p-ISSN: 2963-7856| e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No. 3, Januari-Maret 2026, pp 15993-16002

Segmentasi Pelanggan Grosir Menggunakan K-Means: Analisis Outlier dan Ketidakseimbangan Data

N Tahta Phudjashakty[1*], Hasbi Firmansyah[2], Wahyu Asriyani[3], Ali Sofyan[4]

[1,2,4] Fakultas Teknik dan Ilmu Komputer, Prodi Informatika, Universitas Pancasakti Tegal, Jl. Halmahera KM. 01, Mintaragen, Tegal Tim., Kota Tegal, Jawa Tengah, 52121, Indonesia.
[3]Fakultas Keguruan dan Ilmu Pendidikan, Prodi Pendidikan Bahasa dan Sastra Indonesia, Universitas Pancasakti Tegal, Jl. Halmahera KM. 01, Mintaragen, Tegal Tim., Kota Tegal, Jawa Tengah, 52121, Indonesia.
E-mail: pujasakti31@gmail.com
* Corresponding Author

https://doi.org/10.31004/jerkin.v4i3.4771

ARTICLE INFO

Article history:
Received: 21 Dec 2025
Revised: 27 Dec 2025
Accepted: 02 Jan 2026

Kata Kunci:
Segmentasi Pelanggan, K-Means, Pelanggan Grosir, Pencilan, Data Tidak Seimbang.

Keywords:
Customer Segmentation, K-Means, Wholesale Customers, Outliers, Imbalanced Data.

ABSTRACT

Penelitian ini bertujuan melakukan segmentasi pelanggan grosir menggunakan algoritma K-Means serta mengkaji pengaruh outlier dan ketidakseimbangan data terhadap hasil clustering. Data yang digunakan adalah Wholesale Customers Dataset dari UCI Machine Learning Repository yang berisi 440 pelanggan dengan delapan atribut numerik pembelian tahunan. Tahap pra-pengolahan meliputi eksplorasi data, deteksi outlier menggunakan Z-Score dan boxplot, penanganan nilai ekstrem dengan winsorizing, serta normalisasi Z-Score agar skala atribut sebanding. Jumlah cluster ditentukan menggunakan Elbow Method. Penerapan K-Means dengan $k = 2$menghasilkan dua cluster yang sangat timpang, yaitu 437 pelanggan pada Cluster 0 dan 3 pelanggan pada Cluster 1. Cluster 0 merepresentasikan pelanggan reguler dengan pola pembelian mendekati rata-rata, sedangkan Cluster 1 berisi pelanggan dengan pembelian sangat tinggi, terutama pada kategori Frozen dan Delicassen. Evaluasi menggunakan average within centroid distance dan Davies–Bouldin Index menunjukkan bahwa setelah penanganan outlier dan normalisasi, struktur cluster menjadi lebih stabil dan lebih mudah diinterpretasikan. Hasil segmentasi ini dapat dimanfaatkan untuk merancang strategi pemasaran dan layanan berbeda antara pelanggan reguler dan pelanggan high spender, sekaligus menegaskan pentingnya pra-pengolahan data pada penerapan K-Means.

This study aims to segment wholesale customers using the K-Means clustering algorithm and to examine the impact of outliers and data imbalance on the clustering results. The data are taken from the Wholesale Customers Dataset of the UCI Machine Learning Repository, consisting of 440 customers with eight numerical attributes representing annual purchase amounts. The preprocessing steps include exploratory data analysis, outlier detection using Z-Score and boxplot visualization, handling of extreme values with winsorizing, and Z-Score normalization to make the attribute scales comparable. The number of clusters is determined using the Elbow Method. Applying K-Means with $k = 2$produces two highly imbalanced clusters, with 437 customers in Cluster 0 and 3 customers in Cluster 1. Cluster 0 represents regular customers whose purchasing patterns are close to the overall average, while Cluster 1 consists of customers with very high purchases, especially in Frozen and Delicassen categories. Evaluation using the average within centroid distance and the Davies–Bouldin Index shows that, after outlier handling and normalization, the cluster structure becomes more stable and easier to interpret. The resulting segmentation can support differentiated marketing and service strategies for regular and high-spending customers and highlights the importance of proper preprocessing when applying K-Means.

How to Cite: N Tahta Phudjashakty, et al (2025), Segmentasi Pelanggan Grosir Menggunakan K-Means: Analisis Outlier dan Ketidakseimbangan Data, 4(3). https://doi.org/10.31004/jerkin.v4i3.4771

PENDAHULUAN

Di era digital saat ini, data transaksi yang dimiliki perusahaan bukan lagi sekadar catatan administratif, tetapi sudah menjadi aset penting yang dapat dimanfaatkan untuk mendukung pengambilan keputusan strategis. Hal ini juga berlaku bagi perusahaan yang bergerak di bidang perdagangan grosir. Tanpa pemahaman yang baik tentang perilaku pelanggan, perusahaan akan kesulitan menyusun strategi pemasaran yang tepat sasaran, mengatur stok barang, dan menjaga hubungan dengan pelanggan yang bernilai tinggi (Julian et al., 2023).

Salah satu pendekatan yang banyak digunakan untuk memahami pola perilaku pelanggan adalah segmentasi pelanggan, yaitu mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan kemiripan karakteristik atau pola pembeliannya (Aggarwal, 2017). Dengan adanya segmentasi, perusahaan dapat memberikan perlakuan yang berbeda untuk tiap segmen, misalnya program promosi khusus, penawaran paket tertentu, atau layanan yang lebih personal untuk pelanggan-pelanggan penting.

