第二次作业

舒明

2024-11-28

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

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the 2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (*the Big Bang theory* website, April 17, 2012).

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

```
## min : 13.3
```

```
## max : 16.5
```

b. Compute the mean, median, and mode.

```
## mean : 15.04286
```

```
## median : 15
```

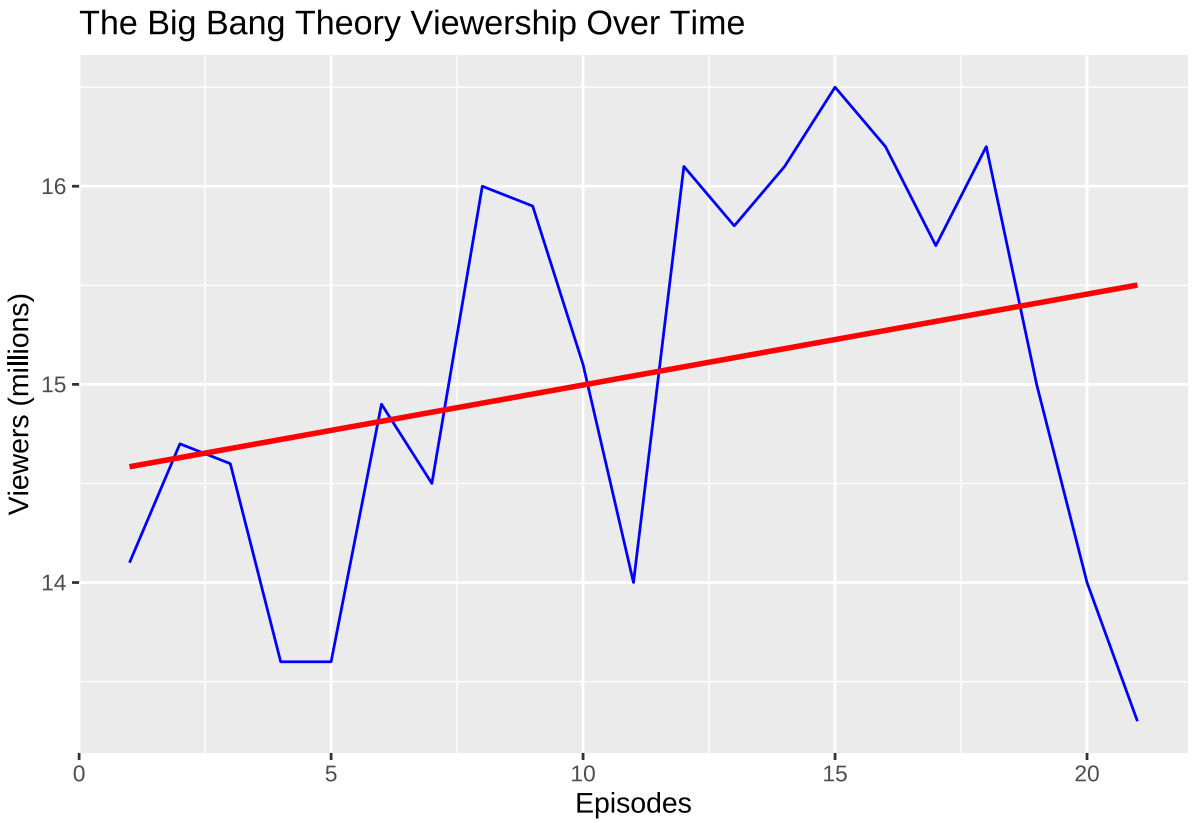
```
## mode : 13.6
```

c. Compute the first and third quartiles.

```
## Q1 : 14.1
```

```
## Q3 : 16
```

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




```
##
##  Mann-Kendall trend test
##
## data:  data$`Viewers (millions)`
## z = 1.4219, n = 21, p-value = 0.1551
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
##   48.0000000 1092.6666667    0.2307799
```

收视率从短期看波动较大，有涨有跌，呈现出不稳定趋势；长期看呈现出轻微的上升趋势——但这种趋势极大概率是由于随机波动造成的，而不是一个显著的长期趋势。

解释：

1、收视率和观众人数

我理解的收视率（*viewership*）和观众人数（*viewers*）是两个概念。

收视率的常见统计方式如下：

$$收视率 = (实际观看人数 / 总潜在观众人数) \times 100\%$$

本题的观众人数（*viewers*）是实际观看人数，但每集的总潜在观众人数未知。如果总潜在观众人数基本固定，那么可认为数据集的时间区间（*September 22, 2011 ~ April 5, 2012*）内，收视率（*viewership*）

趋势等同于观众人数（*viewers*）趋势。只有基于这个前提，才能讨论收视率的趋势；否则无法得出结论。

2、折线图和拟合线

折线图可直观地看出短期趋势，有涨有跌；长期趋势从拟合线（线性）上看，呈上涨趋势，但有待进一步检验是否存在显著趋势。

3、*Mann-Kendall* 趋势检验

由于数据集是时序数据，我们可以通过 *Mann-Kendall* 趋势检验来判断是否存在显著趋势。

检验假设如下：

H_0 ：数据中不存在趋势。H_1 ：数据中存在趋势。

如果检验的 p 值低于一定的显着性水平（一般为 *0.05*），则有统计显著性证据表明时间序列数据中存在趋势。而从检验结果来看，S 值（*48.0000000*）为正，代表趋势为上升趋势；*tau* 值（*0.2307799*）接近于 *0*，表示趋势较弱或几乎不存在；p 值（*0.1551*）大于 *0.05*，表明在统计上我们不能拒绝零假设，即数据中没有显著的趋势。

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

CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.

a. Show the frequency distribution.

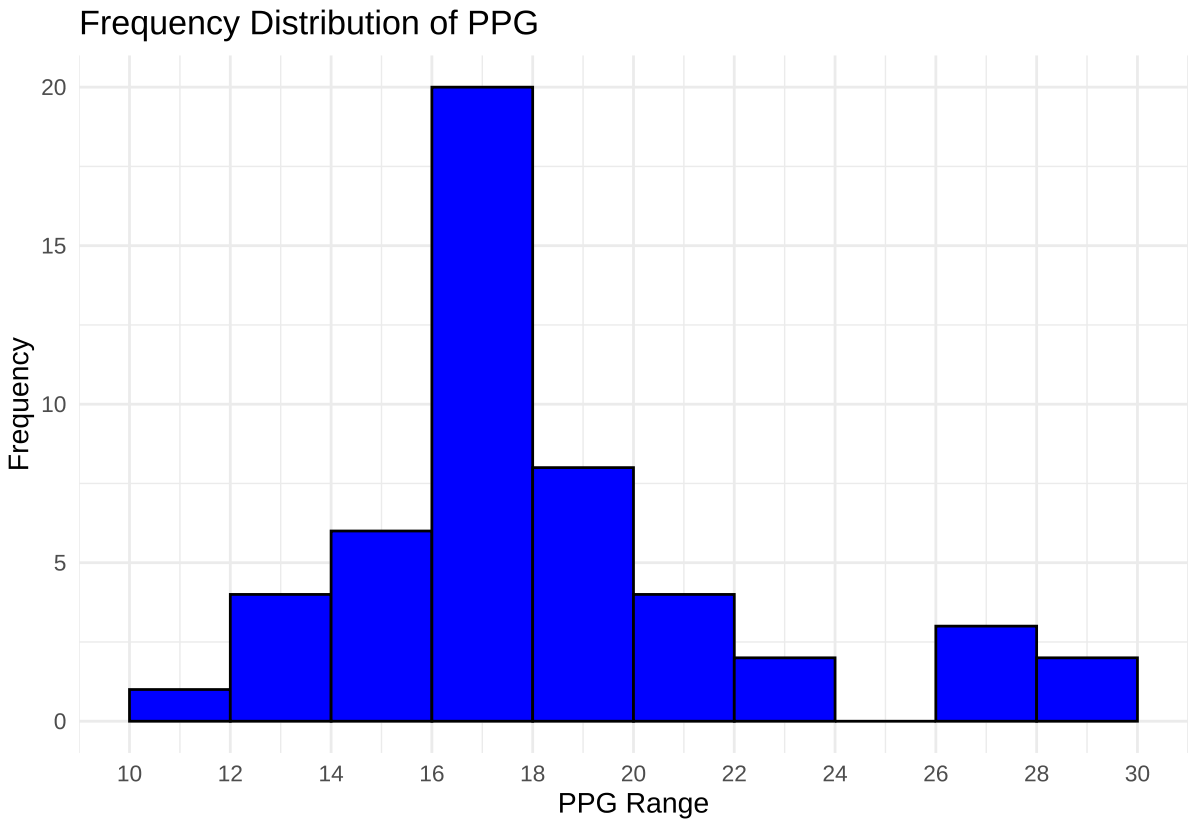


表 1: Frequency Distribution of PPG

PPG Range	Frequency
[10,12]	1
(12,14]	4
(14,16]	6
(16,18]	20
(18,20]	8
(20,22]	4
(22,24]	2
(24,26]	0
(26,28]	3
(28,30]	2

b. Show the relative frequency distribution.

