IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

K-Means Retail Product Sales Clustering via Elbow and Davies-Bouldin

Ni Wayan Wardani*[1], Gede Surya Mahendra[2], Putu Sugiartawan[3], Aniek Suryanti Kusuma[4], Ni Luh Wiwik Sri G[5], I Nyoman Agus Suarya Putra[6]
[1,5] Departement of Informatics, Indonesian Institute of Business and Technology, Indonesia
[2]Departement of Information System, Ganesha University of Education, Indonesia
[3,4] Master's Program in Informatics, Indonesian Institute of Business and Technology, Indonesia
[6]Departement of Visual Communication Design, Indonesian Institute of Business and Technology, Indonesia
e-mail: ***[1]niwayan.wardani@instiki.ac.id**, [2]gmahendra@undiksha.ac.id, [3]putu.sugiartawan@instiki.ac.id, [4]anieksuryanti@instiki.ac.id, [5]wiwik@instiki.ac.id, [6]nyomansuarya@instiki.ac.id

Abstrak

Peningkatan skala bisnis ritel menuntut pelaku usaha mengelompokkan produk laris dan kurang laris secara akurat guna mengoptimalkan strategi penyesuaian stok. Penelitian ini bertujuan mengklasterisasi data transaksi penjualan barang serta menentukan jumlah kelompok optimal menggunakan algoritma K-Means. Dataset sekunder sejumlah 7.373 record transaksi dianalisis berdasarkan atribut jumlah transaksi dan total penjualan. Penentuan cluster terbaik dilakukan melalui metode Elbow berbasis Sum of Square Error dan divalidasi menggunakan Davies-Bouldin Index. Hasil pengujian menunjukkan bahwa kedua metode tersebut secara konsisten menghasilkan jumlah cluster optimal k sama dengan 4 dengan nilai Davies-Bouldin Index sebesar 0,47. Nilai tersebut mengindikasikan struktur klaster terpisah dengan sangat baik. Pengelompokan akhir menghasilkan empat kategori performa produk, yaitu sangat laris sebanyak 2 produk, laris sebanyak 20 produk, cukup laris sebanyak 171 produk, dan kurang laris sebanyak 7.180 produk. Evaluasi ini membuktikan bahwa kombinasi K-Means, Elbow, dan Davies-Bouldin Index efektif memberikan wawasan berbasis data untuk mendukung pengambilan keputusan manajemen persediaan barang pada bisnis ritel.

Kata kunci*— Klasterisasi, K-Means, Metode Elbow, Davies-Bouldin Index*

Abstract

The expansion of the retail business requires business operators to accurately group best-selling and slow-moving products in order to optimize inventory adjustment strategies. This study aims to cluster sales transaction data and determine the optimal number of clusters using the K-Means algorithm. A secondary dataset consisting of 7,373 transaction records was analyzed based on the attributes of transaction count and total sales. The best cluster number was determined using the Elbow method based on the Sum of Squared Errors and validated using the Davies-Bouldin Index. The test results show that both methods consistently produced an optimal number of clusters, k, equal to 4, with a Davies-Bouldin Index value of 0.47. This value indicates a very well-separated cluster structure. The final clustering resulted in four product performance categories: 2 "best-selling" products, 20 "selling well" products, 171 "moderately selling" products, and 7,180 "poorly selling" products. This evaluation demonstrates that the

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

combination of K-Means, the Elbow method, and the Davies-Bouldin Index effectively provides data-driven insights to support inventory management decision-making in the retail business.

Keywords— *Clustering, K-Means, Elbow Method, Davies-Bouldin Index*

1. INTRODUCTION

The growth of the retail business in unit and retail sales of goods to consumers continues to increase, driving a corresponding rise in consumer demand [1]. To meet this demand, retail businesses must be able to distinguish between best-selling and slow-moving products so that appropriate sales strategies can be formulated, such as adjusting stock levels and increasing sales transactions [2]. Manually classifying thousands of sales transaction records is impractical, so a systematic, data-driven approach is required to group products according to their sales performance.

