

Mapping the Distribution of Pak Iwan's Tofu Sales Area Using the K-Means Method Based on Sales Volume

Dimas Arya Bintara ^{1*}, Muhammad Amin ², Ahmad Muhazir ³

^{1*,2,3} Information Systems Study Program, Faculty of Computer Science, Universitas Royal, Asahan Regency, North Sumatra Province, Indonesia

Email: aryabintaradimas@gmail.com ^{1*}, muhammadamin@gmail.com ², ahmadmuhazir@gmail.com ³

Article history:

Received September 24, 2026

Revised October 12, 2026

Accepted October 15, 2026

Abstract

The rapid development of information technology has reshaped many sectors, particularly business activities. In this study, raw sales data are processed into useful information through the grouping of distribution areas based on product sales levels. A quantitative approach is applied by using historical distribution and sales data as the basis of analysis. The findings show that K-Means can organize multidimensional data in a systematic and objective manner, producing regional classifications that are more reliable than decisions based only on intuition or prior experience. The clustering results classify seven marketing areas into three groups. Cluster 1, categorized as Very Popular, includes Air Joman and Silau Laut. Cluster 2, categorized as Popular, consists of Sidokeno. Cluster 3, categorized as Less Popular, includes Sei Kamah, Pulo Bandring, East Kisaran, and West Kisaran.

Keywords:

K-Means Clustering; Data Mining; Distribution Area Mapping; Product Sales Level.

1. INTRODUCTION

Digital transformation has fundamentally reshaped how businesses operate and make decisions. What was once considered a supporting tool has now become an operational necessity. Modern enterprises can no longer rely solely on intuition; they must be able to convert transaction data into measurable, justifiable decisions. Yet in practice, many businesses particularly small enterprises continue operating the old way. Sales records accumulate in systems, but rarely are they leveraged effectively. Business owners make decisions about distribution based on years of experience rather than what the numbers actually reveal (Tegar & Sari, 2025). The food industry presents unique challenges distinct from other sectors. Products like tofu have extremely short shelf lives. A single day's delay in distribution can mean the difference between a product that sells and one that spoils. When distribution decisions lack solid analytical backing, the result is waste on one side and lost sales on the other. Both problems are equally damaging. Excess stock in certain areas represents wasted capital; insufficient stock in others means disappointed customers and foregone revenue. For businesses operating on thin profit margins, such distribution errors can be catastrophic (Mutiah et al., 2024).

Pak Iwan's tofu business faces challenges similar to many small and medium enterprises. Decisions about how much product to send to each sales area still depend heavily on habit and personal judgment. Yet each area has distinct characteristics different population sizes, different purchasing power, different consumption preferences. What sells well in one area may not sell equally well in another. More troubling, sales patterns can shift over time, yet these changes often go undetected until it is too late. Consequently, distribution planning perpetually lags behind market reality. For perishable goods like tofu, this lag is not merely an operational issue it is a genuine financial problem that directly affects which areas receive adequate supply and which do not. The data needed to make better distribution decisions already exists. Daily sales records from each distribution area reflect actual market behavior and regional demand. The problem is that raw sales data does not speak for itself. Numbers must be organized, grouped, and analyzed in meaningful ways to reveal which areas are performing well and which require attention. This is where data mining becomes relevant. Rather than allowing sales records to remain static archives, data mining enables us to uncover hidden patterns in regional sales, distinguish strong-performing areas from weak ones, and identify opportunities for better supply

allocation. For businesses with numerous distribution points and varying regional demand, this approach is far more reliable than memory or general impressions (Windarto, 2017).

Clustering is one of the most suitable techniques for mapping sales distribution patterns. Unlike methods requiring predefined labels or categories, clustering groups data based on their natural similarities in sales behavior. The result is clusters that reflect actual patterns within regional sales data. For distribution area analysis, this means identifying which areas exhibit comparable sales volumes and patterns without making assumptions beforehand. K-Means is among the most practical clustering algorithms for this purpose. It is simple, fast, and effective at handling numerical data like sales volumes across multiple regions. Previous research has demonstrated that K-Means can be applied across diverse contexts from classifying student learning styles, grouping households by socioeconomic facilities, to optimizing resource distribution proving the method's adaptability for mapping and categorizing sales areas (Anwar & Wulandari, 2025; Mutiah et al., 2024; Wibowo & Habanabakize, 2022; Harliana et al., 2021). Why is K-Means appropriate for mapping Pak Iwan's tofu sales distribution? Because the distribution of tofu products depends on measurable sales volumes across different areas. The product range includes fried tofu, jumbo fried tofu, semi-cooked tofu, and raw tofu. Each category likely shows different sales volumes depending on regional preferences and consumption habits. By processing sales volume data from each distribution area, the algorithm can classify areas into clusters reflecting their sales intensity. The result is a clear map: which areas absorb products well, which show weak sales response, and which fall in between. This data-driven approach to mapping aligns with current business trends, where machine learning and analytical systems are used for market segmentation, location optimization, and operational planning (Maulani et al., 2025; Miranda & Sriani, 2025; Wu et al., 2024).

The benefits of mapping distribution areas based on sales volume extend beyond mere data grouping. When sales areas are mapped and categorized according to their actual sales performance, business owners gain a solid foundation for supply priority decisions. High-volume areas can receive more frequent and larger deliveries. Moderate-volume areas require closer monitoring to determine whether sales are stable or fluctuating. Low-volume areas may need different distribution strategies perhaps adjusted quantities, modified delivery frequency, or more targeted promotion. Without such mapping, distribution decisions tend to become inconsistent and inefficient. The result is excess stock accumulating in some areas while others face shortages. Research shows that clustering-based mapping of sales areas can deliver real benefits for business sustainability, particularly regarding supply allocation optimization and inventory management (Bandyopadhyay et al., 2021). Another equally important aspect of mapping sales distribution is reducing bias in decision-making. Business owners frequently rely on personal judgment when deciding which areas deserve priority in product distribution. Experience certainly has value, but experience also carries bias, especially when market conditions and sales patterns shift rapidly. Data-driven mapping helps balance intuition with measurable evidence about actual sales volumes. K-Means works by comparing each area objectively against its nearest cluster center based on sales data. This process makes the resulting area classifications more transparent and easier to explain to stakeholders. In practice, this means distribution planning becomes more disciplined and delivery allocation more efficient. Previous studies show that implementing K-Means for mapping sales areas produces more accurate categorizations, and optimization through additional algorithms can further enhance the quality of area classifications (Naufaldy et al., 2026; Ningtyas & Pribadi, 2023; Harliana et al., 2021).