Dalam konteks data mining, algoritma K-Means Clustering menjadi salah satu metode yang paling sering digunakan untuk melakukan segmentasi karena sifatnya yang sederhana, efisien, dan cocok untuk data numerik berdimensi banyak (Aggarwal, 2015). Di Indonesia, berbagai penelitian telah memanfaatkan K-Means untuk mengelompokkan pelanggan berdasarkan data transaksi maupun model Recency, Frequency, Monetary (RFM), dan menunjukkan bahwa K-Means mampu mengidentifikasi kelompok pelanggan potensial dan pelanggan bernilai tinggi.

Meski demikian, banyak studi lebih menonjolkan aspek penerapan algoritma dan hasil segmentasinya, sementara dampak outlier dan ketidakseimbangan data terhadap hasil clustering sering kali kurang dibahas secara mendalam. Padahal, dalam data transaksi nyata, sangat wajar jika terdapat pelanggan yang pembeliannya jauh lebih besar daripada pelanggan lain. Pelanggan-pelanggan ini berperan sebagai outlier dan dapat menyebabkan centroid bergeser jauh dari posisi yang mewakili mayoritas pelanggan (Aggarwal, 2017). Di sisi lain, distribusi yang timpang (imbalanced) juga dapat membuat sebagian kecil pelanggan ekstrem membentuk cluster kecil yang sangat berbeda, sementara sebagian besar pelanggan terkumpul dalam satu cluster besar.

Beberapa literatur menyarankan agar tahap pra-pengolahan data, seperti penanganan outlier dan normalisasi, dilakukan dengan serius sebelum menerapkan algoritma clustering (Tan et al., 2019). Penelitian-penelitian segmentasi pelanggan di Indonesia pun menunjukkan bahwa kombinasi K-Means dengan persiapan data dan model perilaku seperti RFM mampu menghasilkan cluster yang lebih bermakna secara bisnis.

Berangkat dari latar belakang tersebut, penelitian ini difokuskan pada dua hal utama. Pertama, menerapkan algoritma K-Means untuk melakukan segmentasi pelanggan pada Wholesale Customers Dataset. Kedua, mengkaji lebih dalam bagaimana outlier dan ketidakseimbangan data memengaruhi struktur cluster yang terbentuk, terutama sebelum dan sesudah dilakukan penanganan outlier dan normalisasi. Harapannya, hasil penelitian ini tidak hanya menghasilkan segmentasi pelanggan yang berguna secara praktis, tetapi juga memberikan gambaran yang lebih jelas tentang peran kualitas data dalam penerapan K-Means untuk segmentasi pelanggan grosir(Oktavian et al., 2025).

METODE

Data Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *clustering* berbasis algoritma K-Means untuk membentuk kelompok pelanggan berdasarkan pola pembelian. Data penelitian diambil dari Wholesale Customers Dataset yang tersedia pada UCI Machine Learning Repository. Dataset ini terdiri dari 440 baris data pelanggan dengan delapan atribut numerik, yaitu *Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper,* dan *Delicassen*. Setiap atribut merepresentasikan total pembelian tahunan dalam satuan moneter untuk kategori produk terkait.

Data diunduh dalam format CSV dan selanjutnya diolah menggunakan dua perangkat utama, yaitu Microsoft Excel dan RapidMiner Studio. Microsoft Excel digunakan untuk perhitungan manual dan pemahaman alur algoritma secara bertahap, sedangkan RapidMiner digunakan untuk menjalankan algoritma K-Means secara otomatis pada seluruh dataset sehingga hasilnya dapat dievaluasi secara menyeluruh.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Eksplorasi Data

Tahap awal yang dilakukan adalah eksplorasi data. Pada tahap ini, dilakukan pemeriksaan terhadap tipe data untuk memastikan bahwa seluruh atribut yang digunakan merupakan data numerik, pengecekan keberadaan nilai kosong (*missing values*), serta perhitungan statistik deskriptif seperti nilai minimum, maksimum, rata-rata, dan standar deviasi pada masing-masing atribut. Hasil eksplorasi ini kemudian dijadikan dasar penyusunan tabel statistik deskriptif pada bagian hasil, sekaligus memberikan gambaran awal mengenai sebaran dan rentang nilai tiap atribut sebelum memasuki tahapan *preprocessing*.

Deteksi dan Penangan Outlier

Setelah karakteristik umum data diketahui, penelitian dilanjutkan dengan tahapan deteksi dan penanganan outlier. Deteksi outlier dilakukan dengan mengombinasikan analisis Z-Score dan visualisasi boxplot (Aggarwal, 2017). Nilai Z-Score digunakan untuk mengukur seberapa jauh suatu nilai menyimpang dari rata-rata, dengan rumus:

$$\boldsymbol{Z = \frac{x - \mu}{\sigma}} \tag{1}$$

di mana $\boldsymbol{x}$adalah nilai asli data, $\boldsymbol{\mu}$adalah nilai rata-rata, dan $\boldsymbol{\sigma}$adalah standar deviasi dari atribut yang bersangkutan. Data dengan nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3 dianggap sebagai kandidat outlier karena berada jauh di luar pola umum distribusi data.

Hasil perhitungan Z-Score kemudian diperkuat dengan visualisasi boxplot sehingga atribut yang memiliki nilai ekstrem dapat diidentifikasi dengan lebih mudah. Nilai-nilai yang sangat jauh dari rentang utama pada boxplot menunjukkan keberadaan pelanggan dengan pembelian yang tidak biasa, baik sangat tinggi maupun sangat rendah jika dibandingkan mayoritas pelanggan lain(Penulis, 2025).

