Name: Jingyi Yuan
UNI: jy2736
Class: STAT W4240

Homework 01

Problem 1

(a).

In the R script:

```
setwd("/Users/jingyiyuan/Desktop/Data Mining/R")
college = read.csv("College.csv",header = T)
```

In the command window:

```
> setwd("/Users/jingyiyuan/Desktop/Data Mining/R")
> college = read.csv("College.csv",header = T)
>
```

The data is:

```
Global Environment
Data
college        777 obs. of 19 variables
```

(b).

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

R Data Editor

Copy Paste Quit

	row.names
1	Abilene Christian University
2	Adelphi University
3	Adrian College
4	Agnes Scott College
5	Alaska Pacific University
6	Albertson College
7	Albertus Magnus College
8	Albion College
9	Albright College
10	Alderson-Broaddus College
11	Alfred University
12	Allegheny College
13	Allentown Coll. of St. Francis de Sales
14	Alma College
15	Alverno College
16	American International College
17	Amherst College
18	Anderson University
19	Andrews University
20	Angelo State University
21	Antioch University
22	Appalachian State University
23	Aquinas College
24	Arizona State University Main campus
25	Arkansas College (Lyon College)

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

R Data Editor

Copy Paste Quit

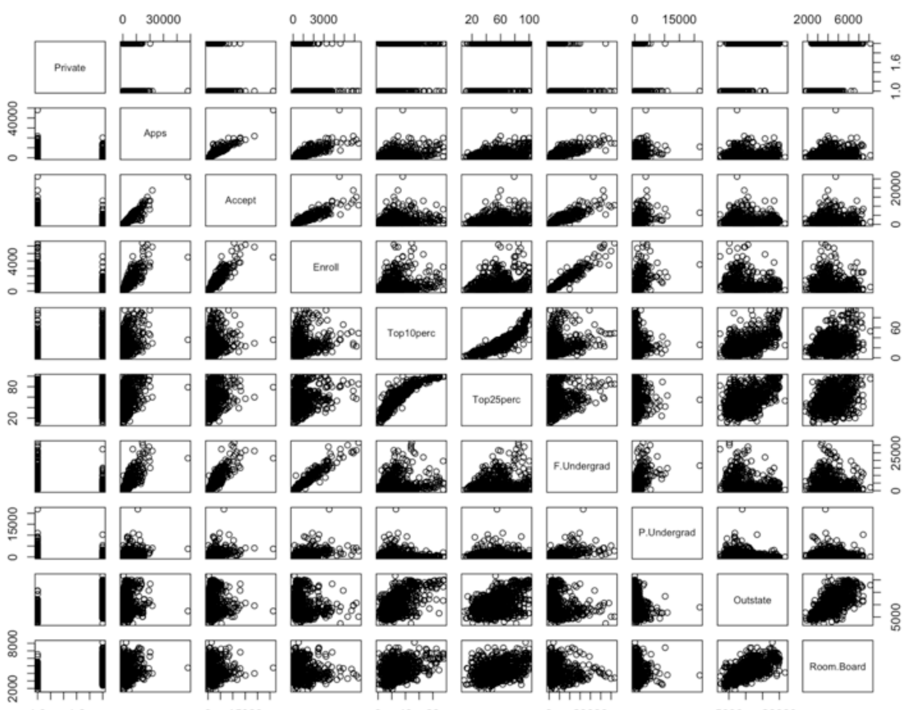
	row.names	Private	Apps	Accept	Enroll
1	Abilene Christian University	Yes	1660	1232	721
2	Adelphi University	Yes	2186	1924	512
3	Adrian College	Yes	1428	1097	336
4	Agnes Scott College	Yes	417	349	137
5	Alaska Pacific University	Yes	193	146	55
6	Albertson College	Yes	587	479	158
7	Albertus Magnus College	Yes	353	340	103
8	Albion College	Yes	1899	1720	489
9	Albright College	Yes	1038	839	227
10	Alderson-Broaddus College	Yes	582	498	172
11	Alfred University	Yes	1732	1425	472
12	Allegheny College	Yes	2652	1900	484
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290
14	Alma College	Yes	1267	1080	385
15	Alverno College	Yes	494	313	157
16	American International College	Yes	1420	1093	220
17	Amherst College	Yes	4302	992	418
18	Anderson University	Yes	1216	908	423
19	Andrews University	Yes	1130	704	322
20	Angelo State University	No	3540	2001	1016
21	Antioch University	Yes	713	661	252
22	Appalachian State University	No	7313	4664	1910
23	Aquinas College	Yes	619	516	219
24	Arizona State University Main campus	No	12809	10308	3761
25	Arkansas College (Lyon College)	Yes	708	334	166

(c).

