



PERBANDINGAN SEGMENTASI PELANGGAN MENGGUNAKAN ALGORITMA K-MEANS BERDASARKAN ANALISIS RFM DAN PERILAKU PEMBELIAN PADA PT XYZ

Maulida Shifa Annisa¹, Sischa Wahyuning Tyas²

^{1,2}Sains Data, Universitas Pembangunan Nasional “Veteran” Jawa Timur

E-mail: maulidashifa26@gmail.com, sischa_wahyuning.sada@upnjatim.ac.id

ABSTRACT

Customer segmentation is an approach to understanding customer characteristics and supporting the development of more targeted marketing strategies. This study aims to segment customers of PT XYZ based on Recency, Frequency, Monetary (RFM) analysis and purchasing behavior using the K-Means algorithm. The dataset consists of sales transactions recorded from March to June 2026. This study compares an RFM-based model with a model that combines RFM and purchasing behavior variables. The optimal number of clusters is evaluated using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. The evaluation results indicate that the RFM model with two clusters provides the best clustering performance. The segmentation produces two customer groups with distinct characteristics: Cluster 0 consists of customers with relatively low transaction activity and purchasing diversity, while Cluster 1 consists of customers with higher transaction frequency, transaction value, and purchasing diversity. These findings indicate that RFM analysis can be used to identify customer characteristics and support the development of differentiated marketing strategies, such as repeat-purchase promotions for Cluster 0 and loyalty programs for Cluster 1.

Keywords: *Customer Segmentation, K-Means, RFM, Purchasing Behavior, Marketing Strategy.*

ABSTRAK

Segmentasi pelanggan merupakan salah satu pendekatan untuk memahami karakteristik pelanggan dan mendukung penyusunan strategi pemasaran yang lebih terarah. Penelitian ini bertujuan mengelompokkan pelanggan PT XYZ berdasarkan analisis *Recency, Frequency, Monetary* (RFM) dan perilaku pembelian menggunakan algoritma K-Means. Data yang digunakan merupakan data transaksi penjualan periode Maret-Juni 2026. Penelitian membandingkan model berbasis RFM dengan model yang menggabungkan RFM dan variabel perilaku pembelian. Penentuan jumlah klaster dilakukan dengan mempertimbangkan nilai *Silhouette Coefficient, Davies-*

Article History

Received: Oktober 2026

Reviewed: Oktober 2026

Published: Oktober 2026

Plagiarism Checker No
237.u873247.fh8374

Copyright : Author
Publish by : Kohesi



This work is licensed
under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)



Bouldin Index, dan *Calinski-Harabasz Index*. Hasil evaluasi menunjukkan bahwa model RFM dengan dua kluster menghasilkan kualitas pengelompokan terbaik, dengan nilai *Silhouette Coefficient* sebesar 0,4513, *Davies-Bouldin Index* sebesar 0,8814, dan *Calinski-Harabasz Index* sebesar 1363,874. Segmentasi menghasilkan Cluster 0 sebanyak 821 pelanggan dan Cluster 1 sebanyak 482 pelanggan. Cluster 0 memiliki aktivitas transaksi dan keragaman pembelian yang relatif rendah, sedangkan Cluster 1 menunjukkan frekuensi dan nilai transaksi serta keragaman pembelian yang lebih tinggi. Hasil penelitian ini dapat digunakan sebagai dasar penyusunan strategi pemasaran yang berbeda untuk setiap kelompok, seperti promosi pembelian ulang bagi Cluster 0 dan program loyalitas bagi Cluster 1.

Kata Kunci: Segmentasi Pelanggan, K-Means, RFM, Perilaku Pembelian, Strategi Pemasaran.

1. PENDAHULUAN

Data transaksi penjualan merupakan catatan aktivitas penjualan yang memuat informasi mengenai waktu transaksi, produk yang dibeli, jumlah pembelian, dan nilai transaksi. Data tersebut dapat dimanfaatkan untuk mengidentifikasi karakteristik dan pola pembelian pelanggan sebagai dasar pengambilan keputusan pemasaran[1]. Namun, data transaksi yang hanya disajikan dalam bentuk laporan penjualan belum secara langsung menggambarkan perbedaan perilaku antar pelanggan. Oleh karena itu, diperlukan segmentasi pelanggan untuk mengelompokkan pelanggan berdasarkan kemiripan karakteristik pembelian sehingga pola transaksi dapat dianalisis secara lebih terstruktur[2].

Analisis *Recency, Frequency, Monetary* (RFM) digunakan untuk mengukur perilaku pelanggan berdasarkan waktu transaksi terakhir, frekuensi transaksi, dan nilai pembelian. Christy *et al.*[3] mengkaji RFM untuk segmentasi pelanggan, sedangkan Awaliyah *et al.* [2] menerapkan algoritma K-Means pada transaksi ritel daring. Anitha dan Patil [4] juga menggabungkan RFM dan K-Means untuk menganalisis perilaku pembelian pelanggan. Penelitian tersebut menjadi landasan penerapan RFM dan K-Means dalam segmentasi pelanggan.

Meskipun RFM dapat menggambarkan aktivitas pembelian pelanggan, pendekatan tersebut belum secara langsung merepresentasikan keragaman produk, ukuran rata-rata keranjang belanja, dan keragaman merek yang dibeli. Oleh karena itu, penelitian ini membandingkan segmentasi berbasis RFM dengan segmentasi yang menambahkan variabel *Product Diversity*, *Average Basket Size*, dan *Brand Diversity*. Perbandingan ini dilakukan untuk mengevaluasi apakah penambahan karakteristik perilaku pembelian dapat meningkatkan kualitas pengelompokan pelanggan.

Penelitian ini menggunakan data transaksi penjualan produk perawatan rambut (*haircare*) PT XYZ selama periode Maret hingga Juni 2026. Algoritma *K-Means Clustering* diterapkan pada kedua konfigurasi variabel, kemudian kualitas kluster dievaluasi menggunakan *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Penelitian ini bertujuan membandingkan



kualitas kedua pendekatan dan mengidentifikasi karakteristik setiap klaster yang terbentuk sebagai dasar analitis dalam mendukung penyusunan strategi pemasaran.

