PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 903-909
E-ISSN: 3109-8886

Penerapan Analisis RFM dan K-Means Clustering untuk Segmentasi Pelanggan E-Commerce: Studi Kasus Dataset Olist Brazil

[1]Raybima Yoga Rajata, [2]Abimanyu Andika Aulia, [3]Fajar Nugroho

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Banten

[1]raybima9@gmail.com, [2]abimanyuandikaaulia@gmail.com, [3]fajarn318@gmail.com

Abstract

The ability to understand individual customer behavior has become a critical competitive factor for e-commerce platforms, particularly in emerging markets characterized by structural dominance of one-time buyers. This study applies RFM (Recency, Frequency, Monetary) analysis combined with K-Means Clustering to segment 93,357 unique customers from the Olist Brazilian E-Commerce dataset (2016–2018) within the CRISP-DM framework. Data preprocessing included filtering by delivered status, tiered missing value handling, duplicate elimination, and normalization using log1p transformation followed by StandardScaler. The optimal number of clusters was determined through a combination of the Elbow Method and Silhouette Score across K=2 to K=10; K=2 was selected as it yielded the highest Silhouette Score (0.7028), with a substantial margin over K=3 (0.4144). The final model achieved a Silhouette Score of 0.7139 and a Davies-Bouldin Index of 0.4709, indicating well-separated and compact clusters. Two segments were identified: Loyal Customer (3.0%, n=2,801) with an average monetary value of BRL 260.05 and frequency of 2.11 orders, and At Risk (97.0%, n=90,556) with an average monetary value of BRL 137.96 and frequency of 1.00. The dominance of the At Risk segment reflects the structural characteristic of the Brazilian e-commerce market, which is predominantly composed of one-time buyers a finding consistent with segmentation literature in emerging market contexts. Differentiated marketing strategies are proposed for each segment to maximize retention of high-value customers while reactivating dormant ones.

Keywords: *customer segmentation, RFM, K-Means Clustering, e-commerce, CRISP-DM, Olist*

Abstrak

Kemampuan memahami perilaku pelanggan secara individual menjadi faktor kritis dalam persaingan platform e-commerce, khususnya di pasar berkembang yang didominasi pembelian satu kali (one-time buyer). Penelitian ini menerapkan analisis RFM (Recency, Frequency, Monetary) yang dikombinasikan dengan algoritma K-Means Clustering untuk mensegmentasi 93.357 pelanggan unik dari dataset Olist Brazilian E-Commerce (2016–2018) dalam kerangka CRISP-DM. Preprocessing data mencakup filter status delivered, penanganan missing value bertingkat, eliminasi duplikat, serta normalisasi menggunakan transformasi log1p dilanjutkan StandardScaler. Jumlah klaster optimal ditentukan melalui kombinasi Elbow Method dan Silhouette Score pada rentang K=2 hingga K=10; K=2 dipilih karena menghasilkan Silhouette Score tertinggi (0,7028) dengan selisih yang signifikan terhadap K=3 (0,4144). Model akhir menghasilkan Silhouette Score 0,7139 dan Davies-Bouldin Index 0,4709, mengindikasikan pemisahan klaster yang kompak dan terdefinisi dengan baik. Dua segmen teridentifikasi: Loyal Customer (3,0%, n=2.801) dengan rata-rata monetary BRL 260,05 dan frequency 2,11 order, serta At Risk (97,0%, n=90.556) dengan rata-rata monetary BRL 137,96 dan frequency 1,00. Dominasi segmen At Risk mencerminkan karakteristik struktural pasar e-commerce Brazil yang didominasi one-time buyer temuan yang konsisten dengan literatur segmentasi di pasar berkembang. Berdasarkan profil tiap segmen, rekomendasi strategi pemasaran yang berbeda dirumuskan untuk memaksimalkan retensi pelanggan bernilai tinggi sekaligus mengaktifkan kembali pelanggan yang tidak aktif.

Kata Kunci: *segmentasi pelanggan, RFM, K-Means Clustering, e-commerce, CRISP-DM, Olist*

A. PENDAHULUAN

Persaingan di industri e-commerce global telah mendorong platform-platform besar untuk bergeser dari pendekatan pemasaran massal menuju strategi yang lebih personal dan berbasis data. Brazil, sebagai pasar e-commerce terbesar di Amerika Latin, mencatatkan nilai transaksi digital sebesar USD 49,2 miliar pada 2022 tumbuh lebih dari dua kali lipat dibanding 2018 didorong oleh meluasnya penetrasi internet dan adopsi pembayaran digital [1]. Di tengah

Diterima Redaksi: xx-xx-xxx | Selesai Revisi: xx-xx-2022 | Diterbitkan Online: xx-xx-xxx

ekspansi ini, kemampuan memahami perilaku pelanggan secara individual menjadi faktor pembeda yang krusial bagi keberlangsungan platform.

