MAPÚA MALAYAN
DIGITAL COLLEGE

A college under Mapúa Malayan Colleges Laguna

Project Data Preparation and Exploratory Analysis

Prepared and Presented by:

Winston Ace Lao
Sherilou Lopez
Rhona Lynne Bansas

Bachelor of Science in Information Technology
2nd Term, S.Y 2024-2025

TABLE OF CONTENTS

Section

1. Introduction

1.1. Project Overview

The goal of this project is to explore and analyze the preprocessed datasets of **FinMark Corporation** using **Exploratory Data Analysis (EDA)** and **clustering techniques**. Through EDA, we will look for patterns, relationships, and trends in the data to help create personalized financial products. The main focus, however, is to **segment FinMark's customers** into different groups based on their financial behaviors, preferences, and demographics using clustering methods.

Customer segmentation is important for **FinMark Corporation** because it helps the company better understand the specific needs and behaviors of different customer groups. By identifying these groups, FinMark can develop financial products and services that are more relevant to each segment, leading to higher customer satisfaction. Segmentation also improves marketing strategies by allowing FinMark to target the right offers to the right customers, making campaigns more effective and building customer loyalty. In the end, clustering helps FinMark make informed decisions that improve customer engagement and support business growth.

1.2. Dataset Overview

DataSet	Key Data	Number of Records	Important Features
Customer Feedback Data	Customer_ID, Satisfaction_Score, Feedback_Comments, Likelihood_to_Recommend	5050	Satisfaction_Score, Likelihood_to_Recommend
Product Offering Data	Product_ID, Product_Name, Product_Type, Risk_Level, Target_Age_Group, Target_Income_Group	15	Product_Type, Risk_Level, Target_Age_Group (where available)
Transaction Data	Transaction_ID, Customer_ID, Transaction_Date, Transaction_Amount, Transaction_Type	5050	Transaction_Amount, Transaction_Type, Transaction_Date

Additional Insights:

- **Customer Feedback Data**:
  - **Key**: Satisfaction_Score is crucial for sentiment analysis.
  - **Insight**: Likelihood_to_Recommend reveals customer loyalty.
- **Product Offering Data**:
  - **Key**: Product_Type and Risk_Level help segment customers by product suitability.
  - **Limitation**: Missing Target_Age_Group limits demographic insights.
- **Transaction Data**:
  - **Key**: Transaction_Amount identifies high-value customers.
  - **Insight**: Transaction_Type and Transaction_Date show spending patterns and trends.

2. Data Cleaning

2.1.Handling Missing Data

Features with Missing Values:

- Satisfaction_Score in dataset1 had 101 missing values. This feature is important for understanding customer satisfaction, and missing values here could impact analyses related to customer experience.
- Target_Age_Group in dataset2 had 15 missing values. Age grouping is essential for segmentation analysis, and missing data could affect age-based trends and insights.
- Transaction_Amount in dataset3 had 100 missing values. This column plays a key role in analyzing transaction behavior, and missing values could distort trends in customer spending.

How Missing Values Were Handled:

Satisfaction_Score (Numerical):

- Missing values were filled with the median of the column. The median is chosen because it is not affected by outliers, making it a good choice for numerical data.

Target_Age_Group (Categorical):

- Rows with missing values were dropped. For categorical data, it’s better to drop missing values than to guess the missing category.

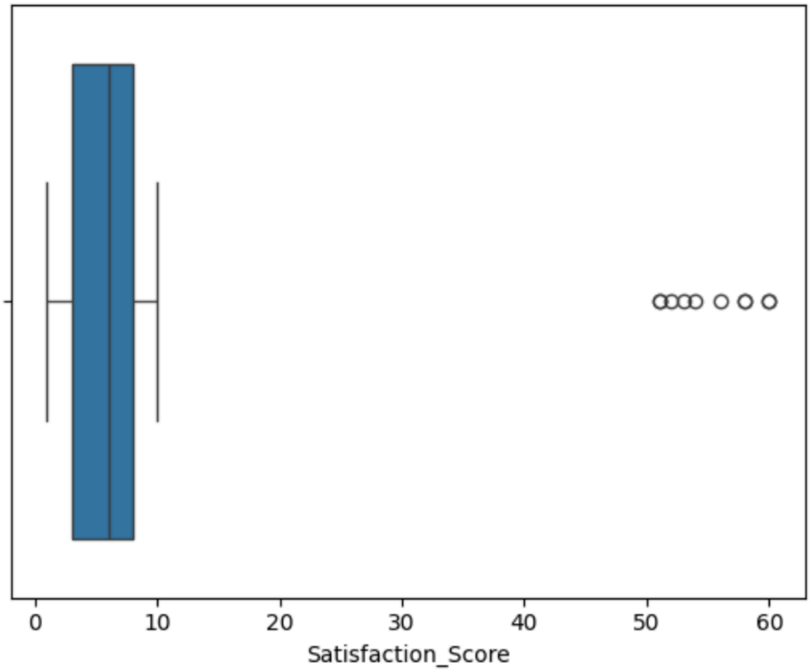
Transaction_Amount (Numerical):

