**Chapter 2: Statistical Learning**

**1(a)**

Better. As the level of flexibility increases, the curve fit the observed data more closely and a better fit would be obtained with the large sample size.

**1(b)**

Worse. The flexible statistical learning method would overfit the data with extremely large number of predictors p and small number of observations n.

**1(c)**

Better. If the relationship between the predictors and response is highly non-linear, an inflexible statistical learning method would lead to high training MSE, while a flexible method may fit the data better.

**1(d)**

Worse. A flexible statistical learning method would fit the noise in the error terms and increase variance.

**2(a)**

Regression problem. We are interested in inference, say, how the CEO salary is affected as profit, number of employees and industry change.

n - 500 firms in the US

p - profit, number of employees, industry

**2(b)**

Classification problem. We are interested in prediction. To predict a new product and will be a success or a failure.

n - 20 similar products previously launched

p - price charged, marketing budget, comp. price, ten other variables

**2(c)**

Regression problem. We are interested in inference, that is to say, we want to know how the % change in the US dollar in relation to the weekly changes in the world stock markets.

n - 52 weeks of 2012 weekly data

p - % change in US market, % change in British market, % change in German market

**3(a)**

> set.seed(1)

> x1=rnorm(100,1,1)

> x2=rnorm(100)

> x=cbind(x1,x2)

> eps=rnorm(100)

> y=x1^2-2\*x2^2+eps

> train=sample(1:nrow(x),nrow(x)/2)

> test=(-train)

> y.test=y[test]

> y.train=y[train]

> fit2=list()

> fit2.pred=list()

> bias=rep(NA,10)

> variance=rep(NA,10)

> irreducible=rep(NA,10)

> training\_mse=rep(NA,10)

> test\_mse=rep(NA,10)

> for (i in 1:10) {

+ fit2[[i]]=lm(y~poly(x1,i)+poly(x2,i),data=data.frame(x),subset = train)

+ fit2.pred[[i]]=predict(fit2[[i]],data.frame(x[test,]))

+ training\_mse[i]=mean((fit2[[i]]$fitted.values-y.train)^2)

+ test\_mse[i]=mean((fit2.pred[[i]]-y.test)^2)

+ bias[i]=mean((fit2[[i]]$fitted.values-y\_true[train])^2)

+ irreducible[i]=mean((eps-mean(eps))^2)

+ variance[i]=test\_mse[i]-bias[i]-irreducible[i]

+ }

> plot(c(1:10),training\_mse,type="l",col="red",xlim = c(0,15),xlab = "Degree of Freedom",ylab = "")

> axis(1,at=c(1:15),labels = c(1:15))

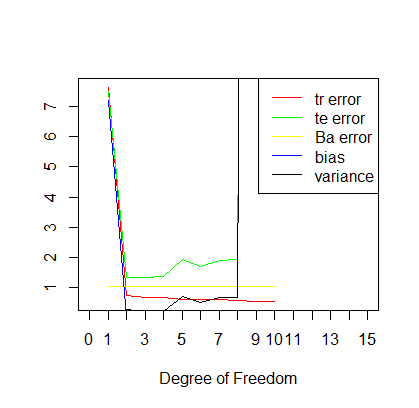
> lines(c(1:10),test\_mse,type="l",col="green")

> lines(c(1:10),irreducible,type="l",col="yellow")

> lines(c(1:10),bias,type="l",col="blue")

> lines(c(1:10),variance,type="l",col="black")

> legend("topright",c("tr error","te error","Ba error","bias","variance"),lty=c(1,1,1,1,1),col = c("red","green","yellow","blue","black"),text.width = 3)



**4(a)**

1 O2O优惠券使用预测

以优惠券盘活老用户或吸引新客户进店消费是O2O的一种重要营销方式。然而随机投放的优惠券对多数用户造成无意义的干扰。对商家而言，滥发的优惠券可能降低品牌声誉，同时难以估算营销成本。个性化投放是提高优惠券核销率的重要技术，它可以让具有一定偏好的消费者得到真正的实惠，同时赋予商家更强的营销能力。

现有O2O场景相关的丰富数据，希望通过分析建模，精准预测用户是否会在规定时间内使用相应优惠券。

2 市民出行选乘公交预测

基于海量公交数据记录，希望挖掘市民在公共交通中的行为模式。以市民出行公交线路选乘预测为方向，期望通过分析广东省部分公交线路的历史公交卡交易数据，挖掘固定人群在公共交通中的行为模式，分析推测乘客的出行习惯和偏好，从而建立模型预测人们在未来一周内将会搭乘哪些公交线路，为广大乘客提供信息对称、安全舒适的出行环境，用数据引领未来城市智慧出行。