Data mining techniques, particularly clustering, have been widely applied to group data based on similarity without relying on labeled classes. The K-Means algorithm is among the most frequently used clustering methods because of its computational efficiency and ease of implementation. Attribute-reduction techniques using information gain have been shown to optimize K-Means cluster formation by removing irrelevant variables [3]. The combination of the Elbow method with K-Means has also been applied to measure students' readiness for a national examination, where the Elbow method determined the optimal number of clusters prior to grouping [4]. In the retail domain, the optimal number of clusters for customer segmentation in a retail store has been determined through internal validation methods to support marketing decisions [5], [6]. K-Means has further been used to cluster clothing sales data to improve sales-strategy formulation [7], while the Elbow method has been applied to determine the best number of clusters for industrial production centers [8]. Beyond the Elbow method, internal validity measures such as the Silhouette Score and Davies-Bouldin Index (DBI) have been compared to evaluate K-Means and DBSCAN clustering results on performance data, showing that DBI is sensitive to cluster compactness and separation [9].

Recent studies published in IJCCS further demonstrate the breadth of K-Means applications across domains. Preprocessing techniques such as normalization and standardization combined with K-Means have improved anomaly-detection performance on payment transaction logs, using DBI as the primary validity measure [10]. Principal Component Analysis (PCA) has been combined with K-Means to reduce dimensionality and improve clustering efficiency for car-customer segmentation [11]. A multi-metric validation approach combining Silhouette Score, DBI, and the Calinski-Harabasz Index has been used to cluster high-school students' career interests, reporting a DBI of 1.88 at the optimal cluster count [12]. K-Means combined with semantic similarity measures has also been applied to estimate hospitalization costs, supported by normalization to improve cluster quality [13]. Elsewhere, K-Means has been compared against Otsu thresholding for detecting tuberculosis bacilli in the HSV color space [14], and applied to analyze and predict palm-oil production clusters in Aceh Utara [15]. Although these studies confirm that K-Means clustering is effective across diverse domains, most sales-related applications rely on a single validity measure and rarely compare the Elbow method directly against DBI on the same dataset, leaving the reliability of the resulting cluster count only partially verified.

This study addresses that gap by applying K-Means clustering to 7,373 product sales transaction records obtained from a public retail dataset, determining the optimal number of clusters through the Elbow method, and independently verifying that result using the Davies-

Bouldin Index. Comparing both methods on the same dataset provides stronger evidence for the chosen number of clusters than relying on either method alone, and the resulting clusters are interpreted directly in terms of product sales performance to support retail stock and sales-strategy decisions.

The main contributions of this study are threefold. First, it applies K-Means clustering to segment retail product sales data into distinct groups based on transaction volume and total sales. Second, it determines the optimal number of clusters using the Elbow method while independently validating the result using the Davies-Bouldin Index. Finally, it interprets the resulting clusters across product sales performance tiers, ranging from slow-moving to best-selling items, to directly support strategic inventory and sales-strategy decisions.

Table 1 Comparison of the Proposed Method with Previous Studies

Methods Used	Other Similarities	Result Performance
Normalization, standardization, PCA, K-Means++, Elbow, Kneedle algorithm, DBI evaluation [10]	K-Means clustering validated with DBI	84 anomalies identified; improved clustering quality with lowest DBI values
Data preprocessing, PCA, K-Means, Elbow Method, Silhouette Score [11]	Dimensionality reduction to improve clustering efficiency	Silhouette Score = 0.7789 with 8 optimal clusters
K-Means, Elbow Method, Silhouette Score, DBI, Calinski-Harabasz Index [12]	Elbow and DBI used jointly for cluster validation	k = 3 optimal; DBI = 1.88
K-Means and C-Means comparison on production/land-area data [15]	Cluster-based segmentation of production data	800.4 ha productive land vs. 343.6 ha less-effective land identified
Proposed Research: Data cleaning, K-Means, Elbow Method (SSE), DBI validation	Retail product sales-performance clustering	k = 4 optimal; DBI = 0.47, confirmed by both Elbow and DBI

