高级统计方法 第3次作业:

序号：27 姓名：王琪瑞 学号：20202241014 班级：软2002

**概念**

1.问题（略）

在表 3.4 中，“TV”的原假设是在存在广播的情况下，广告和报纸广告、电视广告对销售没有影响。 同样，在电视和报纸广告出现的情况下，广播对销售没有影响。电视和广播的低 p 值表明零假设对电视和广播来说是错误的。 报纸的高 p 值表明零假设对报纸是正确的。

3.问题（略）

得到的模型为：

Y=50+20\*(GPA)+0.07\*(IQ)+35\*(Gender)+0.01\*(GPA\*IQ)-10\*(GPA\*Gender)

（a）问题（略）

当IQ和GPA一定时，Y=常量+35\*(Gender)-10(GPA\*Gender)

男性：Y=常量

女性：Y=常量+35-10GPA

结论iii正确。女性的收入与随GPA增加而减少，所以在GPA足够高的情况下，男性的平均收入高于女性。

（b）问题（略）

GPA=4.0，IQ=110，Gender=1。

Y=50+20\*4.0+0.07\*110+35+0.01\*4.0\*110-10\*4.0=137.1

（c）问题（略）

错误。计算p值的方式来反映二者之间是否存在统计上的显著性，而不是只通过系数来判断。

4.问题（略）

（a）问题（略）

三次回归的训练RSS应该**小于**线性回归的训练RSS。因为三次回归会更紧密地对数据进行拟合。

（b）问题（略）

三次回归的测试RSS应该**大于**线性回归的测试RSS。因为三次回归训练的过度拟合比线性回归产生更大的误差。

（c）问题（略）

三次回归的训练RSS应该**小于**线性回归的训练RSS。具有更高光滑度的模型会比低光滑度的模型拥有更好的拟合效果来降低训练RSS。

（d）问题（略）

无法判断。X与Y既非线性，也无法判断其非线性程度。

如果相比三次回归更接近线性，则线性回归测试RSS可能低于三次回归。

如果相比线性更接近三次回归，则三次回归测试RSS可能低于线性回归。

**应用**

**3.7习题：**

8.问题（略）

（a）问题（略）

脚本：

library(ISLR)

write.csv(Auto,"Auto.csv")

auto = read.csv("Auto.csv", header=T, na.strings="?")

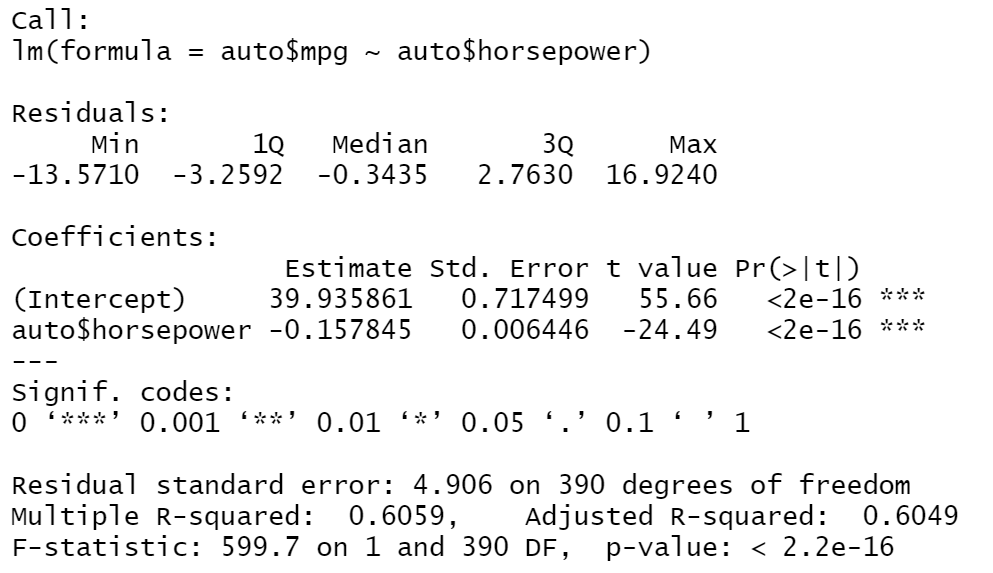
auto = na.omit(auto)

attach(auto)

lm.fit=lm(mpg~horsepower)

summary(lm.fit)

输出结果：



i．预测变量和响应变量是有关系的。将回归系数为0作为零假设，得出的F值统计量远大于1，且p值接近0，我们可以拒绝零假设，并指出horsepower和 mpg 之间存在统计显著性关系。

ii．由上述结果可以得出，该模型的RSE为4.906，由此可计算得误差百分比约为20.92%；此外还可得到该模型的R2的统计量为：0.6059；表示该模型可以对约有60.59%的变异做出解释；从上以及第一问可以看出，相关性还是可以的；

iii．通过结果的系数看出，预测变量和响应变量是负相关的。

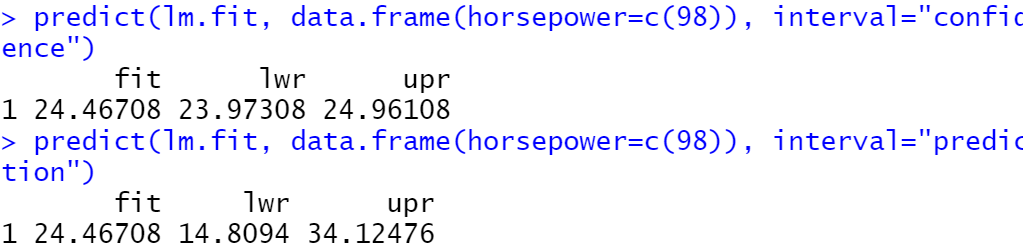
iv．

脚本：

predict(lm.fit, data.frame(horsepower=c(98)), interval="confidence")

predict(lm.fit, data.frame(horsepower=c(98)), interval="prediction")

