Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 5 No. 1 (2025) 12 - 21 E-ISSN: 2809-8064

Integration of RFM Method and K-Means Clustering for Customer Segmentation Effectiveness

Nafissatus Zahro[1*], Nadia Annisa Maori [2], Gentur Wahyu Nyipto Wibowo [3]

[1*,2,3] Informatics Engineering, Sains and Technology, Universitas Islam Nahdlatul Ulama Jepara

[1*]1211240001137@unisnu.ac.id, [2]nadia@unisnu.ac.id, [3]gentur@unisnu.ac.id

Abstract

This research integrates the RFM (Recency, Frequency, Monetary) method with the K-Means Clustering algorithm to segment customers based on transaction data. The RFM method is used to evaluate customer behavior based on the time of the last transaction, purchase frequency, and total transaction value, while K-Means divides customers into three optimal clusters based on similar characteristics. The analysis results show that the first cluster (59% of customers) consists of customers with low transaction activity, the second cluster (41% of customers) includes customers with moderate activity and medium transaction value contribution, while the third cluster (1% of customers) contains high-value customers with significant transaction frequency and value. The third cluster provides strategic opportunities for the development of loyalty programs, while the other clusters require specific strategies to enhance customer activity and retention. The integration of these two methods has proven effective in supporting more targeted and strategic data-driven customer segmentation.

Keywords: *Customer Segmentation, RFM, K-Means Clustering, Data Analysis, Marketing Strategy, Silhouette Score*

1. Introduction

Bisnis In the current digital era, information technology has influenced various aspects of human life, including the business world[1]. In the tight business competition, companies are required to continuously innovate by adopting the right strategies, including the use of data mining technology to analyze customer data [2]. Customer segmentation becomes an important step in understanding customer preferences and behaviors, while also supporting product development, reducing the risk of marketing failures, and enabling the identification of more homogeneous customer groups based on their transaction patterns [3].

Data Mining is used to find desired patterns for the purpose of extracting useful information from large databases. These patterns are identified using tools that can provide valuable and in-depth data analysis, which can be further pursued using other decision support tools. Data Mining is one of the most common techniques in KDD, but it is a very important technique for finding meaningful patterns in large data sets [4].

Data Mining is used to find desired patterns for the purpose of extracting useful information from large databases. These patterns are identified using tools that can provide valuable and in-depth data analysis, which can be further pursued using other decision support tools. Data Mining is one of the most common techniques in KDD, but it is a very important technique for finding meaningful patterns in large data sets [5].

The methods in data mining are[6]:

1. Estimation To estimate an unknown value, such as predicting someone's income when information about that person is known. Examples of the methods are Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM).
2. Prediction To estimate future values, such as predicting stock levels one year ahead. Examples of the methods are Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM).
3. Classification It is the process of discovering a model or function that explains or distinguishes concepts or data classes, with the aim of being able to estimate the class of an object whose label is unknown. Examples of methods include Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, Adaptive Credal C4.5), Naive Bayes (NB), K-

Received: 20-11-2024 | Accepted: 22-01-2025 | Published: 06-02-2025

Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Logistic Regression (LogR).

4. Clustering That is, clustering identifies data that has certain characteristics. Examples of the methods are K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means (FCM).
5. Association, also known as market basket analysis, is a function that identifies product items that are likely to be purchased by consumers together with other products. Examples of the methods are FP-Growth, A Priori, Coefficient of Correlation, Chi Square.

One of the relevant techniques in data mining is the RFM (Recency, Frequency, Monetary) model, which helps companies analyze customer behavior based on three main parameters: the recency of the last transaction, transaction frequency, and the amount of money spent by customers. This method is effective in segmenting customers based on their purchasing patterns, providing a strong foundation for more targeted decision-making[6].

However, although the RFM method is useful, this technique often lacks flexibility in generating highly specific segmentation. As a complement, the K-Means Clustering algorithm is used to group data into several clusters based on certain characteristic similarities. This algorithm excels in providing significant cluster visualizations, but it requires well-structured initial data to produce optimal clusters [7][8]. The integration between the RFM method and K-Means offers the potential to address the shortcomings of each method. In this integration, customer behavior-based data generated by RFM becomes the input used by the K-Means algorithm to create more precise and in-depth customer clusters.

In the research by Satria Ardi Perdana, Sara Famayla Florentin, and Agus Santoso, customer segmentation was conducted using the K-Means clustering algorithm to group Alfagift application customers based on five categories: age, gender, purchase frequency, payment type, and city. This segmentation process follows the CRISP-DM (Cross Industry Standard Process for Data Mining) method, which consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The Elbow Method is used to determine the optimal number of clusters based on the calculation of SSE (Sum Squared Error), with the optimal k being three clusters. The first cluster consists of 7,219 customers, the second cluster has 6,902 customers, and the third cluster has 5,371 customers. The results of this clustering are then interpreted to support more precise and efficient marketing strategy decision-making[11].

