Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA36338

Customer Segmentation in E-Commerce Based on the Recency, Frequency, and Monetary (RFM) Model and K-Means Clustering

Siyu Huang

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
syhuangsiyu@gmail.com

Abstract. Reliable customer segmentation is difficult when studies report final clusters without showing how data cleaning and cluster-number choices affect validity. This study examines customer segmentation in e-commerce using the Recency, Frequency, and Monetary (RFM) model and K-Means clustering. The Online Retail II dataset contains 1,067,371 invoice-line records from 1 December 2009 to 9 December 2011. After removing missing customer identifiers, cancelled invoices, and non-positive quantities or prices, 805,549 valid purchase rows and 5,878 customers remained. RFM indicators were log-transformed, standardized, and clustered with candidate values from k = 2 to k = 8. Although k = 2 achieved the highest silhouette score (0.440), k = 4 was retained as a more interpretable exploratory solution with a positive silhouette score (0.365). The four segments reveal strong value concentration: high-value loyal customers represent 19.0% of customers but 72.2% of valid spending, while dormant low-value customers represent 35.0% of customers but only 4.0% of spending. The results show how a transparent RFM-based workflow can support retention and reactivation priorities while requiring experimental validation before managerial deployment.

Keywords: Customer Segmentation, e-commerce, RFM model, K-Means clustering, customer value

1. Introduction

E-commerce firms collect detailed transaction records, but these data do not automatically identify which customers are active, valuable, or at risk of becoming inactive. Customer segmentation addresses this problem by converting invoice-level behavior into interpretable customer groups. Prior research shows that RFM-based segmentation can provide a practical basis for this conversion while retaining a clear link to recency, frequency, and monetary value [1].

Traditional statistical applications of RFM emphasize interpretable summaries of customer value. Cheng and Chen combined RFM indicators with rough-set theory to classify customer value, whereas Fader, Hardie, and Lee linked RFM behavior to customer lifetime value through iso-value curves [2, 3]. These approaches are transparent and require relatively few variables, but fixed scores

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA36338

and value summaries may oversimplify heterogeneous purchasing patterns and do not automatically reveal natural group structure.

Recent research has also paired RFM with clustering. A review of 105 e-commerce segmentation studies found that RFM was the most common way to represent customers and K-Means the method used most often [4]. The algorithm still matters. On a UK retail dataset, John, Shobayo, and Ogunleye obtained a higher silhouette score from a Gaussian mixture model than from K-Means [5]. Alamsyah et al. used RFM, the Elbow method, and K-Means to produce interpretable retail clusters [6]. K-Means is efficient, but skewed variables, unequal scales, outliers, and the selected number of clusters can change its results. Before clustering, the RFM variables are log-transformed and standardized as z-scores. Candidate cluster numbers are then compared with the Elbow curve and silhouette score.

This study uses RFM and K-Means to segment customers in Online Retail II. It documents the customer-level steps and compares alternative cluster numbers. The resulting groups are discussed as exploratory patterns, not causal effects.

2. Method

2.1. Data source and analytical sample

Online Retail II from the UCI Machine Learning Repository supplies the transaction records used here [7]. They cover about two years, from 1 December 2009 to 9 December 2011. The dataset predates current platform conditions, but it is public and large enough for this analysis. It also spans two complete years and contains the invoice dates, customer identifiers, quantities, and prices required to reproduce the RFM measures and compare them with earlier segmentation studies.

Customer-level aggregation requires an ID. Records with missing customer IDs were removed, along with cancelled invoices and rows containing non-positive quantities or prices. The remaining sample has 805,549 valid purchase rows for 5,878 customers in 41 countries, representing GBP 17.74 million in valid purchase spending. Country, average order value, transaction-row count, and purchased units are used only to describe the sample.