For small enterprises like Pak Iwan's tofu business, mapping distribution areas through sales volume analysis is not merely "nice to have" it is a genuine necessity. Profit margins in the food business are razor-thin. Small errors in how products are distributed across areas directly impact the bottom line. Sending too much product to low-sales areas means spoilage and total loss. Sending too little to high-sales areas means lost revenue and disappointed customers. Neither scenario is theoretical both occur regularly in the field. Therefore, the need for structured analytical tools to map and optimize distribution is not academic. It is a response to concrete operational problems that affect business viability and the ability to serve each sales area appropriately. Sound management information systems form the foundation for effective distribution mapping. When sales data from each area are stored systematically, they become easier to access, compare, and analyze. Well-organized systems facilitate tracking of sales performance across different areas and time periods. Properly structured sales data can then be processed for clustering analysis and distribution area mapping. The relationship between data recording and analysis is critical because the quality of distribution mapping depends on data quality from the start. Messy, inconsistent records make subsequent mapping analysis difficult and unreliable; consistent, structured sales records support more trustworthy area classifications and distribution recommendations. Research demonstrates that implementing K-Means for mapping sales distribution in retail and food distribution contexts has proven effective at identifying area patterns and providing actionable recommendations for supply allocation (Nasyuha et al., 2022).

This study aims to map Pak Iwan's tofu sales distribution areas using the K-Means method, with analysis based on sales volume data. The specific objective is to transform raw sales records from each distribution area into structured, actionable findings about area performance. These findings will support the categorization of sales areas and improved distribution planning. The expected outcome is a clear map showing the sales potential of each distribution area, enabling the business owner to make more informed decisions regarding supply allocation to different areas, production planning based on area demand, and prioritization of delivery

routes. Through this mapping approach, distribution decisions move beyond assumptions and habits, instead relying on real, measurable sales volume patterns from each sales area.

2. RESEARCH METHOD

A research method section has one job: explain how a study moves from raw material to findings people can actually trust. For research dealing with sales records and area mapping, the method is more than just a procedural checklist—it determines whether data become actionable intelligence or stay as numbers on a spreadsheet. This study uses a quantitative approach because the primary material consists of historical distribution and sales volume data from multiple areas. A quantitative design makes sense here because mapping sales distribution relies on numerical records that can be systematically cleaned, compared, and processed across different regions (Wahyudi & Silfia, 2022; Almaripat et al., 2025).

The study starts by examining the operational challenges facing Tahu Pak Iwan's distribution system. The main issue is not a shortage of sales data, but rather that available sales volume data from each area remain underutilized. Daily sales records from different distribution areas sit in archives, never converted into a map showing which areas perform well and which do not. This creates a practical gap: management cannot clearly see sales patterns across areas or understand which distribution regions generate strong demand and which do not. When product allocation to each area still relies on habit or memory, mismatches between supply and actual area demand become inevitable. For a tofu business, where shelf life is measured in days, such distribution mismatches quickly translate into quality degradation, poor stock turnover, and lost revenue. A clustering approach offers a way to map areas based on sales volume, moving beyond guesswork and replacing it with pattern-based area grouping anchored in actual sales data (Sonatha et al., 2025; Putri et al., 2023).

Data come from Tahu Pak Iwan's internal sales records across all distribution areas, supplemented by direct observation of the distribution process. The variables included in the area mapping analysis are the quantity of tofu sold per area, the frequency of orders from each sales region, and daily remaining stock levels by area. These variables were selected because they directly show how products move from production through to each market area. If sales volume data are incomplete, inconsistent, or recorded in different formats across areas, the mapping results will drift away from actual conditions. The collection stage therefore requires careful handling to ensure consistent recording across all distribution areas. The final output's accuracy—particularly the quality of area mapping—depends on how thoroughly sales volume data are gathered and validated across all areas from the beginning (Assegaf et al., 2025; Lathifah et al., 2026).

Once sales records from all areas are collected, data go through cleaning to remove incomplete entries, duplicates, or invalid values that could interfere with area mapping. After cleaning, the sales volume data are restructured into a format suitable for K-Means processing. This study uses three clusters to map sales areas—Very Popular, Popular, and Less Popular—to categorize distribution areas based on sales volume. Three clusters were chosen to keep the area map readable while still distinguishing areas with different sales volume levels. Too many clusters obscure area patterns; too few hide meaningful differences between areas. The goal is not to create complicated classifications, but to build a practical area map that reflects actual sales volume behavior across regions. K-Means has proven effective for product sales segmentation and area classification when sales volume records contain measurable similarities and differences across groups (Siregar, 2024; Rismayadi et al., 2021).

At this stage, the study designs a web-based information system to process sales volume data and display the area mapping and clustering results. The system allows users to input transaction data from each area, run the K-Means process, and view the resulting area map in readable format. The system workflow mirrors the algorithm's logic: centroid initialization, distance calculation based on sales volume, cluster assignment for each area, and final area grouping. System structure is defined using UML diagrams—Use Case, Activity, Sequence, and Class Diagrams—which specify system behavior, user interactions, and data flow for the mapping process. Clear design matters because a mapping method that works well in theory can become impractical if the system interface is poorly built. In decision-support applications for distribution and marketing strategy development, usability determines whether the area map actually gets used in daily operations (Almaripat et al., 2025; Anitha & Patil, 2022).

The built system is tested using actual sales volume data from Tahu Pak Iwan across all distribution areas to determine whether the K-Means algorithm maps and groups areas correctly based on sales volume. Testing goes beyond checking for program errors. It examines whether the area grouping logic aligns with real-world sales patterns and whether the resulting area map makes practical sense. System results are compared against actual distribution conditions to assess whether the area classifications and cluster labels reflect genuine sales volume differences. If the area map diverges significantly from observed sales behavior, the mapping process requires review. Testing serves as a checkpoint that separates a working prototype from a reliable area mapping tool. Research on K-Means clustering applications confirms that testing is essential because area clustering results must be dependable before guiding distribution decisions (Assegaf et al., 2025; Siregar, 2018).

