Name: Tanu Verma

Reg no: 12101448

College: Lovely Professional University

Topic: Telcom Churn Prediction

# INTRODUCTION

## Definition:

Churn Customer refers to the number of existing customers who may leave the service provider over a given period. These customers can be called as churners. The main aim of churn is to predict the churnable customers at the earliest, to identify the reason for churning. The primary goal of churn analysis is to identify and anticipate churnable consumers as soon as possible. This will help us to rectify the issues of the customer. This will be helpful to satisfy the customer needs and will continue to use that service. This will help to meet the needs of the customers, and they will continue to utilize the service.

### Objective of the Report:

The primary objective of this report is to predict customer churn in the telecom industry using machine learning techniques. Churn, in the context of telecommunications, refers to the phenomenon where customers discontinue their subscription or switch to a competitor. Accurately predicting churn allows telecom companies to identify at-risk customers and implement targeted retention strategies, ultimately reducing customer attrition and increasing profitability.

# Data Description

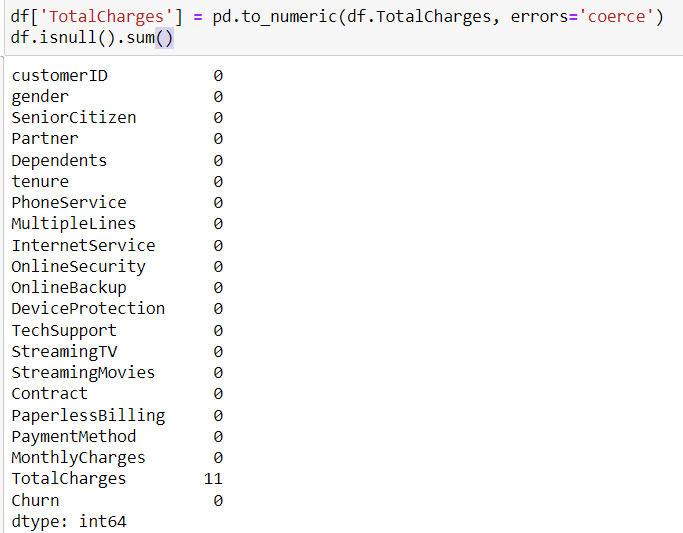
## Overview of the dataset used

* Source of the data: https://www.kaggle.com/datasets/blastchar/telco-
* Description of features/variables:

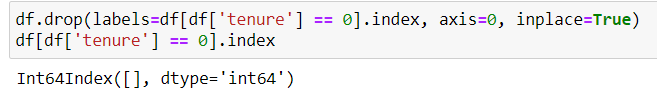
## Data pre-processing steps

* Handling missing values

The dataset used for the telecom churn prediction analysis consists of 7,043 customer records. Ensuring data quality is crucial for building reliable machine learning models, and part of this process involves handling missing values appropriately.Upon examining the dataset for missing values, it was found that there were 11 missing values in the “TotalCharges” column. These missing values were identified using the following code:



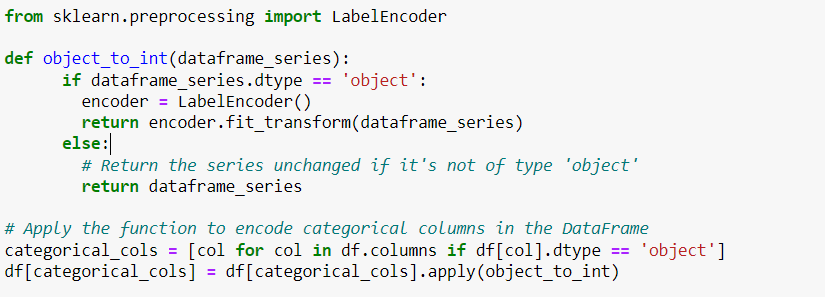
Additionally, it was observed that some records had a tenure value of 0. These records were also dropped from the dataset to avoid inconsistencies, as a tenure of 0 might indicate new customers who have not yet been billed, which could skew the analysis. The removal of these records was carried out using the following code:



* Encoding categorical variables

In the telecom churn prediction analysis, it was essential to encode the categorical variables into numerical values to make them suitable for machine learning models. Categorical variables often contain non-numeric data which needs to be converted into a numeric format before they can be utilized by most machine learning algorithms. To achieve this, the LabelEncoder from the sklearn.preprocessing module was used.

