高级统计方法 第2次作业:

序号： 03 姓名： 潘晨楷 学号： 20212241116 班级：软2107

**概念**

1.问题(略)

（a）问题（略）

容易欠拟合，光滑度高的学习模型更好。

（b）问题（略）

容易过拟合，光滑度低的学习模型更好。

（c）问题（略）

由题意光滑度低的学习模型可能会欠拟合，所以光滑度高的模型更好。

（d）问题（略）

光滑度低的学习模型更好，因为光滑度高的学习模型方差也更大。

2.问题（略）

（a）问题（略）

回归模型，推断。

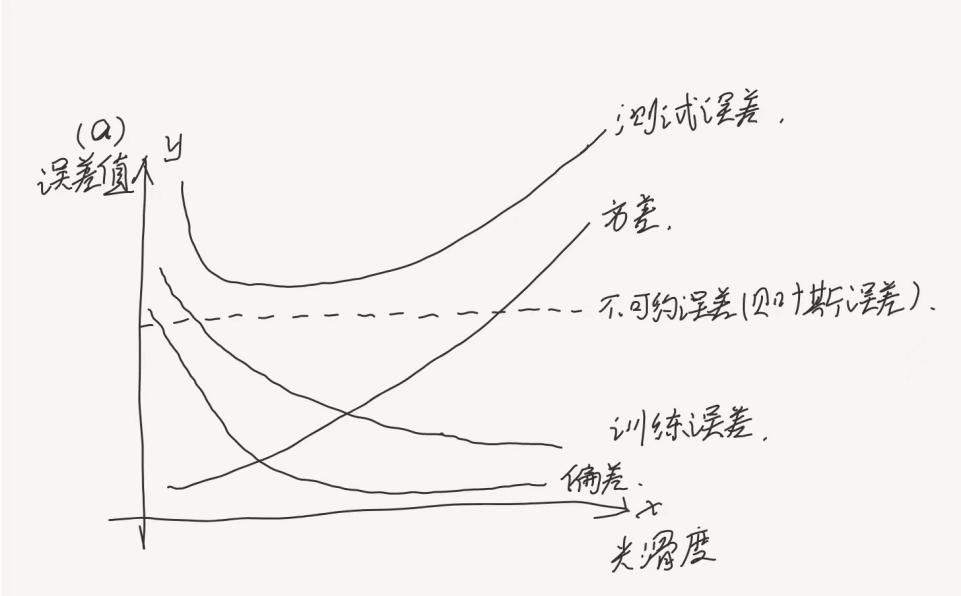
（b）问题（略）

分类器，预测。

（c）问题（略）

回归模型，预测。

1. 问题（略）
2. 问题（略）



测试误差：其为测试数据与模型预测输出之间的误差，表现为U型曲线，开始随着光滑度的增加越来越接近拟合效果更好，之后随着光滑度增加会发生过拟合，拟合效果变差。

方差：在统计描述中，方差用来计算每一个变量（观察值）与总体均数之间的差异。随着光滑度升高而增加，容易发生过拟合现象。

贝叶斯（或不可约）误差：是定义上误差的下限，测试误差由于输出中的误差导致的不可约误差。当训练误差低于不可约误差时就发生了过拟合。

偏差（平方）：偏差指的是为了选择一个模型逼近真实函数而被带入的误差，随着光滑度的提高而变得越来越小。

训练误差：描述训练的样本数据与训练模型输出之间的误差，用来协助拟合模型，因为这些数据本来就是用来拟合模型的，所以预测精度一般会比较高。同时，随着光滑度增加，其误差越来越小，拟合性也越来越强。

1. 问题（略）

学习模型的光滑度提高，方差会变大，偏差会变小。当增加一类模型的灵活性使得偏差下降明显，方差上升趋势小于偏差，则测试误差就会下降。当光滑度的增加对偏差的下降影响较小，而使方差上升明显，则测试误差增加。而训练误差会随着光滑度的增加而一直减小。

4.问题（略）

（a）问题（略）

1.垃圾邮件过滤器：在这个例子中，响应变量是电子邮件是否为垃圾邮件（1表示是垃圾邮件，0表示不是垃圾邮件）。预测变量可能是邮件的发件人、主题、内容等特征。目标是推断一封电子邮件是否为垃圾邮件，以便用户可以选择删除或不打开这些邮件。

该应用的目标是预测，预测变量的值。

1. 信用评分：在这个例子中，响应变量是个人的信用评分（例如，300-850分）。预测变量可能包括收入、债务水平、信用历史长度、还款记录等。目标是预测一个人的信用风险，以便金融机构可以决定是否批准贷款或调整利率。

该应用的目标是预测，预测变量的值。

1. 疾病诊断：在这个例子中，响应变量是患者是否患有某种疾病（例如，阳性表示患病，阴性表示未患病）。预测变量可能包括患者的年龄、性别、家族病史、症状严重程度等。目标是预测患者是否患有特定疾病，以便医生可以提供适当的治疗建议。

