D208- Predictive Modeling Performance Assessment

Task 2: Logistic Regression Modeling

Western Governs University

Table of Contents

Part I: Research Question 3

A1. Research Question 3
A2. Goals 3

Part II: Method Justification 3

B1. Summary of Assumptions 3
B2. Tool Benefits 4
B3. Appropriate Technique 5

Part III: Data Preparation 5

C1. Data Cleaning 5
C2. Summary Statistics 7
C3. Visualizations 9
C4. Data Transformation 24
C5. Prepared Data Set 27

Part IV: Model Comparison and Analysis 27

D1. Initial Model 27
D2. Justification of Model Reduction 29
D3. Reduced Linear Regression Model 30
E1. Model Comparison 31
E2. Output and Calculation 34
E3. Code 37

Part V: Data Summary and Implications 37

F1. Results 37
F2. Recommendations 40

Part VI: Demonstration 41

G. Panopto Demonstration 41
H. Sources of Third-Party Code 41
I. Sources 41

Part I: Research Question

A1. Research Question:

What factors contribute to the churn rate?

A2. Goals

Customers can choose from multiple telecommunication services as their providers. Customer churn is the percentage of customers who discontinued their services from a provider within a specific time frame. The churn rate in the data set is documented by customers who discontinued their services within the last month. By examining the reasons behind customer churn stakeholders can identify patterns in customer behaviors and preferences for increased customer retention. This can lead to decreased churn rates and increased business profits. This analysis will focus on the following variables: MonthlyCharge, Tenure, Yearly_equip_failure, Children, Age, Income, Email, Contacts, TechSupport, outage_sec_perweek, Population Gender, and Techie.

Part II: Method Justification

B1. Summary of Assumptions

A logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and the independent variable(s). In this analysis, the dependent variable is a categorical variable and has binary values of 0 and 1. The logistic regression model uses the sigmoid function to estimate the probability of the dependent variable being in a particular category based on the values of the independent variable (Lacrose & Lacrose, 2019).

There are several assumptions that can be made when preparing to complete a logistical regression model. When completing a logical regression model one assumption is that there will be a binary outcome. Because this model uses a categorical variable as the dependent variable the outcome should be binary without only two possible outcomes. Another assumption of a logistic regression model is linearity. The relationship between the independent variable and the logit of the outcome variable should be linear. Another assumption is no multicollinearity. The independent variables should not be highly corrected with each other. Another assumption is that

the logistic regression model assumes that there is no perfect separation of the outcome variable by the independent variable.

B2. Tool Benefits

Data cleaning for the churn data set was completed utilizing Python. Python is an open-sourced programing language used for analysis and development. Python has a consistent syntax that makes coding and debugging user-friendly for beginners. Python is flexible and has the ability to import packages and to tailor data. The following packages were imported and used for their advantages. Below are the packages/libraries used for the analysis.

```
# Standard data science imports
import numpy as np
import pandas as pd

# Visualization libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Statistics packages
import statsmodels.api as sm
import statistics
from scipy import stats
from scipy.stats import f_oneway

# Scikit-learn
import sklearn
from sklearn.preprocessing import OneHotEncoder
from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# Ignore Warning Code
import warnings
warnings.filterwarnings('ignore')
```

B3. Appropriate Technique

An appropriate technique to answer the question 'which factors impact churn rate?' is a logistic regression model. Logistic regression is an appropriate technique to answer the question and to determine which independent variables impact the dependent variable, churn. Churn is a categorical variable with binary data. The logistic regression method is appropriate to analyze churn data because it is a binary classification method designed to model the probability of a binary outcome variable like churn. In the analysis, the outcome variable is whether a customer has changed or not. The logistic regression model can be used to predict the likelihood of churn based on the relationship between the outcome variable and the independent variables. such as customer tenure, monthly charges, and other demographic variables. Logistic regression can also help to identify the most important predictors of churn which can be used to develop strategies to reduce churn and retain customers.

Part III: Data Preparation

C1. Data Cleaning

The methods discussed below are used to clean the raw data set of the churn data. Cleaning data is important in drawing accurate conclusions when analyzing data. It allows for different sorting options, filtering, and modification of the data set. Using the methods helps detect duplicates while maintaining the integrity of the data. It is necessary to detect and treat duplicate data because duplicate entries can lead to miscalculations or misrepresentations of the data. Python was used to detect duplicate data, missing values, outliers, and any other data quality issues in the churn data set. Utilizing the “import pandas as pd” command Panda was imported into Python. The read_cvs() function was used to read the churn data on my local hard drive. This file was assigned to the variable “df” for easy reference. To determine the data types included in the churn data the df.info(file_path). The data type of each variable is needed information due to certain functions working only with specific functions. This includes the column names and the number of non-null values for each column. Once datatypes are known the data could be cleaned. Cleaning data included detecting duplicates, and identifying missing values, and outliers.

To determine if there were duplicate entries in the data the df.duplicated() function was completed. This function returns columns with TRUE or FALSE values. If the column returns a TRUE value there are duplicate records but if a FALSE value is returned there are no duplicates. The results of this function indicated all FALSE values meaning there were no duplicate values in the data set. To verify and count all entries that were FALSE the print(df.duplicated().value_counts()) was used. If duplicates were present the df.drop_duplicates() function could have been used to drop duplicate values.

The next step included determining if there were missing values. Missing values are usually represented in the form of nan, null, or none in the dataset. The df.isnull().sum() function was used. This function counted how many missing values were present in each column. There were no missing variables in the data set. Once duplicate and missing values were detected and treated outliers were then determined for all quantitative variables. All the quantitative variables from the churn data were plotted on boxplots to visualize outliers. Outliers need to be detected and treated because outliers can provide incorrect/inconsistent collusions from the data. Outliers come from data entry errors, measurement errors, experimental errors, sampling errors, or novelties in the data (Lacrose & Lacrose, 2019). Outliers were detected and removed utilizing z-score measures. The shape of the data set was seen before and after the removal of outliers to insure outliers was in fact removed from the data set. The code df.hist(figsize = (15,15)) was used to visualize the histograms of each column in the data. The df.drop function was used to drop the columns in the data set there were not useful. The df.shape function was used to see the dimensions of the data frame as a tuple containing the number of rows and the number of columns. The df.head function was used to display the data set with all the remaining columns.

There were two types of data from this data set, quantitative and qualitative data. Qualitative or categorical data (i.e. yes/no) requires re-expression or encoding of numbers to perform statical modeling (Lacrose & Lacrose, 2019). One hot encoding was thus used to transform categorical data into nominal data to be used in mathematical models like a logistic regression model. With one hot encoding, each categorical variable is represented as a binary vector where all elements are zero except the element corresponding to the category which is set to one. One hot encoding was used to change the dependent variable (churn) and the independent variables techie, techsupport, and gender to numerical values utilizing dummy columns. Original columns were dropped and

dummy columns remained. The df.column fx was used to visualize the remaining columns left in the data set.

C2. Summary Statistics

Listed below is a description of the dependent variable (Churn) and the independent variables used in this analysis.

Variable Name	Data Type	Description	Example
Population	Quantitative	Population of customer residence	8165
Children	Quantitative	Number of children of customer	5
Age	Quantitative	Age of customer	30
Income	Quantitative	Customer annual income reported	64256.81
Gender	Qualitative	Customer gender	Male
Outage_sec_perweek	Quantitative	Avg number of seconds per week of system outages in customer's neighborhood	12.63069124
Email	Quantitative	Number of emails sent to customer over past year	10
Contacts	Quantitative	Number of times customer contacted technical support	3
Yearly_equip_failure	Quantitative	Number of times customer's equipment failed and replaced	0
Techie	Qualitative	If customer considers themselves technically inclined	No
TechSupport	Qualitative	If customer has technical support add-on	Yes
Tenure	Quantitative	# of months customer has stayed with provider	10.06019902
MonthlyCharge	Quantitative	Amount charged to customer monthly	160.8055418
Churn	Qualitative	If the customer discontinued services in the last month	Yes

Below there is a summary of the statistics for the dependent variable and the independent variables. It is beneficial to get the summary statics when running a logistic regression model because it provides information about the relationship between the independent variables and the dependent variable. This was done with code df.describe() and can be seen below. The dependent variable in this analysis was 'Churn'. The continuous independent variables include

children, tenure, yearly_equip_failure, monthlycharge, age, income, email, contacts, outage_sec_perweek, population, and gender. The summary statistic:

df.describe():

	Population	Children	Age	Income	Outage_sec_perweek	Email	Contacts	Yearly_equip_failure	Tenure	MonthlyCharge
count	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000
mean	8508.409274	1.941453	53.161341	38329.400298	10.009065	12.021676	0.941676	0.374749	34.423473	172.783589
std	11759.988903	1.890668	20.634274	25123.528844	2.926500	3.011651	0.900798	0.582945	26.445276	42.990050
min	0.000000	0.000000	18.000000	348.670000	1.144796	3.000000	0.000000	0.000000	1.005104	79.978860
25%	727.250000	0.000000	35.000000	19041.117500	8.031398	10.000000	0.000000	0.000000	7.892645	139.979239
50%	2750.000000	1.000000	53.000000	32778.475000	10.016014	12.000000	1.000000	0.000000	29.772986	167.484705
75%	11838.750000	3.000000	71.000000	52280.437500	11.961618	14.000000	2.000000	1.000000	61.389790	202.443300
max	52967.000000	8.000000	89.000000	124025.100000	18.851730	21.000000	3.000000	2.000000	71.999280	290.160419

To see the statistical breakdown df.describe() was used. This method is included in the data profiling states which is a type of data transformation that involves analyzing and summarizing data to gain insight into the characteristics and quality of the dataset. This is also useful for quickly getting a picture of the distribution of the data in a column, including its range, central tendency, and dispersion (Massaron, 2016). This method returns a series object that includes the following summary statics for each variable in the data frame:

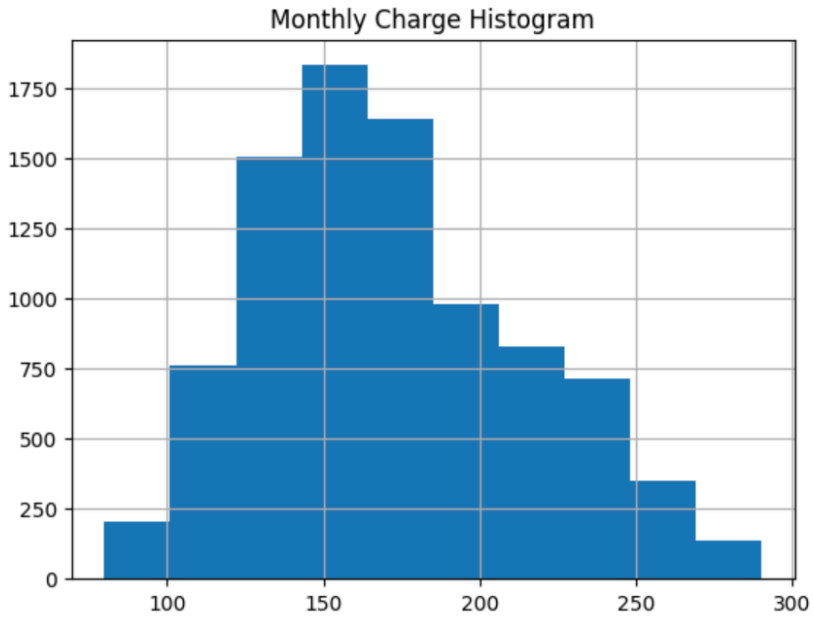
- count: the number of non-missing values in the column
- mean: the arithmetic mean of the values in the column
- std: the standard deviation of the values in the column
- min: the smallest value in the column
- 25% the 25th percentile of the values in the column
- 50% the 50th percentile (median) of the values in the column
- 75% the 75th percentile of the values in the column
- max: the largest value in the column

