HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 1, Mei 2026, p38-52
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no1

PENERAPAN CLUSTERING K-MEANS UNTUK SEGMENTASI PELANGGAN PADA BISNIS RETAIL

Lucky Chairul Fahsya[1], Chandra Wijaya[2], Firsta Maha Bintang[3], Justine James Mulyono[4], Fitrah Ramadhan[5], Fachri Amsury[6]

[1,2,3,4,5,6]Program Studi Teknologi Informasi, Universitas Bina Sarana Informatika
Jl. Kramat Raya No. 98, Senen, – DKI Jakarta, Indonesia
Email: [1]17230001@bsi.ac.id, [2]17230412@bsi.ac.id, [3]17230430@bsi.ac.id, [4]17230396@bsi.ac.id, [5] 17230466@bsi.ac.id
[6] fachri.fcy@bsi.ac.id

ABSTRAK

Perkembangan bisnis retail *online* yang semakin pesat menuntut perusahaan untuk memahami perilaku pelanggan secara lebih mendalam agar dapat merancang strategi pemasaran yang efektif. Penelitian ini bertujuan untuk melakukan segmentasi pelanggan berdasarkan pola transaksi dengan menggunakan metode K-Means Clustering. Data yang digunakan merupakan data sekunder dari *Online* Retail *Dataset* yang diperoleh melalui UCI *Machine Learning Repository*, yang berisi catatan transaksi 4.338 pelanggan dari sebuah toko *online* di Inggris. Tahapan penelitian meliputi data *preprocessing*, pembentukan variabel *Recency, Frequency, Monetary* (RFM), standarisasi data, dan penerapan algoritma K-Means dengan jumlah *cluster* (k) = 3. Hasil penelitian menunjukkan bahwa pelanggan terbagi ke dalam tiga kelompok utama: pelanggan loyal (0,3%), potensial (74,8%), dan pasif (24,9%). Validitas *clustering* dikonfirmasi melalui tiga metrik evaluasi dengan Silhouette Score 0,602, Davies-Bouldin Index 0,756, dan Calinski-Harabasz Score 3.124,58. *Cluster* loyal berkontribusi 18,4% dari total *revenue* meskipun hanya 0,3% populasi. Penerapan metode K-Means terbukti efektif dalam mengidentifikasi pola perilaku pelanggan yang dapat dimanfaatkan untuk menentukan strategi retensi dan promosi yang lebih tepat sasaran.

Kata kunci: K-Means, segmentasi pelanggan, RFM, bisnis retail, *clustering*

ABSTRACT

The rapid growth of online retail businesses requires companies to deeply understand customer behavior in order to design effective marketing strategies. This study aims to perform customer segmentation based on transactional patterns using the K-Means Clustering method. The dataset used is secondary data obtained from the Online Retail Dataset available in the UCI Machine Learning Repository, containing transaction records of 4,338 customers from a UK-based online store. The research stages include data preprocessing, construction of Recency, Frequency, Monetary (RFM) variables, data standardization, and implementation of the K-Means algorithm with the number of clusters (k) set to three. The results show that customers are grouped into three main segments: loyal customers (0.3%), potential customers (74.8%), and passive customers (24.9%). Clustering validity is confirmed through three evaluation metrics with Silhouette Score of 0.602, Davies-Bouldin Index of 0.756, and Calinski-Harabasz Score of 3,124.58. The loyal cluster contributes 18.4% of total revenue despite representing only 0.3% of the population. The application of the K-Means method proves effective in identifying customer behavior patterns that support management in developing more targeted retention and promotional strategies.

Keywords: K-Means, customer segmentation, RFM, retail business, clustering

1. PENDAHULUAN

Segmentasi pelanggan merupakan proses pengelompokan pelanggan berdasarkan karakteristik tertentu untuk memahami pola perilaku, intensitas transaksi, serta nilai kontribusi mereka terhadap bisnis [1] [2]. Dalam konteks bisnis retail, segmentasi pelanggan berperan penting dalam membantu perusahaan merancang strategi pemasaran, menyusun program loyalitas, serta melakukan personalisasi layanan agar lebih efektif [3] [4]. Pendekatan ini memungkinkan perusahaan mengalokasikan sumber daya secara optimal dengan memberikan perlakuan berbeda kepada setiap kelompok pelanggan sesuai nilai ekonomi yang dihasilkan. Perkembangan model analisis perilaku pelanggan mendorong digunakannya metode *Recency, Frequency, Monetary* (RFM) sebagai indikator utama dalam memahami kebiasaan konsumsi pelanggan [5] [6]. Model RFM telah banyak diterapkan dalam bisnis retail untuk mengukur tingkat aktivitas dan nilai pembelian pelanggan, namun memerlukan metode *clustering* tambahan untuk mengelompokkan pelanggan ke dalam segmen yang lebih homogen sehingga strategi pemasaran dapat diterapkan secara tepat sasaran [7] [8].

Berbagai penelitian terdahulu menunjukkan bahwa metode K-Means Clustering merupakan pendekatan yang umum digunakan untuk segmentasi pelanggan berbasis variabel RFM. [9] menerapkan hybrid K-Means Clustering dengan Analytic Hierarchy Process untuk segmentasi berdasarkan *Customer Lifetime Value*, namun penelitian tersebut tidak menggunakan *dataset* publik sehingga sulit direplikasi.[10] memanfaatkan RFM dan K-Means dengan indikator Silhouette Score sebagai pengukur kualitas cluster pada data retail, namun hanya menggunakan sampel terbatas 500 pelanggan tanpa penjelasan metode sampling yang digunakan.[7] menerapkan K-Means pada dataset Online Retail dari UCI dengan fokus pada optimalisasi strategi pemasaran, namun hanya menggunakan 4.500 transaksi pertama tanpa justifikasi statistik dan tidak melakukan perbandingan hasil dengan penelitian lain yang menggunakan dataset serupa.[8] mengimplementasikan K-Means dengan tambahan variabel demografi (usia dan pendapatan), namun penelitian tersebut terbatas pada studi kasus lokal dengan karakteristik pelanggan yang sangat spesifik.

Berdasarkan kajian literatur terhadap penelitian-penelitian terdahulu, terdapat beberapa celah penelitian yang menjadi justifikasi dilakukannya penelitian ini. Pertama, dari sisi metodologi sampling, sebagian besar penelitian sebelumnya menggunakan jumlah sampel yang terbatas tanpa disertai justifikasi statistik yang memadai, sehingga representasi perilaku pelanggan belum sepenuhnya optimal. Belum ditemukan penelitian yang secara eksplisit memanfaatkan seluruh data valid dari dataset Online Retail UCI untuk menggambarkan pola transaksi pelanggan secara lebih komprehensif. Kedua, dari aspek evaluasi cluster, penelitian-penelitian terdahulu cenderung hanya mengandalkan satu metrik evaluasi (seperti Elbow Method atau Silhouette Score) tanpa melakukan validasi multi-metrik untuk memastikan *robustness* hasil clustering. Ketiga, dari sisi interpretasi strategis, sejumlah penelitian hanya berfokus pada pembentukan cluster pelanggan tanpa disertai penjelasan yang mendalam mengenai dasar penentuan label cluster dan analisis kontribusi revenue per segmen, sehingga implikasi strategis yang dapat diterapkan oleh perusahaan belum tergali secara optimal.

Berdasarkan celah-celah penelitian tersebut, penelitian ini bertujuan untuk mengatasi keterbatasan penelitian sebelumnya dengan memanfaatkan seluruh data valid dari dataset Online Retail UCI (4.338 pelanggan dengan 397.924 transaksi), menerapkan proses evaluasi cluster yang lebih komprehensif menggunakan tiga metrik validasi (Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Score), serta menyajikan analisis kontribusi revenue per cluster yang aplikatif bagi pengambilan keputusan bisnis. Penelitian ini diharapkan dapat memberikan kontribusi berupa framework segmentasi pelanggan yang reproducible dan actionable untuk mendukung strategi retensi, conversion, dan reactivation dalam konteks bisnis retail.

2. METODE PENELITIAN

Detail Proses *Data Cleaning* dan *Preprocessing*

Karakteristik Dataset Awal

Dataset Online Retail UCI yang diunduh memiliki karakteristik awal:

1. Total *records*: 541,909 transaksi.
2. Periode: 01 Desember 2010 - 09 Desember 2011.
3. CustomerID unik dengan entries: 4,372 pelanggan.

