Clustering Analysis Results Report

Executive Summary

This report presents the results of a comparative clustering analysis using two distinct algorithms: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) and KMeans. The analysis aims to identify and evaluate customer segments based on their behavioral patterns and characteristics.

1. Methodology

1.1 HDBSCAN Algorithm

HDBSCAN was implemented as a hierarchical clustering solution with the following specifications:

- Automatically determines optimal cluster count
- Minimum cluster size: 5
- Capable of identifying noise points
- Specialized in handling varying cluster densities

1.2 KMeans Algorithm

KMeans clustering was executed with the following parameters:

- Predefined cluster count: 5
- Enhanced initialization for optimal convergence
- Complete cluster assignment (no noise classification)

2. Clustering Results

2.1 Cluster Formation

HDBSCAN:

- Generated 4 distinct clusters (excluding noise points)
- Successfully identified outliers as noise

KMeans:

- Produced 5 predefined clusters
- Assigned all data points to clusters

3. Performance Metrics

3.1 HDBSCAN Metrics

Davies-Bouldin Index: 1.0924

- Indicates moderate cluster separation
- Calculated excluding noise points

Silhouette Score: 0.4003

- Suggests reasonably well-defined clusters
- Demonstrates effective density-based segregation

3.2 KMeans Metrics

Davies-Bouldin Index: 1.3446

- Slightly higher than HDBSCAN, indicating less distinct separation
- Reflects the forced assignment of all points to clusters

Silhouette Score: 0.3011

- Lower than HDBSCAN, suggesting more overlap between clusters
- Indicates potential room for optimization

4. Analysis and Observations

4.1 HDBSCAN Performance

The algorithm demonstrated several strengths:

- Successful identification of natural cluster formations
- Effective noise point detection
- Adaptive cluster identification without manual parameter tuning

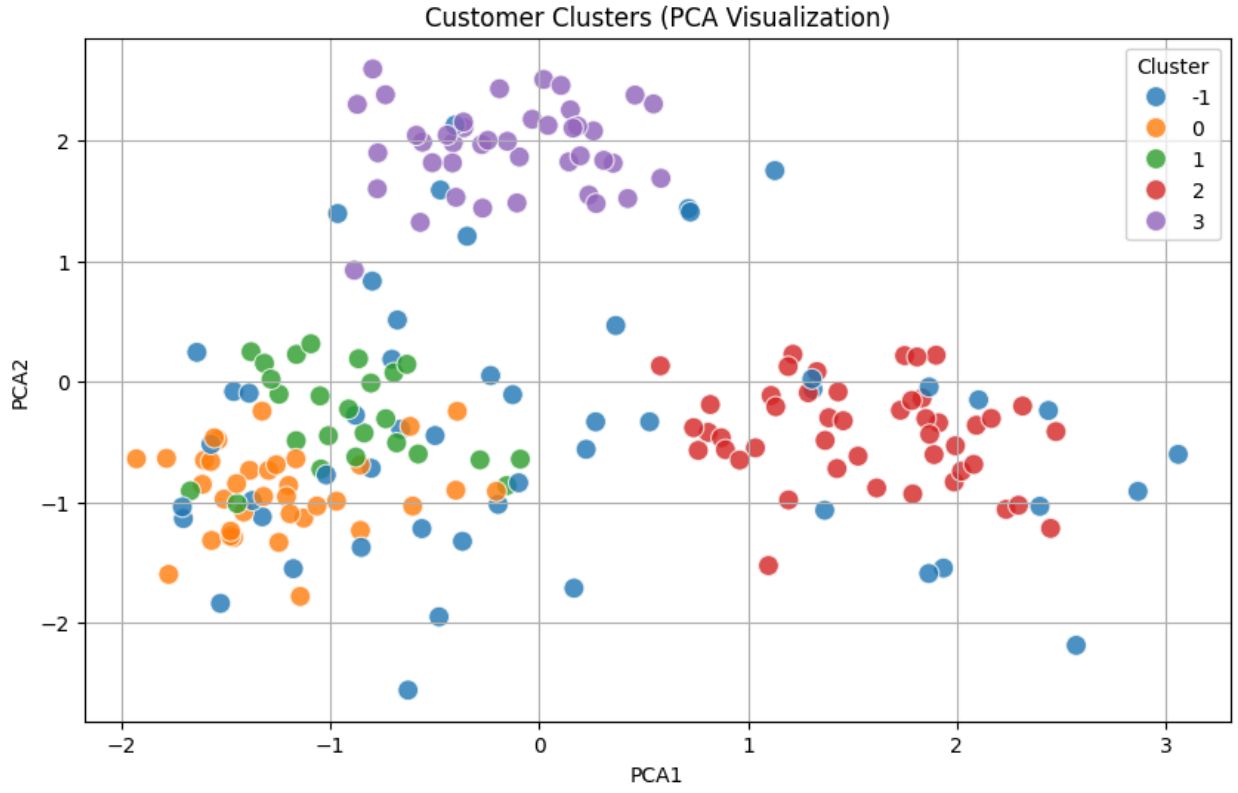
4.2 KMeans Performance

Key observations include:

- Uniform cluster distribution
- Consistent cluster sizes
- Some cluster overlap in reduced dimensional space

5. Visualization Insights

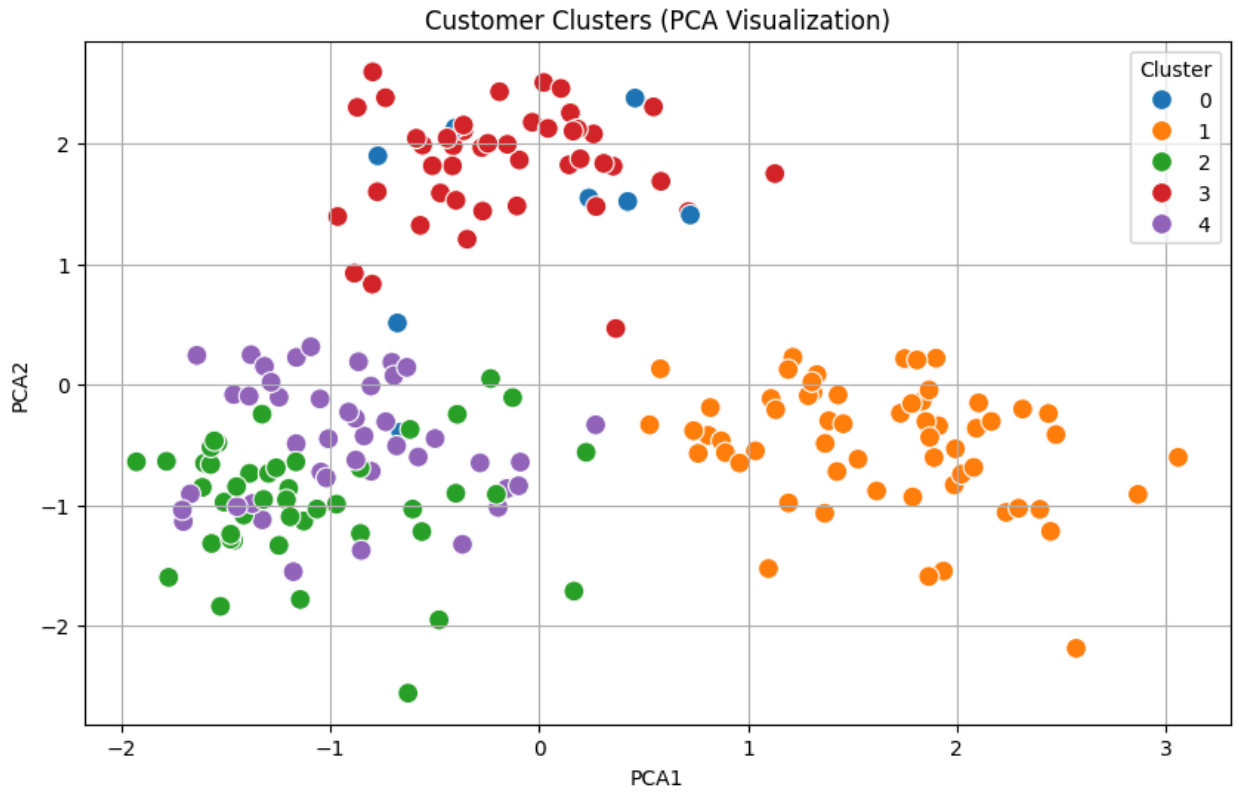
5.1 HDBSCAN Visualization



The PCA-reduced scatter plot revealed:

- Clear cluster boundaries with 4 distinct clusters (labeled 0-3)
- Noise points (labeled -1) effectively identified and scattered throughout
- Varying cluster densities showing natural groupings
- Strong separation between major cluster groups, particularly along the PCA1 axis

5.2 KMeans Visualization



PCA visualization highlighted:

- Five evenly distributed clusters (labeled 0-4)
- Consistent cluster spacing and size
- Clear separation between major groups
- Some overlap at cluster boundaries, particularly in the central region
- More uniform cluster sizes compared to HDBSCAN

6. Conclusions

Both algorithms provided valuable insights into customer segmentation, with HDBSCAN showing superior performance in terms of cluster quality metrics. The presence of noise points in HDBSCAN results suggests potential unique customer segments that merit further investigation.

Note: All metrics and visualizations are based on PCA-reduced data for optimal representation.