第二次陈老师作业

-李梓青

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.

b. Compute the mean, median, and mode.

c. Compute the first and third quartiles.

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

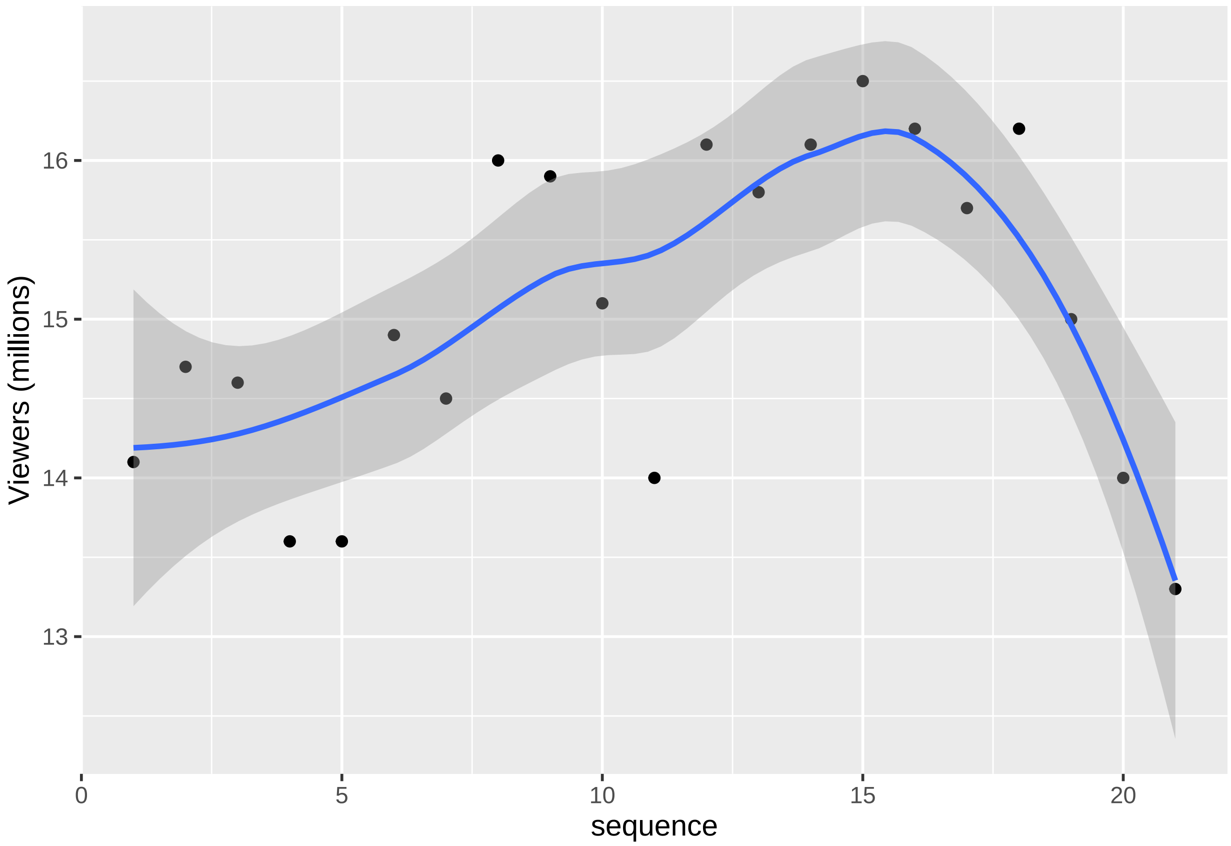
```
##    Air Date          Viewers (millions)
##  Length:21          Min.   :13.30
##  Class :character   1st Qu.:14.10
##  Mode  :character   Median :15.00
##                     Mean   :15.04
##                     3rd Qu.:16.00
##                     Max.   :16.50
```

```
## [1] "list"
```

(a) 观众最少是 13.3 亿元，最多是 16.5 亿元

(b) 观众的均值是 15.04 亿，中位数是 15 亿，mode 是 list

(c)first quartiles 是 14.1 亿, third quartiles 是 16 亿



(d) 在 2011–2012 season 的 21 场节目中，直到 2012 年 2 月 9 号前收视率都是上升趋势，随后收视率下降

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. Show the frequency distribution.

```
##     Value Frequency
## 1 (11,13]         2
## 2 (13,15]         5
## 3 (15,17]        14
## 4 (17,19]        16
## 5 (19,21]         4
## 6 (21,23]         3
## 7 (23,25]         1
## 8 (25,27]         2
## 9 (27,29]         3
```

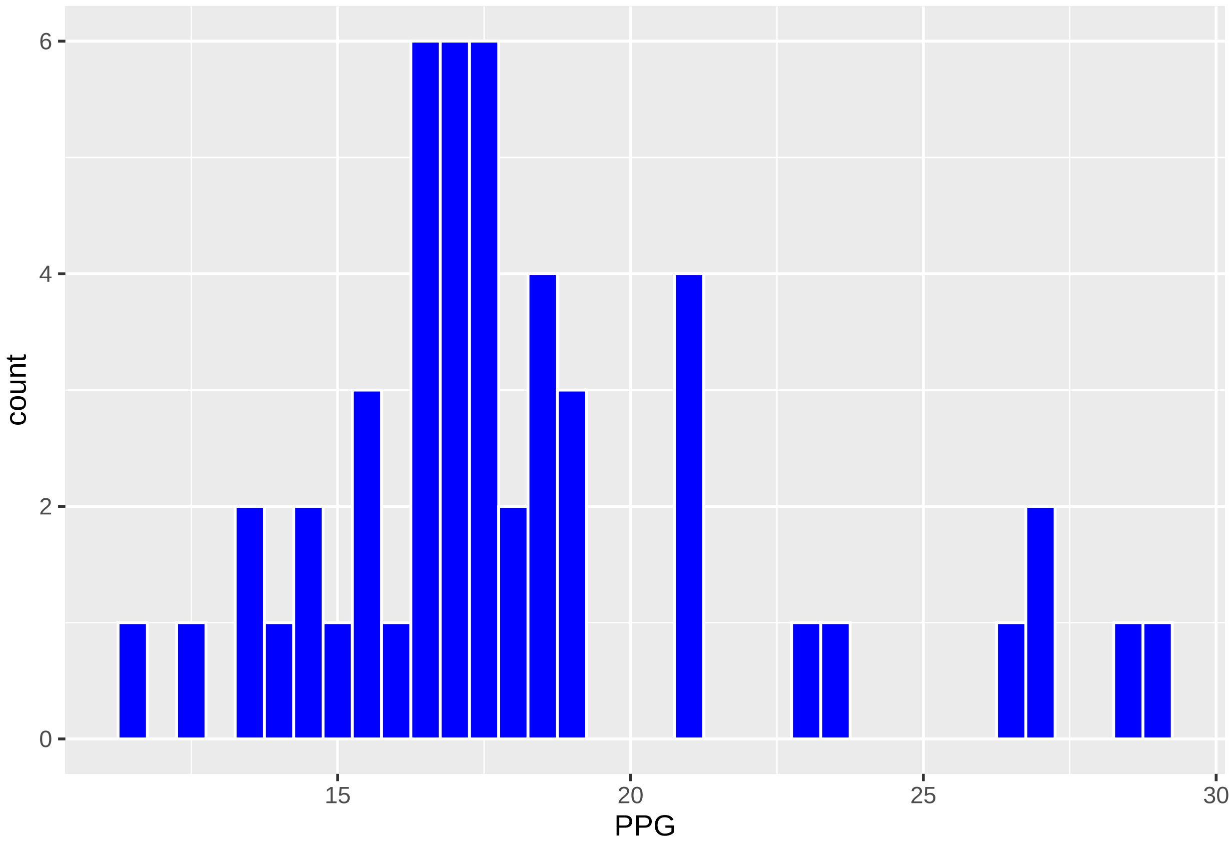
b. Show the relative frequency distribution.

```
##   Interval Relative_Frequency
## 1  (11,13]               0.04
## 2  (13,15]               0.10
## 3  (15,17]               0.28
## 4  (17,19]               0.32
## 5  (19,21]               0.08
## 6  (21,23]               0.06
## 7  (23,25]               0.02
## 8  (25,27]               0.04
## 9  (27,29]               0.06
```

