Telecom Churn Analysis

Javier Lopez

D207, Exploratory Data Analysis

Table of Contents

A1. Research Question

What variable(s) are related to churn?

A2. Benefits

The data dictionary states that " it costs ten times more to acquire a new customer than to retain an existing one." Determining each customer's risk of churn is crucial to reducing operating costs overall. To get to a point where the risk can be calculated for each customer, we must first identify what key factors impact churn. By conducting our analysis, we will be able to understand customer characteristics related to churn. In turn, stakeholders can utilize the data to achieve the understanding necessary to develop action plans that will reduce customer churn.

A3. Relevant Data

Our data set consists of 10,000 observations from a telecommunication company with 49 variables. Once cleaned, there were 10,000 observations with 18 variables remaining. Within the data set, churn is our target variable, and we will analyze the variables listed below as our independent variables. Because all of the remaining variables are nominal, we will conduct the chi-square independence test to identify if a relationship exists.

Variable	Data Type	Description	Example
Area	String, Nominal	Character string indicating the area type, based on census data	rural, urban, suburban
Marital	String, Nominal	Character string indicating customer's marital status at sign-up	Married
Gender	String, Nominal	Character string indicating customer's gender self-identification	Male, Female, Nonbinary
Churn	String, Nominal	Character string indicating if the customer discontinued service within the last month	yes, no
Techie	String, Nominal	Character string indicating whether customer considers themselves technically inclined	yes, no
Contract	String, Nominal	Character string indicating the customer's contract term	month-to-month, one year, two year
Port_modem	String, Nominal	Character string indicating whether customer has a portable modem	yes, no
Tablet	String, Nominal	Character string indicating whether customer owns a tablet	yes, no
InternetService	String, Nominal	Character string indicating customer's internet service provider	DSL, fiber optic, none
Phone	String, Nominal	Character string indicating whether customer has a phone service	yes, no
Multiple	String, Nominal	Character string indicating whether customer has multiple lines	yes, no
OnlineSecurity	String, Nominal	Character string indicating whether customer has an online security add-on	yes, no
OnlineBackup	String, Nominal	Character string indicating whether customer has an online backup add-on	yes, no
DeviceProtection	String, Nominal	Character string indicating whether customer has a device protection add-on	yes, no
TechSupport	String, Nominal	Character string indicating whether customer has a technical support add-on	yes, no
StreamingTV	String, Nominal	Character string indicating whether customer has streaming TV	yes, no

Variable	Data Type	Description	Example
StreamingMovies	String, Nominal	Character string indicating whether customer has streaming movies	yes, no
PaperlessBilling	String, Nominal	Character string indicating whether customer has paperless billing	yes, no
PaymentMethod	String, Nominal	Character string indicating customer's payment method	electronic check

B1. Python Code

```
## Loop through each variable
for col in df.columns:
    if col != 'Churn':
        ## Create  and view contingency table
        contingency = pd.crosstab(df['Churn'], df[col])
        display(contingency)
        ## Run chi2 test of independence
        statistic, pvalue, dof, frequencies = chi2_contingency(contingency)
        ## Store results in DataFrame
        chi2['statistic'][col] = statistic
        chi2['pvalue'][col] = pvalue
        if pvalue <= 0.05:
            chi2['results'][col] = 'reject'
        else:
            chi2['results'][col] = 'accept'
```

B2. Chi-Square Test Results

	critical	statistic	pvalue	results
Area	5.991465	2.439074	0.295367	accept
Marital	9.487729	5.565781	0.234008	accept
Gender	5.991465	7.880065	0.019448	reject
Techie	3.841459	44.114794	0.0	reject
Contract	5.991465	718.591581	0.0	reject
Port_modem	3.841459	0.628905	0.427757	accept
Tablet	3.841459	0.064075	0.800168	accept
InternetService	5.991465	87.462046	0.0	reject
Phone	3.841459	6.711745	0.009578	reject
Multiple	3.841459	173.037988	0.0	reject
OnlineSecurity	3.841459	1.76975	0.183413	accept
OnlineBackup	3.841459	25.281554	0.0	reject
DeviceProtection	3.841459	31.653203	0.0	reject
TechSupport	3.841459	3.461224	0.062824	accept
StreamingTV	3.841459	528.651861	0.0	reject
StreamingMovies	3.841459	835.414013	0.0	reject
PaperlessBilling	3.841459	0.462398	0.496505	accept
PaymentMethod	7.814728	9.437373	0.024007	reject

