Customer Churn Prediction – Data Preparation Report

# 1. Introduction

The objective of this task is to prepare a reliable dataset for developing a predictive model of customer churn. We performed three major steps:  
1. Data Gathering & Integration – collected and merged relevant datasets.  
2. Exploratory Data Analysis (EDA) – understood distributions, relationships, and potential issues.  
3. Data Cleaning & Preprocessing – handled missing values, outliers, scaling, and encoding.  
  
This ensures the final dataset is clean, consistent, and suitable for machine learning.

# 2. Data Gathering

We worked with the file Customer\_Churn\_Data\_Large.xlsx, which contains multiple sheets:  
  
• Churn\_Status – CustomerID, ChurnStatus (target).  
• Customer\_Demographics – Age, Gender, MaritalStatus, IncomeLevel.  
• Transaction\_History – Transactions, aggregated into spend-related features.  
• Customer\_Service – Customer support interactions, aggregated to per-customer metrics.  
• Online\_Activity – Online usage and login recency.  
  
Final Master Dataset: 1000 rows × 17 features.

# 3. Exploratory Data Analysis (EDA)

EDA was performed on customer\_churn\_master.csv. Outputs were saved in:  
- Plots → eda\_plots/  
- Summaries → eda\_summaries/  
  
Key Insights:  
- Churn Distribution: ~20% churn rate.  
- Demographics: Young (<25) and older (>60) customers churn more.  
- Service: High unresolved complaints linked to churn.  
- Online: Longer inactivity correlated with churn.  
- Transactions: Lower spend & fewer categories linked to churn.  
- Correlation: Spend-related variables highly correlated.

# 4. Data Cleaning & Preprocessing

Performed in clean\_preprocess.py. Output: customer\_churn\_cleaned.csv.  
  
Steps:  
1. Missing Values – Median imputation for numerics, mode/unknown for categoricals.  
2. Outliers – Winsorization using IQR.  
3. Scaling – Standardization (z-score).  
4. Encoding – One-hot encoding for small categories, frequency encoding otherwise.  
  
Final Cleaned Dataset: (1000, ~40 features).

# 5. Deliverables

- customer\_churn\_master.csv → raw integrated dataset.  
- eda\_plots/ → visualizations.  
- eda\_summaries/ → statistical summaries.  
- customer\_churn\_cleaned.csv → cleaned dataset ready for modeling.

# 6. Next Steps

1. Model Training & Evaluation (Logistic Regression, Random Forest).  
2. Feature Importance (SHAP, permutation importance).  
3. Handle class imbalance (class weights, SMOTE).