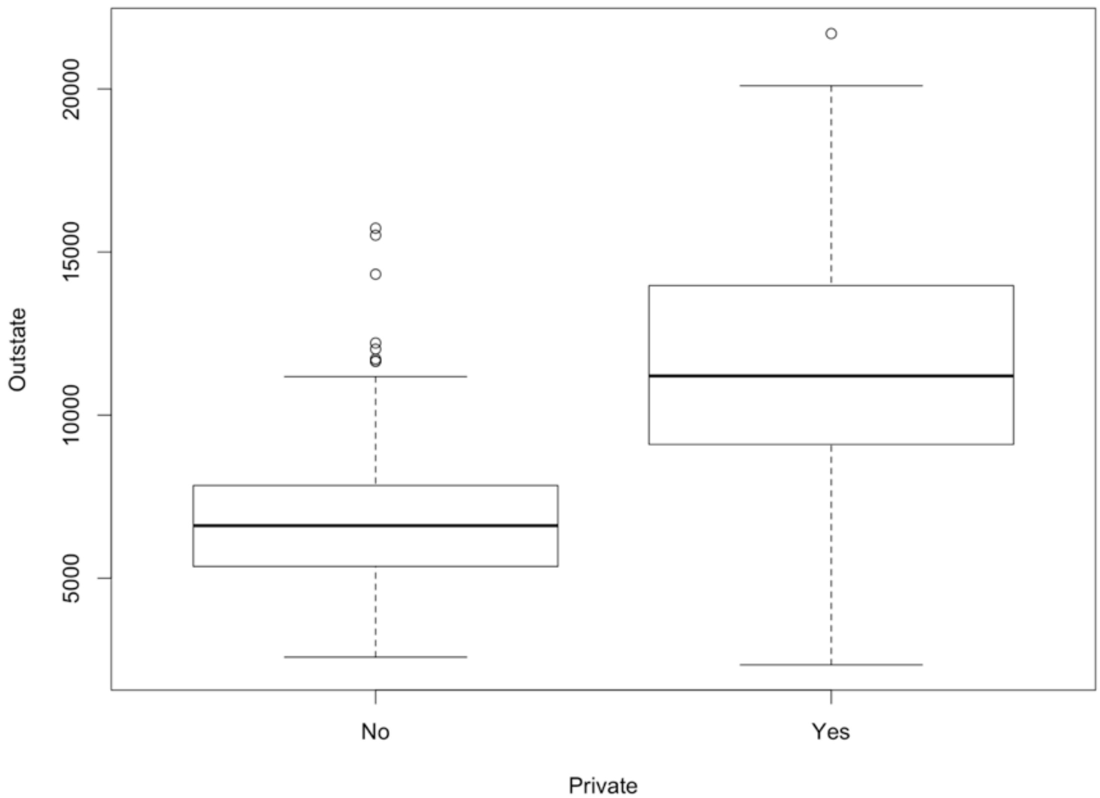
i. summary(college)

```
 Private        Apps           Accept          Enroll
 No :212   Min.   :   81   Min.   :   72   Min.   :  35
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
           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.

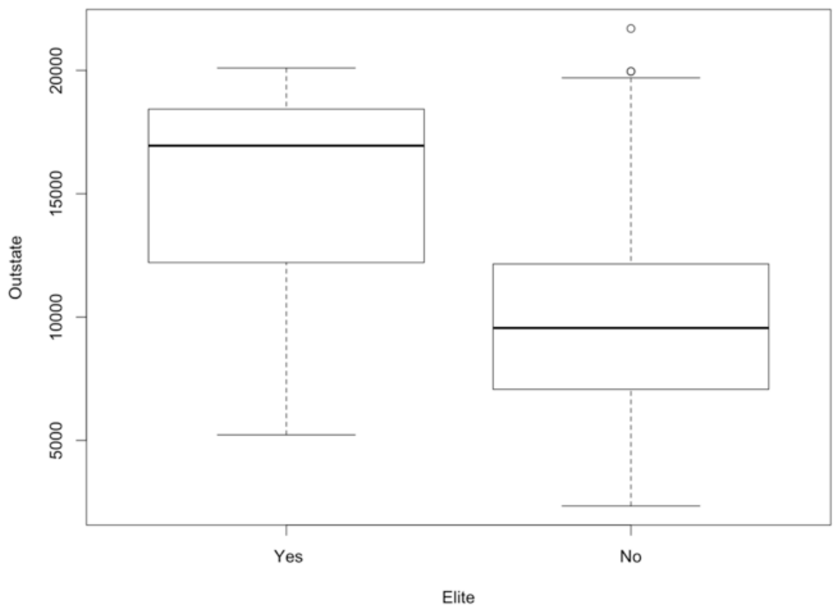


iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private



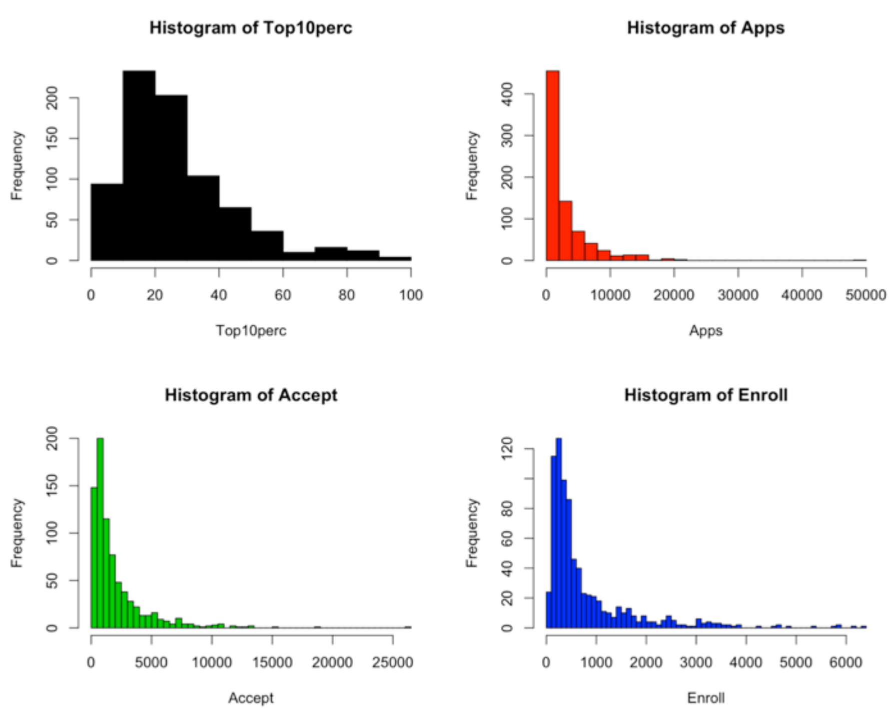
iv. Create a new qualitative variable, called Elite.