截图：



预测值是24.46707，置信区间是(23.97,24.96)，预测区间是(14.81,34.12)。

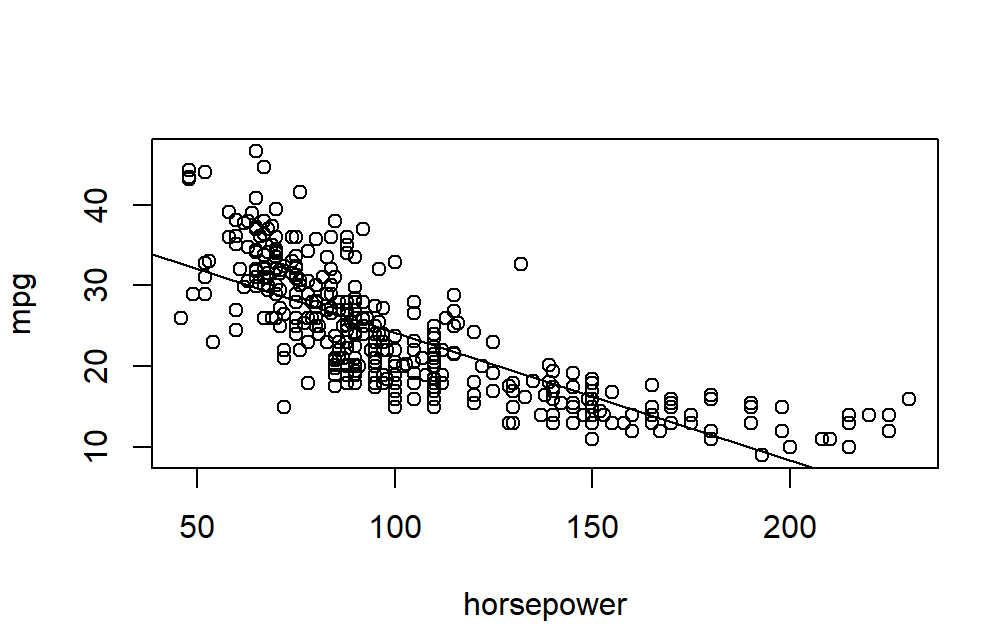
（b）问题（略）

脚本：

plot(horsepower, mpg)

abline(lm.fit)

截图：



（c）问题（略）

脚本：

par(mfrow=c(2,2))

plot(lm.fit)

截图：



1. 从拟合值-残差图可以得到，残差和拟合值基本无关，模型具有一定准确性。
2. 从Normal Q-Q图可以得到残差基本符合正态分布。
3. 从拟合值-方差图可以得到方差处于一个基本确定的状态。
4. 从杠杆图中可以看出基本没有极端的样本点。

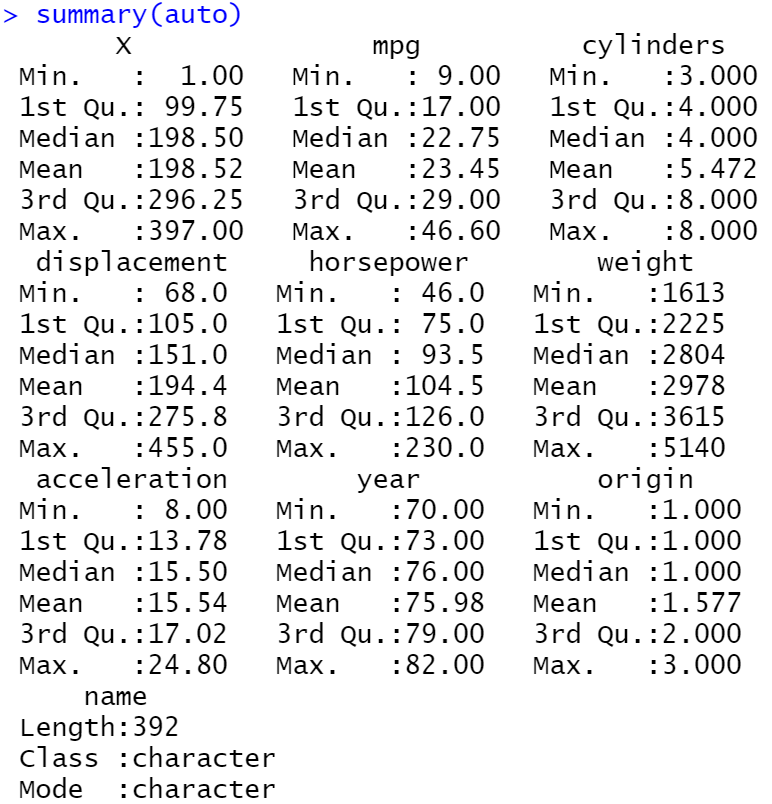
**2.4习题：**

9.问题（略）

（a）问题（略）

上一题已经去除了缺失数据。

调用summary函数可知



定量变量是:

mpg, cylinders, displacement, horsepower, weight, acceleration, year;

定性变量是;

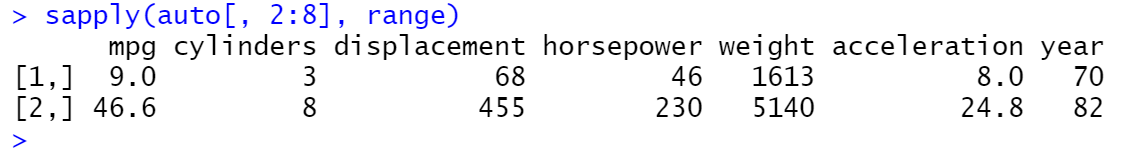
origin, name

（b）问题（略）

脚本：

sapply(auto[, 2:8], range)