- Missing values were filled with the median, just like Satisfaction_Score, to avoid the influence of outliers.

Why These Methods Were Chosen:

- For **numerical features**, we used the **median** to fill missing values because it's less affected by extreme values.
- For **categorical features**, we dropped rows with missing values to avoid making incorrect assumptions.

2.2. Outlier Detection and Treatment



The outlier analysis for the **Satisfaction_Score** column shows that there are **no outliers**. Here's a summary:

- **Lower Bound**: -4.5
- **Upper Bound**: 15.5
- **Outliers Below**: 0
- **Outliers Above**: 0
- **Total Outliers**: 0

This confirms that all values are within the acceptable range after applying the **IQR method,** indicating no outliers in the dataset.

2.3. Scaling and Normalization

Features Requiring Scaling or Normalization

- **Satisfaction_Score**
- **Transaction_Amount**

These features had different units and ranges, which could create issues during analysis or modeling. For example, the **Satisfaction_Score** may range from 1 to 10, while **Transaction_Amount** may range from a few dollars to thousands. When features have different scales, it can affect machine learning algorithms and statistical analysis, especially for methods that rely on distances (like clustering or regression).

Satisfaction_Score and **Transaction_Amount** have different value ranges, so scaling them ensures that neither dominates the other during analysis.

Scaling Method Used:

The method used for scaling was **StandardScaler**. This technique performs **standardization**, which involves:

- **Removing the mean** of the feature (shifting the distribution to center around zero)
- **Scaling to unit variance** (dividing by the standard deviation)

After applying **StandardScaler**, the features are transformed so that they have:

- A **mean of 0**
- A **standard deviation of 1**

This transformation makes the data more comparable across features, and it ensures that they are on a consistent scale, making them easier to work with in machine learning models.

Here’s how the transformed data looks:

- **Satisfaction_Score_Scaled** and **Transaction_Amount_Scaled** are the scaled versions of the original features, with their values adjusted based on the **StandardScaler**.

StandardScaler was used because it standardizes the data by transforming it to have a mean of 0 and a standard deviation of 1, which is effective when features are measured on different scales and units.

3. Feature Engineering

3.1. New Features Created

High_Satisfaction (in dataset1_cleaned):

- Calculated by scaling the Satisfaction_Score using StandardScaler and creating a binary feature where scores above 0 are labeled as 1 (high satisfaction) and scores below or equal to 0 as 0 (low satisfaction).
- This feature helps segment customers into two groups: those with high satisfaction and those without, enabling targeted strategies to improve customer experience.

Transaction_Year,Transaction_Month,Transaction_DayOfWeek (in dataset3_cleaned):

- Extracted from the Transaction_Date column to provide temporal insights into customer transactions.
- These features allow segmentation based on transaction timing, helping identify patterns such as customers who transact more frequently during specific months or days of the week.

Scaled Features:

- Satisfaction_Score_Scaled and Transaction_Amount_Scaled were created by standardizing the respective columns using StandardScaler.
- These features ensure that numerical data is on the same scale, which is essential for clustering algorithms to treat all features equally.

3.2. Transformation of Categorical Features

Transaction_Type (from dataset3_cleaned):

- **Transformation Method**: One-hot encoding was applied to this feature. Each unique transaction type (e.g., Purchase, Bill Payment, Investment) was converted into a separate binary column. For instance, if there are three transaction types, three new columns were created: Transaction_Type_Purchase, Transaction_Type_Bill_Payment, and Transaction_Type_Investment. Each of these columns contains binary values (0 or 1), indicating whether that transaction type is associated with a particular record.

