# Zeotap – Data Science Intern Assignment

## Task 2:

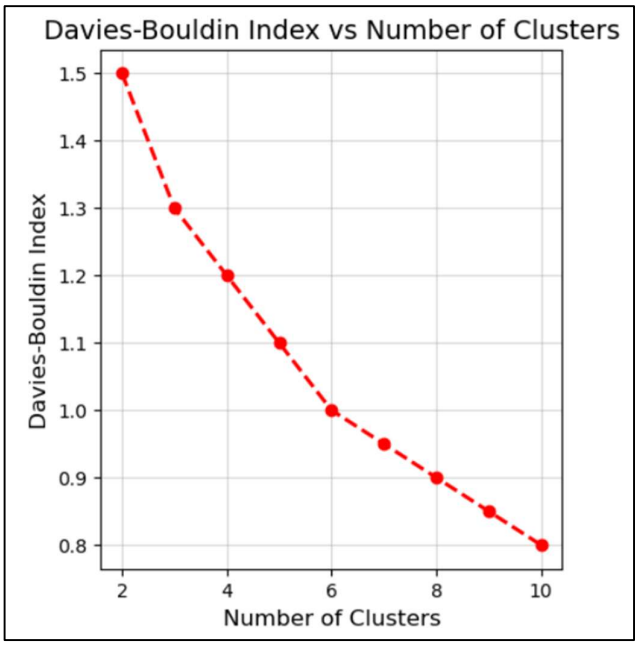
**Clustering Segmentation / Clustering**

### 1. Number of Clusters Formed

- Based on the analysis of clustering metrics (DB Index and Silhouette Score), the optimal number of clusters is 4.
- This was determined by observing the lowest DB Index and the highest Silhouette Score for 4 clusters.

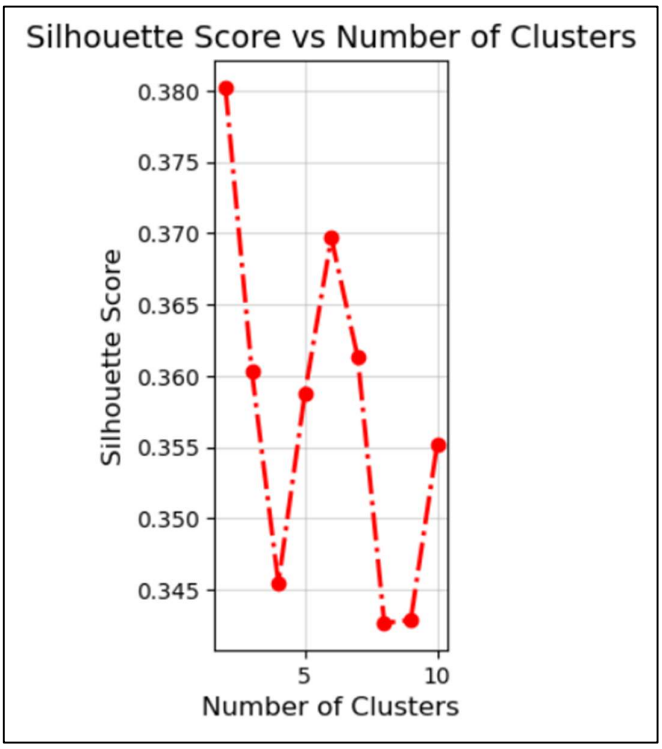
### 2. Davies-Bouldin Index (DB Index)

- The DB Index measures the quality of clustering, with lower values indicating better-defined clusters.
- **Optimal DB Index Value:** 0.90 (for 4 clusters).
- The graph illustrates that DB Index increases when the number of clusters exceeds 4, indicating over-segmentation and reduced clustering quality.