截图：



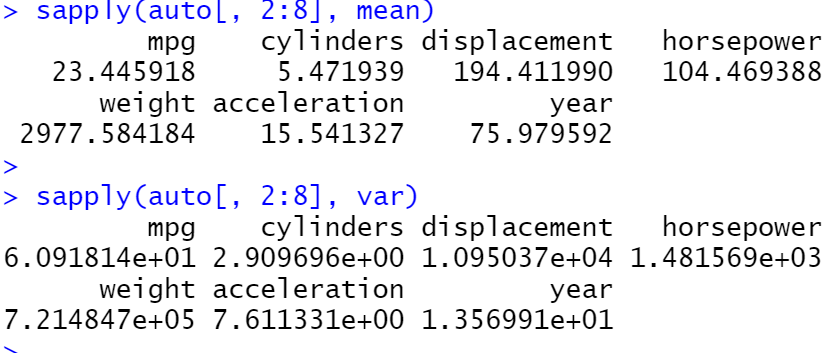
（c）问题（略）

脚本：

sapply(auto[, 2:8], mean) //均值

sapply(auto[, 2:8], var) //方差

截图：



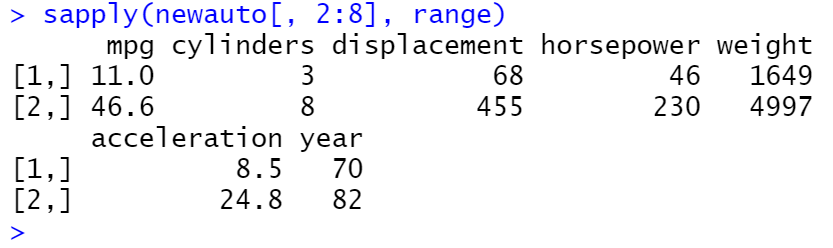
（d）问题（略）

脚本：

newauto = auto[-(10:85),]

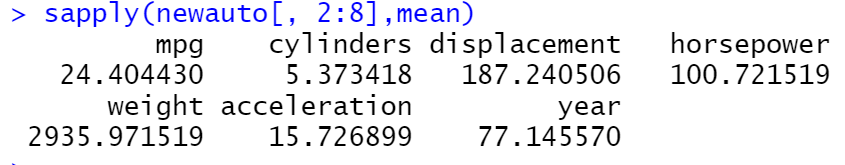
取值范围：

sapply(newauto[, 2:8], range)



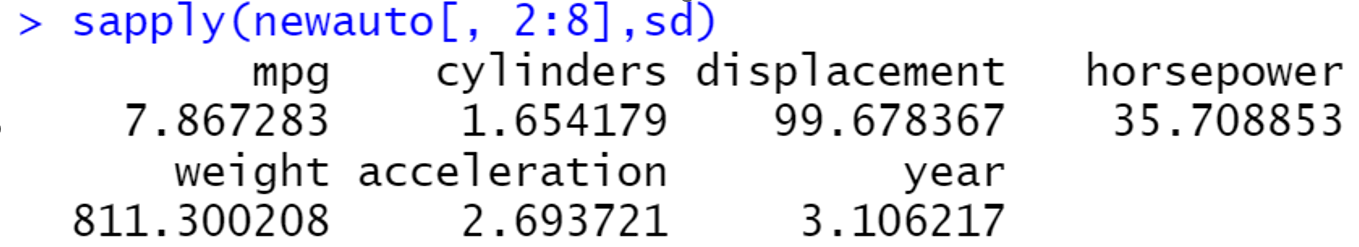
均值：

sapply(newauto[, 2:8],mean)



标准差：

sapply(newauto[, 2:8],sd)



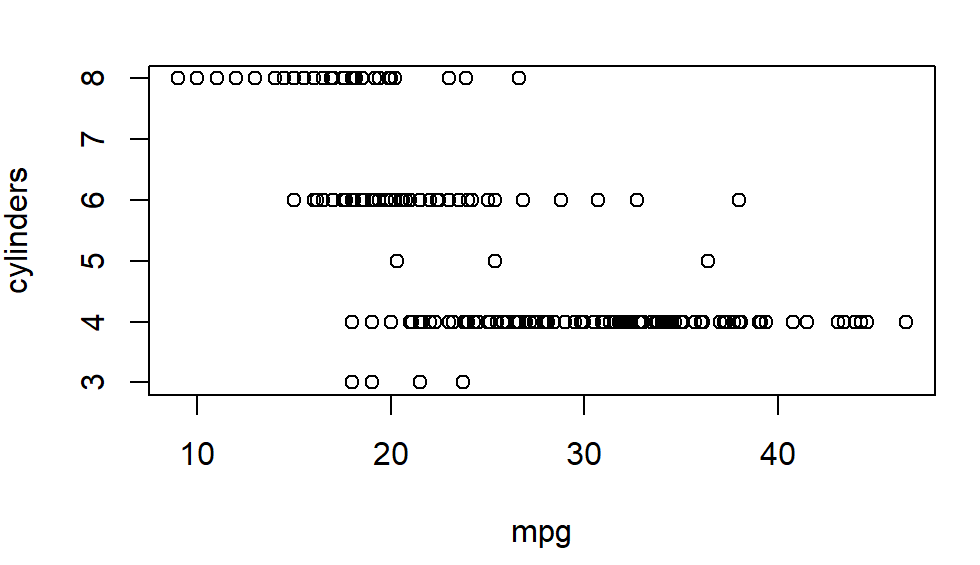
（e）问题（略）

脚本：

par(mfrow=c(1,1))

plot(mpg,cylinders)