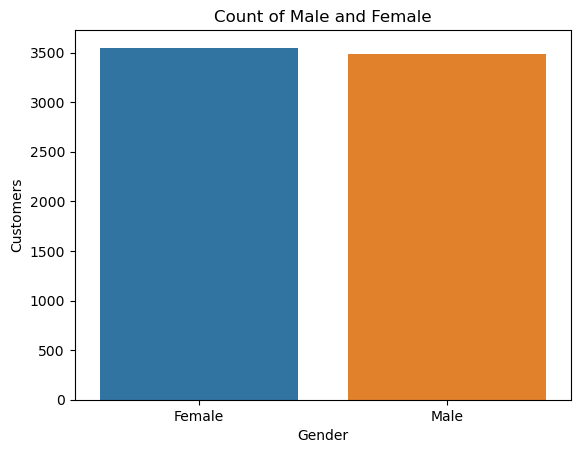
A custom function object\_to\_int was defined to transform the categorical columns in the DataFrame. This function checks if the data type of a series is 'object', which indicates a categorical variable, and then applies LabelEncoder to convert it into numerical values. If the data type is not 'object', the series is returned unchanged.



# Exploratory Data Analysis (EDA)

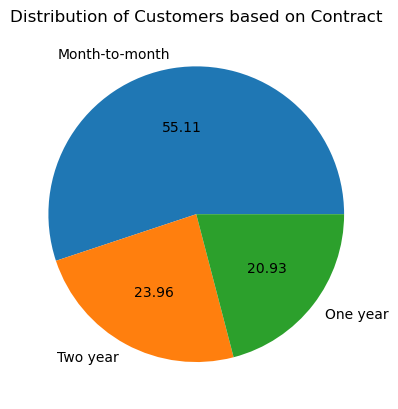
## Gender Distribution

Upon analyzing the dataset, it was found that the number of male and female customers is almost equal. This balanced distribution suggests that any gender-specific patterns observed in churn behavior can be considered representative of the overall customer base, without significant bias towards either gender.



#### Distribution of Customers Based on Contract

Analyzing the distribution of customers based on their contract type is crucial for understanding customer preferences and identifying potential factors that may influence churn. The provided pie chart illustrates the distribution of customers across three different contract types: Month-to-month, One year, and Two year

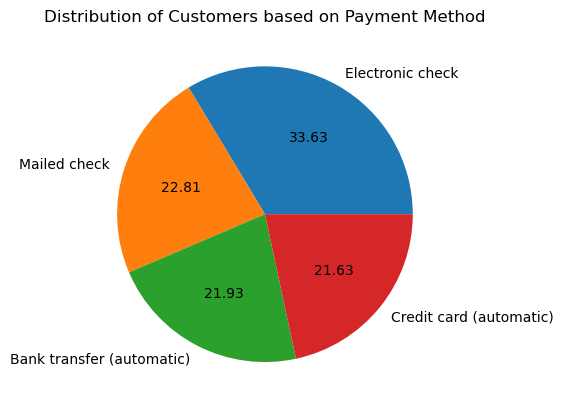
.

The pie chart shows the following distribution:

* **Month-to-month**: 55.11%
* **One year**: 20.93%
* **Two year**: 23.96%

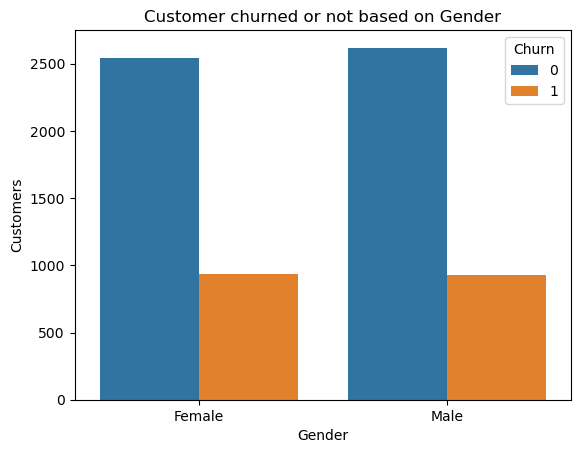
#### Distribution of Customers Based on Payment Method

Understanding the distribution of customers based on their payment methods provides valuable insights into customer preferences and behavior. This information can help telecom companies tailor their services and address potential churn more effectively. The pie chart below illustrates the distribution of customers across four different payment methods: Electronic check, Mailed check, Bank transfer (automatic), and Credit card (automatic).