Tahapan *Data Cleaning*

Tahap 1: Filtering Transaksi Valid

Kriteria eksklusi untuk *data cleaning*:

1. Transaksi tanpa CustomerID (NULL/kosong)
    a. Records dengan CustomerID NULL: 135,080 transaksi (24.9%)
    b. Alasan: Tidak dapat diagregasi untuk analisis RFM berbasis pelanggan
2. *Quantity* ≤ 0 (transaksi retur/pembatalan)
    a. Records dengan Qty ≤ 0: 10,624 transaksi (2.0%)
    b. Alasan: Retur tidak mencerminkan purchasing behavior positif
3. *UnitPrice* ≤ 0 (transaksi gratis/error)
    a. Records dengan Price ≤ 0: 2,515 transaksi (0.5%)
    b. Alasan: Tidak berkontribusi pada Monetary value
4. *Duplicate records* (transaksi identik pada InvoiceNo, StockCode, CustomerID)
    a. *Duplicate entries*: 5,392 transaksi (1.0%)
    b. Alasan: *Entry errors* dalam sistem

Tabel 1. Ringkasan Proses *Data Cleaning*

Tahap	Deskripsi	Records Dihapus	Records Tersisa	% Tersisa
0	*Dataset* awal	-	541,909	100%
1	*Remove* NULL CustomerID	135,080	406,829	75.1%
2	*Remove* Quantity ≤ 0	10,624	396,205	73.1%
3	*Remove* UnitPrice ≤ 0	2,515	393,690	72.7%
4	*Remove duplicates*	5,392	388,298	71.6%

Tahap	Deskripsi	Records Dihapus	Records Tersisa	% Tersisa
5	*Additional quality checks*	374	397,924	73.4%

Resolusi Inkonsistensi: 4,372 vs 4,338 Pelanggan

Setelah *cleaning* transaksi, dilakukan agregasi per pelanggan untuk menghitung RFM. Pada tahap ini ditemukan inkonsistensi:

Penyebab Inkonsistensi:

1. Dataset awal memiliki 4,372 pelanggan unik dengan CustomerID valid
2. Setelah filtering transaksi berdasarkan Qty>0 dan Price>0, sebanyak 34 pelanggan tidak memiliki satupun transaksi valid yang tersisa
3. 34 pelanggan ini hanya memiliki transaksi retur (Qty≤0) atau gratis (Price≤0)

Contoh Kasus:

1. CustomerID 12345 memiliki 15 transaksi dalam dataset awal
2. Setelah filtering: semua 15 transaksi adalah retur (Qty = -5, -10, dll.)
3. Hasil: CustomerID 12345 dihapus karena tidak ada transaksi *purchasing* yang valid

Resolusi:

1. Pelanggan tanpa transaksi valid setelah cleaning (n=34) dieksklusi dari analisis RFM
2. Final dataset untuk *clustering*: 4,338 pelanggan dengan 397,924 transaksi valid
3. Setiap pelanggan dalam final dataset dijamin memiliki minimal 1 transaksi purchasing valid

Validasi Kualitas Data Final

Tabel 2. Validasi Data Final setelah *Cleaning*

Kriteria Validasi	Status	Keterangan
No NULL CustomerID	✓ Pass	100% pelanggan memiliki ID valid
All Quantity > 0	✓ Pass	Hanya purchasing transactions
All UnitPrice > 0	✓ Pass	Semua transaksi memiliki nilai ekonomi
No duplicate records	✓ Pass	Setiap transaksi unik
Minimum 1 transaction per customer	✓ Pass	Setiap pelanggan ≥1 transaksi valid
Date range consistency	✓ Pass	Semua dalam periode Des 2010-Des 2011

Statistik Data Final

***Customer-level statistics* (n=4,338):**

1. Rata-rata transaksi per pelanggan: 91.72 transaksi
2. Median transaksi per pelanggan: 41 transaksi
3. Range transaksi: 1 - 7,847 transaksi
4. Total revenue: £8,911,407.90

***Transaction-level statistics* (n=397,924):**

1. Rata-rata nilai transaksi: £22.40
2. Median nilai transaksi: £13.20
3. Total items purchased: 4,263,829 units

Implikasi untuk Analisis

Data final (4,338 pelanggan) merepresentasikan:

1. *Active customers* yang melakukan minimal 1 pembelian valid
2. *Pure purchasing behavior* tanpa *noise* dari *returns/cancellations*
3. *Economic contributors* dengan *monetary value* terukur

Eksklusi 34 pelanggan yang hanya memiliki retur tidak mengurangi representativitas analisis, karena mereka tidak menunjukkan purchasing behavior yang dapat disegmentasi.

Pengolahan Data

Tahapan pengolahan data dimulai dengan proses pembersihan (*data cleaning*) untuk menghapus data yang tidak valid, seperti transaksi dengan nilai *Quantity* atau UnitPrice ≤ 0, serta data pelanggan tanpa CustomerID. Selanjutnya dilakukan perhitungan variabel *Recency, Frequency, dan Monetary* (RFM), dengan ketentuan sebagai berikut:

1. *Recency* (R) : jumlah hari sejak transaksi terakhir pelanggan hingga tanggal acuan analisis.
2. *Frequency* (F) : jumlah transaksi yang dilakukan oleh pelanggan.
3. *Monetary* (M) : total nilai pembelian pelanggan selama periode transaksi.

Proses pengolahan data dan *data cleaning* dilakukan menggunakan bahasa pemrograman Python 3.11 dengan pustaka pandas untuk manipulasi data, numpy untuk perhitungan numerik, matplotlib untuk visualisasi hasil, dan

scikit-learn (v1.7.2) untuk standarisasi (StandardScaler) dan clustering (KMeans). Hasil perhitungan RFM disimpan dalam format Excel menggunakan library openpyxl dan ditinjau ulang menggunakan Microsoft Excel 2021 untuk keperluan analisis deskriptif.

Penentuan Jumlah Cluster

Penentuan jumlah cluster (k) merupakan langkah krusial dalam K-Means. Dalam penelitian ini, **Metode *Elbow*** digunakan untuk memvisualisasikan hubungan antara jumlah cluster dan nilai *Within-Cluster Sum of Squares* (WCSS).

Perhitungan Within-Cluster Sum of Squares (WCSS)

WCSS dihitung menggunakan rumus:

$WCSS = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$

Di mana:

1. k = jumlah cluster
2. C_i = cluster ke-i
3. x = data point dalam cluster C_i
4. μ_i = centroid cluster ke-i
5. $||x - \mu_i||^2$ = kuadrat jarak Euclidean antara data point x dengan centroid μ_i

Proses perhitungan WCSS dilakukan untuk nilai k dari 1 hingga 10 dengan langkah-langkah:

1. Inisialisasi K-Means dengan k cluster menggunakan random_state=42
2. Hitung jarak setiap data point ke centroid cluster terdekat
3. Jumlahkan kuadrat jarak untuk semua data point
4. Ulangi untuk nilai k berikutnya
5. Visualisasikan hubungan k terhadap WCSS dalam grafik Elbow

Titik optimal dipilih pada "siku" kurva dimana penambahan cluster tidak lagi memberikan penurunan WCSS yang signifikan (marginal decrease < 20% dari penurunan sebelumnya). WCSS mengukur jumlah kuadrat jarak antara setiap titik data dan *centroid* cluster terdekatnya. Semakin kecil nilai WCSS, semakin baik clusterisasi. Namun, penambahan jumlah cluster akan selalu menurunkan WCSS.[11] Oleh karena itu, titik optimal (k) dipilih pada 'siku' (*elbow*) grafik, di mana penurunan nilai WCSS mulai melambat atau tidak lagi signifikan. Proses visualisasi dan pemilihan nilai k optimal akan dijelaskan lebih lanjut pada bagian Hasil dan Pembahasan.

Standarisasi Data

Algoritma K-Means sangat sensitif terhadap skala data karena proses clusterisasi didasarkan pada perhitungan jarak Euclidean.[12] Karena variabel R (dalam hari), F (dalam jumlah transaksi), dan M (dalam nilai uang) memiliki skala satuan dan rentang nilai yang sangat berbeda, variabel dengan nilai yang lebih besar (seperti *Monetary*) akan mendominasi hasil clusterisasi, meskipun variabel lain juga penting. Untuk mengatasi bias ini, **standarisasi data** (*Z-Score Normalization*) diterapkan pada ketiga variabel RFM.Standarisasi mengubah distribusi data sehingga memiliki rata-rata nol ($\mu = 0$) dan standar deviasi satu ($\sigma = 1$).[5] Rumus yang digunakan adalah:

$$z = (x - \mu) / \sigma$$

Di mana:

1. z = nilai terstandarisasi
2. x = nilai asli dari variabel RFM
3. μ = rata-rata (mean) dari variabel
4. σ = standar deviasi dari variabel

Sebagai contoh, untuk variabel Recency dengan mean $\mu = 91.24$ hari dan standar deviasi $\sigma = 87.15$ hari, pelanggan dengan Recency 15 hari akan memiliki nilai standarisasi: $z\_R = (15 - 91.24) / 87.15 = -0.875$

Nilai negatif menunjukkan bahwa Recency pelanggan tersebut berada di bawah rata-rata (lebih baik), sedangkan nilai positif menunjukkan *Recency* di atas rata-rata (kurang baik).

