SEGMENTASI PELANGGAN MENGGUNAKAN ANALISIS RFM DENGAN ALGORITMA K-MEANS CLUSTERING

1. PENDAHULUAN

Perkembangan dunia saat ini sudah memasuki era baru, yaitu era digital. Ketika pintu gerbang zaman digital sudah dibuka, bisnis pun bisa mencapai jangkauan lintas negara. Hal ini sekaligus membuat persaingan bisnis semakin maju dan ketat. Ada banyak cara atau strategi yang bisa digunakan, termasuk dengan menggunakan berbagai platform yang mendukung seperti kegiatan promosi secara online. Cukup dengan menggunakan perangkat komputer, laptop, gadget atau smartphone yang terhubung ke jaringan internet. Lalu, tinggal mengakses media sosial yang sangat berguna untuk mengenalkan dan mempromosikan, bahkan menjual produk dan layanan bisnis. Selain itu juga, bisa memanfaatkan website hingga menggunakan sistem promosi seperti endorsement dan influencer [1].

Recency adalah mengukur nilai pelanggan berdasarkan pembelian paling akhir yang dilakukan pelanggan. Data terpenting yang diperlukan untuk menghitung nilai recency adalah tanggal pembelian terakhir. Nilai recency berkaitan dengan jarak antara tanggal terakhir transaksi dengan periode analisis, sehingga semakin dekat proses pembelian terakhir maka pelanggan tersebut semakin loyal. Frequency adalah mengukur nilai pelanggan berdasarkan seberapa sering pelanggan tersebut melakukan transaksi. Semakin sering melakukan transaksi ini memungkinkan bahwa pelanggan tersebut merupakan pelanggan potensial. Monetary adalah mengukur nilai pelanggan berdasarkan jumlah besaran transaksi yang dikeluarkan pelanggan dalam satu periode. Semakin banyak jumlah besaran uang yang dikeluarkan pelanggan maka nilai M semakin besar [2].

Analisis RFM memberikan wawasan strategis yang berharga untuk meningkatkan efektivitas pemasaran dan retensi pedagang pada transaksi gateway pembayaran pedagang. Top Merchants, dengan kinerja tertinggi di seluruh dimensi RFM, menjadi fokus strategi pemasaran untuk meningkatkan nilai transaksi dan retensi pelanggan. Rekomendasi yang diusulkan antara lain meluncurkan program loyalitas eksklusif, personalisasi konten pemasaran, informasi terkini, acara eksklusif, strategi kolaborasi dengan influencer, evaluasi pelanggan berkelanjutan, dan mengidentifikasi peluang cross-selling dan up-selling. Strategi ini diharapkan dapat mempererat hubungan dengan pelanggan Top Merchant, memaksimalkan nilai transaksi, dan mencapai retensi pelanggan yang tinggi [3].

Dalam sebuah penelitian ditekankan bahwa RFM analysis memberikan perspektif yang komprehensif tentang customer lifetime value karena mencakup aspek behavioral dan transaksional pelanggan. Hal ini didukung oleh penelitian Khajvand et al. (2011) yang menunjukkan bahwa segmentasi berbasis RFM dapat meningkatkan ROI kampanye pemasaran hingga 20-30% dibandingkan dengan pendekatan *mass marketing* [4].

Algoritma K-Means dipilih karena memiliki hasil clustering yang lebih baik dibandingkan metode lainnya. Jumlah segmen optimum didapatkan dengan menggunakan metode Elbow dan Silhouette Coefficient. MacQueen (1967) memperkenalkan algoritma K-Means sebagai metode *partitioning clustering* yang mengoptimalkan jarak dalam

cluster dan memaksimalkan jarak antar cluster. Keunggulan K-Means terletak pada kemudahan implementasi, skalabilitas untuk dataset besar, dan interpretabilitas hasil yang tinggi [5]

Penelitian ini menggunakan algoritma K-Means untuk menjalankan clustering yang performanya akan dibandingkan dengan algoritma K-Medoids mengacu pada nilai silhouette, Calinski-Harabasz Index, dan DaviesBouldin Index dalam melakukan segmentasi pelanggan berdasarkan atribut RFM. Namun, performa algoritma K-Means sangat dipengaruhi oleh *preprocessing data,* khususnya *feature scaling*. Ditekankan pentingnya normalisasi data dalam clustering karena perbedaan skala antar variabel dapat menyebabkan bias dalam perhitungan jarak. Oleh karena itu, penelitian ini akan mengeksplorasi berbagai teknik *feature scaling* untuk mendapatkan hasil clustering yang optimal [6]

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengimplementasikan segmentasi pelanggan menggunakan analisis RFM dengan algoritma *K-Means clustering*. Penelitian ini akan membandingkan efektivitas berbagai teknik *feature scaling* dan memberikan rekomendasi segmentasi pelanggan yang dapat diaplikasikan dalam strategi pemasaran.

2. METODE PENELITIAN

2.1 Dataset dan Deskripsi Data

Dataset yang digunakan dalam penelitian ini adalah Olist Dataset yang terdiri dari 4 dataset yang saling berkaitan satu dengan yang lain.4 Dataset itu antara lain sebagai berikut: Olist Order.csv, Olist Order Items.csv, Olist Customer.csv, Olist Geolocation.csv.

Dataset terbut diperoleh dari https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Dataset ini berisi informasi transaksi pelanggan yang mencakup customer ID, tanggal transaksi, dan nilai transaksi. Dataset terdiri dari jumlah record di bawah ini dengan perkiraan 96.096 pelanggan unik dalam periode 4 Oktober 2016 sampai 16 Oktober 2018.

Tabel 1. Tabel banyak baris dari dataset

Nama Dataset	*Jumlah Record*
Customers	99441
Geo Location	1000163
Orders	99441
Orders Items	112650

Struktur data utama meliputi:

1. Dataset Pelanggan (Customers)

Kolom di sini berfokus pada siapa dan di mana pelanggan berada.

- customer_id: Kunci penghubung teknis ke dataset pesanan. Setiap pesanan baru akan memiliki customer_idyang berbeda, meskipun pelanggannya sama.

- customer_unique_id: Ini adalah kolom paling penting untuk analisis pelanggan. Kolom ini mengidentifikasi satu pelanggan secara unik di semua pesanannya. Sangat penting untuk menghitung jumlah pelanggan unik, loyalitas, dan Customer Lifetime Value (CLV).
- customer_zip_code_prefix dan customer_state: Penting untuk analisis geografis. Berguna untuk memetakan distribusi pelanggan dan memahami pasar regional.

2. Dataset Pesanan (Orders)

Ini adalah pusat dari semua data transaksi.

- order_id: Kunci utama (Primary Key) yang unik untuk setiap pesanan. Kolom ini adalah penghubung utama ke dataset lain seperti order_items.
- customer_id: Kunci penghubung (Foreign Key) ke dataset pelanggan untuk mengetahui siapa yang melakukan pesanan.
- order_purchase_timestamp: Kolom waktu terpenting. Menandai kapan pesanan dibuat, menjadi dasar untuk semua analisis berbasis waktu seperti tren penjualan harian, bulanan, atau tahunan.
- order_status: Sangat penting untuk memfilter data. Analisis sering kali hanya fokus pada pesanan yang berstatus delivered (terkirim) dan mengabaikan yang canceled (dibatalkan).

3. Dataset Item Pesanan (Order Items)

Kolom di sini merinci isi dari setiap pesanan.

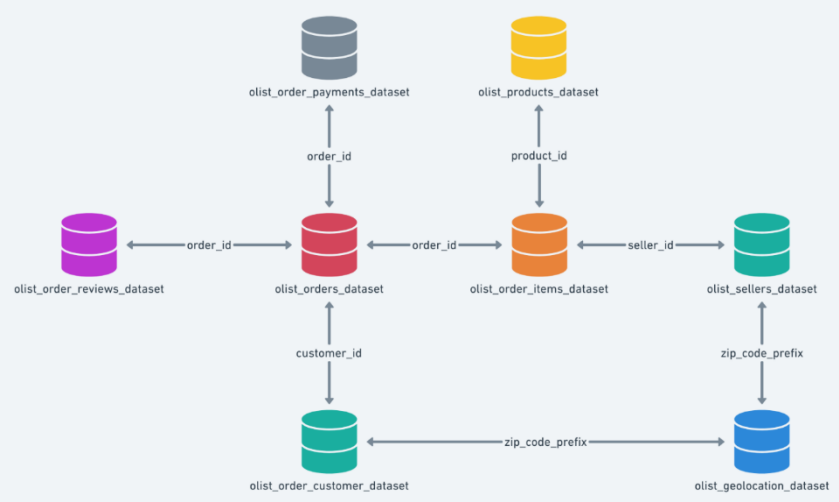
- order_id: Kunci penghubung ke dataset pesanan, memastikan setiap item terhubung ke transaksi yang benar.
- product_id dan seller_id: Penghubung ke data produk dan penjual. Sangat penting untuk menganalisis produk mana yang paling laku dan penjual mana yang memiliki performa terbaik.
- price dan freight_value: Kolom finansial inti. price adalah harga produk dan freight_value adalah biaya pengiriman. Menjumlahkan kolom ini sangat penting untuk menghitung total pendapatan (omzet) dan profitabilitas.

4. Dataset Geolokasi (Geolocation)

Dataset ini berfungsi sebagai kamus untuk menerjemahkan kode pos menjadi koordinat geografis.

