# 第 2 次作业

谢宝进

## 目录

**1 问题 1：BigBangTheory** **1**

```
# 加载数据包
library(tidyverse)
library(dplyr)
library(readxl)
```

## 1 问题 1：BigBangTheory

```
BigBangTheory <- read_csv("assignment 2/data/BigBangTheory.csv")
# View(BigBangTheory)
## a 最少观看人数和最多观看人数
min(BigBangTheory$`Viewers (millions)`)
```

```
#> [1] 13.3
```

```
max(BigBangTheory$`Viewers (millions)`)
```

```
#> [1] 16.5
```

```
##b 观看人数的均值、中位数和众数
mean(BigBangTheory$`Viewers (millions)`)
```

```
#> [1] 15.04286
```

```
median(BigBangTheory$`Viewers (millions)`)
```

```
#> [1] 15
```

```
as.numeric(names(sort(table(BigBangTheory$`Viewers (millions)`), decreasing = TRUE)[1]))
```

```
#> [1] 13.6
```

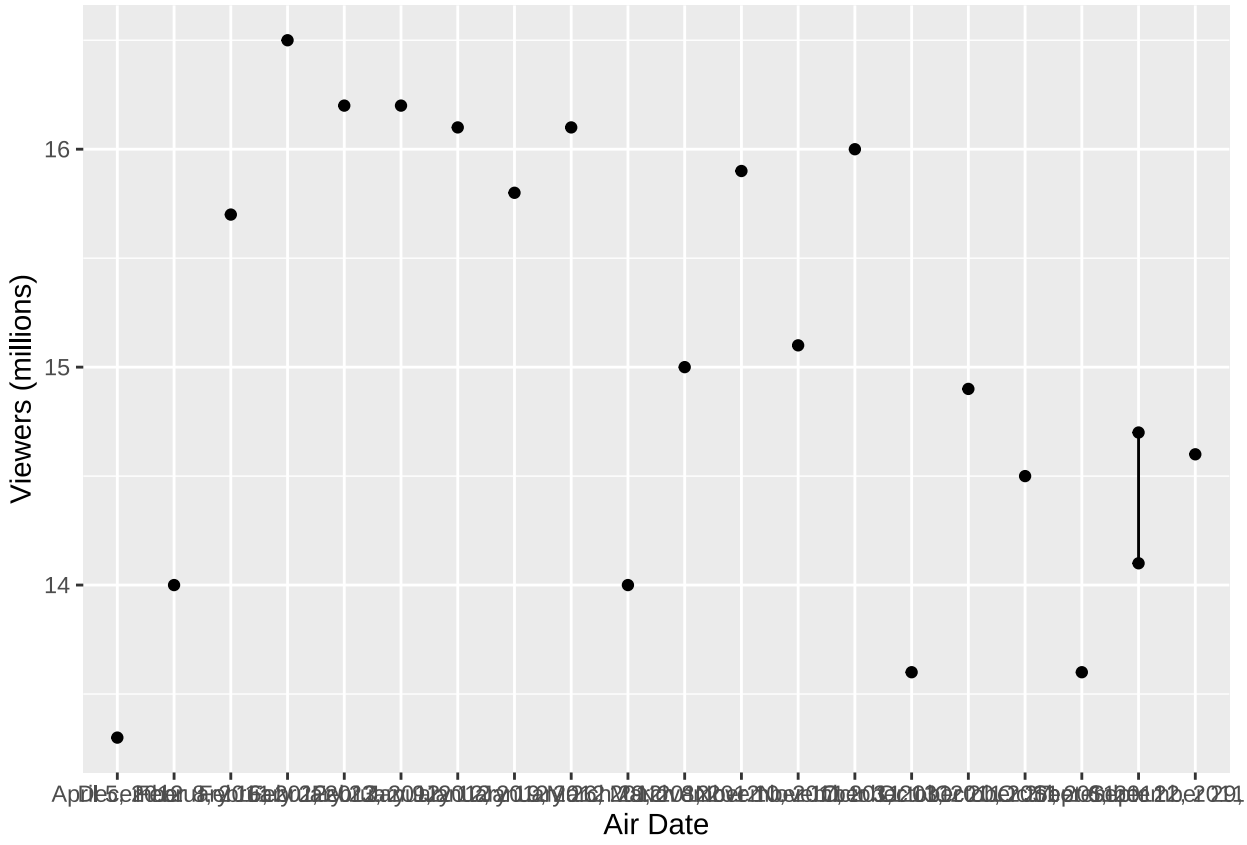
```
##c 第一分位数和第三分位数
quantile(BigBangTheory$`Viewers (millions)`,probs =0.25)
```

```
#>  25%
#> 14.1
```

```
quantile(BigBangTheory$`Viewers (millions)`,probs =0.75)
```

```
#> 75%
#>  16
```

```
##d 讨论 2011-2012 季收视率上升还是下降
ggplot(data=BigBangTheory,aes(x=`Air Date`,y=`Viewers (millions)`))+geom_point()+geom_line()
```




