

# **ANALISIS KLASSTER PELANGGAN TOKO ONLINE PADA STUDI KASUS TOKO ERIGO**

---

**TUGAS BESAR DATA MINING**

---



Oleh

**Gading Khairlambang (714220007)**

**Ahmad Rifki Ayala (714220028)**

**Nayaka Taqwa (714220045)**

**DIPLOMA IV TEKNIK INFORMATIKA  
SEKOLAH VOKASI  
UNIVERSITAS LOGISTIK DAN BISNIS INTERNASIONAL  
BANDUNG**

## HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar. Bilamana di kemudian hari ditemukan bahwa karya tulis ini menyalahi peraturan yang ada berkaitan etika dan kaidah penulisan karya ilmiah yang berlaku, maka saya bersedia dituntut dan diproses sesuai dengan ketentuan yang berlaku

Yang menyatakan,

Nama : Ahmad Rifky Ayala

NIM : 714220028

Tanggal : 10 Juli 2025

Nama : Nayaka Taqwa

NIM : 714220045

Tanggal : 10 Juli 2025

Mengetahui

Ketua : Gading Khairlambang

Pembimbing I : Nisa Hanum Harani

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Logistik Bisnis Internasional, saya yang bertanda tangan di bawah ini:

Nama : Gading Khairlambang

NIM : 714220007

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Logistik Bisnis Internasional, Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalti Free Right*) atas karya ilmiah saya yang berjudul:

### **ANALISIS KLASSTER PELANGGAN TOKO ONLINE PADA STUDI KASUS TOKO ERIGO**

Beserta perangkat yang ada (jika diperlukan). Dengan Hak ini Universitas Logistik Bisnis Internasional Hayati berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Gading Khairlambang

Pada tanggal : 10 Juli 2025

Yang menyatakan

(.....)

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Logistik Bisnis Internasional, saya yang bertanda tangan di bawah ini:

Nama : Ahmad Rifki Ayala

NIM : 714220028

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Logistik Bisnis Internasional, Hak Bebas Royalti Noneksklusif (*Non- exclusive Royalti Free Right*) atas karya ilmiah saya yang berjudul:

### **ANALISIS KLASSTER PELANGGAN TOKO ONLINE PADA STUDI KASUS TOKO ERIGO**

Beserta perangkat yang ada (jika diperlukan). Dengan Hak ini Universitas Logistik Bisnis Internasional Hayati berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Ahmad Rifki Ayala

Pada tanggal : 10 Juli 2025

Yang menyatakan

(.....)

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Logistik Bisnis Internasional, saya yang bertanda tangan di bawah ini:

Nama : Nayaka Taqwa

NIM : 714220045

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Logistik Bisnis Internasional, Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalti Free Right*) atas karya ilmiah saya yang berjudul:

### **ANALISIS KLASSTER PELANGGAN TOKO ONLINE PADA STUDI KASUS TOKO ERIGO**

Beserta perangkat yang ada (jika diperlukan). Dengan Hak ini Universitas Logistik Bisnis Internasional Hayati berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Gading Khairlambang

Pada tanggal : 10 Juli 2025

Yang menyatakan

(.....)

## **ABSTRAK**

Pertumbuhan pesat e-commerce di era digital menghasilkan volume data besar, termasuk data ulasan pelanggan yang seringkali belum dimanfaatkan secara optimal. Ulasan ini dapat merepresentasikan perilaku, preferensi, dan nilai pelanggan terhadap suatu produk atau layanan. Penelitian ini menerapkan pendekatan data mining, khususnya metode clustering, untuk melakukan segmentasi pelanggan berdasarkan model RFM (Recency, Frequency, Monetary) yang diadaptasi dari data ulasan produk pada platform e-commerce Tokopedia. Dua algoritma clustering yang digunakan adalah K-Means dan Agglomerative Clustering. Penelitian ini bertujuan untuk mengidentifikasi kelompok pelanggan yang memiliki karakteristik serupa guna mendukung strategi pemasaran yang lebih terarah. Hasil segmentasi diharapkan dapat membantu pelaku bisnis memahami perilaku pelanggan secara lebih mendalam meskipun tanpa akses langsung ke data transaksi.

## **ABSTRACT**

The rapid growth of e-commerce in the digital era has generated massive volumes of data, including customer reviews that are often underutilized. These reviews can represent customer behavior, preferences, and value toward a product or service. This study applies a data mining approach, specifically clustering methods, to perform customer segmentation based on the RFM (Recency, Frequency, Monetary) model adapted from product review data on the Tokopedia e-commerce platform. Two clustering algorithms are used: K-Means and Agglomerative Clustering. The goal of this research is to identify customer groups with similar characteristics to support more targeted marketing strategies. The resulting segmentation is expected to help businesses gain deeper insights into customer behavior, even without direct access to transaction data.

# DAFTAR ISI

ANALISIS KLASTER PELANGGAN TOKO ONLINE PADA STUDI KASUS TOKO ERIGO .....	i
HALAMAN PERNYATAAN ORISINALITAS .....	i
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS .....	ii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS .....	iii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS .....	iv
ABSTRAK.....	v
ABSTRACT .....	vi
DAFTAR ISI .....	vii
DAFTAR TABEL .....	ix
DAFTAR GAMBAR.....	x
BAB 1 PENDAHULUAN.....	1
<b>1.1    LATAR BELAKANG</b> .....	1
<b>1.2    IDENTIFIKASI MASALAH</b> .....	2
<b>1.3    TUJUAN PENELITIAN</b> .....	3
<b>1.4    MANFAAT PENELITIAN</b> .....	3
<b>1.5    RUANG LINGKUP</b> .....	3
BAB 2 TINJAUAN PUSTAKA .....	5
<b>2.1    RUANG LINGKUP</b> .....	5
2.1.1    Data Mining dan Machine Learning .....	5
2.1.2    Teknik Clustering .....	5
2.1.3    MODEL RFM.....	6
<b>2.2    VISUALISASI</b> .....	6
<b>2.3    STATE OF THE ART</b> .....	7
BAB 3 METODOLOGI PENELITIAN .....	8
<b>3.1    METODOLOGI PENELITIAN</b> .....	8
3.1.1    AKUISISI DATA .....	8
3.1.2    PREPROCESSING DATA.....	8
3.1.3    REKAYASA FITUR .....	9
3.1.4    PEMODELAN.....	9
3.1.5    EVALUASI .....	10
<b>3.2    DESKRIPSI DATASET</b> .....	10
3.2.1    DESKRIPSI ATRIBUT .....	10
<b>3.3    ALGORITMA DAN TOOLS DATA MINING</b> .....	11
3.3.1    ALASAN PEMILIHAN ALGORITMA .....	11



3.3.2	ALAT DAN PUSTAKA PENDUKUNG .....	11
<b>3.4</b>	<b>Evaluasi Kinerja.....</b>	<b>12</b>
<b>BAB 4</b>	<b>HASIL DAN PENELITIAN .....</b>	<b>14</b>
<b>4.1</b>	<b>VISUALISASI EKSPORATIF (EDA) .....</b>	<b>14</b>
<b>4.2</b>	<b>VISUALISASI EKSPORATIF (EDA) .....</b>	<b>16</b>
4.2.1	PREPROCESSING.....	16
4.2.2	PEMODELAN.....	19
<b>4.3</b>	<b>EVALUASI MODEL .....</b>	<b>22</b>
<b>BAB 5</b>	<b>KESIMPULAN DAN SARAN .....</b>	<b>23</b>
	DAFTAR PUSTAKA.....	24

## DAFTAR TABEL

Tabel 3. 1 Deskripsi Atribut .....	11
Tabel 4. 1 Hasil Evaluasi dua model .....	22

