Solution for MEM Assignment r2

邵瑞瑞

2024-11-29

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

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.30   14.10   15.00   15.04   16.00   16.50
```

The minimum is 13.3, the maximum number of viewers is 16.5

b. Compute the mean, median, and mode.

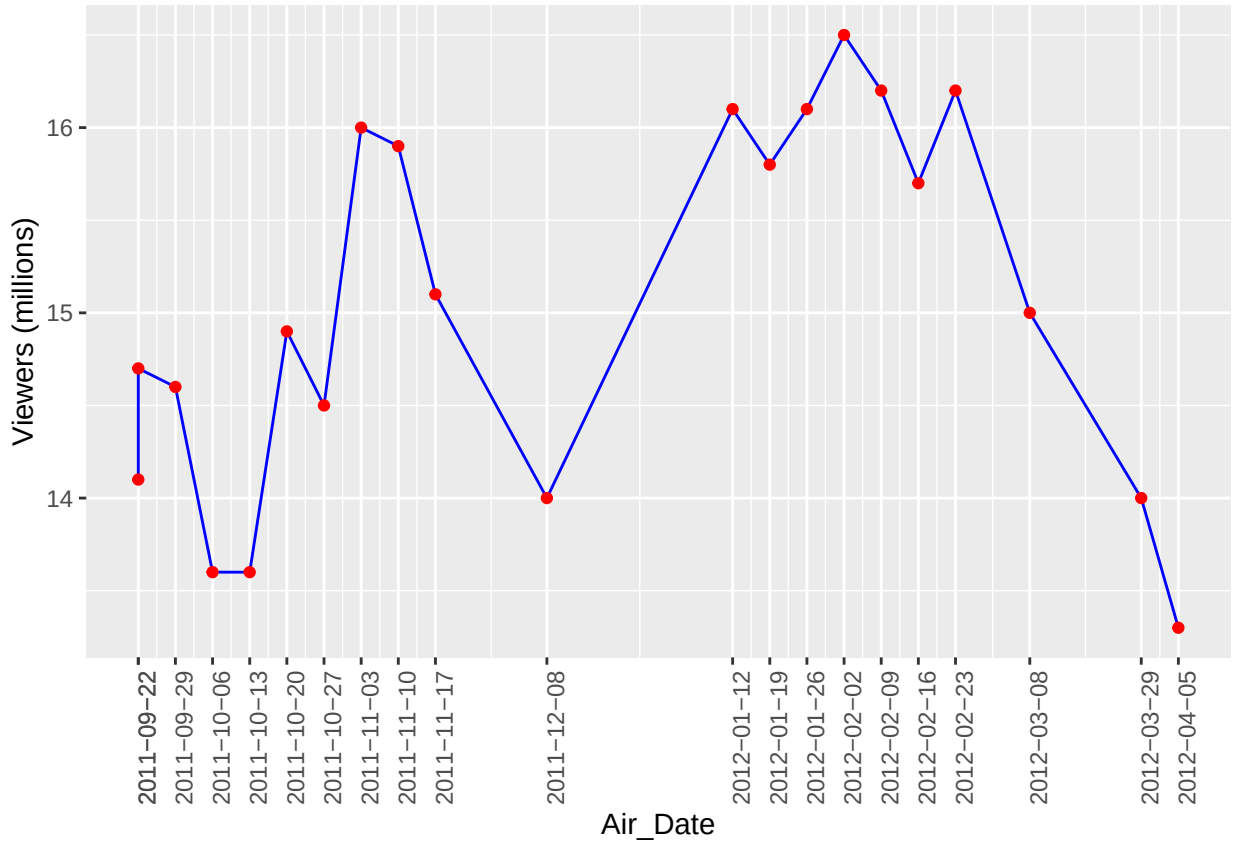
```
## [1] 13.6 14.0 16.1 16.2
```

Mean=15.04; median = 15.00; mode = 13.6; 14.0; 16.1; 16.2

c.Compute the first and third quartiles.

1st Qu.=14.10, 3rd Qu.=16.00

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




```
## average_change == -0.04 /n
```

```
## average_percent_change == -0.1135253
```

During the 2011-2012 period, viewership declined.
1.From the line chart, we can observe the trend in viewership. The number of viewers continuously declined from the episode on 2012-02-23 to the episode on 2012-04-05, with the episode on 2012-04-05 reaching the lowest viewership level in history.
2.On average, the viewership of each episode is 40,000 fewer than the previous episode, showing a declining trend. The average decline in viewership per episode relative to the previous episode is approximately 11.35%. Overall, viewership during the 2011-2012 period experienced a significant decrease.

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

a. Show the frequency distribution.

