zengfeixue homework2

zengfeixue

2024-11-03

```r
# 第一题
library(readxl)
BigBangTheory<- read.csv("C:/Users/pc/Desktop/BigBangTheory.csv")
#a.Compute the minimum and the maximum number of viewers.
max_viewer<-max(BigBangTheory$Viewers)
min_viewer<-min(BigBangTheory$Viewers)
cat('最大值:', max_viewer, '\n')
```

```
## 最大值: 16.5
```

```r
cat('最小值:', min_viewer, '\n')
```

```
## 最小值: 13.3
```

```r
#b.Compute the mean, median, and mode.
mean_viewer <- mean(BigBangTheory$Viewers)
median_viewer <- median(BigBangTheory$Viewers.)
mode_viewer <- names(which.max(table(BigBangTheory$Viewers)))
cat('平均值:', mean_viewer, '\n')
```

```
## 平均值: 15.04286
```

```r
cat('中位数:', median_viewer, '\n')
```

```
## 中位数:
```

```
cat('众数:', mode_viewer, '\n')
```

```
## 众数: 13.6
```

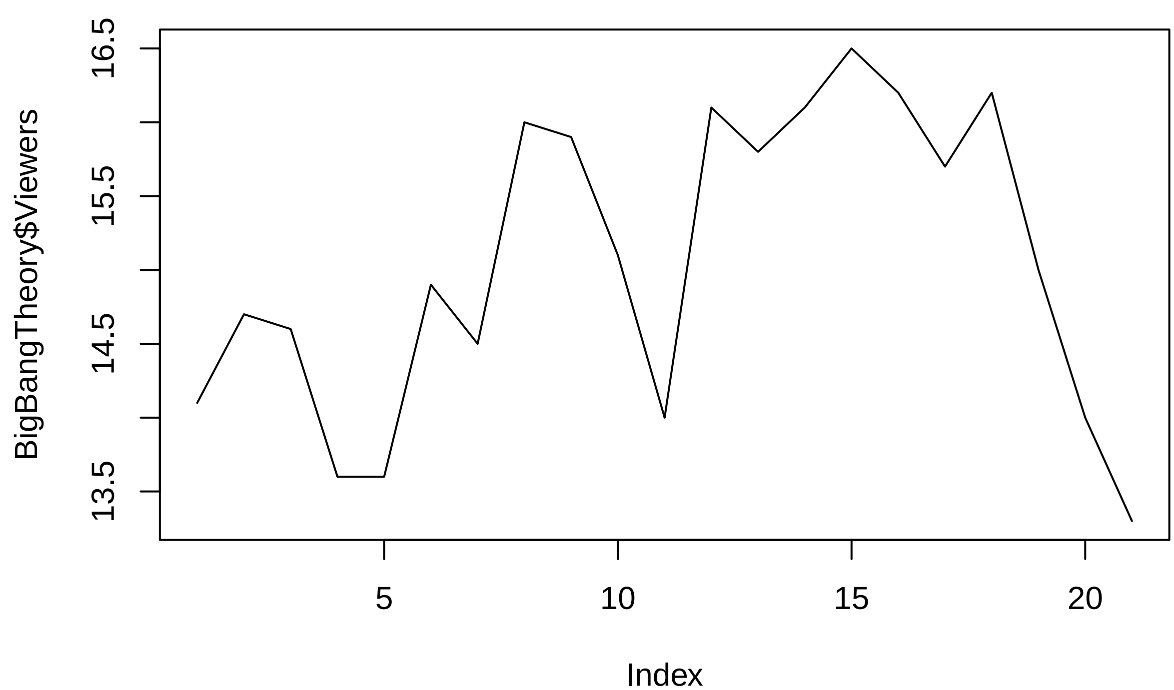
```
#c.Compute the first and third quartiles.
q1_viewers <- quantile(BigBangTheory$Viewers, 0.25)
q3_viewers <- quantile(BigBangTheory$Viewers, 0.75)
cat('q1:', q1_viewers, '\n')
```

```
## q1: 14.1
```

```
cat('q3:', q3_viewers, '\n')
```

```
## q3: 16
```

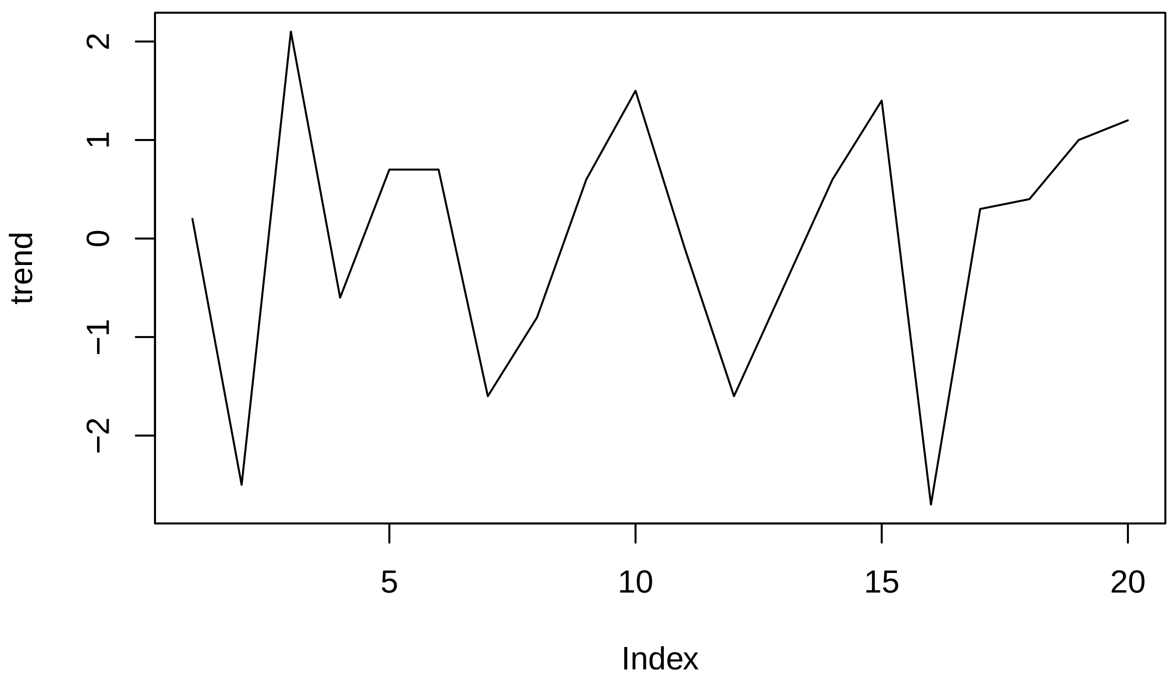
```
#d. has viewership grown or declined over the 2011-2012 season? Discuss.
plot(BigBangTheory$Viewers, type = "l")
```



```
df <- BigBangTheory[order(BigBangTheory$Air.Date), ]
trend <- diff(df$Viewers)
trend
```

```
##  [1]  0.2 -2.5  2.1 -0.6  0.7  0.7 -1.6 -0.8  0.6  1.5 -0.1 -1.6 -0.5  0.6  1.4
## [16] -2.7  0.3  0.4  1.0  1.2
```

```
plot(trend, type = "l")
```




```
# 结论：收视率呈波动状态，12 年之后下降趋势较明显。
```

```
# 第二题
nba<- read.csv("C:/Users/pc/Desktop/NBAPlayerPts.csv")
#a. Show the frequency distribution.
breaks <- seq(10, 30, by = 2)# 定义分组区间
grouped <- cut(nba$PPG, breaks = breaks, include.lowest = TRUE)
freq_table <- table(grouped)# 进行分组并计算频率
print(freq_table)
```

```
## grouped
## [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
```

