JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

Product Segmentation for Apparel MSMEs Using K-Means and CRISP-DM Approach

Rahmat Hidayat [1)]
[1)] University of People
imjustrahmat2722@gmail.com

Submitted : 13 February 2026 | **Accepted** : 8 March 2026 | **Published** : 31 March 2026

Apparel Micro, Small, and Medium Enterprises (MSMEs) generate substantial sales data that remains largely underutilised for strategic decision-making. This study applies K-Means Clustering within the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to analyse 360 sales records (2021–2023) across three key variables: quantity sold, unit price, and total turnover. Using the Silhouette Score ($k = 2$, score = 0.42) to determine the optimal cluster count, the analysis identified two distinct product segments: Cluster 0 (high-performing products with elevated sales volume, premium pricing, and substantial turnover) and Cluster 1 (standard products with lower sales volume, reduced pricing, and modest turnover). The principal contribution of this study is a structured, replicable CRISP-DM-based product segmentation methodology tailored for resource-constrained apparel MSMEs, offering a practical data-driven alternative to intuition-based decision-making. The findings provide actionable guidance for inventory optimisation and differentiated marketing strategies. Future work may extend the framework by incorporating external variables such as seasonal trends or customer demographics to refine segmentation precision.

Keywords: K-Means Clustering, CRISP-DM, Sales Analysis, Apparel MSME, Data Mining, Silhouette Score.

INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) constitute a vital component of the Indonesian economy, contributing significantly to employment generation, poverty alleviation, and economic growth (Indonesia, 2008; Ramadani et al., 2025). The apparel sector, in particular, represents a substantial portion of MSMEs in Indonesia, characterized by diverse product offerings, fluctuating market demands, and intense competition. Despite their economic importance, many apparel MSMEs operate with limited analytical capabilities, relying predominantly on intuition and experience rather than data-driven insights for strategic decision-making (Abdul-Azeez et al., 2024). This reliance on traditional approaches often results in suboptimal inventory management, inefficient marketing strategies, and missed opportunities for business growth.

In the contemporary business landscape, data has emerged as a critical asset for organizational success. Apparel MSMEs frequently generate substantial volumes of sales data encompassing information about products sold, quantities, pricing, and revenue. However, this data often remains underutilized, stored in disparate systems without systematic analysis or strategic application (Dewata et al., 2020). The inability to extract meaningful insights from sales data represents a significant competitive disadvantage, particularly in an era where data-driven decision-making has become imperative for sustainable business performance (Abdul-Azeez et al., 2024). This gap between data availability and data utilization presents both a challenge and an opportunity for MSMEs seeking to enhance their operational efficiency and market competitiveness.

Data mining techniques offer promising solutions for transforming raw sales data into actionable business intelligence. Among various data mining methodologies, clustering algorithms have demonstrated particular effectiveness in identifying patterns and grouping similar entities based on their characteristics (Wei et al., 2024). Clustering enables businesses to segment their product portfolios, customer bases, or operational processes into distinct groups that exhibit homogeneous properties within clusters while maintaining heterogeneity between clusters. This segmentation capability facilitates more nuanced understanding of business dynamics and supports targeted strategic interventions (Wani et al., 2023).

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

K-Means clustering has emerged as one of the most widely adopted unsupervised machine learning algorithms for customer and product segmentation across various industries (Ma, 2024; Salsabila et al., 2025). The algorithm's popularity stems from its computational efficiency, scalability, and interpretability, making it particularly suitable for MSME contexts where technical resources may be limited. K-Means operates by partitioning a dataset into K distinct clusters, where each observation belongs to the cluster with the nearest mean value, serving as the cluster's centroid (Harmain et al., 2022). The algorithm iteratively refines cluster assignments until convergence, resulting in well-defined groups that minimize within-cluster variance while maximizing between-cluster separation.

Numerous studies have demonstrated the effectiveness of K-Means clustering in retail and e-commerce contexts. Huang et al. (2020) successfully applied K-Means combined with the RFM (Recency, Frequency, Monetary) model for customer segmentation, enabling businesses to identify high-value customer segments and tailor marketing strategies accordingly. Similarly, Siagian et al. (2021) employed K-Means with an extended LRFM model (Length, Recency, Frequency, Monetary) for e-commerce customer segmentation, demonstrating improved targeting capabilities. In the Indonesian context, Jordy et al. (2023) utilized K-Means with RFM analysis for customer value segmentation at AVANA Indonesia, yielding actionable insights for customer relationship management. Furthermore, Nugroho et al. (2024) applied K-Means clustering to sales transaction data, successfully identifying distinct customer segments that informed targeted marketing initiatives.

While customer segmentation has received considerable attention in the literature, product segmentation using clustering techniques represents an equally valuable yet less explored application domain. Haris Munandar (2024) demonstrated the application of K-Means for selecting superior products in retail environments, highlighting the algorithm's capability to differentiate product performance based on sales metrics. Product segmentation enables businesses to identify star performers, underperforming items, and products requiring strategic intervention, thereby facilitating more efficient inventory management and resource allocation (Barrera et al., 2023). This approach proves particularly relevant for apparel MSMEs that typically manage diverse product portfolios with varying performance characteristics.

The successful implementation of data mining initiatives requires not only appropriate algorithmic selection but also systematic methodological frameworks that ensure rigor and reproducibility. The Cross-Industry Standard Process for Data Mining (CRISP-DM) has emerged as the de facto standard methodology for data mining projects, providing a structured approach that encompasses business understanding, data understanding, data preparation, modeling, evaluation, and deployment phases (Nursahid et al., 2025). This comprehensive framework ensures that analytical projects remain aligned with business objectives while maintaining technical rigor throughout the process. The CRISP-DM methodology has been successfully applied across various domains, demonstrating its versatility and effectiveness in translating business problems into data mining solutions.

