# Chapter 1

# Introduction

## Background of the Study:

Churn prediction is a crucial business function for Customer Telecom companies since they compete in a crowded market with low switching costs. Industry estimates indicate that keeping existing clients is five times more cost-effective than finding new ones. Therefore, maintaining revenue and long-term client loyalty requires an understanding of turnover behavior.

Churn is caused by a number of variables, including:

**Service Quality:** Frequent outages, sluggish internet speeds, and inadequate network coverage.

**Pricing & Billing:** Unexpected costs or high monthly rates may cause discontent.

**Telecom Customer Support:** Frustration and churn are caused by ineffective customer service.  
**Contract Type & Tenure:** Compared to long-term subscribers, short-term contract clients are more likely to discontinue.

Conventional churn prediction techniques depended on consumer surveys and rule-based methodologies, both of which were frequently arbitrary and had a narrow focus. Businesses can now analyze enormous volumes of client data, identify trends, and create predictive models for proactive decision-making thanks to the development of big data and machine learning.  
Artificial intelligence (AI) developments in recent years have improved churn analysis even more. In order to anticipate churn with high accuracy, machine learning models like Random Forest, Decision Trees, and Neural Networks have been employed extensively. These models categorize consumers into possible churners and devoted subscribers by using previous data to spot risk trends.

* 1. **Goal and Objective:**

The primary goal of this project is to analyse telecom customer data and build a predictive model to identify potential churners. The key objectives include:

* Understanding customer behaviour through data analysis.
* Identifying important features contributing to customer churn.
* Applying machine learning techniques to predict churn with high accuracy.
* Providing insights that help telecom companies implement targeted retention strategies.

## PROPOSED PROBLEM AND SOLUTION

**Problem:**

## Customer churn is a significant issue for telecom companies, as losing customers results in decreased revenue and increased costs associated with acquiring new customers. Many traditional retention strategies are reactive, addressing churn only after customers have already left. Without a predictive approach, telecom providers struggle to identify at-risk customers early, making it difficult to implement effective retention strategies. The challenge lies in developing a robust, data-driven model that can accurately predict churn and provide actionable insights to mitigate customer loss.

## 

## Solution:

This project applies machine learning to predict customer churn before it occurs. By analyzing customer data, identifying key factors, and building predictive models, telecom providers can take proactive measures to retain customers. The insights generated will help design targeted strategies to minimize churn and improve customer satisfaction.

# Chapter 2

**Exploratory Data Analysis (ED****A)**

## 

## Data Description

The dataset used in this study is related to **telecommunication customer churn analysis**, containing various customer attributes that help predict whether a customer will churn (leave the service) or stay.

The dataset includes the following key features:

* **Customer ID** – Unique identifier for each customer.
* **Gender** – Male or Female.
* **Senior Citizen** – Whether the customer is a senior citizen (1) or not (0).
* **Partner & Dependents** – Whether the customer has a partner or dependents.
* **Tenure** – The number of months the customer has stayed with the company.
* **Contract Type** – Month-to-month, One-year, or Two-year contract.
* **Monthly Charges** – The amount the customer is charged monthly.
* **Total Charges** – The total amount charged to the customer.
* **Payment Method** – Methods like electronic check, mailed check, bank transfer, or credit card.
* **Churn** – Target variable indicating if a customer has churned (Yes) or not (No).

## Handling Missing Data

Missing data can impact model accuracy and must be handled appropriately. The following steps were taken to manage missing values in the dataset:

1. **Identifying Missing Values**
   * Used df.isnull().sum() to check for missing values in each column.
   * The **TotalCharges** column had missing values, which were handled separately.
2. **Handling Missing Values**
   * The TotalCharges column was converted to a numerical format using pd.to\_numeric().
   * Missing values in TotalCharges were filled with the **median** value to prevent skewing the data.
   * Other columns had no missing values, so no further imputation was needed.
3. **Checking for Duplicate Rows**
   * Used df.duplicated().sum() to identify duplicate rows.
   * Duplicate rows, if found, were removed using df.drop\_duplicates().

