Clustering Analysis Report

Clustering Analysis Report

---

1. Introduction

Clustering customers based on transaction data to uncover behavior patterns.

2. Data Preprocessing

Merged data from Customers.csv, Products.csv, and Transactions.csv. Encoded Region and calculated Avg_Transaction_Value.

3. Number of Clusters Formed

The optimal number of clusters was determined to be 5 based on the following:

- **Davies-Bouldin Index (DBI): A DBI value of 0.4963 for 5 clusters indicated good cohesion and separation, supporting this choice.**
- **Domain Knowledge and Interpretability: Five clusters provided meaningful, distinguishable customer segments in terms of purchasing behavior, transaction volume, and regional factors, making the results interpretable.**
- **Comparison with Other Options: Fewer clusters resulted in overly broad groups, while more clusters led to smaller, less coherent groups. Thus, 5 clusters struck the best balance between granularity and interpretability.**
- **Other Methods (Elbow): Although the elbow method was considered, the small size of the data allowed us to manually test and validate clustering for all possible values.**

4. Clustering and Evaluation with Davies-Bouldin Index (DBI)

Initially, when clustering the data using all available features, we observed that the Davies-Bouldin Index (DBI) remained above 1, indicating weak clustering performance. This high DBI value was likely due to multicollinearity among several features, especially between product categories like Category_Books, Category_Clothing, Category_Electronics, Category_Home Decor, and transaction-related variables such as TotalValue and Number_of_Transactions. The strong

correlations among these features made it difficult for the clustering algorithm to find distinct customer groups, resulting in poor cluster separation.

To address the issue of multicollinearity, we reduced the feature set by removing highly correlated variables. We dropped redundant product categories and transaction variables that were closely related, such as Category_Books and Category_Clothing, which led to a reduction in multicollinearity and improved the clustering process.

Results After Feature Reduction:

- DBI Improvement: With the reduced feature set, the DBI significantly improved, dropping to a value between 0.6 and 0.7. This indicated a clear improvement in clustering performance, as the reduced feature set allowed for better separation of customer groups.
- Better Cluster Separation: The reduced feature set also provided more meaningful customer segments. The clusters were now more distinct and interpretable, reflecting different customer behaviors, transaction volumes, and product preferences.

To further enhance the clustering performance, we applied Principal Component Analysis (PCA) to the standardized data. PCA helped to reduce the dimensionality of the dataset while retaining a substantial amount of variance, simplifying the data without losing important information.

PCA Results:

- Explained Variance Ratio: The first principal component accounted for 20.6% of the total variance in the data, allowing for efficient dimensionality reduction without compromising the dataset's information content.
- Total Variance Retained: After applying PCA, the total variance retained was also 20.6%, which was a good balance between reducing the complexity of the data and retaining its essential characteristics.

After applying PCA, we ran KMeans clustering again on the data with reduced dimensions. The DBI value after this step dropped to 0.4963, which indicated a significant improvement in clustering quality, showing better cluster cohesion and separation. This reduced DBI value suggested that the clusters formed after PCA were more compact and better separated, making the clustering results more meaningful and interpretable.

Overall, the combination of feature reduction and PCA significantly enhanced the clustering performance. The final clustering solution, with a DBI of 0.4963, demonstrated improved customer segmentation with clear, distinct groups that

reflected meaningful differences in customer behavior, transaction patterns, and product preferences.

5.Other Relevant Clustering Metrics

Other relevant clustering metrics include the Silhouette Score, which measures how well-separated the clusters are and how similar points within a cluster are to each other. While we didn’t explicitly calculate this due to the small dataset size, it is a metric worth considering for future evaluations.