D206 Performance Assessment

Darian Gurrola

Course Instructor: Dr. William Sewell

A1.

The telecommunications industry is competitive and full of options for the modern consumer. Because of this, internet service providers face the challenge of pricing to not only draw in new customers, but retain their existing customer base (Shukla, 2022). In this analysis, we will be exploring whether or not monthly charges have a significant influence on customer churn.

A2.

The stakeholders in the organization could benefit greatly from analyzing the relationship between monthly charges and customer churn. Understanding this relationship would allow the organization to develop better pricing strategies and prevent existing customers from switching to a competitor. It could also help the organization attract new customers and further increase revenue.

A3.

To perform this analysis, I will be using the “MonthlyCharge” and “Churn” variables from the churn dataset.

B1.

B2.

Please see the python code below. The output indicates that there is a p-value of 0.0.

```python
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns

df_churn = pd.read_csv('churn_clean.csv')

#Create two arrays for the possible responses in Churn column
a = df_churn[df_churn["Churn"] == "Yes"].MonthlyCharge
b = df_churn[df_churn["Churn"] == "No"].MonthlyCharge

#Generate t-test using Scipy library
stats.ttest_ind(a, b)
```

```
Ttest_indResult(statistic=40.18947672237426, pvalue=0.0)
```

B3.

I chose to perform a two-sample t-test because I wanted to see if the mean monthly charge between two samples is equal. The two samples being reviewed are customers that have been churned and customers that have been retained. The t-test is an effective method of analyzing the relationship between a categorical and a numerical variable (Bevans, 2020).

C.

The variable "MonthlyCharge" has a normal distribution, with most of the values concentrated near the mean. The variable "Tenure" appears to have a bi-modal distribution, as most of the values are concentrated near the upper and lower ends of the histogram.

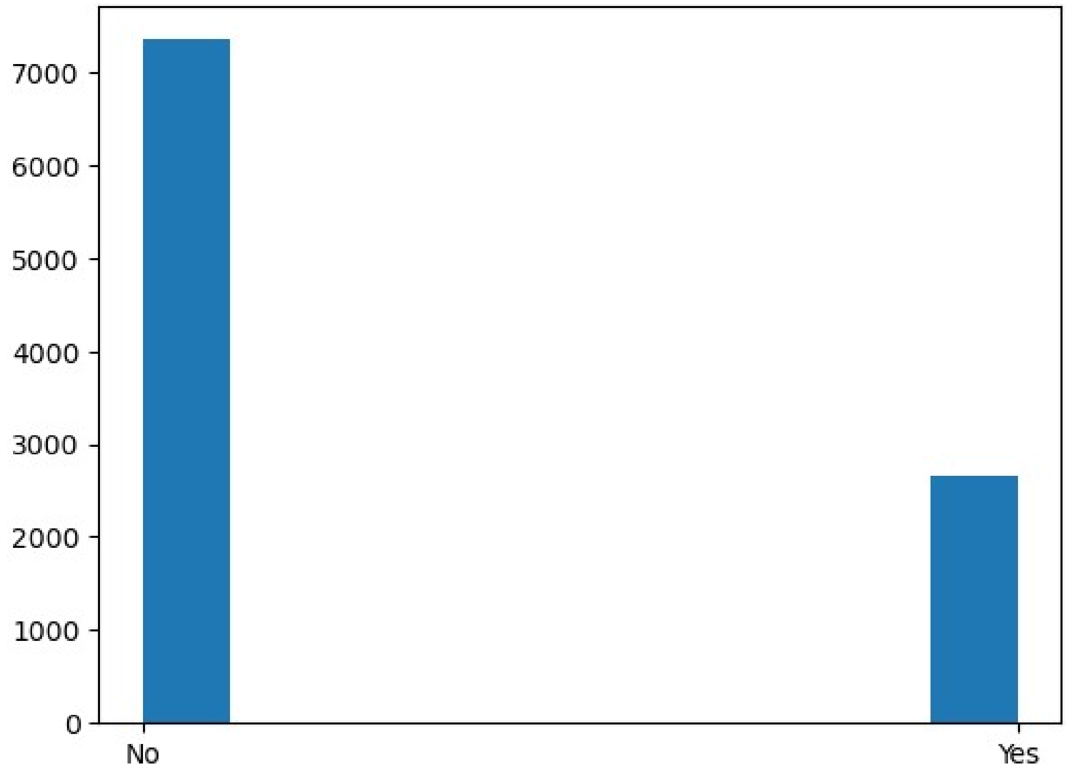
Both categorical variables, "Churn" and "Techie", have skewed distributions. Theses variables have a far greater distribution of "No" samples to "Yes" samples.