In this study, the combination of RFM techniques and K-Means Clustering is expected to produce more systematic and in-depth customer segmentation for data analysis purposes. The main focus of this research is to evaluate the effectiveness of integrating the RFM method with K-Means Clustering in improving the accuracy of customer segmentation. The combination of these two methods is designed to analyze customer purchasing behavior in more detail, thereby facilitating the process of identifying customer patterns. The Silhouette Score is used in this study to determine the optimal value in the segmentation performed. The RFM (Recency, Frequency, Monetary) method was chosen due to its advantages in measuring customer behavior based on three main parameters: the time of the last transaction, transaction frequency, and transaction amount. On the other hand, K-Means Clustering is an effective algorithm for grouping data into several clusters based on characteristic similarities. By combining RFM techniques and K-Means Clustering, this research aims to evaluate how the integration of these two methods can enhance the quality of customer segmentation while providing deeper insights into customer behavior patterns. Based on these objectives, the main question posed in this research is: How can the RFM method and K-Means Clustering be integrated for customer segmentation?

2. Research Methods

However, in this analysis, the selection of the optimal number of clusters is crucial to ensure the quality of the resulting segmentation. For that reason, the Silhouette Coefficient is used as an evaluation method to assess the strength and quality of the formed clusters. The Silhouette Coefficient provides a value range between -1 and 1, where a value close to 1 indicates that the formed cluster is very good, while a value close to -1 indicates a poor cluster, where objects in the cluster are more similar to objects in other clusters. Therefore, the Silhouette Coefficient helps in determining the most optimal number of clusters based on the evaluation of the quality of the formed clusters. [9]. This research method begins with the START phase, followed by Data Collection to gather relevant data. After that, the Data Cleaning stage is carried out to clean the data from missing or inconsistent values. Next, in the Data Transformation stage, the cleaned data will be transformed and processed using the RFM (Recency, Frequency, Monetary) model to identify customer behavior patterns.

After the transformation stage, the Data Mining/Clustering stage is conducted, where data clustering techniques such as K-Means are used to group data into clusters based on similarity of characteristics. Then, Pattern Evaluation is conducted to assess the

quality of the formed clusters using the Silhouette Coefficient as a measure of cluster strength. The next step is Knowledge Representation, where the clustering results are visualized and further analyzed. The research process concludes with the END stage.

Figure 1 shows the flowchart of the entire research process.

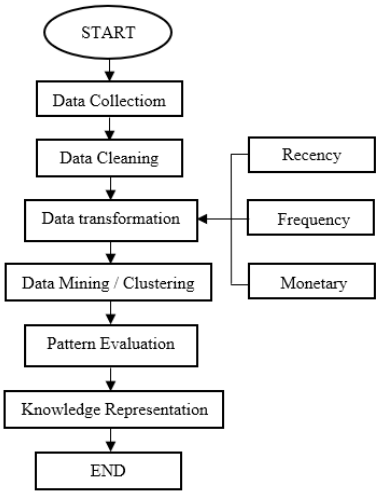


Figure 1. Flowchart Research

2.1. *Data Collection*

At this stage, data is collected through interviews and observations. The data used for customer segmentation is the furniture sales transaction data from ADAPTA.Id in Jepara in 2022, with a total of 2,252 transactions from 928 customers. After data collection, the selection of relevant attributes was carried out from the initial 29 attributes.

Table 1. initial sample data

Username (pembeli)	Waktu Pesanan Dibuat	Total Payment	…	No. Telepon
Mustova15	2022-01-02	74.160	…	******78
Srengenge30498	2022-01-01	67.980	…	******20
Zayd.babywear	2022-01-01	57.000	…	******86
	……	……	…	******08
……. 3idrisefen	2022-12-30	21.000	…	******62

2.2. *Dataset Selection*

At this stage, the selection of relevant attributes for customer segmentation analysis using the RFM (Recency, Frequency, Monetary) method is carried out. Out of the total 29 attributes available in the 2022 furniture sales transaction dataset, 3 main attributes deemed most relevant for the analysis were selected, namely Username (buyer), Order Creation Time, and Total Payment. These attributes were chosen because they can provide a clear picture of customer behavior in terms of transaction recency, purchase frequency, and the amount of money spent.

Table 2. sample data awal

Username (pembeli)	Order Creation Time	Total Payment
Mustova15	2022-01-02 16:37	74.160
Srengenge30498	2022-01-01 09:52	67.980
Zayd.babywear	2022-01-01 12:56	57.000
…….	……	……
3idrisefen	2022-12-30 17:55	21.000

2.2. *Data Cleaning*

At this initial stage, data cleaning is carried out to ensure that the data used is free from incomplete, empty, or duplicate values. The steps taken include:

1. Deleting rows or columns that have null or empty values.
2. Deleting duplicate data.

2.3. *Data Transformation*

Data transformation is carried out to convert data into a format that can be processed in the data mining process. The steps in this stage include:

1. Conversion of categorical data into numerical data for clustering computation purposes.
2. Aggregation of sales transaction data based on the Recency, Frequency, and Monetary (RFM) model.

This model analyzes customer value and behavior based on.

