

Terbit online pada laman : <http://teknosi.fti.unand.ac.id/>

## Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |



Studi Kasus

# Segmentasi Pelanggan menggunakan Metode *K-means Clustering* Berdasarkan Model RFM pada Bisnis *Food and Beverage* (Studi Kasus: *Coffee shop X*)

Anggita Nurlathifa<sup>a,\*</sup>, Diarany Sucahyati<sup>b</sup>

<sup>a</sup>Akuntansi Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294, Indonesia

<sup>b</sup>Akuntansi Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294, Indonesia

### INFORMASI ARTIKEL

Sejarah Artikel:

Diterima Redaksi: 13 Juli 2025

Revisi Akhir: 1 September 2025

Diterbitkan Online: 28 Desember 2025

### KATA KUNCI

*Big Data,*  
*K-means,*  
*Clustering,*  
*Bisnis,*  
*RFM*

### KORESPONDENSI

E-mail: 21013010272@student.upnjatim.ac.id\*

### ABSTRACT

Data tidak bisa dipisahkan dari bidang industri yang semakin maju. Kemajuan ini diikuti dengan banyaknya usaha yang tumbuh yang membuat persaingan usaha juga semakin ketat. Salah satu cara untuk sebuah usaha dapat bersaing dengan baik adalah dengan melakukan analisis data sebagai salah satu cara membuat keputusan yang baik sehingga usaha dapat berkembang dengan baik. Beberapa analisis yang bisa dilakukan adalah dengan memahami segmentasi pelanggan dan analisis transaksi sebagai upaya mengoptimalkan pemasaran, produk, dan operasi. Penelitian ini mengklasifikasikan pelanggan *Coffee Shop X* menggunakan segmentasi RFM (*Recency, Frequency, Monetary*) dengan algoritma *K-means*. *K-means* digunakan untuk mengelompokkan pelanggan berdasarkan skor RFM dengan *Google Colab*. Digunakan juga metode *Elbow* dan validasi *cluster* (*Silhouette Width, Dunn index, dan Connectivity*) untuk menentukan jumlah *cluster* yang optimal. Hasil menunjukkan bahwa segmentasi optimal terdiri dari 2 atau 3 cluster. Pelanggan terbagi menjadi golongan pasif, aktif, dan loyal, masing-masing dengan strategi pemasaran yang berbeda seperti edukasi produk, kampanye promosi, dan program loyalitas.

## 1. PENDAHULUAN

Industri *Food and Beverage* merupakan sektor industri yang selalu mengalami peningkatan dan memiliki peran penting dalam pertumbuhan ekonomi. Menurut Kementerian Keuangan hal ini didukung dengan perubahan gaya hidup di pusat-pusat perkotaan Indonesia yang sebagian besar mengikuti tren pasar yang sudah mapan, dengan pekerja kantoran memiliki lebih sedikit waktu untuk memasak, atau kurang berminat untuk melakukannya, namun menuntut makanan yang meningkatkan kesehatan [1]. Ini dibuktikan dengan menjamurnya bisnis yang bergerak di sektor *food and beverage*. Banyaknya bisnis pada bidang yang sama ini membuat persaingan sangat ketat. Oleh karena itu, diperlukan

strategi penjualan yang sesuai agar bisnis tersebut bisa terus berkembang dengan baik mengikuti pertumbuhan persaingan antar bisnis yang selalu meningkat. Salah satu metode yang patut digunakan, terutama dengan dukungan era digital saat ini adalah menggunakan analisis *big data*.

Dalam era digital yang berkembang pesat, optimalisasi *big data* telah mulai diaplikasikan di berbagai lini kehidupan masyarakat, terlebih lagi aspek vital seperti perusahaan di berbagai industri. Pemanfaatan *Big Data* terbukti memberikan kontribusi positif bagi kinerja perusahaan, antara lain melalui kemampuan analitiknya dalam mengidentifikasi tren pasar, memahami perilaku konsumen, serta menemukan peluang pengembangan

produk baru dengan lebih terarah dan efektif [2]. Terkait hal ini muncul pertanyaan, bagaimana perusahaan di berbagai industri memanfaatkan *big data* untuk meningkatkan pengambilan keputusan bisnis? *Big data* pada dasarnya berarti kumpulan data yang terlalu besar untuk sistem pemrosesan data tradisional, dan oleh karena itu memerlukan teknologi pemrosesan baru [3]. Dengan adanya integrasi data menggunakan *cloud server* secara digital sebagai dampak dari era revolusi industri 5.0, perusahaan kini dapat mengakses data berukuran besar, memudahkan perusahaan untuk mengambil *insight* yang lebih luas dalam manajemennya. Data kini tidak hanya membantu perusahaan memahami tren dan pola perilaku konsumen, tetapi juga memberikan wawasan yang mendalam untuk pengambilan keputusan bisnis yang lebih baik. Namun, pengambilan *insight* dari *big data* tidak sepenuhnya tugas yang mudah. Terlebih lagi dikarenakan kapasitasnya yang memerlukan teknologi pemrosesan baru. Kebutuhan akan pengolahan informasi yang akurat sangat dibutuhkan oleh organisasi bisnis, untuk dapat memberikan arahan yang jelas dalam proses bisnis [4]. Oleh karena itu diperlukan metode untuk mengambil *insight* yang tepat dan efektif dari data, atau yang lebih sering dikenal dengan istilah analisis data.