1. METODE PENELITIAN

2.1 Sumber Data

Penelitian ini menggunakan data sekunder berupa laporan transaksi penjualan produk perawatan rambut (*haircare*) PT XYZ selama periode 1 Maret-30 Juni 2026. Data diperoleh melalui dokumentasi perusahaan dan terdiri atas 34.562 baris transaksi sebelum pengolahan data, termasuk satu baris ringkasan yang tidak merepresentasikan transaksi individual.

2.2 Variabel Penelitian

Variabel penelitian terdiri atas enam fitur numerik yang dibagi menjadi dua kelompok, yaitu RFM dan perilaku pembelian. Ketiga fitur RFM digunakan sebagai konfigurasi dasar, sedangkan tiga fitur perilaku pembelian ditambahkan pada konfigurasi kedua untuk mengetahui perbedaan kualitas segmentasi.

Tabel 2.1 Variabel Penelitian

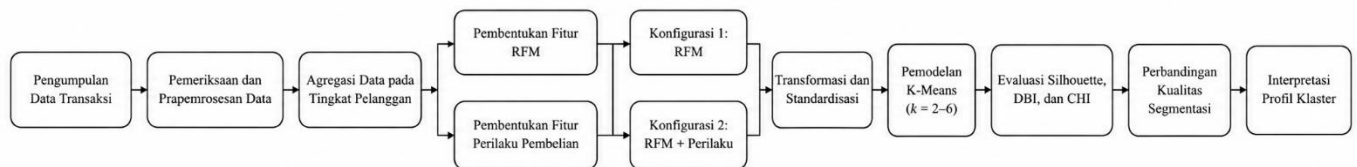
Kelompok	Variabel	Definisi operasional	Skala
RFM	<i>Recency</i>	Jumlah hari sejak transaksi terakhir pelanggan hingga tanggal acuan analisis.	Rasio
RFM	<i>Frequency</i>	Jumlah faktur transaksi unik yang dilakukan pelanggan.	Rasio
RFM	<i>Monetary</i>	Total nilai transaksi pelanggan selama periode pengamatan.	Rasio
Perilaku pembelian	<i>Product Diversity</i>	Jumlah kode produk unik (Item Code) yang dibeli pelanggan.	Rasio
Perilaku pembelian	<i>Average Basket Size</i>	Rata-rata total kuantitas produk (Qty) per faktur transaksi pelanggan.	Rasio
Perilaku pembelian	<i>Brand Diversity</i>	Jumlah merek atau submerek unik (Sub Brand) yang dibeli pelanggan.	Rasio

Variabel *Product Diversity*, *Average Basket Size*, dan *Brand Diversity* digunakan untuk menggambarkan keragaman produk, ukuran rata-rata keranjang belanja, dan keragaman merek berdasarkan catatan transaksi. Variabel perilaku tersebut dihitung dari transaksi penjualan

nonretur. Identitas pelanggan digunakan untuk proses agregasi, tetapi tidak menjadi fitur masukan model.

2.3 Langkah-langkah Penelitian

Penelitian dilakukan melalui beberapa tahapan. Alur penelitian secara garis besar disajikan pada Gambar 2.1 berikut.



Gambar 2.1 Diagram Alur Penelitian

Tahapan penelitian dimulai dengan pemeriksaan dan prapemrosesan data transaksi, dilanjutkan dengan agregasi data pada tingkat pelanggan dan pembentukan fitur RFM serta perilaku pembelian. Selanjutnya, fitur ditransformasikan dan distandardisasi sebelum pemodelan K-Means pada dua konfigurasi, yaitu RFM dan RFM dengan tambahan fitur perilaku pembelian, menggunakan jumlah kluster $k=2$ hingga $k=6$, $random_state=42$, dan $n_init=20$. Hasil pengelompokan kemudian dievaluasi menggunakan *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI) untuk membandingkan kualitas segmentasi dan menginterpretasikan karakteristik setiap kluster.

2.4 Transformasi dan Standardisasi Data

Transformasi dan standardisasi dilakukan untuk mengurangi kemencengan distribusi data serta menyetarakan skala antarvariabel sebelum pemodelan *K-Means*[5]. Transformasi logaritmik dapat dituliskan sebagai:

$$x' = \ln(1 + x) \quad (2.1)$$

Standardisasi dilakukan menggunakan *StandardScaler* dengan rumus:

$$z = \frac{x' - \mu}{\sigma} \quad (2.2)$$

Keterangan: x merupakan nilai awal fitur, x' nilai setelah transformasi, (z) nilai setelah standardisasi, (μ) rata-rata, dan (σ) simpangan baku. Transformasi diterapkan sesuai karakteristik data dan implementasi pada kode penelitian.

2.5 Algoritma K-Means Clustering

K-Means digunakan untuk mengelompokkan pelanggan berdasarkan kemiripan karakteristik pembelian [6]. Algoritma ini membentuk kluster dengan meminimalkan jumlah kuadrat jarak setiap data terhadap pusat klasternya [7]. Fungsi objektif K-Means dirumuskan sebagai berikut:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (2.3)$$



Keterangan: (J) merupakan total kuadrat jarak dalam klaster, (k) jumlah klaster, (C_j) anggota klaster ke- (j) , (x_i) data pelanggan, dan (mu_j) pusat klaster ke- (j) . Pemodelan dilakukan pada konfigurasi RFM dan RFM dengan fitur perilaku pembelian.

2.6 Evaluasi Kualitas Klaster

Evaluasi dilakukan menggunakan tiga metrik untuk membandingkan kualitas segmentasi. *Silhouette Score* mengukur kekompakan dan pemisahan klaster [8], dengan rumus:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.4)$$

Keterangan: $(a(i))$ merupakan rata-rata jarak data ke klasternya sendiri, sedangkan $(b(i))$ merupakan rata-rata jarak terkecil terhadap klaster lain. Nilai yang lebih tinggi menunjukkan pemisahan klaster yang lebih baik.