Why Was It Necessary to Transform These Categorical Features?

- **Clustering algorithms**, such as **KMeans**, rely on numerical data to calculate the distances between data points. Since categorical features are non-numeric, they cannot be directly used in such algorithms.
- **One-hot encoding** ensures that each category is represented as a separate feature, thus preventing any unintended ordinal relationship between categories that could arise if label encoding were used. For example, label encoding might imply an order between categories, which is not appropriate for many categorical features.
- This transformation enables the clustering algorithm to treat each category as an independent feature, allowing for more accurate and meaningful clustering results by properly representing the categorical data.

How Does This Transformation Help Improve Clustering?

- **Numerical Representation**: By converting categorical features into numerical formats through one-hot encoding, clustering algorithms can incorporate these features into their distance calculations. This leads to more meaningful and accurate clusters since the algorithm can now factor in all available features, including categorical ones.

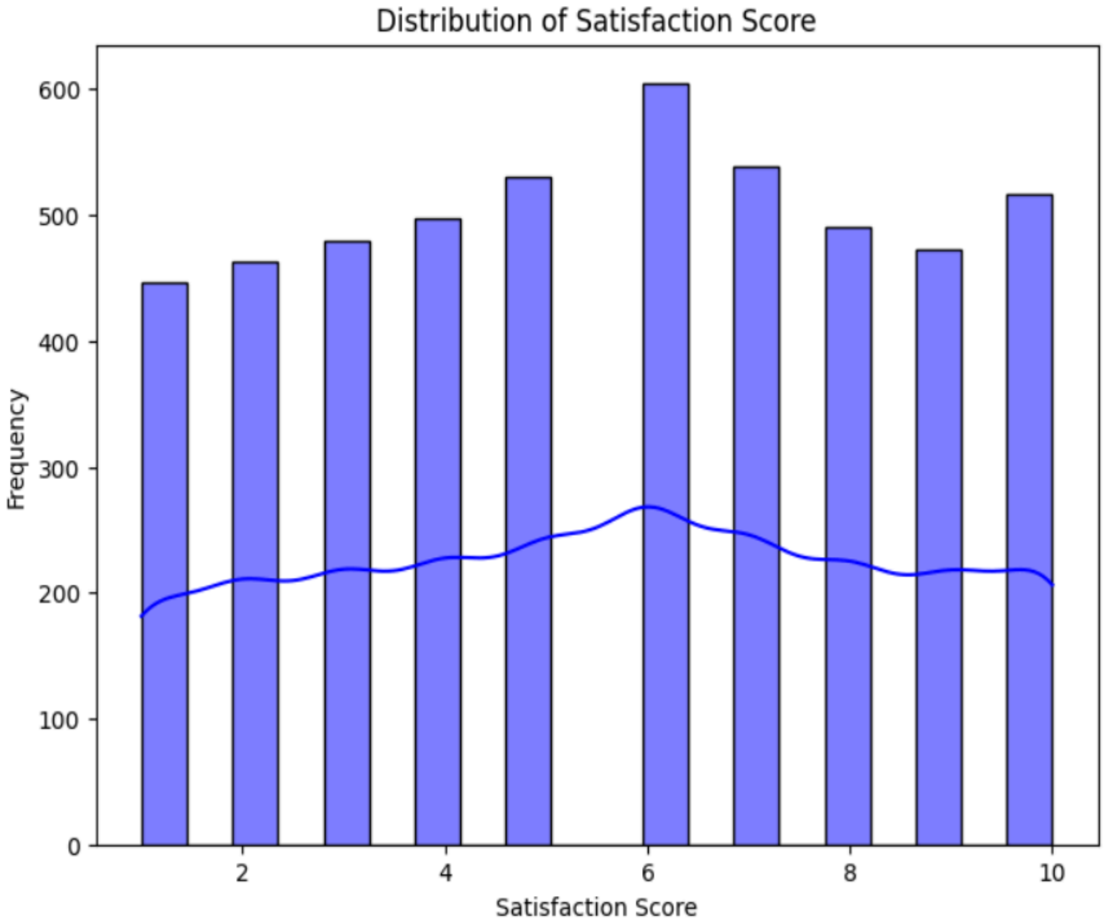
- **Equal Treatment of Categories**: One-hot encoding ensures that each category is treated equally, avoiding any unintended bias or misinterpretation of the data. Without this transformation, some categorical features might be wrongly interpreted as ordinal, which could distort clustering results.
- **Better Segmentation**: Including transformed categorical features allows the clustering algorithm to account for these variables in customer behavior analysis. This results in more actionable and precise customer segments. For instance, clusters may emerge based on transaction preferences, such as identifying customers who primarily make purchases versus those who focus on investments.

