NBM3 TASK 2 LOGISTIC REGRESSION MODELING

Table of Contents

Part I. Research Question

A1. Research Question: What are the key factors that influence customer churn in the telecommunications industry, and how can we predict whether a customer will churn within the next month?

Relevance: The telecommunications industry depends on keeping its customers for business because the cost of acquiring new ones is tremendous. With the insights of prediction built into customer churn, companies are helped to put plan in action to retain every last customer instead of losing them and reduce rates at which customers are interested in other products that competitors have compared to them. This question is directly linked to the company's goal of reducing churn and improving overall customer retention.

A2. Define the Goals of the Data Analysis

Goals:

1. **Identify Key Factors:**
    - Determine which variables/features such as Tenure, MonthlyCharge, etc... are most strongly associated with customer churn.
    - Assess the importance and impact of these factors on churn.
2. **Develop a Predictive Model:**
    - Use logistic regression to create a model that predicts the likelihood of a customer churning based on their features or variables.
    - Train the model using data to learn patterns associated with churn.
3. **Evaluate Model Performance:**
    - Assess the accuracy, precision, recall, and F1 score of the logistic regression model to ensure it reliably predicts churn.
    - Use a confusion matrix to analyze the model's performance and identify any areas for improvement.
4. **Implement Predictive Insights:**
    - Apply the model to new customer data to predict churn probability.
    - Provide actionable insights and recommendations for customer retention strategies based on the model's predictions.
5. **Informed Decision-Making:**
    - Enable the telecommunications company to make informed decisions on where to focus retention efforts.

- Prioritize customers who are at high risk of churning for targeted interventions, such as personalized offers or enhanced support.

Part II: Method Justification

B1. Four Assumptions of a Logistic Regression Model

1. **Binary Dependent Variable:**
   - The dependent variable should be binary. In this case, the dependent variable is Churn, which has two possible outcomes: yes (the customer churned) or no (the customer did not churn).
2. **Independence of Observations:**
   - The observations should be independent of each other. This means that the churn status of one customer should not influence the churn status of another customer.
3. **Linearity of Independent Variables and Log Odds:**
   - There should be a linear relationship between the independent variables and the log odds of the dependent variable. This means that the effect of each predictor variable on the log odds of the outcome should be linear.
4. **No Perfect Multicollinearity:**
   - The independent variables should not be perfectly correlated with each other. Perfect multicollinearity can make it difficult to determine the effect of each independent variable.

B2. Two Benefits of Using Python

1. **Extensive and powerful Libraries:**
   - Python has powerful libraries like pandas for data manipulation, numpy for numerical operations, scikit-learn for machine learning, and statsmodels for statistical analysis. These libraries provide a wide range of functions and methods to handle data cleaning, transformation, modeling, and evaluation efficiently.
2. **Ease of Visualization:**
   - Python offers libraries such as matplotlib, seaborn, and plotly that allow for the creation of insightful plots and charts. These visualizations help in understanding the data distribution, identifying patterns, and presenting the results of the analysis in a clear and interpretable manner.

B3. Why Logistic Regression is an Appropriate Technique to Analyze the Research Question

The research question focuses on predicting customer churn, which is a binary outcome (churn or no churn). Logistic regression is specifically designed to handle binary dependent variables.

Part III: Data Preparation

C1. Data Cleaning Goals and Steps

Cleaning the raw churn dataset is essential to ensure accurate and reliable analysis. Proper data cleaning facilitates better sorting, filtering, and modification of the dataset maintaining its integrity and improving the quality of the results. Here is how the data-cleaning process is carried out:

Step 1: Importing and Initializing Data

The dataset was imported into Python using the pandas library with the command import pandas as pd. The dataset was loaded into a DataFrame using df = pd.read_csv('churn_clean.csv'). Initial inspection of the data types and structure was done using df.info() to understand the variables and their data types, as well as to identify the presence of any non-null values.

Step 2: Identifying and Handling Duplicates

To detect duplicate entries, the function df.duplicated() was utilized. This function returns TRUE for duplicate rows and FALSE otherwise. The results showed no duplicate entries, confirmed by the count of FALSE values using print(df.duplicated().value_counts()).

Step 3: Handling Missing Values

Missing values in the dataset were identified using df.isnull().sum(). There is no missing value in the dataset.

Step 5: Removing the Irrelevant Columns

Columns deemed irrelevant such as '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', were dropped from the dataset using df.drop().

C2. Summary Statistics

Below is a summary of the statistics for both the dependent variable(Churn) and the independent variables that are 'Population', 'Children', 'Age', 'Income', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge', Gender', 'Techie', and 'TechSupport'. Understanding these statistics is crucial when running a logistic regression model, as it provides insights into the relationships between the variables.

The summary statistics for numerical independent variables were obtained using the code df.describe(), and for categorical variables, the loop iterates over each categorical variable and prints the frequency counts and percentages of each category. The results are shown below.

```
Summary Statistics for Numerical Variables:
          Population    Children           Age         Income  \
count   10000.000000  10000.0000  10000.000000   10000.000000
mean     9756.562400      2.0877     53.078400   39806.926771
std     14432.698671      2.1472     20.698882   28199.916702
min         0.000000      0.0000     18.000000     348.670000
25%       738.000000      0.0000     35.000000   19224.717500
50%      2910.500000      1.0000     53.000000   33170.605000
75%     13168.000000      3.0000     71.000000   53246.170000
max    111850.000000     10.0000     89.000000  258900.700000

       Outage_sec_perweek         Email      Contacts
Yearly_equip_failure  \
count        10000.000000  10000.000000  10000.000000
10000.000000
mean            10.001848     12.016000      0.994200
0.398000
std              2.976019      3.025898      0.988466
0.635953
min              0.099747      1.000000      0.000000
0.000000
25%              8.018214     10.000000      0.000000
0.000000
50%             10.018560     12.000000      1.000000
0.000000
75%             11.969485     14.000000      2.000000
1.000000
max             21.207230     23.000000      7.000000
6.000000

             Tenure  MonthlyCharge
count  10000.000000   10000.000000
mean      34.526188     172.624816
std       26.443063      42.943094
min        1.000259      79.978860
25%        7.917694     139.979239
50%       35.430507     167.484700
75%       61.479795     200.734725

max       71.999280     290.160419
```

Summary statistics for Categorical Variables:

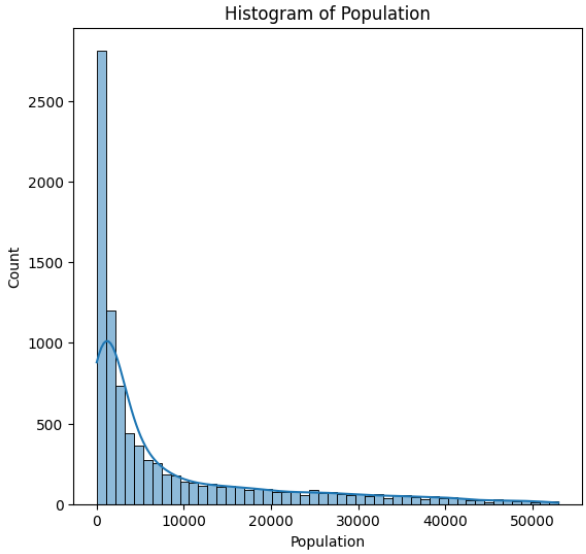
```
Summary for 'Gender':
Counts:
Gender
Female       5025
Male         4744
Nonbinary     231
Name: count, dtype: int64

Percentages:
Gender
Female       50.25
Male         47.44
Nonbinary     2.31
Name: proportion, dtype: float64

--------------------------------------------------

Summary for 'Churn':
Counts:
Churn
No     7350
Yes    2650
Name: count, dtype: int64

Percentages:
Churn
No     73.5
Yes    26.5
Name: proportion, dtype: float64

--------------------------------------------------

Summary for 'Techie':
Counts:
Techie
No     8321
Yes    1679
Name: count, dtype: int64

Percentages:
Techie
No     83.21
Yes    16.79
Name: proportion, dtype: float64

--------------------------------------------------

Summary for 'TechSupport':
Counts:
TechSupport
No     6250
Yes    3750
Name: count, dtype: int64

Percentages:
TechSupport
```

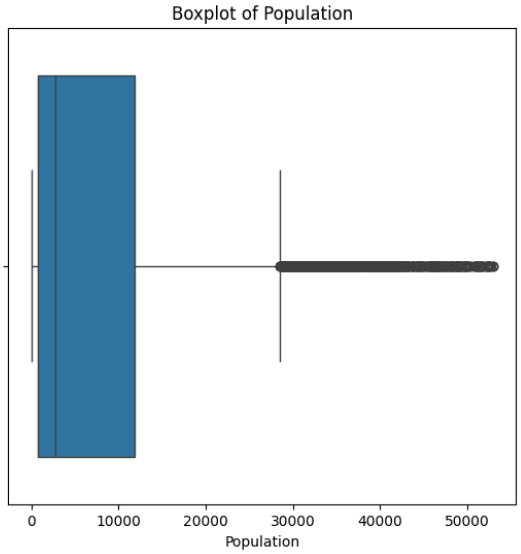
```
No      62.5
Yes     37.5
Name: proportion, dtype: float64


--------------------------------------------------
```

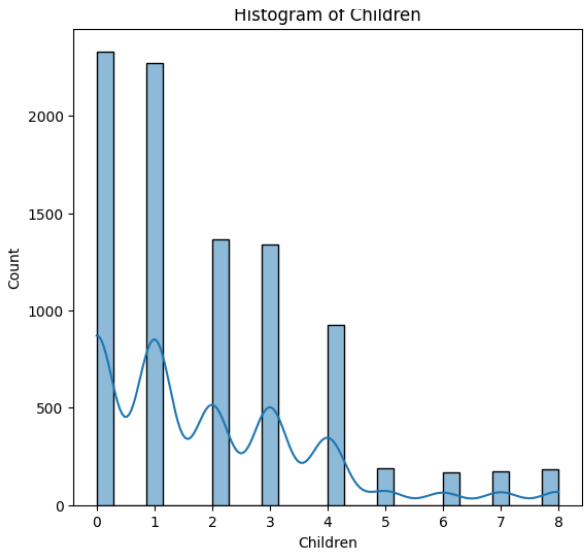
C3. Visualizations

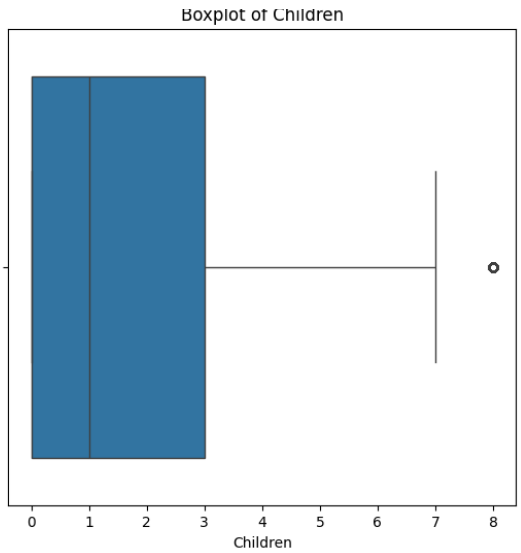
Univariate statistics involve the statistical analysis of a single variable at a time (Bruce et al., 2020). Below are the distributions of all independent variables in this analysis.

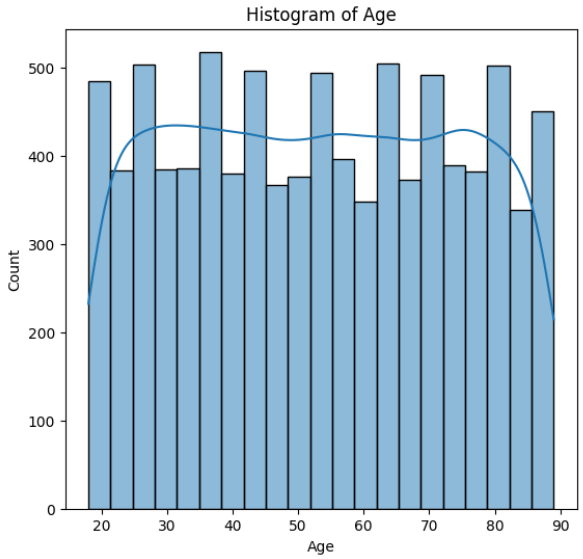
For the numerical independent variables, both histograms and boxplots are used. Below are the visualizations.

