JM GROUP

PRESENTATION - 2023

CUSTOMER
A N A L Y T I C S

ANALYZE THE CUSTOMER CHURN RATE OF TELECOMMUNICATIONS COMPANIES

JM GROUP

Tuấn Nguyên

Nhóm trưởng

Bích Vân

Nhật Quyên

Thúy Anh

MỤC LỤC

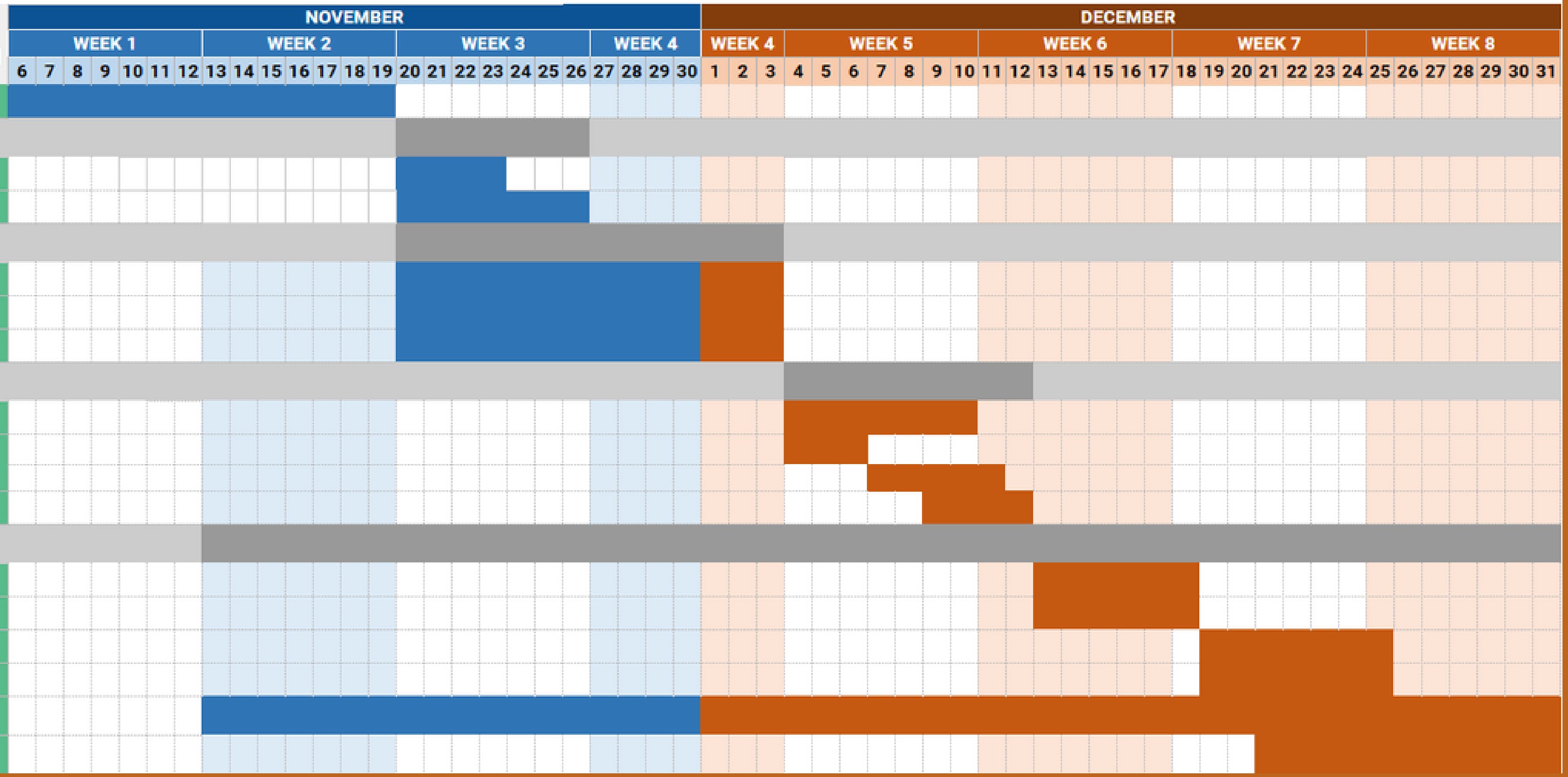
I.
INTRODUCE

Gantt chart

GANTT CHART

PROJECT	Analyze the customer churn rate of telecommunications companies
GROUP	JM GROUP
STAR DATE	6/11/2023
END DATE	25/12/23

NO	JOB TITLE	PIC	STAR DATE	END DATE	TIME PERIODS	PERCENT COMPLETION
0	Picking data modules, Define problems	All	06/11/23	19/11/23	13	100%
1	**Project overview**					
1.1	Reason for choosing the topic	Nhat Quyen	20/11/23	23/11/23	3	100%
1.2	Objectives, Target, and Scope of the Project	Thuy Anh	17/11/23	26/11/23	9	100%
2	**Theoretical basis**				0	
2.1	Customer Analytics	Nhat Quyen	20/11/23	03/12/23	13	100%
2.2	Methods and Tools	Thuy Anh	20/11/23	03/12/23	13	100%
2.3	Measurement Metrics	Bich Van	20/11/23	03/12/23	13	100%
3	**Data preparation**					
3.1	Collection data	Thuy Anh	04/12/23	10/12/23	6	100%
3.2	Exploratory Data Analysis (EDA)	Tuan Nguyen	04/12/23	06/12/23	2	100%
		Bich Van	07/12/23	11/12/23	4	
3.3	Bivariate Analysis	Tuan Nguyen	09/12/23	12/12/23	3	100%
4	**Telco Churn Analytics**					
4.1	Model Prediction	Tuan Nguyen	13/12/23	18/12/23	5	100%
4.2	Visualize with Power BI	Bich Van	13/12/23	18/12/23	5	100%
5	**Solution**	Nhat Quyen	19/12/23	25/12/23	6	100%
6	**Conclusion**	Thuy Anh	19/12/23	25/12/23	6	100%
7	**Slide**	Thuy Anh	13/12/23	31/12/23	18	100%
8	**Word (Final)**	All	21/12/23	31/12/23	10	100%



1.1 Choosing project purpose

- Choosing the topic "Telco Customer Churn" is based on the actual challenges that the telecommunications industry is facing.
- Research not only helps predict customer behavior but also provides cost optimization opportunities.

1.1 Choosing project purpose

- Using advanced technology like artificial intelligence and machine learning to build more accurate prediction models.
- The results of the research are not only theoretical but also have practical applications for telecommunications businesses.

1.2 Purpose, Object and Scope of Research

Objective 1: Develop a Churn Behavior Prediction Model.

Objective 2: Analyze Factors Influencing the Decision to Leave.

Objective 3: Propose an Optimal Retention Strategy.

II. THEORETICAL BASIS

DEFINITION

Customer analytics is an important area of modern business strategy, placing customers at the center of every decision. It is not just about collecting and processing data, but also about a deep analysis process to understand customer behavior and needs.

Customer analytics not only helps businesses identify current customers but also predict future trends.

IMPORTANCE

Multi-dimensional

Companies put so much effort into customer analytics is its ability to provide a detailed of customer behavior and desires

Predict future

customer analysis also provides tools to predict future customer behavior

Engagement & relationships

This not only creates a positive customer experience but also drives loyalty and strengthens brand positioning.

Long-term strategy

Customer analysis is an indispensable tool for shaping a business's long-term strategy.

QUY TRÌNH HOẠT ĐỘNG

Collecting and Processing Data

Data Analysis

Building Customer Profile

Prediction and Optimization

Interaction and Feedback

METHODS AND TOOLS

LOGISTIC REGRESSION

Logistic regression is a widely used statistical method in predictive analysis. In the context of customer analysis, logistic regression is often used to predict the probability of an event occurring based on one or more independent variables.

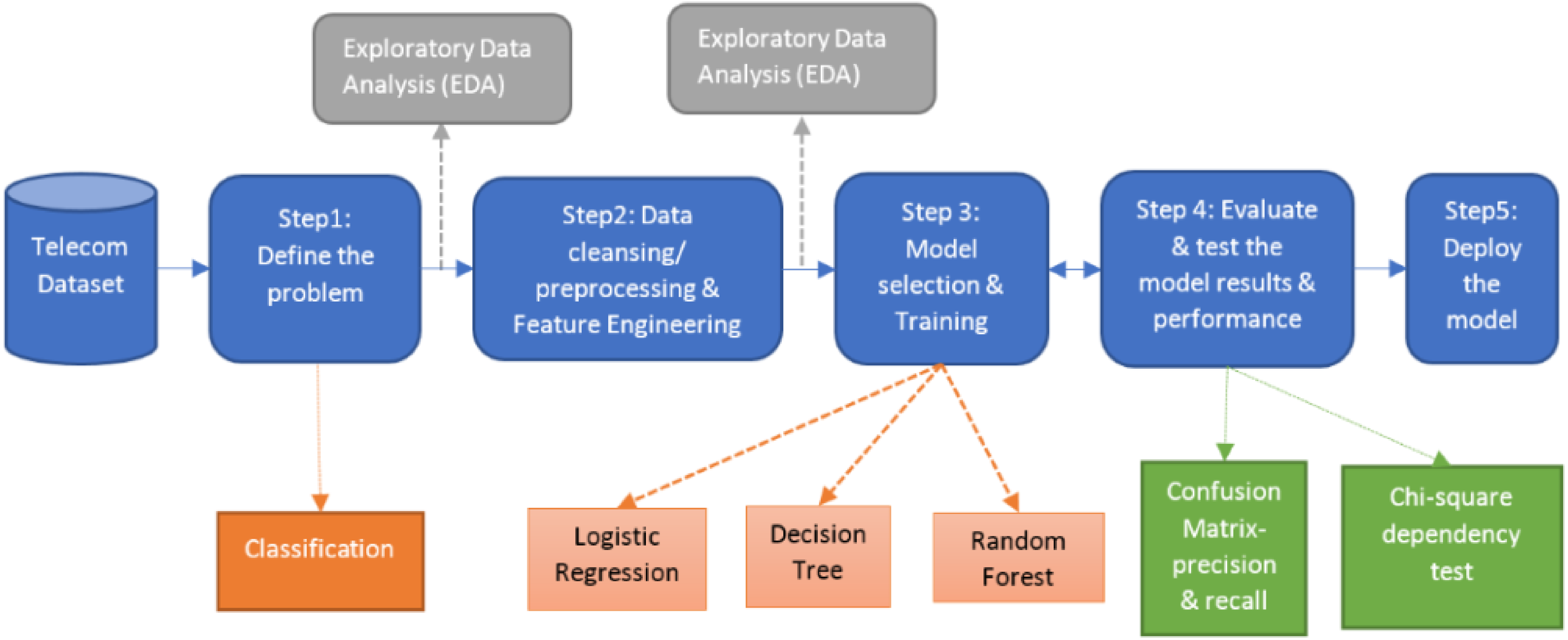
LOGISTIC REGRESSION

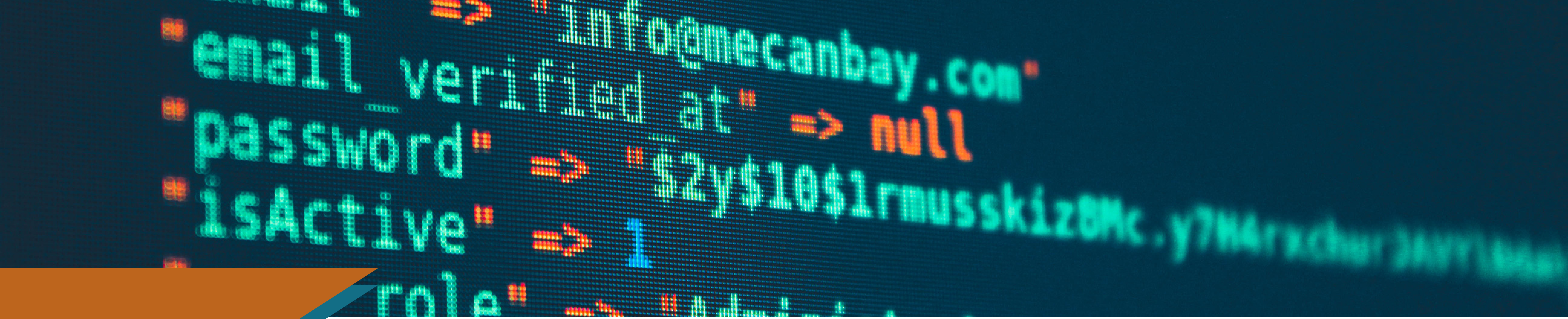
It can be used to predict whether a customer will purchase a product, whether a customer will continue to choose a business, etc. This information can then be used to formulate strategies and policies to retain customers, calculate the risk of their departure, predict the potential customer growth in the future, or classify customers into specific groups.

