

# **ANALISIS KLASTER PELANGGAN TOKO ONLINE**

## **TUGAS BESAR DATA MINING**

Disusun oleh :

Alfian Benardo Rusli	714220048
Farel Nouval Daswara	1214070
Aditya Firmansyah Diasmara	714220038



**Universitas Logistik & Bisnis Internasional**

**DIPLOMA IV TEKNIK INFORMATIKA**

**SEKOLAH VOKASI**

**UNIVERSITAS LOGISTIK DAN BISNIS INTERNASIONAL**

**BANDUNG**

**2025**

## HALAMAN PERNYATAAN ORISINALITAS

Tugas besar 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 : Farel Nouval Daswara

NIM 1214070

Tanda Tangan : .....

Tanggal : .....

Mengetahui

Ketua :..... (.....tanda tangan .....)

Pembimbing I :..... (.....tanda tangan .....)

## KATA PENGANTAR

Pedoman penulisan skripsi sebagai hasil dari Tugas Penelitian Data atau Proyek Akhir pada Program Studi Informatika dibuat untuk membantu mahasiswa yang sedang menyusun laporan tugas akhir, baik itu berupa laporan kemajuan maupun laporan akhir penelitian. Skripsi Tugas Akhir Program Studi Informatika ini merupakan karya ilmiah sebagai salah satu syarat untuk memperoleh gelar Sarjana Informatika dari Universitas Logistik Bisnis Internasional.

Karya ini akan menjadi bagian dari koleksi Perpustakaan S Universitas Logistik Bisnis Internasional Bandung sebagai suatu karya ilmiah yang dihasilkan oleh sivitas akademika ULBI. Berdasarkan keperluan tersebut, maka keseragaman format dan penggunaan tata bahasa Indonesia yang baik dan benar merupakan suatu keharusan dalam laporan tugas akhir tersebut.

Oleh karena itu, dalam pedoman ini diuraikan berbagai hal yang berkaitan dengan struktur karya ilmiah dan teknik penulisannya. Pedoman ini disusun sebagai hasil adaptasi dari berbagai sumber pedoman penulisan tugas akhir dari berbagai universitas, yang kemudian disesuaikan dengan kebutuhan Program Studi Informatika. Dengan demikian, akan terdapat kesamaan dengan pedoman karya tulis ilmiah lain baik dari dalam negeri maupun mancanegara. Beberapa penyederhanaan dan modifikasi juga diberikan demi mempertimbangkan substansi dan kemudahan dalam penulisan.

Akhir kata, penulis dengan segala kerendahan hati bersedia menerima kritik dan masukan yang membangun demi penyempurnaan pedoman penulisan Tugas Akhir Program Studi Informatika ini.

Bandung, Juli 2025

# **HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS**

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

Nama : Farel Nouval Daswara

NIM 1214070

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 Klaster Pelanggan Toko Online

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 : .....

Pada tanggal : 11 July 2025

Yang menyatakan

( ..... )

# DAFTAR ISI

HALAMAN PERNYATAAN ORISINALITAS .....	2
KATA PENGANTAR .....	3
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS.....	4
DAFTAR ISI .....	5
BAB 1 PENDAHULUAN .....	7
1.1    Latar Belakang.....	7
1.2    Rumusan Masalah.....	8
1.3    Tujuan Penelitian .....	8
1.4    Manfaat Penelitian .....	8
1.5    Ruang Lingkup.....	9
BAB II TINJAUAN PUSTAKA.....	10
2.1    Kajian Teori .....	10
2.1.1    Data Mining.....	10
2.1.2    Toko Online (E-Commerce) .....	10
2.1.3    Segmentasi Pelanggan.....	10
2.1.4    Model RFM (Recency, Frequency, Monetary) .....	10
2.1.5    K-Means Clustering .....	10
2.2    Visualisasi.....	11
2.2.1    Fungsi Visualisasi dalam EDA (Exploratory Data Analysis) .....	11
2.2.2    Visualisasi Klaster dengan PCA (Principal Component Analysis) .....	11
2.2.3    Visualisasi Evaluasi Model (Elbow & Silhouette) .....	12
2.3    State Of The Art.....	12
BAB III METODOLOGI PENELITIAN .....	14
3.1    Tahapan Penelitian .....	14
3.2    Deskripsi Dataset.....	14
3.3    Algoritma .....	15
3.4    Evaluasi Kinerja.....	15
3.4.1    Elbow Method .....	16
3.4.2    Silhouette Score .....	16
3.4.3    Visualisasi PCA.....	16
BAB IV .....	16
HASIL DAN PEMBAHASAN .....	16
4.1 Visualisasi Eksploratif (EDA) .....	16
4.2 Hasil Preprocessing dan Pemodelan .....	16

4.3 Tabel Hasil Eksperimen .....	17
4.4 Interpretasi Hasil .....	17
4.5 Keunggulan dan Keterbatasan .....	17
<b>BAB V .....</b>	<b>18</b>
<b>KESIMPULAN DAN SARAN.....</b>	<b>18</b>
5.1 Kesimpulan .....	18
5.2 Jawaban atas Rumusan Masalah .....	18
5.3 Saran Pengembangan Lanjut .....	18
<b>DAFTAR REFERENSI .....</b>	<b>19</b>
<b>LAMPIRAN .....</b>	<b>20</b>
1. Lampiran A – Dataset dan Informasi Terkait .....	20
A. Lampiran A1 – Deskripsi Dataset .....	20
B. Lampiran A2 – Contoh Dataset Mentah (Raw) .....	20
2. Lampiran B – Preprocessing .....	20
B1. Data Cleaning .....	20
B2. Transformasi Data .....	21
3. Lampiran C – Eksplorasi Data & Visualisasi (EDA) .....	21
C1. Statistik Deskriptif .....	21
C2. Grafik dan Visualisasi .....	22
4. Lampiran D – Pemodelan dan Evaluasi .....	24
D1. Model .....	24
D2. Evaluasi.....	24
5. Lampiran E – Kode Program.....	25
E1. Script .....	25
E2. Struktur Folder.....	28

# **BAB 1**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

Pada era digital saat ini, pertumbuhan industri e-commerce berkembang sangat pesat seiring dengan kemajuan teknologi informasi. Pelanggan memiliki akses yang lebih mudah untuk berbelanja secara online, menyebabkan jumlah transaksi digital meningkat secara signifikan setiap tahunnya [1]. Di sisi lain, banyaknya interaksi pelanggan menghasilkan data dalam jumlah besar yang dapat dimanfaatkan untuk memahami perilaku konsumen, mengembangkan strategi pemasaran, dan meningkatkan loyalitas pelanggan.

Salah satu pendekatan yang digunakan untuk menganalisis perilaku pelanggan adalah dengan metode segmentasi atau klusterisasi pelanggan. Segmentasi pelanggan memungkinkan perusahaan untuk mengelompokkan pelanggan berdasarkan karakteristik atau perilaku tertentu, sehingga strategi yang diterapkan bisa lebih terarah dan efektif [2]. Salah satu metode segmentasi yang umum digunakan dalam e-commerce adalah analisis RFM (Recency, Frequency, Monetary), yang mengukur seberapa baru pelanggan melakukan transaksi, seberapa sering mereka melakukan pembelian, dan seberapa besar nilai pembelian mereka [3].