该应用的目标是预测，预测变量的值。

1. 问题（略）
2. 房价预测：在这个例子中，响应变量是房屋的售价。预测变量可能包括房屋的面积、地理位置、建造年份、周边设施等。目标是预测未来的房价走势，以便购房者和开发商可以做出更明智的决策。
3. 销售预测：在这个例子中，响应变量是产品的销售额。预测变量可能包括市场需求、竞争对手的价格、促销活动、季节性因素等。目标是预测未来一段时间内的销售额，以便企业可以制定合适的生产计划和营销策略。
4. 股票价格预测：在这个例子中，响应变量是股票的价格。预测变量可能包括公司的利润、市场趋势、经济指标、政策变化等。目标是预测未来一段时间内的股票价格变动，以便投资者可以做出更明智的投资决策。
5. 问题（略）
6. 客户细分：聚类分析可以用于将客户划分为不同的群体，以更好地了解他们的需求和偏好。例如，一家服装零售商可以使用聚类分析来根据客户的购买记录将他们分为不同的群体，然后针对不同的群体制定不同的营销策略。
7. 文档分类：聚类分析可以用于将文档自动分配到不同的类别中。例如，在自然语言处理中，聚类分析可以用于将新闻文章自动分类为政治、经济、体育等不同的主题。
8. 图像分割：聚类分析可以用于将图像中的像素分组，从而更好地理解图像的内容。例如，在医学图像处理中，聚类分析可以用于将肿瘤细胞与正常组织区分开来。

5.问题（略）

一个光滑度高的回归模型或者分类模型，能够更好的拟合非线性模型，偏差更小。但是模型越光滑，所需要计算的参数就越多，而且容易过拟合，方差更大。

当我们更想预测而不是推断的时候，我们优先考虑光滑度高的模型。

当我们更像推断而不是预测的时候，我们优先考虑光滑度低的模型。

1. 问题（略）
2. 参数方法是一种基于模型估计的两阶段方法。优点是，它把估计f的问题简化到估计一组参数，对f假设一个具体的参数形式将简化对f的估计，因为估计参数是更为容易的，不需要拟合任意一个函数f。缺点是，选定的模型并非与实际的f形式上一致，而且还有过拟合的可能情况。
3. 非参数方法不需要对函数f的形式实现做明确说明的假设。相反，这类方法追求的接近数据点的估计，估计函数在去粗和光滑处理后尽量可能与更多的数据点接近。优点是，不限定函数f的具体形式，可以更大的范围选择更适宜的f形状的估计。缺点是，无法将估计f的问题简单到对少数参数进行估计的问题，所以往往需要大量的观察点。
4. 问题（略）
5. 问题（略）
6. 3 2. 2 3. 4. 5. 6.
7. 问题（略）

Green，当K = 1时最近的点是obs.5 ，响应变量是Green

1. 问题（略）

Red，当K = 3时最近的三个点是obs.、obs.5、obs.6，两个Red一个Green， 所以预测结果为Red

1. 问题（略）

小。1/K越大，光滑度越高，也就越非线性。所以K应该小。

、、、、、、

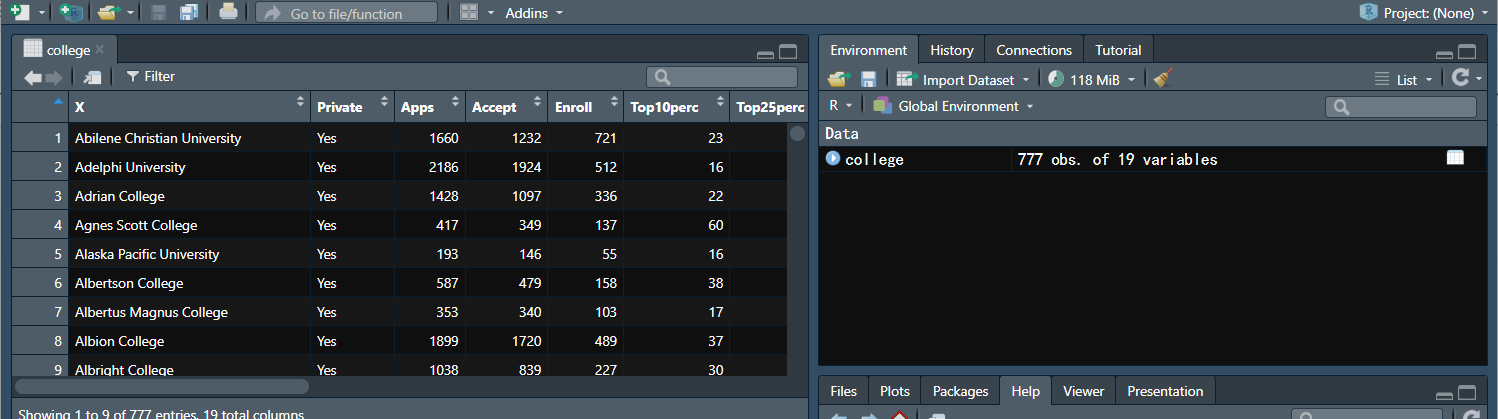
**应用**

8.问题（略）

（a）问题（略）

脚本：college=read.csv("E:\\study\\College.csv")