- geolocation_zip_code_prefix: Kunci penghubung ke dataset pelanggan. Ini memungkinkan Anda untuk mengaitkan lokasi pelanggan dengan data lintang (latitude) dan bujur (longitude).
- geolocation_lat dan geolocation_lng: Koordinat geografis. Sangat berdampak untuk analisis spasial seperti visualisasi peta, menghitung jarak antara pelanggan dan penjual, serta optimalisasi logistik.

Struktur dari data yang kami miliki:



2.2 Analisis RFM (Recency, Frequency, Monetary)

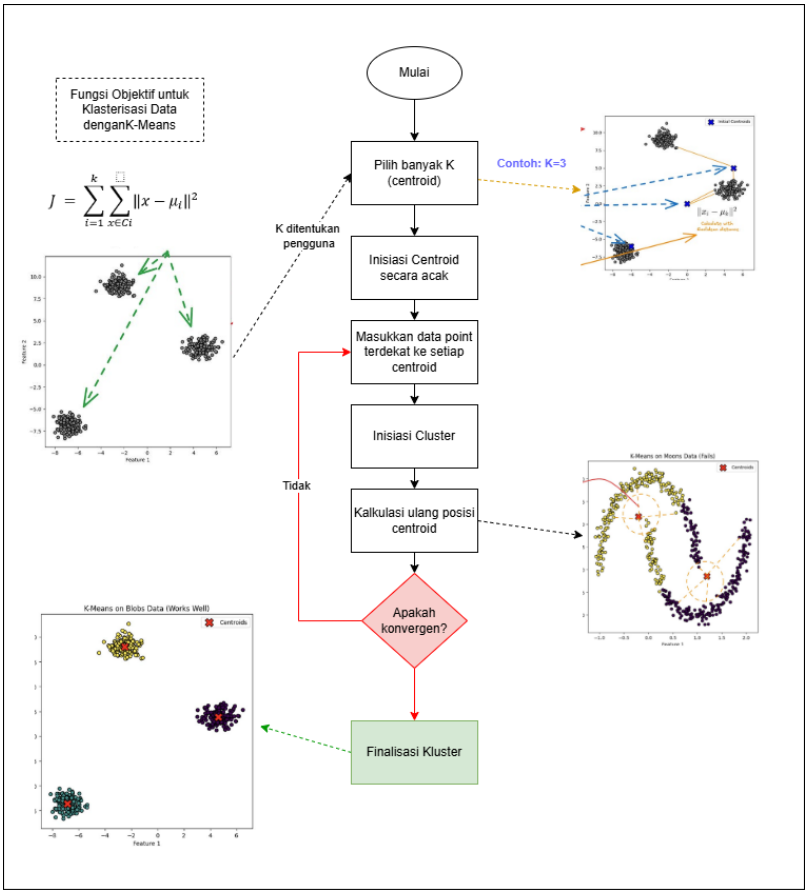
RFM analysis merupakan teknik segmentasi pelanggan yang mengukur tiga dimensi perilaku pelanggan:

- Recency (R): mengukur kebaruan transaksi terakhir pelanggan, dihitung sebagai selisih hari antara tanggal transaksi terakhir dengan tanggal referensi (biasanya tanggal analisis dilakukan). Nilai recency yang lebih kecil menunjukkan pelanggan yang lebih aktif.
- Frequency (F): mengukur frekuensi transaksi pelanggan dalam periode tertentu, dihitung sebagai total jumlah transaksi yang dilakukan pelanggan. Nilai frequency yang lebih tinggi menunjukkan tingkat loyalitas yang lebih baik.
- Monetary (M): mengukur nilai total yang dihabiskan pelanggan, dihitung sebagai jumlah seluruh nilai transaksi pelanggan. Nilai monetary yang lebih tinggi menunjukkan kontribusi ekonomi yang lebih besar.

  Formulasi matematis untuk setiap komponen RFM:

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

2.3 Algoritma K-Means Clustering



K-Means clustering adalah algoritma partitioning yang membagi data ke dalam k cluster berdasarkan kemiripan. Algoritma ini bekerja dengan meminimalkan *Within-Cluster Sum of Squares* (WCSS) melalui iterasi optimasi posisi *centroid*.

Langkah-langkah algoritma K-Means:

1. Inisialisasi: Pilih k centroid awal secara acak
2. Assignment: Masukkan setiap *data point* ke *centroid* terdekat berdasarkan jarak *Euclidean*
3. Update: Hitung ulang posisi *centroid* sebagai rata-rata semua *data point* dalam cluster
4. Konvergensi: Ulangi langkah 2-3 hingga posisi centroid tidak berubah atau mencapai iterasi maksimum

Fungsi objektif K-Means:

$$J = \sum_{i=1}^{k} \sum_{x \in Ci} \|x - \mu_i\|^2$$

Dengan keterangan:

- J = fungsi objektif yang diminimalkan

- k = jumlah cluster
- Ci = cluster ke-i
- x = data point
- μ_i = centroid cluster ke-i
- $\|x - \mu_i\|^2$: Jarak Euclidean distance kuadrat antara titik data x dan centroid cluster-nya

2.4 Perhitungan Jarak Euclidean

Jarak Euclidean digunakan untuk mengukur similarity antar data point dalam ruang multidimensi. Untuk data RFM tiga dimensi, jarak Euclidean dihitung sebagai:

$$d(p,q) = \sqrt{[(p_1 - q_1)^2 + (p_2 + q_2)^2 + (p_3 + q_3)^2]}$$

Dimana p dan q adalah dua data point dengan koordinat (p_1,p_2,p_3) dan (q_1,q_2,q_3) yang merepresentasikan nilai R, F, dan M.

2.5 Teknik Feature Scaling

Karena ketiga komponen RFM memiliki skala dan satuan yang berbeda, feature scaling menjadi krusial untuk memastikan tidak ada variabel yang mendominasi perhitungan jarak. Penelitian ini mengeksplorasi empat pendekatan:

2.5.1 Standardisasi (Z-Score Normalization)

Standardisasi mentransformasi data sehingga memiliki mean = 0 dan standard deviation = 1:

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

Dimana μ adalah mean dan σ adalah standard deviation dari variabel.

2.5.2 Min-Max Normalization

Normalisasi min-max mentransformasi data ke rentang [0,1]:

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

2.5.3 Log Transformation

Transformasi logaritma mengurangi skewness dan menstabilkan variance:

$$x_{log} = \log(x + 1)$$

Penambahan konstanta 1 untuk menghindari log(0).

2.6 Penentuan Jumlah Cluster Optimal

Penentuan jumlah cluster dilakukan dengan menggunakan silhuoette score tertinggi

2.7 Metrik Evaluasi Clustering

2.7.1 Silhouette Score

Silhouette Score mengukur seberapa mirip suatu objek dengan cluster-nya sendiri dibandingkan dengan cluster lain:

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

Dimana:

- a(i) = rata-rata jarak dari data point i ke semua data point lain dalam cluster yang sama
- b(i) = rata-rata jarak dari data point i ke semua data point dalam cluster terdekat

Nilai berkisar antara -1 hingga 1, dimana nilai yang lebih tinggi menunjukkan clustering yang lebih baik.

2.7.2 Dunn Index

Dunn Index mengukur rasio antara jarak minimum antar cluster dengan diameter maksimum cluster:

$$\text{DI} = \frac{\min\big(\delta(Ci, Cj)\big)}{\max\big(\Delta(Ck)\big)}$$

Dimana δ(Ci,Cj) adalah jarak antar cluster dan Δ(Ck) adalah diameter cluster. Nilai yang lebih tinggi menunjukkan clustering yang lebih baik.

2.7.3 Calinski-Harabasz Index

Calinski-Harabasz Index mengukur rasio antara dispersi antar-klaster dan dispersi dalam-klaster:

$$\text{CH} = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{n-k}\right)}$$

Dimana:

- SSB = Jumlah Kuadrat Antar cluster
- SSW = Jumlah Kuadrat dalam kelompok
- k = jumlah cluster
- n = jumlah data point

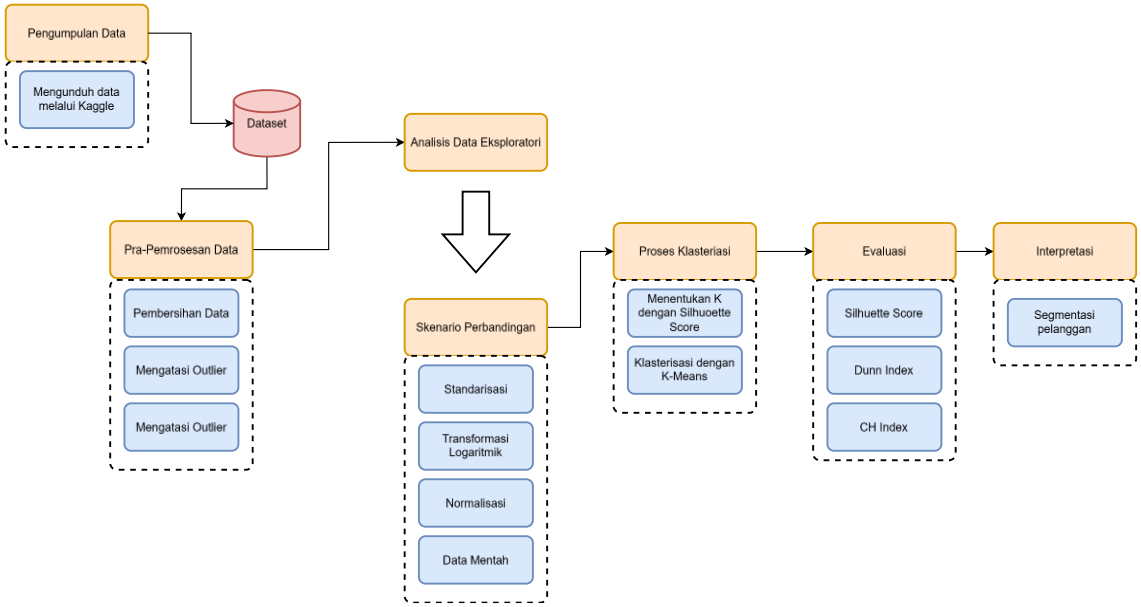
Nilai yang lebih tinggi menunjukkan clustering yang lebih baik.

