Statistical Learning and Data mining

Homework 1

M052040003 鍾冠毅

8.a. See the appendix.

8.b.

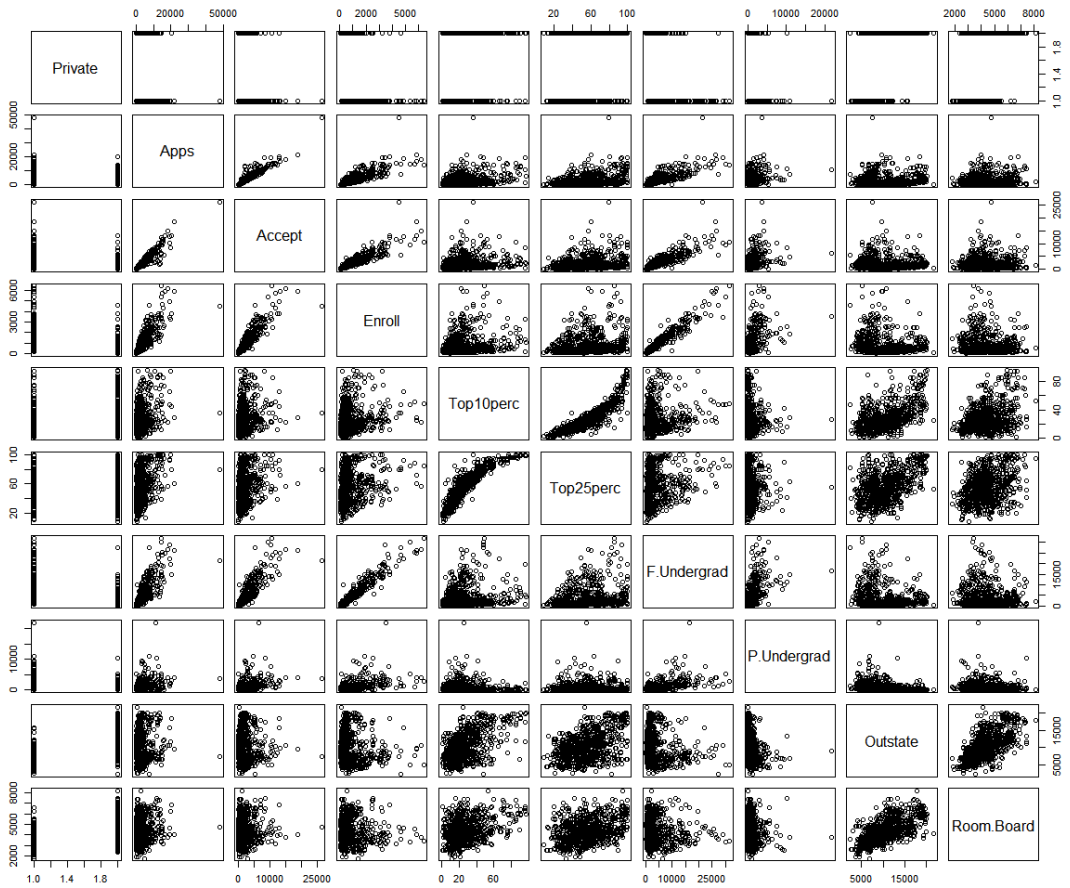
	row.names	X	Private	Apps
1	Abilene Christian University	Abilene Christian University	Yes	1660
2	Adelphi University	Adelphi University	Yes	2186
3	Adrian College	Adrian College	Yes	1428
4	Agnes Scott College	Agnes Scott College	Yes	417
5	Alaska Pacific University	Alaska Pacific University	Yes	193
6	Albertson College	Albertson College	Yes	587
7	Albertus Magnus College	Albertus Magnus College	Yes	353
8	Albion College	Albion College	Yes	1899
9	Albright College	Albright College	Yes	1038
10	Alderson-Broaddus College	Alderson-Broaddus College	Yes	582
11	Alfred University	Alfred University	Yes	1732
12	Allegheny College	Allegheny College	Yes	2652
13	Allentown Coll. of St. Francis de Sales	Allentown Coll. of St. Francis de Sales	Yes	1179
14	Alma College	Alma College	Yes	1267
15	Alverno College	Alverno College	Yes	494
16	American International College	American International College	Yes	1420
17	Amherst College	Amherst College	Yes	4302
18	Anderson University	Anderson University	Yes	1216
19	Andrews University	Andrews University	Yes	1130