Untuk mengurangi pengaruh nilai ekstrem tersebut tanpa harus menghapus data pelanggan dari dataset, penelitian ini menerapkan teknik winsorizing (capping). Secara konsep, nilai yang berada di bawah persentil ke-5 ($\boldsymbol{P_5}$) dan di atas persentil ke-95 ($\boldsymbol{P_{95}}$) disesuaikan agar berada pada batas tersebut. Penyesuaian nilai secara sederhana dapat dinyatakan sebagai berikut:

$$\boldsymbol{x}_{\text{new}} = \begin{cases} \boldsymbol{P_5}, & \boldsymbol{x < P_5} \\ \boldsymbol{x}, & \boldsymbol{P_5 \le x \le P_{95}} \\ \boldsymbol{P_{95}}, & \boldsymbol{x > P_{95}} \end{cases} \tag{2}$$

dengan $\boldsymbol{x}_{\text{new}}$adalah nilai setelah penyesuaian, sedangkan $\boldsymbol{P_5}$dan $\boldsymbol{P_{95}}$masing-masing adalah persentil ke-5 dan ke-95. Dengan cara ini, informasi umum tetap dipertahankan, tetapi pengaruh titik-titik yang sangat ekstrem terhadap proses pembentukan cluster dapat diminimalkan.

Normalisasi Data

Seluruh variabel transaksi pada dataset memiliki rentang dan skala nilai yang berbeda-beda. Beberapa atribut, seperti *Fresh* dan *Grocery*, memiliki nilai pembelian yang relatif besar, sedangkan atribut lain seperti *Delicassen* cenderung memiliki nilai yang lebih kecil. Jika perbedaan skala ini dibiarkan, atribut berskala besar akan lebih dominan dalam perhitungan jarak sehingga dapat menyebabkan bias pada hasil clustering.

Untuk mengatasi hal tersebut, dilakukan normalisasi Z-Score pada seluruh atribut numerik. Normalisasi ini menggunakan rumus yang sama dengan perhitungan Z-Score, yaitu:

$$\boldsymbol{Z = \frac{x - \mu}{\sigma}} \tag{3}$$

Setelah dinormalisasi, setiap atribut memiliki rata-rata mendekati nol dan simpangan baku sekitar satu. Dengan demikian, kontribusi masing-masing atribut dalam perhitungan jarak menjadi lebih seimbang. Proses normalisasi mula-mula diuji dan divalidasi menggunakan Excel sebagai bagian dari perhitungan manual, kemudian dilanjutkan dan diterapkan secara konsisten di RapidMiner menggunakan operator normalisasi yang setara.

Penentuan Jumlah Cluster

Pemilihan jumlah cluster ($\boldsymbol{k}$) pada K-Means tidak dilakukan secara acak, melainkan melalui pendekatan Elbow Method dan pertimbangan kualitas cluster. Beberapa nilai k dicoba, lalu untuk masing-masing nilai dihitung Sum of Squared Errors (SSE). Secara umum, SSE dirumuskan sebagai:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{m_j} \|x_i - c_j\|^2 \quad (4)$$

dengan $k$adalah jumlah cluster, $m_j$adalah jumlah data pada cluster ke-j, $x_i$adalah data ke-i dalam cluster ke-j, dan $c_j$adalah centroid cluster ke-j. Nilai SSE menunjukkan tingkat kekompakan data di sekitar centroid. Semakin kecil SSE, semakin rapat data di dalam cluster tersebut.

Ketika jumlah cluster terus ditambah, SSE cenderung menurun; namun penurunan tersebut tidak selalu signifikan. Oleh karena itu, nilai SSE diplot terhadap nilai k, lalu diamati titik di mana penurunan SSE mulai melandai. Titik tersebut dikenal sebagai "siku" atau elbow dan digunakan sebagai acuan pemilihan k yang efisien (Aggarwal, 2015).

Selain SSE, penelitian juga merujuk pada konsep Silhouette Coefficient sebagai ukuran tambahan untuk menilai kualitas pemisahan cluster. Silhouette untuk data ke-i dapat dituliskan sebagai:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (5)$$

Dengan $a(i)$adalah rata-rata jarak data $i$terhadap seluruh anggota cluster-nya sendiri, dan $b(i)$adalah rata-rata jarak data $i$terhadap anggota cluster terdekat selain cluster-nya sendiri. Nilai Silhouette yang mendekati 1 menunjukkan bahwa data berada dalam cluster yang tepat, sedangkan nilai yang mendekati 0 atau bernilai negatif menunjukkan bahwa posisi data berada di perbatasan atau cenderung tidak sesuai dengan cluster tempatnya berada. Berdasarkan hasil analisis SSE dan pertimbangan interpretasi bisnis, jumlah cluster yang digunakan dalam penelitian ini adalah k = 2.

Implementasi Algoritma K-Means

Setelah jumlah cluster ditetapkan, algoritma K-Means diimplementasikan menggunakan dua pendekatan, yaitu perhitungan manual dan pemodelan otomatis. Perhitungan manual dilakukan di Microsoft Excel untuk sejumlah data contoh agar alur algoritma dapat dipahami secara jelas. Dalam perhitungan manual tersebut, centroid awal dipilih secara acak, kemudian jarak antara setiap data dengan masing-masing centroid dihitung menggunakan jarak Euclidean dengan rumus:

$$d(x, c) = \sqrt{\sum_{n=1}^{p} (x_n - c_n)^2} \quad (6)$$

Di mana $d(x, c)$adalah jarak antara data $x$dan centroid c, $x_n$adalah nilai variabel ke-n pada data x, $c_n$adalah nilai variabel ke-n pada centroid c, dan $p$adalah jumlah variabel. Setiap data kemudian dialokasikan ke cluster dengan jarak terkecil terhadap centroid(Pramudiansyah & Munte, 2021).

