Solution for MEM Assignment r2

2024281050957_ 夏梦

2024-11-30

目录

Question1:BigBangTheory. **4**
a. Compute the minimum and the maximum number of viewers. 4
b. Compute the mean, median, and mode. 4
c. Compute the first and third quartiles. 4
d. has viewership grown or declined over the 2011–2012 season? Discuss. 5

Question2: NBAPlayerPts. **6**
a. Show the frequency distribution. 6
b. Show the relative frequency distribution. 7
c. Show the cumulative percent frequency distribution. 8
d. Develop a histogram for the average number of points scored per game. 8
e. Do the data appear to be skewed? Explain. 9
f. What percentage of the players averaged at least 20 points per game? 9

Question3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500. **9**
a. How large was the sample used in this survey? 9
b. What is the probability that the point estimate was within ±25 of the population mean? 10

Question 4: Young Professional Magazine **10**
a. Develop appropriate descriptive statistics to summarize the data. 10
b. Develop 95% confidence intervals for the mean age and household income of subscribers. 10
c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children. 10
d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data. 11

e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children? . . . 13
f. Comment on the types of articles you believe would be of interest to readers of Young Professional. 14

Question 5: Quality Associate, Inc. 14
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. . . . 14
b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable? . . . 15
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. . . . 15
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? . . . 16

Question 6 16
a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008. . . . 16
b. Provide a 95% confidence interval for the difference in proportions. . . . 16
c. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier? . . . 17

Question 7: Air Force Training Program 17
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? . . . 17
b. Comment on any difference between the population means for the two methods. Discuss your findings. 18
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. . 18
d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain. . . . 19
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? . . . 19

Question 8 20
a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis. . . . 20
b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables? . . . 20

c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s). 20
d. Test for a significant relationship at the .05 level of significance. 21
e. Did the estimated regression equation provide a good fit? Explain. 22
f. Provide an interpretation for the slope of the estimated regression equation. 23
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. 23

Question 9 **23**

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？ 24
b. 通过均值比较的方式验证上述不同是否显著。 24
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。 25
d. 根据上一步预测的结果，对尚未流失（不流失 =1）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。 26

Question1:BigBangTheory.

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

The minimum number of viewers is 13.3;
The maximum number of viewers is 16.5.

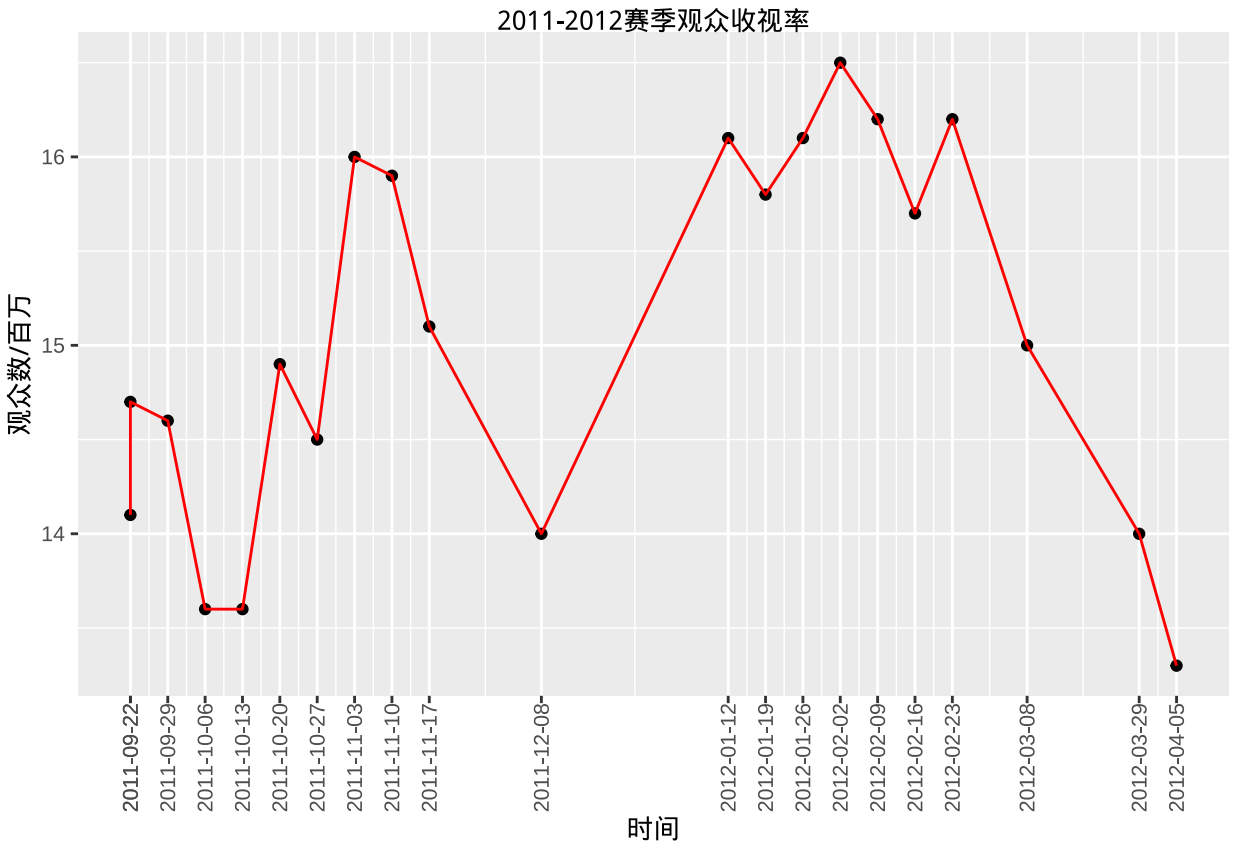
b. Compute the mean, median, and mode.

mean: 15.0429;
median: 15;
mode: 13.6.

c. Compute the first and third quartiles.

The first quartiles is 14.1;
The third quartiles is 16.

