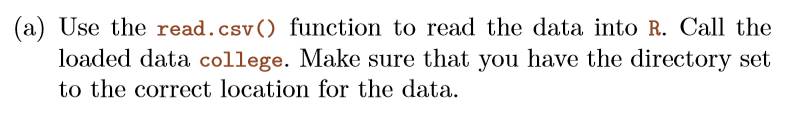
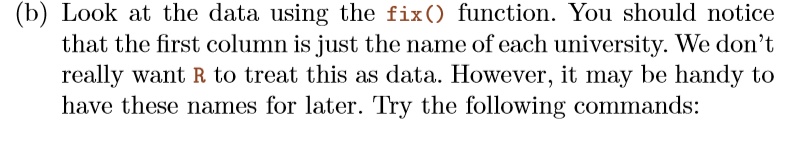
Hw1

M072040019 應數碩一統計組: 梅瀚中

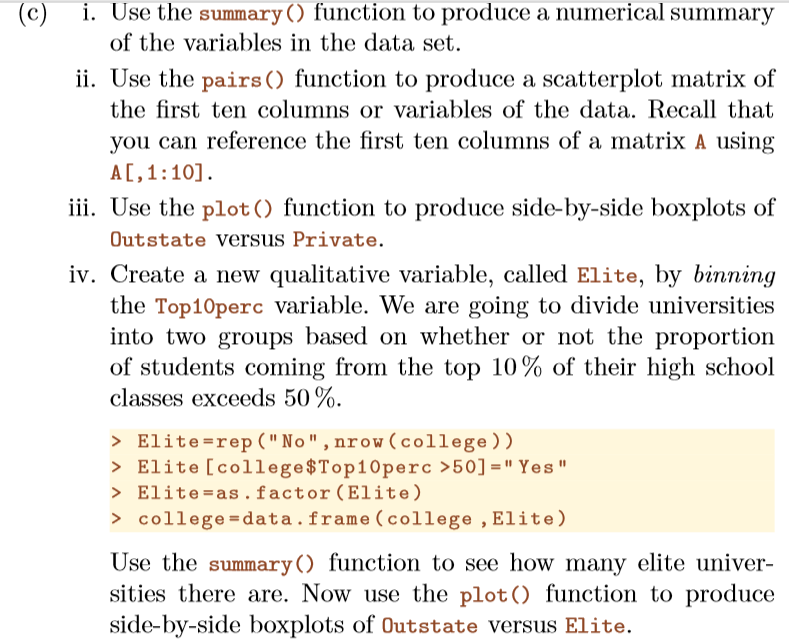


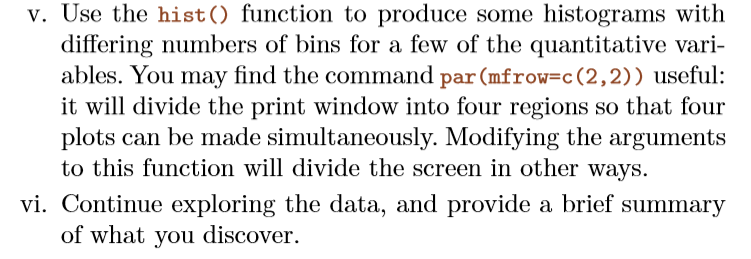
> College=read.csv("C:/Users/user.LAPTOP-N880RBTF/Desktop/統計/machine learning hw/美惠老師的/hw1/College.csv",header=T,sep=",")



