Uranus : Jurnal Ilmiah Teknik Elektro, Sains dan Informatika
Volume. 3 Nomor. 2 Juni 2025
e-ISSN: 3031-996X; p-ISSN: 3031-9951, Hal. 164-173
DOI: https://doi.org/10.61132/uranus.v3i2.1214
Tersedia: https://journal.arteii.or.id/index.php/Uranus

Penerapan *K-Means Clustering* untuk Segmentasi Pelanggan

Fathoni Dwi Atmoko
Program Studi Teknik Informatika, Universitas Nahdlatul Ulama, Indomesia
**Penulis Korespondensi: fathonidwiatmoko03@gmail.com*

Alamat: Taman Fajar, Kec. Purbolinggo, Kabupaten Lampung Timur, Lampung 34192

Abstract*. Public transportation, with Transjakarta as its main pillar, requires a deep understanding of customer behavior to improve service quality and maintain loyalty. This study aims to segment Transjakarta customers using data mining techniques, specifically the K-Means Clustering algorithm, based on the RFM (Recency, Frequency, Monetary/Value) behavioral model. 37,900 rows of raw transaction data were processed into a clean database, resulting in 1,917 unique customers for analysis. The RFM metrics were then normalized using Min-Max Scaler. The optimal number of clusters was evaluated using the Elbow Curve and Silhouette Score Methods, which led to the determination of k = 4 clusters. The segmentation results identified four customer groups requiring specific strategies: Cluster 3 (Champions) with high R, F, and V (requiring rewards and retention); Cluster 0 (Active, Low Value) with high R and F but low V (requiring upsells and cross-sells); Cluster 1 (Potential/At-Risk); and Cluster 2 (Dormant/Lost). Preliminary analysis (EDA) showed that nearly half of customers (49.3%) used Bank DKI cards, dominated by the productive age group (25–45 years old), with the Rusun Kapuk Muara–Penjaringan route being the busiest. The main managerial recommendation is to strengthen the partnership with Bank DKI and optimize services in this busy corridor.*

Keywords: *Clustering; RFM; Data Analysis; K-Means; Customer Segmentation.*

Abstrak. Transportasi publik, dengan Transjakarta sebagai pilar utamanya, memerlukan pemahaman mendalam tentang perilaku pelanggan untuk meningkatkan kualitas layanan dan mempertahankan loyalitas. Penelitian ini bertujuan melakukan segmentasi pelanggan Transjakarta menggunakan teknik data mining, khususnya algoritma *K-Means Clustering*, berdasarkan model perilaku RFM (*Recency*, *Frequency*, *Monetary/Value*). Data transaksi mentah yang berjumlah 37.900 baris diolah menjadi basis data yang bersih, menghasilkan 1.917 pelanggan unik untuk dianalisis. Metrik RFM kemudian dinormalisasi menggunakan *Min-Max Scaler*.Penentuan jumlah klaster optimal dievaluasi menggunakan Metode *Elbow Curve* dan *Silhouette Score*, yang mengarahkan pada penetapan *k=4* klaster. Hasil segmentasi mengidentifikasi empat kelompok pelanggan yang memerlukan strategi spesifik: Klaster 3 (Champions) dengan R, F, dan V Tinggi (memerlukan Reward & Retention); Klaster 0 (Aktif, Nilai Rendah) dengan R dan F Tinggi tetapi V Rendah (memerlukan *Upsell & Cross-sell*); Klaster 1 (Potensial/Berisiko); dan Klaster 2 (*Dormant/Lost*). Analisis awal (EDA) menunjukkan bahwa hampir separuh pelanggan (49,3%) menggunakan kartu Bank DKI, didominasi oleh kelompok usia produktif (25–45 tahun), dengan rute Rusun Kapuk Muara–Penjaringan sebagai yang tersibuk. Rekomendasi manajerial utamanya adalah memperkuat kemitraan dengan Bank DKI dan mengoptimalkan layanan di koridor sibuk tersebut.

Kata kunci: Analisis Data ; *Clustering*; RFM; *K-Means* ;Segmentasi Pelanggan.

1. LATAR BELAKANG

Transportasi publik merupakan urat nadi bagi kota metropolitan seperti Jakarta. Transjakarta, sebagai salah satu pilar utama, melayani jutaan perjalanan setiap bulannya. Di tengah meningkatnya pilihan mobilitas, memahami perilaku dan kebutuhan pelanggan menjadi krusial untuk mempertahankan loyalitas dan meningkatkan kualitas layanan. Pendekatan "satu ukuran untuk semua" tidak lagi memadai di era yang didorong oleh data.

Naskah Masuk: 02 Juni 2025; Revisi: 15 Juni 2025; Diterima: 28 Juni 2025; Terbit: 30 Juni 2025;

Perusahaan di berbagai industri telah beralih ke *Customer Relationship Management* (CRM) yang canggih, di mana segmentasi pelanggan adalah fondasi utamanya. Dengan mengelompokkan pelanggan ke dalam segmen-segmen yang homogen, perusahaan dapat mengidentifikasi pelanggan paling berharga, pelanggan yang berisiko beralih (churn), dan pelanggan baru yang potensial.

Metode K-Means bermanfaat dalam berbagai situasi, menurut studi sebelumnya. Pendekatan ini efektif mengidentifikasi kategori klien berdasarkan tren penggunaan kartu kredit, sebagaimana dibuktikan oleh (Alhamdani et al., 2021). Studi lain (Anam et al., 2024) menggunakan K-Means untuk mengkaji tren penggunaan gawai pada anak usia dini, sementara (Rohman & Wibowo, 2024) menemukan bahwa K-Means mengungguli pendekatan KMedoids dalam segmentasi konsumen mal. Temuan ini menunjukkan bahwa K-Means bekerja dengan baik dengan data yang besar dan beragam, tetapi penelitian lebih lanjut diperlukan ketika menangani data konsumen yang lebih rumit.

Penelitian ini mengusulkan penerapan teknik data mining, khususnya *K-Means Clustering*, untuk melakukan segmentasi pelanggan Transjakarta. Model segmentasi didasarkan pada metrik perilaku yang telah teruji, yaitu RFM (*Recency, Frequency, Monetary*). *Recency* mengukur kapan terakhir kali pelanggan menggunakan layanan, *Frequency* mengukur seberapa sering mereka menggunakannya, dan *Monetary* (dalam konteks ini, *payAmount*) mengukur total pengeluaran mereka.

