**Statistical Learning HW1_106070038 杜葳葳**

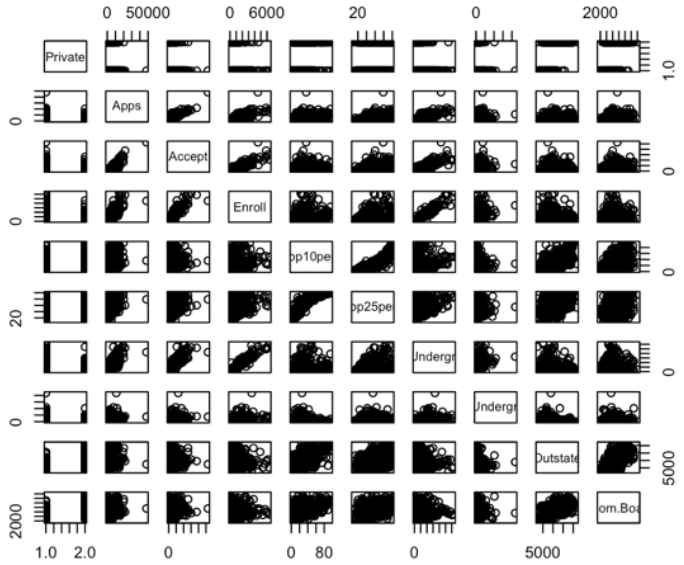
8. (a) 讀入 csv 檔，並將資料存在變數 college

(b) 將各大學的名稱設為 row names，接著把 x 那欄刪除，因此第一欄為 private

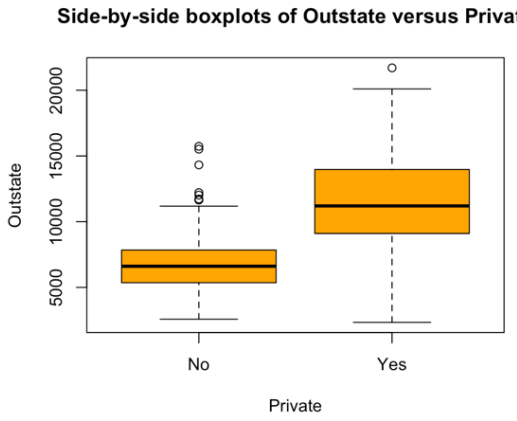
(c) i. 列出每個欄位的最小值、第一四分位數、中位數、平均值、第三四分位數、最大值

```
   Private             Apps           Accept          Enroll       Top10perc
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
 Mode  :character   Median : 1558   Median : 1110   Median : 434   Median :23.00
                    Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
                    Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate       Room.Board
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
      Books           Personal         PhD            Terminal       S.F.Ratio
 Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
 1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
 Median : 500.0   Median :1200   Median : 75.00   Median : 82.0   Median :13.60
 Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
 3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
 Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
  perc.alumni        Expend        Grad.Rate
 Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :21.00   Median : 8377   Median : 65.00
 Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :64.00   Max.   :56233   Max.   :118.00
```

ii. 先將 Private 那欄的 type 轉成 categorical，再將前十個欄位兩兩畫出散佈圖（scatterplot）

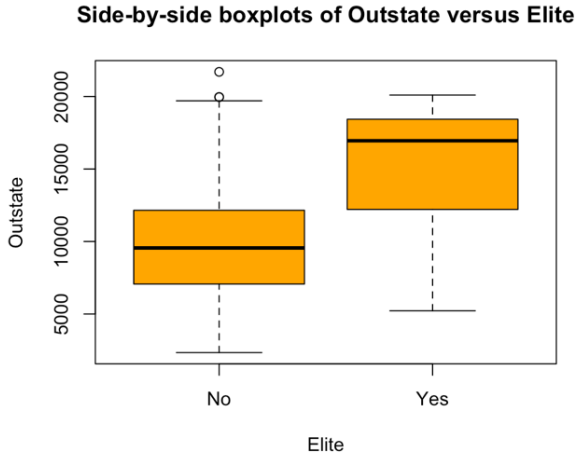



iii. 進行 Outstate 和 Private 兩組數據特徵的盒方圖比較，可看出最大最小值、Q1、Q3、中位數、離群值，可由圖中推論，私立大學 outstate tuition 的分布範圍較大、且平均值也較大

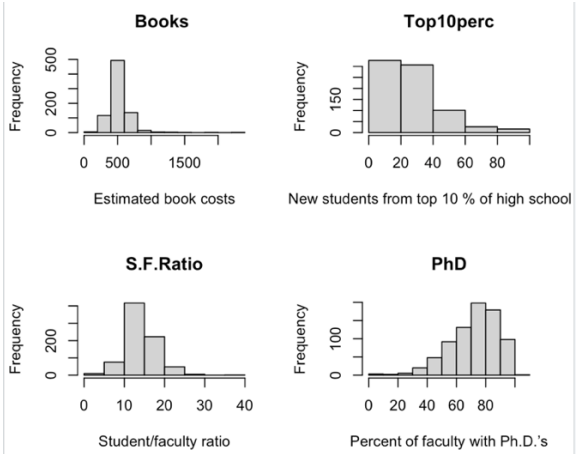


iv. 新增一欄 Elite 並將值全部設為”No”，接著將符合 Top10perc >50 條件的設為”Yes”，將 Elite 那欄的資料型別轉為 categorical，將 college 和 Elite 兩個 dataframe 合併。共有 **78 所學校**符合條件，最後類似 iii.，將 Outstate 和 Elite 兩組數據特徵的箱形圖做比較

```
> summary(Elite)
 No Yes
699  78
```