Davies-Bouldin Index (DBI) mengukur kemiripan antarklaster berdasarkan penyebaran data dan jarak antarpusat klaster[9]:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (2.5)$$

Keterangan: (S_i) dan (S_j) merupakan ukuran penyebaran klaster, sedangkan (M_{ij}) merupakan jarak antara pusat kedua klaster. Nilai DBI yang lebih rendah menunjukkan hasil pengelompokan yang lebih baik.

Calinski-Harabasz Index (CHI) membandingkan dispersi antarklaster dengan dispersi dalam klaster[10], menggunakan rumus:

$$CHI = \frac{B_k / (k-1)}{W_k / (n-k)} \quad (2.6)$$

Keterangan: (B_k) merupakan dispersi antarklaster, (W_k) dispersi dalam klaster, (k) jumlah klaster, dan (n) jumlah pelanggan. Nilai CHI yang lebih tinggi umumnya menunjukkan struktur klaster yang lebih baik. Ketiga metrik tersebut digunakan secara bersama untuk membandingkan konfigurasi RFM dan RFM dengan fitur perilaku pembelian.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pengumpulan data menghasilkan dataset transaksi penjualan yang digunakan sebagai dasar analisis segmentasi pelanggan. Ringkasan karakteristik dataset disajikan pada Tabel 3.1 berikut.

Tabel 3.1 Ringkasan Dataset Penelitian

Indikator	Jumlah
Baris transaksi detail	34.562
Jumlah kolom	13
Pelanggan unik sebelum pemisahan kanal	1.307
Invoice unik	6.254

Kode produk unik	238
Sub-brand unik	31

Berdasarkan Tabel 3.1, dataset penelitian terdiri atas 34.562 baris transaksi detail dan 13 kolom, dengan 1.307 pelanggan unik sebelum pemisahan kanal. Dataset juga mencakup 6.254 invoice, 238 kode produk, dan 31 sub-brand. Keragaman data tersebut dapat digunakan untuk membentuk fitur RFM dan karakteristik perilaku pembelian pelanggan sebagai dasar analisis segmentasi.

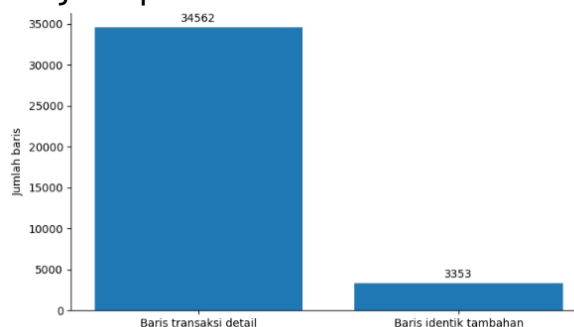
3.2 Prapemrosesan Data

Pemeriksaan kualitas dilakukan untuk mengidentifikasi nilai kosong, baris duplikat, ketidaksesuaian nilai transaksi, dan penanda retur. Ringkasan hasil pemeriksaan kualitas data disajikan pada Tabel 3.2 berikut.

Tabel 3.2 Hasil Pemeriksaan Kualitas Data

Indikator	Hasil
Kolom yang diperiksa	13
Nilai kosong pada seluruh kolom	0
Ketidakesuaian perhitungan Total	0 baris
Baris duplikat tambahan	3.353
Baris transaksi bertanda retur	170
Invoice bertanda retur	121

Berdasarkan Tabel 3.2, pemeriksaan terhadap 13 kolom menunjukkan tidak adanya nilai kosong maupun ketidaksesuaian perhitungan total transaksi. Namun, ditemukan 3.353 baris duplikat identik, termasuk 1.573 baris pada data pelanggan individual, serta 170 baris transaksi retur yang tercatat dalam 121 invoice. Hasil pemeriksaan duplikasi pada keseluruhan dataset dan data pelanggan individual disajikan pada Gambar 3.1 berikut.



Gambar 3.1 Pemeriksaan Baris Identik pada Dataset

Gambar 3.1 menunjukkan bahwa dari 34.562 baris transaksi detail, terdapat 3.353 baris identik tambahan. Baris tersebut tetap dipertahankan dalam analisis utama karena belum dapat dipastikan sebagai kesalahan pencatatan. Pengaruh duplikasi terhadap hasil segmentasi selanjutnya perlu diuji melalui analisis sensitivitas.

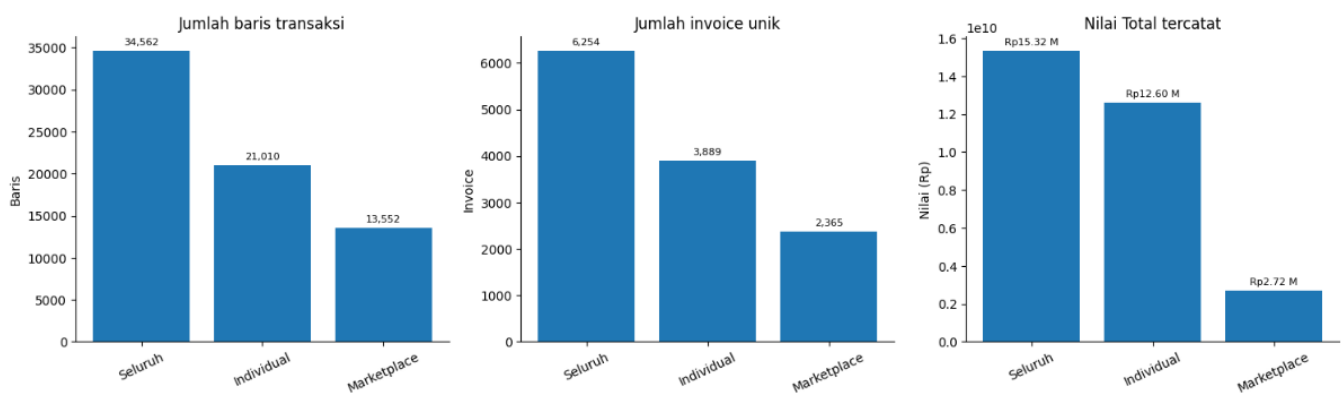
Selanjutnya, data dipisahkan menjadi transaksi pelanggan individual dan marketplace yang mencakup Shopee, TikTok, dan Tokopedia. Pemisahan ini dilakukan karena label marketplace

menunjukkan saluran penjualan, bukan identitas pembeli individual. Perbandingan cakupan data disajikan pada Tabel 3.3 berikut.