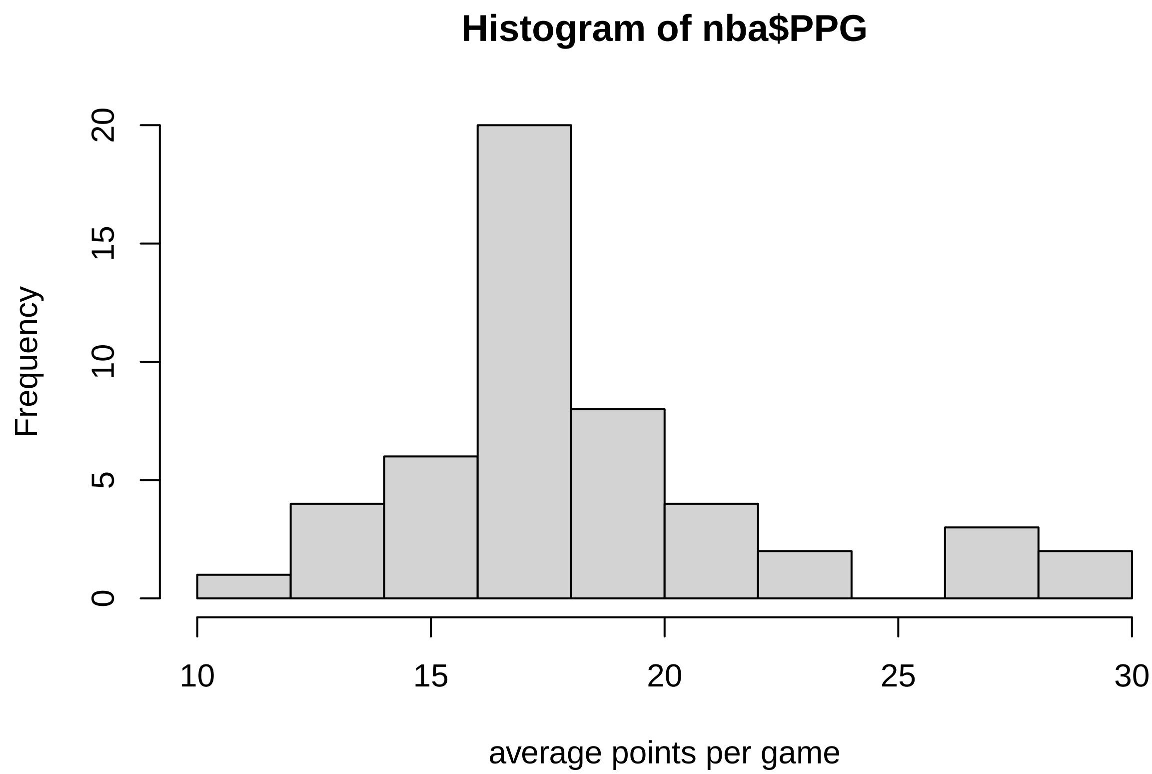
```r
#b.Show the relative frequency distribution.
rel_freq_table <- prop.table(freq_table)
print(rel_freq_table)
```

```
## grouped
## [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
```

```r
#c. Show the cumulative percent frequency distribution.
cum_freq_table <- cumsum(freq_table)
cum_percent_freq_table <- (cum_freq_table / sum(freq_table)) * 100
print(cum_percent_freq_table)
```

```
## [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
##       2      10      22      62      78      86      90      90      96     100
```

```r
#d. Develop a histogram for the average number of points scored per game.
hist(nba$PPG, breaks = 10, xlab = 'average points per game')
```




```
#e. Do the data appear to be skewed? Explain.
mean_ppg <- mean(nba$PPG)
median_ppg <- median(nba$PPG)
if (mean_ppg > median_ppg) {
  skewness <- '右偏'
} else {
  skewness <- '左偏'
}
cat('这组数据', skewness, '\n')
```

```
## 这组数据 右偏
```

```
#f. What percentage of the players averaged at least 20 points per game?
points_20<-nba$PPG >= 20
sum_20 <- sum(points_20)
total_players <- length(nba$PPG)
percentage <- (sum_20 / total_players)
cat( '平均得分至少为 20 分的球员的比例为',percentage,'\n')
```

```
## 平均得分至少为20分的球员的比例为 0.22
```

```r
# 第三题
# 标准误差 =20, 总体标准差 =500, 算样本。
#a. How large was the sample used in this survey?
se <- 20
sigma <- 500
n <- (sigma / se)^2
n
```

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

```r
#b. What is the probability that the point estimate was within ±25 of the population mean?
z1 <- 25 / se
z2 <- -25 / se
probability <- pnorm(z1) - pnorm(z2)
probability
```