```
## PPG_binned
## [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
```

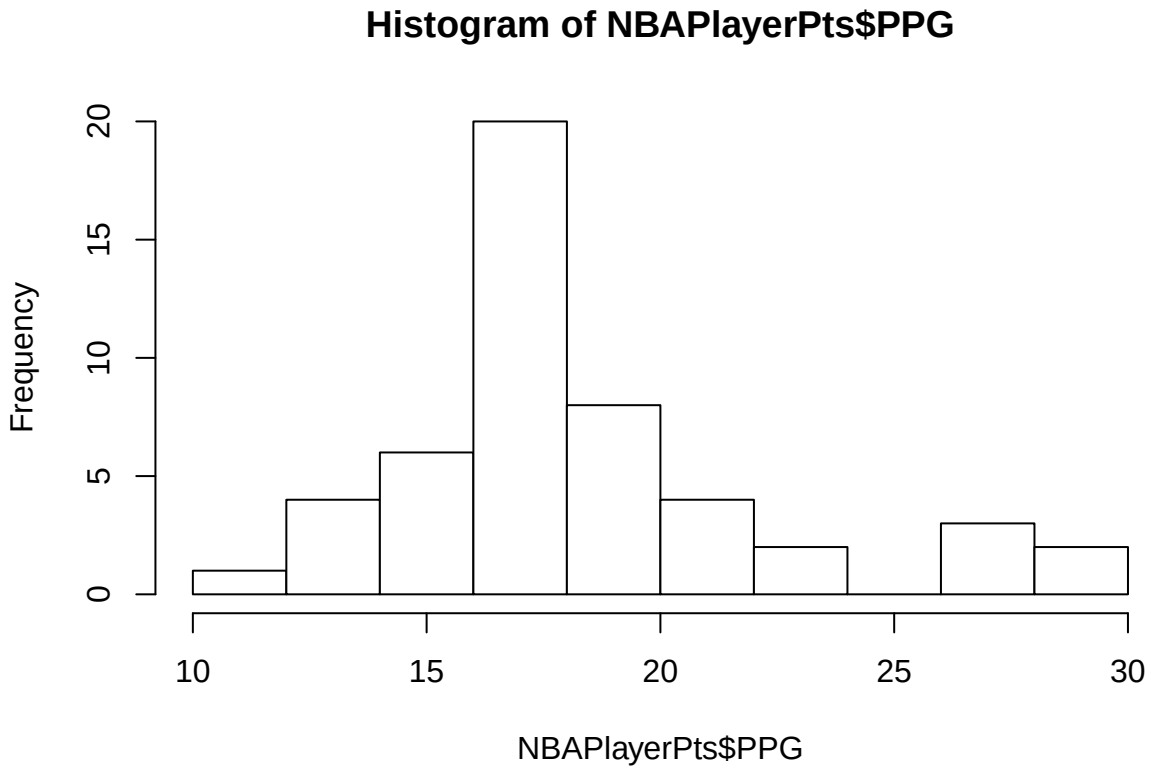
b. Show the relative frequency distribution.

```
## PPG_binned
## [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
```

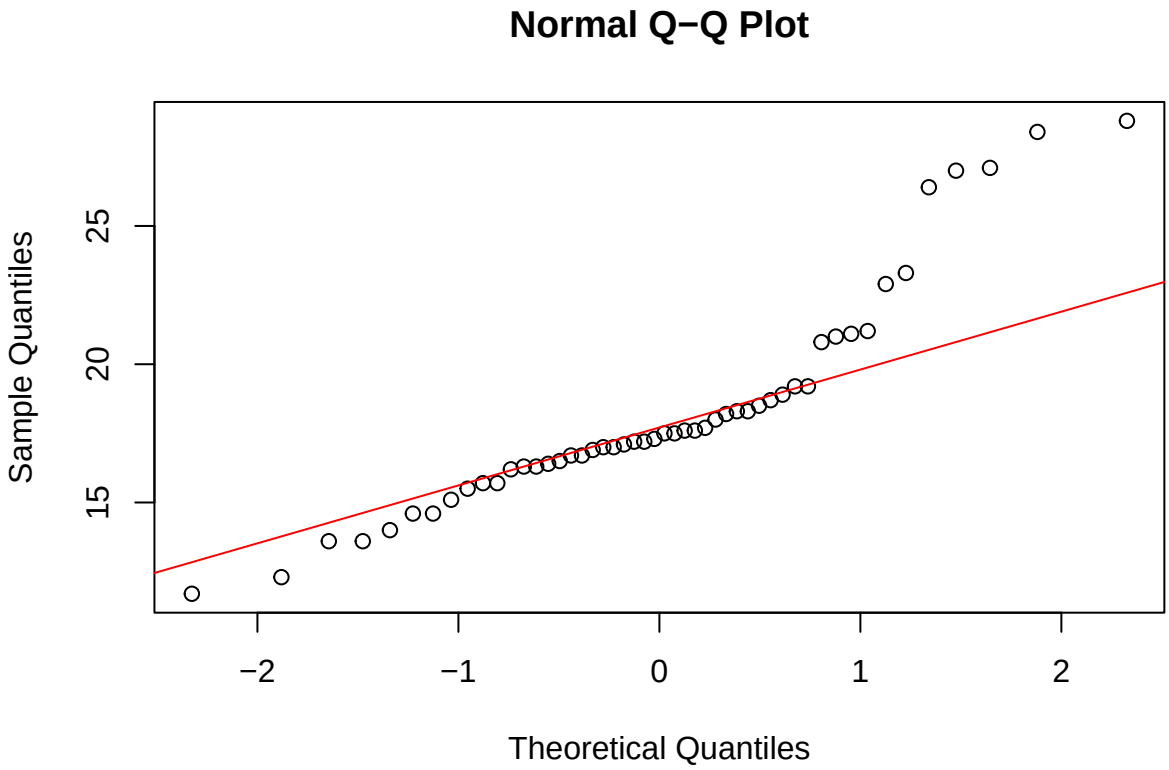
c. Show the cumulative percent frequency distribution.