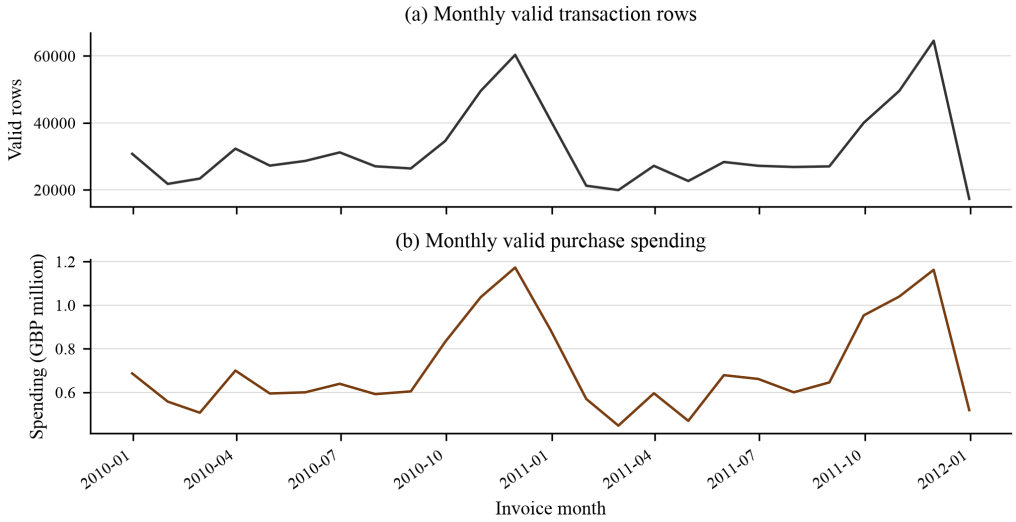


Figure 1. Monthly valid transaction rows and purchase spending after cleaning (picture credit: original)

DOI: 10.54254/2755-2721/2026.BA36338

The exclusions produce a purchase-oriented sample and determine what the clusters represent. Customers without identifiers are absent from the customer-level analysis, and cancelled transactions do not enter an estimate of net value after returns. Monetary measures valid purchase spending, not complete profitability or lifetime value. This boundary is kept explicit when the clusters are interpreted so that their business meaning is not overstated.

Monthly activity is uneven in Figure 1. Valid transaction volume and spending are higher near the end of 2010 and 2011. The observations are treated as dated transactions rather than a static customer list.

2.2. Indicator construction and transformation

Recency, Frequency, and Monetary are the variables used for clustering. For each customer, Recency counts the days from the most recent valid purchase to 10 December 2011. This reference date is the day after the last transaction in the file, giving every customer the same post-observation point without adding later behavioral data. Frequency counts distinct valid invoices. Monetary is total Quantity multiplied by Price across valid invoice lines.

$$R_i = \text{reference date} - \text{last purchase date}; \tag{1}$$

$$F_i = \text{distinct invoices};$$

$$M_i = \sum Quantity \times Price$$

The raw RFM distributions are strongly skewed (Figure 2 and Table 1). The median customer purchased three times and spent GBP 898.92, but the maximum Monetary value exceeded GBP 608,000. Before clustering, each indicator was transformed with ln(1+x) and standardized as a z-score. The logarithm reduces the influence of customers with very high frequency or spending while preserving their rank order. Without this step, a small number of wholesale-like customers could dominate Euclidean distance in K-Means. The +1 term allows zero-valued observations to be transformed. Standardization then puts Recency, Frequency, and Monetary on comparable scales, improving numerical stability and making the segment profiles easier to interpret.

