

# Analisis Pola dan Pengelompokan Tingkat Banjir DKI Jakarta dengan K-Means Clustering



# INFINITY



Khoirul Amar  
Sidik



Balqis Dwian Fitri  
Zamzami



Mujadid Choirus  
Surya



# Outline

- About Data
- Problem Statement
- Hipotesis
- Penentuan Metodologi dan Variabel
- Analisis Awal
- Analisis Mendalam
- Kesimpulan dan Rekomendasi
- Daftar Pustaka

# About Dataset

Data Kejadian Bencana Banjir di Provinsi DKI Jakarta Tahun 2020.  
Dataset ini berisi tentang Data Rekapitulasi Bulanan Kejadian Banjir di Provinsi DKI Jakarta Tahun 2020.

The screenshot shows the Jakarta Open Data website. At the top, there is a logo for 'JAKARTA open data' with the tagline 'Berbagi Data untuk Transparansi'. Below the logo, the navigation bar includes 'Home / Organisasi / Badan Penanggulangan Bencana Daerah'. The main content area has a title 'Organisasi' with a sub-section for 'Badan Penanggulangan Bencana Daerah'. It features a large image of the DKI Jakarta skyline at sunset with the text 'PROVINSI DKI JAKARTA' overlaid. Below the image, the text reads: 'Badan Penanggulangan Bencana Daerah' and 'Badan yang mengurus dan menanggulangi segala bencana di'. A megaphone icon is visible in the bottom left corner.

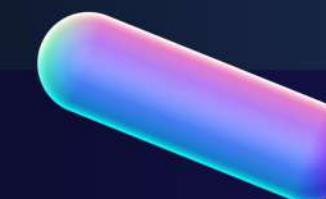
Data columns (total 20 columns):			
#	Column	Non-Null Count	Dtype
0	kota_administrasi	1006 non-null	object
1	kecamatan	1006 non-null	object
2	kelurahan	1006 non-null	object
3	rw	1006 non-null	object
4	jumlah_terdampak_rw	1006 non-null	int64
5	jumlah_terdampak_rt	1006 non-null	int64
6	jumlah_terdampak_kk	1006 non-null	object
7	jumlah_terdampak_jiwa	1006 non-null	int64
8	ketinggian_air	1006 non-null	object
9	tanggal_kejadian	1006 non-null	object
10	lama_genangan	1006 non-null	int64
11	jumlah_meninggal	1006 non-null	int64
12	jumlah_hilang	1006 non-null	int64
13	jumlah_luka_berat	1006 non-null	int64
14	jumlah_luka_ringan	1006 non-null	int64
15	jumlah_pengungsi_tertinggi	1006 non-null	int64
16	jumlah_tempat_pengungsian	1006 non-null	int64
17	nilai_kerugian	1006 non-null	int64
18	Tahun	1006 non-null	object
19	Bulan	1006 non-null	int64

Dataset :



# Problem Statement

- Bagaimana menampilkan distribusi data berdasarkan kota administrasi, kecamatan, dan kelurahan untuk mendapatkan ringkasan menarik tentang banjir di Jakarta?
- Bagaimana menampilkan grafik untuk menunjukkan tren tinggi rata-rata ketinggian air, puncak kejadian banjir setiap bulan selama tahun 2020 di berbagai lokasi di Jakarta?
- Bagaimana menggunakan algoritma K-Means untuk melakukan clustering pada data berdasarkan atribut-atribut yang relevan seperti tinggi rata-rata ketinggian air, jumlah jiwa terdampak, jumlah tempat pengungsian, dan lama genangan?
- Bagaimana melakukan visualisasi grafik batang dan evaluasi model untuk membandingkan nilai atribut-atribut yang relevan antara kelompok-kelompok hasil clustering?



# Hipotesis

- Distribusi data berdasarkan kota administrasi, kecamatan, dan kelurahan akan membantu mengidentifikasi daerah-daerah yang paling terdampak banjir dan membutuhkan perhatian lebih dalam upaya penanggulangan banjir.
- Grafik yang menunjukkan tren tinggi rata-rata ketinggian air selama tahun 2020 di berbagai lokasi di Jakarta akan mengungkapkan pola dan periode tertentu di mana banjir lebih sering terjadi.
- Penggunaan algoritma K-Means dalam melakukan clustering pada data akan mengidentifikasi kelompok-kelompok daerah dengan karakteristik serupa dalam hal dampak banjir, mempermudah perencanaan dan pengambilan keputusan.
- Visualisasi grafik batang dan evaluasi model akan memberikan pemahaman yang lebih jelas tentang perbedaan dan persamaan antara kelompok-kelompok daerah dalam hal dampak banjir, membantu dalam pemahaman situasi dan pengambilan tindakan yang tepat.

# Penentuan Metodelogi dan Variabel

- *Data Preprocessing*
- *Exploratory Data Analysis (EDA)*
- *K-means Clustering*
- *Principal Component Analyst (PCA)*
- *Silhouette Score*

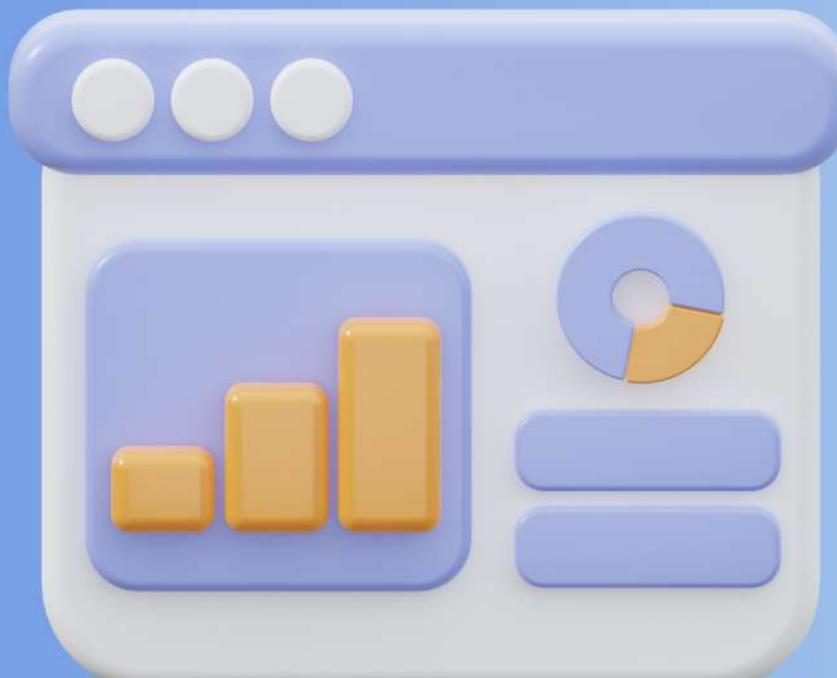


# Data Preprocessing

*Data preprocessing* adalah tahap penting dalam analisis data yang melibatkan transformasi data mentah menjadi bentuk yang lebih terstruktur, bersih, dan siap digunakan untuk analisis lebih lanjut. Proses ini melibatkan serangkaian langkah untuk membersihkan, mengubah, dan mengintegrasikan data agar sesuai dengan kebutuhan analisis yang akan dilakukan.