```
## [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
```

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



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



1.The histogram of the PPG data shows a longer right tail, indicating that the data is right-skewed (positively skewed).
2.The Q-Q plot reveals that the data points on the right side deviate significantly from the diagonal line, suggesting the presence of skewness in the data.

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

```
## 22%
```

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?

```
## The sample size is 625
```

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

```
## The probability is 0.79
```

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

a. Develop appropriate descriptive statistics to summarize the data.

Table 1: Data summary

Name	Professional
Number of rows	410
Number of columns	8
Column type frequency:	
factor	4
numeric	4
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Gender	0	1	FALSE	2	Mal: 229, Fem: 181
Real Estate Purchases	0	1	FALSE	2	No: 229, Yes: 181
Broadband Access	0	1	FALSE	2	Yes: 256, No: 154
Have Children	0	1	FALSE	2	Yes: 219, No: 191

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	74459.51	34818.21	16200	51625	66050	88775	322500	

()|0|1|28538.29|15810.83|0|18300|24800|34275|133400| ___||*NumberofTransactions*|0|1|5.97|3.10|0|4|6|7|21| __||*Household*

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

```
## [1] 29.72153 30.50286
## attr(,"conf.level")
## [1] 0.95
```

95% confidence that the mean age of subscribers is between 29.72 and 30.50 years of age.

```
## [1] 71079.26 77839.77
## attr(,"conf.level")
## [1] 0.95
```

95% confidence that the mean household income of subscribers is between $71,079 and $77,840.

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:  numer_Broadband out of n, 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:  numer_children out of n, 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
```

95% confidence intervals for the proportion of subscribers who have broadband access at home is between 0.58 and 0.67.

95% confidence intervals for the proportion of subscribers who who have children is between 0.48 and 0.58.

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

Yes.

1.Age: Online brokers typically achieve better advertising effectiveness when targeting people aged 25-55. Young Professional subscribers are aged between 19 and 42, with an average age of 30, and 75% of subscribers are older than 28, which places them within the target demographic.

2.Investment needs or interests: The average total value of financial investments is $28,538, and 75% of subscribers have a total financial investment value exceeding $18,300. The average number of transactions is 6 per year, with 75% of subscribers making more than 4 transactions annually. These data demonstrate the Young Professional subscribers' strong investment needs and interest.

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

Yes.

1.From the proportion of households with children and parents' age group: The analysis suggests that 53% of Young Professional subscribers have children, with the proportion exceeding half. Additionally, the Young Professional subscriber group primarily consists of young individuals, which indicates that these parents' children are likely of school age.

2.From household income level: The average household income of Young Professional subscribers is $74,460, and 75% of households have an income greater than $51,625, placing them in the middle-to-upper income level.

3.From the broadband installation status: The analysis indicates that 62% of Young Professional subscribers have broadband installed in their homes, indicating they have access to internet connectivity.

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

```
##
##  1-sample proportions test with continuity correction
##
## data:  numer_Gender out of n, null probability 0.5
## X-squared = 5.3878, df = 1, p-value = 0.02028
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5089269 0.6070321
## sample estimates:
##         p
## 0.5585366
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  numer_Real_Estate_Purchases out of n, null probability 0.5
## X-squared = 5.3878, df = 1, p-value = 0.02028
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3929679 0.4910731
## sample estimates:
##         p
## 0.4414634
```

Additional analysis inference: 65% of Young Professional subscribers are young males, and 44% of subscribers have the intention to purchase a home. Based on the above analysis, the types of articles that Young Professional subscribers are interested in are as follows:

1.Investment and wealth management

2.Real estate market trends, first-time homebuyer guides

3.Internet technology, the latest digital products, smart homes

4.Children's education

5.Health and lifestyle

6.Professional growth and career development

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.

