MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

Analysis of sales patterns and product segmentation using k-means clustering and linear regression

Sania Nur Lathifah [a,1*], Satria Ridha Parama [a,2], Robin Keanu Joan [a,3], Syams Kurniawan Hidayat [a,4], Arif Darmawan [b,5]

[a] STMIK AMIKOM Surakarta, Kartasura, Surakarta, Postal Code 57163, Indonesia
[b] Universitas Islam Negeri Syarif Hidayatullah Jakarta, Postal Code 15412, Indonesia
[1]sania.130525@mhs.amikomsolo.ac.id; [2] satria.130531@mhs.amikomsolo.ac.id; [3] robin.130529@mhs.amikomsolo.ac.id;
[4] Syamskh@dosen.amikomsolo.ac.id,[5]arif.darmawan22@mhs.uinjkt.ac.id

Article Info

Article History:

Received: Jun 01, 2026
Revised: Jun 11, 2026
Accepted: Jul 23, 2026

Keywords :
Data Mining; K-Means Clustering; Linear Regression; Product Segmentation; Sales Prediction

Abstract

Sales transaction data contain valuable information that can support business decision-making, including product segmentation and sales forecasting. However, many organizations still utilize transaction data solely for operational purposes, limiting its potential to generate strategic business insights. This study proposes an integrated framework combining K-Means Clustering and Linear Regression to analyze product sales patterns and predict future sales trends. A quantitative experimental approach was employed using a publicly available sales dataset obtained from Kaggle. Data preprocessing involved data cleaning, feature selection, and Min-Max normalization to ensure consistent feature scales prior to the clustering process. The optimal number of clusters was determined using the Elbow Method, resulting in four product clusters representing distinct sales characteristics. Subsequently, an independent Linear Regression model was developed for each cluster using an 80:20 train-test split to predict sales trends. Model performance was evaluated using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The experimental results demonstrate that the proposed framework successfully grouped products into four meaningful clusters and achieved satisfactory prediction performance, with the highest R^2 value reaching 0.9215, indicating a strong relationship between historical sales patterns and predicted values within the corresponding cluster. The integration of clustering and regression provides more comprehensive analytical insights than applying prediction directly to the entire dataset, facilitating a better understanding of product characteristics and sales behavior. Therefore, the proposed approach can support data-driven decision-making in inventory management, product prioritization, and marketing strategy development. Future research is recommended to evaluate the proposed framework using larger datasets and to compare its performance with other clustering and regression algorithms.

1. Introduction

The development of information technology and business digitalization has increased the volume of sales transaction data generated by various trade sectors. The transaction data stores important information about sales patterns, product characteristics, and consumer behavior that can be used as a basis for business decision-making. However, in practice, most business actors still use sales data only as an operational archive, so the potential information contained in it has not been utilized optimally. As a result, companies often have difficulty identifying sales patterns, determining the right marketing strategies, managing inventory, and predicting future sales trends. This condition shows the importance of

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

applying data analysis techniques that are able to transform transaction data into valuable information to support more objective and data-based decision-making[1] [2]

One of the widely used approaches to analyzing sales data is data mining and machine learning, which are capable of extracting hidden patterns from large data sets. One of the algorithms that is widely applied is K-Means Clustering, which is an unsupervised learning method that groups data based on the level of similarity of characteristics without the need for class labels. Various studies have shown that K-Means Clustering is effectively used in various cases, such as MSME customer segmentation [3], grouping of sales regions[1], analysis of drug sales patterns[2] , marketing budget segmentation [4], as well as food menu grouping based on sales level [5]. What these studies have in common is the use of K-Means to produce homogeneous data sets that facilitate the analysis process and support more targeted decision-making.

In addition to data segmentation, sales forecasting is also an important aspect in supporting business planning because it can help companies forecast market demand and strategize operations more effectively. One of the widely used methods is Linear Regression, which is a supervised learning algorithm that models relationships between variables to predict values based on historical data. Previous research has shown that Linear Regression is able to provide good prediction results on various problems, such as the prediction of boarding house rental prices [6] as well as predictions of goods sales [7]. Unlike K-Means, which focuses on the data segmentation process, Linear Regression plays a role in modeling the tendency to change values so that the two methods have complementary characteristics when applied in an integrated manner.