	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc
1	Abilene Christian University	Yes	1660	1232	721	23	52
2	Adelphi University	Yes	2186	1924	512	16	29
3	Adrian College	Yes	1428	1097	336	22	50
4	Agnes Scott College	Yes	417	349	137	60	89
5	Alaska Pacific University	Yes	193	146	55	16	44
6	Albertson College	Yes	587	479	158	38	62
7	Albertus Magnus College	Yes	353	340	103	17	45
8	Albion College	Yes	1899	1720	489	37	68
9	Albright College	Yes	1038	839	227	30	63
10	Alderson-Broaddus College	Yes	582	498	172	21	44
11	Alfred University	Yes	1732	1425	472	37	75
12	Allegheny College	Yes	2652	1900	484	44	77
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64
14	Alma College	Yes	1267	1080	385	44	73
15	Alverno College	Yes	494	313	157	23	46
16	American International College	Yes	1420	1093	220	9	22
17	Amherst College	Yes	4302	992	418	83	96
18	Anderson University	Yes	1216	908	423	19	40
19	Andrews University	Yes	1130	704	322	14	23

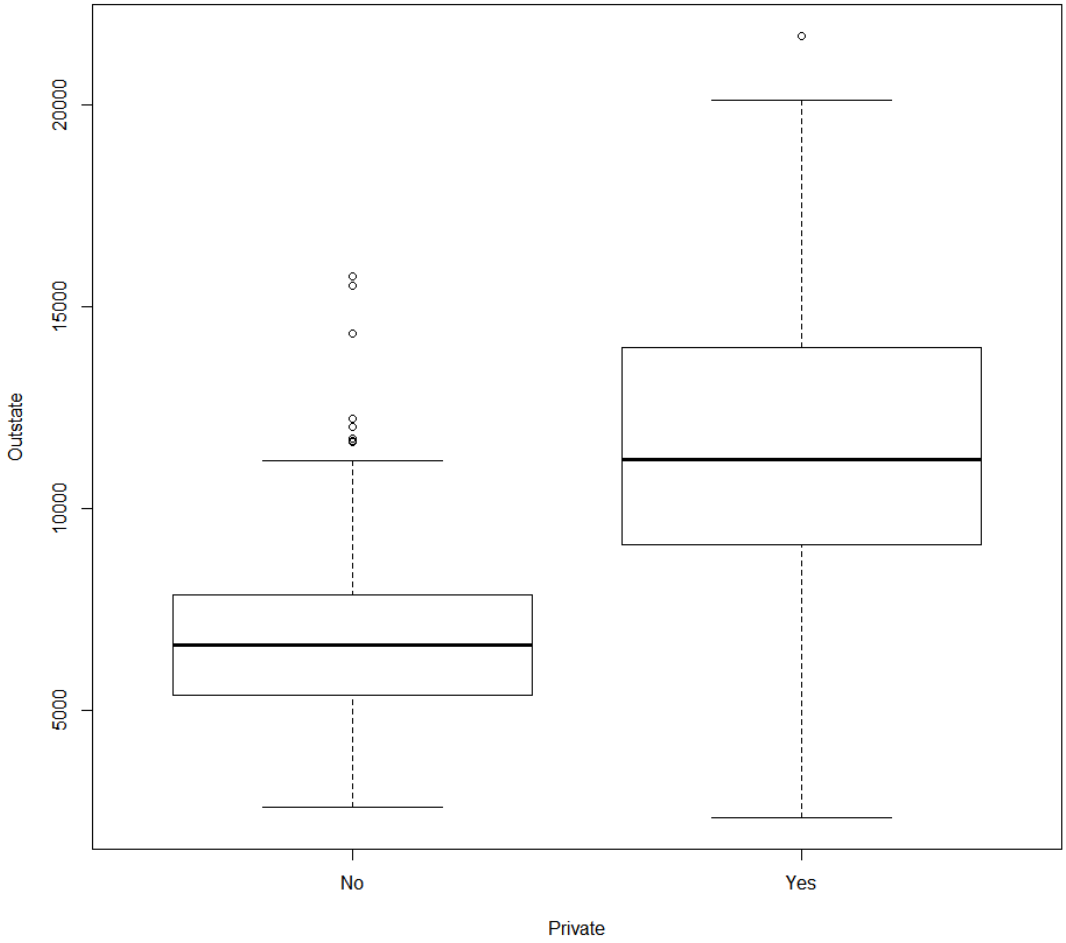
8.c. (i)

```
> summary(college[, 2:length(college[1,])])
      Apps           Accept          Enroll       Top10perc       Top25perc
 Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
 Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
 Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
 3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
 Max.   :48094   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
```

