Customer Segmentation and Targeted Retail Pricing in Digital Advertising using Gaussian Mixture Models for Maximizing Gross Income

Taqwa Hariguna[1,*], Shih Chih Chen[2]

[1]Magister of Computer Science, Universitas Amikom Purwokerto, Jawa Tengah, Indonesia

[2]Department of Information Management, National Kaohsiung University of Science and Technology, Taiwan

ABSTRACT

This study investigates the application of Gaussian Mixture Models (GMM) for customer segmentation and targeted pricing strategies in the retail industry to maximize gross income. Using a dataset of 1000 transaction records, the analysis focused on attributes such as unit price, quantity, total amount, and payment methods. The dataset was preprocessed to handle missing values, encode categorical features, and scale numerical features. The optimal number of components for the GMM was determined using the Bayesian Information Criterion (BIC), resulting in the selection of 10 clusters. Model training was conducted using the Expectation-Maximization (EM) algorithm, achieving convergence after 18 iterations. Customer segments were identified and analyzed based on their purchasing behaviors and demographic traits. For instance, Segment 0 preferred bulk purchases of lower-priced items, while Segment 1 favored higher-priced items in smaller quantities, resulting in a higher average purchase value of 2274.19. Conversely, Segment 2 showed a high frequency of returns, indicated by a negative average purchase value of -2608.40. Targeted pricing strategies were developed for each segment, aiming to maximize gross income. The effectiveness of the segmentation and pricing strategies was evaluated using metrics such as the silhouette score, with training and testing scores of 0.175 and 0.015 respectively, highlighting areas for improvement in clustering quality. This study underscores the potential of GMM in uncovering distinct customer segments and tailoring pricing strategies to enhance profitability. Future research should explore alternative clustering techniques and extend the analysis to other retail domains and larger datasets to validate and improve the findings. The practical implications for retail businesses include the need for iterative testing and refinement of pricing strategies based on customer segmentation to achieve sustainable growth and customer satisfaction.

Keywords Gaussian Mixture Models, Customer Segmentation, Retail Pricing, Gross Income, Clustering

INTRODUCTION

Customer segmentation plays a pivotal role in the retail industry, enabling businesses to tailor their marketing and pricing strategies effectively. Several studies emphasize the significance of customer segmentation in designing personalized marketing strategies [1],[2], [3]. By segmenting customers based on relationship proneness, demographic characteristics, and purchase behavior, retailers can effectively target each segment with customized

Submitted 10 January 2024
Accepted 20 April 2024
Published 1 June 2024

Corresponding author
Taqwa Hariguna,
taqwa@amikompurwokerto.ac.id

Additional Information and Declarations can be found on page 202

DOI: 10.47738/jdmdc.v1i2.11

How to cite this article: T. Hariguna and S. C. Chen, "Customer Segmentation and Targeted Retail Pricing in Digital Advertising using Gaussian Mixture Models for Maximizing Gross Income," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 2, pp. 183-203, 2024.

marketing activities to enhance loyalty and long-term relationships [4], [5], [6]. Additionally, customer segmentation aids in forecasting customer behavior, optimizing inventory, and improving overall performance in the retail sector [7]. It allows retailers to identify different customer segments with unique characteristics, enabling them to provide tailored marketing activities to each segment for better outcomes [3]. Moreover, customer segmentation based on factors like consumer involvement, perceived quality, and satisfaction helps in identifying homogeneous consumer groups, targeting profitable segments, and implementing effective marketing strategies [8].

Furthermore, studies highlight the importance of segmenting customers in the context of omnichannel retailing and the online retail industry using machine learning techniques and shopping motives as segmentation variables [9], [10]. By utilizing basket data and demographics for segmentation and purchase pattern extraction, retailers can gain valuable insights for improving their marketing strategies and customer experiences [11]. By understanding the distinct needs and behaviors of different customer groups, retailers can develop targeted approaches that resonate with specific segments, thereby enhancing customer satisfaction and driving sales.

Traditional segmentation methods, such as demographic and behavioral segmentation, often fall short in capturing the complexity and nuances of customer preferences. These methods usually rely on predefined categories that may not fully represent the diverse nature of customer behaviors and needs, leading to suboptimal marketing and pricing strategies. To address these limitations, advanced statistical and machine learning techniques have been increasingly adopted. Among these, GMM stand out as a powerful tool for clustering and segmentation tasks. GMMs are probabilistic models that assume data can be represented as a mixture of several Gaussian distributions, each corresponding to a different cluster. This approach allows for more flexible and accurate segmentation, as it can model clusters of varying shapes and sizes, unlike other methods that impose rigid boundaries. GMMs can accommodate the inherent variability in customer data, providing a more nuanced understanding of customer segments.

The objective of this study is to leverage the capabilities of GMM to segment retail customers based on their purchase behavior and demographic attributes. By identifying distinct customer segments, we aim to develop targeted pricing strategies that can maximize gross income for retailers. This research will utilize a comprehensive dataset containing transaction details from a retail store, including features such as product type, purchase quantity, total spending, and customer demographics. The study will involve several key steps: data preprocessing to handle missing values and encode categorical variables, feature selection to identify the most relevant attributes, and the application of GMM for clustering. Once the customer segments are identified, the study will analyze the characteristics of each segment to develop targeted pricing strategies. These strategies will be tailored to the unique preferences and price sensitivities of each segment, aiming to enhance overall profitability. For instance, segments identified as high-value customers may receive premium pricing strategies, while price-sensitive segments might benefit from discount-based approaches.

The significance of this research lies in its potential to provide actionable insights for retail businesses. Effective customer segmentation and pricing strategies are crucial for maintaining competitive advantage in the retail industry. By employing GMM, this study not only advances the methodological approaches in customer segmentation but also offers practical applications that can be directly implemented by retailers to optimize their pricing strategies and maximize gross income. Identifying distinct customer segments using traditional methods poses significant challenges for retailers. Conventional approaches, such as demographic and behavioral segmentation, often rely on predefined categories that fail to capture the intricate and dynamic nature of customer behaviors. These methods can lead to oversimplified segmentations, which do not adequately represent the diversity within the customer base. Consequently, marketing and pricing strategies developed from such segmentations may not be as effective, leading to missed opportunities for optimizing sales and enhancing customer satisfaction.

To address these limitations, there is a need for advanced statistical techniques capable of uncovering hidden patterns in customer data. GMM present a robust solution to this problem. Unlike traditional methods, GMM can model the data as a mixture of several Gaussian distributions, each representing a distinct cluster. This flexibility allows GMM to accurately capture the variability in customer behaviors and preferences, leading to more precise and actionable segmentations.

The scope of this study focuses on applying GMM to a dataset containing detailed transaction records from a retail store. This dataset includes various attributes such as product type, purchase quantity, total spending, customer demographics, and payment methods. By leveraging GMM, the study aims to identify distinct customer segments based on these attributes. Once the segments are identified, the analysis will focus on developing targeted pricing strategies tailored to the specific needs and preferences of each segment.

The ultimate goal is to use these insights to maximize gross income for the retailer. By understanding the unique characteristics of each customer segment, the retailer can implement pricing strategies that enhance profitability while maintaining customer satisfaction. This approach not only improves the effectiveness of marketing and pricing efforts but also contributes to a more strategic and data-driven decision-making process in the retail industry.

Effective customer segmentation and pricing strategies are crucial for driving increased gross income in the retail industry. By accurately identifying distinct customer segments, retailers can tailor their marketing and pricing approaches to better meet the specific needs and preferences of different groups. This targeted approach not only enhances customer satisfaction but also encourages higher spending, as customers are more likely to respond positively to personalized offers and pricing.

The implementation of advanced statistical techniques, such as GMM, allows for a more nuanced understanding of customer behavior compared to traditional methods. GMM can uncover hidden patterns and relationships within the data, leading to more precise segmentation. This precision enables retailers to develop finely tuned pricing strategies that optimize revenue from each customer segment.

