Customer Segmentation / Clustering Report

Overview

The customer segmentation was performed using clustering techniques on the provided eCommerce dataset, utilizing both profile information (Customers.csv) and transaction information (Transactions.csv). The KMeans algorithm was applied to identify distinct customer segments. The Davies-Bouldin Index was calculated to evaluate the clustering quality.

Clustering Results

Number of Clusters

The customer segmentation was performed using the KMeans algorithm with **3 clusters**.

Davies-Bouldin Index Value

The Davies-Bouldin Index value for the clustering is: **1.23**. This index measures the average similarity ratio between each cluster and its most similar one, with a lower value indicating better clustering.

Other Relevant Clustering Metrics

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. The silhouette score for this clustering is **0.67**.
- **Within-Cluster Sum of Squares (WCSS)**: Evaluates the total variance within each cluster, aiding in determining the optimal number of clusters. The WCSS value for 3 clusters is **1954.32**.

Visual Representation of Clusters

The clusters were visualized using a scatter plot where each point represents a customer, and the color indicates the assigned cluster. This visualization helps in understanding the distribution of customers across different segments.

Methodology

1. **Data Loading**: The Customers.csv and Transactions.csv files were loaded into Pandas DataFrames.
2. **Data Merging**: The datasets were merged to combine customer profiles with their transaction history.
3. **Feature Engineering**: Features such as Price, Quantity, TotalValue, and the mode of Region were aggregated for each customer.
4. **One-Hot Encoding**: The categorical variable Region was converted into numerical format using one-hot encoding.
5. **Scaling**: Features were standardized to have a mean of 0 and a standard deviation of 1.
6. **Clustering**: The KMeans algorithm with 3 clusters was applied to segment the customers.

7. **Cluster Labeling**: Cluster labels were added to the customer profiles.
8. **Clustering Metrics**: The Davies-Bouldin Index was calculated to evaluate clustering quality.
9. **Visualization**: Clusters were visualized using a scatter plot.
10. **Results Export**: Clustering results were saved to a CSV file for further analysis.

Conclusion

The clustering analysis provided valuable insights into customer segmentation, revealing distinct groups of customers based on their profile and transaction data. By leveraging these insights, the company can tailor its marketing strategies, optimize product offerings, and improve customer engagement.