3 (2) August 2025
Data Science Insights

Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Cluster Analysis of Superstore Data using K-Means and k-Medoids for Product Delivery Insights

Intan Chintia Sarumaha[1], Ajeng Cahyani Foureshtree[2], Angela Jocelyn[3], Jeffri Santoso[4], Fernando Hutabarat[5]

Information System, Pelita Indonesia Institute of Business and Technology, Pekanbaru City, Indonesia
email: [1]intan.chintia@student.pelitaindonesia.ac.id, [2]ajeng.cahyani@student.pelitaindonesia.ac.id, [3]angela.jocelyn@student.pelitaindonesia.ac.id,[4]jeffri.santoso@student.pelitaindonesia.ac.id, [5]fernando.hutabarat@student.pelitaindonesia.ac.id

** Correspondence*

A R T I C L E   I N F O

Article history:
Received August 18, 2024
Revised January 20,2025
Accepted July 05, 2025
Available online August 07, 2025

Keywords:
Clustering
K-Means
K-Medoids
Market Trends

Please cite this article in IEEE style as:
I. chintia Sarumaha, A. C. Foureshtree, A. Jocelyn, J. Santoso, and F. Hutabarat, "Cluster Analysis of Superstore Data using K-Means and K-Medoids for Product Delivery Insights", Data Science Insights, vol. 3, no. 2, pp. 83–93, Aug. 2025.

A B S T R A C T

It is difficult to overcome the challenge of understanding the relationship between consumer patterns and overall market trends and improve the company's operational efficiency through optimizing the delivery process. Utilizing sales data from Super Store available on the Kaggle website, this study aims to identify predictable consumer patterns using cluster analysis, as well as explore how to improve delivery efficiency based on a better understanding of consumer needs and preferences. This research utilizes K-Means and K-Medoids clustering methods to group product subcategories into three categories: best-selling, in-selling, and not-selling. The process of data transformation, exploratory analysis, model building, as well as cluster performance evaluation were conducted with the help of analytical tools such as Microsoft Excel, Tableau, and RapidMiner. The results show that the K-Medoids algorithm provides more accurate clustering performance compared to K-Means, with a Davies-Bouldin Index value of -0.867 for K-Medoids and -0.519 for K-Means. This shows that K-Medoids is more suitable in describing the characteristics of existing data. The most in-demand cluster results are in the sub-category of machines and copiers products.

Correspondence:
Intan Chintia Sarumaha
Department of Information System, Pelita Indonesia Institute of Business and Technology, Pekanbaru City, Indonesia

1. Introduction

One important aspect of achieving a sound business strategy is understanding consumer behavior and needs. Changing purchasing patterns and evolving preferences make understanding consumer patterns key to forecasting market trends and identifying potential growth opportunities. Data is one of the most valuable assets for companies to identify opportunities and face challenges in a changing business environment [1].

Analysis of consumer patterns from sales data can provide valuable insights into consumer behavior. However, there is still a need to understand the relationship more deeply between consumer patterns and overall market trends. On the other hand, increasing competition and pressure to improve operational efficiency are driving companies to optimize delivery processes based on enhanced consumer information.

This research aims to investigate the relationship between consumer patterns and market trends. By utilizing the available sales data, this study will employ cluster analysis to identify predictable consumer patterns and leverage them to forecast future market trends. The clustering method or analysis is utilized to group objects based on observed indicators, ensuring that objects within the same group are highly similar compared to those in different groups. Different groups [2]. Additionally, this research will also explore ways to enhance delivery efficiency by gaining a deeper understanding of consumer needs and preferences.

The results of this research will provide valuable insights for companies in designing more effective and efficient sales strategies, increasing customer satisfaction, and increasing company profitability.

2. Research Methods

Research methods are steps taken by researchers to collect research data. This data is processed and analyzed scientifically to achieve research objectives. The steps in this research include data collection techniques, data transformation, exploratory data analysis, model building, performance clustering, and clustering results as shown in Figure 1.

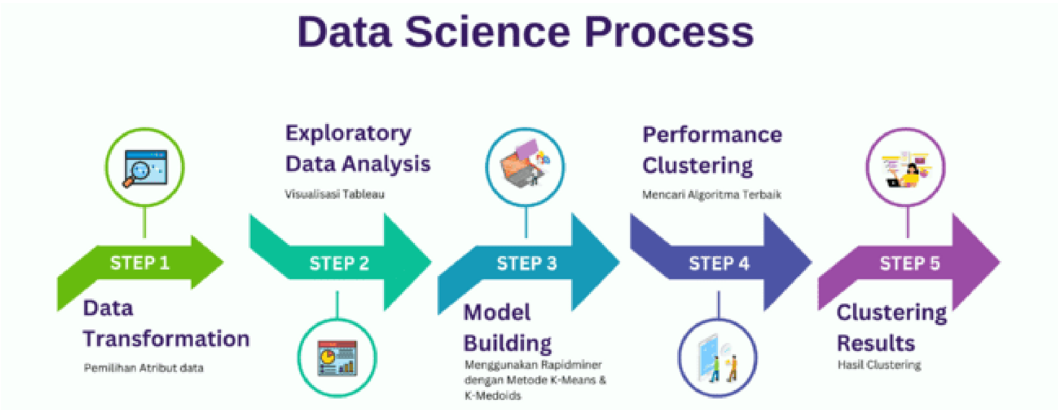

Figure 1. Data Science Process

2.1 Data Collection Technique

Data for this analysis was obtained through data download techniques from the Kaggle website (https://www.kaggle.com/datasets/laibaanwer/superstore-sales-dataset) [5]. The dataset used is the SuperStore Sales Dataset which was downloaded on April 25, 2024 at 12:00 WIB. This dataset has version 1.0 and was chosen because it contains store sales data that can be used to analyze sales trends and consumer behavior. This data will later be used to build a machine learning model to predict future sales.