Segmentasi pelanggan merupakan salah satu teknik analitik yang paling mapan dalam manajemen hubungan pelanggan (*Customer Relationship Management*/CRM). Dengan mengelompokkan pelanggan berdasarkan kesamaan perilaku, perusahaan dapat mengalokasikan sumber daya pemasaran secara proporsional memberikan perhatian lebih pada pelanggan bernilai tinggi sekaligus merancang strategi pemulihan bagi mereka yang menunjukkan tanda-tanda *churn* [2]. Kotler & Keller [2] mencatat bahwa biaya akuisisi pelanggan baru secara umum lima hingga tujuh kali lebih mahal dibanding biaya mempertahankan pelanggan yang ada, sehingga efektivitas segmentasi berdampak langsung pada efisiensi anggaran pemasaran.

Di antara berbagai kerangka segmentasi yang tersedia, analisis RFM (*Recency*, *Frequency*, *Monetary*) telah menjadi standar dalam pemasaran berbasis database sejak diperkenalkan Hughes [3] dalam konteks *direct marketing* pada 1994. Kekuatan RFM terletak pada kemampuannya menangkap tiga dimensi perilaku transaksi yang paling informatif secara simultan, tanpa membutuhkan data demografis atau survei tambahan. Christy et al. [4] menunjukkan bahwa pendekatan berbasis RFM yang dikombinasikan dengan teknik *clustering* menghasilkan segmen yang lebih stabil dan dapat diinterpretasikan secara bisnis dibanding metode segmentasi berbasis skor tunggal.

K-Means merupakan algoritma *clustering* yang paling umum diterapkan bersama RFM, terutama karena efisiensi komputasinya pada dataset berukuran besar dan kemudahan interpretasinya [5]. Penerapan teknik *data mining* untuk pengambilan keputusan berbasis data telah berkembang melampaui domain pemasaran termasuk dalam konteks manajemen mutu dan prediksi kinerja [6] yang memperkuat relevansi pendekatan serupa dalam konteks segmentasi pelanggan e-commerce. Meski demikian, pemilihan jumlah klaster yang tepat tetap menjadi tantangan metodologis yang sering diselesaikan secara tidak konsisten dalam literatur. Penelitian ini menggunakan kombinasi *Elbow Method* dan *Silhouette Score* untuk memastikan pemilihan K yang dapat dipertanggungjawabkan secara kuantitatif.

Sejumlah penelitian telah menerapkan kombinasi RFM dan K-Means pada dataset e-commerce, namun sebagian besar menggunakan dataset dari pasar maju (Tiongkok, Amerika Serikat) dengan karakteristik pembelian ulang yang lebih tinggi. Penelitian pada konteks pasar berkembang yang strukturnya didominasi *one-time buyer* masih terbatas [4]. Dataset Olist Brazilian E-Commerce, yang mencakup lebih dari 99.000 transaksi dari periode 2016–2018 [8], menawarkan kesempatan untuk menguji apakah pendekatan RFM + K-Means menghasilkan segmentasi yang bermakna secara bisnis meski dalam kondisi distribusi transaksi yang sangat tidak seimbang.

Penelitian ini bertujuan untuk: (1) membangun profil RFM dari data transaksi Olist menggunakan kerangka CRISP-DM, (2) menentukan jumlah segmen optimal melalui evaluasi *Silhouette Score* dan *Elbow Method*, serta (3) menghasilkan rekomendasi strategi pemasaran yang dapat ditindaklanjuti berdasarkan karakteristik tiap segmen. Temuan penelitian diharapkan memberikan kontribusi praktis bagi manajer pemasaran platform e-commerce di pasar berkembang, sekaligus memperkaya literatur segmentasi pelanggan berbasis data dalam konteks distribusi transaksi yang *highly skewed*.

B. METODE

Penelitian ini mengikuti kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri atas enam fase berurutan: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment [7]. Pendekatan ini dipilih karena sifatnya yang iteratif dan terstruktur, sehingga setiap keputusan teknis dalam proses analisis dapat ditelusuri kembali ke tujuan bisnis yang telah ditetapkan sejak awal.

