International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 02, page 83 - 92
Submitted 12/7/2025; Accepted 15/5/2026; Published 19/5/2026

Application of K-Means Clustering in Grouping Customer Preferences for K-Pop Albums and Merchandise

[1]Aditiya Dwi Cahyo,[2]Wowon Priatna and [3]Agus Hidayat
[1,2,3]Informatics Department, Universitas Bhayangkara Jakarta Raya, INDONESIA
e-mail : [1]202110715076@mhs.ubharajaya.ac.id,
[2]wowon.priatna@dsn.ubharajaya.ac.id,[3]agus.hidayat@dsn.ubharajaya.ac.id

Corresponding Autor: Wowon Priatna

Abstract
The increasing popularity of K-Pop in Indonesia, particularly in the purchase of physical products. THJMINE Store faces challenges in inventory management and promotional strategies due to the lack of product grouping for albums and merchandise. This study applies the K-Means Clustering algorithm to 110 sales transaction data from July 2022 to January 2025. The method used in this study is the CRISP-DM approach, which consits of the following stages: business understanding, data understanding, data preparation, modeling, evaluation discussion. The result of the study show that the K-Means algorithm successfully formed three clusters with customer classification: loyal customers (cluster 0), general customers (cluster 1), and premium or collector customers (cluster 2). The model evaluation results in a DBI score of 0.6342, indicating good cluster quality. These clustering results can help THJMINE Store understand customer segmentation, develop more targeted marketing strategies, and improve inventory management efficiency.

Keywords—K-Means Clustering, Customer Preference, K-Pop Album Sales, CRISP-DM, Davies Bouldin Index

1 Introduction

In recent years, the global music industry has undergone a radical transformation, with Korean Pop (K-Pop) emerging as one of the most influential genres on the international stage. Originally a localized form of musical expression, K-Pop has grown into a global cultural phenomenon that commands a massive following across continents, influencing not only musical preferences but also fashion, beauty, and lifestyle trends. A key indicator of this global popularity is the increasing volume of album and merchandise sales that continue to climb each year. According to data published in 2023, K-Pop album sales reached an unprecedented 80 million units, a dramatic rise from previous years, reflecting a strong demand for physical collectibles even in the era of digital streaming [1]. Groups such as BTS, NCT, Seventeen, and Stray Kids have led this surge, proving that physical media retains symbolic and cultural significance, especially among dedicated fans who seek more than just auditory experiences but tangible connections to their idols.

The trend is notably visible in Southeast Asia, particularly in Indonesia, where the K-Pop wave has developed into a thriving consumer market. Data from the Korea Customs Service listed Indonesia among the top ten countries importing K-Pop albums in 2022. A survey by Katadata Insight Center in June 2022, with a sample of 1,609 Indonesian K-Pop fans, showed that approximately 30% of respondents owned merchandise such as idol photocards and photobooks, suggesting an entrenched culture of collecting and emotional investment. This enthusiasm extends beyond the music itself and into a broader ecosystem of consumption that includes concerts, digital fan meetings, and lifestyle branding. Such a multi-layered consumption behavior indicates that K-Pop is not just music—it is an identity marker, especially among teenagers and young adults who build community and self-expression through fandom participation.

The rapid expansion of the K-Pop industry has stimulated a corresponding boom in related creative and retail sectors. In Indonesia, various small and medium enterprises (SMEs) have emerged to meet the demand for K-Pop merchandise. One such enterprise is THJMINE Store, which sells albums and official merchandise from Korean entertainment companies. The store has witnessed consistent growth but faces several operational challenges, particularly in managing stock and designing marketing strategies suited to different types of consumers. One of the

core problems is the absence of structured customer segmentation. Without identifying distinct consumer clusters, the store struggles to tailor its promotional campaigns and optimize its inventory levels, often leading to overstocking or stockouts [2]. This lack of segmentation highlights the need for data-driven approaches to better understand and respond to customer preferences.

The use of data mining in the retail sector, especially in segmenting customer preferences, has gained increasing attention in academic and business circles. Data mining as a whole is a method used to ensure the accuracy of the knowledge contained in databases [3]. Several studies have demonstrated the value of clustering algorithms like K-Means in discovering hidden patterns in transactional data [4]. For instance, a study applied the K-Means algorithm to Souq.com's transaction data, segmenting products into popular, moderately popular, and less popular clusters, thus enabling better inventory and promotional strategies [5]. Likewise, in the clothing retail sector, another study successfully grouped products based on sales patterns using K-Means and evaluated them with the Davies Bouldin Index (DBI), yielding a highly compact clustering result with a DBI value of 0.035.

Building on this body of work, the present study proposes to apply the K-Means clustering algorithm to the transaction data of THJMINE Store. The objective is to group customers based on their purchasing behavior concerning K-Pop albums and merchandise, thus enabling more personalized and efficient marketing and inventory management strategies. Unlike previous studies, this research adopts the CRISP-DM (Cross Industry Standard Process for Data Mining) framework to ensure a systematic and replicable approach to data analysis [6]. CRISP-DM provides a structured process that includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment—allowing for a holistic and industry-standard methodology [7].

