

Enhancing marketing efficiency through data-driven customer segmentation with machine learning approaches

Fanindia Purnamasari, Umayra Ramadhani Putri Nasution, Marischa Elveny

Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received Oct 4, 2024

Revised Mar 28, 2025

Accepted Jul 2, 2025

Keywords:

Clustering

Customer behavior

Customer segmentation

Machine learning

Marketing efficiency

ABSTRACT

The importance of understanding consumer behavior in transaction data has become a key to improving marketing efficiency. This study aims to explore the application of machine learning (ML) techniques for data-driven consumer segmentation, focusing on improving product marketing strategies. This work addresses the limitations in the existing literature, especially in terms of handling high-dimensional data that can reduce segmentation quality. Previously, various studies have used clustering algorithms such as K-means without considering dimensionality reduction, which often leads to decreased accuracy and long computation time. In this study, we propose a new approach that combines principal component analysis (PCA) for dimensionality reduction and K-means clustering for consumer segmentation based on purchasing behavior. Experimental results show that using PCA to reduce data dimensionality significantly improves segmentation quality with an inertia score of 1,455,650 and a silhouette score of 0.486366. By implementing this method, we can group consumers into three segments based on frequently purchased product categories and the most common payment methods. These findings provide a scalable, data-driven segmentation framework that can be applied to improve marketing effectiveness by providing special discounts on various products based on the payment method used.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Fanindia Purnamasari

Department of Information Technology, Faculty of Computer Science and Information Technology

Universitas Sumatera Utara

Medan, Indonesia

Email: fanindia@usu.ac.id

1. INTRODUCTION

Customer segmentation help the company to improve the profits. Acquiring new consumers is not quite as important as keeping old customer According to the Pareto principle, a company's customers contribute to 20% of its revenue, which is higher than that of other customers [1]. The company can use customer segmentation to marketing budgeting customize marketing strategies, observe trends, organize product development, design advertising campaigns, and provide appropriate products by utilizing a range of distinctive customer attributes [2]. An organization may no longer have to promote products at random because there is a direct association between the products being promoted and interested clients are more likely to make frequent purchases. To identify the regions where markets, customers, and transactions are most prevalent, first used location distribution heatmaps [3]. It delivered to the right location in the shortest time and at the least cost to consumers. The market distribution strategy permeates all aspects of the organization's actions and encompasses its supply, production, promotion, and distribution environments [4].

In this era, machine learning (ML) has become an effective approach for improving customer segmentation, particularly when it comes to unsupervised learning methods like clustering algorithms. ML algorithms have the ability to identify undiscovered trends and customers based on actual behavioral tendencies by evaluating vast amounts of customer data [5], [6]. Businesses can adapt their marketing campaigns to the unique requirements and preferences of various consumer segments as a result of these data-driven insights. Customer segmentation or customer clustering is strategic decision making for business to seek sustainable growth and customer satisfaction. Clustering is used to identify patterns, such as the top selling product and the preferred payment method based on customer transaction [7]. K-means is one of the most frequently used in customer clustering [8], [9]. However, some clustering algorithms such as K-means often encounter problems when applied to data with high dimensions or features. Some problems are decreased classification accuracy, poor quality of cluster, and long computing times. Dimension reduction is one strategy that can be used to preserve optimal algorithm performance. There are two approaches to dimensionality reduction: feature extraction and feature selection [10]–[13].

According to Christy *et al.* [14], segmentation helps organizations better understand to customer needs and identify future customers characteristics. The researchers used recency, frequency, monetary (RFM) analysis for segmentation and adapted it to other algorithms like K-means and RFM K-means [15]–[17]. In e-commerce, the user behavior will be observed in their activities using website [18]–[20]. Some companies often segment customer based on e-commerce checkout abandonment rates. Companies might offer discounts to encourage customers to make a purchase rather than filling their cart without checkout, and also companies provide support to customers with review questions about products, payment, and quality. The data which contains purchased products can be further analyzed to develop product promotion, product operation strategies for customers in each cluster [21], [22]. Clustering is the procedure of grouping a set of data into groups that exhibit similar characteristics [23]. A cluster is a collection of observations that are similar within the same cluster but differ from observations in other clusters. K-means clustering is one of the clustering methods which simple and popular way to segment a dataset into K different clusters. To perform K-means clustering, the number of K (cluster) should be determined. Determining the number of clusters is crucial part for managing these groups. The contribution is using principal component analysis (PCA) as dimensionality reduction in high dimensionality data as explained in the highlighted sentence

In this study, we are more focuses on the application of K-means for customer segmentation by using PCA as a dimensionality reduction step, to handle the problem of high dimensional data that can degrade the quality of segmentation. Therefore, the main contribution of this research is the application of dimensionality reduction technique (PCA) and the use of K-means for customer segmentation based on purchasing behavior. While the previous studies have explored the impact of K-means clustering by using fixed k- clustered. The parameter in customer demography and behavior such as age, annual income, spending score, purchase, history, and quantity [24].

2. METHOD

This section presents the methodology used in this study. The aim is to segment customers depending on their behavior in purchasing transaction. We performed three stages namely data analysis to identify the outlier from dataset and to check missing value. Second stage was PCA implementation to reduce data, and the third K-means algorithm implementation then analyzed the interpretation of clustering result [24]. The research method as shown as Figure 1.

