Evaluating K-Means and DBSCAN for E-Commerce Customer Segmentation

Lily Puspa Dewi[1], Jonathan Febrian Pratignyo[2], Gregorius Satia Budhi[3]

[1,2,3]Department of Informatics, Petra Christian University, East Java, Indonesia
E-mail: [1*]lily@petra.ac.id, [2]jofebrian01@gmail.com, [3]greg@petra.ac.id

(Received: 17 Aug 2026, revised: 4 Oct 2026, 16 Oct 2026, 12 Nov 2026, accepted: 20 Nov 2026)

Abstract

Customer segmentation based on transaction histories requires consideration of both internal cluster quality and customer coverage. This study evaluated K-Means and DBSCAN using Recency, Frequency, and Monetary (RFM) features from an anonymized Indonesian e-commerce platform. After screening, 3,868 retained orders recorded between June 2021 and June 2023 were aggregated into 2,692 customer profiles. Both algorithms used the same min–max-normalized RFM matrix. Internal validation employed the Silhouette Coefficient, Davies–Bouldin index, and Calinski–Harabasz index. K-Means with K = 3 produced a full-sample Silhouette Coefficient of 0.4169. DBSCAN with ε = 0.13 and min_samples = 5 identified four clusters covering 94.3% of customers and labelled 153 customers as noise. On the matched subset of 2,539 non-noise customers, the Silhouette Coefficients were 0.4350 for K-Means and 0.2017 for DBSCAN. The smallest K-Means cluster contained 15.0% of customers and had the highest mean Frequency and Monetary. Among the DBSCAN clusters, the highest mean Frequency and Monetary occurred in different groups, while the noise group had the highest reported mean Monetary overall. These findings indicate that internal validation should be interpreted alongside customer coverage and the original purchasing profiles of assigned and unassigned customers.

Keywords: Customer Segmentation, K-Means, DBSCAN, RFM Analysis, Silhouette Coefficient

I. INTRODUCTION

Bank Indonesia reported e-commerce transaction values in Indonesia reached IDR 476 trillion in 2022, representing an annual growth of 18.7% [1]. This growth has occurred alongside the increasing adoption of digital technologies, particularly cashless payment systems [2]. Therefore, companies need to adjust their strategies to stay relevant in the competitive e-commerce environment.

Platform switching provides further context for examining customer behavior. A 2022 survey by Meta and Bain & Company reported that approximately 40% of respondents across six Southeast Asian countries had changed their most-used online platform during the three months preceding the survey [3]. Although platform switching does not necessarily indicate customer churn, it highlights the relevance of understanding customer purchasing patterns.

Samosir highlights the role of customer relationship management in maintaining customer relationships and encouraging repeat purchases through personalized services, regular communication, and the use of customer data [4]. One commonly used approach to support customer retention was customer segmentation, which involves dividing buyers into groups based on particular characteristics or behavioral patterns [5]. Once the segments are defined, they can be used to develop more relevant marketing strategies. In an Indonesian e-commerce case study, Sanjaya and Kalifia developed a decision-tree model that used customer preferences to support precision marketing [6].

Although customer segmentation helps companies, many e-commerce firms still find difficulty in applying data-driven segmentation [7]. Conventional segmentation techniques based on demographic factors may not be able to capture the complexity and behavior of online consumers [8]. Chang and Mukherjee reviewed how machine learning can support the analysis of diverse consumer data, including structured records, text, audio, and images [9]. For transaction records, clustering provides a means of identifying groups of customers with similar purchasing patterns without requiring predefined segment labels [10], [11].

In addition, determining the quality of segmentation results is another issue, since inappropriate cluster formation can lead to misleading insights and suboptimal decision-making [12]. There are numerous validity metrics for clustering, but the Silhouette Coefficient is an internal validation metric that evaluates both within-cluster cohesion and between-cluster

DOI: 10.34148/teknika.v15i3.1547

CC BY SA

separation [13]. K-Means and DBSCAN offer different approaches to grouping customer profiles: K-Means partitions observations around centroids, whereas DBSCAN identifies dense regions and can leave observations unassigned to a cluster. Comparing these approaches on the same RFM representation allows their cluster profiles and customer coverage to be examined within a common analytical setting [14], [15].

Previous customer segmentation studies have focused on the representation of customer behavior and the selection of clustering algorithms. By utilizing K-Means, Samidi et al. [16] showed that the representation of transactional behavior influenced the resulting cluster quality. Aslantaş et al. [17] adjusted RFM-related features to connected home-appliance usage and measured K-Means segments using a customer-value method. These studies presented that RFM-based segmentation applications can have divergent operational definitions, feature construction, and business interpretation.

