第二次作业-solution

霍超　　2024281050968　　2024-11-30

第一题：生活大爆炸观众数据分析

a."最小观众数量：　13.3"；"最大观众数量：　16.5"

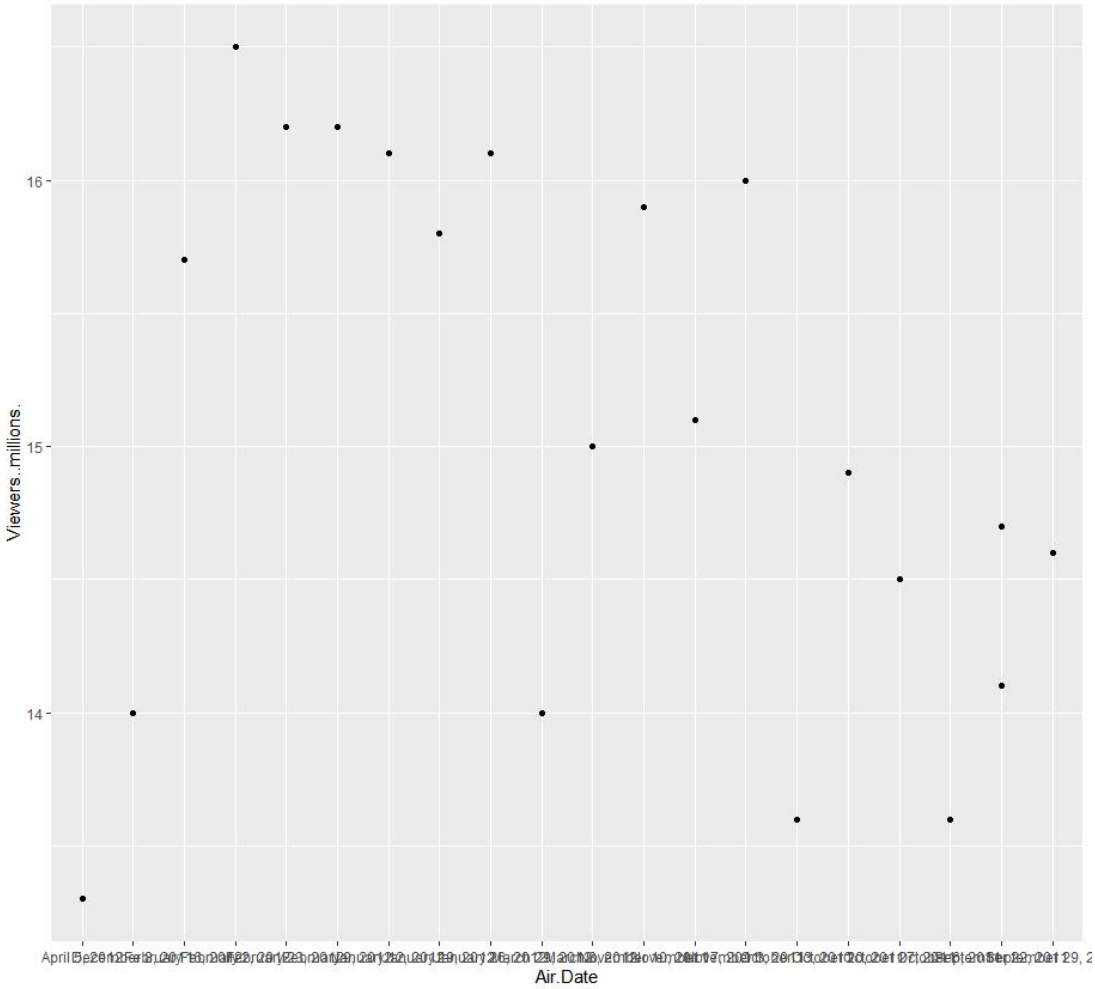
b. "均值（mean）: 15.0428571428571"

"中位数（median）: 15"

"众数（mode）: 13.6"

c.第一个四分位数：14.1；第三个四分位数：16

d.



