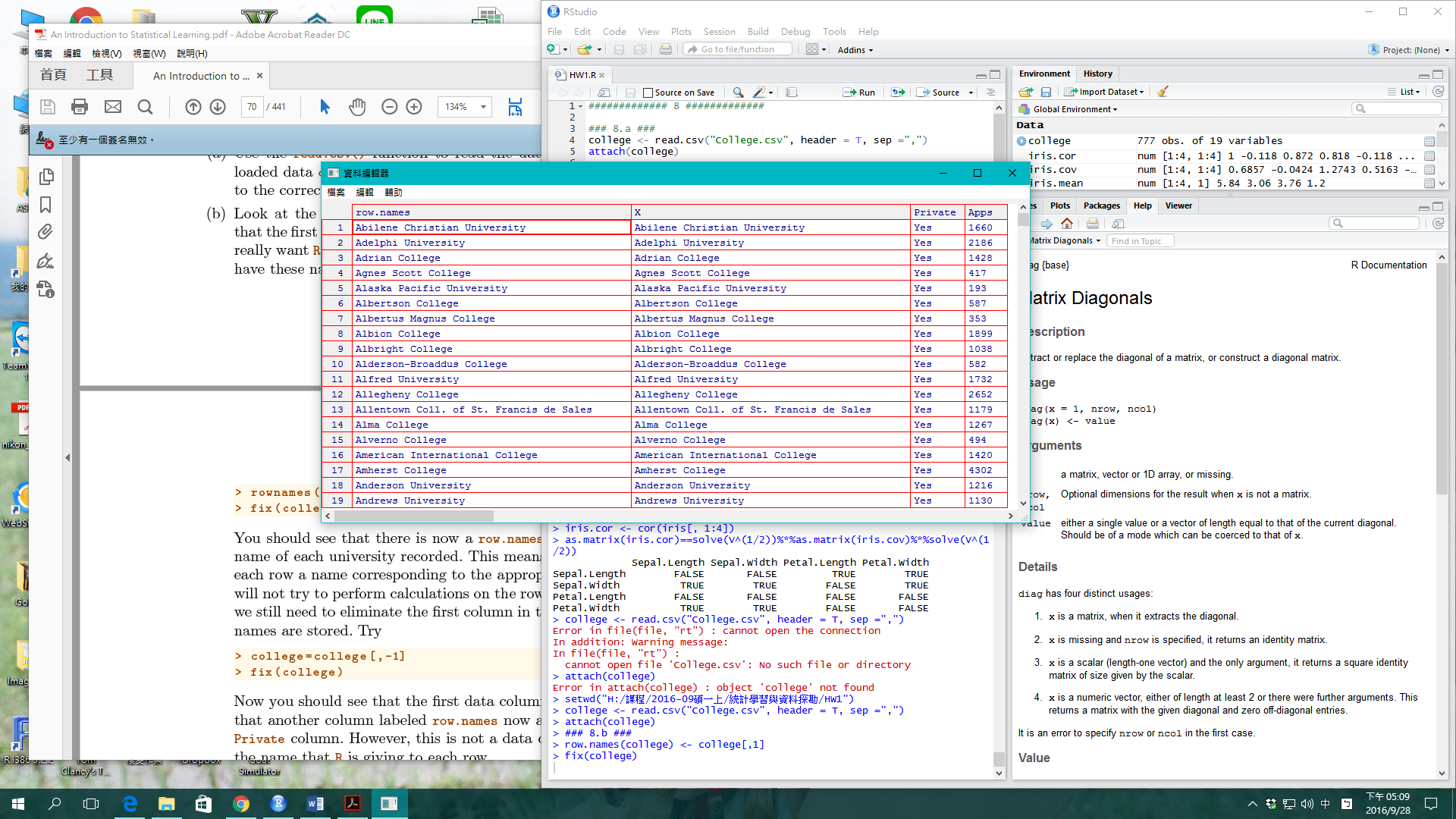
Statistical Learning and Data mining

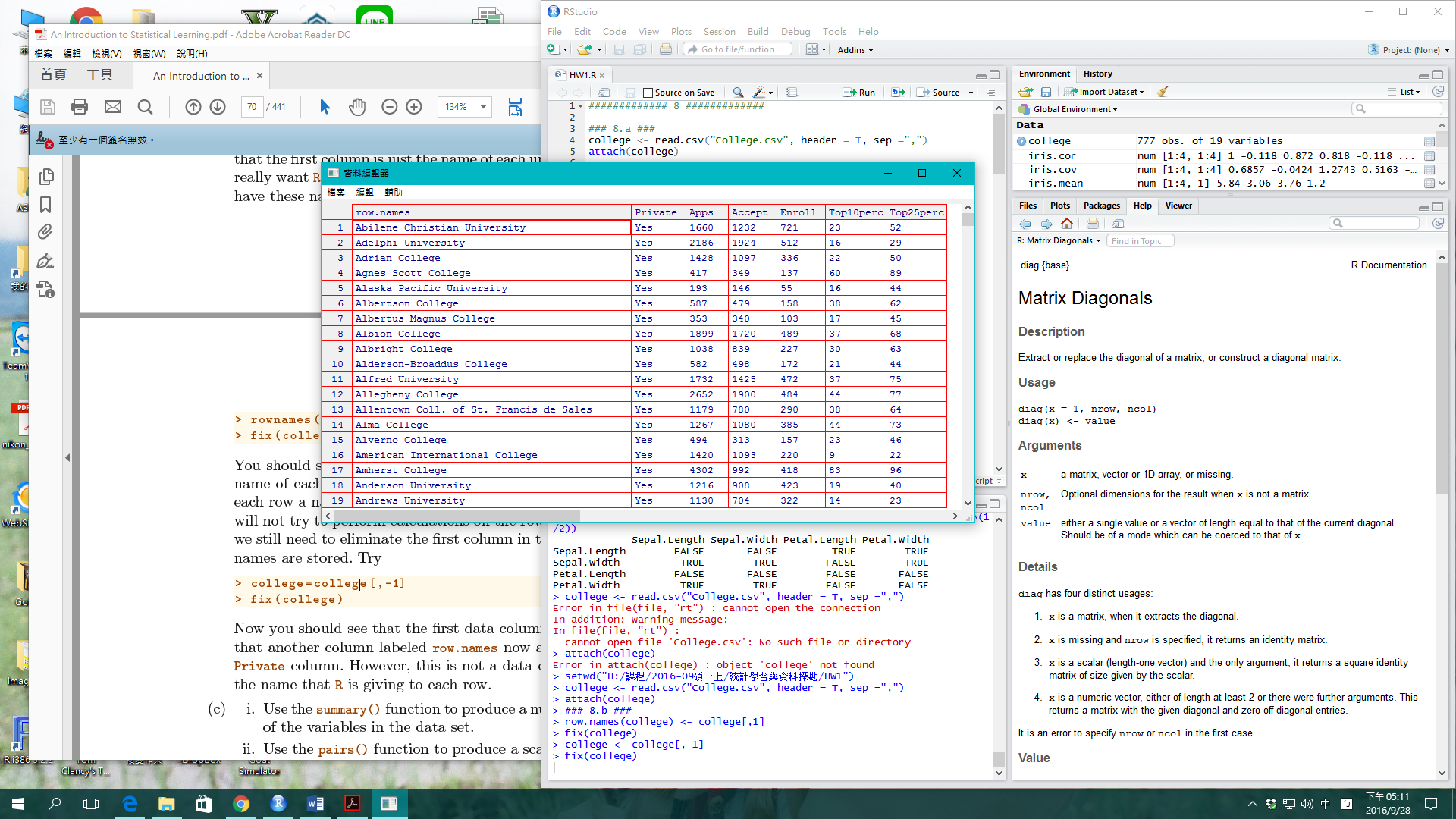
Homework 1

