Customer Churn Analysis: Data Preparation and Exploratory Insights:

Summary of the Data Sets Selected and Rationale for Inclusion:

- Customer Data: Contains demographic information (e.g., age, gender, marital status, income level) crucial for understanding customer segments and potential risk factors related to churn.
- Transaction Data: Records each transaction's details, including the amount spent and product category. This data reveals customer spending patterns and preferences, which can impact loyalty.
- Service Data: Tracks customer interactions with the service team, categorized by interaction type and resolution status. Helps assess how customer support experiences might influence churn.
- Activity Data: Logs each customer's engagement level (e.g., login frequency, service usage). Indicates the level of customer activity and engagement with the platform, potentially predicting churn.
- Churn Data: A binary indicator of whether the customer has churned or not. This is the target variable for building predictive models.

Visualizations and Statistical Summaries from EDA:

Customer Data:

- Age Distribution: Displayed a fairly even spread, providing insight into age-based customer segmentation.
- Income Level Counts: Categories ('Low', 'Medium', 'High') were relatively balanced, indicating diversity in economic segments.
- Gender and Marital Status: Proportions across gender and marital status revealed slight variations with potential churn implications.

Transaction Data:

- Transaction Counts and Total Amount by Product Category: Groceries had the highest spending, followed closely by Books, providing insights into product category popularity.
- Monthly Trends: Spending peaked in July and December, with potential seasonal effects on spending habits.
- Heatmap by Category and Month: Highlighted consistent spending patterns in each product category across months, with notable spikes in specific months.

Service Data:

- Interaction Type Distribution: Feedback interactions were most frequent, followed by complaints and inquiries, shedding light on customer interaction trends.
- Resolution Status: High proportion of unresolved cases, potentially influencing churn.
- Interaction Type vs. Resolution Status: Complaints were more likely unresolved, indicating an area of concern for customer satisfaction.
- Heatmap by Month and Interaction Type: Monthly distribution revealed peaks in feedback during summer months, with some seasonality in complaints and inquiries.

Activity Data:

- Service Usage Counts: Usage was distributed across Mobile App, Website, and Online Banking, with Online Banking slightly leading.
- Login Frequency by Service: Online Banking had the highest login frequency, reflecting its popularity and engagement level.
- Average Monthly Login Frequency by Service: Heatmap analysis showed peaks in login activity, especially on Online Banking, indicating high engagement through this channel.

Churn Status:

- Churn Rate: 20.4% churn rate, indicating room for improvement in customer retention strategies.
- Churn by Gender and Marital Status: Slightly higher churn rates among single and widowed customers, with potential links to relationship status and loyalty.

Merged Data:

- Amount Spent by Income Level: Similar spending levels across income categories, indicating stable spending patterns regardless of income.
- Service Usage vs. Churn: Higher churn rates in Mobile App users, suggesting a possible target for churn prevention efforts.
- Resolution Status vs. Churn: Higher churn among customers with unresolved issues, underscoring the importance of effective resolution.
- Interaction Type vs. Churn: Complaint-related churn was notably higher, indicating a key area to address for retention.
- Heatmap by Churn, Gender, and Marital Status: Gender and marital status appeared to impact churn likelihood, with males and single or widowed customers more likely to churn.

Data Cleaning and Preprocessing Steps:

- Data Merging: Combined the Customer, Transaction, Service, Activity, and Churn datasets using the 'CustomerID' column, ensuring that each record includes a complete view of customer attributes and behaviors.
- Missing Values Handling: Replaced missing categorical values with 'unknown' to retain all records without imputing potentially misleading values. This choice maintains data integrity, especially in categorical fields where NaN could represent unobserved behavior.
- Data Type Conversion: Ensured consistency in data types for each feature (e.g., converting date columns to datetime format).
- Target Variable Encoding: Churn status is retained as a binary indicator, suitable for classification models.

Preprocessing for Model Building:

- Encoding Categorical Variables: Used one-hot encoding for categorical variables (Gender, MaritalStatus, IncomeLevel, ProductCategory, InteractionType, ResolutionStatus, ServiceUsage) to prepare them for machine learning algorithms.
- Scaling Numerical Features: Applied standard scaling to numerical variables (Age, AmountSpent, LoginFrequency) to ensure they’re on a comparable scale, enhancing model performance.

Final Cleaned and Preprocessed Dataset:

The cleaned dataset is now ready, featuring scaled and encoded variables, with no missing values and consistent data types. This dataset can now be directly used for model training.