Although various studies have applied K-Means Clustering and Linear Regression in the analysis of sales data, most studies still use the two methods separately or focus only on a single analysis objective. Research by Henry Adam et al. [7] has combined K-Means Clustering and Linear Regression for sales predictions, but the study focuses more on the implementation of the method without providing an in-depth analysis of the characteristics of each cluster as the basis for the interpretation of the prediction results. On the other hand, most other studies only utilize K-Means for customer segmentation [3] , [8], [4] or product segmentation [2], [5] without integrating segmentation results into the sales prediction process. This condition shows that there are still limitations in research that is able to connect the product segmentation process with sales predictions in one integrated analysis framework. In fact, the integration of the two processes has the potential to produce more comprehensive information about product characteristics as well as sales trends so that it can support business decision-making more effectively.

Based on the research gap, this study proposes the integration of K-Means Clustering and Linear Regression in the analysis of sales data. The K-Means Clustering algorithm is used to group products based on the characteristics of their sales patterns, while Linear Regression is used to predict sales trends in each product group based on historical data. The determination of the number of clusters is carried out using the Elbow method so that the segmentation process can be carried out more objectively before entering the prediction modeling stage.

This research provides three main contributions. First, develop an integrated analysis framework that integrates K-Means Clustering and Linear Regression to produce product segmentation as well as sales predictions in one analysis flow. Second, evaluating the quality of segmentation results using the Elbow method and evaluating the performance of the prediction model using the determination coefficient (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). Third, providing an interpretation of the characteristics of each cluster along with the results of its sales predictions so as to produce more comprehensive information in supporting product management, marketing strategy development, and data-based sales planning. Thus, this research is expected to contribute to the development of the application of data mining and machine learning in sales data analysis.

2. Research Methodology

This study uses a quantitative method with an experimental design to assess the implantation of K-Means Clustering and Linear Regression algorithms in sales data analysis. The quantitative method was chosen because of the research's focus on processing numerical data generated from sales transactions, while the experimental design aimed to test the performance of two algorithms through a series of

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

structured analysis steps. The research stages include dataset collection, data characteristics analysis, data preprocessing, product segmentation using K-Means Clustering, sales prediction through Linear Regression, model evaluation, and interpretation of analysis results. The overall research flow is illustrated in Figure 1.

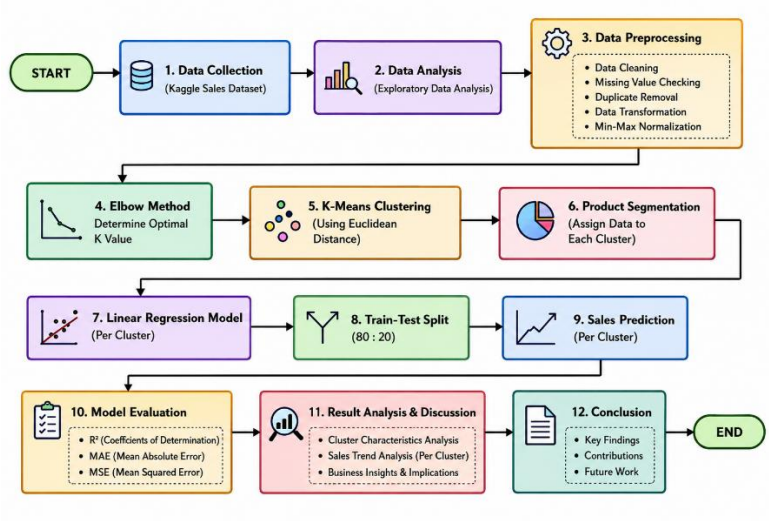

Fig. 1. Research Flow

This study uses Sales Dataset obtained from the Kaggle platform as a data source. The dataset was chosen because it was public, easily accessible, had attributes relevant to the sales analysis, and allowed other researchers to replicate the research. The dataset contains 49 sales transactions with five main attributes, namely Date, Product, Quantity, Unit Price, and Total Sales. The transaction covers the period of January 2023 and describes the sales activity of various products. The following table describes the attributes used.

Table 1. Sales Dataset

Attributes	Data Type	Description
Date	Date	Transaction Date
Product	String	Product Name
Quantity	Integer	Number of Products Sold
Unit Price	Float	Unit Price of Principal
Total Sales	Float	Total Sales Value

This study utilizes two groups of variables. Variables for the clustering process (Quantity, Unit Price, Total Sales), these three attributes were chosen because they can describe the sales characteristics of each product. The variable for the regression process, the independent variable (X) uses the transaction time index, while the dependent variable (Y) uses the Total Sales value. Time representation is used to identify patterns of change in sales based on historical data.

The initial stage of the research was carried out through the analysis of the characteristics of the dataset to understand the data structure and determine the attributes used in the modeling process. Analysis includes data type identification, descriptive statistics, data distribution, and data quality checking. Next, the data preprocessing stage is carried out so that the dataset is ready to be used in the machine learning process. The pre-processing stages include:

1. Missing value checks;
2. Removal of duplicate data;
3. Transform data types when needed;
4. Selection of research attributes;

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

5. Numerical data normalization.

The results of the examination showed that the dataset did not contain missing values or duplicate data so that all data could be used in the analysis process.

