Project Title	**Predicting Customer Churn in the Telecom Industry**
Technologies Used	**SQL, Power BI**
Domain	**Telecom Industry**

Data Preprocessing Documentation

Data preprocessing is a crucial step in the data analysis that ensures the dataset is clean, consistent, and suitable for analysis or model development. This step aims to transform raw data into a format that can be effectively used for analysis, ensuring that it is free from errors, missing values, and irrelevant features.

Steps Involved in Data Preprocessing:

Tools Used: Excel, SQL, MS Docs, PPT Deck, Power BI

Data Collection

- Collected the dataset provided by the "**Guvi Team"**
- Conducted a thorough review of the dataset, which contains **7044 rows and 38 columns**

Data Overview

This dataset is ideal for analysing factors that contribute to customer churn in the telecom industry. The dataset consists of the following columns.

- Customer ID: Unique identifier for each customer
- Gender: Customer's gender
- Age: Customer's age
- Married: Marital status of the customer
- Number of Dependents: Number of people dependent on the customer
- City: Customer's city of residence
- Zip Code: Postal code of the customer's address
- Latitude: Geographic latitude of the customer's location
- Longitude: Geographic longitude of the customer's location
- Number of Referrals: How many new customers this customer has referred
- Tenure in Months: How long the customer has been with the company
- Offer: Type of offer the customer has accepted
- Phone Service: Whether the customer has phone service
- Average Monthly Long-Distance Charges: Average monthly charges for long-distance calls
- Multiple Lines: Whether the customer has multiple phone lines
- Internet Service: Whether the customer has internet service
- Internet Type: Type of internet connection (e.g., DSL, Fiber optic, Cabel)
- Average Monthly GB Download: Average monthly data usage in gigabytes
- Online Security: Whether the customer has online security service
- Online Backup: Whether the customer has online backup service
- Device Protection Plan: Whether the customer has a device protection plan
- Premium Tech Support: Whether the customer has premium tech support
- Streaming TV: Whether the customer uses TV streaming services
- Streaming Movies: Whether the customer uses movie streaming services
- Streaming Music: Whether the customer uses music streaming services
- Unlimited Data: Whether the customer has an unlimited data plan
- Contract: Type of contract the customer has
- Paperless Billing: Whether the customer uses paperless billing
- Payment Method: Customer's preferred payment method
- Monthly Charge: Monthly amount charged to the customer

- Total Charges: Total amount charged to the customer to date
- Total Refunds: Total amount refunded to the customer
- Total Extra Data Charges: Total charges for extra data usage
- Total Long-distance Charges: Total charges for long-distance calls
- Total Revenue: Total revenue generated from the customer
- Customer Status: Current status of the customer (Churned, Joined, stayed)
- Churn Category: Category of churn if the customer has churned
- Churn Reason: Specific reason for churn if the customer has churned

This dataset provides a comprehensive view of customers in a telecom company, covering various aspects

- *Demographics*: Includes personal information like **age, gender, marital status, and location**
- *Service Usage*: Details on the types of services used (**phone, internet, streaming**) and the extent of usage (**data consumption, long-distance charges**)
- *Customer Value*: Information on tenure, referrals, and revenue generated
- *Product Details*: Specifics about the **customer's plan, including contract type, add-on services, and billing preferences**
- *Financial Information*: **Various charges, refunds, and total revenue** from each customer
- *Churn Information*: Whether a customer has churned, the category of churn, and the specific reason

Data Preprocessing

- **Average Monthly Long-Distance Charges**: 682 missing values
- **Multiple Lines**: 682 missing values
- **Internet Type**: 1526 missing values
- **Average Monthly GB Download**: 1526 missing values
- **Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data**: 1526 missing values in each column
- **Churn Category and Churn Reason**: 5174 missing values

Data Importation Using MySQL Workbench

Database Creation:

  - Opened MySQL Workbench and connected to a local server host
  - Created a new database using the syntax:

  CREATE DATABASE churn;

Data Importation:

- After creating the database, I right-clicked on the table and selected **Table Data Import Wizard**
- Browsed to the data file path and followed the steps:
  - Clicked **Next** to proceed
  - Clicked **Finish** to complete the import process
- Refreshed the workspace from the top right corner to confirm that the data was successfully imported into the database

Data Type Verification:

- Checked the data types of each column to ensure they were correctly defined for further analysis

Data Imputation:

- Filled missing values in the dataset using appropriate methods for each column:
  - **Average Monthly Long Distance**: Imputed missing values using the average

- **Multiple Lines**: Imputed missing values using the mode
- **Internet Type**: Imputed missing values using the mode
- **Average Monthly GB Download**: Imputed missing values using the mean
- **Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data**: Each column filled using the mode
- **Churn Category and Churn Reason**: Dropped these columns due to irrelevance or high missing value counts

Rationale for Not Normalizing Features

In the data preprocessing phase, "**normalization and standardization**" of features were not applied due to the following reasons

- **Interpretability of Original Scales**: For certain features, such as price or age, the actual scale holds significant meaning in the context of the analysis. Maintaining the original scales allows for more interpretable results, providing clearer insights into how these features relate to the outcome of interest

Next Steps: Insights and Analysis

- With the data thoroughly pre-processed, the next phase focuses on uncovering valuable insights from the dataset
- By ensuring data quality through **cleaning, imputation, and feature selection**, the dataset is now well-prepared for analysis. The exploration of patterns and trends will follow, providing actionable insights to support business decisions

This concludes the data preprocessing phase, and the insights derived from the dataset will be documented in the subsequent analysis report.