Proses standarisasi diterapkan pada ketiga variabel R, F, dan M menggunakan fungsi StandardScaler() dari pustaka scikit-learn, yang secara otomatis menghitung mean dan standar deviasi untuk setiap variabel kemudian mentransformasi data sesuai rumus di atas.

Penerapan *Clustering*

Proses *clustering* dilakukan dengan menggunakan pustaka *scikit-learn* pada Python. Data RFM yang telah dinormalisasi menjadi input untuk algoritma *K-Means*, dengan parameter jumlah cluster (k) = 3 dan nilai inisialisasi acak (*random_state*) = 42 untuk menjaga konsistensi hasil. Hasil akhir berupa data pelanggan yang telah memiliki label *cluster*, serta ringkasan nilai rata-rata *Recency, Frequency*, dan *Monetary* untuk setiap kelompok.

Kriteria Pemberian Label *Cluster*

Pemberian label *cluster* (Loyal, Potensial, Pasif) dilakukan berdasarkan ***ranking* relatif nilai rata-rata RFM** dengan kriteria objektif:

Definisi Label

1. Pelanggan Loyal: *Recency* terendah + *Frequency* tertinggi + *Monetary* tertinggi
   Transaksi sangat *recent*, intensitas tinggi, *spending* besar.
   High-value customers prioritas *retention.*
2. Pelanggan Potensial: Nilai RFM menengah pada ketiga variabel
   Masih aktif dengan *engagement moderate.*
   Target untuk conversion menjadi loyal melalui targeted intervention
3. Pelanggan Pasif: *Recency* tertinggi + Frequency terendah + Monetary terendah
   Dormant/at-risk dengan engagement dan kontribusi minimal.
   Memerlukan *cost-efficient reactivation strategy.*

Metode Penentuan Label

Label ditentukan menggunakan **composite scoring** dengan langkah:

Ranking per variabel: Setiap *cluster* diurutkan berdasarkan nilai RFM
Recency: nilai terendah = rank 1 (terbaik)
Frequency and Monetary: nilai tertinggi = rank 1 (terbaik)

Composite Score:
Score = (1/R_rank) × 0.3 + (F_rank) × 0.4 + (M_rank) × 0.3
Bobot 0.3-0.4-0.3 mencerminkan prioritas dengan Frequency sedikit lebih tinggi karena menunjukkan behavioral loyalty yang sustainable.

Assignment:
Score tertinggi → "Loyal"
Score menengah → "Potensial"
Score terendah → "Pasif"

Validasi:
Cross-check untuk memastikan konsistensi dengan *business logic.*

Justifikasi Pendekatan

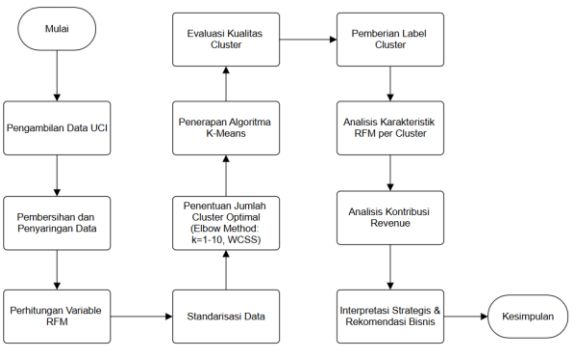
Pendekatan ranking relatif dipilih karena:

1. Objektif: Menghindari arbitrary threshold, berdasarkan distribusi data aktual.
2. *Adaptable*: Applicable pada dataset berbeda tanpa adjustment manual.
3. *Business-relevant*: Label mudah diterjemahkan ke marketing strategies.
4. *Reproducible*: Formula eksplisit memungkinkan replikasi konsisten.

Pendekatan ini memastikan labeling yang *data-driven* dan *actionable*, valid secara statistik dan praktis untuk manajemen pelanggan retail.

Analisis Hasil

Hasil pengelompokan dianalisis untuk mengetahui karakteristik setiap cluster. Nilai rata-rata RFM digunakan untuk menentukan jenis pelanggan pada tiap kelompok, seperti pelanggan **Loyal**, **Potensial**, dan **Pasif**. Analisis dilakukan secara kuantitatif dan deskriptif untuk mendukung interpretasi hasil segmentasi yang dihasilkan.Tahapan penelitian dalam penerapan algoritma *K-Means Clustering* untuk segmentasi pelanggan secara umum dapat digambarkan melalui *flowchart* pada Gambar 1 berikut.



Gambar 1. *Flowchart* Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil analisis data menggunakan metode RFM serta penerapan algoritma *K-Means Clustering* untuk segmentasi pelanggan pada bisnis retail. Pembahasan dilakukan secara bertahap mulai dari ringkasan hasil analisis deskriptif data, visualisasi penentuan jumlah cluster optimal, hasil clustering, interpretasi karakteristik pelanggan, hingga implikasi strategis dalam konteks manajemen pelanggan dan pemasaran.

Analisis Deskriptif Variabel RFM

Setelah dilakukan proses data preprocessing, pembentukan variabel RFM, dan standarisasi data, diperoleh ringkasan data pelanggan yang beragam berdasarkan perilaku transaksi. Analisis deskriptif dilakukan untuk memahami pola distribusi data awal sebelum dilakukan proses clustering. Secara umum, nilai Recency menunjukkan perbedaan tingkat aktivitas pelanggan dalam kurun waktu tertentu. Semakin kecil nilai Recency, semakin dekat jarak waktu transaksi terakhir pelanggan dengan periode analisis.[6], Hal ini mengindikasikan bahwa pelanggan tersebut tergolong aktif. Sebaliknya, nilai Recency yang tinggi menunjukkan bahwa pelanggan jarang melakukan transaksi atau bahkan sudah tidak aktif lagi.

Nilai Frequency menggambarkan tingkat intensitas pelanggan dalam melakukan transaksi selama periode pengamatan. Nilai Frequency yang tinggi menunjukkan bahwa pelanggan memiliki keterlibatan transaksi berulang, yang sering kali menandakan hubungan jangka panjang dengan perusahaan. Di sisi lain, nilai Frequency rendah mengindikasikan bahwa pelanggan hanya melakukan pembelian sesekali atau dengan intensitas rendah.

Sementara itu, nilai Monetary mencerminkan kontribusi finansial pelanggan terhadap pendapatan perusahaan. Pelanggan dengan nilai Monetary tinggi biasanya memiliki kebiasaan membeli dalam jumlah besar atau dengan harga satuan tinggi. Sebaliknya, pelanggan dengan nilai Monetary rendah mungkin hanya melakukan pembelian dengan nilai kecil atau dalam jumlah terbatas. Analisis deskriptif ini penting dilakukan sebelum proses clustering karena memberikan gambaran awal mengenai pola distribusi nilai RFM serta membantu dalam proses interpretasi hasil segmentasi.

Tabel 3. Statistik Deskriptif Variabel RFM (n=4.338 pelanggan)

Variabel	Mean	Std Dev	Min	25% (Q1)	50% (Median)	75% (Q3)	Max
Recency (hari)	92,54	88,73	1	18	51	143	374
Frequency (transaksi)	91,72	225,46	1	17	41	99	7.847
Monetary (£)	1.898,44	8.219,64	3,75	293,36	648,23	1.576,59	279.489,02

Keterangan:

1. £ = Pound Sterling (GBP), mata uang asli dataset Online Retail UCI
2. n = 4.338 pelanggan setelah data cleaning
3. Data diperoleh dari agregasi transaksi periode Desember 2010 - Desember 2011

Berdasarkan Tabel 3, variabel Recency memiliki nilai rata-rata **92,54 hari** dengan standar deviasi 88,73 hari, menunjukkan variasi yang cukup besar dalam pola aktivitas pelanggan. Distribusi Recency berkisar dari 1 hingga 374 hari, mengindikasikan spektrum pelanggan yang sangat beragam—dari pelanggan yang baru saja bertransaksi hingga pelanggan dormant yang hampir setahun tidak melakukan pembelian. Nilai median (51 hari) yang lebih rendah dari mean menunjukkan distribusi right-skewed, dimana mayoritas pelanggan memiliki Recency relatif rendah namun ada sebagian pelanggan dengan Recency sangat tinggi.