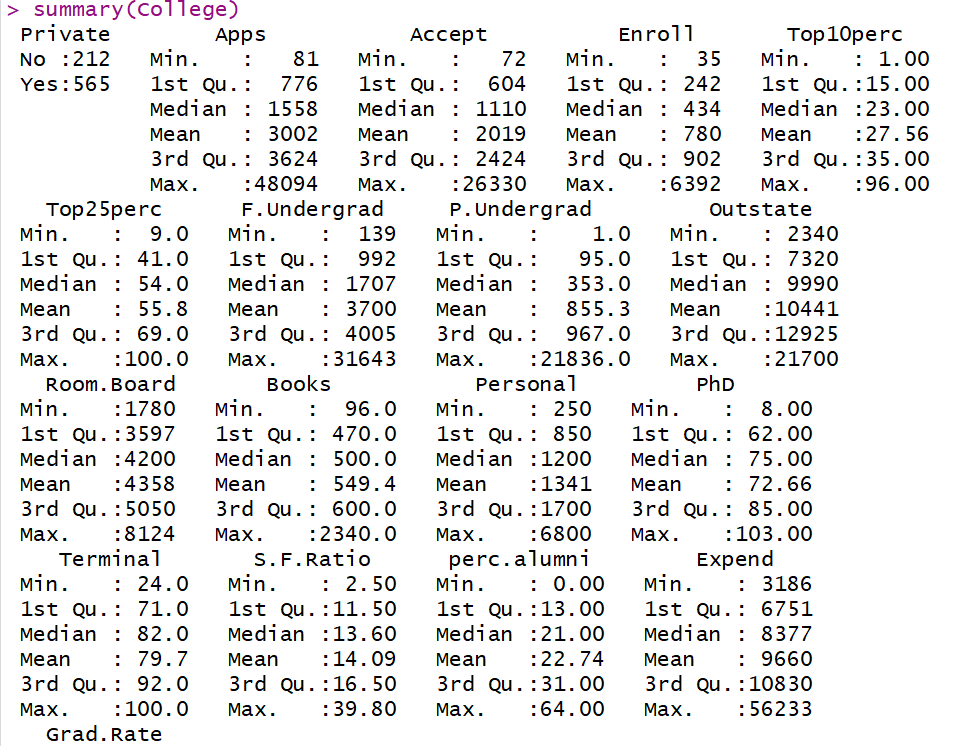
> College=College[ ,-1]

> fix(College)





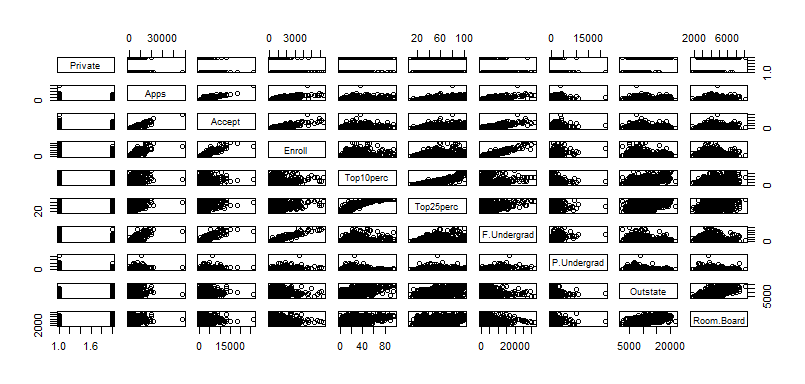
(I)



(II)

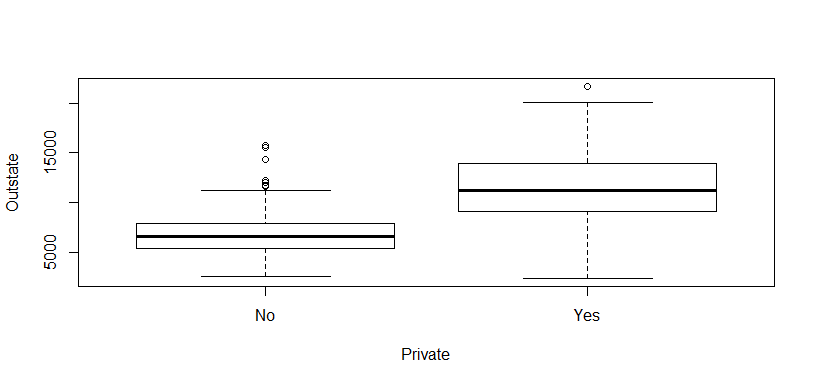
> A=pairs(College[,1:10])

> A



(III)

> plot(formula = Outstate~Private,data=College)



(iv) > Elite=rep("No",nrow(College))

> Elite[College$Top10perc >50]="Yes"

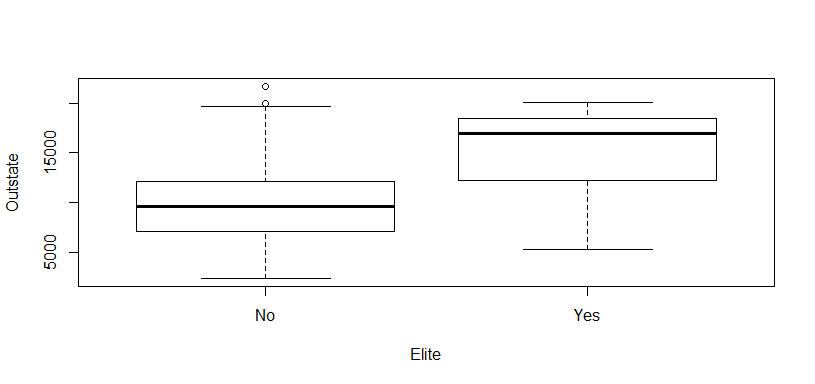
> Elite=as.factor(Elite)