B3. Justification: Analysis Technique

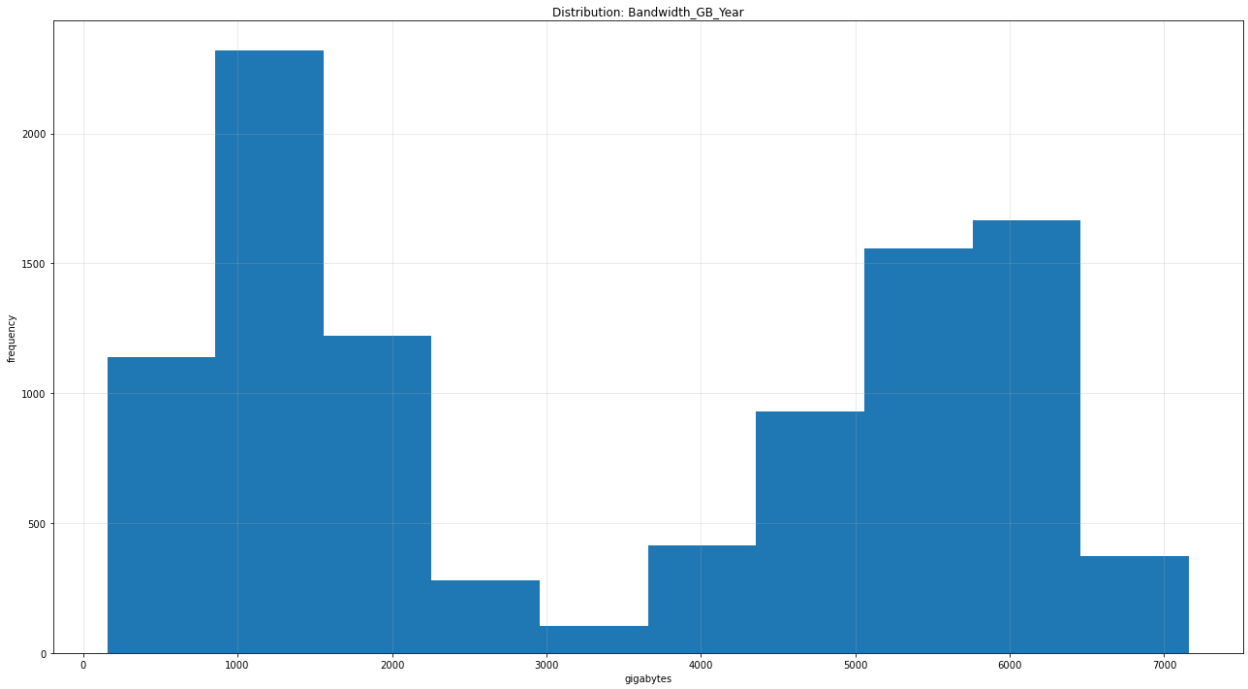
We used the chi-square test of independence to test the relationship between churn and our independent variables because they were all nominal variables. We ran our test using the chi-squared contingency function from the scipy package. Given a choice between chi-square, t-test, and ANOVA, the chi-square test of independence matched what we were looking to accomplish and required no normality test.