A critical component of any data mining initiative involves data preprocessing and preparation. The quality of clustering results depends fundamentally on the quality and appropriateness of input data (Nursahid et al., 2025). Data normalization, in particular, plays a crucial role in K-Means clustering, as the algorithm's distance-based mechanism can be unduly influenced by variables with larger scales (Harmain et al., 2022). Several normalization techniques exist, including min-max normalization, z-score standardization, and decimal scaling, each with distinct characteristics and appropriate use cases. The selection of normalization method can significantly impact clustering outcomes, necessitating careful consideration based on data characteristics and analytical objectives.

One of the primary challenges in K-Means clustering involves determining the optimal number of clusters. Unlike supervised learning methods where target variables guide model training, clustering algorithms require practitioners to specify the number of clusters a priori or through systematic evaluation (Nasution & Hasibuan, 2020). Several methods have been proposed for optimal cluster determination, including the Elbow method, which identifies the point where adding additional clusters yields diminishing returns in variance reduction, and the Silhouette Score, which measures cluster cohesion and separation (Hidayati et al., 2021). The Silhouette Score, ranging from -1 to 1, provides a comprehensive assessment of clustering quality, with higher values indicating better-defined clusters. This metric has gained widespread adoption due to its intuitive interpretation and computational efficiency.

While K-Means clustering offers numerous advantages, alternative clustering algorithms warrant consideration for specific use cases. K-Medoids clustering, for instance, utilizes actual data points as cluster centers rather than computed means, providing greater robustness to outliers (Afari, 2023; He et al., 2022). Several studies have conducted comparative analyses between K-Means and K-Medoids algorithms, yielding mixed results depending on dataset characteristics and evaluation criteria (Fahrudin & Rindiyani, 2024; Mirantika & Rijanto, n.d.; Syamfithriani et al., 2023). Agustin et al. (2025) demonstrated that K-Medoids can achieve superior performance in retail customer segmentation when outliers are prevalent. Similarly, Henderi et al. (2024) optimized the Davies-Bouldin index through K-Medoids application, highlighting the algorithm's potential for improved cluster compactness and separation. Nevertheless, K-Means typically exhibits superior computational

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

efficiency, particularly for large datasets, making it the preferred choice for many business applications (Mirantika & Rijanto, 2024).

The integration of clustering algorithms with established business frameworks enhances their practical utility. The RFM model, which segments customers based on Recency of purchase, Frequency of transactions, and Monetary value, has been extensively combined with clustering techniques to yield actionable customer insights (Aslantaş et al., 2023; Serwah et al., 2023). Farisi and Supatmi (2024) successfully implemented K-Means clustering with RFM attributes to enhance donor retention at Masjid Nusantara, demonstrating the approach's applicability beyond commercial contexts. Maniyara et al. (2024) utilized RFM-based ranking techniques for effective customer segmentation, while Yunita et al. (2025) employed K-Means clustering with RFM metrics to develop targeted marketing strategies. These studies collectively demonstrate the synergistic potential of combining analytical techniques with domain-specific frameworks to generate business value.

Beyond customer and product segmentation, clustering techniques have found applications in diverse business contexts. Janardhanan and Muthalagu (2020) applied machine learning algorithms, including clustering methods, for market segmentation aimed at profit maximization. Niu (2021) developed an intelligent evaluation model combining K-Means and Self-Organizing Maps (SOM) algorithms for e-commerce transaction volume assessment, demonstrating the potential for algorithmic integration. Hung and Dat (2020) employed dynamic time warping distance metrics for customer behavior clustering based on balance history, illustrating advanced distance measure applications. These innovative applications underscore the versatility of clustering methodologies and their potential for addressing complex business challenges across various domains.

The transformation of raw data into business intelligence represents a critical capability for modern organizations. Hadad and Keren (2022) developed a decision-making support system module incorporating customer segmentation and ranking capabilities, exemplifying how clustering outputs can be integrated into broader decision support frameworks. This integration enables real-time or near-real-time analytical capabilities that support agile decision-making processes. Yu (2022) implemented a weighted naive Bayes algorithm for real-time sales forecasting in smart city e-commerce contexts, highlighting the convergence of clustering, classification, and predictive analytics techniques. These developments suggest a trajectory toward increasingly sophisticated analytical ecosystems that leverage multiple algorithmic approaches synergistically.

Despite the demonstrated potential of data mining techniques, their implementation in MSME contexts presents unique challenges. Resource constraints, including limited technical expertise, computational infrastructure, and financial capital, can impede adoption of sophisticated analytical approaches. Furthermore, data quality issues, including incomplete records, inconsistent formats, and measurement errors, frequently plague MSME data systems, potentially compromising analytical outcomes (Dewata et al., 2020). The absence of standardized data collection and management practices exacerbates these challenges, necessitating substantial data cleaning and preprocessing efforts prior to analytical application. Additionally, the interpretation and operationalization of analytical insights require business acumen and domain knowledge that may not always be readily available in MSME settings.

