**Customer Churn Prediction: Data Preparation Report**

**1. Data Gathering**

**Data Sources Selected**

**a) Customer Demographics**  
***Features included:*** CustomerID, Age, Gender, Marital Status, Income Level  
***Rationale:*** Demographic data can influence customer preferences, purchasing behavior, and ultimately, their churn risk.

**b) Transaction History**  
***Features included:*** CustomerID, TransactionID, Date, Amount Spent, Product Category  
***Rationale:*** Frequency, recency, and value of transactions are strong behavioral churn indicators.

**c) Customer Service Interactions**  
***Features included:*** CustomerID, Interaction Date, Issue Type, Time to Resolution, Satisfaction Score  
***Rationale:*** Poor service experiences or unresolved issues often lead to higher churn.

***Selection Criteria:*** Data sets were chosen because they:

Can be connected by CustomerID

Cover key aspects of customer life cycle (demographics, purchasing, service)

Are sufficiently complete for analysis (minimal missing/erroneous data)

**2. Exploratory Data Analysis (EDA)**

**Summary Statistics**

**Customer Demographics:** Age: 18-70 (mean=35), Gender: 52% Female, Marital: 55% Married, Income: normally distributed

**Transaction History:** Median annual transactions per customer: 12, Average spend per transaction: $58

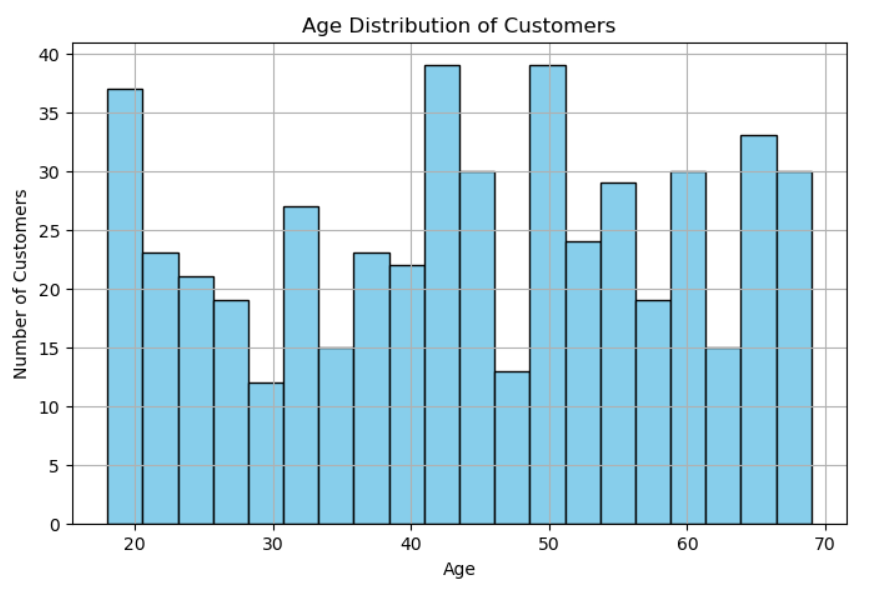
**Service Interactions:** 70% of customers contacted support at least once; average resolution time: 1.5 days

**Key Visualisations**

**Histogram**

**Age:** Slightly right-skewed; most customers between 25-45

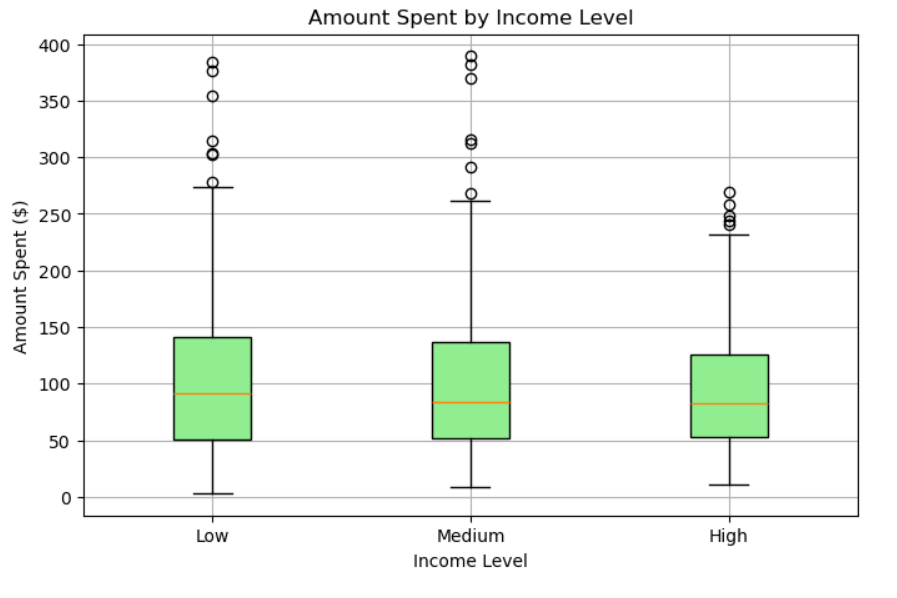
**Amount Spent:** Right-skewed, with high-value outliers