Langkah-langkah *Data Preprocessing*:

1. *Data Cleaning*
2. *Data Integration*
3. *Data Transformation*
4. *Data Reduction*
5. *Data Discretization*



# Exploratory Data Analysis (EDA)

***Exploratory Data Analysis (EDA)*** adalah suatu pendekatan dalam analisis data yang bertujuan untuk memahami dan menggali informasi penting dari data secara visual dan deskriptif. EDA digunakan untuk mengidentifikasi pola, hubungan, anomali, dan tren yang terdapat dalam data tanpa membuat asumsi atau hipotesis sebelumnya. Dalam EDA, data dieksplorasi melalui berbagai teknik visualisasi dan statistik deskriptif.

Tujuan utama EDA adalah untuk mendapatkan pemahaman awal yang baik tentang data sebelum melakukan analisis yang lebih lanjut atau membangun model prediksi. Dengan melakukan EDA, kita dapat mengidentifikasi data yang hilang, outlier, serta memahami distribusi, hubungan antara variabel, dan kecenderungan yang ada. Hal ini membantu dalam pengambilan keputusan yang lebih baik dan memberikan wawasan yang berharga tentang data.

# K-means Clustering

K-Means adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya.

Langkah-langkah umum dari algoritma K-Means sebagai berikut:

1. Menentukan banyak k-cluster yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat cluster awal (*centroid*) sebanyak k-cluster.
3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak Eucledian
4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).
5. Mengupdate nilai centroid. Nilai centroid baru diperoleh dari rata-rata cluster
6. Melakukan perulangan dari langkah 2 hingga 5 hingga anggota tiap cluster tidak ada yang berubah
7. Jika langkah 6 telah terpenuhi, maka nilai rata-rata pusat cluster ( $\mu_j$ ) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data

# Principal Component Analyst (PCA)

Principal Component Analysis (PCA) adalah suatu teknik statistik yang secara linear mengubah bentuk sekumpulan variabel asli menjadi kumpulan variabel yang lebih kecil yang tidak berkorelasi yang dapat mewakili informasi dari kumpulan variabel asli (Dunteman. 1989:7).

Tujuan PCA adalah untuk menjelaskan bagian dari variasi dalam kumpulan variabel yang diamati atas dasar beberapa dimensi. Dari variabel yang banyak dirubah menjadi sedikit variabel.

Tujuan khusus PCA:

- Untuk meringkas pola korelasi antar variabel yang diobservasi.
- Mereduksi sejumlah besar variabel menjadi sejumlah kecil faktor,
- Memberikan sebuah definisi operasional (sebuah persamaan regresi) dimensi pokok penggunaan variabel yang diobservasi

# Silhouette Score

***Silhouette Score*** : matriks evaluasi untuk mengukur seberapa dekat setiap sampel dalam satu cluster dengan sampel dalam cluster lain. Nilai yang lebih tinggi berarti hasil clustering lebih baik.

Skor Silhouette berkisar antara -1 dan 1. Skor yang mendekati 1 menunjukkan bahwa titik data jauh lebih dekat dengan titik-titik dalam cluster yang sama dibandingkan dengan titik-titik di cluster lain, yang berarti titik data tersebut sudah dikelompokkan dengan baik. Sebaliknya, skor yang mendekati -1 menunjukkan bahwa titik data lebih dekat dengan titik-titik di cluster lain dibandingkan dengan titik-titik dalam cluster yang sama.

# Analisis Awal (EDA)

- Statistika Deskriptif
- Kota yang sering terjadi bencana banjir
- Top 10 Jumlah Kejadian Bencana Banjir Berdasarkan Kelurahan
- Banyak kejadian banjir setiap bulan tahun 2020 di Jakarta
- Ketinggian air maksimal dan minimal berdasarkan kota
- Menampilkan persebaran data kolom "avg\_ketinggian" menggunakan Box Plot
- Menampilkan hubungan atau pola antara 'jumlah\_terdampak\_jiwa' dan 'avg\_ketinggian' menggunakan scatter plot
- Menampilkan korelasi antara variabel 'jumlah\_terdampak\_jiwa', 'avg\_ketinggian', 'lama\_genangan' dan 'jumlah\_tempat\_pengungsian'.