TOOLS

To build a logistic regression model, it is necessary to prepare the required tools, specifically by installing basic libraries such as NumPy, Pandas, scikit-learn, matplotlib, and seaborn in the Python programming language.

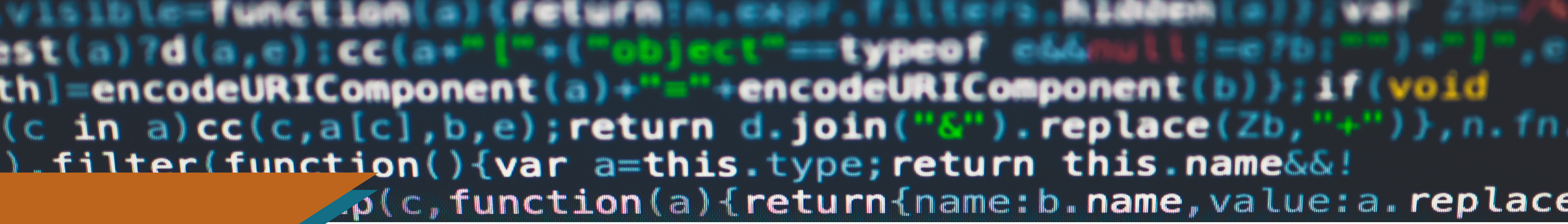
PREDICTING CHURN PROBABILITY





ANALYSIS OF FACTOR INFLUENCE

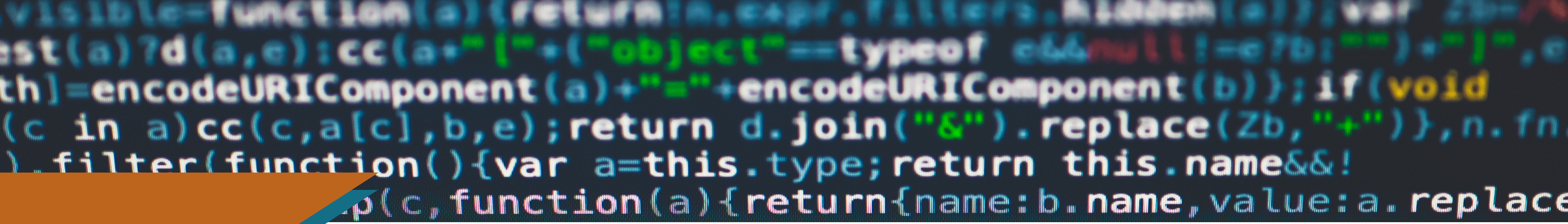
Logistic regression models can help determine the level of influence of specific factors (e.g., price, product quality) on customer purchasing decisions.
After evaluating the model, an additional step is needed: extracting the coefficients of each independent variable and printing them to gain the most comprehensive overview.



BENEFITS

Probability Prediction
ROC Curve and AUC
Impact Analysis
Handling Independent Variables

Understandable and Visual
Adjustable Decision Threshold
Easy Implementation and Integration
Handling Overfitting



LIMITATIONS

- Linear Assumption
- Sensitivity to Noise and Outliers
- Multicollinearity
- Potential for Overfitting
- Inability to Handle Nonlinear Phi Relationships
- Inappropriateness for Modeling Complex Relationships
- Threshold Decision Requirement

MEASUREMENT METRICS

CUSTOMER LIFETIME VALUE

Definition

Customer lifetime refers to the period from when a customer starts purchasing or using a company's products/services until they completely discontinue.

COHORT ANALYSIS

Cohort Analysis is a behavioral analysis method used to examine the effectiveness of marketing strategies for a specific target segment related to marketing costs and customer lifetime value.

App Launched ↓ % Active users after App Launches →

Cohort	Users	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Jan 25	1,098	100%	33.9%	23.5%	18.7%	15.9%	16.3%	14.2%	14.5%	Retention over user lifetime		12.1%
Jan 26	1,358	100%	31.1%	18.6%	14.3%	16.0%	14.9%	13.2%	12.9%			
Jan 27	1,257	100%	27.2%	19.6%	14.5%	12.9%	13.4%	13.0%	10.8%	11.4%		
Jan 28	1,587	100%	26.6%	17.9%	14.6%	14.8%	14.9%	13.7%	11.9%			
Jan 29	1,758	100%	26.2%	20.4%	16.9%	14.3%	12.7%	12.5%				
Jan 30	1,624	100%	26.4%	18.1%	13.7%	15.4%	11.8%					
Jan 31	1,541	100%	23.9%	19.6%	15.0%	14.8%						
Feb 01	868	100%	24.7%	16.9%	15.8%							
Feb 02	1,143	Retention over product lifetime		18.5%								
Feb 03	1,253											
All Users	**13,487**	100%	27.0%	19.2%	15.4%	14.9%	14.0%	13.3%	12.5%	13.1%	12.2%	12.1%

CLV CALCULATION FORMULA IN THE TRADITIONAL APPROACH

$$CLV = \$M\left[\frac{r}{1+d-r}\right]$$

Where

- M: Annual profit rate per customer
- r: Annual retention rate
- d: Discount rate

CLV CALCULATION FORMULA BASED ON INITIAL PROFIT

$$CLV_{alternative} = \$M\left[\frac{1+d}{1+d-r}\right]$$

This formula incorporates an initial cash flow M, which is the certain amount received at the beginning of a customer relationship.

RETENTION RATE

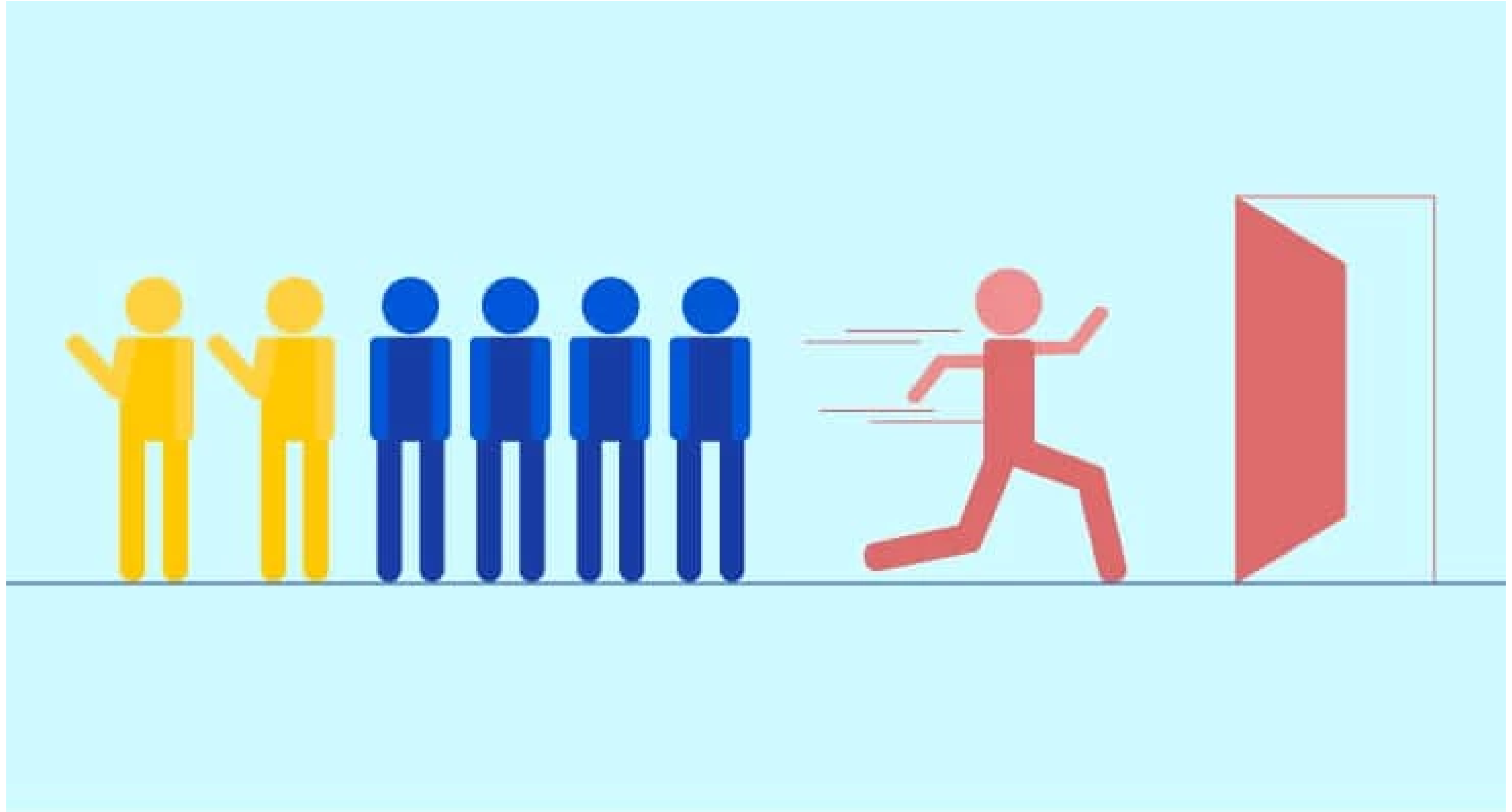
Customer retention is a crucial driver for a company's financial success.

To measure the economic benefits of increasing the retention rate:

- Spreadsheet Model
- Cash Flow Model
- Direct Assumption Analysis on CLV Formula

CHURN RATE

Churn Rate is the percentage of customers who stop using your company's products or services within a specified period.



CHURN RATE FORMULA

CHURN FORMULA

customers who left

to get %

$$\frac{120}{40\,000 + 1\,250} \times 100 = 0{,}29\%$$

customers at the beginning of the period (month)

new customers aquired during that period (month)

ROLE OF CHURN RATE

Evaluate Growth and Profitability

Customer Behavior Analysis

Identify Key Customers

Evaluate Marketing Campaigns

III. DATA PREPARATION

COLLECTION DATA

- Overview: detailed information about customer behavior and service status in the telecommunications industry.
- Size: 977.5 kB
- Type: CSV
- Column: 21 columns (features). The "Churn" column is target.
- Rows: The raw data contains 7043 rows (customers).