After test results are reviewed, the system undergoes refinement to improve the area mapping display and functionality. If calculations run slowly, code is optimized. If the area map interface is hard to read, layout is

adjusted. Implementation aims to make the system accessible for business managers who need quick access to area mapping results. Tables and graphs help users interpret the area distribution map without having to examine raw sales volume numbers one by one. Such a system only proves valuable when it bridges analytical results and practical distribution decision-making. K-Means clustering applications have shown that this method can transform transaction records into structured area maps and marketing strategies that are easier to act on, especially when the distribution map is displayed clearly and consistently (Putri et al., 2023; Wahyudi & Silfia, 2022).

3. RESULTS AND DISCUSSION

3.1. Result

The Results and Discussion section presents the findings obtained from processing Tahu Pak Iwan's product distribution data using the K-Means algorithm. The analysis was carried out on product sales data from each distribution area to produce regional groupings based on sales characteristics. The clustering results were then interpreted to determine sales distribution patterns, identify areas with high, medium, or low demand, and evaluate the current distribution conditions. The discussion not only explains the grouping results obtained, but also describes the relationship between the analysis results and operational conditions in the field, so as to provide a clearer picture of the effectiveness of product distribution. The following is the explanation.

3.1.1. Analysis of the Running System

Tahu Pak Iwan's distribution system is still managed conventionally and relies on intuition, past experience, and manual records. Although daily sales transactions are documented, the data remain static archives and have not been used to map market potential or product demand accurately. As a result, differences in tofu sales across regions are not clearly identified, causing inefficiency in product allocation. This condition often leads to excess stock or shortages, which is risky because tofu has a short shelf life and is easily damaged. The current system also depends on physical documents, manual processing, and handwritten order notes, so sales data cannot be transformed into useful information for decision-making. Therefore, a more structured and data-based approach is needed to improve distribution accuracy and support better operational control.

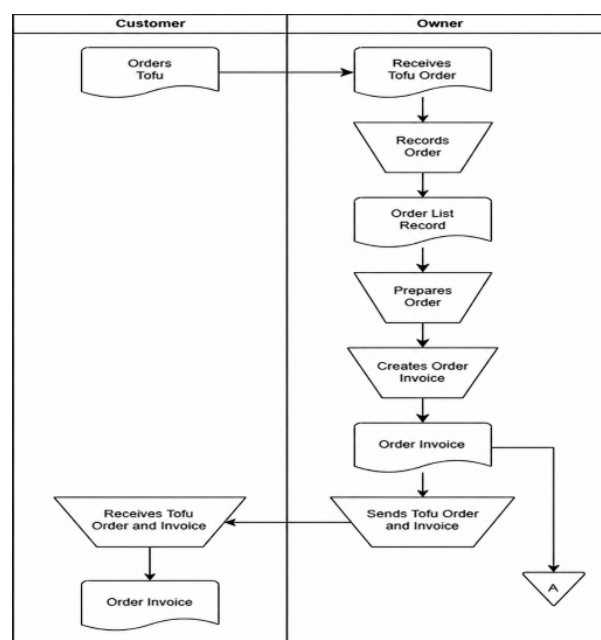


Figure 1. Current System Flow

3.1.2. Proposed System Analysis

Proposed system analysis is designed to replace the weaknesses of the current manual distribution process at Tahu Pak Iwan by using a data mining approach with the K-Means algorithm. The system processes daily sales transaction data, calculates the distance of each record to the centroid, and groups distribution areas into three categories: very popular, popular, and less popular. The proposed flow is built around two user roles, admin and owner. The admin logs in, manages master data such as customers, distribution areas, and tofu product types, then enters sales summaries for analysis. The clustering results are stored in the database and displayed through the system. After processing, the owner logs in to review the dashboard, which presents real-time mapping results and detailed clustering information, including centroid distances. The system also

provides printable reports to support decision-making and improves distribution efficiency, accuracy, and control.

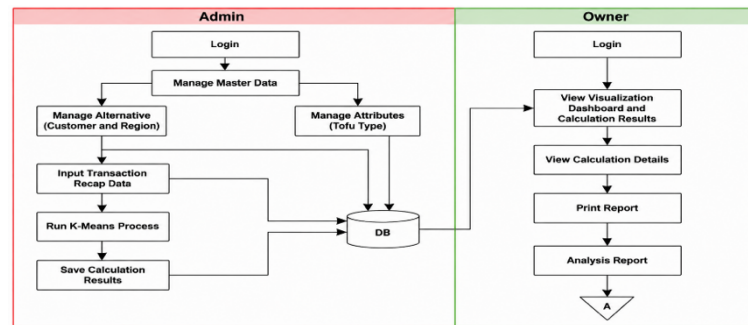


Figure 2. Proposed System Flow

3.1.3. Use Case Diagram

The system process to be designed is illustrated through a Use Case Diagram. Several activities are carried out during the operation of this system, such as logging in as a user, processing master criteria and alternative data, managing transaction recap data, performing K-Means calculations, viewing clustering results, and printing clustering results. The proposed Use Case Diagram design for this system can be seen in the figure 3.

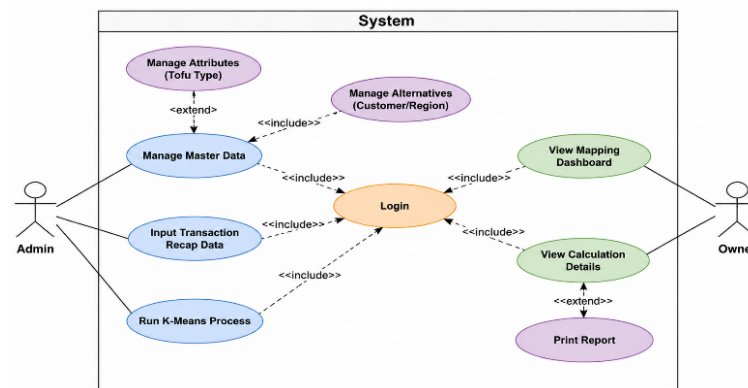


Figure 3. Use Case Diagram

3.1.4. Use of the K-Means Clustering Method

In this study, the K-Means Clustering algorithm was applied to group Tahu Pak Iwan's distribution areas based on sales levels for Hollow Tofu, Jumbo Tofu, Half-Cooked Tofu, and Raw Tofu. The method was selected because it allows the data to be classified objectively through the comparison of sales patterns across regions, without relying on prior group definitions. The process started with collecting sales transaction data from the operational database, then continued with preprocessing to clean the records and combine daily entries into total sales per region. The recapitulation of total sales volume after the pre-processing process can be seen in Table 1 below. After that, the algorithm calculated the distance between each data point and the centroid using the Euclidean Distance method until the cluster positions became stable. The final grouping produced three sales categories: very popular, popular, and less popular. These results give the business owner a more reliable basis for deciding distribution priorities, arranging stock more carefully, and improving operational efficiency in line with actual sales conditions.

