# Exploratory Data Analysis Performance Assessment

John Foster

Department of Information Technology, Western Governors University

D207: Exploratory Data Analysis

Professor William Sewell

April 13, 2023

# Exploratory Data Analysis Performance Assessment

## A1: Question for Analysis

The research question that I chose to explore in this analysis is as follows: Is there a statistical relationship between the average weekly number of seconds that a given customer's neighborhood experiences system outages and whether the customer discontinued service in the last month? If our research question seeks to determine the presence of a statistical relationship between *Outage\_sec\_perweek* and *Churn,* then we can identify a clear hypothesis for testing*.* Our hypothesis is that there is a statistically significant difference between the distributions of *Outage\_sec\_perweek* in churned versus non-churned customers. However, our actual testing is better performed on the inverse of our hypothesis than on the hypothesis itself. We call this second hypothesis a null hypothesis, and in this case, it would be expressed as an absence of statistically significant differences between the distributions of *Outage\_sec\_perweek* in churned versus non-churned customers. Using exploratory data analysis methods such as T-Tests, chi-square testing, and ANOVA testing, we can test our null hypothesis, and by extension, our main hypothesis, thereby providing an answer to our research question.

## A2: Benefit from Analysis

The insight generated through an analysis of this research question could be of significant value to stakeholders of this organization. The question is a simple exploration of how local outages might be related to service cancellation. If there is a relationship between these two variables, this might inform a business decision of whether to invest in additional infrastructure or personnel to retain more customers through operational initiatives to minimize outage times. On the other hand, if there is no relationship, it might indicate an opportunity to audit budgets for departments associated with minimizing those times, which might result in reduced operational overhead.

## A3: Data Identification

The process of answering our research question requires only two variables from the entire *churn\_clean* dataset. These two variables are the *Outage\_sec\_perweek* and *Churn* columns. The *Outage\_sec\_perweek* column has a data type of 64-bit floating point numbers and is a direct measure of elapsed outage time. The *Churn* column is an object data type, essentially functioning as a string-based categorical variable, but since the only unique values are “Yes” and “No,” this feature could also be conceptualized as a boolean variable. This variable is an indicator of whether or not a customer has discontinued service in the past month.

## B1: Code

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from scipy.stats import ttest\_ind

df = pd.read\_csv('churn\_clean.csv')

#perform a t-test using scipy.stats.ttest\_ind on the 'Tenure' and 'Churn' columns

x = ttest\_ind(df[df['Churn'] == 'Yes']['Outage\_sec\_perweek'], df[df['Churn'] == 'No']['Outage\_sec\_perweek'])

#Return our our T-test results.

#The first value is the difference between means of the two groups (mean(a) - mean(b)) divided by the standard error.

#The second value is the p-value

print(x)

if x[1] > .05:

    print('P value is above .05, so we are unable reject the null hypothesis')

else:

    print('P value is below .05, so we can reject the null hypothesis')

#Look at the distribution of 'Outage\_sec\_perweek' column for churned and non-churned customers

df[df['Churn'] == 'Yes']['Outage\_sec\_perweek'].plot(kind='hist', rot=0, bins=30, title='Outage\_sec\_perweek for Churned Customers')

plt.show()

df[df['Churn'] == 'No']['Outage\_sec\_perweek'].plot(kind='hist', rot=0, bins=30, title='Outage\_sec\_perweek for Non-Churned Customers')

plt.show()

#Get the standard deviation of the 'Outage\_sec\_perweek' column for churned and non-churned customers

print(df[df['Churn'] == 'Yes']['Outage\_sec\_perweek'].std())

print(df[df['Churn'] == 'No']['Outage\_sec\_perweek'].std())