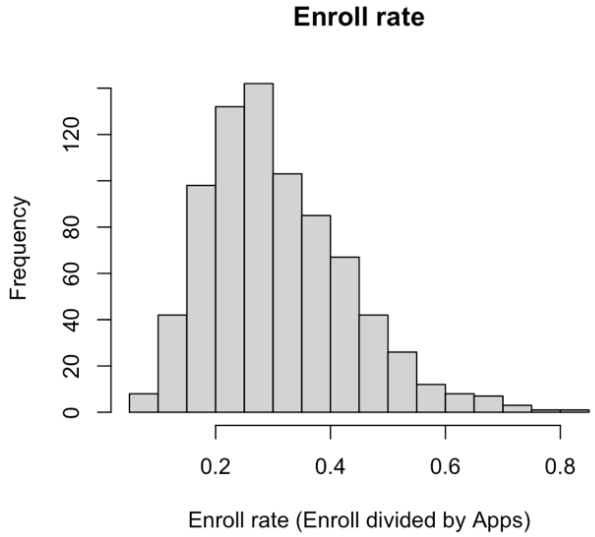

v. 可用參數 breaks 自行設定 bin 的數目，並使用 par(mfrow=c(2,2))切分畫面、同時呈現四個直方圖




vi.

- 入學率：使用各大學的入學人數（Enroll）除以申請人數（Apps），得到各大學的入學率，並分析錄取率的各項統計量（如下），與將分佈畫成直方圖，在 777 所學校中入學率的中位數為 0.29152

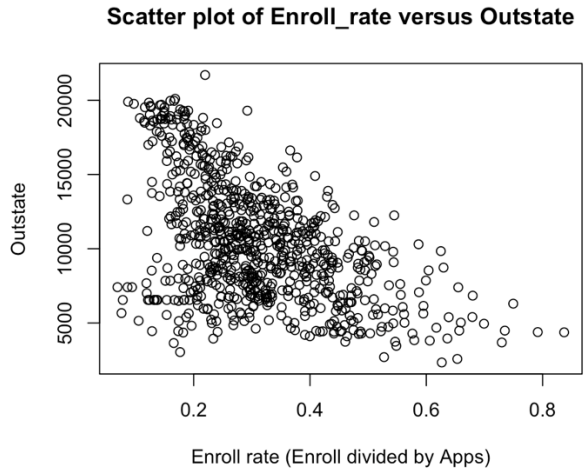
```
  Enroll_rate
Min.   :0.06892
1st Qu.:0.22011
Median :0.29152
Mean   :0.30937
3rd Qu.:0.38268
Max.   :0.83705
```

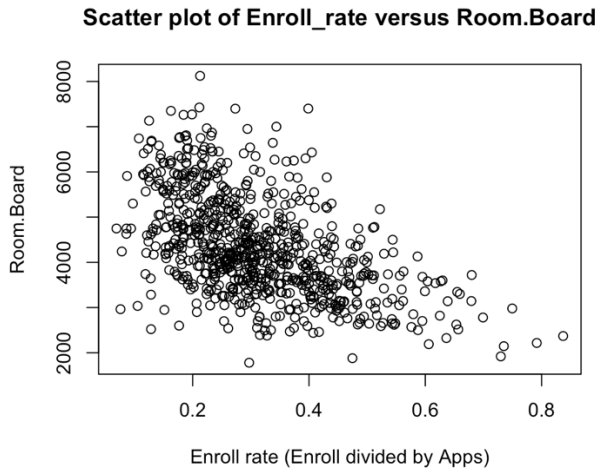


- **探討入學率與其他因素的相關性**：假設入學率可以做為該校熱門程度的指標，想藉由入學率與其他因素的相關性，找出有關聯的指標

| | Top10perc | Top25perc | Outstate | F.Undergrad | P.Undergrad | Books | Room.Board |
|---|---|---|---|---|---|---|---|
| **和 Enroll_rate 的 correlation** | **-0.35734** | **-0.34880** | **-0.50719** | 0.00642 | 0.05752 | -0.06216 | **-0.51989** |
| | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
| **和 Enroll_rate 的 correlation** | 0.17307 | **-0.34474** | **-0.35160** | 0.24702 | -0.22756 | **-0.40566** | **-0.38587** |

由上表可知，**Enroll_rate 和 Top10perc、Top25perc、Outstate、Room.Board、PhD、Terminal、Expend、Grad.Rate 為中度負相關**，將負相關程度最高的 Outstate 和 Room.Board 繪製成散佈圖如下：



