8.

（a）

代码：

#(a)

college <- read.csv("F:/学习资料/高级统计方法/作业/College.csv")

（b）

代码：

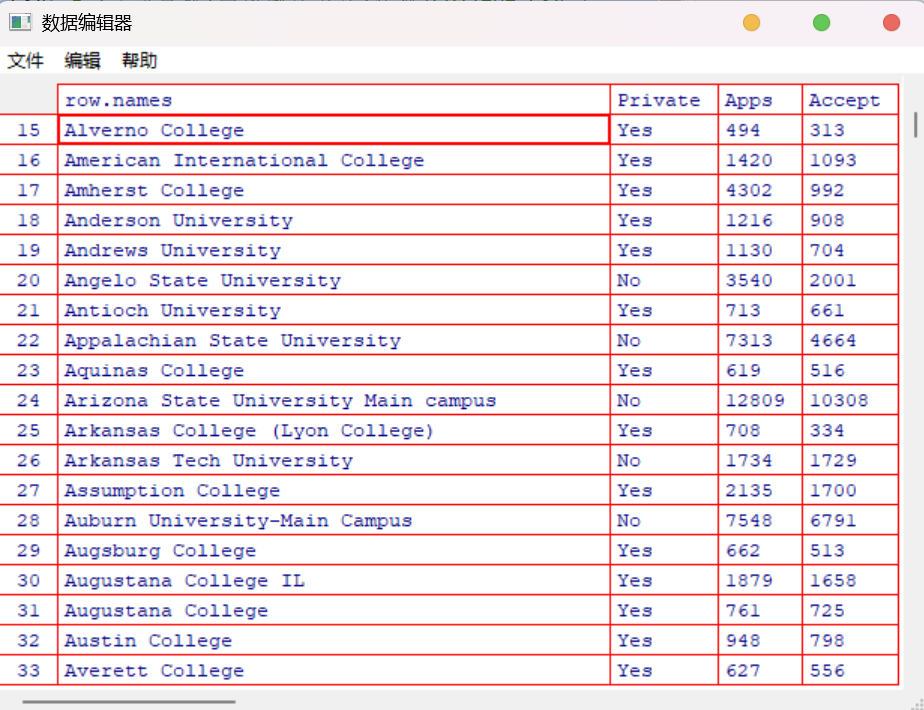
#(b)

rownames(college) <- college[,1]

college <- college[,-1]

fix(college)

运行结果：



（c）

代码：

#1

summary\_college <- summary(college)

summary\_college

#2

pairs(college[,2:10])

#3

ggplot(college, aes(x=college$Private, y=college$Outstate)) +

geom\_boxplot() + labs(x = "学校类型", y = "外州学生学费", title = "公立与私立学校的外州学费比较")

#4

college = read.csv("F:/学习资料/高级统计方法/作业/College.csv")

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

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

Elite = as.factor(Elite)

college = data.frame(college,Elite)

summary\_Elite = summary(college$Elite)

summary\_Elite

plot(college$Elite, college$Outstate)

#5

par(mfrow = c(2, 2))

hist(college$Expend, breaks = 20, col = 2)

hist(college$Grad.Rate, breaks = 20, col = 3)

hist(college$Room.Board, breaks = 20, col = 4)

hist(college$Books, breaks = 20, col = 5 )

par(mfrow = c(1, 1))