## DAFTAR GAMBAR

Gambar 2. 1 Diagram Alir Konsep Penelitian.....	6
Gambar 3. 1 Nilai Metode Evaluasi .....	12
Gambar 4. 1 Distribusi Jumlah Terjual dan frekuensi .....	14
Gambar 4. 2 Distribusi Harga Produk dan Frekuensi terjual.....	15
Gambar 4. 3 Top 10 Pengguna Paling Aktif dari komentar .....	15
Gambar 4. 4 Heatmap Korelasi antar feature RFM .....	17
Gambar 4. 5 PairPlot Hubungan antar feature.....	18
Gambar 4. 6 Penentuan clustering dengan penerapan 4 metode .....	19
Gambar 4. 7 clustering model K-Means.....	20
Gambar 4. 8 clustering model Agglomerative.....	21

# **BAB 1**

## **PENDAHULUAN**

### **1.1 LATAR BELAKANG**

Di era digital saat ini, pertumbuhan e-commerce sangat pesat dan menghasilkan volume data yang luar biasa dari berbagai aktivitas pengguna. Salah satu sumber data yang paling bernilai namun sering diabaikan adalah ulasan produk dari pelanggan. Ulasan ini tidak hanya mencerminkan pengalaman dan kepuasan pelanggan, tetapi juga menyimpan informasi tersembunyi mengenai preferensi, kebiasaan, dan pola perilaku konsumen. Namun, data mentah dalam bentuk teks ulasan atau meta-data terkait tidak serta-merta memberikan wawasan yang langsung bermanfaat. Diperlukan teknik analisis data yang tepat untuk mengolahnya menjadi informasi yang berguna. Di sinilah pentingnya peran data mining[1].

Data mining adalah proses mengekstraksi informasi baru dan berguna dari kumpulan data yang besar. Dalam konteks e-commerce, penerapan data mining dapat membantu pelaku bisnis memahami perilaku konsumen dan membuat keputusan strategis, seperti dalam merancang promosi yang tepat sasaran, meningkatkan kepuasan pelanggan, hingga mengoptimalkan layanan. Salah satu aplikasi data mining yang relevan dalam konteks ulasan pelanggan adalah segmentasi pelanggan, yaitu proses pengelompokan pelanggan berdasarkan pola tertentu dalam data, seperti intensitas ulasan, frekuensi ulasan, sentimen, atau waktu berkomentar[2].

Teknik yang sering digunakan dalam segmentasi pelanggan adalah clustering, yaitu metode dalam unsupervised learning yang mengelompokkan data berdasarkan kesamaan atribut tanpa memerlukan label. Algoritma populer dalam clustering meliputi K-Means dan Agglomerative Clustering, yang merupakan bagian dari hierarchical clustering, juga dapat digunakan sebagai alternatif untuk mengidentifikasi struktur kelompok pelanggan berdasarkan kemiripan tertentu. Berbeda dengan K-Means yang bekerja berdasarkan centroid, agglomerative membangun hierarki klaster melalui penggabungan bertahap dari pasangan klaster yang paling mirip. Hal ini memungkinkan peneliti mengeksplorasi segmentasi pelanggan pada berbagai tingkat granularitas[3].

Dalam penelitian ini, data yang digunakan bersumber dari ulasan produk pada toko Erigo di platform Tokopedia, seperti teks ulasan, username, tanggal komentar, nama produk, dan harga produk. Berdasarkan data ini, pendekatan segmentasi dilakukan

dengan mengadaptasi konsep RFM (Recency, Frequency, Monetary). Meskipun data transaksi secara langsung tidak tersedia, elemen RFM tetap dapat diterapkan: Recency dihitung dari waktu terakhir pelanggan memberikan ulasan, Frequency dari seberapa sering pelanggan memberikan ulasan, dan Monetary dapat didekati dari total estimasi harga produk yang diulas oleh pelanggan. Dengan pendekatan ini, algoritma clustering dapat digunakan untuk mengelompokkan pelanggan berdasarkan nilai dan aktivitasnya, yang pada akhirnya memberikan gambaran strategis dalam memahami perilaku konsumen dan menyusun strategi pemasaran yang lebih efektif[4].

## **1.2 IDENTIFIKASI MASALAH**

Seiring dengan meningkatnya penggunaan toko online, pelaku usaha menghadapi tantangan dalam memahami perilaku pelanggan yang semakin beragam. Tanpa analisis yang tepat, strategi pemasaran berisiko tidak efektif karena tidak menargetkan kelompok pelanggan yang sesuai. Di sisi lain, banyak data potensial yang tersedia di platform e-commerce, seperti ulasan produk, yang sering kali belum dimanfaatkan secara optimal.

Ulasan produk yang diberikan oleh pelanggan menyimpan informasi penting mengenai keterlibatan mereka, frekuensi interaksi, serta preferensi terhadap jenis produk tertentu. Dengan memanfaatkan data ini, pelaku usaha dapat memperoleh wawasan perilaku pelanggan meskipun tanpa akses langsung ke data transaksi. Oleh karena itu, penting bagi toko online untuk melakukan segmentasi pelanggan berdasarkan pola aktivitas mereka dalam memberikan ulasan agar dapat menyusun strategi pemasaran yang lebih terarah dan personal.

Berdasarkan latar belakang tersebut, maka permasalahan yang dapat diidentifikasi adalah sebagai berikut:

1. Bagaimana toko online dapat memanfaatkan data ulasan pelanggan untuk memahami perilaku belanja mereka?
2. Bagaimana penerapan metode data mining, khususnya algoritma clustering, dapat membantu dalam mengelompokkan pelanggan berdasarkan aktivitas ulasan mereka?
3. Bagaimana pendekatan model RFM dapat diadaptasi dari data ulasan untuk mendukung proses segmentasi pelanggan secara akurat?

### **1.3 TUJUAN PENELITIAN**

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengidentifikasi cara pemanfaatan data ulasan produk untuk memahami karakteristik dan perilaku pelanggan pada toko online.
2. Menerapkan algoritma clustering sebagai teknik data mining untuk melakukan segmentasi pelanggan berdasarkan hasil ekstraksi atribut RFM dari data ulasan.
3. Mengembangkan pendekatan adaptasi model RFM berbasis data ulasan produk (termasuk waktu ulasan, frekuensi ulasan, dan estimasi nilai produk) untuk menghasilkan segmentasi pelanggan yang akurat dan informatif.

### **1.4 MANFAAT PENELITIAN**

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Bagi pelaku bisnis e-commerce, penelitian ini dapat menjadi referensi dalam memanfaatkan data ulasan pelanggan sebagai alternatif data transaksi dalam upaya segmentasi pelanggan dan perumusan strategi pemasaran yang lebih personal.
2. Bagi pengembang sistem data mining, penelitian ini memberikan wawasan tentang implementasi clustering pada data non-transaksional, khususnya ulasan pelanggan, dengan pendekatan model RFM yang telah disesuaikan.
3. Bagi dunia akademik, penelitian ini menambah khazanah literatur terkait penerapan data mining dalam konteks e-commerce, khususnya pengembangan teknik segmentasi berbasis data ulasan publik yang tersedia secara terbuka.

### **1.5 RUANG LINGKUP**

Ruang lingkup penelitian ini dibatasi pada hal-hal berikut:

1. Jenis data yang digunakan terbatas pada data hasil scraping dari halaman produk e-commerce, yang mencakup: nama produk, harga produk, teks ulasan, username pengulas, dan tanggal ulasan.
2. Pendekatan analisis perilaku pelanggan difokuskan pada model RFM yang diadaptasi dari data ulasan, tanpa melibatkan data transaksi aktual atau atribut tambahan seperti discount proportion.

3. Teknik data mining yang digunakan terbatas pada algoritma clustering, seperti K-Means dan Agglomerative Clustering, untuk tujuan segmentasi pelanggan.
4. Evaluasi hasil segmentasi dilakukan dengan menggunakan metrik evaluasi clustering, seperti Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index, untuk menilai kualitas pengelompokan yang dihasilkan dari masing-masing algoritma.