Setelah pembagian cluster awal dilakukan, nilai centroid baru dihitung sebagai rata-rata seluruh data yang berada di dalam cluster tersebut. Secara matematis, centroid baru untuk cluster ke-j dapat dirumuskan sebagai:

$$c_j^{\text{baru}} = \frac{1}{m} \sum_{i=1}^{m} x_i \quad (7)$$

Dengan m adalah jumlah anggota cluster ke-j dan $x_i$adalah data ke-i pada cluster tersebut. Proses penghitungan jarak dan pembaruan centroid ini diulang hingga model dianggap konvergen, yaitu ketika perubahan posisi centroid antar iterasi sudah sangat kecil. Secara konseptual, kondisi konvergensi dapat dinyatakan sebagai:

$$\left\| c_j^{baru} - c_j^{lama} \right\| < \varepsilon \quad (8)$$

dengan εadalah nilai ambang batas perubahan yang sangat kecil.

Selain perhitungan manual, algoritma K-Means juga dijalankan secara otomatis menggunakan RapidMiner Studio. Pada tahap ini, data yang telah melalui proses pembersihan, penanganan outlier,

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

dan normalisasi dimasukkan ke dalam operator K-Means di RapidMiner. Konfigurasi jumlah cluster ditetapkan sebesar k = 2 dengan metrik jarak Euclidean. RapidMiner kemudian menghasilkan informasi mengenai jumlah anggota cluster, nilai average within centroid distance, nilai Davies–Bouldin Index, serta nilai centroid untuk setiap atribut pada masing-masing cluster. Hasil inilah yang kemudian dianalisis lebih lanjut pada bagian hasil dan pembahasan(Nugraha et al., 2025).

Analisis Hasil Cluster

Tahap akhir dalam metode penelitian ini adalah analisis terhadap hasil clustering yang diperoleh. Analisis dilakukan dengan memperhatikan dua aspek utama. Aspek pertama adalah struktur dan distribusi anggota cluster, termasuk seberapa seimbang atau timpang jumlah anggota pada masing-masing cluster serta bagaimana indikator evaluasi internal, seperti average within centroid distance dan Davies–Bouldin Index, menggambarkan kualitas pemisahan cluster yang dihasilkan.

Aspek kedua adalah interpretasi nilai centroid pada setiap atribut untuk masing-masing cluster. Nilai centroid ini digunakan untuk menyusun profil pelanggan di tiap cluster, misalnya untuk membedakan segmen pelanggan reguler dengan segmen pelanggan ekstrem atau *high spender* berdasarkan besaran dan pola pembelian tahunan. Hasil analisis ini kemudian dikaitkan dengan konteks bisnis dan dibandingkan secara singkat dengan temuan pada penelitian segmentasi pelanggan berbasis K-Means di Indonesia, sehingga posisi penelitian ini menjadi lebih jelas di antara studi-studi sejenis (Maskanah et al., 2020).

HASIL DAN PEMBAHASAN

Gambaran Umum Dataset

Langkah awal yang dilakukan adalah melihat lebih dekat karakteristik tiap atribut melalui statistik deskriptif. Hasilnya dirangkum dalam Tabel 1 yang memuat nilai minimum, maksimum, rata-rata, dan standar deviasi untuk setiap atribut pembelian.

Tabel 1. Statistik Deskriptif Dataset Pelanggan Grosir

Atribut	Cluster_0	Cluster_1
Channel	0,005	-0,690
Region	-0,001	0,159
Fresh	-0,008	1,183
Milk	-0,014	2,097
Grocery	-0,003	0,241
Frozen	-0,058	8,459
Detergents_Paper	0,003	-0,494
Delicassen	-0,042	6,108

Dari Tabel 1, terlihat bahwa atribut seperti *Fresh* dan *Grocery* memiliki rata-rata dan nilai maksimum yang cukup tinggi dibandingkan atribut lain, sedangkan *Delicassen* cenderung lebih kecil. Selain itu, untuk beberapa atribut, selisih antara nilai minimum dan maksimum sangat besar. Kondisi ini memberi sinyal bahwa ada sebagian kecil pelanggan dengan pembelian sangat besar dibanding mayoritas pelanggan lain. Secara statistik, inilah yang nantinya akan muncul sebagai outlier, yang jika tidak ditangani dapat mengganggu pembentukan cluster (Ramadhona et al., 2022).

Deskripsi Dataset

Penelitian ini menggunakan data transaksi pelanggan grosir dari Wholesale Customers Dataset yang berisi 440 pelanggan dengan delapan atribut numerik, yaitu *Channel*, *Region, Fresh*, *Milk, Grocery*, *Frozen, Detergents_Paper*, dan *Delicassen*. Masing-masing atribut menggambarkan total pembelian tahunan pelanggan pada kategori produk terkait, sehingga cukup representatif untuk menggambarkan pola konsumsi di segmen grosir.

Ringkasan statistik deskriptif untuk tiap atribut sebelumnya telah disajikan dalam Tabel 1. Dari tabel tersebut terlihat bahwa beberapa atribut, seperti *Fresh* dan *Grocery*, memiliki rata-rata dan nilai maksimum yang jauh lebih besar dibanding atribut lain, sementara *Delicassen* cenderung memiliki nilai rata-rata lebih kecil. Rentang nilai yang lebar ini memberi indikasi awal bahwa terdapat pelanggan-pelanggan dengan pembelian sangat besar, yang nantinya akan muncul sebagai outlier dan memengaruhi struktur cluster jika tidak ditangani dengan baik(Pangestu et al., 2023).

Setelah proses normalisasi Z-Score diterapkan, nilai-nilai atribut berubah menjadi skala baku dengan rata-rata mendekati nol. Contoh hasil normalisasi untuk sepuluh baris pertama ditampilkan pada Gambar 1.

