《数理统计分析》

实验报告

|  |  |
| --- | --- |
| 实验名称 | 数据处理 |
| 班 级 | 221 |
| 学生姓名 | 高鑫宇 |
| 学生学号 | 2022131012 |
| 指导教师 |  |
| 成 绩 |  |

2024年11月16日

成都信息工程大学 人工智能学院

**一、实验目的：**

学习、理解基于R语言数据处理和分析的基本操作，和对该内容的具体实践。

**二、实验环境：**

Windows10操作系统、R、RStudio

**三、实验步骤：**

1. 根据以下要求，针对上次实验生成的电影数据集（电影编号和学号结尾两位数字相同），对数据进行处理和分析：

1）. 使用下面的函数，观察评分和时长变量数据性质。

文本

描述已自动生成

代码：

EDA <- function(x) {

par(mfrow=c(2,2), mar=c(4,3,2,1), cex=0.8)

hist(x, main='Histgram', ylab='')

plot(x, main='Dotchart', ylab='')

boxplot(x, horizontal=T, main='Boxplot')

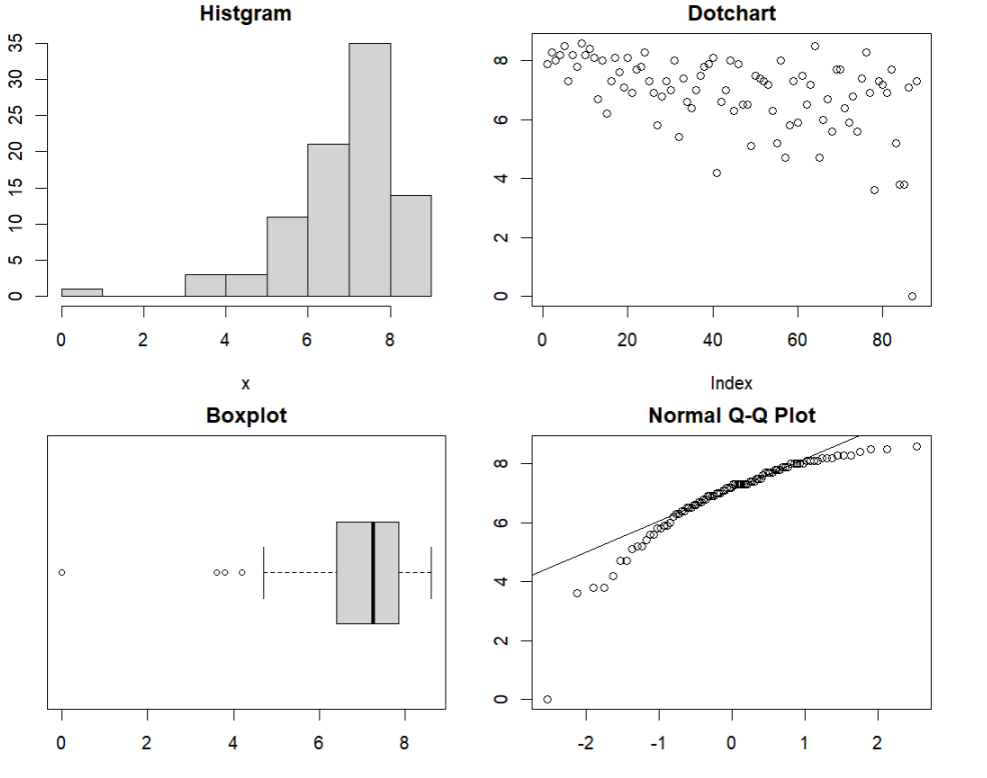
qqnorm(x, xlab='', ylab='');qqline(x)