2.1. Data analysis

In this study we used data contain information about 2,500 instances consumer yearly transaction encompasses 9 features of yearly segmentation such as age, annual income, spending score, purchase history, product category, quantity, unit price, total price, and payment method. But in this study, we considered 6 features including age, annual income, spending score, purchase history, product category, and payment method. In data analysis, first we identified the outlier from four column in data distribution, namely age, annual income, spending score and purchase history. Outliers are observations statistically significant different from the bulk of the data. The process to identify the outliers using boxplot with the interquartile range (IQR) [25]. The identifying outlier process as explained below:

Calculate the first quartile (Q1) and third quartile (Q3):

- a) Q1 is the 25th percentile of the data (the value below which 25% of the data falls).
- b) Q3 is the 75th percentile of the data (the value below which 75% of the data falls).
- i) Calculate the IQR:

$$IQR = Q3 - Q1 \quad (1)$$

ii) Determine the lower bound and upper bound:

$$\begin{aligned} \text{Lower Bound} &= Q1 - 1.5 \times IQR \\ \text{Upper Bound} &= Q3 + 1.5 \times IQR \end{aligned} \quad (2)$$

iii) Identify outliers

iv) Position data point less than the lower bound or greater than the upper bound is considered an outlier.

Based on the box plot visualization the outlier was not found as Figure 2. Then we performed data pre-processing is used for further analysis on PCA and K-means clustering. The data were first normalized using min-max normalization into the range [0,1]. Normalized data means each feature of data has equal weight, which prevents larger-scale features from dominating the analysis results.

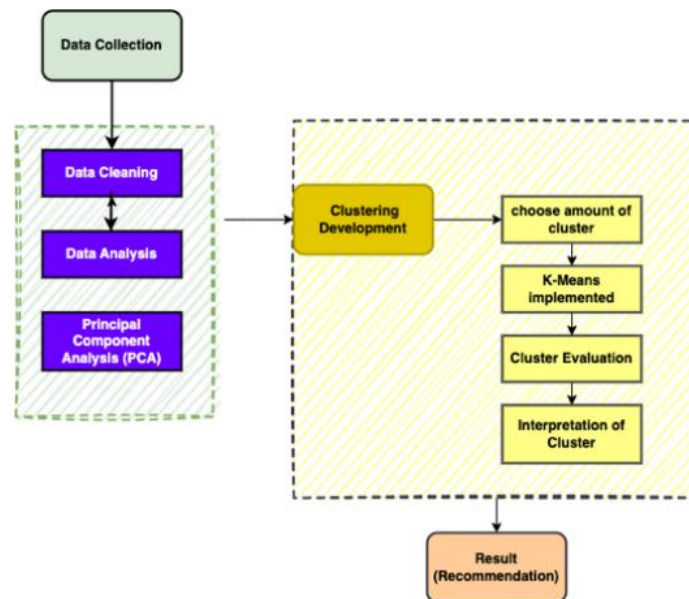


Figure 1. Research method

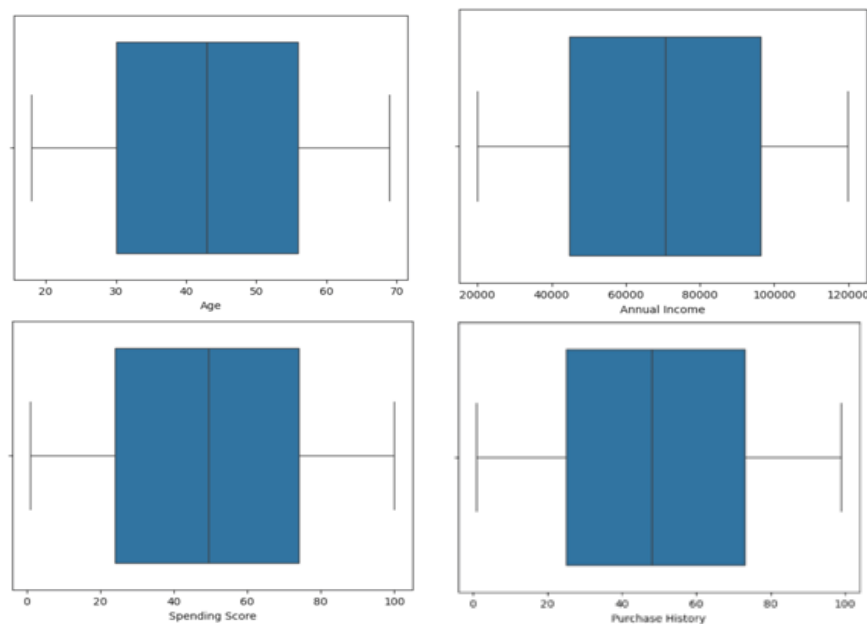


Figure 2. Data distributions to identify the outlier

2.2. Principal component analysis

PCA is used to reduce the complexity of data. After identifying the outlier, we performed PCA due to K-means clustering is sensitivity with high dimensional data. First, we calculate the covariance to identify among features in data. Then perform eigen decomposition which a technique in linear algebra used to decompose matrices into eigenvectors and eigenvalues. To apply PCA we used formulas as [26]:

i) Data centering

$$X_{centerd} = X - \mu \quad (3)$$

Where:

X is data. matrix

μ mean of data in each feature

ii) Covariance matrix

$$C = \frac{1}{n-1} X_{centered}^T X_{centered} \quad (4)$$

Where:

n is amount of data,

$X_{centered}$ is data that has been reduced to the mean of data

Covariance is a measure of how well two variables correlate with one another. Its diagonal contains every variance and contains all potential covariance pairs between the m variables, while smaller variance values may indicate the noise in these data the huge variance values are significant since they correlate to the intriguing dynamics in these data.

iii) Eigen decomposition

$$C = V \Lambda V^T \quad (5)$$

Where:

V is eigen vector matrix (principal components),

Λ is a diagonal matrix containing eigenvalues (the variance explained by each principal component).

iv) Projection into principal components

$$Z = X_{centered} V \quad (6)$$

Where: Z is data that has been projected to a new space.

In this study, we run principal components as 2, they are the two primary components used by PCA for reducing the dimension of the data. Reducing the mean values from the initial dataset is the first step in the PCA analysis process, in order to provide the original dataset's analysis phase equal weights and appropriate normalization. In Figure 3 shows that the PC1 and PC2 axes represent the two of PCA components, indicates that the spread and distribution of customers in a simpler feature space. Based on suitability analysis, the data is divided into several clusters have a strong relationship with other cluster. In contrast, other clusters appear more dispersed indicates that greater variation among customers within those clusters.

2.3. Implementation of K-means clustering

The data were first normalized using min–max normalization into the range [0,1]. The determination number of clusters was not fixed in early stage, but using the silhouette coefficient. The first center is chosen by random, and following points are chosen with a probability proportional to the squared distance from the nearest center. This study used $10 \times$ run from random initial positions the result with the lowest within-cluster sum of squares will be used and max iteration 300 in order to the algorithm reach convergence condition. The iterative process aims to reduce the distance between data points and centroids in each cluster, moving the centroids to better describe the cluster center. The K-means steps as explained below [24]:

i) Centroid initialization

Select k initial centroids randomly from the dataset, where k represents the required number of clusters.

ii) Calculate Euclidean distance

For every data point x_i , compute the Euclidean distance to each centroid C_j . The equation for Euclidean distance is:

$$d(x_i, C_j) = \sqrt{\sum_{n=1}^N (x_{in} - C_{jn})^2} \quad (7)$$

Where:

x_i is the data point to $-i$,

C_j is the centroid to $-j$,

x_{in} and C_{jn} is the value of the component $-n$ from data and centroid.

iii) Assign data points to nearest clusters

Each data point x_i will be allocated to the cluster whose centroid exhibits the minimal distance, as determined by the Euclidean distance computation.

iv) Update Centroid Position

After all data has been allocated to clusters, revise the position of each centroid by computing the mean of all data points contained inside that cluster:

$$C_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \quad (8)$$

Where:

S_j is a set of data points classified into clusters j ,

C_j is the new centroid for the cluster j .

v) Repeat process

Continue executing steps 2 to 4 until the centroids exhibit negligible changes or a predetermined iteration limit is attained. The K-means technique generates k clusters, with each cluster including data points nearest to their corresponding centroids.

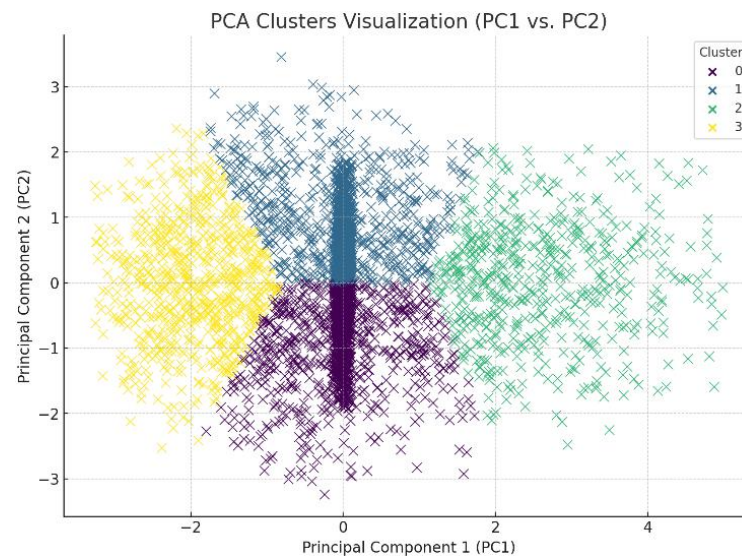


Figure 3. Optimal component in PCA analysis

Then, we evaluate the cluster result by utilizing the inertia and silhouette score. The inertia and silhouette score calculations are performed, where inertia measures the compactness of the cluster. The lower the inertia value, the better the clustering in terms of compactness (data points are closer to their respective cluster centroids). The higher the Silhouette coefficient approach to 1 indicates the better the cluster whereas score closes to 0 indicate overlapping clusters, can be seen in Table 1.