#6

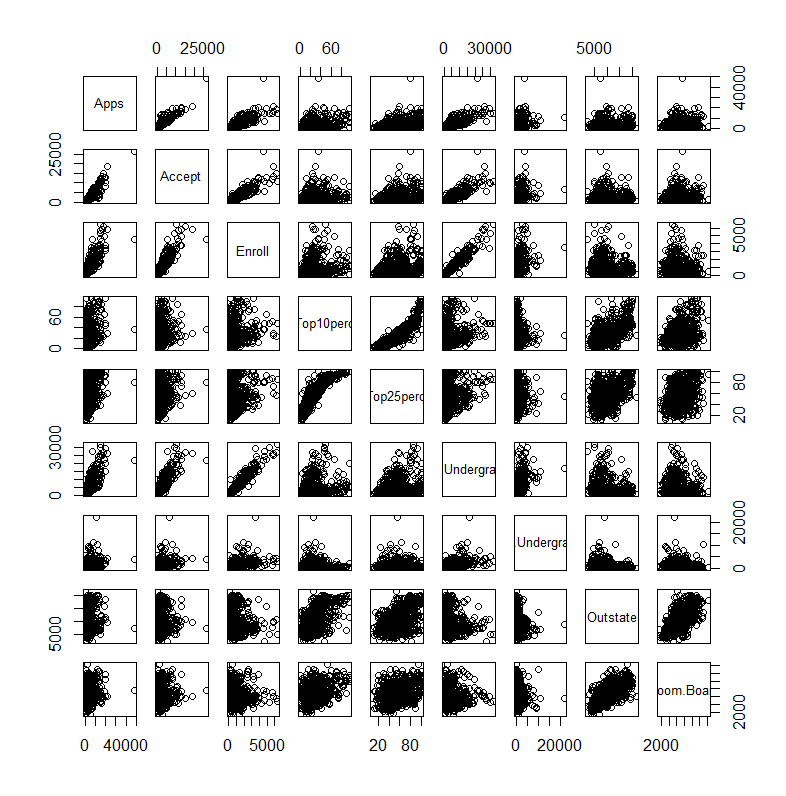
par(mfrow = c(1,2))

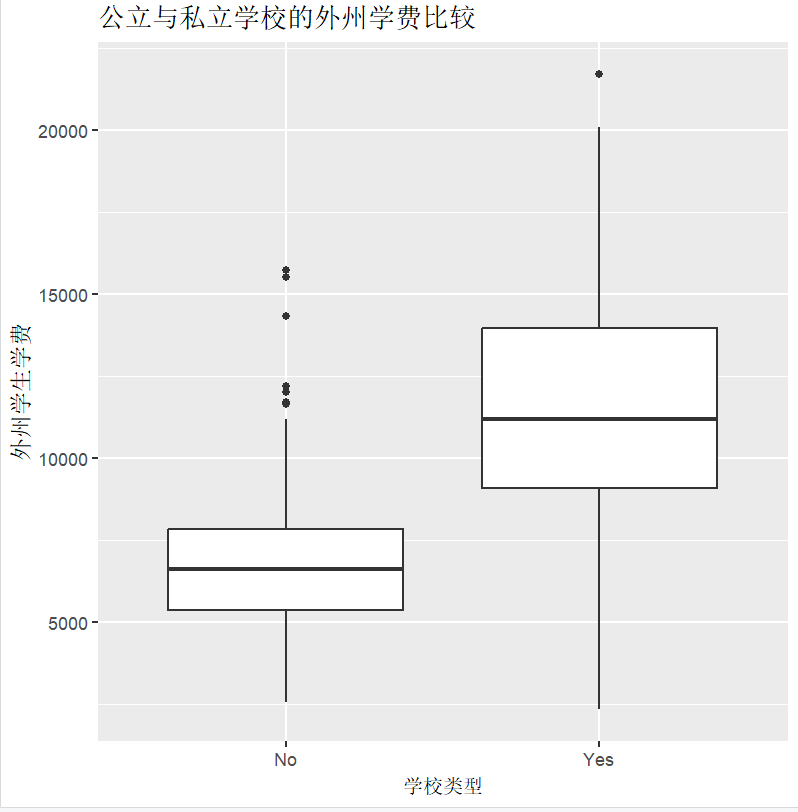
plot(college$Accept,college$Enroll)