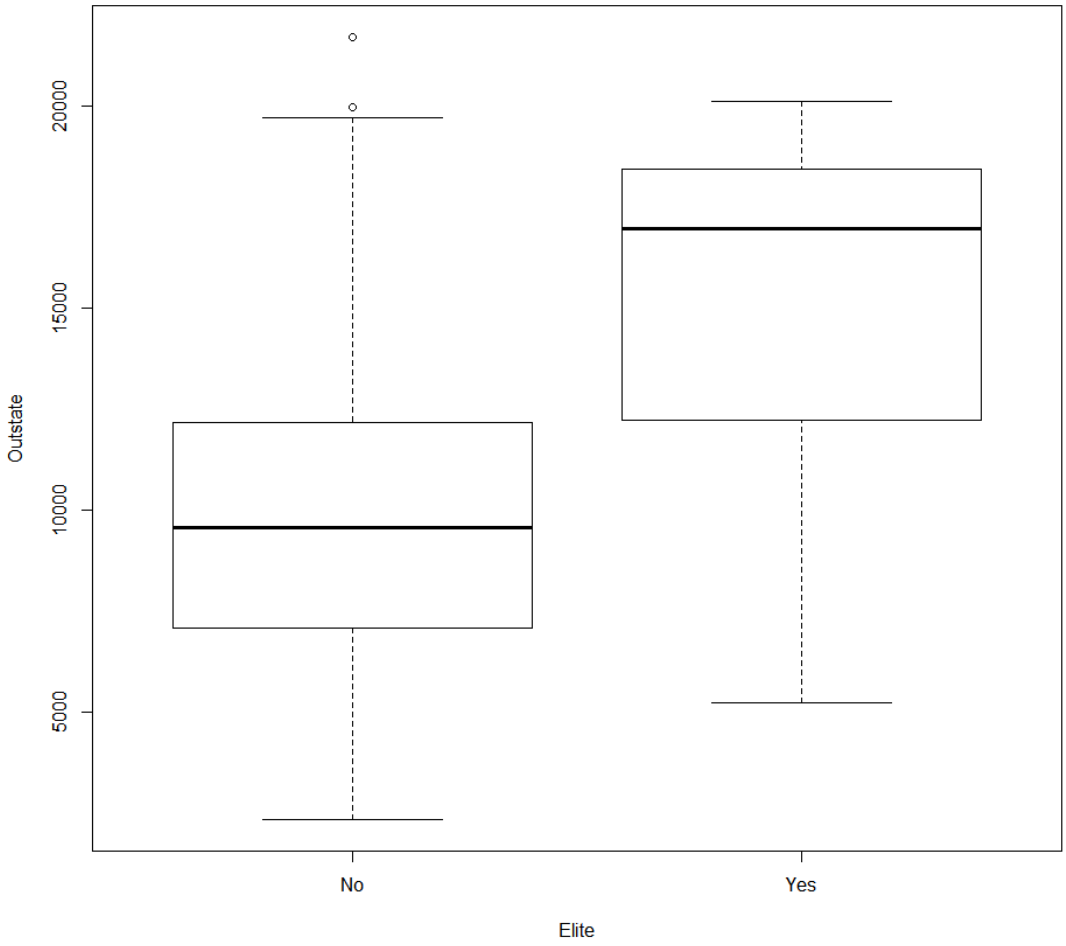
(ii)