2. METHODS

2.1 Research Stages

This study follows a structured clustering pipeline consisting of five distinct stages. First, data is collected from a public retail sales dataset. Second, data preprocessing is performed to remove missing and duplicate records. Third, exploratory data analysis is conducted to examine the data distribution and attribute correlations. Fourth, K-Means clustering is executed, where the optimal number of clusters is determined through the Elbow method. Finally, cluster validity is evaluated using the Davies-Bouldin Index. All processing was implemented in Python using the Jupyter Notebook environment with the pandas, matplotlib, seaborn, and scikit-learn libraries.

2.2 Dataset

The dataset used in this study was obtained from a public repository on Kaggle, consisting of 7,373 records of product sales transactions [16]. Attribute still in Bahasa Indonesia. Each record represents one product and consists of five attributes, as summarized in Table 2.

Table 2 Dataset Attributes

Attribute	Description	Data Type
kode_barang	Product code	Object (string)
nama_barang	Product name	Object (string)
jumlah_transaksi	Number of sales transactions	Integer

total_penjualan	Total quantity sold	Integer
rata_rata	Average sales per transaction	Float

2.3 Data Preprocessing

Before clustering, the dataset was checked for missing values and duplicate records. The check revealed 30 missing values in the nama_barang attribute, which were removed using the dropna() function; a subsequent check confirmed no remaining missing values in that attribute. A duplicate check using the duplicated() function found no fully duplicated rows. The two numerical attributes used for clustering, jumlah_transaksi and total_penjualan, contained no missing values, so all 7,373 product records were retained for the clustering process.

2.4 Determining the Optimal Number of Clusters

The Elbow method identifies the optimal number of clusters (k) by plotting the Sum of Squared Error (SSE), also known as the Within-Cluster Sum of Squares, against a range of k values and locating the point where the rate of decrease slows sharply, forming an "elbow" in the curve, a heuristic first proposed for identifying natural groupings in multivariate data. SSE is computed using Equation (1):

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

where C_i is cluster i, μ_i is the centroid of cluster i, and x is a data point assigned to cluster i. In this study, SSE was computed for k = 2 through k = 10 using the jumlah_transaksi and total_penjualan attributes.

2.5 K-Means Clustering

K-Means, originally introduced as an iterative method for partitioning multivariate observations into k groups, groups data points into k clusters by minimizing the distance between each data point and its assigned cluster centroid, measured using Euclidean distance as shown in Equation (2):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

The algorithm iterates through the following steps until convergence: (1) initialize k centroids; (2) assign each data point to the nearest centroid based on Equation (2); (3) recompute each centroid as the mean of the data points assigned to it; and (4) repeat steps 2 and 3 until no data point changes cluster membership. Based on the Elbow method result, k = 4 was used for the clustering process, applied to the jumlah_transaksi and total_penjualan attributes.

2.6 Cluster Validity Evaluation

Cluster validity was evaluated using the Davies-Bouldin Index (DBI), originally proposed as an internal validity measure based on the ratio of within-cluster to between-cluster distances [18], computed as shown in Equation (3):

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (3)$$

where S_i and S_j denote the average within-cluster distance of cluster i and j, respectively, and M_{ij} denotes the distance between the centroids of cluster i and j. A lower DBI value indicates better clustering quality, as it reflects clusters that are compact internally and well separated from

one another. DBI was computed for k = 2 through k = 10 to independently verify the cluster count identified by the Elbow method.

3. RESULTS AND DISCUSSION

3.1 Dataset Overview

Descriptive statistics for the two numerical attributes used in clustering are summarized in Table 3.