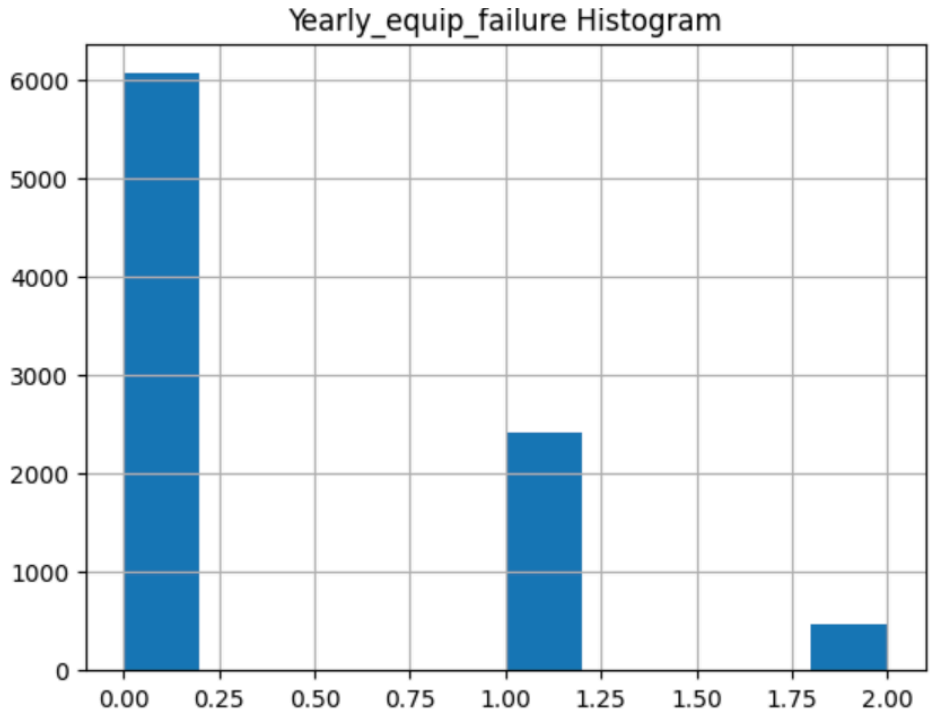
The count for the data set was 8950 observations for each variable. For the variable 'population' the minimum value is 0 which suggests that there may be some missing or errors in the dataset. The average population is 8508.41. When looking at the 'children' variable the mean value is 1.941453 meaning that on average each customer had 2 children. The standard deviation of 1.890668 suggests that there is considerable variation in the number of children among the

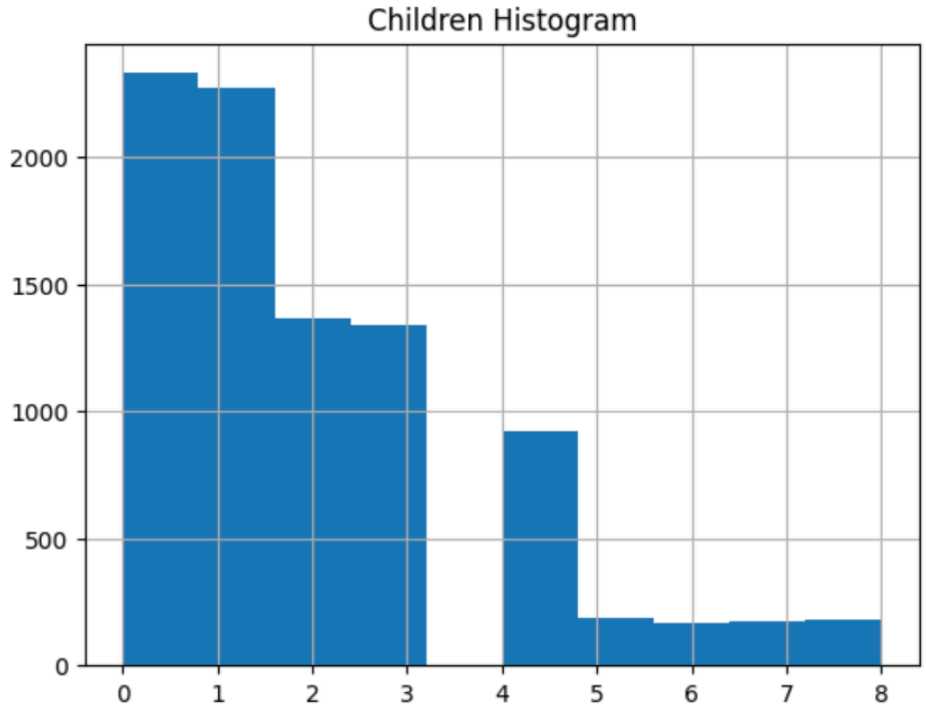
customers. The mean age is 53.1 with a minimum age of 18 years old and maximum age is 89 years old. 75% of the customers were 71 years or less. The average income is 38329.40 with a minimum income 348.67 and the maximum income 124025.10. The average for outage_sec_perweek was 10.009065. The variable email has a mean of 10.0090065. The minimum was 3 and the maximum was 21. The majority of customers received 8 to 14 emails per week. For the variable contacts, there was an average number of contacts per customer of around 12. 50% of the customers have between 10 and 14 contacts. For the variable yearly_equip_failure the average yearly equipment failure is 0.37. Most customers did not experience any equipment failure and 75% of customers experiences failures once or less. The average tenure is 34.42 months. The minimum tenure is 1.01 months and the maximum is 72 months. The average monthly change for customers was $173 (172.783589). Monthly charges ranged from $80 (79.978860) and a maximum of $290 (290.160419).

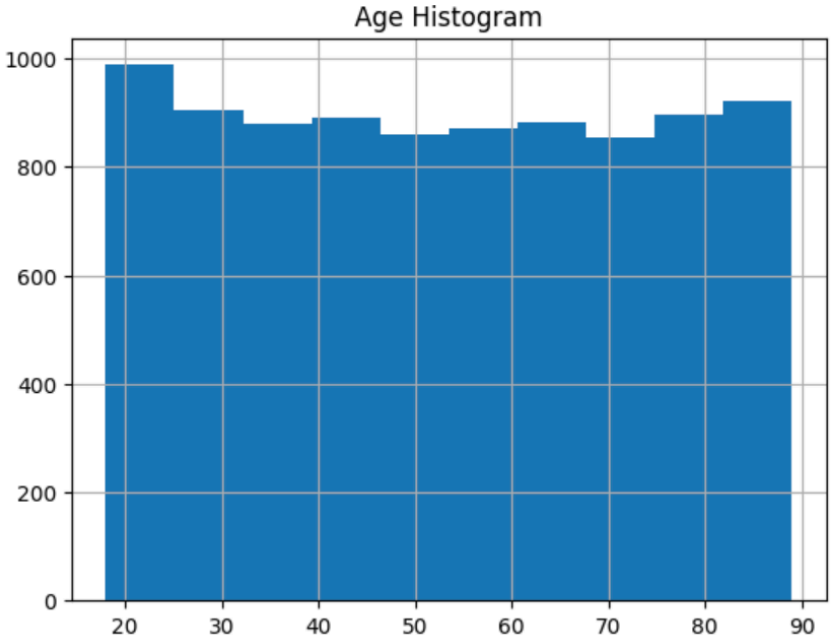
C3. Visualizations

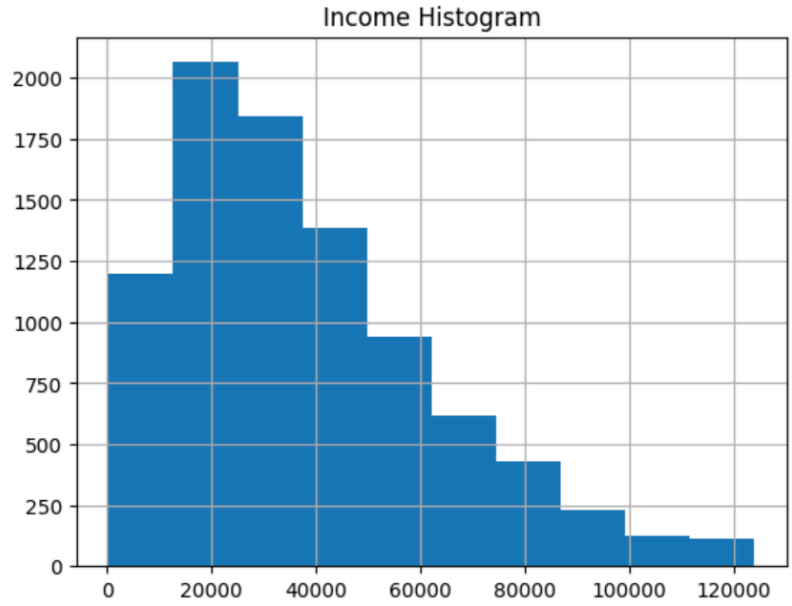
Univariate statistics is the statical analysis of a single variable at one time (Bruce et al., 2020). Below is the distribution of all independent variables in this analysis. A histogram was used to visualize the distribution of each variable.