## B2: Output

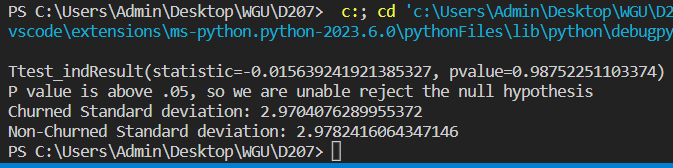
PS C:\Users\Admin\Desktop\WGU\D207> c:; cd 'c:\Users\Admin\Desktop\WGU\D207'; & 'C:\Users\Admin\AppData\Local\Programs\Python\Python310\python.exe' 'c:\Users\Admin\.vscode\extensions\ms-python.python-2023.6.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '59766' '--' 'C:\Users\Admin\Desktop\WGU\D207\scriptnew.py'

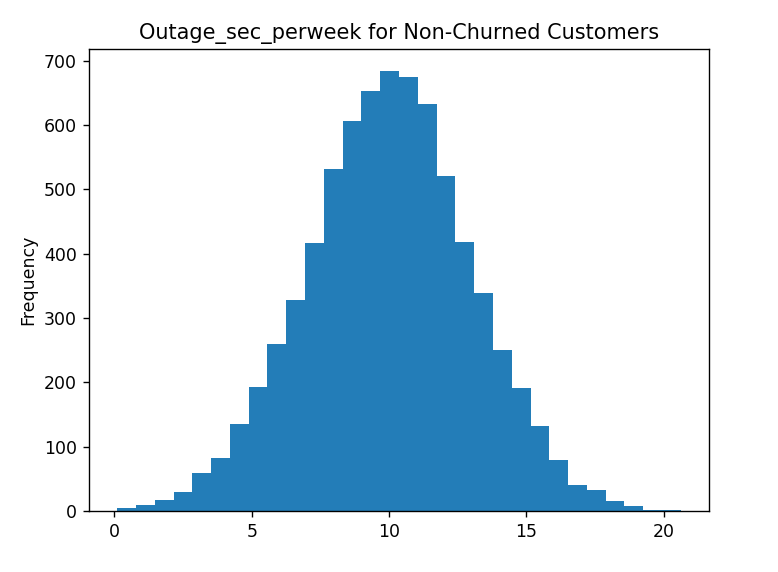
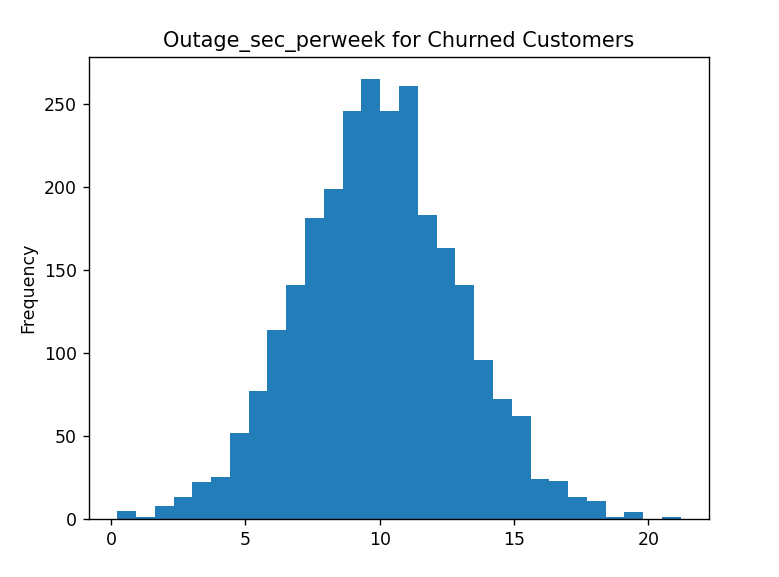
Ttest\_indResult(statistic=-0.015639241921385327, pvalue=0.98752251103374)

P value is above .05, so we are unable reject the null hypothesis

Churned Standard deviation: 2.9704076289955372

Non-Churned Standard deviation: 2.9782416064347146





## B3: Justification

Performing exploratory and graphical data analysis can be a very effective method for uncovering relationships between variables when it is unclear at the outset exactly what an evaluation will uncover (Larose & Larose, 2019). For our analysis, we have used the T-Test method to determine whether a correlation exists between the *Outage\_sec\_perweek* and *Churn* variables. Based on the fact that we are analyzing a numeric variable with a categorical variable, this narrows our potential methods of analysis to T-Test and ANOVA. I have chosen the T-Test method over the ANOVA method because our categorical variable, the *Churn* column, only has two possible values (“Yes” and “No”), making it the more appropriate choice.

