Market Segmentation via K-Means Algorithm and RFM Analysis (Case Study: Microbusiness Sales Transactions)

Diva Anggelina[1*], Jimmy Tjen[2]

[1] Department of Digital Business, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
[2]Department of Informatics, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
*Email: 22430153@widyadharma.ac.id [1], jimmy.tjen@mathmods.eu [2]

Abstract. The increasing number of microbusinesses that are emerging fosters intense competition across various industries. In such a competitive environment, business owners must implement effective strategies to ensure the success and sustainability of their businesses. Understanding customers plays a crucial role in formulating effective marketing strategies, as it allows businesses to accommodate the unique needs and preferences of their target customers. One strategy that has proven to be highly effective in achieving this is customer segmentation. Customer segmentation helps business owners understand their customers by grouping them based on shared characteristics. By categorizing customers into distinct segments, businesses can gain deeper insights into their customer base and create a better strategies to enhance customer satisfaction and loyalty. K-Means is a frequently used method for clustering and is often combined with RFM (Recency, Frequency, Monetary) analysis to assess and classify customer behavior. This paper focuses on segmenting customers based on sales transaction data related to weaving product variations. By employing the K-Means with RFM analysis, three customer segments were identified: segment one, segment two, and segment three. By analyzing the characteristics of the three customer segments, business owners can design and implement targeted marketing strategies tailored to the specific needs and preferences of each segment.

Keywords: customer segmentation, K-Means Clustering, market segmentation, RFM Analysis

Introduction

The number of microbusinesses in Indonesia has increased over the years, indicating a growing public interest in entrepreneurship and small enterprises. Data from the Central Statistics show that the number of micro businesses in Indonesia was 4.181.128 in 2023, an increase from the previous years. This expansion demonstrates the vital role microbusinesses play in promoting community development, job creation, and economic resilience. however in this competitive business, understand customers is crucial to winning strategies that will eventually boost the business profitability.

The customers who buy our products are highly diverse, with differences in age, income, occupation, and shopping behavior (Shirole, Salokhe, & Jadhav, 2021) This diversity encourages segmentation, customer segmentation helps grouping customers into segments based on the same characteristics, enabling business owners to better understand their customer base and create the right business strategy (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021). The most commonly used method for customer segmentation by researchers is K-Means clustering. Many researcher use K-Means algorithm for segmentation including (Nandapala & Jayasena, 2020) (Wu, Yau, Ong, & Chong, 2021) (Mu, Chu, Tian, Feng, & Weisong, 2021).

This research focuses on segmenting customers of a microbusiness that sells various woven products using K-Means method combined with RFM analysis. The most commonly used behaviour segmentation model is the RFM model (Ernawati, Baharin, & Kasmin, 2021). Combining the K-Means algorithm with RFM analysis

achieves a high accuracy of 95% demonstrating the efficiency and effectiveness of this method for customer segmentation (Sarkar, Puja, & Chowdhury, 2024)

Previous research by (Wu, et al., 2020) combined RFM with K-Means using online transaction data, resulting in four customer groups. (Shirole, Salokhe, & Jadhav, 2021) used data from a UK online retail store to identify four clusters based on customer characteristics. The results helped the company target their customers and maintaining good relationships with customers. (Xian, Keikhosrokiani, XinYing, & Li, 2022) used historical sales data, and their results can provide an effective marketing mix and recommending specific products to both new and existing customer segments. (Akande, Akande, Asani, & Dautare, 2024) applied K-Means clustering to group customers into categories and approach highly accuracy and precision.

This study can serve as a reference for small businesses in determining customer segments and implementing appropriate marketing strategies to help empower microbusinesses to survive and thrive in a competitive market. It also emphasize the integration of innovation in efficient business management and sustainability to supporting and achievement of the sustainable development goals (SDGs).

