**Customer Churn Analysis Report**

**1. Introduction**

The purpose of this project was to explore a customer dataset, clean and preprocess it, and analyze behavioral features to identify key factors related to customer churn. The work focused on data preparation, exploratory data analysis (EDA), and uncovering insights that could inform churn prediction models.  
  
**Data Source**All Datasets were used as variables were used for aggregation and feature engineering, columns that were considered to be dropped had meaning once aggregation was made.

**1. Transactional Data (Purchases, Spending, Categories)**

* **Columns included**: total\_amount\_spent, number\_of\_transactions, no\_electronics, no\_furniture, no\_clothing, no\_groceries, no\_books.
* **Why chosen**:
  + Spending patterns are a direct signal of customer value.
  + High-value customers who suddenly reduce their transactions are strong churn indicators.
  + Breaking down spending into categories provides insights into product preferences and whether changes in product engagement are linked to churn.
* **Value added**: This dataset helped quantify **monetary value and purchasing preferences**, which are key churn predictors.

**2. Customer Lifecycle Data**

* **Columns included**: account\_age, days\_since\_last\_purchase, recency.
* **Why chosen**:
  + The longer a customer has been inactive (high recency), the higher their risk of churn.
  + Account age provides a baseline: new customers churn differently from long-tenured ones.
  + Days since last purchase captures short-term disengagement.
* **Value added**: These features measure **temporal engagement** and allowed the model to capture behavioral shifts over time.

**3. Customer Engagement Data**

* **Columns included**: login\_frequency, days\_since\_last\_login, no\_interactions.
* **Why chosen**:
  + Logins and interactions are strong signals of platform activity.
  + A decline in engagement metrics usually precedes churn.
  + Helps differentiate between customers who stopped buying but still browse, versus those completely inactive.
* **Value added**: Captured **digital engagement intensity**, providing another churn risk dimension.

4**. Feedback & Interaction Data**

* **Columns included**: no\_feedbacks, feedback\_rate.
* **Why chosen**:
  + Feedback frequency indicates satisfaction/dissatisfaction.
  + Customers who provide feedback may be more invested, while those with low or declining feedback rates may be disengaging.
* **Value added**: Brought in **qualitative engagement signals**, complementing the numerical and behavioral metrics.

**5. Engineered Features (Aggregations)**

* **Columns created**: average\_spent, spending\_frequency.
* **Why chosen**:
  + Raw totals (like total amount spent) needed normalization against account age and transaction counts.
  + average\_spent gave a per-transaction view, useful for spotting customers who buy less per order.
  + spending\_frequency normalized number of transactions by account age to fairly compare new and old accounts.
* **Value added**: Created **balanced features** that allowed the model to detect subtle behavioral changes.

**2. Data Cleaning and Preprocessing**

The raw dataset required several steps before analysis:

* **Missing Values:** Checked and handled missing values. Imputation was performed where appropriate, and invalid entries were removed.I had to mostly input values when I joined the dataset along tables
* **Data Type Conversion:** Boolean columns (e.g., True/False) were converted into binary format (1/0). Dates were standardized into datetime format to calculate recency-based features.
* **Scaling:** minmaxscaler was applied to normalize numeric features for modeling.
* **Encoding:** Categorical features were encoded using one-hot encoding.(I only encoded ser

Result: A **cleaned and preprocessed dataset** was created, ready for analysis.

**3. Feature Engineering**

To better capture customer behavior, aggregated and derived features were created:

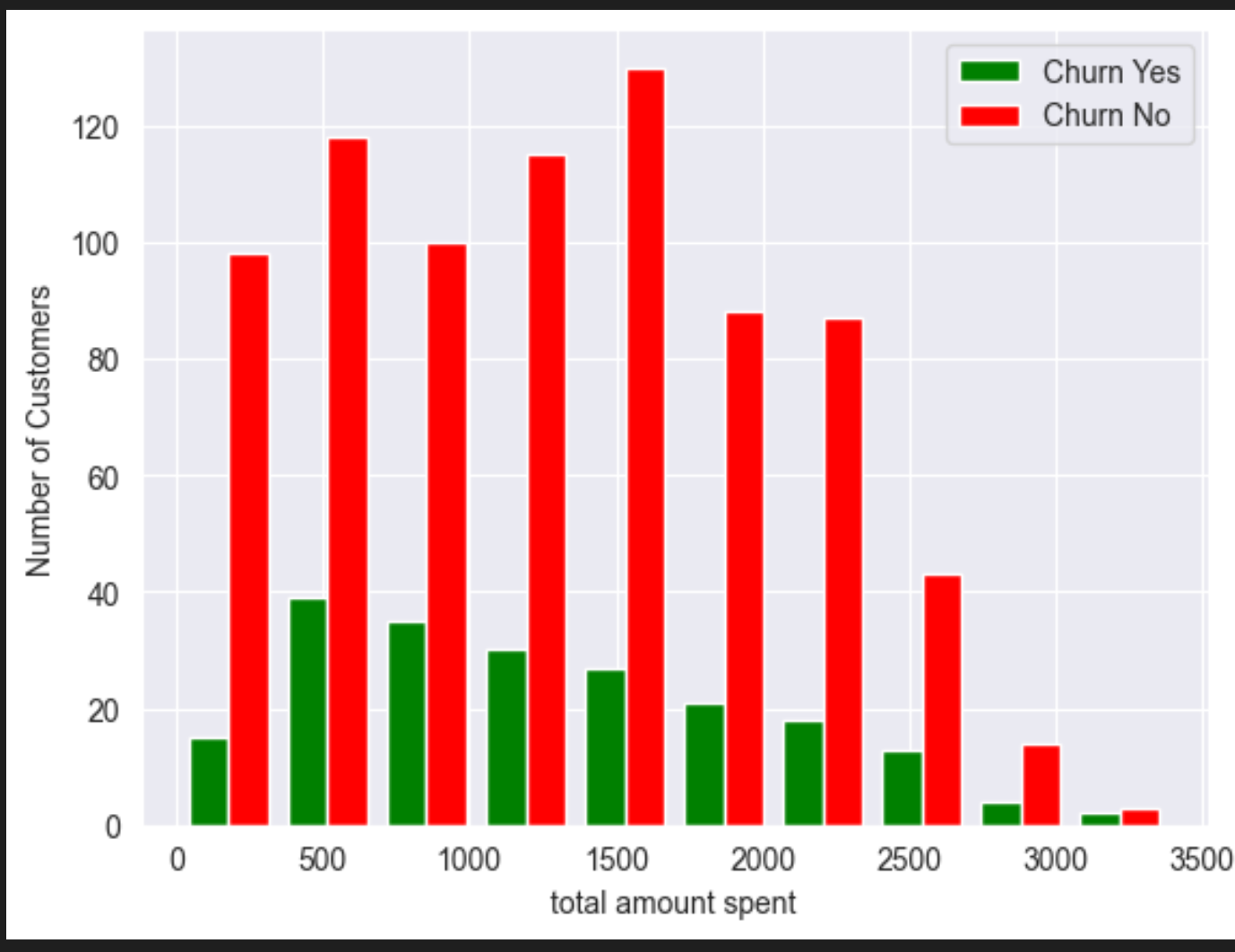
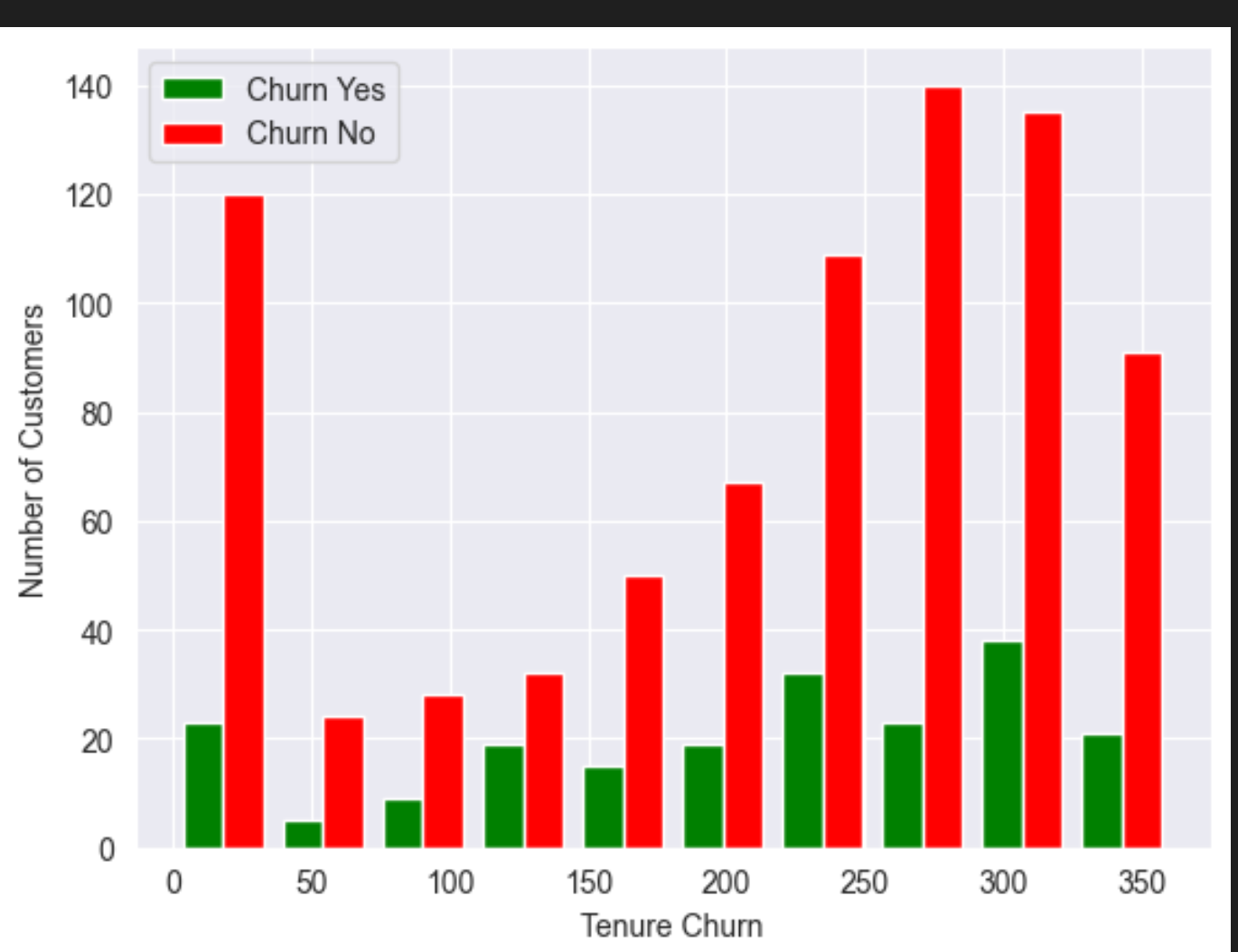
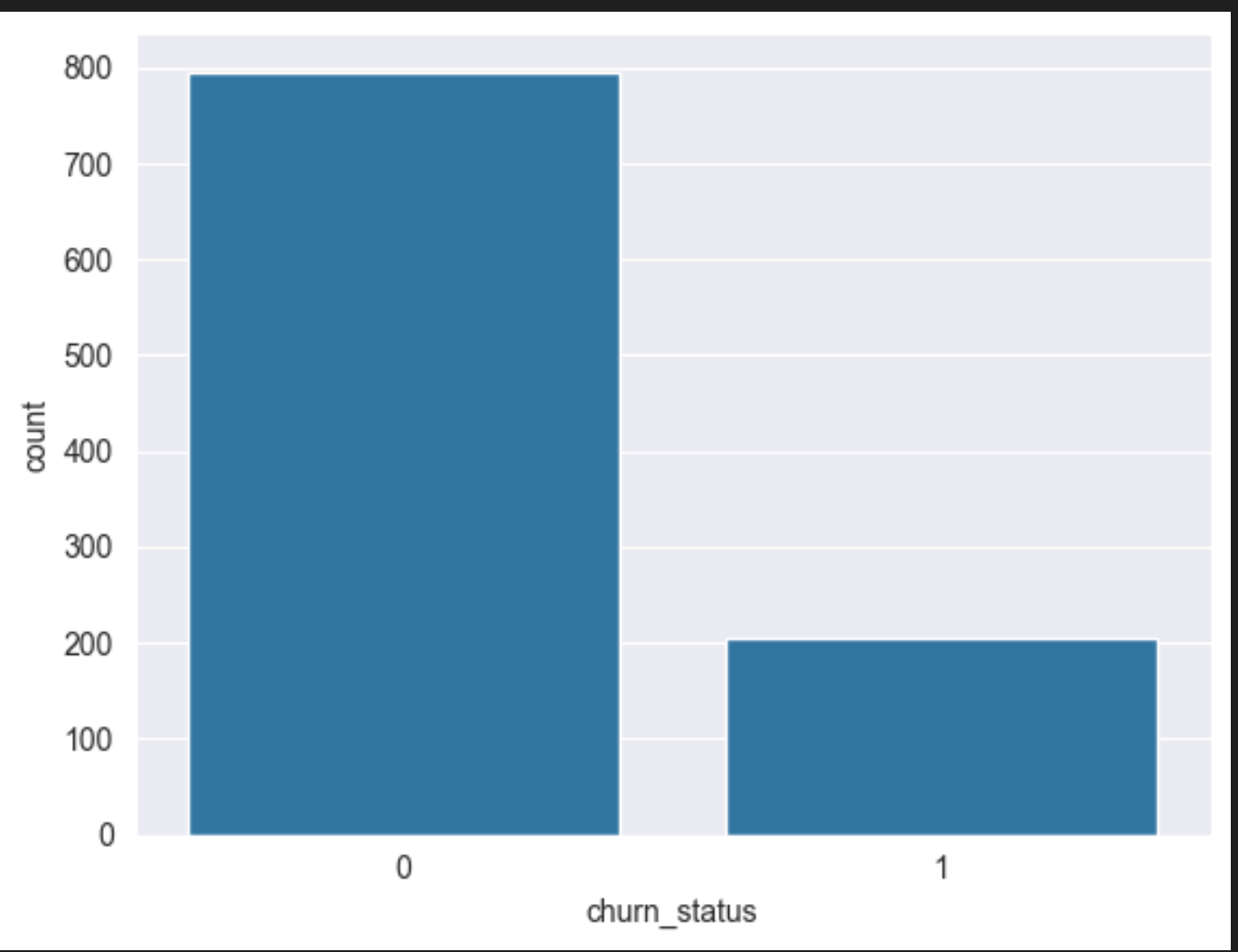
