Customer Segmentation Analysis

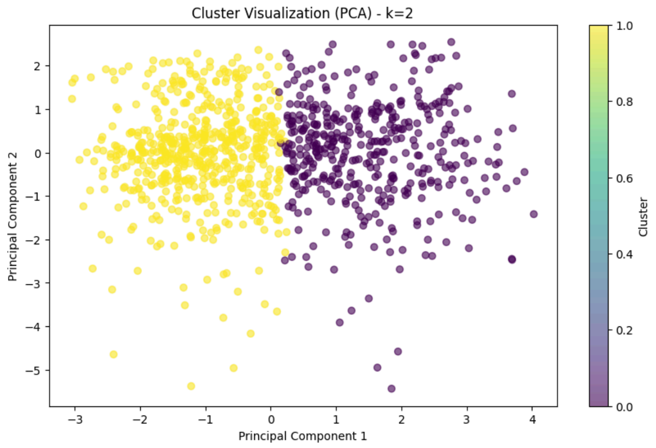
Executive Summary

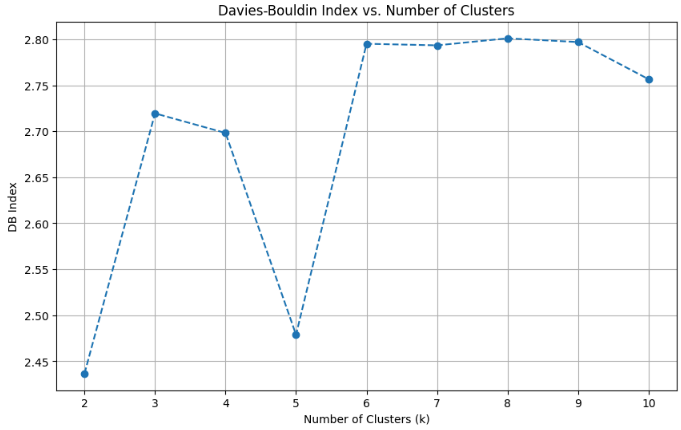
This analysis explores customer segmentation using five clustering techniques: **K-Means, K-Medoids, DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models (GMM)**. The dataset contains **111 features**, capturing customer behavior in transactions, geographic distributions, and product preferences. Feature preprocessing includes one-hot encoding for categorical variables.

Clustering Methodologies & Key Findings

1. K-Means Clustering

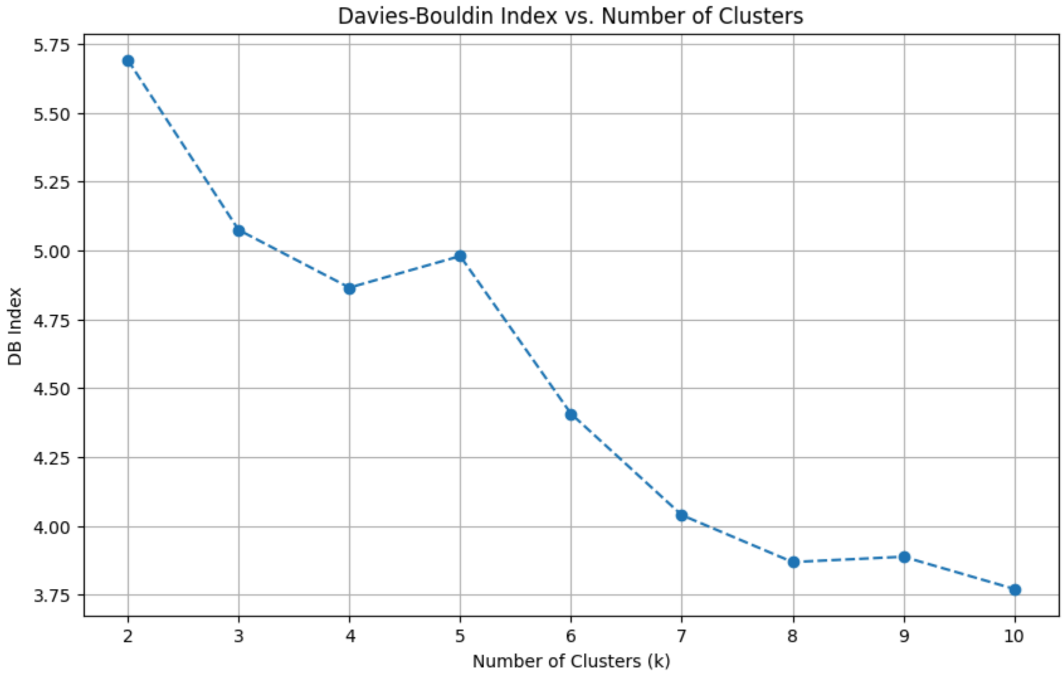
- **Optimal Clusters:** 2
- **Validation Method:** Davies-Bouldin (DB) Index, optimal score = **2.4367**
- **Visualization:** PCA reveals two distinct clusters, some overlap in the center.
- **Strengths:** Fast, simple, effective for spherical clusters.
- **Limitations:** Sensitive to outliers, assumes equally sized clusters.

