Project Akhir Mata Kuliah Pengantar Data Mining
Pemetaan Karakteristik Kejahatan di London: Studi Klaster Berdasarkan Data MPS LSOA

Anggota Tim : Salma Dewi Nataya (23/518235/PA/22243)
Melani Sulistiawati (23/523106/PA/22507)

Dosen Pengampu : Dr. Adhitya Ronnie Effendie, S.Si., M.Si., M.Sc.

Program Studi Statistika
Departemen Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Gadjah Mada

2025

Daftar Isi

BAB I
Pendahuluan

1.1 Latar Belakang

Kejahatan merupakan tantangan sosial yang terus berkembang, terutama di kota metropolitan seperti London yang memiliki kepadatan dan keragaman penduduk tinggi. Menurut data dari Metropolitan Police Service (MPS) yang diakses melalui portal data London, tercatat lebih dari 10 juta kasus kejahatan terjadi selama periode April 2010 hingga Maret 2023. Kejahatan dengan frekuensi tertinggi di antaranya adalah *violence against the person, theft, dan vehicle offences.* Tingginya angka ini menandakan bahwa kejahatan bukan hanya isu keamanan, tetapi juga menjadi indikator penting dalam perencanaan wilayah dan kebijakan publik.

Studi ini menggunakan dataset MPS LSOA Level Crime (Historical) yang mencatat jumlah kasus kejahatan di wilayah administrasi terkecil di London, yaitu Lower Super Output Area (LSOA). Data ini mencakup 4.988 LSOA yang tersebar di 33 borough di London, dengan rincian jumlah kasus berdasarkan kategori kejahatan utama (*major categories*) seperti *burglary, robbery, drug offences,* hingga *possession of weapons.* Melalui agregasi data dari kategori minor ke kategori major, diperoleh total kejadian kejahatan untuk setiap LSOA selama periode 2019 hingga awal 2023. Pengolahan data ini bertujuan untuk mengidentifikasi distribusi dan pola kejahatan secara spasial, yang menjadi dasar penting dalam mendukung perumusan kebijakan berbasis bukti (*evidence-based policy making*).

Dalam studi ini, dilakukan analisis *clustering* terhadap wilayah LSOA berdasarkan total kasus dari masing-masing kategori kejahatan utama. Pendekatan ini memungkinkan pengelompokan wilayah-wilayah dengan karakteristik kejahatan yang mirip, sehingga dapat diidentifikasi klaster dengan tingkat kejahatan tinggi, sedang, atau rendah. Dengan membandingkan berbagai metode *clustering* seperti K-Means, K-Medians, *Agglomerative Clustering*, dan Gaussian Mixture Model (GMM), studi ini tidak hanya mengidentifikasi klaster wilayah, namun juga mengevaluasi keefektifan dan kesesuaian metode dalam konteks spasial-kriminal. Dengan hasil *clustering* ini, pembuat kebijakan dapat merancang strategi penanganan yang lebih terarah, seperti peningkatan patroli di daerah rawan atau program rehabilitasi di daerah dengan kasus *drug offences* yang tinggi.

1.2 Rumusan Masalah

1. Bagaimana pola hubungan antarjenis kejahatan yang terjadi di wilayah Kota London berdasarkan korelasinya?

2. Bagaimana wilayah-wilayah di Kota London dapat dikelompokkan berdasarkan kemiripan pola kriminalitas yang dimiliki?

3. Metode *clustering* manakah yang paling optimal dalam mengelompokkan wilayah berdasarkan data kriminalitas, dilihat dari segi performa metrik dan interpretabilitas hasil?

4. Apakah hasil pengelompokan wilayah berdasarkan *clustering* konsisten dengan pola-pola hubungan antarkejahatan yang ditemukan dalam analisis eksploratif?

BAB II
Landasan Teori

2.1 *Clustering*

Clustering merupakan teknik eksplorasi yang secara luas digunakan dalam analisis data di berbagai bidang, termasuk dalam penggabungan dan penyatuan informasi. *Clustering* secara adaptif mengelompokkan data ke dalam beberapa klaster, dimana anggota dalam satu klaster memiliki kemiripan yang sama. Dari banyaknya metode *clustering* K-Means, K-Medians, *Agglomerative*, dan *Gaussian Mixture Model* (GMM) merupakan beberapa metode penting yang telah berhasil diterapkan dalam berbagai situasi. *Clustering* dapat dibagi secara garis besar menjadi 3 langkah utama, yaitu sebagai berikut:

1. Eksplorasi dan Pra-pemrosesan Awal

   Eksplorasi awal dilakukan untuk memahami karakteristik data, termasuk distribusi jumlah kasus kejahatan di tiap kategori dan wilayah. Data tidak mengandung nilai kosong, namun menunjukkan variasi yang tinggi antar kategori. Oleh karena itu, dilakukan pra-pemrosesan berupa standarisasi nilai-nilai numerik agar seluruh variabel berada pada skala yang setara. Langkah ini penting untuk memastikan hasil clustering tidak bias terhadap kategori dengan jumlah kasus yang besar.

2. Pemilihan Nilai K Optimal

   Pemilihan nilai K optimal memerlukan keseimbangan antara kesederhanaan dan kemudahan interpretasi (dengan jumlah klaster yang sedikit) serta kemampuan untuk menangkap struktur data yang kompleks (dengan jumlah klaster yang lebih banyak). Pemilihan nilai K yang tidak tepat dapat menyebabkan hasil klasterisasi menjadi terlalu sederhana atau terlalu kompleks, sehingga mengurangi makna dan kegunaan dari hasil klasterisasi dalam menghasilkan wawasan yang bermakna. Untuk mengatasi permasalahan ini, beberapa teknik telah dikembangkan dan digunakan selama ini, antara lain, Metode Elbow (*Elbow Method*), *Silhouette Score* , Metode Gap Statistics (*Gap Statistics Method*). Dalam studi ini akan digunakan *silhoutte score* dalam penentuan nilai k optimal. Analisis Silhouette merupakan metodologi yang secara kuantitatif mengukur seberapa baik suatu titik data cocok dengan klaster tempat ia ditempatkan dibandingkan dengan klaster tetangganya. Metode ini memberikan nilai skor silhouette untuk setiap titik data, dengan rentang nilai dari -1 hingga 1:

   - Skor mendekati +1 menunjukkan bahwa titik data tersebut sangat cocok dengan klasternya.
   - Skor mendekati 0 menunjukkan bahwa titik data berada di batas antara dua klaster.
   - Skor mendekati -1 mengindikasikan bahwa titik data kemungkinan lebih cocok berada di klaster lain.

   Dengan demikian, *Silhouette Score* memungkinkan kita untuk mengevaluasi struktur klasterisasi dan memilih nilai K yang menghasilkan rata-rata skor silhouette tertinggi, yang berarti struktur klaster terbaik.

3. Evaluasi Hasil Klaster

   Setelah proses klasterisasi dilakukan, langkah selanjutnya adalah mengevaluasi kualitas klaster yang terbentuk. Evaluasi bertujuan untuk menilai seberapa baik model klaster dalam memisahkan data ke dalam kelompok yang homogen secara internal dan heterogen antar klaster. Selain itu, evaluasi juga memperhatikan interpretabilitas hasil, yaitu apakah masing-masing klaster memiliki karakteristik yang jelas dan berbeda berdasarkan jumlah kasus kejahatan di tiap kategori.

Berikut ini merupakan penjelasan lebih lanjut mengenai beberapa algoritma analisis *clustering* yang telah dsebutkan sebelumnya.

2.1.1 Algoritma K-Means

K-Means adalah salah satu algoritma *clustering* yang paling populer dan banyak digunakan dalam analisis data. Algoritma ini bertujuan untuk membagi data menjadi sejumlah k klaster berdasarkan kemiripan fitur dengan cara meminimalkan jarak antar data dalam klaster dan memaksimalkan jarak antar klaster. K-Means bekerja dengan mencari *centroid* (titik pusat) dari setiap klaster dan mengelompokkan data

berdasarkan kedekatan jarak ke *centroid* tersebut. Tujuan utama algoritma ini adalah untuk meminimalkan fungsi objektif berupa jumlah kuadrat jarak antara setiap titik data dengan *centroid* klasternya, sehingga menghasilkan pengelompokan yang kompak dan terpisah dengan baik.

Fungsi objektif K-Means secara matematis dapat dituliskan sebagai berikut.

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{1}$$

dengan

- k: jumlah klaster,
- S_i: klaster ke-i,
- x_j: titik data dalam klaster S_i,
- μ_i: centroid dari klaster S_i,
- $\|x_j - \mu_i\|$: jarak Euclidean antara titik data dan *centroid*.

Proses *clustering* dengan K-Means dapat dijelaskan dalam beberapa tahap berikut.

1. Tentukan jumlah klaster k.
2. Inisialisasi centroid $\mu_1, \mu_2, \ldots, \mu_k$ secara acak.
3. Tetapkan setiap titik data x_j ke klaster dengan centroid terdekat berdasarkan jarak Euclidean.
$$S_i = \{x_j : \|x_j - \mu_i\|^2 \leq \|x_j - \mu_l\|^2, \forall l = 1, \ldots, k\} \tag{2}$$
4. Hitung ulang centroid tiap klaster sebagai rata-rata titik-titiknya.
$$\mu_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j \tag{3}$$
5. Ulangi langkah 3 dan 4 sampai centroid tidak berubah signifikan (konvergen).

Keuntungan dari k-means *clustering* dalam aplikasi machine learning, yaitu mudah dipahami dan dipraktikkan. Selain itu, K-means *clustering* dirancang dengan pendekatan iteratif yang sederhana secara komputasi. Beberapa tantangan umum yang terkait dengan k-means *clustering* meliputi ketergantungan pada parameter input yang diatur dengan benar. Menginisialisasi c*centroid* dan jumlah kluster yang tepat sangat cocok untuk mendapatkan hasil kluster yang bermakna. Inisialisasi *centroid* yang buruk dapat menyebabkan peningkatan waktu berjalan dan penugasan kluster berkualitas rendah. K-means bekerja secara efektif ketika kumpulan data berisi kluster yang ukurannya serupa dan tidak ada outlier atau variasi kepadatan yang menonjol. K-means berkinerja buruk ketika kumpulan data mengandung banyak variasi atau memiliki banyak dimensi.

2.1.2 Algoritma K-Median

K-medians mencoba mengatasi kelemahan K-means yang sangat sensitif terhadap outlier (nilai ekstrem), dengan menggunakan metrik disimilasi (ukuran ketidaksamaan) yang berbeda. Alih-alih menggunakan jarak Euclidean (L2 norm), K-medians biasanya menggunakan selisih absolut, yang juga dikenal sebagai L1 norm atau jarak Manhattan. Penggunaan L1 norm jauh lebih tidak sensitif terhadap outlier, karena outlier hanya berkontribusi sesuai jaraknya yang sebenarnya ke pusat klaster, bukan kuadrat dari jaraknya seperti dalam jarak Euclidean. Agar hasilnya lebih andal, pusat klaster ditentukan menggunakan median bukan mean (rata-rata), karena median juga lebih tahan terhadap outlier. Sehingga, pada akhirnya, yang ingin dioptimalkan dalam algoritma ini adalah penjumlahan total dari jarak absolut antara setiap titik data ke pusat klaster terdekatnya.

Secara matematis, fungsi objektif yang diminimalkan dalam algoritma K-Medians adalah:

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - m_i\|_1 \tag{4}$$

dengan:

- k: jumlah klaster,
- S_i: klaster ke-i (himpunan titik data),
- x_j: titik data dalam klaster S_i,
- m_i: median dari klaster S_i,
- $\|x_j - m_i\|_1$: jarak Manhattan antara titik data dan pusat klaster.

Langkah-langkah pelaksanaan algoritma K-Medians adalah sebagai berikut:

1. Tentukan jumlah klaster k.
2. Inisialisasi pusat klaster $m_1, m_2, \ldots, m_k$ secara acak.
3. Tetapkan setiap titik data x_j ke klaster dengan pusat klaster terdekat berdasarkan jarak Manhattan:

$$S_i = \{x_j : \|x_j - m_i\|_1 \leq \|x_j - m_l\|_1, \quad \forall l = 1, \ldots, k\} \tag{5}$$

4. Hitung ulang pusat klaster sebagai *median* dari titik-titik dalam klaster:

$$m_i = \text{median}\,\{x_j : x_j \in S_i\} \tag{6}$$

5. Ulangi langkah 3 dan 4 hingga pusat klaster konvergen (tidak berubah) atau jumlah iterasi maksimum tercapai.

2.1.3 Gaussian Mixture Model (GMM)

GMM adalah model probabilistik yang mengasumsikan bahwa semua titik data dihasilkan dari campuran sejumlah terbatas distribusi Gaussian dengan parameter yang tidak diketahui. Model ini sering digunakan untuk pembelajaran tanpa pengawasan (*unsurvised learning*) dan klasterisasi lunak (soft clustering) pada data tidak berlabel. Gaussian Mixture Model (GMM) termasuk dalam model berbasis *clustering* atau model distribusi(Deofanny dkk., 2022). Menurut Lin et al. (2019), parameter GMM diestimasikan menggunakan algoritme *Expectation Maximization* (EM) secara iteratif sehingga dapat terklasterkan dengan karakter yang serupa. Untuk Gaussian Mixture Model, diasumsikan bahwa sampel yang diberikan adalah realisasi dari vektor acak yang distribusinya merupakan campuran dari beberapa distribusi.

Fungsi distribusi probabilitas GMM didefinisikan sebagai:

$$\pi(x) = \sum_{k=1}^{K} p_k \cdot \mathcal{N}(X|\mu_k, \Sigma_k) \tag{7}$$

dengan fungsi kepadatan Gaussian:

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \tag{8}$$

dan batasan:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{9}$$

Gaussian Mixture Model memiliki tiga parameter, yaitu rata-rata, ragam, dan bobot campuran dari semua komponen GMM. Parameter tersebut akan diestimasi dan dimaksimalkan menggunakan algoritme Expectation Maximization (EM) dengan dua tahap, yaitu Expectation Step (E-step) dan Maximization Step (M-Step).Algoritma EM digunakan untuk mengoptimalkan parameter GMM melalui dua tahap:

- E-step:

  Menghitung probabilitas bahwa data x_i terasuk dalam klaster C_k berdasarkan parameter saat ini:

$$Q(\theta^*, \theta^{(t)}) = \sum_{k=1}^{K} \sum_{i=1}^{N} [\ln \pi_k + \ln \mathcal{N}(x_i|\mu_k, \Sigma_k)] \cdot p(z_i = 1|x_i, \theta^{(t)}) \tag{10}$$

- M-step:

  Memperbarui parameter menggunakan *maximum likelihood*:

$$\pi_k^* = \frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \theta^{(t)})}{N} \tag{11}$$

$$\mu_k^* = \frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \theta^{(t)}) \cdot x_i}{\sum_{i=1}^{N} p(z_i = 1|x_i, \theta^{(t)})} \tag{12}$$

$$\Sigma_k^* = \frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \theta^{(t)})(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} p(z_i = 1|x_i, \theta^{(t)})} \tag{13}$$

Gaussian Mixture Model (GMM) merupakan pendekatan yang powerful dalam analisis data tidak berlabel, terutama karena kemampuannya untuk memodelkan distribusi kompleks melalui kombinasi linear distribusi Gaussian. Dengan algoritma EM, GMM secara iteratif mengestimasi parameter rata-rata, ragam, dan bobot campuran sambil memaksimalkan likelihood data. Kelebihan utama GMM terletak pada fleksibilitasnya (melalui pemilihan matriks kovarians) dan output probabilitas keanggotaan klaster. Namun, model ini juga memiliki tantangan, seperti sensitivitas terhadap inisialisasi dan komputasi yang kompleks untuk dataset besar.

