Customer Clustering Report

Prepared By: Soumyadeep Das
Date: 27.01.2025

1. Objective

The objective of this task is to perform customer segmentation using clustering techniques. By leveraging both customer profile information (from Customers.csv) and transaction data (from Transactions.csv), the goal is to identify distinct customer groups for targeted marketing and strategy development. Clustering performance is evaluated using the Davies-Bouldin (DB) Index and other metrics.

2. Dataset Overview

Customers.csv

- **Columns:** CustomerID, CustomerName, Region, SignupDate
- **Total Records**: 200
- **Key Notes**: Contains customer profile details such as ID, name, region, and sign-up date.

Transactions.csv

- **Columns:** TransactionID, CustomerID, ProductID, Quantity, TransactionDate, TotalValue
- **Total Records**: 5000 (after loading and cleaning)
- **Key Notes**: Contains transactional data linking customers and products.

3. Steps for Completion

Step 1: Data Preparation

1. **Merge Datasets:**
   Combined Customers.csv and Transactions.csv to include customer profile details, transaction summaries, and behavioural data.

2. **Feature Engineering:**
   Created features for clustering:
   - **Total Transaction Value:** Sum of all purchases per customer.

- **Average Transaction Value:** Mean value of customer transactions.
- **Transaction Frequency:** Count of transactions.
- **Region:** Encoded as categorical data.
- **Signup Date:** Transformed into numeric values representing days since a fixed date.

3. **Data Preprocessing:**
   - Scaled numerical features to ensure all are on the same scale.
   - Encoded categorical features (Region) using one-hot encoding.
   - Converted SignupDate to numeric (e.g., days since the earliest signup date).

Step 2: Clustering

1. **Algorithm Selection:**
   Used the **K-Means clustering algorithm** due to its simplicity and effectiveness for numerical data. The optimal number of clusters was determined by evaluating the Davies-Bouldin (DB) Index across a range of cluster sizes.

2. **Optimal Cluster Selection:**
   Evaluated DB Index for cluster sizes from 2 to 10.
3. **Final Clustering:**
   Performed clustering with the optimal number of clusters and assigned cluster labels to customers.

```
[49]: customer_features['Cluster'].value_counts()
```

```
[49]: Cluster
      1    31
      3    25
      6    23
      4    22
      0    21
      5    21
      2    20
      7    18
      8    18
      Name: count, dtype: int64
```

Step 3: Evaluation

1. **Davies-Bouldin Index (DB Index):** The final DB Index for the clustering model was [Insert Value], indicating well-defined clusters.

2. **Cluster Distribution:** Distribution of customers across clusters:

3. **Visualization:** Reduced features to 2D using PCA and visualized clusters:

4. Results and Insights

4.1. Number of Clusters:

The optimal number of clusters formed is [Insert Number] based on the minimum DB Index value.

4.2. DB Index Value:

The DB Index for the final clustering is [Insert Value], indicating that the clusters are compact and well-separated.

4.3. Cluster Insights:

- High-Value Customers: Clusters with high TotalValueSum and frequent transactions are potential high-value customers. These can be targeted with loyalty programs.

