商务统计第二次作业

程迪 2023281051004

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

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the 2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (*the Big Bang theory* website, April 17, 2012).

a. Compute the minimum and the maximum number of viewers.

```
## # A tibble: 1 x 2
##   minimum maximum
##     <dbl>   <dbl>
## 1    13.3    16.5
```

b. Compute the mean, median, and mode.

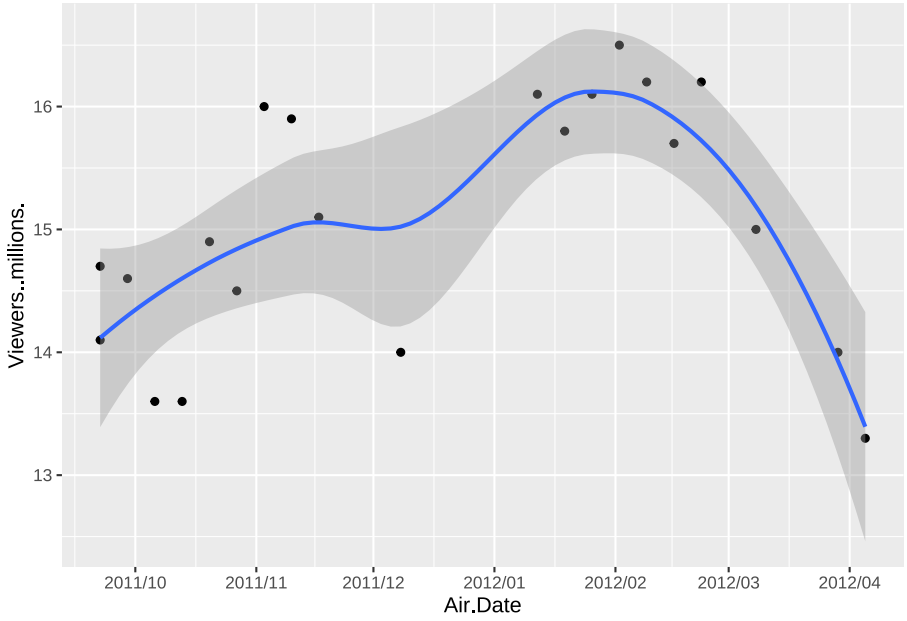
```
## # A tibble: 1 x 3
##    mean median mode
##   <dbl>  <dbl> <chr>
## 1  15.0     15 numeric
```

c. Compute the first and third quartiles.

```
## # A tibble: 1 x 2
##      Q1    Q3
##   <dbl> <dbl>
## 1  14.1    16
```

d. has viewership grown or declined over the 2011–2012 season? Discuss.

答：2011-2012 年年间呈现出现先增长后下降的趋势，于 2012 年 2 月左右达到峰值



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

CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.

```
## # A tibble: 50 x 4
##     Rank Player                   PPG ppg_class
##    <dbl> <chr>                  <dbl>     <int>
##  1     1 LeBron James, MIA       27           9
##  2     2 Kevin Durant, OKC       28.8        10
##  3     3 James Harden, HOU       26.4         9
##  4     4 Kobe Bryant, LAL        27.1         9
##  5     5 Russell Westbrook, OKC  22.9         7
##  6     6 Carmelo Anthony, NY     28.4        10
##  7     7 David Lee, GS           19.2         5
##  8     8 Stephen Curry, GS       21           6
##  9     9 LaMarcus Aldridge, POR  20.8         6
## 10    10 Paul George, IND        17.6         4
## # i 40 more rows
```

a. Show the frequency distribution.

```
##
##  1  2  3  4  5  6  7  9 10
##  1  4  6 20  8  4  2  3  2
```

