第二次作业

Code

your name

2024-11-29

1 导入需要用到的包

2 第一题：生活大爆炸观众数据分析

3 数据导入

- 数据概览 各变量的简短信息：

```
## Rows: 21
## Columns: 2
## $ `Air Date`           <chr> "September 22, 2011", "September 22, 2011", "Sept…
## $ `Viewers (millions)` <dbl> 14.1, 14.7, 14.6, 13.6, 13.6, 14.9, 14.5, 16.0, 1…
```

各变量的简短统计：

```
##    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
```

- 计算观众人数的最小值和最大值。

```
## 观众人数的最大值是： 16.5 百万
```

```
## 观众人数的最小值是： 13.3 百万
```

- 计算平均数、中位数和众数。

```
## 观众人数的平均数是： 15.04286 百万
```

```
## 观众人数的中位数是： 15 百万
```

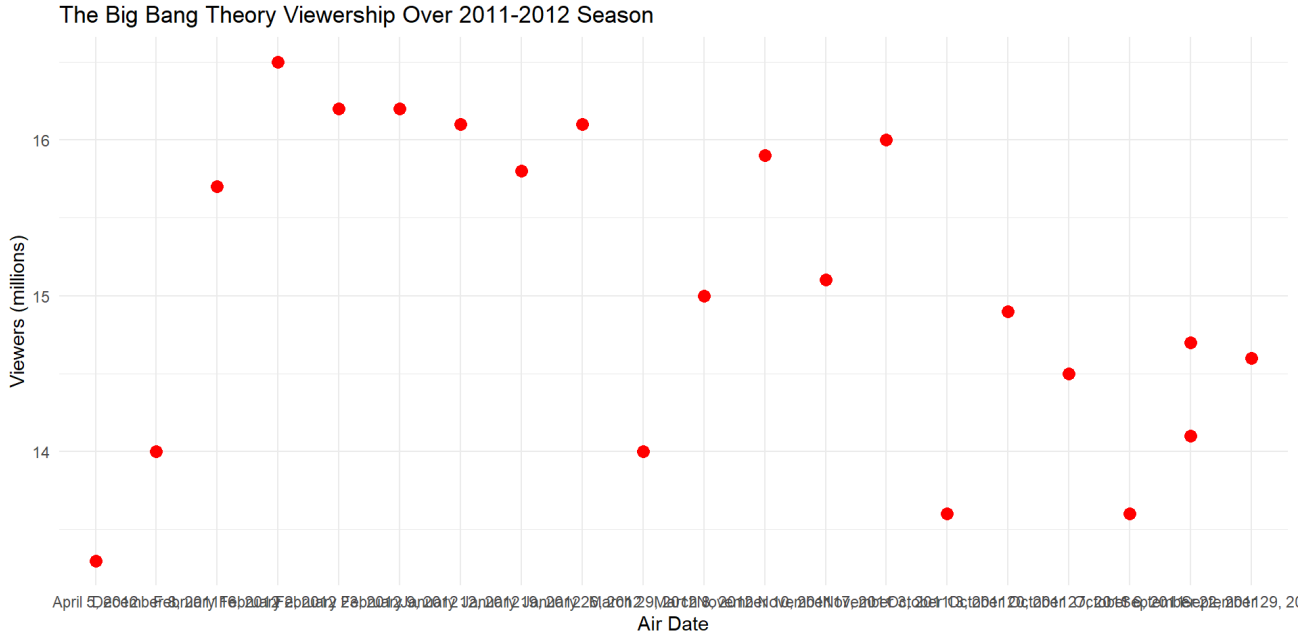
```
## 观众人数的众数是： 13.6 百万
```

- 计算第一和第三四分位数。

```
## 第一四分位数（Q1）是： 14.1 百万
```

```
## 第三四分位数（Q3）是： 16 百万
```

- 2011-2012季度的观众人数是增长还是下降了？ 从图中可以看出来，整体上是下降了



4 球员得分问题

- 展示频率分布。

```
##
## (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
```

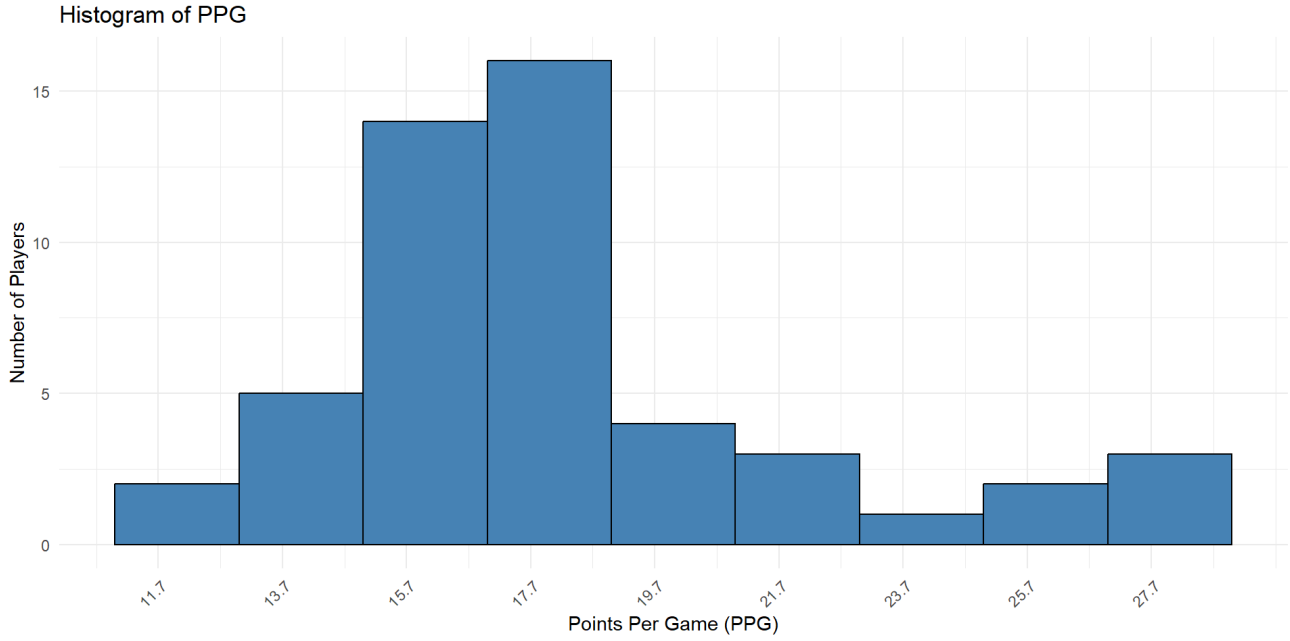
- 展示相对频率分布。

```
##
## (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
```