C. Univariate Analysis

Continuous Variables

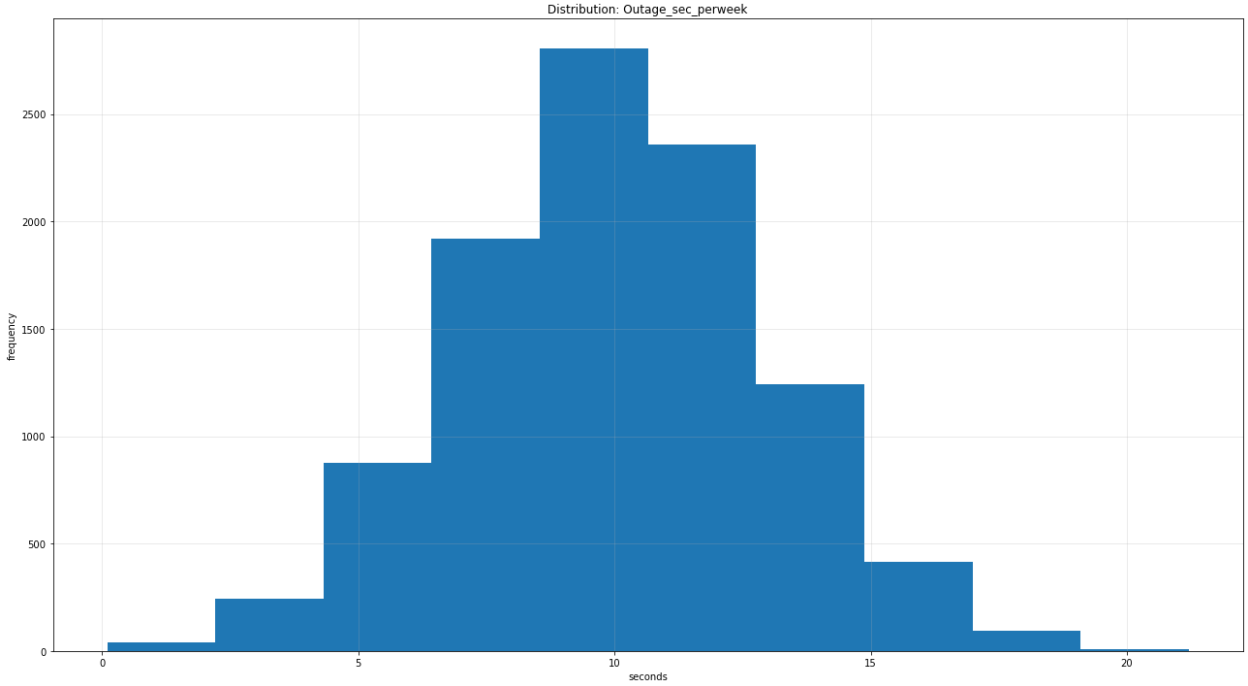
We analyzed the Bandwidth_GB_Year and Outage_sec_perweek continuous variables and found that they have different distributions. The distribution, as shown below, for Bandwidth_GB_Year is bimodal, with the two modes at around 1,000 gigabytes and 6,000 gigabytes. On the other hand, Outage_sec_perweek has a normal distribution with a mode at around 10 seconds. The distribution for Outage_sec_perweek can also be found in the graph below.

Bandwidth_GB_Year



```
count    10000.000000
mean      3392.341550
std       2185.294852
min        155.506715
25%       1236.470827
50%       3279.536903
75%       5586.141370
max       7158.981530
Name: Bandwidth_GB_Year, dtype: float64
```

Outage_sec_perweek

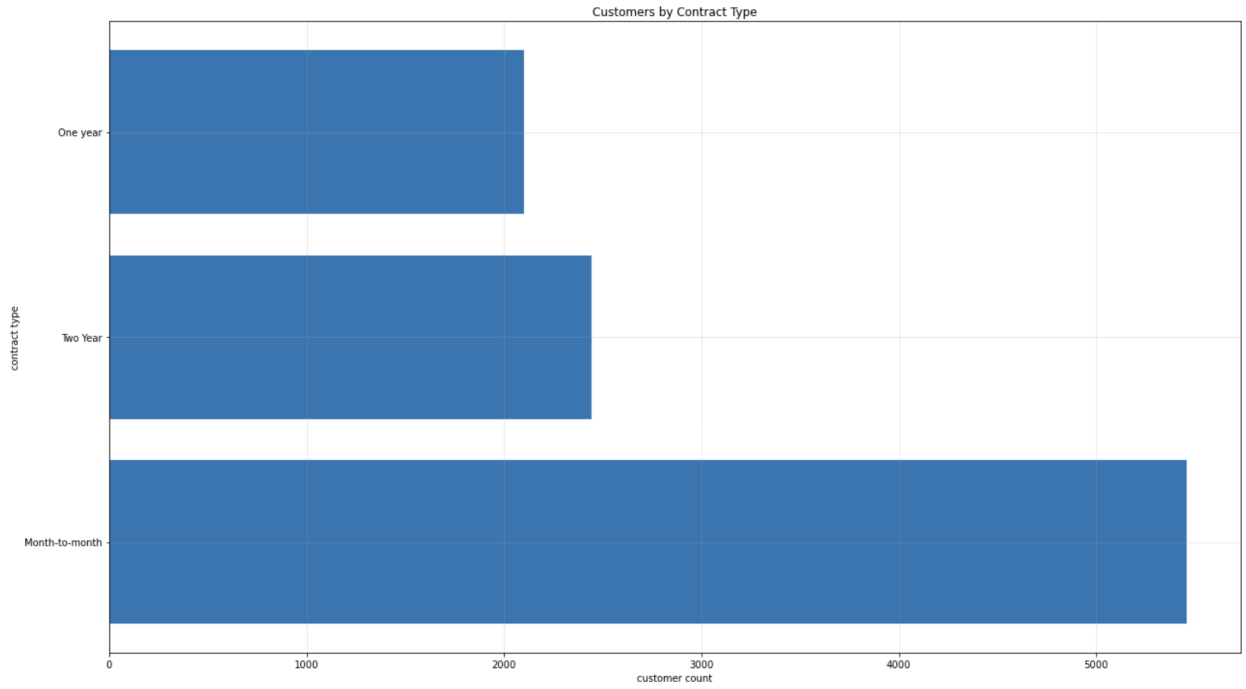



```
count    10000.000000
mean        10.001848
std          2.976019
min          0.099747
25%          8.018214
50%         10.018560
75%         11.969485
max         21.207230
Name: Outage_sec_perweek, dtype: float64
```

