Customer Churn EDA Report

1. Data Sets Selected and Rationale

Data Sources:

- Transaction_History: Contains customer purchase transactions, including amount spent and product category.

Rationale: Transaction patterns and spending behavior are strong indicators of customer engagement and potential churn.

- Customer_Service: Records of customer service interactions, including type and resolution status.

Rationale: Frequent or unresolved complaints may signal dissatisfaction and higher churn risk.

- Online_Activity: Tracks customer login frequency, last login date, and service usage type.

Rationale: Reduced or changing online activity can precede churn.

- Churn_Status: Binary indicator of whether a customer has churned.

Rationale: This is the target variable for prediction.Integration:All datasets were merged on `CustomerID` to create a unified view of each customer's behavior and status.

2. Exploratory Data Analysis (EDA)

2.1 Statistical Summaries

- Numerical Features:
  - TotalSpent, AvgSpent, NumTransactions, NumProductCategories, NumInteractions, NumInteractionTypes, NumUnresolved, AvgLoginFrequency, NumServiceUsageTypes
  - *Summary statistics (mean, median, std, min, max) were computed for each feature.*
- Categorical Features:
  - ProductCategory, InteractionType, ResolutionStatus, ServiceUsage
  - *Frequency counts and unique value analysis were performed.*

2.2 Visualizations

- Distribution Plots:
  - Histograms of key numerical features, colored by churn status.
- Correlation Heatmap:
  - Shows relationships between numerical features.
- Bar Plots:
  - Counts of categorical features by churn status.

Example visualizations (saved in the `plots/` directory):

- `plots/distribution_TotalSpent.png`
- `plots/correlation_matrix.png`
- `plots/categorical_ResolutionStatus.png`

3. Data Cleaning and Preprocessing

3.1 Handling Missing Values

- Strategy:
  - Numerical columns: Imputed with median values.
  - Categorical columns: Imputed with mode (most frequent value).
- Rationale:
  - Median and mode imputation preserves the distribution and avoids bias from outliers or rare categories.

3.2 Outlier Detection and Treatment

- Approach:
  - Outliers were identified using box plots and statistical thresholds.
  - Extreme outliers were capped or transformed as appropriate.

3.3 Feature Engineering

- Aggregations:
  - Transaction, service, and online activity data were aggregated per customer to create summary features.
- Encoding:
  - Categorical variables were one-hot encoded where necessary.
- Scaling:
  - Numerical features were standardized using `StandardScaler` to ensure consistent scale for modeling.

---

4. Cleaned and Preprocessed Data Set

- The final dataset (`processed_customer_churn.csv`) includes:
  - All engineered features
  - No missing values
  - All features numeric and ready for machine learning algorithms