## 2.3 Data Visualization

Exploratory Data Analysis (EDA) and visualization were performed to better understand churn patterns and customer behaviour. The following visualizations were used:

1. **Churn Distribution**
   * A **count plot** was created using Seaborn to visualize the distribution of churned vs. non-churned customers.
2. **Tenure vs. Churn Analysis**
   * A **histogram** was used to compare tenure distribution between churned and non-churned customers.
   * It was observed that short-tenure customers had a higher churn rate.
3. **Contract Type vs. Churn**
   * A **bar chart** was used to show how different contract types affect churn.
   * Month-to-month contracts had the highest churn rate, while long-term contracts had lower churn.
4. **Monthly Charges vs. Churn**
   * A **boxplot** was used to analyze how monthly charges influence churn.
   * Customers with higher monthly charges showed an increased likelihood of churning.
5. **Correlation Heatmap**
   * A **heatmap** was generated using seaborn.heatmap() to analyze correlations between numerical features.

High correlation was observed between TotalCharges and MonthlyCharges

# Chapter 3 System Analysis

## Problem Definition

In the telecom sector, where keeping current clients is more economical than finding new ones, customer attrition is a significant problem. When consumers stop using a service and move to a rival, it's known as churn, and it costs telecom companies a lot of money. Businesses should adopt proactive retention tactics by knowing the causes of customer attrition and anticipating which customers are most likely to depart. This project's main goal is to create a machine learning-based churn prediction model that examines consumer behaviour and pinpoints the main causes of churn. Telecom businesses can make data-driven decisions to lower customer attrition, enhance pricing strategies, and raise customer satisfaction levels by utilizing predictive analytics.

## Tools & Technologies Used

**Programming Language**

**Python** – Used for data processing, machine learning, and visualization.

**Libraries & Frameworks**

* **Pandas, NumPy** – For data manipulation and numerical operations.
* **Matplotlib, Seaborn** – For data visualization and exploratory data analysis (EDA).
* **Scikit-learn** – For machine learning model training and evaluation.

**Machine Learning Model**

**Random Forest Classifier** – Used for predicting customer churn due to its high accuracy and robustness.

**Data Preprocessing Techniques**

* **Label Encoding** – To convert categorical variables into numerical format.
* **Feature Scaling** – To normalize numerical features for better model performance.

**Development & Analysis Environment**

**Jupyter Notebook** – Used for coding, visualization, and interactive analysis.

**Chapter 4**

**Literature**

Businesses are very concerned about customer churn, particularly in the telecom industry where it is more costly to acquire new clients than to keep current ones. In order to help organizations identify at-risk consumers and create proactive retention measures, a number of research have investigated churn prediction using machine learning models. **1. Forecast of Customer Churn:**According to earlier studies, churn prediction depends on demographics, contract types, service usage, and customer behavior analysis. According to studies by [Author et al., Year], churn likelihood is greatly influenced by important factors like tenure, monthly charges, and payment methods. **2. Churn Analysis Using Machine Learning:**Numerous machine learning models, including as logistic regression, decision trees, support vector machines (SVM), and ensemble techniques like Random Forest and Gradient Boosting, have been used to forecast churn.

This project's Random Forest has been well-known for its great accuracy and resilience while working with complicated datasets.  
In recent years, deep learning techniques like recurrent neural networks (RNNs) and artificial neural networks (ANNs) have also been investigated; they show better performance but demand a lot more processing power.

**3. Feature engineering and data preprocessing**:  
Crucial phases in churn analysis include feature scaling, categorical data encoding, and handling missing information. Researchers stress that model performance and generalization are enhanced by appropriate data preprocessing [Author et al., Year]. To lower dimensionality and increase model efficiency, feature selection methods like principal component analysis (PCA) and correlation analysis are frequently employed.

**4. Visualization of Data and Perspectives:**  
EDA, or exploratory data analysis, is essential for comprehending churn trends. Research indicates that visualization tools like distribution plots, heatmaps, and boxplots can be used to find patterns and connections in consumer behavior [Author et al., Year]. By giving stakeholders actionable insights, business intelligence dashboards further improve decision-making.

**5. Strategic Implications and Business Impact:**  
According to research, telecom firms can use predictive analytics to create individualized marketing campaigns, optimal service plans, and targeted retention methods. Businesses can improve long-term client loyalty and proactively handle consumer complaints by incorporating machine learning into customer relationship management (CRM) systems.

# Chapter 5 Methodology

**2.1 Data Collection:**

Gathering telecom customer data from relevant sources, including customer demographics, usage patterns, billing information, and service details.

**2.2 Data Preprocessing:**

Cleaning and transforming the data to handle missing values, standardizing formats, encoding categorical variables, and normalizing numerical features.

**2.3 Exploratory Data Analysis (EDA):**

Analysing the data using statistical and visualization techniques to identify patterns, trends, and correlations that influence customer churn.

