huyanxiaoxi

hyyn

目录

```
---
documentclass: ctexart
output: rticles::ctex
---
```

```
devtools::install_github(c('rstudio/rmarkdown', 'yihui/tinytex'))
tinytex::install_tinytex()
```

1 Question #1: BigBangTheory. (Attached Data: BigBangTheory)

```
bigbang <- read.csv("BigBangTheory.csv")
```

计算观众数量的最小值和最大值，计算均值、中位数和众数，计算第一和第三四分位数

```
min_viewers <- min(bigbang$Viewers..millions., na.rm = TRUE)
max_viewers <- max(bigbang$Viewers..millions., na.rm = TRUE)
mean_viewers <- mean(bigbang$Viewers..millions., na.rm = TRUE)
median_viewers <- median(bigbang$Viewers..millions., na.rm = TRUE)
mode_viewers <- modeest::mfv(bigbang$Viewers..millions., na.rm = TRUE)
first_quartile <- quantile(bigbang$Viewers..millions., 0.25, na.rm = TRUE)
third_quartile <- quantile(bigbang$Viewers..millions., 0.75, na.rm = TRUE)
```

查看分析结果

```
result1 <- data.frame(最小值 = min_viewers, 最大值 = max_viewers, 均值 = mean_viewers, 中位数 = medi
result2 <- data.frame(众数 = mode_viewers)
result1
```

```
#>       最小值 最大值      均值 中位数 第一四分位数 第三四分位数
#> 25%   13.3   16.5 15.04286     15          14.1           16
```

```
result2
```

```
#>   众数
#> 1 13.6
#> 2 14.0
#> 3 16.1
#> 4 16.2
```

问题 1.2:has viewership grown or declined over the 2011–2012 season? Discuss.

```
library(tidyverse)
library(lubridate)
df <- read.csv("BigBangTheory.csv", stringsAsFactors = FALSE, fileEncoding = "UTF-8")
```

将 Air.Date 列转换为日期格式

```
df$Air.Date <- mdy(df$Air.Date)
```

提取年份和季度作为标识、筛选出 2011 到 2012 年的数据

```
df$Year <- year(df$Air.Date)
df$Quarter <- quarter(df$Air.Date)
df_subset <- df %>% filter(Year >= 2011 & Year <= 2012)
```

按照年份和季度分组，并计算平均收视率

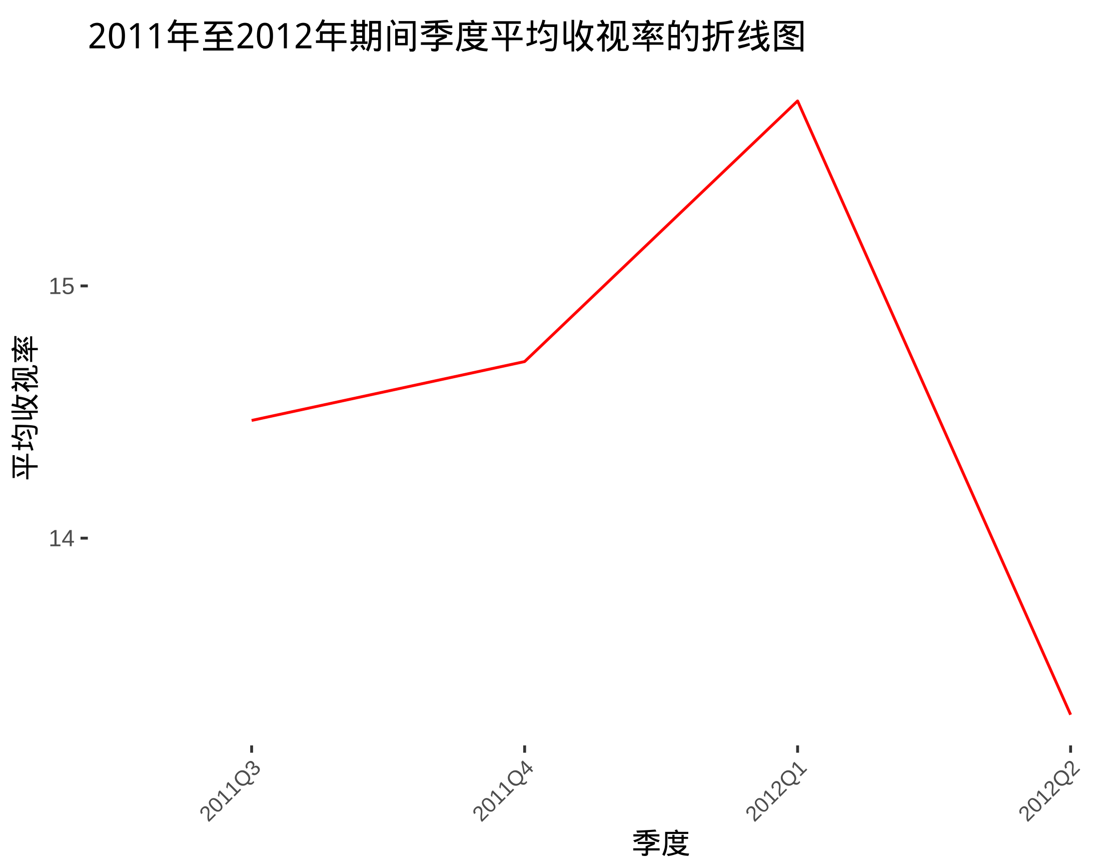
```
quarterly_data <- df_subset %>%
  group_by(Year, Quarter) %>%
  summarise(ave_viewer = mean(Viewers..millions.),.groups = "drop") %>%
  mutate(Quarter_Label = paste(Year, "Q", Quarter, sep = ""))

print(quarterly_data)
```

```
#> # A tibble: 4 x 4
#>    Year Quarter ave_viewer Quarter_Label
#>   <dbl>   <int>      <dbl> <chr>
#> 1  2011       3       14.5 2011Q3
#> 2  2011       4       14.7 2011Q4
#> 3  2012       1       15.7 2012Q1
#> 4  2012       2       13.3 2012Q2
```