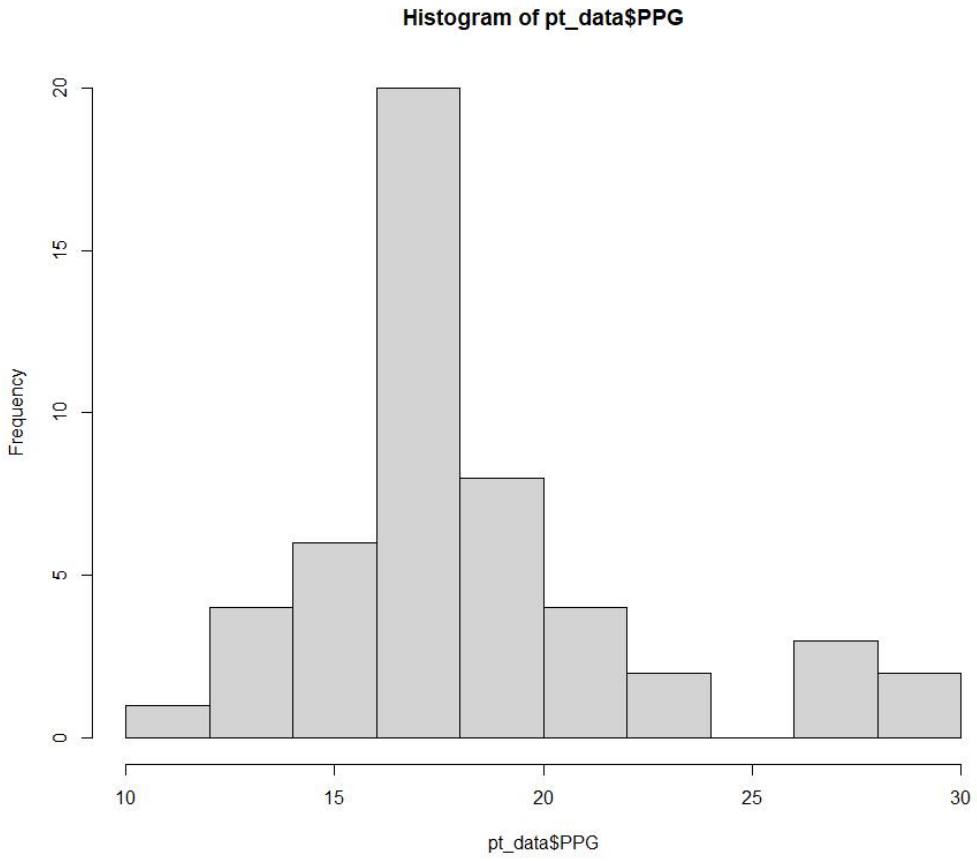
根据图像看出，观众数量波动较大，有不太明显的下降趋势

第二题：NBA 球员得分

a.(10,15] (15,20] (20,25] (25,30]

7　　32　　6　　5

直方图如下：



数据偏态系数为 1.12＞0，且从直方图观察，数据为正偏
PPG 至少为 20 的球员比例：22%

第三题：计算
a.n=625
b.0.79

第四题：青年专业杂志
a.描述性统计
tibble [410 × 10] (S3: tbl_df/tbl/data.frame)
 $ Age : num [1:410] 38 30 41 28 31 32 32 26 26 34 ...
 $ Gender : chr [1:410] "Female" "Male" "Female" "Female" ...
 $ Real Estate Purchases? : chr [1:410] "No" "No" "No" "Yes" ...
 $ Value of Investments ($): num [1:410] 12200 12400 26800 19600 15100 39700 21900 41900 16100 18400 ...
 $ Number of Transactions : num [1:410] 4 4 5 6 5 3 2 2 4 11 ...
 $ Broadband Access? : chr [1:410] "Yes" "Yes" "Yes" "No" ...
 $ Household Income ($) : num [1:410] 75200 70300 48200 95300 73300 ...
 $ Have Children? : chr [1:410] "Yes" "Yes" "No" "No" ...
 $ ...9 : logi [1:410] NA NA NA NA NA NA ...
 $ ...10 : chr [1:410] NA NA NA NA ...

```
      Age                Gender          Real Estate Purchases? Value of Investments ($)
 Min.   :19.00   Length:410         Length:410         Min.   :     0
 1st Qu.:28.00   Class :character   Class :character   1st Qu.: 18300
 Median :30.00   Mode  :character   Mode  :character   Median : 24800
 Mean   :30.11                                         Mean   : 28538
 3rd Qu.:33.00                                         3rd Qu.: 34275
 Max.   :42.00                                         Max.   :133400
 Number of Transactions Broadband Access?  Household Income ($) Have Children?
 Min.   : 0.000         Length:410         Min.   : 16200      Length:410
 1st Qu.: 4.000         Class :character   1st Qu.: 51625      Class :character
 Median : 6.000         Mode  :character   Median : 66050      Mode  :character
 Mean   : 5.973                            Mean   : 74460
 3rd Qu.: 7.000                            3rd Qu.: 88775
 Max.   :21.000                            Max.   :322500
    ...9                ...10
 Mode:logical   Length:410
 NA's:410       Class :character
                Mode  :character
```

