2nd_assignment

Cai Yadong

2024-11-07

The second homework for MEM

Question 1: 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.

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

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

b. Compute the mean, median, and mode.

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

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

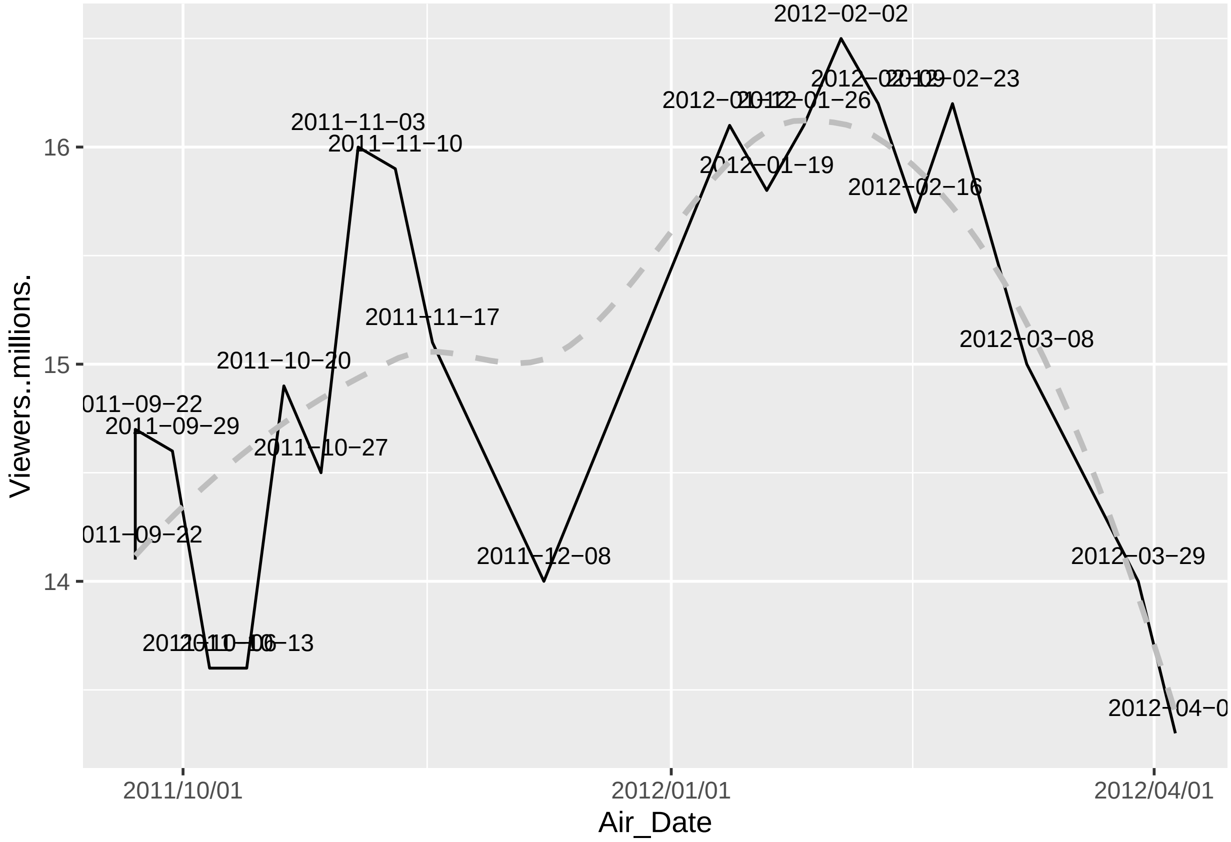
```
## [1] 13.6 14.0 16.1 16.2
```

c. Compute the first and third quartiles.

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

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

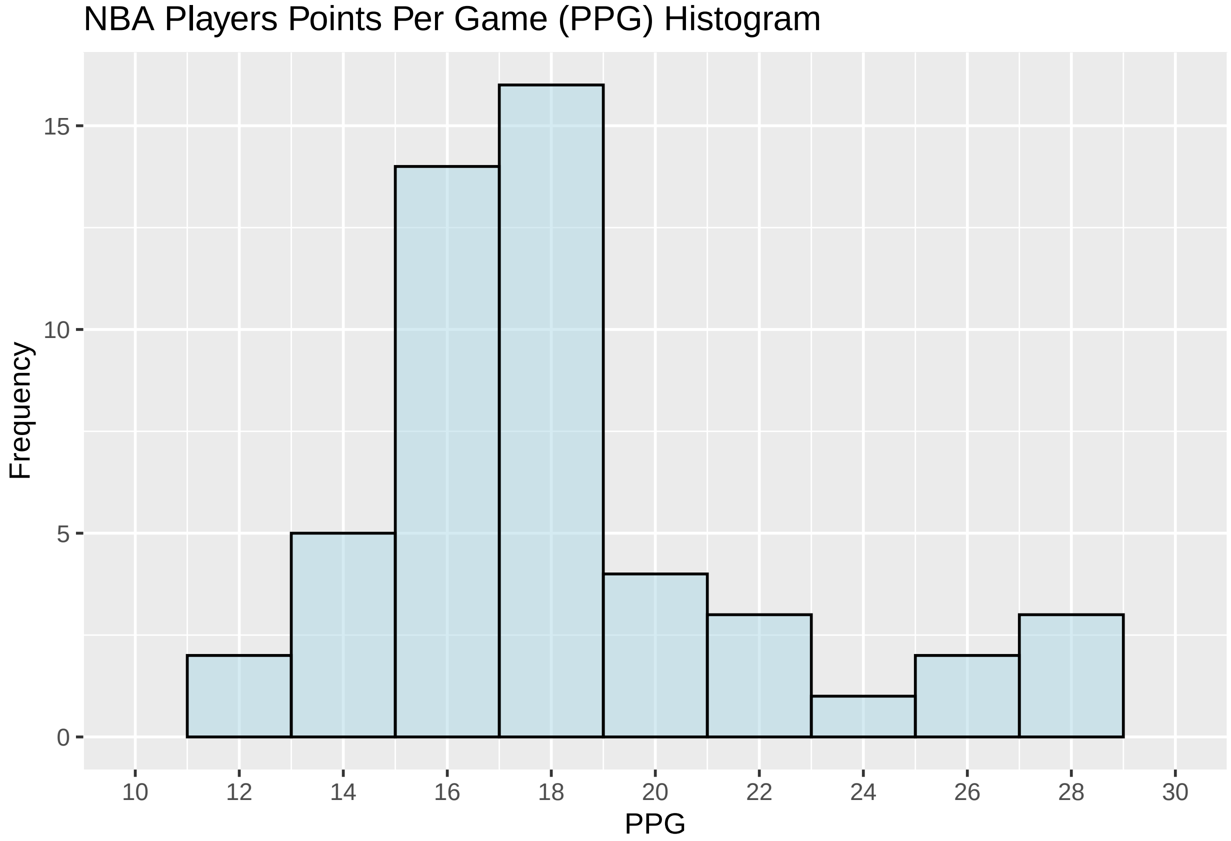
Overall, viewership has grown over a long time from Sep 2011 to Feb 2012 and declined quickly in Mar and Apr. You can know it based on the grey trend dashed line in below chat. However, during this period, there were also several fluctuations in the number of viewers. In the beginning, it has declined from Sep to Oct in 2011. Then it has grown from Oct 20 to Nov, and got a higher viewer record at 16 millions on Nov 3, 2011. After that, it has declined again until a lower record at 14 millions on Dec 8, 2011. In the period of Christmas and New Year's Day, there is no "Big bang theory" scheduled. After that, it has remained a higher viewers in Jan and Feb, 2012. Especially on Feb 2, 2012, it got a highest viewer record at 16.5 millions. With the end of winter and warming weather, the viewers has been declined again and again until at 13.3 millions on Apr 5, 2012.



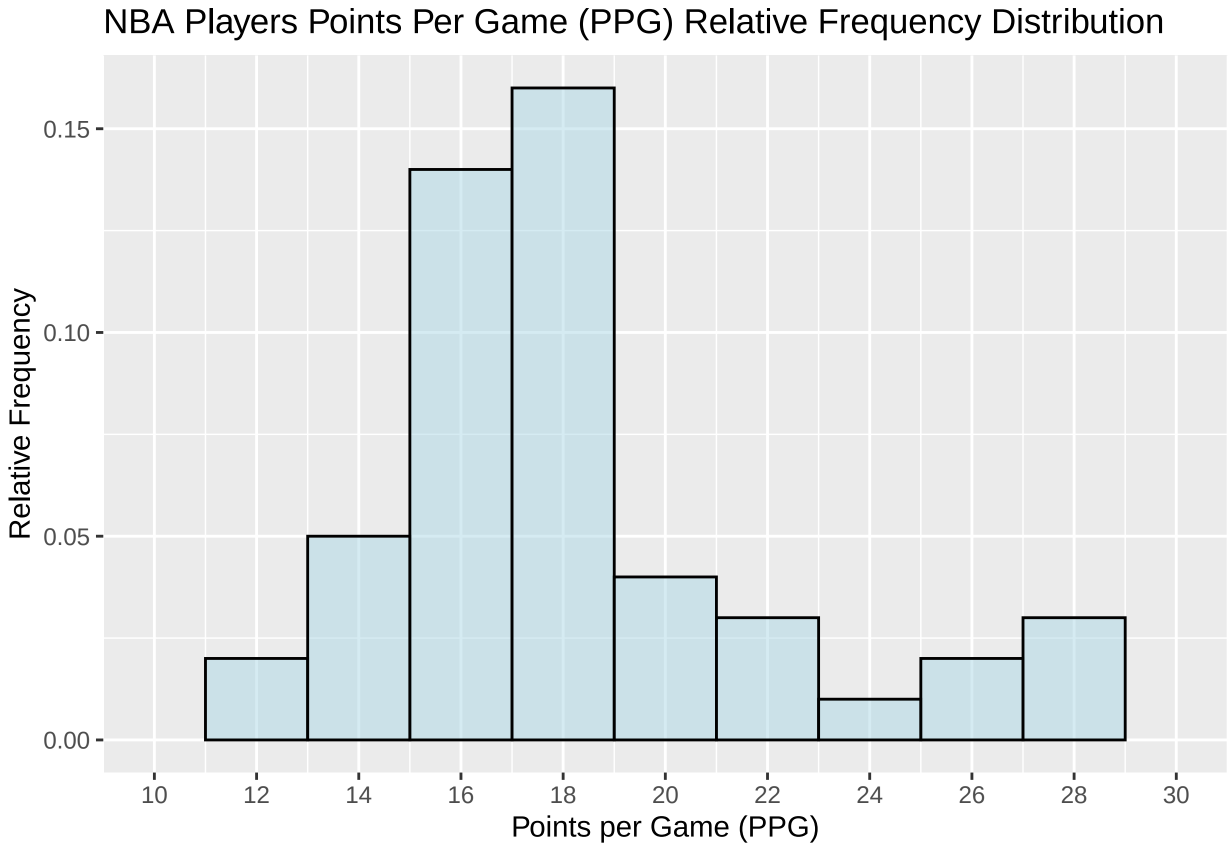
Question 2: NBAPlayerPts.

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.

