Solution for MEM Assignment r2

胡浛

2024-11-30

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

```
## a. Minimum Viewers: 13.3 , Maximum Viewers: 16.5
```

```
## b. Mean = 15.04286 , Median = 15 , Mode = 13.6
```

```
## c. Q1 = 14.1 , Q2 =  16
```

```
## d. p值>0.05，回归系数不显著，从给出的数据没有看出有明显的上升或下降趋势
```

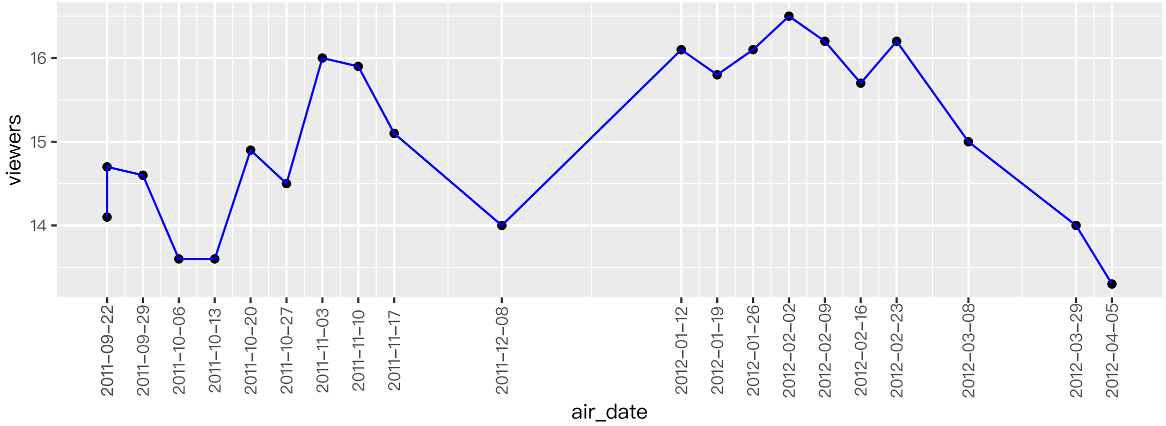


图 1: viewer 观众数变化趋势图

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

a. Show the frequency distribution.

ppg 频率直方分布图如下图

```
## ppg_groups
## [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
##       1       3       7      19       9       4       2       0       3       2
```

b. Show the relative frequency distribution.

表 1: ppg 相对频率分布表

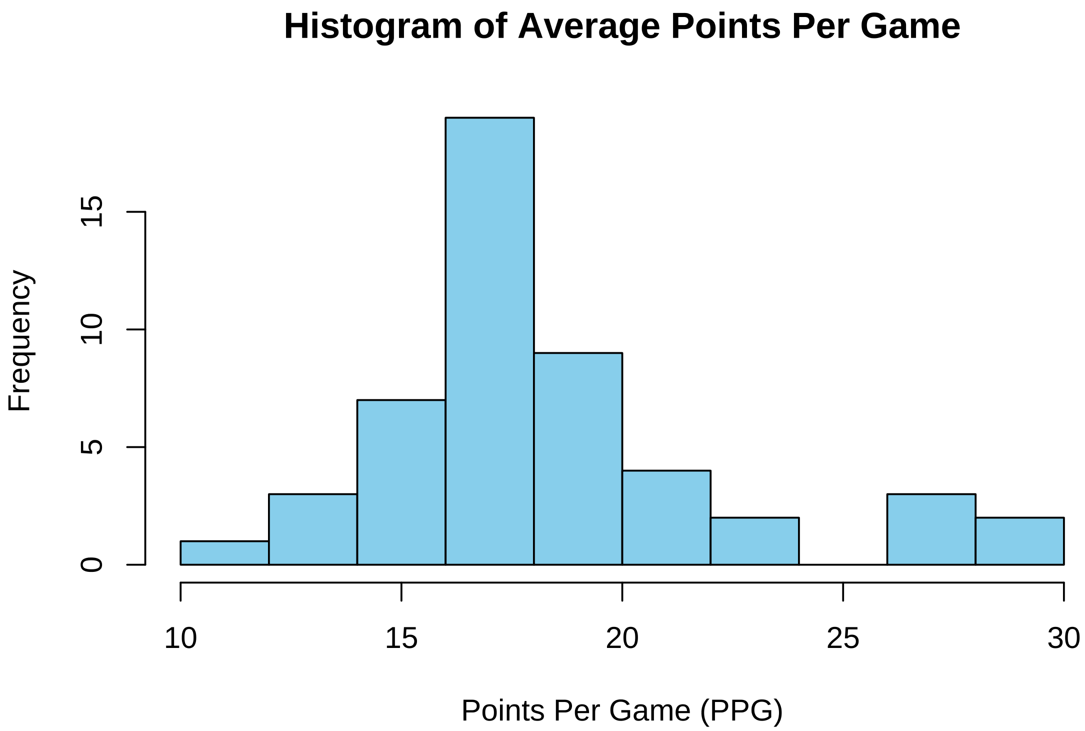
ppg 分组	相对频率 (%)
[10,12)	0.02
[12,14)	0.06
[14,16)	0.14
[16,18)	0.38
[18,20)	0.18
[20,22)	0.08
[22,24)	0.04
[24,26)	0.00
[26,28)	0.06
[28,30)	0.04

