



Analisis Pola Perjalanan Penumpang LRT di Palembang Menggunakan Metode K-Means Clustering Berdasarkan Data Gate dan Waktu Perjalanan

Rahman Ajie¹, Siti Sa'uda^{2*}, Ahkmad Ipandy³

¹⁻²Program Studi Teknik Informatika, Universitas Bina Darma, Indonesia

³Kementerian Perhubungan, Indonesia

*Penulis Korespondensi: sitisauda@binadarma.ac.id

Abstract. This study analyzes travel patterns of Single Trip ticket users of the Light Rail Transit (LRT) in Palembang City using the K-Means Clustering method within the Knowledge Discovery in Databases (KDD) framework. The dataset comprises 9,377 Single Trip ticket transaction records from the BPKARSS ticketing system on August 6, 2025. Modeled attributes include origin station code (gate-in), destination station code (gate-out), departure time in decimal format, and estimated travel duration derived through feature engineering. The Elbow Method identified $K=3$ as the selected cluster count based on the inflection point of the WCSS curve, further supported by a Silhouette Score of 0.3044 (moderate category), indicating discernible but partially overlapping clusters. K-Means with K-Means++ initialization produced three passenger segments: Cluster 0: Short-Distance Travel Pattern (39.8%; $n=3,735$), Cluster 1: Medium-Distance Travel Pattern (29.0%; $n=2,724$), and Cluster 2: Long-Distance Travel Pattern (31.1%; $n=2,918$). The results indicate that all cluster centroids fall within the midday period (12:45–13:45 WIB), which differs from conventional peak-hour assumptions. These findings offer an empirical basis for consideration in operational policy formulation.

Keywords: K-Means Clustering; Knowledge Discovery in Databases; Palembang LRT; Passenger Segmentation Travel Pattern.

Abstrak. Penelitian ini menganalisis pola perjalanan penumpang pengguna tiket Single Trip Light Rail Transit (LRT) Kota Palembang menggunakan metode K-Means Clustering dalam kerangka Knowledge Discovery in Databases (KDD). Data yang digunakan berjumlah 9.377 catatan transaksi tiket Single Trip yang bersumber dari sistem ticketing Balai Pengelola Kereta Api Ringan Sumatera Selatan (BPKARSS) pada tanggal 6 Agustus 2025. Atribut pemodelan meliputi kode stasiun asal (gate-in), kode stasiun tujuan (gate-out), waktu keberangkatan dalam format desimal, serta estimasi durasi perjalanan yang direkayasa melalui feature engineering. Penentuan jumlah kluster menggunakan Elbow Method menghasilkan nilai $K=3$ sebagai titik patahan kurva WCSS, yang dikonfirmasi melalui Silhouette Score sebesar 0,3044 (kategori sedang/moderate). Algoritma K-Means dengan inisialisasi K-Means++ menghasilkan tiga segmen penumpang: Kluster 0: Pola Perjalanan Jarak Pendek (39,8%; $n=3.735$), Kluster 1: Pola Perjalanan Menengah (29,0%; $n=2.724$), dan Kluster 2: Pola Perjalanan Jarak Jauh (31,1%; $n=2.918$). Hasil analisis menunjukkan bahwa titik sentroid ketiga kluster berada pada rentang siang hari (pukul 12:45–13:45 WIB), berbeda dari asumsi jam puncak konvensional (pagi dan sore hari). Temuan ini dapat menjadi pertimbangan empiris dalam perumusan kebijakan operasional.

Kata kunci: K-Means Clustering; LRT Palembang; Penemuan Pengetahuan Dalam Basis Data; Pola Perjalanan; Segmentasi Penumpang.

1. LATAR BELAKANG

Pertumbuhan urbanisasi di kota-kota besar Indonesia mendorong kebutuhan terhadap sistem transportasi massal yang efisien. Di Kota Palembang, Light Rail Transit (LRT) Sumatera Selatan yang diresmikan pada tahun 2018 untuk mendukung penyelenggaraan Asian Games merupakan sistem transportasi berbasis rel pertama yang beroperasi di Indonesia (Gustriansyah et al., 2025; Jamilah & Amanah, 2024). Sistem ini melayani koridor sepanjang 23,4 kilometer yang menghubungkan 13 stasiun dari Bandara Internasional Sultan Mahmud Badaruddin II di ujung utara hingga Stasiun DJKA di kawasan Ogan Permata Indah di ujung selatan. Dalam operasionalnya, sistem ticketing LRT menggunakan dua skema pembayaran:

Kartu Uang Elektronik (KUE) untuk komuter reguler dan tiket Single Trip berbasis kertas untuk penumpang insidental yang dibeli langsung di loket stasiun. Operasional harian LRT Palembang menghasilkan volume data transaksi yang cukup besar melalui sistem ticketing di loket stasiun. Setiap transaksi merekam atribut stasiun asal, stasiun tujuan, dan waktu keberangkatan penumpang. Berbeda dengan pengguna KUE yang menunjukkan pola komuter yang relatif teratur, pengguna tiket Single Trip cenderung mencakup wisatawan, pengunjung sesekali, dan penumpang insidental dengan variasi waktu keberangkatan serta pilihan rute yang lebih heterogen (Gustriansyah et al., 2025; Hagi et al., 2023). Heterogenitas pola ini menjadi tantangan bagi manajemen dalam memprediksi kepadatan stasiun dan merancang kebijakan operasional yang tepat sasaran. Tanpa analisis berbasis data yang sistematis, keputusan terkait penjadwalan staf loket dan alokasi sumber daya berpotensi bersifat kurang terarah (Qin et al., 2025).

