Clustering Report

1 Overview

I performed **customer segmentation** by combining **customer profile** data (Customers.csv) with **transaction** data (Transactions.csv). The combined dataset was transformed into a feature matrix capturing each customer's total spending in various categories, total number of purchases, and days since signup (among other features).

2 Methodology

- **Clustering Algorithm** : K-Means (from 2 to 10 clusters)
- **Feature Scaling** : Used StandardScaler to normalize the feature distributions.
- **Evaluation Metrics** :
    - **Davies-Bouldin (DB) Index** – lower values indicate better cluster separation.
    - **Silhouette Score** – higher values indicate more distinct clusters.

3 Results

1. **Davies-Bouldin Index vs. k**
    - The line chart shows that the DB Index is highest (~2.2) at k=2–3, drops significantly around k=5 (~1.4–1.5), and then fluctuates slightly for k > 5.
2. **Silhouette Score vs. k**
    - The silhouette score peaks at k=5 (about 0.26), suggesting relatively better cohesion and separation of clusters.
3. **Number of Clusters Chosen**
    - Based on both the **lowest DB Index** and a reasonably **high Silhouette Score**, **k=5** was selected as the final number of clusters.
4. **DB Index Value**
    - With **k=5**, the final Davies-Bouldin Index is around **1.50** (exact value may vary slightly by random seed).
5. **Other Clustering Metrics**
    - Silhouette Score for the chosen solution is around **0.26**, indicating moderate separation among clusters.
6. **Cluster Visualization (PCA Projection)**
    - A 2D scatter plot using Principal Component Analysis (PCA) reveals how the five clusters are distributed in a reduced-dimensional space.
    - Though PCA compresses high-dimensional data, the clusters show discernible groupings with partial overlap in some regions.

4 Conclusion and Potential Business Actions

- **Cluster 0** : May consist of customers who are moderate-to-high spenders in certain categories.
- **Cluster 1** : Possibly includes customers who favor different product lines or have lower overall spend.

- **Cluster 2, 3, 4** : Each grouping might have unique patterns in category preference, purchase frequency, or region distribution.

A deeper **profiling** of each cluster (e.g., average spend in Electronics, average days since signup) can guide **tailored marketing campaigns** and **personalized recommendations**.