D207: Exploratory Data Analysis

Performance Assessment

Logan Rosemeyer

Western Governor's University

D207: Exploratory Data Analysis

Dr. David Gagner

December 28, 2022

Contents

A1: Question For Analysis

Is the number of times the customer contacted technical support related to whether the customer churned?

A2: Benefit From Analysis

Stakeholders in the organization can benefit from knowing if contacting customer support is related to churning because they can better predict whether a customer will churn. This gives the organization a huge advantage because they will be able to pinpoint which customers are likely to churn and target them to reduce churn rate.

A3: Data Identification

To answer the question from A1, I will need the churn variable as well as the contacts variable. My data size is 10000 rows and 2 columns. The dependent variable is the contacts variable. The independent variable is the churn variable ("What are independent..., n.d.). I will use a one-sided t-test because I have a categorical variable, churn, with two levels, yes and no. I am using one-sided rather than two sided because I want to see if the mean contacts for the did not churn group is less than the mean contact for the did churn group. In other words, I want to see if customers who did not churn had less customer support contacts than those who did churn. ("What statistical analysis…, n.d.)

B1: Code

Code has been attached in a separate file.

B2: Output

The t test-statistic is -0.85583, the degrees of freedom is 4675.9, the p-value is 0.1961. The 95% confidence interval for difference in the means is -Inf to 0.0177. The mean for group No is 0.9891156, while the mean for group Yes is 1.0083019.