截图：



（b）问题（略）

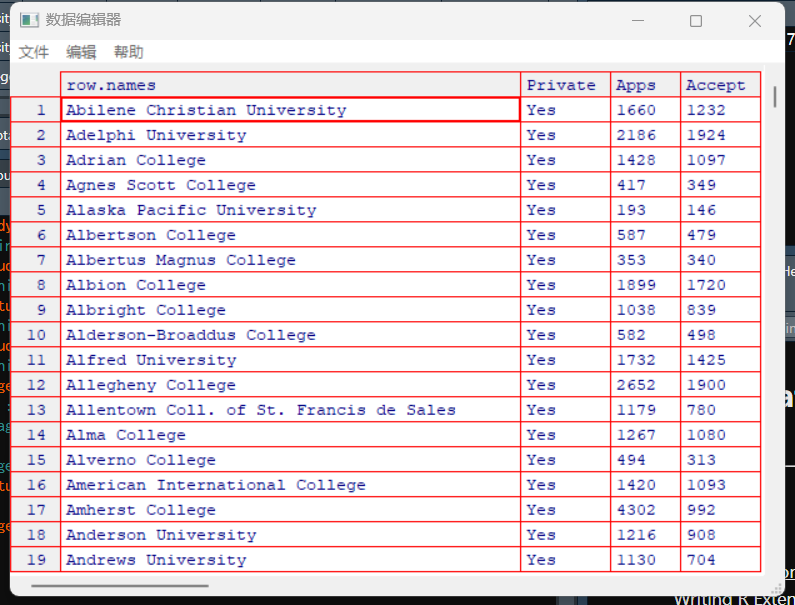
> rownames(college)=college[,1]

> fix(college)



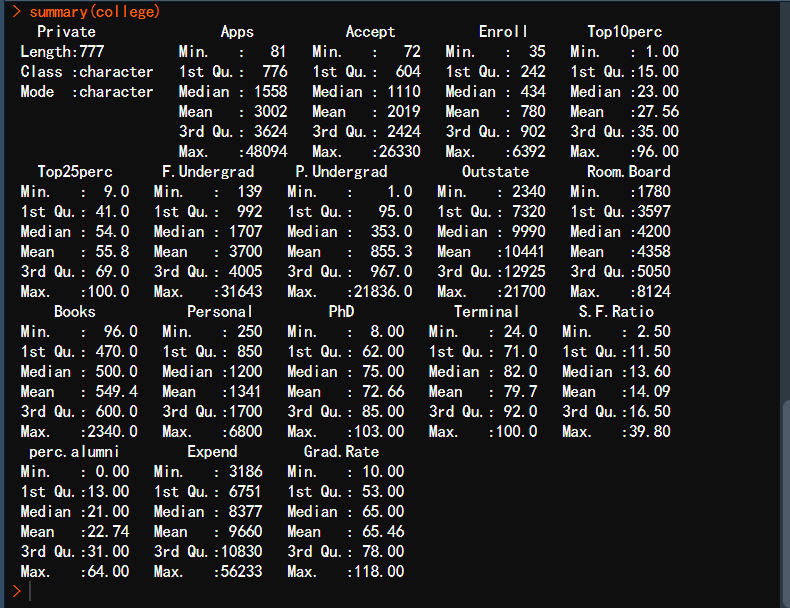
> college=college[,-1]

> fix(college)



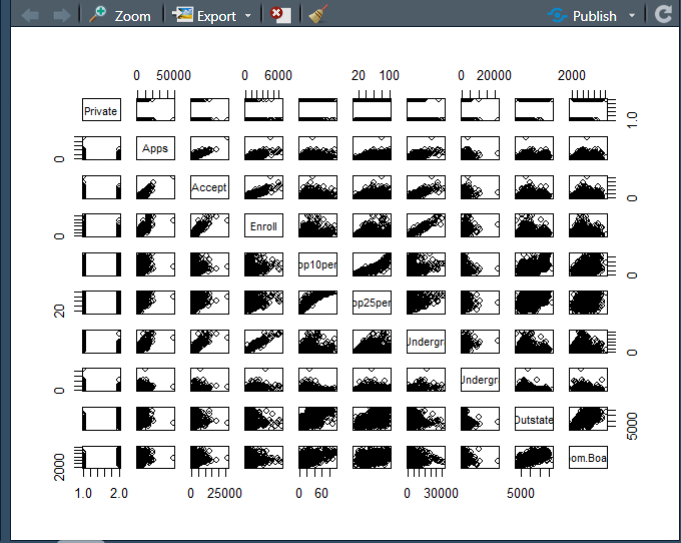
（c）问题（略）

i.summary(college)