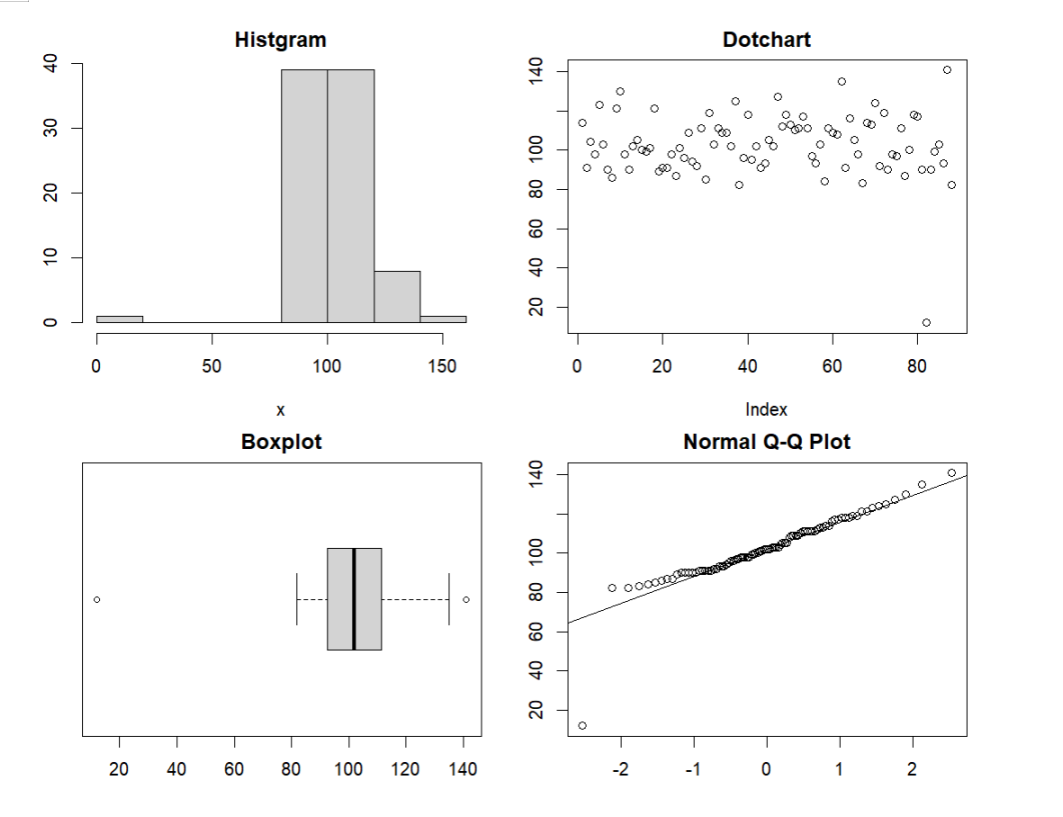
par(mfrow=c(1,1))

}

EDA(df$avgRate)

EDA(df$runtime)





2）. 结合下面的代码和图片，分析观察评分与评分人数、评分与五星人数之间的关系，并依据评分分组数据进行显示。

图表, 散点图

描述已自动生成

代码：

with(df, {

par(mfrow=c(1,2), cex=0.8) # 设置1行2列的图形区域

# 评分与评分人数的关系

plot(avgRate, rateNum, main="评分与评分人数关系", xlab="评分", ylab="评分人数", col='blue', pch=19)

# 评分与五星人数的关系

plot(avgRate, five, main="评分与五星人数关系", xlab="评分", ylab="五星人数", col='red', pch=17)

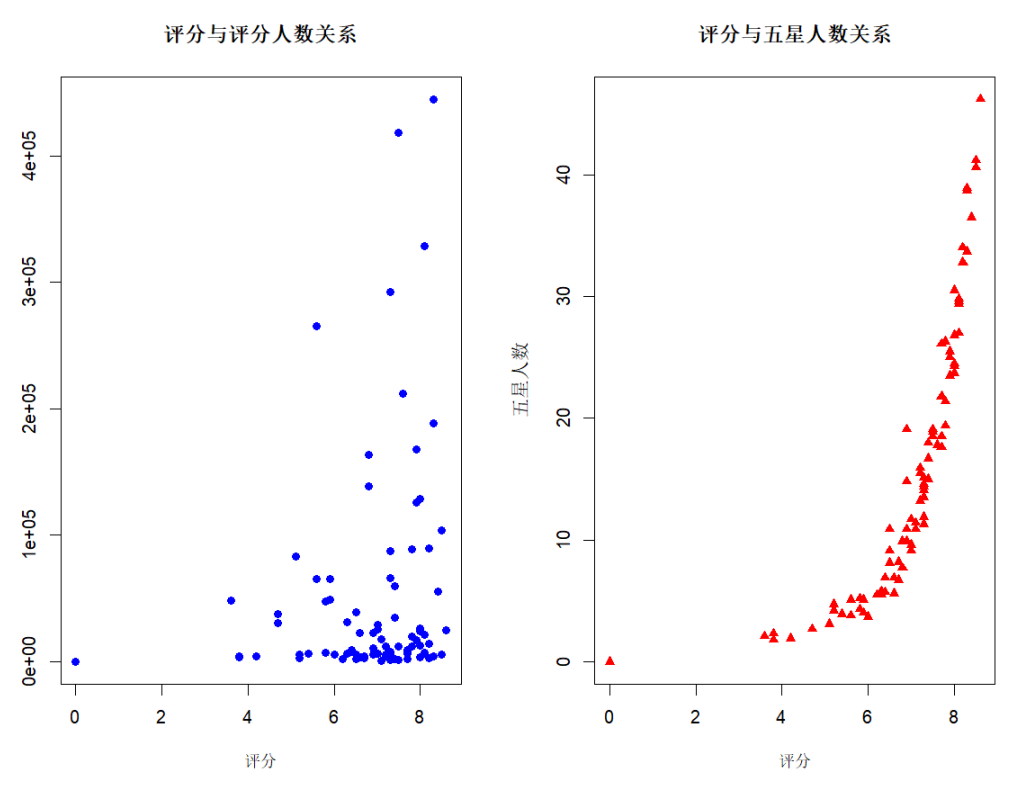
# 通过legend()函数为分组加上图例

# 假设有一个评分等级的分组字段，例如：评分为1星、2星等

legend(3, max(rateNum), levels(as.factor(avgRate)), pch=1:3, bty='n', title="评分分组")

par(mfrow=c(1,1)) # 恢复为默认单图显示

})