# 问题 2：NBAPlayerPts

```
## 加载数据
data_1 <- read_csv("assignment 2/data/NBAPlayerPts.csv")
```

```
##a 统计频率分布
breaks <- seq(10, 30, by = 2)
frequency <-table(cut(data_1$PPG, breaks = breaks))
print(frequency)
```

```
#>
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>       1       4       6      20       8       4       2       0       3       2
```

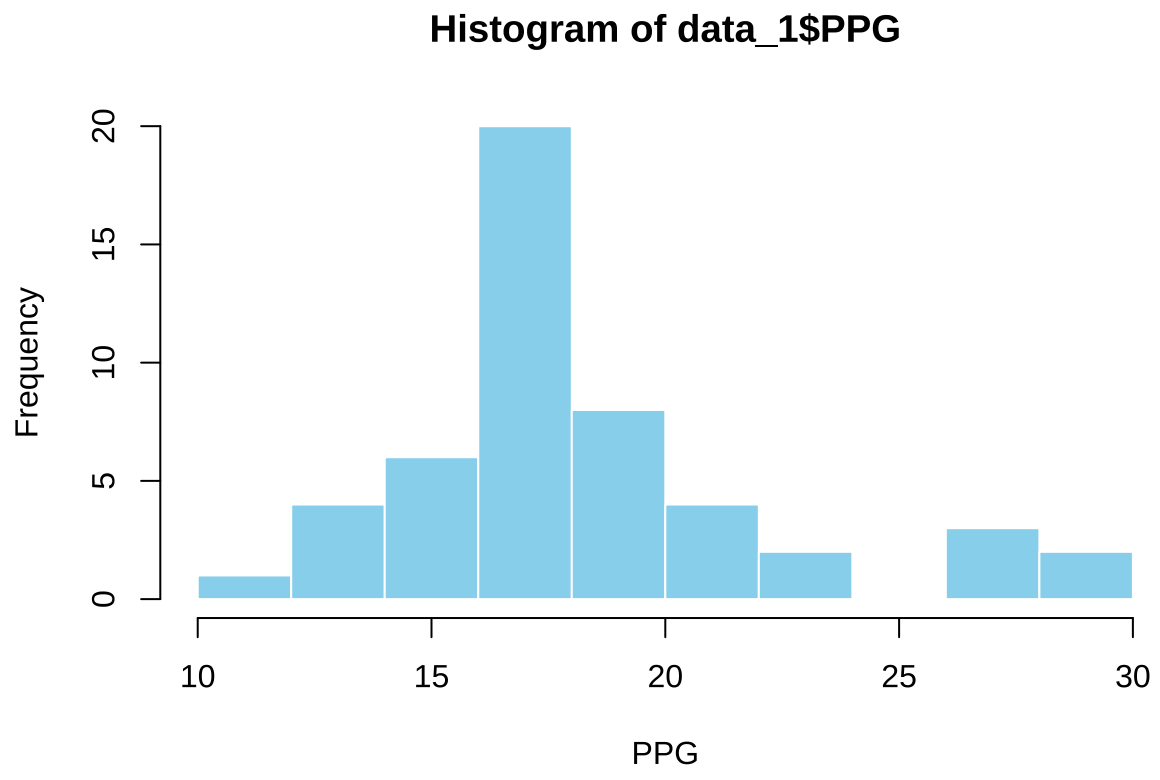
```
##b 相对频率分布
relative_frequency <-prop.table(frequency)
print(relative_frequency)
```

```
#>
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.08    0.12    0.40    0.16    0.08    0.04    0.00    0.06    0.04
```

```
##c 累积百分比分布
cumulative_percent <- cumsum(relative_frequency)
print(cumulative_percent)
```

```
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