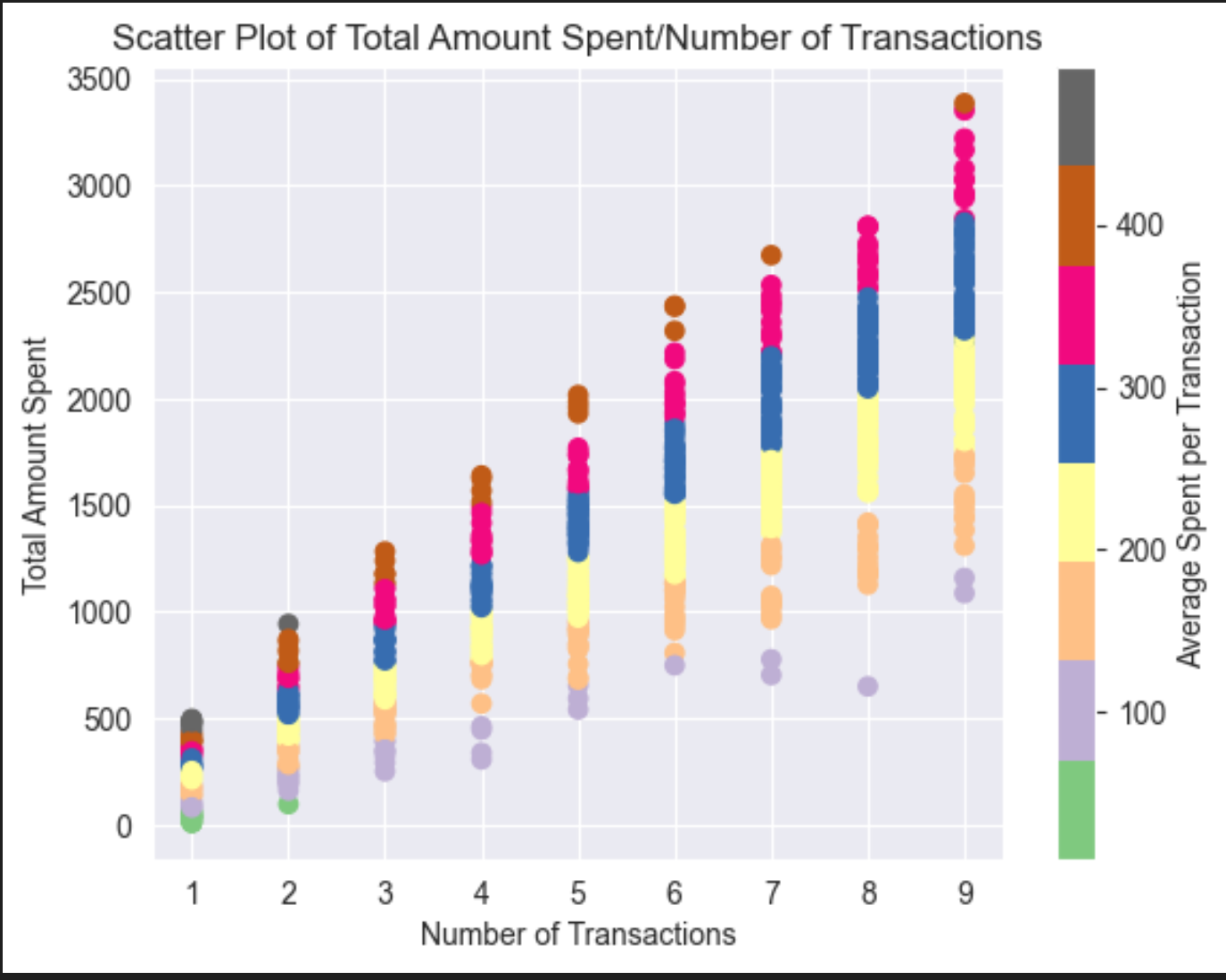
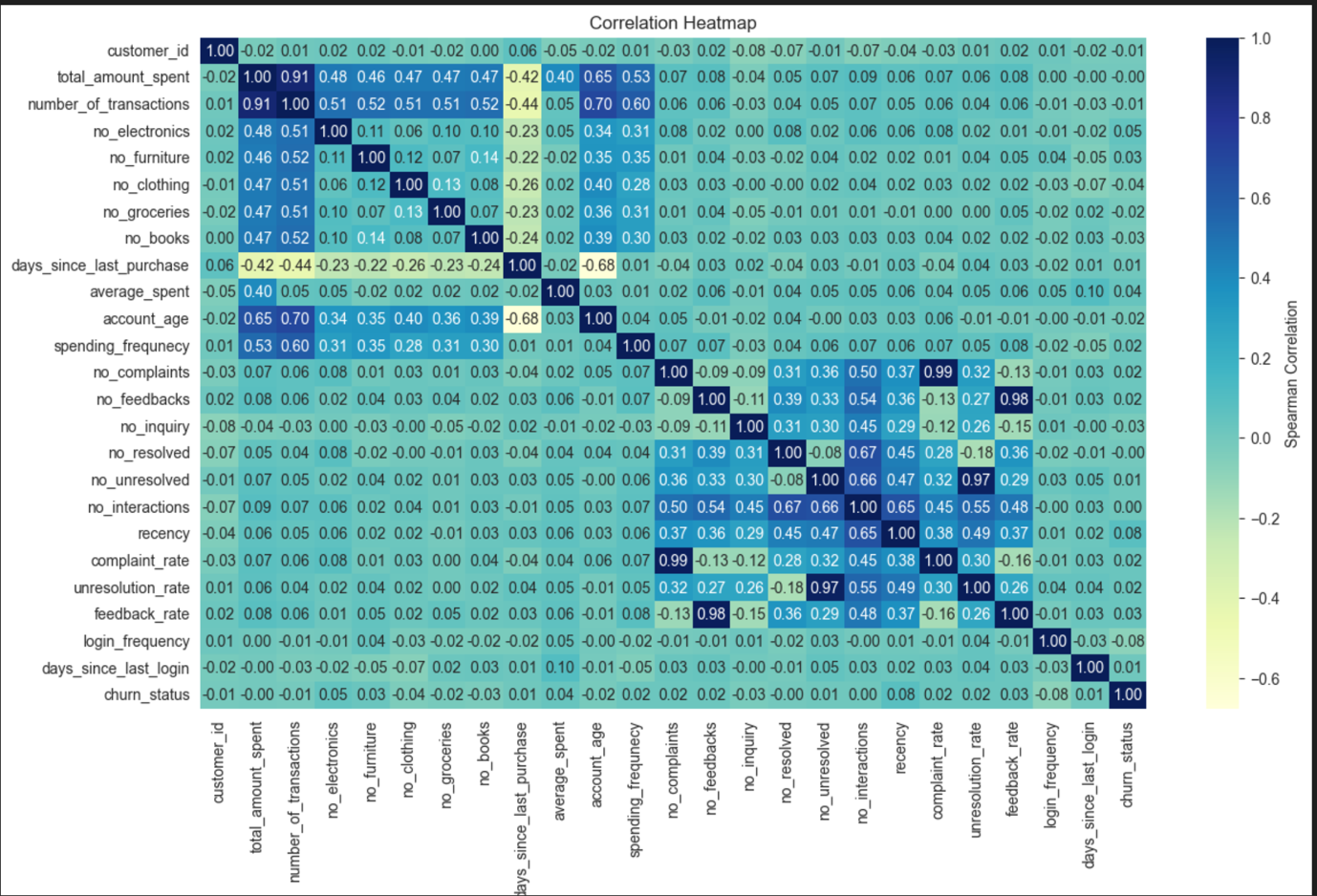
* **Total amount spent** per customer
* **Number of transactions**
* **Average amount spent per transaction**
* **Spending frequency** (transactions / account\_age)
* **Days since last login / last purchase** (recency measures)
* **Complaint rate, feedback rate, unresolved rate**
* **Interaction intensity** (interactions / account\_age)