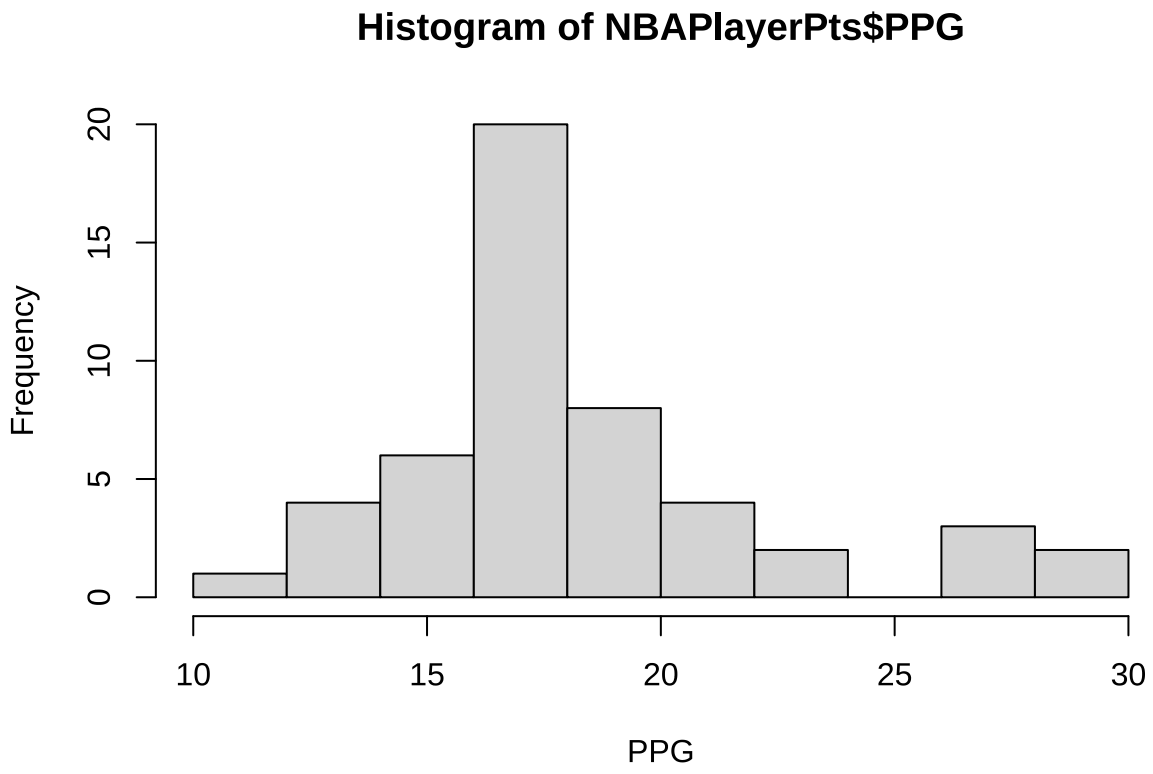
d. has viewership grown or declined over the 2011–2012 season? Discuss.



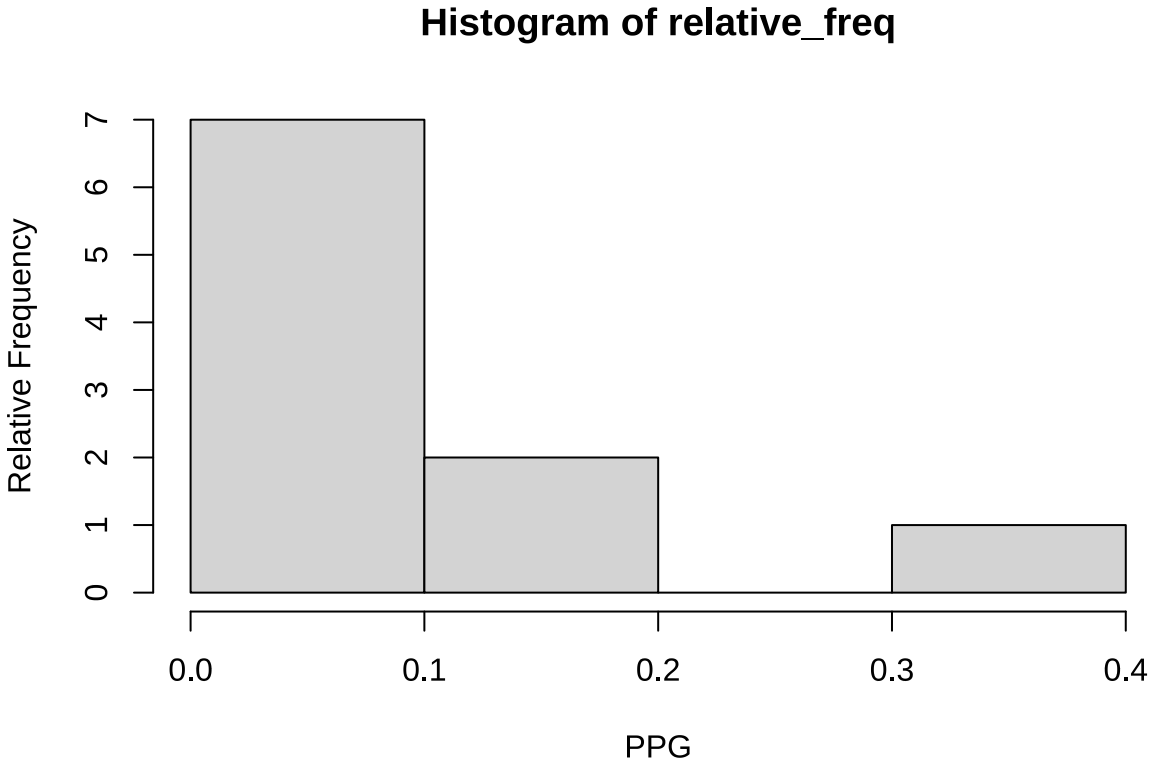
- 可以看到：在 2011 年赛季初期，收视率大致在 1400 万-1500 万左右。到了 2012 年初，收视率上升到了 1600 万左右，但随后在赛季末又有所下降。
- 结论：综合来看，在 2011-2012 赛季期间，收视率并非呈现单一的上升或下降趋势。而是先有一定增长，在 2012 年初达到较高水平后又出现下降。其变化受到多种因素影响，如剧集内容、同期竞争节目、播出时段等。仅从这一赛季的数据不能简单判定其长期的收视率走向，若要更全面了解，还需分析更多季节的数据以及相关影响因素。

Question2: NBAPlayerPts.

a. Show the frequency distribution.