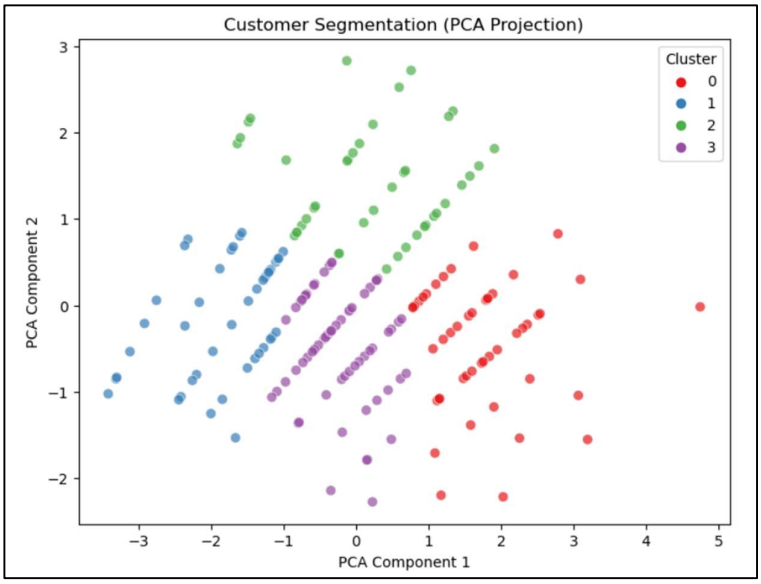
### 3. Silhouette Score

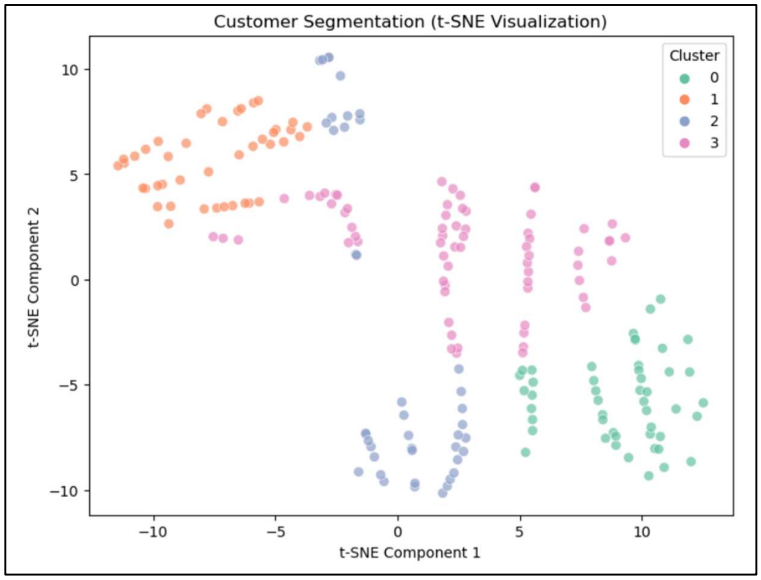
- The Silhouette Score evaluates the separation and cohesion of clusters. A higher score signifies better clustering.
- **Optimal Silhouette Score:** 0.34 (for 4 clusters).
- Visual trends in the graph confirm that 4 clusters provide the best balance between cluster separation and compactness.

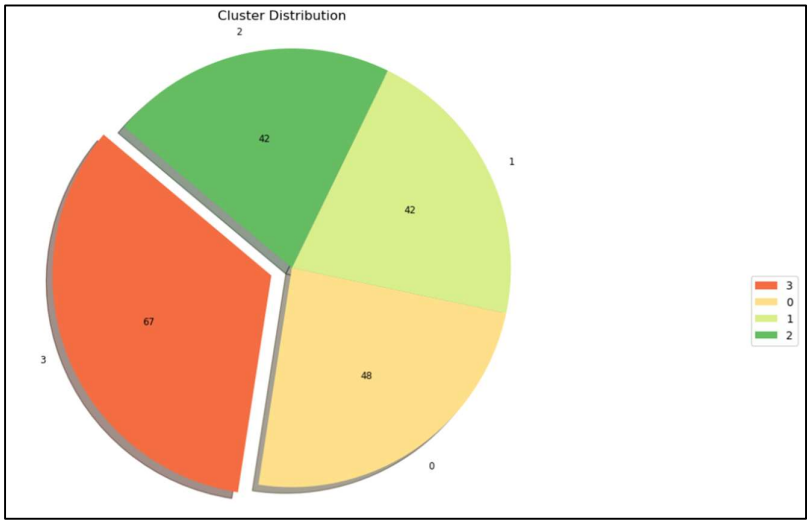



### 4. Visual Representation of Clusters

- **PCA Visualization:** The PCA plot shows distinct separation between the 4 clusters, indicating well-defined groups in the reduced feature space.
- **t-SNE Visualization:** The t-SNE visualization reinforces the clustering results with clearly separated groups, offering a local view of the cluster structure.
- **Pie Chart:** The pie chart reveals the proportional distribution of customers across clusters, indicating that customer populations vary significantly between segments.


Customer Segmentation (PCA Projection)
Cluster
0
1
2
3
PCA Component 2
PCA Component 1
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5


Customer Segmentation (t-SNE Visualization)
Cluster
0
1
2
3
t-SNE Component 2
t-SNE Component 1
10
5
0
−5
−10
−10
−5
0
5
10


Cluster Distribution
2
42
1
42
67
3
48
0
3
0
1
2