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

```r
# 第四题
pro<- read.csv("C:/Users/pc/Desktop/Professional.csv")
#a. Develop appropriate descriptive statistics to summarize the data. 总结数据
str(pro)
```

```
## 'data.frame':    410 obs. of  14 variables:
##  $ Age                        : int  38 30 41 28 31 32 32 26 26 34 ...
##  $ Gender                     : chr  "Female" "Male" "Female" "Female" ...
##  $ Real.Estate.Purchases.     : chr  "No" "No" "No" "Yes" ...
##  $ Value.of.Investments....: int  12200 12400 26800 19600 15100 39700 21900 41900 16100 18400 .
##  $ Number.of.Transactions     : int  4 4 5 6 5 3 2 2 4 11 ...
##  $ Broadband.Access.          : chr  "Yes" "Yes" "Yes" "No" ...
##  $ Household.Income....       : int  75200 70300 48200 95300 73300 123400 73900 54300 93100 60100
##  $ Have.Children.             : chr  "Yes" "Yes" "No" "No" ...
##  $ X                          : logi  NA NA NA NA NA NA ...
##  $ X.1                        : chr  "" "" "" "" ...
##  $ X.2                        : logi  NA NA NA NA NA NA ...
##  $ X.3                        : logi  NA NA NA NA NA NA ...
##  $ X.4                        : logi  NA NA NA NA NA NA ...
##  $ X.5                        : logi  NA NA NA NA NA NA ...
```

```r
summary(pro)
```

```
##       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.         X               X.1
##  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
##    X.2            X.3            X.4            X.5
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:410       NA's:410       NA's:410       NA's:410
##
##
##
##
```

```r
#b.Develop 95% confidence intervals for the mean age and household income of subscribers. 订阅者平均
age_conf <- t.test(pro$Age, conf.level = 0.95)# 计算平均年龄的 95% 置信区间
cat(" 平均年龄的 95% 置信区间为: ", age_conf$conf.int[1], " 至", age_conf$conf.int[2], "\n")
```

```
## 平均年龄的95%置信区间为:  29.72153 至 30.50286
```

```r
income_conf <- t.test(pro$Household.Income...., conf.level = 0.95)# 计算家庭收入的 95% 置信区间
cat(" 家庭收入的 95% 置信区间为: ", income_conf$conf.int[1], " 至", income_conf$conf.int[2], "\n")
```

```
## 家庭收入的95%置信区间为:  71079.26 至 77839.77
```

```r
# c.Develop 95% confidence intervals for the proportion of subscribers who have broadband access a
# 计算拥有宽带接入的比例的 95% 置信区间
broadband_count <- sum(pro$Broadband.Access.== "Yes")
total_count <- nrow(pro)
broadband_conf <- prop.test(broadband_count, total_count, conf.level = 0.95)
cat(" 拥有宽带接入的比例的 95% 置信区间为: ", broadband_conf$conf.int[1], " 至", broadband_conf$conf.
```

```
## 拥有宽带接入的比例的95%置信区间为:  0.5753252 至 0.6710862
```

```r
# 计算拥有孩子的比例的 95% 置信区间
children_count <- sum(pro$Have.Children. == "Yes")
children_conf <- prop.test(children_count, total_count, conf.level = 0.95)
cat(" 拥有孩子的比例的 95% 置信区间为: ", children_conf$conf.int[1], " 至", children_conf$conf.int[2
```

```
## 拥有孩子的比例的95%置信区间为:  0.4845521 至 0.5830908
```

```r
#d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclus
broadband_access_proportion <- mean(pro$Broadband.Access. == "Yes")
cat(" 拥有宽带接入的比例为: ", broadband_access_proportion, "\n")
```

```
## 拥有宽带接入的比例为:  0.6243902
```

```r
# 为 online brokers 开发 95% 置信区间
ci_broadband <- prop.test(sum(pro$Broadband.Access. == "Yes"), nrow(pro), conf.level = 0.95)
print(ci_broadband)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(pro$Broadband.Access. == "Yes") out of nrow(pro), null probability 0.5
## X-squared = 24.88, df = 1, p-value = 6.1e-07
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.5753252 0.6710862
## sample estimates:
##         p
## 0.6243902
```

```
if (ci_broadband$conf.int[1] > 0.5) {
  cat(" 是一个好广告，因为大多数订阅者都有宽带接入。\n")
} else {
  cat(" 不是好广告，因为拥有宽带接入的订阅者比例不高。\n")
}
```

```
## 是一个好广告，因为大多数订阅者都有宽带接入。
```

```
#e. Would this magazine be a good place to advertise for companies selling educational software an
children_proportion <- mean(pro$Have.Children. == "Yes")
if (children_proportion > 0.5) {
  cat(" 适合在这里打广告，因为多数订阅者有孩子。\n")
} else {
  cat(" 不适合打广告，因为拥有孩子的订阅者比例不高。\n")
}
```

```
## 适合在这里打广告，因为多数订阅者有孩子。
```

```
#f. Comment on the types of articles you believe would be of interest to readers of Young Professi
# 读者可能对以下内容感兴趣:
#1. 个人理财和投资建议，因为许多读者都有投资。2. 科技和互联网相关的最新动态，因为大多数读者都有宽带接入。
```

```
# 第五题
quality<- read.csv("C:/Users/pc/Desktop/Quality.csv")
# 总体均值 =12, 总体标准差 = 0.21, n = 30
#a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what
mu <- 12
sigma <- 0.21
n <- 30
sample_means <- apply(quality, 2, mean)# 计算每个样本的均值
z_scores <- (sample_means - mu) / (sigma / sqrt(n))# 计算 z 检验统计量
critical_value <- qnorm(0.995)# 确定临界值
hypothesis_results <- ifelse(abs(z_scores) > critical_value, "Reject H0", "Fail to reject H0")# 执
data.frame(Sample = names(sample_means), Mean = sample_means, Z_Score = z_scores, Decision = hypot
```

```
##            Sample     Mean    Z_Score          Decision
## Sample.1 Sample.1 11.95867 -1.0780571 Fail to reject H0
## Sample.2 Sample.2 12.02867  0.7476848 Fail to reject H0
## Sample.3 Sample.3 11.88900 -2.8951049         Reject H0
## Sample.4 Sample.4 12.08133  2.1213382 Fail to reject H0
```

```
#b. compute the standard deviation for each of the four samples. does the assumption of .21 for th
sample_sds <- apply(quality, 2, sd)
sample_sds
```

```
##  Sample.1  Sample.2  Sample.3  Sample.4
## 0.2203560 0.2203560 0.2071706 0.2061090
```

```
assumed_sigma <- 0.21
comparison <- data.frame(Sample = names(sample_sds), Sample_SD = sample_sds, Assumed_SD = assumed_
comparison
```

```
##            Sample Sample_SD Assumed_SD
## Sample.1 Sample.1 0.2203560       0.21
## Sample.2 Sample.2 0.2203560       0.21
## Sample.3 Sample.3 0.2071706       0.21
## Sample.4 Sample.4 0.2061090       0.21
```

```
#c. upper and lower control limits
alpha <- 0.01
z_alpha <- qnorm(1 - alpha/2)
upper_limit <- mu + z_alpha * (sigma / sqrt(n))
lower_limit <- mu - z_alpha * (sigma / sqrt(n))
data.frame(Lower_Control_Limit = lower_limit, Upper_Control_Limit = upper_limit)
```

```
##   Lower_Control_Limit Upper_Control_Limit
## 1            11.90124            12.09876
```

```
# 只要新样本均值在 11.53 和 12.47 之间，过程就被认为是正常运行的。如果样本均值超过 12.47 或低于 11.53，

#d. discuss the implications of changing the level of significance to a larger value. what mistake
# 定义两个不同的显著性水平
alpha_small <- 0.01  # 较小的显著性水平，例如 0.01
alpha_large <- 0.10 # 较大的显著性水平，例如 0.10
```

```r
# 计算两个显著性水平下的 z 值
z_alpha_small <- qnorm(1 - alpha_small/2)
z_alpha_large <- qnorm(1 - alpha_large/2)
# 计算两个显著性水平下的控制限
upper_limit_small <- mu + z_alpha_small * (sigma / sqrt(n))
lower_limit_small <- mu - z_alpha_small * (sigma / sqrt(n))

upper_limit_large <- mu + z_alpha_large * (sigma / sqrt(n))
lower_limit_large <- mu - z_alpha_large * (sigma / sqrt(n))

# 打印控制限
cat(" 较小显著性水平下的控制限：\n")
```

```
## 较小显著性水平下的控制限：
```

```r
cat(" 下控制限：", lower_limit_small, "\n")
```

```
## 下控制限： 11.90124
```

```r
cat(" 上控制限：", upper_limit_small, "\n\n")
```

```
## 上控制限： 12.09876
```

```r
cat(" 较大显著性水平下的控制限：\n")
```

```
## 较大显著性水平下的控制限：
```

```r
cat(" 下控制限：", lower_limit_large, "\n")
```

```
## 下控制限： 11.93694
```

```r
cat(" 上控制限：", upper_limit_large, "\n")
```

```
## 上控制限： 12.06306
```

```r
# 结论：增加显著性水平可能会在错误地拒绝零假设，这可能导致不必要的纠正措施，并使控制限变宽，从而降低质量控
```

```r
# 第六题
occ<- read.csv("C:/Users/pc/Desktop/Occupancy.csv")
#a. Estimate the proportion of units rented during the first week of March 2007 and the first week
library(dplyr)
```

```
##
## 载入程序包：'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
rented_2007 <- sum(occ$`2007-03-01 00:00:00` == "Yes")
rented_2008 <- sum(occ$`2008-03-01 00:00:00` == "Yes")
total_units <- nrow(occ)
proportion_2007 <- rented_2007 / total_units
proportion_2008 <- rented_2008 / total_units
cat("2007 年 3 月第一周出租单元的比例为：", proportion_2007, "\n")
```

```
## 2007年3月第一周出租单元的比例为： 0
```

```r
cat("2008 年 3 月第一周出租单元的比例为：", proportion_2008, "\n")
```

```
## 2008年3月第一周出租单元的比例为： 0
```

```r
#b. Provide a 95% confidence interval for the difference in proportions.
n_2007 <- sum(occ$Rented[occ$Month == "March 2007"])
n_2008 <- sum(occ$Rented[occ$Month == "March 2008"])
p_diff <- proportion_2008 - proportion_2007
se_diff <- sqrt((proportion_2007 * (1 - proportion_2007) / n_2007) + (proportion_2008 * (1 - propo
ci_lower <- p_diff - 1.96 * se_diff
ci_upper <- p_diff + 1.96 * se_diff
cat("95% 置信区间为：", "(", ci_lower, ", ", ci_upper, ")\n")
```

```
## 95%置信区间为： ( NaN ,  NaN )
```

```r
#c. On the basis of your findings, does it appear March rental rates for 2008 will be upfrom those
# 2008 年出租率很可能上涨。根据 b 95% 比例置信区间为不包含 0，说明两者比例是有差异的，则 2008 年 3 月出租
```

```r
# 第七题
train<- read.csv("C:/Users/pc/Desktop/Training.csv")
#a. use appropriate descriptive statistics to summarize the training time data for each method. wh
library(kableExtra)
```

```
##
## 载入程序包：'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
skimr::skim(train) %>%
  kable() %>%
  kable_styling()
```

