

Analisis Segmentasi Pelanggan *Marketplace* Toko Retail XYZ di Yogyakarta Berbasis RFM dan Algoritma *K-Means Clustering*

Rezky Amymaztura Adz Dzaariyat Erzam¹, Irfan Pratama²

^{1,2}Sistem Informasi, Teknologi Informasi, Universitas Mercu Buana Yogyakarta

¹211210100@student.mercubuana-yogya.ac.id*, ²irfanp@mercubuana-yogya.ac.id

Abstract

The development of the shopee marketplace in Indonesia has made one of XYZ's online retail stores who are engaged in the fashion sector in Yogyakarta need an effective marketing strategy to be on target. The store's shopee transaction data only ends up as a sales archive and has not been processed optimally, so it has difficulty in grouping and understanding customer behavior. The research is aimed at segmenting customer behavior using an analytical Customer Relationship Management (CRM) approach with the RFM (Recency, Frequency, Monetary) model and K-Means Clustering Algorithm. The research used 85,781 transaction data for the 2025 period, which was then carried out in the preprocessing and outlier cleaning stage with the Interquartile Range (IQR) so that it became 42,222 data. A combined evaluation was carried out to determine the optimal cluster using the Elbow, Silhouette Score and Davies-Bouldin Index methods. The study determined the result of $k=3$ as the optimal number of clusters with elbow points on the graph shown on the Elbow graph and validated using the Silhouette Score with the highest value of 0.4979 and the minimum value of the Davies-Bouldin Index of 0.7429. The majority of customers have the characteristics of One-time buyers, based on the relatively same frequency values so that the value of the recency and monetary variables affects segmentation more. The resulting segments are Cluster 0 (new customer / potential), Cluster 1 (at risk customer), Cluster 2 (lost customer). Research has succeeded in converting thousands of transaction data into strategic information by describing the characteristics of each customer segment so that companies can implement appropriate marketing strategies.

Keywords: clustering, k-means, RFM, segmentation, shopee.

Abstrak

Perkembangan *marketplace* shopee di Indonesia membuat salah satu toko *retail online* XYZ yang bergerak di bidang fashion di Yogyakarta memerlukan strategi pemasaran yang efektif agar tepat sasaran. Data transaksi shopee toko tersebut hanya berakhir menjadi arsip penjualan dan belum diolah secara optimal sehingga mengalami kesulitan dalam mengelompokkan dan memahami perilaku pelanggan. Penelitian ditujukan untuk melakukan segmentasi perilaku pelanggan menggunakan pendekatan Customer Relationship Management (CRM) analitik dengan model RFM (Recency, Frequency, Monetary) dan algoritma *k-means clustering*. Penelitian menggunakan 85.781 data transaksi periode 2025, yang kemudian dilakukan tahap *preprocessing* dan pembersihan *outlier* dengan *Interquartile Range* (IQR) sehingga menjadi 42.222 data. Evaluasi gabungan dilakukan untuk menentukan kluster optimal menggunakan metode *Elbow*, *Silhouette Score* dan *Davies-Bouldin Index*. Penelitian menetapkan hasil $k=3$ sebagai jumlah kluster optimal dengan titik siku pada grafik yang ditunjukkan pada grafik *Elbow* serta divalidasi menggunakan *Silhouette Score* dengan nilai paling tinggi 0,4979 dan nilai minimum *Davies-Bouldin Index* 0,7429. Mayoritas pelanggan memiliki karakteristik *one-time buyers*, didasari oleh nilai *frequency* yang relatif sama sehingga nilai variabel *recency* dan *monetary* lebih mempengaruhi segmentasi. Segmen yang dihasilkan adalah Kluster 0 (*new customer / potential*), Kluster 1 (*at risk customer*), Kluster 2 (*lost customer*). Penelitian berhasil mengubah ribuan data transaksi menjadi informasi yang strategis dengan menggambarkan karakteristik setiap segmen pelanggan sehingga perusahaan dapat menerapkan strategi pemasaran dengan tepat.

Kata kunci: klastering, k-means, RFM, segmentasi, shopee.

©This work is licensed under a Creative Commons Attribution -ShareAlike 4.0 International License

1. Pendahuluan

Salah satu Platform E-commerce dan Marketplace yang berkembang pesat beberapa tahun terakhir yaitu Shopee yang juga merupakan marketplace terbesar dan populer di kalangan masyarakat. Shopee banyak memberikan promo dan diskon menarik yang menjadikan minat belanja online pelanggan menjadi cukup tinggi [1]. Semakin meningkatnya penggunaan marketplace, penting bagi pelaku usaha untuk memiliki strategi pemasaran yang tepat sasaran guna mempertahankan potensi pelanggan dengan melakukan penerapan teknologi untuk promosi yang lebih efektif [2]. Perkembangan tersebut membuat pelaku usaha berupaya

memaksimalkan penjualan mereka di bidang online. Oleh sebab itu strategi pemasaran dan promosi yang dilakukan harus tepat. Ditengah banyaknya persaingan di industri fashion, toko retail online XYZ yang juga merupakan salah satu brand lokal dengan ratusan produk mulai dari dompet, tas dan jam tangan bahkan cabang tokonya sudah banyak dibuka menyeluruh di beberapa daerah di Indonesia.

Memiliki loyalitas pelanggan yang cukup tinggi, Toko retail online XYZ memerlukan analisis mengenai segmentasi pelanggan, yaitu membagi pelanggan menjadi beberapa kelompok yang setiap kelompoknya memiliki perilaku pembelian atau karakteristik yang serupa (ho-

mogen) [3]. Oleh sebab itu, penelitian ini bertujuan untuk membangun efektifitas dan efisiensi penerapan strategi promosi dan pemasaran untuk perusahaan [4].