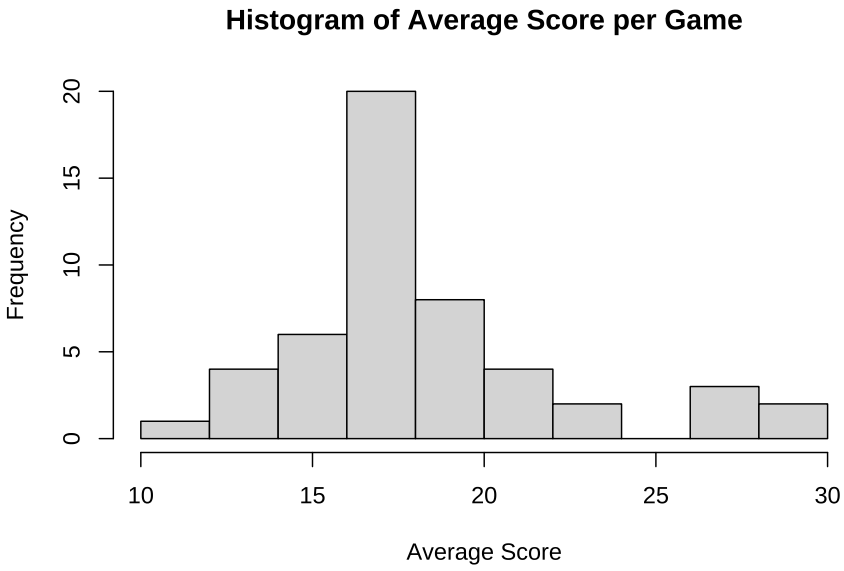
b. Show the relative frequency distribution.

```
##
##    1    2    3    4    5    6    7    9   10
## 0.02 0.08 0.12 0.40 0.16 0.08 0.04 0.06 0.04
```

c. Show the cumulative percent frequency distribution.

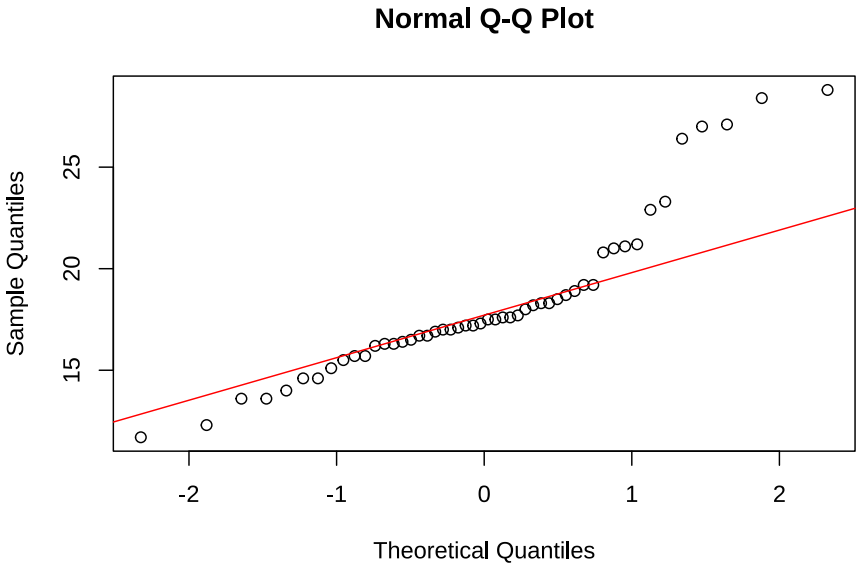
```
## [1] "2.00 %"   "10.00 %"  "22.00 %"  "62.00 %"  "78.00 %"  "86.00 %"  "90.00 %"
## [8] "96.00 %"  "100.00 %"
```

d. Develop a histogram for the average number of points scored per game.



e. Do the data appear to be skewed? Explain.

根据 qq 图，数据存在右偏



f. What percentage of the players averaged at least 20 points per game?

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

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

a. How large was the sample used in this survey?

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

样本量为 625

b. What is the probability that the point estimate was within ±25 of the population mean?

```
## [1] 0.03987761
```

概率为 3.98%

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

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *young Professionals*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

1. What is your age?

2. Are you: Male_____ Female_____

3. Do you plan to make any real estate purchases in the next two years? Yes_____ No_____

4. What is the approximate total value of financial investments, exclusive of your home, owned by you or members of your household?

