**Comprehensive Report on Data Gathering, EDA, and Data Cleaning Processes**

**1. Summary of the Datasets Selected and Rationale for Their Inclusion**

**Datasets Overview**

For this analysis, we have utilized a comprehensive dataset that contains various customer attributes and interaction data. The dataset includes both customer demographic information and transactional/interaction behaviors that are key to understanding customer churn, which is the target variable.

The dataset consists of the following columns:

* **Customer Demographics**: Age, Gender, Marital Status, Income Level
* **Transaction Information**: Transaction ID, Amount Spent
* **Interaction Information**: Interaction ID, Login Frequency, Interaction Types (e.g., Complaint, Feedback, Inquiry)
* **Service Usage**: Usage of Mobile App, Online Banking, Website
* **Churn Status**: Whether the customer has churned (1) or not (0)

**Rationale for Inclusion**

The selected features are highly relevant for building a model to predict customer churn:

* **Demographic features (Age, Gender, Marital Status, Income)**: These are typical predictors in customer behavior modeling. They help us understand how different customer segments behave and whether they are more likely to churn.
* **Transaction features (AmountSpent)**: The amount spent by customers can help us identify whether higher spending correlates with lower churn rates.
* **Interaction features (InteractionType, ResolutionStatus)**: These help capture the quality of customer service interactions and could be indicative of a customer's likelihood to churn.
* **Service Usage**: Customers who use multiple service channels (Mobile App, Online Banking, Website) might have different churn behaviors compared to those who use only one service channel.

These features provide a good mix of demographic, behavioral, and transactional information that will allow us to build a robust model for churn prediction.

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

**2.1 Visualizations**

**2.1.1 Distribution of Age**

A histogram was plotted to show the distribution of customer ages.

* **Insight**: The distribution of ages is relatively balanced, though there are more customers in the 30-60 age range compared to the extremes (20s and 70s).

**2.1.2 Amount Spent by Age Group**

A boxplot was created to visualize the variation of amount spent across different age groups.

* **Insight**: Younger customers tend to spend slightly less, while middle-aged groups (30-50) show a larger spread in spending. Older customers (60-70) also show a considerable amount of variation, suggesting that age might have a slight impact on spending.

**2.1.3 Churn Status by Gender and Marital Status**

A bar plot was used to compare churn rates across different gender and marital status groups.

* **Insight**: Marital status appears to have some relationship with churn, with married customers having lower churn rates compared to divorced or single customers.

**2.1.4 Heatmap of Correlations**

A correlation matrix heatmap was generated to visualize how features correlate with each other, especially with the churn status.

* **Insight**: The correlation between **Age** and **Amount Spent** is weak. **Service usage (Mobile App, Online Banking, Website)** is negatively correlated with each other, suggesting that customers use one service more than others. **Resolution Status** and **Interaction Types** have strong relationships with each other and churn.

**2.2 Statistical Summaries**

**2.2.1 Descriptive Statistics**

* **Age**: The average age of customers is around 45 years with a relatively even spread across all age groups.
* **Amount Spent**: The mean amount spent per transaction is moderate, with some high-spending outliers (indicating a possible skewed distribution).
* **Interaction Frequency**: Customers tend to interact moderately, with some groups showing high interaction frequencies due to complaints or inquiries.

Key metrics like **mean**, **median**, **standard deviation**, and **IQR** (interquartile range) were calculated to understand the central tendency and spread of the data.

**3. Data Cleaning and Preprocessing Steps**

**3.1 Missing Data Handling**

* **Missing Values**: Initially, there were missing values in a few columns. We used **mean imputation** for continuous variables like **AmountSpent**, and **mode imputation** for categorical variables like **Gender**, **Marital Status**.
* After imputation, we verified that no columns had missing data left.

**3.2 Duplicate Entries**

* We identified and removed any duplicate rows based on **TransactionID** and **InteractionID**, ensuring that each entry is unique and relevant.

**3.3 Outliers**

* **AmountSpent** was found to have extreme outliers (particularly in the high spending range). We used a **log transformation** to normalize the distribution and handle the skewness.

**3.4 Encoding Categorical Variables**

* **Gender**: We used **One-Hot Encoding** to represent **Gender** as two separate columns: **Gender\_F** and **Gender\_M**.
* **Marital Status**: We used **One-Hot Encoding** for marital status, creating separate binary columns for **MaritalStatus\_Divorced**, **MaritalStatus\_Married**, **MaritalStatus\_Single**, and **MaritalStatus\_Widowed**.

**3.5 Feature Scaling**

* Since we have features with different units (e.g., **Age** vs **AmountSpent**), we applied **Standard Scaling** to continuous features like **Age**, **AmountSpent**, and **LoginFrequency**. This ensures that all features contribute equally during model training.

**3.6 Feature Selection**

* We used **correlation analysis** to drop highly correlated features. For example, **Gender\_F** and **Gender\_M** were kept as one feature, and we combined similar service usage features to create new ones (e.g., **ServiceUsage\_Overall** representing overall service usage).

**3.7 Final Data Cleaning**

* After handling missing values, duplicates, outliers, and encoding categorical variables, the final cleaned dataset was ready for analysis and modeling. The shape of the final dataset was significantly reduced (after removing redundant features), and it was well-prepared for machine learning tasks.

**4. Cleaned and Preprocessed Dataset Ready for Model Building**

The final preprocessed dataset includes:

* **Customer demographics**: Age, Gender, Marital Status (One-Hot Encoded), Income Level (One-Hot Encoded)
* **Transaction details**: AmountSpent (log-transformed), Transaction ID (removed)
* **Interaction details**: Interaction ID (removed), Interaction Types (One-Hot Encoded)
* **Service usage**: Combined service usage (Mobile App, Online Banking, Website)

**Summary of the Preprocessing Steps:**

* Handling missing data (imputation)
* Removing duplicates
* Encoding categorical variables
* Normalizing continuous features (scaling)
* Log transformation of skewed variables (AmountSpent)
* Feature selection based on correlations

The cleaned dataset is now ready for model building, where we can apply machine learning algorithms to predict customer churn. The target variable (churn status) is ready to be used as the dependent variable for classification tasks.

**5. Conclusion and Next Steps**

This report details the essential data gathering, exploratory data analysis, and cleaning steps necessary to prepare a dataset for churn prediction. The next steps in the project would involve:

* Model selection (e.g., Logistic Regression, Random Forest, XGBoost)
* Model evaluation using metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**
* Fine-tuning the model and making predictions on new customer data.

By following the above steps, the dataset has been transformed into a suitable format for predictive modeling and churn analysis. The report serves as a clear foundation for further model development and performance optimization.