Task 3: Customer Segmentation / Clustering

Objective

The goal of this project was to perform customer segmentation using clustering techniques. By analyzing transaction and demographic data, customers were grouped into clusters to identify patterns and behaviors.

Data Preparation

1. **Datasets Used:**
   - **Customers Dataset:** Contains demographic details like CustomerID and Region.
   - **Transactions Dataset:** Includes transaction details such as TransactionID and TotalValue.
1. **Data Merging and Aggregation:**
   - The two datasets were merged on CustomerID.
   - Transactional data was aggregated to calculate:
     - **Total Spent:** Sum of transaction values per customer.
     - **Total Transactions:** Count of transactions per customer.
1. **Feature Engineering:**
   - The Region column was label-encoded to convert categorical data into numerical format.
   - Features (Region, Total Spent, and Total Transactions) were normalized using StandardScaler to ensure consistency.

Clustering Approach

1. **Algorithm Used:**
   - **KMeans Clustering**: A widely used partition-based clustering method.
1. **Evaluation Metrics:**
   - **Davies-Bouldin Index (DB Index):** Measures cluster compactness and separation. Lower values indicate better clustering quality.
   - **Silhouette Score:** Evaluates how similar an object is to its cluster compared to others. Higher values indicate better-defined clusters.
1. **Optimal Cluster Selection:**
   - A range of cluster counts (2 to 10) was evaluated.

- The optimal number of clusters was determined based on the **minimum Davies-Bouldin Index** value.

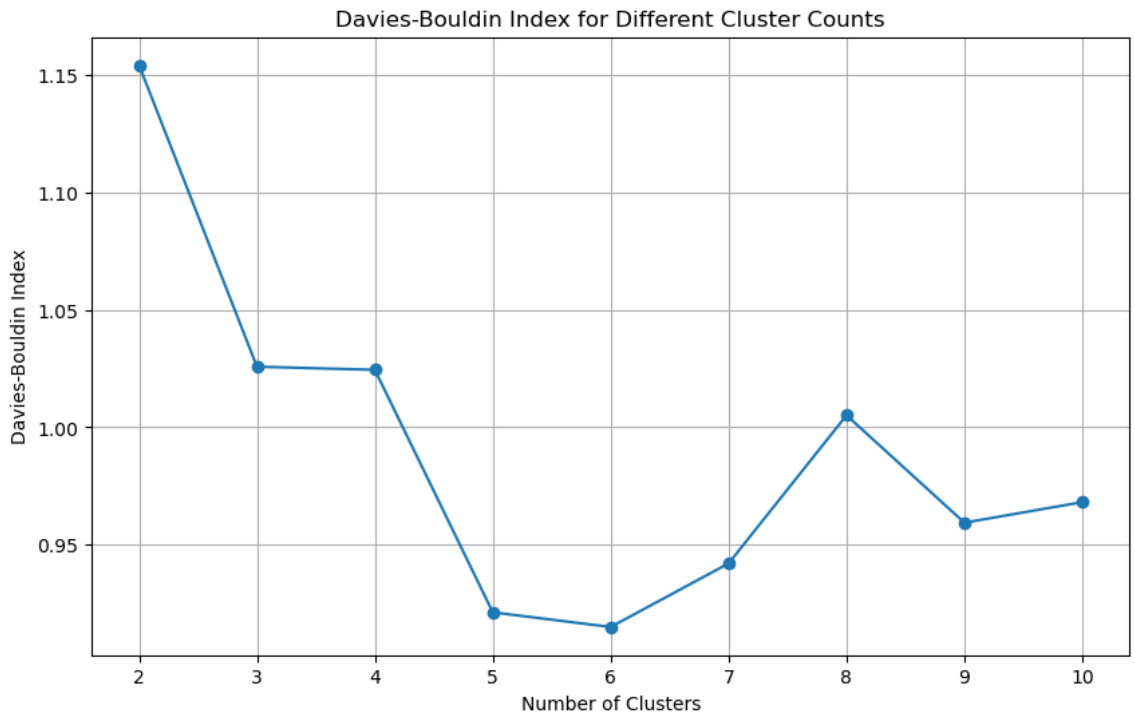
Clustering Results

1. **Optimal Number of Clusters:**
   - The optimal number of clusters was identified as **6**.
1. **Davies-Bouldin Index:**
   - The DB Index for the optimal clustering solution was **0.9149**, indicating high-quality clustering.
1. **Silhouette Score:**
   - The Silhouette Score for the optimal clustering was **0.3370**, suggesting moderate cluster separation and cohesion.