Because the K-Means algorithm uses Euclidean distance calculations that are sensitive to differences in attribute scales, normalization is carried out using the Min-Max Scaling method on the Quantity, Unit Price, and Total Sales attributes.

Normalization is carried out using Equation (1).

$$\frac{X' = X - X_{MIN}}{X_{MAX} - X_{MIN}}$$

Description :

X = original value
Xmin = minimum value
Xmax= maximum value
X′ = normalization results

Normalization aims to equalize the range of values of all attributes so that there are no attributes that dominate the grouping process.

The data from the pre-processing results is then analyzed using the K-Means Clustering method to separate products based on their sales characteristics. K-Means is an unsupervised learning algorithm that divides data into groups based on the degree of similarity, using Euclidean distances to the center of the cluster. [9]The implementation was carried out using the Scikit-Learn library with the parameters as shown in Table 2.

Table 2. Preposes Results

Parameters	Value
Squirt	k-means++
n_clusters	Elbow Method Results
max_iter	300
n_init	10
random_state	42
Distance	Euclidean Distance

The k-means++ method is used to generate a more stable centroid initialization thereby reducing the likelihood of convergence on local solutions. The parameter random_state = 42 is used so that the clustering process can be reproduced.

The distance between the data and the centroid is calculated using Euclidean Distance.

$$d(x,c)=\sqrt{\sum_{i=1}^{n}(x_{i-}c_i)^2}$$

The determination of the number of clusters was carried out using the Elbow Method by evaluating the change in the Within Cluster Sum of Squares (WCSS) value at several K values.[10] used as a basis in determining the number of clusters that provide the best balance between cluster homogeneity and model complexity.

The clustering results are then used to separate the data into several groups of products that have similar sales characteristics. Each cluster is then analyzed independently at the prediction stage using Linear Regression [11]. After the clustering process is completed, a Linear Regression model is built

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

separately on each cluster so that the prediction model can study the sales patterns of product groups that have similar characteristics. The model was developed using LinearRegression from the Scikit-Learn library with the following configuration.

Table 3. Linear Regression

Parameters	Value
Models	Linear Regression
Estimation Method	Ordinary Least Squares (OLS)
Fit_intercept	True
Copy_X	True
Positive	False
Train-Test Split	80 : 20
Evaluation Metrics	R2, MAE, MSE

Data sharing uses a ratio of 80% training data and 20% test data because these proportions are widely used in machine learning modeling to achieve a balance between the training process and model evaluation.

The Linear Regression Model is expressed as follows.

$$\gamma = \beta_o + \beta_1 x + \varepsilon$$

The model parameters were estimated using the Ordinary Least Squares (OLS) method which minimized the number of residual squares.

$$\min \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

This approach results in optimal parameter estimation so that the model is able to describe the linear relationship between transaction time and total sales in each cluster.

The performance of the model was evaluated using three evaluation metrics, namely Coefficient of Determination (R^2), Mean Absolute Error (MAE)[12], and Mean Squared Error (MSE).

The determination coefficient is used to measure the model's ability to explain data variations.

$$R^2 = 1 - \frac{SS_{res}}{ss_{tot}}$$

The average prediction error was calculated using MAE.

$$\text{MAE} = |\frac{1}{n} \sum | \, y_i - \hat{y_i}$$

While MSE is used to measure the square mean of prediction error.

$$\text{MSE} = \frac{1}{n} \sum (yi - \hat{yi})^2$$

R^2 values that are closer to 1 indicate better predictive ability, while smaller MAE and MSE values indicate lower prediction error rates.

The final stage of the research is to integrate the segmentation results and prediction results to obtain more comprehensive information about product characteristics and sales trends. The analysis is carried out through several stages, namely:

1. Interpret the characteristics of each cluster based on centroid values;
2. Compare actual sales patterns with predicted results in each cluster;
3. Evaluate model performance using R^2, MAE, and MSE values;
4. Comparing the results of the research with previous research;

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

5. Identify the implications of results on marketing strategies, inventory management, and business decision-making.

3. Results and Discussion

This section explains in detail the results obtained from each stage of the research methodology that has been implemented.

3.1 Data Analysis and Preprocessing Result

The pre-processing results show that the dataset is free of missing values and duplicate data so that the entire data can be used in the next stage of analysis. In addition, normalization using the Min-Max Scaling method succeeded in equalizing the range of values of the Quantity, Unit Price, and Total Sales attributes, so that there were no attributes that dominated the distance calculation process in the K-Means Clustering algorithm.