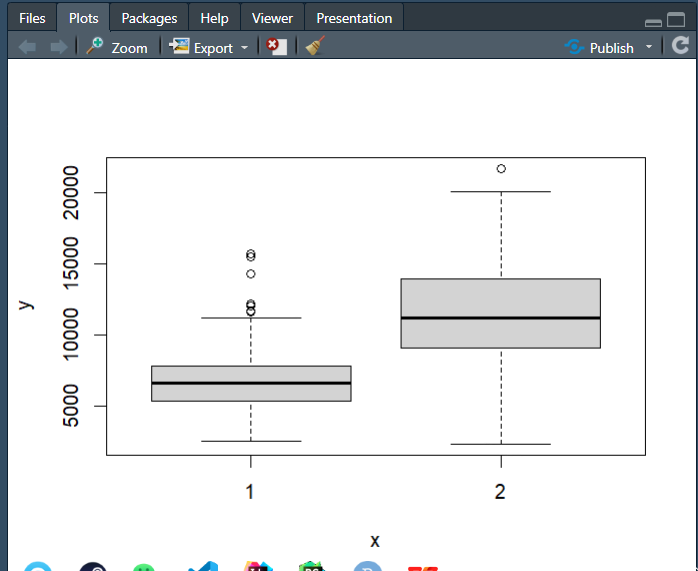


ii. > college[,1] = as.numeric(factor(college[,1]))

> pairs(college[,1:10])



iii.> plot(factor(college$Private),college$Outstate) （1：No，2：Yes）



iv. > plot(college$Outstate,college$Private)

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

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

> Elite=as.factor(Elite)

> college=data.frame(college,Elite)

> summary(Elite)

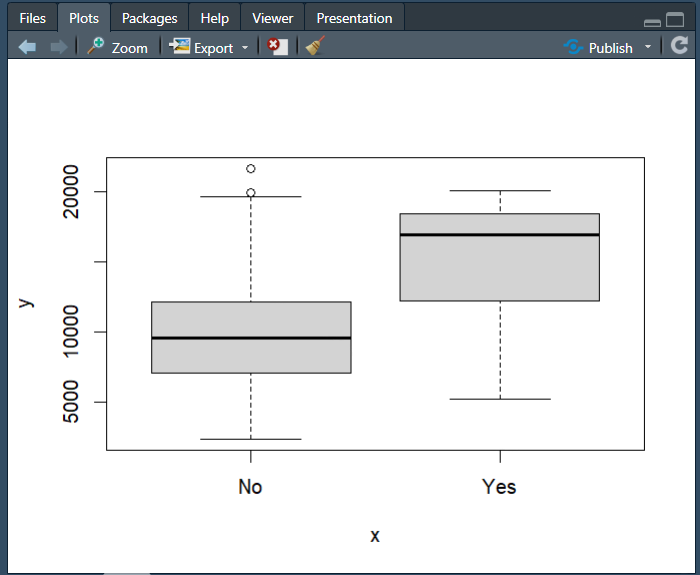
显示：

No Yes

699 78

说明有78个精英大学

plot(college$Elite,college$Outstate)



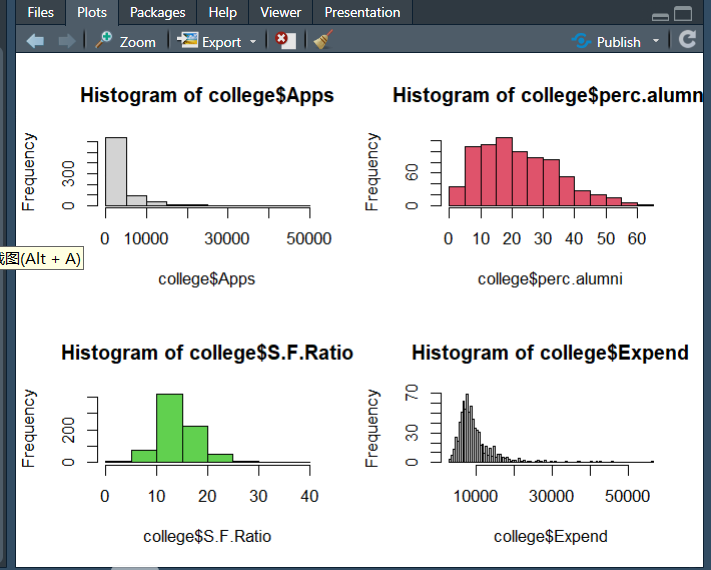
V. > par(mfrow=c(2,2))

> hist(college$Apps)

> hist(college$perc.alumni,col=2)

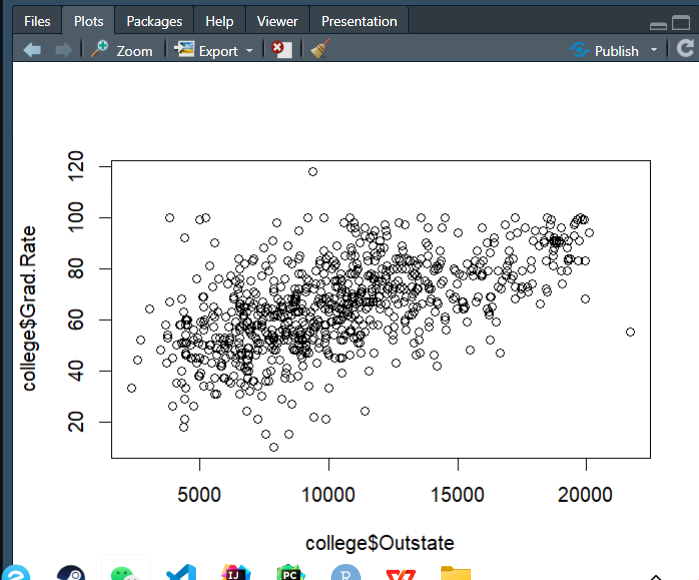
> hist(college$S.F.Ratio,col=3,breaks=10)