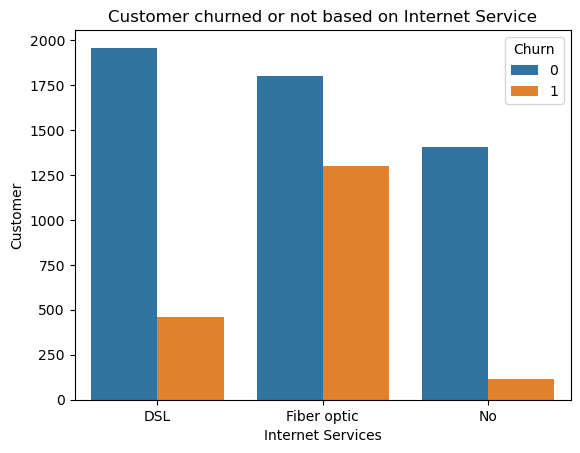


## Churn Analysis Based on Gender

1. **Non-Churned Customers**:
   * Both female and male customers show a similar pattern in the number of non-churned customers.
   * Approximately 2,500 female and 2,500 male customers did not churn, indicating a nearly equal retention rate across genders.
2. **Churned Customers**:
   * The number of churned customers is also similar between genders.
   * There are about 1,000 churned customers for both females and males, suggesting that the likelihood of churn is not significantly different between genders.



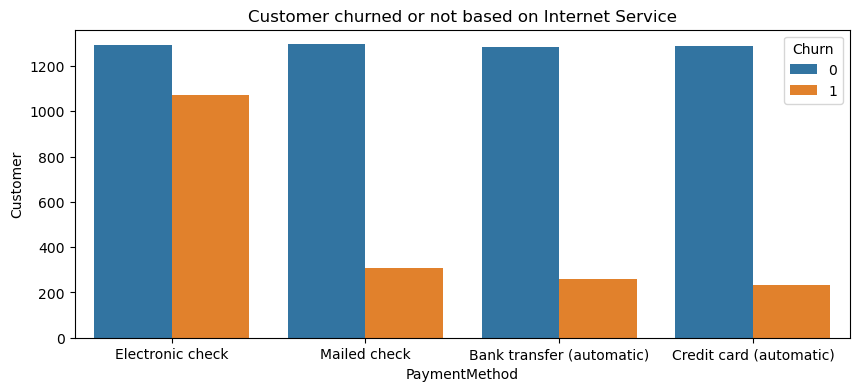
## Churn Analysis Based on Internet Service



The bar chart above illustrates the number of customers who have churned (left the service) versus those who have stayed, segmented by the type of internet service they use. The data is categorized into three types of internet services: DSL, Fiber Optic, and No Internet Service.

* **DSL Internet Service:**
  + **Non-churned Customers:** Approximately 2000
  + **Churned Customers:** Approximately 500
  + **Observation:** The majority of DSL customers remain with the service, with a churn rate of about 20%.
* **Fiber Optic Internet Service:**
  + **Non-churned Customers:** Approximately 1800
  + **Churned Customers:** Approximately 1250
  + **Observation:** A significant number of Fiber Optic customers have churned, indicating a higher churn rate compared to DSL. The churn rate is approximately 40%.
* **No Internet Service:**
  + **Non-churned Customers:** Approximately 1500
  + **Churned Customers:** Approximately 100
  + **Observation:** Customers without internet service show a very low churn rate, likely due to other factors unrelated to internet service.

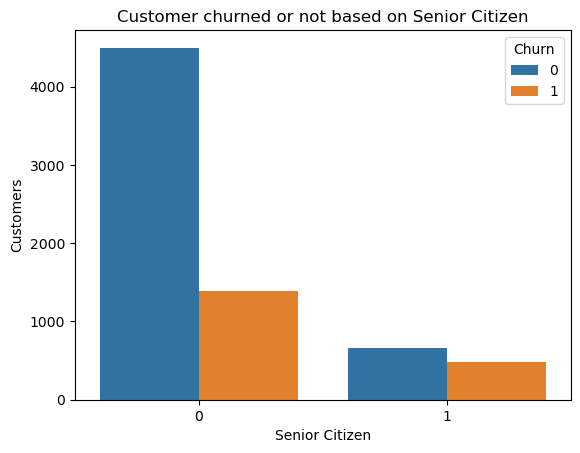
## Churn Analysis Based on Payment Method



The bar chart above shows the number of customers who have churned versus those who have stayed, segmented by their payment method. The payment methods are categorized as Electronic Check, Mailed Check, Bank Transfer (Automatic), and Credit Card (Automatic).