2.1.4 Agglomerative Hierarchical Clustering

Algoritma klasterisasi hierarki adalah salah satu metode pembelajaran tanpa pengawasan (unsupervised learning) yang umum digunakan untuk membagi suatu dataset ke dalam kelompok-kelompok pada berbagai tingkat. Berbeda dengan K-Means, algoritma ini tidak memerlukan jumlah klaster yang ditentukan di awal. Sebaliknya, algoritma ini membangun struktur hierarkis dengan cara menghitung kemiripan atau jarak antar sampel. Proses pembentukan klaster direpresentasikan dalam bentuk struktur pohon yang disebut dendrogram. Dengan menghitung matriks jarak atau kemiripan yang mengukur kedekatan antara setiap pasangan klaster, dua klaster yang paling dekat menurut kriteria linkage yang dipilih akan digabung menjadi satu klaster. Dalam konteks jaringan sensor nirkabel (*Wireless Sensor Networks atau WSN*), algoritma ini dapat digunakan untuk mengelompokkan node sensor menjadi beberapa klaster terpisah guna memudahkan manajemen, agregasi data, dan transmisi informasi. Untuk penjelasan lebih lanjut mengenai langkah -langkah klasterisasi menggunakan *agglomerative hierarchical clustering* adalah sebagai berikut:

1. Inisialisasi

   Pada tahap awal, setiap node sensor dianggap sebagai satu kategori terpisah, yaitu:

$$C^0 = \{C_1^0, C_2^0, \ldots, C_n^0\} \tag{14}$$

   Superskrip 0 menyatakan iterasi awal. Dihitung matriks kedekatan antar kategori D^0, di mana elemen d_{ij}^0 menunjukkan jarak antara kategori i dan j.

2. Penggabungan Klaster Terdekat

   Temukan nilai terkecil dalam matriks D^t. Jika nilai tersebut menunjukkan jarak antara klaster C_i dan C_j, maka gabungkan kedua klaster tersebut menjadi satu. Kedekatan antar klaster setelah penggabungan dihitung menggunakan metode *single linkage* (menggunakan jarak terpendek antar dua elemen di klaster berbeda):

$$d(C_i, C_j) = \min_{x \in C_i,\, y \in C_j} d(x, y) \tag{15}$$

   dengan $d(x, y)$ adalah jarak Euclidean antara dua node.

3. Pembaruan Matriks

   Perbarui daftar klaster dan matriks kedekatan menjadi D^{t+1}, lalu kembali ke langkah 2.

4. Penghentian

   Tetapkan batas ambang jarak τ, dan jika nilai terkecil dalam matriks jarak saat iterasi melebihi τ, maka proses dihentikan. Klaster yang terbentuk menjadi hasil akhir dari klasterisasi.

   Agar algoritma ini efektif dalam sistem terdistribusi seperti *Wireless Sensor Network* (WSN), beberapa penyesuaian dilakukan:

- Hitung jarak maksimum dan minimum Euclidean antar node sensor.
- Misal terdapat dua node i dan j dengan m atribut (misalnya: longitude, latitude, jarak ke node pusat, jumlah tetangga, dan sisa energi), maka jarak Euclidean antar node adalah:

$$d(i,j) = \sqrt{\sum_{k=1}^{m}(x_i^k - x_j^k)^2} \tag{16}$$

- Untuk mengurangi komunikasi data serta kompleksitas ruang dan waktu, atribut dapat dibagi menjadi s subkategori, sehingga jarak total antar node dapat dihitung sebagai:

$$d(i,j) = \sum_{l=1}^{s} d_l(i,j) \tag{17}$$

  di mana $d_l(i,j)$ adalah jarak untuk subkategori ke-l.
- Jarak terpendek antar dua node dalam satu kategori didefinisikan sebagai:

$$d_{\min}(i,j) = \min \text{ path antara } i \text{ dan } j \tag{18}$$

- Batas atas (*upper bound*) jarak dalam setiap klaster dapat dihitung berdasarkan dendrogram. Total jalur terpendek antar atribut dalam satu kategori menjadi batas maksimum jarak antar dua node.

2.2 *Principal Component Analysis* (PCA)

Teknik reduksi dimensi semakin penting untuk meningkatkan efisiensi dan kinerja model *clustering*, terutama saat berhadapan dengan data berdimendi tinggi. Penelitian terbaru telah mengeksplorasi berbagai metode untuk mengurangi dimensi sebelum data dimasukan ke dalam pemodelan. Salah satu pendekatan umum adalah menggunakan metode linear, seperti *Principal Component Analysis* (PCA), teknik ini digunakan untuk mengekstrasi fitur utama (Surono et al.2023). PCA merupakan teknik statsitik yang fleksibel, digunakan untuk meringkas dataset yang berisi banyak baris dan variabel ke dalam unsur - unsur dasarnya yang disebut komponen utama (*principal component*). Komponen - komponen ini merupakan kombinasi linear optimal dari variabel asli yang mampu menangkap varians maksimum dari keseluruhan variabel. Dengan proses ini, PCA menyajikan pendekatan aproksimasi terhadap data asli dengan hanya menyoroti beberapa komponen yang paling signifikan (Greenacre et al, 2022) Langkah -langkah dalam PCA meliputi:

1. Standarisasi variabel

   Tahap awal adalah menstandarkan variabel menggunakan persamaan berikut:

$$z = \frac{x - \mu}{\sigma} \tag{19}$$

2. Menghitung Matriks Korelasi (C)

   Matriks korelasi dihitung dari dataset yang telah distandarisasi:

$$C = \frac{Z^T Z}{n} \tag{20}$$

   Dimana Z adalah dataset yang telah distandaisasi dan Z^T adalh transposenya.

3. Menentukan Nilai Eigen dan Eigen Vektor

   Selanjutnya, cari nilai eigen (λ) dan vektor eigen (V) dari matriks korelasi C.

$$CV = \lambda V \tag{21}$$

4. Mengurutkan Vektor Eigen

   Urutkan vektor - vekrtor eigen berdasrakan nilai eigen tertinggi ke terendah.

5. Memilih dan Membentuk dataset Baru

   Pilih sejumlah vektor eigen teratas, lalu bangun dataset baru berdasarkan representasi baru ini.

   Setelah dataset dibentuk dengan vektor eigen yang dipilih, langkah akhir dari PCA adalah menggunakan data yang telah ditransformasikan ini untuk analisis lanjutan, seperti *clustering* maupun visualisasi.

2.3 Metrik Evaluasi Model *Clustering*

Kualitas dari klasterisasi potensial dievaluasi menggunakan indeks validitas klaster (*Cluster Validity Index*/CVI), yang juga berperan dalam menentukan jumlah klaster optimal. Perlu dicatat bahwa indeks validitas klaster seperti Davies-Bouldin Index, Silhouette Score, atau Dunn Index digunakan tidak hanya untuk menilai validitas jumlah klaster, tetapi juga untuk memberikan wawasan mengenai kualitas klaster itu sendiri. Berbagai indeks validitas klaster internal yang digunakan dalam algoritma klasterisasi adalah sebagai berikut:

- Silhoutte Score

  Silhouette Score mengukur seberapa mirip suatu objek (data) dengan klaster tempatnya berada dibandingkan dengan klaster terdekat lainnya. Nilai skor berkisar antara -1 hingga +1. Dapat digunakan untuk mengevaluasi hasil klasterisasi secara independen dari algoritmanya. Berguna untuk membandingkan beberapa metode klasterisasi.

- Davies Bouldin Index (DBI)

  Indeks ini mengukur seberapa mirip sebuah klaster dengan klaster lainnya berdasarkan rasio antara jarak dalam klaster dan jarak antar klaster.Nilai DBI yang lebih rendah menunjukkan klaster yang lebih baik (kompak dan terpisah). Kelebihannya mudah dihitung dan populer digunakan serta efisien untuk klaster bulat dan data terdistribusi seragam.

- Dunn Index

  Mengukur rasio antara jarak minimum antar klaster dengan jarak maksimum dalam klaster. Artinya, seberapa jauh klaster satu sama lain dan seberapa padat tiap klaster. Nilai Dunn Index yang lebih tinggi menunjukkan klaster yang lebih baik (lebih terpisah dan kompak). Dapat mendeteksi klaster berbentuk arbitrer (tidak beraturan).Cocok untuk dataset dengan kepadatan tinggi.

- Calinski-Harabaz Index

  Mengukur rasio antara variasi antar klaster dengan variasi dalam klaster. Semakin besar variasi antar klaster dibanding dalam klaster, semakin baik.Nilai indeks yang lebih tinggi berarti pembagian klaster lebih efektif.Cepat dan efisien, cocok untuk evaluasi awal klaster. Tidak bergantung langsung pada metrik jarak tertentu.

BAB III
Metodologi

3.1 Data

Data yang digunakan dalam *project* ini bersumber dari portal resmi milik Pemerintah Kota London, yaitu London Datastore. Dataset yang diunduh berjudul **MPS LSOA Level Crime (Historical)** dan mencakup data kejahatan yang tercatat di wilayah London dari tanggal 1 April 2010 hingga 31 Maret 2023. Data ini dikumpulkan oleh Metropolitan Police Service (MPS) dan disusun berdasarkan tingkat wilayah administratif terkecil di Inggris, yaitu Lower Super Output Area (LSOA).

Pada data asli, setiap baris mewakili satu kombinasi unik antara LSOA, kategori utama kejahatan (*major category*), dan subkategori kejahatan (*minor category*), serta jumlah kasus kejahatan yang terjadi setiap bulannya. Data mentah memiliki total 113072 observasi dengan 160 kolom. Untuk keperluan analisis clustering, data dimodifikasi dengan melakukan agregasi jumlah kasus kejahatan berdasarkan kategori utama (*major category*) untuk setiap LSOA.

	LSOA Code	LSOA Name	Borough	Major Category	Minor Category	201903	201904	201905	201906	201907	...	202205	202206	202207	202208	202209	202210	202211	202212	202301	202302
0	E01000006	Barking and Dagenham 016A	E09000002	ARSON AND CRIMINAL DAMAGE	ARSON	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	E01000006	Barking and Dagenham 016A	E09000002	ARSON AND CRIMINAL DAMAGE	CRIMINAL DAMAGE	1	2	0	1	0	...	0	0	2	1	0	0	0	0	0	0
2	E01000006	Barking and Dagenham 016A	E09000002	BURGLARY	BURGLARY BUSINESS AND COMMUNITY	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	E01000006	Barking and Dagenham 016A	E09000002	BURGLARY	BURGLARY IN A DWELLING	1	0	3	1	0	...	0	0	0	0	0	0	2	0	0	0
4	E01000006	Barking and Dagenham 016A	E09000002	DRUG OFFENCES	POSSESSION OF DRUGS	2	2	0	0	0	...	2	0	0	1	0	4	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
113067	E01035722	Westminster 024G	E09000033	VEHICLE OFFENCES	INTERFERING WITH A MOTOR VEHICLE	0	0	0	1	0	...	0	0	1	0	0	0	0	0	0	0
113068	E01035722	Westminster 024G	E09000033	VEHICLE OFFENCES	THEFT FROM A VEHICLE	2	0	5	1	0	...	0	1	1	0	3	1	1	1	1	0
113069	E01035722	Westminster 024G	E09000033	VEHICLE OFFENCES	THEFT OR UNAUTH TAKING OF A MOTOR VEH	0	0	0	0	0	...	1	0	0	0	0	1	0	0	0	1
113070	E01035722	Westminster 024G	E09000033	VIOLENCE AGAINST THE PERSON	VIOLENCE WITH INJURY	0	0	0	1	1	...	1	1	2	1	0	1	2	0	3	0
113071	E01035722	Westminster 024G	E09000033	VIOLENCE AGAINST THE PERSON	VIOLENCE WITHOUT INJURY	0	2	2	2	2	...	1	2	3	3	2	3	2	2	2	0

113072 rows × 160 columns

Gambar 1: *Preview* Data Asli

3.2 Modifikasi Data

Sebelum dilakukan analisis *clustering*, data asli perlu dimodifikasi agar sesuai dengan tujuan analisis, yaitu mengelompokkan wilayah berdasarkan total kasus kejahatan menurut kategori utama (*major category*). Data asli berbentuk *long format* dengan setiap baris mewakili kombinasi unik dari `LSOA Code`, `Major Category`, `Minor Category`, dan jumlah kasus kejahatan per bulan. Selanjutnya, dilakukan beberapa langkah transformasi sebagai berikut.

1. **Menghitung total kasus per baris**
   Kolom-kolom bulan yang mencatat jumlah kasus di setiap periode dikombinasikan untuk menghasilkan satu kolom baru, yaitu `Total`, yang merepresentasikan total kejadian kejahatan per baris data.

2. **Mengagregasi total kasus berdasarkan wilayah dan kategori utama**
   Data kemudian dikelompokkan berdasarkan `LSOA Code` dan `Major Category`, kemudian dijumlahkan total kasusnya untuk masing-masing kombinasi. Ini menghasilkan jumlah total kejadian kejahatan dari setiap kategori besar untuk masing-masing LSOA.

3. **Transformasi ke bentuk *wide format***
   Data yang sudah digabungkan dipivot agar setiap baris mewakili satu LSOA, dan setiap kolom mewakili satu kategori kejahatan utama. Nilai pada sel merupakan total kasus kejahatan dalam kategori tersebut. Kolom yang tidak memiliki kasus di suatu LSOA diisi dengan nol (`0`).

4. **Mengurutkan kolom agar lebih rapi**
   Kolom `LSOA Code` dan `Borough` diletakkan di awal tabel, diikuti oleh kolom-kolom kategori kejahatan untuk memudahkan pembacaan dan analisis lebih lanjut.