> hist(college$Expend,breaks=100)



vi. 1.探究学费高的大学毕业率如何：

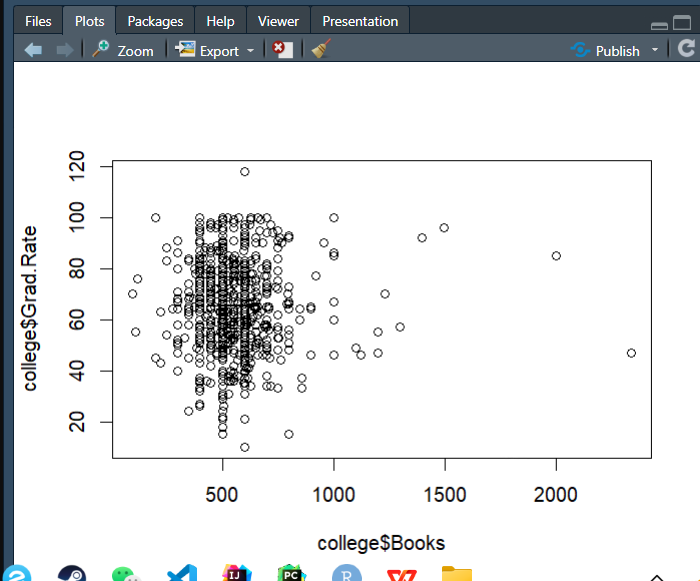
plot(college$Outstate,college$Grad.Rate)



发现学费高的大学毕业率也高

1. 探究大学藏书和学生毕业率的关系：

plot(college$Books,college$Grad.Rate)



发现藏书越多的大学毕业率很高，藏书低的大学的毕业率从高到低都有。

9.问题（略）

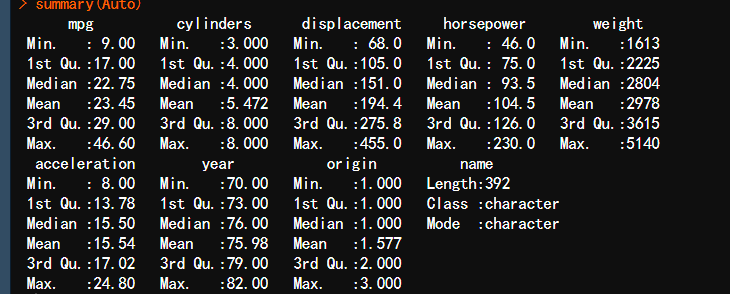
> Auto=read.csv("E:\\study\\Auto.csv",header = TRUE,na.strings = "?")

> Auto=na.omit(Auto)

> dim(Auto)

[1] 392 9

> summary(Auto)



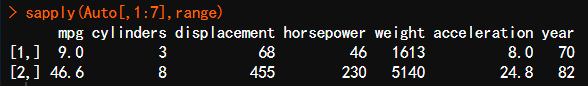
（a）问题（略）

定量：mpg、cylinders、dislacement、horsepower、weight、acceleration、year

定性：origin、name

（b）问题（略）

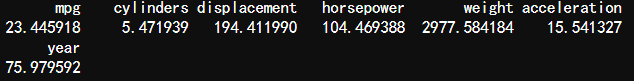
sapply(Auto[,1:7],range)



（c）问题（略）

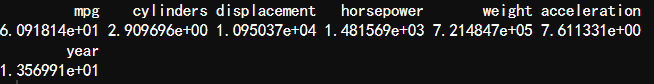
均值：

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



方差：

sapply(Auto[,1:7],var)



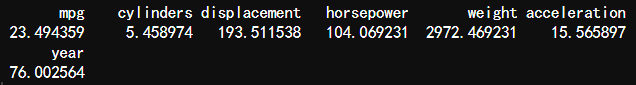
（d）问题（略）

> newAuto=Auto[-c(10,85),]

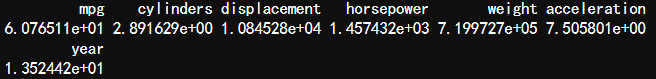
> sapply(newAuto[,1:7],range)