Segmentasi pelanggan adalah salah satu aspek yang perlu diperhatikan untuk menentukan kebijakan terutama dalam bidang penjualan. Untuk melakukan segmentasi pelanggan ini dapat digunakan konsep penggabungan tiga kriteria yaitu *Recency*, *Frequency*, *Monetary* (RFM) yang diperkenalkan oleh (Bult & Wansbeek, 1995 dalam SPM. Ernawati et al. (2021) menyatakan bahwa model *Recency*, *Frequency*, *Monetary* (RFM) merupakan pendekatan segmentasi perilaku pelanggan yang paling banyak digunakan karena sederhana, mudah diterapkan, dan efektif menjelaskan perilaku pembelian berdasarkan data transaksi [6]. Penggunaan ketiga kriteria pemilihan ini dengan tepat membantu pelaku usaha fokus pada pola pembelian yang penting [7]. Segmentasi pelanggan ini dapat dilakukan melalui berbagai cara salah satunya adalah *k-means*. *K-means clustering* merupakan salah satu algoritma *clustering* paling klasik dan biasa disebut sebagai unsupervised learning yang digunakan untuk mengelompokkan data ke dalam sejumlah *cluster* atau kelas berdasarkan kemiripan. Algoritma ini memerlukan pengetahuan sebelumnya tentang jumlah *cluster* (*k*) [8]. Keunggulan *k-means* adalah sederhana, efisien, dan implementasinya yang mudah [9] *K-Means clustering* memungkinkan segmentasi pelanggan secara cerdas berdasarkan pola perilaku pembelian, sehingga memudahkan perusahaan dalam merumuskan strategi yang tepat untuk setiap segmen pasar [10].

Penelitian terdahulu oleh Savitri et al. (2018) telah mengaplikasikan metode *K-Means* dengan model RFM (*Recency*, *Frequency*, *Monetary*) untuk segmentasi pelanggan pada sektor klinik kecantikan, namun penerapannya masih terbatas pada pengelompokan pelanggan tanpa memberikan rekomendasi strategi pemasaran yang spesifik untuk tiap *cluster*. Selain itu, fokus penelitian yang dominan pada sektor klinik kecantikan membuat generalisasi hasil ke sektor lain dengan karakteristik perilaku konsumen yang berbeda masih rendah [11]. Oleh karena itu, penelitian ini mengisi kesenjangan tersebut dengan menerapkan analisis pada sektor *food and beverage* menambahkan atribut relevan di luar RFM, serta

mengintegrasikan hasil segmentasi dengan rekomendasi strategi pemasaran yang tepat sasaran bagi setiap *cluster* pelanggan.

Model RFM (*Recency*, *Frequency*, *Monetary*) adalah pendekatan yang digunakan untuk menganalisis dan mengelompokkan pelanggan berdasarkan perilaku transaksi mereka. Model ini terdiri dari tiga variabel utama:

1. *Recency* (R): jarak waktu sejak transaksi terakhir pelanggan. Semakin kecil nilai *Recency*, semakin aktif pelanggan tersebut.
2. *Frequency* (F): jumlah transaksi yang dilakukan pelanggan dalam periode tertentu. Semakin sering, maka semakin tinggi nilainya.
3. *Monetary* (M): total nilai uang yang dibelanjakan oleh pelanggan. Semakin besar nilainya, semakin bernilai pelanggan tersebut bagi perusahaan.

Sebelum dilakukan proses *clustering*, nilai-nilai RFM biasanya dinormalisasi agar berada pada skala yang sebanding. Salah satu metode yang umum digunakan adalah normalisasi *Min-Max* [12], dengan rumus:

$$x' = \frac{x - \min_a}{\max_a - \min_a} (\text{newmax} - \text{newmin}) + \text{newmin} \dots (1)$$

Keterangan:

- $x'$  = nilai setelah dinormalisasi
- $x$  = nilai asli
- $\min_a$  dan  $\max_a$  = nilai minimum dan maksimum dari variabel
- $\text{newmax}$  dan  $\text{newmin}$  = batas bawah dan atas skala baru, biasanya 0 dan 1

Dalam uji performa membutuhkan tiga metode dalam package *CValid()* yakni *Dunn index*, *Silhouette width*, dan *Connectivity*. *Dunn index* merupakan metrik untuk memvalidasi hasil pengelompokkan dengan mengukur diameter *cluster* dan jarak antara dua *cluster* [11]. *Dunn index* merupakan rasio antara jarak minimum antara objek data dalam *cluster* yang berbeda ( $d_{\min}$ ) dan jarak terbesar dalam *cluster* ( $d_{\max}$ ) [8]. Semakin tinggi nilai *Dunn index* semakin baik pengelompokkannya [13]. Nilai *Dunn index* menurut [14], dapat diperoleh pada persamaan berikut:

$$D_C = \min_{C_k, C_l \in C, C_k \neq C_l} \frac{\left( \min_{i \in C_k, l \in C_l} \text{dist}(i, l) \right)}{\max_{C_m \in C} d(C_m)} \dots (2)$$

*Silhouette width* merupakan ukuran yang digunakan untuk mengevaluasi kualitas pengelompokkan, di mana nilai yang lebih tinggi menunjukkan bahwa objek berada dalam kluster yang tepat dan terpisah dengan baik dari kluster lain [15]. *Silhouette width* adalah metrik yang sangat berguna untuk mengevaluasi kinerja hasil pengelompokkan. Nilai ini mengukur bagaimana titik data dikelompokkan dan menghitung jarak rata-rata yang tersedia antara *cluster* yang berbeda. Lebar koefisien ini selalu terletak pada interval berikut  $[-1, 1]$  yang menyiratkan titik data supergrup dengan nilai mendekati 1 dan titik data grup bawah dengan nilai mendekati -1. Oleh karena itu, koefisien untuk titik data  $i$ , didefinisikan dari persamaan berikut:

$$S(i) = \frac{(y_i - x_i)}{\max(y_i, x_i)} \dots (3)$$

Di mana  $x$  dan  $y$  menunjukkan jarak rata-rata antara titik data dalam *cluster* yang sama dan jarak rata-rata antara titik data dalam *cluster* tetangga terdekat yang dapat dinyatakan sebagai berikut:

$$y_i = \min_{C_k \in C_i} \sum_{j \in C_k} \frac{\text{dist}(i, j)}{nC_k} \dots (4)$$

Di mana  $C$  menunjukkan *cluster* dengan titik data  $i$ ,  $\text{dist}(i, j)$ , menunjukkan jarak antara titik data  $i$  dan  $j$ , kemudian  $nC_k$  menunjukkan kardinalitas *cluster*  $C$ .