Table 1. Pre-processing Data (Total Accumulated Sales)

Buyer Name	Region	Hollow	Jumbo	Half-baked	Raw	Total
Mrs. Mar	Joman Water	9 1,230	42 .755	8,870	5 2,485	1 95,340
Mr. Milan	Sea Glare	5 6,910	5.575	5 7,490	57 .345	17 7,320
Mr. Putra	Sidokeno	85 .510	9. 865	6. 010	3.005	10 4,390
Mr. Men	Sei Kamah	11 .150	2 8,455	3. 150	3,005	45,760
Mr. Parno	Bandring Island	5,430	14,155	2,006	1,861	23,425
Mr. Yoga	East Range	11,150	5,575	3,150	3,005	22,880
Mrs. Tina	West Range	5,430	7,005	2,006	2,433	16,874

After the data is cleaned and accumulated, the next stage is Transformation, which is converting the data into a mathematical variable format to match the Euclidean Distance formula in the K-Means algorithm. Each

region is initialized as an Alternative () and each type of tofu product is initialized as a Criteria / Attribute (). The details of the alternative data are presented in Table 2.

Table 2. Alternative Data Transformations

Alternative	Description (Region)	X1 (Hollow)	X2 (Jumbo)	X3 (Set. Ripe)	X4 (Raw)
A1	Joman Water	91,230	42,755	8,870	52,485
A2	Sea Glare	56,910	5,575	57,490	57,345
A3	Sidokeno	85,510	9,865	6,010	3,005
A4	Sei Kamah	11,150	28,455	3,150	3,005
A5	Bandring Island	5,430	14,155	2,006	1,861
A6	East Range	11,150	5,575	3,150	3,005
A7	West Range	5,430	7,005	2,006	2,433

To map Tahu Pak Iwan's distribution effectively, the analysis employed three clusters (K=3) to separate regions by their actual sales behavior. Cluster C1 groups Very Popular areas—regions where customers consistently order in large quantities and inventory moves quickly. Cluster C2 captures Popular regions, where sales remain steady but lack the urgency of C1. Cluster C3 identifies Less Popular areas, where demand remains modest and stock turnover slows. Rather than relying on assumptions, this classification lets sales data speak directly. By dividing regions into these three tiers, the owner gains a practical foundation for deciding where to prioritize deliveries, allocate resources, and adjust production schedules based on real market behavior rather than guesswork. The next step is to randomly determine the initial centroids from the alternative data set. For this first iteration, three initial data points were selected, representing the high, middle, and low values of the dataset: alternatives A1, A3, and A7. The initial centroid values can be seen in Table 3.

Table 3. Determination of Initial Centroid (Iteration 1)

Centroid	X1	X2	X3	X4	Information
C1	91,230	42,755	8,870	52,485	Taken from the value of A1
C2	85,510	9,865	6,010	3,005	Taken from A3 value
C3	5,430	7,005	2,006	2,433	Taken from the value of A7

Each regional sales data point is assigned to the nearest cluster by measuring the distance between the data and each cluster center. The Euclidean Distance formula calculates how far each region's sales profile lies from the three initial centroids. Once the distance calculations are completed for iteration 1, the algorithm identifies the minimum distance value among C1, C2, and C3 for every region. Each region is then grouped into the cluster that produces the smallest distance, ensuring that areas with similar sales patterns are placed together. This assignment process forms the foundation for subsequent iterations and cluster refinement.

Table 4. First Iteration K-Means Calculation (Minimum & Cluster)

Alternative	C1	C2	C3	Minimum	Cluster
A1	0.00	59757.18	105792.34	0.00	1
A2	70340.16	80245.81	93520.25	70340,16491	1
A3	59757.18	0.00	80233.07	0	2
A4	95384.93	76701.87	22236.39	22236.39	3
A5	103872.54	80302.87	7172.84	7172.843509	3
A6	101371.34	74538.54	6033.18	6033.18	3
A7	105792.34	80233.07	0.00	0	3

Based on the distance between each data point and the initial centroid, each data point is placed in the cluster with the shortest distance. The first iteration shows that tofu is grouped into three clusters based on its sales volume. The clustering results from the first iteration are then used to calculate new centroids, which will be used in subsequent iterations until no cluster changes occur.

Table 5. New Cluster Centers of the Second Iteration

Alternative	C1	C2	C3
A1	1		
A2	1		
A3		1	
A4			1
A5			1
A6			1
A7			1

After the distance calculation process is carried out in the first iteration, each data point has been grouped into the cluster with the closest distance to the centroid. The next step is to calculate a new centroid for each cluster based on the average attribute values of the cluster members. The new centroid calculation is performed using the following average formula.

Table 6. New Centroid C1 (A1, A2)

Attribute	Calculation	Results
Hollow	$(91230 + 56910) / 2$	74070
Jumbo	$(42755 + 5575) / 2$	24165
Half-baked	$(8870 + 57490) / 2$	33180
Raw	$(52485 + 57345) / 2$	54915

The second iteration begins again by calculating the distance to the new centroid as in step 6 above. The formula used is also the same as in equation above. The results of the distance calculation in this second iteration can be seen in the following table 7.

Table 7. K-Means Calculation (Minimum & Clust)

Alternative	C1	C2	C3	Minimum	Cluster
A1	35170.08	59757.18	101232.79	35170.08	1
A2	35170.08	80245.81	91904.76	35170.08	1
A3	61385.83	0.00	77397.39	0.00	2
A4	87027.46	76701.87	14951.02	14951.02	3
A5	92726.45	80302.87	3024.20	3024.20	3
A6	88887.36	74538.54	8735.01	8735.01	3
A7	93445.67	80233.07	7393.60	7393.60	3

At this stage, the cluster results are interpreted by providing information for each tofu alternative. Cluster interpretation is based on the final centroid value. As seen, C1 → Very popular, C2 → Popular, and C3 → Less popular. Therefore, the interpretation can be seen in Table 8 below.

