**Week 2**

Aastha Giri

Department of IT, Westcliff University

TECH 400: Artificial Neural Network and Deep Learning

Prof. Subodh Acharya

November 4, 2024

# **Abstract**

In order to obtain valuable insights into customer behavior and churn-causing factors, this analysis examines a subscription-based customer dataset. Numerous trends and connections between consumer demographics, spending behaviors, and service usage are investigated through a thorough exploratory data analysis. Through distribution visualization, correlation analysis, and categorical and numeric feature analysis, this study reveals significant tendencies that can help comprehend customer retention.

Table of Contents

[Abstract 1](#_Toc181619583)

[Introduction 2](#_Toc181619584)

[Methods 3](#_Toc181619585)

[Conclusion 19](#_Toc181619586)

[References 20](#_Toc181619587)

**Table of Figures**

[figure 1: Google Drive mounting and library imports 3](#_Toc181619210)

[figure 2: Initial Data Inspection 4](#_Toc181619211)

[figure 3: Checking Missing Values 5](#_Toc181619212)

[figure 4: Summary Statistics 6](#_Toc181619213)

[figure 5: Converting Numeric and Categorical Features 6](#_Toc181619214)

[figure 6: Visualizing Distributions of Numerical Features with KDE 7](#_Toc181619215)

[figure 7: Count Plots of Categorical Features 8](#_Toc181619216)

[figure 8: Churn Count Visualization 10](#_Toc181619217)

[figure 9: Churn Analysis by Subscription Type 11](#_Toc181619218)

[figure 10: Encoding Categorical Variables for Analysis 12](#_Toc181619219)

[figure 11: Heatmap of Feature Correlations 13](#_Toc181619220)

[figure 12: Analyzing Total Spend by Subscription Type and Churn 14](#_Toc181619221)

[figure 13: Box Plots of Support Calls and Payment Delay by Churn Status 15](#_Toc181619222)

[figure 14: Box Plot of Age Distribution by Subscription Type 16](#_Toc181619223)

[figure 15: Analyzing Last Interaction by Churn and Contract Length 17](#_Toc181619224)

[figure 16: Outliers Detection using Boxplots 18](#_Toc181619225)

# **Introduction**

Exploratory Data Analysis (EDA) is a crucial first step in comprehending and getting data ready for additional analysis, particularly in domains that deal with customer retention and insights (LIM, 2024). We use an EDA on a dataset of customer subscriptions in this work to find trends and insights that can be used to forecast and comprehend customer behavior, including churn, or the phenomenon where customers stop using a service. This research looks at a range of numerical and categorical characteristics, including age, subscription type, and interaction history, in an effort to find important patterns and connections that support customer retention and satisfaction. Additionally, this investigation provides a basis for developing statistical models that can tackle important business issues like lowering attrition and improving consumer involvement.

# **Methods**

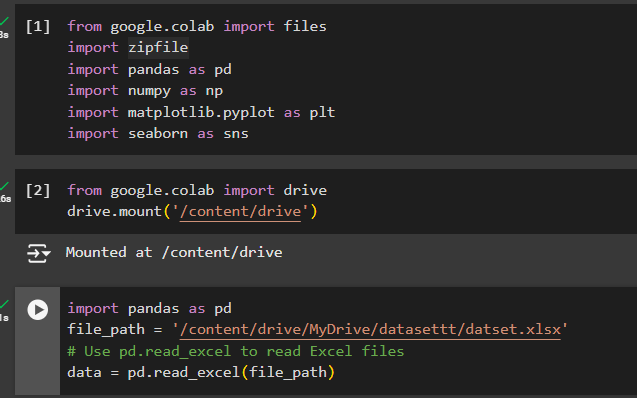


figure 1: Google Drive mounting and library imports

I began implementing the library collections that were required for EDA, such as pandas and numpy. I then accessed Google Drive in order to get the required files.

