DATA SCIENCE ASSIGNMENT

E-Commerce Transactional Dataset

Task 3 : Customer Segmentation/ Clustering

Clustering:

- **Clustering** is a **data analysis technique** used to **group data points** with similar characteristics into **clusters or groups.**
- Each cluster contains **data points**, that are similar to each other.
- Helps to identify the **patterns and structures** in data without any prior labels or categories.

Need of Clustering:

- **Understand customer behaviour** by grouping them based on purchasing patterns or preferences.
- **Personalize marketing strategies** for different customer groups.
- **Identify high-value customers** or potential churners.
- **Optimize resource allocation** and decision-making.

Clustering Methods:

- K-Means Clustering
- Hierarchal Clustering
- Density-Based Clustering (DBSCAN)
- Gaussian Mixture Models (GMM)

For the above Clustering methods, we used **K-Means Clustering Method**.

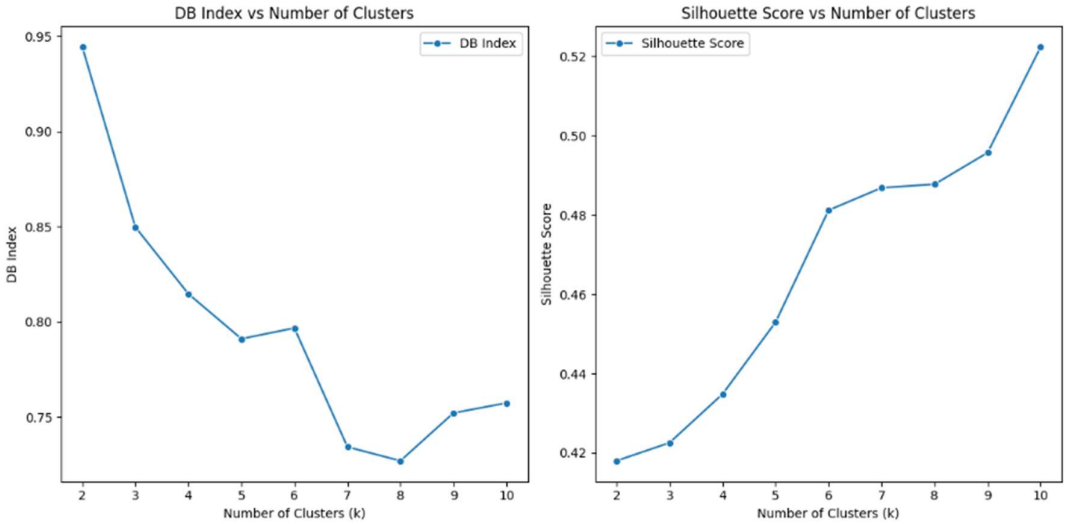
Why Use K-Means Clustering?

- It is **simple**, **efficient**, and works well with structured data.
- It performs well for **large datasets**, such as our e-commerce data.
- The number of clusters can be **pre-defined**, allowing control over segmentation granularity.
- It provides **clear and interpretable clusters**, making it suitable for **customer segmentation**.

Findings:

1. **Number of Clusters:** The optimal number of clusters identified is **8**. This means we segmented customers into 8 different groups based on their profiles and purchasing behaviours.
2. **DB Index Value:** The **Davies-Bouldin Index** (DB Index) measures the separation and compactness of the clusters. A lower DB Index indicates better clustering. The **final DB Index** is **0.73**, which indicates a moderate separation between the clusters. This value suggests that the clusters are reasonably well-separated, but there may still be some overlap in customer behaviour.
3. **Silhouette Score:** The **Silhouette Score** quantifies how similar customers are within their own clusters compared to other clusters. A higher score indicates better-defined clusters. The **final Silhouette Score** is **0.49**, indicating that the clusters have a fair level of cohesiveness, though there may be room for improvement in terms of separating certain customer groups more distinctly.

Visualizations:



The visualizations (line plots) of the **DB Index** and **Silhouette Score** versus the number of clusters helped in determining the optimal number of clusters. Based on the DB Index, the best cluster count was found to be **8**, aligning with the pattern observed in the plots.