使用 ggplot2 绘制图表

```
ggplot(quarterly_data, aes(x = Quarter_Label, y = ave_viewer, group = 1)) +
    geom_line(color = "red") +
    labs(title = "2011 年至 2012 年期间季度平均收视率的折线图",
         x = " 季度",
         y = " 平均收视率") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank())
```



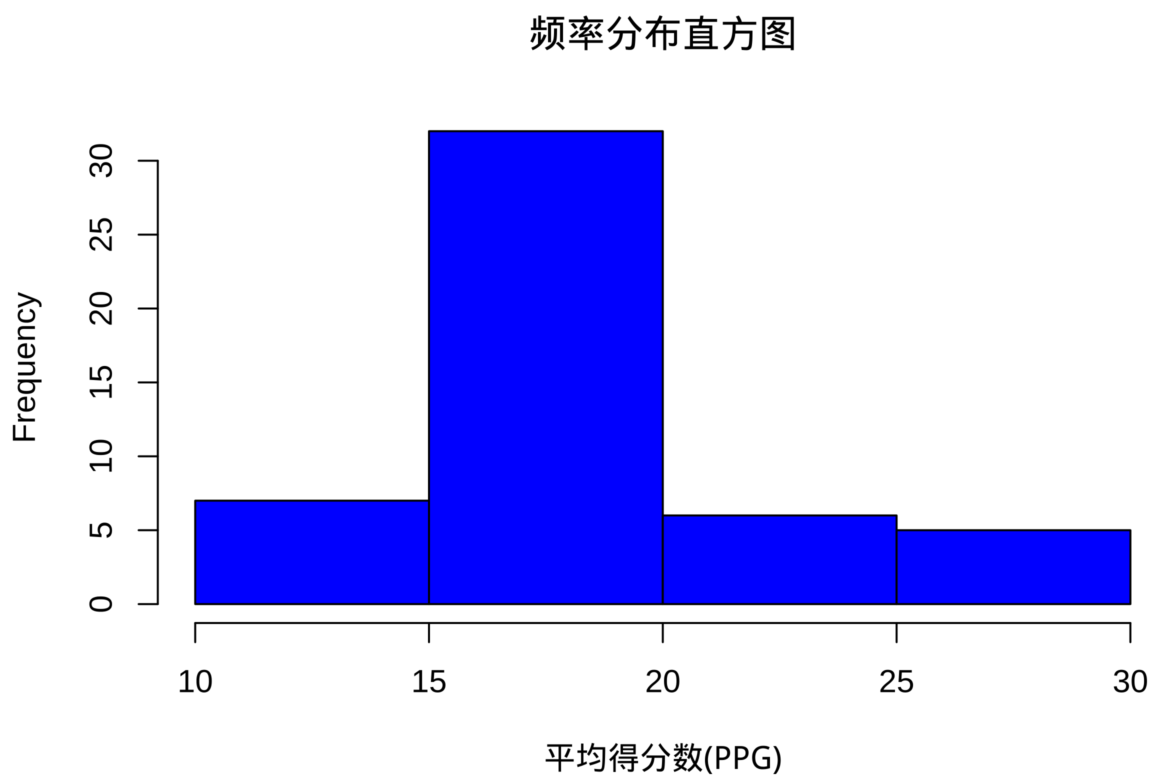
2 Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

加载数据

```
nba <- read.csv("NBAPlayerPts.csv")
```

a. 显示频率分布

```
hist(nba$PPG, breaks = seq(10, 30, by = 5), xlab = '平均得分数 (PPG)', main = '频率分布直方图', col
```