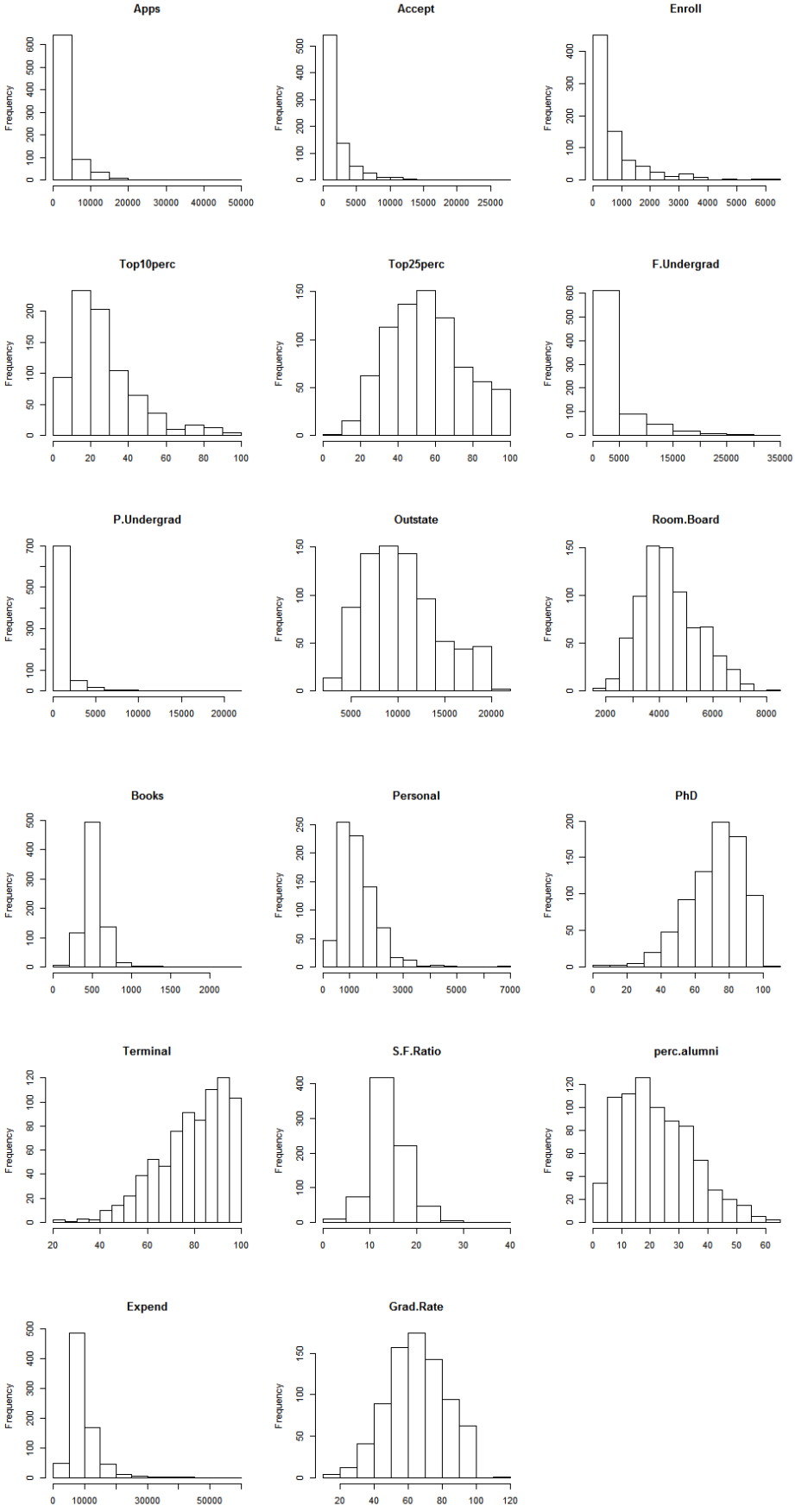
(iii)



(iv) Elite: 78, not Elite: 699



(v)


Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate
Frequency

(vi) 從(ii)圖中可以發現，某幾對變數明顯呈正相關。進一步的研究除了可以討論兩兩變數之間的關係以外，亦可以使用群集分析，將相似數值表現的學校歸唯一類，並進一步探討之間的關係。由(iii)、(iv)之兩盒鬚圖可以發現，在 Private 與 Elite 兩個變數中，yes 的 outstate 都明顯比 no 多，若要進一步了解是否有顯著差異，則可以使用假設檢定之結果作為判斷依據。

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

9.b.

9.c.

