

## SEGMENTASI PELANGGAN MENGGUNAKAN METODE RFM DAN K-MEANS CLUSTERING UNTUK MENDUKUNG STRATEGI PEMASARAN PADA ONLINE RETAIL

Selfy Anggraeni <sup>a\*</sup>, Zalfa Salsabila<sup>b</sup>, Mulky Baihaqi Hakim <sup>c</sup>

<sup>a</sup> Fakultas Ilmu Komputer/ Sistem Informasi; [selfyanggraeni21@gmail.com](mailto:selfyanggraeni21@gmail.com), Universitas Pamulang; Jl. Raya Puspittek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten

<sup>b</sup> Fakultas Ilmu Komputer/ Sistem Informasi; [zalfasalsaa079@gmail.com](mailto:zalfasalsaa079@gmail.com), Universitas Pamulang; Jl. Raya Puspittek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten

<sup>c</sup> Fakultas Ilmu Komputer/ Sistem Informasi; [mulkybhakim@gmail.com](mailto:mulkybhakim@gmail.com), Universitas Pamulang; Jl. Raya Puspittek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten

\* Penulis Korespondensi: Selfy Anggraeni

### ABSTRACT

*The rapid growth of e-commerce has generated massive customer transaction data that can support business decision-making. However, many companies have yet to fully leverage this data for understanding customer behavior and developing targeted strategies. This study aims to perform customer segmentation on the Online Retail II dataset using the Recency, Frequency, Monetary (RFM) approach combined with the K-Means Clustering algorithm in order to generate customer profiles that can serve as the basis for business strategy development.*

*The dataset used in this study consists of 1,067,371 customer transaction records with eight main attributes. The preprocessing stage included the removal of missing values in Customer ID, duplicate records, invalid transactions, and return transactions. Subsequently, RFM values were calculated for 5,878 unique customers. To improve data quality, outlier removal was performed using the Interquartile Range (IQR) method, resulting in 5,182 customers being retained for the clustering process. Prior to modeling, the data were normalized using the Min-Max Scaling technique. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, which indicated that  $k = 3$  was the best configuration with a Silhouette Score of 0.4780.*

*The clustering results identified three customer segments with distinct characteristics. Cluster 0 had an average Recency of 91.53 days, Frequency of 2.88 transactions, and Monetary value of 776.59, representing potential customers. Cluster 1 exhibited the highest Recency value of 479.40 days along with the lowest Frequency and Monetary values, indicating at-risk customers. Cluster 2 showed the lowest Recency value of 69.31 days and the highest Frequency and Monetary values of 8.80 transactions and 2,943.38 respectively, representing loyal and high-value customers (Champions). Based on the characteristics of each segment, different marketing strategies can be implemented, including loyalty programs, upselling strategies, and customer reactivation campaigns.*

*The findings demonstrate that the combination of the RFM method and K-Means Clustering is effective in generating customer segmentation and supporting data-driven customer intelligence to improve marketing strategy effectiveness in the retail industry.*

**Keywords:** Data Mining; Customer Segmentation; RFM; K-Means Clustering; Customer Intelligence; Online Retail

### Abstrak

Pertumbuhan industri e-commerce menghasilkan volume data transaksi pelanggan yang sangat besar. Data ini berpotensi dimanfaatkan sebagai sumber informasi untuk mendukung pengambilan keputusan bisnis. Namun, banyak perusahaan belum memanfaatkan data tersebut secara optimal untuk memahami perilaku pelanggan. Penelitian ini bertujuan untuk melakukan segmentasi pelanggan pada dataset Online Retail II menggunakan pendekatan *Recency, Frequency, Monetary* (RFM) yang dikombinasikan dengan algoritma K-Means Clustering guna menghasilkan profil pelanggan yang dapat digunakan sebagai dasar penyusunan strategi bisnis.

Dataset yang digunakan terdiri dari 1.067.371 data transaksi pelanggan dengan delapan atribut utama. Tahap preprocessing dilakukan melalui penghapusan missing value pada Customer ID, data duplikat, transaksi tidak valid, serta transaksi retur. Selanjutnya dilakukan perhitungan nilai RFM terhadap 5.878 pelanggan unik. Untuk meningkatkan kualitas data, dilakukan penghapusan outlier menggunakan metode Interquartile Range (IQR) sehingga diperoleh 5.182 pelanggan yang digunakan dalam proses clustering. Sebelum pemodelan, data dinormalisasi menggunakan metode Min-Max Scaling. Penentuan jumlah cluster optimal dilakukan menggunakan Elbow Method dan *Silhouette Score*, yang menunjukkan bahwa nilai  $k = 3$  merupakan konfigurasi terbaik dengan nilai *Silhouette Score* sebesar 0,4780.

Hasil clustering menghasilkan tiga segmen pelanggan dengan karakteristik yang berbeda. Cluster 0 memiliki nilai *Recency* rata-rata 91,53 hari, *Frequency* 2,88 transaksi, dan *Monetary* sebesar 776,59 sehingga dapat dikategorikan sebagai pelanggan potensial. Cluster 1 memiliki nilai *Recency* tertinggi yaitu 479,40 hari dengan *Frequency* dan *Monetary* terendah, sehingga dikategorikan sebagai pelanggan berisiko (At-Risk). Cluster 2 memiliki nilai *Recency* terendah sebesar 69,31 hari serta nilai *Frequency* dan *Monetary* tertinggi masing-masing sebesar 8,80 transaksi dan 2.943,38, sehingga dikategorikan sebagai pelanggan loyal atau bernilai tinggi (Champions). Berdasarkan karakteristik setiap segmen, perusahaan dapat menerapkan strategi pemasaran yang berbeda, seperti program loyalitas, strategi upselling, dan kampanye reaktivasi pelanggan. Hasil penelitian menunjukkan bahwa kombinasi metode RFM dan K-Means Clustering mampu menghasilkan segmentasi pelanggan yang efektif serta mendukung pembentukan customer intelligence berbasis data untuk meningkatkan efektivitas strategi pemasaran pada industri retail.

**Kata Kunci:** Data Mining; Customer Segmentation; RFM; K-Means Clustering; Customer Intelligence; Online Retail

## 1. PENDAHULUAN

Perkembangan teknologi informasi dan internet telah mendorong pertumbuhan bisnis online retail secara signifikan, dengan penetrasi e-commerce yang terus meningkat setiap tahunnya. Aktivitas transaksi yang dilakukan melalui platform digital menghasilkan data pelanggan dalam jumlah besar yang dapat dimanfaatkan untuk mendukung pengambilan keputusan bisnis, sebagaimana ditunjukkan oleh penelitian yang menganalisis data clickstream e-commerce untuk memahami perilaku navigasi dan keputusan pembelian konsumen secara lebih mendalam [1]. Pemanfaatan data pelanggan menjadi semakin penting karena perusahaan dituntut untuk memahami perilaku konsumen guna meningkatkan daya saing dan mempertahankan loyalitas pelanggan [1], [2].

Segmentasi pelanggan memungkinkan perusahaan mengidentifikasi kelompok pelanggan yang memiliki kebutuhan, karakteristik, dan nilai bisnis yang berbeda. Dengan adanya segmentasi tersebut, perusahaan dapat menyusun strategi pemasaran yang lebih personal, tepat sasaran, dan efisien sehingga mampu meningkatkan kepuasan pelanggan serta efektivitas kegiatan pemasaran [2].

Salah satu metode yang banyak digunakan dalam segmentasi pelanggan adalah *Recency*, *Frequency*, *Monetary* (RFM). Metode ini mengukur perilaku pelanggan berdasarkan waktu transaksi terakhir (*Recency*), frekuensi pembelian (*Frequency*), dan nilai transaksi yang dihasilkan (*Monetary*). Analisis RFM mampu memberikan gambaran mengenai tingkat loyalitas dan kontribusi pelanggan terhadap perusahaan sehingga dapat digunakan untuk mengidentifikasi pelanggan potensial maupun pelanggan yang berisiko tidak aktif [3], [4].

Selain analisis RFM, teknik clustering dalam data mining juga banyak dimanfaatkan untuk mengelompokkan pelanggan berdasarkan kemiripan karakteristik transaksi. Algoritma K-Means merupakan salah satu metode clustering yang populer karena memiliki proses komputasi yang sederhana, efisien, dan mampu menangani data dalam jumlah besar [9]. Kombinasi metode RFM dan K-Means memungkinkan pembentukan kelompok pelanggan yang lebih representatif sehingga perusahaan dapat memahami pola perilaku pelanggan secara lebih akurat [3].