Table 1. Cluster evaluation results

Number of clusters (K)	Inertia	Silhouette score
2	10.394.179	0.311896
3	1.455.650	0.486366
4	0.504450	0.247399

3. RESULTS AND DISCUSSION

The results of this study is highlight the significant impact that ML-based customer segmentation has on marketing efficiency. In this section, we will discuss the key findings of clustering result based on analysis of PCA, distribution of clustering result, interpretation of clustering results and business insight from cluster.

3.1. Data analysis

We perform descriptive statistic to help understand and explore the data. Descriptive statistics, such as mean, median, standard deviation, minimum, and maximum values, provide an overview of the distribution, and central tendency. Table 2 show the descriptive statistics of the dataset used in this study. This dataset consists of 2,500 customer data entries covering 6 features, including age, annual income, spending score, purchase history, product category, and payment method. This data is real customer transaction data collected for segmentation analysis based on purchase behavior.

Table 2. Descriptive statistics for data

Feature	Count	Mean	Standard deviation	Min	Max
Age	4096	44.05	15.08	18.0	69.0
Annual income	4096	69910.37	28870.58	20060	119993.0
Spending score	4096	50.97	28.86	1.0	100.0
Purchase History	4096	50.33	28.32	1.0	99.0
Quantity	4096	5.03	1.99	1.0	9.0

3.2. Principal component analysis

Two components of PCA (PC1 and PC2) explain about 42.57% of the total variance in the data. This means that despite using two dimensions, it still retains quite significant information from the original data. The high variance explained by these two components indicates that the data can be effectively represented in two dimensions for visual analysis and interpretation. Each value in PC1 and PC2 represents a projection of the customer data into the new feature space. Positive or negative values in PC1 and PC2 indicate where the data is located relative to the mean in that component. High negative values in PC1 and positive in PC2 (first row, PC1=-2.264738, PC2=0.829777) indicate that the data has characteristics far below the mean for the PC1 dimension but above the mean for the PC2 dimension. Conversely, values closer to 0 in both components indicate that the data has characteristics close to the mean. can be seen in the Table 3.

Table 3. PCA analysis with two components (PC1 and PC2)

PC1	PC2
-2.264.738	0.829777
0.439067	1.393.833
0.946629	1.247.088
-2.092.568	-1.084.797
0.083697	-1.782.019

3.3. Application of K-means algorithm

Cluster is a label to identify each cluster in the K-means analysis. In Centroid X and Centroid Y, the coordinates of the initial centroid are initialized randomly. This centroid is the initial centre point of each cluster and will then be adjusted through the iteration of the K-means as shown as Table 4. The closest cluster centroid based on Euclidean distance data is assigned to this cluster in the initial iteration as shown as Table 5. Table 6 explains the centroid is updated after the data is assigned to the closest cluster; the centroid position is recalculated based on the average position of all data points in the cluster. Centroid X and Centroid Y, the new centroid position in PCA coordinates (PC1 and PC2).

We found that the changing in the average distance of each centroid from the previous iteration to the current iteration in the converging process. The iteration process ends when the centroid change becomes into small value. The result as shown as Table 7. The final results are shown as Table 8, where the final clustering results show that after the convergence iteration, each data point is assigned to a cluster based on its proximity to the nearest centroid.

Figure 4 shows the results of K-means clustering on PCA data with two principal components (PC1 and PC2). The colored dots represent data in different clusters (Clusters 0, 1, and 2). Different dot shapes (round, square, and diamond) are used to distinguish each cluster. The cluster centroids are shown with a

large red 'X', indicating the centre of each cluster. This visualization provides a better idea of how the data is distributed in two-dimensional space, as well as how each cluster is centred around the centroid. Future studies may explore to another method to ensure the initial centroid is more dispersed, thus achieving convergence.

Table 4. Initialization of centroid

Cluster	Centroid X	Centroid Y
0	0.5	1.2
1	-1.0	0.7
2	1.5	-1.5

Table 5. Result of determining nearest centroid in initial cluster

Data Point	PC1	PC2	Initial cluster	Nearest centroid
1	-1.868	-1.284	0	1
2	-0.532	-0.943	2	1
3	0.548	1.311	3	0
4	-0.313	1.829	3	0
5	3.082	-0.865	1	2

Table 6. Updated centroid results after initial iteration

Cluster	Centroid X	Centroid Y
0	0.1175	1.570
1	-1.200	-0.943
2	2.315	-0.865

Table 7. The result in converge process

Iteration	Centroid 0 (X, Y)	Centroid 1 (X, Y)	Centroid 2 (X, Y)	After changing
1	(0.1175, 1.570)	(-1.200, -0.943)	(2.315, -0.865)	0.35
2	(0.1200, 1.568)	(-1.180, -0.950)	(2.320, -0.860)	0.02
...
Final	(0.1210, 1.567)	(-1.178, -0.951)	(2.322, -0.859)	<0.01

Table 8. The result in final clustering

Data point	PC1	PC2	Final cluster
1	-1.868	-1.284	1
2	-0.532	-0.943	1
3	0.548	1.311	0
4	-0.313	1.829	0
5	3.082	-0.865	2