Table 1 Descriptive Statistics of Clustering Attributes

Attribute	Count	Mean	Min	Max
jumlah_transaksi	7,373	38	1	6,457
total_penjualan	7,373	59	1	10,363

To examine relationships between attributes prior to clustering, a correlation heatmap was computed across the three numerical attributes, as shown in Figure 1.

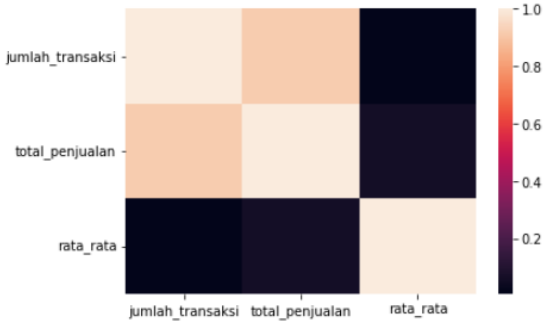


Figure 1 Correlation Heatmap of Numerical Attributes

As shown in Figure 1, jumlah_transaksi and total_penjualan exhibit a strong positive correlation, consistent with the expectation that products with more sales transactions also accumulate higher total sales quantities. In contrast, rata_rata shows a weak correlation with both jumlah_transaksi and total_penjualan, indicating that average sale size per transaction varies largely independently of transaction volume. This pattern supports the choice of jumlah_transaksi and total_penjualan as the primary clustering attributes, since they jointly capture the dominant covariation in the dataset; the role of rata_rata is examined further in Section 3.5.

3.2 Optimal Number of Clusters (Elbow Method)

The Elbow method was applied to the SSE values computed for k = 2 through k = 10. As shown in Figure 2, SSE decreases sharply up to k = 4, after which the rate of decrease becomes noticeably flatter, forming a clear elbow point at k = 4. This point is taken as the optimal number of clusters for the K-Means algorithm.

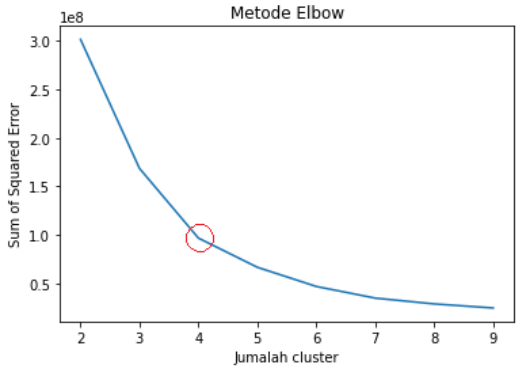


Figure 2 Elbow Method Result (SSE vs. Number of Clusters)

3.3 Cluster Validity (Davies-Bouldin Index)

DBI was computed independently for k = 2 through k = 10 to verify the Elbow method result. As shown in Figure 3, DBI reaches its minimum value of 0.47 at k = 4, before increasing again at higher values of k. Since a lower DBI indicates better-defined and more compact clusters, this result confirms k = 4 as the optimal number of clusters, consistent with the Elbow method.

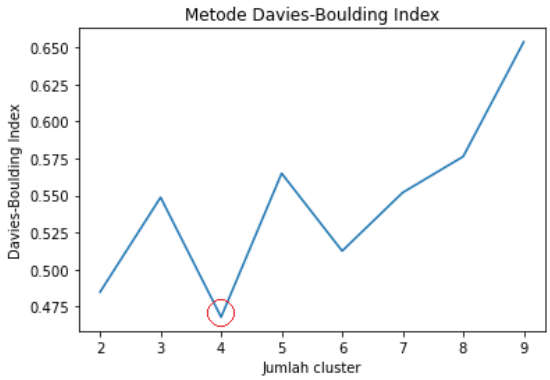


Figure 3 Davies-Bouldin Index Result (DBI vs. Number of Clusters)

3.4 Clustering Results

Applying K-Means with k = 4 to the jumlah_transaksi and total_penjualan attributes produced the cluster structure shown in Figure 4, where the star markers indicate cluster centroids.