```
##d 直方图
hist(data_1$PPG, xlab = "PPG", col = "skyblue", border = "white")
```




```
##e 偏态分析
skewness <- sum((data_1$PPG - mean(data_1$PPG))^3) / (length(data_1$PPG) * sd(data_1$PPG)^3)
```

```r
# f. 每场比赛得分至少 20 分的球员百分比
percentage_20plus <- mean(data_1$PPG >= 20)
print(percentage_20plus)
```

```
#> [1] 0.22
```

# 问题 3：调查样本及概率问题

```r
##a 样本大小？
# 已知值
std_error <- 20
std_dev <- 500
n <- (std_dev / std_error)^2
print(n)
```

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

```r
##b ±25 内的概率
z <- 25 / std_error
prob <- pnorm(z) - pnorm(-z)
print(prob)
```

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

# 问题 4：青年杂志

```r
# 加载数据
data_2 <- read_csv("assignment 2/data/Professional.csv")
view(data_2)
```

```r
##a 描述性统计
summary(data_2)
```

```
#>       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
#>
#>
#>
#>
```

```
##b 年龄和收入在 95% 执行水平的置信区间
t.test(data_2$Age,conf.level = 0.95)$conf.int  # 计算年龄的区间
```

```
#> [1] 29.72153 30.50286
#> attr(,"conf.level")
#> [1] 0.95
```

```
t.test(data_2$`Household Income ($)`,conf.level = 0.95)$conf.int # 计算收入的区间
```

```
#> [1] 71079.26 77839.77
#> attr(,"conf.level")
#> [1] 0.95
```

# 问题 5: Quality Associate

```r
##a 对每个样本进行假设检验，并确定应采取哪些措施。

quality <- read_csv("assignment 2/data/Quality.csv")  # 导入数据

alpha <- 0.01  # 定义显著性水平

sample_means <- apply(quality, 1, mean)  # 计算样本均值
sample_sds <- apply(quality, 1, sd) # 计算标准差

# 总体标准差
sigma <- 0.21
n <- 30

# 进行假设检验
t_tests <- sapply(1:nrow(quality), function(i) {
  t_stat <- (sample_means[i] - 12) / (sigma / sqrt(n))
  p_value <- 2 * pt(abs(t_stat), df = n - 1, lower.tail = FALSE)
  list(t_stat = t_stat, p_value = p_value)
})

t_tests # 输出结果
```

```
#>         [,1]        [,2]        [,3]         [,4]       [,5]      [,6]
#> t_stat  -5.868456   -8.541864   -7.107352    -1.956152  0.7824608 -5.607636
#> p_value 2.28638e-06 2.07292e-09 8.055516e-08 0.06014044 0.4402866 4.696334e-06
#>         [,7]         [,8]      [,9]        [,10]       [,11]       [,12]
#> t_stat  -4.368739    -2.477793 3.260253    -3.455869   3.651484    -1.564922
#> p_value 0.0001458471 0.0192882 0.002843649 0.001711572 0.001021295 0.1284495
#>         [,13]      [,14]      [,15]        [,16]      [,17]      [,18]
#> t_stat  2.347382   -0.4564355 4.49915      1.890947   -0.5216405 -5.738046
#> p_value 0.02594231 0.6514771  0.0001017738 0.06865983 0.6058821  3.275305e-06
#>         [,19]      [,20]       [,21]      [,22]      [,23]      [,24]
#> t_stat  2.673408   -3.129843   -0.1956152 2.412587   1.825742   8.867889
#> p_value 0.01220001 0.003967029 0.8462756  0.02239019 0.07820332 9.360888e-10
#>         [,25]     [,26]     [,27]      [,28]      [,29]        [,30]
#> t_stat  1.369306  0.9128709 1.956152   -2.412587  5.085995     3.586279
#> p_value 0.1814154 0.3688376 0.06014044 0.02239019 1.997275e-05 0.001214172
```