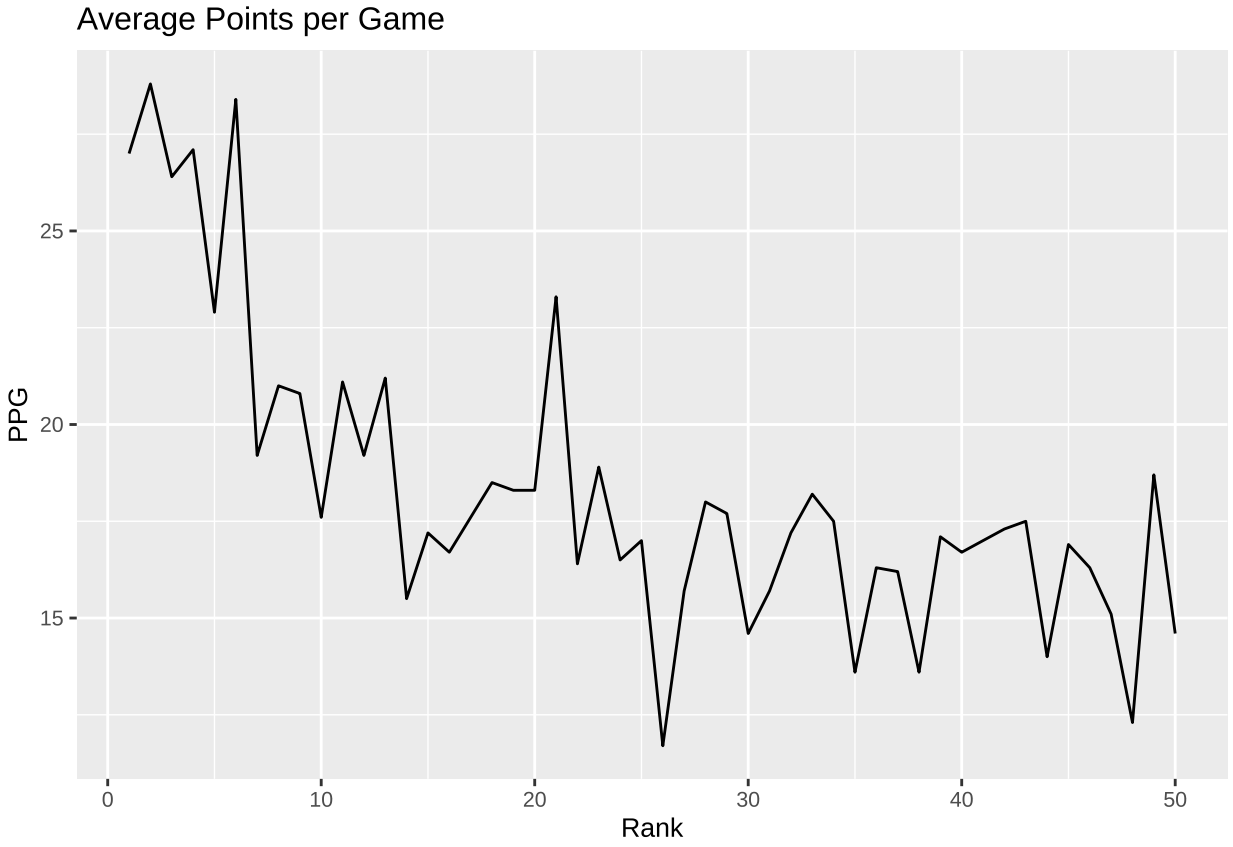


b. Show the relative frequency distribution.

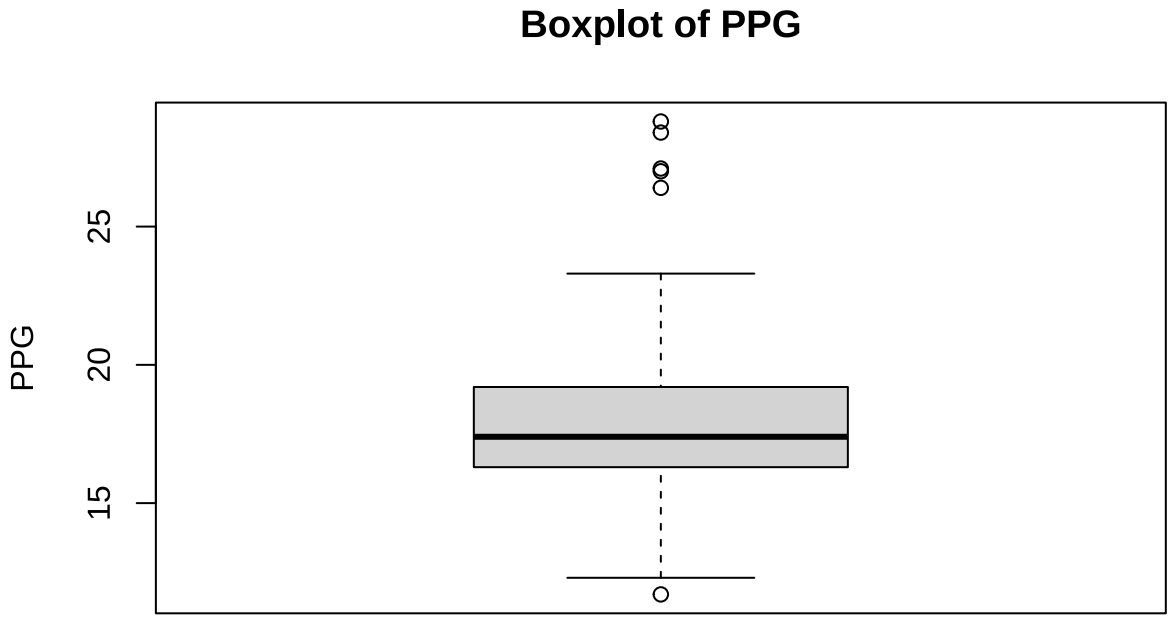


c. Show the cumulative percent frequency distribution.

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



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




```
## [1] 1.124
```

- 可以看到：偏度大于 0，右偏；
- 结论：大部分球员的得分集中在较低的区间，而高分段的球员相对较少，所以数据呈现右偏态（正偏态）。这意味着得分较高的球员相对较少，而得分较低的球员较多，导致分布的右侧（高分端）有较长的尾巴。

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

平均每场比赛得分至少 20 分的球员有 11 个

Question3: 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?

The sample size is 625.

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

The probability that the point estimate was within ±25 of the population mean is: 0.7887.

Question 4: Young Professional Magazine

a. Develop appropriate descriptive statistics to summarize the data.

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts	num
factor	gender	0	1	FALSE	2	Mal: 229, Fem: 181	
factor	real_estate	0	1	FALSE	2	No: 229, Yes: 181	
factor	has_broadband	0	1	FALSE	2	Yes: 256, No: 154	
factor	have_children	0	1	FALSE	2	Yes: 219, No: 191	
numeric	age	0	1	NA	NA	NA	
numeric	investments	0	1	NA	NA	NA	
numeric	num_trans	0	1	NA	NA	NA	
numeric	income	0	1	NA	NA	NA	

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

```
## [1] 29.72 30.50
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 71079 77840
## attr(,"conf.level")
## [1] 0.95
```