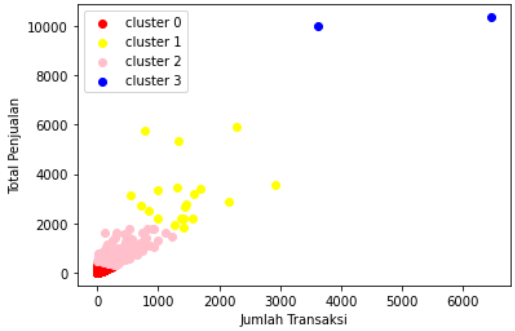


Figure 4 K-Means Clustering Result with Centroids (k = 4)

The distribution of products across the four clusters is summarized in Table 4, and the transaction and sales profile of each cluster is summarized in Table 5.

Table 4 Distribution of Products in Each Cluster

No	Cluster	Sales Category	Number of Products	Percentage
1	Cluster 0	Slow-moving (kurang laris)	7,180	97.38%
2	Cluster 1	Best-selling (laris)	20	0.27%
3	Cluster 2	Moderately selling (cukup laris)	171	2.32%
4	Cluster 3	Top-selling (sangat laris)	2	0.03%

Table 5 Transaction and Sales Profile of Each Cluster

Cluster	Min Trans.	Max Trans.	Mean Trans.	Mean Sales
0	1	290	28	35
1	544	2,919	1,402	3,165
2	4	1,223	356	693
3	3,619	6,457	5,038	10,177

3.5 Robustness Check with an Additional Attribute

Section 3.1 showed that rata_rata correlates weakly with the two primary clustering attributes. To verify that omitting it does not change the optimal cluster count, both the Elbow method and DBI were recomputed using all three attributes, jumlah_transaksi, total_penjualan, and rata_rata, over an extended range of k = 2 to k = 15.

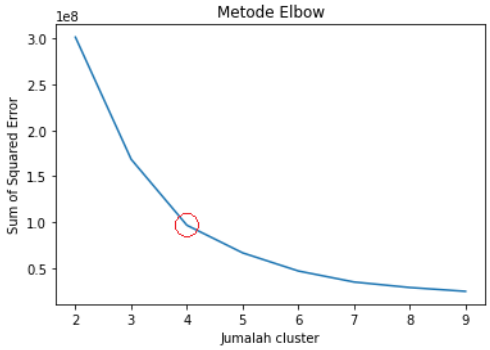


Figure 5 Elbow Method Result with Three Attributes (k = 2 to 15)

As shown in Figure 5, SSE again decreases sharply up to k = 4, after which the curve flattens, reproducing the same elbow point identified using only the two primary attributes.

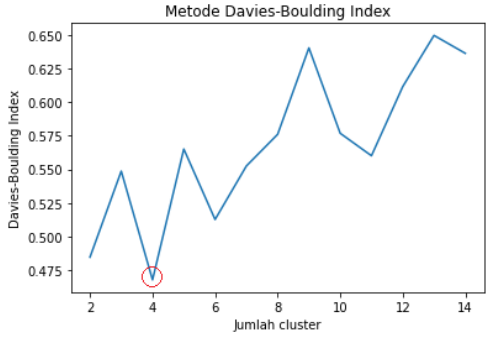


Figure 6 Davies-Bouldin Index Result with Three Attributes (k = 2 to 15)

Figure 6 shows that DBI computed with three attributes also reaches its minimum at k = 4. Table 7 summarizes the outcome of both parameter settings.

Table 7 Elbow and DBI Results Using Two vs. Three Attributes

Parameter Set	Attributes Used	k Range Tested	Elbow k	DBI k
2-Attribute	jumlah_transaksi, total_penjualan	2–10	4	4
3-Attribute	jumlah_transaksi, total_penjualan, rata_rata	2–15	4	4

Both parameter settings converge on k = 4 despite the wider search range used in the three-attribute test, indicating that the optimal cluster count is stable with respect to the inclusion of rata_rata and is not an artifact of the narrower k = 2–10 range used in the primary analysis. While the clustering and interpretation reported in Sections 3.2–3.4 use jumlah_transaksi and total_penjualan for clarity of visualization, this robustness check shows that the underlying number of clusters does not depend on that choice.