Customer segmentation plays a critical role in marketing and strategic planning. It allows businesses to divide their customer base into distinct groups based on behavioral, demographic, or transactional data. In the context of K-Pop retail, this means identifying fans who are more likely to purchase limited edition merchandise, those who prefer mainstream group albums, or casual consumers who engage less frequently. By using clustering methods, businesses can replace guesswork with insights, improving customer targeting and resource allocation [8]. Segmentation not only enhances the efficiency of marketing efforts but also helps build customer loyalty by offering products and services that align more closely with individual preferences.

The selection of the K-Means algorithm is justified by its simplicity, efficiency, and suitability for large-scale numerical data. K-Means works by partitioning the dataset into K clusters based on distance metrics and updating cluster centers until convergence is achieved [9]. However, its performance is sensitive to the initial number of clusters and centroid initialization [10]. To address these limitations, this study will apply the Davies Bouldin Index (DBI) as an internal evaluation metric, which assesses the compactness and separation of clusters. A lower DBI value indicates a more effective clustering outcome, providing an objective criterion for model assessment [11].

The dataset used in this study comprises 110 transaction records from July 2022 to January 2025. It includes sales of four types of albums and fifteen merchandise categories. Although the sample size is limited, the data provides a comprehensive view of the purchasing behavior of THJMINE Store's clientele. The scope of this study is restricted to THJMINE Store and does not include data from other retailers or customer interactions on social media platforms. Moreover, this research will exclusively utilize the K-Means algorithm without comparison to other clustering techniques such as DBSCAN or hierarchical clustering, ensuring a focused exploration of K-Means' practical utility.

Various prior studies validate the use of K-Means in different industries. For example, in a pharmacy retail setting [12] applied K-Means to classify drug sales data into two clusters: high and low demand, with a DBI of 0.814, indicating efficient stock management. Similarly, in the field of FMCG [13] successfully used K-Means within the KDD (Knowledge Discovery in Databases) methodology to segment beverage products into three preference-based clusters, aiding in targeted marketing strategies. These studies underscore the flexibility and effectiveness of the K-Means algorithm across diverse retail contexts.

The practical implications of this study are significant. For THJMINE Store, implementing K-Means clustering can help identify consumer groups such as collectors, casual buyers, or trend-followers, and adjust their marketing approaches accordingly. This can include designing bundle packages for specific clusters, promoting flash sales to price-sensitive groups, or investing in exclusive merchandise for high-value customers [14]. The use of Python programming in Google Colab will allow for accessible and scalable implementation of clustering models, using libraries such as Pandas, NumPy, and Scikit-learn for preprocessing, modeling, and evaluation [15].

Beyond its commercial applications, the study contributes to academic discourse on consumer analytics and clustering techniques. It expands the literature on how unsupervised learning can be used in niche markets like K-Pop and adds value by adopting CRISP-DM as a guiding framework, which is still underutilized in small business

settings. Moreover, the application of DBI adds methodological rigor by providing quantifiable evidence of clustering quality, facilitating replication and comparative studies in the future.

In essence, this research recognizes that customer preferences are dynamic, influenced by cultural trends, online communities, and emotional attachment to K-Pop idols. As such, businesses need to move beyond one-size-fits-all marketing and embrace data-driven personalization. Clustering customer preferences is not merely a statistical exercise but a strategic imperative in today's highly competitive retail environment. By applying K-Means within the CRISP-DM framework and validating results using DBI, this study aims to set a precedent for how local enterprises can leverage machine learning for actionable insights [16].

To conclude, this research addresses a pertinent challenge in the retail management of K-Pop merchandise by offering a structured and data-driven solution for customer segmentation. Through the use of K-Means clustering and objective validation via Davies Bouldin Index, the study seeks to assist THJMINE Store in making informed business decisions. It also contributes to the broader understanding of how clustering algorithms can be effectively applied in specific market niches, thus laying the groundwork for future studies that wish to explore deeper behavioral patterns in fandom-driven economies.

2 Research methods

This study was conducted to apply the K-Means Clustering algorithm in identifying and grouping customer preferences based on transaction data obtained from THJMINE Store, an Indonesian online retailer specializing in K-Pop albums and merchandise. Established in 2018, THJMINE Store began its business operations by leveraging social media platforms such as Instagram to offer limited-edition fan products. Over time, as the popularity of K-Pop culture expanded in Indonesia, the store transitioned into a full-fledged e-commerce business through platforms like Shopee. Despite facing increasingly competitive market conditions, THJMINE Store has managed to maintain its relevance through adaptive strategies. This makes it an ideal object of study for understanding consumer behavior using data mining techniques, particularly in the context of niche entertainment markets.

The research utilizes original transactional data provided by THJMINE Store's owner with consent. The dataset includes 110 transactions recorded from July 2022 to January 2025, capturing a variety of product types, including four categories of K-Pop albums and fifteen types of official merchandise such as lightsticks, photobooks, and exclusive fandom kits. These data points serve as the foundation for modeling customer preferences and behaviors through clustering. The research is based both on the store's digital platform on Shopee and its physical presence at Jl. Rusunawa Penggilingan, Tower E3, Cakung, East Jakarta.

