第二次作业

郑艾轩

2024-11-16

目录

1 第一题 **3**

1.1 a. Compute the minimum and the maximum number of viewers. 3

1.2 b. Compute the mean, median, and mode 3

1.3 c. Compute the first and third quartiles 3

1.4 d. has viewership grown or declined over the 2011–2012 season? Discuss. 3

2 第二题 **4**

2.1 a. Show the frequency distribution. 4

2.2 b.Show the relative frequency distribution. 4

2.3 c. Show the Cumulative Percent Frequency Distribution 5

2.4 d. Develop a Histogram for the Average Number of Points Scored per Game 5

2.5 e. Do the data appear to be skewed? Explain. 5

3 第三题 **6**

3.1 a. How large was the sample used in this survey? 6

3.2 b. What is the probability that the point estimate was within ±25 of the population mean? 6

4 第四题 **6**

4.1 a. Develop appropriate descriptive statistics to summarize the data. 6

4.2 b. Develop 95% confidence intervals for the mean age and household income of subscribers. 7

4.3 c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children. 7

4.4 d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data. 8

4.5 e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children? 8

4.6 f. Comment on the types of articles you believe would be of interest to readers of Young Professional. 8

5 第五题 **9**

5.1 a. 对每个样本在 0.01 的显著性水平下进行假设检验 9

5.2 b. 计算四个样本的标准差 9

5.3 c. 计算控制限 9

5.4 d. 讨论将显著性水平更改为较大值的影响 9

6 第六题 **9**

6.1 a. 估计 2007 年 3 月第一周和 2008 年 3 月第一周已出租单元的比例 9

6.2 b. 为比例差异提供 95% 的置信区间 9

6.3 c. 根据发现判断 2008 年 3 月的租金率是否会比前一年有所上升 10

7 第七题 **10**

7.1 a. 使用恰当的描述性统计量来汇总每种教学方法的训练时间数据 10

7.2 b. 对两种教学方法的总体均值之间的差异进行评论 11

7.3 c. 计算每种教学方法的标准差和方差。针对两种教学方法的总体方差是否相等进行假设检验 12

7.4 d. 关于这两种教学方法之间的任何差异，你能得出什么结论？ 12

7.5 e. 建议其他可能需要的数据或测试 12

8 第八题 **13**

8.1 a. 绘制散点图 13

8.2 b. 观察散点图判断关系 13

8.3 d. 检验显著性 14

9 第九题 **16**

9.1 a. 可视化探索流失客户与非流失客户的行为特点 16

9.2 b. 均值比较验证不同是否显著 17

9.3 c. 建立回归方程进行预测 18

9.4 d. 对尚未流失的客户进行流失可能性排序并给出前 100 名用户 ID 列表 19

1 第一题

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

```
## [1] 13.3
```

```
## [1] 16.5
```

1.2 b. Compute the mean, median, and mode

```
## [1] 15.04286
```

```
## [1] 15
```

