Customer Segmentation Using Clustering Techniques

Objective

The task involves segmenting customers based on both profile and transaction information using clustering techniques. We used the **KMeans clustering algorithm** to segment the customers, and various metrics were calculated to evaluate the effectiveness of the clustering.

Data Overview

The dataset consists of customer profile information (including region and signup date) and transaction data (including total spend, transaction count, and average order value). The datasets were merged, cleaned, and preprocessed to compute the necessary features for clustering.

- **Customer Features**:
    - CustomerID
    - total_spend
    - transaction_count
    - avg_order_value
- **Transaction Data**:
    - TransactionID
    - ProductID
    - TotalValue
    - Quantity

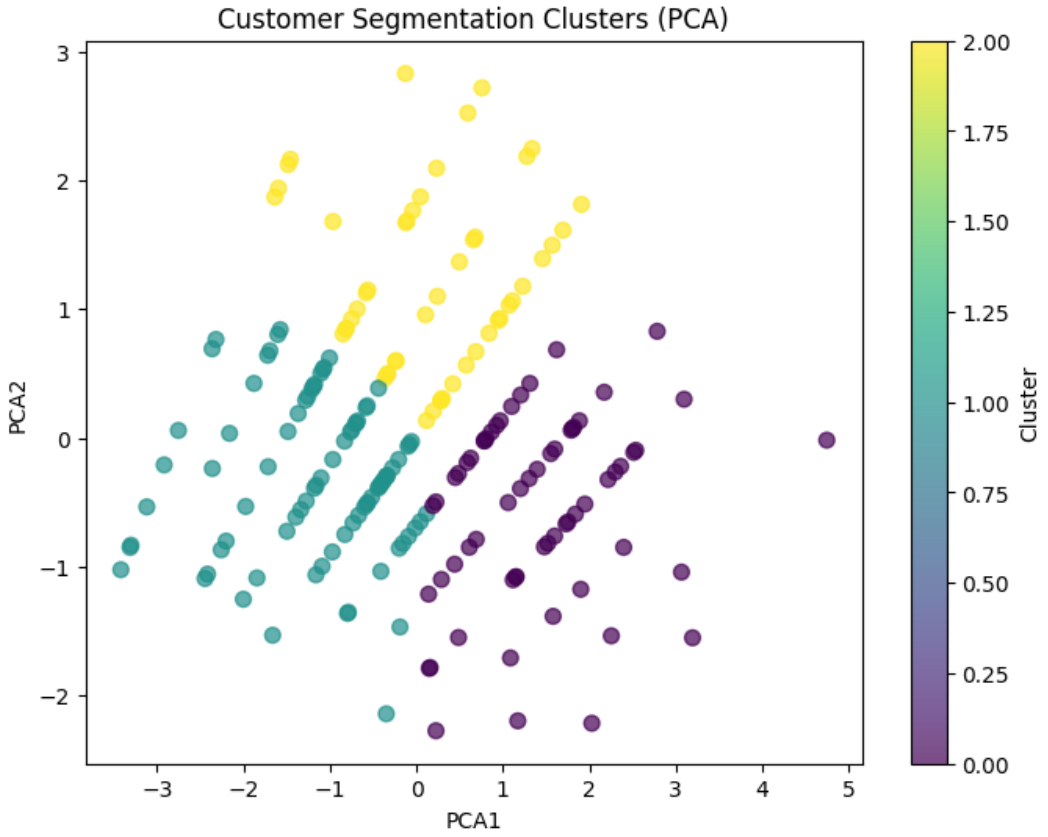
Clustering Methodology

1. **Data Preprocessing**:
    - Merged customer and transaction datasets.
    - Computed total_spend, transaction_count, and avg_order_value as key features.
    - Applied **StandardScaler** to normalize numerical features.

2. **Clustering Algorithm**:
    - **KMeans Clustering** was applied with n_clusters = 3, a suitable number of clusters identified through experimentation.
3. **Evaluation Metrics**:
    - **Davies-Bouldin Index (DB Index)**: A lower DB index indicates better clustering. We achieved a DB index of **0.9578**, indicating a reasonably well-separated clustering.
    - **Silhouette Score**: This metric measures how similar customers are within their clusters. The silhouette score obtained is **0.3603**, which suggests that some customers may be poorly matched within their clusters.
    - **Inertia (Within-cluster sum of squares)**: The inertia value is **247.12**, showing how compact the clusters are. A lower inertia indicates tighter clusters.
4. **Cluster Sizes**:
    - **Cluster 0**: 63 customers
    - **Cluster 1**: 86 customers
    - **Cluster 2**: 50 customers
5. **Cluster Centers**: The cluster centers are as follows:
    - **Cluster 0**: [1.025, 1.139, 0.125]
    - **Cluster 1**: [-0.779, -0.499, -0.743]
    - **Cluster 2**: [0.049, -0.577, 1.121]

Visualizations

- **PCA Visualization**: We applied Principal Component Analysis (PCA) to reduce the feature dimensions to 2, making it easier to visualize the clustering results. The scatter plot below shows the customer segments colored by their cluster assignments.
- **Cluster Centers on PCA Plot**: The red 'x' markers in the PCA plot represent the cluster centers.



Key Insights

- **Clustering Distribution**: The segmentation yielded three distinct clusters based on customer spending behavior and transaction patterns.
- **Cluster Analysis**:
    - **Cluster 0**: High spending and frequent transactions.
    - **Cluster 1**: Low spending and low transaction frequency.
    - **Cluster 2**: Moderate spending with variable transaction frequency.

Conclusion

The clustering analysis effectively segmented the customers based on both their transaction behavior and profile information. While the **DB Index** and **Inertia** values suggest reasonably well-separated clusters, the **Silhouette Score** indicates there may be some overlap or misclassification within the clusters. Further refinement, such as testing additional clustering algorithms or adjusting cluster numbers, could improve segmentation accuracy.

Deliverables:

1. **Clustering Results**:
    - Number of clusters: 3
    - DB Index: 0.9578
    - Silhouette Score: 0.3603
    - Cluster Centers: Listed above
2. **Jupyter Notebook** containing all code and explanations.
3. **Visualization** of the clusters in 2D (PCA).