查看并 print 结果

```
hist_df <- hist(nba$PPG, breaks = seq(10, 30, by = 5),
                plot = FALSE)$counts
print('频率分布')
```

```
#> [1] "频率分布"
```

```
print(hist_df)
```

```
#> [1]  7 32  6  5
```

b. 显示相对频率分布

```
hist_freq <- hist(nba$PPG, breaks = seq(10, 30, by = 5),
                  plot = FALSE)
hist_freq$density <- hist_freq$counts / sum(hist_freq$counts) * 4
```

查看并 print 结果

```
print('相对频率分布')
```

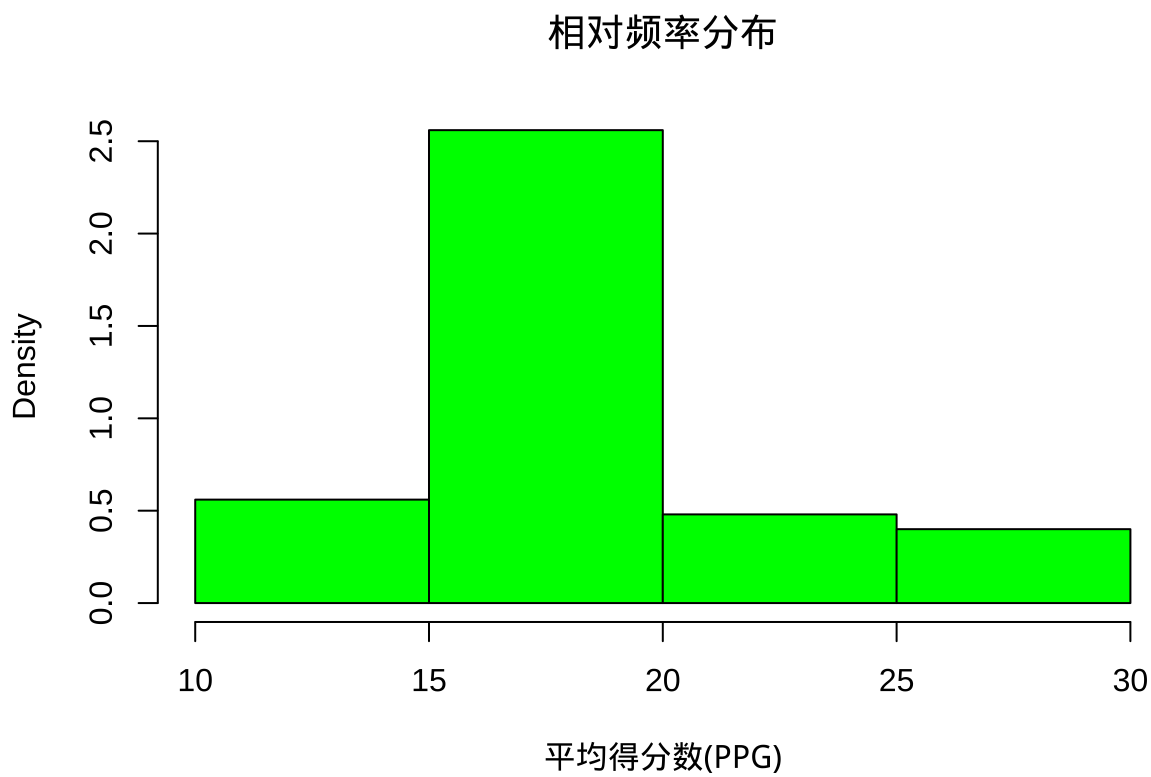
```
#> [1] "相对频率分布"
```

```
print(round(hist_freq$density, 2))
```

```
#> [1] 0.56 2.56 0.48 0.40
```

绘制相对频率分布直方图

```
plot(hist_freq, freq = FALSE, xlab = '平均得分数 (PPG)',
     main = '相对频率分布', col = 'green')
```



c. 显示累积百分比频率分布

```
``` r
hist_cum <- hist(nba$PPG, breaks = seq(10, 30, by = 5),
                 plot = FALSE)
hist_cum$counts <- hist_cum$counts / sum(hist_cum$counts)
```

求累计比例