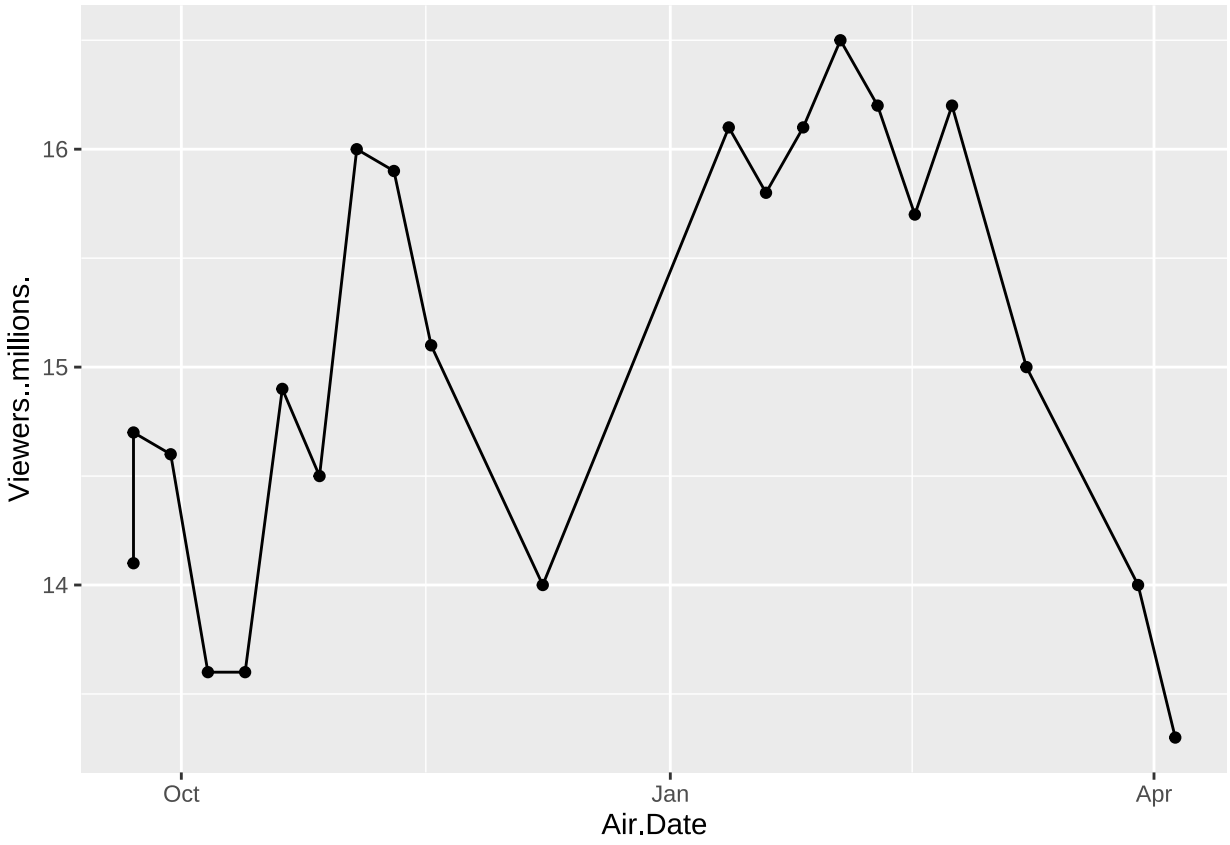
```
## [1] 13.6
```

1.3 c. Compute the first and third quartiles

```
##  25%  75%
## 14.1 16.0
```

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

```
## [1] "en_US.UTF-8"
```



结论：收视率呈波动状态，11 年末有明显下降趋势，12 年初回升至 16.5 万观众之后又出现下降的走势

2 第二题

2.1 a. Show the frequency distribution.

```
##
## (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
```

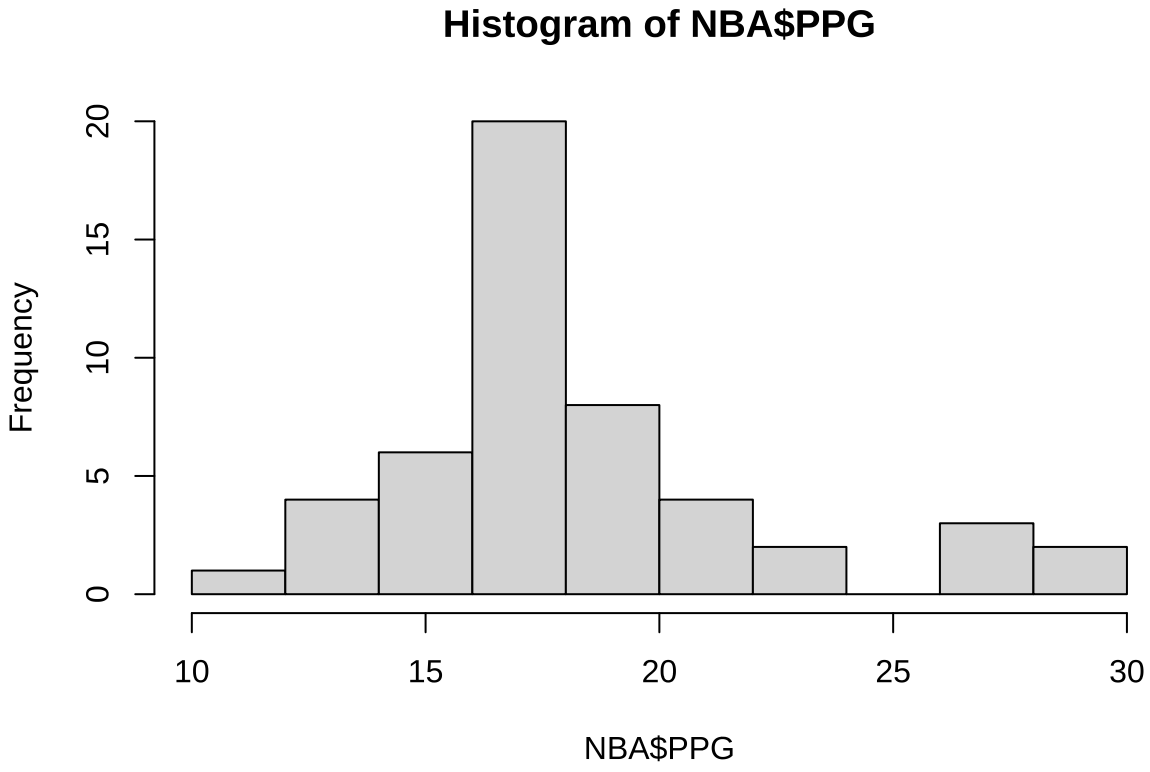
2.2 b.Show the relative frequency distribution.

```
##
## (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
```

