Task 3: Customer Segmentation / Clustering

1. Objective

The objective of this task was to segment customers 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`).

The segmentation was performed using the KMeans clustering algorithm, with clustering quality evaluated through two metrics:

- **Davies-Bouldin Index (DB Index)**
- **Silhouette Score**

---

2. Data Preprocessing

To prepare the data for clustering, the following steps were performed:

- **Dataset Merging**:
  Merged `Customers.csv` and `Transactions.csv` based on `CustomerID` to combine customer profile and transaction history.
- **Feature Engineering**:
  - **Transaction Features**:
    - Total spend (`total_spent`)
    - Number of transactions (`num_transactions`)
    - Number of distinct products bought (`num_products_bought`)
    - Average transaction value (`avg_transaction_value`)
  - **Profile Features**:
    - Days since signup (`days_since_signup`)
- **Data Standardization**:
  Numerical features were standardized using `StandardScaler` to ensure all features contributed equally to the clustering process.

---

3. Clustering Methodology

- **Algorithm Used**: KMeans
- **Cluster Range**: Models were evaluated for clusters ranging from 2 to 10.
- **Evaluation Metrics**:
  - **Davies-Bouldin Index (DB Index)**:
    Indicates the compactness and separation of clusters. A **lower DB Index** represents better clustering.

- **Silhouette Score**:
  Measures how similar customers are within clusters compared to other clusters. A **higher Silhouette Score** indicates better clustering.

---

4. Evaluation Metrics

- **Davies-Bouldin Index (DB Index)**:
  - Calculated for each cluster configuration.
  - The clustering configuration with the lowest DB Index value was considered the best, as it indicates well-separated and compact clusters.
- **Silhouette Score**:
  - Provided insights into the similarity within clusters and distinctness between clusters.
  - Higher scores indicated more coherent clustering.

---

5. Results

- **Optimal Number of Clusters**:
  The best clustering model used **10 clusters**, based on evaluation metrics.
- **Metrics**:
  - **Davies-Bouldin Index**: 1.0740
  - **Silhouette Score**: 0.2799

These metrics indicate reasonable cluster separation and internal consistency, although the Silhouette Score suggests potential overlap for certain clusters.

---

6. Visualizations

1. **DB Index vs. Number of Clusters**:
   A line plot showing the DB Index across different cluster numbers.
   - **Observation**: Lower DB Index values were observed with higher numbers of clusters, indicating better separation.
2. **Silhouette Score vs. Number of Clusters**:
   A line plot illustrating the Silhouette Score across different cluster numbers.
   - **Observation**: Higher scores were generally observed with intermediate cluster numbers.
3. **PCA Visualization of Clusters**:
   Using PCA, for dimensionality reduction, we visualized the clusters in 2D, where each

point represents a customer, and different colors correspond to different clusters.

---

7. Conclusion

- **Number of Clusters**: The best clustering model used **10 clusters**.
- **Davies-Bouldin Index**: 1.0740 (good cluster separation).
- **Silhouette Score**: 0.2799 (moderate internal consistency).

The segmentation revealed meaningful customer groups with distinct behaviors, enabling actionable insights.

---

8. Clustered Data Output

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

---

9. Next Steps

- **Refinements**:
  Explore additional features or alternative clustering algorithms (e.g., DBSCAN, Agglomerative Clustering).
- **Applications**:
  - Targeted marketing campaigns.
  - Personalized product recommendations.
  - Behavior analysis for retention strategies.