10. (a) 506 rows, 14 columns

dataset 為波士頓各城鎮的住房數據，rows 共有 506 個，**每一個 row 是一個城鎮地區**。

columns 有 14 個，依序為:

crim:城鎮人均犯罪率，zn:住宅用地超過 25,000 sq.ft.的比例，indus:非零售業的比例，chas:查爾斯河虛擬變數，nox:氮氧化物的濃度，rm:每個住宅平均房間數，age:1940 年前建商自己擁有的比例，dis:到達五個波士頓就業中心的加權距離，rad:到達放射狀公路的係數，tax:每 10,000 美元所有財產價值的稅率，ptratio:城鎮師生比，black:按 1000(Bk - 0.63)^2 計算的城鎮黑人比例，lstat:地位較低的人的百分比，medv:平均 1,000 美元的自住戶中位數

| | crim | zn | indus | chas | nox | rm | age | dis | rad | tax | ptratio | black | lstat | medv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00632 | 18.0 | 2.31 | 0 | 0.5380 | 6.575 | 65.2 | 4.0900 | 1 | 296 | 15.3 | 396.90 | 4.98 | 24.0 |
| 2 | 0.02731 | 0.0 | 7.07 | 0 | 0.4690 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 396.90 | 9.14 | 21.6 |
| 3 | 0.02729 | 0.0 | 7.07 | 0 | 0.4690 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 392.83 | 4.03 | 34.7 |
| 4 | 0.03237 | 0.0 | 2.18 | 0 | 0.4580 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 394.63 | 2.94 | 33.4 |
| 5 | 0.06905 | 0.0 | 2.18 | 0 | 0.4580 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 396.90 | 5.33 | 36.2 |
| 6 | 0.02985 | 0.0 | 2.18 | 0 | 0.4580 | 6.430 | 58.7 | 6.0622 | 3 | 222 | 18.7 | 394.12 | 5.21 | 28.7 |
| 7 | 0.08829 | 12.5 | 7.87 | 0 | 0.5240 | 6.012 | 66.6 | 5.5605 | 5 | 311 | 15.2 | 395.60 | 12.43 | 22.9 |
| 8 | 0.14455 | 12.5 | 7.87 | 0 | 0.5240 | 6.172 | 96.1 | 5.9505 | 5 | 311 | 15.2 | 396.90 | 19.15 | 27.1 |
| 9 | 0.21124 | 12.5 | 7.87 | 0 | 0.5240 | 5.631 | 100.0 | 6.0821 | 5 | 311 | 15.2 | 386.63 | 29.93 | 16.5 |
| 10 | 0.17004 | 12.5 | 7.87 | 0 | 0.5240 | 6.004 | 85.9 | 6.5921 | 5 | 311 | 15.2 | 386.71 | 17.10 | 18.9 |
| 11 | 0.22489 | 12.5 | 7.87 | 0 | 0.5240 | 6.377 | 94.3 | 6.3467 | 5 | 311 | 15.2 | 392.52 | 20.45 | 15.0 |
| 12 | 0.11747 | 12.5 | 7.87 | 0 | 0.5240 | 6.009 | 82.9 | 6.2267 | 5 | 311 | 15.2 | 396.90 | 13.27 | 18.9 |
| 13 | 0.09378 | 12.5 | 7.87 | 0 | 0.5240 | 5.889 | 39.0 | 5.4509 | 5 | 311 | 15.2 | 390.50 | 15.71 | 21.7 |
| 14 | 0.62976 | 0.0 | 8.14 | 0 | 0.5380 | 5.949 | 61.8 | 4.7075 | 4 | 307 | 21.0 | 396.90 | 8.26 | 20.4 |
| 15 | 0.63796 | 0.0 | 8.14 | 0 | 0.5380 | 6.096 | 84.5 | 4.4619 | 4 | 307 | 21.0 | 380.02 | 10.26 | 18.2 |
| 16 | 0.62739 | 0.0 | 8.14 | 0 | 0.5380 | 5.834 | 56.5 | 4.4986 | 4 | 307 | 21.0 | 395.62 | 8.47 | 19.9 |
| 17 | 1.05393 | 0.0 | 8.14 | 0 | 0.5380 | 5.935 | 29.3 | 4.4986 | 4 | 307 | 21.0 | 386.85 | 6.58 | 23.1 |
| 18 | 0.78420 | 0.0 | 8.14 | 0 | 0.5380 | 5.990 | 81.7 | 4.2579 | 4 | 307 | 21.0 | 386.75 | 14.67 | 17.5 |
| 19 | 0.80271 | 0.0 | 8.14 | 0 | 0.5380 | 5.456 | 36.6 | 3.7965 | 4 | 307 | 21.0 | 288.99 | 11.69 | 20.2 |
| 20 | 0.72580 | 0.0 | 8.14 | 0 | 0.5380 | 5.727 | 69.5 | 3.7965 | 4 | 307 | 21.0 | 390.95 | 11.28 | 18.2 |