3.6 Discussion

The Elbow method and the Davies-Bouldin Index independently converge on the same optimal cluster count, k = 4, which strengthens confidence in the resulting cluster structure beyond what either method could establish alone. The obtained DBI value of 0.47 is notably lower, and therefore better, than the DBI values reported in comparable K-Means clustering studies, such as 1.88 for career-interest clustering [12], indicating that the four clusters identified in this study are relatively compact and well separated.

The cluster distribution reflects a typical retail sales pattern: the great majority of products (97.38%) fall into Cluster 0, the slow-moving category, while only a small number of products drive disproportionately high transaction volumes and sales totals in Clusters 1 and 3. This imbalance is consistent with the underlying business reality of retail sales, where a small set of best-selling products accounts for a large share of total transactions. From a managerial perspective, Cluster 0 represents candidates for stock reduction or promotional strategies, while Clusters 1 and 3 represent products that warrant prioritized restocking to avoid missed sales opportunities. Cluster 2 represents a moderately selling group that may benefit from targeted marketing to shift toward the best-selling categories.

These findings show that combining the Elbow method with the Davies-Bouldin Index provides a more reliable basis for determining the number of clusters in retail sales data than relying on a single validity measure, addressing the gap identified in prior K-Means clustering studies that typically use only one validation method.

3.7 Comparison with Related Clustering Studies

To place these results in a broader context, Table 6 compares the domain, validation approach, and reported outcome of this study against other recent K-Means clustering studies applied to product- and customer-level retail data.

Table 6 Comparison with Related K-Means Clustering Studies on Retail Data

Domain	Validation Approach	Reported Outcome	Domain
Car-customer segmentation [11]	PCA + K-Means, Elbow + Silhouette	Silhouette = 0.7789, k = 8	Car-customer segmentation
Student career-interest clustering [12]	K-Means, Elbow + Silhouette + DBI + CHI	DBI = 1.88, k = 3	Student career-interest clustering
Retail customer segmentation (RFM) [17]	K-Means with combined internal and external validity indices	Cluster validity assessed jointly across multiple indices, including DBI	Retail customer segmentation (RFM)
E-commerce customer profiling (RFM) [18]	K-Means on RFM scores	Customer profiles derived directly from RFM-based cluster centroids	E-commerce customer profiling (RFM)

Domain	Validation Approach	Reported Outcome	Domain
E-commerce segmentation (extended RFM) [19]	Clustering algorithms on an extended RFM feature set	Segment profiles derived from extended behavioral attributes	E-commerce segmentation (extended RFM)
Supermarket transaction segmentation [20]	K-Means, Elbow + Silhouette	Three transaction segments identified	Supermarket transaction segmentation
Proposed Research: Retail product sales-performance clustering	K-Means, Elbow (SSE) + DBI, cross-validating each other	k = 4; DBI = 0.47, lowest DBI among the compared studies	Retail product sales-performance clustering

Table 6 shows that most related studies rely on the Elbow method or the Silhouette Score to select k, using DBI, when reported at all, as one of several supporting indices rather than as an independent cross-check. RFM-based studies cluster customers using behavioral purchase attributes [17], [18], [19], whereas this study and cluster at the product- or transaction-record level [11], [12], [20]. The comparatively low DBI obtained here suggests that, for a two-attribute transaction-volume/sales-total feature space, K-Means can produce well-separated clusters, though this should be interpreted alongside the pronounced size imbalance discussed below rather than as evidence of uniformly higher clustering quality across all compared domains.

