Task-3 : Task 3: Customer Segmentation / Clustering

1. Number of Clusters Generated:

The KMeans algorithm determined 3 different clusters of customers, both from the profile and the transactional information.

The number of clusters was set at 3 (optimal_clusters).

2. Cluster Distribution:

The number of customers spread across the three clusters is

Cluster 0: 78 customers

Cluster 1: 104 customers

Cluster 2: 17 customers

These counts demonstrate how the customers are categorized according to the similarities in their profiles and transaction history.

3. PCA Components:

PCA was applied to the dataset to reduce the dimensionality, and the first two components were obtained for the purpose of visualizing and understanding the data distribution in a 2D space.

The PCA Components for each customer are as follows:

CustomerID	PCA Component 1	PCA Component 2	Cluster
C0001	0.031957	-0.196448	1
C0002	-0.855328	-1.128049	1
C0003	-0.820778	0.403622	1
C0004	1.812192	-0.751722	0
C0005	-0.944593	-0.066417	1

The first two variables explain the highest amount of data variance and provide the most for distinguishing between groups.

4. Cluster Centroids in PCA-reduced Space:

The cluster center for each in the 2D PCA-reduced space is,

Cluster	PCA Component 1	PCA Component 2
Cluster 1	1.351880	0.379131
Cluster 2	-0.555523	-0.447274
Cluster 3	-2.804250	0.996718

Cluster centroids are points that represent the average of all customers from a given group in the 2D space. These show where the "centre" of every group is.

Cluster 1 is plotted at (1.35, 0.38), meaning the customers for this cluster lie scattered in the vicinity of the area.

Cluster 2 is plotted at (-0.56, -0.45). This means that it is a denser cluster around this central point.

Cluster 3 is plotted way at the extreme at (-2.80, 0.99) which is distinct from other clusters.

5. Davies-Bouldin Index:

The Davies-Bouldin Index, DBI, is a measure that uses to measure the quality of a cluster. The smaller the value of DBI the greater is the separation between clusters.

DB Index: *1.0973*

A DBI value of 1.0973 suggests that while the clusters are reasonably well-separated, there is likely to be some overlap between clusters, especially Cluster 2 and Cluster 1, since they are so

close in the 2D PCA space. A lower DBI score would mean that the clusters are more well-defined, but this score is fairly good for the number of clusters chosen.

Conclusion:

Clusters: The KMeans produced three clusters as follows: 78, 104, and 17 customers in Clusters 0, 1, and 2, respectively.

PCA: The PCA transformation is effective at reducing the feature space to two components, thereby aiding in the visualization of customer segmentation in a 2D space.

Centroids: The centroids in PCA space reveal the central tendency of each cluster and the distance between them gives the distinctiveness of each group.

Davies-Bouldin Index: The DB Index is 1.0973, which indicates a fair but not perfect separation of clusters. A lower score might improve the quality of the clustering.

This cluster analysis offers insight into customer segmentation, so businesses can focus on particular groups of customers by tailoring their strategies according to their transaction behavior and profile.