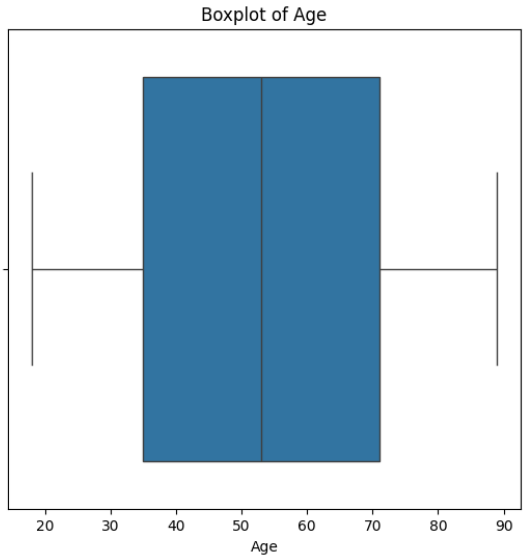

Histogram of Population
Count
Population
0
500
1000
1500
2000
2500
0
10000
20000
30000
40000
50000

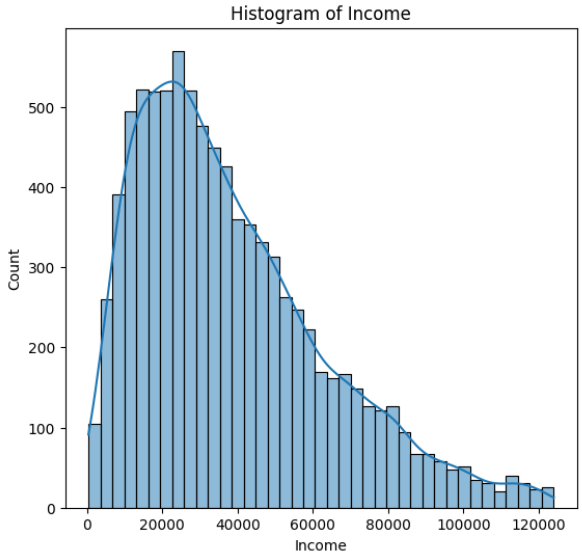

Boxplot of Population
Population
0
10000
20000
30000
40000
50000

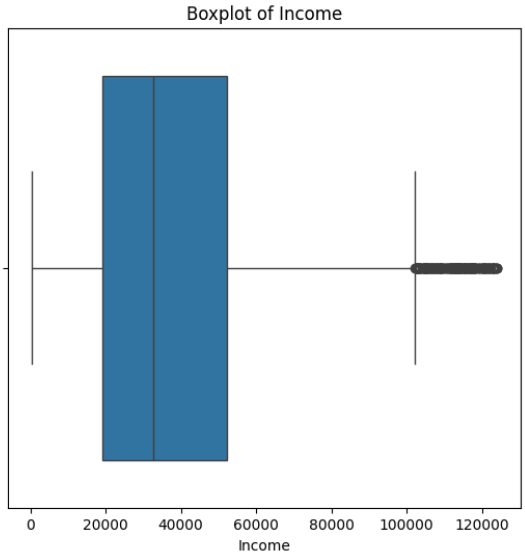

Histogram of Children
Count
Children
0
500
1000
1500
2000
0
1
2
3
4
5
6
7
8

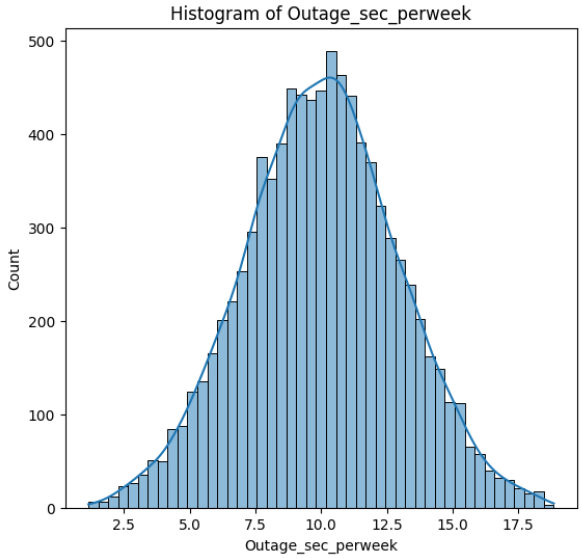

Boxplot of Children
Children
0
1
2
3
4
5
6
7
8

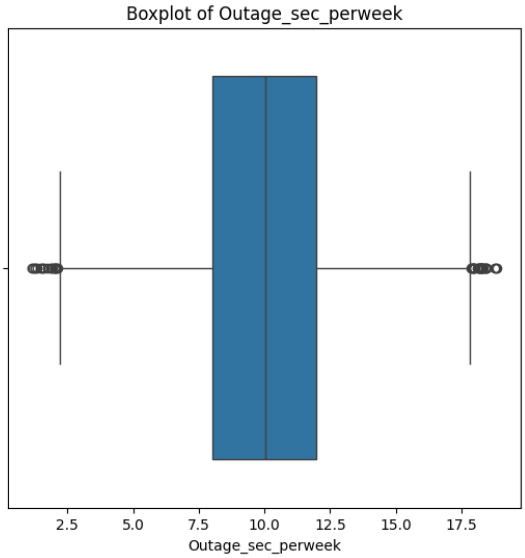

Histogram of Age
Count
Age
0
100
200
300
400
500
20
30
40
50
60
70
80
90

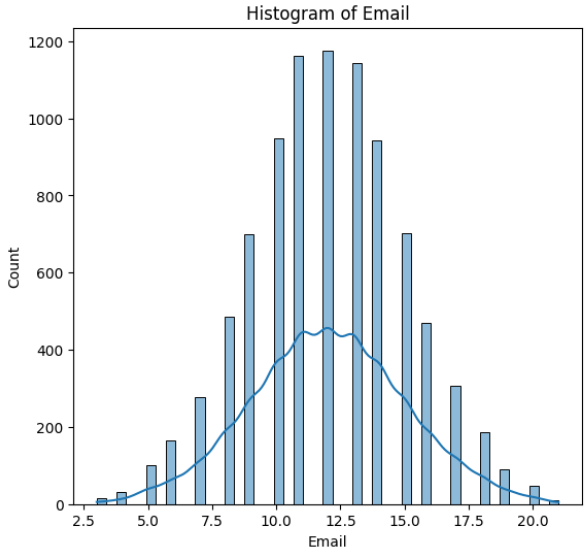

Boxplot of Age
Age
20
30
40
50
60
70
80
90

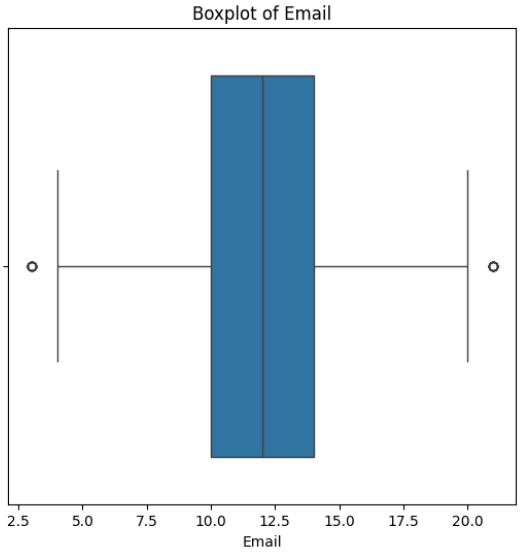

Histogram of Income
Count
Income
0
100
200
300
400
500
0
20000
40000
60000
80000
100000
120000

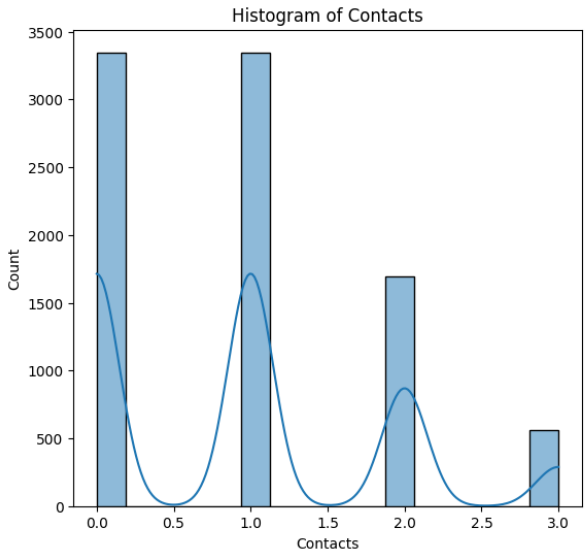

Boxplot of Income
Income
0
20000
40000
60000
80000
100000
120000

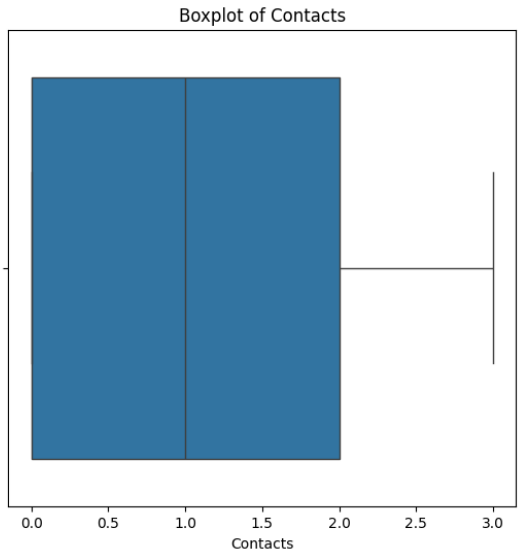

Histogram of Outage_sec_perweek
Count
Outage_sec_perweek
0
100
200
300
400
500
2.5
5.0
7.5
10.0
12.5
15.0
17.5

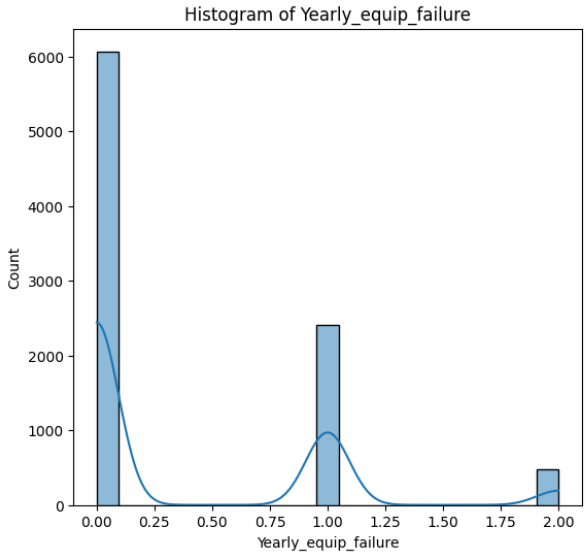

Boxplot of Outage_sec_perweek
Outage_sec_perweek
2.5
5.0
7.5
10.0
12.5
15.0
17.5

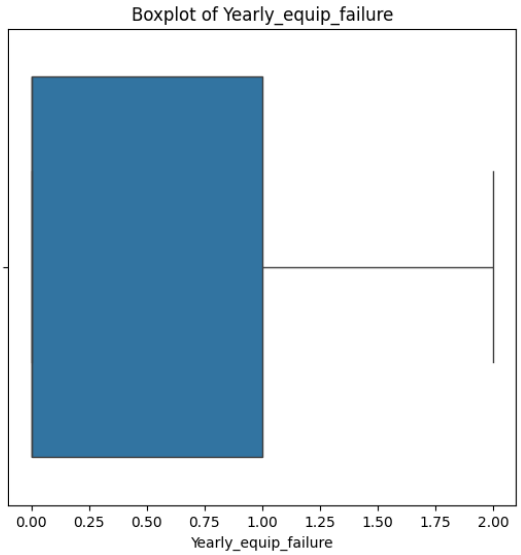

Histogram of Email
Count
Email
0
200
400
600
800
1000
1200
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

