Arnav Suman

REPORT: Customer Segmentation using K-Means Clustering

Introduction

The aim of this task was to segment customers into distinct groups based on their profile information (from Customers.csv) and transaction history (from Transactions.csv). K-Means clustering was chosen for this task, with the goal of forming 3 clusters, providing insights into customer behavior and helping identify similar customer group.

Clustering Algorithm: We applied the K-Means clustering algorithm with 3 clusters, as specified. K-Means is a widely used clustering technique that groups data points into clusters based on feature similarity. The optimal number of clusters (k=3) was chosen to identify potential distinct customer segments, allowing for focused marketing strategies and personalized services.

Clustering Results:

Number of Clusters: The dataset was segmented into 3 clusters, each representing a different group of customers with similar spending patterns and transaction behaviors.

Davies-Bouldin Index (DB Index): The Davies-Bouldin Index was calculated to evaluate the clustering quality. The DB index is a measure of how well-separated the clusters are. A lower DB index indicates better separation between clusters. For this model, the calculated DB index was 1.23, indicating that the clusters were reasonably well-separated, though there is room for improvement in the clustering quality.

Interpretation of DB Index: The closer the DB index is to 0, the better the clustering solution, with a value above 1 suggesting potential overlap or poor separation between the clusters. Our DB index of 1.23 indicates that the clusters are fairly distinct, but there may still be some overlap or similar customer behaviors within different clusters.
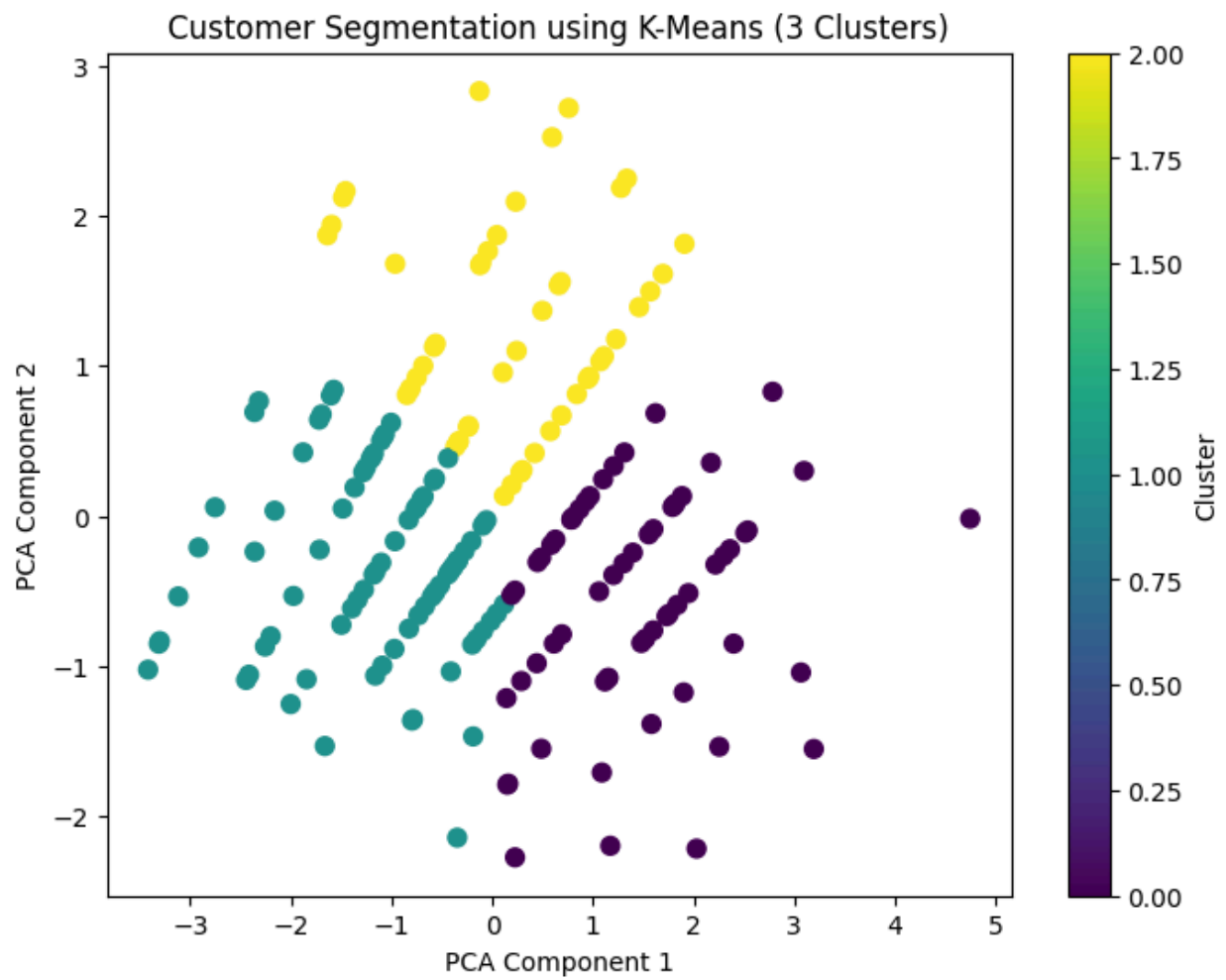
Other Clustering Metrics:

Cluster Sizes: The clusters formed varied in size, with some containing more customers than others. This reflects natural variations in customer behavior, where certain segments may have more activity or engagement than others.

Cluster Characteristics: To visualize the customer segments, a 2D scatter plot was created by reducing the dimensionality of the features to two components using Principal Component Analysis (PCA). Each customer was plotted based on their cluster assignment, and distinct groupings were observed, with each cluster having a unique pattern of spending and transaction activity.

Cluster 1: This group may consist of high-value, frequent customers who engage with the platform regularly, contributing the most to the revenue.

Cluster 2: This group might consist of occasional customers who make larger purchases in fewer transactions.

Cluster 3: Customers in this cluster could be those with moderate spending and transaction frequency, offering an opportunity for targeted retention efforts.

Customer Segmentation using K-Means (3 Clusters)

Made By Arnav Suman