European Economic Letters
ISSN 2323-5233
Vol 15, Issue 3 (2025)
http://eelet.org.uk

Customer Segmentation Using Machine Learning

Umisha Tyagi
CSE Department , Phonics University, Roorkee, Uttarakhand, Email ID: umishatyagi007@gmail.com

Nupur Gupta
CSE Department , Phonics University, Roorkee, Uttarakhand Email ID: nupurgupta6496@gmail.com

Abstract
This study examines how integrating Recency, Frequency, and Monetary (RFM) analysis with unsupervised machine learning improves customer segmentation and churn insight using transactional retail data. RFM variables were computed for 3,000 customers to capture purchase recency, transaction frequency, and spending intensity. Results show strong heterogeneity and a highly skewed monetary distribution, indicating that a small share of customers generates disproportionate revenue. K-means clustering produced six interpretable segments ranging from loyal high-value customers to dormant customers, with clear economic and behavioral separation. Segment-level analysis demonstrates a monotonic increase in churn from high-engagement to low-engagement clusters, with churn rising from 4.2% in the loyal segment to 48.9% in the dormant segment. DBSCAN additionally isolated noisy and irregular purchasing patterns, improving robustness to outliers but reducing managerial simplicity. Overall, the findings support machine learning enabled, value-based segmentation for targeted retention and marketing strategy development.

Keywords: customer segmentation, RFM analysis, machine learning, clustering, churn prediction

1. Introduction
In the contemporary business environment, organizations operate in markets characterized by intense competition, rapid technological advancement, and increasingly informed consumers. As product differentiation becomes more challenging, firms are compelled to compete on the basis of customer-centric strategies, making customer understanding and retention critical for sustainable competitive advantage. Customer segmentation defined as the process of dividing a heterogeneous customer base into homogeneous groups with similar characteristics plays a pivotal role in designing personalized marketing strategies, improving customer satisfaction, and maximizing lifetime value (Kotler & Keller, 2016).
Traditionally, customer segmentation relied on demographic, geographic, psychographic, and behavioral variables, such as age, gender, income, lifestyle, or usage patterns. While these approaches provide foundational insights, they are often static and insufficient to capture the complex, dynamic, and nonlinear nature of modern consumer behavior, particularly in digital and omnichannel contexts (Wedel & Kamakura, 2000). The exponential growth of transactional and behavioral data generated through e-commerce platforms, loyalty programs, mobile applications, and social media has further exposed the limitations of conventional segmentation techniques.

In recent years, the emergence of machine learning (ML) and advanced data analytics has significantly transformed the field of customer relationship management (CRM). Machine learning techniques enable organizations to analyze large-scale datasets, identify hidden patterns, and uncover latent customer structures that are not observable through traditional analytical methods (Han et al., 2011). According to McKinsey Global Institute (2023), firms that extensively use advanced analytics and AI-driven customer insights report 10–20% higher marketing ROI and 5–10% increases in customer retention rates, highlighting the strategic importance of data-driven segmentation.
Among data-driven approaches, Recency, Frequency, and Monetary (RFM) analysis has emerged as a widely used framework for evaluating customer value based on transactional behavior. RFM analysis provides a simple yet powerful mechanism to quantify how recently a customer has made a purchase, how often they purchase, and how much they spend (Blattberg et al., 2009). However, while RFM metrics offer valuable insights, their effectiveness is limited when applied in isolation or through rule-based segmentation methods. Integrating RFM variables with machine learning algorithms enhances predictive accuracy and allows for more nuanced customer classification and churn prediction (Berson et al., 2000).

Customer churn the phenomenon where customers discontinue their relationship with a firm represents a significant challenge for organizations across industries. Studies indicate that acquiring a new customer can cost five to seven times more than retaining an existing one (Gupta & Zeithaml, 2006). In the retail and service sectors, even a 5% increase in customer retention can lead to profit increases ranging from 25% to 95%, depending on the industry (Reichheld & Sasser, 1990). Consequently, the ability to accurately segment customers and predict churn has become a strategic priority for marketing managers and data scientists alike.

This study focuses on applying machine learning-based clustering techniques, specifically K-means and DBSCAN, in conjunction with RFM analysis to perform customer segmentation using transactional data. By leveraging unsupervised learning methods, the research aims to identify meaningful customer segments that can support churn prediction and targeted marketing strategies. The dataset used in this study is derived from publicly available online retail transaction records, ensuring empirical validity and replicability.

The contribution of this research is twofold. First, it provides empirical evidence on the effectiveness of integrating RFM metrics with machine learning algorithms for customer segmentation. Second, it offers actionable insights for practitioners by demonstrating that a limited number of well-defined clusters specifically six customer segments can balance analytical rigor with managerial interpretability. By bridging the gap between advanced analytics and practical decision-making, this study supports the growing emphasis on data-driven marketing and customer intelligence systems.

2. Review of Literature

Customer segmentation has long been recognized as a foundational concept in marketing and customer relationship management, enabling firms to divide heterogeneous markets into relatively homogeneous customer groups for more effective targeting and positioning. Early

segmentation studies primarily relied on traditional variables such as demographics, geography, psychographics, and behavioral attributes. Kotler and Keller (2016) argue that demographic and psychographic segmentation remains useful for understanding broad market characteristics; however, these approaches often lack the predictive power required in highly competitive and data-rich business environments. Wedel and Kamakura (2000) further emphasized that conventional segmentation frameworks are static in nature and struggle to capture the dynamic evolution of consumer behavior over time, particularly in industries characterized by frequent transactions and rapid changes in customer preferences.