C1.

Please see the distributions below for 2 continuous variables ("MonthlyCharge", "Tenure") and 2 categorical variables ("Churn", "Techie").

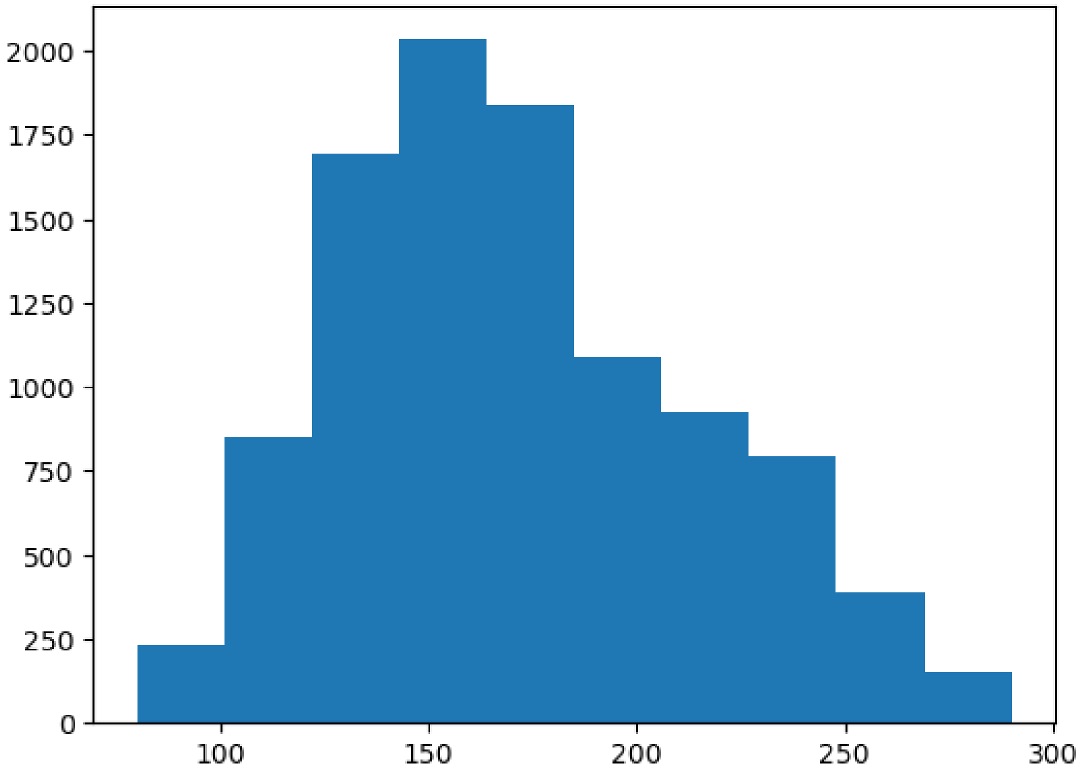
```
#Univariate distribution of Churn
plt.hist(df_churn['Churn'])
```

```
(array([7350.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,
        2650.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```



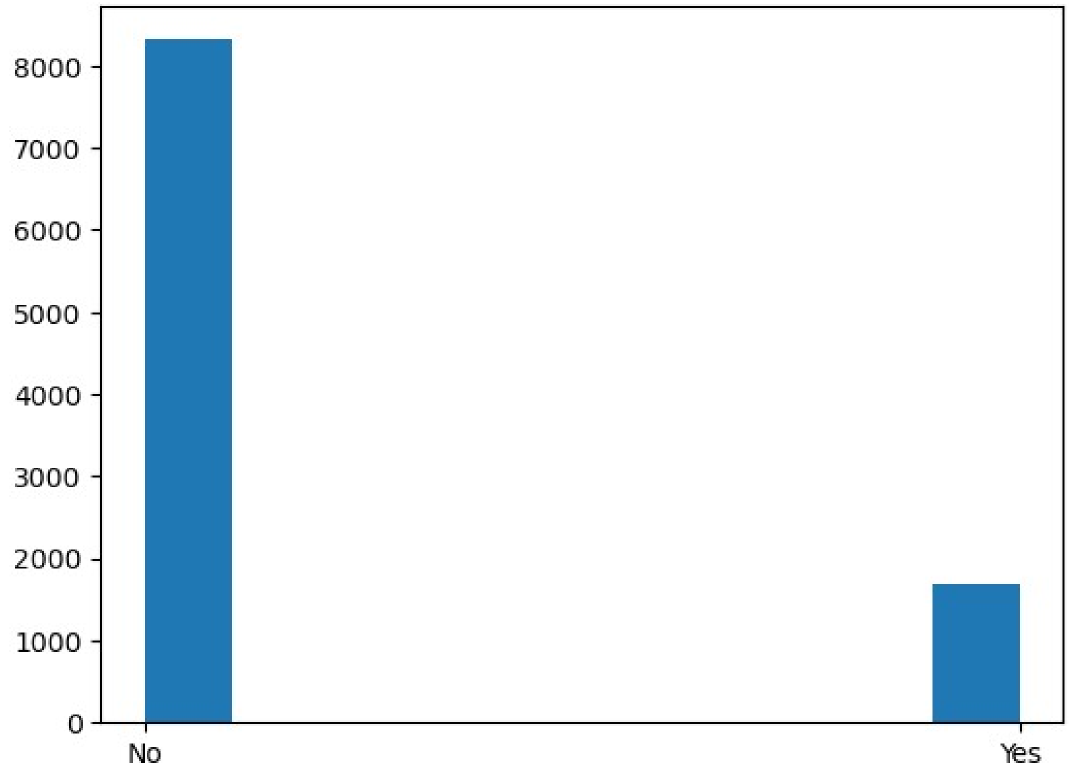
```
#Univariate distribution of MonthlyCharge
plt.hist(df_churn['MonthlyCharge'])