Boston {MASS} R Documentation

## Housing Values in Suburbs of Boston

### Description

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

### Usage

```
Boston
```

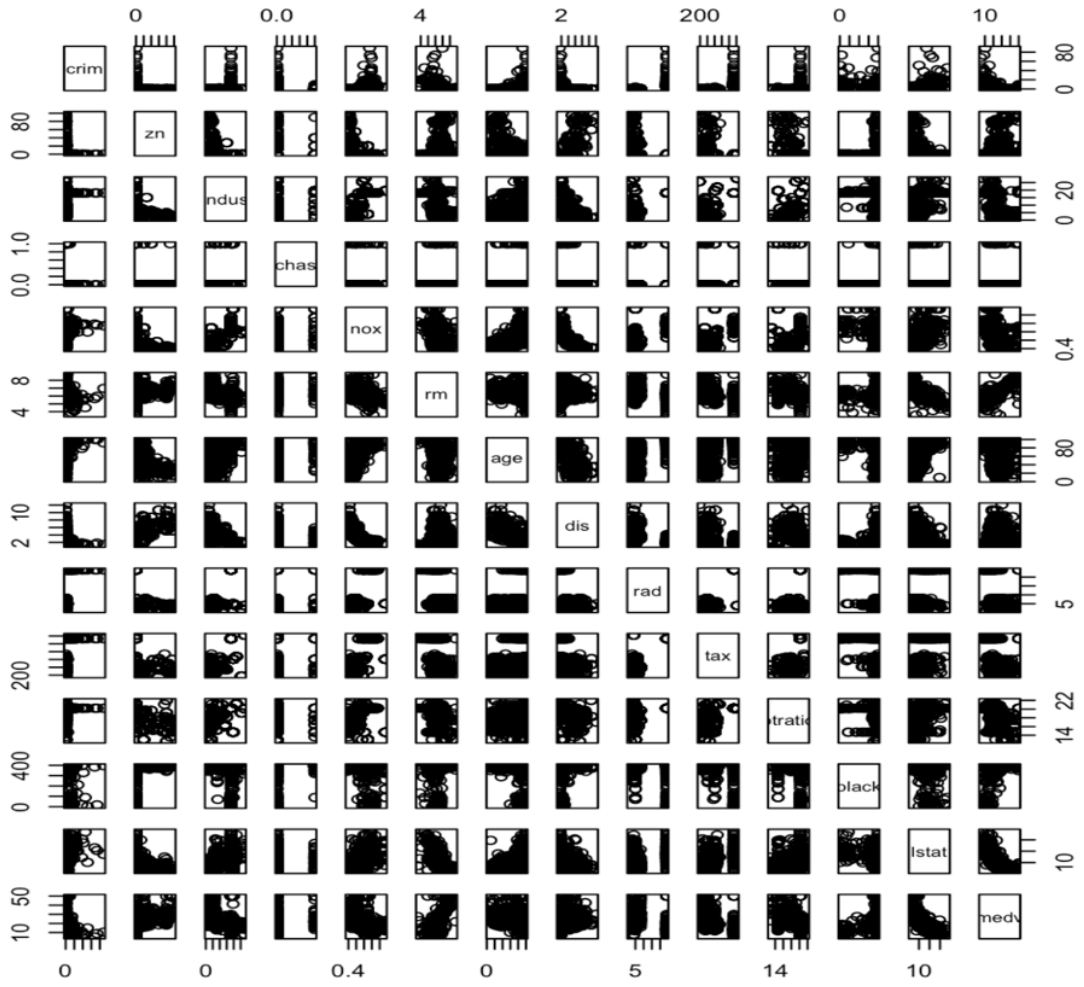
### Format

This data frame contains the following columns:

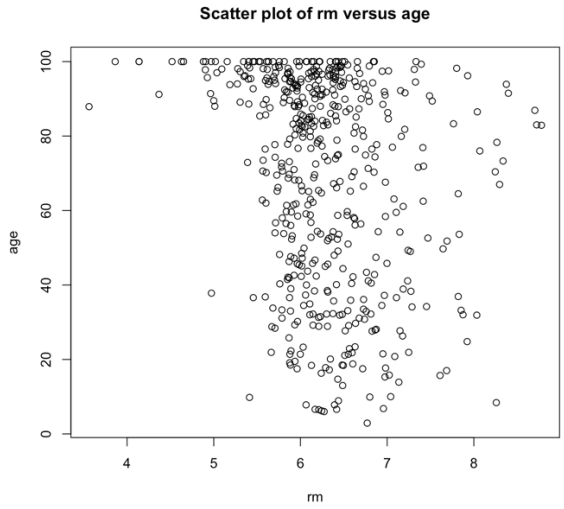
`crim`
per capita crime rate by town.

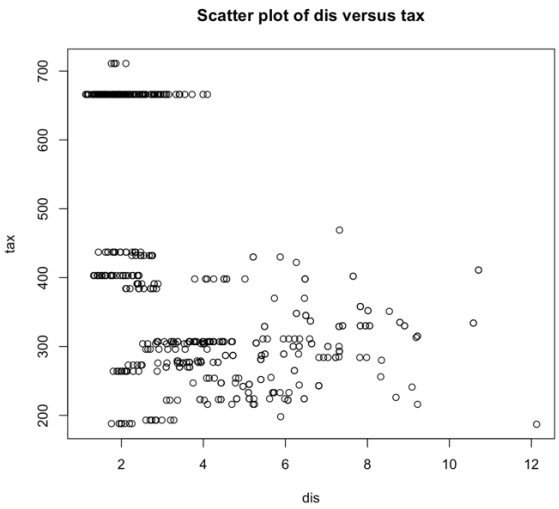
`zn`
proportion of residential land zoned for lots over 25,000 sq.ft.