Recency (R): The time span since the last transaction.

$$Recency = Current\ date - Last\ transaction\ date \quad (1)$$

Frequency (F): The number of transactions in a certain period.

$$Frequency = \frac{Total\ number\ of\ transactions}{Time\ period} \quad (2)$$

Monetary (M): The total value of purchases in a certain period.

$$Monetary = Total\ value\ of\ purchases \quad (3)$$

2.4. *Data Mining / Clustering*

At this stage, data mining techniques are applied to discover patterns or models from the processed data, using the K-Means clustering method. The steps taken in this data mining stage include:

2.4.1 *Normalization.*

Normalization the RFM data is normalized to ensure that all features are on the same scale, which is important for clustering algorithms like K-Means, which are sensitive to data scale.

$$Value\ (RFM) = \frac{Value\ RFM\ of\ data\ \ i}{value\ \ RFM\ max} \quad (4)$$

2.4.2. *Determining the Optimal Number of Clusters*

The Silhouette Coefficient is used to optimize a cluster so that the clustering process can be considered good and optimal [11]. Calculating the Silhouette Coefficient from the known values of $a(xi)$ and $b(xi)$ for each data point.

$$S_{xi} = \frac{(b_{xi} - a_{xi})}{\max\{a_{xi}, b_{xi}\}} \quad (5)$$

Description:

s_{x_i} = Silhouette Value.

a_{x_i} = Average distance between data points within the same cluster.

b_{x_i} = Average distance to the nearest next cluster.

2.4.3. *Analyzing Data with the K-Means Clustering Algorithm*

Analyzing Data with K-Means Clustering Algorithm The data obtained in the data collection phase is processed and analyzed using the K-Means Clustering algorithm. The steps of the algorithm are as follows [12]:

Here are the steps in performing clustering using the K-Means method:

1. Determining the number of clusters to be created.
2. Determining the initial centroid points (cluster center points) for the clustering process.
1. The distance that will be used to determine the data with the cluster center uses Euclidean Distance. In the calculation to determine the distance of the data to each cluster center, Euclidean distance can be used as follows:

$$d(X, C) = \sqrt{\sum_{i=1}^{n}(X_i - C_i)^2} \quad (6)$$

Explanation:

$d(X, C)$: Euclidean distance between data point X and cluster center$C$

X_i: The data value in the to attribute - i

C_i: The cluster center value of the to attribute - i

n: The number of attributes or features in the data.

2. Grouping data into clusters with the minimum distance from each data point to each centroid. Updating the centroid value with the new centroid value obtained based on the average of the data within the cluster.

$$C_j^{(t+1)} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i \quad (2.7)$$

Explanation:

$C_j^{(t+1)}$: New center point (centroid) for the th cluster- j

X_i: Data points that are in the th cluster - j

n_j : Number of data points in the th cluster - j

$\sum_{i=1}^{n_j} X_i$: The sum of all data points in the th cluster - j

3. Repeat steps 3 and 4 until they no longer meet the criteria. Satisfactory criteria can include the number of iterations or changes in the centroid position in successive iterations.

2.8. *Evaluation*

The evaluation process is the final step in the data processing stages, aimed at assessing the quality of the clustering results obtained using the applied algorithm. This evaluation is conducted by calculating the Silhouette Coefficient, which is used to measure the consistency and compactness of the data within each cluster.

To support the evaluation process, Google Colab software is used as the main platform for data processing and analysis. Google Colab enables the application of clustering algorithms, the calculation of evaluation metrics, and the efficient visualization of cluster results. The results of this evaluation provide an overview of the quality of the customer segmentation produced, which can be used to develop more effective marketing strategies.

2.9. *Knowledge Representation*

At this stage, the results of the clustering analysis are presented in a format that allows stakeholders to easily understand and apply the insights gained. Knowledge representation is the final step in the data mining process aimed at conveying useful information from the analyzed data.

```
plt.figure(figsize=(8, 5))
plt.scatter(rfm_normalized[:, 0], rfm_normalized[:, 1], c=rfm_table['Cluster'], cmap='viridis', alpha=0.5)
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], marker='X', s=200, c='red', label='Centroids'
plt.title('Clustering Hasil RFM')
plt.xlabel('Recency (Normalized)')
plt.ylabel('Frequency (Normalized)')
plt.legend()
plt.colorbar(label='Cluster')
plt.show()
```

Figure 2. Scatter Plot Clustering Results of RFM

3. Results and Discussion

3.1. *Data Collection*

Furniture sales transaction data from ADAPTA.Id in Jepara in 2022, with a total of 2,252 transactions from 928 customers. After data collection, the selection of relevant attributes was carried out from the initial 29 attributes.

```
Nama Kolom dalam Data:
Index(['No. Pesanan', 'Status Pesanan', 'Status Pembatalan/ Pengembalian',
       'No. Resi', 'Opsi Pengiriman', 'Antar ke counter/ pick-up',
       'Pesanan Harus Dikirimkan Sebelum (Menghindari keterlambatan)',
       'Waktu Pengiriman Diatur', 'Waktu Pesanan Dibuat',
       'Waktu Pembayaran Dilakukan', 'Metode Pembayaran', 'SKU Induk',
       'Nama Produk', 'Nomor Referensi SKU', 'Nama Variasi', 'Harga Awal',
       'Harga Setelah Diskon', 'Jumlah', 'Returned quantity',
       'Total Harga Produk', 'Total Diskon', 'Diskon Dari Penjual',
       'Diskon Dari Shopee', 'Berat Produk', 'Jumlah Produk di Pesan',
       'Total Berat', 'Voucher Ditanggung Penjual', 'Cashback Koin',
       'Voucher Ditanggung Shopee', 'Paket Diskon',
       'Paket Diskon (Diskon dari Shopee)',
       'Paket Diskon (Diskon dari Penjual)', 'Potongan Koin Shopee',
       'Diskon Kartu Kredit', 'Ongkos Kirim Dibayar oleh Pembeli',
       'Estimasi Potongan Biaya Pengiriman',
       'Ongkos Kirim Pengembalian Barang', 'Total Pembayaran',
       'Perkiraan Ongkos Kirim', 'Catatan dari Pembeli', 'Catatan',
       'Username (Pembeli)', 'Nama Penerima', 'No. Telepon',
       'Alamat Pengiriman', 'Kota/Kabupaten', 'Provinsi',
       'Waktu Pesanan Selesai'],
      dtype='object')
```