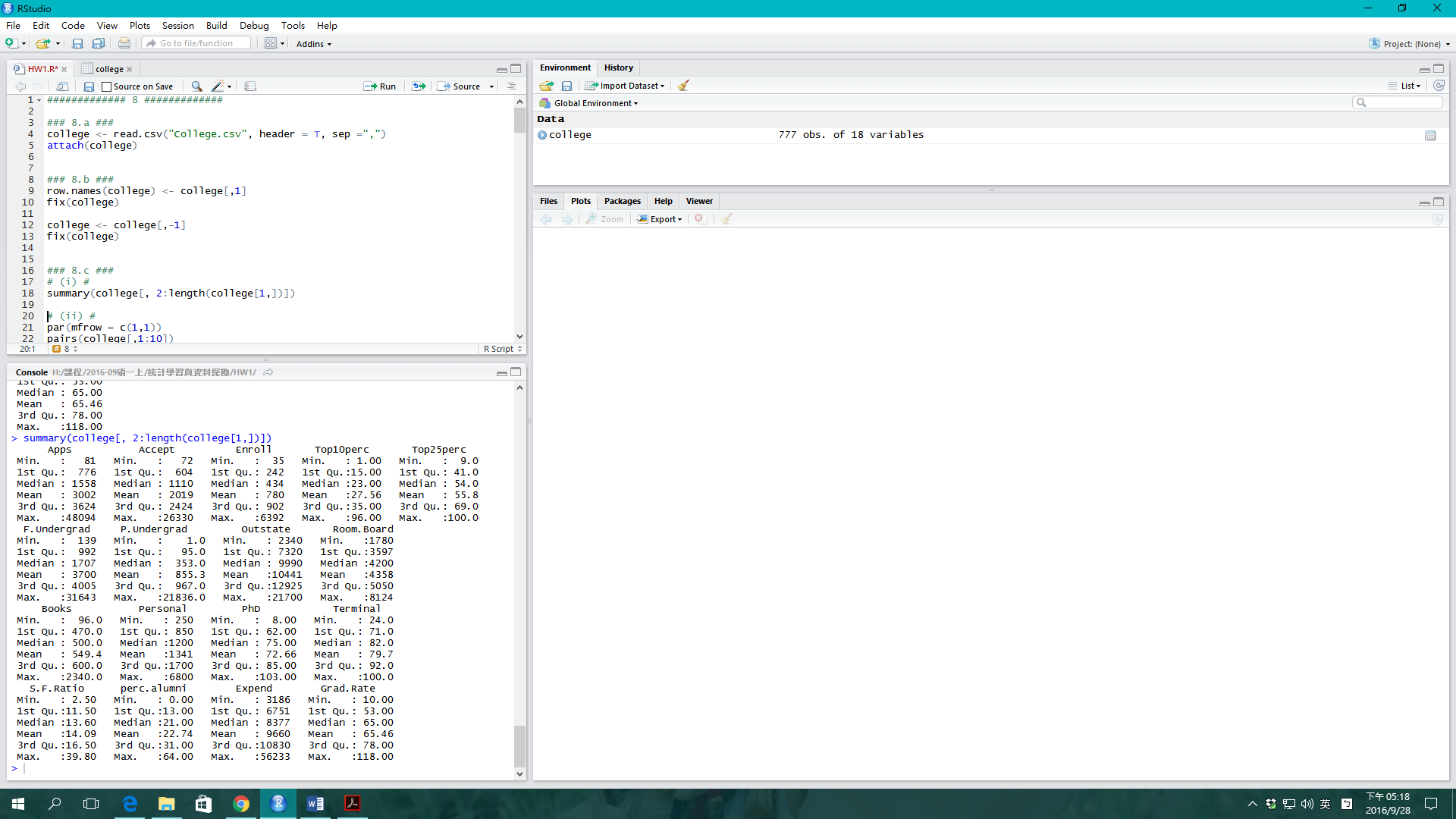
M052040003 鍾冠毅

8.a. See the appendix.

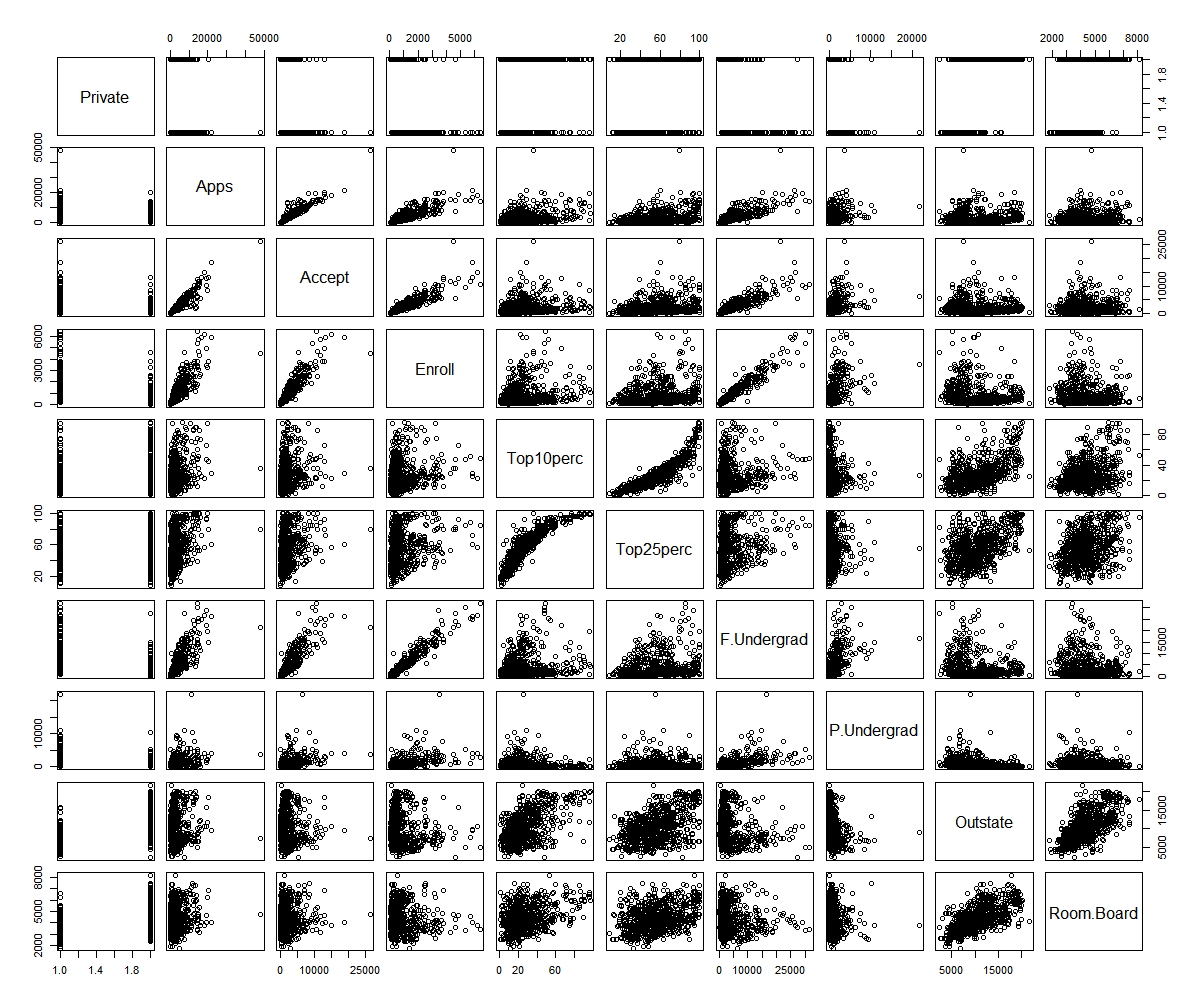
8.b.