EXPLORATORY DATA ANALYSIS

EDA

Import required libraries

```
#import the required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.ticker as mtick
import matplotlib.pyplot as plt
%matplotlib inline
```

STATISTIC ANALYSIS

```
[ ]  telco_base_data.shape
```

```
(7043, 21)
```

```
[ ]  telco_base_data.columns.values
```

```
array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

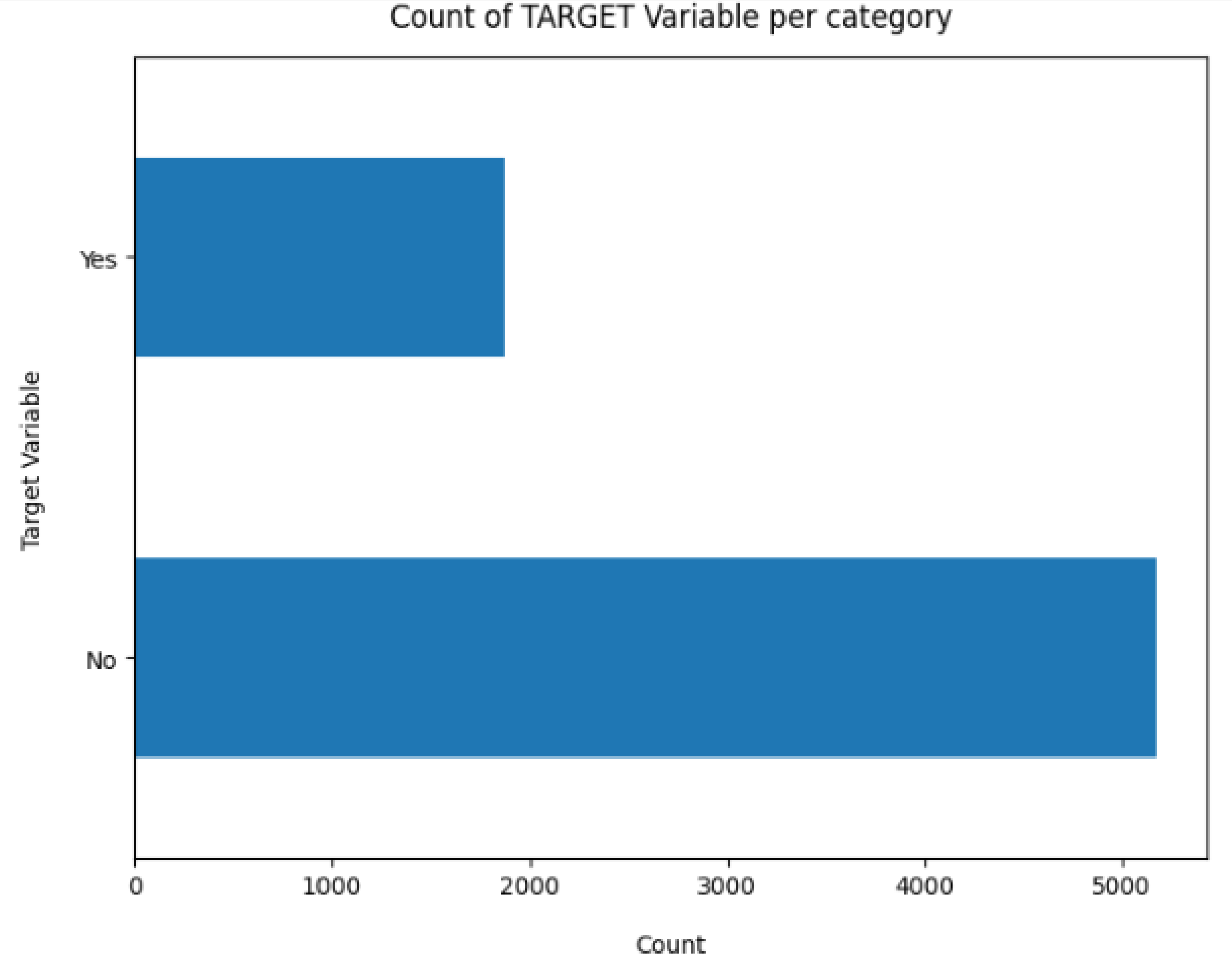
This dataframe has 7043 rows and 21 columns

STATISTIC ANALYSIS

Columns are mostly objects and do not have null values

```
[ ] telco_base_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

STATISTIC ANALYSIS



STATISTIC ANALYSIS

Highly imbalanced data, ratio = 73:27

```
100*telco_base_data['Churn'].value_counts()/len(telco_base_data['Churn'])
```

```
No     73.463013
Yes    26.536987
Name: Churn, dtype: float64
```

```
telco_base_data['Churn'].value_counts()
```

```
No     5174
Yes    1869
Name: Churn, dtype: int64
```

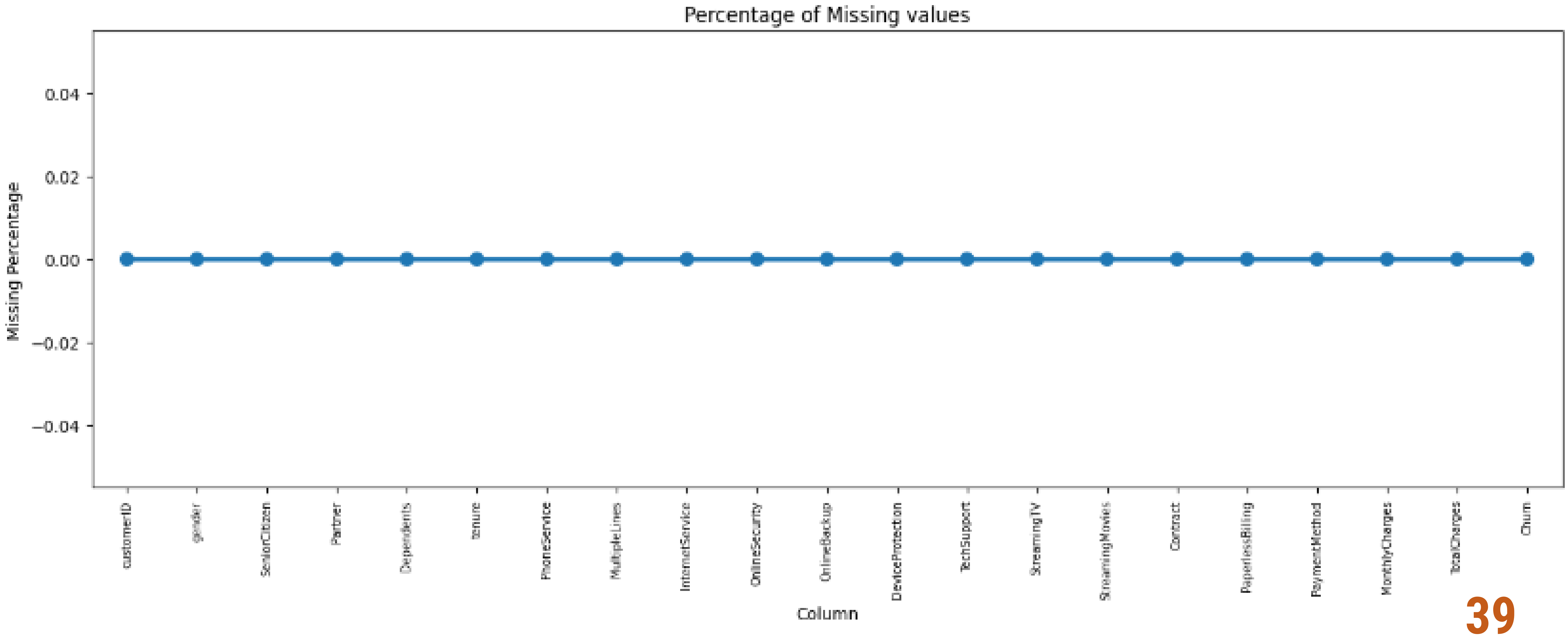
Data Cleaning

```
# show the column with null values
telco_base_data.isnull().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

No Null cases were recorded

Data Cleaning



Data Cleaning

```
telco_data.TotalCharges = pd.to_numeric(telco_data.TotalCharges, errors='coerce')
telco_data.isnull().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

The % of these records compared to the total dataset is very low

Data Cleaning

Divide customers into segments based on tenure (length of tenure)

```
# Group the tenure in bins of 12 months
labels = ["{0} - {1}".format(i, i + 11) for i in range(1, 72, 12)]
telco_data['tenure_group'] = pd.cut(telco_data.tenure, range(1, 80, 12), right=False, labels=labels)
telco_data['tenure_group'].value_counts()
```

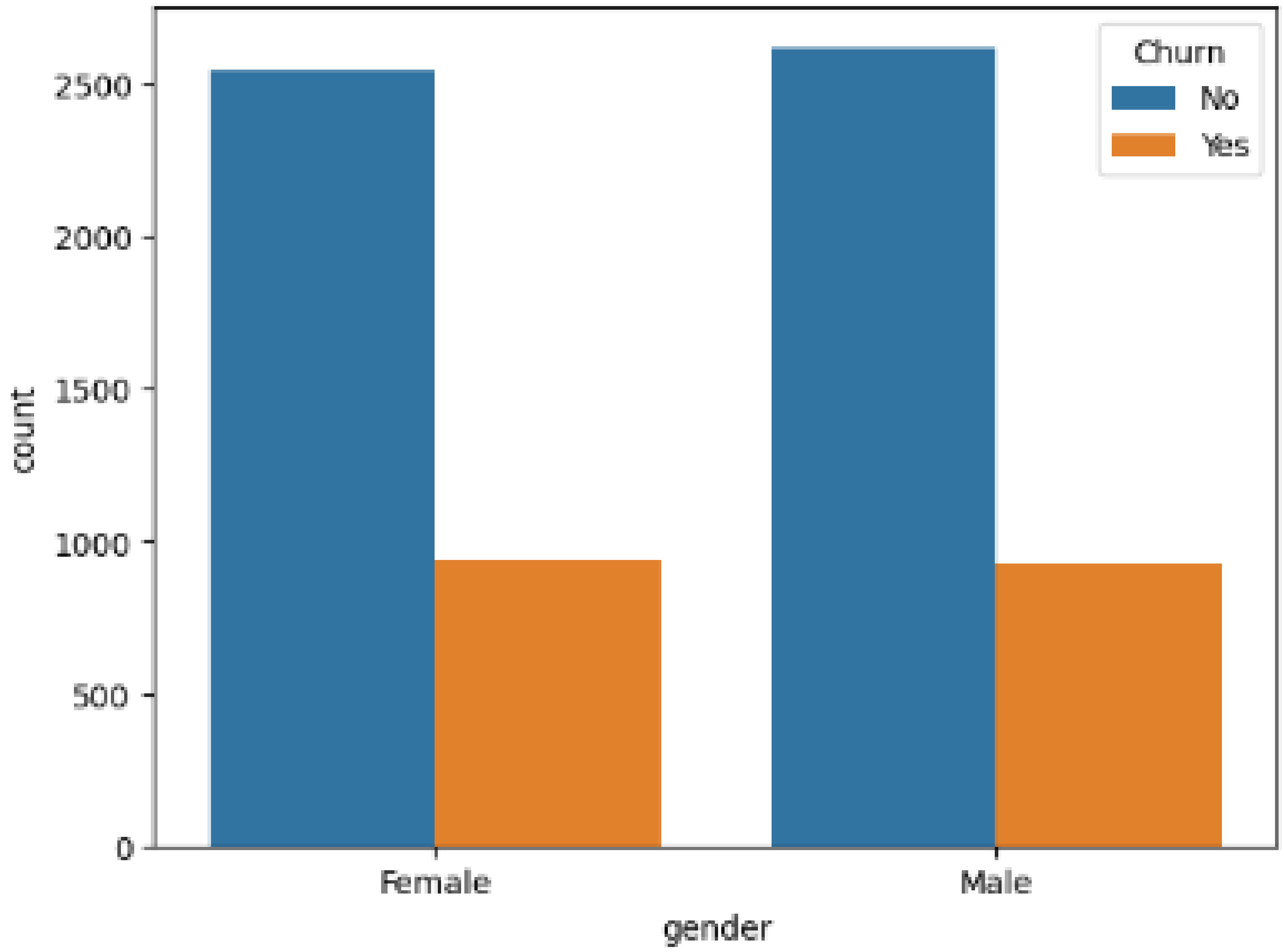
```
1 - 12      2175
61 - 72     1407
13 - 24     1024
25 - 36      832
49 - 60      832
37 - 48      762
Name: tenure_group, dtype: int64
```