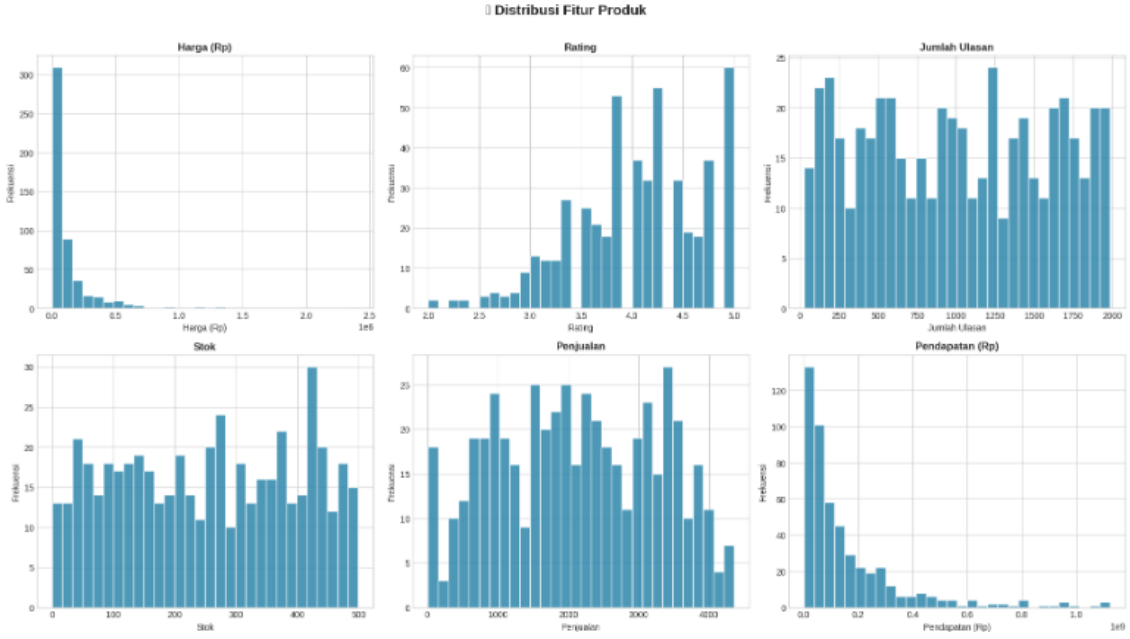

Fig. 2. Distribution of Product

Figure 2 shows that the data distribution after normalization maintains the original data distribution pattern even though it is at a uniform range of values. This indicates that the normalization process only changes the scale of the data without changing the relationships between objects in the dataset. This condition is very important because the K-Means algorithm uses Euclidean distance calculations that are sensitive to the difference in scale between attributes. Therefore, normalization contributes to producing a more objective and stable clustering process.

These findings are in line with the research of Zamzani et al. [3] and Desryani et al. [1] which states that normalization of data before the clustering process is able to improve the quality of grouping by reducing the influence of attributes that have a larger range of values.

3.2 K-Means Clustering Results

The determination of the number of clusters was carried out using the Elbow Method by evaluating the change in the Within Cluster Sum of Squares (WCSS) value at several K values. Thus, four clusters are selected as the optimal number of clusters because it provides a balance between the homogeneity of the data in the cluster and the complexity of the model.

In addition to considering the WCSS value, the selection of four clusters also results in more interpretable product segmentation. A smaller number of clusters causes multiple products with different characteristics to be in the same group, while a larger number of clusters results in segmentation that is too detailed and less effective to be used as a basis for decision-making.

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

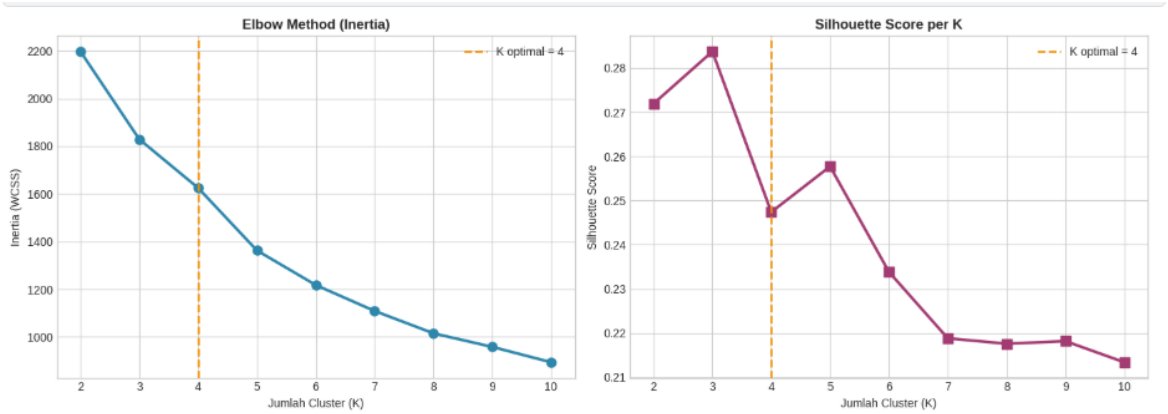

Fig. 3. Elbow Method (Inertia) and Silhouette Score per K (K optimal = 4)

