Telco customer churn

A fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3.

Introduction to the Dataset

This dataset contains customer data from a fictional telecommunications company in California, with 7,043 observations and 33 variables. It includes customer demographics (e.g., gender, location, age), service subscriptions (e.g., phone, internet, streaming), billing preferences, and churn-related information (e.g., churn label, churn reason, churn score). Key metrics like Customer Lifetime Value (CLTV) and Total Charges provide insights into customer value and behavior.

As part of our project, we used this dataset to predict customer churn and analyze the factors influencing it. By studying patterns in customer behavior, we aimed to identify at-risk customers and develop strategies to improve retention. This analysis helps the company take proactive measures to reduce churn and enhance customer satisfaction

1. Data Collection

We took this dataset from Kaggle and ensured it contained key customer features such as demographics, usage patterns, and subscription details.

This dataset is widely used for analyzing customer churn and includes information about customers' services, account details, and whether they churned or not.

2. Data Description:

Categorization of Data Features

In our dataset, we classify the features into qualitative (categorical) and quantitative (numerical) variables based on their nature and usage.

Qualitative (Categorical) Variables

Categorical variables represent characteristics that describe customers and their service usage. These can be further divided into ordinal (ordered) and nominal (not ordered) variables:

- Ordinal Variables (Ordered categories):
    - Tenure Months: Represents how long a customer has been with the company in months.
    - Churn Score: A ranking that indicates the likelihood of a customer churning.
    - Contract Type: Represents contract commitments (Month-to-Month, One-Year, Two-Year).
- Nominal Variables (No inherent order):
    - Customer Identifiers: CustomerID, Country, State, City, Lat Long.
    - Demographics: Gender (Male/Female), Senior Citizen (Yes/No), Partner (Yes/No), Dependents (Yes/No).
    - Service-Related Features: Phone Service, Multiple Lines, Internet Service (DSL, Fiber optic, None).
    - Additional Services: Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies.
    - Billing and Payment: Paperless Billing (Yes/No), Payment Method.
    - Churn Indicators: Churn Label (Yes/No), Churn Reason.

Quantitative (Numerical) Variables

Numerical variables capture measurable values, which can be categorized as continuous or discrete:

- Continuous Variables (Can take any value within a range):
    - Latitude & Longitude: Geographical location coordinates of customers.
    - Monthly Charges: The monthly fee paid by the customer.

    - **Total Charges: The total amount paid by the customer over their tenure.**

- **Discrete Variables (Whole numbers, countable):**
    - **Count: A counter column for specific events or occurrences.**
    - **Zip Code: Customer's postal zip code.**
    - **Churn Value: A binary indicator (0 = No, 1 = Yes) representing customer churn.**
    - **Customer Lifetime Value (CLTV): An estimated measure of a customer's total worth to the company.**

Data Exploration:-

1. Missing Values

- The dataset has **5,174 missing values** in the Churn Reason column, which means only **1,869 churned customers** have a recorded reason for leaving.
- No other columns have missing values.

2. Numerical Data Summary

- Tenure Months: Ranges from **0 to 72 months**, with a median of **29 months**.
- Monthly Charges: Varies between **$18.25 and $118.75**, with an average of **$64.76**.
- Churn Score: Ranges from **5 to 100**, indicating customer churn likelihood.

3. Categorical Data Insights

- Gender, Senior Citizen, Partner, Dependents, and Paperless Billing are **binary** (Yes/No).
- Internet Service, Contract, and Payment Method have **3–4 unique values** each.

- City and Lat Long have many unique values, indicating a diverse customer base.

Data Processing Steps

To ensure the dataset is clean and ready for analysis, we performed the following preprocessing steps:

1. Handling Missing Values

- The Churn Reason column had 5,174 missing values, which were replaced with "Unknown" since it's categorical.
- The Total Charges column had 11 missing values, which were imputed with the median value to maintain numerical integrity.

2. Data Type Conversions

- The Total Charges column was stored as an object (string) instead of a numerical format. It was converted to a float for proper analysis.
- Churn Label, Senior Citizen, Partner, Dependents, Phone Service, and Paperless Billing were converted from Yes/No to binary (0/1) values for easier processing.

3. Encoding Categorical Variables

- Binary encoding: Converted Gender (Male = 1, Female = 0) and other Yes/No columns to binary.
- One-hot encoding: Transformed multi-category columns into separate indicator variables:
    - Internet Service (DSL, Fiber Optic, No Internet)
    - Contract (Month-to-month, One year, Two years)
    - Payment Method (Electronic Check, Mailed Check, Bank Transfer, Credit Card)

4. Dropping Unnecessary Columns

- The following columns were dropped as they did not contribute to churn prediction:
    - CustomerID (unique identifier, not useful for analysis).
    - Location-based data (City, State, Zip Code, Country, Latitude, Longitude) as they were not required for churn prediction.
    - Lat Long (duplicated geographical information).