Sejumlah penelitian telah mengkaji pola perjalanan penumpang transportasi publik menggunakan teknik data mining (Hagi et al., 2023). Menerapkan metode clustering pada data Automatic Fare Collection (AFC) LRT Jakarta, sementara (Chen et al., 2022) mengintegrasikan model RFM dengan K-Means untuk membangun profil penumpang transit. (Khairunnisa & Hendrawan, 2025; Qin et al., 2025) memanfaatkan K-Means untuk segmentasi periode operasional dan manajemen mobilitas penumpang di terminal bus. Meskipun relevan secara metodologis, penelitian-penelitian tersebut beroperasi pada sistem ticketing digital berbasis smartcard yang secara otomatis merekam waktu keluar (tap-out), sehingga durasi perjalanan aktual tersedia tanpa rekayasa fitur tambahan. Sistem tiket Single Trip berbasis kertas di LRT Palembang tidak merekam tap-out, sehingga menimbulkan tantangan metodologis yang berbeda dan belum dibahas dalam literatur yang ada (Arief Rahman Ramadhan & Anita Diana, 2025; Saputra & Yusuf, 2024). Selain itu, belum terdapat studi yang secara spesifik mengkaji segmentasi penumpang tiket loket non-smartcard di konteks LRT Indonesia.

Penelitian ini mengisi celah tersebut melalui tiga kontribusi. Pertama, analisis difokuskan secara eksklusif pada pengguna tiket Single Trip, segmen yang relatif belum banyak dikaji dalam studi data mining transportasi Indonesia. Kedua, penelitian ini mengembangkan fitur Durasi_Estimasi melalui rekayasa fitur (feature engineering) berbasis selisih indeks posisi geografis stasiun asal dan tujuan dikalikan rata-rata waktu tempuh antara stasiun, sebagai respons terhadap ketiadaan data tap-out pada sistem tiket kertas. Ketiga, penelitian ini menerapkan kerangka Knowledge Discovery in Databases (KDD) secara

sistematis mencakup seleksi data, preprocessing, transformasi, data mining, serta interpretasi hasil untuk mengolah 9.377 catatan transaksi tiket Single Trip menjadi segmentasi penumpang yang bermakna secara operasional (Llatas et al., 2024).

Algoritma K-Means Clustering dipilih sebagai metode utama karena kemampuannya mengelompokkan data tanpa label kategori yang telah ditentukan sebelumnya (unsupervised learning), yang sesuai dengan karakteristik data transaksi tiket loket (Chen et al., 2022; Hagi et al., 2023; Khairunnisa & Hendrawan, 2025; Qin et al., 2025). Jumlah kluster optimal ditentukan menggunakan Elbow Method berbasis nilai Within-Cluster Sum of Squares (WCSS). Penelitian ini bertujuan: (1) menerapkan tahapan KDD untuk memproses dan menyiapkan data transaksi Single Trip; (2) mengimplementasikan K-Means Clustering dengan validasi Elbow Method untuk menentukan jumlah kluster optimal; dan (3) mengekstraksi profil karakteristik tiap kluster sebagai landasan rekomendasi manajerial bagi Balai Pengelola Kereta Api Ringan Sumatera Selatan (BPKARSS).

2. METODE PENELITIAN

Sumber dan Karakteristik Data

Data yang digunakan dalam penelitian ini merupakan data primer operasional berupa log transaksi tiket Single Trip yang diperoleh dari sistem ticketing BPKARSS. Data mencakup seluruh transaksi pada tanggal 6 Agustus 2025 di 13 stasiun LRT Palembang, dengan total 9.377 baris catatan transaksi. Penelitian ini secara eksplisit mengeksklusi data transaksi Kartu Uang Elektronik (KUE).

Tahapan Knowledge Discovery in Databases (KDD)

Penelitian ini mengikuti kerangka kerja KDD secara sistematis sebagaimana diilustrasikan dalam alur berikut:

Tabel 1. Alur Tahapan Knowledge Discovery in Databases (KDD) yang Diterapkan dalam Penelitian. Tahapan meliputi: Seleksi Data → Preprocessing → Transformasi & Feature Engineering → Normalisasi → Data Mining (K-Means) → Interpretasi & Evaluasi.

No	Tahapan
1	Data Selection
2	Preprocessing
3	Transformation
4	Feature Engineering
5	Normalisasi
6	Data Mining (K-Means)
7	Interpretation & Evaluation

Preprocessing dan Pembersihan Data

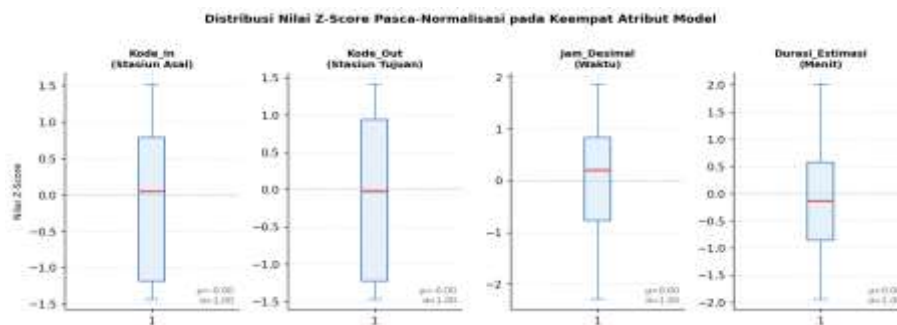
Pemeriksaan missing values dilakukan menggunakan fungsi `isnull().sum()` pada library Pandas Python. Hasil pemeriksaan menunjukkan tidak ditemukan nilai Null/NaN pada ketiga atribut kunci (Stasiun Awal, Stasiun Tujuan, dan Waktu Penjualan). Verifikasi melalui Missing Value Heatmap menunjukkan tidak terdapat indikator data hilang, konsisten dengan hasil pemeriksaan `isnull().sum()`. Tingkat kelengkapan data pada ketiga atribut kunci tercatat sebesar 100%. Setelah proses konversi format waktu dilakukan, baris dengan nilai tidak valid pada atribut `Jam_Desimal`, `Kode_In`, atau `Kode_Out` dieliminasi menggunakan fungsi `dropna()` sebagai langkah sanitasi tambahan, menghasilkan dataset final sebesar 9.377 baris yang seluruhnya siap diproses ke tahap transformasi.