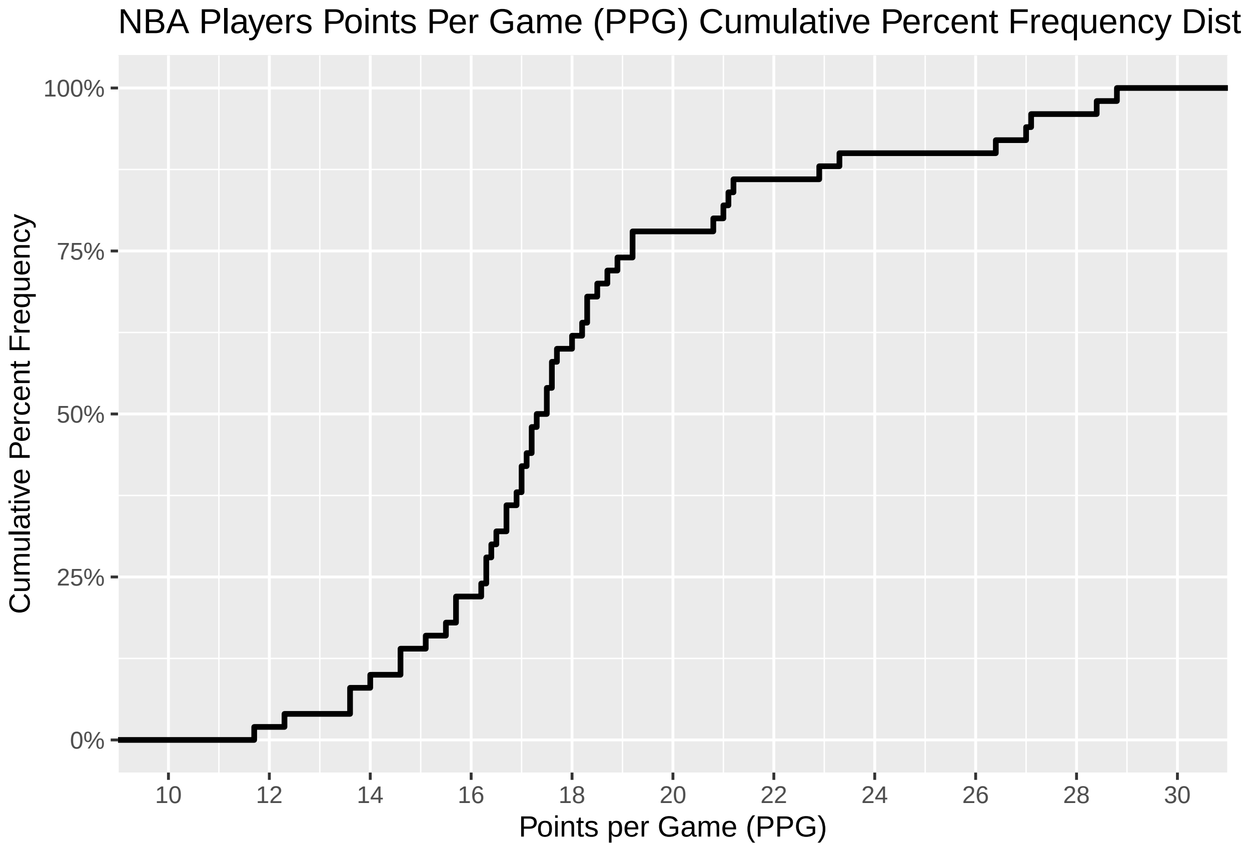
a. Show the frequency distribution.



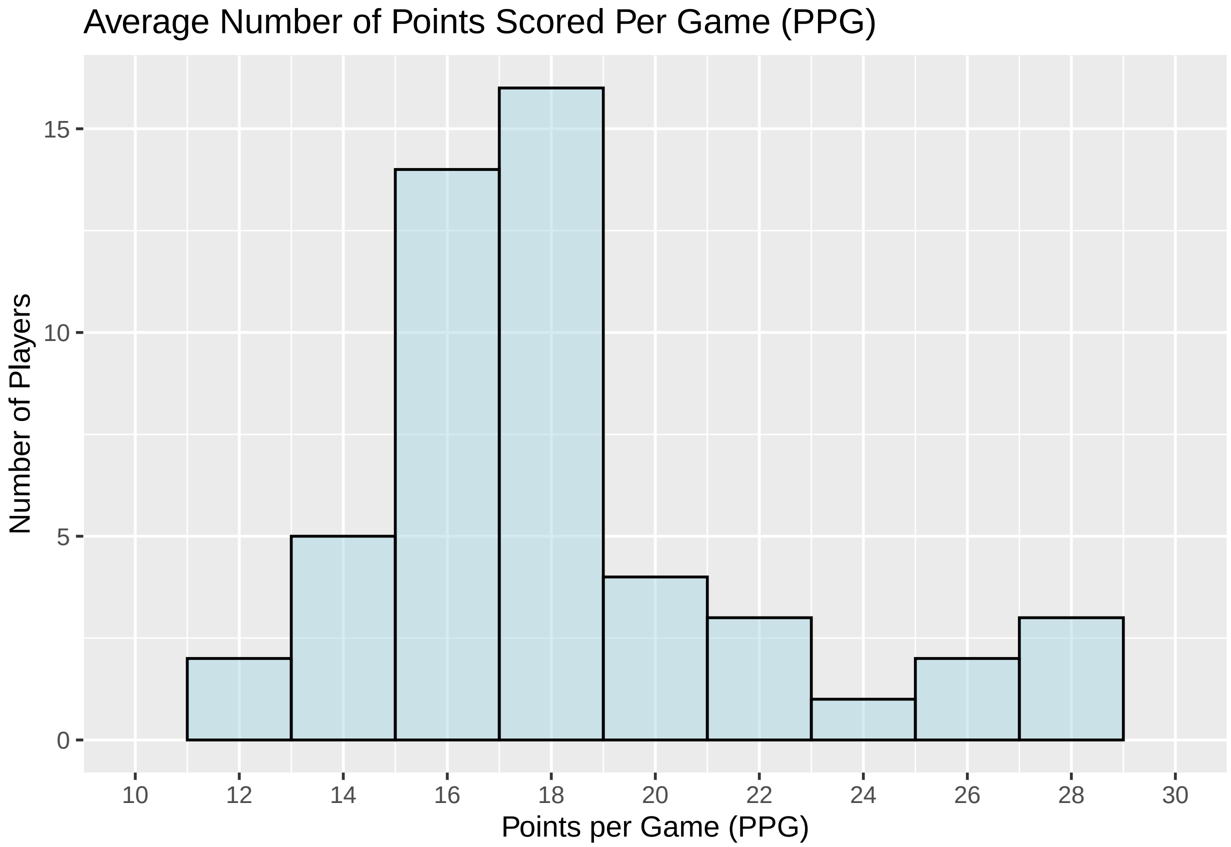
b. Show the relative frequency distribution.



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.

The frequency distribution is right skewed as its skewness(see below) is over 0.

```
## [1] 1.124025
```

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

```
## [1] "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?

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

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

```
## [1] "79%"
```

Question 4: Young Professional Magazine

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:

What is your age?

Are you: Male_____________ Female________________

Do you plan to make any real estate purchases in the next two years?

Yes_________ No_________

What is the approximate total value of financial investments, exclusive of your

home, owned by you or members of your household?

How many stock/bond/mutual fund transactions have you made in the past year?

Do you have broadband access to the Internet at home? Yes_________ No_________

Please indicate your total household income last year. ________________

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.

View cleaned table “professional_cleaned”

```
## 'data.frame':    410 obs. of  8 variables:
##  $ Age                     : int  38 30 41 28 31 32 32 26 26 34 ...
##  $ Gender                  : chr  "Female" "Male" "Female" "Female" ...
##  $ Real.Estate.Purchases.  : chr  "No" "No" "No" "Yes" ...
##  $ Value.of.Investments....: int  12200 12400 26800 19600 15100 39700 21900 41900 16100 18400 ...
##  $ Number.of.Transactions  : int  4 4 5 6 5 3 2 2 4 11 ...
##  $ Broadband.Access.       : chr  "Yes" "Yes" "Yes" "No" ...
##  $ Household.Income....    : int  75200 70300 48200 95300 73300 123400 73900 54300 93100 60100 ...
##  $ Have.Children.          : chr  "Yes" "Yes" "No" "No" ...
```

Summary “Age”, “Value.of.Investments....”, “Number.of.Transactions”, “Household.Income....”

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   19.00   28.00   30.00   30.11   33.00   42.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   18300   24800   28538   34275  133400
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   6.000   5.973   7.000  21.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16200   51625   66050   74460   88775  322500
```

Check percentage of “Male” and “Female” of “Gender”

```
## [1] "56%"
```

Check percentage of “Yes” and “No” of “Real.Estate.Purchases.”

```
## [1] "44%"
```

Check percentage of “Yes” and “No” of “Broadband.Access.”

```
## [1] "62%"
```

Check percentage of “Yes” and “No” of “Have.Children.”

```
## [1] "53%"
```

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
```

```
## [1] 71079.26 77839.77
## 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] 0.5753252 0.6710862
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 0.4845521 0.5830908
## attr(,"conf.level")
## [1] 0.95
```

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

Yes, “Young Professional” be a good advertising outlet for online brokers, especially for financial investments brokers.

In general, young people have three major investment tendencies, with the required amounts ranging from large to small: real estate(house), financial investments, and children education.

Firstly, “Young Professional” described the family situation, willingness to buy a house, annual income, and financial investment willingness of young people (20-40), as well as the feasibility of online transactions.

Secondly, from the results of the survey, we can know the investment tendencies for young people and their actual affordability online.

Investment Tendencies:

a, The investment tendencies on real estate 44% of young people have the investment tendencies on real estate(house).(95% confidence intervals from 39% ~ 49%) b, The investment tendencies on Financial Median of the approximate total value of Financial investment from young people is 24,800. (Range 0 ~ 133K) Median of No. of Transactions on Financial investment is 6 in the past year. c, The investment tendencies on children education 53% of young people have children. (95% confidence intervals from 48% ~ 58%)

Nearly half of the people have the willingness to invest in real estate, the vast majority have a tendency and practical experience in Financial investment, and more than half of them invest in children education with responsibility.

Actual Affordability Online:

Median of total household income is 66050. 62% of young people have the ability to invest online. (95% confidence intervals from 57% ~ 67%)

Basically, for most of young people, they are not affordable to invest in real estate with current total household income. However, they have the ability and willing in Financial investment. If they have children, they can consider investing a small portion of their income in children education.

In a summary, “Young Professional” be a excellent advertising outlet for online brokers, especially for financial investments brokers. On the other hand, for Children education brokers, it’s also a good advertising outlet. They can also find their potenial customer from there.

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

Exactly, yes. As I said from last question, more than half of young parents have ability to invest on their children with variant purpose. Educational software and computer games are two of the items.

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

Types of Articles of Interest to Readers of “Young Professional”

Young professionals, typically in the early to mid-stages of their careers, have a keen interest in a variety of topics that can support their professional growth, financial well-being, and children education. Here are some article themes that could be particularly appealing to this demographic:

Career Development and Skill Enhancement to increase personal income

Career Planning: Guidance on setting and achieving both short-term and long-term career goals, including strategies for career advancement and personal branding. Skill Enhancement: Information on the latest vocational skills training resources, such as online courses, seminars, and workshops, to stay relevant and competitive. Leadership Development: Tips and insights on improving leadership skills, including effective team management, project management, and communication techniques. Industry Trends: Analysis of the latest trends and developments within specific industries, helping young professionals to stay informed and maintain their edge in the job market.

Personal Finance and Investment

Financial Planning: Advice on creating a personal budget, managing debt, and establishing an emergency fund to build a strong financial foundation. Introduction to Investment: An overview of basic investment concepts, such as stocks, bonds, and mutual funds, along with practical steps on how to start investing. Real Estate Investment: Discussion on the pros and cons of buying versus renting, and strategies for building wealth through real estate investments. Retirement Planning: Early retirement planning tips, including the importance of starting a retirement account and choosing the right pension or retirement plan.

Children Education and Family Financial Planning

Online Education Resources: Recommendations for high-quality online education platforms and applications that can help children supplement their learning outside of school hours. Extracurricular Tutoring: Exploration of the benefits and considerations of enrolling children in extracurricular tutoring, and advice on selecting the right tutoring services. Education Savings Plans: Introduction to various education savings plans, such as the 529 Plan in the USA, and other financial instruments like education insurance, to help parents prepare for their children future education costs. Financial Planning for Education: Strategies for integrating education expenses into the family budget, ensuring that there is adequate funding for children education while maintaining overall financial stability.

These articles would not only provide valuable information but also actionable steps that young professionals can take to enhance their careers, manage their finances, and support their families' educational needs. By addressing these key areas, "Young Professional" can become a go-to resource for those looking to navigate the complexities of modern professional and personal life.

Question 5: Quality Associate, Inc.

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. H0:u = 12H1:u <> 12 Corrective action will be taken any time H0 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.1
## $Sample.1$p_value
## [1] 0.3127296
##
## $Sample.1$conclusion
## [1] "Do not reject H0"
##
##
## $Sample.2
## $Sample.2$p_value
## [1] 0.4818209
##
## $Sample.2$conclusion
## [1] "Do not reject H0"
##
##
## $Sample.3
## $Sample.3$p_value
## [1] 0.006468822
##
## $Sample.3$conclusion
## [1] "Reject H0"
##
##
## $Sample.4
## $Sample.4$p_value
## [1] 0.03905895
##
## $Sample.4$conclusion
## [1] "Do not reject H0"
```

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  Sample.4
## 0.2203560 0.2203560 0.2071706 0.2061090
```

```
## Sample.1 Sample.2 Sample.3 Sample.4
##     TRUE     TRUE     TRUE     TRUE
```

From above comparison result, the assumption of 0.21 for the population standard deviation appears reasonable.

c. compute limits for the sample mean x around =12 such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if x exceeds the upper limit or if 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 (UCL): 12.11502
```

```
## Lower Control Limit (LCL): 11.88498
```

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?