Figure 3. Results of Attribute Data Collection

3.2. *Data selection*

The column selection process is carried out by choosing the relevant columns for RFM analysis, namely 'Username (Buyer)', 'Order Creation Time', and 'Total Payment'. This is an important initial step in building an accurate RFM metric.

	Username (Pembeli)	Waktu Pesanan Dibuat	Total Pembayaran
0	mustova15	2022-01-02 16:37	74.160
1	srengenge30498	2022-01-01 09:52	67.980
2	zayd.babywear	2022-01-01 12:56	57.000
3	dianaprilia21kartini	2022-01-04 12:15	103.727
4	renal.a.p	2022-01-01 19:02	28.652

Figure 4. Results of data selection/column selection

3.2. *Pembersihan data*

Data cleaning is the process of ensuring that the data used in analysis is free from errors, duplicates, and irrelevant data. Common steps taken in data cleaning include converting data formats, handling missing values, and ensuring data consistency and accuracy so that the analysis conducted is more valid and reliable.

```
[ ] rfm_data['Waktu Pesanan Dibuat'] = pd.to_datetime(rfm_data['Waktu Pesanan Dibuat'], format='%Y-%m-%d %H:%M', errors='coerce')

<ipython-input-17-7bf5c81c492d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  rfm_data['Waktu Pesanan Dibuat'] = pd.to_datetime(rfm_data['Waktu Pesanan Dibuat'], format='%Y-%m-%d %H:%M', errors='coerce')
```

Figure 4. Convert to Date and Time

```
[14] rfm_data['Total Pembayaran'] = pd.to_numeric(rfm_data['Total Pembayaran'].fillna(0))
```

Figure 5. Convert to Numeric

```
[11] rfm_data = rfm_data.dropna(subset=['Waktu Pesanan Dibuat'])
```

Figure 6. Handling Missing Values

3.3. *Transformasi Data*

This table shows the results of the Recency analysis, which measures the number of days since a buyer last made a transaction. The username "mustova15" has a Recency score of 363 days, indicating that the last purchase was made 363 days ago. The lowest Recency value in the table is 0 days (for example, "ajirogomigi"), indicating that the buyer has just made a transaction. Recency results are in Figure 7.

	Username (Pembeli)	Recency
0	mustova15	363.0
1	srengenge30498	364.0
2	zayd.babywear	364.0
3	dianaprilia21kartini	361.0
4	renal.a.p	364.0
...	...	...
2188	zakizulfikar287	3.0
2189	naylafairuzia	1.0
2191	muhammadandry466	1.0
2192	ajirogomigi	0.0
2193	3idrisefen	1.0

929 rows × 2 columns

Figure 7. Results of Recency

This table presents the results of the Frequency analysis, which indicates how many times a buyer has made transactions within a specific time period. All buyers listed in this table excerpt have a Frequency value of 2, meaning each of them has made two transactions. Frequency results are in Figure 8.

	Username (Pembeli)	Frequency
0	0809dani	2
1	0907jhoeben	2
2	098777_654321	2
3	0uqlpkarir	2
4	123evinursanti	2
...	...	...
923	zidamawan	2
924	zizaahhh	2
925	zuba.shoes	2
926	zuhara8	2
927	zumrotibintirahmatullah	2

928 rows × 2 columns

Figure 8. Results of frequency

Tabel ini menyajikan hasil analisis Moneter, yang mengukur total nilai transaksi yang dilakukan oleh setiap pembeli selama periode waktu tertentu. Misalnya, "0809dani" memiliki total jumlah pembelian 249.250, sedangkan "zizaahhh" memiliki nilai pembelian yang lebih tinggi yaitu 501.400. Hasil moneter ada pada Gambar 9.

	Username (Pembeli)	Monetary
0	0809dani	249.250
1	0907jhoeben	135.960
2	098777_654321	136.000
3	0uqlpkarir	135.960
4	123evinursanti	158.080
...	...	...
923	zidamawan	53.168
924	zizaahhh	501.400
925	zuba.shoes	313.000
926	zuhara8	268.000
927	zumrotibintirahmatullah	152.000

928 rows × 2 columns

Gambar 9. Hasil Moneter

The result in the image below is RFM data that has been normalized using Min-Max Scaling, where the values of Recency, Frequency, and Monetary have been transformed into a range of 0 to 1. Each value indicates how recent, frequent, and large the customer's transactions are. This normalization ensures that each feature has an equal weight for further analysis, such as in the K-Means clustering algorithm. The results of the RFM data normalization process are shown in Figure 10.

```
array([[0.99725275, 0.        , 0.01199357],
       [1.        , 0.        , 0.01056634],
       [1.        , 0.        , 0.00803058],
       ...,
       [0.00274725, 0.        , 0.05883807],
       [0.        , 0.        , 0.01126379],
       [0.00274725, 0.        , 0.01407206]])
```