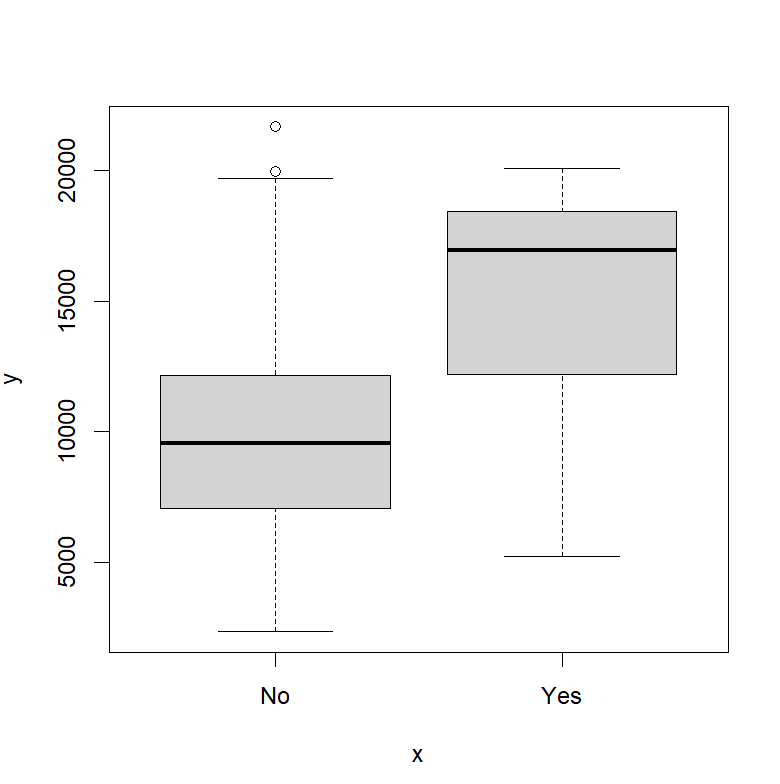
plot(college$Top25perc,college$Top10perc)

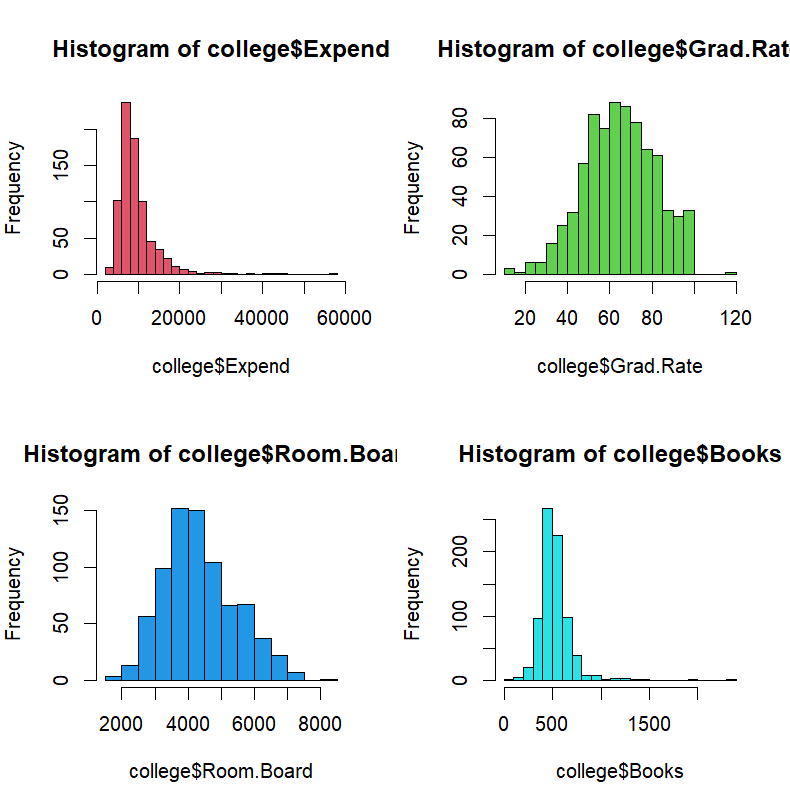
par(mfrow = c(1,2))