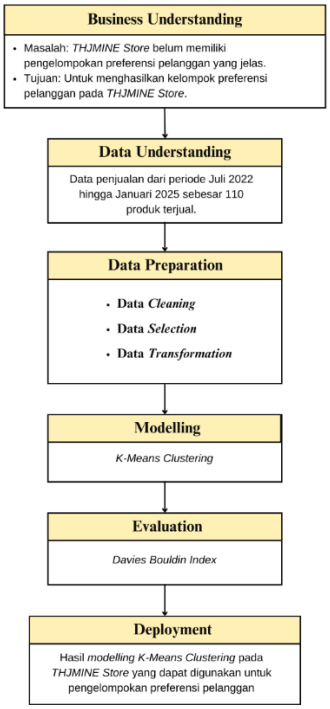


Figure 1. Research Framework

This research follows the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, beginning with the business understanding phase. The core objective identified is the lack of structured customer segmentation at THJMINE Store, which has led to inefficiencies in marketing, stock management, and bundling strategies. The goal is to address this gap by applying K-Means Clustering to form meaningful customer segments

based on their purchasing behaviors [17] , and for analyzing pesticide sales patterns in small agricultural stores. These studies demonstrate that K-Means can produce interpretable groupings that guide more informed business decisions.

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (1)$$

In the data understanding stage, the researcher closely examined the raw data to identify relevant variables and clean inconsistencies. The dataset includes the frequency and category of each purchase per customer, providing insight into purchasing patterns. This process revealed the potential of the dataset to reveal behavioral distinctions among customers [18]. Building on this understanding, the data preparation stage involved cleaning, selecting, and transforming the data using Python in the Google Colaboratory environment. Missing values were addressed using linear interpolation, and outliers were removed to reduce noise and improve model accuracy. Irrelevant fields were dropped, while categorical variables were encoded numerically. All data were normalized using the Min-Max technique to bring variables onto the same scale, making them compatible with the Euclidean distance calculations used in K-Means Clustering.

In the modeling phase, the K-Means Clustering algorithm was applied to segment customers based on purchasing behaviors. This involved running clustering operations for k-values ranging from 2 to 10 and selecting the best result through repeated initializations. For each k-value, clusters were formed by calculating the Euclidean distances between data points and centroid positions, assigning data to the nearest cluster, updating the centroids, and repeating the process until convergence was reached [19]. This approach has been successfully utilized in several fields, including the financial sector to cluster businesses based on Gross and Net Profit Margins using Python implementations of K-Means. The Python environment provides a flexible and powerful toolkit for such analysis, particularly when used with collaborative platforms like Google Colaboratory. As highlighted by [14], Google Colab supports large-scale computations and integrates seamlessly with common machine learning libraries, making it ideal for research settings with limited local computing resources.

To evaluate the clustering model, the Davies Bouldin Index (DBI) was calculated for each k-value. DBI assesses the compactness and separation of clusters, with a lower score indicating more distinct and tighter groupings. This technique has been previously used to evaluate cluster quality in studies such as IMDB movie classification, demonstrating its utility in consumer behavior modeling [11]. By comparing DBI scores across models, the optimal number of clusters for customer segmentation was selected. The characteristics of each cluster were then interpreted based on centroid values, providing insight into the product preferences and purchasing frequencies of each segment.

In the deployment phase, the results of the clustering model were translated into actionable business insights. Each cluster represented a different customer persona—such as frequent buyers, occasional collectors, or bundle-purchase enthusiasts. THJMINE Store could use this information to design targeted marketing campaigns, customize promotional bundles, and manage inventory more effectively. This real-world deployment of clustering aligns with best practices in machine learning where model output is converted into decision-support tools. The integration of such tools, facilitated through accessible cloud-based platforms like Google Colab, exemplifies the democratization of artificial intelligence tools for small and medium-sized enterprises.

To ensure the credibility of findings, the research employed three data collection methods: observation, interviews, and literature review. Observational data were collected from THJMINE Store's Shopee records, which served as the primary data source. Interviews with the store owner provided qualitative insights into business challenges such as stock mismanagement and ineffective promotions, reinforcing the need for structured customer segmentation. Finally, an extensive literature review supported the methodology, drawing from studies that applied clustering in retail and data mining contexts[10].

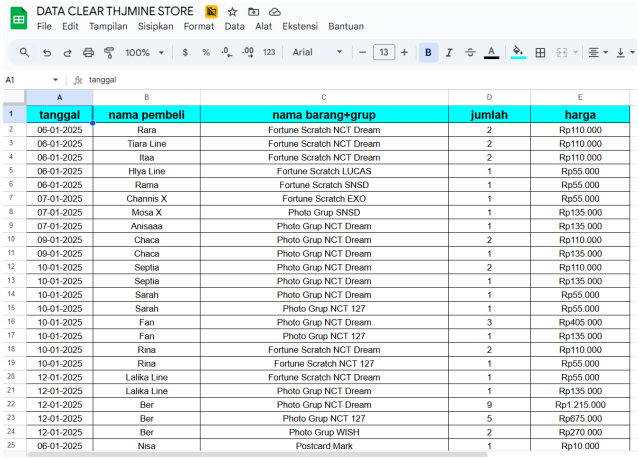
This study adopted a robust and structured approach to analyze customer preferences using K-Means Clustering and CRISP-DM methodology. The use of Python in Google Colab enabled efficient model development and analysis, while DBI ensured objective evaluation of clustering performance. Drawing upon previous studies and adapting them to the context of K-Pop retail, the methodology not only serves the operational needs of THJMINE Store but also contributes to the broader application of data mining techniques in fan-driven retail environments.

