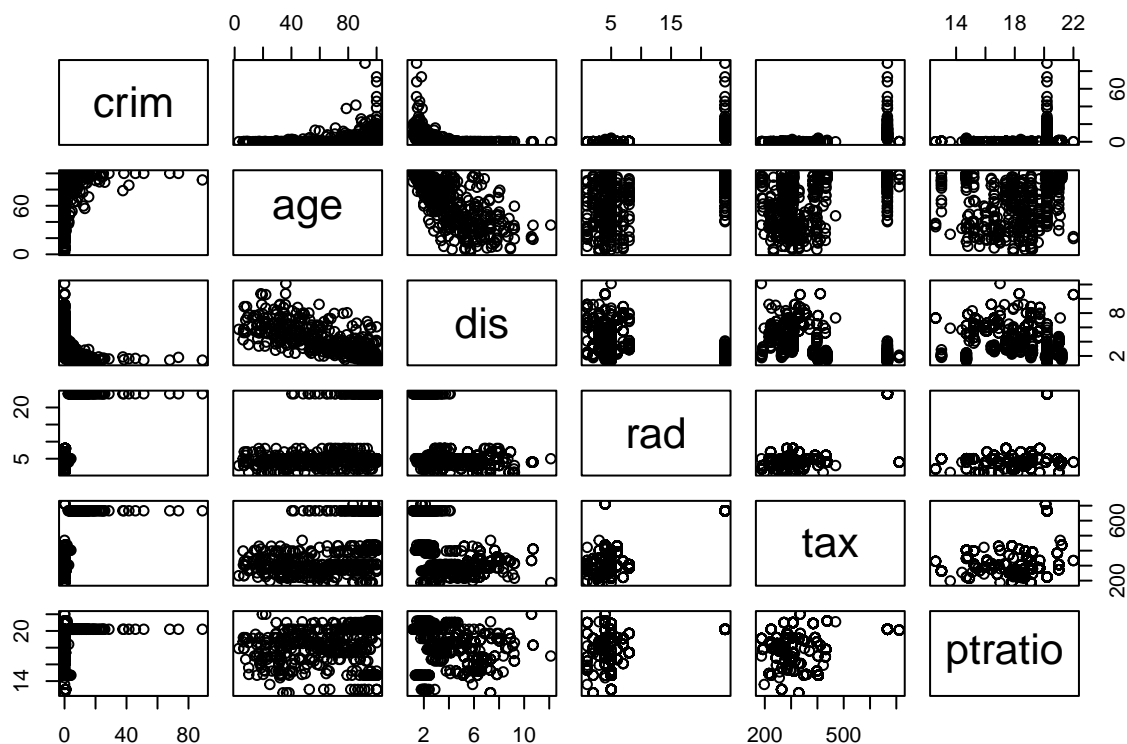
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

Answer

```
pairs(Boston[1:6])
```

```
pairs(Boston[c(1,7:11)])
```

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

Answer

```
cor(Boston[,1],Boston[,2:11])
```

```
##               zn      indus        chas        nox         rm        age        dis
## [1,] -0.2004692 0.4065834 -0.05589158 0.4209717 -0.2192467 0.3527343 -0.3796701
##            rad        tax     ptratio
## [1,] 0.6255051 0.5827643 0.2899456
```

The highest correlation to the Crime rate appears to be `rad` index of accessibility to radial highways and `tax` full-value property-tax rate per \$10,000. As these increase, the crime rate increases.

(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.

Answer

```
par(mfrow = c(2,2))
hist(Boston$crim,main = "",xlab = "Crime")
hist(Boston$tax, main = "",xlab = "Tax")
hist(Boston$ptratio, main = "",xlab = "PTratio")
```
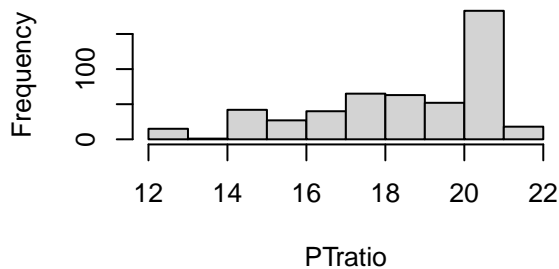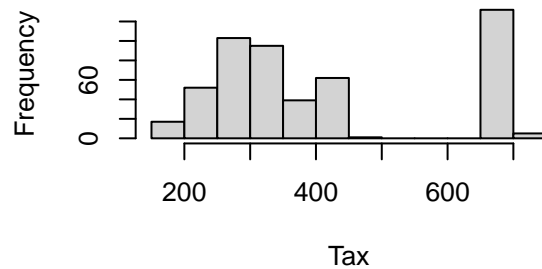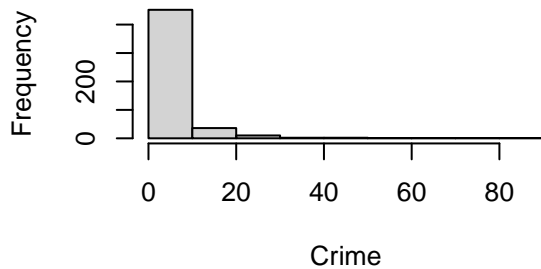
The crime rate is skewed right. If you use the 1.5 IQR rule, there are some towns that are outliers.

Property tax seems to jump between 500 and 700. This shows bimodal.

Parent-teacher ratio is skewed left.

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

Answer

```
sum(Boston$chas)
```

[1] 35

35 Suburbs bound the Charles River.

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

Answer

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

[1] 19.05

The median is 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.

```
which.min(Boston$medv) #This is the observation that has the lowest median value.
```

[1] 399

```
Boston[which.min(Boston$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
```

(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.

Answer

```
length(which(Boston$rm > 7)) #Number of suburbs average more than 7 rooms
```

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

```
length(which(Boston$rm > 8)) #Number of suburbs average more than 8 rooms
```

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

```
Boston[which(Boston$rm > 8),]
```

```
##         crim zn indus chas    nox    rm  age    dis rad tax ptratio lstat medv
## 98   0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0  4.21 38.7
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7  3.32 50.0
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7  2.88 50.0
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4  4.14 44.8
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4  4.63 50.0
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4  3.13 37.6
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4  2.47 41.7
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4  3.95 48.3
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1  3.54 42.8
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0  5.12 50.0
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0  5.91 48.8
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0  7.44 50.0
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2  5.29 21.9
```