(b) 所有 column 的矩陣散佈圖




因爲此圖太小，故將兩組參數特別獨立出來比較，如下頁:



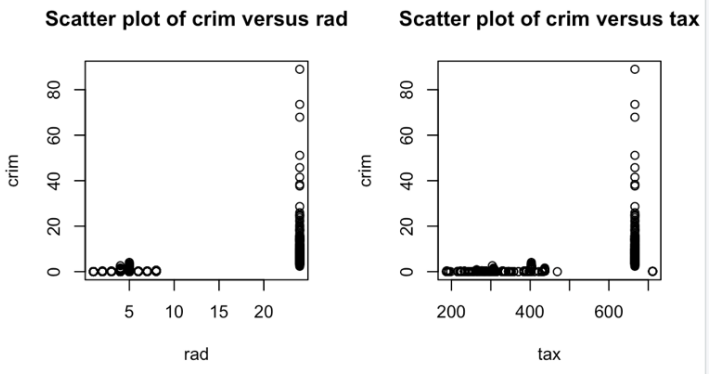

rm(平均房間數) v.s age(建商自己擁有的比例): 平均房間數較少（區間 4-5）的城鎮、房屋大多是建商自己擁有，其餘多數城鎮的房間數皆集中在區間 6-7

dis(到達就業中心的加權距離) v.s tax(稅率): 稅率高的皆距離市中心較近（區間 1-4）

(c) 計算各個參數和 crim 的 correlation

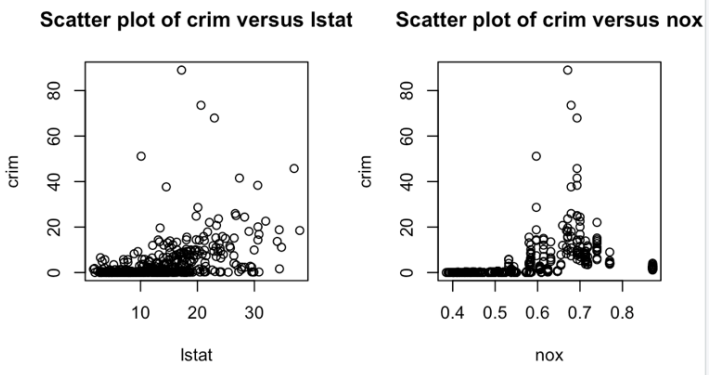
```
         zn       indus        chas         nox          rm         age         dis
-0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670  0.35273425 -0.37967009
        rad         tax     ptratio       black       lstat        medv
 0.62550515  0.58276431  0.28994558 -0.38506394  0.45562148 -0.38830461
```