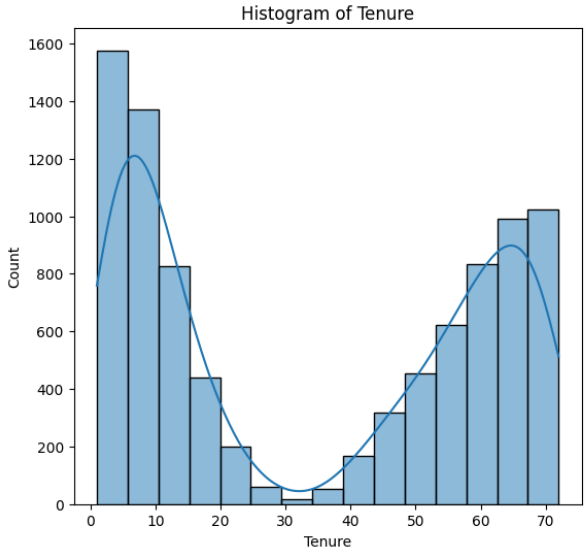

Boxplot of Email
Email
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

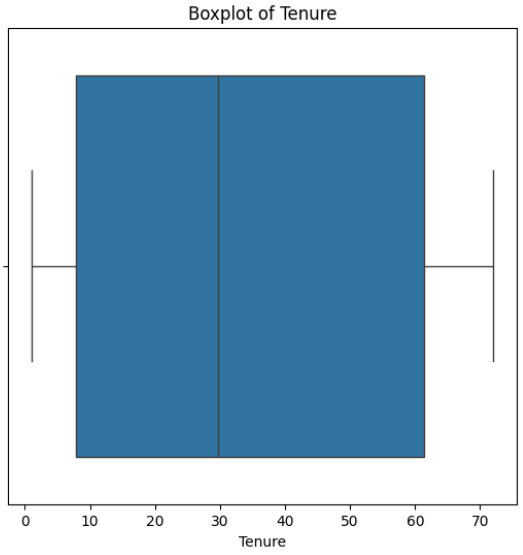

Histogram of Contacts
Count
Contacts

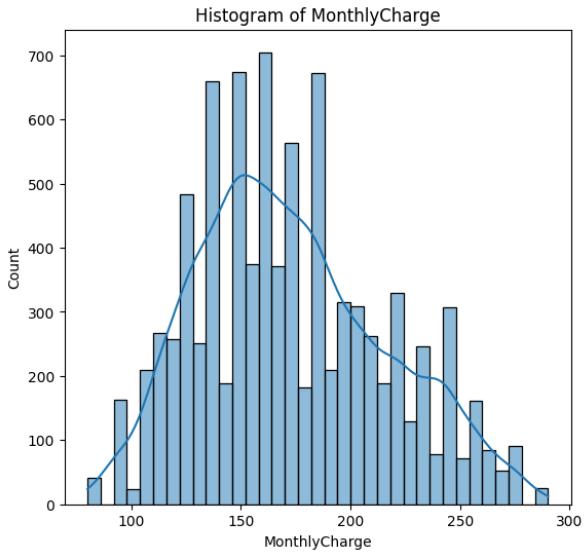

Boxplot of Contacts
Contacts

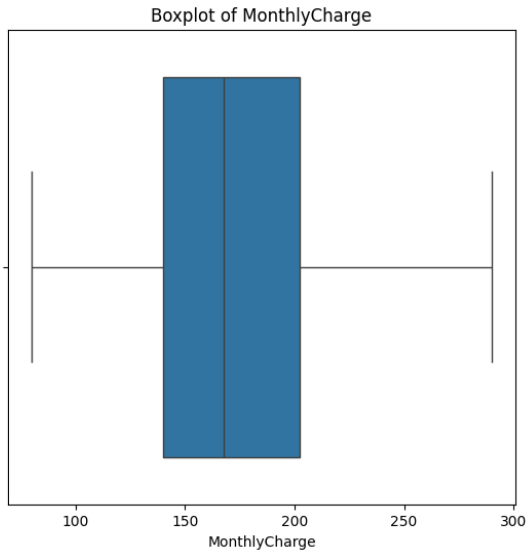

Histogram of Yearly_equip_failure
Count
Yearly_equip_failure


Boxplot of Yearly_equip_failure
Yearly_equip_failure

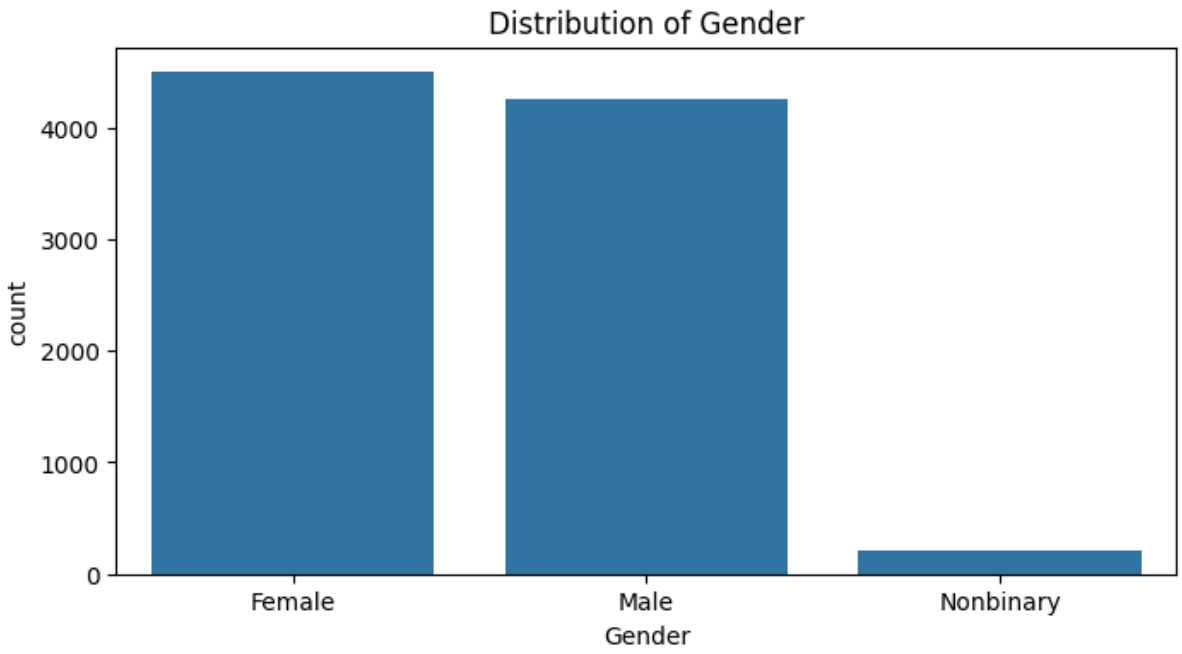

Histogram of Tenure
Count
Tenure

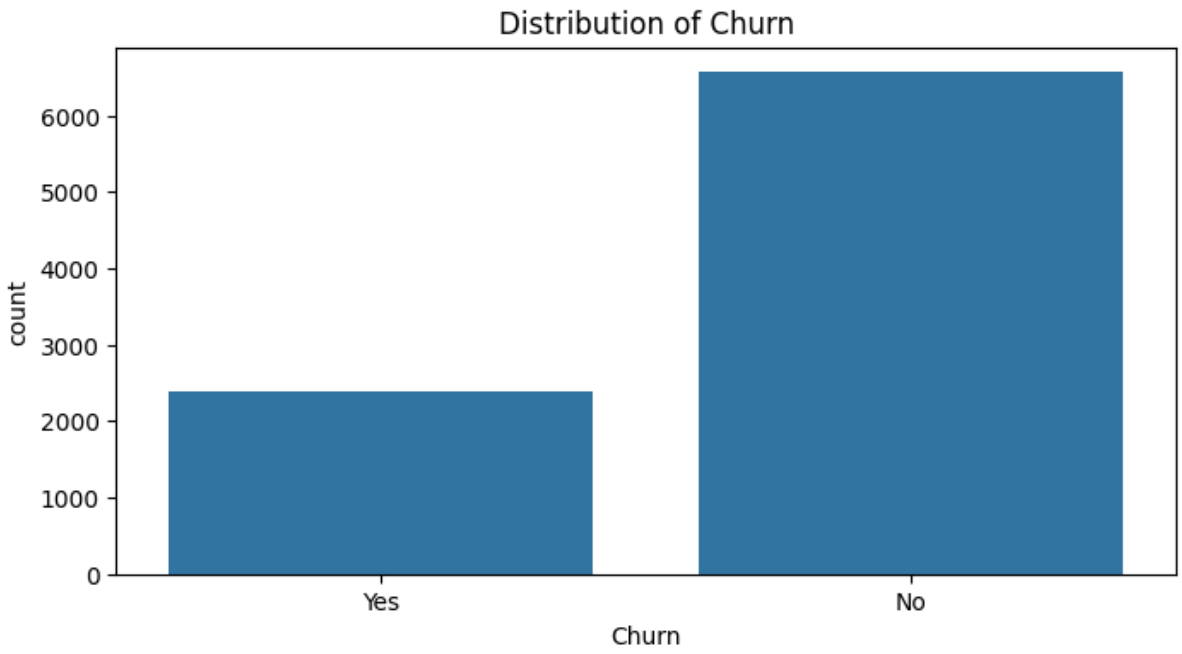

Boxplot of Tenure
Tenure