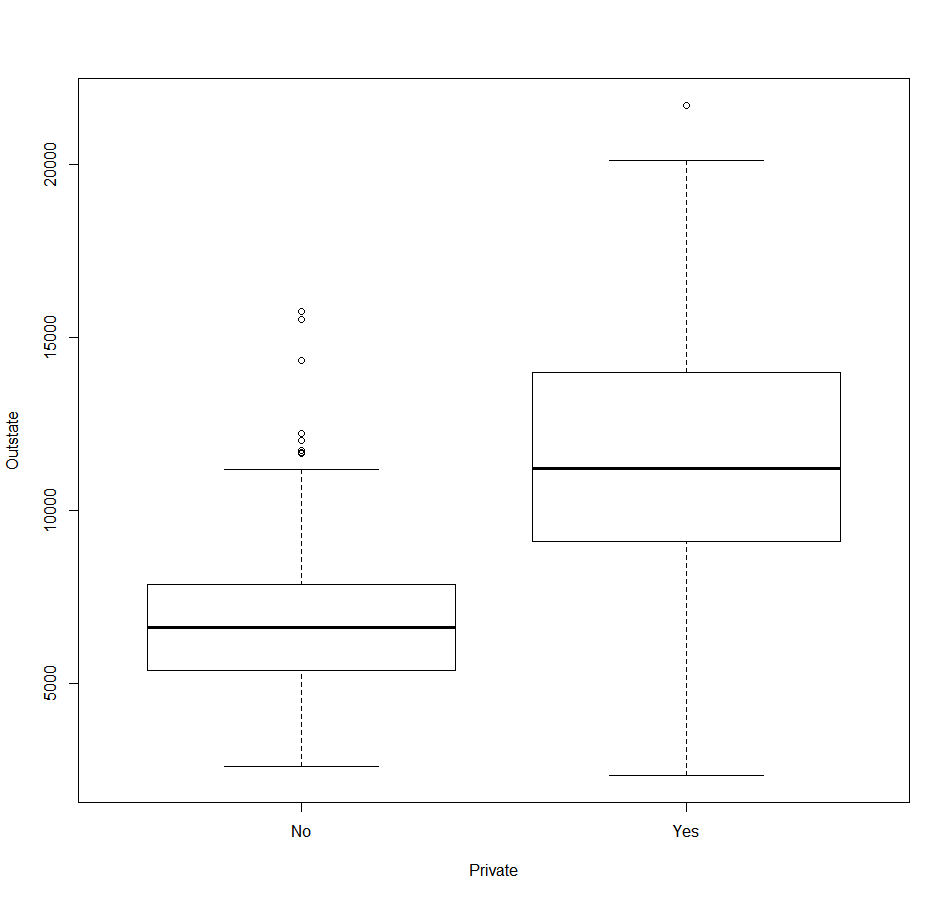




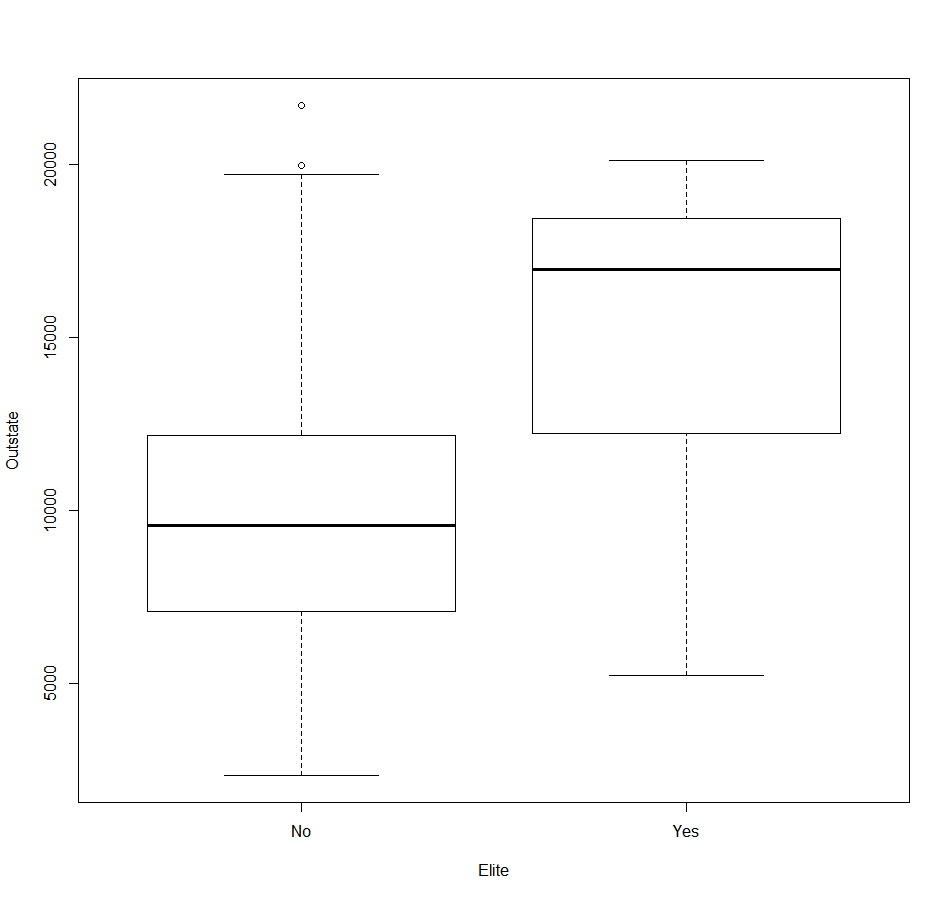
8.c. (i)

(ii)