Table 4: Data summary

Name	Quality
Number of rows	30
Number of columns	4
Column type frequency:	
numeric	4
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sample 1	0	1	11.96	0.22	11.52	11.81	11.96	12.15	12.32	
Sample 2	0	1	12.03	0.22	11.59	11.88	12.02	12.21	12.39	
Sample 3	0	1	11.89	0.21	11.36	11.75	11.92	12.00	12.22	
Sample 4	0	1	12.08	0.21	11.64	11.98	12.08	12.23	12.47	

```
##
##  One Sample t-test
##
## data:  Quality$`Sample 1`
## t = -1.0274, df = 29, p-value = 0.3127
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.84777 12.06956
## sample estimates:
## mean of x
##  11.95867
```

For Sample 1, p = 0.3127 > 0.01, we cannot reject the null hypothesis. The process is considered to be operating normally, and no corrective action is needed.

```
##
##  One Sample t-test
##
## data:  Quality$`Sample 2`
## t = 0.71255, df = 29, p-value = 0.4818
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.91777 12.13956
## sample estimates:
## mean of x
##  12.02867
```

For Sample 2, p = 0.4818 > 0.01, we cannot reject the null hypothesis. The process is considered to be operating normally, and no corrective action is needed.

```
##
##  One Sample t-test
##
## data:  Quality$`Sample 3`
## t = -2.9346, df = 29, p-value = 0.006469
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.78474 11.99326
## sample estimates:
## mean of x
##    11.889
```

For Sample 3, p = 0.006469 < 0.01, we reject the null hypothesis, and corrective action may be needed.

```
##
##  One Sample t-test
##
## data:  Quality$`Sample 4`
## t = 2.1614, df = 29, p-value = 0.03906
## alternative hypothesis: true mean is not equal to 12
## 99 percent confidence interval:
##  11.97761 12.18506
## sample estimates:
## mean of x
##  12.08133
```

For Sample 4, p = 0.03906 > 0.01, we cannot reject the null hypothesis. The process is considered to be operating normally, and no corrective action is needed.

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

H0: = 0.21 H1: != 0.21

```
## p_value_sample1 = 0.3229211
```

```
## p_value_sample2 = 0.3229211
```

```
## p_value_sample3 = 0.5059715
```

```
## p_value_sample4 = 0.5213745
```

For the four samples above, the p-values are all greater than 0.05. There is no significant difference between the sample standard deviations and the hypothesized population standard deviation. Therefore, the assumption that the population standard deviation is 0.21 is reasonable.

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.

```
## Upper Control Limit : 12.09877
```

```
## Lower Control Limit : 11.90123
```

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?

Increasing the significance level increases the chance of rejecting the null hypothesis, which in turn increases the probability of making a Type I error.

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.

```
## 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 0.4666667
```

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

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## 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
```

The 95% confidence interval for the difference in proportions is: -0.226151510 to -0.007181823.

Since p-value < 0.05, it indicates that there is a significant difference between the proportions for the two years.

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

Yes.Since the confidence interval does not include 0,it indicates that there is a significant difference between the proportions for the two years. Besides, the proportion of units rented during the first week of March 2008 exceeds that of the first week of March 2007.

#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?

Table 6: 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	

Similarities:

1.The average duration for both methods is similar.

2.The time distribution histograms for both methods are close to a normal distribution.

3.The learning time for 50% of the students in both groups is 76.

Differences:

1.The standard deviation for the Current group is larger than that of the Proposed group, indicating that the learning time for the Proposed group is more consistent.

2.The Proposed group has a smaller range, indicating that the learning times are more concentrated.

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

```
##
##  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
```

The 95 percent confidence interval and a p-value > 0.05 indicate that there is no significant difference between the means of the two groups.

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.