```
> Elite = rep ("No",nrow(college ))
> Elite [college$Top10perc >50]=" Yes"
> Elite = as.factor (Elite)
> college = data.frame(college ,Elite)
> summary(Elite)
 Yes   No
  78  699
> plot(Elite, Outstate, xlab = "Elite", ylab = "Outstate")
```



v.

```
par(mfrow=c(2,2))
hist(Top10perc,col = 1, breaks = 10)
hist(Apps,col = 2, breaks = 20)
hist(Accept,col = 3, breaks = 40)
hist(Enroll,col = 4, breaks = 80)
```



vi.

```
#continue explore more
admission_rate = Accept/Apps
range(admission_rate)
```

```
> admission_rate = Accept/Apps
> range(admission_rate)
[1] 0.1544863 1.0000000
```

Explore the data by using the command ”admission_rate = Accept/Apps” to get the admission rate of each school. Using range() function to know the admission range and when applying, students can take the rate as a reference.

A brief summary: summary() function produces a numerical summary of each column thus we can clearly get the min, max, median etc. values of a particular data set. The pairs() function creates a scatterplot matrix for every pair of the first 10 columns and the relationship can be analyzed through each picture. The hist() function is used to plot histograms, which can tell us the frequency of numbers, and the distribution of the numbers in each column is shown clearly.

Problem 2

(a).

Mpg, cylinders, displacement, horsepower, weight and acceleration are quantitative. Year, name and origin are qualitative.

(b).

```
> setwd("/Users/jingyiyuan/Desktop/Data Mining/R")
> Auto = read.csv("Auto.csv", na.string = "?", header = T)
> Auto <- na.omit(Auto)
> attach(Auto)
> range(mpg)
[1]  9.0 46.6
> range(cylinders)
[1] 3 8
> range(displacement)
[1]  68 455
> range(horsepower)
[1]  46 230
> range(weight)
[1] 1613 5140
> range(acceleration)
[1]  8.0 24.8
```

(c).

