

Clustering the Sales Levels of Thrift Clothing on the Shopee Platform Using the K-Means Algorithm

Putra Dwi Kurniawan¹, Theresia Dwiati Wismarini²

Informatics Engineering Study Program, Faculty of Information and Industrial Technology
Universitas Stikubank (Unisbank) Semarang

The Shopee e-commerce platform has recorded exponential transaction growth, with Indonesia becoming its largest global market, while the thrift clothing segment has emerged as one of the significant sales categories within the platform. Thrift clothing sellers face challenges in objectively identifying daily sales performance patterns because transaction data are highly fluctuating. This study applies the K-Means algorithm within the Knowledge Discovery in Databases framework to cluster daily thrift clothing sales levels based on the variables of Total Sales and Total Orders. Primary data were obtained from the Shopee Seller Center of a thrift clothing store during the period from December 2025 to May 2026, resulting in 182 daily records, which were filtered into 168 active operational days after excluding 14 inactive days. Numerical variables were normalized using the Min-Max Normalization method before being processed with K-Means in RapidMiner Studio, while the optimal number of clusters was determined using the Elbow Method. Cluster quality evaluation was conducted using the Davies Bouldin Index in RapidMiner and the Silhouette Score, which was independently validated using Orange Data Mining. The results indicate that $K = 3$ is the optimal number of clusters, with a Davies Bouldin Index of 0.693 and a Silhouette Score of 0.512, both categorized as good. The clustering process produced three categories: Low Cluster (101 days, 60.1%, average IDR 134,139/day), Medium Cluster (50 days, 29.8%, average IDR 362,131/day), and High Cluster (17 days, 10.1%, average IDR 873,521/day). The High Cluster was concentrated in the April May 2026 period and contributed 31.9% of the total six-month revenue. These findings provide an empirical basis for thrift clothing sellers to formulate more targeted promotional strategies, inventory management, and marketing budget allocation based on historical data patterns.

Keywords: K-Means; clustering; thrift clothing; Shopee; data mining; Knowledge Discovery in Databases.

This is an open-access article under the [CC BY-NC](#) license



Corresponding Author:

Putra Dwi Kurniawan

Informatics Engineering Study Program, Faculty of Information and Industrial Technology Universitas Stikubank (Unisbank) Semarang
putradwi3014@mhs.unisbank.ac.id

1. Introduction

The rapid advancement of information and communication technology has significantly transformed the global trading system, including in Indonesia. Electronic commerce (e-commerce) has become a key driver of the country's digital economy, with its growth continuing to increase substantially. According to the 2023 E-Commerce Statistics Survey published by Statistics Indonesia (BPS), the number of e-commerce businesses in Indonesia reached 3.82 million in 2023, representing a 27.40% increase compared to the previous year. During the same period, the total value of national e-commerce transactions reached IDR 1,100.87 trillion, reflecting a 40% increase from IDR 783 trillion recorded in 2022.

The growth of Indonesia's e-commerce sector continued in 2024, with the number of e-commerce businesses increasing to 4.4 million, up 15.3% from 2023, while total transaction value reached IDR 1,288.93 trillion. This demonstrates that digital commerce has become a vital component of the national economy, supporting businesses of all sizes. Among Indonesian marketplaces, Shopee remains the dominant platform, recording an average of 124.9 million monthly visits from Indonesian users in 2023 [1]. Globally, Shopee had 295 million active users in 2023, with Indonesia contributing approximately 103 million users. In

2024, Shopee processed 10.9 billion gross orders with a Gross Merchandise Value (GMV) of USD 100.5 billion, highlighting its strategic role for online businesses, including thrift clothing sellers.

Within this ecosystem, the thrift clothing market has become an important business segment. Thrift clothing refers to wearable second-hand apparel sold at more affordable prices than new products [2]. Beyond affordability, thrifting has evolved into a lifestyle associated with sustainable fashion and self-expression, particularly among Generation Z and millennials[3]. A GoodStats survey found that 49.4% of Indonesians had purchased thrift clothing, while only 16.1% expressed no interest [4]. The market potential is also substantial: the global second-hand apparel market was valued at USD 42.5 billion in 2023 and is projected to reach USD 113.2 billion by 2032. In the Asia-Pacific region, the market is expected to grow from USD 54 billion in 2024 to USD 164.5 billion by 2034. In Indonesia, imports of used clothing increased dramatically from 7 tons in 2021 to 3,600 tons in 2024, with 1,800 tons imported between January and August 2025, reflecting strong and growing demand.

Despite this market potential, thrift sellers often struggle to understand daily sales patterns because sales fluctuate due to factors such as weekends, holidays, Shopee promotional campaigns, fashion trends, and seasonal effects. Although sellers possess daily transaction data, including sales value and total orders, these data provide limited insight without appropriate analytical methods. Consequently, operational decisions regarding inventory, promotions, advertising budgets, and resource allocation are often based on intuition rather than data[5].

Data mining provides an effective solution by extracting valuable knowledge from large datasets [6]. Within the Knowledge Discovery in Databases (KDD) framework, data mining involves data selection, preprocessing, transformation, modeling, evaluation, and knowledge interpretation [7]. One suitable technique is clustering, which groups data with similar characteristics. Among clustering methods, K-Means is widely used because of its simplicity, efficiency, and ability to produce interpretable clusters using Euclidean Distance [8]). In this study, K-Means is applied to classify daily thrift clothing sales into high-, medium-, and low-sales clusters, enabling objective, data-driven analysis of sales patterns.

Previous studies have demonstrated the effectiveness of K-Means in e-commerce sales analysis. [9] successfully clustered 18,119 Shopee transaction records using the CRISP-DM framework to identify popular and less popular products. [10] applied K-Means within the KDD framework to Shopee supplier data and identified $K = 3$ as the optimal cluster solution with a Davies-Bouldin Index (DBI) of 0.213, effectively classifying suppliers by transaction performance. Similarly, [11]) employed K-Means with the Recency, Frequency, and Monetary (RFM) model, producing three product clusters with a Sum of Squared Error (SSE) of 0.0041. However, no previous study has specifically examined daily sales clustering for thrift clothing on Shopee. Given the unique characteristics of thrift products including highly volatile demand, dependence on promotional events, limited and unique inventory, and strong sensitivity to rapidly changing social media trends this study addresses an important research gap and offers both originality and practical significance.

