# REPORT ON CLUSTERING (TASK 3)

1. Number of Clusters Formed

- Optimal Number of Clusters (k): 7
    - This was determined using the Silhouette Score, where the highest score was achieved at k=7.

2. Clustering Metrics

- Davies-Bouldin Index (DB Index): 0.905
    - A lower DB Index indicates that clusters are compact and well-separated. With a value below 1, the clustering quality is excellent.
- Silhouette Score: 0.414
    - A positive Silhouette Score closer to 1 indicates that the clusters are well-defined and reasonably distinct.
    - A value of 0.414 suggests moderate separation between clusters.

3. Methodology

Data Preprocessing:

1. RFM Features:
    - Recency: Calculated as the number of days since the customer's last transaction.
    - Frequency: Number of transactions made by each customer.
    - Monetary: Total monetary value of transactions.
2. Transformations:
    - Log-transformed the Monetary feature to reduce skewness.
    - Standardized all numerical features using StandardScaler for uniform scaling.
3. Categorical Encoding:
    - One-hot encoding was applied to the Region feature to handle categorical data.

Clustering Approach:

- K-Means Clustering was used to segment customers into 7 clusters.

- The optimal k=7 was chosen based on:
    - The Elbow Method for observing the Sum of Squared Errors (SSE).
    - The Silhouette Score, which peaked at k=7.
    - The Davies-Bouldin Index, indicating better cluster compactness.

**4**. PCA Visualization

- A PCA-based 2D scatter plot was created to visualize clusters in reduced dimensions.
- The clusters are reasonably distinct in the 2D space, suggesting meaningful segmentation.

Other relevant clustering methods:

-Cluster Size Distribution:

- Number of data points in each cluster.
- This helps understand if clusters are balanced or if there are outliers/skewed clusters.

-Inertia (Sum of Squared Distances):

- Measures how tightly the points are grouped around their centroids.
- Lower inertia values indicate better clustering.

-Cluster Centroids:

- Describes the central characteristics of each cluster.
- Useful for interpreting cluster patterns.

-Inter-Cluster and Intra-Cluster Distances:

- Measures the distance between different clusters and the spread within each cluster.
- Helps verify if clusters are distinct and compact.