数据包括了 8 个变量，年龄 Q1 分位数是 28，均值 30，Q3 分位数 33，最大年龄 42 岁

金融投资 Q1 分位数是 18300，均值 28538 大于中位数 24800，Q3 分位数 34275，呈右偏分布

家庭收入 Q1 分位数是 51625，均值 74460，Q3 分位数 88775，呈右偏分布

b.年龄 95%置信区间上限：30.5，下限：29.7
收入 95%置信区间上限：7784，下限：7108
c.有宽带客户 95%置信区间上限：0.67，下限：0.58
有孩子客户 95%置信区间上限：0.58，下限：0.49
d.是的。Young Professional 杂志是一个好广告渠道。理由如下：
1. 62.4%的杂志订阅客户家里有宽带，可以进行线上交易。有宽带客户 95%置信区间是（0.58,0.67）
2. 订阅客户家庭金融投资均值 28538 美元，家庭收入均值 74460 美元，Q3 分位数 88775 美元
这些条件对于在线经纪人来说，是一个很好的目标客户
e.是的。理由如下：
1. 杂志订阅客户平均年龄为 30 岁，比较年轻，95%的置信区间是（29.7,30.5）
2.超过 53.4%的订阅客户有孩子，有孩子 95%的置信区间是（0.49,0.58）
3.订阅者家庭平均收入 74460 美元，收入较高，属于高质量群体，有付费能力
因此这本杂志是为销售教育软件和幼儿电脑游戏的公司做广告的好地方

f.该部分读者特点为：年轻（平均年龄 30 岁），家庭平均收入较高，过半数有孩子，有投资经验，因此推测可能对科技、投资、育儿、旅游、金融等文章感兴趣

第五题：质量助理公司

a.t1=2.608,p1=0.014 > alpha=0.01,不拒绝 H0，无需采取行动

b.sd_sample1=0.206

sd_sample2=0.175

sd_sample3=0.183

sd_sample4=0.190

仅有第一个样本的标准差接近 0.21，其他与总体标准差为 0.21 不一致，因此假设不太合理

c.在 95%的置信水平下，样本均值为 12 的置信区间上限=12.079　下限=11.921

d.常见的显著性水平包括 0.05（5%）和 0.01（1%）。当显著性水平增加时，意味着我们变得更容易拒绝零假设，即更容易做出错误的决定,增加 I 类错误（更容易错误的拒绝实际为真的假设）

第六题：

a.

2007 年 3 月第一周租用比例为 0.35

2008 年 3 月第一周租用比例为 0.44

b.95%的置信区间为：-0.092——0.092

c.是的，两者有差异，且 2008 年 3 月出租率大于 2007 年

第七题：

a.

Current:

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

Propsoed:

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

通过比较，常规训练和拟提议的训练，均值接近，中位数相同，其他分位数相差不大，整体区别不大

b.

```
      Welch Two Sample t-test

data:    sl_data$Current and sl_data$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

p-value=0.5481 远大于 0.05，不拒绝原假设，即在 95%的显著性水平上，两组无差异

c.
current: sd=3.945 方差=15.562
proposed:sd=2.506　方差=6.282

方差检验结果：

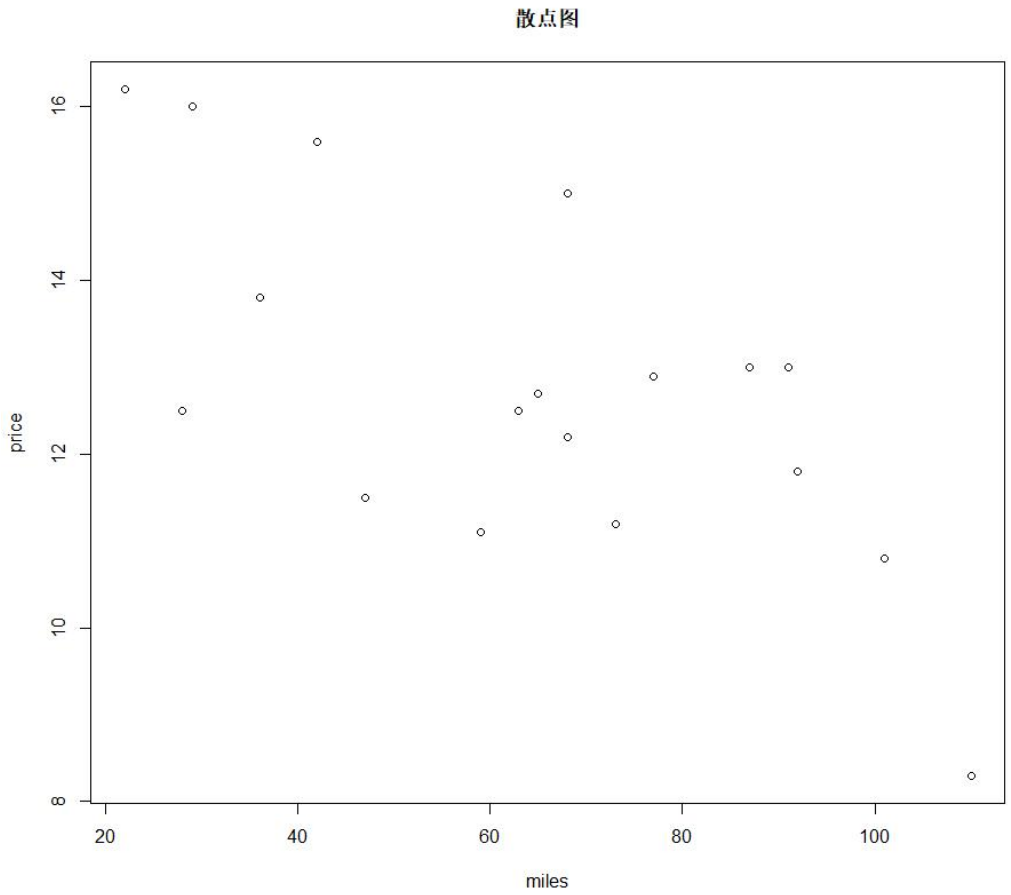
F test to compare two variances

data:　sl_data$Current and sl_data$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

根据方差检验 p=0.000578＜0.05，即在 95%的显著性水平下，两组标准差或方差有显著性差异

d.根据 t 检验数据，两种方法在训练平均时间上比较接近，差异的 95%的置信区间是-1.55 到 0.83 小时
根据方差检验数据，两种方法在标准差或方差上有明显的差异，差异的 95%置信区间为 1.49 到 4.13，也就是拟提议的训练方法具有更低的方差或标准差，此种训练对于不同的学生更稳定，不同的学生都有可能在用时接近的情况下完成培训，用时波动较小
综上所述，更推荐拟提议的训练方法
e.建议学生在不同的训练方法下统计通过考试的相关数据

第八题：
a.散点图如下



b.价格和里程接近负相关关系，即里程越大，价格约便宜
c.price= 16.47 + -0.06 * miles
d.结果如下：