Hasil akhir modifikasi berupa dataset yang berisi total kasus kejahatan untuk masing-masing kategori kejahatan utama pada setiap LSOA. Data hasil modifikasi hanya memiliki 4988 observasi (mewakili 4988 LSOA) dan 12 kolom.

Berikut adalah penjelasan mengenai variabel-variabel dalam data hasil modifikasi.

- **LSOA Code** berisi kode unik yang merepresentasikan wilayah LSOA.
- **Borough** berisi kode dari borough atau wilayah administratif tempat LSOA tersebut berada (contoh: `E09000002` untuk Barking and Dagenham).
- **ARSON AND CRIMINAL DAMAGE** adalah total kasus kejahatan yang termasuk dalam kategori pembakaran dan perusakan properti.
- **BURGLARY** adalah total kasus pembobolan, baik rumah tinggal maupun tempat usaha.
- **DRUG OFFENCES** adalah jumlah kasus terkait pelanggaran narkoba, termasuk kepemilikan dan penyelundupan.
- **MISCELLANEOUS CRIMES AGAINST SOCIETY** mencakup kejahatan lain terhadap masyarakat seperti penipuan, pemalsuan, atau pelanggaran peradilan.
- **POSSESSION OF WEAPONS** adalah kasus kepemilikan senjata tajam atau senjata api ilegal.
- **PUBLIC ORDER OFFENCES** merupakan kejahatan yang mengganggu ketertiban umum seperti kerusuhan atau ujaran kebencian.
- **ROBBERY** adalah total kasus perampokan properti milik pribadi maupun bisnis.
- **THEFT** mencakup kasus pencurian termasuk pencurian sepeda, dari toko, atau dari individu.
- **VEHICLE OFFENCES** adalah pelanggaran terkait kendaraan bermotor seperti pencurian kendaraan atau perusakan.
- **VIOLENCE AGAINST THE PERSON** merupakan kejahatan terhadap individu seperti penyerangan atau kekerasan fisik.

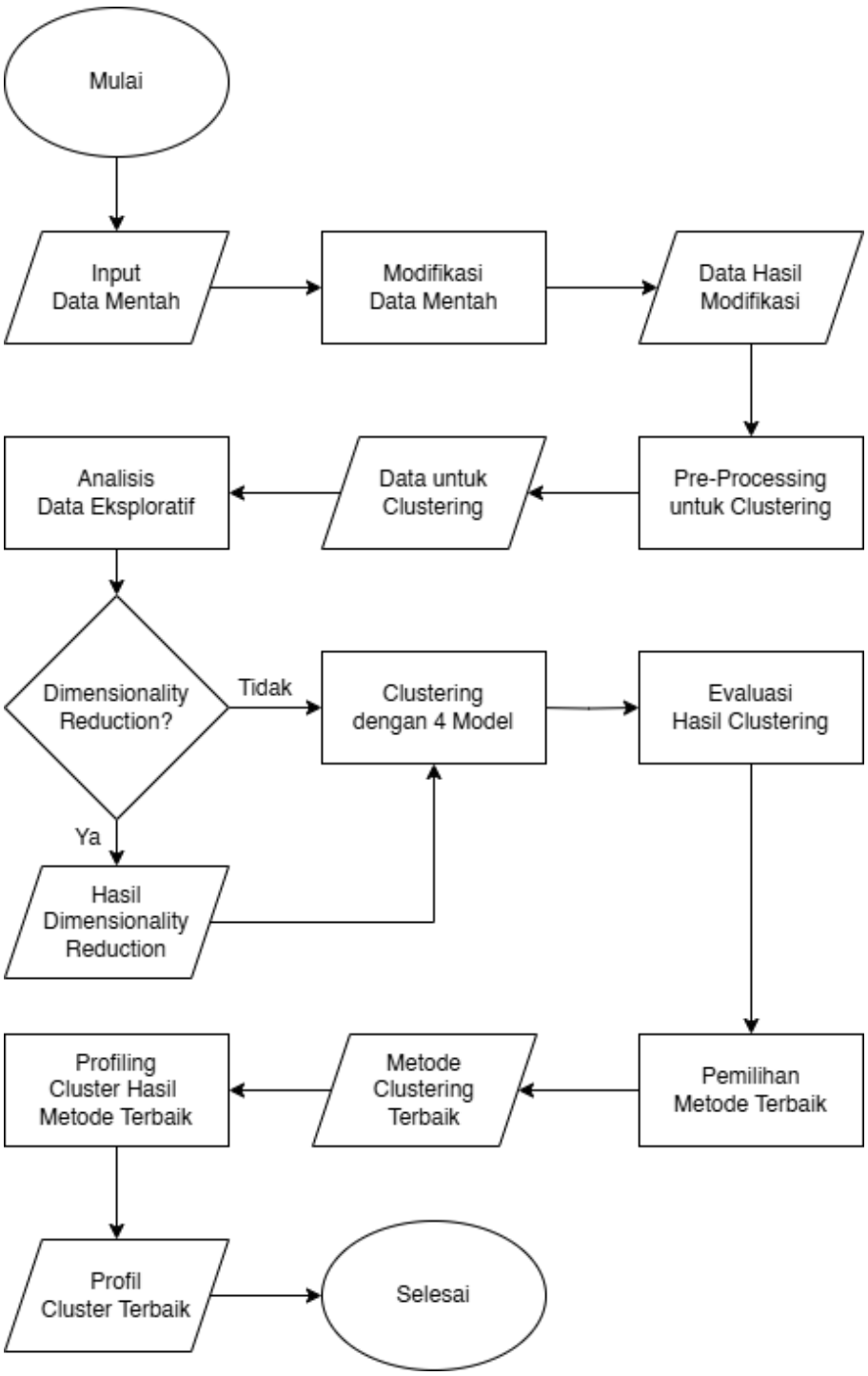
	LSOA Code	Borough	ARSON AND CRIMINAL DAMAGE	BURGLARY	DRUG OFFENCES	MISCELLANEOUS CRIMES AGAINST SOCIETY	POSSESSION OF WEAPONS	PUBLIC ORDER OFFENCES	ROBBERY	THEFT	VEHICLE OFFENCES	VIOLENCE AGAINST THE PERSON
0	E01000006	E09000002	372.0	568.0	480.0	96.0	24.0	156.0	264.0	440.0	1152.0	1412.0
1	E01000007	E09000002	1468.0	688.0	2584.0	296.0	280.0	1456.0	1328.0	5224.0	1152.0	6392.0
2	E01000008	E09000002	848.0	508.0	840.0	184.0	80.0	372.0	324.0	2296.0	1676.0	2392.0
3	E01000009	E09000002	1092.0	732.0	1944.0	148.0	112.0	680.0	1004.0	2480.0	1220.0	3844.0
4	E01000011	E09000002	508.0	420.0	364.0	104.0	32.0	160.0	252.0	528.0	676.0	1988.0
...	...	...	...	...	...	...	...	...	...	...	...	...
4983	E01035718	E09000033	1616.0	2364.0	9136.0	344.0	472.0	2236.0	3364.0	32224.0	5468.0	7624.0
4984	E01035719	E09000033	288.0	280.0	184.0	36.0	32.0	268.0	104.0	1272.0	224.0	996.0
4985	E01035720	E09000033	408.0	340.0	352.0	36.0	32.0	252.0	88.0	2348.0	380.0	912.0
4986	E01035721	E09000033	1380.0	908.0	8748.0	312.0	360.0	1772.0	604.0	14048.0	1384.0	4920.0
4987	E01035722	E09000033	400.0	720.0	440.0	40.0	100.0	308.0	308.0	1880.0	868.0	1460.0

4988 rows × 12 columns

Gambar 2: *Preview* Data Hasil Modifikasi

3.3 Skema Analisis

Alur analisis digambarkan dalam *flowchart* pada Gambar 3.



Gambar 3: Skema Analisis

Pada setiap bagian *clustering*, dilakukan penentuan jumlah *cluster* terbaik dengan metode *Elbow* (untuk K-Means) atau *Silhouette* (untuk K-Median, Agglomerative, dan GMM) serta visualisasi hasil *clustering* dengan PCA.

BAB IV
Hasil Analisis

4.1 Pra-pemrosesan Data

Sebelum dilakukan analisis *clustering*, data terlebih dahulu diproses dan dinormalisasi agar sesuai dengan kebutuhan analisis dan untuk menghindari bias akibat skala atau distribusi variabel. Langkah-langkah pra-pemrosesan yang dilakukan adalah sebagai berikut:

1. **Menghitung total kasus kejahatan per wilayah (Total Case)**
   Dibuat sebuah variabel baru bernama `Total Case` yang merepresentasikan jumlah total seluruh kasus kejahatan di tiap LSOA. Nilai ini diperoleh dengan menjumlahkan seluruh kasus dari berbagai kategori kejahatan yang terjadi di wilayah tersebut.

2. **Mengubah nilai absolut menjadi proporsi**
   Untuk setiap kategori kejahatan, nilai absolut jumlah kasus diubah menjadi proporsi terhadap total kasus di masing-masing LSOA. Transformasi ini dilakukan untuk menyesuaikan analisis dengan perbandingan karakteristik antar wilayah, terutama karena setiap LSOA dapat memiliki ukuran populasi dan tingkat kejahatan yang berbeda. Dengan menggunakan proporsi, kita dapat memahami distribusi jenis kejahatan di suatu wilayah tanpa dipengaruhi oleh jumlah total kejahatan yang mungkin sangat besar atau kecil. Langkah ini juga penting untuk menghindari dominasi variabel dengan nilai absolut besar dalam proses *clustering*.

3. **Menghapus variabel non-relevan untuk clustering**
   Variabel `Borough` dan `Total Case` dihapus dari data karena tidak relevan untuk proses *clustering*. Variabel `Borough` bersifat kategorikal dan tidak merepresentasikan karakteristik kejahatan, sementara `Total Case` sudah diwakili secara implisit dalam proporsi tiap jenis kejahatan yang telah dihitung sebelumnya.

4. **Menjadikan `LSOA Code` sebagai indeks**
   Variabel `LSOA Code` digunakan sebagai indeks pada data agar setiap baris tetap dapat ditelusuri ke wilayah asalnya tanpa memengaruhi proses perhitungan atau analisis statistik.

5. **Standarisasi data**
   Seluruh variabel numerik yang akan digunakan dalam *clustering* distandarisasi agar memiliki rata-rata nol dan standar deviasi satu. Proses ini penting untuk memastikan bahwa semua variabel memiliki kontribusi yang seimbang dalam pengukuran jarak antar objek pada algoritma *clustering*, serta mencegah variabel dengan skala besar mendominasi hasil akhir.

Dengan tahapan ini, data menjadi optimal untuk dianalisis lebih lanjut menggunakan metode *clustering* yang dapat mengelompokkan wilayah berdasarkan pola distribusi jenis kejahatan yang terjadi. Data hasil pra-pemrosesan ditunjukkan pada Gambar 4.

LSOA Code	ARSON AND CRIMINAL DAMAGE	BURGLARY	DRUG OFFENCES	MISCELLANEOUS CRIMES AGAINST SOCIETY	POSSESSION OF WEAPONS	PUBLIC ORDER OFFENCES	ROBBERY	THEFT	VEHICLE OFFENCES	VIOLENCE AGAINST THE PERSON
E01000006	0.074940	0.114424	0.096696	0.019339	0.004835	0.031426	0.053183	0.088638	0.232071	0.284448
E01000007	0.070347	0.032969	0.123826	0.014184	0.013418	0.069772	0.063638	0.250335	0.055204	0.306306
E01000008	0.089076	0.053361	0.088235	0.019328	0.008403	0.039076	0.034034	0.241176	0.176050	0.251261
E01000009	0.082378	0.055220	0.146651	0.011165	0.008449	0.051298	0.075739	0.187085	0.092034	0.289982
E01000011	0.100954	0.083466	0.072337	0.020668	0.006359	0.031797	0.050079	0.104928	0.134340	0.395072
...	...	...	...	...	...	...	...	...	...	...
E01035718	0.024920	0.036454	0.140883	0.005305	0.007279	0.034481	0.051875	0.496916	0.084320	0.117567
E01035719	0.078176	0.076004	0.049946	0.009772	0.008686	0.072747	0.028230	0.345277	0.060803	0.270358
E01035720	0.079254	0.066045	0.068376	0.006993	0.006216	0.048951	0.017094	0.456099	0.073815	0.177156
E01035721	0.040074	0.026368	0.254036	0.009060	0.010454	0.051458	0.017540	0.407945	0.040190	0.142874
E01035722	0.061312	0.110362	0.067443	0.006131	0.015328	0.047210	0.047210	0.288167	0.133047	0.223789