* **Electronic Check:**
  + **Non-churned Customers:** Approximately 1200
  + **Churned Customers:** Approximately 1000
  + **Observation:** There is a significant churn rate among customers using electronic checks, with nearly half of these customers churning.
* **Mailed Check:**
  + **Non-churned Customers:** Approximately 1300
  + **Churned Customers:** Approximately 300
  + **Observation:** Customers who pay by mailed check exhibit a lower churn rate compared to those using electronic checks, with a churn rate of about 20%.
* **Bank Transfer (Automatic):**
  + **Non-churned Customers:** Approximately 1300
  + **Churned Customers:** Approximately 300
  + **Observation:** Similar to mailed checks, bank transfer (automatic) users show a low churn rate, around 20%.
* **Credit Card (Automatic):**
  + **Non-churned Customers:** Approximately 1300
  + **Churned Customers:** Approximately 300
  + **Observation:** Customers using automatic credit card payments also demonstrate a low churn rate, about 20%.

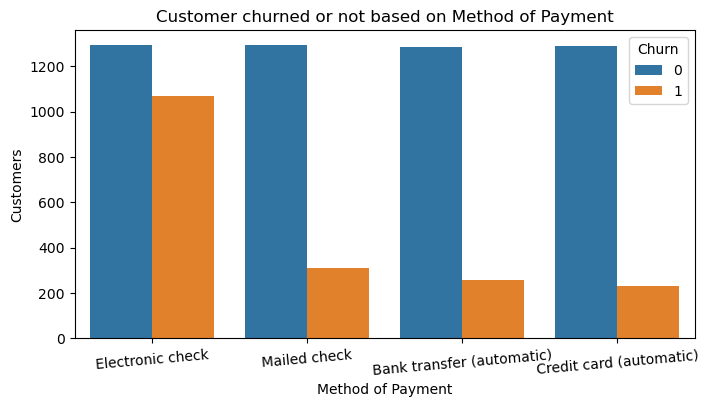
## Churn Analysis based on Senior Citizen



The bar chart above illustrates the number of customers who have churned versus those who have stayed, segmented by whether they are senior citizens or not.

* **Non-Senior Citizens (0):**
  + **Non-churned Customers:** Approximately 4500
  + **Churned Customers:** Approximately 1500
  + **Observation:** A larger portion of non-senior citizens remains with the service, indicating a lower churn rate.
* **Senior Citizens (1):**
  + **Non-churned Customers:** Approximately 500
  + **Churned Customers:** Approximately 250
  + **Observation:** Senior citizens have a higher churn rate compared to non-senior citizens, though the total number of senior citizen customers is significantly lower.

## Churn Analysis based on Method of Payment

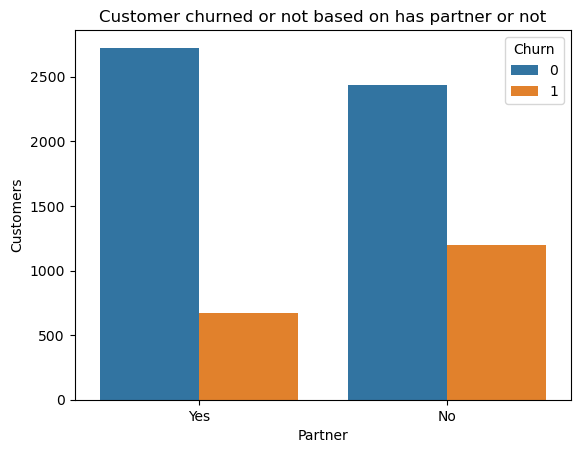


The bar chart provides a visual representation of customer churn in relation to different methods of payment. The y-axis represents the number of customers, while the x-axis indicates the various payment methods. The payment methods included in the chart are Electronic Check, Mailed Check, Bank Transfer (automatic), and Credit Card (automatic). Each method is subdivided into two categories: customers who churned (indicated in orange) and those who did not churn (indicated in blue).