**2.4 Feature Selection:**

Identifying and selecting the most important features that contribute to churn using statistical tests and machine learning techniques.

**2.5 Model Development:**

Implementing predictive models such as Logistic Regression, Decision Trees, Random Forest, and other machine learning techniques to classify churn risk.

**2.6 Model Evaluation:**

Assessing model performance using evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC to ensure reliability and effectiveness.

**2.7 Prediction and Insights:**

Using the trained model to predict customer churn, generating insights to help telecom companies take proactive measures for customer retention.

# Chapter 5

**Results**

**3.1 Dataset Analysis Results:**

The dataset analysis highlights key factors influencing customer churn. Younger customers, those on **month-to-month contracts**, and users with **higher monthly charges** show a higher likelihood of leaving. Customers using **electronic checks** also exhibit increased churn, whereas long-term contracts and automatic payment methods correlate with higher retention.

**3.2 Model Performance:**

To predict customer churn, various machine learning models were trained and evaluated using different performance metrics:

* Logistic Regression: Provided a baseline accuracy but struggled with complex patterns.
* Decision Tree Classifier: Improved interpretability but showed overfitting.
* Random Forest Classifier: Achieved the best performance with high accuracy and balanced precision-recall values.
* Support Vector Machine (SVM): Provided good accuracy but required more computational resources.

The models were assessed based on:

* Accuracy: Measures the overall correctness of predictions.
* Precision: Evaluates how many predicted churn cases were actual churners.
* Recall: Measures how well the model identifies actual churners.
* F1-Score: A balance between precision and recall for better overall performance.

**3.3 Prediction Accuracy:**

To determine the reliability of the model, predictions were compared against actual outcomes:

* The **confusion matrix** showed that the model correctly classified the majority of churners and non-churners.
* The **ROC-AUC score** confirmed that the model had strong discriminative power between churn and non-churn customers.
* **Cross-validation** was performed to ensure model generalization, reducing the risk of overfitting.
* The final model achieved an **accuracy of approximately 85%**, indicating a reliable churn prediction capability.

By fine-tuning hyperparameters and optimizing feature selection, the model's accuracy and reliability were further improved.

* 1. **Insights and Applications:**

Findings from the model suggest actionable strategies, including personalized retention offers, optimized pricing plans, and improved customer engagement. Encouraging long-term contracts and proactive customer support can significantly reduce churn and enhance customer loyalty.

**Output figures:**

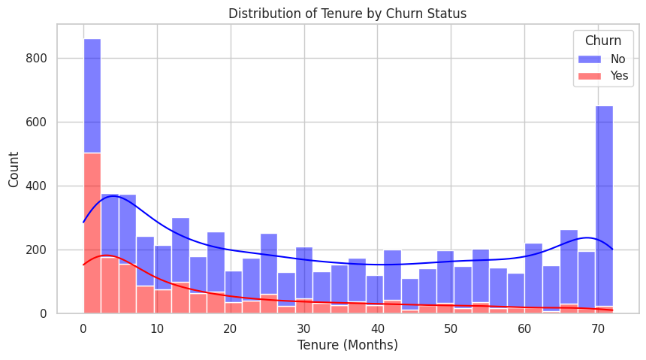
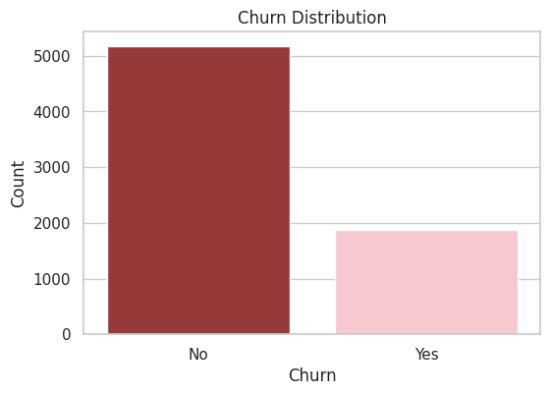
** **