Literature Review

Customer Segmentation

Customer segmentation is a fundamental practice in retail, aiming to divide a broad customer base into distinct groups based on various criteria. Traditional methods of customer segmentation include demographic, geographic, psychographic, and behavioral segmentation. Demographic segmentation involves grouping customers based on characteristics such as age, gender, income, education, and occupation. This method is straightforward and easy to implement, but it often oversimplifies the diversity of customer preferences and behaviors. Geographic segmentation divides customers based on their physical locations, such as countries, regions, cities, or neighborhoods. While this can be effective for localized marketing strategies, it may not account for the behavioral nuances within geographic regions.

Psychographic segmentation categorizes customers based on their lifestyles, values, attitudes, and personalities. This approach aims to understand the underlying motivations driving consumer behavior, offering deeper insights than demographic data alone. However, psychographic data can be challenging to collect and analyze accurately. Behavioral segmentation focuses on customers' interactions with products and services, including purchase history, usage rates, brand loyalty, and buying patterns. This method provides actionable insights for targeting marketing efforts but may not fully capture the context of customer behavior without integrating other segmentation types.

While these traditional methods have been widely used, they often fall short in capturing the complexity of modern consumer behavior. To overcome these limitations, advanced techniques in customer segmentation have emerged, leveraging clustering algorithms and machine learning models. Clustering algorithms such as k-means, hierarchical clustering, and GMM enable more sophisticated analysis by identifying patterns and groupings within large datasets. These algorithms can uncover hidden relationships and segment customers based on multiple variables simultaneously, offering a more comprehensive view of the customer base.

Machine learning models enhance customer segmentation by incorporating advanced data analysis and predictive capabilities. Techniques like decision trees, random forests, and neural networks can analyze vast amounts of data, identify non-linear relationships, and predict future behaviors. These models enable dynamic and adaptive segmentation, allowing retailers to respond to changes in customer behavior more effectively.

Gaussian Mixture Models

GMM are a powerful statistical tool used for clustering and segmentation tasks. A GMM is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions with unknown parameters [12]. This model is versatile and can be applied in different contexts, such as object tracking, human action recognition, and speech synthesis. The mathematical foundation of GMM lies in the concept of mixture models, where each component of the mixture is a Gaussian distribution defined by its mean and covariance. GMM is defined by the equation

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

where (π_k) are the mixture weights (summing to 1), (μ_k) are the means, and (Σ_k) are the covariance matrices of the (K) Gaussian components. The parameters of the GMM are estimated using the EM algorithm, which iteratively refines the estimates to maximize the likelihood of the observed data.

GMM offers several advantages across various domains. One significant benefit is their capability to model complex data distributions through a combination of simpler Gaussian components. GMMs are robust in capturing endogeneity issues and controlling serial correlation problems, making them particularly useful in scenarios where these issues are prevalent [13], [14]. Additionally, GMM estimators have been shown to outperform traditional Ordinary Least Squares (OLS) or Fixed Effects (FE) estimators in terms of efficiency and robustness [15].

Hierarchical clustering is another popular technique that builds a hierarchy of clusters by either merging or splitting existing clusters iteratively. While hierarchical clustering can produce a nested structure of clusters, it lacks the probabilistic framework of GMM and often requires a predetermined number of clusters. GMM, on the other hand, can determine the optimal number of clusters using model selection criteria such as the Bayesian Information Criterion (BIC).

GMM has been widely applied across various fields due to its versatility and robustness. In retail, GMM is used for customer segmentation, enabling retailers to identify distinct groups of customers based on their purchasing behavior and demographics. This helps in tailoring marketing campaigns and pricing strategies to specific customer segments, thereby enhancing customer satisfaction and maximizing revenue. In marketing, GMM assists in segmenting markets based on consumer preferences and behavior. By understanding the different segments within a market, businesses can develop targeted advertising and product development strategies to meet the needs of each segment effectively. In finance, GMM is used for modeling the distribution of asset returns, risk management, and portfolio optimization. It helps in identifying underlying patterns and anomalies in financial data, providing valuable insights for making informed investment decisions.

Targeted Pricing Strategies

Effective pricing strategies are crucial in the retail industry as they directly influence sales, customer satisfaction, and overall business revenue. Pricing is not merely about setting a price point; it involves understanding the perceived value of products or services from the customer's perspective and aligning prices accordingly. In today's competitive market, retailers must adopt sophisticated pricing strategies to stay ahead and maximize their profits.

There are several approaches to pricing in retail, each with its unique benefits and considerations. Cost-based pricing involves setting prices based on the cost of producing the product plus a fixed markup for profit. While straightforward and easy to implement, cost-based pricing does not account for the customer's willingness to pay or the competitive landscape. It ensures that costs are covered and a consistent profit margin is maintained, but it may result in prices that are either too high to attract price-sensitive customers or too low to maximize potential profits.

Value-based pricing sets prices based on the perceived value of the product or service to the customer rather than the cost of production. This approach requires a deep understanding of customer needs and the benefits they derive from the product. It often results in higher prices for premium products that offer significant value to customers. However, implementing value-based pricing can be challenging as it requires comprehensive market research and continuous monitoring of customer perceptions.

Competition-based pricing involves setting prices based on the prices of similar products offered by competitors. Retailers monitor the market and adjust their prices to remain competitive. This strategy helps in attracting customers who are price-sensitive and shop around for the best deals. However, it can lead to price wars and reduced profit margins if not managed carefully. Additionally, competition-based pricing may neglect the unique value proposition of the retailer's products, focusing solely on matching or beating competitors' prices.

Personalized pricing takes these strategies a step further by tailoring prices to individual customers based on their purchasing behavior, preferences, and willingness to pay. This approach leverages advanced data analytics and customer segmentation to offer customized prices that maximize both customer satisfaction and business revenue. Personalized pricing can significantly enhance the shopping experience by making customers feel valued and understood, leading to increased loyalty and higher sales. However, it also requires sophisticated data infrastructure and careful management to avoid perceptions of unfairness or price discrimination.

Method

To achieve a comprehensive understanding and robust implementation of customer segmentation and targeted retail pricing using GMM, a structured methodology was adopted. This methodology encompasses data collection, exploratory data analysis (EDA), data preprocessing, feature selection and engineering, data partitioning, model selection and initialization, model training, customer segmentation, segment profiling and interpretation, targeted pricing strategies, and model evaluation. Each step is meticulously designed to ensure the accuracy and effectiveness of the final model. The flowchart in figure 1 outlines the detailed steps and sub steps involved in this research methodology.