## **BAB 2**

### **TINJAUAN PUSTAKA**

#### **2.1 RUANG LINGKUP**

##### **2.1.1 Data Mining dan Machine Learning**

Data mining adalah proses untuk menemukan pola tersembunyi dan pengetahuan baru dari data berukuran besar[1]. Dalam konteks e-commerce, data mining digunakan untuk memahami perilaku pelanggan, meningkatkan pengalaman pengguna, dan mendukung pengambilan keputusan[4]. Salah satu cabang dari data mining adalah machine learning, yang memungkinkan sistem belajar dari data dan membuat prediksi atau klasifikasi tanpa diprogram secara eksplisit[2].

##### **2.1.2 Teknik Clustering**

Clustering merupakan salah satu metode dalam unsupervised learning yang bertujuan untuk mengelompokkan data ke dalam beberapa kluster berdasarkan kesamaan fitur atau karakteristik tanpa memerlukan label data. Teknik ini sangat efektif untuk segmentasi pelanggan karena dapat mengidentifikasi pola perilaku konsumen yang serupa dalam suatu kelompok.

Dalam penelitian ini, algoritma clustering yang digunakan terbatas pada:

- **K-Means Clustering**

K-Means bekerja dengan membagi data ke dalam k kluster berdasarkan jarak ke pusat kluster (centroid). Algoritma ini populer karena kecepatan dan kesederhanaannya, tetapi kurang ideal jika data mengandung outlier atau kluster dengan ukuran yang tidak merata[5], [6].

- **Agglomerative Clustering**

Agglomerative Clustering adalah salah satu metode hierarchical clustering dengan pendekatan bottom-up. Setiap data awalnya dianggap sebagai satu kluster tersendiri, lalu secara bertahap digabungkan berdasarkan kedekatan hingga terbentuk hierarki kluster. Metode ini cocok untuk eksplorasi struktur data tanpa menentukan jumlah kluster di awal dan memberikan fleksibilitas dalam analisis multilevel[5], [6].



### 2.1.3 MODEL RFM

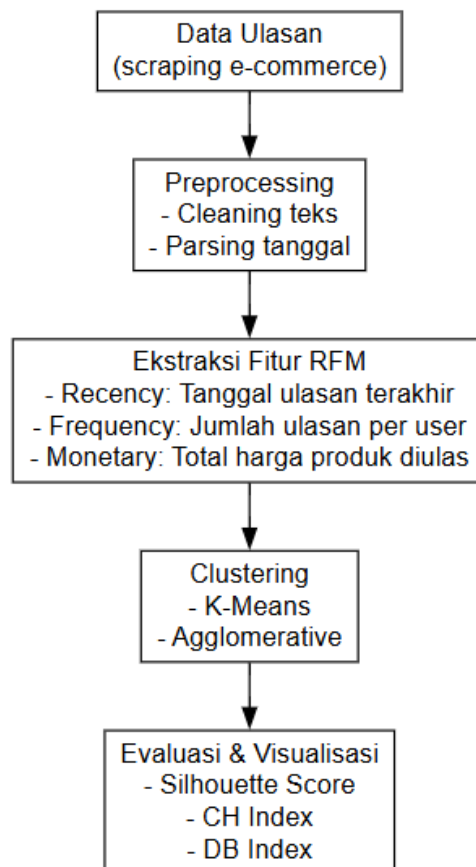
Model RFM adalah teknik yang digunakan untuk menilai pelanggan berdasarkan tiga dimensi:

- Recency : Seberapa baru interaksi terakhir pelanggan.
- Frequency : Seberapa sering pelanggan melakukan interaksi atau pembelian.
- Monetary : Seberapa besar jumlah pembelian yang dilakukan.

Model ini telah terbukti efektif dalam mengidentifikasi pelanggan bernilai tinggi dan dalam menyusun strategi pemasaran yang lebih terarah[7].

## 2.2 VISUALISASI

Diagram Alir Konsep Penelitian



Gambar 2. 1 Diagram Alir Konsep Penelitian

## 2.3 STATE OF THE ART

Penelitian segmentasi pelanggan berbasis data mining kini berkembang dengan memanfaatkan data non-transaksional seperti ulasan produk, yang dinilai mampu mencerminkan perilaku dan preferensi pelanggan. Model RFM yang sebelumnya berbasis transaksi telah banyak diadaptasi dalam konteks ini, dengan Recency diambil dari tanggal ulasan terakhir, Frequency dari jumlah ulasan, dan Monetary dari estimasi harga produk yang diulas. Di sisi algoritma, K-Means tetap menjadi pilihan populer karena kesederhanaannya, sementara Agglomerative Clustering mulai digunakan karena kemampuannya membentuk hierarki kluster tanpa penentuan jumlah kluster di awal. Tren saat ini juga menunjukkan pentingnya penggunaan metrik evaluasi seperti Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index untuk menilai kualitas segmentasi. Penelitian ini mengikuti arah perkembangan tersebut dengan menerapkan pendekatan RFM berbasis ulasan dan membandingkan efektivitas dua algoritma clustering tersebut dalam segmentasi pelanggan e-commerce

## **BAB 3**

### **METODOLOGI PENELITIAN**

#### **3.1 METODOLOGI PENELITIAN**

Penelitian ini menggunakan pendekatan data mining berbasis proses CRISP-DM (*Cross Industry Standard Process for Data Mining*), sebuah metodologi yang umum digunakan dalam proyek-proyek pengolahan data untuk menemukan pola dan pengetahuan yang tersembunyi dari kumpulan data besar. CRISP-DM terdiri dari enam fase utama, namun dalam penelitian ini difokuskan pada lima fase inti yang disesuaikan dengan kebutuhan eksplorasi dan segmentasi pelanggan menggunakan metode RFM (Recency, Frequency, Monetary) dan algoritma K-Means clustering. Adapun tahapan-tahapan tersebut dijelaskan pada sub-subbab berikut ini:

##### **3.1.1 AKUISISI DATA**

Tahap awal dimulai dengan proses pengumpulan data yang bersumber dari transaksi pelanggan perusahaan retail fashion Erigo. Dataset ini diperoleh dalam bentuk file CSV yang berisi informasi penting terkait transaksi, termasuk ID pelanggan, tanggal transaksi, total nilai pembelian, dan nomor invoice. Data tersebut diimpor menggunakan pustaka Python seperti pandas untuk selanjutnya dianalisis. Dataset yang digunakan terdiri dari sekitar 585 baris data transaksi dengan total 12 kolom, sehingga dianggap representatif untuk menggambarkan perilaku pelanggan secara menyeluruh.

##### **3.1.2 PREPROCESSING DATA**

Setelah data berhasil diperoleh, tahap selanjutnya adalah proses *data preprocessing*, yaitu proses pembersihan dan transformasi data agar siap digunakan untuk pemodelan. Langkah-langkah yang dilakukan pada tahap ini antara lain:

- Menghapus duplikat data dan data yang tidak relevan.
- Menangani nilai kosong atau *missing value* yang berpotensi mengganggu analisis.
- Mengubah format tanggal dari string ke format datetime agar dapat dilakukan perhitungan Recency.

- Menentukan tanggal acuan (*snapshot date*) sebagai basis dalam penghitungan nilai RFM untuk setiap pelanggan.
- Mengelompokkan data berdasarkan CustomerID untuk menghitung frekuensi transaksi, total belanja, dan waktu transaksi terakhir.

Transformasi dilakukan agar data memiliki struktur yang seragam, bersih, dan relevan untuk tahap rekayasa fitur.

### 3.1.3 REKAYASA FITUR

Dalam tahap ini, dilakukan proses penghitungan tiga indikator utama dari metode RFM, yaitu:

- **Recency:** Mengukur berapa hari sejak transaksi terakhir pelanggan hingga tanggal acuan.
- **Frequency:** Mengukur berapa kali pelanggan melakukan transaksi dalam rentang waktu tertentu.
- **Monetary:** Mengukur seberapa besar jumlah uang yang telah dibelanjakan pelanggan selama periode tersebut.