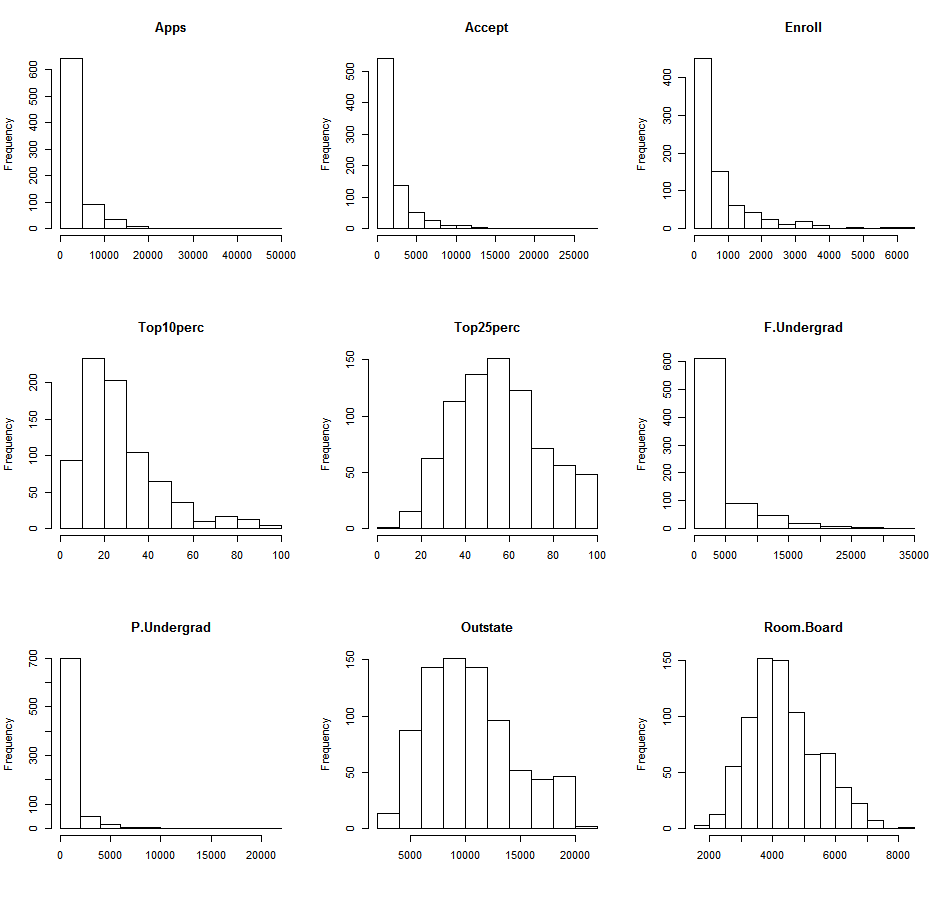


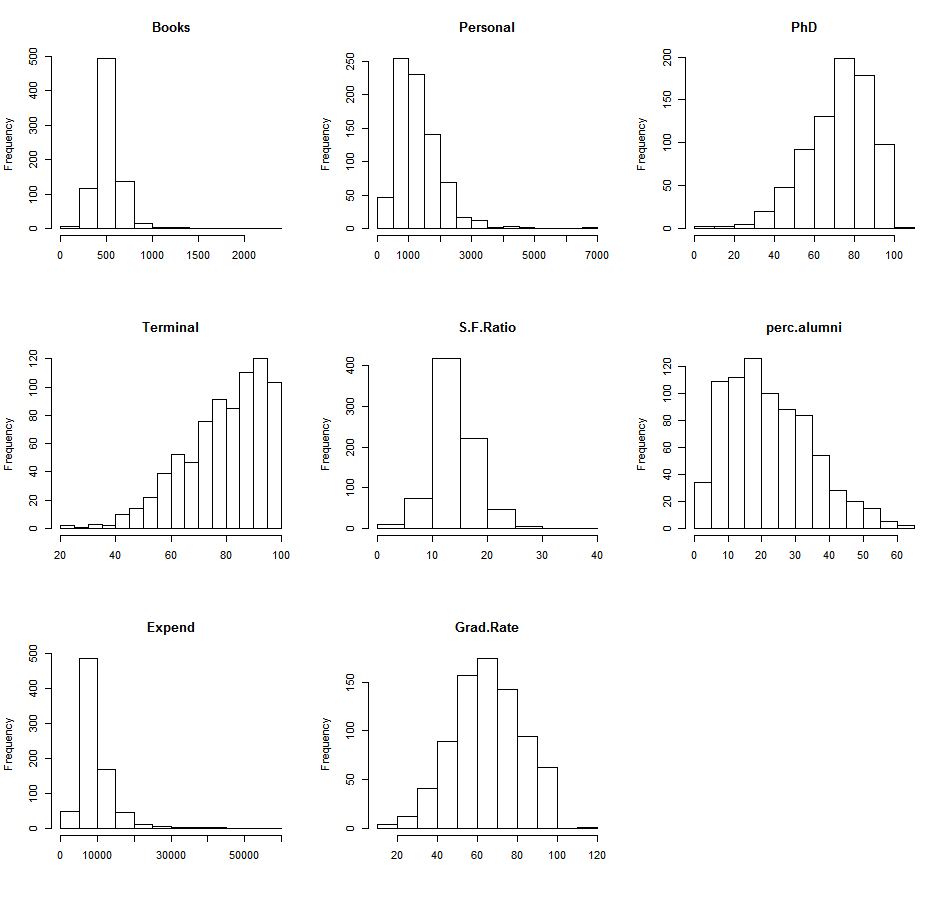
(iii)

(iv) Elite: 78, not Elite: 699



(v)





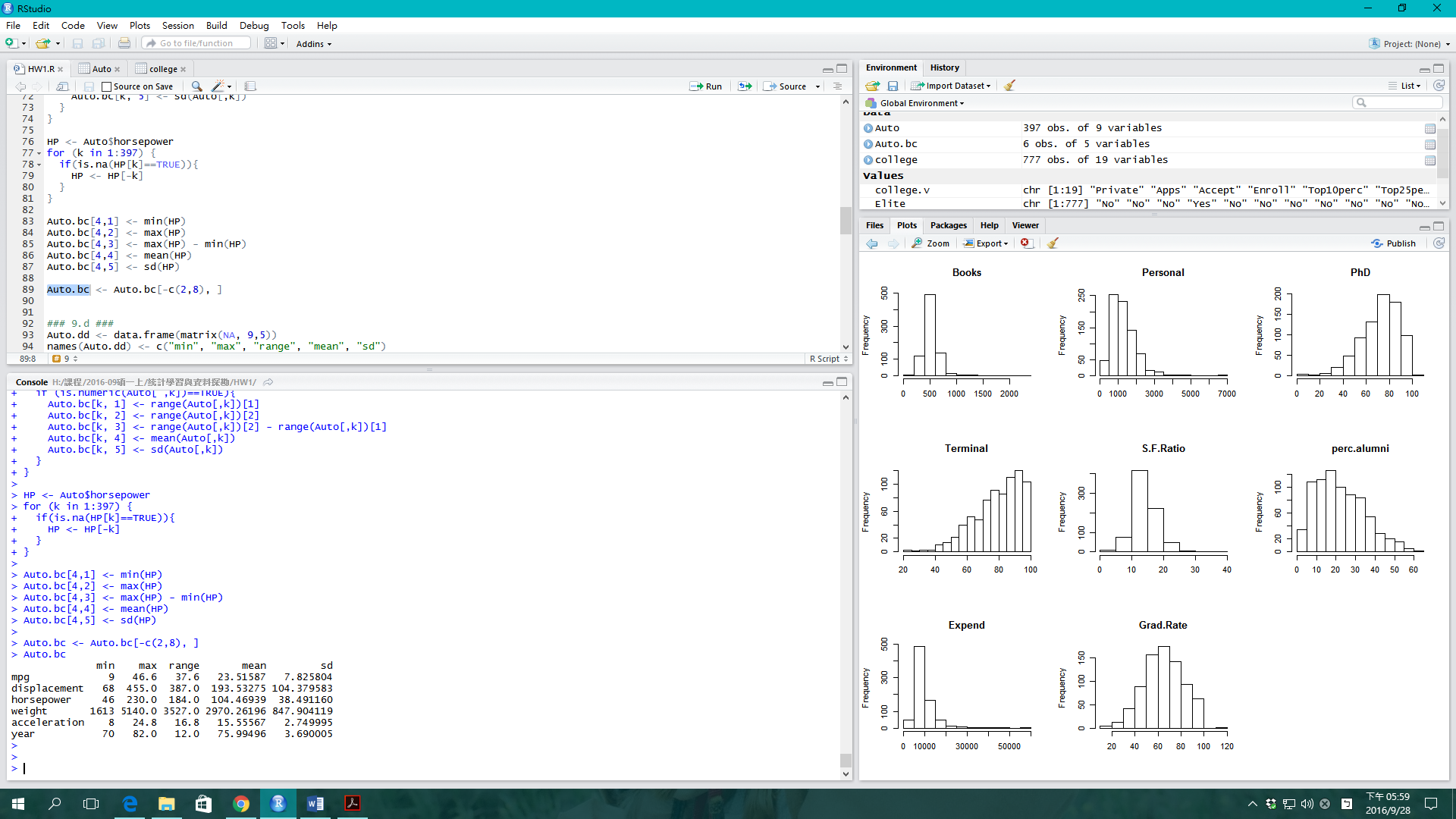
(vi) 從(ii)圖中可以發現，某幾對變數明顯呈正相關。進一步的研究除了可以討論兩兩變數之間的關係以外，亦可以使用群集分析，將相似數值表現的學校歸唯一類，並進一步探討之間的關係。由(iii)、(iv)之兩盒鬚圖可以發現，在Private與Elite兩個變數中，yes的outstate都明顯比no多，若要進一步了解是否有顯著差異，則可以使用假設檢定之結果作為判斷依據。

