Report: Analysis of Telco Customer Churn Dataset

Chosen Data set is Customer churn, it is a critical metric for businesses, especially in the telecom sector, where retaining customers is often more cost-effective than acquiring new ones. This report analyzes the Telco Customer Churn Dataset to identify patterns, key drivers of churn, and actionable insights to reduce customer attrition.

1. Data Cleaning

1.1 Dataset Loading and Inspection

- Dataset loaded: "WA_Fn-UseC_-Telco-Customer-Churn.csv"
- Initial shape: 7,043 rows and 21 columns
- Data types: 18 object columns, 2 integer columns (SeniorCitizen, tenure), and 1 float column (MonthlyCharges)

1.2 Handling Missing Values

- Initial check revealed no explicit missing values
- Further inspection found 11 blank strings in the TotalCharges column
- Converted TotalCharges from object to numeric type, which turned blank strings into NaN values
- Imputed missing TotalCharges values with the median value
- Verified all missing values were successfully addressed

1.3 Removing Duplicate Records

- Checked for duplicate records in the dataset
- No duplicates were found; dataset maintained its original shape
- Dataset shape after duplicate check: 7,043 rows

1.4 Outlier Detection and Treatment

- Created box plots to visualize outliers in MonthlyCharges, TotalCharges, and tenure
- Applied IQR method to identify potential outliers
- Found no significant outliers that needed removal in MonthlyCharges
- Dataset shape remained unchanged after outlier treatment

1.5 Standardizing Categorical Values

- Standardized InternetService values by removing spaces (e.g., "Fiber optic" to "Fiberoptic")
- Modified PaymentMethod values by removing "(automatic)" from the text

- Verified the changes with unique value checks
- Final standardized values:
    - InternetService: ['DSL', 'Fiberoptic', 'No']
    - PaymentMethod: ['Electronic check', 'Mailed check', 'Bank transfer', 'Credit card']
- Cleaned dataset exported as "telco_churn_cleaned.csv"

2. Exploratory Data Analysis (EDA)

2.1 Univariate Analysis

2.1.1 Numerical Variables

- **Tenure**:
    - Mean: 32.37 months
    - Standard deviation: 24.56 months
    - Range: 0-72 months
    - Distribution: Histogram showed somewhat uniform distribution across values
- **MonthlyCharges**:
    - Mean: $64.76
    - Standard deviation: $30.09
    - Range: $18.25-$118.75
    - Distribution: Slightly bimodal distribution observed in histogram
- **TotalCharges**:
    - Mean: $2,281.92
    - Standard deviation: $2,265.27
    - Range: $18.80-$8,684.80
    - Distribution: Right-skewed distribution as shown in histogram

2.1.2 Categorical Variables

- **Churn Distribution**:
    - No: 73.46%
    - Yes: 26.54%
- **Contract Types**:
    - Month-to-month: 3,875 customers
    - Two year: 1,695 customers
    - One year: 1,473 customers
- **Payment Methods**:
    - Distribution visualized through bar plots showing counts across different methods

2.2 Bivariate Analysis

2.2.1 Numerical vs Numerical

- **MonthlyCharges vs TotalCharges**:
    - Scatter plot showed positive relationship between the two variables
    - Points colored by Churn status showed some clustering patterns
- **Correlation Matrix**:
    - Tenure and TotalCharges: Strong positive correlation (0.83)
    - MonthlyCharges and TotalCharges: Moderate positive correlation (0.65)
    - Tenure and MonthlyCharges: Weak positive correlation (0.24)

2.2.2 Categorical vs Numerical

- **Churn vs MonthlyCharges**:
    - Box plot showed differences in MonthlyCharges distribution between churned and non-churned customers
- **Contract vs Tenure**:
    - Violin plot revealed distinct tenure distributions across different contract types
    - Month-to-month contracts had lower tenure values
    - Two-year contracts had higher tenure values

2.2.3 Categorical vs Categorical

- **Churn vs Contract**:
    - Stacked bar plot showed relationship between contract types and churn status
    - Month-to-month contracts had higher proportion of churned customers
    - Two-year contracts had lower proportion of churned customers

2.3 Multivariate Analysis

2.3.1 Pair Plot

- Created a pair plot using a sample of 200 rows
- Visualized relationships between tenure, MonthlyCharges, and TotalCharges
- Points colored by Churn status to identify patterns

2.3.2 Heatmap for Categorical Interactions

- Converted Churn to numeric (Yes: 1, No: 0)
- Created a pivot table of churn rates by InternetService and Contract
- Visualized the interaction with a heatmap
- Color intensity (blue) indicated higher churn rates

2.3.3 Faceted Analysis

- Created a facet grid of Tenure vs MonthlyCharges
- Faceted by Churn status and Contract type
- Showed the relationship between variables across different segments

The data cleaning process successfully addressed missing values, checked for duplicates, and standardized categorical variables. The exploratory data analysis revealed important insights about the dataset's distributions and relationships between variables.

Key observations from the analysis:

- The dataset has a churn rate of 26.54%
- Contract type appears to have a strong relationship with churn
- Tenure, MonthlyCharges, and TotalCharges show varying degrees of correlation
- The interaction between InternetService type and Contract type appears to influence churn rates

This analysis provides a foundation for understanding the factors that may contribute to customer churn in the telecommunications company.