c. Show the cumulative percent frequency distribution.

表 2: ppg 累计频率分布表

ppg 分组	累计频率 (%)
[10,12)	2
[12,14)	8
[14,16)	22
[16,18)	60
[18,20)	78
[20,22)	86
[22,24)	90
[24,26)	90
[26,28)	96
[28,30)	100

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



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

Skewness=1.1240253 > 0, 数据右偏

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

场均得分至少为 20 分的球员占比: 22%.

Question #3:

a. How large was the sample used in this survey?
b. What is the probability that the point estimate was within ±25 of the population mean?

```
## 样本量为: 625
## 样本均值在总体均值 ±25 范围内的概率为: 0.7887005
```

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

a. Develop appropriate descriptive statistics to summarize the data.

```
##       age            gender    real_estate_purchases  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
##  transactions_count access_broadband     income       have_childred
##  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
```

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

平均年龄 95% 置信区间为 [29.7226869, 30.5017033], 平均年收入 95% 的置信区间为 $[7.1089258 \times 10^4, 7.7829766 \times 10^4]$。

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

95% confidence intervals for the proportion of subscribers who have broadband access at home is [0.577514, 0.6712665];

95% confidence intervals for the proportion of subscribers who have children is [0.4858615, 0.5824312]

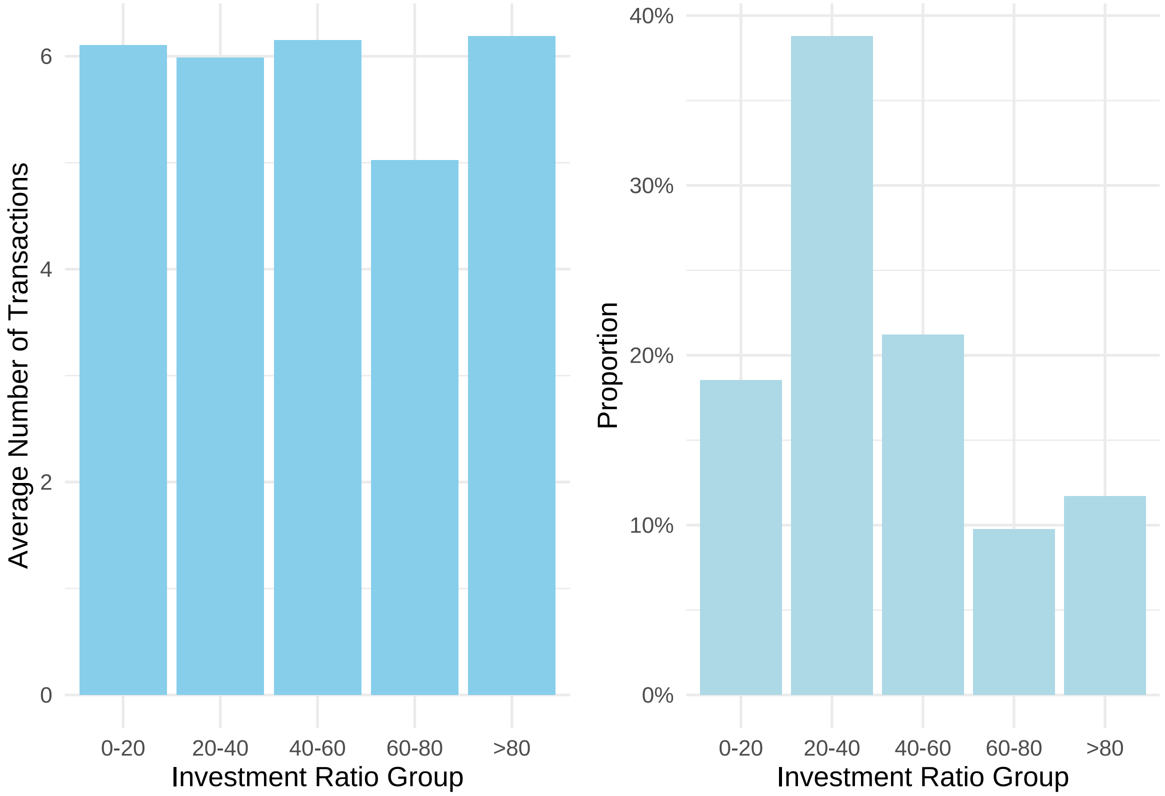
d. Would *Young Professional* be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

Yes. 如下图所示，按金融投资金额占家庭年收入的比值分为 5 组

1) 占总人数 80% 以上的人 Young Professional，金融投资占年收入比值 >20%，这部分人群都是潜在的广告客户

2) 接近 62.4% 的人群允许了广告投放

3) 每组的平均金融交易次数没有因投资比例而呈现明显的差异，可能是因为缺少专业的 brokers



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

Yes. 有小孩的人群占比为 53.4%。