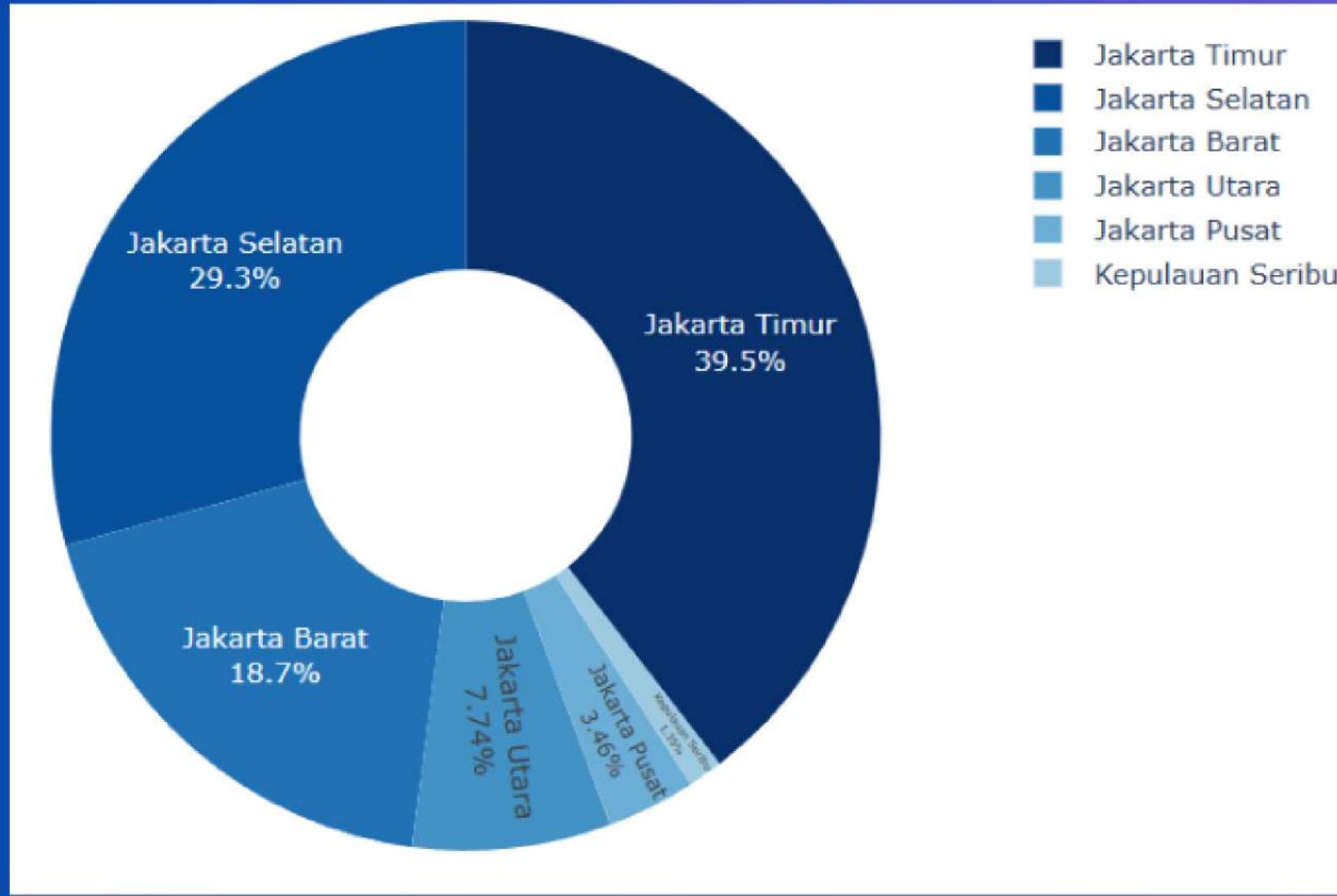
# Statistika Deskriptif

	avg_ketinggian	jumlah_terdampak_jiwa	jumlah_tempat_pengungsian	lama_genangan
count	603.000000	603.000000	603.000000	603.000000
mean	43.398380	250.854063	0.019900	0.031509
std	28.668943	873.885891	0.206864	0.192903
min	1.500000	0.000000	0.000000	0.000000
25%	20.000000	0.000000	0.000000	0.000000
50%	35.000000	0.000000	0.000000	0.000000
75%	55.000000	88.000000	0.000000	0.000000
max	180.000000	13450.000000	3.000000	2.000000

Analisis deskriptif menunjukkan analisis untuk data numerik dimana kita bisa mendapatkan informasi seperti count, mean, standar deviasi, nilai minimum, nilai maksimum, dan nilai kuartil.

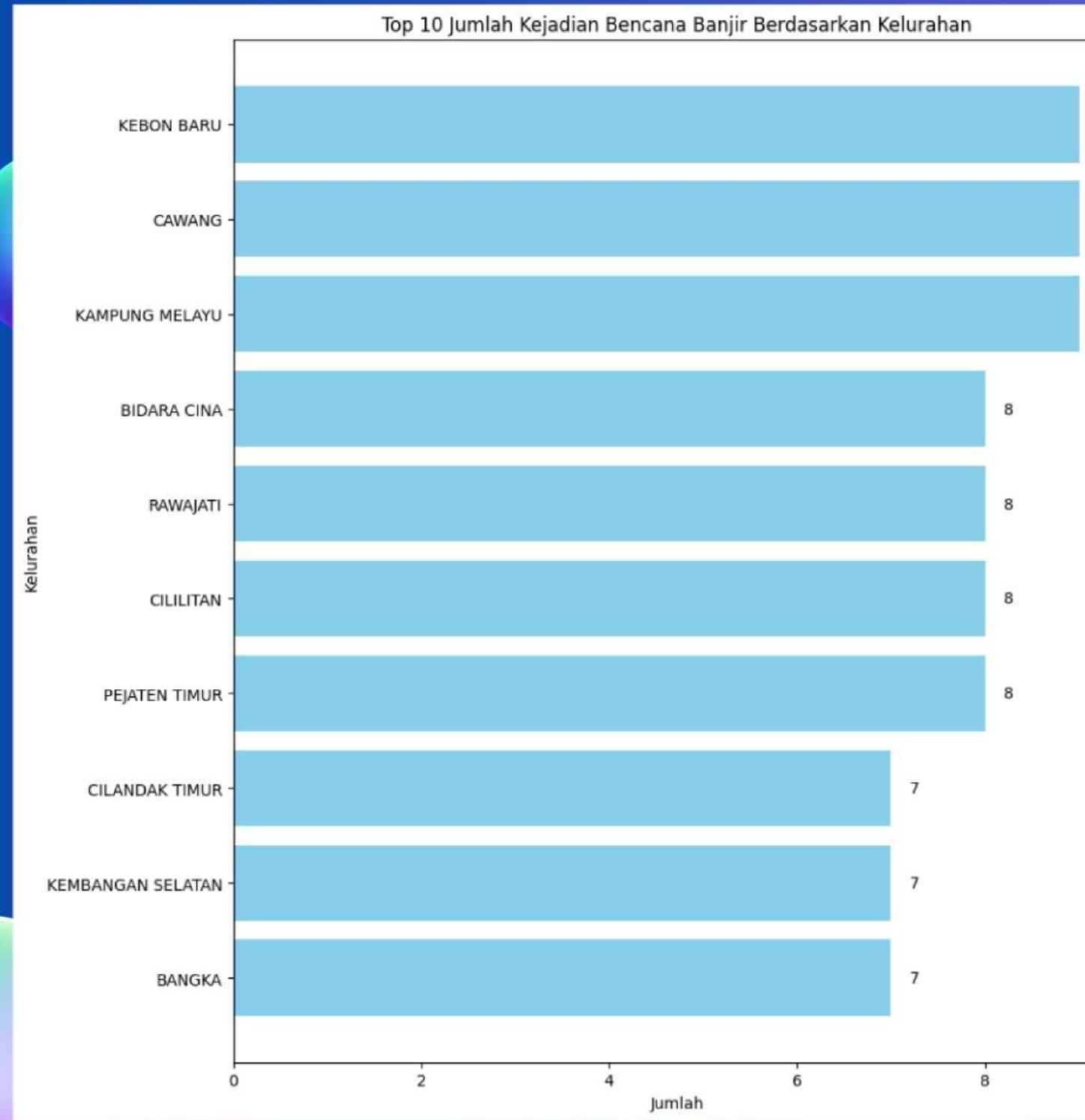
Hal menarik dari hasil analisis ini adalah kita mengetahui rata-rata ketinggian banjir adalah 43,398 cm dengan ketinggian maksimum yaitu 180 dan ketinggian minimum 1,5 cm. Rata-rata jumlah terdampak sebesar 250 orang namun dengan standar deviasi yang cukup besar yaitu 873,885 yang mengartikan bahwa data jumlah terdampak tidak berdistribusi normal.