Beberapa penelitian terdahulu menunjukkan bahwa integrasi RFM dan K-Means Clustering mampu menghasilkan segmentasi pelanggan yang efektif dalam mendukung strategi pemasaran. Abbasimehr dan

Bahrini [5] mengembangkan kerangka segmentasi pelanggan berbasis RFM dan clustering untuk menghasilkan segmentasi yang lebih dinamis. Penelitian Alzami dkk. [4] menunjukkan bahwa penerapan RFM dan K-Means pada data e-commerce dapat menghasilkan kelompok pelanggan dengan karakteristik yang berbeda sehingga membantu proses pengambilan keputusan pemasaran. Selain itu, penelitian Sarkar dkk. [2] dan Kasem dkk. [6] menunjukkan bahwa segmentasi pelanggan berbasis data mampu meningkatkan efektivitas strategi pemasaran serta mendukung pengelolaan hubungan pelanggan secara lebih optimal.

Meskipun berbagai penelitian telah menerapkan metode RFM dan K-Means untuk segmentasi pelanggan, implementasi pada data online retail masih perlu dikaji lebih karena karakteristik transaksi pelanggan online retail yang bersifat musiman dan memiliki variasi produk yang tinggi. Oleh karena itu, penelitian ini bertujuan untuk menerapkan metode RFM dan K-Means Clustering pada data transaksi online retail guna mengelompokkan pelanggan berdasarkan perilaku pembelian mereka. Hasil segmentasi yang diperoleh diharapkan dapat memberikan informasi yang bermanfaat bagi pengambilan keputusan strategis dalam penyusunan strategi pemasaran yang lebih efektif, meningkatkan loyalitas pelanggan, serta mendukung pertumbuhan bisnis perusahaan.

## 2. TINJAUAN PUSTAKA

### 2.1 Data Mining

Data mining merupakan proses ekstraksi pola, hubungan, atau pengetahuan yang tersembunyi dari sekumpulan data dalam jumlah besar dengan memanfaatkan berbagai teknik, seperti statistik, *machine learning*, kecerdasan buatan, dan sistem basis data. Menurut Han, Pei, dan Kamber [12], data mining merupakan salah satu tahapan dalam proses *Knowledge Discovery in Databases* (KDD). KDD adalah proses yang lebih luas yang mencakup tahapan pemilihan data (*data selection*), pembersihan data (*data cleaning*), transformasi data (*data transformation*), proses data mining, serta evaluasi dan interpretasi hasil. Dengan demikian, data mining berfokus pada proses menemukan pola atau pengetahuan yang bernilai dari data yang telah dipersiapkan sebagai bagian dari keseluruhan proses KDD.

Dalam konteks bisnis, data mining digunakan untuk menganalisis perilaku pelanggan, memprediksi tren pasar, mengidentifikasi peluang bisnis, mendukung personalisasi layanan, serta membantu pengambilan keputusan berbasis data. Dengan memanfaatkan data transaksi yang tersimpan dalam sistem informasi perusahaan, data mining mampu menghasilkan informasi dan pola yang bernilai sehingga perusahaan dapat menyusun strategi pemasaran yang lebih efektif, meningkatkan kepuasan pelanggan, serta memperkuat daya saing perusahaan [12].

### 2.2 Customer Intelligence

Customer Intelligence merupakan proses pengumpulan, pengolahan, dan analisis data pelanggan yang bertujuan untuk memperoleh pemahaman yang lebih mendalam mengenai karakteristik, kebutuhan, dan perilaku pelanggan. Informasi yang dihasilkan dari proses tersebut dapat dimanfaatkan untuk mendukung pengambilan keputusan bisnis yang lebih tepat, berbasis data, serta membantu perusahaan menyusun strategi pemasaran yang lebih efektif [6].

Menurut Kasem dkk. [6], Customer Intelligence membantu perusahaan dalam mengidentifikasi pelanggan bernilai tinggi, memahami pola pembelian pelanggan, serta mengembangkan strategi pemasaran yang lebih personal dan strategi retensi pelanggan yang efektif. Dengan demikian, Customer Intelligence menjadi salah satu komponen penting dalam meningkatkan efektivitas *Customer Relationship Management* (CRM), mempertahankan loyalitas pelanggan, serta membangun hubungan jangka panjang dengan pelanggan. Dalam penelitian ini, Customer Intelligence dimanfaatkan sebagai dasar untuk menginterpretasikan hasil segmentasi pelanggan sehingga setiap kelompok pelanggan dapat diberikan rekomendasi strategi pemasaran yang sesuai dengan karakteristiknya.

### 2.3 Customer Segmentation

Customer Segmentation merupakan proses mengelompokkan pelanggan ke dalam beberapa kelompok berdasarkan karakteristik, kebutuhan, perilaku, atau nilai transaksi yang memiliki kesamaan sehingga perusahaan dapat menyusun strategi pemasaran yang lebih tepat sasaran. Menurut Kotler dan Armstrong [13], segmentasi pelanggan membantu perusahaan memahami perbedaan karakteristik pelanggan agar produk, layanan, dan strategi pemasaran dapat disesuaikan dengan kebutuhan masing-masing segmen. Secara umum, segmentasi pelanggan dapat dilakukan berdasarkan aspek demografis, geografis, psikografis, dan perilaku. Dalam penelitian ini, segmentasi dilakukan berdasarkan perilaku transaksi pelanggan menggunakan pendekatan *Recency, Frequency, Monetary* (RFM).

Segmentasi pelanggan memungkinkan perusahaan mengalokasikan sumber daya pemasaran secara lebih efisien, meningkatkan retensi pelanggan, mengidentifikasi peluang pasar baru, serta menyusun strategi pemasaran yang lebih personal. Menurut Alzami dkk. [4], segmentasi pelanggan memungkinkan perusahaan mengidentifikasi pelanggan yang memberikan kontribusi tinggi terhadap pendapatan, pelanggan potensial yang dapat dikembangkan menjadi pelanggan loyal, serta pelanggan yang berisiko berhenti melakukan transaksi. Hasil penelitian Widiyanto dan Witanti [7] juga menunjukkan bahwa segmentasi pelanggan berbasis RFM dan K-Means dapat digunakan sebagai dasar penyusunan strategi pemasaran yang lebih tepat sasaran.

## 2.4 Recency Frequency Monetary (RFM)

Abbasimehr dan Bahrini [5] menyatakan bahwa metode *Recency, Frequency, Monetary* (RFM) mampu menghasilkan segmentasi pelanggan yang lebih dinamis dan mudah diinterpretasikan karena berfokus pada perilaku transaksi aktual pelanggan. Temuan serupa juga dilaporkan oleh Pangestu dkk. [8] yang menunjukkan bahwa kombinasi model RFM dan algoritma K-Means mampu menghasilkan segmentasi pelanggan yang jelas berdasarkan perilaku transaksi. Dalam metode RFM, *Recency* menunjukkan jarak waktu sejak transaksi terakhir pelanggan hingga tanggal referensi yang digunakan dalam penelitian dan dihitung dalam satuan hari, *Frequency* menunjukkan jumlah transaksi yang dilakukan pelanggan selama periode pengamatan, sedangkan *Monetary* menunjukkan total nilai transaksi yang dihasilkan oleh setiap pelanggan selama periode pengamatan.

Rumus *Recency*:

$$R = \text{Reference Date} - \text{Last Transaction Date} \quad (1)$$

Rumus *Recency* digunakan untuk menghitung selisih antara tanggal referensi (*Reference Date*) dengan tanggal transaksi terakhir pelanggan (*Last Transaction Date*). Dalam penelitian ini, tanggal referensi yang digunakan adalah 31 Desember 2011, yaitu tanggal terakhir pada periode dataset sehingga nilai *Recency* dihitung dalam satuan hari.

*Frequency* menunjukkan jumlah transaksi yang dilakukan oleh setiap pelanggan selama periode pengamatan. Semakin tinggi nilai *Frequency*, semakin sering pelanggan melakukan transaksi sehingga menunjukkan tingkat aktivitas pembelian yang lebih tinggi.

Rumus *Frequency*:

$$F = \Sigma \text{Transaksi Pelanggan Dalam Periode Pengamatan} \quad (2)$$

*Monetary* menunjukkan total nilai transaksi yang dihasilkan oleh setiap pelanggan selama periode penelitian. Nilai *Monetary* diperoleh dari akumulasi hasil perkalian antara jumlah produk (*Quantity*) dan harga (*Price*) pada setiap transaksi yang dilakukan oleh pelanggan yang sama.

Rumus *Monetary*:

$$M = \Sigma(\text{Quantity} \times \text{Price}) \quad (3)$$

Dengan:

$R = \text{Recency}$

$F = \text{Frequency}$

$M = \text{Monetary}$

*Quantity* = jumlah produk yang dibeli

*Price* = harga produk

## 2.5 K-Means Clustering

K-Means Clustering merupakan salah satu algoritma *clustering* yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik. Algoritma ini membagi data ke dalam sejumlah cluster yang ditentukan terlebih dahulu oleh pengguna sebelum proses clustering dimulai. Menurut John dkk. [9], K-

Means menjadi salah satu algoritma clustering yang paling banyak digunakan karena memiliki proses komputasi yang sederhana, efisien, dan mampu menangani data dalam jumlah besar.