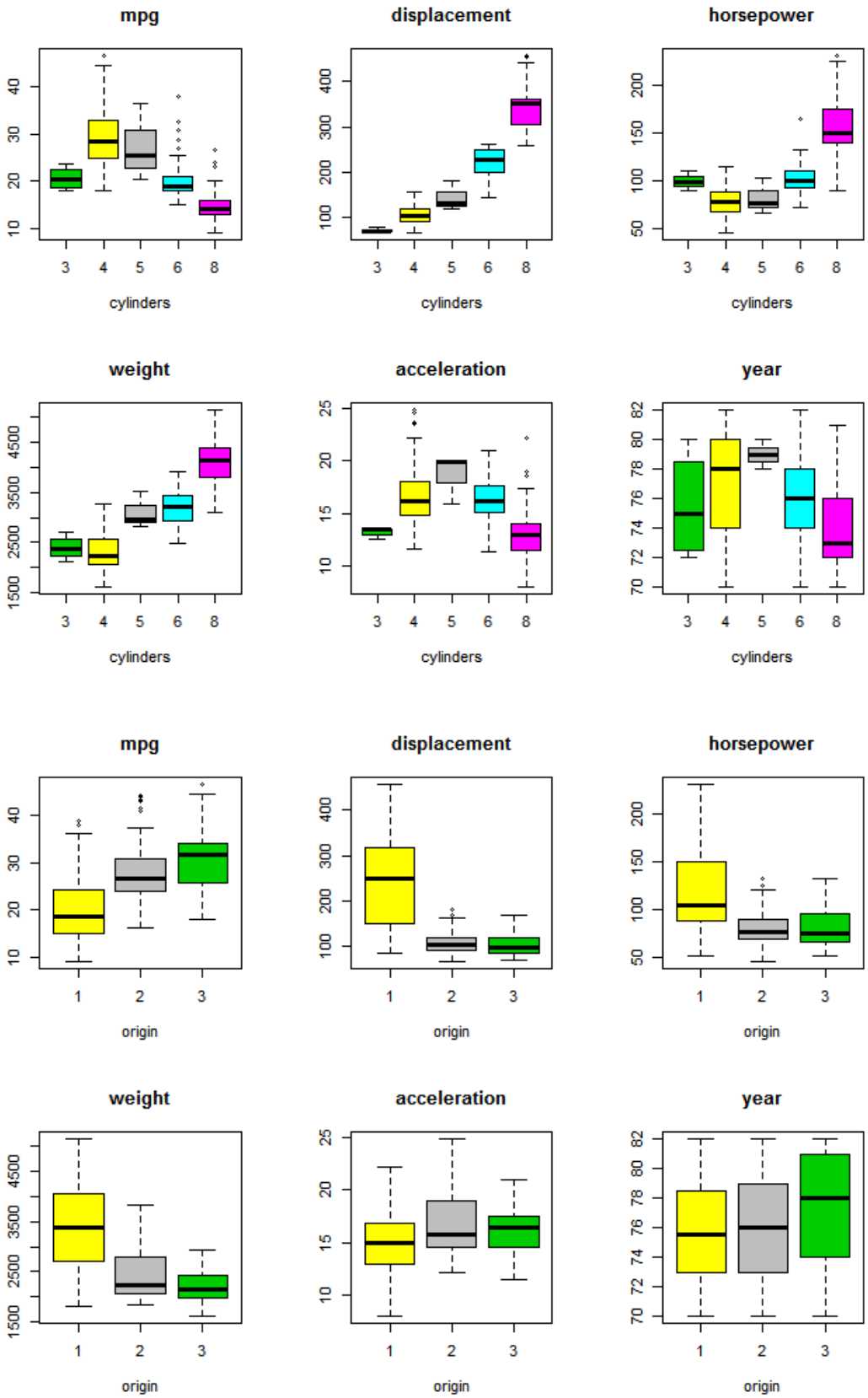
```
> Auto.bc
               min    max  range        mean          sd
mpg              9   46.6   37.6    23.51587   7.825804
displacement    68  455.0  387.0   193.53275 104.379583
horsepower      46  230.0  184.0   104.46939  38.491160
weight        1613 5140.0 3527.0 2970.26196 847.904119
acceleration     8   24.8   16.8    15.55567   2.749995
year            70   82.0   12.0    75.99496   3.690005
```

9.d.

```
> Auto.dd
                 min    max  range        mean          sd
mpg             11.0   46.6   35.6    24.43863   7.908184
displacement    68.0  455.0  387.0   187.04984  99.635385
horsepower      46.0  230.0  184.0   100.95584  35.895567
weight        1649.0 4997.0 3348.0 2933.96262 810.642938
acceleration     8.5   24.8   16.3    15.72305   2.680514
year            70.0   82.0   12.0    77.15265   3.111230
```

range 與 sd 因為樣本數減少而下降，mean 之增減則無依定規則。

9.e.


mpg
displacement
horsepower
weight
acceleration
year
cylinders
origin

可以發現汽缸的數量影響其他述職的表現，以排氣量來說汽缸的數量越多，其值越高。由年份來看，五汽缸的年份較其他汽缸數之車輛新，可以推估五汽缸引擎較晚被發明與應用。另外，汽缸數越高並不代表加速越快，但是馬力可能有較好的表現。

以來原來說，來源一的汽車在排氣量、馬力普遍有較好的表現，但是車身重量也比較高。

9.f. 預測模型可以使用線性迴歸預測。

```
> mpg.lm <- lm(mpg ~ factor(cylinders) + displacement + horsepower
+ + weight + acceleration + year + factor(origin))
> summary(mpg.lm)

Call:
lm(formula = mpg ~ factor(cylinders) + displacement + horsepower +
    weight + acceleration + year + factor(origin))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6797 -1.9373 -0.0678  1.6711 12.7756

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
factor(cylinders)4  6.722e+00  1.654e+00   4.064 5.85e-05 ***
factor(cylinders)5  7.078e+00  2.516e+00   2.813  0.00516 **
factor(cylinders)6  3.351e+00  1.824e+00   1.837  0.06701 .
factor(cylinders)8  5.099e+00  2.109e+00   2.418  0.01607 *
displacement        1.870e-02  7.222e-03   2.590  0.00997 **
horsepower         -3.490e-02  1.323e-02  -2.639  0.00866 **
weight             -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
acceleration        2.598e-02  9.304e-02   0.279  0.78021
year                7.370e-01  4.892e-02  15.064  < 2e-16 ***
factor(origin)2     1.764e+00  5.513e-01   3.200  0.00149 **
factor(origin)3     2.617e+00  5.272e-01   4.964 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.098 on 380 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8425
F-statistic: 191.1 on 11 and 380 DF,  p-value: < 2.2e-16
```

可以看到大部分的變數，對於 mpg 有顯著之影響，若要有更佳的預測模型，則可以選用其他廣義線性迴歸，並選用 stepwise 等方法挑選解釋變數。

10.a.