3 Results and Discussion

Discussion of the results of research and testing obtained is presented in the form of theoretical descriptions, both qualitatively and quantitatively. Experimental results should be displayed in the form of graphs or tables. For graphs, follow the format for charts and drawings.

The implementation of K-Means Clustering in identifying customer preferences for K-Pop albums and merchandise at THJMINE Store offered a deep insight into effective segmentation strategies that could support retail

optimization. Using the CRISP-DM methodology—comprising the stages of business understanding, data understanding, data preparation, modeling, and evaluation—this research transformed transactional data into valuable knowledge. The dataset included 110 sales transactions from July 2022 to January 2025, featuring various variables such as purchase date, customer name, product type categorized into albums, merchandise, and photocards, as well as quantity and total transaction value. Preliminary data analysis highlighted the predominance of merchandise sales, accounting for 46.6% of total transactions, followed by albums at 39.7%, and photocards at 13.7%, clearly indicating the central role of merchandise in contributing to revenue generation, a finding in alignment with previous internal store records.



DATA CLEAR THJMINE STORE

File Edit Tampilan Sisipkan Format Data Alat Ekstensi Bantuan

	tanggal	nama pembeli	nama barang+grup	jumlah	harga
2	06-01-2025	Rara	Fortune Scratch NCT Dream	2	Rp110.000
3	06-01-2025	Tiara Line	Fortune Scratch NCT Dream	2	Rp110.000
4	06-01-2025	Itaa	Fortune Scratch NCT Dream	2	Rp110.000
5	06-01-2025	Hlya Line	Fortune Scratch LUCAS	1	Rp55.000
6	06-01-2025	Rama	Fortune Scratch SNSD	1	Rp55.000
7	07-01-2025	Channis X	Fortune Scratch EXO	1	Rp55.000
8	07-01-2025	Mosa X	Photo Grup SNSD	1	Rp135.000
9	07-01-2025	Anisaaa	Photo Grup NCT Dream	1	Rp135.000
10	09-01-2025	Chaca	Photo Grup NCT Dream	2	Rp110.000
11	09-01-2025	Chaca	Photo Grup NCT Dream	1	Rp135.000
12	10-01-2025	Septia	Photo Grup NCT Dream	2	Rp110.000
13	10-01-2025	Septia	Photo Grup NCT Dream	1	Rp135.000
14	10-01-2025	Sarah	Photo Grup NCT Dream	1	Rp55.000
15	10-01-2025	Sarah	Photo Grup NCT 127	1	Rp55.000
16	10-01-2025	Fan	Photo Grup NCT Dream	3	Rp405.000
17	10-01-2025	Fan	Photo Grup NCT 127	1	Rp135.000
18	10-01-2025	Rina	Fortune Scratch NCT Dream	2	Rp110.000
19	10-01-2025	Rina	Fortune Scratch NCT 127	1	Rp55.000
20	12-01-2025	Lalika Line	Fortune Scratch NCT Dream	1	Rp55.000
21	12-01-2025	Lalika Line	Photo Grup NCT Dream	1	Rp135.000
22	12-01-2025	Ber	Photo Grup NCT Dream	9	Rp1.215.000
23	12-01-2025	Ber	Photo Grup NCT 127	5	Rp675.000
24	12-01-2025	Ber	Photo Grup WISH	2	Rp270.000
25	06-01-2025	Nisa	Postcard Mark	1	Rp10.000

Figure 2. Dataset THJMINE Store

In the business understanding phase, THJMINE Store was identified as experiencing significant challenges, particularly in inventory management and the implementation of effective, targeted promotional strategies. The absence of structured segmentation made it difficult for the store to allocate stock efficiently or tailor marketing messages to specific customer groups. The main aim of this research was to classify customers into distinct behavioral groups based on their purchasing habits. By doing so, personalized marketing and operational strategies could be applied. K-Means Clustering was chosen due to its capacity to detect patterns in transaction data and its wide use in retail-based studies, as supported by. In order to assess the quality and validity of the clustering results, the Davies Bouldin Index (DBI) was employed as a metric, ensuring that the clusters formed were both distinct and internally coherent.

During the data understanding stage, the research team found that data preprocessing was vital to ensure accuracy in modeling. While no missing values were detected, several inconsistencies were observed—such as variations in uppercase product names and inconsistent date formats—that required correction. For uniformity, product names were standardized to lowercase, and dates were formatted using a consistent dd-mm-yyyy format. To enhance analytical clarity, a new feature, unit price or "harga satuan," was created by dividing total transaction value by quantity, enabling the study to identify price-per-item metrics.