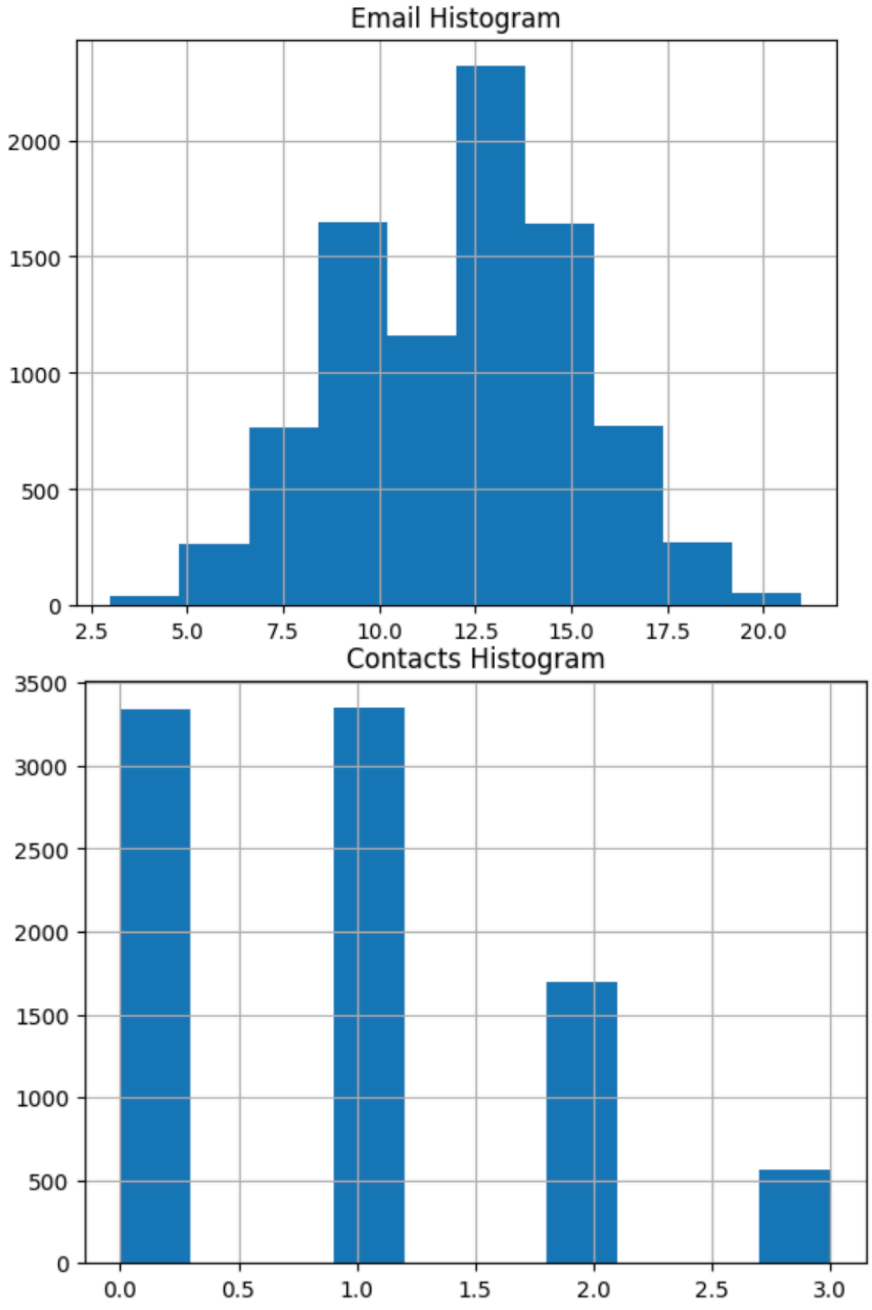


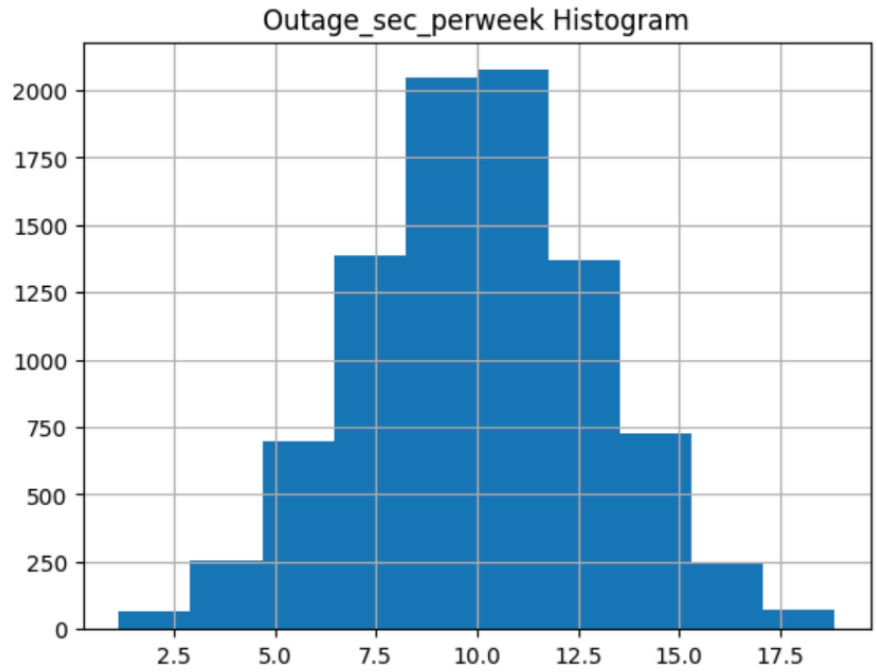

Yearly_equip_failure Histogram
6000
5000
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

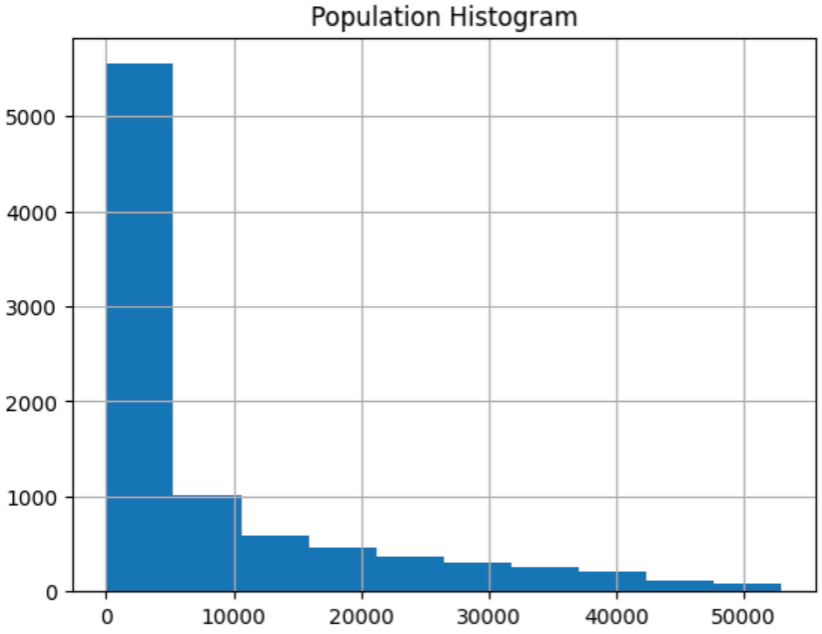

Children Histogram
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8

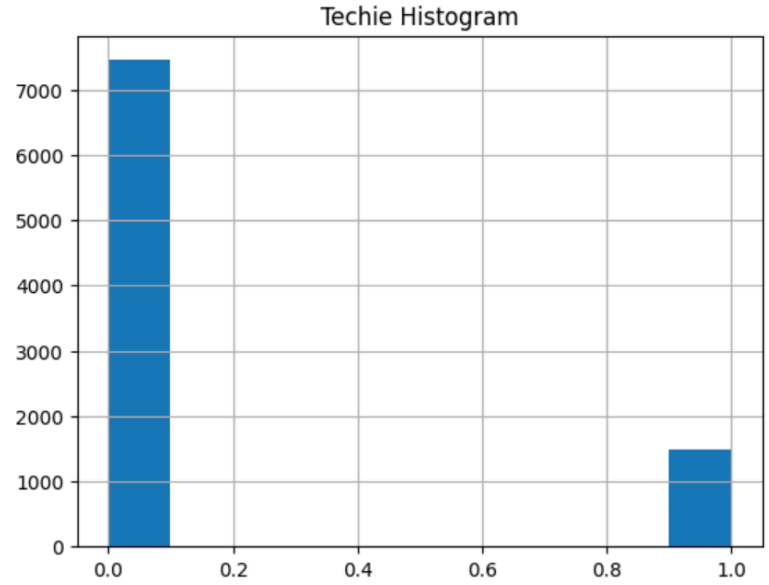

Age Histogram
1000
800
600
400
200
0
20
30
40
50
60
70
80
90

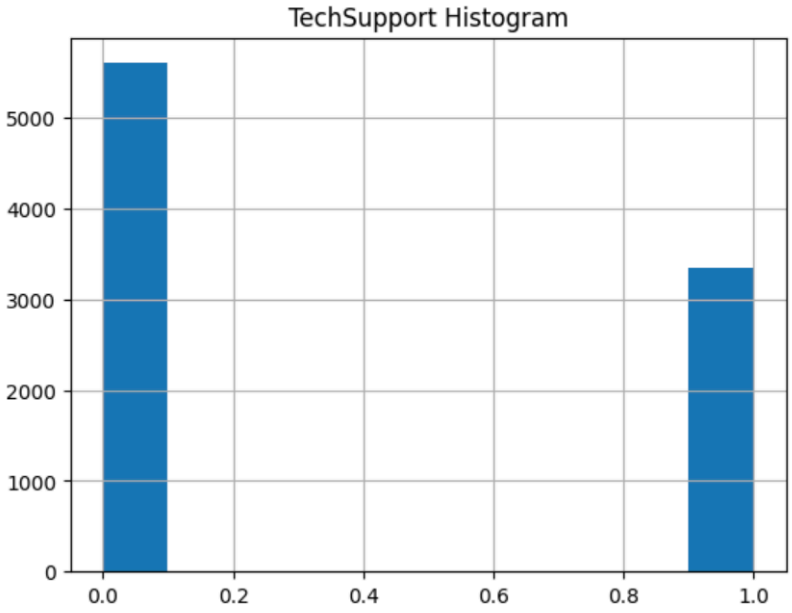

Income Histogram
2000
1750
1500
1250
1000
750
500
250
0
0
20000
40000
60000
80000
100000
120000

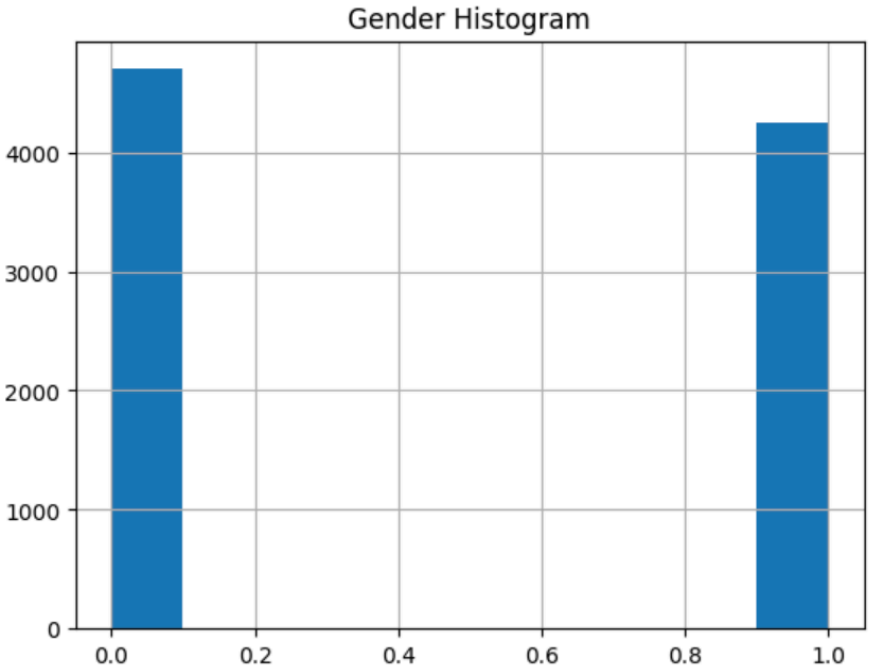

Email Histogram
2000
1500
1000
500
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Contacts Histogram
3500
3000
2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

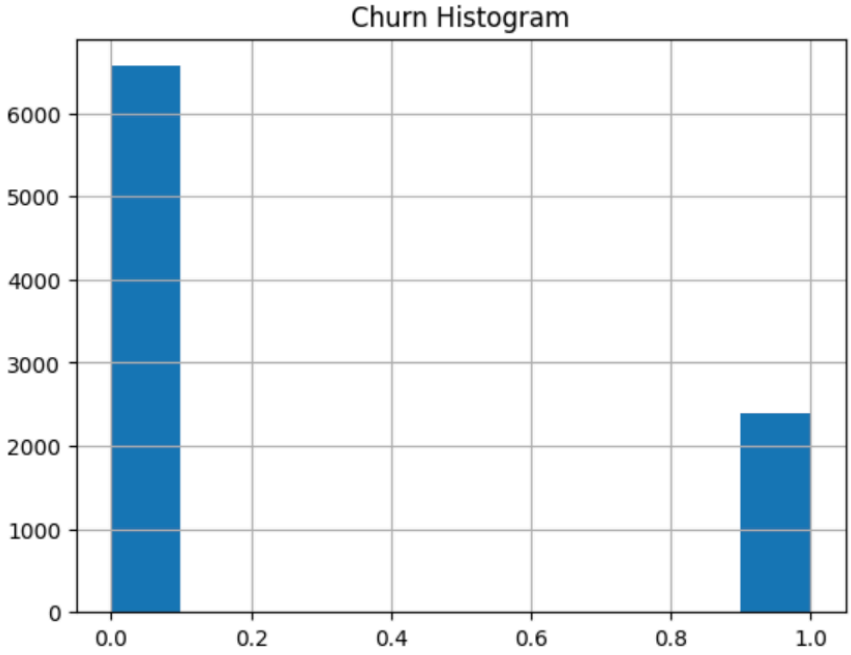

Outage_sec_perweek Histogram
2000
1750
1500
1250
1000
750
500
250
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Population Histogram
5000
4000
3000
2000
1000
0
0
10000
20000
30000
40000
50000