3待测微生物种类判别

DNA是多数生物的遗传物质，DNA上的碱基（A，T，C和G）就储藏了遗传信息，不同物种的DNA序列在序列长度和碱基组成上差异显著。所以我们能够通过DNA序列的比较分析，来判断DNA序列是来自哪些物种。由于测序技术限制，我们只能得到一定长度的DNA序列片段。通过DNA序列片段与已知的微生物DNA序列进行比较，可以确定DNA片段的来源微生物，进而确定待测微生物种类。

期望在相关数据基础上，建立分析方法，在计算资源消耗尽量小的情况下，尽可能快地给出准确的结果，以满足临床诊断需求。

**4(b)**

1 农产品价格预测分析

农产品价格受市场影响的程度特别大,特别是受农产品的供求关系影响较大,同时价格本身又受自然条件、社会和经济条件的影响,特别是国际市场的影响。从价格本身来看,受供求、季节等发生波动,受外界各种影响比较多,这就造成了价格预测的困难。但从长期看,农产品价格随着时间的推移仍然呈现一定规律性。价格预测是大数据的精华所在，通过大量的历史数据分析，预测未来的价格走势，为决策者提供更有力的数据支持。

希望通过分析价格历史数据，对要求预测的农产品接下来固定时间的价格进行预测。并尽可能多的使用与价格有影响的其他数据以提高预测的准确率。

2 基于多源数据的青藏高原湖泊面积预测

全球气候变化对青藏高原的湖泊水储量有很大影响，因此精确的估计青藏高原湖泊面积变化对于研究气候变化变得很重要。海量多源异构数据和大数据处理与挖掘技术给湖泊面积变化研究带来新的解决思路；如何通过多源数据对青藏高原的湖泊面积进行预测，将大数据技术应用到全球气候变化研究中来成为一项新的挑战。

希望通过研究青藏高原湖泊面积变化的多种影响因素，构建青藏高原湖泊面积预测模型。

3 学生成绩排名预测

学生的校园行为数据，可以挖掘用户作息规律、兴趣爱好等，精准地预测学生之间的相对排名。通过对这些日常行为的建模来预测学生的学业成绩，可以实现提前预警学生的异常情况，并进行适当的干预，因而对学生的培养、管理工作将会起到极其重要的作用。从某高校的某个学院随机抽取一定比例学生，提供这些学生在三个学期的图书馆进出记录、一卡通消费记录、图书馆借阅记录、以及综合成绩的相对排名。这一部分数据将作为训练数据。我们从另外的某学院随机抽取一定比例的学生，然后提供他们在三个学期的图书馆进出记录、一卡通消费记录、图书借阅记录、以及前两个学期的成绩排名。

希望通过借助大数据相关的挖掘技术和基础算法，预测第三学期的成绩排名。

**4(c)**

1 基于用户位置信息的商业选址

随着信息技术的快速发展，移动设备和移动互联网已经普及到千家万户。在用户使用移动网络时，会自然的留下用户的位置信息。随着近年来GIS地理信息技术的不断完善普及，结合用户位置和GIS地理信息将带来创新应用。如百度与万达进行合作，通过定位用户的位置，结合万达的商户信息，向用户推送位置营销服务，提升商户效益。

2 国家电网用户画像

随着电力体制改革向纵深推进，售电侧逐步向社会资本放开，当下的粗放式经营和统一式客户服务内容及模式，难以应对日益增长的个性化、精准化客户服务体验要求。如何充分利用现有数据资源，深入挖掘客户潜在需求，改善供电服务质量，增强客户黏性，对公司未来发展至关重要。

对电力服务具有较强敏感度的客户对于电费计量、供电质量、电力营销等各方面服务的质量及方式上往往具备更高的要求，成为各级电力公司关注的重点客户。经过多年的发展与沉淀，目前国家电网积累了全网4亿多客户档案数据和海量供电服务信息，以及公司营销、电网生产等数据，可以有效的支撑海量电力数据分析。

因此，国家电网公司希望通过大数据分析技术，科学的开展电力敏感客户分析，以准确地识别敏感客户，并量化敏感程度，进而支撑有针对性的精细化客户服务策略，控制电力服务人工成本、提升企业公众形象。

3 网站关键词来源聚类整和

以领域特征明显的词和短语作为聚类对象，在分类系统的大规模层级分类语料库中，利用文本分类的特征提取算法进行词语的领域聚类，通过控制词语频率的影响，分别获取领域通用词和领域专类词。

