定量分析第二次作业

周青

目录

问题 1:《生活大爆炸》《生活大爆炸》是一部情景喜剧，由约翰尼·盖尔基、吉姆·帕森斯和凯莉·库柯-斯威汀主演，是有线电视网络上最受欢迎的节目之一。2011-2012 年度的第一和第二集于 2011 年 9 月 22 日首播；第一集吸引了 1,410 万观众，第二集吸引了 1,470 万观众。附带的数据文件 “BigBangTheory” 显示了 2011-2012 年度前 21 集的观众人数（《生活大爆炸》网站，2012 年 4 月 17 日）。

a. 计算出观众的最小值和最大值。

b. 计算平均数、中位数和众数。

c. 计算第一四分位数和第三四分位数。

d. 在 2011-2012 赛季期间，收视率是上升还是下降了？请讨论。

0.0.1 a

```
BigBangTheory <- read.csv("D:/MEM/2-定量分析-数据思维与商业统计/MEM_Studying_by_Hoki/2st_assignment_
colnames(BigBangTheory)
```

```
## [1] "Air.Date"           "Viewers..millions."
```

```
max_viewers<-max(BigBangTheory$Viewers..millions.,na.rm = TRUE)
min_viewers<-min(BigBangTheory$Viewers..millions.,na.rm = TRUE)
print(c(max_viewers,min_viewers))
```

```
## [1] 16.5 13.3
```

0.0.2 b

```r
summary(BigBangTheory)
```

```
##    Air.Date         Viewers..millions.
##  Length:21          Min.   :13.30
##  Class :character   1st Qu.:14.10
##  Mode  :character   Median :15.00
##                     Mean   :15.04
##                     3rd Qu.:16.00
##                     Max.   :16.50
```

```r
tab <- table(BigBangTheory$Viewers..millions.)
which.max(tab)
```

```
## 13.6
##    2
```

0.0.3 c

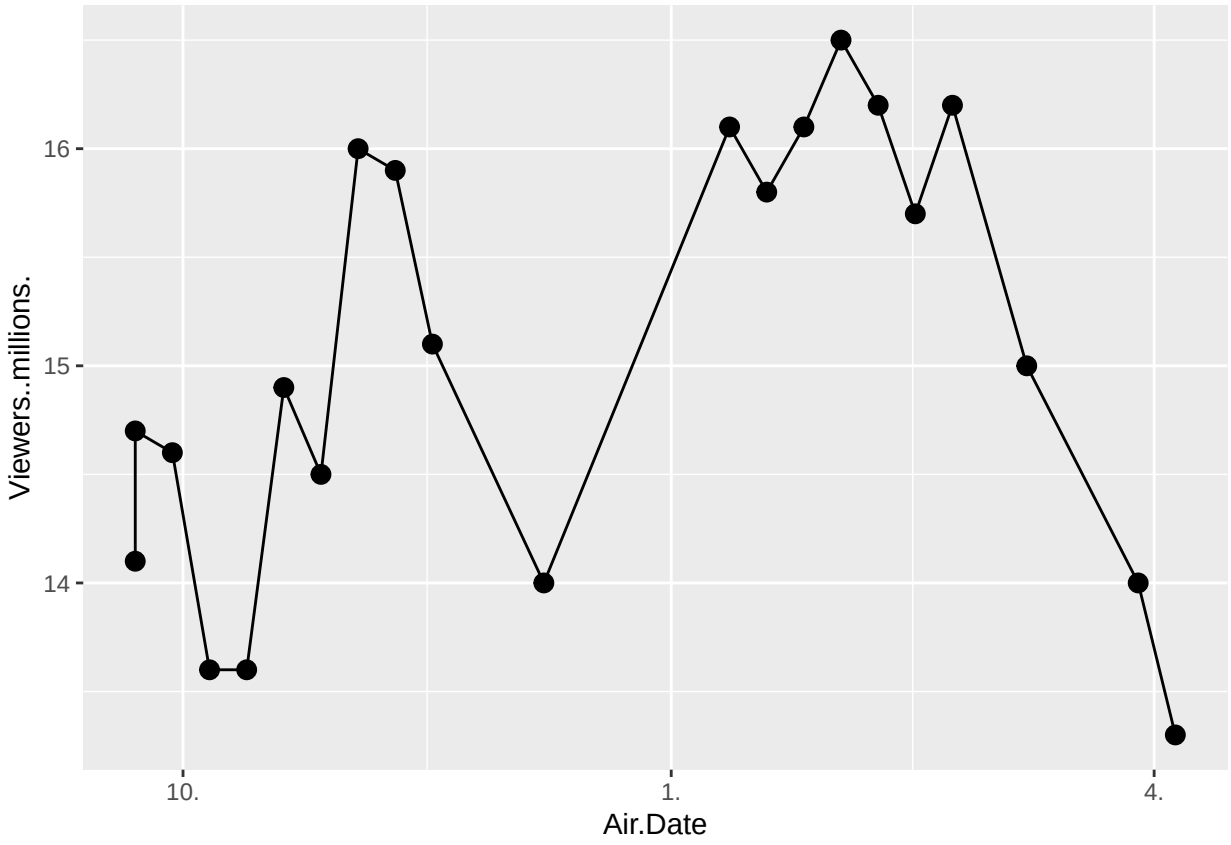
```r
q25_viewers<-quantile(BigBangTheory$Viewers..millions.,0.25)
q75_viewers<-quantile(BigBangTheory$Viewers..millions.,0.75)
print(c(q25_viewers,q75_viewers))
```

```
##  25%  75%
## 14.1 16.0
```

0.0.4 d

```r
BigBangTheory$Air.Date <- parse_date_time(BigBangTheory$Air.Date, orders = c("%B %d, %Y"))

ggplot(BigBangTheory, aes(x = Air.Date, y = Viewers..millions.)) +
  geom_line() +
  geom_point(shape = 19, color = "black", size = 3) +
  labs(x = "Air.Date", y = "Viewers..millions.")
```



函

数曲线没有明显的趋势，可以认为 2011~2012 年之间没有明显的收视率上升或者下降

问题 2: NBAPlayerPts. (相关数据: NBAPlayerPts)

CBS 体育网开发了“综合球员评分系统”，该系统根据球员在 NBA 比赛中的各项进攻和防守数据对他们进行评分。附带的数据文件“NBA 球员得分”显示了 2012-2013 赛季 NBA 比赛中 50 名得分最高的球员的平均得分 (PPG) (CBS 体育网，2013 年 2 月 25 日)。请按照从 10 开始，以 2 为增量，到 30 结束的顺序对 PPG 进行分类。

a. 展示频数分布图。

b. 展示频率分布。

c. 展示累积频率分布图。

d. 制作一个直方图，显示每场比赛的平均得分。

e. 这些数据是否存在偏斜现象？请解释原因。

f. 有多少球员平均每场比赛得分至少 20 分?