Previous studies have demonstrated the effectiveness of K-Means for analyzing e-commerce sales data, but their analytical focus differs from the problem addressed in this study. [12], for example, clustered 18,119 Shopee transaction records to identify popular and less popular products, while [13] focused on supplier performance using Shopee supplier data. [14] combined K-Means with the Recency, Frequency, and Monetary (RFM) model to classify products for inventory prediction. These studies primarily emphasize product-, supplier-, or customer-related segmentation rather than identifying temporal variations in daily sales performance within a specific and highly dynamic product category. Therefore, the research gap is not limited to the absence of a study applying K-Means to thrift clothing, but also concerns the lack of an

analytical approach that treats the day of operation as the fundamental unit of observation for understanding fluctuations in thrift clothing sales.

The novelty of this study lies in the development of a daily sales clustering approach specifically adapted to the operational characteristics of thrift clothing businesses on Shopee. Unlike conventional e-commerce products that can generally be restocked in relatively standardized quantities, thrift clothing is characterized by limited and non-uniform inventory, product uniqueness, rapidly changing consumer preferences, and demand that can fluctuate substantially across short periods. Consequently, aggregating transactions solely at the product or customer level may not adequately reveal when the business experiences low-, medium-, or high-performance sales conditions. By using each active sales day as an observation and jointly clustering Total Sales and Total Orders, this study captures daily variations in both monetary performance and transaction volume [15]. The approach therefore extends the application of K-Means from conventional transaction or product segmentation toward temporal performance segmentation at the operational level of a thrift clothing store.

Furthermore, the analytical contribution is strengthened by integrating Min-Max normalization, Elbow-based determination of the optimal cluster number, K-Means clustering, and two complementary internal validation measures, namely the Davies Bouldin Index and Silhouette Score. The exclusion of inactive store days during preprocessing also ensures that the resulting clusters represent actual operating conditions rather than periods without sales activity. This design enables the resulting clusters to function not merely as descriptive labels, but as interpretable indicators of different daily business conditions that can subsequently be associated with promotional timing, inventory preparation, and marketing resource allocation. Thus, the contribution of this study is the formulation and empirical testing of a data-mining workflow for identifying daily sales-performance patterns in a volatile, limited-inventory, and trend-sensitive thrift clothing environment. This provides a more context-sensitive application of K-Means and demonstrates how unsupervised learning can be used to transform raw daily marketplace transactions into operationally meaningful sales segments.

2. Methods

Data Type and Source

This study uses primary data obtained directly from the researcher's thrift clothing store operating on the Shopee platform. The data were collected through the Export Report feature in Shopee Seller Center, a web-based platform for managing stores and accessing sales reports. The reports were downloaded in .xlsx format for subsequent data processing and analysis.

Dataset Attributes

The dataset consists of three attributes obtained from Shopee Seller Center: Transaction Date, Total Sales, and Total Products Sold. Transaction Date serves only as an observation identifier, while Total Sales (IDR) and Total Products Sold (units) are used as the clustering variables in the K-Means algorithm. These two variables were selected because they comprehensively represent daily sales performance by capturing both revenue and sales volume, enabling meaningful clustering of daily thrift clothing sales.

Data Observation Period

The dataset covers the period from December 2025 to May 2026, representing six consecutive months of daily sales. This period includes different sales conditions, such as the year-end season, the beginning of the new year, Ramadan, and the post-Eid period, providing sufficient variation in daily sales patterns. A

maximum of 182 daily records could be obtained, with each active sales day representing one observation used in the clustering process.

Tools Used

This study employed several software tools to support data collection, preprocessing, K-Means clustering, cluster quality evaluation, and research documentation, covering all stages from extracting daily sales data from Shopee Seller Center to validating the clustering results.

RapidMiner Studio was selected because it provides comprehensive data analysis features that support the implementation of the K-Means clustering algorithm. Its visual interface enables a systematic workflow, including data loading, normalization, clustering, and result evaluation, without requiring complex programming.

In addition, Orange Data Mining was used as a validation tool to calculate the Silhouette Score and generate the Silhouette Plot, allowing cluster quality to be assessed in terms of both intra-cluster cohesion and inter-cluster separation. The combined use of these two tools enhances the reliability of the clustering evaluation.

3. Result and Discussion

Data Transformation Results

Before the cleaned dataset of 168 records was processed using the K-Means algorithm in RapidMiner Studio, the two input variables were normalized using the Normalize operator with the Range Transformation method, applying Min-Max Normalization to scale all values into the [0,1] range. Normalization was necessary because the variables had different scales. Total Sales ranged from IDR 13,000 to IDR 1,363,000, while Total Orders ranged from 1 to 13 units. Without normalization, Total Sales would dominate the Euclidean Distance calculation. Therefore, normalization ensured that both variables contributed equally to the clustering process.

Table 1. Sample Results of Min-Max Normalization for Five Representative Records

Date	Total Sales	Total Orders	Total_Sales_Norm	Total_Orders_Norm
02-12-2025	IDR 53,000	1	0,029	0,000
12-12-2025	IDR 523,099	11	0,377	0,833
04-05-2026	IDR 1,363,000	6	1,000	0,416
03-05-2026	IDR 1,266,000	13	0,928	1,000
13-12-2025	IDR 17,000	1	0,003	0,000

Based on Table 1, May 4, 2026 obtained a normalized value of 1.000 for the Total Sales variable because it recorded the highest sales in the cleaned dataset (IDR 1,363,000). Meanwhile, May 3, 2026 received a normalized value of 1.000 for the Total Orders variable, representing the highest number of orders (13 units). After normalization, the output was passed to the K-Means Clustering operator to group daily sales based on similar patterns.

Determining the Optimal Number of Clusters Using the Elbow Method

The optimal number of clusters was determined using the Elbow Method in RapidMiner Studio. The K-Means algorithm was executed with K values ranging from 2 to 8, and each solution was evaluated using the Average Within Centroid Distance generated by RapidMiner's cluster evaluation operator. This metric measures the average distance between cluster members and their centroids; lower values indicate more compact clusters. The evaluation results for each K value are presented in Table 2 and illustrated in Figure 1.

