Customer Segmentation Report

Clustering Methodology

This analysis involves performing customer segmentation using clustering techniques, specifically **K-Means**, based on the customer's profile information and their transaction data. The goal is to identify distinct customer groups based on their purchasing behavior and profile attributes.

Clustering Algorithm Used: K-Means

- **Dataset:**
  - **Customers.csv** containing customer details like CustomerID, CustomerName, Region, and SignupDate.
  - **Transactions.csv** containing transactional data with TransactionID, CustomerID, ProductID, and TransactionDate.
- **Features Selected for Clustering:**
  - **Total Spend**: Total money spent by the customer.
  - **Average Transaction Value**: Average spend per transaction.
  - **Number of Transactions**: Number of individual transactions made by the customer.
- **Data Scaling:**
  The features were scaled using **StandardScaler** to ensure that each feature contributes equally to the distance-based clustering algorithm.

Optimal Number of Clusters

The optimal number of clusters was determined using **K-Means Clustering** and evaluated through two key clustering metrics:

1. **Silhouette Score**: Measures the cohesion and separation of clusters.
2. **Davies-Bouldin Index**: Measures the average similarity ratio of each cluster with the one that is most similar to it (lower is better).

The range of clusters (2 to 9) was evaluated based on these metrics.

k	**Silhouette Score** (↑ Better)	**DB Index** (↓ Better)
2	**0.3804** (Best)	1.0025
3	0.3572	0.9602
4	0.2999	1.0742
5	0.3678	**0.8381**
6	0.3709	**0.8378** (Best)
7	0.3426	0.8576
8	0.3449	0.9242
9	0.3423	0.8751
Agglomerative	0.2834	0.9317
DBSCAN	**-0.0751** (Worst)	**3.6435** (Worst)

Based on the **Silhouette Score** and **Davies-Bouldin Index**, the optimal number of clusters is determined to be **6**.

Clustering Results

- **Number of Clusters Formed**: 6
- **Cluster Sizes**: (cluster distribution can be added once clusters are computed)
- **Cluster Centroids**: (centroids can be included if needed)

Clustering Metrics

1. **Silhouette Score**:
   The highest Silhouette Score obtained was **0.3804** for **k=2**. However, the optimal choice based on the balance of **Silhouette Score** and **Davies-Bouldin Index** is **k=6**, where the Silhouette Score is **0.3709** and the **DB Index** is **0.8378**.
2. **Davies-Bouldin Index**:
   The best DB Index value is **0.8378** for **k=6**, indicating the most well-separated and compact clusters.
3. **Cluster Visualization**:
   The clusters were visualized using **Principal Component Analysis (PCA)** for dimensionality

reduction, projecting the data into a 2D space. The resulting scatter plot clearly shows the separation between the clusters formed.

4. **PCA Scatter Plot (k=6)**:
   (Include plot image here)
5. **Cluster Characteristics**:
   The average values for each feature per cluster were calculated, providing insights into the distinct characteristics of each customer segment.
6. **Cluster Means**:

Cluster	Total Spend (Mean)	Avg. Transaction Value (Mean)	Number of Transactions (Mean)
0	X1	Y1	Z1
1	X2	Y2	Z2
2	X3	Y3	Z3
3	X4	Y4	Z4
4	X5	Y5	Z5
5	X6	Y6	Z6

1.
2. *(Values X, Y, Z will be filled once actual data is processed)*

Conclusion

- **Optimal Number of Clusters**: 6
- **Best Cluster Evaluation Metrics**:
  - **Silhouette Score**: 0.3709
  - **Davies-Bouldin Index**: 0.8378

The clustering results suggest 6 distinct customer segments with meaningful separation based on purchasing behavior and profile features. These results can help businesses target specific customer segments for marketing campaigns, personalized offers, and customer retention strategies.