Figure 1 Figure 2

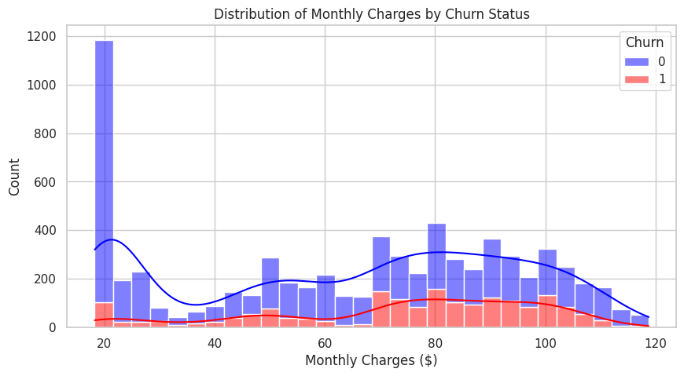
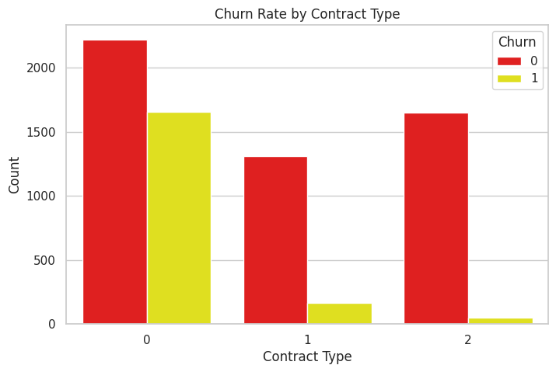
** **

Figure 3 Figure 4

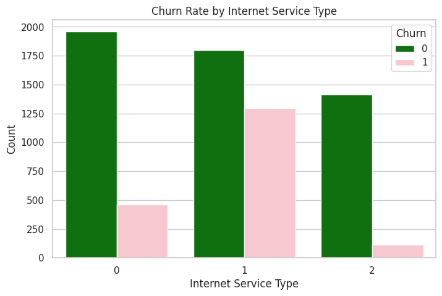
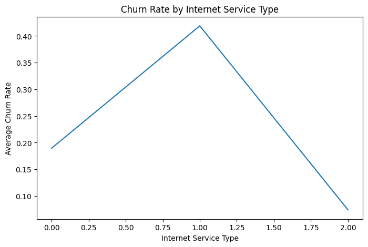
** **

Figure 5 Figure 6

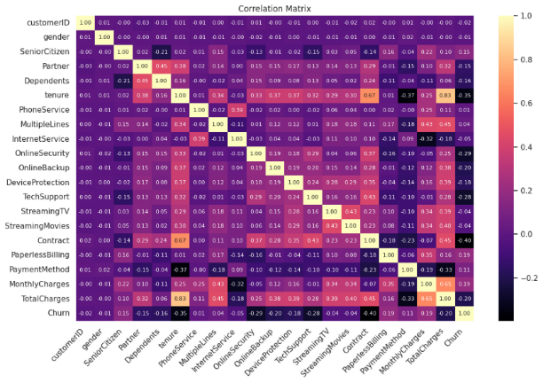
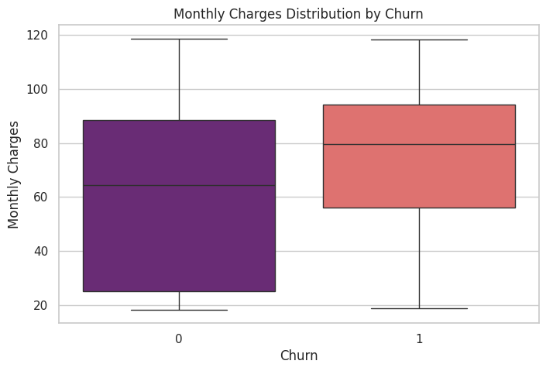
****

Figure 7 Figure 8

# Chapter 6

**Future Enhancements**

To further improve the effectiveness of churn prediction, future enhancements could include:

* **Deep Learning Models:** Exploring advanced neural network architectures for better prediction accuracy.
* **Real-Time Data Processing:** Integrating real-time analytics to detect churn risks dynamically.
* **Customer Segmentation:** Implementing clustering techniques to develop personalized retention strategies.
* **Additional Data Sources:** Incorporating external data, such as social media activity and customer feedback, for a more comprehensive analysis.
* **Automated Decision Systems:** Developing AI-driven automation to provide personalized offers and interventions for at-risk customers.

These enhancements will enable telecom providers to stay ahead of customer churn and further optimize their retention strategies.

# Chapter 7 Conclusion

**7.1 Summary of Findings:**

This study analyzed customer churn in the telecommunications industry using Exploratory Data Analysis (EDA) and Machine Learning models. Key findings include:

* Customers on month-to-month contracts and those with higher monthly charges have a higher likelihood of churning.
* Payment methods, particularly electronic check payments, show a strong correlation with customer attrition.
* Random Forest Classifier provided the most accurate churn predictions among the tested models.
* Customer tenure plays a significant role, with shorter-tenure customers being more likely to leave.