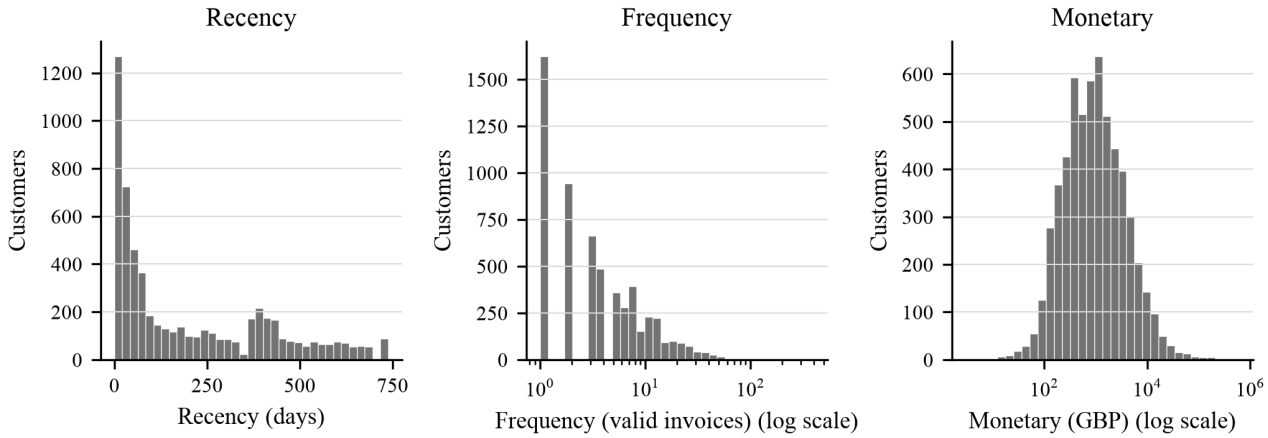


Figure 2. Raw customer-level RFM distributions before transformation (picture credit: original)

DOI: 10.54254/2755-2721/2026.BA36338

Table 1. Customer-level RFM descriptive statistics before transformation

Indicator	Mean	Median	Std. dev.	Min	Max
Recency	200.87	95.00	209.35	0.00	738.00
Frequency	6.29	3.00	13.01	1.00	398.00
Monetary	3,018.62	898.91	14,737.73	2.95	608,821.65

2.3. Clustering procedure

Figure 3 summarizes the workflow. K-Means was applied to standardized log-RFM variables using k-means++ initialization, 50 initialization runs, a maximum of 300 iterations per run, and a fixed random seed of 42. Candidate solutions from k = 2 to k = 8 were evaluated by inertia and the average silhouette score [8]. Segment names were assigned after comparing raw and standardized RFM profiles.

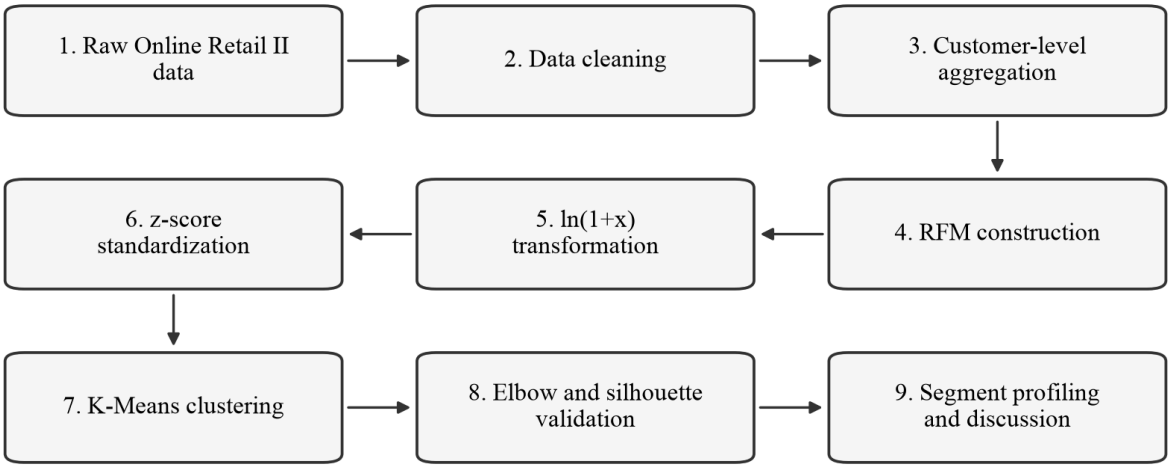


Figure 3. Analytical workflow for RFM-based customer segmentation (picture credit: original)

3. Results and discussion

3.1. Cluster-number validation

Figure 4 and Table 2 report the cluster-number evidence. Figure 4(a) plots inertia, the within-cluster sum of squared distances, against the candidate number of clusters from 2 to 8; the curve declines as k increases because adding clusters reduces compactness error. Figure 4(b) plots the average silhouette score over the same k values; larger positive values indicate more cohesive and better-separated clusters. Inertia decreases visibly up to four clusters and then shows smaller marginal gains. The two-cluster solution has the highest silhouette score (0.440), while k = 4 has a lower but positive score (0.365) and a local improvement over k = 3.