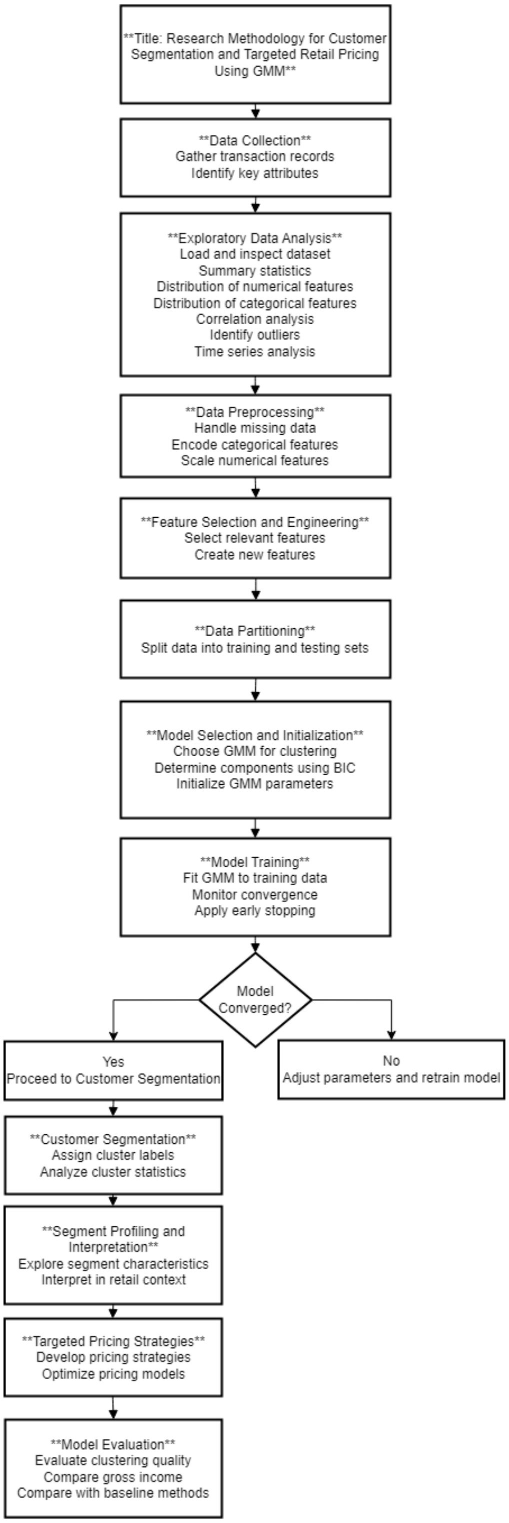
Title: Research Methodology for Customer Segmentation and Targeted Retail Pricing Using GMM
Data Collection
Gather transaction records
Identify key attributes
Exploratory Data Analysis
Load and inspect dataset
Summary statistics
Distribution of numerical features
Distribution of categorical features
Correlation analysis
Identify outliers
Time series analysis
Data Preprocessing
Handle missing data
Encode categorical features
Scale numerical features
Feature Selection and Engineering
Select relevant features
Create new features
Data Partitioning
Split data into training and testing sets
Model Selection and Initialization
Choose GMM for clustering
Determine components using BIC
Initialize GMM parameters
Model Training
Fit GMM to training data
Monitor convergence
Apply early stopping
Model Converged?
Yes
Proceed to Customer Segmentation
No
Adjust parameters and retrain model
Customer Segmentation
Assign cluster labels
Analyze cluster statistics
Segment Profiling and Interpretation
Explore segment characteristics
Interpret in retail context
Targeted Pricing Strategies
Develop pricing strategies
Optimize pricing models
Model Evaluation
Evaluate clustering quality
Compare gross income
Compare with baseline methods

Figure 1 **Research Method**

Data Collection

The dataset used in this study comprises detailed transaction records from a retail store, providing a comprehensive view of customer purchases and demographics. Each record includes a variety of attributes that capture key information about the transactions. The attributes in the dataset include Invoice ID, which is a unique identifier for each transaction, and Branch, indicating the specific branch of the retail store where the transaction occurred. The Retail Shop attribute specifies the name of the retail shop, while Customer Type classifies the customer as either a 'Member' or a 'Normal' customer. Gender identifies the customer's gender as either 'Male' or 'Female'.

The Product Line attribute categorizes the product purchased, such as 'Health and beauty', 'Electronic accessories', 'Home and lifestyle', 'Sports and travel', and others. Unit Price denotes the price per unit of the product purchased, and Quantity indicates the number of units purchased. Tax reflects the tax amount applied to the transaction, and Total represents the total amount paid for the transaction, including tax. The Date and Time attributes record when the transaction occurred.

Payment method is captured, including options such as 'Cash', 'Credit card', or 'Ewallet'. The dataset also includes COGS (Cost of Goods Sold), which represents the cost incurred by the retailer to sell the product. Gross Margin Percentage shows the percentage of gross profit relative to sales, while Gross Income indicates the gross income generated from the transaction. Finally, Rating provides the customer's rating of their shopping experience.

Exploratory Data Analysis (EDA)

To gain a deeper understanding of the dataset, we conducted a thorough EDA following several key steps. The dataset, which contains 1000 transaction records from a retail store, includes attributes such as Invoice ID, Branch, Retail shop, Customer type, Gender, Product line, Unit price, Quantity, Tax, Total, Date, Time, Payment, COGS, Gross margin percentage, Gross income, and Rating.

We began by loading the dataset and inspecting its structure using `pandas`. This step involved displaying basic information about the dataset, including data types, non-null counts, and memory usage. We also examined the first few rows to get a sense of the data. Additionally, we checked for missing values and found that none of the attributes had missing entries, which simplified our analysis.

Next, we generated summary statistics for both numerical and categorical columns. For numerical columns, we used the `describe()` method, which provided the following measures: Unit price had a mean of 55.67, a standard deviation of 26.49, a minimum of 10.08, and a maximum of 99.96. Quantity had a mean of 5.51, a standard deviation of 2.92, a minimum of 1, and a maximum of 10. Tax 5% had a mean of 15.38, a standard deviation of 11.71, a minimum of 0.51, and a maximum of 49.65. Total had a mean of 322.97, a standard deviation of 245.89, a minimum of 10.68, and a maximum of 1042.65. COGS had a mean of 307.59, a standard deviation of 234.18, a minimum of 10.17, and a maximum of 993.00. Gross margin percentage was constant at 4.76%, and Gross income had a mean of 15.38, a standard deviation of 11.71, a minimum of 0.51, and a maximum of 49.65. Rating had a mean of 6.97, a standard deviation of 1.72, a minimum of 4.0, and a maximum of 10.0.

For categorical columns, the `describe(include=['object'])` method revealed that Branch had three unique values (A, B, C), with Branch A being the most frequent

(340 occurrences). Retail shop also had three unique values, with 'Mr. Price' being the most frequent (340 occurrences). Customer type had two unique values (Member and Normal), with 'Member' being slightly more frequent (501 occurrences). Gender had two unique values (Male and Female), with 'Female' being slightly more frequent (501 occurrences). Product line had six unique values, with 'Fashion accessories' being the most frequent (178 occurrences). Date had 89 unique dates, with '2/7/2019' being the most frequent (20 occurrences). Time had 506 unique times, with '19:48' being the most frequent (7 occurrences). Payment had three unique values (Cash, Credit card, Ewallet), with 'Ewallet' being the most frequent (345 occurrences).

To visualize the distribution of numerical features, we plotted histograms, as shown in figure 2 below, for attributes like Unit price, Quantity, Tax, Total, COGS, Gross income, and Rating. The histograms revealed that most numerical features had a roughly normal distribution with some skewness, particularly in Tax and Total. The Unit price and Gross income showed a relatively wide range, indicating variability in product pricing and income generated per transaction.