- 展示累积百分比频率分布。

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

- 为每场比赛平均得分开发一个直方图。



- 数据是否看起来有偏斜？请解释。 是有偏斜，原因可能是nba球员中极端优秀的是少数，然后又有很高门槛，所以集中在偏左的区域
- 有多少百分比的球员平均每场比赛得分至少为20分？

```
## [1] 22
```

5 问题3：一位研究人员通过报告调查结果来说明平均数的标准误差是20。总体标准差是500

- 这项调查中使用的样本有多大？

```
## The sample size used in this survey was:  625
```

- 点估计在总体均值±25范围内的概率是多少？

```
## The probability that the point estimate was within ±25 of the population mean is:  0.2112995
```

6 问题4 《年轻专业人士》杂志受众分析

- 开发适当的描述性统计数据来总结数据。

```
##  [1] "Age"                      "Gender"
##  [3] "Real.Estate.Purchases."   "Value.of.Investments...."
##  [5] "Number.of.Transactions"   "Broadband.Access."
##  [7] "Household.Income...."     "Have.Children."
##  [9] "X"                        "X.1"
## [11] "X.2"                      "X.3"
## [13] "X.4"                      "X.5"
```

```
##      Mean Median       SD Min Max
## 1 30.1122     30 4.024023  19  42
```

```
##       Mean Median       SD   Min    Max
## 1 74459.51  66050 34818.21 16200 322500
```

```
## Mode of Age:  32
```

```
## Mode of Household Income:  60300
```

- 为订阅者的平均年龄和家庭收入开发95%置信区间。

```
## 95% CI for Mean Age: [ 29.72268 ,  30.50171 ]
```

```
## 95% CI for Mean Household Income: [ 71089.2 ,  77829.83 ]
```

- 为拥有家庭宽带接入的订阅者比例和有孩子的订阅者比例开发95%置信区间。

```
## Proportion of subscribers with broadband access:  0.6243902
```

```
## 95% CI for broadband access: [ 0.5774559 ,  0.6713246 ]
```

```
## Proportion of subscribers with children:  0.5341463
```

```
## 95% CI for having children: [ 0.4858016 ,  0.5824911 ]
```

-《年轻专业人士》会是在线经纪人的好广告渠道吗？用统计数据证明你的结论。

为了评估《年轻专业人士》杂志是否是在线经纪人的好广告渠道，我们需要考虑订阅者的特征，特别是那些可能对在线经纪服务感兴趣的特征。以下是一些可能相关的特征：

家庭收入：高收入家庭可能更有可能投资和使用在线经纪服务。 金融投资价值：拥有较高金融投资价值的订阅者可能更可能对在线经纪服务感兴趣。 股票/债券/共同基金交易数量：频繁交易的订阅者可能更倾向于使用在线经纪服务。 宽带接入：拥有宽带接入的订阅者更可能更可能使用在线服务。

```
## Proportion of subscribers potentially interested in online brokerage services:  0.07073171
```

```
## 95% CI for the proportion: [ 0.04588482 ,  0.0955786 ]
```

- 这本杂志会是为销售教育软件和针对幼儿的电脑游戏的公司做广告的好地方吗？ 为了评估《年轻专业人士》杂志是否适合为销售教育软件和针对幼儿的电脑游戏的公司做广告，我们需要考虑订阅者的家庭状况，特别是他们是否有孩子。以下是使用R语言进行评估的步骤：

确定有孩子的订阅者比例。 计算这个比例的95%置信区间。

```
## Proportion of subscribers with children: 0.5341463
```

```
## 95% CI for the proportion of subscribers with children: [ 0.4858016 ,  0.5824911 ]
```

- 评论您认为《年轻专业人士》的读者会感兴趣的文章类型。 根据《年轻专业人士》杂志的目标受众——近期大学毕业生和职业生涯初期的专业人士，我们可以推断出一些可能吸引他们的文章类型,比如职业发展与规划等。

7 问题5： 质量伙伴公司

- 对每个样本在0.01的显著性水平上进行假设检验，并确定是否应采取任何措施。为每个测试提供p值。

```
## # A tibble: 4 × 5
##   Sample   Sample.Mean Sample.Size t.Value p.Value
##   <chr>          <dbl>       <int>   <dbl>   <dbl>
## 1 Sample.1        12.0          30  -1.03  0.313
## 2 Sample.2        12.0          30   0.713 0.482
## 3 Sample.3        11.9          30  -2.93  0.00647
## 4 Sample.4        12.1          30   2.16  0.0391
```

```
## # A tibble: 4 × 6
##   Sample   Sample.Mean Sample.Size t.Value p.Value Corrective.Action
##   <chr>          <dbl>       <int>   <dbl>   <dbl> <chr>
## 1 Sample.1        12.0          30  -1.03  0.313   No Action Needed
## 2 Sample.2        12.0          30   0.713 0.482   No Action Needed
## 3 Sample.3        11.9          30  -2.93  0.00647 Take Action
## 4 Sample.4        12.1          30   2.16  0.0391  No Action Needed
```

- 计算四个样本的每个标准差。假设总体标准差为0.21是否合理?

```
## # A tibble: 4 × 2
##   Sample   Sample.SD
##   <chr>        <dbl>
## 1 Sample.1     0.220
## 2 Sample.2     0.220
## 3 Sample.3     0.207
## 4 Sample.4     0.206
```

```
## Assumed Population SD: 0.21
```

```
## # A tibble: 4 × 2
##   Sample   Sample.SD
##   <chr>        <dbl>
## 1 Sample.1     0.220
## 2 Sample.2     0.220
## 3 Sample.3     0.207
## 4 Sample.4     0.206
```

- 计算样本均值x围绕μ = 12的界限，以便只要新样本均值在这些界限内，流程就被认为是运行令人满意的。如果x超过上限或x低于下限，将采取纠正措施。这些界限被称为质量控制的上下限和下控制限。

```
## Upper Control Limit (UCL): 12.10667
```

```
## Lower Control Limit (LCL): 11.89333
```

- 讨论将显著性水平更改为较大值的含义。如果显著性水平增加，可能会增加哪种错误或错误?