2.2 Data Transformation

Data transformation is used to change data in a form that is suitable for the data mining process. Data transformation is used to change the attribute data type from nominal to numerical data type with the Nominal to Numerical tool to match the required data type [6]. Some techniques for data transformation are normalization, attribute selection, and discretization. Data normalization technique is a technique used in data mining to convert the value of a data set into a common scale [7]. Feature selection aims to select the most relevant subset of attributes from the original data. This technique reduces the number of features involved in determining a target class value [8]. Discretization technique refers to the process of transforming continuous data into a set of data intervals [9]. Discretization can help in identifying patterns that are not visible in the original data. This whole process improves data quality so that the analysis or model built becomes more accurate and efficient.

The data transformation used in this research is Attribute selection. Attribute selection is the process of selecting the most relevant and informative subset of attributes from a larger data set. The goal is to reduce the dimensionality of the data and improve the performance of the machine learning model.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an analytical approach to data to create a summary of the data so that it is easy to understand [10]. Researchers can apply various statistical and visualization techniques to uncover hidden patterns, trends, and relationships in the data. In this phase, it is very important to remain critical and not immediately draw conclusions from what is seen. Researchers should consider various possibilities and look for relevant evidence according to their research objectives. Hasty analysis can lead to erroneous conclusions. Therefore, each finding should be carefully examined and validated.

In this process, the authors visualized data from the super store to find out various useful knowledge to better understand market trends. We used Tableau to facilitate the completion of this research. This tool makes it easy for researchers to process and analyze data effectively. Using Tableau as a data visualization tool has several significant advantages. Among them is the high flexibility in processing and visualizing data from various sources.

2.4 Model Building

The process of building this model requires deep thought and understanding of the data and algorithms to be used. In this study, researchers will use RapidMiner as software as well as the K-Means and K-Medoids algorithms for data analysis. In the context of the K-means algorithm, clusters usually refer to groups of data that are similar or like each other. K-means is one of the most used clustering methods in data analysis and machine learning. It works by randomly defining a cluster center (centroid), then calculating the distance between each data and the centroid. Each data will be grouped into the cluster that has the closest centroid. After all data is grouped, a new centroid will be recalculated based on the average of the data in one cluster [11].

Besides using the K-means algorithm, this research also compares with other algorithms such as K-Medoids. K-Medoids is like K-means, but uses actual data points as cluster centers, making it more robust to outliers. This research sets three groups, namely most in-demand, in-demand, and not in-demand. The determination of these groups aims to categorize products based on sales levels. By comparing the two algorithms, it is expected to find the most effective clustering method for this sales analysis.

2.5 Performance Clustering

Based on the results of the implementation of several algorithms, performance results are obtained which aim to compare which algorithm is more accurate later. This performance evaluation is carried out using metrics such as the Davies-Bouldin Index for both algorithms, namely K-means and K-Medoid. The Davies-Bouldin Index metric is defined as the average ratio of within- and between-cluster distances for each cluster to its nearest neighbor cluster [12]. By using the same metric for both algorithms, it allows for a more fair and consistent comparison in evaluating the quality of cluster.

It can be seen later whether there is a significant difference in the way the two algorithms cluster the data. The best performance is determined not only by the value of the metric but also by the ease of interpretation of the clustering results. Thus, a final decision on which algorithm is more accurate and effective can be made by considering various evaluation factors.

2.6 Clustering Results

This stage is the last stage where the results of the clustering process will be obtained. The clustering results will group the products into three categories: best-selling, in-selling, and not-selling. With this information, the super store company can know which product subcategories have the highest sales or are most in demand. This helps the company to identify ongoing market trends. This knowledge is invaluable for making more informed business decisions.

Market trends inform us about what products need to be developed for customers, cost how much, distribution should be done through which channels, thus reducing uncertainty or business losses while in the process of developing new products or services [14]. By knowing market trends, companies can focus their efforts on products that are most in demand by customers. The best-selling product subcategories can get more attention in terms of promotion and inventory. Meanwhile, the less-selling products can be re-evaluated for improvements or additional promotions. Thus, the company can improve its overall sales performance.

3. Result and Discussion

This section will explain the results of the research method. starting from the results of data collection, data transformation, exploratory data analysis, model building, performance clustering, and clustering results.

3.1 Data Transformation Result

In this study, we used feature selection techniques to choose the most relevant subset of attributes from the original data, reducing data dimensions and enhancing model efficiency. Table 1 shows the Superstore data before attribute selection.

	A	B	C	D	E	F	G	H	I	J
1	order_id	order_date	ship_date	ship_mode	customer_name	segment	state	country	market	region
2	AG-2011-2040	1/1/2011	6/1/2011	Standard Class	Toby Braunhardt	Consumer	Constantine	Algeria	Africa	Africa
3	IN-2011-47883	1/1/2011	8/1/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania
4	HU-2011-1220	1/1/2011	5/1/2011	Second Class	Annie Thurman	Consumer	Budapest	Hungary	EMEA	EMEA
5	IT-2011-3647632	1/1/2011	5/1/2011	Second Class	Eugene Moren	Home Office	Stockholm	Sweden	EU	North
6	IN-2011-47883	1/1/2011	8/1/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania
7	IN-2011-47883	1/1/2011	8/1/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania
8	CA-2011-1510	2/1/2011	6/1/2011	Standard Class	Magdelene Morse	Consumer	Ontario	Canada	Canada	Canada
9	IN-2011-79397	3/1/2011	3/1/2011	Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania
10	ID-2011-80230	3/1/2011	9/1/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania
11	IZ-2011-4680	3/1/2011	7/1/2011	Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA
12	IN-2011-65159	3/1/2011	7/1/2011	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia
13	IN-2011-65159	3/1/2011	7/1/2011	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia
14	ES-2011-4869686	3/1/2011	7/1/2011	Standard Class	Dorothy Dickinson	Consumer	England	United Kingdom	EU	North
15	IN-2011-33652	3/1/2011	9/1/2011	Standard Class	Dennis Pardue	Home Office	Sarawak	Malaysia	APAC	Southeast Asia
16	ID-2011-80230	3/1/2011	9/1/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania
17	MX-2011-160234	3/1/2011	7/1/2011	Standard Class	Stewart Visinsky	Consumer	Guatemala	Guatemala	LATAM	Central
18	IR-2011-770	3/1/2011	7/1/2011	Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA
19	ID-2011-80230	3/1/2011	9/1/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania
20	ID-2011-80230	3/1/2011	9/1/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania
21	ID-2011-12596	3/1/2011	8/1/2011	Standard Class	Chris McAfee	Consumer	Nakhon Ratchasima	Thailand	APAC	Southeast Asia