Table 2. Average Within Centroid Distance Values for Each K Value Using the Elbow Method

K Value	Average Within-Centroid Distance	Reduction from Previous K
K = 2	0.033	
K = 3	0.019	42.4%
K = 4	0.013	31.6%
K = 5	0.009	30.8%
K = 6	0.008	11.1%
K = 7	0.007	12.5%
K = 8	0.006	14.3%

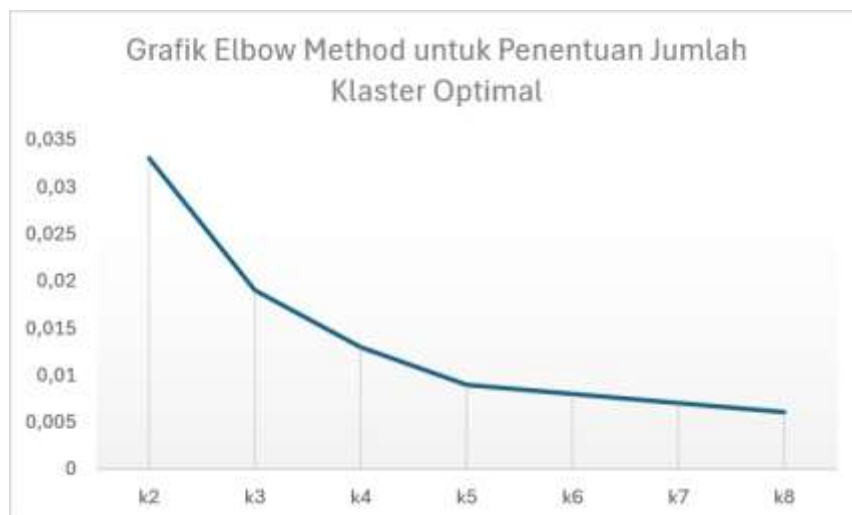


Figure 1. Elbow Method Graph for Determining the Optimal Number of Clusters

Figure 1 shows that the largest reduction in Average Within Centroid Distance occurred from K = 2 to K = 3 (0.033 to 0.019), identifying K = 3 as the optimal number of clusters. This choice also provides a practical business classification of low, medium, and high sales.

K-Means Clustering Implementation

The K-Means algorithm was implemented in RapidMiner Studio through a workflow consisting of data loading, normalization, clustering, and cluster evaluation, as illustrated in Figure 2.

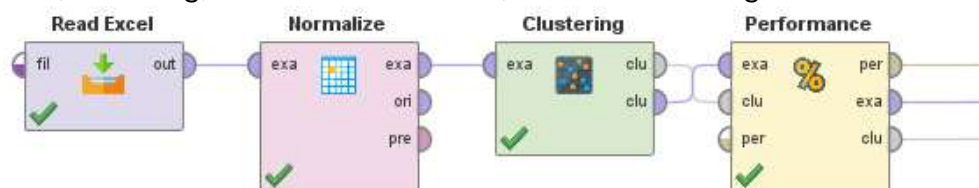


Figure 2. K-Means Clustering Process Flow in RapidMiner Studio

Figure 2 presents the RapidMiner Studio workflow used for clustering. The process consists of four main operators: Read Excel to load the 168-record dataset, Normalize to apply Min-Max normalization (range 0 1), K-Means Clustering with K = 3, 10 max runs, 300 maximum iterations, and Euclidean Distance, followed by Performance (Cluster Distance Performance) to evaluate the clustering results using the Davies-Bouldin Index (DBI)

Table 3. K-Means Parameters Used in RapidMiner Studio

Parameter	Configured Value	Description
Number of Clusters (k)	3	Used to create three daily sales clusters: Low, Medium, and High.

Parameter	Configured Value	Description
Max Runs	10	Specifies that the K-Means algorithm is executed multiple times to obtain more stable clustering results.
Maximum Iterations	300	Defines the maximum number of centroid update iterations until the clustering process reaches convergence.
Distance Metric	Euclidean Distance	Used to calculate the distance between data points and centroids during the K-Means clustering process.
Additional Evaluation Metric	Davies Bouldin Index (DBI)	Used to validate the quality of the resulting clusters.

K-Means Clustering Results (K = 3)

Applying the K-Means algorithm with K = 3 to the 168-record dataset in RapidMiner Studio produced three clusters with distinct final centroids. The clusters were labeled low, medium, and high based on the normalized centroid values of Total Sales and Total Orders, where higher centroid values indicate better daily sales performance. The clustering results are presented in Table 4 and visualized in Figure 3.

Table 4. Final Centroid Positions of the Three Clusters Based on Normalized Values

RapidMiner Output Label	Cluster Interpretation	Number of Records	Centroid Sales (Normalized)	Centroid Order (Normalized)
cluster_0	Low Cluster	101 (60.1%)	0.090	0.115
cluster_2	Medium Cluster	50 (29.8%)	0.259	0.382
cluster_1	High Cluster	17 (10.1%)	0.637	0.735



Figure 3. Scatter Plot of K-Means Clustering Results K=3 on Daily Sales Data in RapidMiner Studio

Figure 3 illustrates the distribution of 168 daily sales records in a normalized two-dimensional space. The clusters are automatically labeled by RapidMiner as cluster_0 (low sales), cluster_2 (medium sales), and cluster_1 (high sales) based on their centroid values. The results show that most observations belong to the low-sales cluster, while the high-sales cluster contains the fewest records.

Cluster Quality Evaluation

Cluster quality was assessed using two internal validation metrics: the Davies-Bouldin Index (DBI) calculated in RapidMiner Studio and the Silhouette Coefficient (SC) calculated and visualized in Orange Data Mining, providing complementary validation of the clustering results.

Davies-Bouldin Evaluation Results

The Davies-Bouldin Index (DBI) was computed using the Cluster Distance Performance operator in RapidMiner Studio to evaluate the compactness and separation of the clusters. The evaluation results are presented in Figure 4.

Davies Bouldin

Davies Bouldin: 0.693

Figure 4. Cluster Distance Performance Evaluation Results on RapidMiner Studio

As shown in Figure 4, the Davies-Bouldin Index (DBI) is 0.693, indicating good clustering quality, as values below 1 reflect compact and well-separated clusters. Therefore, the K-Means model with $K = 3$ is suitable for analyzing daily thrift clothing sales patterns on Shopee.

Silhouette Score Validation

To further validate the clustering results, the Silhouette Score was calculated using the normalized 168-record dataset in Orange Data Mining. The validation workflow consisted of File, Select Columns, K-Means ($K = 3$), and Silhouette Plot, which produced both the Silhouette Score and a visualization of cluster quality.



Figure 5. Silhouette Score Validation Process Flow

Figure 5 illustrates the Silhouette Score validation workflow in Orange Data Mining, consisting of File, Select Columns, K-Means ($K = 3$), and Silhouette Plot. The analysis used normalized data with 10 reruns and 300 maximum iterations to ensure stable clustering. The evaluation produced a Silhouette Score of 0.512, indicating good cluster quality (score > 0.5). This result suggests that the clusters are reasonably compact and well separated, confirming that the K-Means model ($K = 3$) is appropriate for analyzing daily thrift clothing sales patterns.