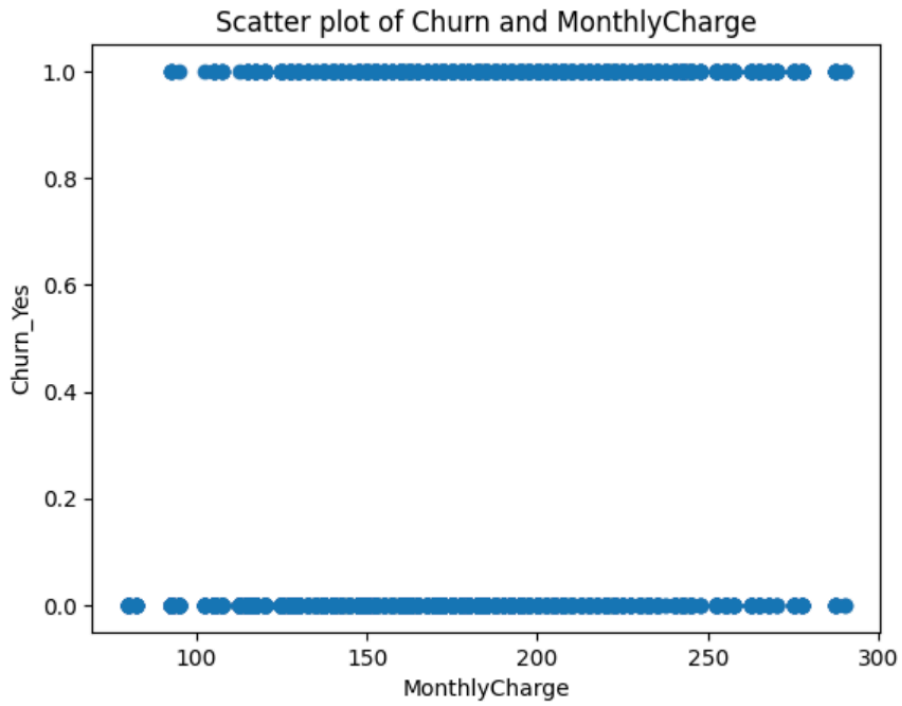

Techie Histogram
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

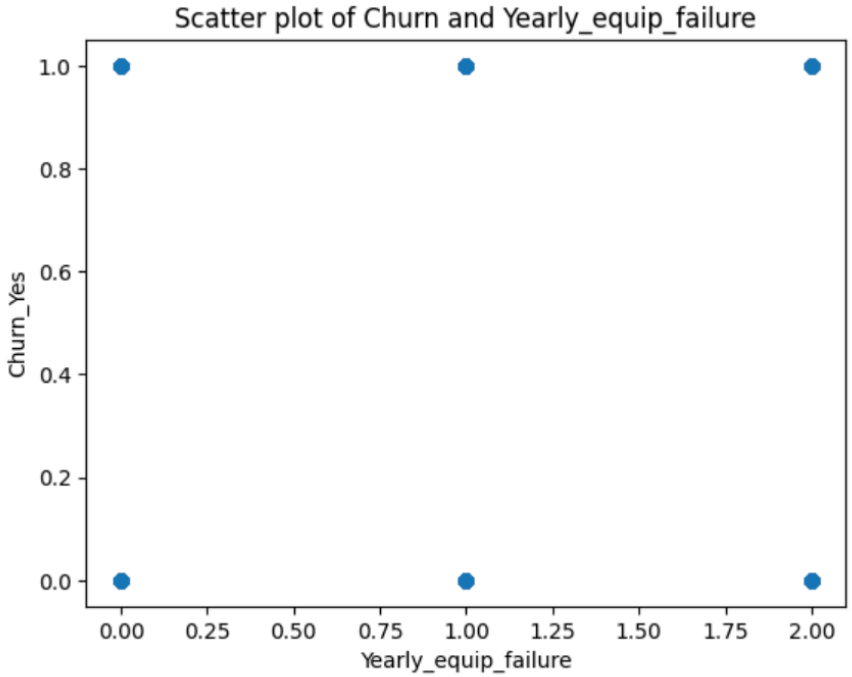

TechSupport Histogram
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

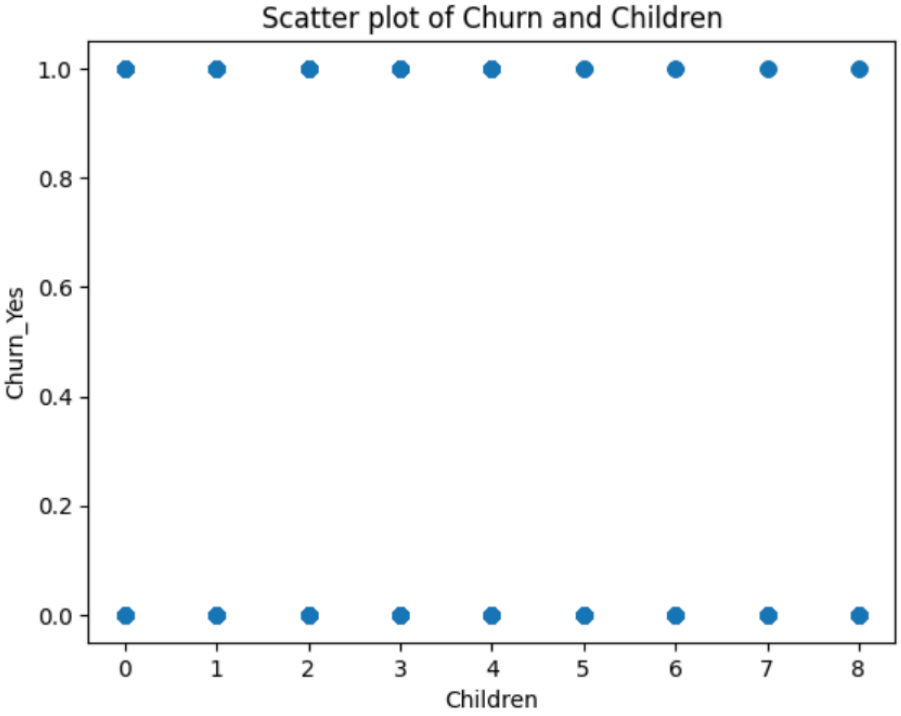

Gender Histogram
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

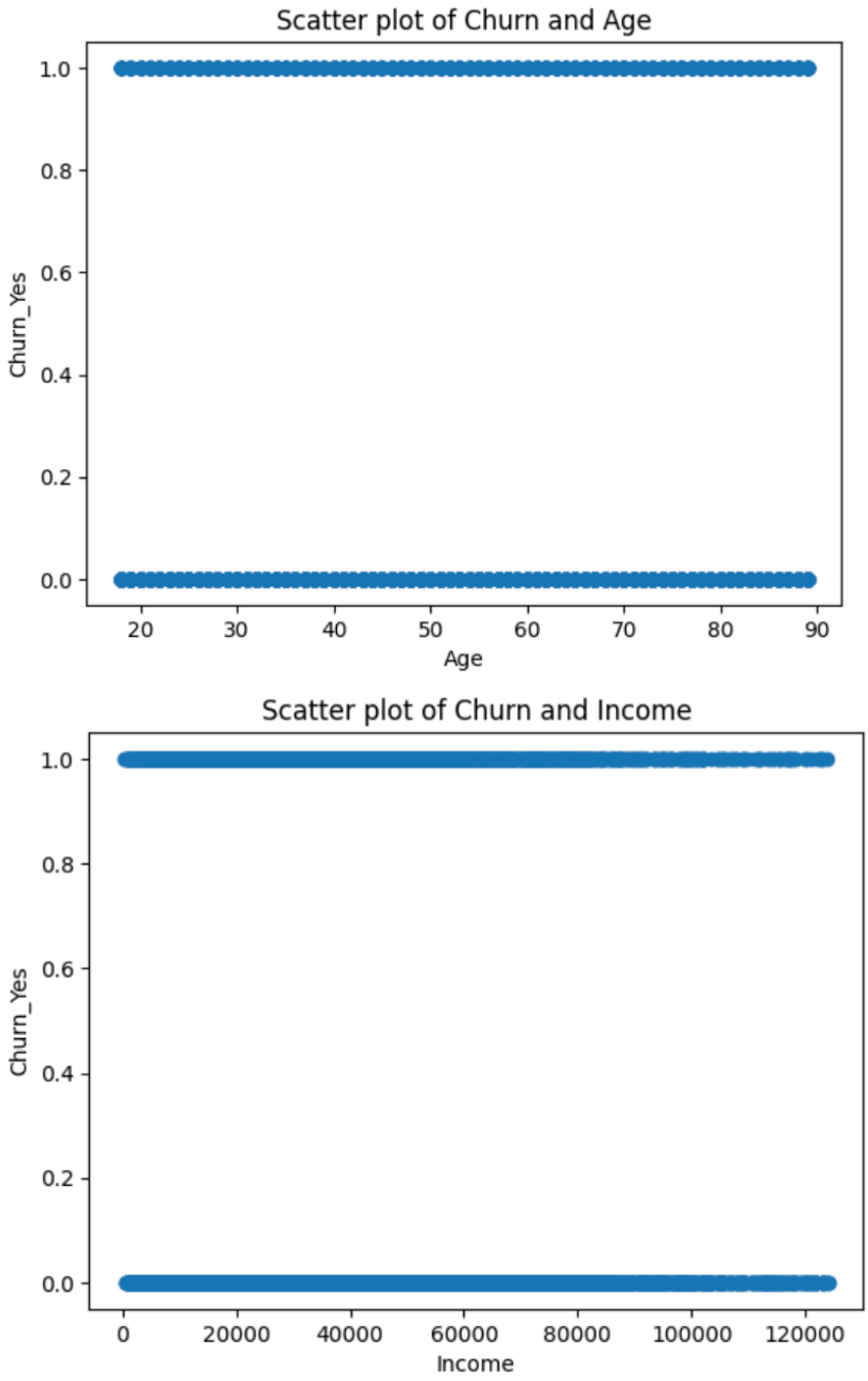

Churn Histogram
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

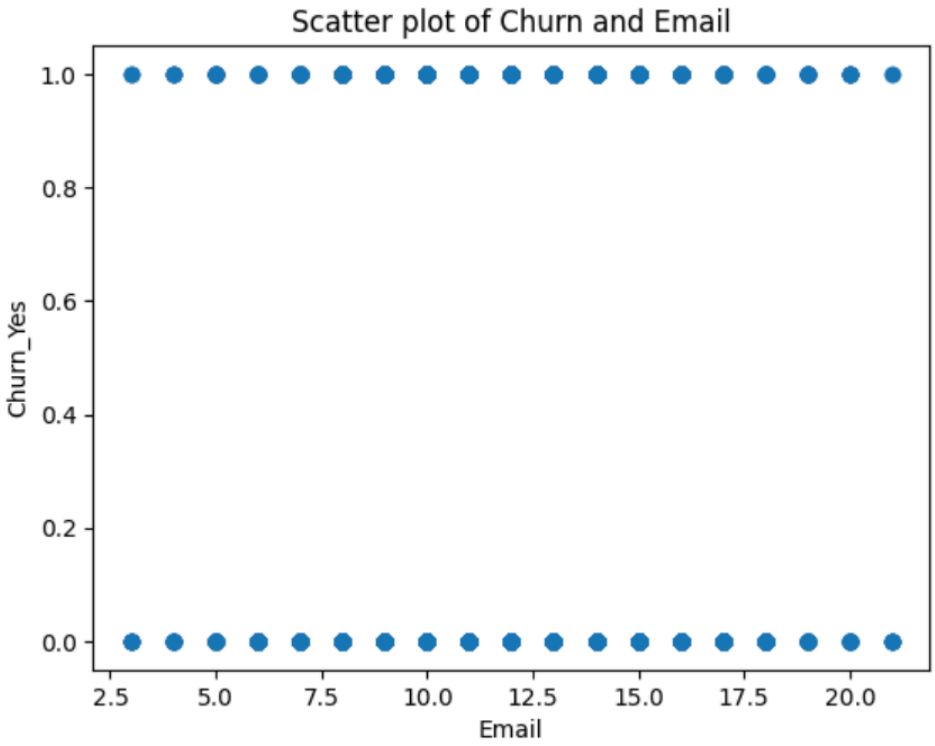
Bivariate statistical analysis refers to the statistical analysis of two variables at once (Bruce et al., 2020). A scatterplot and was created for each independent variable to examine the relationship with the dependent variable ('DummyChurn'). The scatterplots below show that there is no correlation between the 'DummyChurn' and independent variables. A chi-square test was used as a bivariate analysis for the categorical variables: 'DummyGender', 'DummyTechie', and 'DummyTechSupport' against the dependent variable, 'DummyChurn'. There was no correlation between the dependent and independent variables evident.