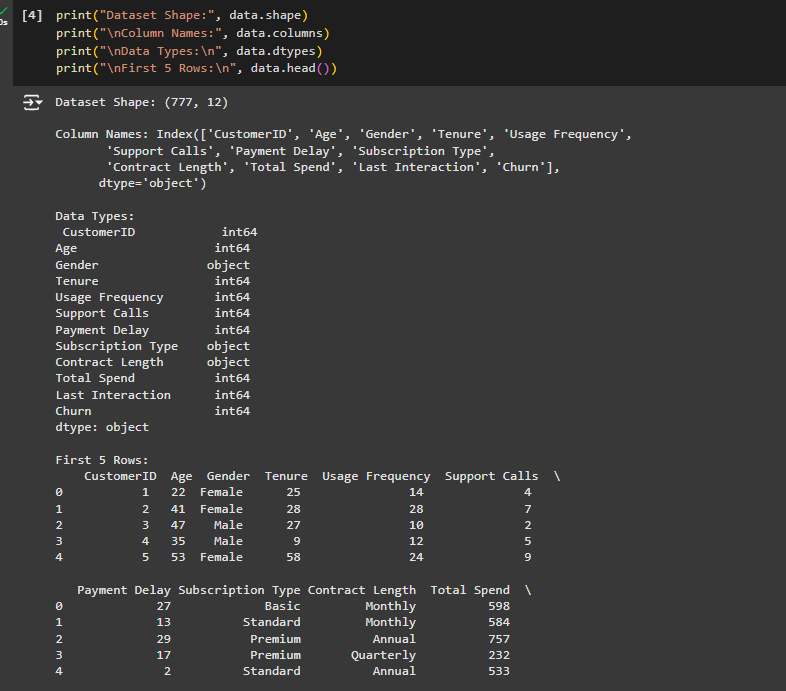


figure 2: Initial Data Inspection

Here, I performed a simple exploring data analysis on a data DataFrame. It first prints the shape of the dataset, showing that it contains 12 columns and 777 rows. It then displays all of the titles of the columns and displays the data type for each one, showing that some, like Gender and Subscription Type, are object (categorical), while the majority are int64 (numerical). The code's last section shows the dataset's first five rows, giving a brief overview and facilitating visual inspection of the column distribution and content, featuring variables such as Age, Tenure, and Total Spend. This first EDA stage aids in comprehending the structure of the dataset prior to additional analysis.

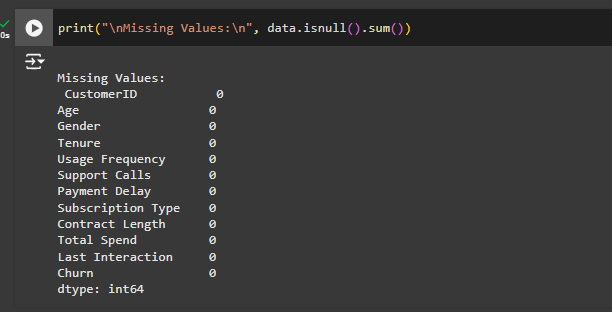


figure 3: Checking Missing Values

I looked for any missing values in any of the DataFrame data's columns. The output indicates that there's no missing values for any of the columns, proving that the data set is full.

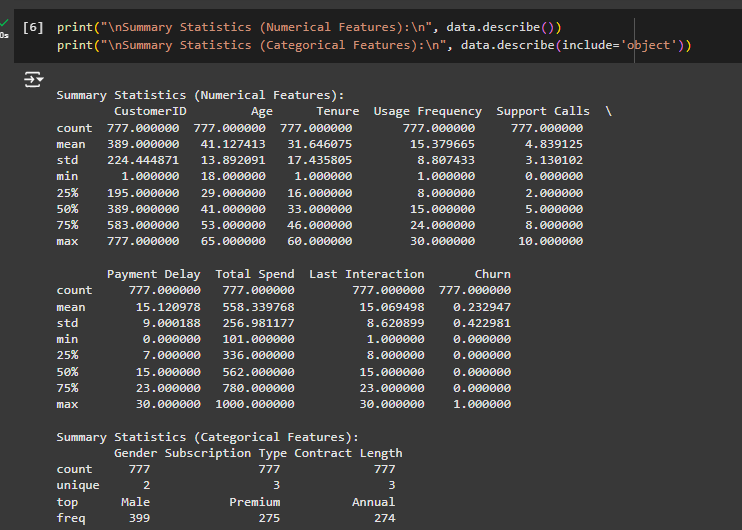


figure 4: Summary Statistics

This code provides a summary of the numerical and categorical columns in the dataset. In order to comprehend distribution and dispersion, numerical data is displayed using metrics such as mean, standard deviation, and quartiles. It shows the frequency, most prevalent category, count, and number of unique values for categorical data. This makes it easier to spot important traits and trends in the data.