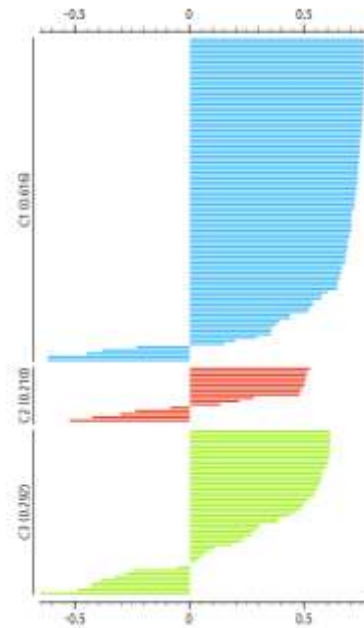


Figure 6. Visualization of Silhouette Score Results of K-Means Clustering with K=3

Figure 6 presents the Silhouette Score distribution for each cluster. Although a few observations have low silhouette values, the overall score of 0.512 indicates that the clustering results are reliable and support the three-cluster solution (low, medium, and high sales).

Cluster Characteristics and Descriptive Analysis

After clustering, a descriptive statistical analysis was performed using the original Total Sales and Total Orders values (converted back from normalized data) to facilitate interpretation in terms of IDR and units sold. The descriptive statistics for each cluster are presented in Table 5.

Table 5. Summary of Descriptive Statistics per Cluster Based on Actual Values

Cluster	Number of Days (%)	Average Sales per Day	Average Orders per Day	Minimum Sales	Maximum Sales
Low	101 (60.1%)	IDR 134,139	2.38 orders	IDR 13,000	IDR 400,000
Medium	50 (29.8%)	IDR 362,131	5.58 orders	IDR 172,100	IDR 693,000
High	17 (10.1%)	IDR 873,521	9.82 orders	IDR 523,099	IDR 1,363,000
Total	168 (100%)	IDR 276,812	4.08 orders	IDR 13,000	IDR 1,363,000

As shown in Table 5, the Low Cluster contains 101 days (60.1%), with an average daily sales value of IDR 134,139 and 2.38 orders. The Medium Cluster includes 50 days (29.8%), averaging IDR 362,131 and 5.58 orders per day. The High Cluster, although the smallest with 17 days (10.1%), achieved the highest performance, averaging IDR 873,521 and 9.82 orders per day. These results demonstrate that K-Means effectively distinguished daily sales into three meaningful performance levels.

Low Cluster (Cluster_0)

The Low Cluster represents 101 days (60.1%) with sales ranging from IDR 13,000 to IDR 400,000. It contributed IDR 13,548,074 (29.1%) of total revenue, indicating that most operating days experienced relatively low sales.

Medium Cluster (Cluster_2)

The Medium Cluster consists of 50 days (29.8%), with sales ranging from IDR 172,100 to IDR 693,000. It generated IDR 18,106,538 (38.9%) of total revenue, making it the largest contributor during the observation period.

High Cluster (Cluster_1)

The High Cluster includes 17 days (10.1%) with the strongest sales performance. It contributed IDR 14,849,851 (31.9%) of total revenue, despite representing the fewest observations. Most high-performing days occurred during April May 2026, particularly May 3 4, 2026, suggesting that the post-Eid period was the store's peak sales season. These findings provide valuable insights for planning promotions and inventory management.

Analysis of Cluster Proportion and Distribution

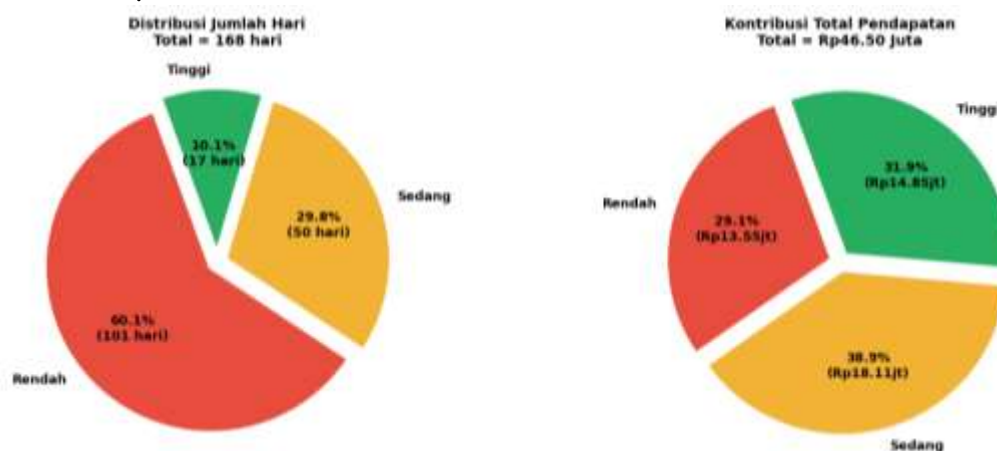


Figure 7. Pie Chart: Proportion of Number of Days (Left) and Revenue Contribution (Right) per Cluster

Figure 7 compares the proportion of operating days and revenue contribution across the three clusters. Although the Low Cluster accounted for the largest share of operating days (60.1%), it contributed only 29.1% of total revenue, indicating relatively low daily sales performance. In contrast, the High Cluster represented only 10.1% of operating days but generated 31.9% of total revenue. This highlights the significant impact of high-performing days and suggests that their patterns should be used to optimize promotional strategies, inventory management, and sales timing (Febrianty et al., 2023).

