Customer Segmentation Using PCA: A Comparative Analysis of GMM and K-Means Clustering

Aditi Jha and Jungho Park

Abstract

This study explores customer segmentation using Principal Component Analysis (PCA) for dimensionality reduction, followed by clustering with Gaussian Mixture Models (GMM) and K-Means. The objective is to identify and compare customer behavior patterns across clusters formed by these two clustering techniques. PCA is used to reduce the dataset to its most significant features, ensuring effective clustering while maintaining interpretability. GMM and K-Means are evaluated based on cluster characteristics, visualization, and silhouette scores, with K-Means yielding better-defined clusters and a higher silhouette score of 0.30 compared to GMM's 0.17. Key findings revealed that K-Means provided more distinct and interpretable customer segments, making it a more effective approach to segmentation in this dataset. The paper concludes with a discussion on the strengths and limitations of each method and their practical implications for customer segmentation.

Introduction and Background

Customer segmentation plays a crucial role in modern data-driven marketing and customer relationship management, enabling businesses to identify diverse customer needs and develop targeted strategies. By categorizing customers based on their behaviors, preferences, and demographics, segmentation facilitates personalized marketing campaigns, product recommendations, and enhanced customer retention. Additionally, it aids businesses in allocating resources efficiently, improving operational performance and customer satisfaction. Machine learning techniques, particularly clustering algorithms, have become essential for uncovering natural groupings within large datasets, offering deeper insights into customer behavior. This project focuses on leveraging Principal Component Analysis (PCA) for dimensionality reduction and applying Gaussian Mixture Models (GMM) and K-Means clustering to identify distinct customer segments, aiming to evaluate and compare the effectiveness of these two widely used clustering methods.

Over the years, PCA has been extensively employed for dimensionality reduction in customer segmentation, particularly for high-dimensional data. Bandyopadhyay et al. (2020) highlights PCA's ability to distill complex features like income and purchase price into principal components that retain essential variance while reducing complexity. In their study, K-Means is used to cluster these components, with the elbow method determining the optimal number of clusters aligned with customer purchasing behaviors, showcasing PCA's utility in simplifying data and K-Means' effectiveness in creating meaningful groups. Similarly, GMM has proven to be highly effective in retail, where it identifies distinct customer segments to optimize pricing and marketing strategies. Hariguna et al. (2024) demonstrates GMM's application in a retail dataset, using metrics like the Bayesian Information Criterion (BIC) and silhouette scores to refine clusters ranging from bulk buyers of low-cost items to premium product buyers. By integrating PCA's data compression with a comparison of K-Means and GMM clustering approaches, this project offers a comprehensive evaluation of these techniques, supporting data-driven customer engagement strategies and enhanced decision-making. Figure 1 shows the workflow of this comparative analysis.

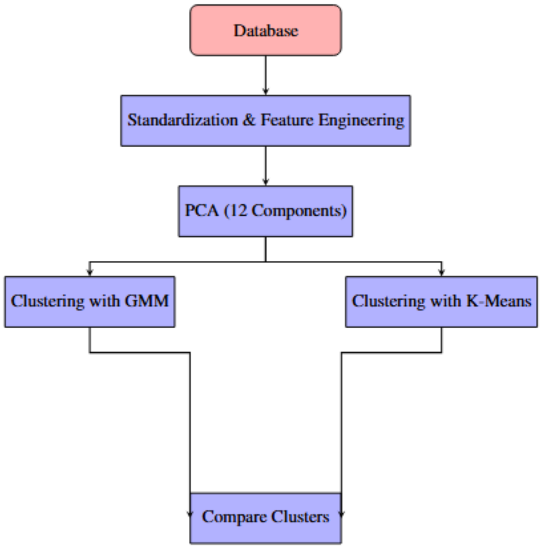

Figure 1: Workflow of this project

Data Preparation and Methodology

Data Preparation

The dataset utilized in this project originates from customer transaction records of a gift and holiday product shop. It contains diverse features capturing customer purchasing behavior, geographical distribution, and product preferences. Initial preprocessing steps included handling missing values, removing duplicates, and filtering for relevance, resulting in a clean dataset of 401,604 records, for total of 20 attributes.

To better analyze customer segmentation, feature engineering is performed to derive variables such as Total Spending, Purchase Frequency, Average Order Value, and Product Variety Index. These features are designed to encapsulate key behavioral traits of customers. Additionally, geographical indicators (e.g., Asia-Pacific, Europe) and product-related variables (e.g., mentions of “red,” “bag,” and “design”) are included to reflect broader purchase patterns.