Row No.	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
1	1.447	0.590	0.053	0.523	-0.041	-0.589	-0.044	-0.066
2	1.447	0.590	-0.391	0.544	0.170	-0.270	0.086	0.089
3	1.447	0.590	-0.447	0.408	-0.028	-0.137	0.133	2.241
4	-0.690	0.590	0.100	-0.623	-0.393	0.688	-0.498	0.093
5	1.447	0.590	0.839	-0.052	-0.079	0.174	-0.232	1.299
6	1.447	0.590	-0.205	0.334	-0.297	-0.496	-0.228	-0.026
7	1.447	0.590	0.010	-0.352	-0.103	-0.534	0.054	-0.347
8	1.447	0.590	-0.350	-0.114	0.155	-0.289	0.092	0.369
9	-0.690	0.590	-0.477	-0.291	-0.195	-0.545	-0.244	-0.275
10	1.447	0.590	-0.474	0.718	1.150	-0.394	0.953	0.203

Gambar 1. Hasil normalisasi Z-Score pada atribut pelanggan

Gambar 1 memperlihatkan bahwa nilai untuk setiap atribut sudah berada pada kisaran yang relatif seragam, umumnya antara sekitar -1 hingga 1. Misalnya, pada baris ke-1 nilai *Channel* sebesar 1,447, *Region* sebesar 0,590, dan *Fresh* sebesar 0,053, sedangkan pada beberapa baris lain terdapat nilai negatif seperti -0,690 pada *Channel* atau -0,589 pada *Frozen*. Nilai-nilai tersebut menandakan seberapa jauh suatu pembelian berada di atas atau di bawah rata-rata secara terstandarisasi. Dengan skala yang sudah seragam seperti ini, tidak ada lagi atribut yang terlalu mendominasi perhitungan jarak dalam K-Means hanya karena perbedaan satuan atau rentang nilai.

Analisis Outlier

Setelah gambaran umum data diperoleh, langkah berikutnya adalah mengidentifikasi adanya nilai-nilai ekstrem yang berpotensi menjadi outlier. Deteksi dilakukan menggunakan perhitungan Z-Score untuk setiap atribut dan visualisasi boxplot. Data dengan nilai Z-Score lebih besar dari 3 atau lebih kecil dari -3 diperlakukan sebagai kandidat outlier karena dianggap terlalu jauh dari rata-rata.

Hasil visualisasi boxplot (tidak ditampilkan di sini) menunjukkan bahwa beberapa atribut, terutama *Fresh*, *Milk*, *Grocery*, *Frozen*, dan *Delicassen*, memiliki sejumlah titik yang terletak jauh di luar *whisker* bagian atas. Titik-titik ini merepresentasikan pelanggan dengan pembelian sangat besar pada kategori produk tertentu. Apabila nilai-nilai ekstrem tersebut langsung dimasukkan ke proses K-Means tanpa penyesuaian, centroid cluster dapat tertarik ke arah pelanggan ekstrem tersebut sehingga cluster yang terbentuk menjadi bias dan kurang mencerminkan pola pembelian mayoritas(Wahyuni et al., 2023).

Untuk mengurangi efek ini, diterapkan teknik winsorizing (capping) pada nilai-nilai yang berada di luar persentil tertentu. Nilai yang terlalu rendah dinaikkan ke batas bawah, sedangkan nilai yang terlalu tinggi dipotong ke batas atas yang wajar. Dengan cara ini, informasi mengenai pelanggan tetap dipertahankan, tetapi pengaruhnya terhadap rata-rata dan perhitungan jarak menjadi lebih terkendali.

Analisis Ketidakseimbangan Data

Selain permasalahan outlier, dataset juga menunjukkan karakter ketidakseimbangan distribusi antar pelanggan. Mayoritas pelanggan memiliki pembelian yang relatif rendah hingga menengah, sedangkan hanya segelintir pelanggan yang memiliki pembelian sangat tinggi. Pola ini terlihat dari perbandingan nilai minimum, maksimum, dan rata-rata pada Tabel 1 yang menunjukkan perbedaan cukup mencolok, serta dari histogram distribusi yang cenderung miring ke kanan (*right-skewed*)(Iqbal et al., 2025).

Ketidakseimbangan ini berimplikasi langsung pada hasil clustering. Struktur data seperti ini secara alami akan mendorong terbentuknya satu cluster besar yang menampung hampir semua pelanggan dengan pola pembelian "biasa", dan satu cluster kecil yang berisi pelanggan-pelanggan dengan pola pembelian ekstrem. Pola tersebut memang terbukti pada hasil K-Means, di mana salah satu cluster berisi 437 pelanggan dan cluster lainnya hanya berisi 3 pelanggan. Situasi ini menegaskan bahwa isu *imbalanced data* tidak bisa diabaikan ketika menerapkan algoritma clustering pada data transaksi dunia nyata(Ramadhan, 2023).

Penentuan Jumlah Cluster

Penentuan jumlah cluster (k) dilakukan menggunakan Elbow Method dengan menghitung Sum of Squared Errors (SSE) untuk beberapa kandidat nilai k. Ketika nilai SSE diplot terhadap k, terlihat

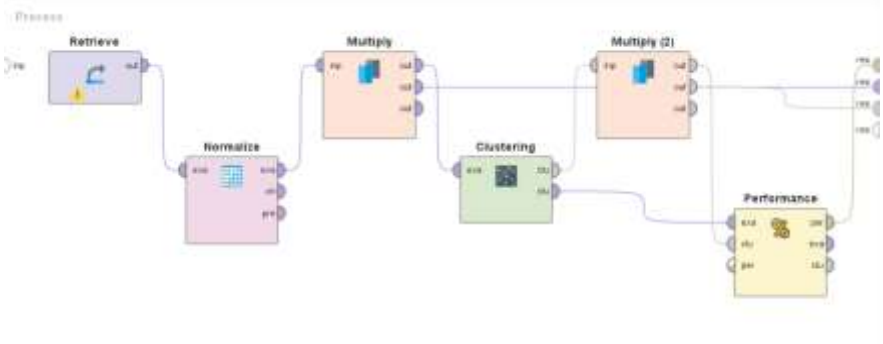
ISSN 2963-7856 (print), ISSN 2961-9890 (online)

bahwa penurunan SSE cukup tajam ketika jumlah cluster dinaikkan dari satu menjadi dua, namun penurunannya mulai melandai pada k > 2. Pola tersebut menunjukkan bahwa penambahan jumlah cluster di atas dua tidak lagi memberikan penurunan kesalahan yang signifikan(Silamantha & Hadiono, 2024).