截图：



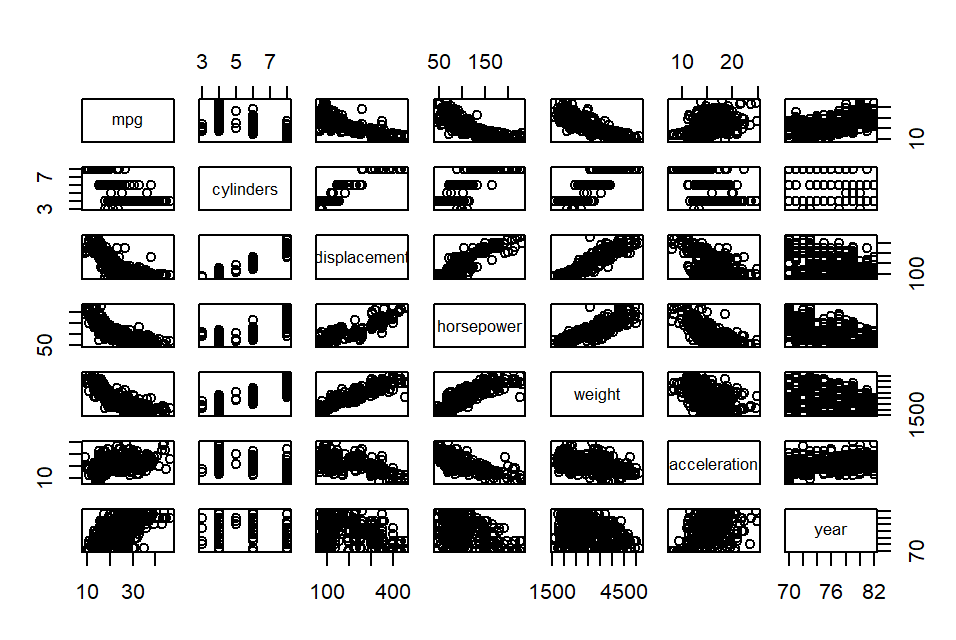
更多的气缸，更少的mpg，二者成负相关。

（f）问题（略）

脚本：

pairs(auto[,2:8])

截图：

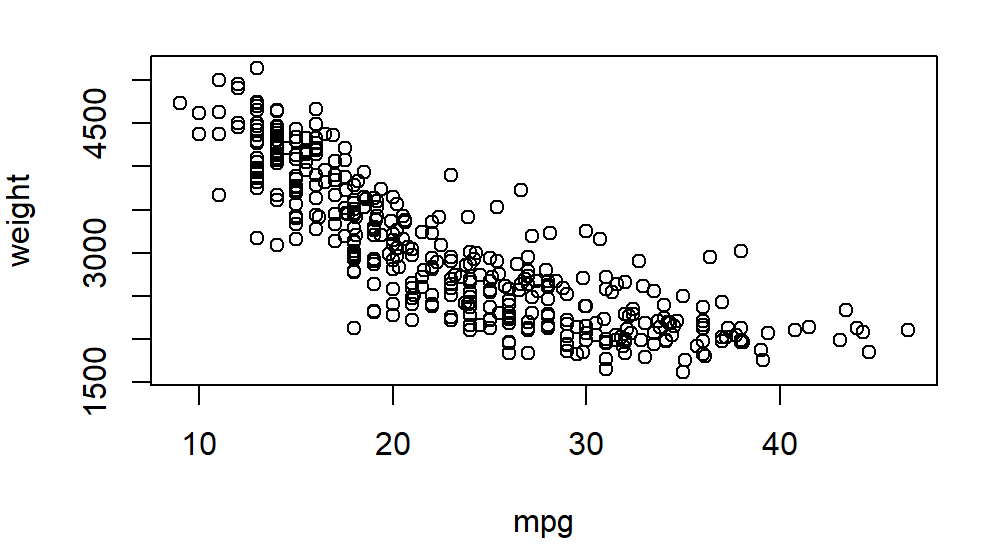


从散点图看，weight与mpg有着较好的线性关系。

脚本：

plot(mpg,weight)

结果：



可以看出，在mpg小于25左右的样子，有着良好的线性关系，而大于25的部分拟合效果会变差。

10.问题（略）

（a）问题（略）

脚本：

library(MASS)

?Boston

?dim(Boston)

结果：



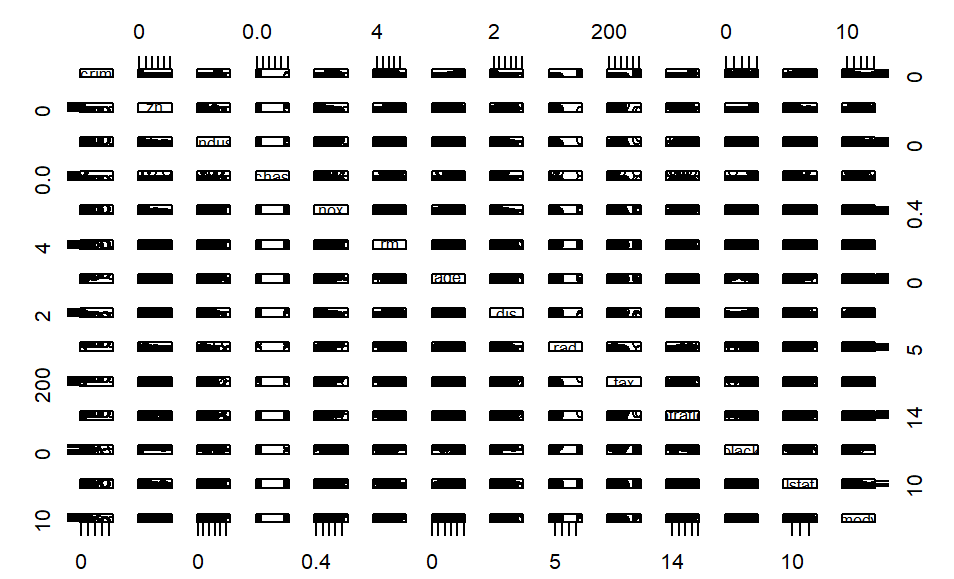
一共统计了506行， 14列。

（b）问题（略）

脚本：

pairs(Boston)