Despite the extensive literature on clustering techniques in retail and e-commerce, three specific gaps remain unaddressed. First, the majority of studies focus on customer segmentation (e.g., using RFM models) rather than product performance categorisation; studies that do examine product-level clustering predominantly target large retail or e-commerce platforms with abundant data and technical resources. Second, the apparel MSME sector in Indonesia represents a structurally distinct operational context characterised by limited data infrastructure, small transaction volumes, and constrained analytical capacity—conditions that have not been specifically examined in prior product segmentation research. Third, while CRISP-DM has been validated in various domains, its explicit application as a structuring framework for data mining initiatives in Indonesian MSME settings remains scarce, with Nursahid et al. (2025) being one of the few examples and focused on healthcare classification rather than business segmentation. This study directly addresses all three gaps by applying K-Means clustering within the full CRISP-DM pipeline to product sales data from an Indonesian apparel MSME. In doing so, it offers two distinct contributions not found in combination in prior work: (1) a methodological contribution demonstrating the feasibility and effectiveness of CRISP-DM as a systematic guide for data mining in resource-constrained MSME environments, and (2) a substantive contribution providing empirical evidence of identifiable product performance tiers in the apparel sector that can directly inform inventory and marketing decisions without requiring specialist analytical infrastructure.

This study aims to address the identified research gap by examining sales data from an apparel MSME using K-Means clustering within the CRISP-DM framework. The specific research objectives include: (1) to systematically analyze apparel sales data following CRISP-DM methodology, encompassing business understanding, data understanding, data preparation, modeling, evaluation, and deployment phases; (2) to determine the optimal number of product clusters based on sales performance metrics using the Silhouette Score

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

evaluation criterion; (3) to identify and characterize distinct product performance segments based on sales volume, unit price, and total turnover; and (4) to translate analytical findings into actionable recommendations for marketing strategy development and inventory management optimization. By achieving these objectives, this research seeks to demonstrate how data mining techniques can be effectively applied in MSME contexts to generate practical business value.

The anticipated contributions of this research encompass both theoretical and practical dimensions. From a theoretical perspective, this study extends the application of the CRISP-DM framework to Indonesian MSME contexts, providing insights into methodological adaptation for resource-constrained environments. Additionally, the research contributes to the growing body of literature on product segmentation using clustering techniques, complementing the predominant focus on customer segmentation in existing studies. From a practical standpoint, this research provides apparel MSME stakeholders with a systematic approach to sales data analysis that can inform strategic decision-making regarding product portfolio management, pricing strategies, and inventory optimization. The demonstration of accessible analytical techniques suitable for MSME implementation may encourage broader adoption of data-driven decision-making practices in this critical economic sector.

The remainder of this paper is structured as follows: The Literature Review section provides a comprehensive examination of relevant theoretical foundations and empirical studies, encompassing CRISP-DM methodology, K-Means clustering algorithms, evaluation metrics, and applications in retail and e-commerce contexts. The Method section details the research design, data collection procedures, preprocessing techniques, clustering implementation, and evaluation approaches employed in this study. The Result section presents the findings from each CRISP-DM phase, including data exploration outcomes, optimal cluster determination, and cluster characterization. The Discussion section interprets these findings in relation to existing literature, explores their business implications, acknowledges research limitations, and proposes directions for future research. Finally, the Conclusion section synthesizes the key findings and their implications for apparel MSME management and data mining practice.

LITERATURE REVIEW

The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a comprehensive framework for structuring data mining projects, encompassing six sequential phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This methodology ensures systematic alignment between analytical activities and organizational objectives while maintaining technical rigor throughout the process (Nursahid et al., 2025). The framework's versatility has been demonstrated across diverse domains, from healthcare applications in heart disease classification (Nursahid et al., 2025) to retail analytics and customer segmentation initiatives. The business understanding phase establishes project objectives and success criteria, while data understanding involves exploratory analysis to assess data quality and identify patterns. Data preparation, often the most time-intensive phase, encompasses data cleaning, transformation, and feature engineering activities essential for effective modeling. The modeling phase applies selected algorithms, followed by rigorous evaluation against predetermined criteria, and culminates in deployment where insights are operationalized into business processes (Hidayat, 2025).

K-Means clustering represents a foundational unsupervised learning algorithm that partitions data into K distinct groups by minimizing within-cluster variance. The algorithm operates iteratively, initially selecting K random centroids, assigning each data point to its nearest centroid based on Euclidean distance, recalculating centroids as the mean of assigned points, and repeating until convergence or maximum iterations are reached (Harmain et al., 2022; Salsabila et al., 2025). The computational efficiency and scalability of K-Means have contributed to its widespread adoption in business analytics, particularly for customer and product segmentation applications. However, the algorithm exhibits sensitivity to initial centroid selection, potentially converging to local optima rather than global solutions (Nasution & Hasibuan, 2020). Various initialization strategies, including K-Means++ and multiple random initializations with selection of the best outcome, have been proposed to mitigate this limitation. Data normalization emerges as a critical preprocessing step, as K-Means' distance-based mechanism can be disproportionately influenced by variables with larger scales, necessitating standardization techniques such as min-max normalization or z-score standardization to ensure equitable feature contribution (Harmain et al., 2022).

Determining the optimal number of clusters constitutes a fundamental challenge in unsupervised clustering applications. The Silhouette Score has emerged as a prominent evaluation metric, quantifying both cluster cohesion (how similar objects are within the same cluster) and separation (how distinct clusters are from each other), with values ranging from -1 to +1 where higher scores indicate better-defined clusters (Hidayati et al., 2021). Alternative metrics include the Davies-Bouldin Index, which measures the average similarity ratio of each cluster with its most similar cluster, with lower values indicating better clustering (Henderi et al., 2024), and the Elbow

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

method, which plots within-cluster sum of squares against the number of clusters to identify the point of diminishing returns. Comparative analyses have demonstrated that different validation metrics may yield divergent optimal cluster numbers for identical datasets, necessitating domain knowledge and business context to guide final selection. The Silhouette Score's intuitive interpretation and consideration of both cohesion and separation have contributed to its preference in practical applications, particularly in business contexts where stakeholder communication of analytical results is essential (Hidayati et al., 2021).