> College=data.frame(College ,Elite)

> fix(College)

> summary(College)

> plot(formula = Outstate~Elite,data=College)



(v)

> attach(College)

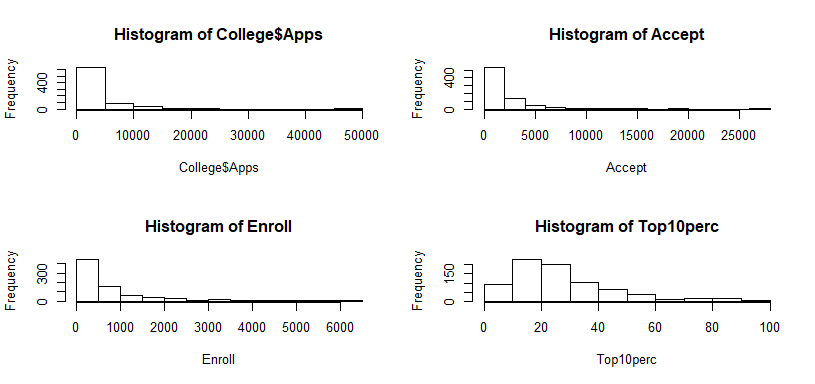
> par(mfrow=c(2,2))

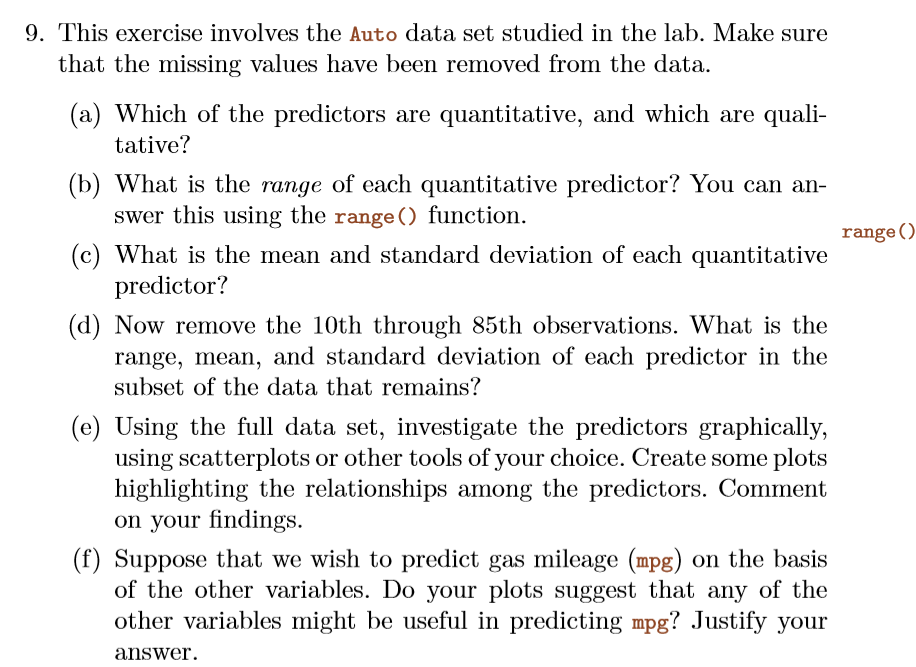
> hist(College$Apps)

> hist(Accept)

> hist(Enroll)

> hist(Top10perc)