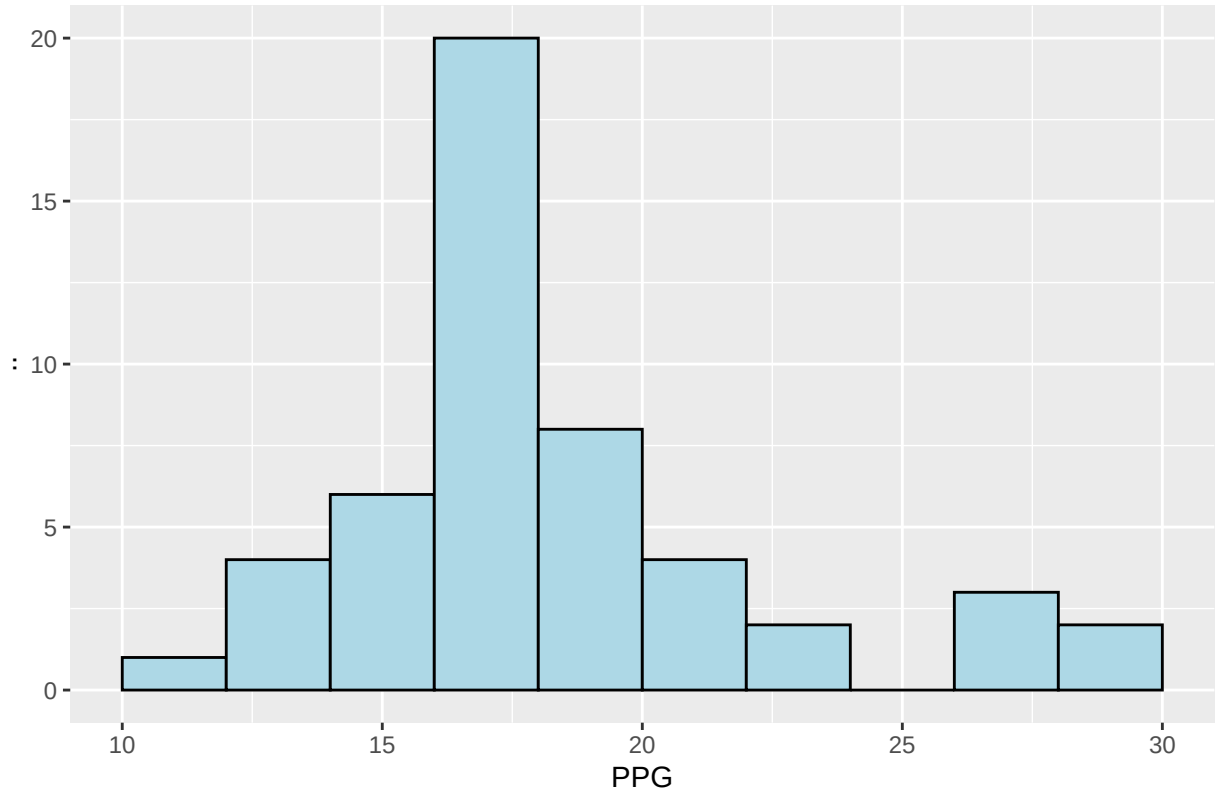
0.0.5 a

```
data_PPG<-read_csv("D:/MEM/2-定量分析-数据思维与商业统计/MEM_Studying_by_Hoki/2st_assignment_eda_Hok
```

```
## Rows: 50 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Player
## dbl (2): Rank, PPG
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
start_value <- 10#
increment <- 2
end_value <- 30
breaks_vector <- seq(start_value, end_value, by = increment)
frame_PPG <- data.frame(PPG = data_PPG$PPG)
ggplot(frame_PPG, aes(x = PPG)) +
  geom_histogram(breaks = breaks_vector, fill = "lightblue", color = "black") +
  labs(title = " 频数分布直方图", x = "PPG", y = " 频数")
```

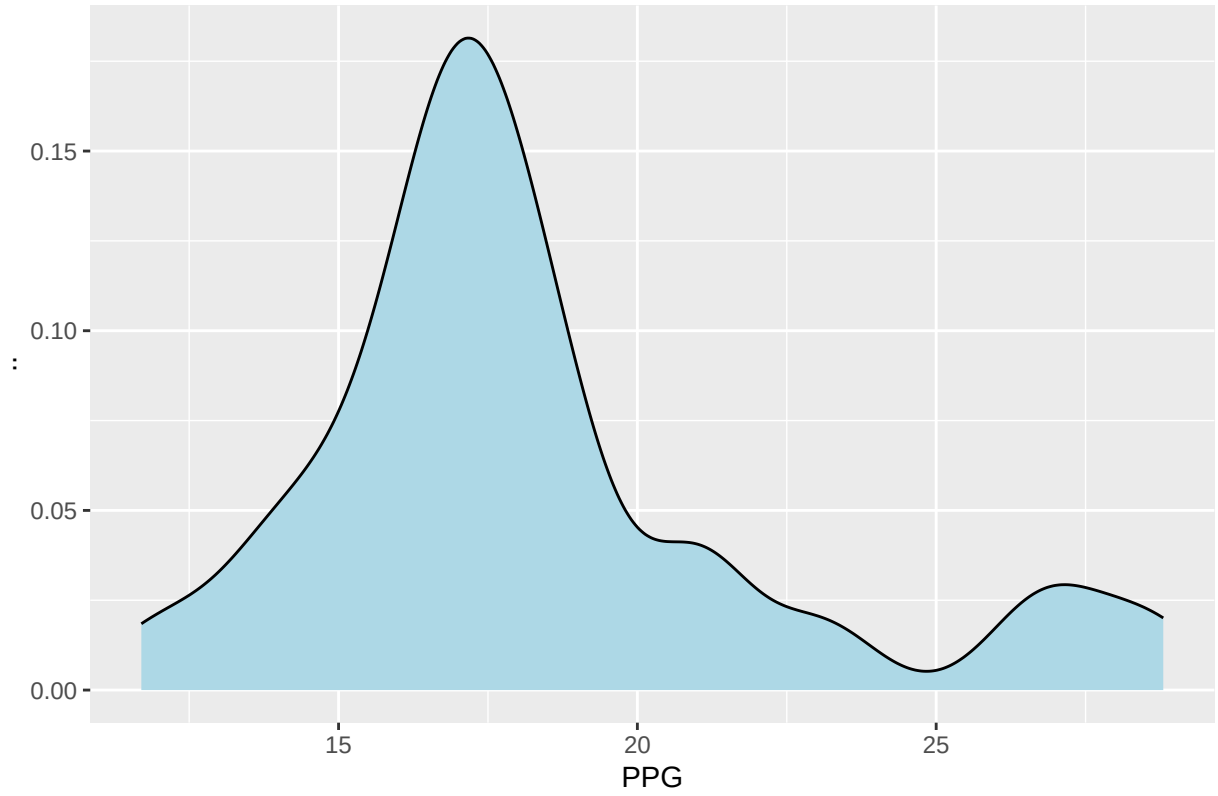
.....



0.0.6 b

```
ggplot(frame_PPG, aes(x = PPG)) +
  geom_density(fill = "lightblue", color = "black") +
  labs(title = " 概率密度曲线", x = "PPG", y = " 密度")
```

.....



:

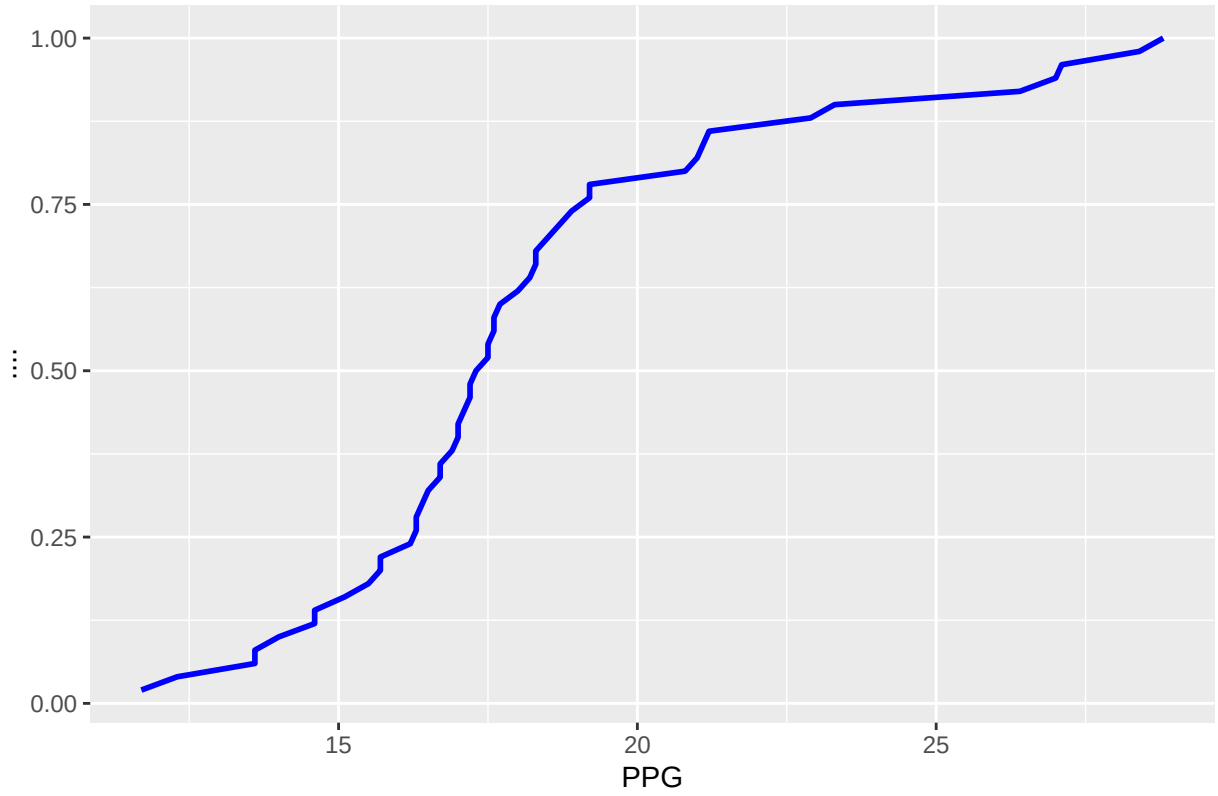
0.0.7 c

```
frame_PPG <- frame_PPG %>%
  arrange(PPG) %>%
  mutate(cumulative_frequency = cumsum(rep(1 / nrow(frame_PPG), nrow(frame_PPG))))

ggplot(frame_PPG, aes(x = PPG, y = cumulative_frequency)) +
  geom_line(color = "blue", size = 1) +
  labs(title = " 累积频率分布图", x = "PPG", y = " 累积频率")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

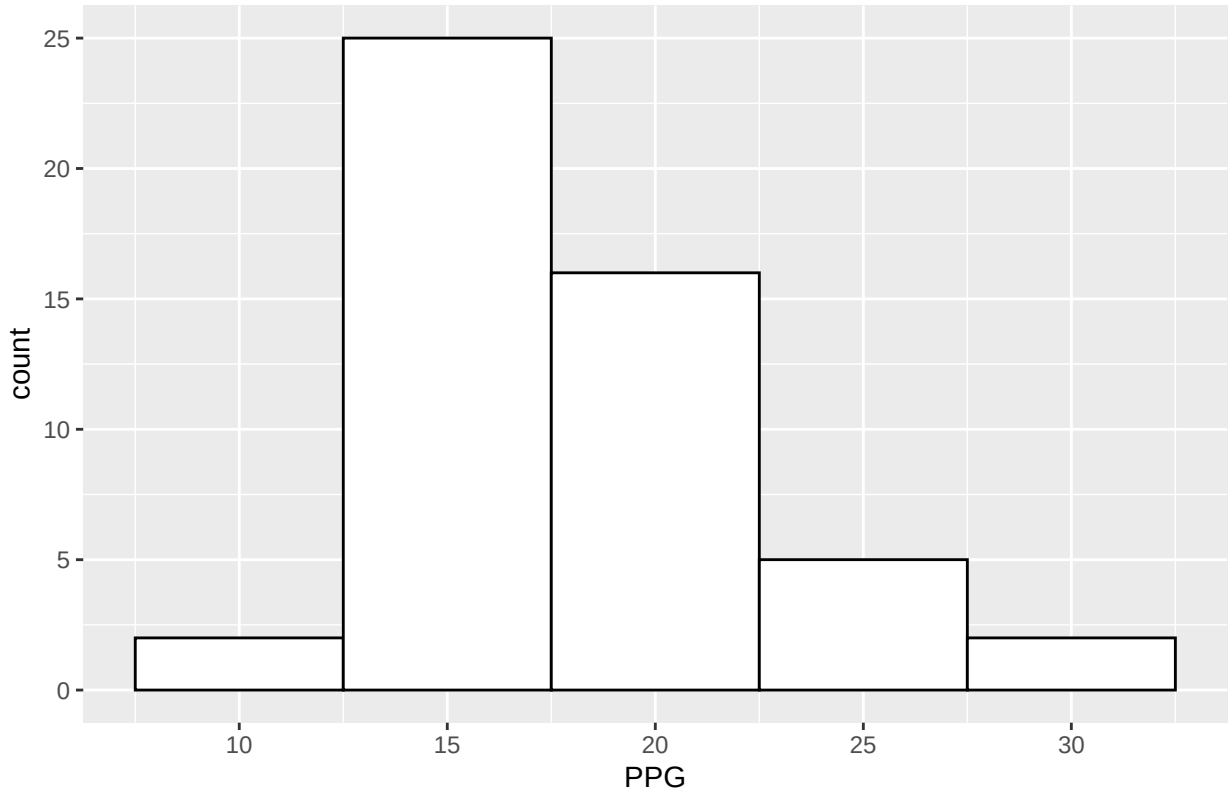
.....



0.0.8 d

```
ggplot(data_PPG,aes(PPG)) +
  geom_histogram(binwidth = 5,color = "black",fill="white") +
  labs(title = " 比赛平均得分", x = "PPG")
```

.....



0.0.9 e

根据频率分布图可知，数据存在右偏的情况，原因可能是因为比赛成绩一般的情况较为集中，比赛成绩优异的情况比较分散。

0.0.10 f

```
num_20<-data_PPG%>%
  filter(PPG > 20) %>%
  nrow()
pre_20<-num_20/nrow(data_PPG)*100
print(pre_20)
```

```
## [1] 22
```

问题 3: 研究人员在报告调查结果时声明平均值的标准误差为 20。总体标准差是 500。

a. 这次调查的样本量有多大?

b. 点估计值在总体均值 ± 25 以内的概率是多少?

0.0.11 a

```r
sigma_q3 <- 500
SE_q3 <- 20
n_q3 <- (sigma_q3 / SE_q3)^2

print(n_q3)
```

```
## [1] 625
```

0.0.12 b

```r
# 计算下限和上限对应的标准正态分布的 z 值
z1 <- (-25) / SE_q3
z2 <- 25 / SE_q3

# 计算概率
probability <- pnorm(z2) - pnorm(z1)
print(probability)
```

```
## [1] 0.7887005
```

问题 4: 青年专业杂志（附件数据：专业）

《年轻专业人士》杂志的目标受众是刚毕业的大学生，他们在商业/职业生涯中处于头 10 年。这本杂志出版两年来相当成功。现在出版商有意扩大杂志的广告基础。潜在的广告商不断询问“年轻专业人士”订户的人口统计数据和兴趣。为了收集这些信息，该杂志委托进行了一项调查，以建立其订户的档案。调查结果将用于帮助杂志选择感兴趣的文章，并向广告商提供订户的概况。作为杂志的新员工，你被要求帮助分析调查结果。

以下是部分调查问题：

1. 你多大了?
2. 你是：男 _____ 女 _____
3. 你打算在未来两年内购买房地产吗?

Yes_____ No_____

4. 除去你的投资，金融投资的总价值大概是多少

你或你的家人拥有的房子?

5. 在过去一年中，你进行了多少笔股票/债券/共同基金交易?
6. 你家里有宽带上网吗? Yes_____ No_____
7. 请注明你去年的家庭总收入。_____
8. 你有孩子吗? Yes_____ No_____

标题为“专业”的文件包含对这些问题的回答。

管理报告:* * * *

准备一份总结调查结果的管理报告。除了统计总结，讨论杂志如何利用这些结果来吸引广告商。你也可以评论一下杂志的编辑如何利用调查结果来确定读者感兴趣的话题。你的报告应该解决以下问题，但不要把你的分析局限于这些领域。

a. 制定适当的描述性统计来总结数据。

b. 为用户的平均年龄和家庭收入制定 95% 置信区间。

c. 制定家中有宽带接入的用户比例和有子女的用户比例的 95% 置信区间。

d. *Young Professional* 会是在线经纪人的一个很好的广告渠道吗? 用统计数据证明你的结论。

e. 这本杂志会是为那些销售教育软件和儿童电脑游戏的公司做广告的好地方吗?

f. 评论你认为读者会对 *Young Professional* 感兴趣的文章类型。

0.0.13 a

```
data_professional<-read_csv("assignment-2-BruceZhaoR/data/Professional.csv")
```

```
## New names:
## Rows: 410 Columns: 14
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): Gender, Real Estate Purchases?, Broadband Access?, Have Children?, ... dbl
## (4): Age, Value of Investments ($), Number of Transactions, Household In... lgl
## (5): ...9, ...11, ...12, ...13, ...14
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
```

```
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
summary(data_professional)
```