*Connectivity* menunjukkan jumlah total baris  $n$  (titik data atau observasi) dan kolom  $m$  dalam satu set data. Nilai-nilai tersebut selalu dianggap sebagai numerik (misalnya, nilai parameter fisiologis). Misalkan  $Y_{ni}(j)$  dan  $x_i Y_{ni}(j)$  masing-masing adalah tetangga terdekat  $j^{\text{th}}$  dari titik data  $i$  dan nol, jika  $i$  dan  $j$  berada dalam *cluster* yang sama, dan kemudian  $i/j$  jika tidak. *Connectivity* diukur untuk *cluster* tertentu  $C = \{C_1, C_2, \dots, C_k\}$  dengan titik data  $n$  menggunakan persamaan menurut [14]

Gambar 1.

### 3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data transaksi pelanggan pada sebuah *coffee shop* selama periode tiga bulan terakhir di tahun 2019. Data diperoleh dari sistem pencatatan internal toko dan terdiri dari atribut-atribut seperti transaction id, transaction date, customer id, gender, birthdate, generation, brought product, price, dan store. Data tersebut mencerminkan perilaku belanja pelanggan dan menjadi dasar

$$C = \sum_{i=1}^n \sum_{j=1}^p x_i Y_{ni}(j) \dots (5)$$

Data yang digunakan pada penelitian ini adalah data *dummy* yang diperoleh dari pihak akselerator sebagai fasilitas pelatihan analisis data. Data ini berisikan informasi tentang penjualan di semua *outlet* sebuah kedai kopi pada bulan April tahun 2019. Penelitian ini diharapkan dapat memberikan gambaran untuk pelaku bisnis dan data analis untuk menemukan cara menggunakan datanya sebagai sarana pengambilan keputusan untuk keberlangsungan bisnis kedai kopi yang di era sekarang sudah sangat menjamur.

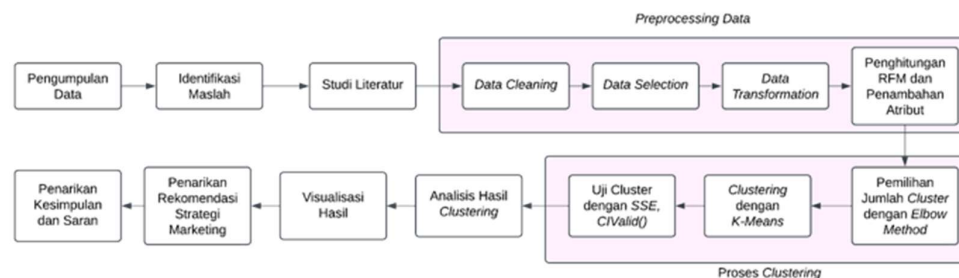
## 2. METODE

3. Alur penelitian atau metodologi dari penelitian yang dilakukan dapat dilihat pada

dalam membangun model segmentasi berbasis RFM (*Recency, Frequency, Monetary*).

### 3.2. Identifikasi Masalah

Masalah yang terjadi adalah untuk memahami karakteristik pelanggan *coffee shop* dengan melakukan segmentasi pelanggan berdasarkan perilaku pelanggan. Oleh karena itu, permasalahan utama dalam penelitian ini adalah bagaimana mengelompokkan pelanggan dengan optimal menggunakan pendekatan RFM dan metode *clustering K-means* agar didapatkan segmentasi yang relevan untuk penerapan strategi bisnis.



Gambar 1. Alur Penelitian Segmentasi Pelanggan dengan *K-means*

### 3.3. Studi Literatur

Mempelajari teori-teori yang mendukung metode segmentasi pelanggan menggunakan pendekatan RFM, metode *clustering* yang digunakan (*K-means*), yang dikenal sebagai salah satu teknik unsupervised learning paling populer untuk segmentasi karena kemampuannya mengelompokkan data berdasarkan kesamaan fitur. Selain itu, dipelajari juga metode *Elbow* dan validasi *cluster* (*Silhouette Width*, *Dunn index*, dan *Connectivity*) untuk menentukan jumlah *cluster* yang optimal.

### 3.4. Preprocessing Data

Dilakukan tahap preprocessing data untuk memastikan kualitas data menggunakan *Google Colab*, yakni penanganan *missing value* dengan mengisi *Recency* dengan nilai nol untuk pelanggan yang tidak ada transaksi ulang, pembersihan data dengan penghapusan nilai kosong dan konversi format tanggal untuk menghitung jarak *Recency*, dan pembuatan variabel RFM.

### 3.5. Proses Clustering

Metode *K-means* digunakan untuk mengelompokkan pelanggan berdasarkan skor RFM dengan *Google Colab*. Penentuan jumlah *cluster* ( $k$ ) dilakukan dengan menggunakan *Elbow Method* untuk melihat titik optimal dari penurunan nilai *Sum of Squared Errors* (*SSE*). Setelah diperoleh nilai  $k$  optimal, proses *clustering* dilakukan, dan masing-masing pelanggan diberikan label *cluster* sesuai dengan hasil pengelompokannya. Validasi model *clustering* dilakukan menggunakan metrik *Silhouette Width*, *Dunn index*, dan *Connectivity*.