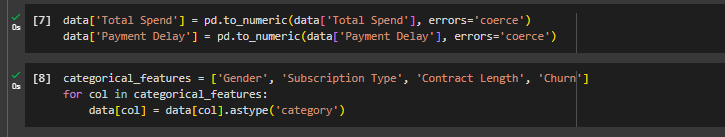
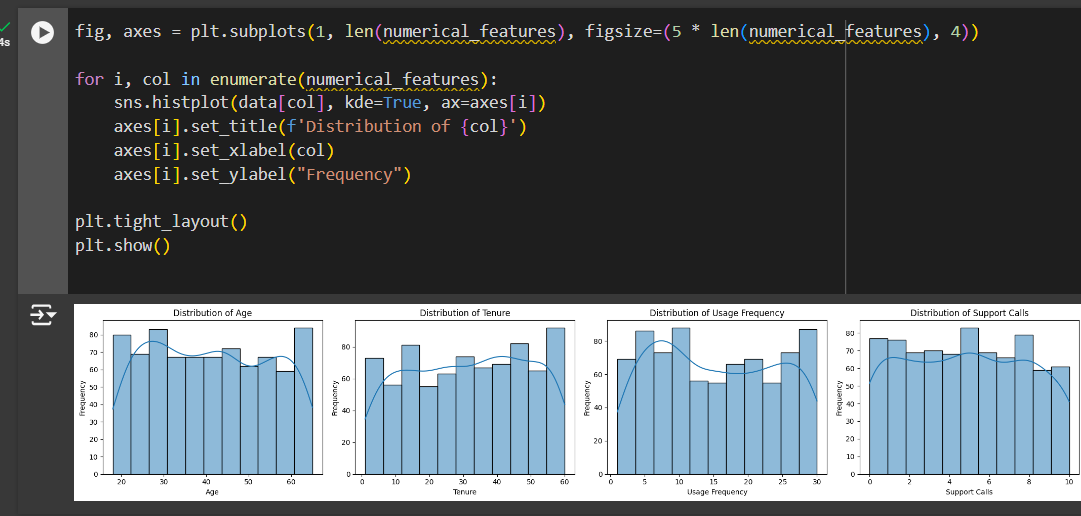


figure 5: Converting Numeric and Categorical Features

In order to handle any non-numeric values, this code first turns the two columns "Total Spend" and "Payment Delay" into numeric format and sets them to "missing." After that, it converts four more columns—"Gender," "Subscription Type," "Contract Length," and "Churn"—to a categorical format, which facilitates analysis, particularly when they reflect separate categories.



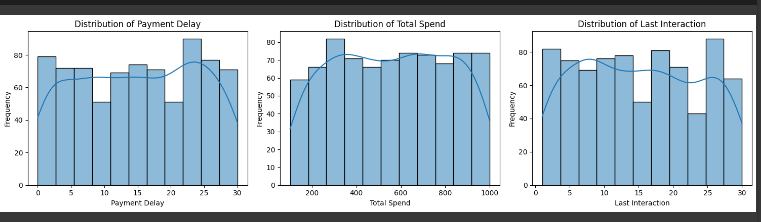
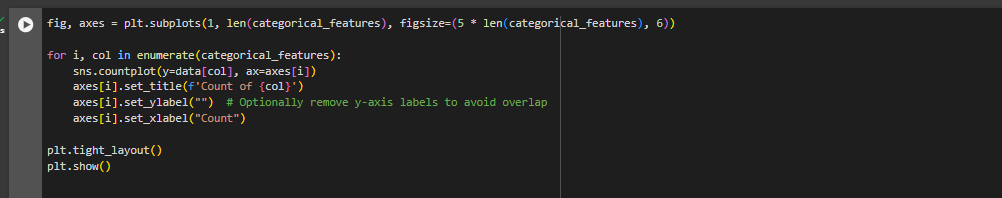


figure 6: Visualizing Distributions of Numerical Features with KDE

By producing a sequence of histograms with KDE (Kernel Density Estimation) curves superimposed, this tool illustrates the variance of many numerical traits in a dataset. In order to provide information about the distribution of a particular numerical feature and spot any trends or abnormalities, each histogram displays how its values are distributed throughout several ranges.

In bar graph:

* Age: User age distribution, showing the most common age ranges.
* Tenure: Indicates how long a user has been using the service; typical lengths are indicated.
* Usage Frequency: Indicates trends in high and low usage by reflecting how frequently users engage.
* Support Calls: Shows how frequently consumers require assistance by displaying the volume of customer support interactions.
* Payment Delay: Indicates whether late payments are common by highlighting trends in late payments.
* Total Spend: The distribution of the entire amount spent, showing how much is typically spent.
* Last Interaction: Indicates the most recent interaction a user has had.