Other studies have investigated alternative clustering approaches. Monalisa et al. [18] examined the clusters of retail customer transactions using DBSCAN, incorporating demographic information into the RFM-based interpretation. Transformed RFM data was used by Wong et al. [19] to compare K-Means and agglomerative hierarchical clustering. Comparisons have also been made directly with K-Means and DBSCAN. John et al. [20] tested both algorithms in combination with other clustering methods on UK retail data. These studies, when taken together, establish a precedent for segmentation and comparative clustering using RFM. Differences in datasets, feature representations and preprocessing procedures, as well as algorithm configurations make it challenging to transfer their findings to another customer population. In Table 1, the selected studies are listed, and their scope is determined in relation to present study.

An internal validation score alone does not indicate how many customers receive cluster assignments or characterize those left unassigned. Reporting these aspects together is therefore necessary to interpret the practical consequences of the selected clustering configurations.

This study examines K-Means and DBSCAN for customer segmentation using the same min–max-normalized RFM representation. The analysis focuses on internal cluster quality, customer coverage, and purchasing profiles under the selected parameter configurations, with explicit consideration of DBSCAN noise observations. The findings provide evidence for the observed dataset rather than establishing the general superiority of either algorithm.

Table 1. Comparison of Previous Studies and Present Study

Study	Dataset	Features	Algorithms	Scope relative to the present study
Samidi et al. [16]	Indonesian stall-based e-commerce; 273,454 transaction rows.	RFM	K-Means	no DBSCAN comparison
Aslantaş et al. [17]	Connected home-appliance usage data.	RFM	K-Means	no DBSCAN comparison
Monalisa et al. [18]	212 Mart; 1,205 customers.	RFM	DBSCAN	no K-means comparison
John et al. [20]	UK Online Retail; 541,909 source records.	RFM	K-Means, DBSCAN, GMM, BIRCH, agglomerative.	Different customer population and preprocessing workflow.
Present study	Indonesian e-commerce; 2,692 customer profiles.	Min–max-normalized RFM.	K-Means, DBSCAN.	Shared RFM representation; full-sample and matched-subset internal validation; explicit customer coverage and noise profiling.

II. METHODS

Figure 1 summarizes the research procedure, comprising transaction screening, customer-level aggregation, RFM feature construction, descriptive analysis, normalization, clustering, internal validation, and customer profiling. Both algorithms used the same normalized RFM feature matrix. Internal validation included the Silhouette Coefficient, Davies–Bouldin index, and Calinski–Harabasz index. Evaluation populations and the treatment of DBSCAN noise observations are specified below. Customer profiles were interpreted using RFM values in their original measurement units.

The source dataset contained 6,512 transaction records from an anonymized Indonesian e-commerce platform, covering June 2021 to June 2023. User_ID, Invoice_ID, Transaction_Date, and Amount were used to construct the RFM features, while Order_Status was used to identify cancelled and returned orders during screening. Before customer-level aggregation, the records were screened for missing identifiers, duplicates, cancelled or returned orders, and zero or negative transaction amounts. Exclusions were applied sequentially, with each excluded record counted only under the first applicable criterion to prevent double counting. Table 2 summarizes the data screening and aggregation process.