Figure 10. Results of the RFM data normalization process

3.4. *Data mining*

3.4.1. *Determination of the Number of Clusters (k)*

The results show the average silhouette score for various numbers of clusters (n_clusters) in the clustering analysis. The silhouette score measures how well the data is clustered, with a value range between -1 and 1, where higher values indicate better clustering. Based on these results, the best number of clusters is 3, with the highest average silhouette score of 0.575, indicating the most optimal cluster separation compared to other cluster numbers. On the other hand, a higher number of clusters, such as 8 or 10, have lower silhouette scores, indicating poorer cluster separation. the optimal number of clusters in Figures 11 and 12.

```
For n_clusters = 2, the average silhouette_score is 0.5717018205380217
For n_clusters = 3, the average silhouette_score is 0.5751805004403845
For n_clusters = 4, the average silhouette_score is 0.5566669778428956
For n_clusters = 5, the average silhouette_score is 0.5093599860344113
For n_clusters = 6, the average silhouette_score is 0.5141482561786495
For n_clusters = 7, the average silhouette_score is 0.5062337828324304
For n_clusters = 8, the average silhouette_score is 0.4884807702452004
For n_clusters = 9, the average silhouette_score is 0.5074333499210048
For n_clusters = 10, the average silhouette_score is 0.4915298192574071
```

Figure 11. Results of the Process for Determining the Optimal Number of Clusters

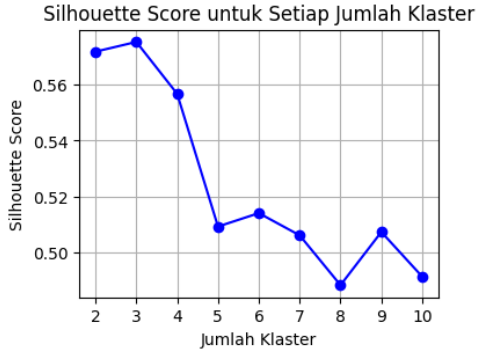
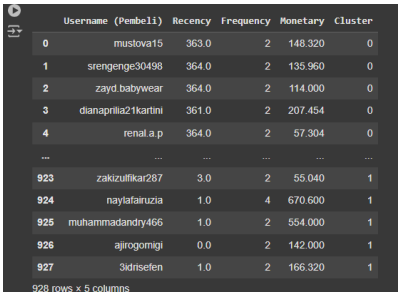


Figure 12. Results Silhouette Scores for Different Values of k

3.4.2. *Clustering Process Results*

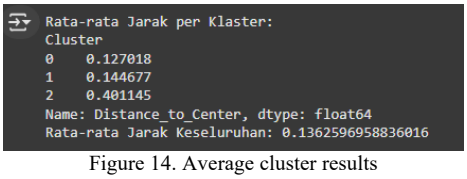
Based on the clustering results, three clusters have formed with different numbers of customers. Cluster 0 consists of 544 customers, with an average distance to the cluster center of 0.127, indicating that customers in this cluster are closer to the center. Cluster 1 has 377 customers, with an average distance to the cluster center of 0.1447, slightly farther than cluster 0. Meanwhile, Cluster 2 consists of only 7 customers, with an average distance to the cluster center that is much greater, namely 0.4011, indicating that customers in this cluster are more separated from the center. Overall, the average distance of all customers to the cluster center is 0.1363. the results are shown in figures 13 and 14



	Username (Pembeli)	Recency	Frequency	Monetary	Cluster
0	mustova15	363.0	2	148.320	0
1	srengenge30498	364.0	2	135.960	0
2	zayd.babywear	364.0	2	114.000	0
3	dianaprilia21kartini	361.0	2	207.454	0
4	renal.a.p	364.0	2	57.304	0
...	...	...	...	...	...
923	zakizulfikar287	3.0	2	55.040	1
924	naylafairuzia	1.0	4	670.600	1
925	muhammadandry466	1.0	2	554.000	1
926	ajirogomigi	0.0	2	142.000	1
927	3idrisefen	1.0	2	166.320	1

928 rows × 5 columns

Figure 13. Cluster results

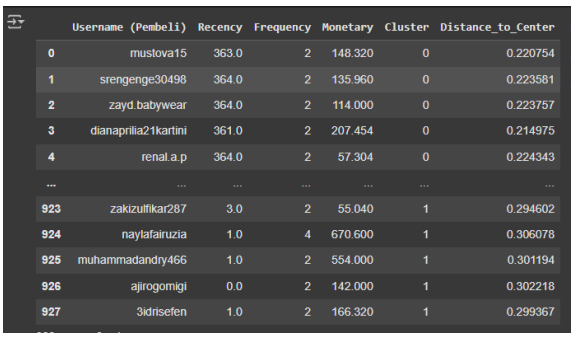


```
Rata-rata Jarak per Klaster:
Cluster
0    0.127018
1    0.144677
2    0.401145
Name: Distance_to_Center, dtype: float64
Rata-rata Jarak Keseluruhan: 0.1362596958836016
```

Figure 14. Average cluster results

3.4.3. *Results of Euclidean Distance Calculation in K-Means Clustering*

The results show that each customer has a Distance_to_Center value that indicates how far they are from the formed cluster center. For example, the customer with ID mustova15 has a distance of 0.220754 from cluster center 0, while the customer zakizulfikar287 has a distance of 0.294602 from cluster center 1. The average distance between customers and the cluster center shows higher variation among customers in smaller clusters, while customers in larger clusters tend to have smaller distances. This analysis provides an overview of the extent to which customers are distributed within each cluster and how well the separation between clusters is achieved.