1. Dataset

Data yang digunakan bersumber dari dataset publik Olist Brazilian E-Commerce yang tersedia di platform Kaggle [8]. Dataset ini mencakup transaksi e-commerce di Brazil sepanjang periode September 2016 hingga Oktober 2018, tersebar dalam lima tabel relasional: *orders*, *order_items*, *order_payments*, *customers*, dan *products*. Dari kelima tabel tersebut, informasi yang relevan untuk konstruksi fitur RFM diekstrak melalui proses *join* berdasarkan kunci *order_id* dan *customer_id*, menghasilkan 99.441 baris data order sebagai titik awal analisis.

2. Preprocessing Data

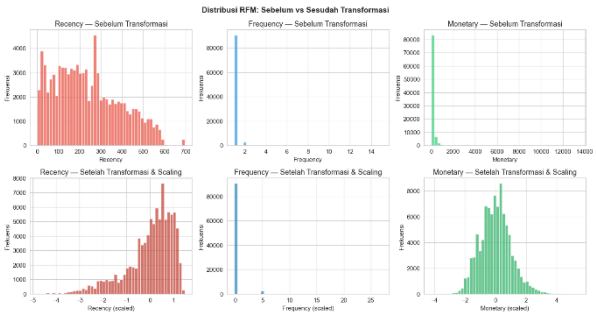
Tahap persiapan data diawali dengan pemfilteran berdasarkan kolom order_status, di mana hanya transaksi berstatus *"delivered"* yang dipertahankan. Langkah ini penting untuk memastikan bahwa nilai *Monetary* yang dihitung benar-benar mencerminkan transaksi yang telah selesai secara penuh, bukan pesanan yang masih dalam proses atau dibatalkan. Setelah filter diterapkan, dilakukan penanganan *missing value* secara bertingkat: variabel kritis seperti order_delivered_customer_date dan customer_id dihapus baris-nya jika kosong, sedangkan variabel non-kritis diimputasi menggunakan nilai median. Pemeriksaan duplikat pada level order_id dan filter nilai total_payment ≤ 0 turut dilakukan sebagai langkah validasi, namun tidak ditemukan rekaman yang memenuhi kondisi tersebut dalam dataset ini.

Tanggal referensi (*snapshot date*) yang digunakan untuk menghitung *Recency* ditetapkan sebagai satu hari setelah tanggal transaksi terakhir dalam dataset, mengikuti konvensi umum dalam literatur RFM [4]. Nilai *Monetary* dalam penelitian ini dihitung berdasarkan total_price yakni akumulasi harga produk per pelanggan bukan dari payment_value, agar mencerminkan nilai intrinsik pembelian tanpa distorsi dari biaya pengiriman atau metode pembayaran.

3. Konstruksi Fitur RFM

Setelah preprocessing selesai, tiga fitur utama dikonstruksi pada level pelanggan unik. *Recency* didefinisikan sebagai selisih hari antara tanggal *snapshot* dan tanggal transaksi terakhir pelanggan nilai yang lebih kecil mengindikasikan aktivitas yang lebih baru. *Frequency* dihitung sebagai total jumlah order yang telah diselesaikan oleh setiap pelanggan dalam periode observasi. *Monetary* merupakan total akumulasi total_price dari seluruh transaksi pelanggan yang bersangkutan.

Distribusi ketiga fitur ini sebelum transformasi menunjukkan kemiringan positif (*right-skewed*) yang kuat, terutama pada *Frequency* yang memiliki lonjakan ekstrem pada nilai 1 mencerminkan dominasi pelanggan yang hanya bertransaksi satu kali. Kondisi ini umum ditemui pada dataset e-commerce dan berpotensi mengganggu kinerja algoritma berbasis jarak seperti K-Means jika tidak ditangani [5]. Oleh karena itu, transformasi log1p diterapkan pada ketiga fitur untuk mereduksi skewness, dilanjutkan dengan standarisasi menggunakan StandardScaler agar tiap dimensi memiliki mean nol dan standar deviasi satu. Perbandingan distribusi sebelum dan sesudah transformasi dapat dilihat pada Gambar 1.



Gambar 1. Distribusi Fitur RFM Sebelum dan Sesudah Transformasi Log1p dan StandarScaler

4. Pemodelan K-Means Clustering

Penelitian ini mengimplementasikan K-Means Clustering dengan inisialisasi *centroid* menggunakan metode k-means++ [9] dipilih karena metode ini terbukti mengurangi risiko konvergensi ke solusi lokal suboptimal dibanding inisialisasi acak standar, yang menjadi perhatian khusus pada dataset berdistribusi tidak seimbang seperti Olist. Parameter *max_iter* ditetapkan pada 300 dengan *n_init* sebesar 10 untuk memastikan stabilitas hasil di setiap eksekusi.