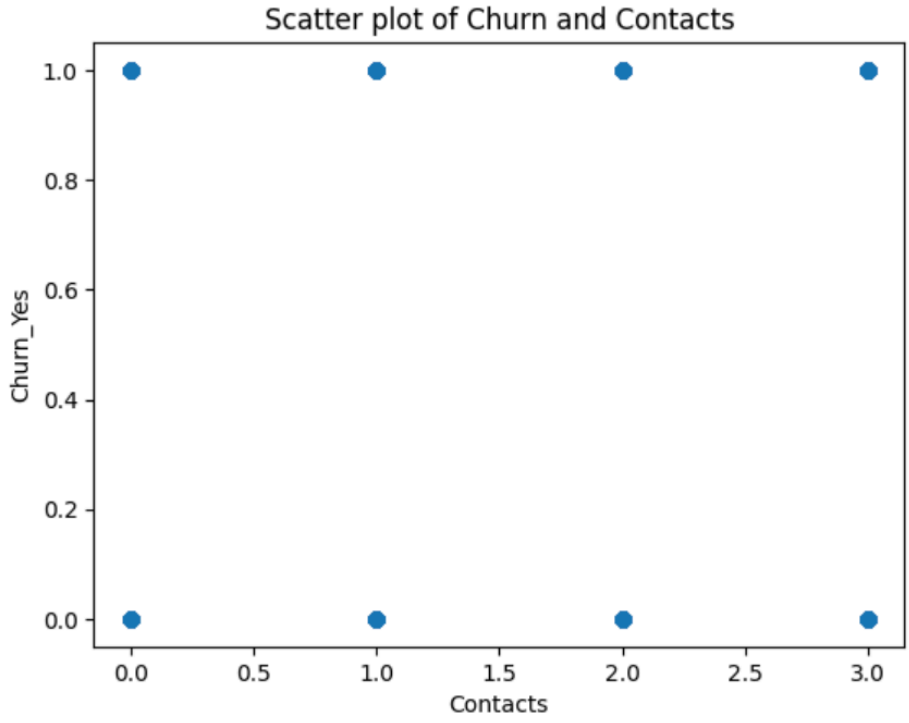


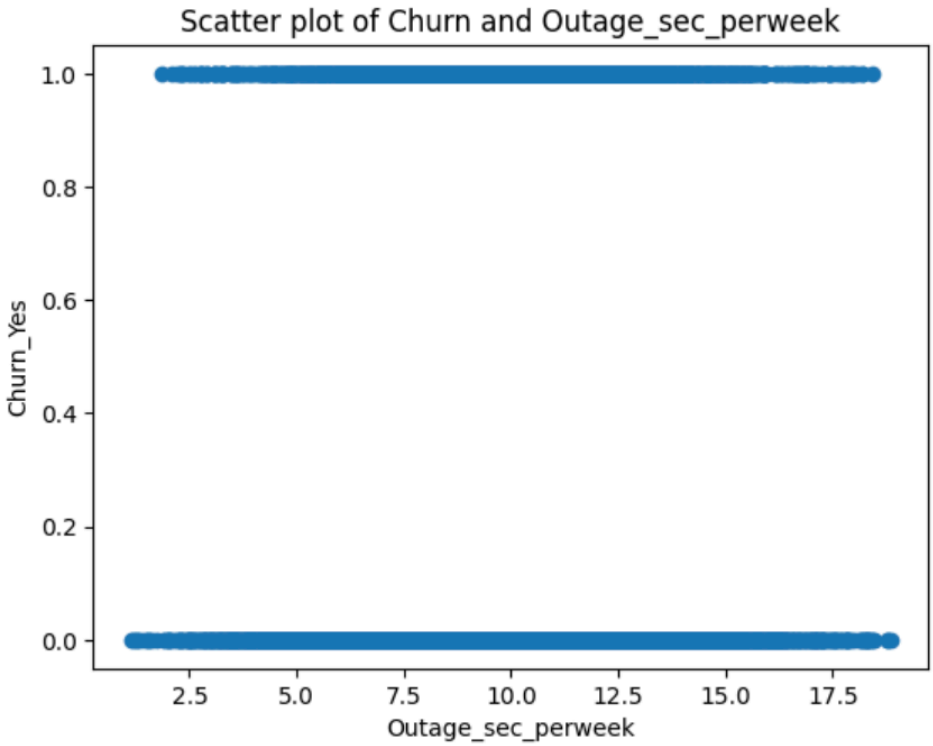

Scatter plot of Churn and Yearly_equip_failure
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Yearly_equip_failure

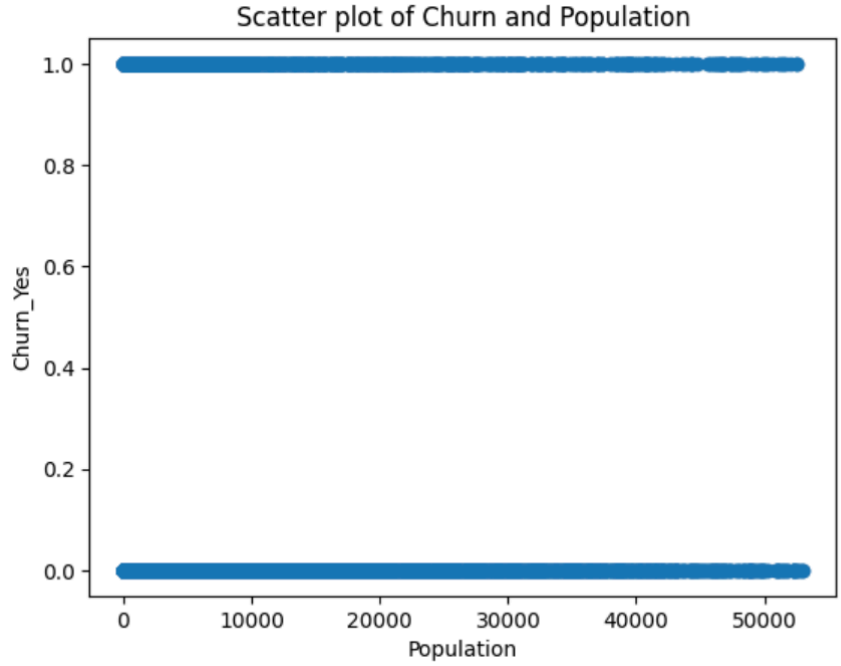

Scatter plot of Churn and Children
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
Children

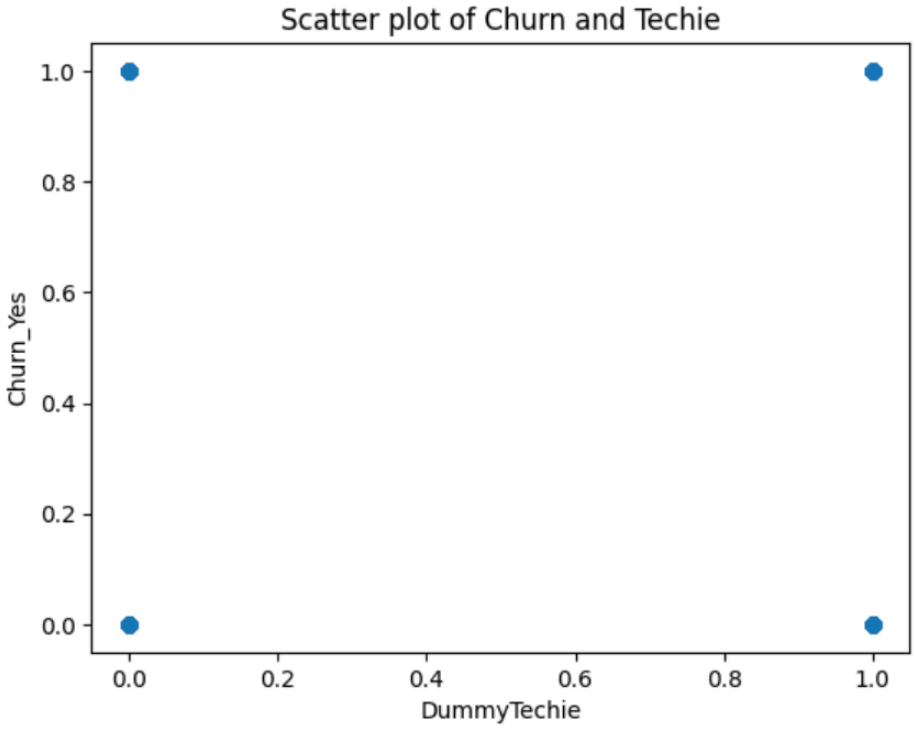

Scatter plot of Churn and Age
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
80
90
Age
Scatter plot of Churn and Income
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
20000
40000
60000
80000
100000
120000
Income

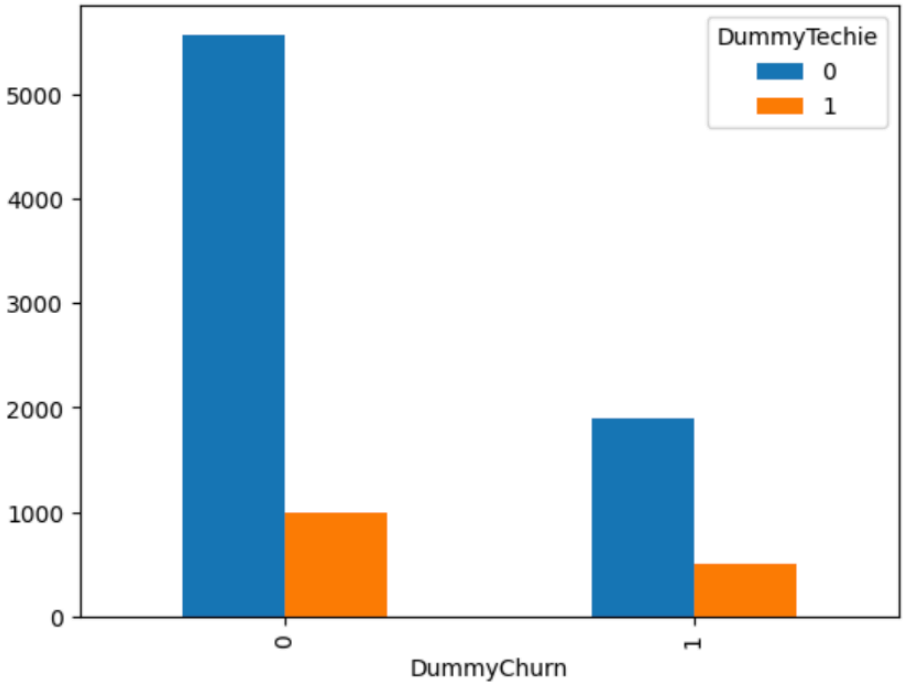

Scatter plot of Churn and Email
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Email

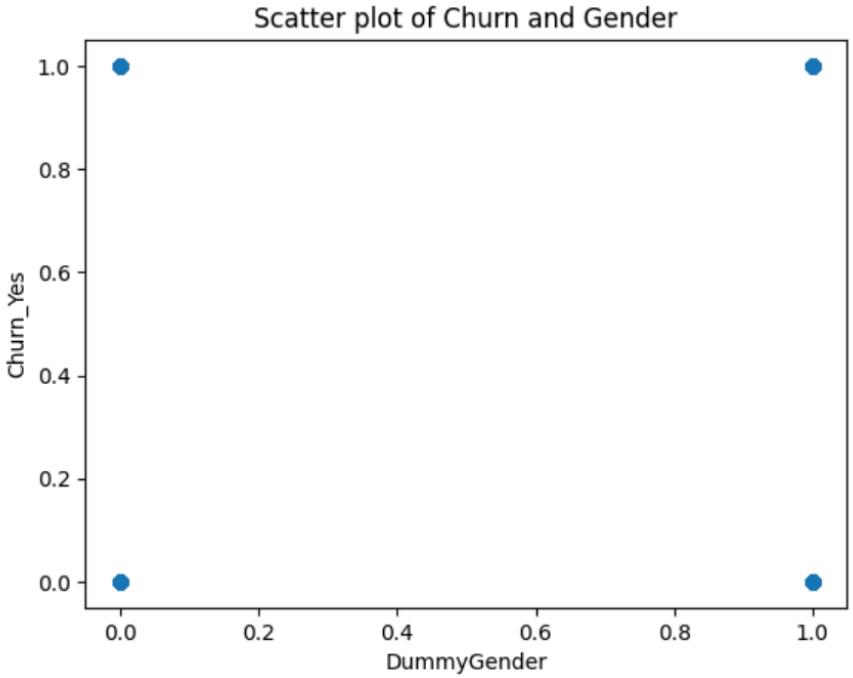

Scatter plot of Churn and Contacts
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Contacts