Empirical applications of K-Means clustering in retail and e-commerce contexts have demonstrated substantial business value across customer segmentation, product categorization, and market analysis domains. Huang et al. (2020) successfully integrated K-Means with the RFM model for customer segmentation, enabling targeted marketing strategies based on customer value tiers. Similarly, Siagian et al. (2021) employed an extended LRFM framework combined with K-Means for e-commerce customer classification, yielding actionable insights for customer relationship management. In the Indonesian MSME context, Jordy et al. (2023) applied K-Means clustering with RFM analysis at AVANA Indonesia, successfully identifying distinct customer value segments that informed strategic marketing initiatives. Product-focused applications include Haris Munandar's (2024) superior product selection methodology using K-Means and K-Medoids algorithms, which enabled systematic identification of high-performing products based on sales metrics. Ma (2024) demonstrated K-Means' effectiveness in optimizing marketing strategies for tobacco companies through data-driven customer insights, while Niu (2021) developed an intelligent e-commerce transaction volume evaluation model combining K-Means with Self-Organizing Maps. These diverse applications underscore clustering algorithms' versatility in generating actionable business intelligence from transactional data (Hidayat, 2026).

Comparative studies examining K-Means against alternative clustering algorithms, particularly K-Medoids, have yielded nuanced insights regarding algorithmic selection criteria. K-Medoids utilizes actual data points as cluster centers rather than computed means, providing enhanced robustness to outliers and noise (Afari, 2023; He et al., 2022). Fahrudin and Rindiyani (2024) conducted systematic comparisons of K-Means and K-Medoids for RFM-based customer segmentation, finding that optimal algorithm selection depends on dataset characteristics and business priorities. Agustin et al. (2025) demonstrated K-Medoids' superior performance in retail customer segmentation when outliers are prevalent, while Mirantika and Rijanto (2024) found K-Means exhibited better computational efficiency for large datasets. The integration of clustering algorithms with the RFM model has proven particularly effective, with studies by Aslantaş et al. (2023), Serwah et al. (2023), and Maniyara et al. (2024) demonstrating enhanced customer segmentation accuracy through combined approaches. Yunita et al. (2025) successfully employed K-Means with RFM metrics for developing targeted marketing strategies, while Farisi and Supatmi (2024) extended the application to non-commercial contexts, improving donor retention through data-driven segmentation. These findings collectively indicate that effective clustering implementation requires careful consideration of algorithm characteristics, data properties, and business objectives to achieve optimal outcomes.

METHOD

This research utilizes sales data from a convection MSME (Micro, Small, and Medium Enterprise) spanning the period from 2021 to 2025. The dataset comprises 360 transaction records featuring six attributes: Year, Month, Product, Quantity Sold, Unit Price, and Turnover. The analysis focuses on three core variables—Quantity Sold, Unit Price, and Turnover—as they are considered the most representative indicators of sales performance. Quantity Sold reflects consumer demand, Unit Price relates to pricing strategies affecting competitiveness, and Turnover represents the resulting financial contribution. By examining these three variables, product performance can be analyzed in terms of sales volume, economic value, and pricing strategy. The dataset specifications are summarized in Table 1.

Table 1
Research Dataset Specifications

Attribute	Data Type	Description
Year	Integer	Sales year (2021–2025)
Month	Integer	Sales month (1–12)
Product	String	Name of the convection product
Quantity Sold	Integer	Number of products sold
Unit Price	Integer	Price per unit (Rp)
Turnover	Integer	Total revenue (Rp)

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

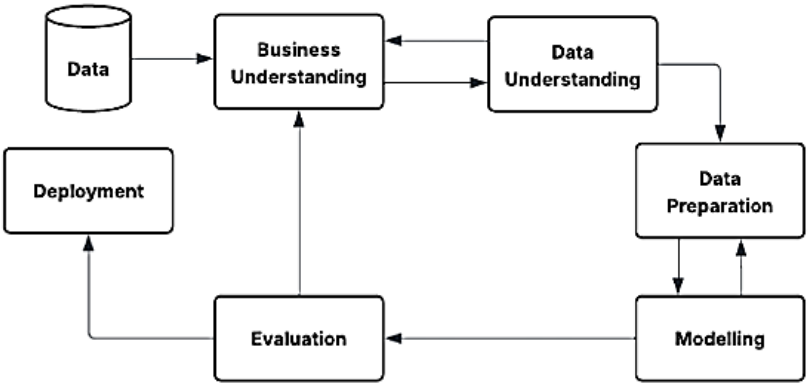

Figure 1. Research Stages with the CRISP-DM Approach

In the business understanding stage, the main research objective was determined: to group MSME convection products based on their sales performance. The grouping results are expected to be used by MSMEs to develop more targeted marketing strategies and support more efficient inventory management. Therefore, the decision-making process is no longer based solely on intuition, but also on objective data.

The data understanding stage involves an initial analysis of the dataset to obtain a comprehensive overview of the data characteristics. The exploration process is carried out by calculating descriptive measures such as the average, minimum, and maximum values for each variable. Furthermore, the data distribution is examined through simple visualizations to detect trends and potential anomalies. The results of this exploration are crucial because they will form the basis for further modeling.

The data preparation stage involves cleaning the dataset to address missing values and potential inconsistencies. Then, relevant core variables are selected: Quantity Sold, Unit Price, and Turnover. To ensure these variables can be analyzed in a balanced manner, normalization was performed using the StandardScaler. This normalization process is necessary because turnover has a much larger scale than the quantity sold or unit price, and without adjustment, it can dominate the clustering results [13]. The next stage is modeling, which is performed using the K-Means Clustering algorithm. This algorithm was chosen because it works efficiently in handling relatively large datasets and produces easily interpretable clustering [14].