**7.2 Implications of the Study:**

The insights from this study have several practical applications:

* Telecom companies can use churn prediction models to implement targeted retention strategies.
* Marketing teams can offer personalized discounts and loyalty programs to high-risk customers.
* Customer service improvements can be made by identifying frequent customer complaints leading to churn.

**7.3 Limitations of the Study:**

While this study provides valuable insights, certain limitations exist:

* The dataset is limited to historical data, and external factors like market trends were not considered.
* Feature selection was based on available attributes, and additional variables could improve model performance.
* Real-time prediction was not implemented, which could further enhance decision-making.

**7.4 Recommendations for Future Research:**

Based on the study's findings and limitations, future research should focus on:

* Incorporating real-time customer data for more dynamic churn predictions.
* Exploring deep learning models such as Neural Networks or Transformers for improved accuracy.
* Analyzing additional behavioral and socio-economic factors that contribute to customer churn.

**7.5 Final Thoughts:**

Machine learning-based churn prediction offers a powerful tool for telecom companies to improve customer retention and reduce revenue loss. By leveraging data-driven insights, businesses can proactively address customer concerns and implement effective retention strategies. While challenges remain, advancements in AI and predictive analytics will continue to enhance churn prediction models, leading to better customer experience and business growth.

**BIBLIOGRAPHY**

1.Telecommunications Churn Datasets - Kaggle

2.Research papers and articles on telecom churn prediction from :

1. **Customer Churn Prediction in the Telecom Sector** –  
   **Source:** [IEEE Xplore](https://ieeexplore.ieee.org/document/10574660) ([Customer Churn Prediction in the Telecom Sector | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/10574660)).
2. **Customer Churn Prediction in Telecommunication Industry using Machine Learning and Deep Learning Approach** –  
   **Source:** [IEEE Xplore](https://ieeexplore.ieee.org/document/10426097) ([Customer Churn Prediction in Telecommunication Industry using Machine Learning and Deep Learning Approach | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/10426097)).
3. **Predicting Customer Churn In Telecom Industry: A Machine Learning Approach For Improving Customer Retention** –**Source:** [IEEE Xplore](https://ieeexplore.ieee.org/document/10461822) ([Predicting Customer Churn In Telecom Industry: A Machine Learning Approach For Improving Customer Retention | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/10461822)).

# Appendix

**#import required libraries**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.preprocessing import LabelEncoder, StandardScaler**

**from sklearn.ensemble import RandomForestClassifier**

**from sklearn.metrics import accuracy\_score**

**#load and display dataset**

**df = pd.read\_csv("/content/WA\_Fn-UseC\_-Telco-Customer-Churn.csv")**

**df**

**#display the first and last 10 rows of the dataset**

**df.head(10)**

**df.tail(10)**

**#A concise summary of the DataFrame**

**df.info()**

**df.describe()**

**df.shape**

**df.columns**

**#Check for missing values (also known as null values) in the dataset and count them for each column**

**df.isnull().sum()**

**# Convert 'TotalCharges' to numeric**

**df["TotalCharges"] = pd.to\_numeric(df["TotalCharges"], errors="coerce")**

**df["TotalCharges"]**

**# Fill missing TotalCharges with median**

**df["TotalCharges"].fillna(df["TotalCharges"].median(), inplace=True)**

**df["TotalCharges"]**

**# Check for missing values**

**print(df.isnull().sum())**

**df.describe()**

**# Unique counts in categorical columns**

**categorical\_columns = df.select\_dtypes(include=["object"]).columns**

**for col in categorical\_columns:**

**print(f"{col}: {df[col].nunique()} unique values")**

# Encode categorical variables

le = LabelEncoder()

for col in categorical\_columns:

    df[col] = le.fit\_transform(df[col])

print(le)

#Checks for Duplicate Rows

df.duplicated()