3）. 使用table函数统计每种语言的电影数，并使用barplot，pie绘制图形（左右排列），按不同颜色（非默认）区分。

代码：

language\_count <- table(df$language)

par(mfrow=c(1, 2), mar=c(4, 4, 2, 1))

# 绘制条形图，设置不同颜色

barplot(language\_count,

main="每种语言的电影数",

col=rainbow(length(language\_count)),

ylab="电影数",

xlab="语言",

border="white")

# 绘制饼图，设置不同颜色

pie(language\_count,

main="每种语言的电影比例",

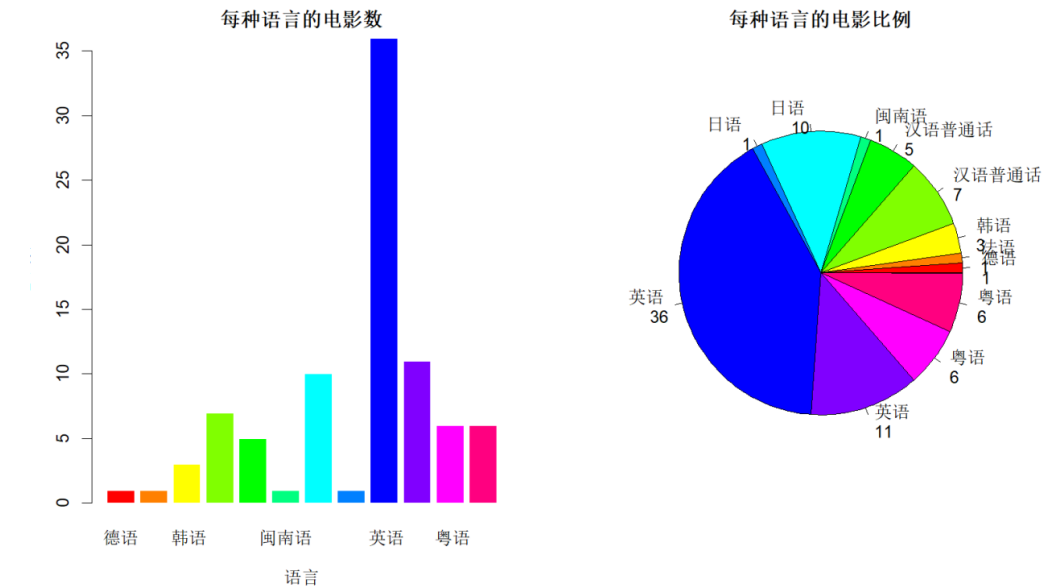
col=rainbow(length(language\_count)),

labels=paste(names(language\_count), "\n", language\_count),

radius=1.2)

# 恢复为单个图形窗口

par(mfrow=c(1, 1))



4）. 针对3生成的数据，求解每种语言的电影数占比，并将生成的数据和3中的数据左右拼接。

代码：

Ftab <- function(x) {

f = table(x)

s = sum(f)

p = f / s \* 100

t = cbind('Counts' = f, 'Percent(%)' = round(p, 2))

print(rbind(t, 'Total' = c(s, 100)))

invisible(f)

}

language\_count <- table(df$language)

total\_count <- sum(language\_count)

percentages <- language\_count / total\_count \* 100

# 创建一个数据框，包含语言、电影数和百分比

language\_data <- data.frame(

Language = names(language\_count),

Counts = as.integer(language\_count),

Percent = round(percentages, 2)

)

# 设置画布，按左右排列

par(mfrow=c(1, 2), mar=c(4, 4, 2, 1))

# 绘制条形图，设置不同颜色

barplot(language\_count,

main="每种语言的电影数",

col=rainbow(length(language\_count)),

ylab="电影数",

xlab="语言",

border="white")

# 在条形图上添加百分比标签

text(x = seq\_along(language\_count),

y = language\_count,

labels = paste(round(percentages, 1), "%"),

pos = 3, cex = 0.8, col = "black")

# 绘制饼图，设置不同颜色

pie(language\_count,

main="每种语言的电影比例",

col=rainbow(length(language\_count)),

labels = paste(names(language\_count), "\n", round(percentages, 1), "%"),

radius=1.2)

# 恢复为单个图形窗口

par(mfrow=c(1, 1))

