D207: Exploratory Data Analysis

A1. Question for Analysis

Are the number of seconds per week of system outages in the customer's neighborhood (outage_sec_perweek) significant to churn?

A2. Benefits from Data Analysis

Performing data analysis allows the telecommunications company and its stakeholders to have a better idea of what is affecting churn. The reason I would like to look into the specific variable of outages is because I have personally swapped from one telecommunications company to another because of the outages I would experience in my area. It would not only happen often, but the outages would occur for long periods of time. Then, I would communicate to my neighbors how much better I liked the telecommunications company I switched to, and the neighbors would switch over too. Perhaps the outages could have an impact on customers discontinuing their service. If this is the case, the company can focus on the customers who experience the outages and look into why there is outages occurring for the longer amounts of time.

A3. Identify the Data

The relevant data needed to answer my research question as listed in part A1 are the "Outage_sec_perweek" and "Churn" variables from the churn_clean.csv file. The "Outage_sec_perweek" variable is the number of seconds per week of system outages in the customer's neighborhood which has continuous numeric data. An example would be 7.978322947 from row 1. The "Churn" variable is the dependent variable and binary categorical because it has two options, "Yes" and "No".

- H_0 = Churn and the number of outages are independent (have no relationship) [This is the null hypothesis].

- H_1 = Churn and the number of outages are dependent (have a relationship) [This is the alternative hypothesis].

I will set an alpha of 0.05.

B1 & B2. Code & its Output

In [1]:
```
# Import the necessary packages & libraries
import pandas as pd
import numpy as np
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt
```

In [2]:
```
# Load the data set into the pandas data frame by using read_csv command
df = pd.read_csv(r'C:\Users\ashle\Downloads\D207 Churn Dataset\churn_clean.csv',
                 usecols=['Churn',
                          'Outage_sec_perweek'])
```

In [3]:
```
# Run a two-sample t-test by creating groups of customer's churn with the number of seconds per week of system outages in the customer's neighborhood
churn_yes = df[df['Churn'] == 'Yes'].Outage_sec_perweek
churn_no = df[df['Churn'] == 'No'].Outage_sec_perweek
```

In [4]:
```
# Print t-test results
ttest_result = stats.ttest_ind(churn_yes,
                               churn_no)
print(ttest_result)
```

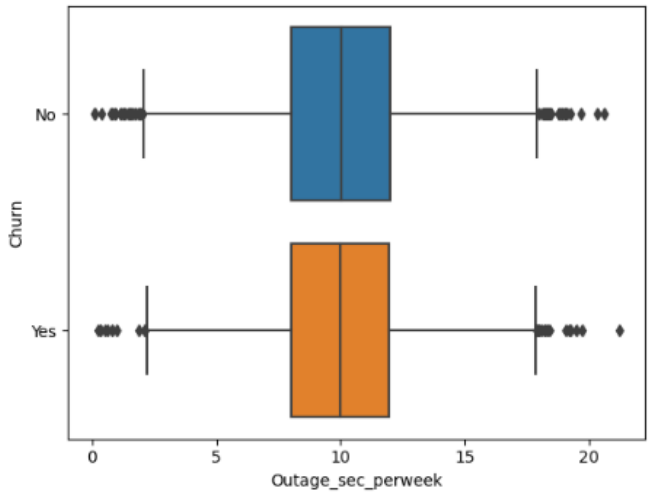
```
TtestResult(statistic=-0.015639241921385327, pvalue=0.98752251103374, df=9998.0)
```

In [5]:
```
# Determine whether to accept or reject the null hypothesis using alpha = 0.05
alpha = 0.05
if (ttest_result[1] < alpha):
    print(f'This is the t-test p-value: {str(ttest_result[1])} with an alpha of {str(alpha)}. \nP-value falls within 0.05 alpha: \nReject the null hypothesis.')
else:
    print(f'This is the t-test p-value: {str(ttest_result[1])} with an alpha of {str(alpha)}. \nP-value does not fall within 0.05 alpha: \nAccept the null hypothesis.')
```

```
This is the t-test p-value: 0.98752251103374 with an alpha of 0.05.
P-value does not fall within 0.05 alpha:
Accept the null hypothesis.
```