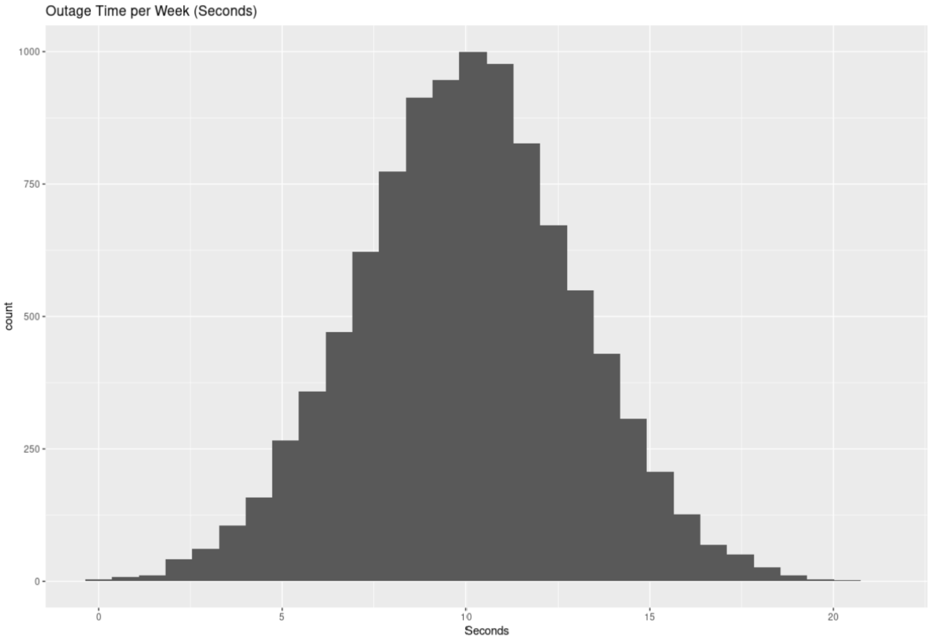
B3: Justification

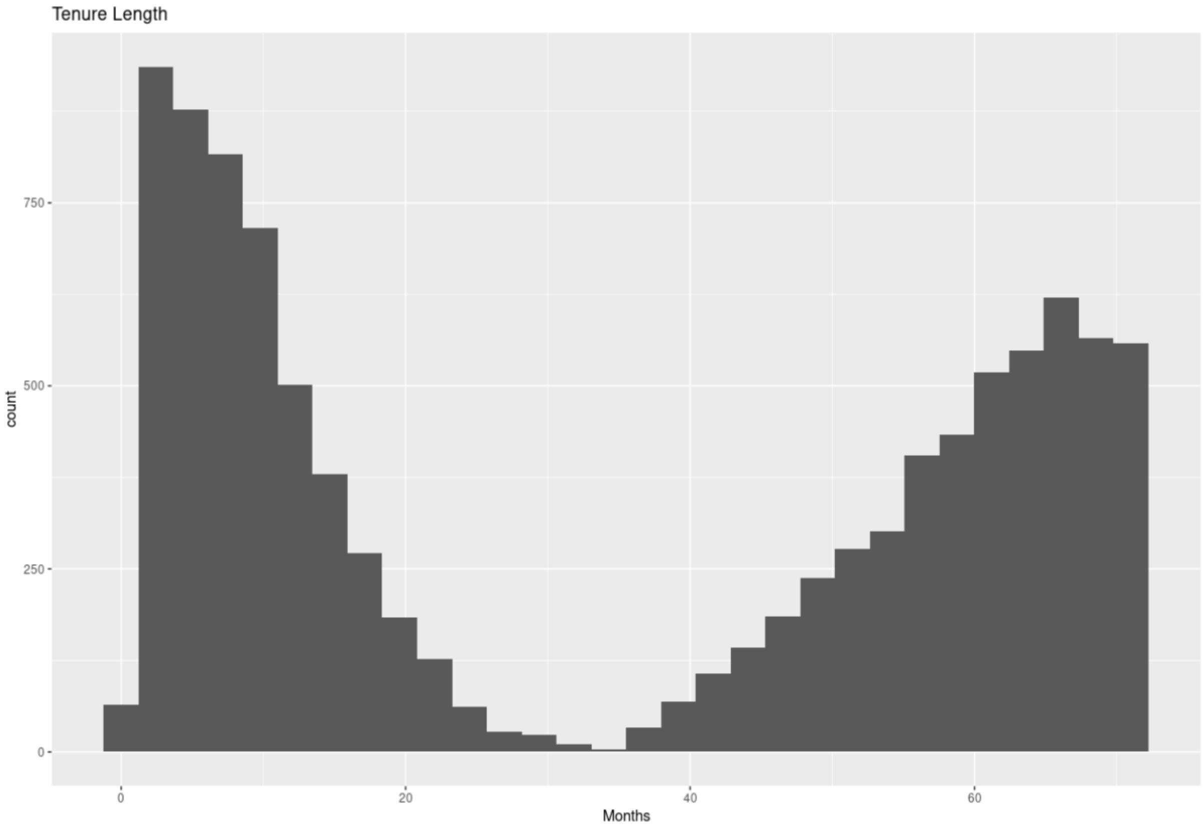
A t-test was selected because I have one numeric and one categorical variable. The categorical variable has two levels, which is why I picked a t-test over ANOVA. This test enables us to see how the mean for each churn group will differ and if there is enough evidence to say that the two means are statistically different.