```
> mean(cylinders)
[1] 5.471939
> sd(cylinders)
[1] 1.705783
> mean(displacement)
[1] 194.412
> sd(displacement)
[1] 104.644
> mean(horsepower)
[1] 104.4694
> sd(horsepower)
[1] 38.49116
> mean(weight)
[1] 2977.584
> sd(weight)
[1] 849.4026
> mean(acceleration)
[1] 15.54133
> sd(acceleration)
[1] 2.758864
```

(d).

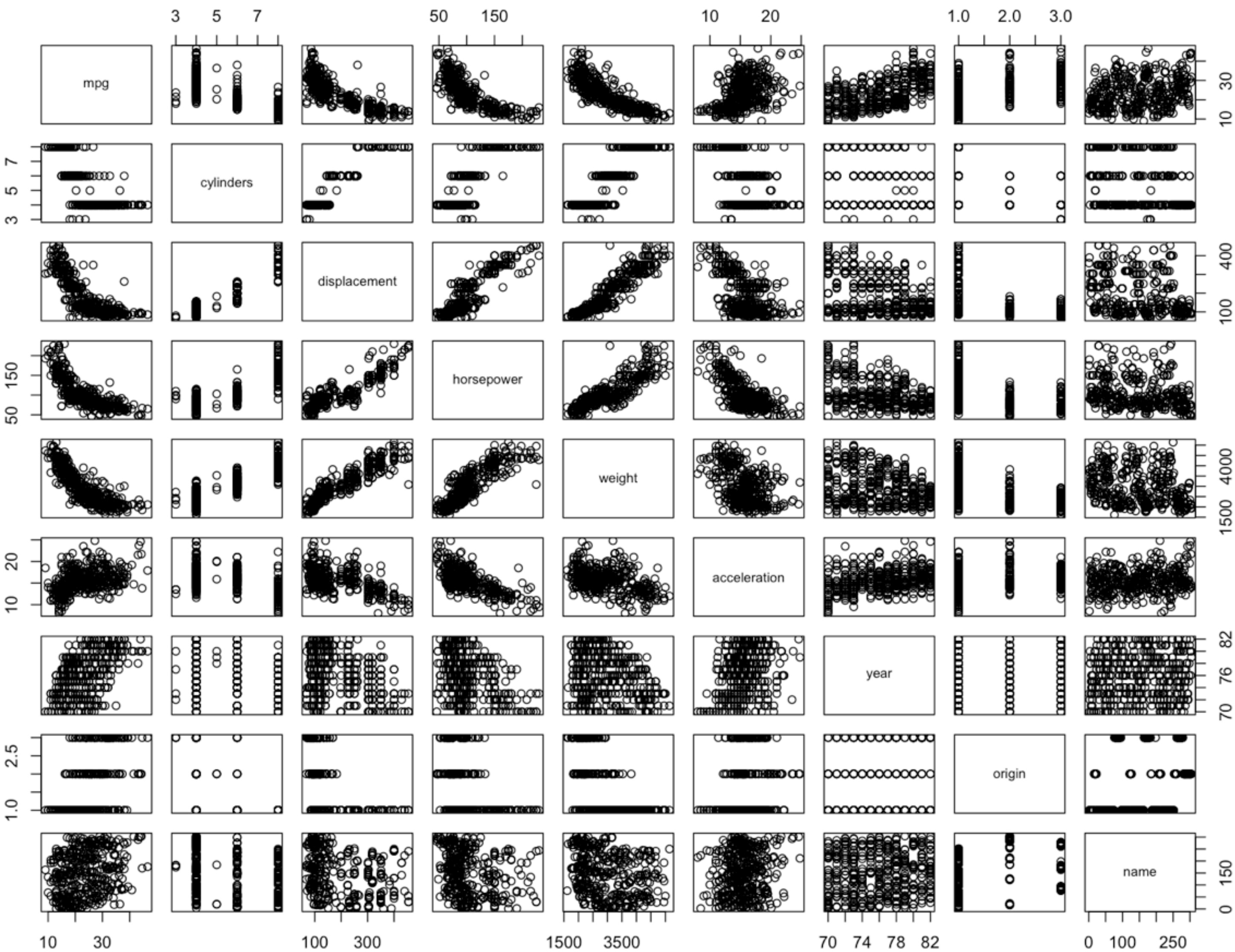
```
> Auto_01 = Auto[10:85,]
> Auto_01 = na.omit(Auto_01)
> range(mpg)
[1]  9.0 46.6
> mean(mpg)
[1] 23.44592
> sd(mpg)
[1] 7.805007
> range(cylinders)
[1] 3 8
> mean(cylinders)
[1] 5.471939
> sd(cylinders)
[1] 1.705783
> range(displacement)
[1]  68 455
> mean(displacement)
[1] 194.412
> sd(displacement)
[1] 104.644
> range(horsepower)
[1]  46 230
> mean(horsepower)
[1] 104.4694
> sd(horsepower)
[1] 38.49116
```