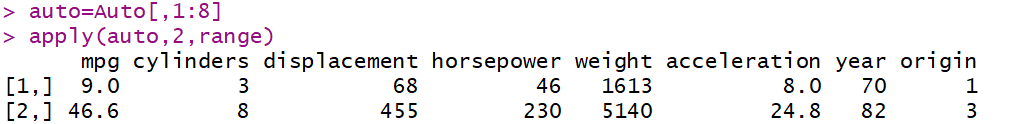




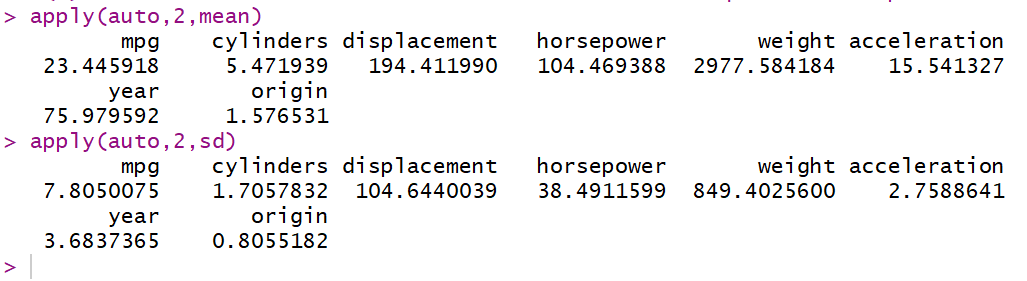
(a)

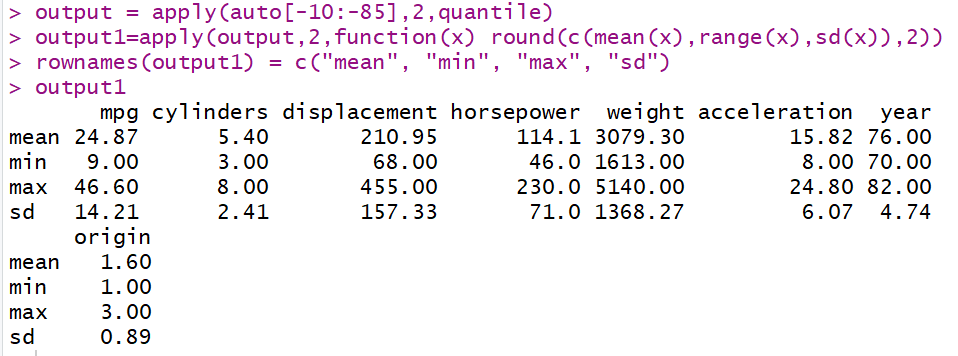
“mpg" ,"cylinders" , "displacement","horsepower" ,"weight","acceleration", "year" ,"origin" 都是數值資料 , "name"類別資料

(b)



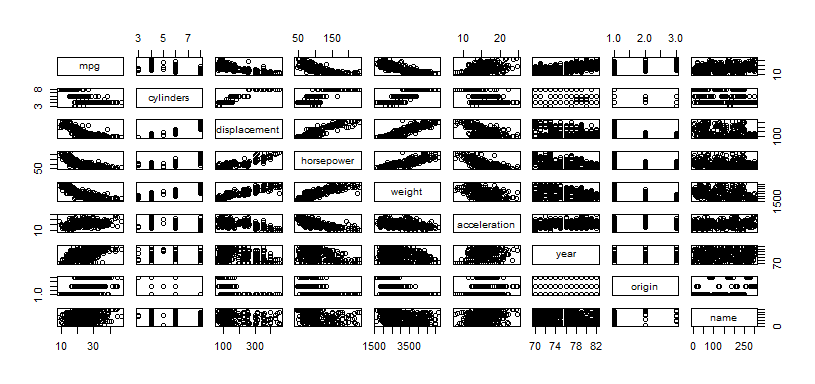
(c)



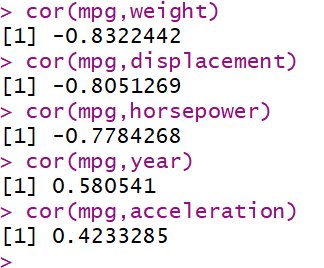
(d)