**5**

Advantages: 与实际观测的数据更接近，偏差更小，对非线性模型拟合得更好。

Disadvantages：易受到噪声影响，方差大，易出现过度拟合的情况，解释性没有less flexible approach强。

当数据存在明显非线性关系且观测数据量相比预测变量较多时，用more flexible approach更好，当数据是线性关系时用less flexible approach更好。

如果我们的主要目的是prediction，那么用more flexible approach更好，如果我们的主要目的是inference，那么用less flexible approach更好。

**6**

Parametric statistical learning approach: it reduces the problem of estimating f down to one of estimating a set of parameters.

Non-parametric statistical learning approach: it does not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly

Advantages: Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters.

Disadvantages: the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f, then our estimate will be poor. A more flexible models can address this problem but may lead to overfitting meanwhile.

**7(a)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs | X1 | X2 | X3 | distance | Y |
| 1 | 0 | 3 | 0 | 3 | Red |
| 2 | 2 | 0 | 0 | 2 | Red |
| 3 | 0 | 1 | 3 |  | Red |
| 4 | 0 | 1 | 2 |  | Green |
| 5 | -1 | 0 | 1 |  | Green |
| 6 | 1 | 1 | 1 |  | Red |

**7(b)**

Green. (0,0,0)距离第5个点最近。

**7(c)**

Red. (0,0,0)距离第2，5，6三个点最近，Red的概率为，比Green概率大。

**7(d)**

Small. K值越小，模型越flexible，K值越大决策边界越接近线性。

**8(a)**

>College=read.csv("C:\\Users\\25106\\Desktop\\ISLR\\R数据集\\College.csv")

**8(b)**

> fix(College)

> rownames(College)=College[,1]

> fix(College)

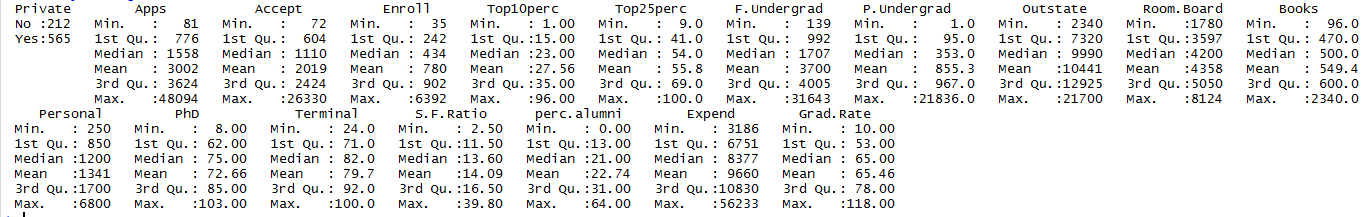
> College=College [,-1]

> fix(College)

**8(c)**

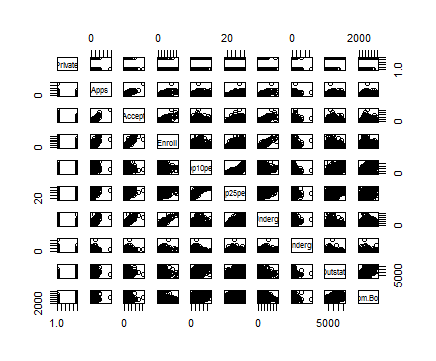
**i**

> summary(College)



**ii**

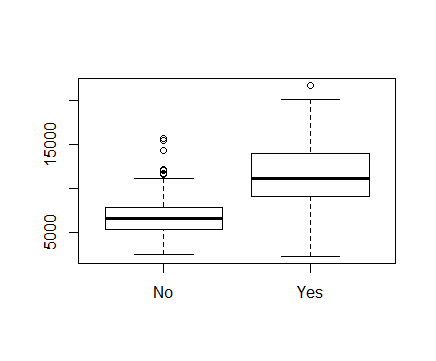
> pairs(College[,1:10])



**iii**

> attach(College)

> plot(Private,Outstate)



**iv**

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

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

> Elite=as.factor(Elite)

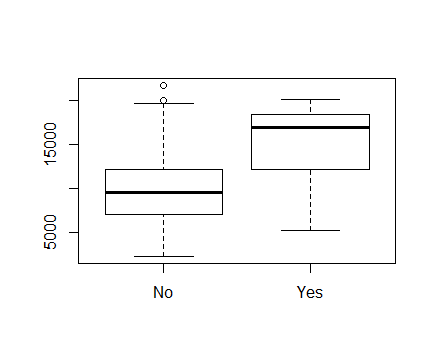
> College=data.frame(College,Elite)