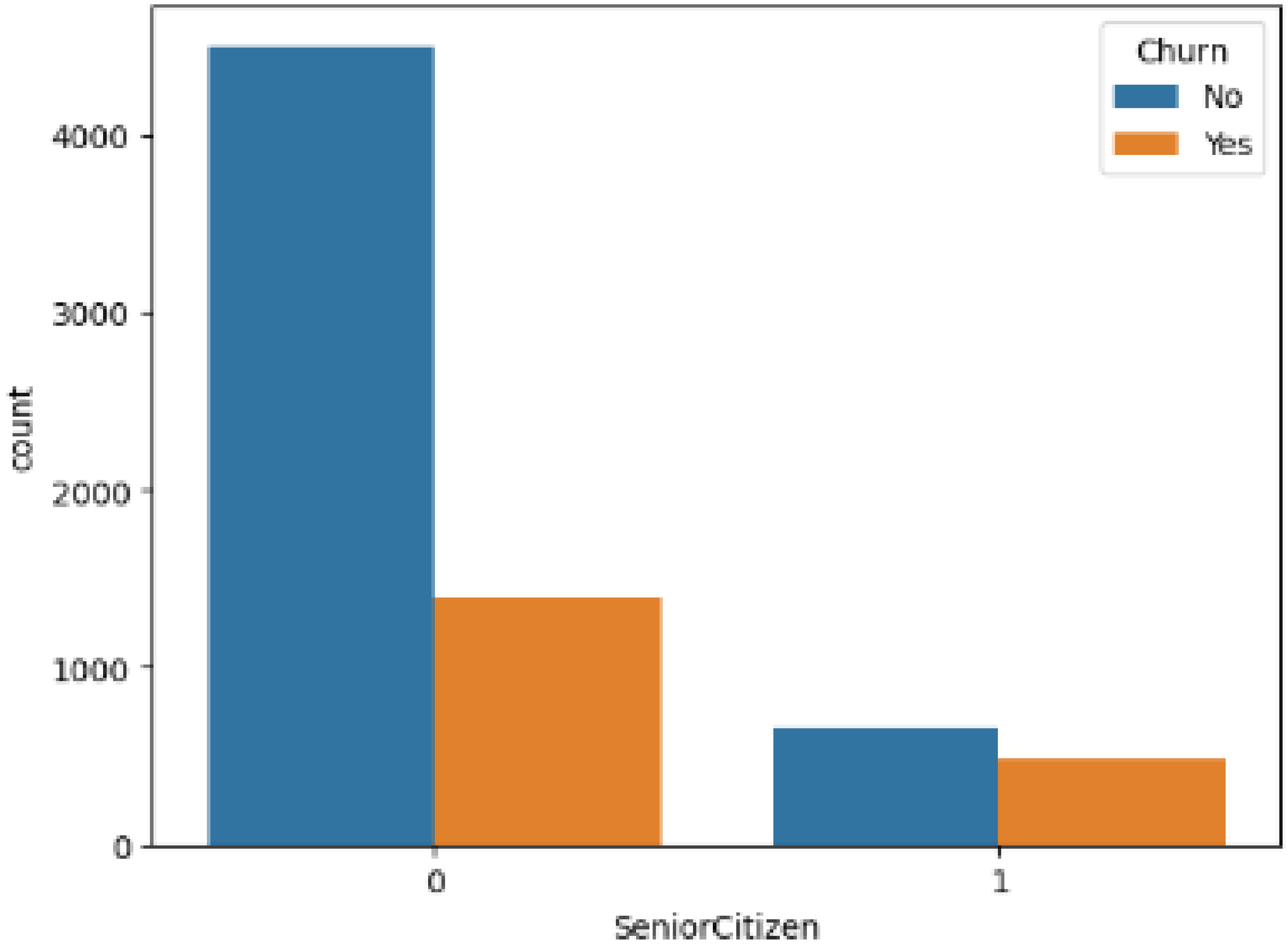
Data Cleaning

Delete unnecessary columns to continue processing the data file

```
#drop column customerID and tenure
telco_data.drop(columns= ['customerID','tenure'], axis=1, inplace=True)
telco_data.head()
```