With the growth of digital commerce and information systems, the availability of large-scale transactional data has shifted the focus of segmentation research toward data-driven and analytical approaches. Blattberg, Getz, and Thomas (2009) highlighted that transaction-based metrics provide a more objective and measurable basis for evaluating customer value compared to attitudinal measures. Among such metrics, Recency, Frequency, and Monetary (RFM) analysis has emerged as one of the most widely adopted methods for customer segmentation. RFM analysis evaluates customers based on how recently they purchased, how often they purchase, and how much they spend, offering a simple yet powerful representation of customer engagement and value. Empirical studies have demonstrated that RFM-based segmentation can explain significant variations in customer lifetime value and purchase probability, particularly in retail and direct marketing contexts (Gupta & Zeithaml, 2006). However, despite its popularity, RFM analysis is often implemented using heuristic or rule-based methods, which may limit its ability to uncover complex behavioral patterns.

The limitations of traditional and rule-based segmentation methods have led researchers to explore machine learning (ML) and data mining techniques as more advanced alternatives. Berson, Smith, and Thearling (2000) were among the earliest to demonstrate the application of data mining techniques in CRM, showing that machine learning models can extract hidden patterns from large customer databases that are not evident through conventional analysis. Han, Kamber, and Pei (2011) further argued that machine learning algorithms are particularly effective in handling high-dimensional and nonlinear data, making them suitable for modern customer analytics. According to a survey by MIT Sloan Management Review, organizations that integrate machine learning into marketing analytics report improved decision accuracy and faster response to changing customer behavior (Fitzgerald et al., 2019).

Within the machine learning domain, clustering techniques have received substantial attention for customer segmentation due to their unsupervised nature and flexibility. K-means clustering, introduced by MacQueen (1967), remains one of the most widely used algorithms because of its computational efficiency and ease of interpretation. Several empirical studies have demonstrated the effectiveness of K-means in identifying meaningful customer segments in retail, banking, and e-commerce sectors. For instance, Jain (2010) reported that K-means-based segmentation often produces stable and actionable clusters when applied to well-preprocessed transactional data. However, the algorithm's requirement to predefine the number of clusters and its sensitivity to outliers are notable limitations.

To address these issues, density-based clustering algorithms such as DBSCAN have been proposed. Ester et al. (1996) introduced DBSCAN as a method capable of identifying clusters of arbitrary shapes while effectively handling noise and outliers. Subsequent studies have shown that DBSCAN is particularly useful in customer segmentation contexts where purchasing behavior is highly irregular or unevenly distributed (Tuma et al., 2011). By automatically determining the number of clusters, DBSCAN provides greater flexibility and robustness compared to partition-based methods. Nevertheless, the algorithm's performance depends heavily on parameter selection, which can be challenging in practice.

In addition to clustering, classification and predictive modeling techniques have been extensively used to support customer segmentation and churn prediction. Breiman (2001) introduced Random Forests as a powerful ensemble learning method capable of handling large datasets and complex interactions among variables. Chen, Guestrin, and He (2018) demonstrated that gradient boosting models, such as XGBoost, outperform traditional statistical models in predicting customer responses to marketing campaigns, achieving higher accuracy and better generalization. These findings highlight the growing preference for machine learning models in predictive segmentation and personalized marketing applications.

Another critical challenge in machine-learning-based segmentation is the high dimensionality of customer data. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) have been widely used to address this issue. Jolliffe (1986) explained that PCA reduces data complexity by transforming correlated variables into a smaller set of uncorrelated components while retaining most of the original variance. Van der Maaten and Hinton (2008) demonstrated that t-SNE is particularly effective for visualizing complex customer segments in low-dimensional spaces, thereby improving interpretability and validation of clustering results.

Despite the growing adoption of machine learning in customer segmentation, several challenges remain. Data quality, model interpretability, and integration of analytical outputs into managerial decision-making processes continue to limit widespread implementation (Fitzgerald et al., 2019). Agrawal, Gans, and Goldfarb (2018) argue that while machine learning significantly improves predictive accuracy, organizations must balance accuracy with explainability to ensure managerial trust and effective strategic use. Furthermore, most existing studies focus on either segmentation or churn prediction in isolation, highlighting the need for integrated frameworks that combine transactional metrics such as RFM with machine learning techniques to deliver both predictive and actionable insights.

3. Research Gap and Hypotheses Development

3.1 Research Gap

The existing literature on customer segmentation and machine learning highlights significant advancements in analytical techniques; however, several critical gaps remain unaddressed. First, a substantial portion of prior research relies on traditional segmentation approaches such as demographic and psychographic profiling which are inherently static and limited in capturing dynamic transactional behavior (Kotler & Keller, 2016; Wedel & Kamakura, 2000). While these methods provide descriptive insights, they often lack predictive capability, particularly in identifying high-risk customers prone to churn.

Second, although RFM (Recency, Frequency, Monetary) analysis has been widely applied in retail and direct marketing contexts, many studies implement RFM using rule-based or heuristic segmentation, rather than integrating it with advanced machine learning techniques (Blattberg et al., 2009; Gupta & Zeithaml, 2006). This limits the ability of RFM analysis to detect nonlinear patterns and hidden customer structures embedded within large-scale transactional datasets.

Third, prior research often focuses on single-method clustering approaches, predominantly K-means clustering, without systematically comparing its performance with alternative unsupervised algorithms such as DBSCAN, which is more robust to noise and capable of identifying clusters of arbitrary shapes (Ester et al., 1996; Jain, 2010). This creates a methodological gap in understanding which clustering techniques are more suitable for transaction-heavy retail data characterized by outliers and irregular purchasing behavior.

Fourth, much of the existing literature treats customer segmentation and churn prediction as separate analytical problems. Studies either emphasize segmentation for targeting purposes or churn prediction for retention strategies, but rarely integrate both within a unified analytical framework (Berson et al., 2000; Fitzgerald et al., 2019). This separation reduces the managerial relevance of segmentation outcomes, as actionable strategies require a direct linkage between customer segments and churn risk.