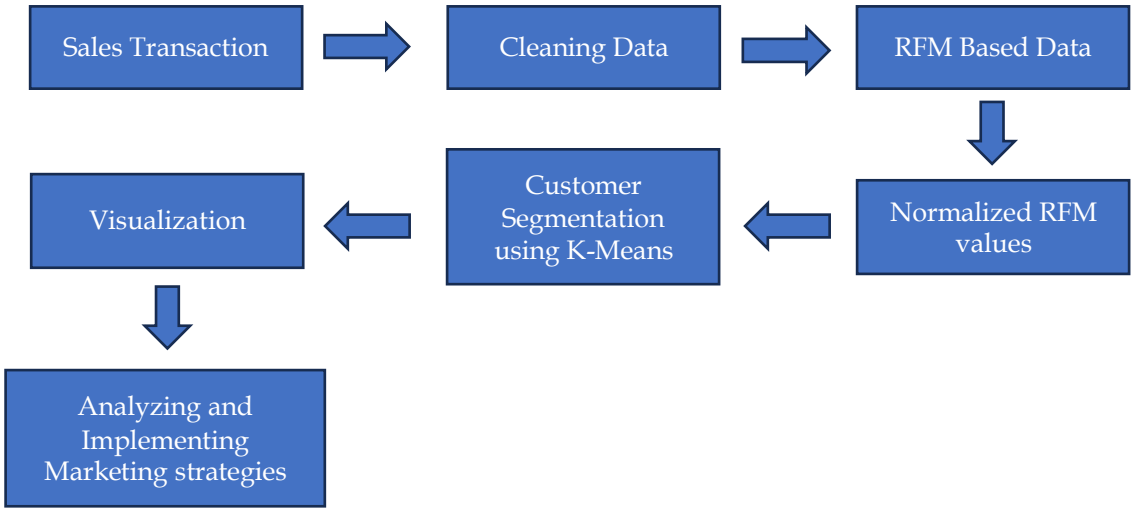
This paper is organized into the following section: section II presents the methods used in this study, including the dataset, data preparation, RFM model, the application of the K-Means clustering, and visualization into 2D graph, section III result and finding, focusing to analyzing the segments and providing marketing strategies for each segment. The final section is conclusion, which highlights the contributions of this research to businesses, especially microbusinesses, by helping them understand their customers and implement the right strategies for each customer segment.

Methods

DATA UNDERSTANDING

The data used in this research consists of sales transaction from microbusiness brand "*Arti Borneo*", which specialized in selling of various combinations of "*Tenun Ikat Sintang*," a traditional woven fabric of the *Sintang Dayak* community. The dataset includes 197 sales transactions collected from January 17, 2024 to November 5, 2024, with a focus on a sample of 45 transactions used for analysis.

DATA PREPARATION



Cleaning Data

The data used in this research were collected through manual spreadsheet recordings. In this phase duplicate and negative transaction data were removed, missing values were added, and the raw dataset was organized.

RFM Model

The primary metrics in this research are the RFM (Recency, Frequency, and Monetary) parameters, which are used for customer segmentation. The three of RFM are as follows:

1. Recency (R) : calculate time since the customers last purchase during the research period.
2. Frequency (F) : counts the number of purchases made by each customer during the research period.
3. Monetary (M) : indicates the total amount spent by the customer during the research period.

Table 2: some of the RFM values

Customer_id	Recency	Frequency	Monetary
A001	293	1	Rp 1.100.000
A003	292	2	Rp 143.000
A004	144	2	Rp 4.600.000
A005	282	2	Rp 290.000
A006	206	3	Rp 195.000
A007	260	1	Rp 275.000
A008	196	2	Rp 750.000
A009	232	1	Rp 70.000
A010	206	1	Rp 150.000

RFM Normalization

This step is normalizes the RFM values to a 0-1 scale. Using the following formula for normalization:

$$X_{normalized} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Table 2: some of the RFM normalizes values

R	F	M
1.0000	0	0.2177
0.9960	0.1111	0.0257
0.4064	0.1111	0.9198
0.9562	0.1111	0.0552
0.6534	0.2222	0.0361
0.8685	0	0.0522
0.6135	0.1111	0.1474
0.7570	0	0.0110
0.6534	0	0.0271

MODELING

In this research, we use K-Means clustering method, with the following formula:

$$Ci = {}^{1}/_{M} \sum j = 1^{m} \text{ Xj} \quad (2)$$

Where:

Ci: the centroid of cluster i

M: the total number of point i

j: indexes each data point in the cluster

Formula for the Euclidean distance:

$$\mathrm{d}(p,q) = \sqrt{(p1 - q1)^2} + (p2 - q2)^2 \quad (3)$$

Where:

P: the coordinates of the first point

q: the coordinates of the second point

The first step in clustering is to determine the number of clusters. To optimize this number, we use silhouette analysis. For a single data point i , the silhouette score is defined as follows:

$$\mathrm{s}(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (4)$$

Where:

$s(i)$: represents the silhouette coefficient for the data point i

a(i): represents the average distance between point i and all other points in the same cluster

b(i): represents the average distance between point i and all point in the nearest cluster

Average silhouette score for k:

$$S_k = \frac{1}{N}\sum_{i=1}^{N} s(i) \quad (5)$$

Where:

N: The total number of data points

$s(i)$: the silhouette score for data point i

$$k_{optimal} = \arg\max_k S_k \quad (6)$$

(6)The optimal number of clusters is the k that produces the highest average silhouette score(S_k)

After determining the optional k, the K-Means algorithm was applied to segment the customers into clusters based on their RFM values.