DOI: 10.54254/2755-2721/2026.BA36338

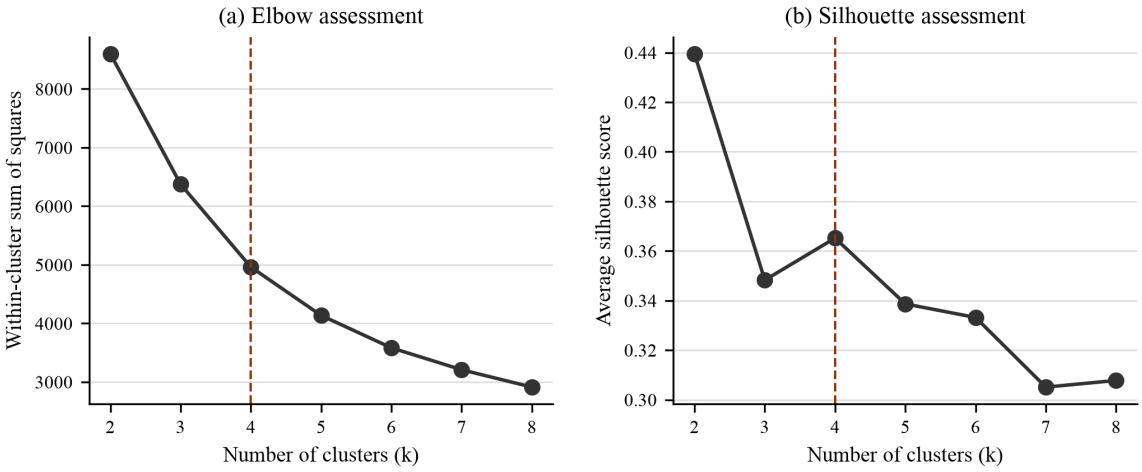


Figure 4. Cluster-number diagnostics: (a) inertia; (b) average silhouette score (picture credit: original)

Table 2. Cluster-number validation results

k	Inertia	Average silhouette score
2	8,600.1	0.440
3	6,373.7	0.348
4	4,960.3	0.365
5	4,136.0	0.339
6	3,582.5	0.333
7	3,208.6	0.305
8	2,912.3	0.308

For this reason, k=4 is retained as an exploratory compromise rather than a unique statistical optimum. The four-cluster solution provides more interpretable behavioral contrasts than k=2, especially between recent regular customers and historically valuable but inactive customers.

This choice follows the purpose of the study. If the aim were only numerical compactness, the two-cluster solution would be attractive because it separates broad active and inactive customers. However, such a solution is too coarse for differentiated marketing discussion. The four-cluster solution keeps statistical separation at an acceptable level while producing customer groups with clearer managerial meanings.

3.2. Segment profiles and managerial interpretation

Table 3 summarizes the reproduced four-cluster profile. High-value loyal customers represent 19.0% of customers but account for 72.2% of valid purchase spending. At-risk valuable customers have high historical spending but much longer recency. Recent regular customers are active but lower in spending, while dormant low-value customers form the largest segment and contribute only 4.0% of spending. Even after logarithmic compression, the high-value segment remains strongly top-heavy: its mean spending of GBP 11,466.01 is more than twice its median of GBP 5,356.03, indicating a Pareto-like concentration in which a small number of super-high-spending customers pull up the segment average.