Disamping itu, data transaksi penjualan pelanggan *online* pada toko *retail online* XYZ saat ini hanya berakhir menjadi kumpulan catatan dan arsip administratif penjualan saja sehingga belum diolah secara optimal oleh perusahaan menjadi hasil strategis untuk penjualan dan segmentasi pelanggan. Karena belum adanya sistematisa pelanggan tersebut menyebabkan perusahaan mengalami kesulitan untuk menentukan pengelompokan pelanggan dengan loyalitas tinggi bahkan pelanggan yang berisiko hilang (*churn*), sehingga perusahaan akan kehilangan peluang dalam mempertahankan pelanggan.

CRM (*Customer Relationship Management*) adalah strategi dalam bisnis untuk mengelola hubungan antar organisasi dengan pelanggan [5]. Implementasi CRM (*Customer Relationship Management*) yang digunakan pada penelitian ini yaitu CRM analitik berbasis data *mining* yang mengutamakan proses mengolah data dengan menghasilkan pengetahuan baru yang dilakukan melalui pendekatan RFM (*Recency, Frequency, Monetary*). Model RFM dilakukan dalam pengelompokan pelanggan berdasar pola pembelian yang meliputi pengukuran jarak transaksi terakhir pelanggan (*Recency*), intensitas pembelian pelanggan (*frequency*) dan total pengeluaran pelanggan untuk berbelanja (*monetary*) [6]. Parameter RFM akan digabungkan dengan algoritma *K-Means clustering* serta akan dikelompokkan ke dalam segmen berdasarkan persamaan perilaku atau karakteristiknya.

Penelitian yang dilakukan Muhamad Fikri Fadhillah dan rekan-rekan yang berjudul, “*Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM)*”, menghasilkan bahwa penggunaan variabel RFM dan algoritma *Clustering* didapatkan, bahwa dari data penjualan PT. SID Jawa Timur pada rentang Waktu dari Januari 2021- Desember 2023 terdiri dari 118.314 transaksi penjualan dan mencakup 1.570 pelanggan yang dilakukan *clustering* menggunakan *K-Means* dengan hasil terbaik dengan urutan nilai *silhouette score* 0.364, DBI 0.93 dan *calinski-harabaz* 1303.6. Data tersebut kemudian dilakukan proses *clustering* menjadi 3 klaster pelanggan yaitu klaster 0 berisi 694 pelanggan dan dikategorikan *loyal customer*, klaster 1 berisi 383 pelanggan dengan kategori *adequate customer* dan klaster 2 berisi 494 pelanggan dengan kategori pelanggan *churn* atau *lost customer*[7].

Teknik yang diterapkan pada data *mining* adalah algoritma *K-Means Clustering*. *K-Means Clustering* digunakan sebagai bagian utama dalam pemodelan yang dapat menghasilkan kelompok (*cluster*) pada dataset. Beberapa pengelolaan dan pengaturan dalam *K-Means* seperti pengaturan jumlah Klaster (K), penentuan letak awal (*centroid*) dan waktu proses selesai harus di sesuaikan supaya kinerja algoritma berjalan opti-

mal [4]. Pembelajaran tanpa pengawasan (*Unsupervised Learning*) bertujuan untuk melakukan pengelompokkan data yang tidak memiliki label ke dalam kelompok yang memiliki persamaan karakteristik tertentu. *K-Means* merupakan salah satu algoritma yang sangat populer digunakan karena *K-Means clustering* membagi dataset dengan kesamaan ke dalam beberapa *cluster*[6].

Penelitian ini mencakup tahap pra-pemrosesan data transaksi shopee toko *retail online* XYZ untuk menghasilkan nilai RFM, menerapkan metode *K-Means clustering* segmentasi pelanggan dan menghasilkan jumlah *cluster* menggunakan evaluasi gabungan dengan *Elbow Method* dan kemudian divalidasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index (DBI)*. Analisis yang dihasilkan diharapkan dapat menentukan karakteristik pada setiap segmen pelanggan untuk memperoleh keputusan strategis bagi manajemen perusahaan. Selain dapat mengubah kumpulan data transaksi menjadi sebuah informasi yang bermanfaat dan menguntungkan guna mengoptimalkan biaya dan strategi pemasaran serta promosi agar lebih tepat sasaran, penelitian ini juga memiliki manfaat dalam kontribusi di bidang CRM serta integrasi pendekatan RFM dan *K-Means Clustering* untuk membantu perusahaan menghasilkan segmentasi pelanggan yang akurat sehingga efektivitas pemasaran meningkat [8].

Selanjutnya pada penelitian yang dilakukan oleh Muhammad Fahri Ilman, Feri Suliarta dengan judul, “*Segmentasi Pelanggan Perbankan Menggunakan Metode Clustering K-Means*”, mendapatkan hasil pengelompokkan pelanggan menjadi dua *cluster* yaitu klaster 0 yang memiliki pelanggan dengan pengeluaran rata-rata yang lebih tinggi namun frekuensi transaksi rendah dibanding klaster 1. Oleh sebab itu pelanggan yang memiliki nilai lebih tinggi dari *cluster* 0 membutuhkan strategi khusus dalam menentukan promosi dan pemasaran.

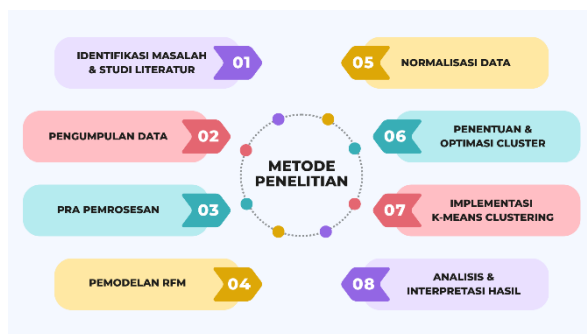
Ada beberapa metode *clustering* lain yang digunakan untuk melakukan pengelompokkan data. Pada penelitian yang dilakukan Yoel Michael Sihombing dan Eka Sri Rahayu, yang berjudul “*Implementasi Fuzzy C-Means Clustering dalam Segmentasi Pelanggan dengan RFM Model*,”. Penggunaan *Fuzzy C-Means* sebagai metode utama *clustering* untuk mengelompokkan data berdasar derajat keanggotaan. Metode ini digunakan apabila terdapat kondisi ketidakpastian pada batas antar kelompok atau data sehingga *Fuzzy C-Means* memiliki fleksibilitas dalam menentukan kelompok meliputi pengolahan citra, pengenalan pola serta analisis data[9].