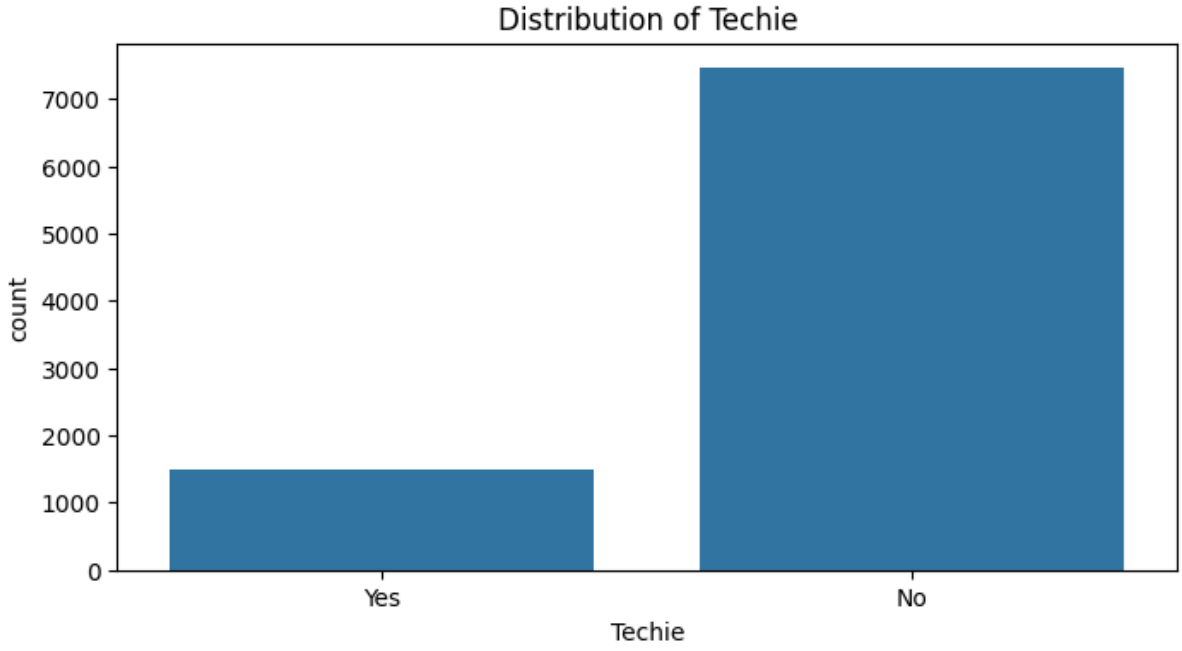


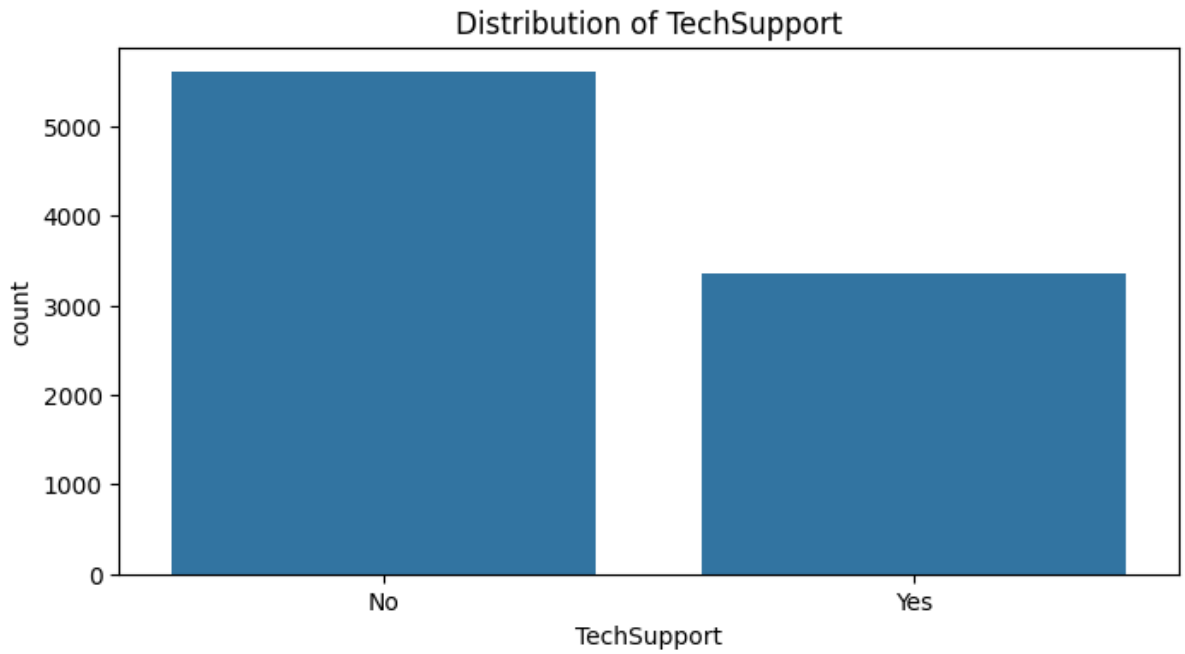


For categorical variables, bar plots are the most informative.




Distribution of Churn
6000
5000
4000
3000
2000
1000
0
count
Yes
No
Churn


Distribution of Techie
7000
6000
5000
4000
3000
2000
1000
0
count
Yes
No
Techie

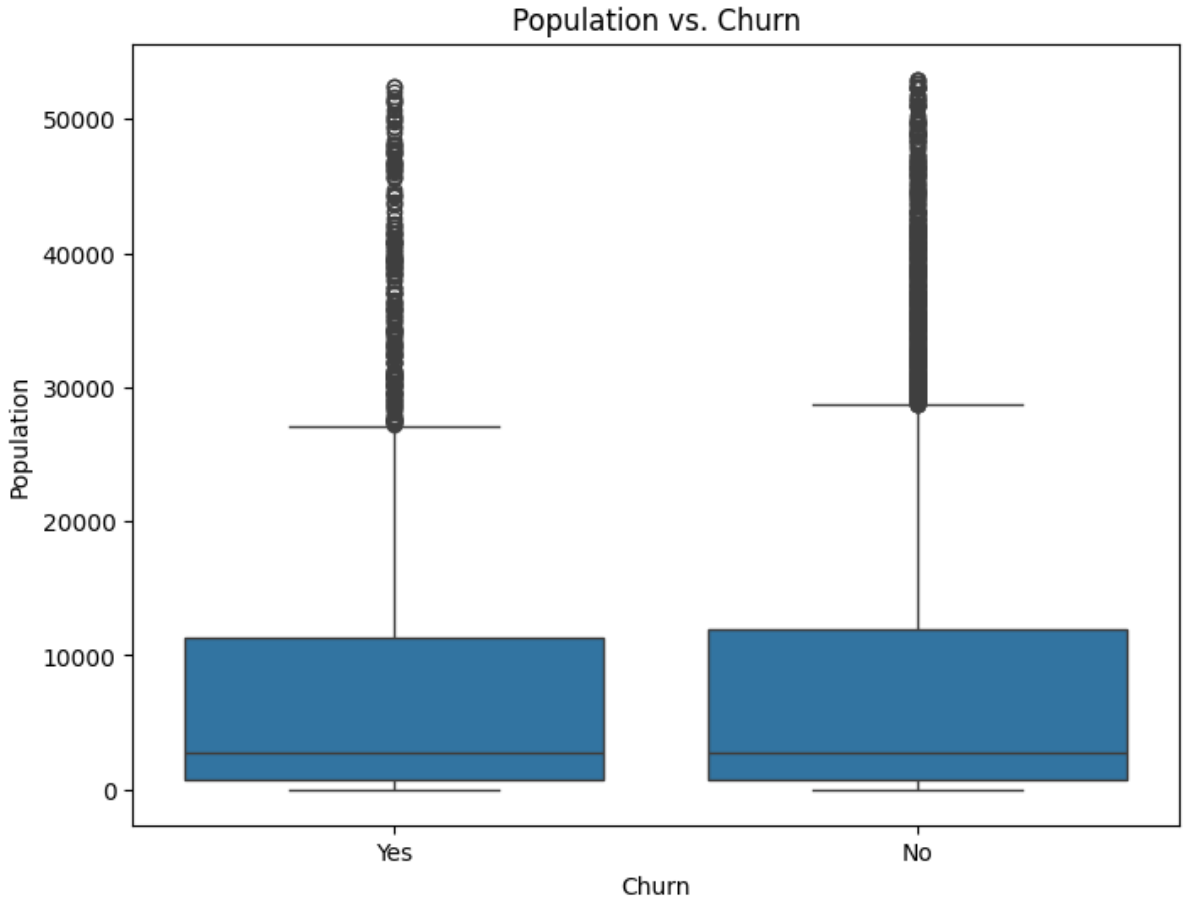


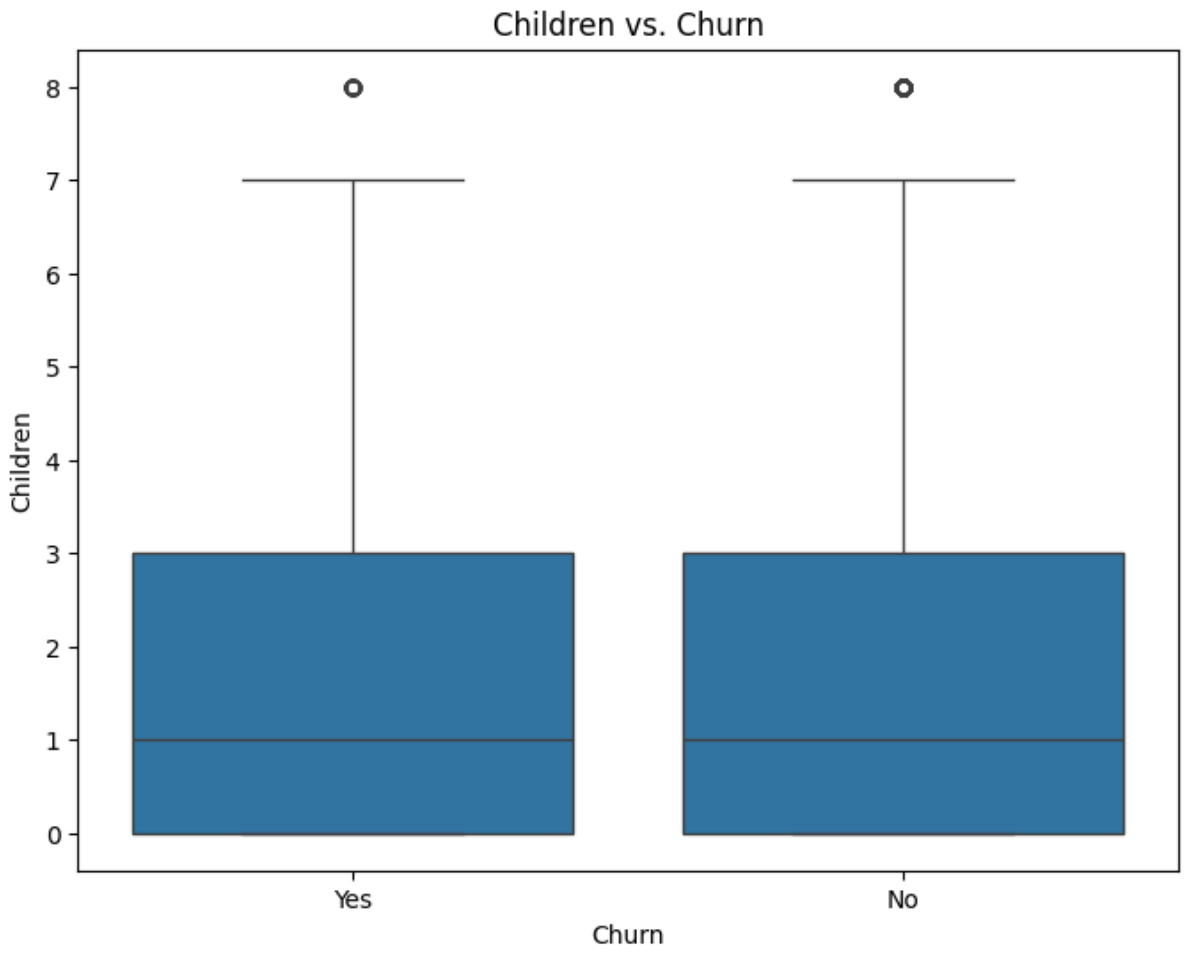
Bivariate Visualization with Churn:

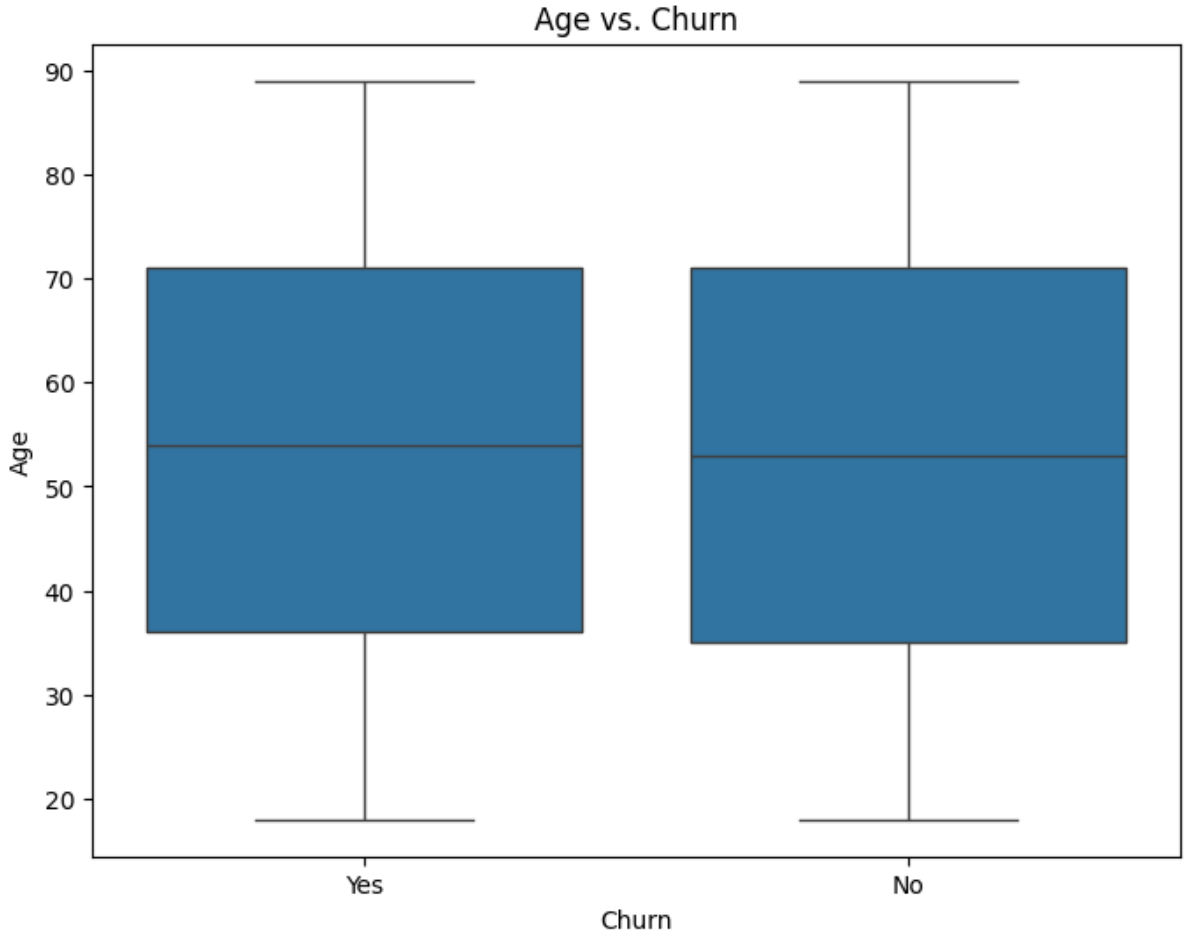
Bivariate statistical analysis involves examining the relationship between two variables simultaneously (Bruce et al., 2020).