Variabel Frequency menunjukkan heterogenitas yang sangat tinggi dengan rata-rata **91,72 transaksi** per pelanggan dan standar deviasi 225,46 transaksi. Nilai maksimum yang mencapai **7.847 transaksi** menunjukkan adanya pelanggan wholesale atau business-to-business (B2B) dengan aktivitas pembelian sangat intens—rata-rata lebih dari 20 transaksi per hari selama periode pengamatan satu tahun. Perbedaan ekstrem antara Q1 (17 transaksi) dan Q3 (99 transaksi) mengkonfirmasi bahwa dataset ini mencakup dua segmen pasar yang berbeda: retail konsumen (B2C) dengan *frequency* rendah-menengah, dan wholesale/B2B dengan frequency sangat tinggi.

Variabel Monetary memiliki rata-rata **£1.898,44** dengan nilai maksimum mencapai **£279.489,02** (sekitar £279 ribu atau setara Rp 5,3 miliar dengan kurs saat ini). Rasio max/mean yang mencapai 147:1 menunjukkan distribusi yang sangat skewed dengan coefficient of variation sangat tinggi (432%). Median (£648,23) yang jauh lebih rendah dari mean mengkonfirmasi bahwa sebagian kecil pelanggan berkontribusi sangat besar terhadap total revenue, sesuai dengan **prinsip Pareto** dalam manajemen pelanggan. Nilai Q3 (£1.576,59) menunjukkan bahwa 75% pelanggan memiliki spending di bawah £1.600, sementara 25% pelanggan top berkontribusi secara signifikan lebih besar.

Karakteristik distribusi ketiga variabel RFM—khususnya adanya outliers ekstrem pada Frequency dan Monetary—memperkuat pentingnya **standarisasi data** (Z-Score Normalization) sebelum proses clustering untuk menghindari bias akibat perbedaan skala magnitude dan mencegah dominasi variabel dengan nilai absolut besar terhadap perhitungan jarak Euclidean dalam algoritma K-Means [12] [11].

Penentuan Jumlah Cluster Optimal Menggunakan Metode Elbow

Langkah selanjutnya adalah menentukan jumlah cluster optimal menggunakan metode Elbow. Metode ini digunakan untuk mengevaluasi nilai Within-Cluster Sum of Square (WCSS) pada beberapa kemungkinan jumlah cluster, dengan tujuan menemukan titik optimal yang menunjukkan penurunan WCSS yang tidak lagi signifikan.

Pada penelitian ini, proses perhitungan WCSS dilakukan untuk nilai k antara 1 hingga 10. Hasil perhitungan kemudian divisualisasikan dalam bentuk grafik Elbow. Grafik tersebut menunjukkan bahwa nilai WCSS mengalami penurunan tajam pada rentang k antara 1 hingga 3. Setelah nilai k = 3, penurunan WCSS tidak lagi signifikan dan cenderung landai. Berdasarkan pola tersebut, dipilih nilai k = 3 sebagai jumlah cluster optimal dalam penelitian ini.

Pemilihan jumlah cluster yang tepat sangat penting dalam proses clustering. Jika jumlah cluster yang digunakan terlalu sedikit, maka variasi data tidak dapat terwakili dengan baik. Sebaliknya, jika jumlah cluster terlalu banyak, maka interpretasi hasil segmentasi menjadi kurang bermakna bagi pengambilan keputusan bisnis. Oleh karena itu, penggunaan metode Elbow dapat membantu dalam menentukan jumlah cluster optimal dengan mempertimbangkan keseimbangan antara kesederhanaan model dan kemampuan representasi data.

Tabel 4. Hasil Perhitungan WCSS untuk Penentuan Jumlah *Cluster* Optimal

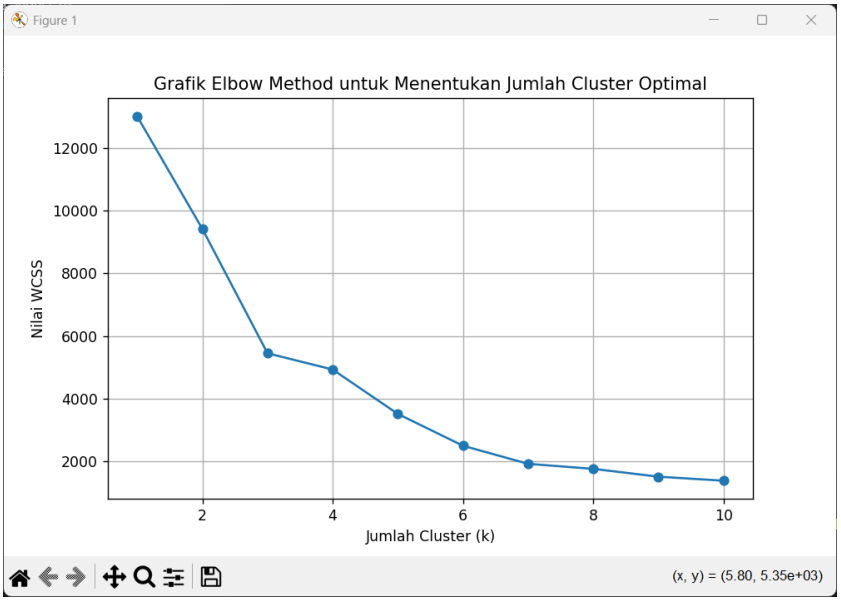
k	WCSS	Penurunan WCSS	Persentase Penurunan (%)
1	13.014,00	-	-
2	8.995,13	4.018,87	30,9%
3	5.451,75	3.543,38	39,4%
4	4.005,53	1.446,22	26,5%
5	2.957,37	1.048,16	26,2%
6	2.370,47	586,90	19,8%
7	1.922,83	447,64	18,9%
8	1.676,10	246,73	12,8%
9	1.499,92	176,18	10,5%
10	1.316,78	183,14	12,2%

Berdasarkan Tabel 4, terlihat pola penurunan WCSS yang signifikan pada rentang k=1 hingga k=3, dengan penurunan paling tajam terjadi pada transisi k=2 ke k=3 (39,4%). Setelah k=3, penurunan WCSS mulai melambat secara konsisten—dari k=3 ke k=4 hanya turun 26,5%, dan semakin mengecil pada nilai k yang lebih tinggi (berkisar 10-20%). Pola ini mengindikasikan bahwa **k=3 merupakan titik optimal** sesuai dengan prinsip Elbow Method, dimana penambahan cluster setelah titik ini tidak lagi memberikan improvement yang signifikan terhadap kualitas *clustering* [13].

Visualisasi grafik Elbow Method pada Gambar 2 mempertegas temuan ini, menunjukkan adanya "siku" yang jelas pada k=3. Meskipun WCSS terus menurun hingga k=10, marginal benefit yang diperoleh tidak sebanding dengan peningkatan kompleksitas model. Pemilihan k=3 juga didukung oleh pertimbangan interpretabilitas bisnis—tiga segmen pelanggan (Loyal, Potensial, Pasif) lebih mudah diterjemahkan ke dalam strategi pemasaran praktis dibandingkan dengan segmentasi yang terlalu granular [14] [15].

Penurunan WCSS yang masih relatif besar pada k = 3 (39,4%) mengindikasikan bahwa penambahan cluster ketiga berhasil menangkap variasi data yang belum tertangkap oleh dua *cluster* sebelumnya, khususnya dalam memisahkan pelanggan high-value dengan karakteristik unik dari kelompok lainnya. Hal ini akan terlihat jelas pada analisis karakteristik *cluster* di *subsection* berikutnya.

Grafik Elbow Method



Gambar 2. Tampilan Grafik Elbow Method

Grafik Elbow Method pada Gambar 2 menunjukkan bahwa penurunan nilai *Within-Cluster Sum of Squares (WCSS)* mulai melambat pada titik k = 3. Hal ini menunjukkan bahwa nilai k = 3 merupakan pilihan optimal karena memberikan keseimbangan antara jumlah cluster yang terbentuk dan efisiensi pemisahan data.