Categorical Variables

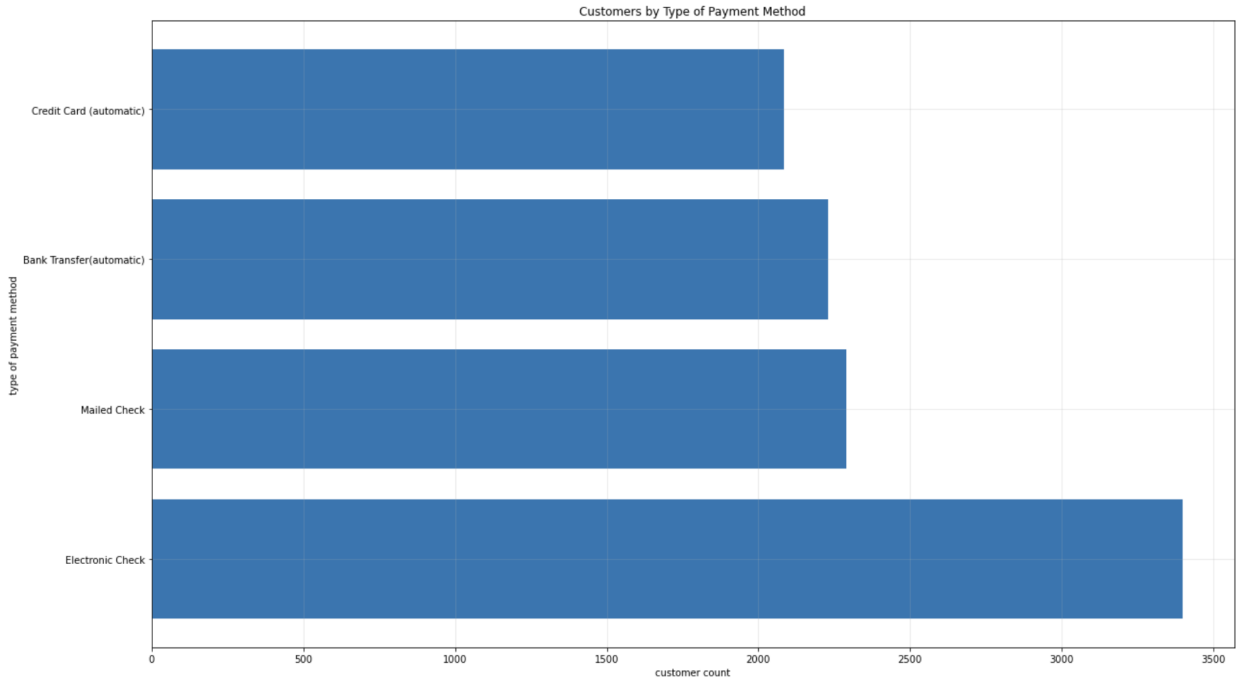
We analyzed two categorical variables, both nominal variables, PaymentMethod and Contract. Through our analysis, we found that most customers prefer less commitment with the month-to-month contract type at 54.56%. However, customers who commit to a longer-term contract sign up for two years more frequently than one-year contracts. The exact figures for each category are in the graph and summary below. Regarding payment methods, over a third of customers prefer to pay via electronic check, 33.98%, while 22.9% prefer to mail a check. The summary with exact figures for each payment type and their distribution can be seen in the graph below.