Proses K-Means dimulai dengan menentukan jumlah cluster, memilih centroid awal secara acak atau menggunakan metode tertentu, menghitung jarak setiap data terhadap centroid menggunakan Euclidean Distance, kemudian menempatkan setiap data ke dalam cluster dengan centroid terdekat. Selanjutnya centroid diperbarui berdasarkan rata-rata anggota pada masing-masing cluster. Proses ini dilakukan secara iteratif hingga posisi centroid tidak mengalami perubahan yang signifikan atau telah mencapai kondisi konvergen, sehingga menghasilkan pengelompokan data yang stabil.

Rumus Euclidean Distance:

$$d(x, C_k) = \sqrt{[(x_1 - C_{k1})^2 + (x_2 - C_{k2})^2 + \dots + (x_n - C_{kn})^2]} \quad (4)$$

Dengan:

$d(x, C_k)$  = jarak antara data dan centroid

$x$  = titik data

$C_k$  = centroid

$n$  = jumlah atribut

Setelah seluruh data dikelompokkan, centroid baru dihitung menggunakan rata-rata seluruh anggota cluster.

Rumus centroid:

$$C_k = \frac{\sum_{i=1}^n X_i}{n} \quad (5)$$

Dengan:

$C_k$  = centroid cluster ke- $k$

$X_i$  = data ke- $i$  pada cluster

$n$  = jumlah anggota cluster

Proses ini dilakukan secara berulang hingga centroid tidak mengalami perubahan yang signifikan.

## 2.6 Elbow Method

Elbow Method merupakan metode yang digunakan untuk menentukan jumlah cluster optimal pada algoritma K-Means dengan menghitung nilai *Within Cluster Sum of Squares* (WCSS) atau *inertia* untuk beberapa jumlah cluster yang berbeda. Jumlah cluster optimal ditentukan berdasarkan titik siku (*elbow point*), yaitu titik ketika penurunan nilai *inertia* mulai melambat secara signifikan. Kondisi tersebut menunjukkan bahwa penambahan jumlah cluster tidak lagi memberikan peningkatan kualitas pengelompokan yang berarti.

Rumus WCSS:

$$\sum_{k=1}^K \sum_{X_i \in C_k} (X_i - C_k)^2 \quad (6)$$

Dengan:

$X_i$  = data ke- $i$

$C_k$  = centroid cluster ke- $k$

Pada rumus WCSS, notasi  $X_i \in C_k$  menunjukkan bahwa data  $X_i$  merupakan anggota dari cluster  $C_k$ , sehingga perhitungan jarak hanya dilakukan terhadap data yang berada pada cluster tersebut. Meskipun Elbow Method dapat digunakan untuk menentukan jumlah cluster optimal, pemilihan titik *elbow* bersifat subjektif sehingga perlu didukung oleh metode evaluasi lain, seperti *Silhouette Score*, untuk memperoleh hasil clustering yang lebih akurat.

## 2.7 Silhouette Score

*Silhouette Score* merupakan metode evaluasi clustering yang digunakan untuk mengukur kualitas hasil pengelompokan data. Nilai *Silhouette Score* berada pada rentang -1 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan bahwa data berada pada cluster yang sesuai dan memiliki tingkat pemisahan yang baik terhadap cluster lain. Sebaliknya, nilai yang mendekati -1 menunjukkan bahwa data kemungkinan berada pada cluster yang kurang sesuai [3].

Menurut Wilbert dkk. [3], *Silhouette Score* dapat digunakan untuk mengukur tingkat kedekatan data dengan cluster tempat data tersebut berada serta tingkat perbedaannya terhadap cluster lain. Oleh karena itu, metode ini sering digunakan sebagai salah satu indikator dalam menentukan jumlah cluster optimal dan mengevaluasi performa algoritma clustering.

Secara umum, nilai *Silhouette Score* di atas 0,25 menunjukkan kualitas clustering yang cukup baik, nilai di atas 0,50 menunjukkan kualitas clustering yang baik, sedangkan nilai di atas 0,70 menunjukkan kualitas clustering yang sangat baik [3].

Rumus *Silhouette Score*:

$$S(i) = \left[ \frac{b(i) - a(i)}{\max[a(i), b(i)]} \right] \quad (7)$$

Dengan:

$a(i)$  = rata-rata jarak data terhadap anggota cluster yang sama

$b(i)$  = rata-rata jarak data terhadap cluster terdekat lainnya

Interpretasi nilai *Silhouette Score* adalah sebagai berikut:

1. Nilai mendekati 1 menunjukkan kualitas cluster sangat baik.
2. Nilai mendekati 0 menunjukkan cluster saling tumpang tindih.
3. Nilai mendekati -1 menunjukkan data berada pada cluster yang kurang sesuai.

## 2.8 Cross Industry Standard Process for Data Mining (CRISP-DM)

Cross Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu metodologi yang paling banyak digunakan sebagai standar industri dalam pelaksanaan proyek data mining. Metodologi ini dikembangkan untuk menyediakan kerangka kerja yang sistematis dalam proses penggalian data, mulai dari pemahaman permasalahan bisnis hingga implementasi hasil analisis [10]. Menurut Chapman dkk. [10], CRISP-DM terdiri atas enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*.

Dalam penelitian ini, metodologi CRISP-DM digunakan sebagai kerangka kerja untuk melakukan segmentasi pelanggan menggunakan metode RFM dan algoritma K-Means Clustering. Setiap tahapan dilakukan secara sistematis mulai dari identifikasi permasalahan bisnis, pengolahan data, pembentukan model, hingga evaluasi hasil clustering. Pada tahap *Deployment*, hasil segmentasi dimanfaatkan untuk menyusun rekomendasi strategi pemasaran berdasarkan karakteristik masing-masing segmen pelanggan. Rekomendasi tersebut dapat menjadi dasar pengambilan keputusan bisnis serta diintegrasikan dengan sistem *Customer Relationship Management* (CRM) untuk mendukung pelaksanaan strategi pemasaran yang lebih efektif.

## 2.9 Penelitian Terdahulu

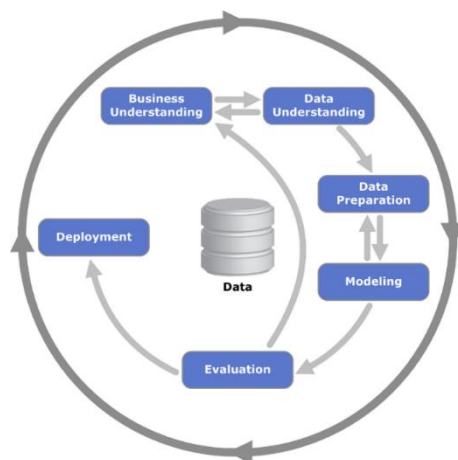
Berbagai penelitian terdahulu menunjukkan bahwa kombinasi metode *Recency*, *Frequency*, *Monetary* (RFM) dan algoritma K-Means Clustering merupakan pendekatan yang efektif dalam melakukan segmentasi pelanggan pada berbagai bidang bisnis. Abbasimehr dan Bahrini [5] mengembangkan kerangka segmentasi pelanggan berbasis RFM dan teknik clustering untuk menghasilkan segmentasi yang lebih dinamis. Penelitian Alzami dkk. [4], Sarkar dkk. [2], Kasem dkk. [6], Widiyanto dan Witanti [7], Pangestu dkk. [8], serta Perdana dkk. [11] juga menunjukkan bahwa penerapan RFM dan K-Means mampu menghasilkan kelompok pelanggan dengan karakteristik yang berbeda sehingga dapat mendukung penyusunan strategi pemasaran, meningkatkan efektivitas *Customer Relationship Management* (CRM), dan membantu pengambilan keputusan bisnis.

Meskipun demikian, implementasi metode RFM dan K-Means pada dataset Online Retail II masih menarik untuk dikaji karena karakteristik transaksi pelanggan pada *online retail* bersifat dinamis, memiliki variasi produk yang tinggi, serta dipengaruhi oleh pola pembelian pelanggan yang terus berubah. Selain itu, belum banyak penelitian yang mengaitkan hasil segmentasi pelanggan dengan konsep Customer Intelligence sebagai dasar penyusunan rekomendasi strategi pemasaran berbasis data.

Berdasarkan penelitian-penelitian tersebut, penelitian ini menerapkan metode RFM dan algoritma K-Means Clustering pada dataset Online Retail II untuk menghasilkan segmentasi pelanggan yang dapat dimanfaatkan sebagai dasar pembentukan Customer Intelligence serta penyusunan strategi pemasaran yang lebih efektif sesuai karakteristik setiap segmen pelanggan.

### 3. METODOLOGI PENELITIAN

Penelitian ini menggunakan metodologi Cross Industry Standard Process for Data Mining (CRISP-DM) sebagai kerangka kerja dalam proses segmentasi pelanggan. Metodologi ini terdiri dari enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [10].



Gambar 1. Tahapan CRISP-DM

Sumber: Diadaptasi dari Chapman et al. (2000).