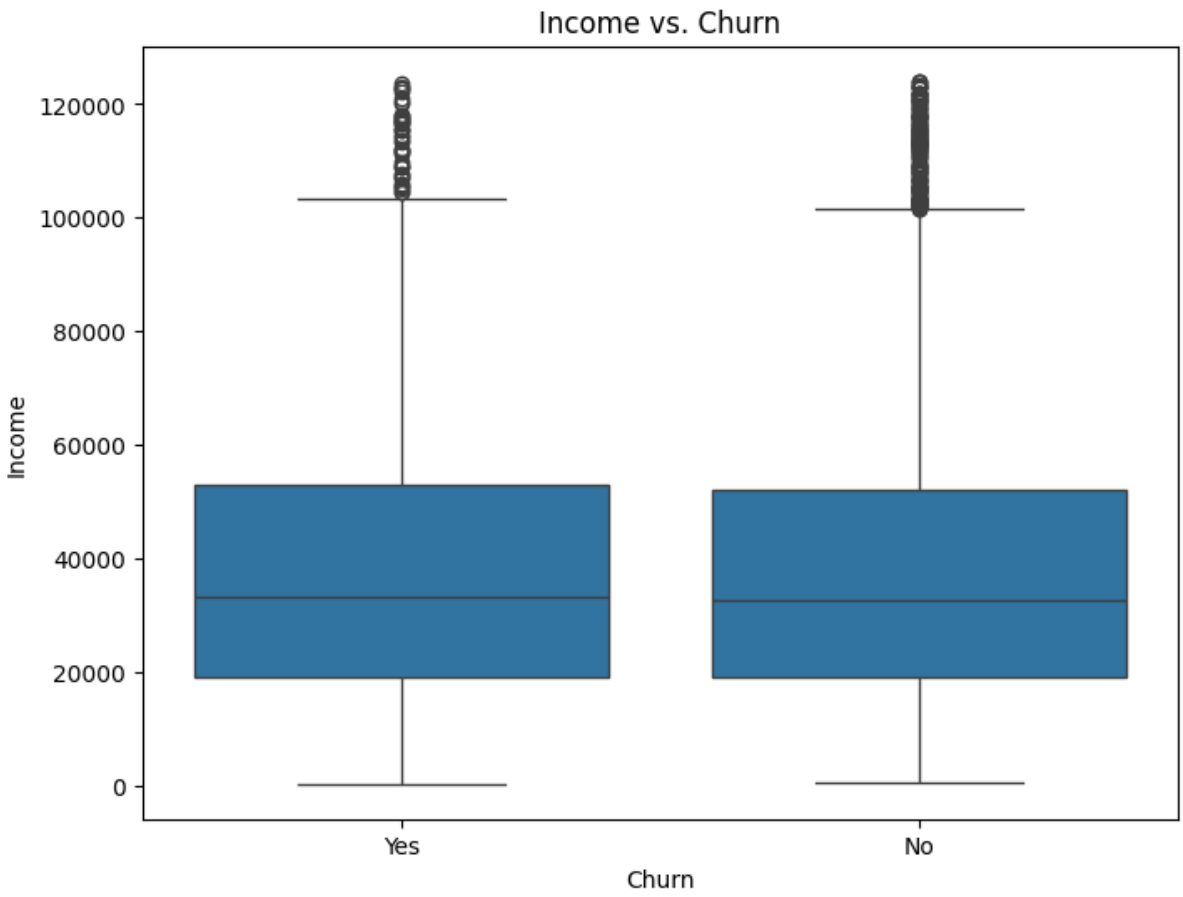
1. Numerical Variables vs. Churn

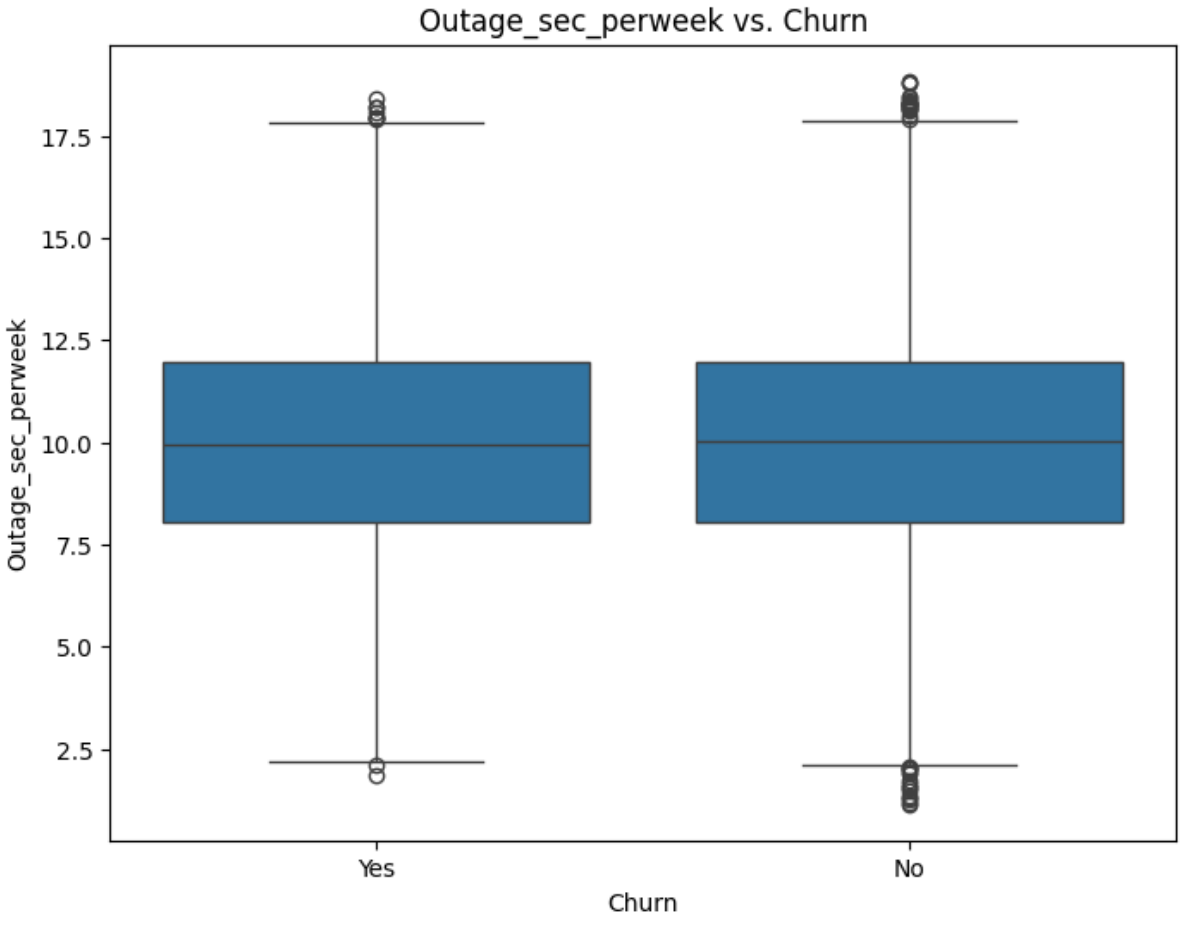
Boxplots show the distribution of numerical variables across the Churn categories.