```
> range(weight)
[1] 1613 5140
> mean(weight)
[1] 2977.584
> sd(weight)
[1] 849.4026
> range(acceleration)
[1]  8.0 24.8
> mean(acceleration)
[1] 15.54133
> sd(acceleration)
[1] 2.758864
```

(e).

By using the pairs() function, we can see the relationship between the 9 predictors. As we can see, there is negative relationship between mpg and weight, which means the heavier, the less mpg. Another negative relationship is between mpg and cylinders. The bigger the cylinders are, the less mpg is.

There is positive relationship between weight, horsepower and displacement (which means a negative relationship between mpg and horsepower, mpg and displacement). Another positive relationship that can be spotted is between year and mpg. Maybe as time goes by, the development of technology, the mpg becomes higher. Relationships between other predictors are not very obvious.



(f).

Not all variables are useful. As I analyzed above, the relationships between mpg and weight, horsepower, cylinders and displacement are all negative, and the relationship between mpg and year is positive. So only those variables that have a relationship with mpg count. Those do not have an obvious relationship have little impact on mpg thus would not be included while predicting mpg.

Problem 3

(a).

There are 506 rows and 14 columns in the datasets.

According to the documents in "Help", this data frame contains the following columns:

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

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

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per \$10,000.

ptratio: pupil-teacher ratio by town.