# 调用Ftab函数显示电影数和百分比

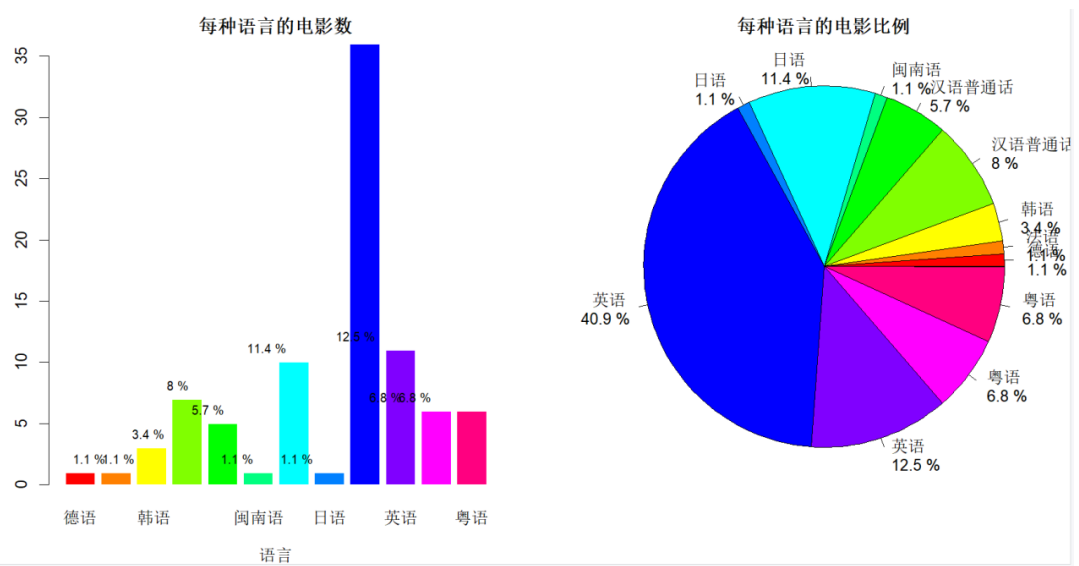
Ftab(df$language)

# 合并原始数据和统计数据

df\_with\_stats <- merge(df, language\_data, by.x = "language", by.y = "Language")

# 查看拼接后的数据框

print(df\_with\_stats)



5）. 试分用mean, median, Var, sd, summary函数求评分变量的相关统计量。

图形用户界面, 文本

描述已自动生成

代码：

mean\_rating <- mean(df$avgRate, na.rm = TRUE) #用于忽略 NA 值

cat("Mean (平均值): ", mean\_rating, "\n")

# 计算评分的中位数

median\_rating <- median(df$avgRate, na.rm = TRUE)

cat("Median (中位数): ", median\_rating, "\n")

# 计算评分的方差

var\_rating <- var(df$avgRate, na.rm = TRUE)

cat("Variance (方差): ", var\_rating, "\n")

# 计算评分的标准差

sd\_rating <- sd(df$avgRate, na.rm = TRUE)

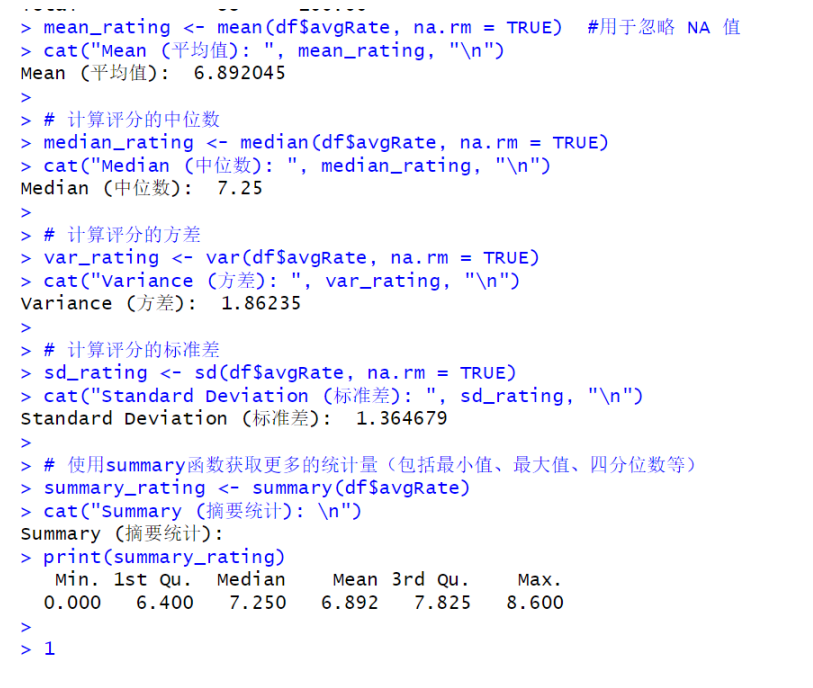
cat("Standard Deviation (标准差): ", sd\_rating, "\n")

# 使用summary函数获取更多的统计量（包括最小值、最大值、四分位数等）

summary\_rating <- summary(df$avgRate)

cat("Summary (摘要统计): \n")

print(summary\_rating)



6）. 试使用下面的函数生成相应数据。将函数生成的直方图与评分分组变量数据的barplot进行对比（相同的分组数目），观察它们的区别和联系。

图形用户界面, 应用程序, Word

描述已自动生成

代码：

Freq <- function(x, Bs=10) { #自定义计量频数表函数

h=hist(x, breaks=Bs, xlab='', main='') # 使用 hist() 函数生成直方图

f=h$counts # 获取频数（每个区间内的数据个数

p=f/sum(f)\*100 # 计算频率百分比：

cp=cumsum(p) # 计算累积频率

freq=data.frame(m=h$mids, f=f, p=p, cp=round(cp,2))

names(freq)=c('Median', 'Counts', 'Frequency(%)', 'CumsumFreq(%)')

freq

}

# 使用 Freq 函数生成频数表

freq\_data <- Freq(df$avgRate, Bs=5)