5. Feature Scaling

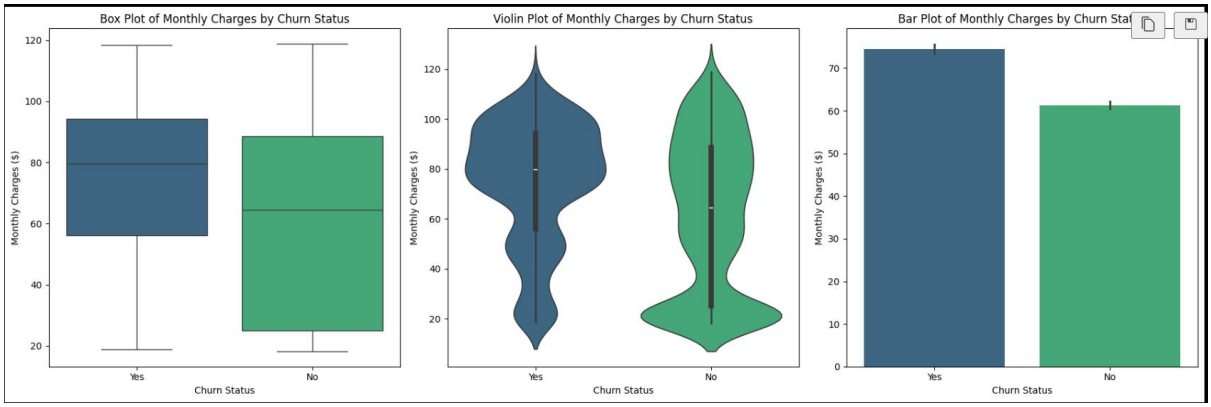
- Min-Max Scaling was applied to normalize numerical values between 0 and 1, ensuring consistent feature ranges:
    - Tenure Months
    - Monthly Charges
    - Total Charges
    - Churn Score
    - CLTV (Customer Lifetime Value)

4. Data Visualization & Insights

Monthly Charges & Churn Analysis

A **1x3 grid of subplots** was created to examine the relationship between **Monthly Charges** and **Churn Status** using the following visualizations:

- **Box Plot**: Highlights the distribution of monthly charges for churned and non-churned customers, showcasing medians, quartiles, and potential outliers.
- **Violin Plot**: Combines a box plot with a density estimate to provide deeper insights into charge distribution.
- **Bar Plot**: Compares the average monthly charges between churned and non-churned customers.



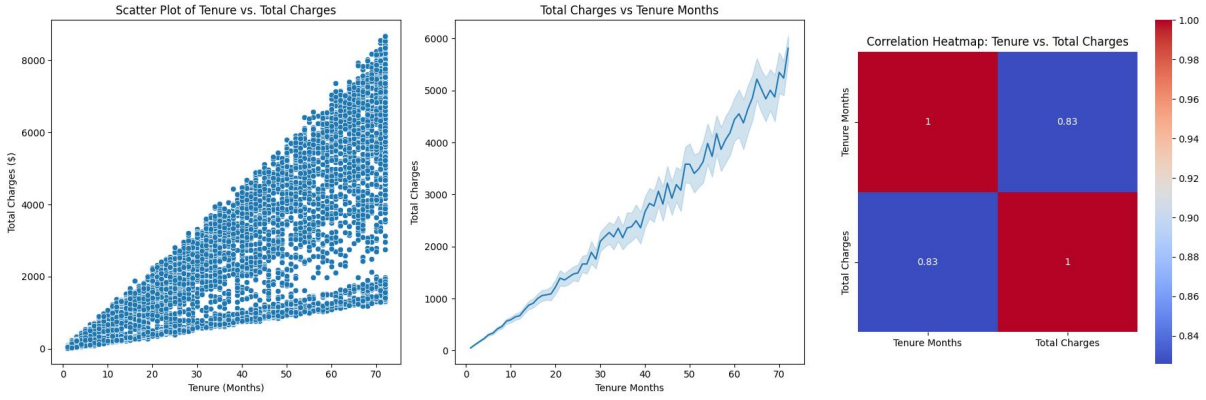
Key Insight

Customers who churn tend to have **higher monthly charges** than those who stay. This suggests that pricing plays a critical role in customer retention—higher costs may lead customers to explore alternative providers or more affordable plans.

Tenure & Total Charges Analysis

A **1x3 grid of subplots** was used to analyze how **Tenure Months** and **Total Charges** interact:

- **Scatter Plot**: Displays individual data points to detect trends, clusters, or anomalies.
- **Line Plot**: Illustrates how total charges increase over time with tenure.
- **Correlation Heatmap**: Quantifies the strength and direction of the relationship between tenure and total charges.