Selain mempertimbangkan bentuk kurva SSE, aspek interpretasi bisnis juga diperhatikan. Dengan k = 2, pelanggan dapat dikelompokkan secara intuitif ke dalam dua segmen besar: satu segmen pelanggan reguler dan satu segmen pelanggan dengan perilaku pembelian yang ekstrem. Oleh karena itu, nilai k = 2 dipilih sebagai jumlah cluster yang digunakan dalam penelitian ini.

Proses Clustering di RapidMiner

Secara implementasi, proses pemodelan K-Means di RapidMiner disusun dalam bentuk rangkaian operator yang saling terhubung. Rangkaian proses tersebut ditunjukkan pada Gambar 2.



Gambar 2. Rangkaian proses K-Means pada RapidMiner

Gambar 2 memperlihatkan bahwa data terlebih dahulu diambil dari repositori menggunakan operator *Retrieve*, kemudian dinormalisasi melalui operator *Normalize*. Setelah itu, data digandakan dengan operator *Multiply* sehingga satu salinan dapat langsung dikirim ke proses clustering (*Clustering*) dan salinan lainnya digunakan untuk keperluan pembandingan atau visualisasi. Hasil dari operator *Clustering* kemudian diteruskan ke operator *Performance*, yang menghitung metrik kualitas cluster seperti average within centroid distance dan Davies–Bouldin Index. Rangkaian ini memastikan bahwa proses pra-pengolahan, pembentukan cluster, dan evaluasi dapat dilakukan secara terstruktur dan konsisten.

Hasil Clustering dan Evaluasi Kulitas Model

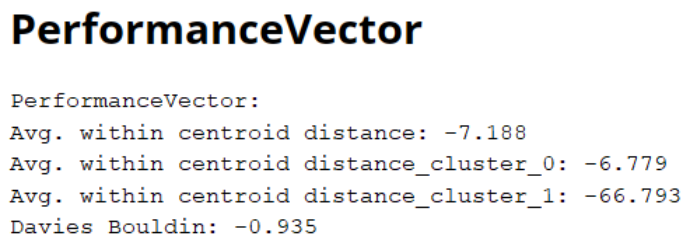
Setelah seluruh tahapan pra-pengolahan selesai, algoritma K-Means dijalankan pada data yang telah dinormalisasi. Hasil akhir pembentukan cluster di RapidMiner menghasilkan dua kelompok, yaitu Cluster 0 dan Cluster 1. Informasi jumlah anggota tiap cluster dapat dilihat pada Gambar 3.

Cluster Model

```
Cluster 0: 437 items
Cluster 1: 3 items
Total number of items: 440
```

Gambar 3. Output Cluster Model dari RapidMiner

Gambar 2 menunjukkan bahwa Cluster 0 berisi 437 pelanggan, sedangkan Cluster 1 hanya berisi 3 pelanggan, dengan total keseluruhan 440 pelanggan. Distribusi ini menegaskan bahwa Cluster 0 adalah cluster mayoritas yang berisi hampir seluruh pelanggan dengan pola pembelian relatif normal, sedangkan Cluster 1 adalah cluster minoritas yang mengelompokkan pelanggan-pelanggan dengan karakter pembelian sangat berbeda dan cenderung ekstrem. Untuk menilai kualitas cluster yang terbentuk, RapidMiner menghasilkan keluaran PerformanceVector, yang dirangkum pada Gambar 4.



Gambar 4. Output PerformanceVector K-Means pada RapidMiner

Gambar 4 menampilkan nilai average within centroid distance secara keseluruhan sebesar -7,188, dengan rincian sekitar -6,779 untuk Cluster 0 dan -66,793 untuk Cluster 1. Semakin kecil (dalam arti mendekati nol jika mengabaikan tanda negatif akibat implementasi perangkat lunak) nilai rata-rata jarak ke centroid, semakin rapat data di dalam cluster. Nilai untuk Cluster 0 yang relatif lebih kecil dibanding Cluster 1 menunjukkan bahwa pelanggan di Cluster 0 memiliki jarak ke centroid yang lebih homogen. Sebaliknya, nilai yang sangat besar secara absolut pada Cluster 1 menggambarkan bahwa ketiga pelanggan di cluster tersebut memiliki variasi yang cukup tinggi dan benar-benar jauh dari centroid, sehingga cluster ini memang berisi pelanggan yang sangat ekstrem(Rahma et al., 2025).

Selain itu, pada Gambar 4 juga ditampilkan nilai Davies–Bouldin Index (DBI) sebesar -0,935. Meskipun dalam teori DBI lazimnya bernilai positif, implementasi tertentu seperti di RapidMiner dapat menghasilkan nilai negatif. Namun, prinsip interpretasinya tetap sama, yaitu semakin mendekati nol, pemisahan cluster cenderung semakin baik. Nilai -0,935 pada penelitian ini menunjukkan bahwa dua cluster yang terbentuk masih memiliki kualitas pemisahan yang cukup layak untuk digunakan dalam analisis segmentasi pelanggan.

