Clustering Report

Prepared By: Soumita Sahu
Date: 27.01.2025

1. Objective

The goal of this analysis is to perform clustering on customer and transaction data using the K-Means algorithm. The clustering aims to group customers based on their transactional behavior and derive meaningful insights to inform business strategies.

2. Data Overview

2.1 Dataset Description

- **Customers Dataset:** Contains information on customer demographics and regions.
- **Transactions Dataset:** Contains details of customer transactions, such as Transaction ID, Total Value, and Transaction Date.

2.2 Data Merging and Preprocessing

- **Data Merging:** The datasets were merged on CustomerID.
- **Derived Feature:**
  - DaysSinceFirstTransaction was calculated to capture the recency of transactions.
- **Dropped Columns:** Columns unrelated to clustering (e.g., TransactionDate) were removed.
- **Feature Scaling:** Used StandardScaler to standardize numerical features for clustering.

Here's a professional PDF-style clustering report based on the operations and code you've executed. The report covers all the necessary details:

Clustering Report

Prepared By: [Your Name]
Date: [Insert Date]

1. Objective

The goal of this analysis is to perform clustering on customer and transaction data using the K-Means algorithm. The clustering aims to group customers based on their transactional behavior and derive meaningful insights to inform business strategies.

2. Data Overview

2.1 Dataset Description

- **Customers Dataset:** Contains information on customer demographics and regions.

- **Transactions Dataset:** Contains details of customer transactions, such as Transaction ID, Total Value, and Transaction Date.

2.2 Data Merging and Preprocessing

- **Data Merging:** The datasets were merged on CustomerID.
- **Derived Feature:**
  - DaysSinceFirstTransaction was calculated to capture the recency of transactions.
- **Dropped Columns:** Columns unrelated to clustering (e.g., TransactionDate) were removed.
- **Feature Scaling:** Used StandardScaler to standardize numerical features for clustering.

3. Clustering Methodology

3.1 Algorithm

- **Clustering Algorithm:** K-Means
- **Number of Clusters:** 3
- **Features Used:**
  - TotalValue
  - TransactionID
  - DaysSinceFirstTransaction

3.2 Metrics Calculated

1. **Davies-Bouldin Index (DBI):** Measures the quality of clustering. Lower values indicate better clustering.
2. **Silhouette Score:** Measures how similar an object is to its cluster compared to others. Higher values indicate well-separated clusters.

4. Clustering Results

4.1 Key Metrics

Metric	**Value**
Number of Clusters	3
Davies-Bouldin Index	0.8092321837458051
Silhouette Score	0.4178923169126942

4.2 Cluster Analysis Interpretation:

1. **Davies-Bouldin Index (DBI) – 0.8092:**

- A lower DBI value indicates better clustering. The value you've got (0.809) is decent but suggests that there is room for improvement. Ideally, you want the DBI value to be closer to 0 for a good clustering result.

2. **Silhouette Score – 0.4179:**

    - The silhouette score ranges from -1 to +1. A score closer to +1 indicates well-separated clusters, while a score close to 0 indicates that the clusters are overlapping or poorly separated. A score of 0.4179 is reasonable but also indicates that the clusters are not perfectly well-separated.

5. Visualization

5.1 Pairplot of Clusters

- The pairplot below visualizes the clusters based on the selected features (TotalValue, TransactionID, and DaysSinceFirstTransaction).

5.2 Cluster Distribution

- Each cluster's distribution across features is represented in the pairplot, highlighting the separation among clusters.

6. Business Insights

1. **Distinct Customer Segments:**
    - Customers are grouped into three clusters based on transactional patterns.
    - Example: One cluster may represent high-value frequent shoppers, while another represents low-value, infrequent shoppers.
2. **Target Marketing Opportunities:**
    - Tailored marketing campaigns can be designed for high-value clusters to enhance retention and boost revenue.
3. **Operational Efficiency:**
    - Understanding clusters can help allocate resources efficiently (e.g., delivery routes, inventory).
4. **Promotional Campaigns:**
    - Clusters with low transaction values and frequency can be targeted with discounts and promotions to encourage more activity.
5. **Strategic Planning:**
    - Insights from clustering can guide decisions on product offerings, pricing strategies, and customer engagement plans.

7. **Conclusion**

The clustering analysis has produced 3 clusters, which were evaluated using the Davies-Bouldin Index (DBI) and Silhouette Score.

- **Davies-Bouldin Index**: The DBI value of **0.809** suggests that while the clusters are somewhat distinct, there is still room for improvement in terms of cluster separation and cohesion.
- **Silhouette Score**: The score of **0.418** indicates that the clusters are reasonably separated but could benefit from refinement.

While these metrics suggest that the clustering is not perfect, they provide an initial segmentation of the customer base, which can still be valuable for targeted marketing and decision-making. Further analysis, including testing different numbers of clusters and potentially using different clustering algorithms, could help improve the results.