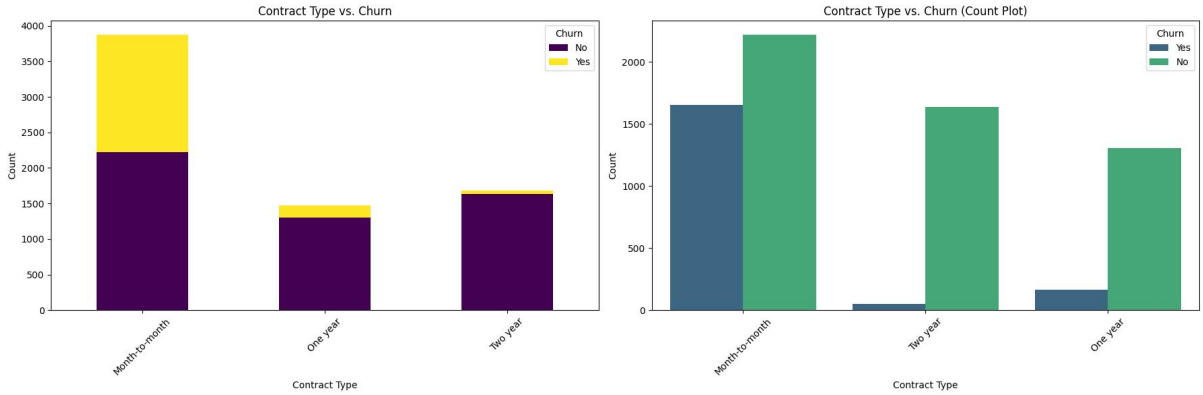
Key Insights

- A **strong positive correlation (~0.83)** indicates that **longer-tenured customers** naturally accumulate higher total charges.
- Some **long-term customers** have **lower charges**, possibly due to discounts or downgrades.
- **Short-tenured customers** with **high charges** may have premium plans or additional services.
- Data gaps could indicate missing billing information.

Contract Type & Churn Analysis

A **1x2 grid of subplots** was used to compare **Contract Types** against **Churn Status** using:

- **Stacked Bar Plot**: Shows the proportion of churned vs. non-churned customers across different contract types.
- **Count Plot**: Directly compares churn rates for each contract type.



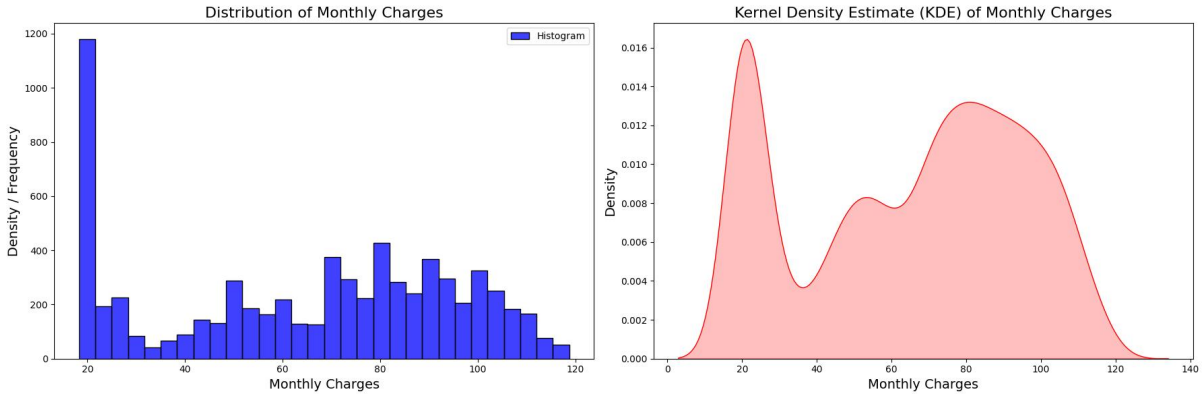
Key Insights

- **Month-to-month contracts** have the **highest churn rate (42.7%)** due to their flexibility.
- **One-year contracts** show **moderate churn (11.3%)**, indicating a slightly stronger commitment.
- **Two-year contracts** have the **lowest churn rate (2.8%)**, suggesting that longer commitments increase customer retention.

Monthly Charges Distribution Analysis

A **1x2 grid of subplots** was used to analyze the distribution of **Monthly Charges** through:

- **Histogram**: Displays the frequency distribution of charges.
- **Kernel Density Estimate (KDE) Plot**: Provides a smoothed view of the distribution.