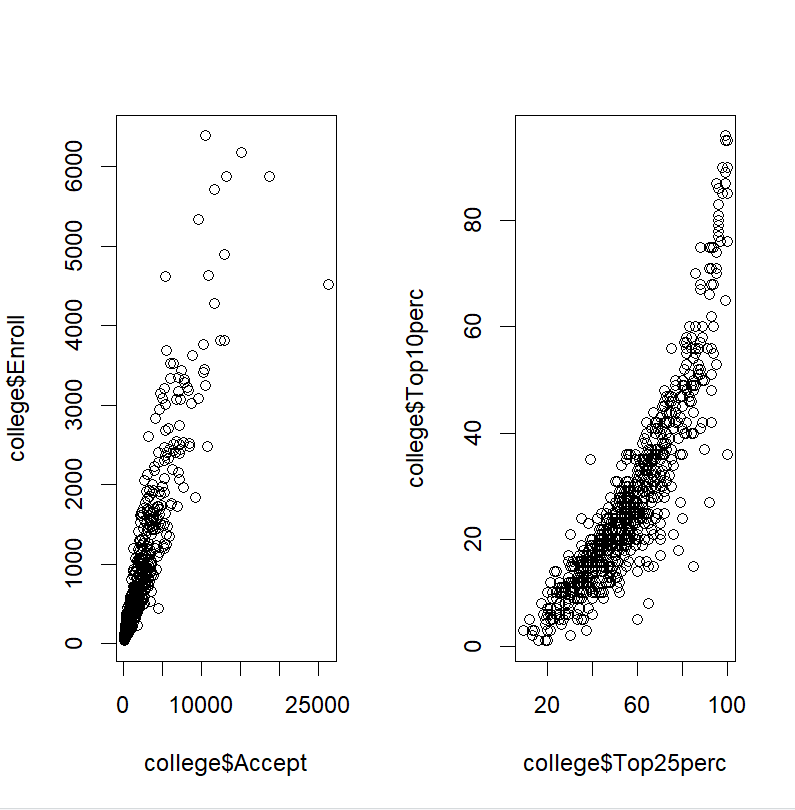
运行结果：











结论：

1、一般来说，录取的学生越多，入学的学生也越多，所以应该存在正相关关系。但是，入学率（Enroll/Accept）因学校而异。一些学校可能录取很多学生但只有少数入学，而另一些学校可能录取较少但入学率较高。

2、点应该沿着一条从左下到右上的趋势线分布。如果点分散，表明入学率变化很大。如果点聚集在一条线附近，表明入学率相对稳定。

3、第二张图探索新生中来自前25%高中毕业生的比例与前10%高中毕业生的比例之间的关系。由于前10%的学生是前25%的子集，我们预期看到一个强烈的正相关关系。即，拥有更多前25%学生的学校，通常也拥有更多前10%的学生。

9.

（a）

定量预测变量有：mpg（每英里油耗）、cylinders（气缸数）、displacement（排量）、horsepower（马力）、weight（重量）、acceleration（加速度）、year（年份）

定性预测变量有：origin（产地，分为美国、欧洲、日本三类）；name为车型标识（唯一值，一般不用于预测分析）

（b）

mpg range: 9 46.6

cylinders range: 3 8

displacement range: 68 455

horsepower range: 46 230

weight range: 1613 5140

acceleration range: 8 24.8

year range: 70 82

（c）

mpg mean: 23.44592 variance: 60.91814

cylinders mean: 5.471939 variance: 2.909696

displacement mean: 194.412 variance: 10950.37

horsepower mean: 104.4694 variance: 1481.569

weight mean: 2977.584 variance: 721484.7

acceleration mean: 15.54133 variance: 7.611331

year mean: 75.97959 variance: 13.56991

（d）

mpg range: 9 to 46.6 ; mean: 23.49436 ; sd: 7.795198

cylinders range: 3 to 8 ; mean: 5.458974 ; sd: 1.700479

displacement range: 68 to 455 ; mean: 193.5115 ; sd: 104.1407

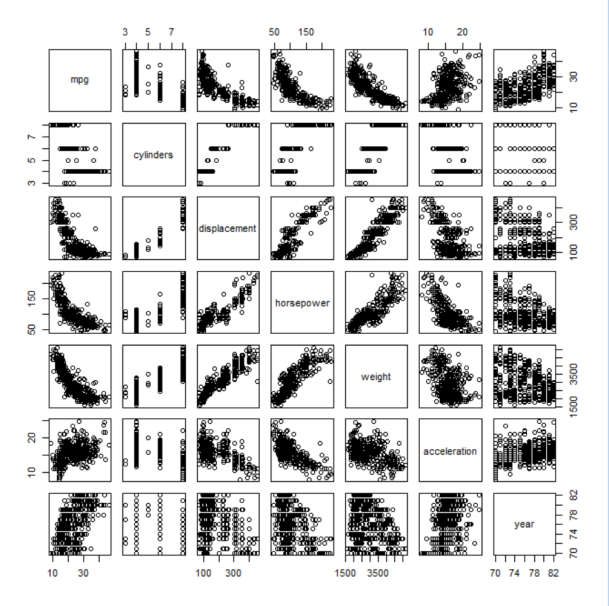
horsepower range: 46 to 230 ; mean: 104.0692 ; sd: 38.17633