### 3.6. Visualisasi Hasil

Hasil segmentasi divisualisasikan menggunakan dengan pustaka *matplotlib* pada *Google Colab*, dalam bentuk grafik yang terdiri dari jumlah pelanggan per *cluster*, distribusi gender, generasi, dan lama keanggotaan di tiap *cluster*, serta rata-rata nilai RFM di tiap *cluster* untuk membantu memahami karakteristik

pelanggan di setiap *cluster* dan menjadi dasar menyusun strategi pemasaran yang tepat sasaran.

#### 4. HASIL

Data yang didapatkan berisi data pelanggan dalam waktu 3 bulan terakhir di tahun 2019. Data terdiri dari beberapa atribut, yakni *transaction id*, *transaction date*, *customer id*, *gender*, *birthdate*, *generation*, *brought product*, *price*, *store*. Dari data tersebut, kemudian diolah untuk mendapatkan insight nilai RFM pada sub-bab selanjutnya.

##### 4.1. Analisis Recency, Frequency, Monetary (RFM)

*Recency* merupakan rentang waktu sejak transaksi terakhir dihitung dari perbedaan tanggal transaksi terakhir dan sebelumnya, lalu disesuaikan dengan nilai referensi (29 hari) yang ada pada data. Dari analisis ini, pelanggan yang sering kembali (jarak antara hari melakukan antar transaksi singkat) memiliki nilai *Recency* yang lebih tinggi. Cuplikan hasil dari analisis *Recency* pada penelitian ini ditunjukkan pada Tabel 1. Tabel *Recency* Pelanggan.

Tabel 1. Tabel *Recency* Pelanggan

<i>Custo mer ID</i>	<i>Second Last Transactio n Date</i>	<i>Last Transaction Date</i>	<i>Days Differen ce</i>	<i>Recency</i>
1	2019-04-29	2019-04-29	0	29
2	2019-04-25	2019-04-26	1	28
3	2019-04-23	2019-04-24	1	28
...	...	...	...	...
8499	2019-04-23	2019-04-29	6	23
8500	2019-04-21	2019-04-26	5	24
8501	2019-04-27	2019-04-28	1	28

Pada Tabel 1. Tabel *Recency* Pelanggan, terdapat beberapa atribut yakni *Customer ID* yang menandai tiap individu pelanggan yang berbeda, *Second Last Transaction Date* yang menunjukkan tanggal transaksi kedua dari terakhir yang dilakukan tiap pelanggan, *Last Transaction Date* yang menunjukkan tanggal transaksi terakhir yang dilakukan, serta *Recency* yang dihitung dari rumus berikut:

$$\text{Recency} = \text{Nilai Referensi} - \Delta \text{ transaction date}$$

Dengan  $\Delta \text{ transaction date}$  adalah jarak (hari) dari transaksi terakhir dan transaksi terakhir-1. Sementara itu, Tabel 2. Tabel *Frequency* Pelanggan memiliki atribut *Customer ID* dan *Frequency*, yang diambil dari berapa kali pelanggan melakukan transaksi di toko. Tabel 3. Tabel *Monetary* Pelanggan berisi data ID pelanggan dan *Monetary*, atau berapa banyak uang yang dihabiskan satu pelanggan di toko.

Tabel 2. Tabel *Frequency* Pelanggan

<i>Customer ID</i>	<i>Frequency</i>
1	8
2	21

3	31
...	...
8499	12
8500	21
8501	6

Tabel 3. Tabel *Monetary* Pelanggan

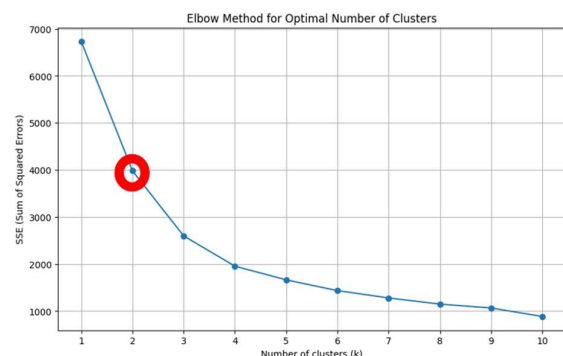
<i>Customer ID</i>	<i>Monetary</i>
1	29.20
2	90.35
3	166.40
...	...
8499	58.10
8500	90.85
8501	28.75

##### 4.2. Data Pre-processing

Dilakukan proses pengecekan nan value jika ada pelanggan yang tidak kembali ke *coffee shop* dalam satu rentang waktu tersebut. Dari sana, ditemukan 7 pelanggan tidak kembali sehingga dilakukan *fillna()* dengan nilai 0 atau pelanggan tidak kembali, karena semakin kecil rentang kembali, semakin besar nilai *Recency*. Selain itu, dilakukan data *clean up* seperti pengisian *missing value* sesuai dengan konteks yang diberikan, serta pengubahan format data.

##### 4.3. Penentuan Jumlah Cluster (k) menggunakan Elbow Method

Penentuan jumlah *cluster* optimal dilakukan menggunakan *Elbow Method* dengan mengamati nilai SSE (Sum of Squared Errors) untuk berbagai nilai *k* dari 1 hingga 9. Berdasarkan Gambar 2. Grafik Nilai *Elbow Method*, Grafik Nilai *Elbow Method* dan Tabel 6. Perhitungan SSE dan Selisih untuk tiap *Cluster*, penurunan nilai SSE paling signifikan terjadi saat *k*=2 (dari 6735.00 ke 3981.45) dan *k*=3 (ke 2598.52). Setelah *k*=3, penurunan nilai SSE semakin kecil. Oleh karena itu, kandidat terbaik untuk jumlah *cluster* adalah *k*=2 dan *k*=3, dikarenakan penurunan SSE yang tidak signifikan setelah *k*=3.