## C: Univariate Statistics

The continuous variables that we have selected for univariate statistical analysis are the *Age* and *Children* columns. To examine their distributions, we will be utilizing box plots. After examining the box plots (contained in section C1), we have identified that the *Age* column is most appropriately described as a uniform distribution, and the *Children* column has a positively skewed distribution. The *Age* column had to be further explored via histogram to determine if it was a normal or uniform distribution, as this differentiation is difficult to visually surmise from a box plot alone. The categorical variables that I have chosen are the *InternetService* and *Marital* columns. We have chosen to use bar charts to examine the distributions of these variables. However, this format creates a few challenges when attempting to describe a distribution of the different categories. The *InternetService* column can at least be ordered in such a way as to rank the quality of internet service (*None, DSL, Fiber Optic*), but the *Marital* column is not a direct measure of a feature, because it contains five categories that cannot be logically ordered. However, if we can accept that the categories of this feature are unordered, we can disregard the ordering that occurs through our visualization method and see that all five categories are around the 2000 count mark in our histogram, meaning they each represent roughly one-fifth of the total dataset size. With this in mind, we can describe the *Marital* column as having a uniform distribution. The *InternetService* column, on the other hand, appears to have a negative distribution when appropriately ordered. The code for our univariate analyses can be found below.

#Box plots of the 'Age' and 'Children' columns

df['Age'].plot(kind='box', rot=0)

plt.show()

df['Children'].plot(kind='box', rot=0)

plt.show()

#Histogram of the 'Age' column

df['Age'].plot(kind='hist', rot=0, bins=20, title='Age')

plt.show()

#Bar charts of the InternetService and Marital columns

df['InternetService'].value\_counts().plot(kind='bar', rot=0, title='Internet Service').invert\_xaxis()

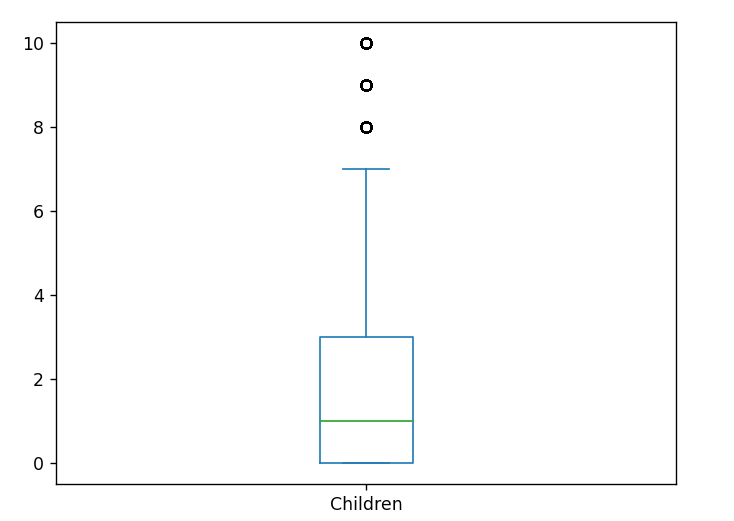
plt.show()