# 输出频数表

print(freq\_data)

# 绘制直方图

hist(df$avgRate, breaks=5, col="lightblue", main="Histogram of avgRate", xlab="Rating", ylab="Frequency")

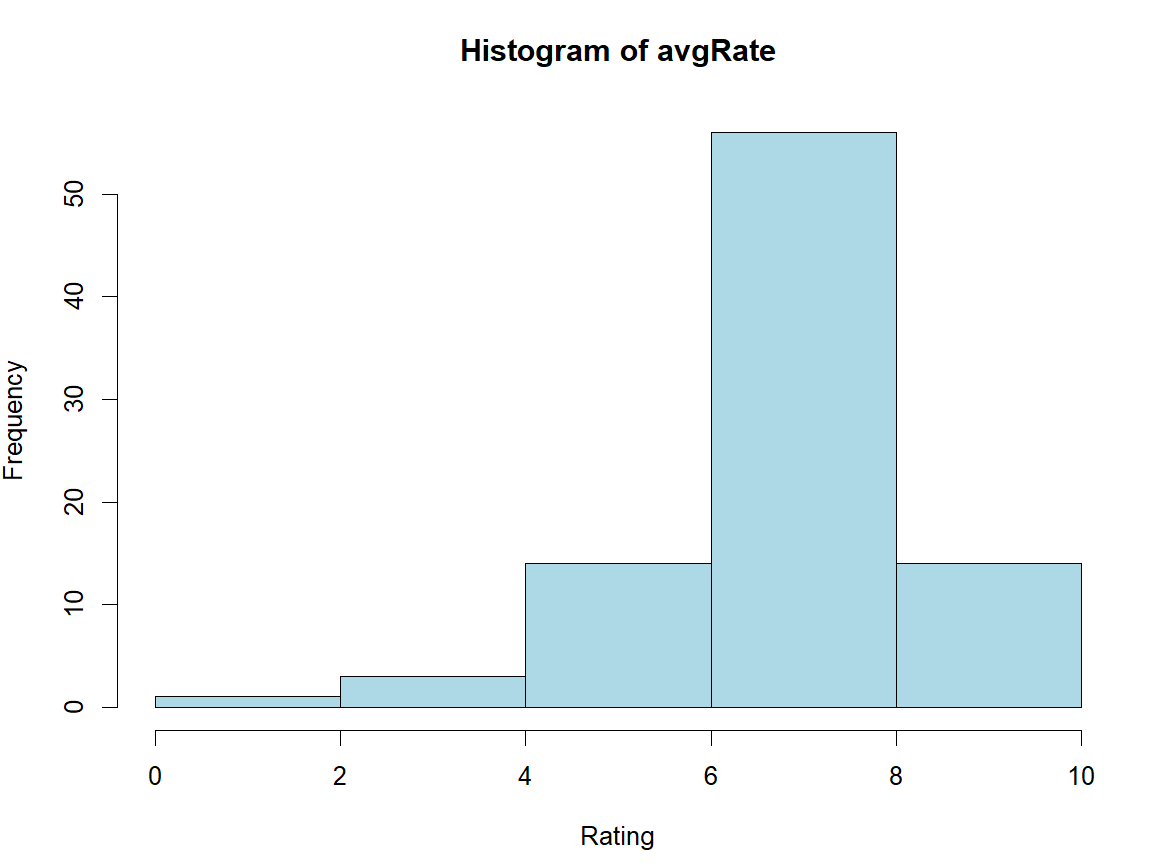
# 使用cut()将评分数据分成5个区间

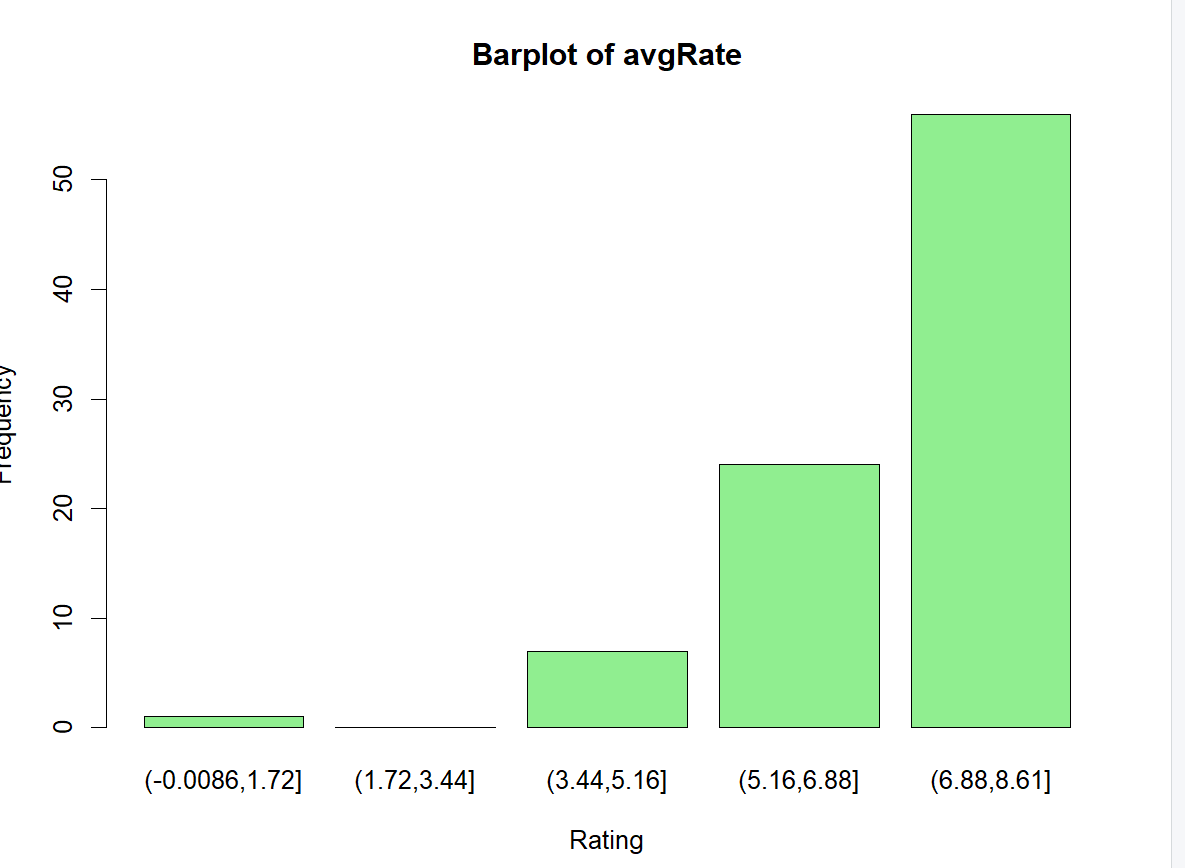
rating\_groups <- cut(df$avgRate, breaks=5)

# 生成评分分组的条形图

rating\_table <- table(rating\_groups)

barplot(rating\_table, col="lightgreen", main="Barplot of avgRate", xlab="Rating", ylab="Frequency")





区别: 直方图适用于连续数据，展示数据在不同区间的分布，条形相连，强调数据的连续性。条形图适用于离散数据，展示不同分组或类别的频数，条形之间有间隙，强调每个分组的独立性。

