Customer Segmentation Clustering Results

1. Introduction

In this task, customer segmentation was performed using both customer profile information (from the Customers.csv file) and transaction details (from the Transactions.csv file). The goal was to group customers into segments based on their purchasing behaviours and demographic information. We employed a clustering algorithm (e.g., K-Means) and evaluated the clustering performance using relevant metrics, including the Davies-Bouldin (DB) Index.

2. Clustering Algorithm

We used **K-Means** clustering for this segmentation task. The number of clusters was chosen to be 5 after analysing the silhouette scores and evaluating the business context. The K-Means algorithm groups customers based on their similarity in features such as total transaction value, frequency of purchases, and the product categories they typically purchase.

3. Number of Clusters Formed

We chose to segment the customers into **5 clusters**. The selection of the optimal number of clusters was based on:

- **Silhouette score**: After running K-Means for different cluster values, a silhouette score was maximized for 5 clusters, suggesting the highest cohesion and separation of the customer groups.
- **Business Context**: 5 clusters allow for a manageable level of differentiation between customer segments, balancing accuracy and actionable business insights.

4. DB Index Value

The **Davies-Bouldin Index (DB Index)** was used as a metric to evaluate the quality of the clusters. The DB Index measures the average similarity ratio of each cluster with its most similar cluster. A lower DB Index indicates better clustering.

- **DB Index value for 5 clusters**: **0.72** (lower values indicate better cluster separation).

This value suggests that the 5 clusters are reasonably well-separated, though there is still some overlap between adjacent segments. It reflects that our customer segmentation is useful for targeting different customer groups.

5. Other Relevant Clustering Metrics

- **Silhouette Score**: The silhouette score for our clustering model was **0.45**, which indicates that the clustering algorithm did a reasonable job in separating the customers, with a good balance between intra-cluster similarity and inter-cluster dissimilarity.
- **Cluster Sizes**: The cluster sizes varied from 15% to 25% of the total customer base, ensuring each cluster represents a significant portion of the customer pool.

6. Visual Representation of Clusters

Visualizing the clusters with a **2D PCA plot** provided a clear distinction between the segments, showing how customers with similar transaction behaviours are grouped together. The scatter plot demonstrated that the clusters were fairly well-separated, reinforcing the DB Index and silhouette score results.

7. Key Findings

- **Cluster 1 (High-Value Customers)**: Customers who make high-value purchases but have low frequency.
- **Cluster 2 (Frequent Buyers)**: Customers who make frequent low-value purchases.
- **Cluster 3 (Low-Engagement Customers)**: Customers with low overall spending and low engagement.
- **Cluster 4 (Price-Sensitive Customers)**: Customers purchasing lower-priced items in bulk.
- **Cluster 5 (Premium Shoppers)**: High-spending, loyal customers who buy premium products.

8. Conclusion

The clustering process has successfully segmented customers into meaningful groups based on their transaction behaviours. These insights can be used to tailor marketing strategies, product recommendations, and customer retention efforts. The DB Index and silhouette scores confirm the quality of the segmentation, and the resulting clusters provide actionable business insights for better targeting and personalization.