Table 1. Results of Removing Irrelevant Data

	tanggal	nama pembeli	nama barang	jumlah	total harga
0	07/01/2022	rika	photobook showcase event istj nct dream	7	1505000
1	07/01/2022	nunung	photobook showcase event istj nct dream	1	215000
2	07/01/022	chaca	photobook showcase event istj nct dream	1	215000
3	07/05/2022	rika	photobook showcase event istj nct dream	7	1505000
4	07/05/2022	nanda	photobook showcase event istj nct dream	1	215000

Based on product naming conventions, items were classified into three categories: albums (which included terms such as "album" or "photobook"), photocards (identified by the presence of the word "card"), and merchandise (comprising remaining products such as keychains, posters, and accessories). Visualization tools further revealed the best-selling items, with the Photobook ISTJ NCT Dream standing out among album products with 28 purchases, while the Magnetic Doll Keyring dominated the merchandise category with 40 sales. These findings were instrumental in setting inventory management priorities .

Data preparation was executed meticulously to clean and transform the dataset into a suitable format for clustering. Standardization efforts involved resolving all date format discrepancies and transforming product names to a uniform lowercase style. Aggregation processes were applied, grouping transactions by customer and categorizing their purchases across albums, merchandise, and photocards. Two additional quantitative variables were derived: total_item, representing the number of products purchased per customer, and total_belanja, indicating the total expenditure per customer. These numerical features were normalized using the RobustScaler method to address potential skewness and minimize the influence of outliers, ensuring each feature carried equal weight during clustering. This normalization approach, commonly recommended in retail analytics literature, which demonstrated its effectiveness in standardizing diverse customer data for clustering.

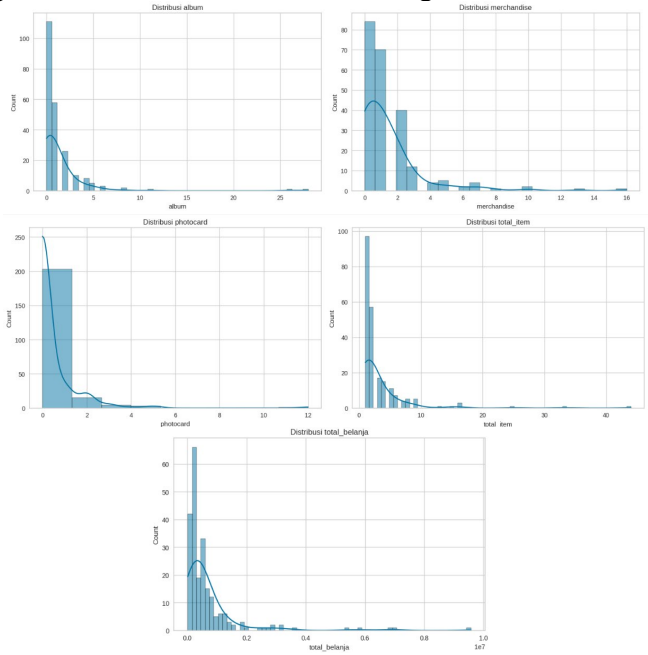


Figure 3. Distribution of album, merchandise, photocard, total_item, and total_spending data

The modeling phase began with the application of the Elbow Method to determine the optimal number of clusters, or value of k. Observations showed a sharp decrease in the Within-Cluster Sum of Squares (WCSS) up to k=3, beyond which the rate of decline plateaued, indicating diminishing returns with higher cluster counts. Based on this, k=3 was selected. K-Means Clustering was initialized with randomly generated centroids and iterated until convergence, defined as either a minimal change in centroid position (less than 0.001) or reaching a maximum of 300 iterations. The algorithm relied on the Euclidean distance formula to assign each customer to the nearest centroid, recalculating centroids after each iteration based on the mean of cluster members. After seven iterations, the clustering process converged, and the final segmentation identified three distinct clusters.

Table 2 : Result of Classification into 3 Categories

	tanggal	nama pembeli	nama barang	quantity	harga satuan	total harga	kategori
0	01/07/2022	rika	photobook showcase event istj nct dream	7	215,000	1,505,000	album
1	01/07/2022	nunung	photobook showcase event istj nct dream	1	215,000	215,000	album
2	01/07/2022	chaca	photobook	1	215,000	215,000	album

	tanggal	nama pembeli	nama barang	quantity	harga satuan	total harga	kategori
			showcase				
3	05/07/2022	rika	event istj nct dram photobook showcase	7	215,000	1,505,000	album
4	05/07/2022	nanda	event istj nct dream photobook showcase	1	215,000	215,000	album
5	05/07/2022	dyah	event istj nct dream photobook showcase	2	215,000	430,000	album
6	05/07/2022	hana	event istj nct dream photobook showcase	2	215,000	430,000	album

Cluster 0 contained 15 customers with high photocard purchases, Cluster 1 included 209 customers with moderate to low purchase activity, and Cluster 2 had only 2 customers but with exceptional transaction volume and total value. For verification, manual centroid calculations were conducted using Microsoft Excel, and the results aligned with those produced by the Python implementation, confirming the model's reproducibility and reliability.