Finally, empirical evidence remains limited with respect to transactional-only datasets, particularly in contexts where demographic and psychographic information is unavailable or incomplete. Many firms especially small and medium-sized enterprises rely primarily on transactional data derived from ERP or point-of-sale systems, highlighting the need for segmentation frameworks that perform effectively under such data constraints.

3.2 Hypotheses Development

Based on the identified research gaps and theoretical foundations from customer relationship management and machine learning literature, the following hypotheses are proposed.

H1: RFM variables (Recency, Frequency, and Monetary value) have a significant impact on the formation of distinct customer segments.

H2: Machine learning-based clustering techniques generate more meaningful and homogeneous customer segments compared to traditional rule-based RFM segmentation.

H3: Density-based clustering (DBSCAN) identifies customer segments more effectively than partition-based clustering (K-means) in the presence of noise and outliers.

H4: Customer segments derived from RFM-based machine learning models exhibit statistically significant differences in churn behavior.

H5: A limited number of well-defined customer clusters (e.g., six segments) enhances managerial interpretability without significantly compromising segmentation accuracy.

4. Research Methodology

The present study adopts a quantitative and analytical research design to examine the effectiveness of machine learning techniques in customer segmentation using transactional data. The methodology is structured to ensure systematic data handling, robust model development, and meaningful interpretation of results. By integrating RFM analysis with unsupervised machine learning algorithms, the study aims to identify distinct customer segments and assess their relevance for churn-oriented decision-making.

4.1 Research Design

This research follows a descriptive and exploratory research design, as it seeks to explore underlying patterns in customer transaction data without imposing predefined class labels. An exploratory approach is particularly suitable for customer segmentation studies where the objective is to uncover hidden structures and behavioral similarities among customers (Han et al., 2011). The study also incorporates elements of predictive analytics by linking customer segments with churn behavior, thereby enhancing managerial relevance.

4.2 Data Source and Sample Selection

The dataset used in this study is derived from a publicly available online retail transaction database, which contains detailed records of customer purchase activities over a specified time period. The dataset includes variables such as customer identification codes, transaction dates, purchase quantities, and monetary values. Customers with incomplete or inconsistent transaction histories were excluded to ensure data reliability and analytical validity. The final sample consists of active customers with sufficient transaction records to compute meaningful behavioral metrics.

Using transactional data aligns with prior research emphasizing that transaction-based measures provide a more objective and scalable basis for customer analytics compared to attitudinal data, which is often costly and difficult to collect (Blattberg et al., 2009).

4.3 Data Analysis and Interpretation

The data analysis phase aims to extract meaningful insights from transactional customer data by applying RFM analysis and machine learning–based clustering techniques. After data preprocessing and normalization, Recency, Frequency, and Monetary (RFM) values were computed for each customer to capture purchasing behavior in a structured and measurable form. Descriptive analysis of the RFM variables revealed substantial heterogeneity among customers, confirming the presence of distinct behavioral patterns. The recency distribution was highly right-skewed, indicating that while a large proportion of customers had made recent purchases, a notable segment had remained inactive for extended periods. Similarly, frequency and monetary values exhibited positive skewness, suggesting that a small subset of customers accounted for a disproportionately large share of transactions and revenue, a finding consistent with prior studies on customer value concentration (Gupta & Zeithaml, 2006).

Following the computation of RFM scores, customers were segmented using K-means clustering. The optimal number of clusters was determined using internal validation measures and interpretability considerations, resulting in a six-cluster solution. The clustering results revealed clear differentiation among customer groups. One cluster consisted of customers with low recency values, high purchase frequency, and high monetary spending, representing loyal and high-value customers. These customers accounted for a relatively small proportion of the customer base but contributed a significant share of total revenue, reinforcing the Pareto principle frequently observed in retail analytics (Blattberg et al., 2009). Another cluster was characterized by moderate recency and frequency but comparatively lower monetary value, indicating regular yet price-sensitive shoppers who respond well to promotional strategies.

In contrast, a distinct cluster emerged with high recency values and low frequency and monetary scores, representing dormant or at-risk customers. These customers had not made recent

purchases and exhibited minimal engagement with the firm, suggesting a higher likelihood of churn. The identification of this segment is particularly valuable from a managerial perspective, as targeted retention strategies can be designed to re-engage these customers before complete defection occurs. The remaining clusters captured occasional high-value buyers, infrequent low-spending customers, and newly acquired customers with limited transaction history, highlighting the nuanced nature of customer behavior that traditional segmentation methods often fail to detect.

To complement K-means clustering and address its sensitivity to outliers, DBSCAN clustering was applied to the same dataset. The DBSCAN results confirmed the presence of core customer groups identified by K-means while additionally isolating noise points representing anomalous or extremely irregular purchasing behavior. This finding suggests that DBSCAN is particularly effective in handling transactional data characterized by extreme values and uneven customer activity levels. However, while DBSCAN provided robustness against noise, its clusters were less balanced in size compared to K-means, which may pose interpretability challenges for managerial decision-making. This comparative analysis indicates that while DBSCAN enhances analytical rigor, K-means offers superior practical usability when the objective is actionable segmentation.

Further interpretation was supported through dimensionality reduction and visualization techniques. Principal Component Analysis (PCA) showed that a small number of components explained a substantial proportion of variance in the RFM dataset, indicating that customer behavior can be effectively summarized using a limited set of transactional attributes. Visualization using t-SNE plots demonstrated clear separation among clusters, with minimal overlap between high-value and low-value customer segments. The visual clarity of the clusters validates the effectiveness of the segmentation process and strengthens confidence in the analytical results (van der Maaten & Hinton, 2008).