Table 1. Super Store Data

	K	L	M	N	O	P	Q	R	S	T	U
1	product_id	category	sub_category	product_name	sales	quantity	discount	profit	shipping_cost	order_priority	year
2	OFF-TEN-10000025	Office Supplies	Storage	Tenex Lockers, Blue	408	2	0	106,14	35,46	Medium	2011
3	OFF-SU-10000618	Office Supplies	Supplies	Acme Trimmer, High Speed	120	3	0,1	36,036	9,72	Medium	2011
4	OFF-TEN-10001585	Office Supplies	Storage	Tenex Box, Single Width	66	4	0	29,64	8,17	High	2011
5	OFF-PA-10001492	Office Supplies	Paper	Enermax Note Cards, Premium	45	3	0,5	-26,055	4,82	High	2011
6	FUR-FU-10003447	Furniture	Furnishings	Eldon Light Bulb, Duo Pack	114	5	0,1	37,77	4,7	Medium	2011
7	OFF-PA-10001968	Office Supplies	Paper	Eaton Computer Printout Paper, 8.5 x 11	55	2	0,1	15,342	1,8	Medium	2011
8	TEC-OKI-10002750	Technology	Machines	Okidata Inkjet, Wireless	314	1	0	3,12	24,1	Medium	2011
9	OFF-AP-10000304	Office Supplies	Appliances	Hoover Microwave, White	276	1	0,1	110,412	125,32	Critical	2011
10	TEC-CO-10004182	Technology	Copiers	Hewlett Wireless Fax, Laser	912	4	0,4	-319,464	107,1	Low	2011
11	FUR-NOV-10002791	Furniture	Chairs	Novimex Swivel Stool, Set of Two	667	4	0	253,32	81,26	High	2011
12	OFF-ST-10003020	Office Supplies	Storage	Tenex Lockers, Industrial	338	3	0,45	-122,8005	33,75	High	2011
13	FUR-TA-10002797	Furniture	Tables	Chromcraft Round Table, Adjustable Height	211	1	0,55	-70,3995	21,32	High	2011
14	FUR-BO-10000728	Furniture	Bookcases	Dania Corner Shelving, Traditional	854	7	0	290,43	12,56	Medium	2011
15	TEC-CO-10000594	Technology	Copiers	Hewlett Fax and Copier, Laser	193	1	0	50,13	10,4	Medium	2011
16	FUR-CH-10000214	Furniture	Chairs	Hon Rocking Chair, Set of Two	159	2	0,4	-95,676	10,07	Low	2011
17	TEC-PH-10002647	Technology	Phones	Nokia Headset, VoIP	195	4	0	44,88	8,43	Medium	2011
18	OFF-BRE-10003081	Office Supplies	Appliances	Breville Coffee Grinder, Black	123	2	0	42,9	8,41	Medium	2011
19	TEC-AC-10002881	Technology	Accessories	Belkin Numeric Keypad, Bluetooth	69	2	0,4	3,42	8,34	Low	2011
20	FUR-CH-10000666	Furniture	Chairs	SAFCO Chairmat, Black	69	2	0,4	-26,412	8,17	Low	2011
21	OFF-ST-10002066	Office Supplies	Storage	Smead File Cart, Blue	135	2	0,47	-45,9018	7,74	Medium	2011

Table 1. Continuation of Super Store Data

For this study, we used only two data attributes: sub_category and sales. Figure 2 illustrates the data after attribute selection.

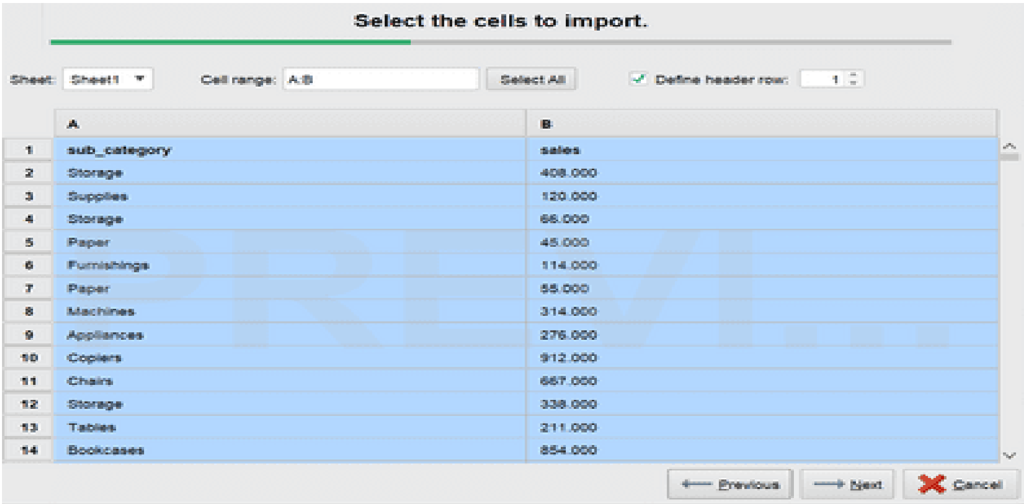

Figure 2. Data after attribute selection.