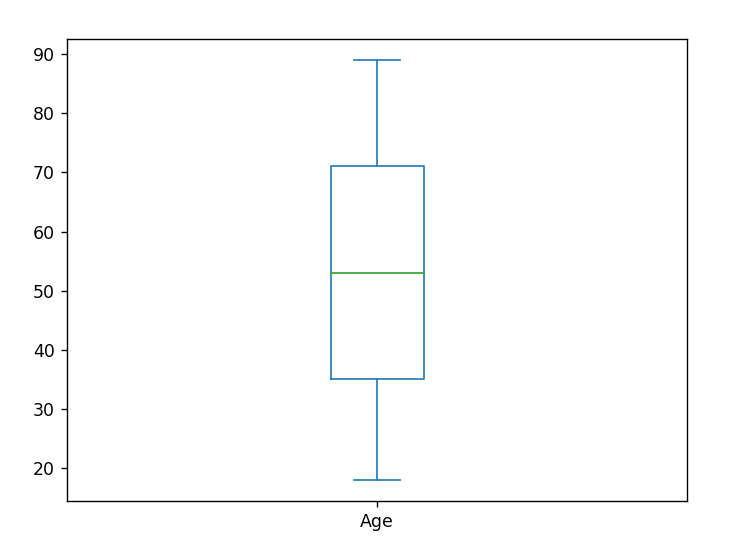
df['Marital'].value\_counts().plot(kind='bar', rot=0, title='Marital Status')

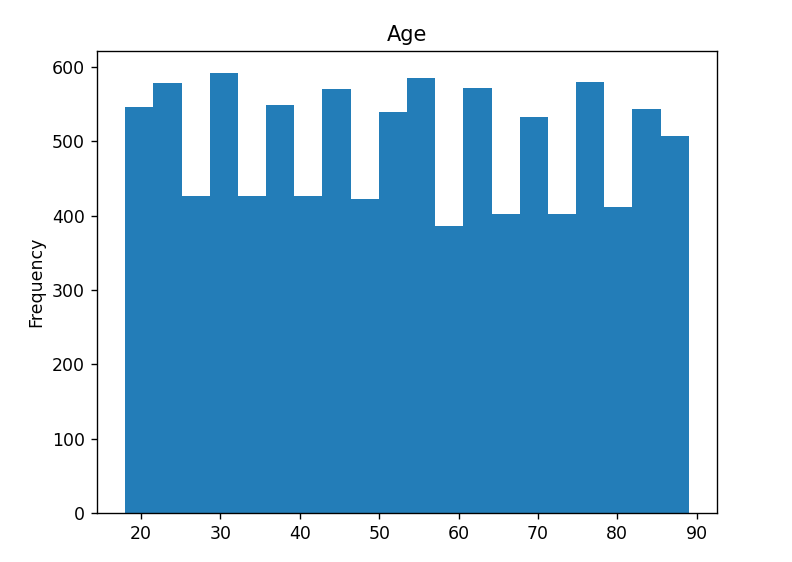
plt.show()

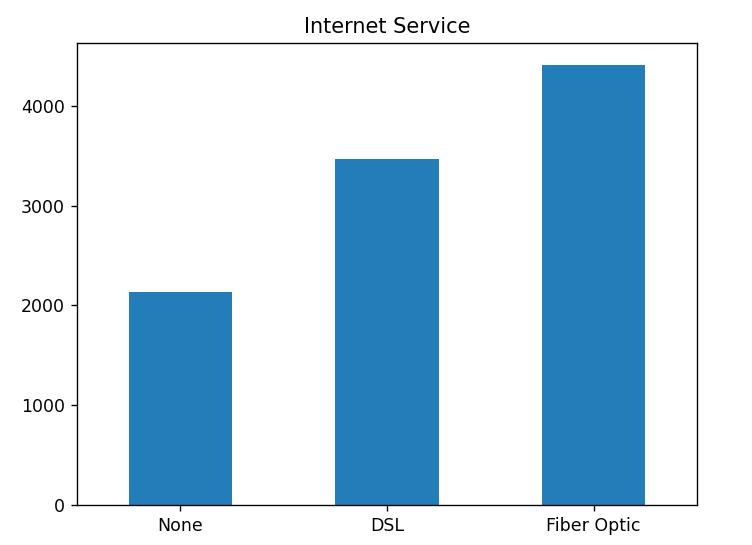
Code paraphrased from: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html>