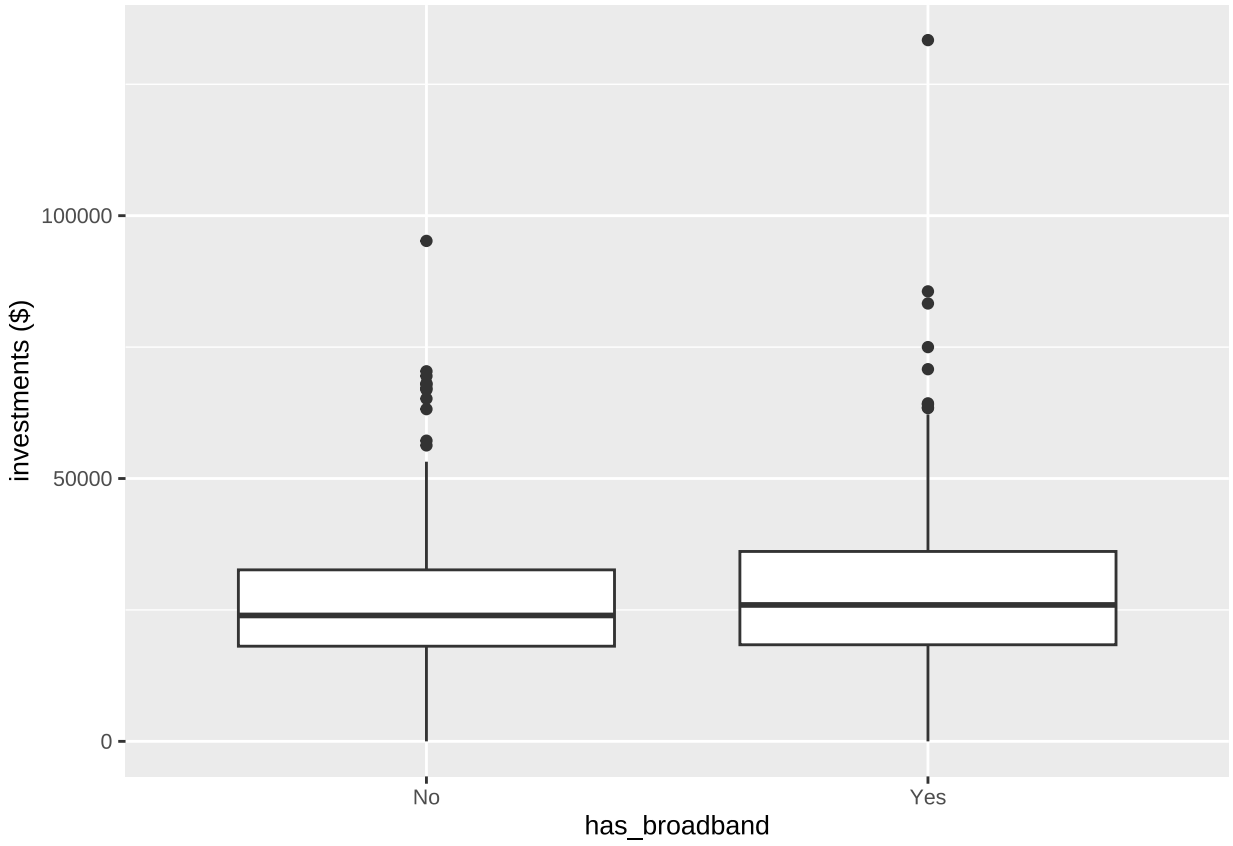
c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```
## [1] 0.5753 0.6711
## attr(,"conf.level")
## [1] 0.95
```

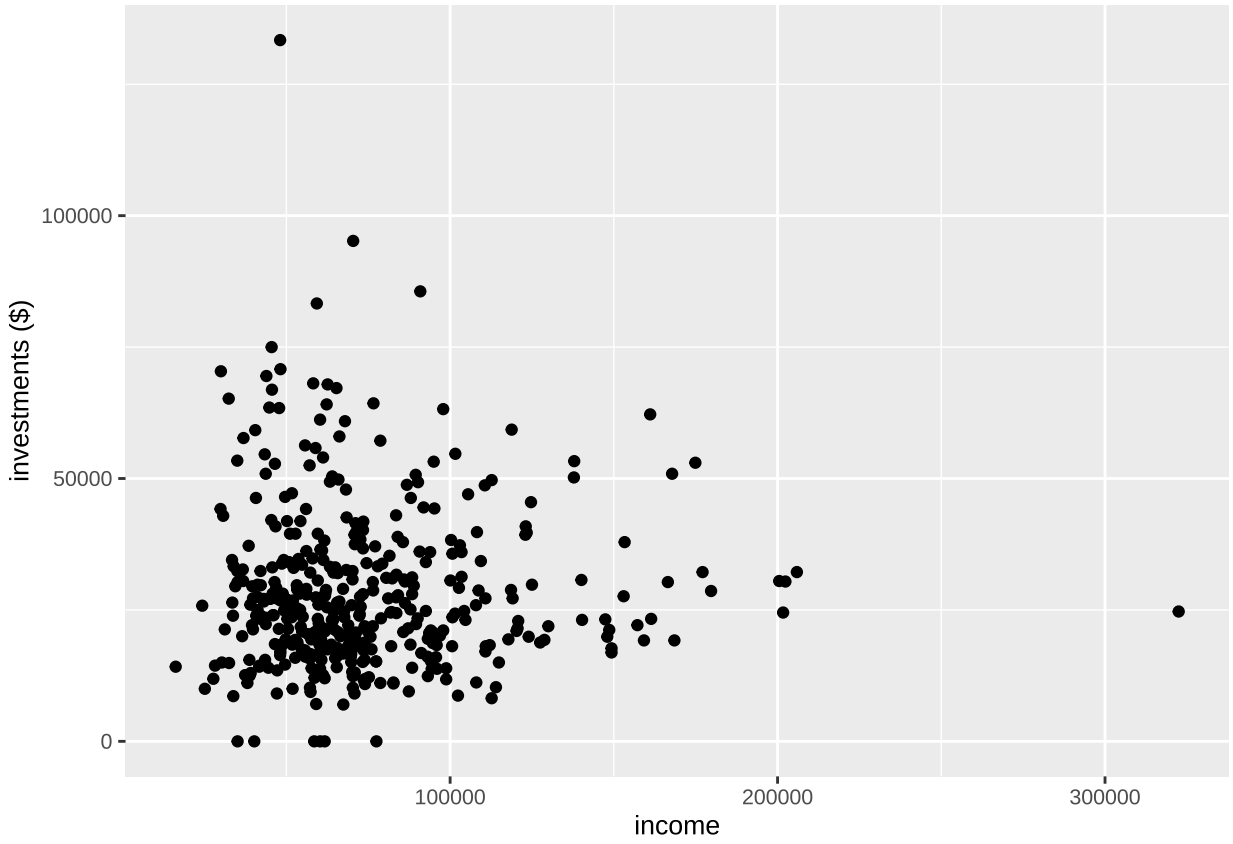
```
## [1] 0.4846 0.5831
## attr(,"conf.level")
## [1] 0.95
```

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

```
##       age          gender    real_estate  investments       num_trans
##  Min.   :19.0   Female:181   No :229     Min.   :     0   Min.   : 0.00
##  1st Qu.:28.0   Male  :229   Yes:181     1st Qu.: 18300   1st Qu.: 4.00
##  Median :30.0                           Median : 24800   Median : 6.00
##  Mean   :30.1                           Mean   : 28538   Mean   : 5.97
##  3rd Qu.:33.0                           3rd Qu.: 34275   3rd Qu.: 7.00
##  Max.   :42.0                           Max.   :133400   Max.   :21.00
##  has_broadband     income       have_children
##  No :154       Min.   : 16200   No :191
##  Yes:256       1st Qu.: 51625   Yes:219
##                Median : 66050
##                Mean   : 74460
##                3rd Qu.: 88775
##                Max.   :322500
```




```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  investments by has_broadband
## W = 18415, p-value = 0.3
## alternative hypothesis: true location shift is not equal to 0

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  investments by have_children
## W = 22088, p-value = 0.3
## alternative hypothesis: true location shift is not equal to 0
```