The modeling process was carried out by testing several variations in the number of clusters, ranging from $k = 2$ to $k = 10$, to find the clustering structure that best fits the data. The random_state parameter was set to ensure consistent replication of the results, while n_init was used to minimize the influence of the initial centroid selection. Mathematically, the K-Means algorithm aims to minimize the distance between the data and the cluster center [15], as shown in Equation 1.

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

The evaluation phase is conducted to assess the quality of the clustering results. This evaluation utilizes the Silhouette Score metric, which measures the proximity of data points to their assigned cluster relative to other clusters. Silhouette values range from -1 to 1; a value approaching 1 indicates superior cluster separation, while a value near 0 suggests overlapping clusters. The final stage is deployment, where the analytical results are presented through summary tables and various visualizations. These visualizations include two-dimensional scatterplots to illustrate data distribution across clusters, three-dimensional plots to depict inter-variable interactions, and heatmaps to display the mean value differences for each cluster. All visualizations are provided in the Results and Discussion section to support the interpretation of the identified patterns.

The methodology applied in this study is designed to yield systematic analytical results, yet it possesses certain constraints. The primary limitations include a relatively small number of variables and the reliance on a single evaluation metric. Consequently, the obtained results should be interpreted with caution, taking these limitations into account.

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

RESULT

The research dataset consists of 360 sales records from a convection MSME collected over the 2021–2023 period, focusing on three primary variables: Quantity Sold, Unit Price, and Turnover. The descriptive statistics for these variables are presented in Table 2.

Table 2
Descriptive Statistics of Research Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
Quantity Sold	22	994	514.36	280.08
Unit Price	50,106	249,551	150,979.02	59,447.19
Turnover	1,852,470	240,096,100	78,812,390	55,741,820

Based on Table 2, it is evident that product turnover varies significantly, as indicated by the high standard deviation value. This suggests the presence of top-performing products that contribute much more substantially than others. To determine the most appropriate number of clusters, the Silhouette Score was calculated for a range of $k = 2$ to $k = 10$. The test results are illustrated in Figure 2.

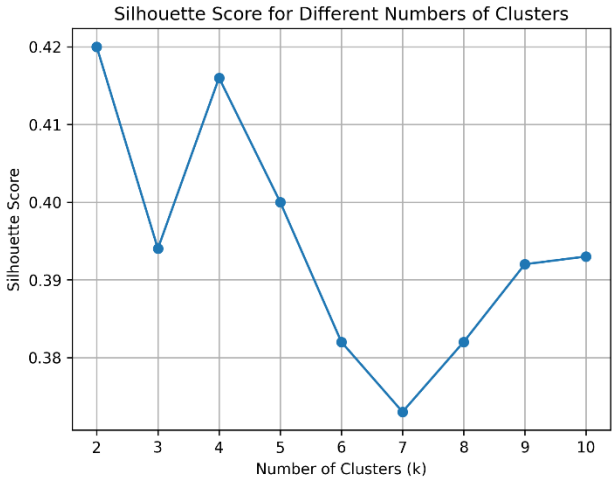

Figure 2. Silhouette Score for various numbers of clusters

As shown in Figure 2, the highest value was obtained at $k = 2$ with a score of 0.4202. Consequently, the number of clusters used in this analysis is two. The details for each cluster are presented in Table 3 below :

Table 3
Silhouette Score for each Cluster

Cluster (k)	Silhouette Score
2	0.4202
3	0.3939
4	0.4164
5	0.4001
6	0.3824
7	0.3729
8	0.3821
9	0.3921
10	0.3928

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

After the clustering process, a summary of the characteristics of each cluster was obtained as shown in Table 4.

Table 4
Average Variables in Each Cluster

Cluster	Total Sold (Average)	Unit Price (Average)	Turnover (Average)
0	742.07	178,190.04	128,102,800
1	315.11	127,169.38	35,683,260

According to Table 4, Cluster 0 represents high-performing data with higher sales, premium pricing, and substantial turnover, whereas Cluster 1 shows lower values across these metrics. The clustering outcomes are further detailed through various visualizations: Figure 3 (scatterplot) displays data distribution, Figure 4 shows 3D variable relationships, and Figure 5 (heatmap) compares the mean of each variable between clusters.

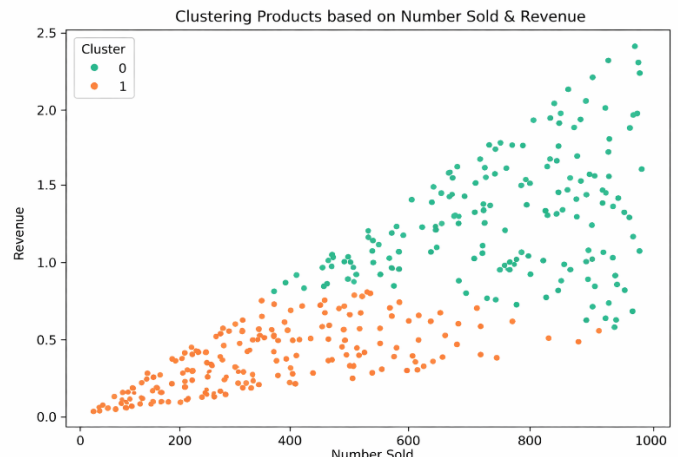

Figure 3. Scatterplot of clustering results based on number of items sold and turnover