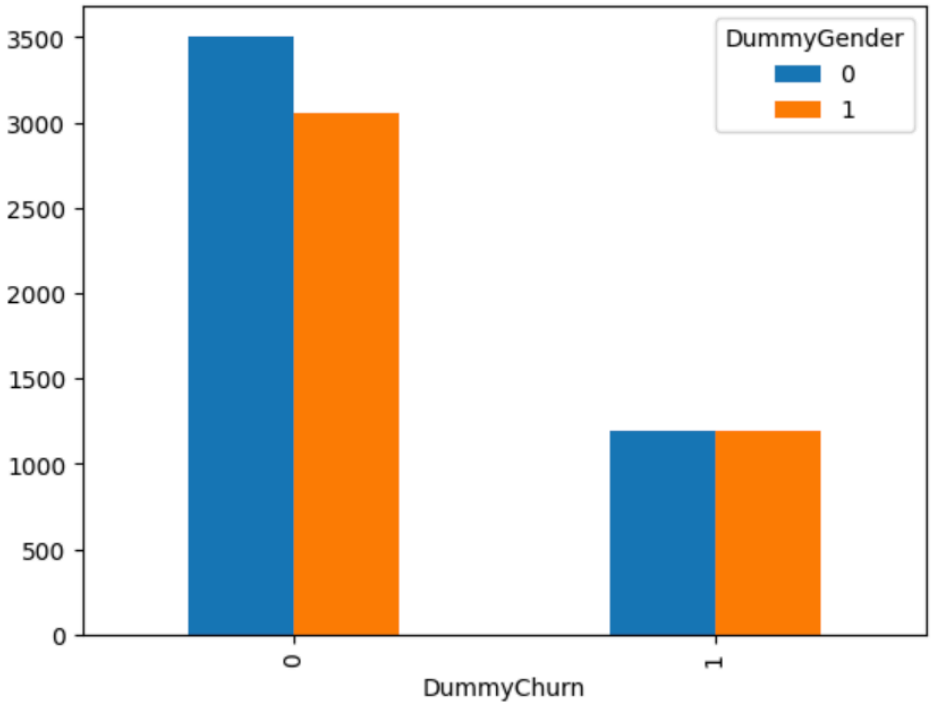

Scatter plot of Churn and Outage_sec_perweek
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Outage_sec_perweek

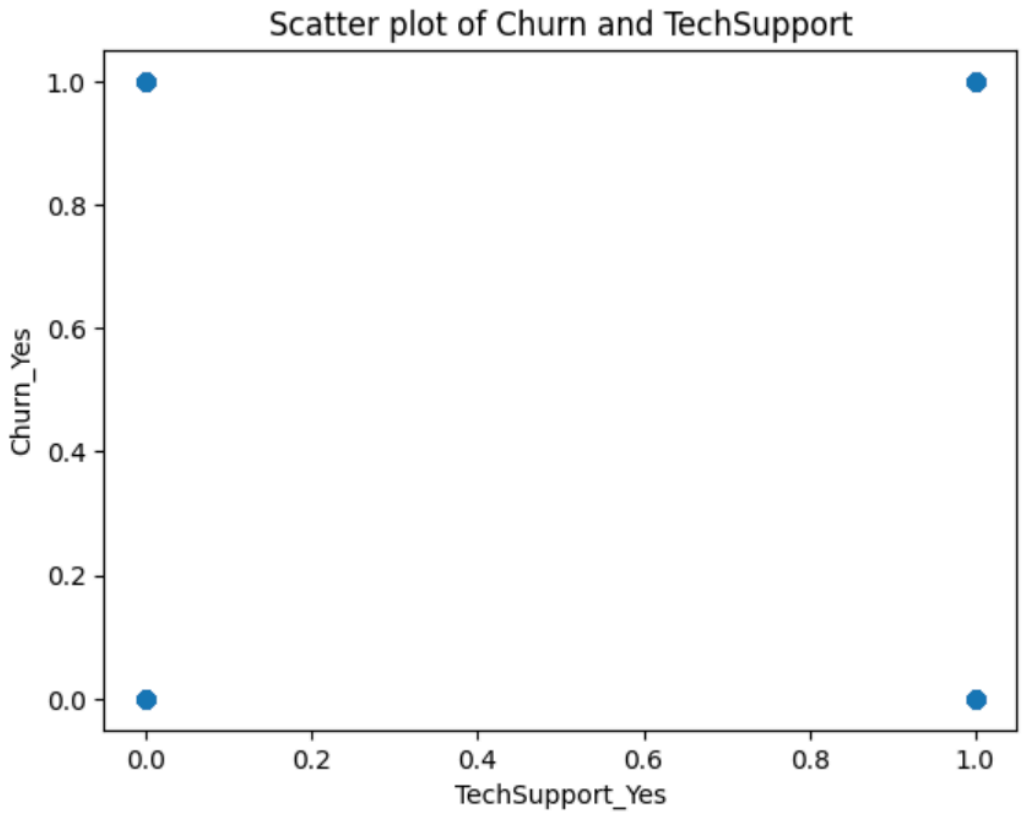

Scatter plot of Churn and Population
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
50000
Population

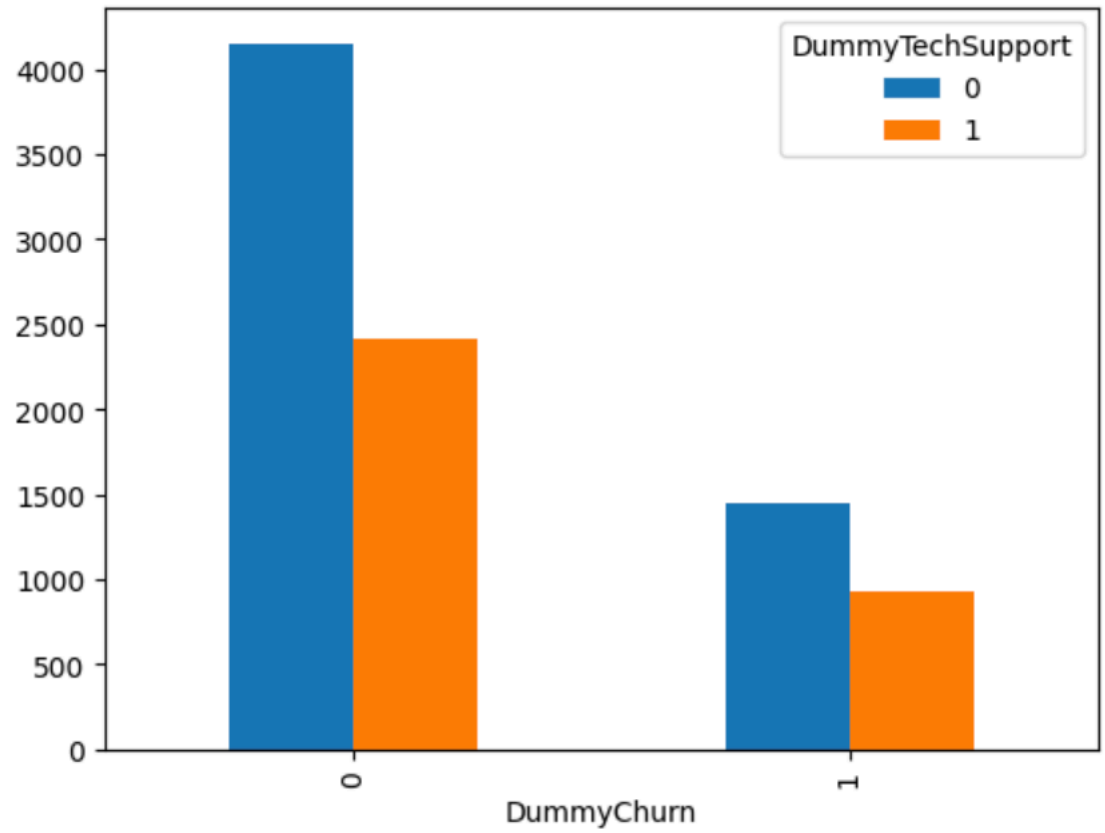

Scatter plot of Churn and Techie
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
DummyTechie


DummyTechie
0
1
5000
4000
3000
2000
1000
0
0
1
DummyChurn


Scatter plot of Churn and Gender
Churn_Yes
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
DummyGender


DummyGender
0
1
3500
3000
2500
2000
1500
1000
500
0
0
1
DummyChurn


Scatter plot of Churn and TechSupport
Churn_Yes
TechSupport_Yes
0.0
0.2
0.4
0.6
0.8
1.0


DummyTechSupport
0
1
DummyChurn
0
500
1000
1500
2000
2500
3000
3500
4000

C4. Data Transformation

To determine which factors impact customer tenure in relation to the dependent variable 'churn' a logistic regression model was created. The regression model measures the independent variable in correlation to the dependent variable to see which has the most impact. Data wrangling was completed to examine the data for further analysis. Data wrangling is the process of cleaning, transforming, and organizing raw data from various sources into a consistent format for analysis (Lacrose & Lacrose, 2019). An important step in the data wrangling process is data transformation. Data transformation is a part of the data cleaning process that involves converting data into a standard format and/or applying calculations to generate new variables (Lacrose & Lacrose, 2019). The data transformation process included first filtering outliers. Z-scores were calculated. The outliers were filtered out based on the z-score criterion.

1. Checking for outliers and removal:

```
print(df.shape)
df = df[(np.abs(stats.zscore(df.select_dtypes(include=np.number))) < 3).all(axis=1)]
print(df.shape)
```

2. Drop the less meaningful columns

```
df = df.drop(columns=['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State', 'County',
'Zip', 'Lat', 'Lng',
                'TimeZone', 'Job','Marital', 'Contract', 'Port_modem', 'Tablet', 'InternetService',
'Phone', 'Multiple',
                'OnlineSecurity', 'OnlineBackup', 'Area','DeviceProtection', 'StreamingTV',
'StreamingMovies', 'PaperlessBilling',
                'PaymentMethod', 'Bandwidth_GB_Year', 'Item1', 'Item2',
                'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8'])
```

3. Display the dimension of dataframe

```
df.shape
```

4. display data set with all the columns

```
df.head()
```

5. Validate there are no nulls

```
df.isnull().sum()
```

To transform categorical variables into numeric variables the encoding process was completed. This can help identify potential issues with the data and evaluate the effectiveness of different transformation methods. Label encoding was applied to the categorical variables. constructing a logistic regression model from all predictors that were identified above. Dummy variables are used to represent categorical variables as numerical values. Categorical variables, such as gender, cannot be included directly in a logistic regression model because they are not numerical. Categorical variables can be converted into dummy variables, which are numerical variables that represent a category (Massaron, 2016). Dummy variables are useful because they allow one to compare the effect of different categories on the probability of a certain outcome.