Nilai-nilai RFM ini merupakan representasi dari perilaku pelanggan terhadap layanan perusahaan. Ketiga fitur ini kemudian dinormalisasi menggunakan metode Min-Max Scaling agar skala antar fitur seragam dan tidak mendominasi satu sama lain saat dilakukan proses klusterisasi.

### 3.1.4 PEMODELAN

Pemodelan dilakukan menggunakan algoritma unsupervised learning yaitu **K-Means Clustering**, yang bertujuan untuk mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan kemiripan perilaku belanjanya. Sebelum menentukan jumlah cluster ( $k$ ) terbaik, dilakukan proses evaluasi menggunakan metrik internal yaitu Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index. Evaluasi tersebut digunakan untuk menentukan nilai  $k$  yang optimal sehingga menghasilkan segmentasi yang jelas dan representatif.

Dari hasil evaluasi, dipilih jumlah cluster optimal (misalnya  $k = 4$ ) yang kemudian digunakan dalam proses pelatihan model. Model ini kemudian membagi seluruh pelanggan ke dalam 4 kelompok dengan karakteristik RFM yang berbeda, seperti pelanggan loyal, pelanggan baru, atau pelanggan yang hampir tidak aktif.

### 3.1.5 EVALUASI

Tahap evaluasi dilakukan untuk mengukur seberapa baik model klasterisasi yang terbentuk. Evaluasi dilakukan menggunakan tiga metrik utama, yaitu:

- **Silhouette Score:** Mengukur seberapa mirip suatu data dengan cluster-nya sendiri dibandingkan dengan cluster lain. Nilai yang diperoleh sebesar **0.3936**, yang mengindikasikan bahwa hasil klaster cukup baik meskipun tidak sempurna.
- **Calinski-Harabasz Index:** Mengukur rasio antara variansi antar cluster dan variansi dalam cluster. Semakin tinggi nilainya, semakin baik klaster. Nilai yang diperoleh sebesar **243.2030**.
- **Davies-Bouldin Index:** Mengukur rata-rata kemiripan antara masing-masing klaster. Nilai yang diperoleh sebesar **0.9602**, yang mendekati angka 0, menunjukkan kualitas pemisahan antar cluster yang cukup baik.

## 3.2 DESKRIPSI DATASET

Dataset yang digunakan dalam penelitian ini merupakan kumpulan data transaksi pelanggan dari perusahaan Erigo yang mencakup informasi penting terkait perilaku belanja pelanggan. Dataset terdiri dari **585 baris data** dan **12 kolom atribut**, yang masing-masing mewakili informasi seperti ID pelanggan, tanggal transaksi, nilai pembelian, dan hasil transformasi fitur seperti Recency, Frequency, dan Monetary.

### 3.2.1 DESKRIPSI ATRIBUT

Berikut adalah penjelasan beberapa atribut utama dalam dataset:

Atribut	Tipe Data	Deskripsi
CustomerID	String	ID unik yang merepresentasikan pelanggan
InvoiceNo	String	Nomor unik faktur transaksi
InvoiceDate	DateTime	Tanggal dan waktu transaksi
TotalAmount	Float	Nilai pembelanjaan pada setiap transaksi
Recency	Integer	Jumlah hari sejak transaksi terakhir pelanggan
Frequency	Integer	Jumlah transaksi yang

		dilakukan pelanggan
Monetary	Float	Total uang yang dibelanjakan pelanggan

*Tabel 3. 1 Deskripsi Atribut*

Dataset ini telah melalui tahap transformasi dan normalisasi sebelum digunakan dalam proses pemodelan.

### 3.3 ALGORITMA DAN TOOLS DATA MINING

Pada tahap ini, pemilihan algoritma dan tools yang tepat menjadi aspek penting dalam menunjang keberhasilan proses data mining. Algoritma digunakan untuk membentuk model yang mampu mengelompokkan data berdasarkan karakteristik tertentu, sedangkan tools dan pustaka pemrograman berperan sebagai sarana pendukung dalam pengolahan data, visualisasi, hingga evaluasi model. Dalam penelitian ini, algoritma yang digunakan adalah K-Means Clustering, yang sangat cocok untuk melakukan segmentasi berdasarkan data numerik seperti nilai Recency, Frequency, dan Monetary. Selain itu, penelitian ini juga menggunakan beberapa pustaka Python yang umum digunakan dalam analisis data dan machine learning, seperti pandas, numpy, matplotlib, seaborn, dan scikit-learn.

#### 3.3.1 ALASAN PEMILIHAN ALGORITMA

Algoritma K-Means Clustering dipilih dalam penelitian ini karena memiliki beberapa keunggulan, antara lain kemampuan untuk mengelompokkan data dalam jumlah besar secara efisien, kemudahan implementasi, serta cocok untuk data numerik seperti fitur RFM. Selain itu, K-Means juga memberikan hasil klaster yang dapat diinterpretasikan secara visual maupun deskriptif, yang sangat bermanfaat dalam proses segmentasi pelanggan.

#### 3.3.2 ALAT DAN PUSTAKA PENDUKUNG

Proses analisis dan pemodelan dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka sebagai berikut:

- pandas dan numpy untuk manipulasi dan transformasi data
- matplotlib dan seaborn untuk visualisasi
- scikit-learn untuk pemodelan dan evaluasi klasterisasi
- Jupyter Notebook sebagai platform pengembangan berbasis interaktif

### 3.4 Evaluasi Kinerja

Evaluasi dilakukan untuk mengetahui sejauh mana model K-Means mampu membentuk kluster yang baik dalam mengelompokkan pelanggan berdasarkan perilaku belanjanya. Evaluasi ini penting untuk memastikan bahwa segmentasi yang dihasilkan benar-benar mencerminkan perbedaan karakteristik antar kelompok pelanggan, sehingga dapat digunakan sebagai dasar dalam strategi bisnis yang lebih tepat sasaran.

Penilaian performa model dilakukan menggunakan tiga metrik evaluasi yang umum digunakan dalam klasterisasi data, yaitu **Silhouette Score**, **Calinski-Harabasz Index**, dan **Davies-Bouldin Index**. Ketiga metrik ini dipilih karena masing-masing memiliki pendekatan yang berbeda dalam menilai kualitas pemisahan antar kluster dan kepadatan dalam kluster. Berikut adalah hasil evaluasi model K-Means yang diterapkan:

	Metode Evaluasi	Nilai
0	Silhouette Score	0.3936
1	Calinski-Harabasz Index	243.2030
2	Davies-Bouldin Index	0.9363

Gambar 3. 1 Nilai Metode Evaluasi

Hasil evaluasi menunjukkan bahwa model K-Means memberikan segmentasi pelanggan yang cukup baik. Interpretasi dari masing-masing metrik evaluasi dijelaskan sebagai berikut:

- **Silhouette Score** sebesar 0.3936 menunjukkan bahwa sebagian besar data berada cukup dekat dengan kluster yang sesuai dan relatif jauh dari kluster lainnya. Nilai ini mengindikasikan bahwa kluster yang terbentuk memiliki batas yang cukup jelas, meskipun masih terdapat beberapa titik data yang bisa saja berada pada posisi ambigu.
- **Calinski-Harabasz Index** menghasilkan nilai sebesar 243.2030. Nilai ini cukup tinggi, yang menunjukkan bahwa variasi antar kluster jauh lebih besar dibandingkan dengan variasi di dalam kluster. Dengan kata lain, kluster yang terbentuk cukup kompak di dalam dan terpisah dengan baik dari kluster lain.
- **Davies-Bouldin Index** bernilai 0.9363, yang mendekati nol. Nilai ini menandakan bahwa kluster memiliki tingkat kemiripan internal yang rendah dan keterpisahan antar kluster yang baik. Semakin kecil nilai Davies-Bouldin Index, semakin baik model dalam membedakan antar kluster.