> summary(Elite)

No Yes

699 78

> plot(Elite,Outstate)



**v**

opar<-par(no.readonly = TRUE)

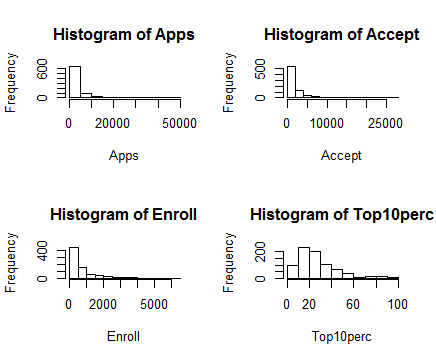
> par(mfrow=c(2,2))

> hist(Apps)

> hist(Accept)

> hist(Enroll)

> hist(Top10perc,breaks = 10)

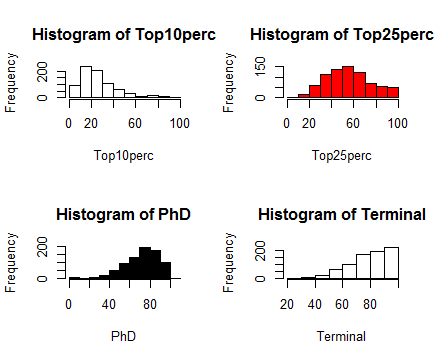


> hist(Top10perc,breaks = 10)

> hist(Top25perc,breaks = 10,col = 2)

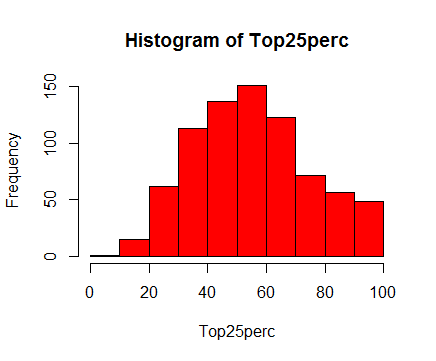
> hist(PhD,breaks = 8,col = 1)

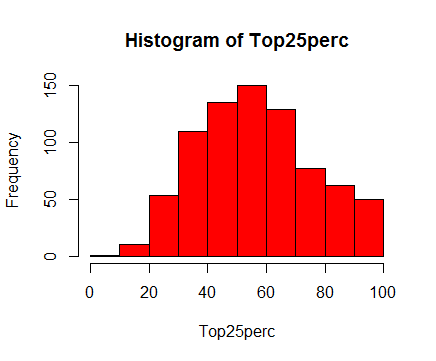
> hist(Terminal,breaks = 6)



>hist(Top25perc,breaks = 10,col = 2,right = T)

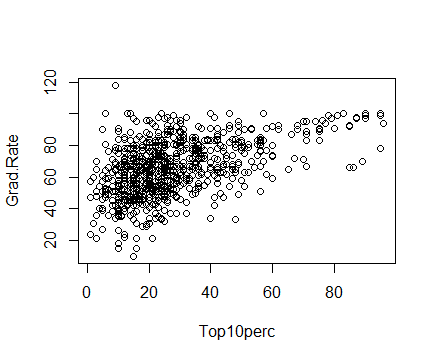
> hist(Top25perc,breaks = 10,col = 2,right = F)





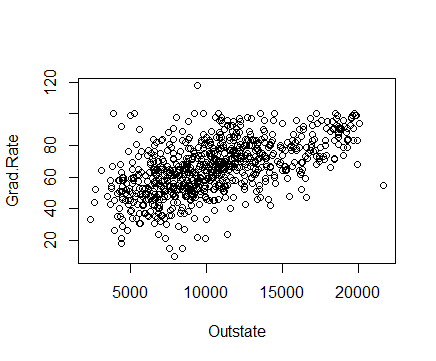
vi

> plot(Top10perc,Grad.Rate)



从前10%的高中录取的学生所占比例越大，学校学生毕业率倾向于越高，不过不是必然的。

plot(Outstate, Grad.Rate)



Out-state学费越高，学校学生毕业率倾向于越高，同样也不是必然的。

**9(a)**

> Auto = read.csv("C:\\Users\\25106\\Desktop\\ISLR\\R数据集\\Auto.csv",na.strings="?")