Before applying PCA, the data is standardized to ensure all variables contributed equally, a critical step given the varying scales of the original features. After standardization, PCA is applied to reduce the dimensionality of the dataset while retaining its core variance.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining the most significant variance. PCA is employed to simplify the dataset, addressing the curse of dimensionality and enhancing computational efficiency. The method reduces correlated features into a set of orthogonal components, preserving the variance in the original data.

Steps in PCA Implementation

- Features are standardized.
- The covariance matrix of the dataset is computed, and eigenvectors/eigenvalues are derived.
- Eigenvectors are ranked based on their eigenvalues (variance captured), and the principal components corresponding to the highest eigenvalues are selected (Bandyopadhyay and Mandal 2020).

Based on the cumulative variance explained in Figure 2, 12 principal components are selected. These components captures approximately 85-90% of the total variance, balancing meaningful data retention with dimensionality reduction. Diminishing returns in variance gain after 12 components supports this choice.

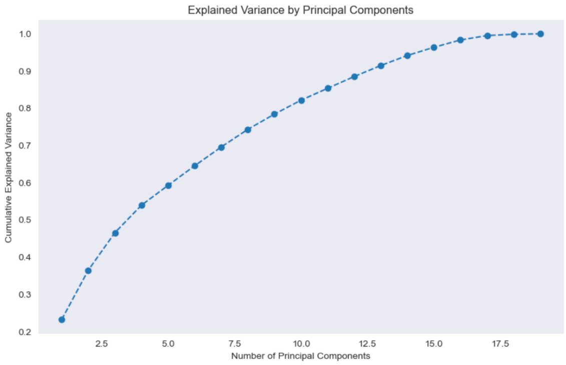

Figure 2: Explained Variance by Principal Components.

Gaussian Mixture Models (GMM)

Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that models data as a mixture of multiple Gaussian distributions, each defined by its mean and covariance. It provides a flexible and probabilistic approach to clustering, allowing clusters to overlap. GMM employs the Expectation-Maximization (EM) algorithm, which iteratively estimates the parameters of these Gaussian distributions to maximize the likelihood of the observed data. The EM algorithm alternates between assigning probabilities of data points belonging to each Gaussian component (Expectation step) and refining the parameters of these components based on these probabilities (Maximization step) (Hariguna and Chen 2024).

The probability density of a data point x in GMM is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- π_k: Mixture weight for the k^{th} Gaussian component (sums to 1).
- μ_k: Mean vector of the k^{th} Gaussian component.
- Σ_k: Covariance matrix of the k^{th} Gaussian component.
- $\mathcal{N}(x|\mu_k, \Sigma_k)$: Gaussian probability density function for component k.

This flexibility allows GMM to form soft clusters, where each point belongs to multiple clusters with varying probabilities. To determine the optimal number of clusters, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are used. BIC and AIC are statistical measures that evaluate model fit while penalizing model complexity (Hariguna and Chen 2024).

The BIC is calculated using the formula:

$$BIC = -2 \cdot \log(L) + k \cdot \log(n)$$

where L is the likelihood of the model, k is the number of parameters, and n is the number of data points.

The AIC is given by:

$$AIC = -2 \cdot \log(L) + 2k$$

Lower BIC and AIC scores indicate a better balance between model accuracy and simplicity. These criteria penalize overfitting by discouraging excessive cluster numbers. Based on the BIC/AIC plot (Figure 3), 4 clusters are selected, as they minimize both metrics and provide meaningful segmentation.

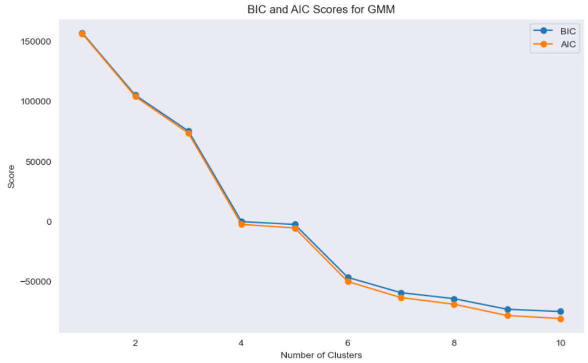

Figure 3: BIC and AIC Scores for GMM.

K-Means Clustering

K-Means is an unsupervised learning algorithm used for partitioning data into *k* clusters, minimizing the within-cluster variance. The algorithm iteratively assigns points to the nearest cluster centroid and updates centroids until convergence. The objective of K-Means is to minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\min_{j}||x_i - m_j||^2$$

- x_i: Data point.
- m_j: Centroid of cluster j.
- $||x_i - m_j||^2$: Squared Euclidean distance between x_i and m_j (Bandyopadhyay and Mandal 2020).