When the level of significance is increased, essentially, it is easier to reject the null hypothesis (H0). This has several important implications: Increased probability of Type I Error Decreased probability of Type II

Error As increases, the statistical power of the test also increases In some fields, such as medical research, a higher might lead to more frequent false positive results, which can have serious consequences.

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.

```
## [1] "35%"
```

```
## [1] "46.67%"
```

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

```
## [1] 0.2849421 0.4209231
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 0.3854464 0.5496202
## attr(,"conf.level")
## [1] 0.95
```

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

No, the data does not appear that March rental rates for 2008 will be up from those a year earlier.

From the proportion of units rented during the first week of March 2007 and the first week of March 2008, it seems like March rental rates for 2008 will be up from those a year earlier.

However, there are some na or unknown data in 2008. If we compare the total number of rented rooms between 2007 and 2008, they are same as 70(see below data). We cannot determine these unknown data, so we cannot firmly believe March rental rates for 2008 will be up from those a year earlier.

Total March rental rooms in March 2007

```
## [1] 70
```

Total March rented rooms in March 2008

```
## [1] 70
```

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 instruction 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?

From below descriptive statistics, the Median score for these two methods are same. However, the mean score of proposed method is a litter higher than the other. On the other side, the range for proposed method is lower than current method.

```
##     Current         Proposed
##  Min.   :65.00   Min.   :69.00
##  1st Qu.:72.00   1st Qu.:74.00
##  Median :76.00   Median :76.00
##  Mean   :75.07   Mean   :75.43
##  3rd Qu.:78.00   3rd Qu.:77.00
##  Max.   :84.00   Max.   :82.00
```

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

The population mean for proposed method is a little higher than the one for current method.The range of score for proposed method is lower. We can have a basic conclusion, the proposed method is better as the scores are higher and more stable.

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.

Compute the standard deviation for current and proposed method

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

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

Compute the variance for current and proposed method

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

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

Hypothesis test about the equality of population variances for the two training methods H0:S1 = S2, H1: S1 <> S2

```
##
##  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
```

Evan the standard deviation and variance of proposed method is lower than the standard deviation and variance of current method, F test result shows, p-value 0.000578 is less than 0.05. H0 is rejected(H0:S1 = S2, H1: S1 <> S2), the population variances for these two methods are not equal.

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

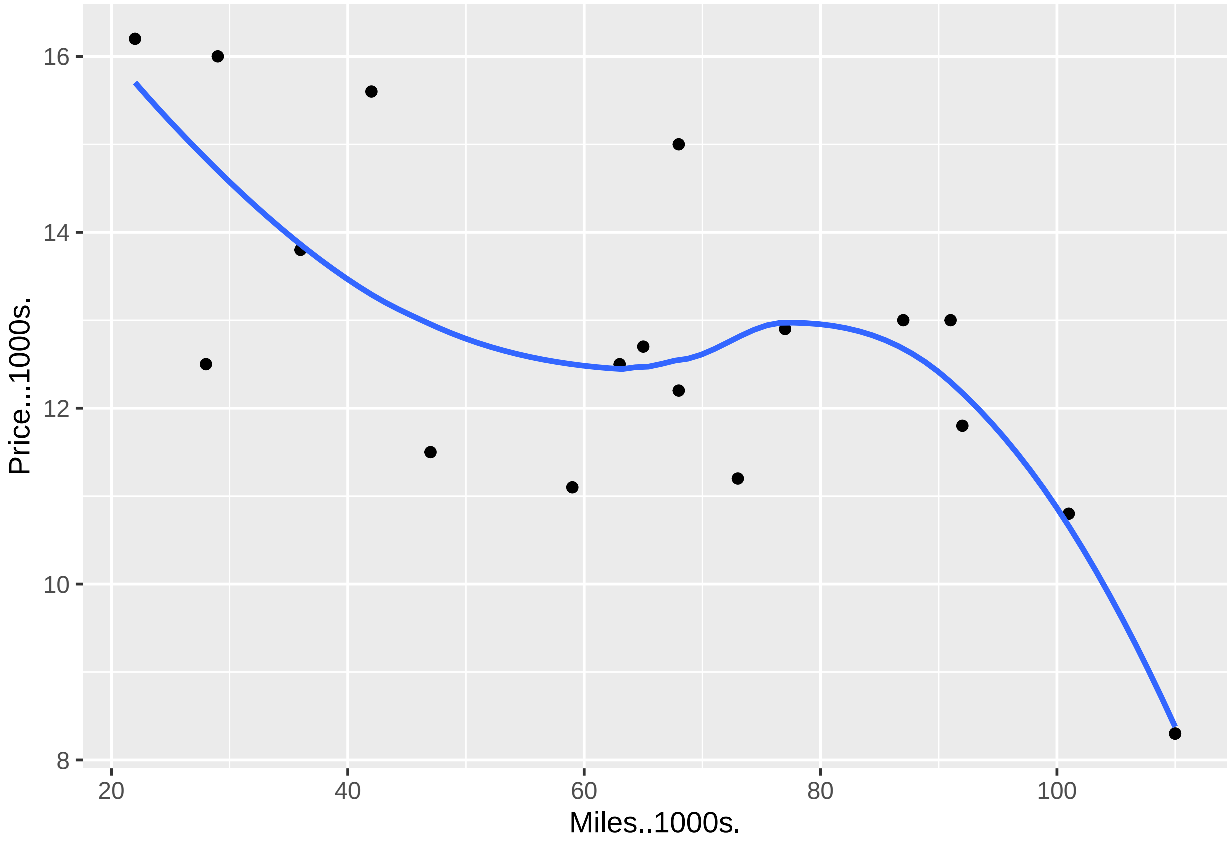
The proposed method is better than current one. Not only the score of students are more stable with similar mean score, the proposed method can save time for some students.

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?

A same group of students take part in two times for current method and proposed method. This testing can remove the variance differences between samples themselves.

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?

There appears to be a negative relationship between price and miles that can be approximated by a straight line. An argument could be made the relationship maybe curve linear at some point.

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
```

Based on the above output, we can write the regression equation: Price...1000s. = 16.47 - 0.059 * Miles..1000s.

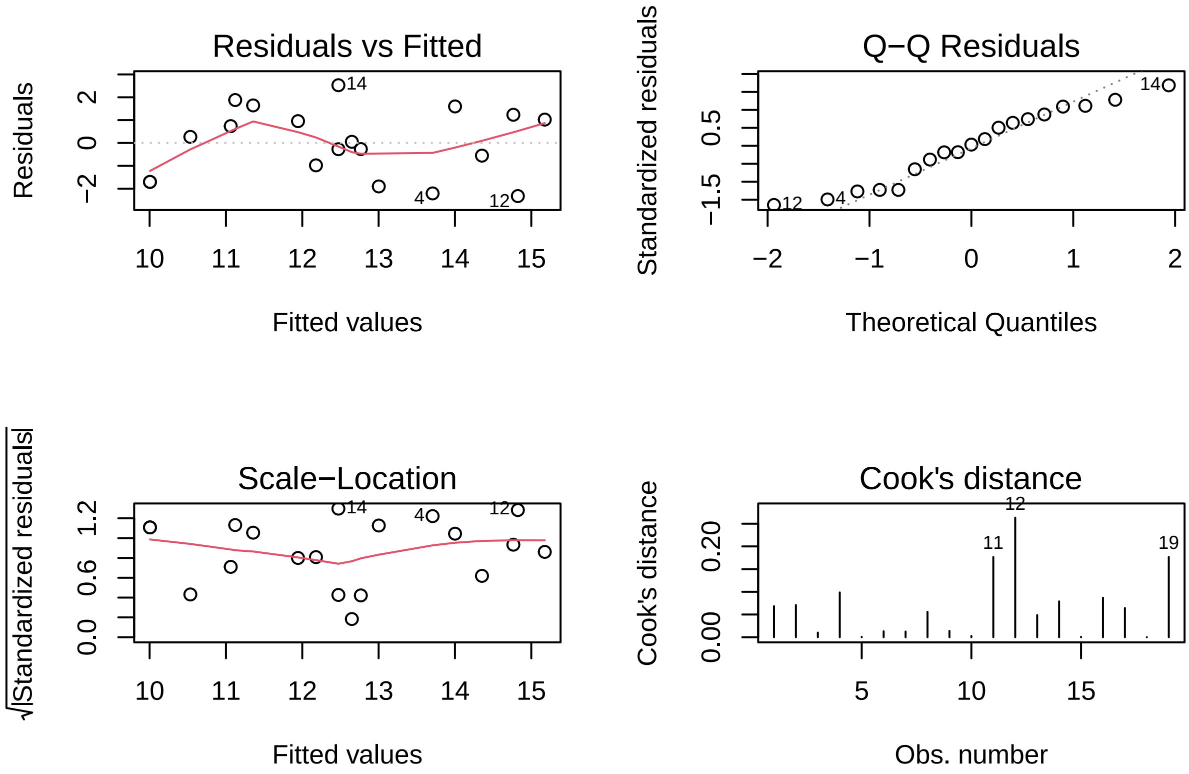
d. Test for a significant relationship at the .05 level of significance.

Setup H0: 1 = 0 no relationship between price and miles, H1: 1 <> 0 there is relationship between price and miles

```
## At the .05 level of significance reject H0. There is relationship between price and miles
## P value: 0.000347511
```

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

R-squared: 0.5387, indicating that the model explained 53.87% of the price changes. Although not very high, it is still a reasonable value. Adjusted R-squared: 0.5115, also reasonable. F-statistic: 19.85, p-value 0.0003475, indicating overall significance of the model. Significance of coefficients: The p-values of intercept and mileage are both very small, indicating that they are both significant. Residual standard error: 1.541, relatively small, indicating that the difference between the predicted and actual values of the model is small



However, from Residual analysis results, there may be two concerns. The model systematically underestimates the value of the dependent variable. There may be a nonlinear relationship between the dependent variable and the independent variable.

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

Slope -0.059 indicates that for every additional 1000 miles of driving, the price of the car is expected to decrease by $59. This slope is statistically significant, reflecting the negative impact of mileage on car prices

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.

According to the regression equation, the predicted price is