6. Create dummy variables in order to encode categorical, yes/no data points into 1/0 numerical values.

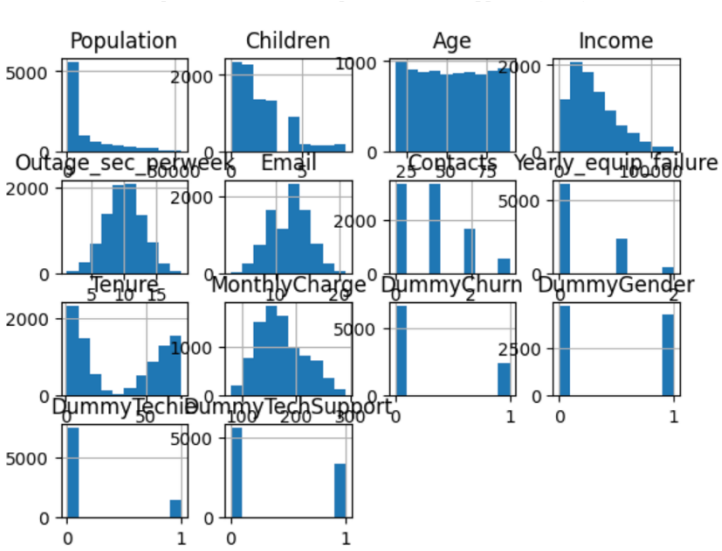
```
df['DummyChurn'] = [1 if v == 'Yes' else 0 for v in df['Churn']]
df['DummyGender'] = [1 if v == 'Male' else 0 for v in df['Gender']]
df['DummyTechie'] = [1 if v == 'Yes' else 0 for v in df['Techie']]
df['DummyTechSupport'] = [1 if v == 'Yes' else 0 for v in df['TechSupport']]
```

7. # Drop original categorical features from dataframe

```
df = df.drop(columns=['Gender', 'Churn', 'Techie','TechSupport'])
```

Histograms are a useful way to understand the distribution of a set of numeric data. A histogram is a graph that displays the frequency of data points in a set of data, divided into intervals. Histograms for each of the columns were visualized to quickly be able to see the range of values that the data set covers and how frequently each value occurs.

8. df.hist()



9. df.columns

```
Index(['Population', 'Children', 'Age', 'Income', 'Outage_sec_perweek',
   'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge',
       'DummyChurn', 'DummyGender', 'DummyTechie', 'DummyTechSupport'],
      dtype='object')
```

To gain further insights into the distribution of the data visualization techniques shown above were used. To complete the data transformation process data profiling using code df.nunique() that involved calculating the number of unique values. By calculating the unique values one can see the diversity and distribution of the data and identify any potential issues that need to be addressed before the data can be used for analysis or modeling (Massaron, 2016).

df.nunique():

```
Population              5414
Children                   9
Age                       72
Income                  8945
Outage_sec_perweek      8940
Email                     19
Contacts                   4
Yearly_equip_failure       3
Tenure                  8948
MonthlyCharge            748
DummyChurn                 2
```

```
DummyGender                2
DummyTechie                2
DummyTechSupport           2
intercept                  1
dtype: int64
```

C5. Prepared Data Set

The new data frame was saved to a new file and attached as a csv file.

```
# Prepared dataset saved to new file
df.to_csv('prepared_churn.csv', index=False)
```

Part IV: Model Comparison and Analysis

D1. Initial Model

A logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and the independent variable(s). Below is the initial logistic regression model utilizing the independent (MonthlyCharge, Tenure, Yearly_equip_failure, Children, Age, Income, Email, Contacts, TechSupport, outage_sec_perweek, Population Gender, Techie) and dependent variable (churn).

Logistic Regression Model (Zach, 2020):

```
df['intercept'] = 1
churn_logit = sm.Logit(df['DummyChurn'], df[['intercept','Population', 'Children', 'Age',
'Income', 'Outage_sec_perweek',
            'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge',
            'DummyGender', 'DummyTechie', 'DummyTechSupport']])
        results = churn_logit.fit()
        results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.337066
         Iterations 8
```

Logit Regression Results

Dep. Variable:	DummyChurn	**No. Observations:**	8950
Model:	Logit	**Df Residuals:**	8936
Method:	MLE	**Df Model:**	13
Date:	Sun, 19 Mar 2023	**Pseudo R-squ.:**	0.4187
Time:	14:05:49	**Log-Likelihood:**	-3016.7
converged:	True	**LL-Null:**	-5189.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-5.5314	0.264	-20.965	0.000	-6.049	-5.014
Population	-1.853e-06	2.76e-06	-0.672	0.501	-7.25e-06	3.55e-06
Children	-0.0046	0.017	-0.266	0.790	-0.038	0.029
Age	0.0020	0.002	1.291	0.197	-0.001	0.005
Income	3.261e-07	1.28e-06	0.254	0.800	-2.19e-06	2.84e-06
Outage_sec_perweek	-0.0027	0.011	-0.247	0.805	-0.024	0.019
Email	0.0082	0.011	0.761	0.447	-0.013	0.030
Contacts	0.0293	0.036	0.818	0.413	-0.041	0.099
Yearly_equip_failure	-0.0094	0.055	-0.170	0.865	-0.118	0.099
Tenure	-0.0755	0.002	-39.574	0.000	-0.079	-0.072
MonthlyCharge	0.0338	0.001	35.176	0.000	0.032	0.036
DummyGender	0.1914	0.065	2.960	0.003	0.065	0.318
DummyTechie	0.5733	0.084	6.847	0.000	0.409	0.737
DummyTechSupport	-0.2087	0.067	-3.118	0.002	-0.340	-0.078

variance_inflation_factor:

```
# Define the independent and dependent variables
X = df[['Population', 'Children', 'Age', 'Income', 'Outage_sec_perweek',
       'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge',
       'DummyGender', 'DummyTechie', 'DummyTechSupport']]
y = df['DummyChurn']

# Fit a logistic regression model
model = sm.Logit(y, X).fit()

# Calculate VIFs for each independent variable
```

```
vifs = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

# Print the VIFs for each variable
for i, col in enumerate(X.columns):
    print(f'{col}: {vifs[i]:.2f}')

# Add a constant term to the independent variables
X = sm.add_constant(X)
```

```
Optimization terminated successfully.
         Current function value: 0.364938
         Iterations 7
Population: 1.52
Children: 2.02
Age: 6.81
Income: 3.18
Outage_sec_perweek: 10.31
Email: 12.38
Contacts: 2.06
Yearly_equip_failure: 1.40
Tenure: 2.61
MonthlyCharge: 13.09
DummyGender: 1.89
DummyTechie: 1.20
DummyTechSupport: 1.62
```

D2. Justification of Model Reduction

To answer the research question what factors contribute to the churn rate? model evaluation metrics can be used to justify a reduced initial model and improve the accuracy in predicting the dependent variable. One possible selection procedure is the backward elimination method. This involves removing independent variables from the model one at a time based on their statical significance until only statically significant variables remain (Massaron, 2016). This approach is justified since the goal is to identify the most important independent variables while minimizing the number of irrelevant variables. Based on this initial model summary some of the predictor variables are not significantly associated with the outcome variable. Some of the independent variables have a p-value greater than 0.05 indicating no significance. These included "Population," "Children," "Age," "Income," "Outage_sec_perweek," "Email," "Contacts," and "Yearly_equip_failure.". These independent variables were removed from the model to obtain a reduced model.

To check for multicollinearity among the independent variables the variance inflation factor for each variable can be calculated (Massaron, 2016). The VIF measures how much the variance of the estimated regression coefficient is increased due to collinearity. A VIF greater than 5 or 10 indicates that the variable is highly collinear with other variables and may need to be removed from the model (Massaron, 2016). After running the initial model "Outage_sec_perweek (10.31) and 'MonthlyCharge' (13.09) had extremely high VIFs which could indicate high collinearity and thus were in the initial variable to be removed from the logistic regression model.

D3. Reduced Logistic Regression Model

Based on the VIF results above, there is high multicollinearity among the variables in the model. To address this issue removing one or more of the highly correlated variables can be done. Also removing variables with high p-values can increase fit of model. The variables "Population," "Children," "Age," "Income," "Outage_sec_perweek," "Email," "Contacts," and "Yearly_equip_failure" and 'MonthlyCharge' were removed to form the reduced model.

Reduced Logistic Regression Model:

```
df['intercept'] = 1
churn_logit = sm.Logit(df['DummyChurn'], df[['intercept', 'Tenure',
      'DummyGender', 'DummyTechie', 'DummyTechSupport']])
results = churn_logit.fit()
results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.440288
         Iterations 7
```

Logit Regression Results

Dep. Variable:	DummyChurn	**No. Observations:**	8950
Model:	Logit	**Df Residuals:**	8945
Method:	MLE	**Df Model:**	4
Date:	Sun, 19 Mar 2023	**Pseudo R-squ.:**	0.2407
Time:	21:50:02	**Log-Likelihood:**	-3940.6
converged:	True	**LL-Null:**	-5189.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.2229	0.052	4.309	0.000	0.122	0.324
Tenure	-0.0563	0.001	-38.731	0.000	-0.059	-0.053
DummyGender	0.1515	0.055	2.738	0.006	0.043	0.260
DummyTechie	0.4430	0.072	6.129	0.000	0.301	0.585
DummyTechSupport	0.1139	0.057	1.999	0.046	0.002	0.226

E1. Model Comparison

The initial regression model and reduced regression model can be compared for a full analysis of the independent variables on the dependent variable. To answer the research question what factors contribute to the churn rate? model evaluation metrics can be used to justify a reduced initial model and improve the accuracy in predicting the dependent variable. In examining the initial model there are several pieces of information that can help evaluate the model's performance. The interaction value indicates how many iterations the optimization algorithm took to converge (Zach, 2020). This initial model converged after 8 interactions, which is good. The reduced model had 7 interactions.

The initial model included 13 independent variables while the reduced model only included 4 independent variables. The initial model had a higher pseudo R-squared value of 0.4187 compared to the reduced model which had a pseduo r-squared of 0.2407. Both models had a significant likelihood rate test (LLR p-value<0.05). The initial model may indicate more variability. Both modeled have p-vales less than 0.05, indicating that there are statistically significant independent variables in predicting the dependent variable. the reduced model may be preferred over the initial model due to its simplicity and only included the most significant

independent variables. Here is a summary of the coefficients and p-values in the initial model vs the reduced model:

Initial Model		
Variable	**Coefficient**	**P-value**
Intercept	-5.5314	0
Population	-1.85E-06	0.501
Children	-0.0046	0.79
Age	0.002	0.197
Income	3.26E-07	0.8
Outage_sec_perweek	-0.0027	0.805
Email	0.0082	0.447
Contacts	0.0293	0.413
Yearly_equip_failure	-0.0094	0.865
Tenure	-0.0755	0
MonthlyCharge	0.0338	0
DummyGender	0.1914	0.003
DummyTechie	0.5733	0
DummyTechSupport	-0.2087	0.002

Reduced Model		
Variable	**Coefficient**	**P-value**
Intercept	0.2229	0
Tenure	-0.0563	0
DummyGender	0.1515	0.006
DummyTechie	0.443	0
DummyTechSupport	0.1139	0.046

All coefficients in the reduced model are statistically significant ($p<0.05$). For 'tenure' the coefficient for tenure changed from -0.0755 in the initial model to 0.0563 in the reduced model. This shows the effect of tenure on churn is still negative but the effect is not as strong as in the reduced model as it is in the initial model. Gender and Techie have smaller coefficients int he reduced model compared to the initial model. This suggests that the impact of these two variables on the probability of churn has decreased in the reduced model.

Akaike Information Criterion (AIC) is a model evaluation metric used to measure the relative quality of a statical model while taking into account both the goodness of fit and the complexity of the model (Massaron, 2016). A lower AIC value indicates a better model fit with a

larger difference in AIC values indicating a greater difference in model quality. This can be used to compare models to select the best model.

AIC for initial model:

```
# Fit the initial model
df['intercept'] = 1
initial_model = sm.Logit(df['DummyChurn'], df[['intercept', 'Population', 'Children', 'Age',
'Income', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure',
'MonthlyCharge', 'DummyGender', 'DummyTechie', 'DummyTechSupport']])
initial_results = initial_model.fit()

# Calculate the AIC
initial_aic = initial_results.aic
print("Initial model AIC:", initial_aic)
```

```
Initial model AIC: 6061.485297988196
```

AIC for reduced model:

```
# Define the reduced model
X_reduced = df[['Tenure', 'DummyGender', 'DummyTechie', 'DummyTechSupport']]
y = df['DummyChurn']

# Fit the reduced model
X_reduced = sm.add_constant(X_reduced)
model_reduced = sm.Logit(y, X_reduced).fit()

# Compute the AIC for the reduced model
AIC_reduced = model_reduced.aic
print("AIC for the reduced model:", AIC_reduced)
```

```
AIC for the reduced model: 7891.159598437633
```

The AIC measures the goodness of fit of the logistic model. The initial model has an AIC of 6061.49 and the reduced model has an AIC of 7891.16. This difference indicates the initial model was a better fit for the data than the reduced model due to the lower AIC. However, the models have a different number of variables so the comparison should be interpreted with caution.