> Auto=na.omit(Auto)

quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year

qualitative: name, origin

**9(b)**

> sapply(Auto[1:7], range)

mpg cylinders displacement horsepower weight acceleration year

[1,] 9.0 3 68 46 1613 8.0 70

[2,] 46.6 8 455 230 5140 24.8 82

**9(c)**

> sapply(Auto[1:7], mean)

mpg cylinders displacement horsepower weight acceleration year

23.445918 5.471939 194.411990 104.469388 2977.584184 15.541327 75.979592

> sapply(Auto[1:7], median)

mpg cylinders displacement horsepower weight acceleration year

22.75 4.00 151.00 93.50 2803.50 15.50 76.00

**9(d)**

> Auto1=Auto[-(10:85),]

> sapply(Auto1[1:7], range)

mpg cylinders displacement horsepower weight acceleration year

[1,] 11.0 3 68 46 1649 8.5 70

[2,] 46.6 8 455 230 4997 24.8 82

> sapply(Auto1[1:7], mean)

mpg cylinders displacement horsepower weight acceleration year

24.404430 5.373418 187.240506 100.721519 2935.971519 15.726899 77.145570

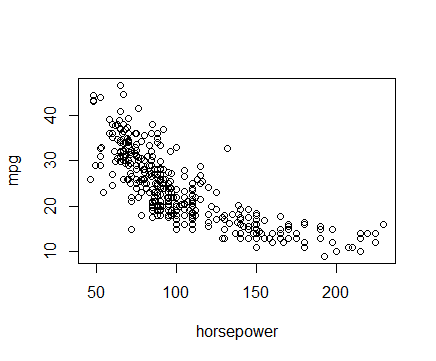
> sapply(Auto1[1:7],sd)

mpg cylinders displacement horsepower weight acceleration year

7.867283 1.654179 99.678367 35.708853 811.300208 2.693721 3.106217

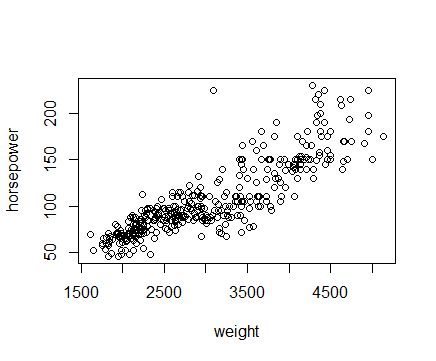
**9(e)**

plot(horsepower,mpg)



车马力越大，英里每加仑越小，即油耗越大。

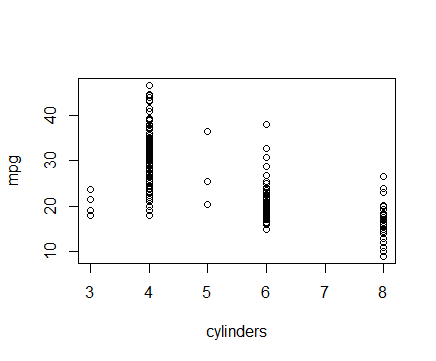
> plot(weight,horsepower)



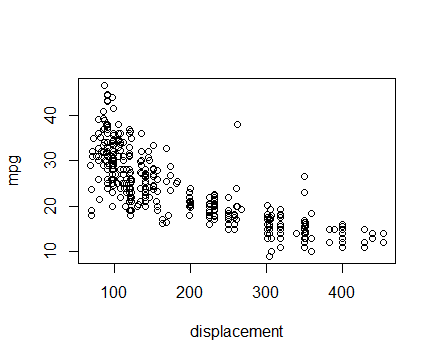
车越重，马力越大。

**9(f)**

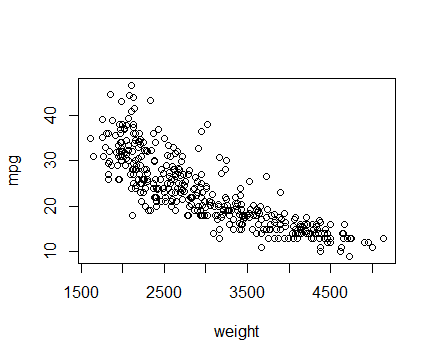
> plot(cylinders,mpg)



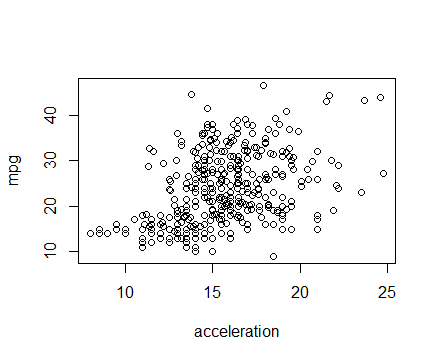
> plot(displacement,mpg)



> plot(weight,mpg)



> plot(acceleration,mpg)



气缸越多，马力越大，发动机排量越大，车越重，从0到60英里/小时的加速时间越短，英里每加仑小，即油耗越大。

**10(a)**

> dim(Boston)

[1] 506 14

506行，14列

> ?Boston

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（全额财产税税率）.

ptratio pupil-teacher ratio by town.