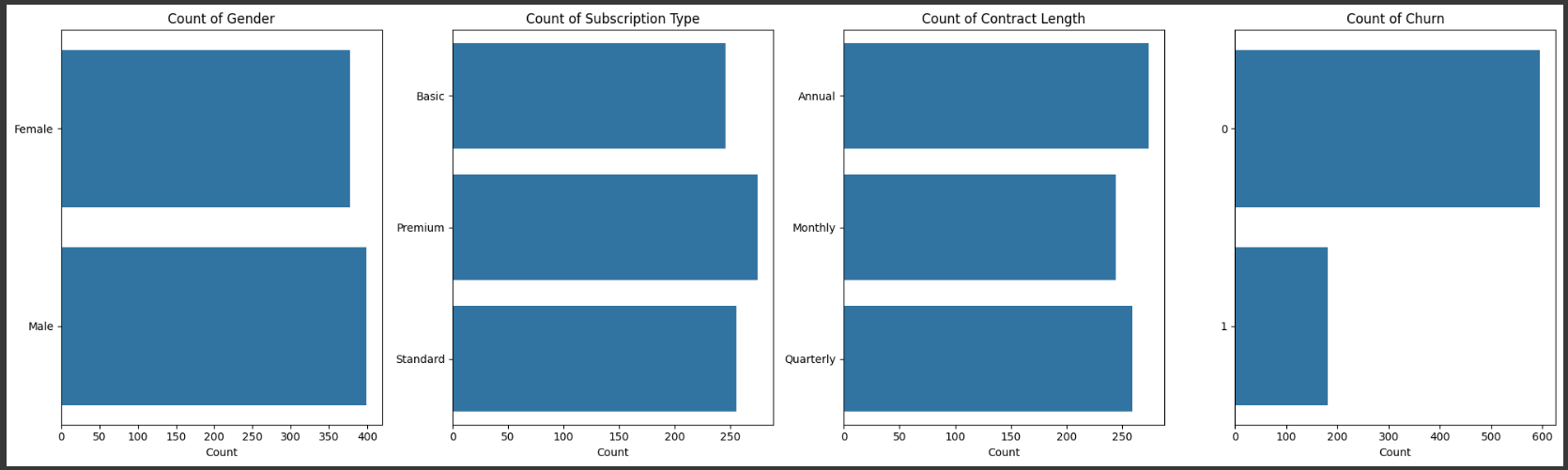


figure 7: Count Plots of Categorical Features

Using Seaborn and Matplotlib, the code creates bar charts for every categorical feature in the dataset. As it goes over each feature, it sets headings and labels for readability and creates a count plot. Lastly, it shows the charts collectively and modifies the spacing.

In the graph:

1. Count of Gender:Displays the "Male" and "Female" category distribution. At over 400 apiece, the counts for the sexes are comparable.
2. Subscription Type Count: Shows the number of "Basic," "Premium," and "Standard" subscribers. With the highest count, "Premium" is followed by "Basic" and "Standard," each of which have about 250.
3. Contract Length Count: The allocation of "Annual," "Monthly," and "Quarterly" contracts is shown. The numbers are evenly distributed, with 200–250 in each category.
4. Count of Churn: Represents a churn status of either 1 (churned) or 0 (non-churned). The proportion of non-churned (0) versus churned (1) clients is significantly higher.

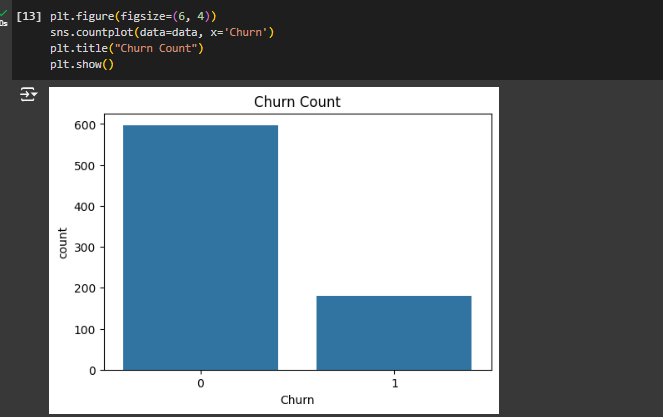


figure 8: Churn Count Visualization

The code creates a plot using a bar chart to display the pattern of customer attrition. The plot scale is set for readability, and then a count plot showing the proportion of customers that have churned against those who haven't is created using Seaborn. After labeling the chart with a title, the plot is eventually shown.

Bar explanation:

Customers who churned (1) and those who didn't (0) are shown in the following bar chart. This graphic illustrates churn dispersion, as the title "Churn Count" makes clear. Since more consumers did not churn, the threshold for churn value 0 is noticeably larger. On the other hand, the turnover value 1 bar is shorter, indicating that fewer consumers decided to quit. The dataset's imbalance among churned and non-churned clients can be seen with the aid of this graphic.

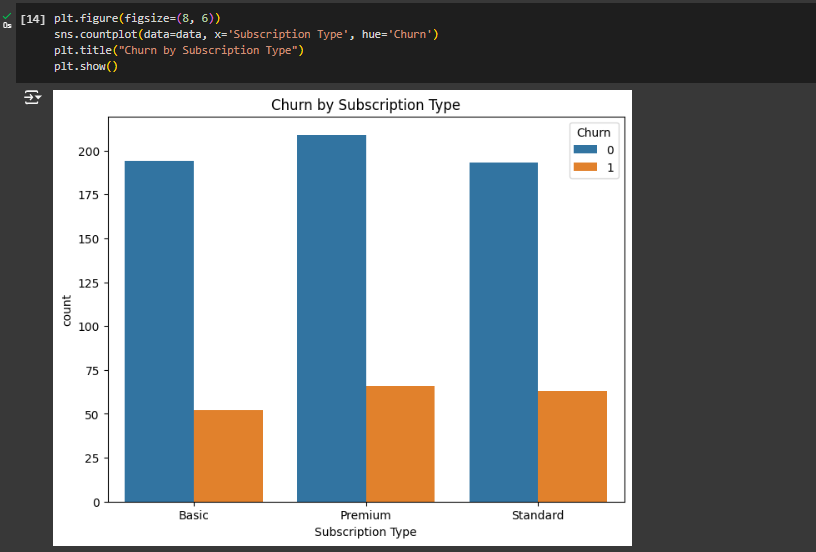


figure 9: Churn Analysis by Subscription Type

Using color to distinguish between churned and non-churned clients, the code generates a bar chart that displays the churn distribution across various subscription kinds. It shows the chart and assigns a title.