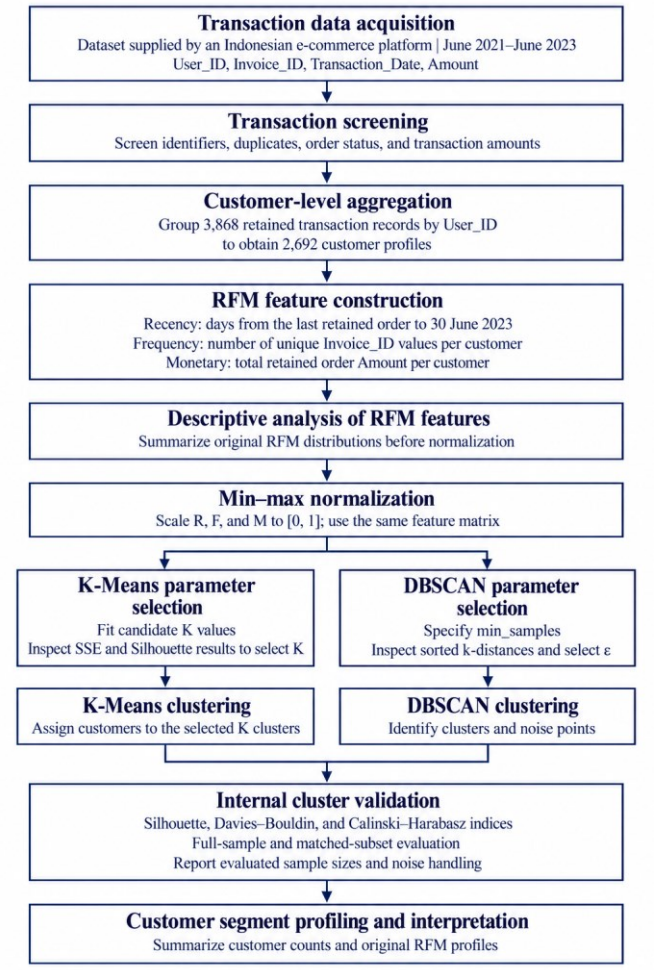


Figure 1. Flow Diagram

Table 2. Transaction Data Screening and Customer Aggregation

Data	Count	Treatment
Initial transaction records	6,512	Source data before cleaning.
Records with missing User_ID	420	Excluded because customer-level aggregation required a valid customer identifier.
Records with missing Invoice_ID	310	Excluded because order identity could not be verified.
Duplicate transaction records	850	Redundant copies excluded; one valid record retained for each duplicated transaction.
Cancelled or returned orders	560	Excluded to restrict the analysis to completed, non-returned purchases. Order_Status= Cancelled or returned.
Records with zero Amount	284	Excluded because they did not represent positive purchase expenditure.
Records with negative Amount	220	Records with negative Amount values were excluded from the retained transaction dataset.
Valid transaction records after cleaning	3,868	Represent distinct orders, with one retained record per Invoice_ID.
Unique customers included in RFM analysis	2,692	Customers obtained by aggregating valid orders by User_ID.

Following screening, 3,868 valid transaction records associated with 2,692 unique customers were retained. These records were aggregated by User_ID to construct customer-level RFM features.

Recency was calculated as the number of calendar days between 30 June 2023 and the date of each customer's most recent retained order using Equation 1.

$$R_i = \text{days}(t_{\text{ref}} - t_{i\,last}) \quad (1)$$

where t_{ref} was June 30, 2023, and $t_{i\,last}$ was the most recent purchase date of customer i. A transaction on the reference date yielded $R_i = 0$; lower values therefore indicated more recent purchasing activity.

Frequency was defined as the number of distinct retained orders per customer during the observation period. Monetary was defined as the total value of those orders. Their formulas is shown in Equation 2:

$$F_i = \mid \mathcal{O}_i \mid$$
$$M_i = \sum_{o \in \mathcal{O}_i} v_o \quad (2)$$

where $\mathcal{O}_i$ denoted the set of distinct orders retained after screening for customer i, and v_o denoted the value of order o in Indonesian rupiah. Each unique order contributed one count to the customer's Frequency, and its transaction amount contributed to the customer's Monetary value.

Before normalization, customer-level RFM values were summarized using the minimum, maximum, mean, median, first and third quartiles, interquartile range, and skewness. Quartiles were calculated using Excel's QUARTILE.INC function, and skewness was calculated using the SKEW function. The interquartile range (IQR) was calculated as the difference between the third and first quartiles. Valid extreme customer profiles were retained, and no customers were excluded solely because of high Frequency or Monetary values.

The RFM features were normalized using min–max scaling before clustering using Equation 3:

$$z_{ij} = \frac{x_{ij} - x_{j\,min}}{x_{j\,max} - x_{j\,min}} \quad (3)$$

DOI: 10.34148/teknika.v15i3.1547
CC BY SA

where x_{ij} represented the value of feature j for customer i, and $x_{j\,min}$ and $x_{j\,max}$ showed minimum and maximum values of that feature across the customer-level dataset. Min–max normalization placed each feature within the range [0, 1], and both algorithms used the same normalized feature matrix. This transformation adjusts numerical ranges but does not remove distributional skewness or eliminate the influence of extreme values. Descriptive statistics and customer profiles were reported in the original measurement units.

K-Means models were fitted to the normalized RFM matrix for K values from 2 to 10 using scikit-learn. The models used k-means++ initialization, n_init = 20, and random_state = 42. Inertia, expressed as the within-cluster sum of squared errors (SSE), and the Silhouette Coefficient were calculated for each candidate solution using all 2,692 customers. The selected K was determined by considering the SSE pattern and Silhouette results.

DBSCAN was applied to the min–max-normalized RFM matrix of 2,692 customers. The min_samples parameter was set to 5, requiring a neighborhood to contain at least five customers, including the customer being evaluated, for that customer to qualify as a core point. The neighborhood radius ε = 0.13 was selected through visual inspection of the knee region in the sorted neighbor-distance curve shown in Figure 2. The selected parameters define the density threshold used to identify customer groups within the shared normalized RFM representation.