In [6]:
```
# Generate a boxplot of the relationship between churn and the number of seconds per week of system outages in the customer's neighborhood
sns.boxplot(y='Churn',
            x='Outage_sec_perweek',
            data=df)
```

Out[6]:
```
<Axes: xlabel='Outage_sec_perweek', ylabel='Churn'>
```



B3. Justification of Analysis Technique

T-tests are able to measure 1 or 2 groups and can test 1 numeric and 1 categorical variable. As mentioned in part A3, these are the exact data types I will be using in this analysis. The "Churn" variable is binary and categorical because it has two options, "Yes" and "No" and the "Outage_sec_perweek" has continuous numeric data since it consists of numbers in its data. Therefore, t-tests would be a beneficial analysis technique to compare the two churn groups to the numeric data to determine if there is a relationship.

C1. Univariate Statistics

For univariate statistics, I will be looking at one variable at a time without any comparisons and identifying the distribution of two continuous variables and two categorical variables. I will be creating histograms of each variable and exploring the distribution of the following two continuous variables: The outages, in seconds, per week in the customer's neighborhood (Outage_sec_perweek) and the amount the customer was charged monthly (MonthlyCharge). I will also be exploring the binary distribution of the following two categorical variables: Whether or not (yes/no) the customer cancelled their service within the past month (Churn) and whether or not (yes/no) the customer is tech-savvy according to a provided questionnaire (Techie).

Two continuous variables:

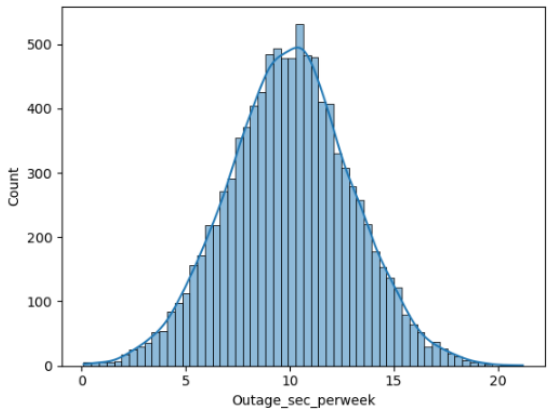
- Outage_sec_perweek – The distribution of the graph is normal. This is because the graph has a "bell-shape" and looks symmetrical. I used the describe() command to provide me with the count, mean, standard deviation, minimum, maximum, and each of the percentiles of the data. The mean is 10.001848. The standard deviation is 2.976019. The minimum plot is 0.099747 and the maximum plot is 21.207230.
- MonthlyCharge - The distribution of the graph is right-skewed. This is because the distribution is longer on the right side of its peak than on its left. I used the describe() command to provide me with the count, mean, standard deviation, minimum, maximum, and each of the percentiles of the data. The mean is 172.624816. The standard deviation is 42.943094. The minimum plot is 79.978860 and the maximum plot is 290.160419.

This is shown in the screenshots below. Two categorical variables:

```
In [7]: # Create a dataframe for the univariate statistics
df_univ = pd.read_csv(r'C:\Users\ashle\Downloads\D207 Churn Dataset\churn_clean.csv',
                      usecols=['Outage_sec_perweek',
                               'MonthlyCharge',
                               'Churn',
                               'Techie'])
```