Berdasarkan keterbatasan pengolahan data administratif yang ada di toko *retail* XYZ serta landasan dari berbagai penelitian terdahulu, penggabungan model RFM dan algoritma *K-Means clustering* dinilai sebagai solusi untuk efisiensinya dalam mengolah data berskala besar tanpa label (*unsupervised learning*). Oleh sebab itu, penelitian ini bertujuan melakukan segmentasi per-

ilaku pembeli produk fashion pada shopee toko *retail XYZ* di Yogyakarta secara akurat. Melalui integrasi variabel RFM, penelitian ini diharapkan dapat mengelompokkan karakteristik unik dari setiap kelompok pelanggan, mengidentifikasi pelanggan yang berisiko hilang (*churn*), serta membuat rekomendasi strategi *Customer Relationship Management* (CRM) analitik yang tepat sasaran demi meningkatkan efektivitas promosi dan retensi pelanggan perusahaan.

2. Metode Penelitian

Tahap dan proses penelitian dilakukan secara sistematis. Pada *data mining*, kerangka kerja yang digunakan mencakup pengumpulan data (*data collection*), *data transformation* dan *data analysis* [10] Proses dan tahapan yang dilakukan pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Tahap Penelitian

2.1. Identifikasi Masalah & Studi Literatur

Identifikasi masalah berisi latar belakang penelitian dan rumusan masalah yang di dapat, dimana data transaksi penjualan shopee toko *retail online XYZ* belum dimanfaatkan secara optimal untuk analisis dalam memahami karakteristik pelanggan. Oleh karena itu penelitian ini dilakukan untuk menganalisis data guna menghasilkan segmentasi pelanggan berdasarkan kesamaan karakteristik sehingga perusahaan dapat menerapkan strategi pemasaran yang sesuai pada masing-masing kelompok. Sementara studi literatur dilakukan dengan pengumpulan teori, konsep dan evaluasi dari penelitian terdahulu yang relevan dengan topik penelitian ini. Landasan teori meliputi variabel RFM (*Recency, Frequency, Monetary*), algoritma *K-Means Clustering* yang digunakan sebagai metode utama pengelompokan data, validitas metode, parameter yang digunakan serta teori terdahulu yang menganalisis terkait segmentasi pelanggan.

2.2. Perangkat Penelitian

Implementasi pemodelan dalam mengolah data penelitian ini dilakukan menggunakan lingkungan pengembangan berbasis *cloud* dengan *google colab* dan bahasa pemrograman *python* serta pustaka (*library*) yang digunakan diantaranya yaitu *pandas & numPy, scikit-learn* guna normalisasi data, evaluasi klaster dan implementasi *K-Means* serta *matplotlib & seaborn* untuk visualisasi hasil analisis.

2.3. Pengumpulan Data

Sumber data penelitian yang digunakan yaitu data transaksi penjualan yang diperoleh dari shopee toko XYZ yang di ekspor melalui *Seller Centre* pada periode 1 Januari – 31 Desember 2025 dengan total sebanyak 85.781 baris data transaksi. Dikarenakan atribut dataset asli cukup banyak, maka dilakukan pemilihan atribut yang berkaitan dengan proses transformasi RFM yang di deskripsikan pada Tabel 1 dibawah ini:

Tabel 1. Deskripsi Atribut

Nama Atribut	Deskripsi	Variabel RFM
Username	ID unik pelanggan shopee	Identify
No. Pesanan	Kode unik transaksi	Frequency
Waktu Pesanan Dibuat	Menentukan nilai Recency	Recency
Total Pembayaran	Menentukan nilai Monetary	Monetary

Kemudian mengenai gambaran data mentah yang digunakan penelitian ini, ditampilkan dalam beberapa baris data pada Tabel 2.

Tabel 2. Dataset Pelanggan Toko Retail XYZ

Username	Waktu Pesanan	No. Pesanan	Total Bayar
.aira28	2025-01-01	250101AJ...	145.000
.ekka	2025-01-05	250105BX...	118.000
.2017	2025-01-10	250110ZL...	125.000
.marshanda04	2025-01-12	250112KM...	65.000
...	...	...	...
dessykelele	2025-12-31	251231SM..	134.800
syafiqaputri..	2025-12-31	251231TY..	102.500
yanara1103	2025-12-31	2512317H..	172.630

2.4. Pra Pemrosesan Data

Sebelum diolah ke dalam model RFM dan *clustering*, dilakukan tahap pra pemrosesan atau *pre-processing* data guna memastikan kualitas data. Tahapan yang dilakukan meliputi *filtering* data yaitu tahap untuk mengambil transaksi yang hanya berstatus “selesai” dan menghapus atribut-atribut yang tidak diperlukan dalam menghitung RFM seperti alamat pengiriman atau catatan pesanan. Kemudian data *cleaning* yaitu mengatasi *missing values* (data kosong) dan menghapus data duplikat. *Data aggregation* dilakukan untuk mentransformasi data mentah menjadi profil perilaku pelanggan dengan mengelompokkan berdasar atribut *username*. Pada tahap ini setiap variabel dilakukan perhitungan berdasar nomor pesanan. *Recency* menghitung selisih hari pelanggan melakukan pembelian, *frequency* menghitung kemunculan nomor pesanan unik setiap *username* dengan mempertahankan frekuensi pembelian berulang serta *monetary* menghitung total akumulasi pembelanjaan pelanggan. serta konversi tipe data untuk mengubah tipe data kolom tanggal menjadi *datetime* untuk memudahkan menghitung selisih hari dan variabel nominal harga bertipe numerik.