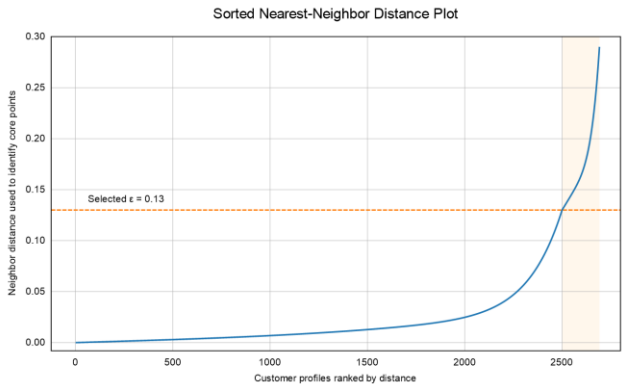


Figure 2. Sorted Nearest-Neighbor Distance Plot

The comparison therefore concerned the selected configurations rather than establishing the best attainable performance of either algorithm.

Internal cluster quality was evaluated using the Silhouette Coefficient, Davies–Bouldin index, and Calinski–Harabasz index. The Silhouette Coefficient was defined as Equation 4.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (4)$$

where $a(i)$was the mean distance from observation $i$to the other observations in its cluster, and $b(i)$was the lowest mean distance from observation $i$to another cluster. The overall coefficient was the mean of the individual coefficients across the evaluated observations. Higher values indicated that observations were, on average, closer to members of their own cluster than to members of the nearest alternative cluster.

Both algorithms were fitted to the full normalized RFM dataset of 2,692 customers. K-Means was evaluated on the full dataset and additionally on the 2,539 customers assigned to non-noise DBSCAN clusters. For this matched-subset evaluation, the original full-data K-Means labels were retained, and all three internal validation indices were recalculated using the same observations for both algorithms. DBSCAN noise observations were excluded from these indices but retained for descriptive profiling. Because the matched subset was defined by DBSCAN assignments, the comparison was conditional on the customers covered by the selected DBSCAN configuration.

To provide a broader assessment of internal cluster quality, the Davies–Bouldin (DB) and Calinski–Harabasz (CH) indices were also calculated. The DB index measures the average similarity between each cluster and its most similar one, where a lower value indicates better cluster separation [21]. The CH index, also known as the variance ratio criterion, evaluates the ratio of between-cluster dispersion to within-cluster dispersion, with higher scores denoting denser and well-separated clusters [22].

Customer groups were characterized using their sizes and the mean, median, and interquartile range of the original RFM values. Lower Recency indicated more recent purchasing, while higher Frequency and Monetary indicated more orders and greater cumulative expenditure. Noise observations were summarized separately and were not assumed to form a homogeneous customer segment. Business implications were treated as hypotheses because campaign response, retention, and profitability were not evaluated.

III. RESULTS

This section presents the customer-level RFM distributions, the selection and results of the clustering configurations, and their internal validation using the Silhouette Coefficient, Davies–Bouldin index, and Calinski–Harabasz index. Customer profiles and DBSCAN coverage are reported alongside the validation results.

Customer-level aggregation yielded 2,692 profiles for analysis. Table 3 summarizes the RFM distributions before normalization using the minimum, maximum, mean, median, first and third quartiles, interquartile range, and skewness. Recency is expressed in days, Frequency in distinct orders, and Monetary in Indonesian rupiah. Lower Recency indicates more recent purchasing activity, whereas higher Frequency and Monetary indicate more orders and greater cumulative expenditure, respectively.

Frequency ranged from one to eight orders, with a median of one and an IQR of one. Its positive skewness indicated a distribution concentrated toward lower order counts with a tail toward higher frequencies. Monetary ranged from IDR 17,000 to IDR 1,462,640. Its mean of IDR 308,899 exceeded the median of IDR 175,056.50, consistent with positive skewness.

These characteristics were considered when interpreting the mean-based cluster profiles. The descriptive statistics do not, by themselves, establish the existence of distinct clusters.

Table 3. Descriptive Statistics of Customer-Level RFM Features

Statistic	Recency	Frequency	Monetary
Minimum	0	1	17,000
Maximum	394	8	1,462,640
Mean	166	1.44	308,899
Median	155	1	175,056.50
Q1	68	1	45,912.00
Q3	255	2	478,113.25
IQR	187	1	432,201.25
Skewness	0.260	1.53	1.282

A. K-Means Clustering

Among the evaluated values of K, K = 3 produced the highest Silhouette Coefficient (0.4169). SSE decreased from 231.75 at K = 2 to 143.84 at K = 3, followed by smaller reductions for larger K values. Together, these results supported selecting the three-cluster configuration. Table 4 reports the candidate solutions, and Figure 3 presents the corresponding elbow plot.