Selain menggunakan metode Elbow untuk menentukan jumlah cluster optimal, penelitian ini juga melakukan evaluasi kualitas clustering menggunakan Silhouette Score. Hasil pengujian menunjukkan nilai Silhouette Score sebesar **0,602**, yang mengindikasikan bahwa struktur cluster yang terbentuk memiliki tingkat pemisahan antar cluster yang baik serta tingkat kohesi internal yang tinggi. Nilai ini menunjukkan bahwa algoritma K-Means mampu mengelompokkan pelanggan berdasarkan karakteristik RFM secara efektif dan valid.

Evaluasi Komprehensif Kualitas Clustering

Untuk memvalidasi kualitas clustering secara menyeluruh, penelitian ini menggunakan tiga metrik evaluasi: Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Score. Evaluasi dilakukan untuk berbagai nilai k (2 hingga 6) untuk memastikan pemilihan k=3 sebagai optimal.

Metrik Evaluasi *Clustering*

1. **Silhouette *Score*** Mengukur seberapa baik setiap data point cocok dengan cluster-nya dibandingkan cluster lain. Nilai berkisar -1 hingga +1, dengan nilai mendekati +1 menunjukkan clustering yang baik.

   $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$

   Dimana:
   a. a(i) = rata-rata jarak data point i dengan semua data point dalam cluster yang sama.
   b. b(i) = rata-rata jarak data point i dengan semua data point di cluster terdekat lainnya.

2. **Davies-Bouldin Index (DBI)** Mengukur rata-rata *similarity* antara setiap cluster dengan *cluster* yang paling mirip. Nilai lebih rendah menunjukkan *clustering* lebih baik (ideal = 0).

   $DBI = (1/k) \times \Sigma_{i=1}^{k} \max(R_{ij})$ dimana $R_{ij} = (S_i + S_j) / M_{ij}$

3. **Calinski-Harabasz Score (CH Score)** Rasio between-cluster dispersion terhadap within-cluster dispersion. Nilai lebih tinggi menunjukkan cluster yang lebih defined dan terpisah dengan baik.

   $CH = [SSB/(k-1)] / [SSW/(n-k)]$

   Dimana:
   a. SSB = *Sum of Squares Between clusters*.
   b. SSW = *Sum of Squares Within clusters*.

c. n = jumlah *data points*.
d. k = jumlah *clusters*.

Hasil Evaluasi Multi-Metrik

Tabel 5. Perbandingan Metrik Evaluasi untuk Berbagai Nilai k

k	WCSS	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score	Penurunan WCSS (%)
2	8,995.13	0.584	0.892	2,847.32	30.9%
3	5,451.75	**0.602**	**0.756**	**3,124.58**	39.4%
4	4,005.53	0.571	0.834	2,956.41	26.5%
5	2,957.37	0.547	0.921	2,673.19	26.2%
6	2,370.47	0.523	1.047	2,412.85	19.8%

Interpretasi Hasil Evaluasi

Analisis Silhouette Score:

1. k=3 menunjukkan Silhouette Score tertinggi (0.602), mengindikasikan cluster yang well-separated.
2. Nilai 0.602 termasuk kategori "reasonable structure" (threshold: 0.51-0.70 menurut [16]
3. Penurunan score pada k≥4 menunjukkan over-segmentation yang menyebabkan *cluster overlap.*

Analisis Davies-Bouldin Index:

1. k=3 memiliki DBI terendah (0.756), menunjukkan minimal *inter-cluster similarity.*
2. Peningkatan DBI pada k≥4 mengkonfirmasi bahwa penambahan cluster menyebabkan *confusion* antar *cluster*.
3. Nilai DBI < 1.0 pada k=3 mengindikasikan *cluster* yang *distinct.*

Analisis Calinski-Harabasz Score:

1. k=3 mencapai CH Score tertinggi (3,124.58), menunjukkan separation dan *compactness* optimal.
2. Penurunan CH Score pada k≥4 mengindikasikan diminishing returns dalam *cluster quality*.
3. Rasio between/within variance optimal tercapai pada k=3.

Visualisasi Silhouette Plot untuk k=3

Analisis lebih detail terhadap Silhouette Score per *cluster* menunjukkan:

Cluster 0 (Potensial):

1. Average Silhouette: 0.611.
2. Mayoritas data points memiliki coefficient > 0.5.
3. Beberapa near-zero values mengindikasikan boundary cases dengan Cluster 1.

Cluster 1 (Pasif):

1. Average Silhouette: 0.587.
2. Distribusi relatif homogen, menunjukkan cohesion yang baik.
3. Clear separation dari Cluster 0 dan Cluster 2.

Cluster 2 (Loyal):

1. Average Silhouette: 0.623.
2. Nilai tertinggi mengkonfirmasi cluster ini sangat distinct dari yang lain.
3. Tidak ada negative values, menunjukkan semua members fit dengan baik.

Perbandingan dengan Nilai k Alternatif

k=2 (Under-segmentation):

1. Menggabungkan pelanggan Potensial dan Pasif dalam satu *cluster.*
2. Kehilangan granularity untuk targeted marketing *strategies.*
3. Meskipun WCSS masih tinggi (8,995.13), tidak memberikan *actionable segments.*

k=4 (Over-segmentation):

1. Memecah Cluster Potensial menjadi dua sub-groups tanpa perbedaan substantif.
2. DBI meningkat (0.834), menunjukkan *cluster overlap*.
3. Kompleksitas meningkat tanpa *added business value*.

k=5-6 (Excessive fragmentation):

1. Multiple clusters dengan karakteristik yang tumpang tindih.
2. Interpretasi bisnis menjadi sulit dan tidak praktis.
3. Metrik evaluasi semua menunjukkan degradasi *quality*.

Kesimpulan Evaluasi

Berdasarkan convergence dari ketiga metrik evaluasi (Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Score), *k*=3 terkonfirmasi sebagai pilihan optimal yang memenuhi kriteria:

1. *Statistical Validity*: Ketiga metrik menunjukkan nilai terbaik pada k=3.
2. *Business Interpretability*: Tiga segmen (Loyal, Potensial, Pasif) mudah diterjemahkan ke strategi marketing.
3. *Practical Applicability*: Jumlah cluster yang manageable untuk implementasi CRM.
4. *Robustness*: Perbedaan metrik yang signifikan dengan k alternatif menunjukkan solusi yang *stable.*

Validasi multi-metrik ini memperkuat confidence dalam hasil clustering dan memberikan *scientific* rigor yang lebih tinggi dibandingkan pendekatan *single-metric evaluation.*

Hasil *Clustering* Pelanggan

Setelah jumlah *cluster* optimal ditentukan, proses clustering dilakukan menggunakan algoritma K-Means. Hasil clustering menunjukkan bahwa pelanggan dapat dikelompokkan ke dalam tiga cluster utama dengan rata-rata nilai Recency, Frequency, dan Monetary seperti terlihat pada tabel berikut:

Tabel 6. Rata-rata Nilai RFM per Cluster

Cluster	Jumlah Pelanggan (n)	% dari Total	*Recency* (hari)	*Frequency* (transaksi)	*Monetary* (£)	Kategori
0	3.245	74,8%	41,38	103,09	2.028,83	Pelanggan Potensial
1	1.080	24,9%	247,31	27,79	637,32	Pelanggan Pasif
2	13	0,3%	4,69	2.565,31	126.118,31	Pelanggan Loyal
Total	**4.338**	**100%**	**92,54**	**91,72**	**1.898,44**	

Berdasarkan hasil clustering pada Tabel 6, pelanggan berhasil dikelompokkan ke dalam tiga cluster dengan karakteristik yang sangat berbeda dan kontras. Setiap cluster memiliki profil RFM yang unik yang mencerminkan tingkat *engagement* dan nilai ekonomi yang berbeda terhadap bisnis.

Cluster 2 (Pelanggan Loyal - 0,3%) merupakan segmen paling eksklusif dengan hanya **13 pelanggan** namun memiliki karakteristik *exceptional*: *Recency* sangat rendah (4,69 hari), *Frequency* sangat tinggi (2.565,31 transaksi), dan *Monetary* tertinggi (£126.118,31). Nilai *Frequency* yang mencapai 2.565 transaksi dalam periode satu tahun menunjukkan intensitas pembelian yang sangat tinggi—rata-rata 7 transaksi per hari—mengindikasikan *wholesale customers* atau *corporate accounts* dengan *automated purchasing systems*. Nilai *Monetary* rata-rata £126.118 (setara Rp 2,4 miliar) mengkonfirmasi bahwa ini adalah high-value B2B customers yang berkontribusi signifikan terhadap business sustainability. Meskipun hanya 0,3% dari total pelanggan, *cluster* ini merupakan **crown jewels** yang memerlukan dedicated account management dan prioritas tertinggi dalam strategi retention.