Rekayasa Fitur (Feature Engineering) dan Transformasi Data

Terdapat empat proses transformasi utama yang diterapkan: (1) Label Encoding Stasiun Geospasial: Ketiga belas nama stasiun dikonversi menjadi indeks integer 1 hingga 13 berdasarkan urutan geografis dari utara ke selatan (BANDARA=1 sampai DJKA=13), sehingga selisih antar-indeks merepresentasikan jarak spasial relatif. (2) Konversi Waktu ke Desimal: Atribut Waktu Penjualan dikonversi dari format string HH:MM:SS menjadi bilangan desimal (`Jam_Desimal`) menggunakan rumus: $\text{Jam_Desimal} = \text{Jam} + (\text{Menit} / 60)$. Sebagai contoh, pukul 14:30:00 dikonversi menjadi 14,50. (3) Feature Engineering Durasi_Estimasi: Sistem tiket Single Trip berbasis kertas tidak merekam waktu keluar (tap-out), sehingga durasi perjalanan aktual tidak tersedia sebagai atribut langsung dalam dataset. Ketidadaan data tap-out ini merupakan keterbatasan struktural sistem pencatatan tiket loket yang berbeda dari sistem smartcard/AFC, dan merupakan salah satu tantangan metodologis yang juga dijumpai dalam penelitian transportasi berbasis tiket kertas lainnya (Hagi et al., 2023; Saputra & Yusuf, 2024). Untuk mengatasi keterbatasan ini, fitur Durasi_Estimasi direkayasa menggunakan pendekatan proksi deterministik berbasis jarak indeks antar stasiun. Nilai konstanta 3,5 menit per segmen antara stasiun ditetapkan berdasarkan tiga pertimbangan yang saling menguatkan:

Jadwal operasional resmi LRT Palembang: Dokumen jadwal perjalanan yang diperoleh dari BPKARSS mencatat total waktu tempuh koridor penuh (13 stasiun, 12 segmen antara stasiun) berkisar antara 42–45 menit pada kondisi operasional normal, sehingga rata-rata per segmen adalah $42/12 = 3,50$ menit hingga $45/12 = 3,75$ menit (Gustriansyah et al., 2025; Jamilah & Amanah, 2024). Observasi lapangan selama magang di BPKARSS: Pencatatan waktu perjalanan pada beberapa rute sampel selama periode magang menunjukkan bahwa waktu tempuh antar stasiun yang berdekatan pada kondisi operasional reguler umumnya

berkisar antara 3–4 menit, konsisten dengan informasi jadwal resmi di atas. Prinsip konservatisme estimasi: Nilai 3,5 menit dipilih sebagai batas bawah dari rentang empiris tersebut agar estimasi tidak melebihi-lebihkan durasi aktual, sehingga bias searah yang mungkin timbul bersifat minimal dan dapat diprediksi. Fitur ini direkayasa menggunakan rumus: $\text{Durasi_Estimasi (menit)} = |\text{Indeks_Tujuan} - \text{Indeks_Asal}| \times 3,5$. Perlu ditekankan bahwa rumus ini mengasumsikan keseragaman waktu tempuh antar setiap pasang stasiun, yang merupakan simplifikasi dari kondisi nyata di mana jarak fisik dan kecepatan operasional dapat bervariasi. Meskipun demikian, dalam konteks algoritma K-Means yang dioperasikan pada ruang fitur yang dinormalisasi dengan Z-Score, Durasi_Estimasi berfungsi sebagai proksi ordinal yang merepresentasikan panjang relatif perjalanan antar-penumpang, bukan pengukuran absolut durasi. Konsekuensinya, bias akibat asumsi keseragaman bersifat sistematis dan tidak mengubah urutan relatif antar penumpang dalam ruang fitur, sehingga tidak mempengaruhi validitas hasil klasterisasi sebagai pemisahan segmen berdasarkan pola perjalanan relatif. Implikasi lebih lanjut dari keterbatasan pendekatan ini dibahas pada sub-bagian Hasil dan Pembahasan. Normalisasi Z-Score: Seluruh atribut dinormalisasi menggunakan Z-Score Standardization (StandardScaler dari scikit-learn) dengan rumus $z = (x - \mu) / \sigma$, agar seluruh fitur memiliki bobot matematis yang setara dalam perhitungan jarak Euclidean.



Gambar 1. Distribusi Nilai Z-Score Pasca-Normalisasi (StandardScaler) pada Keempat Atribut Model.

Interpretasi: Normalisasi Z-Score berhasil menyetarakan skala keempat atribut dengan rata-rata mendekati 0 dan standar deviasi 1 pada seluruh fitur. Variasi outlier terlihat pada atribut Durasi_Estimasi, konsisten dengan keberagaman jarak tempuh antar-penumpang dalam dataset.

Tabel 2. Kamus Atribut dan Transformasi Data K-Means.

Atribut Asal	Tipe Asal	Tipe Transformasi	Keterangan
Stasiun Awal	String	Integer (1–13)	Label Encoding berbasis posisi geografis utara-selatan (nama stasiun disesuaikan dengan format kode sistem ticketing BPKARSS)
Stasiun Tujuan	String	Integer (1–13)	Label Encoding berbasis posisi geografis utara-selatan
Waktu Penjualan	Datetime	Float (Jam_Desimal)	Konversi: Jam + (Menit/60)
Durasi_Estimasi	(fitur baru)	Float (menit)	$ \text{Indeks_Tujuan} - \text{Indeks_Asal} \times 3,5$ menit/segmen (proksi ordinal jarak relatif; konstanta dari jadwal resmi BPKARSS & observasi lapangan; lihat 3.4)

Konfigurasi Parameter K-Means

Model K-Means final dieksekusi menggunakan scikit-learn dengan konfigurasi parameter sebagaimana tercantum pada Tabel 3 berikut.

Tabel 3. Konfigurasi Parameter Komputasi K-Means (scikit-learn).