```
## # A tibble: 2 x 3
##   have_childred count  prop
##   <fct>         <int> <dbl>
## 1 No              191  46.6
## 2 Yes             219  53.4
```

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

感兴趣的文章类型

1) 财经, 人群中投资比、收入均比较高
2) 育儿，有小孩的人群超过总人数的 50%
3) 房产，有房产交易的占比 44%
4) 运动，人群中青年人居多

```
## # A tibble: 2 x 3
##   real_estate_purchases count  prop
##   <fct>                 <int> <dbl>
## 1 No                      229  55.9
## 2 Yes                     181  44.1
```

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

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.

the p-value for each test as below

```
## Sample.1 p值为 0.2810083 不需要整改
## Sample.2 p值为 0.4546503 不需要整改
## Sample.3 p值为 0.003790318 需要整改
## Sample.4 p值为 0.03389336 不需要整改
```

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

每个样本的标准差与预估标准差 0.21 的偏差均小于 5%, 可以认为总体的标准差为 0.21 是比较合理的

```
## Sample.1 Sample.2 Sample.3 Sample.4
## 4.931445 4.931445 1.347336 1.852858
## [1] 11.90124 12.09876
```

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.

lower and upper control limits is 11.9012412, 12.0987588

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?

显著性水平增加，允许拒绝原假设的证据要求越低，测试更倾向于拒绝 H_0。α =0.01 调整为 α=0.05 时，对假设检验的要求更宽松, 第一类错误增加。

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

```
## [1] 11.92485 12.07515
```

Question #6

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

```
## March.2007 March.2008
##  0.3500000  0.4666667
```

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

```
## 出租比例差异的95%置信区间: [ 0.01301325 , 0.2203201 ]
```

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

是的。置信区间的下限为正数，说明 2008 年出租率显著高于 2007 年。因为在该置信区间内，所有可能的差异值都表明 2008 年出租率高于 2007 年

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

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?