Tabel 3.3 Perbandingan Cakupan Data Berdasarkan Kanal

Cakupan	Baris transaksi	Customer unik	Invoice unik	Total tercatat
Seluruh transaksi	34.562	1.307	6.254	Rp15.317.593.473
Pelanggan individual	21.010	1.304	3.889	Rp12.602.256.623
Marketplace	13.552	3	2.365	Rp2.715.336.850

Perbandingan jumlah transaksi, invoice unik, dan nilai transaksi berdasarkan kanal ditampilkan pada Gambar 3.2 berikut.



Gambar 3.2 Perbandingan Cakupan Transaksi Berdasarkan Kanal

Data pelanggan individual mencakup 21.010 baris transaksi, sedangkan marketplace mencakup 13.552 baris. Penelitian ini menggunakan data pelanggan individual sebagai dasar segmentasi, sementara data marketplace hanya digunakan untuk menggambarkan cakupan penjualan.

Pemeriksaan retur menemukan 170 baris bertanda RET dari 121 invoice pada keseluruhan dataset. Pada data pelanggan individual, terdapat delapan baris retur dari satu invoice dan satu pelanggan, dengan nilai Total sebesar Rp0. Transaksi retur tidak digunakan untuk menghitung Recency, Frequency, Product Diversity, Average Basket Size, dan Brand Diversity. Sementara itu, Monetary dihitung berdasarkan nilai pada kolom Total. Karena nilai kedelapan baris retur tersebut Rp 0, transaksi tersebut tidak memengaruhi nilai Monetary pelanggan terkait. Dari 1.304 pelanggan individual, satu pelanggan tidak memiliki tanggal transaksi penjualan valid untuk menghitung Recency. Dengan demikian, sebanyak 1.303 pelanggan digunakan dalam pemodelan. Ringkasan data pemodelan disajikan pada Tabel 3.4 berikut.

Tabel 3.4 Ringkasan Data untuk Pemodelan

Tahap	Jumlah
Transaksi pelanggan individual	21.010
Identitas pelanggan sebelum pembentukan fitur	1.304
Pelanggan yang digunakan dalam pemodelan	1.303
Fitur pada model RFM	3



Tabel 3.4 menunjukkan bahwa unit analisis penelitian adalah pelanggan, bukan baris transaksi. Setiap pelanggan direpresentasikan melalui fitur yang dihitung dari riwayat pembeliannya. Fitur tersebut digunakan dalam dua konfigurasi, yaitu model RFM dengan tiga fitur serta model RFM dengan tambahan tiga fitur perilaku pembelian. Kedua model kemudian dibandingkan untuk menentukan konfigurasi yang menghasilkan segmentasi pelanggan lebih baik.

3.3 Pembentukan Fitur Pelanggan

Pembentukan fitur dilakukan melalui agregasi data transaksi berdasarkan identitas pelanggan, sehingga setiap pelanggan direpresentasikan dalam satu baris data. Agregasi menghasilkan enam fitur dari 1.303 pelanggan, yaitu *Recency*, *Frequency*, *Monetary*, *Product Diversity*, *Average Basket Size*, dan *Brand Diversity*.

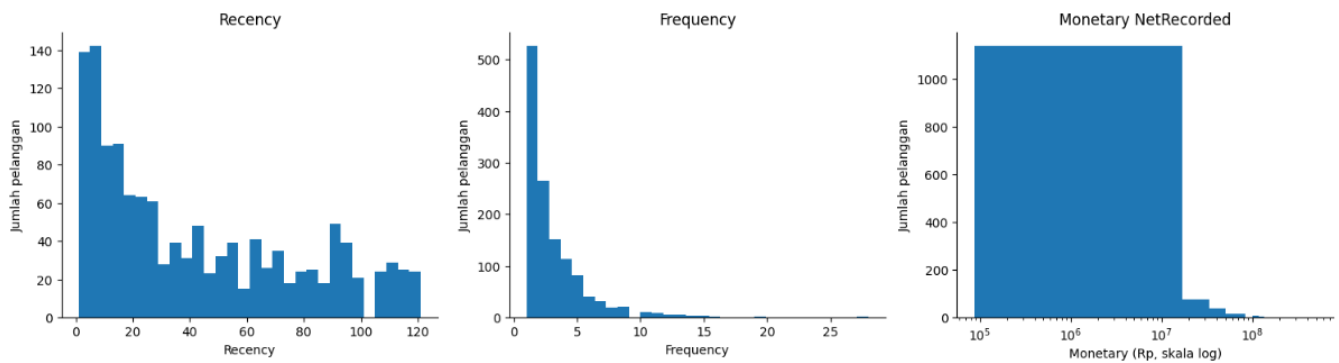
Recency menunjukkan jarak hari antara tanggal acuan 1 Juli 2026 dan transaksi terakhir pelanggan, *Frequency* merupakan jumlah invoice unik, sedangkan *Monetary* adalah total nilai transaksi. Adapun *Product Diversity* menghitung jumlah kode produk unik, *Average Basket Size* menunjukkan rata-rata kuantitas produk per invoice, dan *Brand Diversity* menghitung jumlah sub-brand unik yang dibeli.

Statistik deskriptif keenam fitur disajikan pada Tabel 3.5 untuk menggambarkan karakteristik pembelian pelanggan sebelum proses segmentasi.

Tabel 3.5 Statistik Deskriptif Fitur Pelanggan

Fitur	Minimum	Median	Maksimum
Recency (hari)	1	29	121
Frequency (invoice)	1	2	28
Monetary (Rp)	87.000	2.310.000	502.613.013
Product Diversity (item)	1	5	80
Average Basket Size (unit)	1	8	244,2
Brand Diversity (sub-brand)	1	3	22

Berdasarkan Tabel 4.5, terdapat variasi yang cukup besar pada karakteristik pembelian pelanggan. Nilai *Recency* mencapai 121 hari, sedangkan *Frequency* berkisar antara 1 hingga 28 invoice. Nilai *Monetary* juga memiliki rentang yang lebar, dengan nilai maksimum sebesar Rp502.613.013. Variasi serupa terlihat pada jumlah produk, ukuran keranjang pembelian, dan jumlah sub-brand yang dibeli. Distribusi tiga fitur RFM divisualisasikan pada Gambar 3.3 berikut.