These features enriched the dataset, allowing deeper behavioral insights into churn.

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

The EDA phase focused on comparing **churners vs. non-churners** across features:

* **Boxplots:**
  + Churners often had higher feedback rates but irregular transaction behavior.
  + Non-churners clustered at moderate transaction and spending ranges.
  + Churners had a higher days since last purchase and last login and vice versa
  + Churners tended to fall most likely with accounts with short life span and total and average spent were smaller comapred to Non Churners
* **Correlation Heatmap (Spearman):**
  + No feature had a strong single correlation with churn (maximum ~0.02).
  + Spending-related features (e.g., total\_spent, avg\_spent, transactions) were strongly correlated with each other, indicating potential multicollinearity.
* **Outlier Analysis:**
  + Very few extreme spenders were detected. These were retained, as they represented realistic high-value customers.



**5. Key Findings**

From the EDA, the following insights emerged:

1. **Customer engagement is critical:** Features like days\_since\_last\_login and days\_since\_last\_purchase distinguish churners more than raw spending totals.
2. **Churners exhibit irregular behavior:** Higher variability in feedback and recency suggests unstable engagement.
3. **Spending patterns alone are not enough:** While spending correlates with account activity, churn is more influenced by **recency and frequency of interactions**.
4. **Class imbalance exists:** Churn cases are much fewer than non-churn, making prediction harder this must be addressed in modeling.

**6. Deliverable Summary**

* A **cleaned and preprocessed dataset** was produced with:
  + Standardized numeric features
  + Encoded categorical variables
  + Outliers checked and retained where valid
  + Engineered behavioral features
* EDA revealed:
  + Recency and engagement features are stronger churn indicators than raw spend.
  + Spending-related features are highly correlated with each other but weakly correlated with churn.
  + Churners show distinct behavior in login frequency and feedback activity.