black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat: lower status of the population (percent).

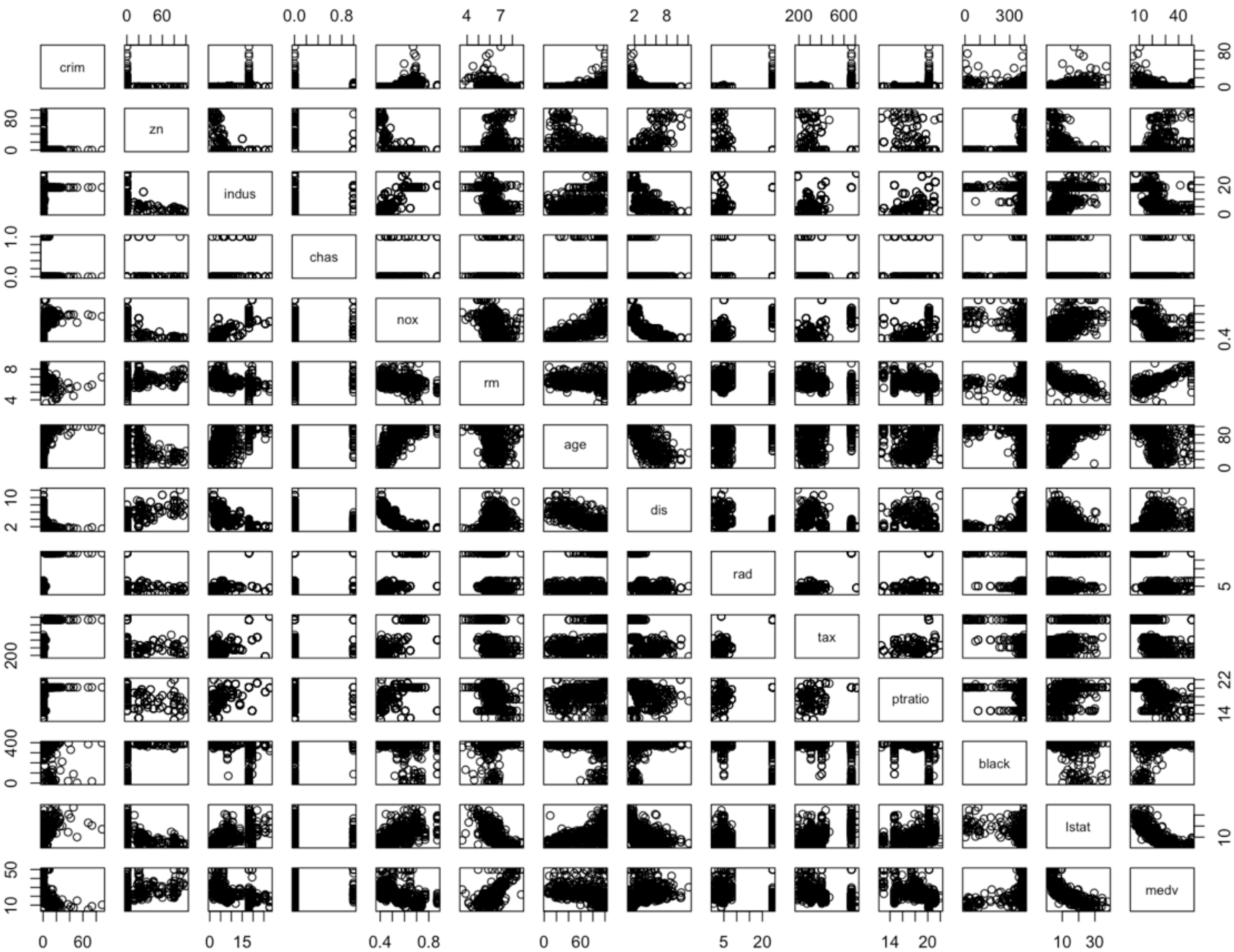
medv: median value of owner-occupied homes in \$1000s.

The rows mean different datasets, which is, the data of some suburbs.

(b).

There is negative relationship between dis and nox, which means the larger the weighted mean of distances to five Boston employment centres is, the lower nitrogen oxides concentration is. Another negative relationship is between Istat and medv, which means the lower status of the population is, the smaller median value of owner-occupied homes in \$1000s will be.

Also, there is positive relationship between rm and medv. The average number of rooms per dwelling increases as median value of owner-occupied homes in \$1000s becomes larger.

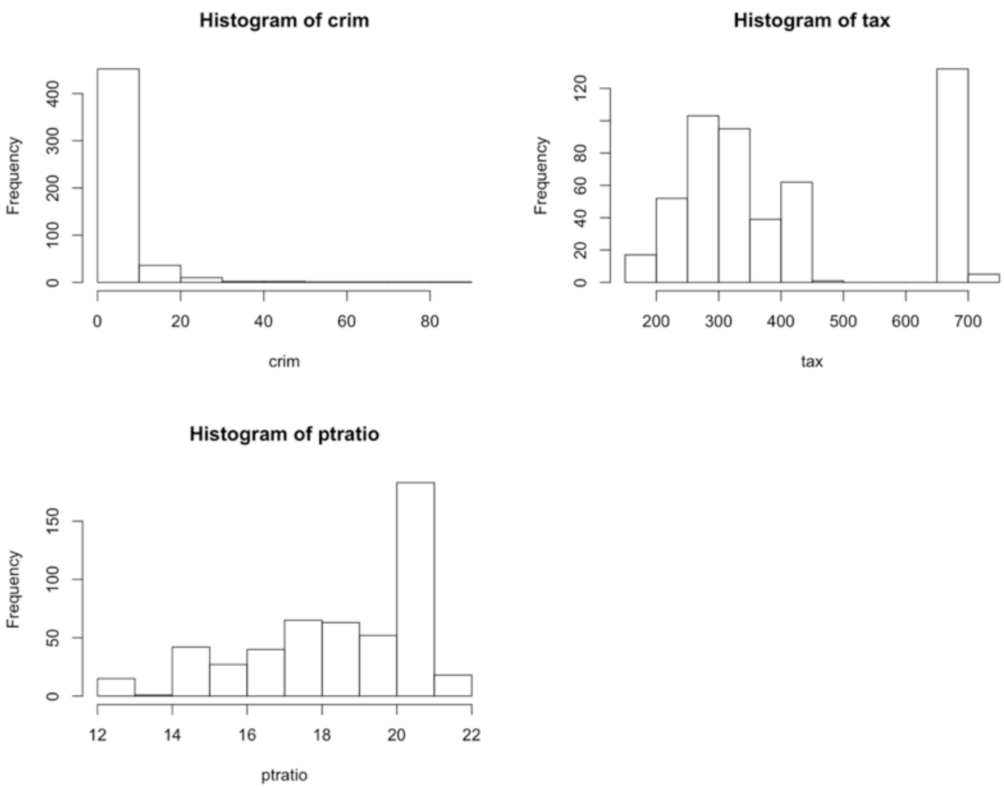


(c).

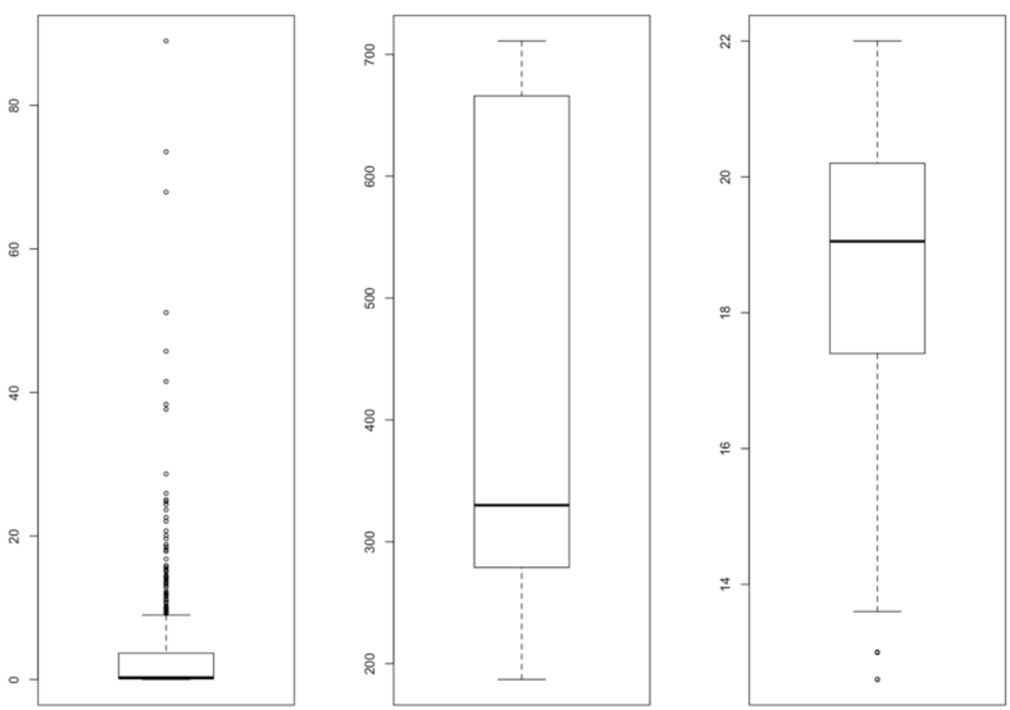
The per capita crime rate by town has a positive relationship with age (proportion of owner-occupied units built prior to 1940), and a negative relationship with dis (weighted mean of distances to five Boston employment centres) and medv (median value of owner-occupied homes in \$1000s).

(d).

The range of crime rates is 0.00632 to 88.97620, which means some suburb has a particularly high crime rate 88.97620. The tax range is 187 to 711, and the Pupil-teacher ratios is 12.6 to 22.0, which are all wide spread without distinctive high or low rates.



Histogram of crim
Frequency
0
100
200
300
400
0
20
40
60
80
crim
Histogram of tax
Frequency
0
20
40
60
80
120
200
300
400
500
600
700
tax
Histogram of ptratio
Frequency
0
50
100
150
12
14
16
18
20
22
ptratio



0
20
40