(array([ 230.,  853., 1695., 2034., 1842., 1089.,  926.,  791.,  388.,
         152.]),
 array([ 79.97886  , 100.9970159, 122.0151718, 143.0333277,
164.0514836,
        185.0696395, 206.0877954, 227.1059513, 248.1241072,
269.1422631,
        290.160419 ]),
 <BarContainer object of 10 artists>)
```



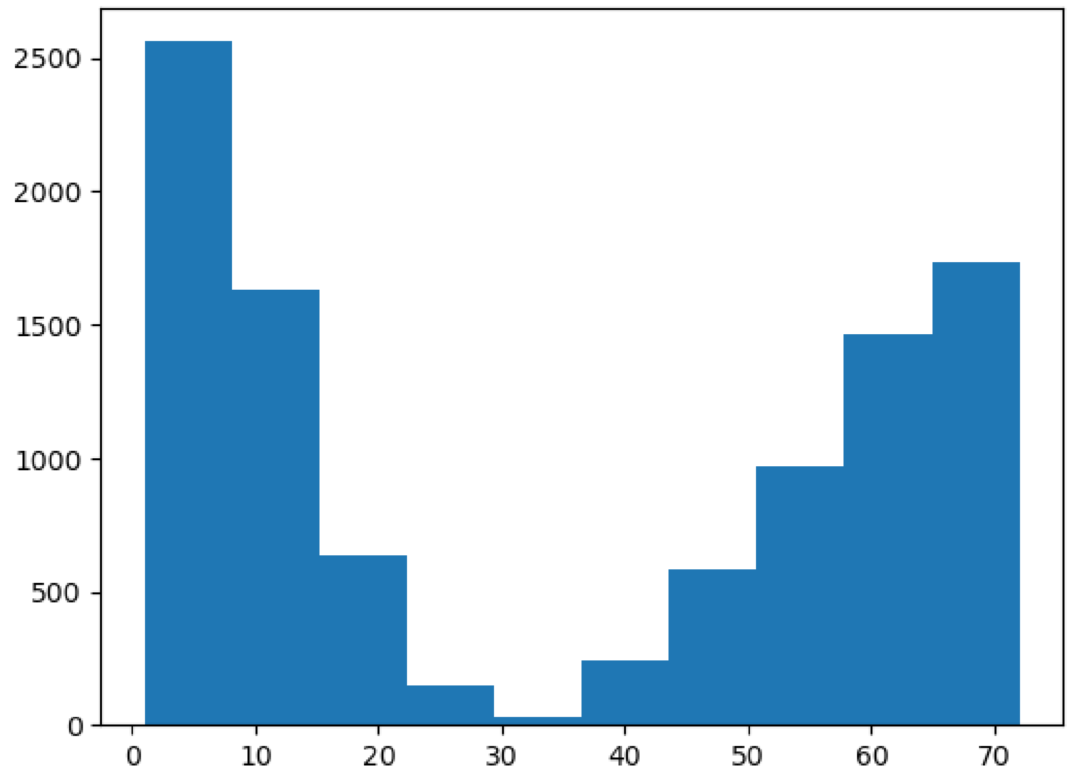
```
#Univariate distribution of Techie
plt.hist(df_churn["Techie"])