Bar graph:

Within the Basic, Premium, and Standard subscription types, the bar chart displays the pattern of churn. The two bars for each subscription type are orange for churned and blue for non-churned. The taller non-churned bar across all categories suggests that more consumers stayed rather than left. The churn rates are visually comparable for all kinds, with Premium having a larger total count, then Basic and Standard. This graph makes it easier to determine whether some subscription kinds are more inclined to churn.

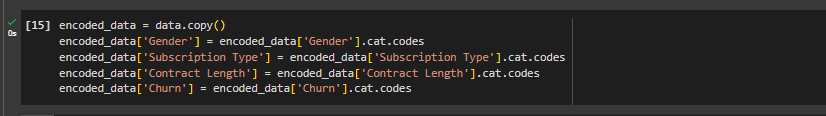


figure 10: Encoding Categorical Variables for Analysis

The original dataset is copied, and category columns are changed to numerical codes using this code. It converts the values for each categorical feature (such as "Gender," "Subscription Type," "Contract Length," and "Churn") into numerical codes, preparing the data for machine learning models that need numerical input.

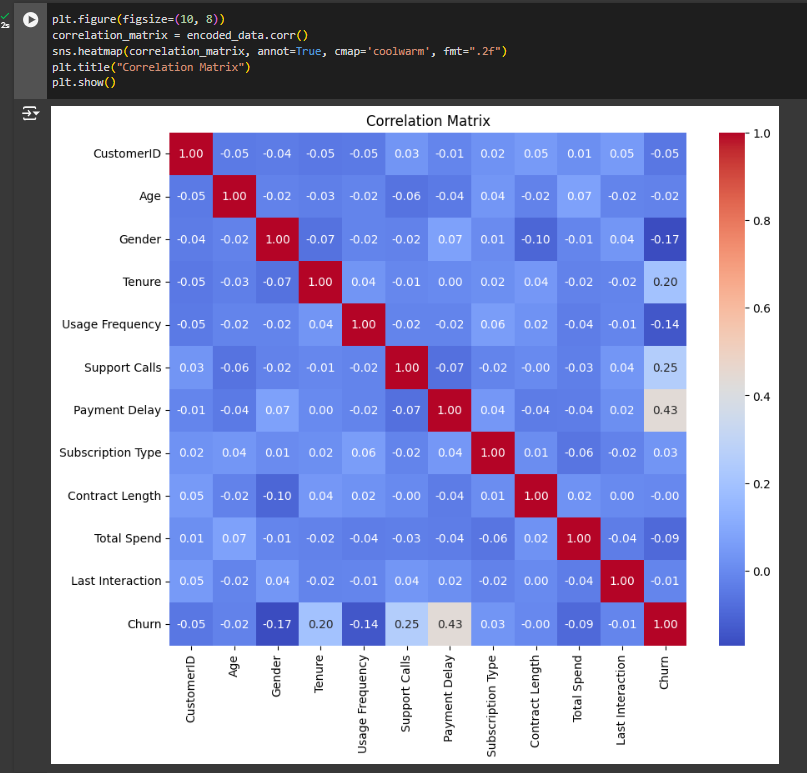


figure 11: Heatmap of Feature Correlations

A correlation matrix for a dataset is computed and shown using the code. With annot=True to display the numbers, cmap='coolwarm' for the color's intensity, and fmt=".2f" to limit values to two decimals, it utilizes sns.heatmap to generate a color-coded grid that displays the correlation values across variables.

The matrix shows correlation values, which range from -1 to 1, between two variables. Whereas negative numbers (in blue) demonstrate an inverse association, positive values (in red) suggest a direct relationship. As an example, there is a moderately positive connection (0.43) between "Churn" and "Payment Delay," indicating that consumers who experience payment delays are more inclined to leave. Because "CustomerID" is largely used as a unique identifier instead of a predictor, it does not exhibit a significant connection with other variables (values near 0).