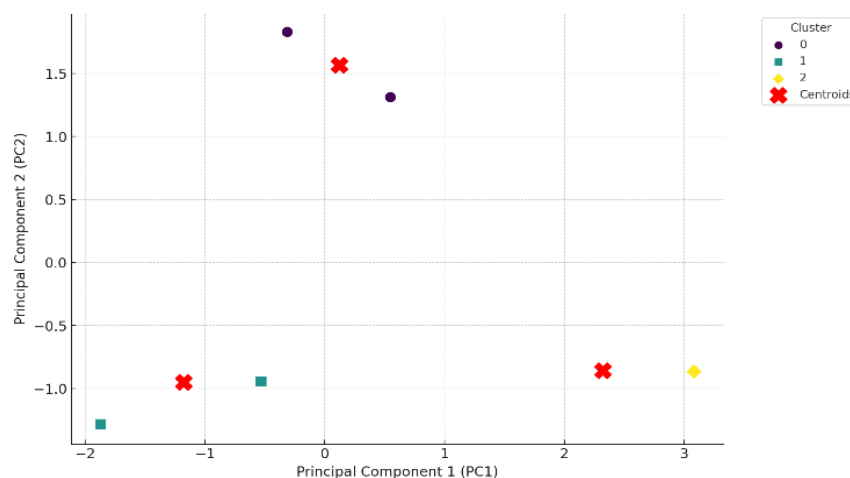


Figure 4. Enhanced K-means clustering visualization with PCA components

3.4. Distribution of clustering

Then, we obtained four cluster based on the distribution of the data feature. Cluster 0, high spending score (0.5176), but age and annual income are close to the average. Customers in this cluster tend to have high spending with above-average product prices (unit price: 0.342). Cluster 1, low spending score (-0.5409) with above average income and age. Customers in this cluster are more careful in shopping or focus on lower-priced products. Cluster 2, very high quantity (1.1958) and highest total price (1.9570). These customers tend to buy in large quantities and focus on higher priced products. Cluster 3, very low quantity (-1.0293) and low total price (-1.22), indicating customers with low purchase frequency and a tighter budget. Table 9 helps in determining specific marketing strategies for each customer segment. PCA cluster distribution helps to reduce the complexity of the original data into two dimensions. The results can be seen in Table 10.

Table 9. Average profile for each cluster based on features

Cluster	Age	Annual income	Spending score	Purchase history	Quantity	Unit price	Total price
0	-0.52	-0.04	0.5176	-0.0910	-0.3070	0.342	-0.08
1	0.51	0.03	-0.5409	0.1039	0.3443	-0.25	-0.08
2	0.01	-0.10	0.0217	0.0266	1.1958	12.26	19.57
3	0.00	0.10	0.0278	-0.0467	-1.0293	-11.5	-1.22

Table 10. K-means clustering results on PCA data with four clusters

PC1	PC2	Cluster
-1.868.169	-1.283.976	0
-0.531638	-0.943265	2
0.548007	1.311.395	3
-0.312829	1.828.949	3
3.081.836	-0.865296	1

We obtained data distribution by clustering. Figure 5 shows the number of products per category in each cluster, providing an overview of the product category preferences such as electronics, fashion, and groceries among the clusters. Figure 6 illustrates the most commonly used payment methods in each cluster, such as credit card, debit card, and PayPal. This helps understand the dominant payment methods among customers in a particular cluster. Figure 7 shows the average total price per product category in each cluster. This visualization provides insight into the price trends and product preferences by category in each cluster. Figure 8 illustrates the distribution of transaction IDs against transaction dates in each cluster. This visualization provides information about transaction date and frequency of transactions occur among the clusters.