5. How many stock/bond/mutual fund transactions have you made in the past year?

6. Do you have broadband access to the Internet at home? Yes_____ No_____

7. Please indicate your total household income last year. _____

8. Do you have children? Yes_____ No_____

The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

a. Develop appropriate descriptive statistics to summarize the data.

```
##       Age           Gender    Real.Estate.Purchases. Value.of.Investments....
##  Min.   :19.00   Female:181   No :229                Min.   :     0
##  1st Qu.:28.00   Male  :229   Yes:181                1st Qu.: 18300
##  Median :30.00                                       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         No :154           Min.   : 16200       No :191
##  1st Qu.: 4.000         Yes:256           1st Qu.: 51625       Yes:219
##  Median : 6.000                           Median : 66050
##  Mean   : 5.973                           Mean   : 74460
##  3rd Qu.: 7.000                           3rd Qu.: 88775
##  Max.   :21.000                           Max.   :322500
```

这个数据集包括了年龄、性别、房地产购买情况、投资价值、交易数量、宽带接入、家庭收入和是否有孩子等多个维度的信息。首先，从年龄上看，数据集中的最小年龄是 19 岁，最大年龄是 42 岁，平均年龄是 30.11 岁。其中，19 岁到 28 岁之间的人数最多，其次是 28 岁到 33 岁之间的人。在性别方面，女性占 181 人，男性占 229 人。在房地产购买情况方面，有 181 人计划 2 年内购买房产，而没有意愿的人有 229 人。这表明大约 47% 的人在调查前已经购买了房地产。在投资方面，平均投资价值为 28538 元，最小投资价值为 0 元，最大投资价值为 133400 元。此外，平均交易数量为 5.973 次，最小交易数量为 0 次，最大交易数量为 21 次。在宽带接入方面，有 256 人表示他们已经在家中使用了宽带接入，而有 154 人没有使用宽带接入。最后，在家庭收入方面，平均家庭收入为 74460 元，最小家庭收入为 16200 元，最大家庭收入为 322500 元。此外，有 219 人表示他们有孩子，而有 191 人表示他们没有孩子。

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

```
## [1] 29.72269 30.50170
```

```
## [1] 71089.26 77829.77
```

对于年龄字段，95% 的置信区间是 [29.72,30.50]，意味着我们有 95% 的把握认为这个数据集的平均年龄在 29.72 岁到 30.50 岁之间。对于家庭收入字段，95% 的置信区间是 [71089.26,77829.77]，意味着我们有 95% 的把握认为这个数据集的家庭平均收入在 71089.26 元到 77829.77 元之间

c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```
## [1] 0.4858025 0.5824902
```

```
## [1] 0.5774567 0.6713237
```

有 95% 的把握认为，在这个数据集中，有孩子的用户的比例在 0.4858025 到 0.5824902 之间，有宽带的用户在 0.5774567 到 0.6713237 之间。

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

有 95% 的把握认为，在这个数据集中，该杂志受众意愿买房的的比例在 0.3933396 到 0.4895872 之间，并不算一个很好的广告渠道

```
## [1] 0.3933396 0.4895872
```

e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

```
## [1] 0.2955068 0.3874201
```

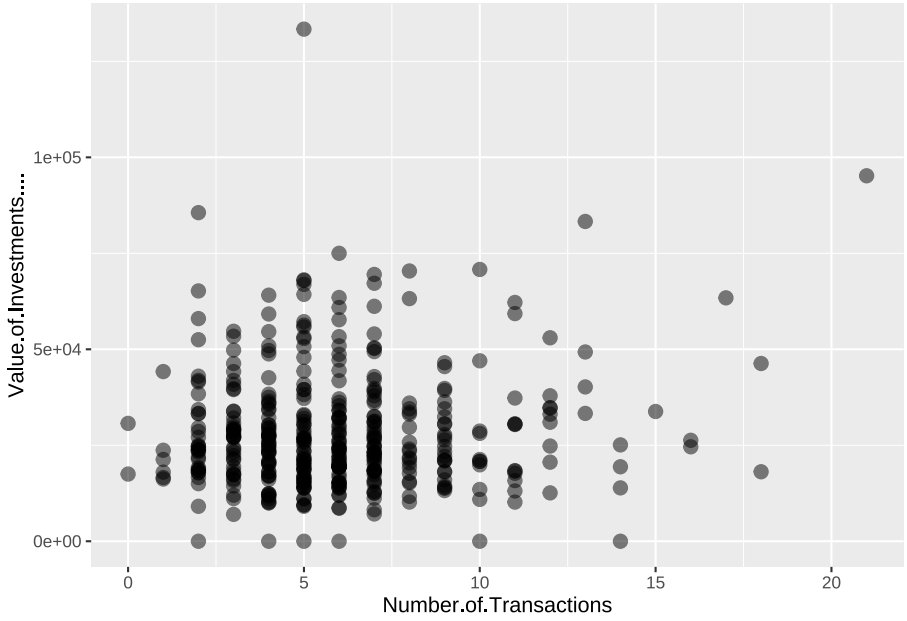
有 95% 的把握认为，在这个数据集中，有孩子且有宽带的用户的比例在 0.2955068 到 0.3874201 之间，因此不适合作为销售教育软件和儿童电脑游戏的公司做广告

f. Comment on the types of articles you believe would be of interest to readers of *Young Professional*. 评论您认为《青年专业人士》读者感兴趣的文章类型。

有 95% 的把握认为，在这个数据集中，用户使用的投资种类在的比例在 5.673019 到 6.273322 之间，投资金额在 27007.87 到 30068.71 之间。因此，该杂志适合投资、理财类的文章