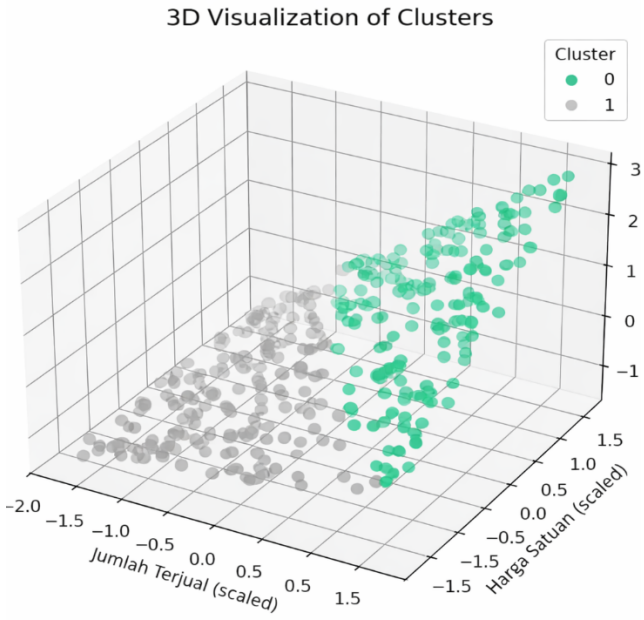

Figure 4. Three-dimensional visualization of clustering results

JCEIT: Journal of Computer Engineering and Information Technology
Volume 2, Number 2, March 2026
Page: 93-104
ISSN 3089-106X
(online)

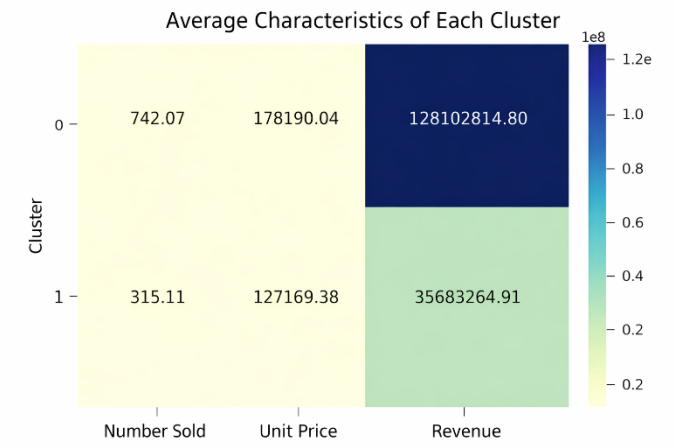

Figure 5. Heatmap of the average characteristics of each cluster

DISCUSSIONS

The clustering results indicate that convective MSME goods are categorised into two groups with unique characteristics. Cluster 0 consists of high-revenue products characterised by comparatively elevated prices and substantially greater turnover. These can be classified as premium products; although they do not prevail in quantity, their impact on corporate revenue is significant. In contrast, Cluster 1 comprises products characterised by low sales volume, reduced pricing, and modest turnover, designated as regular products that sustain sales diversity and market accessibility. This interpretation corresponds with Kotler's market segmentation theory, wherein categorising products into discrete categories facilitates more efficient marketing techniques. Premium items can be strategically positioned to elevate brand image and cultivate consumer loyalty via value-added services. Simultaneously, conventional items may be subjected to market penetration techniques, including bundling or discount campaigns, to attain a wider client demographic. Moreover, these findings align with the Pareto Principle (80/20 rule), indicating that a limited fraction of items usually accounts for the majority of revenue. In this study, Cluster 0 serves as the income foundation, whereas Cluster 1 sustains transaction volume. Consequently, both organisations provide complementary functions in promoting company sustainability. In comparison to prior research on retail MSMEs, these findings reveal notable disparities between high and low-turnover items. This research provides a competitive advantage by employing the CRISP-DM paradigm, facilitating a methodical analysis from business comprehension to implementation. Notwithstanding its advantages, the Silhouette Score of 0.42 signifies that the quality of the clusters remains suboptimal. Future investigations may include external variables, like seasonal trends or distribution channels, and evaluate outcomes utilising various techniques such as DBSCAN or Hierarchical Clustering.

CONCLUSION

This study demonstrates that K-Means Clustering, structured within the CRISP-DM framework, can effectively convert raw sales data into actionable product intelligence for apparel MSMEs. With an optimal Silhouette Score of 0.42 at $k = 2$, the analysis produced two well-differentiated clusters: high-performing products with elevated sales volume, premium pricing, and substantial turnover (Cluster 0), and standard products with lower values across all three dimensions (Cluster 1). The principal contribution of this work is methodological: it establishes a structured, accessible, and replicable data mining pipeline suitable for resource-constrained MSME environments—one that does not require specialist infrastructure and can be directly operationalised for product portfolio management and inventory optimisation. From a practical standpoint, the two-cluster segmentation provides a clear basis for differentiated business strategies: premium products warrant investment in quality maintenance and brand reinforcement, while standard products are suited to volume-driven approaches such as bundling and competitive pricing. Future research should extend the framework by incorporating additional variables—such as seasonal patterns, regional demand, or promotional activity—and explore alternative algorithms such as DBSCAN or hierarchical clustering to assess whether richer segmentation structures can be identified in similar MSME datasets.

REFERENCES

Abdul-Azeez, O., Ihechere, A. O., & Idemudia, C. (2024). Enhancing business performance: The role of data-driven analytics in strategic decision-making. International Journal of Management and Entrepreneurship Research, 6(7), 2066–2081. https://doi.org/10.51594/ijmer.v6i7.1257

Afari, I. S. (2023). K-medoids customer segmentation algorithm by utilizing customer relationship management. Journal of Computer Science and Information Technology. https://doi.org/10.35134/jcsitech.v9i2.69

Agustin, E. W., Uthami, K., Ulfa, A. I., Efrizoni, L., & Rahmaddeni. (2025). Optimization of customer segmentation in the retail industry using the K-medoid algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science. https://doi.org/10.57152/malcom.v5i3.1977

Aslantaş, G., Gençgül, M., Rumelli, M., Özsaraç, M., & Bakirli, G. (2023). Customer segmentation using K-means clustering algorithm and RFM model. Deu Muhendislik Fakultesi Fen ve Muhendislik. https://doi.org/10.21205/deufmd.2023257418

Azzaria, C., Daniati, E., & Ristyawan, A. (2025). Peningkatan akurasi deteksi liver disease melalui hyperparameter tuning pada algoritma random forest. Indonesian Journal of Computer Science Research, 4(2), 139–147.

