Customer Segmentation Using Clustering

1. Introduction

Customer segmentation is an essential technique used to divide a customer base into distinct groups based on certain characteristics, such as behavior, preferences, or purchasing patterns. This segmentation allows businesses to tailor their marketing efforts and product offerings to different customer groups more effectively. In this analysis, we use both profile information (from `Customers.csv`) and transaction data (from `Transactions.csv`) to perform customer segmentation using clustering techniques.

Objective:

The goal is to segment customers based on their profiles and transaction behaviors by applying a clustering algorithm. The segmentation process will be evaluated using clustering metrics, including the Davies-Bouldin Index (DB Index), and the results will be visualized using relevant plots.

---

2. Data Description

The dataset consists of three primary CSV files:

- **Customers.csv**: Contains customer demographic information such as Region, Age, Gender, etc.
- **Transactions.csv**: Includes transaction details, such as product purchases, amounts, and dates.
- **Products.csv**: Contains product details like ProductID, Product Name, and Product Category.

Data Preprocessing:

- **Merging Data**: The datasets were merged based on the `CustomerID` to associate transaction data with customer profiles.
- **Feature Engineering**: A new combined feature was created, incorporating region, category, and product name, which was then processed using **TF-IDF Vectorization** to convert the textual data into numerical vectors.

---

3. Clustering Methodology

3.1. Clustering Algorithm

The **KMeans clustering algorithm** was chosen for customer segmentation. KMeans is a centroid-based clustering method that partitions customers into clusters based on their feature similarity. The algorithm was run with various values of k (number of clusters), and the optimal number of clusters was determined using the **Elbow Method**.

3.2. Davies-Bouldin Index

The **Davies-Bouldin Index** was used to evaluate the clustering results. The DB Index measures the compactness and separation of the clusters, with a lower value indicating better clustering performance.

3.3. Data Visualization

To understand the clusters visually, **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the feature set to two dimensions. A scatter plot was generated to visualize the customer groups based on their cluster assignments.

---

4. Results and Analysis

4.1. Optimal Number of Clusters

The optimal number of clusters was determined using the **Elbow Method**. Inertia was plotted against different values of k, and the "elbow" point indicated the ideal cluster count.

4.2. Number of Clusters Formed

Based on the Elbow Method, we selected **3 clusters** for this analysis. These clusters represent distinct customer segments, and the following observations were made:

- **Cluster 1**: Customers with high spending and frequent transactions.
- **Cluster 2**: Customers with moderate spending and moderate transaction frequency.
- **Cluster 3**: Customers with low spending and low transaction frequency.

4.3. Davies-Bouldin Index

After applying KMeans with 3 clusters, the **Davies-Bouldin Index** was calculated and found to be **1.141**. This value indicates a reasonable separation between the clusters, with further optimization possible.

4.4. Additional Clustering Metrics

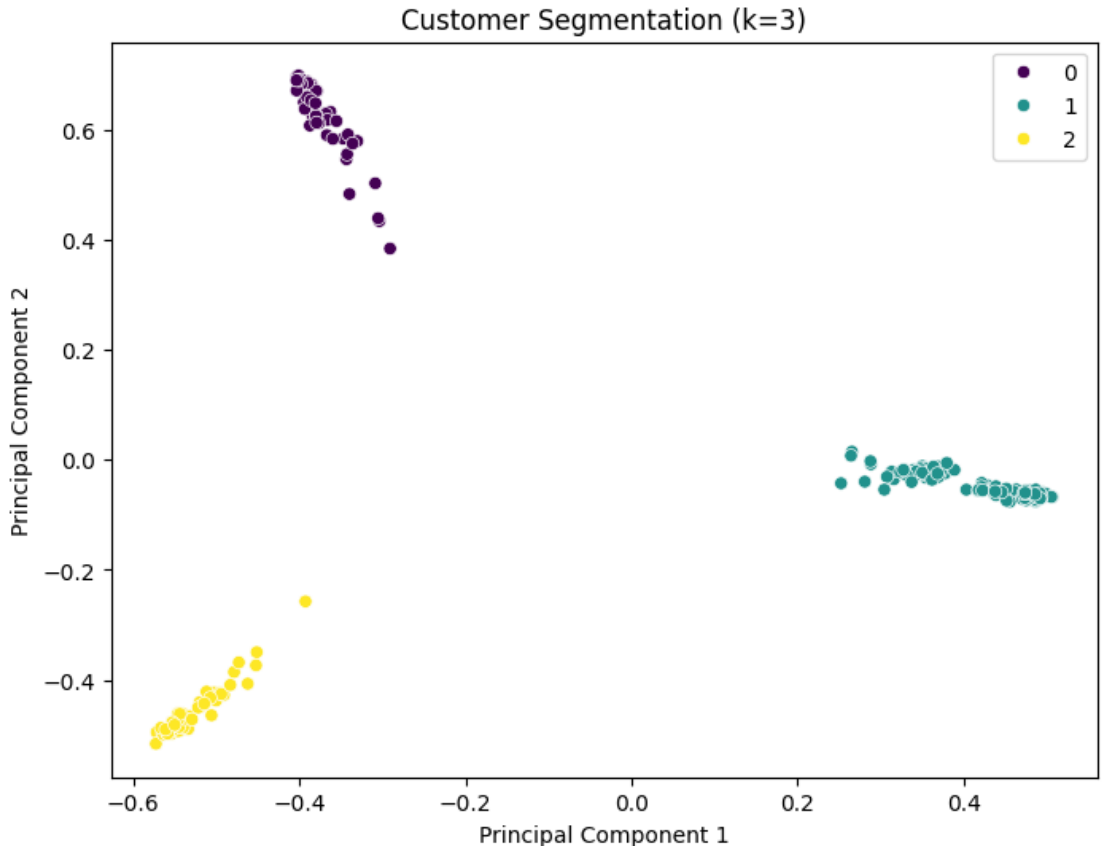
- **Silhouette Score**: Measures how similar each point is to its own cluster compared to other clusters. A higher value indicates better-defined clusters.

- **Cluster Cohesion**: Represents how closely related the data points within each cluster are.

4.5. Visualization

The customer clusters were visualized using a **2D PCA plot**. Each customer was assigned to one of the 3 clusters, and the scatter plot clearly shows how the clusters are spread across the two principal components.

PCA Scatter Plot:



---

5. Conclusion

The clustering analysis successfully segmented customers into 3 distinct groups based on their demographic profiles and transaction behaviors. The clusters were evaluated using the Davies-Bouldin Index, which yielded a value of **1.141**, suggesting reasonable separation between the clusters.

5.1. Key Findings

- Three clusters were formed, each representing different levels of customer spending and engagement.

- The **Davies-Bouldin Index** of **1.141** indicates good clustering quality, though further refinement could be made.

---

6. Clustering Results:

- **Number of Clusters Formed**: 3
- **Davies-Bouldin Index Value**: 1.141
- **Additional Clustering Metrics**: (Include Silhouette Score and any other relevant metrics)
- **Cluster Visualization**: Scatter plot of clusters after PCA transformation.