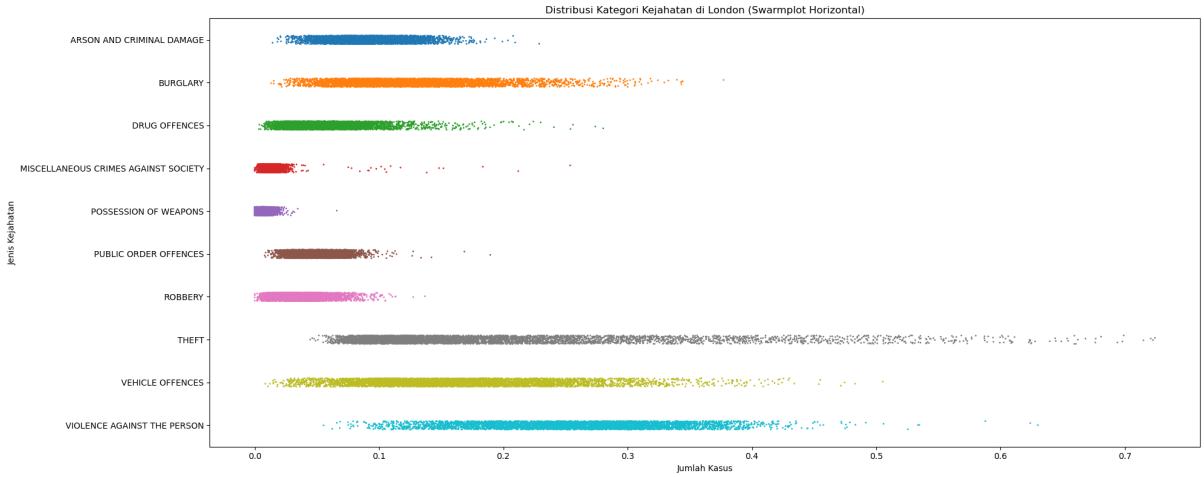
4988 rows × 10 columns

Gambar 4: *Preview* Data Hasil Pra-Pemrosesan

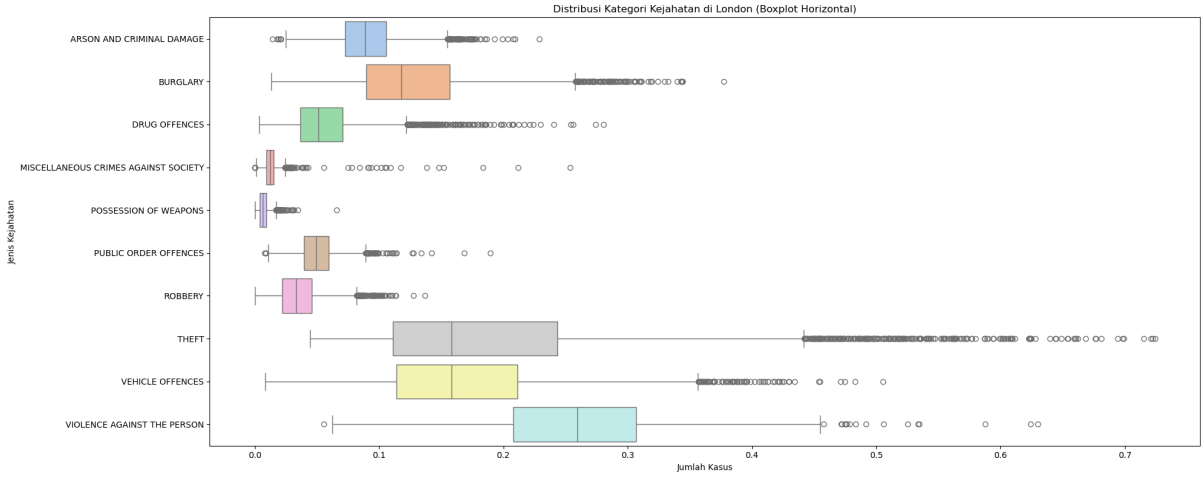
4.2 Analisis Data Eksploratif

Analisis data eksploratif bertujuan untuk memahami struktur, pola, dan karakteristik umum dari data sebelum dilakukan pemodelan lebih lanjut. Tahapan ini penting untuk memperoleh gambaran awal mengenai distribusi data, hubungan antar variabel, serta potensi keberadaan *outlier* atau pola yang tidak biasa. Analisis eksploratif difokuskan pada pemeriksaan variasi jumlah dan proporsi kejahatan menurut kategori utama kejahatan di berbagai wilayah LSOA dan hubungannya.

4.2.1 Distribusi Proporsi Jenis Kejahatan



Gambar 5: *Swarmplot* Distribusi Proporsi Jenis Kriminalitas



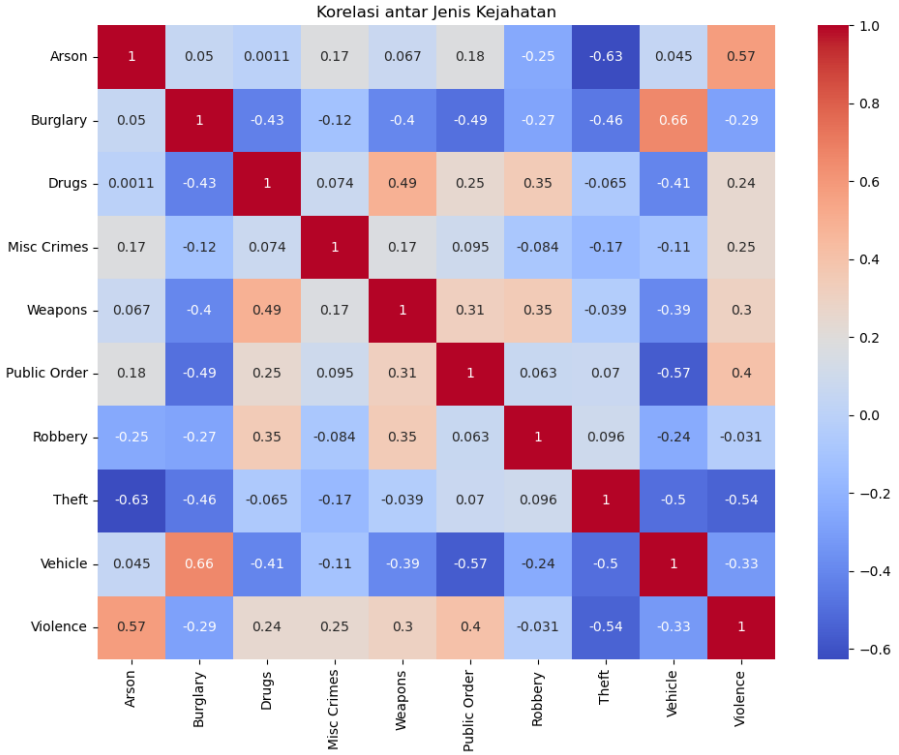
Gambar 6: *Boxplot* Distribusi Proporsi Jenis Kriminalitas

Berdasarkan visualisasi *swarmplot* pada Gambar 5 dan *boxplot* pada Gambar 6, kategori kejahatan dengan proporsi kasus terendah di wilayah-wilayah London adalah **Possession of Weapons** dan **Miscellaneous Crimes Against Society**. Hal ini terlihat dari titik-titik pada *swarmplot* yang terkonsentrasi di sebelah kiri (sekitar 0 hingga 0,01), menunjukkan bahwa sebagian besar LSOA (*Lower Layer Super Output Area*) memiliki proporsi yang sangat kecil untuk dua jenis kejahatan ini. Selain itu, *boxplot* memperkuat temuan tersebut dengan menunjukkan median yang sangat rendah, distribusi yang sempit, dan adanya sedikit *outlier*. Artinya, kejahatan seperti kepemilikan senjata dan pelanggaran sosial lainnya memang relatif jarang terjadi dan penyebarannya tidak merata di seluruh wilayah London.

Sebaliknya, kejahatan dengan proporsi kasus tertinggi adalah **Theft**, diikuti oleh **Vehicle Offences** dan **Violence Against the Person**. Pada *swarmplot*, ketiga kategori ini menunjukkan persebaran titik yang lebih luas dan menjangkau hingga lebih dari 0,7 untuk *Theft*, mengindikasikan bahwa beberapa

wilayah memiliki konsentrasi kasus pencurian yang sangat tinggi. Sementara itu, *boxplot* menunjukkan median yang tinggi dan rentang interkuartil yang lebar, terutama pada *Theft*, yang menandakan bahwa pencurian merupakan jenis kejahatan yang umum terjadi di banyak area dengan variasi yang cukup besar antar wilayah. Kategori *Vehicle Offences* dan *Violence Against the Person* juga menunjukkan pola yang serupa, meskipun dengan sebaran yang sedikit lebih rendah dibandingkan *Theft*.

4.2.2 Korelasi antarjenis Kejahatan



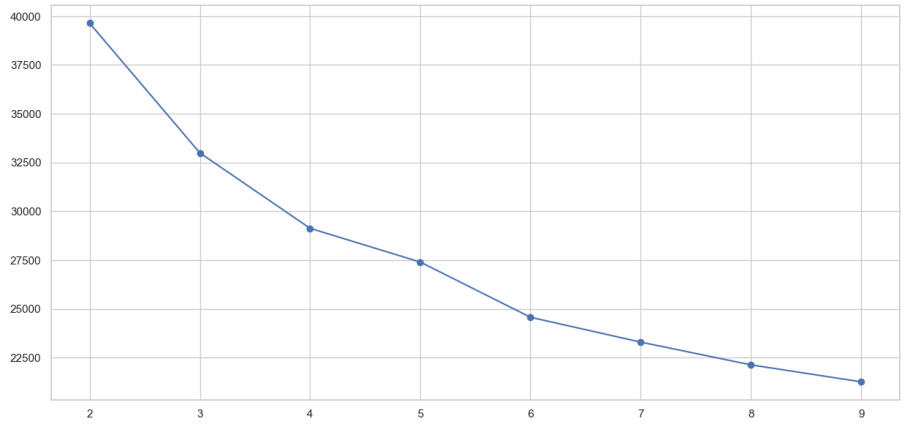
Gambar 7: *Heatmap* Korelasi antarjenis Kriminalitas

Berdasarkan visualisasi *heatmap* pada Gambar 7, diketahui bahwa beberapa jenis kejahatan cenderung terjadi bersamaan di wilayah yang sama, yang ditunjukkan oleh nilai korelasi positif yang tinggi. Misalnya, **Burglary** dan **Vehicle Offences** memiliki korelasi sebesar 0,66, yang merupakan salah satu nilai korelasi tertinggi, menandakan bahwa area dengan banyak kasus pembobolan rumah biasanya juga memiliki banyak kasus pencurian kendaraan. Selain itu, **Arson** (pembakaran) dan **Violence Against the Person** juga menunjukkan korelasi positif yang cukup tinggi ($r = 0,57$), menunjukkan bahwa daerah dengan banyak insiden pembakaran juga cenderung memiliki banyak kasus kekerasan terhadap individu.

Sebaliknya, terdapat pula pasangan jenis kejahatan yang cenderung tidak terjadi bersamaan, sebagaimana ditunjukkan oleh nilai korelasi negatif. Sebagai contoh, **Theft** memiliki korelasi negatif kuat dengan **Arson** ($r = -0,63$) dan **Violence Against the Person** ($r = -0,54$), menunjukkan bahwa wilayah dengan banyak kasus pencurian justru cenderung memiliki lebih sedikit kasus pembakaran atau kekerasan terhadap orang. Selain itu, **Public Order Offences** juga menunjukkan korelasi negatif dengan **Vehicle Offences** ($r = -0,57$), yang menandakan ketidakterkaitan antara pelanggaran ketertiban umum dan pencurian kendaraan.

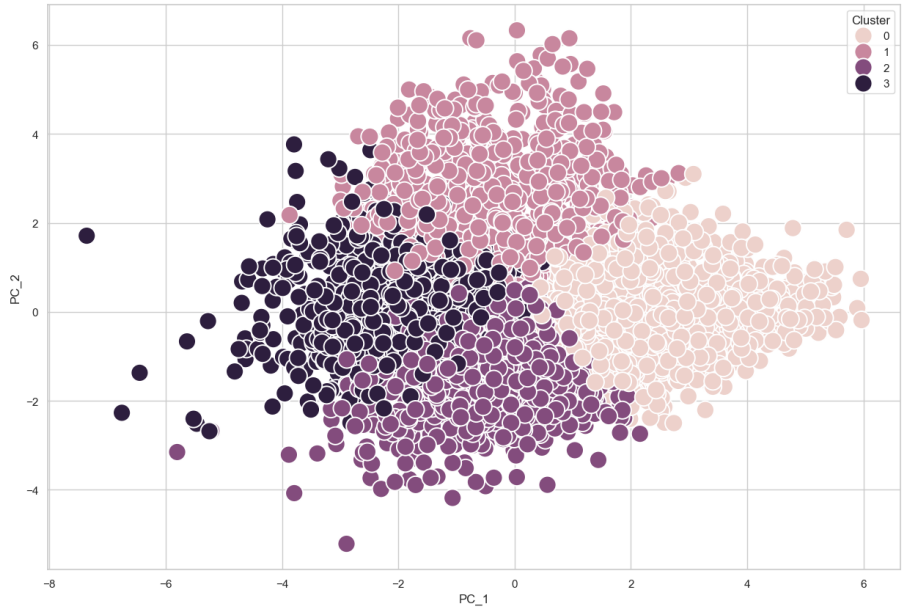
4.3 *Clustering* tanpa *Dimensionality Reduction*

4.3.1 *Clustering* dengan K-Means



Gambar 8: Penentuan Jumlah *Cluster* K-Means

Dengan *Elbow Method*, berdasarkan grafik nilai *inertia* pada Gambar 8, diketahui titik yang diawali penurunan tajam dan dilanjutkan dengan penurunan landai adalah titik 4. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 4 untuk K-Means. Selanjutnya, dilakukan *clustering* menggunakan metode K-Means untuk 4 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 9.



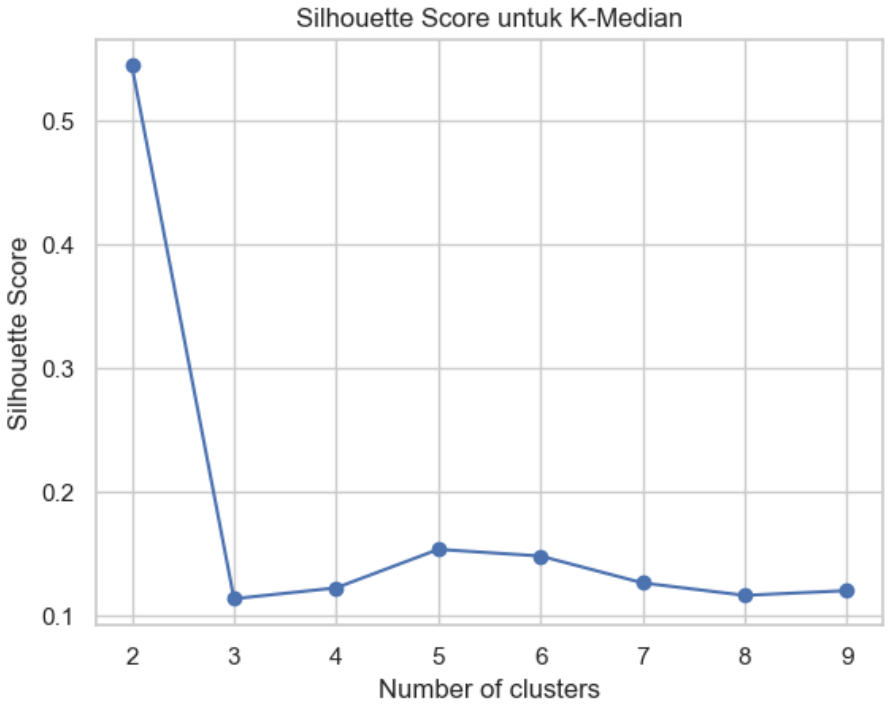
Gambar 9: Visualisasi *Cluster* K-Means

Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 1320 untuk *cluster* 0;
- 806 untuk *cluster* 1;
- 1557 untuk *cluster* 2; dan
- 1305 untuk *cluster* 3.