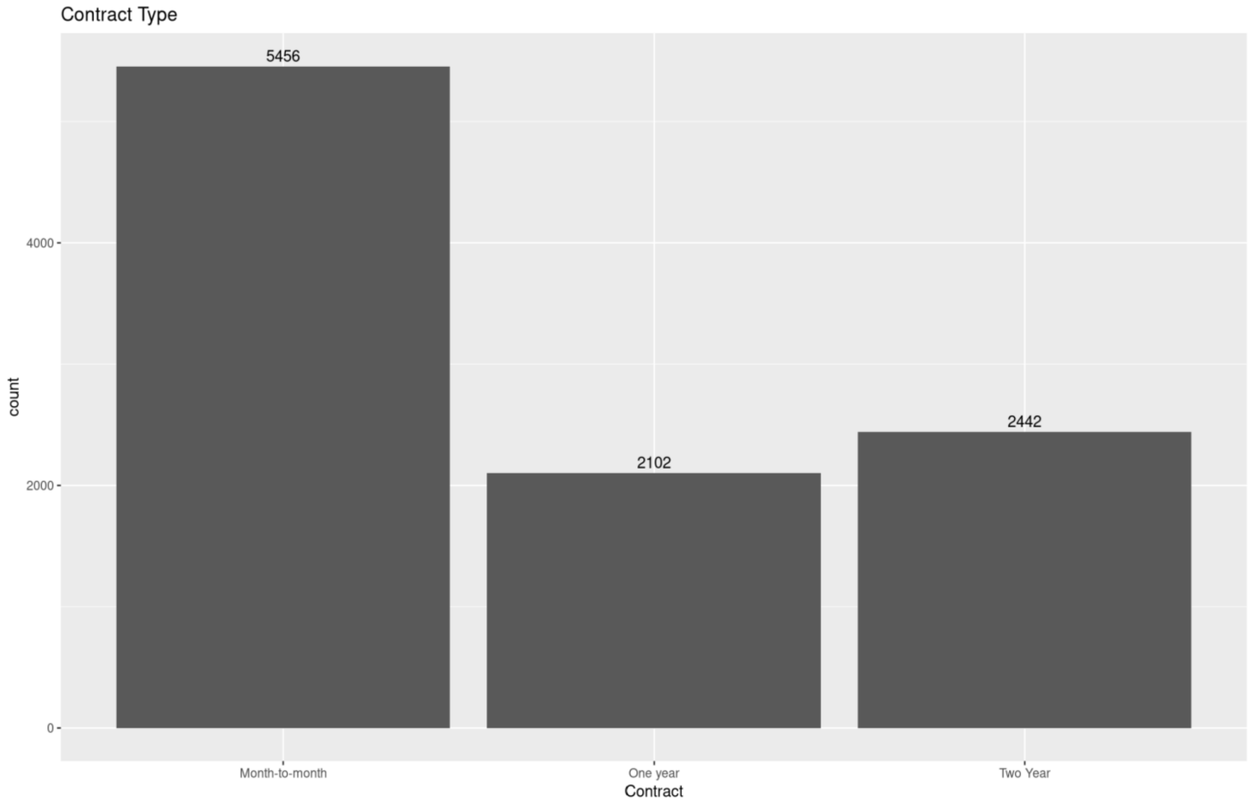
C: Univariate Statistics

The outage per week in seconds variable has a very normal distribution centered around 10 seconds per week on average. The tenure length histogram shows a bimodal distribution with peaks around 1 month and 65 months. The fact that there is such a dip at 37 months should be very concerning for the business. This means they were either picking up nearly no new customers 37 months ago, or nearly all of the ones they picked up had left already. Nearly half of all contract types are month to month, while one year and two year contracts each count for about a quarter of total contracts. Electronic check is the most common payment method, while the other three types are relatively close together.