c. Show the cumulative percent frequency distribution.

```
##         Cumulative_Relative_Frequency
## (11,13]                          0.04
## (13,15]                          0.14
## (15,17]                          0.42
## (17,19]                          0.74
## (19,21]                          0.82
## (21,23]                          0.88
## (23,25]                          0.90
## (25,27]                          0.94
## (27,29]                          1.00
```

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



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

数据有一点右偏。图形的尾巴有向右延申，但是右边的分数没左边紧密。

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

```
##      n
## 1 0.22
```

球员中至少场均得分 20 分的比例是 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
```

在这次调查中使用的样本有 625 个

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

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

点估计值在总体均值的 ± 值 25 范围内的概率是 78.87%。

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.

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

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

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

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

f. Comment on the types of articles you believe would be of interest to readers of *Young Professional*.

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

```
## [1] "Gender"
```

```
##
## Female   Male
##    181    229
```

```
## [1] "Real Estate Purchases?"
```

```
##
##  No Yes
## 229 181
```

```
## [1] "Broadband Access?"
```

```
##
##  No Yes
## 154 256
```

```
## [1] "Have Children?"
```

```
##
##  No Yes
## 191 219
```

这些受调查的人一共又 410 人，其中，（1）年龄最小为 19 岁，最大 42 岁，均值为 30.11 岁，中位数是 30 岁；

（2）44.15% 是女性，有 181 人，55.85% 是男性，有 229 人；

（3）有 181 人最近两年内购买房产的想法，占 44.15%；

（4）家庭金融资产最小为 0，最大为 133400 美元，均值为 28538 美元，中位数为 24800 美元；

（5）过去一年股票交易最少为 0 次，最多为 21 次，均值 5.973 次，中位数 6 次；

（6）62.44% 的人用宽带上网，有 256 人；

（7）去年的家庭年收入最小 16200 美元，最多 322500 美元，均值 74460 美元，中位数 66050 美元；

（8）53.4% 的人有孩子，有 219 人。

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

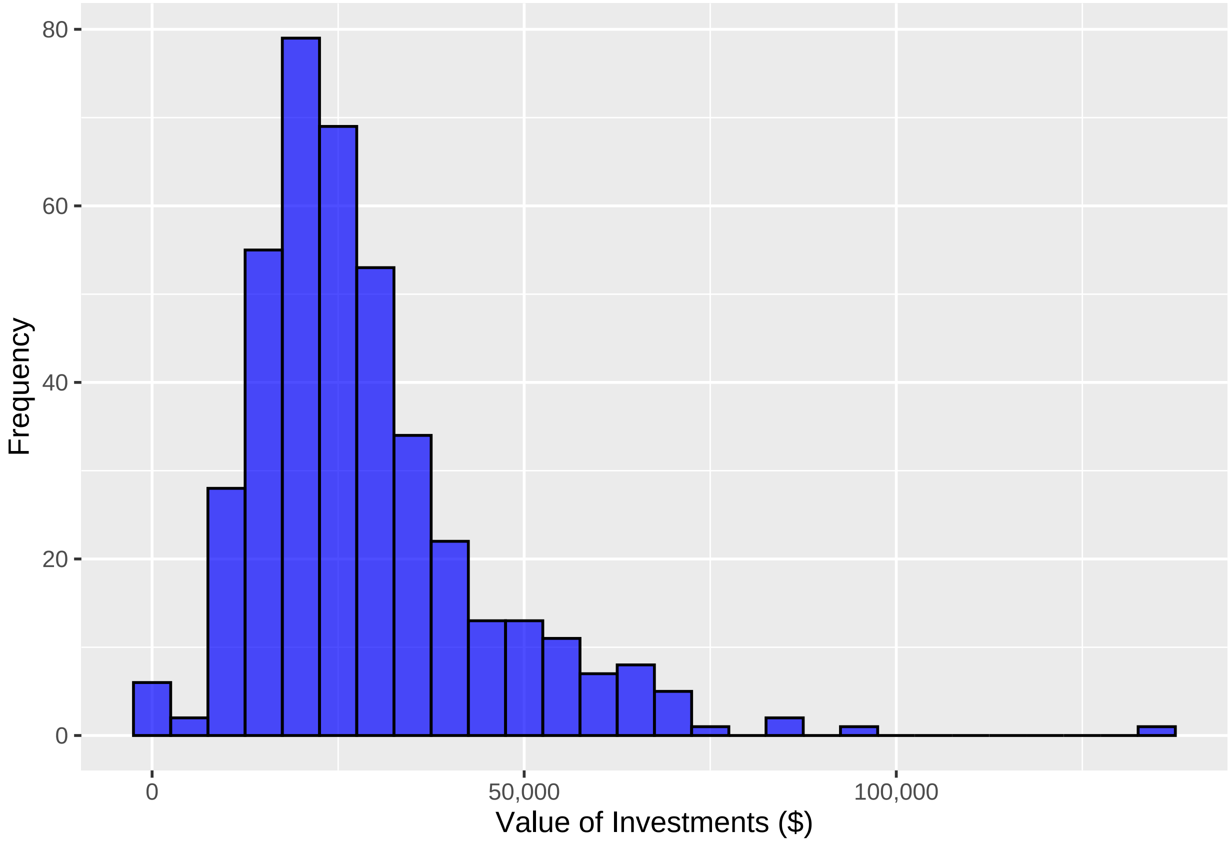
```
## [1] 71079.26 77839.77
## attr(,"conf.level")
## [1] 0.95
```

平均年龄 95% 的置性区间为 [29.7,30.5]; 家庭收入 95% 的置性区间为 [71079.26,77839.77]

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

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

在家有宽带接入的用户比例 95% 的置信区间为 [57.53%,67.11%]; 和有孩子的用户比例 95% 的置信区间为 [48.46%,58.31%]



我认为 *Young Professional* 会成为在线经纪人的一个很好的广告渠道，有网络的人超过一大半且如上图所示，绝大多数人有进行金融投资。

```
## [1] 0.2960551 0.3898927
## attr(,"conf.level")
## [1] 0.95
```

我认为这本杂志是为销售教育软件和电脑游戏的公司做广告的不错的地方，一半以上的人有电脑有宽带，可以玩游戏；其中 30% 左右的人已经有小孩，剩下的人总是还要有小孩的，所以不管是做游戏软件还是教育软件都不错。

根据以上的分析，我认为这本杂志读者是高净值人群，不管是做房地产广告还是金融投资广告，又或者游戏广告和孩子的广告都很合适。

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.