结果：



可以得到：

crim与年龄，dis，rad，tax，ptratio有关

zn与indus, nox, age, lstat有关

indus与年龄，dis有关

nox与年龄，dis有关

dis与lstat有关

lstat与医疗有关

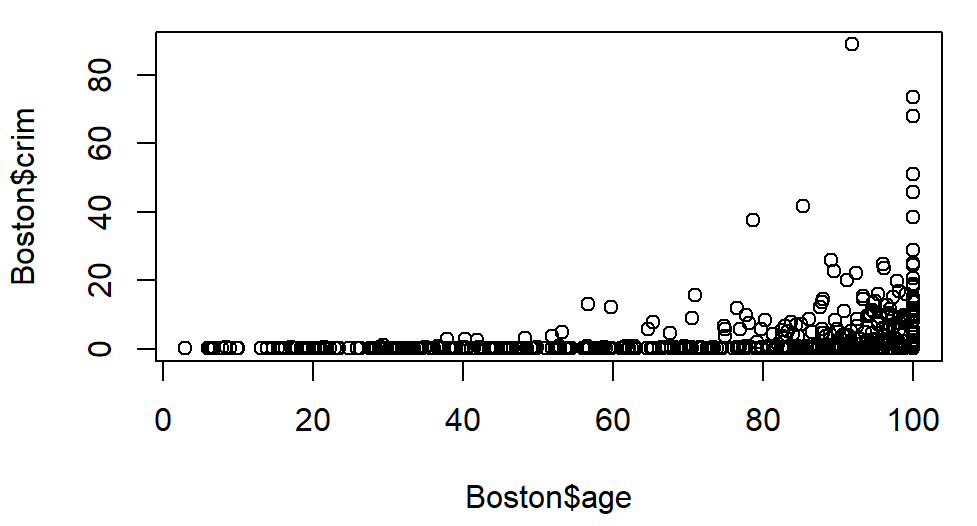
（c）问题（略）

将上一问与crim有关的变量做散点图。

脚本：

plot(Boston$age, Boston$crim)

截图：

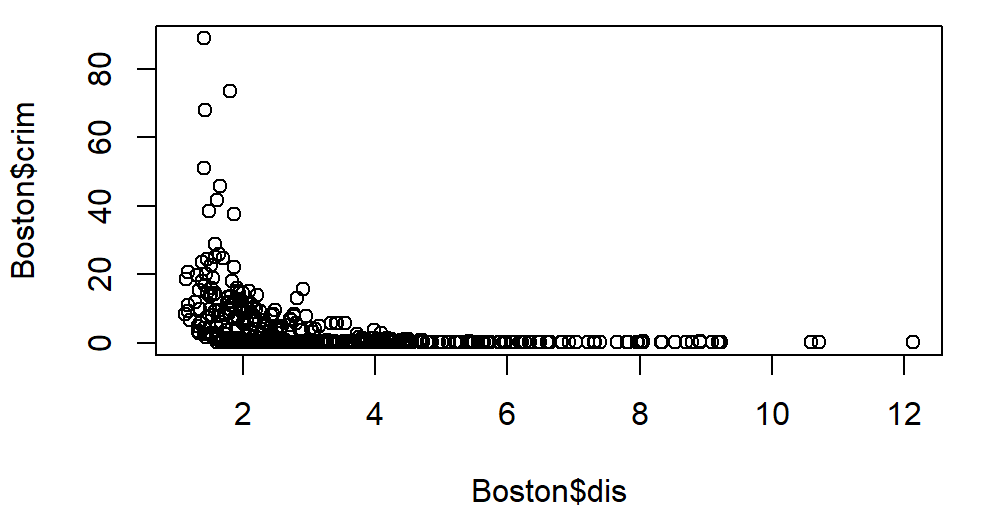


年份越老的郊区的犯罪比例越高，呈正相关。

脚本：

plot(Boston$dis, Boston$crim)

截图：

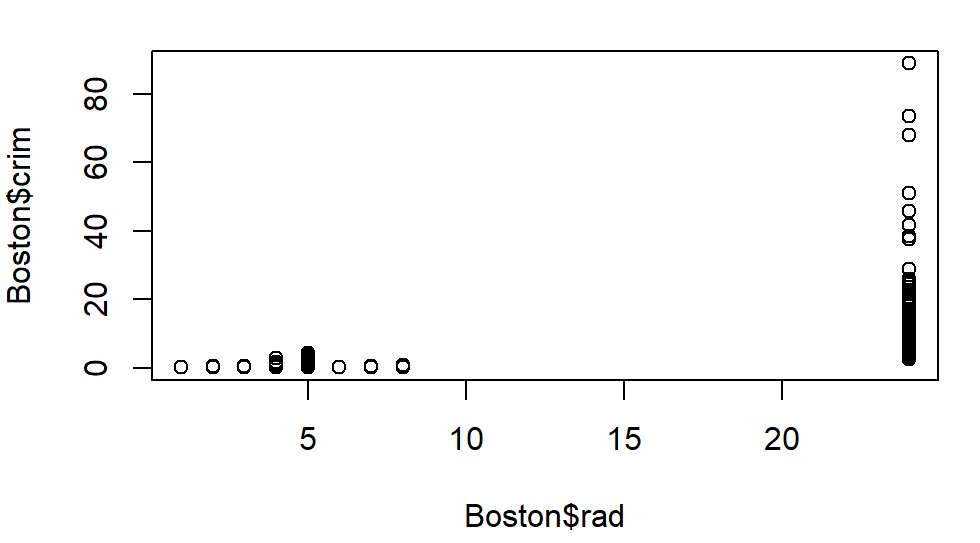


到五个就业中心的加权平均距离越近，犯罪率越高，呈负相关。

脚本：

plot(Boston$rad, Boston$crim)