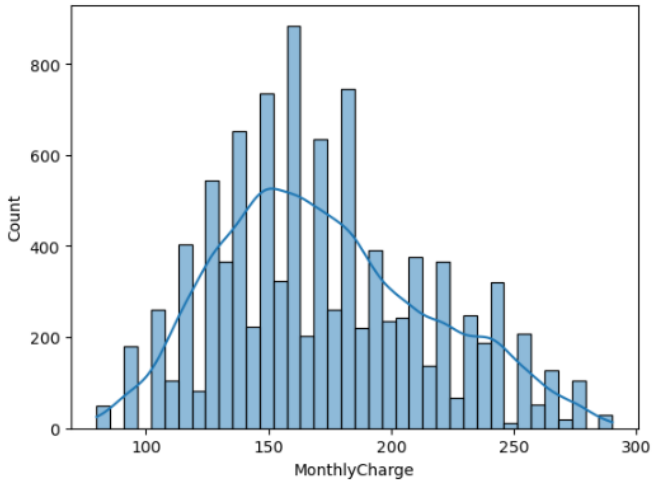
```
In [8]: # Create histograms of the continuous variables
sns.histplot(x='Outage_sec_perweek',
             data=df_univ,
             kde=True)
df_univ.Outage_sec_perweek.describe()
```

```
Out[8]: count    10000.000000
mean        10.001848
std          2.976019
min          0.099747
25%          8.018214
50%         10.018560
75%         11.969485
max         21.207230
Name: Outage_sec_perweek, dtype: float64
```




```
In [9]: sns.histplot(x='MonthlyCharge',
             data=df_univ,
             kde=True)
df_univ.MonthlyCharge.describe()
```

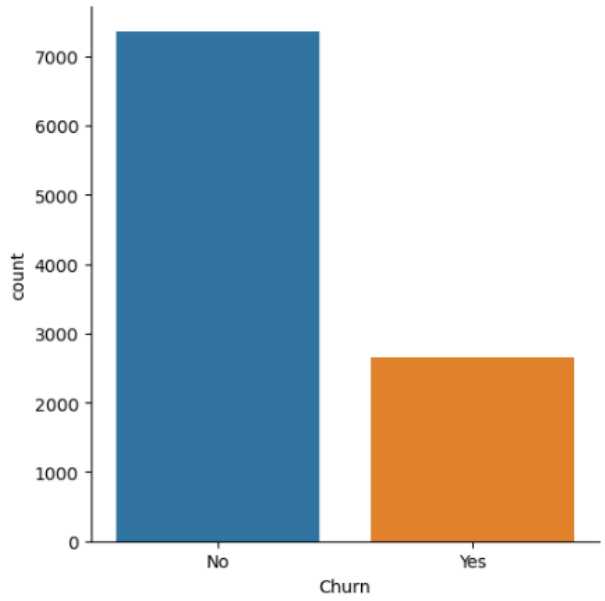
```
Out[9]: count    10000.000000
mean       172.624816
std         42.943094
min         79.978860
25%        139.979239
50%        167.484700
75%        200.734725
max        290.160419
Name: MonthlyCharge, dtype: float64
```



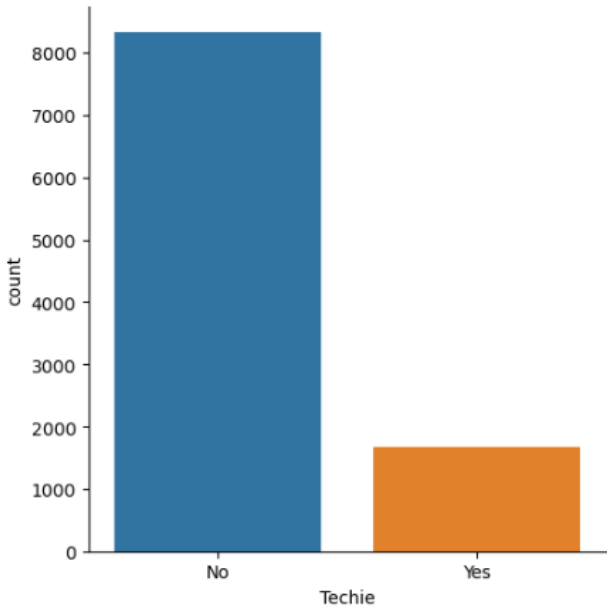
- Churn – The distribution of the graph is Bernoulli. This is because there are only two options (binary) that can only be answered with yes and no. Based on the graph, there were less customers that cancelled their services within the month than customers that kept their services. I used the value_counts() command to further prove my findings. According to value_counts(), there were 7350 people who did not churn and 2650 people who did churn.
- Techie – The distribution of the graph is Bernoulli. This is because there are only two options (binary) that can only be answered with yes and no. Based on the graph, there were more customers that were not tech-savvy than customers that are good with technology. I used the value_counts() command to further prove my findings. According to value_counts(), there were 8321 people who did not consider themselves to be techies and 1679 people who were considered techies.