```
## Sample 1 :
##   t-统计量: -1.027
##   自由度: 29
##   p-值: 0.313
##
##   无法拒绝零假设
##
## Sample 2 :
##   t-统计量: 0.713
##   自由度: 29
##   p-值: 0.482
##
##   无法拒绝零假设
##
## Sample 3 :
##   t-统计量: -2.935
##   自由度: 29
##   p-值: 0.006
##
##   在显著性水平0.01下拒绝零假设
##
## Sample 4 :
##   t-统计量: 2.161
##   自由度: 29
##   p-值: 0.039
##
##   无法拒绝零假设
```

样本一、二、四无法拒绝原假设，无需采取行动，样本三在显著性水平0.01下拒绝原假设，应采取改进措施。

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

```
## Sample 1 标准差: 0.2204
## Sample 2 标准差: 0.2204
## Sample 3 标准差: 0.2072
## Sample 4 标准差: 0.2061
```

```
## 样本一的标准差和假设的标准差对比: TRUE
```

```
## 样本二的标准差和假设的标准差对比: TRUE
```

```
## 样本三的标准差和假设的标准差对比: TRUE
```

```
## 样本四的标准差和假设的标准差对比: TRUE
```

综上我觉得 0.21 的标准差设置还是合理的。

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.

```
## Upper Control Limit (UCL): 12.08
```

```
## Lower Control Limit (LCL): 11.966
```

设置这个样本的上下限为 [11.966,12.08]

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?

当显著性水平增加时，意味着我们更容易拒绝零假设，更容易得出“统计显著”的结论, 即更容易犯“第一类错误”。

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.

```
## 估计2007年3月第一周单位被出租的比例: 0.348
```

```
## 2008年3月第一周单位被出租的比例: 0.464
```

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

```
## [1] TRUE
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] 0.2834667 0.4189548
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 0.3827392 0.5462946
## attr(,"conf.level")
## [1] 0.95
```

2007 年 3 月第一周出租比例 95% 的置信区间为 [0.283,0.419];2008 年 4 月第一周出租比例 95% 的置信区间为 [0.383,0.546]

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

```
## [1] 0.2768834
```

```
## [1] -0.0462485
```

两个年份的出租比例的差异的置信区间差异包含 0，2008 年的出租比例不一定比 2007 年的高。但是大概率会高。至于出租租金，还需要更多的信息才能得出。

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
```

使用当前的方法的话，学生学习时长的最小值为 65，最大值为 84，均值为 75.07，中位数为 76；使用计算机辅助方法，学生学习时长最小值为 69，最大值为 82，均值为 75.43，中位数为 76.

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

```
## 比较两种方法的总体均值:
```

```
##
##  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
```

t = -0.60268, 当前学习方法的学习时长均值小于使用计算机辅助方法的学习时长均值，但是两者相差不是很显著。95% 的置信区间包含零又进一步证实这一点。

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.

```
## 方差比较的假设检验:
```

```
##
##  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
```

两个变量的方差齐性检验的 F 值为 2.4773，对应的 p 值为 0.000578。p 值小于 0.05 的显著性水平，所以拒绝原假设（即两个样本来自的总体方差相等），认为两个变量的方差不相等。此外，方差比率的 95% 置信区间在 1.486267 到 4.129135 之间。方差比率的值约为 2.477296。即使用当前的学习方法，学生的学习时长方差较大。

d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain.

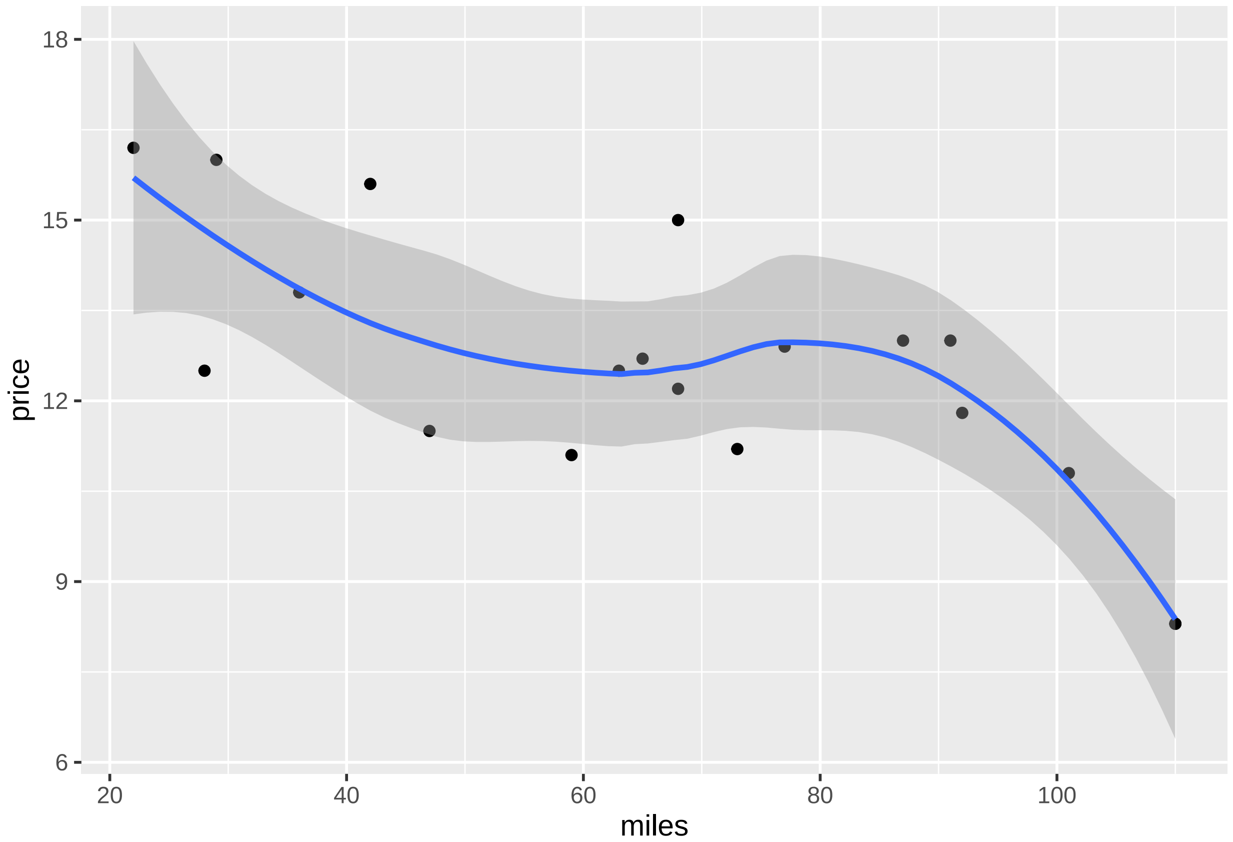
使用计算机辅助学习更集中，节省了学习慢者的学习时间，提高了组织效率，但是也略微拖累了天赋型学生的学习时间。

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?

