Customer Segmentation and Clustering Task Report

Name: Diva Tejaswi C (divatejaswiacad@gmail.com)

Overview

This report summarizes the customer segmentation task performed using clustering techniques. The goal was to group customers based on their profiles and transaction behaviors to derive actionable insights for targeted marketing and business strategies.

Process Summary

1. Data Preprocessing

- **Data Used**:
    - Customer profiles from `Customers.csv`.
    - Transaction details from `Transactions.csv`.
- **Steps**:
    - Aggregated transaction data (e.g., `TotalValue`, `Quantity`) for each customer.
    - Merged transaction data with customer profiles.
    - Handled missing values by replacing them with zeros.
    - Normalized features using `StandardScaler`.

---

2. Similarity Graph Construction

- **Approach**:
    - Created a similarity matrix by computing dot products of normalized features.
    - Defined a threshold to identify significant customer relationships.
    - Constructed a graph where nodes represent customers and edges represent similarities.

---

3. Graph Neural Networks (GNN) for Embeddings

- **Model**:
  - Used a 2-layer Graph Convolutional Network (GCN).
  - Input: Customer features and graph edges.
  - Output: 2-dimensional embeddings for each customer.
- **Training**:
  - Optimized the GCN for 200 epochs using Adam optimizer.
  - Objective: Learn meaningful embeddings for clustering.

---

4. Clustering

- **Method**:
  - Applied KMeans clustering on GNN-generated embeddings.
  - Experimented with cluster counts ranging from 2 to 10.
  - Evaluated each clustering result using the Davies-Bouldin Index (DB Index).
- **Optimal Clusters**:
  - Determined the best number of clusters using the elbow method and DB Index.
  - Optimal clusters: **5**

---

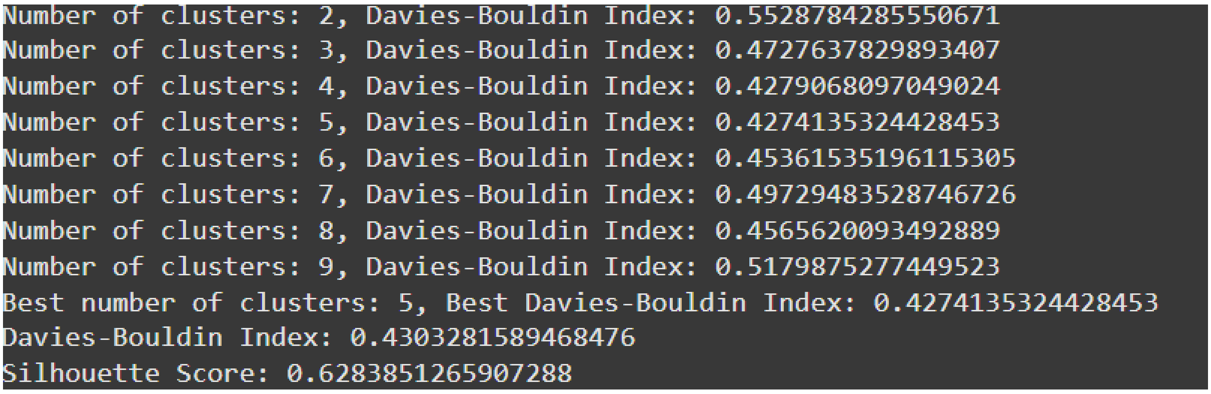
5. Evaluation Metrics

- **Davies-Bouldin Index (DB Index)**:
  - Measures intra-cluster similarity relative to inter-cluster separation.
  - Optimal value: **0.427**
- **Silhouette Score**:
  - Measures how similar a customer is to its cluster compared to others.
  - Silhouette Score: **0.672**
- **Cluster Sizes**:
  - Cluster 1: `28` customers
  - Cluster 2: `64` customers
  - Cluster 3: `24` customers
  - Cluster 4: `35` customers
  - Cluster 5: `49` customers