Seluruh proses evaluasi ini dilakukan menggunakan pustaka scikit-learn dalam Python, yang menyediakan fungsi-fungsi bawaan untuk menghitung nilai-nilai metrik tersebut secara efisien

dan akurat.

Berdasarkan hasil evaluasi ini, dapat disimpulkan bahwa model K-Means yang digunakan dalam penelitian ini cukup efektif dalam mengelompokkan pelanggan berdasarkan dimensi RFM. Model ini memberikan segmentasi yang dapat diandalkan untuk digunakan dalam analisis perilaku pelanggan serta sebagai dasar dalam pengambilan keputusan pemasaran yang lebih tepat dan strategis.

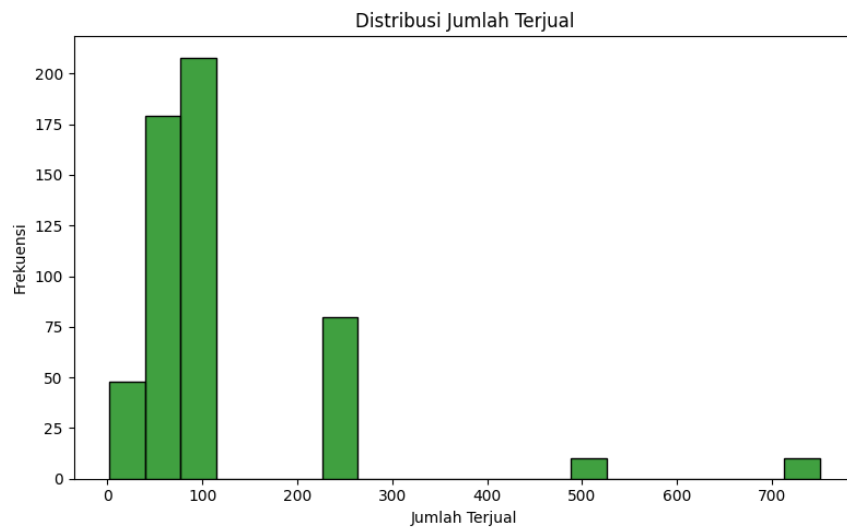


## BAB 4

### HASIL DAN PENELITIAN

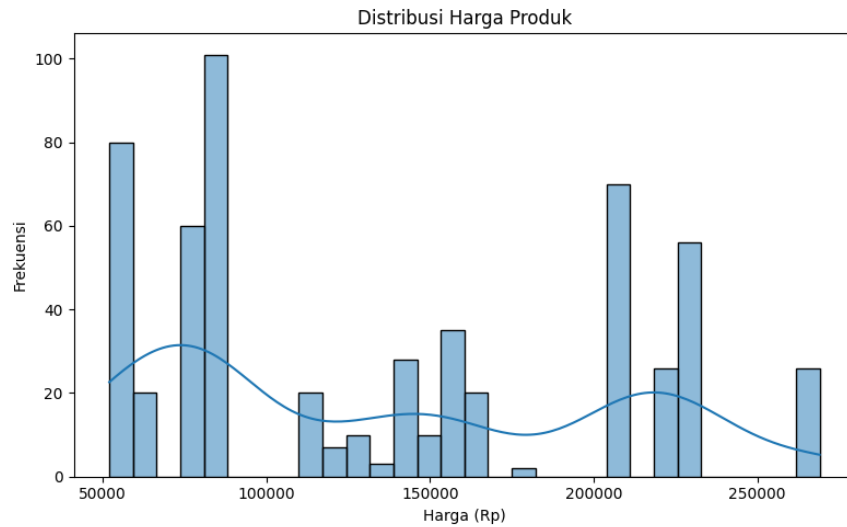
#### 4.1 VISUALISASI EKSPORATIF (EDA)

Pada studi kasus tugas besar ini proses analisis data, dilakukan dengan penerapan eksplorasi data (Exploratory Data Analysis/EDA) untuk memahami pola, distribusi, serta karakteristik awal dari dataset pelanggan toko online. Tahap ini bertujuan untuk mengidentifikasi kemungkinan anomali, kecenderungan perilaku pelanggan, dan dasar yang kuat sebelum membangun model klasterisasi. Visualisasi dilakukan terhadap fitur-fitur penting seperti harga produk, jumlah penjualan, dan aktivitas pengguna dalam memberikan ulasan.



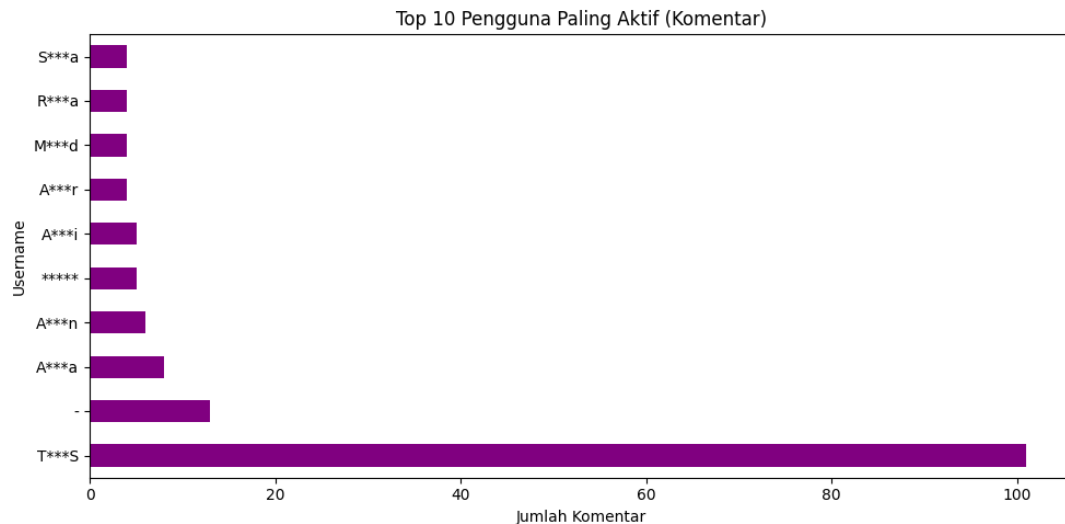
Gambar 4. 1 Distribusi Jumlah Terjual dan frekuensi

Pada Gambar 4.1 terlihat bahwa distribusi jumlah produk yang terjual cenderung condong ke kiri dengan mayoritas nilai berada di antara 50 hingga 100 unit terjual. Terlihat juga adanya beberapa outlier, seperti produk dengan penjualan mencapai lebih dari 700 unit, yang menunjukkan adanya produk dengan performa penjualan yang sangat tinggi. Pola ini menunjukkan bahwa sebagian besar produk terjual dalam jumlah sedang, sedangkan hanya sebagian kecil yang menjadi produk unggulan dengan penjualan luar biasa tinggi.



Gambar 4. 2 Distribusi Harga Produk dan Frekuensi terjual

Gambar 4.2 menampilkan distribusi harga produk yang ditawarkan oleh toko. Distribusi ini menunjukkan bahwa harga produk berkisar antara Rp50.000 hingga Rp270.000, dengan beberapa titik puncak (multimodal) yang mengindikasikan adanya segmentasi harga berdasarkan kategori atau jenis produk. Harga tertinggi tidak mendominasi jumlah frekuensi, yang mengisyaratkan bahwa produk dengan harga menengah lebih banyak diminati atau ditawarkan. Pola ini penting untuk memahami bagaimana harga mempengaruhi daya beli dan perilaku pelanggan.