```
## [1] 12.93
```

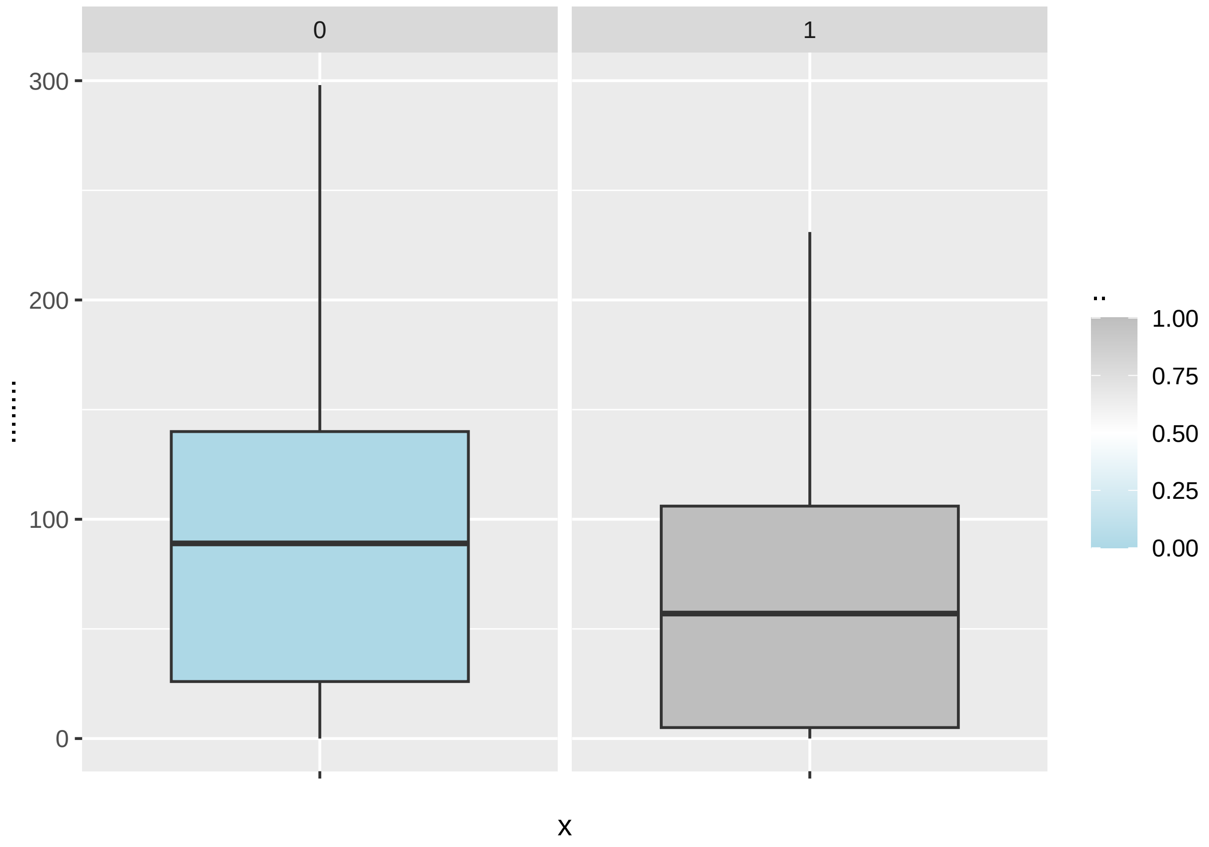
thousand dollars.

In second-hand car market, sellers expect a certain degree of negotiation space. I can use the predicted price as a reference point and adjust the quotation based on specific conditions such as market conditions, vehicle conditions, personal preferences, etc

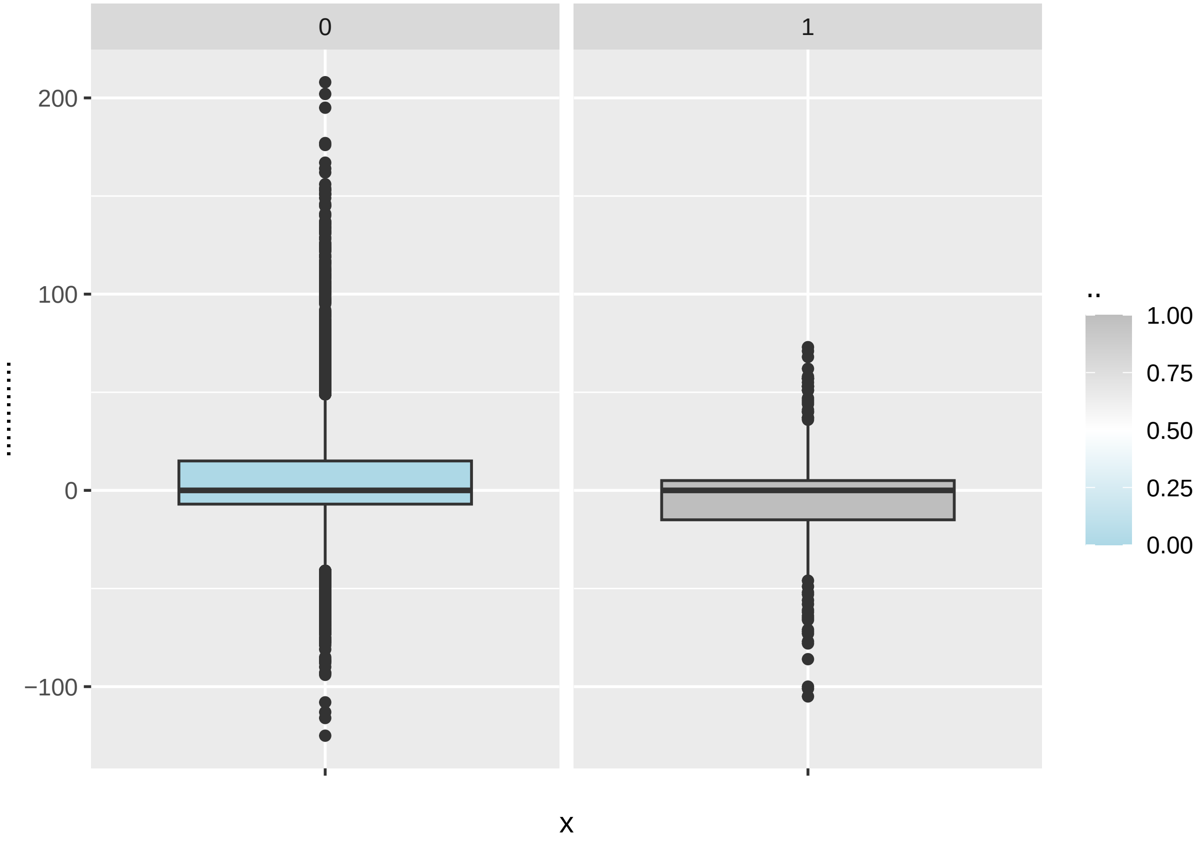
Question 9: Attachment WE.xlsx is customer data of an Internet service provider that provides website services. The data includes the performance of 6347 customers on 11 indicators. Among them, 0 in the "churned" indicator