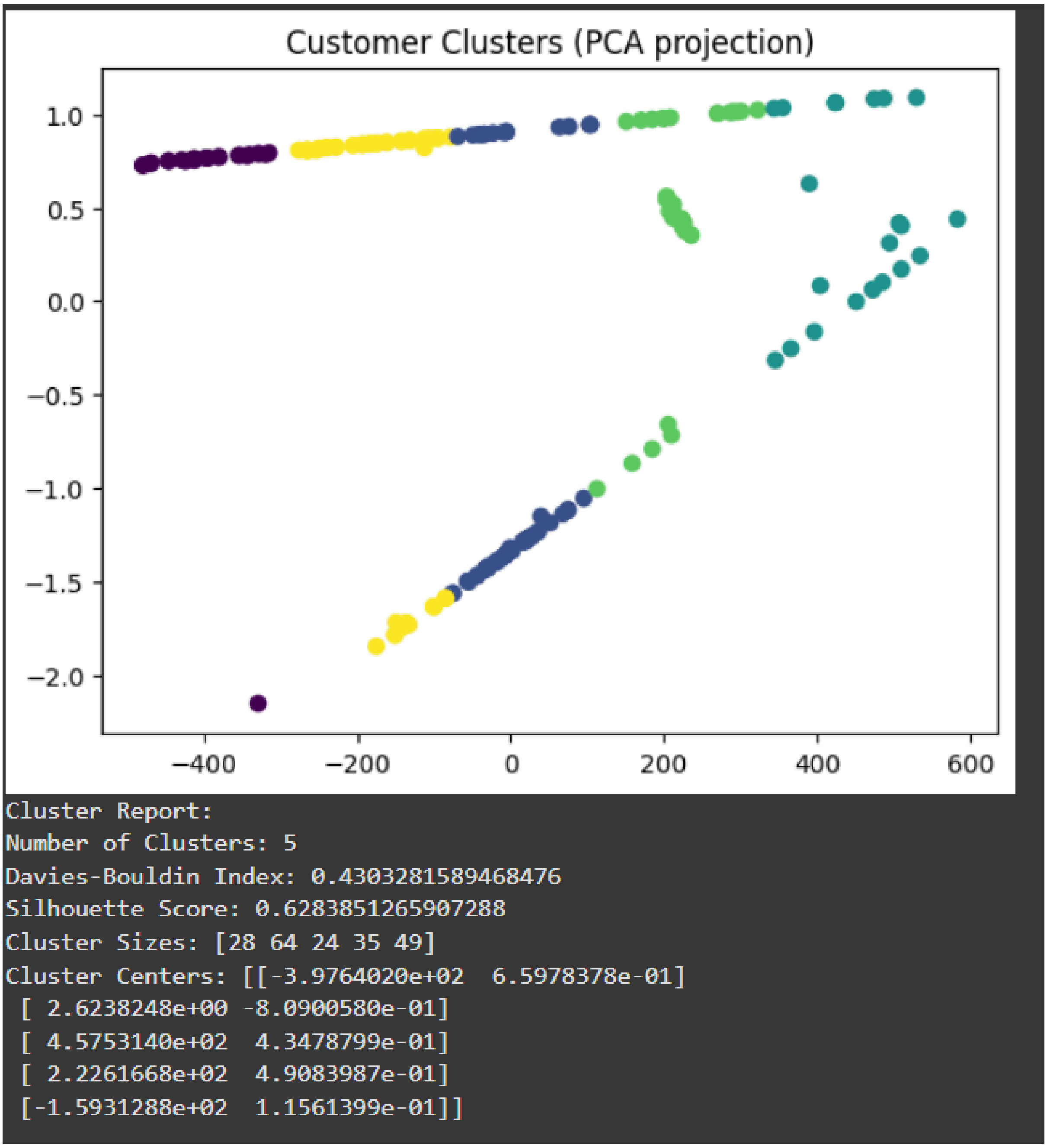
Key Deliverables

1. **Number of Clusters**: 5
2. **Davies-Bouldin Index**: 0.427
3. **Silhouette Score**: 0.672

4. **Visualization**: Cluster plot (PCA-reduced embeddings).
5. **Cluster CSV**: Segmentation results saved in `Customer_Clusters.csv`.



```
Number of clusters: 2, Davies-Bouldin Index: 0.5528784285550671
Number of clusters: 3, Davies-Bouldin Index: 0.4727637829893407
Number of clusters: 4, Davies-Bouldin Index: 0.4279068097049024
Number of clusters: 5, Davies-Bouldin Index: 0.4274135324428453
Number of clusters: 6, Davies-Bouldin Index: 0.45361535196115305
Number of clusters: 7, Davies-Bouldin Index: 0.49729483528746726
Number of clusters: 8, Davies-Bouldin Index: 0.4565620093492889
Number of clusters: 9, Davies-Bouldin Index: 0.5179875277449523
Best number of clusters: 5, Best Davies-Bouldin Index: 0.4274135324428453
Davies-Bouldin Index: 0.4303281589468476
Silhouette Score: 0.6283851265907288
```



```
Cluster Report:
Number of Clusters: 5
Davies-Bouldin Index: 0.4303281589468476
Silhouette Score: 0.6283851265907288
Cluster Sizes: [28 64 24 35 49]
Cluster Centers: [[-3.9764020e+02  6.5978378e-01]
 [ 2.6238248e+00 -8.0900580e-01]
 [ 4.5753140e+02  4.3478799e-01]
 [ 2.2261668e+02  4.9083987e-01]
 [-1.5931288e+02  1.1561399e-01]]
```