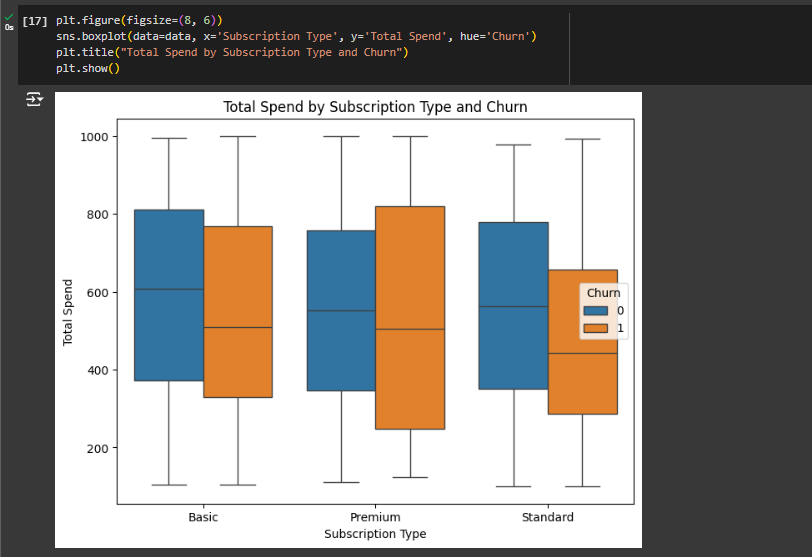


figure 12: Analyzing Total Spend by Subscription Type and Churn

The figure displays the "Total Spend" by churn status and subscriber type (Basic, Premium, Standard), with orange representing churned customers and blue representing non-churned ones. The median is displayed as a line inside each box, while the margins show the interquartile range. Outliers are points, while whiskers show the spread. The median spend of churned customers is typically lower throughout all subscription types, indicating a connection between spending and turnover.

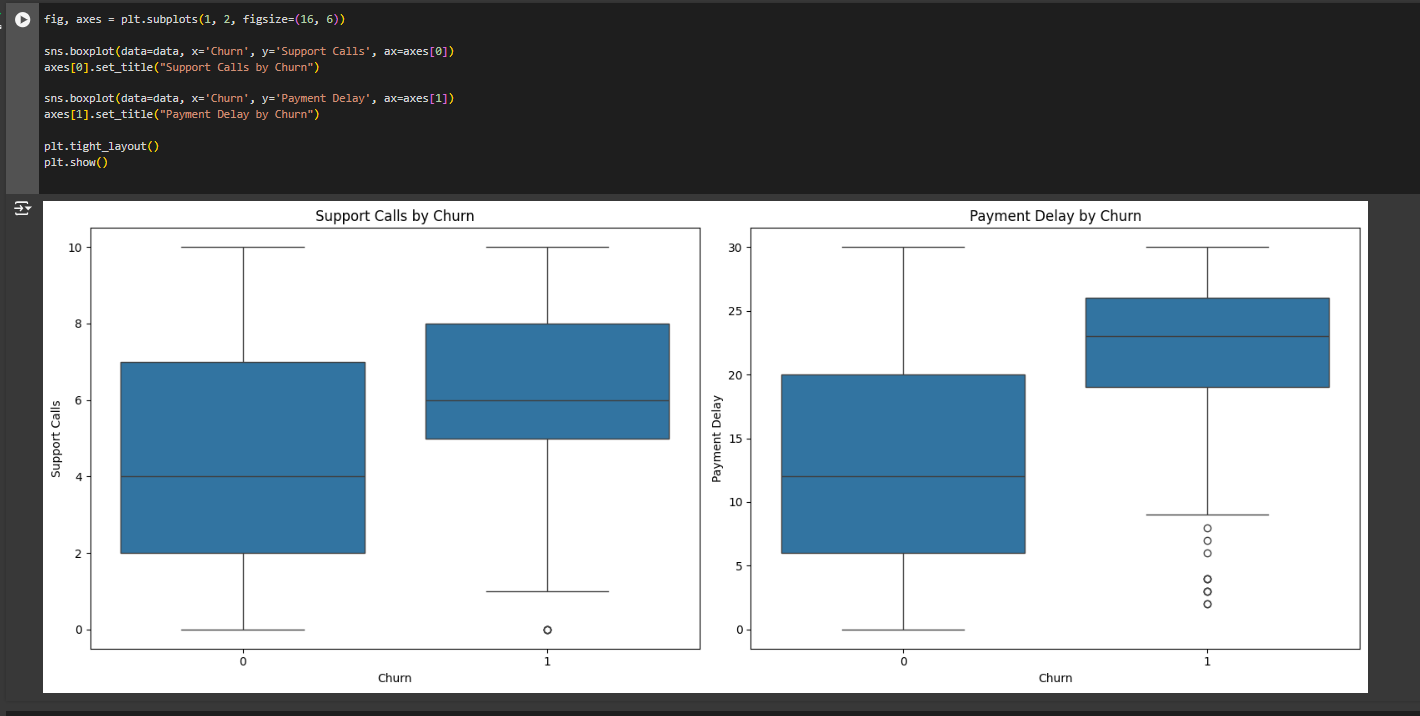


figure 13: Box Plots of Support Calls and Payment Delay by Churn Status

The box plots display how "Support Calls" and "Payment Delay" are distributed according to churn status.