The selection of K = 4 results in more representative segmentation than fewer or more clusters. If the number of clusters is less than four, several products with different characteristics are still incorporated into one group, so the segmentation information becomes less specific. Conversely, if the number of clusters is increased, the division of data becomes too detailed, making it difficult to interpret and less effective as a basis for business decision-making.

Furthermore, the clustering results are visualized using Principal Component Analysis (PCA) as shown in Figure 4 below:

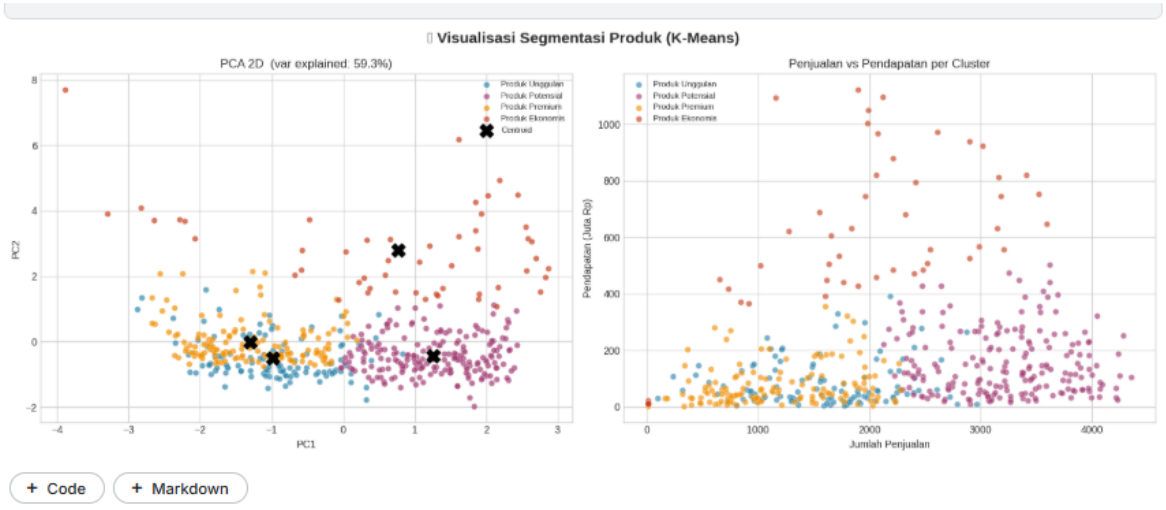

Fig. 4. Visualization of K-Means Clustering Using PCA

The PCA visualization in Figure 4 shows that most of the data forms relatively separate groups. Although there are still several points that are close to each other due to almost similar sales characteristics, the pattern of separation between clusters is quite clear. This shows that the K-Means algorithm is able to identify similarities in product characteristics based on the attributes of Quantity, Unit Price, and Total Sales[13]. These results are consistent with the research of Fajar et al. [2] and Triyandana et al. [5] who report that K-Means is effectively used to segment products based on sales patterns so as to produce information that is easier to interpret as the basis for developing business strategies.

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

The characteristics of each cluster produced are summarized in Table 4 below:

Table 4. Characteristics of K-Means Clustering Results

Cluster	Average Sales	Average Price	Characteristics
0	Low	Intermediate	Products with low sales volume and medium price
1	Height	Height	Superior products with high sales volume and premium prices
2	Intermediate	Low	Reasonably priced products with a medium sales volume
3	Low-Medium	Very Low	Cheap products with limited sales turnover and relatively small profit contribution

1. Cluster 0 contains products with a mid-to-lower sales rate and a price in the middle category. This characteristic indicates that the products in this group have a lower sales contribution than other clusters, so further evaluation of marketing strategies and stock management is needed.
2. Cluster 1 recorded the highest average prices and sales, so it was included in the category of superior products. Products in this group have the potential to be the main contributor to revenue, so they must be prioritized in managing stock to remain available.
3. Cluster 2 includes fairly low-priced products, but sales are at a medium level. This situation indicates an opportunity to increase sales through promotional strategies or marketing improvements.
4. Cluster 3 is a product segment with a fairly low price and sales volume. Due to its smaller contribution to sales, this segment can be a major focus in assessing product portfolios as well as optimizing inventory.