Gambar 2. Grafik Nilai *Elbow Method*

#### 4.4. Hasil Clustering Model K-means

Dari Sub Bab 3.3, didapatkan *cluster* paling ideal untuk dianalisis adalah k=2 dan k=3. Masing hasilnya dijabarkan sebagai berikut:

##### 4.4.1. Hasil Clustering K-means dengan k=2

Dari hasil *clustering* dengan k=2, diperoleh dua *cluster* yang masing-masing memiliki karakteristik berbeda. Tabel 4. Cuplikan Hasil *Clustering* dengan k=2 menunjukkan cuplikan tampilan hasil *clustering* yang kemudian akan ditelaah pada bab selanjutnya. Nilai tiap kolom adalah ID Pelanggan yang membedakan satu pelanggan dengan pelanggan lain, nilai RFM, dan juga pengclusterannya setelah dilakukan *clustering* dengan model *K-means*, di mana hasilnya terdapat 2 *cluster*, 0 dan 1.

Tabel 4. Cuplikan Hasil *Clustering* dengan k=2

ID	Recency	Frequency	Monetary	Cluster
0	29	8	29.0	0
1	28	21	90.35	1
2	28	31	166.40	1
...	...	...	...	...
2240	19	12	66.00	0
2241	24	9	39.75	0
2242	23	12	58.10	0
2243	24	21	90.85	1
2244	28	6	28.75	0

##### 4.4.2. Hasil Clustering K-means dengan k=3

Hasil pada Tabel 5 yang ditampilkan, memiliki bentuk yang sama dengan Tabel 4, termasuk ID pelanggan dan nilai RFM untuk tiap pelanggannya. Yang membedakan adalah di bagian *cluster*, di mana pada k=3, nilai *cluster*nya terbagi menjadi 0,1, dan 2.

Tabel 5. Cuplikan Hasil *Clustering* dengan k=3

ID	Recency	Frequency	Monetary	Cluster
0	29	8	29.0	1
1	28	21	90.35	2
2	28	31	166.40	2
...	...	...	...	...
2240	19	12	66.00	0
2241	24	9	39.75	1
2242	23	12	58.10	1
2243	24	21	90.85	2
2244	28	6	28.75	1

#### 4.5. Uji Performa Model

##### 4.5.1. SSE

Tabel 6 menunjukkan bahwa SSE terus menurun seiring bertambahnya jumlah *cluster*. Namun, penurunan terbesar terjadi pada k=2 dan k=3. Setelah k=3, penurunan SSE semakin kecil, yang berarti pembagian *cluster* setelah nilai tersebut memberikan peningkatan yang tidak signifikan sehingga tidak layak untuk digunakan sebagai pemilihan hyperparameter analisis.

Tabel 6. Perhitungan SSE dan Selisih untuk tiap *Cluster*

Cluster	SSE	Selisih
1	6735.00	-
2	3981.45	2753.55
3	2598.52	1382.93
4	1955.01	643.51
5	1663.99	291.02
6	1435.60	228.39
7	1278.22	157.38
8	1148.92	129.3
9	1066.12	82.8

##### 4.5.2. CValid()

Validasi model dilakukan menggunakan beberapa metrik: *Silhouette*, *Dunn index*, dan *Connectivity*. Hasil dari pengukuran metrik ini ditunjukkan pada Tabel 7. Dari penyajian data tersebut, dapat dilihat bahwa *Silhouette* Score tertinggi ada pada k=3 (0.425) yang menunjukkan pemisahan *cluster* yang baik. Sementara itu, *Dun* Index tertinggi pada k=2 (1.615). Hal ini menunjukkan bahwa jarak antar *cluster* optimal untuk k=2. Kemudian, nilai *Connectivity* yang paling tinggi ada pada k=3 (891.367), dengan diikuti dengan nilai *Silhouette* dan *Dunn index* yang baik. Hal ini menunjukkan koneksi antar objek dalam *cluster* yang terbaik.

Tabel 7. Perhitungan Validasi Menggunakan CValid()

Cluster	Validation Methods		
	Silhouutte	Dunn index	Connectivity
2	0.385	1.615	1190.412
3	0.425	1.587	891.367
4	0.373	1.157	734.524
5	0.338	0.976	640.305
6	0.342	0.993	531.579
7	0.349	0.967	479.393
8	0.329	0.769	472.284
9	0.332	0.704	447.423

Dengan demikian, baik k=2 maupun k=3 memiliki kelebihan dan kekurangan masing-masing yang tergantung daripada tujuan segmentasi yang diinginkan. Untuk segmentasi yang lebih



sederhana dan jelas,  $k=2$  cukup efektif, sedangkan  $k=3$  memberikan segmentasi yang lebih rinci dan mendalam.

## 5. PEMBAHASAN

### 5.1. Clustering K-means dengan $k=2$

Dari hasil rata-rata yang dirangkum dalam Tabel 8, didapatkan bahwa *cluster* 0 memiliki rata-rata *Recency* lebih rendah (24.09), frekuensi transaksi yang lebih sedikit (8.03), dan pengeluaran (*Monetary*) lebih rendah (35.93). Sementara itu, *cluster* 1 memiliki nilai rata-rata *Recency* lebih tinggi (26.93), frekuensi transaksi lebih banyak (15.74), dan pengeluaran yang lebih besar (74.44). Dalam hal ini, *cluster* 0 berisi pelanggan dengan karakteristik belanja yang lebih pasif, bertransaksi lebih jarang dan dengan pengeluaran yang lebih rendah. Sementara itu, *cluster* 1 terdiri dari pelanggan yang lebih aktif dan loyal, dengan frekuensi kunjungan dan pengeluaran yang tinggi. Tabel penyajian *clustering* untuk hasil  $k=2$  ditunjukkan pada Tabel 8 Tabel Rata-rata RFM untuk *Cluster*=2.

