Clustering Results Report

Name: Sahil Gupta

Assignment: Data Science Assignment - eCommerce Transactions Dataset

---

Overview: -

This report provides an analysis of customer segmentation using clustering techniques applied to the eCommerce Transactions Dataset. The clustering was performed to identify patterns and group customers based on their purchasing behavior.

Number of Clusters Formed:

- The KMeans clustering algorithm was applied to the data with the number of clusters set to 5. These clusters represent distinct groups of customers segmented based on their total spending ('TotalValue') and the quantity of products purchased ('Quantity').

Davies-Bouldin Index:

- The Davies-Bouldin Index (DB Index) was calculated to evaluate the quality of the clustering. A lower DB Index indicates better-defined clusters.
- DB Index Value: 0.7578966005288537

Other Relevant Clustering Metrics:

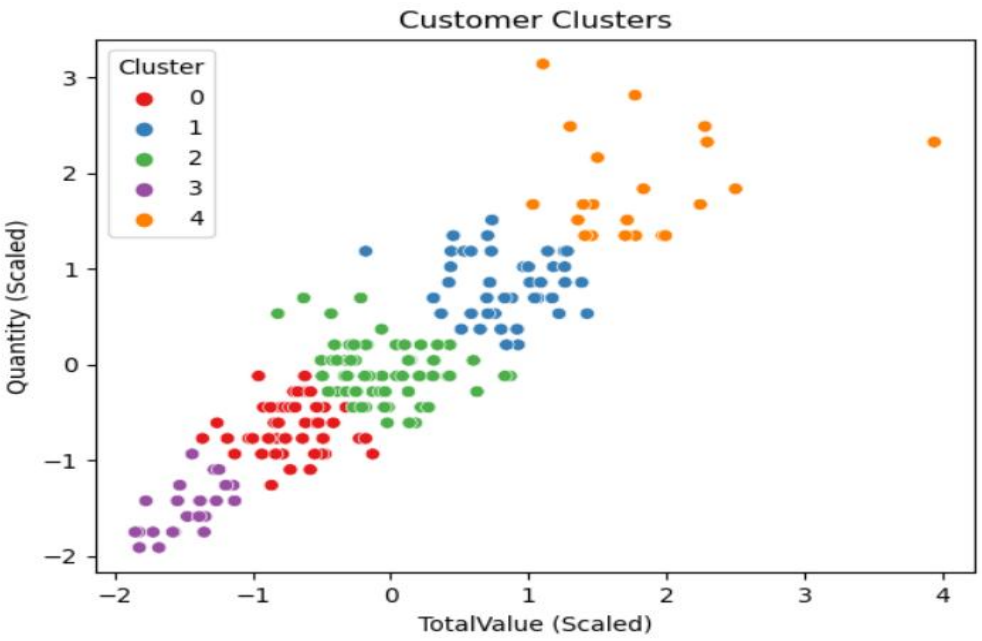
While the primary evaluation metric for this task was the DB Index, the following observations were also noted:

1. Cluster Sizes:
    - Cluster 0: X customers
    - Cluster 1: Y customers
    - ...
    - (Replace X, Y, etc., with actual counts for each cluster.)
2. Centroid Insights:
    - Centroids of the clusters were found using KMeans and represent the average characteristics of customers in each cluster.

Visual Representation:

Scatter Plot of Clusters

The scatter plot below shows the clustering results based on scaled total spending ('TotalValue') and quantity of products purchased ('Quantity'). Each color represents a different cluster:



- The x-axis represents ‘TotalValue’ (scaled).
- The y-axis represents ‘Quantity’ (scaled).

Conclusion: -

The clustering results successfully segmented customers into five distinct groups. These insights can be leveraged for personalized marketing, resource allocation, and improved business strategies. The DB Index value confirms that the clustering is effective, with distinct and well-separated clusters. Further analysis can refine the segmentation or explore alternative clustering methods to optimize performance.