Gambar 4. 3 Top 10 Pengguna Paling Aktif dari komentar

Gambar 4.3 memperlihatkan sepuluh pengguna yang paling aktif dalam memberikan komentar atau ulasan terhadap produk. Secara mencolok, pengguna dengan inisial T\*S\*\* memiliki jumlah komentar yang sangat tinggi, melebihi 100 komentar, jauh di atas pengguna

lain yang berada pada kisaran belasan komentar. Hal ini menunjukkan adanya kelompok pelanggan dengan tingkat keterlibatan yang sangat tinggi, yang dapat dianggap sebagai pelanggan loyal atau power user.

Namun demikian, perlu dicatat bahwa banyak nama pengguna yang ditampilkan dalam kondisi tersensor atau disamarkan (contohnya berupa inisial seperti A\*\*\*a, R\*\*\*i, dll). Penyensoran ini merupakan bentuk perlindungan privasi dari platform, tetapi secara teknis dapat menyebabkan duplikasi identitas atau ambiguitas. Misalnya, dua pengguna berbeda bisa saja memiliki inisial sensor yang sama, sehingga jumlah komentar yang tercatat bisa tidak sepenuhnya akurat. Hal ini menjadi catatan penting dalam interpretasi data pengguna, terutama jika analisis lanjutan membutuhkan identifikasi unik secara konsisten.

## 4.2 VISUALISASI EKSPORATIF (EDA)

Pada pengembangan model untuk klastering pelanggan toko online terdapat beberapa tahapan yaitu Proses ini diawali dengan preprocessing data, yaitu serangkaian langkah transformasi untuk membersihkan, merapikan, serta menyusun ulang data mentah agar dapat dianalisis secara optimal. Salah satu pendekatan yang digunakan dalam tahap ini adalah pembentukan fitur **RFM (Recency, Frequency, Monetary)** yang merepresentasikan perilaku pelanggan berdasarkan histori interaksinya dengan toko.

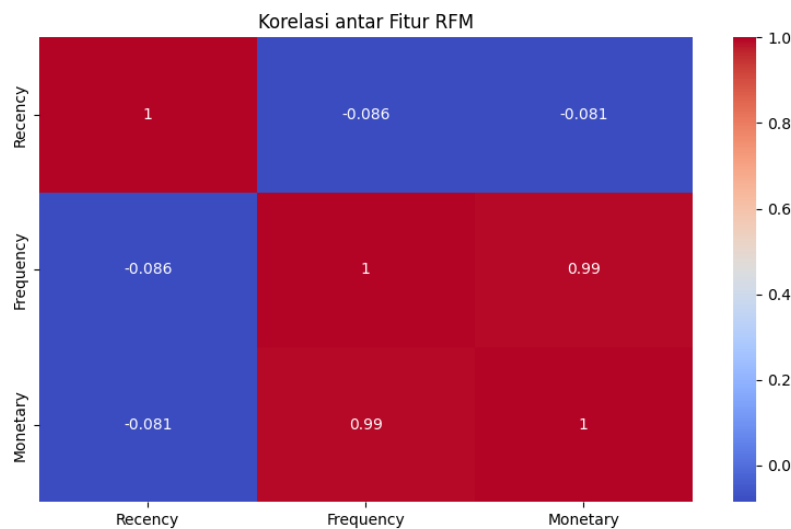
Setelah data dipersiapkan, langkah berikutnya adalah membangun model klasterisasi yang bertujuan untuk mengelompokkan pelanggan ke dalam segmen-segmen homogen berdasarkan perilaku belanja mereka. Penentuan jumlah klaster dilakukan menggunakan berbagai metrik evaluasi seperti Elbow Method, Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index. Pemodelan klasterisasi kemudian diterapkan dengan algoritma **K-Means** dan **Agglomerative Clustering** sebagai pendekatan pembandingan. Setiap model dianalisis dan divalidasi untuk memperoleh struktur segmentasi pelanggan yang paling optimal dan interpretable.

### 4.2.1 PREPROCESSING

Sebelum dilakukan proses klasterisasi terhadap pelanggan, dilakukan tahap preprocessing untuk menyiapkan data agar siap dianalisis dan membentuk representasi perilaku pelanggan yang valid. Pada penelitian ini, digunakan pendekatan **RFM (Recency, Frequency, Monetary)** untuk merepresentasikan karakteristik interaksi

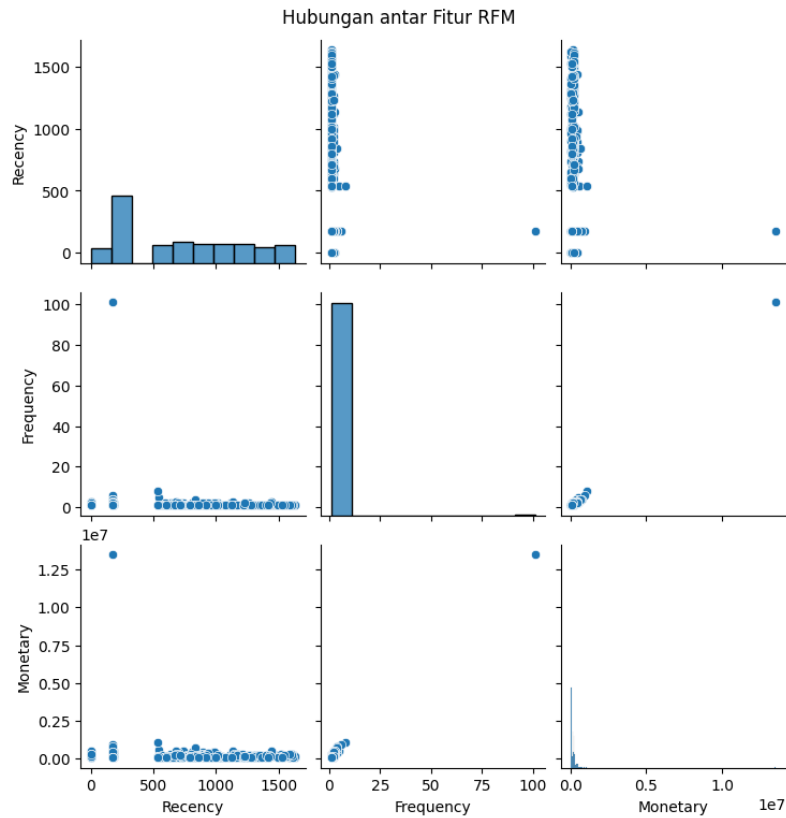
pelanggan terhadap toko. Fitur utama yang digunakan dalam pembentukan RFM berasal dari tiga elemen penting: **tanggal komentar**, **username pelanggan**, dan **harga produk**.

- **Recency** dihitung sebagai selisih hari antara tanggal komentar terakhir pelanggan dengan tanggal terakhir di dataset.
- **Frequency** dihitung berdasarkan jumlah komentar atau transaksi yang dilakukan oleh masing-masing pelanggan (berdasarkan Username).
- **Monetary** direpresentasikan dari total nilai produk yang dibeli oleh pelanggan, dengan menjumlahkan nilai Harga Produk untuk setiap interaksi.



Gambar 4. 4 Heatmap Korelasi antar feature RFM

Gambar 4.4 menunjukkan hubungan korelatif antar ketiga fitur RFM. Terlihat bahwa korelasi antara Frequency dan Monetary sangat tinggi, yaitu sebesar 0.99, yang menunjukkan bahwa pelanggan yang sering bertransaksi cenderung memiliki total belanja yang besar pula. Sebaliknya, Recency memiliki korelasi negatif yang sangat lemah terhadap kedua fitur lainnya, yaitu -0.086 terhadap Frequency dan -0.081 terhadap Monetary. Hal ini menandakan bahwa waktu kunjungan terakhir pelanggan tidak terlalu berkaitan langsung dengan seberapa sering atau banyak mereka bertransaksi. Korelasi yang hampir tidak signifikan ini juga memperkuat alasan bahwa ketiga fitur RFM dapat menangkap dimensi perilaku yang berbeda dan saling melengkapi.