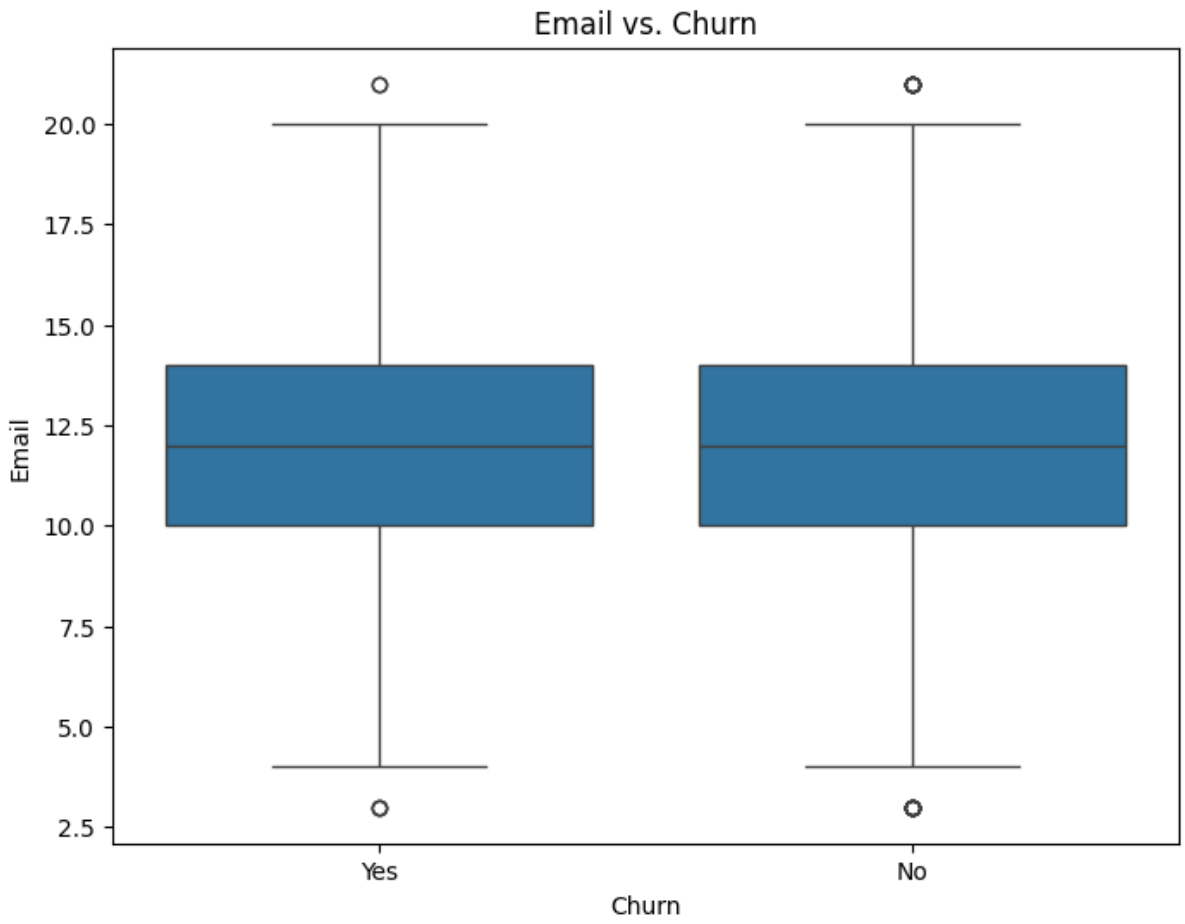


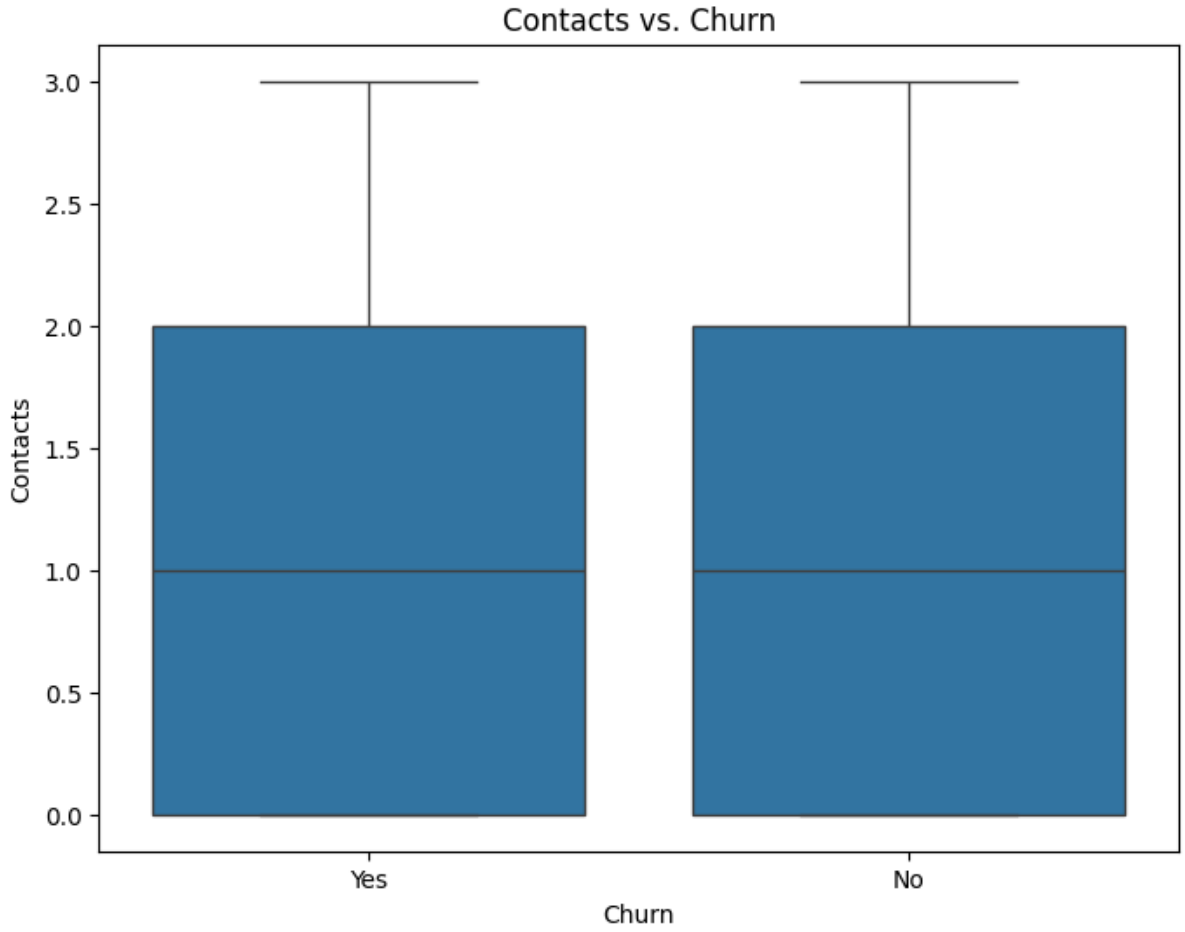

Children vs. Churn
Children
0
1
2
3
4
5
6
7
8
Yes
No
Churn

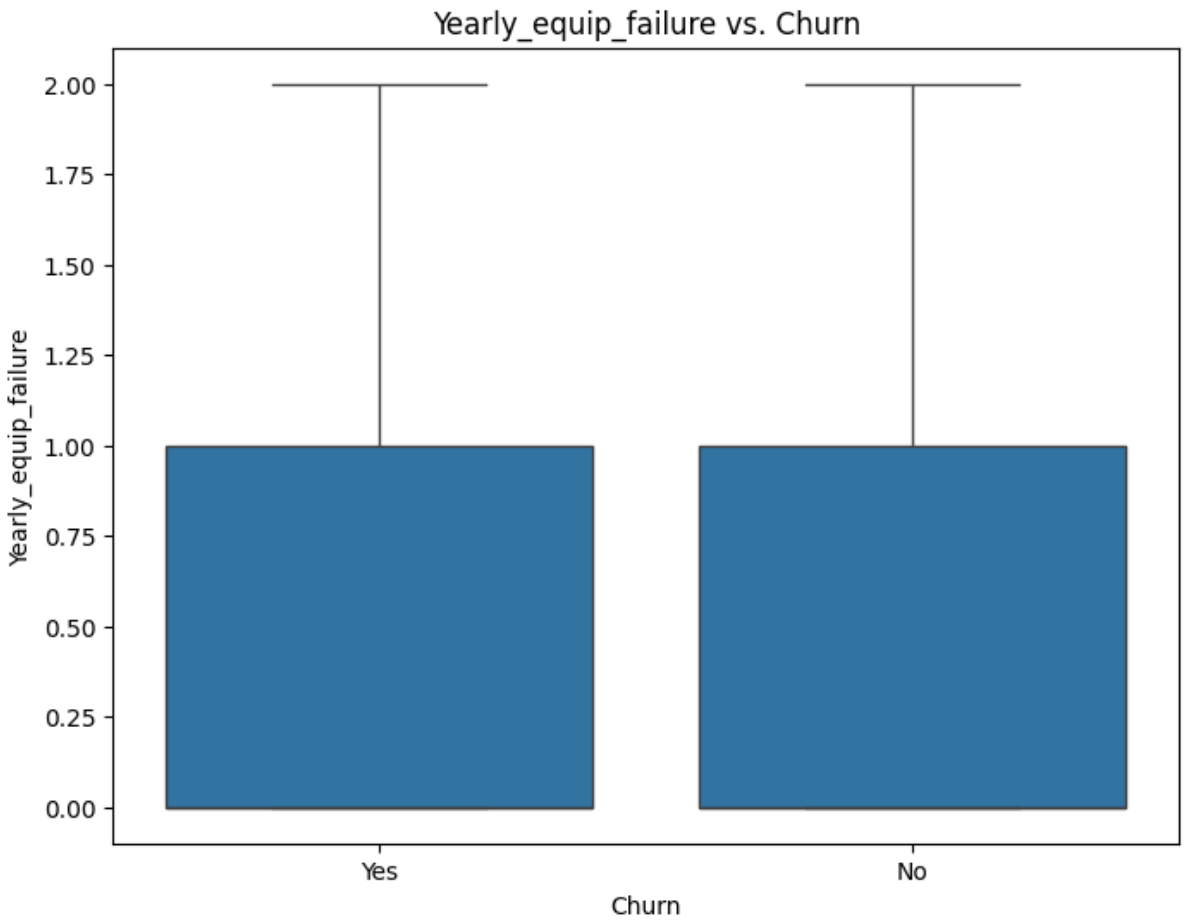

Age vs. Churn
Age
20
30
40
50
60
70
80
90
Yes
No
Churn

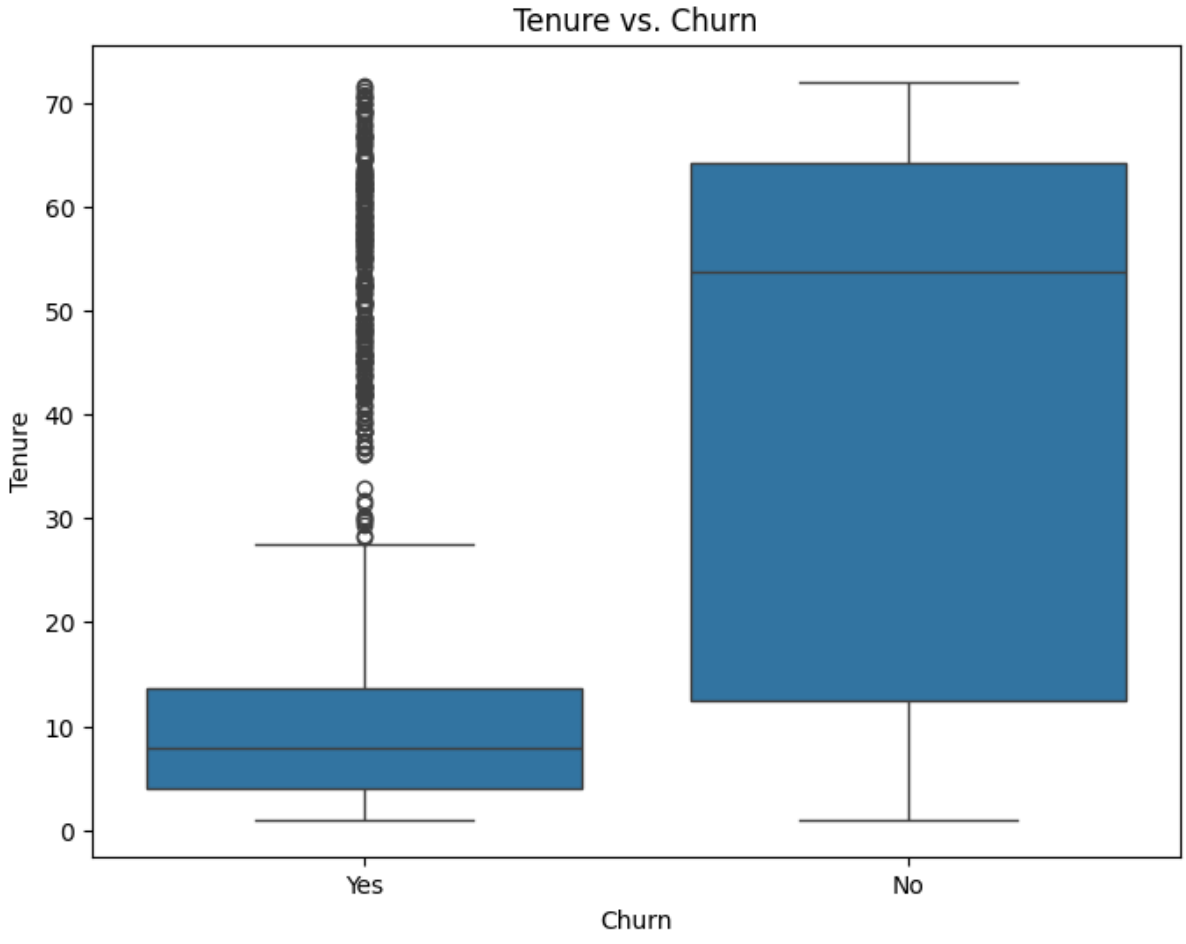

Income vs. Churn
Income
120000
100000
80000
60000
40000
20000
0
Yes
No
Churn

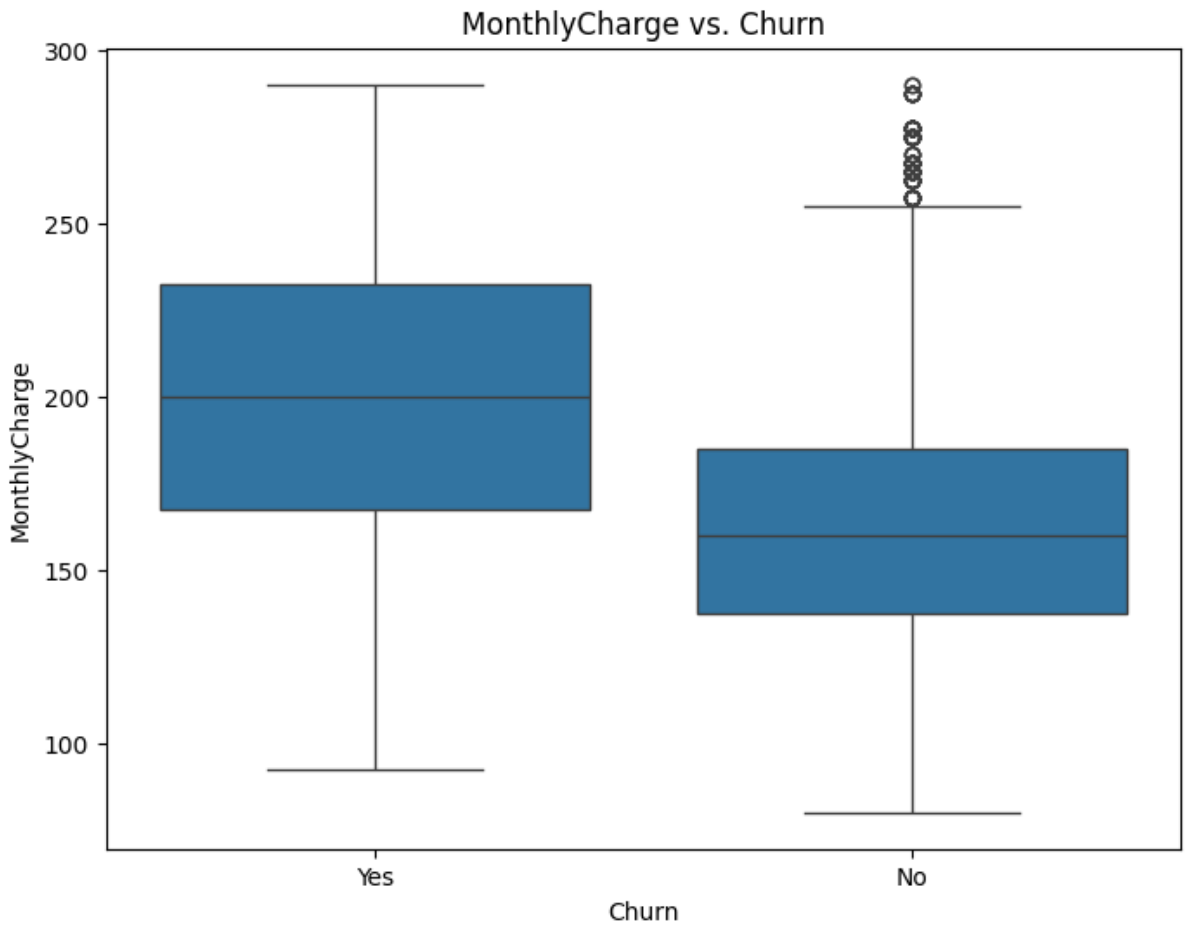

Outage_sec_perweek vs. Churn
Outage_sec_perweek
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Yes
No
Churn


Email vs. Churn
Email
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Yes
No
Churn


Contacts vs. Churn
Contacts
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Yes
No
Churn

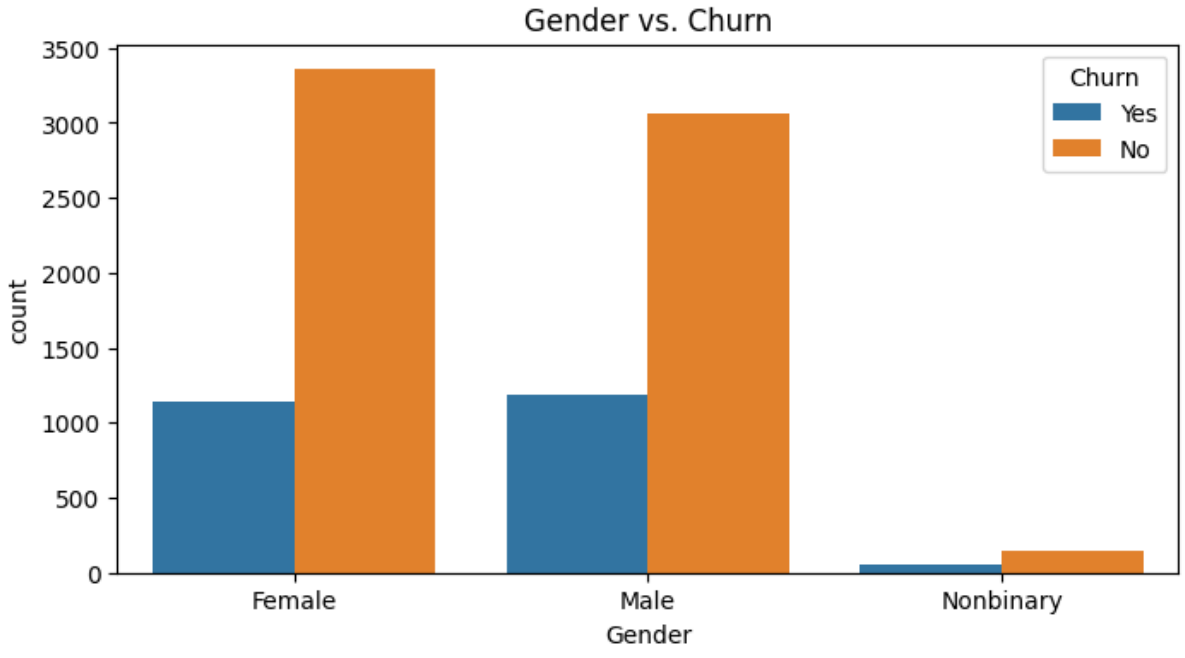

Yearly_equip_failure vs. Churn
Yearly_equip_failure
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Yes
No
Churn

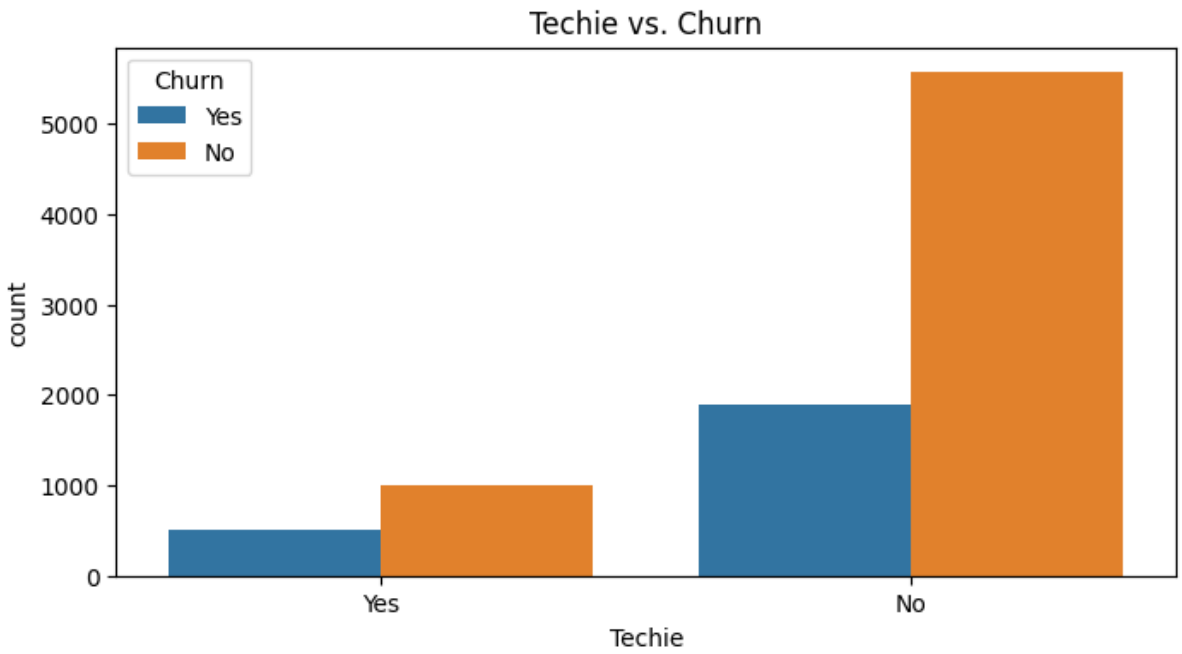

Tenure vs. Churn
Tenure
70
60
50
40
30
20
10
0
Yes
No
Churn

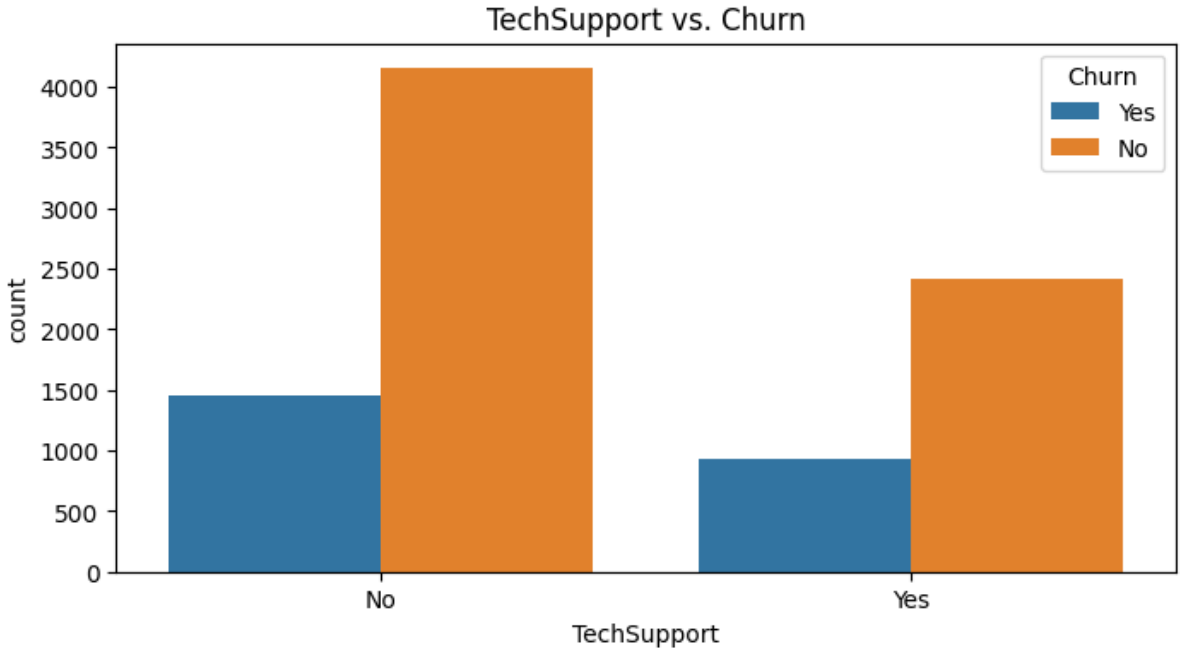


2. Categorical Variables vs. Churn

For categorical variables, we use bar plots to visualize the relationship with Churn.




Techie vs. Churn
Churn
Yes
No
count
0
1000
2000
3000
4000
5000
Yes
No
Techie


TechSupport vs. Churn
Churn
Yes
No
count
0
500
1000
1500
2000
2500
3000
3500
4000
No
Yes
TechSupport

C4. Data Transformation

The data transformation process aimed to prepare the dataset to ensure it aligns with the research question and facilitates logistic regression modeling. The primary objectives were:

1. **Convert Categorical Variables to Numerical Representations:** Logistic regression requires numerical inputs. Therefore, categorical variables such as Churn, Techie, TechSupport, and Gender were transformed into numerical formats.
2. **Feature Engineering:** Create new features to capture potential interactions or relationships between variables. For example, an interaction term between Techie and TechSupport was added to explore their combined influence on customer churn.
3. **Ensure Consistency in Data Types:** Ensure all columns in the dataset are numeric and formatted correctly to avoid issues during the logistic regression modeling process. This included converting boolean variables to integers and verifying all numerical data types.

Steps Taken

Step 1: Convert Categorical Variables to Numerical Representations

Binary categorical variables (Churn, Techie, TechSupport) were encoded using LabelEncoder, converting them into integer values (1 for "Yes" and 0 for "No"). The Gender variable, which includes Male, Female, and Nonbinary categories, was one-hot encoded, resulting in two binary columns: Gender_Male and Gender_Nonbinary.