```
## [1] 5.673019 6.273322
```

```
## [1] 27007.87 30068.71
```



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

Quality associates, inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. in one particular application, a client gave Quality associates a sample of 800 observations taken during a time in which that client's process was operating satisfactorily. the sample standard deviation for these data was .21; hence, with so much data, the population standard deviation was assumed to be .21. Quality associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. by analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. when the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. the design specification indicated the mean for the process should be 12. the hypothesis test suggested by Quality associates follows.

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

Corrective action will be taken any time H_0 is rejected.

Data are available in the data set Quality.

Managerial Report

a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what action, if any, should be taken. Provide the p-value for each test.

```
## [1] 0.3127296
```

```
## [1] 0.4818209
```

```
## [1] 0.006468822
```

```
## [1] 0.03905895
```

4 个样本中，sample1，2，4 计算的 p 值均大于 0.01，符合零假设 =12；sample3 计算的结果小于 0，不符合假设

b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

```
## [1] 0.2134979
```

4 组样本数据的平均方差为 0.2134979，与给定的总体标准差比较接近，是合理的

c. compute limits for the sample mean $\overline{x}$ around $\mu = 12$ such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if $\overline{x}$ exceeds the upper

limit or if $\overline{x}$ is below the lower limit, corrective action will be taken. these limits are referred to as upper and lower control limits for quality control purposes.

```
## [1] 11.87981 12.03752
```

```
## [1] 11.94981 12.10752
```

```
## [1] 11.81487 11.96313
```

```
## [1] 12.00758 12.15509
```

d. discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased? 由于显著性水平提高，对于任何给定的样本数据，更大的差异或变化可能会被检测到。这可能会导致过度敏感的结论或错误的发现，一些重要的信息被忽略，可能会导致第二类错误的概率增加，从而降低了假设检验的可靠性。

Question #6: Vacation occupancy rates were expected to be up during March 2008 in Myrtle Beach, South Carolina (*the sun news,* February 29, 2008). Data in the file Occupancy (Attached file **Occupancy**) will allow you to replicate the findings presented in the newspaper. The data show units rented and not rented for a random sample of vacation properties during the first week of March 2007 and March 2008.

a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
## [1] 0.35
```

```
## [1] 0.4666667
```

b. Provide a 95% confidence interval for the difference in proportions.

```
## [1] 0.2837307 0.4162693
```

```
## [1] 0.3865620 0.5467713
```

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

```
##
##  Welch Two Sample t-test
##
## data:  occupy$`March 2008` by occupy$`March 2007`
## t = 1.5617, df = 120.66, p-value = 0.121
## alternative hypothesis: true difference in means between group 0 and group 1 is not
## 95 percent confidence interval:
##  -0.03484688  0.29517512
## sample estimates:
## mean in group 0 mean in group 1
##       0.5161290       0.3859649
```

2008 年和 2007 年均值没有显著差异，不能确认 08 年之后的的入住率会有显著增长

Question #7: **Air Force Training Program** (data file: Training)

An air force introductory course in electronics uses a personalized system of instruction whereby each student views a videotaped lecture and then is given a programmed instruc-tion text. the students work independently with the text until they have completed the training and passed a test. Of concern is the varying pace at which the students complete this portion of their training program. Some students are able to cover the programmed instruction text relatively quickly, whereas other students work much longer with the text and require additional time to complete the course. The fast students wait until the slow students complete the introductory course before the entire group proceeds together with other aspects of their training.