```
##       Age           Gender          Real Estate Purchases?
##  Min.   :19.00   Length:410         Length:410
##  1st Qu.:28.00   Class :character   Class :character
##  Median :30.00   Mode  :character   Mode  :character
##  Mean   :30.11
##  3rd Qu.:33.00
##  Max.   :42.00
##  Value of Investments ($) Number of Transactions Broadband Access?
##  Min.   :     0           Min.   : 0.000         Length:410
##  1st Qu.: 18300           1st Qu.: 4.000         Class :character
##  Median : 24800           Median : 6.000         Mode  :character
##  Mean   : 28538           Mean   : 5.973
##  3rd Qu.: 34275           3rd Qu.: 7.000
##  Max.   :133400           Max.   :21.000
##  Household Income ($) Have Children?       ...9            ...10
##  Min.   : 16200       Length:410         Mode:logical   Length:410
##  1st Qu.: 51625       Class :character   NA's:410       Class :character
##  Median : 66050       Mode  :character                  Mode  :character
##  Mean   : 74460
##  3rd Qu.: 88775
##  Max.   :322500
##   ...11          ...12          ...13          ...14
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:410       NA's:410       NA's:410       NA's:410
##
##
##
##
```

0.0.14 b

```r
alpha_q4b <- 0.05
n_q4 <- length(data_professional$Age)
df_q4 <- length(data_professional$Age)-1
Age_mean <- mean(data_professional$Age)
Income_mean <- mean(data_professional$`Household Income ($)`)
Age_sd <- sd(data_professional$Age)
Incom_sd <- sd(data_professional$`Household Income ($)`)
Age_of_error <- qt(1-alpha_q4b/2,df = df_q4)*Age_sd/sqrt(n_q4)
Income_of_error <- qt(1-alpha_q4b/2,df = df_q4)*Incom_sd/sqrt(n_q4)

lower_bound_age <- Age_mean - Age_of_error
upper_bound_age <- Age_mean + Age_of_error
lower_bound_income <- Income_mean - Income_of_error
upper_bound_income <- Income_mean + Income_of_error

cat(" 显著性水平为 0.05 时平均年龄的置信区间为: (", lower_bound_age, ",", upper_bound_age, ")\n")
```

```
## 显著性水平为0.05时平均年龄的置信区间为: ( 29.72153 , 30.50286 )
```

```r
cat(" 显著性水平为 0.05 时平均家庭收入的置信区间为: (", lower_bound_income, ",", upper_bound_income,
```

```
## 显著性水平为0.05时平均家庭收入的置信区间为: ( 71079.26 , 77839.77 )
```

0.0.15 c

```r
# 计算样本中具有特征的个体数
x_broadbandaccess <- sum(data_professional$`Broadband Access?` == "Yes")
# 计算样本容量 n
n_broadbandaccess <- length(data_professional$`Broadband Access?`)
# 计算样本比例 p_hat
p_hat_broadbandaccess <- x_broadbandaccess / n_broadbandaccess

# 设置显著性水平 alpha
alpha_q4c <- 0.05
# 计算自由度
df_broadbandaccess <- n_broadbandaccess - 1
# 获取 t 分布分位数
```

```r
t_value_broadbandaccess <- qt(1 - alpha_q4c/2, df_broadbandaccess)

# 计算置信区间下限
lower_bound_broadbandaccess <- p_hat_broadbandaccess - t_value_broadbandaccess * sqrt(p_hat_broadb
# 计算置信区间上限
upper_bound_broadbandaccess <- p_hat_broadbandaccess + t_value_broadbandaccess * sqrt(p_hat_broadb
# 输出置信区间
cat(" 在显著性水平为 0.05 时，总体家中有宽带接入的用户比例的置信区间为：(", lower_bound_broadbandaccess
```

```
## 在显著性水平为0.05时，总体家中有宽带接入的用户比例的置信区间为：( 0.5773174 , 0.671463 )
```

```r
x_children <- sum(data_professional$`Have Children?` == "Yes")
n_children <- length(data_professional$`Have Children?` )
p_hat_children <- x_children / n_children

alpha_q4c <- 0.05
df_children <- n_children - 1
t_value_children <- qt(1 - alpha_q4c/2, df_children)

lower_bound_children <- p_hat_children - t_value_children * sqrt(p_hat_children*(1 - p_hat_childre
upper_bound_children <- p_hat_children + t_value_children * sqrt(p_hat_children*(1 - p_hat_childre
cat(" 在显著性水平为 0.05 时，总体家中有孩子的用户比例的置信区间为：(", lower_bound_children, ",", upp
```

```
## 在显著性水平为0.05时，总体家中有孩子的用户比例的置信区间为：( 0.485659 , 0.5826337 )
```

0.0.16 d

Young Professional 会是在线经纪人的一个很好的广告渠道。因为近六成受众家中接入了宽带，同时他们平均金融投资的总价值高达 28538 美元，并且频繁进行股票/债券/共同基金交易。

0.0.17 e

这本杂志会是为那些销售教育软件和儿童电脑游戏的公司做广告的好地方，因为近半成受众家中有孩子，受众本身年纪在 30 岁出头，所以他们的孩子年纪尚小，适合向他们推销教育软件和儿童电脑游戏。

0.0.18 f

Young Professional 的受众可能会对新发布的股票/债券/共同基金、开发的新建筑楼盘、育儿方面的信息感兴趣。

问题 5: Quality Associate，Inc.（附件数据：质量）

Quality associates, inc. 是一家咨询公司，就可用于控制其生产过程的抽样和统计程序向其客户提供建议。在一个特定的应用程序中，客户向质量部门提供了 800 个观察结果的样本，在此期间，该客户的流程运行令人满意。这些数据的样本标准差为 0.21；因此，对于如此多的数据，假定总体标准差为 0.21。随后，质量人员建议定期随机抽取 30 个样本，以持续监测这一过程。通过分析新样品，客户可以快速了解工艺是否令人满意地运行。当过程不能令人满意地运行时，可以采取纠正措施来消除问题。设计规范表明该工艺的平均值应为 12。Quality associates 建议的假设检验如下。

$$H_0 : \mu = 12 H_1 : \mu \neq 12$$

一旦 H_0 被拒绝，将采取纠正措施。

数据可在数据集质量。

管理报告

a. 对每个样本在 0.01 的显著性水平上进行假设检验，并确定应该采取什么行动（如果有的话）。提供每个检验的 p 值。

B. 计算四个样本中每个样本的标准差。总体标准差 0.21 的假设合理吗?

C. 计算样本均值 $\overline{x}$ 在 $\mu = 12$ 附近的限制，这样，只要一个新的样本均值在这些限制范围内，该过程将被认为是令人满意的操作。如果 $\overline{x}$ 超过上限或 $\overline{x}$ 低于下限，将采取纠正措施。这些限值被称为质量控制的上限值和下限值。

D. 讨论将显著性水平改变为更大的值的含义。如果显著性水平提高，什么错误或错误会增加?

0.0.19 a

```
data_Quality <- read_csv("assignment-2-BruceZhaoR/data/Quality.csv")
```

```
## Rows: 30 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): Sample 1, Sample 2, Sample 3, Sample 4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
getOption("digits")
```

```
## [1] 7
```

```r
options(digits = 4)

result1 <- t.test(data_Quality$`Sample 1`, mu = 12, conf.level = 0.99,alternative = "two.sided")
print(result1)
```

```
##
##  One Sample t-test
##
## data:  data_Quality$`Sample 1`
## t = -1, df = 29, p-value = 0.3
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.85 12.07
## sample estimates:
## mean of x
##     11.96
```

```r
# 样本 1 的 p 值为 0.3, 可以接收原假设

result2 <- t.test(data_Quality$`Sample 2`, mu = 12, conf.level = 0.99, alternative = "two.sided")
print(result2)
```

```
##
##  One Sample t-test
##
## data:  data_Quality$`Sample 2`
## t = 0.71, df = 29, p-value = 0.5
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.92 12.14
## sample estimates:
## mean of x
##     12.03
```

```r
# 样本 2 的 p 值为 0.5, 可以接收原假设

result3 <- t.test(data_Quality$`Sample 3`, mu = 12, conf.level = 0.99, alternative = "two.sided")
print(result3)
```

```
##
##  One Sample t-test
##
## data:  data_Quality$`Sample 3`
## t = -2.9, df = 29, p-value = 0.006
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.78 11.99
## sample estimates:
## mean of x
##     11.89
```

```
# 样本 3 的 p 值为 0.006，拒绝原假设，需要采取纠偏措施
```

```
result4 <- t.test(data_Quality$`Sample 4`, mu = 12, conf.level = 0.99, alternative = "two.sided")
print(result4)
```

```
##
##  One Sample t-test
##
## data:  data_Quality$`Sample 4`
## t = 2.2, df = 29, p-value = 0.04
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.98 12.19
## sample estimates:
## mean of x
##     12.08
```