(array([8321.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,
        1679.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```



```
#Univariate distribution of Tenure
plt.hist(df_churn["Tenure"])

(array([2560., 1634.,  636.,  146.,   32.,  239.,  583.,  969., 1465.,
        1736.]),
 array([ 1.00025934,  8.10016141, 15.20006347, 22.29996554, 29.3998676
,
        36.49976967, 43.59967174, 50.6995738 , 57.79947587,
64.89937793,
        71.99928   ]),
 <BarContainer object of 10 artists>)
```



D.

I performed bivariate analysis using two continuous variables(Bandwidth_GB_Year and Outage_sec_perweek) and two categorical variables(StreamingMovies and PaperlessBilling).

To analyze the relationship between Bandwidth_GB_Year and Outage_sec_perweek, I used the plot() function from the matplotlib library and generated a scatterplot. As we can see below, the values seem to be distributed into two areas. Most of the values in Outage_sec_perweek fall within 5 and 15 seconds. Bandwidth_GB_Year values are concentrated between the rough range of 500 and 2500, and the range of 4000 and 7000 GB. There are relatively few values distributed between these two concentrated areas.

To analyze the relationship between StreamingMovies and Bandwidth_GB_Year, I used the boxplot function from seaborn. The boxplots appear very similar with a few exceptions. For the "Yes" response in streaming movies, the middle 50% of values are distributed between roughly 1300GB and 6700GB. The "No" response is only slightly different, with the middle 50% of values being distributed between 1000GB and 6400GB. The upper and lower whiskers for these boxplots are nearly identical as well.

I also used a boxplot function to display the relationship between PaperlessBilling and Bandwidth_GB_Year. The distribution of values in Bandwidth_GB_Year is nearly identical in both the "Yes" and "No" columns of PaperlessBilling. The middle 50% of values in both boxplots is distributed within roughly 1200GB and 6600GB. The only difference between these bivariate representations is that the mean value in "Yes" is slightly higher than in "No".