Table 4. SSE for candidate K-Means solutions

K	SSE	Silhouette Coefficient
2	231.75	0.3701
3	143.84	0.4169
4	118.84	0.3578
5	99.71	0.3571
6	84.95	0.3641
7	75.32	0.3752
8	68.13	0.3760
9	63.38	0.3713
10	58.12	0.3623

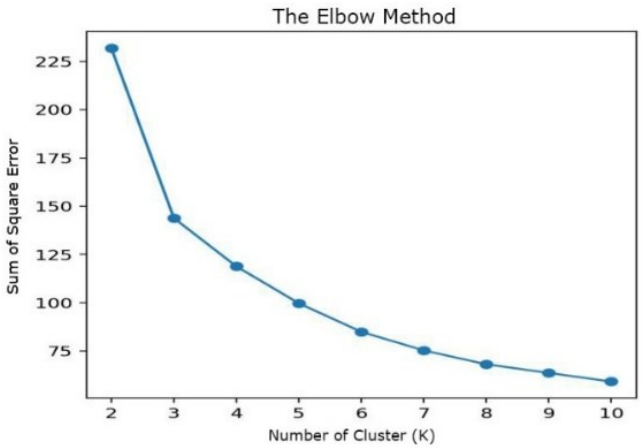


Figure 3. Elbow Plot for K-Means

The final K-Means model was fitted to the normalized RFM variables, producing the three-dimensional cluster distribution shown in Figure 4.

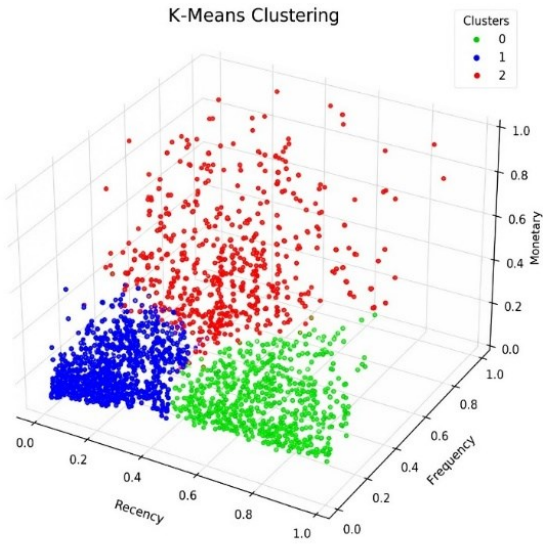


Figure 4. K-Means Clustering of Customer Behavior

The final K-Means model produced three customer segments, as summarized in Table 5.

Table 5. K-Means Customer Segment Profiles

Cluster	Mean Recency	Mean Frequency	Mean Monetary	Customers
0	250	1.18	239,955	1179 (43.8%)
1	82	1.33	218,851	1109 (41.2%)
2	152	2.50	757,288	404 (15%)

Cluster 0 was the largest segment, comprising 1,179 customers (43.8%). It had the highest mean Recency, at 250 days, indicating the longest average interval since the most recent retained order. Its mean Frequency was 1.18 orders per customer, and its mean Monetary was IDR 239,955.

Cluster 1 comprised 1,109 customers (41.2%). It had the lowest mean Recency, at 82 days, indicating the most recent purchasing activity on average. Its mean Frequency was 1.33 orders per customer, slightly higher than that of Cluster 0, while its mean Monetary was the lowest among the three clusters, at IDR 218,851.

Cluster 2 was the smallest segment, comprising 404 customers (15.0%). It had the highest mean Frequency, at 2.50 orders per customer, and the highest mean Monetary, at IDR 757,288. Its mean Recency was 152 days, between those of Clusters 0 and 1. The segment therefore combined relatively frequent purchasing with higher cumulative expenditure during the observation period.

Clusters 0 and 1 accounted for 85.0% of customers and had lower mean Frequency and Monetary than Cluster 2. They differed particularly in Recency: Cluster 1 purchased more recently on average, whereas Cluster 0 had the longest interval since the most recent retained order.

DOI: 10.34148/teknika.v15i3.1547
TEKNIKA, Volume 15(3), November 2026, pp. 609-616
ISSN 2549-8037, EISSN 2549-8045

B. DBSCAN Clustering

Using ε = 0.13 and min_samples = 5, DBSCAN identified four clusters containing 2,539 customers and labelled the remaining 153 customers as noise. Figure 5 shows the distribution of DBSCAN assignments in the three-dimensional RFM space.