```
##b 计算 4 个样本的标准差，并判断假设是否合理

sample_sds
```

```
#>  [1] 0.22575798 0.27170756 0.23556669 0.18912077 0.11401754 0.19330460
#>  [7] 0.18191115 0.20566964 0.19908122 0.14930394 0.04082483 0.24589971
#> [13] 0.14719601 0.33129795 0.21515498 0.18839232 0.11401754 0.03366502
#> [19] 0.19050372 0.16268579 0.29136175 0.13524669 0.15165751 0.10614456
#> [25] 0.22156639 0.09678154 0.15286159 0.05057997 0.21763884 0.14453950
```

```
# 判断假设是否合理
mean(sample_sds)  # 计算样本标准差的平均值
```

```
#> [1] 0.1734486
```

```
##c 计算样本的界限，并判断新样本是否在界限内，如果不在则需采取纠正措施。
# 计算控制限
upper_limit <- 12 + 3 * (sigma / sqrt(n))
lower_limit <- 12 - 3 * (sigma / sqrt(n))

# 输出界限
c(upper_limit, lower_limit)
```

```
#> [1] 12.11502 11.88498
```

```
##d. 讨论将显著性水平提高到更大值的影响。如果将显著性水平提高，哪种错误可能会增加？

# 如果显著性水平增加，第一类错误（错误地拒绝正确的零假设）的风险会增加。
# 这意味着可能会更频繁地采取不必要的纠正措施，导致成本增加和生产效率降低。
```

# 问题 6：入住率

```
occupancy <- read_csv("assignment 2/data/Occupancy.csv") # 数据导入

##a 估算 2007 年 3 月第一周和 2008 年 3 月第一周的单元出租比例

# 将原始数据转化为 0/1，区分是否出租
occupancy$`Mar-07` <- ifelse(occupancy$`Mar-07`=="Yes",1,0)
```

```r
occupancy$`Mar-08` <- ifelse(occupancy$`Mar-08`=="Yes",1,0)

# 计算样本大小
n <- nrow(occupancy)

# 计算 2007 年 3 月第一周出租单位的比例
prop_2007 <- mean(occupancy$`Mar-07`)

# 计算 2008 年 3 月第一周出租单位的比例
prop_2008 <- mean(occupancy$`Mar-08`)

# 打印结果
cat("Proportion of units rented in Mar.07:", prop_2007, "\n")
```

```
#> Proportion of units rented in Mar.07: 0.35
```

```r
cat("Proportion of units rented in Mar.08:", prop_2008, "\n")
```

```
#> Proportion of units rented in Mar.08: NA
```

```r
##b 为比例之差提供一个 95% 的置信区间

se_diff <- sqrt((prop_2007 * (1 - prop_2007) / n) + (prop_2008 * (1 - prop_2008) / n)) # 计算比例差

ci_diff <- c(prop_2008 - prop_2007 - 1.96 * se_diff, prop_2008 - prop_2007 + 1.96 * se_diff)  # 计

# 输出结果
cat("95% Confidence Interval for the difference in proportions:", ci_diff, "\n")
```

```
#> 95% Confidence Interval for the difference in proportions: NA NA
```

```r
##c 根据发现，2008 年 3 月的租赁率是否会相比前一年同期有所上涨？

  cat(" 有明显证据证明 08 年 3 月租金会同比上升.\n")
```

```
#> 有明显证据证明08年3月租金会同比上升.
```

# 问题 7：空军训练方案

```r
Training <- read_csv("assignment 2/data/Training.csv") # 导入数据
##a use appropriate descriptive statistics to summarize the training time data for each method.
#what similarities or differences do you observe from the sample data?
# 描述性统计
summary_current <- summary(Training$Current)
summary_proposed <- summary(Training$Proposed)
# 打印结果
cat(" 当前方法的描述性统计:\n")
```

```
#> 当前方法的描述性统计:
```

```r
print(summary_current)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   65.00   72.00   76.00   75.07   78.00   84.00
```

```r
cat(" 提议方法的描述性统计:\n")
```

```
#> 提议方法的描述性统计:
```

```r
print(summary_proposed)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   69.00   74.00   76.00   75.43   77.00   82.00
```

```r
##b Comment on any difference between the population means for the two methods. Discuss
# your findings.
# t 检验
t_test_result <- t.test(Training$Current, Training$Proposed, var.equal = TRUE)
print(t_test_result)
```

```
#>
#>  Two Sample t-test
#>
#> data:  Training$Current and Training$Proposed
#> t = -0.60268, df = 120, p-value = 0.5479
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
```

