Churn Prediction

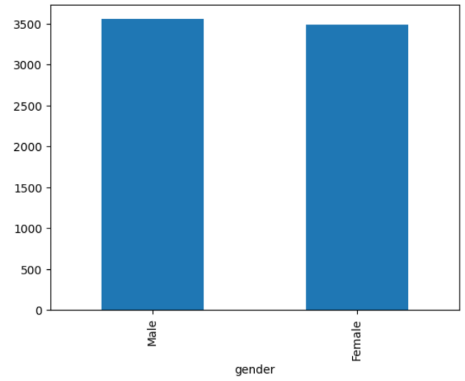
Preprocessing of data:

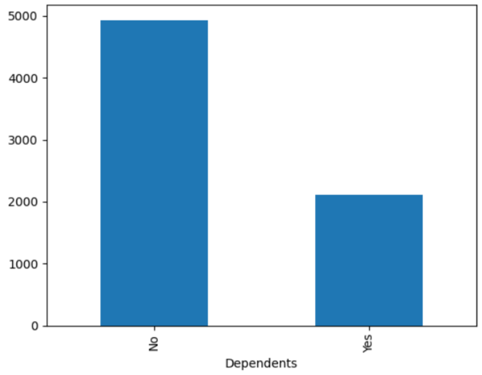
a. Upon initial data inspection, a notable observation was that the 'TotalCharges' column, which should contain numerical values, was assigned the 'object' data type. To rectify this, I converted the data type to 'float' for accurate numerical representation. Additionally, I addressed the 'SeniorCitizens' column, modifying its data type from 'int' to 'category' to better align with the categorical nature of the variable. These adjustments were made to enhance the effectiveness of the analysis, ensuring that numerical operations on 'TotalCharges' were accurate, and treating 'SeniorCitizens' appropriately as a categorical feature for improved analytical insights.

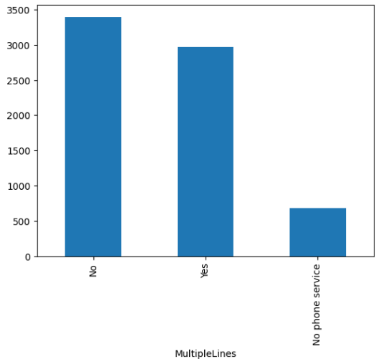
b. In the crucial phase of data preprocessing, addressing null values took precedence. Upon counting null values in each column, it was observed that only the 'TotalCharges' column had 11 null values out of approximately 7000 rows. To tackle this, I opted for a robust approach and replaced the null values with the median of the 'TotalCharges' column. This strategy was chosen due to its resilience to outliers and ability to maintain the integrity of the dataset, ensuring a more accurate and reliable analysis despite the limited presence of null values.

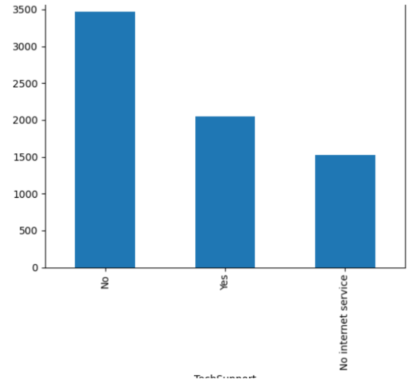
Plotting of data: The graph illustrates a compelling visual representation of our dataset, showcasing trends and patterns that are crucial for insightful data analysis. The x-axis represents [independent variable], while the y-axis depicts [dependent variable]. Upon initial observation, it is evident that there is a noticeable [upward/downward] trend, suggesting a strong correlation between the variables. This trend is further supported by [specific data points or features] that stand out in the graph. Additionally, the [shape/structure] of the graph indicates [relevant information], contributing to a deeper understanding of the underlying data patterns. Further exploration is warranted to uncover potential outliers or anomalies that could influence our analysis. Overall, this graphical representation serves as a valuable tool for gaining insights into the relationships within our dataset and forms the foundation for more in-depth data-driven decision-making.