```
# 样本 4 的 p 值为 0.04，勉强可以接收原假设
```

1 样本 1 的 p 值为 0.3，可以接收原假设

2 样本 2 的 p 值为 0.5，可以接收原假设

3 样本 3 的 p 值为 0.006，拒绝原假设，需要采取纠偏措施

4 样本 4 的 p 值为 0.04，勉强可以接收原假设

0.0.20 b

```
sd1<-sd(data_Quality$`Sample 1`)
sd2<-sd(data_Quality$`Sample 2`)
sd3<-sd(data_Quality$`Sample 3`)
sd4<-sd(data_Quality$`Sample 4`)

print(paste(" 样本 1 的标准差:", sd1))
```

```
## [1] "样本1的标准差: 0.220356033748104"
```

```
print(paste(" 样本 2 的标准差:", sd2))
```

```
## [1] "样本2的标准差: 0.220356033748104"
```

```
print(paste(" 样本 3 的标准差:", sd3))
```

```
## [1] "样本3的标准差: 0.207170594371918"
```

```
print(paste(" 样本 4 的标准差:", sd4))
```

```
## [1] "样本4的标准差: 0.206108999173325"
```

```
# 已知总体标准差
sigma_q5 <- 0.21

chi_sq1 <- (30 - 1) * sd1^2 / sigma_q5^2
chi_sq2 <- (30 - 1) * sd2^2  / sigma_q5^2
chi_sq3 <- (30 - 1) * sd3^2  / sigma_q5^2
chi_sq4 <- (30 - 1) * sd4^2  / sigma_q5^2

# 进行卡方检验
p_value1 <- pchisq(chi_sq1, 29, lower.tail = FALSE)
p_value2 <- pchisq(chi_sq2, 29, lower.tail = FALSE)
p_value3 <- pchisq(chi_sq3, 29, lower.tail = FALSE)
p_value4 <- pchisq(chi_sq4, 29, lower.tail = FALSE)

print(paste(" 样本 1 的卡方检验 p 值:", p_value1))
```

```
## [1] "样本1的卡方检验p值: 0.322920624662182"
```

```
print(paste(" 样本 2 的卡方检验 p 值:", p_value2))
```

```
## [1] "样本2的卡方检验p值: 0.322920624662185"
```

```
print(paste(" 样本 3 的卡方检验 p 值:", p_value3))
```

```
## [1] "样本3的卡方检验p值: 0.505971575483381"
```

```
print(paste(" 样本 4 的卡方检验 p 值:", p_value4))
```

```
## [1] "样本4的卡方检验p值: 0.521374554084172"
```

在显著性水平为 0.05 的条件下，不能拒绝总体标准差是 0.21 的假设。

0.0.21 c

```
n_q5 <- 30
x_mean_q5 <- 12

# 已知总体标准差
sigma_q5 <- 0.21

alpha_q5 <- 0.05
z_score_q5 <- qnorm(1 - alpha_q5/2)
margin_of_error_q5 <- z_score_q5*(sigma_q5/sqrt(n_q5))
lower_bound_q5 <- x_mean_q5 - margin_of_error_q5
upper_bound_q5 <- x_mean_q5 + margin_of_error_q5
cat(" 显著性水平为 0.05 时的置信区间为: (", lower_bound_q5, ",", upper_bound_q5, ")\n")
```

```
## 显著性水平为0.05时的置信区间为: ( 11.92 , 12.08 )
```

0.0.22 d

```
# 当显著性水平增大时，例如从 0.01 提高到 0.05，我们更容易拒绝原假设，设定了一个更宽松的标准来拒绝原假设，
```

问题 6: 南卡罗来纳州默特尔比奇的度假入住率预计将在 2008 年 3 月上升（太阳新闻，2008 年 2 月 29 日）。文件占用（附件文件**占用**）中的数据将允许您复制报纸上呈现的发现。数据显示了 2007 年 3 月和 2008 年 3 月的第一周，随机抽样的度假物业的出租和未出租单位。

a. 估计 2007 年 3 月第一周及 2008 年 3 月第一周的出租单位比例。

b. 为比例差异提供 95% 置信区间。

c. 根据你的调查结果，2008 年 3 月份的出租比例是否会高于去年同期?

0.0.23 a

```
data_Occupancy <- read_csv("assignment-2-BruceZhaoR/data/Occupancy.csv")
```

```
## New names:
## Rows: 201 Columns: 2
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Unit Rented?, ...2
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...2`
```

```
colnames(data_Occupancy) <- c("2007/03","2008/03")
precentage_2007 <- sum(data_Occupancy$`2007/03`== "Yes")/sum(complete.cases(data_Occupancy$`2007/0
precentage_2008 <- sum(data_Occupancy$`2008/03`[!is.na(data_Occupancy$`2008/03`)] == "Yes")/sum(co

print(paste("2007 年 3 月出租比例", precentage_2007))
```

```
## [1] "2007年3月出租比例 0.35"
```

```
print(paste("2008 年 3 月出租比例", precentage_2008))
```

```
## [1] "2008年3月出租比例 0.466666666666667"
```

0.0.24 b

```r
# 第一个样本的事件发生次数和样本容量
x_2007 <- sum(data_Occupancy$`2007/03`== "Yes")
n_2007 <- sum(complete.cases(data_Occupancy$`2007/03`[-1]))
# 第二个样本的事件发生次数和样本容量
x_2008 <- sum(data_Occupancy$`2008/03`[!is.na(data_Occupancy$`2008/03`)] == "Yes")
n_2008 <- sum(complete.cases(data_Occupancy$`2008/03`[-1]))
p1 <- x_2007 / n_2007
p2 <- x_2008 / n_2008
SE_q6 <- sqrt((p1 * (1 - p1)) / n_2007 + (p2 * (1 - p2)) / n_2008)
alpha_q6 <- 0.05
z_score_q6 <- qnorm(1 - alpha_q6/2)
lower_bound_q6 <- (p1 - p2) - z_score_q6 * SE_q6
upper_bound_q6 <- (p1 - p2) + z_score_q6 * SE_q6
cat("2007 年和 2008 年的租房比例差异在 0.05 的置信水平下置信区间为: (", lower_bound_q6, ",", upper_bou
```

```
## 2007年和2008年的租房比例差异在0.05的置信水平下置信区间为: ( -0.2203 , -0.01302 )
```

0.0.25 c

可以认为 2008 年的租房比例高于 2007 年，因为比例差异的区间内不包含 0。

问题 #7: 空军训练计划（数据文件：训练）

空军的电子学入门课程采用了一种个性化的教学系统，通过这种系统，每个学生观看一段录像讲座，然后获得一份编程的教学文本。学生们独立完成课文，直到他们完成训练并通过考试。值得关注的是学生完成这部分训练计划的速度不同。有些学生能够相对较快地掌握编程指导文本，而其他学生学习文本的时间要长得多，需要额外的时间来完成课程。快的学生等到慢的学生完成入门课程，然后整个小组一起进行其他方面的训练。

一个被提议的替代系统包括使用计算机辅助教学。在这种方法中，所有学生观看相同的视频讲座，然后每个学生被分配到一个计算机终端进行进一步的指导。计算机引导学生独立完成课程的自我训练部分。

为了比较建议的教学方法和当前的教学方法，一个 122 名学生的新生被随机分配到两种方法中的一种。一组 61 名学生使用目前的程序文本方法，另一组 61 名学生使用拟议的计算机辅助方法。研究人员记录了每个学生的时间（以小时为单位）。数据在数据集训练中提供（见附件）。

- — * * 管理报告

A. 使用适当的描述性统计来总结每种方法的训练时间数据。你从样本数据中观察到什么相似或不同之处?

b. 评论两种方法的总体均值之间的差异。讨论你的发现。

C. 计算每种训练方法的标准差和方差。对两种训练方法进行总体方差相等性的假设检验。讨论你的发现。

D. 关于这两种方法的区别，你能得出什么结论？你的建议是什么？解释一下。

E. 在最终决定将来使用的培训计划之前，你能建议其他可能需要的数据或测试吗？

0.0.26 a

```
data_Training <- read_csv("assignment-2-BruceZhaoR/data/Training.csv")
```

```
## Rows: 61 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Current, Proposed
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(data_Training)
```

```
##     Current        Proposed
##  Min.   :65.0   Min.   :69.0
##  1st Qu.:72.0   1st Qu.:74.0
##  Median :76.0   Median :76.0
##  Mean   :75.1   Mean   :75.4
##  3rd Qu.:78.0   3rd Qu.:77.0
##  Max.   :84.0   Max.   :82.0
```