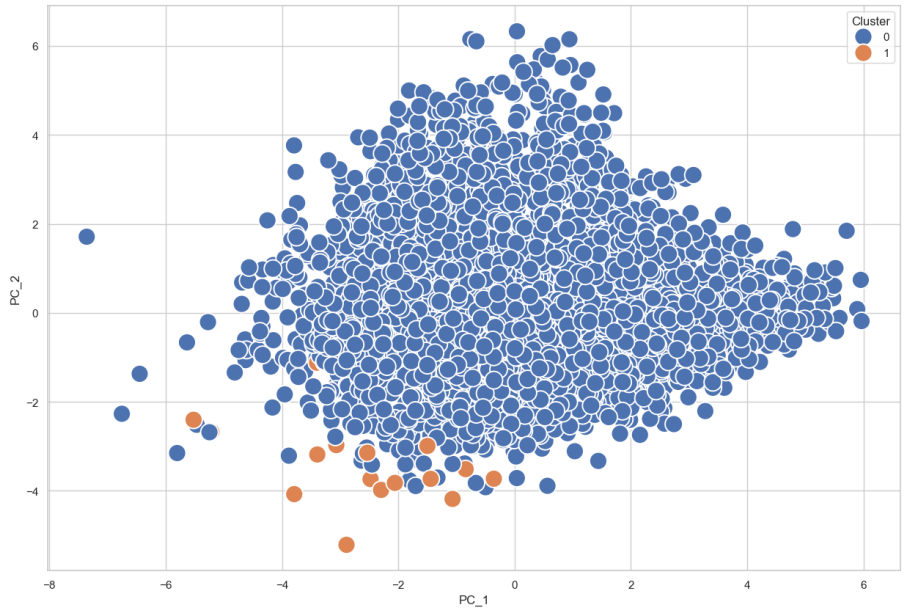
Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode K-Means relatif seimbang dan tidak terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.3.2 *Clustering* dengan K-Median



Gambar 10: Penentuan Jumlah *Cluster* K-Median

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 10, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 2. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 2 untuk K-Median. Selanjutnya, dilakukan *clustering* menggunakan metode K-Median untuk 2 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 11.



Gambar 11: Visualisasi *Cluster* K-Median

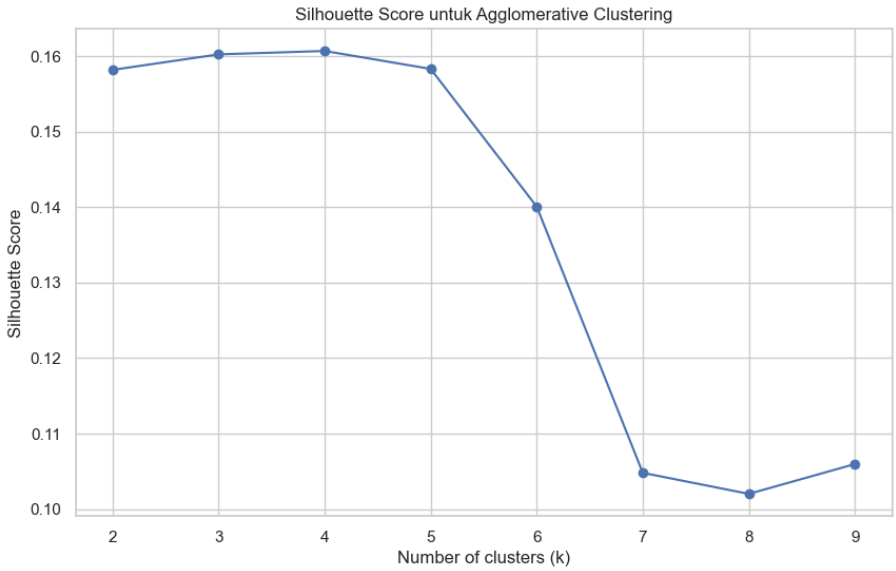
Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 4969 untuk *cluster* 0; dan

- 19 untuk *cluster* 1.

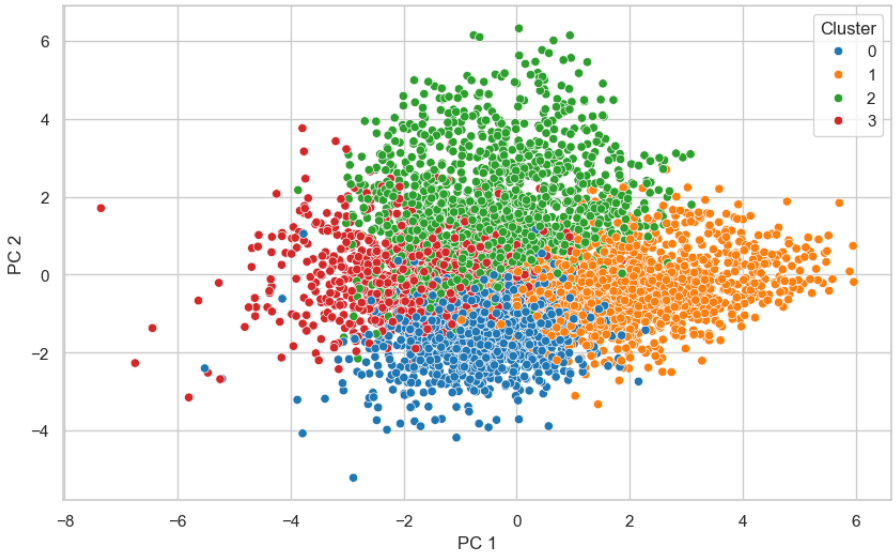
Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode K-Median sangat timpang dan terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.3.3 *Clustering* dengan Agglomerative



Gambar 12: Penentuan Jumlah *Cluster* Agglomerative

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 12, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 4. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 4 untuk Agglomerative. Selanjutnya, dilakukan *clustering* menggunakan metode Agglomerative untuk 4 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 13.



Gambar 13: Visualisasi *Cluster* Agglomerative

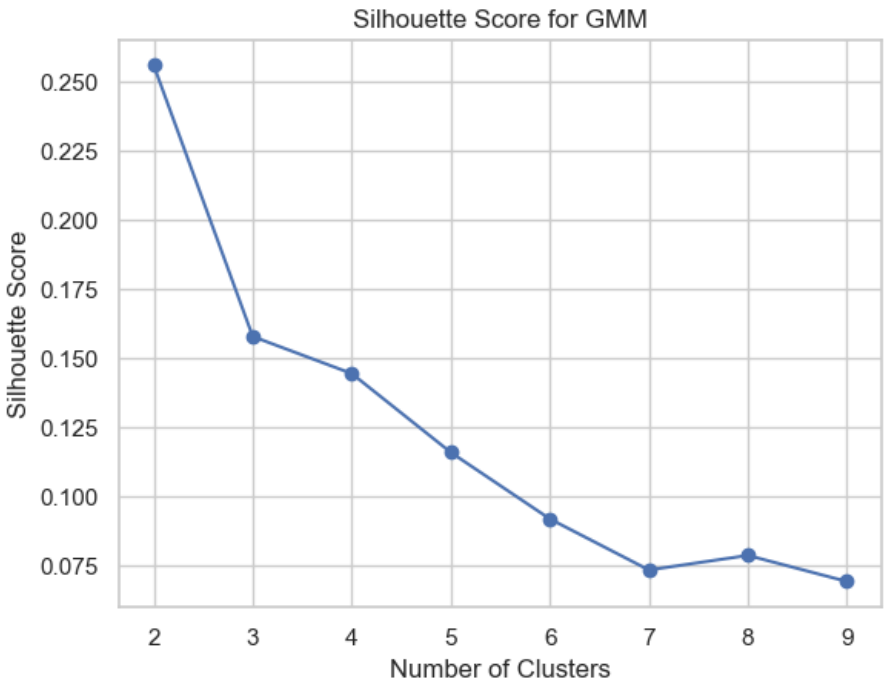
Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 1160 untuk *cluster* 0;
- 1427 untuk *cluster* 1;

- 1484 untuk *cluster* 2; dan
- 917 untuk *cluster* 3.

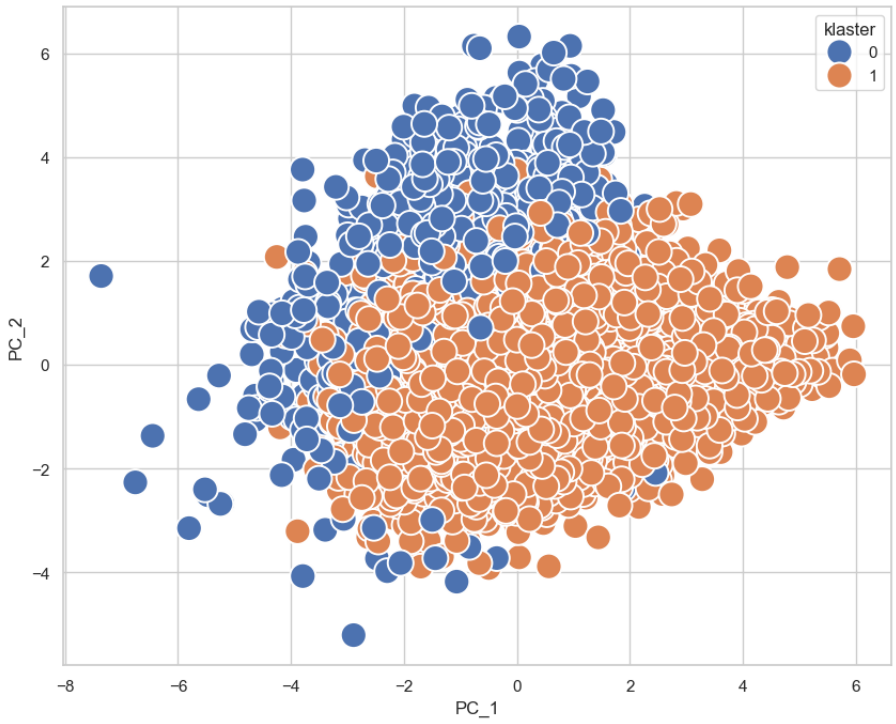
Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode Agglomerative relatif seimbang dan tidak terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.3.4 *Clustering* dengan Gaussian Mixture Model (GMM)



Gambar 14: Penentuan Jumlah *Cluster* GMM

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 14, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 2. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 2 untuk GMM. Selanjutnya, dilakukan *clustering* menggunakan metode GMM untuk 2 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 15.



Gambar 15: Visualisasi *Cluster* GMM

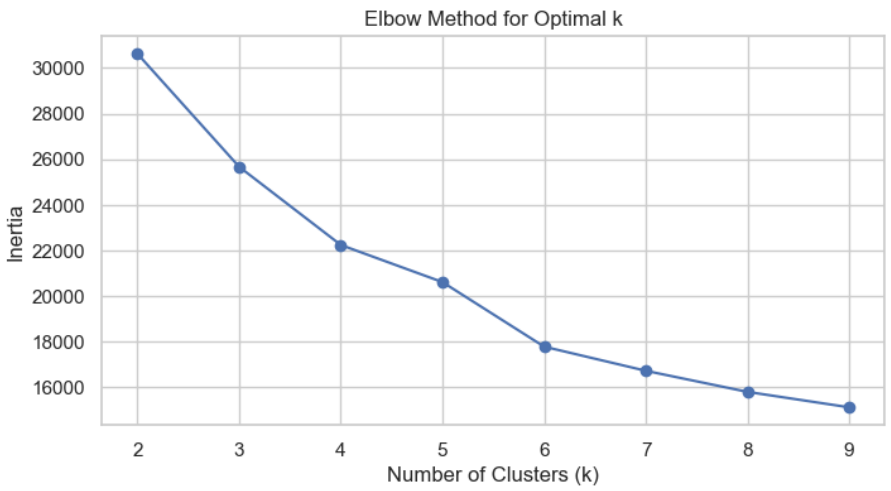
Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 815 untuk *cluster* 0; dan
- 4173 untuk *cluster* 1.

Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode GMM sangat timpang dan terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.4 *Clustering* dengan *Dimensionality Reduction*

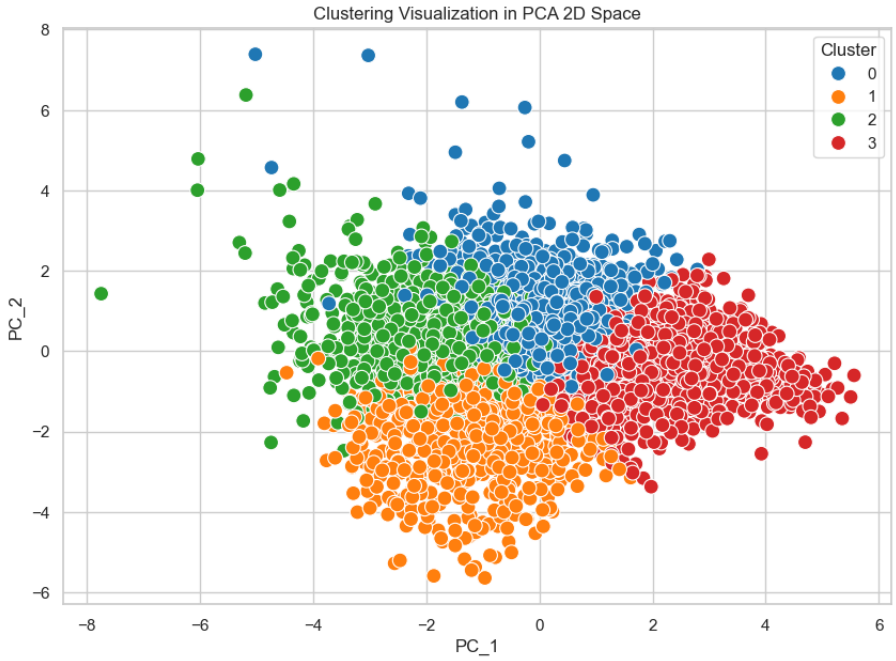
4.4.1 *Clustering* dengan K-Means



Gambar 16: Penentuan Jumlah *Cluster* K-Means

Dengan *Elbow Method*, berdasarkan grafik nilai *inertia* pada Gambar 16, diketahui titik yang diawali penurunan tajam dan dilanjutkan dengan penurunan landai adalah titik 4. Dengan demikian, ditetapkan

jumlah *cluster* optimal adalah 4 untuk K-Means. Selanjutnya, dilakukan *clustering* menggunakan metode K-Means untuk 4 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 17.



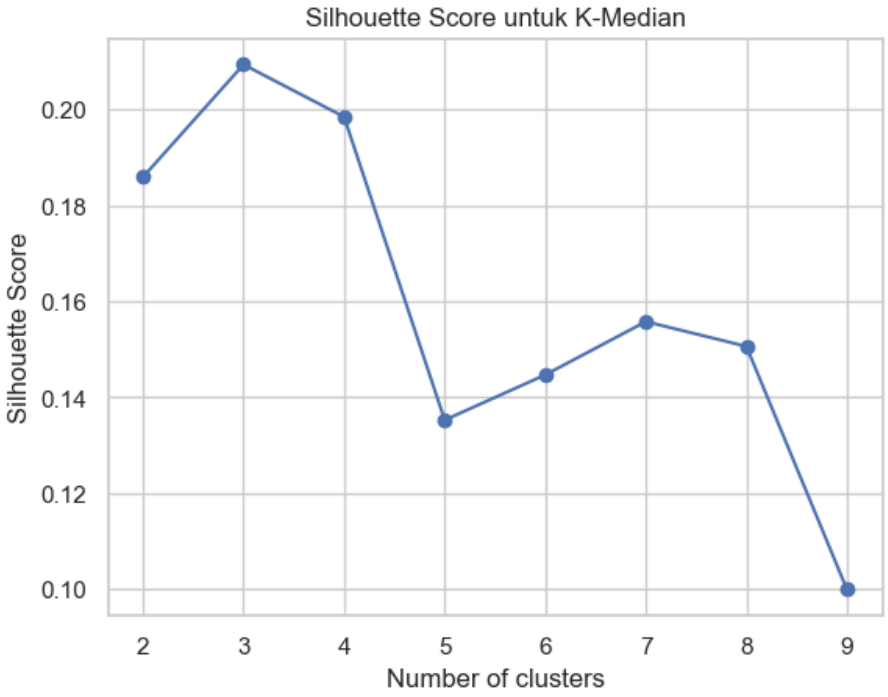
Gambar 17: Visualisasi *Cluster* K-Means

Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 1532 untuk *cluster* 0;
- 787 untuk *cluster* 1;
- 1275 untuk *cluster* 2; dan
- 1394 untuk *cluster* 3.

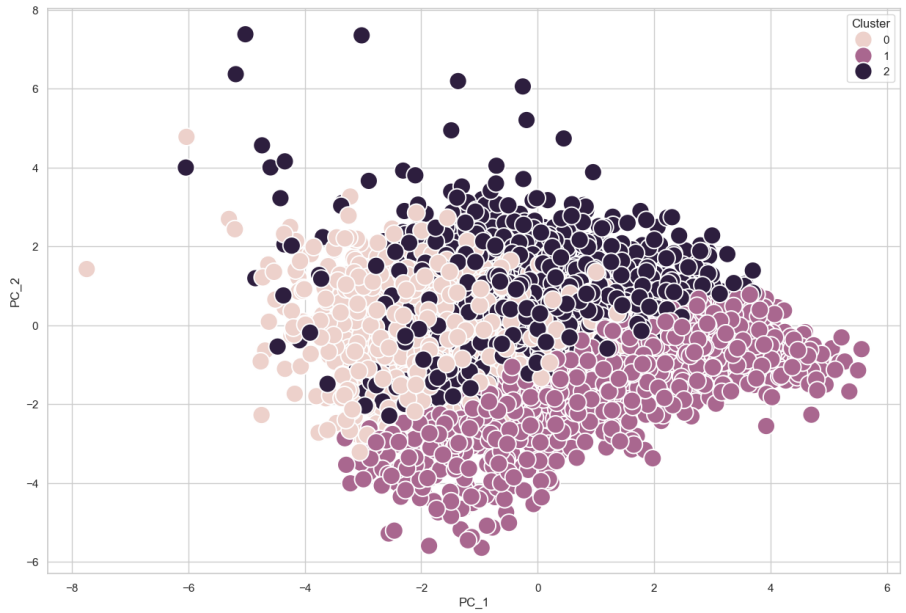
Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode K-Means relatif seimbang dan tidak terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.4.2 *Clustering* dengan K-Median



Gambar 18: Penentuan jumlah *Cluster* K-Median

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 18, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 5. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 5 untuk K-Medians. Selanjutnya dilakukan *clustering* menggunakan 5 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 19



Gambar 19: Visualisasi *cluster* K-Median

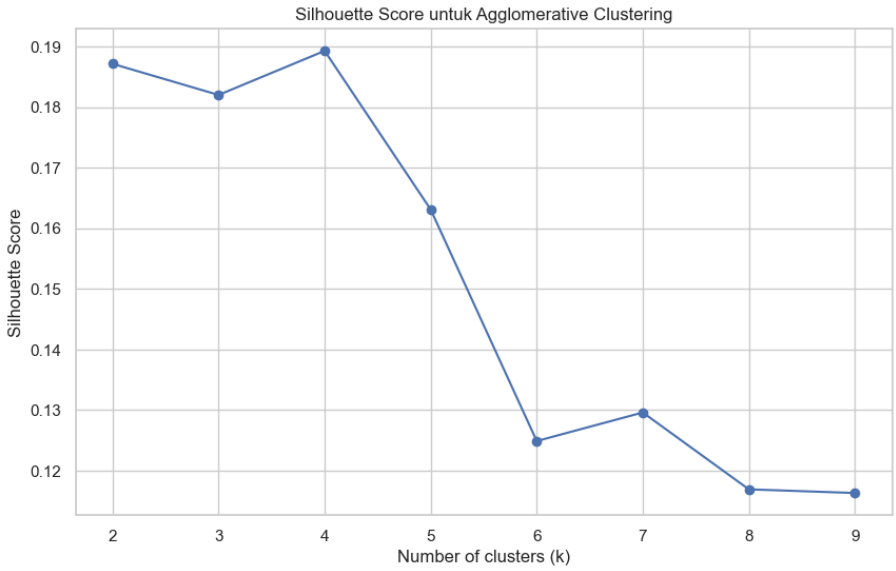
Berdasarkan hasil *clustering* dengan metode K-Median, diperoleh jumlah anggota *cluster* sebanyak

- 1555 untuk *cluster* 0;

- 1305 untuk *cluster* 1; dan
- 2128 untuk *cluster* 2.

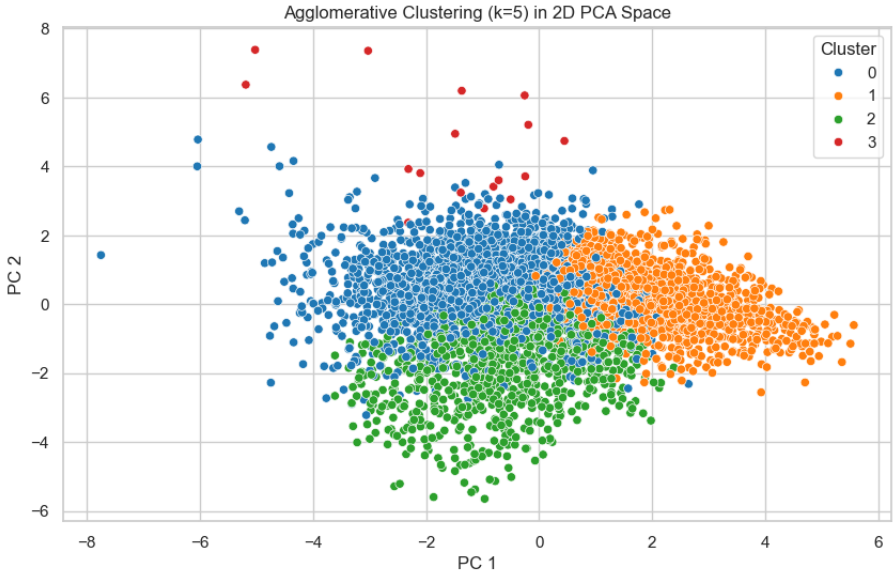
Distribusi ini menunjukkan bahwa terdapat ketidakseimbangan jumlah anggota antar *cluster*, di mana *cluster* 2 memiliki jumlah anggota yang lebih besar dibandingkan dua *cluster* lainnya. Hal ini bisa mengindikasikan adanya pola atau densitas data yang lebih tinggi pada area tertentu dari ruang fitur, sehingga menghasilkan *cluster* yang lebih besar secara alami.

4.4.3 *Clustering* dengan Agglomerative



Gambar 20: Penentuan Jumlah *Cluster* Agglomerative

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 20, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 4. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 4 untuk Agglomerative. Selanjutnya, dilakukan *clustering* menggunakan metode Agglomerative untuk 4 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 21.



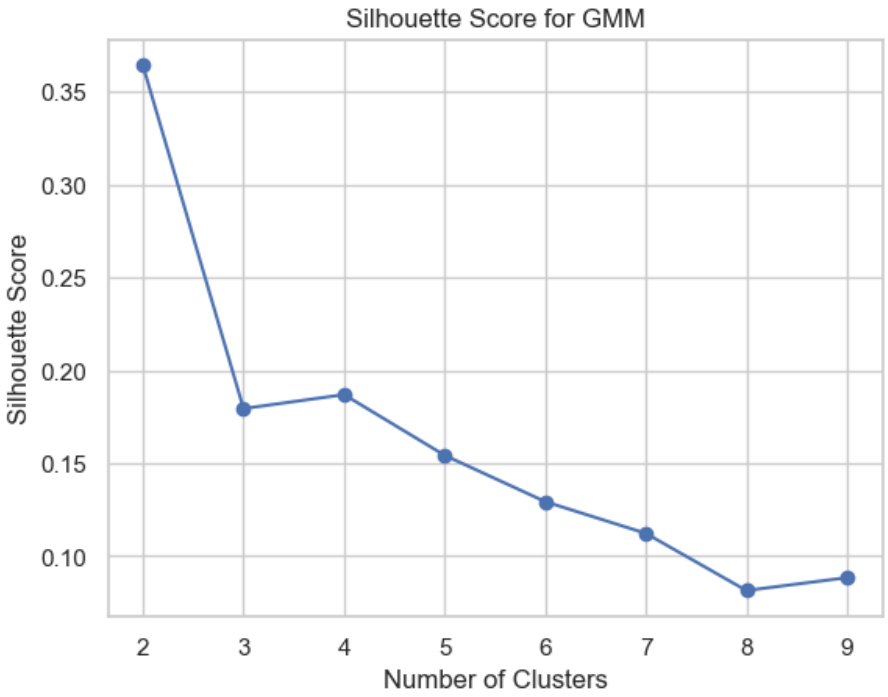
Gambar 21: Visualisasi *Cluster* Agglomerative

Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 2554 untuk *cluster* 0;
- 1404 untuk *cluster* 1;
- 1012 untuk *cluster* 2; dan
- 18 untuk *cluster* 3.

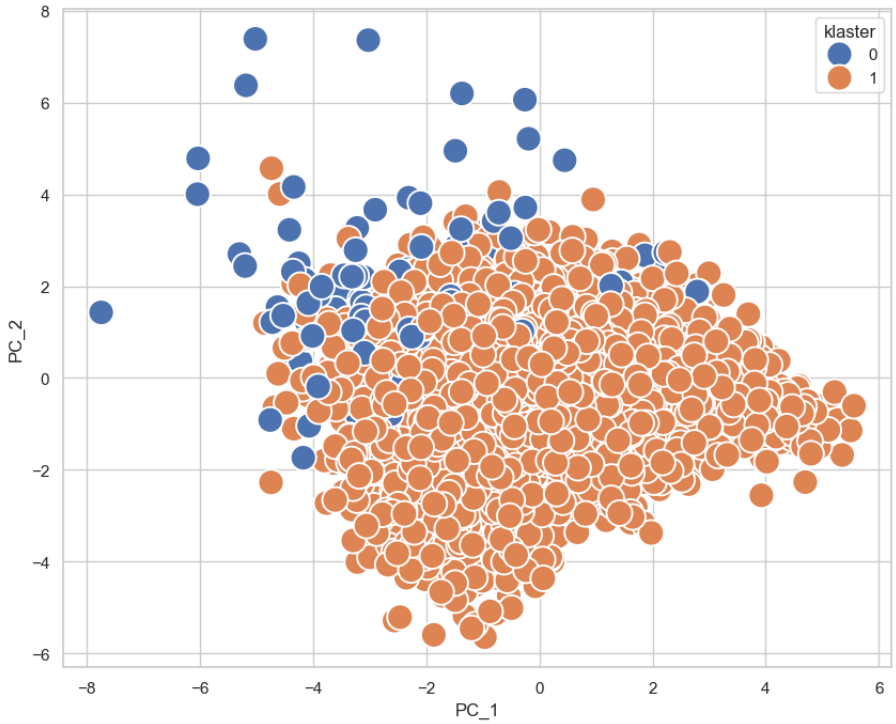
Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode Agglomerative relatif timpang untuk *cluster* 3.

4.4.4 *Clustering* dengan Gaussian Mixture Model (GMM)



Gambar 22: Penentuan Jumlah *Cluster* GMM

Dengan *Silhouette Method*, berdasarkan grafik nilai *silhouette* pada Gambar 22, diketahui titik yang memiliki nilai paling tinggi adalah adalah titik 2. Dengan demikian, ditetapkan jumlah *cluster* optimal adalah 2 untuk GMM. Selanjutnya, dilakukan *clustering* menggunakan metode GMM untuk 2 *cluster* dan diperoleh visualisasi hasil *cluster* pada Gambar 23.



Gambar 23: Visualisasi *Cluster* GMM

Berdasarkan hasil *clustering* dengan metode K-Means, diperoleh jumlah anggota *cluster* sebanyak

- 224 untuk *cluster* 0; dan
- 4764 untuk *cluster* 1.

Distribusi jumlah anggota dalam masing-masing *cluster* menunjukkan bahwa hasil *clustering* dengan metode GMM sangat timpang dan terdapat dominasi jumlah yang ekstrem pada salah satu *cluster*.

4.5 Evaluasi Metode *Clustering*

Evaluasi bertujuan untuk menilai kualitas pemisahan cluster serta kepadatan *cluster* yang dihasilkan oleh beberapa metode *clustering*, yaitu KMeans, KMedians, Agglomerative, dan Gaussian Mixture Model (GMM) dengan dan tanpa *dimensionality reduction*. Setiap metode akan dievaluasi menggunakan 4 metrik: *Silhouette Score*, Indeks Davies-Bouldin, Indeks Calinski-Harabasz, dan Indeks Dunn. Evaluasi model untuk data tanpa dan dengan *dimensionality reduction* secara berurutan disajikan pada Tabel 1 dan 2.

Tabel 1: Hasil Evaluasi Model *Clustering* tanpa *Dimensionality Reduction*

Metode	**Silhouette Score**	**Davies-Bouldin Index**	**Calinski-Harabasz Index**	**Dunn Index**
KMeans	0,1975	1,5723	**1183,6891**	0,0127
KMedians	**0,6945**	**0,6567**	316,3443	**0,1801**
Agglomerative	0,1060	1,7291	694,5922	0,0291
GMM	0,2469	2,0946	615,3756	0,0149

Tabel 2: Hasil Evaluasi Model *Clustering* dengan *Dimensionality Reduction*

Metode	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Dunn Index
KMeans	0,2222	**1,4261**	**1364,9831**	0,0053
KMedians	0,1638	1,7924	932,5984	0,0110
Agglomerative	0,1163	1,5916	868,0172	**0,0147**
GMM	**0,3651**	2,4129	216,0541	0,0176

Secara keseluruhan, model KMedians tanpa d*imensionality reduction* memberikan hasil evaluasi terbaik berdasarkan sebagian besar metrik, terutama Silhouette Score yang tinggi dan Davies-Bouldin Index yang rendah, yang menunjukkan cluster yang lebih *compact* dan terpisah dengan baik. Namun, hasil clustering KMedians tersebut menghasilkan proporsi kelas yang sangat timpang sehingga berisiko menimbulkan bias dan kurang representatif untuk analisis selanjutnya.