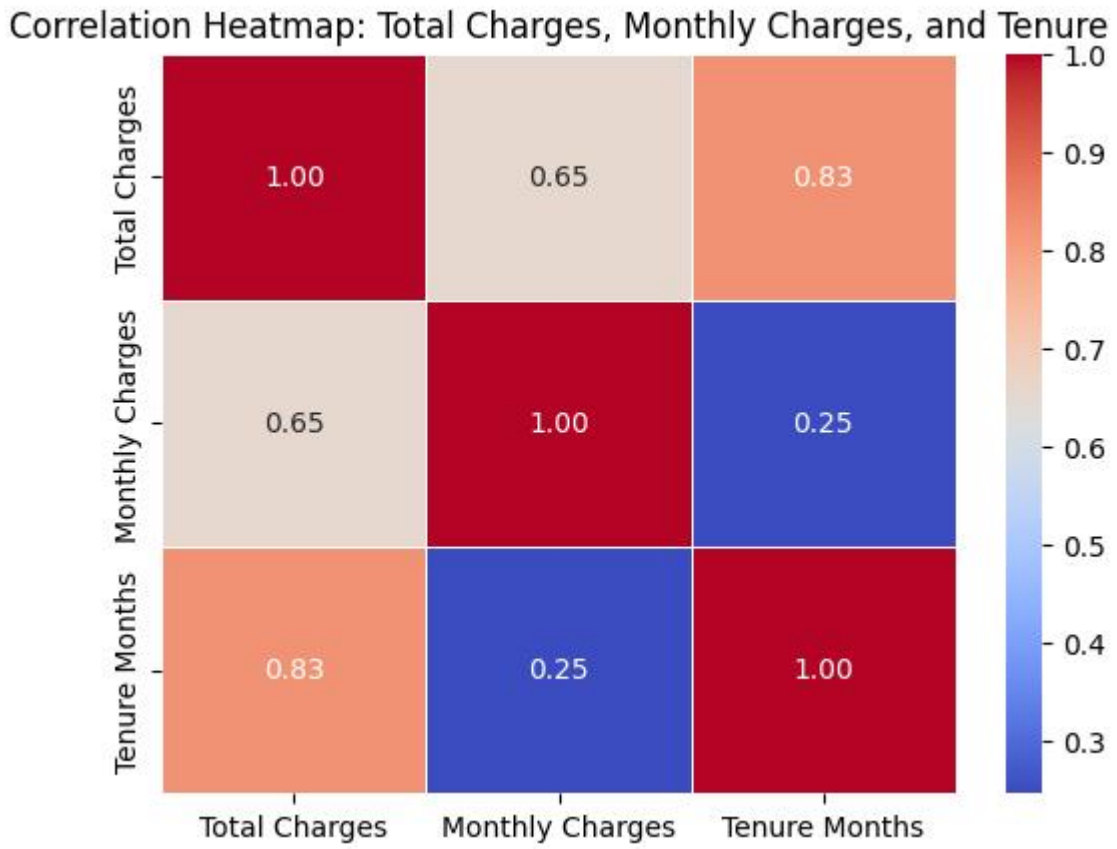


Key Insights

- Monthly charges range from **$18.25 to $118.75**, with an **average of $64.76** and a **standard deviation of $30.09**.
- The **bimodal or right-skewed distribution** suggests two major customer segments:
    - **Low-cost plans** (~$20–$40)
    - **Higher-cost plans** (~$70–$100)
- The KDE plot confirms that more customers fall into the **mid-to-premium pricing range ($70–$90)**.

Correlation Analysis

A **heatmap** was created to analyze relationships between **Tenure, Monthly Charges, and Total Charges**.