The screenshots of the graphs are shown below.

```
In [10]: # Create a distribution plot of the categorical variables
         sns.catplot(x='Churn',
                     data=df_univ,
                     kind='count')
         df_univ.Churn.value_counts()
Out[10]: Churn
         No     7350
         Yes    2650
         Name: count, dtype: int64
```




```
In [11]: sns.catplot(x='Techie',
                     data=df_univ,
                     kind='count')
         df_univ.Techie.value_counts()
Out[11]: Techie
         No     8321
         Yes    1679
         Name: count, dtype: int64
```



D1. Bivariate Statistics

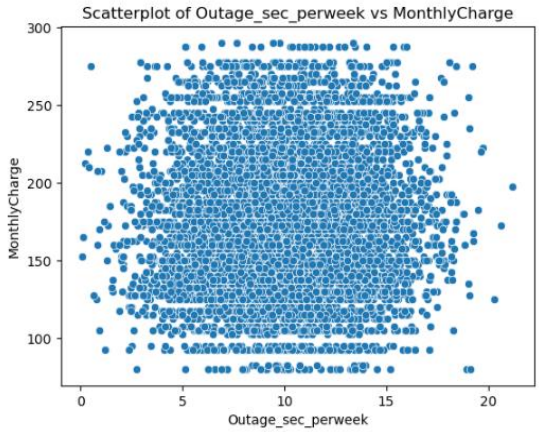
I will be using the same continuous variables (Outage_sec_perweek and MonthlyCharge) and categorical variables (Churn and Techie) as I did for univariate statistics. However, for bivariate statistics, I will be comparing the two variables instead of just summarizing one single variable at a time like I did for the univariate statistics in part C1.

- Continuous variables: Based on the scatterplot, it seems that there is no correlation between MonthlyCharges and the Outage_sec_perweek since there isn't a noticeable trend of the plots. However, there is some variability in the monthly charge for any given length of outages (in seconds). For example, the scatterplot shows very few plots between 0 to 5 seconds of outages and 15 to 20 seconds of outages in comparison to plots 5 to 15. There are many plots between 5 to 15 in which the monthly charge has a large range from $100 per month to about $275 per month. This further proves my point that the two variables have no correlation to each other. Refer to the describe() command to retrieve any necessary information in Part C1.

```
In [12]: # Create a dataframe for the bivariate statistics
df_biv = pd.read_csv(r'C:\Users\ashle\Downloads\D207 Churn Dataset\churn_clean.csv',
                     usecols=['Outage_sec_perweek',
                              'MonthlyCharge',
                              'Churn',
                              'Techie'])
```

```
In [13]: # Create a scatterplot of continuous variables (Outages_sec_perweek and MonthlyCharge)
sns.scatterplot(x='Outage_sec_perweek',
                y='MonthlyCharge',
                data=df_biv)
plt.title('Scatterplot of Outage_sec_perweek vs MonthlyCharge')
```