```r
hist_cum$accumulative_count <- cumsum(hist_cum$counts)
```

将小数转换为百分比并查看结果

```r
print('累积百分比频率分布')
```

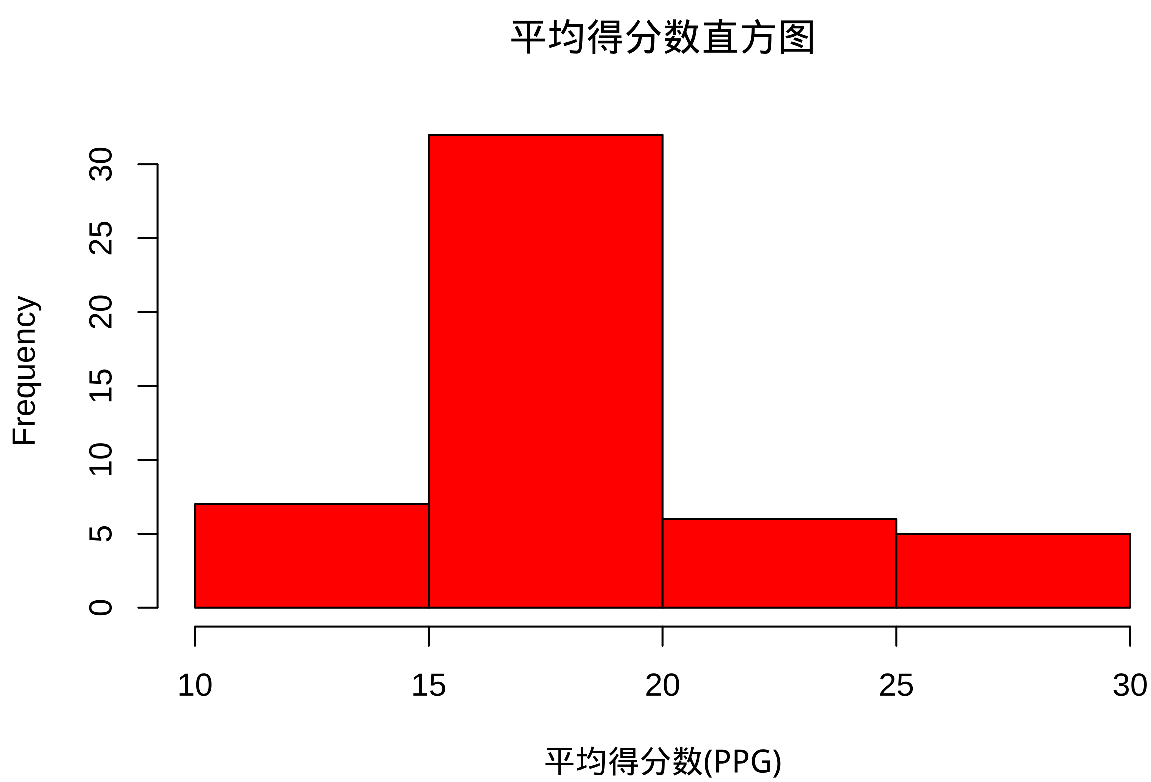
```
#> [1] "累积百分比频率分布"
```
```

```
print(round(hist_cum$accumulative_count * 100, 2))
```

```
#> [1]  14  78  90 100
```

d. 展示平均得分数直方图

```
hist(nba$PPG, breaks = seq(10, 30, by = 5),
     xlab = '平均得分数 (PPG)', main = '平均得分数直方图', col = 'red')
```



e. It seems skewed rightly, for it has a long tail to the right.

f. 1-78% = 22%.

3 Question #3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500

```
simga <- 500
SE <- 20
```

根据公式计算样本 n

```
n <- (simga/SE)^2
```

print 样本 n

```
print(paste(" 样本量为: ",n))
```

```
#> [1] "样本量为:  625"
```

b. 计算人口在 ±25 内的概率转化成标准正态分布

```
standard <- simga/sqrt(n)
z_low <- (-25)/standard
z_upr <- 25/standard
```

使用 pnorm 函数计算数值、结果百分比转换

```
probability <- pnorm(z_upr)-pnorm(z_low)
probability_percentage <- round(probability*100,1)
print(paste(" 人口在 ±25 内概率为:",probability_percentage,"%"))
```

```
#> [1] "人口在±25内概率为: 78.9 %"
```

4 Question #4: Young Professional Magazine (Attached Data: Professional)

```r
professional <- read.csv('Professional.csv')
```

#1.Develop appropriate descriptive statistics to summarize the data.

```r
summary(professional)
```

```
#>       Age           Gender          Real.Estate.Purchases Value.of.Investments
#>  Min.   :19.00   Length:410         Length:410            Min.   :     0
#>  1st Qu.:28.00   Class :character   Class :character      1st Qu.: 18300
#>  Median :30.00   Mode  :character   Mode  :character      Median : 24800
#>  Mean   :30.11                                            Mean   : 28538
#>  3rd Qu.:33.00                                            3rd Qu.: 34275
#>  Max.   :42.00                                            Max.   :133400
#>  Number.of.Transactions Broadband.Access   Household.Income Have.Children
#>  Min.   : 0.000         Length:410         Min.   : 16200   Length:410
#>  1st Qu.: 4.000         Class :character   1st Qu.: 51625   Class :character
#>  Median : 6.000         Mode  :character   Median : 66050   Mode  :character
#>  Mean   : 5.973                            Mean   : 74460
#>  3rd Qu.: 7.000                            3rd Qu.: 88775
#>  Max.   :21.000                            Max.   :322500
```