在统计假设检验中，显著性水平（也称为α水平或第一类错误的概率）是指拒绝零假设（H0）时出错的概率。显著性水平通常设定为0.05、0.01或0.10，这取决于研究领域和研究者对错误容忍度的偏好。

将显著性水平更改为较大值的含义包括:

增加第一类错误的风险：第一类错误是指错误地拒绝了真实的零假设。当显著性水平增加时，我们更有可能犯这种错误。例如，如果我们将显著性水平从0.05增加到0.2，我们拒绝零假设的概率增加，即使零假设实际上是真的。

降低检验的统计功效：统计功效是指正确拒绝错误的零假设（H1）的概率。显著性水平增加，统计功效降低，意味着我们更有可能未能检测到实际存在的效果或差异。

增加第二类错误的风险：第二类错误是指未能拒绝一个假的零假设。虽然增加显著性水平并不直接影响第二类错误，但通常为了保持一定的统计功效，我们可能会减少样本大小或增加效应大小，这间接增加了第二类错误的风险。

8 问题6:度假房产已出租和未出租的单元情况

- 估算 2007 年 3 月第一周和 2008 年 3 月第一周期间已出租单元的比例。

```
## 2007年3月第一周出租单位的比例为： 0.3482587
```

```
## 2008年3月第一周出租单位的比例为： 0.3482587
```

- 给出这两个比例差异的 95% 置信区间。

```
## 2007年和2008年出租单位比例差异的95%置信区间为： -0.09314374 0.09314374
```

- 根据你的研究结果，2008 年 3 月的出租率看起来是否会比一年前（即 2007 年 3 月）有所上升?

```
## 没有足够的证据表明2008年3月的出租率与2007年3月有显著差异。
```

9 问题7:空军训练计划

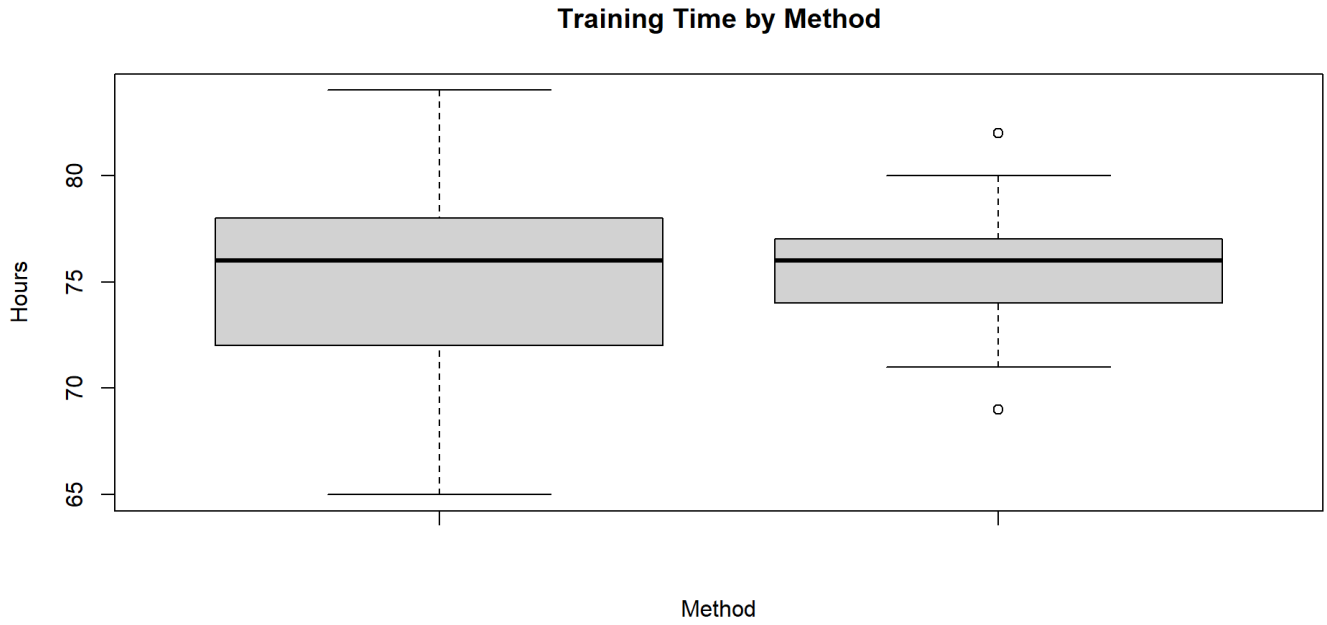
- 使用适当的描述性统计来总结每种方法的训练时间数据。你从样本数据中观察到什么相似之处或差异?

```
## Current Method Descriptive Statistics:
```

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

```
##
## Proposed Method Descriptive Statistics:
```

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



- 对两种方法的总体均值之间的差异发表评论。讨论你的发现。 均值差不多

```
## Current Method Mean Training Time:  NA  hours
```

```
## Proposed Method Mean Training Time:  NA  hours
```

```
##
## T-test results for comparing means:
```

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

- 计算每种训练方法的标准差和方差。对两种训练方法的总体方差是否相等进行假设检验。讨论你的发现。 第一种标准差和方差大，第二种标准差和方差小。所以第二种训练方式更稳定。

```
## Current Method Standard Deviation:  NA
```

```
## Current Method Variance:  NA
```

```
## Proposed Method Standard Deviation:  NA
```

```
## Proposed Method Variance:  NA
```

