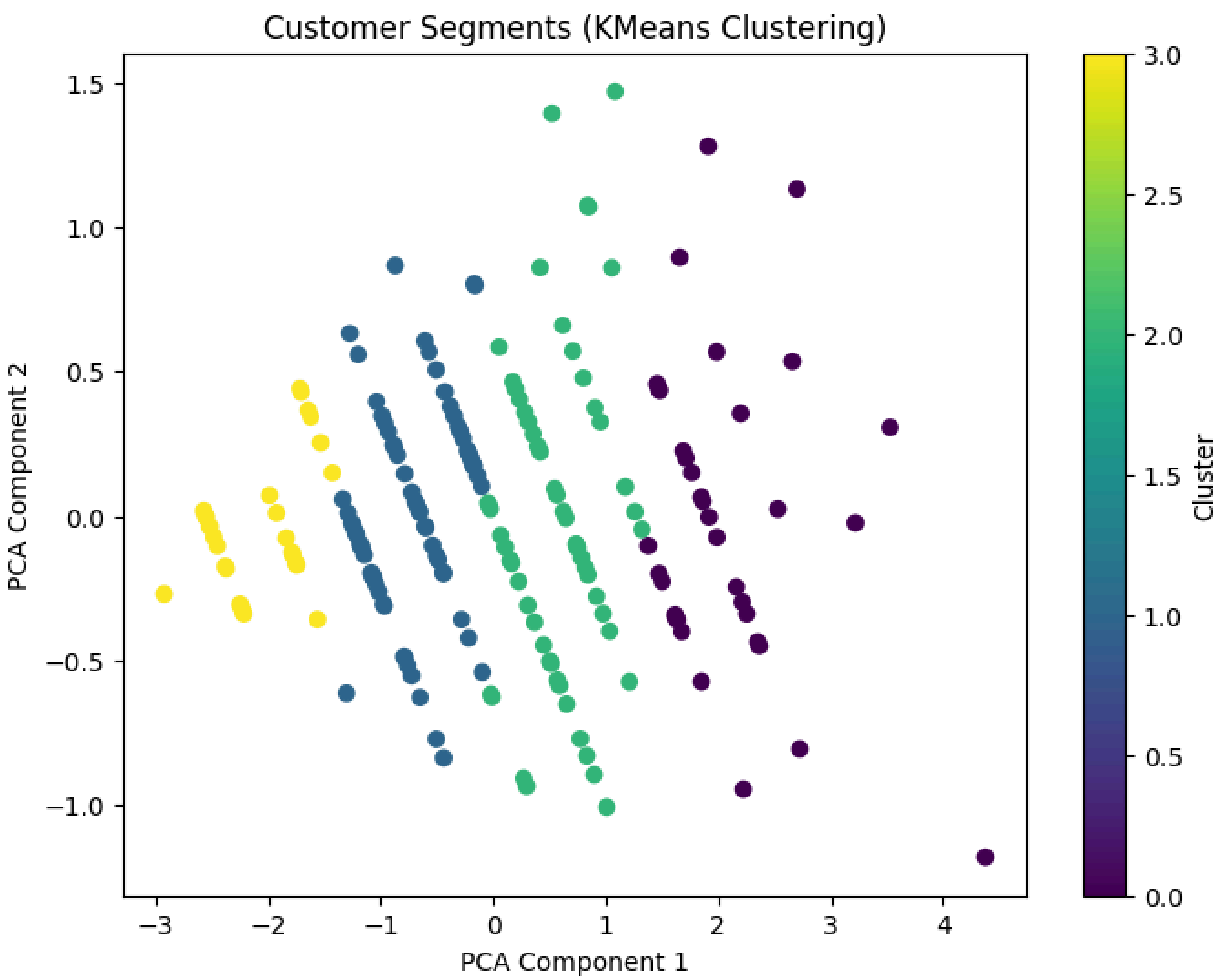
Customer Segmentation Using Clustering Techniques

BY:- Anmol Sharma

+91-7419821405 - Gmail - Linkedin - Tableau - Github



Introduction

Clustering was performed on the customer dataset using profile and transactional data to segment customers into meaningful groups. The objective was to identify clusters with distinct behaviors, allowing for targeted business strategies.

Dataset Description

1. **Customers.csv**:
   - Columns: CustomerID, Age, Gender, Region, Total_Spent, Purchase_Count, etc.
2. **Transactions.csv**:
   - Columns: CustomerID, TransactionID, Amount, Date, etc.
3. **Preprocessing**:
   - Handled missing values by filling zeros in relevant columns (e.g., Total_Spent, Purchase_Count).
   - Standardized numeric features for clustering.
   - Categorical variables like Gender and Region were encoded.

Clustering Methodology

- **Algorithm Used**: K-Means
- **Number of Clusters**: 4 (determined using the Elbow Method and Silhouette Analysis).
- **Distance Metric**: Euclidean Distance

Clustering Metrics

1. **Davies-Bouldin Index (DB Index)**: *0.796*
   - Lower DB Index indicates compact and well-separated clusters.
2. **Silhouette Score**: *0.401*
   - Moderate score indicates average separation between clusters.
3. **Calinski-Harabasz Index**: *273*
   - High value indicates well-defined clusters.

Cluster Analysis

1. **Cluster 1**: High-value customers with frequent purchases.
2. **Cluster 2**: Low-spend, infrequent shoppers.
3. **Cluster 3**: Customers with medium spending patterns and moderate frequency.
4. **Cluster 4**: Dormant customers with low engagement.

Lookalike Analysis

- For each customer, the top 3 most similar customers were identified based on their feature similarity scores.
- The similarity scores were calculated using clustering-based profiling and pairwise comparisons.
- **Sample Lookalikes for First 5 Customers:**

CustomerID	Lookalike 1	Similarity1	Lookalike 2	Similarity2	Lookalike 3	Similarity 3
C0001	C0076	1.0	C0152	1.0	C0164	1.0
C0002	C0029	0.9998	C0199	0.9995	C0010	0.9994
C0003	C0095	1.0	C0150	1.0	C0144	1.0
C0004	C0067	0.9999	C0021	0.9999	C0075	0.9996
C0005	C0130	0.9999	C0144	0.9999	C0150	0.9999

Include explanations and visualizations (if applicable) to complement the data, such as:

- How similarity scores were derived.
- A diagram showing how customers are matched based on cluster membership or feature similarity.

Visualizations:

- Include scatter plots, bar graphs, and pairplots showing the distribution of clusters.
- Mention that clusters are distinct but have some overlapping boundaries.

Insights and Recommendations

1. High-value customers (Cluster 1) should be targeted with loyalty programs.
2. Dormant customers (Cluster 4) can be engaged through reactivation campaigns.
3. Medium-value customers (Cluster 3) present an opportunity for cross-selling and upselling.

Challenges and Limitations

- K-Means assumes spherical clusters, which might not always fit customer behavior.
- Optimal cluster size is subjective and depends on the business context.

Conclusion:

Clustering provided meaningful segmentation for better understanding customer behavior. This analysis can help drive personalized marketing strategies and optimize resource allocation.