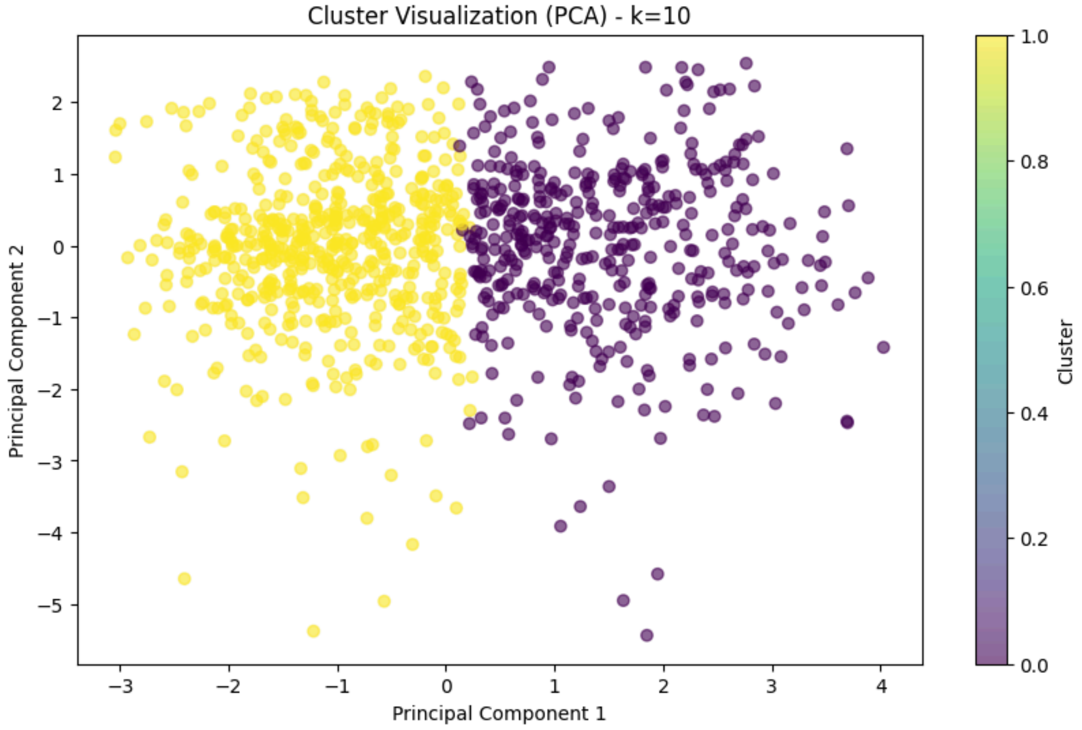




2. K-Medoids Clustering

- **Optimal Clusters:** 10
- **Validation Method:** Davies-Bouldin (DB) Index, best score = **3.7704**
- **Visualization:** PCA indicates complex structure, overlapping clusters.
- **Strengths:** Robust to outliers, interpretable medoids as cluster centers.
- **Limitations:** Computationally expensive, struggles with poorly defined clusters.