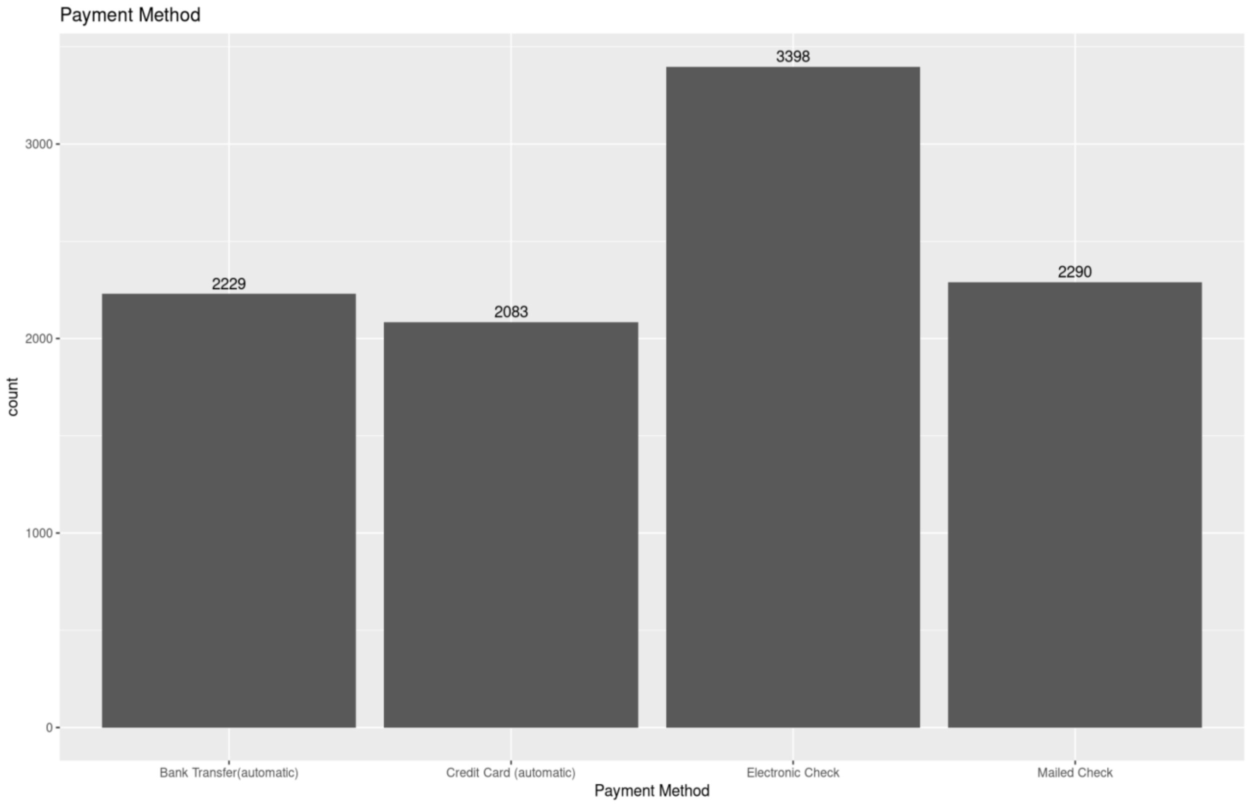
C1: Visual of Finding







Contract Type

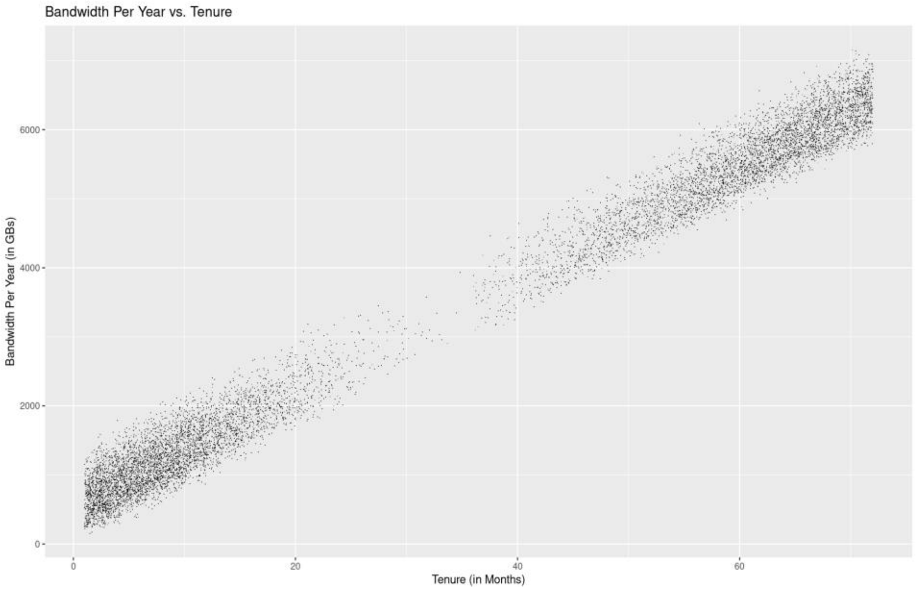


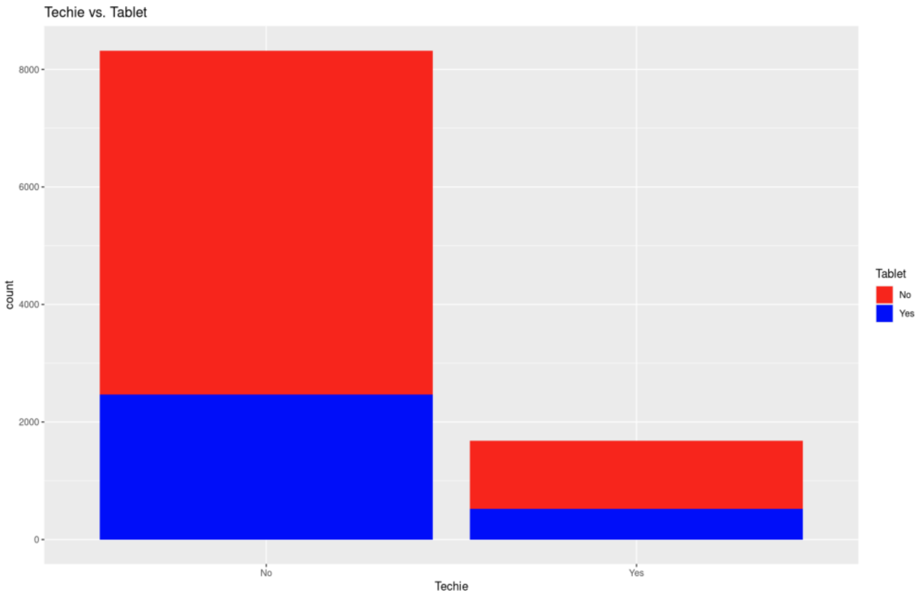
Payment Method

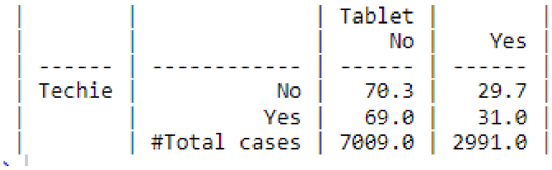
D: Bivariate Statistics

Looking at the distribution between tenure and bandwidth per year, there is a positive linear relationship. This relationship would be worth diving deeper into. I also represented the distribution between whether the customer considers themselves a techie and whether the customer has a tablet. Most customers do not consider themselves to be techies, and most customers, techie or not, do not own a tablet. I also represented this with a two way proportion table. This table show the percent of techies that have a tablet, and the percent of non-techies that have a tablet. Both rows have about a 70/30 split in favor of not having a tablet.

D1: Visual of Findings







```
|        |              | Tablet |        |
|        |              |     No |    Yes |
| ------ | ------------ | ------ | ------ |
| Techie |           No |   70.3 |   29.7 |
|        |          Yes |   69.0 |   31.0 |
|        | #Total cases | 7009.0 | 2991.0 |
```

E1: Results of Analysis

With a high p-value, we were unable to reject the null hypothesis which was that the two means of customer service contacts are equal regardless of the customer churning. Our alternative hypothesis was that the customer service contacts were higher in churned customers. We cannot say that there is a significant difference in the two means.

E2: Limitations of Analysis