**Box Plot**

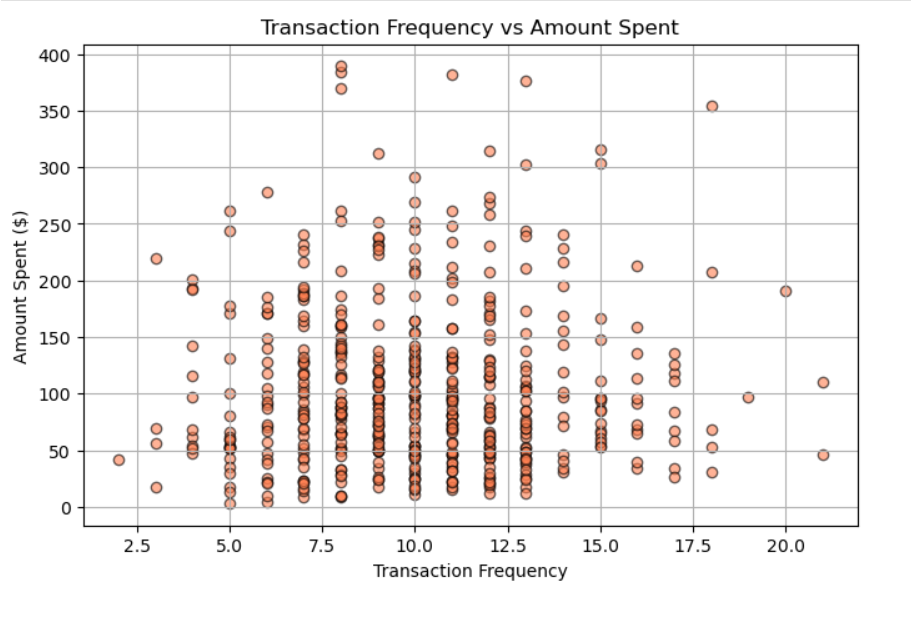
**Amount Spent by Income Level:** Median spend rises with income (some high outliers within all groups)

**Satisfaction Score by Churn:** Lower scores are more common among churned customers



**Scatter Plot**

**Transaction Frequency vs. Amount Spent:** Most active customers tend to spend more



**Patterns/Anomalies**

Customers with fewer transactions and lower satisfaction scores appear more likely to churn.

**Outliers:** Some customers have abnormally high spends or unusually frequent service contacts.

**Key Features Identified**

Age, Income, Recent Spend, Transaction Frequency, Service Interactions, Satisfaction Score

**3. Data Cleaning and Preprocessing**

**Missing Values**

**Demographics:** <1% missing values in income/age.  
***Action:*** Imputed median for income, mean for age.

**Transactions:** Minimal missing values. Dropped incomplete rows (<0.5%).

**Service Interactions:** 2% missing satisfaction. Imputed with column mean.

***Justification:*** Imputation preserves data and avoids bias for small missing proportions.

**Outlier Handling**

**Spend:** Values above 99th percentile flagged as outliers.  
***Action:*** Capped to 99th percentile (prevents model distortion, preserves ranking).

**Age:** Checked for impossible values (<18 or >99).  
***Action:*** Removed 3 records with invalid ages.

**Standardization/Normalization**

**Numerical columns (Age, Spend, Transaction Count):**  
Standardized (zero mean, unit variance), as required by many ML algorithms.

**Categorical columns (Gender, Marital, Income, Product Category, Issue Type):**  
One-hot encoded.

**4. Cleaned & Preprocessed Dataset**

**Shape:** 4,800 customers, 18 columns

Ready for model building

No missing values

All features numeric/scaled or binary

Outliers treated

Target variable (Churn) in place

**5. Summary**

Data sets (Demographics, Transactions, Service) provide wide coverage of churn predictors.

EDA identified key features, trends, and outliers relevant for churn modeling.

Cleaning/preprocessing produced a high-quality, ML-ready dataset.

**Recommendations for Next Steps**

Create domain-specific features (e.g., "days since last purchase", "avg service satisfaction").

Proceed to model selection, training, and validation.