```
## [1] 0.003026
```

- 可以看到：Young Professional 的订阅者的平均年龄在 30 岁左右，打算在未来两年内购买房地产的占 0.4415，除去房产，金融投资部分的平均金额为 $28538, 在过去的一年内，每个人平均至少有 6 笔股票/债券/共同基金交易；有 0.6244 的人使用宽带上网；去年的家庭总收入平均值为 $74460；
- 年轻专业人士在数据集中占较大比例，并且他们的平均收入、投资价值和交易数量都较高，那么"Young Professional" 将是一个针对在线经纪人的有吸引力的广告渠道。这些指标表明他们更有可能参与投资活动，并且可能对在线经纪服务感兴趣；
- 结论：因此 Young Professional 会是在线经纪人的一个很好的广告渠道。

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

我认为这本杂志会是为那些销售教育软件和儿童电脑游戏的公司做广告的好地方；原因如下：- Young Professional 的订阅者的平均年龄在 30 岁左右；- 有小孩的人群占比 0.5341；- 因此他们的孩子应该正是教育软件和儿童电脑游戏的受众人群；

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

我认为 Young Professional 的读者们对理财投资类、儿童教育类、技术创新与行业动态类、工作与生活平衡类的文章感兴趣;

Question 5: Quality Associate, Inc.

Quality associates, inc. 是一家咨询公司，就可用于控制其生产过程的抽样和统计程序向其客户提供建议。在一个特定的应用程序中，客户向质量部门提供了 800 个观察结果的样本，在此期间，该客户的流程运行令人满意。这些数据的样本标准差为 0.21；因此，对于如此多的数据，假定总体标准差为 0.21。随后，质量人员建议定期随机抽取 30 个样本，以持续监测这一过程。通过分析新样品，客户可以快速了解工艺是否令人满意地运行。当过程不能令人满意地运行时，可以采取纠正措施来消除问题。设计规范表明该工艺的平均值应为 12。Quality associates 建议的假设检验如下。

$$H_0 : \mu = 12 \; H_a : \mu \neq 12$$

一旦 H_0 被拒绝，将采取纠正措施。

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.

```
## $s1
## $s1$p
## [1] 0.281
##
##
## $s2
## $s2$p
## [1] 0.4547
##
##
## $s3
## $s3$p
## [1] 0.00379
##
##
```

```
## $s4
## $s4$p
## [1] 0.03389
```

- 第一个样本的 p-value = 0.28，第二个样本的 p-value = 0.45，第三个样本的 p-value = 0.0038，第四个样本的 p-value = 0.034。
- 因此，样本三、四的 p 值小于 0.01，则拒绝零假设，表明该样本的平均值显著不同于 12，可能需要采取纠正措施。样本一、二的 p 值大于或等于 0.01，则不拒绝零假设，表明没有足够的证据表明该样本的平均值与 12 有显著差异。

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

```
## $s1
## [1] 0.2204
##
## $s2
## [1] 0.2204
##
## $s3
## [1] 0.2072
##
## $s4
## [1] 0.2061
```

```
## 总体标准差的 95% 置信区间为：( 0.2002 , 0.2208 )
```

- 假设总体的标准差为 0.21 是合理，因为四个样本的标准差都在总体标准差的置信区间内;

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.

```
## [1] 11.9 11.9
```

- 均值 =12 的置信区间是 (12, 12)

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?

```
## [1] 11.92 11.92
```

```
## [1] 11.94 11.94
```

- 第一类错误的概率增大。如果将增大到 0.05，那么我们错误地拒绝原假设的概率就增加到了 5%。在这种情况下，我们会更容易发现过程存在问题（即拒绝原假设），但同时也增加了误判的风险，即实际过程是令人满意的，却因为增大而错误地认为过程不令人满意。
- 如果犯第一类错误的概率增加，可能会导致不必要的纠正措施。例如，质量控制团队可能会对实际上运行良好的工艺进行调整，这不仅浪费了资源（如时间、人力、物力），还可能引入新的问题或干扰原本正常的工艺。可能会打断生产流程，增加生产成本，并且还可能影响产品的稳定性和质量，如果过度调整反而可能导致产品质量下降或者生产效率降低等问题。
- 结论：增大显著性水平，置信区间会变小；

Question 6

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

The proportion of units rented during the first week of March 2007 is 0.35;
The proportion of units rented during the first week of March 2008 is 0.4667;

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

两比例之差的区间估计:

$$\sigma_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{p_1 * (1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}}$$

$$ME = Z_{\alpha/2} \times \sigma_{\hat{p_1}-\hat{p_2}}$$

```
## [1] -0.22032 -0.01302
```

The interval is （-0.2203, -0.013）.

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

假设检验:

$$H_0 : P_{2008} - P_{2007} >= 0; H_a : P_{2008} - P_{2007} < 0$$

原假设为真，$P_{2008} = P_{2007} = P$,

$$\bar{p} = \frac{n_1\bar{p_1} + n_2\bar{p_2}}{n_1 + n_2}$$

$$\sigma_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

检验统计量为:

$$z = \frac{\bar{p_1} - \bar{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

```
## [1] 0.9863
```

```
## [1] 0.01373
```

- 对于单侧检验 p 值为 $0.0137 < 0.05$ ，在 0.05 的显著性水平下，拒绝 H0，认为 2008 年 3 月的入住比例显著比 2007 年 3 月的高。入住率提高，那么租金会上涨。

Question 7: Air Force Training Program

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.   :65.0   Min.   :69.0
##  1st Qu.:72.0   1st Qu.:74.0
##  Median :76.0   Median :76.0
##  Mean   :75.1   Mean   :75.4
##  3rd Qu.:78.0   3rd Qu.:77.0
##  Max.   :84.0   Max.   :82.0
```

```
## [1] 3.945
```

```
## [1] 2.506
```

- 可以看到中位数、均值均相差不大；实验组的最短用时和最长用时介于标准组的最短用时和最长用时之间，即实验组内学生的用时差距较小，实验组的标准差小于标准组；

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

```
## [1] 75.07
```

```
## [1] 75.43
```

```
##
##  Welch Two Sample t-test
##
## data:  Training$Current and Training$Proposed
## t = -0.6, df = 102, p-value = 0.5
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5477  0.8263
## sample estimates:
## mean of x mean of y
##     75.07     75.43
```

- 标准组的均值为 75.1，实验组的均值为 75.4；因此两者差异不大；
- 用 t 检验来检测下假设结果,p-value=0.5, 大于 0.05，表明没有足够的证据拒绝零假设，即两种方法的均值没有显著差异。

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.

```
## $Current
## [1] 3.945
##
## $Proposed
## [1] 2.506
```

```
## $Current
## [1] 15.56
##
```

```
## $Proposed
## [1] 6.282

##
##  F test to compare two variances
##
## data:  Training$Current and Training$Proposed
## F = 2.5, num df = 60, denom df = 60, p-value = 0.0006
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.486 4.129
## sample estimates:
## ratio of variances
##              2.477
```