2.5. Pemodelan Variabel RFM

Data transaksi yang sebelumnya sudah dibersihkan, selanjutnya ditransformasi ke dalam variabel RFM, dian-

taranya yaitu *recency* untuk mengukur rentang waktu atau jarak hari dengan transaksi terakhir, *frequency* untuk menghitung total frekuensi atau seberapa banyak pelanggan melakukan pembelian dan *monetary* yaitu menghitung total akumulasi atau total nilai ekonomi yang dikeluarkan oleh pelanggan.

2.6. Normalisasi Data

Ketiga variabel RFM memiliki satuan yang berbeda, yaitu *recency* (nilai hari), *frequency* (nilai jumlah) dan *monetary* (nilai rupiah). Sebelumnya data dibersihkan terlebih dahulu dari *oulier* menggunakan *Interquartile Range* (IQR), kemudian tahap akhir yaitu normalisasi. Proses ini untuk mempercepat pembelajaran yaitu dengan transformasi logaritma dengan *min-max scaling* untuk menyamakan skala variabel dengan mengubah rentang nilai 0 hingga 1 agar seimbang sehingga algoritma dapat bekerja secara efektif.

Log-transform dengan *min-max scaling*:

$$X_{norm} = \frac{\ln(X + 1) - \ln(X_{min} + 1)}{\ln(X_{max} + 1) - \ln(X_{min} + 1)} \quad (1)$$

Nilai X merepresentasikan nilai asli variabel sebelum transformasi, sedangkan nilai X_{norm} menunjukkan hasil akhir proses normalisasi yang berada pada skala 0 hingga 1. X_{min} menyatakan nilai minimum variabel data asli sementara X_{max} yaitu nilai maksimum variabel data asli dan simbol $\ln$ menunjukkan operasi logaritma natural.

Penggunaan transformasi logaritma ini ditujukan untuk mengatasi ketidakseimbangan distribusi data yang ekstrem pada variabel *frequency* dan *monetary* serta rentang nilai dengan skala yang cukup besar agar nilai lebih rapat dan terdistribusi normal dan tidak bias saat melakukan perhitungan jarak *euclidean* pada algoritma K-Means.

2.7. Penentuan Kluster Optimal

Penentuan jumlah kluster optimal dalam implementasi algoritma K-Means dilakukan dengan menguji menggunakan rentang nilai $k=2$ hingga $k=10$ menggunakan tiga metode evaluasi metrik yaitu elbow method yang digunakan untuk mencari titik siku dengan melihat penurunan inertia (*within-cluster sum of squares*) [11]. *Silhouette score* didukung dengan nilai paling tinggi yang dihasilkan dengan cara mengukur kedekatan antar data dalam kluster [12].

Rumus *silhouette score* :

$$SC = \frac{1}{n} \sum_{i=1}^n s(i) \quad (2)$$

Pada persamaan 2, SC merupakan nilai *silhouette score* yang merupakan nilai rata-rata seluruh kluster dan n menunjukkan total data yang digunakan.

Davies-Bouldin Index (DBI) menghitung rata-rata kesamaan *cluster* yaitu jarak dalam *cluster* dengan jarak antar *cluster* dari nilai 0 hingga 1, dimana jika nilai

semakin mendekati 0 maka pengelompokkan yang dihasilkan semakin baik [13]. Pengukuran nilai minimum pemisahan antar kluster sebagai validasi guna memastikan objektivitas, maka jumlah kluster yang dihasilkan juga semakin optimal seperti yang dirumuskan pada Persamaan 3 berikut:

Rumus *Davies-Bouldin Index*:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + M_i}{M_j} \right) \quad (3)$$

Nilai DBI menunjukkan indeks *Davies-Bouldin* yang didapatkan, variabel k yaitu jumlah kluster yang terbentuk, kemudian S menunjukkan nilai sebaran objek pada kluster serta M sebagai jarak pemisah antar kluster.

2.8. Implementasi Algoritma K-Means

Jumlah kluster optimal (k) yang sudah di dapatkan, kemudian diimplementasi menggunakan algoritma *K-Means Clustering* dengan beberapa tahapan. Langkah awal dalam K-Means yaitu menentukan jumlah pusat kluster (*centroid*) menggunakan perhitungan *euclidean distance* seperti ditunjukkan pada Persamaan 4.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

$d(x, y)$ merepresentasikan jarak antar data (x) dengan pusat kluster (y). Kemudian dilakukan proses pengelompokan data ke *centroid* terdekat dan menghitung nilai rata-rata seluruh data pada kluster dan dilakukan secara berulang (iterasi) hingga posisi *centroid* tidak mengalami perubahan.

2.9. Analisis & Interpretasi Hasil

Proses terakhir yaitu analisis dan interpretasi hasil dan *profiling* dari kluster yang telah dihasilkan. Sebelumnya RFM yang di normalisasi, dikembalikan terlebih dahulu ke nilai asli untuk menghitung rata-rata setiap kluster dan analisis karakteristik untuk memahami perilaku belanja pelanggan. Hasil analisis kluster selanjutnya diberi label segmentasi seperti pelanggan baru/loyal, potensial dan berisiko yang digunakan untuk analisis rekomendasi dalam strategi pemasaran perusahaan.

3. Hasil dan Pembahasan

Analisis data dan tahapan proses penelitian yang telah dilakukan, diperoleh hasil penelitian sebagai berikut :

3.1. Analisis Data Transaksi

Dari data mentah sebanyak 85.781 baris data, dilakukan tahap *filtering* untuk mengambil data pesanan yang hanya berstatus “selesai” sehingga diperoleh 67.773 baris data. Kemudian data diproses melalui tahap agregasi untuk proses pengelompokkan data transaksi berdasarkan username pelanggan.

Perhitungan *frequency* dilakukan berdasarkan jumlah nomor pesanan unik (*nunique*) pada setiap akun username sehingga menghasilkan sebanyak 50.850 baris data pelanggan tunggal. Kemudian data ditransformasi ke dalam variabel RFM (*Recency*, *Frequency*, dan *Monetary*).