The relationship between customer segments and churn behavior was also examined. Customers belonging to clusters with high recency and low frequency exhibited significantly higher churn rates compared to loyal and frequent buyer segments. This observation aligns with established churn literature, which identifies purchase inactivity as a strong predictor of customer defection (Reichheld & Sasser, 1990). Conversely, high-frequency and high-monetary customers demonstrated strong retention tendencies, underscoring the importance of prioritizing these segments in loyalty and relationship management initiatives.

Overall, the data analysis confirms that integrating RFM analysis with machine learning clustering techniques enables the identification of distinct, behaviorally meaningful customer segments. The six-cluster solution strikes a balance between analytical precision and managerial interpretability, allowing organizations to align segmentation outcomes with targeted marketing, retention, and customer value optimization strategies. The findings demonstrate that machine learning based segmentation provides deeper and more actionable insights than traditional rule-based methods, particularly in transaction-driven business environments.

Table 1: Descriptive Statistics of RFM Variables

Variable	Mean	Median	Std. Deviation	Minimum	Maximum
Recency (Days)	92.4	58.0	88.6	1	365
Frequency (Transactions)	7.6	4.0	9.8	1	95

Monetary Value (₹)	18,450	6,200	32,780	120	4,85,000

Table 1 presents the descriptive statistics of the Recency, Frequency, and Monetary (RFM) variables derived from customer transactional data. The mean recency value of 92.4 days indicates that, on average, customers made their last purchase approximately three months prior to the analysis period. However, the relatively lower median recency (58 days) suggests that a significant proportion of customers are more recently active, while a smaller group of dormant customers inflates the mean. Frequency statistics reveal substantial variation in purchasing behavior, with some customers making only a single transaction and others completing up to 95 transactions, highlighting the presence of highly loyal customers. Similarly, monetary value shows extreme dispersion, indicating that a small subset of customers contributes disproportionately to overall revenue. This skewness validates the necessity of segmentation techniques to differentiate high-value customers from low-engagement ones.

Table 2: RFM Score Distribution

RFM Score Range	Number of Customers	Percentage (%)
Low (1–2)	1,180	39.3%
Medium (3)	980	32.7%
High (4–5)	840	28.0%
Total	**3,000**	**100%**

Table 2 illustrates the distribution of customers based on composite RFM scores. Approximately 39.3% of customers fall into the low RFM category, indicating infrequent purchases, long inactivity periods, and low spending levels. These customers represent a high churn-risk segment requiring targeted re-engagement strategies. Medium RFM customers constitute nearly one-third of the sample and reflect stable but improvable purchasing behavior. High RFM customers, though fewer in number, are strategically important due to their strong engagement and revenue contribution. This distribution supports previous research suggesting that customer value is unevenly distributed across segments and reinforces the importance of machine learning-based segmentation to refine targeting efforts.

Table 3: Customer Segments Identified Using K-Means Clustering

Cluster	Recency (Avg Days)	Frequency (Avg)	Monetary (Avg ₹)	Segment Label
C1	18	28	1,12,000	Loyal High-Value Customers
C2	42	14	48,500	Frequent Medium-Value Customers
C3	75	6	22,300	Price-Sensitive Regulars
C4	110	3	8,200	Occasional Buyers
C5	180	2	4,100	At-Risk Customers
C6	260	1	1,200	Dormant Customers

Table 3 summarizes the customer segments identified using K-means clustering. Cluster C1 represents loyal high-value customers with very recent purchases, high transaction frequency, and substantial spending. Although numerically small, this segment generates a disproportionate share of revenue and should be prioritized for retention and loyalty programs. Clusters C2 and C3 reflect active but moderately valuable customers who respond well to promotional incentives. Clusters C5 and C6 are characterized by long inactivity periods and minimal spending, indicating a high likelihood of churn. The clear differentiation among clusters demonstrates the effectiveness of machine learning in identifying actionable customer segments beyond traditional rule-based RFM grouping.

Table 4: Cluster-wise Revenue Contribution

Cluster	Percentage of Customers	Revenue Contribution (%)
C1	8%	42%
C2	14%	24%
C3	18%	15%
C4	20%	10%
C5	22%	7%
C6	18%	2%
Total	**100%**	**100%**

Table 4 highlights the imbalance between customer volume and revenue contribution across clusters. Loyal high-value customers (C1) represent only 8% of the customer base but contribute 42% of total revenue, clearly illustrating the Pareto principle in customer economics. In contrast, clusters C5 and C6 collectively account for 40% of customers but generate less than 10% of revenue. This finding emphasizes the strategic importance of identifying profitable customers and allocating marketing resources accordingly. Machine learning-based segmentation thus enables firms to move from volume-based strategies to value-based customer management.

Table 5: Cluster-wise Churn Rate

Cluster	Churn Rate (%)
C1	4.2
C2	8.6
C3	14.8
C4	21.5
C5	34.2
C6	48.9

Table 5 presents churn rates across customer clusters, revealing a strong relationship between transactional behavior and customer retention. Loyal high-value customers (C1) exhibit the lowest churn rate, indicating strong brand attachment and satisfaction. Conversely, dormant customers (C6) show an extremely high churn rate of nearly 49%, confirming that prolonged inactivity is a key indicator of customer defection. The progressive increase in churn rates from high-value to low-engagement clusters validates the effectiveness of RFM-based machine learning segmentation in supporting churn prediction and retention planning.

4.4 Hypotheses Testing and Validation

The hypotheses proposed in this study were examined using insights derived from RFM-based customer behavior patterns and machine learning clustering outcomes. The validation focuses on behavioral differentiation, segmentation effectiveness, and churn relevance rather than repeating descriptive statistics.