```
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
```

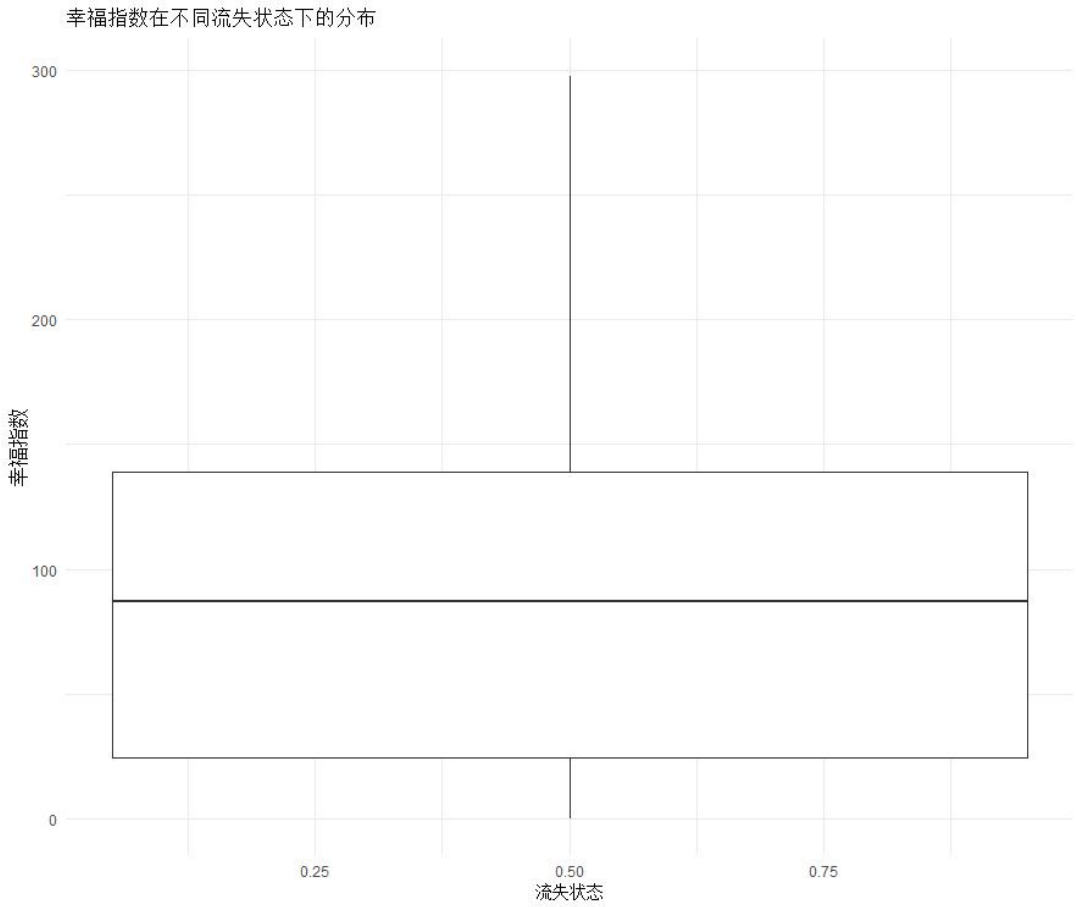
p=0.0003475 小于 0.05，即在 95%的显著性水平下 price 和 miles 有显著性关系

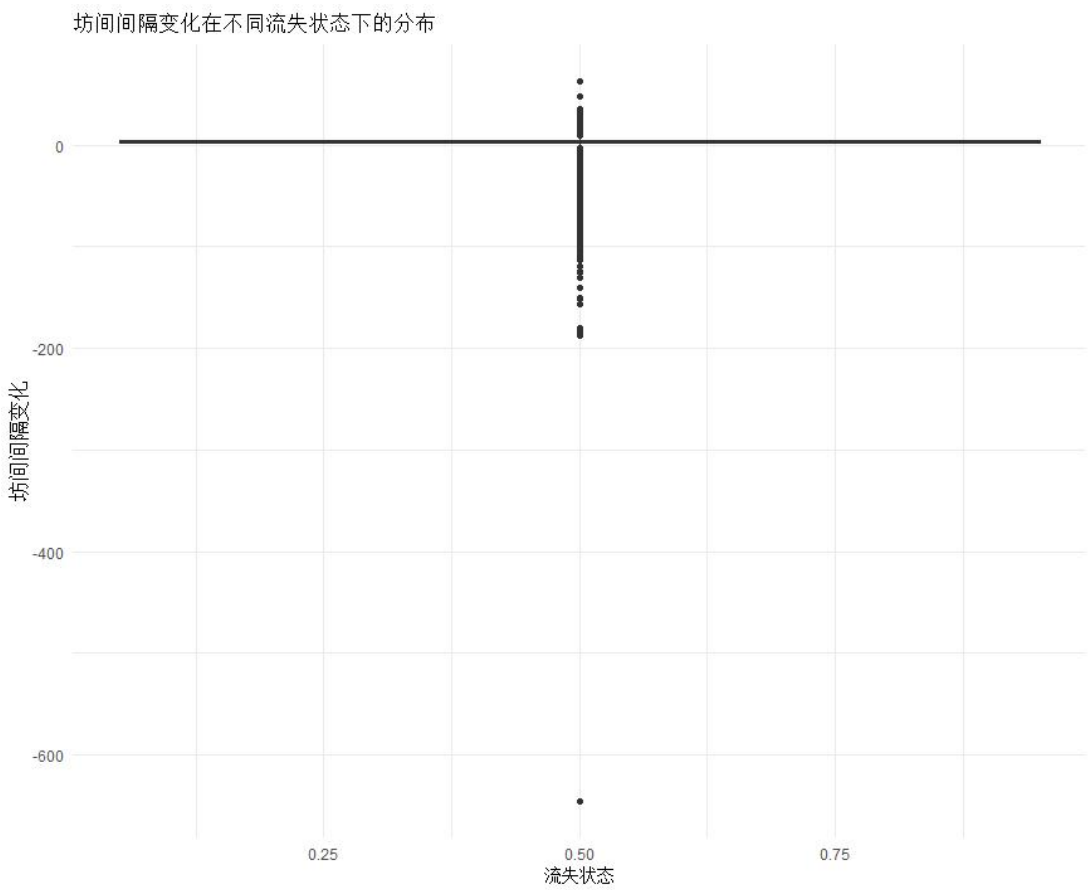
e.拟合一般，匹配水平为 51.15%
f.斜率指每增加 1000miles，价格下降 60 美元