4. Exploratory Data Analysis (EDA)

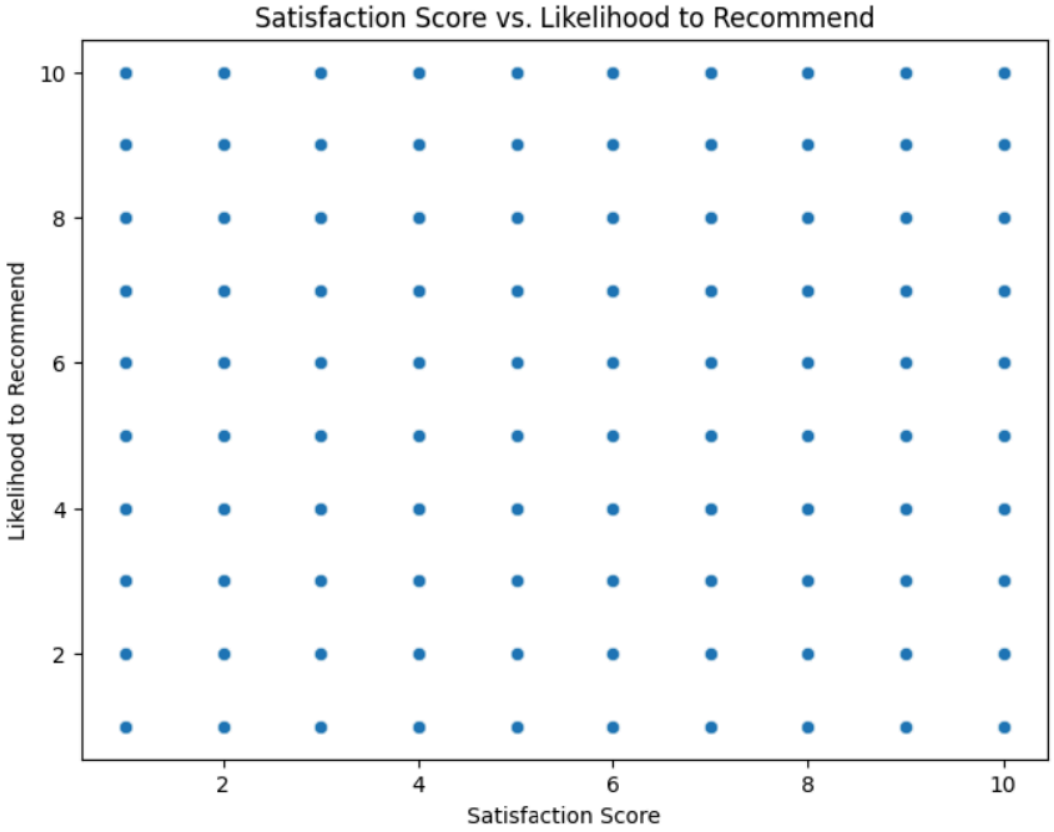
4.1. Overview of the EDA Approach

In the context of **Finmark**, **Exploratory Data Analysis (EDA)** is essential for understanding and preparing financial data for customer segmentation. By identifying key patterns in transaction amounts, types, and frequencies, EDA helps highlight the most important financial features for accurate segmentation. It also plays a critical role in detecting data issues such as **outliers** and **missing values**, ensuring that the data is clean and reliable for further analysis. Through this process, EDA provides insights that guide the transformation of data, making it ready for clustering algorithms. By focusing on relevant financial behaviors, such as high-value transactions or frequent bill payments, EDA enables **Finmark** to create meaningful and actionable customer segments. This leads to more targeted marketing, tailored product offerings, and improved customer service strategies.

