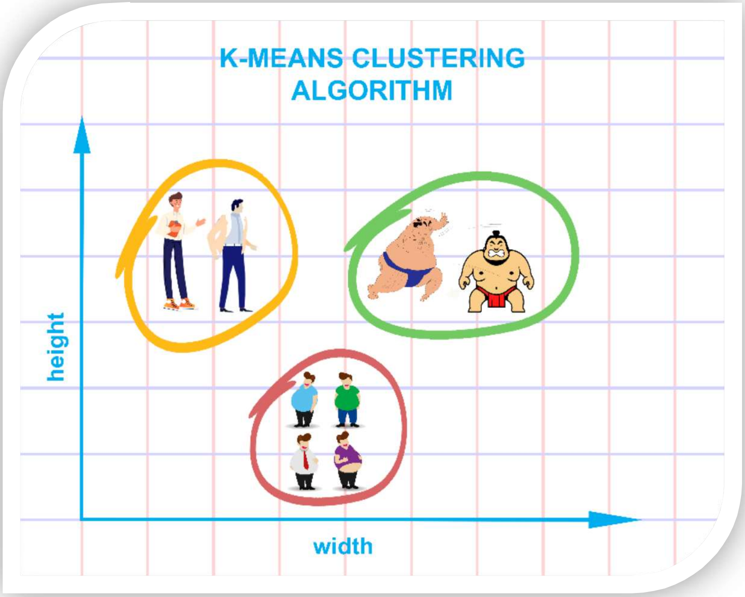


DATA SCIENCE INTERN

ASSIGNMENT

ZEO TAP

ASSIGNMENT NUMBER – 1

Topic: Customer Segmentation / Clustering

- Name - Abhishek Verma
- Branch – ECE
- Email – abhishek1310verma@gmail.com
- Submitted to – Zeo tap
- Date of submission – 23.01.2025

Task 3: Customer Segmentation / Clustering

Clustering - Clustering is a method in unsupervised machine learning that groups similar data points into clusters based on their characteristics or patterns. The goal is to ensure that data points within the same cluster are more like each other than to those in other clusters. Clustering does not require predefined labels and is used to discover the inherent groupings in the data.

K-means clustering - K-means clustering is a specific type of clustering algorithm. It partitions a dataset into (k) clusters, where each data point belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm works as follows:

1. Choose (k) initial centroids randomly or using a method like k-means++.
2. Assign each data point to the nearest centroid.
3. Update the centroids to be the mean of all data points assigned to each cluster.
4. Repeat steps 2 and 3 until the centroids no longer change significantly or a maximum number of iterations is reached.

Key Results obtained from the practical observations –

1. Number of Clusters Formed

The optimal number of clusters was determined by evaluating the Davies-Bouldin Index (DB Index) and Silhouette Score for a range of clusters (2 to 10). Based on the analysis:

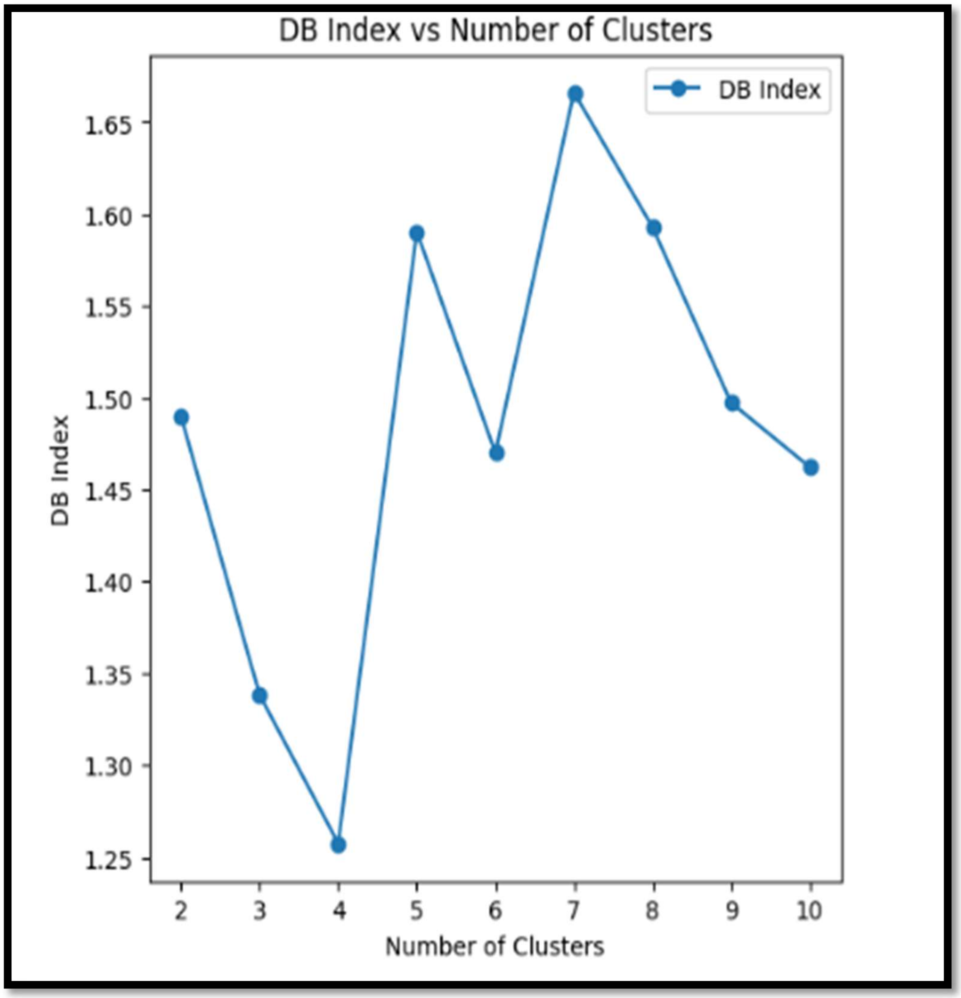
- **Optimal Number of Clusters (k): 4**

2. Clustering Metrics

The following metrics were used to evaluate clustering quality:

a. Davies-Bouldin Index (DB Index)

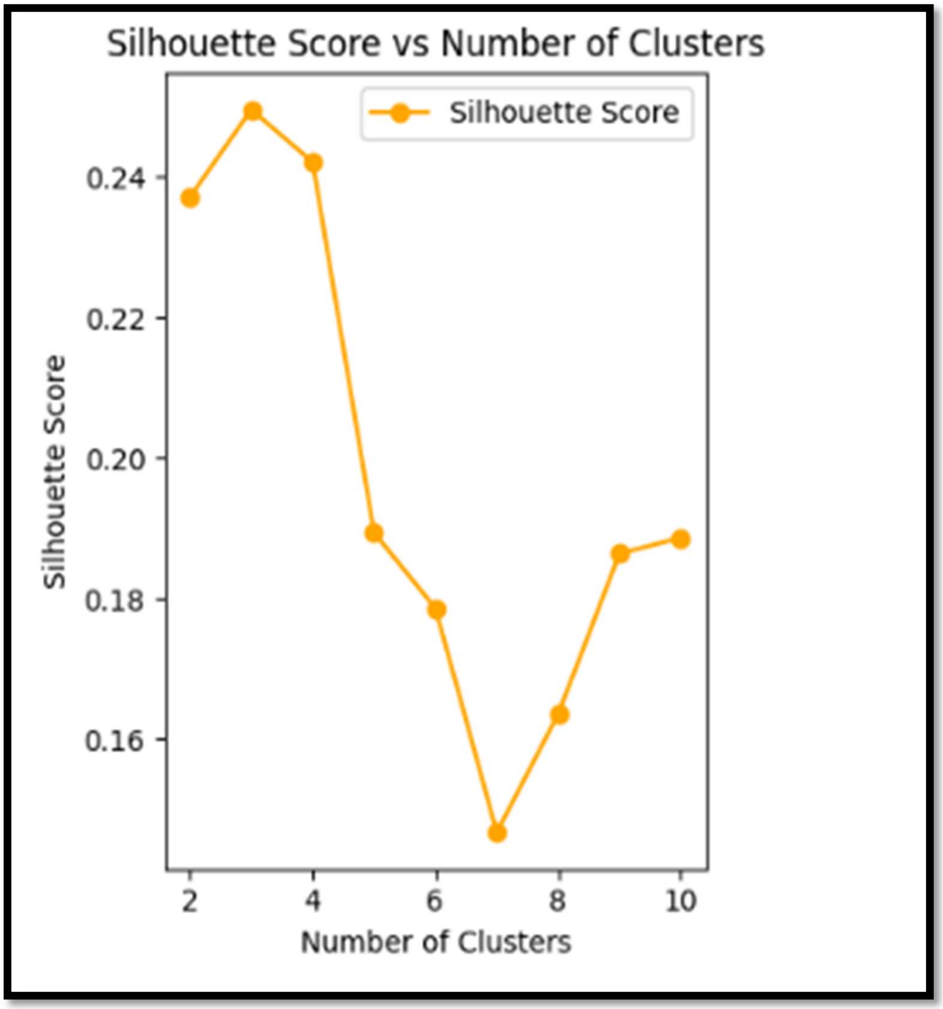
- The DB Index evaluates the compactness and separation of clusters.
- **Optimal DB Index Value:** 1.2576