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25
numeric	Current	0	1	75.06557	3.944907	65	72
numeric	Proposed	0	1	75.42623	2.506385	69	74

```r
str(train)
```

```
## 'data.frame':    61 obs. of  2 variables:
##  $ Current : int  76 76 77 74 76 74 74 77 72 78 ...
##  $ Proposed: int  74 75 77 78 74 80 73 73 78 76 ...
```

```r
summary(train)
```

```
##     Current         Proposed
##  Min.   :65.00   Min.   :69.00
##  1st Qu.:72.00   1st Qu.:74.00
##  Median :76.00   Median :76.00
##  Mean   :75.07   Mean   :75.43
##  3rd Qu.:78.00   3rd Qu.:77.00
##  Max.   :84.00   Max.   :82.00
```

```r
# 这两种方法从统计上来看均值、中位数等数据都差异不大，二者没有明显区别

#b. Comment on any difference between the population means for the two methods. Discuss your findi
t.test(train$Current,train$Proposed)
```

```
##
##  Welch Two Sample t-test
##
## data:  train$Current and train$Proposed
## t = -0.60268, df = 101.65, p-value = 0.5481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5476613  0.8263498
## sample estimates:
## mean of x mean of y
##  75.06557  75.42623
```

```r
# p-value = 0.5481, 在 0.05 的显著性水平下，两种方法之间无显著差异。

#c. compute the standard deviation and variance for each training method. conduct a hypothesis tes
var_current <- var(train$Current)
sd_current <- sd(train$Current)
var_proposed <- var(train$Proposed)
sd_proposed <- sd(train$Proposed)
var.test(train$Current,train$Proposed,conf.level = 0.95)
```

```
##
##  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
```