Tujuan dari penelitian ini adalah untuk (1) Mengolah data transaksi mentah Transjakarta menjadi metrik RFM yang bermakna; (2) Menerapkan algoritma *K-Means* untuk mengidentifikasi cluster pelanggan yang berbeda; dan (3) Menganalisis karakteristik setiap cluster untuk memberikan rekomendasi manajerial yang dapat ditindaklanjuti

2. KAJIAN TEORITIS

Segmentasi Pelanggan

Segmentasi adalah proses membagi pelanggan menjadi beberapa klaster dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran (Harani et al., 2020). Segmentasi dapat didasarkan pada demografi (usia, jenis kelamin), geografi, psikografi, atau perilaku. Segmentasi perilaku, yang digunakan dalam penelitian ini, dianggap paling kuat karena didasarkan pada interaksi aktual pelanggan dengan layanan. Segmentasi pelanggan adalah teknik untuk mengidentifikasi berbagai jenis pelanggan agar dapat lebih memahami mereka dan membuat keputusan yang lebih menguntungkan. Hasil segmentasi pelanggan ini dapat digunakan sebagai panduan untuk menyusun strategi pemasaran, melakukan penjualan

silang produk baru untuk setiap kelompok, dan menciptakan produk untuk kelompok pelanggan yang paling berharga (Khajvand & Tarokh, 2011).. Dalam konsep pemasaran, segmentasi sangat penting dalam pemasaran relasional karena membuat hubungan pelanggan lebih menarik dan mengarah pada pemahaman yang lebih baik tentang kebutuhan pelanggan (Adiana, 2018).

Analisis RFM

RFM, singkatan dari *Recency, Frequency, Monetary*, adalah model yang banyak digunakan untuk mengidentifikasi perilaku pelanggan dan merepresentasikan karakteristik perilaku pelanggan (Zakariyya, 2020).. Menurut (Khobzi et al., 2014), RFM adalah model segmentasi perilaku yang populer.

a) *Recency* (R): Mengukur kebaruan. Pelanggan yang baru saja bertransaksi lebih mungkin untuk merespons promosi daripada pelanggan yang sudah lama tidak aktif.
b) *Frequency* (F): Mengukur frekuensi. Pelanggan yang sering bertransaksi lebih loyal dan berharga.
c) *Monetary* (M/V): Mengukur nilai moneter. Pelanggan yang menghabiskan lebih banyak uang memiliki nilai lebih tinggi bagi perusahaan.

Dalam konteks Transjakarta, model ini diadaptasi: '*Recency*' adalah waktu sejak perjalanan terakhir, '*Frequency*' adalah jumlah total perjalanan, dan '*Monetary*' (disebut 'Value' dalam analisis) adalah total *payAmount* yang dibayarkan.

K-Means Clustering

Salah satu teknik pembelajaran tanpa pengawasan yang paling sering digunakan untuk pengelompokan adalah K-Means. Program ini membagi kumpulan data menjadi k klaster yang telah ditentukan sebelumnya secara iteratif. Tujuannya adalah untuk meminimalkan varians dalam setiap cluster (meminimalkan *Sum of Squared Distances* / SSD antara titik data dan centroid cluster mereka). Dengan mengelompokkan klien berdasarkan ciri dan perilaku, algoritma K-Means, sebuah teknik penambangan data yang memungkinkan penyedia layanan untuk menawarkan solusi yang lebih sesuai dan relevan. (Harani et al., 2020).

Penentuan Jumlah Cluster Optimal

Memilih nilai 'k' yang tepat sangat penting. Dua metode umum digunakan:

1. **Metode *Elbow*:** Metode ini memplot nilai SSD (atau *inertia*) terhadap jumlah cluster (k). 'Siku' (*elbow*) pada grafik menunjukkan titik di mana penambahan cluster baru tidak lagi memberikan penurunan SSD yang signifikan. Menurut merliana dan santoso

(Merliana & Santoso, 2015), Dengan mengukur persentase klaster yang membentuk siku di lokasi tertentu, metode siku menghasilkan informasi untuk mengidentifikasi jumlah klaster yang optimal.

2. ***Silhouette Score*:** Metode ini menilai kekohesifan suatu objek (kemiripan dengan klasternya sendiri) dibandingkan dengan klaster lain (keterpisahan). Pengelompokan yang lebih baik dan lebih padat ditunjukkan dengan skor yang lebih tinggi, yang berkisar antara -1 hingga 1. Salah satu teknik untuk mengevaluasi kualitas klaster yang dihasilkan oleh proses pengelompokan adalah siluet.. (Aditya et al., 2020).

3. METODE PENELITIAN

Sumber Data

Data diperoleh dari transaksi pelanggan Transjakarta yang berisi 37.900 baris dan 22 kolom, yang kemudian pembersihan data**,** terdapat nilai yang hilang (null) di beberapa kolom krusial seperti *corridorName* dan *tapOutStops*. Metode *dropna()* diterapkan untuk menghapus semua baris yang mengandung nilai null, menghasilkan dataset yang bersih (*df_droped*) dengan 31.730 catatan transaksi.

Tahap Pra-Pemrosesan

1. Menghapus nilai hilang (*missing values*).
2. Konversi tipe data kolom waktu *tap in* dan *tap out*.
3. Penambahan variabel:
    a) usia pelanggan
    b) jam *tap-in* dan *tap-out*
    c) durasi perjalanan
4. Basis data pelanggan unik (*customer*) dibuat dengan menghapus duplikat berdasarkan payCardName, menghasilkan 1.917 pelanggan unik untuk dianalisis.

Perhitungan RFM

Untuk setiap 1.917 pelanggan unik, variabel RFM dihitung:

1. *Recency*, dihitung sebagai selisih hari antara tanggal transaksi terakhir di seluruh dataset (*max_date*) dan tanggal *tapOutTime* pelanggan.
2. *Frequency*, dihitung menggunakan *value_counts()* pada kolom *payCardName* untuk mendapatkan jumlah total transaksi per pelanggan.
3. *Value*, dihitung dengan menjumlahkan (*sum()*) seluruh *payAmount* untuk setiap *payCardName*.

Ketiga metrik ini kemudian digabungkan ke dalam satu DataFrame (clv atau data).