联系：两者都用于展示数据的频数分布，帮助分析数据集中在哪些区域或类别。都可以显示数据的频率或百分比，只是表现方式不同。

7）. 按照不同的语言分类得出评分数据的的基本统计量（summary函数）和boxplot箱式图和stripchart点带图（对角线排列）。

代码：

summary\_stats <- tapply(df$avgRate, df$language, summary)

print(summary\_stats)

# 绘制箱式图和点带图（对角线排列）

par(mfrow = c(1, 2)) # 设置画布为1行2列

# 绘制箱式图

boxplot(avgRate ~ language, data = df, main = "Boxplot by Language",

xlab = "Language", ylab = "Avg Rating", col = "lightblue")

# 绘制点带图

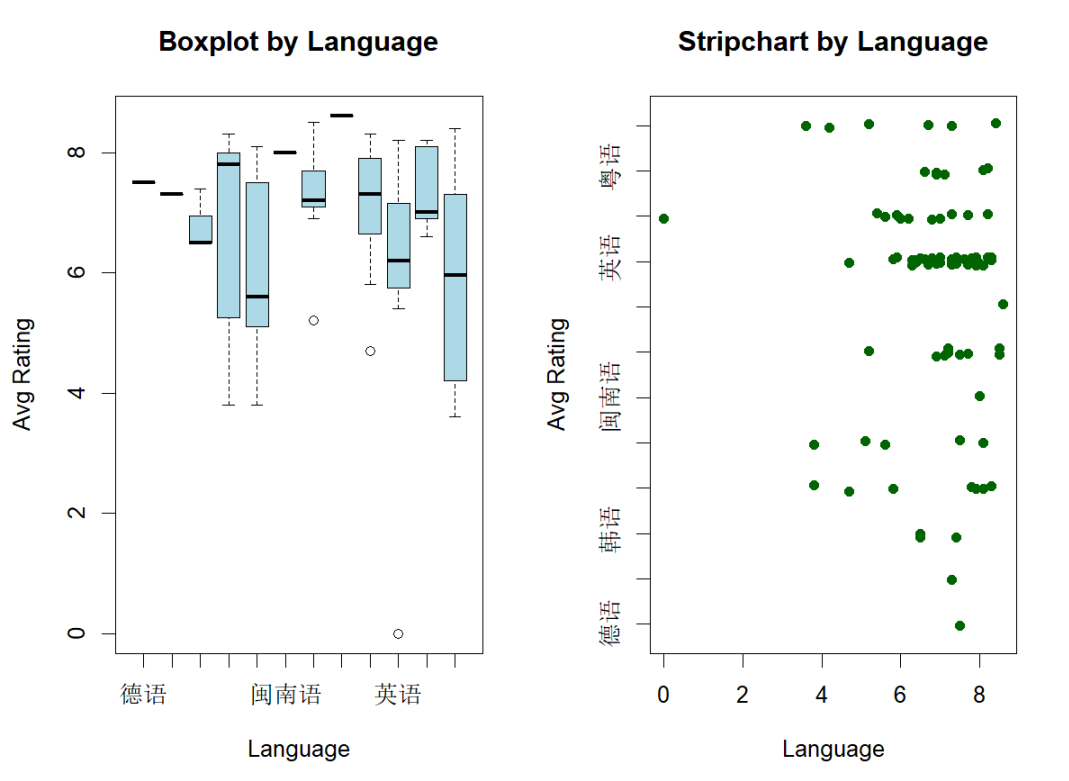
stripchart(avgRate ~ language, data = df, method = "jitter",