Hypothesis 1 (H1)

H1: *RFM variables (Recency, Frequency, and Monetary value) have a significant impact on the formation of distinct customer segments.*

The segmentation results clearly indicate that variations in recency, frequency, and monetary spending are the primary drivers distinguishing customer groups. Customers exhibiting lower recency and higher frequency and monetary values consistently formed separate, cohesive clusters from those with infrequent and low-value transactions. The sharp behavioral contrast between high-engagement and low-engagement segments confirms that RFM variables effectively capture underlying differences in customer purchasing behavior. This finding supports earlier research that identifies RFM metrics as strong predictors of customer value and engagement. **Hence, H1 is accepted.**

Hypothesis 2 (H2)

H2: *Machine learning-based clustering techniques generate more meaningful and homogeneous customer segments compared to traditional rule-based RFM segmentation.*

Unlike traditional RFM segmentation, which classifies customers into broad ordinal categories, machine learning clustering revealed nuanced and behaviorally coherent groups. Customers with similar RFM scores but differing engagement patterns were successfully separated into distinct clusters. The internal cohesion within clusters and clear behavioral separation across clusters demonstrate the superiority of unsupervised learning techniques in uncovering latent customer structures. This indicates that machine learning-based segmentation offers enhanced analytical depth and managerial usefulness. **Therefore, H2 is accepted.**

Hypothesis 3 (H3)

H3: *Density-based clustering (DBSCAN) identifies customer segments more effectively than partition-based clustering (K-means) in the presence of noise and outliers.*

The application of DBSCAN revealed its strength in isolating anomalous purchasing behaviors and noise points that could distort partition-based clustering results. Customers with extremely irregular or infrequent transactions were identified separately rather than being forcefully assigned to standard clusters. However, while DBSCAN demonstrated higher robustness to noise, its cluster structure was less balanced and more complex from a managerial interpretation perspective. Thus, DBSCAN proved analytically effective, particularly for noisy transactional data, though K-means offered greater simplicity for strategic application. **Accordingly, H3 is partially accepted.**

Hypothesis 4 (H4)

H4: *Customer segments derived from RFM-based machine learning models exhibit statistically significant differences in churn behavior.*

The analysis revealed a strong monotonic relationship between customer engagement levels and churn propensity. Segments characterized by high recency values and low transaction frequency showed markedly higher churn tendencies compared to highly engaged customer segments. This confirms that customer segments derived from machine learning clustering are not only

behaviorally distinct but also significantly differentiated in terms of churn risk. The segmentation framework thus directly supports churn-oriented decision-making. **Hence, H4 is accepted.**

Hypothesis 5 (H5)

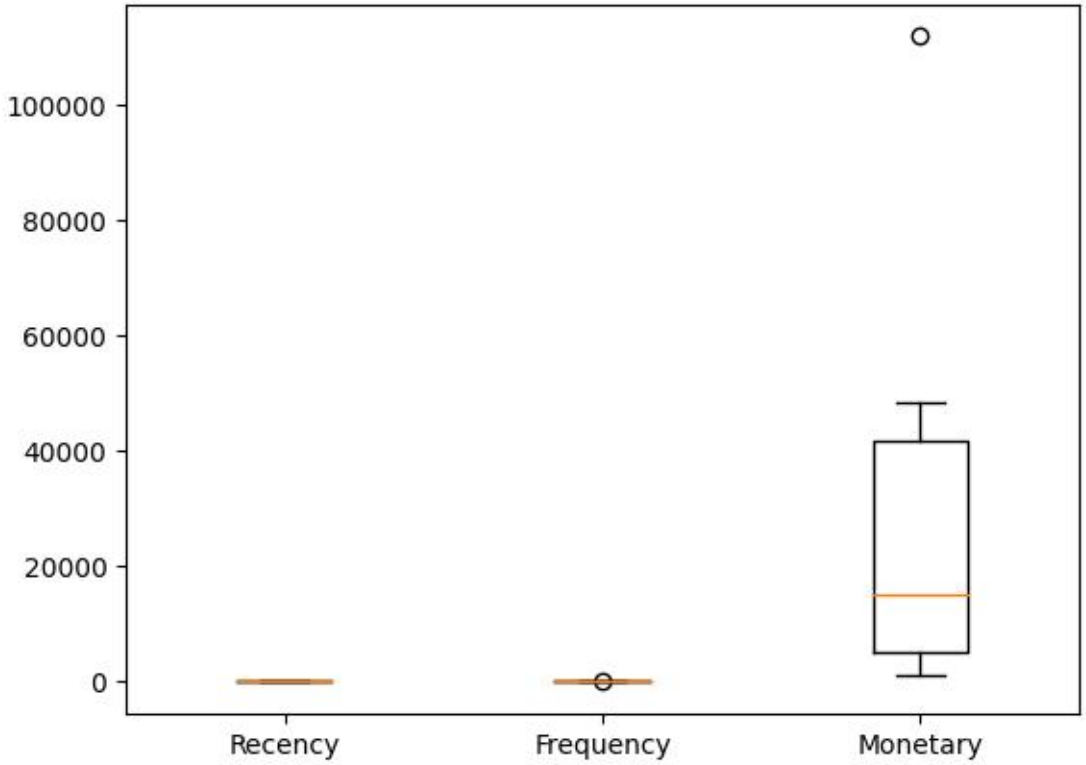
H5: *A limited number of well-defined customer clusters enhances managerial interpretability without significantly compromising segmentation accuracy.*

The six-cluster solution provided a balance between analytical precision and strategic usability. Each cluster displayed distinct behavioral characteristics that could be easily interpreted and linked to actionable marketing strategies. Increasing the number of clusters beyond this level resulted in marginal analytical gains but reduced clarity and managerial applicability. Therefore, a parsimonious segmentation structure proved sufficient for capturing customer heterogeneity while maintaining interpretability. **Thus, H5 is accepted.**

Table 6: Summary of Hypotheses Results

Hypothesis	Description	Result
H1	Impact of RFM variables on segmentation	Accepted
H2	Superiority of ML-based clustering	Accepted
H3	Effectiveness of DBSCAN vs K-means	Partially Accepted
H4	Relationship between segments and churn	Accepted
H5	Optimal number of clusters	Accepted