- F 检验的零假设是两个样本的方差相等。p-value = 0.0006, 小于 0.05，表明两个训练方法的方差存在显著差异，表明两种方法在变异性上有所不同。

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

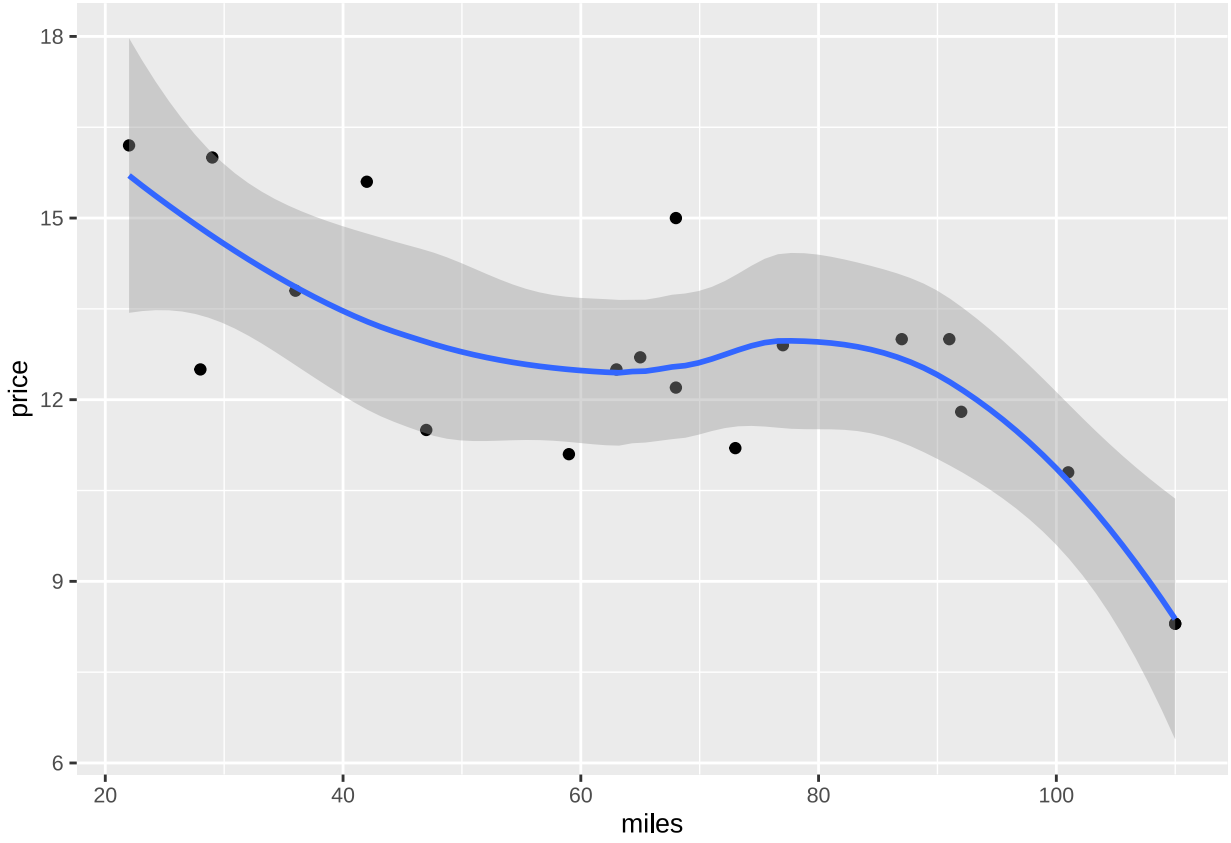
- 结论两种方法的均值相似, 但 Proposed 方法的方差更低, 表明 Proposed 方法更稳定, 可以考虑使用 Proposed 方法。
- 建议：建议基于上述分析结果，如果 Proposed 方法在均值上表现更好且方差没有显著增加，那么可以推荐使用 Proposed 方法。如果 Proposed 方法的方差显著增加，需要进一步调查原因。如果两种方法在统计上没有显著差异，可以维持现状，同时继续监控过程性能。

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?

- 可扩展性测试：测试培训方法的可扩展性，确定它们是否适用于不同规模的团队或组织。
- 对比实验：将两种方法应用于相似但独立的群体，以更准确地比较效果。

Question 8

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?

- 两者之间是负相关的关系，当一辆车的行驶里程达到一定程度时，其价值会变得非常小

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

```
##
## Call:
## lm(formula = price ~ miles, data = Camry)
##
```

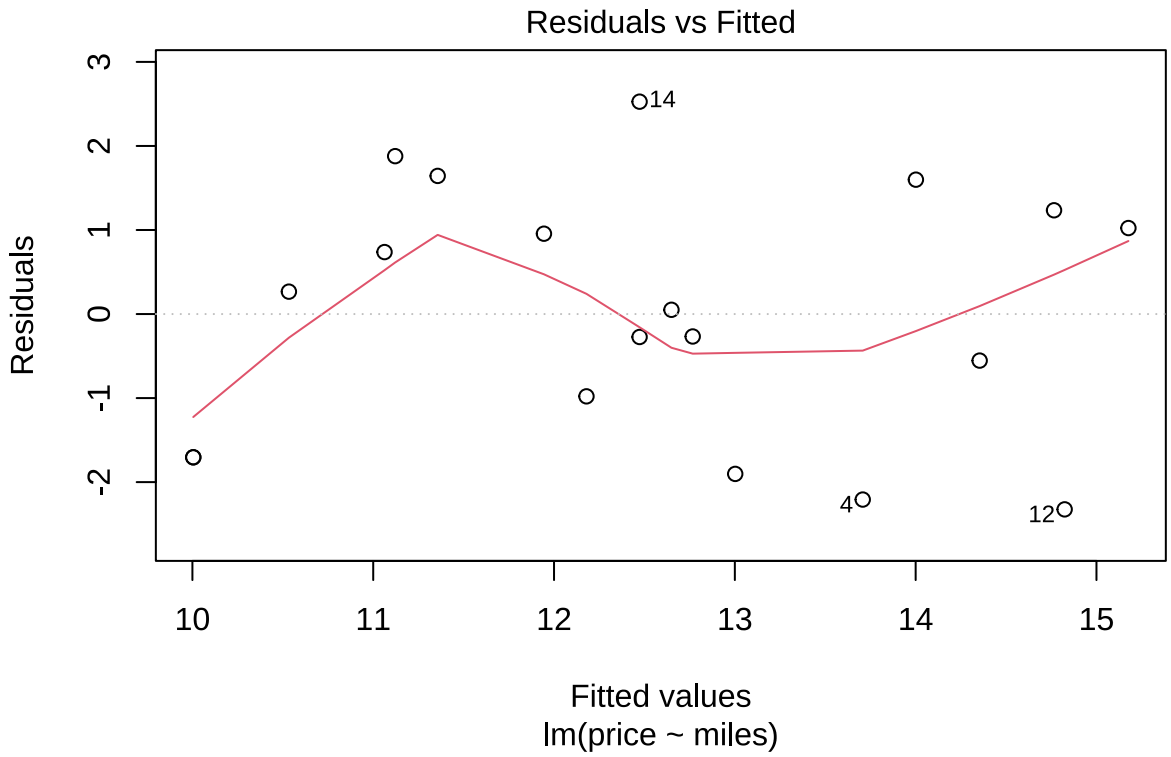
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3241 -1.3419  0.0506  1.1290  2.5269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.4698     0.9488   17.36    3e-12 ***
## miles        -0.0588     0.0132   -4.46  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.54 on 17 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.512
## F-statistic: 19.8 on 1 and 17 DF,  p-value: 0.000348