Untuk menentukan nilai K yang optimal, dua pendekatan evaluasi diterapkan secara bersamaan pada rentang K=2 hingga K=10. *Elbow Method* memvisualisasikan nilai inertia terhadap K , titik di mana kurva mulai melandai mengindikasikan penambahan klaster tidak lagi memberikan pengurangan inertia yang substansial. Karena titik infleksi *Elbow* sering ambigu secara visual, *Silhouette Score* [10] digunakan sebagai penentu akhir yang bersifat kuantitatif: nilai K dengan *Silhouette Score* tertinggi dipilih sebagai K optimal.

5 Evaluasi Model

Evaluasi akhir model menggunakan dua metrik yang saling komplementer. *Silhouette Score* [10] mengukur kekompakan dan separabilitas klaster pada skala −1 hingga 1, di mana nilai mendekati 1 mengindikasikan klaster yang terdefinisi dengan baik. *Davies-Bouldin Index* [11] dipilih sebagai metrik kedua karena sensitif terhadap rasio dispersi intra-klaster terhadap jarak antar-*centroid* , karakteristik yang relevan untuk dataset dengan distribusi tidak seimbang seperti yang dihadapi dalam penelitian ini. Kombinasi keduanya digunakan karena kedua metrik kadang memberikan sinyal yang berbeda; konsistensi antara keduanya memperkuat kepercayaan terhadap stabilitas hasil segmentasi.

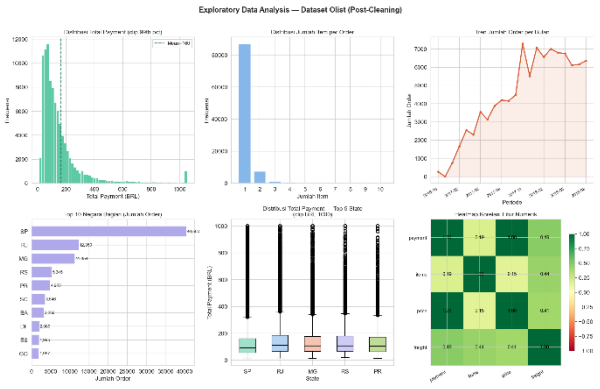
C. HASIL DAN PEMBAHASAN

1. Eksplorasi Data Awal

Sebelum konstruksi fitur RFM dilakukan, pemahaman terhadap karakteristik dataset secara menyeluruh menjadi langkah yang tidak bisa dilewati. Dari 99.441 baris data order yang tersedia, sebanyak 96.478 berstatus *delivered* dan setelah penanganan *missing value* serta eliminasi duplikat, total pelanggan unik yang masuk ke tahap analisis berjumlah 93.357.

Rata-rata nilai pembayaran per transaksi berada di angka BRL 160, dengan distribusi yang sangat miring ke kanan (*right-skewed*) sebagian kecil transaksi bernilai sangat tinggi menarik rata-rata jauh dari nilai tengah yang sesungguhnya. Mayoritas order, lebih dari 85%, hanya berisi satu item produk. Pola ini mengisyaratkan bahwa platform Olist lebih banyak digunakan untuk pembelian tunggal yang bersifat insidental, bukan untuk belanja rutin dalam jumlah besar.

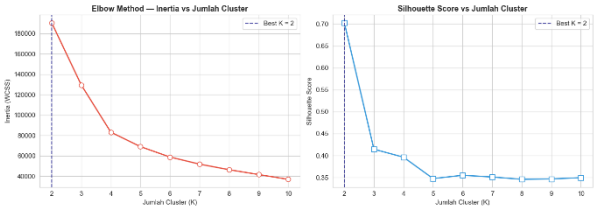
Berdasarkan analisis distribusi order dalam dataset Olist, negara bagian São Paulo (SP) mendominasi volume transaksi dengan 40.500 order, diikuti Rio de Janeiro (RJ) sebesar 12.350 dan Minas Gerais (MG) sebesar 11.354. Konsentrasi transaksi di tiga negara bagian ini mencerminkan distribusi populasi dan infrastruktur digital Brazil yang masih sangat terpusat di wilayah tenggara. Tren temporal menunjukkan pertumbuhan order yang konsisten dari akhir 2016 hingga puncaknya di akhir 2017, kemudian stabil sepanjang 2018. Analisis korelasi antar fitur numerik mengungkapkan hubungan yang sangat kuat antara payment_value dan price ($r \approx 1{,}00$), sementara freight_value berkorelasi sedang dengan payment_value ($r = 0{,}49$). Ringkasan eksplorasi data ini tersaji pada Gambar berikut.