Data Exploration

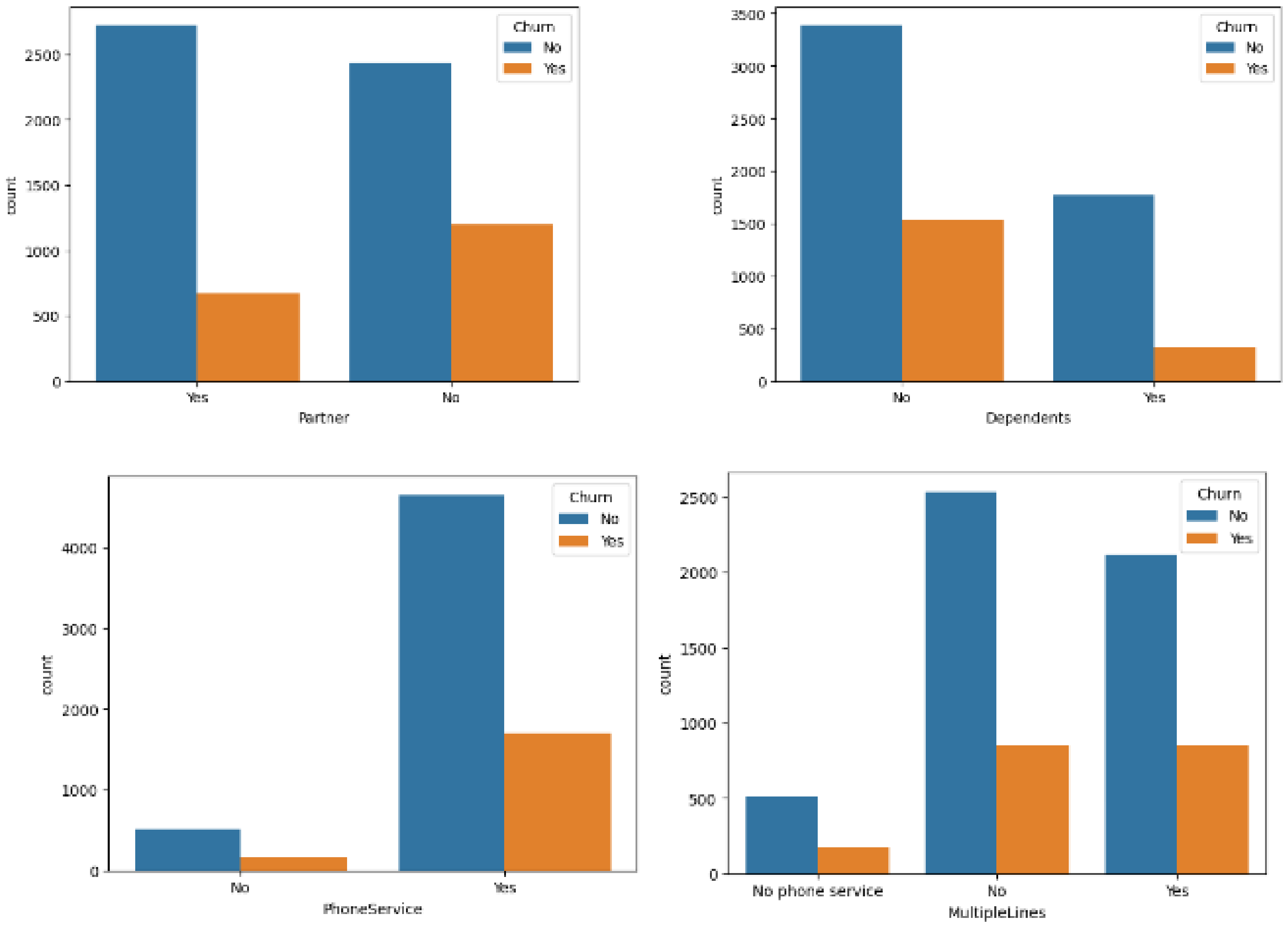




Histogram distribution of individual predictors based on Churn

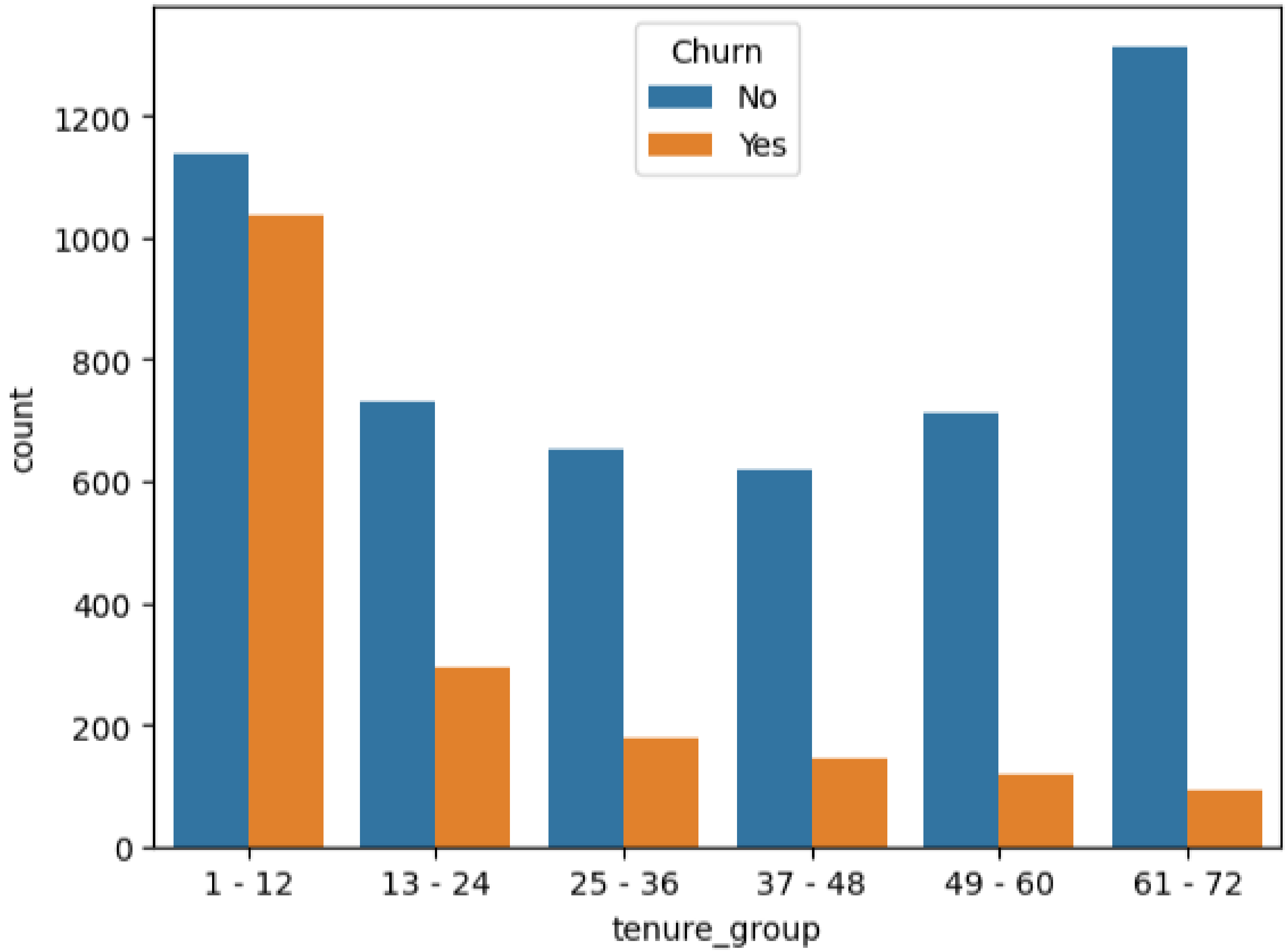
Data Exploration

The influence of variables Dependents, PhoneService, MultipheLines, Partner

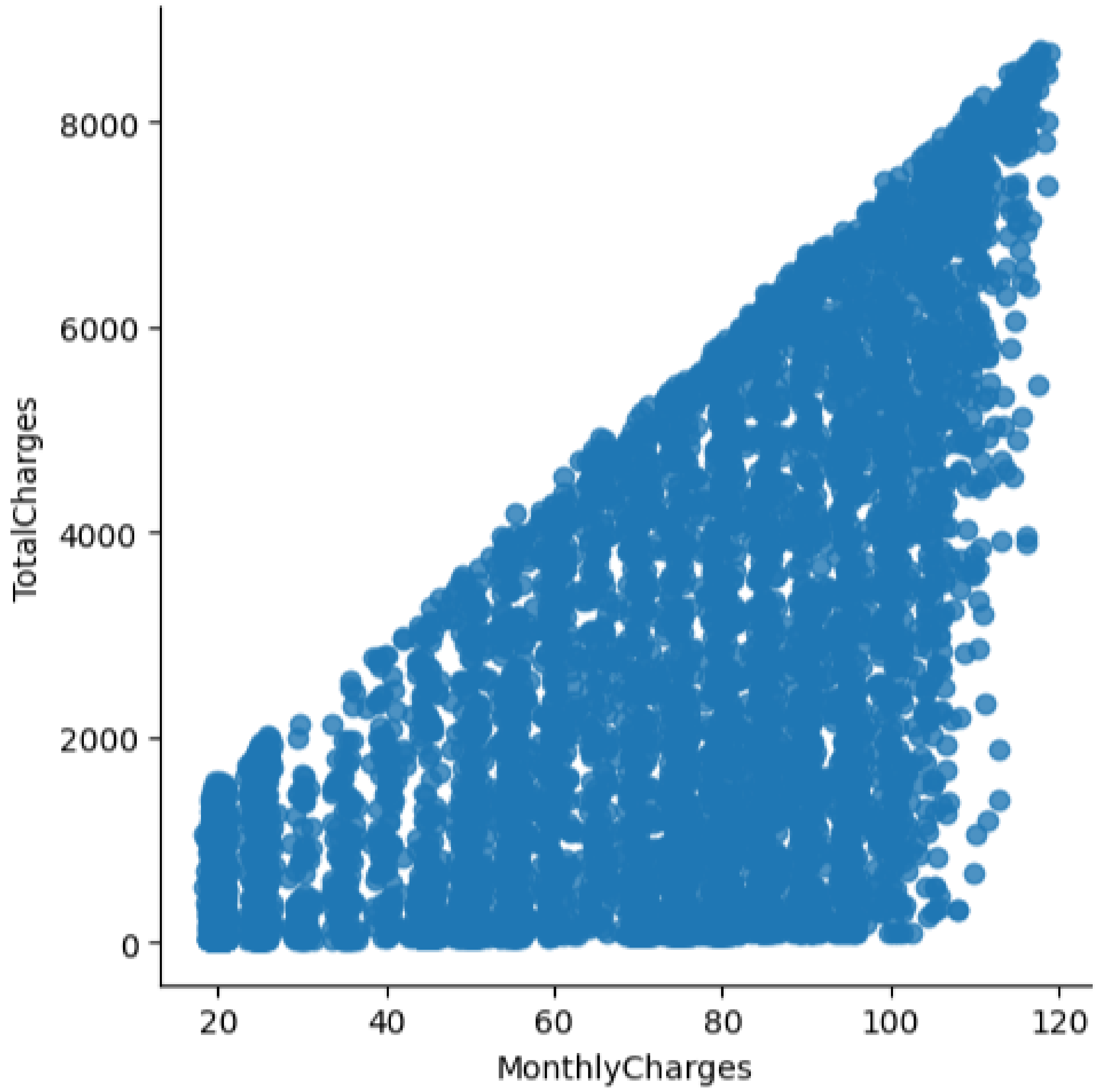


Data Exploration

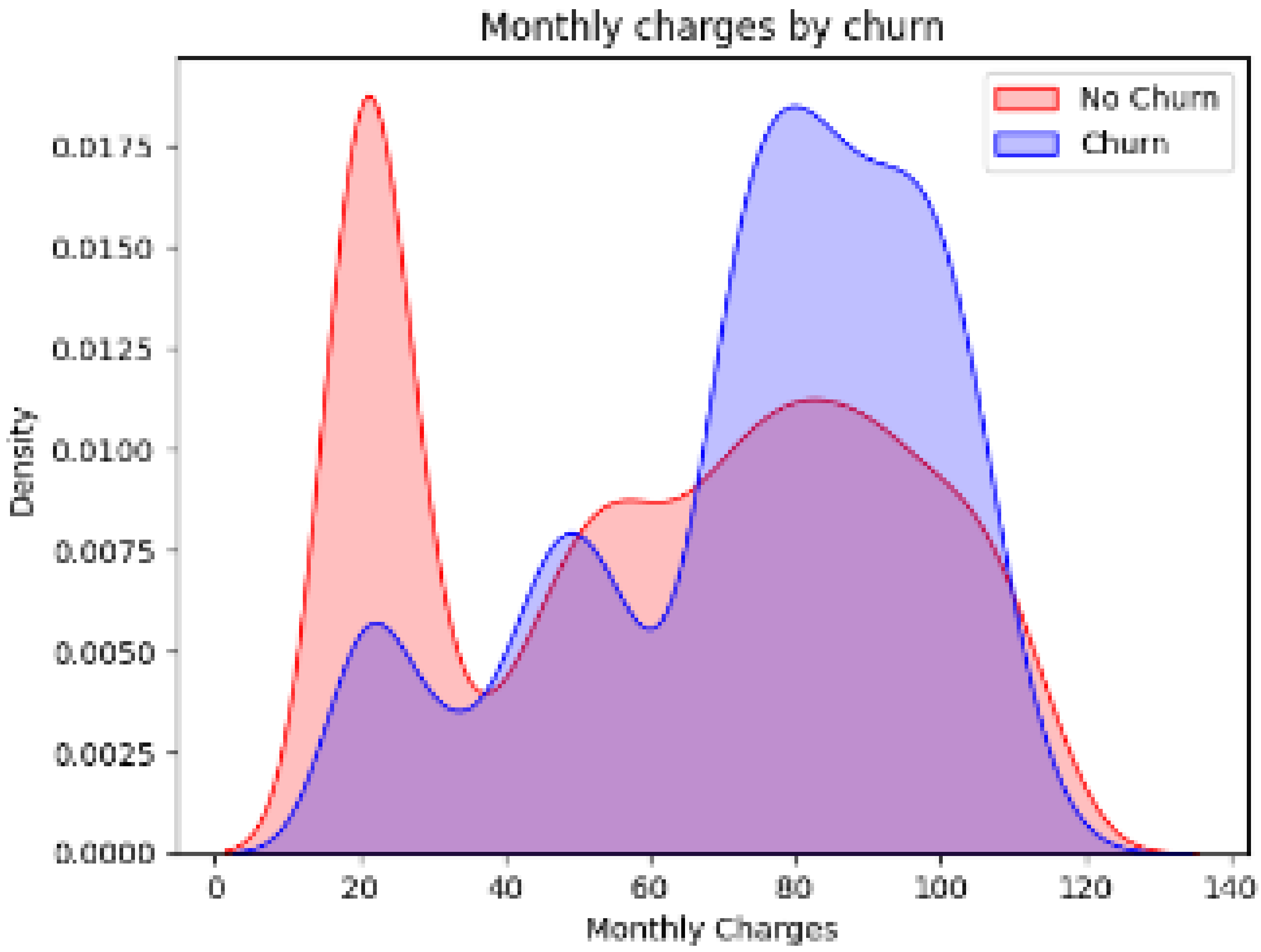
There is a clear change over time

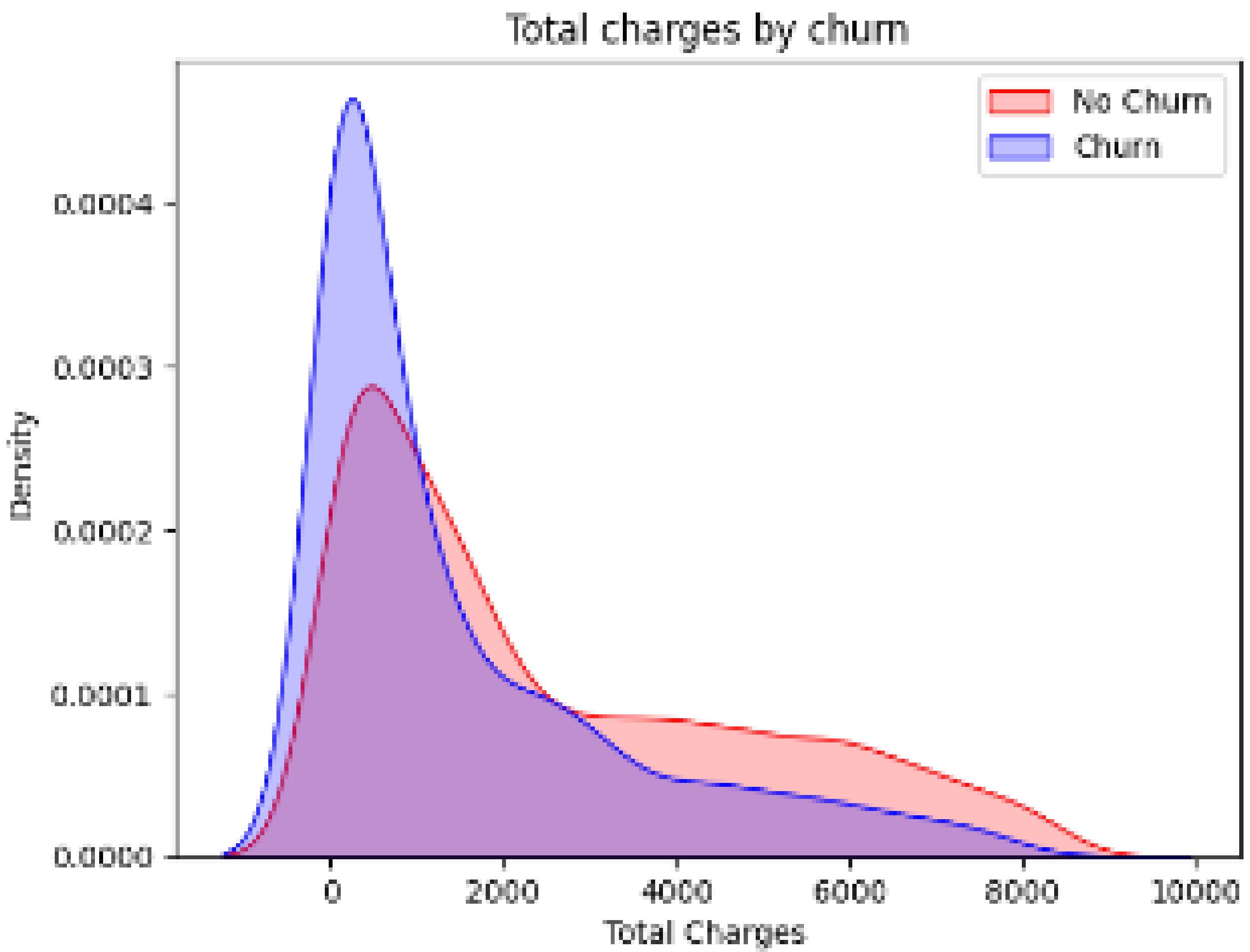


Relationship between Monthly Charges and Total Charges



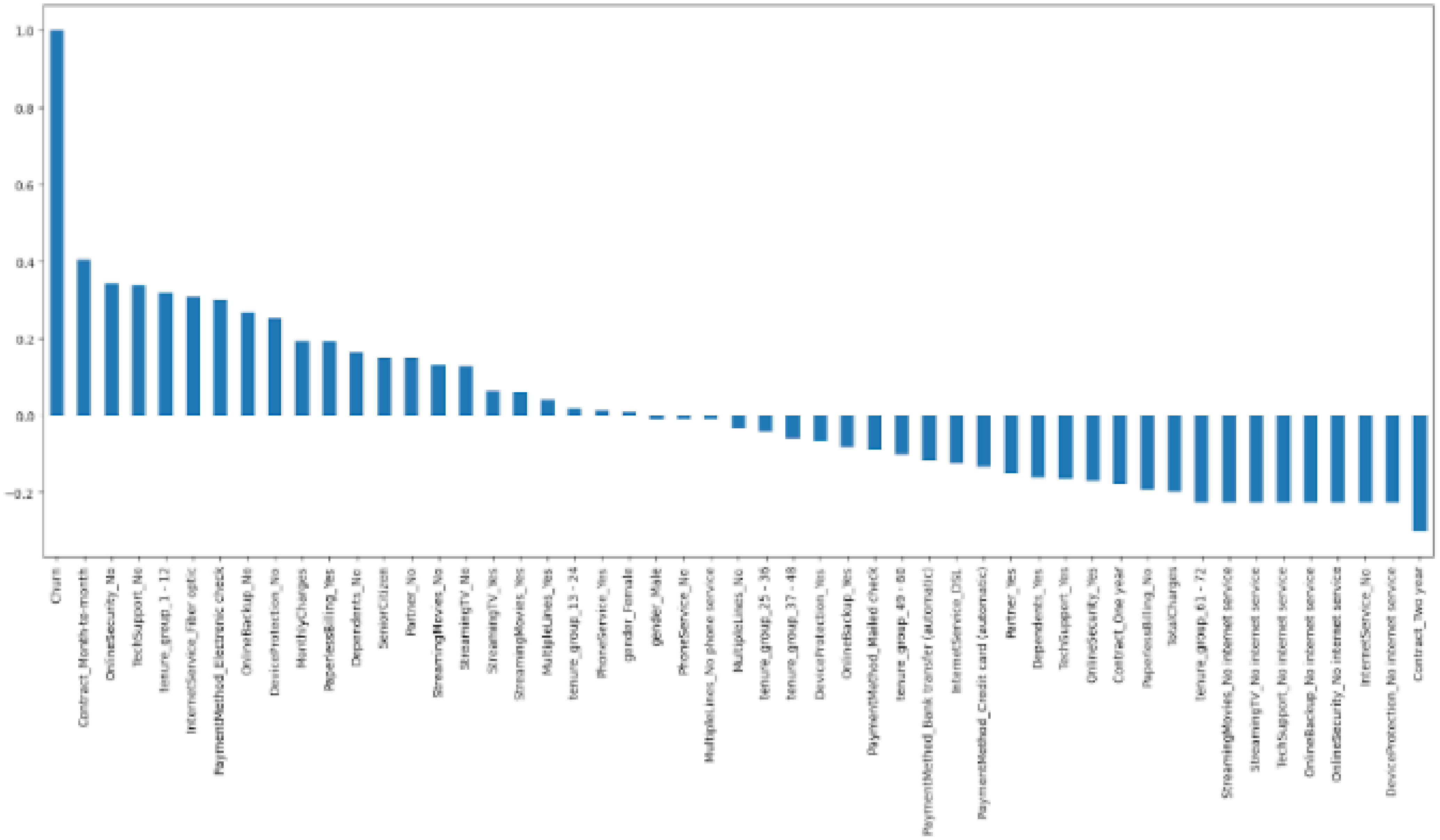
Relationship between Monthly Charges and Total Charges