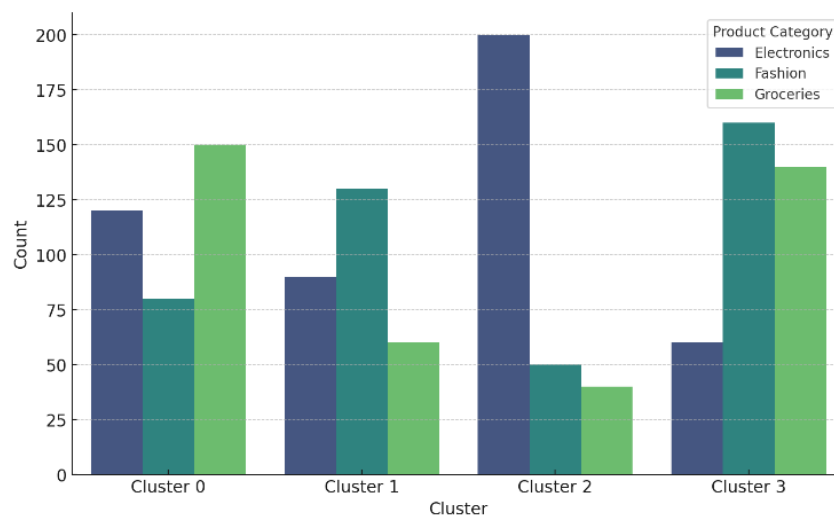


Figure 5. Distribution of product category by clustering

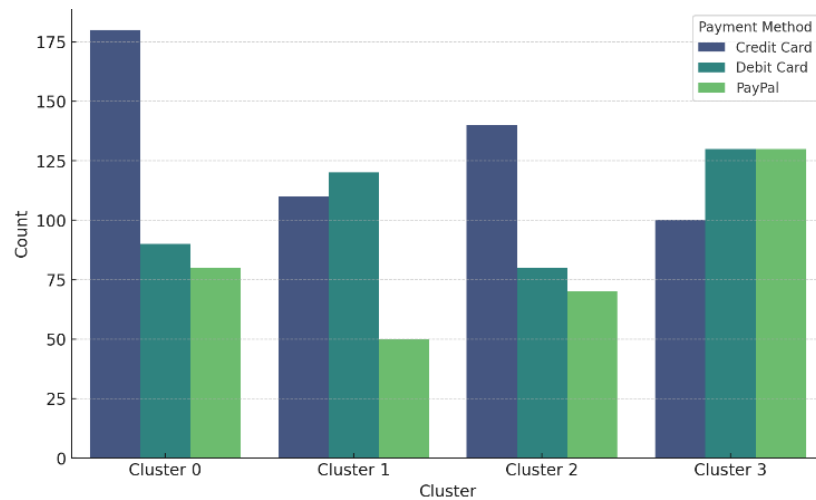


Figure 6. Distribution of payment method by clustering

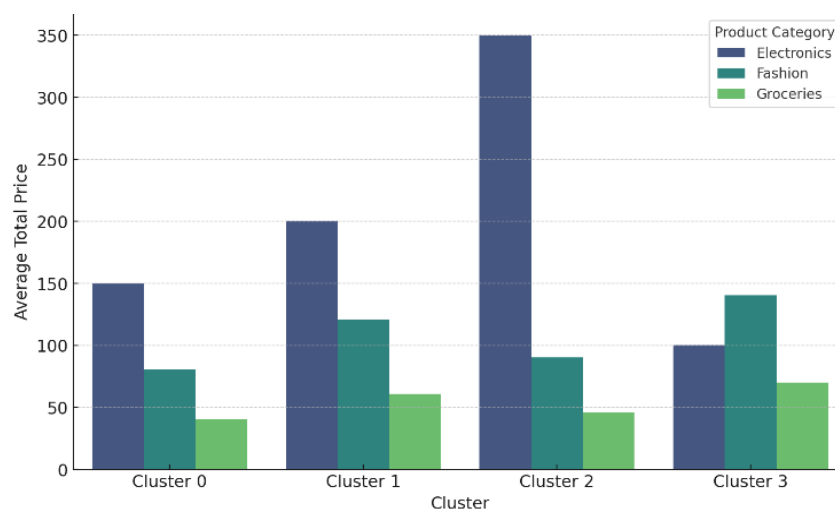


Figure 7. Distribution of product category and average total price by clustering

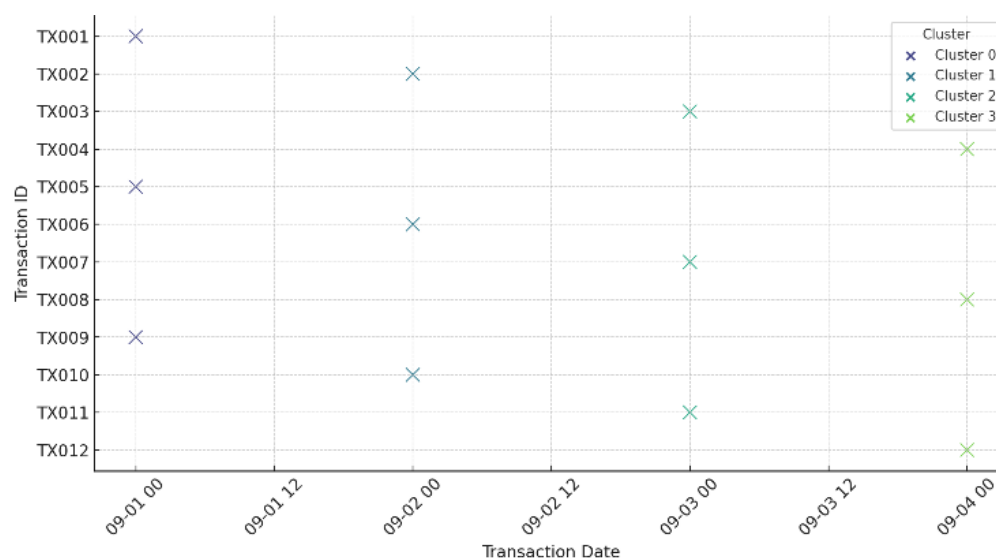


Figure 8. Transaction date and transaction ID clustering

3.5. Interpretation of clustering result

Understanding the characteristics that describe each cluster is crucial for interpreting results of clustering analyses, such as those produced by K-means. Table 11 provide information to understand the relationship between among average PC 1 and PC2 to most common product category and most common payment. Tables 12 and 13 provide information to understand cluster characteristics, business insights, and relevant strategy recommendations. The tables provide clustering results, unique characteristics of each cluster, and how these results can be used for more effective marketing strategies.