我觉得可以让两边的学生交换学习新的内容，让第一次使用当前方法的学生第二次使用计算机辅助方法，让第一次使用计算机辅助的学生第二次使用当前方法，排除学生本身学习能力的影响。

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.



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

二手车车辆的价格是随着开过的里程的增加而减少的。

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

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 16.46975503 0.94876415 17.359167 2.986768e-12
## miles       -0.05877393 0.01319216 -4.455218 3.475110e-04
```

设 Y 为价格，X 为使用过的里程数。模拟出来的回归方程是 Y=-0.0588X+16.47

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

```
## 
## Call:
## lm(formula = price ~ 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 ***
## 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
```

在 0.05 的显著性水平下，里程和价格之间存在显著的关系

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

R-平方值为 0.5387，说明模型对价格的变异性有一定的解释能力，但并不是非常高。因此，我们可以说估计的回归方程提供了一定程度上的拟合，但不是完美的

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

回归方程中斜率的估计值是 -0.05877。这个斜率表示单位里程增加一千英里时，价格变化的速率。

在这个具体的例子中，由于斜率的估计值是负数，我们可以解释为：每增加一千英里的里程，车辆的价格平均下降约 0.05877 千美元。这是因为斜率反映了自变量（里程）对因变量（价格）的影响方向和程度。

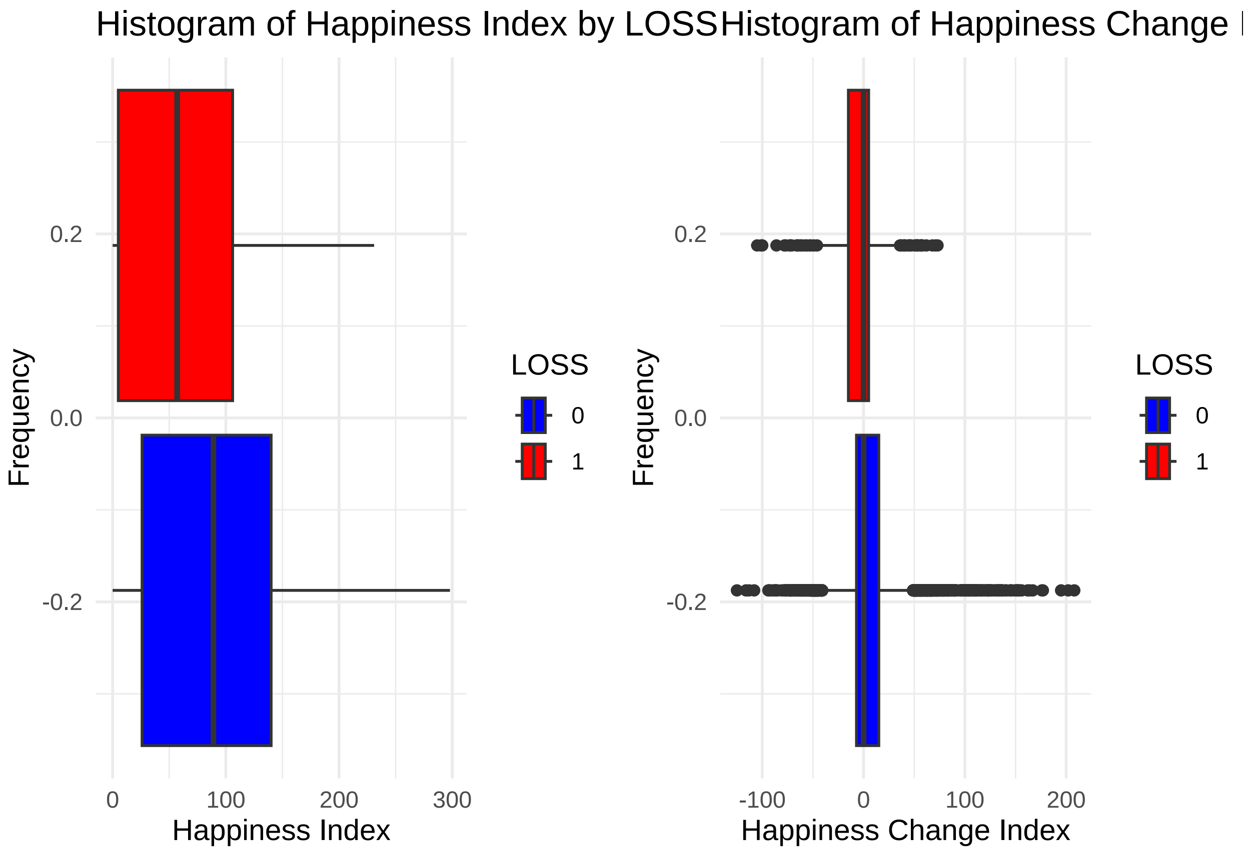
因此，在这个情境下，负的斜率表明里程的增加与价格的降低之间存在负相关关系。

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.