Gambar 2. Hasil Eksplorasi Data Awal Dataset Olist

2. Pemilihan Jumlah Cluster Optimal

Pengujian nilai K dilakukan pada rentang 2 hingga 10 dengan mencatat inertia dan *Silhouette Score* di setiap iterasi. Seperti terlihat pada Gambar 3, kurva inertia menurun tajam dari K=2 ke K=3, kemudian melandai secara bertahap pola klasik yang membentuk "siku" namun tanpa titik infleksi yang benar-benar tegas. Ambiguitas pada *Elbow Method* ini justru memperkuat urgensi penggunaan metrik kuantitatif sebagai penentu akhir.



Gambar 3. Elbow Method dan Silhouette Score untuk Pemilihan Jumlah Cluster Optimal

Silhouette Score memberikan sinyal yang jauh lebih jelas. Nilai tertinggi diraih pada K=2 sebesar 0,7028, lalu turun drastis ke 0,4144 pada K=3 dan terus menurun hingga 0,3458 pada K=8. Penurunan sebesar hampir 0,29 poin dari K=2 ke K=3 merupakan selisih yang terlalu besar untuk diabaikan. Berdasarkan argmax *Silhouette Score* ini, K=2 ditetapkan sebagai jumlah klaster optimal. Seluruh hasil pengujian K disajikan pada Tabel 1.

Tabel 1. Hasil Evaluasi Inertia dan Silhouette Score untuk K=2 hingga K=10

K	Inertia	Silhouette Score
2	190.704,14	0,7028 ← optimal
3	129.654,05	0,4144
4	83.292,18	0,3962
5	69.091,55	0,3474
6	58.881,03	0,3550
7	51.894,82	0,3511
8	46.419,11	0,3458
9	41.672,05	0,3467
10	37.001,17	0,3495

3. Profil dan Karakteristik Cluster

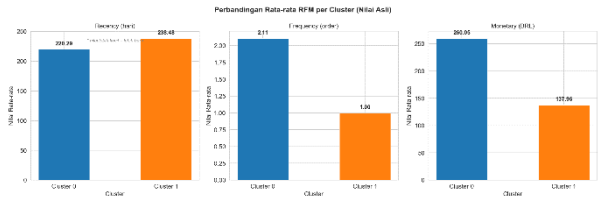
Evaluasi akhir model menghasilkan *Silhouette Score* sebesar 0,7139 dan *Davies-Bouldin Index* sebesar 0,4709. Perbedaan antara nilai *Silhouette Score* pada tahap seleksi K (0,7028) dan model final (0,7139) disebabkan oleh perbedaan inisialisasi *random seed* antar eksekusi keduanya berada dalam rentang yang konsisten dan tidak mengubah keputusan pemilihan K. Nilai DBI sebesar 0,4709, yang berada jauh di bawah ambang 1,0, mengindikasikan bahwa jarak rata-rata intra-klaster relatif kecil dibanding jarak antar-*centroid* [11] konsisten dengan nilai *Silhouette Score* yang tinggi.

Pelabelan klaster dilakukan berdasarkan interpretasi nilai rata-rata RFM pada skala asli (sebelum transformasi). Klaster 0 diberi label **Loyal Customer**, sementara Klaster 1 dilabeli **At Risk**. Profil lengkap kedua segmen tersaji pada Tabel 2.

Tabel 2. Profil Rata-rata RFM per Segmen (Nilai Asli)

Cluster	Label	n	%	Avg Recency (hari)	Avg Frequency	Avg Monetary (BRL)
C0	Loyal Customer	2.801	3,0%	220,29	2,11	260,05
C1	At Risk	90.556	97,0%	238,48	1,00	137,96

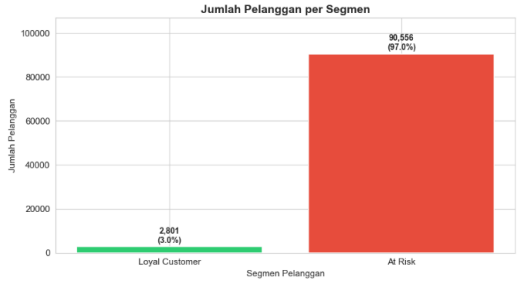
Segmen *Loyal Customer* (C0) mencakup 2.801 pelanggan hanya 3,0% dari total basis pengguna namun menunjukkan profil RFM yang secara konsisten lebih unggul di seluruh dimensi. Nilai rata-rata *Monetary* segmen ini sebesar BRL 260,05, hampir dua kali lipat dibanding segmen *At Risk* yang hanya BRL 137,96. *Frequency* rata-rata sebesar 2,11 order mengkonfirmasi bahwa pelanggan di segmen ini telah melakukan lebih dari satu kali transaksi sebuah pencapaian yang tidak trivial di pasar e-commerce yang didominasi *one-time buyer*. Perbandingan visual ketiga dimensi RFM antar segmen dapat dilihat pada Gambar 4.