Based on Table 5, the Linear Regression model for Cluster 1 provides the most superior performance with an R^2 value of 0.9215, indicating that about 92.15% of the sales variation can be explained by the model. Low MAE and MSE values also indicate relatively small prediction errors, so the model can replicate historical data patterns fairly accurately. The better performance in Cluster 1 is thought to stem from the more uniform nature of the data compared to other clusters. Products in this group show fairly stable sales patterns, so the linear relationship between time and total sales can be mapped more precisely. In contrast, slightly lower R^2 values in Cluster 0 and Cluster 3 indicate greater data variation in both groups, so the linear model is not yet fully able to explain all sales fluctuations. This finding is consistent with the work of Henry Adam et al. [7], which states that Linear Regression can provide satisfactory predictions on sales data with relatively consistent historical patterns.

3.3 Data Analysis and Preprocessing

After the products were grouped using K-Means Clustering, Linear Regression was applied to predict the sales trend of each cluster. This approach allows the model to learn sales patterns that are more homogeneous than if all data were analyzed as a group. The prediction results are shown in Figure 5.

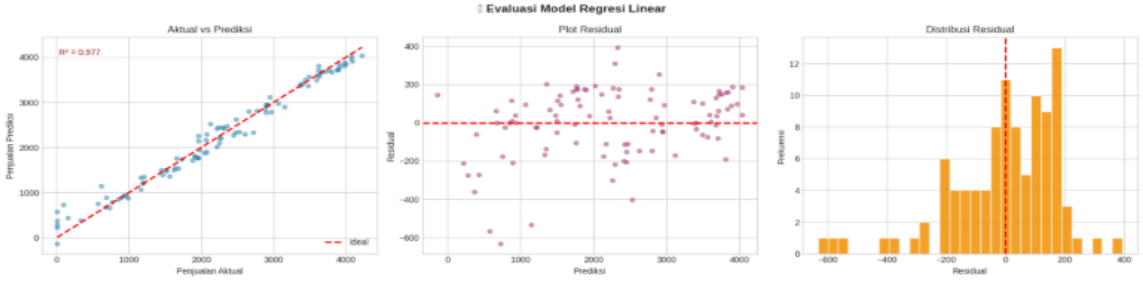

Fig. 5. Actual Sales Trend Graph and Linear Regression Prediction

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

Figure 5 shows that the prediction line is able to follow the actual data pattern on each cluster. This indicates that the Linear Regression model can capture the trend of changes in sales based on historical data. The analysis per cluster also shows that each product group has different sales patterns, [14]So that the use of one model for all data has the potential to produce less accurate predictions. The performance evaluation of the Linear Regression model is summarized in Table 5 as follows:

Table 5. Performance Evaluation of Linear Regression Models

Evaluation Metrics	Cluster 0	Cluster 1	Cluster 2	Cluster 3
MAE	0.0821	0.0643	0.0752	0.0887
MSE	0.0104	0.0079	0.0091	0.0112
R2	0.8734	0.9215	0.8956	0.8592

Based on Table 5, Cluster 1 performed best with an R^2 value of 0.9215, which shows that the model was able to explain about 92.15% of the sales variation in the flagship product group. The lowest MAE and MSE values also indicate that the prediction error in these clusters is relatively small. Meanwhile, Cluster 0, Cluster 2, and Cluster 3 also showed R^2 values above 0.85, so the model was still able to provide a good prediction even though the sales patterns in the three clusters were more varied.

These results show that the application of Linear Regression after the clustering process is able to produce more accurate predictions because each model is built on similar product characteristics. Thus, the integration of K-Means Clustering and Linear Regression provides added value compared to the use of Linear Regression directly on the entire dataset, [15] Because it is able to produce more specific predictions for each product segment and support the preparation of more appropriate business strategies.

3.5 Analysis Results

The integration of K-Means Clustering and Linear Regression results in a more comprehensive analysis in evaluating sales data. K-Means Clustering is used to group products based on their sales characteristics, while Linear Regression predicts the sales trend of each cluster. This approach allows the analysis to be carried out on groups of products that have similar characteristics so that the prediction results become more specific.