#2.Develop 95% confidence intervals for the mean age and household income of subscribers

```r
age_t <- t.test(professional$Age,conf.level = 0.95)$conf.int
print(paste(" 年龄 95% 的置信区间为: [", round(age_t[1], 2), ", ", round(age_t[2], 2), "]"))
```

```
#> [1] "年龄95%的置信区间为: [ 29.72 ,  30.5 ]"
```

```r
names <- names(professional)
names[7] <- "Household.Income"
colnames(professional) <- names
income_t <- t.test(professional$Household.Income,conf.level = 0.95)$conf.int
print(paste(" 收入 95% 的置信区间为: [", round(income_t[1], 2), ", ", round(income_t[2], 2), "]"))
```

```
#> [1] "收入95%的置信区间为: [ 71079.26 ,  77839.77 ]"
```

#3.Develop 95% confidence intervals for the proportion of subscribers who have broadband #access at home and the proportion of subscribers who have children # 总用户数

```r
total_users <- 410
```

有宽带接入的用户数

```r
broadband_users <- sum(grepl("Yes", professional$Broadband.Access))
```

有孩子的用户数

```r
children_users <- sum(grepl("Yes", professional$Have.Children))
```

计算有宽带接入的用户比例的 95% 置信区间

```r
ci_broadband <- prop.test(x = broadband_users, n = total_users)
```

计算有孩子的用户比例的 95% 置信区间、打印结果

```r
ci_children <- prop.test(x = children_users, n = total_users)
cat("95% Confidence Interval for Broadband Access:\n")
```

```
#> 95% Confidence Interval for Broadband Access:
```

```r
print(ci_broadband$conf.int)
```

```
#> [1] 0.5753252 0.6710862
#> attr(,"conf.level")
#> [1] 0.95
```

```r
cat("\n95% Confidence Interval for Having Children:\n")
```

```
#>
#> 95% Confidence Interval for Having Children:
```

```r
print(ci_children$conf.int)
```

```
#> [1] 0.4845521 0.5830908
#> attr(,"conf.level")
#> [1] 0.95
```

#4. Would Young Professional be a good advertising outlet for online brokers? Justify your #conclusion with statistical data.

```
library(dplyr)
```

```
# 删除缺失值
professionals <- na.omit(professional)
```

5 定义年轻专业人士

```
young_professionals <- filter(professionals, Age < 35)
```

6 描述性统计分析

```
summary_stats <- young_professionals %>%
  summarise(
    Average_Investments = mean(Value.of.Investments , na.rm = TRUE),
    Average_Transactions = mean(Number.of.Transactions, na.rm = TRUE),
    Household_Income = mean(Household.Income, na.rm = TRUE)
  )
```

7 相关性分析

```
correlation <- cor(young_professionals$Value.of.Investments, young_professionals$Number.of.Transac
```