3.2 Exploratory Data Analysis Result

The exploratory data analysis results will be visualized using Tableau. This aims to uncover market trends and is also useful for analyzing the most effective shipping methods for Superstore. Below are some of the visualizations.

3.2.1 Sales Profitability by Product Category

This section visualizes sales profitability based on product subcategories. It utilizes the attributes sub_category, year, and profitability.

Figure 3

Figure 3. Continuation

Figure 3 shows that the highest profitability of the company is from the sale of copier products in 2014, with a profit of 104,049 USD. Conversely, the company experienced a loss in 2014 with the sub_category of tables, amounting to -30,546 USD. This graph provides information and warnings to the company to analyze the causes of these losses, such as high production costs, non-market-trend-following designs in the tables sub_category, leading to a lack of consumer interest, and also suggests collecting feedback from consumers

3.2.2 Average Discount on Various Product Sub-Categories

This section visualizes the average discount in each product subcategory using the attributes sub_category and discount (average). Knowing the average discount helps the company assess its pricing and discount strategies. For instance, an average discount of 11.90% for accessories might be considered too high or low, allowing the company to adjust its discount strategy accordingly.

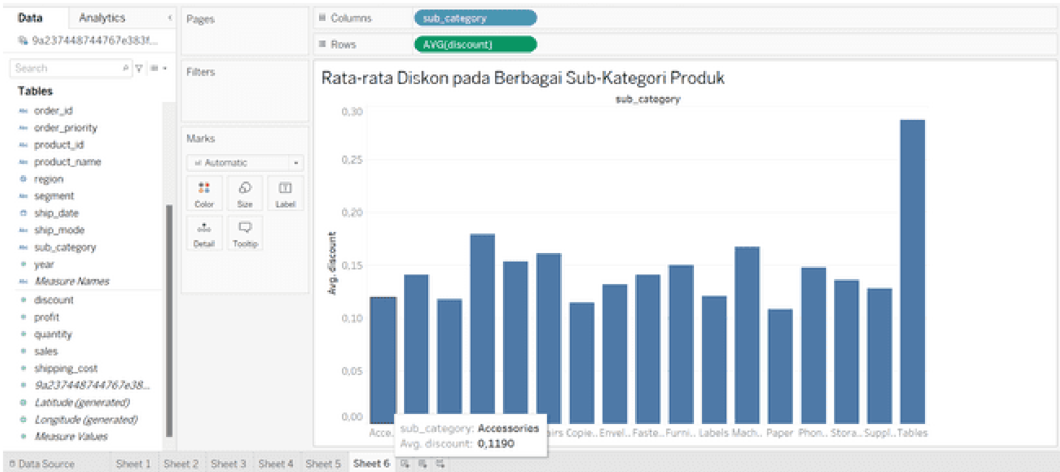

Figure 4

Figure 4 shows that the subcategory of accessories typically gets an average discount of 0.1190, meaning customers usually receive an 11.90% price reduction from the original price.

Higher or lower discounts can affect sales volume. The average discount impacts profit margins. If the discount is too high, profit margins may decrease. The company must ensure that the discounts offered remain profitable.

Compared to the average discount on the tables subcategory, which is 28.77%, there is a significant difference. The following points should be noted, and actions should be taken based on this comparison:

a. The tables sub-category has a much higher average discount (28.77%) compared to the accessories sub-category (11.90%).
b. Profitability: Higher discounts on tables products may reduce profit margins more than accessories.
c. Product Demand: High discounts might be necessary to boost or maintain tables product sales, indicating that demand for these products is more price-sensitive.

d. Competition: High discounts could also signify tougher competition in the market for tables products, forcing

Actions that can be taken include:

e. Reducing Discounts: If analysis shows that lower discounts can still maintain or increase sales, consider reducing discounts on tables products.

f. Discount Variation: Experiment with discount variations over specific periods or for certain customer segments to find the optimal discount that maximizes profits.

3.2.3 Sales Data by Region and Customer Segment

This section visualizes sales data by region and customer segment using the attributes region, segment, and sales.

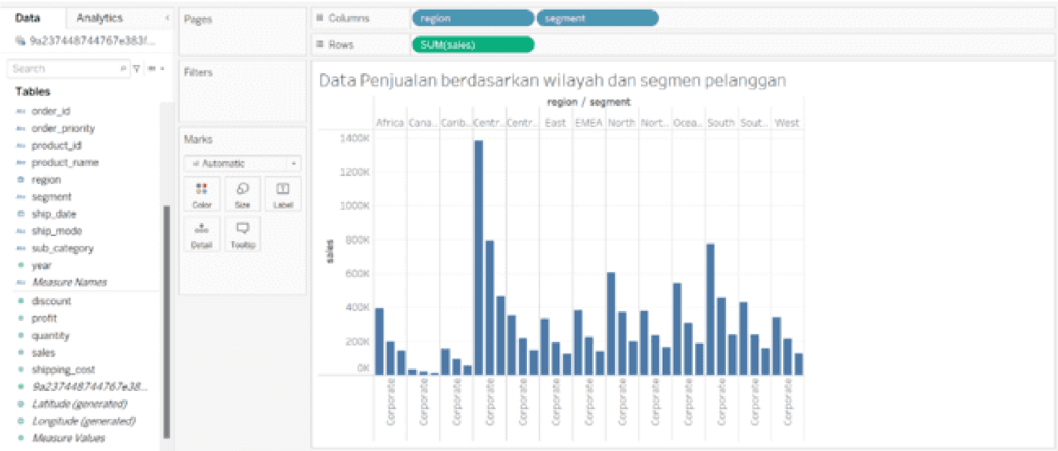

Figure 5

Figure 5 provides valuable information related to market trends. We can see that the highest sales are in the Central region in the consumer segment, with sales amounting to 1,381,649 USD. The company should focus more on marketing and distribution strategies in this region to further maximize sales. Marketing strategies may include product adjustments, promotions, and services to better meet the needs of the dominant customer segment.