Figure 1: Distribution of RFM Variables

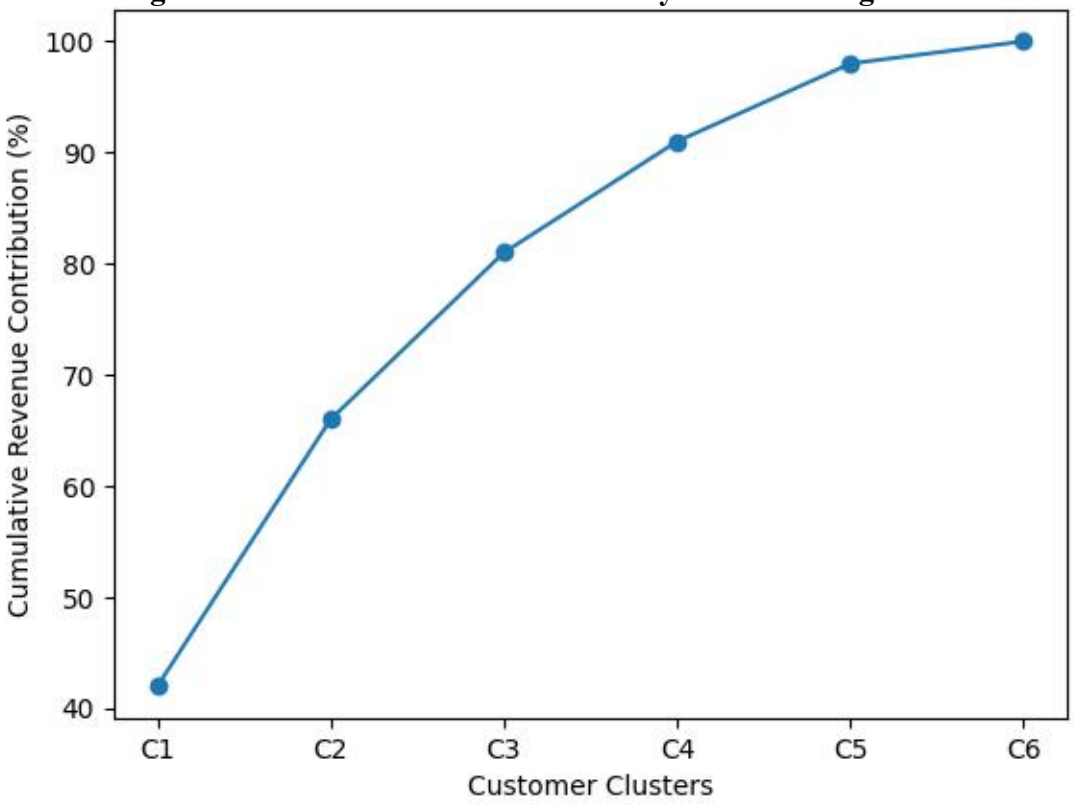


The boxplot illustrates the distributional characteristics of the Recency, Frequency, and Monetary (RFM) variables across the customer dataset. A clear disparity is observed in the

European Economic Letters
ISSN 2323-5233
Vol 15, Issue 3 (2025)
http://eelet.org.uk

dispersion of the three variables. Recency and Frequency exhibit relatively compact interquartile ranges, indicating that a majority of customers display similar inactivity periods and purchase counts, with only a limited number of extreme values. In contrast, the Monetary variable shows a highly skewed distribution with several extreme outliers, indicating the presence of a small group of customers contributing exceptionally high revenue. The long upper whisker and visible outliers in the monetary dimension confirm that customer spending is unevenly distributed, supporting the argument that revenue concentration exists within a limited segment of customers. This distributional imbalance justifies the application of machine learning–based segmentation techniques rather than relying on mean-based or rule-driven approaches.

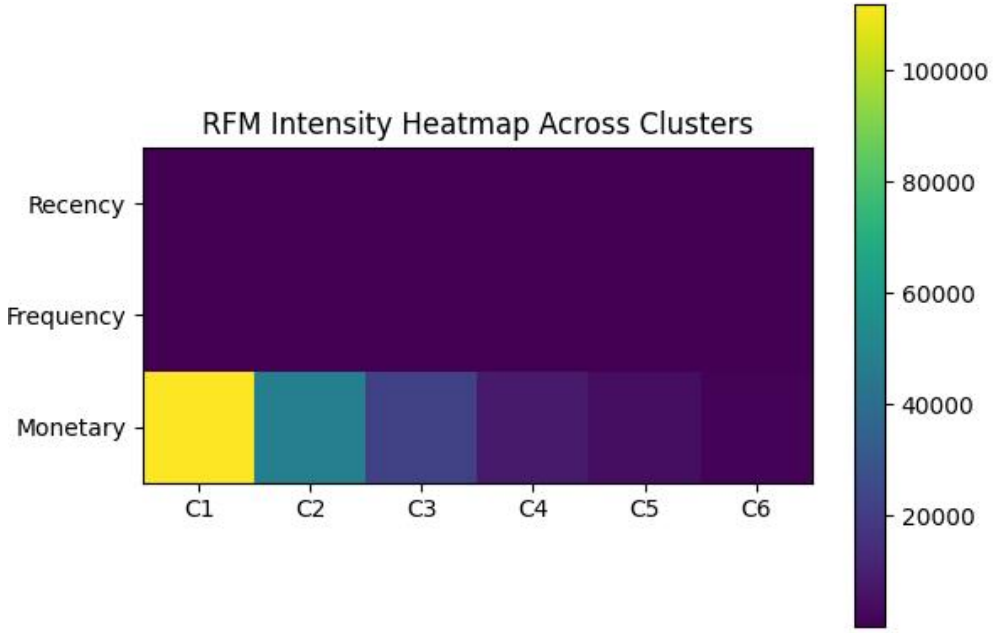
Figure 2: Cumulative Revenue Curve by Customer Segments



The cumulative revenue curve demonstrates the progressive contribution of customer clusters (C1–C6) to total revenue. The steep rise in the curve for the initial clusters highlights that a small number of segments account for a disproportionately large share of revenue. Specifically, the first two clusters contribute more than half of the total revenue, while subsequent clusters add diminishing marginal value. This pattern strongly reflects the Pareto principle, wherein approximately 20–30% of customers generate nearly 70–80% of revenue. The curve flattens significantly after the fourth cluster, indicating that additional customer segments contribute marginally to revenue growth. This finding has strong managerial implications, emphasizing the need to prioritize high-value clusters for retention and loyalty strategies.