* Churned customers (1) typically receive more support calls versus non-churned customers (0) in the "Support Calls" plot, indicating a connection between high support usage and churn.
* The "Payment Delay" figure shows that, with a few severe exceptions, churned customers also typically experience longer payment delays. These trends suggest that increased assistance requirements and late payments may be linked to customer attrition.

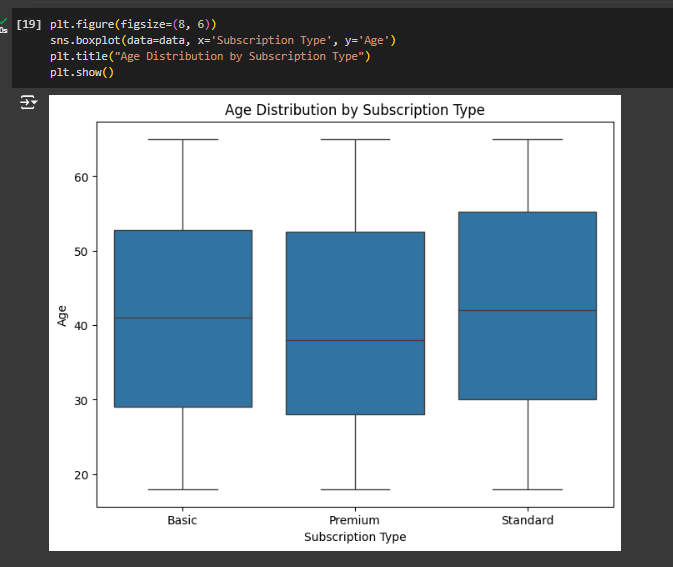


figure 14: Box Plot of Age Distribution by Subscription Type

A box plot comparing the age distributions of the Basic, Premium, and Standard subscription classes is produced by the code. A fixed plot size guarantees improved clarity, while the x-axis displays subscription types and the y-axis ages. The plot is presented together with a title that explains its goal.

The age distribution for the Basic, Premium, and Standard subscription categories is displayed in a box plot. The interquartile range, or middle 50% of ages, is displayed on the box margins, while the line inside each box indicates the median age for each kind. The total age range is depicted by whiskers, and any points outside of them would be considered outliers. With a median age of about 40 to 45 for each group, the plot shows comparable age distributions across subscription kinds.

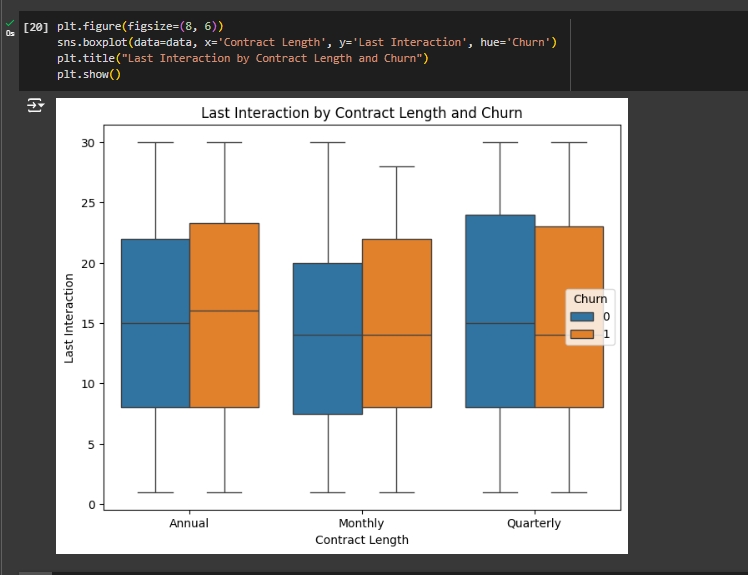


figure 15: Analyzing Last Interaction by Churn and Contract Length

The graph displays the breakdown of customers' "Last Interaction" days according to their churn status and length of contract. Each category of contract term has box plots that are colored orange for churned customers and blue for non-churned customers. Each box shows the interquartile range (box margins) and median (center line), and whiskers indicate the data's distribution. Plotting the relationship between contract type and churn likelihood and the timing of the most recent interaction is shown.