Normalisasi

Data dinormalisasi menggunakan *Min-Max Scaler*, karena variabel R, F, dan V memiliki rentang nilai yang berbeda, penskalaan diperlukan agar *K-Means* tidak bias terhadap satu variabel. *MinMaxScaler* dari *Scikit-learn* digunakan untuk mengubah skala ketiga fitur ini ke rentang 1 dan 0

Penentuan Jumlah Klaster

Penentuan nilai *k* (jumlah klaster) yang optimal adalah langkah krusial dalam algoritma *K-Means*, karena nilai yang tidak tepat dapat menghasilkan klaster yang tidak representatif atau terlalu spesifik. Dalam penelitian ini, penentuan *k* dilakukan dengan mengevaluasi performa model *K-Means* untuk berbagai nilai *k* (dari 2 hingga 8) menggunakan dua metode utama: Metode *Elbow* dan *Silhouette Score*.

Metode Elbow

Metode /Teknik *Elbow* digunakan untuk memvisualisasikan total variasi dalam klaster (internal variability). Total variasi ini diukur dengan *inertia* atau *Sum of Squared Distances* (SSD), jumlah kuadrat jarak antara setiap titik data ke centroid klaster terdekatnya.

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

Di mana *k* adalah jumlah klaster, C_i adalah klaster *i*, *x* adalah titik data, dan μ_i adalah centroid dari klaster *i*.

Nilai *k* yang optimal diidentifikasi pada titik di mana penambahan klaster baru tidak lagi memberikan penurunan SSD yang signifikan. Titik ini akan terlihat seperti "siku" (*elbow*) pada grafik SSD vs. *k*.

Metode Silhouette Score

Untuk memvalidasi dan menyempurnakan hasil dari Metode *Elbow*, digunakan *Silhouette Score*. Skor ini mengukur seberapa mirip sebuah objek dengan klaster tempatnya berada (kohesi) dibandingkan dengan klaster terdekat lainnya (separasi).

Skor Siluet (*S(i)*) untuk sebuah titik data *i* dihitung sebagai:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Di mana:

- *a(i):* Jarak rata-rata antara *i* dan semua titik data lainnya di klaster yang sama.
- *b(i) :* Jarak rata-rata antara *i* dan semua titik data di klaster terdekat berikutnya (klaster tetangga).

Rata-rata dari semua *S(i)* menghasilkan *Silhouette Average Score* untuk seluruh model. Skor berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa titik data terkelompokkan dengan baik dan terpisah jauh dari klaster lain.

4. HASIL DAN PEMBAHASAN

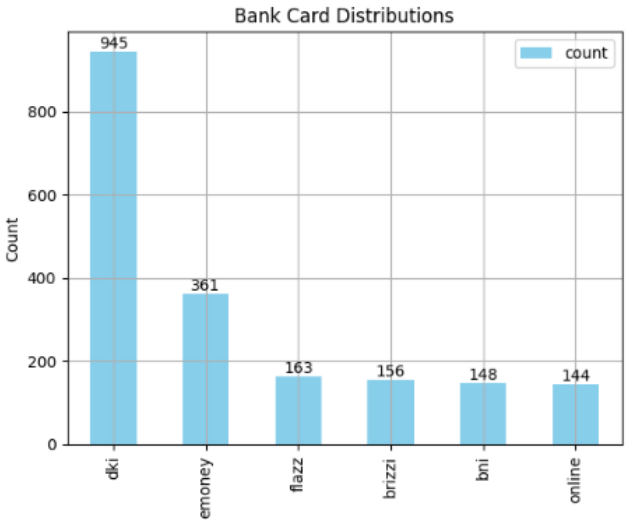
Analisis Eksploratif Data (EDA)

Analisis deskriptif pada 1.917 pelanggan unik dilakukan untuk mendapatkan pemahaman awal mengenai karakteristik demografi dan transaksi pengguna layanan Transjakarta sebelum proses segmentasi.

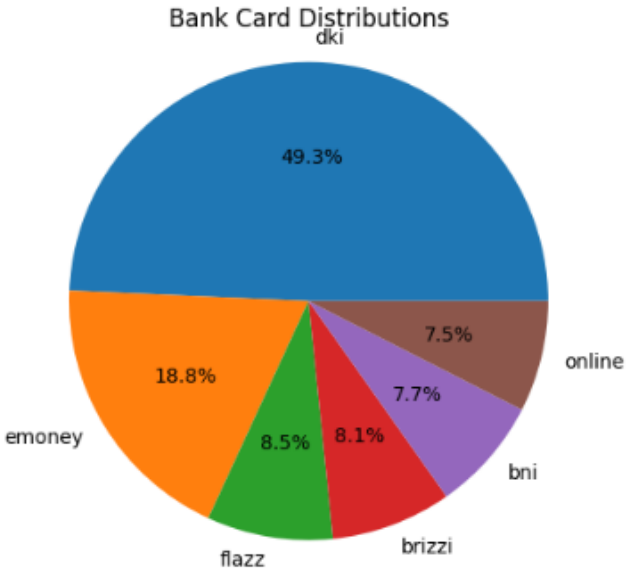
a. Distribusi Bank Penerbit Kartu

Analisis terhadap sumber penerbitan kartu pembayaran (`payCardBank`) mengungkapkan adanya dominasi yang jelas dari bank tertentu dalam ekosistem pembayaran Transjakarta. Dari total 1.917 pelanggan unik, didapatkan bahwa kartu dari Bank DKI ('dki') adalah yang paling dominan, mencakup 945 pelanggan atau 49,3% dari keseluruhan basis pelanggan.

Dominasi ini diikuti oleh penggunaan kartu 'emoney' dengan 361 pelanggan (18,8%). Sementara itu, kartu-kartu lainnya seperti 'flazz', 'brizzi', dan 'tapcash' memiliki pangsa pasar yang jauh lebih kecil. Temuan ini sangat krusial, menunjukkan bahwa hampir separuh pelanggan Transjakarta memiliki afiliasi dengan Bank DKI. Hal ini dapat dijadikan dasar bagi manajemen Transjakarta untuk memperkuat kemitraan strategis dengan Bank DKI, misalnya melalui program loyalitas bersama atau layanan eksklusif, untuk meningkatkan retensi dan *customer value* dari segmen pelanggan terbesar ini. Hal ini ditunjukkan pada visualisasi bar chart dan pie chart *Bank Card Distribution* dibawah ini.



