**Data Preparation Documentation**

**Overview**

Data Preparation section aims to prepare the raw data for customer churn analysis in a telecommunications company. Data cleaning and preprocessing steps were performed to ensure quality and suitability for machine learning models.

**Dataset Used**

* **Source:** WA\_Fn-UseC\_-Telco-Customer-Churn.csv
* **Rows:** 7,043
* **Columns:** 21 (before cleaning)

**Preparation Steps**

**1. Data Loading**

* The dataset was loaded from a CSV file using pandas.
* Required libraries such as matplotlib and seaborn were imported for visualization.

**2. Initial Exploration**

* Displayed the first 5 rows to understand the structure.
* Checked the dataset dimensions (shape) and data types (info).
* Removed the unnecessary column “customerID” since it does not affect modeling.

**3. Missing Values Check**

* No missing values were found in the dataset (isnull().sum()).

**4. Unique Values & Variable Inspection**

* Inspected unique values in each column to identify categorical and numerical features.
* Observed that the column “TotalCharges” was stored as object (string) even though it represents numeric values.

**5. Data Processing**

* Converted “TotalCharges” to numeric (float) using pd.to\_numeric with errors='coerce'.
* Replaced NaN values resulting from conversion with 0.

**6. Distribution Analysis**

* Created boxplots for each numerical variable to detect outliers.
* Used sns.countplot to visualize the target variable Churn, where class imbalance was observed:
  + No: 5,174
  + Yes: 1,869

**Results**

* The dataset is now ready for the next stage of analysis and modeling.
* The data type of “TotalCharges” was corrected.
* No missing values require additional handling.