Table 8. Cluster Interpretation

Alternative	Cluster	Region Name	Information
A1	1	Joman Water	Very Popular
A2	1	Sea Glare	Very Popular
A3	2	Sidokeno	Best Seller
A4	3	Sei Kamah	Less Popular
A5	3	Bandring Island	Less Popular
A6	3	East Range	Less Popular
A7	3	West Range	Less Popular

3.1.5. System Implementation

System implementation is the stage of translating the interface design and program logic into operational code. The system was built using PHP version 8, the C14 Framework, and MySQL or MariaDB database, then deployed on a local server for initial testing. The implementation produced several functional modules. The login interface serves as the primary security gateway, requiring users to enter their registered username and password before accessing the system. Once authenticated, users are directed to the dashboard, which displays a comprehensive summary of key information including the total number of alternatives (regions), registered attributes (product types), and system administrators. This two-tier interface ensures both security and quick access to operational data, allowing the business owner to monitor system status at a glance while maintaining control over who can access sensitive information.

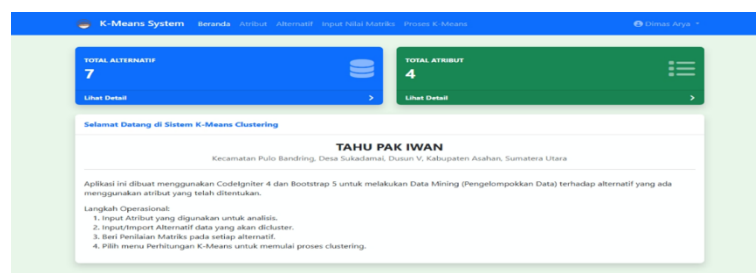


Figure 4. Home Page

The attribute management page provides administrators with full control over the system's data structure. Admins can add, edit, or delete attributes as needed to reflect changes in product types or sales criteria. Additionally, the alternative data management feature offers two flexible approaches: administrators can import bulk data directly from CSV files for efficiency, or manually enter data one by one using the dedicated "Add Data" function. This dual-method approach accommodates different operational workflows, whether handling large datasets or making incremental updates to the system.

Kode	Nama Alternatif	Keterangan	Cluster	Aksi
A01	Air Joman	Distribusi tahu wilayah Air Joman	C1	Tambah
A02	Silau Laut	Distribusi tahu wilayah Silau Laut	C1	Tambah
A03	Sidokeno	Distribusi tahu wilayah Sidokeno	C1	Tambah
A04	Sei Kamah	Distribusi tahu wilayah Sei Kamah	C1	Tambah
A05	Pulo Bandring	Distribusi tahu wilayah Pulo Bandring	C1	Tambah
A06	Kisaran Timur	Distribusi tahu wilayah Kisaran Timur	C1	Tambah
A07	Kisaran Barat	Distribusi tahu wilayah Kisaran Barat	C1	Tambah

Figure 5. Manage Alternative Data View

The matrix input page allows administrators to enter evaluation values for each alternative regional sales data. Each value corresponds to the attributes previously registered in the system, creating a structured dataset ready for analysis. Once the matrix is complete, the system transitions to the K-Means calculation interface, which displays the algorithmic processing in real-time. The calculation process begins by specifying the desired number of clusters and determining the initial centroids based on the input data. This sequential workflow ensures that users can clearly observe how raw sales data transforms into meaningful cluster assignments, making the technical process transparent and verifiable throughout each computational stage.

Alternatif / Atribut	A1	A2	A3	A4
A01 - Air Joman	91,200.000	42,753.000	88,700.000	52,481.000
A02 - Silau Laut	580,100.000	5,575.000	5,749,000.000	5,744,500.000
A03 - Sidokeno	855,100.000	9885.000	60,100.000	3,000.000
A04 - Sei Kamah	1,110,000.000	204,550.000	3,100.000	3,000.000
A05 - Pulo Bandring	5,400.000	1,415,000.000	2,000.000	184,100.000
A06 - Kisaran Timur	1,110,000.000	5,575.000	3,100.000	3,000.000
A07 - Kisaran Barat	5,400.000	7,000.000	2,000.000	243,000.000

Figure 6. Matrix Value Input Display

After the algorithm completes its iterations, the system displays the final cluster assignments alongside their business implications. Each region receives a classification Very Popular, Popular, or Less Popular—determined by the K-Means calculations. Rather than leaving numbers on a screen, the interface translates these results into actionable guidance: which areas deserve more frequent deliveries, where stock should be reduced, and which regions need strategic attention. This approach lets the owner move directly from analysis to decision-making without guesswork.

Hasil Clustering (K = 3)

Iterasi ke-1

Titik Pusat Centroid	A1	A2	A3	A4
C1	91,200.0000	42,753.0000	88,700.0000	52,481.0000
C2	85,510.0000	5,575.0000	6,010.0000	3,000.0000
C3	54,30.0000	7,000.0000	2,000.0000	2,431.0000

Iterasi ke-2

Titik Pusat Centroid	A1	A2	A3	A4
C1	74,070.0000	2,415.0000	3,100.0000	5,431.5.0000
C2	85,510.0000	5,575.0000	6,010.0000	3,000.0000
C3	85,510.0000	7,107.0000	2,000.0000	2,431.0000

Iterasi ke-3

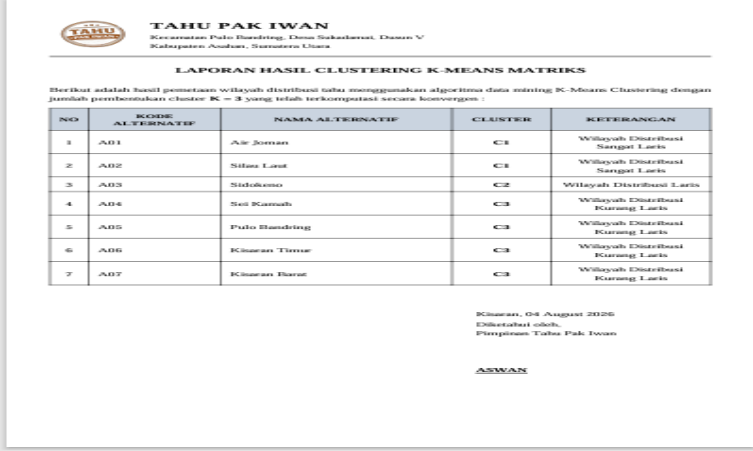
Titik Pusat Centroid	A1	A2	A3	A4
C1	74,070.0000	2,415.0000	3,100.0000	5,431.5.0000
C2	85,510.0000	5,575.0000	6,010.0000	3,000.0000
C3	85,510.0000	7,107.0000	2,000.0000	2,431.0000