The Elbow Method is used to identify the number of clusters. It evaluates the inertia (sum of squared distances to cluster centers) for increasing cluster counts. Figure 4 reveals an “elbow” at 4 or 5 clusters, indicating a balance between cluster compactness and model simplicity.

To compare GMM and K-Means, three criteria are used: Silhouette Scores, Cluster Plots, and Cluster Characteristics. Silhouette scores measure how well points fit within their clusters compared to other clusters. Higher scores signify better-defined clusters.

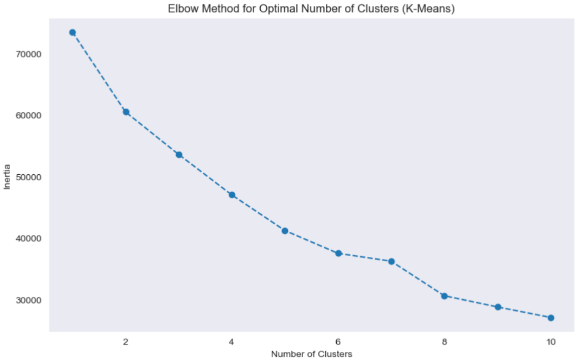

Figure 4: Elbow Method for Optimal Number of Clusters

Results

Gaussian Mixture Models (GMM)

Figure 5 illustrates the GMM clustering results projected onto the first two principal components, whereas Figure 6 shows the 3D GMM clusters for first three principal components. These plots reveal moderately separated clusters, reflecting the probabilistic nature of GMM, which accommodates overlapping data distributions. GMM identifies four distinct customer segments with unique purchasing behaviors.

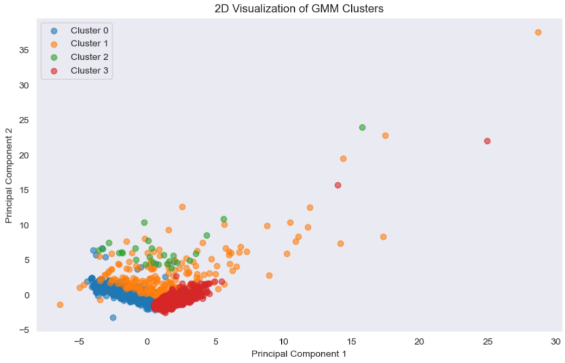

Figure 5: 2D Visualization of GMM Clusters

Cluster 0 represents moderate buyers with diverse purchasing habits. These customers exhibit moderate purchase frequency and spending, with high product variety and decent average order values. They are predominantly from ”Other Regions” but also have notable representation from Europe and North America, showing a slight preference for ”red” products.

Cluster 1 comprises high-value, focused buyers who demonstrate high product diversity and significant engagement with trendy items such as ”bag,” ”red,” ”heart,” and ”design.” They predominantly belong to the European region, with some presence in Asia-Pacific, and exhibit the

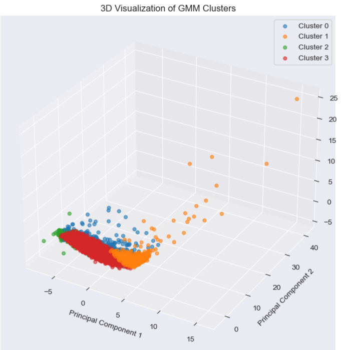

Figure 6: 3D Visualization of GMM Clusters

highest purchase frequency and spending among all clusters.

Cluster 2 consists of bulk buyers from Asia-Pacific regions. These customers have extremely high total spending and purchase quantities but low product diversity. They show minimal interest in popular product categories, reflecting a preference for large, bulk purchases with limited variety.

Cluster 3 includes minimal buyers with the lowest spending, purchase frequency, and product diversity. These customers primarily belong to Europe and ”Other Regions” and exhibit low engagement across all top product categories, representing infrequent, low-value purchases.

K-Means Clustering

Figure 7 shows the K-Means clustering results projected onto the first two principal components and Figure 8 shows the 3D K-Means clusters for the first three principal components, revealing distinctly separated clusters. This separation highlights the deterministic approach of K-Means, which creates hard boundaries between clusters. Like GMM, K-Means also identified four customer groups with distinct characteristics.