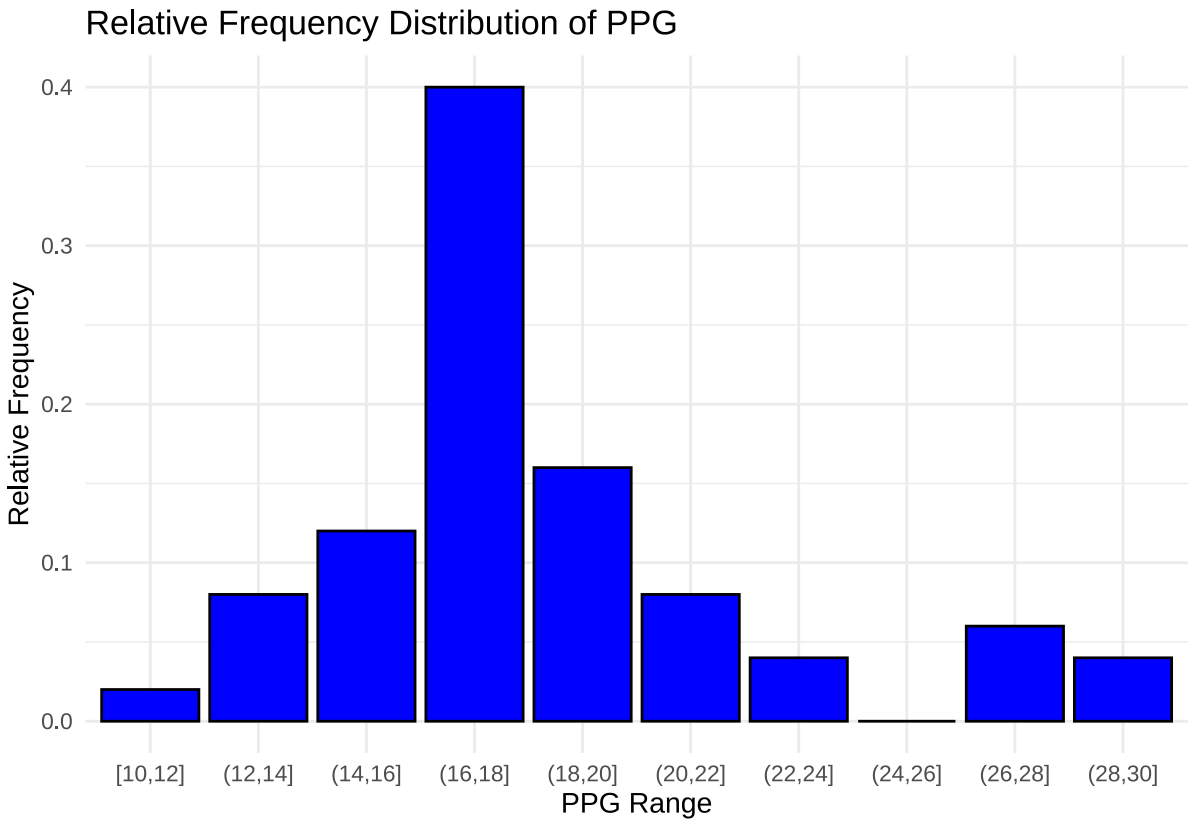


表 2: Relative Frequency Distribution of PPG

PPG Range	Frequency	Relative_Frequency
[10,12]	1	0.02
(12,14]	4	0.08
(14,16]	6	0.12
(16,18]	20	0.40
(18,20]	8	0.16
(20,22]	4	0.08
(22,24]	2	0.04
(24,26]	0	0.00
(26,28]	3	0.06
(28,30]	2	0.04

c. Show the cumulative percent frequency distribution.

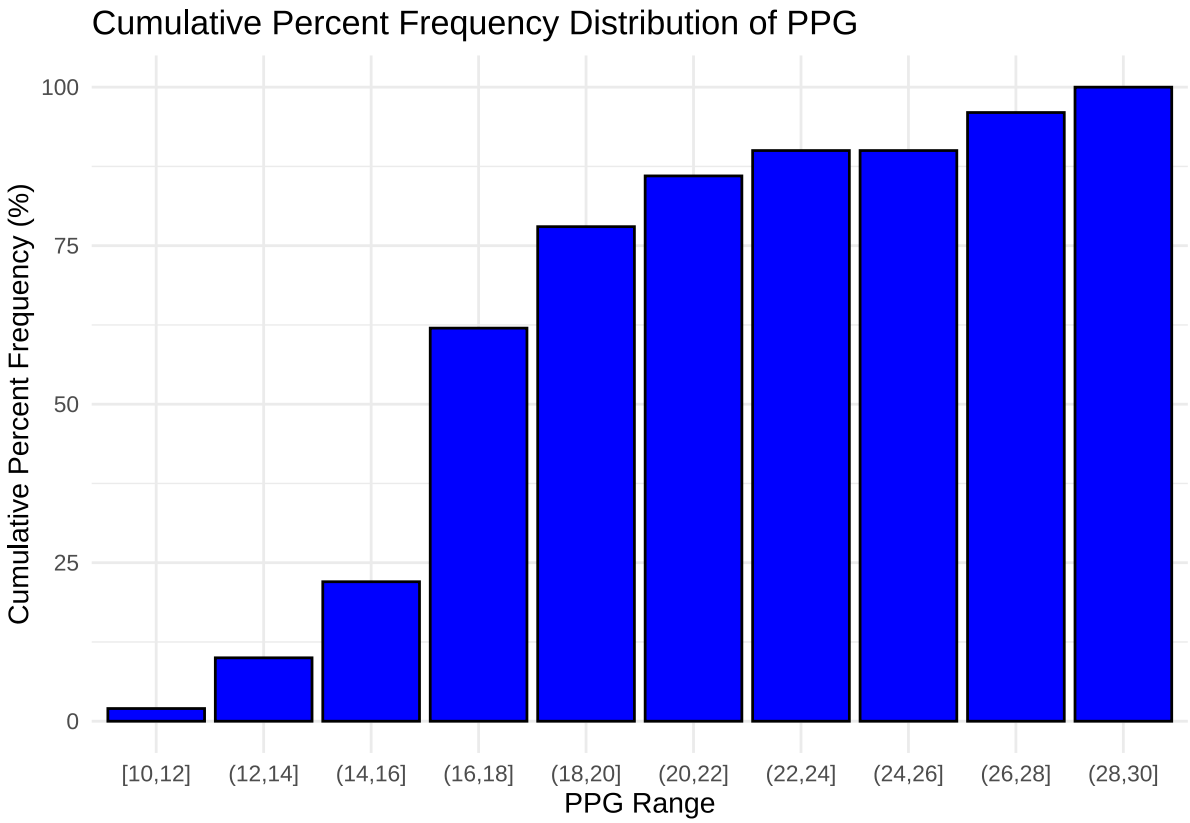
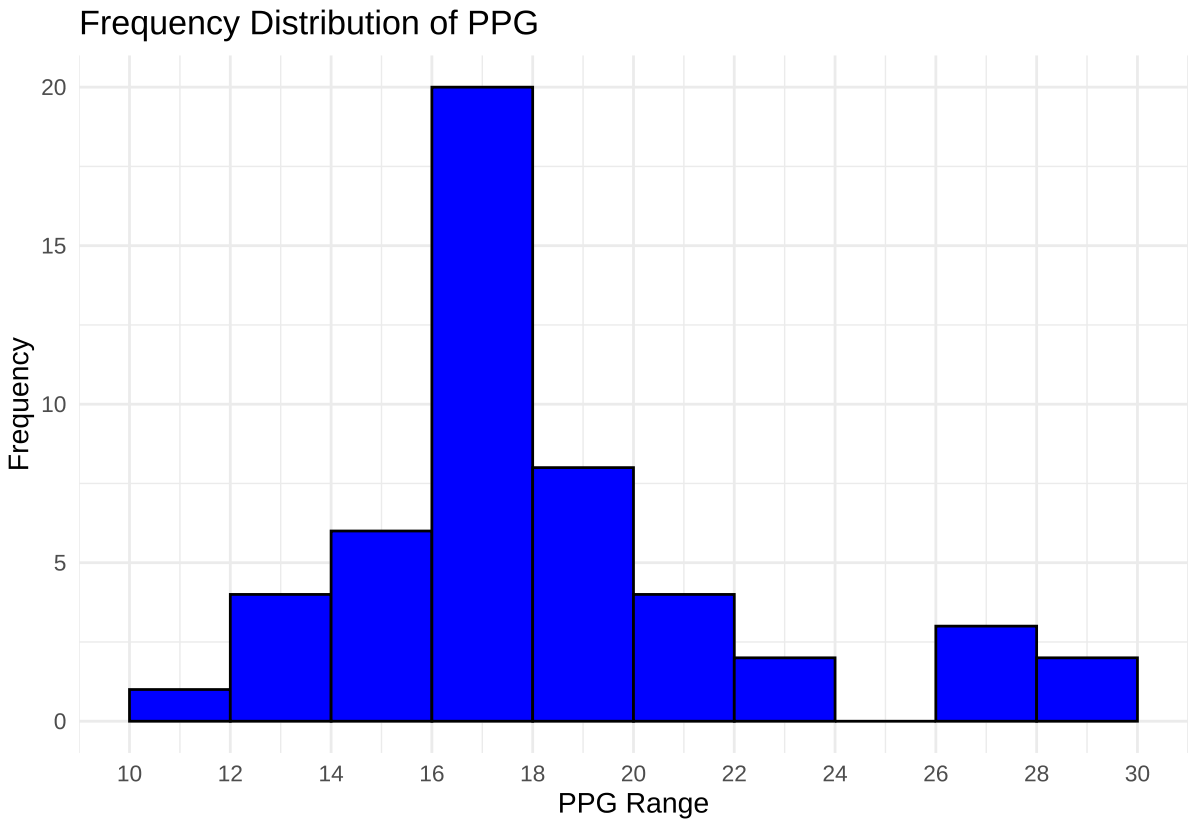