Tabel 8 Tabel Rata-rata RFM untuk *Cluster*=2

<i>Cluster</i>	<i>Recency</i>	<i>Frequency</i>	<i>Monetary</i>
0	24.096253	8.036003	35.926076
1	26.936652	15.740950	74.438314

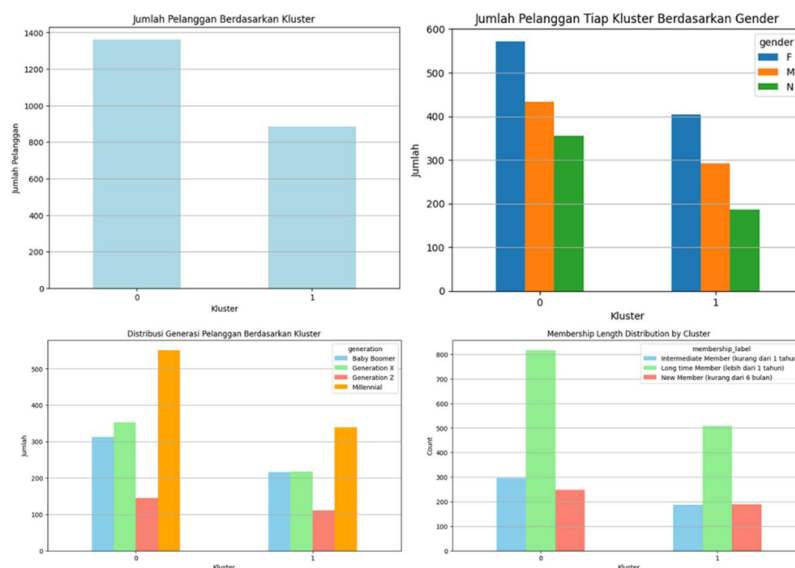
Setelah melihat dari statistik deskripsi yakni nilai rata-rata dari hasil *clustering*, dilakukan analisis yang lebih mendalam pada Gambar 3. Didapatkan visualisasi karakteristik dari *cluster*=2. Di dalamnya terdapat empat grafik yang memberikan informasi tambahan mengenai karakteristik demografis dan status keanggotaan pelanggan pada masing-masing *cluster*. Dari grafik Jumlah Pelanggan Berdasarkan *Cluster*, ditunjukkan bahwa

bahwa *Cluster* 0 memiliki jumlah pelanggan lebih banyak dibandingkan *Cluster* 1. Informasi ini menunjukkan bahwa mayoritas pelanggan dikelompokkan ke dalam *Cluster* 0.

Pada grafik Jumlah Pelanggan Tiap *Cluster* berdasarkan Gender, dibagi gender pelanggan berdasarkan Female (F) untuk perempuan dan Male (M) untuk laki-laki, serta Not Specified (N) untuk yang tidak dispesifikasi. Dapat dilihat bahwa *Cluster* 0 memiliki dominasi pelanggan perempuan (F), disusul laki-laki (M), dan sisanya tidak menyebutkan gender (N). Sementara itu, dengan jumlah total yang lebih sedikit, *Cluster* 1 menunjukkan pola yang serupa. Kemudian, Grafik Distribusi Generasi Pelanggan menunjukkan bahwa pada *Cluster* 0, generasi milenial mendominasi, diikuti Gen X dan Gen Z. Sementara itu, pada *Cluster* 1, proporsinya lebih seimbang dibandingkan *Cluster* 0. Hal ini menunjukkan mayoritas pelanggan berasal dari generasi muda hingga paruh baya. Terakhir, Grafik Lama Keanggotaan menunjukkan *Cluster* 0 didominasi oleh Intermediate Member (anggota kurang dari 1 tahun), dengan jumlah signifikan pada New Members (kurang dari 6 bulan). Dibandingkan dengan *Cluster* 1, yang lebih banyak diisi Intermediate Member dan sebagian kecil *Long time Member* (lebih dari 1 tahun). Pola ini mengindikasikan banyak pelanggan baru yang ada di *Cluster* 0, yang masih menjajaki atau mencoba layanan.

Dari Tabel 8 dan

Gambar 3, perlu dilakukan strategi fokus pada edukasi produk, diskon frekuensi pembelian, dan reminder agar mereka naik kelas menjadi pelanggan aktif dengan penerapan poin loyalitas untuk tiap pembelian pada *Cluster* 0 untuk meningkatkan aktivitasnya. Sementara itu, pada *Cluster* 1, dapat diterapkan strategi loyalitas dengan program eksklusif serta reward poin tinggi. Dapat juga diberikan benefit tambahan seperti akses spesial *early promo* atau *event member*. Pelanggan ini memiliki nilai tinggi pada bisnis sehingga perlu didahulukan feedback-feedbacknya.



Gambar 3. Grafik Segmentasi Pelanggan dengan 2 *cluster* untuk beberapa atribut.

Dari

Gambar 3 dapat disimpulkan bahwa *Cluster* 0 adalah kelompok pelanggan yang jumlahnya lebih besar, didominasi oleh

perempuan, generasi muda (terutama Milenial), dan pelanggan yang tergolong baru. Di sisi lain, *Cluster* 1, meskipun jumlahnya lebih sedikit, memiliki proporsi pelanggan yang lebih loyal atau

lebih lama bergabung, serta profil demografis yang relatif serupa namun sedikit lebih stabil. Kondisi ini memperkuat analisis sebelumnya bahwa *Cluster 0* cenderung terdiri dari pelanggan dengan frekuensi dan pengeluaran yang rendah, sementara *Cluster 1* adalah pelanggan aktif dan loyal. Oleh karena itu, strategi yang berbeda perlu diterapkan pada masing-masing *cluster* untuk mengoptimalkan retensi dan pertumbuhan pelanggan, seperti yang telah disampaikan sebelumnya.