Gambar 3.3 Distribusi Fitur RFM Pelanggan

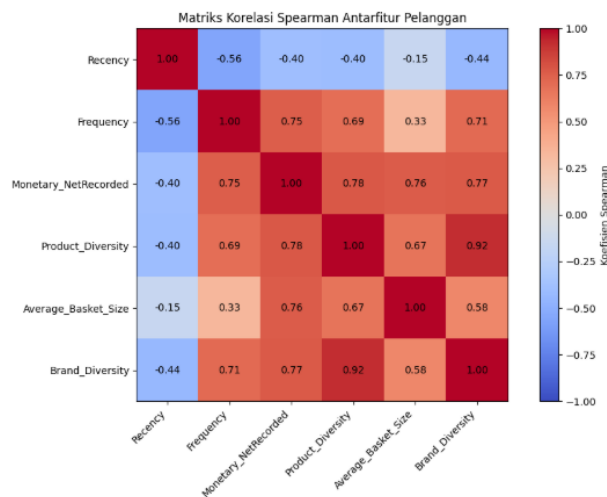
Gambar 3.3 menunjukkan bahwa fitur Recency tersebar pada rentang sekitar 0-120 hari, dengan jumlah pelanggan cenderung lebih banyak pada nilai yang rendah. Fitur Frequency sangat terkonsentrasi pada nilai 1-3 transaksi, sedangkan pelanggan dengan frekuensi lebih tinggi jumlahnya semakin sedikit. Sementara itu, Monetary menunjukkan sebagian besar pelanggan memiliki nilai transaksi pada kisaran 10^5 hingga 10^7 rupiah, dengan sebagian kecil mencapai nilai yang lebih tinggi. Perbedaan skala dan kemencengan distribusi ini menjadi pertimbangan dalam transformasi dan standardisasi fitur sebelum pemodelan K-Means.

3.4 Analisis Korelasi Antarfitur

Analisis korelasi Spearman dilakukan untuk mengetahui hubungan antarfitur yang digunakan dalam segmentasi pelanggan. Hasil analisis disajikan pada Tabel 3.6 dan Gambar 3.4 berikut.

Tabel 3.6. Matriks Korelasi Spearman Antarfitur Pelanggan

Fitur	Recency	Frequency	Monetary	Product Diversity	Average Basket Size	Brand Diversity
Recency	1,000	-0,556	-0,397	-0,404	-0,152	-0,444
Frequency	-0,556	1,000	0,753	0,692	0,334	0,707
Monetary	-0,397	0,753	1,000	0,781	0,760	0,766
Product Diversity	-0,404	0,692	0,781	1,000	0,669	0,916
Average Basket Size	-0,152	0,334	0,760	0,669	1,000	0,584
Brand Diversity	-0,444	0,707	0,766	0,916	0,584	1,000



Gambar 3.4. Heatmap Korelasi Spearman Antarfitur Pelanggan

Berdasarkan hasil analisis, korelasi tertinggi terdapat pada *Product_Diversity* dan *Brand_Diversity* sebesar 0,916, yang menunjukkan hubungan positif sangat kuat. *Monetary_NetRecorded* juga memiliki korelasi positif kuat dengan *Product_Diversity* (0,781), *Brand_Diversity* (0,766), dan *Average_Basket_Size* (0,760). Sementara itu, *Recency* berkorelasi negatif dengan seluruh fitur lainnya, dengan korelasi terkuat terhadap *Frequency* sebesar -0,556.

Dengan menggunakan ambang korelasi absolut $\geq 0,70$, ditemukan enam pasangan fitur yang memiliki hubungan relatif kuat. Temuan ini menunjukkan adanya keterkaitan antarfitur, khususnya antara keragaman produk dan merek. Meskipun demikian, fitur tidak dihapus secara otomatis karena setiap variabel merepresentasikan aspek perilaku pembelian yang berbeda. Hasil korelasi selanjutnya digunakan sebagai informasi pendukung dalam mengevaluasi kualitas clustering berdasarkan Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index.

3.4 Transformasi dan Standardisasi Data

Transformasi dilakukan menggunakan fungsi $\log(1 + x)$ pada lima fitur nonnegatif dan *signed-log* pada fitur *Monetary*. Hasilnya menunjukkan penurunan nilai absolut *skewness* pada seluruh fitur. Hasil tersebut menunjukkan bahwa transformasi mengurangi kemencengan distribusi data. Hasil transformasi disajikan pada Tabel 3.7 berikut.

Tabel 3.7 Perbandingan Skewness Sebelum dan Sesudah Transformasi

Fitur	Sebelum	Sesudah
Recency (hari)	0,665	-0,575
Frequency (invoice)	3,207	0,988
Monetary (Rp)	8,953	0,423
Product Diversity (item)	2,447	0,453
Average Basket Size (unit)	4,146	0,535
Brand Diversity (sub-brand)	1,466	0,390

Selanjutnya, standardisasi dilakukan menggunakan *StandardScaler* agar seluruh fitur memiliki skala yang sebanding. Hasil standardisasi menunjukkan bahwa rata-rata setiap fitur mendekati 0 dan standar deviasi sebesar 1 disajikan pada Tabel 3.8 berikut.