2.8 Eksperimen dan Skenario

Penelitian ini melakukan empat skenario eksperimen:

1. Skenario 1: K-Means dengan Standardisasi
2. Skenario 2: K-Means dengan Min-Max Normalization
3. Skenario 3: K-Means dengan Log Transformation
4. Skenario 4: K-Means dengan Raw Data (tanpa feature scaling)

Setiap skenario akan dievaluasi menggunakan keempat metrik evaluasi dan dibandingkan performanya untuk menentukan pendekatan terbaik.



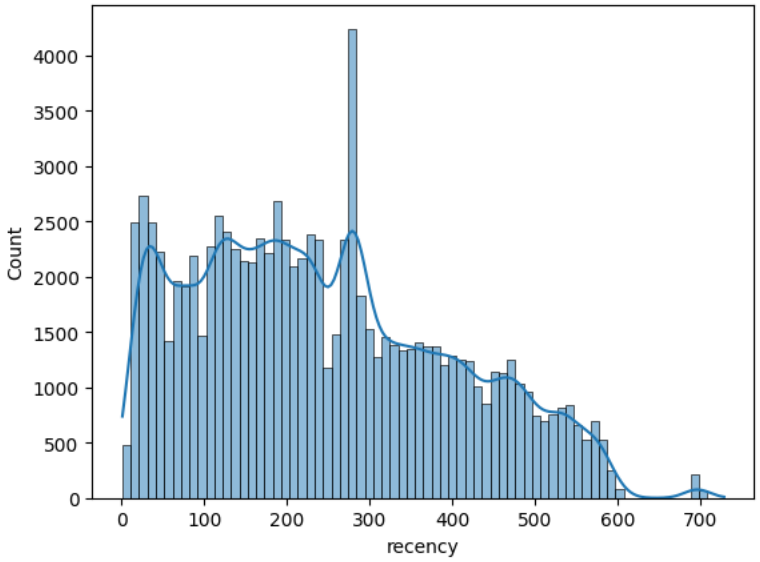
Gambar 1. Flowchart metodologi penelitian

3. METODE PENELITIAN

3.1 Hasil Preprocessing Data

3.1.1 Eksplorasi Data Awal

Dataset yang seperti digunakan dalam penelitian ini terdiri dari [jumlah] record transaksi dari [jumlah] pelanggan unik. Analisis deskriptif menunjukkan distribusi data sebagai berikut:



Gambar 2. Distribusi data recency

Statistik deskriptif dataset:

- Total transaksi: 94488
- Periode data: Recency
- Rata-rata nilai transaksi: 243.85
- Standar deviasi nilai transaksi: 153.16\

3.1.2 Pembentukan Variabel RFM

Setelah melakukan *preprocessing*, diperoleh dataset RFM dengan karakteristik:

Recency:

- Mean: 239 hari
- Median: 222 hari
- Std: 153 hari
- Range: 0 - 695 hari

Frequency:

- Mean: 1 transaksi
- Median: 1 transaksi
- Std: 0.18 transaksi
- Range: 1 - 15 transaksi

Monetary:

- Mean: 160.99 USD
- Median: 105.00 USD
- Std: 219.73 USD
- Range: 0.85 - 13664.08 USD

[Tampilkan Gambar 3: Distribusi variabel RFM]

3.1.3 Deteksi dan Penanganan Outlier

Outlier pada data numerik dianalisis menggunakan pendekatan IQR (Interquartile Range), dengan langkah-langkah sebagai berikut:

1. Menghitung kuartil bawah (Q1) dan kuartil atas (Q3) untuk setiap fitur numerik.
2. Menghitung rentang antar kuartil:

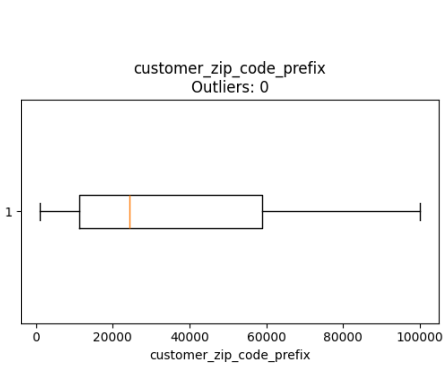
$$IQR = Q3 - Q1$$

3. Menentukan batas bawah dan atas untuk deteksi outlier:

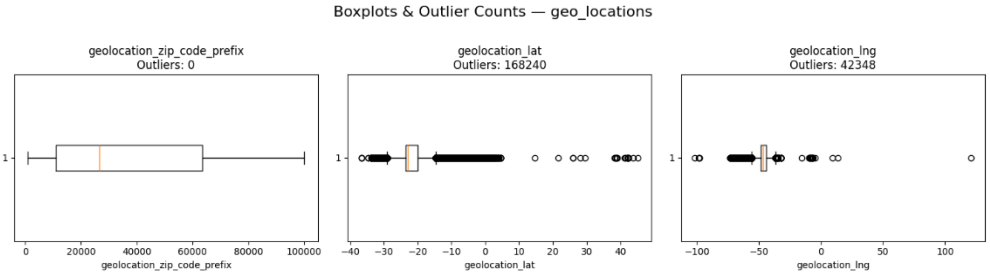
$$Lower\ Bound = Q1 - 1.5 \times IQR$$

$$Upper\ Bound = Q3 + 1.5 \times IQR$$

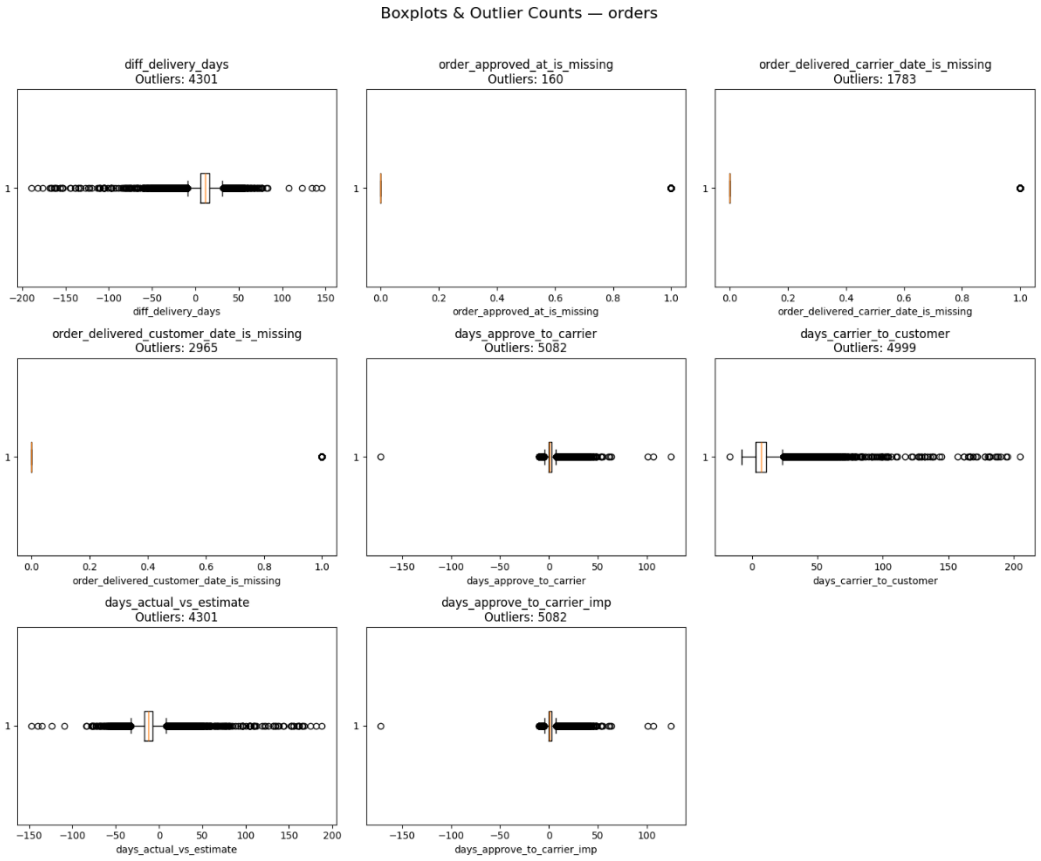
4. Observasi yang berada di luar rentang tersebut dikategorikan sebagai **outlier**.