2.3 c. Show the Cumulative Percent Frequency Distribution

```
## (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
```

2.4 d. Develop a Histogram for the Average Number of Points Scored per Game



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

```
## [1] "右偏"
```

结论：每场平均得分的平均数大于中位数，该组数据右偏

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

3 第三题

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

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

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

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

4 第四题

4.1 a. Develop appropriate descriptive statistics to summarize the data.

```
##       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
##
##
##
##
```

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

```
##
##  One Sample t-test
##
## data:  Young_Professional$Age
## t = 151.52, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  29.72153 30.50286
## sample estimates:
## mean of x
##   30.1122

##
##  One Sample t-test
##
## data:  Young_Professional$Household.Income.
## t = 43.302, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  71079.26 77839.77
## sample estimates:
## mean of x
##  74459.51
```

4.3 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-sample proportions test with continuity correction
##
## data:  broadband_counts["Yes"] out of sum(broadband_counts), 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

##
##  1-sample proportions test with continuity correction
##
## data:  children_counts["Yes"] out of sum(children_counts), null probability 0.5
## X-squared = 1.778, df = 1, p-value = 0.1824
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4845521 0.5830908
## sample estimates:
##         p
## 0.5341463
```

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

```
## [1] "《青年专业人士》可能是网络经纪商的一个不错的广告投放渠道，因为平均投资价值较高。"
```

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

```
## [1] "对于销售幼儿教育软件和电脑游戏的公司来说，这本杂志可能是一个做广告的好地方，因为有超过一半
```

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

```
## [1] 28538.29
```

```
## [1] 74459.51
```

从以上数据中可以看出《青年专业人士》杂志读者大多在 30 岁左右，且大部分事业有成，支出和收入都很不错，他们可能会对投资与理财，生活方式与健康的文章内容更感兴趣，杂志可以多往这些方面涉及

5 第五题

5.1 a. 对每个样本在 0.01 的显著性水平下进行假设检验

```
## 样本 1 ： 不拒绝原假设H0。p值为 0.2810083样本 2 ： 不拒绝原假设H0。p值为 1.54535样本 3 ： 拒绝原假
```

5.2 b. 计算四个样本的标准差

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

5.3 c. 计算控制限

```
## [1] 11.90124 12.09876
```

5.4 d. 讨论将显著性水平更改为较大值的影响

增大显著性水平会增加犯第一类错误（弃真错误）的概率, 更容易错误地拒绝原假设，导致不必要的纠正措施

6 第六题

6.1 a. 估计 2007 年 3 月第一周和 2008 年 3 月第一周已出租单元的比例

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

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

6.2 b. 为比例差异提供 95% 的置信区间

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(occupancy$X7.Mar == "Yes"), sum(occupancy$X8.Mar == "Yes")) out of c(length(occupa
```

```
## X-squared = 4.3872, df = 1, p-value = 0.03621
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.226151510 -0.007181823
## sample estimates:
##    prop 1    prop 2
## 0.3500000 0.4666667
```