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	num
numeric	Current	0	1	75.06557	3.944907	65	72	76	78	
numeric	Proposed	0	1	75.42623	2.506385	69	74	76	77	

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

```
## [1] 75.06557
```

```
## [1] 75.42623
```

```
##
##  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
```

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.

```
## Method 1 - Variance: 15.5623 SD: 3.944907
```

```
## Method 2 - Variance: 6.281967 SD: 2.506385
```

```
##
##  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
```

Current 和 Proposed 的方差存在显著差异，Proposed 的方差更小。

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

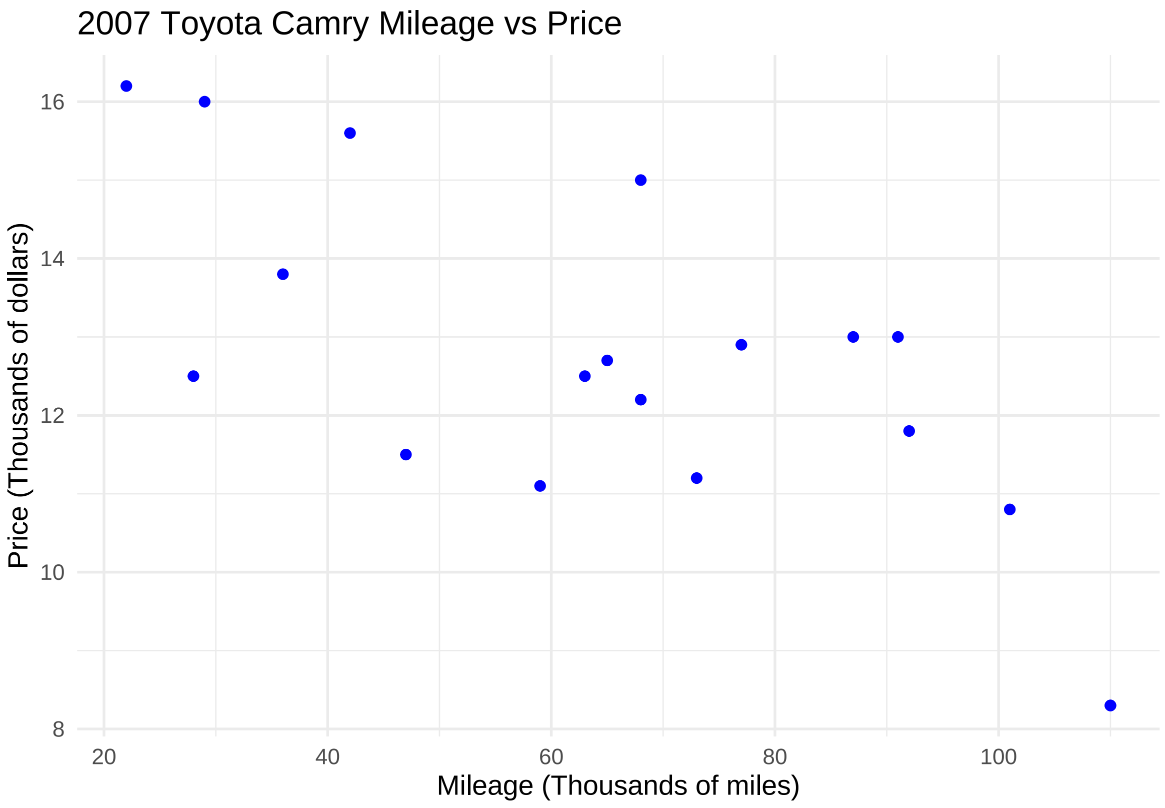
学习通过两种方式培训的时长均值不存在显著差异，但 Proposed 的方差更小，说明时长分布相对密集，为了减少其他同学的等待时长，选择 Proposed 的方式更佳。

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?

为了做出全面的决策，除了培训时长外，还需要考虑培训后的效果，完成培训后，记录每个学生在培训后的表现，分析两组的培训效果是否存在差异。

Q8

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?

