Comprehensive Report: Data Gathering, EDA, and Data Cleaning

1. Data Gathering

Data Sets Selected

The following datasets were selected from the Excel file `Customer_Churn_Data_Large.xlsx`:

1. **Customer_Demographics**: Contains customer profile information such as age, gender, marital status, and income level.
2. **Transaction_History**: Includes purchase records with transaction ID, date, amount spent, and product category.
3. **Customer_Service**: Captures customer support interactions, their type, and resolution status.
4. **Online_Activity**: Tracks login activity, frequency, and service usage (e.g., mobile app, website).
5. **Churn_Status**: Specifies whether a customer has churned (1) or remains active (0).

Rationale for Inclusion

These datasets were chosen because they provide a comprehensive view of customer behavior, including demographic details, purchasing patterns, service interactions, online activity, and churn status. This information is critical for understanding the factors influencing customer churn and building predictive models.

2. Exploratory Data Analysis (EDA)

Visualizations

1. **Churn Distribution**:
   - A bar chart was created to visualize the distribution of churned vs. active customers.
   - Insights: The dataset contains both churned and active customers, allowing for balanced analysis.
2. **Age Distribution by Churn Status**:
   - A histogram was plotted to show the age distribution of customers segmented by

churn status.
    - Insights: Certain age groups show higher churn rates.
3. **Transaction Amount by Churn Status**:
    - A boxplot was used to compare transaction amounts for churned and active customers.
    - Insights: Churned customers tend to spend less on average.
4. **Customer Service Interaction Types by Churn Status**:
    - A count plot was created to analyze the types of customer service interactions segmented by churn status.
    - Insights: Certain interaction types, such as complaints, are more common among churned customers.
5. **Login Frequency vs. Service Usage**:
    - A scatter plot was used to explore the relationship between login frequency and service usage, segmented by churn status.
    - Insights: Active customers tend to have higher login frequencies and service usage.

Statistical Summaries

- **Customer_Demographics**:
    - Age ranges from 18 to 69 years, with an average of 43.27.
    - Gender and marital status distributions were analyzed.
- **Transaction_History**:
    - Average amount spent per transaction is $250.71, with a range of $5.18 to $499.86.
    - The dataset contains 5054 transactions from 1000 unique customers.
- **Online_Activity**:
    - Customers log in on average 25.91 times, with a minimum of 1 and a maximum of 49.

3. Data Cleaning and Preprocessing

Steps Taken

1. **Handling Missing Values**:
    - Missing values were identified in the `InteractionType` and `ResolutionStatus` columns of the `Customer_Service` dataset.
    - Rows with missing values were retained for now, as they may still provide useful information.
    - **Further Action**: Depending on the modeling approach, these missing values may need to be imputed (e.g., with the most frequent value) or addressed during the model building phase.
2. **Standardization of Numerical Features**:

   - Numerical columns (`Age`, `AmountSpent`, `LoginFrequency`) were standardized using `StandardScaler` to ensure consistent scaling.
   - **Rationale**: Standardization was applied to these features to ensure that they contribute equally to distance-based models (e.g., k-nearest neighbors) and gradient-based models (e.g., support vector machines, logistic regression).
3. **One-Hot Encoding of Categorical Variables**:
   - Categorical columns (`Gender`, `MaritalStatus`, `ProductCategory`, `InteractionType`, `ResolutionStatus`) were one-hot encoded to prepare them for machine learning models.
   - **Note**: One-hot encoding creates binary columns for each category, preventing the model from assuming any ordinal relationship between the categories.
4. **Merging Data Sets**:
   - Relevant columns from all datasets were merged into a single dataframe (`df_final`) using `CustomerID` as the key.
   - **Handling Inconsistencies**: It's important to ensure that the `CustomerID` is clean and consistent across all datasets to avoid any data loss or misalignment during the merge.
5. **Feature Engineering**:
   - Created a new dataset (`df_clv`) to calculate Customer Lifetime Value (CLV) by summing the total amount spent by each customer.
   - **Potential Issues**: The current CLV calculation is a simplification. A more accurate CLV would typically consider factors like purchase frequency, average purchase value, and customer lifespan.
   - **Further Exploration**: Explore other potentially relevant features, such as:
     - Recency of last purchase
     - Frequency of transactions
     - Average customer service interaction time
     - Time since the customer's first interaction

4. Cleaned and Preprocessed Data Set

Final Data Set

The cleaned and preprocessed dataset (`df_final`) contains the following features:

- **CustomerID**: Unique identifier for each customer.
- **Gender**: One-hot encoded gender information.
- **MaritalStatus**: One-hot encoded marital status.
- **ProductCategory**: One-hot encoded product categories.

- **InteractionType**: One-hot encoded customer service interaction types.
- **ResolutionStatus**: One-hot encoded resolution statuses.
- **Age**: Standardized age of the customer.
- **AmountSpent**: Standardized transaction amounts.
- **LoginFrequency**: Standardized login frequency.
- **ChurnStatus**: Target variable indicating whether the customer has churned.

Ready for Model Building

The final dataset is now ready for model building, with all features cleaned, standardized, and encoded. This ensures compatibility with machine learning algorithms and facilitates accurate predictions.

5. Conclusion

The data gathering, EDA, and cleaning processes have provided a robust foundation for predictive modeling. The insights gained from EDA will guide feature selection and model development, while the cleaned dataset ensures high-quality input for machine learning algorithms.

Additional Considerations:

- **Data Validation**: Implement data validation checks to ensure data quality and consistency.
- **Feature Selection**: Consider feature selection techniques to reduce dimensionality and improve model performance.
- **Class Imbalance**: If the `ChurnStatus` is imbalanced, techniques like oversampling or undersampling may be necessary.
- **Iterative Process**: Data cleaning and preprocessing is often an iterative process. As you build and evaluate models, you may need to revisit these steps.