# Kota yang sering terjadi bencana banjir



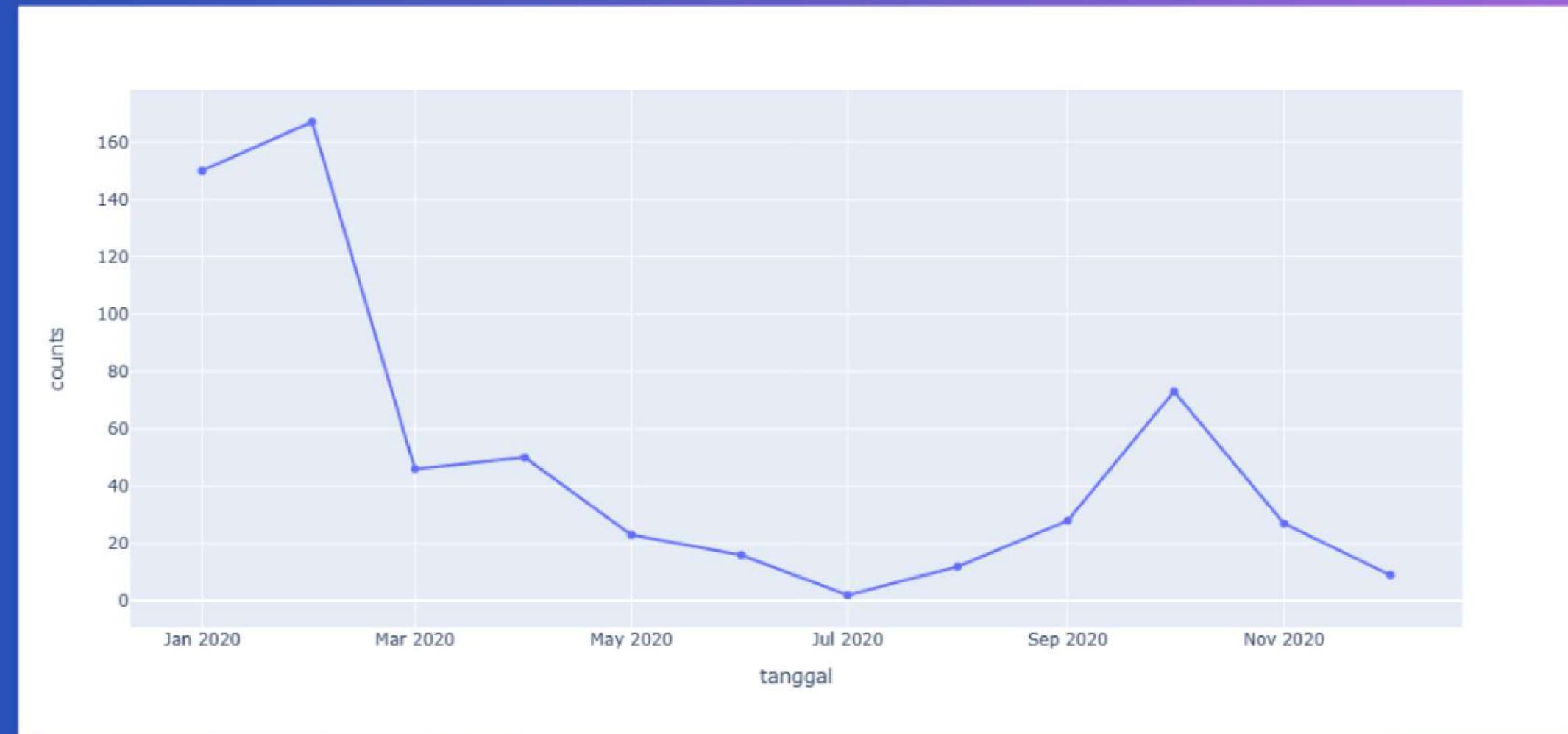
Analisis pie chart disamping merepresentasikan proporsi kota dengan tingkat kejadian banjir paling sering. Dimana Jakarta Timur menjadi kota paling sering terjadi bencana banjir sebesar 39,5% selanjutnya adalah Jakarta Selatan sebesar 29,3% dan Jakarta Barat Sebesar 18,7%. Kota dengan proporsi kejadian banjir paling kecil adalah Kepulauan Seribu sebesar 1,35%. Data proporsi ini mengacu pada banyak kejadian banjir di DKI Jakarta pada tahun 2020.

# Top 10 Jumlah Kejadian Bencana Banjir Berdasarkan Kelurahan



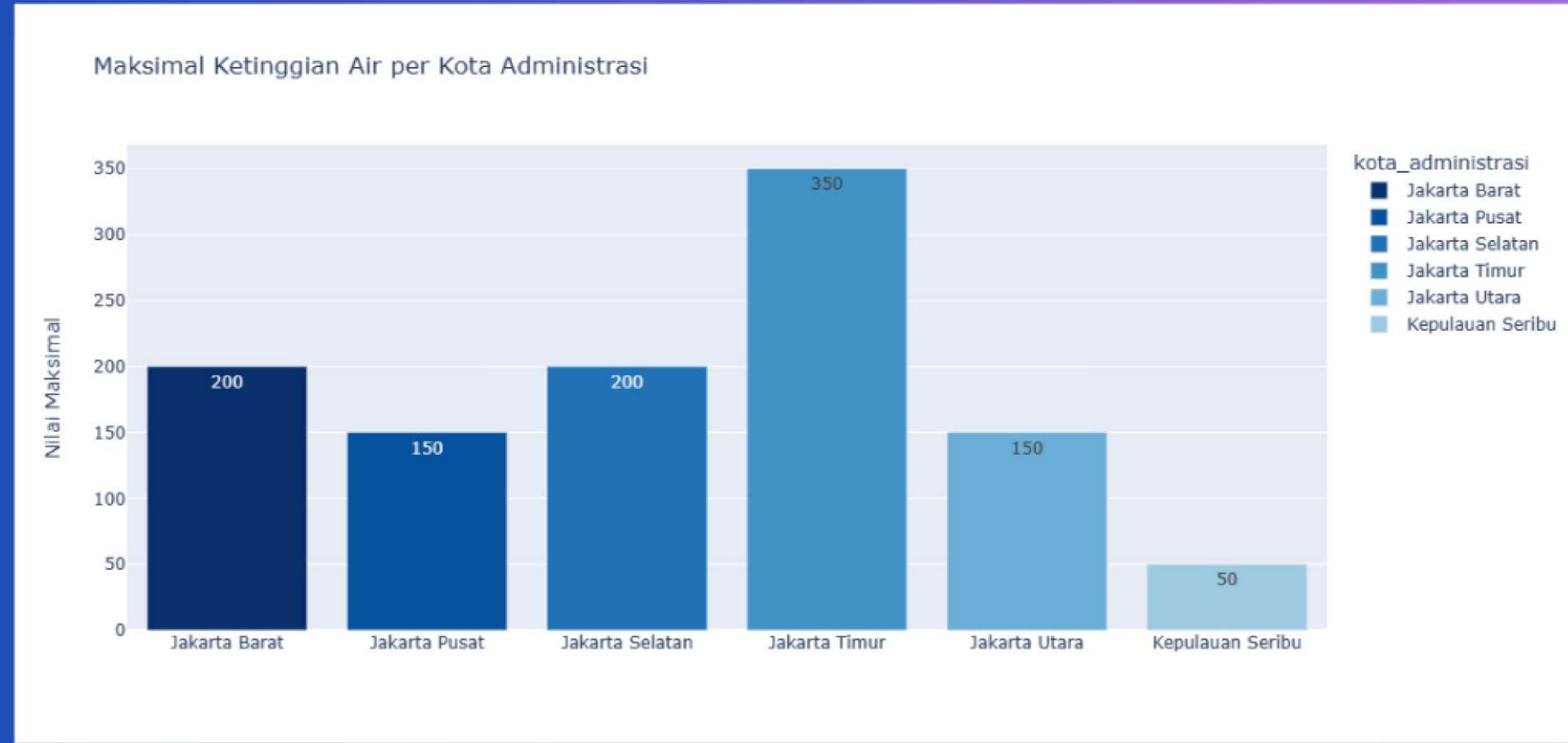
Analisis bar chart horizontal disamping menunjukkan Top 10 kelurahan dengan kejadian banjir paling sering ditahun 2020. Dapat terlihat bahwa kelurahan Kebon Baru, Cawang dan Kampung Melayu menjadi kelurahan dengan tingkat kejadian banjir paling sering di tahun 2020 sebanyak 9 kali. Hal ini dapat menjadi perhatian khusus bagi pemerintah setempat.

