Clustering Analysis

1. Data Import and Feature Engineering

The code begins by importing customer, product, and transaction datasets. Key customer metrics are derived using transaction data:

- TotalRevenue: Total spending of each customer.
- TransactionCount: Number of transactions per customer.
- AvgOrderValue: Average value of orders.

Customer region information is merged to include geographical segmentation. The Region column is encoded using one-hot encoding for numerical processing.

2. Data Standardization

Numerical features are standardized using StandardScaler to ensure consistent scaling across features, preventing biases during clustering.

3. Clustering and Evaluation

To determine the optimal number of clusters, K-Means clustering is applied iteratively for cluster counts ranging from 2 to 10. Two key metrics are calculated:

- Davies-Bouldin Index (DB Index): Lower values indicate better-defined clusters.
- Silhouette Score: Higher values reflect well-separated clusters.

The optimal cluster count is selected based on the lowest DB Index.

4. Final Clustering and Visualization

K-Means clustering is applied using the optimal cluster count. Principal Component Analysis (PCA) reduces data to two dimensions for visualization. A scatterplot illustrates customer clusters with PCA components as axes.

5. Key Findings

- Optimal Number of Clusters: The number of clusters minimizing theDB Index.
- DB Index for Optimal Clustering: Indicates cluster compactness andseparation.
- Silhouette Scores: Provides comparative insights into clusterquality for different counts.

Insights and Benefits of Customer Segmentation

Customer clustering enables businesses to identify distinct groups based on purchasing behavior and region. These segments can inform tailored marketing campaigns, personalized offers, and resource allocation strategies.