曲线整体是一个向下倾斜的趋势，说明随着里程的增加，价格可能在下降。

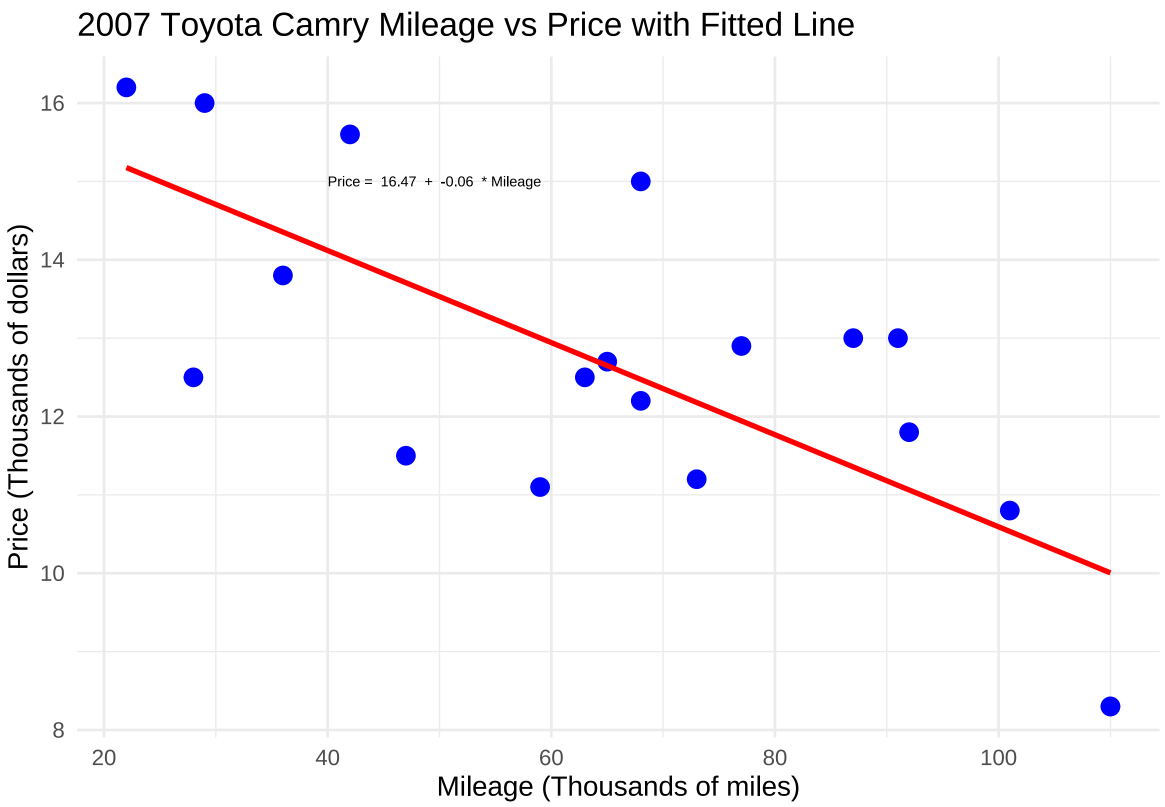
c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

```
##
## Call:
## lm(formula = Price ~ Mileage, data = camry_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 ***
## Mileage     -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

## 回归方程为: Price =  16.46976  +  -0.05877393  * Mileage

## 预测里程为60,000英里的价格为:  12.94332 千美元
```