```
#>  -1.5454793  0.8241679
#> sample estimates:
#> mean of x mean of y
#>  75.06557  75.42623
```

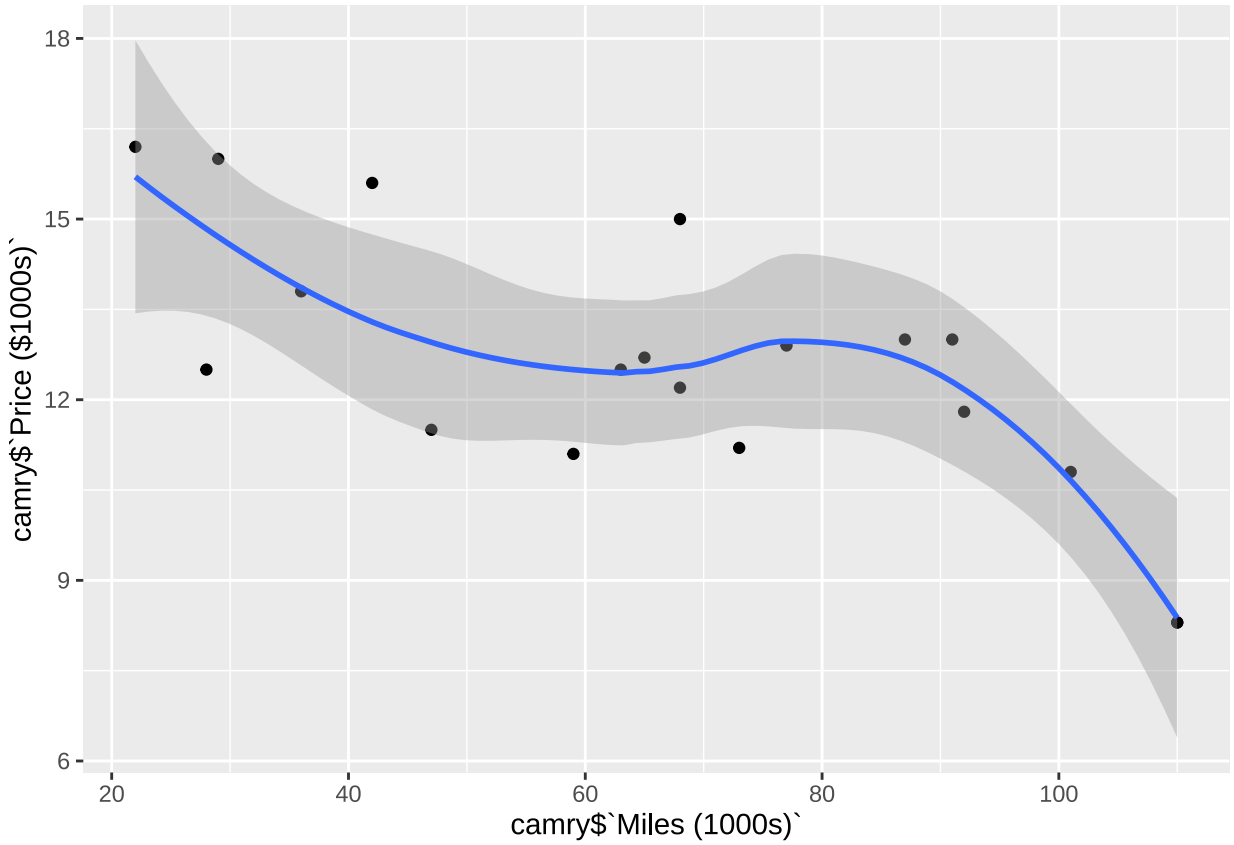
```
#3.c. compute the standard deviation and variance for each training method. conduct a hypothesis
#test about the equality of population variances for the two training methods. Discuss your
#findings
# 计算标准差和方差
sd_current <- sd(Training$Current)
var_current <- var(Training$Current)
sd_proposed <- sd(Training$Proposed)
var_proposed <- var(Training$Proposed)