Key Observations

1. **Electronic Check:**
   * **Churned:** Approximately 1,000 customers
   * **Not Churned:** Approximately 1,200 customers
   * **Insight:** A significant portion of customers using electronic checks tend to churn, suggesting potential issues or dissatisfaction with this payment method.
2. **Mailed Check:**
   * **Churned:** Around 300 customers
   * **Not Churned:** Around 1,200 customers
   * **Insight:** Mailed checks show a lower churn rate compared to electronic checks. The majority of customers using this method do not churn, indicating a stable customer base with this payment method.
3. **Bank Transfer (Automatic):**
   * **Churned:** Around 200 customers
   * **Not Churned:** Around 1,200 customers
   * **Insight:** Automatic bank transfers have the lowest churn rate among all payment methods, suggesting that customers who use this method are more likely to remain subscribed.
4. **Credit Card (Automatic):**
   * **Churned:** Around 200 customers
   * **Not Churned:** Around 1,200 customers
   * **Insight:** Similar to automatic bank transfers, automatic credit card payments also exhibit a low churn rate, indicating high customer retention.

## Churn Analysis based on having a Partner

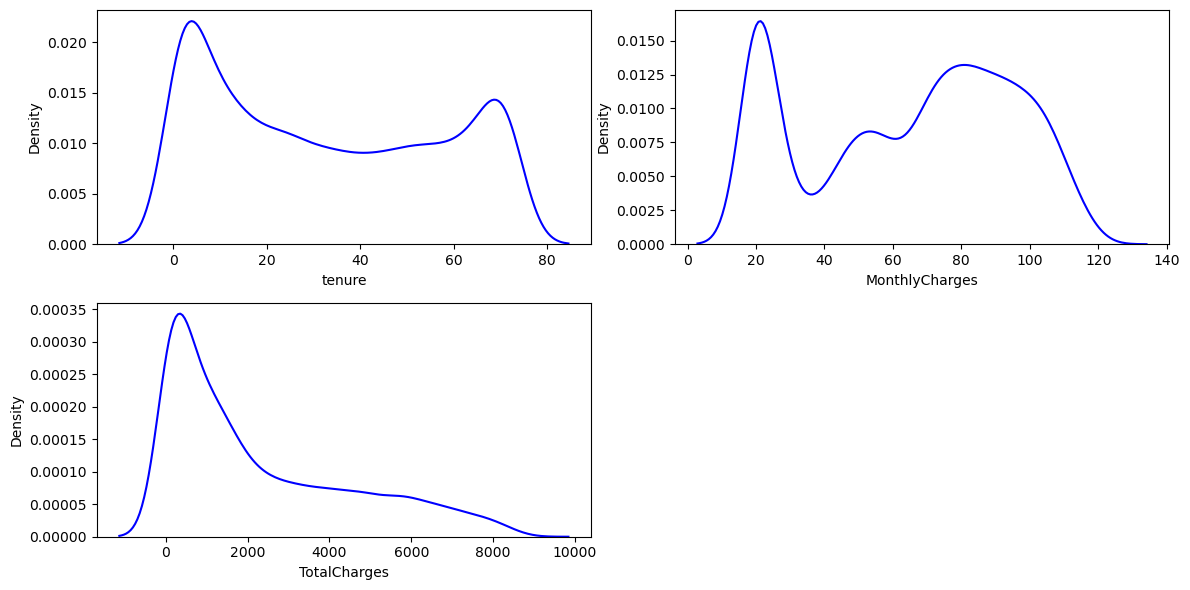


The bar chart illustrates customer churn in relation to whether customers have a partner. The y-axis represents the number of customers, while the x-axis categorizes them based on their partner status (Yes or No). Each category is further divided into two groups: customers who churned (shown in orange) and those who did not churn (shown in blue).

Key Observations

1. **Customers with a Partner:**
   * **Churned:** Approximately 500 customers
   * **Not Churned:** Approximately 2,600 customers
   * **Insight:** The majority of customers who have a partner do not churn. This indicates a high retention rate among partnered customers.
2. **Customers without a Partner:**
   * **Churned:** Around 1,300 customers
   * **Not Churned:** Around 2,200 customers
   * **Insight:** There is a higher churn rate among customers without a partner compared to those with a partner. However, the number of customers who do not churn is still significant, though lower than the partnered group.

### Analysis of Customer Distribution Based on Tenure, Monthly Charges, and Total Charges



The density plots provide insights into the distribution of customers based on three variables: tenure, monthly charges, and total charges. Each plot represents the density of customers for the respective variable.