9.a. Quantitative: mpg, displacement, horsepower, weight, acceleration, year

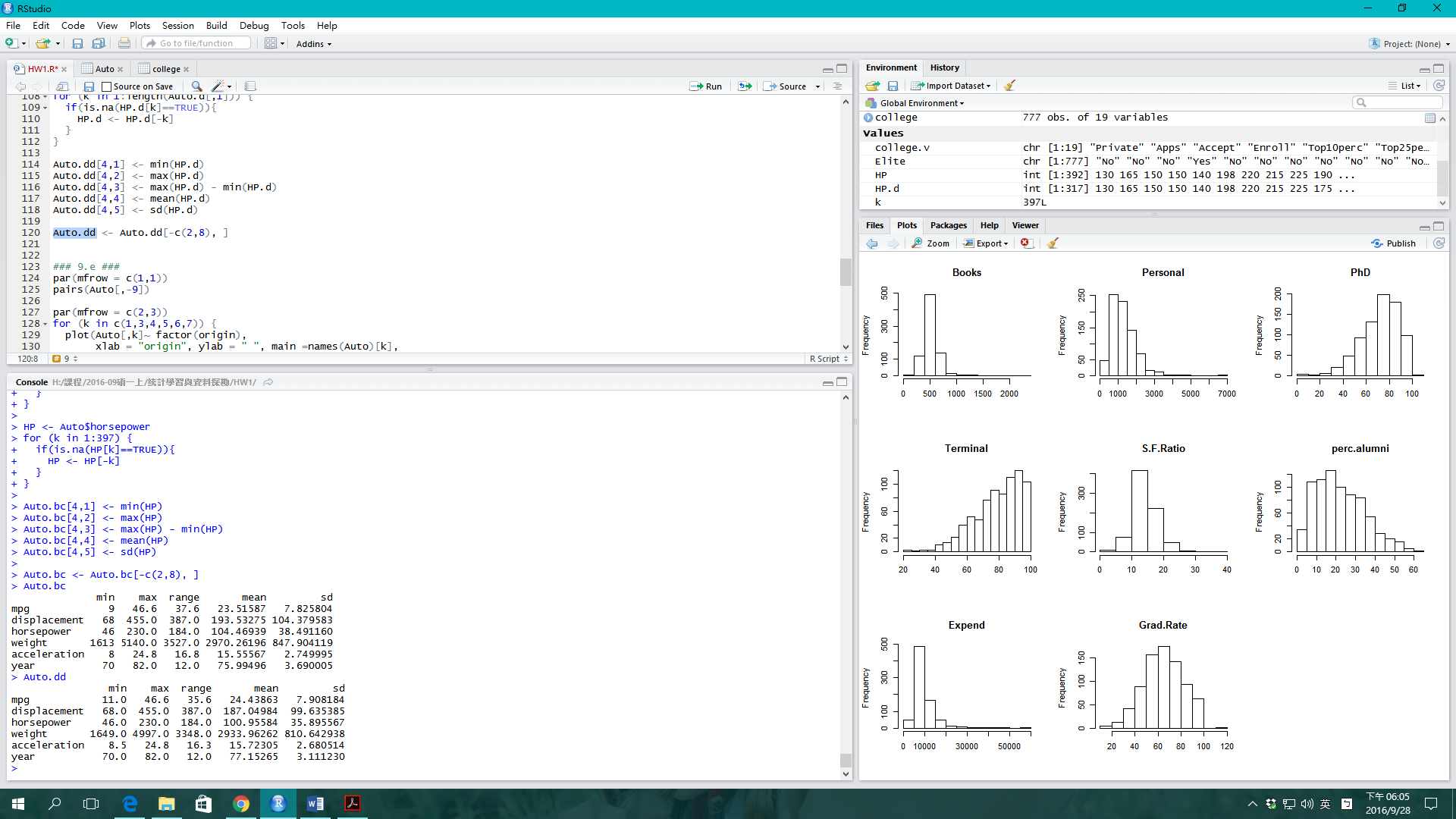
Qualitative: cylinders, origin, name

9.b.

9.c.

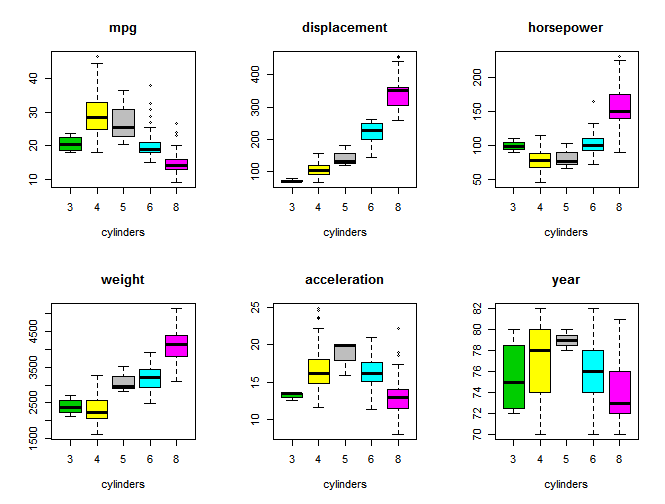


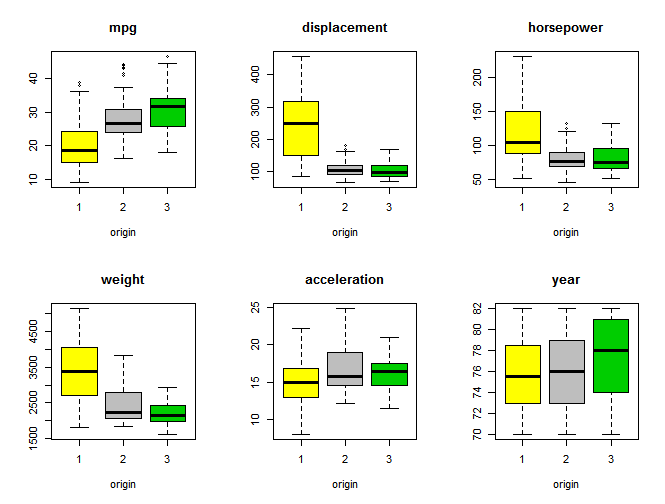
9.d.