6.3 c. 根据发现判断 2008 年 3 月的租金率是否会比前一年有所上升

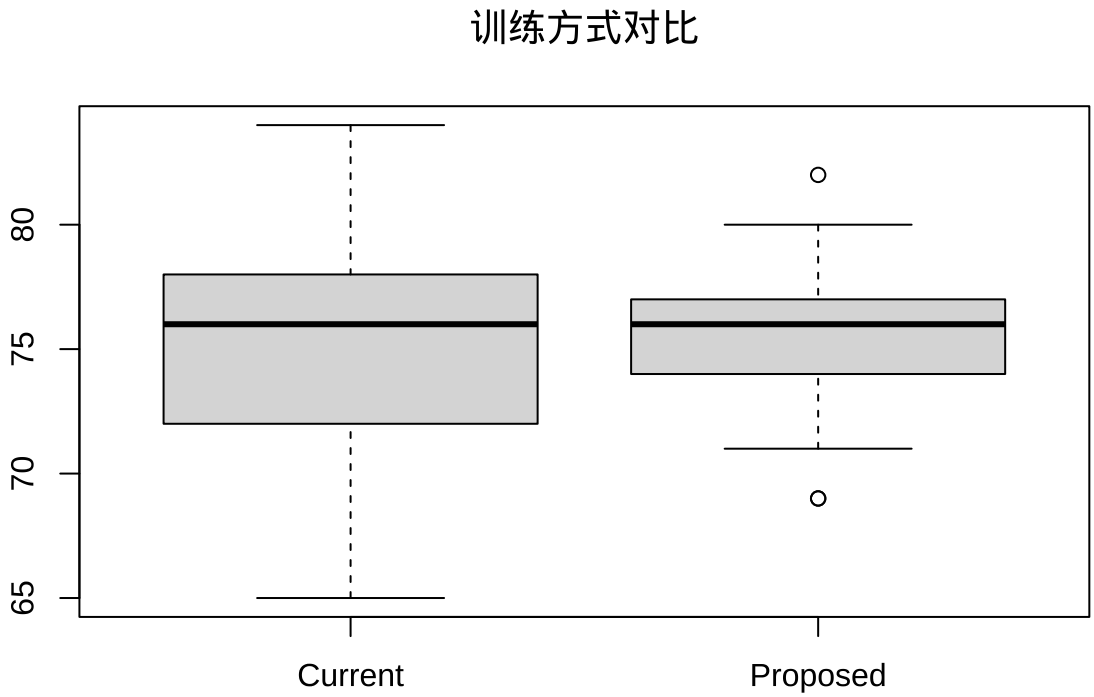
置信区间上限为-0.007，区间不包含 0，且 2008 年的比例大于 2007 年的比例，2008 年 3 月的租金率会比前一年会有所上升

7 第七题

7.1 a. 使用恰当的描述性统计量来汇总每种教学方法的训练时间数据

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

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



两种训练方法训练时间中位数相同，现有的训练方式训练时间最大值和最小值相差较大，数据波动较大；提议的训练方式训练时间最大值和最小值相差较小，数据波动较小

7.2 b. 对两种教学方法的总体均值之间的差异进行评论

```
##
##  Welch Two Sample t-test
##
## data:  current and 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
```

7.3 c. 计算每种教学方法的标准差和方差。针对两种教学方法的总体方差是否相等进行假设检验

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

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

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

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

```
##
##  F test to compare two variances
##
## data:  current and 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
```

```
## function (x, ...)
## UseMethod("var.test")
## <bytecode: 0x000001b5bcfbb8e8>
## <environment: namespace:stats>
```