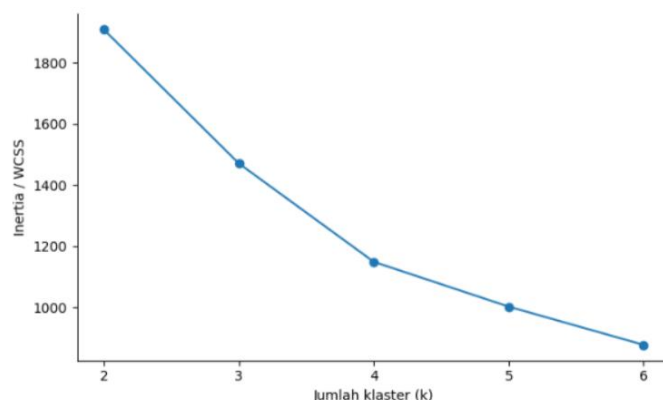
Tabel 3.8 Statistik Data Sesudah Standardisasi

Fitur	Mean	Standar Deviasi
Recency (hari)	0,000	1,000
Frequency (invoice)	0,000	1,000
Monetary (Rp)	0,000	1,000
Product Diversity (item)	0,000	1,000
Average Basket Size (unit)	0,000	1,000
Brand Diversity (sub-brand)	0,000	1,000

Berdasarkan Tabel 3.8, nilai mean sebesar 0,000 menunjukkan bahwa rata-rata data setelah standardisasi berada pada titik nol, sedangkan standar deviasi sebesar 1,000 menunjukkan bahwa penyebaran data telah disesuaikan terhadap standar deviasi awal. Nilai tersebut merupakan hasil matematis standardisasi, bukan nilai setiap pelanggan. Proses ini menyamakan skala fitur agar fitur dengan nilai numerik besar, seperti Monetary, tidak mendominasi perhitungan jarak dalam algoritma K-Means.

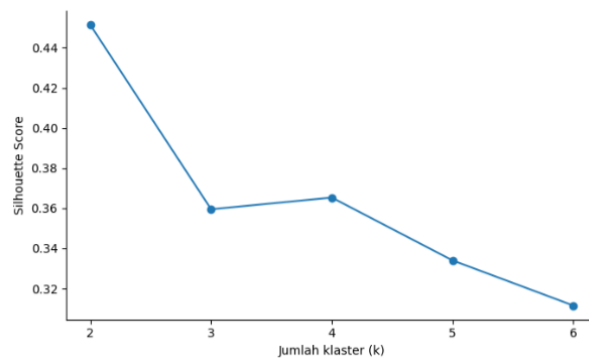
3.5 Penentuan Jumlah Kluster

Penentuan jumlah kluster dilakukan dengan menguji nilai $k = 2$ hingga $k = 6$ pada model RFM dan RFM + perilaku pembelian. Penentuan model terbaik mempertimbangkan grafik *Elbow Method* serta nilai *Silhouette Score*, *Davies-Bouldin Index (DBI)*, dan *Calinski-Harabasz Index (CHI)*.



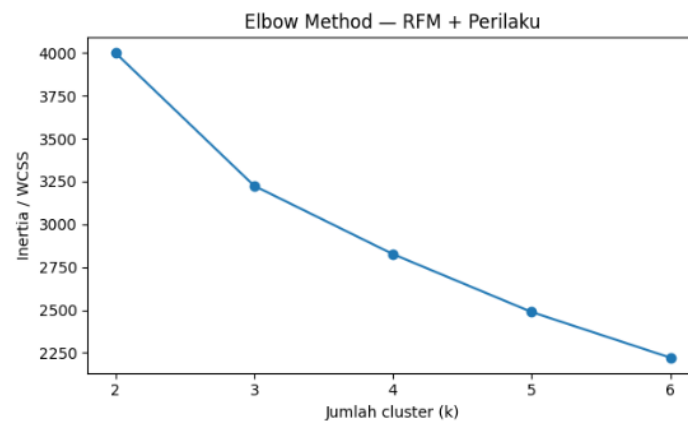
Gambar 3.5 Elbow Method Model RFM

Berdasarkan Gambar 3.5, grafik Elbow Method pada model RFM menunjukkan penurunan nilai inertia yang cukup tajam pada $k = 2$ hingga $k = 4$, kemudian mulai melandai pada $k = 4$ hingga $k = 6$. Pola ini mengindikasikan bahwa titik siku (elbow) berada pada $k = 4$, sehingga empat kluster dapat dipertimbangkan sebagai kandidat jumlah kluster optimal berdasarkan metode ini.



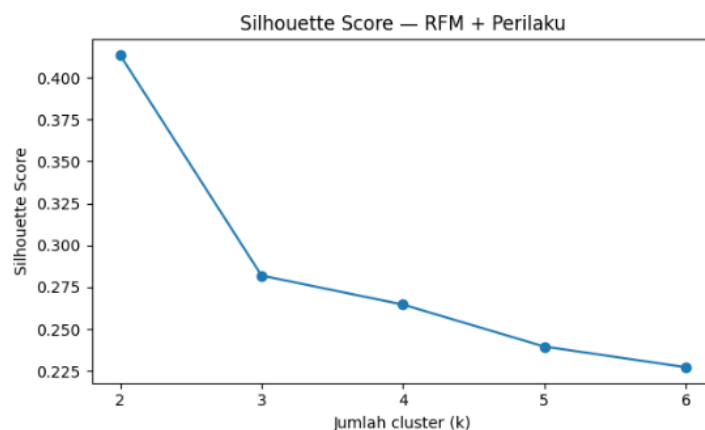
Gambar 3.6 Silhouette Score Model RFM

Berdasarkan Gambar 3.6, nilai Silhouette Score tertinggi diperoleh pada $k = 2$, yaitu sebesar 0,4513. Nilai tersebut menunjukkan bahwa dua kluster menghasilkan kualitas pemisahan terbaik dibandingkan jumlah kluster lainnya pada model RFM.



Gambar 3.7 Elbow Method Model RFM + Perilaku Pembelian

Berdasarkan Gambar 3.7, grafik Elbow Method pada model RFM + perilaku pembelian menunjukkan penurunan inertia paling tajam dari $k = 2$ ke $k = 3$. Setelah itu, penurunan inertia cenderung lebih landai hingga $k = 6$. Pola ini mengindikasikan bahwa titik siku (elbow) berada pada $k = 3$, sehingga tiga kluster dapat dipertimbangkan sebagai kandidat jumlah kluster optimal berdasarkan metode ini.



