Clustering Results Report

1. Overview

Task: To segment clients based on transaction and demographic data.

Algorithm: The K-Means clustering algorithm was used to group the clients based on three important characteristics:

- Total Value Spent
- Days Since Signup
- Total Amount Purchased

To ensure comparability across different scales, these were first merged at the customer level and then standardized. To determine the optimal number of clusters, the clustering process was evaluated in terms of several measures, such as inertia, silhouette coefficient, and the Davies-Bouldin index, DB Index.

2. Optimal number of clusters:

The following key results were obtained from the performance analysis of the clustering:

- The inertia (Elbow Method) technique displays the sum of squared distances between data points and their designated cluster centroids. Naturally, as the number of clusters rises, inertia falls. The "elbow" approach usually shows the point at which the inertia is no longer significantly reduced by adding more clusters. According to this point, there should be about five clusters.

- Silhouette Score: This score measures the similarity of an individual data point to its own cluster compared to other clusters. A larger silhouette score indicates clusters are compact and fairly well separated. Five clusters obtained the highest silhouette scores, which means this is the optimal number.

Thus the optimal number of clusters is 5.

3. Davies-Bouldin index

The Davies-Bouldin index is another measure of the effectiveness of clustering, with a lower score indicating better clustering (meaning clusters are well separated and have minimal overlap). The calculated Davies-Bouldin Index for the best 5 clusters is **0.872**.

This value is fairly low, meaning that the clusters are well separated and different from each other, and the 5-cluster solution was successful.

4. Clustering Metrics Summary

The metrics of the clustering process are as follows:

- Five is the optimal number of clusters.
- The DB Index of the Davies-Bouldin is 0.872, indicating good separation of the clusters.
- Silhouette Score: Five clusters had the highest silhouette score, which confirms that this is the number of clusters with the ideal amount.

5. Cluster Information

In the final clusters, TotalValue average spending, Quantity average numbers purchased, and DaysSinceSignup average days since signup were analyzed.

- Cluster 0 includes customers who are heavy spenders and buy with higher frequencies.
- Cluster 1 holds those who are moderate spenders and have been a relatively long time since their signup.
- Cluster 2 is relatively low spenders and buy moderately and have been with the company for a short time.
- Cluster 3 is comprised of low spenders with the short signup history.
- Cluster 4 is very high spenders that have huge numbers of purchase amount and short signup history.

6. Representing the Clusters

- Scatter Plot of Clusters: A graphical representation of the clusters based on Total Value (expenditure) and Quantity Bought was developed, with each cluster represented in a unique color. This provides an easy illustration of how the customer segments differ in terms of their buying habits.
- Cluster Summary Bar Plot: A bar plot was used to understand the average value for different metrics such as Spend, Quantity, and Days Since Signup across different clusters. This gives a more holistic view of the characteristics of each cluster.

7. Conclusion

The clustering analysis has successfully categorized customers into 5 different groups. With the help of several metrics, including the Davies-Bouldin Index (0.872) and Silhouette Score, it was verified that the clusters are well formed. All these results can be used for focused marketing, customer behavior analysis, and other services based on different segments of customers.