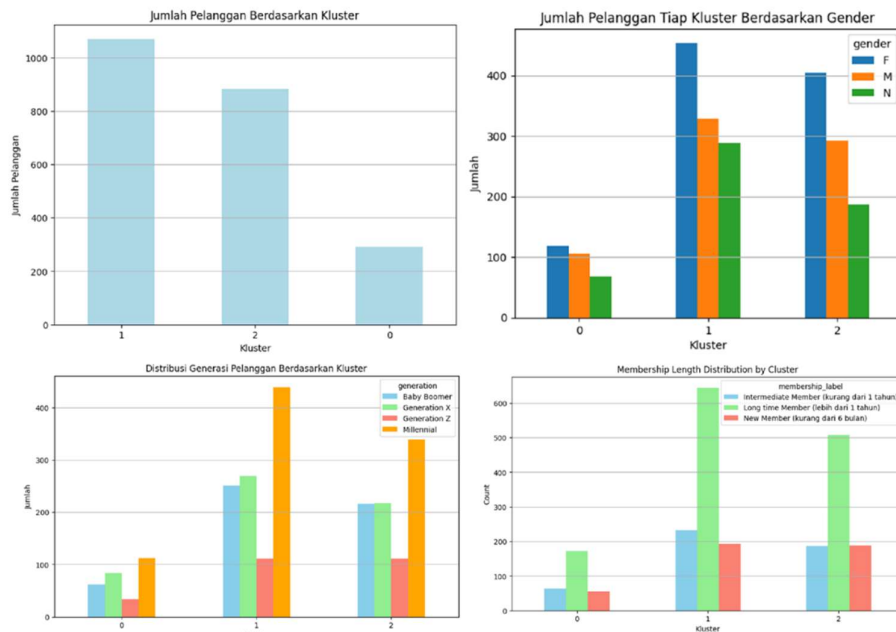
## 5.2. Clustering K-means dengan $k=3$

Sementara itu, pada clustering dengan  $k=3$ , diperoleh tiga *cluster* dengan karakteristik rata-rata yang ditunjukkan pada Tabel 9. *Cluster 0* memiliki *Recency* paling rendah (14.75), frekuensi rendah (7.47), dan pengeluaran rendah (35.93). Sementara itu, *cluster 1* memiliki *Recency* sedang (26.64), frekuensi sedikit lebih tinggi (8.19), dan *Monetary* menengah (36.62). Terakhir, *cluster 2* memiliki *Recency* tinggi (26.95), frekuensi tinggi (15.74), dan *Monetary* tertinggi (74.44).

Tabel 9. Tabel Rata-rata RFM untuk *Cluster=3*

Cluster	Recency	Frequency	Monetary
0	14.746575	7.472603	35.926076
1	26.639252	8.194393	36.62267
2	26.950170	15.744054	74.43954

Pada Gambar 4, didapatkan visualisasi karakteristik dari *cluster=3*, yang terdiri dari distribusi pelanggan berdasarkan jumlah, gender, generasi, dan lama keanggotaan, sama seperti sebelumnya pada *Cluster=2*. Dari grafik Jumlah Pelanggan, ditunjukkan bahwa *Cluster 1* memiliki jumlah pelanggan terbanyak, diikuti dengan *Cluster 2*, dan jumlah paling sedikit yang jauh di *Cluster 3*. Hal ini menunjukkan bahwa sebagian besar pelanggan ada pada dua *cluster* utama, yakni *Cluster 1* dan 2. Sementara itu, *Cluster 0* menjadi kelompok yang eksklusif atau minoritas dalam *clustering* ini. Kemudian, dari Grafik Jumlah Pelanggan Tiap *Cluster* berdasarkan Gender, ditemukan bahwa setiap *cluster* memiliki dominasi oleh perempuan (F), diikuti dengan laki-laki (M), dan netral (N). Grafik Distribusi Generasi Pelanggan menunjukkan bahwa *Cluster 1* dan 2 didominasi oleh generasi milenial dan gen X, sementara *Cluster 0* memiliki jumlah yang kecil di semua generasi, meskipun ada dominasi di milenial. Pada grafik selanjutnya, yakni distribusi lama keanggotaan, di semua *cluster*, intermediate member (anggota kurang dari 1 tahun) mendominasi, dengan pola serupa terutama pada *Cluster 0* dan 2.



Gambar 4. Grafik Segmentasi Pelanggan dengan 3 *cluster* untuk beberapa atribut

Dari informasi pada Gambar 4 dan Tabel 9, ditemukan bahwa *Cluster 0* adalah kelompok kecil dengan karakteristik yang khas, dimana kemungkinan pelanggan dengan pengeluaran dan frekuensi rendah, dengan *Recency* yang juga rendah (menunjukkan kurang kembali). Pelanggan pada *cluster* ini dapat menjadi target kampanye promosi atau reaktivasi, serta edukasi produk. Sementara itu, *Cluster 1* berisi pelanggan terbanyak dan

mewakili pelanggan aktif atau reguler, merupakan pelanggan dengan aktivitas sedang, tidak terlalu sering tapi tidak terlalu pasif. Mereka memiliki potensi untuk ditingkatkan menjadi pelanggan loyal. Secara jumlah, *cluster* ini paling banyak anggotanya, sehingga strategi kampanye dan umpan-balik paling cocok dilakukan pada *cluster* ini. Selain itu, strategi upsell dan reminder cocok diterapkan pada *Cluster* ini. *Cluster 2*, adalah

pelanggan paling loyal dan berharga, karena sering, rutin, dan banyak berbelanja. Pelanggan dengan *cluster* ini cocok diberi program loyalitas atau reward eksklusif, serta mengikuti trend yang lebih cocok untuk Generasi Milenial seperti yang ditunjukkan pada Gambar 3.