60
80
200
300
400
500
600
700
14
16
18
20
22

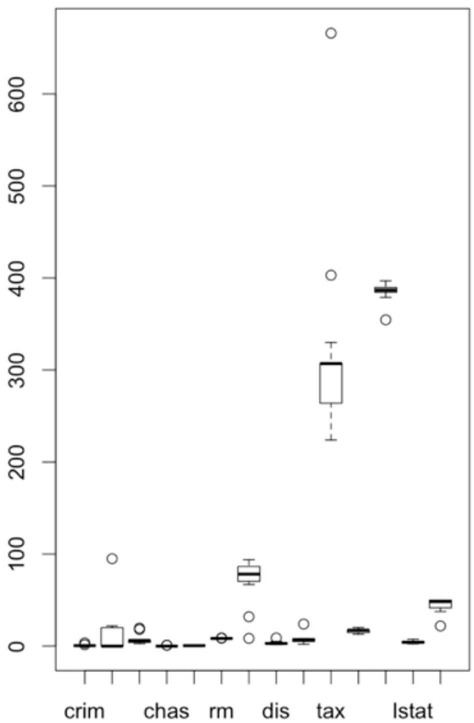
(e).
35
(f).
19.5
(g).
The 399th and 406th suburbs of Boston have the lowest median value of owner-occupied homes. The values of the other predictors for those two suburbs are shown in the picture below.

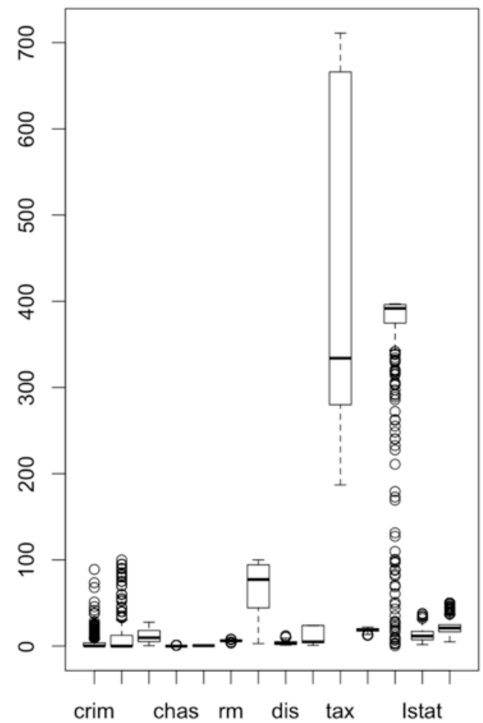
```
       crim zn indus chas   nox    rm age    dis rad tax ptratio
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2
     black lstat medv
399 396.90 30.59    5
406 384.97 22.98    5
```

The ranks of those values compared to the overall ranges for those predictors are shown below. Except for “black”, other predictors are shown to be in a similar rank, that is, only “black” have a distinctive difference in rank. For example, in 506 rows, “crim” of the two row are 500 and 504, whose rank is pretty much close to each other considering the total number of the rows.

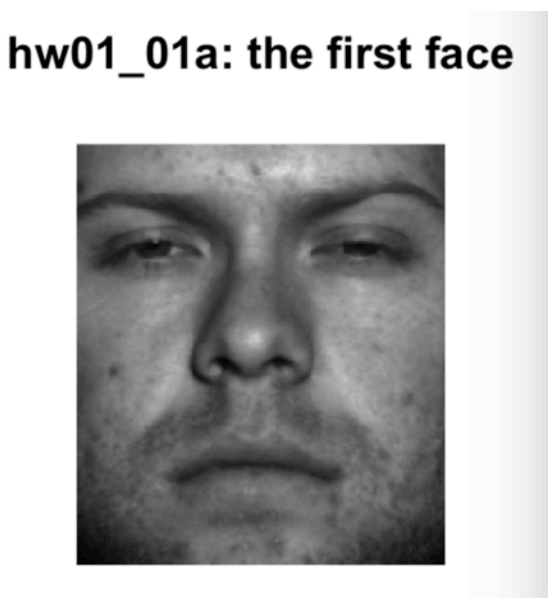
```
     [,1]  [,2]  [,3] [,4]  [,5] [,6] [,7] [,8]  [,9] [,10] [,11]
[1,]  500 186.5 383.5  236 427.5   39  485   29 440.5 435.5 380.5
[2,]  504 186.5 383.5  236 427.5   69  485   21 440.5 435.5 380.5
     [,12] [,13] [,14]
[1,]   446   495   1.5
[2,]   177   455   1.5
```

(h).
There are 64 suburbs average more than seven rooms per dwelling, and 13 suburbs more than 8. According to the boxplot of the dwellings have more than 8 rooms and those have less than 8 rooms, one predictor that has a particular difference is “tax”. Those who have more rooms are with less tax. Another predictor that has an ambiguous difference is “black”.





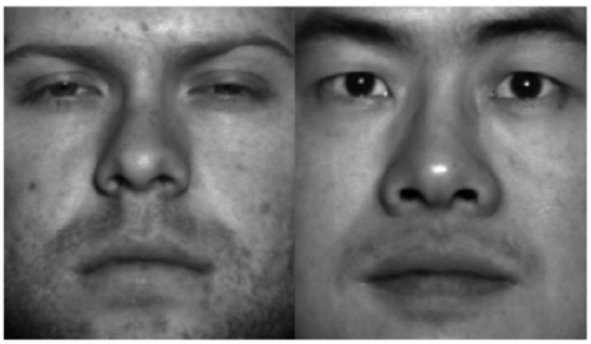
Problem 4
(a).
plot(face_01)



