Customer Segmentation Report

Clustering Logic and Metrics

- **Clustering Algorithm**: KMeans clustering was used to segment customers based on both profile and transaction information. The customer profiles were augmented with transaction data such as total spend, total transactions, and quantity of items purchased. The data was standardized before applying the clustering algorithm to ensure that each feature contributed equally to the distance calculations.
- **Clustering Metrics**: We evaluated the performance of the clustering results using two primary metrics:
    1. **Davies-Bouldin Index (DB Index)**: The DB Index measures the compactness and separation of the clusters. Lower values indicate better clustering results. A lower DB Index suggests that the clusters are more distinct and well-separated.
    2. **Silhouette Score**: The Silhouette Score quantifies how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A higher silhouette score indicates better-defined clusters.

Results

- For different values of clusters (2 to 10), the following DB Index and Silhouette Scores were obtained:

Number of Clusters	DB Index	Silhouette Score
2	1.635	0.236
3	1.415	0.308
4	0.927	0.479

5	0.870	0.480
6	0.821	0.501
7	0.733	0.491
8	0.718	0.489
9	0.710	0.481
10	0.710	0.469

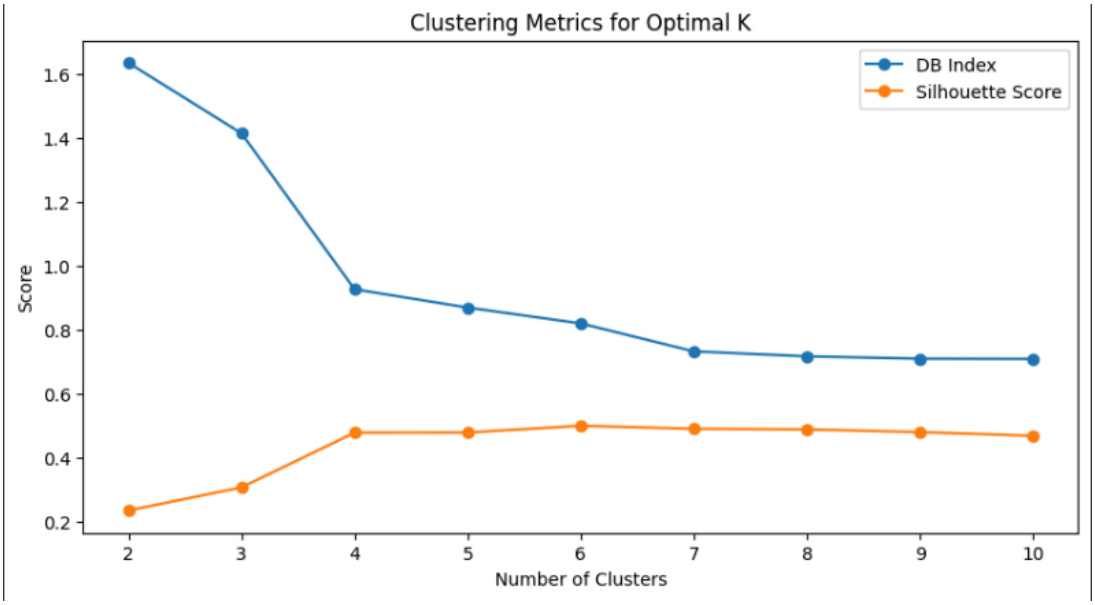
- **Optimal Number of Clusters**: Based on the DB Index, **10 clusters** was determined to be optimal since it achieved one of the lowest DB Index values (0.710), indicating better cluster separation.
- **Final Clustering Metrics**:
  - **Optimal Clusters**: 10
  - **Minimum DB Index**: 0.710 (for 9 and 10 clusters)
  - **Maximum Silhouette Score**: 0.501 (for 6 clusters)

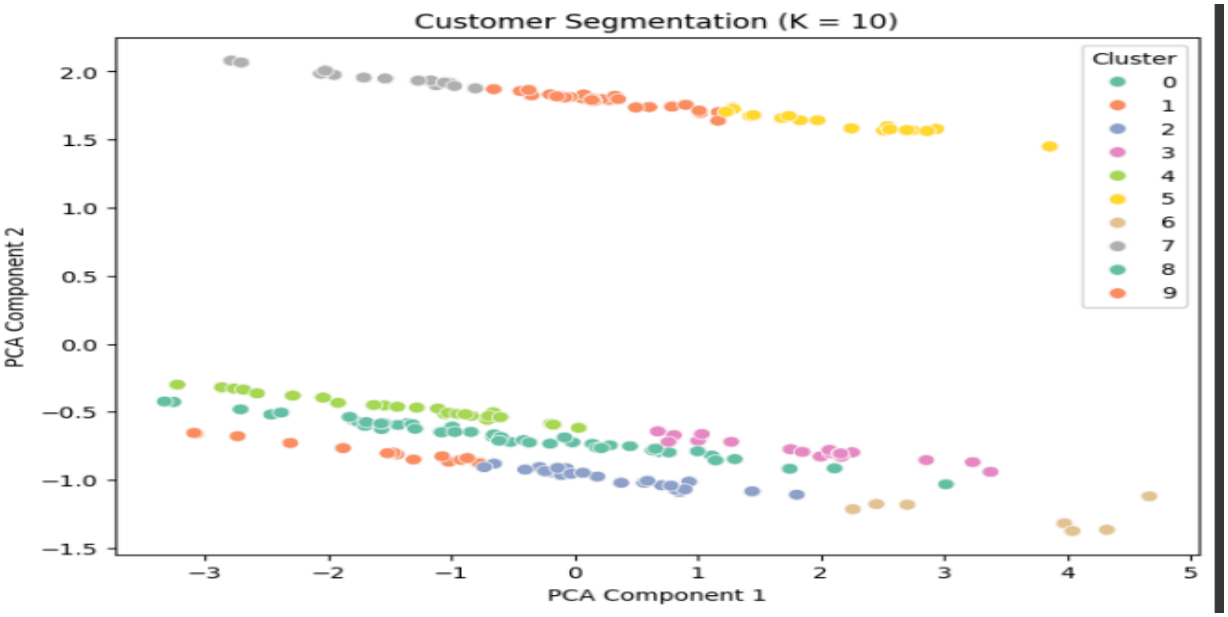
Visual Representation of Clusters

1. **Clustering Metrics**: A plot of the DB Index and Silhouette Score for each number of clusters is shown below, which helps visualize the clustering performance for different values of k.
   (This is just a placeholder for your plot)
2. **PCA Visualization**: A 2D scatter plot of the clusters after performing PCA (Principal Component Analysis) shows how the customers are distributed across the clusters. The plot visualizes the customer segments in the first two principal components of the data.
   (This is just a placeholder for your plot)

Conclusion

- **Optimal Clusters**: The optimal number of clusters for segmenting the customers was determined to be **10**.
- **DB Index**: The lowest DB Index (0.710) was achieved for both 9 and 10 clusters, suggesting that these clusters are well-separated.
- **Silhouette Score**: The Silhouette Score showed a slightly higher score for 6 clusters (0.501), but considering the DB Index, 10 clusters were chosen as optimal for the final model.
- The PCA visualization further confirms that the customer data is well-distributed across the chosen clusters.




Customer Segmentation (K = 10)
Cluster
0
1
2
3
4
5
6
7
8
9
PCA Component 2
PCA Component 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3
4
5