1694585046972

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



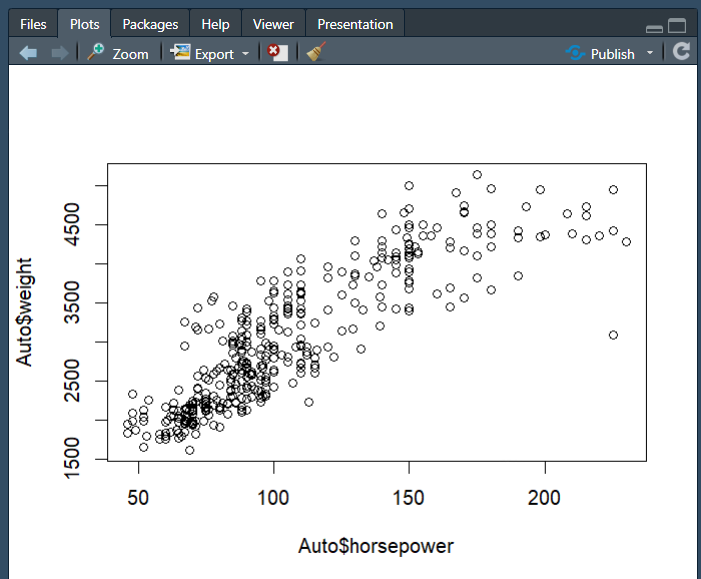
> sapply(newAuto[,1:7],var)



1. 问题（略）

探究车辆马力和重量之间的关系：

plot(Auto$horsepower,Auto$weight)

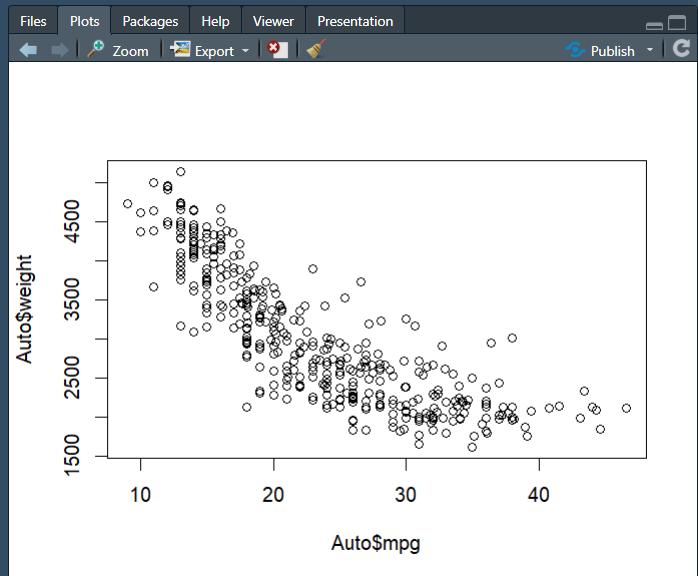


可以发现，马力和重量大体呈现线性关系，马力越大的车，车的重量越大。

1. 问题（略）

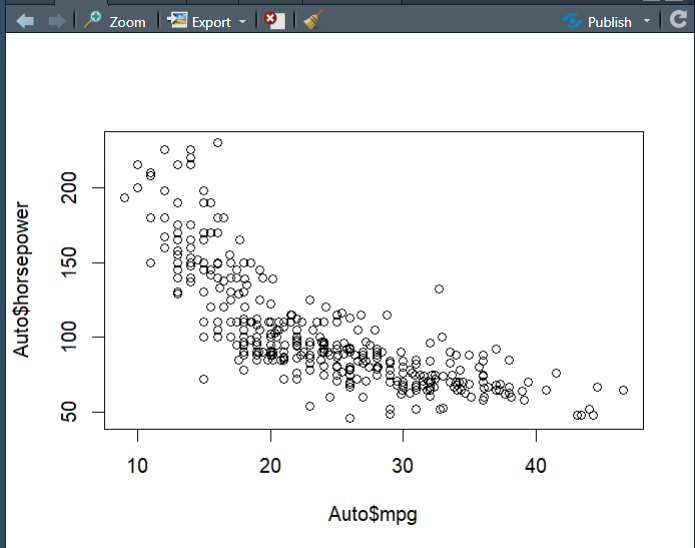
mpg与weight之间的关系：

plot(Auto$mpg,Auto$weight)



mpg与horsepower之间的关系：

plot(Auto$mpg,Auto$horsepower)



从两幅图中不难看出mpg和weight、horsepower都有关系，大体上呈线性关系。mpg越高，weight和horsepower都越小。

1. 问题（略）
2. 问题（略）

> library(MASS)

> ?Boston

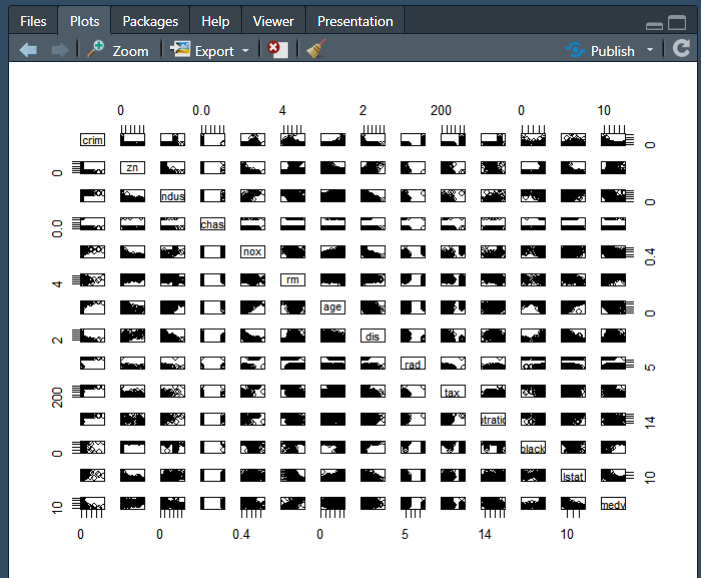
> dim(Boston)