Berdasarkan Gambar 1, penelitian ini mengikuti enam tahapan CRISP-DM yang terdiri dari Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Setiap tahapan dilakukan secara sistematis untuk menghasilkan segmentasi pelanggan yang dapat mendukung penyusunan strategi pemasaran berbasis customer intelligence.

#### 3.1 Business Understanding

Pertumbuhan bisnis *online retail* menghasilkan data transaksi pelanggan dalam jumlah besar yang berpotensi dimanfaatkan sebagai sumber informasi untuk mendukung pengambilan keputusan bisnis. Namun, tanpa adanya segmentasi pelanggan yang terstruktur, perusahaan akan mengalami kesulitan dalam mengidentifikasi pelanggan bernilai tinggi, memahami karakteristik pelanggan, serta menyusun strategi pemasaran yang efektif. Oleh karena itu, diperlukan pendekatan analisis yang mampu mengelompokkan pelanggan berdasarkan perilaku transaksi sehingga strategi bisnis dapat disusun secara lebih tepat sasaran.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengelompokkan pelanggan berdasarkan perilaku transaksi menggunakan metode *Recency*, *Frequency*, *Monetary* (RFM) dan algoritma K-Means Clustering. Hasil segmentasi diharapkan dapat membantu perusahaan dalam memahami karakteristik pelanggan, menyusun strategi pemasaran yang lebih efektif, meningkatkan loyalitas pelanggan, serta mempertahankan pelanggan bernilai tinggi.

Rumusan masalah dalam penelitian ini adalah:

Bagaimana profil pelanggan yang dihasilkan melalui segmentasi menggunakan metode *Recency*, *Frequency*, *Monetary* (RFM) dan algoritma K-Means Clustering pada dataset Online Retail II, serta bagaimana hasil segmentasi tersebut dapat dimanfaatkan untuk menyusun strategi pemasaran yang lebih efektif?.

### 3.2 Data Understanding

Dataset yang digunakan dalam penelitian ini adalah Dataset Online Retail II yang diperoleh melalui platform Kaggle. Dataset ini bersifat publik sehingga dapat digunakan untuk keperluan penelitian dan telah banyak dimanfaatkan pada berbagai penelitian mengenai segmentasi pelanggan, analisis perilaku konsumen, serta penerapan metode data mining. Dataset mencakup transaksi penjualan pada perusahaan *online retail* yang berlangsung selama periode Desember 2009 hingga Desember 2011, sehingga dinilai sesuai untuk penelitian segmentasi pelanggan menggunakan metode RFM dan algoritma K-Means Clustering.

Tabel 1. Deskripsi Atribut Dataset

| Atribut     | Keterangan                |
|-------------|---------------------------|
| Invoice     | Nomor transaksi           |
| StockCode   | Kode produk               |
| Description | Nama produk               |
| Quantity    | Jumlah produk yang dibeli |
| InvoiceDate | Tanggal transaksi         |
| Price       | Harga produk              |
| Customer ID | Identitas pelanggan       |
| Country     | Negara pelanggan          |

Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Berdasarkan hasil eksplorasi data, dataset terdiri atas 1.067.371 transaksi dengan delapan atribut utama. Setelah dilakukan identifikasi awal, diketahui bahwa transaksi berasal dari berbagai negara dengan mayoritas pelanggan berasal dari United Kingdom. Selain itu, dataset memiliki 5.878 pelanggan unik yang selanjutnya digunakan dalam proses pembentukan nilai *Recency*, *Frequency*, dan *Monetary* (RFM) setelah melalui tahapan *preprocessing*.

### 3.3 Data Preparation

Tahap *data preparation* dilakukan untuk menghasilkan data yang siap digunakan dalam proses segmentasi pelanggan. Pada tahap ini dilakukan pembersihan data, pembentukan fitur RFM, penanganan *outlier*, serta normalisasi data.

#### 3.3.1 Data Cleaning

Proses *data cleaning* dilakukan untuk meningkatkan kualitas data sebelum dianalisis. Tahapan yang dilakukan meliputi penghapusan data yang memiliki nilai kosong (*missing value*) pada atribut Customer ID, penghapusan data duplikat, penghapusan transaksi yang memiliki nilai *Quantity* atau *Price* tidak valid, serta penghapusan transaksi retur (*cancelled transaction*). Transaksi retur diidentifikasi melalui nilai *Quantity* yang bernilai negatif maupun nomor Invoice yang diawali dengan huruf "C". Setelah seluruh tahapan *data cleaning* dilakukan, diperoleh data yang lebih bersih dan siap digunakan pada tahap pembentukan nilai RFM.

#### 3.3.2 Pembentukan Fitur RFM

Setelah proses pembersihan data selesai dilakukan, tahap berikutnya adalah pembentukan fitur *Recency*, *Frequency*, dan *Monetary* (RFM) untuk setiap pelanggan. Nilai *Recency* menunjukkan jarak waktu sejak transaksi terakhir pelanggan, *Frequency* menunjukkan jumlah transaksi yang dilakukan pelanggan, sedangkan *Monetary* menunjukkan total nilai transaksi yang dihasilkan oleh setiap pelanggan selama periode pengamatan. Ketiga atribut tersebut digunakan sebagai variabel utama dalam proses segmentasi pelanggan menggunakan algoritma K-Means Clustering.

#### 3.3.3 Penanganan Outlier

Nilai RFM yang diperoleh masih mengandung beberapa data ekstrem (*outlier*) yang dapat memengaruhi hasil clustering. Oleh karena itu dilakukan deteksi dan penghapusan *outlier* menggunakan metode Interquartile Range (IQR). Pada metode ini, data dikategorikan sebagai *outlier* apabila memiliki nilai lebih kecil dari  $Q1 - 1,5 \times IQR$  atau lebih besar dari  $Q3 + 1,5 \times IQR$ , dengan  $IQR = Q3 - Q1$ . Proses ini bertujuan untuk mengurangi pengaruh nilai yang terlalu tinggi atau terlalu rendah sehingga data menjadi lebih representatif sebelum dilakukan proses clustering.

#### 3.3.4 Normalisasi Data

Atribut *Recency*, *Frequency*, dan *Monetary* memiliki rentang nilai yang berbeda sehingga diperlukan proses normalisasi. Pada penelitian ini digunakan metode Min-Max Scaling karena mampu mengubah seluruh atribut ke dalam rentang nilai yang sama, yaitu 0 hingga 1, tanpa mengubah pola distribusi data secara signifikan. Dengan demikian, tidak terdapat atribut yang mendominasi proses perhitungan jarak pada algoritma K-Means.

### 3.4 Modeling

Tahap pemodelan dilakukan menggunakan algoritma K-Means Clustering untuk mengelompokkan pelanggan berdasarkan nilai *Recency*, *Frequency*, dan *Monetary* (RFM) yang telah dinormalisasi. Penentuan jumlah cluster dilakukan menggunakan Elbow Method dan *Silhouette Score* sehingga diperoleh jumlah cluster yang paling sesuai dengan karakteristik data.

Setelah jumlah cluster optimal diperoleh, algoritma K-Means digunakan untuk membentuk kelompok pelanggan berdasarkan tingkat kemiripan karakteristik transaksi. Proses pemodelan dilakukan secara iteratif melalui penentuan centroid awal, perhitungan jarak setiap data terhadap centroid menggunakan Euclidean Distance, pengelompokan data ke dalam cluster terdekat, serta pembaruan centroid hingga diperoleh hasil clustering yang stabil.

### 3.5 Evaluation

Tahap Evaluation dilakukan untuk menentukan jumlah cluster optimal serta mengevaluasi kualitas hasil clustering yang dihasilkan oleh algoritma K-Means. Pada penelitian ini, proses evaluasi dilakukan menggunakan Elbow Method dan *Silhouette Score* secara bersama-sama sehingga jumlah cluster yang dipilih tidak hanya optimal berdasarkan pola penurunan *inertia*, tetapi juga memiliki kualitas pengelompokan yang baik.

#### 3.5.1 Elbow Method

Elbow Method digunakan untuk menentukan jumlah cluster optimal berdasarkan nilai *inertia* atau *Within Cluster Sum of Squares* (WCSS) yang diperoleh pada setiap jumlah cluster. Pada penelitian ini dilakukan pengujian jumlah cluster  $k = 2$  hingga  $k = 10$ . Jumlah cluster terbaik ditentukan berdasarkan titik siku (*elbow point*), yaitu titik ketika penurunan nilai *inertia* mulai melambat sehingga penambahan cluster tidak lagi memberikan peningkatan yang signifikan.

#### 3.5.2 Silhouette Score

*Silhouette Score* digunakan untuk mengevaluasi kualitas cluster yang terbentuk berdasarkan tingkat kedekatan data dengan cluster tempat data tersebut berada serta tingkat pemisahannya terhadap cluster lain. Nilai *Silhouette Score* digunakan sebagai pendukung hasil Elbow Method sehingga jumlah cluster yang dipilih tidak hanya optimal secara visual, tetapi juga memiliki kualitas pengelompokan yang baik secara statistik. Semakin tinggi nilai *Silhouette Score*, semakin baik kualitas segmentasi yang dihasilkan.