churned customers, 1 indicates non-churned customers, and the meanings of other indicators depend on the variable.

a. By using visualization to explore and compare the behavioral characteristics of churned customers versus non-churned customers, can you identify significant differences between the two groups across several key metrics?

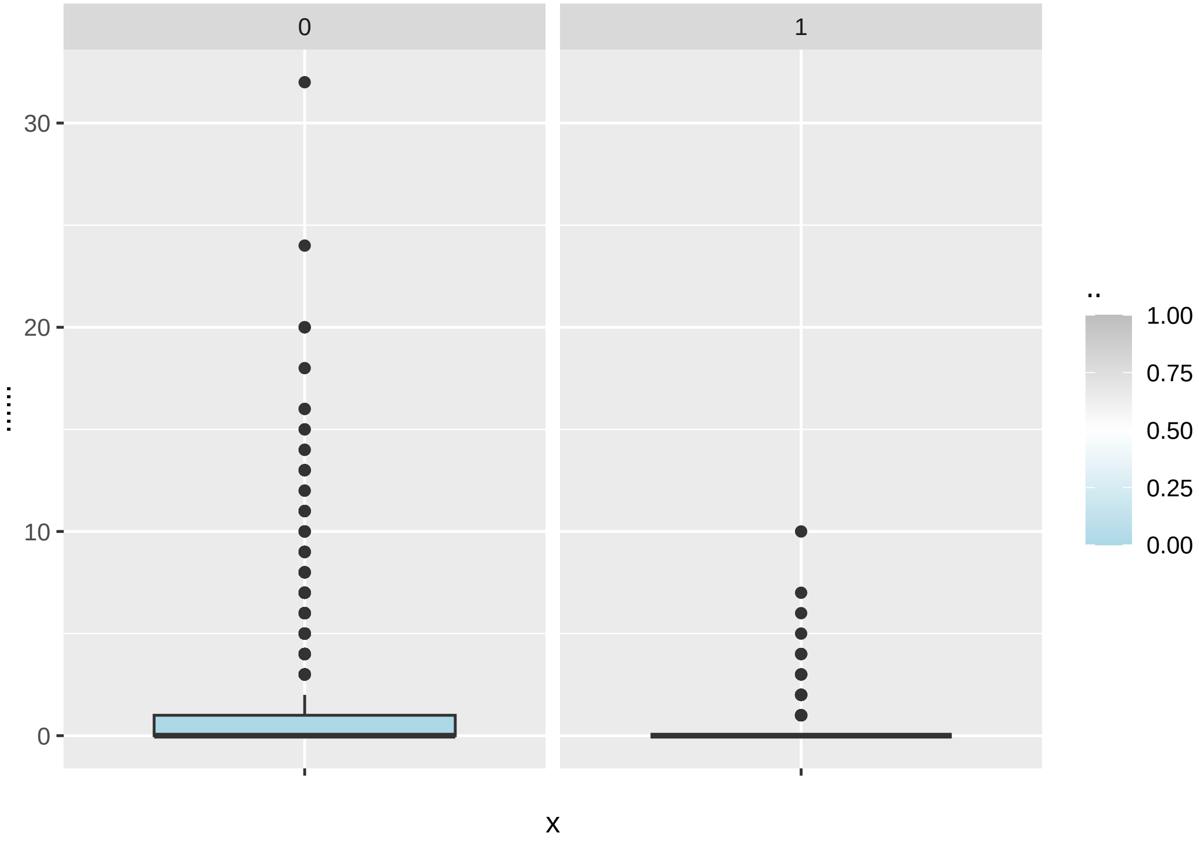
Through the following visual comparisons, there are significant differences between the behaviors of churned and non-churned customers in terms of the current month's customer happiness index, the change in the customer