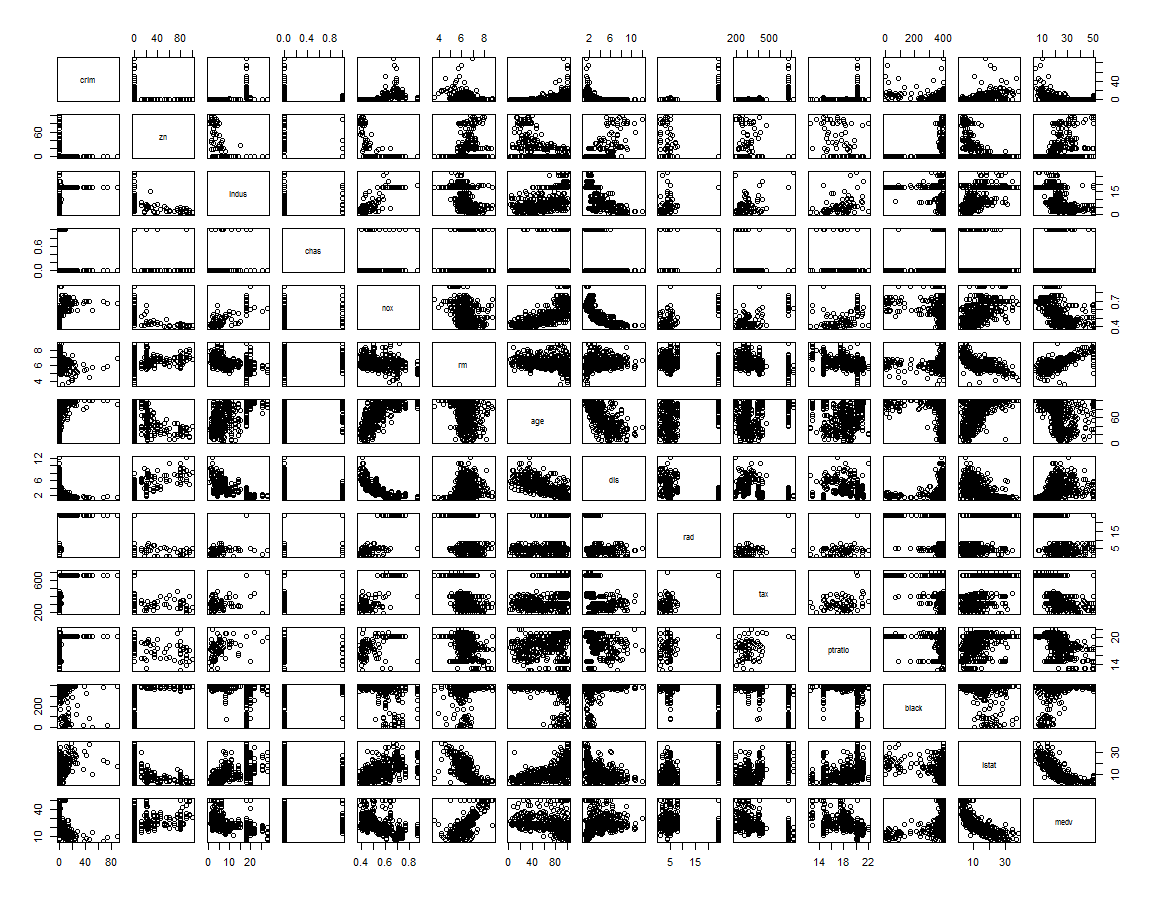
black 1000(Bk - 0.63)^2 where *Bk* is the proportion of blacks by town.

lstat lower status of the population (percent).