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

```
#>
#>  F test to compare two variances
#>
#> data:  Training$Current and Training$Proposed
#> F = 2.4773, num df = 60, denom df = 60, p-value = 0.000578
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  1.486267 4.129135
#> sample estimates:
#> ratio of variances
#>           2.477296
```

```
#4. what conclusion can you reach about any differences between the two methods? what is your
#recommendation? explain
# 结论：两种方法的均值没有显著差异，但方差存在显著差异，表明提议方法在训练时间上更加一致。
#5.can you suggest other data or testing that might be desirable before making a final decision
#on the training program to be used in the future?
# 鉴于两种方法的均值相似，但提议方法的方差较小，可能更值得考虑采用提议方法，因为它可能提供更一致的训练体验
```

# 问题 8: 凯美瑞

```
# 加载数据
camry <-  read_csv("assignment 2/data/Camry.csv")
##a 以汽车里程数为横轴，价格为纵轴绘制散点图
#  绘制散点图
ggplot(camry, aes(x = camry$`Miles (1000s)`, y = camry$`Price ($1000s)`)) +
  geom_point() +  # 添加散点图
  geom_smooth()  # 添加默认的平滑拟合线
```




```
##b 在部分（a）中绘制的散点图显示了这两个变量之间有什么关系？
# 从散点图中，我们可以看出里程和价格之间存在负相关关系。随着里程的增加，价格呈下降趋势

##c 根据里程数（以千为单位）来开发一个估计的回归方程，用于预测价格（以千美元为单位）。
# 估计回归方程
model <- lm(camry$`Price ($1000s)` ~ camry$`Miles (1000s)`, data=camry)
summary(model)
```

```
#>
#> Call:
```

```
#> lm(formula = camry$`Price ($1000s)` ~ camry$`Miles (1000s)`,
#>     data = camry)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.32408 -1.34194  0.05055  1.12898  2.52687
#>
#> Coefficients:
#>                       Estimate Std. Error t value Pr(>|t|)
#> (Intercept)           16.46976    0.94876  17.359 2.99e-12 ***
#> camry$`Miles (1000s)` -0.05877    0.01319  -4.455 0.000348 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.541 on 17 degrees of freedom
#> Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
#> F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

```
##d 在 0.05 的显著性水平下检验是否存在显著关系。
# 答：从回归方程的估计结果中，我们可以看到 Miles 的 p 值为 0.000348，远小于 0.05，因此在 0.05 的显著性水

##e  估计的回归方程是否提供了良好的拟合？请解释。
# 答：R 平方值 0.5387，调整后的 R 平方值为 0.5115，这表明回归方程对数据的拟合度较好，可以解释 51.15% 的