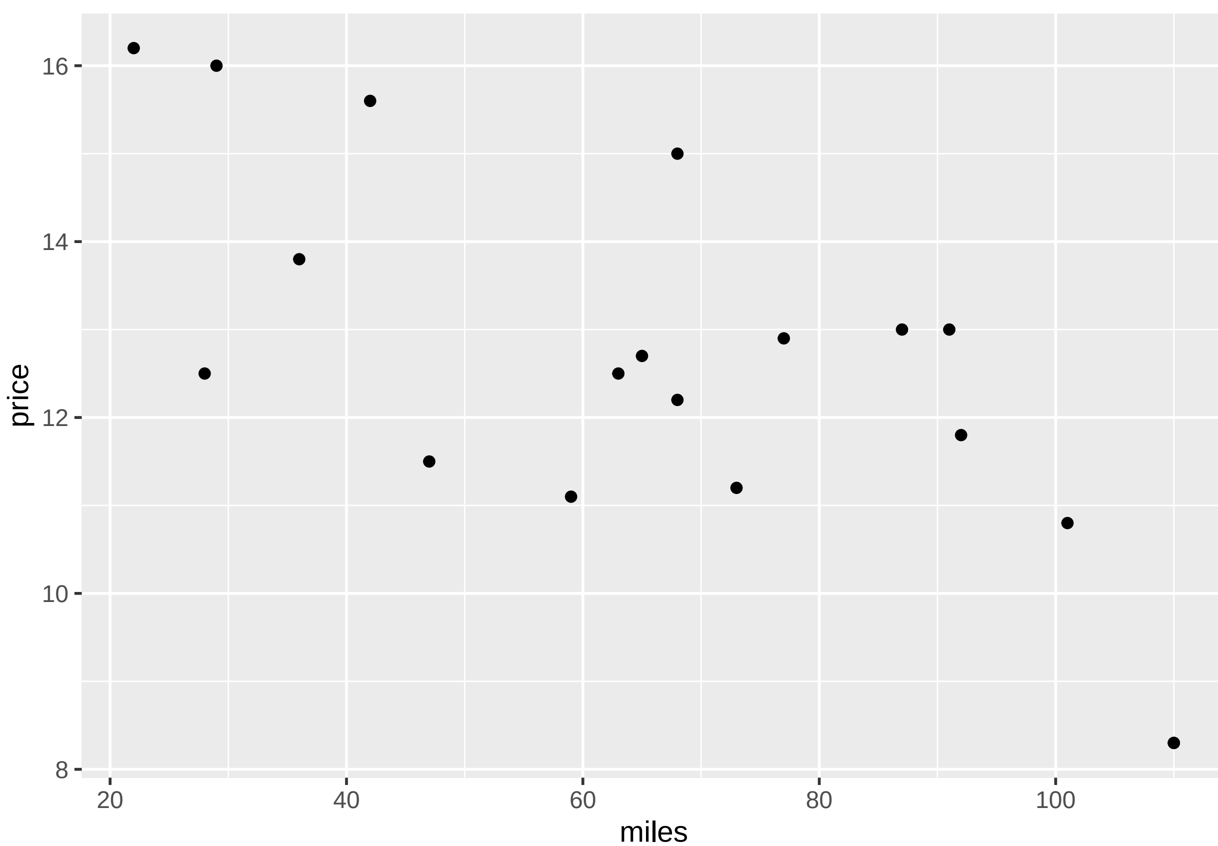
```
#p-value = 0.000578, 在 0.05 显著性水平下，两种方法的标准差或方差具有显著性差异

#d. what conclusion can you reach about any differences between the two methods? what is your reco
# 更推荐提议的方法（Proposed）。这两种方法在平均数上没有显著差异，但提议的方法在方差和标准差上显著性更低。

#e.can you suggest other data or testing that might be desirable before making a final decision on
# 目前的统计方法只是从时间上进行统计，得出了更推荐提议方法的结论。但学习效果是否也一样还有待检验，我认为还
```

```
# 第八题
carmy<- read.csv("C:/Users/pc/Desktop/Camry.csv")%>%
  rename(miles = `Miles..1000s.`,
         price = `Price...1000s.`)

#a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the ver
library(ggplot2)
ggplot(carmy,aes(miles,price))+
  geom_point()
```



```r
#b. what does the scatter diagram developed in part (a) indicate about the relationship between th
# 根据散点图，凯美瑞汽车的里程和价格大致呈负相关，里程越长，价格越低。

#c. Develop the estimated regression equation that could be used to predict the price ($1000s) giv
model_carmy<-lm(price ~ miles, data = carmy)
summary(model_carmy)
```

```
##
## Call:
## lm(formula = price ~ miles, data = carmy)
##
## 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 ***
## 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
```

```r
cat(" 价格 = ", coef(model_carmy)[1], " + ", coef(model_carmy)[2], " 里程\n")
```

```
## 价格 =  16.46976  +  -0.05877393 里程
```

```r
#coef(model) 用于提取模型的系数，即截距和斜率

#d. Test for a significant relationship at the .05 level of significance.
summary(model_carmy)
```

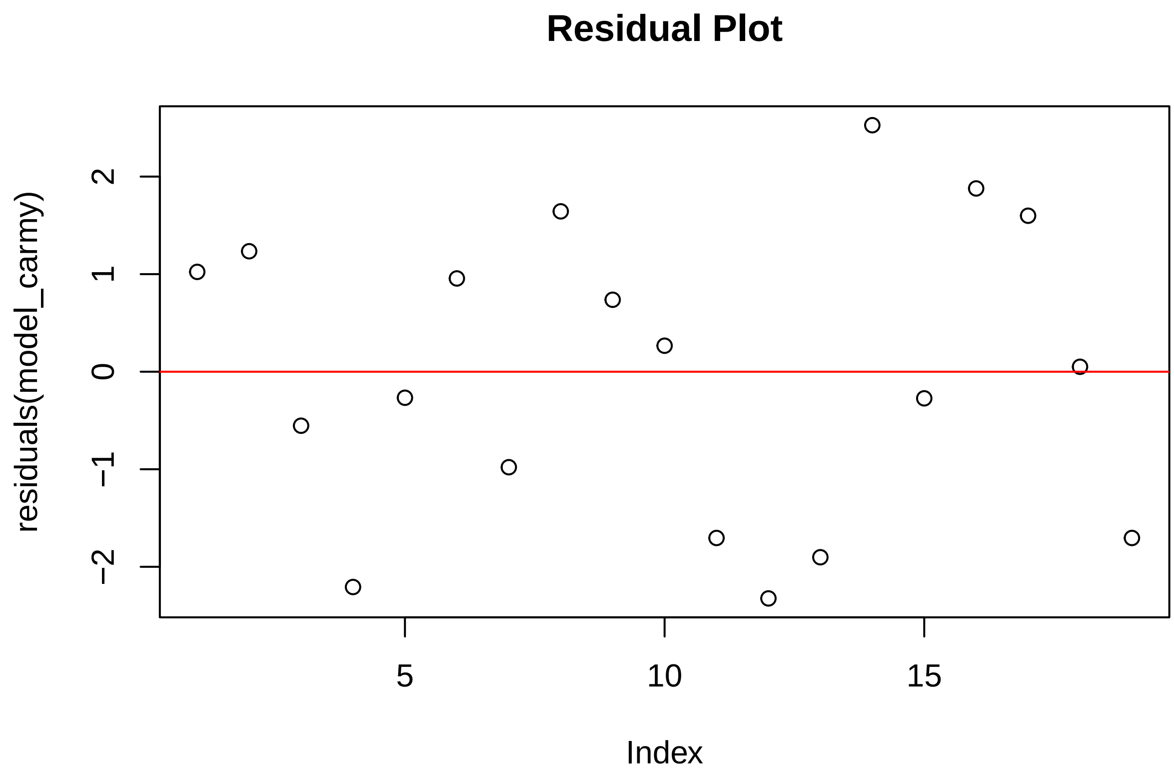
```
##
## Call:
## lm(formula = price ~ miles, data = carmy)
```

```
##
## 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 ***
## 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
```

```
#p-value: 0.0003475<0.05, 是显著的

#e. Did the estimated regression equation provide a good fit? Explain.
#R-squared: 0.5115, 有 51.15%, R 平方值越接近 1, 表示模型拟合得越好

plot(residuals(model_carmy), main = "Residual Plot")
abline(h = 0, col = "red")# 绘制残差图
```




```
# 残差图随机分布在 0 周围，没有明显的模式（如曲线或系统性偏差）
# 根据 R 平方值和残差图，基本可得出结论：这个回归方程提供了良好的拟合

#f. Provide an interpretation for the slope of the estimated regression equation.
slope <- coef(model_carmy)[1]
cat(" 回归方程的斜率是: ", slope, " (意味着每增加 1000 英里, 价格将会下降", slope, " 美元).\n")
```

```
## 回归方程的斜率是:  16.46976  (意味着每增加 1000英里, 价格将会下降 16.46976  美元).
```

```
#g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven
predicted_price <-16.46976-0.05877393*60
predicted_price
```

```
## [1] 12.94332
```

```
# 预测价格为 12.94 美元，回归模型只有 55% 的准确率，并不能完全代表实际，在真实出价中还要考虑车的其他因素。
```