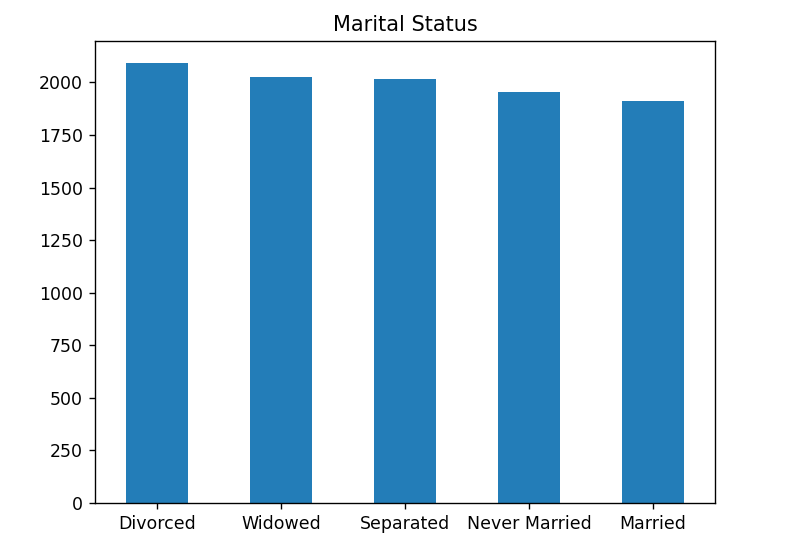
## C1: Visual of Findings











## D: Bivariate Statistics

We have chosen to create a scatter plot to evaluate the bivariate statistical relationship between two continuous variables in our dataset: *Age* and *Children.* As we can see from the plot in section D1, there does not appear to be any directed grouping in the distribution of data points between these two variables. Each additional single child along the y-axis shows an even line of parental ages, fading as we increase the number of children through the application of the *alpha* argument in our plot function. Based on this, we can conclude that there is no apparent statistical relationship between these two variables using our chosen method of bivariate statistical analysis.

We have chosen to create stacked bar charts in order to explore our two categorical variables: *Gender* and *Techie*. However, as we can see from the stacked bar chart comprised of the value counts across these respective variables, it is difficult to visually recognize meaningful relationships between these different variables if they are not normalized (Larose & Larose, p. 49). Because of this, we have also normalized our bar chart to compare the ratios of both *Techie* variables across each gender. Doing this reveals that the ratio appears fairly similar across each gender, despite the differing total counts of each gender in the dataset. Therefore, we can conclude that there is no apparent statistical relationship between these two variables using our chosen method of bivariate statistical analysis. The code for this section of the analysis is as follows:

#make a correlation plot out of the 'Age' and 'Children' columns

df[['Age', 'Children']].plot(kind='scatter', x='Age', y='Children', alpha=.01)

plt.show()

#Make a stacked bar chart out of the Gender and Techie columns

crosstab\_01 = pd.crosstab(df['Gender'], df['Techie']).sort\_index(axis=1, ascending=False)

crosstab\_01.plot(kind='bar', stacked = True, rot=0)

crosstab\_norm = crosstab\_01.div(crosstab\_01.sum(1), axis = 0)

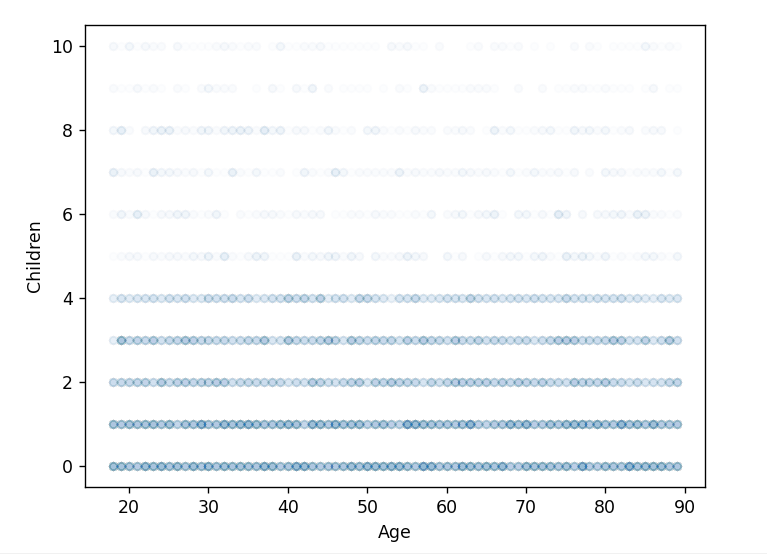
plt.show()