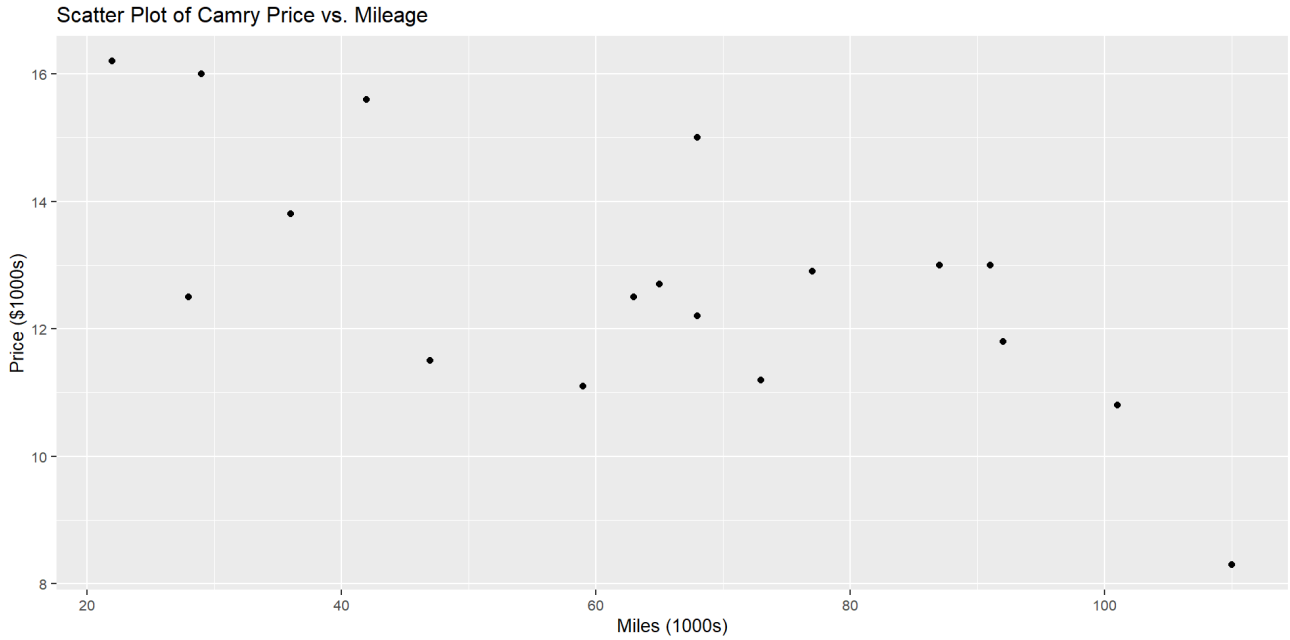
```
##
## F-test results for comparing variances:
```

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

- 你能否得出关于两种方法之间差异的结论？你的建议是什么？解释。 第一种方法能够训练出特别优秀的，但是效果不稳定。第二种方法成果比较稳定。 我建议取长补短，选用第二种方法为主，但借鉴第一种方法优化成绩的点。
- 在做出未来要使用的培训计划的最终决定之前，你能建议其他数据或测试吗？ 最好还有训练效率的数据。

10 问题8：2007年款凯美瑞的汽车里程与销售价格之间的关系

- 绘制一个散点图，将汽车里程放在水平轴上，价格放在垂直轴上。



- 在部分(a)中开发的散点图表明了这两个变量之间的关系是什么？

两个变量之间似乎存在一种可以通过直线近似的负相关关系。也可以认为这种关系可能是曲线的，因为到了某个时候，汽车的里程数太多，其价值会变得非常小。

- 开发一个估计的回归方程，用于预测给定里程（千英里）的价格（千美元）。

```
## Estimated Regression Equation:
```

```
## Price =  16.46976  +  -0.05877393  * Miles
```

```
##
## Call:
## lm(formula = Price ~ Miles, data = data)
##
## 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
##   (因为不存在，1个观察量被删除了)
## 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的显著性水平下测试关系的显著性。

```
## P-value for the slope: 0.000347511
```

```
## The relationship between miles and price is significant at the 0.05 level.
```

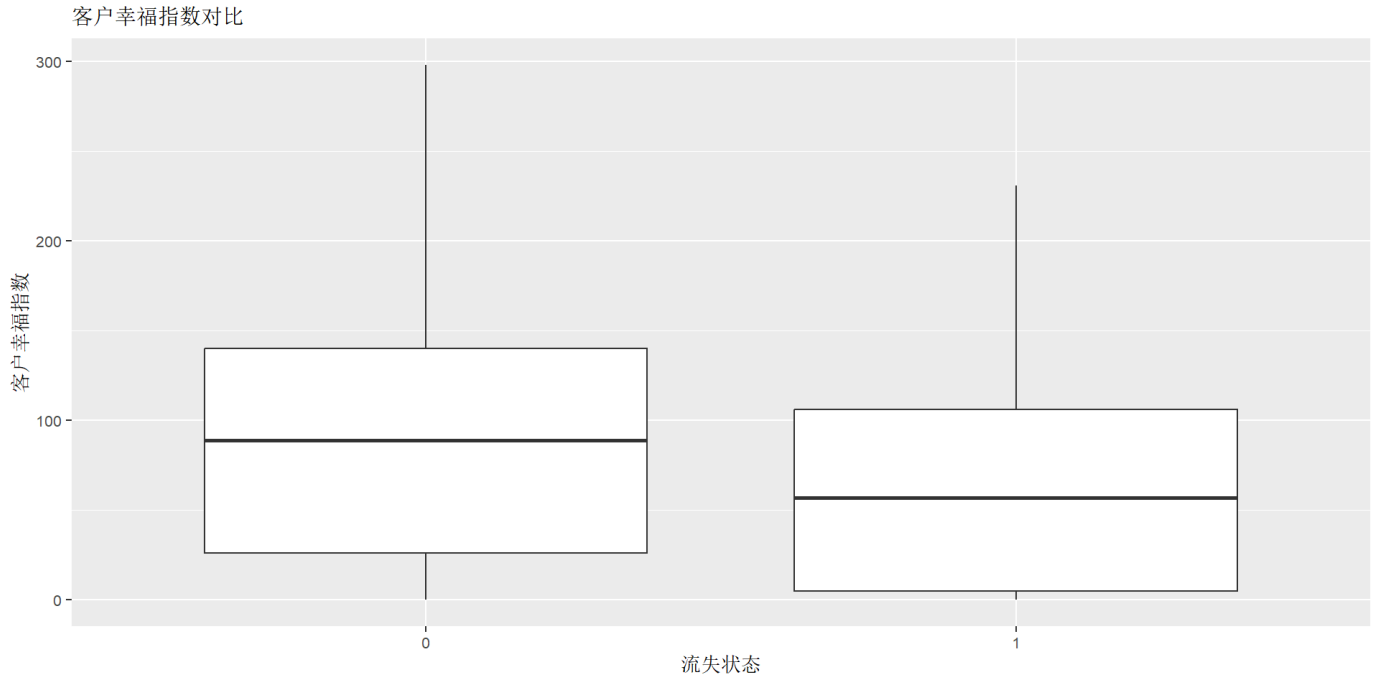
- 估计的回归方程是否提供了良好的拟合？解释。 没有很好的拟合，因为有其他更重要的影响因素，比如原车价。
- 解释估计回归方程的斜率。 表示年份的折价率。
- 假设你正在考虑购买一辆已经行驶了60,000英里的2007年款二手车凯美瑞。使用部分(c)中开发的估计回归方程，预测这辆车的价格。这是你会向卖家提供的价格吗？ 价格太低了，不是我愿意的价格。