Gambar 3. Outlier Costumer Zip Code



Gambar 2. Outlier Geo Locations



Gambar 3. Outlier Orders

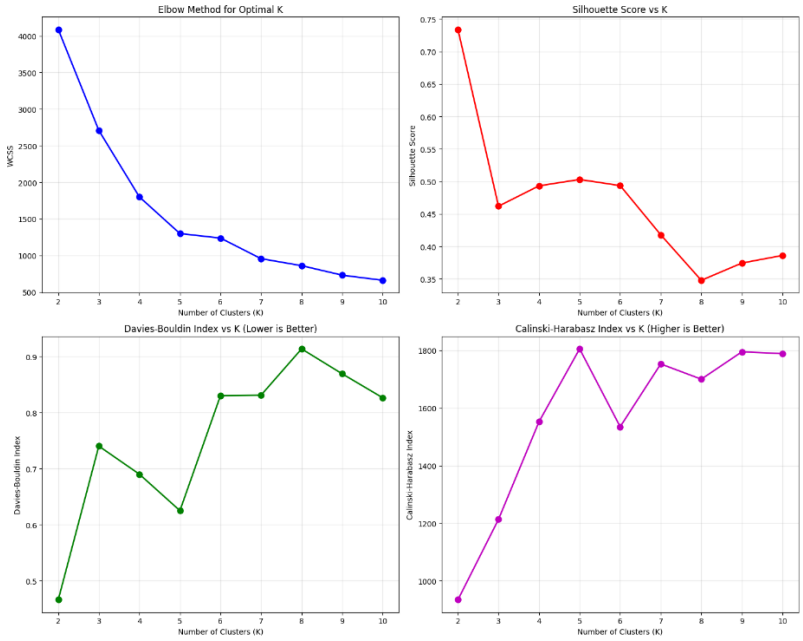
Gambar 4. Outlier Order Items

3.2 Hasil Preprocessing Data

Hasil Clustering per Skenario

3.3.1 Skenario 1: Standardisasi

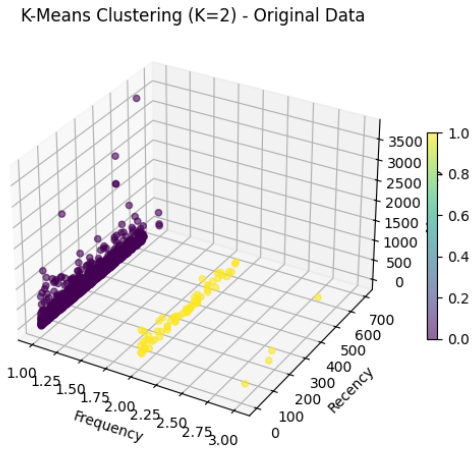
- Jumlah cluster optimal



Gambar 4. Nilai Kluster optimal untuk setiap metrik

Mengambil kluster optimal dari metrik Silhuoette Score, yaitu K = 2.

- Silhouette Score: 0.7341
- Dunn Index: 0.4663
- Calinski-Harabasz Index: 934.3349
- WCSS: 4088.2097

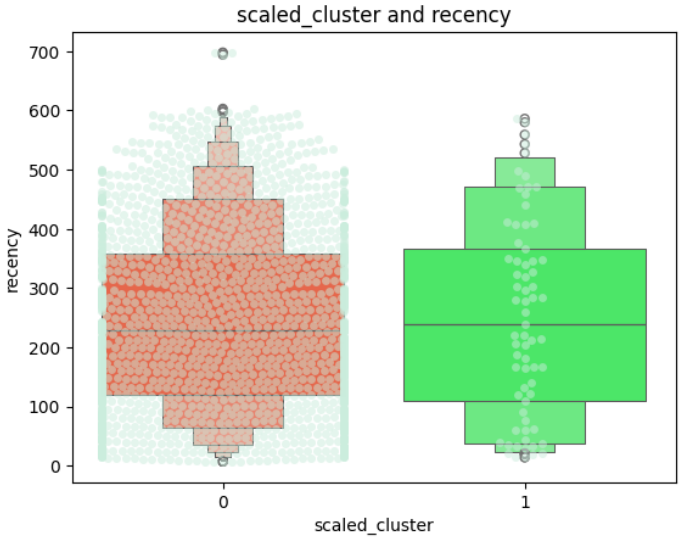


Gambar 5. Visualisasi hasil clustering skenario standarisasi

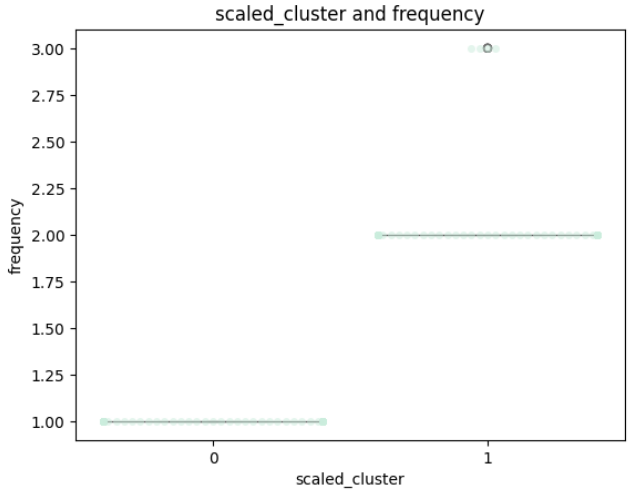
Distribusi data per cluster

Cluster 0	: 1931 data points (96.5%)
Cluster 1	: 69 data points (3.5%)

Karakteristik setiap cluster:



Gambar 6. Boxplot bagaimana data recency tersebar di Skenario 1 (standarisasi)



Gambar 7. Boxplot bagaimana data frequncy tersebar di Skenario 1 (standarisasi)

Gambar 8. Boxplot bagaimana data monetary tersebar di Skenario 1 (standarisasi)

- Cluster 0: *High Value, Low Engagement*
  Karakteristik:

- Recency (R): Setara dengan cluster lain, menunjukkan bahwa pelanggan di cluster ini melakukan transaksi terakhir dalam waktu yang hampir sama dengan pelanggan di cluster lainnya.
- Frequency (F): Rendah. Pelanggan dalam cluster ini jarang melakukan pembelian berulang.
- Monetary (M): Tinggi. Meskipun jarang bertransaksi, nilai transaksi yang dilakukan pelanggan di cluster ini cukup besar.

Interpretasi:

Cluster ini dapat diidentifikasi sebagai pelanggan yang berbelanja dengan nominal besar, namun tidak sering bertransaksi. Mereka mungkin hanya melakukan pembelian pada momen tertentu atau bersifat seasonal, tetapi saat mereka membeli, mereka mengeluarkan uang dalam jumlah besar. Mereka adalah pelanggan bernilai tinggi dari segi kontribusi revenue, tetapi dengan keterlibatan yang rendah.

- Cluster 1: *Engaged but Low Value*

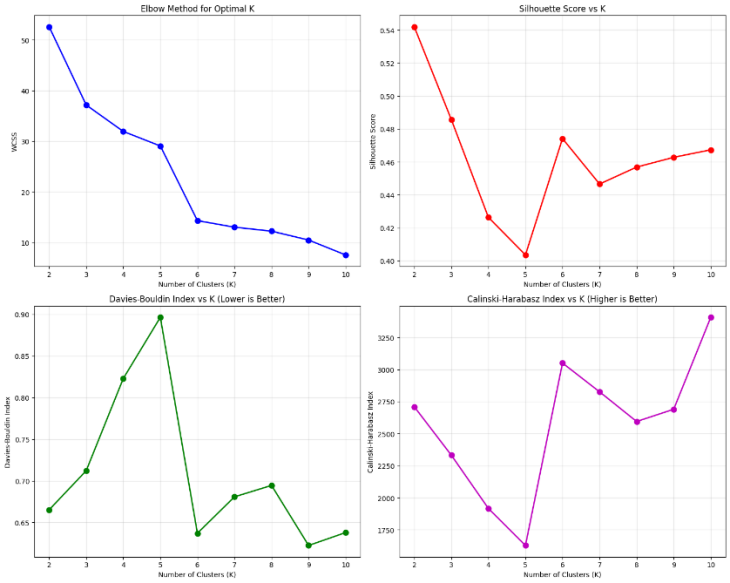
  Karakteristik:
  - Recency: Setara dengan cluster lain, menunjukkan waktu transaksi terakhir tidak berbeda signifikan.
  - Frequency (F): Tinggi (berada pada cluster ke-2 atau ke-3 dalam skala frekuensi).
  - Monetary (M): Lebih rendah dibandingkan dengan Cluster 0.

  Interpretasi Kasar:

  Cluster ini berisi pelanggan yang sering melakukan transaksi, namun dengan nilai pembelian yang relatif kecil. Mereka menunjukkan loyalitas atau engagement yang tinggi, namun dari sisi revenue, kontribusinya masih lebih rendah dibandingkan Cluster 0. Segmen ini cocok untuk strategi upselling atau cross-selling karena mereka sudah aktif berinteraksi.

3.3.2 Skenario 2: *Min-Max Normalization*

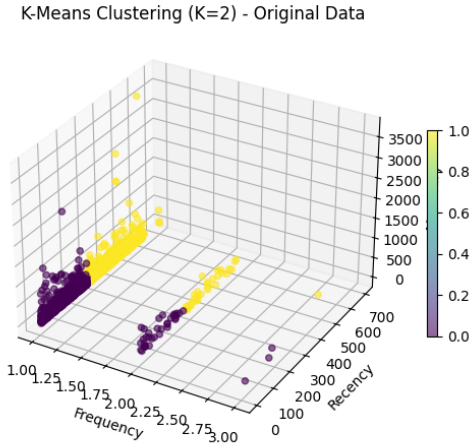
- Jumlah *cluster* optimal



Gambar 9. Nilai Kluster optimal untuk setiap metrik

Mengambil kluster optimal dari metrik Silhuoette Score, yaitu K = 2.