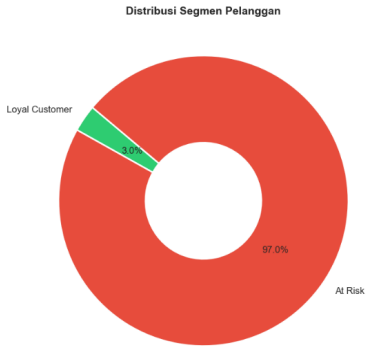
Gambar 4. Perbandingan Rata-Rata RFM per Cluster (Nilai Asli)

Segmen *At Risk* (C1) mencakup 90.556 pelanggan atau 97,0% dari seluruh basis pengguna yang dianalisis. Nilai *Frequency* rata-rata tepat di angka 1,00 yang berarti hampir seluruh pelanggan di segmen ini hanya bertransaksi satu kali sepanjang periode observasi. *Recency* rata-rata sebesar 238,48 hari menunjukkan bahwa transaksi terakhir

mereka terjadi hampir delapan bulan yang lalu, jauh melampaui ambang waktu yang lazim diasosiasikan dengan risiko *churn* dalam literatur CRM [2]. Distribusi jumlah pelanggan per segmen divisualisasikan pada Gambar 5 dan Gambar 6.



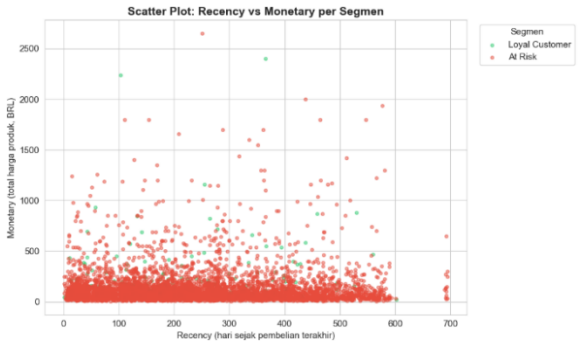
Gambar 5. Distribusi Jumlah Pelanggan per Segmen



Gambar 6. Proporsi Distribusi Segmen Pelanggan

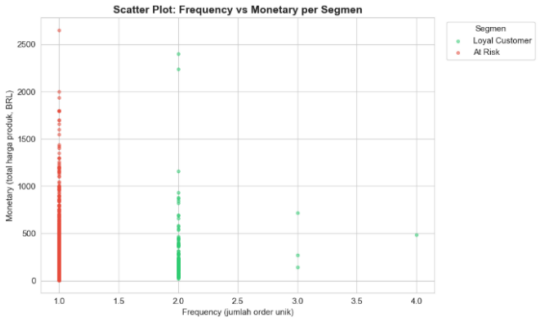
4. Analisis Karakteristik Sebaran Pelanggan

Jika ditelaah lebih dalam melalui ruang dua dimensi, perbedaan antara kedua segmen menjadi semakin gamblang. Pada Gambar 7, sebaran pelanggan di ruang *Recency* vs *Monetary* memperlihatkan bahwa segmen *At Risk* (titik merah) tersebar hampir merata di seluruh rentang *recency* dengan nilai *monetary* yang relatif rendah dan terkonsentrasi. Segmen *Loyal Customer* (titik hijau), meski jumlahnya jauh lebih sedikit, menampilkan variasi *monetary* yang lebih lebar mengindikasikan heterogenitas nilai transaksi yang lebih tinggi di dalam segmen ini.



Gambar 7. Sebaran Pelanggan pada Ruang Recency vs Monetary per Segmen

Gambar 8 memperjelas dinamika *frequency*. Titik-titik merah (*At Risk*) menumpuk hampir seluruhnya pada nilai *Frequency* = 1, membentuk kolom vertikal yang padat di sisi kiri grafik. Sebaliknya, titik-titik hijau (*Loyal Customer*) tersebar pada rentang *Frequency* 1 hingga 4, dengan sebaran *monetary* yang lebih bervariasi di tiap level frekuensi. Pola ini secara visual mengkonfirmasi temuan statistik pada Tabel 2.