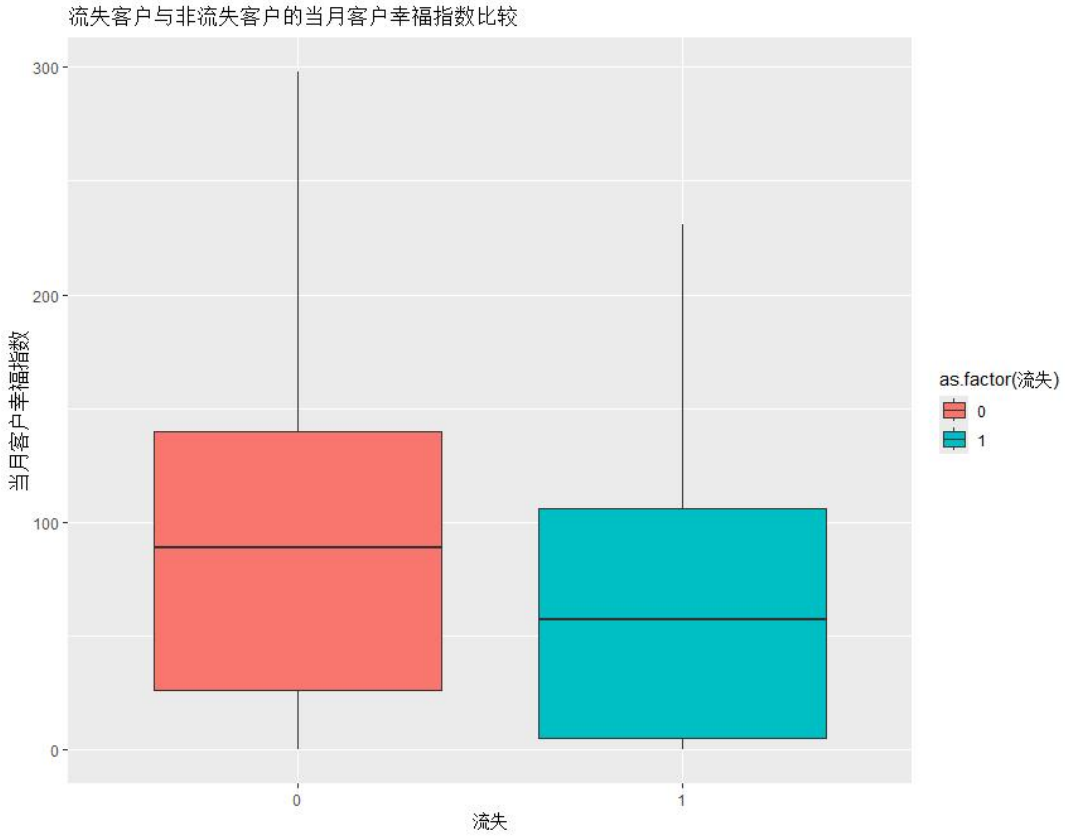
g.根据回归方程，当 miles=60（即 60000miles），price=12.87（千美元）。
不一定是给卖家的价格，因为该回归方程匹配水平只有 51%，准确度不是很高

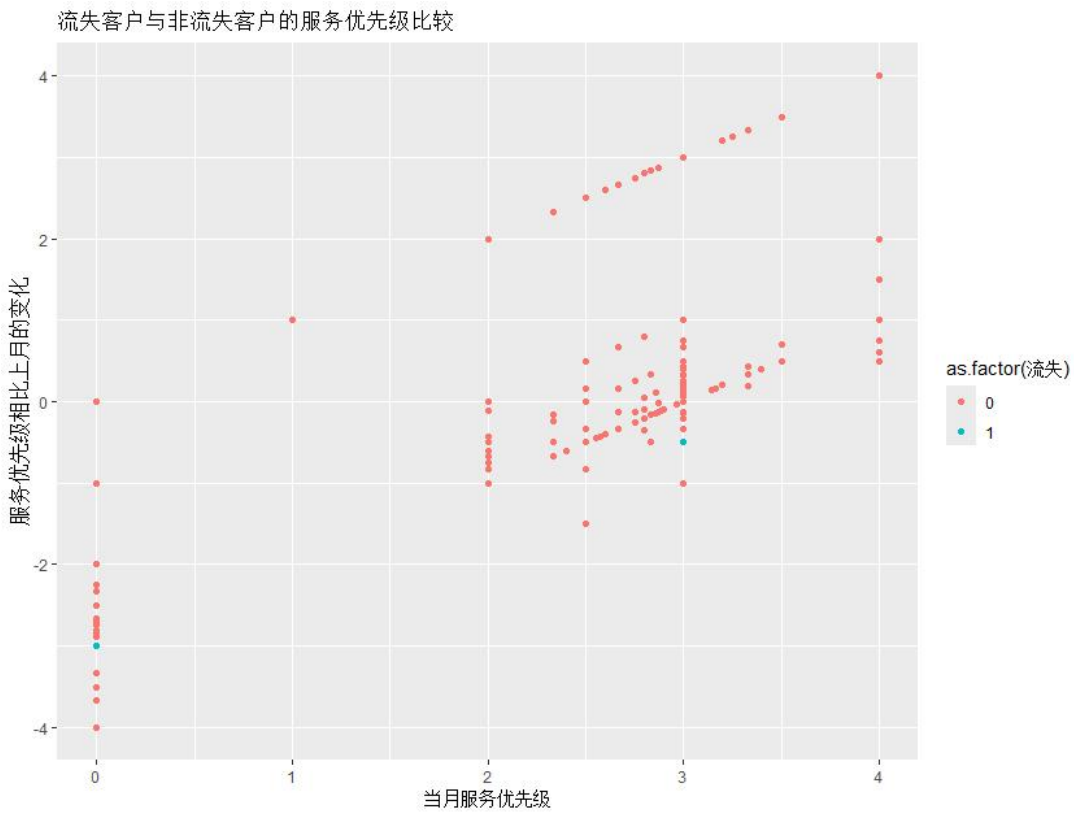
第九题：

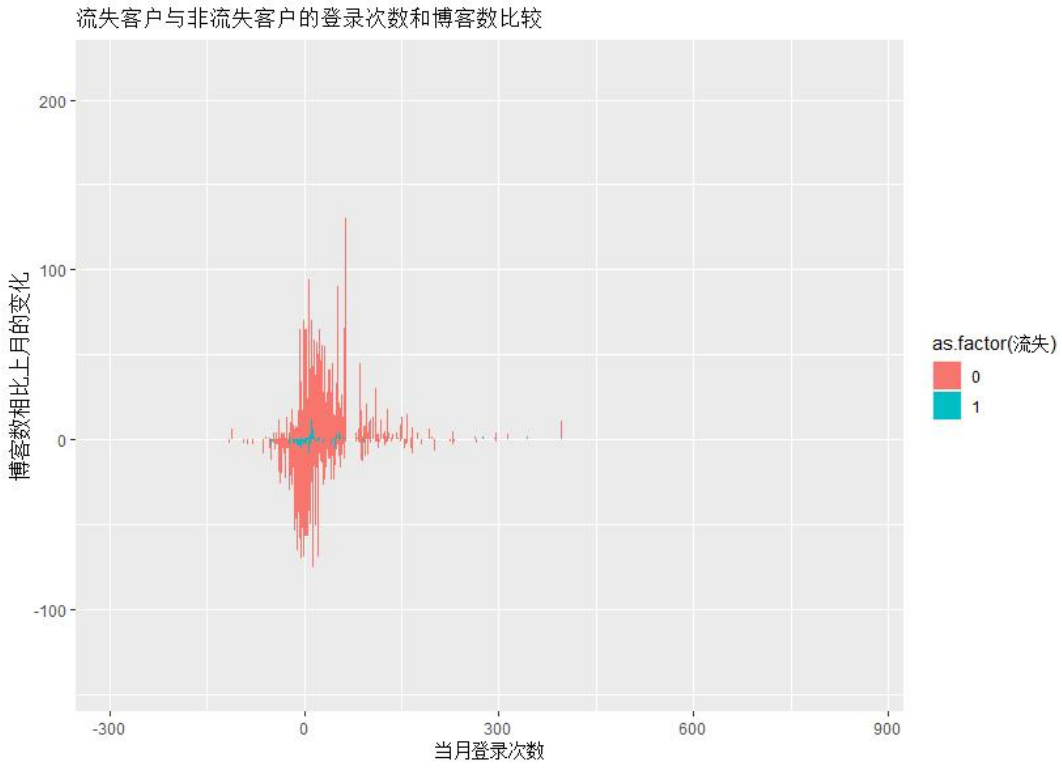
a.




坊间间隔变化在不同流失状态下的分布
0
-200
-400
-600
坊间间隔变化
0.25
0.50
0.75
流失状态


流失客户与非流失客户的当月客户幸福指数比较
300
200
100
0
当月客户幸福指数
0
1
流失
as.factor(流失)
0
1





b.基本指标：

当月客户幸福指数

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	24.50	87.00	87.32	139.00	298.00

当月客户支持

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.7063  1.0000 32.0000
```

