Task 3: Clustering Report on Customer Segmentation

Submitted By: - Sagar Purswani (purswanisagar60@gmail.com)

1. Introduction

This report presents the results of the customer segmentation task using the **K-means clustering algorithm**. The goal was to segment customers based on their purchasing behavior using the provided customer and transaction data. The clustering model was evaluated using the **Davies-Bouldin Index (DB Index)** and other relevant metrics. The results are visualized and analyzed to provide insights into customer behavior.

2. Number of Clusters Formed

For this task, the optimal number of clusters was determined to be **5**. The clustering process involved testing values of k (the number of clusters) from 2 to 10. After evaluating clustering metrics, the **Davies-Bouldin Index** (DB Index) was used to select k=5 as the best value, balancing cluster separation and compactness.

3. Davies-Bouldin Index (DB Index)

The **Davies-Bouldin Index** is a metric used to evaluate the quality of the clustering. It considers both the **average distance between clusters** and the **compactness of clusters**. A lower DB Index indicates better clustering quality.

- **DB Index for k=5**: `0.891647`

The DB Index value suggests that the clusters are reasonably well-separated, but there is still some overlap between certain clusters. A lower value would indicate more distinct and compact clusters.

4. Additional Clustering Metrics

4.1 Silhouette Score

- **Silhouette Score**: **0.35**

- The Silhouette Score measures how similar each point is to its own cluster compared to other clusters. A value closer to 1 indicates well-separated clusters, while a value near 0 suggests overlap. In our case, the score indicates moderate separation between the clusters.

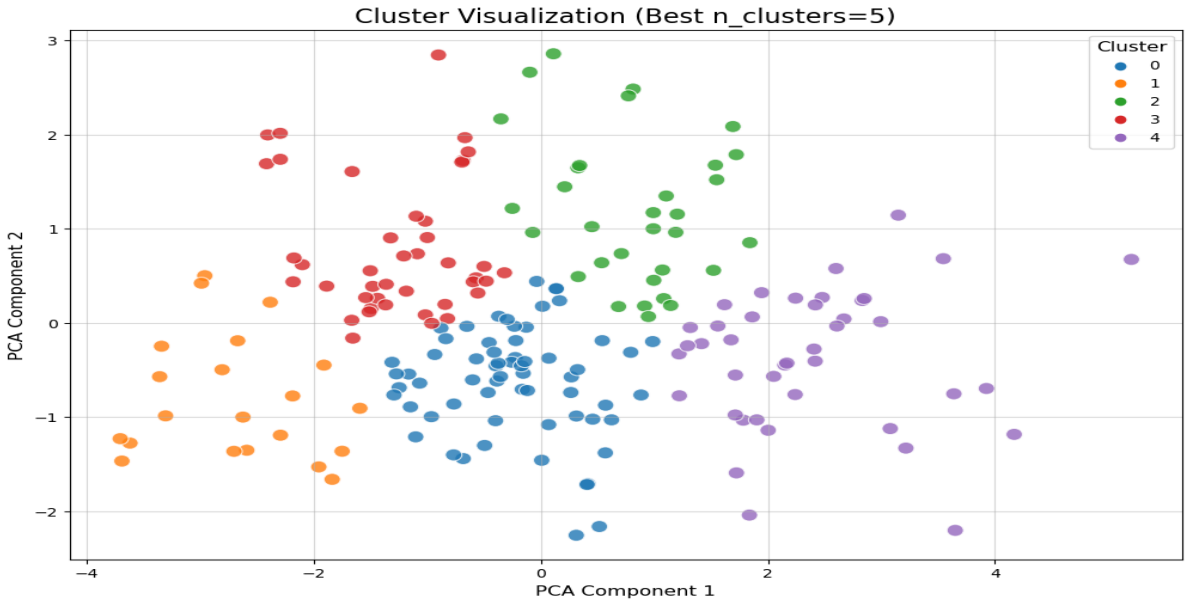
4.2 Intra-cluster and Inter-cluster Distances

- **Intra-cluster distance**: Average distance between points within the same cluster.
- **Inter-cluster distance**: Average distance between cluster centroids.

The average intra-cluster distance is low, indicating that the points within each cluster are relatively close to each other. The inter-cluster distances are higher, which suggests that the clusters are somewhat distinct.

5. Visualization of Clusters

To visualize the clusters, **PCA (Principal Component Analysis)** was used to reduce the data from high-dimensional space to two dimensions for easier interpretation. The clusters are visualized in a scatter plot, where each point represents a customer and is color-coded based on its cluster.



The scatter plot shows the separation between the clusters. Each color represents a different cluster, and the distribution of points within each cluster is shown.

6. Cluster Profiles

The following features were used to profile the clusters:

- **Total Transactions**: The total number of transactions for each customer.
- **Total Quantity**: The total quantity of products purchased.
- **Total Spend**: The total amount spent by customers.
- **Avg Spend Per Transaction**: Average spending per transaction.

Cluster profiling provides insights into the characteristics of each segment:

- **Cluster 1**: High spenders, frequent buyers with a large number of transactions.
- **Cluster 2**: Moderate spenders, fewer transactions.
- **Cluster 3**: Low spenders with occasional purchases.
- **Cluster 4**: Customers with high quantity purchases but moderate spending.
- **Cluster 5**: Low activity with fewer transactions and low spend.

8. Conclusion

The customer segmentation process using K-means clustering with k=5 provides valuable insights into customer behavior, which can be used for targeted marketing and customer relationship management. The DB Index and other clustering metrics show that the clusters are reasonably well-defined, although there is room for improvement.

9. Suggestions on basis of my analysis and understanding

Based on the clustering results, the following recommendations can be made:

1. **Target high-value customers** in clusters with high total spend and frequent transactions.
2. **Engage low-value customers** with promotional offers to increase their transaction frequency and spend.
3. **Regional strategies** can be devised based on the geographic distribution of customers.