Ringkasan Akhir Clustering

Cluster	Alternatif
Cluster C1 (Ditambah Distribusi Sangat Leri)	A01 - Air Joman
Cluster C2 (Ditambah Distribusi Leri)	A03 - Sidokeno
Cluster C3 (Ditambah Distribusi Sangat Leri)	A02 - Silau Laut
Cluster C4 (Ditambah Distribusi Sangat Leri)	A04 - Sei Kamah
Cluster C5 (Ditambah Distribusi Sangat Leri)	A05 - Pulo Bandring
Cluster C6 (Ditambah Distribusi Sangat Leri)	A06 - Kisaran Timur
Cluster C7 (Ditambah Distribusi Sangat Leri)	A07 - Kisaran Barat

Figure 7. Calculation Results Display

The results visualization page allows administrators to review the complete cluster assignments for all tofu distribution areas. The interface displays each region's classification alongside relevant sales metrics, making patterns immediately visible. Users can examine the data on-screen or generate printed reports for field distribution, meetings, or record-keeping purposes. This dual functionality ensures that cluster findings reach decision-makers in their preferred format, whether digital or hardcopy, supporting both operational efficiency and documentation requirements.



TAHU PAK IWAN
Kecamatan Pulau Bandring, Desa Subakharat, Dusun V
Kabupaten Andam, Sumatera Utara

LAPORAN HASIL CLUSTERING K-MEANS MATRIKS

Berikut adalah hasil pemetaan wilayah distribusi tahu menggunakan algoritma data mining K-Means Clustering dengan jumlah pembentukan cluster K = 3 yang telah terkomputasi secara konvergen :

NO	KODE ALTERNATIF	NAMA ALTERNATIF	CLUSTER	KETERANGAN
1	A01	Air Joman	C1	Wilayah Distribusi Sangat Laris
2	A02	Silau Lant	C1	Wilayah Distribusi Sangat Laris
3	A03	Sidokeno	C2	Wilayah Distribusi Laris
4	A04	Sei Kamah	C3	Wilayah Distribusi Kurang Laris
5	A05	Pulau Bandring	C3	Wilayah Distribusi Kurang Laris
6	A06	Kecamatan Teras	C3	Wilayah Distribusi Kurang Laris
7	A07	Kecamatan Barau	C3	Wilayah Distribusi Kurang Laris

Kecamatan, 04 August 2026
Disetujui oleh,
Pengelola Tahu Pak Iwan

ASWAN

Figure 8. Printed View of Calculation Results Report

3.2. Discussion

Patterns that remained hidden in Tahu Pak Iwan's distribution network now become visible after K-Means analysis. Two regions Joman Water and Sea Glare emerge as the business volume drivers, commanding the majority of sales across all product types. One region Sidokeno maintains steady, reliable demand. Four remaining regions show significantly lower sales volumes. This three-tier structure is not random categorization; it reflects how customers in different areas actually purchase tofu. Understanding this distinction matters because it reshapes how the business should operate. Similar findings have appeared in other product sales analyses, where K-Means clustering effectively separates regional performance and supports strategic decision-making (Assegaf et al., 2025; Almaripat et al., 2025).

Clustering results do more than organize data into neat boxes. They expose operational realities that conventional distribution practices have masked. Joman Water and Sea Glare are not simply "high-volume" regions in an abstract sense. These are areas where customers place large, consistent orders; where inventory moves quickly; where the business can confidently allocate greater production capacity. The fact that these two regions absorb a large share of total sales indicates they represent Tahu Pak Iwan's core market. Overlooking this concentration would be a strategic mistake. This sales segmentation approach aligns with established practices in data mining, where clustering algorithms identify distinct market segments based on purchasing patterns. Research on sales data classification shows that K-Means effectively separates customer groups with different purchase behaviors, allowing businesses to adjust their strategies accordingly (Wahyudi & Silfia, 2022; Sonatha et al., 2025). In Tahu Pak Iwan's case, clustering reveals that regional demand is not uniform a finding that should reshape how the business allocates resources and manages inventory. Conversely, the four regions grouped as "Less Popular" Sei Kamah, Bandring Island, East Range, and West Range show sales volumes roughly one-third to one-half of the very popular regions. This is not market failure; it reflects genuine differences in local demand, customer preferences, or market saturation. Treating these regions as having equal potential to Joman Water would produce poor resource allocation. Studies on product grouping and sales analysis confirm that recognizing such differences is essential for effective inventory management and distribution optimization (Putri et al., 2023; Hasanah & Humaidi, 2020). The middle cluster containing Sidokeno deserves separate attention. With sales volumes between the very popular and less popular tiers, Sidokeno represents a growth opportunity. This region shows enough demand to justify regular supply, yet its sales remain below the top tier. This positioning suggests Sidokeno could move into the very popular category if distribution improves, product quality increases, or local marketing efforts intensify. Unlike less popular regions, Sidokeno does not need a different strategy it needs better execution of the existing one.

Tahu Pak Iwan's conventional distribution approach has sent roughly equal quantities to all regions, adjusted only by the owner's memory of which areas "usually" order more. This approach fails on multiple fronts. First, it ignores actual regional demand differences. Second, it creates unnecessary risk: large shipments to low-demand areas increase spoilage risk, while small shipments to high-demand areas risk stockouts and lost sales. Third, it wastes management attention by treating all regions as equally important when clearly, they are not. The clustering results suggest a more rational approach grounded in data-driven marketing strategy development. Very popular regions like Joman Water and Sea Glare should receive priority in delivery scheduling. These areas warrant more frequent deliveries perhaps every two to three days to ensure fresh stock

and rapid turnover. Production planning should allocate a larger daily output share to these regions. Stock levels require close monitoring to prevent shortages that disappoint reliable customers. Research on sales segmentation and marketing strategy shows that such differentiated approaches, based on cluster analysis, significantly improve operational efficiency and customer satisfaction (Almaripat et al., 2025; Lathifah et al., 2026). The popular region (Sidokeno) should receive regular scheduled deliveries perhaps twice weekly with inventory levels adjusted based on observed demand patterns. Less popular regions could operate differently: perhaps weekly deliveries of smaller quantities, with closer attention to customer feedback about which products sell and which accumulate as excess stock. This differentiated approach recognizes that not all regions require identical service levels or management attention. Studies comparing clustering algorithms demonstrate that K-Means provides reliable segmentation foundation, though ongoing refinement may enhance results further (Siregar, 2024; Fitrianti Fahrudin & Rindiyani, 2024).