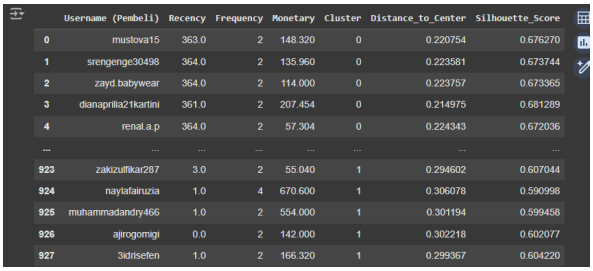


	Username (Pembeli)	Recency	Frequency	Monetary	Cluster	Distance_to_Center
0	mustova15	363.0	2	148.320	0	0.220754
1	srengenge30498	364.0	2	135.960	0	0.223581
2	zayd.babywear	364.0	2	114.000	0	0.223757
3	dianaprilia21kartini	361.0	2	207.454	0	0.214975
4	renal.a.p	364.0	2	57.304	0	0.224343
...	...	...	...	...	...	...
923	zakizulfikar287	3.0	2	55.040	1	0.294602
924	naylafairuzia	1.0	4	670.600	1	0.306078
925	muhammadandry466	1.0	2	554.000	1	0.301194
926	ajirogomigi	0.0	2	142.000	1	0.302218
927	3idrisefen	1.0	2	166.320	1	0.299367

Figure 15. Results of the Euclidean Distance Calculation

3.4.3. *Results of Cluster Quality Evaluation Using Silhouette Score*

In the clustering results, the Silhouette Score is used to evaluate how well the separation between the formed clusters is. This value measures how well each customer is placed in the correct cluster, with a range of values between -1 and 1, where higher values indicate better clustering. For example, the customer mustova15 who is in Cluster 0 has a Silhouette Score of 0.676270, indicating that this customer is well-clustered. On the other hand, the customer zakizulfikar287 who is in Cluster 1 has a lower Silhouette Score of 0.607044, indicating a slightly weaker separation compared to other customers. Overall, the higher Silhouette Score in Cluster 0 indicates that the separation between customers in that cluster is clearer and better compared to other clusters. The result is shown in Figure 16.



	Username (Pembeli)	Recency	Frequency	Monetary	Cluster	Distance_to_Center	Silhouette_Score
0	mustova15	363.0	2	148.320	0	0.220754	0.676270
1	srengenge30498	364.0	2	135.960	0	0.223581	0.673744
2	zayd.babywear	364.0	2	114.000	0	0.223757	0.673365
3	dianaprilia21kartini	361.0	2	207.454	0	0.214975	0.681289
4	renal.a.p	364.0	2	57.304	0	0.224343	0.672036
...	...	...	...	...	...	...	...
923	zakizulfikar287	3.0	2	55.040	1	0.294602	0.607044
924	naylafairuzia	1.0	4	670.600	1	0.306078	0.590998
925	muhammadandry466	1.0	2	554.000	1	0.301194	0.599458
926	ajirogomigi	0.0	2	142.000	1	0.302218	0.602077
927	3idrisefen	1.0	2	166.320	1	0.299367	0.604220

Figure 16. Silhouette Score Calculation Results

3.5. *Visualization and Graphics*

3.5.1. *Visualization of Recency Distribution with Histogram*

This histogram illustrates the distribution of customer Recency, which measures the time since the last transaction by the customer. The X-axis shows the range of recency values, while the Y-axis shows the number of customers within each range. Based on the histogram, it can be seen that the recency values of the customers are fairly evenly distributed, with some customers having recently transacted and others who have not interacted for a long time. This helps in understanding the distribution of customer interaction time with the company as shown in Figure 17.

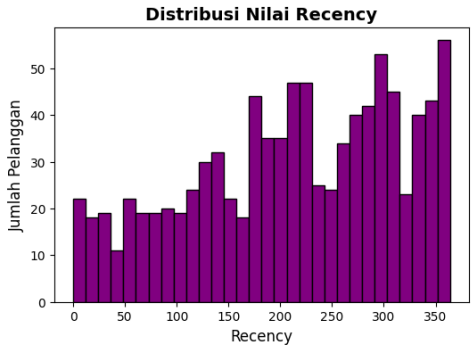
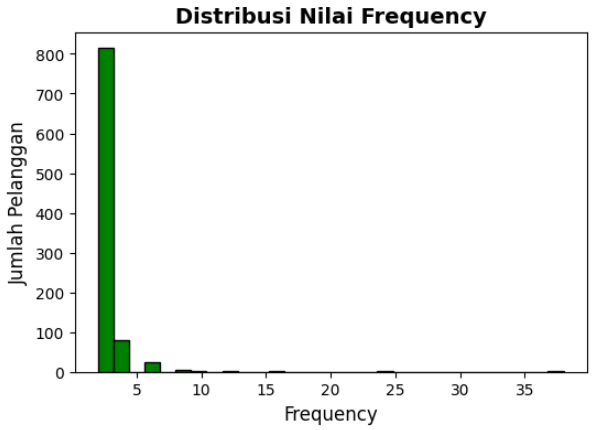