Beyond validity comparison, the product-level clusters identified in this study could be combined with transaction-level techniques such as market basket analysis, which has been used to uncover co-purchase patterns and cross-selling opportunities in retail transaction data [21], [22]. Linking the best-selling and top-selling clusters identified here with association-rule mining on the same transactions could reveal which slow-moving products are frequently purchased alongside best-sellers, offering an additional, complementary basis for stock bundling and promotional strategy beyond cluster membership alone.

3.7 Limitations

Several limitations should be considered when interpreting these results. First, although Section 3.5 confirms that the optimal cluster count is stable when rata_rata is included, the cluster assignment, visualization, and profiling reported in Sections 3.2–3.4 still rely on only two attributes; product-category information was not available in the dataset and could reveal additional structure not captured here. Second, the dataset originates from a single public Kaggle source, so the generalizability of the specific cluster boundaries to other retail settings has not been established. Third, both the Elbow method and DBI are internal validity measures computed from the data itself, without an external ground truth; while their agreement on k = 4 is reassuring, it does not by itself guarantee that the clusters align with a business-meaningful segmentation. Finally, the pronounced imbalance in cluster size, ranging from 2 products in Cluster 3 to 7,180 in Cluster 0, means that DBI, which is sensitive to within-cluster scatter, may be disproportionately influenced by the very small clusters; this reinforces the value of cross-checking DBI against the Elbow method, as done in this study, rather than relying on DBI alone.

4. CONCLUSIONS

This study applied the K-Means algorithm to cluster 7,373 retail product sales transaction records based on the number of transactions and total sales. The Elbow method and the Davies-Bouldin Index were used together to determine and independently validate the optimal number of clusters, both converging on k = 4 with a DBI value of 0.47, indicating good cluster quality. The

resulting clusters identify four distinct product sales performance categories, top-selling, best-selling, moderately selling, and slow-moving, with the slow-moving category accounting for the large majority of products. These results provide data-driven insights that can support retail businesses in making stock-management and sales-strategy decisions. Future research is recommended to incorporate additional internal validity measures, such as the Silhouette Score and the Dunn Index, and to include additional attributes, such as customer membership data, to further improve the reliability and granularity of the clustering results.

REFERENCES

[1] H. Matuan, E. Dude, A. Mallo, H. Yowey, Y. S. Patey, and H. Sutejo, "Application of the K-Means Method for Grouping Product Data Based on Sales Level," *J. Ilm. Sist. Inf.*, vol. 5, no. 1, pp. 292–305, Jan. 2026, doi: 10.51903/53pfrd78.

[2] H. Mamiya, K. Crowell, C. L. Mah, A. Quesnel-Vallée, A. Verma, and D. L. Buckeridge, "Characterizing Co-Purchased Food Products with Soda, Fresh Fruits, and Fresh Vegetables Using Loyalty Card Purchasing Data in Montréal, Canada, 2015–2017," *Int. J. Behav. Nutr. Phys. Act.*, vol. 22, no. 1, p. 19, Feb. 2025, doi: 10.1186/s12966-024-01701-8.

[3] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means," *JEPIN J. Edukasi Dan Penelit. Inform.*, vol. 6, no. 1, pp. 48–53, Apr. 2020, doi: 10.26418/jp.v6i1.37606.

[4] N. T. Hartanti, "Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 6, no. 2, pp. 82–89, Aug. 2020, doi: 10.25077/TEKNOSI.v6i2.2020.82-89.

[5] S. I. Murpratiwi, I. G. A. Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *J. Pendidik. Teknol. Dan Kejuru.*, vol. 18, no. 2, pp. 152–163, Sep. 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.

[6] S. Kajornkasirat, P. Gunglin, K. Puangsuwan, and N. Kaewsuwan, "Customer Segmentation Using Association Rule Mining on Retail Transaction Data," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 14, no. 3, pp. 1919–1929, Jun. 2025, doi: 10.11591/ijai.v14.i3.pp1919-1929.

[7] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.

[8] G. Wang, "Customer Segmentation in the Digital Marketing Using a Q-Learning Based Differential Evolution Algorithm Integrated with K-Means Clustering," *PLOS ONE*, vol. 20, no. 2, p. e0318519, Feb. 2025, doi: 10.1371/journal.pone.0318519.