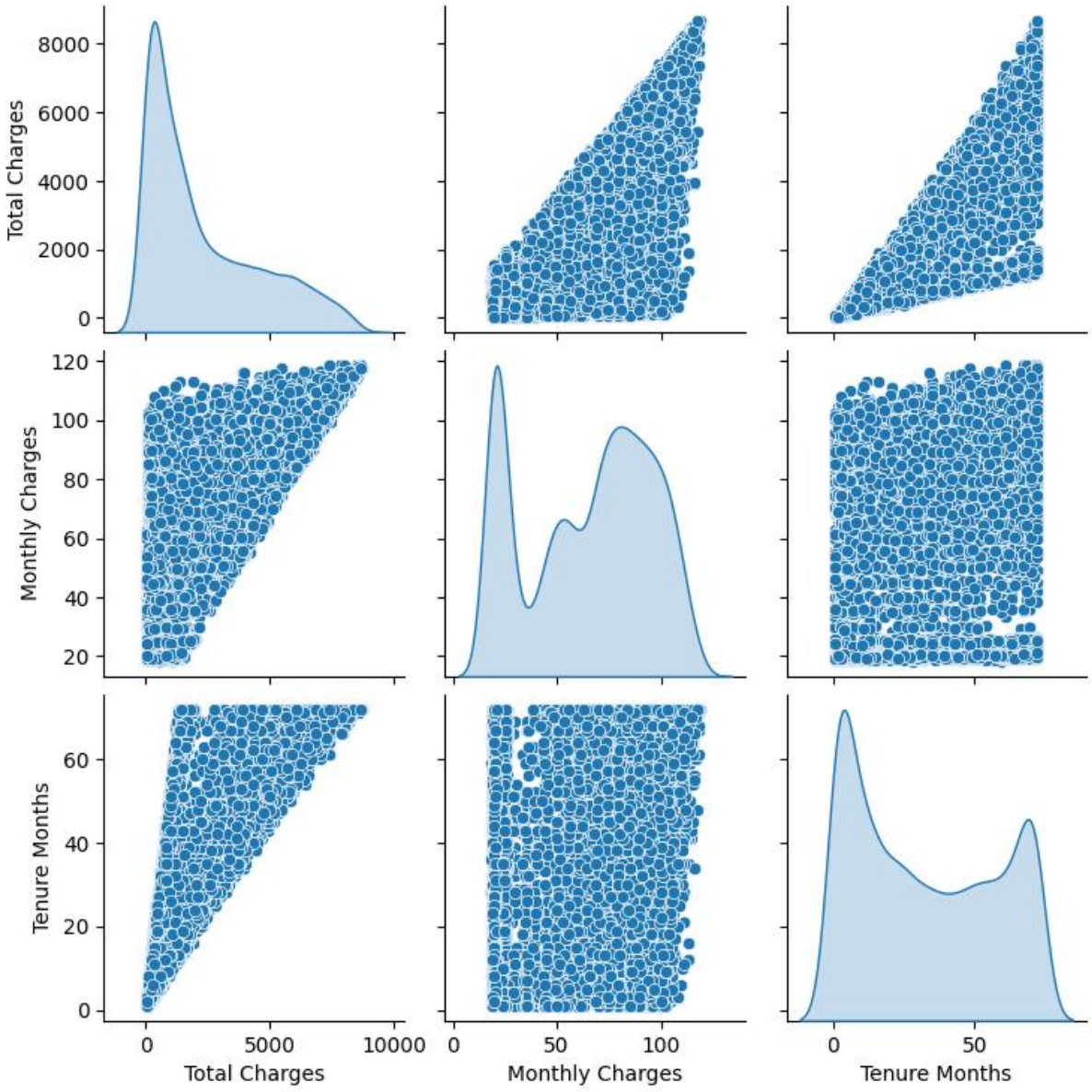


Key Insights

- **Total Charges vs. Tenure (0.83 – Strong Positive Correlation)** → Longer tenure results in higher total charges.
- **Total Charges vs. Monthly Charges (0.65 – Moderate Positive Correlation)** → Higher monthly charges generally lead to higher total charges, but tenure impacts this relationship.
- **Monthly Charges vs. Tenure (0.25 – Weak Correlation)** → Longer tenure does not necessarily mean higher monthly charges, indicating that customers may be on legacy or discounted plans.

Pairwise Feature Analysis

A **pairplot (scatterplot matrix)** was created to examine relationships between **Tenure, Monthly Charges, and Total Charges.**