- Silhouette Score: 0.5420
- Dunn Index: 0.6652
- Calinski-Harabasz Index: 2709.5608
- WCSS: 52.5686

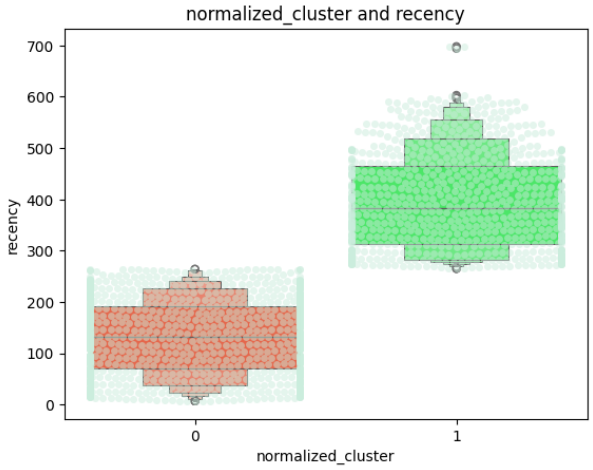


Gambar 10. Visualisasi hasil clustering skenario normalisasi

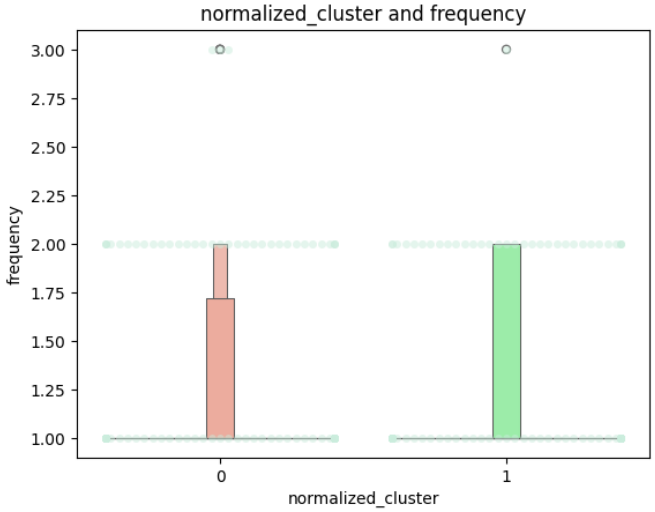
Distribusi data per cluster

Cluster 0	: 1130 data points (56.5%)
Cluster 1	: 870 data points (43.5%)

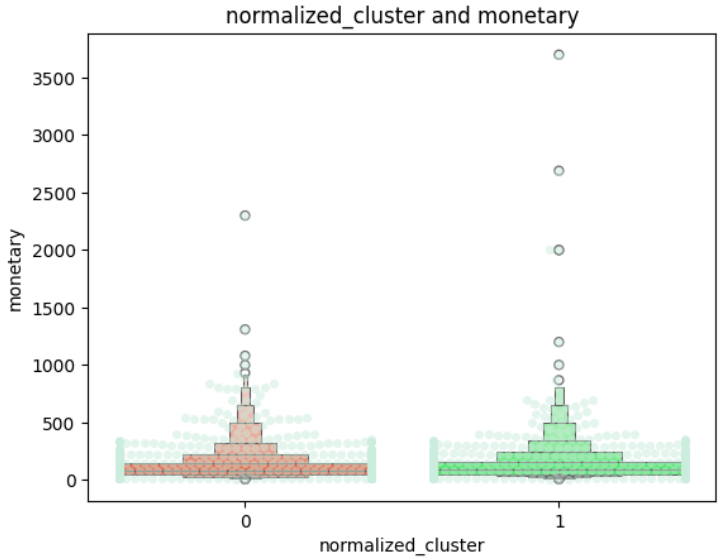
Karakteristik setiap cluster:



Gambar 11. Boxplot bagaimana data recency tersebar di Skenario 2 (normalisasi)



Gambar 12. Boxplot bagaimana data frequncy tersebar di Skenario 2 (normalisasi)



Gambar 13. Boxplot bagaimana data monetary tersebar di Skenario 2 (normalisasi)

- Cluster 0: *High Value, Low Engagement*

Karakteristik: Pelanggan Aktif

- Recency: Rendah. Artinya, pelanggan dalam cluster ini baru-baru ini melakukan pembelian, sehingga menunjukkan tingkat keterlibatan yang masih tinggi dengan bisnis.
- Frequency: Setara dengan cluster lain. Pelanggan dalam cluster ini memiliki jumlah transaksi yang mirip dibandingkan dengan cluster lainnya, sehingga tidak menunjukkan kecenderungan lebih sering ataupun lebih jarang dalam berbelanja.
- Monetary: Setara dengan cluster lain. Nilai total pembelian dari pelanggan ini berada pada tingkat yang sebanding, tanpa dominasi yang signifikan terhadap kontribusi pendapatan secara keseluruhan.

Interpretasi:

Pelanggan pada cluster ini masih aktif dan menunjukkan potensi tinggi untuk mempertahankan hubungan jangka panjang. Karena keterlibatan mereka masih tinggi secara waktu, strategi yang dapat diterapkan adalah program loyalitas, rekomendasi produk berbasis histori pembelian, atau pemberian insentif untuk meningkatkan frekuensi transaksi lebih lanjut.

- Cluster 1: Pelanggan Tidak Aktif

  Karakteristik:

  - Recency: Tinggi. Artinya, pelanggan dalam cluster ini sudah cukup lama tidak melakukan pembelian, yang dapat mengindikasikan risiko penurunan minat atau potensi *churn*.
  - Frequency: Setara dengan cluster lain. Secara historis, pelanggan dalam cluster ini memiliki frekuensi pembelian yang setara dengan pelanggan aktif, yang berarti mereka pernah cukup *engaged*.
  - Monetary: Setara dengan cluster lain. Meskipun saat ini tidak aktif, pelanggan ini pernah memberikan kontribusi pendapatan yang sama besar dengan pelanggan lain.

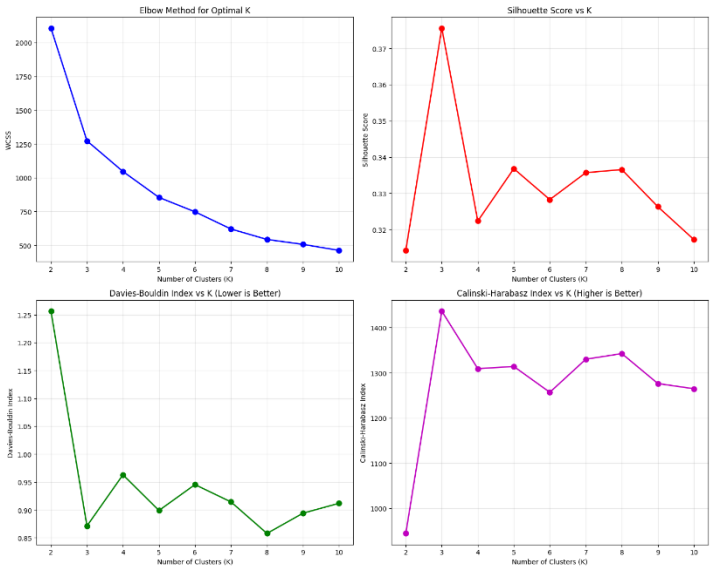
  Interpretasi:

  Pelanggan dalam cluster ini termasuk dalam kategori tidak aktif, namun dengan riwayat interaksi yang baik. Oleh karena itu, mereka merupakan target ideal untuk strategi *re-engagement*, seperti penawaran personal, diskon eksklusif untuk comeback, atau kampanye *"kami merindukan Anda"* guna membangkitkan kembali loyalitas mereka.

3.3.3 Skenario 3: Log Transformation

Hasil clustering dengan log transformation:

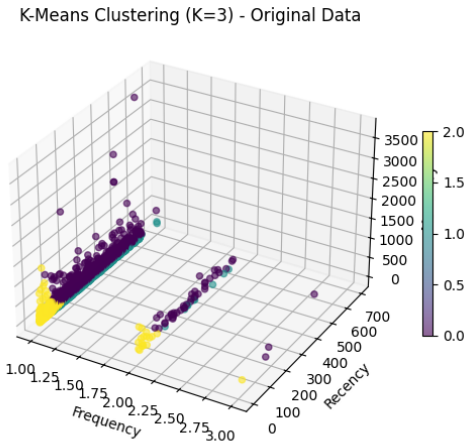
- Jumlah *cluster* optimal



Gambar 14. Nilai Kluster optimal untuk setiap metrik untuk Skenario 3

Mengambil kluster optimal dari metrik Silhuoette Score, yaitu K = 3.