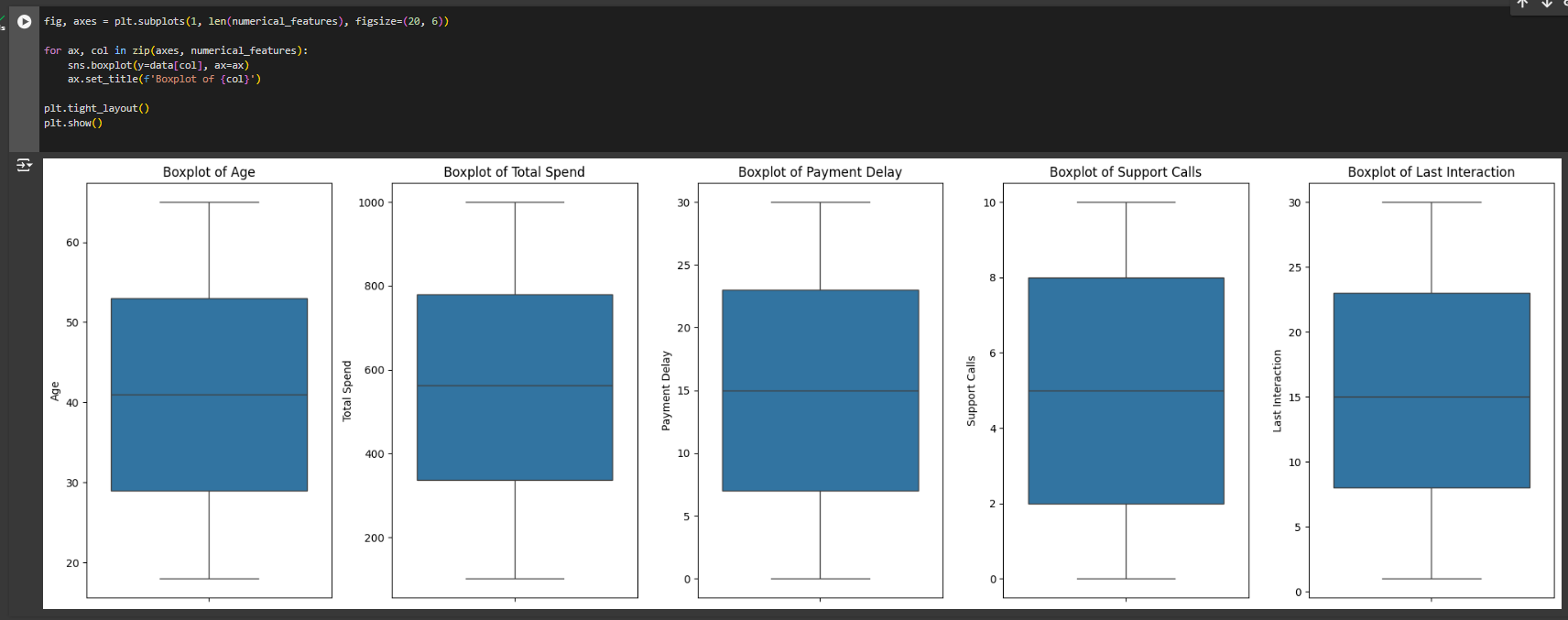
****

figure 16: Outliers Detection using Boxplots

The output displays a number of box plots for different numerical values, giving a visual representation of the distribution of each attribute. Each figure shows the data's median, interquartile range (IQR), and overall distribution, which aids in determining the variability and central tendency of each feature.

* Age: A median age of roughly 40 years is displayed by the IQR. No outliers exist in the whisker age range of 20 to 60 years.
* Total Spend: The median of the IQR for total spending is around 500. Whiskers show high variability without outliers, extending from close to 0 to 1000.
* Payment Delay: The IQR indicates a median of roughly 15 days for payment delays. With a range of 0 to 30 days, whiskers exhibit variability free of outliers.
* Support Calls: Four to six calls make up the median of the IQR for support calls. With a range of 0 to 10 calls, whiskers show moderate variability free of outliers.
* last Interaction: There is a median of 15 to 20 days in the IQR for days after the last interaction. With whiskers ranging from 0 to 30 days, interactions can vary.

# **Conclusion**

The subscription customer dataset's Exploratory Data Analysis (EDA) provided important insights into the variables influencing customer behavior and attrition. We discovered important trends, including spending habits, usage frequency, and the influence of support interactions, that are related to customer retention by examining both numerical and categorical data. Significant behavioral variations between churned and retained customers were shown by visualizations, especially across subscription kinds and contract durations. These results provide a strong basis for creating predictive models to identify high-risk clients and putting plans in place to improve client retention and satisfaction. All things considered, this analysis emphasizes how important data-driven insights are to comprehending and resolving customer attrition, which helps businesses make decisions that will maintain long-term profitability and customer loyalty.

**Github Link:** <https://github.com/AasthaGiri/Week-2-EDA->

# **References**

LIM, S. Z. (2024, April 6). Mastering Exploratory Data Analysis (EDA): Everything You Need To Know. *Medium:* <https://medium.com/data-and-beyond/mastering-exploratory-data-analysis-eda-everything-you-need-to-know-7e3b48d63a95>