Gambar 8. Sebaran Pelanggan pada Ruang Frequency vs Monetary per Segmen

5. Diskusi: Dominasi Segmen At Risk sebagai Temuan Struktural

Proporsi 97,0% pelanggan pada segmen *At Risk* mungkin tampak tidak proporsional pada pembacaan pertama, namun temuan ini konsisten dengan karakteristik struktural pasar e-commerce di negara berkembang. Christy et al. [4] mencatat bahwa *repeat purchase rate* pada platform e-commerce di pasar berkembang secara konsisten berada di bawah 20%, sementara Kotler & Keller [2] menegaskan bahwa dominasi pembeli satu kali merupakan tantangan struktural yang tidak dapat diselesaikan semata-mata melalui akuisisi pelanggan baru. Dengan demikian, distribusi segmen yang dihasilkan model bukan artefak dari pemilihan K=2, melainkan cerminan empiris dari kondisi pasar yang sesungguhnya.

Dari perspektif *Customer Lifetime Value* (CLV), segmen *Loyal Customer* menunjukkan nilai *monetary* rata-rata BRL 260,05 hampir dua kali lipat segmen *At Risk* (BRL 137,96) dengan proporsi hanya 3,0% dari total basis pengguna. Ketidakproporsionalan antara ukuran segmen dan nilai yang dihasilkan ini relevan secara strategis: Kotler & Keller [2] mencatat bahwa biaya akuisisi pelanggan baru umumnya lima hingga tujuh kali lebih tinggi dibanding biaya retensi, sehingga mempertahankan 2.801 pelanggan *Loyal Customer* berpotensi memberikan efisiensi anggaran yang signifikan dibanding upaya akuisisi dengan skala setara.

Perlu dicatat bahwa segmentasi dengan K=2 menghasilkan pembagian yang bersifat dikotomis *loyal* versus *at risk* tanpa gradasi segmen menengah. Pada dataset dengan distribusi *repeat purchase* yang lebih merata, penambahan jumlah klaster kemungkinan akan mengungkap segmen dengan karakteristik transisional yang bernilai strategis. Keterbatasan ini menjadi salah satu arah pengembangan yang relevan untuk penelitian selanjutnya.

6. Rekomendasi Strategi Pemasaran

Perbedaan profil RFM yang signifikan antara kedua segmen menuntut pendekatan pemasaran yang berbeda secara fundamental. Rekomendasi berikut dirumuskan berdasarkan prinsip-prinsip CRM yang dikemukakan Kotler & Keller [2] dan disesuaikan dengan karakteristik tiap segmen.

Bagi segmen *Loyal Customer*, prioritas utama adalah retensi dan penguatan loyalitas. Pelanggan dengan *frequency* rata-rata 2,11 order telah melewati ambang kepercayaan pertama sebuah kondisi yang dalam literatur CRM diasosiasikan dengan peningkatan probabilitas pembelian ulang [2]. Program loyalitas berbasis tingkatan (*tiered loyalty program*) dengan insentif proporsional terhadap nilai transaksi lebih efektif dibanding diskon generik karena memperkuat persepsi pengakuan nilai pelanggan. Personalisasi rekomendasi produk berbasis riwayat transaksi dan layanan prioritas merupakan komponen tambahan yang disarankan untuk mempertahankan engagament segmen ini. Program referral juga berpotensi mengubah pelanggan loyal menjadi kanal akuisisi organik yang biaya konversinya lebih rendah dibanding iklan berbayar [2].

Bagi segmen *At Risk*, tantangan utama adalah minimnya riwayat transaksi yang dapat dijadikan sinyal preferensi nilai *frequency* rata-rata 1,00 berarti hampir tidak ada data perilaku kedua yang tersedia. Strategi *win-back* berbasis insentif finansial dengan urgensi temporal (*time-limited offer*) disarankan sebagai pendekatan awal, mengacu pada prinsip psikologi keputusan konsumen mengenai *loss aversion* dan *scarcity* [2]. Komunikasi melalui kanal dengan tingkat keterbacaan tinggi (email, SMS, atau *push notification*) dengan pesan yang merujuk pada kategori produk yang pernah dibeli sebelumnya dapat meningkatkan relevansi pesan. Mekanisme *feedback* pasca-reaktivasi misalnya survei singkat disarankan untuk membangun data preferensi yang selama ini tidak tersedia bagi segmen ini, sekaligus membuka interaksi dua arah yang dapat mendorong keterlibatan lebih lanjut.