weight range: 1613 to 5140 ; mean: 2972.469 ; sd: 848.5121

acceleration range: 8 to 24.8 ; mean: 15.5659 ; sd: 2.739672

year range: 70 to 82 ; mean: 76.00256 ; sd: 3.677556

（e）



mpg与weight：强负相关（重量越大，油耗越低）；

mpg与year：强正相关（年份越新，油耗越低，技术进步）；

mpg与cylinders、horsepower：负相关（气缸数/马力越多，油耗越高）；

其他变量（如displacement与weight）也存在明显正相关。

（f）

从图形和统计量可得到以下线索：

year（正相关）、weight（负相关）、cylinders（负相关）、horsepower（负相关）与mpg存在显著线性趋势。

验证相关性（如cor(Auto[c("mpg", "year", "weight", "cylinders", "horsepower")])）：

mpg与year相关系数≈0.58（强正）；

mpg与weight相关系数≈-0.83（强负）；

mpg与cylinders相关系数≈-0.78（负）；

mpg与horsepower相关系数≈-0.78（负）。

结论：这些变量（尤其是year、weight）可作为预测mpg的有效特征。

10.

（a）

506行，代表波士顿郊区样本

14列，每列代表一个预测特征

（b）

强正相关：如 rm（平均房间数）与 medv（房价中位数），房间数越多，房价越高；dis（到就业中心距离）与 lstat（低地位人口比例），离就业中心越远，低地位人口占比越高。

强负相关：如 lstat与 medv，低地位人口占比越高，房价中位数越低。

非线性/弱相关：部分变量（如 crim与 zn）无明显的线性趋势，需结合领域知识或进一步建模分析。

（c）

正相关显著变量：

rad（高速可达性）：相关系数≈0.6，高速路网密集区域交通便利，可能滋生犯罪活动；

tax（财产税）：相关系数≈0.58，高税率区域经济结构或人口特征（如商业集中、低收入群体多）可能导致犯罪率高；

lstat（低地位人口比例）：相关系数≈0.35，贫困集中区域社会资源匮乏，犯罪风险上升；

nox（氮氧化物浓度）：相关系数≈0.42，工业污染严重区域环境差、社会问题多，犯罪率高。

结论：crim与“社会经济地位（lstat）、工业/交通特征（indus/rad）、环境质量（nox）、政策因素（tax）”等变量正相关，这些因素共同影响犯罪率。

（d）

犯罪率（crim）：临河郊区数量少（仅35个），crim均值低于整体（整体crim均值≈3.61），可能因临河区域多为高端社区，治安更好。

税率（tax）：临河郊区tax均值与整体接近，无明显“特别高”特征。

师生比（ptratio）：ptratio均值略低于整体（整体中位数≈19.05），学校资源可能更优。

（e）

35个

（f）

19.05

（g）

找到medv最小的行对应第399行。

crim=18.72（高犯罪）、indus=27.74（工业用地占比极高）、rad=24（高速路网密集）、tax=666（高财产税）、lstat=30.59（低地位人口占比极高）、rm=5.454（平均房间数极少）、age=94.3（房龄极老）等。这些变量在总体中均处于高百分位（如lstat、crim、indus远高于均值），负面因素叠加导致房价中位数（medv=5.0千美元）极低。

（h）

房间数超过7的郊区数：sum(Boston$rm > 7)→ 约64个；超过8的郊区数：sum(Boston$rm > 8)→ 约13个。

超过8个房间的郊区特征：

房价：中位数远高于整体（多在30千美元以上），因大户型房屋多为高端住宅；

社会经济：lstat（低地位人口比例）低、ptratio（师生比）小（学校资源优）、dis（到就业中心距离）近（通勤便利）、nox（氮氧化物浓度）低（环境好）、crim（犯罪率）低（社区治安优）；

产业结构：indus（工业用地比例）低，以住宅、商业为主。