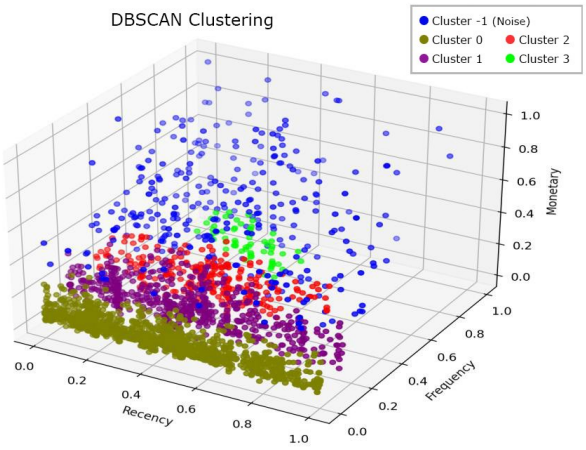


Figure 5. DBSCAN Clustering Results

Table 6 presents the customer counts and RFM profiles of the four DBSCAN clusters. Customers labelled as noise (−1) are summarized separately and are not interpreted as a single homogeneous segment. Noise observations were excluded from internal validation but retained for descriptive profiling. Percentages are based on all 2,692 customers and may not come to 100% because of rounding.

Cluster 0 was the largest DBSCAN cluster, comprising 1,201 customers (44.6%). Its mean Recency was 166.42 days, mean Frequency was 1.00 order per customer, and mean Monetary was IDR 110,000. It had the lowest mean Monetary among the four identified clusters.

Cluster 1 comprised 658 customers (24.4% of the dataset). Its mean Recency was 163.15 days, mean Frequency was 1.00 order per customer, and mean Monetary was IDR 280,450. Although its mean Frequency matched that of Cluster 0, its mean cumulative expenditure was higher.

Cluster 2 comprised 421 customers (15.6%). It had the lowest mean Recency, at 155.08 days, and the highest mean Frequency among the four identified clusters, at 3.00 orders per customer. Its mean Monetary was IDR 450,200.

Cluster 3 was the smallest identified cluster, comprising 259 customers (9.6%). It had the highest mean Monetary among the four clusters, at IDR 620,150, but its mean Frequency of 2.00 orders was lower than that of Cluster 2. Its mean Recency of 190.21 days indicated the longest average interval since the most recent retained order.

DBSCAN labelled 153 customers (5.7%) as noise. Their mean Recency was 157.80 days, mean Frequency was approximately 1.49 orders per customer, and mean Monetary was IDR 1,076,850. The latter exceeded the mean Monetary of every identified cluster. These averages describe the noise observations collectively but do not establish that they form a homogeneous customer segment.

C. Comparative Internal Validation

K-Means obtained a Silhouette Coefficient of 0.4169 on the full sample of 2,692 customers. On the matched subset of 2,539 customers assigned to non-noise DBSCAN clusters, the coefficient was 0.4350 for K-Means and 0.2017 for DBSCAN. The Davies–Bouldin and Calinski–Harabasz indices were evaluated on the same matched subset, as reported in Table 7.

DB denotes the Davies–Bouldin index, and CH denotes the Calinski–Harabasz index. Lower DB and higher Silhouette and CH values indicate better results under their respective criteria. A dash indicates an index not calculated for the reported evaluation population.

Table 6. DBSCAN Customer Segment Profiles

Cluster	Mean Recency	Mean Frequency	Mean Monetary	Customers
0	166.42	1	110,000	1201 (44.6%)
1	163.15	1	280,450	658 (24.4%)
2	155.08	3	450,200	421 (15.6 %)
3	190.21	2	620,150	259 (9.6%)
-1 (noise)	157.80	1.49	1,076,850	153 (5.7%)

Table 7. Comparative Internal Cluster Validation

Algorithm	Evaluation population	N	Silhouette	DB	CH
K-Means, K = 3	Full sample	2,692	0.4169	-	-
K-Means, K = 3	DBSCAN non-noise subset	2,539	0.4350	0.85	1,850.2
DBSCAN, ε=0.13, min_samples=5	Non-noise subset	2,539	0.2017	1.45	496,24

DB = Davies-Bouldin
CH= Calinski-Harabasz

IV. DISCUSSION

The selected K-Means configuration obtained a higher Silhouette Coefficient and Calinski–Harabasz index and a lower Davies–Bouldin index than DBSCAN on the 2,539 non-noise customers. These results favor K-Means under the reported internal validation criteria. However, the indices assess particular geometric properties and should not be interpreted as establishing universal algorithm superiority. The matched subset was defined by DBSCAN assignments, so the

comparison applies to the customers covered by that configuration. Coverage and the original RFM profiles therefore remain necessary for interpreting the results.