截图：

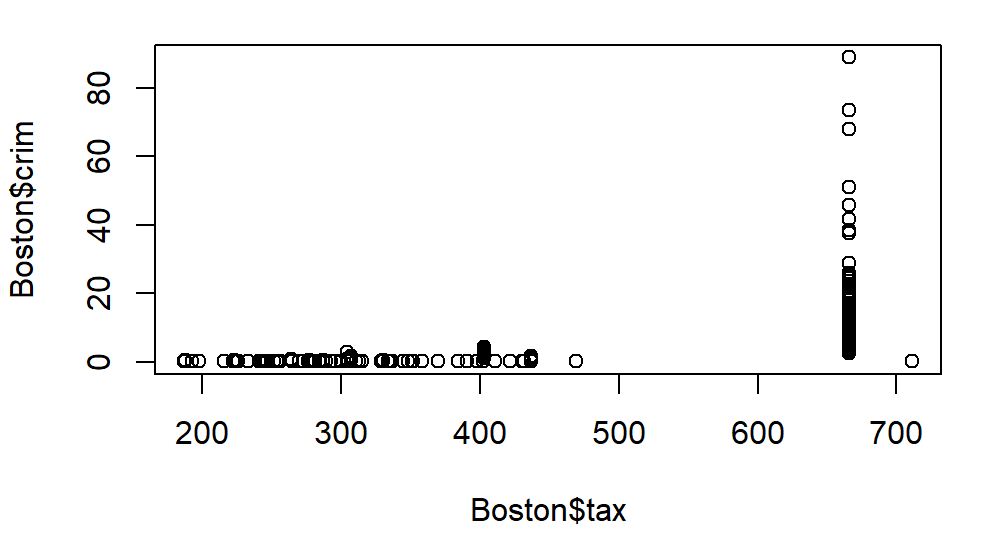


辐射状高速公路的可达性指数越高，犯罪率越高，呈正相关。

脚本：

plot(Boston$tax, Boston$crim)

截图：

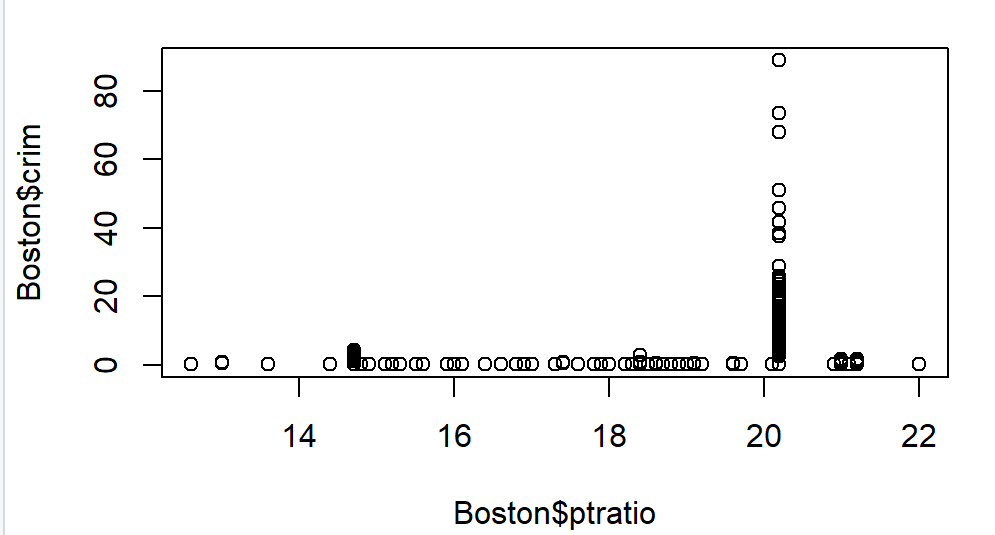


更高的税率，更多的犯罪，呈正相关。

脚本：

plot(Boston$ptratio, Boston$crim)

截图：



学生比老师比例越高，犯罪率越高，呈正相关。

（d）问题（略）

脚本：

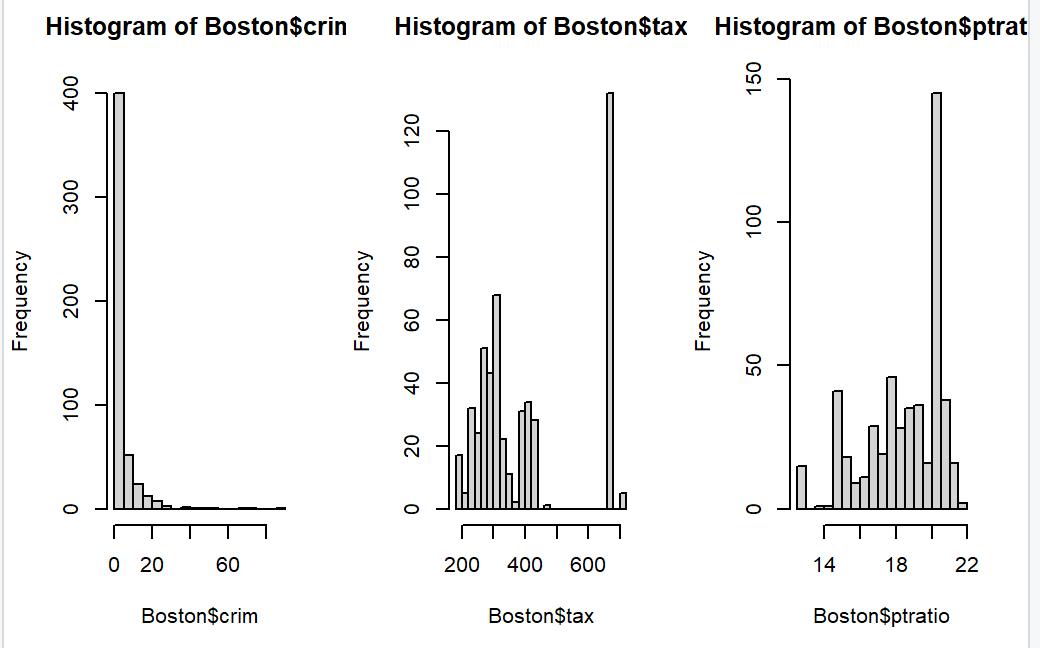
par(mfrow=c(1,3))

hist(Boston$crim, breaks=25)

hist(Boston$tax, breaks=25)

hist(Boston$ptratio, breaks=25)

截图：



绝大部分街区都有着很低的人均犯罪率，有少数地区人均犯罪率大于20%，且有极少数街区接近80%；

一部分街区税率都在200到400之间，另一部分集中在650到700之间。

大部分街区的学生比老师的比例集中在20附近。

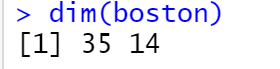
（e）问题（略）

脚本：

boston = subset(Boston, chas == 1)

dim(boston)

截图：



有35个街区在查尔斯河岸附近。

（f）问题（略）

脚本：

median(Boston$ptratio)