Visualizations

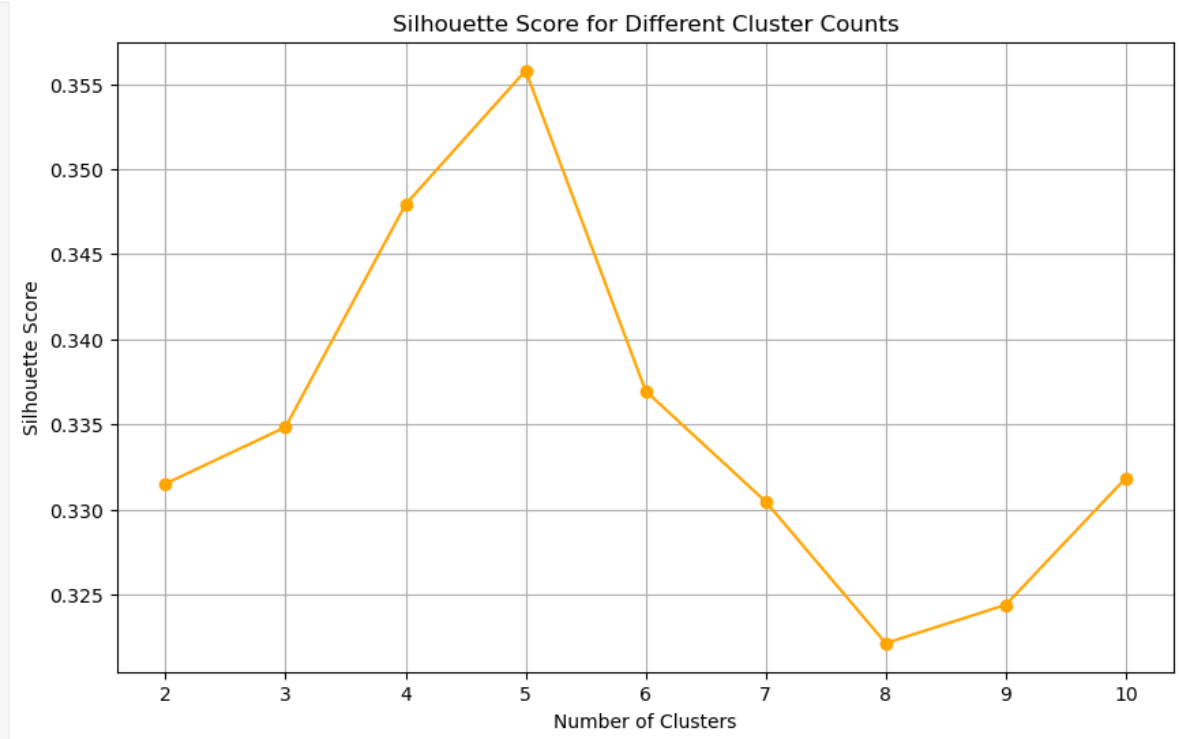
1. **Davies-Bouldin Index for Different Cluster Counts**

The DB Index was computed for cluster counts ranging from 2 to 10. The graph clearly shows that the index is minimized at **6 clusters**, making it the optimal choice.



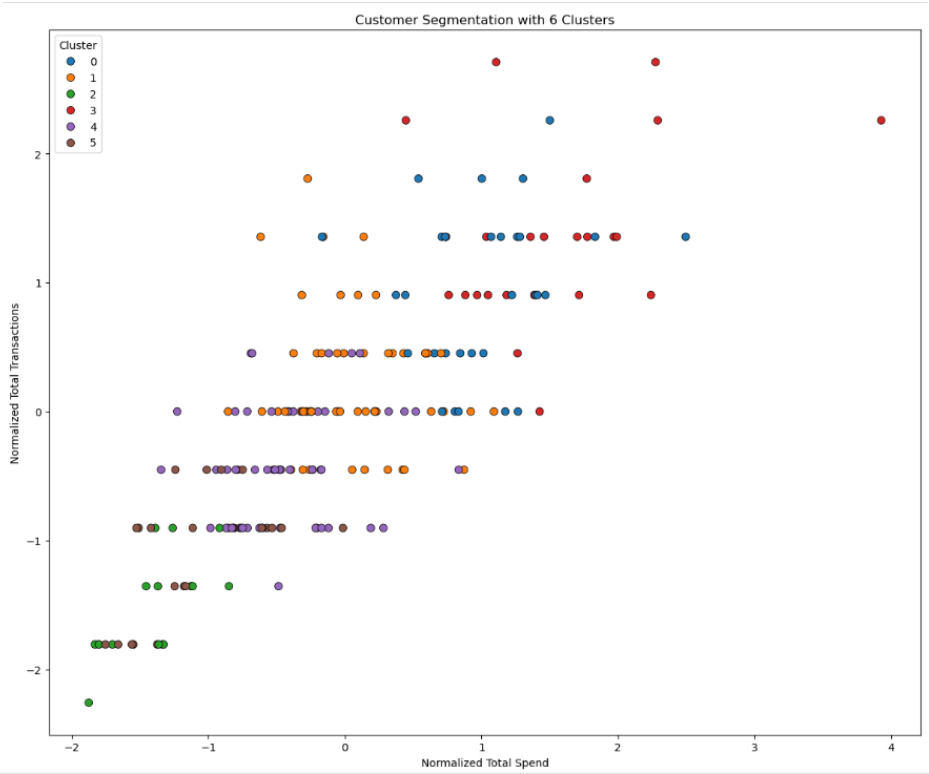
2. **Silhouette Score for Different Cluster Counts**