1694586724519

Boston数据集有506行，14列。每行代表一行数据，包括城镇人口犯罪率、城镇非零售商用土地的比例等数据。每列代表一个属性，比如crim：城镇人口犯罪率，ZN：住宅用地超过25000sq.ft.的比例。

1. 问题（略）

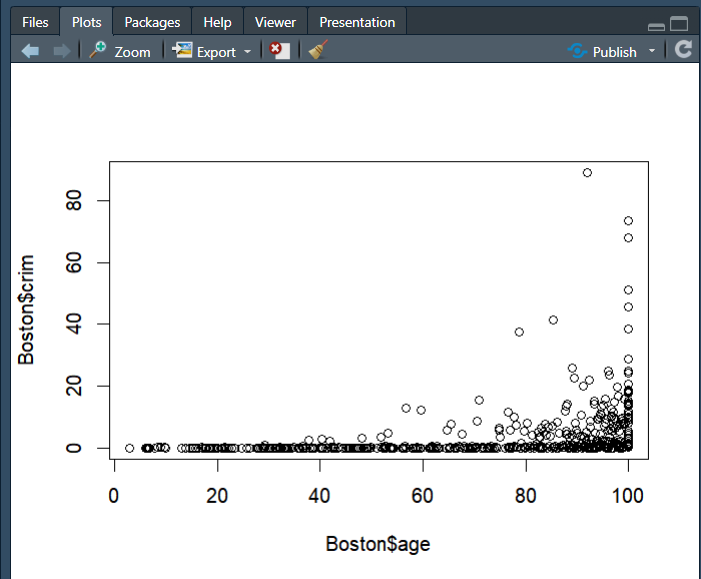
pairs(Boston)



发现每两个属性之间都有着很紧密的关系，例如chas属性：查理斯河空变量(如果边界是河流,则为1;否则为0)，基本上属性值为1和属性值为0的数据一样多。

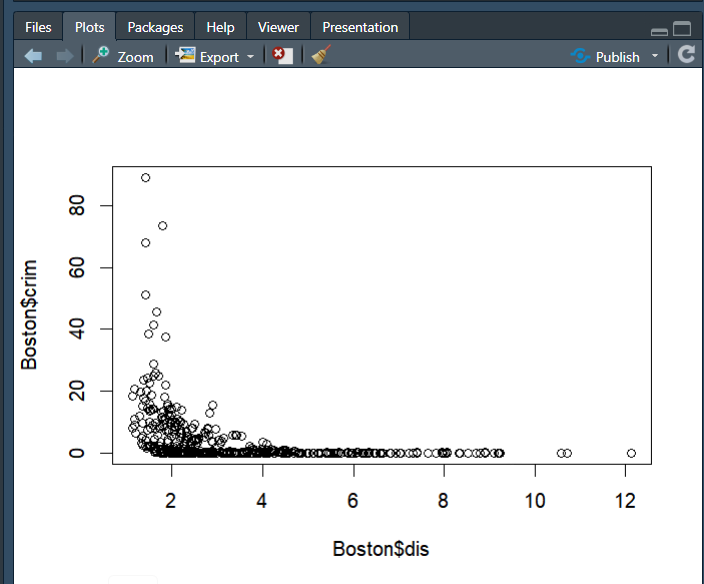
1. 问题（略）

plot(Boston$age,Boston$crim)



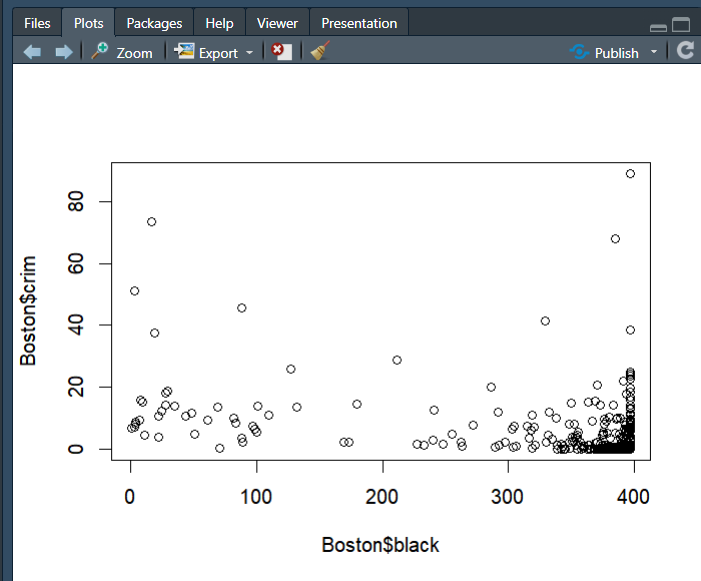
发现年龄越大，犯罪的比例越大。

plot(Boston$dis,Boston$crim)