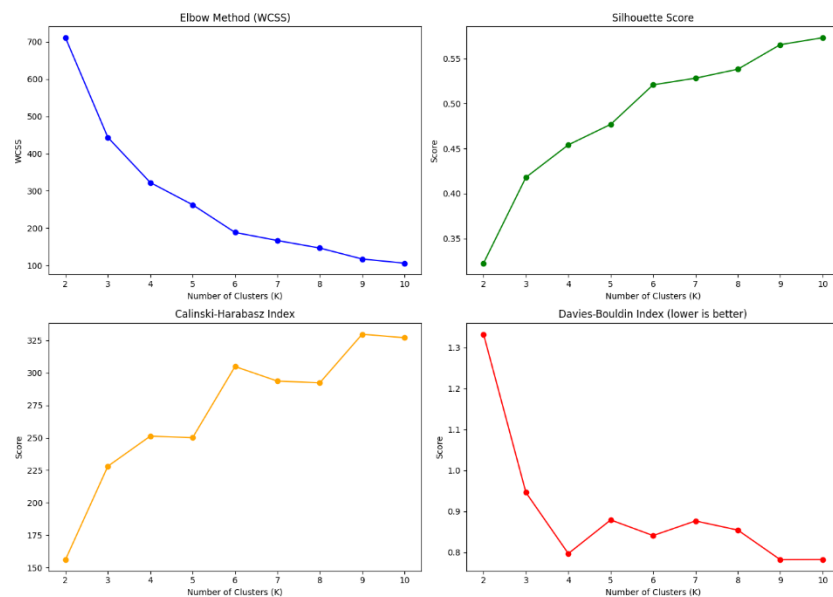


Gambar 4. 5 PairPlot Hubungan antar feature

Gambar 4.5 menampilkan hubungan visual antar fitur RFM dalam bentuk pairplot. Terlihat bahwa distribusi nilai pada fitur Frequency dan Monetary sangat ekstrem, dengan outlier yang cukup dominan. Beberapa pelanggan memiliki nilai Monetary hingga di atas 10 juta rupiah, dan ada pula yang sangat sering melakukan transaksi dibanding mayoritas lainnya. Distribusi dari Recency juga menunjukkan dominasi pelanggan yang baru-baru ini aktif. Pola ini menunjukkan bahwa sebagian besar pelanggan memiliki aktivitas pembelian terbatas, namun terdapat beberapa pelanggan yang sangat aktif dan bernilai tinggi. Outlier ini penting untuk dianalisis lebih lanjut karena bisa mempengaruhi hasil klusterisasi dan dapat menjadi indikasi adanya pelanggan strategis atau premium.

#### 4.2.2 PEMODELAN

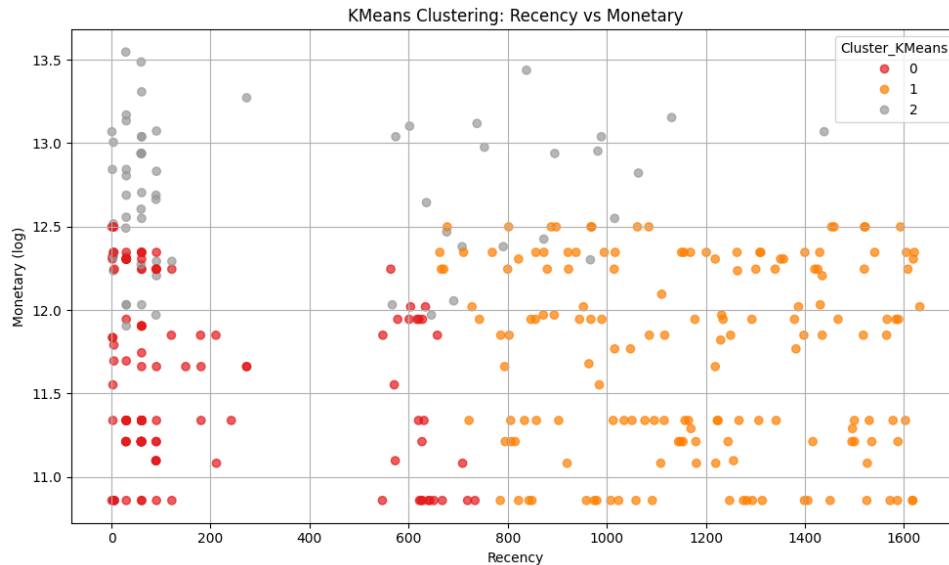
Penelitian ini menerapkan dua metode klustering, yaitu **KMeans Clustering** dan **Agglomerative Clustering**, dengan terlebih dahulu menentukan jumlah kluster optimal (K) menggunakan empat pendekatan evaluasi: *Elbow Method*, *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*. Proses pemodelan ini bertujuan untuk mengelompokkan pelanggan berdasarkan perilaku pembelian mereka yang direpresentasikan dalam variabel RFM (Recency, Frequency, Monetary). Model dievaluasi secara kuantitatif dan visual untuk menilai kualitas segmentasi yang dihasilkan.



Gambar 4. 6 Penentuan clustering dengan penerapan 4 metode

Gambar 4.6 menyajikan hasil evaluasi terhadap berbagai jumlah kluster (K) menggunakan empat metrik utama: Elbow Method, Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index. Berdasarkan Elbow Method, titik siku yang jelas terlihat pada K=3 menunjukkan bahwa penurunan nilai inertia (WCSS) menjadi tidak signifikan setelah jumlah kluster tersebut, mengindikasikan bahwa K=3 merupakan titik optimal. Silhouette Score mengalami peningkatan konsisten dan mencapai nilai yang cukup tinggi pada K=3, yakni sekitar 41–46%, yang mencerminkan kualitas pemisahan kluster yang baik. Calinski-Harabasz Index juga menunjukkan lonjakan tajam hingga K=3 sebelum akhirnya cenderung datar, yang menandakan bahwa struktur kluster paling solid ditemukan pada jumlah kluster tersebut. Sementara itu,

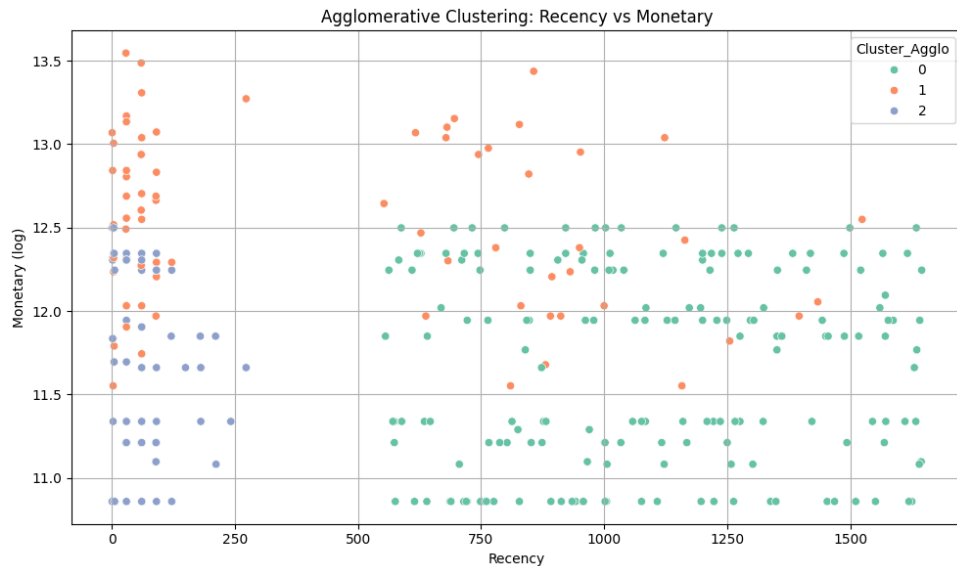
Davies-Bouldin Index mencapai nilai minimum pada  $K=3$  hingga  $K=4$ , dengan skor terbaik (terendah) berada di  $K=3$ , yaitu di bawah 0.9. Berdasarkan kesesuaian keempat metrik evaluasi ini, maka ditetapkan bahwa jumlah kluster optimal yang digunakan dalam pemodelan, baik untuk metode K-Means maupun Agglomerative, adalah  $K=3$ .