表 3: Cumulative Percent Frequency of PPG

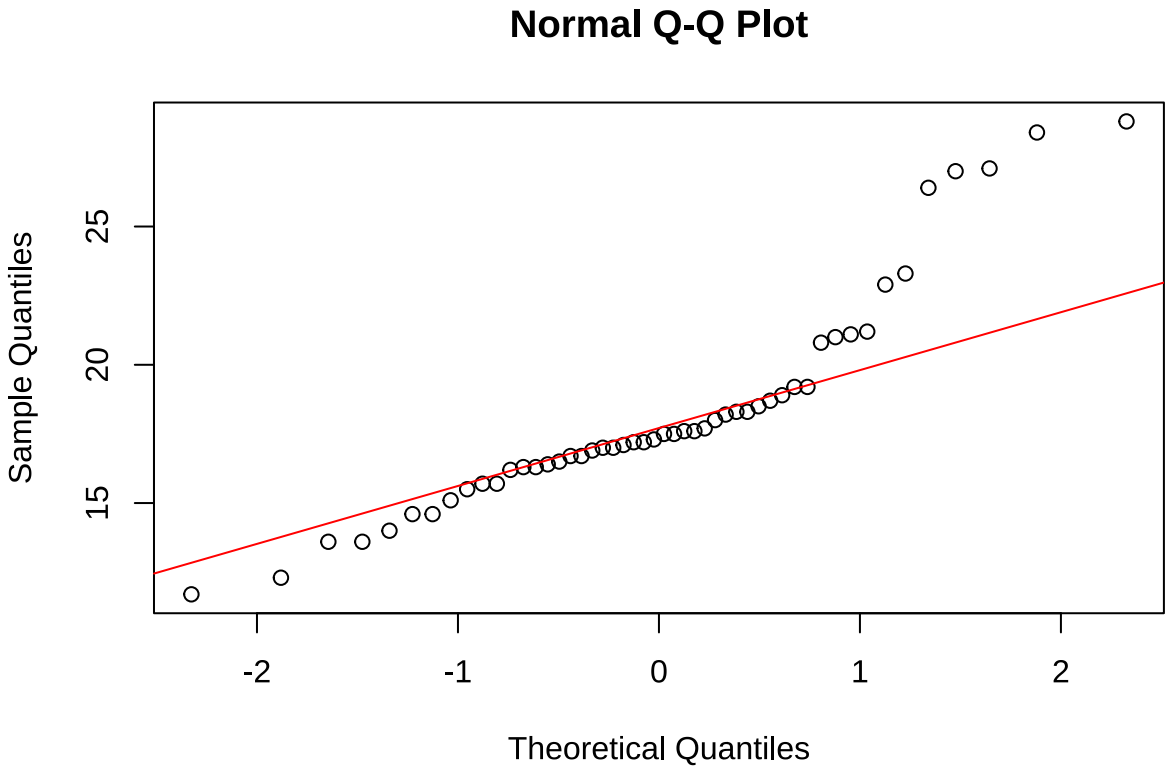
PPG Range	Frequency	Relative_Frequency	Cumulative_Percent_Frequency
[10,12]	1	0.02	2%
(12,14]	4	0.08	10%
(14,16]	6	0.12	22%
(16,18]	20	0.40	62%
(18,20]	8	0.16	78%
(20,22]	4	0.08	86%
(22,24]	2	0.04	90%
(24,26]	0	0.00	90%
(26,28]	3	0.06	96%
(28,30]	2	0.04	100%

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



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

```
## skewness:  1.124025
```



场均得分数据整体呈右偏（正偏）——即大多数球员的得分集中在较低区间，而少数球员（明星球员）得分非常高，导致整体分布右偏。

解释：

- 偏度值：偏度值大于 *0*（*1.124025*），表明数据呈右偏（正偏）。这意味着数据的长尾在右侧，即极端值偏向于较大的数值。
- *Q-Q* 图：红色的对角线表示如果样本数据完全符合正态分布，样本分位数应该落在这条线上——如果数据点大致沿着参考线排列，则样本数据接近正态分布。但从图中可以看到，在左侧（负的理论分位数）和右侧（正的理论分位数）都有明显的偏离，尤其是右侧的偏离更为明显。这进一步支持了数据向右偏斜的观点。

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

```
## 22 %
```

解释：从累计频率分布也可直观看出（*100%* 减去 *(18,20]* 对应的累计频率 *78%*）。

Question #3: 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?

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

解释：

标准误差（*Standard Error, SE*）的公式：

$$SE = \frac{\sigma}{\sqrt{n}}$$

其中 SE 是标准误差，σ 是总体标准差，n 是样本大小。则：

$$n = \left(\frac{\sigma}{SE}\right)^2$$

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

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

解释：

为了找到点估计（样本均值）在总体均值 *±25* 范围内的概率，考虑 *n=625* 是一个相当大的样本，我们可以直接使用正态分布的性质。 首先，算范围 *±25* 内的 *z -score*：

$$z = \frac{\bar{x} - \mu}{SE}$$

其中： $\bar{x}$ 是样本均值，μ 是总体均值， SE 是均值的标准误差。
对于 $x = \mu + 25$：

$$z_1 = \frac{(\mu + 25) - \mu}{20} = \frac{25}{20} = 1.25$$

对于 $x = \mu - 25$:

$$z_2 = \frac{(\mu - 25) - \mu}{20} = \frac{-25}{20} = -1.25$$

故只需求z 在 *-1.25* 和 *1.25* 之间的概率为: $P(-1.25 < Z < 1.25) = P(1.25) - P(-1.25)$

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

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *young Professionals*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

1. What is your age?
2. Are you: Male_____ Female_____
3. Do you plan to make any real estate purchases in the next two years?

   Yes_____ No_____
4. What is the approximate total value of financial investments, exclusive of your

   home, owned by you or members of your household?
5. How many stock/bond/mutual fund transactions have you made in the past year?
6. Do you have broadband access to the Internet at home? Yes_____ No_____
7. Please indicate your total household income last year. _____
8. Do you have children? Yes_____ No_____

The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of

interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

a. Develop appropriate descriptive statistics to summarize the data.