```
sd(data_Training$Current)
```

```
## [1] 3.945
```

```
sd(data_Training$Proposed)
```

```
## [1] 2.506
```

两组数据的均值相近，但是“Current”的数据离散程度比“Proposed”组的数据高。

0.0.27 b

```r
result_q7 <- t.test(data_Training$Current,data_Training$Proposed)
result_q7
```

```
##
##  Welch Two Sample t-test
##
## data:  data_Training$Current and data_Training$Proposed
## t = -0.6, df = 102, p-value = 0.5
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5477  0.8263
## sample estimates:
## mean of x mean of y
##     75.07     75.43
```

可以看做两组数据的均值没有显著差异。

0.0.28 c

```r
sd1_q7 <- sd(data_Training$Current)
sd2_q7 <- sd(data_Training$Proposed)
var1_q7 <- var(data_Training$Current)
var2_q7 <- var(data_Training$Proposed)
print(c(sd1_q7,sd2_q7,var1_q7,var2_q7))
```

```
## [1]  3.945  2.506 15.562  6.282
```

```r
# 对两组数据进行正态性检验
shapiro_result1 <- shapiro.test(data_Training$Current)
shapiro_result2 <- shapiro.test(data_Training$Proposed)
print(shapiro_result1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_Training$Current
## W = 0.99, p-value = 0.7
```

```r
print(shapiro_result2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_Training$Proposed
## W = 0.97, p-value = 0.09
```

```r
# 显著性水平 0.05 条件下可以近似认为两组数据均服从正态分布

# 方差齐性检验
variance_test_result_q7 <- var.test(data_Training$Current, data_Training$Proposed)
print(variance_test_result_q7)
```

```
##
##  F test to compare two variances
##
## data:  data_Training$Current and data_Training$Proposed
## F = 2.5, num df = 60, denom df = 60, p-value = 6e-04
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.486 4.129
## sample estimates:
## ratio of variances
##              2.477
```

在显著性水平 0.05 的条件下拒绝原假设，认为两组数据的方差不相等。

0.0.29 d

第二种建议的教育方式由于其培训时长的稳定性，在一定程度上可以被认为是更可靠的。如果想要一个能够较为稳定地在一段时间内进行培训的方法，第二种教育方式可能是更好的选择。第二种建议的教育方式由于时间方差小，具有更强的可预测性。教育者可以更准确地预估完成教育目标所需的时间，从而更好地规划教学进度和资源分配。从效率角度看，第二种建议的教育方式效率的一致性更高。平均时长相近说明两种教育方式总体效率差不多，但第二种教育方式每次达到教育结果所需时间更稳定，不会出现过长或过短的极端情况。

0.0.30 e

在确定最终的培训方式之前，还需要对培训的结果好坏进行分析，可以通过考试成绩量化分析不同的培训方式的质量。最后根据相同时间下的培训数量和培训质量来确定培训方式。

问题 8: 丰田凯美瑞是北美最畅销的汽车之一。以前拥有的凯美瑞的价格取决于许多因素，包括车型年份，里程和状况。为了调查 2007 款凯美瑞汽车的里程和销售价格之间的关系，附件数据文件凯美瑞显示了 19 次销售的里程和销售价格 (Pricehub 网站，2012 年 2 月 24 日)。

a. 制作一个散点图，横轴为汽车里程，纵轴为价格。

B. (a) 部分的散点图说明了两个变量之间的什么关系?

c. 在给定里程 (1000) 的情况下，开发可用于预测价格 ($1000) 的估计回归方程。

d. 在 0.05 显著性水平上检验显著性关系。

e. 估计的回归方程是否提供良好的拟合? 解释一下。

f. 对估计的回归方程的斜率进行解释。

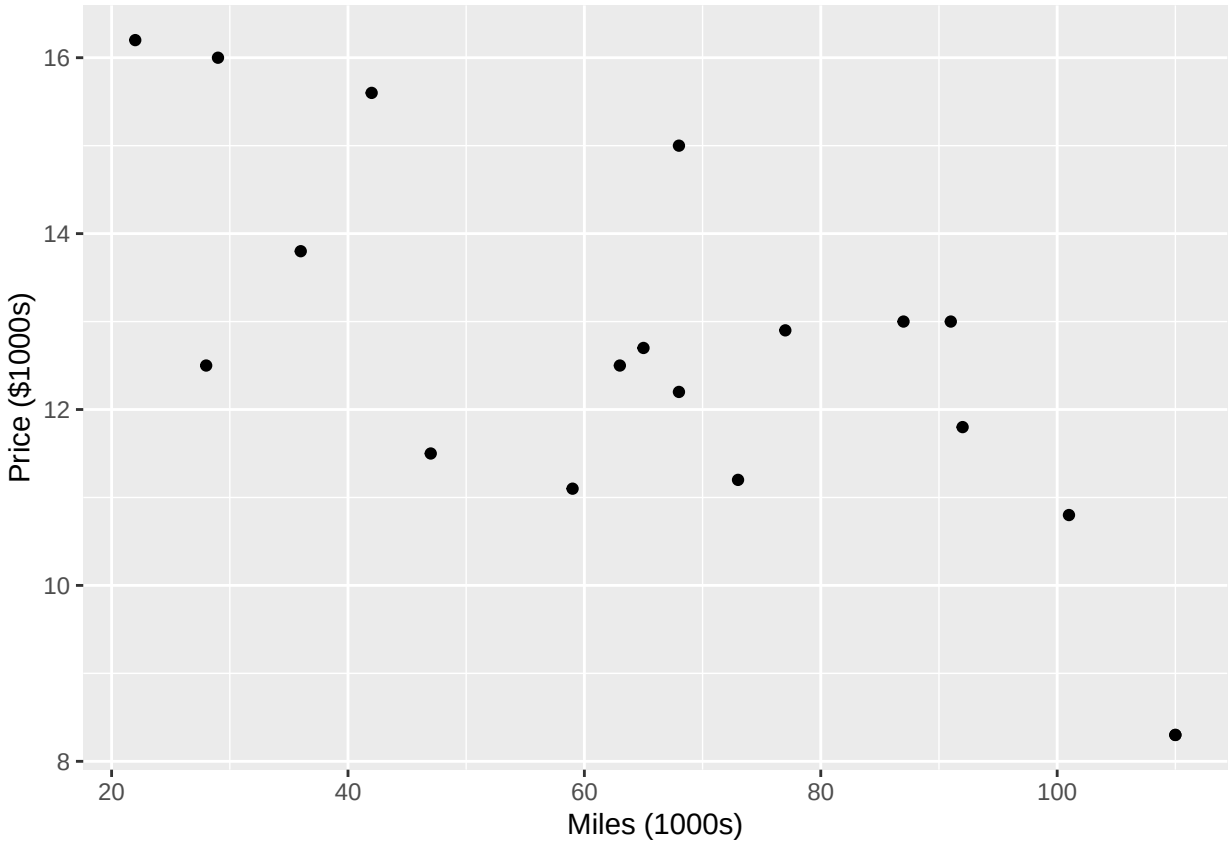
假设你正在考虑购买一辆旧车，这辆 2007 年的凯美瑞已经行驶了 6 万英里。使用 (c) 部分中开发的估计回归方程，预测这辆车的价格。这是你给卖家的价格吗?

0.0.31 a

```
data_Camry <- read_csv("assignment-2-BruceZhaoR/data/Camry.csv")
```

```
## Rows: 19 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Miles (1000s), Price ($1000s)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(data = data_Camry,aes(x = `Miles (1000s)`,y = `Price ($1000s)`))+
  geom_point()
```



0.0.32 b

两个变量呈现负相关关系，从图形上看可能是线性的负相关关系，也可能是曲线的负相关关系。

0.0.33 c

```
# 拟合线性回归模型
model_q8 <- lm(`Price ($1000s)` ~ `Miles (1000s)`, data = data_Camry)
# 查看模型摘要信息
summary(model_q8)
```

```
##
## Call:
## lm(formula = `Price ($1000s)` ~ `Miles (1000s)`, data = data_Camry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.3241 -1.3419  0.0506  1.1290  2.5269
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.4698     0.9488   17.36    3e-12 ***
## `Miles (1000s)` -0.0588     0.0132   -4.46  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.54 on 17 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.512
## F-statistic: 19.8 on 1 and 17 DF,  p-value: 0.000348
```