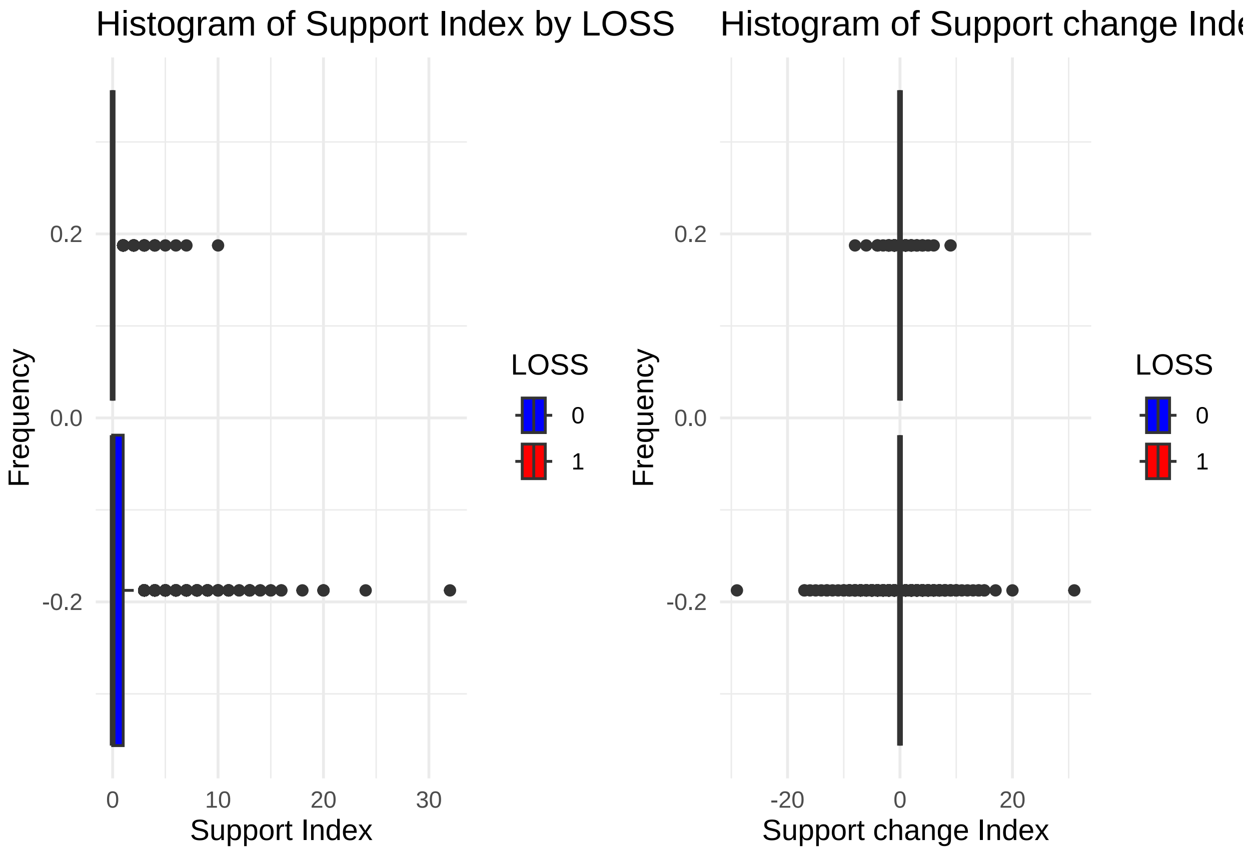
当 X=60,Y= 12.942。这个会是我像卖家卖的价格，和图 a 我模拟出来的数据很接近，并且从实际上看，有的人用比这更低的价格卖出去过。要是能以 12.942 千美元的价格卖出去，那太好了。

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

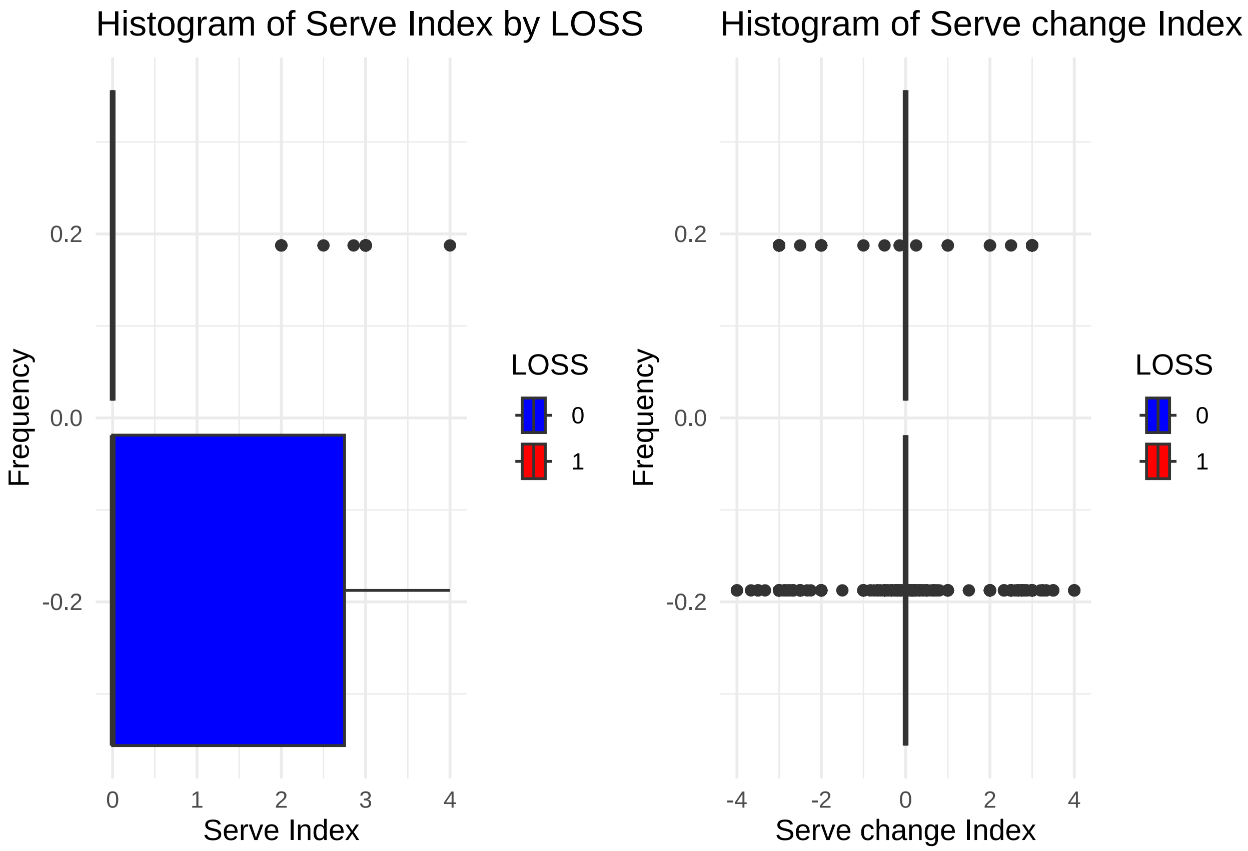
a. 通过可视化探索流失客户与非流失客户的行为特点（或特点对比），你能发现流失与非流失客户行为在哪些指标有可能存在显著不同？



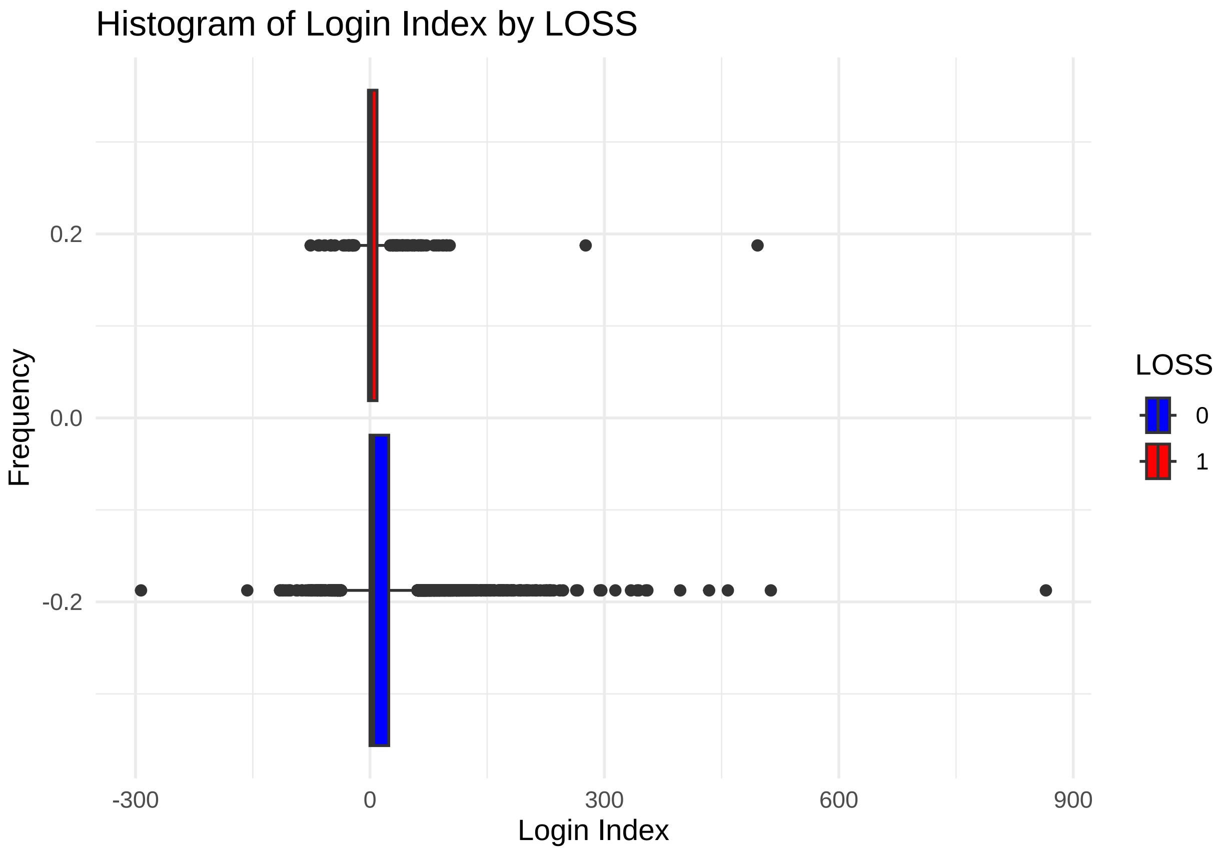
从图片上看当月客户幸福指数和客户幸福指数相比上月变化对客户流失影响还是很显著的，流失客户的指数均值明显更低。



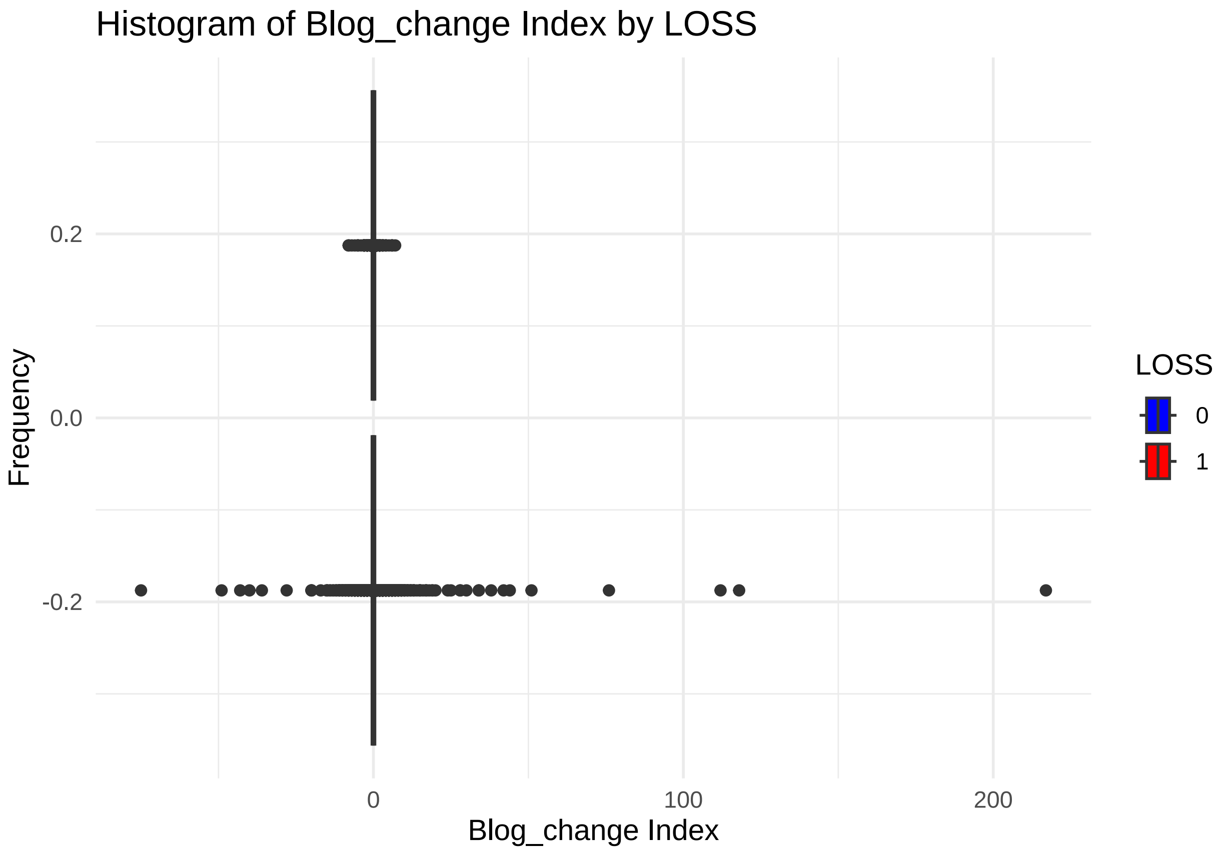
从图片上看当月客户支持对客户流失影响还是很显著的，流失客户的当月客户支持均值明显更低，整体靠左。而客户支持相比上月的变化不是很显著。



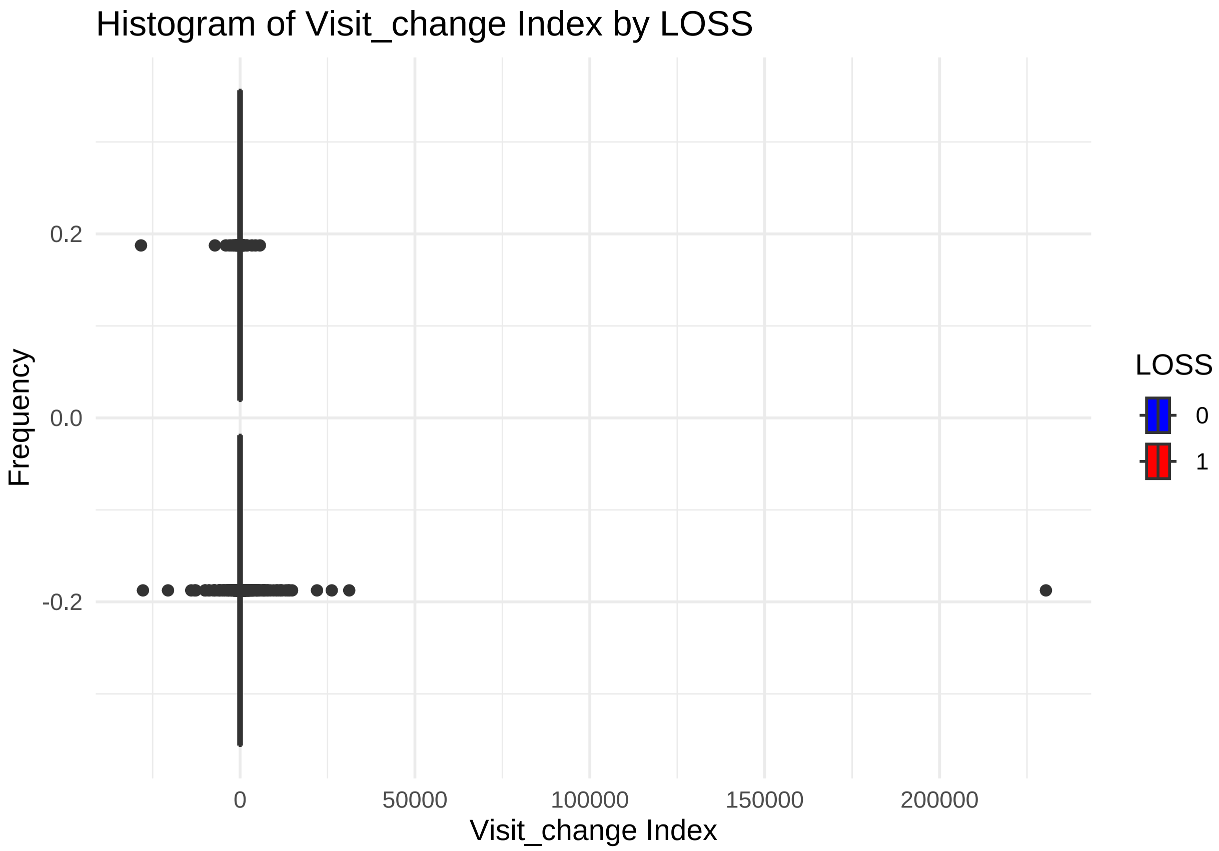
从图片上看当月服务优先级对客户流失影响还是很显著的，流失客户的当月服务优先级均值几乎为 0。而服务优先级相比上月的变化不是很显著。



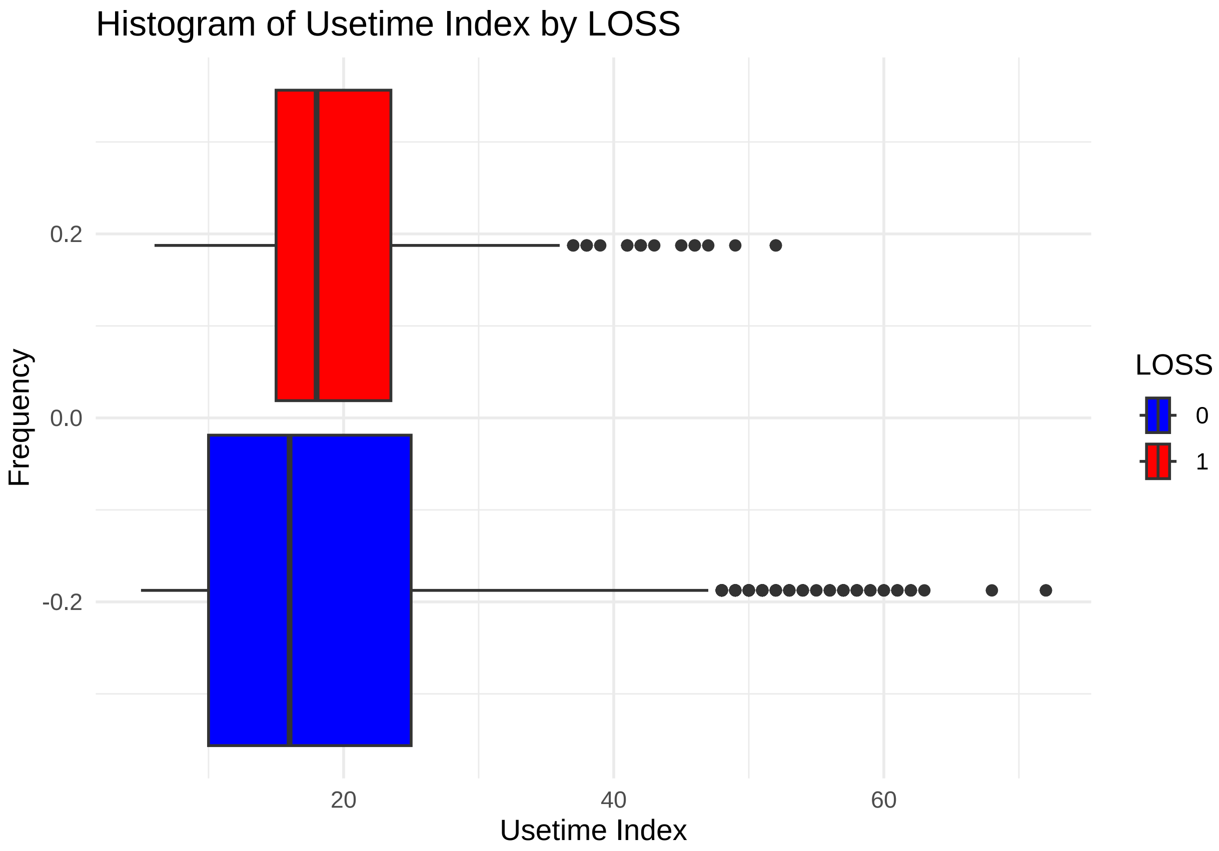
从图片上看当月登录次数影响还是很显著的，均值靠后。



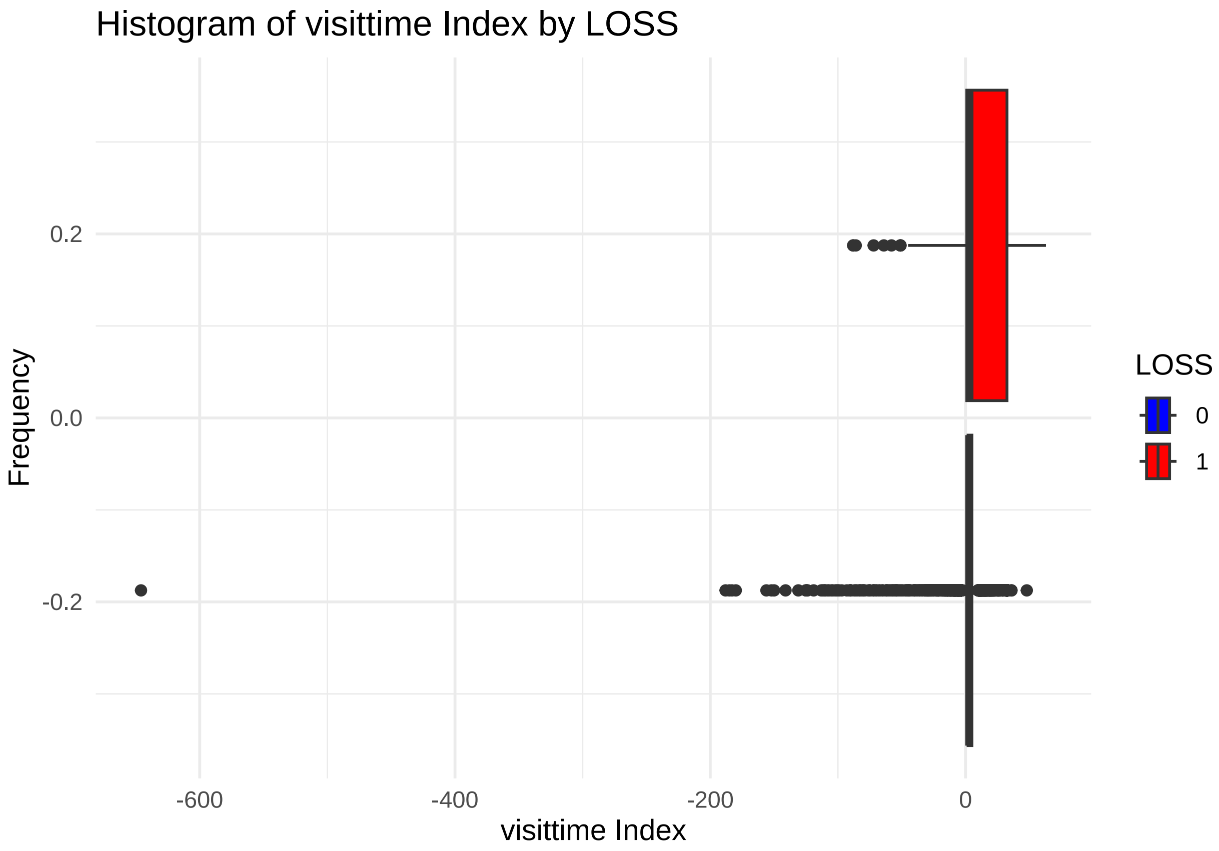
从图片上看博客数相比上月的变化均值对比不是很显著的，但是极值变化对比很明显。



从图片上看访问次数相比上月的增加影响不是很显著的。



从图片上看客户使用期限影响比较显著的。流失客户的使用期限均值更长。



从图片上看访问间隔变化影响比较显著的。流失客户的访问间隔时间均值更长。

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:  Happiness by Loss
## 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
```

```
##
##  Welch Two Sample t-test
##
## data:  Happy_change by Loss
## t = 5.7835, df = 365.71, p-value = 1.571e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   6.116137 12.417972
## sample estimates:
## mean in group 0 mean in group 1
##        5.530212       -3.736842
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Support by Loss