happiness index compared to the previous month, the current month's service priority, and the duration of customer usage.



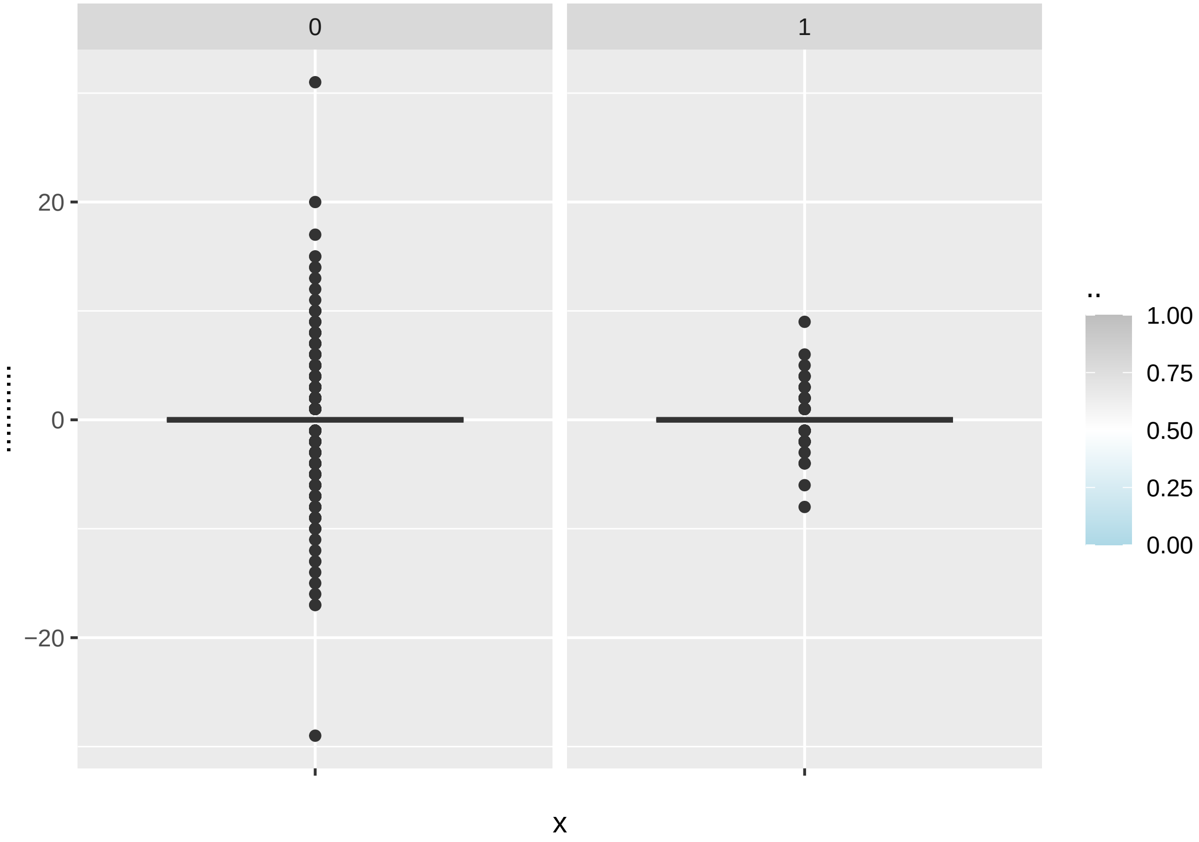
```
## NULL
```



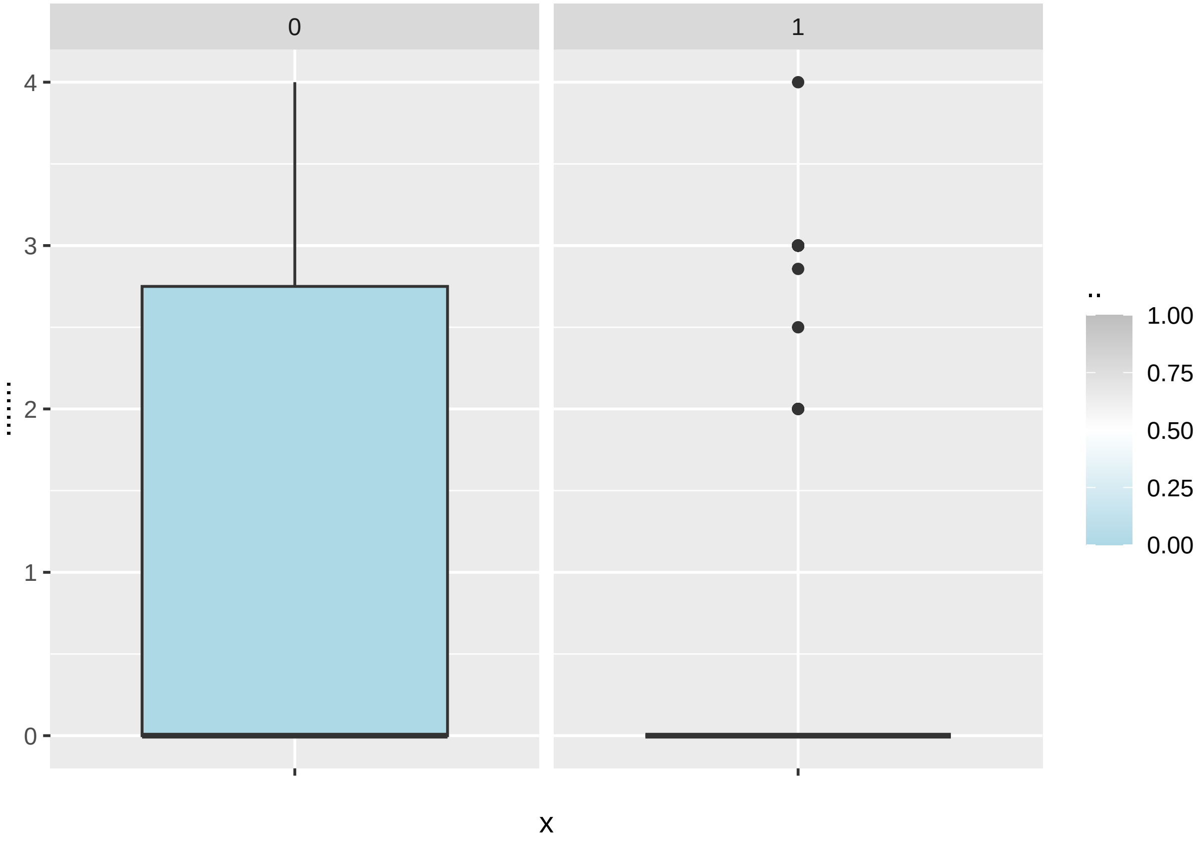
```
## NULL
```



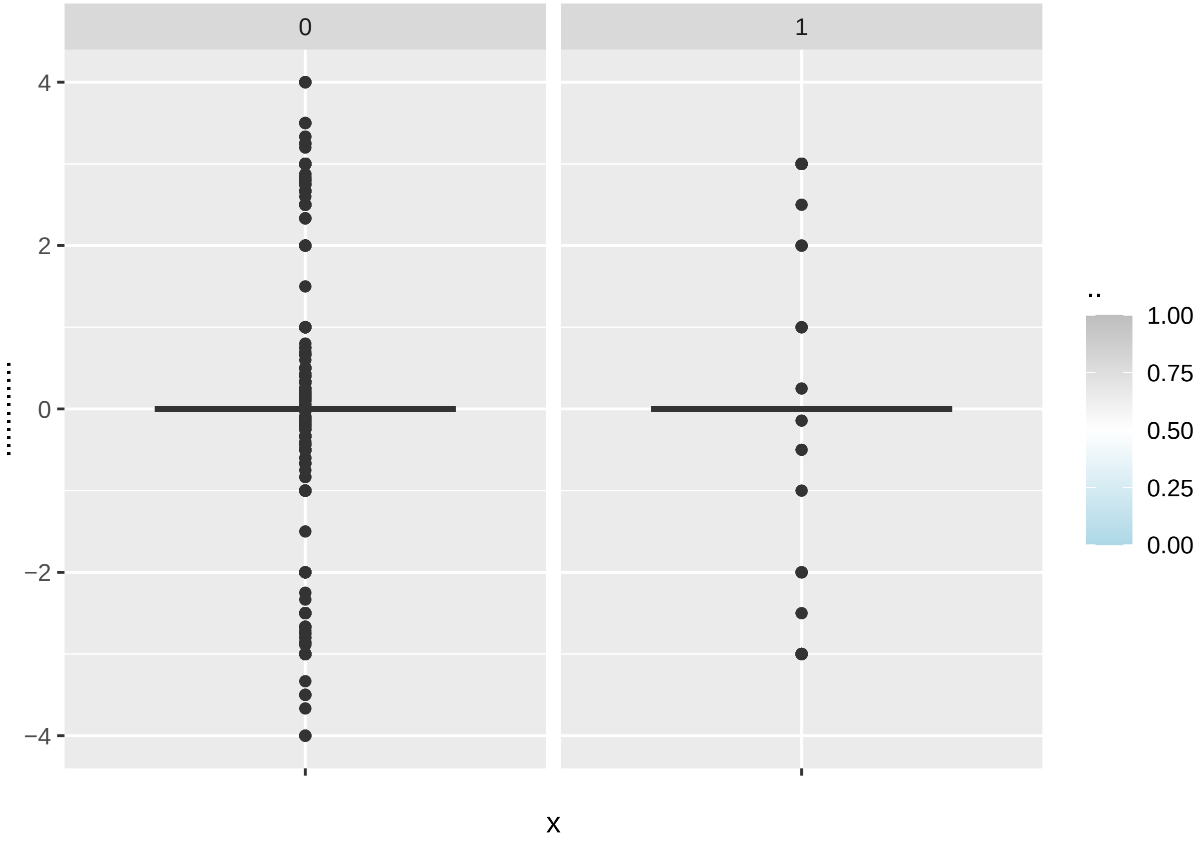
```
## NULL
```



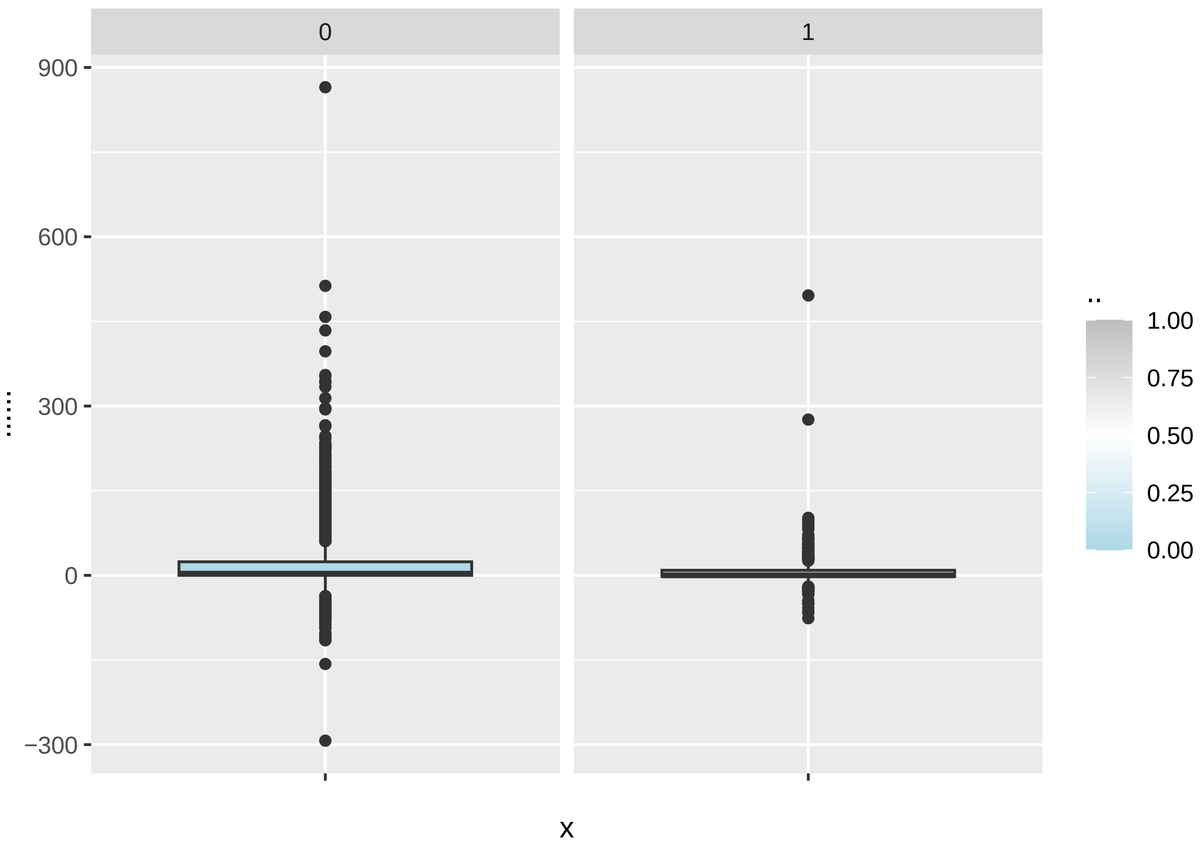
```
## NULL
```



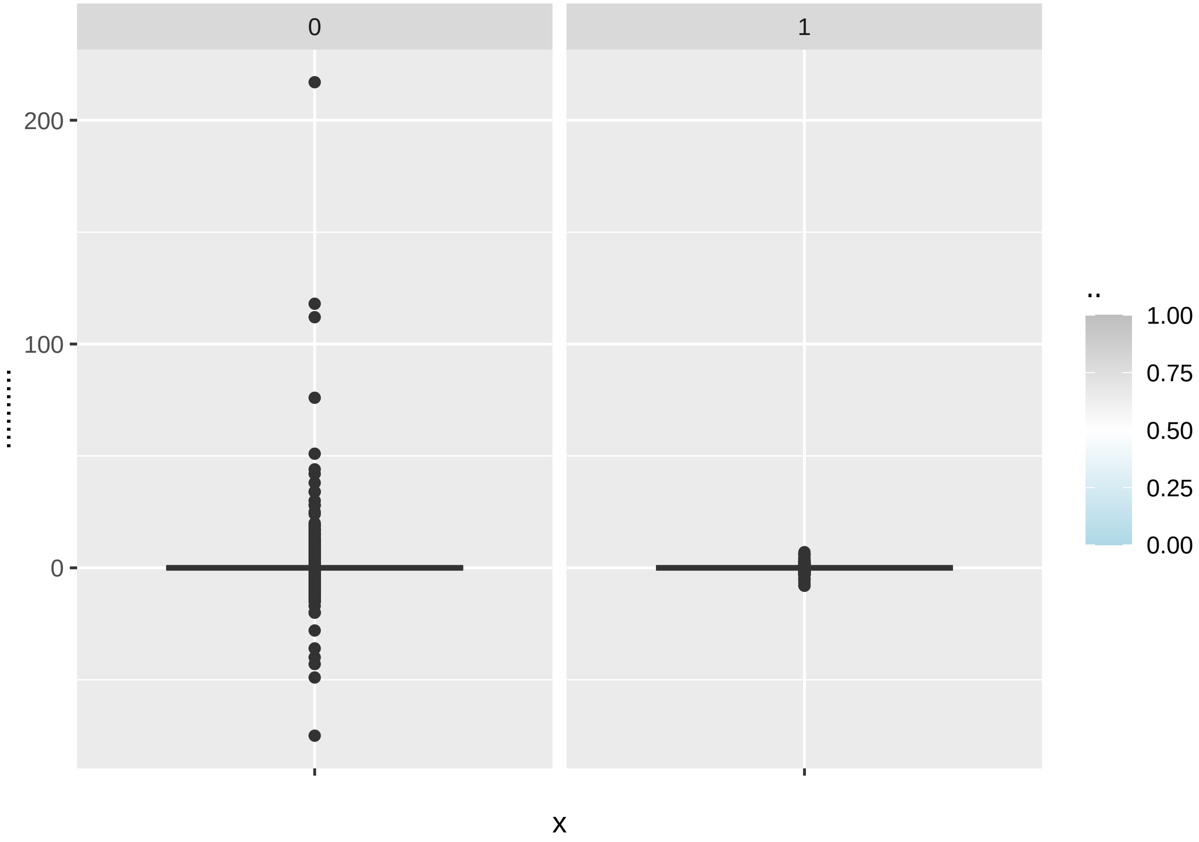
```
## NULL
```



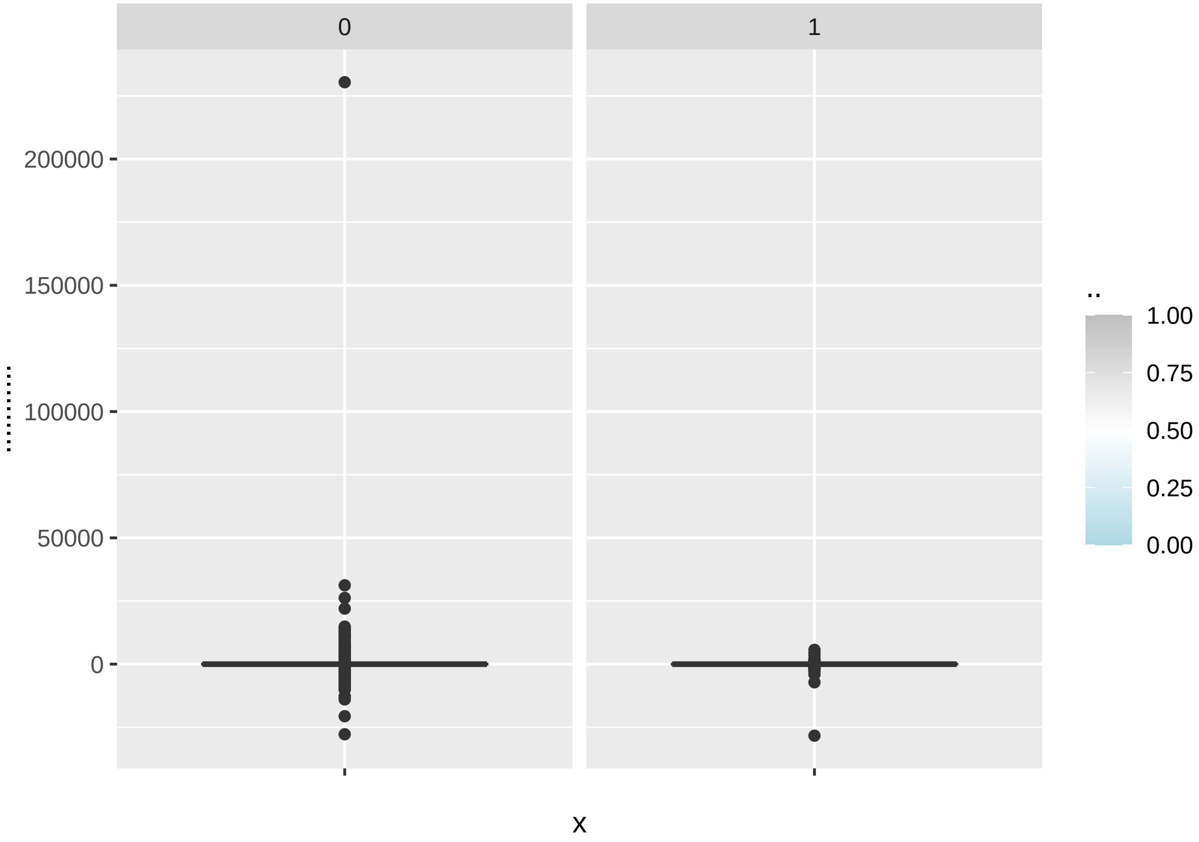
```
## NULL
```



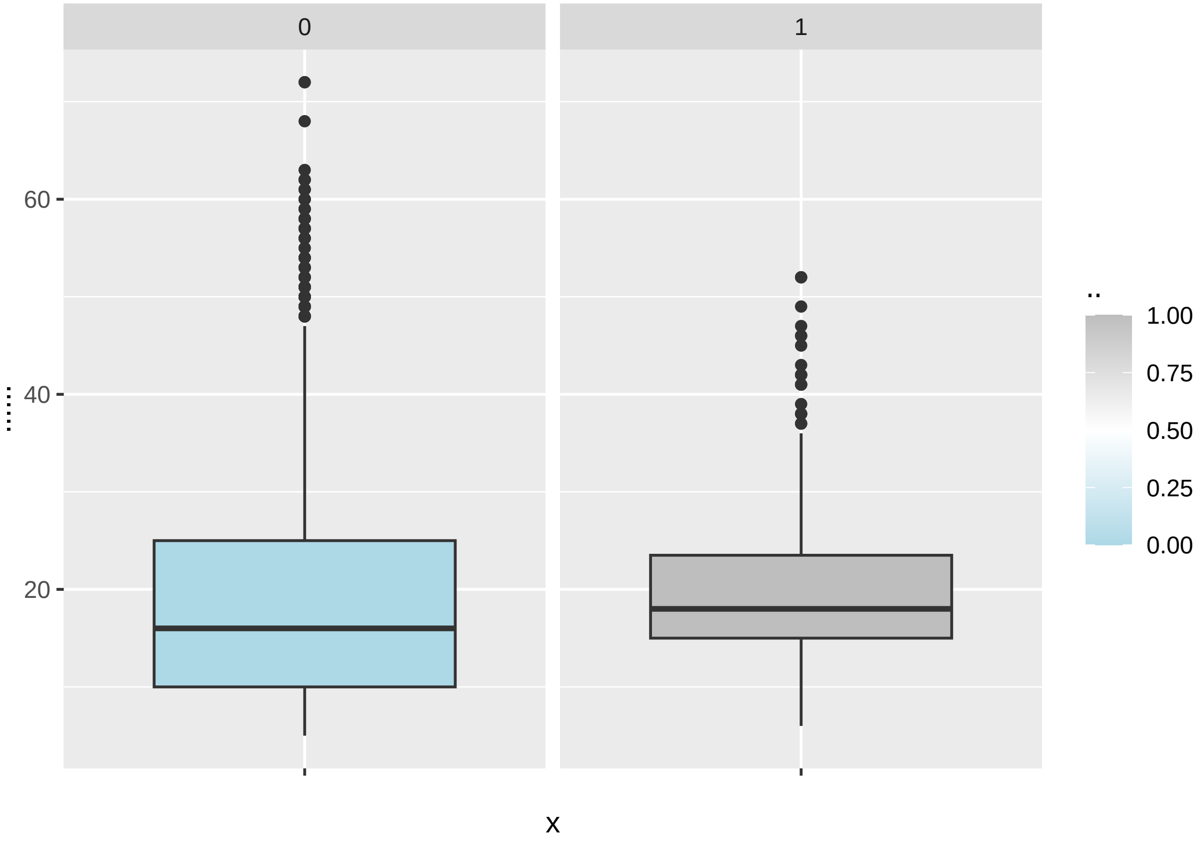
```
## NULL
```



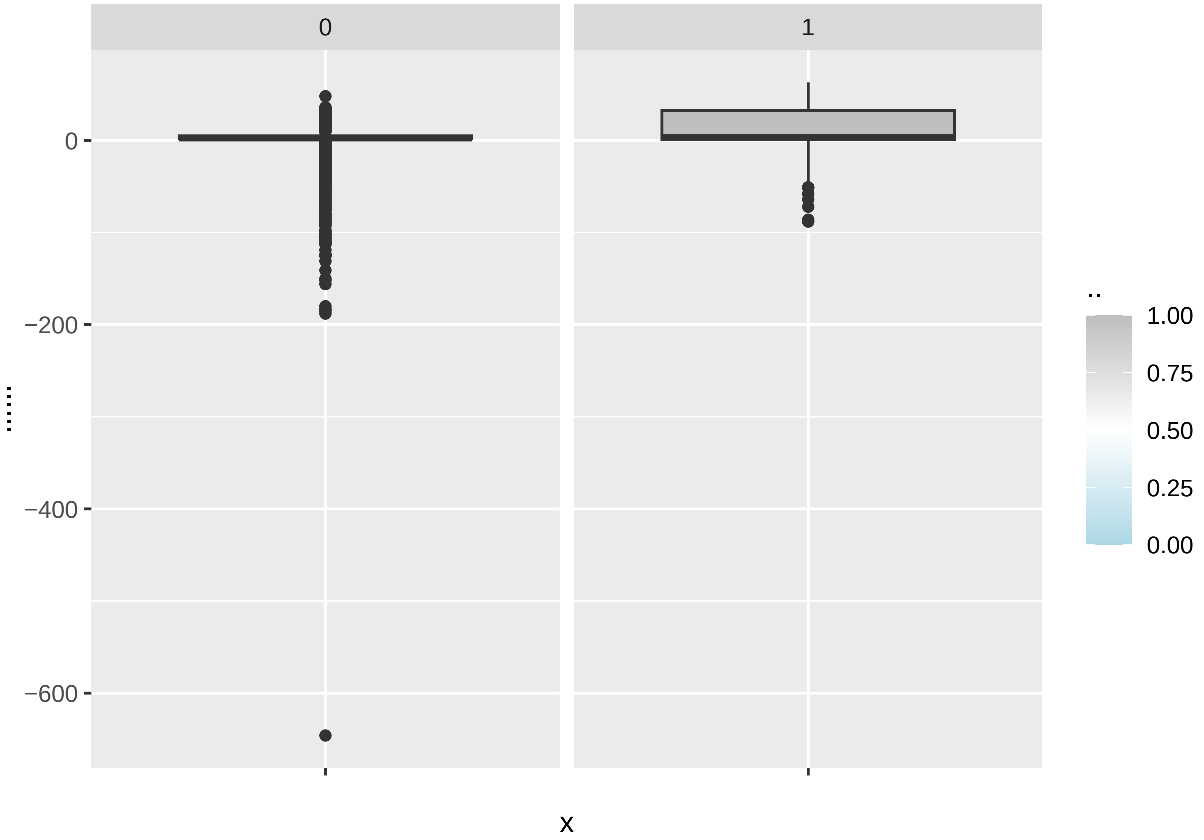
```
## NULL
```