Table 11. Profile cluster analysis

Cluster	Average PC1	Average PC2	Most common product category	Most common payment method
0	0.1200	1.568	Electronics	Credit card
1	-1.180	-0.950	Fashion	Debit card
2	2.320	-0.860	Groceries	PayPal

Table 12. Business insight from clusters

Cluster	Business Insights
0	Customers tend to buy electronic products and use credit cards. Marketing strategy: discount offers for electronics and credit card promotions.
1	Customer have a preference for fashion and use debit cards. Strategy: campaign exclusive offers for fashion products with debit card instalments.
2	Customer tend to buy wholesale product purchases. Strategy: offer subscription promotions or large purchase discounts.

Table 13. Marketing strategies recommendation from clusters

Cluster	Marketing Strategies
0	Provide special discounts on electronic purchases with credit cards. Launch cashback program or reward points campaigns for credit card users.
1	Provide free instalments or additional discounts on fashion purchases for customer who using debit card. Create a loyalty programme for fashion product.
2	Provide a subscription program to buyer who buy wholesale product. Offer special discounts for bulk or recurring purchases.

4. CONCLUSION

In this study demonstrates the significant of role ML models may convert unstructured consumer data into useful insights. This model that helps businesses stay competitive in a market. We implemented PCA analysis in dimension reducing, using k-means clustering, then evaluate by using inertia and silhouette score with 1.455.650 and 0.486366 respectively. K-means clustering is success to distinct customer into several groups based on preference and history transaction, which enables more accurate marketing allocation, which enhances customer engagement, and maximizes sales performance. Future studies may investigate the use of larger and more diverse datasets, incorporating additional factors like long-term customer behavior or product preferences. In addition, using the K-means ++ approach is expected to provide the advantage of selecting an ideal centroid at the initial iteration and speeding up the convergence process.

FUNDING INFORMATION

This research was financially supported by Universitas Sumatera Utara through a research grant under the Grant Number 3641/UN5.2.14.D/PT.01.03/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Fanindia Purnamasari	✓	✓		✓	✓	✓		✓		✓			✓	
Umayra Ramadhani		✓	✓	✓		✓		✓	✓			✓		
Putri Nasution														
Marischa Elveny	✓		✓	✓			✓			✓	✓		✓	

C : C onceptualization	I : I nterpretation	Vi : V isualization
M : M ethodology	R : R esources	Su : S upervision
So : S oftware	D : D ata Curation	P : P roject administration
Va : V alidation	O : Writing - O riginal Draft	Fu : F unding acquisition
Fo : F ormal analysis	E : Writing - Review & E diting	

CONFLICT OF INTEREST STATEMENT

The authors of this paper disclose any financial, personal, or professional relationships that might present a conflict of interest in order to promote impartial and equitable decision-making. Competing political, personal, religious, ideological, academic, or intellectual interests are examples of non-financial competing interests. The authors hereby certify that none of the work described in this publication could have been influenced by any known competing financial interests or personal relationships. The authors therefore state that they have no conflicts of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals involved in this study.

ETHICAL APPROVAL

Human use research has been authorized by the authors' institutional review board or comparable committee and has adhered with all applicable national rules and institutional procedures in line with the Declaration of Helsinki's tenets.

DATA AVAILABILITY

The data that support the findings of this study are available from the lead author on reasonable request. For further questions or data access, please contact the lead author.




REFERENCES

- [1] S. Dwivedi and A. Singh, "A study of customer segmentation based on RFM analysis and K-means," in *Lecture Notes in Networks and Systems*, vol. 731 LNNS, 2024, pp. 347–355.
- [2] S. Arefin *et al.*, "Retail industry analytics: unraveling consumer behavior through RFM segmentation and machine learning," in *2024 IEEE International Conference on Electro Information Technology (eIT)*, May 2024, pp. 545–551, doi: 10.1109/eIT60633.2024.10609927.
- [3] D. Jaiswal, V. Kaushal, P. K. Singh, and A. Biswas, "Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market," *Benchmarking: An International Journal*, vol. 28, no. 3, pp. 792–812, Mar. 2021, doi: 10.1108/BIJ-05-2020-0247.
- [4] R. Sethuraman, J. C. Gázquez-Abad, and F. J. Martínez-López, "The effect of retail assortment size on perceptions, choice, and sales: review and research directions," *Journal of Retailing*, vol. 98, no. 1, pp. 24–45, Mar. 2022, doi: 10.1016/j.jretai.2022.01.001.
- [5] M. Alkhayrat, M. Aljnnidi, and K. Aljoumaa, "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA," *Journal of Big Data*, vol. 7, no. 1, p. 9, Dec. 2020, doi: 10.1186/s40537-020-0286-0.
- [6] J. M. John, O. Shobayo, and B. Ogunleye, "An exploration of clustering algorithms for customer segmentation in the UK retail market," *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.
- [7] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4995–5005, Mar. 2024, doi: 10.1007/s00521-023-09339-6.
- [8] R. Punhani, V. P. S. Arora, S. Sabitha, and V. Kumar Shukla, "Application of clustering algorithm for effective customer segmentation in e-commerce," in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Mar. 2021, pp. 149–154, doi: 10.1109/ICCIKE51210.2021.9410713.
- [9] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.
- [10] J. M. Spoor, "Improving customer segmentation via classification of key accounts as outliers," *Journal of Marketing Analytics*, vol. 11, no. 4, pp. 747–760, Dec. 2023, doi: 10.1057/s41270-022-00185-4.
- [11] S. Tasoulis, N. G. Pavlidis, and T. Roos, "Nonlinear dimensionality reduction for clustering," *Pattern Recognition*, vol. 107, p. 107508, Nov. 2020, doi: 10.1016/j.patcog.2020.107508.
- [12] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 1, pp. 56–70, May 2020, doi: 10.38094/jastt1224.
- [13] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: constant-size coresets for K-means, PCA, and projective clustering," *SIAM Journal on Computing*, vol. 49, no. 3, pp. 601–657, Jan. 2020, doi: 10.1137/18M1209854.
- [14] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – an effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, Dec. 2021, doi: 10.1016/j.jksuci.2018.09.004.