- Silhouette Score: 0.3756
- Dunn Index: 0.8710
- Calinski-Harabasz Index: 1436.3006
- WCSS: 1272.5558

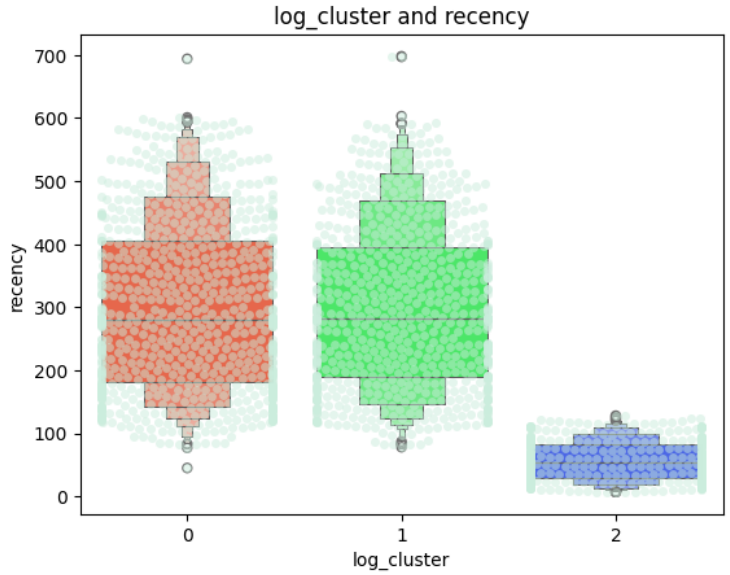


Gambar 15. Visualisasi hasil clustering skenario log

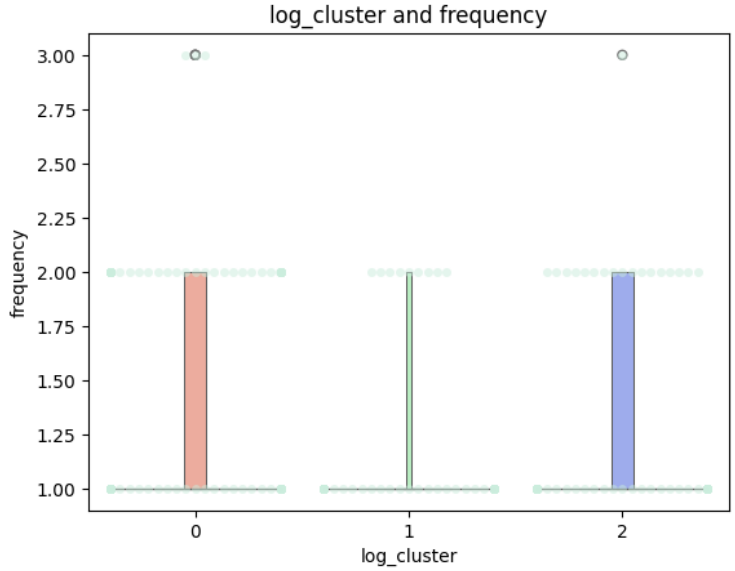
Distribusi data per cluster

Cluster 0	: 823 data points (41.1%)
Cluster 1	: 749 data points (37.5%)
Cluster 2	: 428 data points (21.4%)

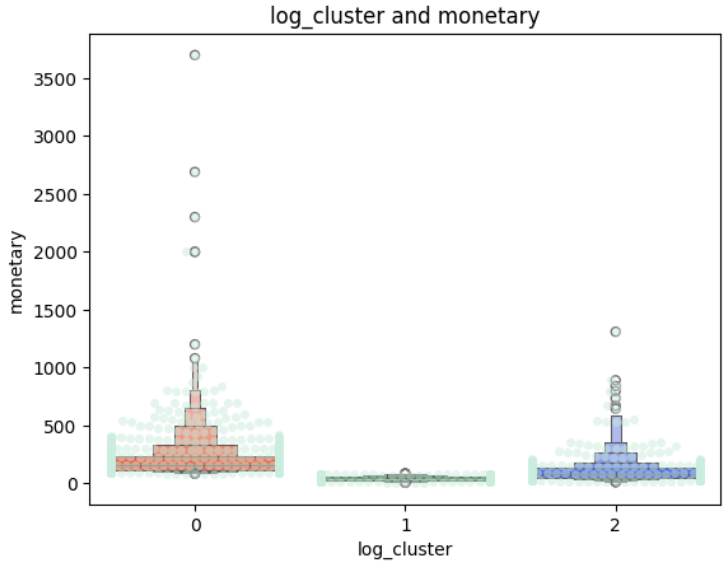
Karakteristik setiap cluster:



Gambar 16. Boxplot bagaimana data recency tersebar di Skenario 3 (log)



Gambar 17. Boxplot bagaimana data frequncy tersebar di Skenario 3 (log)



Gambar 18. Boxplot bagaimana data monetary tersebar di Skenario 3 (transformasi log)

- Cluster 0: Pelanggan Potensial Bernilai Tinggi namun Tidak Aktif
  Karakteristik: Pelanggan Aktif
  - Recency: Tinggi. Pelanggan dalam cluster ini sudah cukup lama tidak melakukan transaksi, yang menandakan tingkat keterlibatan yang rendah saat ini.
  - Frequency: Setara. Jumlah transaksi yang dilakukan serupa dengan cluster lain.
  - Monetary: Tertinggi di antara semua cluster. Pelanggan dalam kelompok ini pernah menghasilkan nilai pembelian yang tinggi, meskipun aktivitasnya kini menurun.

  Interpretasi:
  Cluster ini berisi pelanggan yang memiliki riwayat pembelian bernilai tinggi, namun saat ini tidak lagi aktif. Mereka memiliki potensi besar untuk menghasilkan revenue apabila berhasil diaktivasi kembali. Oleh karena itu, strategi yang disarankan adalah *re-engagement* dengan penekanan pada nilai premium, seperti promo eksklusif, produk edisi terbatas, atau program pelanggan VIP.
- Cluster 1: Pelanggan Tidak Aktif
  Karakteristik:
  - *Recency*: Tinggi. Pelanggan dalam cluster ini lama tidak bertransaksi, menunjukkan bahwa mereka tidak aktif dalam periode waktu terakhir.
  - *Frequency*: Setara. Jumlah transaksi historis tidak berbeda signifikan dibandingkan cluster lain.
  - *Monetary*: Terendah dari seluruh cluster. Kontribusi total pembelian pelanggan dalam cluster ini paling kecil.

  Interpretasi:

Cluster ini merupakan segmen berisiko rendah tetapi tidak prioritas, karena baik dari sisi nilai maupun keterlibatan sangat minim. Pelanggan dalam kelompok ini tidak memberikan kontribusi besar terhadap pendapatan dan memiliki kecenderungan untuk churn. Strategi promosi pada segmen ini perlu diperhitungkan secara efisien, seperti menggunakan kampanye massal atau reminder otomatis berbiaya rendah, tanpa alokasi sumber daya berlebih.

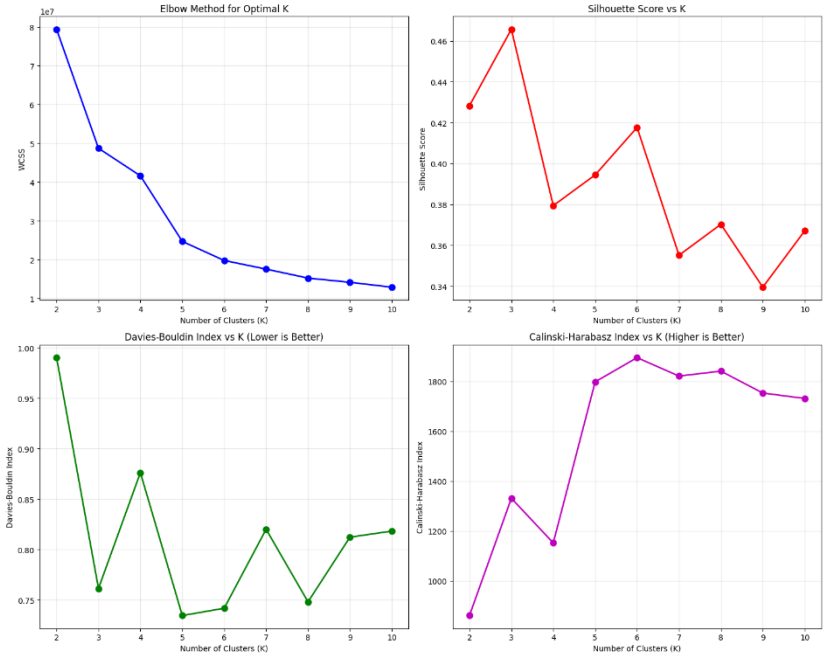
- Cluster 2: Pelanggan Aktif dengan Nilai Menengah
  Karakteristik:
  - *Recency*: Rendah. Pelanggan dalam cluster ini baru-baru ini melakukan transaksi, yang menandakan bahwa mereka masih aktif dan engaged.
  - *Frequency*: Setara. Sama dengan dua cluster lainnya.
  - *Monetary*: Menengah. Total nilai pembelian pelanggan di cluster ini berada di antara Cluster 0 dan Cluster 1.

  Interpretasi:
  Cluster ini merupakan segmen pelanggan aktif dengan nilai pembelian moderat. Mereka menunjukkan keterlibatan terkini dan memiliki potensi untuk ditingkatkan nilainya. Oleh karena itu, pendekatan yang disarankan adalah strategi upselling dan loyalty program, dengan tujuan meningkatkan nilai transaksi sambil mempertahankan frekuensi dan keterlibatan yang sudah baik.