A proposed alternative system involves use of computer-assisted instruction. In this method, all students view the same videotaped lecture and then each

is assigned to a computer terminal for further instruction. The computer guides the student, working independently, through the self-training portion of the course.

To compare the proposed and current methods of instruction, an entering class of 122 students was assigned randomly to one of the two methods. one group of 61 students used the current programmed-text method and the other group of 61 students used the proposed computer-assisted method. The time in hours was recorded for each student in the study. Data are provided in the data set training (see Attached file).

Managerial Report

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?

```
##     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
```

```
## [1] 3.944907
```

```
## [1] 2.506385
```

两种方法的中位数、四分位数和最大值都几乎相同，说明两种方法在分布上没有显著的偏态或异常值。

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

```
##
##  Welch Two Sample t-test
##
## data:  training$Current and training$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
```

通过 t 经验，p-value 为 0.5481，大于 0.05，因此我们不能拒绝原假设，即两个样本的均值是相等的

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.

```
## [1] 3.944907
```

```
## [1] 2.506385
```

```
## [1] 15.5623
```

```
## [1] 6.281967
```

```
##
##  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
```

F 值是 2.4773，表示两组数据的方差比例。num df 和 denom df 的值都是 60，分别代表了分子和分母的自由度。p-value 是 0.000578，这个值小于常用的显著性水平 0.05，因此我们有理由拒绝零假设，即认为这两组数据的方差是不相等的。95% 的置信区间是 1.486267 到 4.129135，这个区间不包含 1，也支持了我们拒绝零假设的结论。sample estimates 下面的 ratio of ariance 是 2.477296，这个值大于 1，进一步证实了两组数据的方差存在显著差异

d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain. 在目前的实验样本中，两种方法在分布上没有显著的偏态或异常值；只是对比"Current"方法，"Proposed" 培训时间数据标准差更小，相对更集中。可能的原因是新教学方法应用降低学习难度，缩小学生差距

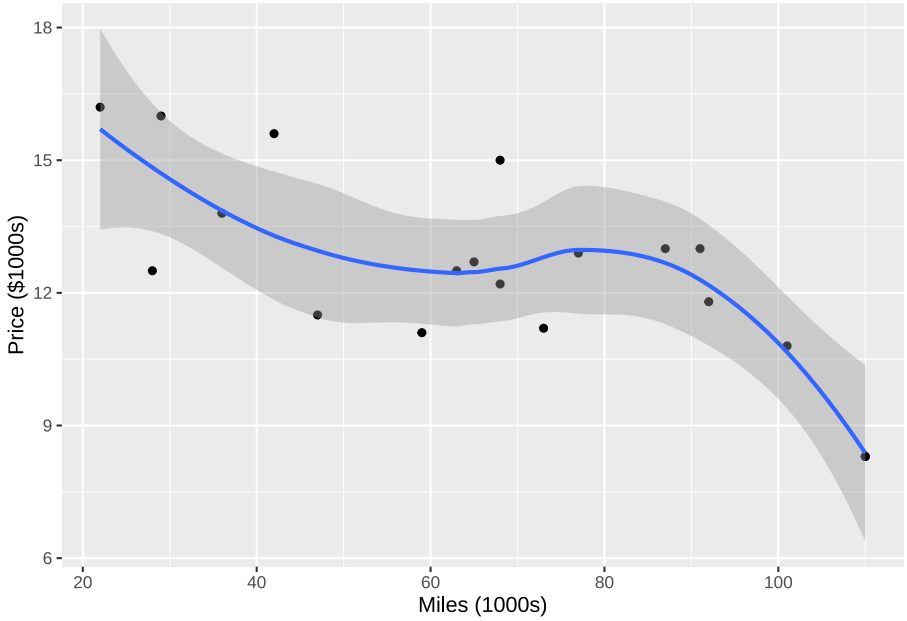
e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future? 建议加长测试时间，加大样本量

Question #8: The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, Attached data file Camry show the mileage and sale price for 19 sales (Pricehub website, February 24, 2012).

