Clustering Analysis Report

The report is divided into two parts: clustering performed without PCA (Principal Component Analysis) and clustering performed after dimensionality reduction using PCA.

Part 1: Clustering Without PCA

1. **Data Preparation:**
   - Customer and transaction data were merged using CustomerID as the key.
   - Relevant features such as total spending (TotalValue), transaction count (TransactionID), and customer tenure (SignupDuration) were aggregated at the customer level to create a feature set for clustering.
2. **Feature Scaling:**
   - The features were standardized using StandardScaler to ensure that all variables contribute equally to the clustering process, as they originally had different scales.
3. **Clustering:**
   - K-Means clustering was applied to the standardized features, varying the number of clusters (kk) from 2 to 10. The algorithm iteratively assigned customers to clusters by minimizing the within-cluster variance.
4. **Evaluation Metrics:**
   - **Davies-Bouldin Index (DB Index):** Measures the ratio of within-cluster dispersion to between-cluster separation. Lower values indicate better clustering.
   - **Silhouette Score:** Assesses the cohesion of points within the same cluster and separation between clusters. Higher values indicate better clustering.
   - These metrics were computed for each value of kk, and the optimal number of clusters was selected based on the lowest DB Index.

Results

1. **Optimal Number of Clusters:**
   - Based on the DB Index, the optimal number of clusters is **4**.
2. **Metrics:**
   - **DB Index for Optimal Clusters:** 0.8822
   - **Silhouette Score for Optimal Clusters:** 0.3501
   - **DB Index Scores for Different kk:** [1.1653,0.9976,0.8822,0.9694,0.9410,1.0600,1.0764,1.0336,1.0738][1.1653, 0.9976, 0.8822, 0.9694, 0.9410, 1.0600, 1.0764, 1.0336, 1.0738]
   - **Silhouette Scores for Different kk:** [0.3343,0.3388,0.3501,0.3135,0.3188,0.2909,0.2947,0.3007,0.2829][0.3343, 0.3388, 0.3501, 0.3135, 0.3188, 0.2909, 0.2947, 0.3007, 0.2829]

Key Observations

- At k=4k=4, both the DB Index and Silhouette Score indicate well-defined clusters.
- Clustering was performed on the original feature set without reducing its dimensionality, retaining all original information.

Part 2: Clustering After PCA

1. **Dimensionality Reduction:**
   - Principal Component Analysis (PCA) was used to reduce the original feature space to 2 dimensions while preserving most of the variance. This step is useful for simplifying the dataset and eliminating potential multicollinearity among features.
2. **Feature Scaling:**
   - Before applying PCA, features were standardized using StandardScaler, ensuring equal contribution of variables to the PCA transformation.
3. **PCA Transformation:**
   - The first two principal components were retained, accounting for the majority of the variance in the data. These components served as the input for clustering.
4. **Clustering:**
   - K-Means clustering was applied to the PCA-reduced dataset, varying the number of clusters (kk) from 2 to 10.
   - The Davies-Bouldin Index (DB Index) was computed for each kk, and the optimal number of clusters was selected based on the lowest DB Index.

Results

1. **Optimal Number of Clusters:**
   - Based on the DB Index, the optimal number of clusters is **7**.
2. **Metrics:**
   - **DB Index for Optimal Clusters:** 0.7595
   - **DB Index Scores for Different kk:** [1.1034,0.9932,0.8734,0.8631,0.8195,0.7595,0.7813,0.8121,0.8456][1.1034, 0.9932, 0.8734, 0.8631, 0.8195, 0.7595, 0.7813, 0.8121, 0.8456]

Key Observations

- PCA-based clustering resulted in a better DB Index (0.7595 vs. 0.8822) compared to clustering without PCA, suggesting that dimensionality reduction improved cluster compactness and separation.
- However, reducing the data to 2 dimensions may have sacrificed some information, which could impact interpretability in domain-specific contexts.

Comparison and Conclusion

1. **Number of Clusters:**
   - Without PCA: k=4k=4
   - With PCA: k=7k=7
2. **DB Index:**
   - Without PCA: 0.8822
   - With PCA: 0.7595 (improved clustering quality)
3. **Silhouette Score (without PCA):**
   - Indicates moderate cluster cohesion and separation at k=4k=4.
4. **Insights:**

- Dimensionality reduction using PCA improved clustering quality, as evidenced by the lower DB Index. However, for interpretability, clustering without PCA retains the full feature space, which might be more useful in domain-specific applications.

Deliverables

- The detailed clustering process, evaluation metrics, and visualizations are documented and available in the notebook.
- Outputs include:
  - Optimal cluster numbers (kk) for both approaches.
  - Metrics like DB Index and Silhouette Score.
  - Cluster visualizations for both approaches.