Gambar 1. Grafik Distribusi Kartu Bank



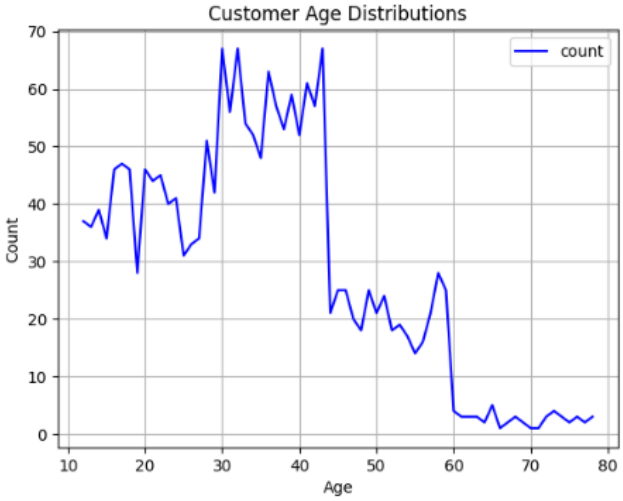
Gambar 2. Pie Chart Distribusi Kartu Bank

b. Distribusi Usia Pelanggan

Analisis distribusi usia pelanggan (age) memberikan wawasan demografis yang penting mengenai target pasar utama Transjakarta. Berdasarkan visualisasi histogram "*Customer* Age Distributions", ditemukan bahwa mayoritas pengguna layanan berada dalam rentang usia produktif.

Pola distribusi menunjukkan konsentrasi tertinggi pada usia antara 25 hingga 45 tahun, dengan puncaknya berada di sekitar usia akhir 20-an hingga awal 30-an. Hal ini secara kuat mengindikasikan bahwa Transjakarta didominasi oleh segmen pekerja profesional dan komuter muda yang menggunakan layanan untuk perjalanan rutin (kerja atau pendidikan).

Sebaliknya, terdapat penurunan tajam jumlah pelanggan pada kelompok usia di atas 60 tahun. Distribusi ini mengimplikasikan bahwa strategi pemasaran dan pengembangan layanan perlu difokuskan untuk memenuhi kebutuhan mobilitas harian kelompok usia produktif. Selain itu, temuan ini juga dapat menjadi dasar untuk merancang layanan khusus atau promosi yang ditargetkan pada segmen usia muda atau, sebaliknya, meningkatkan aksesibilitas dan kemudahan bagi pelanggan lanjut usia untuk memperluas basis pengguna. Seperti yang ditampilkan pada grafik dibawah ini.

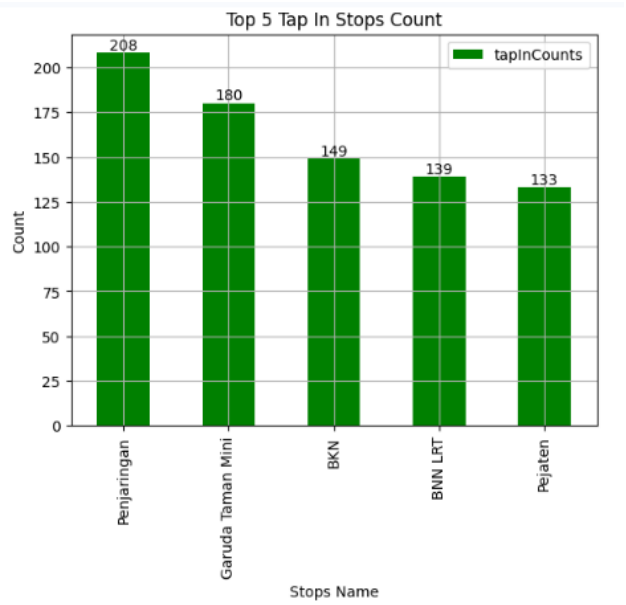


Gambar 3. Grafik Distribusi Usia

c. Pola perjalanan Halte *Tap-In* Tersibuk

Analisis pola perjalanan dimulai dengan mengidentifikasi lima halte pemberangkatan (*tap-in*) dengan frekuensi transaksi tertinggi. Hasil yang divisualisasikan pada Grafik gambar 6 menunjukkan adanya konsentrasi aktivitas pengguna pada halte-halte tertentu, yang mengindikasikan lokasi padat penduduk atau sentra mobilitas.

Halte Penjaringan menduduki peringkat pertama sebagai titik tap-in paling sibuk, mencatat total 208 perjalanan. Dominasi Penjaringan mengimplikasikan pentingnya halte ini sebagai titik entry utama bagi komuter. Halte sibuk berikutnya adalah Garuda Taman Mini dengan 180 perjalanan, diikuti oleh BKN dengan 149 perjalanan, dan BNN LRT dengan 139 perjalanan.



Gambar 4. Grafik Halte *Tap-in* Terbanyak

Tap In terbanyak terjadi pada halte Penjaringan pada jam Pukul 17:00 (5 Sore), halte Halte Garuda Taman Mini menunjukkan puncak aktivitas yang jelas pada Pagi Hari, yaitu sekitar Pukul 06:00 (6 Pagi), halte BKN, BNN LRT, Cibubur Junction pada jam 17:00 (5 Sore). Seperti yang terlihat pada tabel dibawah ini.

Tabel 1. Waktu Tap-in Terbanyak

	tapInStopsName	tapInHour	tapInCounts
0	Penjaringan	17	46
1	Garuda Taman Mini	6	42
2	BKN	17	36
3	BNN LRT	17	36
4	Cibubur Junction	17	33
...	...	...	...
8226	Yado III	7	1
8229	Yayasan Perguruan Rakyat 2	15	1
8233	ACC Simatupang	17	1
8235	Yos Sudarso Kodamar	11	1
8236	AKR Tower	14	1