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

```
##  Miles (1000s)    Price ($1000s)
##  Min.   : 22.00   Min.   : 8.30
```

```
##  1st Qu.: 44.50   1st Qu.:11.35
##  Median : 68.00   Median :12.50
##  Mean   : 66.74   Mean   :12.55
##  3rd Qu.: 89.00   3rd Qu.:13.40
##  Max.   :110.00   Max.   :16.20
```



b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

根据图像，价格随着行驶里数逐渐下降

c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

```
##
## Call:
## lm(formula = `Price ($1000s)` ~ `Miles (1000s)`, data = camry)
##
```

```
## Coefficients:
##     (Intercept)  `Miles (1000s)`
##        16.46976         -0.05877

##
## Call:
## lm(formula = `Price ($1000s)` ~ `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 ***
## `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
```

拟合曲线为 y=-0.05877x+16.46976

d. Test for a significant relationship at the .05 level of significance.

F 统计量的 p 值为 0.0003475，远小于 0.05，是显著的

e. Did the estimated regression equation provide a good fit? Explain.

虽然模型的拟合度不是非常好（R-squared 和调整 R 平方都只略超过 50%），但是该模型是显著的，并且能够为 Price($1000s) 提供一个合理的预测

f. Provide an interpretation for the slope of the estimated regression equation.