Parameter	Nilai	Deskripsi
n_clusters	3	Jumlah klaster hasil penentuan Elbow Method dan konfirmasi Silhouette Score
init	k-means++	Inisialisasi centroid heuristik untuk mengurangi risiko local optimum
random_state	42	Seed untuk reproducibility hasil klasterisasi
n_init	10	Jumlah pengulangan untuk memilih inersia terendah

Evaluasi Kualitas Klaster: Silhouette Score

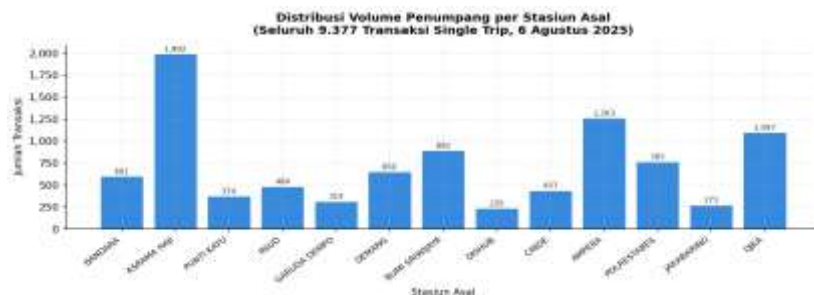
Selain penentuan jumlah klaster melalui Elbow Method, kualitas klaster yang dihasilkan dievaluasi menggunakan Silhouette Score sebagai metrik validasi internal (internal validity index). Berbeda dengan Elbow Method yang hanya mengukur kompaktness intra-klaster melalui WCSS, Silhouette Score mengukur dua dimensi kualitas secara bersamaan: (1) seberapa kohesif setiap titik data terhadap klaster asalnya, dan (2) seberapa terpisah titik data tersebut dari klaster lain yang terdekat (Saputra & Yusuf, 2024). Untuk setiap titik data x_i , Silhouette Score $s(i)$ didefinisikan sebagai: $s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\} \dots (4)$. di mana $a(i)$ adalah rata-rata jarak titik x_i terhadap seluruh titik dalam klaster yang sama (intra-cluster distance), dan $b(i)$ adalah rata-rata jarak minimum x_i terhadap titik-titik di klaster lain yang terdekat (nearest-cluster distance). Nilai $s(i)$ berkisar antara -1 dan $+1$: nilai mendekati $+1$ mengindikasikan titik data berada jauh dari batas antar-klaster; nilai mendekati 0 mengindikasikan titik data berada di dekat batas klaster; dan nilai negatif mengindikasikan kemungkinan penugasan klaster yang keliru. Rata-rata seluruh nilai $s(i)$ menghasilkan Silhouette Score keseluruhan model yang digunakan sebagai ukuran kualitas partisi klaster

secara global. Dalam penelitian ini, Silhouette Score dihitung menggunakan fungsi `silhouette_score` dari modul `sklearn.metrics` pada Python, dieksekusi pada ruang fitur yang telah dinormalisasi (post-StandardScaler) menggunakan `metric='euclidean'` yang konsisten dengan metrik jarak yang digunakan dalam algoritma K-Means. Hasil evaluasi ini dilaporkan pada sub-bagian Hasil dan Pembahasan Evaluasi Kualitas. Nilai Silhouette Score yang diperoleh sebesar 0,3044 (overall) dengan rincian per klaster: Klaster 0 = 0,1745; Klaster 1 = 0,3804; dan Klaster 2 = 0,3996.

3. HASIL DAN PEMBAHASAN

Analisis Deskriptif Dataset

Dataset terdiri dari 9.377 transaksi tiket Single Trip yang mencakup pergerakan penumpang di 13 stasiun LRT Palembang pada satu hari operasional. Dari hasil Exploratory Data Analysis (EDA), distribusi volume penumpang menunjukkan variasi antar stasiun, dengan Stasiun Asrama Haji tercatat sebagai salah satu titik dengan frekuensi transaksi tertinggi, baik sebagai stasiun asal maupun tujuan.



Gambar 2. Distribusi Volume Penumpang per Stasiun Asal (n = 9.377 Transaksi, 6 Agustus 2025).

Interpretasi: Stasiun Asrama Haji mencatat volume transaksi tertinggi sebagai stasiun asal, diikuti Stasiun DJKA dan Jakabaring. Distribusi ini mencerminkan pola kepadatan penggunaan layanan LRT yang tidak merata antar stasiun pada hari yang dianalisis.

Penanganan Missing Values

Sebagaimana diuraikan pada tahap preprocessing (sub dibawah ini), pemeriksaan missing values pada ketiga atribut kunci tidak menemukan nilai kosong dengan data completeness 100%. Setelah sanitasi pasca-konversi menggunakan `dropna()`, dataset final yang digunakan dalam proses klasterisasi berjumlah 9.377 transaksi.



Gambar 3. Heatmap Missing Values pada Atribut Kunci Dataset Single Trip LRT Palembang. Warna hijau penuh mengindikasikan data completeness 100% pada seluruh atribut.

Interpretasi: Tidak ditemukan satu pun nilai kosong (missing value) pada seluruh atribut kunci sebelum maupun sesudah rekayasa fitur. Kondisi ini mengonfirmasi kualitas data primer yang diperoleh dari sistem ticketing BPKARSS.

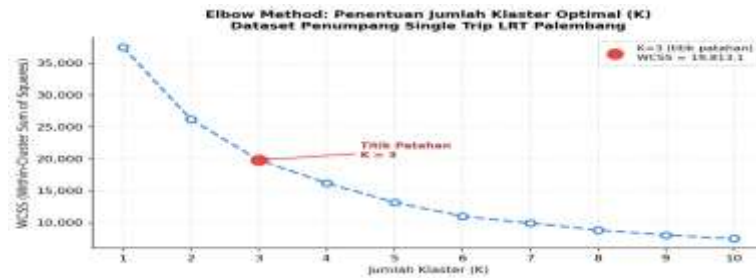


Gambar 4. Distribusi Waktu Keberangkatan Penumpang Single Trip (Format Desimal). Rentang sentroid klaster (12:45–13:45 WIB) ditandai dengan area berbayang merah.