4.2. Descriptive Statistics and Visualizations

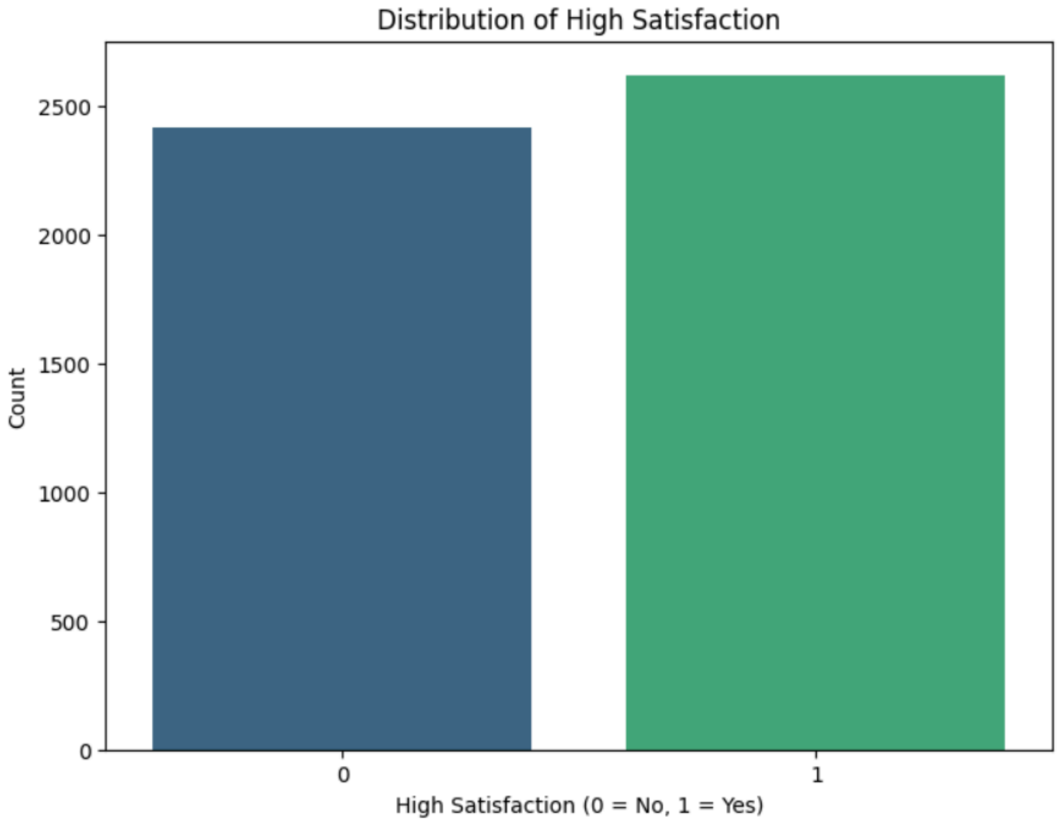


- A histogram with a KDE plot to show the distribution of satisfaction scores.



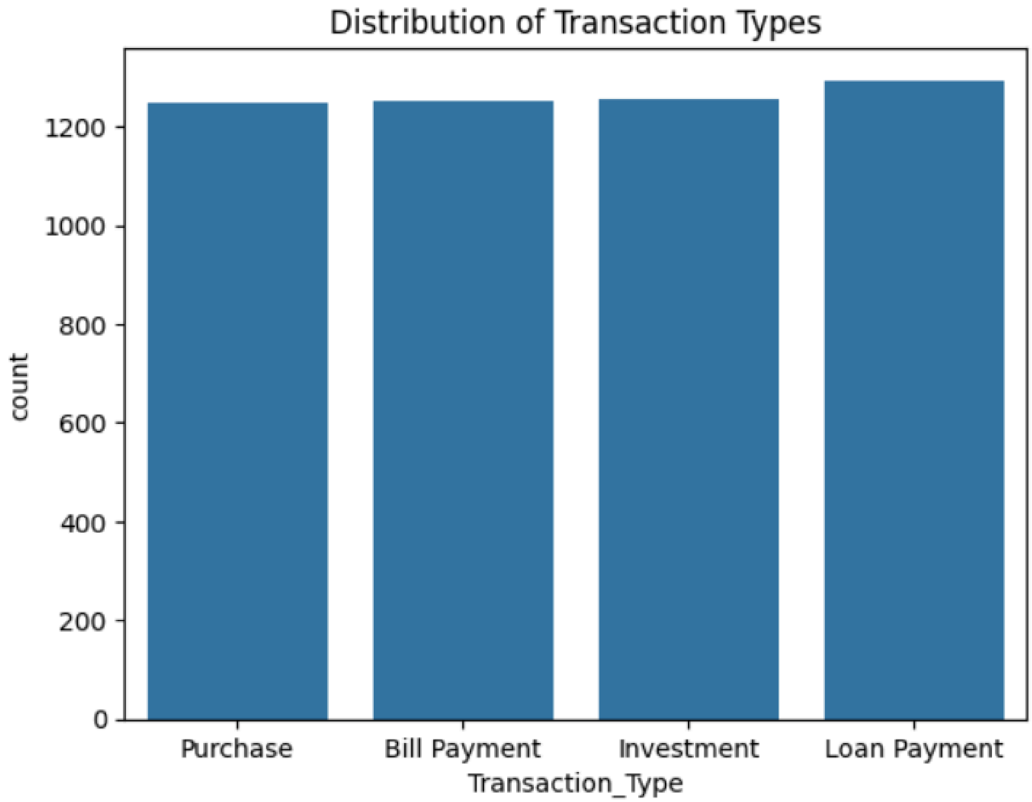
The **Satisfaction Score vs. Likelihood to Recommend (Scatter Plot)** provides the following insights:

- **Purpose**: Visualizes the relationship between customer satisfaction and their likelihood to recommend.
- **Key Insight**: Identifies if higher satisfaction correlates with a higher likelihood of recommending the service/product.
- **Pattern Observation**: The distribution of points may show trends, clusters, or any correlation (positive or negative) between the two variables.



The **Distribution of High Satisfaction (Bar Chart)** provides the following insights:

- **Purpose**: Compares the number of customers with high satisfaction (1) vs. those without (0), based on the scaled satisfaction score.
- **Key Insight**: Shows the proportion of highly satisfied customers, providing a snapshot of overall customer satisfaction.
- **Utility**: This visualization helps to assess whether the customer base is generally satisfied or if improvements are needed to increase high satisfaction rates.



The **Distribution of High Satisfaction (Bar Chart)** provides the following insights:

- **Purpose**: Compares the number of customers with high satisfaction (1) vs. those without (0), based on the scaled satisfaction score.
- **Key Insight**: Shows the proportion of highly satisfied customers, providing a snapshot of overall customer satisfaction.
- **Utility**: This visualization helps to assess whether the customer base is generally satisfied or if improvements are needed to increase high satisfaction rates.

5. Key Patterns and Insights

5.1. Identified Patterns

Key Patterns Observed:

Satisfaction Score Distribution: The satisfaction scores are distributed across a range, with a concentration around certain values, indicating varying levels of customer satisfaction. This suggests that there are different groups of customers, some more satisfied than others, which is crucial for segmentation.

