Customer Segmentation Clustering Report

Objective:

The goal of this analysis is to segment customers based on both transaction behaviors and profile information. By applying clustering techniques, we can identify distinct customer groups that could assist in targeted marketing, customer retention strategies, and decision-making.

Data Overview:

- **Customer Data** (Customers.csv): Contains CustomerID, CustomerName, Region, and SignupDate.
- **Transaction Data** (Transactions.csv): Contains TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, and Price.

Clustering Approach:

We used **K-Means clustering** to create customer segments based on their transaction patterns and profile data. The features used for clustering were:

- **Transaction Features**: Total spent, total quantity purchased, average transaction value, number of transactions.
- **Profile Features**: Days since signup, and region (encoded as categorical variables).

Clustering Algorithm:

- **Algorithm**: K-Means Clustering
- **Number of Clusters**: We evaluated the performance of different cluster numbers (2 to 10) using the **Davies-Bouldin Index** and **Silhouette Score** to select the optimal number of clusters.
- **Optimal Number of Clusters**: Based on clustering metrics, **4 clusters** were found to be optimal.

Clustering Results:

1. **Number of Clusters**:
    - The final clustering solution formed **4 clusters.**
2. **Davies-Bouldin Index**:
    - The Davies-Bouldin Index measures the compactness and separation between clusters. A lower DB Index indicates better-defined clusters.
    - The final DB Index value was **1.36.** This indicates a reasonable separation between the clusters, though there is still room for improvement in defining the clusters more clearly.

3. **Silhouette Score**:
    - The **Silhouette Score** evaluates how well-separated the clusters are. A higher score suggests better-defined clusters.
    - The final Silhouette Score was **0.25**, which is relatively low. This suggests that the clusters are not as well-separated as they could be, and there might be some overlap between the groups.
4. **Other Clustering Metrics**:
    - The **Elbow Method** and **Silhouette Scores** were used to validate the choice of 4 clusters.
        - The **Elbow Method** showed a clear "elbow" at k=4, suggesting it as a reasonable choice.
        - While the **Silhouette Score** was not very high, it still indicated that the clustering performed reasonably well.

Cluster Visualization:

We used **Principal Component Analysis (PCA)** for dimensionality reduction to visualize the clusters in 2D space. The clusters were color-coded to highlight the segmentation.

Visualization Summary:

- The **PCA plot** showed that the clusters are somewhat distinct but have some overlap, which corresponds to the moderate DB Index and lower Silhouette Score.
- The clusters each exhibited unique transaction patterns, but there was some ambiguity between the groups, indicating potential areas for refinement in segmentation.

Conclusion:

- **Number of Clusters**: 4
- **Final Davies-Bouldin Index**: 1.36
- **Final Silhouette Score**: 0.25

The segmentation process identified **4 distinct customer groups.** Although the clusters are somewhat distinct, the relatively low Silhouette Score indicates that further improvements could be made, such as tuning features or exploring other clustering algorithms (e.g., DBSCAN or hierarchical clustering). The segmentation results are still valuable for understanding customer behaviors but may require refinement for more precise groupings.