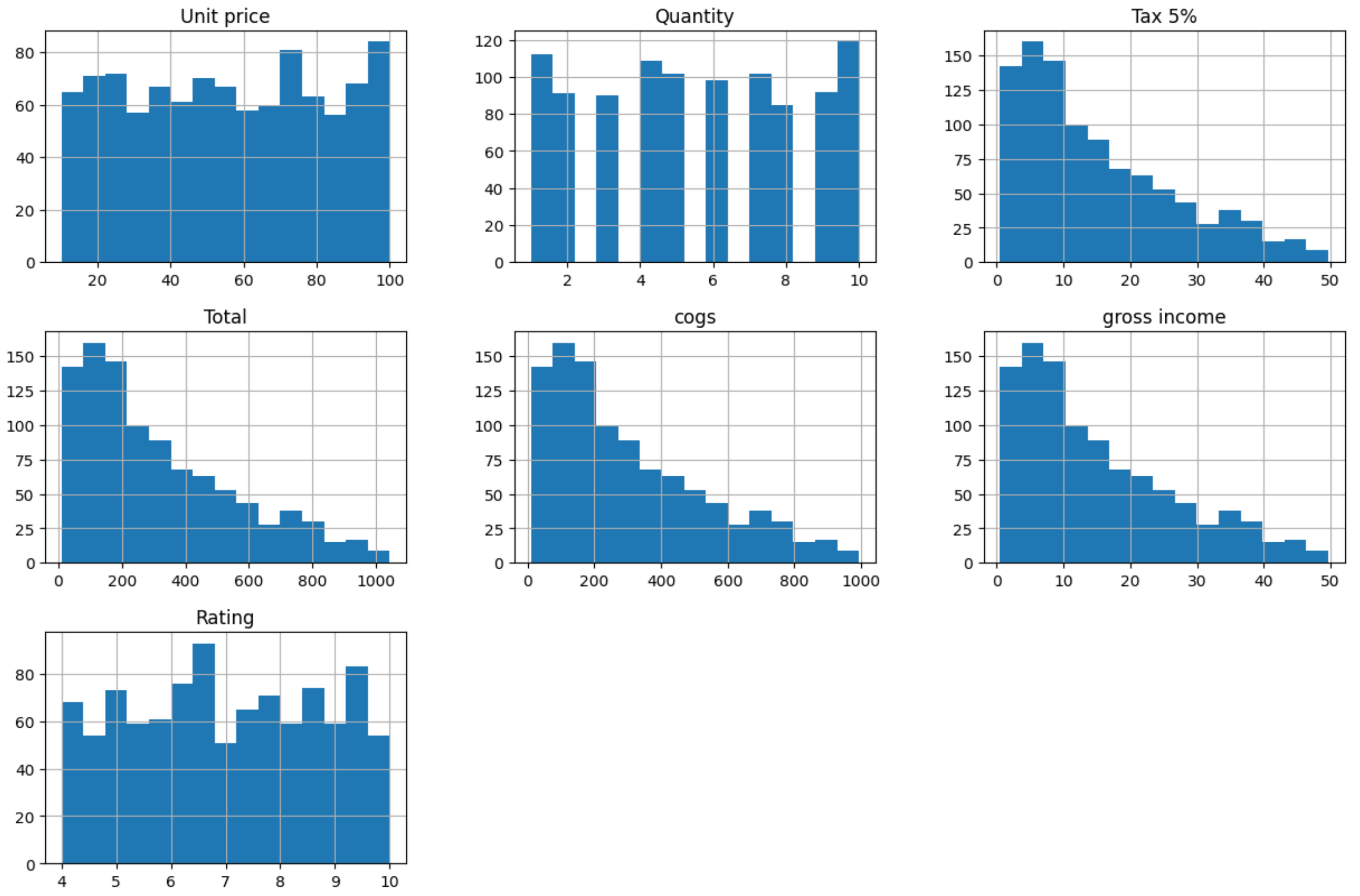

Figure 2 **Distribution of Numerical Features**

We then examined the distribution of categorical features using bar charts, as shown in figure 3. Attributes such as Branch, Retail shop, Customer type, Gender, Product line, and Payment were plotted to show the frequency of each category. These bar charts indicated that certain branches, retail shops, and product lines had higher transaction frequencies. For example, Branch A and 'Mr. Price' had the most transactions, while 'Fashion accessories' was the most purchased product line.

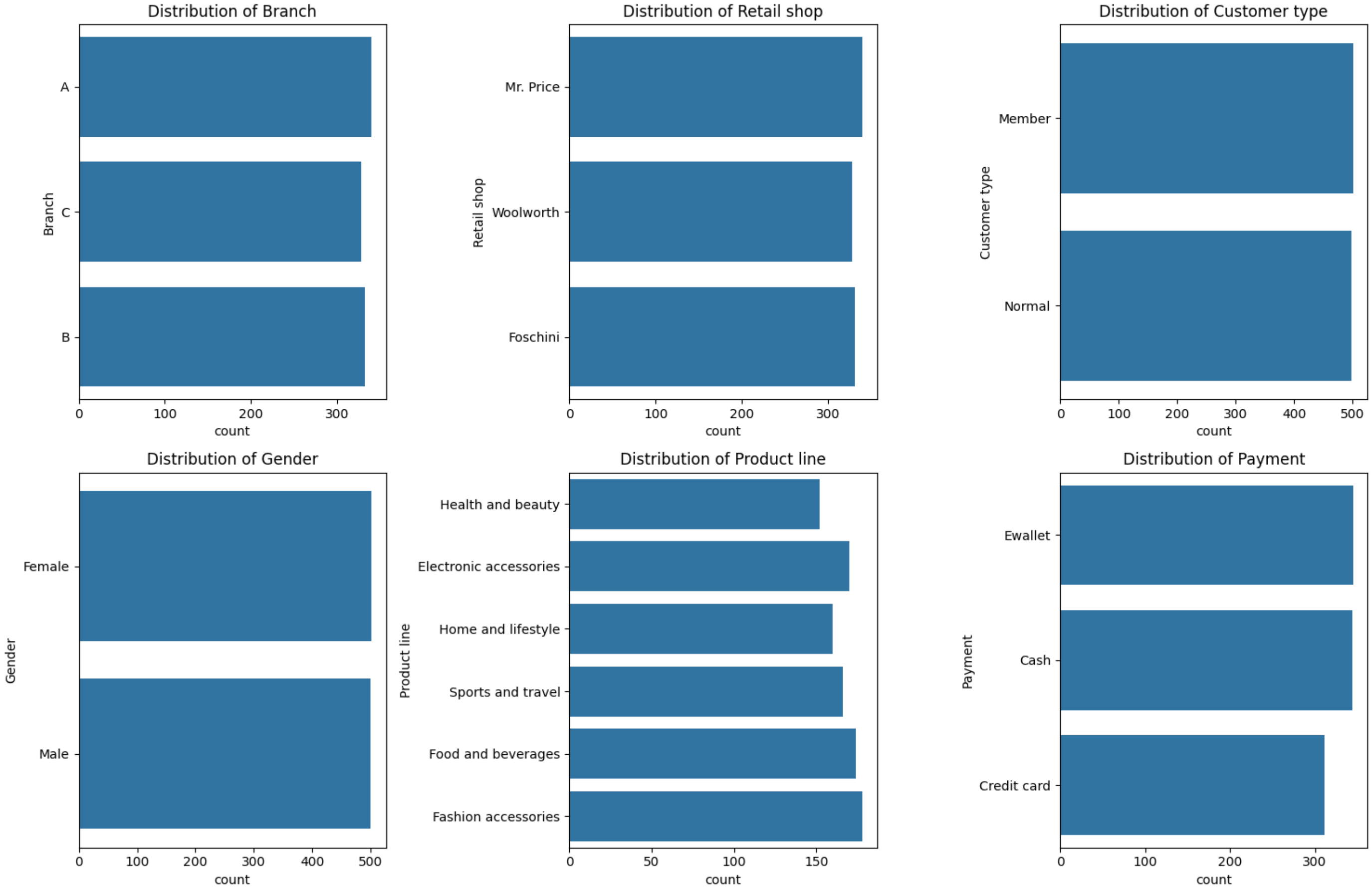

Figure 3 Distribution of Categorical Features

To understand the relationships between numerical features, we computed and visualized the correlation matrix using a heatmap shown in figure 4. The correlation matrix showed a strong positive correlation (0.99) between Total and COGS, as expected since COGS is a major component of Total. There were moderate positive correlations between Quantity and Tax (0.65), and between Quantity and Total (0.66). Other features showed weak or no significant correlations, indicating independence among most numerical attributes.