VISUALIZATION

To visualize the clustering results, a 2D cluster graph were created, displaying the customers RFM values and their cluster assignments. These visualizations provide a clear representation of how the customers are grouped based on their purchase behavior. Finally, the results were interpreted in the context of customer segmentation.

Result and Finding

This is the silhouette graph, indicating that the optimal number of clusters is 3

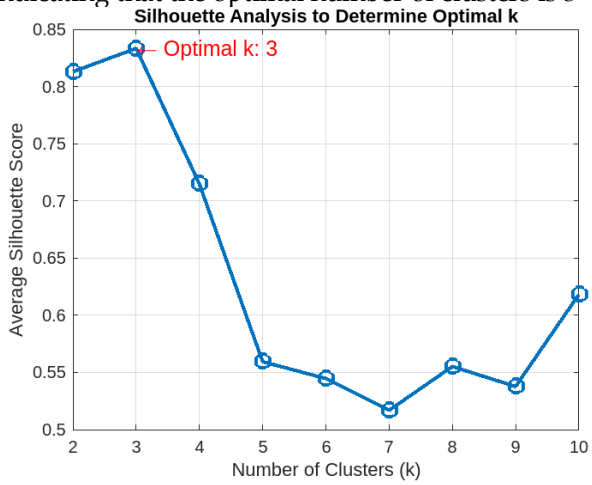
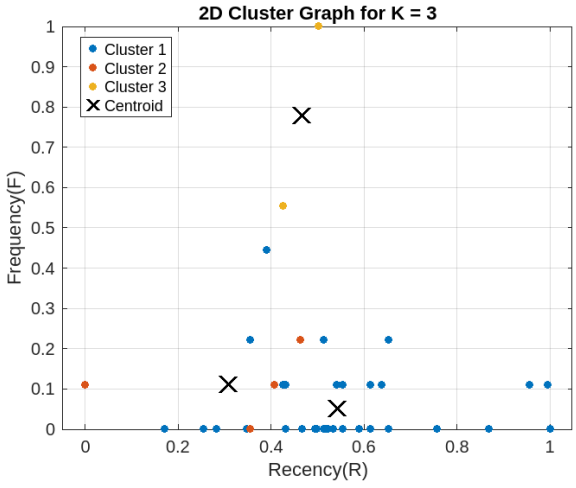
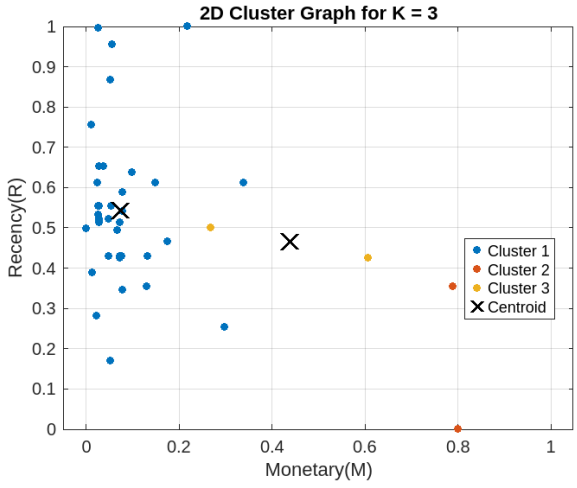