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

斜率（Mileage）的 p 值为 0.0003，远小于 0.05，说明我们拒绝零假设，即里程与价格之间存在显著的负相关关系

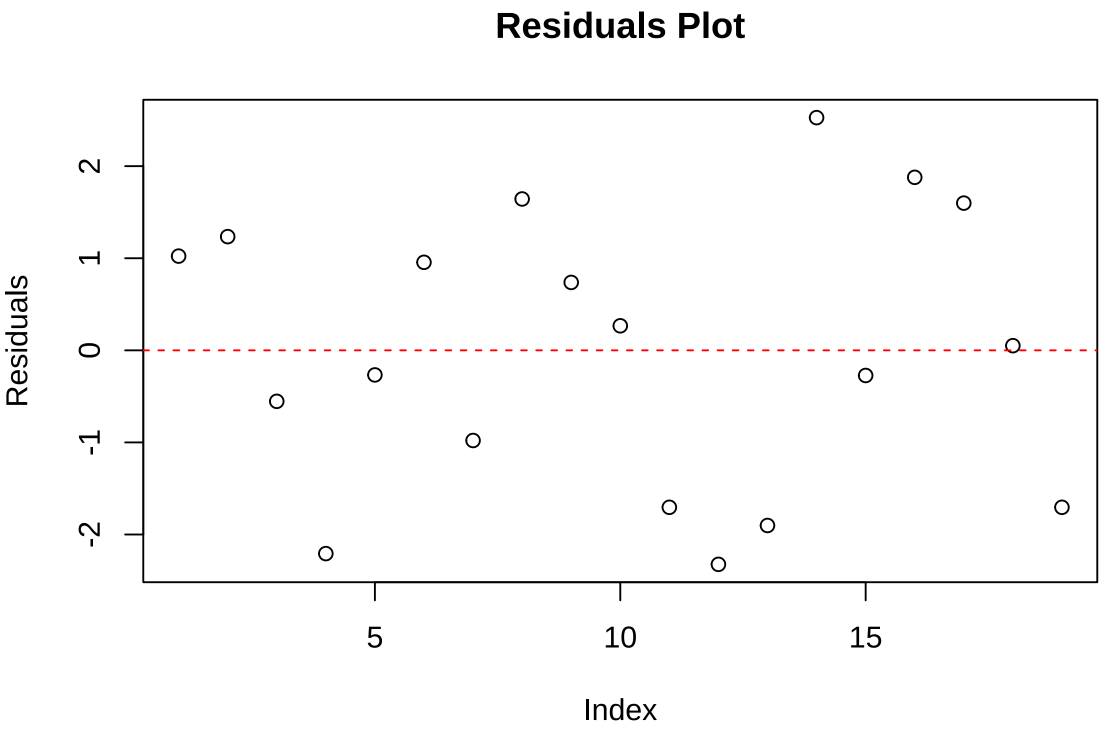
```
##
## Call:
## lm(formula = Price ~ Mileage, data = camry_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.46976    0.94876  17.359 2.99e-12 ***
## Mileage     -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
```

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

```
## R-squared: 0.5386574
```

```
## F-statistic: 19.84897
```

```
## F p-value: 0.000347511
```



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

这意味着每行驶1000英里，汽车的价值下降大约60美元。

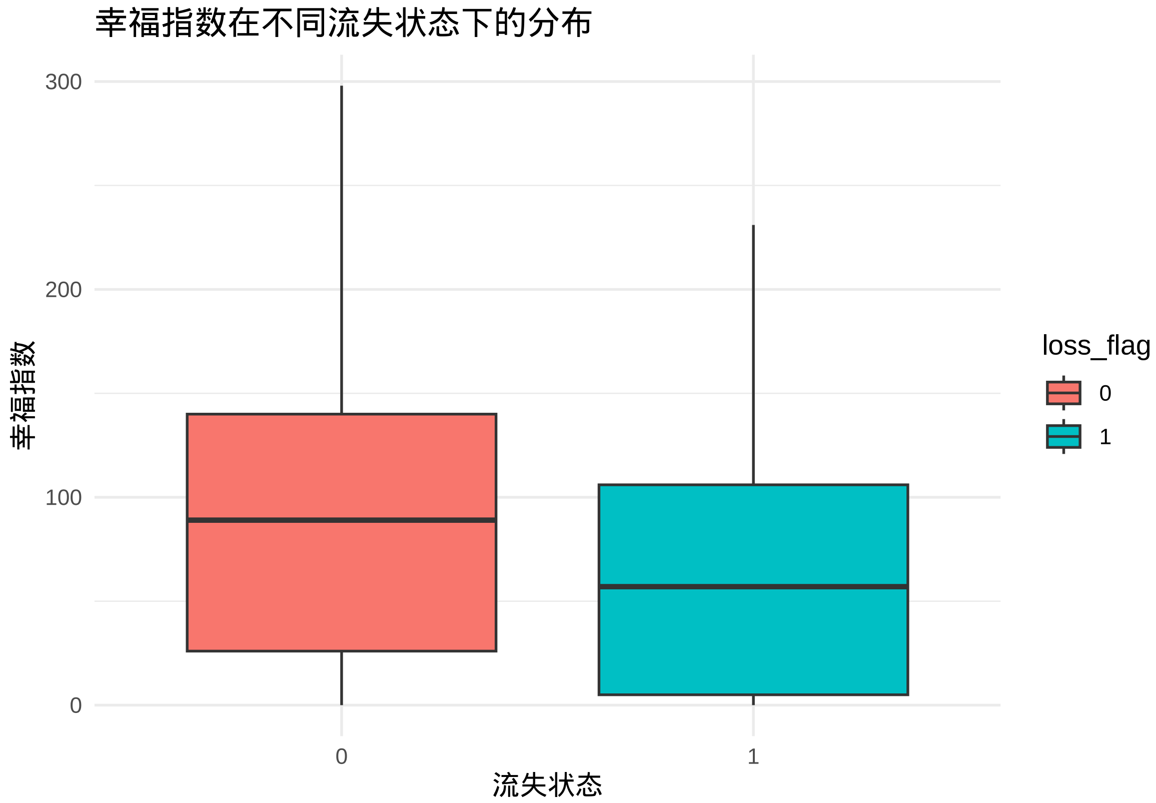
g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller.