main = "Stripchart by Language",

xlab = "Language", ylab = "Avg Rating",

pch = 19, col = "darkgreen")

par(mfrow = c(1, 1)) # 恢复为单个画布



8）. 使用聚合函数求出连续型变量（包括：评分，时长，评分人数，五星、四星、三星、二星、一星）的基本统计量。

代码：

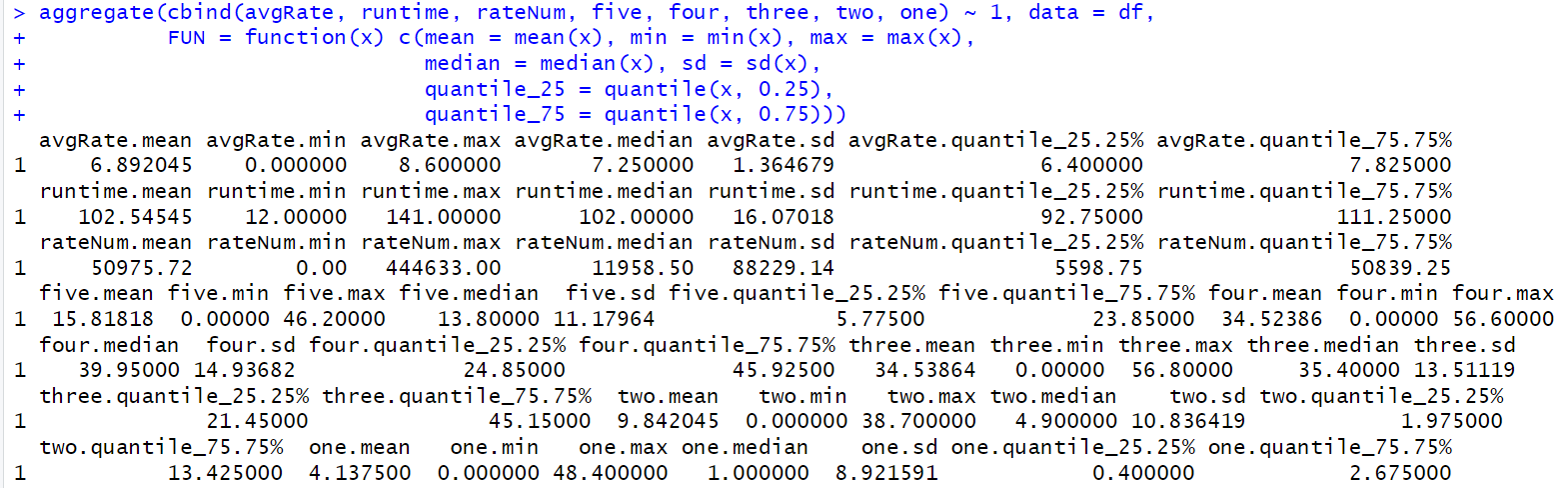
aggregate(cbind(avgRate, runtime, rateNum, five, four, three, two, one) ~ 1, data = df,

FUN = function(x) c(mean = mean(x), min = min(x), max = max(x),

median = median(x), sd = sd(x),

quantile\_25 = quantile(x, 0.25),

quantile\_75 = quantile(x, 0.75)))



9）. 依据评分分组数据和时长分组数据求出评分、评分人数和时长的均值和标准差。

代码：

summary\_by\_rating <- aggregate(cbind(avgRate, rateNum, five, runtime) ~ avgRate,

data = df,

FUN = function(x) c(mean = mean(x), sd = sd(x)))

# 按时长 (runtime) 分组，计算评分、评分人数、五星人数和时长的均值和标准差

summary\_by\_runtime <- aggregate(cbind(avgRate, rateNum, five, runtime) ~ runtime,