3.2.4 Distribution of the Most Effective Shipping Methods

This section visualizes which shipping methods are most effective and frequently used by customers, using the attributes ship_mode and quantity.

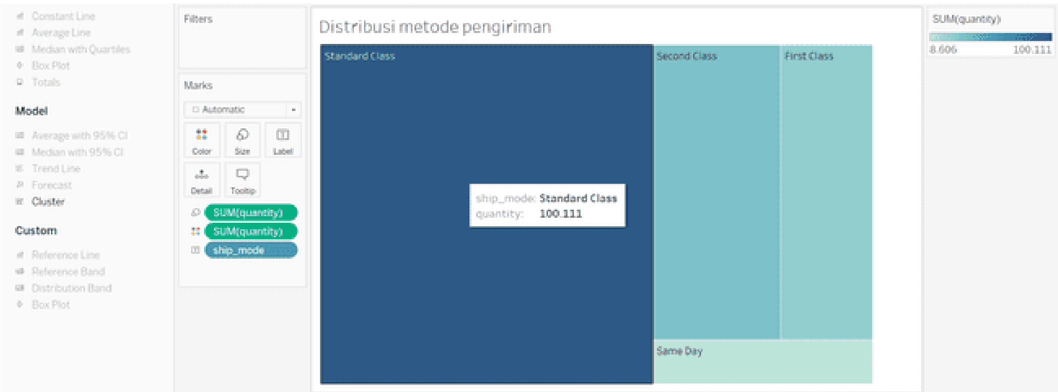

Figure 6

Figure 6 provides valuable information about the most effective shipping methods. We can see that the Standard Class method is preferred by customers, with a quantity of 100,111 shipments, whereas the Same Day shipping method is rarely used. The company can take several strategic actions. First, it can enhance the efficiency and capacity of the Standard Class service to ensure customer satisfaction remains high. Second, the company could promote or offer incentives for less popular shipping methods, such as

Same Day, to encourage their usage. Third, the company can reevaluate the costs and benefits of each shipping method to optimize resources and improve profitability.

3.3 Model Building Result

After reviewing the visualization result from Tableau, proceed to data modeling using the K-Means and K-Medoids algorithms. Figures 7 and 8 illustrate that the operator is utilized to read the dataset in this study. Subsequently, apply K-Means and K-Medoids clustering to model the existing dataset. Cluster distance performance is then used to evaluate the best clustering result. This study categorize the data into 3 groups: in demand, in demand, and not in demand.