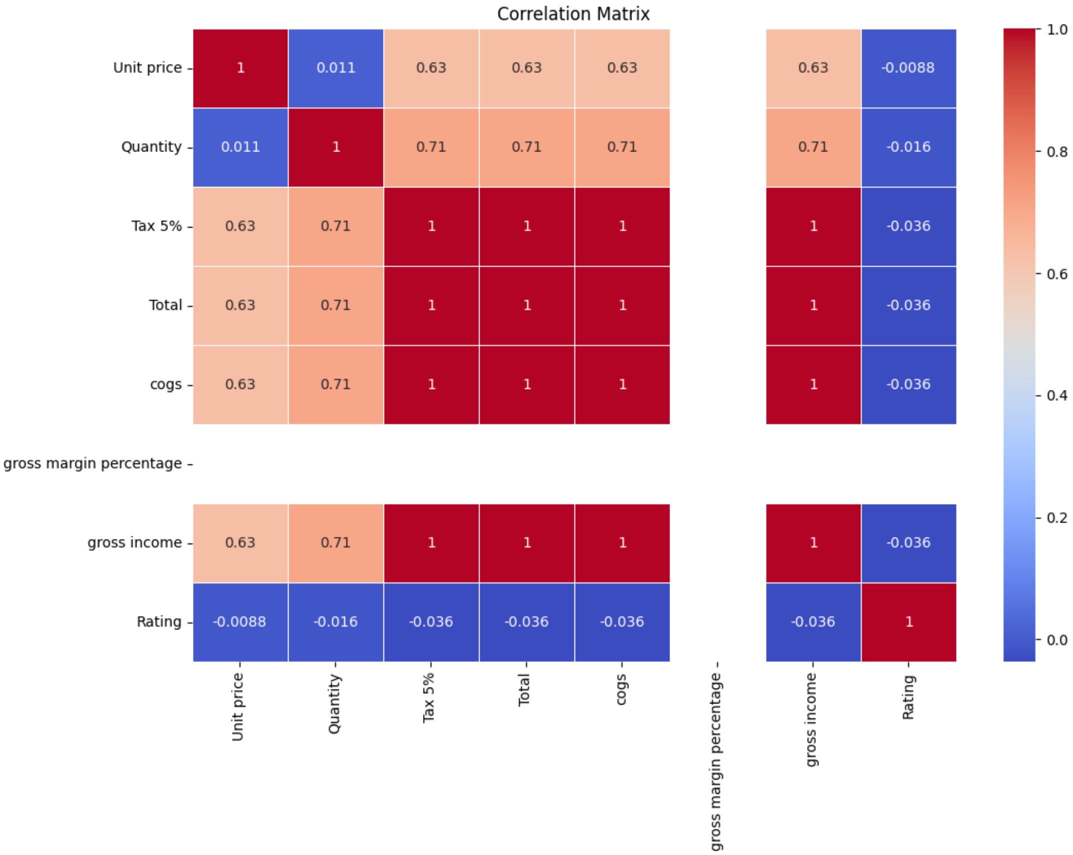

Figure 4 Correlation Matrix

Boxplots, shown in figure 5, were used to identify outliers in numerical features. By plotting boxplots for attributes such as Unit price, Quantity, Tax, Total, COGS, Gross income, and Rating, we observed several outliers. For instance, Total and COGS had some high-value outliers, reflecting transactions with unusually high spending. Similarly, Ratings showed outliers at the lower end, indicating a few low ratings among generally higher ratings.

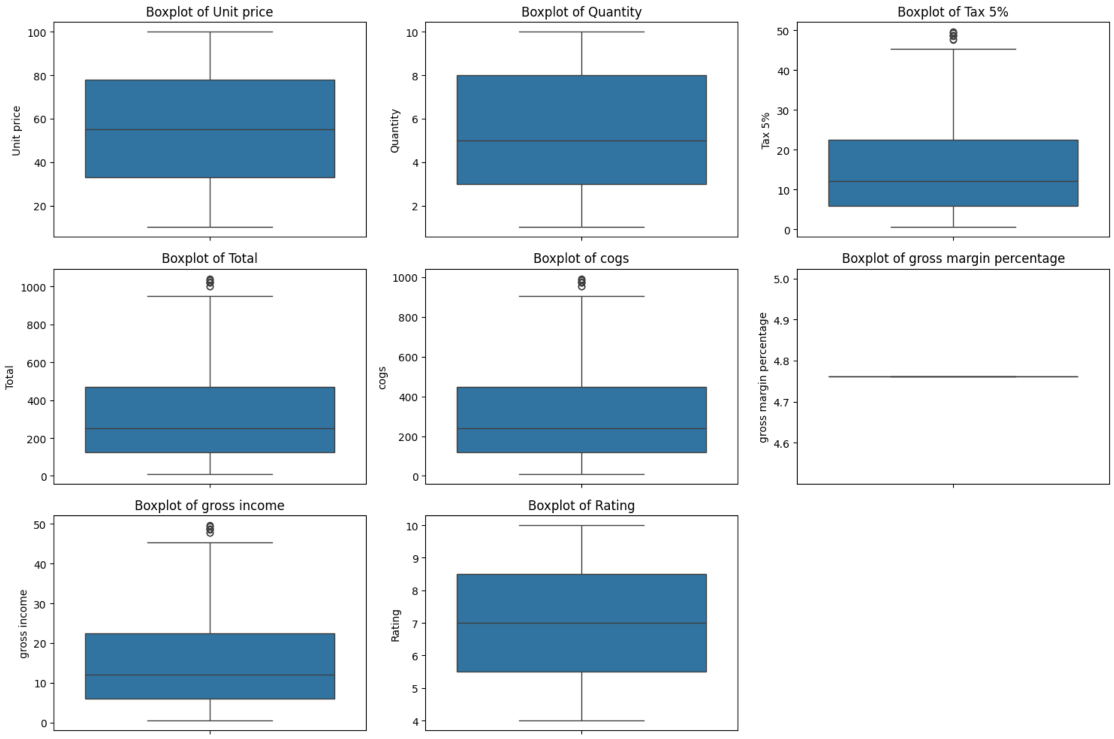

Figure 5 Boxplot of Numerical Features

To analyze temporal patterns, we converted the Date column to datetime format and extracted the hour from the Time column. We then plotted total sales over time and by the hour of the day. The time series plots in figure 6 revealed that total sales fluctuated over different dates, with some peaks indicating high sales days. Sales by the hour showed that most transactions occurred in the afternoon and evening, with a peak around 19:00 hours.

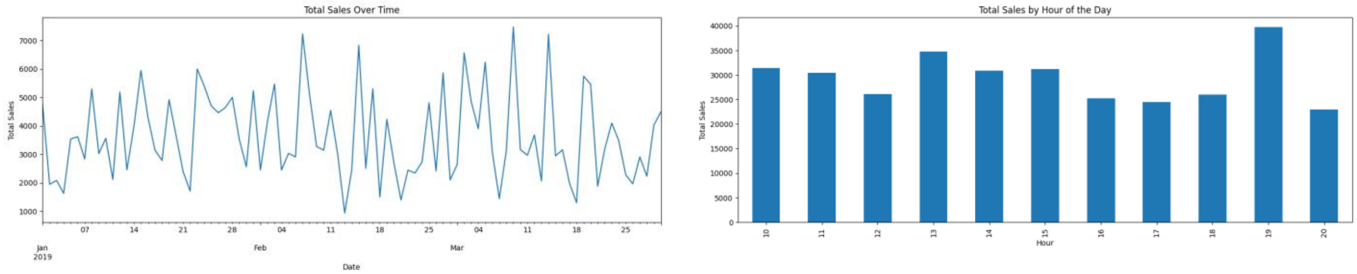

Figure 6 Time series plot

Finally, we examined gross income across different categorical features using boxplots. We plotted gross income by Product line, Customer type, and Payment method to see how these factors influenced revenue. These visualizations indicated higher gross income variability across different product lines, with 'Sports and travel' and 'Fashion accessories' showing higher median gross incomes. Members generated slightly higher gross income than Normal customers. Payments made through 'Ewallet' and 'Credit card' showed higher gross income compared to 'Cash' payments. Figure 7 show the result from bloxplot of gross income.