## 6. KESIMPULAN

Dari hasil analisis segmentasi pelanggan berdasarkan nilai RFM pada *Coffee shop X* menggunakan *K-means* didapatkan beberapa kesimpulan, yaitu yang pertama adalah nilai *Recency*, *Frequency*, dan *Monetary* (RFM) tiap pelanggan di *Coffee shop X* bisa diklasifikasikan dengan baik menggunakan *K-means* dengan *cluster* terbaik 2 atau 3. Dari hasil *clustering* tersebut, didapatkan segmentasi pelanggan dengan nilai evaluasi terbaik dan mampu mewakili keseluruhan pelanggan.

Selanjutnya, dari *clustering* dengan  $k=2$ , didapatkan segmentasi pelanggan sejumlah 2. Golongan pertama adalah golongan dengan karakteristik belanja yang lebih pasif, bertransaksi lebih jarang dan dengan pengeluaran yang lebih rendah (*new-comer*), dan golongan pelanggan yang lebih aktif dan loyal, dengan frekuensi kunjungan dan pengeluaran yang tinggi (*loyal-customer*). Strategi marketing kampanye bisa diterapkan pada golongan pertama, sementara strategi loyalitas dapat diterapkan pada golongan kedua. Kedua golongan tersebut cenderung diisi oleh jenis pelanggan yang serupa, yakni perempuan generasi milenial sehingga strategi bisnis dapat menyasar ke dua sasaran tersebut.

Yang terakhir, dari *clustering* dengan  $k=3$ , ditemukan hasil segmentasi menjadi 3 golongan, yakni golongan pertama dengan karakteristik RFM cenderung rendah, yang cocok ditangani dengan strategi marketing edukasi produk. Selain itu, terdapat golongan kedua yang merepresentasikan pelanggan aktif dengan aktivitas sedang dan merupakan mayoritas jumlah pelanggan. Golongan ini cocok dengan strategi marketing kampanye dan feed-back atau umpan balik. Terakhir, golongan ketiga yakni pelanggan loyal yang dapat terus dipertahankan dengan menjaga kualitas dan program reward loyalitas/membership. Pengisi ketiga golongan ini mayoritas sama dengan *clustering*  $k=2$  yakni generasi milenial perempuan sehingga target strategi marketing bisa mempertimbangkan dua aspek tersebut.

## DAFTAR PUSTAKA

- [1] Kementerian Keuangan, "Kondisi Industri Pengolahan Makanan dan Minuman." Accessed: Jul. 05, 2025. [Online]. Available: <https://www.djkn.kemenkeu.go.id/kanwil-suluttenggomallut/baca-artikel/15588/Kondisi-Industri-Pengolahan-Makanan-dan-Minuman-di-Indonesia.html>
- [2] Muh. M. Sari and R. P. Avianto, "Big Data dalam Bisnis: Studi Literatur dan Penerapannya di Indonesia," *Jurnal Inovasi Informatika*, vol. 7, no. 2, pp. 25–36, 2025.
- [3] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013. [Online]. Available: <https://books.google.co.id/books?id=EZAtAAAAQBAJ>
- [4] E. S. Mulyanta and R. G. Andi, *How to Make Money in Big Data: Implementasi Big Data untuk UMKM*. Penerbit Andi, 2022. [Online]. Available: <https://books.google.co.id/books?id=OAt0EAAAQBAJ>
- [5] J. R. Bult and T. Wansbeek, "Optimal Selection for Direct Mail," *Marketing Science*, vol. 14, no. 4, pp. 378–394, Nov. 1995, doi: [10.1287/mksc.14.4.378](https://doi.org/10.1287/mksc.14.4.378).
- [6] E. Ernawati, S. S. K. Baharin, and F. Kasmin, "A review of data mining methods in RFM-based customer segmentation," *J Phys Conf Ser*, vol. 1869, no. 1, p. 012085, Apr. 2021, doi: [10.1088/1742-6596/1869/1/012085](https://doi.org/10.1088/1742-6596/1869/1/012085).
- [7] Y.-L. Chen, M.-H. Kuo, S.-Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data," *Electron Commer Res Appl*, vol. 8, no. 5, pp. 241–251, Oct. 2009, doi: [10.1016/j.elerap.2009.03.002](https://doi.org/10.1016/j.elerap.2009.03.002).
- [8] S. Manochandar, M. Punniyamoorthy, and R. K. Jeyachitra, "Development of new seed with modified validity measures for k-means clustering," *Comput Ind Eng*, vol. 141, p. 106290, Mar. 2020, doi: [10.1016/j.cie.2020.106290](https://doi.org/10.1016/j.cie.2020.106290).
- [9] Bao Chong, "K-means clustering algorithm: a brief review," *Academic Journal of Computing & Information Science*, vol. 4, no. 5, 2021, doi: [10.25236/AJCIS.2021.040506](https://doi.org/10.25236/AJCIS.2021.040506).
- [10] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: [10.3390/su14127243](https://doi.org/10.3390/su14127243).
- [11] A. D. Savitri, F. A. Bachtiar, and N. Y. Setiawan, "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus: Belle Crown Malang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2018.
- [12] H. Junaedi, H. Budianto, I. Maryati, and Y. Melani, "Data transformation pada data mining," 2011.
- [13] T. Gupta and S. P. Panda, "Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 10–13, doi: [10.1109/COMITCon.2019.8862199](https://doi.org/10.1109/COMITCon.2019.8862199).
- [14] M. Ambigavathi and D. Sridharan, "Analysis of Clustering Algorithms in Machine Learning for Healthcare Data," 2020, pp. 117–128, doi: [10.1007/978-981-15-6634-9\\_12](https://doi.org/10.1007/978-981-15-6634-9_12).
- [15] P. Bombina, D. Tally, Z. B. Abrams, and K. R. Coombes, "SillyPutty: Improved clustering by optimizing the silhouette width," *PLoS One*, vol. 19, no. 6, p. e0300358, Jun. 2024, doi: [10.1371/journal.pone.0300358](https://doi.org/10.1371/journal.pone.0300358).