# Banyak kejadian banjir setiap bulan tahun 2020 di Jakarta



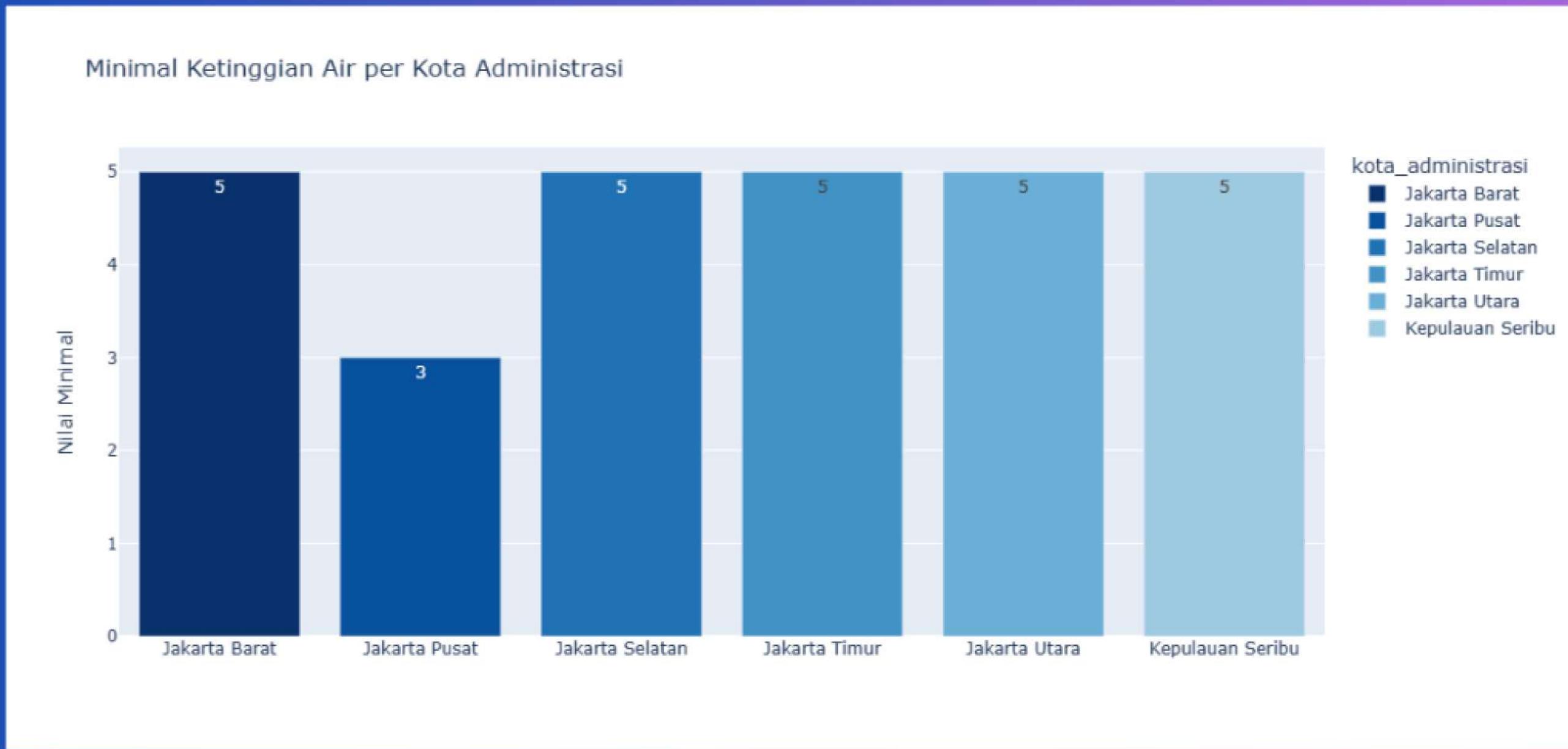
Line chart diatas menunjukan banyak kejadian banjir setiap bulan pada tahun 2020. Kita bisa melihat kejadian puncak banyak kejadian bencana banjir di Jakarta adalah pada bulan Februari dengan total lebih dari 160 kejadian atau tepatnya 167 kejadian. Kemudian grafik mulai menurun sampai bulan Juli dan kembali meningkat pada bulan september - agustus. Hal ini bisa menjadi perhatian pemerintah untuk antisipasi bencana banjir berdasarkan pola kejadian banjir setiap bulan di Jakarta pada tahun 2020.

