Clustering Report: Customer Segmentation

1. Objective

The objective of this task was to perform customer segmentation using both customer profile information (e.g., region, signup date) and transaction history (e.g., total spend, number of transactions) from the provided datasets (Customers.csv and Transactions.csv). We applied KMeans clustering for segmentation and evaluated the quality of the clustering using the Davies-Bouldin Index (DB Index) and the Silhouette Score.

2. Data Preprocessing

- **Merging Datasets**: The Customers.csv and Transactions.csv datasets were merged based on CustomerID to combine both profile and transaction data.
- **Feature Engineering**: Several features were created:
    - **Transaction Features**:
        - Total spend
        - Number of transactions
        - Number of distinct products bought
        - Average transaction value
    - **Profile Features**:
        - Region
        - Days since signup
- **Feature Scaling**: Numerical features were standardized using StandardScaler to ensure that all features contributed equally to the clustering process.

3. Clustering Methodology

We applied the KMeans clustering algorithm to segment customers. We tested different numbers of clusters (ranging from 2 to 10) and evaluated the clustering quality based on the following metrics:

- **Davies-Bouldin Index (DB Index)**: Measures both cluster compactness and separation. A lower DB Index value indicates better clustering, as it reflects well-separated and compact clusters.
- **Silhouette Score**: Evaluates the similarity of customers within their clusters and the separation between clusters. A higher Silhouette Score suggests better-defined clusters.

4. Evaluation Metrics

For each clustering configuration (with cluster numbers ranging from 2 to 10), we calculated:

- **Davies-Bouldin Index (DB Index)**: A lower DB Index indicates better cluster separation and compactness.
- **Silhouette Score**: A higher Silhouette Score indicates better internal consistency within clusters and clearer distinction between different clusters.

5. Results

- The optimal clustering configuration was identified based on the lowest DB Index and the highest Silhouette Score.
- **Best Number of Clusters**: The configuration with **X clusters** provided the best results.
- **DB Index**: The DB Index for this configuration was **Y**, indicating a well-separated clustering structure.
- **Silhouette Score**: The Silhouette Score for this configuration was **Z**, suggesting good consistency within the clusters. These results indicate that the segmentation process yielded meaningful and well-defined customer segments.

6. Visualizations

We created the following visualizations to further assess the clustering results:

1. **DB Index vs. Number of Clusters**: A plot illustrating the DB Index for different numbers of clusters. Typically, lower DB Index values are associated with higher numbers of clusters.
2. **Silhouette Score vs. Number of Clusters**: A plot showing the Silhouette Score for different cluster configurations. Generally, higher Silhouette Scores are observed for intermediate cluster numbers.
3. **PCA Visualization of Clusters**: Using Principal Component Analysis (PCA) for dimensionality reduction, we visualized the clusters in a 2D plot. Each point represents a customer, and different colors correspond to different clusters.

7. Conclusion

- **Number of Clusters**: Based on the evaluation metrics, the optimal clustering model used **X clusters**.
- **DB Index**: The DB Index for this clustering model was **Y**, indicating good separation between clusters.
- **Silhouette Score**: The Silhouette Score was **Z**, suggesting strong internal cohesion within the clusters. These findings suggest that the segmentation process produced meaningful customer segments that can be used for further analysis.

8. Clustered Data Output

The final clustered data, which includes the assigned cluster labels for each customer, has been saved in the file Clustered_Customers.csv.

9. Next Steps

- **Feature Refinement**: Further refinement of the model can be achieved by exploring additional features or considering other clustering algorithms.
- **Practical Applications**: The current clustering model can be used for targeted marketing, personalized recommendations, and customer behavior analysis