The K-Means solution separated customers by both purchasing activity and Recency. Cluster 2 combined the highest mean Frequency and Monetary, despite containing only 15.0% of customers. Clusters 0 and 1 had lower mean purchasing frequency and expenditure, but differed substantially in Recency. The more recent activity of Cluster 1 did not coincide with the highest cumulative expenditure. These profiles describe observed behavior rather than future profitability or purchasing power.

DBSCAN provided a complementary description by assigning 94.3% of customers to four clusters and leaving 5.7% as noise. Among the assigned clusters, Cluster 2 had the highest mean Frequency, whereas Cluster 3 had the highest mean Monetary and the longest mean Recency. The highest purchasing frequency and highest cumulative expenditure therefore occurred in different clusters. The noise observations had a higher mean Monetary value than any of the four clusters. Thus, the noise label did not correspond to low expenditure and should not be used on its own to exclude customers from further business analysis. These observations require individual examination rather than interpretation as a homogeneous customer segment.

A conclusion based solely on the reported internal indices would favor the selected K-Means configuration. Considering coverage and noise profiling qualifies that interpretation: DBSCAN left 153 customers unassigned, and their expenditure summaries indicated that noise status did not imply low observed purchasing value. Thus, internal cluster quality and the business relevance of unassigned customers address different aspects of the segmentation result.

These profiles could inform hypotheses for differentiated customer engagement. Customers with recent but infrequent purchases could be considered for repeat-purchase initiatives, while customers with higher historical expenditure and longer intervals since their last purchase could be examined for re-engagement opportunities. These are potential applications rather than demonstrated outcomes. Their usefulness would need to be evaluated through campaign response, subsequent purchasing behaviour, and profitability measures.

The contribution of this study is the joint interpretation of internal validation, customer coverage, and purchasing profiles under a shared RFM representation. Matched-subset validation enables comparison on a common evaluation population, while noise profiling shows why observations excluded from cluster validity calculations should remain visible in customer analysis.

The findings are limited to one platform, one observation period, and the reported feature representation and parameter configurations. The matched evaluation subset was defined by DBSCAN, and the results do not establish general algorithm superiority. Customer behavior over time and the effects of segmentation on subsequent purchasing, retention, or profitability were not evaluated.

V. CONCLUSION

This study evaluated K-Means and DBSCAN using min–max-normalized RFM features from 3,868 retained orders and 2,692 customer profiles. K-Means formed three clusters, whereas DBSCAN assigned 94.3% of customers to four clusters and labelled 153 customers as noise. On the matched subset, K-Means obtained a Silhouette Coefficient of 0.4350 compared with 0.2017 for DBSCAN, with the Davies–Bouldin and Calinski–Harabasz indices also favoring the selected K-Means configuration.

The purchasing profiles provided information beyond internal validation scores. The smallest K-Means cluster had the highest mean Frequency and Monetary. For DBSCAN, the highest mean Frequency and Monetary occurred in different clusters, while noise observations had a higher mean Monetary than any identified cluster. These results support interpreting validation scores together with customer coverage and the original RFM profiles.

The findings are specific to the dataset, preprocessing, and configurations examined. The profiles may inform hypotheses for differentiated customer engagement, but their effectiveness for retention or profitability requires prospective evaluation.

REFERENCES

[1] *Laporan Perekonomian Indonesia 2022*. Bank Indonesia, 2022. [Online]. Available: https://www.bi.go.id/id/publikasi/laporan/Pages/LPI_2022.aspx

[2] Fadhilah, "Digitalisasi Sistem Pembayaran Dan Dampaknya Terhadap Inklusi Keuangan Di Indonesia," *Jurnal Ilmu Ekonomi dan Bisnis*, vol. 2, no. 2, pp. 13–18, 2025, doi: 10.70134/jukoni.v2i2.457

[3] *Southeast Asia's Digital Consumers: A new stage of evolution*. Meta and Bain & Company, 2022. [Online]. Available: https://www.bain.com/globalassets/noindex/2022/meta_bain_syncsea_2022.pdf#page=45

[4] C. A. Samosir, "Strategi Customer Relationship Management Untuk Meningkatkan Retensi Pelanggan Di Era Digital," *Repeater: Publikasi Teknik Informatika dan Jaringan*, vol. 3, no. 1, pp. 84–97, 2025, doi: 10.62951/repeater.v3i1.369.