Figure 1. Graph of the optimal number with silhouette method





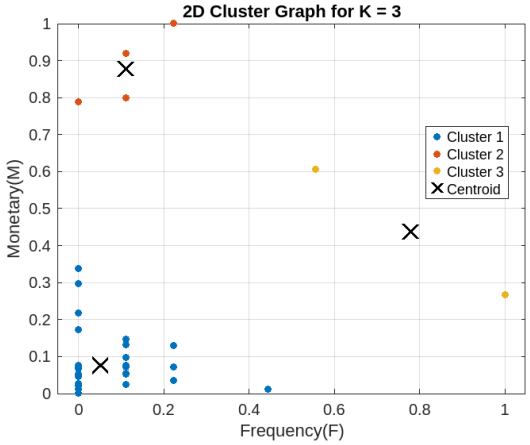


Figure 2. 2D Cluster Graph

Figure 2 shows three customer segments: segment one is located at centroid one, segment two at centroid two segment three at centroid three. The R values is derived from the time since the customer's last purchase. The higher the R value, the more it indicates that the customer has made a recent purchase and is still engaged with the business. This customer is a potential candidate for various promotions aimed at encouraging continued transactions and fostering loyalty. The F value represents the frequency of purchases made by a customer. The higher the F value, the more frequently the customer buys from the business, indicating a strong purchasing behavior. The M values is based on the total amount of money a customer spends during the period. The higher the M values, the more the customer has spent, indicating a higher level of investment in the business.

RESULT ANALYSIS

From the graphs in Figure 2, we can identify three customer segments:

1. Segment one : customers in this segment have an average to low F value, ranging from 0.2 and 1 on the R value scale, and average to low M value. This segment consists of customers who purchase infrequently and have a lower overall spend. The recency varies, meaning some customers in this group may have bought recently, while others have not purchased in a while.
2. Segment two : customers in this segment have a high M value, a low F value, and an average to low R value. This segment consists of customers who spend a lot when they make a purchase but buy infrequently and may not have purchased recently. They show potential to be high-value customers, but their low frequency means they are not engaged enough to become loyal or frequent buyers.
3. Segment three : customer in this segment have an average to high F value, an average R value, and an average M value. This segment consists of customers who buy regularly, are fairly engaged, and spend moderately. They appear to be consistent, middle-tier customers who not exhibit extreme behaviors.

Marketing Strategies for three of the segments:

1. Segment one: This segment consists of customers who have a lower overall spend, purchase infrequently, and show varied recency. Since they are lower spenders, offering them with cost saving strategies such as promotions, discounts, and product bundling to make them spend more on affordable products. To enhance their shopping experience, a simple loyalty program can be implemented, such as awarding points that can be collected and redeemed for gifts, discounts, or price cuts.
2. Segment two: since they are high spenders, they can be targeted with personalized offers that highlight the value of their previous purchases. Providing them with VIP access to new products or special events can make them feel more valued. Exclusive promotions or time-limited offers can also encourage them to make repeat purchases and engage more frequently. As they do not seem overly price-sensitive, offering high-quality products can be an effective strategy.

3. Segment three: Since these customers are somewhat engaged, they show potential for further growth. To maintain their interest, engage them with a variety of new products and special promotions. Utilize email marketing campaigns or remarketing ads to reach out to this segment with relevant product and special offers that align with their purchasing patterns.

Conclusion

Based on the research conducted, it can be concluded that RFM Analysis and the K-Means Algorithm were used to segment customers in a microbusiness selling various types of *Tenun Ikat Sintang* over nearly a year resulting in three customer segments. These segments have been analysed, providing strategies base on the characteristic of each segment. In this paper we also determined the optimal number of centroids, resulting in three clusters. The customer RFM graph has been displayed and analyzed, and marketing strategies for each segment has been carried out.

The results of this research will provided to business owners to help them implement the proposed marketing strategies. Additionally, this research advances our understanding of customer segmentation for traditional products, such as *Tenun Ikat Sintang*. Since this research uses data collected manually, it demonstrates that even simple data sales from small businesses can be analyzed to generate effective marketing strategies. This enables small business owners to compete by implementing the right strategies, ultimately helping microbusinesses grow and develop.

References

Nandapala, E., & Jayasena, K. (2020). The practical approach in Customers segmentation by using the K-Means Algorithm. *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, 344--349.

Akande, O. N., Akande, H. B., Asani, E. O., & Dautare, B. T. (2024). Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, 1-8.

Mu, Y. L., Chu, X., Tian, D., Feng, J., & Weisong. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing, 113*, 107924.

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using rfm model and k-means clustering. *Int. J. Sci. Res. Sci. Technol, 8*, 591--597.

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). [Retracted] An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering, 2020*, 8884227.

Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access, 9*, 62118-62136.

Xian, Z., Keikhosrokiani, P., XinYing, C., & Li, Z. (2022). An RFM model using K-means clustering to improve customer segmentation and product recommendation. *Handbook of Research on Consumer Behavior Change and Data Analytics in the Socio-Digital Era*, 124--145.

Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies, 6*, 54--60.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking ,Äì An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences, 33*, 1251-125.

Ernawati, E., Baharin, S. S., & Kasmin, F. (2021, apr). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series, 1869*, 012085.