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.

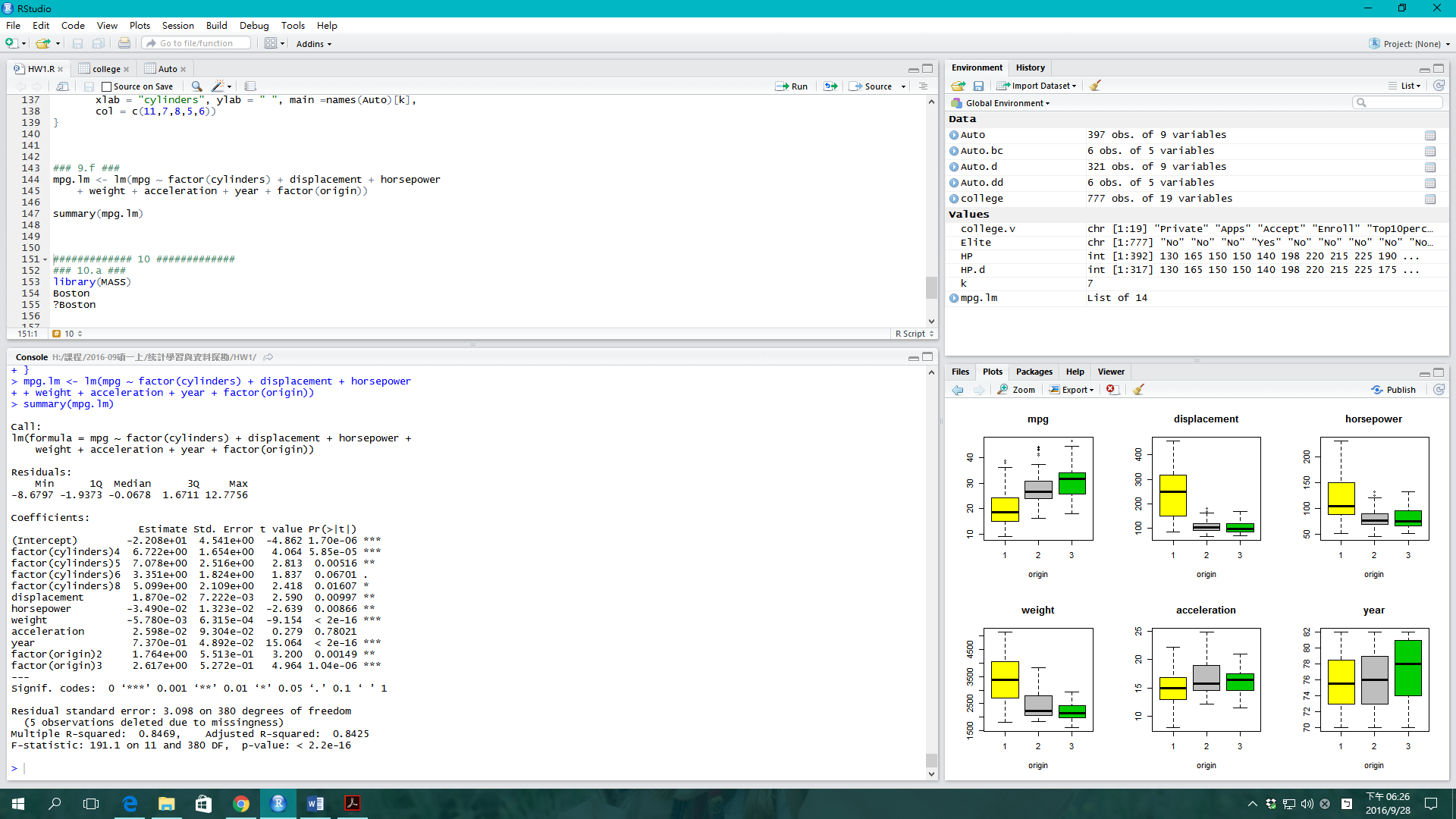




可以發現汽缸的數量影響其他述職的表現，以排氣量來說汽缸的數量越多，其值越高。由年份來看，五汽缸的年份較其他汽缸數之車輛新，可以推估五汽缸引擎較晚被發明與應用。另外，汽缸數越高並不代表加速越快，但是馬力可能有較好的表現。