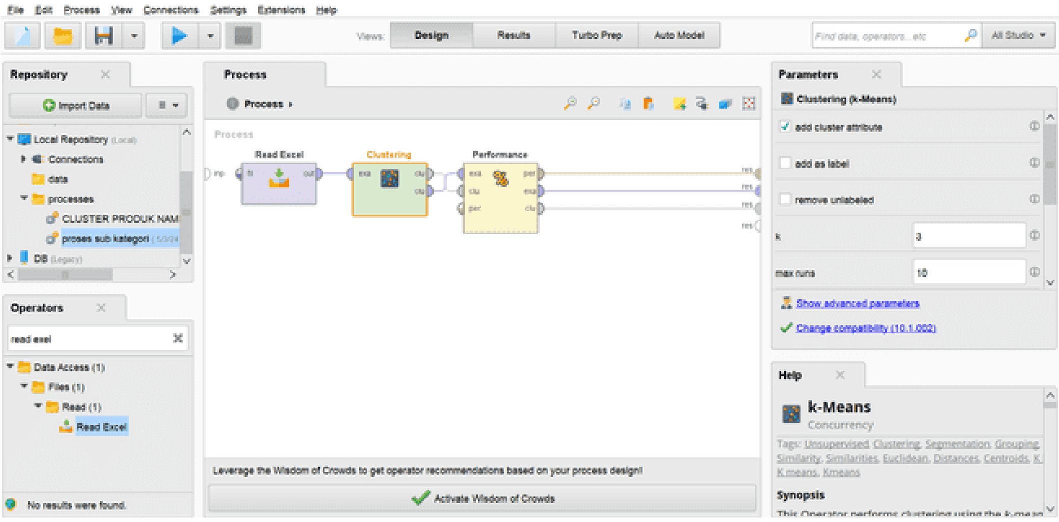

Figure 7. K-Means Algorithm Model Using RapidMiner

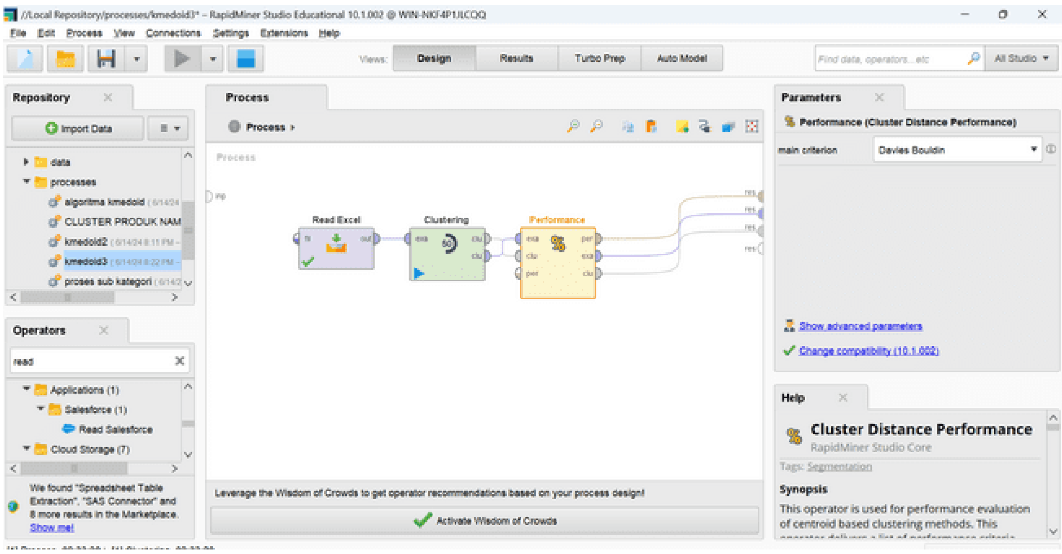

Figure 8. K-Medoids Algorithm Model Using RapidMiner

3.4 Clustering Performance Results

Based on the result of the implementing of the K-Means algorithm in RapidMiner, the following performance (Figure 9) results were obtained:

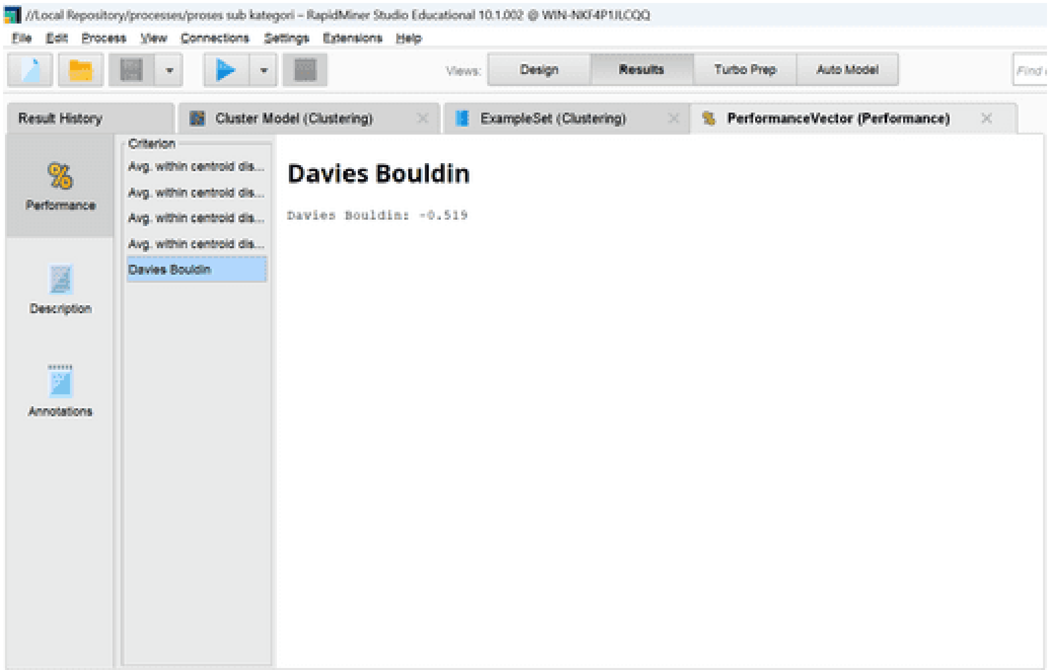

Figure 9. K-Means Performance Data

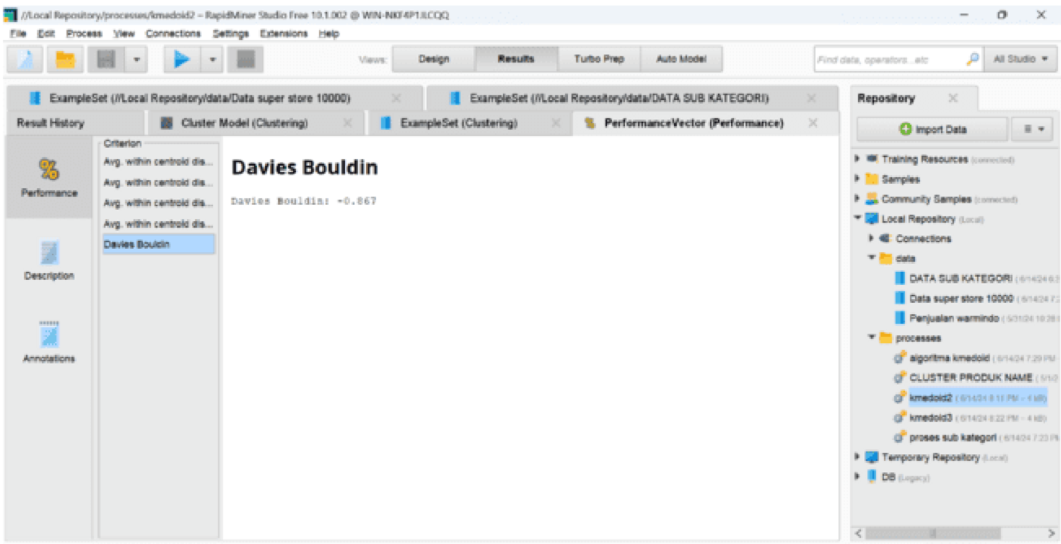

Figure 10. K-Medoids Performance Data

Based on Figure 9, which present performance data using the K-Means algorithm, it is explained that the performance result of the Davies-Bouldin Index (DBI) is -0,519. In Figure 10, using the K-Medoids algorithm, the performance result is -0,867. Therefore, it can be concluded that the clustering performed by the K-Means algorithm is superior to K-Medoids. K-Means demonstrates more accurate and better performance compared to K-Medoids because, according to the principle of the Davies-Bouldin Index (DBI), the ideal value in data mining is to be smaller or closer to zero [15]. Thus, it indicates better clustering quality.