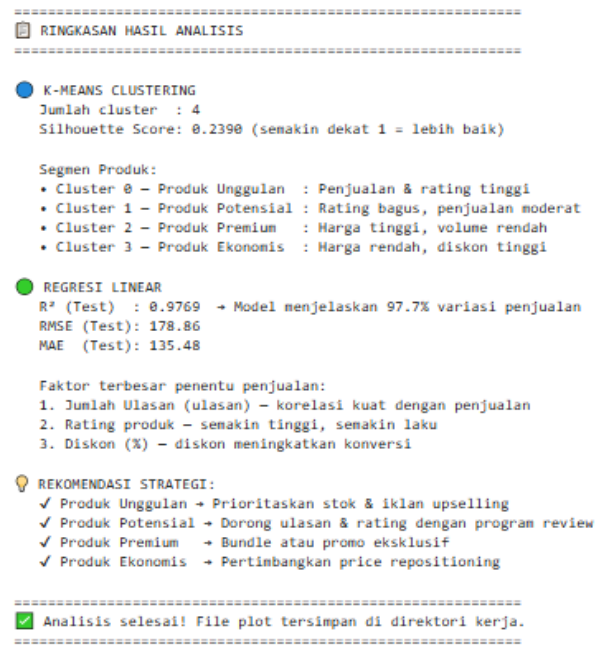

Fig. 6. Sales Trend Prediction per K-Means Yield Cluster

MIoSA: Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, pp. 86-97
E-ISSN: XXXX-XXXX

Figure 6 shows that each cluster has a different sales trend pattern. These differences indicate that each product group requires a different management strategy. Products in clusters with high sales need to be prioritized in stock management to maintain product availability, while products in clusters with low sales require evaluation of promotion strategies, pricing, and inventory management to be more efficient. Meanwhile, products in clusters with mid-sales have the potential to be improved through more appropriate promotional programs or marketing strategies.

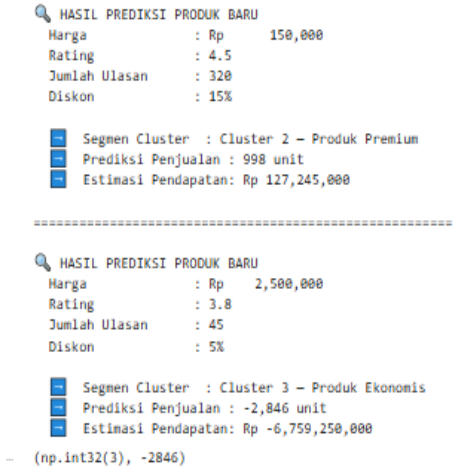

Fig. 7. Prediction Results

Figure 7 shows that the predicted results are able to follow the actual sales data patterns in each cluster. This shows that the Linear Regression model is able to capture the trend of sales changes after the data is grouped by its characteristics. Grouping first using K-Means produces more homogeneous data so that the prediction model can provide more representative results for each product segment.

The combination of K-Means Clustering with Linear Regression offers advantages over the use of each method separately. K-Means groups products based on similar sales patterns, while Linear Regression leverages these patterns to create more targeted prediction models for each group. This approach not only displays sales trends, but also provides insight into the characteristics of each product segment. In application, these findings can be used as a basis for formulating data-driven business strategies, such as setting stock management priorities for key products, designing promotional programs for goods with growth potential, and assessing products with low sales contributions. Thus, the combination of the two algorithms adds more value than the use of clustering or prediction methods separately.

This research has several limitations. First, the amount of data used is quite limited so the model's ability to generalize still needs to be tested on larger datasets. Second, the analysis only utilizes the K-Means Clustering and Linear Regression algorithms, so there is no comparison with other methods that can provide better performance. Third, the variables presented are still limited to sales attributes only, not including external factors such as seasons, promotions, or consumer behavior. Therefore, follow-up research is recommended to use a broader dataset, add more diverse variables, and evaluate performance with other machine learning algorithms.

4. Conclusion

This study is intended to combine K-Means Clustering with Linear Regression to help analyze patterns and forecast sales based on data. The findings of the study showed that the K-Means Clustering method successfully separated the products into four groups (K = 4) according to their sales characteristics, while Linear Regression provided a good prediction in each group with the highest coefficient of determination (R^2) value reaching 0.9215. These results confirm that segmenting the product before starting the prediction results in a more focused and representative analysis than applying the prediction model to the entire dataset at once. The main contribution of this research lies in the application of a framework that combines segmentation and prediction processes, resulting in a more

complete understanding of product characteristics and their sales behavior. This approach can be used as a basis for supporting business decision-making, such as designing marketing strategies, managing inventory, and determining product priorities based on the characteristics of each segment. This study still has limitations related to the volume of data and the use of relatively simple algorithms, so the results cannot be generalized widely. Therefore, it is recommended that further research take advantage of larger datasets, compare performance with clustering algorithms or other predictions, and test proposed approaches across different business sectors to obtain more comprehensive findings.