Interpretasi: Distribusi waktu menunjukkan pola multimodal dengan konsentrasi pada siang hari. Area berbayang (12:45–13:45 WIB) mencakup posisi sentroid ketiga klaster, mengindikasikan dominasi perjalanan di luar jam puncak konvensional pada pengguna tiket Single Trip.

Validasi Jumlah Klaster: Elbow Method

Untuk menentukan nilai K secara objektif, algoritma K-Means diuji coba pada rentang K=1 hingga K=10 dengan merekam nilai Within-Cluster Sum of Squares (WCSS) di setiap iterasi: $WCSS = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - c_j\|^2$. Hasil Elbow Curve menunjukkan penurunan WCSS yang cukup tajam dari K=1 menuju K=3. Melewati K=3, penurunan WCSS menjadi lebih landai. Titik patahan pada K=3 ini diinterpretasikan sebagai jumlah klaster yang sesuai untuk dataset penumpang Single Trip. Konvergensi algoritma pada 9.377 data rata-rata tercapai pada iterasi ke-14. Untuk mengevaluasi kualitas partisi yang dihasilkan secara lebih komprehensif, penentuan K melalui Elbow Method dilengkapi dengan perhitungan Silhouette Score sebagaimana diuraikan pada sub-bab dibawah ini.



Gambar 5. Kurva Elbow Method: Penentuan Jumlah Kluster Optimal (K=1–10).

Titik patahan teridentifikasi pada K=3 (ditandai lingkaran merah).

Interpretasi: Penurunan WCSS yang cukup tajam terjadi dari K=1 menuju K=3. Melewati K=3, laju penurunan melambat secara visual, mengindikasikan bahwa penambahan kluster tidak lagi memberikan gain kompaktness yang proporsional. K=3 dipilih sebagai konfigurasi model final.

Evaluasi Kualitas Kluster: Silhouette Score

Kualitas partisi kluster yang dihasilkan dievaluasi menggunakan Silhouette Score sebagai metrik validasi internal (internal validity index). Evaluasi dilakukan menggunakan fungsi `silhouette_score` dari modul `sklearn.metrics` dengan parameter `metric='euclidean'`, konsisten dengan metrik jarak yang digunakan dalam algoritma K-Means.

Tabel 4. Silhouette Score per Kluster.

Kluster	Label Segmen	Silhouette Score
Klaster 0	Pola Perjalanan Jarak Pendek	0,1745
Klaster 1	Pola Perjalanan Menengah	0,3804
Klaster 2	Pola Perjalanan Jarak Jauh	0,3996
Rata-rata (Overall)	—	0,3044

Hasil evaluasi menunjukkan bahwa nilai Silhouette Score keseluruhan model sebesar 0,3044, yang mengindikasikan kualitas pemisahan kluster pada kategori cukup (moderate clustering structure). Berdasarkan hasil per kluster pada Tabel 4, Klaster 0 memperoleh nilai 0,1745 yang menunjukkan tingkat separasi relatif rendah, mengindikasikan adanya kedekatan karakteristik dengan kluster lain. Sementara itu, Klaster 1 dan Klaster 2 memperoleh nilai masing-masing sebesar 0,3804 dan 0,3996, yang menunjukkan kualitas pengelompokan lebih baik dengan tingkat kohesi internal yang lebih stabil. Hasil ini menunjukkan bahwa metode K-Means berhasil membentuk segmentasi pola perjalanan penumpang, meskipun masih terdapat kemungkinan overlap karakteristik antar segmen akibat dinamika mobilitas pengguna transportasi publik (Chen et al., 2022; Hagi et al., 2023; Saputra & Yusuf, 2024).

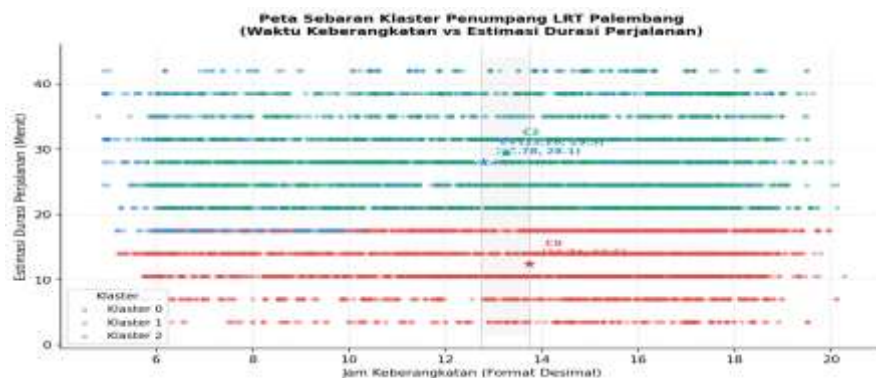
Dikombinasikan dengan titik patahan Elbow Method pada K=3, nilai Silhouette Score sebesar 0,3044 memberikan landasan evaluatif yang memadai untuk menetapkan K=3 sebagai konfigurasi klasterisasi yang dipilih dalam penelitian ini.

Profiling Klaster Penumpang

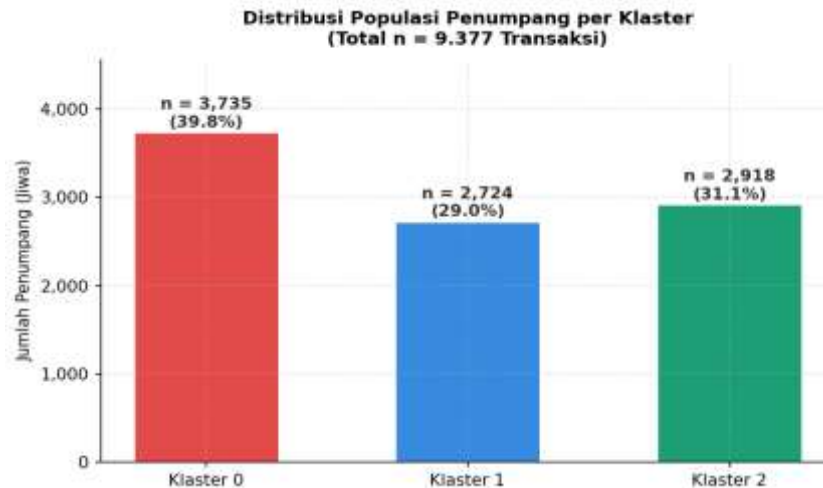
Setelah K=3 dipilih berdasarkan titik patahan Elbow Curve, seluruh 9.377 data transaksi dilabeli ke dalam tiga klaster. Nilai centroid akhir dikonversi kembali ke skala asli (de-normalisasi) untuk memudahkan interpretasi. Tabel 5 merangkum karakteristik deskriptif ketiga klaster yang terbentuk.