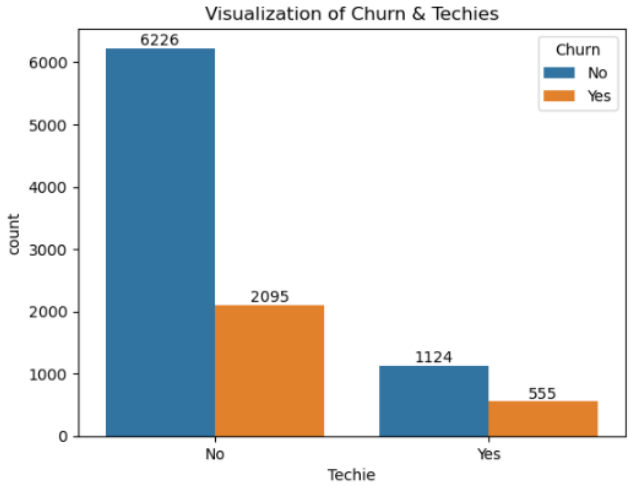
```
Out[13]: Text(0.5, 1.0, 'Scatterplot of Outage_sec_perweek vs MonthlyCharge')
```



- Categorical variables: I created a count plot to compare the customer churn versus customers who consider themselves to be a techie or not. According to the graph, the majority of customers did not churn (kept their services) and were not techies. After that, I did the value_counts() command to provide the count of each bar. For example, the tallest bar represents 6226 customers. As already mentioned, these 6226 customers did not churn and were not techies. The smallest bar represents 555 techies who did churn. There does not seem to be a correlation between these two variables. However, it is good that there are so many people who do not feel tech-savvy that kept their telecommunication services.

```
In [14]: # Create a countplot of categorical variables (Churn and Techie)
ax = sns.countplot(x='Techie', hue= 'Churn', data= df_biv)
for container in ax.containers:
    ax.bar_label(container)
plt.title('Visualization of Churn & Techies')
df_biv.value_counts(["Techie", "Churn"])
```

```
Out[14]: Techie  Churn
No      No       6226
        Yes      2095
Yes     No       1124
        Yes       555
Name: count, dtype: int64
```



E1. Hypothesis Test Results

- H_0 = Churn and the number of outages are independent (have no relationship) [This is the null hypothesis].
- H_1 = Churn and the number of outages are dependent (have a relationship) [This is the alternative hypothesis].

The question for analysis is "are the number of seconds per week of system outages in the customer's neighborhood (outage_sec_perweek) significant to churn?" As mentioned above, I used an alpha of 0.05 when testing the null hypothesis. The analysis shows that the p-value ~ 0.99. The p-value is much greater than the alpha. This means we would accept the null hypothesis because there is not enough data to find a statistical relationship between churn and the number of outages. However, I created a visualization of a boxplot, and they look fairly similar. This may be worth looking into for further analysis.

E2. Limitations of the Analysis

T-tests assume that the data has normal distribution and is limited to a maximum of two groups. Otherwise, you would have to use a z-test for larger sample sizes. The test results of the analysis are also very vague. With such a large p-value of 0.99, this just means we don't have enough data against the null, resulting in accepting the null or in other words, failing to reject the null hypothesis.

E3. Recommendations Based on the Results

As mentioned in part E2, the null hypothesis was accepted. Consequently, my only recommendation is to perhaps gather more data and information from even more than the 10,000 customers provided because this may give us a better analysis since there is not enough data or evidence to reject the null hypothesis of there being a relationship between customer churn and the number of outages.

F. Panopto Video

G. Web Sources

[Global Health with Greg Martin]. (2019, June 10). *Statistics made easy ! ! ! Learn about the t-test, the chi square test, the p value and more* [Video]. Youtube.com. https://www.youtube.com/watch?v=I10q6fjPxJ0

University of South Hampton (n.d.). Practical Applications of Statistics in the Social Sciences. Southampton.ac.uk. Retrieved May 21, 2024, from https://www.southampton.ac.uk/passs/gcse_scores/bivariate_analysis/t_test.page#:~:text=In%20a%20t%20test%2C%20like%20in,no%20difference%20in%20the%20population

(2024, February 11). Univariate, Bivariate and Multivariate data and its analysis. Geeksforgeeks.org. Retrieved May 21, 2024, from https://www.geeksforgeeks.org/univariate-bivariate-and-multivariate-data-and-its-analysis/

H. Sources

N/A

I. Professional Communication

Demonstrate professional communication in the content and presentation of your submission.