1. Understanding the Problem

We aim to segment customers based on:

- **Profile information** (e.g., region from Customers.csv).
- **Transaction patterns** (RFM: Recency, Frequency, Monetary from Transactions.csv).

We'll:

1. Create a feature-rich dataset combining customer profile and transaction data.
2. Apply clustering techniques.
3. Evaluate the clusters using metrics like the Davies-Bouldin Index (DBI) and Silhouette Score.
4. Visualize clusters using dimensionality reduction (e.g., PCA).

---

2. Clustering Logic

The following detailed steps form the clustering logic:

Step 1: Data Preparation

1. **Feature Extraction:**
    - From Customers.csv: Include Region (one-hot encoded) and any relevant profile details.
    - From Transactions.csv: Compute RFM metrics:
        - **Recency:** Days since the last transaction.
        - **Frequency:** Number of transactions.
        - **Monetary:** Total spending.
2. **Feature Scaling:**
    - Normalize numerical features (RFM) to standardize the scale using StandardScaler.
3. **Handle Missing Values:**
    - Use imputation (e.g., SimpleImputer) for any missing data in profile or transaction information.

Step 2: Clustering Algorithm

1. **KMeans Clustering:**

- Use the KMeans algorithm for clustering due to its simplicity and interpretability.
- Train models for clusters ranging from **2 to 10**.

2. **Evaluate Optimal Clusters:**
    - Use **Davies-Bouldin Index (DBI):**
        - Lower values indicate better clustering.
    - Use **Silhouette Score:**
        - Higher values indicate distinct and well-defined clusters.

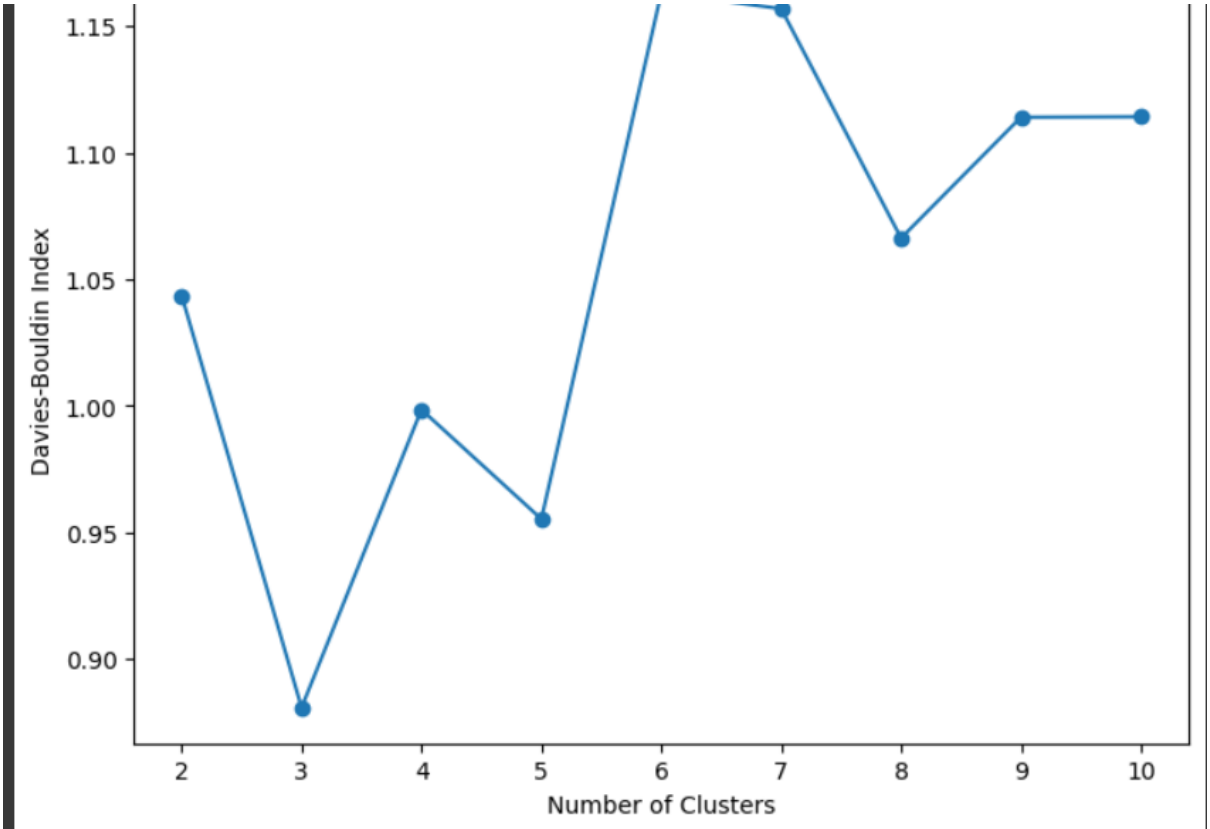
Step 3: Dimensionality Reduction

1. **PCA (Principal Component Analysis):**
    - Reduce dimensions to 2 for visualization.
    - Visualize clusters in a scatter plot.

Step 4: Cluster Profiling

1. Analyze clusters to understand characteristics:
    - What regions dominate each cluster?
    - How do RFM metrics vary across clusters?
    - What product categories or price ranges are associated with specific clusters?


Customer Segments (K=3)
Cluster
0
1
2
PCA Component 2
PCA Component 1
2
1
0
−1
−2
−4
−2
0
2
4



1. Number of Clusters Formed

- **Optimal Number of Clusters:** 3 (determined from the Davies-Bouldin Index graph).

2. Davies-Bouldin Index (DB Index)

- The DB Index is a measure of clustering quality, where lower values indicate better-defined clusters.
- For 3 clusters, the DB Index appears to be **lowest**, around **0.90**, suggesting this is the optimal clustering choice.

3. Other Clustering Metrics

- **Silhouette Score:** 0.383.
  The silhouette score measures how similar an object is to its own cluster compared to other clusters. A score close to 1 indicates well-separated clusters, while a low value (near zero) suggests overlapping clusters.
    - In this case, **0.383** suggests moderately defined clusters, with some potential overlap.
- **Inertia:** 233.137.
  Inertia represents the within-cluster sum-of-squares distance. Lower inertia is better

but must be balanced with the number of clusters. The inertia value supports a balanced choice for **3 clusters.**

Key Insights

- The clustering results show **3 clusters** as the optimal solution, based on the Davies-Bouldin Index (minimum at 3 clusters).
- While the Silhouette Score is moderate, it indicates that clusters are distinguishable but might have slight overlaps.
- Inertia supports the clustering model with 3 clusters, as it balances compactness with the number of clusters.