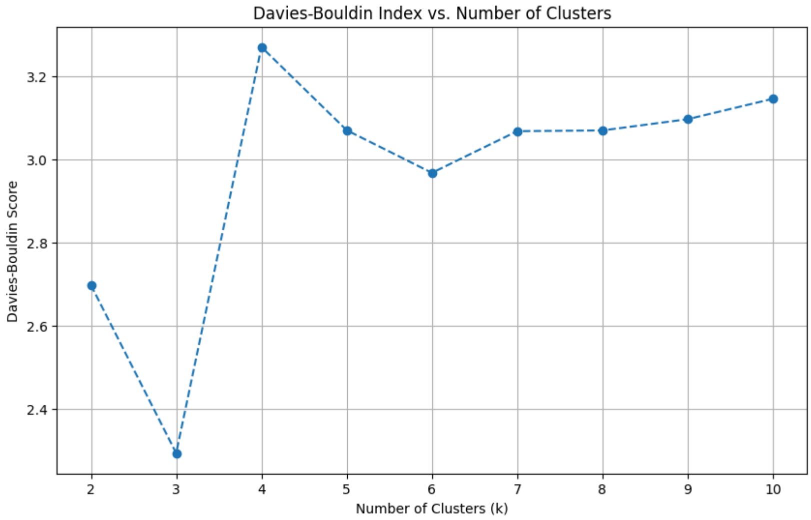


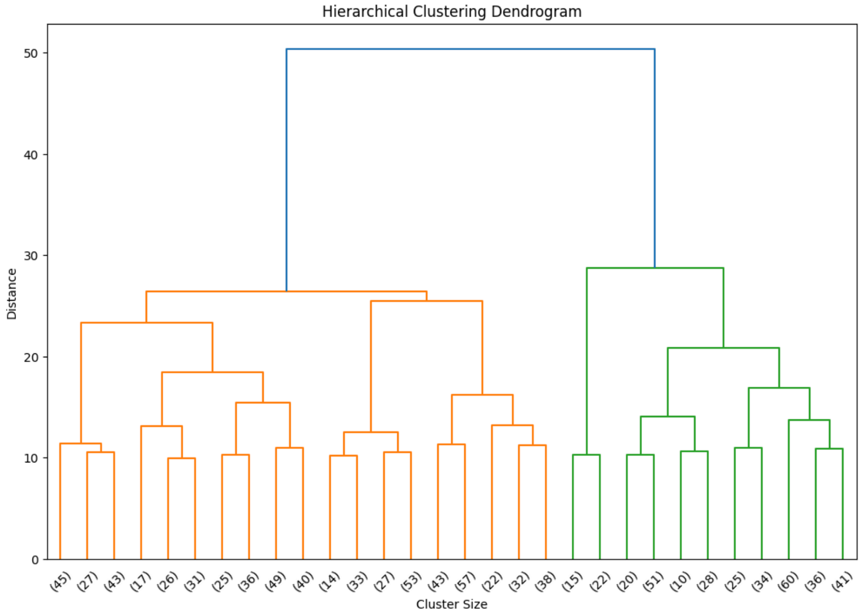
3. DBSCAN (Density-Based Clustering)

- **Cluster Formation:** Failed to form viable clusters (all points labeled noise).
- **Validation Method:** Extensive parameter tuning, but no successful clustering.
- **Reasons for Failure:** High dimensionality and binary feature sparsity.
- **Strengths:** Handles noise well, works for arbitrary cluster shapes.
- **Limitations:** Struggles with high-dimensional, sparse binary data.

4. Hierarchical Clustering

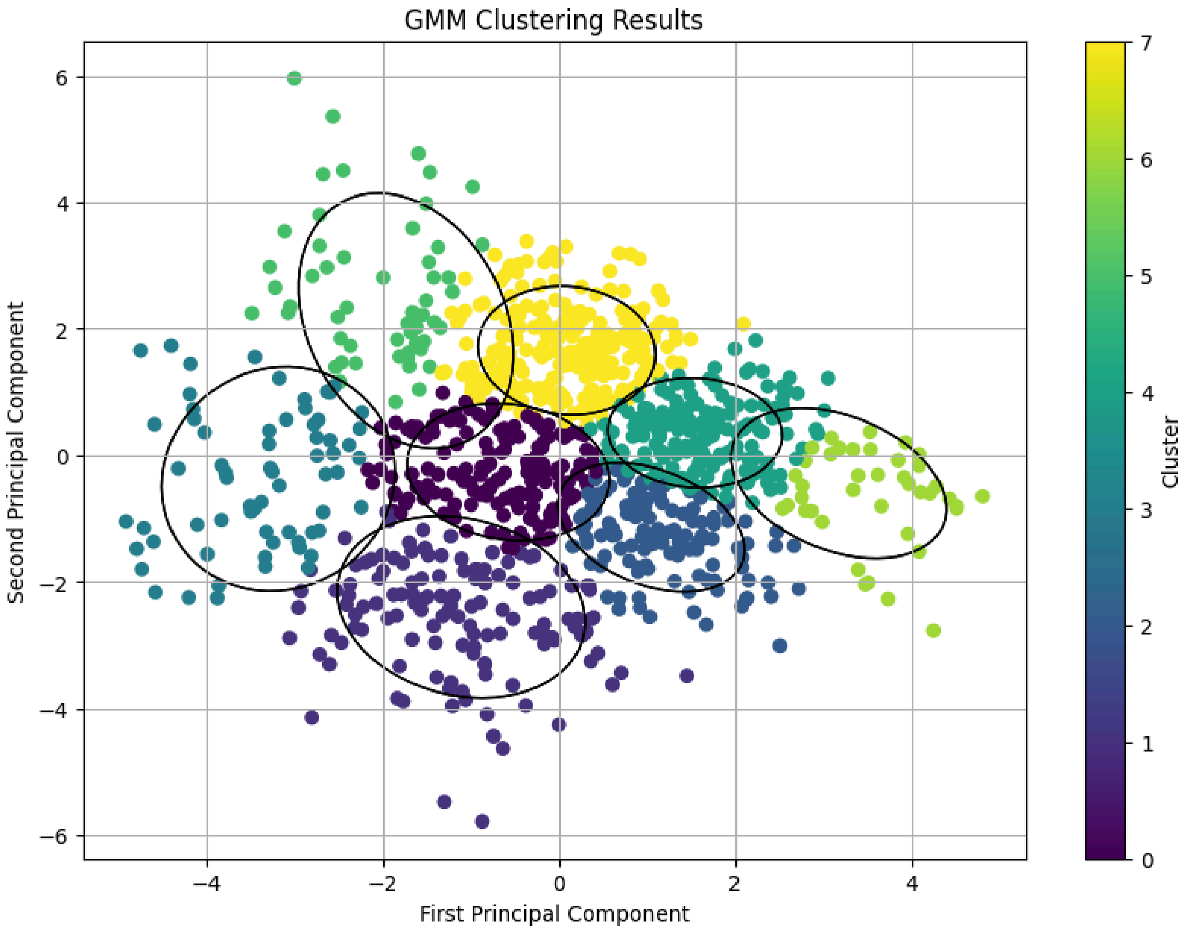
- **Optimal Clusters:** 3
- **Validation Method:** DB Index = **2.29**, dendrogram analysis.
- **Visualization:** Stable hierarchical structure before & after PCA.
- **Strengths:** Provides a full cluster hierarchy, does not require predefined clusters.
- **Limitations:** Computationally expensive for large datasets, linkage choice impacts results.