## t = 5.5099, df = 419.22, p-value = 6.281e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.2269082 0.4785969
## sample estimates:
## mean in group 0 mean in group 1
##       0.7242696       0.3715170
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Support_change by Loss
## t = -0.63198, df = 406.9, p-value = 0.5278
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.19092606  0.09803036
## sample estimates:
## mean in group 0 mean in group 1
##    -0.009296149     0.037151703
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Serve by Loss
## 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 equal to 0
## 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:  Serve_change by Loss
## t = 0.64116, df = 364.49, p-value = 0.5218
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1020692  0.2008252
## sample estimates:
## mean in group 0 mean in group 1
##      0.03268184     -0.01669615

##
##  Welch Two Sample t-test
##
## data:  Login by Loss
## t = 3.5709, df = 362.67, p-value = 0.0004037
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   3.628884 12.525166
## sample estimates:
## mean in group 0 mean in group 1
##        16.13894         8.06192

##
##  Welch Two Sample t-test
##
## data:  Blog_change by Loss
## t = 2.5315, df = 695.95, p-value = 0.01158
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.06133902 0.48529282
## sample estimates:
## mean in group 0 mean in group 1
##       0.1711487      -0.1021672

##
##  Welch Two Sample t-test
##
## data:  Visit_change by Loss
## t = 1.9136, df = 448, p-value = 0.05631
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   -5.463729 410.218457
## sample estimates:
## mean in group 0 mean in group 1
##        106.6096        -95.7678
```

```
##
##  Welch Two Sample t-test
##
## data:  Usetime by Loss
## t = -2.9811, df = 379.9, p-value = 0.003057
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.5461200 -0.5223121
## sample estimates:
## mean in group 0 mean in group 1
##        18.81873        20.35294
```

```
##
##  Welch Two Sample t-test
##
## data:  visittime by Loss
## t = -4.0971, df = 346.03, p-value = 5.215e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -7.362712 -2.586515
## sample estimates:
## mean in group 0 mean in group 1
##        3.511454        8.486068
```