3.3.4 Skenario 4: Raw Data

Hasil clustering dengan raw data:

- Jumlah *cluster* optimal

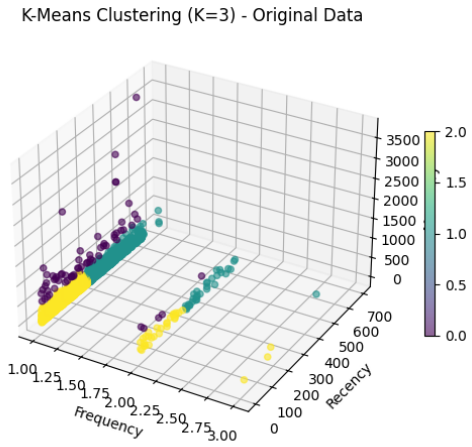


Gambar 19. Nilai Kluster optimal untuk setiap metrik untuk Skenario 4

Mengambil kluster optimal dari metrik Silhuoette Score, yaitu K = 3.

- Silhouette Score: 0.4657
- Dunn Index: 0.8710

- Calinski-Harabasz Index: 1436.3006
- WCSS: 48710032.1096

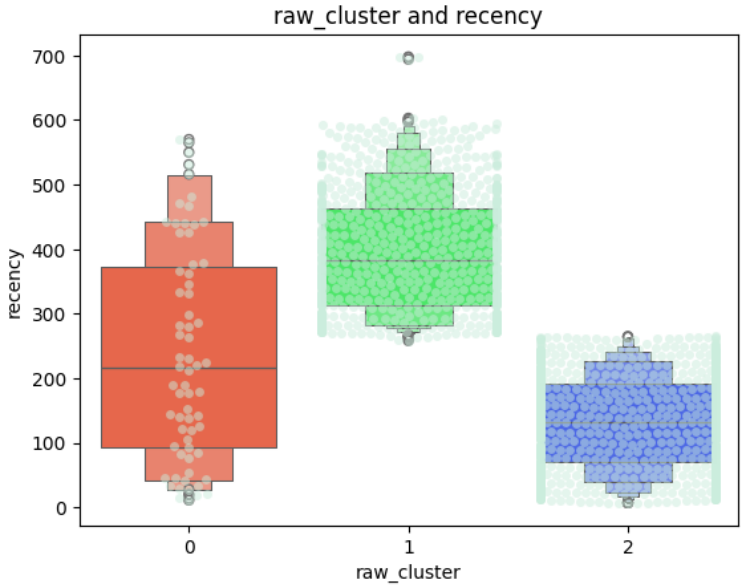


Gambar 20. Visualisasi hasil clustering skenario raw

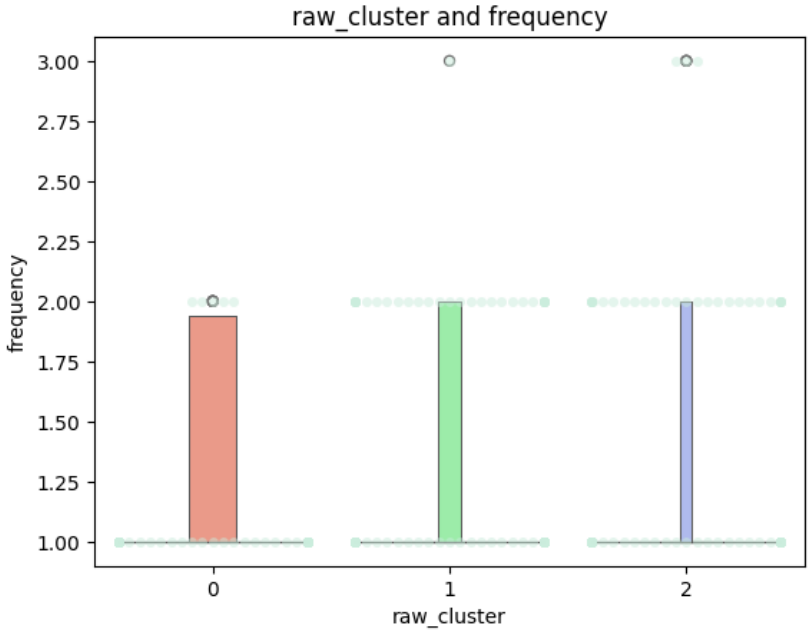
Distribusi data per cluster

Cluster 0	: 66 data points (3.3%)
Cluster 1	: 846 data points (42.3%)
Cluster 2	: 1088 data points (54.4%)

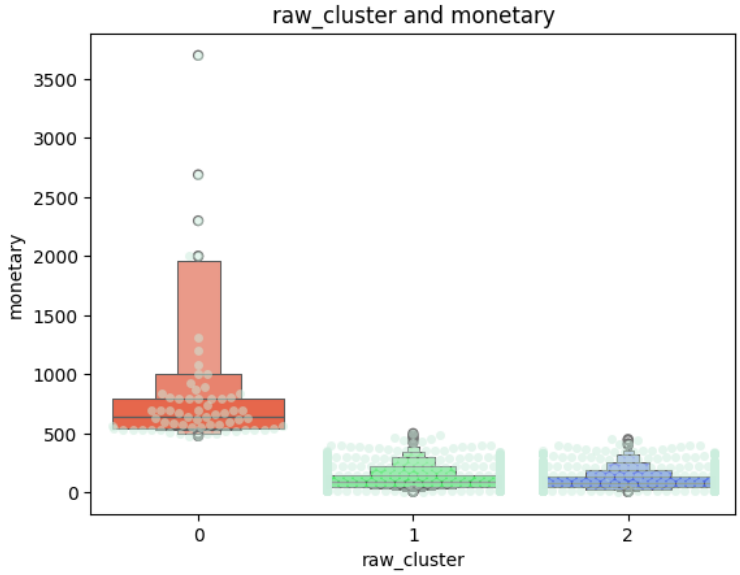
Karakteristik setiap cluster:



Gambar 21. Boxplot bagaimana data recency tersebar di Skenario 4 (raw)



Gambar 22. Boxplot bagaimana data frequncy tersebar di Skenario 4 (raw)



Gambar 23. Boxplot bagaimana data monetary tersebar di Skenario 4 (raw)

- Cluster 0: Pelanggan Semi-Aktif dengan Potensi Nilai Tinggi
  Karakteristik: Pelanggan Aktif
  - Recency: Rentang 0 hingga menengah atas. Artinya, sebagian pelanggan dalam cluster ini baru saja melakukan transaksi, sementara sebagian lainnya berada dalam tahap agak pasif, namun masih dalam batas waktu yang relatif dekat.
  - Frequency: Rentang 1–2. Frekuensi pembelian berada dalam kategori rendah, menandakan bahwa pelanggan belum melakukan pembelian berulang dalam jumlah signifikan.

- Monetary: Rentang menengah bawah hingga tinggi. Total nilai transaksi bervariasi, tetapi secara keseluruhan menunjukkan adanya peluang nilai yang besar dari beberapa pelanggan di dalam cluster ini.

Interpretasi:

Cluster ini merupakan pelanggan semi-aktif yang memiliki nilai transaksi beragam, dari cukup besar hingga tinggi. Walaupun mereka belum menunjukkan frekuensi pembelian yang tinggi, potensi pendapatan dari kelompok ini masih menjanjikan. Strategi yang cocok meliputi penguatan keterlibatan melalui pendekatan personal, edukasi produk, atau promosi eksklusif agar pelanggan terdorong melakukan pembelian ulang dengan nilai tinggi.

- Cluster 1: Pelanggan Lama dengan Nilai Transaksi Rendah

  Karakteristik:

  - Recency: Rentang menengah atas hingga tinggi. Pelanggan dalam kelompok ini telah cukup lama tidak bertransaksi, menunjukkan kecenderungan tidak aktif.
  - Frequency: Rentang 1–3. Meskipun sebagian kecil pelanggan memiliki jumlah pembelian lebih dari sekali, sebagian besar masih berada dalam kategori frekuensi rendah hingga sedang.
  - Monetary: Rentang 0 hingga menengah bawah. Nilai transaksi cenderung rendah dan tidak signifikan terhadap kontribusi total pendapatan.

  Interpretasi:

  Cluster ini merepresentasikan segmen pelanggan dengan kontribusi rendah dan keterlibatan yang sudah menurun. Karena nilai historisnya pun kecil, pelanggan dalam kelompok ini tidak menjadi prioritas utama dalam strategi pemasaran. Namun, mereka tetap dapat menjadi target untuk kampanye promosi berskala besar dengan biaya rendah, misalnya melalui email blast diskon, reminder keranjang, atau konten penarik perhatian untuk mencoba reaktivasi dengan efisien.

- Cluster 2: Pelanggan Aktif dengan Nilai Menengah

  Karakteristik:

  - Recency: Rentang 0 hingga menengah bawah. Sebagian besar pelanggan dalam kelompok ini baru-baru ini melakukan transaksi, menandakan keterlibatan saat ini masih tergolong baik.
  - Frequency: Rentang 1–3. Pelanggan telah melakukan transaksi satu hingga beberapa kali, yang menunjukkan aktivitas yang cukup positif, walaupun belum terlalu sering.
  - Monetary: Rentang 0 hingga menengah bawah. Nilai transaksi yang dihasilkan tergolong rendah, meskipun pelanggan aktif.

  Interpretasi:

  Cluster ini merupakan pelanggan aktif dengan daya beli rendah. Mereka sudah menunjukkan engagement yang baik secara waktu dan frekuensi, namun masih belum menghasilkan nilai pembelian yang tinggi. Strategi pemasaran yang cocok

untuk segmen ini adalah cross-selling atau upselling produk bernilai lebih tinggi, serta penawaran bundling yang dapat meningkatkan average transaction value secara bertahap.

