**Customer Churn Prediction Project**

**Introduction**

In this project, we aim to predict customer churn for a company using machine learning techniques. Customer churn, also known as customer attrition, is the loss of customers or clients. It is an important metric for businesses to understand, as it directly impacts revenue and growth. By predicting customer churn, businesses can implement targeted strategies to retain customers and ultimately improve customer satisfaction.

**Analysis Objectives**

The primary objectives of this analysis are as follows:

1. **Predict Customer Churn:** Develop a predictive model to forecast customer churn based on historical data.
2. **Identify Key Factors:** Determine the significant features that contribute to customer churn, providing insights for potential business improvements.

**Data Collection**

To initiate the analysis, we collected customer data from the provided source. The dataset includes various features related to customer behavior, interactions, and other relevant information. This data will serve as the foundation for building the predictive model and analyzing customer churn.

**DATA SET:** [**https://www.kaggle.com/datasets/blastchar/telco-customer-churn**](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

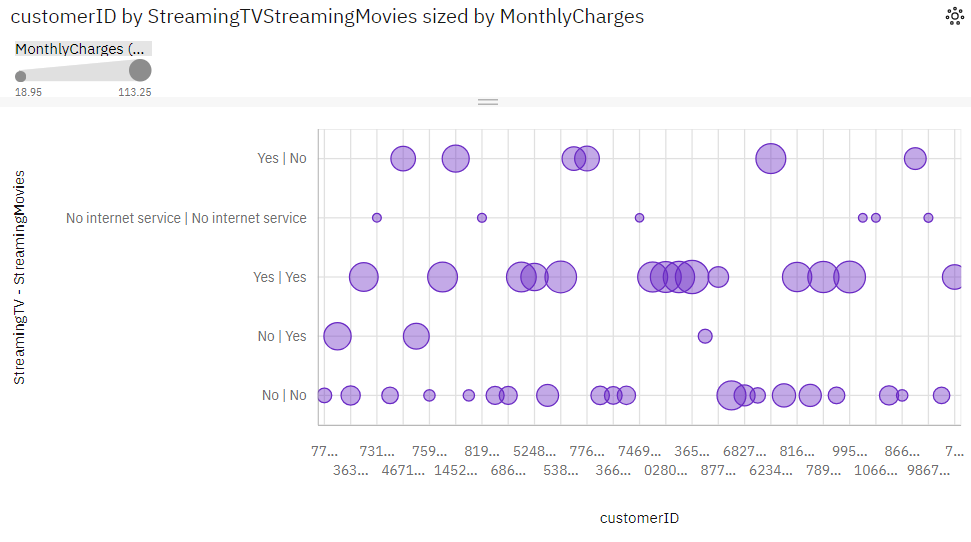
**Data Preprocessing and Cleaning**

The collected data was subjected to a rigorous preprocessing and cleaning procedure to ensure its quality and accuracy. This involved the following steps:

1. **Handling Missing Values:** Missing values in the dataset were either imputed using appropriate methods or removed based on the nature and amount of missing data.
2. **Removing Duplicates:** Duplicate entries, if any, were identified and removed to maintain data integrity.
3. **Data Transformation:** Certain features were transformed or encoded into numerical values to prepare the data for machine learning algorithms.
4. **Outlier Detection and Removal:** Outliers that could potentially distort the predictive model were detected and eliminated.
5. **Data Normalization/Scaling:** To ensure consistency in the features, numerical values were normalized or scaled appropriately.

**Visualization:**

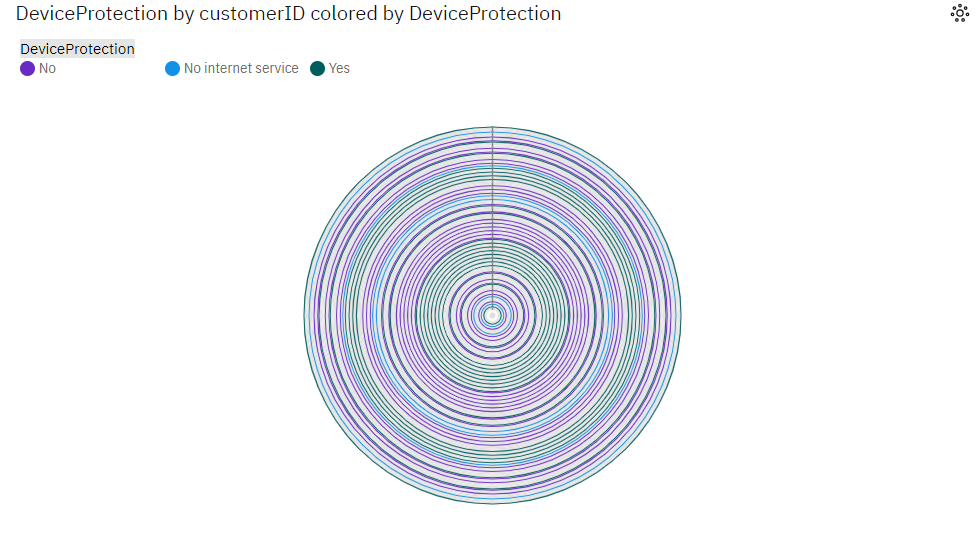
Monthly charges for streaming platform:

****

* **StreamingMovies** **No** has the highest total **MonthlyCharges** due to **MultipleLines** **No**
* **StreamingMovies** **No** has the highest total **MonthlyCharges** due to **StreamingTV** **No**
* **StreamingTV** **Yes** has the highest total **MonthlyCharges** due to **MultipleLines** **Yes**.
* **Yes** has a **MonthlyCharges** of **over a thousand** for **MultipleLines** **Yes**.

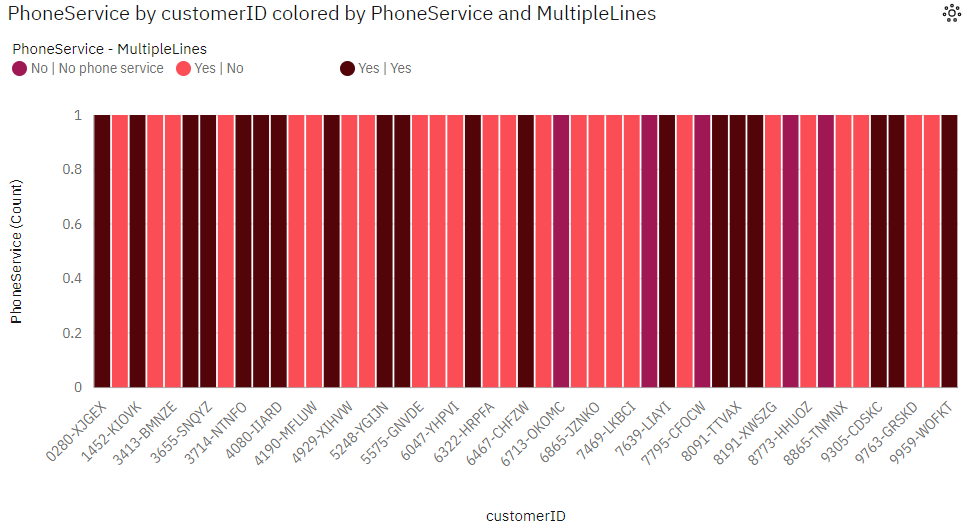
**Device Protection:**

* **No** (**46.9** %) and **Yes** (**40.8** %) are the most frequently occurring categories of **DeviceProtection** with a combined count of **43** items with **DeviceProtection** values (**87.8** % of the total).
* The total number of results for **DeviceProtection**, across all **customerID**, is **49**.

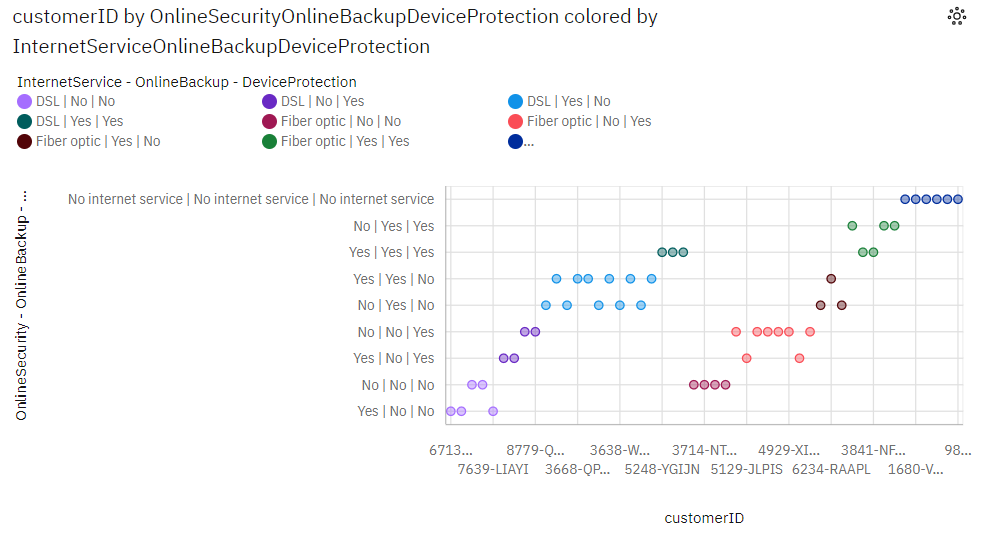


**Phone service and Multiple lines**

* **MultipleLines** **No** has the highest **PhoneService** due to **InternetService** **DSL**.
* **InternetService** **DSL** has the highest **PhoneService** at **23**, out of which **MultipleLines** **No** contributed the most at **12**.
* **Yes** has a **PhoneService** of **13** for **InternetService** **Fiber optic**.
* **Yes|No** (**51** %) and **Yes|Yes** (**38.8** %) are the most frequently occurring categories of **PhoneService - MultipleLines** with a combined count of **44** items with **PhoneService** values (**89.8** % of the total).
* The total number of results for **PhoneService**, across all **customerID**, is **49**.

****

**Online security and online backup:**

****

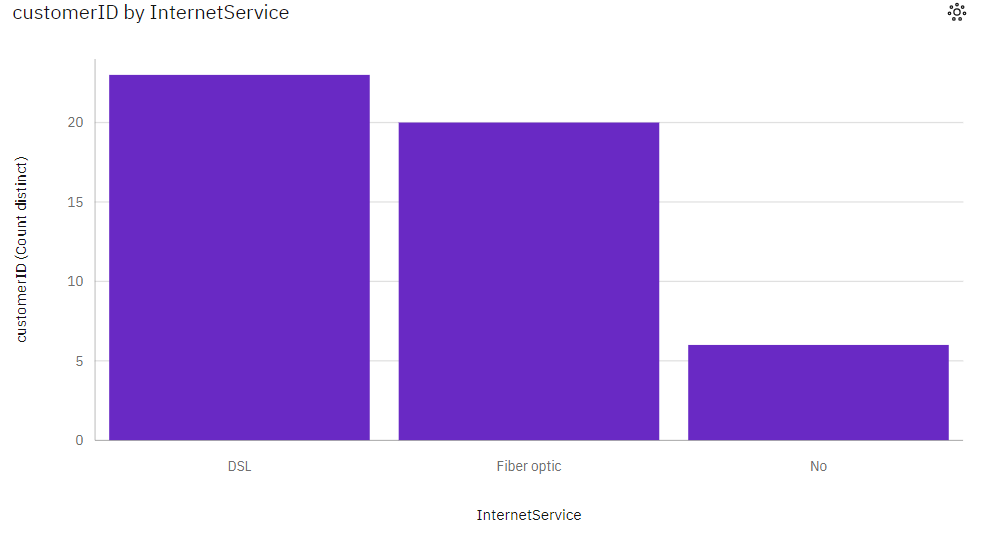
* **InternetService** **DSL** has the highest **DeviceProtection** due to **MultipleLines** **No**.
* **MultipleLines** **No** has the highest **DeviceProtection** at **4**, out of which **InternetService** **No** contributed the most at **2**.

**Dependentd by multiple lines**

****

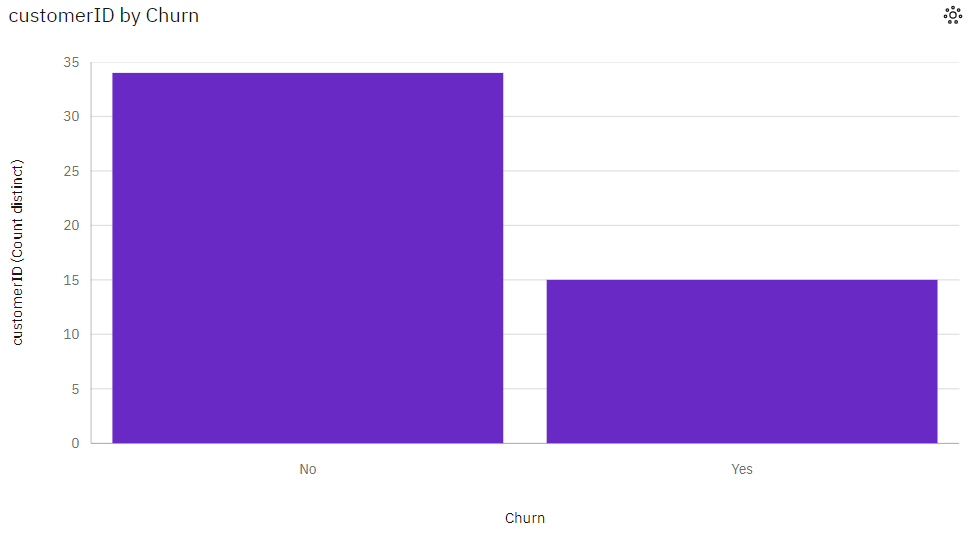
* **MultipleLines** **No** has the highest **Dependents** due to **PaperlessBilling** **No**.
* **PaperlessBilling** **No** has the highest **Dependents** at **6**, out of which **MultipleLines** **No** contributed the most at **2**.

**Internet service**

****

* **InternetService** **DSL** has the highest **customerID** due to **MultipleLines** **No**
* **MultipleLines** **No** has the highest **customerID** at **25**, out of which **InternetService** **DSL** contributed the most at **12**.
* **Fiber optic** has a **customerID** of **13** for **MultipleLines** **Yes**.
* **DSL** (**46.9** %) and **Fiber optic** (**40.8** %) are the most frequently occurring categories of **InternetService** with a combined count of **43** items with **customerID** values (**87.8** % of the total).

**Customer churn rate**

****

* **No** is the most frequently occurring category of **Churn** with a count of **34** items with **customerID** values (**69.4** % of the total).
* **No** exceeds **Yes** in **customerID** by **19**.

**Conclusion**

The dataset is now ready for further analysis and model development. The preprocessing steps undertaken have enhanced the dataset's quality and will contribute to the accuracy of the predictive model for customer churn. The next phase involves model selection, training, evaluation, and eventually utilizing the predictive model to assist the business in reducing customer churn and improving overall performance.