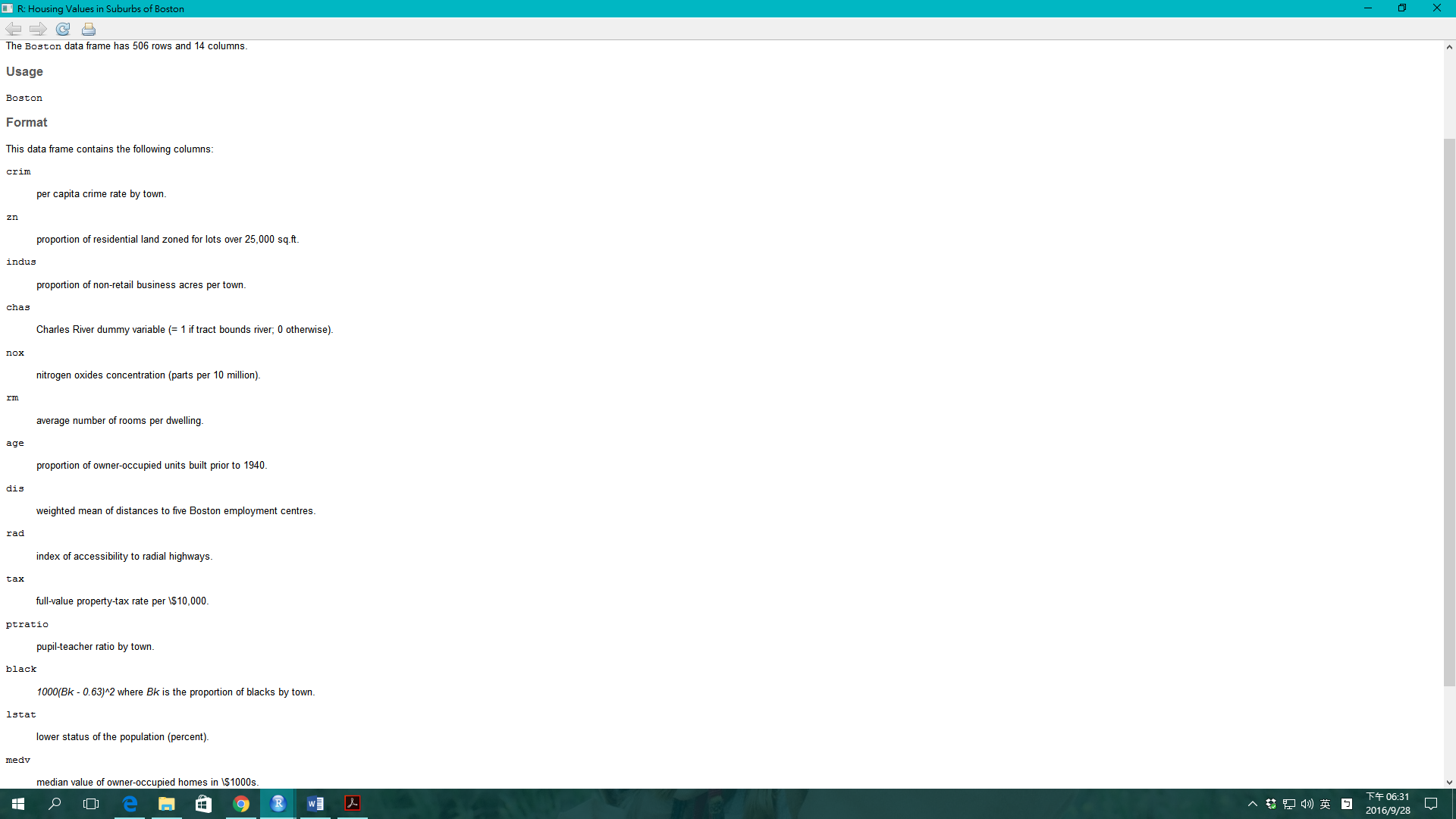
以來原來說，來源一的汽車在排氣量、馬力普遍有較好的表現，但是車身重量也比較高。

9.f. 預測模型可以使用線性迴歸預測。

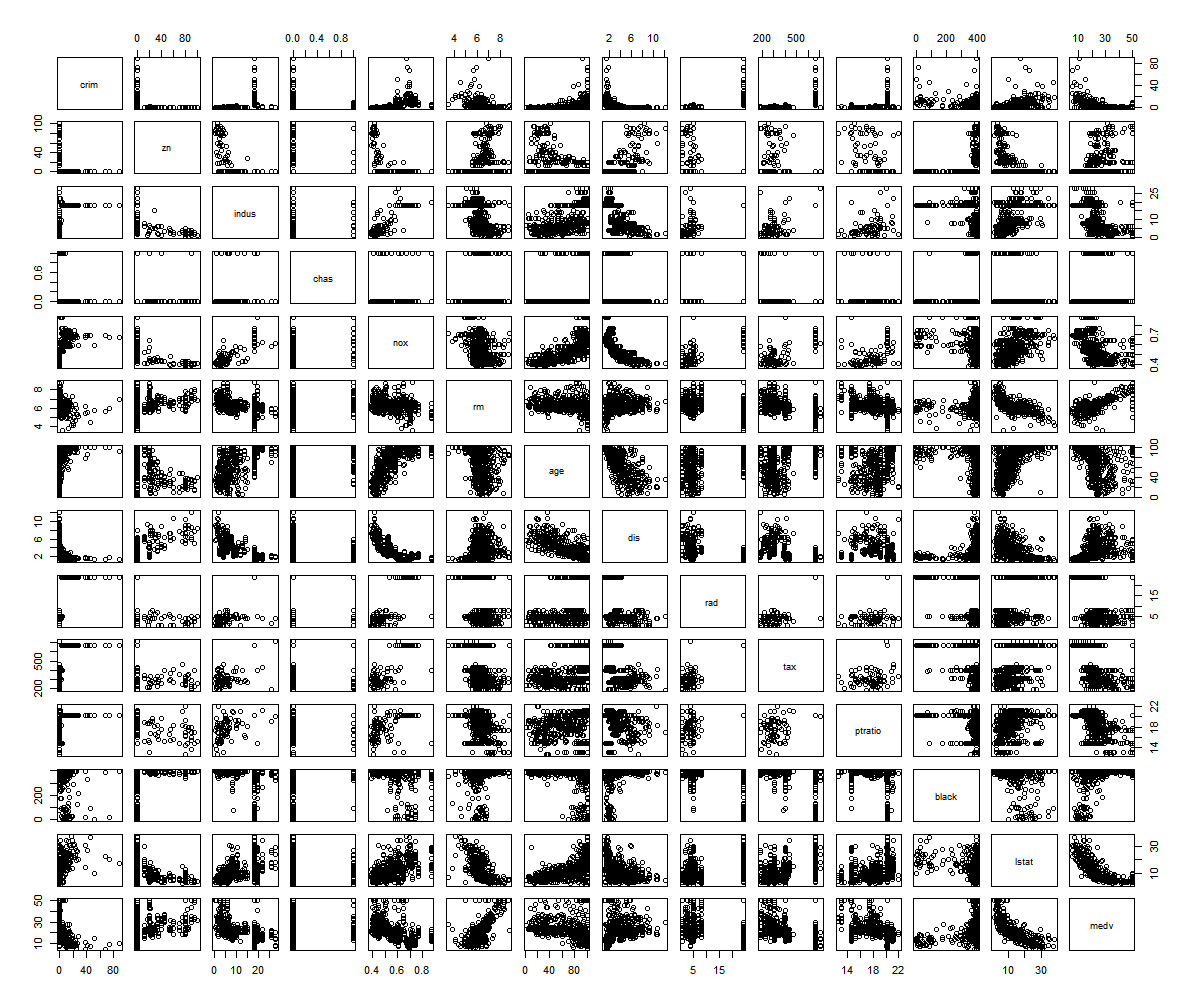


可以看到大部分的變數，對於mpg有顯著之影響，若要有更佳的預測模型，則可以選用其他廣義線性迴歸，並選用stepwise等方法挑選解釋變數。

10.a.



10.b.

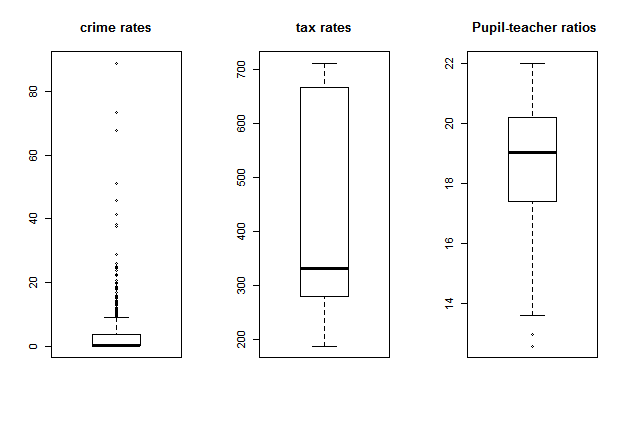


除了幾對變數有明顯的正相關、負相關之線性關係以外，age與lstat兩變數之分散圖可以發現，age越高lstat的值與範圍即越高，若要進行迴歸分析，則建議對lstat取log後，降低同age之lstat數值變異，避免R-square太小。另外以最右下角兩變數lstat與medv，其關係可能不僅是負相關，亦有可能有倒數關係或指數關係，可以另外以其他廣義線性迴歸分析該數據。

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.