# Ketinggian air maksimal berdasarkan kota



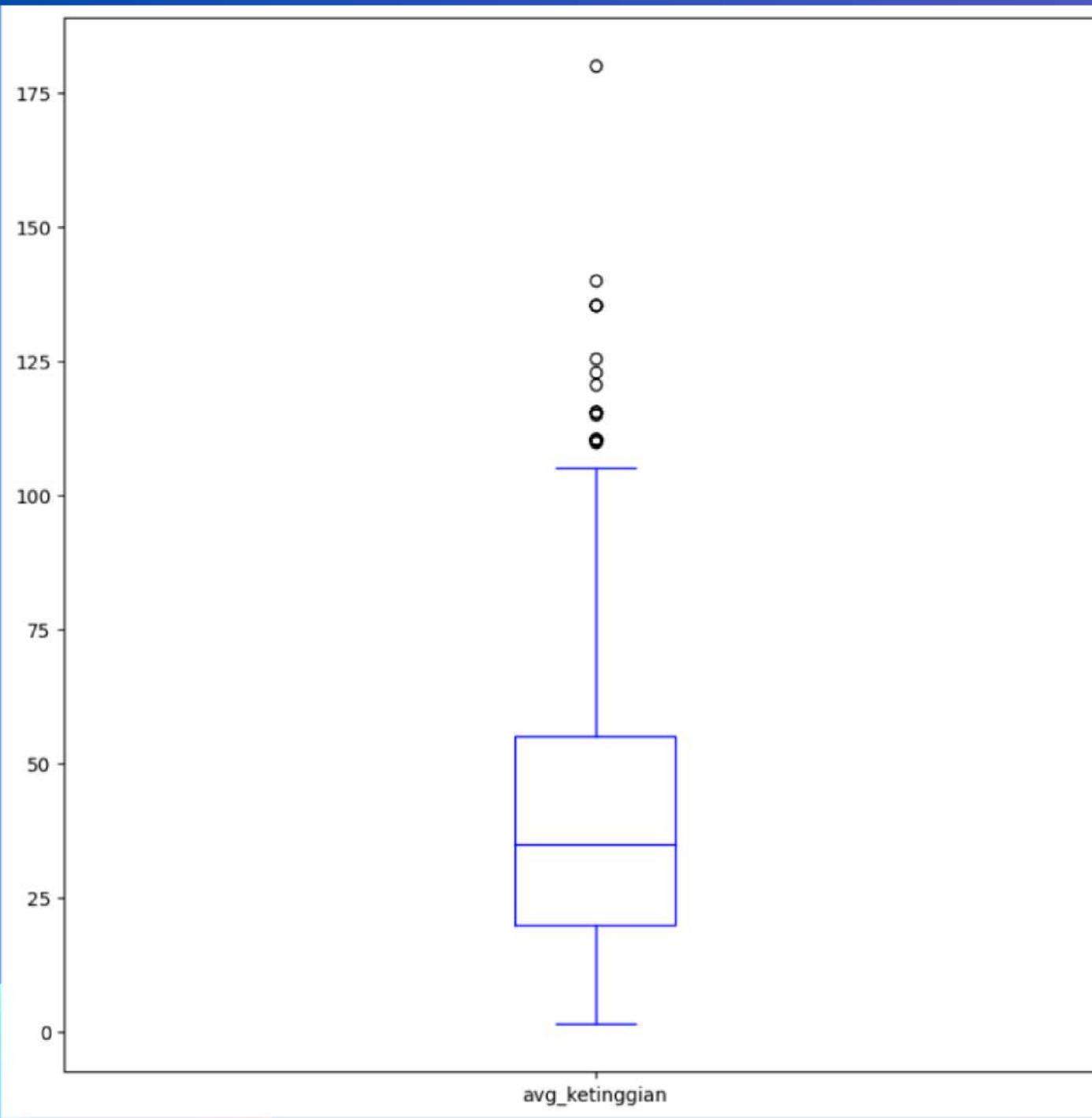
Maksimal ketinggian air pada saat banjir mencapai 350 cm yang terdapat di Kota Jakarta Timur.

# Ketinggian air minimal berdasarkan kota



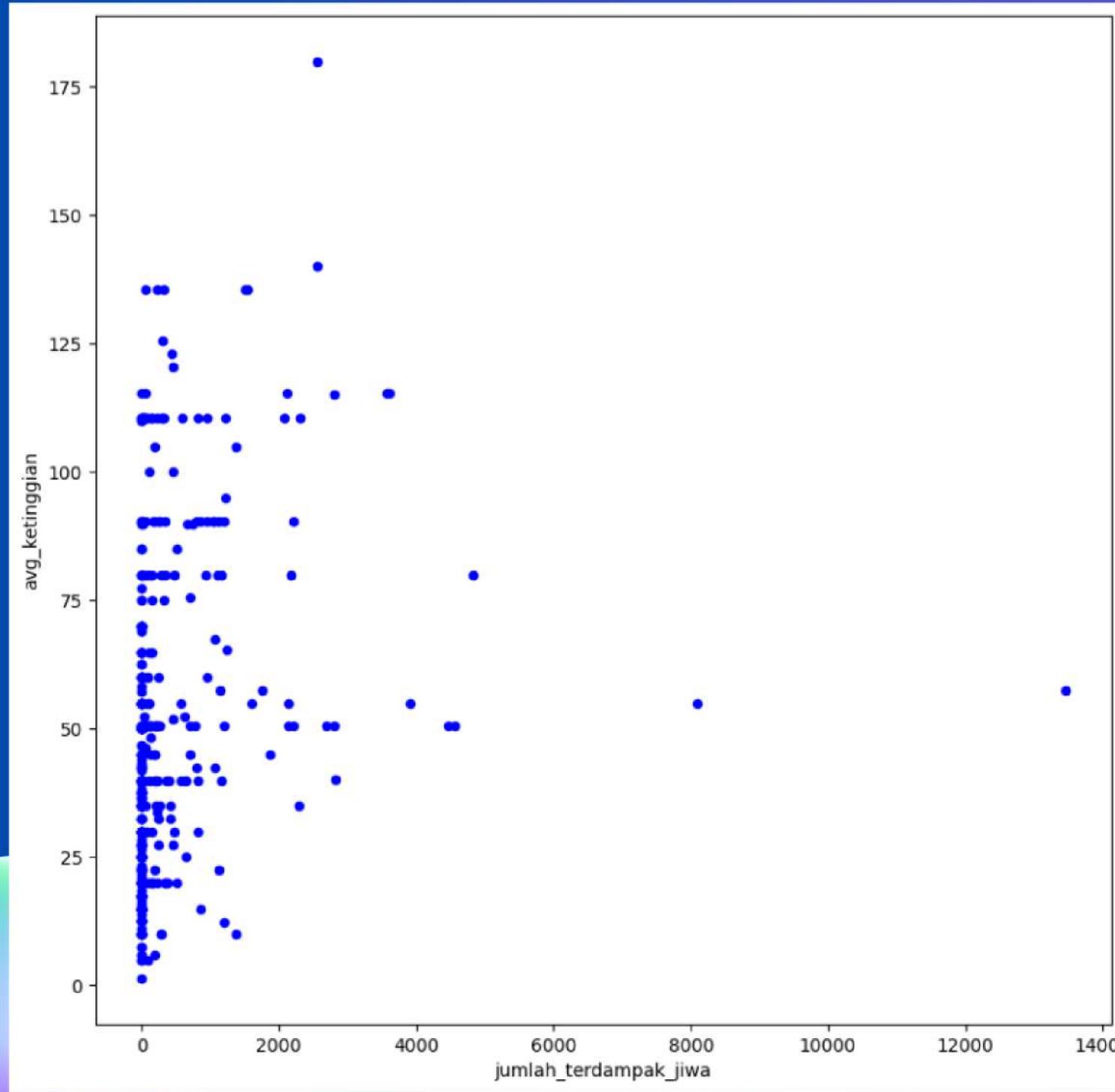
Minimal ketinggian air pada saat banjir sekitar 3 cm yang terdapat di Kota Jakarta Pusat.