当月服务优先级

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.8128  2.6667  4.0000
```

当月登录次数

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-293.00   -1.00    2.00   15.73   23.00  865.00
```

客户使用期限

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0    10.0    16.0    18.9    25.0    72.0
```

访问间隔变化

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-646.000     2.000     2.000     3.765     5.000    63.000
```

```
Call:
glm(formula = kh_data$流失 ~ kh_data$当月客户幸福指数 +
    kh_data$当月客户支持 + kh_data$当月服务优先级 +
    kh_data$当月登录次数 + kh_data$客户使用期限 +
    kh_data$访问间隔变化, family = binomial, data = kh_data)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.844721   0.118986 -23.908  < 2e-16 ***
kh_data$当月客户幸福指数    -0.005787   0.001148  -5.040 4.67e-07 ***
kh_data$当月客户支持        -0.029289   0.070882  -0.413 0.679462
kh_data$当月服务优先级      -0.046603   0.074001  -0.630 0.528845
kh_data$当月登录次数        -0.001789   0.002103  -0.851 0.394898
kh_data$客户使用期限         0.017484   0.005068   3.450 0.000561 ***
kh_data$访问间隔变化         0.013920   0.004017   3.465 0.000530 ***
---
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: 2467.7  on 6340  degrees of freedom
AIC: 2481.7
```

按上述模型的结果，对是否流失的显著变量有

1.当月客户幸福指数 及其相比上月变化与流失概率负相关,用户当月幸福度的提升会降低流失的概率

2.客户使用期限越大，用户流失的概率越高，表明老用户更易流失
3.用户访问间隔越大，用户流失的概率越高，用户越不活跃越易流失