Karena proporsi kelas yang seimbang juga penting agar hasil *clustering* dapat diaplikasikan secara efektif dan interpretasi *cluster* menjadi lebih bermakna, akhirnya diputuskan menggunakan model KMeans. Meskipun K-Means tidak selalu menjadi yang terbaik di semua metrik, hasilnya tetap menunjukkan performa yang cukup baik dan proporsi *cluster* yang lebih merata. Setelah penerapan *dimensionality reduction*, KMeans memiliki kinerja yang stabil dan metrik evaluasi yang paling baik kedua dibanding metode lain.

Dengan mempertimbangkan kualitas *cluster* dan distribusi proporsi kelas, K-Means setelah *dimensionality reduction* dipilih sebagai metode terbaik.

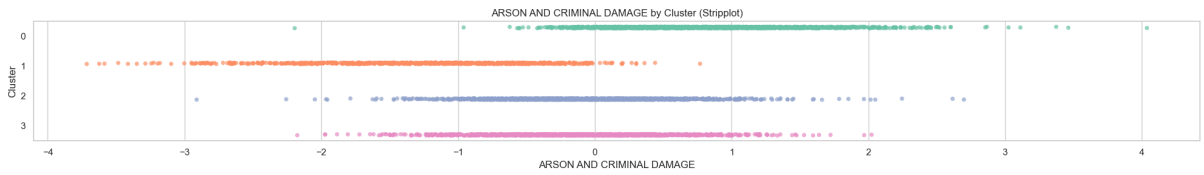
4.6 Profiling *Cluster* dari Metode Terbaik

Profiling *cluster* bertujuan untuk mengidentifikasi dan mendeskripsikan karakteristik khusus dari masing-masing cluster yang terbentuk sehingga memberikan pemahaman yang lebih mendalam terkait pola dan tingkat kriminalitas di setiap cluster. Dalam hal ini, profiling akan menyoroti jenis-jenis tindak kejahatan yang dominan pada setiap *cluster* sehingga diperoleh gambaran distribusi serta intensitas berbagai kategori kriminalitas secara spesifik.

4.6.1 Jenis Kejahatan *Arson and Criminal Damage*

Tabel 3: Statistik *Arson and Criminal Damage* per *Cluster*

Cluster	Min	Max	Mean
0	-2.194625	4.036200	0.811853
1	-3.715393	0.769179	-1.262516
2	-2.911589	2.698202	-0.069299
3	-2.175974	2.023258	-0.116070



Gambar 24: *Swarmplot* Distribusi Proporsi Kejahatan *Arson and Criminal Damage*

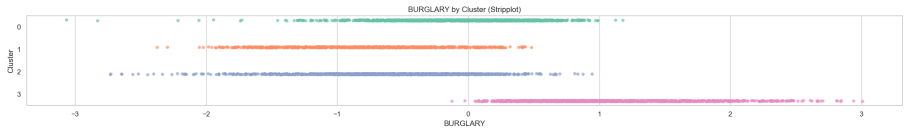
Secara umum, berdasarkan statistik deskriptif pada Tabel 3, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 0. Secara lebih lanjut, pada *swarmplot* Gambar 24 juga diketahui bahwa *cluster* 0 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 0 memiliki insiden pembakaran dan perusakan properti paling tinggi dibanding *cluster* lain.

4.6.2 Jenis Kejahatan *Burglary*

Tabel 4: Statistik *Arson and Criminal Damage* per *Cluster*

Cluster	Min	Max	Mean
0	-3.063893	1.180955	-0.061299
1	-2.372915	0.483857	-0.769123
2	-2.730024	0.945899	-0.609647
3	-0.123445	3.009870	1.059189



Gambar 25: Swarmplot Distribusi Proporsi Kejahatan *Burglary*

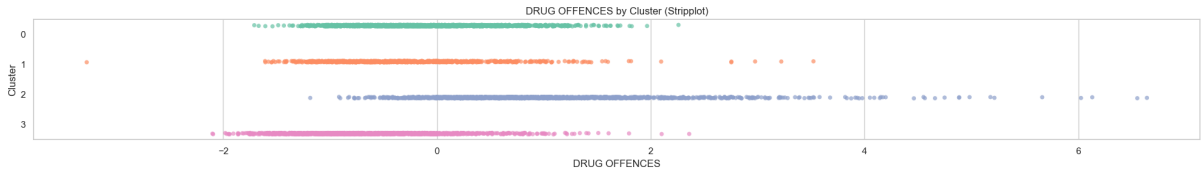
Secara umum, berdasarkan statistik deskriptif pada Tabel 4, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 3. Secara lebih lanjut, pada *swarmplot* Gambar 25 juga diketahui bahwa *cluster* 3 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *Cluster* 3 merupakan area dengan tingkat pencurian rumah paling menonjol dibandingkan *cluster* lainnya.

4.6.3 Jenis Kejahatan *Drug Offences*

Tabel 5: Statistik *Drug Offences* per *Cluster*

Cluster	Min	Max	Mean
0	-1.708544	2.259719	-0.115621
1	-3.276051	3.522526	-0.286243
2	-1.185081	6.642040	1.024169
3	-2.099667	2.359918	-0.648071



Gambar 26: *Swarmplot* Distribusi Proporsi Kejahatan *Drug Offences*

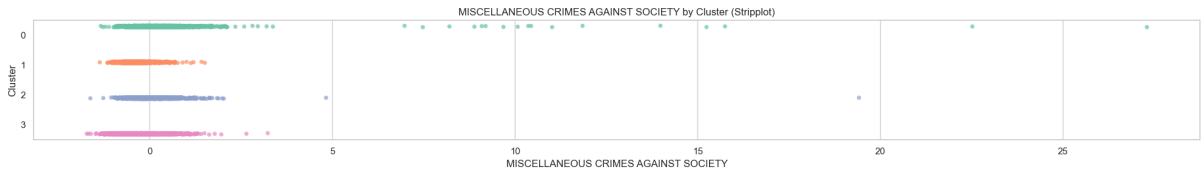
Secara umum, berdasarkan statistik deskriptif pada Tabel 5, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 2. Secara lebih lanjut, pada *swarmplot* Gambar 26 juga diketahui bahwa *cluster* 2 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 2 memiliki tingkat pelanggaran narkoba yang paling menonjol dibandingkan *cluster* lainnya.

4.6.4 Jenis Kejahatan *Miscellaneous Crimes Against Society*

Tabel 6: Statistik *Miscellaneous Crimes Against Society* per *Cluster*

Cluster	Min	Max	Mean
0	-1.335626	27.295280	0.357583
1	-1.363352	1.511905	-0.325373
2	-1.622295	19.414200	0.061964
3	-1.721361	3.232654	-0.265964



Gambar 27: *Swarmplot* Distribusi Proporsi Kejahatan *Miscellaneous Crimes Against Society*

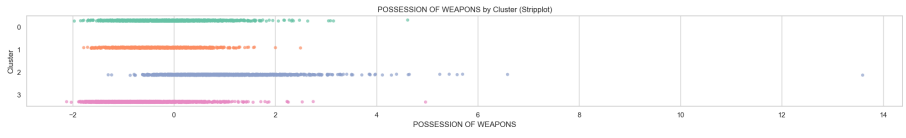
Secara umum, berdasarkan statistik deskriptif pada Tabel 6, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 0. Secara lebih lanjut, pada *swarmplot* Gambar 27 juga diketahui bahwa *cluster* 0 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 0 memiliki tingkat pelanggaran kejahatan sosial lainnya (misalnya prostitusi atau pelanggaran moral) paling menonjol dibandingkan *cluster* lainnya.

4.6.5 Jenis Kejahatan *Possession of Weapons*

Tabel 7: Statistik *Possession of Weapon* per *Cluster*

Cluster	Min	Max	Mean
0	-1.965061	4.615091	-0.103986
1	-1.779025	2.500394	-0.316201
2	-1.295875	13.596200	1.002245
3	-2.117016	4.968136	-0.623893



Gambar 28: *Swarmplot* Distribusi Proporsi Kejahatan *Possession of Weapon*

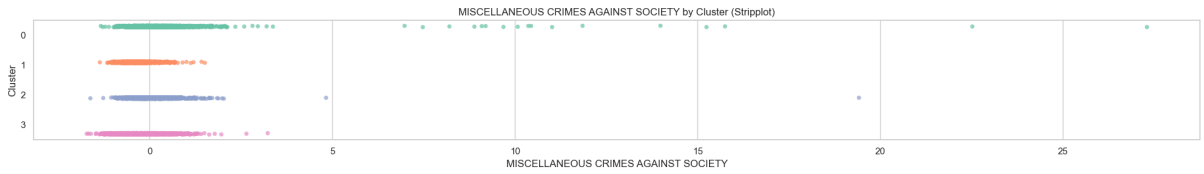
Secara umum, berdasarkan statistik deskriptif pada Tabel 7, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 2. Secara lebih lanjut, pada *swarmplot* Gambar 28 juga diketahui bahwa *cluster* 2 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan wilayah-wilayah yang termasuk *cluster* 2 memiliki Kepemilikan senjata paling menonjol di *cluster* 2 dibandingkan *cluster* lainnya.

4.6.6 Jenis Kejahatan *Public Order Offences*

Tabel 8: Statistik *Public Order Offences* per *Cluster*

Cluster	Min	Max	Mean
0	-2.122693	5.067379	0.496969
1	-1.608819	4.966748	0.195992
2	-2.000889	3.184184	0.337872
3	-3.111305	0.498459	-0.965845



Gambar 29: *Swarmplot* Distribusi Proporsi Kejahatan *Public Order Offences*

Secara umum, berdasarkan statistik deskriptif pada Tabel 8, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 0. Secara lebih lanjut, pada *swarmplot* Gambar 29 juga diketahui bahwa *cluster* 0 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 0 memiliki tingkat gangguan ketertiban umum paling menonjol dibandingkan *cluster* lainnya.

4.6.7 Jenis Kejahatan *Robbery*

Tabel 9: Statistik *Robbery* per *Cluster*

Cluster	Min	Max	Mean
0	-1.979769	1.748851	-0.483236
1	-1.770548	3.429881	-0.030317
2	-1.587310	5.757080	0.957122
3	-2.045067	3.858896	-0.327227

Gambar 30: *Swarmplot* Distribusi Proporsi Kejahatan *Robbery*

Secara umum, berdasarkan statistik deskriptif pada Tabel 9, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 2. Secara lebih lanjut, pada *swarmplot* Gambar 30 juga diketahui bahwa *cluster* 2 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 2 memiliki tingkat perampokan yang paling menonjol dibandingkan *cluster* lainnya.

4.6.8 Jenis Kejahatan *Theft*

Tabel 10: Statistik *Theft* per *Cluster*

Cluster	Min	Max	Mean
0	-2.450911	1.405676	-0.441913
1	0.338662	4.210675	1.753341
2	-1.417275	2.197634	-0.069477
3	-2.055650	1.583445	-0.440665

Gambar 31: *Swarmplot* Distribusi Proporsi Kejahatan *Theft*

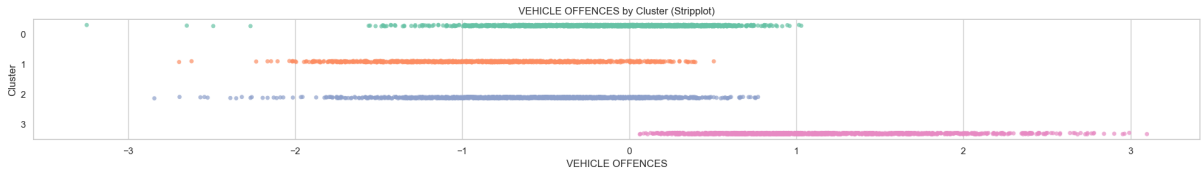
Secara umum, berdasarkan statistik deskriptif pada Tabel 10, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 1. Secara lebih lanjut, pada *swarmplot* Gambar 31 juga diketahui bahwa *cluster* 1 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 1 memiliki tingkat pencurian yang paling menonjol dibandingkan *cluster* lainnya.

4.6.9 Jenis Kejahatan *Vehicle Offences*

Tabel 11: Statistik *Vehicle Offences* per *Cluster*

Cluster	Min	Max	Mean
0	-3.246795	1.029836	-0.111177
1	-2.693851	0.506316	-0.837009
2	-2.841967	0.771235	-0.551305
3	0.063178	3.098503	1.098970



Gambar 32: *Swarmplot* Distribusi Proporsi Kejahatan *Vehicle Offences*

Secara umum, berdasarkan statistik deskriptif pada Tabel 11, diketahui rata-rata proporsi paling tinggi untuk kasus kejahatan ini berada pada *cluster* 3. Secara lebih lanjut, pada *swarmplot* Gambar 32 juga diketahui bahwa *cluster* 3 memiliki rentang proporsi yang lebih tinggi untuk jenis kejahatan ini dibandingkan *cluster* lain.

Dengan demikian, dapat disimpulkan bahwa wilayah-wilayah yang termasuk *cluster* 3 memiliki tingkat pelanggaran terkait kendaraan yang paling menonjol dibandingkan *cluster* lainnya.

4.7 Hasil Akhir *Cluster*

Berdasarkan analisis klasterisasi terbaik (K-Means pada data dengan *dimensional reduction*) wilayah berdasarkan jenis kriminalitas di Kota London, terbentuk 4 *cluster* dengan karakteristik kejahatan yang berbeda-beda. Adapun kesimpulan dari masing-masing *cluster* adalah sebagai berikut.

Cl. 1 menonjol pada kasus *arson and criminal damage* serta *possession of weapons*.
Wilayah-wilayah pada *Cluster* 0 memiliki proporsi kasus *arson and criminal damage* (pembakaran dan perusakan properti) serta *possession of weapons* (kepemilikan senjata) yang paling tinggi dibandingkan klaster lainnya. Hal ini menandakan bahwa klaster ini terdiri dari area-area dengan tingkat vandalisme dan kepemilikan senjata ilegal yang cukup signifikan.

Cl. 2 menonjol pada kasus *public order offences* dan *robbery*.
Wilayah-wilayah pada *Cluster* 1 memiliki proporsi kasus *public order offences* (gangguan ketertiban umum) dan *robbery* (perampokan) yang paling tinggi. Hal ini menandakan bahwa klaster ini mencakup area yang rawan terhadap konflik sosial, unjuk rasa yang berujung kekerasan, serta tindak kejahatan dengan kekerasan untuk mengambil barang dari korban secara langsung.

Cl. 3 menonjol pada kasus *drug offences*, *miscellaneous crimes against society*, dan *theft*.
Wilayah-wilayah pada *Cluster* 2 memiliki proporsi kasus *drug offences* (pelanggaran narkoba), *miscellaneous crimes against society* (kejahatan sosial lainnya), dan *theft* (pencurian umum) yang paling tinggi. Hal ini menandakan bahwa area dalam klaster ini sangat terdampak oleh peredaran dan penyalahgunaan narkoba, pelanggaran hukum sosial seperti prostitusi ilegal dan imigrasi, serta pencurian tanpa kekerasan seperti pencopetan dan kehilangan barang.

Cl. 4 menonjol pada kasus *burglary* dan *vehicle offences*.
Wilayah-wilayah pada Cluster 3 memiliki proporsi kasus *burglary* (pencurian rumah) dan *vehicle offences* (pelanggaran terkait kendaraan) yang paling tinggi dibanding klaster lainnya. Hal ini menandakan bahwa klaster ini terdiri dari area yang rawan terhadap pencurian properti dengan masuk secara ilegal ke dalam bangunan serta pelanggaran hukum yang berkaitan dengan kendaraan, seperti pencurian atau perusakan kendaraan.

BAB V
Kesimpulan

5.1 Kesimpulan

Clustering bertujuan untuk mengelompokkan wilayah-wilayah di Kota London berdasarkan kesamaan pola kriminalitas yang dimiliki. Dengan menggunakan teknik ini, dapat diidentifikasi kelompok wilayah yang memiliki karakteristik kejahatan yang serupa sehingga mempermudah dalam memahami distribusi jenis kejahatan di berbagai area. Hasil dari analisis ini diharapkan dapat menjadi dasar bagi pihak berwenang dalam merancang strategi pencegahan dan penanganan kejahatan yang lebih tepat sasaran sesuai dengan karakteristik masing-masing wilayah.

Sebelum dilakukan *clustering*, diperoleh informasi terkait pola data dengan analisis eksploratif menggunakan visualisasi *heatmap*. Diketahui terdapat beberapa kasus kejahatan yang saling berkorelasi. Korelasi positif paling tinggi ditemukan antara tipe kejahatan *burglary* dan *vehicle offences* yang menandakan bahwa area dengan banyak kasus pembobolan rumah biasanya juga memiliki banyak kasus pencurian kendaraan. Selanjutnya, diketahui tipe kejahatan *arson* juga berkorelasi positif cukup terhadap tipe kejahatan *violence against the person* yang menunjukkan bahwa daerah dengan banyak insiden pembakaran juga cenderung memiliki banyak kasus kekerasan terhadap individu.

Sebaliknya, terdapat pula tipe kejahatan yang berkorelasi negatif. Korelasi negatif paling tinggi ditemukan antara tipe kejahatan *theft* dengan *arson* dan *violence against the person* yang menunjukkan bahwa wilayah dengan banyak kasus pencurian justru cenderung memiliki lebih sedikit kasus pembakaran atau kekerasan terhadap orang. Selain itu, diketahui korelasi negatif yang cukup kuat pula antara tipe kejahatan *public order offences* dan *vehicle offences* yang menandakan bahwa pelanggaran ketertiban umum cenderung bertolak belakang dengan pencurian kendaraan.

Berdasarkan informasi yang diperoleh dari analisis data eksploratif, dilakukan analisis lebih lanjutn *clustering* untuk mengelompokkan daerah-daerah dengan pola-pola yang mirip. Dalam hal ini, dilakukan *clustering* dengan 4 jenis model, yaitu K-Means, K-Median, Agglomerative Hierarchical, dan Gaussian Mixture Model. *Clustering* dilakukan pada data dengan 2 skema berbeda: tanpa *dimensional reduction* dan dengan *dimensional reduction*. Metode *clustering* selanjutnya dibandingkan dan dievaluasi menggunakan 4 metrik terpilih, yaitu *Silhouette Score*, Indeks Davies-Bouldin, Indeks Calinski-Harabasz, dan Indeks Dunn.

Berdasarkan hasil evaluasi, secara metrik, diperoleh model terbaik adalah K-Median yang diterapkan pada data tanpa *dimensional reduction*. Akan tetapi, model ini memiliki sebaran proporsi *cluster* yang sangat timpang. Padahal, proporsi kelas yang seimbang juga penting agar hasil *clustering* dapat diaplikasikan secara efektif dan interpretasi *cluster* menjadi lebih bermakna. Dengan demikian, dipilih model terbaik kedua berdasarkan metrik sebagai model acuan, yaitu K-Means yang diterapkan pada data dengan *dimensional reduction*. Model ini memiliki selisih metrik yang dekat dengan model terbaik pertama dan proporsi kelas yang jauh lebih seimbang sehingga dapat digunakan sebagai acuan *clustering*.

Berdasarkan analisis klasterisasi terbaik (K-Means pada data dengan *dimensional reduction*) wilayah berdasarkan jenis kriminalitas di Kota London, terbentuk 4 *cluster* dengan karakteristik kejahatan yang berbeda-beda. Adapun kesimpulan dari masing-masing *cluster* adalah sebagai berikut.

Cl. 1 menonjol pada kasus *arson and criminal damage* serta *possession of weapons*.
Wilayah-wilayah pada *Cluster* 0 memiliki proporsi kasus *arson and criminal damage* (pembakaran dan perusakan properti) serta *possession of weapons* (kepemilikan senjata) yang paling tinggi dibandingkan klaster lainnya. Hal ini menandakan bahwa klaster ini terdiri dari area-area dengan tingkat vandalisme dan kepemilikan senjata ilegal yang cukup signifikan.

Cl. 2 menonjol pada kasus *public order offences* dan *robbery*.
Wilayah-wilayah pada *Cluster* 1 memiliki proporsi kasus *public order offences* (gangguan ketertiban umum) dan *robbery* (perampokan) yang paling tinggi. Hal ini menandakan bahwa klaster ini mencakup area yang rawan terhadap konflik sosial, unjuk rasa yang berujung kekerasan, serta tindak kejahatan dengan kekerasan untuk mengambil barang dari korban secara langsung.

Cl. 3 menonjol pada kasus *drug offences*, *miscellaneous crimes against society*, dan *theft*.
Wilayah-wilayah pada *Cluster* 2 memiliki proporsi kasus *drug offences* (pelanggaran narkoba), *miscellaneous crimes against society* (kejahatan sosial lainnya), dan *theft* (pencurian umum) yang paling tinggi. Hal ini menandakan bahwa area dalam klaster ini sangat terdampak oleh peredaran

dan penyalahgunaan narkoba, pelanggaran hukum sosial seperti prostitusi ilegal dan imigrasi, serta pencurian tanpa kekerasan seperti pencopetan dan kehilangan barang.

Cl. 4 menonjol pada kasus *burglary* dan *vehicle offences*.
Wilayah-wilayah pada Cluster 3 memiliki proporsi kasus *burglary* (pencurian rumah) dan *vehicle offences* (pelanggaran terkait kendaraan) yang paling tinggi dibanding klaster lainnya. Hal ini menandakan bahwa klaster ini terdiri dari area yang rawan terhadap pencurian properti dengan masuk secara ilegal ke dalam bangunan serta pelanggaran hukum yang berkaitan dengan kendaraan, seperti pencurian atau perusakan kendaraan.

Secara umum, hasil *clustering* menunjukkan konsistensi yang cukup baik dengan informasi awal yang diperoleh dari analisis data eksploratif. Dalam analisis korelasi, diketahui bahwa terdapat korelasi positif yang kuat antara *burglary* dan *vehicle offences*. Hal ini tercermin dalam hasil *clustering* karena keduanya sama-sama mendominasi *Cluster* 3. Ini menunjukkan bahwa wilayah dengan banyak pencurian rumah juga cenderung mengalami kasus pencurian atau pelanggaran terkait kendaraan. Selain itu, korelasi negatif antara *public order offences* dan *vehicle offences* juga terlihat pada hasil klasterisasi. Cluster 1 menampung banyak kasus *public order offences*, sementara *vehicle offences* mendominasi *Cluster* 3, yang memperkuat temuan bahwa kedua jenis kejahatan ini cenderung tidak terjadi bersamaan dalam satu area.

Dengan demikian, dapat disimpulkan bahwa hasil *clustering* cukup selaras dengan pola-pola hubungan antar jenis kejahatan yang teridentifikasi dalam analisis eksploratif. *Clustering* berhasil mengelompokkan wilayah-wilayah yang memiliki profil kejahatan serupa dan mencerminkan struktur korelasi yang sebelumnya diamati.

5.2 Saran

Berdasarkan hasil analisis *clustering* yang didasarkan pada profil jenis kejahatan di tiap wilayah, serta keterkaitannya dengan pola korelasi antar variabel kriminalitas, dapat dirumuskan beberapa saran strategis sebagai berikut.

1. Pendekatan Penanganan Berbasis Profil Kriminalitas Setiap *cluster* menunjukkan pola dominan kejahatan yang berbeda. Oleh karena itu, intervensi kebijakan keamanan harus bersifat tailored atau disesuaikan secara spesifik untuk setiap *cluster*:

   - *Cluster* 0: Penekanan pada *preventive surveillance* (pengawasan pencegahan) untuk kejahatan berbasis kekerasan properti seperti pembakaran dan kepemilikan senjata ilegal.
   - *Cluster* 1: Fokus pada penguatan peran aparat dalam menjaga ketertiban umum dan meminimalkan konflik sosial. Diperlukan kehadiran aparat secara aktif di titik-titik rawan unjuk rasa dan kerumunan.
   - *Cluster* 2: Implementasi program rehabilitasi narkoba, kampanye kesadaran sosial, serta peningkatan sistem pelaporan untuk pencurian kecil agar respons penegakan hukum lebih cepat.
   - *Cluster* 3: Peningkatan pengawasan lingkungan tempat tinggal dan sistem keamanan kendaraan seperti CCTV, patroli rutin, dan edukasi publik terkait modus kejahatan properti.

2. Penguatan Program Pencegahan Berbasis Komunitas

   Masyarakat memiliki peran penting dalam mendeteksi potensi kejahatan lebih awal. Untuk itu, setiap klaster dapat menjadi basis penerapan program *Community Policing*, yaitu kemitraan antara warga dan aparat keamanan dalam menjaga lingkungan masing-masing.

3. Integrasi Data dan Kebijakan Wilayah

   Hasil *clustering* dapat dijadikan dasar dalam penyusunan kebijakan tata ruang dan keamanan wilayah. Misalnya, daerah dengan dominasi pencurian (*Cluster* 2 dan 3) dapat diusulkan untuk diberi penerangan jalan tambahan, sistem alarm lingkungan, atau pos keamanan terpadu.

4. Pemanfaatan Teknologi untuk Pengawasan dan Prediksi

   Karena hasil *clustering* ini merepresentasikan pola data kriminal historis, maka dapat digunakan untuk mengembangkan sistem *early warning* berbasis machine learning agar polisi dapat memprediksi jenis kejahatan dominan yang berpotensi muncul di wilayah tertentu berdasarkan ciri-ciri *cluster*-nya.

DAFTAR PUSTAKA

[1] Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences, 34*, 1785–1793. https://doi.org/10.1016/j.jksuci.2019.12.011

[2] Fariz, T. K. N., & Basha, S. S. (2024). Enhancing solar radiation predictions through COA optimized neural networks and PCA dimensionality reduction. *Energy Reports, 12*, 341–359. https://doi.org/10.1016/j.egyr.2024.06.025

[3] Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon, 11*, e41953. https://doi.org/10.1016/j.heliyon.2025.e41953

[4] Lai, X., Deng, X., Tang, X., Gao, F., Han, X., & Zheng, Y. (2022). Soft clustering of retired lithium-ion batteries for the secondary utilization using Gaussian mixture model based on electrochemical impedance spectroscopy. *Journal of Cleaner Production, 339*, 130786. https://doi.org/10.1016/j.jclepro.2022.130786

[5] Omran, M. G., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. *Intelligent Data Analysis, 11*(6), 583–605. https://doi.org/10.3233/ida-2007-11602

[6] Oti, E. U., & Olusola, M. O. (2024). Overview of agglomerative hierarchical clustering methods. *British Journal of Computer Networking and Information Technology, 7*(2), 14–23. https://doi.org/10.52589/bjcnit-cv9poogw

[7] Rahmawati, T., Wilandari, Y., & Kartikasari, P. (2024). Analisis perbandingan silhouette coefficient dan metode elbow pada pengelompokkan provinsi di Indonesia berdasarkan indikator IPM dengan K-medoids. *Jurnal Gaussian, 13*(1), 13–24. https://doi.org/10.14710/j.gauss.13.1.13-24

[8] Reynolds, D. (n.d.). Gaussian mixture models [Lecture notes]. MIT Lincoln Laboratory. http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/refs/GMM_Tutorial_Reynolds.pdf

[9] Sihombing, P. R. (2021). Implementation of K-Means and K-Medians clustering in several countries based on Global Innovation Index (GII) 2018. *Advance Sustainable Science Engineering and Technology, 3*(1), 0210107. https://doi.org/10.26877/asset.v3i1.8461

[10] Song, J., Wang, K., Bian, T., Li, W., Dong, Q., Chen, L., Xue, G., & Wu, X. (2025). A novel heat load prediction algorithm based on fuzzy C-mean clustering and mixed positional encoding informer. *Applied Energy, 388*, 125709. https://doi.org/10.1016/j.apenergy.2025.125709

[11] Ummami, R., & Winarno, B. (2023). Gaussian Mixture Model dengan Algoritme Expectation Maximization untuk Pengelompokan Data Distribusi Air Bersih di Jawa Barat. *PRISMA, Prosiding Seminar Nasional Matematika, 6*, 745–750. https://journal.unnes.ac.id/sju/index.php/prisma/

[12] Van Der Maaten, L., Postma, E., Van Den Herik, J., Tilburg centre for Creative Computing, & Tilburg University. (2009). Dimensionality Reduction: A Comparative Review (TiCC TR 2009–005). TiCC, Tilburg University. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

[13] Zambrano, J. (2017). Gaussian Mixture Model – method and application. https://doi.org/10.13140/RG.2.2.32667.77602

[14] Zhu, Q., Tang, X., & Elahi, A. (2021). Application of the novel harmony search optimization algorithm for DBSCAN clustering. *Expert Systems with Applications, 178*, 115054. https://doi.org/10.1016/j.eswa.2021.115054

LAMPIRAN

Data dan *Notebook* analisis lengkap dapat diakses melalui GitHub.
(github.com/salmanataya/pdm-clustering apabila *hyperlink* tidak dapat diakses)