Barrera, F., Segura, M., & Maroto, C. (2023). Multicriteria sorting method based on global and local search for supplier segmentation. International Transactions in Operational Research, 31, 3108–3134. https://doi.org/10.1111/itor.13288

Daniati, E., & Utama, H. (2019). Clustering K-Means for criteria weighting with improvement result of alternative decisions using SAW and TOPSIS. In 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 73–78). IEEE. https://doi.org/10.1109/ICITISEE48480.2019.9003858

Dewata, E., Sari, Y., & Jauhari, H. (2020). Penyusunan laporan keuangan terkomputerisasi berdasarkan SAK EMKM pada UMKM konveksi. Intervensi Komunitas, 2(1), 11–16. https://doi.org/10.32546/ik.v2i1.676

Fahrudin, N. F., & Rindiyani, R. (2024). Comparison of K-medoids and K-means algorithms in segmenting customers based on RFM criteria. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202448402008

Farisi, I. M., & Supatmi, S. (2024). Implementation of the K-means clustering technique using RFM attributes to enhance donor retention at Masjid Nusantara. In 2024 International Conference on Informatics Engineering, Science & Technology (INCITEST) (pp. 1–6). https://doi.org/10.1109/INCITEST64888.2024.11121472

Hadad, Y., & Keren, B. (2022). A decision-making support system module for customer segmentation and ranking. Expert Systems, 40. https://doi.org/10.1111/exsy.13169

Haris Munandar, M. (2024). Application of data mining in selecting superior products using the K-means and K-medoids algorithm methods. JURTEKSI (Jurnal Teknologi dan Sistem Informasi). https://doi.org/10.33330/jurteksi.v10i4.3196

Harmain, A., Paiman, P., Kurniawan, H., Kusrini, K., & Maulina, D. (2022). Normalisasi data untuk efisiensi K-Means pada pengelompokan wilayah berpotensi kebakaran hutan dan lahan berdasarkan sebaran titik panas. Teknik Teknologi Informasi dan Multimedia, 2(2), 83–89. https://doi.org/10.46764/teknimedia.v2i2.49

He, Y., Xu, Z., & Liu, N. (2022). Research on K-medoids algorithm with probabilistic-based expressions and its applications. Applied Intelligence, 52, 12016–12033. https://doi.org/10.1007/s10489-021-02937-8

Henderi, H., Fitriana, L., Iskandar, I., Astuti, R., Arifandy, M. I., Hayadi, B., Mesran, M., Chin, J., & Kurniawan, A. (2024). Optimization of Davies-Bouldin index with K-medoids algorithm. AIP Conference Proceedings. https://doi.org/10.1063/5.0225220

Hidayati, R., Zubair, A., Pratama, A. H., & Indana, L. (2021). Analisis silhouette coefficient pada 6 perhitungan jarak K-Means clustering. Techno.Com, 20(2), 186–197. https://doi.org/10.33633/tc.v20i2.4556

Huang, Y., Zhang, M., & He, Y. (2020). Research on improved RFM customer segmentation model based on K-means algorithm. In 2020 5th International Conference on Computational Intelligence and Applications (ICCIA) (pp. 24–27). https://doi.org/10.1109/ICCIA49625.2020.00012

Hung, P. D., & Dat, D. Q. (2020). Customer behavior clustering based on balance history using dynamic time warping distance. International Journal of Machine Learning and Computing. https://doi.org/10.18178/ijmlc.2020.10.1.903

Indonesia, R. (2008). Undang-Undang Republik Indonesia Nomor 20 Tahun 2008 tentang Usaha Mikro, Kecil, dan Menengah. https://peraturan.bpk.go.id/Details/39653/uu-no-20-tahun-2008

Janardhanan, S., & Muthalagu, R. (2020). Market segmentation for profit maximization using machine learning algorithms. Journal of Physics: Conference Series, 1706. https://doi.org/10.1088/1742-6596/1706/1/012160

Jordy, M., Triayudi, A., & Sholihati, I. D. (2023). Analisis segmentasi recency dan customer value pada AVANA Indonesia dengan algoritma K-means dan model RFM. Journal of Information System Research (JOSH). https://doi.org/10.47065/josh.v4i2.2950

Kurnia, O. D., et al. (2024). Analisis perbandingan algoritma Naïve Bayes dengan K-Nearest Neighbor (KNN) pada dataset mobile price classification. Prosiding Seminar Nasional Inovasi Teknologi (SEMNAS INOTEK), 8, 2549–7952.

Ma, M. (2024). The application of K-means algorithm-based data mining in optimizing marketing strategies of tobacco companies. International Journal of Advanced Computer Science and Applications. https://doi.org/10.14569/ijacsa.2024.0151186

Maniyara, K., Shah, K., Gaidhane, V. H., & Wanjari, R. (2024). An effective customer segmentation using RFM ranking techniques. In 2024 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom) (pp. 592–597). https://doi.org/10.1109/MoSICom63082.2024.10882067

Maulana, A., Ristyawan, A., Ndun, A. R., & Ristyawan, A. (2025). Prediksi volume sampah perkotaan berbasis data spasial menggunakan random forest di DKI Jakarta. Prosiding SEMNAS INOTEK, 9, 1667–1672.