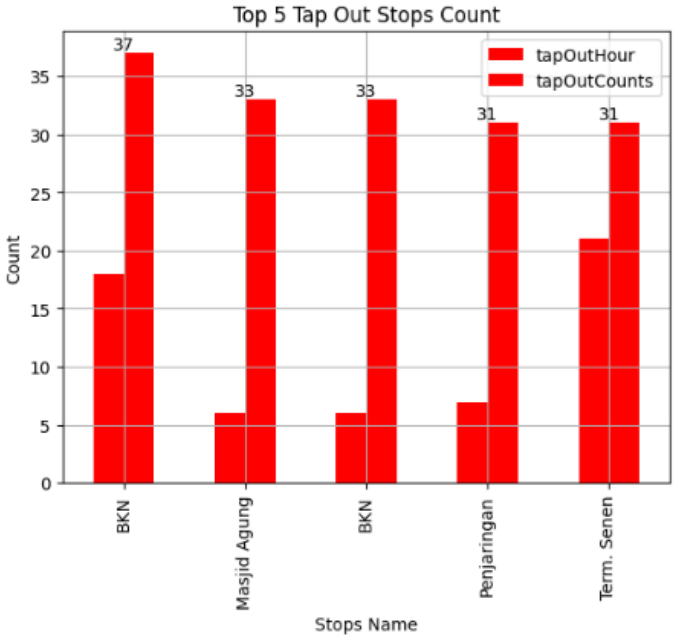
2428 rows × 3 columns

Sedangkan, analisis menunjukkan bahwa halte BKN pada pukul 18.00 menjadi titik Tap Out tersibuk dengan 37 transaksi, mencerminkan puncak aktivitas pulang kerja. Dua titik berikutnya yaitu Masjid Agung pukul 06.00 dan BKN pukul 06.00 mencatat masing-masing

33 transaksi, menegaskan adanya lonjakan aktivitas pada jam keberangkatan pagi. Pada posisi selanjutnya, Penjaringan pukul 07.00 dan Terminal Senen pukul 21.00 sama-sama mencatat 31 transaksi, menunjukkan pola sibuk pagi hari serta tingginya mobilitas pengguna hingga malam hari. Terlihat pada tabel dibawah

Tabel 2. Waktu Tap-out Tersibuk

	stopsName	tapOutHour	tapOutCounts
0	BKN	18	37
1	Masjid Agung	6	33
2	BKN	6	33
3	Penjaringan	7	31
4	Term. Senen	21	31



Gambar 5. Grafik Halte *Tap-out* Terbanyak

d. Data Rute Teratas

Pola yang paling dominan adalah rute bolak-balik (pulang-pergi) antara Rusun Kapuk Muara dan Penjaringan. Rute dari Rusun Kapuk Muara ke Penjaringan merupakan yang tersibuk dengan 108 transaksi, sementara rute baliknya dari Penjaringan ke Rusun Kapuk Muara berada di posisi kedua dengan 103 transaksi, menegaskan koneksi harian yang sangat kuat di koridor ini.

Secara ringkas, rute teratas ini didominasi oleh pergerakan pulang-pergi yang sangat padat antara Rusun Kapuk Muara dan Penjaringan, diikuti oleh pergerakan dari area timur (Garuda Taman Mini dan Cibubur) menuju BKN.

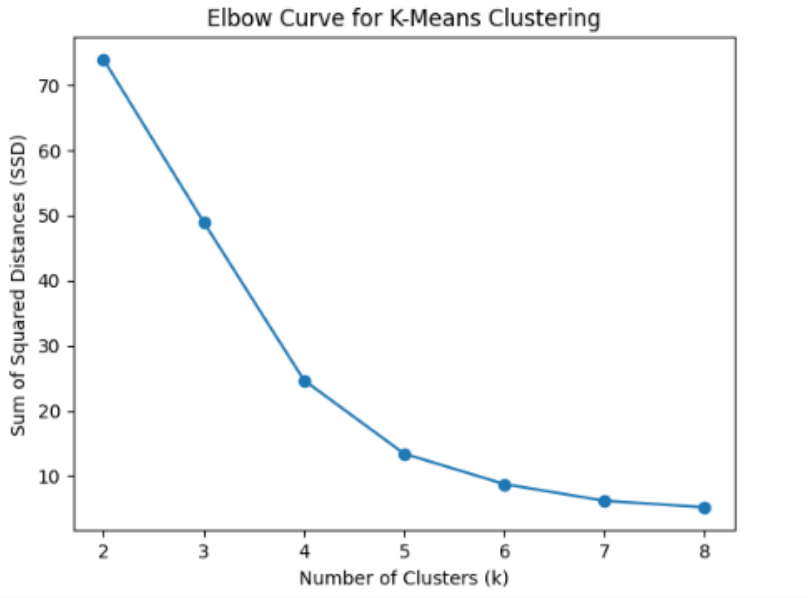
Tabel 3. Daftar Rute Tersibuk

	tapInStopsName	tapInStopsLat	tapInStopsLon	tapOutStopsName	tapOutStopsLat	tapOutStopsLon	TransactionCount
0	Rusun Kapuk Muara	-6.135667	106.76299	Penjaringan	-6.126306	106.79203	108
1	Penjaringan	-6.126306	106.79203	Rusun Kapuk Muara	-6.135667	106.76299	103
2	Garuda Taman Mini	-6.290154	106.88116	BKN	-6.257751	106.87000	92
3	Tanah Merdeka Arah Timur	-6.307866	106.87389	Kampung Rambutan	-6.309885	106.88216	79
4	Cibubur Junction	-6.368735	106.89366	BKN	-6.257751	106.87000	75
5	Simpang Danau Sunter Utara Barat	-6.139853	106.85663	Jembatan Item	-6.130078	106.85492	74
6	Garuda Taman Mini	-6.290154	106.88116	Pinang Ranti	-6.291075	106.88634	71
7	Rusun Penjaringan	-6.130702	106.79487	Penjaringan	-6.126306	106.79203	65
8	Penggilingan	-6.214132	106.93961	Rusun Komarudin	-6.208781	106.94252	62
9	Penjaringan	-6.126306	106.79203	Rusun Penjaringan	-6.130702	106.79487	61

Penentuan Jumlah Cluster (k)

Metode *Elbow* Curve (SSD)

Grafik *Elbow* Curve atau Grafik Kurva Siku menunjukkan hubungan antara jumlah klaster (k) dan Jumlah Jarak Kuadrat (SSD). Tujuannya adalah menemukan "siku" pada grafik, di mana penurunan SSD mulai melambat secara signifikan, yang menunjukkan bahwa penambahan klaster tidak memberikan peningkatan informasi yang substansial.