```
## sd_Current = 3.944907
```

```
## sd_Proposed = 2.506385
```

```
## variance_Current = 15.56229
```

```
## variance_Proposed = 6.281966
```

```
##
##  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
```

The 95 percent confidence interval and a p-value < 0.05 indicate that there is a significant difference between the variances of the two methods.

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

It is recommended to adopt the computer-assisted instruction method.

Reason: There is no significant difference between the means of the two groups, but the Proposed group has a smaller range, standard deviation, and variance. Moreover, there is a significant difference in the variances between the two groups, with the Proposed group showing more concentrated and stable learning times. This indicates that the computer-assisted instruction method can improve the efficiency of students who initially progress more slowly, ensuring the overall pace of learning.

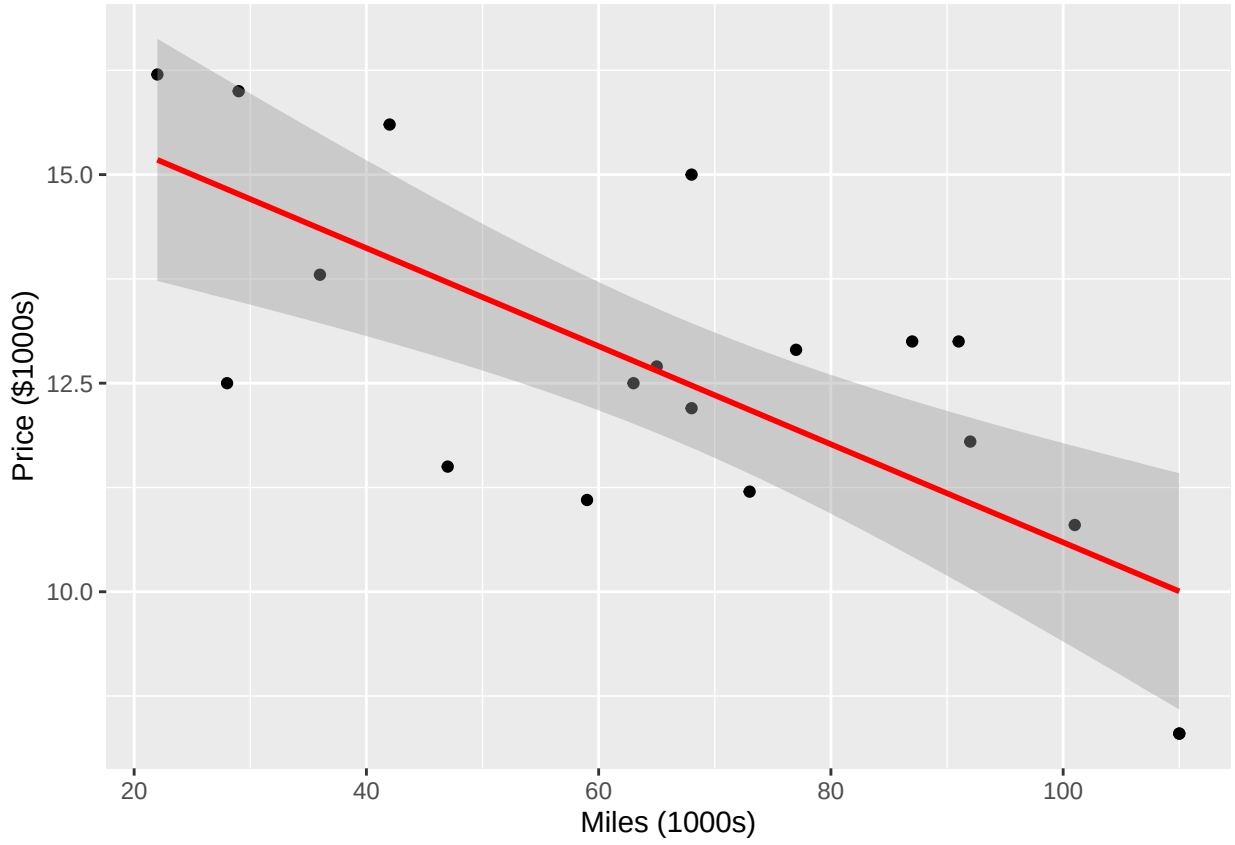
e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future?

1.Data on the subsequent exam scores of the two groups can be introduced to evaluate whether there is a difference in the effectiveness of the two methods.

2.Different classes can be selected for additional 2-3 grouping experiments.

#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?

From the scatter plot and the added regression line, it can be observed that Price and Miles exhibit a linear negative correlation. This means that as the car's mileage increases, the resale price decreases.

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

y=16.46976-0.05877x

y:price ($1000s), x:miles (1000s)

```
Camry_model  <- lm(Camry$`Price ($1000s)`~Camry$`Miles (1000s)`,  Camry )
summary(Camry_model)
```

```
##
## Call:
## lm(formula = Camry$`Price ($1000s)` ~ Camry$`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 ***
## Camry$`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
```

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

The mileage has a significant impact on the price, meaning that the mileage has a statistically significant effect on the sales price of the Toyota Camry.

The p-value for the t-test of the intercept is extremely small (much smaller than 0.05), so we can reject the null hypothesis (that the intercept is zero), indicating that the intercept is significant.

The p-value for the mileage coefficient is also smaller than 0.05, so we can reject the null hypothesis (that the mileage coefficient is zero).

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

The estimated regression equation provides a good fit.

The residual standard error is 1.541, which is relatively small compared to the price magnitude, indicating that the model fits the data well.

The R^2 value is 0.5387, suggesting that the model has a certain level of explanatory power.

The F-statistic is 19.85, and the p-value is less than 0.05, indicating that the overall regression model is significant, meaning that the mileage as an independent variable has a meaningful effect on predicting the price.

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

The mileage coefficient is -0.05877. This coefficient indicates that for every additional 1,000 miles of mileage, the price of the Camry decreases by approximately 0.05877 thousand dollars (or about 58.77 dollars). The negative value suggests a negative correlation between mileage and price, meaning that the higher the mileage, the lower the price.

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.