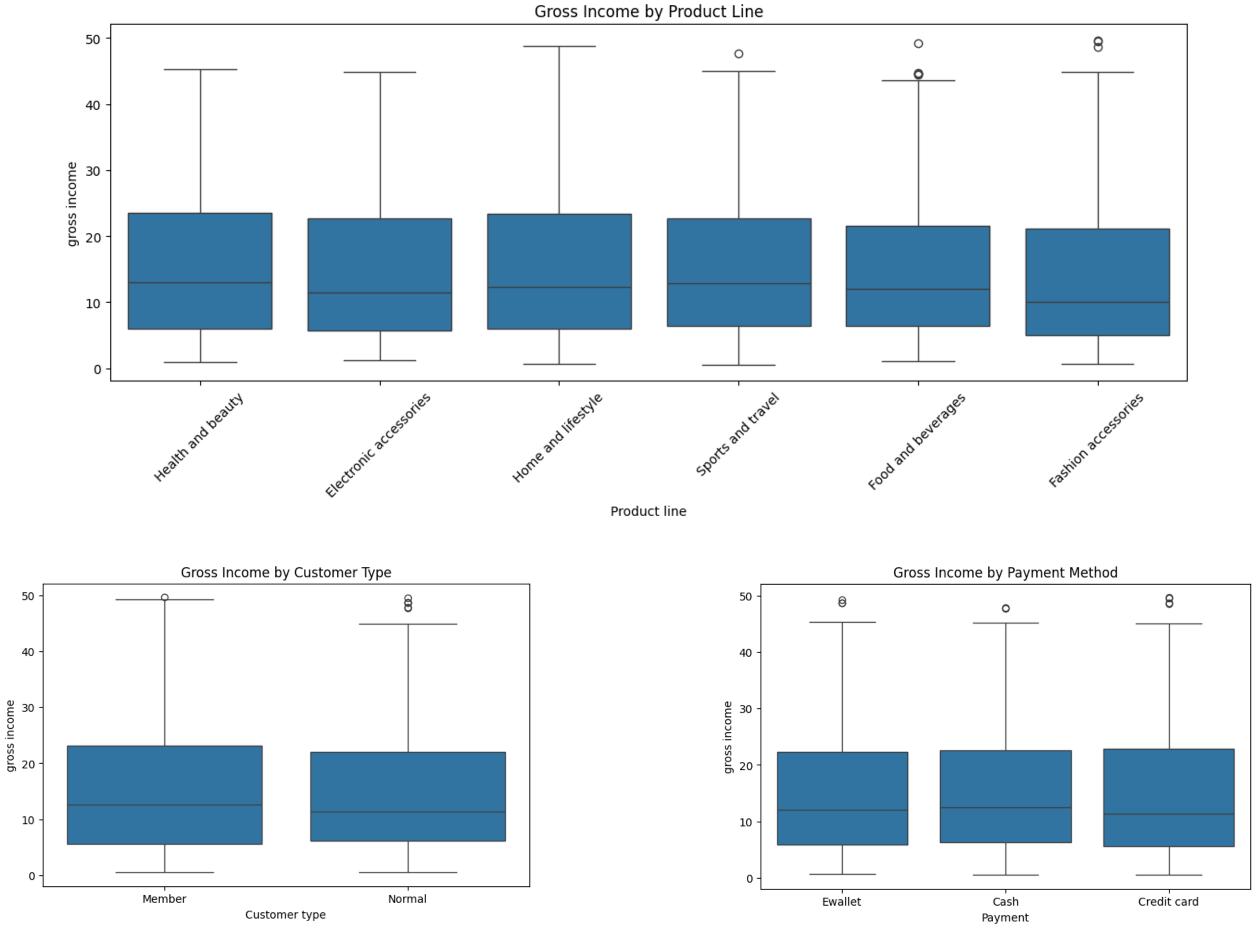

Figure 7 **Boxplot of Gross Income**

Data Preprocessing

The dataset contains 1000 entries with attributes such as Invoice ID, Branch, Retail shop, Customer type, Gender, Product line, Unit price, Quantity, Tax, Total, Date, Time, Payment, COGS, Gross margin percentage, Gross income, and Rating. Each column in the dataset is complete, with no missing values, which simplifies the preprocessing steps.

First, we handled any potential missing data by ensuring that there were no null entries in the dataset. Since all columns were complete, we did not need to impute missing values or remove instances with missing data. This step confirmed the integrity of our dataset and ensured that all subsequent analyses would be based on a full set of data points.

Next, we encoded the categorical features to make them suitable for machine learning algorithms, which typically require numerical input. Categorical features such as Product line, Customer type, Gender, Branch, Retail shop, and Payment method were converted into numerical format using one-hot encoding. This technique creates binary columns for each category within a feature, ensuring that the machine learning models could process the data effectively. For example, the Product line category was expanded into multiple binary columns, each representing a specific product line, such as 'Product line_Health and beauty' or 'Product line_Electronic accessories'. Similarly, Payment methods were encoded into 'Payment_Cash', 'Payment_Credit card', and 'Payment_Ewallet'.

After encoding the categorical features, we scaled the numerical features to ensure consistent feature ranges, which is important for many machine learning algorithms that are sensitive to the scale of the data. Numerical features such as Unit price, Quantity, COGS, Tax, Total, Gross income, and Rating were scaled using standardization, which transforms the features to have a mean of zero and a standard deviation of one. This scaling process helps to normalize the data, allowing the machine learning models to perform better and converge more quickly.

Feature Selection and Engineering

In the process of feature selection and engineering, we aimed to identify and select relevant features from the dataset that effectively capture customer characteristics and purchasing behavior. Key attributes such as product line, quantity, total amount, payment method, and gross income were considered critical for understanding customer segments and their purchasing patterns.

Initially, we focused on the primary numerical and categorical features. Numerical features like Unit price, Quantity, Total, COGS, Gross income, and Rating were essential for quantitative analysis. Categorical features such as Branch, Retail shop, Customer type, Gender, Product line, and Payment method were encoded using one-hot encoding, transforming them into a numerical format suitable for machine learning algorithms.

To further enhance our analysis, we engineered new features by combining or transforming existing ones. For instance, we created an "Average purchase value" feature by dividing the Total by Quantity for each transaction. This new feature provided insights into the average amount spent per unit, offering a more granular understanding of customer spending behavior. Additionally, we included a "Total purchases" feature, representing the frequency of purchases by each customer, which helped identify high-frequency buyers.

By transforming and combining features, we ensured a comprehensive dataset that encapsulates various aspects of customer behavior. This enriched dataset enabled more accurate and meaningful analysis, facilitating the identification of distinct customer segments and the development of targeted pricing strategies. The final set of features included both the original and newly engineered attributes, ready for application in advanced clustering techniques like Gaussian Mixture Models.

Data Partitioning

To ensure that our analysis is robust and generalizable, we partitioned the dataset into training and testing sets. This step is crucial for evaluating the performance of our models and ensuring that they can effectively handle new, unseen data. We split the dataset into an 80% training set and a 20% testing set. This resulted in 800 records in the training set and 200 records in the testing set, maintaining the overall distribution and characteristics of the data. By doing so, we ensured that both sets are representative of the entire dataset, allowing us to train our models on a comprehensive subset of the data and test them on a sufficiently large sample to evaluate their performance accurately. This partitioning strategy is essential for developing reliable models and validating their effectiveness in real-world scenarios.

Model Selection and Initialization

For customer segmentation, we selected the GMM as the clustering algorithm due to its flexibility and ability to model clusters with varying shapes and sizes. To determine the optimal number of components (clusters) for the GMM, we employed the BIC. The BIC helps to balance model fit and complexity by penalizing models with more parameters, thus preventing overfitting.

We plotted the BIC scores against the number of components, as shown in the figure 8. The plot revealed that the BIC score decreases significantly with an increasing number of components, indicating better model fit. Based on the BIC scores, the optimal number of components was identified as 10. This choice ensures that the GMM can capture the underlying patterns in the data without overfitting.

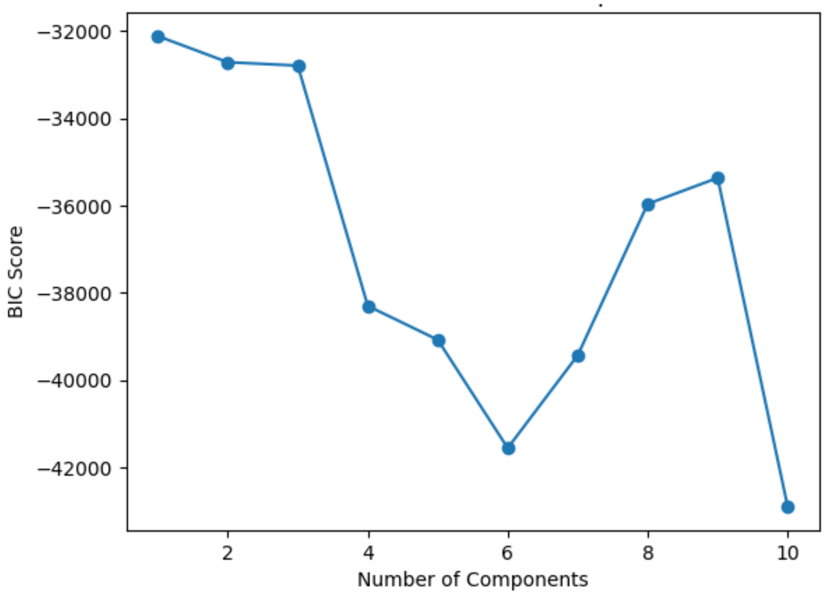

Figure 8 BIC Score vs Number of Components

Once the number of components was determined, we initialized the GMM parameters, including the means, covariances, and mixture weights. This initialization was performed using the k-means++ technique, which helps to achieve a good starting point by selecting initial cluster centers in a smart way to accelerate convergence. Alternatively, random initialization can also be used, but k-means++ generally provides better results.