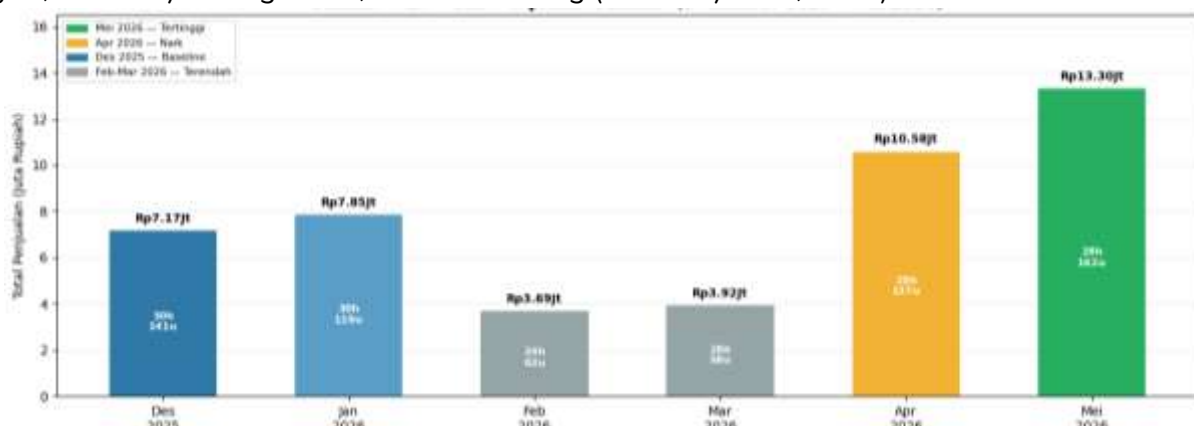


Figure 8. Total Monthly Sales After Data Cleansing for the Period December 2025 May 2026

Table 6. Summary of Total Sales and Orders per Month After Data Cleansing

Month	Number of Days	Total Sales (IDR)	Total Orders (Units)	Remarks
December 2025	30 days	IDR 7,166,779	141	National Online Shopping Day (12.12); moderate performance
January 2026	30 days	IDR 7,847,989	119	Relatively stable performance
February 2026	24 days	IDR 3,691,008	69	Second-lowest performance
March 2026	28 days	IDR 3,922,999	58	Lowest performance
April 2026	28 days	IDR 10,577,598	137	Performance began to improve during the Eid al-Fitr period
May 2026	28 days	IDR 13,298,090	162	Highest performance
Total	168 days	IDR 46,504,463	686	

ased on Table 6 and Figure 8, the highest monthly revenue was recorded in May 2026 at IDR 13,298,090 (28.6%), followed by April 2026 with IDR 10,577,598 (22.7%). March 2026 recorded the lowest revenue at IDR 3,922,999 (8.4%). Overall, the cleaned dataset generated IDR 46,504,463 in total revenue from 686 orders across 168 active operating days. The increase in sales during April and May 2026 supports the clustering results, indicating that high-sales days were concentrated toward the end of the observation period.

Discussion of Clustering Results

The temporal clustering pattern shows that High Cluster days were concentrated in April May 2026 (post-Eid period) rather than during the 12.12 shopping campaign. This suggests that high thrift clothing sales were influenced not only by platform-wide promotions but also by internal factors such as inventory availability, product variety, and product quality, consistent with the findings of Zamani dkk (2024). In contrast, the Medium Cluster was distributed throughout the observation period, while the Low Cluster was concentrated in February March 2026, when sales were weakest. Despite representing only 10.1% of operating days, the High Cluster generated 31.9% of total revenue, with an average daily sales value of IDR 873,521, compared to IDR 362,131 for the Medium Cluster and IDR 134,139 for the Low Cluster. This demonstrates the substantial impact of a relatively small number of high-performing days on overall revenue.

The clustering model showed good performance, with a Davies-Bouldin Index (DBI) of 0.693 and a Silhouette Score of 0.512, indicating compact and well-separated clusters suitable for daily sales analysis Ramadhan dkk (2026). These findings are consistent with previous studies applying K-Means to sales data (Dermawan dkk., 2023; Harjono et al., 2023). Unlike earlier research, this study focuses on daily temporal clustering of thrift clothing sales and excludes inactive store days during preprocessing to better represent actual business activity.

Business Strategy Recommendations Based on Clustering Results

The K-Means clustering results classified daily sales into high-, medium-, and low-sales clusters, each with distinct sales performance and revenue contributions. These differences provide a practical basis for developing targeted business strategies, as summarized in Table 7.

Table 7. Business Strategy Recommendations Based on K-Means Clustering Results

Cluster	Characteristics	Pattern Findings	Strategic Recommendations
High	Average daily sales of IDR 873,521 with an average of 9.82 orders per day; represents 17 days (10.1%) of the total observation period.	Predominantly observed during April May 2026, particularly after the Eid al-Fitr holiday.	Identify the factors contributing to increased sales, such as adequate inventory availability and product variety; replicate the effective strategies implemented during April May; prepare inventory in advance for high-potential periods; and consider using paid advertising during periods with strong sales potential.
Medium	Average daily sales of IDR 362,131 with an average of 5.58 orders per day; represents 50 days (29.8%) of the total observation period.	Distributed across multiple periods and serves as a transitional category between low and high performance.	Use this cluster as the primary target for improvement toward the High Cluster; increase promotional activities; update product listings more consistently; and monitor the factors that distinguish medium-performance days from low-performance days.
Low	Average daily sales of IDR 134,139 with an average of 2.38 orders per day; represents 101 days (60.1%) of the total observation period.	Most frequently observed during February March 2026.	Evaluate the causes of low sales during February March, including possible inventory shortages; use low-demand periods for inventory replenishment and product research; and implement limited promotional campaigns to increase daily sales activity.

The recommendations in Table 7 indicate that each cluster requires a different business strategy based on its sales characteristics. For the High Cluster, the priority is to maintain and replicate the factors driving strong sales, particularly during April May 2026, by optimizing inventory availability, product variety, and promotional timing. For the Medium Cluster, the focus should be on improving performance through consistent product updates and regular promotional activities. For the Low Cluster, efforts should concentrate on identifying the causes of weak sales, especially during February March 2026, when inventory limitations were observed. Thus, the clustering results provide not only a sales classification but also a practical basis for data-driven operational decision-making.

The K-Means clustering results demonstrate that daily thrift clothing sales can be meaningfully differentiated into three performance levels, namely Low, Medium, and High. The three clusters were formed based on the combined characteristics of Total Sales and Total Orders after Min-Max normalization, rather than on a predetermined sales threshold. This distinction is important because the resulting categories represent patterns emerging from the structure of the observed data rather than classifications imposed by the researcher. The Low Cluster consisted of 101 days (60.1%), the Medium Cluster comprised 50 days (29.8%), and the High Cluster contained 17 days (10.1%). The progressive increase in both sales and order centroids across the three groups indicates that the clustering process successfully captured differences in daily commercial performance.

The formation of the Low Cluster can be interpreted as an indication of the store's dominant baseline sales condition. This cluster accounted for 60.1% of active operating days, with average daily sales of IDR

134,139 and an average of 2.38 orders. The relatively large proportion of observations in this cluster suggests that low sales were not isolated events but represented the most frequently occurring operating condition during the observation period. This finding is consistent with the monthly sales pattern, particularly during February and March 2026, when the store recorded relatively weak sales performance. The existence of a large low-sales cluster is important from a business perspective because it indicates that the store's overall revenue is not generated uniformly across operating days. Instead, most days operate at a relatively modest sales level, while a smaller number of high-performing days contribute disproportionately to total revenue.