```
## Predicted price for a 2007 Camry with 60,000 miles: $ 12.94332  (in thousands)
```

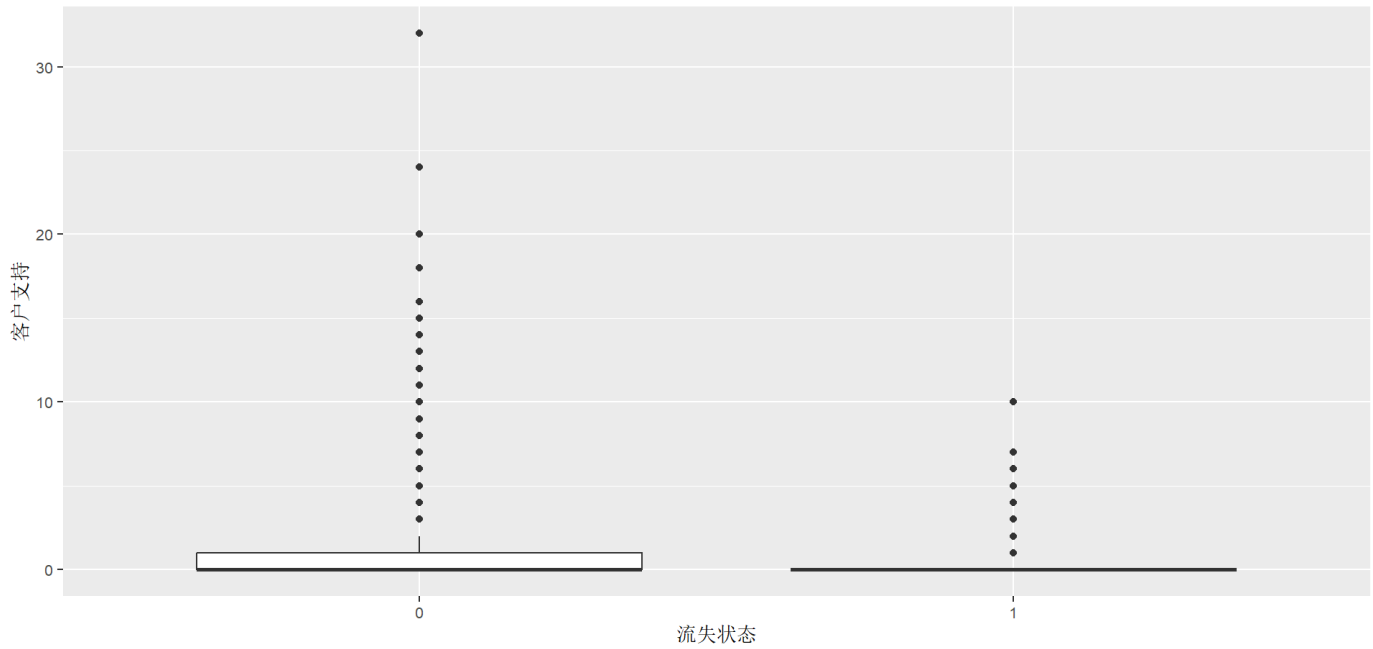
11 问题9：服务商的客户流失数据

- 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？

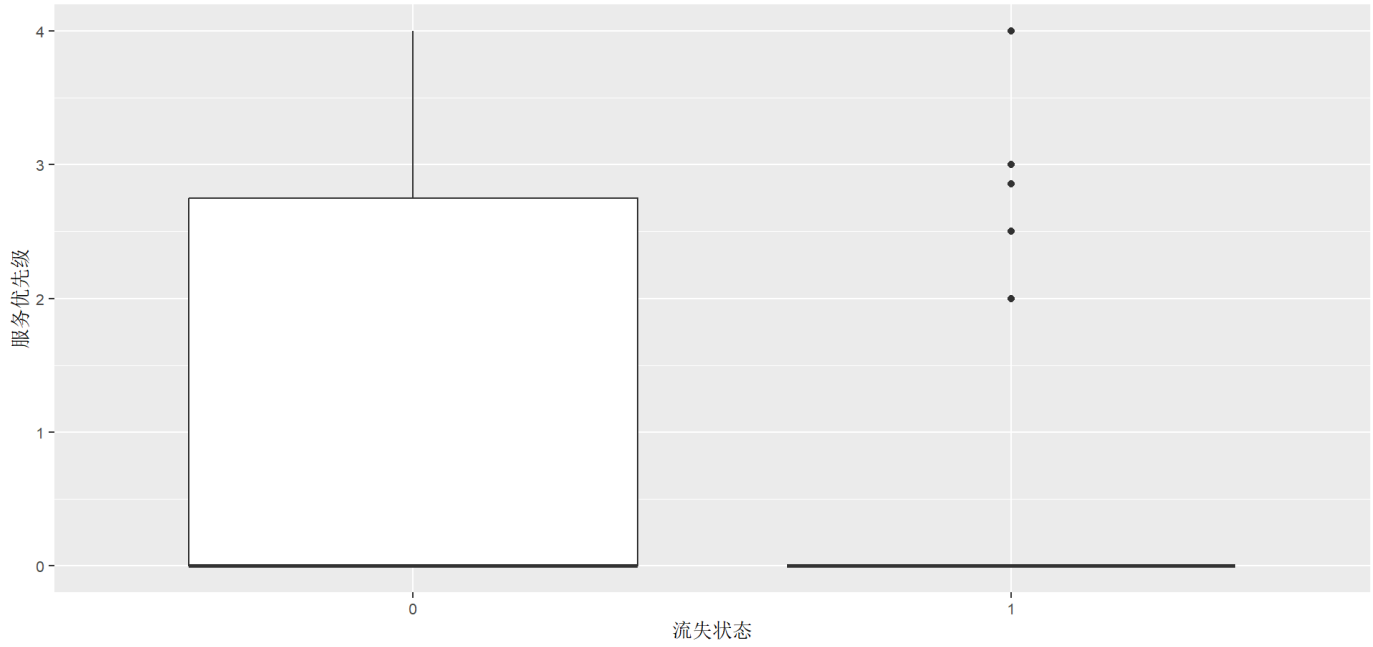
```
##       客户ID          流失          当月客户幸福指数 客户幸福指数相比上月变化
##  Min.   :   1   Min.   :0.00000   Min.   :  0.00   Min.   :-125.000
##  1st Qu.:1588   1st Qu.:0.00000   1st Qu.: 24.50   1st Qu.:  -8.000
##  Median :3174   Median :0.00000   Median : 87.00   Median :   0.000
##  Mean   :3174   Mean   :0.05089   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
```



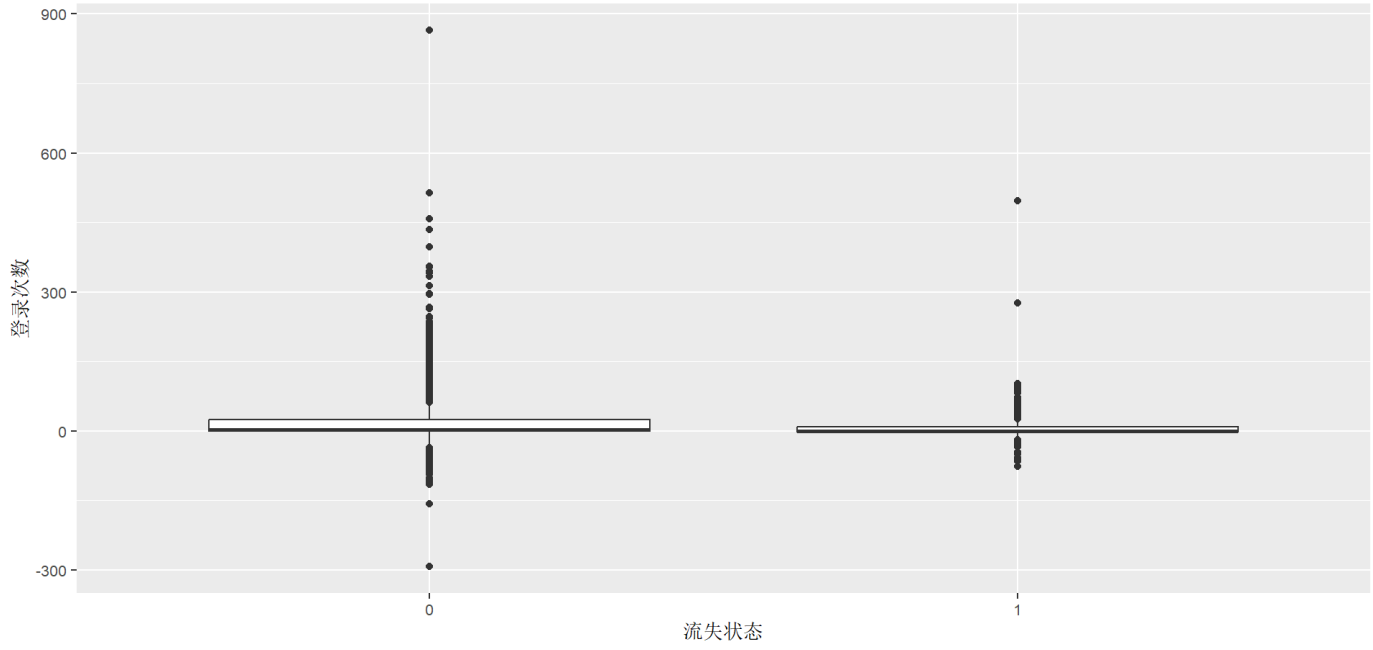