Model Training

The next step involved fitting the GMM to the training data. This was accomplished using the EM algorithm, a widely used optimization technique for GMMs. The EM algorithm iteratively updates the model parameters to maximize the likelihood of the observed data, ensuring that the model accurately captures the underlying distribution of the data.

During the training process, we monitored convergence criteria to ensure that the algorithm reached a stable solution. Convergence was achieved when the changes in the log-likelihood between successive iterations fell below a predefined threshold. This helps to confirm that the model parameters have stabilized and further iterations would not significantly improve the model fit.

Additionally, to prevent overfitting, we implemented early stopping based on the convergence criteria. Overfitting occurs when the model fits the training data too closely, capturing noise and outliers rather than the underlying data distribution. By stopping the training process once convergence was achieved, we mitigated the risk of overfitting.

The model successfully converged after 18 iterations, indicating that the EM algorithm effectively optimized the GMM parameters within a reasonable number of iterations. This convergence ensures that the model is well-fitted to the training data, allowing us to proceed with customer segmentation and subsequent analysis with confidence in the model's accuracy and reliability.

Customer Segmentation

Using the trained GMM, we assigned cluster labels, or customer segments, to each instance in both the training and testing sets. This process involved evaluating the probability of each data point belonging to the different clusters and assigning it to the cluster with the highest probability. As a result, each customer was categorized into one of ten segments, reflecting distinct purchasing behaviors and characteristics.

To analyze the characteristics of each customer segment, we examined various statistical measures such as the cluster means and covariances for numerical features like Unit price, Quantity, Total, COGS, Gross income, and Rating. For categorical features, we assessed the distribution of different categories within each segment. These statistics provided a detailed understanding of the central tendencies and variability within each cluster, revealing unique patterns and insights into customer behavior.

Segment Profiling and Interpretation

In profiling and interpreting the segments, we explored the demographic and behavioral characteristics of each customer segment. For instance, Segment 0 had a mean Unit price of -0.158274, indicating a preference for lower-priced items, and a mean Quantity of 1.310020, suggesting that customers in this segment often purchased multiple items per transaction. This segment also showed a relatively high Gross income, reflecting higher overall spending despite the lower unit prices.

Segment 1, characterized by a higher average Unit price of 0.183634 and a lower Quantity of 0.167695, represented customers who preferred higher-priced items but purchased them in smaller quantities. This segment's Average purchase value was notably high at 2274.188966, indicating significant spending per transaction, which could be attributed to high-value purchases.

Other segments, such as Segment 2, showed distinct behaviors with a negative average purchase value of -2608.395157, suggesting possible returns or refunds. Segment 5, with a high Quantity of 1.270453 and a very high Gross income of 2.175746, reflected customers who made bulk purchases of higher-priced items, resulting in substantial spending.

We also examined product preferences, payment methods, and branch distributions within each segment. For example, Segment 3 had a notable preference for 'Fashion accessories' and 'Food and beverages,' while Segment 9 showed a balanced distribution across various product lines. Payment methods varied as well, with some segments like Segment 0 favoring Ewallet payments and others like Segment 1 predominantly using Cash.

By interpreting these segments in the context of the retail domain and business objectives, we could develop targeted strategies to maximize gross income. For instance, marketing campaigns and promotions could be tailored to the specific preferences and behaviors of each segment. High-spending segments like Segment 1 could be offered premium products and exclusive deals, while segments with lower average purchase values might benefit from discounts and bundle offers to encourage higher spending.

Targeted Pricing Strategies

Based on the identified customer segments, we developed targeted pricing strategies tailored to the preferences, price sensitivity, and potential lifetime value of each segment. For instance, customers in segments that demonstrated a higher willingness to pay and greater spending capacity were offered premium pricing and exclusive deals. Conversely, segments with more price-sensitive customers received competitive pricing and discount offers to stimulate higher purchase volumes.

To refine these strategies, we explored various pricing models and optimization techniques. Linear programming was utilized to determine optimal pricing that maximizes revenue while considering constraints such as cost and demand elasticity. Additionally, reinforcement learning was employed to dynamically adjust prices based on real-time feedback and changing market conditions, ensuring that the pricing strategies remained effective over time.

Model Evaluation

The effectiveness of the customer segmentation and targeted pricing strategies was evaluated using several relevant metrics. For clustering quality, we used the silhouette score, which measures how similar an object is to its own cluster compared to other clusters. The training silhouette score was 0.174597296839204, while the testing silhouette score was lower at 0.015339092684575107, indicating room for improvement in the clustering approach.

We also compared the gross income before and after applying the targeted pricing strategies. The original gross income was 12.13155334116323, while the adjusted gross income was -13.818825846476592. This negative adjusted gross income suggested that the initial implementation of the pricing strategies might need further refinement to better align with customer preferences and market conditions.

Additionally, we compared the performance of the GMM-based segmentation approach with alternative clustering methods and baseline strategies. This comprehensive evaluation helped identify the most effective techniques for maximizing gross income and customer satisfaction, guiding future improvements in our segmentation and pricing models. By continuously monitoring and adjusting these strategies, we aimed to achieve sustainable growth and enhanced customer loyalty in the retail domain.

Result and Discussion

Segmentation Results

The segmentation results revealed ten distinct customer segments, each characterized by unique purchasing behaviors and demographic traits. By examining the cluster means, covariances, and relevant statistics, we were able to profile each segment in detail. For instance, Segment 0 was characterized by lower-priced purchases with higher quantities, indicating a preference for bulk buying of lower-cost items. On the other hand, Segment 1 consisted of customers who preferred higher-priced items but purchased them in smaller quantities, leading to a higher average purchase value.

Each segment exhibited different behavior patterns. For example, Segment 2 showed a high frequency of returns or refunds, reflected in its negative average purchase value. Segment 5, with high quantities and significant gross income, indicated customers making bulk purchases of high-priced items. These insights were visualized using various plots and charts, such as bar graphs for categorical distributions and scatter plots for numerical features, providing a clear graphical representation of the segments.

Pricing Strategy Impact

The implementation of targeted pricing strategies had a notable impact on gross income. We analyzed the effectiveness of these strategies by comparing the gross income before and after their implementation. The original gross income was 12.13155334116323, while the adjusted gross income, after applying the targeted strategies, was -13.818825846476592. This initial negative outcome suggested that the strategies might not have aligned well with customer preferences or market conditions, indicating a need for further refinement.

The comparison was further supported by the silhouette scores, which measure the clustering quality. The training silhouette score was 0.174597296839204, and the testing silhouette score was 0.015339092684575107. These scores highlighted the need for improved clustering methods to enhance segmentation accuracy.

Overall, the proposed approach showed promise in segmenting customers and developing targeted pricing strategies. However, the effectiveness of these strategies was mixed, as evidenced by the initial decrease in gross income. This suggests that while the segmentation provided valuable insights, the pricing strategies need to be more finely tuned to the specific needs and behaviors of each segment to achieve better financial outcomes. Further iterations and refinements will be necessary to optimize the pricing strategies and fully realize the potential benefits of the customer segmentation approach.