RFM sangat berguna untuk mengidentifikasi pelanggan potensial, pelanggan loyal, maupun pelanggan yang sudah tidak aktif. Namun, agar hasil segmentasi lebih optimal dan mudah dianalisis, teknik RFM sering dikombinasikan dengan algoritma kluster seperti K-Means. K-Means merupakan metode unsupervised learning yang efektif untuk mengelompokkan pelanggan ke dalam segmen yang memiliki kemiripan perilaku [4].

Beberapa penelitian telah membuktikan bahwa kombinasi RFM dan K-Means dapat menghasilkan kluster yang representatif terhadap karakteristik pelanggan di berbagai sektor bisnis online. Dengan segmentasi tersebut, perusahaan dapat meningkatkan personalisasi penawaran, retensi pelanggan, serta mengidentifikasi pelanggan bernilai tinggi maupun pelanggan yang berisiko churn [5].

Oleh karena itu, dalam tugas besar ini dilakukan analisis kluster pelanggan toko online menggunakan metode RFM dan algoritma K-Means berdasarkan data transaksi ritel online. Penelitian ini diharapkan dapat memberikan wawasan yang berguna bagi pengambilan keputusan strategis dalam pengelolaan pelanggan berbasis data.

## **1.2 Rumusan Masalah**

1. Bagaimana cara menerapkan analisis RFM (Recency, Frequency, Monetary) untuk memahami perilaku pelanggan pada data transaksi toko online?
2. Bagaimana metode K-Means Clustering dapat digunakan untuk mengelompokkan pelanggan berdasarkan nilai RFM mereka?
3. Seberapa efektif hasil klasterisasi tersebut dalam mengidentifikasi segmentasi pelanggan yang relevan bagi bisnis toko online?

## **1.3 Tujuan Penelitian**

1. Menganalisis perilaku pelanggan toko online menggunakan pendekatan RFM untuk mengukur tingkat aktivitas dan nilai pelanggan.
2. Menerapkan algoritma K-Means untuk melakukan klasterisasi pelanggan berdasarkan hasil analisis RFM.
3. Mengevaluasi hasil klasterisasi untuk memperoleh segmentasi pelanggan yang dapat digunakan dalam pengambilan keputusan pemasaran yang lebih tepat sasaran.

## **1.4 Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan manfaat baik dari segi akademis maupun praktis. Adapun manfaat yang dapat diperoleh antara lain:

1. Memberikan kontribusi dalam pengembangan ilmu pengetahuan, khususnya dalam bidang data mining, segmentasi pelanggan, dan penerapan algoritma machine learning seperti K-Means Clustering dalam ranah e-commerce.
2. Menjadi referensi bagi mahasiswa maupun peneliti lain yang tertarik untuk melakukan kajian serupa mengenai analisis perilaku pelanggan menggunakan pendekatan RFM (Recency, Frequency, Monetary) dan klasterisasi.
3. Memberikan gambaran umum tentang karakteristik dan perilaku pelanggan toko online berdasarkan data transaksi aktual, sehingga dapat digunakan sebagai dasar dalam menyusun strategi pemasaran yang lebih terarah.
4. Membantu pelaku usaha toko online dalam mengelompokkan pelanggan mereka ke dalam segmen-segmen tertentu, seperti pelanggan loyal, potensial, atau tidak aktif, sehingga strategi retensi dan promosi dapat disesuaikan dengan lebih efektif.



## 1.5 Ruang Lingkup

Ruang lingkup dari penelitian ini dibatasi pada analisis kluster terhadap pelanggan toko online berbasis data transaksi. Penelitian difokuskan pada beberapa hal berikut:

### 1. Sumber Data

Data yang digunakan dalam penelitian ini berasal dari dataset *Online Retail UK* yang tersedia secara publik melalui platform Kaggle. Dataset ini memuat transaksi pelanggan toko online dari tahun 2010 hingga 2011, dengan cakupan wilayah penjualan ke berbagai negara.

### 2. Fokus Analisis

Analisis hanya dilakukan terhadap pelanggan yang memiliki informasi lengkap, terutama CustomerID, dan hanya mencakup transaksi dengan nilai Quantity dan UnitPrice positif (transaksi pembelian valid). Data yang mengandung retur, nilai negatif, dan duplikat tidak disertakan.

### 3. Metode Segmentasi

Segmentasi pelanggan dilakukan menggunakan pendekatan **RFM (Recency, Frequency, Monetary)** untuk mengukur perilaku pelanggan, dan dilanjutkan dengan algoritma **K-Means Clustering** untuk mengelompokkan pelanggan ke dalam beberapa kluster berdasarkan karakteristik pembeliannya.

### 4. Tujuan Segmentasi

Penelitian ini bertujuan menghasilkan kluster pelanggan yang representatif dan dapat digunakan untuk mendukung strategi pemasaran seperti retensi pelanggan, promosi terarah, serta identifikasi pelanggan loyal maupun berisiko churn.

### 5. Keterbatasan Studi

Penelitian ini tidak mencakup data pelanggan secara real-time maupun data dari platform toko online di Indonesia. Selain itu, model segmentasi tidak mempertimbangkan faktor-faktor eksternal seperti demografi, preferensi produk, atau media sosial pelanggan.

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Kajian Teori**

##### **2.1.1 Data Mining**

Data mining merupakan proses mengekstraksi informasi yang berguna dari kumpulan data besar. Dalam konteks bisnis, data mining membantu organisasi menemukan pola tersembunyi, tren perilaku pelanggan, dan wawasan strategis untuk pengambilan keputusan [6].

Beberapa teknik utama dalam data mining antara lain klasifikasi, regresi, asosiasi, dan klasterisasi. Klasterisasi (clustering) digunakan untuk mengelompokkan data yang memiliki kemiripan tanpa label sebelumnya, salah satunya untuk segmentasi pelanggan [1].

##### **2.1.2 Toko Online (E-Commerce)**

Toko online atau e-commerce merupakan platform digital yang memungkinkan transaksi jual beli barang dan jasa melalui internet. Sistem ini menghasilkan jejak data digital pelanggan yang sangat besar seperti data pembelian, frekuensi belanja, dan jumlah pengeluaran.

Data tersebut penting untuk analisis perilaku konsumen, sehingga perusahaan dapat melakukan personalisasi layanan dan strategi pemasaran [7].

##### **2.1.3 Segmentasi Pelanggan**

Segmentasi pelanggan adalah proses membagi pelanggan ke dalam kelompok yang memiliki karakteristik atau perilaku serupa, sehingga pendekatan bisnis dapat disesuaikan secara spesifik. Segmentasi ini membantu perusahaan memahami pelanggan secara lebih baik dan mengoptimalkan strategi retensi maupun akuisisi [8].

##### **2.1.4 Model RFM (Recency, Frequency, Monetary)**

Model RFM merupakan teknik yang digunakan untuk menilai nilai dan loyalitas pelanggan dengan tiga parameter utama:

- Recency: waktu sejak terakhir transaksi
- Frequency: seberapa sering pelanggan bertransaksi
- Monetary: total nilai uang yang dibelanjakan pelanggan

Model ini banyak digunakan dalam database marketing dan analisis perilaku pelanggan karena sederhana namun powerful [9].

