Customer Segmentation Using K-Means Clustering

This script implements customer segmentation using K-Means clustering on engineered features derived from transaction and customer datasets. Key insights are derived by evaluating multiple clustering metrics and visualizing the resulting customer segments.

Key Features of the Code

1. Data Loading and Preparation

- **Functions**:
  - `load_and_prepare_data` : Loads customer, product, and transaction data and processes date fields for further analysis.
- Datasets:
  - Customers, products, and transactions are merged and prepared for feature engineering.

---

2. Feature Engineering for Clustering

- **Function**:
  - `create_simple_features` : Constructs features like:
    - `Recency` : Days since the last purchase.
    - `Frequency` : Total number of transactions.
    - `Monetary` : Total value of transactions.
    - `AvgOrderValue` : Average transaction value.
    - `Region` : Encoded as binary features for clustering.
- Output:
  - A feature matrix containing both transactional and regional customer attributes.

---

3. Data Scaling

- The features are scaled using `StandardScaler` for improved clustering performance.

---

4. Optimal Cluster Identification

- **Function**:
  - `find_optimal_clusters` : Computes Davies-Bouldin Index (DB), Silhouette Score, and Calinski-Harabasz (CH) Score for 2 to 10 clusters.
- Metrics:
  - **Davies-Bouldin Index (DB)**: Measures cluster compactness and separation. Lower scores are better.
  - **Silhouette Score**: Indicates the clarity of clusters. Higher scores are better.
  - **Calinski-Harabasz (CH) Score**: Evaluates cluster density and separation. Higher scores are better.
- Visualizations:
  - Line plots for DB Index, Silhouette Score, and CH Score versus the number of clusters.
  - Space for metrics:
    - Optimal clusters: 5.
    - DB Index: 0.988323.
    - Silhouette Score: 0.413089.
    - CH Score: 67.725235.

---

5. Final Clustering and Visualization

- The optimal number of clusters is selected based on the lowest Davies-Bouldin Index.
- **Visualization**:
  - Principal Component Analysis (PCA) reduces features to two dimensions.
  - A scatter plot visualizes clusters in the 2D PCA space.

Outputs

- **Cluster Metrics Summary**:
  - Optimal Clusters: 5.
  - Davies-Bouldin Index: 0.988323.
  - Silhouette Score: 0.413089.
  - Calinski-Harabasz Score: 67.725235.
- **Segmentation Insights**:
  - Cluster descriptions based on Recency, Frequency, and Monetary (RFM) values.

Usage

- **Modules Used**: pandas, numpy, scikit-learn, matplotlib, seaborn, scipy.
- **Instructions**:
    1. Place the datasets (`Customers.csv` , `Products.csv` , `Transactions.csv`) in the `data` folder.
    2. Run the script to generate clustering metrics, visualizations, and customer segmentation insights.