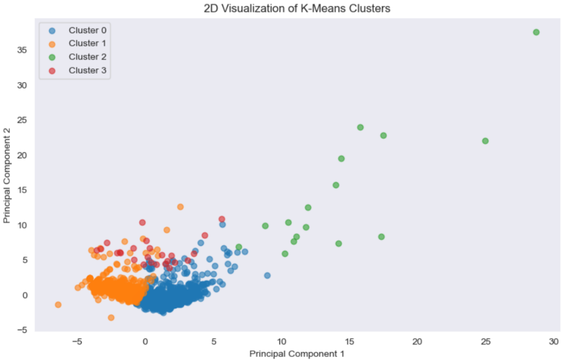

Figure 7: 2D Visualization of K-Means Clusters

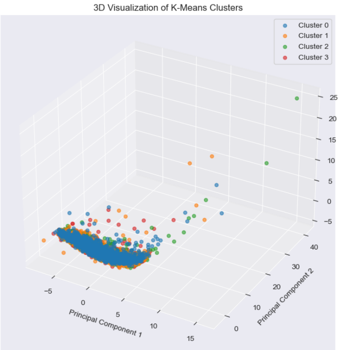

Figure 8: 2D Visualization of K-Means Clusters

Cluster 0 comprises low-spending generalists who exhibit below-average purchase frequency, spending, and quantity. They purchase fewer unique products with a low average order value and are slightly more represented in Europe and North America. None of the product categories stand out as particularly popular for this group.

Cluster 1 includes high-spending specialists who show the highest purchase frequency, spending, and quantity. They purchase a moderate variety of products and are strongly represented in ”Other Regions.” They show moderate interest in items like ”red,” ”design,” and ”vintage,” reflecting focused buying preferences.

Cluster 2 represents medium-spending Asian shoppers, who have moderate spending and quantity but low purchase frequency. They show minimal product variety but have a high average order value. This cluster is dominantly represented in Asia-Pacific regions, with a slight preference for retrospot-themed products.

Cluster 3 is made up of product-focused enthusiasts who exhibit slightly above-average purchase frequency but low spending and quantity. They show high engagement with unique products and significant interest in items like ”bag,” ”heart,” ”design,” and ”pink,” indicating a preference for stylish and diverse product categories.

Comparative Analysis

The comparison between GMM and K-Means clustering is conducted using three main criteria: silhouette scores, cluster plots, and cluster characteristics. Silhouette scores, which measure how well points fit within their clusters relative to others, reveal that K-Means outperform GMM. K-Means achieve a score of 0.30, indicating well-defined and separated clusters, while GMM scored 0.17, suggesting less distinct cluster boundaries and overlap between groups.

Cluster plots for the top two and three principal components provides visual insights into cluster separations. K-Means clusters are distinctly separated, reflecting its deterministic approach and clear boundary assignment for each data point. In contrast, GMM clusters exhibits noticeable overlap, a characteristic of its probabilistic nature, which assumes that data points can belong to multiple clusters with varying probabilities.

The behavioral summaries of the clusters further highlights the differences between the two methods. K-Means provides more straightforward and non-overlapping groupings, enabling easy interpretation of customer traits. For instance, K-Means distinctly identifies low-spending generalists, high-spending specialists, medium-spending Asian shoppers, and product-focused enthusiasts. In contrast, GMM captures overlapping behaviors, such as moderate buyers with diverse habits and bulk buyers with low diversity, offering nuanced insights but less clearly defined boundaries.

Overall, K-Means is better suited for applications requiring clear, non-overlapping segmentation, while GMM's probabilistic approach excels in scenarios needing detailed analysis of overlapping behaviors. Together, they highlight the complementary strengths and limitations of these techniques, offering valuable insights for targeted marketing strategies.

Conclusion and Discussion

This project successfully demonstrates the application of PCA for dimensionality reduction and a comparative analysis of GMM and K-Means clustering for customer segmentation. By leveraging customer transaction data, we identified meaningful customer groups, each with distinct purchasing behaviors, geographical trends, and product preferences. The results highlight K-Means as an effective tool for producing well-separated clusters, while GMM offers valuable insights into overlapping customer traits. Together, these approaches provide a comprehensive framework for understanding customer behaviors and enabling data-driven strategies in marketing and customer relationship management.

Despite its strengths, the project has limitations. The reliance on pre-defined metrics, such as silhouette scores and visual inspection of clusters, may not fully capture the nuances of real-world customer behaviors. Additionally, the dataset lacks variables such as customer demographics or seasonal patterns, which could enrich the segmentation analysis. Future work could involve integrating these additional variables, experimenting with hybrid clustering methods, or applying the framework to other domains, such as e-commerce or subscription-based businesses, to test its versatility and robustness.

References

Bandyopadhyay, S., T. S. S., and Mandal, J. K. 2020. Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society. *Innovations in Systems and Software Engineering* 17(1):45–52.

Hariguna, T., and Chen, S. C. 2024. Customer Segmentation and Targeted Retail Pricing in Digital Advertising using Gaussian Mixture Models for Maximizing Gross Income. *Journal of Digital Marketing and Digital Currency* 1(2):183–203.