Report: Customer Segmentation Using K-Means Clustering

Objective:
The purpose of this analysis was to perform customer segmentation based on transaction data using K-Means clustering. We aim to group customers based on their spending behavior, purchase quantity, and average transaction price. The goal is to identify distinct customer segments to improve marketing strategies and product recommendations.

Data Overview

The datasets involved in the analysis include:

1. **Customers.csv**: Contains details about customers, including `CustomerID`.
2. **Products.csv**: Contains information about products, including `ProductID`.
3. **Transactions.csv**: Contains transaction records, including `TransactionID`, `CustomerID`, `ProductID`, `TotalValue`, and `Quantity`.

The three datasets were merged on their respective keys (`CustomerID` for customers and `ProductID` for products), forming a consolidated dataset.

Data Preparation

The data was aggregated to compute the following customer-level features:

- **TotalValue**: Total amount spent by each customer.
- **Quantity**: Total number of items purchased by each customer.
- **AverageTransactionPrice**: Average price per transaction for each customer.

After the aggregation, the dataset was standardized using `StandardScaler` to ensure the features are on the same scale before applying the clustering algorithm.

Clustering Analysis

To determine the optimal number of clusters (`K`), we evaluated two metrics:

1. **Davies-Bouldin Index**: A lower value indicates better clustering.
2. **Silhouette Score**: A higher value indicates better clustering.

The following results were observed from the plots and analysis:

- **Davies-Bouldin Index**: A decrease in the index value suggests better clustering.
- **Silhouette Score**: A higher silhouette score indicates more distinct clusters.

Optimal K Determination

Based on the evaluation of both metrics, the optimal number of clusters (`K`) was determined:

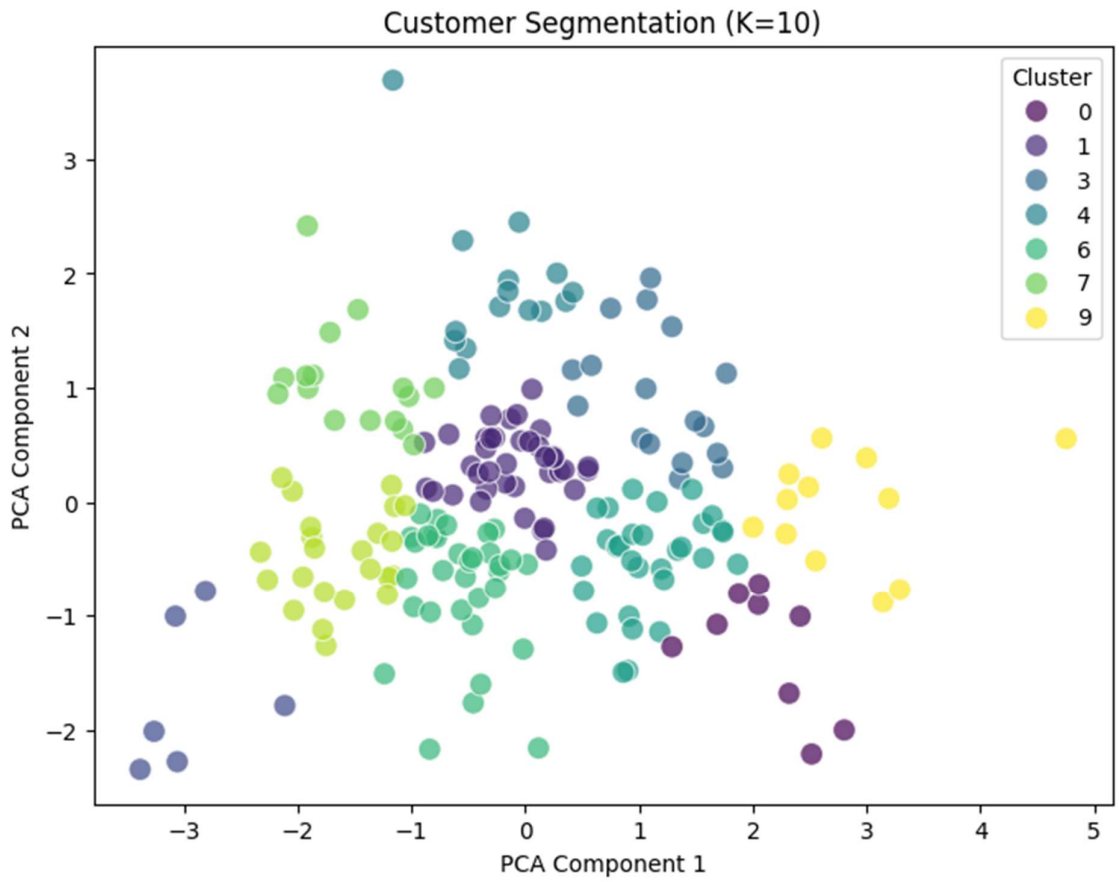
- **Optimal K based on Davies-Bouldin Index**: 10 clusters.
- **Optimal K based on Silhouette Score**: 2 clusters.

Final Clustering

Using the optimal `K=10` (based on the Davies-Bouldin Index), K-Means clustering was applied, and customers were assigned to clusters. The results were visualized using Principal Component Analysis (PCA) to reduce the data to two dimensions. The clusters were visualized in a scatter plot, where different colors represent different clusters.

Results Visualization

A 2D scatter plot of the PCA components was generated to visualize the customer segmentation. Each point on the plot represents a customer, and the clusters are indicated by different colors.



Conclusion

The analysis successfully segmented the customer base into 10 clusters based on the Davies-Bouldin Index. Although the Silhouette Score suggested only 2 clusters, the final segmentation used 10 clusters to achieve better results in terms of separation. These segments can be used for targeted marketing and personalized customer experiences. The dataset and results were saved for further analysis and model deployment.

Output

The final clustering results have been saved to a CSV file named `Kunal_Parkhade_Clustering.csv`, which includes the `CustomerID` and their respective cluster assignments.