客户支持对比

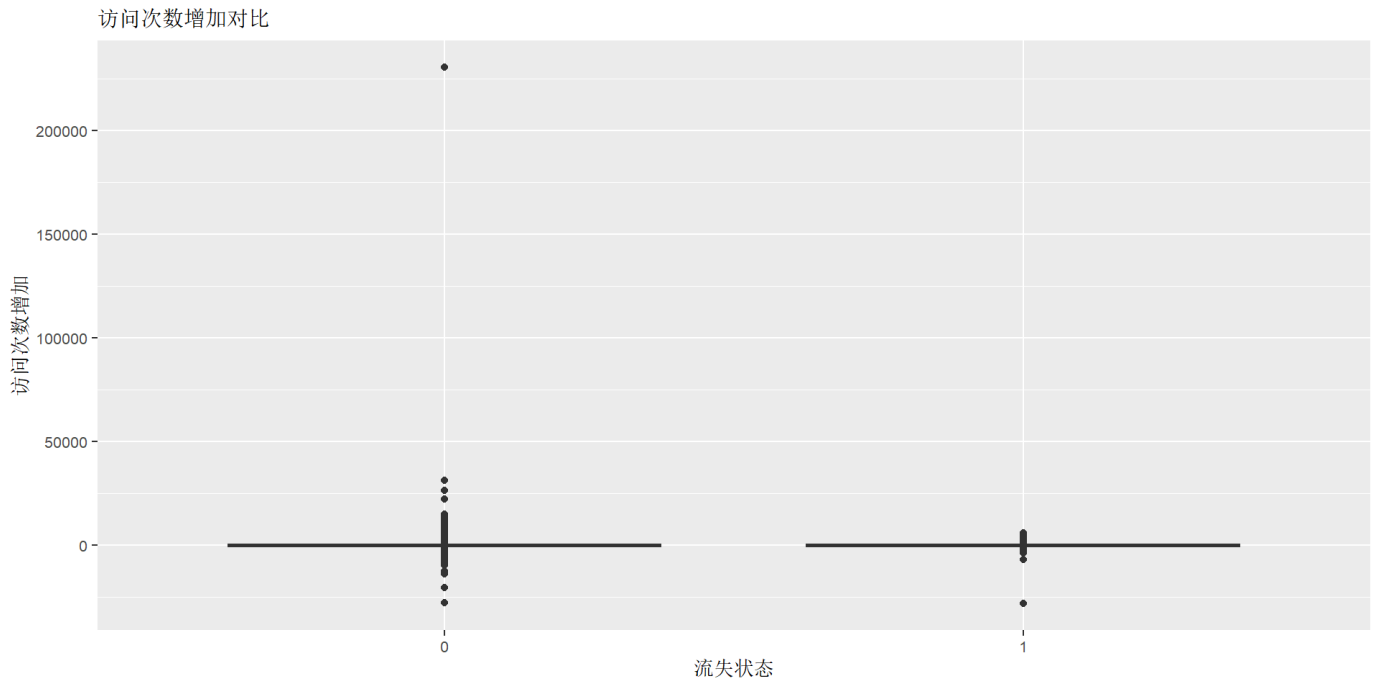


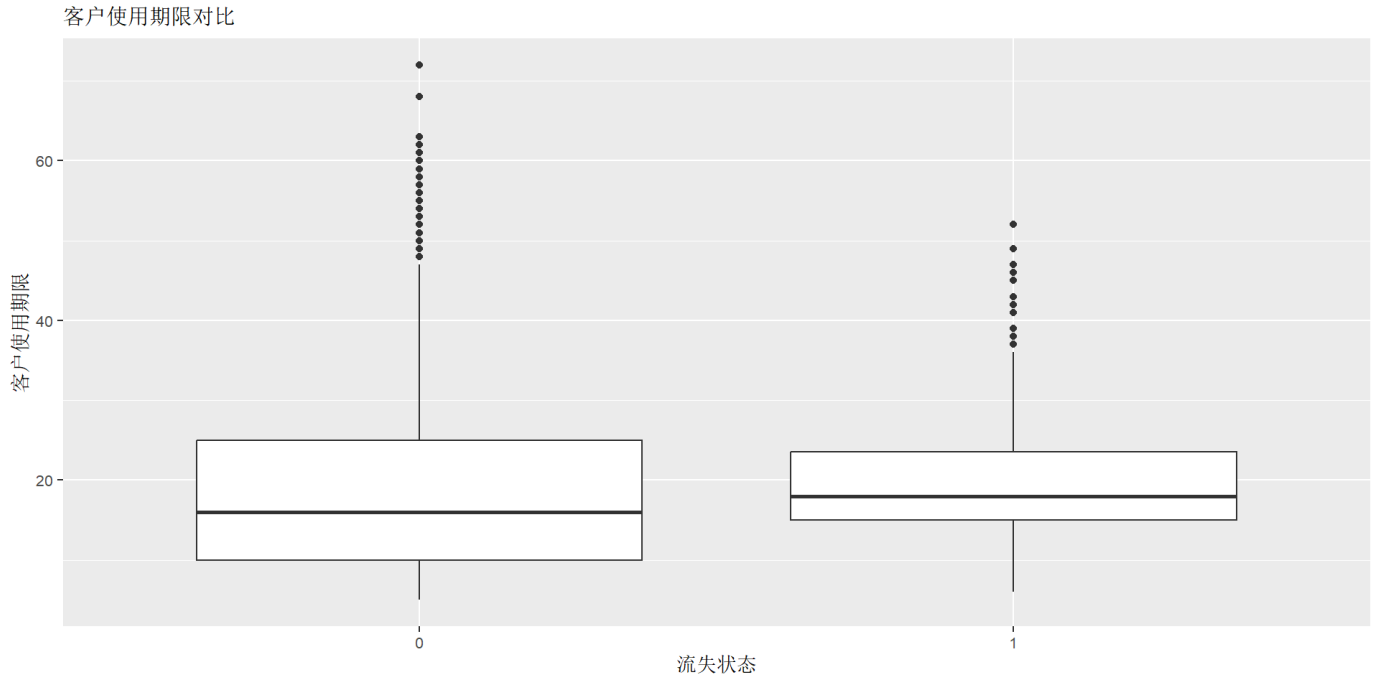
服务优先级对比

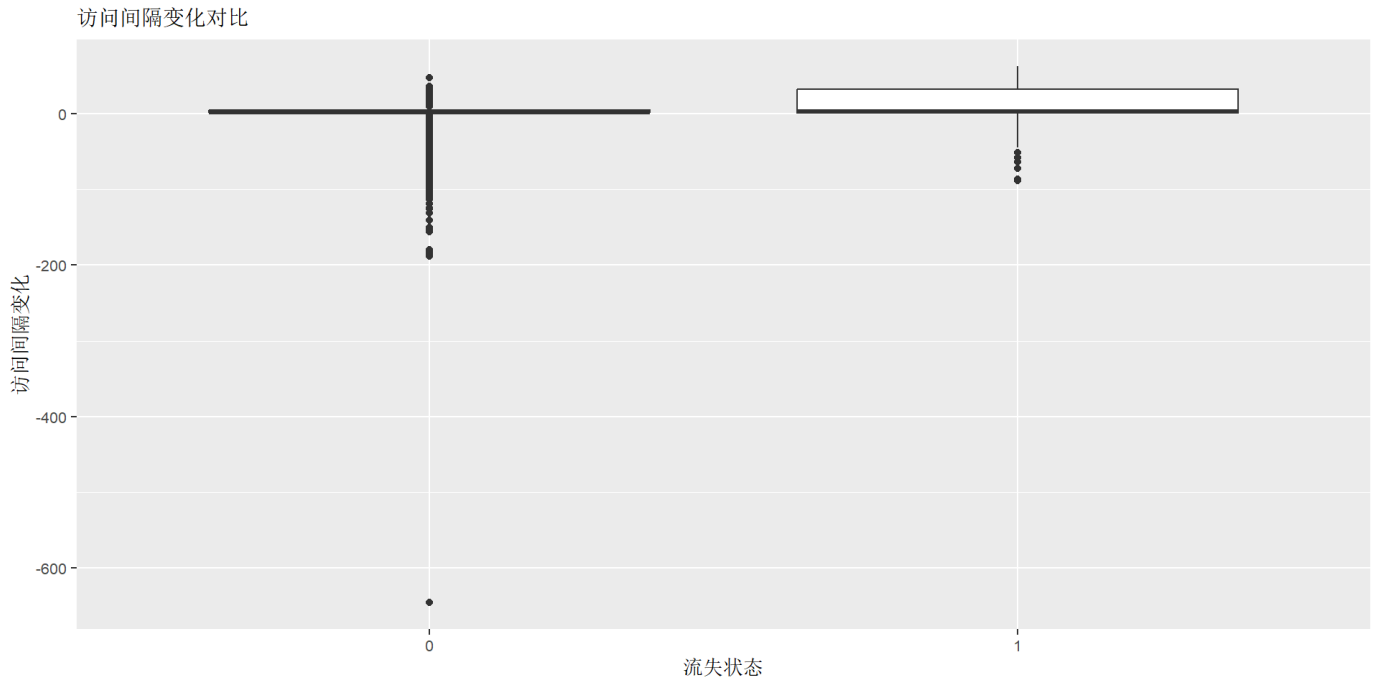


登录次数对比




访问次数增加对比
访问次数增加
200000
150000
100000
50000
0
0
1
流失状态


客户使用期限对比
客户使用期限
60
40
20
0
1
流失状态


访问间隔变化对比
访问间隔变化
0
-200
-400
-600
0
1
流失状态

- 通过均值比较的方式验证上述不同是否显著。