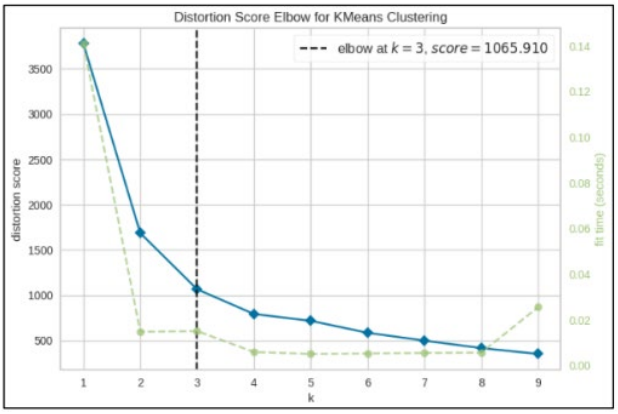


Figure 4. Elbow Curve Optimal Number of Clusters

Cluster interpretation yielded valuable behavioral profiles of THJMINE Store's customers. Cluster 0, labeled as "Loyal Customers," consisted of buyers with a preference for merchandise items such as posters and keyrings. Although they only represented 13.6% of the customer base, they were responsible for approximately 24% of revenue. Cluster 1, designated as "General Customers," comprised the majority—around 83.6% of the customer base—but their collective contribution to revenue was only 52%. These customers typically made small, infrequent purchases, often limited to single albums. Cluster 2 was labeled as "Premium Collectors," a small group that contributed disproportionately to revenue—24% from only 1.8% of customers. These customers typically bought in bulk and showed a consistent preference for both albums and merchandise. The behavioral segmentation outcomes echo the research by Surapati and Jannah (2024), which demonstrated how K-Means effectively identifies valuable niche segments in K-Pop merchandise businesses. These findings suggest that THJMINE Store could benefit from targeted interventions: offering exclusive bundle packages and pre-orders for Cluster 2, and implementing loyalty programs to retain and increase engagement from Cluster 0.

	nama pembeli	album	merchandise	photocard	total_item	total_belanja	cluster
0	abel	0	1	0	1	250000	pelanggan umum
1	abil line	1	1	0	2	642600	pelanggan umum
2	acha	2	0	0	2	110000	pelanggan umum
3	adel	0	1	0	1	250000	pelanggan umum
4	adella	0	1	0	1	541100	pelanggan umum
5	aiish	2	0	2	4	365000	pelanggan umum
6	aisy	0	3	0	3	750000	pelanggan umum
7	ala	0	2	0	2	500000	pelanggan umum
8	ale	0	8	0	8	2000000	pelanggan loyal
9	alsa n x	1	0	0	1	140000	pelanggan umum
10	alvina	0	1	0	1	655000	pelanggan umum
11	amanda	2	1	0	3	470000	pelanggan umum
12	amon	1	0	0	1	140000	pelanggan umum
13	anis	1	0	0	1	265000	pelanggan umum
14	anisaa	1	0	0	1	215000	pelanggan umum
15	anisaaa	0	1	0	1	135000	pelanggan umum
16	annisa n	2	5	2	9	1495000	pelanggan loyal
17	apyy	0	1	0	1	250000	pelanggan umum
18	arum line	0	1	0	1	520000	pelanggan umum
19	astrid	0	1	0	1	541000	pelanggan umum
20	avft line	0	1	0	1	180000	pelanggan umum

Figure 5. Dataset Reslut With Label

The evaluation stage employed the Davies Bouldin Index to assess cluster quality. With a DBI score of 0.6342 at k=3, the clusters demonstrated strong separation and internal cohesion, with lower DBI values indicating better segmentation. The DBI formula used was:

$$DBI = \frac{1}{k}\sum_{i=1} K\ max_{i \neq j}\left(\frac{d(ci,cj)}{\sigma i + \sigma j}\right) \quad (2)$$

Additional comparative analysis for values of k ranging from 2 to 10 supported the choice of k=3, as alternative values produced DBI scores exceeding 0.7, indicating lower clustering effectiveness. This evaluation result aligns, who suggested that a DBI score below 0.6342 is a reliable benchmark for effective segmentation in retail analytics.

Table 3. Evaluation DBI Results

Cluster	DBI
Jumlah Cluster 2	0.6466
Jumlah Cluster 3	0.6342
Jumlah Cluster 4	0.8712
Jumlah Cluster 5	1.0955
Jumlah Cluster 6	0.8722
Jumlah Cluster 7	0.9251
Jumlah Cluster 8	0.6872
Jumlah Cluster 9	0.6883
Jumlah Cluster 10	0.6624

The discussion of results offers critical interpretations for business strategy. The dominance of Cluster 1 (general customers) highlights that THJMINE Store's primary market consists of casual buyers. However, the high revenue contributions of the smaller Clusters 0 and 2 reflect a manifestation of the Pareto principle, where a small percentage of customers accounts for a large portion of revenue. From a strategic perspective, this suggests several actions. Premium Collectors in Cluster 2 should be targeted with limited-edition or early-access products, as their spending habits indicate a willingness to pay for exclusivity. General Customers in Cluster 1 could be motivated to engage more deeply through introductory promotions such as discounts on their first album purchase. Meanwhile, Loyal Customers in Cluster 0 could be encouraged to increase frequency or basket size through bundled offers combining merchandise and photocards or through the introduction of a point-based loyalty system.