**和 crim 正相關程度最高的四個參數： rad、tax、lstat、 nox**



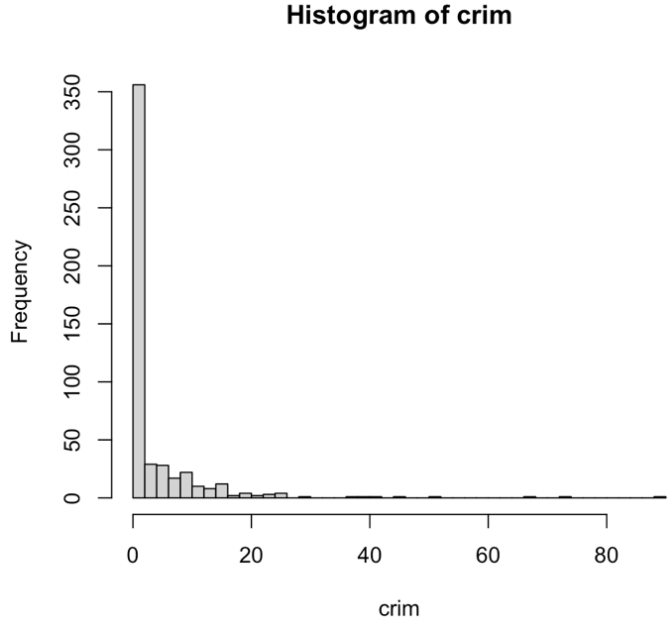






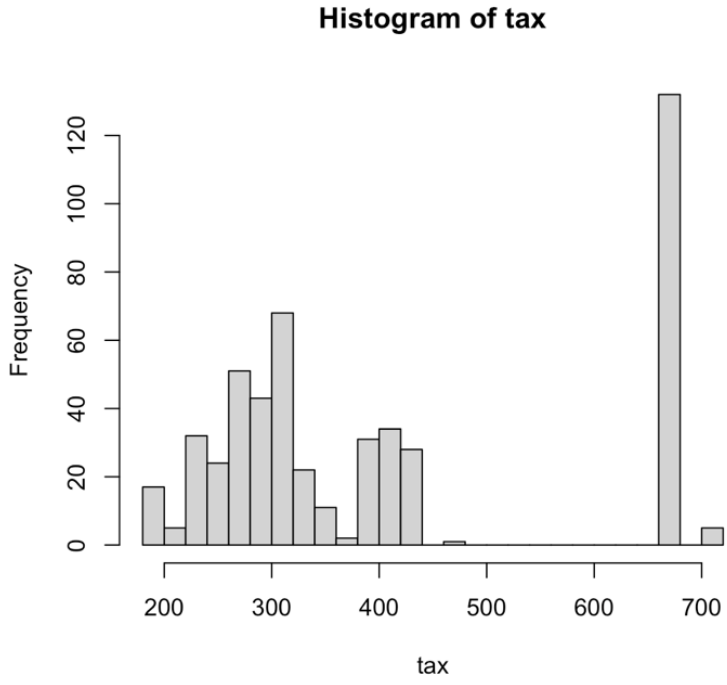
(d) 使用 mean()和 sd()計算平均和標準差，有額外寫一個 function 來計算有多少地區符合 particularly high 的條件

crim rate:假設平均加兩倍標準差 **20.81661（3.61352+2*8.601545）為 particularly high**，呼叫上面所寫的 function、同時傳入參數，總共有 **16 個地區**符合條件



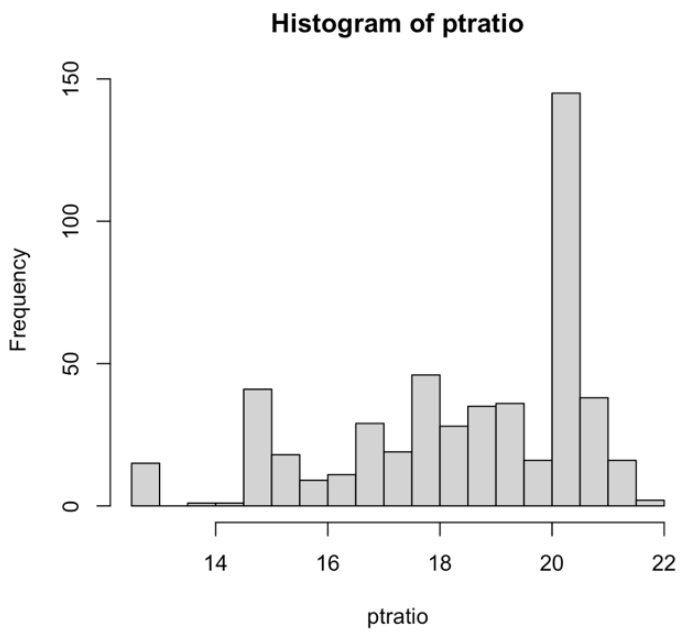

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  0.00632   0.08204   0.25651   3.61352   3.67708  88.97620
```

tax rates: 假設平均加一倍標準差 **576.7743（408.2372+1* 168.5371）為 particularly high**，呼叫上面所寫的 function、同時傳入參數，總共有 **137 個地區**符合條件




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  187.0   279.0   330.0   408.2   666.0   711.0
```

Pupil-teacher rates: 假設平均加一倍標準差 **20.62048（18.45553 +1* 2.164946）為 particularly high**，呼叫上面所寫的 function、同時傳入參數，總共有 **56** 個地區符合條件




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.60   17.40   19.05   18.46   20.20   22.00
```

(e) 35suburbs 環繞查爾斯河

```
> table(Boston[,'chas'])

  0   1
471  35
```

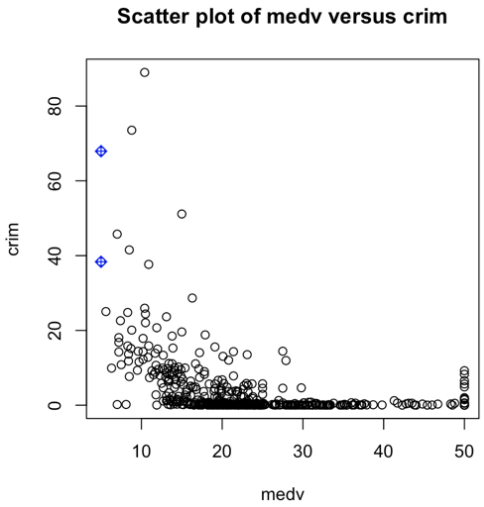
(f) ptratio 的中位數為 19.05

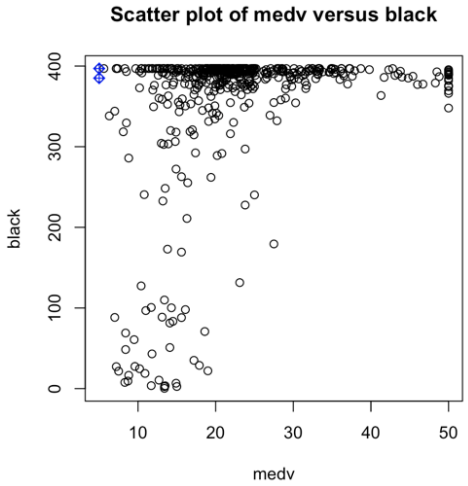
```
> median(Boston[,'ptratio'])
[1] 19.05
```

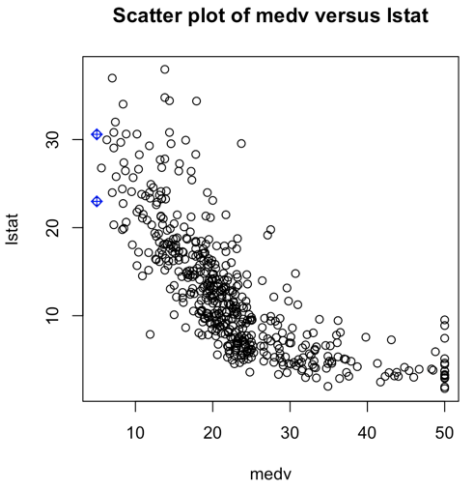
(g) medv 的最小值為 5，有兩個地區的 medv 為 5，以下為該地區的其他參數