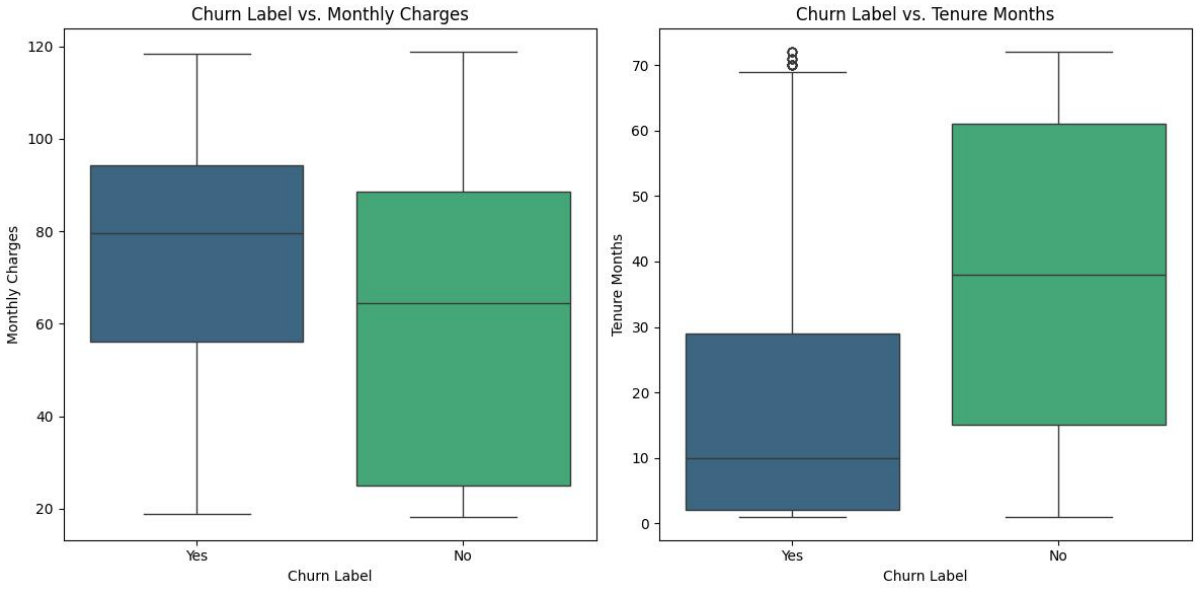


Key Insights

- **Total Charges vs. Tenure (Strong Positive Trend)** → A clear linear relationship, confirming that tenure directly affects total charges.
- **Total Charges vs. Monthly Charges (Moderate Positive Trend)** → More variation exists, indicating that total charges are influenced by both tenure and monthly charges.
- **Monthly Charges vs. Tenure (Weak Correlation)** → No strong trend, suggesting that tenure does not dictate monthly charge levels.

Churn vs. Monthly Charges & Tenure

A **1x2 grid of boxplots** was created to compare **Monthly Charges and Tenure** against **Churn Status**.

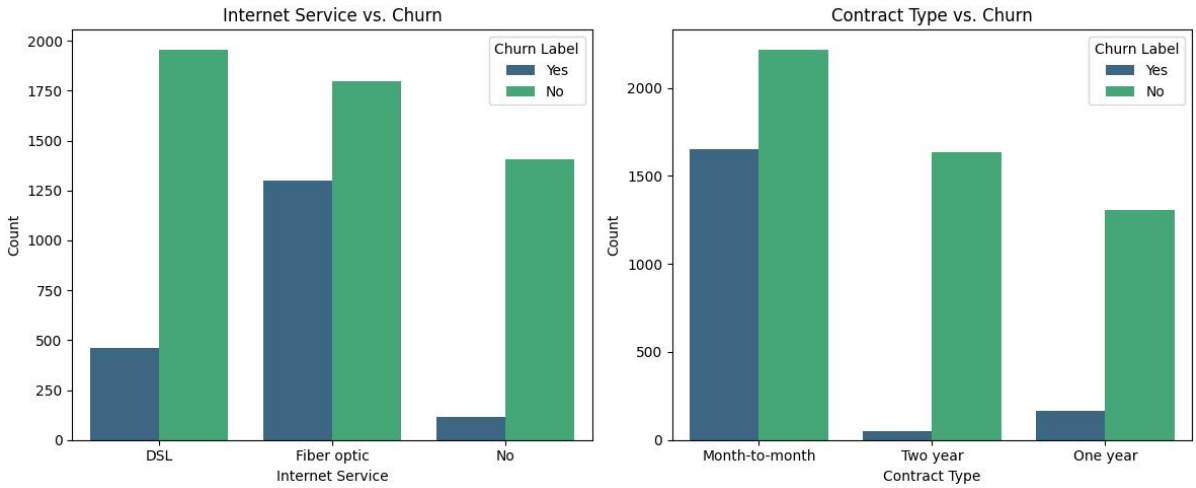


Key Insights

- **Churn vs. Monthly Charges** → Customers who churn **pay higher monthly charges**, indicating dissatisfaction or a perception of low value for money.
- **Churn vs. Tenure** → Churned customers typically have **lower tenure**, meaning new customers are more likely to leave before forming long-term loyalty.

Internet Service & Contract Type vs. Churn

A **1x2 grid of count plots** was used to analyze **Internet Service and Contract Type** in relation to **Churn Status**.