表 4: Data summary

Name	professional
Number of rows	410
Number of columns	14
Column type frequency:	
character	5
logical	5
numeric	4
Group variables	None

Variable type: character

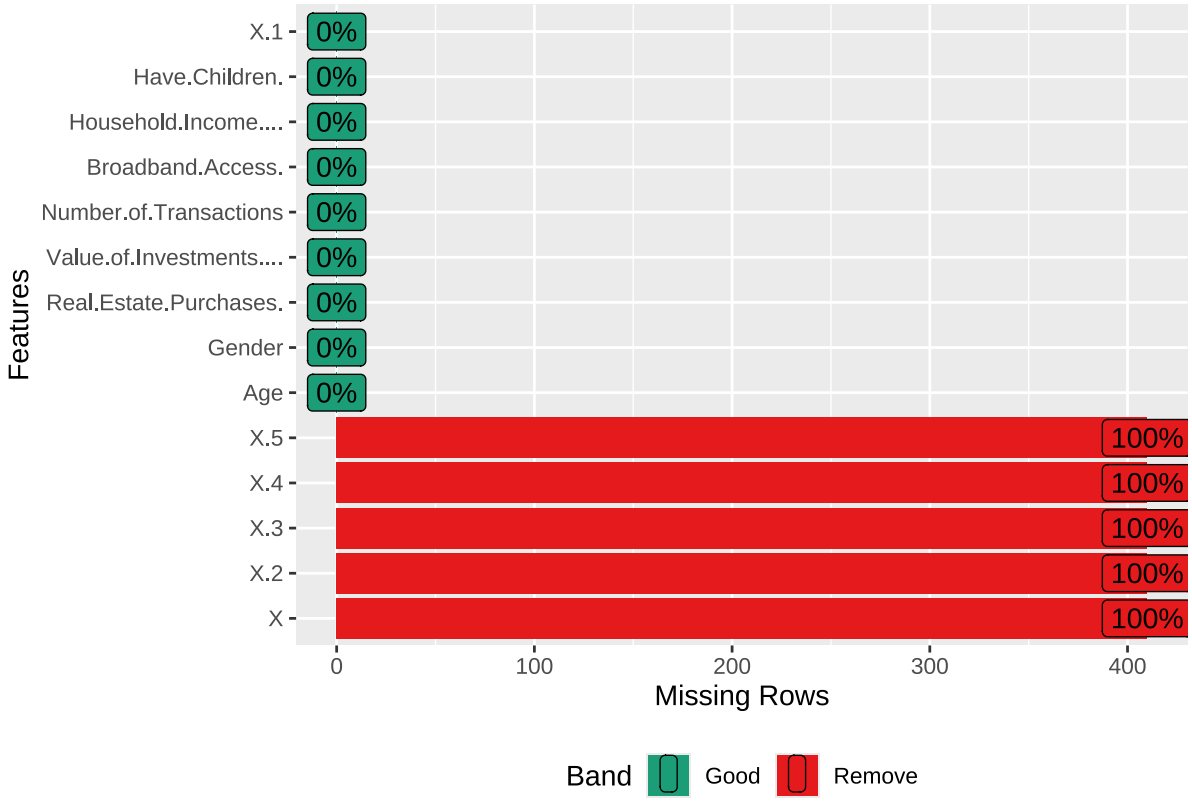
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	4	6	0	2	0
Real.Estate.Purchases.	0	1	2	3	0	2	0
Broadband.Access.	0	1	2	3	0	2	0
Have.Children.	0	1	2	3	0	2	0
X.1	0	1	0	1	409	2	0

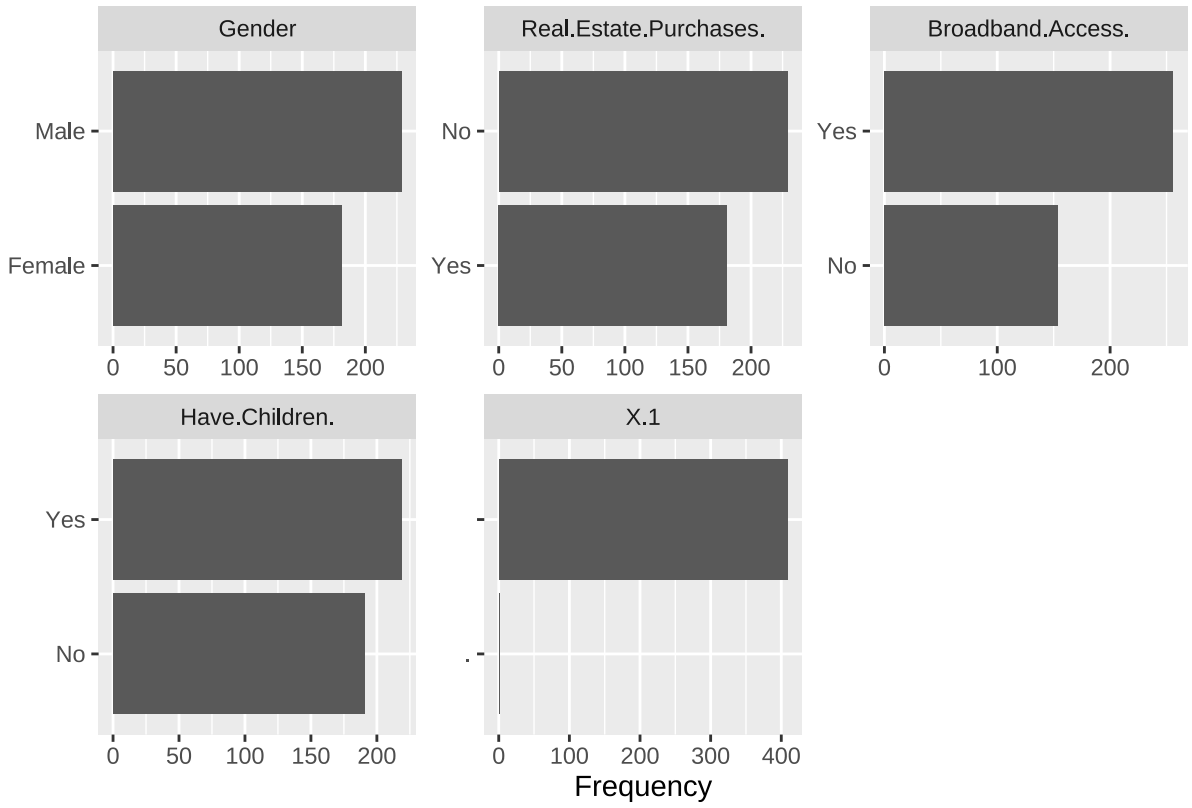
Variable type: logical

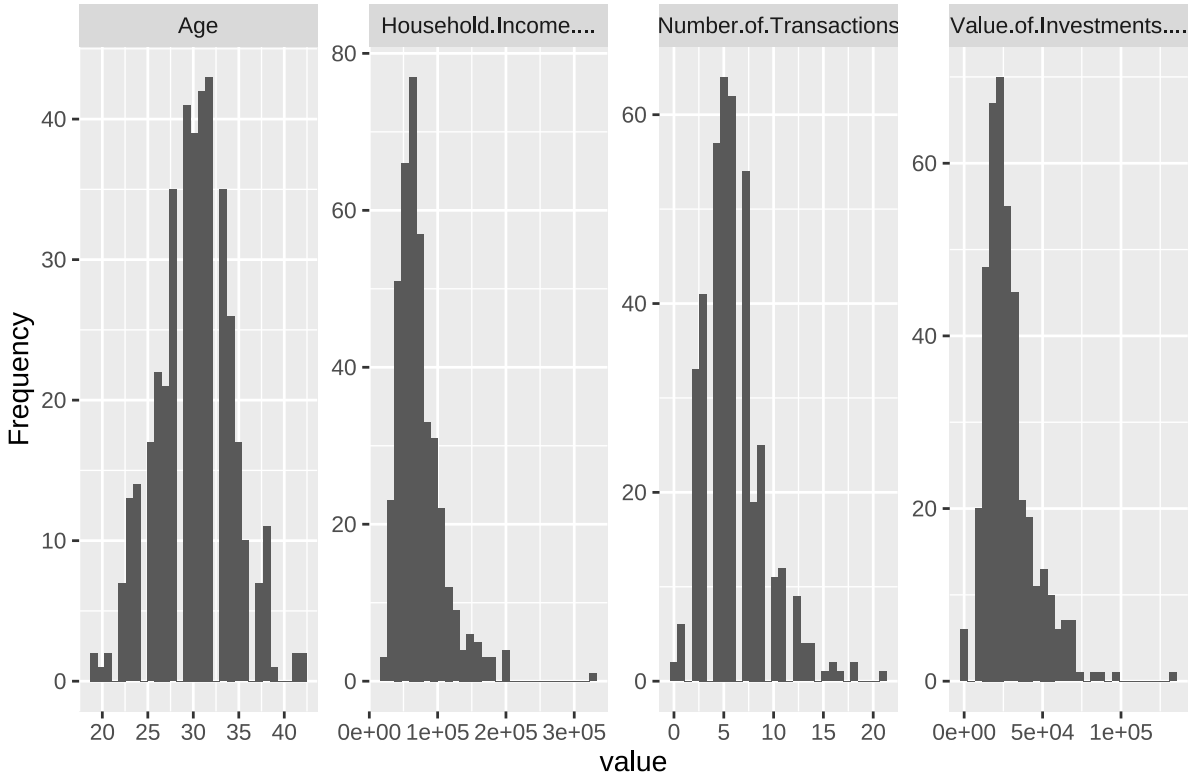
skim_variable	n_missing	complete_rate	mean	count
X	410	0	NaN	:
X.2	410	0	NaN	:
X.3	410	0	NaN	:
X.4	410	0	NaN	:
X.5	410	0	NaN	:

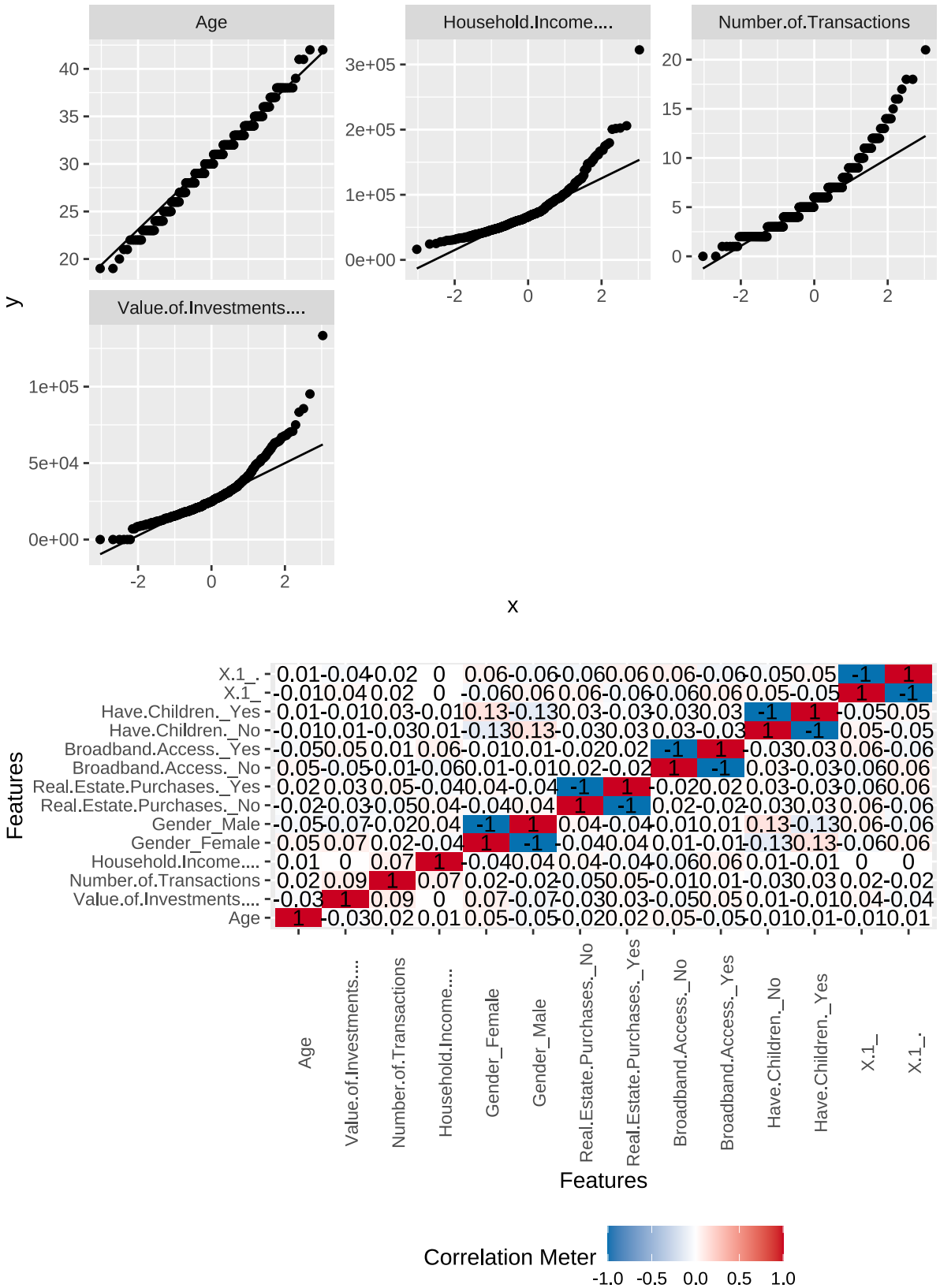
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	30.11	4.02	19	28	30	33	42	
Value.of.Investments....	0	1	28538.29	15810.83	0	18300	24800	34275	133400	
Number.of.Transactions	0	1	5.97	3.10	0	4	6	7	21	
Household.Income....	0	1	74459.51	34818.21	16200	51625	66050	88775	322500	




Gender
Male
Female
0
50
100
150
200
Real.Estate.Purchases.
No
Yes
0
50
100
150
200
Broadband.Access.
Yes
No
0
100
200
Have.Children.
Yes
No
0
50
100
150
200
X.1
.
0
100
200
300
400
Frequency


Age
Household.Income....
Number.of.Transactions
Value.of.Investments....
Frequency
0
10
20
30
40
20
25
30
35
40
0
20
40
60
80
0e+00
1e+05
2e+05
3e+05
0
20
40
60
0
5
10
15
20
0
20
40
60
0e+00
5e+04
1e+05
value



解释：通过 *skimr* 和 *DataExplorer* 两个包可快速做描述性统计。

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

```
##
##  One Sample t-test
##
## data:  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:  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
```

```
## 95% ci for the mean age:  29.72153 30.50286
```

```
## 95% ci for the mean household income of subscribers:  71079.26 77839.77
```

解释：这里使用 *t-test* 而非 *z-test* 做单样本检验——虽然样本足够大（*>30*），但总体方差未知。直接取 *conf.int* 作为结果。

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:  sum(professional$Broadband.Access. == "Yes") out of nrow(professional), null probabi
## 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:  sum(professional$Have.Children. == "Yes") out of nrow(professional), null probabilit
## 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

## 95% ci for the proportion of subscribers who have broadband access at home:   0.5753252 0.6

## 95% ci for the proportion of subscribers who have children:  0.4845521 0.5830908
```

解释：这里使用 *prop.test* 来执行单样本比例检验，直接取 *conf.int* 作为结果。

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

```
##
##  One Sample t-test
##
## data:  professional$Number.of.Transactions
## t = 39.004, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.672129 6.274213
## sample estimates:
## mean of x
##  5.973171