```
> Boston[399,1:14]
       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    20.2 396.9 30.59    5
> Boston[406,1:14]
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
```

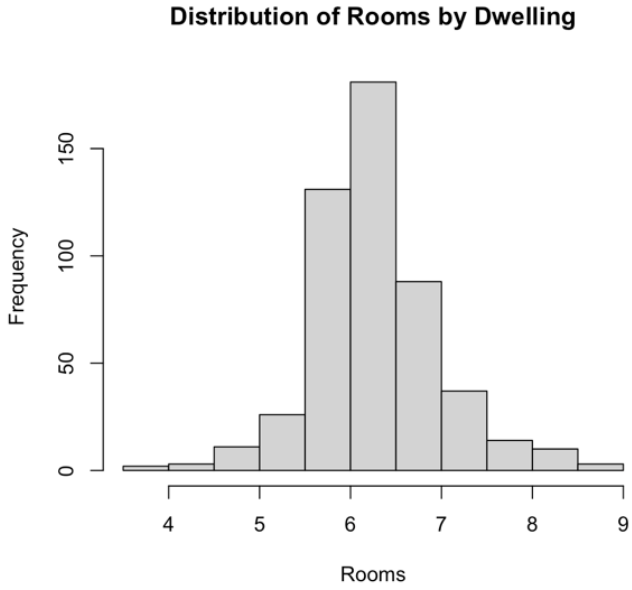
將以上兩個地區用⌖標示在散佈圖上，從圖中可觀察到此二個地區的「人均犯罪率」、「黑人比例」、「地位較低的人的百分比」較其他多數地區高




(h) 下圖為每個住宅平均房間數的分佈




用 counter、for 迴圈和一個判斷式做計算，平均房間數大於 7 的有 **64** 個、大於 8 有 **13** 個，因此推論平均房間數量大於 7 的大多為 7-8 之間（51 個地區），大於 8 的較少（13 個地區）

```
> print(count_7)
[1] 64
```

```
> print(count_8)
[1] 13
```

附錄：**R** 語言程式碼

```
# HW1
install.packages("glmnet")
install.packages("XQuartz")
library(glmnet)
library(datasets)
library(XQuartz)
# 8.(a) 讀入檔案
college = read.csv("/Statistical Learning/College.csv", stringsAsFactors=FALSE)
# 8.(b) 把每個 row 的名字設為 university
rownames (college) = college[,1]
fix (college)
# 把第一個 column 刪掉
college = college[,-1]
fix (college)
# 8.(c) i
summary (college)
# 8.(c) ii
college$Private = as.factor (college$Private)
pairs(college[,1:10])
# 8.(c) iii
Boxplot (college$Outstate ~ college$Private, col = "orange", main = "Side-by-side boxplots of Outstate versus
Private", ylab = "Outstate", xlab = "Private")
# 8.(c) iv
Elite = rep("No", nrow(college))
Elite [college$Top10perc >50] = "Yes"
as.factor (Elite)
college = data.frame (college,Elite)
# there are 78 elite universities
summary (Elite)
summary (college)
boxplot (college$Outstate ~ college$Elite, col = "orange", main = "Side-by-side boxplots of Outstate versus
Elite", ylab = "Outstate", xlab = "Elite")
# 8.(c) v
par (mfrow = c(2,2))
hist (college$Books,main = "Books", xlab = "Estimated book costs", breaks = 9)
hist (college$Top10perc, main = "Top10perc", xlab = "New students from top 10 % of high school class",
breaks = 5)
hist (college$S.F.Ratio, main = "S.F.Ratio", xlab = "Student/faculty ratio")
hist (college$PhD, main = "PhD", xlab = "Percent of faculty with Ph.D.'s")
# 8.(c) vi
## Calculate the accept rate
Enroll_rate = rep (0, nrow(college))
for (i in 1:777) {
   Enroll_rate[i] = college[i,4] / college[i,2]
}
summary (Enroll_rate)
par (mfrow = c(1,1))
college = data.frame (college, Enroll_rate)
```