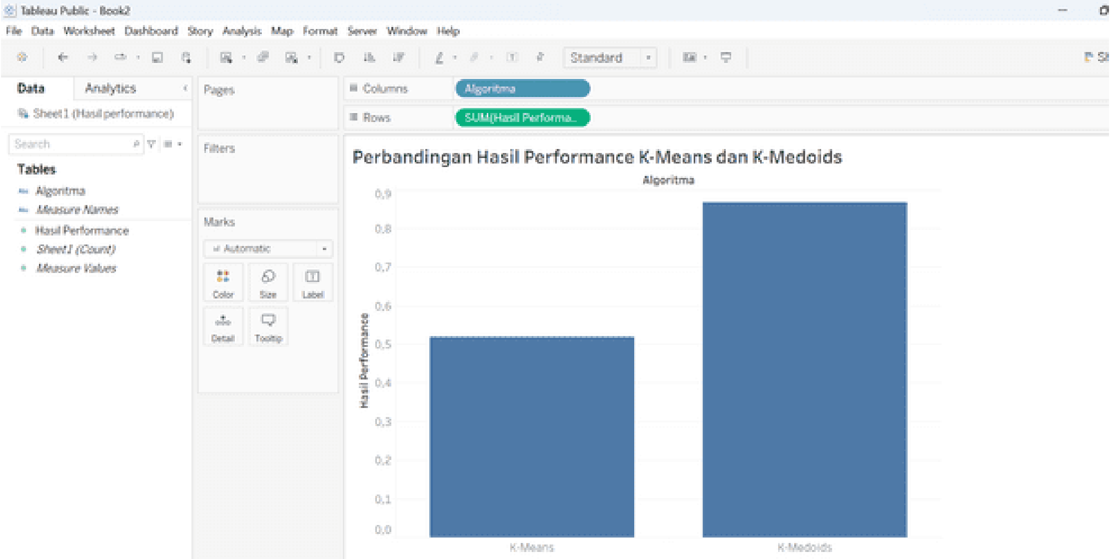

Figure 11. Comparison of K-Means and K-Medoids Performance Results

It can also be seen from Figure 11 that the Tableau visualization result show that the bar chart of K-Means performance result is closer to zero or smaller in value than the bar chart of K-Medoids performance results.

3.5 Clustering Results

Based on the results of the K-Means performance value which is more accurate and faster in clustering. Resulting in Cluster 0 with 45136 items categorized as Unsold, then Cluster 2 with 5465 items categorized as Sold, and finally Cluster 1 with 689 items categorized as Most Sold, and the last. Can be seen from Figure 12 below.

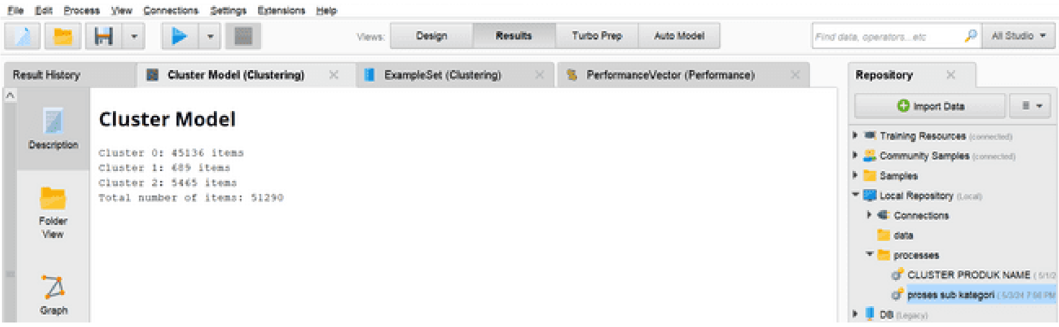

Figure 12. Cluster Model

Furthermore, to identity the subcategories of products that fall under three clusters mentioned earlier, it can be done through visualization using RapidMiner (Figure 13). This platform enables in-depth analysis of cluster data, allowing companies can direct marketing and inventory management strategies with more precision, ensuring that each product sub_category receives the attention it deserves according to market potential and customer demand.