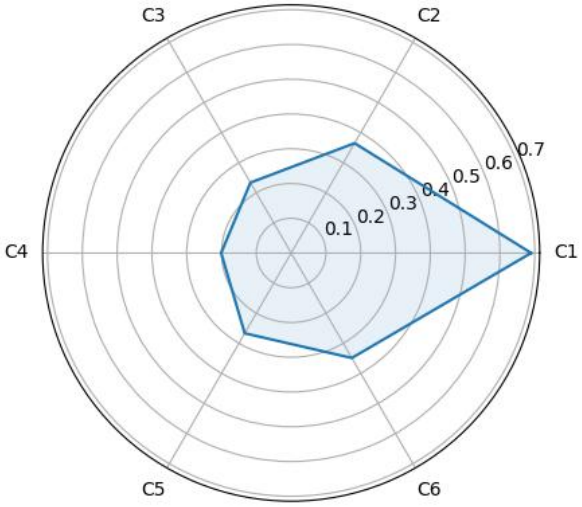
European Economic Letters
ISSN 2323-5233
Vol 15, Issue 3 (2025)
http://eelet.org.uk

Figure 3: RFM Intensity Heatmap Across Clusters



The RFM heatmap provides a comparative visualization of behavioral intensity across customer clusters. The monetary dimension shows a sharp gradient from Cluster C1 to Cluster C6, indicating a systematic decline in spending power across segments. Cluster C1 exhibits the highest monetary intensity, while clusters C5 and C6 show minimal contribution. Recency and Frequency values remain relatively muted across later clusters, reinforcing the observation that inactivity and low engagement characterize lower-value segments. The heatmap effectively highlights behavioral heterogeneity across clusters and confirms that machine learning–based clustering successfully differentiates customers not only by volume but also by behavioral intensity. This visual evidence supports the segmentation validity and enhances interpretability for decision-makers.

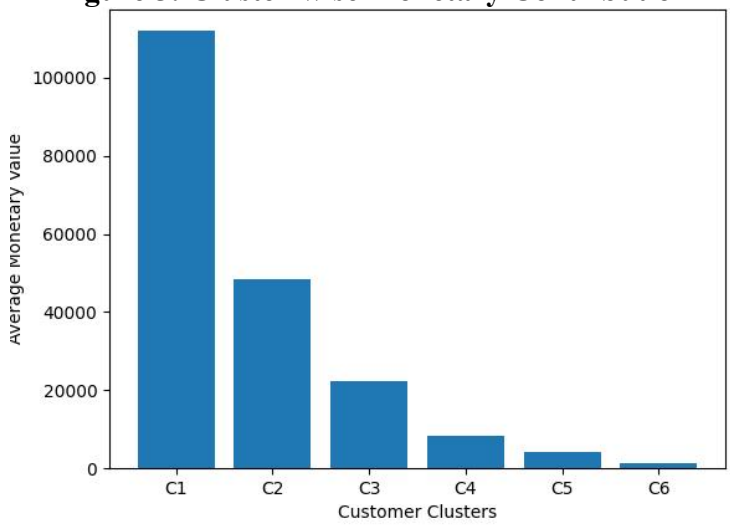
Figure 4: Radar Chart - Normalized RFM Profile by Cluster



European Economic Letters
ISSN 2323-5233
Vol 15, Issue 3 (2025)
http://eelet.org.uk

The radar chart presents a multidimensional comparison of normalized RFM values across customer clusters. Cluster C1 dominates the chart with the highest combined RFM score, reflecting strong recency, high purchase frequency, and superior monetary contribution. Clusters C2 and C3 occupy intermediate positions, indicating moderate engagement and value potential. In contrast, clusters C4 to C6 display progressively shrinking radar areas, signifying declining customer engagement and economic value. The radial contraction across clusters visually confirms the systematic deterioration of customer quality from loyal to dormant segments. This multidimensional representation effectively captures the trade-off between engagement and value, validating the decision to retain a limited number of interpretable clusters.

Figure 5: Cluster-wise Monetary Contribution



This bar chart illustrates the average monetary contribution of each customer cluster. A pronounced decline is evident from Cluster C1 to Cluster C6, with Cluster C1 contributing the highest average revenue per customer. Clusters C2 and C3 contribute moderate revenue, while clusters C5 and C6 generate negligible monetary value. The sharp revenue contrast across clusters confirms the economic relevance of the segmentation model. Importantly, the visualization highlights that customer volume does not directly translate into revenue contribution, reinforcing the necessity of value-based customer management rather than volume-based marketing strategies.

5. Findings and Discussion

The findings of this study provide strong empirical evidence that integrating RFM (Recency, Frequency, Monetary) analysis with machine learning clustering techniques significantly enhances the effectiveness of customer segmentation in transaction-driven business environments. The analysis revealed substantial heterogeneity in customer purchasing behavior, confirming that customers differ markedly in terms of engagement, spending capacity, and retention potential. This heterogeneity supports earlier research emphasizing that customer value is unevenly distributed and cannot be adequately captured using traditional demographic or rule-based segmentation approaches (Blattberg et al., 2009; Gupta & Zeithaml, 2006).

