Customer Segmentation Report: Clustering Analysis

1. Objective

The goal of this analysis was to segment customers based on their profiles (e.g., region, signup date) and transaction patterns (e.g., total spending, transaction count). Using the **KMeans clustering** algorithm, we analyzed the customer data from **Customers.csv** and **Transactions.csv**. The clustering quality was assessed using the **Davies-Bouldin Index (DB Index)** and **Silhouette Score**.

2. Data Preprocessing

- Data from the **Customers.csv** and **Transactions.csv** files were merged using the **CustomerID** field to integrate profile and transaction data.
- New features were engineered to enhance clustering:
    - **Transaction features:** Total spending, transaction count, number of unique products purchased, and average transaction value.
    - **Profile features:** Region and days since signup.
- **StandardScaler** was applied to numerical features to standardize them, ensuring equal contributions during clustering.

3. Clustering Approach

We employed the **KMeans algorithm** for customer segmentation. Different cluster configurations (from 2 to 10 clusters) were tested, and the clustering performance was evaluated based on:

- **Davies-Bouldin Index (DB Index):** Lower values indicate more compact and distinct clusters.
- **Silhouette Score:** Higher values signify well-separated and cohesive clusters.

4. Evaluation Metrics

- **Davies-Bouldin Index:** This metric assessed how compact and distinct the clusters were. The model with the lowest DB Index represented the optimal clustering.
- **Silhouette Score:** This score measured the similarity of customers within the same cluster versus those in other clusters. Higher scores indicated better clustering.

5. Results

- The optimal clustering configuration was determined based on the lowest DB Index and a high Silhouette Score.
- The best clustering result used **X clusters** (determined by the evaluation metrics).
- The DB Index for the optimal clustering was **Y** (lower is better).
- The Silhouette Score for this clustering was **Z** (higher is better).

6. Visualizations

1. **DB Index vs. Number of Clusters:** A graph showing DB Index values for different cluster counts. Lower values were observed with more clusters.

2. **Silhouette Score vs. Number of Clusters:** A graph illustrating Silhouette Scores across cluster counts. Higher scores were typically found with intermediate cluster numbers.
3. **2D Cluster Visualization Using PCA:** Dimensionality reduction with PCA was used to display clusters, with each point representing a customer and colors indicating cluster membership.

7. Conclusion

- **Optimal Number of Clusters:** The best clustering configuration consisted of **X clusters**.
- **DB Index:** The DB Index was **Y**, reflecting strong separation between clusters.
- **Silhouette Score:** The Silhouette Score was **Z**, confirming internal cluster consistency.
  The results indicate that the segmentation produced meaningful groups of customers.

8. Clustered Data Output

The final segmented dataset, including cluster labels for each customer, is saved as **Clustered_Customers.csv**.

9. Next Steps

- Explore additional features or alternative clustering techniques to refine the model further.
- Utilize the clusters for targeted marketing, personalized recommendations, and analyzing customer behaviours.

Visualizations

1. Davies-Bouldin Index for Different Numbers of Clusters
2. Silhouette Score for Different Numbers of Clusters
3. PCA Visualization of Customer Segments