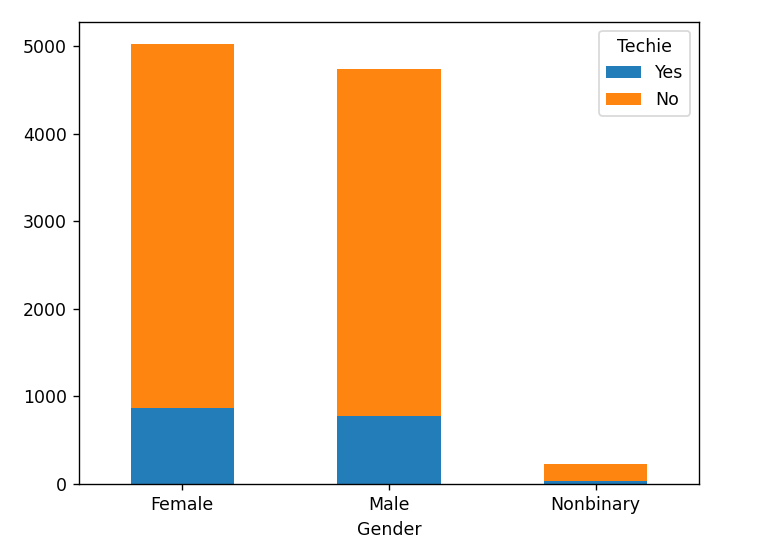
crosstab\_norm.plot(kind='bar', stacked = True, rot=0, title="Normalized Ratios of Techie vs. Non-Techie by Gender")

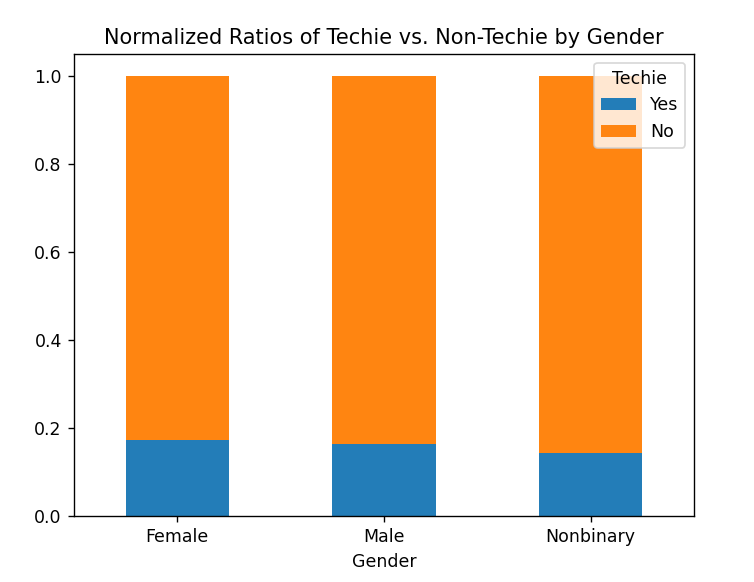
plt.show()

Code paraphrased from: Larose & Larose, 2019, p. 49-51.

## D1: Visual of Findings







## E1: Results of Analysis

The results of our hypothesis testing were fairly clear. The T-Test returned a p-value of over .987. To reject our null hypothesis and thereby support our primary hypothesis of a statistically significant relationship existing between the *Outage\_sec\_perweek* and *Churn* variables, we would look for a p-value below .05. To put these results in perspective, the “p” in p value stands for the word “probability.” What it indicates when applied to null hypothesis testing is the probability of the differences between the means of our tested groups occurring if we expect them to be identical. As we saw from our code, the two groups we refer to are the *Outage\_sec\_perweek* values among churned versus non-churned customers. The greater the differences in mean value between these groups, the less likely the null hypothesis is to be true. This is why a p-value of less than .05 is generally needed to reject a null hypothesis; it means that there is less than a 5 percent chance of the observed differences in means occurring if there is no statistical relationship between the variables. Instead, our distributions were so similar that we calculated a probability of over 98 percent that we would observe the difference that appeared, assuming our null hypothesis is true. This is a conclusive indication that our null hypothesis cannot be rejected, and that there is no statistical evidence to support the assumption that these variables are correlated.