Gambar 6. Kurva *Elbow*

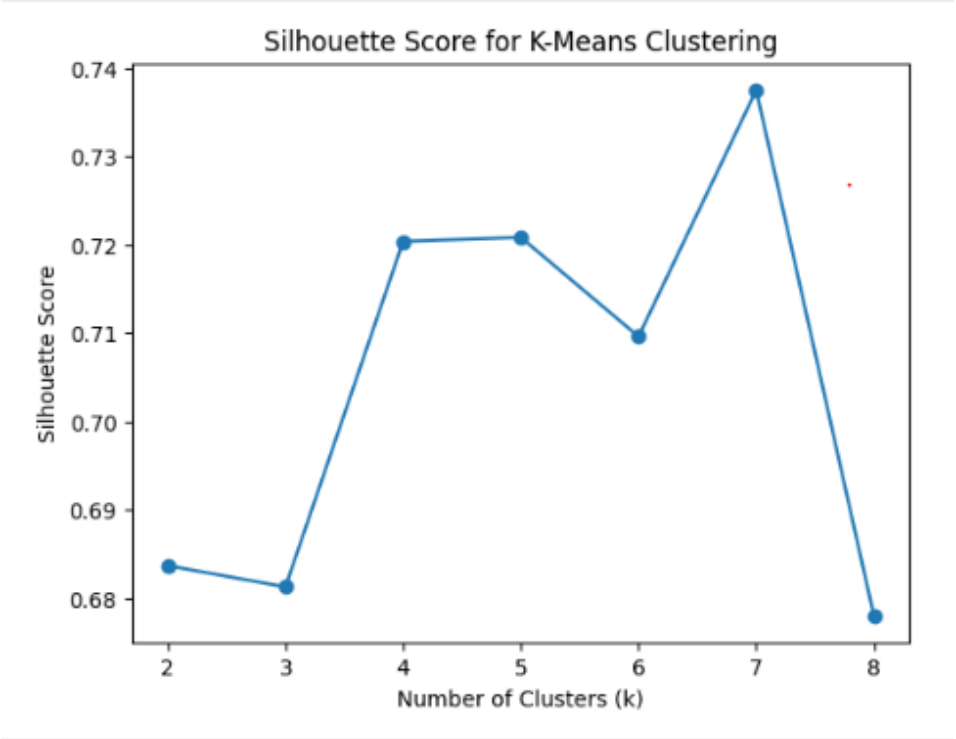
Keterangan :

- Ada penurunan tajam SSD dari *k=2 ke k=3*, dan dari *k=3 ke k=4*.
- Penurunan mulai melambat setelah *k=4* dan *k=5*.

- Titik Siku (*Elbow*): Berdasarkan grafik, titik siku yang paling jelas terjadi pada *k=4* atau *k=5*. Setelah *k=5*, grafik mulai mendatar.

Metode *Silhouette Score*

Grafik *Silhouette Score* atau Grafik Skor Silhouette mengukur seberapa mirip suatu objek dengan klasternya sendiri dibandingkan dengan klaster lain. Skor yang lebih tinggi (mendekati 1) menunjukkan klaster yang padat dan terpisah dengan baik..



Gambar 7. *Silhouette Score*

Keterangan :

- Skor mencapai puncaknya (nilai tertinggi) pada *k=7* dengan nilai skor sekitar 0.738.
- Skor tertinggi kedua dicapai pada *k=4* dan *k=5* (sekitar 0.72).

Penerapan Model *K-Means*

Setelah proses penentuan jumlah klaster optimal (seperti yang dibahas di sub-bab metode penelitian), model *K-Means* final diterapkan.

```
# Perform K-Means Clustering with k=4 (4 clusters)
kmeans = KMeans(n_clusters=4)
data['Cluster'] = kmeans.fit_predict(data[['Recency', 'Frequency', 'Value']])
data
```

	payCardName	Recency	Frequency	Value	Cluster
0	Bajragin Usada	0.931034	0.485714	0.161184	0
1	Gandi Widodo	0.931034	0.485714	0.161184	0
2	Emong Wastuti	0.931034	0.485714	0.161184	0
3	Surya Wacana	0.931034	0.500000	0.165789	0
4	Embuh Mardhiyah	0.931034	0.542857	0.179605	0
...	...	...	...	...	...
1912	Kamila Mahendra	0.310345	0.000000	0.004605	2
1913	Titi Siregar	0.413793	0.000000	0.004605	2
1914	drg. Zahra Nashiruddin	0.896552	0.000000	0.026316	1
1915	Ana Agustina	0.517241	0.000000	0.000000	1
1916	drg. Leo Najmudin	0.620690	0.000000	0.004605	1

1917 rows × 5 columns

Gambar 8. Penerapan Model *K-Means*

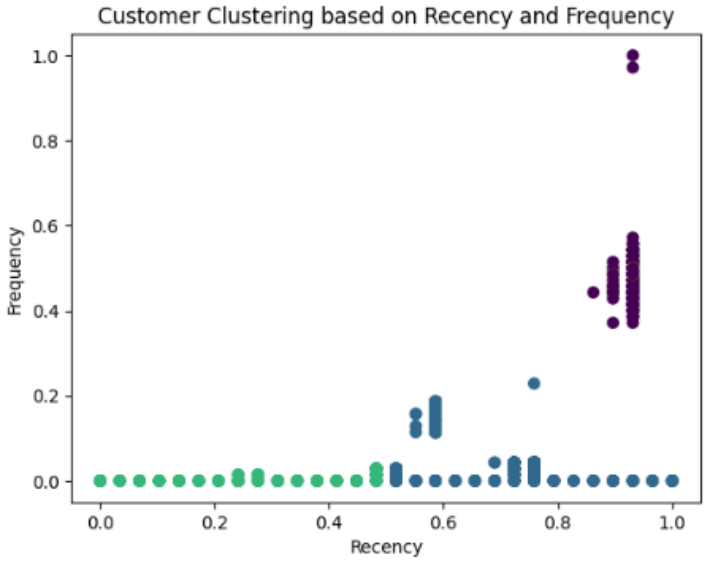
Berdasarkan analisis gambar diatas, jumlah klaster ditetapkan *k=4*. Algoritma *K-Means (n_clusters=4)* kemudian dilatih pada data RFM yang telah dinormalisasi. Hasil dari proses ini adalah penugasan label klaster (0, 1, 2, atau 3) untuk setiap 1.917 pelanggan unik dalam dataset.

Visualisasi Klaster Pelanggan

Untuk memahami bagaimana model memisahkan pelanggan, dilakukan visualisasi data dalam ruang 2D dan 3D.

Visualisasi 2D (*Recency* vs. *Frequency*)

Plot "*Customer Clustering based on Recency and Frequency*" Gambar dibawah ini menunjukkan pemisahan yang jelas antara klaster. Terlihat bagaimana pelanggan dikelompokkan berdasarkan seberapa baru (*Recency*) dan seberapa sering (*Frequency*) mereka menggunakan layanan.