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

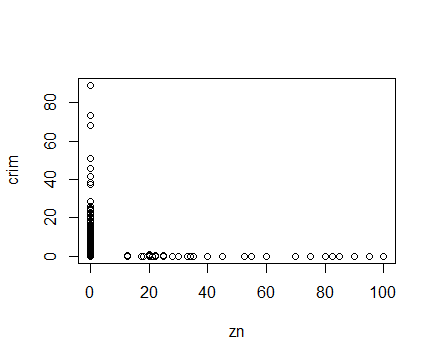
**10(b) (c)**

> pairs(Boston)

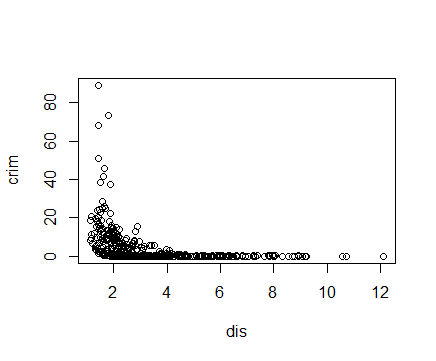


> attach(Boston)

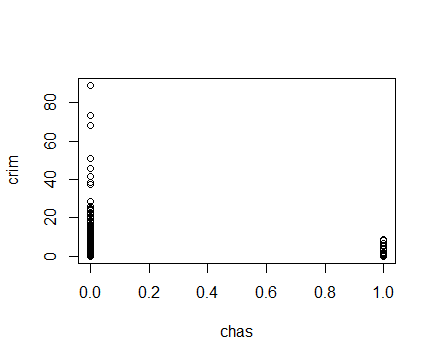
> plot(zn,crim)



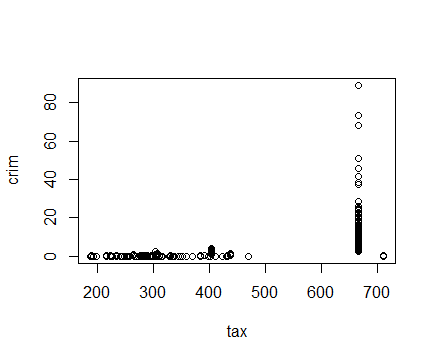
> plot(dis,crim)



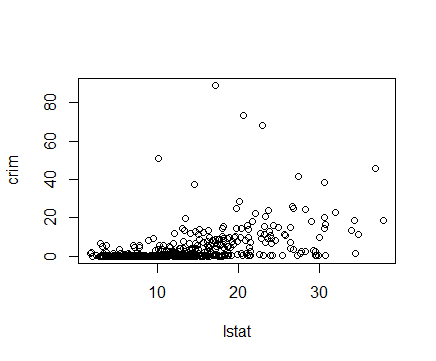
> plot(chas,crim)



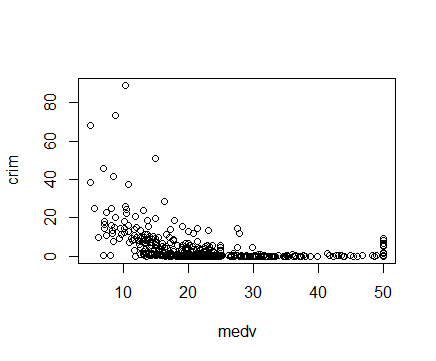
> plot(tax,crim)



> plot(lstat,crim)



> plot(medv,crim)



根据散点图，crim和zn，chas，dis，tax， lstat及medv可能有关。

zn比较小时人均犯罪比例较高。离five Boston employment centres距离越近，即dis越小的地区人均犯罪比例越高。在Charles River附近，即chas=1的地区人均犯罪比例较高。每10000美元的全额财产税税率高，即tax较高的地区人均犯罪比例较高。lstat比例较高的地区人均犯罪比例较高。medv越高的地区人均犯罪比例越高。

**10(d)**

> range(crim)

[1] 0.00632 88.97620

> dim(subset(Boston,crim>20))[1]

[1] 18

> dim(subset(Boston,crim>10))[1]

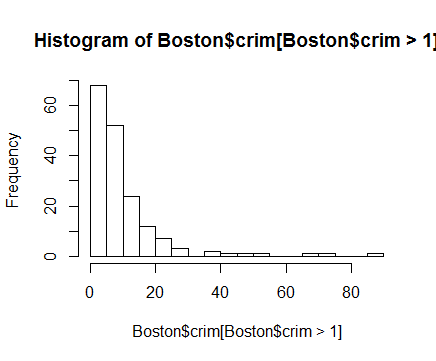
[1] 54

> dim(subset(Boston,crim>1))[1]

[1] 174

人均犯罪比例超过20的有18个郊区，超过10的有54个郊区，超过1的有174个郊区。

> hist(Boston$crim[Boston$crim>1],breaks = 20)