The Medium Cluster represents an intermediate sales condition, consisting of 50 days (29.8%) with average daily sales of IDR 362,131 and 5.58 orders. Scientifically, this cluster can be interpreted as a transitional performance state between the store's routine low-sales condition and its peak-sales condition. Its position between the Low and High clusters indicates that daily sales performance does not simply consist of extreme low and high outcomes, but contains an intermediate pattern that may reflect relatively stable demand or partial success of promotional and merchandising activities. The Medium Cluster also generated the largest proportion of total revenue, contributing IDR 18,106,538 or 38.9% of total revenue. Therefore, although the High Cluster has the highest sales intensity, the Medium Cluster is economically important because its relatively large number of operating days makes it the largest cumulative contributor to revenue.

The High Cluster provides the most notable finding in terms of business concentration. Although it consisted of only 17 days (10.1%) of the total active operating period, the cluster recorded an average daily sales value of IDR 873,521 and 9.82 orders and contributed IDR 14,849,851, equivalent to 31.9% of total revenue. This means that a relatively small proportion of operating days generated almost one-third of the store's total revenue. Such a pattern indicates that sales performance in the thrift clothing business is highly uneven across time. In the context of thrift products, this concentration can plausibly be associated with the characteristics of limited and non-standardized inventory, product uniqueness, changing consumer preferences, and the timing of product exposure or promotional activity. However, because the present study only clusters sales outcomes and does not include explanatory variables such as promotion intensity, product type, price, advertising expenditure, or inventory availability, these factors should be interpreted as possible explanations rather than causal determinants.

The temporal distribution of the clusters provides additional insight into the observed sales dynamics. High Cluster observations were concentrated mainly in April May 2026, whereas Low Cluster observations were more prevalent during February March 2026. May 2026 recorded the highest monthly revenue at IDR 13,298,090, while March 2026 recorded the lowest at IDR 3,922,999. This temporal concentration indicates that the clustering results capture meaningful changes in sales conditions over the observation period rather than merely separating observations randomly in the feature space. The concentration of high-performing days after the Eid al-Fitr period is particularly noteworthy because it suggests that the store's sales performance was influenced by temporal business conditions. Nevertheless, the result should not be interpreted as evidence that the post-Eid period directly caused higher sales. Previous research cited in this study indicates that product variety, price, and product quality can influence purchasing decisions in thrift markets, while the present dataset does not contain these variables. Therefore, the temporal pattern should be regarded as an empirical association that provides a basis for further investigation rather than a causal conclusion.

These findings are broadly consistent with previous applications of K-Means in e-commerce and sales analysis. Dermawan et al. (2023) demonstrated that K-Means could distinguish differences in product sales performance using Shopee transaction data, while Harjono et al. (2023) showed that K-Means could classify sales levels into meaningful performance groups. The present study extends these findings by shifting

the unit of analysis from products or broader sales entities toward individual active operating days within a thrift clothing store. This distinction is important because thrift clothing has characteristics that differ from standardized retail products, particularly limited quantities, product uniqueness, and rapidly changing demand [16]. Consequently, daily clustering provides a perspective that can reveal temporal performance conditions that may be obscured when transactions are aggregated only at the product level [17].

The quality evaluation further supports the interpretation that the three clusters are sufficiently distinguishable. The Davies Bouldin Index (DBI) of 0.693 indicates relatively compact and separated clusters, while the Silhouette Score of 0.512 provides additional evidence that observations generally have greater similarity to their assigned cluster than to alternative clusters. The use of two evaluation measures is valuable because relying on a single metric could provide a limited assessment of clustering quality. The consistency between the DBI and Silhouette results strengthens the conclusion that $K = 3$ provides a reasonable representation of the observed daily sales structure. However, these metrics should not be interpreted as evidence that the three clusters represent universal sales categories for all thrift clothing businesses. They only establish the quality of separation within the dataset analyzed in this study.

From a data-mining perspective, the results demonstrate the usefulness of unsupervised learning for converting raw marketplace transaction records into interpretable operational information [18]. The original data consisted of daily sales values and order volumes that, when examined individually, provide limited information about the broader structure of sales performance. Through K-Means clustering, these observations were transformed into three identifiable performance states. This transformation is particularly relevant for businesses that lack predefined sales categories because the cluster boundaries are derived from the observed distribution of the data. The resulting classification can therefore serve as a descriptive decision-support mechanism for identifying periods requiring different operational responses [19].

The findings also have practical implications. High Cluster periods can be used as reference conditions for examining inventory availability, product variety, promotional activity, and other operational factors that coincided with strong sales. Medium Cluster periods may represent opportunities for improvement because their sales performance is already above the store's baseline but has not reached the high-performance level. Meanwhile, Low Cluster periods can be used to identify periods in which inventory replenishment, product research, listing optimization, or targeted promotional activities may be required. These recommendations should be understood as decision-support implications rather than direct causal prescriptions because the K-Means model identifies similarity patterns but does not explain the causal mechanisms producing those patterns [20].

Despite these contributions, several limitations should be acknowledged. First, K-Means is sensitive to the selection of initial centroids because different starting positions can potentially produce different clustering solutions. Although the present study configured the algorithm with multiple runs and a maximum of 300 iterations to improve stability, the resulting clusters remain dependent on the structure of the observed dataset and the algorithm's initialization process. Therefore, the $K = 3$ solution should be interpreted as the most appropriate solution for the analyzed data based on the Elbow Method and cluster-quality evaluation, rather than as an inherently optimal classification for all thrift clothing sales datasets.

Second, the clustering variables were limited to Total Sales and Total Orders. Although these variables directly represent monetary and transaction-volume performance, they do not capture other factors that may explain daily sales variation, such as product category, price, product quality, inventory availability, promotional expenditure, advertising activity, customer characteristics, weekends, holidays, or Shopee campaign participation. Consequently, the clusters describe the observed sales outcomes but cannot establish why a particular day entered the Low, Medium, or High Cluster. Future research could incorporate

additional explanatory variables to develop a more comprehensive representation of daily sales behavior and potentially compare K-Means with other clustering or temporal analytical techniques.

