Customer Segmentation Clustering Report

Objective

The primary goal of this task was to perform customer segmentation using clustering techniques. This involved utilizing both profile information from the `Customers.csv` file and transaction information from the `Transactions.csv` file. The clustering results aimed to group customers into distinct clusters based on their behavioral and transactional patterns.

Methodology

Data Preparation

1. **Dataset Integration**:
   - The `Customers.csv` and `Transactions.csv` datasets were merged to create a consolidated dataset.
   - Key features such as `total_spent`, `total_transactions`, and `avg_transaction_value` were computed for each customer.
2. **Feature Engineering**:
   - Categorical variables such as `Region` were one-hot encoded to facilitate clustering.
   - Missing values in the transaction data were filled with `0` for customers without transactions.
3. **Feature Scaling**:
   - All numerical features were standardized using `StandardScaler` to ensure that each feature contributed equally to the clustering process.

Clustering Technique

- **K-Means Clustering** was selected as the clustering algorithm due to its simplicity and effectiveness in creating compact, well-separated clusters.
- The **Elbow Method** was employed to determine the optimal number of clusters by analyzing the Davies-Bouldin Index and inertia values.

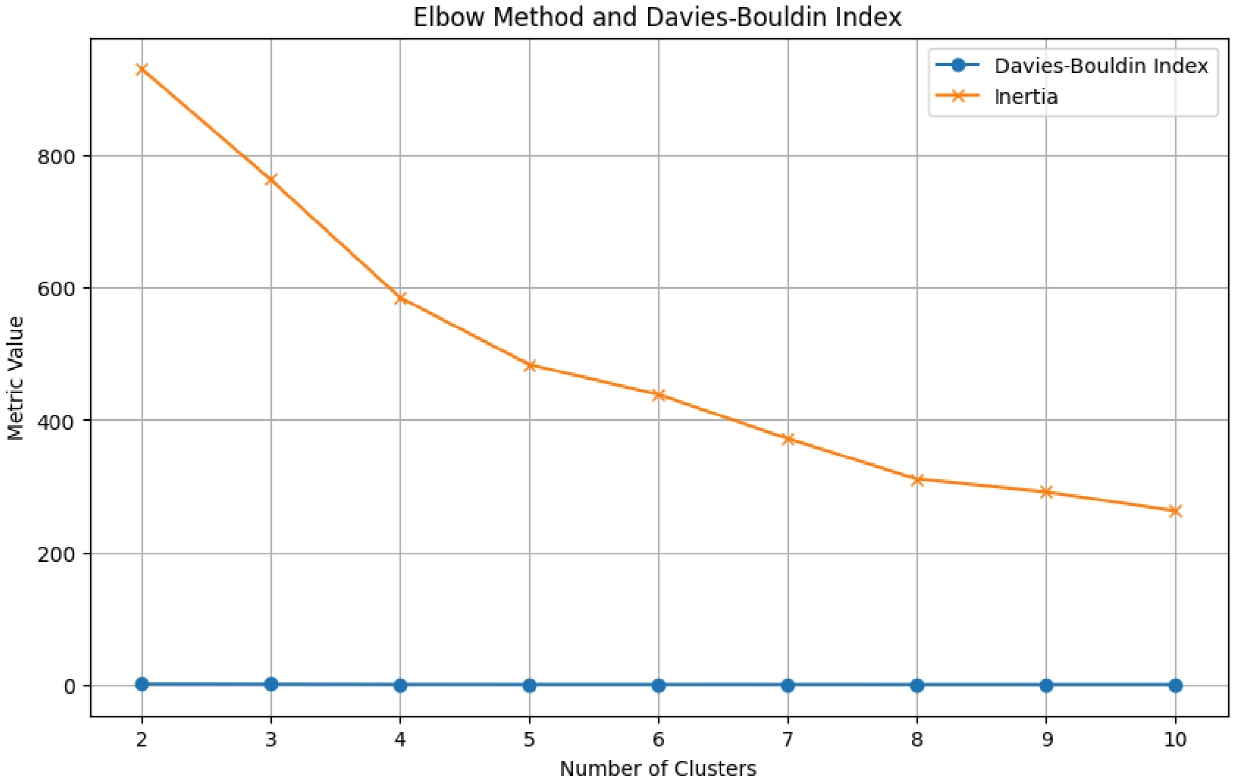

Figure: Plot of Elbow Method and DB Index

Metrics Evaluation

- **Davies-Bouldin Index**: Measures the compactness and separation of clusters (lower is better).
- **Silhouette Score**: Measures the similarity of points within clusters relative to other clusters (higher is better).
- **Calinski-Harabasz Score**: Evaluates cluster dispersion and separation (higher is better).