E2. Output and Calculation

Cross-validation was used to validate the performance of the model. Cross-validation helps to reduce the risk of overfitting by providing a more reliable estimate of a model's performance on the data. The RMSE below indicates that the difference between the predicted and actual values in the dataset is very small leading to a better-fit model. The standard deviation of the RMSE is also small and the model has a consistent level of accuracy across the dataset. The cross-validation scores below are the accuracy scores obtained for each fold of the cross-validation along with the mean scores (GeeksforGeeks, 2023). The cross-validation scores indicate how well the logistic model performed. The mean score is the average accuracy score across the folds. The higher the mean score, the better the model is expected to perform with new data.

Cross-validation code (GeeksforGeeks, 2023):

```
# Define the logistic regression model
logreg = LogisticRegression()

# Define the independent and dependent variables
X = df[[ 'Tenure', 'DummyGender', 'DummyTechie', 'DummyTechSupport']]
y = df['DummyChurn']

# Perform cross-validation on the logistic regression model
scores = cross_val_score(logreg, X, y, cv=5)

# Print the cross-validation scores
print("Cross-validation scores:", scores)
print("Mean score:", scores.mean())
```

```
Cross-validation scores: [0.39497207 0.54972067 0.77374302 0.78044693 0.78
547486]
Mean score: 0.6568715083798884
```

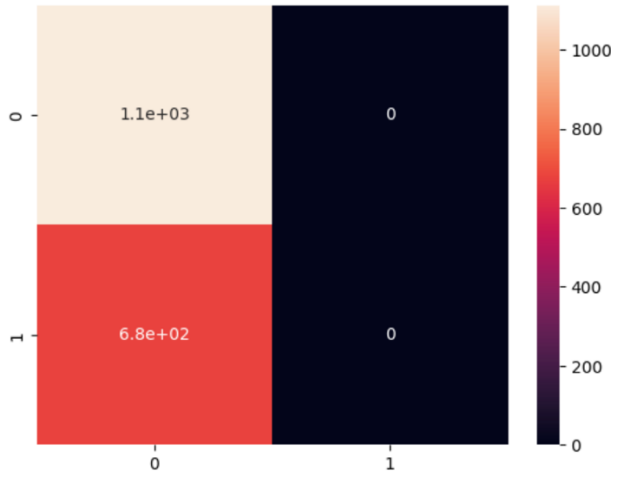
The cross-validation scores range from 0.39 to 0.79 indicating varying degrees of accuracy across the data set. The mean score is 0.66 and provides an estimate of the model's performance.

A model evaluation metric used to evaluate the performance of binary classifiers in the precision, recall, and F1-scores. These can be found in the classification report. These metrics can be used to assess the model's ability to correctly classify instances of each class. These can be calculated after obtaining the confusion matrix.

```
# Import the prepared dataset
matrix_df = pd.read_csv('prepared_churn.csv')
X = matrix_df.iloc[:, 1:-1].values
y = matrix_df.iloc[:, -1].values

# Split the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

# Training the Logistic Regression model on the Training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)


# Predict the Test set results
y_pred = classifier.predict(X_test)

# Display the Confusion Matrix
from sklearn.metrics import confusion_matrix
matrix = confusion_matrix(y_test, y_pred)
print(matrix)
```

```
[[1113    0]
 [ 677    0]]
```

```
heatmap of confusion maxtix
y_predict_test = classifier.predict(X_test)
matrix_new = confusion_matrix(y_test, y_predict_test)
sns.heatmap(matrix_new, annot=True)
```



classification_report:
from sklearn.metrics import classification_report
print(classification_report(y_test, y_predict_test))

	precision	recall	f1-score	support
0	0.62	1.00	0.77	1113
1	0.00	0.00	0.00	677
accuracy			0.62	1790
macro avg	0.31	0.50	0.38	1790
weighted avg	0.39	0.62	0.48	1790

To calculate the accuracy, precision, recall, and F1-score of a model, you can use the following formulas (output shown above in the classification report):

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1-score = 2 * (Precision * Recall) / (Precision + Recall)
(TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives)

Using the confusion matrix [[1113, 0], [677, 0]], the performance metrics can be calculated
TP = 1113
TN = 0
FP = 0
FN = 677
Accuracy = (1113 + 0) / (1113 + 0 + 0 + 677) = 0.6216
Precision = 1113 / (1113 + 0) = 1.0000
Recall = 1113 / (1113 + 677) = 0.6216
F1-score = 2 * (1.0000 * 0.6216) / (1.0000 + 0.6216) = 0.7659

The accuracy of the model is 0.6216, the precision is 1.0000, the recall is 0.6216, and the F1-score is 0.7659.The accuracy of the model at 0.6216 means that 62.16% of the predictions from the model may be correct. This can be misleading if the dataset is imbalanced or if the false positives and false negatives are different (Korstanje, 2022). The precision of the reduced model is 1.0000. This means the predictions made by the model are correct. However, this does not mean the model is performing adequately and could be a prediction for the majority class which is negative instead of the positive class (Korstanje, 2022). The recall of the reduced model is 0.6216 meaning that model is able to identify 62.16% of the actual positive instances in the dataset. The F1-score is the harmonic mean between the precision and the recall ranging from 0 to 1 where 1 is the best (Korstanje, 2022). The F1-scores was 0.7659 suggesting the model is performing reasonably.

E3. Code

An error-free copy of the code used to support the implementation of the logistic regression model using python is attached.

Part V: Data Summary and Implications

F1. Results

a. A regression equation for the reduced model:

A regression equation is a mathematical equation that expresses the relationship between a dependent variable and the independent variables(s). The logistic regression model assumes a linear relationship between the independent variables and the log-odds of the dependent variable. The equation below is a regression equation for the reduced model for the dependent variable 'churn' and the independent variables 'gender', 'techie', and 'techsupport'.

The formula for a logistic regression model can be expressed as (Zach, 2020):

- $\text{Log(odds of churn)} = \beta0 + \beta1X1 + \beta2X2 + ... + \beta kXk$

- log-odd of churn: the natural logarithm of the odds of churn which is the ratio of the probability of churn to the probability of not churn, given the independent variables
- β0: the intercept
- β1, β2, ... βk, etc: are the coefficients for the predictor variables X1, X2, ..., Xk

The logistic function is then applied to the linear predictor variables to obtain the predicted probability of churn rates which is expressed as:

- p = 1 / (1 + e^-(β0 + β1X1 + β2X2 + ... + βkXk))
- p is the predicted probability of churn
- e is the base of the natural logarithm

Regression Equation for reduced model (Zach, 2020):

churn = 0.2229 - 0.0563 * Tenure + 0.1515 * DummyGender + 0.4430 * DummyTechie + 0.1139 * DummyTechSupport

The intercept in 0.2229 which represents the log odds of churn when all the other predictors are equal to zero. Thus, the intercept represents the log odds of churn for a customer with zero tenure, female gender, non-techie, and no tech support. The model predicts that for each one unit increase in tenure, the log odds of churn will decrease by 0.0563. Thus higher tenure is associated with lower churn rates. For each one unit increase in gender, the log odds of churn will increase by 0.1515. This means that females are associated with higher churn. For each one unit increase techie the log-odd of churn will increase by 0.443. For each one unit increase in tech support, the log-odd of churn will increase by 0.1139 so the log odds of churn is 0.1139 higher than for customers who do not use tech support.

b. An interpretation of the coefficients of the reduced model:

The coefficients represent the change in the dependent variable per unit increase of each of the independent variables while holding all other variables constant. The coefficient can be

used to make predictions on new data using the same model. The intercept coefficient is 0.2229 and indicates that if all the independent variables are set to zero, the log-odd of the response variable being equal to 1 (the probability of churn) would be 0.2229.

The coefficient for 'tenure' is -0.0563 indicating that a one-unit increase in 'tenure' is associated with a decrease in the log-odds of churn by 0.0563. For example, when a customer's tenure increases by one month the churn decreases. This suggests that customers who have been with the company for a longer time are less likely to churn.

The coefficient for gender is 0.1515 indicating that a one-unit increase in 'gender' is associated with an increase in the log odds of churn by 0.1515. This indicates male customers are associated with an increase in the log-odds of churn by 0.1515 if holding the other variables constant. This suggests that male customers are more likely to churn compared to female customers.

The coefficient 'techie' is 0.4430, signifying that a rise of one unit in 'techie' is accompanied by a log-odds of churn increase of 0.4430. This could imply that a client who perceives themselves as a tech enthusiast is linked to a log-odds of churn increase of 0.4430. As an illustration, individuals who consider themselves knowledgeable about technology may have a higher likelihood of churning than those who do not consider themselves as such.

The coefficient for 'techsupport' is 0.1139, indicating that if 'techsupport' increases by one unit, the log-odds of churn also increase by 0.1139. This suggests that acquiring tech support add-ons is linked to a rise in the log-odds of churn by 0.1139. It may imply that consumers who have tech support add-ons are more prone to churn than those who do not have these add-ons.

c. The statistical and practical significance of the reduced model:

The statistical significance of the reduced logistic model is indicated by the p-values of the coefficients showing a statistical significance indicated by $p < 0.05$. This means that the independent variables are statistically significant in predicting the probability of the churn rate.

The intercept also has p-value of less than 0.05. This means that overall the logistic regression model is statically significant. The pseudo r-squared value is 0.2407 means the model explains 24.07% of the variance in the outcome variable. The LLR p-value is 0.000 is also less than 0.005 indicating the model is statistically significant. The variable for tech support had a p-value of 0.046 which is close to 0.05 and shows minimal significance. The log-likelihood is -3940.6. The log-likelihood measures the goodness of fit of the model to the date. The higher log-likelihood values the better fit of the model to the data. The LL-null value is -5189.6. This is lower than the log-likelihood indicating the model is an improvement of the null model.

The reduced model has four independent variables where the coefficient and p-values indicate that all four variables are statistically significant and have an impact on the probability of churn. The negative coefficient for tenure suggests that customers who have been with a telecommunication service for a longer time are less likely to churn if all other factors remain the same. The positive coefficient for gender, tech, and tech support suggests these factors increased the likelihood of churn. Thus, the reduced model is practically significant as it provides insights into factors associated with customer churn and can be used to make informed business decisions such as improving tech support to decrease the likelihood of churn.

d. The limitations of the data analysis:

There are several limitations to the reduced logistic model. One limitation of this model is the small data size of the churn data set. A larger data set could provide more data and a better-fit model. The limited number of independent variables used in the analysis may not fully capture all the favors that influence customer churn. This analysis only indicates correlations and not causation between the dependent and independent variables. It is possible that other factors not included in the analysis may be responsible for churn rates. More investigation is required.

F2. Recommendations

A recommended course of action for the business would be to focus on retention strategies for customers with low tenure. This could include promotions to encourage customers

to stay with the company. There should be a further investigation regarding the reasons why gender, techie, and tech support are associated with higher churn rates and address underlying issues. One way to do this would be for those customers with technical issues who are more likely to churn. The business could improve the quality and accessibility of technical support to possibly reduce churn.

Part VI: Demonstration

G. Panopto Demonstration

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d66e09df-d7c5-4929-9249-afcc01749d75

H. Sources of Third-Party Code

GeeksForGeeks. (Jan, 2023). ML | Logistic Regression using Python https://www.geeksforgeeks.org/ml-logistic-regression-using-python/

Korstanje, J. (2022, October 11). *The F1 score | Towards Data Science*. Medium. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Zach. (Oct, 2020). How to Perform Logistic Regression in Python (Step-by-Step) https://www.statology.org/logistic-regression-python/

I. Sources

Daniel T. Larose, & Chantal D. Larose. (2019). Data Science Using Python and R. Wiley.

Massaron, L. (2016). *Regression analysis with python: Learn the art of regression analysis with python*. Packt Publishing.

Zach.(October, 2020). The 6 Assumptions of Logistic Regression. https://www.statology.org/assumptions-of-logistic-regression/