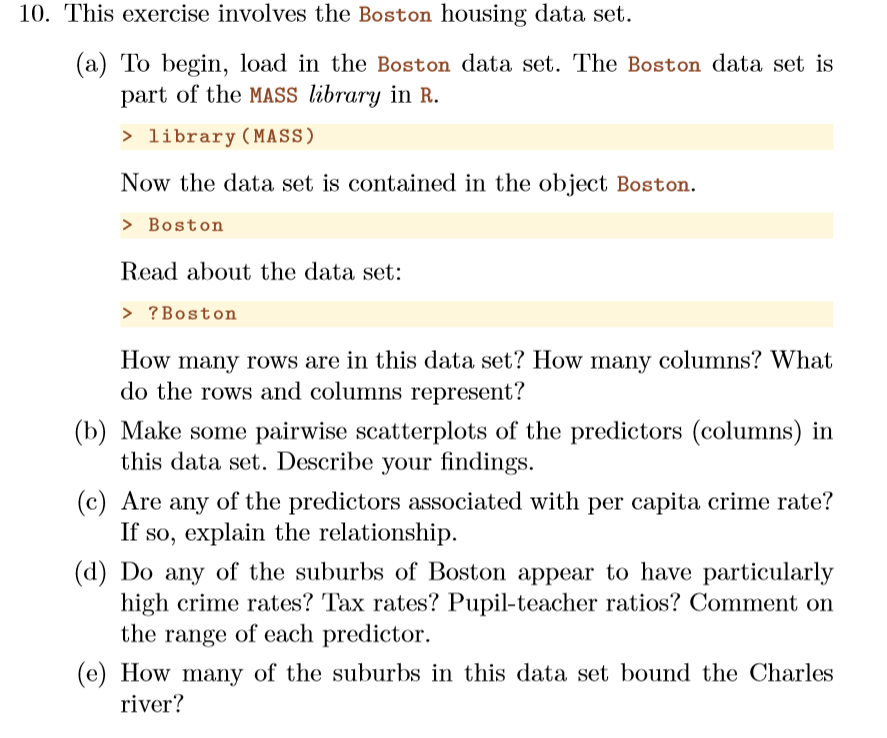
(e)看到某些變數兩兩之間存在正相關,和負相關 ex:msg變數和weight變數存在負相關

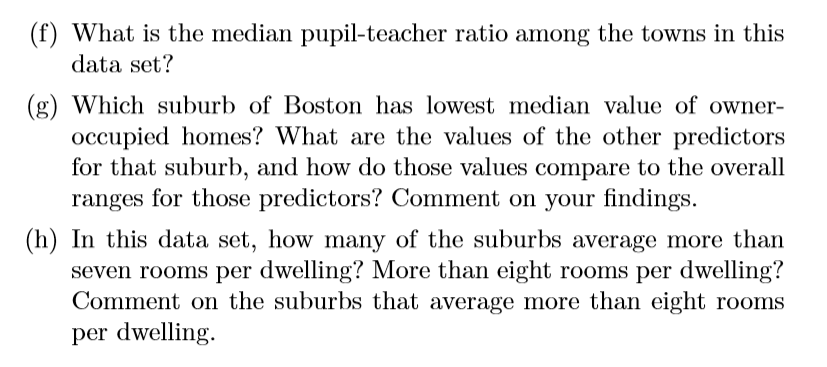
> pairs(Auto)



(f) 認同,displacement,horsepower,weight 與mpg似乎都有很高的負相關係數,year跟acceleration跟mpg有正的相關係數







(a) row:506,代表506個資料 column:14代表14個變數

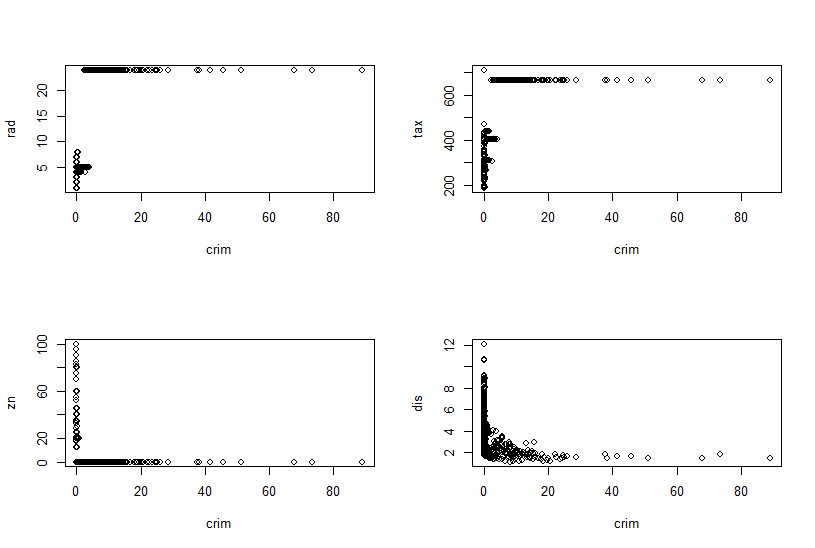
(b), (c)

> plot(crim,rad)

> plot(crim,tax)

> plot(crim,zn)

> plot(crim,dis)