```
## NULL
```



```
## NULL
```



```
## NULL
```

b. Verify whether the aforementioned differences are statistically significant through mean comparison.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   57.00   63.27  106.00  231.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   26.00   89.00   88.61  140.00  298.00
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -105.000  -15.000    0.000   -3.737    5.000   73.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -125.00   -7.00    0.00    5.53   15.00  208.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4996  0.0000  4.0000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.8296  2.7500  4.0000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   15.00   18.00   20.35   23.50   52.00
```

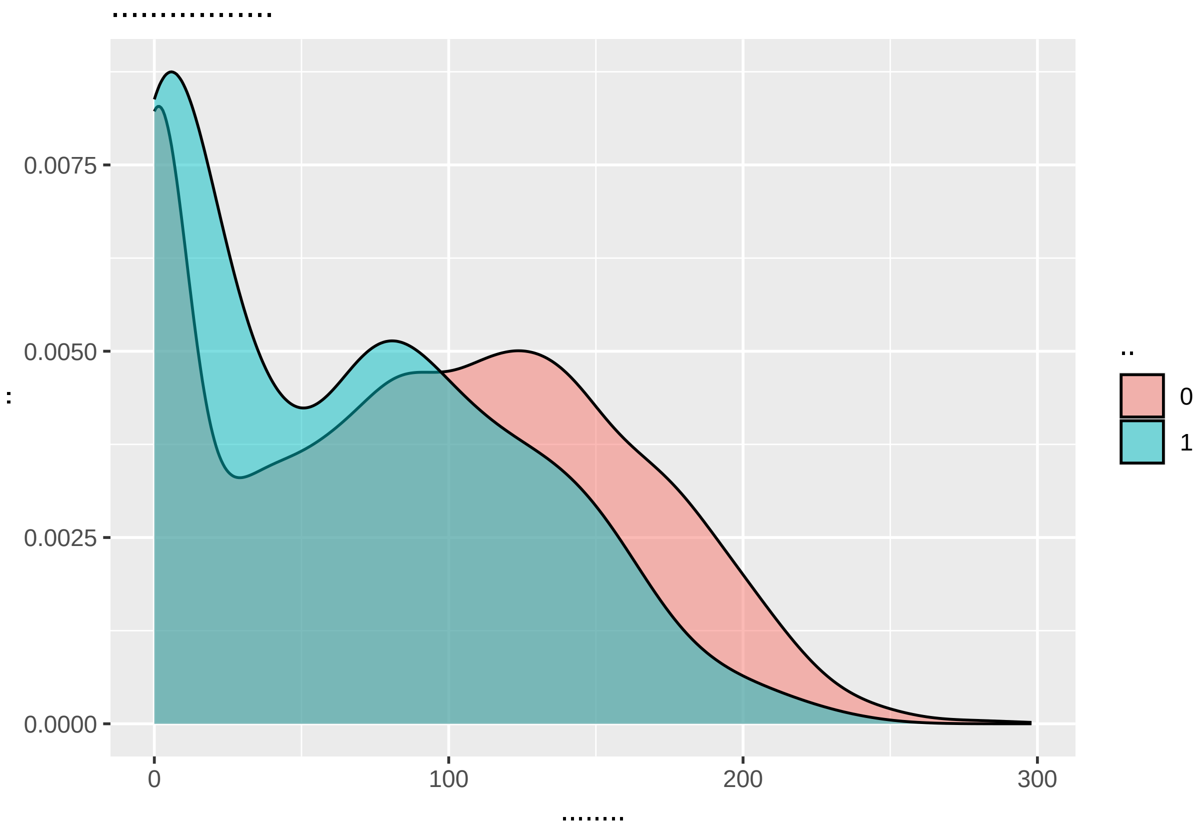
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   10.00   16.00   18.82   25.00   72.00
```

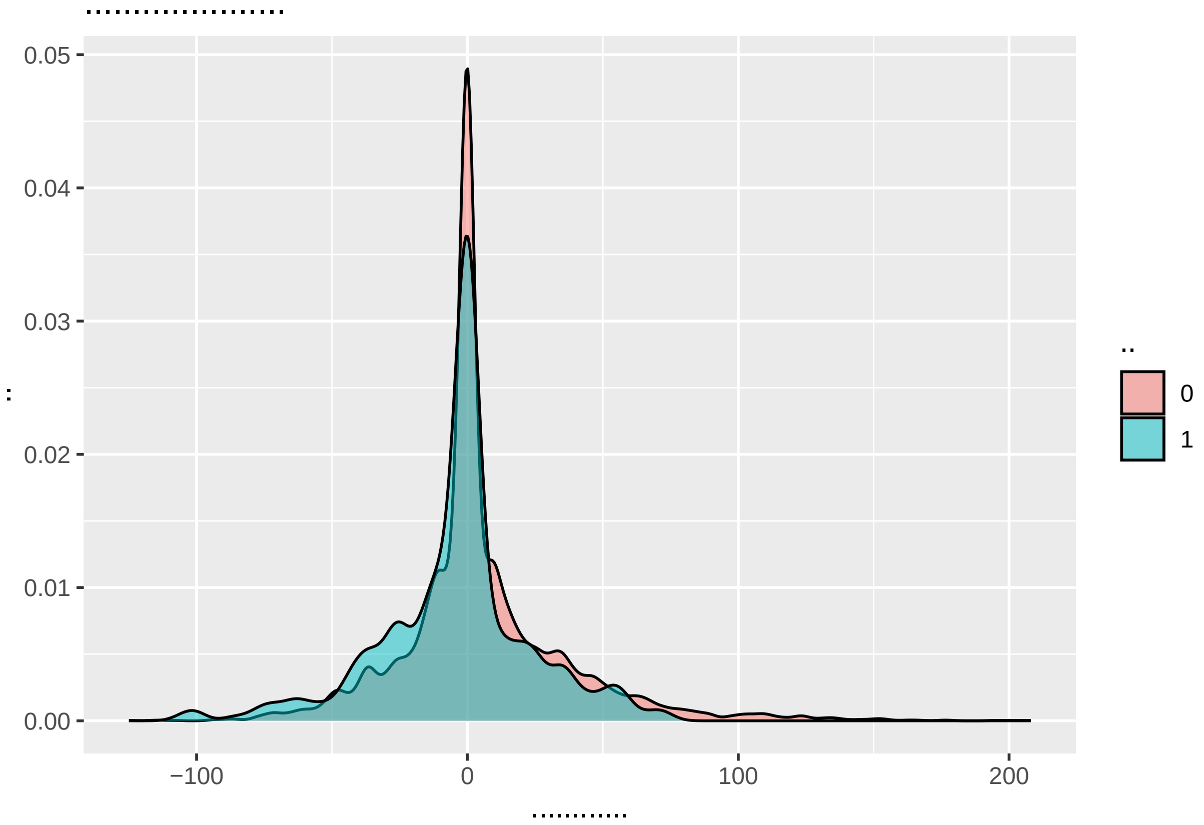
c. Use "churned" as the dependent variable and other variables that you consider important (based steps a and b) as independent variables to establish a regression equation for predicting whether a customer will churn.

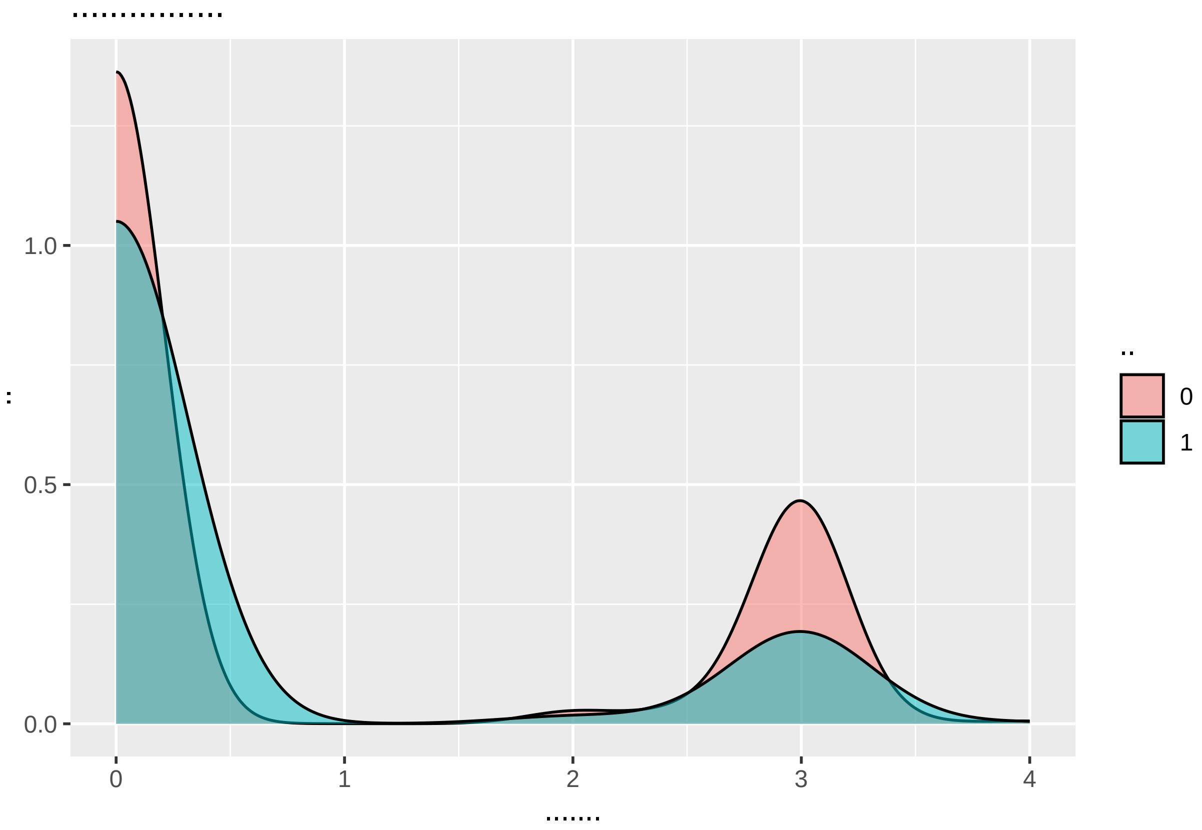
1, Exploratory Data Analysis: a. Examine the relationship between each independent variable and the dependent variable "churn". Review the density plots of each independent variable against the dependent variable "churn". It was found that the current month's customer happiness index, the change in the customer happiness index compared to the previous month, and the current month's service priority have a significant relationship with the dependent variable. (Refer to the detailed figures below for illustration.)

Among other independent variables, the current month's customer support, the change in customer support compared to the previous month, the current month's service priority, the change in service priority compared to the previous month, the change in the number of blog posts compared to the previous month, the increase in visit count compared to the previous month, and the change in visit interval do not show a clear relationship with "churn". The duration of customer usage does not have a significant relationship with "churn". The data for the number of log-in in the current month contains anomalies (negative values) and cannot be used for reference.

b. Select the current month's customer happiness index, the change in the customer happiness index compared to the previous month, and the current month's service priority based on their relationships with the dependent variable "churned".


0.0075
0.0050
0.0025
0.0000
0
100
200
300
0
1


0.05
0.04
0.03
0.02
0.01
0.00
−100
0
100
200
0
1



2, Establish a Logistic Regression Model

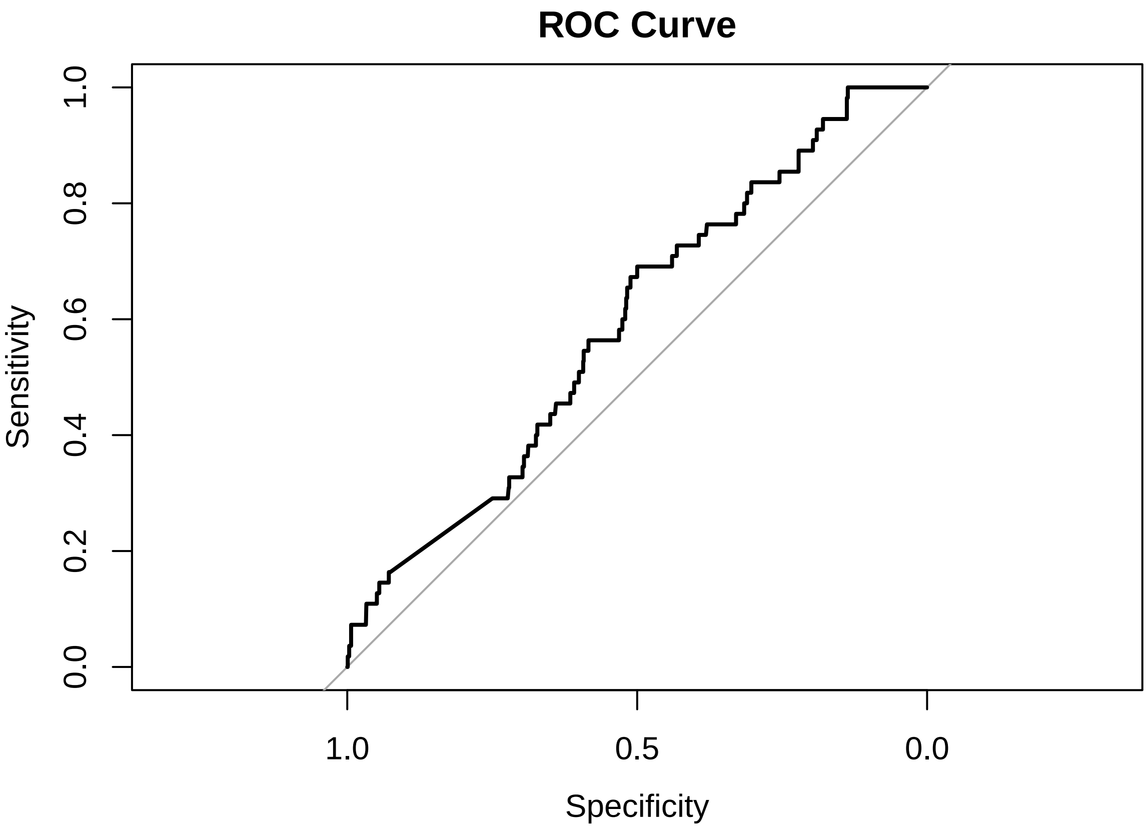
```
##
## Call:
## glm(formula =    ~ ., family = "binomial", data = train_data[,
##     c(selected_variables, " ")])
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.457512   0.093884 -26.176  < 2e-16 ***
##              -0.004845   0.001106  -4.383 1.17e-05 ***
##        -0.009650   0.002392  -4.034 5.48e-05 ***
##                -0.097726   0.060991  -1.602    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2098.3  on 5077  degrees of freedom
## Residual deviance: 2037.3  on 5074  degrees of freedom
## AIC: 2045.3
##
## Number of Fisher Scoring iterations: 6
```

Intercept: -2.502641, Standard Error 0.095908, z-value -26.094, p-value < 2e-16. The intercept is highly significant (p-value < 0.001), indicating that when all independent variables are zero, the predicted log-odds of churn is -2.502641. Current Month's Customer Happiness Index: -0.004991, Standard Error 0.001111, z-value -4.494, p-value 6.99e-06. This variable is highly significant (p-value < 0.001), and the negative coefficient suggests that as the current month's customer happiness index increases, the probability of churn decreases. Change in Customer Happiness Index Compared to the Previous Month: -0.009400, Standard Error 0.002448, z-value -3.840, p-value 0.000123. This variable is also highly significant (p-value < 0.001), and the negative coefficient indicates that a decrease in the customer happiness index compared to the previous month increases the probability of churn. Current Month's Service Priority: -0.049181, Standard Error 0.060061, z-value -0.819, p-value 0.412874. This variable is not significant (p-value > 0.05), and the negative coefficient suggests that the current month's service priority has little effect on customer churn. Residual Deviance: 1993.0, Degrees of Freedom 5075. This is the deviance of the model after including the independent variables. The Residual Deviance is significantly lower than the Null Deviance, indicating that the independent variables make a significant contribution to the model. AIC (Akaike Information Criterion): 2001. A relatively small value, suggesting that the model fits the data well.

3, Model Evaluation