Pada tahap pra-pemrosesan data, dilakukan pembersihan outlier menggunakan metode IQR dengan batas standar $Q1 - 1.5 \times IQR$ untuk batas bawah dan $Q3 + 1.5 \times IQR$ untuk batas atas. Proses ini menghasilkan sebanyak 42.222 baris data yang akan digunakan untuk proses klustering. Terkait penurunan yang cukup signifikan setelah proses pembersihan pencilan (*outlier*) menggunakan metode *Interquartile Range* (IQR) sehingga didapatkan sebanyak 8.628 data atau sekitar 16,9% data yang tereliminasi dari hasil evaluasi isolasi variabel *frequency* yang menunjukkan nilai 1.0. pendekatan ini dilakukan untuk menghindari efek akumulasi pembuangan data yang berlebihan, di mana satu orang pelanggan hanya di eliminasi jika nilainya berada pada luar batas wajar masa transaksi terakhir (*recency*) atau nilai pembelian (*monetary*).

Melalui metode isolasi variabel *frequency*, kelompok pelanggan dengan frekuensi pembelian tinggi tidak dapat dipastikan tidak akan terbuang secara tidak sengaja oleh algoritma. Namun pembersihan IQR pada variabel *monetary* tetap melakukan eliminasi pada beberapa akun dengan nilai transaksi yang ekstrem dan terlampaui besar seperti akun distributor atau reseller. Tahap ini digunakan untuk mempertahankan performa algoritma K-Means agar pusat kluster tidak bias akan nilai ekstrem dan tetap dapat merepresentasi karakteristik segmen pelanggan toko *retail XYZ* dengan lebih akurat.

Hasil evaluasi data setelah pembersihan menunjukkan bahwa nilai pada variabel *frequency* tetap menunjukkan nilai 1.0 dengan *standard deviation* sebesar 0.0. Kondisi ini membuktikan bahwa hal tersebut bukan disebabkan oleh kesalahan pada algoritma IQR namun murni karakteristik dari pelanggan toko *retail XYZ* pada platform Shopee dalam periode penelitian ini yang didominasi pembeli satu kali (*one-time buyers*).

3.2. Transformasi RFM dan Normalisasi Data

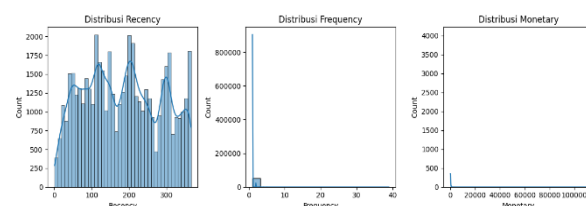
Transformasi variabel RFM yang telah dilakukan dinormalisasi untuk mengubah nilai menjadi rentang 0 hingga 1. Tabel 3 menunjukkan perbandingan antara data sebelum dan sesudah proses normalisasi.

Tabel 3. Perbandingan Nilai RFM

Username	R	F	M	R_Norm	F_Norm	M_Norm
.2017	24	1	125.0	0.48483	0.0	0.80987
.aira28	34	1	145.0	0.98754	0.0	0.84179
.ekka	34	1	118.0	0.98698	0.0	0.79749
.lis-nawati	34	1	78.0	0.98698	0.0	0.70872

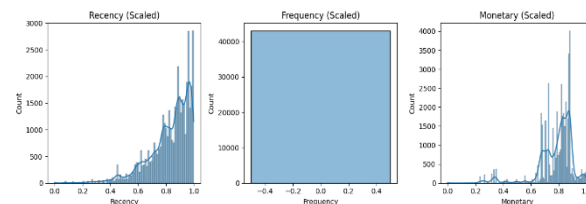
.mar-sanda	35	1	65.0	0.99575	0.0	0.66976
	7			8		4

Diperlukan analisis terhadap persebaran data pada variabel RFM sebelum melakukan pengelompokan dengan Algoritma K-Means. Proses ini digunakan untuk memastikan data telah sesuai dengan kriteria pemrosesan agar hasil kluster tidak bias. Berdasarkan Gambar 2 dibawah ini, terdapat *right-skewed* (kemiringan positif) yang signifikan pada sisi kanan karena penumpukan data pada nilai rendah dengan beberapa *outlier* pada variabel *frequency* dan *monetary* yang menandakan bahwa terdapat segmen pelanggan yang memiliki frekuensi dan total pengeluaran atau transaksi yang tinggi.



Gambar 2. Distribusi data RFM sebelum Transformasi

Untuk mengatasi ketidakseimbangan distribusi dan menyamakan skala variabel seperti *monetary* dengan satuan rupiah yang besar dan *frequency* dengan jumlah satuan kali yang lebih kecil, proses transformasi logaritma serta normalisasi menggunakan *min-max scaling* dilakukan. Hasil transformasi dapat dilihat pada Gambar 3 berikut ini.

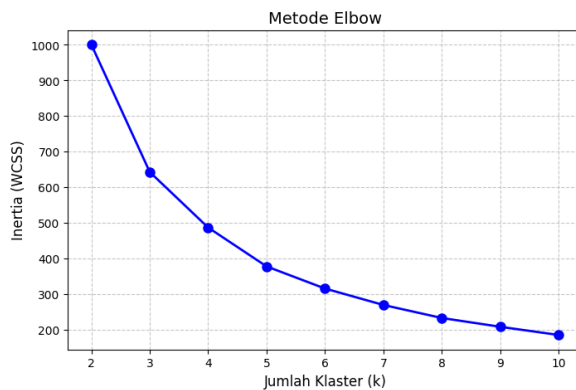


Gambar 3. Distribusi data RFM setelah Transformasi & Scaling

Setelah dilakukan normalisasi, terlihat persebaran data pada variabel *recency* dan *monetary* sudah berada pada skala yang merata yaitu pada rentang 0 hingga 1 sehingga perhitungan algoritma K-Means dalam menghitung jarak *eulclidean* antar variabel menjadi seimbang. Pada variabel *frequency* terdapat pemusatan data yang menunjukkan bahwa mayoritas pelanggan berupa pelanggan yang hanya melakukan satu kali pembelian (*one-time buyers*).