```
> #----- START YOUR CODE BLOCK HERE -----#
> class(face_01)
[1] "pixmapGrey"
attr(,"package")
[1] "pixmap"
> face_01@size
[1] 192 168
> row = face_01@size[1]
> column = face_01@size[2]
> #----- END YOUR CODE BLOCK HERE -----#
```

(b).

plot(faces)



The maximum value that a pixel can take for this type of file is 1, which is white(255).

The minimum is 0, which is black(0).

```
> #----- START YOUR CODE BLOCK HERE -----#
> max(faces_matrix)
[1] 1
> min(faces_matrix)
[1] 0.007843137
> #----- END YOUR CODE BLOCK HERE -----#
```

(c).

The “dir_list_1” contains 38 subjects from "yaleB01" to "yaleB39" with no 14. The “dir_list_2” contains every file in "CroppedYale", including files in folders like "yaleB01_P00_Ambient.pgm" and files in "CroppedYale" like "DEADJOE".

```
> #----- START YOUR CODE BLOCK HERE -----#
> length(dir_list_1)
[1] 38
> length(dir_list_2)
[1] 2547
> #----- END YOUR CODE BLOCK HERE -----#
```

(d).

```
> #----- START YOUR CODE BLOCK HERE -----#
> faces_rows <- array(NA,dim=c(row,column,3,3))
> faces_columns <- array(NA,dim=c(row,column*3,3))
> for (i in 1:3){
+ for (j in 1:3){
+ filename = sprintf("CroppedYale/%s/%s_%s.pgm",dir_list_1[pic_list[i]] ,
dir_list_1[pic_list[i]] , view_list[j])
+ face_all = read.pnm(file = filename)
+ faces_rows[,,i,j] <- matrix(getChannels(face_all),nrow = 192, ncol = 168
)
+ }
+ faces_columns[,,i] = cbind(faces_rows[,,i,1], faces_rows[,,i,2], faces_r
ows[,,i,3])
+ }
Warning messages:
1: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
2: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
3: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
4: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
5: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
6: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
7: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
8: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
9: In rep(cellres, length = 2) : 'x' is NULL so the result will be NULL
> faces_matrix = rbind(faces_columns[,,1], faces_columns[,,2], faces_colum
ns[,,3])
> #----- END YOUR CODE BLOCK HERE -----#
```

hw01_01d: 3x3 grid of faces