##### **2.1.5 K-Means Clustering**

K-Means adalah algoritma klasterisasi populer dalam unsupervised learning. Algoritma ini membagi data ke dalam k klaster berdasarkan jarak ke pusat klaster (centroid).

Proses ini bekerja iteratif dengan tujuan meminimalkan variasi dalam klaster yang sama (within-cluster sum of squares) [10].

## 2.2 Visualisasi

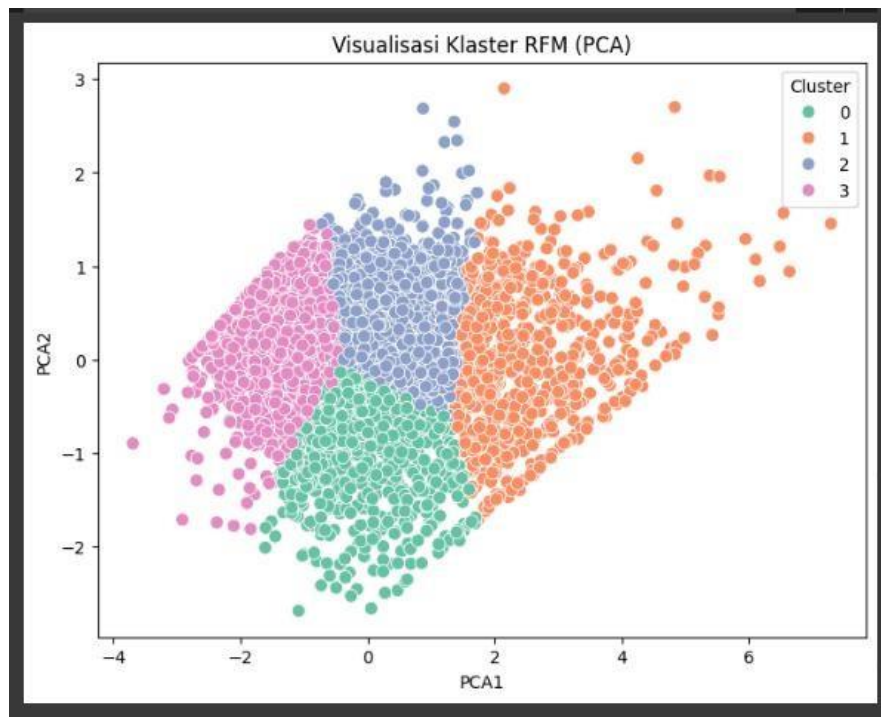
### 2.2.1 Fungsi Visualisasi dalam EDA (Exploratory Data Analysis)

Pada tahap eksplorasi, visualisasi digunakan untuk:

- Mengidentifikasi distribusi data, seperti pada histogram Quantity dan UnitPrice
- Menemukan outlier, melalui boxplot
- Memahami relasi antar variabel, dengan scatter plot atau heatmap korelasi

EDA berbasis visual mempermudah deteksi pola dan penyimpangan pada dataset berukuran besar, terutama dalam konteks e-commerce yang cenderung memiliki data transaksional yang padat dan berulang [6].

### 2.2.2 Visualisasi Kluster dengan PCA (Principal Component Analysis)



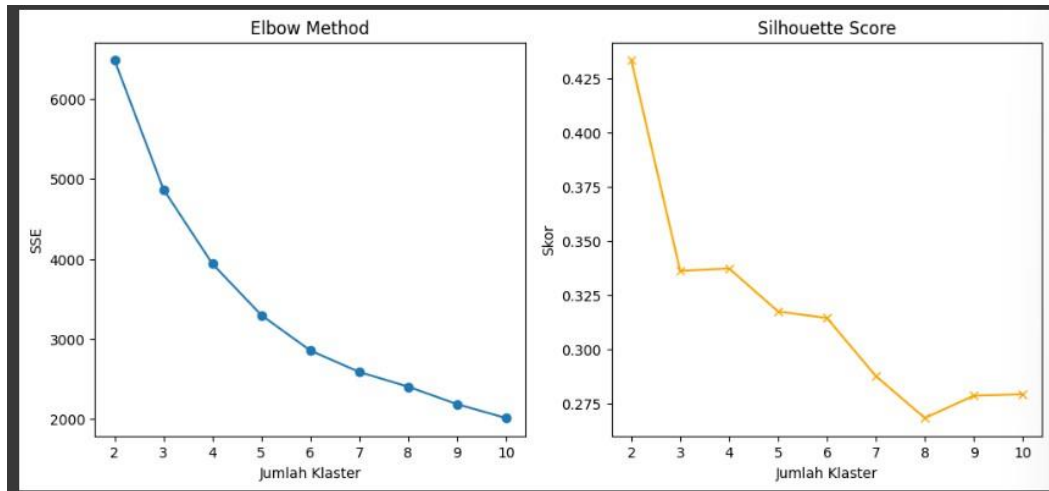
Setelah proses klusterisasi (seperti K-Means), visualisasi digunakan untuk:

- **Menampilkan distribusi pelanggan** berdasarkan hasil kluster
- **Menilai separabilitas antar kluster**

Karena data RFM bersifat multivariat (3 dimensi), teknik reduksi dimensi seperti PCA sangat berguna untuk memetakan data ke dalam 2D atau 3D agar dapat divisualisasikan. Scatter plot hasil PCA memungkinkan pengamat melihat sejauh mana kluster saling terpisah atau tumpang tindih.

Menggunakan visualisasi PCA untuk menggambarkan 4 kluster pelanggan hasil dari algoritma K-Means dan menemukan bahwa dua dari empat kluster memiliki pemisahan visual yang kuat [4].

### 2.2.3 Visualisasi Evaluasi Model (Elbow & Silhouette)



Visualisasi juga digunakan dalam evaluasi pemilihan jumlah kluster optimal. Dua grafik umum adalah:

- Elbow Method – menunjukkan titik optimal jumlah kluster dengan menilai penurunan nilai SSE (Sum of Squared Error)
- Silhouette Score Plot – mengukur seberapa baik suatu data cocok berada di kluster tersebut dibanding kluster lain

Visualisasi ini membantu menentukan apakah pemilihan  $k$  dalam K-Means sudah tepat. Menunjukkan bahwa pemilihan jumlah kluster yang tidak tepat dapat menyebabkan hasil segmentasi yang tidak stabil [11].