### 3.6 Deployment

Tahap Deployment dilakukan melalui interpretasi hasil segmentasi pelanggan serta penyusunan rekomendasi strategi pemasaran berdasarkan karakteristik masing-masing cluster yang dihasilkan. Rekomendasi tersebut diharapkan dapat membantu perusahaan dalam menentukan strategi yang sesuai untuk setiap kelompok pelanggan, seperti program loyalitas bagi pelanggan bernilai tinggi, strategi *cross-selling* bagi pelanggan potensial, serta program reaktivasi bagi pelanggan yang berisiko tidak aktif.

Selain itu, hasil segmentasi pelanggan dapat dimanfaatkan sebagai bagian dari Customer Intelligence dan diintegrasikan dengan sistem Customer Relationship Management (CRM) untuk mendukung pengambilan keputusan bisnis, meningkatkan efektivitas strategi pemasaran, serta memperkuat hubungan jangka panjang dengan pelanggan.

## 4. HASIL DAN PEMBAHASAN

### 4.1 Eksplorasi Dataset

Dataset Online Retail II terdiri dari 1.067.371 transaksi pelanggan dengan delapan atribut utama. Dataset ini digunakan sebagai sumber data untuk membangun model segmentasi pelanggan.

Tabel 2. Dataset Awal

|   | Invoice | StockCode | Description                         | Quantity | InvoiceDate         | Price | Customer ID | Country        |
|---|---------|-----------|-------------------------------------|----------|---------------------|-------|-------------|----------------|
| 0 | 489434  | 85048     | 15CM CHRISTMAS GLASS BALL 20 LIGHTS | 12       | 2009-12-01-07:45:00 | 6.95  | 13085.0     | United Kingdom |
| 1 | 489434  | 79323P    | PINK CHERRY LIGHTS                  | 12       | 2009-12-01-07:45:00 | 6.75  | 13085.0     | United Kingdom |
| 2 | 489434  | 79323W    | WHITE CHERRY LIGHTS                 | 12       | 2009-12-01-07:45:00 | 6.75  | 13085.0     | United Kingdom |
| 3 | 489434  | 22041     | RECORD FRAME 7" SINGLE SIZE         | 48       | 2009-12-01-07:45:00 | 2.10  | 13085.0     | United Kingdom |
| 4 | 489434  | 21232     | STRAWBERRY CERAMIC TRINKET BOX      | 24       | 2009-12-01-07:45:00 | 1.25  | 13085.0     | United Kingdom |

Sumber: Hasil Pengolahan Data Menggunakan Python pada Google Colab (2026).

Berdasarkan hasil eksplorasi, dataset terdiri atas 1.067.371 transaksi dengan delapan atribut utama yang mencakup informasi nomor transaksi (*Invoice*), kode produk (*StockCode*), nama produk (*Description*), jumlah pembelian (*Quantity*), tanggal transaksi (*InvoiceDate*), harga (*Price*), identitas pelanggan (*Customer ID*), dan negara asal pelanggan (*Country*). Setiap baris pada dataset merepresentasikan satu item produk dalam satu transaksi, sehingga satu nomor Invoice dapat muncul pada beberapa baris apabila pelanggan membeli lebih dari satu jenis produk dalam transaksi yang sama.

Selain itu, dataset mencakup transaksi selama periode Desember 2009 hingga Desember 2011 dengan mayoritas transaksi berasal dari United Kingdom. Hasil eksplorasi juga menunjukkan bahwa dataset masih mengandung missing value pada atribut Customer ID, data duplikat, serta transaksi retur sehingga diperlukan proses *preprocessing* sebelum dilakukan analisis lebih lanjut.

#### 4.2 Hasil Pembentukan Nilai RFM

Setelah proses data cleaning selesai dilakukan, dilakukan perhitungan nilai *Recency*, *Frequency*, dan *Monetary* (RFM) untuk setiap pelanggan.

Tabel 3. Perhitungan Nilai RFM

| Customer ID | Recency | Frequency | Monetary |
|-------------|---------|-----------|----------|
| 12346.0     | 326     | 12        | 77556.46 |
| 12347.0     | 2       | 8         | 4921.53  |
| 12348.0     | 75      | 5         | 2019.40  |
| 12349.0     | 19      | 4         | 4428.69  |
| 12350.0     | 310     | 1         | 334.40   |

Sumber: Hasil Pengolahan Data Menggunakan Python pada Google Colab (2026).

Hasil perhitungan menghasilkan nilai *Recency*, *Frequency*, dan *Monetary* (RFM) yang digunakan sebagai dasar dalam proses segmentasi pelanggan. Nilai *Recency* dihitung berdasarkan selisih antara tanggal referensi, yaitu tanggal terakhir pada periode dataset, dengan tanggal transaksi terakhir setiap pelanggan sehingga diperoleh jarak waktu dalam satuan hari. Selanjutnya dilakukan analisis statistik deskriptif terhadap nilai RFM untuk mengetahui karakteristik data sebelum dilakukan proses clustering.

Tabel 4. Statistik Deskriptif RFM

|       | Recency     | Frequency   | Monetary     |
|-------|-------------|-------------|--------------|
| count | 5878.000000 | 5878.000000 | 5878.000000  |
| mean  | 201.331916  | 6.289384    | 2955.904095  |
| std   | 209.338707  | 13.009406   | 14440.852688 |

*Segmentasi Pelanggan Menggunakan Metode RFM dan K-Means Clustering Untuk Mendukung Strategi Pemasaran Pada Online Retail (Selfy Anggraeni)*

|     |            |            |               |
|-----|------------|------------|---------------|
| min | 1.000000   | 1.000000   | 2.950000      |
| 25% | 26.000000  | 1.000000   | 342.280000    |
| 50% | 96.000000  | 3.000000   | 867.740000    |
| 75% | 380.000000 | 7.000000   | 2248.305000   |
| max | 739.000000 | 398.000000 | 580987.040000 |

Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

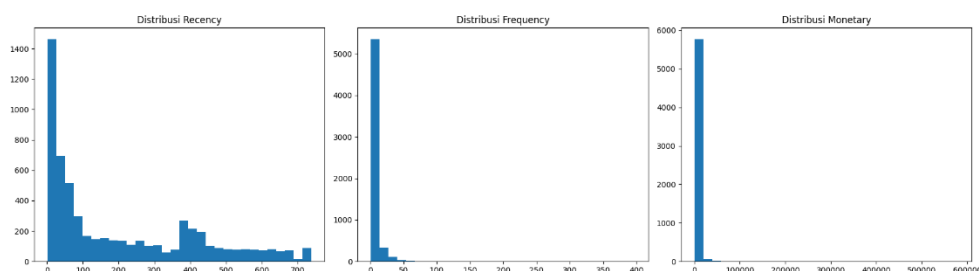
Berdasarkan hasil statistik deskriptif, nilai rata-rata *Recency* sebesar 201,33 hari, *Frequency* sebesar 6,29 transaksi, dan *Monetary* sebesar 2.955,90. Nilai maksimum *Monetary* mencapai 580.987,04, jauh lebih tinggi dibandingkan nilai rata-ratanya. Selain itu, nilai standar deviasi atribut *Monetary* sebesar 14.440,85 menunjukkan adanya variasi nilai transaksi yang sangat besar antar pelanggan. Kondisi tersebut mengindikasikan adanya outlier yang berpotensi memengaruhi proses clustering sehingga diperlukan tahap penanganan outlier sebelum proses pemodelan dilakukan.

### 4.3 Preprocessing Data

Tahap preprocessing dilakukan untuk menghasilkan data yang berkualitas dan siap digunakan pada proses *clustering*. Tahapan yang dilakukan meliputi visualisasi distribusi data, penanganan *outlier*, dan normalisasi data.

#### 4.3.1 Distribusi Data RFM

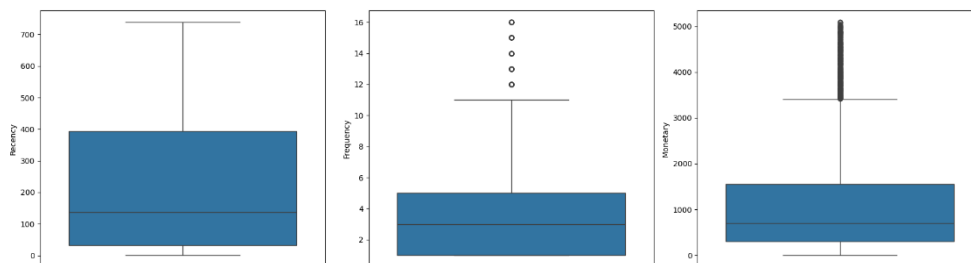
Visualisasi distribusi menunjukkan bahwa atribut *Frequency* dan *Monetary* memiliki distribusi yang cenderung tidak normal dan bersifat right-skewed. Kondisi tersebut menunjukkan bahwa sebagian besar pelanggan memiliki frekuensi transaksi dan nilai pembelian yang relatif rendah, sedangkan hanya sebagian kecil pelanggan yang memiliki frekuensi maupun nilai transaksi yang sangat tinggi. Pola distribusi seperti ini umum dijumpai pada data transaksi *online retail* dan menjadi salah satu alasan dilakukannya penanganan *outlier* sebelum proses clustering.