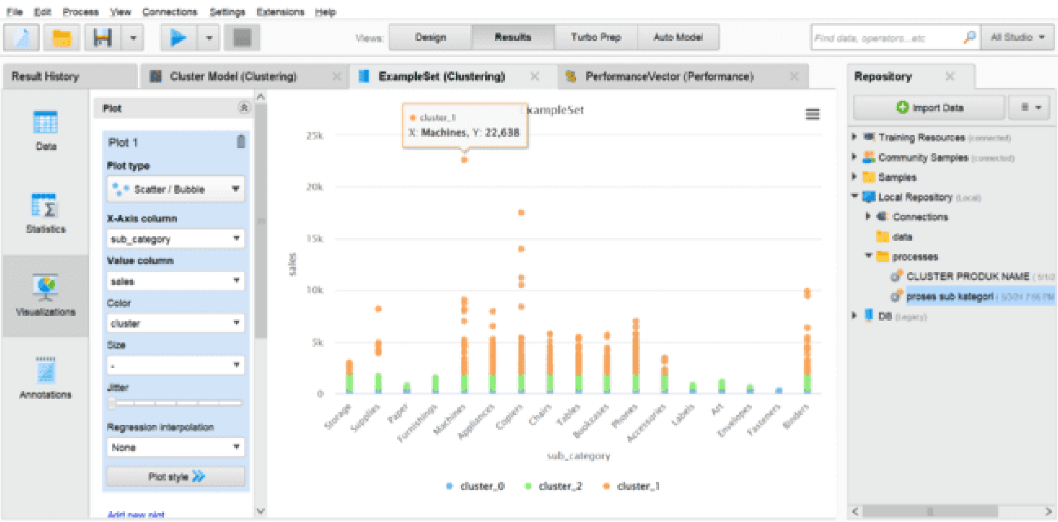

Figure 13. RapidMiner Visualization

From the visualization results using RapidMiner above, it can be seen in Table 2, that the subcategory of machine products dominates in cluster (1), which is the best-selling category, with sales reaching 22,368 USD. In the second position, the copiers product subcategory recorded sales of 17,500 USD. In cluster (2), which is also a hot-selling category, there are subcategories of phones and tablets with sales of USD 2,022 and USD 2,018 respectively. Meanwhile, in cluster (0), which is a non-selling category, the binders and fasteners subcategory shows very low sales of only 6 USD. This information provides a clear view of the sales performance of each product subcategory, enabling the company to take strategic steps in managing inventory and directing marketing efforts more effectively according to the identified market conditions.

No.	Paling Laris Cluster 1	Sales	Laris Cluster 2	Sales	Tidak Laris Cluster 0	Sales
1.	Machines	22.368	Phones	2.022	Binders	6
2.	Copiers	17.500	Tables	2.018	Fasteners	6

Table 2. Clustering Results

4. Conclusion

The conclusion of the report "Cluster Analysis Using K-Means, K-Medoids, and Product Shipment Analysis Methods on Super Store Data" is that this research successfully identified market trends in a Super Store Company through clustering products into three categories: best-selling, in-selling, and not-selling. Using data from the Kaggle website and analysis tools such as Microsoft Excel, Tableau, and RapidMiner, this study found that the K-Means algorithm provides more accurate clustering performance than K-Medoids. This finding shows that the performance value of K-Means is more accurate at -0.519. In terms of time, K-Means is also faster in clustering than K-Medoids. K-Means is more suitable for the characteristics of the data, so this information can be used by companies to optimize marketing strategies and inventory management, focusing on products with the best sales performance. This conclusion emphasizes the importance of structured data analysis in supporting better and timely business decision making.

References

[1] W. Aprilyani *et al.*, "Klasterisasi Data Penjualan Alat Transportasi dengan Rapidminer menggunakan Metode K-Medoid," vol. 8, no. 2, pp. 1348–1353, 2024.

[2] H. Syahputra, "Clustering Tingkat Penjualan Menu (Food and Beverage) menggunakan Algoritma K-Means," *J. KomtekInfo*, vol. 9, pp. 29–33, 2022, doi: 10.35134/komtekinfo.v9i1.274.

[3] A. Nugraha, O. Nurdiawan, and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering untuk Analisa Penjualan pada Toko Yana Sport," *Jati (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 849–855, 2022, doi: 10.36040/jati.v6i2.5755.

[4] M. N. Setiawan, Purwono, and I. A. Ashari, "Terakreditasi Sinta Peringkat 4 Analisa Cluster Data Transaksi Penjualan Minimarket selama Pandemi Covid-19 dengan Algoritma K-means," vol. 3, no. 1, pp. 153–160, 2018.

[5] R. Gustrianda and D. I. Mulyana, "Penerapan Data Mining dalam Pemilihan Produk Unggulan dengan

Metode Algoritma K-Means dan K-Medoids," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 27, 2022, doi: 10.30865/mib.v6i1.3294.
[6] L. Anwer, "SuperStore Sales Dataset," www.kaggle.com. Accessed: Jun. 25, 2024. [Online]. Available: https://www.kaggle.com/datasets/laibaanwer/superstore-sales-dataset
[7] N. Suarna and Y. Arie Wijaya, "Analisa Penerapan Metode Clustering K-Means untuk Pengelompokan Data Transakasi Konsumen (Studi Kasus: Cv. Mitra Indexindo Pratama)," *J. Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1322–1328, 2023.
[8] geeksforgeeks, "Normalisasi Data dalam Data Mining," www.geeksforgeeks.org. Accessed: Jun. 25, 2024. [Online]. Available: https://www.geeksforgeeks.org/data-normalization-in-data-mining/
[9] Trivusi, "Metode-Metode dalam Feature Selection," www.trivusi.web.id. Accessed: Jun. 25, 2024. [Online]. Available: https://www.trivusi.web.id/2019/02/metode-metode-dalam-feature-selection.html
[10] D. Vogiatzis, "Panduan Anda tentang Teknik Transformasi Data," blog.coupler.io. Accessed: Jun. 25, 2024. [Online]. Available: https://blog.coupler.io/data-transformation-techniques/
[11] Y. Ws, "Mengenal Exploratory Data Analysis," eksplorasidata.mipa.ugm.ac.id. Accessed: Jun. 25, 2024. [Online]. Available: https://eksplorasidata.mipa.ugm.ac.id/2021/08/16/mengenal-exploratory-data-analysis/
[12] R. Maoulana, B. Irawan, and A. Bahtiar, "Data Mining dalam Konteks Transaksi Penjualan Hijab dengan menggunakan Algoritma Clustering K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 515–521, 2024, doi: 10.36040/jati.v8i1.8504.
[13] Y. Sopyan, A. D. Lesmana, and C. Juliane, "Analisis Algoritma K-Means dan Davies Bouldin Index dalam mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1464–1470, 2022, doi: 10.47065/bits.v4i3.2697.
[14] Youtap, "Cara Jitu mengetahui Produk Paling Laris Sesuai Tren Pasar," www.youtap.id. Accessed: Jun. 25, 2024. [Online]. Available: https://www.youtap.id/blog/cara-mengetahui-produk-paling-laris
[15] N. Ameliana, N. Suarna, and W. Prihartono, "Analisis Data Mining pengelompokkan Umkm menggunakan Algoritma K-Means Clustering di Provinsi Jawa Barat," vol. 8, no. 3, pp. 3261–3268, 2024.