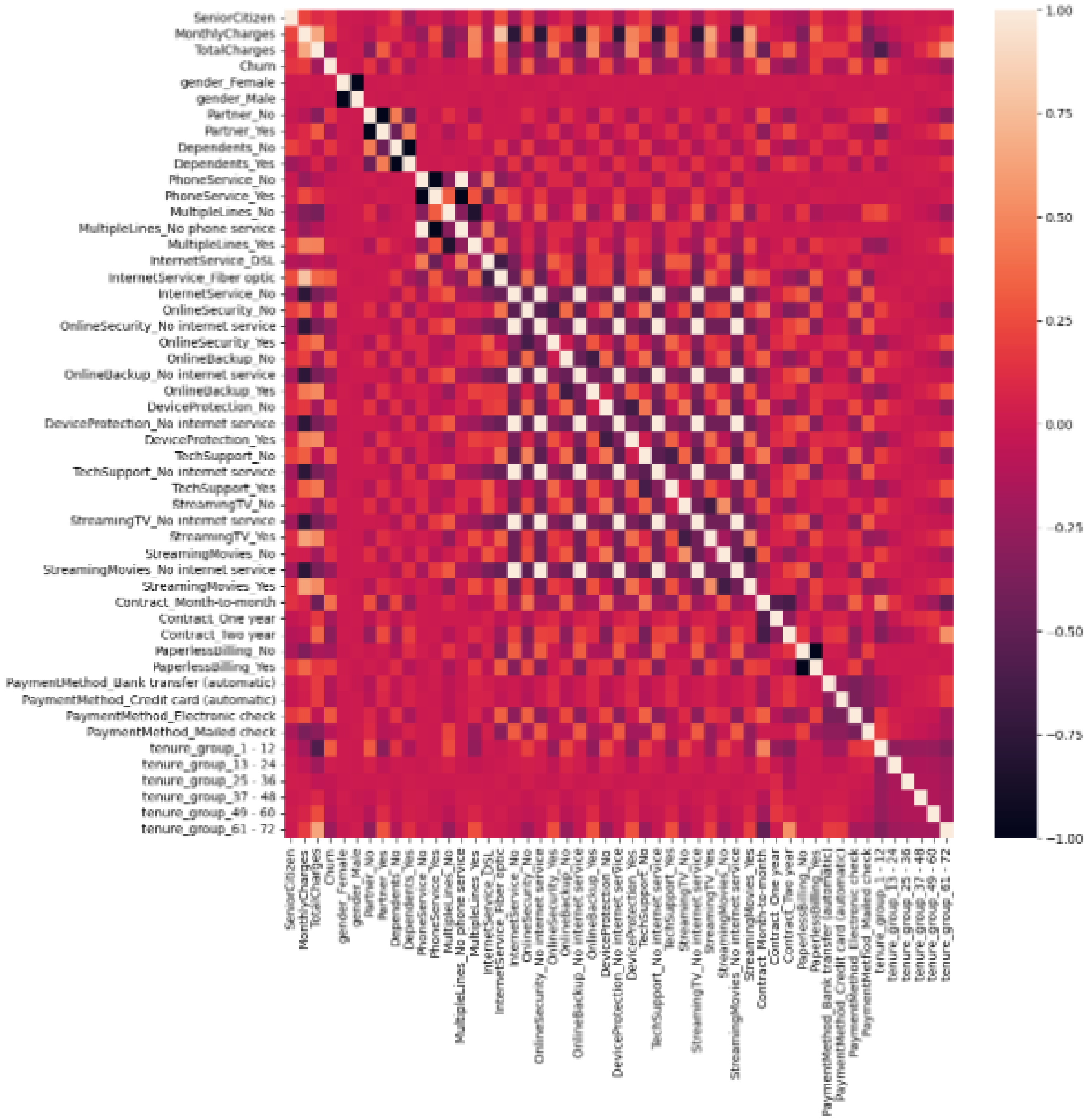
Higher churn with lower Total Fees

Relationship between Monthly Charges and Total Charges



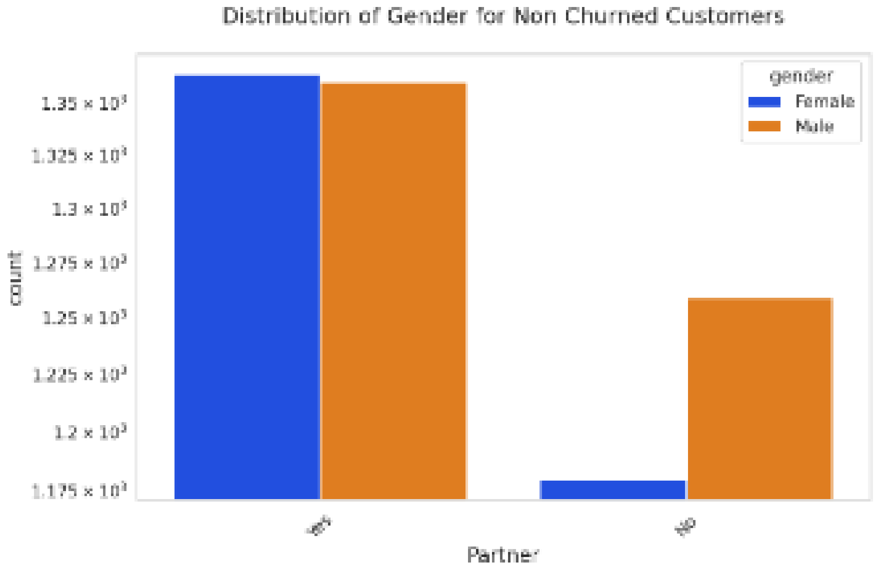
Relationship of all predictors to 'Churn'