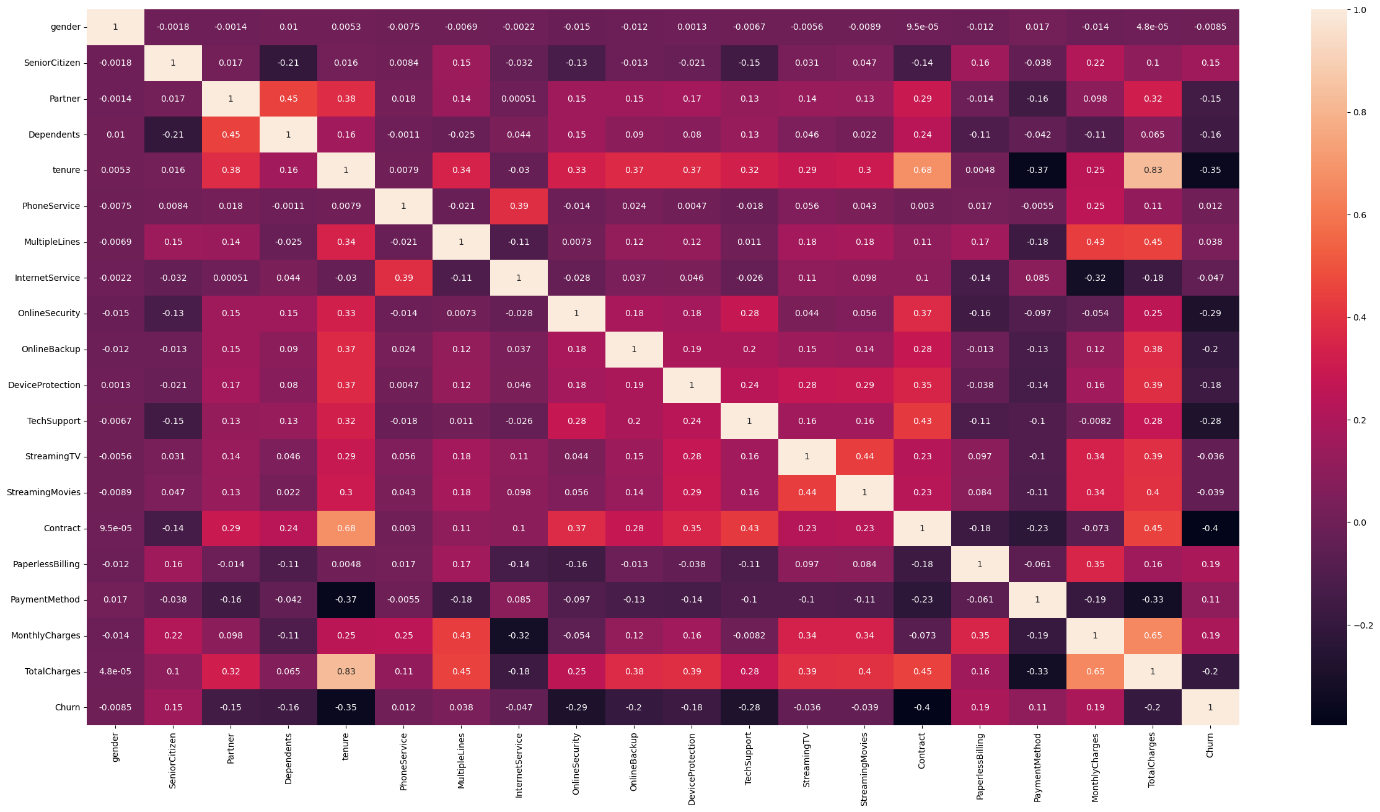
Key Observations

1. **Tenure:**
   * **Peak at Start:** There is a high density of customers with very low tenure, indicating a large number of new customers.
   * **Second Peak:** Another smaller peak is observed at around 70-80 months of tenure.
   * **Insight:** The majority of customers either are new or have been with the company for a long period, with fewer customers in the middle range of tenure.
2. **Monthly Charges:**
   * **Peak at Low Charges:** There is a high density of customers with monthly charges around 20-25 units.
   * **Second Peak:** Another peak is observed at around 70-80 units.
   * **Insight:** Customers tend to cluster around lower and moderate monthly charges, with fewer customers paying high monthly charges.
3. **Total Charges:**
   * **Peak at Low Total Charges:** There is a high density of customers with total charges near the lower end of the scale.
   * **Long Tail:** The distribution has a long tail extending towards higher total charges.
   * **Insight:** Most customers have relatively low total charges, which is consistent with the presence of many new customers (low tenure). A smaller number of customers accumulate higher total charges over time.

