Customer Segmentation Using KMeans

I've used KMeans clustering algorithm to segment customers based on their transactional behavior. Key factors used for segmentation include total spending, number of transactions, average transaction value, recency, frequency, and region.

The optimal number of clusters was determined using the elbow method and evaluation metrics like the Davies-Bouldin Index and Silhouette Score.

Data Used:

The dataset for clustering is a combination of customer profiles, transaction details, and product information. The customer data (`customer_data`) aggregates various features, including:

- **TotalValue:** The total spending of each customer.
- **NumTransactions:** The number of transactions made by each customer.
- **AvgTransactionValue:** The average value of each transaction made by a customer.
- **Recency:** The number of days since the customer's last transaction.
- **Frequency:** The number of distinct transaction days for each customer.
- **Region:** The continent where the customer resides.

Key Results:

1. Number of Clusters Formed:

Using the Elbow Method and Davies-Bouldin Index (DB Index) analysis, I identified that the optimal number of clusters is **3**. This choice was determined by analyzing the inertia values and clustering metrics across different numbers of clusters, ranging from 2 to 10.

2. Clustering Quality Metrics:

After applying KMeans clustering with 3 clusters, the following metrics were obtained:

- **Davies-Bouldin (DB) Index**:
  The **Davies-Bouldin Index** was used to assess the compactness and separation of the clusters.
  - **DB Index**: 1.3997

    A lower DBI indicates better-defined clusters. A DBI of 1.3997 is moderately good, suggesting that the clusters are reasonably compact and well-separated, though there is room for improvement.

- **Silhouette Score**:
  The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to +1, with higher values indicating better-defined clusters.
    - **Silhouette Score**: 0.2364

      This score suggests weak clustering structure, meaning the clusters might not be very well-defined.

- **Calinski-Harabasz Index**:
  The **Calinski-Harabasz Index** evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion.
    - **Calinski-Harabasz Index**: 76.3544

      A higher score indicates better clustering. With a score of 76.35, the clustering shows a moderate distinction between clusters.

Cluster Characteristics:

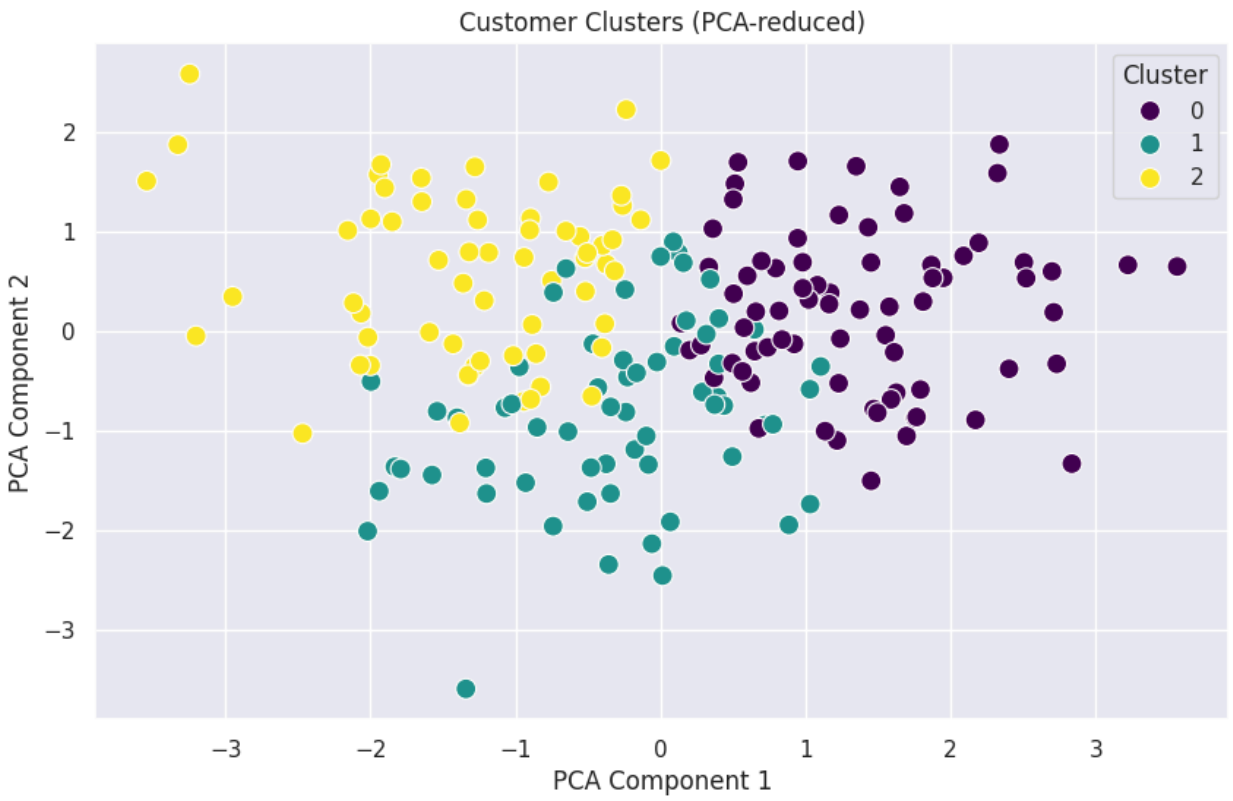
Each cluster represents a distinct group of customers based on their behavior. By examining the cluster centroids (mean values for each feature in a cluster), this can be interpreted :

- **Cluster 1 (High-Spenders)**:
    - High total spending, moderate number of transactions.
    - Higher-than-average transaction value.
    - Customers in this cluster tend to have less frequent but larger transactions.
- **Cluster 2 (Frequent Buyers)**:
    - Customers with a high frequency of transactions but lower average spending.
    - These customers make regular purchases but tend to spend less per transaction.
- **Cluster 3 (Low-Spenders)**:
    - Customers who have lower total spending, less frequent transactions, and higher recency values.
    - These customers are less engaged, possibly indicating dormant accounts or low-value customers.

Cluster Visualization:

To visualize the customer segmentation, **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the data to 2 components for a 2D scatter plot:

- The plot clearly shows the formation of 3 distinct clusters, where each cluster represents different customer groups based on their transactional behavior. The clusters are visually separated, providing a clear understanding of customer segmentation.



Conclusion:

- **3 distinct customer segments** were identified, representing high-spenders, frequent buyers, and low-engagement customers.
- The clustering quality was moderate, with a **Davies-Bouldin Index of 1.3997** and a **Silhouette Score of 0.2364**, indicating reasonably well-defined clusters.
- These results can be further used to tailor marketing strategies, personalize customer interactions, and improve business decision-making.

The customer segmentation provides valuable insights into customer behavior, enabling more targeted campaigns and effective resource allocation.

Attached:

- I've attached a CSV file (customer_clusters.csv) containing customer data with their assigned cluster labels.