7.4 d. 关于这两种教学方法之间的任何差异，你能得出什么结论？

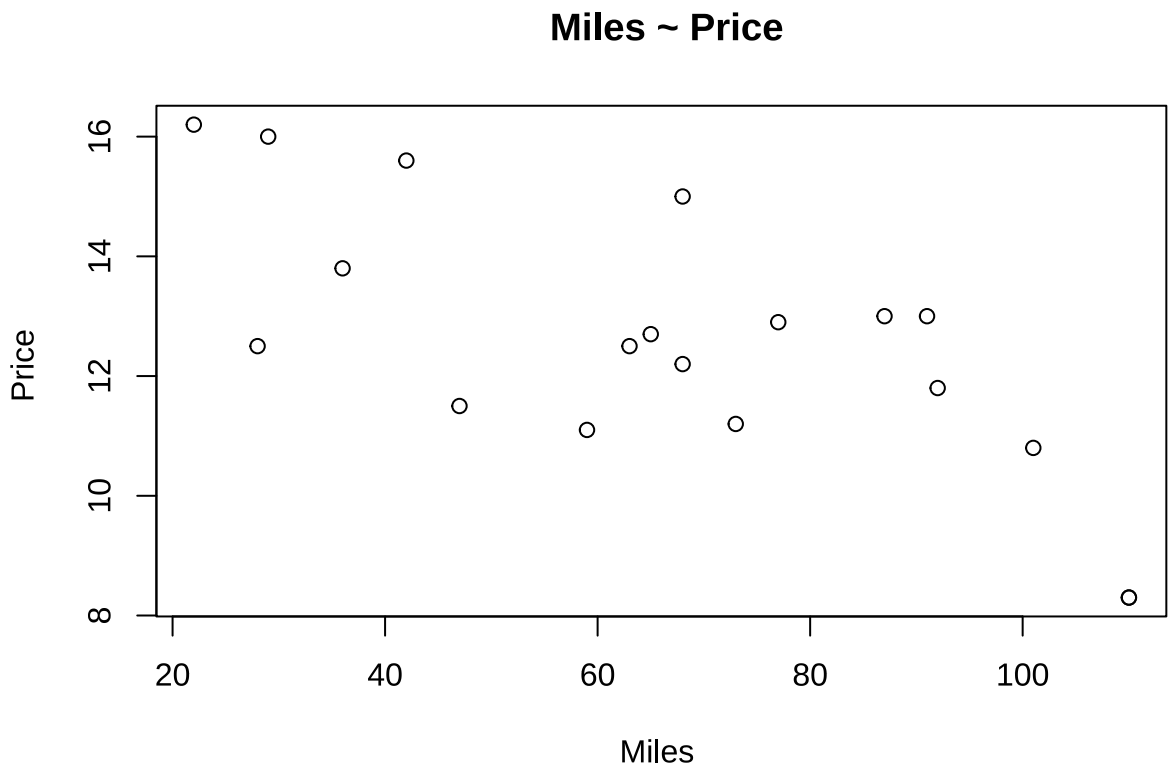
根据假设检验的结果，提议的训练方式训练时间相较于现有的训练方式更稳定更平均，计算机辅助教学更适用于空军训练

7.5 e. 建议其他可能需要的数据或测试

长期效果评估：评估两种方法对学生长期学习效果的影响。成本效益分析：比较两种方法的成本效益。学生满意度调查：收集学生对两种教学方法的反馈，了解他们的偏好和体验。

8 第八题

8.1 a. 绘制散点图



8.2 b. 观察散点图判断关系

从散点图可以看出，随着里程数的增加，价格有下降的趋势，两者存在负相关关系 ## c. 建立线性回归方程

```
##
## 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
```

8.3 d. 检验显著性

```
## [1] 0.000347511
```

在 0.05 的显著性水平下，里程数与价格之间存在显著的线性关系 ## e. 评估拟合优度

```
## [1] 0.5386574
```

行驶里程可以解释汽车价格 53.87% 的变异，模型拟合得较好。## f. 解释斜率回归方程的斜率表示每增加 1000 英里的行驶里程，汽车的价格（以千美元计）平均下降 58.77 美元。## g. 预测价格