```
## Price ($1000s) = 12.94356
```

The price of $12.94356 ($1000s) would not be the offer price, as the vehicle still needs a comprehensive evaluation.

There are many factors that influence the pricing of a used car, including the vehicle's age, mileage, condition (both exterior and mechanical), accident history, and maintenance records.

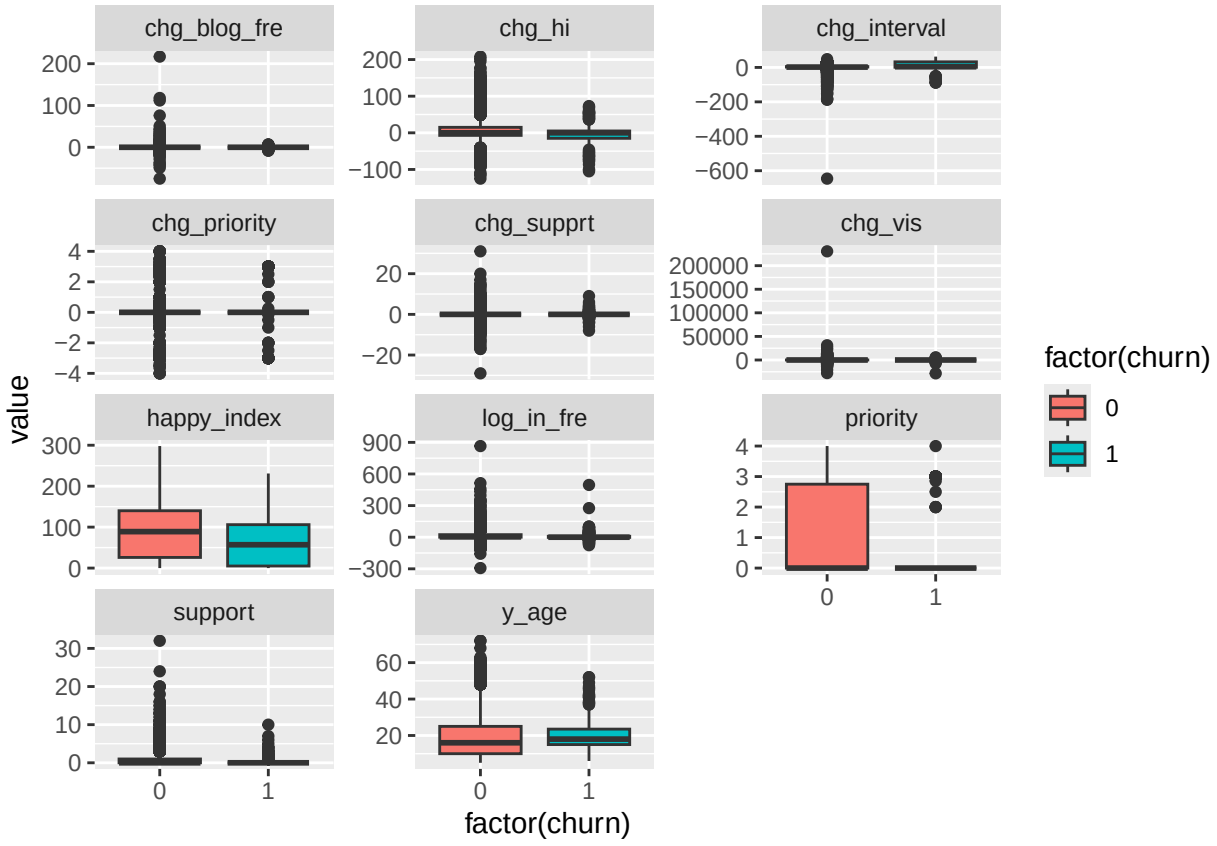
A 2007 Toyota Camry with 60,000 miles, if in good overall condition and without any accident or maintenance history, could potentially be sold at a higher price. $12.94356 ($1000s) could serve as a reference for the offer price.

#Question 9 附件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。

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？