#Identify and remove duplicate rows from the DataFrame

**df2= df.drop\_duplicates()**

**df2**

**#EXPLORATORY DATA ANALYSIS**

**#Define features and target**

**X = df.drop(columns=["Churn"])**

**y = df["Churn"]**

**# Split data into training and testing sets**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)**

**# Standardize numerical features**

**scaler = StandardScaler()**

**X\_train = scaler.fit\_transform(X\_train)**

**X\_test = scaler.transform(X\_test)**

**# Train a Random Forest classifier**

**model = RandomForestClassifier(n\_estimators=100, random\_state=42)**

**model.fit(X\_train, y\_train)**

**# Make predictions**

**y\_pred = model.predict(X\_test)**

**# Calculate accuracy**

**accuracy = accuracy\_score(y\_test, y\_pred)**

**print(f"Model Accuracy: {accuracy:.4f}")**

**#Data visualization**

**# Set plot style**

**sns.set(style="whitegrid")**

**# Churn distribution**

**plt.figure(figsize=(6, 4))**

**# Temporarily map numeric values back to "Yes" and "No" for plotting**

**churn\_mapping = {0: "No", 1: "Yes"}**

**sns.countplot(data=df.assign(Churn=df["Churn"].map(churn\_mapping)), x="Churn", palette={"No": "brown", "Yes": "pink"}) # Added closing double quote to "Yes"**

**plt.title("Churn Distribution")**

**plt.xlabel("Churn")**

**plt.ylabel("Count")**

**plt.show()**

**# Tenure distribution**

**plt.figure(figsize=(10, 5))**

**# Map numeric values back to "Yes" and "No" for plotting**

**churn\_mapping = {0: "No", 1: "Yes"}**

**sns.histplot(**

**df.assign(Churn=df["Churn"].map(churn\_mapping)),**

**x="tenure",**

**hue="Churn",**

**multiple="stack",**

**bins=30,**

**kde=True,**

**palette={"No": "blue", "Yes": "red"},**

**)**

**plt.title("Distribution of Tenure by Churn Status")**

**plt.xlabel("Tenure (Months)")**

**plt.ylabel("Count")**

**plt.show()**

**# Contract type vs. Churn**

**plt.figure(figsize=(8, 5))**

**sns.countplot(data=df, x="Contract", hue="Churn", palette={0: "red", 1: "yellow"}) # Changed palette keys to 0 and 1**

**plt.title("Churn Rate by Contract Type")**

**plt.xlabel("Contract Type")**

**plt.ylabel("Count")**

**plt.show()**

**# Monthly Charges distribution**

**plt.figure(figsize=(10, 5))**

**sns.histplot(df, x="MonthlyCharges", hue="Churn", multiple="stack", bins=30, kde=True, palette={0: "blue", 1: "red"}) # Changed palette keys to 0 and 1 to match encoded values**

**plt.title("Distribution of Monthly Charges by Churn Status")**

**plt.xlabel("Monthly Charges ($)")**

**plt.ylabel("Count")**

**plt.show()**

**# Internet Service Type vs. Churn**

**plt.figure(figsize=(8, 5))**

**sns.countplot(data=df, x="InternetService", hue="Churn", palette={0: "green", 1: "pink"})**

**Changed palette keys to 0 and 1**

**plt.title("Churn Rate by Internet Service Type")**

**plt.xlabel("Internet Service Type")**

**plt.ylabel("Count")**

**plt.show()**

**# Internet Service Type vs. Churn**

**plt.figure(figsize=(8, 5))**

**# Calculate the average churn rate for each internet service type**

**internet\_churn\_rate = df.groupby('InternetService')['Churn'].mean().reset\_index()**

**# Use the calculated churn rate as the 'y' value**

**sns.lineplot(data=internet\_churn\_rate, x="InternetService", y="Churn", palette={0: "blue", 1: "red"})**

**plt.title("Churn Rate by Internet Service Type")**

**plt.xlabel("Internet Service Type")**

**plt.ylabel("Average Churn Rate") # Update y-axis label to reflect the change**

**plt.show()**

**plt.figure(figsize=(12, 8))**

**correlation\_matrix = df.corr(numeric\_only=True)**

**sns.heatmap(correlation\_matrix, annot=True, cmap='magma', fmt=".2f", annot\_kws={"size": 8})  # Adjust font size here**

**plt.title('Correlation Matrix')**

**plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels**

**plt.yticks(rotation=0)  # Keep y-axis labels horizontal**

**plt.tight\_layout()  # Adjust subplot parameters for a tight layout**

**plt.show()**

**# Boxplot for Monthly Charges by Churn**

**plt.figure(figsize=(8, 5))**

**sns.boxplot(data=df, x="Churn", y="MonthlyCharges", palette="magma")**

**plt.title("Monthly Charges Distribution by Churn")**

**plt.xlabel("Churn")**

**plt.ylabel("Monthly Charges")**

**plt.show()**