```
summary (college)
hist (college$Enroll_rate, main = "Enroll rate", xlab = "Enroll rate (Enroll divided by Apps)")
## Calculate the correlation between Enroll_rate and every columns
cor (college$Enroll_rate, college$Top10perc)
cor (college$Enroll_rate, college$Top25perc)
cor (college$Enroll_rate, college$Outstate)
cor (college$Enroll_rate, college$F.Undergrad)
cor (college$Enroll_rate, college$P.Undergrad)
cor (college$Enroll_rate, college$Room.Board)
cor (college$Enroll_rate, college$Books)
cor (college$Enroll_rate, college$Personal)
cor (college$Enroll_rate, college$PhD)
cor (college$Enroll_rate, college$Terminal)
cor (college$Enroll_rate, college$S.F.Ratio)
cor (college$Enroll_rate, college$perc.alumni)
cor (college$Enroll_rate, college$Expend)
cor (college$Enroll_rate,college$Grad.Rate)
plot (college$Enroll_rate, college$Outstate, main = "Scatter plot of Enroll_rate versus Outstate", xlab = "Enroll
rate (Enroll divided by Apps)", ylab = "Outstate")
plot (college$Enroll_rate, college$Room.Board, main = "Scatter plot of Enroll_rate versus Room.Board", xlab
= "Enroll rate (Enroll divided by Apps)", ylab = "Room.Board")
# 10.(a)
library(MASS)
Boston
?Boston
nrow (Boston) # how many rows?
ncol (Boston) # how many cols?
# 10.(b)
pairs (Boston[,1:14])
plot (Boston[,6], Boston[,7], main = "Scatter plot of rm versus age", xlab = colnames (Boston)[6], ylab =
colnames(Boston)[7])
plot (Boston[,8], Boston[,10], main = "Scatter plot of dis versus tax", xlab = colnames(Boston)[8], ylab =
colnames(Boston)[10])
# 10.(c)
cor (Boston)
## correlation with crim
Boston.corr[-1,1]
par (mfrow = c(2,2))
plot (Boston[,'rad'], Boston[,'crim'], main = "Scatter plot of crim versus rad", xlab = 'rad', ylab = 'crim')
plot (Boston[,'tax'], Boston[,'crim'], main = "Scatter plot of crim versus tax", xlab = 'tax', ylab = 'crim')
plot (Boston[,'lstat'], Boston[,'crim'], main = "Scatter plot of crim versus lstat", xlab = 'lstat', ylab = 'crim')
plot (Boston[,'nox'], Boston[,'crim'], main = "Scatter plot of crim versus nox", xlab = 'nox', ylab = 'crim')
# 10.(d)
par (mfrow = c(1,1))
count <- function (col_name, num) {
   counter = 0
   for (i in 1:506) {
      if (Boston[i, col_name] > num)
         counter = counter+1
   }
   print (num)
   return (counter)
```

```
}
## high crim
high_crim = mean (Boston[,'crim']) + 2*sd (Boston[,'crim'])
count ('crim', high_crim)
hist (Boston[,'crim'], main = "Histogram of crim", xlab = 'crim', breaks = 50)
summary (Boston[,'crim'])
## high tax
high_tax = mean (Boston[,'tax']) + 1*sd (Boston[,'tax'])
count ('tax',high_tax)
hist (Boston[,'tax'], main = "Histogram of tax", xlab = 'tax', breaks = 30)
summary (Boston[,'tax'])
## high ptratio
high_ptratio = mean(Boston[,'ptratio'])+1*sd(Boston[,'ptratio'])
count ('ptratio', high_ptratio)
hist (Boston[,'ptratio'], main = "Histogram of ptratio", xlab = 'ptratio', breaks = 30)
summary (Boston[,'ptratio'])
# 10.(e)
table (Boston[,'chas'])
# 10.(f)
median (Boston[,'ptratio'])
# 10.(g)
for (i in 1:506) {
   if (Boston[i,'medv'] == min(Boston[,'medv']))
      min_data <- min_data + Boston[i,1:14]
}
## plot
plot (Boston[,'medv'], Boston[,'crim'], main = "Scatter plot of medv versus crim", xlab = 'medv', ylab = 'crim')
points (x = Boston[399,1:14]$medv, y = Boston[399,1:14]$crim, pch = 9, col = "blue")
points (x = Boston[406,1:14]$medv, y = Boston[406,1:14]$crim, pch = 9, col = "blue")
plot (Boston[,'medv'], Boston[,'lstat'], main = "Scatter plot of medv versus lstat", xlab = 'medv', ylab = 'lstat')
points (x = Boston[399,1:14]$medv, y = Boston[399,1:14]$lstat, pch = 9, col = "blue")
points (x = Boston[406,1:14]$medv, y = Boston[406,1:14]$lstat, pch = 9, col = "blue")
## print other predictors
Boston [399, 1:14]
Boston [406, 1:14]
# 10.(h)
summary (Boston[,'rm'])
hist (Boston[,'rm'], main = "Distribution of Rooms by Dwelling", xlab = "Rooms")
## rm>7
count_7 = 0
for (i in 1:506) {
   if (Boston[i,'rm'] > 7 )
      count_7 = count_7+1
}
print (count_7)
## rm>8
count_8 = 0
for (i in 1:506) {
   if (Boston[i,'rm'] > 8 )
      count_8 = count_8+1
}
print(count_8)
```