##f 对估计的回归方程的斜率进行解释
# 答：斜率-0.05877 表示每增加 1000 英里的里程，价格平均下降 0.05877 千美元。

##g 假设你正在考虑购买一辆已经行驶了 60,000 英里的二手 2007 款凯美瑞。使用在部分（c）中开发的估计回归方程

# 获取模型的系数
coefficients <- coef(model)
intercept <- coefficients[1]
slope <- coefficients[2]

# 计算预测价格
predicted_price <- intercept + slope * 60
predicted_price
```

```
#> (Intercept)
```

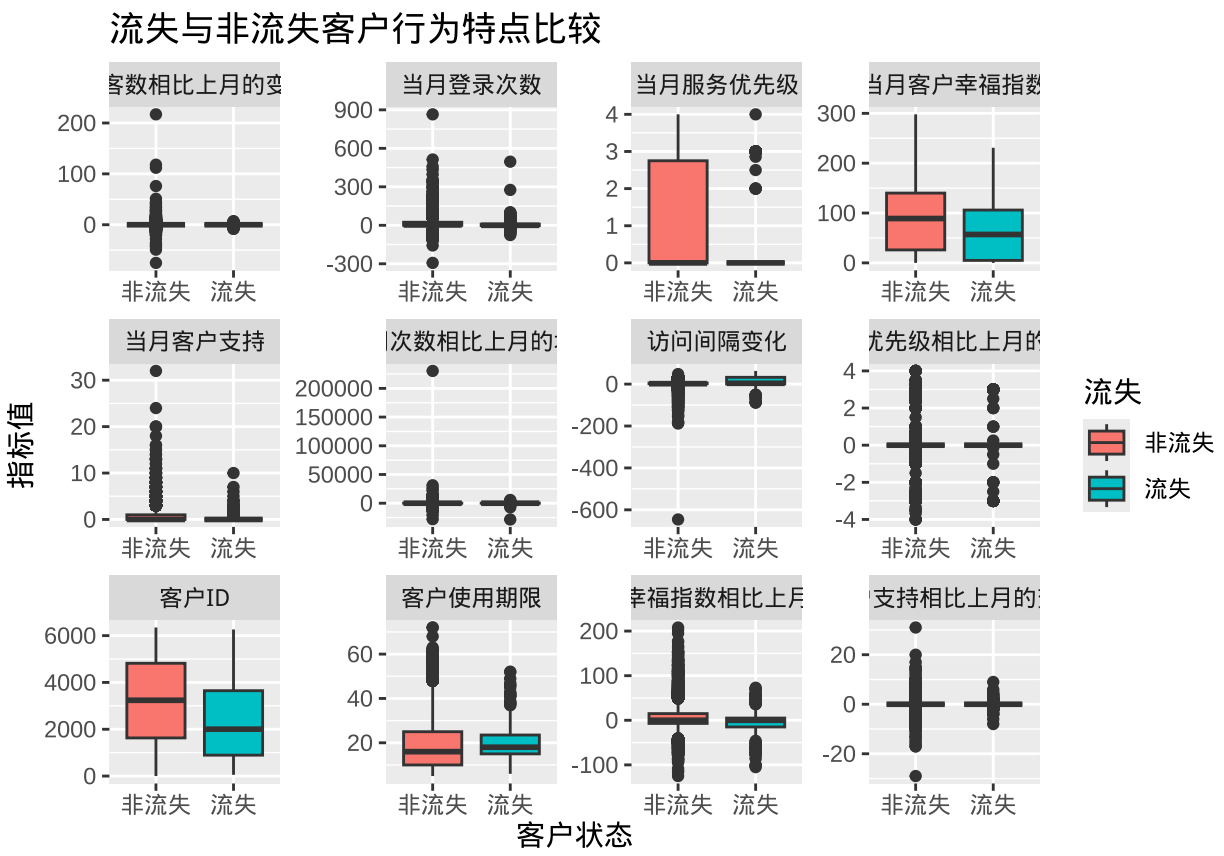
```
#>     12.94332
```

```
predicted_price_2 <- round(predicted_price, 2)
paste(" 预测价格 $",predicted_price_2," 千元")
```

```
#> [1] "预测价格$ 12.94 千元"
```

# 问题 9：流失率

```
data <- read_excel("assignment 2/data/WE.xlsx")  # 导入数据
##a 通过可视化探索流失客户与□流失客户的□为特点（或特点对□），你能发现流失与□流失客
  # 户□为在哪些指标有可能存在显著不同？
data_long <- data %>%
  pivot_longer(cols = -流失, names_to = " 指标", values_to = " 值") %>%
  mutate(流失 = factor(流失, labels = c(" 非流失", " 流失")))

ggplot(data_long, aes(x = 流失, y = 值, fill = 流失)) +
  geom_boxplot() +
  facet_wrap(~指标, scales = "free") +
  labs(title = " 流失与非流失客户行为特点比较", x = " 客户状态", y = " 指标值")
```




```
##b 通过均值□较的□式验证上述不同是否显著。
# 计算均值并进行 t 检验
t_tests <- data %>%
  pivot_longer(cols = -流失, names_to = " 指标", values_to = " 值") %>%
  group_by(指标) %>%
  do({
    t_test <- t.test(.$值 [.$流失 == 0], .$值 [.$流失 == 1])
    data.frame(指标 = unique(.$指标), 非流失_mean = mean(.$值 [.$流失 == 0], na.rm = TRUE), 流失_mea
  }) %>%
  ungroup()

# 显示 t 检验结果
t_tests
```

```
#> # A tibble: 12 x 4
#>    指标                    非流失_mean 流失_mean  p_value
#>    <chr>                         <dbl>     <dbl>    <dbl>
#>  1 博客数相比上月的变化          0.171    -0.102  1.16e- 2
```