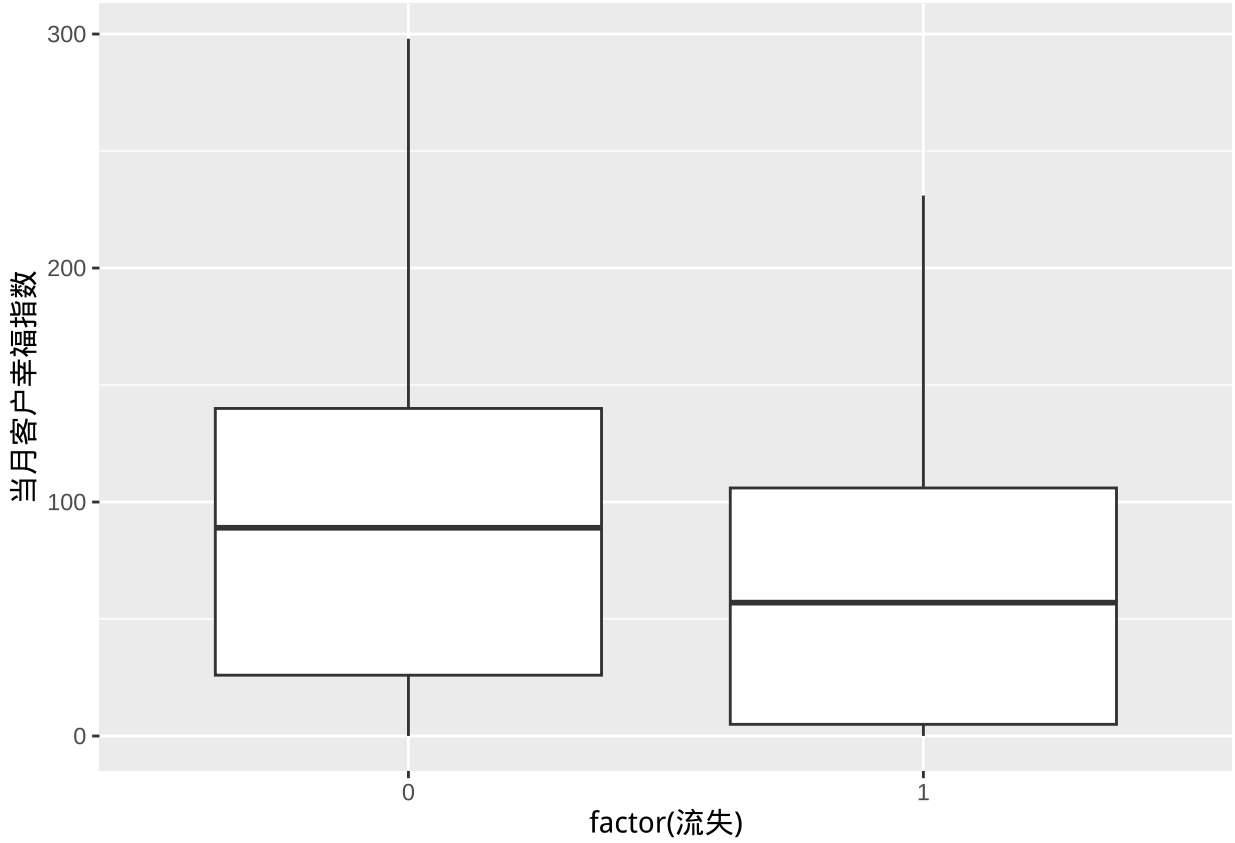
```
predicted_price <- predict(model, newdata = data.frame(Miles..1000s. = 60))
predicted_price
```