Third, the dataset was obtained from a single thrift clothing store and covered six months of active operating days, resulting in 168 observations after inactive days were removed. This relatively focused dataset provides detailed insight into the operational pattern of the observed store, but it limits the generalizability of the findings. The cluster proportions, sales ranges, and temporal patterns may differ across stores because sellers can have different product assortments, prices, target markets, inventory capacities, promotional strategies, and store sizes. Therefore, the results should be considered empirical findings specific to the observed store and period rather than a generalized representation of thrift clothing sellers on Shopee.

Finally, the study is descriptive and exploratory in nature. K-Means identifies groups of observations with similar characteristics, but it does not establish causal relationships between sales performance and factors such as promotions, seasonal events, inventory, or consumer behavior. The observed concentration of High Cluster days in April May 2026, for example, provides an important pattern for further investigation but does not independently prove that the post-Eid period caused the increase in sales. Future studies involving multiple thrift stores, longer observation periods, additional business variables, and comparative clustering methods would therefore be valuable for testing the stability and generalizability of the patterns identified in this study.

Overall, the findings demonstrate that K-Means can provide a useful exploratory framework for identifying heterogeneous daily sales conditions in a thrift clothing business. The contribution of the study lies not only in producing Low, Medium, and High sales categories, but also in demonstrating that daily marketplace transactions can be transformed into empirically validated temporal performance segments that support operational interpretation. At the same time, the findings emphasize that clustering should be regarded as a pattern-discovery technique rather than a causal prediction model. This distinction is important for ensuring that the practical recommendations derived from the clusters remain consistent with the methodological capabilities and limitations of the K-Means approach.

4. Conclusion

Based on the findings and discussion, the following conclusions can be drawn:

1. The Knowledge Discovery in Databases (KDD) framework was successfully implemented through data selection, preprocessing, transformation, data mining, and evaluation/interpretation. The initial dataset obtained from Shopee Seller Center contained 182 records with the attributes Transaction Date, Total Sales, and Total Orders. After excluding 14 inactive store days, 168 active operating days remained for analysis. The data were normalized to the [0,1] range before applying the K-Means algorithm, while the optimal number of clusters ($K = 3$) was determined using the Elbow Method.
2. The K-Means algorithm successfully classified the 168 daily sales records into three clusters, namely Low, Medium, and High. The Low Cluster consisted of 101 days (60.1%), with an average daily sales value of IDR 134,139 and 2.38 orders. The Medium Cluster comprised 50 days (29.8%), with an average daily sales value of IDR 362,131 and 5.58 orders. The High Cluster consisted of 17 days (10.1%), with an average daily sales value of IDR 873,521 and 9.82 orders. Although it contained the fewest observations, the High Cluster contributed 31.9% of the total revenue, demonstrating that a relatively small number of high-performing days had a substantial contribution to overall sales performance.
3. The clustering model demonstrated good internal validation results, with a Davies Bouldin Index (DBI) of 0.6932 and a Silhouette Score of 0.512. These results indicate that the three clusters were reasonably compact and well separated within the analyzed dataset. Therefore, K-Means with $K = 3$

Clustering the Sales Levels of Thrift Clothing on the Shopee Platform Using the K-Means Algorithm. Putra Dwi Kurniawan et.al

provides a suitable exploratory approach for identifying daily thrift clothing sales patterns in the observed store. However, the results should be interpreted as being specific to the analyzed dataset and should not be considered universally applicable to all thrift clothing sellers.

4. The temporal analysis showed that High Cluster observations were concentrated mainly during April May 2026, particularly in the post-Eid period. May 2026 recorded the highest monthly revenue of IDR 13,298,090, whereas March 2026 recorded the lowest revenue of IDR 3,922,999. Across the observation period, the store generated total revenue of IDR 46,504,463 from 686 orders over 168 active operating days. These patterns indicate substantial variation in daily sales performance and provide an empirical basis for differentiating operational strategies according to sales conditions.
5. Despite these findings, this study has several methodological limitations. The dataset was obtained from only one thrift clothing store and covered a six-month observation period, resulting in 168 active operating-day observations. Therefore, the cluster composition, sales ranges, and temporal patterns may not represent other thrift clothing businesses with different product assortments, market characteristics, inventory conditions, pricing strategies, or promotional practices. In addition, the clustering process used only Total Sales and Total Orders as input variables. Other potentially relevant factors, such as product category, price, inventory availability, promotional activity, advertising expenditure, weekends, holidays, and marketplace campaign participation, were not incorporated. Furthermore, K-Means is sensitive to the initial centroid selection, meaning that different initializations may potentially produce different clustering configurations. Although multiple runs and cluster validation measures were employed, this limitation should still be considered when interpreting the stability of the resulting clusters. Finally, because K-Means is an unsupervised clustering method, the resulting clusters describe similarity patterns but do not establish causal relationships explaining why particular days achieve low, medium, or high sales.
6. Future research should address these limitations by expanding the dataset to include multiple thrift clothing stores, longer observation periods, and more diverse business characteristics to improve the robustness and generalizability of the findings. Subsequent studies may also incorporate additional variables related to product characteristics, inventory, pricing, promotions, advertising, temporal events, and customer behavior to obtain a more comprehensive representation of daily sales performance. In addition, alternative clustering algorithms, such as Hierarchical Clustering, DBSCAN, or other partitioning approaches, can be compared with K-Means to evaluate the stability of the identified patterns across different methodological assumptions. Ensemble or comparative clustering evaluations using multiple internal validation metrics may further strengthen the reliability of the cluster solution. These extensions would enable future research to determine whether the three-level sales structure identified in this study remains consistent across different stores, periods, datasets, and clustering techniques.

5. References

- [1] A. Juraidah, A. Hasan, A. Pengaruh Penerapan Etika Bisnis dan Tanggung Jawab Sosial Terhadap Kepercayaan Konsumen, D. Transaksi Jual Beli Online Di Platform Shopee, and A. Juraedah, "Analisis Pengaruh Penerapan Etika Bisnis Dan Tanggung Jawab Sosial terhadap Kepercayaan Konsumen : dalam Transaksi Jual Beli Online di Platform Shopee," *JUSTITIA: Journal of Justice, Law Studies, and Politic*, vol. 1, pp. 1–8, 2025.
- [2] F. Sulianta, K. Ulfah, and E. Amalia, "Revealing Consumer Preferences in the Fashion Industry Using K-Means Clustering," *International Journal of Engineering Continuity*, vol. 3, no. 2, pp. 34–53, Aug. 2024, doi: 10.58291/ijec.v3i2.280.