References

[1] F. Desryani, B. Ginting, and M. Ramadhan, "The Application of Data Mining in the Grouping of Bottled Mineral Water Sales Shop Areas Using the K-Means Clustering Algorithm," vol. 4, no. September, pp. 1243–1257, 2025.

[2] M. Fajar, N. Rahaningsih, and R. Danar Dana, "Analysis of Drug Sales Patterns at An-Naafi Pharmacy Using the K-Means Clustering Method," *JATI (Journal of Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 486–492, 2024, doi: https://doi.org/10.36040/jati.v8i1.8395

[3] M. Imron Zamzani, A. Andini, K. R. K. Rahayu, and N. Muhasdi, "Segmentation Analysis of MSME Customers of Albynha Cake Shop Using RFM and K-Means Methods," *J. Ind. Eng. Technol.*, vol. 2, no. 1, pp. 1–11, 2026, doi: https://doi.org/10.36277/jietech.v2i1.57

[4] N. K. Ahmadi and Herlina, "Segmentation Analysis of Eiger Product Purchase Decisions in Bandar Lampung," J. Manaj. Magister, Vol 03. No.01, January 2017, vol. 03, no. 01, pp. 75–95, 2017.

[5] G. Triyandana, L. A. Putri, and Y. Umaidah, "Application of Data Mining for Food and Beverage Menu Grouping Based on Sales Level Using the K-Means Method," vol. 6, no. 1, pp. 40–46, 2022. doi: https://doi.org/10.30871/jaic.v6i1.3824

[6] M. R. Fahlepi and A. Widjaja, "Application of Multiple Linear Regression Method for Prediction of Boarding House Rental Prices," vol. 1, no. November, pp. 615–629, 2019.

[7] Henry Adam, Tukino, Elfina Novalia, and A. L. Hananto, "Prediction of Goods Sales Using K-Means Method and Linear Regression," Bull. Comput. Sci. Res., vol. 5, no. 4, pp. 298–307, 2025, doi: 10.47065/bulletincsr.v5i4.541. doi: https://doi.org/10.47065/bulletincsr.v5i4.541

[8] Sulistyowati, B. E. Ketherin, A. A. Arifiyanti, and A. Sodik, "Analysis of Consumer Segmentation Using the K-Means Clustering Algorithm of the Department of Information Systems, Adhi Tama Institute of Technology Surabaya," no. September, 2018.

[9] Anggelia Deli, Petronela Kurniati Kondang, Wilnotus Daniel Awil, and Astina Ranti, "Analysis of Marketing and Product Sales Budget Segmentation in the Retail Industry Using K-Means Clustering Based on R Shiny," Instink Inov. Education, Technology. Inf. and Comput., vol. 4, no. 1, pp. 41–54, 2025, doi: https://doi.org/10.30599/9a8f3305

[10] Nayla Salsabila, Karina Aulisari, and Hani Zulfia Zahro, "Application of K-Means Algorithm for Clustering Productivity of Ginger Plants," Infotek J. Inform. and Technology., vol. 8, no. 1, pp. 228–238, 2025, doi: https://doi.org/10.29408/jit.v8i1.28195

[11] A. Gafari and R. Sovia, "An Analysis of Public Satisfaction with Government Services : A Multi-Method Approach Using PCA, K-Means Clustering, and Linear Regression," vol. 30, no. 1, pp. 1–8, 2026, doi: https://doi.org/10.46984/sebatik.v30i1.2742

[12] R. Luthfiansyah and B. Wasito, "Application of Deep Learning Techniques (Long Short Term Memory) and Classical Approaches (Linear Regression) in Prediction of BRI Stock Movements," *J. Inform. and Business*, vol. 12, no. 2, pp. 42–54, 2023, doi: https://doi.org/10.46806/jib.v12i2.1059

[13] N. Aminudin, N. Hidayat, D. Feriyanto, D. Septasari, and I. Awaliyani, "Digital Landscape and Behavior in Indonesia 2024: A National Survey Analysis of Internet Penetration, Cybersecurity Risks, and User Segmentation Using K-Means Clustering and Logistic Regression," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3336–3351, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.5.5117

[14] F. Erza *et al.*, "Object Tracking System on Quadcopter Using Image Segmentation with Hsv Color Detection and Linear Regression Method Based on," vol. 9, no. 7, pp. 1733–1740, 2022, doi: https://doi.org/10.25126/jtiik.2022976808

[15] N. M. Fithryani, A. R. Dikananda, D. Rohman, and K. Cirebon, "K-MEANS ALGORITHM FOR," vol. 13, no. 1, pp. 997–1003, 2025.