8 回归分析

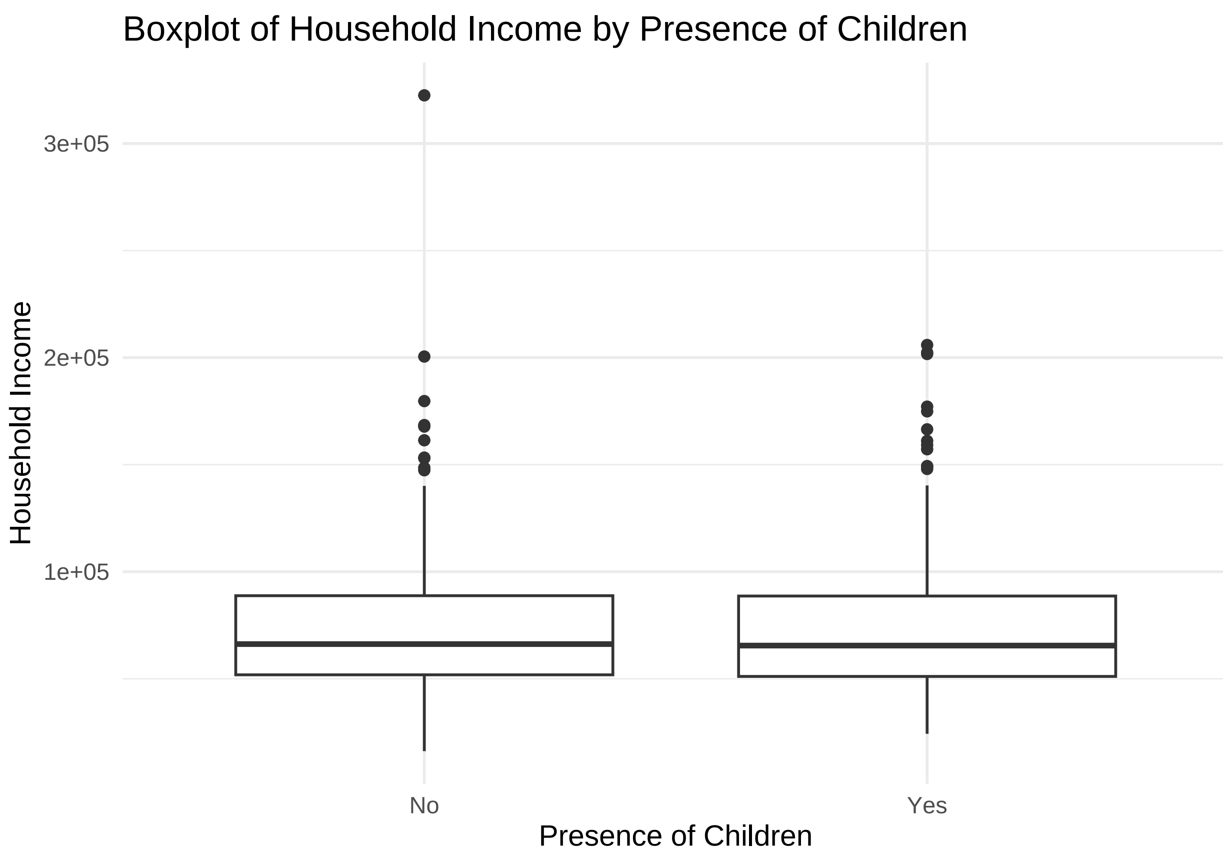
```
model <- lm(Number.of.Transactions ~ Value.of.Investments + Household.Income, data = young_profess
```

9 显示结果

```r
summary(model)
```

```
#>
#> Call:
#> lm(formula = Number.of.Transactions ~ Value.of.Investments +
#>     Household.Income, data = young_professionals)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -6.236 -1.964 -0.408  1.230 14.057
#>
#> Coefficients:
#>                      Estimate Std. Error t value Pr(>|t|)
#> (Intercept)          5.218e+00  4.650e-01  11.221   <2e-16 ***
#> Value.of.Investments 1.521e-05  9.880e-06   1.539    0.125
#> Household.Income     3.938e-06  4.452e-06   0.885    0.377
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 2.991 on 357 degrees of freedom
#> Multiple R-squared:  0.008684,   Adjusted R-squared:  0.00313
#> F-statistic: 1.564 on 2 and 357 DF,  p-value: 0.2108
```

```r
library(ggplot2)
professional$Have.Children <- as.factor(professional$Have.Children)
ggplot(professional, aes(x = Have.Children, y = Household.Income)) +
  geom_boxplot() +
  labs(title = "Boxplot of Household Income by Presence of Children",
       x = "Presence of Children",
       y = "Household Income") +
  theme_minimal()
```



10 Question #5: Quality Associate, Inc. (Attached Data: Quality)

```
quality_data <- read.csv("Quality.csv")
alpha <- 0.01
sample_means <- apply(quality_data, 1, mean)
sample_sds <- apply(quality_data, 1, sd)
sigma <- 0.21
n <- 30
t_tests <- sapply(1:nrow(quality_data), 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
```

```
#>         [,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
```

```
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
```

```
mean(sample_sds)
```

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

```
upper_limit <- 12 + 3 * (sigma / sqrt(n))
lower_limit <- 12 - 3 * (sigma / sqrt(n))
c(upper_limit, lower_limit)
```

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

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。

11 Question #6

```
data <- read.csv("Occupancy.csv")
data$Mar.07 <- ifelse(data$Mar.07=="Yes",1,0)
data$Mar.08 <- ifelse(data$Mar.08=="Yes",1,0)
n <- nrow(data)
```

12 计算 2007 年 3 月第一周出租单位的比例

```
prop_2007 <- mean(data$Mar.07)
```

13 计算 2008 年 3 月第一周出租单位的比例

```
prop_2008 <- mean(data$Mar.08)*4/3
```

14 打印结果

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

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

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

```
#> Proportion of units rented in Mar.08: 0.4666667
```

15 计算比例差异的标准误差

```
se_diff <- sqrt((prop_2007 * (1 - prop_2007) / n) + (prop_2008 * (1 - prop_2008) / n))
```

16 计算 95% 置信区间

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

17 打印结果

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

```
#> 95% Confidence Interval for the difference in proportions: 0.02100854 0.2123248
```

#3. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier?

```
if (ci_diff[1] > 0) {
  cat(" 表明 08 年 3 月租金会同比上升.\n")
} else {
  cat(" 没有明显证据证明 08 年 3 月租金会同比上升.\n")
}
```

```
#> 表明08年3月租金会同比上升.
```

18 Question #7

```
train <- read.csv("Training.csv")
summary_current <- summary(train$Current)
summary_proposed <- summary(train$Proposed)
```

19 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:
```

```
print(summary_current)
```

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

```
cat(" 提议方法的描述性统计:\n")
```

```
#> 提议方法的描述性统计:
```

```
print(summary_proposed)
```

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

#2.Comment on any difference between the population means for the two methods. Discuss your findings.

