TASK 3: Customer Segmentation / Clustering Report

1. Clustering Logic and Methodology:

For this task, we applied KMeans clustering to segment customers based on their profile information and transaction data. The process followed the below steps:

- **Data Preprocessing:**
    - **Merge Data:** Combined the profile data from Customers.csv with the transaction data from Transactions.csv based on the CustomerID.
    - **Feature Engineering:**
        - Aggregated customer data to create features that represent total spend, number of transactions, and average quantity purchased per transaction.
    - **Handling Missing Data:** Missing values were imputed using the mean of the respective feature.
    - **Feature Scaling:** Used StandardScaler to normalize features, as KMeans is sensitive to the scale of the data.
- **Clustering:**
    - We used KMeans clustering and experimented with the number of clusters ranging from 2 to 10.
    - The KMeans algorithm was chosen for its simplicity and efficiency for customer segmentation tasks.
    - KMeans was run multiple times with different cluster values, and the number of clusters was selected based on clustering performance metrics.

2. Clustering Results (for 8 clusters):

- Number of Clusters Formed: 8 clusters were formed after evaluating different cluster numbers.
- **Clustering Metrics:**
    - **Davies-Bouldin Index (DB Index):**
        - DB Index for 8 clusters: 0.9382
        - The Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster, where a lower value indicates better clustering. In this case, a value of 0.9382 suggests moderate cluster separation.

- **Silhouette Score:**
  - Silhouette Score for 8 clusters: 0.3046
  - The Silhouette Score measures how well-separated and cohesive the clusters are. A value closer to 1 indicates good clustering, while a score closer to 0 suggests overlapping clusters. In this case, the score of 0.3046 indicates moderate clustering quality.
- **Inertia:**
  - Inertia for 8 clusters: 116.19
  - Inertia measures the sum of squared distances between each point and its cluster centroid. A lower inertia value indicates more compact and well-defined clusters.

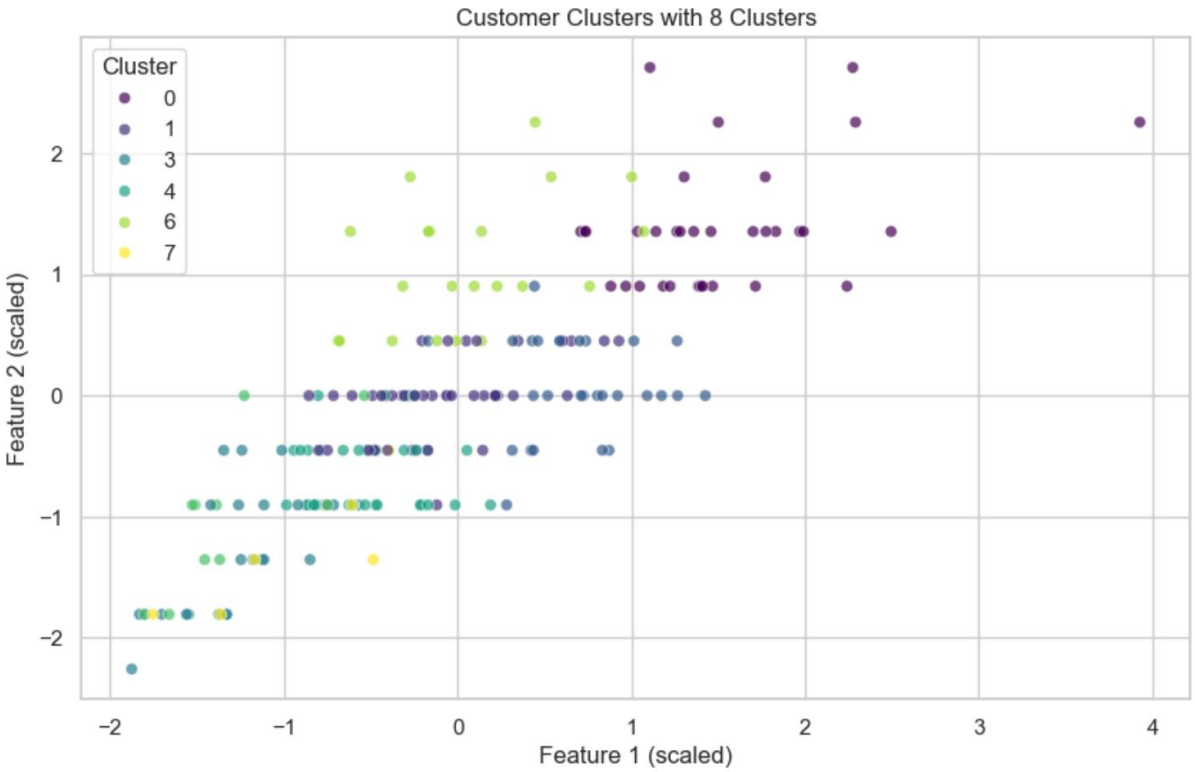
3. Visual Representation:

We visualized the clusters using a scatter plot, where each point represents a customer. The clusters are distinguished by different colors. The x-axis represents the Feature 1 (scaled) while the y-axis represents the Feature 2 (scaled)

- The clusters are well-separated in this visualization, though there is some overlap due to the moderate Silhouette Score and DB Index.

4. Summary of Clustering Metrics for 8 clusters:

- **Number of Clusters: 8**
- **DB Index: 0.9382**
- **Silhouette Score: 0.3046**
- **Inertia: 116.19**



Conclusion:

- The KMeans clustering model effectively segmented the customers into 8 clusters.
- The clustering results show moderate cohesion and separation between the clusters, with the Silhouette Score of 0.3046 indicating that the clustering could still be improved.
- The Davies-Bouldin Index (0.9382) suggests moderate separation between the clusters, while the Inertia (116.19) suggests reasonable compactness within the clusters.