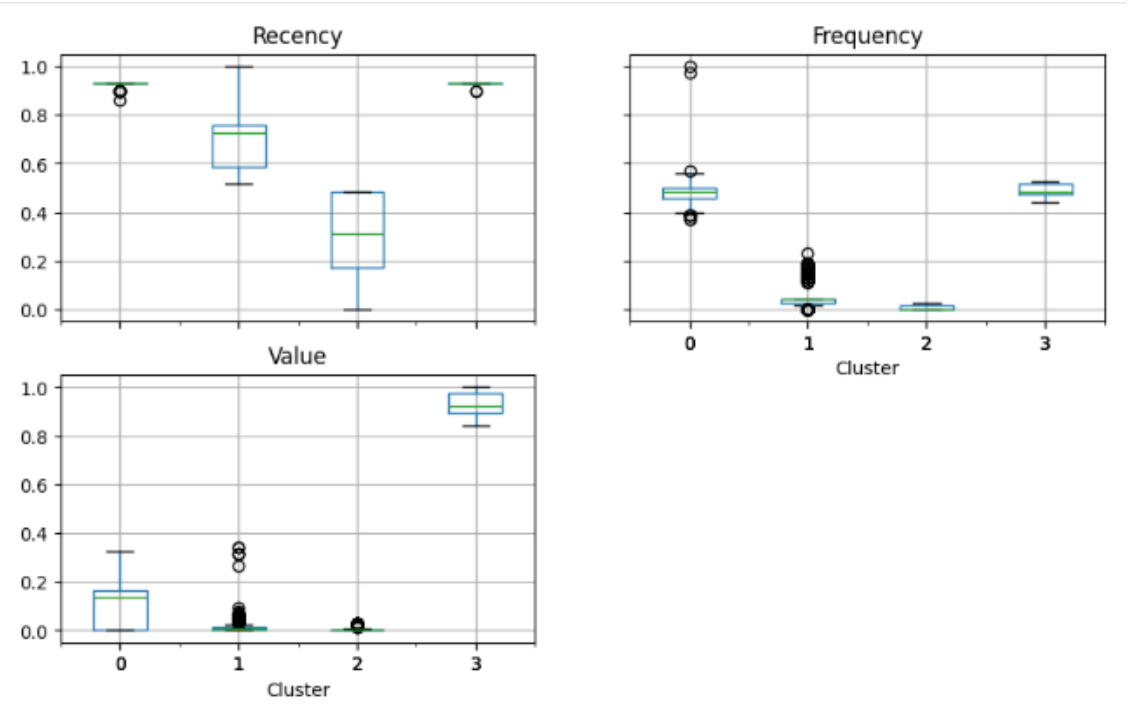
Gambar 9. Visualisasi 2D

Plot "*Customer Clustering based on Recency, Frequency, and Value*" Gambar dibawah ini memberikan gambaran yang lebih komprehensif. Visualisasi ini menegaskan bahwa keempat klaster (ditandai dengan warna berbeda) menempati ruang yang berbeda secara signifikan dalam tiga dimensi perilaku (R, F, dan V), menunjukkan bahwa segmen yang ditemukan memang unik.

Gambar 10. Visualisasi 3D

Analisis Karakteristik Setiap Cluster

Analisis paling krusial dilakukan dengan memeriksa distribusi nilai RFM untuk setiap yang dihasilkan oleh model *K-Means*, seperti yang ditunjukkan oleh visualisasi box plot (kotak garis) dibawah ini.



Gambar 11. Analisis Karakteristik Setiap Cluster

Berdasarkan box plot tersebut, karakteristik dari empat segmen pelanggan dapat diinterpretasikan sebagai berikut:

Tabel ini, yang merupakan output dari perintah data.*groupby('Cluster').mean()*, adalah bagian yang paling penting dalam proses interpretasi, karena mendefinisikan secara kuantitatif profil setiap segmen pelanggan.

Tabel 4. Analisis Karakteristik Setiap Cluster

Cluster	*Recency* (R)	*Frequency* (F)	*Value* (V)	Karakteristik Utama
0	Tinggi (Rata-rata ≈0.93)	Tinggi (Rata-rata ≈0.48)	Rendah/Sedang (Rata-rata ≈0.08)	Pelanggan Aktif Sering (Frequent Buyers/Active but Low Value). Mereka baru saja membeli dan sering membeli, tetapi pembelian mereka bernilai rendah
1	Sedang (Rata-rata ≈0.70)	Rendah (Rata-rata ≈0.06)	Sangat Rendah (Rata-rata ≈0.02)	Pelanggan Potensial atau Berisiko (Potential/At-Risk). Mereka belum terlalu lama membeli, tetapi frekuensi dan nilainya sangat rendah. Mungkin mereka masih baru atau kurang tertarik.
2	Sangat Rendah (Rata-rata ≈0.30)	Sangat Rendah (Rata-rata ≈0.006)	Sangat Rendah (Rata-rata ≈0.006)	Pelanggan Tidur (Dormant/Lost). Sudah lama tidak membeli, frekuensi dan nilai transaksinya juga sangat rendah. Mereka kemungkinan besar telah beralih ke pesaing.
3	Tinggi (Rata-rata ≈0.93)	Tinggi (Rata-rata ≈0.49)	Tinggi (Rata-rata ≈0.93)	Pelanggan Terbaik (Best *Customer*s/Champions). Mereka baru saja membeli, sering membeli, dan menghasilkan nilai tertinggi. Mereka adalah segmen paling berharga.