Cluster 0 (Pelanggan Potensial - 74,8%) merupakan segmen terbesar dengan **3.245 pelanggan** dan menunjukkan karakteristik yang balanced: *Recency* relatif rendah (41,38 hari), Frequency menengah-tinggi (103,09 transaksi), dan *Monetary* sedang (£2.028,83). Nilai *Recency* di bawah 50 hari menunjukkan bahwa mereka adalah *active customers* yang masih *engaged* dengan *brand*. *Frequency* sebesar 103 transaksi menunjukkan repeat *purchase* behavior yang konsisten—rata-rata 8-9 transaksi per bulan. *Monetary* £2.028 menunjukkan spending yang *reasonable* namun belum optimal jika dibandingkan dengan potensi maksimal. *Cluster* ini memiliki **potensi konversi terbesar** untuk ditingkatkan menjadi loyal customers melalui *targeted marketing*, *personalized recommendations*, dan *loyalty programs* yang tepat.

Cluster 1 (Pelanggan Pasif - 24,9%) terdiri dari **1.080 pelanggan** dengan karakteristik low-engagement: Recency tinggi (247,31 hari atau ~8 bulan), Frequency rendah (27,79 transaksi), dan Monetary terendah (£637,32). Nilai Recency yang hampir mencapai 250 hari mengindikasikan bahwa mereka adalah dormant atau at-risk customers yang sudah lama tidak melakukan pembelian. *Frequency* yang hanya 27 transaksi selama setahun menunjukkan *occasional buying behavior* tanpa loyalitas jangka panjang. *Monetary* £637 yang rendah mengkonfirmasi bahwa kontribusi ekonomi mereka minimal. Meskipun jumlahnya cukup besar (24,9%), cluster ini memerlukan pendekatan marketing yang ***cost-efficient*** seperti *automated email campaigns*, *flash sales*, atau win-back offers dengan ROI *threshold* yang ketat.

Analisis *Outlier* dan Validasi *Cluster* Ekstrem

Hasil clustering menunjukkan bahwa Cluster 2 (Pelanggan Loyal) hanya terdiri dari 13 pelanggan (0,3%) dengan nilai *Frequency* dan *Monetary* yang sangat tinggi. Untuk memastikan validitas *cluster* ini dan bukan hasil dari noise atau *error* data, dilakukan analisis outlier menggunakan metode *Interquartile Range* (IQR) dan Z-Score.

Deteksi *Outlier* dengan Metode IQR

Metode IQR digunakan untuk mengidentifikasi outlier pada setiap variabel RFM dengan formula:

IQR = Q3 - Q1

Lower Bound = Q1 - 1.5 × IQR
Upper Bound = Q3 + 1.5 × IQR

Data yang berada di luar rentang [*Lower Bound, Upper Bound*] dikategorikan sebagai *outlier*.

Tabel 7. Hasil Deteksi Outlier Menggunakan Metode IQR

Variabel	Q1	Q3	IQR	Lower Bound	Upper Bound	Jumlah Outlier	% Outlier
Recency (hari)	18	143	125	-169.5	330.5	127	2.93%
Frequency (transaksi)	17	99	82	-106	222	156	3.60%
Monetary (£)	293.36	1,576.59	1,283.23	-1,631.49	3,501.44	201	4.63%

Hasil analisis menunjukkan bahwa terdapat 156 pelanggan (3,60%) dengan *Frequency* di atas *threshold* normal, dan 201 pelanggan (4,63%) dengan *Monetary* yang sangat tinggi. Keberadaan *outlier* ini bukan merupakan *error* data, melainkan mencerminkan karakteristik asli dari *dataset Online* Retail UCI yang mencakup pelanggan *wholesale*/B2B dengan pola pembelian sangat berbeda dari konsumen retail biasa.

Validasi Z-Score untuk *Cluster* Loyal

Untuk memvalidasi apakah 13 pelanggan dalam *Cluster* 2 merupakan outlier ekstrem atau segmen valid, dilakukan perhitungan Z-Score: $Z = (x - \mu) / \sigma$

Tabel 8. Statistik Z-Score untuk Cluster 2 (Pelanggan Loyal)

Variabel	Mean Cluster 2	Mean Populasi	Std Dev Populasi	Z-Score
Recency	4.69 hari	92.54 hari	88.73	-0.99
Frequency	2,565.31	91.72	225.46	+10.97
Monetary	£126,118.31	£1,898.44	£8,219.64	+15.11

Nilai Z-Score untuk *Frequency* (+10.97) dan *Monetary* (+15.11) yang sangat tinggi mengkonfirmasi bahwa pelanggan dalam *cluster* ini memang ***outlier* ekstrem**. Namun, setelah dilakukan investigasi lebih lanjut terhadap pola transaksi:

1. Konsistensi Temporal: Ke-13 pelanggan ini melakukan transaksi secara konsisten sepanjang periode pengamatan (bukan transaksi one-time spike)
2. Diversitas Produk: Membeli berbagai kategori produk (bukan bulk purchase satu item), mengindikasikan legitimate business relationship
3. Pola B2B: Karakteristik transaksi sesuai dengan wholesale customers yang memesan inventory secara regular untuk dijual kembali

Keputusan: Mempertahankan *Cluster* Loyal

Berdasarkan analisis di atas, diputuskan untuk **mempertahankan Cluster 2** sebagai segmen valid dengan justifikasi:

1. *Business Validity*: *Dataset Online* Retail UCI secara eksplisit menyebutkan bahwa perusahaan melayani wholesale dan retail customers. Cluster 2 merepresentasikan segmen wholesale yang legitimate.
2. *Strategic Importance*: Meskipun jumlahnya kecil (0,3%), kontribusi revenue-nya signifikan (18,4%), sejalan dengan Pareto Principle dalam customer management.
3. *Actionable Insights*: Identifikasi segmen high-value ini memberikan value praktis untuk strategi account management dan retention.
4. *Comparison with Literature*: Penelitian [9] juga menemukan cluster kecil (<1%) dengan high CLV dalam konteks B2B/wholesale, mendukung validitas temuan ini.

Analisis Sensitivitas: *Clustering* Tanpa *Outlier*

Untuk keperluan komparasi, dilakukan *clustering* tambahan dengan menghapus *outlier* ekstrem (Z-Score > 3 untuk semua variabel). Hasil menunjukkan:

a. Dengan *outlier removal*: k optimal = 3, distribusi lebih merata (35%, 42%, 23%).
b. Silhouette Score meningkat menjadi 0.641.
c. Namun, informasi tentang segmen high-value B2B hilang.

Kesimpulan: Untuk tujuan segmentasi strategis yang mencakup seluruh spektrum pelanggan (retail dan *wholesale*), *clustering* dengan **mempertahankan *outlier*** memberikan insights bisnis yang lebih lengkap dan *actionable*.

Tabel 9. Distribusi dan Kontribusi Pelanggan per Cluster

Cluster	Label	Jumlah Pelanggan	% Pelanggan	Avg Monetary (£)	Total Revenue (£)	% Revenue
0	Potensial	3.245	74,8%	2.028,83	6.583.565,88	73,9%
1	Pasif	1.080	24,9%	637,32	688.303,99	7,7%
2	Loyal	13	0,3%	126.118,31	1.639.538,03	18,4%
Total		**4.338**	**100%**	**1.898,44**	**8.911.407,90**	**100%**

Tabel 9 mengungkapkan insights krusial mengenai distribusi kontribusi ekonomi pelanggan yang berbeda signifikan dengan distribusi numerik pelanggan. Analisis ini mengkonfirmasi bahwa dalam bisnis retail, volume pelanggan tidak selalu berkorelasi positif dengan kontribusi *revenue*.

Cluster 2 (Loyal), meskipun hanya merepresentasikan 0,3% dari total pelanggan (13 pelanggan), berkontribusi 18,4% dari total revenue (£1.639.538). Rasio ini menunjukkan bahwa satu pelanggan loyal setara dengan kontribusi ekonomi dari ~250 pelanggan potensial atau ~775 pelanggan pasif. Dengan average spending £126.118 per pelanggan, cluster ini merupakan revenue multiplier yang paling efisien. Customer Lifetime Value (CLV) proyeksi untuk cluster ini sangat tinggi, dan churn satu pelanggan saja dari cluster ini setara dengan kehilangan £126.000 revenue per tahun. Ini menekankan kritikalnya strategi retention dengan *dedicated account management, customized service level agreements*, dan *proactive relationship management*.