## 截距: 16.47

## 斜率: -0.05877
```

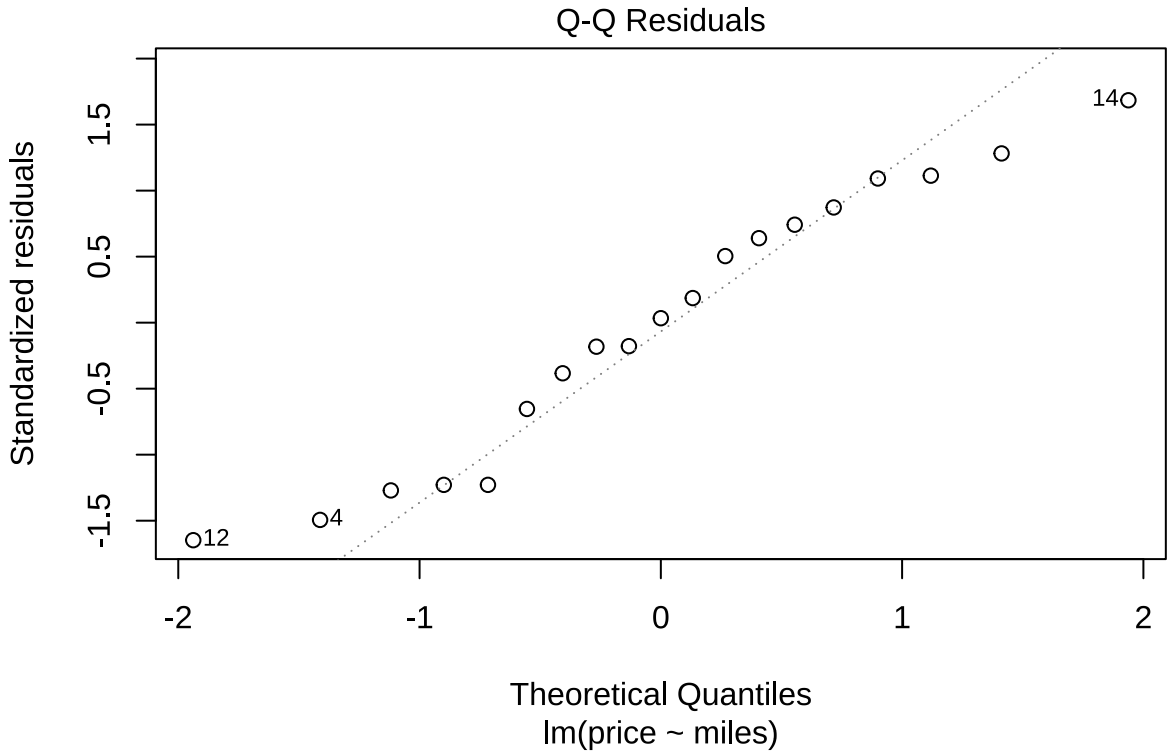
$$Price = 16.47 - 0.05877 * miles$$

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

- p-value: 0.000348 , 小于显著性水平 0.05，则拒绝零假设, 说明 miles 与 pirce 存在一个显著的关系.

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





- R-squared: 0.539, (R 平方值越接近 1，表示模型的拟合度越好)

- 根据上述分析，R 平方值较接近 1，F 统计量的 p 值小于 0.05，且残差图没有显示出明显的模式，我们可以得出结论，估计的回归方程提供了良好的拟合。

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

- x 值增加一个单位时，y 值将减少 0.059。由于数据是以千为单位记录的，所以每增加 1000 英里的车程，预测价格将减少 59.0 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.

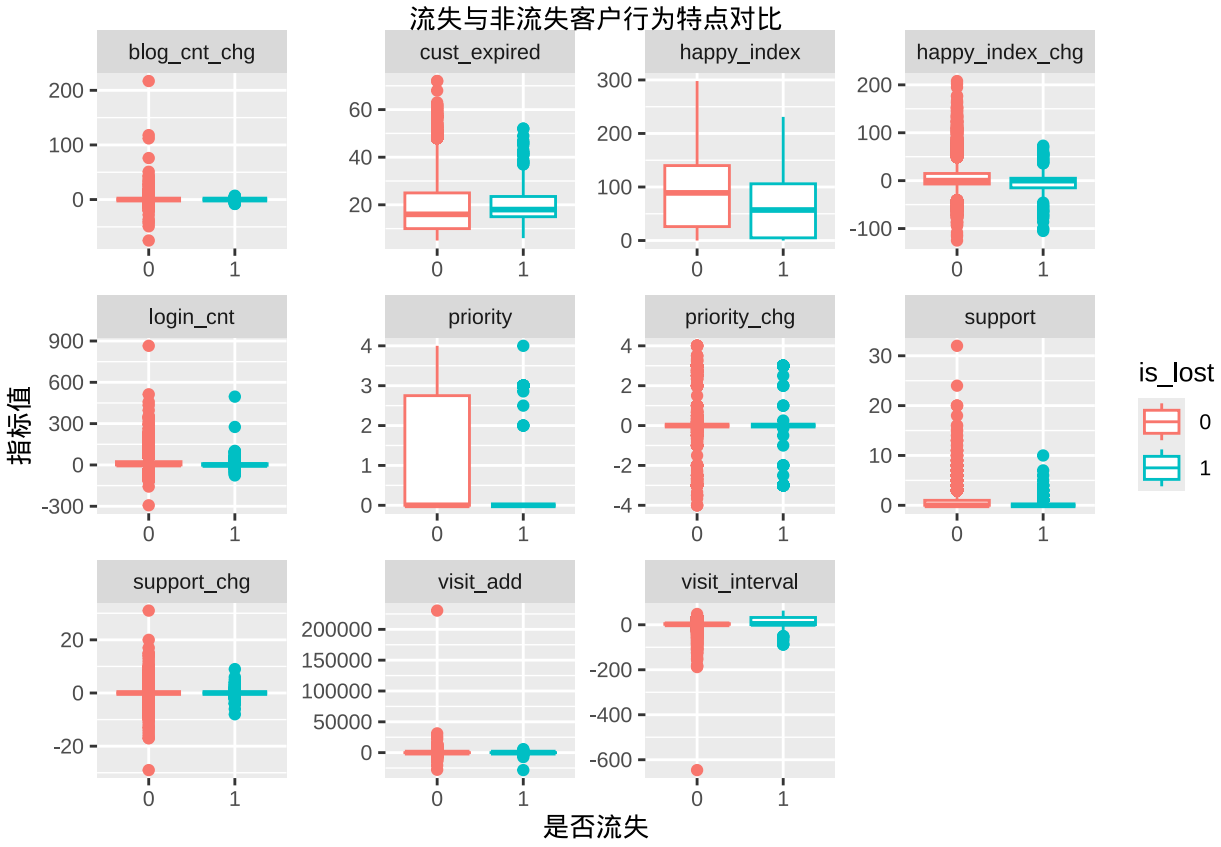
```
##     1
## 12.94
```

- 预测价格是 $= 16.47 - .058877(60) = 12.94$，即 12940 美元，但汽车的价格还会受到各种其他因素（市场行情、汽车使用情况等）的影响，因此该价格只能用来参考；

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？



- 流失与非流失客户行为在 11 个指标都有可能存在显著不同；（箱线图分布均有所不同）

b. 通过均值比较的方式验证上述不同是否显著。