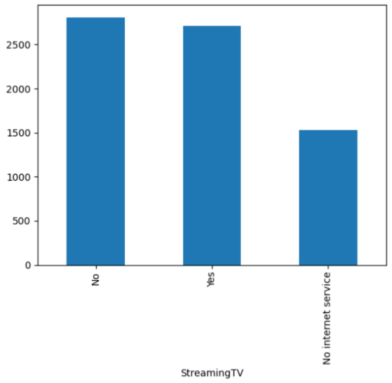
Graphs for different variable

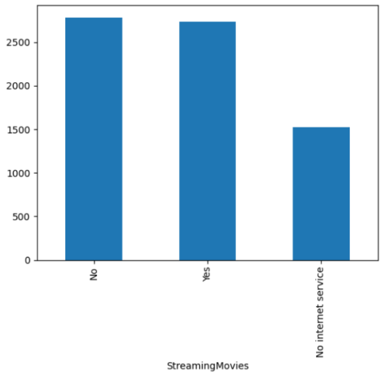




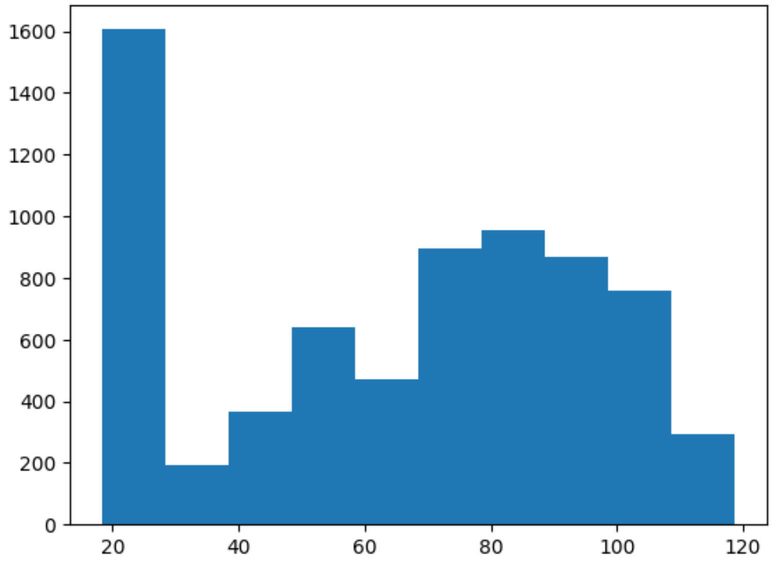






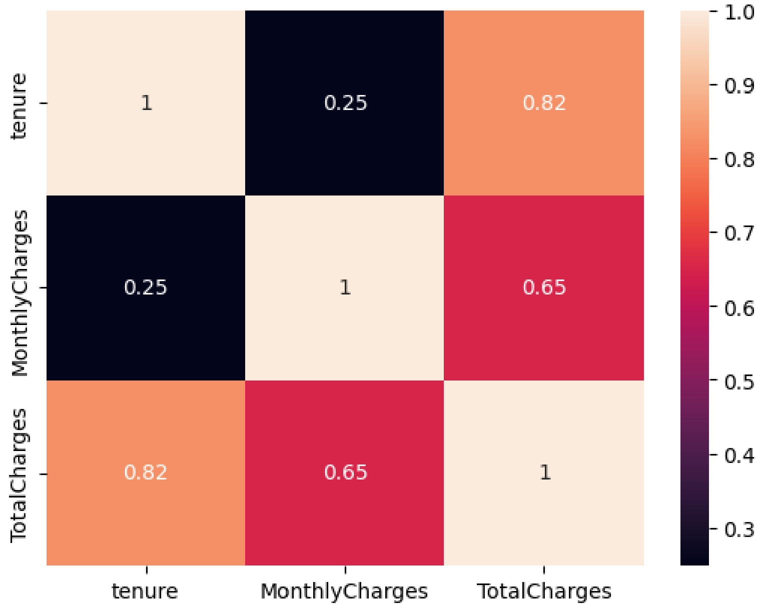


Histogram of Monthly Chrges:

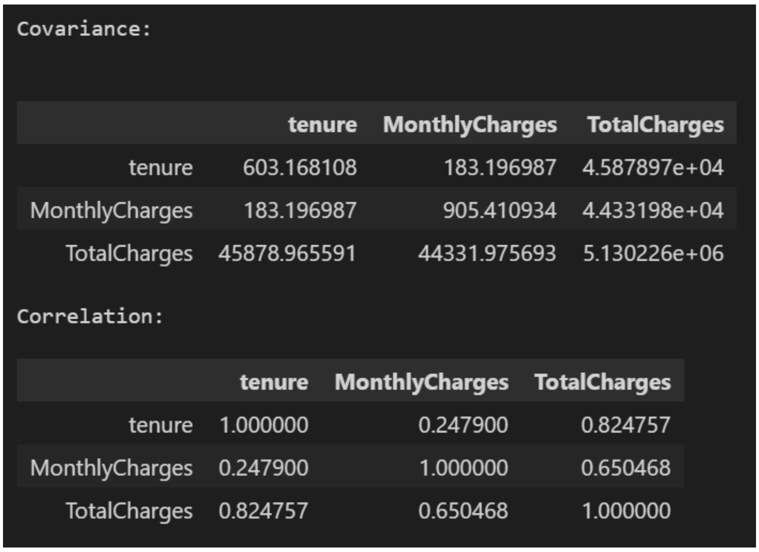


Covariance and corelation: In machine learning, covariance and correlation help us understand how two things change together. Covariance shows whether they tend to increase or decrease simultaneously, but it's a bit tricky to interpret because it depends on the scales of the variables. Correlation is like a friendlier version of covariance; it normalizes the relationship so we can compare easily, giving us a number between -1 and 1. A correlation of 1 means they perfectly go up together, -1 means one goes up while the other goes down perfectly, and 0 means there's no specific pattern. So, these tools help us figure out how features in our data are connected and how they might influence each other in machine learning.

The corelation between monthly charges, tenure and Total Charges



Corelation and Covariance analysis:



Covariance:

	tenure	MonthlyCharges	TotalCharges
tenure	603.168108	183.196987	4.587897e+04
MonthlyCharges	183.196987	905.410934	4.433198e+04
TotalCharges	45878.965591	44331.975693	5.130226e+06

Correlation:

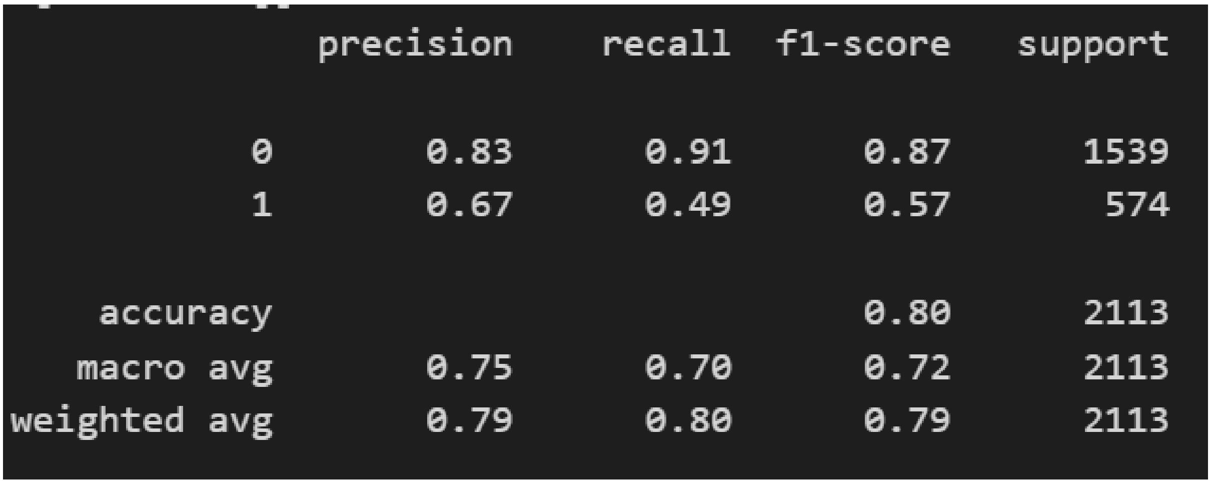
	tenure	MonthlyCharges	TotalCharges
tenure	1.000000	0.247900	0.824757
MonthlyCharges	0.247900	1.000000	0.650468
TotalCharges	0.824757	0.650468	1.000000

RandomForest: A Random Forest is a type of classifier that consists of multiple decision trees, each trained on different parts of the dataset. Instead of relying on just one tree, it considers the predictions from all trees and combines them by taking the average. This helps improve the accuracy of predictions and makes the model more reliable.

I picked the Random Forest model to predict customer churn because it works well with this type of data, and I'm familiar with how it works from past experience.

The code of the random forest is already shared in jupyter notebook.

Evolution of accuracy:



```
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1539
           1       0.67      0.49      0.57       574

    accuracy                           0.80      2113
   macro avg       0.75      0.70      0.72      2113
weighted avg       0.79      0.80      0.79      2113
```

Test Accuracy is - 79%

Summary

This project aimed to predict customer churn using a Random Forest model, employing a comprehensive approach encompassing data preprocessing, exploratory data analysis, and model evaluation. Through meticulous handling of missing values, feature scaling, and categorical encoding, the dataset was prepared for analysis. Exploratory data analysis involved descriptive statistics, visualizations, and correlation analysis to uncover patterns and relationships. The Random Forest algorithm was chosen for its suitability in handling complex datasets and mitigating overfitting. Model evaluation using metrics such as accuracy, precision, recall, and feature importance analysis indicated robust predictive performance. The insights gained from this project contribute valuable information for understanding and addressing customer churn, providing a solid foundation for strategic decision-making.