The most immediate benefit of this clustering is reduced product spoilage. Tofu is perishable with short shelf life. When wrong quantities reach wrong regions, spoilage becomes inevitable. Clustering results provide data-driven justification for adjusting quantities to match actual demand. By sending larger quantities to very popular regions and smaller quantities to less popular regions, the business can reduce total product sitting unsold long enough to spoil. For a business operating on thin margins, this matters significantly. Every spoiled tofu unit represents lost revenue and wasted production capacity. Data mining analysis of sales patterns consistently shows that clustering-based approaches reduce waste and improve inventory turnover across various product categories (Aldo, 2023; Nabila et al., 2021). Similarly, clustering prevents stockouts in high-demand regions. When Joman Water or Sea Glare runs short of inventory, customers may switch to competitors. The clustering results make clear these regions are too valuable to lose to stockouts. By prioritizing deliveries to these areas and maintaining higher safety stock levels, the business protects its most valuable customer relationships.

The web-based system successfully translates the K-Means algorithm into a practical tool business managers can actually use. The system accepts daily sales data, performs clustering calculations, and displays results in formats supporting decision-making. The ability to generate printed reports means cluster findings reach field staff and delivery personnel, not just the owner. This analysis-to-operations integration is crucial. A clustering result sitting only in a computer file does not improve distribution. A clustering result reaching people making daily delivery decisions drives real operational change. Implementation research on data mining systems demonstrates that successful applications require not just accurate algorithms, but also user-friendly interfaces and clear result communication to decision-makers (Siregar, 2018; Ramadhani et al., 2022). The system design also bridges a common gap between data analysis and business practice. Many analytical tools produce technically correct outputs that are practically useless because they are too complex, too abstract, or too disconnected from daily operations. This system avoids that trap by presenting cluster results in terms directly mapping to distribution decisions: which regions get priority, which products move quickly in which areas, and which regions might benefit from different approaches.

The current clustering uses three clusters based on sales volume alone. This is a reasonable starting point, but future refinements could strengthen the analysis. For example, clustering could incorporate additional factors such as product mix (which tofu types each region prefers), seasonality (whether demand fluctuates by season), or customer concentration (whether a few large customers dominate regional sales). Research on advanced clustering techniques suggests that hybrid approaches combining K-Means with other methods or incorporating multiple variables may yield more nuanced insights (Lu & Chang, 2014; Lathifah et al., 2026). The system could also track how cluster assignments change over time, revealing whether regions move between categories or whether the distribution network becomes more or less concentrated. Additionally, clustering results should be validated against qualitative observations from field staff. Do very popular regions actually receive the most frequent customer orders? Do less popular regions show the patterns data suggests? This cross-checking ensures quantitative analysis aligns with operational reality.

4. CONCLUSION

The research demonstrates that data-driven clustering outperforms intuition-based distribution strategies. By aggregating daily sales transactions into regional volume metrics, the pre-processing stage creates a foundation for objective analysis. The K-Means algorithm then processes multidimensional product data Hollow Tofu, Jumbo, Half-Cooked, and Raw to classify regions systematically rather than relying on owner assumptions. The analysis reveals three distinct market tiers across seven regions. Cluster 1 (Very Popular) identifies Air Joman and Silau Laut as high-performing areas deserving premium attention. Cluster 2 (Popular) flags Sidokeno as a moderate performer. Cluster 3 (Less Popular) groups Sei Kamah, Pulo Bandring, East Kisaran, and West Kisaran as underperforming zones requiring strategic intervention. Rather than leaving these numbers buried in spreadsheets, the dashboard translates findings into operational guidance. The system's practical value lies in enabling owners to allocate delivery frequency proportionally to actual demand, reduce inventory waste through demand-aligned stocking, optimize route efficiency, and minimize spoilage from over-supply in weak

markets. By replacing guesswork with measurable data, the owner gains a competitive advantage: resources flow where they generate returns, not where habit dictates.