Code

```
from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder

label_encoder = LabelEncoder()

# Encode binary categorical columns

df['Churn'] = label_encoder.fit_transform(df['Churn'])  # 0 = No, 1 = Yes

df['Techie'] = label_encoder.fit_transform(df['Techie'])

df['TechSupport'] = label_encoder.fit_transform(df['TechSupport'])

# One-hot encode 'Gender'

df = pd.get_dummies(df, columns=['Gender'], drop_first=True)
```

Step 2: Feature Engineering

To capture the combined effect of technical inclination and support on churn, an interaction term (Techie_TechSupport) was created by multiplying the Techie and TechSupport columns. This interaction term provides additional insight into how these variables jointly influence customer behavior.

Code

```
# Creating interaction term between Techie and TechSupport

df['Techie_TechSupport'] = df['Techie'] * df['TechSupport']
```

Step 3: Ensure Consistency in Data Types

To ensure compatibility with logistic regression modeling, all boolean columns were explicitly converted to integer types, and all variables were verified to be numeric. Additionally, missing values were checked and handled by dropping any rows with missing data.

Code

```
# Ensure all boolean columns are converted to integers

X = X.astype({col: 'int' for col in X.select_dtypes(include=['bool']).columns})

# Ensure all data is numeric

X = X.apply(pd.to_numeric, errors='coerce')

y = pd.to_numeric(y, errors='coerce')

# Check for missing data

print("Missing values in X:", X.isnull().sum())

print("Missing values in y:", y.isnull().sum())

# Drop rows with missing values

X = X.dropna()

y = y.loc[X.index]
```

Through these steps, the dataset was successfully transformed to meet the requirements of logistic regression modeling. The categorical variables were numerically encoded, an interaction term was introduced, and the dataset was thoroughly verified for consistency. This ensured that the data was in an optimal format for the subsequent modeling phase.

C5. Prepared Data set:

The dataset is saved as a CSV file for further analysis and is attached with the submission. Check prepared_churn_dataset.csv file.

Code

```
# save the prepared data set

df.to_csv('prepared_churn_dataset.csv', index=False)
```

Part IV: Model Comparison and Analysis

D1. Constructing the Initial Logistic Regression Model:

To identify the key factors influencing customer churn, the initial logistic regression model included all independent variables. This approach allowed us to assess the statistical significance of each variable and its predictive power for customer churn.

The model summary revealed several variables with p-values below the significance threshold (0.05), indicating their importance in predicting churn. These variables included Techie, TechSupport, Tenure, MonthlyCharge, and Gender_Male.

Code

```
import pandas as pd

import statsmodels.api as sm

# Load the prepared dataset

df = pd.read_csv('prepared_churn_dataset.csv')

# Initial Logistic Regression Model

X = df.drop('Churn', axis=1)

y = df['Churn']

X = sm.add_constant(X)


# Ensure all boolean columns are converted to integers

X = X.astype({col: 'int' for col in X.select_dtypes(include=['bool']).columns})


# Ensure all data is numeric

X = X.apply(pd.to_numeric, errors='coerce')

y = pd.to_numeric(y, errors='coerce')


# Check for missing data

print("Missing values in X:", X.isnull().sum())

print("Missing values in y:", y.isnull().sum())


# Drop any rows with missing values
```

```
X = X.dropna()

y = y.loc[X.index]


# Fit the initial logistic regression model

initial_model = sm.Logit(y, X).fit()

print("Initial Model Summary:")

print(initial_model.summary())
```

```
Missing values in X: const                  0
Population              0
Children                0
Age                     0
Income                  0
Outage_sec_perweek      0
Email                   0
Contacts                0
Yearly_equip_failure    0
Techie                  0
TechSupport             0
Tenure                  0
MonthlyCharge           0
Gender_Male             0
Gender_Nonbinary        0
Techie_TechSupport      0
dtype: int64
Missing values in y: 0
Optimization terminated successfully.
         Current function value: 0.337934
         Iterations 8
Initial Model Summary:
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9984
Method:                           MLE   Df Model:                           15
Date:                Tue, 21 Jan 2025   Pseudo R-squ.:                 0.4156
Time:                        17:33:24   Log-Likelihood:                -3379.3
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -5.4723      0.247    -22.168      0.000      -5.956      -4.988
Population           -2.693e-06   2.15e-06     -1.255      0.209    -6.9e-06    1.51e-06
Children                 -0.0063      0.014     -0.434      0.664      -0.035       0.022
Age                       0.0019      0.001      1.271      0.204      -0.001       0.005
Income                8.566e-07   1.08e-06      0.790      0.429   -1.27e-06    2.98e-06
Outage_sec_perweek       -0.0023      0.010     -0.226      0.821      -0.022       0.018
Email                     0.0026      0.010      0.258      0.797      -0.017       0.023
Contacts                  0.0257      0.031      0.836      0.403      -0.035       0.086
Yearly_equip_failure     -0.0274      0.048     -0.568      0.570      -0.122       0.067
Techie                    0.6228      0.100      6.206      0.000       0.426       0.820
TechSupport              -0.2006      0.070     -2.871      0.004      -0.338      -0.064
Tenure                   -0.0747      0.002    -41.738      0.000      -0.078      -0.071
MonthlyCharge             0.0338      0.001     37.183      0.000       0.032       0.036
Gender_Male               0.1671      0.062      2.706      0.007       0.046       0.288
Gender_Nonbinary         -0.1926      0.203     -0.951      0.342      -0.590       0.204
Techie_TechSupport       -0.0459      0.162     -0.283      0.777      -0.363       0.272
==========================================================================================
```

D2. Justification of Model Reduction

Procedure Used: Statistically-Based Feature Selection via P-values

Given the research question, the goal is to identify key factors influencing customer churn. A statistically based feature selection method, particularly the evaluation of p-values, is suitable for this purpose. To refine the model, we applied a statistically-based feature selection method by evaluating the p-values of each variable:

- **Objective:** Remove variables with p-values greater than the significance threshold of 0.05.
- **Process:** Iteratively remove variables with the highest p-values until all remaining variables are statistically significant.

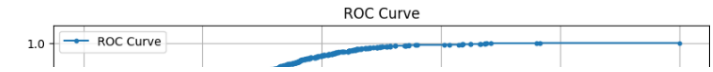
Code

```
# Feature Selection and Reduced Model

current_model = initial_model

significant_level = 0.05

while True:

    p_values = current_model.pvalues

    max_p_value = p_values.max()

    max_p_var = p_values.idxmax()


    if max_p_value > significant_level:

        X = X.drop(columns=max_p_var)

        current_model = sm.Logit(y, X).fit()

    else:

        break

print("Reduced Model Summary:")

print(current_model.summary())
```

D3. Reduced Logistic Regression Model

After applying the feature selection procedure, we developed a reduced logistic regression model that retains only the significant variables. The model was reduced from the original set of independent variables down to five variables, each with a statistically significant p-value.

```
==========================================================================================
Optimization terminated successfully.
         Current function value: 0.337936
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.337940
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.337944
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.337953
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.337970
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.338000
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.338035
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.338084
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.338161
         Iterations 8
Optimization terminated successfully.
         Current function value: 0.338241
         Iterations 8
Reduced Model Summary:
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                10000
Model:                          Logit   Df Residuals:                    9994
Method:                           MLE   Df Model:                           5
Date:                Tue, 21 Jan 2025   Pseudo R-squ.:                 0.4150
Time:                        17:33:26   Log-Likelihood:                -3382.4
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -5.3601      0.157    -34.186      0.000      -5.667      -5.053
Techie          0.6071      0.079      7.684      0.000       0.452       0.762
TechSupport    -0.2064      0.063     -3.269      0.001      -0.330      -0.083
Tenure         -0.0746      0.002    -41.746      0.000      -0.078      -0.071
MonthlyCharge   0.0338      0.001     37.206      0.000       0.032       0.036
Gender_Male     0.1749      0.061      2.870      0.004       0.055       0.294
==============================================================================
Confusion Matrix:
[[6732  618]
 [1019 1631]]

Accuracy: 0.8363
ROC AUC: 0.9018
```



E1. Model Comparison:

Comparison of Initial and Reduced Models

1. Initial Logistic Regression Model

Results:

- Log-Likelihood: -3379.3
- Pseudo R-squared: 0.4156
- Significant Variables: Variables such as Techie, TechSupport, Tenure, MonthlyCharge, and Gender_Male were statistically significant, while others like Population, Children, Age, and Income were not.

2. Reduced Logistic Regression Model

Results:

- Log-Likelihood: -3382.4
- Pseudo R-squared: 0.4150
- Significant Variables: The reduced model retained only five significant variables: Techie, TechSupport, Tenure, MonthlyCharge, and Gender_Male.

E2. Performance Evaluation: Confusion Matrix and Accuracy

Confusion Matrix

The confusion matrix for the reduced model is as follows:

Predicted: No Churn	**Predicted: Churn**
Actual: No Churn	6732
Actual: Churn	1019

- True Positives (TP): 1,631 customers who churned were correctly identified.
- True Negatives (TN): 6,732 customers who did not churn were correctly identified.
- False Positives (FP): 618 customers were incorrectly predicted to churn but did not.
- False Negatives (FN): 1,019 customers who churned were not correctly identified.

Accuracy

- Accuracy: 0.8363
- This accuracy score indicates that 83.63% of the model's predictions (both churn and no churn) were correct.

E3. Complete Error-Free Python Code:

An error-free copy of the code is attached with the submission.

Part V: Data Summary and Implications

F1. Results

a. Regression Equation for the Reduced Model

The reduced logistic regression model can be represented by the following equation:

$$Logit(P(Churn)) = -5.3601 + 0.6071 \times Techie - 0.2064 \times TechSupport - 0.0746 \times Tenure + 0.0338 \times MonthlyCharge + 0.1749 \times Gender\_Male$$

b. Interpretation of Coefficients:

- Techie: Customers who identify as tech-savvy have a higher likelihood of churning (+0.6071).
- TechSupport: Customers with tech support add-ons are less likely to churn (−0.2064).
- Tenure: Longer tenure with the company significantly decreases the likelihood of churn (−0.0746).
- MonthlyCharge: Higher monthly charges are associated with an increased likelihood of churn (+0.0338).
- Gender_Male: Male customers have a slightly higher probability of churning compared to other genders (+0.1749).

c. Statistical and Practical Significance

- Statistical Significance: The variables included in the reduced model are statistically significant, as indicated by their p-values (all below 0.05).
- Practical Significance: The model is practically significant, as it highlights actionable factors (e.g., monthly charges, tech support) that can be targeted to reduce churn.

d. Limitations

The relatively small dataset may limit the model's generalizability. A larger dataset could provide more robust results. Additionally, the reduced model includes a limited number of predictors, which might not capture all factors influencing churn. The model's predictions might be influenced by biases in the data collection process or unobserved variables.

F2. Recommendations

Based on the analysis, it is recommended to:

- Enhance Tech Support Offerings: Since tech support is associated with reduced churn, the company should promote tech support services and potentially offer them as part of bundled packages.
- Target High-Risk Customers: Focus on customers with high monthly charges and those who identify as tech-savvy, as they are more likely to churn. Tailored retention strategies such as personalized offers or loyalty programs can be effective.

- Loyalty Programs: Strengthen loyalty programs to reward long-term customers, as tenure is a significant factor in reducing churn.

By implementing these recommendations, the business can work towards reducing churn rates and improving customer satisfaction.

Part VI: Demonstration

G. Demonstration

Video link: -

H. Web Sources

1. **https://www.tableau.com/learn/articles/what-is-data-cleaning**
2. **https://www.w3schools.com/python/pandas/default.asp**
3. Zach. (Oct, 2020). How to Perform Logistic Regression in Python (Step-byStep) https://www.statology.org/logistic-regression-python/
4. https://www.sciencedirect.com/topics/computer-science/logistic-regression

I. Sources

1. Zach. (Oct, 2020). How to Perform Logistic Regression in Python (Step-byStep) https://www.statology.org/logistic-regression-python/
2. Daniel T. Larose, & Chantal D. Larose. (2019). Data Science Using Python and R. Wiley
3. Panda, N. R. (2022). A review on logistic regression in medical research. *National Journal of Community Medicine*, *13*(04), 265-270.