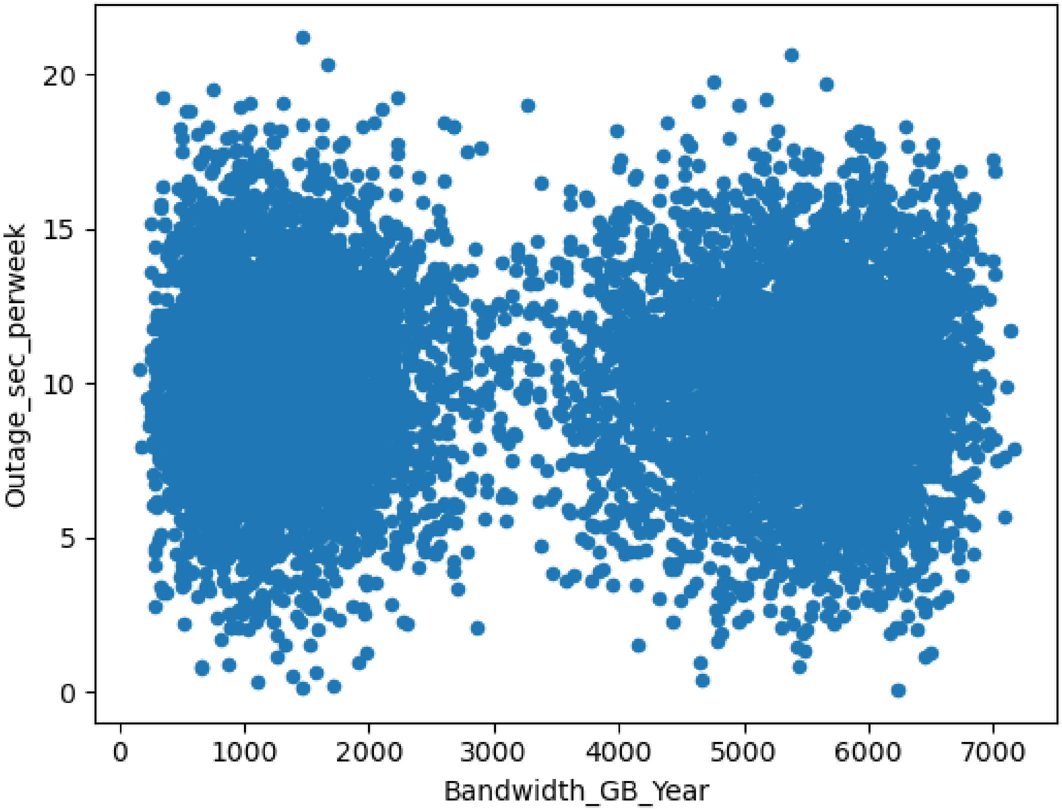
Lastly, I used crosstab() function from pandas to analyze the relationship between the PaperlessBilling and StreamingMovies variables. The greatest distribution of values appears to be in the "Yes" response for both of these variables. The smallest distribution of values appears in the "No" Response for PaperlessBilling and the "Yes" response for StreamingMovies. Over half of the values are distributed in the "Yes" column for PaperlessBilling.

D1.

Please see the bivariate distributions below.

```
#Bivariate distribution of two continuous variables, Bandwidth_GB_Year
and Outage_sec_perweek
df_churn.plot(kind="scatter", x="Bandwidth_GB_Year",
y="Outage_sec_perweek")
```