```
## Accuracy: 0.04334121
```

```
##          Actual
## Predicted   -1    0
##         0 1214   55
```



```
## AUC: 0.5921297
```

Accuracy: 0.9495268 This high accuracy indicates that the model performs very well overall on the test set. Accuracy is the proportion of all correct predictions, including both correctly predicted positive (churned) and negative (non-churned) cases. True Negatives: 1204 These are the actual non-churn customers that the model correctly predicted as non-churn. False Negatives: 64 These are the actual churn customers that the model incorrectly predicted as non-churn. AUC (Area Under the ROC Curve): 0.623832 This AUC value indicates that the model has relatively weak discriminatory power

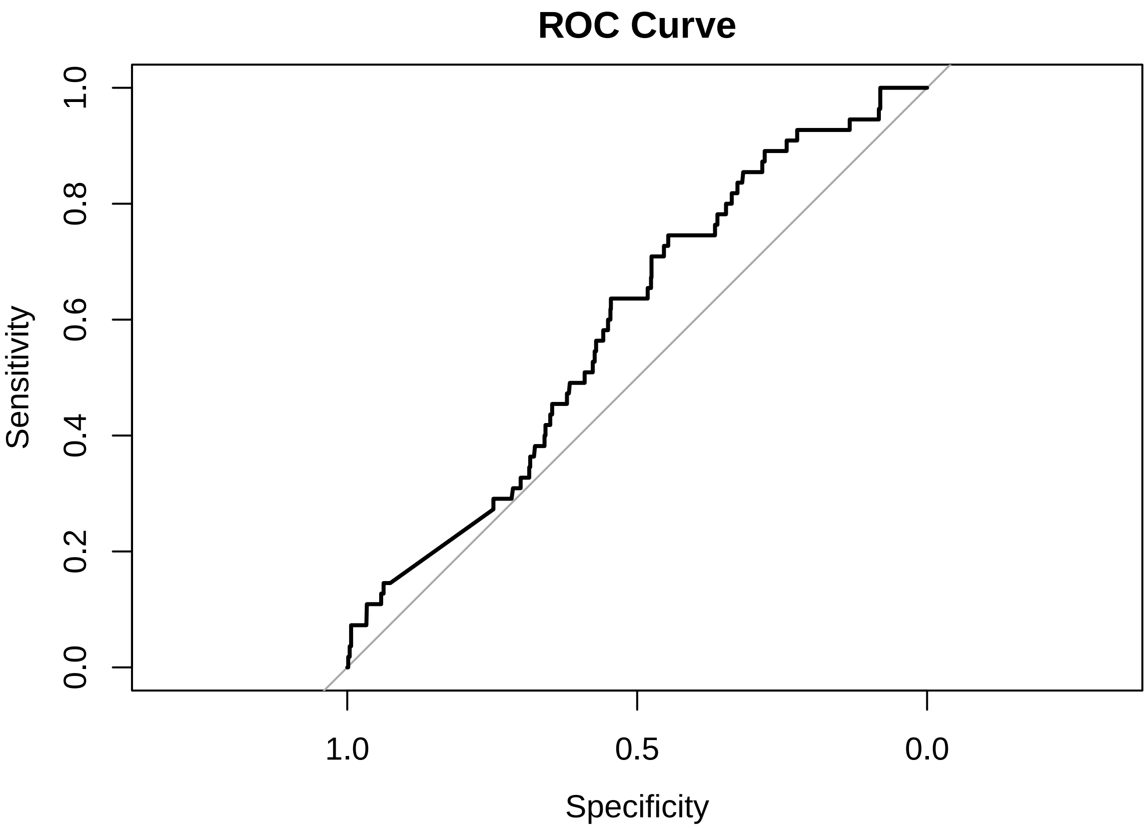
4 Based on the above results, modify the model by removing the insignificant independent variable "current month's service priority" and re-evaluate the model.

```
##
## Call:
## glm(formula =   ~ ., family = "binomial", data = train_data[,
##     c(selected_variables2, " ")])
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.472953   0.093477 -26.455  < 2e-16 ***
##              -0.005468   0.001044  -5.235 1.65e-07 ***
##        -0.010082   0.002354  -4.284 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2098.3  on 5077  degrees of freedom
## Residual deviance: 2040.0  on 5075  degrees of freedom
## AIC: 2046
##
## Number of Fisher Scoring iterations: 6
```

AIC (Akaike Information Criterion): 1999.6, which is lower compared to Model 1.

```
## Accuracy: 0.04334121
```

```
##          Actual
## Predicted   -1    0
##         0 1214   55
```



```
## AUC: 0.5909316
```

Accuracy: 0.9495268 - The accuracy remains the same as in Model 1. True Negatives (, TN) and False Negatives (, FN) are also consistent with Model 1.

AUC (Area Under the Curve): 0.6141313 - This value is slightly lower compared to Model 1.

Conclusion: Model 2 is superior in terms of AIC and is more parsimonious. Although the AUC is slightly lower, in practical applications, a higher accuracy and a lower AIC are often more important. Therefore, Model 2 is selected as the final model.

The logistic regression equation for Model 2 is:

log(P(Churned=0)/P(Churned=1))$= -2.510851 - 0.005302 \times$ Current Month's Customer Happiness Index $- 0.009637 \times$ Change in Customer Happiness Index Compared to the Previous Month

d. Based on the prediction results from the previous step, rank the customers who have not yet churned (churned = 0) by their likelihood of churning, and provide a list of the top 100 user IDs with the highest probability of churning.

```
## # A tibble: 100 x 2
##       ID probability
##    <dbl>       <dbl>
##  1   109       0.229
##  2  1971       0.209
##  3  3340       0.183
##  4   299       0.178
##  5   4191       0.174
##  6   1574       0.166
##  7   2835       0.163
##  8   2481       0.162
##  9   5314       0.161
## 10   2546       0.161
## # i 90 more rows
```