##
##  One Sample t-test
##
## data:  professional$Value.of.Investments....
## t = 36.548, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  27003.33 30073.26
## sample estimates:
## mean of x
##  28538.29
```

是的。从之前的描述性统计和推断性统计我们可以看出，对于读者总体，可得出如下合理推测：

- 投资价值（Value of Investments）、交易次数（Number of Transactions）、家庭收入（Household Income）均呈现明显明显右偏的趋势，证明杂志有相当一部分读者可能是高净值个人，有较高的收入水平和投资活动，他们可能对更复杂的金融产品和服务感兴趣，对于在线经纪人而言，相当于有了大客户基础。
- 年龄（Age）的均值为 30 左右，说明读者群体相对年轻，处于职业生涯的早期阶段，他们更有可能尝试一些新鲜事物，例如在线经纪人服务。
- 宽带接入（Broadband Access）在 6 成左右，说明大多数读者拥有宽带接入，对技术较为熟悉，更可能便捷去使用在线经纪人服务。

综上，我认为《Young Professional》是一个不错的广告渠道，特别是针对在线经纪人的服务。

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

是的。从之前的描述性统计和推断性统计我们可以看出，对于读者总体，可得出如下合理推测：

- 从有小孩的情况（Have Children）看，半数左右的家庭是有小孩的。这意味着有相当一部分读者可能是年轻家庭的父母，他们可能对教育软件和儿童电脑游戏感兴趣。
- 家庭收入（Household Income）的均值（74460）较高，这表明读者群体有较强的购买力，能够负担教育软件和儿童电脑游戏的费用。
- 宽带接入（Broadband Access）在 6 成左右，说明大多数读者可能更倾向于使用数字产品和服务，包括教育软件和在线游戏。
- 年龄（Age）的均值为 30 左右，说明读者群体相对年轻，他们可能更愿意为孩子寻找现代的教育工具和娱乐方式。
- 投资价值（Value of Investments）和交易次数（Number of Transactions）也能看出读者都有一定的投资基础和交易活动，他们可能对新技术和市场趋势较为敏感，这可能包括对教育科技产品的关注，以及对小孩的投资。

综上，我认为《Young Professional》杂志是教育软件和儿童电脑游戏公司的一个不错的广告投放渠道。

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

结合前面的描述性分析，以下三类文章我认为会更吸引读者：

- **职业发展和职场技能**
  鉴于杂志的目标受众是年轻的专业人士，文章可以关注如何提升职业技能、职场晋升策略、工作

与生活平衡等主题。

- **财务管理和投资**
  读者群体对投资和交易有一定的参与度，因此，提供关于个人财务管理、投资策略、退休规划等方面的文章可能会受到欢迎。
- **家庭和育儿**
  有孩子的读者是大多数，文章可以包括育儿技巧、家庭财务管理、如何平衡工作和家庭生活等主题。

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

Quality associates, inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. in one particular application, a client gave Quality associates a sample of 800 observations taken during a time in which that client's process was operating satisfactorily. the sample standard deviation for these data was .21; hence, with so much data, the population standard deviation was assumed to be .21. Quality associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. by analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. when the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. the design specification indicated the mean for the process should be 12. the hypothesis test suggested by Quality associates follows.

$$H_0 : \mu = 12 H_1 : \mu \neq 12$$

Corrective action will be taken any time H_0 is rejected.

Data are available in the data set Quality.

Managerial Report

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.

```
##     sample     p_value
## 1 Sample.1 0.281008276
## 2 Sample.2 0.454650325
## 3 Sample.3 0.003790318
## 4 Sample.4 0.033893355
```

从 p-value 可以看出，**只有 Sample 3 有足够的证据拒绝原假设（p-value<0.01），即有理由认为其平均值与设计规范的平均值 12 有显著差异，需要采取纠正措施**；其余的 Sample（1、2 和 4）没有足够的证据拒绝原假设，不需要采取纠正措施。

解释：数据量为 *30*，且总体标准差已知，用 z 检验。

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

```
##                    Sample.1          Sample.2          Sample.3
## sd        0.220356033748104 0.220356033748104 0.207170594371918
## p_value   0.645841249324365 0.645841249324369  1.01194315096676
## reject_H0             FALSE             FALSE             FALSE
##                    Sample.4
## sd        0.206108999173325
## p_value    1.04274910816834
## reject_H0             FALSE
```

设原假设和备选假设如下：

- $H0 : \sigma = 0.21$
- $H1 : \sigma \neq 0.21$

从卡方检验结果可以看出，所有四个样本的卡方检验的 p 值都大于 0.01，因此不拒绝原假设，认为样本标准差与假设的总体标准差 0.21 没有显著差异。这表明**假设的总体标准差 0.21 是合理的**。

解释：用卡方检验来看样本标准差和假设的总体标准差的差异。

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_limit upper_limit
## 1    11.90124    12.09876
```

解释：查找临界值-> 计算误差边界（$E = Z \times SE$）-> 获得置信区间（$CI = \overline{x} \pm E$），即控制限。

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?

如果显著性水平增加，第一类错误（错误地拒绝真实的原假设）的风险也会增加。结合本题，这意味着更有可能在过程实际上运行良好时采取纠正措施，从而导致不必要的成本和混乱。

Question #6: Vacation occupancy rates were expected to be up during March 2008 in Myrtle Beach, South Carolina (*the sun news,* February 29, 2008). Data in the file Occupancy (Attached file **Occupancy**) will allow you to replicate the findings presented in the newspaper. The data show units rented and not rented for a random sample of vacation properties during the first week of March 2007 and March 2008.

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

表 8: Data summary

Name	occupancy
Number of rows	200
Number of columns	2
Column type frequency:	
character	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
March 2007	0	1.00	2	3	0	2	0
March 2008	50	0.75	2	3	0	2	0

```
## the proportion of units rented during the first week of March 2007: 0.35
```

```
## the proportion of units rented during the first week of March 2008: 0.4666667
```

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

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(occupancy$`March 2007` == "Yes", na.rm = TRUE), sum(occupancy$`March 2008` ==
## 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
```

```
## 95% confidence interval: -0.2261515 -0.007181823
```

解释：检测两个比例是否有显著差异，而不关心差异的方向时，使用双侧检验（*two.sided*）。

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

```
##
```

```
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(occupancy$`March 2007` == "Yes", na.rm = TRUE), sum(occupancy$`March 2008` ==
## X-squared = 4.3872, df = 1, p-value = 0.01811
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.02384624
## sample estimates:
##    prop 1    prop 2
## 0.3500000 0.4666667
```

是的。通过单侧（左侧）检验（less）我们可以看出，备选假设是 "2007 年 3 月的租赁率小于 2008 年 3 月租赁率"，p-value（0.01811）小于显著性水平（0.05），则有理由推翻原假设，支持备选假设——即 **2008 年 3 月租赁率相较 2007 年 3 月有显著提升**。

解释：检测两个比例是否有显著差异，且关心差异的方向时，使用单侧检验（*less or greater*）。

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

An air force introductory course in electronics uses a personalized system of instruction whereby each student views a videotaped lecture and then is given a programmed instruc-tion text. the students work independently with the text until they have completed the training and passed a test. Of concern is the varying pace at which the students complete this portion of their training program. Some students are able to cover the programmed instruction text relatively quickly, whereas other students work much longer with the text and require additional time to complete the course. The fast students wait until the slow students complete the introductory course before the entire group proceeds together with other aspects of their training.

A proposed alternative system involves use of computer-assisted instruction. In this method, all students view the same videotaped lecture and then each is assigned to a computer terminal for further instruction. The computer guides the student, working independently, through the self-training portion of the course.

To compare the proposed and current methods of instruction, an entering class of 122 students was assigned randomly to one of the two methods. one group of 61 students used the current programmed-text method and the other group of 61 students used the proposed computer-assisted method. The time in hours was recorded for each student in the study. Data are provided in the data set training (see Attached file).

Managerial Report

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?

表 10: Data summary

Name	training
Number of rows	61
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Current	0	1	75.07	3.94	65	72	76	78	84	
Proposed	0	1	75.43	2.51	69	74	76	77	82	

相似之处：

- 数据完整度（complete_rate）一致，两组数据中都没有缺失值。
- 中位数（p50）一样，均值（mean）基本一样，从直方图（hist）也能看出，两个变量的中心趋势非常接近，数据分布较为对称。

差异之处：

- 当前方法（Current）相对提议方法（Proposed），最小值（p0）、第一四分位数（p25）更小，第三四分位数（p75）、最大值（p100）更大，标准差（sd）更大。

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

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

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

从两个样本 t-test 的结果可看出，p 值为 0.5481，远大于 0.05，因此没有足够的证据拒绝原假设（均值无差异），故我们认为**当前方法和提议方法的均值没有显著差异。**

解释：使用 t.test() 函数进行独立样本 t-test 时，参数 var.equal 是一个逻辑值，用于指定是否假设两个独立样本的总体方差相等。这是一个重要的假设，因为当两个样本的方差不相等时，t 检验的计算方法会有所不同。因此先通过 f-test（var.test）来检验两个样本的方差是否相等。本题备选假设是方差之比不为 1，即不相等，p 值为 0.000578，远小于 0.05，故认为总体方差不相等，var.equal 设为 FALSE。

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 Proposed
## sd   3.944907 2.506385
## var 15.562295 6.281967
```

基于上一题对方差的 f-test，可以推断出**当前方法和提议方法的总体方差有显著差异，提议方法有更小的标准差和方差。**

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

提议方法更好。虽然两者在平均值的差异不大，但提议方法的方差更小，意味着其波动性更小——即学生更有可能在大致相同的时间完成培训，从而更有可能解决题目中提出的 “快学生等慢学生完成后才能一起继续” 的核心问题。