发现距离波士顿就业中心越近，犯罪率越高。

plot(Boston$black,Boston$crim)



发现城镇中黑人比例越高的地方，犯罪率比比例低的地方要高。

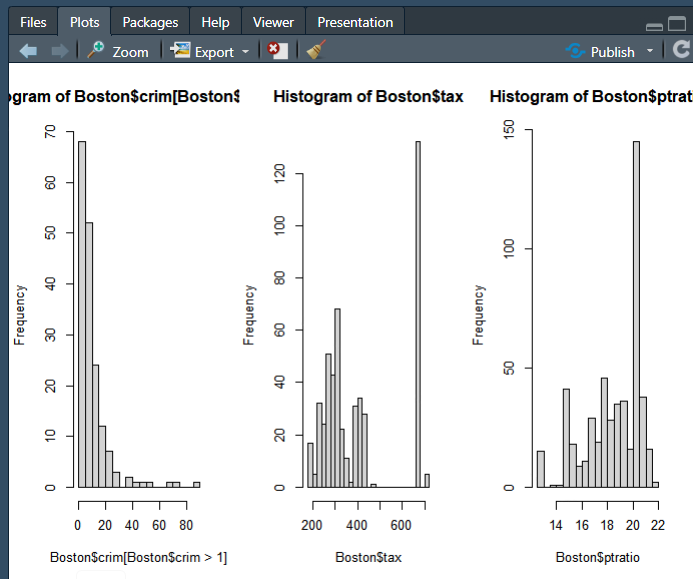
1. 问题（略）

> par(mfrow=c(1,3))

> hist(Boston$crim[Boston$crim>1], breaks=25)

> hist(Boston$tax, breaks=25)

> hist(Boston$ptratio, breaks=25)



大多数城市的犯罪率都很低，但出现了18个郊区犯罪率＞20，达到80以上。

低税率的郊区与660-680的峰值之间存在很大差距。

每一万美元的不动产税率向高比率倾斜，但没有特别高的比率

1. 问题（略）

1694588805610

35个。

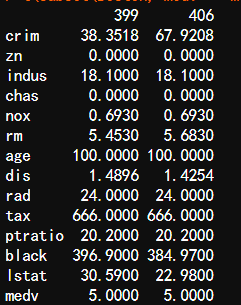
1. 问题（略）

1694588861201

19.05

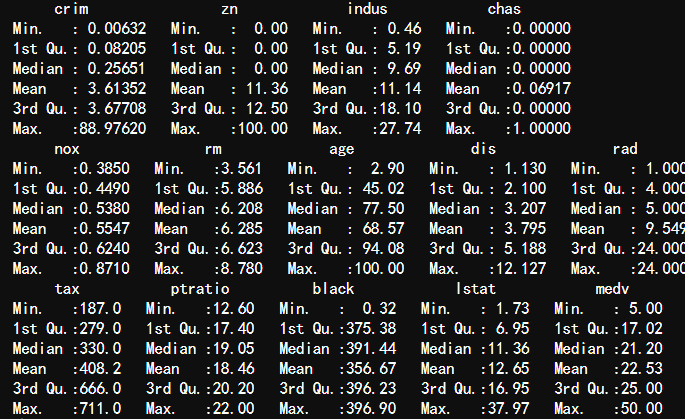
1. 问题（略）

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



中位数最小的波士顿郊区是第399个。

summary(Boston)



结合多方面的因素来看，此社区不是最好的居住地，但同时也不是最糟糕的。

1. 问题（略）

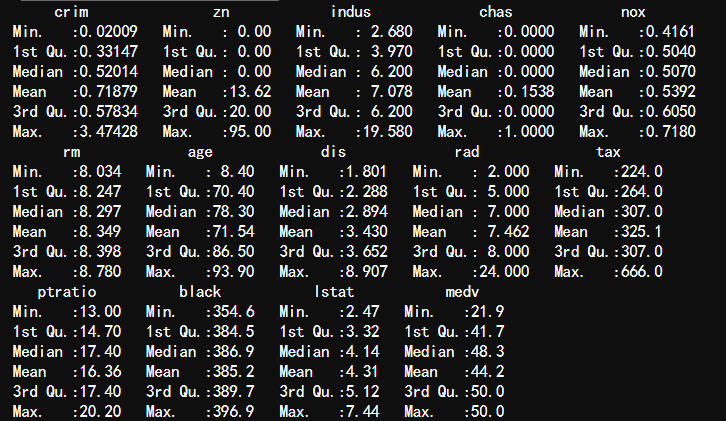
1694589174651

64个超过7的

1694589231220

13个超过8的

summary(subset(Boston, rm > 8))



作为居住地来讲中规中矩。相对较低的犯罪率，较低的地区中有多少房东属于低收入人群。