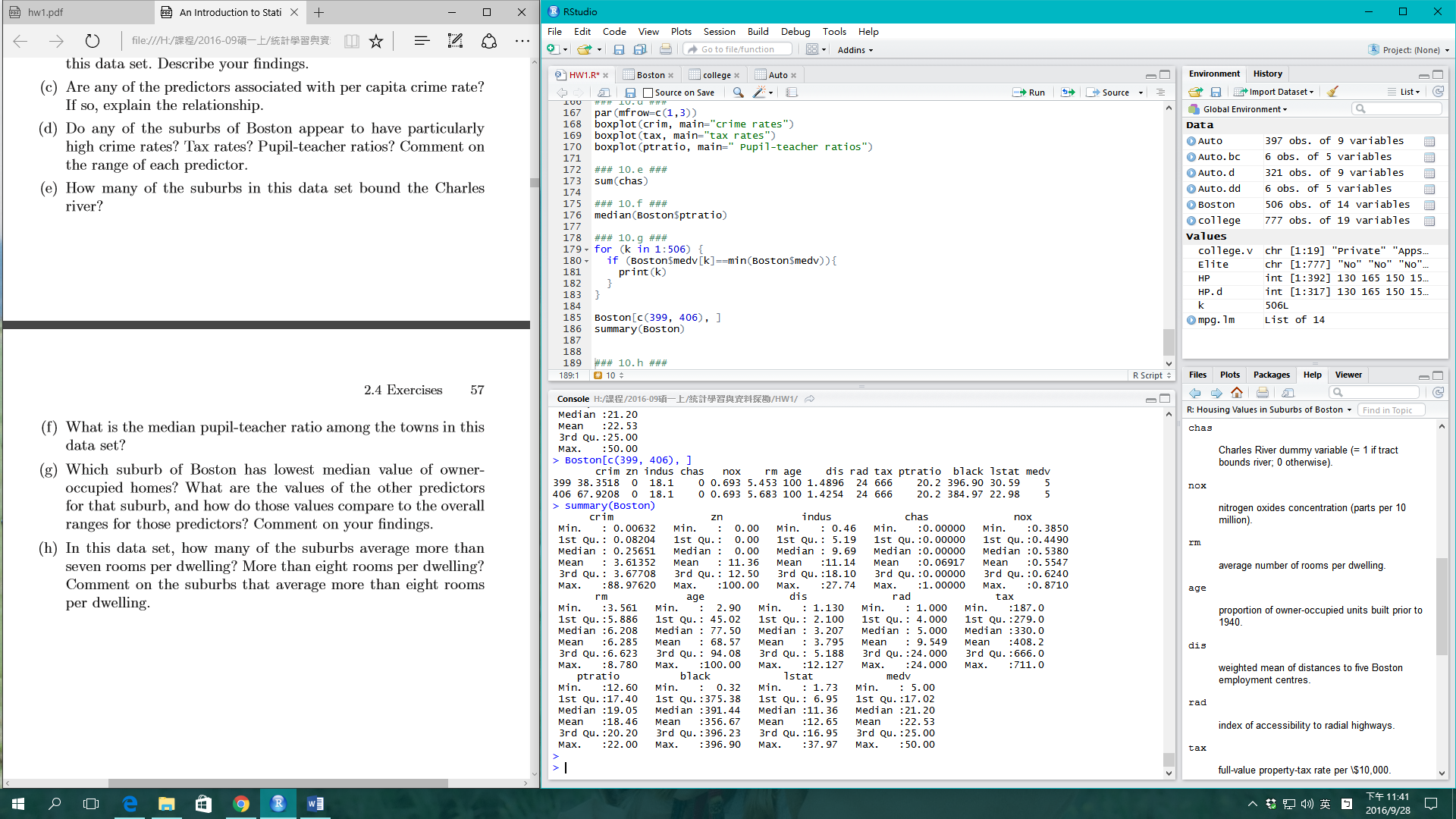


Crime rates 大多呈城市偏低，多在10以內，但是也不少城市有屬於離群值，遠高於10，最多有超過60的城鎮。稅率的部分約介於200至700，且無離群值。師生比的部分大多介於14至22之間，有兩個城鎮屬於偏低之離群值。

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.



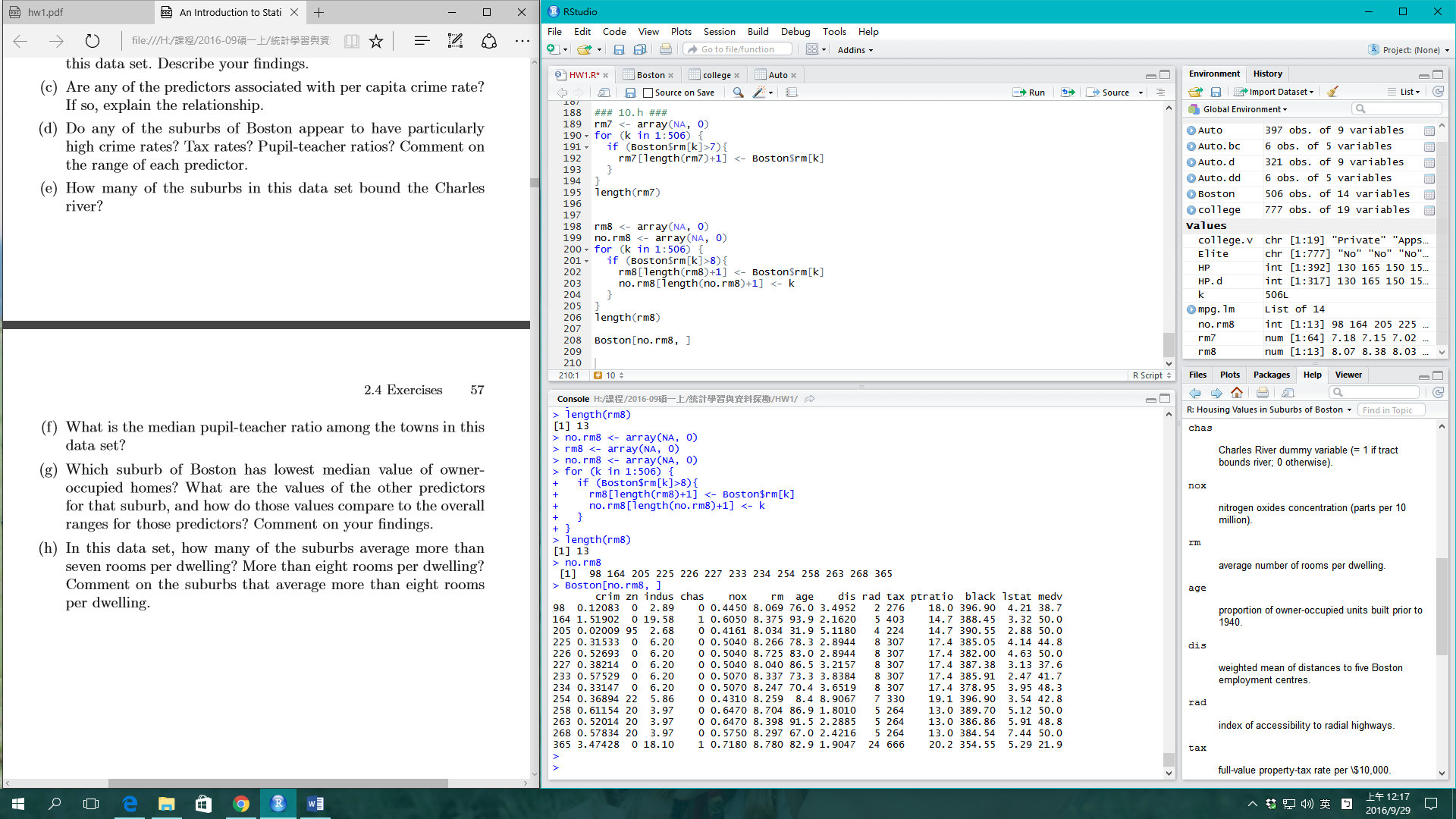
兩城鎮皆大於Q3：crim、indus、nox、age、rad、tax、ptratio、black、lstat

兩城鎮皆大於Q1：zn、chas、rm、medv

10.h.

>7: 64 suburbs

>8: 13 suburbs



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, ]