Figure 17. Recency histogram

3.5.2. *Histogram Frequency Visualization with Histogram*

This histogram shows the distribution of Frequency values, which represent the number of customer

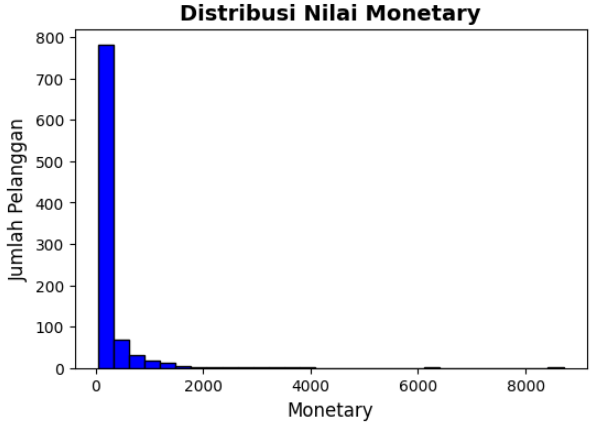
transactions over a specific period. From the graph, it can be seen that most customers make few transactions (between 1 to 5 times). Higher frequencies (above 5) are rare, with the number of customers decreasing drastically as the Frequency value increases. This indicates that the majority of customers only make a small number of transactions.



Gambar 18. Histogram frequency

3.5.2. *Monetary Visualization with Histogram*

This histogram illustrates the distribution of Monetary value, which measures the total purchase value of customers over a specific period. The graph shows that most customers have a low purchase value (below 500). Only a few customers have a higher Monetary value (above 2000), so this distribution also shows inequality, where customers with a significant contribution to the company's revenue are quite rare as shown in chart 17.



Gambar 19. Histogram Monetary

3.5.2. *Visualization of the number of customers per cluster*

This graph shows the distribution of the number of customers in each cluster resulting from K-Means. Cluster 0 has the largest number of customers (544 customers), followed by Cluster 1 (377 customers). Cluster 2 has a very small number of customers (7 customers). This shows that the majority of customers are concentrated in the two main clusters, while the small cluster represents a group of customers with unique characteristics that are less frequently encountered, as shown in the image.

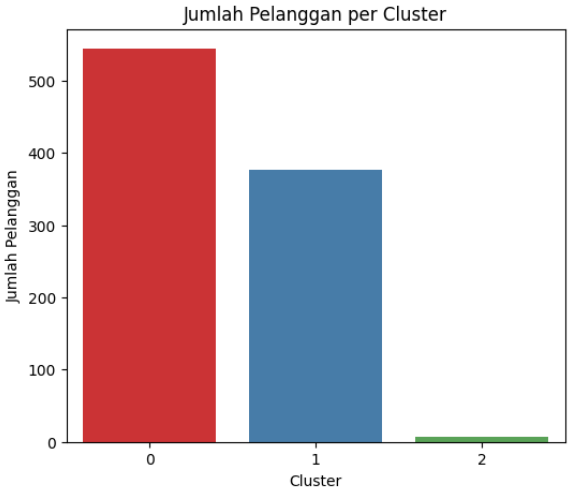


Figure 20. Histogram of customer frequency per cluster

3.5.2. *Correlation Between Variables (Recency, Frequency, Monetary)*

This heatmap shows the correlation relationship between the Recency, Frequency, and Monetary variables. It appears that Frequency and Monetary have a fairly strong positive correlation (0.63), indicating that customers who transact more frequently tend to have a higher total transaction value (Monetary). On the other hand, Recency has a very low correlation with Frequency (-0.06) and Monetary (0.03), indicating that the time since the customer's last transaction is not directly related to the amount or value of their transactions.

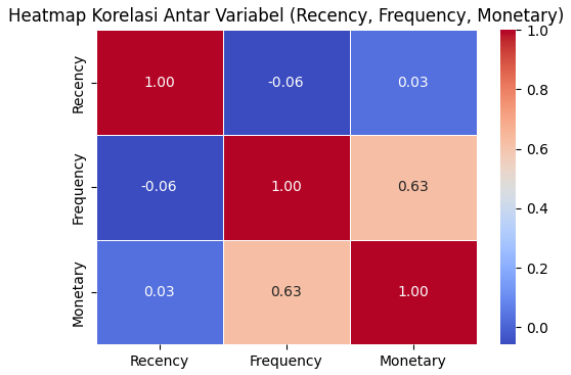


Figure 21. Histogram of customer frequency per cluster

3.5.2. Visualization of the number of customers per cluster

This 3D plot provides an overview of customer distribution across three main dimensions: Recency, Frequency, and Monetary, which form the basis for cluster formation. Each color of the dot represents customers in a specific cluster. The small cluster (yellow dots) appears to have much higher Frequency and Monetary values compared to other clusters, while the majority of customers are in a cluster with higher Recency values and lower Frequency/Monetary values.