# Persebaran data kolom "avg\_ketinggian" menggunakan Box Plot



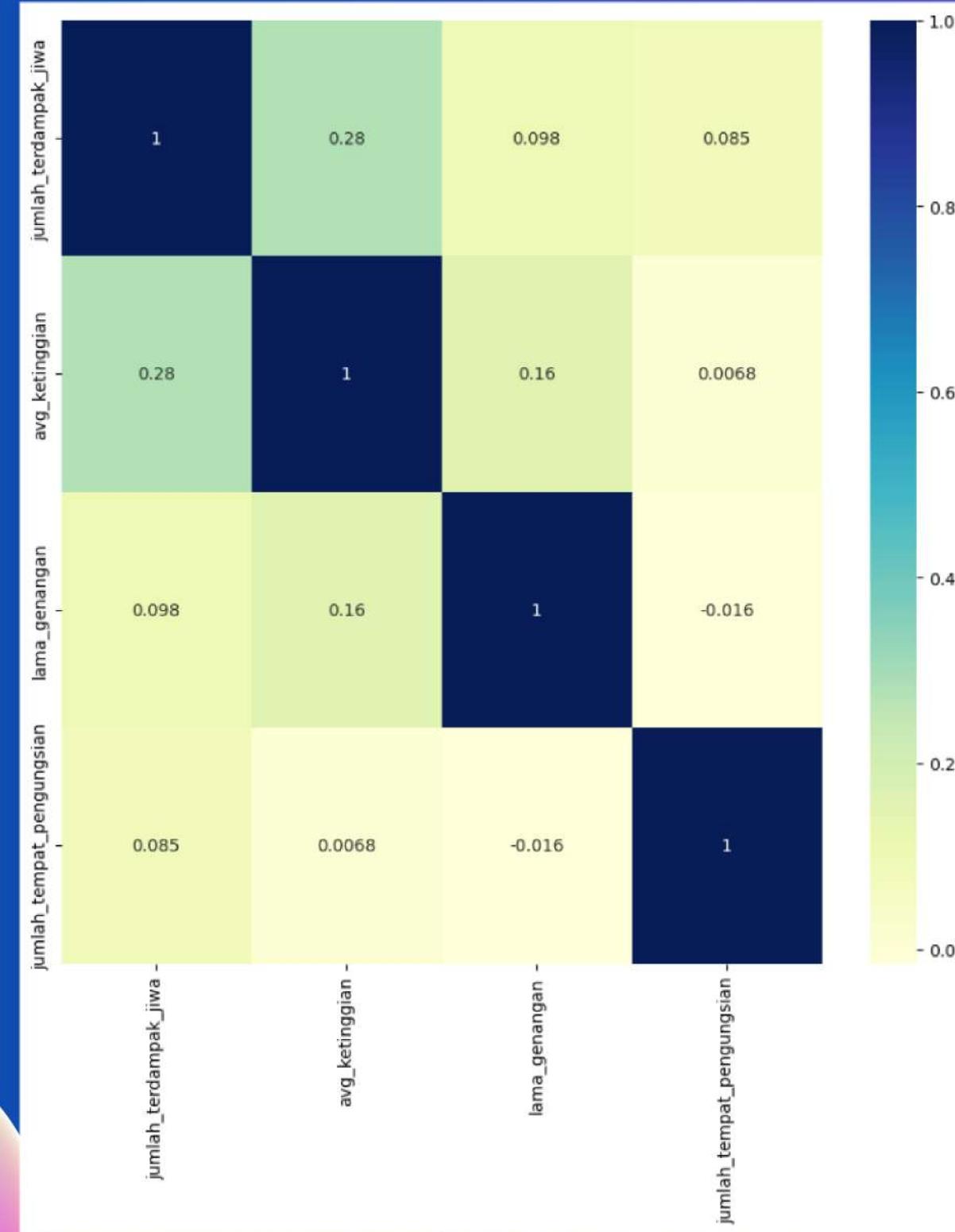
Dengan box plot ini kita dapat melihat distribusi data kolom ketinggian rata-rata air. Terdapat 8 pencilan dengan nilai pencilan maksimum sekitar 180 cm, Q3 dari data sekitar 55, nilai median sebesar 30, Q1 dari data sekitar 20 serta nilai nimimum sebesar 3 cm. Terlihat bahwa nilai median hampir simetris ditengah box. Distribusi simetris pada box plot menunjukkan bahwa data cenderung memiliki nilai-nilai yang seimbang di kedua sisi median.

# Hubungan atau pola antara 'jumlah\_terdampak\_jiwa' dan 'avg\_ketinggian' menggunakan scatter plot



Scatter plot disamping menunjukkan hubungan  
dua buah variabel numerik jumlah terdampak dan  
rata-rata ketinggian air.