Perlu ditekankan bahwa seluruh rekomendasi di atas bersifat proporsional berbasis literatur dan memerlukan validasi empiris melalui eksperimen terkontrol misalnya *A/B testing* pada kampanye retensi berbasis segmen sebelum diterapkan pada skala platform secara penuh.

D. PENUTUP

1. Kesimpulan

Penelitian ini berhasil membangun sistem segmentasi pelanggan berbasis data transaksi Olist Brazilian E-Commerce menggunakan kombinasi analisis RFM dan algoritma K-Means Clustering dalam kerangka CRISP-DM. Dari 93.357 pelanggan unik yang dianalisis, model menghasilkan dua segmen yang terdefinisi dengan baik dikonfirmasi oleh *Silhouette Score* akhir sebesar 0,7139 dan *Davies-Bouldin Index* sebesar 0,4709.

Pemilihan K=2 sebagai jumlah klaster optimal bukan keputusan arbitrer. *Silhouette Score* pada K=2 sebesar 0,7028 unggul dengan selisih yang signifikan dibanding K=3 (0,4144), dan nilai ini tidak tersaingi oleh konfigurasi K manapun hingga K=10. Penurunan tajam setelah K=2 mengindikasikan bahwa struktur alami data hanya mendukung dua kelompok yang benar-benar terpisah secara statistik.

Dua segmen yang terbentuk *Loyal Customer* (3,0%, n=2.801) dan *At Risk* (97,0%, n=90.556) mencerminkan realitas pasar e-commerce Brazil yang sesungguhnya. Dominasi segmen *At Risk* merupakan bukti empiris dari fenomena *one-time buyer* yang menjadi tantangan struktural platform e-commerce di pasar berkembang [4]. Segmen *Loyal Customer*, meski minoritas, memiliki nilai *monetary* rata-rata BRL 260,05 hampir dua kali lipat segmen *At Risk* yang menegaskan potensi CLV yang tidak proporsional dengan ukurannya.

Secara metodologis, transformasi log1p dan StandardScaler diterapkan untuk menangani distribusi RFM yang sangat *skewed*, khususnya pada dimensi *Frequency* yang memiliki lonjakan ekstrem di nilai 1. Tanpa transformasi ini, jarak Euclidean dalam ruang fitur berpotensi didominasi oleh *outlier monetary* sehingga menghasilkan klaster yang kurang bermakna secara bisnis.

2. Saran

Beberapa arah pengembangan layak dipertimbangkan untuk penelitian selanjutnya. Pertama, eksplorasi algoritma *clustering* alternatif seperti DBSCAN [12] atau *Gaussian Mixture Model* (GMM) [13] pada dataset yang sama dapat menjadi pembanding yang menarik kedua algoritma tersebut tidak mengharuskan penentuan jumlah klaster di awal dan lebih toleran terhadap distribusi yang tidak sferis, sehingga berpotensi mengungkap struktur segmen yang tidak tertangkap oleh K-Means.

Kedua, integrasi dimensi tambahan di luar RFM seperti kategori produk, metode pembayaran, atau skor ulasan pelanggan berpotensi menghasilkan segmentasi yang lebih granular dan kaya secara interpretatif. Ketiga, penelitian ini menggunakan data statis dari periode 2016–2018; penerapan segmentasi pada data *streaming* secara *real-time*

akan memberikan nilai operasional yang lebih langsung bagi platform, mengingat perilaku pelanggan e-commerce berubah mengikuti siklus promosi dan musiman.

Terakhir, validasi eksternal terhadap hasil segmentasi melalui eksperimen *A/B testing* pada strategi rekomendasi yang berbeda per segmen akan memperkuat klaim kausalitas yang dalam penelitian ini masih bersifat korelatif.

E. DAFTAR PUSTAKA

[1] Statista. (2023). *E-commerce in Brazil Statistics & Facts*. Statista Research Department.

[2] Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson Education.

[3] Hughes, A. M. (1994). *Strategic Database Marketing*. Probus Publishing.

[4] 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.

[5] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108.

[6] Karimah, M., & Marwati, F. (2024). Sustainability of Quality Management by Implementing Data Mining to Predict Academic Achievement. Journal of Social Science and Business Studies, 2(3), 240–250. https://doi.org/10.61487/jssbs.v2i3.90

[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

[8] Olist. (2018). *Brazilian E-Commerce Public Dataset by Olist* [Dataset]. Kaggle. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

[9] Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

[10] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

[11] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

[12] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

[13] Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of Biometrics*, 741–749.