3.3 Penentuan Kluster Optimal

Pengujian dilakukan dengan penentuan nilai *k* menggunakan pemodelan *k = 2* hingga *k = 10*. Hasil yang didapatkan melalui grafik *elbow method* yaitu penurunan inertia (titik siku) pada *k = 3* yang ditunjukkan oleh Gambar 4.



Gambar 4. Grafik Metode Elbow

Setelah itu evaluasi dilakukan menggunakan metrik *silhouette score* dan *davies-bouldin index*. Hasil evaluasi ditampilkan pada Tabel 4.

Tabel 4. Evaluasi Hasil Klastering

No.	k	<i>Silhouette</i>	DBI
0	2	0.460300	1.017896
1	3	0.497874	0.742912
2	4	0.392881	0.876995
3	5	0.396485	0.823727
4	6	0.402030	0.791522
5	7	0.396691	0.789473
6	8	0.400000	0.801514
7	9	0.370030	0.847478
8	10	0.374435	0.834386

Dari hasil pada Tabel 4, di dapatkan bahwa nilai $k = 3$ oleh *silhouette score* menghasilkan pemisahan antar klaster yang maksimal dengan nilai cukup tinggi yaitu 0.4979 serta *davies-bouldin index* dengan kepadatan internal klaster yang baik sehingga menghasilkan nilai minimum yaitu 0.7429. Dengan demikian jumlah klaster optimal $k = 3$ digunakan sebagai tahap implementasi algoritma *K-Means* pada toko *retail XYZ*.

3.4 Implementasi K-Means dan Analisis Karakteristik

Setelah nilai optimal $k = 3$ di dapatkan, selanjutnya dilakukan pengelompokkan dengan membagi kelompok pelanggan berdasarkan kesamaan karakteristik dan perilaku RFM. Hasil nilai rata-rata RFM pada setiap klaster ditunjukkan pada Tabel 5.

Tabel 5. Profil Karakteristik Rata-Rata RFM

Cluster	<i>Recency</i>	<i>Frequency</i>	<i>Monetary</i>	Jumlah Pelanggan
0	58.320440	1.0	130.742173	9.902
1	229.968829	1.0	131.177622	31.022
2	146.238163	1.0	16.464426	2.112

Berdasar hasil pada Tabel 5, maka analisis kelompok pelanggan dari 3 klaster tersebut yaitu :

Klaster 0 (*new customer/potential*), pelanggan yang memiliki nilai *recency* rendah, rata-rata 58 hari dan *Monetary* dengan rata-rata nilai yang cukup tinggi yaitu Rp. 130.742. Hal tersebut menunjukkan bahwa mereka termasuk pelanggan paling baru dalam melakukan transaksi dari pada klaster lain sehingga masih memiliki potensi menjadi *loyal customer*.

Klaster 1 (*at risk customer*), pelanggan dengan nilai rata-rata *monetary* cukup tinggi yaitu Rp. 131.177 namun juga *recency* yang tinggi yaitu 229 hari atau lama tidak melakukan transaksi dan akan berisiko hilang (*churn*).

Klaster 2 (*lost customer*) yaitu kelompok yang berisi pelanggan dengan profit paling kecil dengan rata-rata *monetary* paling rendah yaitu Rp. 16.464 dan dengan rata-rata *recency* yang cukup tinggi, 146 hari. Kelompok pelanggan ini termasuk ke dalam kelompok pelanggan hilang karena sudah lama tidak melakukan transaksi dan nominal transaksi nya pun rendah.

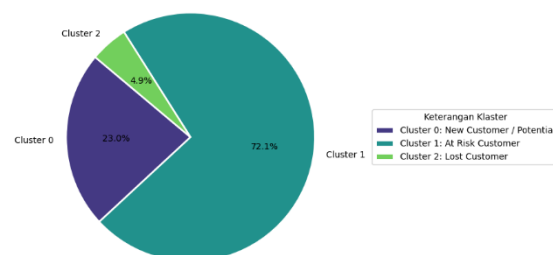
Berdasar karakteristik kelompok yang terbentuk, penelitian ini mengidentifikasi bahwa terdapat keterbatasan metodologis dari perspektif *Customer Relationship Management* (CRM) yaitu pelanggan yang di dominasi oleh *one-time buyers* dan menyebabkan variabel *frequency* kehilangan variasi pada ketiga klaster, sebagaimana ditunjukkan pada Tabel 5.

Kondisi tersebut menyebabkan implementasi algoritma *K-Means* pada penelitian ini hanya bertumpu pada variabel karakteristik *recency* dan *monetary* (RM) dalam memisahkan kelompok pelanggan. Meskipun demikian, penelitian ini tetap mempertahankan penggunaan RFM pada judul dan narasi karena secara metodologis tahapan awal ekstraksi tetap menggunakan variabel *frequency* dan kondisi *frequency* yang tetap bernilai 1.0 dilaporkan secara transparan sebagai batasan dari dataset riil yang digunakan.

3.5 Visualisasi Distribusi Segmen

Proses visualisasi digunakan untuk memaparkan perbedaan karakteristik setiap segmen. Visualisasi dilakukan melalui analisis menggunakan pendekatan visual, yaitu *Pie Chart*, *Boxplot* dan *Scatter Plot*.