5. Gaussian Mixture Model (GMM)

- **Optimal Components:** 8
- **Validation Method:** DB Index = **0.8356**
- **Visualization:** PCA shows complex interactions with some overlap.
- **Strengths:** Models diverse cluster shapes, allows soft assignments.
- **Limitations:** Assumes Gaussian distributions, computationally intensive.



Comparative Analysis

Clustering Method	Optimal Clusters/Components	Validation Method	Best DB Index Score	Strengths	Limitations
K-Means	**2**	DB Index	**2.4367**	Simple, fast, effective for spherical clusters	Sensitive to outliers, assumes spherical & equally sized clusters
K-Medoids	**10**	DB Index	**3.7704**	Robust to outliers, uses real data points as centers	Computationally expensive, struggles with ill-defined clusters
DBSCAN	**FAILED**	Extensive parameter tuning	N/A	Good for varying shapes & sizes, handles noise well	Struggles with high-dimensional & sparse binary data
Hierarchical	**3**	DB Index	**2.29**	Provides full hierarchy, useful for nested structures	Computationally expensive, choice of linkage affects results
GMM	**8**	DB Index	**0.8356**	Models different shapes, probabilistic assignments	Assumes Gaussian distribution, computationally intensive

Key Takeaways & Business Implications

- **K-Means & Hierarchical Clustering:** Useful for high-level business strategies (e.g., premium vs. value-seeking customers).
- **K-Medoids & GMM:** Offer more granular segmentation, beneficial for personalized marketing & targeted strategies.
- **DBSCAN:** Not suitable for high-dimensional, sparse binary datasets.
- **Scalability:** K-Means is the most computationally efficient, while K-Medoids and GMM offer deeper insights at a computational cost.
- **Interpretability:** K-Medoids and GMM provide more meaningful clusters due to their approach to centroids/medoids and probabilistic modeling.

This analysis provides insights into **how different clustering techniques segment customers, balancing computational efficiency, robustness, and interpretability**. Future work could explore **feature selection, alternative clustering validation metrics, or hybrid clustering approaches** to refine segmentation results.

Customer Segmentation Analysis - Summary

Potential for Overlap & Clustering Detail Despite high dimensionality and binary nature, GMM and K-medoids indicate nuanced relationships, suggesting non-strictly categorical data. K-means provides a simple binary segmentation, whereas K-medoids and GMM reveal complex subgroup structures (10 clusters and 8 components).

Dimensionality & Data Sparsity PCA (retaining 85.2% variance across 26 components) highlights persistent clustering challenges. High-dimensional binary features lead to sparse data, making density-based methods like DBSCAN ineffective.

Data Characteristics & Outliers K-medoids' robustness suggests significant outliers affecting clustering-sensitive models. Clustering results (2, 3, 8, 10 clusters) imply natural groupings, with hierarchical clustering identifying a stable 3-cluster hierarchy.

Challenges & Considerations Binary features limit traditional clustering approaches due to ineffective distance metrics. DBSCAN's failure underscores the need for dimensionality reduction or alternative representations. Each method reveals unique insights, reinforcing the multifaceted nature of data structure.

Implications for Data Handling

- **Feature Engineering:** Transformation of binary features into meaningful continuous representations is essential.

- **Clustering Strategy:** A mixed approach (e.g., K-means for segmentation, GMM for probabilistic modeling) may yield better insights.
- **Validation of Results:** Multi-metric validation, such as Davies-Bouldin index, is critical to confirm clustering reliability and mitigate algorithmic biases.