- [3] Arinda Aprilia and Prayitno Prayitno, "Studi Fenomenologis tentang Fear of Missing Out (FOMO) dalam Shopping Lifestyle pada Minat Beli Produk Fashion di Kalangan Generasi Z," *EKONOMIKA45: Jurnal Ilmiah Manajemen, Ekonomi Bisnis, Kewirausahaan*, vol. 13, no. 2, pp. 450–466, Apr. 2026, doi: 10.30640/ekonomika45.v13i2.6039.
- [4] H. Yunnisa, A. Pradesti, U. Islam, and S. Y. Tangerang, "Analisi Campaign Influencer Tyna Dwijayanti Dalam Membangun Brand Awareness Terhadap Produk Fashion Lokal Di Instagram (Studi Kasus Pada Oemah Etnik)," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, pp. 4756–4772, 2024.
- [5] S. Sella and Yohanni Syahra, "Analisis K-Means Clustering pada Data Penempatan Karyawan di PT. Guna Cipta Prima Berbasis Web," *Journal of Science, Technology, and Innovation*, vol. 1, no. 3, pp. 265–275, Apr. 2026, doi: 10.65310/59fhza60.
- [6] I. Suwandi, Z. Fadli, S. Wulandari, and N. Muin, "Pemanfaatan Aplikasi Sensus Penduduk Online Dalam Pencatatan Penduduk," *Jurnal Manajemen Informatika, Sistem Informasi dan Teknologi Komputer (JUMISTIK)*, vol. 2, no. 1, pp. 100–111, Jun. 2023, doi: 10.70247/jumistik.v2i1.28.
- [7] R. Ishak and A. Bengnga, "Clustering Tingkat Pemahaman Mahasiswa Pada Perkuliahan Probabilitas Statistika Dengan Metode K-Means," *Jambura Journal of Electrical and Electronics Engineering*, vol. 4, no. 1, pp. 65–69, Jan. 2022, doi: 10.37905/jjee.v4i1.11997.
- [8] S. W. Harjono, N. W. Utami, and I. G. A. P. D. Putri, "Klasterisasi Tingkat Penjualan pada Startup Panak.id dengan Algoritma K-Means," *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 17, no. 1, pp. 55–66, Jan. 2023, doi: 10.32815/jitika.v17i1.888.
- [9] Miftahul Arif Aldi and Zaehol Fatah, "IMPLEMENTASI K-MEANS CLUSTERING DALAM PENGELOMPOKAN DATA KUNJUNGAN WISATAWAN ASING DI INDONESIA," *Jurnal Ilmiah Multidisiplin Ilmu*, vol. 2, no. 1, pp. 13–19, Feb. 2025, doi: 10.69714/3hhfj353.
- [10] A. muliawan Nur, M. Saiful2, H. Bahtiar, and Muhammad Taufik Hidayat, "Penerapan Algoritma K-Means Clustering Dalam Mengelompokkan Smartphone Yang Rekomendasi Berdasarkan Spesifikasi," *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 2, pp. 478–488, Jul. 2024, doi: 10.29408/jit.v7i2.26283.
- [11] F. M. Sarimole and L. Hakim, "Klasifikasi Barang Menggunakan Metode Clustering K-Means Dalam Penentuan Prediksi Stok Barang," *Jurnal Sains dan Teknologi*, vol. 5, no. 3, pp. 846–854, Feb. 2024, doi: 10.55338/saintek.v5i3.2709.
- [12] M. A. N. Sabarna and R. Ramdani, "Collaborative governance in waste management: a case study in the municipality of Yogyakarta, Indonesia," *Publisia: Jurnal Ilmu Administrasi Publik*, vol. 10, no. 1, pp. 33–45, Apr. 2025, doi: 10.26905/pjiap.v10i1.14408.
- [13] A. Rajsya, P. Purnawansyah, and A. R. Manga', "Rancang Bangun Penerapan Metode Elbow Pada K-Means Untuk Clustering Data Persediaan Barang," *LINIER: Literatur Informatika dan Komputer*, vol. 1, no. 4, pp. 395–403, Dec. 2024, doi: 10.33096/linier.v1i4.2539.
- [14] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 14, no. 2, pp. 277–288, Nov. 2023, doi: 10.24176/simet.v14i2.9630.
- [15] I. Wahyudi, M. B. Sulthan, and L. Suhartini, "ANALISA PENENTUAN CLUSTER TERBAIK PADA METODE K-MEANS MENGGUNAKAN ELBOW TERHADAP SENTRA INDUSTRI PRODUKSI DI PAMEKASAN," *Jurnal Aplikasi Teknologi Informasi dan Manajemen (JATIM)*, vol. 2, no. 2, pp. 72–81, Dec. 2021, doi: 10.31102/jatim.v2i2.1274.
- [16] A. Winarta and W. J. Kurniawan, "OPTIMASI CLUSTER K-MEANS MENGGUNAKAN METODE ELBOW PADA DATA PENGGUNA NARKOBA DENGAN PEMROGRAMAN PYTHON," *JTIK (Jurnal Teknik Informatika Kaputama)*, vol. 5, no. 1, pp. 113–119, Jan. 2021, doi: 10.59697/jtik.v5i1.593.

- [17] C. A. Sri Fastaf and Y. Yamasari, "Analisa Pemetaan Kriminalitas Kabupaten Bangkalan Menggunakan Metode K-Means dan K-Means++," *Journal of Informatics and Computer Science (JINACS)*, vol. 3, no. 04, pp. 534–546, Jul. 2022, doi: 10.26740/jinacs.v3n04.p534-546.
- [18] A. Mira Yunita, R. Rizky, E. Nurafliyan Susanti, and S. Sumarno, "ANALISIS REKOMENDASI BARANG LAKU MENGGUNAKAN METODE WIGHTED PRODUCT PADA CV. PERMATA PUTERA SEJAHTERA," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 3, pp. 4602–4606, May 2025, doi: 10.36040/jati.v9i3.13387.
- [19] A. M. Artiarno, P. Setiaji, and F. Nugraha, "K-Means Clustering untuk Segmentasi Pelanggan: Mengungkap Pola Pembelian Strategi Pemasaran pada Sektor Ritel," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 2, pp. 442–451, Aug. 2025, doi: 10.29408/edumatic.v9i2.30336.
- [20] B. K. Padang and R. Kurniawan, "Clustering Data Penjualan Kopi Sidikalang Menggunakan Metode K-Means," *SENTRI: Jurnal Riset Ilmiah*, vol. 4, no. 8, pp. 1188–1197, Aug. 2025, doi: 10.55681/sentri.v4i8.4400.