Gambar 2. Distribusi Nilai *Recency*, *Frequency*, dan *Monetary* Sebelum Penanganan Outlier  
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

#### 4.3.2 Penanganan Outlier

*Outlier* dihapus menggunakan metode IQR untuk mengurangi pengaruh data ekstrem terhadap proses *clustering*.



Gambar 3. Boxplot Nilai RFM Setelah Penanganan Outlier Menggunakan Metode IQR  
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Hasil boxplot menunjukkan bahwa sebagian besar *outlier* berhasil dikurangi sehingga distribusi data menjadi lebih stabil. Setelah dilakukan penghapusan *outlier* menggunakan metode Interquartile Range (IQR), jumlah pelanggan yang digunakan dalam proses clustering berkurang dari 5.878 menjadi 5.182 pelanggan. Dengan

demikian, sebanyak 696 pelanggan teridentifikasi sebagai *outlier* dan tidak diikutsertakan pada proses pemodelan untuk meningkatkan kualitas hasil clustering.

#### 4.3.3 Normalisasi Data

Setelah penghapusan *outlier*, dilakukan normalisasi menggunakan *Min-Max Scaling*.

Tabel 5. Hasil Normalisasi Data Menggunakan Min-Max Scaling

| <b>Customer ID</b> | <b>Recency</b> | <b>Frequency</b> | <b>Monetary</b> |
|--------------------|----------------|------------------|-----------------|
| 12347.0            | 0.001355       | 0.466667         | 0.966862        |
| 12348.0            | 0.100271       | 0.266667         | 0.396380        |
| 12349.0            | 0.024390       | 0.200000         | 0.869982        |
| 12350.0            | 0.418699       | 0.000000         | 0.065154        |
| 12351.0            | 0.506775       | 0.000000         | 0.058575        |

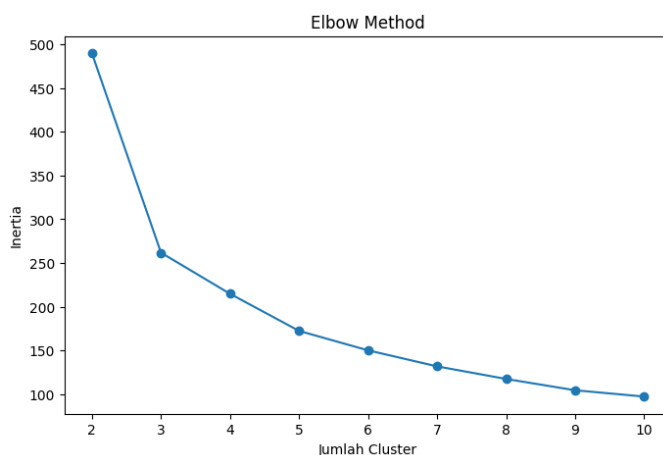
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Proses normalisasi menggunakan metode Min-Max Scaling menghasilkan seluruh atribut berada pada rentang nilai 0 hingga 1, sehingga setiap atribut memiliki skala yang sama dalam proses perhitungan jarak pada algoritma K-Means. Metode Min-Max Scaling dipilih karena mampu mempertahankan pola distribusi data sekaligus menghindari dominasi atribut yang memiliki rentang nilai lebih besar dibandingkan atribut lainnya.

#### 4.4 Penentuan Jumlah Cluster Optimal

##### 4.4.1 Elbow Method

Penentuan jumlah cluster dilakukan menggunakan Elbow Method.



Gambar 4. Grafik Elbow Method untuk Menentukan Jumlah Cluster Optimal

Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Grafik Elbow Method menunjukkan bahwa nilai inertia mengalami penurunan yang cukup signifikan dari k=2 hingga k=3. Setelah k=3, penurunan inertia mulai melambat sehingga nilai k=3 dipertimbangkan sebagai jumlah cluster optimal.

##### 4.4.2 Silhouette Score

Untuk memastikan jumlah cluster terbaik, dilakukan evaluasi menggunakan *Silhouette Score*.

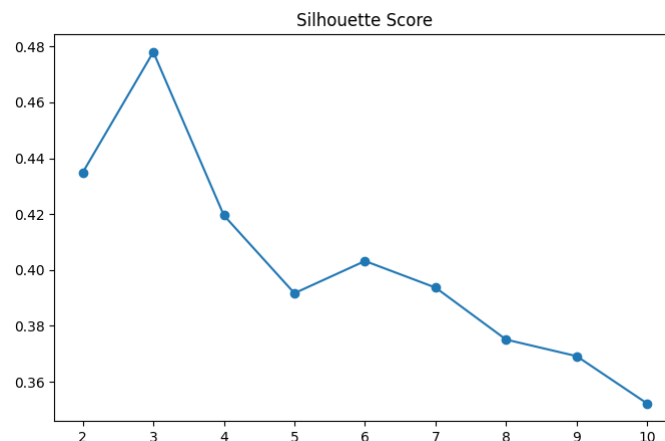
Tabel 6. Hasil Silhouette Score

| <b>Jumlah Cluster (k)</b> | <b>Silhouette Score</b> |
|---------------------------|-------------------------|
| 2                         | 0,4349                  |
| 3                         | 0,4780                  |
| 4                         | 0,4195                  |

*Segmentasi Pelanggan Menggunakan Metode RFM dan K-Means Clustering Untuk Mendukung Strategi Pemasaran Pada Online Retail (Selfy Anggraeni)*

| Jumlah Cluster (k) | <i>Silhouette Score</i> |
|--------------------|-------------------------|
| 5                  | 0,3917                  |
| 6                  | 0,4032                  |
| 7                  | 0,3937                  |
| 8                  | 0,3751                  |
| 9                  | 0,3691                  |
| 10                 | 0,3522                  |

Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

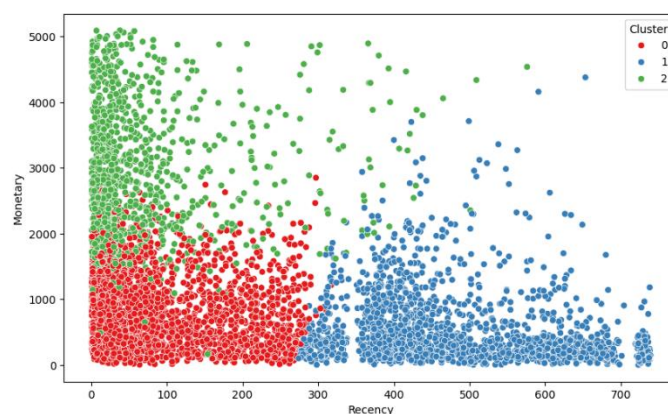


Gambar 5. Grafik Nilai *Silhouette Score* pada Berbagai Jumlah Cluster  
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Hasil pengujian menunjukkan bahwa nilai *Silhouette Score* tertinggi diperoleh pada  $k = 3$ , yaitu sebesar 0,4780. Nilai tersebut menunjukkan bahwa hasil clustering memiliki kualitas yang cukup baik, dengan tingkat homogenitas anggota cluster yang memadai serta pemisahan antarcluster yang cukup jelas. Hasil ini memperkuat temuan pada Elbow Method, sehingga jumlah cluster yang digunakan dalam penelitian ditetapkan sebanyak tiga cluster.

#### 4.5 Hasil K-Means Clustering

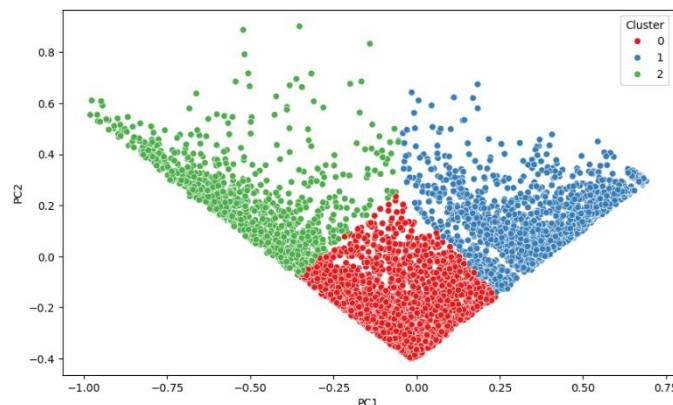
Proses clustering menggunakan K-Means menghasilkan tiga kelompok pelanggan yang memiliki karakteristik berbeda.