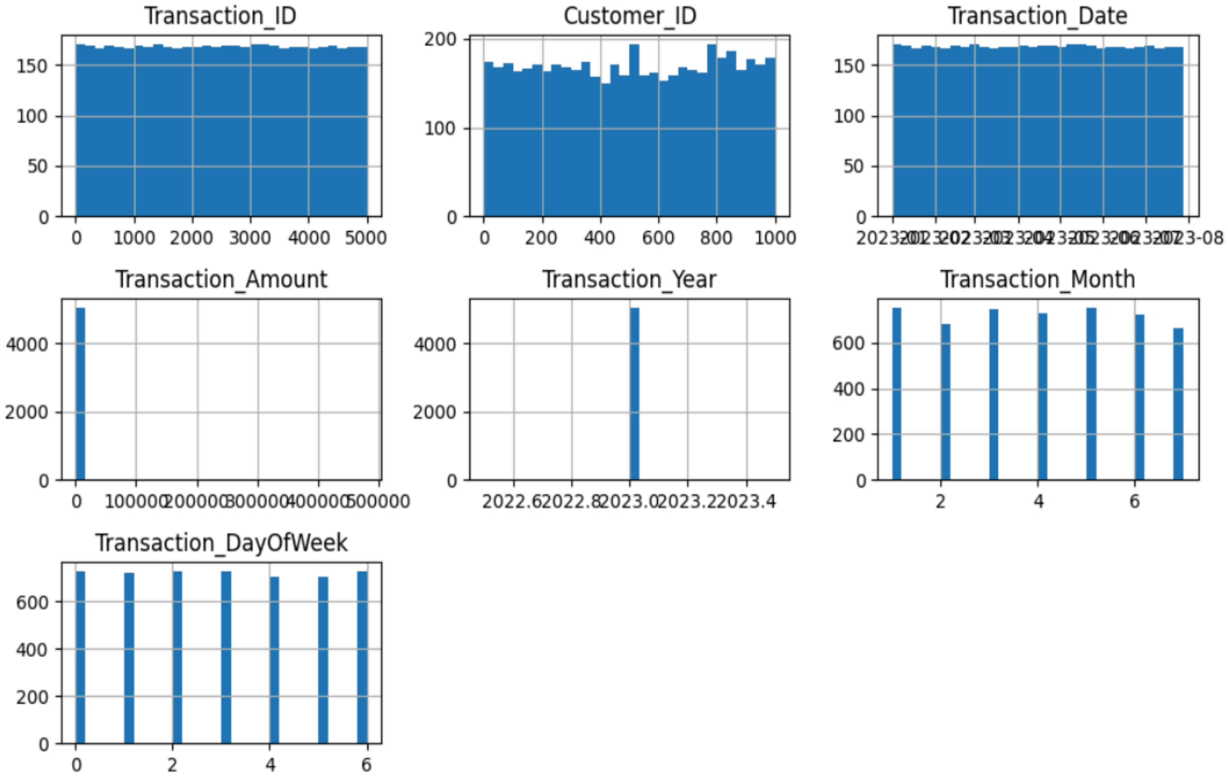
Relationship Between Satisfaction Score and Likelihood to Recommend: There appears to be a positive correlation between satisfaction scores and the likelihood to recommend. Customers with higher satisfaction scores are more likely to recommend the service/product to others. This pattern indicates that satisfied customers are likely to become brand advocates, which is important for identifying loyal customer segments.

High Satisfaction Proportion: A large proportion of customers have high satisfaction scores (scaled scores above 0), indicating a generally positive customer sentiment. This suggests that most customers are satisfied, but it may also indicate a potential group of customers who could be further analyzed for retention strategies or upselling opportunities.

Temporal Patterns in Transactions: Features such as transaction year, month, and day of the week reveal seasonal or weekly trends in customer transactions. For example, some customers may transact more frequently during certain months or specific days of the week, which could inform marketing strategies or product promotions.

Scaled Features: Scaling satisfaction scores and transaction amounts ensures that these features are standardized, making them comparable for clustering or predictive modeling. This standardization removes the bias that may result from features with different magnitudes and helps algorithms treat all features equally, improving the quality of the analysis.

Summary Statistics and Distributions



- The summary statistics for `dataset3_cleaned` show the central tendency, spread, and range of numerical columns. There are no missing values in the dataset after transformations. The histograms visualize the distribution of numerical columns, providing insights into their spread and patterns.

5.2. Feature Importance

The features that typically show the most variation across customers are:

1. **Satisfaction_Score**: There is often a wide range of satisfaction scores, from highly satisfied to dissatisfied customers. This variation reflects differences in customer experiences and perceptions, making it useful for segmentation based on satisfaction levels.
2. **Transaction_Amount**: The transaction amounts can vary greatly, with some customers spending significantly more than others. This feature helps to identify high-value customers, which is key for segmenting based on purchasing behavior and financial contribution.
3. **Likelihood_to_Recommend**: Similar to the Satisfaction_Score, this feature shows variation based on customer loyalty and advocacy. Customers who are more likely to recommend the service/product are typically more engaged and satisfied.
4. **Transaction_Type**: Different types of transactions (e.g., purchases, bill payments, refunds) can indicate different customer behaviors. Variability in this feature can highlight specific needs or preferences.
5. **Product_Type**: Customers may engage with a variety of product types, and differences here can help segment customers based on product preferences and suitability.

These features are crucial for **customer segmentation** because they directly capture key aspects of customer behavior, satisfaction, spending habits, and loyalty. By segmenting customers based on these features, businesses can better tailor their marketing efforts, identify high-value customers, and provide personalized experiences, ultimately improving customer satisfaction and retention.