Gambar 3.8 Silhouette Score Model RFM + Perilaku Pembelian



Gambar 3.8 menunjukkan bahwa nilai Silhouette Score tertinggi pada model RFM + perilaku pembelian juga diperoleh pada $k = 2$, yaitu sebesar 0,4134. Nilai ini lebih rendah daripada model RFM yang mencapai 0,4513. Dengan demikian, berdasarkan Silhouette Score, model RFM menghasilkan pemisahan klaster yang lebih baik. Perbandingan selanjutnya dilakukan menggunakan nilai DBI dan CHI untuk mendukung pemilihan model akhir.

Untuk memperoleh penilaian yang lebih menyeluruh, hasil evaluasi kedua skenario dibandingkan menggunakan Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index. Hasil evaluasi masing-masing kandidat jumlah klaster disajikan pada Tabel 3.9 dan Tabel 3.10 berikut.

Tabel 3.9 Hasil Evaluasi Model RFM

Jumlah klaster (k)	Silhouette Score	DBI	CHI
2	0,4513	0,8814	1.363,874
3	0,3590	1,0690	1.077,833
4	0,3650	0,9910	1.040,790
5	0,3340	1,0130	941,631
6	0,3110	1,0300	897,155

Berdasarkan Tabel 3.9, konfigurasi $k = 2$ menghasilkan Silhouette Score tertinggi sebesar 0,4513, DBI terendah sebesar 0,8814, dan CHI tertinggi sebesar 1.363,874. Hasil ketiga metrik tersebut menunjukkan bahwa dua klaster merupakan konfigurasi terbaik pada model RFM di antara jumlah klaster yang diuji. Dibandingkan konfigurasi lainnya, $k = 2$ memiliki pemisahan klaster yang lebih baik, nilai DBI yang lebih rendah, serta nilai CHI yang lebih tinggi.

Tabel 3.10 Hasil Evaluasi Model RFM + Perilaku Pembelian

Jumlah klaster (k)	Silhouette Score	DBI	CHI
2	0,4134	0,9563	1.240,991
3	0,2820	1,2680	925,754
4	0,2650	1,2800	764,988
5	0,2400	1,2840	695,324
6	0,2270	1,3100	653,720

Berdasarkan Tabel 3.10, konfigurasi $k = 2$ menghasilkan Silhouette Score tertinggi sebesar 0,4134, DBI terendah sebesar 0,9563, dan CHI tertinggi sebesar 1.240,991. Ketiga metrik tersebut menunjukkan bahwa dua klaster merupakan konfigurasi terbaik pada model RFM + perilaku pembelian di antara jumlah klaster yang diuji.

Selanjutnya, perbandingan dilakukan terhadap konfigurasi terbaik dari masing-masing skenario, yaitu $k = 2$. Hasil perbandingan disajikan pada Tabel 3.11 berikut.

Tabel 3.11 Perbandingan Hasil Evaluasi Model RFM dan RFM + Perilaku Pembelian

Indikator evaluasi	RFM ($k=2$)	RFM + perilaku ($k=2$)	Model yang lebih baik
Silhouette Score	0,4513	0,4134	RFM
Davies-Bouldin Index	0,8814	0,9563	RFM
Calinski-Harabasz Index	1.363,874	1.240,991	RFM

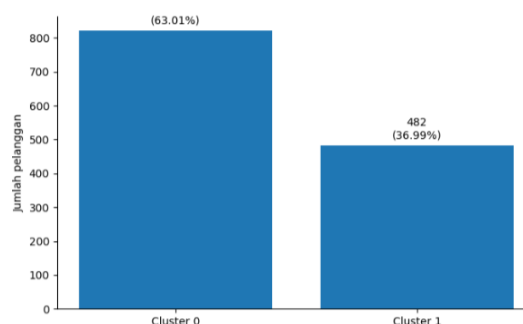
Berdasarkan Tabel 3.11, model RFM menghasilkan kualitas pengelompokan yang lebih baik dibandingkan model RFM + perilaku pembelian. Hal ini ditunjukkan oleh nilai Silhouette Score yang lebih tinggi, yaitu 0,4513 dibandingkan 0,4134, serta nilai DBI yang lebih rendah, yaitu 0,8814 dibandingkan 0,9563. Nilai CHI pada model RFM juga lebih tinggi, yaitu 1.363,874 dibandingkan 1.240,991. Ketiga indikator tersebut menunjukkan bahwa penambahan fitur perilaku pembelian belum meningkatkan kualitas kluster. Oleh karena itu, model K-Means dengan fitur RFM dipilih sebagai model akhir untuk menganalisis karakteristik segmen pelanggan dan menyusun rekomendasi pemasaran bagi PT XYZ.

3.6 Hasil Pengelompokan Pelanggan

Berdasarkan hasil evaluasi, model K-Means dengan fitur RFM dan jumlah kluster ($k=2$) dipilih sebagai model akhir. Pengelompokan menghasilkan dua kluster dengan jumlah anggota yang berbeda, sebagaimana disajikan pada Tabel 3.12 berikut.

Tabel 3.12 Distribusi Pelanggan pada Model RFM

Kluster	Jumlah Pelanggan	Persentase
Cluster 0	821	63,01%
Cluster 1	482	36,99%
Total	1.303	100,00%



Gambar 3.9 Distribusi Jumlah Pelanggan per Kluster



Berdasarkan Tabel 3.12 dan Gambar 3.9, Cluster 0 memiliki jumlah anggota lebih besar, yaitu 821 pelanggan (63,01%), sedangkan Cluster 1 terdiri atas 482 pelanggan (36,99%). Hasil ini menunjukkan bahwa distribusi pelanggan lebih terkonsentrasi pada Cluster 0.

3.7 Profiling Klaster

Profil setiap klaster dianalisis berdasarkan nilai median fitur RFM dan karakteristik perilaku pembelian untuk mengidentifikasi perbedaan antarsegmen pelanggan. Hasil perbandingan median setiap fitur disajikan pada Tabel 3.13 berikut.