# Feature Selection for Churn Prediction

#### Introduction

In the context of predicting customer churn, it's essential to choose the most relevant features to ensure our model performs optimally. Through a comprehensive exploratory data analysis (EDA), we have identified and selected a subset of features that are most indicative of customer churn.



#### Features Dropped

Several features were excluded from the model based on their low correlation with churn and redundancy with other features. These features include:

* **Gender**: Showed no significant correlation with churn.
* **Partner**: Had a low correlation with churn and was not considered a strong predictor.
* **Dependents**: Similar to the partner, dependents had a low impact on churn prediction.
* **PhoneService**: Was found to be redundant due to its high correlation with other features.
* **OnlineSecurity**: Showed multicollinearity with other internet-related services.
* **DeviceProtection**: Also exhibited multicollinearity with similar service features.
* **StreamingTV**: Had high inter-correlation with other streaming services and thus was removed to avoid redundancy.
* **Churn**: This is the target variable and was separated from the feature set.

#### Features Selected

The features selected for building the machine learning model include:

* **SeniorCitizen**: Indicates whether the customer is a senior citizen.
* **tenure**: Represents the number of months the customer has been with the company.
* **MultipleLines**: Indicates if the customer has multiple lines.
* **InternetService**: Type of internet service the customer has (DSL, Fiber optic, No).
* **OnlineBackup**: Whether the customer has online backup service.
* **TechSupport**: Indicates if the customer has tech support service.
* **StreamingMovies**: Whether the customer has streaming movies service.
* **Contract**: The type of contract the customer has (Month-to-month, One year, Two year).
* **PaperlessBilling**: Indicates if the customer is using paperless billing.
* **PaymentMethod**: The payment method used by the customer (Electronic check, Mailed check, Bank transfer, Credit card).
* **MonthlyCharges**: The monthly charges the customer incurs.
* **TotalCharges**: The total charges the customer has incurred.

#### Justification for Selected Features

* **SeniorCitizen**: This demographic information could be a strong indicator of customer behavior and preferences.
* **Tenure**: The length of the customer relationship is often inversely related to churn; longer tenure usually means a lower likelihood of churn.
* **MultipleLines**: Customers with multiple lines might have a higher switching cost, impacting churn.
* **InternetService**: The type of internet service could influence churn due to differences in customer satisfaction and service reliability.
* **OnlineBackup and TechSupport**: These additional services could indicate customer engagement and satisfaction, affecting churn.
* **StreamingMovies**: As a popular service, its presence could impact customer retention.
* **Contract**: The type of contract is directly related to churn; customers on month-to-month plans are more likely to churn than those on longer contracts.
* **PaperlessBilling**: This modern billing method might correlate with customer satisfaction and technological adeptness.
* **PaymentMethod**: Different payment methods could be associated with varying levels of convenience and customer satisfaction.
* **MonthlyCharges and TotalCharges**: Financial aspects are crucial, as higher charges might lead to dissatisfaction and eventual churn.

# Model Performance Report

#### Introduction

After extensive feature engineering and selection, we have trained a machine learning model to predict customer churn. Below is a detailed analysis of the model's performance based on various metrics, including training accuracy, testing accuracy, precision, recall, and F1-score.

#### Model Accuracy

* **Training Accuracy**: 0.804
* **Testing Accuracy**: 0.793

#### Analysis

* **Class 0 (Non-Churn)**:
  + **Precision**: 0.84 - Of the customers predicted not to churn, 84% actually did not churn.
  + **Recall**: 0.88 - Of the customers who did not churn, 88% were correctly identified by the model.
  + **F1-Score**: 0.86 - This balance between precision and recall shows that the model performs well in predicting customers who will not churn.
* **Class 1 (Churn)**:
  + **Precision**: 0.65 - Of the customers predicted to churn, 65% actually did churn.
  + **Recall**: 0.56 - Of the customers who did churn, 56% were correctly identified by the model.
  + **F1-Score**: 0.60 - This lower score indicates that the model is less effective at predicting churn compared to non-churn.