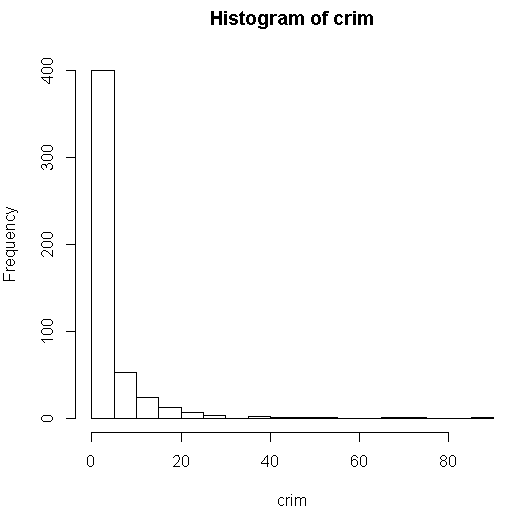
指數愈高的地方，犯罪率愈高。

財產稅愈高的地方，犯罪率愈高。

到波士頓五個就業中心的加權平均值越低的地方,犯罪率越高。

佔地面積超過25,000平方尺的住宅用地比例越低的地方,犯罪率越高。

(d) > hist(crim,breaks=20)



> length(crim[crim>20])

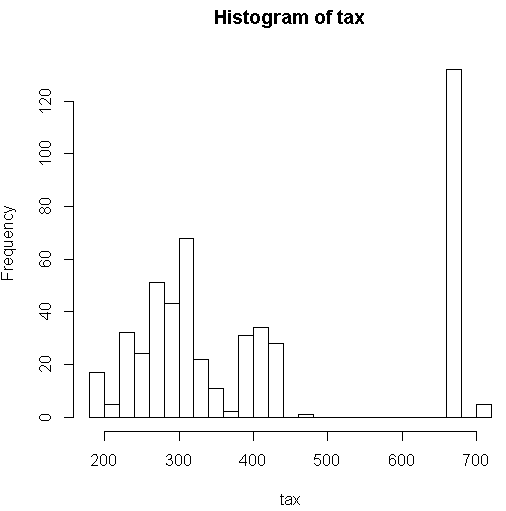
[1] 18

> max(crim)

[1] 88.9762

大部分的郊區犯罪率都低於20%，但是有18個郊區的犯罪率卻超過20%，犯罪率最高的郊區甚至高達88.98%。

> hist(tax,breaks=20)

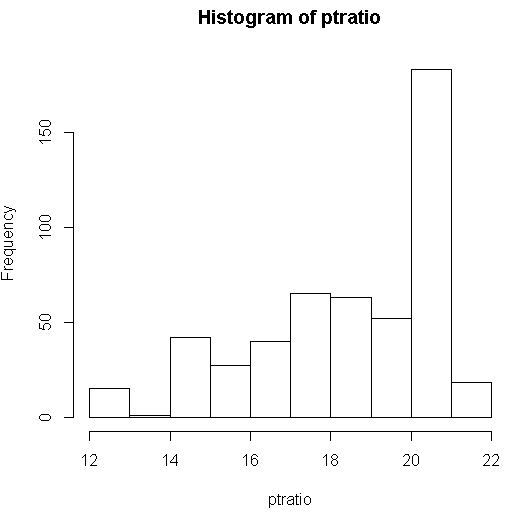


> length(tax[tax>=666])

[1] 137

財產稅大多位在500萬美元內，500-666之間沒有，呈現斷層，但財產稅大於等於666萬元的郊區數量為137個。

> hist(ptratio)



師生比的分佈有稍微向右，並沒有特別大的值出現。

(e)