REFERENCES

- Aldo, D. (2023). Data Mining Sales of Skin Care Products Using the K-Means Method. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, 7(1), 295-304. <https://doi.org/10.33395/sinkron.v8i1.12007>
- Almaripat, M., Faqih, A., & Rinaldy, A. R. (2025). Sales Data Classterization Analysis Using K-Means Method for Marketing Strategy Development. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 4(2), 972–976. <https://doi.org/10.59934/jaiea.v4i2.792>
- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34(5), 1785-1792. <https://doi.org/10.1016/j.jksuci.2019.12.011>
- Anwar, M. S., & Wulandari, S. (2025). Mobile Intelligent System for VARK-Based Student Learning Style Classification Using KNN. *bit-Tech*, 8(2), 2890-2899. <https://doi.org/10.32877/bt.v8i2.3428>
- Assegaf, N. N., Wadu, R. M. B., Oktaviano, M., & Tjahjanto, T. (2025, December). Analysis of Product Sales Patterns at PT Nitto Alam Indonesia Using K-Means Clustering. In *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* (pp. 1377-1382). IEEE. <http://doi.org/10.1109/ICIMCIS68501.2025.11327175>
- Bandyopadhyay, S., Thakur, S. S., & Mandal, J. K. (2021). Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society: S. Bandyopadhyay et al. *Innovations in Systems and Software Engineering*, 17(1), 45-52. <https://doi.org/10.1007/s11334-020-00372-5>
- Fitrianti Fahrudin, N., & Rindiyani, R. (2024). Comparison of K-Medoids and K-Means Algorithms in Segmenting Customers based on RFM Criteria. *E3S Web of Conferences*, 484, 02008. <https://doi.org/10.1051/e3sconf/202448402008>
- Harliana, H., Bhakti, R. M. H., Bachri, O. S., & Efendi, F. S. (2021). Optimasi K-Means dengan Particle Swarm Optimization pada Pengelompokkan Daerah Stunting. *Jurnal Ilmiah Intech: Information Technology Journal of UMUS*, 3(02), 95-101. <https://doi.org/10.46772/intech.v3i02.457>
- Hasanah, U., & Humaidi, B. (2020). Analysis of data mining using k-means clustering algorithm for product grouping. *International Journal of Informatics and Information Systems*, 3(1), 12-22. <https://doi.org/10.47738/ijiis.v3i1.3>
- Lathifah, S. N., Parama, S. R., Joan, R. K., Hidayat, S. K., & Darmawan, A. (2026). Analysis of sales patterns and product segmentation using k-means clustering and linear regression. *Machine Intelligence for Societal Advancement*, 1(3), 86-97.
- Lu, C. J., & Chang, C. C. (2014). A Hybrid Sales Forecasting Scheme by Combining Independent Component Analysis with K-Means Clustering and Support Vector Regression. *The Scientific World Journal*, 2014(1), 624017. <https://doi.org/10.1155/2014/624017>
- Maulani, G., Hasan, F. N., Setiawan, D., Bowo, I. T., Ardhana, V. Y. P., Ramdhani, Y., ... & Safitri, R. (2025). *Machine Learning*. Mega Press Nusantara.
- Miranda, M., & Sriani, S. (2025). Implementation of K-Means Clustering in Grouping Sales Data at Zura Mart. *Journal of Applied Informatics and Computing*, 9(2), 547-555.
- Mutiah, S., Hasnataeni, Y., Fitrianto, A., Erfiani, E., & Jumansyah, L. R. D. (2024). Perbandingan Metode Klastering K-Means dan DBSCAN dalam Identifikasi Kelompok Rumah Tangga Berdasarkan Fasilitas Sosial Ekonomi di Jawa Barat. *Teorema: Teori dan Riset Matematika*, 9(2), 247-260. <http://dx.doi.org/10.25157/teorema.v9i2.16290>

- Nabila, Z., Isnain, AR, Permata, P., & Abidin, Z. (2021). Data mining analysis for clustering of Covid-19 cases in Lampung Province with the k-means algorithm. *Journal of Technology and Information Systems*, 2(2), 100-108. <https://doi.org/10.33365/jtsi.v2i2.868>
- Nasyuha, A. H., Zulham, Z., & Rusydi, I. (2022). Implementation of K-means algorithm in data analysis. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 20(2), 307-313. <http://doi.org/10.12928/telkomnika.v20i2.21986>
- Ningtyas, M. W., & Pribadi, F. S. (2023). Soybean Collect Recommender Based on Distance and Productivity Cluster Using K-means Clustering and Simple Addictive Weighting Method. *Elinvo (Electronics, Informatics, and Vocational Education)*, 8(1), 86–95. <https://doi.org/10.21831/elinvo.v8i1.53208>
- Putri, N., Muhammad, A., & Mahessya, R. A. (2023). Grouping Building Material Sales for Management Optimization with the K-Means Algorithm. *Journal of Computer Scine and Information Technology*, 9(3), 119–124. <https://doi.org/10.35134/jcsitech.v9i3.74>
- Ramadhani, S., Azzahra, D., & Z, T. (2022). Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, 13(1), 24–33. <https://doi.org/10.31849/digitalzone.v13i1.9292>
- Rismayadi, AA, Fatonah, NN, & Junianto, E. (2021). K-Means Clustering Algorithm to Determine Marketing Strategy at CV. Construction Integrates. *Responsive Journal: Science and Informatics Research*, 3(1), 30-36. <https://doi.org/10.51977/jti.v3i1.393>
- Siregar, B. (2024). Comparative Analysis of K-Means and K-Medoids Algorithms for Product Sales Clustering and Customer. *Journal of Mathematics, Computations and Statistics*, 7(2), 360-370. <https://doi.org/10.35580/jmathcos.v7i2.4053>
- Siregar, MH (2018). Data Mining Clusterization of Building Equipment Sales Using the K-Means Method (Case Study at a Building Store). *Journal of Technology and Open Source*, 1(2), 83-91. <https://doi.org/10.36378/jtos.v1i2.24>
- Sonatha, Y., Erianda, A., & Fitri, R. (2025). Sales Segmentation Analysis of Tobacco Products Using the K-Means Clustering Method. *INVOTEK: Jurnal Inovasi Vokasional Dan Teknologi*, 24(2), 139–146. <https://doi.org/10.24036/invotek.v24i2.1221>
- Tegar, T., & Sari, C. M. (2025). The Digital Shift in The Mebel Market: How Marketing Strategies, Brand Equity, Location and Price Perception Influence Consumer Decisions. *Mabny: Journal of Sharia Management and Business*, 5(1), 50-64. <https://doi.org/10.19105/mabny.v5i1.18644>
- Untuk Rekomendasi Lokasi Strategis Umkm Kopi Keliling: Implementation of Weighted K-Means and Heatmap Overlay for Strategic Location Recommendation of Mobile Coffee Msmes. *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, 11(2), 3993-4005.
- Wahyudi, T., & Silfia, T. (2022). Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy In S&R Baby Store. *Journal of Applied Engineering and Technological Science (JAETS)*, 4(1), 93-103. <https://doi.org/10.37385/jaets.v4i1.913>
- Wibowo, A. E., & Habanabakize, T. (2022). K-Means Clustering untuk Klasifikasi Standar Kualifikasi Pendidikan dan Pengalaman Kerja Guru SMK di Indonesia. *Jurnal Dinamika Vokasional Teknik Mesin*, 7(2). <https://doi.org/10.21831/dinamika.v7i2.53848>
- Windarto, A. P. (2017). Penerapan datamining pada ekspor buah-buahan menurut negara tujuan menggunakan k-means clustering method. *Techno.Com*, 16(4), 348–357. <https://doi.org/10.33633/tc.v16i4.1447>
- Wu, L. Y., Liu, F. M., Lin, W. C., Qiu, J. T., Lin, H. Y., & Lin, K. F. (2024). Research on using K-Means clustering to explore high-risk products with ethylene oxide residues and their manufacturers in Taiwan. *Foods*, 13(16), 2510. <https://doi.org/10.3390/foods13162510>