Tabel 3.13 Profil Median Pelanggan pada Setiap Klaster

Fitur	Cluster 0	Cluster 1
Recency (hari)	56	9
Frequency (transaksi)	1	5
Monetary (Rp)	1.220.000	10.314.500
Product Diversity	3	14
Average Basket Size	6	13,82
Brand Diversity	2	7

Berdasarkan Tabel 3.13, Cluster 0 memiliki median *Recency* 56 hari, *Frequency* 1 transaksi, dan *Monetary* Rp1.220.000. Nilai tersebut menunjukkan bahwa pelanggan pada klaster ini cenderung kurang aktif bertransaksi. Selain itu, median *Product Diversity* sebanyak 3 produk, *Average Basket Size* sebesar 6 item, dan *Brand Diversity* sebanyak 2 merek menunjukkan variasi pembelian yang relatif rendah.

Sebaliknya, Cluster 1 memiliki median *Recency* 9 hari, *Frequency* 5 transaksi, dan *Monetary* Rp10.314.500. Klaster ini juga memiliki median *Product Diversity* sebanyak 14 produk, *Average Basket Size* sebesar 13,82 item, dan *Brand Diversity* sebanyak 7 merek. Hasil tersebut menunjukkan bahwa pelanggan Cluster 1 lebih aktif bertransaksi dengan nilai dan variasi pembelian yang lebih tinggi.

Berdasarkan karakteristik tersebut, perusahaan dapat menerapkan promosi pembelian ulang dan rekomendasi produk untuk Cluster 0, sedangkan Cluster 1 dapat dipertahankan melalui program loyalitas dan penawaran khusus. Strategi ini diharapkan dapat meningkatkan aktivitas transaksi sekaligus mempertahankan pelanggan aktif.

4. KESIMPULAN

Penelitian ini menerapkan algoritma K-Means untuk melakukan segmentasi pelanggan PT XYZ berdasarkan analisis *Recency*, *Frequency*, *Monetary* (RFM) dan perilaku pembelian selama



periode Maret-Juni 2026. Hasil evaluasi menunjukkan bahwa model RFM dengan $k = 2$ menghasilkan kualitas klaster lebih baik dibandingkan model RFM + perilaku pembelian, dengan nilai *Silhouette Score* sebesar 0,4513, DBI sebesar 0,8814, dan CHI sebesar 1.363,874. Segmentasi menghasilkan dua kelompok, yaitu Cluster 0 sebanyak 821 pelanggan dengan aktivitas transaksi relatif rendah dan Cluster 1 sebanyak 482 pelanggan dengan frekuensi, nilai, serta keragaman pembelian yang lebih tinggi. Hasil ini menunjukkan bahwa RFM dapat digunakan untuk mengidentifikasi karakteristik pelanggan sebagai dasar penyusunan strategi pemasaran. Perusahaan dapat mendorong pembelian ulang pada Cluster 0 serta mempertahankan pelanggan Cluster 1 melalui program loyalitas dan penawaran khusus.

Penelitian selanjutnya disarankan menggunakan data transaksi dengan rentang waktu lebih panjang serta menambahkan data interaksi, seperti riwayat pencarian produk dan kunjungan halaman, agar karakteristik perilaku pelanggan dapat dianalisis lebih menyeluruh. Evaluasi berkala terhadap strategi pemasaran yang diterapkan juga diperlukan untuk mengetahui efektivitasnya dalam meningkatkan pembelian ulang dan mempertahankan pelanggan aktif.

DAFTAR PUSTAKA

- [1] Oktavian, V. D., Ramadhan, R., & Fadhilla, D. R. (2025). Segmentasi pelanggan berbasis RFM dengan algoritma K-Means pada data transaksi online retail. *Jurnal Riset Informatika dan Teknologi Informasi*, 2(3), 236-243. <https://doi.org/10.58776/jriti.v2i3.156>
- [2] Awaliyah, D. A., Prasetyo, B., Muzayanah, R., & Lestari, A. D. (2024). Optimizing customer segmentation in online retail transactions through the implementation of the K-means clustering algorithm. *Scientific Journal of Informatics*, 11(2), 539-548. <https://doi.org/10.15294/sji.v11i2.6137>.
- [3] Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking-An effective approach to customer segmentation. *Journal of King Saud University Computer and Information Sciences*, 33(10), 1251-1257. <https://doi.org/10.1016/j.jksuci.2018.09.004>.
- [4] Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34(5), 1785-1792. <https://doi.org/10.1016/j.jksuci.2019.12.011>.
- [5] Naufal, R., & Alamsyah (2026). Customer segmentation using RFM analysis and the K-Means clustering algorithm to support data-driven marketing strategies. *Information Technology Education Journal*, 78-95. <https://doi.org/10.59562/intec.v5i3.12821>
- [6] Mado, P. M. K., & Hendry, H. (2025). Implementasi algoritma clustering K-Means untuk segmentasi pelanggan di e-commerce. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, 6(3), 1680-1686. <https://doi.org/10.63447/jimik.v6i3.1563>
- [7] Zai, N., Kenap, A. A., & Hasibuan, A. (2026). Perbandingan Clustering Berbasis RFM dan Implementasi K-Means untuk Segementasi Pelanggan Bisnis Laundry. *SemanTIK: Teknik Informasi*, 12(1), 22-32. <https://doi.org/10.55679/semantik.v12i1.275>
- [8] Utami, I. T., Suryaningrum, F., & Ispriyanti, D. (2023). K-means cluster count optimization with silhouette index validation and davies bouldin index (CASE study: Coverage of pregnant women, childbirth, and postpartum health services in indonesia in 2020). *BAREKENG J. Ilmu Mat. dan Terap*, 17(2), 0707-0716. <https://doi.org/10.30598/barekengvol17iss2pp0707-0716>
- [9] Anfossi, A., & Chicco, D. (2026). An easy guide to the Davies-Bouldin index for



unsupervised internal clustering evaluation. *Discover Computing*, 29(1), 212. <https://doi.org/10.1007/s10791-026-10094-0>

- [10] Freyhult, E., Landfors, M., Önskog, J., Hvidsten, T. R., & Rydén, P. (2010). Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC bioinformatics*, 11(1), 503. <https://doi.org/10.1186/1471-2105-11-503>.