One limitation of the t-test is that you can only use one factor at a time. Another is that the data should have a normal distribution. Data does not always end up in a normal distribution, so another test would have to be performed for non-normal data. The last limitation I will mention is that the data is collected from a random sample of the population. The sample we gather could be much different than the actual population.

E3: Recommended Course of Action

I would not recommend any action in terms of using customer service contacts to predict churn. There is no statistical evidence to prove that these are related. The next action I would recommend is diving deeper into the relationship between tenure length and bandwidth used. There seems to be a very positive relationship between those two variables.

F: Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=10806b46-fcec-487c-abbc-af7a01361a48

G: Sources For Third-Party Code

t.test: Student's t-Test. RDocumentation. Retrieved December 28, 2022. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

Black, Kelly. *Two Way Tables.* Cyclismo. Retrieved December 28, 2022. https://www.cyclismo.org/tutorial/R/tables.html#creating-a-table-from-data

Ggplot2 title: main, axis and legend titles. Statistical tools for high-throughput data analysis. Retrieved December 28, 2022. http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

H: Sources

What are independent and dependent variables? National Center for Education Statistics. Retrieved December 28, 2022. https://nces.ed.gov/nceskids/help/user_guide/graph/variables.asp

What statistical analysis should I use? Statistical analyses using SPSS. UCLA: Statistical Consulting Group. From https://stats.oarc.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/