Rekomendasi

Berikut adalah rekomendasi tindakan yang dapat diambil untuk setiap cluster untuk memaksimalkan retensi, nilai, dan profitabilitas:

Tabel 5. Rekomendasi

Cluster	Karakteristik Cluster	Strategi Manajerial Utama	Tindakan yang Dapat Ditindaklanjuti
0	Aktif, Nilai Rendah (Active but Low Value) (R: Tinggi, F: Tinggi, V: Rendah)	Upsell & Cross-sell (Peningkatan & Penjualan Silang). Dorong mereka untuk meningkatkan nilai rata-rata pesanan mereka.	– -Rekomendasi Produk Premium: Tawarkan produk dengan harga/margin lebih tinggi berdasarkan riwayat pembelian mereka. – -Bundling: Buat paket produk untuk meningkatkan nilai transaksi. – Gratis Ongkir Bersyarat: Tetapkan ambang batas nilai transaksi yang sedikit lebih tinggi untuk mendapatkan pengiriman gratis
1	Potensial atau Berisiko (Potential/At-Risk) (R: Sedang, F: Rendah, V: Sangat Rendah)	Activation & Engagement (Aktivasi & Keterlibatan). Dorong pembelian kedua atau ketiga untuk membangun kebiasaan.	– Kampanye Win-Back Ringan: Kirim penawaran personal setelah periode tanpa pembelian (lag time) untuk mengingatkan mereka. – Edukasi Produk: Kirim konten yang menunjukkan nilai penuh produk atau cara menggunakannya secara optimal. – - Survei Singkat: Tanyakan alasan mereka belum melakukan pembelian lagi
2	Pelanggan Tidur (Dormant/Lost) (R: Sangat Rendah, F: Sangat Rendah, V: Sangat Rendah)	Reactivation atau Abandon (Reaktivasi atau Lepaskan). Lakukan upaya terakhir untuk menarik mereka kembali	– Kampanye Reaktivasi Agresif: Tawarkan diskon besar atau penawaran ""Kembali & Dapatkan"" yang sangat menarik. – -Analisis Biaya/Manfaat: Jika upaya reaktivasi mahal dan tidak efektif, pertimbangkan untuk mengalihkan anggaran pemasaran dari segmen ini ke segmen lain (0, 1, atau 3) yang lebih responsif.
3	Champions/Pelanggan Terbaik (R: Tinggi, F: Tinggi, V: Tinggi)	Reward & Retention (Penghargaan & Retensi). Pertahankan loyalitas mereka, dan dorong mereka untuk menjadi advokat merek	– -Program Loyalitas VIP: Beri preview eksklusif produk baru atau diskon khusus sebagai tanda terima kasih. – -Kumpulkan Ulasan: Minta mereka memberikan testimoni/referensi (program referral) untuk menarik pelanggan baru. – -Layanan Prioritas: Pastikan pengalaman pembelian mereka selalu mulus

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menerapkan algoritma *K-Means Clustering* pada model perilaku pelanggan RFM (*Recency*, *Frequency*, Value) untuk membagi 1.917 pelanggan unik Transjakarta menjadi segmen-segmen yang homogen. Melalui analisis *Elbow* Curve dan *Silhouette Score*, model *K-Means* ditetapkan dengan k=4 klaster. Hasil segmentasi ini memberikan kerangka kerja yang kuat untuk memahami dan menargetkan pelanggan dengan lebih efektif.

Analisis data awal (EDA) menunjukkan bahwa hampir separuh pelanggan (49.3%) menggunakan kartu dari Bank DKI, dan basis pengguna didominasi oleh komuter usia produktif (25 hingga 45 tahun). Dalam hal mobilitas, rute pulang-pergi Rusun Kapuk Muara - Penjaringan adalah yang tersibuk, menegaskan adanya koneksi harian yang sangat padat di koridor tersebut.

Sebagai tambahan pada rekomendasi spesifik klaster RFM, manajemen Transjakarta disarankan untuk memperkuat kemitraan strategis dengan Bank DKI , mengingat dominasi kartu bank tersebut, yang mencakup 49,3% dari basis pelanggan unik. Kemitraan ini dapat diwujudkan melalui program loyalitas bersama atau layanan eksklusif untuk meningkatkan retensi dan *customer value* dari segmen pelanggan terbesar. Selain itu, karena Transjakarta didominasi oleh kelompok usia produktif dan komuter , strategi pemasaran dan pengembangan layanan perlu difokuskan untuk memenuhi kebutuhan mobilitas harian kelompok ini. Terakhir, berdasarkan temuan halte tersibuk, perlu adanya fokus pada optimalisasi kapasitas, frekuensi, dan kenyamanan di koridor yang sangat sibuk, khususnya yang melibatkan Halte Penjaringan dan rute bolak-balik Rusun Kapuk Muara – Penjaringan.

DAFTAR REFERENSI

Adiana, B. E., Soesanti, I., & Permanasari, A. E. (2018). Analisis segmentasi pelanggan menggunakan kombinasi RFM model dan teknik clustering. Jurnal Terapan Teknologi Informasi, 2(1), 23–32.

Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means clustering ujian nasional Sekolah Menengah Pertama di Indonesia tahun 2018/2019. Jurnal Media Informatika Budidarma, 4(1), 51–58.

Alhamdani, F. D. S., Dianti, A. A., & Azhar, Y. (2021). Segmentasi pelanggan berdasarkan perilaku penggunaan kartu kredit menggunakan metode K-Means clustering. JISKA (Jurnal Informatika Sunan Kalijaga), 6(2), 70–77.

Anam, K., Rusyana, R., Nurhakim, B., & Pratama, D. (2024). Analisis tingkat penggunaan gadget pada anak usia dini dengan menggunakan K-Means. Jurnal Informatika dan Rekayasa Perangkat Lunak, 6(1), 281–288.

Harani, N. H., Prianto, C., & Nugraha, F. A. (2020). Segmentasi pelanggan produk digital service Indihome menggunakan algoritma K-Means berbasis Python. Jurnal Manajemen Informatika (JAMIKA), 10(2), 133–146.

Hardiani, T., Sulistyo, S., & Hartanto, R. (2015). Segmentasi nasabah tabungan menggunakan model RFM (Recency, Frequency, Monetary) dan K-Means pada lembaga keuangan mikro. Seminar Nasional Teknologi Informasi dan Komunikasi Terapan, 463–468.

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327–1332.

Khobzi, H., Akhondzadeh-Noughabi, E., & Minaei-Bidgoli, B. (2014). A new application of RFM clustering for guild segmentation to mine the pattern of using banks' e-payment services. Journal of Global Marketing, 27(3), 178–190.

Merliana, N. P. E., & Santoso, A. J. (2015). Analisa penentuan jumlah cluster terbaik pada metode K-Means clustering.

Rohman, N., & Wibowo, A. (2024). Perbandingan metode K-Medoids dan metode K-Means dalam analisis segmentasi pelanggan mall. SINTECH (Science and Information Technology) Journal, 7(1), 49–58.

Zakariyya, R. H. (2020). Customer segmentation by using RFM model and K-Means clustering in PT XYZ. Telkom University, 1–10.