The `Boston` data frame has 506 rows and 14 columns.

Usage

```
Boston
```

Format

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.

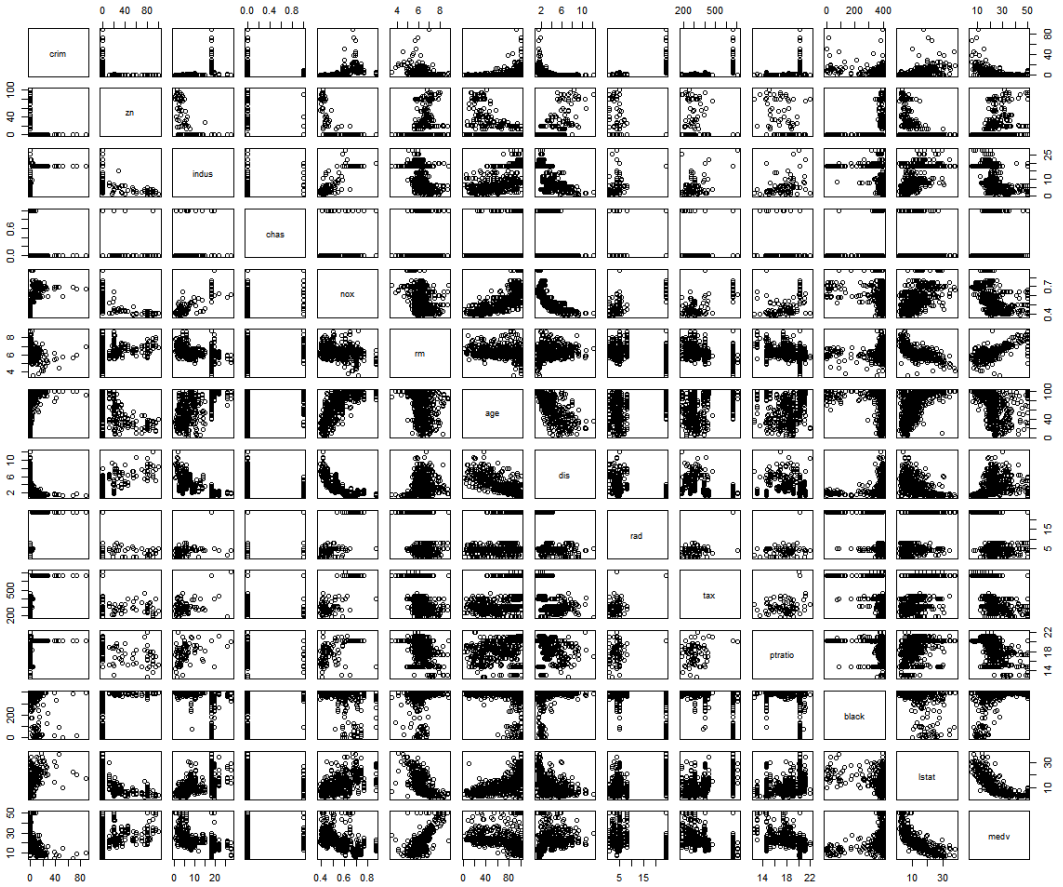
`ptratio`
: pupil-teacher ratio by town.

`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.

10.b.