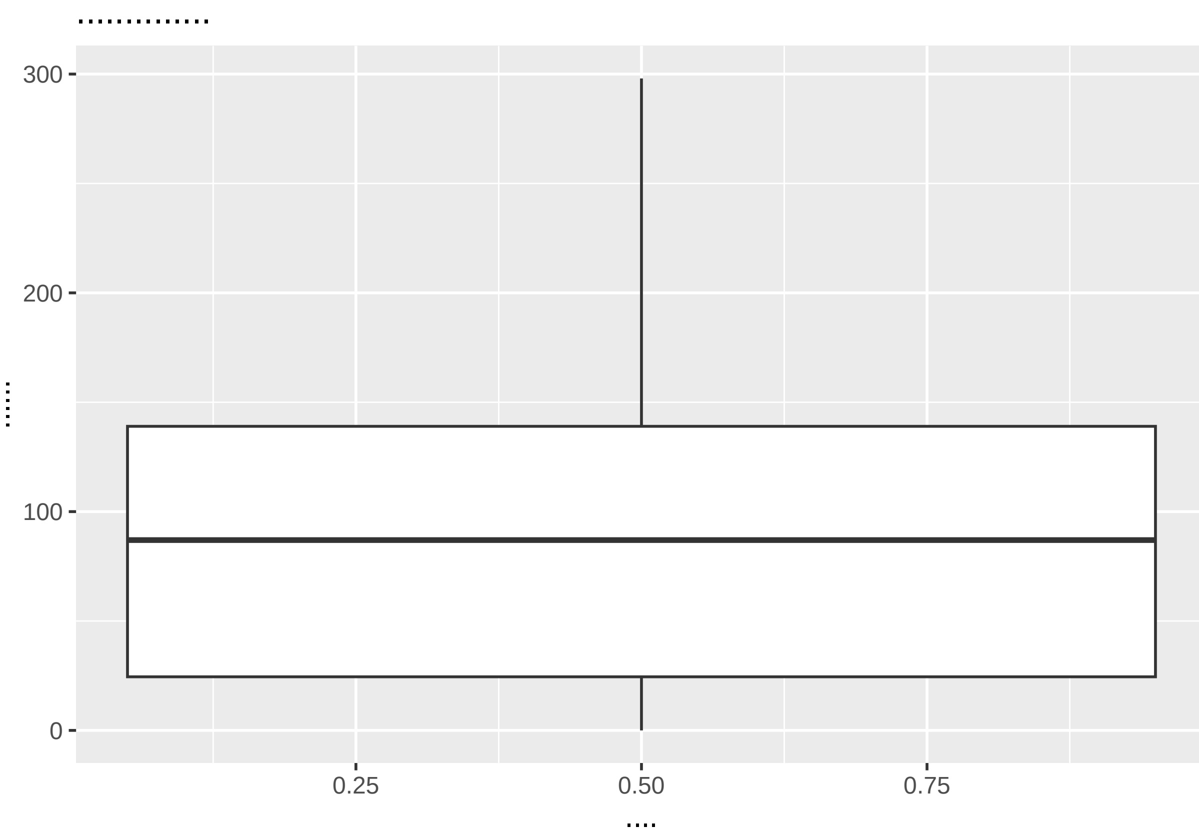
```r
# 第九题
library(readxl)
library(dplyr)
we <- read_excel("C:/Users/pc/Desktop/WE.xlsx") %>%
  rename(
    id = `客户 ID`,
    loose = `流失`,
    happy_index = `当月客户幸福指数`,
    happy_index_var = `客户幸福指数相比上月变化`,
    support = `当月客户支持`,
    support_var = `客户支持相比上月的变化`,
    service = `当月服务优先级`,
    service_var = `服务优先级相比上月的变化`,
    login = `当月登录次数`,
    blog_var = `博客数相比上月的变化`,
    vist_add = `访问次数相比上月的增加`,
    age = `客户使用期限`,
    gap = `访问间隔变化`
  )

#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
glimpse(we)
```

```
## Rows: 6,347
## Columns: 13
## $ id              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ loose           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ happy_index     <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78, 91, 40, 215,~
## $ happy_index_var <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1, 14, 15, 0, ~
## $ support         <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ support_var     <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ service         <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ service_var     <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ login           <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0, 71, 0, 5, 0~
## $ blog_var        <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, 0, 1, 0, 0, 0~
## $ vist_add        <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30, 0, 15, 8658,~
## $ age             <dbl> 72, 72, 60, 68, 62, 63, 62, 51, 61, 61, 58, 61, 62, 62~
## $ gap             <dbl> 33, 33, 33, 2, 33, 2, 2, 8, 9, 16, 2, 33, 2, 33, 2, 33~
```

```
summary_stats <- we %>%
  group_by(loose) %>%
  summarise(across(everything(), list(mean = mean, sd = sd, min = min, max = max)))
library(ggplot2)
ggplot(we, aes(x = loose, y = happy_index)) +
  geom_boxplot() +
  labs(title = " 流失与非流失客户幸福指数比较", x = " 流失状态", y = " 客户幸福指数")# 箱线图：比较流失
```