Tabel 5. Profiling dan Tabulasi Karakteristik Klaster Penumpang.

Atribut / Fitur	Klaster 0	Klaster 1	Klaster 2
Total Penumpang	3.735 jiwa	2.724 jiwa	2.918 jiwa
Persentase Populasi	39,8%	29,0%	31,1%
Centroid Waktu (Jam Desimal)	13,74 (Pk. 13:44 WIB)	12,78 (Pk. 12:46 WIB)	13,26 (Pk. 13:15 WIB)
Rata-rata Durasi Tempuh	12,50 menit	28,08 menit	29,49 menit
Stasiun Keberangkatan Dominan	Bumi Sriwijaya	Ampera	Asrama Haji
Stasiun Tujuan Dominan	Asrama Haji	Asrama Haji	DJKA
Label Segmen	Pola Perjalanan Jarak Pendek	Pola Perjalanan Menengah	Pola Perjalanan Jarak Jauh



Gambar 6. Scatter Plot Sebaran Klaster Penumpang pada Ruang Fitur Waktu Keberangkatan (Jam Desimal) dan Estimasi Durasi Perjalanan (Menit). Bintang (★) menunjukkan posisi sentroid tiap klaster.



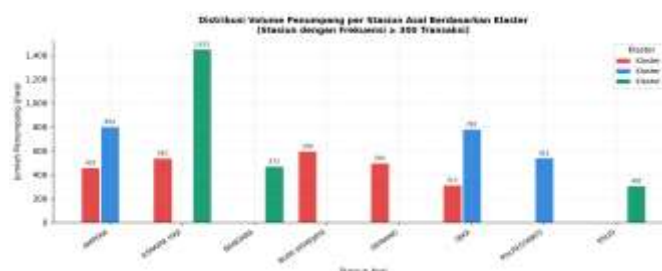
Gambar 7. Distribusi Populasi Penumpang per Klaster (Total n = 9.377 Transaksi).

Deskripsi Profil Tiap Klaster

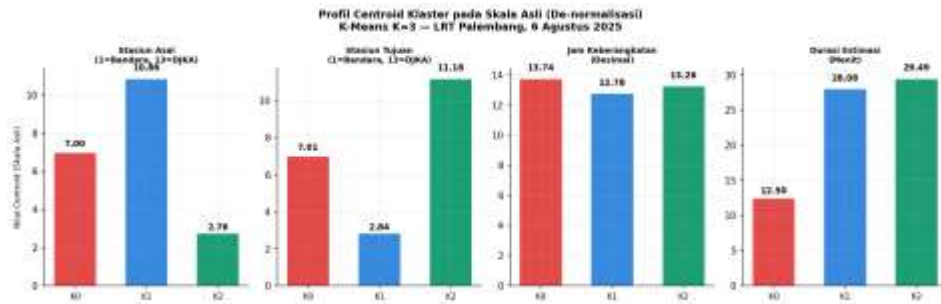
Klaster 0: Pola Perjalanan Jarak Pendek (39,8%): Klaster ini mencakup 3.735 transaksi (39,8% populasi). Karakteristik utamanya adalah durasi perjalanan yang relatif singkat (rata-rata 12,50 menit) dengan titik sentroid waktu pada pukul 13:44 WIB. Pola spasial menunjukkan pergerakan yang cenderung terpusat dari Stasiun Bumi Sriwijaya menuju Stasiun Asrama Haji. Pola ini mengindikasikan kecenderungan perjalanan jarak dekat pada rentang siang hari.

Klaster 1: Pola Perjalanan Menengah (29,0%): Klaster ini mencakup 2.724 transaksi (29,0% populasi) dengan durasi perjalanan menengah (rata-rata 28,08 menit) dan sentroid waktu pada pukul 12:46 WIB. Rute pergerakan cenderung terpusat dari Stasiun Ampera menuju Stasiun Asrama Haji di bagian utara koridor. Pola ini mengindikasikan pergerakan dengan arah dominan selatan-ke-utara pada rentang waktu siang hari.

Klaster 2: Pola Perjalanan Jarak Jauh (31,1%): Klaster ini mencakup 2.918 transaksi (31,1% populasi) dengan rata-rata durasi perjalanan tertinggi di antara ketiga klaster, yakni 29,49 menit. Sentroid waktu berada pada pukul 13:15 WIB. Pola spasial menunjukkan pergerakan dari Stasiun Asrama Haji (ujung utara koridor) menuju Stasiun DJKA (ujung selatan koridor), yang merepresentasikan perjalanan sepanjang koridor secara penuh.



Gambar 8. Distribusi Volume Penumpang pada Stasiun Asal Dominan per Klaster (ditampilkan untuk stasiun dengan frekuensi ≥ 300 transaksi).



Gambar 9. Profil Centroid Klaster pada Skala Asli (De-normalisasi). Visualisasi nilai centroid pada ruang fitur asli menunjukkan perbedaan karakteristik antar klaster berdasarkan stasiun asal, stasiun tujuan, jam keberangkatan, dan durasi estimasi perjalanan penumpang LRT Palembang.

Interpretasi: Centroid de-normalisasi mengonfirmasi perbedaan karakteristik antar klaster: Klaster 0 memiliki indeks stasiun asal dan tujuan yang berdekatan (jarak dekat), Klaster 1 menunjukkan pola arah utara (Kode_Out rendah), dan Klaster 2 mencerminkan perjalanan dari stasiun utara ke selatan koridor (Kode_Out tinggi). Jam keberangkatan ketiga klaster konvergen pada rentang 12,78–13,74 (siang hari).