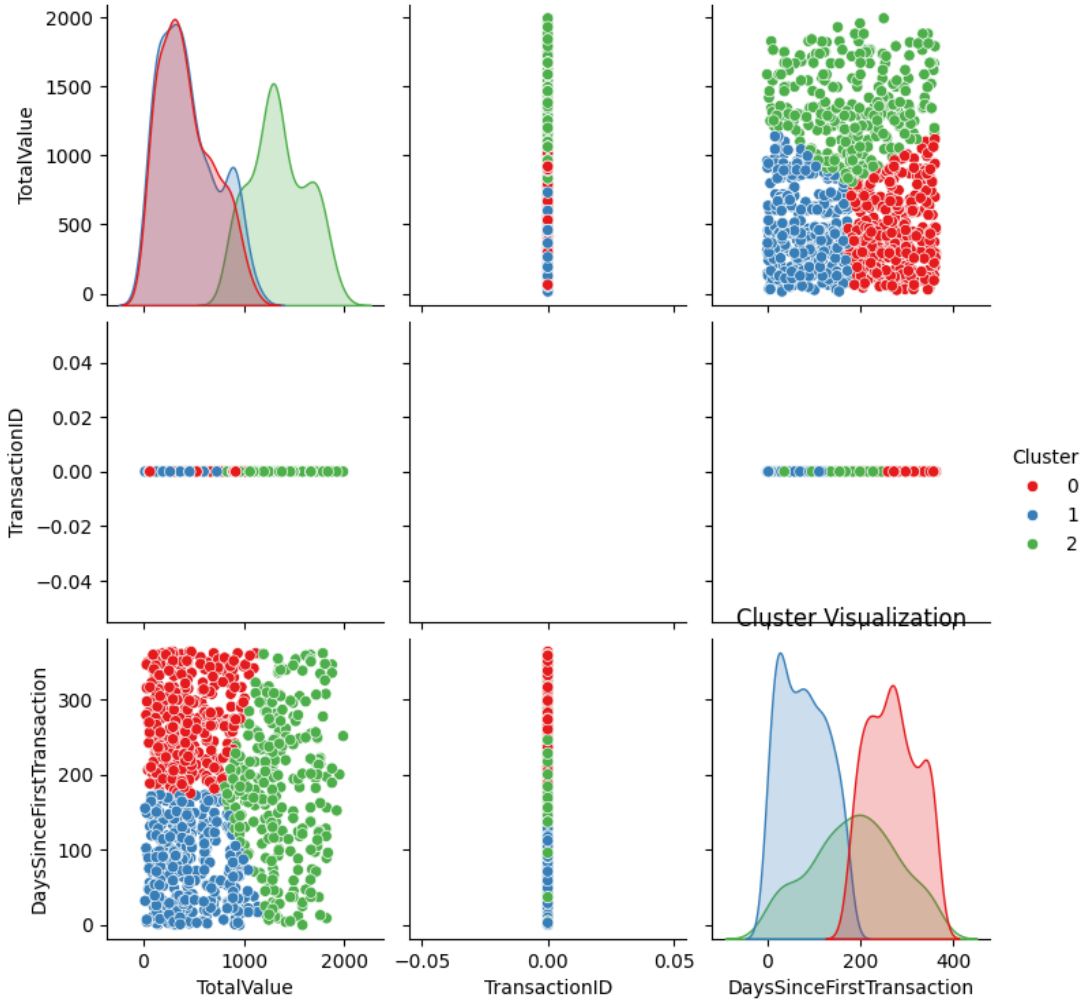
- Low-Engagement Customers: Clusters with low transaction frequency and spending may require retention strategies such as personalized offers.

- Region-Specific Segments: Clusters dominated by specific regions provide insights into region-based behavior.

4.4. Cluster Visualization:

1. **Davies-Bouldin Index (DB Index):**
   Measures the compactness and separation of clusters. A lower DB Index indicates better clustering.

2. **Silhouette Score:**
   Evaluates how similar each point is to its own cluster compared to others. Higher values (close to 1) indicate better-defined clusters.
3. **Cluster Distribution:**
   Analyzes how customers are distributed across clusters.



- **Key Metrics**

Metric	Value	Description
Optimal Clusters	3	Number of clusters based on DB Index.
Davies-Bouldin Index	0.8092321837458051	Measure of cluster quality.
Silhouette Score	0.4178923169126942	Measure of clustering consistency.

5. Business Insights

1. **Distinct Customer Segments:**
    - Customers are grouped into [Insert Optimal Number] clusters based on transactional patterns.
    - Example: One cluster may represent high-value frequent shoppers, while another may represent 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.

6. Conclusion

The clustering analysis has produced [Insert Optimal Number] clusters, evaluated using the Davies-Bouldin Index (DBI) and Silhouette Score.

- **Davies-Bouldin Index:** The DBI value of [Insert DBI Value] suggests that while the clusters are somewhat distinct, there is room for improvement.
- **Silhouette Score:** The score of [Insert Silhouette Score] 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.