```
## $当月客户幸福指数
## $当月客户幸福指数$mean1
## [1] 88.60591
##
## $当月客户幸福指数$mean2
## [1] 63.27245
##
## $当月客户幸福指数$p_value
## [1] 2.041576e-11
##
## $当月客户幸福指数$diff
## [1] 25.33346
##
##
## $当月客户支持
## $当月客户支持$mean1
## [1] 0.7242696
##
## $当月客户支持$mean2
## [1] 0.371517
##
## $当月客户支持$p_value
## [1] 0.0003383435
##
## $当月客户支持$diff
## [1] 0.3527526
##
##
## $当月服务优先级
## $当月服务优先级$mean1
## [1] 0.8295759
##
## $当月服务优先级$mean2
## [1] 0.4995577
##
## $当月服务优先级$p_value
## [1] 1.194977e-05
##
## $当月服务优先级$diff
## [1] 0.3300182
##
##
## $当月登录次数
## $当月登录次数$mean1
## [1] 16.13894
##
## $当月登录次数$mean2
## [1] 8.06192
##
## $当月登录次数$p_value
## [1] 0.000783041
```

```
##
## $当月登录次数$diff
## [1] 8.077025
##
##
## $访问次数相比上月的增加
## $访问次数相比上月的增加$mean1
## [1] 106.6096
##
## $访问次数相比上月的增加$mean2
## [1] -95.7678
##
## $访问次数相比上月的增加$p_value
## [1] 0.2610323
##
## $访问次数相比上月的增加$diff
## [1] 202.3774
##
##
## $客户使用期限
## $客户使用期限$mean1
## [1] 18.81873
##
## $客户使用期限$mean2
## [1] 20.35294
##
## $客户使用期限$p_value
## [1] 0.01607184
##
## $客户使用期限$diff
## [1] -1.534216
##
##
## $访问间隔变化
## $访问间隔变化$mean1
## [1] 3.511454
##
## $访问间隔变化$mean2
## [1] 8.486068
##
## $访问间隔变化$p_value
## [1] 1.22239e-06
##
## $访问间隔变化$diff
## [1] -4.974614
```

- 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 = 流失 ~ 当月客户幸福指数 + 客户使用期限,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.7989543  0.1130826 -24.751  < 2e-16 ***
## 当月客户幸福指数 -0.0075660  0.0009984  -7.578 3.50e-14 ***
## 客户使用期限      0.0227618  0.0046969   4.846 1.26e-06 ***
## ---
## 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: 2485.0  on 6344  degrees of freedom
## AIC: 2491
##
## Number of Fisher Scoring iterations: 6
```

- 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   57   51   55    2   59  121   61   89   76   95
##  [16]  110  137   62   68  154    5   12   42   69  119  146  171  183  190   75
##  [31]  101  123  109 1392  141  142 1393  106 1419 1438   30  194   16   84   64
##  [46] 1395 1478 1520 2235 2240 2255  203 1459 1462 2245 1496  112 1108 1143  158
##  [61] 1893 1908  128 2286  147 1051    9   10   72  130  163 1474 1951 1971   17
##  [76] 2244   63  127   47  113 1091 1141 2047 2062 2070  169 1383  104 1953 1378
##  [91] 2080  133  179  192 1446 2281 2306 1110   41   49
```