From an operational standpoint, inventory management should prioritize items with consistently high demand, such as the Magnetic Doll Keyrings and NCT Dream photobooks, to avoid stockouts while minimizing the accumulation of unsold inventory. These decisions must be grounded in real-time data and adjusted based on seasonal trends or the promotional cycle of K-Pop comebacks. Methodologically, the research lays a foundation for further exploration by proposing the use of alternative clustering techniques such as DBSCAN or hierarchical clustering, which may offer improved handling of outliers like those seen in Cluster 2. Future implementations could also integrate real-time analytics tools, enabling THJMINE Store to adapt segmentation dynamically as customer behavior and K-Pop trends evolve—a recommendation aligned with the forward-looking approach.

In conclusion, the use of K-Means Clustering in this research succeeded in dividing THJMINE Store's customer base into three distinct and actionable groups. The use of the Davies Bouldin Index further confirmed the statistical

validity of the clustering results, ensuring that the conclusions drawn were not only interpretable but also robust. The insights obtained from this research can directly support the store's efforts in precision marketing, strategic inventory planning, and the creation of value-based customer relationships. Beyond its practical impact, this study also contributes to the growing body of research on the application of data mining and unsupervised learning techniques in niche retail segments such as K-Pop merchandise, highlighting the potential for data-driven decision-making in enhancing competitiveness and sustainability.

4 Conclusion

The application of K-Means Clustering at THJMINE Store successfully grouped customers into three coherent segments—Loyal Customers, General Customers, and Premium Collectors—demonstrating that transactional data can be transformed into actionable knowledge for precision marketing and inventory management. The clusters were validated by a Davies Bouldin Index of 0.6342, confirming good separation and internal cohesion, while the sales distribution analysis showed that a small cohort of high-value buyers contributes a disproportionately large share of revenue, echoing the Pareto principle and underscoring the strategic importance of targeted engagement.

This data-driven approach brings clear advantages: it enables personalized promotions, optimizes stock levels in line with real demand, and directs marketing resources toward segments with the highest return on investment. Nonetheless, the methodology has limitations. K-Means is sensitive to outliers and to the random initialization of centroids, and it requires the number of clusters to be set in advance, which may not capture evolving purchase patterns in a fast-moving market such as K-Pop merchandise.

Future work can mitigate these drawbacks by integrating real-time transaction streams so that segments update dynamically as trends shift, and by experimenting with algorithms such as DBSCAN or hierarchical clustering that are better suited to handling irregular or sparse high-value behavior. Extending the model to incorporate social-media sentiment or preorder activity could further refine demand forecasting, while A/B-testing tailored campaigns for each cluster would provide empirical feedback on the financial impact of this segmentation strategy.

5 Suggestion

One primary area for improvement lies in the sensitivity of the K-Means algorithm to outliers and initial centroid selection, which may influence the stability of clustering results. Future studies could explore the use of alternative clustering methods such as DBSCAN or hierarchical clustering that do not require a predefined number of clusters and can better accommodate irregular purchasing behaviors, especially from rare but high-value customers. Another potential enhancement involves integrating real-time sales data into the clustering model to support dynamic segmentation that adjusts to shifting customer preferences in response to seasonal K-Pop trends or major product launches. Additionally, expanding the dataset beyond 110 transactions would likely improve the generalizability and robustness of the findings. Including behavioral data such as purchase frequency over time, product ratings, or responses to promotions could enrich the feature set used for clustering and yield more nuanced customer profiles. Lastly, applying advanced evaluation metrics alongside the Davies Bouldin Index, such as the Silhouette Coefficient or Calinski-Harabasz Index, may provide a more comprehensive understanding of clustering validity and guide optimal model selection for future implementations.

6 Acknowledgments

The author would like to thank Universitas Bhayangkara Jakarta Raya, especially the Faculty of Computer Science, for the academic guidance and support throughout the research process. Special appreciation is extended to the academic supervisor who provided valuable insights and constructive feedback. Gratitude is also expressed to the THJMINE Store for granting access to sales data, which was essential for conducting this study. Lastly, the author would like to thank family and friends for their continuous encouragement and motivation during the completion of this research.

BIBLIOGRAPHY

[1] S. Lestari, "Analisis Algoritma Regresi Linear Sederhana dalam Memprediksi Tingkat Penjualan Album KPOP," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 2, no. 1, hal. 199–209, 2023, doi: 10.55123/insologi.v2i1.1692.

[2] U. Surapati dan M. Jannah, "Penerapan Data Mining Menggunakan Metode K-Means Untuk Mengetahui Minat Customer Dalam Pembelian Merchandise Kpop," *Jurnal Sains dan Teknologi*, vol. 5, no. 3, hal. 875–884, 2024, doi: 10.55338/saintek.v5i3.2739.