使用回归方程预测价格为12.94千美元, 考虑当前二手车市场低迷，可以在这个价格上适当压价。

Q9

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?

1. 相对于非流失客户，流失客户当月幸福指数可能更低
2. 相对于非流失客户，流失客户访问间隔更大，流失客户的访问频率可能更低
3. 相对于非流失客户，流失客户当月支持、当月服务优先级可能更低；当月登录次数、博客数相比上月的变化、访问次数都相对较低



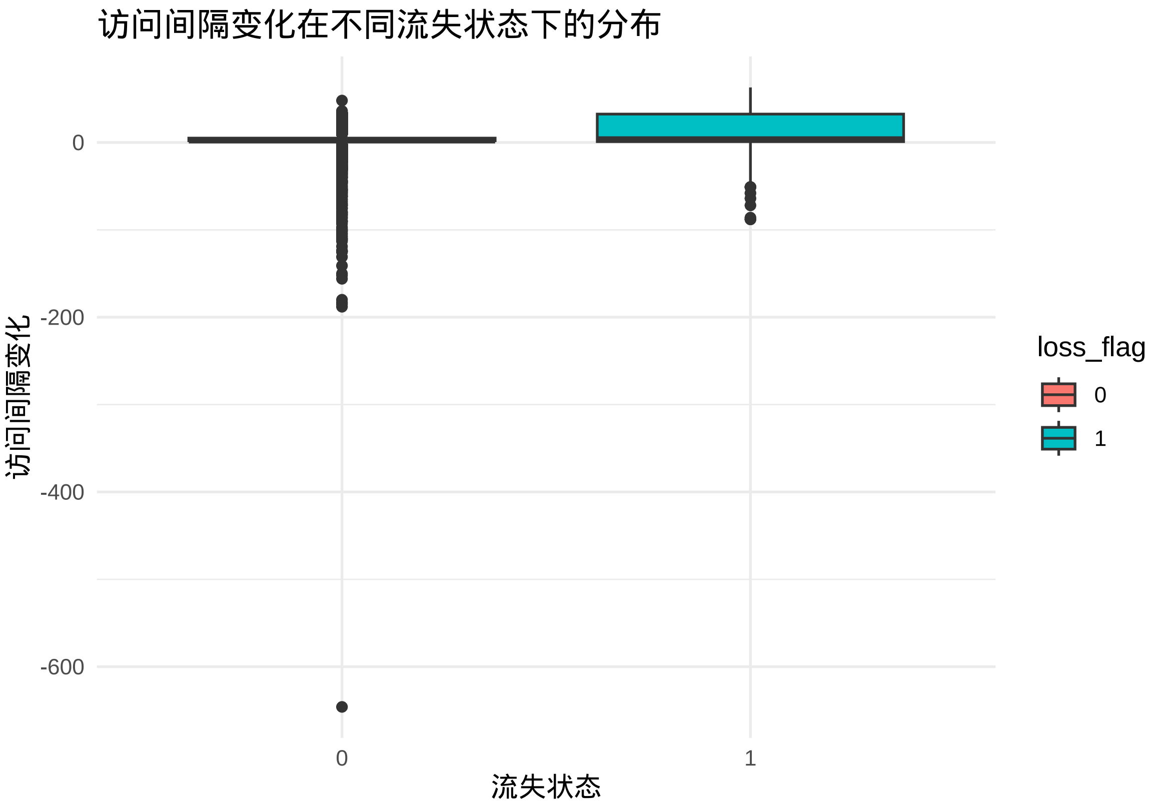


表 4: 流失与非流失客户关键指标均值对比

loss_flag	m_support	m_service_priority	m_login_times	m_visits_change	m_blogs_change
0	0.7242696	0.8295759	16.13894	106.6096	0.1711487
1	0.3715170	0.4995577	8.06192	-95.7678	-0.1021672

b. 通过均值比较的方式验证上述不同是否显著。

表 5: 流失与非流失客户关键指标显著性对比

Variable	Mean (Loss)	Mean (No Loss)	p_value	Significance
happiness_degree	63.2724458	88.6059097	0.0000000	***
happiness_degree_change	-3.7368421	5.5302125	0.0000000	***
support	0.3715170	0.7242696	0.0000001	***
support_change	0.0371517	-0.0092961	0.5277532	NS

service_priority	0.4995577	0.8295759	0.0000004	***
service_priority_change	-0.0166962	0.0326818	0.5218233	NS
login_times	8.0619195	16.1389442	0.0004037	***
blogs_change	-0.1021672	0.1711487	0.0115761	*
visits_change	-95.7678019	106.6095618	0.0563070	NS
use_time	20.3529412	18.8187251	0.0030568	**
visit_interval	8.4860681	3.5114542	0.0000522	***

从上表可以看出，除了 support_change、service_priority_change、visits_change 其他变量都可能是影响客户流失的显著性性指标。

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_flag ~ happiness_degree + happiness_degree_change +