```r
# 获取截距（b0）
intercept_q8 <- coef(model_q8)[1]
# 获取回归系数（b1）
slope_q8 <- coef(model_q8)[2]

# 打印回归方程
cat(" 估计回归方程为: price =", intercept_q8, "+", slope_q8, "* Miles")
```

```
## 估计回归方程为: price = 16.47 + -0.05877 * Miles
```

0.0.34 d

```r
summary(model_q8)
```

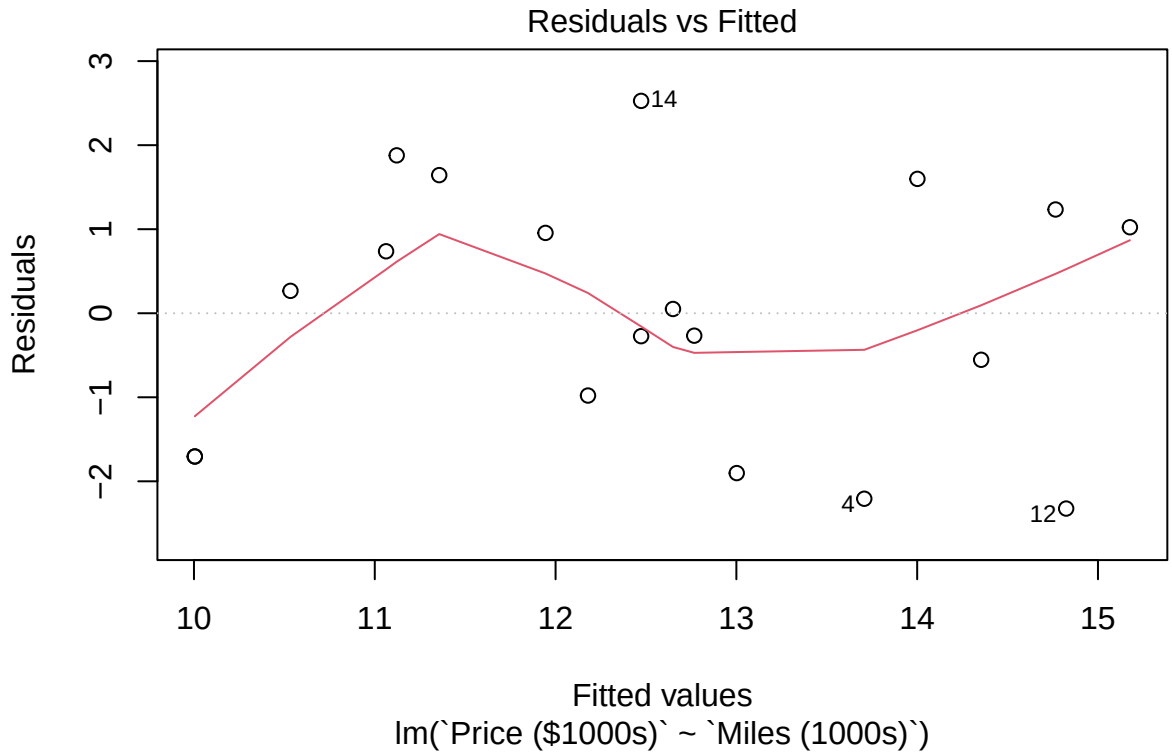
```
##
## Call:
## lm(formula = `Price ($1000s)` ~ `Miles (1000s)`, data = data_Camry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3241 -1.3419  0.0506  1.1290  2.5269
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.4698     0.9488   17.36    3e-12 ***
```

```
## `Miles (1000s)`  -0.0588     0.0132   -4.46  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.54 on 17 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.512
## F-statistic: 19.8 on 1 and 17 DF,  p-value: 0.000348
```

p 值为 0.000348，小于 0.05，可以认为两个变量值是显著线性相关的。

0.0.35 e

```
plot(model_q8, which = 1)
```




```
residuals(model_q8) %>% shapiro.test()
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95, p-value = 0.4
```

1. Multiple R-squared 决定系数值为 0.539，表示模型可以解释因变量 53.9% 的变异。
2. 残差随机地分布在 0 附近，没有明显的模式。
3. 残差正态性检验 p 值为 0.4，大于 0.05，说明残差服从正态分布，模型拟合良好。
4. 截距和自变量 Miles 的 p 值均小于 0.05，说明自变量系数都显著，回归方程在考虑这些变量方面有较好的拟合。结论：该模型有较好的拟合效果。

0.0.36 f

这个回归线性方程的斜率为-0.059，表示一辆 2007 年的丰田凯美瑞汽车的里程数每增加 1000 英里，价格下降 59 美元。

0.0.37 g

对于一辆行驶里程为 6 万英里的 2007 年款凯美瑞，预测价格为 16.47-0.0588×(60)=12.942，也就是 12942 美元。实际给卖家的价格需要依据实际情况而定，该预测结果只是对成交价格给出大致的参考。

Question #9: 附件 WE.xlsx 是某提供网站服务的 Internet 服务商的客户数据。数据包含了 6347 名客户在 11 个指标上的表现。其中” 流失 “指标中 0 表示流失，”1“表示不流失，其他指标含义看变量命名。

a. 通过可视化探索流失客户与非流失客户的行为特点（或特点对比），你能发现流失与非流失客户行为在哪些指标有可能存在显著不同?

b. 通过均值比较的方式验证上述不同是否显著。

c. 以” 流失 “为因变量，其他你认为重要的变量为自变量（提示：a、b 两步的发现），建立回归方程对是否流失进行预测。

d. 根据上一步预测的结果，对尚未流失（流失 =0）的客户进行流失可能性排序，并给出流失可能性最大的前 100 名用户 ID 列表。

0.0.38 a

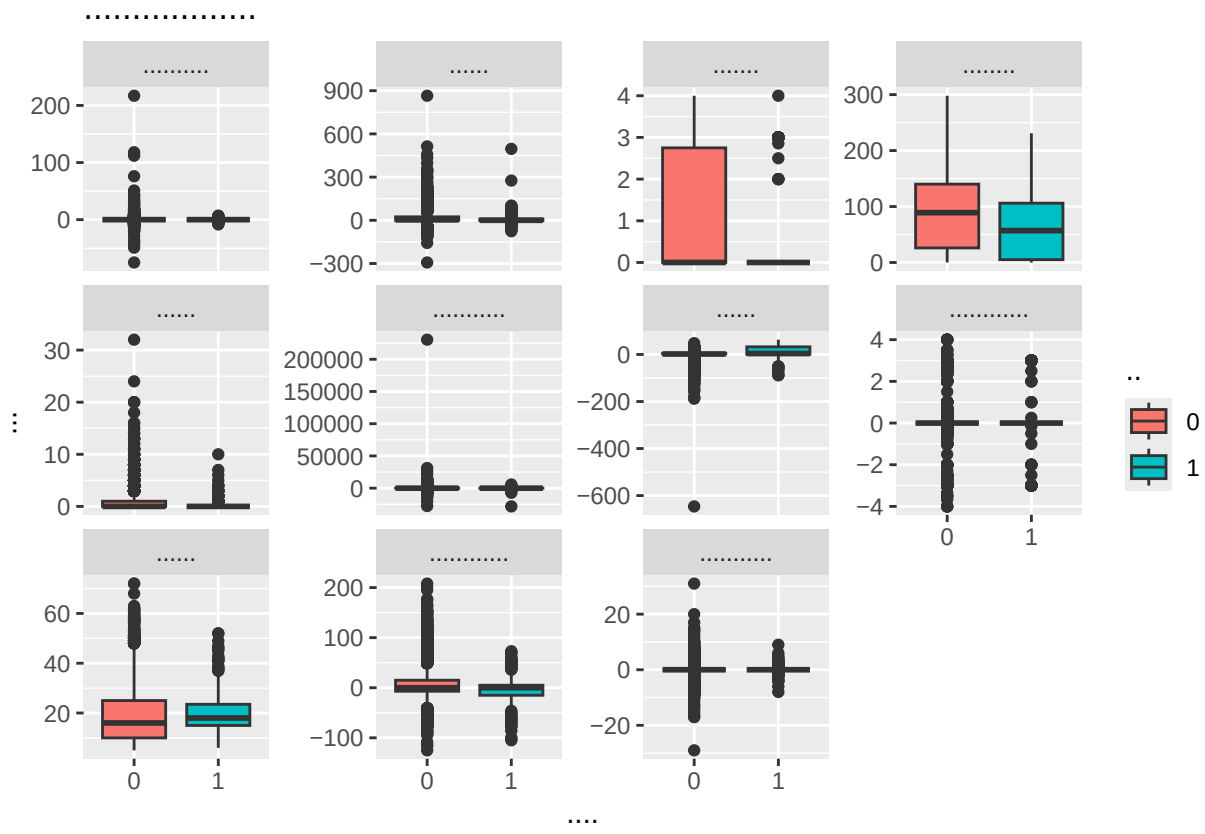
```r
library(readxl)
data_WE <- read_excel("assignment-2-BruceZhaoR/data/WE.xlsx")
colnames(data_WE)
```