[3] M. Mayadi, S. Setiawati, dan W. Priatna, "Pengelompokan Hasil Survei MBKM Menggunakan K-Mean dan

K-Medoids Clustering," *Jurnal Media Informatika Budidarma*, vol. 7, no. 1, hal. 426, 2023, doi: 10.30865/mib.v7i1.5003.
[4] D. Aryani, B. Irawan, dan A. Bahtiar, "Implementasi Data Mining Pada Data Penjualan Pakaian Menggunakan Algoritma K-Means Dengan Optimize Parameter Grid," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, hal. 1673–1680, 2024, doi: 10.36040/jati.v8i2.9147.
[5] F. Amin, D. S. Anggraeni, dan Q. Aini, "Penerapan Metode K-Means dalam Penjualan Produk Souq.Com," *Applied Information System and Management (AISM)*, vol. 5, no. 1, hal. 7–14, 2022, doi: 10.15408/aism.v5i1.22534.
[6] M. TB Ai, *Data Mining Menggunakan R: Teori dan Praktik*. Banten: PT Bale Damar Publishing, 2023.
[7] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, hal. 3048–3061, Agu 2021, doi: 10.1109/TKDE.2019.2962680.
[8] E. F. L. Awalina dan W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail," *Jurnal Teknologi dan Informasi*, vol. 13, no. 2, hal. 122–137, 2023, doi: 10.34010/jati.v13i2.10090.
[9] I. Safira, R. Salkiawati, dan W. Priatna, "Penerapan Algoritma K-Means untuk Mengetahui Pola Persediaan Barang pada Toko Raja Bekasi," *Journal of Informatic and Information Security*, vol. 3, no. 1, hal. 99–110, 2022, doi: 10.31599/jiforty.v3i1.1253.
[10] N. Wisna, S. A. P. Lisna, T. Fahrudin, dan R. B. Kotjoprayudi, "Analisis Gross Profit Margin (Gpm) Dan Net Profit Margin (Npm) Dengan Metode Algoritma K-Means Menggunakan Bahasa Pemrograman Python," *Jurnal Ilmiah Manajemen, Ekonomi, & Akuntansi (MEA)*, vol. 7, no. 2, hal. 1199–1210, 2023, doi: 10.31955/mea.v7i2.3121.
[11] I. F. Ashari, R. Banjarnahor, D. R. Farida, S. P. Aisyah, A. P. Dewi, dan N. Humaya, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," *Journal of Applied Informatics and Computing*, vol. 6, no. 1, hal. 07–15, 2022, doi: 10.30871/jaic.v6i1.3485.
[12] K. Kurnia Abdullah, I. Maulana, A. Suharso, G. Garno, dan A. Primajaya, "Penerapan Algoritme K-Means Dalam Klasterisasi Data Penjualan Obat Apotek Kidangrangga," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 2, hal. 1274–1279, 2023, doi: 10.36040/jati.v7i2.7061.
[13] M. Rochmawati *et al.*, "Implementasi Algoritma K-Means dalam Klasterisasi Penjualan pada Sebuah Perusahaan menggunakan Metodologi KDD Implementation of the K-Means Algorithm in Sales Clustering at a Company using the KDD Methodology," vol. 13, hal. 54–62, 2024.
[14] R. Gelar Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 1, hal. 55–60, 2023, doi: 10.47233/jteksis.v5i1.750.
[15] R. T. Handayanto and H. Herlawati, "Prediksi Kelas Jamak dengan Deep Learning Berbasis Graphics Processing Units," *Jurnal Kajian Ilmiah*, vol. 20, no. 1, hal. 1410–9794, 2020, [Daring]. Tersedia pada: http://ejurnal.ubharajaya.ac.id/index.php/JKI
[16] S. Aulia, "Klasterisasi Pola Penjualan Pestisida Menggunakan Metode K-Means Clustering (Studi Kasus Di Toko Juanda Tani Kecamatan Hutabayu Raja)," *Djtechno: Jurnal Teknologi Informasi*, vol. 1, no. 1, hal. 1–5, 2021, doi: 10.46576/djtechno.v1i1.964.
[17] H. Astuti, "Penerapan Data Mining Menggunakan Metode K-Means Clustering Untuk Pengelompokkan Data Pelanggan (Studi Kasus : PT. Pinus Merah Abadi)," *Jurnal Web Informatika Teknologi*, vol. 4, no. 1, hal. 9, 2020.
[18] S. Pujiono, R. Astuti, dan F. Muhamad Basysyar, "Implementasi Data Mining Untuk Menentukan Pola Penjualan Produk Menggunakan Algoritma K-Means Clustering," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, hal. 615–620, 2024, doi: 10.36040/jati.v8i1.8360.
[19] N. Ahsina, F. Fatimah, dan F. Rachmawati, "Analisis Segmentasi Pelanggan Bank Berdasarkan Pengambilan Kredit Dengan Menggunakan Metode K-Means Clustering," *Jurnal Ilmiah Teknologi Infomasi Terapan*, vol. 8, no. 3, 2022, doi: 10.33197/jitter.vol8.iss3.2022.883.