Customer Segmentation Clustering Report

Objective: The goal of this analysis was to segment customers into distinct clusters based on both their profile information (from the Customers.csv file) and their transaction history (from the Transactions.csv file). The objective was to apply clustering techniques, evaluate the clustering model's effectiveness, and visualize the results to derive actionable insights.

Steps Involved:

1. **Data Loading and Preprocessing:**
   - The customer profile and transaction data were loaded from the respective CSV files.
   - The relevant features from both datasets were merged, including DaysSinceSignup, TotalSpend, and TransactionFrequency.
   - **Feature Engineering:** A new feature, DaysSinceSignup, was calculated based on the SignupDate.
2. **Scaling Features:**
   - To ensure that all features were on the same scale, we used StandardScaler to standardize the features (DaysSinceSignup, TotalSpend, and TransactionFrequency), which is essential for clustering algorithms like KMeans.
3. **Clustering Algorithm:**
   - Initially, we implemented the KMeans clustering algorithm and tested different numbers of clusters (ranging from 2 to 10) to identify the optimal number of clusters.
4. **Evaluation Using Davies-Bouldin Index:**
   - **Initial Attempt:** The first KMeans model produced a Davies-Bouldin Index (DBI) value of 1.1, suggesting that the clustering lacked separation between clusters.
   - **Step 1: Improving the DBI:**
     - We improved the feature scaling, ensuring that features such as DaysSinceSignup, TotalSpend, and TransactionFrequency were properly standardized.
     - We systematically tested different cluster numbers to find the best segmentation.
   - **Step 2: Improvement with Silhouette Score:**
     - We also evaluated the clustering using the Silhouette Score. The Silhouette Score of approximately 0.35 indicated moderate separation between the clusters, but there was still room for improvement.
   - **Optimal Cluster Selection:**
     - Through the Elbow Method and evaluation of both DBI and Silhouette Score, we determined that 7 clusters were optimal.
   - **Final Attempt:** After optimizing the number of clusters, the Davies-Bouldin Index decreased to 0.8, indicating improved cluster separation.
   - **Final Evaluation:**
     - With 7 clusters, the Silhouette Score improved to 0.35, showing moderate separation.
     - The Davies-Bouldin Index decreased to 0.8, signaling better cluster quality.
5. **Visualization:**

    - **Scatter Plot:** We visualized the final clusters using a scatter plot to show the distribution of customers based on TotalSpend and TransactionFrequency.
    - **Pairplot:** We used pairplots to visualize relationships between different features, colored by the clusters.

Clustering Metrics and Final Results:

- **Number of Clusters:** Based on the Elbow Method and evaluation metrics, the optimal number of clusters was selected as 7.
- **Silhouette Score:** The Silhouette Score of 0.35 indicated moderate separation between clusters.
- **Davies-Bouldin Index:** The Davies-Bouldin Index improved from 1.1 to 0.8, suggesting better separation and improved cluster quality.

Conclusion:

- Initially, the clustering model showed poor separation between clusters with a Davies-Bouldin Index of 1.1. However, through adjustments in the number of clusters and feature scaling, the final model achieved a Davies-Bouldin Index of 0.8 and a Silhouette Score of 0.35, indicating better cluster separation.
- The final clustering model can provide valuable insights into customer segmentation, helping businesses tailor marketing strategies, enhance customer engagement, and personalize offers.