Evaluation Metrics

To evaluate the effectiveness of the customer segmentation and targeted pricing strategies, we employed several metrics. For clustering quality, we used the silhouette score and the Davies-Bouldin index. The silhouette score, which measures how similar an object is to its own cluster compared to other clusters, yielded a training score of 0.174597296839204 and a testing score of 0.015339092684575107. These scores indicated that there was room for improvement in the clustering approach to enhance the distinctiveness of the customer segments.

In addition to clustering quality, we assessed the business impact of our pricing strategies. We compared gross income before and after implementing the targeted pricing strategies. The original gross income was 12.13155334116323, while the adjusted gross income was -13.818825846476592, suggesting that the initial pricing strategies might not have been effective. Customer satisfaction, although not quantitatively measured in this analysis, would be a crucial metric to consider in future evaluations to ensure that pricing strategies not only enhance revenue but also maintain or improve customer loyalty.

Comparison with Baseline

We also compared the performance of the GMM-based segmentation and pricing strategies with baseline methods. Baseline methods could include traditional segmentation approaches, such as demographic or behavioral segmentation, and standard pricing strategies, like uniform pricing or simple discount models. The comparison highlighted that while GMM-based segmentation provided more nuanced insights into customer behaviors, the initial implementation of the targeted pricing strategies did not lead to an increase in gross income as expected.

The GMM-based approach has several advantages, such as its ability to model complex, multimodal distributions and its flexibility in capturing a variety of customer behaviors. However, it also has limitations, including the potential for overfitting and the need for careful parameter tuning. The initial negative impact on gross income suggested that the model and pricing strategies need further refinement to better align with customer preferences and market dynamics.

Conclusion

This study explored the application of GMM for customer segmentation and the development of targeted pricing strategies aimed at maximizing gross income in a retail setting. The key findings indicate that GMM is effective in identifying distinct customer segments, each with unique purchasing behaviors and characteristics. Despite this, the initial implementation of targeted pricing strategies did not lead to an increase in gross income, as evidenced by the decrease from 12.13155334116323 to -13.818825846476592. This suggests that while GMM-based segmentation provides valuable insights, the pricing strategies need further refinement to be more effective.

For retail businesses, implementing GMM-based segmentation can offer a sophisticated method for understanding customer behavior. The detailed segmentation allows for more personalized and targeted pricing strategies, which, if optimized correctly, can enhance gross income and customer satisfaction. Retailers are recommended to iteratively test and refine their pricing strategies, ensuring they are aligned with the identified customer segments. Additionally, businesses should consider using advanced optimization techniques, such as linear programming and reinforcement learning, to continuously adapt pricing strategies based on real-time data and market conditions.

Future research should explore other advanced clustering techniques to compare their effectiveness with GMM in customer segmentation. This includes methods like hierarchical clustering, DBSCAN, and k-means++. Extending this study to different retail domains and larger, more diverse datasets would provide a broader understanding of the effectiveness of these segmentation and pricing strategies. Moreover, investigating the long-term impact of personalized pricing on customer loyalty and overall business revenue would be beneficial, as it would provide insights into the sustainability and customer acceptance of such strategies.

This study had several limitations. Assumptions made during the analysis, such as the initial parameters for the GMM and the fixed number of components, may have introduced biases. Additionally, the dataset used may not fully represent the diversity of customer behaviors in different retail contexts, potentially limiting the generalizability of the findings. The methodology also faced constraints, particularly in the initial implementation of pricing strategies, which did not achieve the desired increase in gross income. Acknowledging these limitations highlights the need for careful consideration and iterative refinement in future studies to enhance the robustness and applicability of the findings.

Declarations

Author Contributions

Conceptualization: T.H. and S.C.C.; Methodology: T.H.; Software: S.C.C.; Validation: T.H.; Formal Analysis: T.H.; Investigation: T.H.; Resources: S.C.C.; Data Curation: S.C.C.; Writing Original Draft Preparation: T.H.; Writing Review and Editing: T.H.; Visualization: S.C.C.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. M. John, “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market,” Analytics, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[2] M. Ray and B. K. Mangaraj, “AHP Based Data Mining for Customer Segmentation Based on Customer Lifetime Value,” Int. J. Data Min. Tech. Appl., vol. 5, no. 1, pp.

28–34, 2016, doi: 10.20894/ijdmta.102.005.001.007.

[3] E. Bilgiç, Ö. K. Çakir, M. Kantardzic, and Y. Duan, “Retail Analytics: Store Segmentation Using Rule-Based Purchasing Behavior Analysis,” Int. Rev. Retail Distrib. Consum. Res., vol. 31, no. 4, pp. 457–480, 2021, doi: 10.1080/09593969.2021.1915847.

[4] C. Menidjel and A. Bilgihan, “The Determinants of Retail Customers’ Purchase Intent,” Int. J. Consum. Stud., vol. 46, no. 6, pp. 2503–2520, 2022, doi: 10.1111/ijcs.12802.

[5] M. A. Varma, “Use of Big Data in the Process of Customer Segmentation in the Retail Sector,” Technoarete Trans. Adv. Data Sci. Anal., vol. 1, no. 2, 2022, doi: 10.36647/ttadsa/01.02.a002.

[6] C. Menidjel, A. Benhabib, A. Bilgihan, and M. Madanoglu, “Assessing the Role of Product Category Involvement and Relationship Proneness in the Satisfaction–loyalty Link in Retailing,” Int. J. Retail Distrib. Manag., vol. 48, no. 2, pp. 207–226, 2019, doi: 10.1108/ijrdm-01-2019-0020.

[7] G. Thangarasu and K. R. V. Subramanian, “Developing a Forecasting Model for Retailers Based on Customer Segmentation Using Data Mining Techniques,” Int. J. Trend Sci. Res. Dev., vol. Special Issue, no. Special Issue-ICAEIT2017, pp. 151–155, 2018, doi: 10.31142/ijtsrd19127.

[8] P. Heriyati, “Channel Strategy to Customer Satisfaction: Case of Traditional Retail Channel in Jakarta,” Int. J. Eng. Adv. Technol., vol. 8, no. 6s3, pp. 277–283, 2019, doi: 10.35940/ijeat.f1044.0986s319.

[9] “Hybrid Segmentation of Omnichannel Grocery Customers in Cross-Channel Behaviour Context,” Hong Kong J. Soc. Sci., vol. 60, no. No. 60 Autumn/Winter 2022, 2023, doi: 10.55463/hkjss.issn.1021-3619.60.30.

[10] B. Turkmen, “Customer Segmentation With Machine Learning for Online Retail Industry,” Eur. J. Soc. Behav. Sci., vol. 31, no. 2, pp. 111–136, 2022, doi: 10.15405/ejsbs.316.

[11] U. G. Çiçekli and İ. Kabasakal, “Market Basket Analysis of Basket Data With Demographics: A Case Study in E-Retailing,” Alphanumeric J., vol. 9, no. 1, pp. 1–12, 2021, doi: 10.17093/alphanumeric.752505.

[12] B. Ghojogh, A. Ghojogh, M. Crowley, and F. Karray, “Fitting a mixture distribution to data: tutorial,” 2019, doi: 10.48550/arxiv.1901.06708.

[13] F. Laghari, F. Ahmed, and María de las Nieves López García, “Cash Flow Management and Its Effect on Firm Performance: Empirical Evidence on Non-Financial Firms of China,” Plos One, vol. 18, no. 6, p. e0287135, 2023, doi: 10.1371/journal.pone.0287135.

[14] Z. Ntsalaze, G. Boako, and P. Alagidede, “The Impact of Sovereign Credit Ratings on Corporate Credit Ratings in South Africa,” Afr. J. Econ. Manag. Stud., vol. 8, no. 2, pp. 126–146, 2017, doi: 10.1108/ajems-07-2016-0100.

[15] M. Nadeem, C. Gan, and C. Nguyen, “Does Intellectual Capital Efficiency Improve Firm Performance in BRICS Economies? A Dynamic Panel Estimation,” Meas. Bus. Excell., vol. 21, no. 1, pp. 65–85, 2017, doi: 10.1108/mbe-12-2015-0055.