截图：

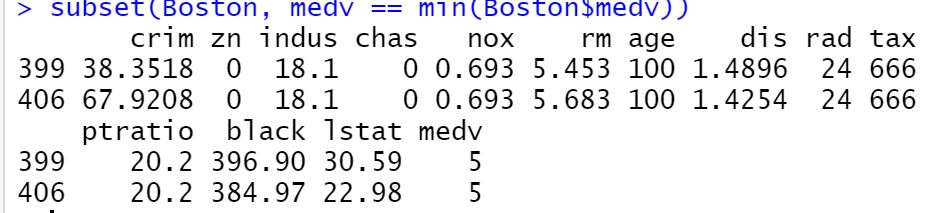


（g）问题（略）

脚本：

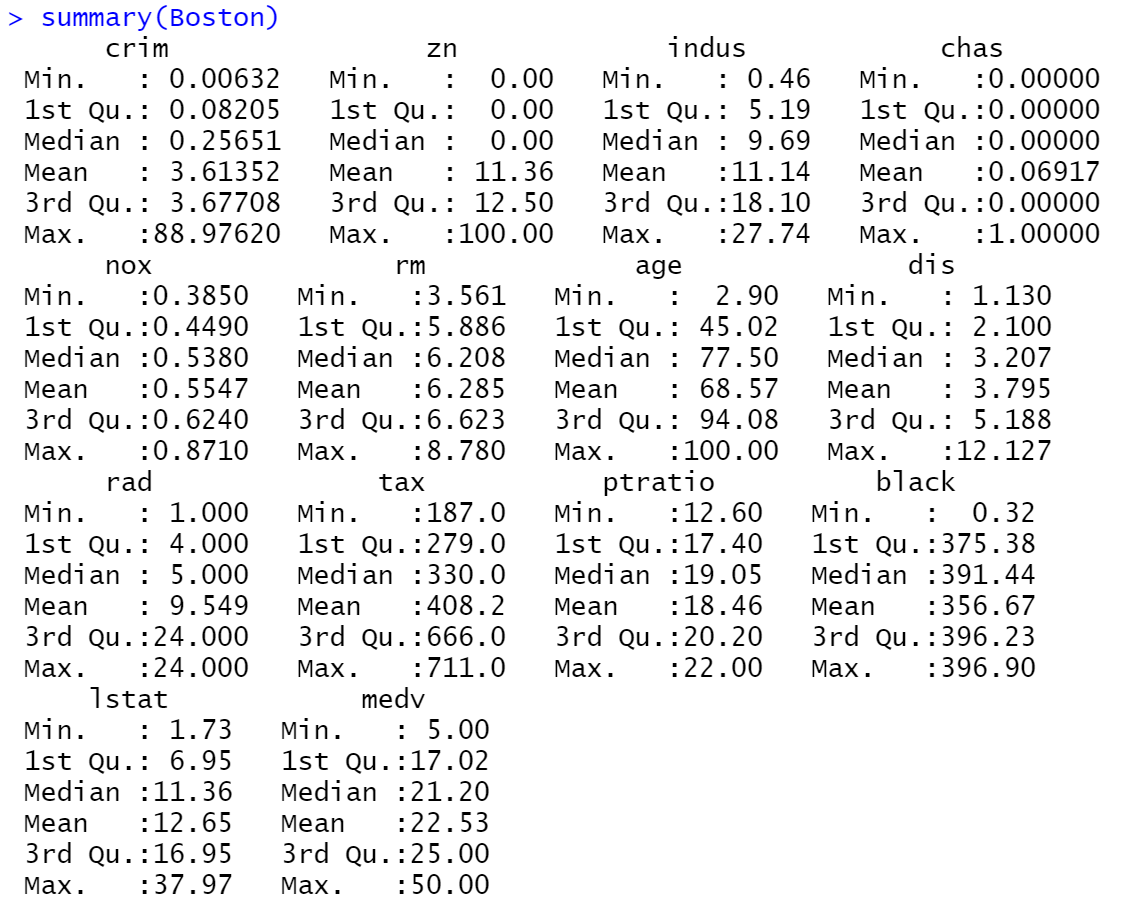
subset(Boston, medv == min(Boston$medv))

截图：



两个郊区是399和406.

summary(Boston)



与总体水平对比：

crim均高于3号四分位点；

zn均位于最小值；

indus均位于3号四分位点附近；

chas均不位于查尔斯河岸附近；

nox均高于3号四分位点；

rm均低于1号四分位点；

age均位于最大值；

dis均低于1号四分位点；

rad均位于最大值；

tax均位于3号四分位点附近；

ptratio均位于3号四分位点附近；

black均高于1号四分位点；

lstat均高于3号四分位点；

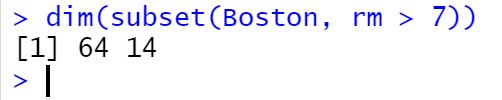
medv均位于最小值；

（h）问题（略）

脚本：

dim(subset(Boston, rm > 7))

截图：

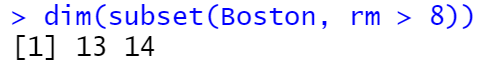


有64个郊区居民平均居住房间超过7。

脚本：

dim(subset(Boston, rm > 8))