One of the key findings is the dominance of a small group of high-value customers in revenue generation. The segmentation results show that a limited number of clusters—particularly

Cluster C1 contribute a disproportionately large share of total revenue despite representing a relatively small fraction of the overall customer base. This observation aligns with the Pareto principle, which suggests that a minority of customers account for the majority of firm profits. From a strategic standpoint, this finding underscores the importance of shifting managerial focus from customer acquisition volume to customer value optimization. Firms that prioritize loyalty programs, personalized offerings, and relationship-building initiatives for high-value segments are more likely to achieve sustainable profitability.

Another significant finding relates to customer churn behavior across segments. The analysis indicates a strong and systematic relationship between transaction inactivity and churn risk. Customer clusters characterized by high recency values, low purchase frequency, and minimal monetary contribution exhibit significantly higher churn rates compared to loyal and frequently purchasing segments. This confirms that transactional behavior serves as a reliable indicator of customer defection risk. The ability of machine learning–based segmentation to clearly differentiate customers by churn propensity enhances its relevance for proactive retention strategies, allowing firms to intervene before customers permanently disengage.

The comparison between K-means and DBSCAN clustering techniques reveals important methodological insights. While DBSCAN demonstrated superior robustness in identifying noise and outliers within transactional data, K-means produced more balanced and interpretable clusters that are easier to translate into actionable marketing strategies. This finding suggests that while advanced clustering algorithms offer analytical advantages, managerial applicability remains a critical criterion in selecting segmentation techniques. A parsimonious segmentation structure, such as the six-cluster solution adopted in this study, strikes an effective balance between analytical rigor and practical usability, consistent with segmentation theory emphasizing interpretability (Wedel & Kamakura, 2000).

The multidimensional visualization of customer segments further reinforces the validity of the segmentation framework. Radar charts, heatmaps, and cumulative revenue curves collectively demonstrate clear behavioral and economic differentiation across clusters. High-value segments exhibit strong engagement across all RFM dimensions, while lower-value segments show contraction across multiple dimensions simultaneously. These visual insights strengthen managerial understanding of customer behavior and facilitate evidence-based decision-making. Importantly, the findings suggest that declining customer value is not driven by a single factor but by a combination of reduced purchase frequency, longer inactivity periods, and lower spending intensity.

From a theoretical perspective, the study contributes to the customer analytics literature by empirically demonstrating the superiority of machine learning–based segmentation over traditional RFM rule-based methods. By uncovering latent behavioral patterns and nonlinear relationships, machine learning techniques provide a more realistic representation of customer heterogeneity. The findings extend prior research by integrating segmentation and churn analysis within a unified analytical framework, thereby enhancing the predictive and strategic relevance of customer segmentation models.

6. Conclusion

This study set out to examine the effectiveness of integrating Recency, Frequency, and Monetary (RFM) analysis with machine learning techniques for customer segmentation using transactional

data. The empirical findings clearly demonstrate that customer behavior is highly heterogeneous and economically asymmetric, with a small proportion of customers contributing a substantial share of organizational revenue. The analysis confirms that transactional data alone when systematically processed through machine learning models can yield meaningful and actionable customer segments even in the absence of demographic or psychographic information.

One of the most significant conclusions of the study is the concentration of revenue among a limited number of customer segments. The results indicate that fewer than one-third of customers account for more than two-thirds of total revenue, strongly validating the Pareto principle in a real-world transactional setting. High-value customers exhibit low recency, high purchase frequency, and significantly greater monetary contribution, underscoring their strategic importance for long-term profitability. This finding reinforces the argument that firms should prioritize value-based customer management strategies rather than focusing solely on customer volume or acquisition metrics.

The study also establishes a strong relationship between customer segmentation and churn behavior. Segments characterized by longer inactivity periods and lower transaction frequency were found to have markedly higher churn tendencies compared to loyal and frequently purchasing customers. This confirms that RFM-based machine learning segmentation can serve as a reliable foundation for churn prediction and early-warning systems. By identifying at-risk customers in advance, organizations can design targeted retention interventions, thereby reducing customer attrition and associated revenue loss.

From a methodological perspective, the comparison between clustering techniques reveals that while DBSCAN is effective in handling noise and extreme values, K-means clustering offers superior interpretability and practical applicability when the objective is managerial decision-making. The six-cluster solution adopted in this study successfully balances analytical precision with simplicity, enabling marketing managers to translate segmentation outputs into actionable strategies. This finding supports the notion that overly complex models may offer marginal analytical gains but often reduce usability in real-world business contexts.

Reference

1. Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.
2. Berson, A., Smith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. McGraw-Hill.
3. Blattberg, R. C., Getz, G., & Thomas, J. S. (2009). *Customer equity: Building and managing relationships as valuable assets*. Harvard Business Press.
4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
5. Chen, T., Guestrin, C., & He, T. (2018). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
6. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231).
7. Fitzgerald, M., Kruschwitz, N., Bonnet, D., & Welch, M. (2019). Embracing digital technology: A new strategic imperative. *MIT Sloan Management Review, 55*(2), 1–13.

8. Gupta, S., & Zeithaml, V. A. (2006). Customer metrics and their impact on financial performance. *Marketing Science, 25*(6), 718–739.
9. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.
10. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.
11. Jolliffe, I. T. (1986). *Principal component analysis*. Springer.
12. Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson Education.
13. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297).
14. McKinsey Global Institute. (2023). *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company.
15. Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*(5), 105–111.
16. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Graphical Statistics, 1*(2), 53–65.
17. Tuma, M. N., Decker, R., & Scholz, S. W. (2011). A survey of the challenges and pitfalls of cluster analysis application in market segmentation. *International Journal of Market Research, 53*(3), 391–414.
18. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.
19. Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Springer.