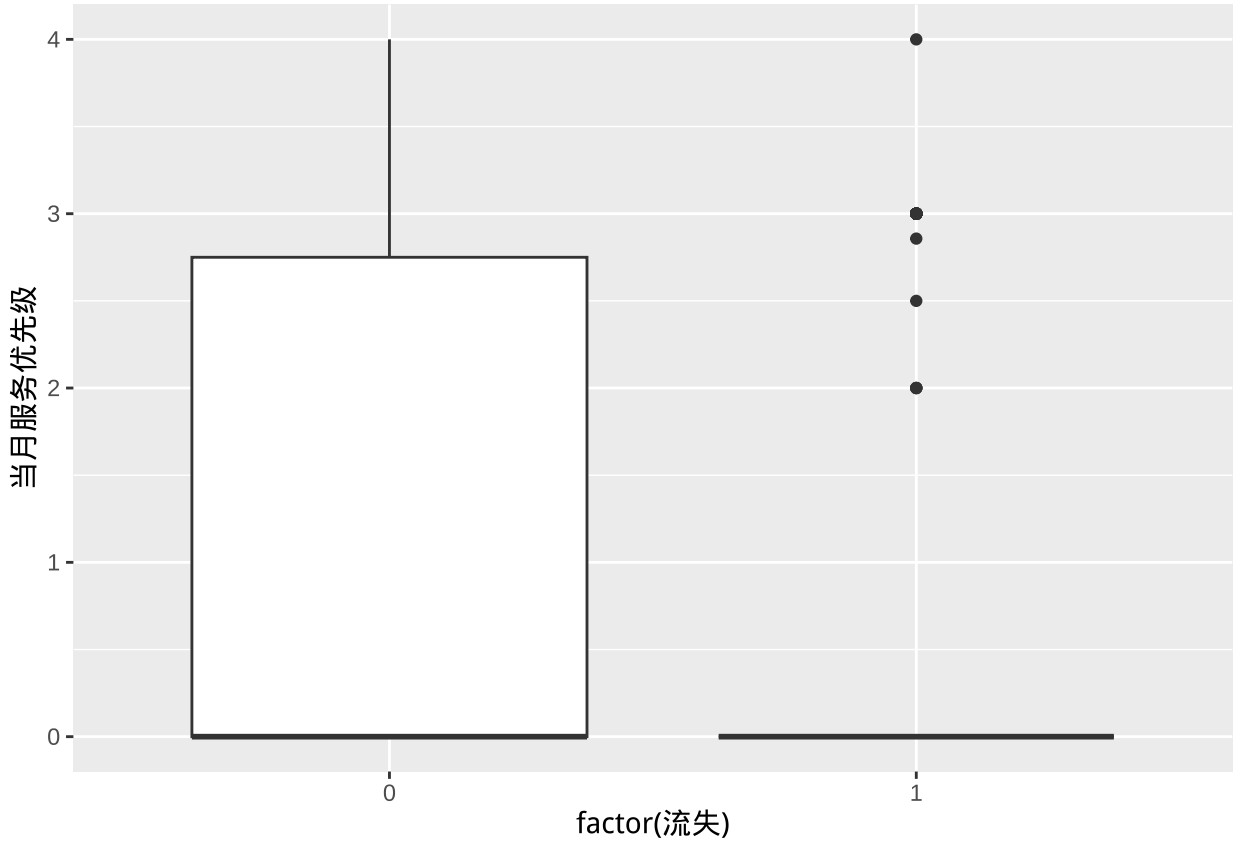
```
##        1
## 12.94332
```

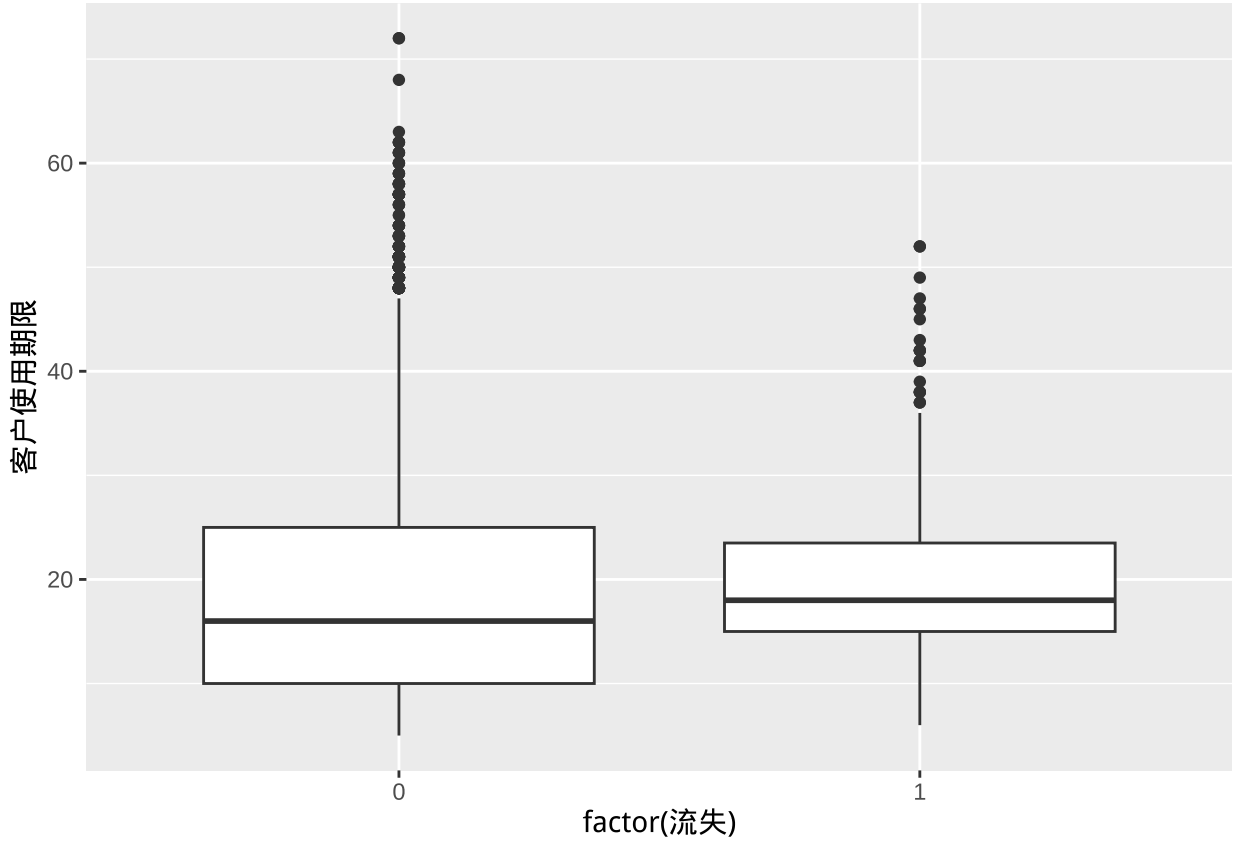
这个价格不一定是实际卖价，实际价格还受其他因素影响，模型不能完全预测价格，只能作为参考