DOI: 10.54254/2755-2721/2026.BA36338

Table 3. Reproduced four-cluster RFM profile

Segment	Customers (n)	Customer share	Mean recency	Mean frequency	Monetary: mean (median), GBP	Revenue share
High-value loyal	1,117	19.0%	24.8	20.0	11,466.01 (5,356.03)	72.2%
At-risk valuable	1,478	25.1%	210.7	5.4	2,148.18 (1,619.85)	17.9%
Recent regular	1,228	20.9%	27.6	3.0	854.99 (741.73)	5.9%
Dormant low-value	2,055	35.0%	393.0	1.4	345.96 (295.13)	4.0%

Figure 5 clarifies the standardized log-RFM basis for the labels. Negative standardized recency indicates more recent purchasing because a smaller raw recency value is preferable. The profile also shows that monetary value alone would obscure important behavioral differences between at-risk valuable and recent regular customers.

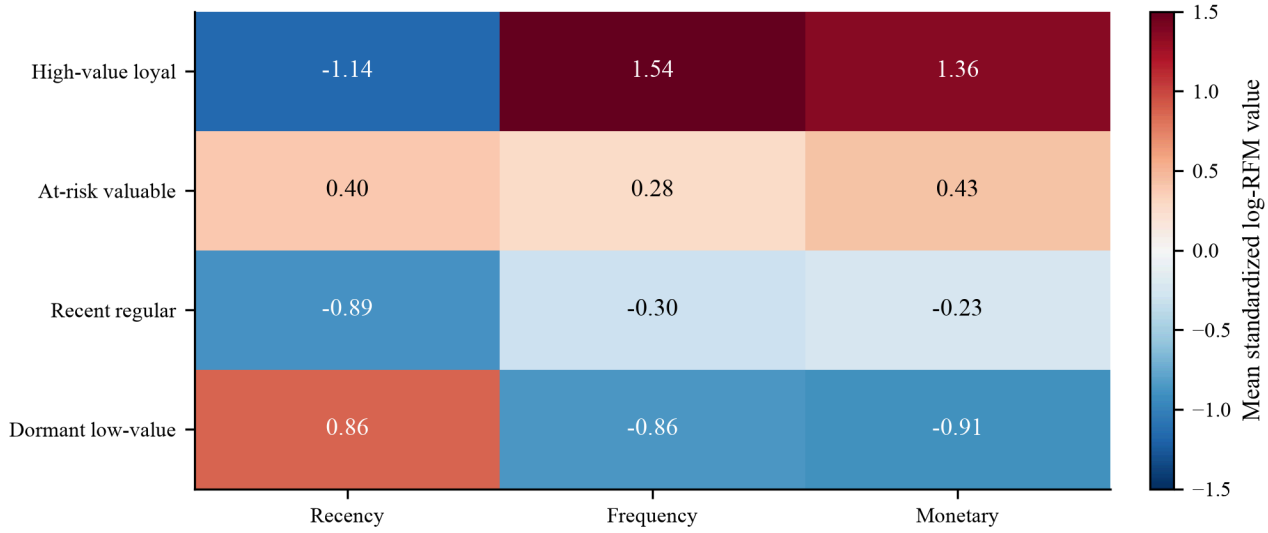


Figure 5. Mean standardized log-RFM values for the four-cluster solution (picture credit: original)

Table 4. Segment interpretation and cautious managerial implications

Segment	Behavioral evidence	Possible strategy	Caution
High-value loyal	Recent purchases; high frequency and monetary value	Loyalty benefits and personalized communication may be considered	Requires monitoring because spending is concentrated
At-risk valuable	High historical value but longer recency	Reactivation tests may be appropriate	Response should be validated experimentally

DOI: 10.54254/2755-2721/2026.BA36338

Table 4. (continued)

Recent regular	Recent activity with moderate frequency and spending	Cross-selling or membership development could be tested	Future value is not guaranteed
Dormant low-value	Long recency, low frequency, and low spending	Low-cost automated communication may be more proportionate	Expensive incentives may be inefficient