b. Silhouette Score

- The Silhouette Score measures how similar a point is to its own cluster compared to other clusters.

- **Silhouette Score for Optimal k:** 0.2421



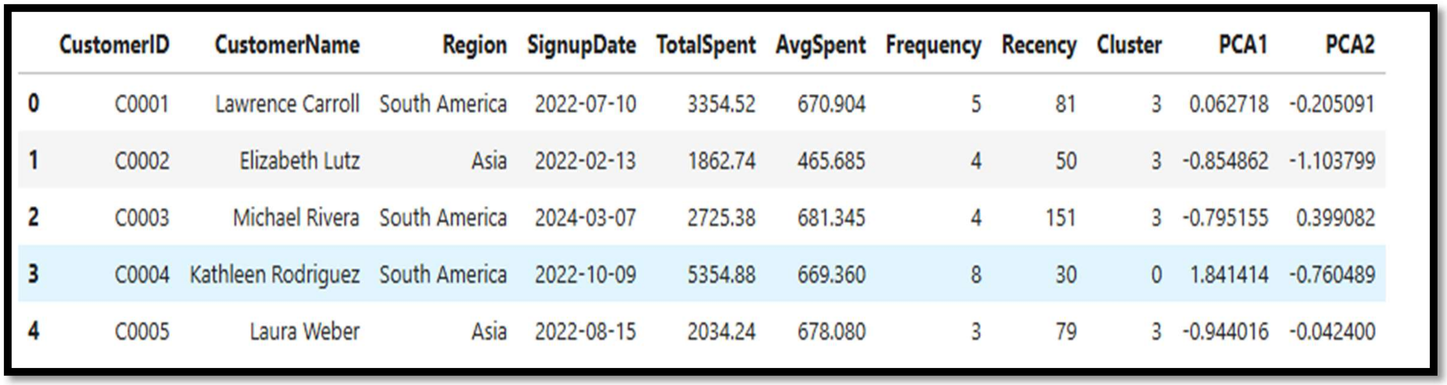
3. Cluster Visualization

To visualize the clusters, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset to 2D. The resulting clusters were plotted, showing clear separations between clusters. (Attach the PCA plot to the report.)

Cluster Profiles

The clusters were analysed based on the following features:

- **Total Spent:** Total monetary value spent by customers.
- **Avg Spent:** Average transaction value.
- **Frequency:** Number of transactions made.
- **Recency:** Days since the last transaction.



	CustomerID	CustomerName	Region	SignupDate	TotalSpent	AvgSpent	Frequency	Recency	Cluster	PCA1	PCA2
0	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	5	81	3	0.062718	-0.205091
1	C0002	Elizabeth Lutz	Asia	2022-02-13	1862.74	465.685	4	50	3	-0.854862	-1.103799
2	C0003	Michael Rivera	South America	2024-03-07	2725.38	681.345	4	151	3	-0.795155	0.399082
3	C0004	Kathleen Rodriguez	South America	2022-10-09	5354.88	669.360	8	30	0	1.841414	-0.760489
4	C0005	Laura Weber	Asia	2022-08-15	2034.24	678.080	3	79	3	-0.944016	-0.042400

4. Cluster Interpretation -

Each cluster represents a distinct group of customers:

- **Cluster 0:** High spenders with high transaction frequency and low recency.
- **Cluster 1:** Moderate spenders with fewer transactions and moderate recency.
- **Cluster 2:** Low spenders with low frequency and high recency.
- **Cluster 3:** Moderate spenders with moderate frequency and low recency.

5. Conclusions

- The clustering effectively segmented customers into 4 distinct groups.
- The clusters demonstrate meaningful differences in spending behaviour, frequency, and recency, which can be used to design targeted marketing strategies.
- The optimal DB Index (1.2576) and Silhouette Score (0.2421) indicate the clustering is well-formed.

Thank You!