```
## Rows: 6,347
## Columns: 13
## $ ID           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ churn        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ happy_index  <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78, 91, 40, 215, 0,~
## $ chg_hi       <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1, 14, 15, 0, 63,~
## $ support      <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ chg_supprt   <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ priority     <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
## $ chg_priority <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ log_in_fre   <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0, 71, 0, 5, 0, 4~
## $ chg_blog_fre <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, 0, 1, 0, 0, 0, 6~
## $ chg_vis      <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30, 0, 15, 8658, 0,~
## $ y_age        <dbl> 72, 72, 60, 68, 62, 63, 62, 51, 61, 61, 58, 61, 62, 62, 6~
## $ chg_interval <dbl> 33, 33, 33, 2, 33, 2, 2, 8, 9, 16, 2, 33, 2, 33, 2, 33, 3~
```

churn	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg_blog_fre	c
0	88.60591	5.530213	0.7242696	-0.0092961	0.8295759	0.0326818	16.13894	0.1711487	1
1	63.27245	-3.736842	0.3715170	0.0371517	0.4995577	-0.0166962	8.06192	-0.1021672	-



The behavior of churned and non-churned customers shows certain differences in terms of the mean values of various indicators and the boxplots.But whether these differences are significant,we need to test.

b. 通过均值比较的方式验证上述不同是否显著

```
## # A tibble: 1 x 11
##   happy_index       chg_hi   support chg_supprt priority chg_priority log_in_fre
##         <dbl>        <dbl>     <dbl>      <dbl>    <dbl>        <dbl>      <dbl>
## 1    2.10e-13 0.0000000157   6.28e-8      0.528  4.38e-7        0.522   0.000404
## # i 4 more variables: chg_blog_fre <dbl>, chg_vis <dbl>, y_age <dbl>,
## #   chg_interval <dbl>
```

通过查看 p.value（<0.05 表明均值有显著性差异）发现：

happy_index、chg_hi、support、priority、log_in_fre、chg_blog_fre、y_age、chg_interval 存在显著性差异

chg_supprt、chg_priority、chg_vis 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 = churn ~ happy_index + chg_hi + support + priority +
```

```
##     log_in_fre + chg_blog_fre + y_age + chg_interval, family = binomial,
##     data = WE_set)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8763327  0.1212590 -23.721  < 2e-16 ***
## happy_index  -0.0051988  0.0011558  -4.498 6.86e-06 ***
## chg_hi       -0.0093063  0.0024124  -3.858 0.000114 ***
## support      -0.0221691  0.0714550  -0.310 0.756369
## priority     -0.0447524  0.0741355  -0.604 0.546072
## log_in_fre    0.0008545  0.0019376   0.441 0.659211
## chg_blog_fre -0.0009717  0.0205099  -0.047 0.962213
## y_age         0.0142559  0.0052396   2.721 0.006513 **
## chg_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
```

```
##  happy_index       chg_hi      support     priority   log_in_fre chg_blog_fre
##     1.500264     1.238816     2.079547     2.079925     1.291954     1.064875
##        y_age chg_interval
##     1.247810     1.198751
```

Although four independent variables in the regression model have p-values greater than 0.05, the primary goal of the model is prediction, and since the VIF values are all between 1 and 5, these variables are still retained. The equation is as follows:

iflose = -2.876333-0.005199*happy_index-0.009306*chg_hi-0.022169*support-0.044752*priority+0.000854*log_in_fre-0.000972*chg_blog_fre+0.014256*y_age+0.016950*chg_interval

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。

ID	churn	churn_prob	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg
1363	1	0.1919757	0	-34	0	0	0	0	-9	
1672	1	0.1802687	2	1	0	0	0	0	0	
299	1	0.1744655	14	-101	0	0	0	0	-4	
2922	1	0.1623822	13	-52	0	0	0	0	-1	
2951	1	0.1618436	20	-39	0	0	0	0	0	
1021	1	0.1589334	12	-73	0	0	0	0	-4	
335	1	0.1563708	0	-64	0	0	0	0	-7	
156	1	0.1541017	8	0	0	0	0	0	0	
1488	1	0.1469821	5	0	0	0	0	0	0	
3340	1	0.1453980	15	-105	0	0	0	0	-50	