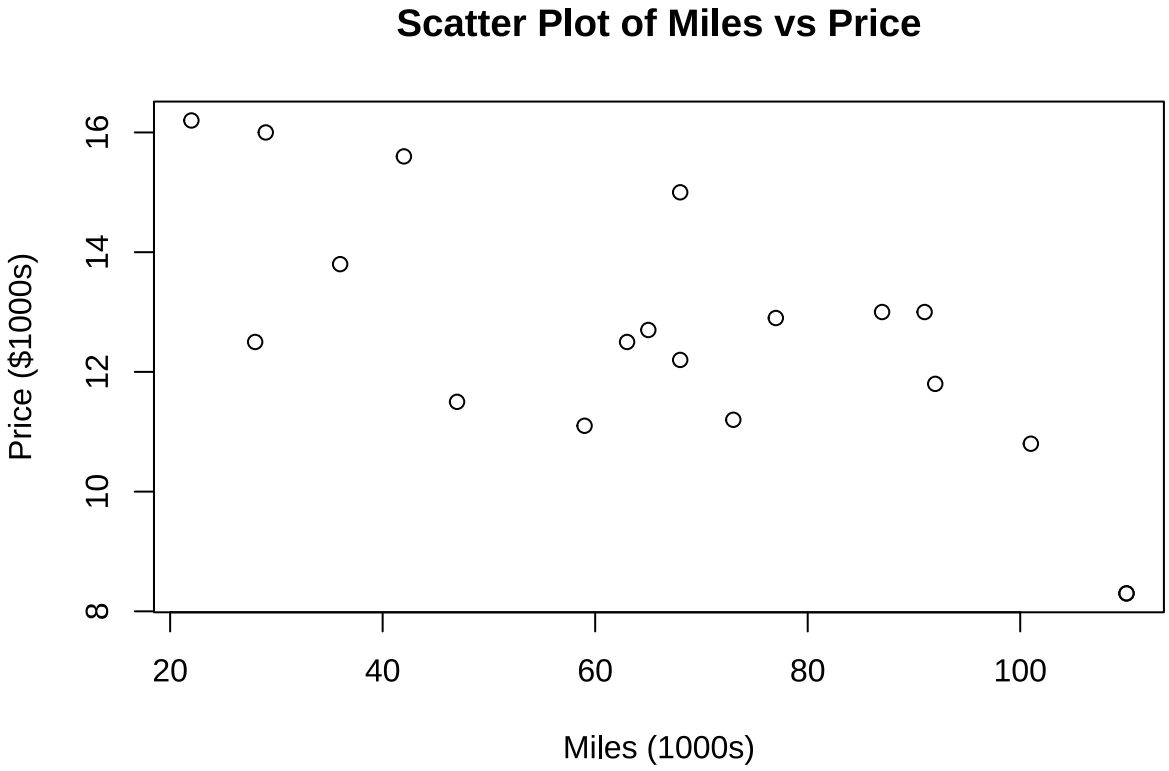
e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future?

建议收集两种方法学习效果（考试成绩）的数据。目前提供的数据只能说明提议方法能够优化培训时间的波动性，但对于学生而言，培训最终的目的还是考试，考试成绩是否会因为提议方法更优尚未可知。因此，**建议在最终决定是否切换到提议方法之前，通过分析考试成绩来评估两种方法的培训效果。**

Question #8: The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition.

To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, Attached data file Camry show the mileage and sale price for 19 sales (Pricehub website, February 24, 2012).

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?

从散点图可以看出，里程和价格之间似乎存在**负相关**关系——**随着里程的增加，价格基本呈下降趋势**。但考虑样本数量（19）较小，对总体数据的关系代表性不足。

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

```
##
## 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:
```

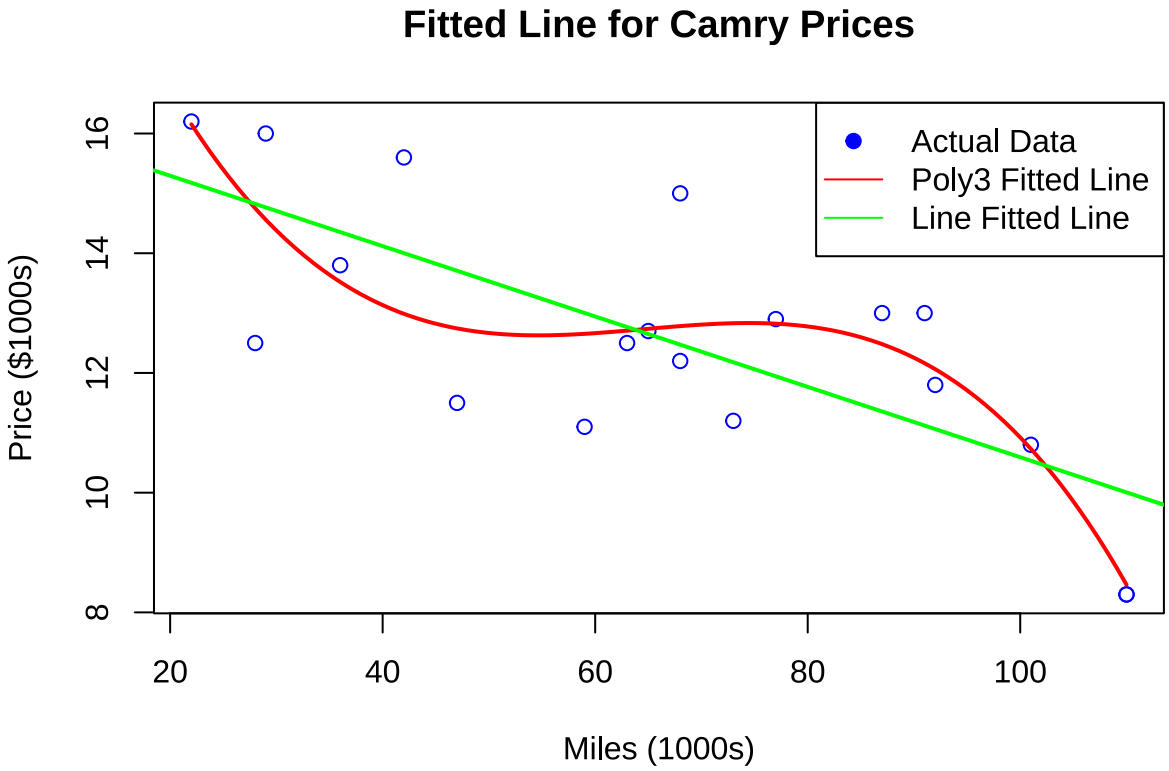
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.46976    0.94876  17.359 2.99e-12 ***
## Miles..1000s. -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

```
##
## Call:
## lm(formula = Price...1000s. ~ poly(Miles..1000s., 2), data = camry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3558 -1.1937 -0.2573  1.5213  2.2195
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               12.5474     0.3568  35.171  < 2e-16 ***
## poly(Miles..1000s., 2)1  -6.8672     1.5551  -4.416 0.000433 ***
## poly(Miles..1000s., 2)2  -1.3031     1.5551  -0.838 0.414384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555 on 16 degrees of freedom
## Multiple R-squared:  0.5581, Adjusted R-squared:  0.5028
## F-statistic:  10.1 on 2 and 16 DF,  p-value: 0.001455
```

```
##
## Call:
## lm(formula = Price...1000s. ~ poly(Miles..1000s., 3), data = camry)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25122 -0.42277 -0.03783  0.39908  2.61250
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               12.5474     0.3082  40.713  < 2e-16 ***
```

```
## poly(Miles..1000s., 3)1  -6.8672     1.3434  -5.112 0.000128 ***
## poly(Miles..1000s., 3)2  -1.3031     1.3434  -0.970 0.347413
## poly(Miles..1000s., 3)3  -3.4091     1.3434  -2.538 0.022746 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.343 on 15 degrees of freedom
## Multiple R-squared:  0.6908, Adjusted R-squared:  0.629
## F-statistic: 11.17 on 3 and 15 DF,  p-value: 0.0004152

##
## Call:
## lm(formula = Price...1000s. ~ log(Miles..1000s.), data = camry)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52271 -1.13498  0.08368  1.10936  2.82915
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         25.7327     3.2443   7.932  4.1e-07 ***
## log(Miles..1000s.)  -3.2141     0.7857  -4.091 0.000762 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.611 on 17 degrees of freedom
## Multiple R-squared:  0.4961, Adjusted R-squared:  0.4664
## F-statistic: 16.73 on 1 and 17 DF,  p-value: 0.0007619

##
## Call:
## lm(formula = Price...1000s. ~ exp(Miles..1000s.), data = camry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2465 -1.0471 -0.1471  0.3765  3.1529
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.305e+01  4.042e-01   32.28   <2e-16 ***
## exp(Miles..1000s.) -8.018e-48  2.104e-48   -3.81   0.0014 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.667 on 17 degrees of freedom
## Multiple R-squared:  0.4606, Adjusted R-squared:  0.4289
## F-statistic: 14.52 on 1 and 17 DF,  p-value: 0.001399

## The model with the lowest AIC is: poly3_model with AIC: 70.64512

## The model with the lowest BIC is: poly3_model with BIC: 75.36732
```