[5] A. Saputra and F. I. Malik, "Segmentasi Pelanggan E-commerce Berbasis Perilaku Transaksi Menggunakan Algoritma K-Means untuk Strategi Pemasaran Bertarget," *COMPUTA: Journal of Computer and Informatics Engineering*, vol. 1, no. 2, Nov. 2025, [Online]. Available: https://computa.ipbcirebon.ac.id/index.php/computa/article/view/20

[6] F. I. Sanjaya and A. D. Kalifia, "Precision Marketing Model using Decision Tree on e-Commerce Case Study Orebae.com", *Journal RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 7, no. 5, pp. 1033-1039, Sep. 2023, doi: 10.29207/resti.v7i5.4531

[7] A. Razia Sulthana, A. Jaiswal, P. Supraja and L. Sairamesh, "Customer Segmentation Using Machine Learning", in *Proc. 3rd International Conference on Advances in Electrical Computing Communication and Sustainable Technologies (ICAECT)*, 2023, doi: 10.1109/ICAECT57570.2023.10117924

[8] R. U. Upadhyay and N. Choudhary, "A Review of Customer Segmentation Methods," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 2, pp. 684–692, Mar. 2023, doi: 10.48175/IJARSCT-8904.

[9] H. Chang and A. Mukherjee, "Machine Learning and Consumer Data," *arXiv.org Cornell University*, Jun. 2023, doi: 10.48550/arXiv.2306.14118.

[10] G. V. D. Ayyappa Babu and M. D. Sathish, "E-Commerce Customer Segmentation Using Machine Learning," *International Journal for Research in Applied Science and Technology*, vol. 10, no. 9, pp. 1275-1277, Sep. 2022, doi: 10.22214/ijraset.2022.46772.

[11] S. S. Mim and D. Logofatu, "A Cluster-based Analysis for Targeting Potential Customers in a Real-world Marketing System," in *Proc. IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2022, pp. 159-166, doi: 10.1109/ICCP56966.2022.10053985.

[12] A. M. Durán-Rosal, P. A. Gutiérrez, F. J. Martínez-Estudillo, and C. Hervás-Martínez, "Simultaneous Optimisation Of Clustering Quality And Approximation Error For Time Series Segmentation," *Information Sciences*, vol. 442-443, pp. 186-201, May 2018, doi: 10.1016/J.INS.2018.02.041.

[13] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, Jun. 2021, doi: 10.3390/E23060759.

[14] R. Punhani, V. P. S. Arora, A. S. Sabitha, and V. K. Shukla, "Segmenting E-Commerce Customer through Data Mining Techniques," *Journal of Physics: Conference Series*, vol. 1714, no. 1, Jan. 2021, doi: 10.1088/1742-6596/1714/1/012026.

[15] N. Gholizadeh, H. Saadatfar, and N. Hanafi, "K-DBSCAN: An Improved DBSCAN Algorithm for Big Data," *The Journal of Supercomputing*, vol. 77, pp. 6214-6235, Jun. 2021, doi: 10.1007/S11227-020-03524-3.

[16] S. Samidi, R. Y. Suladi, and D. Kusumaningsih, "Comparison of the RFM Model's Actual Value and Score Value for Clustering", *Journal RESTI (Rekayasa Sist. Teknol. Inf.),* vol. 7, no. 6, pp. 1430-1438, Dec. 2023, doi: 10.29207/resti.v7i6.5416

[17] G. Aslantaş, M. Gençgül, M. Rumelli, M. Özsaraç, and G. Bakırlı, "Customer segmentation using K-Means clustering algorithm and RFM model," *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi,* vol. 25, no. 74, pp. 491–503, 2023, doi: 10.21205/deufmd.2023257418.

[18] S. Monalisa, Y. Juniarti, E. Saputra, F. Muttakin, and T. K. Ahsyar, "Customer segmentation with RFM models and demographic variable using DBSCAN algorithm," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 21, no. 4, pp. 742–749, Aug. 2023, doi: 10.12928/TELKOMNIKA.v21i4.22759.

[19] C.-G. Wong, G.-K. Tong, and S.-C. Haw, "Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques", *Journal of Telecommunications and the Digital Economy*, vol. 12, no. 3, pp. 97-125, Sep 2024, doi: 10.18080/jtde.v12n3.978.

[20] J. M. John, O. Shobayo, and B. Ogunleye, "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

[21] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227, April 1979, doi: 10.1109/TPAMI.1979.4766909.

[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-Theory and Methods,* vol. 3, no. 1, pp. 1-27, 1974, doi: 10.1080/03610927408827101.