2296	1	0.1448915	0	0	0	0	0	0	0
1069	1	0.1379684	0	0	0	0	0	0	0
3811	1	0.1362973	40	-100	0	-1	0	-2	-16
2653	1	0.1346123	0	0	0	0	0	0	0
3604	1	0.1338222	0	-78	0	0	0	0	-4
1405	1	0.1328413	11	11	0	0	0	0	2
2636	1	0.1313254	0	0	0	0	0	0	0
2077	1	0.1308082	13	-4	0	0	0	0	0
1987	1	0.1297076	0	0	0	0	0	0	0
4292	1	0.1296641	0	-26	0	0	0	0	0
2082	1	0.1268756	5	-25	0	0	0	0	-1
1782	1	0.1265470	31	-25	0	0	0	0	0
2766	1	0.1257334	36	-27	0	0	0	0	0
2166	1	0.1249558	0	0	0	0	0	0	0
2120	1	0.1234054	0	0	0	0	0	0	0
1303	1	0.1231142	0	0	0	0	0	0	0
904	1	0.1230854	0	-29	0	0	0	0	-1
3092	1	0.1226184	15	-3	0	0	0	0	0
2624	1	0.1222536	18	2	0	0	0	0	2
2371	1	0.1218715	0	0	0	0	0	0	0
2928	1	0.1218715	0	0	0	0	0	0	0
1563	1	0.1203540	0	0	0	0	0	0	0
1711	1	0.1188529	0	0	0	0	0	0	0
1532	1	0.1170892	0	0	0	0	0	0	0
891	1	0.1158993	0	0	0	0	0	0	0
945	1	0.1158993	0	0	0	0	0	0	0
947	1	0.1158993	0	0	0	0	0	0	0
948	1	0.1158993	0	0	0	0	0	0	0
2084	1	0.1156708	5	5	0	0	0	0	2
896	1	0.1144465	0	0	0	0	0	0	0
402	1	0.1139450	8	-7	0	0	0	0	0
227	1	0.1130096	0	0	0	0	0	0	0
979	1	0.1130096	0	0	0	0	0	0	0
938	1	0.1127398	0	0	0	0	0	0	0
2902	1	0.1121649	30	1	0	0	0	0	0
257	1	0.1115885	0	0	0	0	0	0	0
317	1	0.1115885	0	0	0	0	0	0	0
363	1	0.1115885	0	0	0	0	0	0	0
371	1	0.1115885	0	0	0	0	0	0	0
523	1	0.1115885	0	0	0	0	0	0	0
543	1	0.1115885	0	0	0	0	0	0	0
548	1	0.1115885	0	0	0	0	0	0	0
787	1	0.1115885	0	0	0	0	0	0	0
1214	1	0.1115885	0	0	0	0	0	0	0
1760	1	0.1115885	0	0	0	0	0	0	0
3312	1	0.1115885	0	0	0	0	0	0	0
3313	1	0.1115885	0	0	0	0	0	0	0
4500	1	0.1115885	0	0	0	0	0	0	0
3569	1	0.1106295	0	-45	0	0	0	0	-15
1659	1	0.1102284	5	5	0	0	0	0	2

3163	1	0.1101830	0	0	0	0	0	0	0
3235	1	0.1101830	0	0	0	0	0	0	0
3349	1	0.1101830	0	0	0	0	0	0	0
3228	1	0.1099191	0	0	0	0	0	0	0
3267	1	0.1099191	0	0	0	0	0	0	0
640	1	0.1097650	10	1	0	0	0	0	0
5312	1	0.1091330	0	-58	0	-1	0	-3	-14
930	1	0.1088256	9	-2	0	0	0	0	0
3772	1	0.1087931	0	0	0	0	0	0	0
3363	1	0.1074185	0	0	0	0	0	0	0
3487	1	0.1072170	22	-9	0	0	0	0	0
2179	1	0.1066551	60	-25	0	0	0	0	0
4280	1	0.1062829	18	-56	0	0	0	0	-6
2707	1	0.1061770	70	-77	0	-1	0	-3	-50
4273	1	0.1060593	0	0	0	0	0	0	0
3861	1	0.1055119	33	-36	0	0	0	0	-4
2189	1	0.1053949	37	0	0	0	0	0	0
4263	1	0.1048756	0	0	0	0	0	0	2
4289	1	0.1048756	0	0	0	0	0	0	2
4291	1	0.1047153	0	0	0	0	0	0	0
453	1	0.1038621	20	-1	0	0	0	0	0
1831	1	0.1037784	65	0	0	0	0	0	0
4680	1	0.1033863	0	0	0	0	0	0	0
3584	1	0.1019346	14	0	0	0	0	0	0
2316	1	0.1015223	17	0	0	0	0	0	3
3317	1	0.1008967	38	-66	0	0	0	0	11
5052	1	0.0995455	0	-40	0	-4	0	-3	-58
1523	1	0.0951064	69	-72	0	0	0	0	-6
1709	1	0.0950096	31	-27	0	0	0	0	-2
1909	1	0.0949265	0	-35	0	0	0	0	-16
387	1	0.0941401	61	-12	0	0	0	0	2
5313	1	0.0935990	0	-22	1	1	3	3	1
412	1	0.0935846	5	-31	0	0	0	0	-1
105	1	0.0935244	26	-13	0	0	0	0	0
4893	1	0.0927391	34	-30	0	-1	0	-3	-3
1823	1	0.0926003	18	-1	0	0	0	0	-2
2003	1	0.0910678	62	13	0	0	0	0	4
1456	1	0.0910644	0	0	0	0	0	0	0
430	1	0.0909350	42	1	0	0	0	0	0
5460	1	0.0885247	0	-43	0	0	0	0	-22