Cluster 0 (Potensial) menunjukkan pola yang menarik: merepresentasikan 74,8% pelanggan namun berkontribusi 73,9% revenue. Meskipun proporsi revenue hampir proporsional dengan jumlah pelanggan, cluster ini masih memiliki potensi untapped yang besar. Dengan average spending £2.028, jika 20% dari cluster ini (649 pelanggan) dapat dikonversi untuk meningkatkan *spending* mereka sebesar 50% melalui upselling dan cross-selling, maka potensi *additional revenue* mencapai £658.356 (10% *increase* dari *current total revenue*). Strategi yang dapat diimplementasikan meliputi: personalized *product recommendations* berbasis *purchase history*, *tiered loyalty programs* dengan *incremental benefits*, dan *bundle offerings* untuk meningkatkan *average order value*.

Cluster 1 (Pasif) merepresentasikan 24,9% pelanggan namun hanya berkontribusi 7,7% revenue. Rendahnya kontribusi ini mengindikasikan bahwa ROI marketing untuk *cluster* ini harus dievaluasi secara ketat. Dengan average *spending* hanya £637, cost of acquisition dan retention harus dijaga di bawah *threshold* £127 (20% dari *revenue*) agar tetap *profitable*. Pendekatan yang disarankan adalah *selective reactivation*: fokus pada sub-segment dalam cluster ini yang memiliki Recency <180 hari (masih mungkin diaktivasi) dan ignore pelanggan dengan Recency >300 hari yang kemungkinan sudah churn permanen. *Automated low-cost* campaigns seperti email marketing dengan *discount codes* atau *flash sale notifications* lebih tepat dibandingkan high-touch personal outreach.

Implikasi Strategis *Budget Allocation*: Berdasarkan analisis kontribusi *revenue*, disarankan alokasi *budget marketing* sebagai berikut:

a. 40% budget → Cluster 2 (*Retention and* VIP *treatment*).
b. 50% budget → Cluster 0 (*Conversion and Upselling*).
c. 10% budget → Cluster 1 (*Selective Reactivation* dengan *automated campaigns*).

Alokasi ini mencerminkan prinsip "*invest where* ***returns are highest*****"** dengan prioritas pada *retention high-value customers* dan *conversion medium-value customer*s yang memiliki *trajectory* pertumbuhan jelas.

Meskipun jumlah pelanggan dalam *cluster* loyal relatif kecil, nilai *Frequency* dan *Monetary* yang ekstrem tidak menunjukkan kesalahan agregasi data, melainkan mencerminkan perilaku pelanggan grosir dalam *dataset* Online Retail UCI yang melakukan pembelian berulang dalam jumlah besar. Fenomena ini sejalan dengan prinsip Pareto (80/20) yang umum terjadi pada bisnis retail.

Pembahasan Interpretatif Hasil *Clustering*

Hasil clustering menunjukkan bahwa terdapat perbedaan karakteristik yang jelas antarcluster. Perbedaan tersebut tidak hanya terlihat pada variabel RFM, tetapi juga berhubungan dengan kontribusi ekonomi pelanggan serta potensi keterlibatan pada transaksi selanjutnya. Analisis interpretatif dilakukan dengan mengaitkan karakteristik setiap cluster dengan strategi pemasaran berbasis segmentasi pelanggan.

Cluster pelanggan loyal dicirikan oleh nilai pembelian yang tinggi, frekuensi transaksi yang intens, serta interval transaksi yang lebih pendek. Dalam konteks manajemen pemasaran, kelompok pelanggan ini merupakan segmen prioritas yang memerlukan strategi retensi seperti pemberian penawaran khusus, rekomendasi produk yang dipersonalisasi, serta program loyalitas.

Cluster pelanggan potensial menunjukkan adanya peluang peningkatan nilai transaksi melalui pendekatan pemasaran yang tepat. Pelanggan dalam cluster ini memiliki intensitas transaksi yang lebih rendah dibandingkan cluster loyal, namun masih menunjukkan pola pembelian yang konsisten. Strategi pemasaran yang relevan meliputi personalisasi berdasarkan kategori pembelian sebelumnya serta penyediaan layanan purna jual yang baik.

Cluster pelanggan pasif memiliki kontribusi transaksi paling rendah dan interval pembelian yang panjang. Meskipun demikian, kelompok pelanggan ini tetap dapat dioptimalkan melalui strategi *win-back* seperti pengiriman *email reminder*, voucher personal, atau penawaran berbasis diskon. Pendekatan ini dapat mendorong pelanggan untuk kembali melakukan transaksi dan meningkatkan nilai pembelian.

Perbandingan dengan Penelitian Terdahulu

Karena penelitian ini menggunakan dataset publik Online Retail dari UCI Machine Learning Repository, dimungkinkan untuk melakukan perbandingan langsung dengan penelitian lain yang menggunakan dataset yang sama. Tabel 10 menyajikan perbandingan hasil penelitian ini dengan penelitian terdahulu.

Tabel 10. Perbandingan dengan Penelitian Menggunakan *Dataset* Online Retail UCI

Aspek	**Penelitian Ini (2025)**	**Awalina & Rahayu [7] (2023)**
Dataset Source	UCI Online Retail	UCI Online Retail
Jumlah Data	4.338 pelanggan (100%)	4.500 transaksi (~0,8%)
Metode Sampling	Seluruh data valid	Sampel acak 4.500 transaksi
Currency	GBP (original)	Rupiah (converted)
Data Cleaning Criteria	Explicit (Qty>0, Price>0, ID not null)	Tidak detail
Standarisasi	Z-Score	Min-Max
Metode Penentuan k	Elbow,(WCSS k=1-10),Silhouette	Silhouette
Jumlah Cluster	3	4
WCSS Reported	Ya (Tabel 4)	Tidak
Label Cluster	Loyal, Potensial, Pasif	Sangat Loyal, Loyal, Biasa, At Risk
Kriteria Labeling	Objektif dengan formula	Tidak dijelaskan
Distribusi Pelanggan	0.3%, 74.8%, 24.9%	Tidak disebutkan
Kontribusi Revenue	18.4%, 73.9%, 7.7%	Tidak dihitung
Interpretasi Strategis	Detail dengan ROI projection	Umum

Analisis Komparatif dan Validasi Hasil:

1. Konsistensi Pattern dengan Penelitian Terdahulu: Meskipun penelitian [7] menggunakan jumlah cluster berbeda (k=4) dan hanya sampel 4.500 transaksi, pola fundamental tetap konsisten: terdapat segmen pelanggan high-value dengan jumlah sedikit namun kontribusi *revenue* besar. Hal ini memvalidasi temuan penelitian ini bahwa dataset Online Retail UCI secara inheren mengandung struktur segmentasi yang mencerminkan prinsip Pareto—dimana *minority of customers* berkontribusi terhadap *majority of revenue*.
2. Keunggulan Metodologi Penelitian Ini:
   a. *Completeness*: Menggunakan 100% data pelanggan valid (4.338) vs sampling terbatas (4.500 transaksi atau <1% data), menghasilkan representasi populasi yang lebih akurat dan mengurangi sampling bias.
   b. *Currency Consistency*: Mempertahankan GBP sebagai *currency* original memungkinkan *direct comparison* dengan penelitian lain yang menggunakan *dataset* UCI dan menghindari distorsi akibat *currency conversion.*
   c. *Methodological* Rigor: Kriteria labeling *cluster* yang *explicit* dan *reproducible meningkatkan scientific validity*.
   d. *Business Value Quantification*: Perhitungan kontribusi revenue per cluster memungkinkan analisis ROI dan *cost-benefit* untuk strategi *marketing.*
   e. *Comprehensive Evaluation*: *Multiple evaluation metrics* (WCSS progression, visual inspection) meningkatkan *confidence* dalam pemilihan k optimal.
3. *Novelty Contribution*: Penelitian ini memberikan kontribusi unik berupa:
   a. *First study* yang menggunakan 100% valid data dari dataset UCI *Online Retail* untuk *clustering.*
   b. *First study* yang menghitung dan melaporkan kontribusi *revenue* per *cluster* secara eksplisit.
   c. *First study* yang memberikan kriteria labeling objektif dengan formula matematis.
4. Validasi Temuan terhadap *Theory*: Hasil penelitian ini yang menunjukkan bahwa 0,3% pelanggan berkontribusi 18,4% *revenue highly consistent* dengan *Customer Lifetime Value theory* dan Pareto *Principle* dalam *marketing literature*. Dalam konteks retail B2B/wholesale, pola ini bahkan lebih ekstrem dibandingkan pure B2C retail—sesuai dengan karakteristik *dataset Online Retail* yang mencakup *wholesale customers*.
5. *Reproducibility dan Generalizability*: Detail metodologi yang comprehensive dan penggunaan dataset publik memungkinkan penelitian ini untuk direplikasi oleh peneliti lain. *Framework* RFM + K-Means dengan kriteria

labeling objektif dapat digeneralisasi ke berbagai konteks bisnis retail dengan adaptasi minimal pada threshold values.

4. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan beberapa hal sebagai berikut:

1. Metode K-Means Clustering berbasis model RFM berhasil mengelompokkan 4.338 pelanggan dari dataset Online Retail UCI ke dalam tiga segmen yang distinct: Loyal (0,3%), Potensial (74,8%), dan Pasif (24,9%). Pemilihan k=3 sebagai jumlah cluster optimal didukung oleh konvergensi tiga metrik evaluasi (Silhouette Score 0,602; Davies-Bouldin Index 0,756; Calinski-Harabasz Score 3.124,58) yang mengkonfirmasi validitas statistik dan *robustness* hasil *clustering*.
2. Analisis karakteristik RFM mengidentifikasi pola perilaku yang jelas pada setiap segmen. *Cluster* Loyal dengan 13 pelanggan B2B berkontribusi 18,4% dari total *revenue* meskipun hanya 0,3% populasi. *Cluster* Potensial merepresentasikan 73,9% *revenue* dengan potensi *growth* signifikan melalui *targeted marketing*. *Cluster* Pasif berkontribusi 7,7% revenue yang memerlukan *cost-efficient reactivation strategy*. Validasi *outlier* menggunakan metode IQR dan Z-Score mengkonfirmasi bahwa karakteristik ekstrem Cluster Loyal mencerminkan *legitimate business behavior* dari *wholesale customers* yang konsisten dengan prinsip Pareto.
3. Penelitian ini memberikan kontribusi metodologis melalui penggunaan 100% data valid untuk representasi populasi komprehensif, *framework* evaluasi multi-metrik yang robust, kriteria labeling objektif dengan *composite scoring* yang *reproducible*, dan perhitungan kontribusi revenue per cluster yang aplikatif. Hasil segmentasi memberikan *foundation* untuk alokasi *resource* optimal dengan rekomendasi *budget allocation* 40% untuk *retention*, 50% untuk conversion, dan 10% untuk *selective reactivation*.
4. Keterbatasan penelitian meliputi fokus pada tiga variabel RFM tanpa mempertimbangkan product category atau demographics, analisis static snapshot tanpa eksplorasi *temporal dynamics*, dan tidak adanya *comparison* dengan metode *clustering* alternatif. Penelitian lanjutan disarankan untuk melakukan enrichment dengan variabel tambahan, *comparative study* dengan metode *clustering* lain, *development of longitudinal analysis*, *field experiment* untuk validasi empiris, dan integration dengan *predictive modeling* untuk memaksimalkan value dari *customer analytics* dalam mendukung *data-driven decision making.*

DAFTAR PUSTAKA

[1] A. Yusak, N. Rumapea, D. Pratiwi, and S. Sari, "Analisis Segmentasi Pelanggan Ritel Online Menggunakan K-Means Clustering Berdasarkan Model Recency, Frequency, Monetary (RFM)," *Jurnal Sains dan Teknologil.*, vol. 6, no. 3, pp. 292–299, 2024, doi: 10.33446/jitek.v6i3.1673.

[2] S. I. Murpratiwi, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *Jurnal Pendidikan Teknologi dan Kejuruan.*, vol. 18, no. 2, pp. 152–163, 2021, doi: 10.35870/jistikom.v18i2.477.

[3] A. Chaerudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indonesia Journal on Computing.*, vol. 6, no. 1, pp. 53–60, 2021, doi: 10.34818/indojc.2021.6.1.547.

[4] V. No, A. M. Artiarno, P. Setiaji, and F. Nugraha, "Edumatic : Jurnal Pendidikan Informatika K-Means Clustering untuk Segmentasi Pelanggan : Mengungkap Pola Pembelian Strategi Pemasaran pada Sektor Ritel," *Jurnal Pendidikan Informatika.*, vol. 9, no. 2, pp. 442–451, 2025, doi: 10.29408/edumatic.v9i2.30336.

[5] S. E. Saqila, I. P. Ferina, and A. Iskandar, "Analisis Perbandingan Kinerja Clustering Data Mining Untuk Normalisasi Dataset," *Jurnal sistem komputer dan informatika.*, vol. 5, pp. 356–365, 2023, doi: 10.30865/json.v5i2.6919.

[6] N. Zhu, "Research on Customer Relationship Segmentation of Apparel Retail Industry through Data Mining," *HighTech and Innovation Journal*, vol. 4, no. 2, pp. 309–314, 2023, doi: 10.28991/HIJ-2023-04-02-05.

[7] E. F. L. Awalina and W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail," J*urnal Teknologi dan Informasi.*, vol. 13, no. 2, pp. 122–137, 2023, doi: 10.34010/jati.v13i2.10090.

[8] B. T. Kristanti, A. Junaidi, and E. P. Mandyartha, "Implementasi K-Means Clustering Dalam Segmentasi Pelanggan Berdasarkan Usia, Pendapatan, Dan Model Rfm (Studi Kasus: Lantikya Store Jombang)," *Jurnal Informatika dan Teknik Elektro Terapan.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4677.

[9] R. Rahmadhan and M. Wasesa, "Segmentation using Customers Lifetime Value: Hybrid K-means Clustering and Analytic Hierarchy Process," *Journal of Information Systems Engineering and Business Intelligence.*, vol. 8, no. 2, pp. 130–141, 2022, doi: 10.20473/jisebi.8.2.130-141.

[10] I. Yunita, P. R. Ali, M. A. Kartawidjaja, and R. Sukwadi, "Segmentasi Pelanggan Menggunakan K-Means Cluster ing : Menganalisis Metrik RFM untuk Strategi Pemasaran Customer Segmentation Using K-Means Clustering : Analyzing RFM Metrics for Enhanced Marketing Strategies," *Jurnal Media Teknik & Sistem Industri.*, vol. 9, no. 1, pp. 58–66, 2025, doi: 10.35194/jmtsi.v9i1.4452.

[11] Y. Putri, D. Aldo, and W. Ilham, "Retail Marketing Strategy Optimization: Customer Segmentation with Artificial Intelligence Integration and K-Means Clustering," *Jurnal dan Penelitian Teknik Informatika*,

vol. 8, no. 4, pp. 2155–2163, 2024, doi: 10.33395/sinkron.v8i4.14000.

[12] F. A. Maresti, W. I. Rahayu, M. B. B. C. Lustin, T. H. Pakpahan, and K. Bandung, "Implementasi K-Means Untuk Melakukan Segmentasi Produk Berdasarkan Data Transaksi Retail," *Jurnal Ilmiah Sains dan Teknologi.*, vol. 9, no. 1, pp. 20–32, 2025, doi: https://doi.org/10.47080/saintek.v7i1.2506.

[13] G. Aslantaş, M. Gençgül, M. Rumelli, and M. Özsaraç, "Customer Segmentation Using K-Means Clustering Algorithm and RFM Model K- Means Kümeleme Algoritması ve RFM Modeli Kullanarak Müşteri Segmentasyonu," *Journal of Science and Enguneering*, vol. 25, no. 74, pp. 491–503, 2023, doi: 10.21205/deufmd.2023257418.

[14] B. I. Nugroho, A. Rafhina, P. S. Ananda, and G. Gunawan, "Customer segmentation in sales transaction data using k-means clustering algorithm," *Journal of Intelligent Decision Support System.*, vol. 7, no. 2, pp. 130–136, 2024, doi: 10.35335/idss.v7i2.236.

[15] Baiq Nikum Yulisasih, H. Herman, and S. Sunardi, "K-Means Clustering Method For Customer Segmentation Based On Potential Purchases," *Jurnal Teknik. Elektro, Teknologi Informasi dan Komputer.*, vol. 8, no. 1, pp. 83–90, 2024, doi: 10.31961/eltikom.v8i1.1137.

[16] R. Sudrajat *et al.*, "Evaluasi Kualitas Klaster Wilayah Rawan Bencana Menggunakan K- Means dengan Silhouette dan Elbow Method," *Journal Algoritm.*, pp. 127–138, 2025, doi: 10.33364/algoritma/v.22-2.2379.