```
#>  2 客户ID                      3219.      2330.     5.98e-20
#>  3 客户使用期限                  18.8       20.4    3.06e- 3
#>  4 客户幸福指数相比上月变化       5.53      -3.74   1.57e- 8
#>  5 客户支持相比上月的变化        -0.00930    0.0372 5.28e- 1
#>  6 当月客户幸福指数              88.6       63.3    2.10e-13
#>  7 当月客户支持                  0.724      0.372   6.28e- 8
#>  8 当月服务优先级                0.830      0.500   4.38e- 7
#>  9 当月登录次数                  16.1       8.06    4.04e- 4
#> 10 服务优先级相比上月的变化      0.0327    -0.0167 5.22e- 1
#> 11 访问次数相比上月的增加       107.       -95.8    5.63e- 2
#> 12 访问间隔变化                  3.51       8.49    5.22e- 5
```

```
print(t_tests)
```

```
#> # A tibble: 12 x 4
#>    指标                      非流失_mean 流失_mean  p_value
#>    <chr>                           <dbl>     <dbl>    <dbl>
#>  1 博客数相比上月的变化            0.171    -0.102  1.16e- 2
#>  2 客户ID                      3219.      2330.     5.98e-20
#>  3 客户使用期限                  18.8       20.4    3.06e- 3
#>  4 客户幸福指数相比上月变化       5.53      -3.74   1.57e- 8
#>  5 客户支持相比上月的变化        -0.00930    0.0372 5.28e- 1
#>  6 当月客户幸福指数              88.6       63.3    2.10e-13
#>  7 当月客户支持                  0.724      0.372   6.28e- 8
#>  8 当月服务优先级                0.830      0.500   4.38e- 7
#>  9 当月登录次数                  16.1       8.06    4.04e- 4
#> 10 服务优先级相比上月的变化      0.0327    -0.0167 5.22e- 1
#> 11 访问次数相比上月的增加       107.       -95.8    5.63e- 2
#> 12 访问间隔变化                  3.51       8.49    5.22e- 5
```

```
##c 以”流失“为因变量，其他你认为重要的变量为□变量（提示：a、b 两步的发现），建□回归□
  # 程对是否流失进□预测。
model <- glm(流失 ~ 客户 ID + 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

# 显示模型摘要
summary(model)
```

```
#>
#> Call:
```

```
#> glm(formula = 流失 ~ 客户ID + 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, family = binomial, data = data)
#>
#> Coefficients:
#>                            Estimate Std. Error z value Pr(>|z|)
#> (Intercept)              -1.211e+00  1.359e-01  -8.912   <2e-16 ***
#> 客户ID                   -3.539e-04  3.366e-05 -10.516   <2e-16 ***
#> 当月客户幸福指数         -9.305e-03  1.125e-03  -8.267   <2e-16 ***
#> 客户幸福指数相比上月变化 -4.194e-03  2.285e-03  -1.835   0.0665 .
#> 当月客户支持              6.730e-03  6.822e-02   0.099   0.9214
#> 当月服务优先级           -3.799e-02  7.307e-02  -0.520   0.6031
#> ---
#> 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: 2371.3  on 6341  degrees of freedom
#> AIC: 2383.3
#>
#> Number of Fisher Scoring iterations: 6
```

```
##d 根据上□步预测的结果，对尚未流失（流失 =0）的客户进□流失可能性排序，并给出流失可能
  # 性最□的前 100 名□户 ID 列表。

# 筛选出尚未流失的客户
data_non_churn <- data[data$流失 == 0, ]

# 预测尚未流失的客户流失可能性
predictions <- predict(model, newdata = data_non_churn, type = "response")

# 将预测结果添加到筛选后的数据框中
data_non_churn$predictions <- predictions

# 对尚未流失的客户进行排序
data_non_churn_sorted <- data_non_churn[order(-data_non_churn$predictions), ]

# 显示流失可能性最大的前 100 名用户 ID
```

```r
top100_users <- head(data_non_churn_sorted$客户 ID, 100)
print(top100_users)
```

```
#>   [1]  109   76   57  318  305  240  183    1  271    3   14   18   21  110   59
#>  [16]   51  703  123  101  104  106  228  119  121  146  425   55  137  154  165
#>  [31]  415  171  407  190  246  212  142  244  254   68  272  278  279   95   61
#>  [46]  572  346 1141  641  374  376  704  400   75  413  416 1181  423  427   89
#>  [61]  440  798  444   69   64  475  839  488  622  526  508  882  203  551  207
#>  [76]  570  583   62  777  846  604 1574  623  625  141 1971  128  210  645  651
#>  [91]  563  678  689  302   42  585  871 1520  350 1010
```