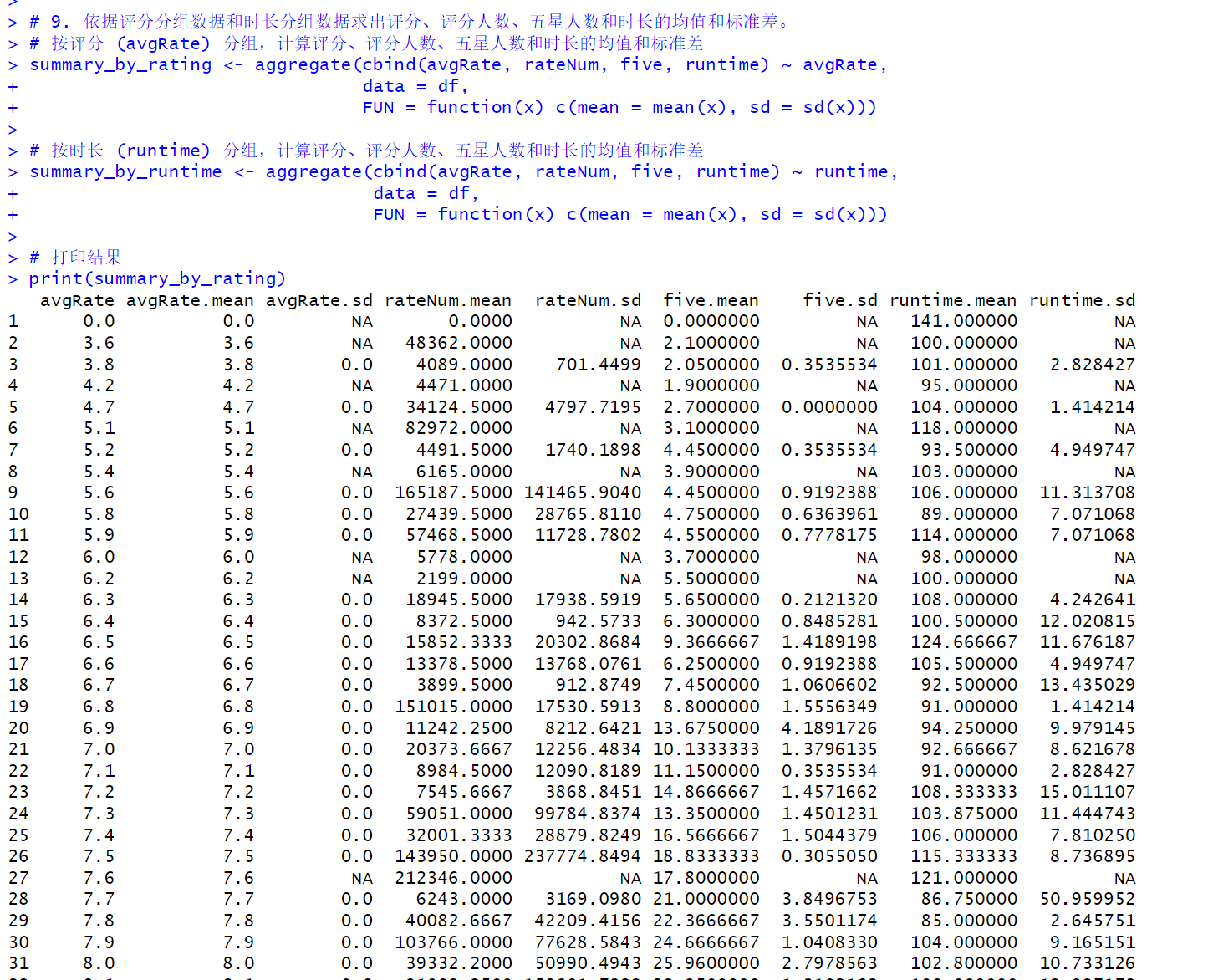
data = df,

FUN = function(x) c(mean = mean(x), sd = sd(x)))

# 打印结果

print(summary\_by\_rating)

print(summary\_by\_runtime)



**四、实验结果：**

1. 完成了什么实验？

本次实验主要集中在基于R语言进行数据处理和分析的基本操作。通过使用 table、barplot、hist、cut 等函数，完成了以下任务：

1).统计数据频数并生成频数表。

2).绘制了不同类型的图形（如条形图、饼图和直方图）。

3).使用 Freq 函数分析评分数据，并将生成的直方图与分组条形图进行对比

2. 实验的结论。

通过这次实验，我对R语言的数据处理和分析有了更深入的理解，尤其是如何使用 table 函数进行数据汇总，如何用 barplot 和 hist 函数可视化数据分布，以及如何通过分组（如 cut 函数）更好地分析离散数据的频数分布

3. 遇到了什么困难，怎么解决的？

数据分组的理解：如何将连续的评分数据合理地分组进行条形图绘制是一个挑战，特别是如何选择合适的分组数目（如区间大小）。

·直方图与条形图的对比：理解直方图和条形图的区别以及它们适用的场景，最初对如何选择合适的图形表示数据不太清晰。

4. 对实验有何认识？

本实验加深了我对数据处理、频数统计和数据可视化的理解。R语言的强大数据处理和可视化能力让我能够高效地从数据中提取信息，并通过图形化的方式清晰展示数据分布，为后续的统计分析和决策提供支持。

附件如下:

