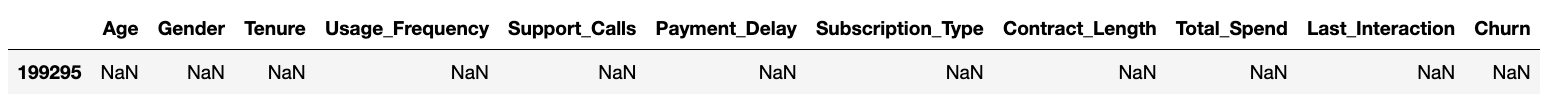
**2.2 Data preparation**

Initially, the Customer ID, which serves as a distinct numerical identifier for each customer within the organization, was eliminated. Subsequently, the vacant space that is present between variable names was substituted with an underscore character ("\_"), which facilitated the process of referencing said variable.

A screenshot of a computer

Description automatically generated

Given the cleanliness of the dataset, few preparatory measures were required, with the sole task of eliminating null values.



Furthermore, in order to investigate and examine the data, we employed a technique of converting a continuous variable into a categorical variable by dividing the age values of customers into distinct intervals of age ranges, specifically 0-20, 21-30, 31-40, 41-50, 51-60, and 61+, with the intention of facilitating the process of visual representation.

A table with numbers and text

Description automatically generated

Subsequently, the string labels were transformed into numerical labels in order to facilitate the execution of the model. In addition, it was necessary to convert other variables such as Gender, Subscription\_Type, and Contract\_Length into numerical labels for the primary analysis.

A screenshot of a table

Description automatically generated

**2.3 Exploratory Data Analysis (EDA)**

**Correlation between features except for the Churn rate**

Firstly, we tested the correlation between features except for the Churn rate.

A diagram of a heatmap

Description automatically generated

The resulting plot shows there are only 2 features (including total spend and support calls) that are highly correlated except for Churn at 0.21. Therefore, we kept all the variables of this dataset for analysis.

**Target Variable Analysis – Churn rate**

In the first place, we analyzed the target variable which is the Churn rate. By analyzing that, we can see most customers of the shop are churning!

**A diagram of a pie chart

Description automatically generatedA chart with a green and blue bar

Description automatically generated with medium confidence**

**Correlation between different features and Churn rate**

Next, we tested to see if there were any notable correlations between different features and churn.

A graph with different colors and numbers

Description automatically generated with medium confidence

From the first plot, we realize that the support calls feature has the highest correlation with Churn.

A graph of different colored bars

Description automatically generated

The second plot shows the features that have the highest correlation with the target variable.

Therefore, we will do the exploratory data analysis related to these 5 important variables.

**Support calls and Churn rate**

**A graph of support calls

Description automatically generatedA graph of a number of people

Description automatically generated with medium confidence**

On average, clients contact the support service three times. There is a notable increase in support calls ranging from 0 to 3 instances per customer. Within this range, the calls appear to have positive outcomes, as evidenced by the churn rate of these customers. As the volume of support calls increases, there is a corresponding increase in churn rates. This relationship is particularly evident when the number of calls exceeds five, as there are no remaining customers who have not churned. This suggests that the support team's effectiveness in resolving customer issues may be limited.

**Total spend and Churn rate**

**A graph of a distribution graph

Description automatically generated with medium confidenceA chart with different colored squares

Description automatically generated**

**A graph of a chart

Description automatically generated with medium confidence**

On average, the expenditure per customer amounts to approximately $650. A significant proportion of consumers who exceed a spending threshold of $500 exhibit a lower likelihood of churning.

**Payment delay and Churn rate**

A diagram of a distribution

Description automatically generatedA graph of a number of blue and green lines

Description automatically generated

It is normally distributed and slightly positively skewed. The churn rate exhibits a notable increase within the initial 20 months but experiences a substantial decline subsequent to a 20-month period of payment deferral.

**Age distribution and Churn rate**

A graph of a number of people

Description automatically generated with medium confidence A chart of a diagram

Description automatically generated with medium confidenceA graph of age distribution

Description automatically generated

The distribution of age follows a normal distribution, albeit with a modest positive skew, indicating a prevalence of younger customers.

A graph of different colored bars

Description automatically generated with medium confidence

Furthermore, we conducted a segmentation of the customer age variable into six distinct categories. Upon analysis, it became evident that a considerable proportion of customers fell within the 40-50 age bracket, while a smaller nevertheless noteworthy cohort was observed within the 20-30 age range. Approximately 50% of consumers within the age range of 20-30 exhibit churn behavior. However, the churn rate gradually decreases for subsequent age groups until reaching the 41-50 age group. Individuals who are beyond the age of 60 exhibit a propensity for churning.

**Last interaction and Churn rate**

A graph with blue lines

Description automatically generatedA diagram of a surfboard

Description automatically generated

It was observed that the final engagement of customers has a uniform distribution but with a small elevation within the range of 0 to 15. There is no substantial difference in the churn rate based on the last engagement, indicating that the last interaction has minimal or no impact on the churn rate.