如果回归指的是线性回归，带入系数和截距可知方程为：$\text{Price} = -0.05877 \times \text{Miles} + 16.46976$

解释：从拟合度最优的角度考虑，线性不一定拟合最优。为此尝试了四种简单的非线性回归，比较最优（最低）*AIC* 和 *BIC* 发现，三次多项式回归方程拟合度上比线性回归更优：

$$\text{Price} = 12.5474 - 6.8672 \cdot \text{Miles} - 1.3031 \cdot \text{Miles}^2 - 3.4091 \cdot \text{Miles}^3$$

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

从 p-value 为 0.0003475 远小于 0.05 可看出，**线性模型整体上是统计显著的。**

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

是的。从 R-squared 为 0.5387 可以看出，**线性模型解释了 53.87% 的因变量变异。**

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

线性回归方程的斜率为-0.05877，这意味着**汽车里程每增加 1000 英里，价格预计下降 58.77 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.

```
## predicted price of the linear model is 12.94332
```

由上述预测结果可知，**2007 款 Camry 行驶了 60,000 英里后的预测价格为 12.94332，即 12,943.32 美元**。这个价格来源于已有数据建模，可以作为向卖家提供的参考价格（技术层面有一定说服力），但维度过于单一，实际购买时还需要考虑其他因素，如实际车况和当下的市场需求等。

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？

表 12: Data summary

Name	churned_we
Number of rows	6024
Number of columns	13
Column type frequency:	
numeric	13
Group variables	None

Variable type: numeric

表 13: Data summary

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
客户 ID	0	1	3219.27	1833.47	1	1629.75	3238.5	4818.25	6347	
流失	0	1	0.00	0.00	0	0.00	0.0	0.00	0	
当月客户幸福指数	0	1	88.61	66.47	0	26.00	89.0	140.00	298	
客户幸福指数相比上月变化	0	1	5.53	30.91	-125	-7.00	0.0	15.00	208	
当月客户支持	0	1	0.72	1.75	0	0.00	0.0	1.00	32	
客户支持相比上月的变化	0	1	-0.01	1.90	-29	0.00	0.0	0.00	31	
当月服务优先级	0	1	0.83	1.33	0	0.00	0.0	2.75	4	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
服务优先级相比上月的变化	0	1	0.03	1.47	-4	0.00	0.0	0.00	4	
当月登录次数	0	1	16.14	42.22	-293	0.00	3.0	24.00	865	
博客数相比上月的变化	0	1	0.17	4.77	-75	0.00	0.0	0.00	217	
访问次数相比上月的增加	0	1	106.61	3210.13	-27768	-12.00	0.0	28.00	230414	
客户使用期限	0	1	18.82	11.26	5	10.00	16.0	25.00	72	
访问间隔变化	0	1	3.51	17.73	-646	2.00	2.0	5.00	48	

Name	non_churned_we
Number of rows	323
Number of columns	13
Column type frequency:	
numeric	13
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
客户 ID	0	1	2329.72	1593.44	54	893.5	2003	3647.0	6263	
流失	0	1	1.00	0.00	1	1.0	1	1.0	1	
当月客户幸福指数	0	1	63.27	57.70	0	5.0	57	106.0	231	
客户幸福指数相比上月变化	0	1	-3.74	27.89	-105	-15.0	0	5.0	73	
当月客户支持	0	1	0.37	1.08	0	0.0	0	0.0	10	
客户支持相比上月的变化	0	1	0.04	1.25	-8	0.0	0	0.0	9	
当月服务优先级	0	1	0.50	1.11	0	0.0	0	0.0	4	
服务优先级相比上月的变化	0	1	-0.02	1.34	-3	0.0	0	0.0	3	
当月登录次数	0	1	8.06	39.46	-76	-2.0	0	9.0	496	
博客数相比上月的变化	0	1	-0.10	1.60	-8	0.0	0	0.0	7	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
访问次数相比上月的增加	0	1	-95.77	1749.30	-28322	-8.0	0	15.0	5664	
客户使用期限	0	1	20.35	8.87	6	15.0	18	23.5	52	
访问间隔变化	0	1	8.49	21.43	-88	1.0	4	32.5	63	

从统计结果能看出，流失客户与非流失客户的结果集，**除了客户 ID（无意义）和流失字段本身之外，其余 11 个字段均有不同，是否显著不同需要进一步测试**。

b. 通过均值比较的方式验证上述不同是否显著。

```
##                              variable     estimate1    estimate2  statistic
## mean of x             当月客户幸福指数  88.605909695  63.27244582  7.6242176
## mean of x1  客户幸福指数相比上月变化   5.530212483  -3.73684211  5.7835224
## mean of x2               当月客户支持   0.724269588   0.37151703  5.5098545
## mean of x3    客户支持相比上月的变化  -0.009296149   0.03715170 -0.6319825
## mean of x4             当月服务优先级   0.829575893   0.49955772  5.1427709
## mean of x5  服务优先级相比上月的变化   0.032681838  -0.01669615  0.6411575
## mean of x6               当月登录次数  16.138944223   8.06191950  3.3603469
## mean of x7      博客数相比上月的变化   0.171148738  -0.10216718  2.5315145
## mean of x8    访问次数相比上月的增加 106.609561753 -95.76780186  1.9136102
## mean of x9               客户使用期限  18.818725100  20.35294118 -2.9811315
## mean of x10              访问间隔变化   3.511454183   8.48606811 -4.0971030
##                   p.value is.significant
## mean of x   2.096694e-13           TRUE
## mean of x1  1.571085e-08           TRUE
## mean of x2  6.280509e-08           TRUE
## mean of x3  5.277532e-01          FALSE
## mean of x4  4.380969e-07           TRUE
## mean of x5  5.218233e-01          FALSE
## mean of x6  7.830410e-04           TRUE
## mean of x7  1.157611e-02           TRUE
## mean of x8  5.630696e-02          FALSE
## mean of x9  3.056793e-03           TRUE
## mean of x10 5.215207e-05           TRUE
```

由 t.test 结果可以看出，**以 0.05 为显著性水平，则：**

- **显著的指标有：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**
- **不显著的指标有：客户支持相比上月的变化、服务优先级相比上月的变化、访问次数相比上月的增**

加

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 = 流失 ~ 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,
##     family = binomial, data = we)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.8763327  0.1212590 -23.721  < 2e-16 ***
## 当月客户幸福指数         -0.0051988  0.0011558  -4.498 6.86e-06 ***
## 客户幸福指数相比上月变化 -0.0093063  0.0024124  -3.858 0.000114 ***
## 当月客户支持             -0.0221691  0.0714550  -0.310 0.756369
## 当月服务优先级           -0.0447524  0.0741355  -0.604 0.546072
## 当月登录次数              0.0008545  0.0019376   0.441 0.659211
## 博客数相比上月的变化     -0.0009717  0.0205099  -0.047 0.962213
## 客户使用期限              0.0142559  0.0052396   2.721 0.006513 **
## 访问间隔变化              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
```

解释：选取均值不同最显著的作为特征来建模。

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。

```
## # A tibble: 100 x 2
##    客户ID predictions
##     <dbl>       <dbl>
##  1   1363       0.192
```

```
##  2   1672       0.180
##  3    299       0.174
##  4   2922       0.162
##  5   2951       0.162
##  6   1021       0.159
##  7    335       0.156
##  8    156       0.154
##  9   1488       0.147
## 10   3340       0.145
## # i 90 more rows
```

解释：本小题描述似乎有误，尚未流失的客户应该为 *(*流失 $=$ *1)*，因为*"1"*表示不流失。按这样理解，可通过 *type = "response"* 返回模型预测的概率值再来倒序输出 *top100*。