```
t_test_result <- t.test(train$Current, train$Proposed, var.equal = TRUE)
print(t_test_result)
```

```
#>
#>  Two Sample t-test
#>
#> data:  train$Current and train$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

```
sd_current <- sd(train$Current)
var_current <- var(train$Current)
sd_proposed <- sd(train$Proposed)
var_proposed <- var(train$Proposed)
var_test_result <- var.test(train$Current, train$Proposed)
print(var_test_result)
```

```
#>
#>  F test to compare two variances
#>
#> data:  train$Current and train$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
```

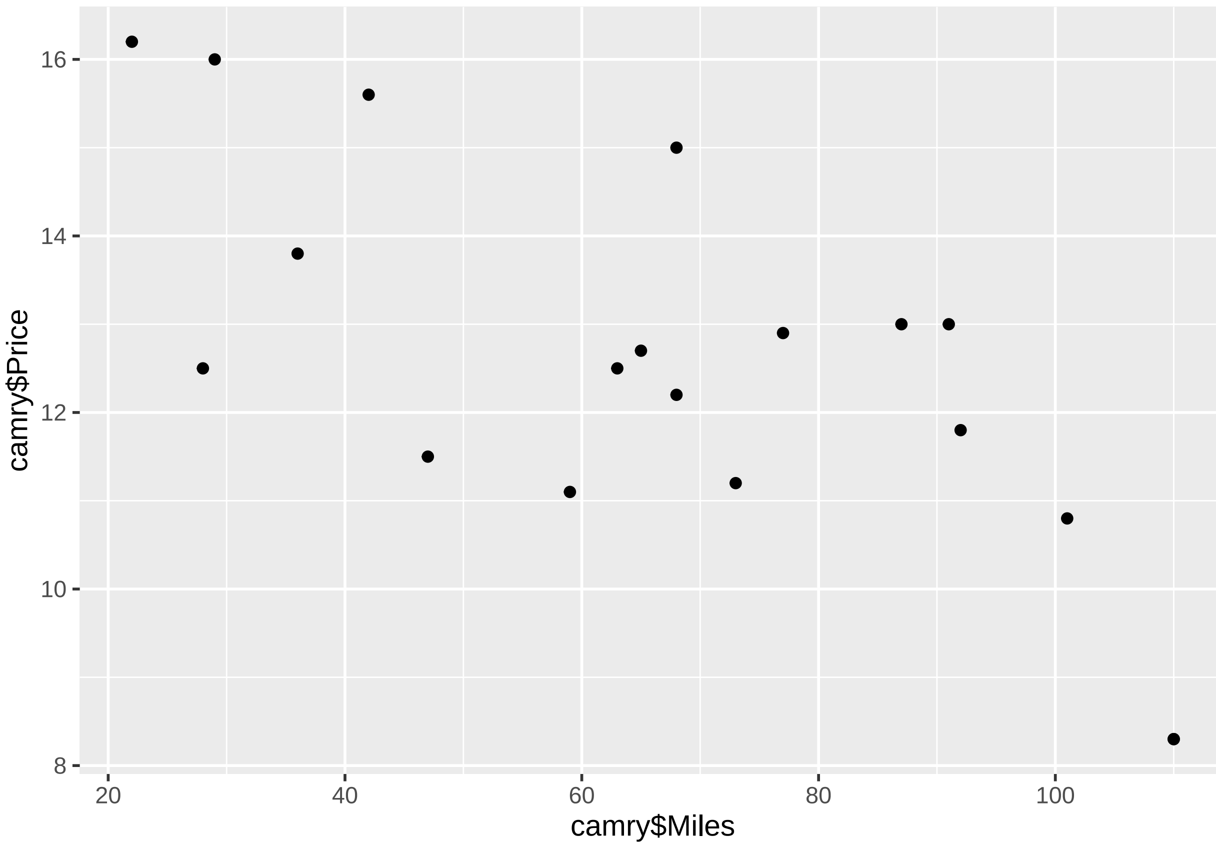
#4. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain 结论：两种方法的均值没有显著差异，但方差存在显著差异，表明提议方法在训练时间上更加一致。#5.can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future? 结论：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。