Gambar 6. Scatter Plot Hasil Pengelompokan Pelanggan Menggunakan Algoritma K-Means  
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Visualisasi *scatter plot* menunjukkan bahwa pelanggan berhasil dikelompokkan ke dalam tiga cluster berdasarkan karakteristik transaksi yang dimiliki. Secara umum, ketiga cluster dapat dibedakan dengan cukup baik, meskipun masih terdapat beberapa data yang berada pada area yang berdekatan. Cluster 2 (Champions)

terlihat memiliki karakteristik yang paling berbeda dibandingkan cluster lainnya, yang menunjukkan bahwa kelompok pelanggan tersebut memiliki pola transaksi yang lebih khas berdasarkan nilai *Recency*, *Frequency*, dan *Monetary*.



Gambar 7. Visualisasi Hasil Clustering Menggunakan Principal Component Analysis (PCA)  
Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Untuk mempermudah visualisasi hasil clustering, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA) sehingga data multidimensi dapat ditampilkan dalam dua dimensi tanpa mengubah karakteristik utama data secara signifikan. Hasil visualisasi menunjukkan bahwa ketiga cluster dapat dipisahkan dengan cukup baik, meskipun masih terdapat sedikit tumpang tindih antara Cluster 0 dan Cluster 1. Kondisi tersebut menunjukkan bahwa kedua cluster memiliki beberapa karakteristik transaksi yang relatif mirip, sedangkan Cluster 2 memiliki pemisahan yang lebih jelas sehingga merepresentasikan kelompok pelanggan dengan karakteristik yang paling berbeda.

#### 4.6 Analisis Profil Cluster

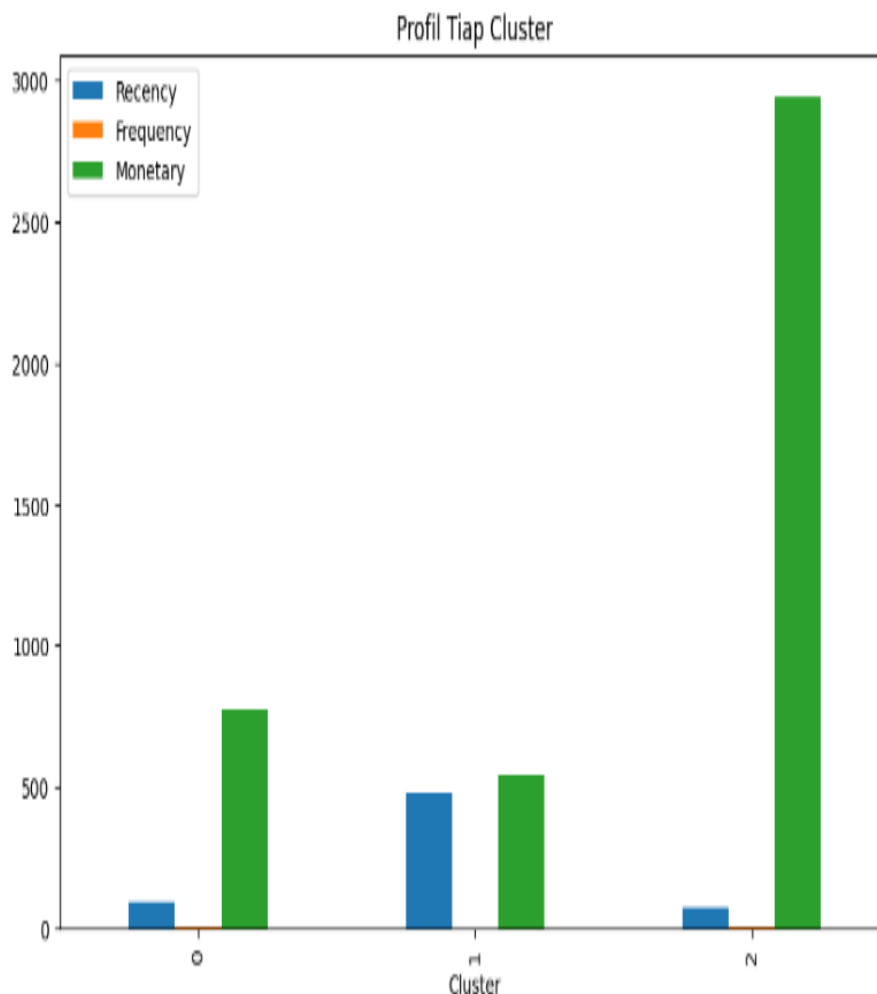
Hasil rata-rata nilai RFM setiap cluster ditunjukkan pada Tabel 7.

Tabel 7. Profil Rata-Rata Cluster

| Cluster | <i>Recency</i> | <i>Frequency</i> | <i>Monetary</i> |
|---------|----------------|------------------|-----------------|
| 0       | 91,53          | 2,88             | 776,59          |
| 1       | 479,4          | 1,9              | 540,73          |
| 2       | 69,31          | 8,8              | 2.943,38        |

Sumber: Hasil Analisis Menggunakan Python pada Google Colab (2026).

Berdasarkan Tabel 7, setiap cluster memiliki karakteristik yang berbeda berdasarkan nilai *Recency*, *Frequency*, dan *Monetary*. Perbedaan karakteristik tersebut menunjukkan adanya variasi perilaku transaksi pelanggan sehingga setiap cluster memerlukan pendekatan pemasaran yang berbeda. Oleh karena itu, masing-masing cluster diinterpretasikan sebagai Potential Customers, At-Risk Customers, dan Champions sesuai dengan pola transaksi yang ditunjukkan oleh nilai RFM.



Gambar 8. Perbandingan Nilai Rata-Rata *Recency*, *Frequency*, dan *Monetary* pada Setiap Cluster  
 Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

#### Cluster 0 (Potential Customers)

Cluster 0 (Potential Customers) memiliki nilai *Recency* yang relatif rendah serta nilai *Frequency* dan *Monetary* pada tingkat menengah. Hal ini menunjukkan bahwa pelanggan dalam cluster ini masih cukup aktif melakukan transaksi dan memiliki potensi untuk berkembang menjadi pelanggan loyal apabila perusahaan menerapkan strategi pemasaran yang tepat, seperti program loyalitas, promosi yang dipersonalisasi, atau strategi *cross-selling*.

#### Cluster 1 (At-Risk Customers)

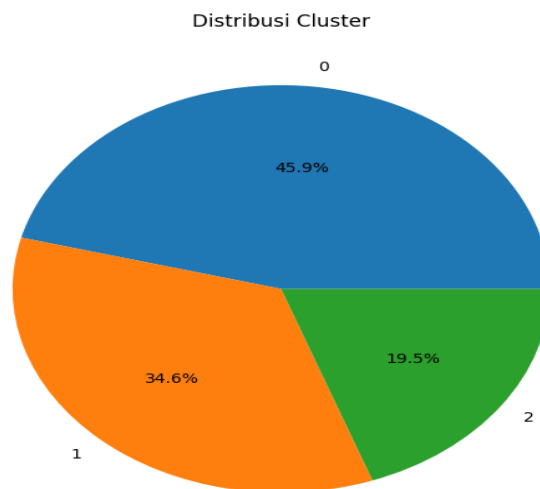
Cluster 1 (At-Risk Customers) memiliki nilai *Recency* tertinggi disertai nilai *Frequency* dan *Monetary* terendah. Kondisi tersebut menunjukkan bahwa pelanggan pada cluster ini sudah lama tidak melakukan transaksi sehingga memiliki risiko tinggi untuk berhenti menjadi pelanggan aktif (*customer churn*). Oleh karena itu, perusahaan perlu melakukan intervensi melalui program reaktivasi, seperti pemberian diskon khusus, email marketing, atau penawaran promosi yang relevan agar pelanggan kembali melakukan transaksi.

#### Cluster 2 (Champions)

Cluster 2 (Champions) memiliki nilai *Recency* terendah serta nilai *Frequency* dan *Monetary* tertinggi. Karakteristik tersebut menunjukkan bahwa pelanggan dalam cluster ini merupakan pelanggan paling loyal dan memberikan kontribusi terbesar terhadap pendapatan perusahaan. Oleh karena itu, perusahaan perlu mempertahankan pelanggan pada cluster ini melalui program loyalitas, penghargaan eksklusif, layanan yang dipersonalisasi, maupun program *referral* sehingga loyalitas pelanggan tetap terjaga.

#### 4.7 Customer Intelligence dan Rekomendasi Strategi Bisnis

Distribusi pelanggan pada setiap cluster ditunjukkan pada Gambar 8.



Gambar 9. Distribusi Pelanggan pada Setiap Cluster Hasil K-Means Clustering

Sumber : Hasil Analisis Menggunakan Python pada Google Colab (2026).

Berdasarkan hasil segmentasi yang diperoleh, setiap cluster memiliki karakteristik perilaku transaksi yang berbeda sehingga memerlukan strategi pemasaran yang berbeda pula. Informasi tersebut menjadi bagian dari Customer Intelligence yang dapat dimanfaatkan perusahaan untuk memahami karakteristik pelanggan serta mendukung penyusunan strategi pemasaran yang lebih tepat sasaran.