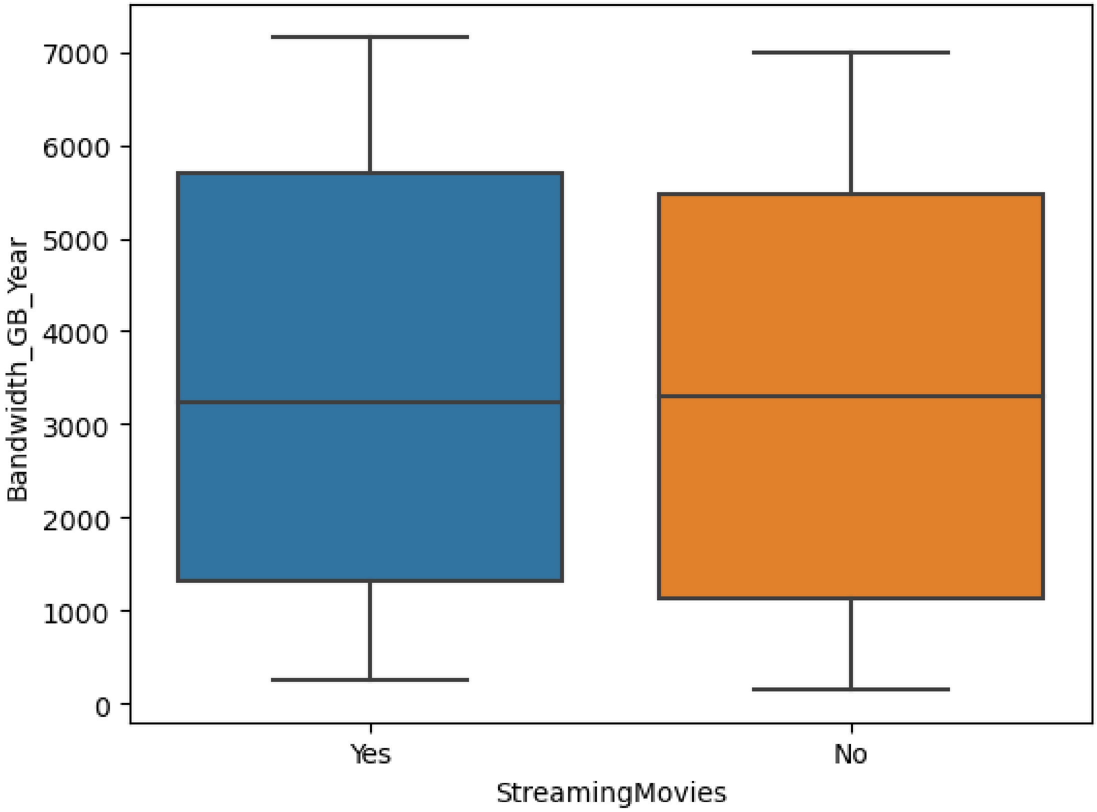
```
<Axes: xlabel='Bandwidth_GB_Year', ylabel='Outage_sec_perweek'>
```



```
# Bivariate distribution between one continuous and one categorical
variable
# Bandwidth_GB_Year and StreamingMovies

sns.boxplot(x="StreamingMovies", y="Bandwidth_GB_Year", data =
df_churn)
```

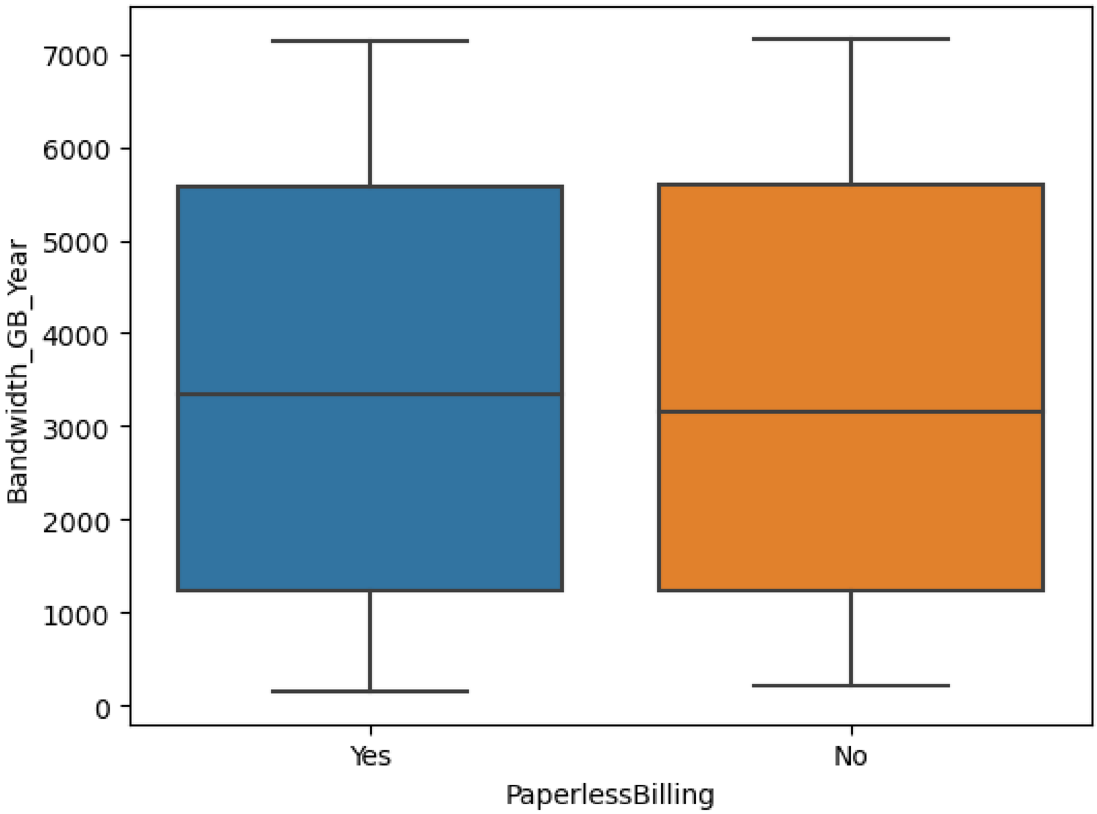
```
<Axes: xlabel='StreamingMovies', ylabel='Bandwidth_GB_Year'>
```



```
# Bivariate distribution between one continuous and one categorical
variable
# Bandwidth_GB_Year and StreamingMovies

sns.boxplot(x="PaperlessBilling", y="Bandwidth_GB_Year", data =
df_churn)

<Axes: xlabel='PaperlessBilling', ylabel='Bandwidth_GB_Year'>
```



```
# Bivariate distribution for two categorical variables
# StreamingMovies and PaperlessBilling

pd.crosstab(df_churn["StreamingMovies"], df_churn["PaperlessBilling"])

PaperlessBilling    No   Yes
StreamingMovies
No                2106  3004
Yes               2012  2878
```

