D207 PA

Jon Isak

D207: Data Cleaning

William Sewell

Contents

[D207 PA 3](#_Toc150486367)

D207 PA

A1. I will look into the relationship between the average number of seconds per week of system outages in the customer’s neighborhood and churn. Outage is an interruption of the service that the customers are paying for and is likely to affect their decision to discontinue their service. Are customers who experience longer outages more likely to churn? This question will identify if customers who experience longer outages are more likely to leave the company, or if there is no difference between customers who leave and average outage time. The null and alternative hypothesis would be: H0: Fail to reject the null hypothesis. There is no significant difference in mean outage duration between churned and non-churned customers. H1: Reject the null hypothesis. There is a significant difference in mean outage duration between churned and non-churned customers.

A2. Understanding whether average outage time affects customers' decisions to leave could help the company retain customers and save money. If outage time influences a customer's decision to leave, then the company can allocate resources to upgrade its infrastructure for improved uptime. If downtime has no impact on churn, the company can save money by prioritizing other aspects that play a more substantial role in customer retention. In either scenario, the analysis will provide valuable insights for the company to make informed decisions and allocate resources effectively.

A3. The data relevant to answering the question include the categorical 'Churn' column, which consists of 'yes' or 'no', and the 'Outage sec per week' column, which consists of numerical values like '7.1'.

B1. The code for the test is atatched.

B2. The results of the t-test were: t-statistic: -0.0156, p-value: 0.9875. My alpha was .05 and since the p-value was higher than the alpha, I decided to fail to reject the null hypothesis because there is no significant difference in average outage duration between churned and non-churned customers (Editor, 2015).

B3. I picked the t-test technique for my analysis. This technique is the best option because I want to compare the means of outage for two independent groups, customers who churned or not. In addition, the data is normally distributed and variance is equal for both groups. The Shapiro-Wilk and Levene's Test showed a p-value higher than .05 indicating a normal distribution. And equal variance.

C. I will identify the distribution of the continuous variables (Outage\_sec\_perweek, tenure) and categorical variables (Churn, Techie). The distribution of 'Outage\_sec\_perweek' is illustrated through a histogram, revealing a bell curve shape indicative of a normal distribution. Additionally, a boxplot for 'Outage\_sec\_perweek' highlights the presence of outliers within the data. Moving on to 'Tenure,' its distribution is visualized using a histogram, showcasing a bimodal distribution, suggesting the presence of two modes within the data. For the categorical variables, 'Churn' and 'Techie,' countplots have been employed to understand the frequency distribution. The countplot for 'Churn' indicates that the majority answer was 'no,' suggesting a low churn rate. Similarly, the countplot for 'Techie' reveals that the majority answered 'no' as well, indicating that a significant portion of the respondents does not identify as tech-savvy. The visuals are included with the code.

D. I will identify the distribution of the continuous variables (Bandwidth\_GB\_Year, MonthlyCharge) and categorical variables (PaymentMethod, PaperlessBilling). For the continuous variables, 'Bandwidth\_GB\_Year' and 'MonthlyCharge,' a scatter plot has been generated to visualize their relationship. The scatter plot reveals two distinct clusters, suggesting potential subgroups within the data. Additionally, a correlation matrix and heatmap for these continuous variables indicate a strong correlation between bandwidth usage and monthly charges. This implies that as bandwidth usage increases, monthly charges tend to rise as well. For the categorical variables, 'PaymentMethod' and 'PaperlessBilling' have been explored using a combination plot and statistical testing. The combination plot illustrates the distribution of paperless billing across different payment methods. Furthermore, it shows that paperless billing is utilized across all payment methods and is the majority form of billing for respondents. To quantify this relationship, a chi-square test was performed on the contingency table of 'PaymentMethod' and 'PaperlessBilling.' The test results reveal a p-value higher than 0.05, indicating a lack of statistically significant correlation between the choice of payment method and the preference for paperless billing. The visuals are included with the code.

E1. The t-test yielded a t-statistic of -0.0156 and a p-value of 0.9875. Given my significance level (alpha) of 0.05, and since the p-value surpasses alpha, I opt not to reject the null hypothesis. Therefore, I conclude that there is no significant difference in the outage duration between churned and non-churned customers.

E2. There may be limitations in the data analysis. The t-test was chosen under the assumption that the data follows a normal distribution, and that variances are equal. The results may be affected if these assumptions are incorrect. The outage data is specific to a certain time frame and does not necessarily reflect future outage patterns. The analysis also primarily focuses on outage duration as the main factor in churn, without taking into consideration other relevant variables, such as reliability, that may influence churn.

E3. The analysis found that there is no statistically significant difference in the duration of outages between customers who churned and those who did not. This suggests that outage duration may not be the only factor that drives customer churn. As a result, the company should allocate resources to discovering other variables that may affect churn and not spend it on upgrading their infrastructure to improve uptime, as the data shows it won't affect churn and bring value to stakeholders.

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c079f63a-56a7-43a3-9f1e-b0b601404be7

G.

matplotlib. (n.d.). *Matplotlib.pyplot.subplot*. Retrieved November 10, 2023, from https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.subplot.html

Pydata. (n.d.). *Pandas.crosstab*. Retrieved November 10, 2023, from https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html

Scipy. (n.d.). *Scipy.stats.chi2\_contingency*. Retrieved November 10, 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2\_contingency.html

Scipy. (n.d.). *Scipy.stats.levene*. Retrieved November 10, 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html

Scipy. (n.d.). *Scipy.stats.shapiro*. Retrieved November 10, 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

Scipy. (n.d.). *Scipy.stats.ttest\_ind*. Retrieved November 10, 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\_ind.html

Yi, M. (n.d.). *A Complete Guide to Histograms*. Chartio. Retrieved November 10, 2023, from https://chartio.com/learn/charts/histogram-complete-guide/

H.

Editor, M. B. (2015, March 19). *Understanding Hypothesis Tests: Significance Levels (Alpha) and P values in Statistics*. https://blog.minitab.com/en/adventures-in-statistics-2/understanding-hypothesis-tests-significance-levels-alpha-and-p-values-in-statistics