Data Cleaning and Preprocessing Report

1. Data Loading and Merging

Data from three branches (Branch1.csv, Branch2.csv, Branch3.csv) was loaded and combined into a single dataset. The ***read_csv*** function was used to load each file, and the ***bind_rows*** function was used to merge the data into a complete dataset.

2. Handling Missing Values

Missing values in numeric columns were identified and replaced with the median value of the respective column. Median imputation is robust to outliers and prevents missing values from skewing data analysis. The ***mutate*** and ***across*** functions were used to replace missing values in all numeric columns with their respective medians.

3. Removing Duplicate Records

Duplicate rows in the dataset were removed. Duplicate records can bias analysis and lead to incorrect results. The ***distinct*** function was used to ensure that each record in the dataset was unique.

4. Detecting and Capping Outliers

Outliers in numeric columns were capped within the interquartile range (IQR). Outliers can distort statistical analysis and model performance. Capping them ensures the data distribution remains representative.

The first quartile (Q1) and third quartile (Q3) were calculated for each numeric column. The IQR (Q3 - Q1) was used to define the acceptable range: [Q1 - 1.5*IQR, Q3 + 1.5*IQR]. Values outside this range were capped to the nearest bound.

5. Assigning Branch Names

Each record was assigned a ***Branch*** identifier (Branch1, Branch2, or Branch3) based on its original source file. Adding branch information allows for branch-level analysis and comparison. The row indices and the respective number of rows in each branch dataset were used to determine the branch label for each record.

6. Ensuring Binary Values for the Target Column

The ***Left*** column, which indicates whether a customer churned, was validated and corrected to ensure it only contains binary values (0 or 1). Binary values are essential for classification tasks and model consistency. Values greater than 1 were capped to 1, and values less than 0 were set to 0.

7. Summary Statistics by Branch

Summary statistics (mean and standard deviation) were calculated for numeric columns, grouped by branch. These statistics provide insights into data distribution and help identify any significant differences between branches. The ***group_by*** and ***summarise*** functions were used to calculate the mean and standard deviation for each numeric column, excluding ***Customer_ID***. For ***Customer_ID***, separate mean and standard deviation statistics were calculated and merged into the results.

8. Saving the Cleaned Data

The cleaned and preprocessed dataset was saved as a new CSV file (Cleaned_Data.csv). The ***write_csv*** function was used to save the data in CSV format.

Results

The cleaned dataset is now ready for analysis, with missing values addressed, duplicates removed, outliers capped, and branch-level identifiers added for deeper insights.