As in Table 4, the managerial interpretation should therefore remain cautious. The analysis suggests testable directions for retention, reactivation, cross-selling, and low-cost communication, but it does not prove that any intervention will change customer behavior. In addition, the dataset is dominated by one retailer and does not include marketing exposure, profit margin, customer acquisition cost, or post-campaign response. Recent studies similarly treat RFM-based clustering as context-dependent decision support: Chalechema et al. use RFM-K-Means for segment identification [9]; Antonius and Fitrianah show that adding discount proportion changes customer insight [10]; Wong, Tong, and Haw find that different clustering algorithms can produce different RFM partitions [11]; and Al Hamid and Wiradinata connect RFM-K-Means clusters to targeted promotions and retention programs [12]. Collectively, these studies support a common scientific conclusion: RFM-based clustering can reveal behaviorally distinct groups for decision support, but its usefulness depends on feature design, algorithm choice, and subsequent validation.

The main practical value of the segmentation is prioritization. High-value loyal customers may justify closer monitoring because a small group contributes most purchase spending. At-risk valuable customers may deserve reactivation tests because their historical value is not matched by recent engagement. Recent regular customers could be examined for development opportunities, whereas dormant low-value customers may require lower-cost communication. These suggestions should be treated as hypotheses for subsequent evaluation.

4. Conclusion

This study transformed Online Retail II transaction histories into customer-level RFM segments using K-Means clustering. The analysis reports data scale, cleaning rules, variable construction, log transformation, cluster validation, and segment interpretation in a transparent workflow. After cleaning, 805,549 valid purchase rows from 5,878 customers were retained. The cluster-number evidence did not identify a single automatic answer: k=2 produced the highest silhouette score, but k=4 preserved more useful behavioral distinctions. The retained four-cluster solution shows substantial value concentration. High-value loyal customers account for 19.0% of customers but 72.2% of valid spending, whereas dormant low-value customers account for 35.0% of customers but only 4.0% of spending. These findings support differentiated but cautious marketing hypotheses, such as loyalty monitoring for high-value customers, reactivation tests for at-risk valuable customers, and low-cost communication for dormant customers. The study also has limitations. The dataset represents one UK non-store retailer, cancellations and missing customer identifiers were excluded, and the analysis is descriptive rather than causal. Future work could compare K-Means with hierarchical, Gaussian mixture, and density-based models; include additional behavioral

DOI: 10.54254/2755-2721/2026.BA36338

variables such as returns, discounts, or product categories; and evaluate proposed interventions through controlled experiments. Overall, RFM-based clustering is useful as a prioritization tool, but managerial action should be tested before full deployment.

References

[1] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

[2] Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

[3] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

[4] Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management, 21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

[5] John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics, 2*(4), 809–823. https://doi.org/10.3390/analytics2040042

[6] Alamsyah, Prasetyo, P. E., Sunyoto, Bintari, S. H., Saputro, D. D., Rohman, S., & Pratama, R. N. (2022). Customer segmentation using the integration of the Recency Frequency Monetary model and the K-Means cluster algorithm. *Scientific Journal of Informatics, 9*(2), 189–196. https://doi.org/10.15294/sji.v9i2.39437

[7] Chen, D., & UCI Machine Learning Repository. (2019). Online retail II [Data set]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5CG6D

[8] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[9] Chalechema, S., Saini, M. L., Perla, I., & Shivanand, A. V. (2023). Customer segmentation using K means algorithm and RFM model. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 393–398). https://doi.org/10.1109/ICCCIS60361.2023.10425556

[10] Antonius, V. H., & Fitrianah, D. (2024). Enhancing customer segmentation insights by using RFM + discount proportion model with clustering algorithms. *International Journal of Advanced Computer Science and Applications, 15*(3), 902–911. https://doi.org/10.14569/IJACSA.2024.0150390

[11] Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy, 12*(3), 97–125. https://jtde.telsoc.org/index.php/jtde/article/view/978

[12] Al Hamid, B., & Wiradinata, T. (2026). RFM-based customer segmentation using K-Means clustering for marketing strategy optimization at Queen Audio. *Journal of Applied Informatics and Computing, 10*(2), 1828–1833. https://doi.org/10.30871/jaic.v10i2.12302