Miles (1000s) 这个预测变量的系数是-0.05877, 意味着每增加 1000 英里, 预计价格会下降 0.05877$/1000

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.

```
## [1] 12.94356
```

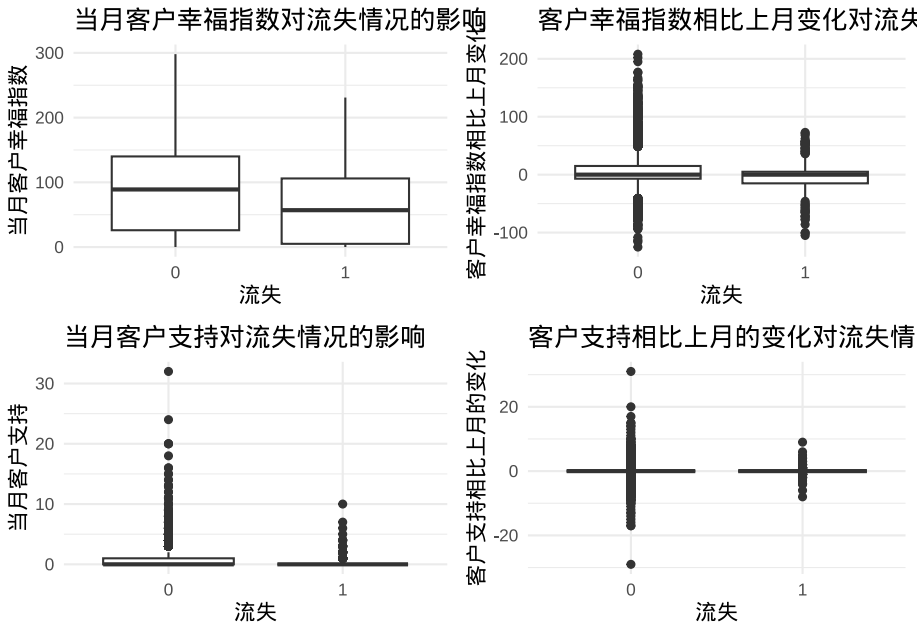
根据模型预测价格约 13,000 美元

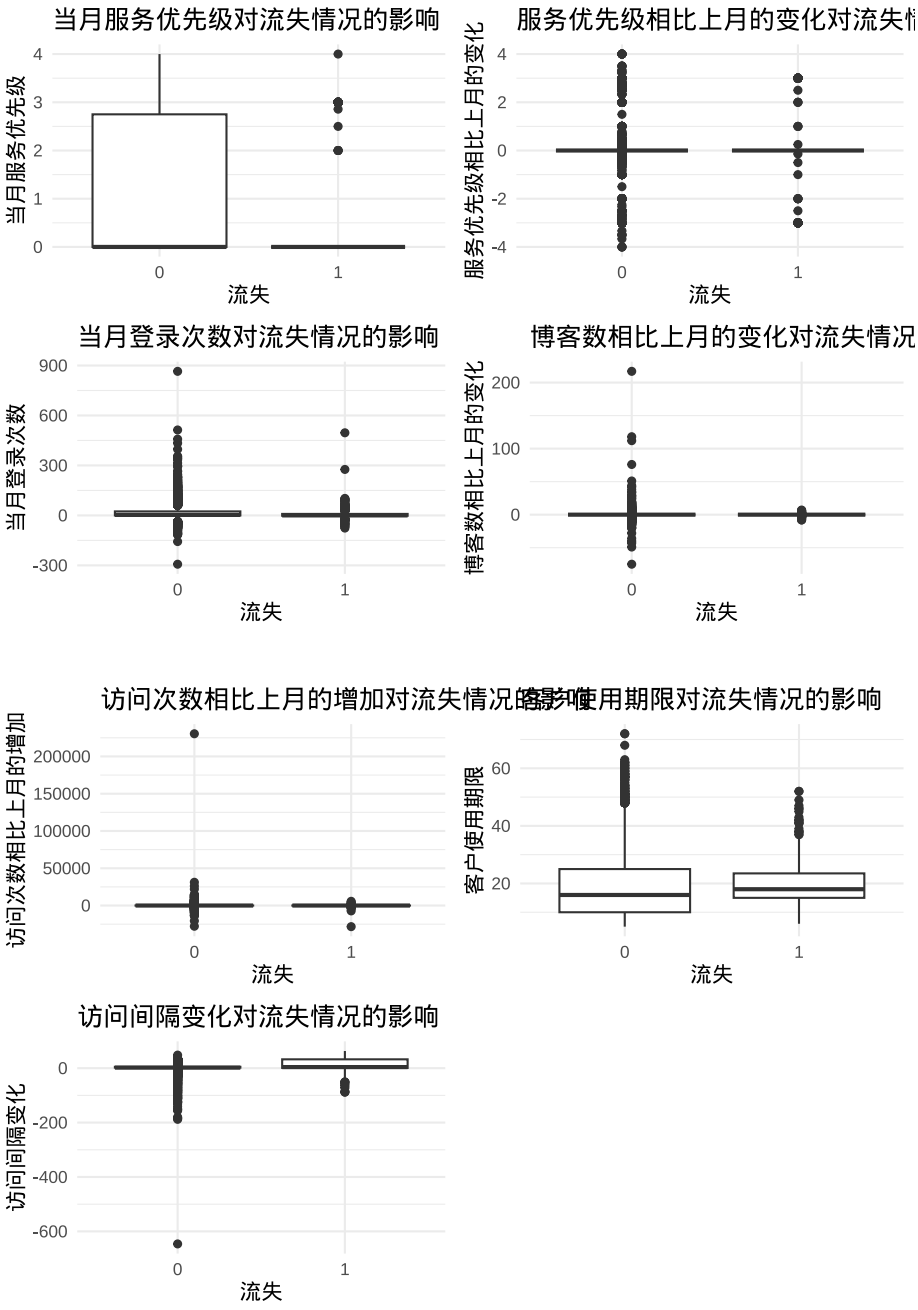
Question #9: 附件 WE.xlsx 是某提供网站服务的 Internet 服务商的客户数据。数据包含了 6347 名客户在 11 个指标上的表现。其中" 流失" 指标中 0 表示流失，"1" 表示不流失，其他指标含义看变量命名。

```
##       客户ID          流失          当月客户幸福指数 客户幸福指数相比上月变化
##  Min.   :   1   Min.   :0.00000   Min.   :  0.00   Min.   :-125.000
##  1st Qu.:1588   1st Qu.:0.00000   1st Qu.: 24.50   1st Qu.:  -8.000
##  Median :3174   Median :0.00000   Median : 87.00   Median :   0.000
##  Mean   :3174   Mean   :0.05089   Mean   : 87.32   Mean   :   5.059
##  3rd Qu.:4760   3rd Qu.:0.00000   3rd Qu.:139.00   3rd Qu.:  15.000
##  Max.   :6347   Max.   :1.00000   Max.   :298.00   Max.   : 208.000
##   当月客户支持     客户支持相比上月的变化 当月服务优先级
##  Min.   : 0.0000   Min.   :-29.000000     Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:  0.000000     1st Qu.:0.0000
##  Median : 0.0000   Median :  0.000000     Median :0.0000
##  Mean   : 0.7063   Mean   : -0.006932     Mean   :0.8128
##  3rd Qu.: 1.0000   3rd Qu.:  0.000000     3rd Qu.:2.6667
##  Max.   :32.0000   Max.   : 31.000000     Max.   :4.0000
##  服务优先级相比上月的变化  当月登录次数      博客数相比上月的变化
##  Min.   :-4.00000         Min.   :-293.00   Min.   :-75.0000
```

```
##  1st Qu.: 0.00000          1st Qu.:  -1.00   1st Qu.:  0.0000
##  Median : 0.00000          Median :   2.00   Median :  0.0000
##  Mean   : 0.03017          Mean   :  15.73   Mean   :  0.1572
##  3rd Qu.: 0.00000          3rd Qu.:  23.00   3rd Qu.:  0.0000
##  Max.   : 4.00000          Max.   : 865.00   Max.   :217.0000
##  访问次数相比上月的增加  客户使用期限    访问间隔变化
##  Min.   :-28322.00      Min.   : 5.0   Min.   :-646.000
##  1st Qu.:   -11.00      1st Qu.:10.0   1st Qu.:   2.000
##  Median :     0.00      Median :16.0   Median :   2.000
##  Mean   :    96.31      Mean   :18.9   Mean   :   3.765
##  3rd Qu.:    27.00      3rd Qu.:25.0   3rd Qu.:   5.000
##  Max.   :230414.00      Max.   :72.0   Max.   :  63.000
```