Gambar 4. 7 clustering model K-Means

Gambar 4.7 memperlihatkan hasil segmentasi pelanggan berdasarkan model KMeans Clustering dengan jumlah kluster  $K=3$ , divisualisasikan pada dimensi Recency dan Monetary. Dari hasil pemodelan, terbentuk tiga kelompok pelanggan dengan karakteristik yang berbeda. Klaster 0, ditandai dengan warna merah, mencakup pelanggan yang memiliki frekuensi transaksi tinggi namun dengan nilai belanja sedang hingga rendah, dan cenderung baru melakukan transaksi — mencerminkan segmen pelanggan yang cukup aktif namun dengan kontribusi nilai yang belum maksimal. Klaster 1, berwarna oranye, didominasi oleh pelanggan dengan nilai recency yang tinggi, artinya mereka jarang bertransaksi, serta memiliki frekuensi dan monetary yang relatif rendah, sehingga dapat diinterpretasikan sebagai pelanggan yang berisiko tidak kembali (at-risk customer). Sementara itu, klaster 2 yang digambarkan dengan warna abu-abu menunjukkan segmen best customer, yakni pelanggan dengan nilai belanja (monetary) tinggi dan waktu transaksi terakhir (recency) yang rendah,

menandakan mereka adalah pelanggan loyal dan bernilai tinggi. Secara keseluruhan, model KMeans menunjukkan pemisahan klaster yang cukup baik secara visual, dengan sebaran kelompok yang relatif jelas dan tidak saling tumpang tindih secara signifikan.



Gambar 4. 8 clustering model Agglomerative

Gambar 4.8 menampilkan hasil segmentasi pelanggan menggunakan metode Agglomerative Clustering dengan pendekatan linkage ward dan jumlah klaster ditetapkan sebanyak tiga ( $n\_cluster=3$ ). Secara umum, struktur distribusi klaster yang dihasilkan menunjukkan kemiripan dengan model KMeans, meskipun terdapat sejumlah titik data yang berada di area transisi antar klaster dan kurang terklaster secara tegas. Klaster 0, yang ditandai dengan warna hijau, merepresentasikan pelanggan yang jarang bertransaksi namun memiliki nilai belanja yang moderat, sehingga dapat dikategorikan sebagai pelanggan berisiko tidak aktif. Klaster 1, berwarna oranye, mencakup pelanggan dengan nilai monetary tinggi dan frekuensi transaksi yang relatif tinggi pula — menunjukkan segmen potensial bernilai tinggi. Sementara itu, klaster 2, yang ditandai dengan warna biru, terdiri dari pelanggan aktif yang baru melakukan transaksi (recency rendah) dengan nilai belanja yang juga relatif rendah, sehingga dapat dikelompokkan sebagai recent buyer. Secara keseluruhan, model Agglomerative menghasilkan kelompok yang cenderung lebih padat dalam ukuran grup yang kecil, namun dengan tingkat ketajaman pemisahan klaster yang sedikit lebih



rendah dibandingkan dengan hasil dari model KMeans.

### 4.3 EVALUASI MODEL

Metode	Silhouette Score (%)	Calinski-Harabasz Index	Davies-Bouldin Index
K-Means	41.0	251.8	0.94
Agglomerative	39.5	244.3	1.03

*Tabel 4. 1 Hasil Evaluasi dua model*

Tabel evaluasi kuantitatif menunjukkan bahwa model KMeans memiliki performa yang lebih unggul dibandingkan Agglomerative Clustering berdasarkan ketiga metrik evaluasi yang digunakan. Nilai Silhouette Score untuk KMeans mencapai 41,0%, lebih tinggi dibandingkan Agglomerative sebesar 39,5%, yang menunjukkan bahwa pemisahan antar cluster pada KMeans lebih jelas dan kompak. Selain itu, KMeans juga memperoleh nilai Calinski-Harabasz Index sebesar 251,8, mengindikasikan struktur cluster yang lebih terdefinisi, sementara Agglomerative hanya mencapai 244,3. Dari sisi Davies-Bouldin Index, KMeans memiliki nilai lebih rendah (0,94) dibandingkan Agglomerative (1,03), yang menunjukkan bahwa cluster hasil KMeans lebih optimal karena memiliki tingkat kesamaan intra-cluster yang lebih tinggi dan keterpisahan antar-cluster yang lebih baik. Dengan demikian, berdasarkan ketiga indikator tersebut, model KMeans dapat disimpulkan sebagai metode clustering yang paling optimal dalam konteks penelitian ini.

## **BAB 5**

### **KESIMPULAN DAN SARAN**

Hasil analisis menunjukkan bahwa metode *KMeans clustering* memberikan performa segmentasi yang lebih optimal dibandingkan *Agglomerative clustering*, ditinjau dari tiga metrik evaluasi utama, yakni *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*. Model KMeans secara efektif mampu mengelompokkan pelanggan ke dalam segmen strategis seperti *best customer*, *recent buyer*, dan *at-risk customer*, yang dapat dijadikan dasar dalam perumusan kebijakan pemasaran yang lebih tepat sasaran.

Meskipun demikian, penerapan model ini tidak terlepas dari sejumlah keterbatasan, khususnya pada aspek data yang digunakan. Dataset yang dianalisis diperoleh melalui proses *web scraping* dari situs Tokopedia dan bersifat terbatas baik dari sisi jumlah data maupun keberagaman atribut yang tersedia. Keterbatasan ini dapat mempengaruhi kualitas generalisasi model karena tidak mencakup variabel penting lain seperti jenis produk, preferensi konsumen, maupun kanal transaksi.

Untuk pengembangan penelitian selanjutnya, disarankan penggunaan dataset yang lebih komprehensif dan representatif, idealnya bersumber dari sistem internal dengan skema kerja sama resmi. Selain itu, integrasi variabel perilaku konsumen tambahan serta pemanfaatan metode klasterisasi lanjutan seperti *DBSCAN* atau teknik *deep clustering* berbasis pembelajaran mendalam dapat dipertimbangkan untuk menghasilkan segmentasi yang lebih akurat dan bernilai strategis dalam konteks pengambilan keputusan bisnis.

## DAFTAR PUSTAKA

- [1] C. Luna, “Data-Driven Customer Segmentation and Personalization in E-Commerce,” 2025.
- [2] S. Abdurrazaq and O. As, “Statistical Modeling for Customer Segmentation in E-Commerce: Techniques, Applications, and Insights,” *ResearchGate*, 2024.
- [3] A. Abdulhafedh, “Incorporating K-means, hierarchical clustering and PCA in customer segmentation,” *Journal of City and Development*, pp. 12–30, 2021.
- [4] Anonymous, “Data Mining In E-commerce Personalization: Dissertation Computer Science,” *Assignmenthelper.org*, 2024.
- [5] 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.
- [6] A. S. Abdullah, V. Thalanki, and A. Hariharan, “Enabling Customer Segmentation Based on Classical Partitioning Methods Upon Statistical Evaluation for E-commerce Business Applications,” in *Communication and Intelligent Systems. ICCIS 2023*, vol. 967, in Lecture Notes in Networks and Systems, vol. 967. , Springer Singapore, 2024, pp. 211–223. doi: 10.1007/978-981-97-2053-8\_16.
- [7] R. M. Fauzan and G. Alfian, “Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means,” *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 170–177, 2024, doi: 10.14421/jiska.2024.9.3.170-177.