9 第九题

9.1 a. 可视化探索流失客户与非流失客户的行为特点







9.2 b. 均值比较验证不同是否显著

```
##
##  Welch Two Sample t-test
##
## data:  当月客户幸福指数 by 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 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:  当月服务优先级 by 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 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:  客户使用期限 by 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 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
```

9.3 c. 建立回归方程进行预测

```
##
## Call:
## glm(formula = 流失 ~ 当月客户幸福指数 + 当月服务优先级 +
##     客户使用期限, family = binomial(), data = we)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.771237   0.114435 -24.217  < 2e-16 ***
## 当月客户幸福指数 -0.006936   0.001076  -6.444 1.17e-10 ***
## 当月服务优先级   -0.082358   0.055273  -1.490    0.136
## 客户使用期限      0.021643   0.004777   4.531 5.88e-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: 2482.7  on 6343  degrees of freedom
## AIC: 2490.7
##
## Number of Fisher Scoring iterations: 6
```

9.4 d. 对尚未流失的客户进行流失可能性排序并给出前 100 名用户 ID 列表

```
##      客户ID
## 1         1
## 14       14
## 3         3
## 18       18
## 21       21
## 56       57
## 51       51
## 2         2
## 54       55
## 58       59
## 116     121
## 59       61
## 73       76
## 91       95
## 105     110
## 132     137
## 60       62
## 5         5
## 12       12
## 148     154
## 42       42
## 66       69
## 114     119
## 141     146
## 164     171
## 176     183
## 183     190
## 72       75
## 97      101
## 118     123
## 104     109
```

```
## 1280   1392
## 136     141
## 137     142
## 1281   1393
## 101     106
## 1306   1419
## 1325   1438
## 30       30
## 86       89
## 16       16
## 81       84
## 62       64
## 65       68
## 1283   1395
## 1362   1478
## 1402   1520
## 2061   2235
## 2066   2240
## 2080   2255
## 195     203
## 1345   1459
## 1348   1462
## 2071   2245
## 1378   1496
## 107     112
## 1006   1108
## 1041   1143
## 151     158
## 123     128
## 10       10
## 1742   1893
## 1755   1908
## 142     147
## 2111   2286
## 69       72
## 125     130
## 17       17
## 156     163
## 1358   1474
```

```
## 61     63
## 1795   1951
## 1814   1971
## 122    127
## 2070   2244
## 47     47
## 108    113
## 162    169
## 990    1091
## 1039   1141
## 1271   1383
## 1884   2047
## 1899   2062
## 1907   2070
## 100    104
## 1797   1953
## 128    133
## 41     41
## 172    179
## 185    192
## 1333   1446
## 1915   2080
## 117    122
## 2106   2281
## 2130   2306
## 1008   1110
## 959    1058
## 964    1063
## 1941   2108
## 2542   2744
```