##     support + service_priority + login_times + blogs_change +
##     use_time + visit_interval, family = binomial, data = we_data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.8763327  0.1212590 -23.721  < 2e-16 ***
## happiness_degree        -0.0051988  0.0011558  -4.498 6.86e-06 ***
## happiness_degree_change -0.0093063  0.0024124  -3.858 0.000114 ***
## support                 -0.0221691  0.0714550  -0.310 0.756369
## service_priority        -0.0447524  0.0741355  -0.604 0.546072
## login_times              0.0008545  0.0019376   0.441 0.659211
## blogs_change            -0.0009717  0.0205099  -0.047 0.962213
## use_time                 0.0142559  0.0052396   2.721 0.006513 ** 
## visit_interval           0.0169505  0.0042787   3.962 7.44e-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: 2452.2  on 6338  degrees of freedom
## AIC: 2470.2
##
## Number of Fisher Scoring iterations: 6
```

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 (happiness_degree) 及其相比上月变化 (happiness_degree_change) 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 (use_time) 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 (visit_interval) 越大，用户流失的概率越高，用户越不活跃越易流失

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。

流失可能性最大的前 100 名用户如下表，针对这些高风险客户，可采取以下措施:

1. 设计促活措施，缩短客户访问间隔，比如定期推送内容、发放优惠券;
2. 持续为老客户提供价值，避免倦怠感，提供老用户的增值服务

表 6: 流失可能性最大的前 10 名用户 ID 列表

customer_id	predicted_prob

109	0.2803438
1971	0.2327480
1	0.2157609
2076	0.2073632
14	0.1926158
76	0.1919250
3	0.1882206
18	0.1882206
21	0.1860521
2244	0.1859914