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?





根据图像，当月客户幸福指数、当月服务优先级、客户使用期限可能存在显著不同

b. 通过均值比较的方式验证上述不同是否显著。

```
##
##  Welch Two Sample t-test
##
## data:  we$当月客户幸福指数 by we$流失
## 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
## 95 percent confidence interval:
##  18.79956 31.86737
## sample estimates:
## mean in group 0 mean in group 1
##        88.60591        63.27245

##
##  Welch Two Sample t-test
##
## data:  we$当月服务优先级 by we$流失
## t = 5.1428, df = 373.13, p-value = 4.381e-07
## alternative hypothesis: true difference in means between group 0 and group 1 is not
## 95 percent confidence interval:
##  0.2038355 0.4562009
## sample estimates:
## mean in group 0 mean in group 1
##       0.8295759       0.4995577

##
##  Welch Two Sample t-test
##
## data:  we$客户使用期限 by we$流失
## t = -2.9811, df = 379.9, p-value = 0.003057
## alternative hypothesis: true difference in means between group 0 and group 1 is not
## 95 percent confidence interval:
##  -2.5461200 -0.5223121
```

```
## sample estimates:
## mean in group 0 mean in group 1
##        18.81873        20.35294
```

三种因素进行 t 检验的结果表明，流失、非流失的均值均不相同，且 95% 的置信区间不包含 0，p 小于 0.05。总之支持了这样的假设：客户的流失与幸福指数、服务优先级、客户使用期限可能存在关联。

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。

```
##
## Call:
## lm(formula = 流失 ~ 当月客户幸福指数 + 当月服务优先级 +
##     客户使用期限, data = wee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14223 -0.06653 -0.05056 -0.03022  1.00473
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.897e-02  5.961e-03   9.894  < 2e-16 ***
## 当月客户幸福指数 -3.175e-04  4.766e-05  -6.660 2.96e-11 ***
## 当月服务优先级   -2.725e-03  2.281e-03  -1.195    0.232
## 客户使用期限       1.156e-03  2.625e-04   4.406 1.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2187 on 6343 degrees of freedom
## Multiple R-squared:  0.01072,    Adjusted R-squared:  0.01025
## F-statistic:  22.9 on 3 and 6343 DF,  p-value: 9.67e-15
```

拟合回归方程表达式为：y = 0.05897 - 0.0003175* x1 - 0.002725* x2

+0.001156* x3 其中：y 表示流失概率，x1 表示当月客户幸福指数，x2 表示当月服务优先级，x3 表示客户使用期限

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。

```
##   [1]    1   14    3   18   21    2   56   51   54   58    5   59   12  116   73
##  [16]   91   60   42  105  132   66  148   72  114  141  164  176  183  118   97
##  [31]   86  104   30   16  136 1280  137   65 1281  101 1306 1325   81   62 1283
##  [46] 1362 1402 2061 2066 2080  195   10  107 1345 1348 2071   17 1378  151   69
##  [61]  125  142  123   61   47  156  122 1006 1041  108   41 2111 1358  162 1742
##  [76] 1755 2070  187 1271  128 1795 1814  172  185   46  117  990   23 1333 1039
##  [91]  110 1884 1899 1907  159   93 2106 2130  959  964
```