[9] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan DBSCAN," *KAKIFIKOM Kumpul. Artik. Karya Ilm. Fak. Ilmu Komput.*, pp. 60–74, Apr. 2024.

[10] R. Alamsyah and N. Rokhman, "Preprocessing Algorithm for K-Means Anomaly Detection on Payment Logs," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 19, no. 2, pp. 211–222, Apr. 2025, doi: 10.22146/ijccs.105290.

[11] A. R. Savira, "Optimizing Clustering Models Using Principle Component Analysis for Car Customers," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 18, no. 2, Apr. 2024, doi: 10.22146/ijccs.94744.

[12] N. Latifah, I. A. Fatia, and S. Adiyono, "Clustering High School Students' Career Interests Using K-Means with Multi-Metric Validation," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 20, no. 2, pp. 175–186, Apr. 2026, doi: 10.22146/ijccs.117507.

[13] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, "The K-Means Clustering Algorithm With Semantic Similarity To Estimate The Cost of Hospitalization," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 13, no. 4, pp. 313–322, Oct. 2019, doi: 10.22146/ijccs.45093.

[14] B. S. Riza, J. Na'am, and S. Sumijan, "Comparison of K-Means Clustering and Otsu Thresholding Methods in the Detection of Tuberculosis Extra Pulmonary Bacilli in the HSV

Color Space," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 16, no. 3, pp. 227–238, Jul. 2022, doi: 10.22146/ijccs.74531.

[15] M. Ula, G. Perdinanta, R. Hidayat, and I. Sahputra, "Analyze the Clustering and Predicting Results of Palm Oil Production in Aceh Utara," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 17, no. 2, Apr. 2023, doi: 10.22146/ijccs.83195.

[16] "Public Dataset: Cluster Penjualan." Accessed: Aug. 05, 2026. [Online]. Available: https://www.kaggle.com/datasets/ipunguhbpwt/clsuterpenjualan

[17] H. J. Wilbert, A. F. Hoppe, A. Sartori, S. F. Stefenon, and L. A. Silva, "Recency, Frequency, Monetary Value, Clustering, and Internal and External Indices for Customer Segmentation from Retail Data," *Algorithms*, vol. 16, no. 9, p. 396, Sep. 2023, doi: 10.3390/a16090396.

[18] O. P. Barus, C. Nathasya, and J. J. Pangaribuan, "The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering," *Math. Model. Eng. Probl.*, vol. 10, no. 1, pp. 298–303, Feb. 2023, doi: 10.18280/mmep.100135.

[19] T. Ozcan, "Customer Segmentation Using an Extended RFM Model and Clustering Algorithms in E-Commerce," *J. Theor. Appl. Electron. Commer. Res.*, vol. 21, no. 5, p. 142, May 2026, doi: 10.3390/jtaer21050142.

[20] F. M. Khoiriyyah, H. Suhendar, and Y. Maulana, "Transaction Segmentation of Supermarket Sales Data for Retail Decision Support Using K-Means Clustering," *J. Intell. Syst. Technol. Inform.*, vol. 2, no. 1, pp. 45–50, Mar. 2026, doi: 10.64878/jistics.v2i1.145.

[21] I. D. Hunyadi, N. Constantinescu, and O.-A. Țicleanu, "Efficient Discovery of Association Rules in E-Commerce: Comparing Candidate Generation and Pattern Growth Techniques," *Appl. Sci.*, vol. 15, no. 10, p. 5498, May 2025, doi: 10.3390/app15105498.

[22] M. M. Alawadh and A. M. Barnawi, "A Survey on Methods and Applications of Intelligent Market Basket Analysis Based on Association Rule," *J. Big Data*, vol. 4, no. 1, pp. 1–25, 2022, doi: 10.32604/jbd.2022.021744.