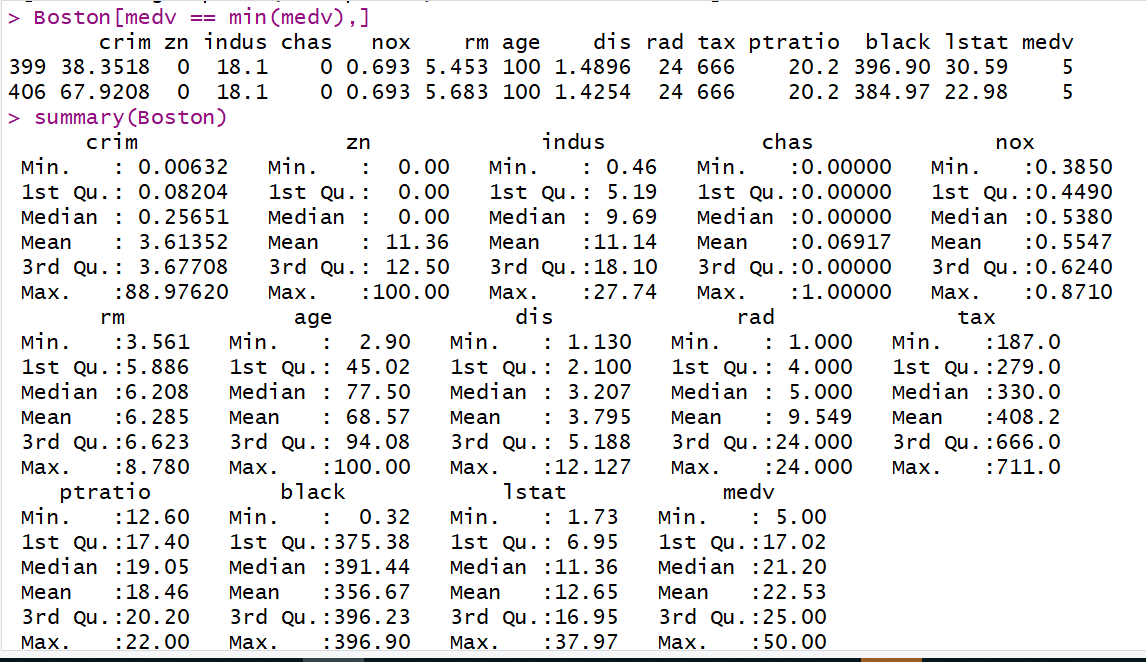
35個郊區。

(f)



師生比的中位數為19.05。

(g)



第399和第406個郊區。

相較於其他郊區 這兩個郊區的:

犯罪率偏高。

佔地面積超過25,000平方尺的住宅用地比例最小。

非零售業面積偏高。

都沒有在河川旁邊。

氮氧化物濃度偏高。

住宅房間數偏低。

1940年以前建造的自住單位比例很高。

距離就業中心很近。

可達徑向高速公路的指標最高。

財產稅偏高。

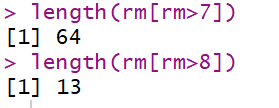
城鎮的師生比例偏高。

黑人比例偏高。

較低的人口狀況（百分比）偏高。

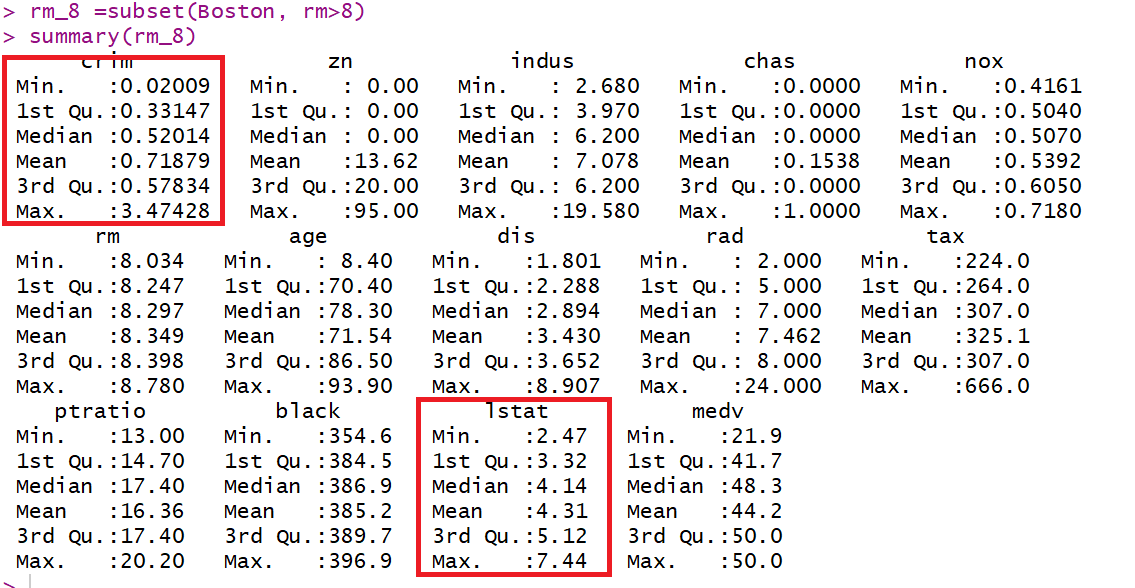
結論:這兩個郊區相較於Boston其他郊區 可被視為不受歡迎的郊區。

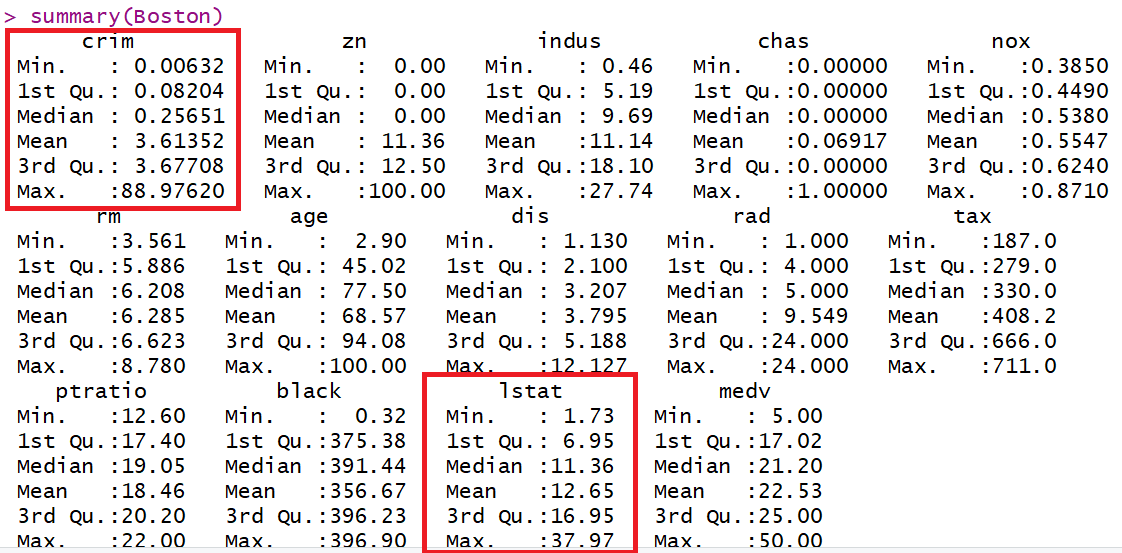
(h)



平均房間數大於七間的有64個郊區。

平均房間數大於八間的有13個郊區。





經比較得知，平均房間數大於八間的這13個郊區，其犯罪率較低,低地位人口較少。

Code:

##8

College=read.csv("C:/Users/user.LAPTOP-N880RBTF/Desktop/統計/machine learning hw/美惠老師的/hw1/College.csv",header=T,sep=",")

College

rownames(College)=College[,1]

fix(College)

College=College[ ,-1]

fix(College)

summary(College)

A=pairs(College[,1:10])

A

plot(formula = Outstate~Private,data=College)

Elite=rep("No",nrow(College))

Elite[College$Top10perc >50]="Yes"

Elite=as.factor(Elite)

College=data.frame(College ,Elite)

fix(College)

summary(College)

plot(formula = Outstate~Elite,data=College)

College

attach(College)

par(mfrow=c(2,2))

hist(College$Apps)

hist(Accept)

hist(Enroll)

hist(Top10perc)

hist(Top25perc)

hist(F.Undergrad)

hist(P.Undergrad)

hist(PhD)

hist(Terminal)

hist(S.F.Ratio)

##9

library(ISLR)

dim(Auto)

data(Auto)

sum(is.na(Auto))

names(Auto)

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

str(Auto)

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

auto=Auto[,1:8]

apply(auto,2,range)

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

apply(auto,2,mean)

apply(auto,2,sd)

#(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?

output = apply(auto[-10:-85],2,quantile)

output1=apply(output,2,function(x) round(c(mean(x),range(x),sd(x)),2))

rownames(output1) = c("mean", "min", "max", "sd")

output1

#(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.

pairs(Auto)

plot(Auto)

#(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.

cor(mpg,weight)

cor(mpg,displacement)

cor(mpg,horsepower)

cor(mpg,year)

cor(mpg,acceleration)

###10

library(MASS)

data(Boston)

#(A)How many rows are in this data set? How many columns? What do the rows and columns represent?

dim(Boston)

#(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your ﬁndings.

names(Boston)

attach(Boston)

plot(crim,rad)

plot(crim,tax)

plot(crim,zn)

plot(crim,dis)

?Boston

#(d)

hist(crim,breaks=20)

length(crim[crim>20])

max(crim)

hist(tax,breaks=20)

length(tax[tax>=666])

hist(ptratio)

#(e)

length(chas[chas>0])

#(f)

median(Boston$ptratio)

#(g)

Boston[medv == min(medv),]

summary(Boston)

?Boston

#(H)

length(rm[rm>7])

length(rm[rm>8])

rm\_8 =subset(Boston, rm>8)

summary(rm\_8)

summary(Boston)