Hidayat, R., & Aminulhaq, F. (2025). A Comparative Analysis of Decision Tree, Logistic Regression, and Support Vector Machine Algorithms in Sentiment Analysis of Threads App Reviews. *Intechno Journal: Information Technology Journal*, *7*(2), 45-55.

Hidayat, R., & Ratnaningsih, D. J. (2025). Analisis Sentimen Program Mbg Menggunakan Algoritma Random Forest Dan Naive Bayes. *Journal of Computing and Informatics Research*, *5*(1), 395-400.

Hidayat, R. (2026). The Effect of Ability and Career Development on Employee Performance: Evidence from the Indonesian Automotive Industry. *Jurnal Prima Manajemen*, *1*(3), 535-547.

Mirantika, N., & Rijanto, E. (2024). Implementasi metode clustering partisi dalam menentukan segmentasi pelanggan. Jurnal Tata Kelola dan Kerangka Kerja Teknologi Informasi. https://doi.org/10.34010/jtk3ti.v10i1.11320

Mirantika, N., & Rijanto, E. (n.d.). Comparative analysis of K-means and K-medoids algorithms in determining customer segmentation using RFM model. Retrieved from https://www.semanticscholar.org/paper/2053b7b55a15bd710794443298214c8305d29458

Nasution, M. Z., & Hasibuan, M. S. (2020). Pendekatan initial centroid search untuk meningkatkan efisiensi iterasi klustering K-Means. Techno.Com, 19(4), 341–352. https://doi.org/10.33633/tc.v19i4.3875

Niu, J. (2021). Intelligent evaluation model of e-commerce transaction volume based on the combination of K-means and SOM algorithms. International Journal of Information and Communication Technology, 18, 189–206. https://doi.org/10.1504/ijict.2021.10034321

Nugroho, B. I., Rafhina, A., Ananda, P. S., & Gunawan, G. (2024). Customer segmentation in sales transaction data using K-means clustering algorithm. Journal of Intelligent Decision Support System (IDSS). https://doi.org/10.35335/idss.v7i2.236

Nursahid, W., Nugroho, B. I., & Syefudin, S. (2025). Optimalisasi preprocessing data menggunakan pendekatan CRISP-DM untuk meningkatkan kualitas klasifikasi penyakit jantung. Jurnal Artificial Intelligence dan Digital Business, 4(3), 3621–3626. https://doi.org/10.31004/riggs.v4i3.2514

Putri, A. Z., Afdal, M., Monalisa, S., & Permana, I. (2023). Penerapan algoritma fuzzy C-means pada segmentasi pelanggan B2B dengan model LRFM. Jurnal Media Informatika Budidarma. https://doi.org/10.30865/mib.v7i3.6150

Ramadani, S., Ramadhani, D. A., Ikrom, M., & Harahap, L. M. (2025). Peran strategis UMKM dalam mendorong pertumbuhan ekonomi berkelanjutan di Indonesia. Jurnal Ekonomi, Bisnis dan Manajemen, 4(1), 158–166. https://doi.org/10.58192/ebismen.v4i1.3183

Salsabila, N., Aulisari, K., & Zahro, H. Z. (2025). Penerapan algoritma K-Means untuk klasterisasi produktivitas tanaman jahe. Infotek: Jurnal Informatika dan Teknologi, 8(1), 228–238. https://doi.org/10.29408/jit.v8i1.28195

Serwah, A. M. A., Khaw, K., Yeng, C. S. P., & Alnoor, A. (2023). Customer analytics for online retailers using weighted K-means and RFM analysis. Data Analytics and Applied Mathematics (DAAM). https://doi.org/10.15282/daam.v4i1.9171

Siagian, R., Sirait, P. S. P., & Halima, A. (2021). E-commerce customer segmentation using K-Means algorithm and length, recency, frequency, monetary model. Jurnal Informatics and Telecommunication Engineering, 5(1), 21–30. https://doi.org/10.31289/jite.v5i1.5182

Syamfithriani, T. S., Mirantika, N., & Trisudarmo, R. (2023). Perbandingan algoritma K-means dan K-medoids untuk pemetaan daerah penanganan diare pada balita di Kabupaten Kuningan. Jurnal Sistem Informasi Bisnis. https://doi.org/10.21456/vol12iss2pp132-139

Wani, A., Priyanka, M., & Prasath, R. (2023). Unleashing customer insights: Segmentation through machine learning. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1–5). https://doi.org/10.1109/WCONF58270.2023.10235136

Wei, X., Zhang, Z., Huang, H., & Zhou, Y. (2024). An overview on deep clustering. Neurocomputing, 590, Article 127761. https://doi.org/10.1016/j.neucom.2024.127761

Windjaya, P. A., & Siregar, B. (2024). Analisis segmentasi pelanggan toko online marketplace berdasarkan RFMTS menggunakan algoritma K-medoids clustering. Mutiara: Multidisciplinary Scientific Journal. https://doi.org/10.57185/mutiara.v2i2.144

Yu, Y. (2022). Real-time sales forecasting algorithm of electronic commerce products in a smart city based on weighted naive Bayes. Journal of Testing and Evaluation. https://doi.org/10.1520/jte20220074

Yunita, I., Ali, P. R., Kartawidjaja, M. A., & Sukwadi, R. (2025). Segmentasi pelanggan menggunakan K-Means clustering: Menganalisis metrik RFM untuk strategi pemasaran. Jurnal Media Teknik dan Sistem Industri, 9(1), 58. https://doi.org/10.35194/jmtsi.v9i1.4452