Temuan Anomali Pola Waktu

Temuan yang perlu diperhatikan adalah posisi sentroid seluruh klaster yang berada pada rentang siang hari (pukul 12:45–13:45 WIB). Pola ini berbeda dari jam puncak konvensional yang umumnya berlangsung pada pagi (06:00–08:00 WIB) dan sore hari (16:00–18:00 WIB) seiring mobilitas pekerja dan pelajar (Jamilah & Amanah, 2024; Qin et al., 2025).

Pola ini dapat diinterpretasikan sebagai indikasi bahwa pengguna tiket Single Trip pada hari yang dianalisis didominasi oleh perjalanan yang berlangsung di luar jam puncak konvensional. Salah satu penjelasan yang mungkin adalah bahwa penumpang dengan pola komuter reguler lebih banyak menggunakan skema pembayaran lain seperti KUE, meskipun hal ini tidak dapat dikonfirmasi dari data Single Trip yang menjadi cakupan penelitian ini. Interpretasi lebih lanjut memerlukan perbandingan dengan data transaksi KUE pada periode yang sama.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menginterpretasikan hasilnya. Pertama, data yang dianalisis hanya mencakup satu hari operasional (6 Agustus 2025), sehingga pola klaster yang teridentifikasi belum tentu merepresentasikan kondisi pada periode lain seperti hari kerja reguler, akhir pekan, maupun musim libur. Kedua, ketiadaan data tap-out pada sistem tiket Single Trip mengharuskan

rekayasa fitur *Durasi_Estimasi* sebagai proksi ordinal jarak perjalanan relatif, sebagaimana diuraikan di sub-bagian Metode Rekayasa Fitur. Nilai konstanta 3,5 menit per segmen didasarkan pada jadwal operasional resmi BPKARSS dan observasi lapangan yang menghasilkan rentang empiris 3,5–3,75 menit; nilai batas bawah dipilih sebagai pendekatan konservatif. Terdapat dua implikasi keterbatasan yang perlu dicatat: pertama, pendekatan ini mengasumsikan keseragaman waktu tempuh antara setiap pasang stasiun, yang merupakan simplifikasi dari kondisi nyata di mana kecepatan operasional, panjang segmen fisik, dan durasi berhenti di peron dapat bervariasi; kedua, pendekatan ini tidak memperhitungkan kondisi tidak normal seperti keterlambatan, perubahan jadwal mendadak, atau penumpukan penumpang yang memperpanjang waktu tunggu. Meskipun demikian, karena *Durasi_Estimasi* digunakan sebagai atribut dalam ruang fitur yang dinormalisasi (Z-Score), bias akibat asumsi keseragaman bersifat sistematis dan proporsional, sehingga tidak mengubah urutan relatif antar penumpang dalam ruang komputasi K-Means. Dengan kata lain, penumpang yang melakukan perjalanan lebih jauh tetap mendapatkan nilai *Durasi_Estimasi* yang secara konsisten lebih tinggi dibandingkan penumpang yang melakukan perjalanan lebih dekat, yang merupakan properti ordinal yang diperlukan dan cukup untuk tujuan segmentasi berbasis clustering ini.

Ketiga, dataset tidak memuat informasi demografi penumpang seperti usia, jenis kelamin, atau tujuan perjalanan, sehingga interpretasi profil kluster sepenuhnya didasarkan pada atribut spasial dan temporal yang tersedia. Profil seperti "perjalanan siang hari" atau "perjalanan lintas kota" merupakan deskripsi berbasis pola data, bukan kategorisasi yang diverifikasi dari data penumpang secara langsung. Selain itu, visualisasi distribusi stasiun asal pada Gambar 4 hanya menampilkan stasiun dengan frekuensi transaksi ≥ 300 sebagai threshold tampilan; seluruh data tetap diikutsertakan dalam proses klasterisasi tanpa filter apapun. Keterbatasan-keterbatasan ini tidak mengurangi validitas metodologis penelitian, namun membuka ruang bagi penelitian lanjutan untuk menggunakan cakupan data multi-hari atau multi-bulan, mengintegrasikan data tap-out dari sistem yang lebih lengkap, serta menggabungkan variabel demografi guna menghasilkan segmentasi yang lebih komprehensif.

4. KESIMPULAN DAN SARAN

Penelitian ini menerapkan metode K-Means Clustering dalam kerangka Knowledge Discovery in Databases (KDD) terhadap 9.377 data transaksi tiket Single Trip LRT Palembang pada tanggal 6 Agustus 2025. Hasil penelitian menunjukkan bahwa penumpang dapat dikelompokkan menjadi tiga klaster berdasarkan empat atribut utama, yaitu stasiun asal (gate-

in), stasiun tujuan (gate-out), waktu perjalanan, dan estimasi durasi perjalanan. Ketiga klaster yang terbentuk mengindikasikan adanya perbedaan pola mobilitas yang dapat menjadi referensi dalam evaluasi pola perjalanan penumpang LRT Palembang.

Tujuan pertama menerapkan tahapan KDD untuk memproses dan menyiapkan data transaksi Single Trip dicapai melalui serangkaian tahapan sistematis yang meliputi seleksi data, pemeriksaan missing values (data completeness 100%), sanitasi pasca-konversi menggunakan dropna(), label encoding stasiun berbasis posisi geografis, konversi waktu ke format desimal, rekayasa fitur Durasi_Estimasi sebagai proksi jarak relatif, dan normalisasi Z-Score pada seluruh atribut. Keseluruhan tahapan ini menghasilkan dataset empat dimensi berjumlah 9.377 baris yang siap dimodelkan oleh algoritma K-Means.