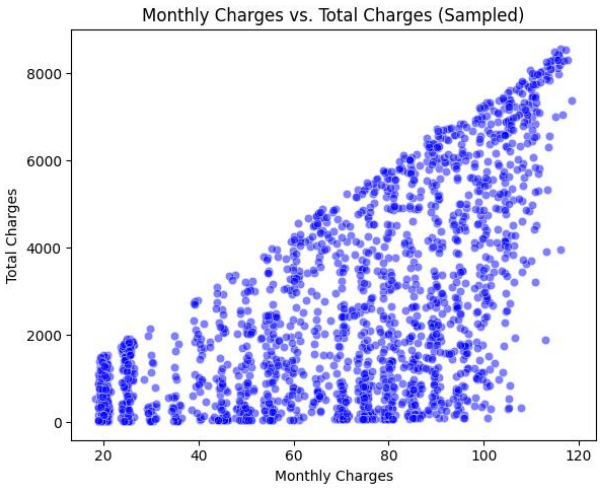
Key Insights

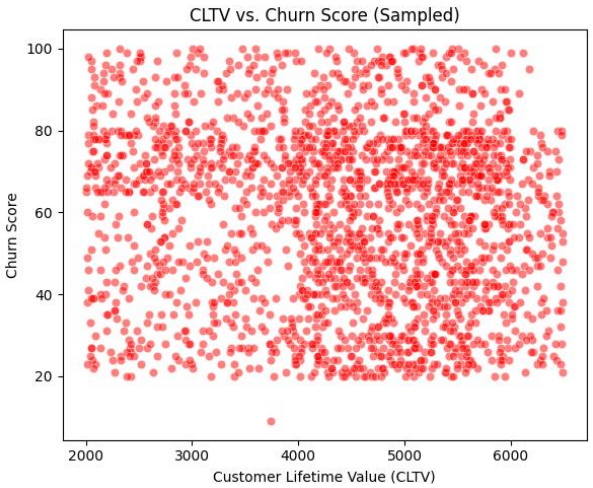
- **Fiber Optic users** have the **highest churn rate**, likely due to pricing or service quality issues.
- **Customers without internet service** have **the lowest churn**, possibly because they only use basic phone services.
- **Month-to-month contracts** drive **higher churn**, while **longer contracts** improve customer retention.

Scatter Plot Analysis: CLTV & Churn Score

A **1x2 grid of scatter plots** was used to explore:

- **Monthly Charges vs. Total Charges** → Identifies how charges accumulate over time.
- **CLTV vs. Churn Score** → Assesses whether high-value customers are at greater risk of churn.





Key Insights

- **Higher monthly charges contribute to higher total charges**, but tenure plays a significant role.
- **Customers with high CLTV and high churn scores** require targeted retention strategies to prevent revenue loss.