1. *Pie Chart* (Proporsi Pelanggan Per Klaster)



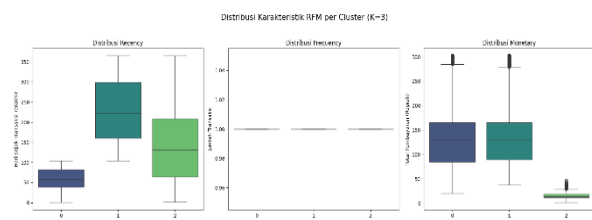
Gambar 5. *Pie Chart* Presentase Distribusi Pelanggan

Gambar *pie chart* tersebut merupakan jumlah distribusi pelanggan pada setiap klaster, dimana Klaster 0 (*new customer / potential*) berisi 23% dari total pelanggan yang merupakan kelompok yang masih memiliki potensi menjadi pelanggan potensial. Klaster 1 (*at risk customer*) merupakan kelompok dengan mayoritas pelanggan yang tidak aktif dengan proporsi 72,1% atau

sebanyak 31.022 pelanggan dan Klaster 2 (*lost customer*) berisi segmen terendah dengan proporsi 4,9% berisi pelanggan yang berisiko hilang karena sudah lama tidak bertransaksi dan total akumulasi belanja yang dilakukan paling rendah.

2. Boxplot (Sebaran Karakteristik RFM)

Gambar 6 menunjukkan hasil analisis sebaran data sebagai berikut :



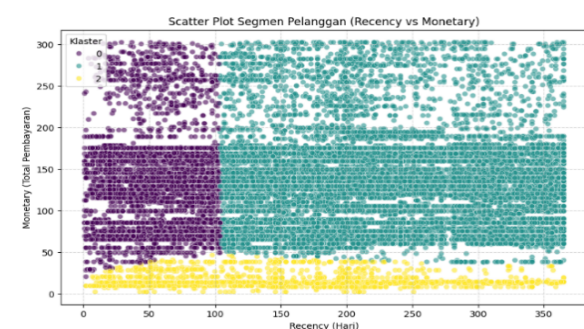
Gambar 6. Boxplot Karakteristik RFM

Pada *distribusi recency* terdapat perbedaan yang signifikan pada Klaster 0 dengan kotak paling rendah yaitu dibawah 100 hari sehingga mereka masih termasuk pelanggan terbaru yang aktif, sedangkan Klaster 1 dengan kotak berada di paling atas yang membuktikan bahwa kelompok pelanggan sudah lama tidak bertransaksi.

Distribusi *frequency* dengan ketiga klaster berada pada garis datar yang sama yaitu angka 1.0. Hal tersebut menunjukkan hasil dari proses penanganan *outlier* dengan metode IQR. Data dibersihkan dari pelanggan dengan pembelian yang terlampau ekstrem agar posisi pusat klaster (*centroid*) tidak bias sehingga *K-Means* dapat mengelompokkan pelanggan utama yaitu pelanggan *retail* secara spesifik.

Kemudian untuk Distribusi *monetary*, Klaster 0 dan 1 berada pada rentang nilai yang hampir sama, sementara Klaster 2 berada di bagian bawah karena merupakan kelompok yang memiliki transaksi terendah.

3. Scatter Plot (Pemisahan Wilayah Klaster)



Gambar 7. Scatter Plot Sebaran Pelanggan

Scatter Plot menunjukkan visualisasi algoritma *K-Means* mengelompokkan pelanggan dalam persebaran dimensi *recency* dan *monetary*, terlihat pada wilayah yang dipisahkan berdasar warna pada Gambar 7. Pada Titik Ungu (Klaster 0) menunjukkan pelanggan dengan nilai ekonomi cukup tinggi. Titik Hijau (Klaster 1) berisi pelanggan yang memiliki *Monetary* yang cukup

tinggi (nilai transaksi besar) namun sudah berisiko hilang, sementara Titik Kuning (Klaster 2) berisi pelanggan dengan frekuensi dan total nilai belanja rendah.

3.6. Analisis Strategi Pemasaran

Dari hasil analisis dan visualisasi yang sudah didapatkan, strategi pemasaran yang dapat disesuaikan oleh toko *retail XYZ* pada setiap segmen pelanggan untuk menghasilkan profit yang optimal dan meminimalisir risiko kehilangan pelanggan di antaranya, Klaster 0 (*new customers / potential*) dapat diberikan loyalitas program atau akses eksklusif pada produk terbaru yang akan *launching* untuk menjaga pelanggan agar frekuensi pembelian meningkat dan menjadi pelanggan loyal dan dapat memberikan penawaran penggunaan *membership* atau dapat melalui fitur *shopee chat broadcast* misalnya mengirimkan pesan, "Spesial untuk pelanggan setia : *Voucher Potongan Rp30.000*". Promo tersebut dapat dikirimkan bertepatan dengan *shopee campaign* bulanan (*double date* seperti 8.8, 9.9). Klaster 1 (*at risk customers*) dapat diberikan penawaran atau diskon khusus untuk mengundang minat belanja pelanggan dan mengaktifkan kembali pelanggan yang sudah lama tidak bertransaksi dengan memanfaatkan fitur *shopee video* untuk membagikan katalog produk terbaru, serta berikan *voucher* seperti "ikuti toko" atau *voucher "next purchase"* khusus agar mereka segera melakukan transaksi ulang, serta Klaster 2 (*lost customers*) dengan memberi promo *bundling* atau cuci gudang dengan menjual produk-produk *slow-moving* dengan harga miring guna menghabiskan stok gudang menggunakan fitur *shopee live* yang dilakukan secara berkala.

4. Kesimpulan

Model RFM dan algoritma *K-Means* berhasil menerapkan pengelompokkan pelanggan toko *retail XYZ*. Melalui pengujian metode *elbow*, *silhouette score* dan *davies-bouldin index*, di dapatkan tiga klaster optimal ($k = 3$). Hasil analisis pengelompokkan pelanggan yaitu klaster 0 (*new customer*) sebanyak 23% mencakup 9.902 pelanggan yang masih baru, aktif dan memiliki potensial belanja yang tinggi. Klaster 1 (*at risk customer*) dengan dominasi sebesar 72,1% berisi pelanggan dengan *monetary* tinggi namun sudah lama tidak aktif serta klaster 2 (*lost customer*) sebesar 4,9% berisi pelanggan dengan nilai transaksi yang cukup rendah.