## E2: Limitations of Analysis

Despite our interpretation of the results, any analysis that utilizes a T-Test comes with several limitations. First, the data evaluated must be measurable on a continuous or ordinal scale. Our continuous variable in this case is *Outage\_sec\_perweek,* and both groups being evaluated share this continuous measure when evaluating means and distributions*.* Second, the data evaluated must be a simple random sample, meaning it has to be directly representative of the total population of data, and sampled randomly. As we will discuss shortly, this may be a factor that affects the validity of our analysis. Third, our compared groups should both display a normal distribution, which allows us to perform our analysis with a p-score cutoff of .05. When we visualize our groups with histograms, we can see that both do appear to display normal distributions. Fourth, a T-Test is appropriately utilized in an analysis of a large sample size. Our sample size is 10,000, so this facet also appears to support our choice of evaluation method. Fifth, the groups compared using a T-Test should have approximately equal standard deviations. As we can see from our code output, our standard deviations for these groups were 2.970 and 2.978; a difference of only eight thousandths. Finally, a T-Test should only be employed to evaluate two distributions, and in our case, the *Churn* variable does only contain two unique categorical values (Sewell, 2021).

As these limitations apply to our own research question, we do have some difficulty establishing conclusively that our T-Test is being performed on a simple random sample. Because our *Churn* variable represents whether a customer has discontinued service in the past month, we run into issues of how to determine whether or not the churn rate this month is commensurate with those of other months. This means that while we can say our results do not indicate a statistical relationship between our selected variables, we are unable to extend that observation beyond the current month without corresponding datasets to evaluate from other months.

In addition, we can identify another unique limitation of our analysis. If we observe the range of values in our histograms, we can see that there are no data points for *Outage\_sec\_perweek* in either group that reach 25. This is an incredibly low range, and means that any correlation, or lack thereof, observed through this analysis may be subject to change if the range of our data were to increase past a certain point.

Because of these limitations, all recommendations made based on our analysis must be understood in context, or more appropriately, subjected to further evaluation before any business decisions are predicated upon them.

## E3: Recommended Course of Action

Based on the results of our analysis and its apparent limitations, we can both respond to our research question and make recommendations for potential actions to take. In response to our research question, we can state that there does not appear to be a statistical relationship between the average weekly number of seconds that a given customer's neighborhood experiences system outages and whether the customer discontinued service in the last month. Because of this, I have two recommendations. The first would be to perform identical analyses on datasets representing other months in order to establish the validity of this month's dataset. The second is to begin exploring avenues to audit and reduce the budgets of departments and operations that directly contribute to the minimization of outages. However, this recommendation must also be presented with the disclaimer that this analysis does not indicate that customer retention will never be affected by outage frequency and duration no matter how high it is, only that it does not appear to in the context of our current performance in this metric, and that this may indicate an opportunity to relax those performance standards until some effect is actually observable.

## F: Video

A brief demonstration video of the Python code, including a description of the software environment used, has been recorded via Panopto and submitted for evaluation in conjunction with this word document in the WGU submission portal.

## G: Sources for Third-Party Code

Larose, C. D, & Larose, D. T. (2019). Data Science Using Python and R. ISBN-13: 978-1-119-52684-1.

<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html>

# H: Sources

Larose, C. D, & Larose, D. T. (2019). Data Science Using Python and R. ISBN-13: 978-1-119-52684-1.

Sewell, W. (2021). D207 Episode 1. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3980b7aa-6bdb-4b38-8356-ae3100e4ff55>