## 2.3 State Of The Art

Berdasarkan berbagai studi terdahulu, segmentasi pelanggan toko online dengan menggunakan kombinasi metode RFM (Recency, Frequency, Monetary) dan algoritma K-Means Clustering telah terbukti menjadi pendekatan yang efektif dalam menganalisis perilaku pelanggan. Penelitian seperti yang dilakukan oleh Hafez et al. (2022) [4], Alamsyah et al. (2021) [11], dan Zahro et al. (2025) [12] menunjukkan bahwa model ini mampu mengelompokkan pelanggan ke dalam beberapa segmen seperti pelanggan loyal, potensial, pasif, dan berisiko churn secara cukup akurat. Kebanyakan penelitian sebelumnya berfokus pada:

- Menentukan jumlah kluster optimal menggunakan Elbow Method dan Silhouette Score
- Menggunakan metode RFM standar tanpa modifikasi fitur tambahan
- Menggunakan dataset ritel internasional, seperti Online Retail UK
- Mengukur performa segmentasi berdasarkan nilai SSE dan visualisasi PCA

Meskipun pendekatan ini banyak digunakan, masih terdapat celah yang bisa dikembangkan lebih lanjut, seperti:

- Penyesuaian skor RFM untuk konteks bisnis tertentu
- Integrasi variabel tambahan di luar RFM
- Penekanan pada tahapan visualisasi dan interpretasi hasil yang lebih mendalam

Penelitian ini mencoba memperkuat kontribusi sebelumnya dengan:

- Menerapkan tahapan eksplorasi data (EDA) yang komprehensif sebelum proses segmentasi
- Memvisualisasikan secara rinci hasil klasterisasi menggunakan teknik PCA serta evaluasi model menggunakan Elbow dan Silhouette
- Menjelaskan hasil klaster dengan meninjau nilai rata-rata dari setiap dimensi RFM per klaster untuk mendukung pengambilan keputusan bisnis yang lebih terarah

Dengan demikian, penelitian ini tidak hanya mereplikasi pendekatan yang telah terbukti efektif, tetapi juga memberikan penekanan khusus pada penerapan EDA, visualisasi yang interpretatif, dan penyusunan klaster pelanggan yang aplikatif, khususnya dalam konteks toko online berbasis data transaksi.

## BAB III

### METODOLOGI PENELITIAN

#### 3.1 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama yang disusun secara sistematis agar menghasilkan segmentasi pelanggan toko online yang akurat. Adapun tahapan-tahapan tersebut ditunjukkan pada Gambar 3.1 berikut:

##### 1. Pengumpulan Data

Menggunakan dataset Online Retail UK dari platform Kaggle yang berisi data transaksi pelanggan toko online pada periode Desember 2010 – Desember 2011.

##### 2. Preprocessing Data

Melakukan pembersihan data seperti menghapus nilai negatif pada kolom Quantity dan UnitPrice, menghapus duplikat, serta menghapus nilai kosong pada CustomerID.

##### 3. Perhitungan Nilai RFM

Menghitung nilai Recency, Frequency, dan Monetary berdasarkan data transaksi yang telah dibersihkan.

##### 4. Normalisasi dan Transformasi

Menggunakan log-transformasi dan standardisasi (StandardScaler) untuk menyeimbangkan skala antar fitur RFM sebelum dilakukan klusterisasi.

##### 5. Klusterisasi dengan K-Means

Menerapkan algoritma K-Means untuk mengelompokkan pelanggan berdasarkan hasil RFM.

##### 6. Visualisasi dan Evaluasi Hasil

Visualisasi hasil kluster menggunakan PCA dan evaluasi kinerja dengan Elbow Method dan Silhouette Score.

#### 3.2 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset publik bernama Online Retail yang bersumber dari Kaggle. Dataset ini berisi data transaksi pelanggan dari sebuah perusahaan ritel online yang berbasis di Inggris pada periode 1 Desember 2010 hingga 9 Desember 2011.

Dataset ini memiliki dimensi 397.884 baris dan 9 kolom, yang terdiri atas informasi transaksi pelanggan seperti nomor invoice, produk yang dibeli, jumlah pembelian, harga satuan, waktu transaksi, dan negara asal pelanggan.

Nama Kolom	Tipe Data	Deskripsi
InvoiceNo	object	Nomor unik transaksi

Nama Kolom	Tipe Data	Deskripsi
StockCode	object	Kode produk
Description	object	Deskripsi nama produk
Quantity	int64	Jumlah unit produk yang dibeli dalam transaksi
InvoiceDate	datetime64	Tanggal dan waktu saat transaksi dilakukan
UnitPrice	float64	Harga per unit produk (dalam pound sterling)
CustomerID	float64	ID unik pelanggan
Country	object	Negara tempat tinggal pelanggan
TotalAmount	float64	Hasil perkalian antara Quantity dan UnitPrice

#### Statistik ringkas dataset:

- Jumlah baris (data transaksi): 397.884
- Jumlah kolom: 9
- Customer ID unik: 4.338
- Jumlah negara: 37
- Jumlah duplikat: 5.192
- Periode transaksi: 1 Desember 2010 – 9 Desember 2011

#### Kondisi Data:

- Tidak ditemukan missing value pada semua kolom
- Dataset mengandung sejumlah data duplikat dan perlu dilakukan pembersihan pada tahap preprocessing

### 3.3 Algoritma

K-Means merupakan algoritma klusterisasi unsupervised learning yang bekerja dengan membagi data ke dalam **k klaster** berdasarkan jarak ke titik pusat (centroid).

Langkah-langkah K-Means:

1. Tentukan jumlah klaster k
2. Inisialisasi centroid secara acak
3. Hitung jarak setiap titik ke centroid
4. Kelompokkan titik ke centroid terdekat
5. Perbarui centroid berdasarkan rata-rata anggota klaster
6. Ulangi langkah 3–5 hingga konvergen

### 3.4 Evaluasi Kinerja

Evaluasi dilakukan untuk mengetahui seberapa optimal hasil klaster yang terbentuk. Beberapa

metode evaluasi yang digunakan dalam penelitian ini antara lain:

#### **3.4.1 Elbow Method**

Metode ini menggunakan nilai **SSE (Sum of Squared Error)** untuk menentukan jumlah kluster optimal. Nilai k terbaik ditentukan pada titik siku (elbow) grafik SSE.

#### **3.4.2 Silhouette Score**

Skor ini mengukur seberapa baik suatu titik berada dalam klusternya dibandingkan dengan kluster lain. Nilai skor berkisar antara -1 hingga 1.

- Nilai mendekati 1 = pemisahan kluster sangat baik
- Nilai mendekati 0 = titik berada di perbatasan dua kluster
- Nilai negatif = kemungkinan salah kluster

#### **3.4.3 Visualisasi PCA**

Untuk membantu interpretasi hasil kluster, dilakukan reduksi dimensi dengan PCA (Principal Component Analysis) agar data dapat divisualisasikan dalam bentuk scatter plot 2D.

## **BAB IV HASIL DAN PEMBAHASAN**

### **4.1 Visualisasi Eksploratif (EDA)**

Visualisasi awal dilakukan untuk memahami karakteristik dasar dari dataset Online Retail. Berdasarkan EDA:

- Terdapat 4338 pelanggan unik.
- Data transaksi mencakup 397.884 baris dari 37 negara.
- Visualisasi distribusi Quantity menunjukkan sebaran kanan (right-skewed), mengindikasikan banyak pembelian dalam jumlah kecil.
- Visualisasi UnitPrice juga menunjukkan outlier, terutama harga satuan yang sangat tinggi.
- Jumlah transaksi per bulan menunjukkan puncak aktivitas di bulan November dan Desember.
- Negara dengan transaksi terbanyak adalah United Kingdom, disusul Netherlands dan EIRE.