Contract



```
Distribution of Contract
Month-to-month: 54.56%, 5456 occurences
Two Year: 24.42%, 2442 occurences
One year: 21.02%, 2102 occurences
```

PaymentMethod



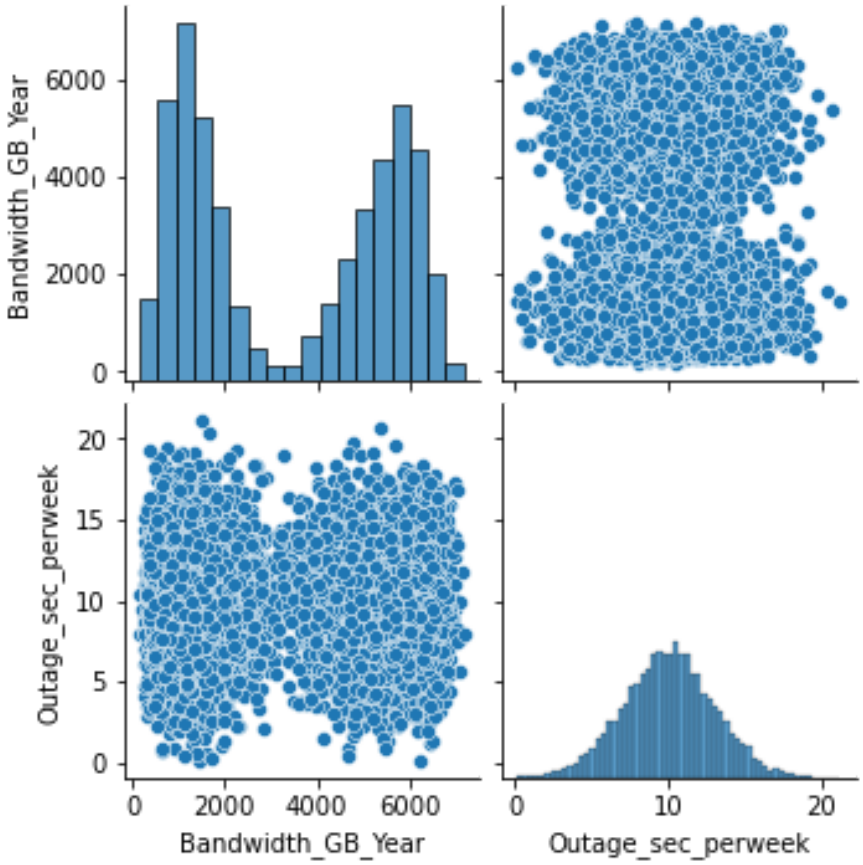

```
Distribution of PaymentMethod
Electronic Check: 33.98%, 3398 occurences
Mailed Check: 22.9%, 2290 occurences
Bank Transfer(automatic): 22.29%, 2229 occurences
Credit Card (automatic): 20.83%, 2083 occurences
```

D. Bivariate Analysis

Continuous Variables

We analyzed Bandwidth_GB_Year and Outage_sec_perweek and found no distinguishable relationship between the two variables. The correlation coefficient

between the two variables of 0.004176 further quantifies the lack of relationshipbetween Bandwidth_GB_Year and Outage_sec_perweek. The two variables can be seen in the graphic below.