3.4 Perbandingan Performa Antar Skenario

Tabel perbandingan metrik evaluasi:

Tabel 2. Perbandingan Performa antar Skenario

Skenario	*Silhouette Score*	*Dunn Index*	*CH Index*	*WCSS*
Standardisasi	0.7341	0.4663	934.3349	4088.2097
Normalization	0.5420	0.6652	2709.5608	52.5686
Log Transform	0.3756	0.8710	1436.3006	1272.5558
Raw Data	0.4657	0.7615	1330.7719	48710032.1096

Berdasarkan hasil evaluasi, skenario 1 memberikan performa clustering terbaik dengan *Silhouette Score* tertinggi 0.7341.

3.5 Interpretasi Segmen Pelanggan

Berdasarkan hasil clustering dengan skenario standardisasi, diperoleh dua segmen utama pelanggan. Meskipun jumlah klaster hanya dua (K=2), keduanya menunjukkan karakteristik yang dapat dipetakan dalam kerangka segmentasi RFM secara bermakna.

3.5.1 Segmen 1: *High Value, Low Engagement*

- Karakteristik: Recency sedang, Frequency rendah, Monetary tinggi
- Jumlah pelanggan: 1931 pelanggan (96.5%)
- Deskripsi: Pelanggan yang jarang melakukan transaksi, tetapi setiap transaksi bernilai besar. Mereka memiliki kontribusi revenue yang tinggi meskipun tingkat engagement-nya rendah.
- Strategi pemasaran:
    - *Reward program* untuk mempertahankan loyalitas pasif
    - *Personalized offers* berdasarkan waktu-waktu mereka cenderung bertransaksi
    - *Exclusive promotions* pada momen-momen khusus untuk mengaktivasi engagement

3.5.2 Segmen 2: *Engaged but Low Value*

- Karakteristik: Recency sedang, Frequency tinggi, Monetary rendah
- Jumlah pelanggan: 69 pelanggan (3.5%)
- Deskripsi: Segmen ini aktif secara frekuensi, namun nilai transaksinya lebih kecil dibandingkan segmen lainnya. Mereka menunjukkan loyalitas yang baik dan merupakan target potensial untuk peningkatan nilai pembelian.
- Strategi pemasaran:
    - *Upselling* untuk meningkatkan rata-rata nilai pembelian
    - *Cross-selling* produk terkait
    - *Loyalty program* untuk mempertahankan engagement dan memperbesar kontribusi *revenue*

3.6 Pembahasan

3.6.1 Pengaruh Feature Scaling terhadap Clustering

Hasil eksperimen dari berbagai skenario feature scaling (Standardisasi, Normalisasi, Log Transformasi, dan Raw) menunjukkan bahwa standardisasi memberikan hasil optimal secara keseluruhan, dengan skor Silhouette tertinggi yaitu 0.7341. Hal ini menandakan bahwa klaster yang terbentuk cukup terpisah dan kohesif.

Standardisasi menjaga distribusi relatif antar fitur tanpa terdistorsi oleh skala nilai yang berbeda, menjadikannya sesuai untuk algoritma seperti K-Means yang berbasis jarak. Hal ini konsisten dengan temuan Jain (2010), yang menekankan pentingnya skala fitur dalam meningkatkan performa algoritma clustering berbasis distance metric.

3.6.2 Validitas Segmentasi RFM

Segmentasi yang dihasilkan dari model clustering pada skenario standardisasi menunjukkan karakteristik pelanggan yang terdefinisi dengan baik.

- Distribusi tidak seimbang, namun tetap informatif: 96.5% dari pelanggan berada di cluster High Value-Low Engagement dan sisanya di cluster Engaged-Low Value.
- Meskipun hanya dua segmen, perbedaan dimensi RFM cukup signifikan dan dapat dimanfaatkan secara strategis.
- Hal ini mendukung literatur Wei et al. (2010) yang menyatakan bahwa RFM analysis mampu memberikan insight segmentatif yang bermakna dalam konteks bisnis.

3.6.3 Implikasi Bisnis

Hasil segmentasi pelanggan ini memberikan landasan kuat untuk implementasi strategi yang lebih terfokus:

- Customer Retention: Segmentasi membantu mengidentifikasi pelanggan dengan engagement rendah namun nilai tinggi, yang dapat menjadi prioritas dalam kampanye retensi.
- Marketing Personalization: Setiap segmen dapat ditargetkan dengan pendekatan pemasaran yang berbeda dan lebih relevan.
- Resource Allocation: Perusahaan dapat memfokuskan alokasi sumber daya marketing untuk memaksimalkan return dari pelanggan bernilai tinggi.
- *Customer Lifetime Value* (CLV): Strategi khusus per segmen dapat meningkatkan CLV secara keseluruhan, terutama melalui peningkatan engagement segmen bernilai tinggi.

3.6.4 Limitasi Penelitian

Beberapa keterbatasan perlu dicatat untuk interpretasi hasil dan pengembangan lanjutan:

- Terbatas pada dimensi RFM saja, tanpa mempertimbangkan preferensi produk atau interaksi lainnya seperti kanal komunikasi yang digunakan.
- Waktu pengumpulan data bersifat snapshot, tidak menangkap pola musiman atau perubahan tren pembelian.

- Asumsi K-Means terhadap bentuk klaster yang sferis, bisa membatasi kemampuan dalam mengenali pola segmentasi yang tidak berbentuk bulat secara geometris.
- Distribusi klaster yang sangat tidak seimbang, meskipun informatif, bisa membatasi aplikasi strategi yang proporsional di seluruh segmen.

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Means untuk melakukan segmentasi pelanggan berbasis analisis RFM, menggunakan dataset transaksi yang mencakup 94.488 entri. Melalui tahapan eksplorasi data, pembersihan, dan evaluasi model secara menyeluruh, empat pendekatan feature scaling yaitu standardisasi, normalisasi min-max, transformasi logaritmik, dan penggunaan data mentah—diuji untuk menentukan metode paling efektif dalam membentuk segmentasi pelanggan yang relevan secara bisnis.

Hasil analisis menunjukkan bahwa pendekatan standardisasi memberikan performa terbaik, tercermin dari nilai *Silhouette Score* tertinggi sebesar 0.7341 dan *Calinski-Harabasz Index* sebesar 934.33. Temuan ini mengindikasikan bahwa standardisasi mampu menjaga proporsi antar fitur dengan baik, sehingga mendukung pembentukan cluster yang lebih optimal pada algoritma berbasis jarak seperti K-Means.

Segmentasi yang dihasilkan dari pendekatan terbaik ini membentuk dua kelompok pelanggan yang cukup kontras. Kelompok pertama mencakup pelanggan dengan nilai transaksi tinggi tetapi frekuensi pembelian rendah (*High Value, Low Engagement*), sementara kelompok kedua terdiri dari pelanggan yang sering bertransaksi namun dengan nilai yang lebih kecil (*Engaged but Low Value*). Meskipun hanya menghasilkan dua segmen, pembagian ini tetap memberikan insight yang kuat dan dapat ditindaklanjuti untuk mendukung keputusan bisnis.

Hasil segmentasi ini sangat potensial untuk digunakan dalam berbagai strategi, seperti mempertahankan pelanggan bernilai tinggi, merancang program upselling dan cross-selling yang lebih terarah, serta mengoptimalkan alokasi sumber daya pemasaran. Validasi terhadap hasil clustering juga telah dilakukan secara menyeluruh menggunakan berbagai metrik—Silhouette Score, Dunn Index, Calinski-Harabasz Index, dan WCSS—yang bersama-sama mengonfirmasi kualitas pemisahan dan kekompakan antar cluster.

DAFTAR PUSTAKA

[1] P. W. Nofiani dan M. C. Mursid, “Pentingnya Perilaku Organisasi dan Strategi Pemasaran dalam Menghadapi Persaingan Bisnis di Era Digital,” *Jurnal Logistik Bisnis*, vol. 11, no. 2, pp. 71–77, Nov. 2021. [Online]. Tersedia: https://ejurnal.poltekpos.ac.id/index.php/logistik/index

[2] R. Y. Firmansah, J. D. Irawan, dan N. Vendyansyah, “Analisis RFM (Recency, Frequency and Monetary) Produk Menggunakan Metode K-Means,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 5, no. 1, pp. 334–341, Mar. 2021.

[3] K. Akhmad, A. Andriyanto, Y. Yearrimia, dan J. Heikal, "RFM Model Analysis to Increase Merchant Retention in Payment Gateway Companies. Case Study: PT F-PAY," *Manajemen Strategis Terkini*, vol. 6, no. 3, pp. 17–26, Sep. 2024. [Online]. Tersedia: https://journalpedia.com/1/index.php/mst

[4] V. Kumar dan W. Reinartz, "Creating Enduring Customer Value," *Journal of Marketing*, vol. 80, no. 6, pp. 36–68, Nov. 2016.

[5] K. Z. Wijaya, A. Djunaidy, dan F. Mahananto, "Segmentasi Pelanggan Menggunakan Algoritma K-Means dan Analisis RFM di Ova Gaming E-Sports Arena Kediri," *Jurnal Teknik ITS*, vol. 10, no. 2, pp. A230–A237, 2021. [Online]. Tersedia: https://ejurnal.its.ac.id/index.php/teknik

[6] D. A. Imanuel dan G. Alfian, "Visualisasi Segmentasi Pelanggan Berdasarkan Atribut RFM Menggunakan Algoritma K-Means untuk Memahami Karakteristik Pelanggan pada Toko Retail Online," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 12, no. 2, pp. 283–292, Apr. 2025. DOI: 10.25126/jtiik.2025128619