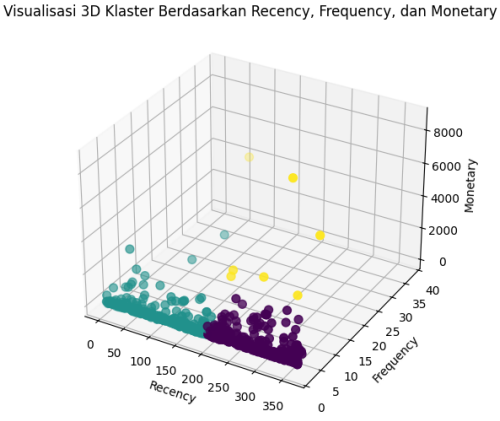


Figure 22. Histogram of customer frequency per cluster

4. Conclusion

From the integration of the RFM (Recency, Frequency, Monetary) method and K-Means Clustering, it can be concluded that the combination of these two methods is very effective for customer segmentation. RFM measures customer characteristics based on the time of the last transaction, frequency, and total transaction value, while K-Means divides customers into three clusters based on similar patterns. With an optimal Silhouette Score (0.575) for 3 clusters, it was found that Cluster 0 dominates with 544 customers who have high Recency, low Frequency, and medium Monetary; Cluster 1 consists of 377 active customers with low Recency, slightly higher Frequency, and medium Monetary; and Cluster 2 includes only 7 high-value customers with much higher Frequency and Monetary. This segmentation allows for the design of targeted marketing strategies, such as customer retention for high-value clusters and reactivation for clusters with high recency, thereby providing strategic insights for business decision-making.

References

[1] R. B. Prasetyo, "Pengaruh E-Commerce dalam Dunia Bisnis," *JMEB J. Manaj. Ekon. Bisnis*, vol. 1, no. 01, pp. 1–11, 2023, doi: 10.59561/jmeb.v1i01.92.

[2] M. Sulaiman, R. Yudistira, R. Narasati, and R. Herdiana, "Penerapan Data Mining dengan Metode Clustering untuk menentukan Strategi Peningkatan Penjualan Berdasarkan Data Transaksi," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, 2024.

[3] P. S. Ananda, E. Sediono, and I. Sembiring, "KMeans Clustering Menggunakan RapidMiner dalam Segmentasi Pelanggan dengan Evaluasi Davies Bouldin Index Untuk Menentukan Jumlah Cluster Paling Optimal," *J. BATIRSI*, vol. 6, no. 2, p. 10, 2023.

[4] P. E. Prakasawati, Y. H. Chrisnanto, and A. I. Hadiana, "Segmentasi Pelanggan Berdasarkan Produk Menggunakan Metode K- Medoids," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, 2019, doi: 10.30865/komik.v3i1.1610.

[5] F. M. Almufqi and A. Voutama, "Perbandingan Metode Data Mining Untuk Memprediksi Prestasi Akademik Siswa," *J. Tek.*, vol. 15, no. 1, pp. 61–66, 2023, doi: 10.30736/jt.v15i1.929.

[6] P. A. Wicaksana, I. B. A. Swamardika, and R. S. Hartati, "Literature Review Analisis Perilaku Pelanggan Menggunakan RFM Model," *Maj. Ilm. Teknol. Elektro*, vol. 21, no. 1, p. 21, 2022, doi: 10.24843/mite.2022.v21i01.p04.

[7] B. Basri, W. Gata, and R. Risnandar, "Analisis Loyalitas Pelanggan Berbasis Model Recency, Frequency, dan Monetary (RFM) dan Decision Tree pada PT. Solo," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, pp. 943–950, 2020, doi: 10.25126/jtiik.2020752284.

[8] N. Ahsina, F. Fatimah, and F. Rachmawati, "Analisis Segmentasi Pelanggan Bank Berdasarkan Pengambilan Kredit Dengan Menggunakan Metode K-Means Clustering," *J. Ilm. Teknol. Infomasi Terap.*, vol. 8, no. 3, 2022, doi: 10.33197/jitter.vol8.iss3.2022.883.

[9] R. J. Kasim, S. Bahri, and S. Amir, "Implementasi Metode K-Means Untuk Clustering Data Penduduk Miskin Dengan Systematic Random Sampling," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 95–101, 2021.

[10] N. Afiasari, N. Suarna, and N. Rahaningsi, "Implementasi Data Mining Transaksi Penjualan

Menggunakan Algoritma Clustering dengan Metode K-Means," *J. SAINTEKOM*, vol. 13, no. 1, pp. 100–110, 2023, doi: 10.33020/saintekom.v13i1.402.

[11] J.-O. Palacio-Niño and F. Berzal, "Evaluation Metrics for Unsupervised Learning Algorithms," 2019, [Online]. Available: http://arxiv.org/abs/1905.05667

[12] I. Virgo, S. Defit, and Y. Yuhandri, "Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering," *J. Sistim Inf. dan Teknol.*, vol. 2, pp. 23–28, 2020, doi: 10.37034/jsisfotek.v2i1.17.