截图：



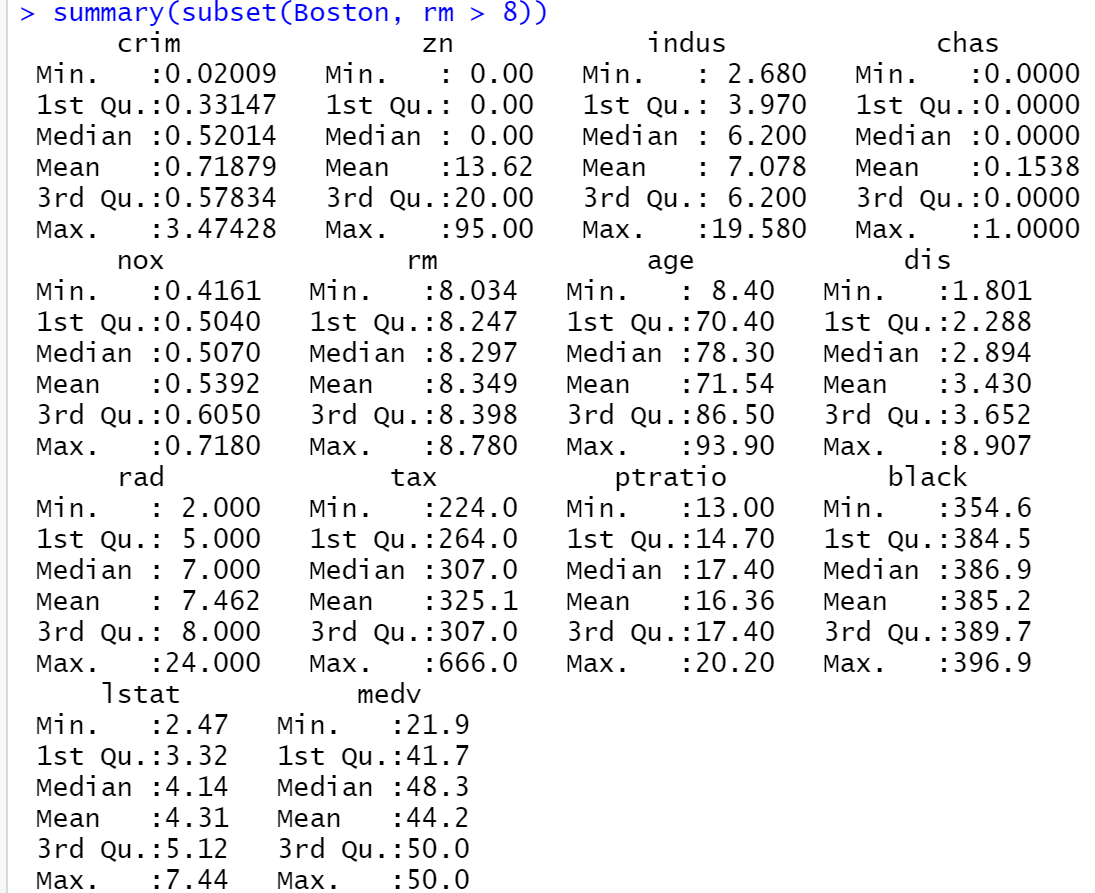
有13个郊区居民平均居住房间超过8。

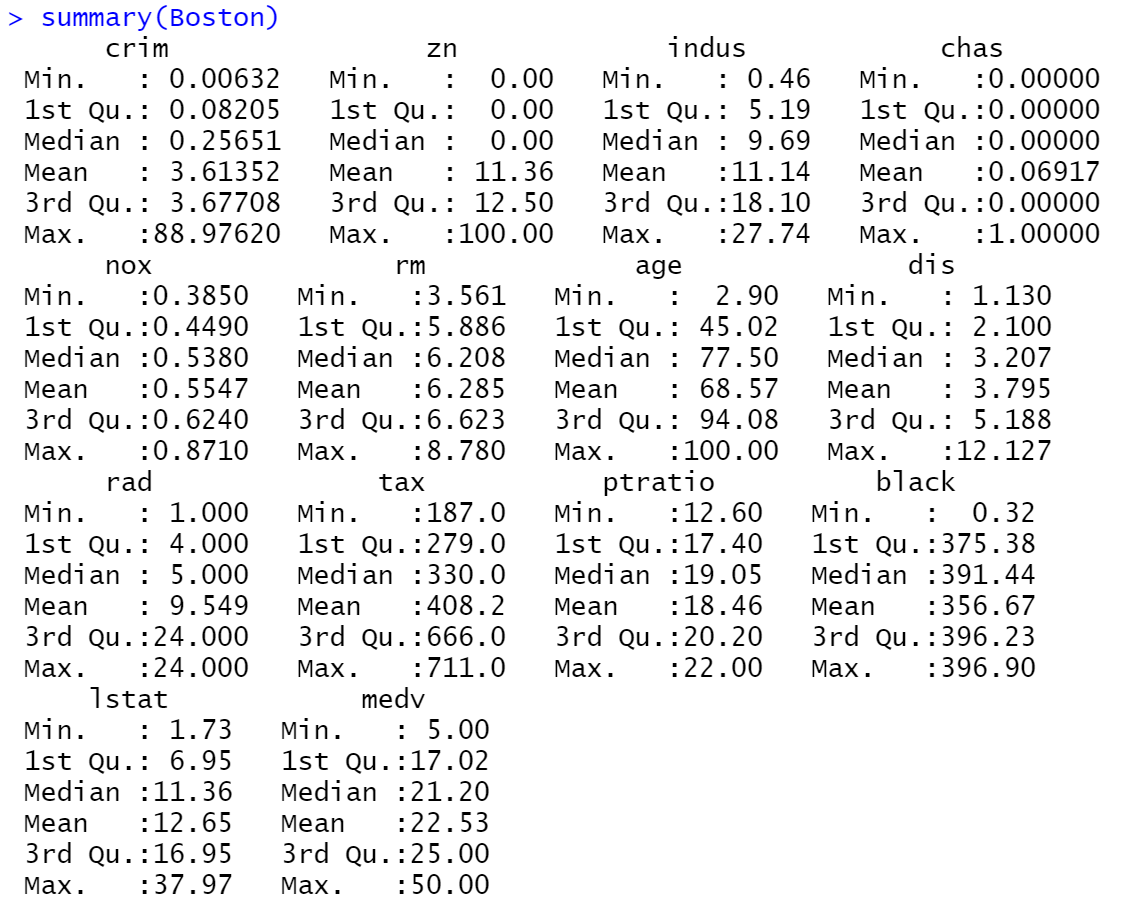
脚本：

summary(subset(Boston, rm > 8))

summary(Boston)

截图：





通过对比可得：

居民平均居住房间超过8的郊区，犯罪率更低，低地位人口比例更低，自住房屋的中值更高，其他特征差异不大。