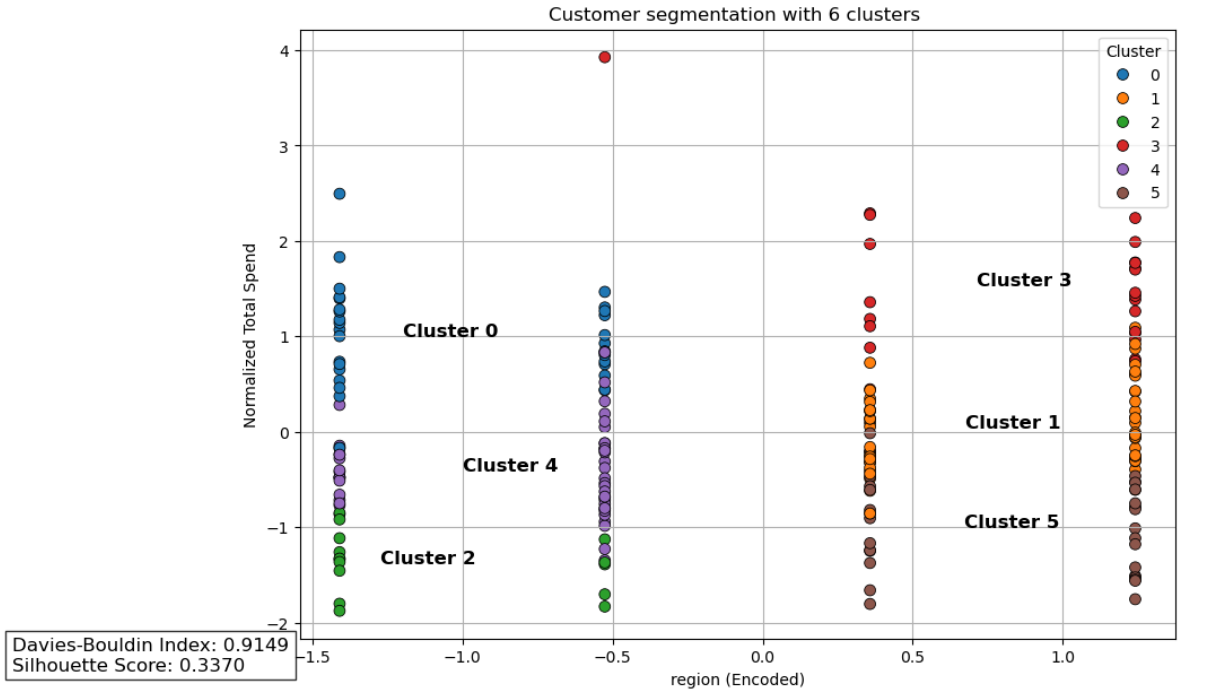
The silhouette scores were also computed for the same range. While the highest score was observed at 5 clusters, **6 clusters** provided a reasonable balance between silhouette score and DB Index.



3. **Cluster Visualization**

A scatter plot of the clusters was created using normalized features (Total Spent vs. Total Transactions). Each cluster is distinctly visualized using different colors. And also another graph plotting region vs Normalized Total Spend






```
Davies-Bouldin Index: 0.9149
Silhouette Score: 0.3370
```

Conclusion

- The customer segmentation analysis successfully grouped customers into **6 clusters** using KMeans.
- The clustering solution is robust, with a low **Davies-Bouldin Index** of **0.9149**.
- Moderate **Silhouette Score** (**0.3370**) indicates room for improvement in cluster cohesion.

Recommendations

- The segmented customer groups can be analyzed further to derive insights for targeted marketing strategies or customer retention programs.
- Additional features like customer demographics, frequency of transactions, or product preferences can be incorporated to enhance clustering quality.