Analisi Centroid dan Karakteristik Cluster

Untuk memahami karakteristik masing-masing cluster, nilai centroid tiap atribut dianalisis dalam skala Z-Score dan disajikan dalam tabel centroid (tidak ditampilkan kembali di sini). Hasilnya menunjukkan bahwa Cluster 0 memiliki nilai centroid yang sangat dekat dengan nol pada hampir semua atribut. Hal ini berarti bahwa pelanggan di cluster ini memiliki pola pembelian yang berada di sekitar rata-rata dataset, baik untuk produk *Fresh, Milk, Grocery, Frozen, Detergents_Paper,* maupun *Delicassen*. Tidak ada atribut yang menonjol jauh di atas maupun jauh di bawah rata-rata, sehingga Cluster 0 dapat diinterpretasikan sebagai segmen pelanggan reguler dengan volume pembelian kecil hingga menengah yang relatif stabil.

Sebaliknya, Cluster 1 menampilkan pola centroid yang sangat kontras. Nilai centroid pada atribut *Frozen* dan *Delicassen* sangat tinggi, yaitu masing-masing sekitar 8,459 dan 6,108 dalam skala Z-Score, sementara pada atribut *Fresh* dan *Milk* juga berada cukup jauh di atas nol. Pola ini mengindikasikan bahwa tiga pelanggan yang tergabung dalam Cluster 1 melakukan pembelian dalam jumlah jauh di atas rata-rata untuk produk-produk tersebut, terutama untuk produk beku dan delicatessen. Menariknya, centroid pada atribut *Detergents_Paper* justru berada di bawah rata-rata, yang mengisyaratkan bahwa pelanggan dalam cluster ini tidak terlalu fokus pada pembelian produk pembersih kertas.

Dengan demikian, Cluster 1 dapat diartikan sebagai segmen pelanggan ekstrem atau high spender yang sangat bergantung pada pasokan produk tertentu dalam jumlah besar. Walaupun hanya terdiri dari tiga pelanggan, kontribusi mereka terhadap total nilai transaksi sangat mungkin signifikan, sehingga secara bisnis mereka layak mendapatkan perlakuan khusus, seperti penawaran harga kontrak, jadwal pengiriman fleksibel, atau prioritas layanan(Awalina & Rahayu, 2025).

Perbandingan dengan Kondisi Sebelum Penanganan Outlier

Pada percobaan awal sebelum dilakukan penanganan outlier dan normalisasi, struktur cluster yang terbentuk cenderung tidak stabil. Posisi centroid mudah bergeser, variasi jarak di dalam cluster cukup besar, dan metrik evaluasi tidak menunjukkan hasil yang memuaskan. Hal ini sejalan dengan sifat K-Means yang sensitif terhadap skala atribut dan nilai-nilai ekstrem.

Setelah outlier ditangani dengan winsorizing dan data dinormalisasi menggunakan Z-Score, struktur cluster yang dihasilkan menjadi lebih teratur dan mudah diinterpretasikan. Mayoritas pelanggan

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

berkumpul dalam satu cluster yang relatif kompak, sementara cluster kedua yang berisi pelanggan ekstrem tetap muncul tetapi kini lebih jelas perannya sebagai segmen khusus. Nilai average within centroid distance dan Davies–Bouldin Index yang dihasilkan memperlihatkan adanya peningkatan kualitas pemisahan cluster. Kondisi ini menunjukkan bahwa tahap pra-pengolahan memiliki peran penting dalam memastikan bahwa K-Means memberikan hasil segmentasi yang lebih stabil dan masuk akal.

Implikasi Segmentasi dan Pengaruh Outlier serta Ketidakseimbangan Data

Secara keseluruhan, hasil clustering menunjukkan bahwa algoritma K-Means mampu mengelompokkan pelanggan grosir ke dalam dua segmen utama yang cukup jelas. Segmen pertama adalah pelanggan reguler yang tergabung dalam Cluster 0 dan mewakili mayoritas pelanggan. Segmen ini cocok menjadi sasaran berbagai program pemasaran massal, promosi rutin, dan program loyalitas umum. Segmen kedua adalah pelanggan ekstrem atau high spender yang tergabung dalam Cluster 1 dan meskipun jumlahnya sangat sedikit, mereka berpotensi memberikan kontribusi penjualan yang besar sehingga perlu strategi pelayanan dan penawaran yang lebih personal.

Secara metodologis, penelitian ini juga menegaskan bahwa outlier dan ketidakseimbangan data memberikan pengaruh besar terhadap hasil K-Means. Tanpa penanganan yang memadai, centroid cenderung tertarik ke arah nilai-nilai ekstrem dan menyebabkan cluster yang terbentuk sulit diinterpretasikan. Tahap pra-pengolahan yang mencakup deteksi dan penanganan outlier serta normalisasi Z-Score terbukti membantu menghasilkan struktur cluster yang lebih konsisten dan selaras dengan realitas bisnis. Dengan demikian, penelitian ini tidak hanya menghasilkan segmentasi pelanggan yang dapat dimanfaatkan perusahaan, tetapi juga memberikan gambaran yang lebih konkret mengenai pentingnya kualitas data dalam penerapan algoritma clustering pada konteks pelanggan grosir.

SIMPULAN

Kesimpulan dari penelitian ini adalah bahwa K-Means cukup efektif digunakan untuk mengelompokkan pelanggan grosir pada Wholesale Customers Dataset, terutama setelah data terlebih dahulu dibersihkan dari outlier dan dinormalisasi dengan Z-Score. Dengan jumlah cluster k = 2, diperoleh dua segmen utama, yaitu kelompok pelanggan reguler yang jumlahnya sangat dominan (Cluster 0) dan kelompok kecil pelanggan dengan pembelian sangat tinggi khususnya pada produk Frozen dan Delicassen (Cluster 1) yang dapat dikategorikan sebagai high spender. Hasil ini menunjukkan bahwa kualitas pra-pengolahan data sangat berpengaruh terhadap bentuk dan interpretasi cluster, sekaligus menegaskan bahwa outlier dan ketidakseimbangan data tidak boleh diabaikan dalam proses segmentasi. Secara praktis, perusahaan dapat memanfaatkan segmentasi ini untuk membedakan strategi, misalnya program promosi massal untuk pelanggan reguler dan layanan atau penawaran khusus untuk pelanggan high spender. Ke depan, penelitian serupa dapat dikembangkan dengan membandingkan K-Means dengan algoritma clustering lain yang lebih robust terhadap outlier, atau dengan menambahkan variabel perilaku lain agar segmentasi pelanggan menjadi lebih kaya dan semakin mendukung pengambilan keputusan bisnis.