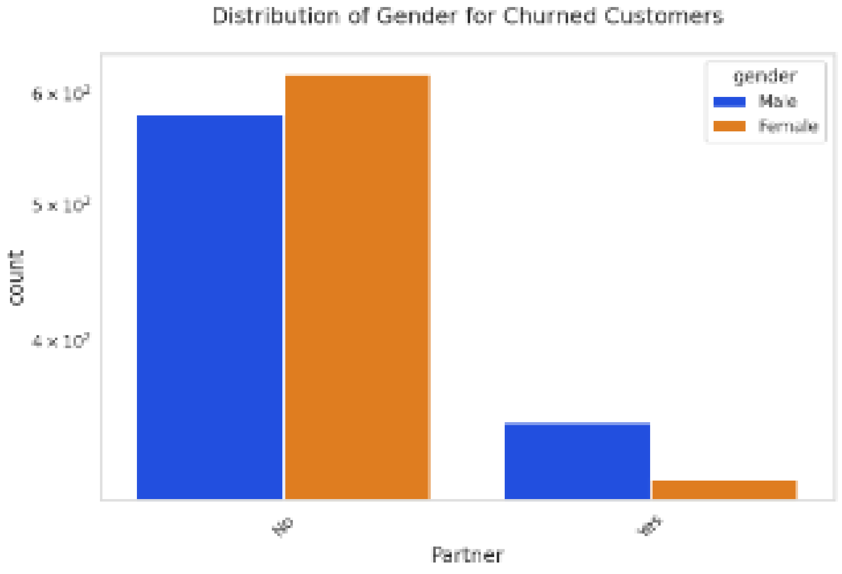
Correlation

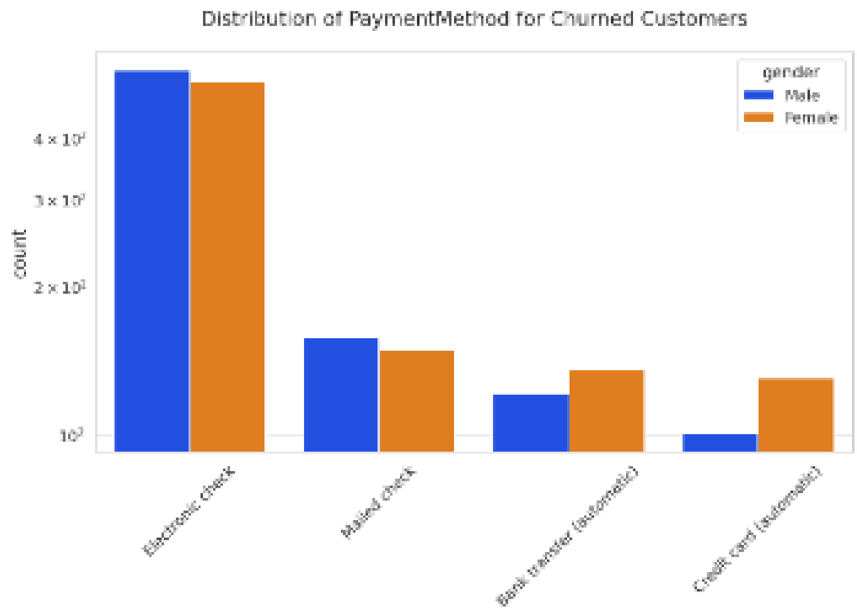


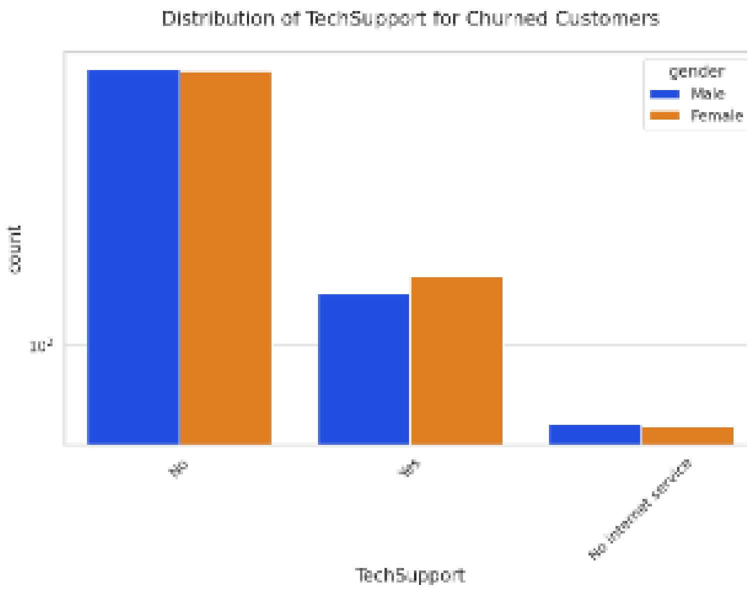
Correlation heatmap

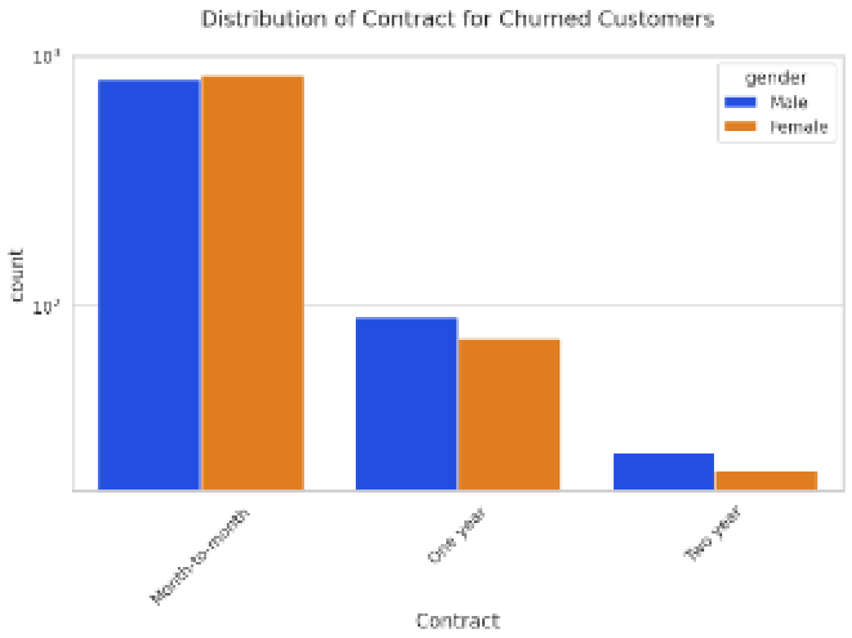
Bivariate Analysis

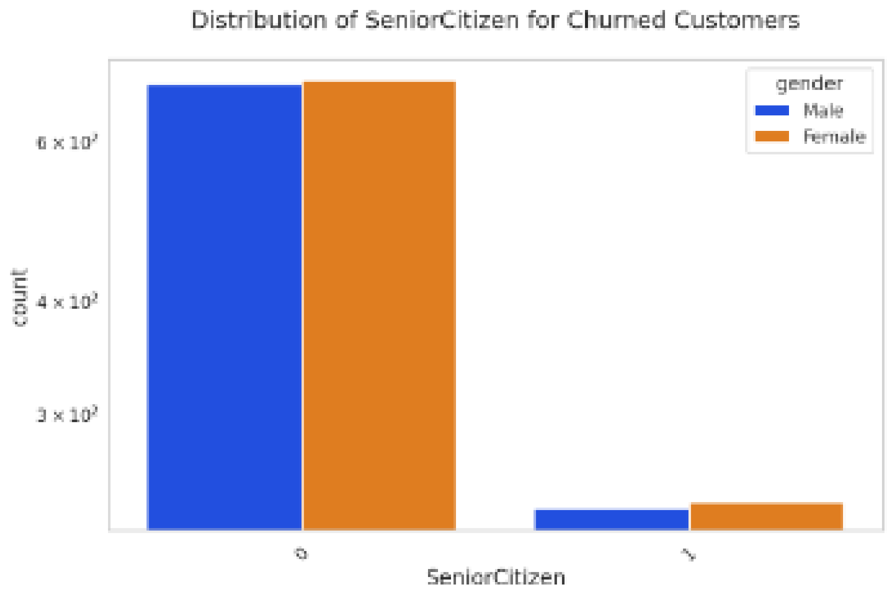












IV
CUSTOMER ANALYSIS

DECISION TREE CLASSIFIER

01

Create a Decision Tree Classifier model (Imbalanced data)

02

Train the Decision Tree model on the original training data

03

Predict the target variable for the test set

04

Calculate the accuracy of the model on the test set

Model Prediction

Decision Tree Classifier

```
print(classification_report(y_test, y_pred, labels=[0,1]))
```

```
              precision    recall  f1-score   support

           0       0.85      0.86      0.85      1034
           1       0.59      0.57      0.58       373

    accuracy                           0.78      1407
   macro avg       0.72      0.71      0.72      1407
weighted avg       0.78      0.78      0.78      1407
```

Step 5: Summarize model evaluations on the unbalanced test set

DECISION TREE CLASSIFIER

06

Call SMOTEENN to handle imbalanced data sets

07

Divide into training set and test set from balanced data

08

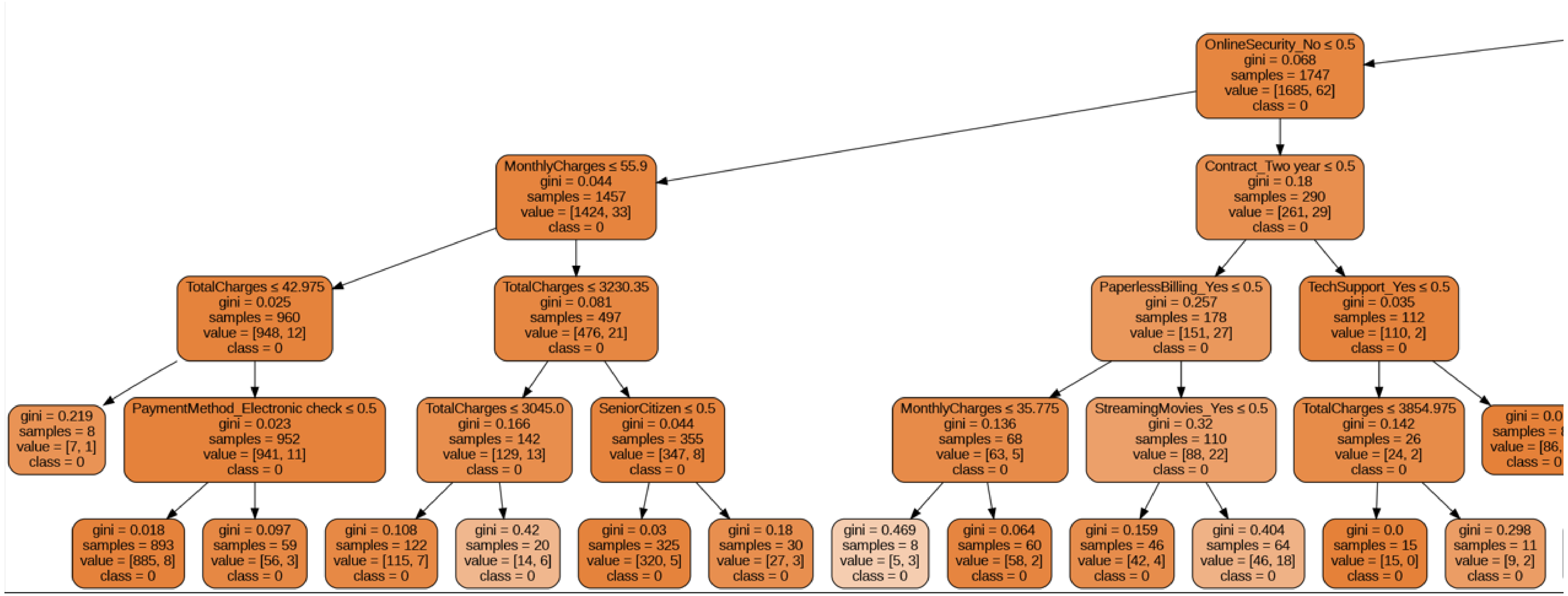
Create a new Decision Tree model for the balanced data

09

Run the new model

Model Prediction

Decision Tree Classifier



Step 10: Display the error matrix of the model on the balanced test set

RANDOM FOREST CLASSIFIER

01

Model Creation and Training

02

Model Prediction and Evaluation

03

Using SMOTEENN for Imbalanced Data

04

Training a Random Forest Model on Augmented Data

4.1. odel Prediction

Random Forest Classifier

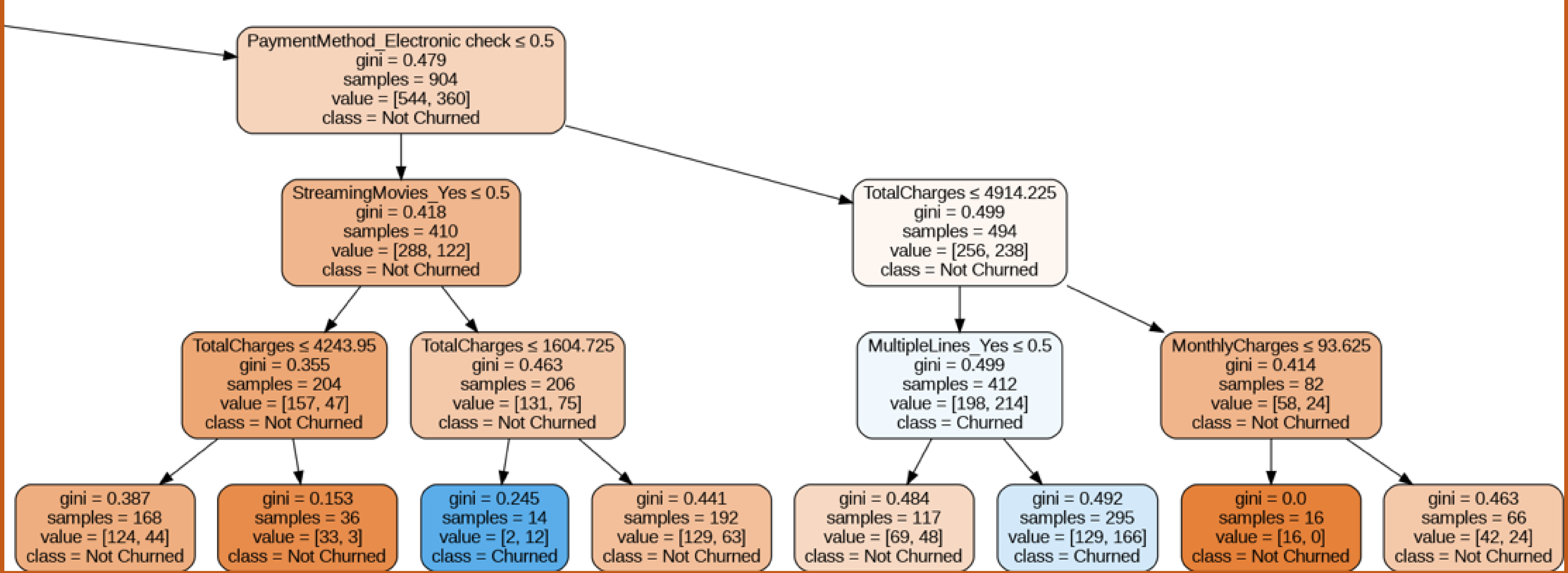
```
print(model_score_r1)
print(metrics.classification_report(yr_test1, yr_predict1))
```

```
0.9316823228010248
              precision    recall  f1-score   support

           0       0.95      0.89      0.92       508
           1       0.92      0.97      0.94       663

    accuracy                           0.93      1171
   macro avg       0.93      0.93      0.93      1171
weighted avg       0.93      0.93      0.93      1171
```