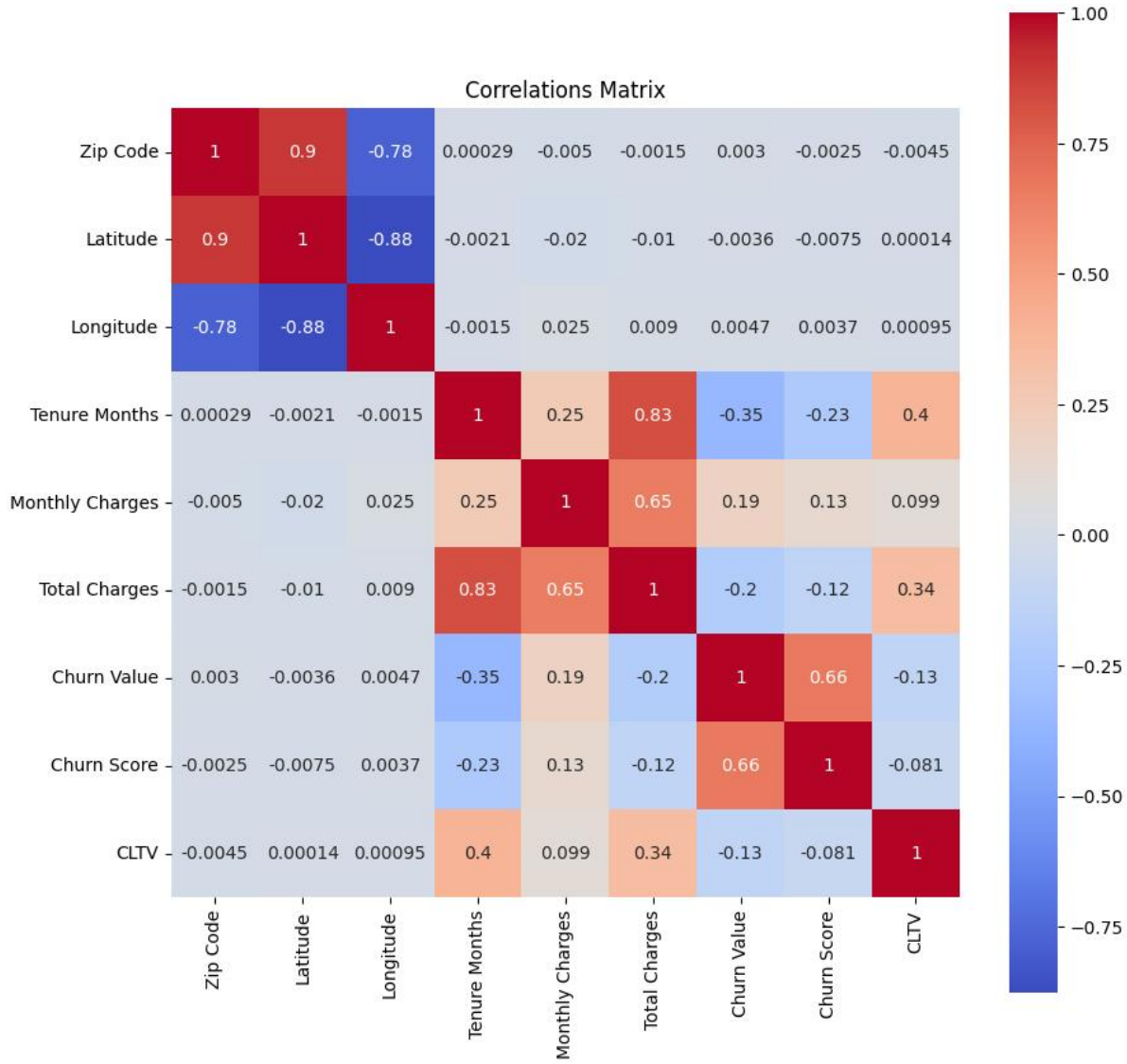
Overall Correlation Analysis

A **comprehensive correlation heatmap** was created to examine relationships between **all numerical features**.

Key Insights

- **Tenure & Total Charges (0.99 - Strongest Correlation)** → Customers who stay longer contribute significantly more revenue.
- **Weak or No Correlation Between Some Variables** → Certain numerical attributes are not strong predictors of one another.
- **Potential Predictors for Churn** → If churn-related features like **Churn Score** or **Monthly Charges** have strong correlations with **Churn Label**, they can serve as key indicators for customer retention strategies.

Conclusion



This analysis highlights several crucial factors influencing customer churn:

- **High monthly charges** correlate with **higher churn rates**, indicating pricing sensitivity among customers.
- **Longer-tenured customers** contribute **higher total revenue**, reinforcing the value of customer retention strategies.
- **Contract type significantly impacts churn**, with month-to-month contracts leading to higher customer attrition.
- **Internet service type affects churn rates**, especially for fiber optic users, suggesting potential service quality concerns.

- **Churn risk is higher for new customers**, emphasizing the need for early engagement and retention efforts.

These insights provide a **data-driven foundation** for developing **pricing strategies, customer retention initiatives, and targeted marketing campaigns** aimed at reducing churn and improving overall customer satisfaction.

Summarization to the notebook and the report

1. Notebook Overview

This notebook takes a deep dive into the Telco Customer Churn dataset, exploring patterns behind customer retention and churn. The analysis follows these key steps:

- Data Cleaning & Preprocessing:
    - Converted the *Total Charges* column into a numeric format.
    - Handled missing values to ensure data consistency.
    - Standardized categorical variables for uniformity.
- Exploratory Data Analysis (EDA):
    - Generated statistical summaries of key features.
    - Created visualizations to uncover trends in churn behavior.
    - Examined correlations between customer tenure, charges, and churn rates.

---

2. Key Insights from Data Exploration

- **Customer Tenure & Charges:**
    - Customers who stay longer accumulate higher total charges.
    - However, monthly charges remain fairly constant across tenure.

- **Churn Analysis:**
    - Customers with higher monthly charges are more likely to churn.
    - Those on long-term contracts (one or two years) are much less likely to leave.
- **Contract & Service Influence:**
    - Month-to-month contracts have the highest churn rate, indicating a lack of long-term commitment.
    - Internet service type plays a major role in churn—fiber-optic users tend to leave more frequently.

Cleaned Dataset

The dataset has been fully processed and is now ready for machine learning and predictive modeling. Key improvements include:

Consistent categorical formats for better analysis.
Missing values handled to avoid data inconsistencies.
Normalized numerical features to improve model accuracy.
Encoded categorical features for seamless machine learning applications.

These transformations ensure better interpretability and usability for predictive analytics.

Interactive Dashboard

1. Overview

To make data exploration more intuitive, we built an interactive dashboard using Dash & Plotly. This tool allows users to visually explore customer churn trends, identifying key drivers of customer retention.

2. Features & Visualizations

- **Monthly Charges Analysis**
    - A histogram and density plot reveal how charges are distributed across customers.
- **Tenure vs. Total Charges**
    - A scatter plot and line plot illustrate customer spending trends over time.
- **Contract Type vs. Churn**
    - A stacked bar chart highlights how contract types impact churn rates.
- **Churn Analysis**
    - Box plots and violin plots compare monthly charges across churned and non-churned customers.
- **Correlation Heatmap**
    - A heatmap uncovers key relationships between numerical features.

---

3. Insights Gained

Customers with higher monthly charges are at a greater risk of churning. Longer tenure reduces churn probability, especially for customers on annual contracts.

Customers with fiber-optic internet experience higher churn rates, possibly due to pricing or competition.

With these insights, businesses can take data-driven actions to reduce churn, improve customer experience, and optimize pricing strategies.