20 Question #8

```
camry <- read.csv("Camry.csv")
```

#a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.

```
ggplot(camry, aes(x = camry$Miles, y = camry$Price)) +
  geom_point()
```



#b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables? # 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. Develop the estimated regression equation that could be used to predict the price ($1000s)given the miles (1000s).

```
model <- lm(camry$Price ~ camry$Miles, data=camry)
summary(model)
```

```
#>
#> Call:
#> lm(formula = camry$Price ~ camry$Miles, 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 -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. Test for a significant relationship at the .05 level of significance. # 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 的显著性水平下，Miles 对 Price 有显著的影响。#e. Did the estimated regression equation provide a good fit? Explain. #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% 的价格变化。

#f. Provide an interpretation for the slope of the estimated regression equation. # 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. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller

```
coefficients <- coef(model)
intercept <- coefficients[1]
slope <- coefficients[2]
```

21 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"
```

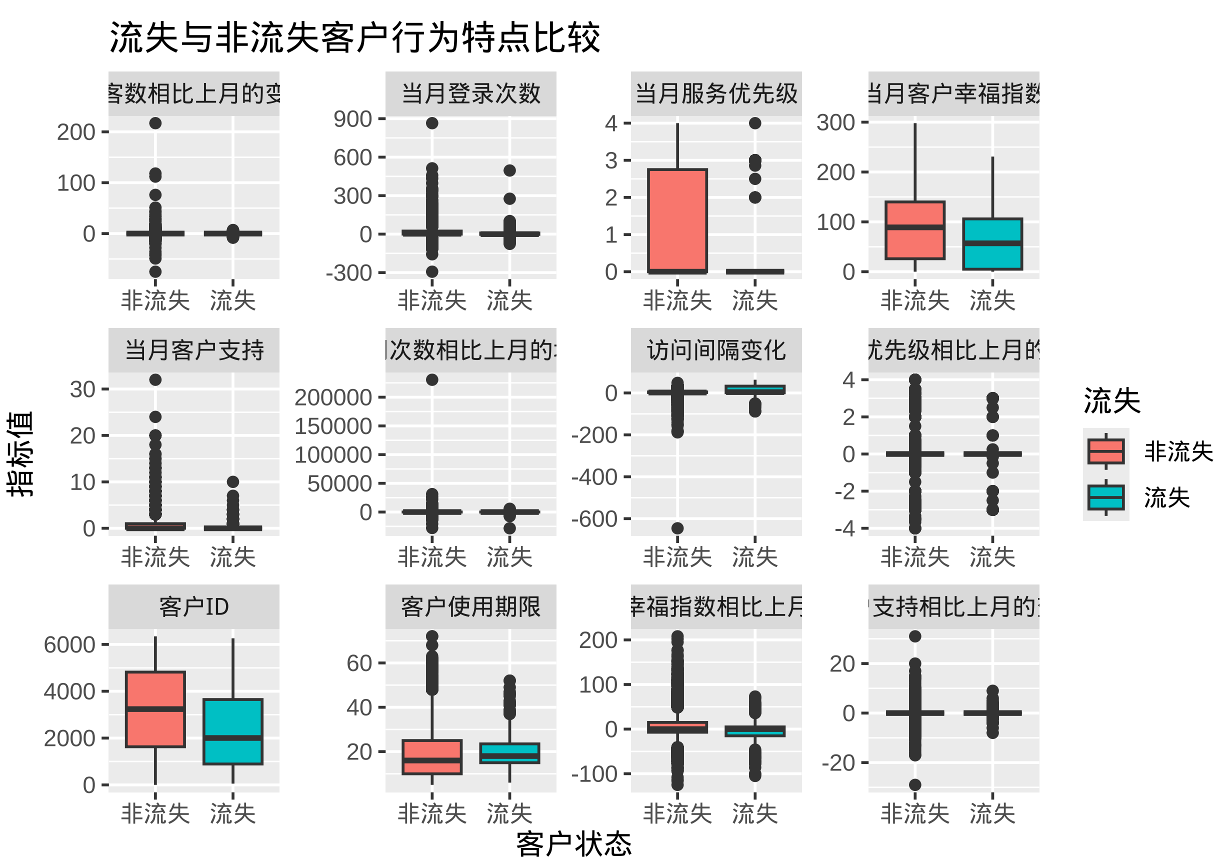
22 Question #9

```
library(readxl)
data <- read_excel("WE.xlsx")
```

#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()
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
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

```r
data_non_churn <- data[data$流失 == 0, ]
predictions <- predict(model, newdata = data_non_churn, type = "response")
data_non_churn$predictions <- predictions
data_non_churn_sorted <- data_non_churn[order(-data_non_churn$predictions), ]
top100_users <- head(data_non_churn_sorted$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
```