Penelitian ini dapat menyimpulkan bahwa mayoritas pelanggan di dominasi oleh pembeli satu kali (*one-time buyers*) karena nilai tetap yang dihasilkan variabel *frequency* yaitu 1.0 pada setiap klaster. Kondisi tersebut berdampak pada model RFM yang diajukan menjadi model *Recency* dan *Monetary* (RM) saja dalam pemisahan segmen pelanggan.

Keterbatasan penelitian ini mencakup keterbatasan metodologis pada variabel *frequency*, di mana berisi pelanggan dengan dominasi ekstrem yaitu pembeli satu kali. Sehingga analisis pembentukan klaster pada algo-

ritma K-Means hanya mengandalkan variabel *recency* dan *monetary* (RM) sehingga segmentasi perilaku loyalitas pelanggan belum digali secara mendalam melalui model RFM utuh.

Berdasarkan keterbatasan yang ditemukan dalam penelitian ini, saran yang dapat dilakukan untuk penelitian berikutnya diantaranya yaitu perbaikan metodologi pembersihan *outlier* yang dapat dilakukan dengan pendekatan alternatif lain seperti *threshold* manual yang lebih longgar, metode penanganan outlier dengan *machine learning* seperti *isolation forest* atau binning kategorikal yang dilakukan terlebih dahulu pada variabel *frequency*. Kemudian dapat melakukan pengembangan model dan variabel dengan menambah variabel eksternal seperti kategori produk serta menguji penerapan metode klastering lainnya seperti *Fuzzy C Means* untuk perbandingan akurasi segmentasi.

Daftar Rujukan

- [1] N. M. Ulfa, A. Faris, and I. Himawan, 'The Effect of Celebrity Endorsers, Service Features, and Promotions for Shopee E-marketplace Purchase Interest', *Indonesian Vocational Research Journal*, vol. 2, no. 1, p. 31, Dec. 2022, Accessed: Jul. 08, 2026. [Online]. Available: <https://doi.org/10.30587/ivjr.v2i1.4780>
- [2] E. Purwanto and A. Andhini, 'Clustering Pelanggan E-Marketplace Shopee Berdasarkan Kategori Produk dan Pendapatan Menggunakan K-Means', *Seminar Nasional AMIKOM Surakarta*, vol. 3, pp. 208–219, Jan. 2025.
- [3] R. Mukhlis Juliyanto, F. Patria, H. Danggo Pamungkas, and F. Surya Nugraha, 'Penerapan Algoritma Clustering untuk Segmentasi Pelanggan Berdasarkan Perilaku Pembelian', *Seminar Nasional AMIKOM Surakarta*, vol. 2, pp. 117–127, Nov. 2024.
- [4] P. Indra Pangestu, T. Iman Hermanto, and D. Irmayanti, 'Analisis Segmentasi Pelanggan Berbasis Model Recency Frequency Dan Monetary (RFM) Menggunakan Algoritma K-Means', *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 11, no. 3, pp. 1486–1492, Oct. 2023, doi: 10.36040/jati.v7i3.7171.
- [5] E. N. Wakhidah, D. A. Sabara, and R. Sinatria, 'Transformasi Customer Relationship Management (CRM) Berbasis Artificial Intelligence untuk Meningkatkan Efektivitas Hubungan Pelanggan', *INVESTASI: Inovasi Jurnal Ekonomi dan Akuntansi*, vol. 4, no. 2, pp. 130–139, May 2026, doi: 10.59696/investasi.v4i2.251.
- [6] A. Jihan, W. Prihartono, and . F., 'Segmentasi Konsumen di Pasarmu.id Menggunakan K-Means Clustering Berdasarkan Model RFM', *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, pp. 780–788, Apr. 2025, doi: 10.23960/jitet.v13i2.6327.
- [7] J. Homepage, M. F. Fadhillah, A. Lovely, A. Suyoso, and I. Puspitasari, 'Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM)', *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 48–56, Nov. 2025, doi: 10.57152/malcom.v5v1.1491.
- [8] Mas'ud Ghania Al-Fathoni and Teti Sofia Yanti, 'Pemasaran Efektif melalui Optimasi RFM Analisis Berdasarkan K-Means Clustering', *Bandung Conference Series: Statistics*, vol. 5, no. 2, pp. 517–526, Aug. 2025, doi: 10.29313/bcss.v5i2.20213.
- [9] Y. michael Sihombing and E. S. Rahayu, 'Implementasi Fuzzy C-Means Clustering dalam Segmentasi Pelanggan dengan RFM Model', *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 4, pp. 2777–2786, Dec. 2025, doi: 10.31004/riggs.v4i4.4034.
- [10] F. Irhamna Rahman, S. Mujadilah, T. Wahyuni, and L. Anas, 'Prediksi Tingkat Kelulusan Menggunakan K-Means Pada Program Studi Informatika Unismuh Makassar', *Jurnal FASILKOM*, vol. 3, no. 13, pp. 504–510, Dec. 2023, Accessed: Jul. 08, 2026. [Online]. Available: <https://doi.org/10.37859/jf.v13i3.6061>
- [11] M. Arif Saifudin, H. Endah Wahanani, and A. Junaidi, 'Implementasi Algoritma Asosiasi Fp-Growth Dan Klasifikasi K-Means Terhadap Pola Pembelian Konsumen Di Marketplace Shopee', *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 1, pp. 764–771, Mar. 2024.
- [12] D. Tatang Warienta *et al.*, 'Optimasi K-Means Menggunakan Algoritma Firefly Untuk Segmentasi Pelanggan pada E-commerce', *Jurnal FASILKOM*, vol. 14, no. 3, pp. 755–785, Dec. 2024, Accessed: Jul. 08, 2026. [Online]. Available: <https://doi.org/10.37859/jf.v14i3.8287>
- [13] D. S. Nugraha, I. Thoib, N. Sururi, F. Bayu F, and B. P. Candra, 'Segmentasi Pelanggan Berbasis Analisis RFM Menggunakan Algoritma K-Means Clustering', *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, vol. 4, no. 2, pp. 1361–1369, Sep. 2025, doi: 10.62712/juktisi.v4i2.641.