UCAPAN TERIMA KASIH

Peneliti menyampaikan ucapan terima kasih kepada semua pihak yang telah membantu dalam pelaksanaan penelitian dan penyusunan artikel ini. Terima kasih khusus disampaikan kepada dosen pembimbing yang telah memberikan arahan, masukan, dan koreksi sejak tahap perumusan masalah hingga penyusunan laporan akhir. Peneliti juga berterima kasih kepada program studi dan institusi yang telah menyediakan sarana, prasarana, serta dukungan akademik selama proses penelitian berlangsung. Tidak lupa, peneliti mengucapkan terima kasih kepada keluarga dan rekan-rekan yang senantiasa memberikan dukungan moral, motivasi, serta bantuan teknis sehingga penelitian ini dapat diselesaikan dengan baik.

REFERENSI

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-47578-3

Awalina, L., & Rahayu, T. (2025). Segmentasi Customer pada Industri Ritel Menggunakan Teknik Clustering K-Means. *Digital Transformation Technology (Digitech)*, *5*(2), xx–xx.

Iqbal, I., Hidayat, N., Gevano, D. P., & Ilahi, A. P. R. (2025). Segmentasi Pelanggan Menggunakan K-Means Clustering Berdasarkan Data Kepribadian dan Pola Konsumsi. *Jurnal Teknik Informatika (JUTIF)*, *6*(5), 3914–3924.

Julian, N. D., Belakang, N., & Others. (2023). Segmentasi Pelanggan Menggunakan Algoritma K-Means pada Jaringan Telekomunikasi untuk Optimalisasi Strategi Pemasaran. *JTT (Jurnal Teknologi Terpadu)*, *xx*(xx), xx–xx.

Maskanah, I., Primajaya, A., & Rizal, A. (2020). Segmentasi Pelanggan Toko Purnama dengan Algoritma K-Means dan Model RFM untuk Perancangan Strategi Pemasaran. *INOVTEK Polbeng - Seri Informatika*, *5*(2), 218–225. https://doi.org/10.35314/isi.v5i2.1443

Nugraha, R. D., Adelia, D. D., & Rivaldi, D. (2025). Segmentasi Pelanggan Retail Berbasis Perilaku Menggunakan Algoritma K-Means Clustering. *Digital Transformation Technology (Digitech)*, *5*(2), 141–149. https://doi.org/10.47709/digitech.v5i2.6340

Oktavian, V. V. D., Ridho, & Daffa. (2025). Segmentasi Pelanggan Berbasis RFM dengan Algoritma K-Means pada Data Transaksi Online Retail. *Jurnal Riset Informatika Dan Teknologi Informasi (JRITI)*, *2*(3), xx–xx.

Pangestu, P. I., Hermanto, T. I., & Irmayanti, D. (2023). Analisis Segmentasi Pelanggan Berbasis Recency Frequency Monetary (RFM) Menggunakan Algoritma K-Means. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(3), 1486–1492. https://doi.org/10.36040/jati.v7i3.7171

Penulis, N. (2025). Implementasi Algoritma K-Means Clustering untuk Segmentasi Pelanggan Berdasarkan Data Transaksi dan Preferensi Pembelian. *Sistemasi: Jurnal Sistem Informasi*, *14*(6), 2751–2767.

Pramudiansyah, A., & Munte, H. (2021). Segmentasi Pelanggan Menggunakan Algoritma K-Means Berdasarkan Model Recency Frequency Monetary. *Jurnal Nasional (Online)*, *7*(2), xx–xx.

Rahma, A. A., Faqih, A., & Rinaldi, A. R. (2025). Optimalisasi Strategi Pemasaran melalui Segmentasi Pelanggan dengan Analisis RFM dan Algoritma K-Means untuk Bisnis Ritel. *JIKO (Jurnal Informatika Dan Komputer)*, *9*(2), xx–xx. https://doi.org/10.26798/jiko.v9i2.1737

Ramadhan, A. G. (2023). Data Mining Untuk Segmentasi Pelanggan dengan Algoritma K-Means: Studi Kasus pada Data Pelanggan di Toko Retail. *Syntax Literate: Jurnal Ilmiah Indonesia*, *8*(10), 5701–5718.

Ramadhona, W., Nugroho, B. I., & Murtopo, A. A. (2022). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means. *Jurnal Minfo Polgan*, *11*(2), 100–104. https://doi.org/10.33395/jmp.v11i2.11797

Silamantha, W. A., & Hadiono, K. (2024). Analisis RFM dan K-Means Clustering untuk Segmentasi Pelanggan pada PT Sanutama Bumi Arto. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer Dan Manajemen)*, *5*(3), 1297–1305. https://doi.org/10.30645/kesatria.v5i3.448

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

Wahyuni, S., Wulansari, T. T., & Fahrullah, F. (2023). Segmentasi Pelanggan Berdasarkan Analisis Recency, Frequency, Monetary Menggunakan Algoritma K-Means pada CV Toedjoe Sinar Group. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, *7*(2), 29–36. https://doi.org/10.30872/jurti.v7i2.8748

ISSN 2963-7856 (print), ISSN 2961-9890 (online)