Step 5: Run the new model

Model Prediction

Random Forest Classifier

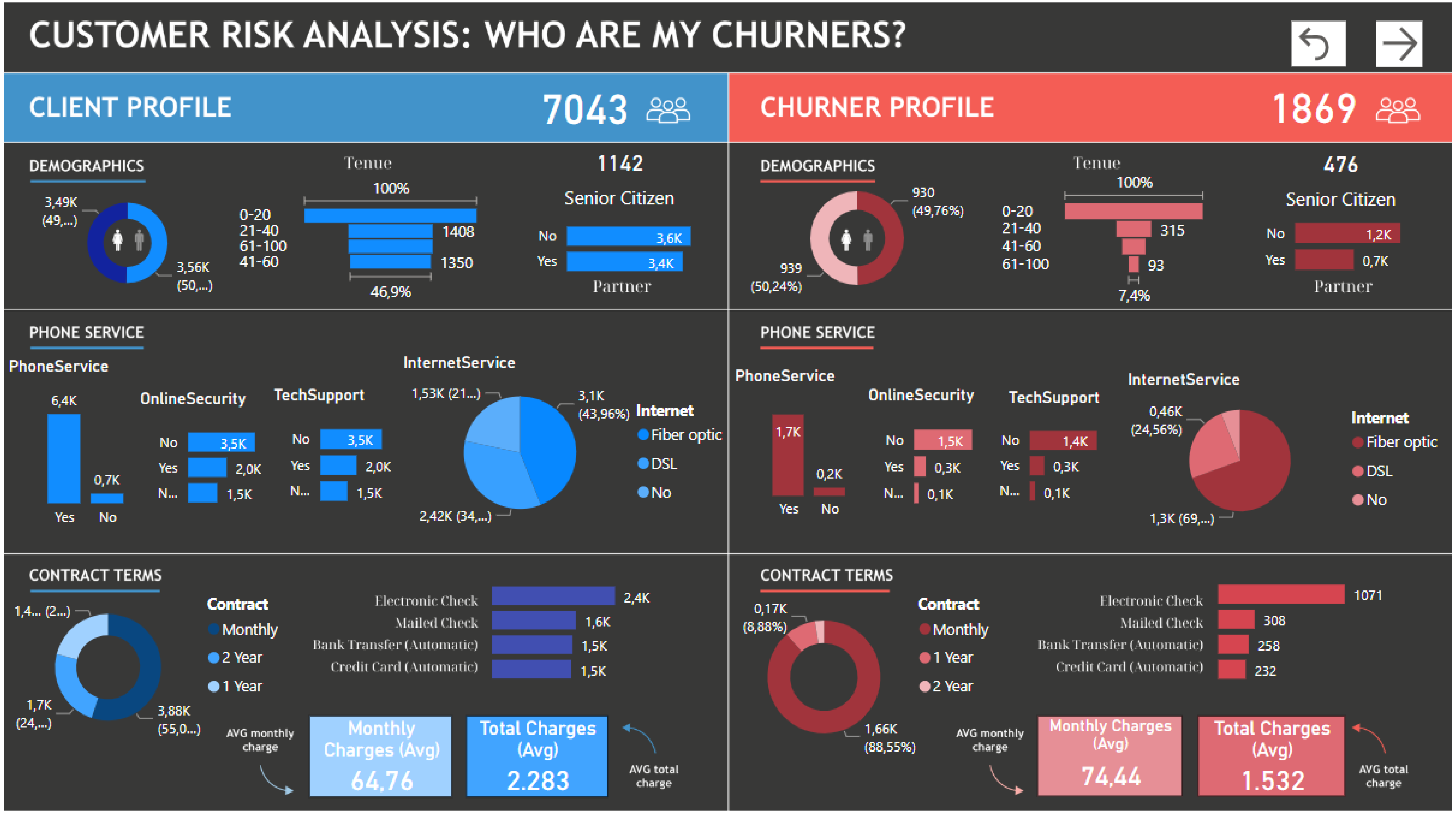


Results

VISUALIZE WITH POWER BI

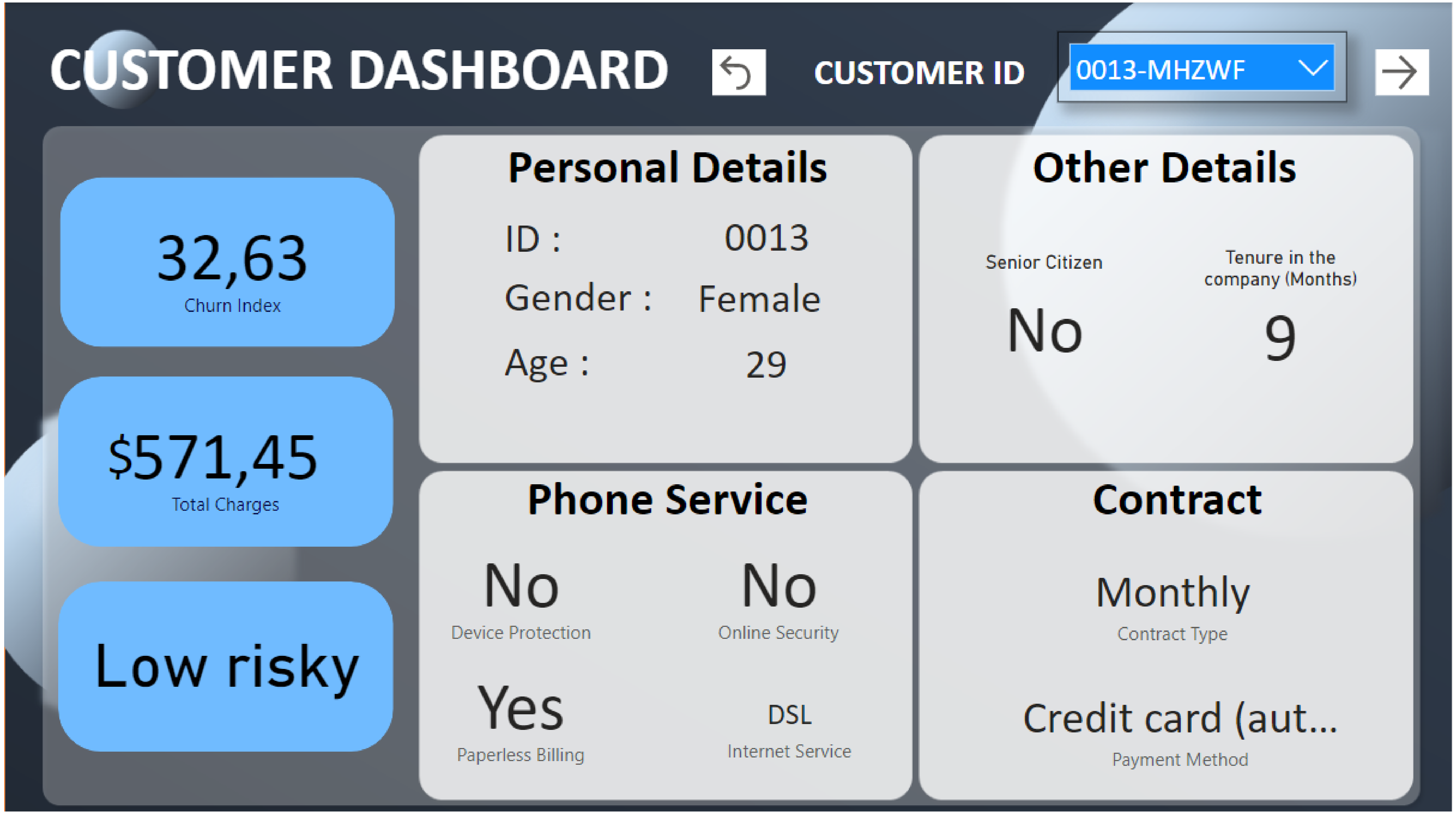
Main screen overview of reports

VISUALIZE WITH POWER BI



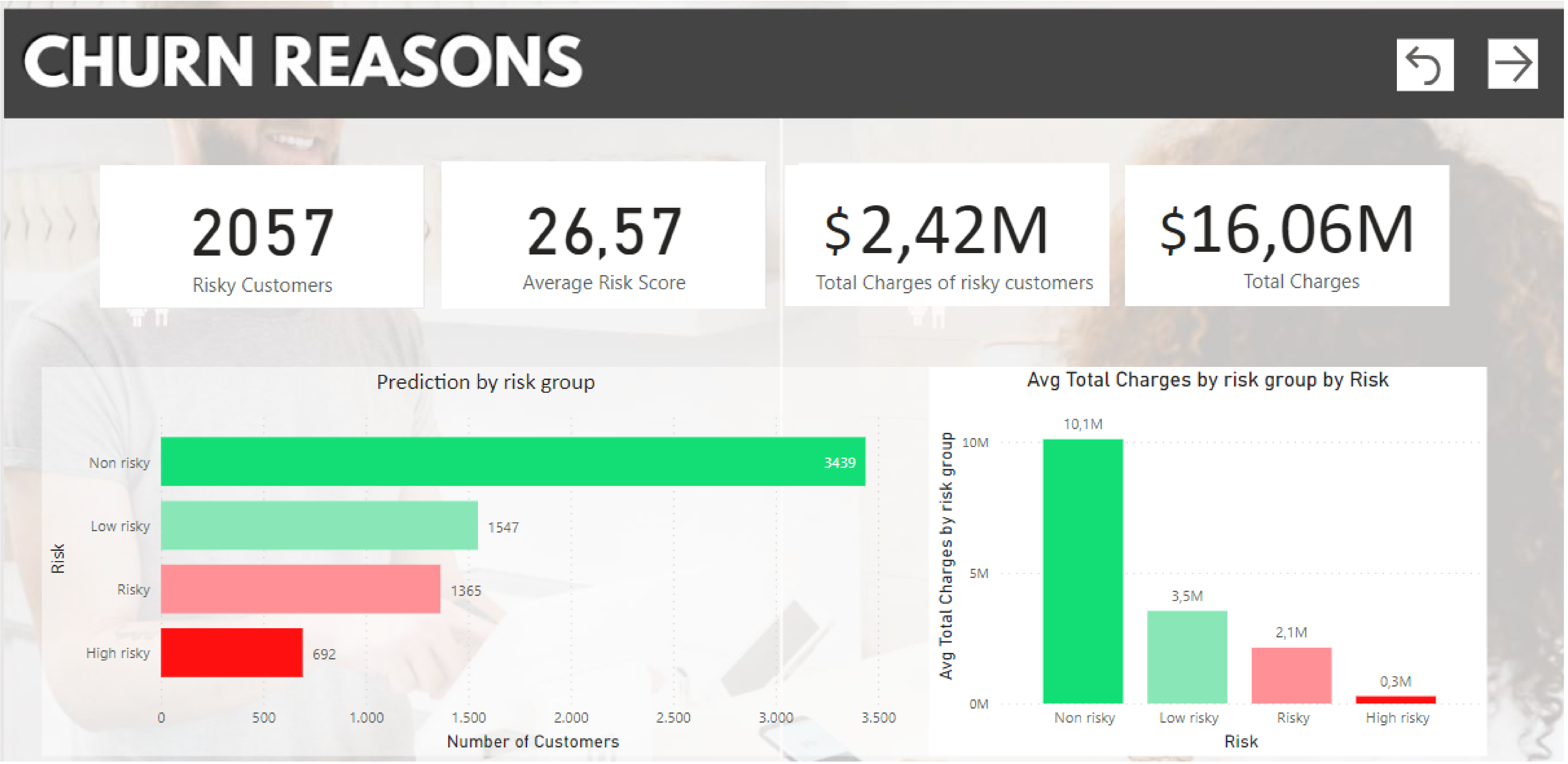
Analyze customer churn rate

VISUALIZE WITH POWER BI



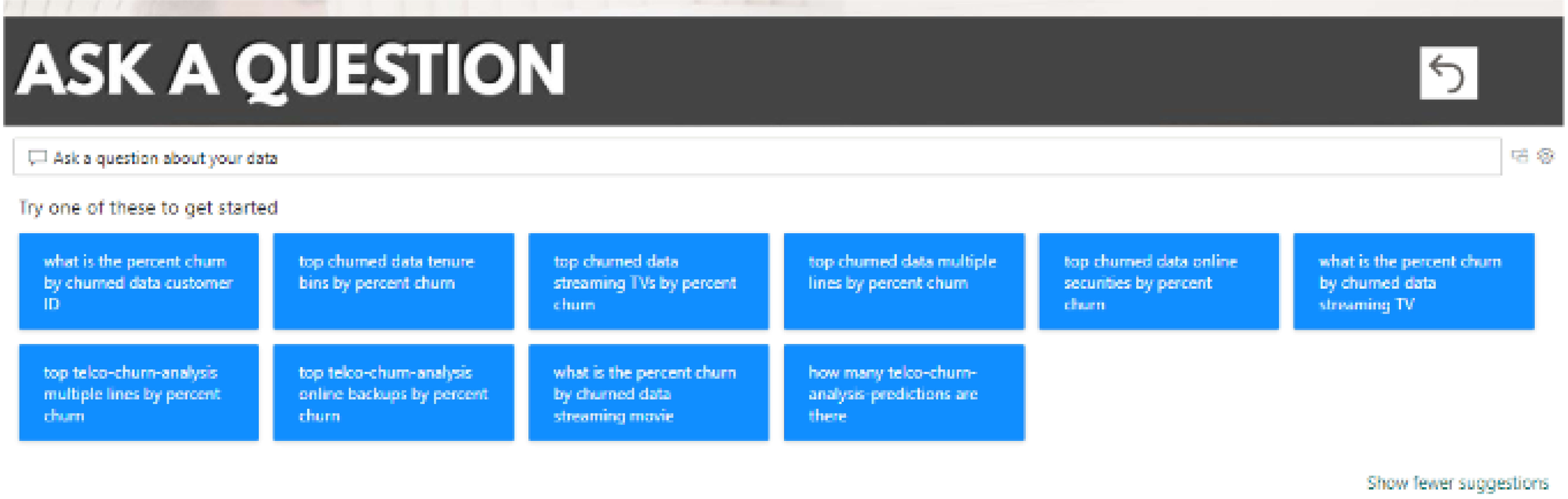
Information and specific assessment of each customer's situation

VISUALIZE WITH POWER BI



Overview of customer churn reasons

VISUALIZE WITH POWER BI



Statistics on customer churn issues

VISUALIZE WITH POWER BI

Top risk categories by percent churn

PRESENTATION - 2023

THANK YOU

FOR YOUR ATTENTION