> range(tax)

[1] 187 711

> dim(subset(Boston,tax>600))[1]

[1] 137

> dim(subset(Boston,tax>500))[1]

[1] 137

> dim(subset(Boston,tax>400))[1]

[1] 200

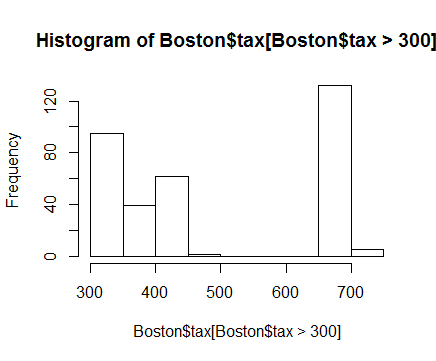
> dim(subset(Boston,tax>300))[1]

[1] 334

> dim(subset(Boston,tax>200))[1]

[1] 489

> hist(Boston$tax[Boston$tax>300],breaks = 10)



tax在600以上的郊区有137个，400到600的比较少，只有63个。300以上共有334个。

> range(ptratio)

[1] 12.6 22.0

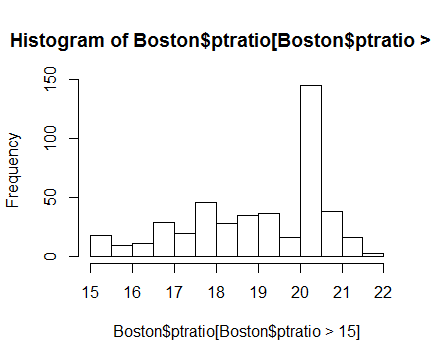
> dim(subset(Boston,ptratio>20))[1]

[1] 201

> dim(subset(Boston,ptratio>15))[1]

[1] 448

> hist(Boston$ptratio[Boston$ptratio>15],breaks = 10)



总体较为平均，比例在20到20.5的郊区最多，共有145个。

**10(e)**

> dim(subset(Boston,chas==1))[1]

[1] 35

35个郊区在Charles river河畔。

**10(f)**

> median(Boston$ptratio)

[1] 19.05

**10(g)**

> t(subset(Boston, medv == min(Boston$medv)))

399 406

crim 38.3518 67.9208

zn 0.0000 0.0000

indus 18.1000 18.1000

chas 0.0000 0.0000

nox 0.6930 0.6930

rm 5.4530 5.6830

age 100.0000 100.0000

dis 1.4896 1.4254

rad 24.0000 24.0000

tax 666.0000 666.0000

ptratio 20.2000 20.2000

black 396.9000 384.9700

lstat 30.5900 22.9800

medv 5.0000 5.0000

> Boston1=Boston[,-4]

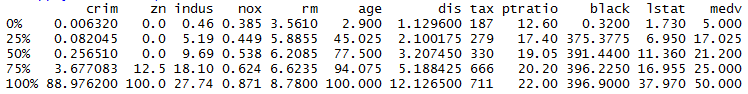
> Boston1=Boston1[,-8]

> sapply(Boston1, mean)





> sapply(Boston1, quantile)



两个城市的crim都很高，zn都为0，indus超过约75%的郊区，nox较高，rm较低，age很高，dis小于不超过25%的郊区，tax，ptratio，black较高，超过约75%的郊区，lstat处于较高水平。

**10(h)**

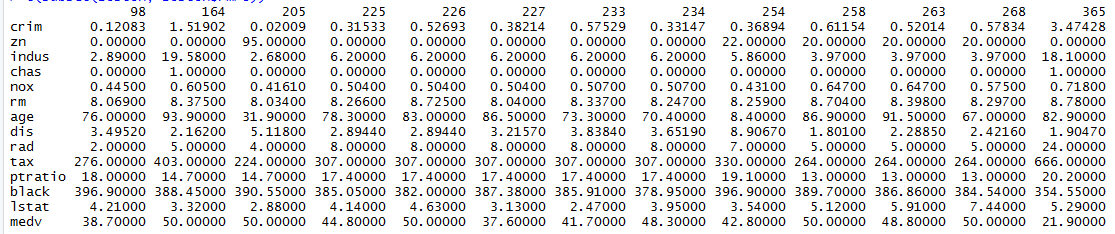
> dim(subset(Boston,rm>7))[1]

[1] 64

> dim(subset(Boston,rm>8))[1]

[1] 13

> t(subset(Boston, Boston$rm>8))



crim较低，除了164和365都不在Charles river河畔，除了365tax都不高。