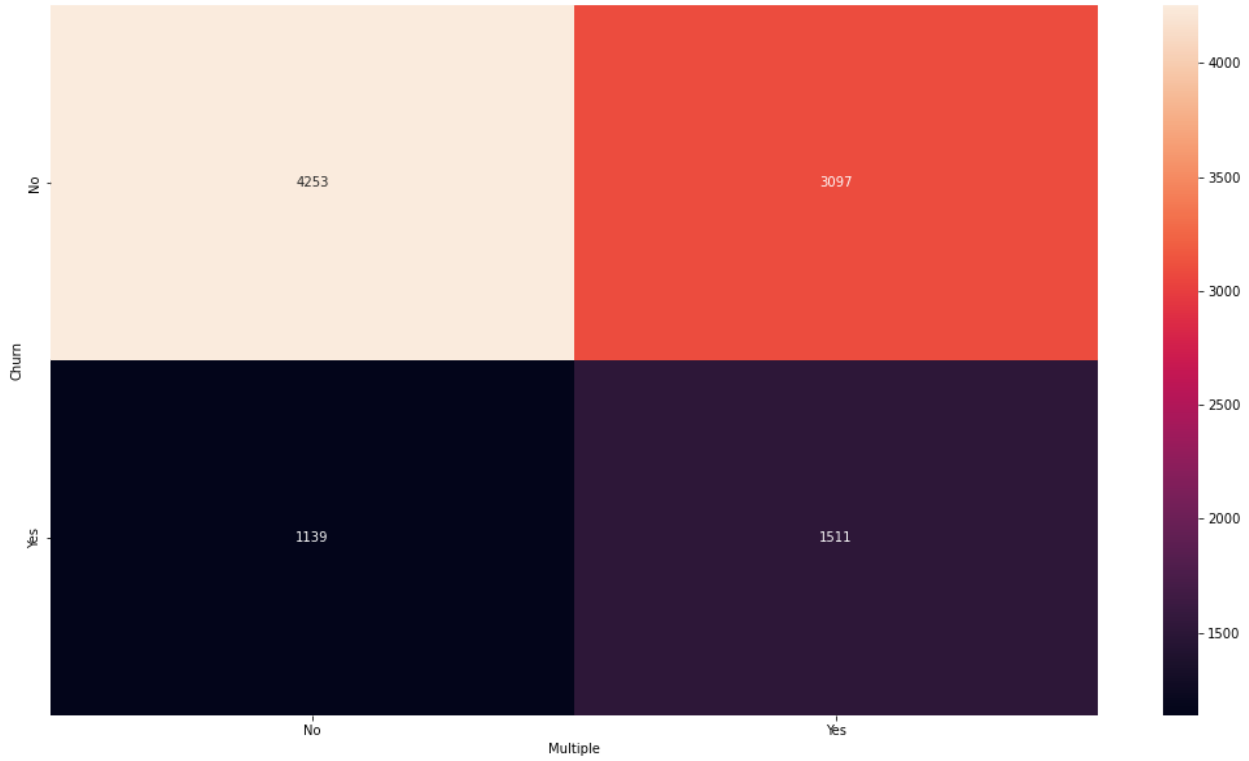


Categorical Variables

To analyze the relationship between the Churn and Multiple categorical variables, we created a contingency table. The table showed us that more customers that had a single service were retained that those that were subscribed to multiple services. When we conducted the chi-square test for the two variables, the results further quantified the relationship between the two variables and helped us to reject the null hypothesis. The

contingency table and heatmap below depict the distributions and relationship between the two variables.

Multiple Churn	No	Yes
No	4253	3097
Yes	1139	1511



E1. Hypothesis Test Results

The hypotheis test we conducted utilized the chi-square test from scipy. The test results showed a significant relationship between Churn and the following variables: Gender, Techie, Contract, InternetService, Phone, Multiple, OnlineBackup, DeviceProtection, StreamingTV, StreamingMovies, PaymentMethod. All of the variables

mentioned had a higher test statistic than the critical value attained from the degrees of freedom, and their pvalue was below the 0.05 threshold set.

E2. Limitations

The biggest limitation in our data analysis was the amount of variables that were tested. A thorough analysis conducted to understand the relationship between continuous and categorical variables can provide further insight into potential correlations between churn and other variables. For example, testing the relationship between churn and tenure or churn and monthly charge can provide insight into the amount of time or money that a customer is willing to spend with the services offered. Furthermore, knowing the individual tenure a customer has had with each service can help dig deeper into the performance of each service being offered to the customer.

E3. Course of Action

Based on the results of the chi-square test we conducted, we would recommend that the service provider assess the targeting that their marketing department is using to attract customers and the retargetting that is used on current customers. I would also recommend that the marketing and sales department work in conjunction with the data department to further understand the best performing services within the provider's offerings and create a promotion strategy that target customer reaching a tenure where they are at highest risk of churn to attract them to additional services or discounts.

G. Third-Party Code

No third party code was used for the analysis conducted.

H. Sources

No external sources were used for the analysis conducted.