- [15] S. S. Ling, C. W. Too, W. Y. Wong, and M. H. Hoo, "Customer relationship management system for retail stores using unsupervised clustering algorithms with RFM modeling for customer segmentation," in *2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, May 2024, pp. 1–6, doi: 10.1109/ISCAIE61308.2024.10576353.
- [16] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-means algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.
- [17] G. Aslantaş, M. Gençgöl, M. Rumelli, M. Özsera, and G. Bakirli, "Customer segmentation using K-means clustering algorithm and RFM model," *Deu Muhendislik Fakültesi Fen ve Muhendislik*, vol. 25, no. 74, pp. 491–503, May 2023, doi: 10.21205/deufmd.2023257418.
- [18] M. A. Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.
- [19] C. S. Vamsee, D. Rakesh, I. Prathyusha, B. Dinesh, and C. Bharathi, "Demographic and psychographic customer segmentation for ecommerce applications," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May 2023, pp. 615–622, doi: 10.1109/ICAAIC56838.2023.10140861.
- [20] A. Wasilewski, "Customer segmentation in e-commerce: a context-aware quality model for comparing clustering algorithms," *Journal of Internet Services and Applications*, vol. 15, no. 1, pp. 160–178, Jul. 2024, doi: 10.5753/jisa.2024.3851.
- [21] A. de Sousa, S. Moro, and R. Pereira, "Cluster-based approaches toward developing a customer loyalty program in a private security company," *Applied Sciences*, vol. 14, no. 1, p. 78, Dec. 2023, doi: 10.3390/app14010078.
- [22] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, Jul. 2020, doi: 10.1016/j.inffus.2020.01.005.
- [23] P. Patel, B. Sivaiah, and R. Patel, "Approaches for finding optimal number of clusters using K-means and agglomerative hierarchical clustering techniques," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCCSP)*, Jul. 2022, pp. 1–6, doi: 10.1109/ICICCCSP53532.2022.9862439.
- [24] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
- [25] Y.-F. Tan, G.-Z. Zhao, C.-P. Ooi, and W.-H. Tan, "Leveraging interquartile range and isolation forest for abnormal power consumption prediction," in *2024 IEEE 6th Advanced Information Management, Communication, Electronic and Automation Control Conference (IMCEC)*, May 2024, pp. 815–819, doi: 10.1109/IMCEC59810.2024.10575711.
- [26] Y.-C. Wang, "Prediction of engine failure time using principal component analysis, categorical regression tree, and back propagation network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 11, pp. 14531–14539, Nov. 2023, doi: 10.1007/s12652-018-0997-7.

BIOGRAPHIES OF AUTHORS






Fanindia Purnamasari    is a lecturer at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara. She received her master's degree in master of information technology from Universiti Kebangsaan Malaysia in 2017 and bachelor's degree in information technology from Universitas Sumatera Utara in 2013. Her research interests are intelligent system, data science, and artificial intelligence. She can be contacted at email: fanindia@usu.ac.id.



Umay Ramadhani Putri Nasution    studied undergraduate education in 2009 at the Universitas Sumatera Utara and earned her M.Kom (master) degree in 2017 in informatics engineering, Universitas Sumatera Utara. During her undergraduate studies, she participated in campus organizations such as the HIMATIF Administration and Sekretariat Coordinator from 2011 to 2012 and participated in the Education membership at UKMI Al Falaq MIPA from 2010 to 2011. In February 2021, she became a lecturer at Universitas Sumatera Utara. She also joined the Information Systems Center team as the person in charge of the official USU website with her colleagues as a programmer. Her research field is related to the fields of machine learning and data analytics. She can be contacted at email: umaya.nst@usu.ac.id.



Marischa Elveny    received (bachelor's) degree in information technology at the Universitas Sumatera Utara and earned her M.Kom (master) degree in informatics engineering, Universitas Sumatera Utara. Doctorate (Ph.D.) at the Universitas Sumatera Utara. Currently working as a lecturer in the Faculty of Computer Science and Information Technology at the Universitas Sumatera Utara. Her research interests are artificial intelligence, data science, and computational intelligence. She can be contacted at email: marischaelveny@usu.ac.id.