Tabel 8. Strategi Bisnis Berdasarkan Hasil Segmentasi Pelanggan

| Cluster             | Karakteristik                                         | Strategi Bisnis                                         |
|---------------------|-------------------------------------------------------|---------------------------------------------------------|
| Potential Customers | Pelanggan cukup aktif dengan nilai transaksi menengah | Loyalty program, cross-selling, voucher pembelian       |
| At-Risk Customers   | Pelanggan lama tidak bertransaksi                     | Kampanye reaktivasi, diskon khusus, email marketing     |
| Champions           | Pelanggan loyal dengan kontribusi tinggi              | Members hip premium, reward eksklusif, program referral |

Sumber : Hasil Analisis Penulis berdasarkan Hasil Segmentasi RFM dan K-Means (2026).

Hasil penelitian menunjukkan bahwa kombinasi metode RFM dan K-Means Clustering mampu menghasilkan segmentasi pelanggan yang efektif sehingga dapat dimanfaatkan sebagai bagian dari Customer Intelligence dalam mendukung pengambilan keputusan bisnis. Informasi mengenai karakteristik setiap segmen memungkinkan perusahaan menyusun strategi pemasaran yang lebih tepat sasaran sesuai dengan perilaku transaksi pelanggan.

Pelanggan pada segmen Champions dapat diprioritaskan melalui program loyalitas, penghargaan eksklusif, dan program *referral* untuk mempertahankan loyalitas mereka. Sementara itu, pelanggan pada segmen

*Segmentasi Pelanggan Menggunakan Metode RFM dan K-Means Clustering Untuk Mendukung Strategi Pemasaran Pada Online Retail (Selfy Anggraeni)*

Potential Customers dapat ditingkatkan menjadi pelanggan loyal melalui strategi *cross-selling*, *up-selling*, maupun promosi yang dipersonalisasi. Adapun pelanggan pada segmen At-Risk Customers memerlukan program reaktivasi, seperti email marketing, pemberian voucher atau diskon khusus, serta promosi yang disesuaikan dengan riwayat transaksi pelanggan agar kembali melakukan pembelian.

Dengan demikian, hasil segmentasi tidak hanya memberikan informasi mengenai karakteristik pelanggan, tetapi juga dapat diintegrasikan dengan sistem Customer Relationship Management (CRM) untuk mendukung pelaksanaan strategi pemasaran yang lebih efektif, meningkatkan loyalitas pelanggan, serta membantu perusahaan dalam mengambil keputusan bisnis berbasis data.

## 5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menerapkan metode *Recency*, *Frequency*, *Monetary* (RFM) dan algoritma K-Means Clustering pada Online Retail II Dataset sehingga menghasilkan tiga segmen pelanggan, yaitu Potential Customers, At-Risk Customers, dan Champions. Segmentasi tersebut memberikan gambaran mengenai karakteristik pelanggan berdasarkan nilai *Recency*, *Frequency*, dan *Monetary*, sehingga dapat dimanfaatkan sebagai dasar dalam penyusunan strategi pemasaran yang lebih tepat sasaran.

Selain memberikan kontribusi dalam menghasilkan segmentasi pelanggan, penelitian ini juga mendukung penerapan Customer Intelligence berbasis data yang dapat membantu perusahaan memahami karakteristik pelanggan, menentukan prioritas strategi pemasaran, meningkatkan loyalitas pelanggan, serta mendukung pengambilan keputusan bisnis secara lebih efektif. Dengan demikian, hasil penelitian ini memberikan manfaat praktis bagi perusahaan dalam mengalokasikan sumber daya pemasaran secara lebih efisien sesuai dengan karakteristik masing-masing segmen pelanggan.

Penelitian selanjutnya disarankan melakukan perbandingan dengan algoritma clustering lain, seperti Hierarchical Clustering atau DBSCAN, untuk mengetahui metode yang paling sesuai dengan karakteristik data transaksi *online retail*. Selain itu, penelitian berikutnya dapat menambahkan atribut lain, seperti preferensi produk, waktu pembelian, maupun data demografis pelanggan agar proses segmentasi menghasilkan karakteristik pelanggan yang lebih komprehensif. Penelitian longitudinal juga dapat dilakukan untuk menganalisis perubahan perilaku pelanggan dari waktu ke waktu sehingga strategi pemasaran dapat disusun secara lebih adaptif.

## Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan selama proses penelitian. Ucapan terima kasih juga disampaikan kepada penyedia Dataset Online Retail II melalui platform Kaggle yang digunakan dalam penelitian ini.

## DAFTAR PUSTAKA

- [1] S.-C. Necula, "Exploring the Impact of Time Spent Reading Product Information on E-Commerce Websites: A Machine Learning Approach to Analyze Consumer Behavior," *Behavioral Sciences*, vol. 13, no. 6, p. 439, 2023, doi: [10.3390/bs13060439](https://doi.org/10.3390/bs13060439).
- [2] M. Sarkar, A. R. Puja, and F. R. Chowdhury, "Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis," *J. Bus. Manag. Stud.*, vol. 6, no. 2, pp. 54–60, 2024, doi: [10.32996/jbms.2024.6.2.5](https://doi.org/10.32996/jbms.2024.6.2.5).
- [3] H. J. Wilbert et al., "Recency, Frequency, Monetary Value, Clustering, and Internal and External Indices for Customer Segmentation from Retail Data," *Algorithms*, vol. 16, no. 9, p. 396, 2023, doi: [10.3390/a16090396](https://doi.org/10.3390/a16090396).
- [4] F. Alzami et al., "Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit," *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 32–44, 2023, doi: [10.33096/ilkom.v15i1.1524.32-44](https://doi.org/10.33096/ilkom.v15i1.1524.32-44).
- [5] H. Abbasimehr and A. Bahrini, "An Analytical Framework Based on the Recency, Frequency, and Monetary Model and Time Series Clustering Techniques for Dynamic Segmentation," *Expert Syst. Appl.*, vol. 192, p. 116373, 2022, doi: [10.1016/j.eswa.2021.116373](https://doi.org/10.1016/j.eswa.2021.116373).
- [6] M. S. E. Kasem, M. Hamada, and I. Taj-Eddin, "Customer Profiling, Segmentation, and Sales Prediction Using AI in Direct Marketing," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4995–5005, 2024, doi: [10.1007/s00521-023-09339-6](https://doi.org/10.1007/s00521-023-09339-6).
- [7] A. T. Widiyanto and A. Witanti, "Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan

- Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global),” *Konstelasi Konvergensi Teknol. dan Sist. Inf.*, vol. 1, no. 1, pp. 204–215, 2021, doi: [10.24002/konstelasi.v1i1.4293](https://doi.org/10.24002/konstelasi.v1i1.4293).
- [8] P. I. Pangestu, T. I. Hermanto, and D. Irmayanti, “Analisis Segmentasi Pelanggan Berbasis Model *Recency Frequency Monetary* (RFM) Menggunakan Algoritma K-Means,” *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, 2023, doi: [10.23960/jitet.v11i3s1.3396](https://doi.org/10.23960/jitet.v11i3s1.3396).
- [9] J. M. John, O. Shobayo, and B. Ogunleye, “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market,” *Analytics*, vol. 2, no. 4, pp. 809–823, 2024, doi: [10.3390/analytics2040042](https://doi.org/10.3390/analytics2040042).
- [10] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0: Step-by-Step Data Mining Guide. Chicago, IL: SPSS Inc., 2000. [Online]. Available: <https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf>. [Accessed: Jun. 16, 2026].
- [11] S. A. Perdana, S. F. Florentin, and A. Santoso, “Analisis Segmentasi Pelanggan Menggunakan *K-Means Clustering* Studi Kasus Aplikasi Alfagift,” *Sebatik*, vol. 26, no. 2, pp. 446–457, 2022, doi: [10.46984/sebatik.v26i2.1991](https://doi.org/10.46984/sebatik.v26i2.1991).
- [12] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 4th ed., 2022.
- [13] Kotler, P., & Armstrong, G. (2021). *Principles of Marketing* (18th Global Edition). Pearson.

#### NOMENKLATUR

R = *Recency* (jarak waktu sejak transaksi terakhir pelanggan)

F = *Frequency* (jumlah transaksi yang dilakukan pelanggan)

M = *Monetary* (total nilai transaksi pelanggan)

*Quantity* = Jumlah produk yang dibeli pelanggan

*Price* = Harga produk

$d(x, C_k)$  = Jarak antara data dan centroid cluster ke-k

x = Titik data

$C_k$  = Centroid cluster ke-k

$X_i$  = Data ke-i pada cluster

n = Jumlah anggota cluster

WCSS = *Within Cluster Sum of Squares*

$a(i)$  = Rata-rata jarak data terhadap anggota pada cluster yang sama

$b(i)$  = Rata-rata jarak data terhadap cluster terdekat lainnya

$S(i)$  = Nilai *Silhouette Score* untuk data ke-i

k = Jumlah cluster