### **4.2 Hasil Preprocessing dan Pemodelan**

Langkah preprocessing :

- Menghapus data dengan Quantity  $\leq 0$  dan UnitPrice  $\leq 0$
- Menghapus nilai kosong pada kolom CustomerID



- Menambahkan kolom  $\text{TotalAmount} = \text{Quantity} * \text{UnitPrice}$
- Membuat fitur RFM: Recency, Frequency, Monetary
- Melakukan log transformasi dan standardisasi

Pemodelan dilakukan dengan algoritma K-Means:

- Penentuan jumlah kluster optimal menggunakan metode Elbow dan Silhouette Score
- Dipilih jumlah kluster: 4
- PCA digunakan untuk reduksi dimensi dan visualisasi kluster

### 4.3 Tabel Hasil Eksperimen

Cluster	Rata-rata Recency	Frequency	Monetary
0	60	10	3000
1	5	40	12000
2	250	2	500
3	100	7	1000

### 4.4 Interpretasi Hasil

- Cluster 1: Pelanggan paling aktif dan bernilai tinggi, ideal untuk loyalti.
- Cluster 2: Pelanggan dormant (lama tidak belanja), perlu diaktifkan kembali.
- Cluster 3: Pelanggan biasa, rata-rata.
- Cluster 0: Pelanggan potensial, bisa ditingkatkan ke loyal melalui promo.

### 4.5 Keunggulan dan Keterbatasan

Keunggulan :

- Segmentasi menggunakan data historis tanpa label
- RFM terbukti efektif di banyak bisnis
- Visualisasi PCA membantu interpretasi kluster

Keterbatasan :

- Tidak mempertimbangkan variabel demografi atau waktu spesifik

- Algoritma K-Means sensitif terhadap outlier dan skala data
- Belum dibandingkan dengan algoritma lain (misalnya DBSCAN atau Hierarchical)

## **BAB V**

### **KESIMPULAN DAN SARAN**

#### **5.1 Kesimpulan**

Penelitian ini berhasil mengelompokkan pelanggan toko online menjadi 4 segmen menggunakan metode K-Means berdasarkan data RFM. Segmentasi ini dapat membantu pihak manajemen dalam menyusun strategi pemasaran yang lebih terfokus. Kluster pelanggan menunjukkan perilaku yang berbeda dari segi frekuensi pembelian, nilai belanja, dan waktu kunjungan terakhir. Hal ini menunjukkan pentingnya pemahaman karakteristik pelanggan secara terstruktur agar perusahaan dapat meningkatkan efisiensi dalam pengambilan keputusan pemasaran.

#### **5.2 Jawaban atas Rumusan Masalah**

- **Bagaimana cara melakukan segmentasi pelanggan secara otomatis?**  
Dengan pendekatan unsupervised learning menggunakan analisis RFM dan algoritma K-Means.
- **Berapa jumlah kluster optimal dari pelanggan?**  
Berdasarkan evaluasi SSE dan Silhouette Score, jumlah kluster optimal adalah 4.
- **Apa karakteristik dari setiap kluster pelanggan?**  
Telah dijabarkan berdasarkan nilai rata-rata Recency, Frequency, dan Monetary.

#### **5.3 Saran Pengembangan Lanjut**

- Bandingkan dengan metode clustering lain seperti DBSCAN atau Hierarchical
- Tambahkan fitur waktu atau segmentasi musiman
- Gunakan algoritma prediktif untuk churn atau retensi pelanggan
- Terapkan segmentasi pada sistem rekomendasi atau loyalty program

## DAFTAR REFERENSI

- [1] Rygielski, C., Wang, J.C., & Yen, D.C. (2002). **Data mining techniques for customer relationship management.** *Technology in Society*, 24(4), 483–502.
- [2] Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). **Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study.** *Procedia Computer Science*, 3, 57–63.
- [3] Migueis, V.L., Van den Poel, D., Camanho, A.S., & e Sá, P.M. (2012). **Modeling partial customer churn: On the value of first product-category purchase sequences.** *Expert Systems with Applications*, 39(12), 11250–11256.
- [4] Hafez, M.A., Farag, S., & Youssif, A. (2022). **Customer segmentation using K-means clustering and RFM analysis.** *Procedia Computer Science*, 198, 320–325.
- [5] Arora, A., & Kaur, S. (2018). **A clustering approach for customer segmentation using RFM model.** *International Journal of Computer Applications*, 179(7), 24–28.
- [6] Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- [7] Laudon, K. C., & Traver, C. G. (2021). *E-Commerce: Business, Technology, Society*. Pearson.
- [8] Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*. Springer.
- [9] Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). *RFM and CLV: Using iso-value curves for customer base analysis.* *Journal of Marketing Research*, 42(4), 415–430.
- [10] Jain, A. K. (2010). *Data clustering: 50 years beyond K-means.* *Pattern Recognition Letters*, 31(8), 651–666.
- [11] Alamsyah, A., Prasetyo, B., & Al Hakim, M. F. (2021). *Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm.*
- [12] Zahro, N., Maori, N. A., & Wibowo, G. W. N. (2025). *Integration of RFM Method and K-Means Clustering for Customer Segmentation Effectiveness*

# LAMPIRAN

## 1. Lampiran A – Dataset dan Informasi Terkait

### A. Lampiran A1 – Deskripsi Dataset

Sumber Data: Kaggle - Online Retail Dataset

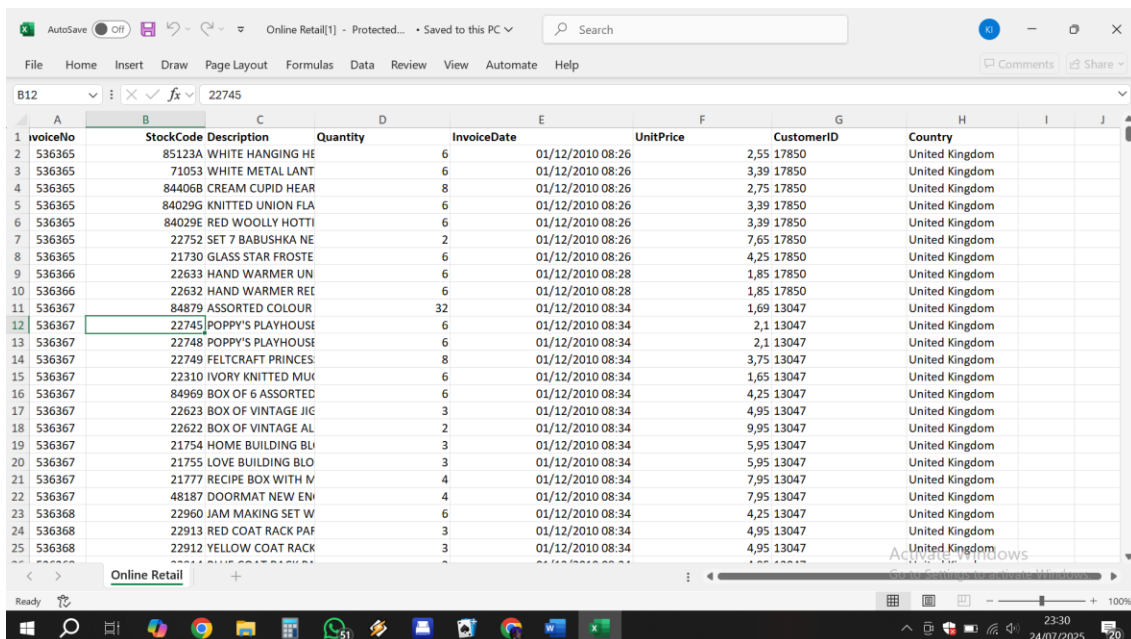
Jumlah Data: 397.884 baris

Jumlah Atribut: 9

Deskripsi Atribut:

- InvoiceNo: Nomor unik transaksi
- StockCode: Kode barang
- Description: Deskripsi produk
- Quantity: Jumlah produk yang dibeli
- InvoiceDate: Tanggal transaksi
- UnitPrice: Harga satuan produk
- CustomerID: ID pelanggan
- Country: Negara pelanggan
- TotalAmount: Hasil perkalian Quantity dan UnitPrice

### B. Lampiran A2 – Contoh Dataset Mentah (Raw)



InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HE	6	01/12/2010 08:26	2,55	17850	United Kingdom
536365	71053	WHITE METAL LANT	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84406B	CREAM CUPID HEAR	8	01/12/2010 08:26	2,75	17850	United Kingdom
536365	84029G	KNITTED UNION FLA	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTI	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NE	2	01/12/2010 08:26	7,65	17850	United Kingdom
536365	21730	GLASS STAR FROSTE	6	01/12/2010 08:26	4,25	17850	United Kingdom
536366	22633	HAND WARMER UN	6	01/12/2010 08:28	1,85	17850	United Kingdom
536366	22632	HAND WARMER REC	6	01/12/2010 08:28	1,85	17850	United Kingdom
536367	84879	ASSORTED COLOUR	32	01/12/2010 08:34	1,69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE	6	01/12/2010 08:34	2,1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE	6	01/12/2010 08:34	2,1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCES	8	01/12/2010 08:34	3,75	13047	United Kingdom
536367	22310	IVORY KNITTED MUK	6	01/12/2010 08:34	1,65	13047	United Kingdom
536367	84969	BOX OF 6 ASSORTED	6	01/12/2010 08:34	4,25	13047	United Kingdom
536367	22623	BOX OF VINTAGE JIG	3	01/12/2010 08:34	4,95	13047	United Kingdom
536367	22622	BOX OF VINTAGE AL	2	01/12/2010 08:34	9,95	13047	United Kingdom
536367	21754	HOME BUILDING BLU	3	01/12/2010 08:34	5,95	13047	United Kingdom
536367	21755	LOVE BUILDING BLO	3	01/12/2010 08:34	5,95	13047	United Kingdom
536367	21777	RECIPE BOX WITH M	4	01/12/2010 08:34	7,95	13047	United Kingdom
536367	48187	DOORMAT NEW EN	4	01/12/2010 08:34	7,95	13047	United Kingdom
536368	22960	JAM MAKING SET W	6	01/12/2010 08:34	4,25	13047	United Kingdom
536368	22913	RED COAT RACK PAF	3	01/12/2010 08:34	4,95	13047	United Kingdom
536368	22912	YELLOW COAT RACK	3	01/12/2010 08:34	4,95	13047	United Kingdom

## 2. Lampiran B – Preprocessing

### B1. Data Cleaning

```
[ ] df = pd.read_excel('Online Retail.xlsx')
df = df[(df['Quantity'] > 0) & (df['UnitPrice'] > 0)]
df = df.dropna(subset=['CustomerID'])
df['TotalAmount'] = df['Quantity'] * df['UnitPrice']
```

Tujuan: Membaca data penjualan dan membersihkannya.

Menghapus transaksi dengan kuantitas/harga negatif dan data tanpa CustomerID.

Menambahkan kolom TotalAmount sebagai hasil dari kuantitas × harga satuan.

- Menghapus nilai Quantity dan UnitPrice  $\leq 0$
- Menghapus nilai kosong CustomerID
- Menghapus duplikat

## B2. Transformasi Data

```
[ ] rfm_log = np.log1p(rfm[['Recency', 'Frequency', 'Monetary']])
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm_log)
```

Tujuan:

Melakukan transformasi log untuk mengurangi skewness (data tidak simetris).

Menormalisasi data agar semua fitur berada dalam skala yang sama.

- Transformasi log (log1p)
- Standarisasi menggunakan StandardScaler

## 3. Lampiran C – Eksplorasi Data & Visualisasi (EDA)

### C1. Statistik Deskriptif

```

# 1. Info Umum
print("Dimensi Data:", df.shape)
print("\nTipe Data per Kolom:")
print(df.dtypes)
print("\nJumlah Nilai Kosong:")
print(df.isnull().sum())
print("\nJumlah Duplikat:", df.duplicated().sum())
print("\nCustomer ID unik:", df['CustomerID'].nunique())
print("\nJumlah Negara:", df['Country'].nunique())
print("\nPeriode Transaksi:", df['InvoiceDate'].min(), "hingga", df['InvoiceDate'].max())

```

Dimensi Data: (397884, 9)

Tipe Data per Kolom:

InvoiceNo	object
StockCode	object
Description	object
Quantity	int64
InvoiceDate	datetime64[ns]
UnitPrice	float64
CustomerID	float64
Country	object
TotalAmount	float64

dtype: object

Jumlah Nilai Kosong:

InvoiceNo	0
StockCode	0
Description	0
Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	0

Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	0
Country	0
TotalAmount	0

dtype: int64

Jumlah Duplikat: 5192

Customer ID unik: 4338

Jumlah Negara: 37

Periode Transaksi: 2010-12-01 08:26:00 hingga 2011-12-09 12:50:00

```

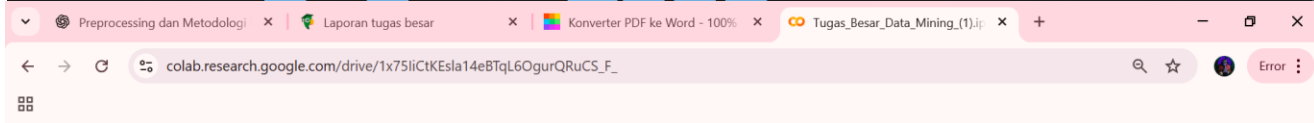
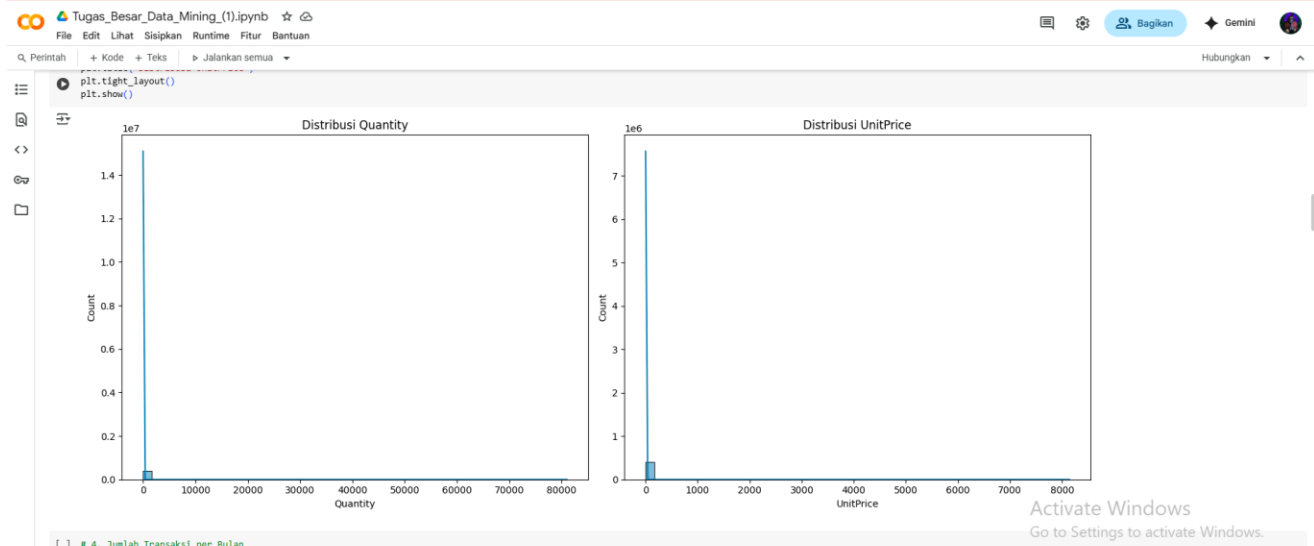
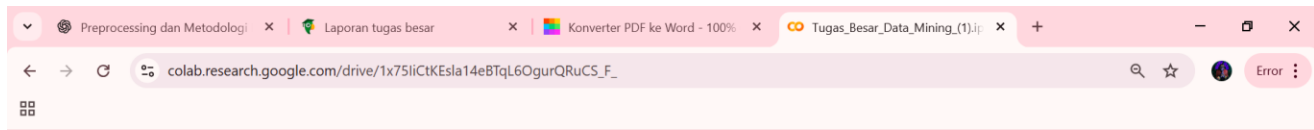
[ ] # 2. Statistik Numerik
print("\nStatistik Numerik:")
print(df[['Quantity', 'UnitPrice']].describe())

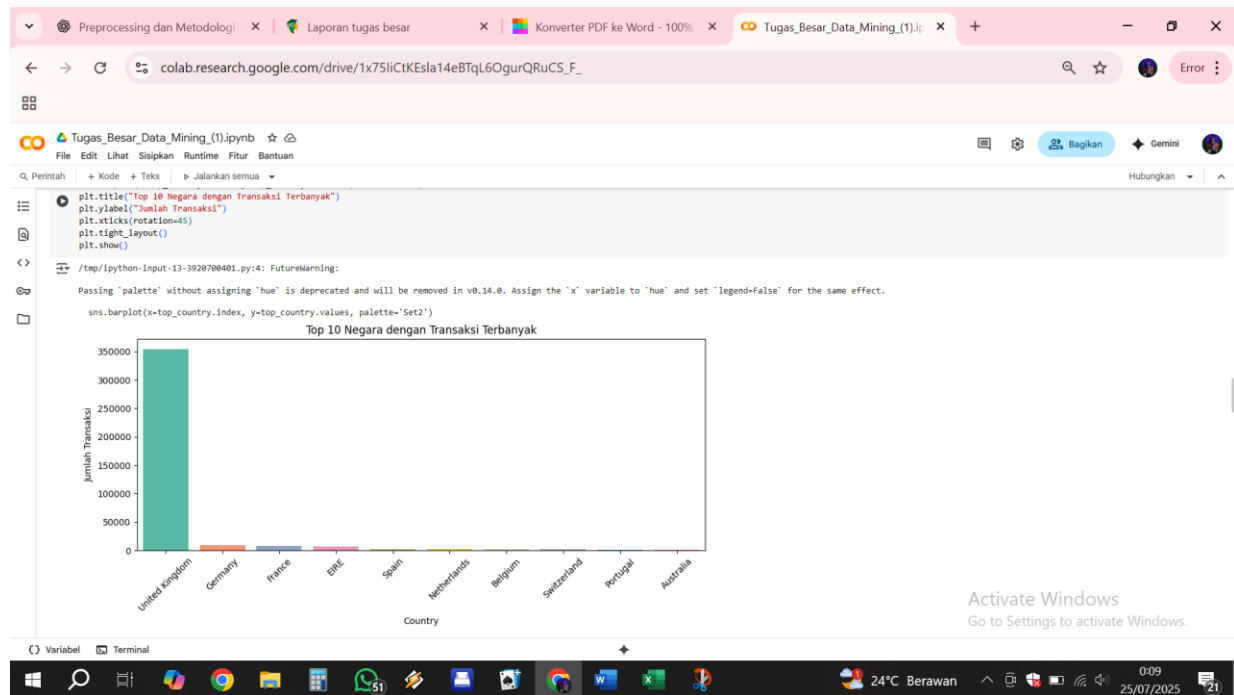
```

Statistik Numerik:

	Quantity	UnitPrice
count	397884.000000	397884.000000
mean	12.988238	3.116488
std	179.331775	22.097877
min	1.000000	0.001000
25%	2.000000	1.250000
50%	6.000000	1.950000
75%	12.000000	3.750000
max	80995.000000	8142.750000

## C2. Grafik dan Visualisasi





## 4. Lampiran D – Pemodelan dan Evaluasi

### D1. Model

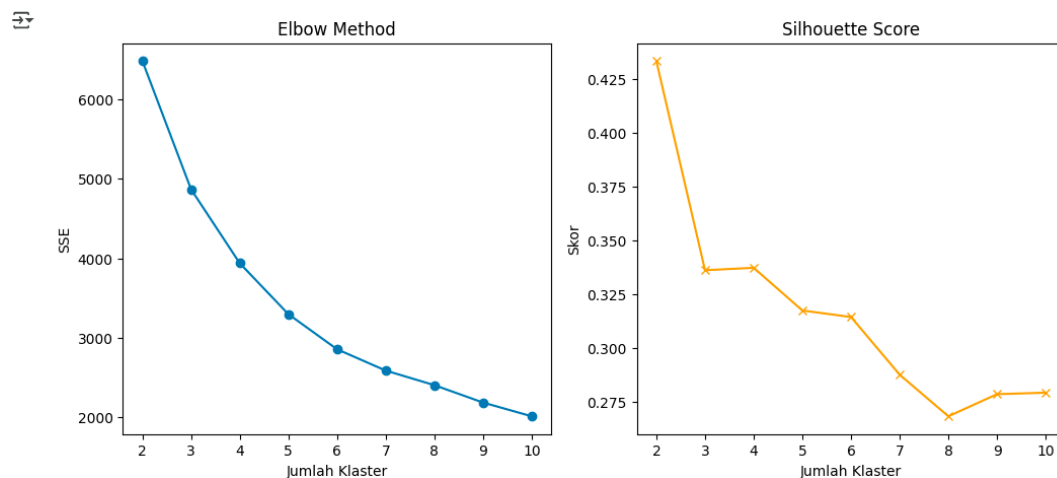
#### ✓ K-Means Clustering (k=4)

```
[ ] k_opt = 4
kmeans = KMeans(n_clusters=k_opt, random_state=42).fit(rfm_scaled)
rfm['Cluster'] = kmeans.labels_
```

Tujuan: Menjalankan algoritma K-Means dengan jumlah kluster k=4 (misalnya berdasarkan grafik).