E1.

The t-test performed in section b has returned a p-value of 0.0. Because this falls below the alpha value of 0.05 we can interpret this to mean that there is a statistically significant difference in monthly charge between customers who have been retained and customers who have been churned. This seems to be a reasonable conclusion because the telecommunications market is competitive and price-sensitive consumers will likely explore other options if they deem a certain monthly charge too high.

E2.

Although the two-sample t-test is a good starting point for exploring the relationship between "Churn" and "MonthlyCharge" there might be better alternatives for a more detailed analysis. A limitation of two sample t-tests is that they make assumptions about factors such as independence, normality, homogeneity of variances, and random sampling (Rani, 2022). Luckily, "MonthlyCharge" does have a normal distribution, but if this was not the case the results of the analysis might be inaccurate.

E3.

In the future, it might be a good idea to perform a more detailed type of analysis. Logistic regression is a useful tool for more in-depth analysis because it can be used to predict a binary outcome based on a numerical input (Edgar et al, 2017). This would help us to better understand the relationship between the "MonthlyCharge" and "Churn" variables. It could greatly benefit the organization by allowing us to predict if any customers are at risk of churn. This could lead to greater customer retention down the line.

F.

Please see the link below for the panopto video recording.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3a18db13-12c5-42c1-ba35-b0e7000d55a5

G.

Shukla, V. (2022, November 22). Is price a significant predictor of the churn behavior during the global pandemic? A predictive modeling on the telecom industry. Journal of Revenue and Pricing Management, 21, 470-483. https://doi.org/10.1504/IJMC.2019.096518

Bevans, R. (2020, January 28). Choosing the Right Statistical Test | Types & Examples. Scribbr. https://www.scribbr.com/statistics/statistical-tests/#:~:text=Comparison%20tests%20look%20for%20differences%20among%20group%20means.,%28e.g.%2C%20the%20average%20heights%20of%20men%20and%20women%29.

Rani, B. (2022, May 5). LIMITATIONS INVOLVED IN A TWOSAMPLE INDEPENDENT T-TEST. International Journal of Creative Research Thoughts (IJCRT), 10(5), 44. https://ijcrt.org/papers/IJCRTQ020009.pdf

Edgar, T. W., & Manz, D. O. (2017). Logistic Regression. ScienceDirect. https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.