Tujuan kedua—mengimplementasikan K-Means Clustering dengan penentuan jumlah klaster yang tervalidasi secara objektif dicapai melalui dua tahapan evaluasi komplementer. Elbow Method pada rentang K=1 hingga K=10 mengidentifikasi titik patahan kurva WCSS pada K=3. Hasil ini dikonfirmasi melalui perhitungan Silhouette Score sebesar 0,3044 (overall), yang berada pada kategori sedang dan mengindikasikan struktur klaster yang dapat dibedakan meskipun terdapat tumpang-tindih parsial yang wajar pada data transportasi publik. Model K-Means dengan inisialisasi K-Means++ dan random_state=42 mencapai konvergensi rata-rata pada iterasi ke-14.

Tujuan ketiga mengekstraksi profil karakteristik tiap klaster sebagai landasan rekomendasi manajerial dicapai melalui de-normalisasi nilai sentroid dan interpretasi distribusi spasiotemporal ketiga segmen yang terbentuk: Klaster 0 (Pola Perjalanan Jarak Pendek, 39,8%; durasi rata-rata 12,50 menit; Bumi Sriwijaya→Asrama Haji), Klaster 1 (Pola Perjalanan Menengah, 29,0%; 28,08 menit; Ampera→Asrama Haji), dan Klaster 2 (Pola Perjalanan Jarak Jauh, 31,1%; 29,49 menit; Asrama Haji→DJKA). Profil ketiga klaster ini telah dirumuskan menjadi rekomendasi operasional pada saran dibawah ini.

Temuan lintas-klaster yang perlu dicatat adalah posisi sentroid seluruh klaster pada rentang siang hari (pukul 12:45–13:45 WIB), yang berbeda dari pola jam puncak konvensional pada pagi dan sore hari. Pola ini mengindikasikan bahwa pengguna tiket Single Trip pada hari yang dianalisis cenderung melakukan perjalanan di luar jam komuter reguler. Secara keseluruhan, hasil segmentasi yang diperoleh dapat dijadikan referensi dalam evaluasi pola mobilitas penumpang LRT Palembang, khususnya untuk mendukung perencanaan sumber daya operasional yang lebih terarah oleh manajemen BPKARSS. Optimalisasi Operasional Siang Hari: Manajemen BPKARSS dapat mempertimbangkan penyesuaian alokasi staf loket

dan ketersediaan mesin Vending Machine di stasiun dengan volume transaksi tertinggi—khususnya Stasiun Bumi Sriwijaya, Ampera, dan Asrama Haji—pada rentang waktu siang hari, berdasarkan pola sentroid klaster yang teridentifikasi dalam penelitian ini. Program Konversi ke KUE: Segmen Klaster 2 yang menunjukkan pola perjalanan jarak jauh dengan tiket kertas berpotensi untuk didorong beralih ke KUE. Penempatan stand edukasi dan penjualan KUE di Stasiun Asrama Haji dan DJKA direkomendasikan untuk mendorong transisi ini. Pengembangan Pemodelan Lanjutan: Penelitian selanjutnya disarankan untuk menggunakan rentang data yang lebih panjang (multi-hari atau multi-bulan) guna menguji konsistensi pola klaster pada periode yang berbeda. Pengintegrasian variabel demografi penumpang, apabila tersedia secara teknis, berpotensi memperkaya interpretasi profil segmen yang dihasilkan. Penerapan Berkala Oleh BPKARSS: Skrip Python yang dikembangkan dalam penelitian ini berpotensi dioperasikan secara berkala oleh pihak pengelola untuk memperbarui segmentasi penumpang seiring bertambahnya data transaksi, sehingga mendukung pengambilan keputusan operasional yang lebih adaptif.

DAFTAR REFERENSI

- Arief Rahman Ramadhan, & Anita Diana. (2025). Penerapan algoritma K-means untuk klasterisasi jumlah penumpang angkutan umum di Jakarta. *Infotek: Jurnal Informatika dan Teknologi*, 8(1), 228–238. <https://doi.org/10.29408/jit.v8i1.28195>
- Chen, A. H. L., Liang, Y. C., Chang, W. J., Siau, H. Y., & Minanda, V. (2022). RFM model and K-means clustering analysis of transit traveller profiles: A case study. *Journal of Advanced Transportation*, 2022, 1108105. <https://doi.org/10.1155/2022/1108105>
- Gustriansyah, R., Puspasari, S., Sanmorino, A., & Suhandi, N. (2025). Forecasting light rail transit passenger demand using parameter-tuned exponential smoothing models. *Journal of Information Engineering and Educational Technology*, 9(2), 101–112. <https://doi.org/10.26740/jieet.v9n2.p101-112>
- Hagi, M., Suroyo, H., Wijaya, A., & Syazili, A. (2023). Data analysis automatic fare collection Light Rail Transit Jakarta using the cluster method. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 4(2), 465–473. <https://doi.org/10.30596/jcositte.v4i2.16413>
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jamilah, W., & Amanah, T. (2024). Performance analysis of public transport feeder Light Rapid Transit (LRT) Palembang South Sumatra. *Atlantis Press International BV*. https://doi.org/10.2991/978-94-6463-386-3_24

- Khairunnisa, H. R., & Hendrawan, A. (2025). LSTM forecasting and K-means clustering for passenger mobility management at bus terminals. *Journal of Information Systems and Informatics*, 7(3), 2129–2146. <https://doi.org/10.51519/journalisi.v7i3.1159>
- Llatas, C., Soust-Verdaguer, B., Torres, L. C., & Cagigas, D. (2024). Application of knowledge discovery in databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design. *Journal of Building Engineering*, 91, 109546. <https://doi.org/10.1016/j.jobbe.2024.109546>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
- Qin, Y., Guo, J., Xu, P., Wang, L., & Xia, B. (2025). Passenger flow-oriented operating period division in urban rail transit: A hybrid SOM and K-means clustering approach. *Symmetry*, 17(11), 11860. <https://doi.org/10.3390/sym17111860>
- Saputra, A., & Yusuf, R. (2024). Perbandingan algoritma DBSCAN dan K-means dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(4), 1346–1361. <https://journal.irpi.or.id/index.php/malcom/article/view/1516>