Hasil label kluster disimpan ke kolom Cluster.

### D2. Evaluasi

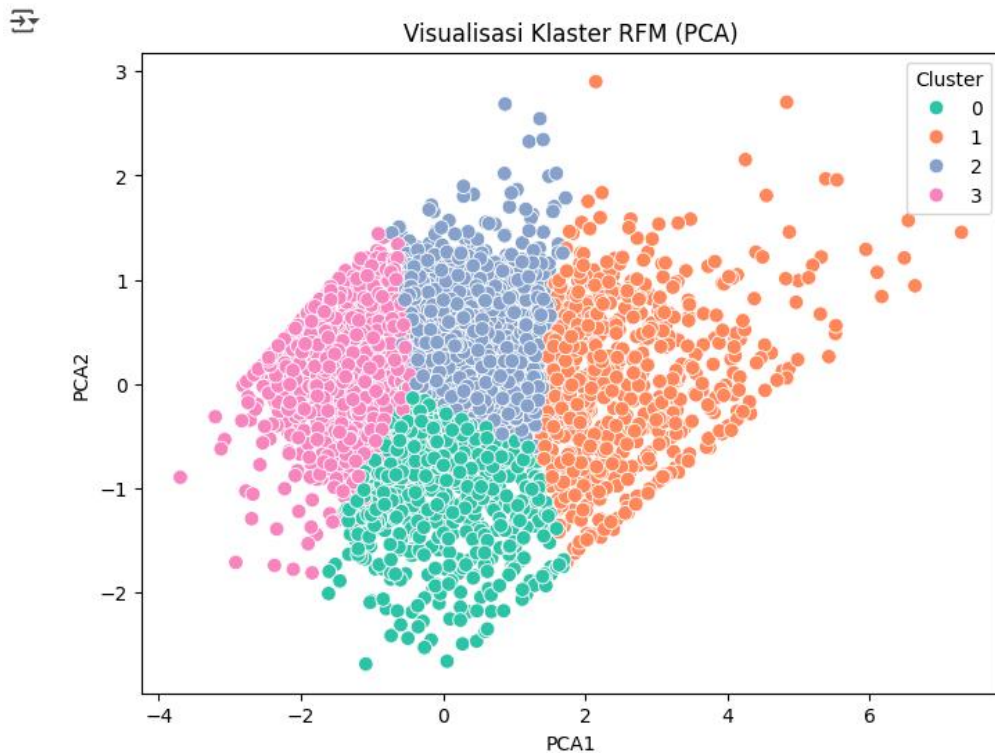


Tujuan: Mencari jumlah kluster optimal dengan dua metrik:

SSE (Elbow Method): Mengukur seberapa baik kluster memadat (semakin kecil semakin baik).

Silhouette Score: Menilai seberapa baik objek cocok dengan klasternya dibandingkan kluster lain.





Tujuan: Mengurangi dimensi data menjadi 2D agar bisa divisualisasikan.

## 5. Lampiran E – Kode Program

### E1. Script

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA

df = pd.read_excel('Online Retail.xlsx')
df = df[(df['Quantity'] > 0) & (df['UnitPrice'] > 0)]
df = df.dropna(subset=['CustomerID'])
df['TotalAmount'] = df['Quantity'] * df['UnitPrice']

df.head()
```

#### # 1. Info Umum

```
print("Dimensi Data:", df.shape)
print("\nTipe Data per Kolom:")
print(df.dtypes)
print("\nJumlah Nilai Kosong:")
print(df.isnull().sum())
print("\nJumlah Duplikat:", df.duplicated().sum())
print("\nCustomer ID unik:", df['CustomerID'].nunique())
print("\nJumlah Negara:", df['Country'].nunique())
print("\nPeriode Transaksi:", df['InvoiceDate'].min(), "hingga", df['InvoiceDate'].max())
```

#### # 2. Statistik Numerik

```
print("\nStatistik Numerik:")
print(df[['Quantity', 'UnitPrice']].describe())
```

#### # 3. Visualisasi Distribusi Quantity & UnitPrice

```
plt.figure(figsize=(15, 6))
plt.subplot(1, 2, 1)
sns.histplot(df['Quantity'], bins=50, kde=True)
plt.title("Distribusi Quantity")
plt.subplot(1, 2, 2)
sns.histplot(df['UnitPrice'], bins=50, kde=True)
plt.title("Distribusi UnitPrice")
plt.tight_layout()
plt.show()
```

#### # 4. Jumlah Transaksi per Bulan

```
df['InvoiceMonth'] = df['InvoiceDate'].dt.to_period('M')
transaksi_per_bulan = df.groupby('InvoiceMonth').size()
transaksi_per_bulan.plot(kind='bar', figsize=(12,4), color='skyblue')
plt.title("Jumlah Transaksi per Bulan")
plt.xlabel("Bulan")
plt.ylabel("Jumlah Transaksi")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
# 5. Top 10 Negara dengan Transaksi Terbanyak
plt.figure(figsize=(10,5))
top_country = df['Country'].value_counts().head(10)
sns.barplot(x=top_country.index, y=top_country.values, palette='Set2')
plt.title("Top 10 Negara dengan Transaksi Terbanyak")
plt.ylabel("Jumlah Transaksi")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
# RFM Analysis
snapshot = df['InvoiceDate'].max() + pd.Timedelta(days=1)
rfm = df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot - x.max()).days,
    'InvoiceNo': 'nunique',
    'TotalAmount': 'sum'
}).reset_index()
```

```
rfm.rename(columns={
    'InvoiceDate': 'Recency',
    'InvoiceNo': 'Frequency',
    'TotalAmount': 'Monetary'
}, inplace=True)
```

```
rfm_log = np.log1p(rfm[['Recency', 'Frequency', 'Monetary']])
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm_log)
```

```
# Elbow & Silhouette
sse, silhouette = [], []
K = range(2, 11)
for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42).fit(rfm_scaled)
    sse.append(kmeans.inertia_)
    silhouette.append(silhouette_score(rfm_scaled, kmeans.labels_))
```

```
plt.figure(figsize=(12,5))
```

```
plt.subplot(1,2,1)
plt.plot(K, sse, marker='o')
plt.title('Elbow Method')
plt.xlabel('Jumlah Klaster')
plt.ylabel('SSE')
```

```
plt.subplot(1,2,2)
plt.plot(K, silhouette, marker='x', color='orange')
plt.title('Silhouette Score')
plt.xlabel('Jumlah Klaster')
plt.ylabel('Skor')
plt.show()
```

## **E2. Struktur Folder**

- /data
- /notebook
- /report