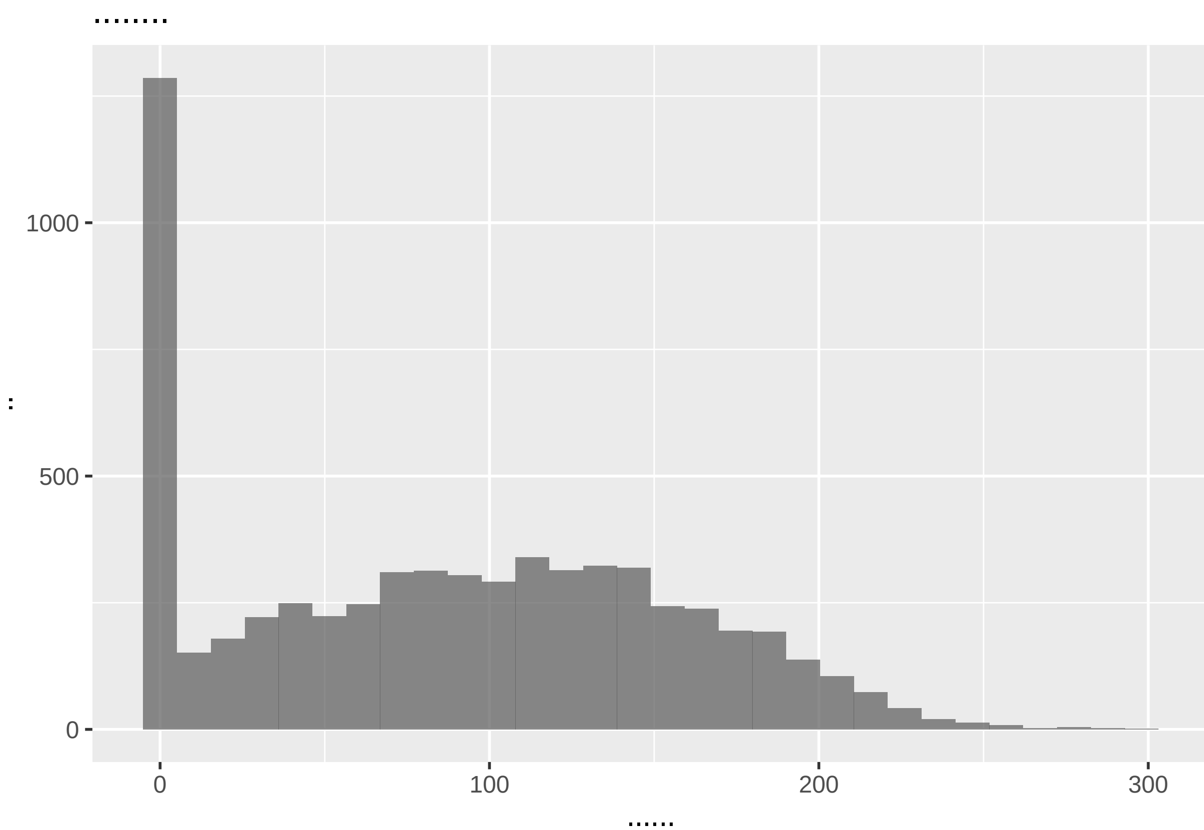
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```




```
# 直方图：查看单个指标的分布
ggplot(we, aes(x =  happy_index, fill = loose)) +
  geom_histogram(bins = 30, alpha = 0.7) +
  labs(title = " 客户幸福指数分布", x = " 客户幸福指数", y = " 频数")
```

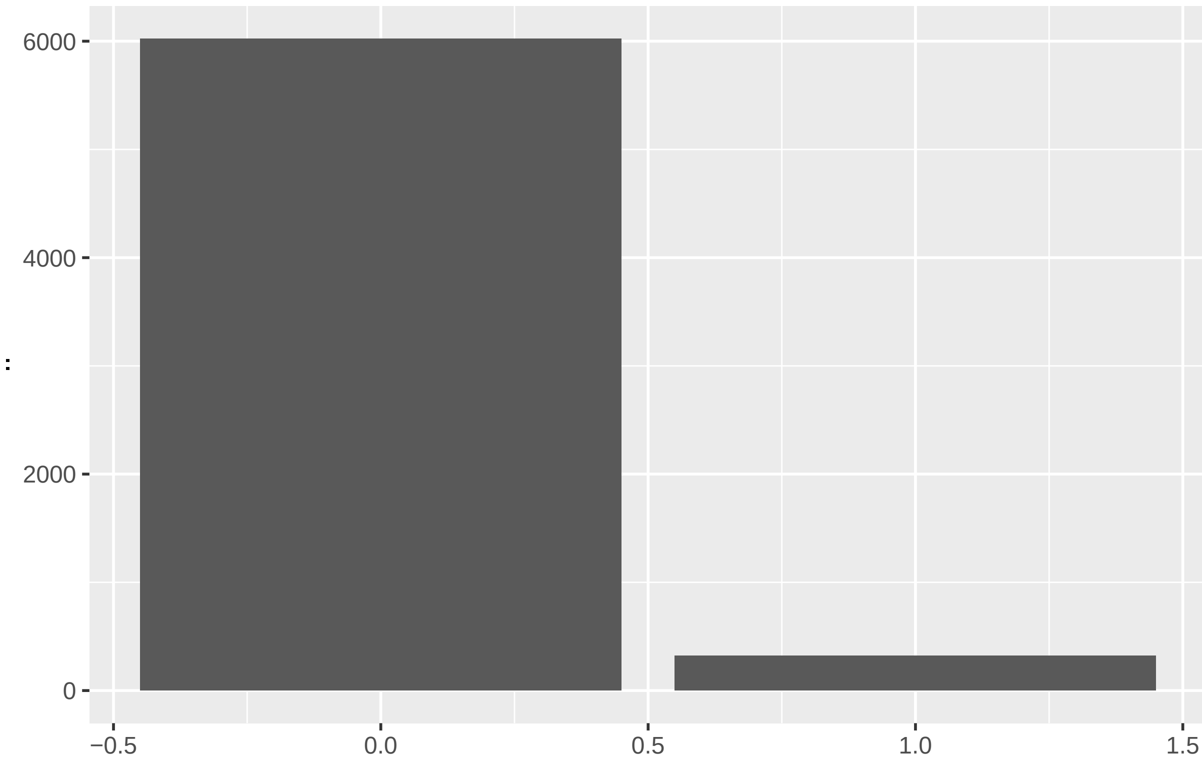
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
```