Happiness：流失和非流失客户在幸福指数上存在显著差异（p-value < 0.05），流失客户的幸福指数均值较低。

Happy_change：流失和非流失客户在幸福指数变化上存在显著差异（p-value < 0.05），流失客户的幸福指数变化均值较低。

Support：流失和非流失客户在支持指数上存在显著差异（p-value < 0.05），流失客户的支持指数均值较低。

Support_change：流失和非流失客户在支持指数变化上差异不显著（p-value > 0.05）。

Serve：流失和非流失客户在服务优先级上存在显著差异（p-value < 0.05），流失客户的服务优先级均值较低。

Serve_change：流失和非流失客户在服务优先级变化上差异不显著（p-value > 0.05）。

Login：流失和非流失客户在登录次数上存在显著差异（p-value < 0.05），流失客户的登录次数均值较低。

Blog_change：流失和非流失客户在博客数变化上存在显著差异（p-value < 0.05），流失客户的博客数变化均值较低。

Visit_change：流失和非流失客户在访问次数变化上差异接近显著（p-value = 0.056），但未达到通常的显著

水平。

Usetime：流失和非流失客户在使用期限上存在显著差异（p-value < 0.05），流失客户的使用期限均值较低。

Visittime：流失和非流失客户在访问间隔变化上存在显著差异（p-value < 0.05），流失客户的访问间隔变化均值较高。

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:
## glm(formula = Loss ~ Happiness + Happy_change + Support + Serve +
##     Login + Blog_change + Visit_change + Usetime + visittime,
##     family = binomial, data = df2)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
## Happiness    -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
## Happy_change -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
## Support      -3.522e-02  7.438e-02  -0.474  0.63581
## Serve        -3.727e-02  7.514e-02  -0.496  0.61985
## Login         9.104e-04  1.952e-03   0.466  0.64098
## Blog_change  -2.357e-05  2.080e-02  -0.001  0.99910
## Visit_change -1.170e-04  4.069e-05  -2.877  0.00401 **
## Usetime       1.418e-02  5.260e-03   2.696  0.00701 **
## visittime     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
```

Loss =-0.005225 Happiness -0.009501 Happy_change -0.03522 Support -0.03727 Serve + 0.0009104 Login -0.00002357 Blog_change -0.000117 Visit_change + 0.01418 Usetime + 0.017 visittime-2.874

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。

```
##        ID
## 1   2287
## 2    109
## 3   1971
## 4   2025
## 5      1
## 6    929
## 7   2076
## 8     76
## 9     14
## 10    18
## 11     3
## 12  2244
## 13    21
## 14  1287
## 15  1929
## 16  1459
## 17    51
## 18   128
## 19   183
## 20    59
## 21    55
## 22   121
## 23  2240
## 24  1520
## 25  2599
## 26  1236
## 27   137
## 28  1862
## 29  2080
## 30  1143
## 31   154
## 32  1286
## 33  2546
## 34   146
## 35   119
## 36   171
## 37   190
## 38    42
## 39     5
## 40     2
## 41   123
## 42   101
## 43  1616
## 44    95
## 45  2680
## 46  2838
## 47    61
## 48  2289
## 49  1438
## 50  1392
## 51  2481
## 52  2924
## 53   106
## 54  3671
## 55   203
## 56  1393
## 57    69
## 58  1574
## 59  1204
## 60    68
## 61  2255
## 62  1395
## 63  1478
## 64  2235
## 65    89
## 66   798
## 67  1141
## 68  2739
## 69    62
## 70  4245
## 71  1151
## 72  2830
## 73  1693
## 74  3042
## 75    12
## 76   142
## 77  1908
## 78    10
## 79   868
## 80  2286
## 81  3076
## 82     57
## 83   2242
## 84   1951
## 85   3124
## 86   1019
## 87   1110
## 88   2062
## 89   2903
## 90   2913
## 91   2047
## 92    104
## 93   1953
## 94   2656
## 95   1155
## 96   2744
## 97   1446
## 98   2306
## 99    163
## 100   240
```