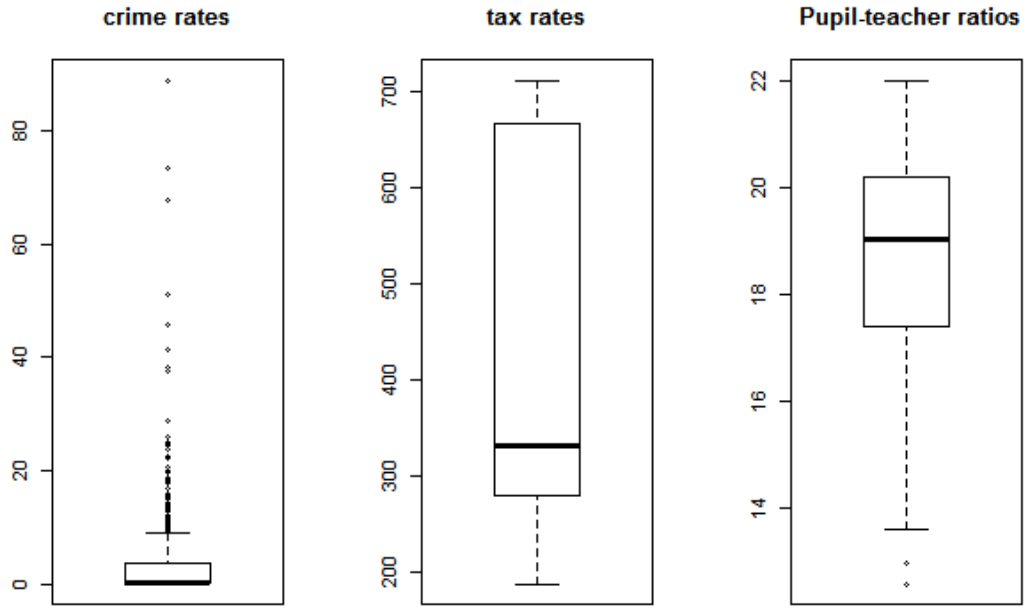


除了幾對變數有明顯的正相關、負相關之線性關係以外，age 與 lstat 兩變數之分散圖可以發現，age 越高 lstat 的值與範圍即越高，若要進行迴歸分析，則建議對 lstat 取 log 後，降低同 age 之 lstat 數值變異，避免 R-square 太小。另外以最右下角兩變數 lstat 與 medv，其關係可能不僅是負相關，亦有可能有倒數關係或指數關係，可以另外以其他廣義線性迴歸分析該數據。

10.c.

對 age 取對數，對 dis 取對數或是倒數，都有可能有較高的線性關係。

10.d.



Crime rates 大多呈城市偏低，多在 10 以內，但是也不少城市有屬於離群值，遠高於 10，最多有超過 60 的城鎮。稅率的部分約介於 200 至 700，且無離群值。師生比的部分大多介於 14 至 22 之間，有兩個城鎮屬於偏低之離群值。

10.e. 35 suburbs.

10.f. 19.5 p/t

10.g. suburb no. 399 and no. 406 有最低的 median value of owner-occupied homes.

```
> Boston[c(399, 406), ]
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
> summary(Boston)
      crim                zn             indus            chas              nox        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710  
       rm              age              dis              rad              tax        
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0  
 1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0  
 Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0  
 Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2  
 3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0  
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0  
    ptratio          black            lstat            medv      
 Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
```

兩城鎮皆大於 Q3：crim、indus、nox、age、rad、tax、ptratio、black、lstat

兩城鎮皆大於 Q1：zn、chas、rm、medv

10.h.

>7: 64 suburbs

>8: 13 suburbs

```
> Boston[no.rm8, ]
       crim zn indus chas    nox    rm  age    dis rad tax ptratio  black lstat medv
98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90  4.21 38.7
164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45  3.32 50.0
205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55  2.88 50.0
225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05  4.14 44.8
226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00  4.63 50.0
227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38  3.13 37.6
233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91  2.47 41.7
234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95  3.95 48.3
254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90  3.54 42.8
258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70  5.12 50.0
263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86  5.91 48.8
268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54  7.44 50.0
365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55  5.29 21.9
```