```
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
# 条形图：比较流失与非流失客户在分类指标上的比例
ggplot(we, aes(x = loose, fill = loose)) +
  geom_bar() +
  labs(title = " 流失与非流失客户比例", x = " 流失状态", y = " 比例")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
# 统计检验：t 检验或卡方检验
# 以客户幸福指数为例进行 t 检验
t_test_result <- t.test( happy_index ~ loose, data = we)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  happy_index by loose
## t = 7.6242, df = 369.36, p-value = 2.097e-13
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  18.79956 31.86737
## sample estimates:
## mean in group 0 mean in group 1
##        88.60591        63.27245
```

```r
# 以服务优先级为例进行卡方检验
chi_squared_result <- chisq.test(table(we$loose, we$service))
```

```
## Warning in chisq.test(table(we$loose, we$service)):
## Chi-squared近似算法有可能不准
```

```r
print(chi_squared_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(we$loose, we$service)
## X-squared = 28.334, df = 26, p-value = 0.3422
```

```r
# 流失与非流失客户在客户幸福指数、客户支持次数、服务优先级、登录和活跃度、访问次数变化以及客户使用期限等指


#b. 通过均值比较的方式验证上述不同是否显著。
non_loose <- we[we$loose == 0, ]
loose <- we[we$loose == 1, ]
指标 <- c("happy_index", "happy_index_var", "support", "support_var", "service", "service_var", "lo
# 进行 t 检验并输出结果
t_test_results <- sapply(指标, function(x) {
  # 确保列是数值型的
  if(is.numeric(non_loose[[x]]) && is.numeric(loose[[x]])) {
    t.test(non_loose[[x]], loose[[x]], var.equal = TRUE) # 假设方差相等
  } else {
    NULL # 如果不是数值型，返回 NULL
  }
})
print(t_test_results)
```

```
##             happy_index                    happy_index_var
## statistic   6.715168                       5.274251
## parameter   6345                           6345
## p.value     2.041576e-11                   1.377109e-07
## conf.int    numeric,2                      numeric,2
## estimate    numeric,2                      numeric,2
```

```
## null.value  0                              0
## stderr      3.772573                       1.757037
## alternative "two.sided"                    "two.sided"
## method      " Two Sample t-test"           " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]" "non_loose[[x]] and loose[[x]]"
##             support                        support_var
## statistic   3.585978                       -0.4346473
## parameter   6345                           6345
## p.value     0.0003383435                   0.6638332
## conf.int    numeric,2                      numeric,2
## estimate    numeric,2                      numeric,2
## null.value  0                              0
## stderr      0.09836997                     0.1068633
## alternative "two.sided"                    "two.sided"
## method      " Two Sample t-test"           " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]" "non_loose[[x]] and loose[[x]]"
##             service                        service_var
## statistic   4.381985                       0.5919949
## parameter   6345                           6345
## p.value     1.194977e-05                   0.5538751
## conf.int    numeric,2                      numeric,2
## estimate    numeric,2                      numeric,2
## null.value  0                              0
## stderr      0.07531249                     0.08340948
## alternative "two.sided"                    "two.sided"
## method      " Two Sample t-test"           " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]" "non_loose[[x]] and loose[[x]]"
##             login                          blog_var
## statistic   3.360347                       1.026795
## parameter   6345                           6345
## p.value     0.000783041                    0.304556
## conf.int    numeric,2                      numeric,2
## estimate    numeric,2                      numeric,2
## null.value  0                              0
## stderr      2.403628                       0.2661835
## alternative "two.sided"                    "two.sided"
## method      " Two Sample t-test"           " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]" "non_loose[[x]] and loose[[x]]"
##             vist_add                       age
```

```
## statistic   1.124055                          -2.407925
## parameter   6345                              6345
## p.value     0.2610323                         0.01607184
## conf.int    numeric,2                         numeric,2
## estimate    numeric,2                         numeric,2
## null.value  0                                 0
## stderr      180.0422                          0.6371528
## alternative "two.sided"                       "two.sided"
## method      " Two Sample t-test"              " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]" "non_loose[[x]] and loose[[x]]"
##             gap
## statistic   -4.856669
## parameter   6345
## p.value     1.22239e-06
## conf.int    numeric,2
## estimate    numeric,2
## null.value  0
## stderr      1.024285
## alternative "two.sided"
## method      " Two Sample t-test"
## data.name   "non_loose[[x]] and loose[[x]]"
```

```r
# 根据 p.value 的值，除客户支持相比上月的变化和服务优先级相比上月的变化外，其他变量均显著



#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
we_model <- glm(loose ~ happy_index + happy_index_var + support + service + login
              + blog_var + vist_add  + age + gap,
             data = we,
             family = binomial)
```

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

```r
summary(we_model)
```

```
##
## Call:
## glm(formula = loose ~ happy_index + happy_index_var + support +
##     service + login + blog_var + vist_add + age + gap, family = binomial,
```

```
##     data = we)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
## happy_index     -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
## happy_index_var -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
## support         -3.522e-02  7.438e-02  -0.474  0.63581
## service         -3.727e-02  7.514e-02  -0.496  0.61985
## login            9.104e-04  1.952e-03   0.466  0.64098
## blog_var        -2.357e-05  2.080e-02  -0.001  0.99910
## vist_add        -1.170e-04  4.069e-05  -2.877  0.00401 **
## age              1.418e-02  5.260e-03   2.696  0.00701 **
## gap              1.700e-02  4.277e-03   3.975 7.03e-05 ***
## ---
## 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: 2465.9
##
## Number of Fisher Scoring iterations: 6
```

```
# 当月客户幸福指数、客户幸福指数相比上月变化、访问间隔变化在 0.001 显著性水平上是显著的。访问次数相比上月

#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
non_loose <- we[we$loose == 0, ]
# 计算流失可能性得分，这里我们使用客户幸福指数的下降作为指标
non_loose$churn_score <- -non_loose$happy_index_var
# 对非流失客户按流失可能性得分进行降序排序
sorted_non_loose <- non_loose[order(non_loose$churn_score, decreasing = TRUE), ]
# 提取流失可能性最大的前 100 名用户 ID 列表
top_100 <- sorted_non_loose[1:100, c("id")]
print(top_100)
```

```
## # A tibble: 100 x 1
##       id
```

```
##    <dbl>
##  1   109
##  2  4191
##  3  1971
##  4  3823
##  5  2481
##  6  2903
##  7  5189
##  8  3577
##  9  1481
## 10  1574
## # i 90 more rows
```