Clustering Segmentation Report:

Objective:

The goal of this analysis was to segment customers into distinct groups based on specific behavioral parameters, such as total transactions, spending patterns, customer activity duration, and region. This process helps to identify customer groups with similar behaviors.

The following steps were undertaken by me to complete the assignment:

Step 1: Data Normalization:

To ensure the features used in clustering were on a similar scale, a Min-Max Normalization approach was applied. This method scaled all numerical values of the dataset between 0 and 1.The values of the various columns were of varying ranges and therefore, the normalization became essential because K-Means relies on the distance between data points, and having features on different scales could skew the results.

Step 2: Clustering Method:

K-Means Clustering was chosen for segmenting the customers into groups. K-Means is an iterative algorithm that partitions the dataset into K distinct clusters based on the similarity of the data points. However, before moving ahead in the clustering process, the challenge was to identify the optimal value of K (the number of clusters from a given range of 2 to 10).

Step 3: Evaluation Metrics:

To determine the best value of K, two metrics were used:
1) DB Score: The DB Score measures how well the clusters are separated and how compact they are. It is to be noted that a lower DB Score indicates that the clusters are well-separated and compact. In this analysis, the lowest DB score was observed at K=10, indicating that this is the most optimal number of clusters in terms of compactness and separation.

2) The Silhouette Score evaluates how well each point fits into its assigned cluster. It compares the average distance between a point and others within the same cluster to the average distance to points in the nearest different cluster. The Silhouette score ranges from -1 to 1, where higher values indicate better-defined clusters. The highest Silhouette Score was observed at K=10, suggesting that 10 clusters offer the best grouping of customers in terms of intra-cluster similarity and inter-cluster dissimilarity.

Given the low DB score and high Silhouette score at K=10, it was concluded that 10 clusters is the most optimal number of clusters for this dataset.

Step 4: Principal Component Analysis (PCA):

To better understand the clustering segmentation, Principal Component Analysis (PCA) was applied. PCA is a dimensionality reduction technique that converts multiple features in the dataset into two principal components, allowing for visualization of the clusters in a 2D space.

By reducing the dimensions to just two axes, it becomes easier to visualize how the clusters are distributed and how distinct each group is from the others. The 2D plot of the clusters showed clear separations between the customer groups, confirming the effectiveness of the clustering segmentation.

Conclusion:

- K-Means Clustering with K=10 clusters was identified as the most optimal segmentation for this dataset, based on both the DB Score and Silhouette Score.
- Using the PCA helped visualize the segmentation by reducing the feature space to two principal components.
- This customer segmentation can be used for more targeted strategies, such as personalized marketing or recommendations, as it identifies distinct customer groups with similar behaviors and preferences.