```
##                t_value p_value mean_group_0 mean_group_1 Significant
## happy_index     7.6242  0.0000    88.605910     63.27245         Yes
## happy_index_chg 5.7835  0.0000     5.530212     -3.73684         Yes
## support         5.5099  0.0000     0.724270      0.37152         Yes
## support_chg    -0.6320  0.5278    -0.009296      0.03715          No
## priority        5.1428  0.0000     0.829576      0.49956         Yes
## priority_chg    0.6412  0.5218     0.032682     -0.01670          No
## login_cnt       3.5709  0.0004    16.138944      8.06192         Yes
## blog_cnt_chg    2.5315  0.0116     0.171149     -0.10217         Yes
## visit_add       1.9136  0.0563   106.609562    -95.76780          No
## cust_expired   -2.9811  0.0031    18.818725     20.35294         Yes
```

```
## visit_interval  -4.0971  0.0001     3.511454      8.48607         Yes
```

- 在显著性水平为 0.05 的条件下，流失客户与非流失客户在当月客户幸福指数、客户幸福指数相比上月变化、当月客户支持、当月服务优先级、当月登录次数、博客数相比上月的变化、客户使用期限、访问间隔变化这几个指标下均值有显著性差异。
- 客户支持相比上月的变化 (support_chg)、服务优先级相比上月的变化（priority_chg)、访问次数相比上月的增加（visit_add）这几个指标下均值没有显著性差异;

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 = is_lost ~ happy_index + happy_index_chg + support +
##     priority + login_cnt + blog_cnt_chg + cust_expired + visit_interval,
##     family = binomial, data = selected_vars)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.876333   0.121259  -23.72  < 2e-16 ***
## happy_index     -0.005199   0.001156   -4.50  6.9e-06 ***
## happy_index_chg -0.009306   0.002412   -3.86  0.00011 ***
## support         -0.022169   0.071455   -0.31  0.75637
## priority        -0.044752   0.074135   -0.60  0.54607
## login_cnt        0.000854   0.001938    0.44  0.65921
## blog_cnt_chg    -0.000972   0.020510   -0.05  0.96221
## cust_expired     0.014256   0.005240    2.72  0.00651 **
## visit_interval   0.016950   0.004279    3.96  7.4e-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
##
## Number of Fisher Scoring iterations: 6
```

- Intercept (截距): -2.876333 表示当所有自变量取值为 0 时，客户流失的基准概率。这个值对应的概率可以通过将其转化为概率值（使用 logit = exp(Intercept) / (1 + exp(Intercept))）计算出来，约为 0.05334（即客户流失的概率是 53.34%）。
- p value (p 值) 小于 0.05，说明截距对模型是显著的。
- 每个指标的回归系数表示该指标对客户流失概率的影响程度。回归系数的符号（正负）告诉我们自变量与客户流失之间的关系：正系数：表示该指标的值增加会增加客户流失的概率（即流失的可能性增加）。负系数：表示该指标的值增加会减少客户流失的概率（即流失的可能性降低）。
- 以 happy_index_chg 指标为例：回归系数为 -0.009306，p 值为 0.00011，小于 0.05，说明客户幸福指数相比上月变化对客户流失有显著的负向影响，即每增加 1 单位，客户流失的概率会下降；可能表明该指标的增加有助于降低流失率（例如，客户满意度较高时流失率较低）

d. 根据上一步预测的结果，对尚未流失（不流失 **=1**）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	is_lost	predictions
1363	1	0.1920
1672	1	0.1803
299	1	0.1745
2922	1	0.1624
2951	1	0.1618
1021	1	0.1589
335	1	0.1564
156	1	0.1541
1488	1	0.1470
3340	1	0.1454
2296	1	0.1449
1069	1	0.1380
3811	1	0.1363
2653	1	0.1346
3604	1	0.1338
1405	1	0.1328
2636	1	0.1313
2077	1	0.1308
1987	1	0.1297
4292	1	0.1297
2082	1	0.1269

1782	1	0.1265
2766	1	0.1257
2166	1	0.1250
2120	1	0.1234
1303	1	0.1231
904	1	0.1231
3092	1	0.1226
2624	1	0.1223
2371	1	0.1219
2928	1	0.1219
1563	1	0.1204
1711	1	0.1189
1532	1	0.1171
891	1	0.1159
945	1	0.1159
947	1	0.1159
948	1	0.1159
2084	1	0.1157
896	1	0.1144
402	1	0.1139
227	1	0.1130
979	1	0.1130
938	1	0.1127
2902	1	0.1122
257	1	0.1116
317	1	0.1116
363	1	0.1116
371	1	0.1116
523	1	0.1116
543	1	0.1116
548	1	0.1116
787	1	0.1116
1214	1	0.1116
1760	1	0.1116
3312	1	0.1116
3313	1	0.1116

4500	1	0.1116
3569	1	0.1106
1659	1	0.1102
3163	1	0.1102
3235	1	0.1102
3349	1	0.1102
3228	1	0.1099
3267	1	0.1099
640	1	0.1098
5312	1	0.1091
930	1	0.1088
3772	1	0.1088
3363	1	0.1074
3487	1	0.1072
2179	1	0.1067
4280	1	0.1063
2707	1	0.1062
4273	1	0.1061
3861	1	0.1055
2189	1	0.1054
4263	1	0.1049
4289	1	0.1049
4291	1	0.1047
453	1	0.1039
1831	1	0.1038
4680	1	0.1034
3584	1	0.1019
2316	1	0.1015
3317	1	0.1009
5052	1	0.0995
1523	1	0.0951
1709	1	0.0950
1909	1	0.0949
387	1	0.0941
5313	1	0.0936
412	1	0.0936

105	1	0.0935
4893	1	0.0927
1823	1	0.0926
2003	1	0.0911
1456	1	0.0911
430	1	0.0909
5460	1	0.0885

- 以上是尚未流失可能性最高的 30 名客户的表格。应针对这些客户采取客户保留措施。