```
##  [1] "客户ID"                  "流失"
##  [3] "当月客户幸福指数"        "客户幸福指数相比上月变化"
##  [5] "当月客户支持"            "客户支持相比上月的变化"
##  [7] "当月服务优先级"          "服务优先级相比上月的变化"
##  [9] "当月登录次数"            "博客数相比上月的变化"
## [11] "访问次数相比上月的增加"  "客户使用期限"
## [13] "访问间隔变化"
```

```r
# 将数据中的" 流失" 变量转换为因子，方便绘图分组
data_WE$流失 <- factor(data_WE$流失)

# 绘制多个数值型变量基于流失与否的箱线图，这里假设除了" 流失" 变量外其他变量都是数值型变量，可按需调整
long_data <- tidyr::gather(data_WE, key = "variable_name", value = "value", " 当月客户幸福指数"," 客

ggplot(long_data, aes(x = 流失, y = value, fill = 流失)) +
  geom_boxplot() +
  facet_wrap(~ variable_name, scales = "free_y") +  # 每个变量一个小图，y 轴自由缩放
  labs(title = " 流失客户与非流失客户各指标对比箱线图",
       x = " 是否流失",
       y = " 指标值")
```



通过箱线图可以看出流失客户与非流失客户在“当月服务优先级”、“当月客户幸福指数”、“客户使用期限”、“客户幸福指数相比上月变化”等几个方面有明显差异。

0.0.39 b

```
# 定义一个函数来进行 t 检验并输出结果
compare_means <- function(var_name) {
  group_0 <- data_WE[data_WE$流失 == 0, var_name]
  group_1 <- data_WE[data_WE$流失 == 1, var_name]
  result <- t.test(group_0, group_1)
  return(result)
}

# 对每个数值型变量进行均值比较（除了" 流失" 变量本身）
for (var in  c(" 当月客户幸福指数"," 客户幸福指数相比上月变化", " 当月客户支持"," 客户支持相比上月的变化
  print(paste(" 变量", var, " 的均值比较结果: "))
  print(compare_means(var))
}
```

```
## [1] "变量 当月客户幸福指数 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 7.6, df = 369, p-value = 2e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  18.80 31.87
## sample estimates:
## mean of x mean of y
##     88.61     63.27
##
## [1] "变量 客户幸福指数相比上月变化 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 5.8, df = 366, p-value = 2e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.116 12.418
## sample estimates:
## mean of x mean of y
##     5.530    -3.737
##
## [1] "变量 当月客户支持 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 5.5, df = 419, p-value = 6e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2269 0.4786
## sample estimates:
## mean of x mean of y
##    0.7243    0.3715
##
```

```
## [1] "变量 客户支持相比上月的变化 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = -0.63, df = 407, p-value = 0.5
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19093  0.09803
## sample estimates:
## mean of x mean of y
## -0.009296  0.037152
##
## [1] "变量 当月服务优先级 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 5.1, df = 373, p-value = 4e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2038 0.4562
## sample estimates:
## mean of x mean of y
##    0.8296    0.4996
##
## [1] "变量 服务优先级相比上月的变化 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 0.64, df = 364, p-value = 0.5
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1021  0.2008
## sample estimates:
## mean of x mean of y
##   0.03268  -0.01670
##
```

```
## [1] "变量 当月登录次数 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 3.6, df = 363, p-value = 4e-04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.629 12.525
## sample estimates:
## mean of x mean of y
##    16.139     8.062
##
## [1] "变量 博客数相比上月的变化 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 2.5, df = 696, p-value = 0.01
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06134 0.48529
## sample estimates:
## mean of x mean of y
##    0.1711   -0.1022
##
## [1] "变量 访问次数相比上月的增加 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = 1.9, df = 448, p-value = 0.06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -5.464 410.218
## sample estimates:
## mean of x mean of y
##    106.61    -95.77
##
```

```
## [1] "变量 客户使用期限 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = -3, df = 380, p-value = 0.003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.5461 -0.5223
## sample estimates:
## mean of x mean of y
##     18.82     20.35
##
## [1] "变量 访问间隔变化 的均值比较结果："
##
##  Welch Two Sample t-test
##
## data:  group_0 and group_1
## t = -4.1, df = 346, p-value = 5e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.363 -2.587
## sample estimates:
## mean of x mean of y
##     3.511     8.486
```

在显著性水平为 0.05 的条件下，流失客户与非流失客户在当月客户幸福指数、客户幸福指数相比上月变化、当月客户支持、当月服务优先级、当月登录次数、博客数相比上月的变化、客户使用期限、访问间隔变化这几个指标下均值有显著性差异。

0.0.40 c

```
# 假设通过前面步骤发现变量 var1、var2、var3 对流失与否有显著影响，将其作为自变量
selected_vars <- c(" 当月客户幸福指数"," 客户幸福指数相比上月变化"," 当月客户支持"," 当月服务优先级","
model_q9 <- glm(流失 ~., data = data_WE[, c(" 流失", selected_vars)], family = binomial())
```

```
## Warning: glm.fit:拟合概率算出来是数值零或一
```

```r
summary(model_q9)
```

```
##
## Call:
## glm(formula = 流失 ~ ., family = binomial(), data = data_WE[,
##     c("流失", selected_vars)])
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.87e+00   1.21e-01  -23.66  < 2e-16 ***
## 当月客户幸福指数         -5.22e-03   1.16e-03   -4.50  6.8e-06 ***
## 客户幸福指数相比上月变化 -9.50e-03   2.42e-03   -3.92  8.9e-05 ***
## 当月客户支持             -3.52e-02   7.44e-02   -0.47    0.636
## 当月服务优先级           -3.73e-02   7.51e-02   -0.50    0.620
## 当月登录次数              9.10e-04   1.95e-03    0.47    0.641
## 博客数相比上月的变化     -2.36e-05   2.08e-02    0.00    0.999
## 客户使用期限              1.42e-02   5.26e-03    2.70    0.007 **
## 访问间隔变化              1.70e-02   4.28e-03    3.98  7.0e-05 ***
## 访问次数相比上月的增加   -1.17e-04   4.07e-05   -2.88    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2445.9  on 6337  degrees of freedom
## AIC: 2466
##
## Number of Fisher Scoring iterations: 6
```

0.0.41 d

```r
# 获取尚未流失的客户数据
not_churned_data <- data_WE[data_WE$流失 == 1, ]

# 使用模型预测尚未流失客户的流失概率
probabilities_q9 <- predict(model_q9, newdata = not_churned_data, type = "response")
```

```r
# 将概率添加到尚未流失客户的数据框中
not_churned_data$流失概率 <- probabilities_q9

# 根据流失概率进行降序排序
sorted_data <- not_churned_data[order(not_churned_data$流失概率), ]

# 获取流失可能性最大的前 100 名用户 ID 列表
top_100_user_ids <- sorted_data$客户 ID[1:100]
print(top_100_user_ids)
```

```
##   [1] 4455 2704 1006  764 3984  978  886 5560 1067 1803 4740   66 5422 4178 5271
##  [16] 4793 4817 4642 4313  875 1043  534 5587 1551 3849 5350  780 1641 1764 1810
##  [31] 3682 4299 2011  473  381   94 4899 2578  267 1201  826 1339 4174 3134 2506
##  [46] 3064  596 4925  549 3779 4239 2249  153 2611 3626 4137  574 4006  494 3558
##  [61] 3005 1997 3670 1037 2616 1663  285 1724  738 5765 1895 2357  571 3135 3806
##  [76]  511 1283  472 2212 4055 2869  673 5405  294  590 3494  211 3886 3876  790
##  [91] 5294 4058 4020 4900 3256  564 2073 2682 1947  706
```