Results

Number of Clusters Formed

- The optimal number of clusters determined was **4**, as derived from the elbow plot and Davies-Bouldin Index evaluation.

Clustering Metrics

- **Davies-Bouldin Index**: **1.1926**
  - This relatively low value indicates compact and well-separated clusters.
- **Silhouette Score**: **0.3197**
  - A moderate score, suggesting room for improvement in the separation between clusters.
- **Calinski-Harabasz Score**: **68.6363**
  - This value reflects moderate dispersion and separation of the clusters.

Visualization

1. **Elbow Method Plot**:
   - The elbow plot showed a steady decrease in Davies-Bouldin Index and inertia, with an optimal cluster number of **4**.
2. **PCA-Based Scatter Plot**:
   - A 2D scatter plot using Principal Component Analysis (PCA) demonstrated the distinct separation of the 4 clusters.

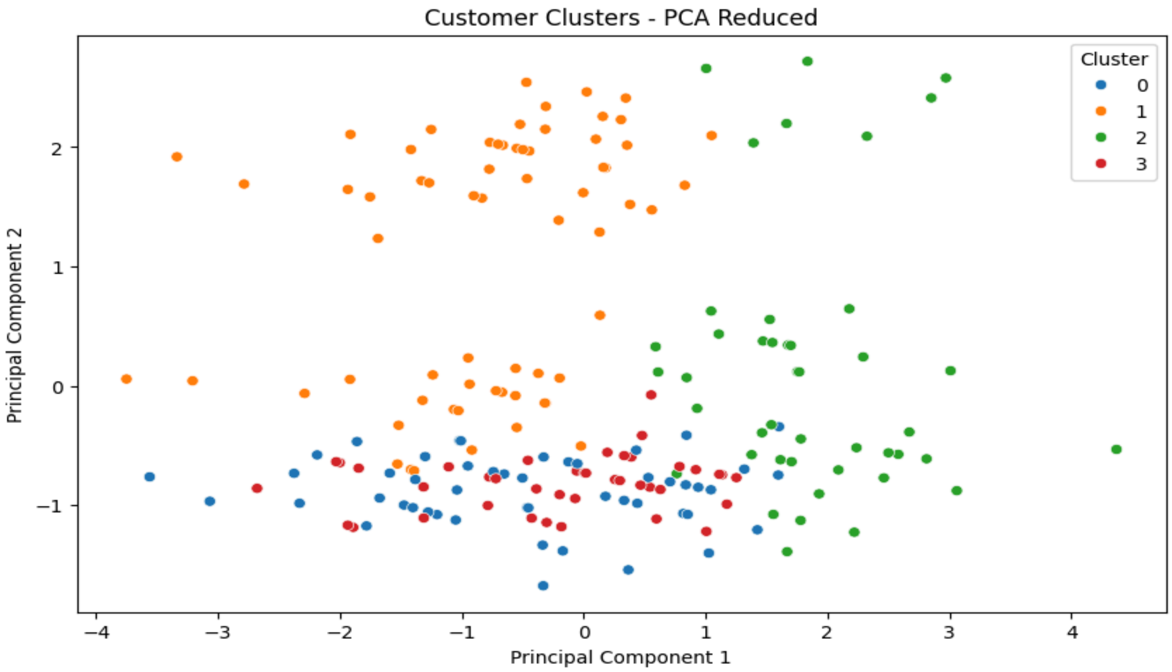

Figure: Plot of Customer Cluster

Process Explanation

1. The datasets were preprocessed by merging customer profiles with transaction data.
2. Features were scaled to ensure equal contribution to clustering.
3. The elbow method was used to determine the optimal cluster count, balancing compactness and separation of clusters.
4. K-Means clustering was performed with **4 clusters**, achieving the best Davies-Bouldin Index.
5. Clustering results were visualized using PCA for interpretation.

Conclusion

The customer segmentation process successfully identified 4 clusters with:

- A **Davies-Bouldin Index** of **1.1926430643192663**, indicating compact and well-separated clusters.
- Moderate **Silhouette Score** and **Calinski-Harabasz Score**, suggesting reasonable cluster quality.

Further refinement could involve exploring alternative clustering algorithms or additional feature engineering to improve the separation between clusters.