Black 的值皆大於平均值 356.67，甚至接近 Q3 值 396.9。lstat 大多小於 Q1。Age 除了 no.254 以外，其他都偏高。

Appendix

```
############## 8 ##############

### 8.a ###
college <- read.csv("College.csv", header = T, sep =",")
attach(college)


### 8.b ###
row.names(college) <- college[,1]
fix(college)

college <- college[,-1]
fix(college)


### 8.c ###
# (i) #
summary(college[, 2:length(college[1,])])
```

```
# (ii) #
par(mfrow = c(1,1))
pairs(college[,1:10])

# (iii) #
plot(Outstate~Private)

# (iv) #
Elite <- rep("No", nrow(college))
Elite[college$Top10perc >50] <- "Yes"
Elite <- as.vector(Elite)
summary(college)
college <- data.frame(college, Elite)
plot(Outstate ~ factor(Elite), xlab="Elite")

# (v) #
length(college[1,])
par(mfrow = c(3,3))

college.v <- names(college)
for (k in 2:18) {
   hist(college[,k], main = college.v[k], xlab = " ")
}

par(mfrow = c(1,1))

# (vi) #



############## 9 ##############
Auto <- read.csv("Auto.csv", header = T, sep =",")
attach(Auto)

### 9.a ###
# quantitative:
# quatlitative: cylinders, origin, name
```

```
### 9.b.c ###
Auto.bc <- data.frame(matrix(NA, 9,5))
names(Auto.bc) <- c("min", "max", "range", "mean", "sd")
row.names(Auto.bc) <- names(Auto)
Auto.bc <- Auto.bc[-9, ]

for (k in 1:length(Auto[1,])) {
   if (is.numeric(Auto[ ,k])==TRUE){
      Auto.bc[k, 1] <- range(Auto[,k])[1]
      Auto.bc[k, 2] <- range(Auto[,k])[2]
      Auto.bc[k, 3] <- range(Auto[,k])[2] - range(Auto[,k])[1]
      Auto.bc[k, 4] <- mean(Auto[,k])
      Auto.bc[k, 5] <- sd(Auto[,k])
   }
}

HP <- Auto$horsepower
for (k in 1:397) {
   if(is.na(HP[k]==TRUE)){
      HP <- HP[-k]
   }
}

Auto.bc[4,1] <- min(HP)
Auto.bc[4,2] <- max(HP)
Auto.bc[4,3] <- max(HP) - min(HP)
Auto.bc[4,4] <- mean(HP)
Auto.bc[4,5] <- sd(HP)

Auto.bc <- Auto.bc[-c(2,8), ]


### 9.d ###
Auto.d <- Auto[-c(10:85), ]
Auto.dd <- data.frame(matrix(NA, 9,5))
names(Auto.dd) <- c("min", "max", "range", "mean", "sd")
row.names(Auto.dd) <- names(Auto.d)
Auto.dd <- Auto.dd[-9, ]
```

```
for (k in 1:length(Auto.d[1,])) {
   if (is.numeric(Auto.d[ ,k])==TRUE){
      Auto.dd[k, 1] <- range(Auto.d[,k])[1]
      Auto.dd[k, 2] <- range(Auto.d[,k])[2]
      Auto.dd[k, 3] <- range(Auto.d[,k])[2] - range(Auto.d[,k])[1]
      Auto.dd[k, 4] <- mean(Auto.d[,k])
      Auto.dd[k, 5] <- sd(Auto.d[,k])
   }
}

HP.d <- Auto.d$horsepower
for (k in 1:length(Auto.d[,1])) {
   if(is.na(HP.d[k]==TRUE)){
      HP.d <- HP.d[-k]
   }
}

Auto.dd[4,1] <- min(HP.d)
Auto.dd[4,2] <- max(HP.d)
Auto.dd[4,3] <- max(HP.d) - min(HP.d)
Auto.dd[4,4] <- mean(HP.d)
Auto.dd[4,5] <- sd(HP.d)

Auto.dd <- Auto.dd[-c(2,8), ]


### 9.e ###
par(mfrow = c(1,1))
pairs(Auto[,-9])

par(mfrow = c(2,3))
for (k in c(1,3,4,5,6,7)) {
   plot(Auto[,k]~ factor(origin),
        xlab = "origin", ylab = " ", main =names(Auto)[k],
        col = c(7, 8, 11))
}
```

```
par(mfrow = c(2,3))
for (k in c(1,3,4,5,6,7)) {
   plot(Auto[,k]~ factor(cylinders),
          xlab = "cylinders", ylab = " ", main =names(Auto)[k],
          col = c(11,7,8,5,6))
}



### 9.f ###
mpg.lm <- lm(mpg ~ factor(cylinders) + displacement + horsepower
      + weight + acceleration + year + factor(origin))

summary(mpg.lm)



############## 10 ##############
### 10.a ###
library(MASS)
Boston <- Boston
attach(Boston)
?Boston


### 10.b ###
pairs(Boston)



### 10.c ###

### 10.d ###
par(mfrow=c(1,3))
boxplot(crim, main="crime rates")
boxplot(tax, main="tax rates")
boxplot(ptratio, main=" Pupil-teacher ratios")
```

```
### 10.e ###
sum(chas)

### 10.f ###
median(Boston$ptratio)

### 10.g ###
for (k in 1:506) {
   if (Boston$medv[k]==min(Boston$medv)){
      print(k)
   }
}

Boston[c(399, 406), ]
summary(Boston)

### 10.h ###
rm7 <- array(NA, 0)
for (k in 1:506) {
   if (Boston$rm[k]>7){
      rm7[length(rm7)+1] <- Boston$rm[k]
   }
}
length(rm7)


rm8 <- array(NA, 0)
no.rm8 <- array(NA, 0)
for (k in 1:506) {
   if (Boston$rm[k]>8){
      rm8[length(rm8)+1] <- Boston$rm[k]
      no.rm8[length(no.rm8)+1] <- k
   }
}
length(rm8)

Boston[no.rm8, ]
```