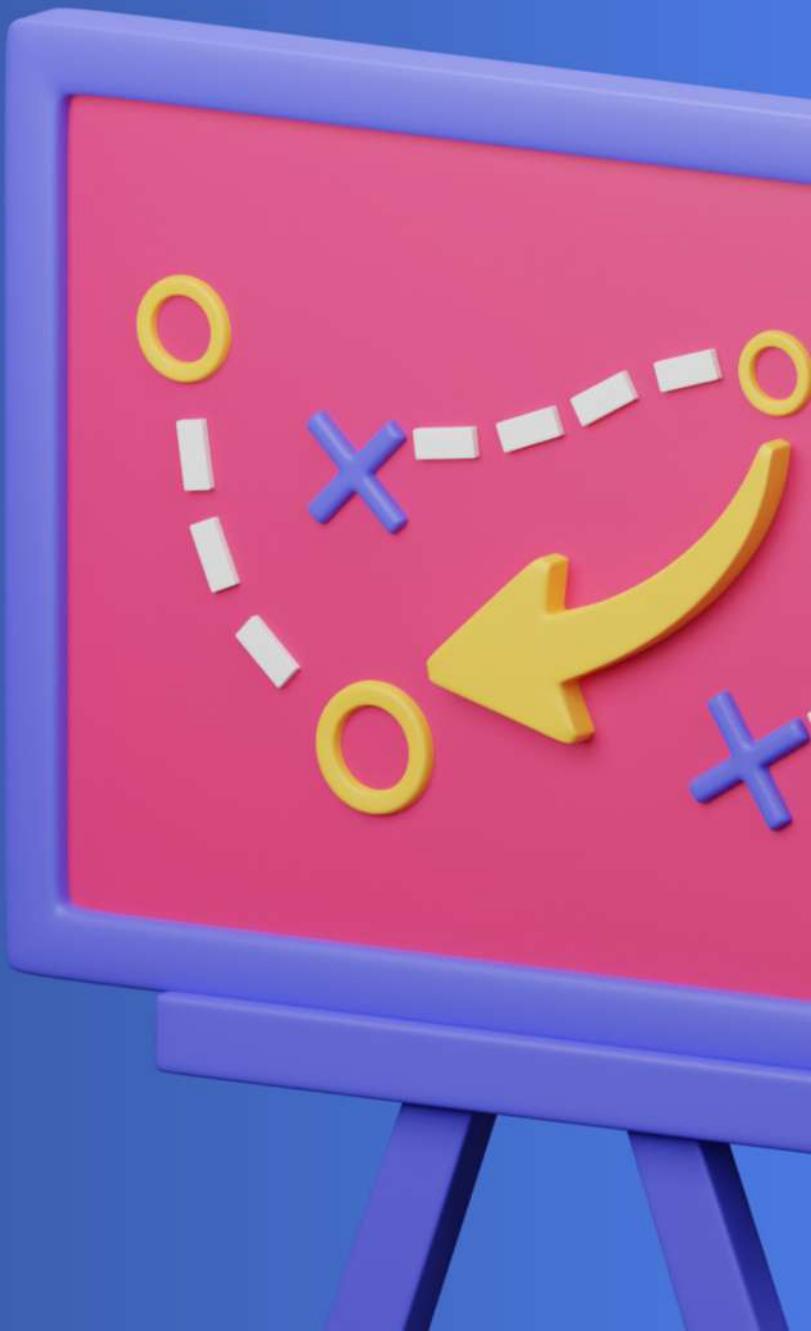
# Korelasi antar variabel



Grafik korelasi menunjukkan hubungan antara dua varaiel. Dimana nilainya diantara -1 sampai dengan 1. Semakin mendekati -1 atau 1, kedua variabel memiliki hubungan yang kuat (saling mempengaruhi). Terlihat bahwa korelasi antara avg\_ketinggian dan jumlah\_terdampak\_jiwa memiliki korelasi yang lebih besar dibandingkan dengan nilai korelasi variabel lainnya, yaitu sebesar 0,28.

# Analisis Mendalam (Modeling)

- Pemilihan Metode dan Variabel
- *Feature Engineering*
- *Hyper-parameter Tuning*
  - *K-means Clustering* tanpa PCA
  - *K-means Clustering* dengan PCA
- Evaluasi Model



# Pemilihan Metode dan Variabel

Metode yang kami pilih adalah *K-means Clustering* dengan tujuan untuk melakukan pengelompokan data tingkat banjir di Jakarta pada tahun 2020.

Berdasarkan hasil eksplorasi data sebelumnya kami memilih variabel avg\_ketinggian, jumlah\_terdampak\_jiwa, lama\_genangan dan jumlah\_tempat\_pengungsian yang digunakan untuk melakukan clustering. Hal ini atas pertimbangan korelasi yang lebih tinggi dibandingkan dengan variabel lainnya. Dengan begitu pola hubungan antara variabel saat dilakukan clustering akan menghasilkan pengelompokan yang mendekati sesuai.

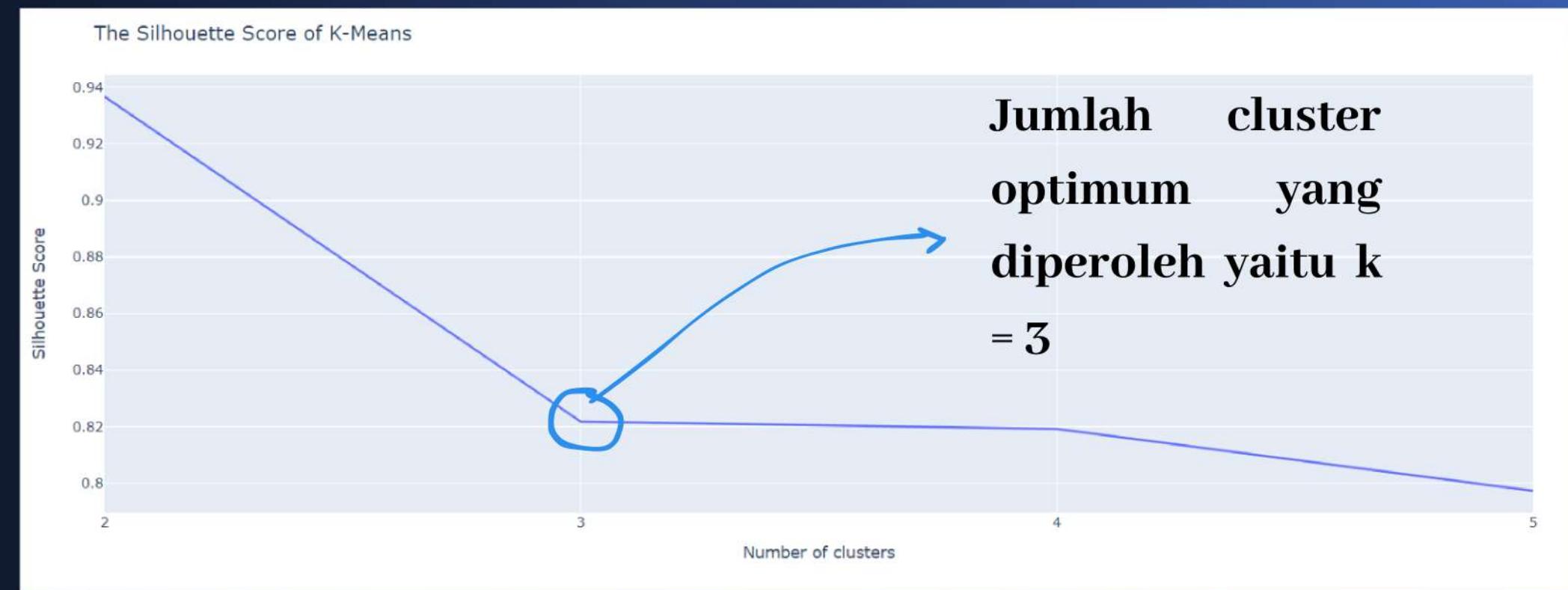
# *Feature Engineering*

*Feature Engineering* ini sudah kami lakukan pada saat data pre-processing dimana kami menambahkan fitur (variabel) baru yaitu avg\_ketinggian yang merupakan kolom rata-rata ketinggian air banjir dan menghapus beberapa variabel yang dirasa kurang relevan dengan tujuan dalam perumusan masalah, sehingga fitur-fitur yang digunakan menjadi lebih representatif, dan mengandung informasi yang berguna.

# K-means Clustering tanpa PCA

K-means clustering tanpa PCA artinya kita akan menggunakan semua data dari variabel yang sudah didefinisikan sebelumnya yaitu avg\_ketinggian, jumlah\_terdampak\_jiwa, lama\_genangan dan jumlah\_tempat\_pengungsian untuk dilakukan proses clustering.

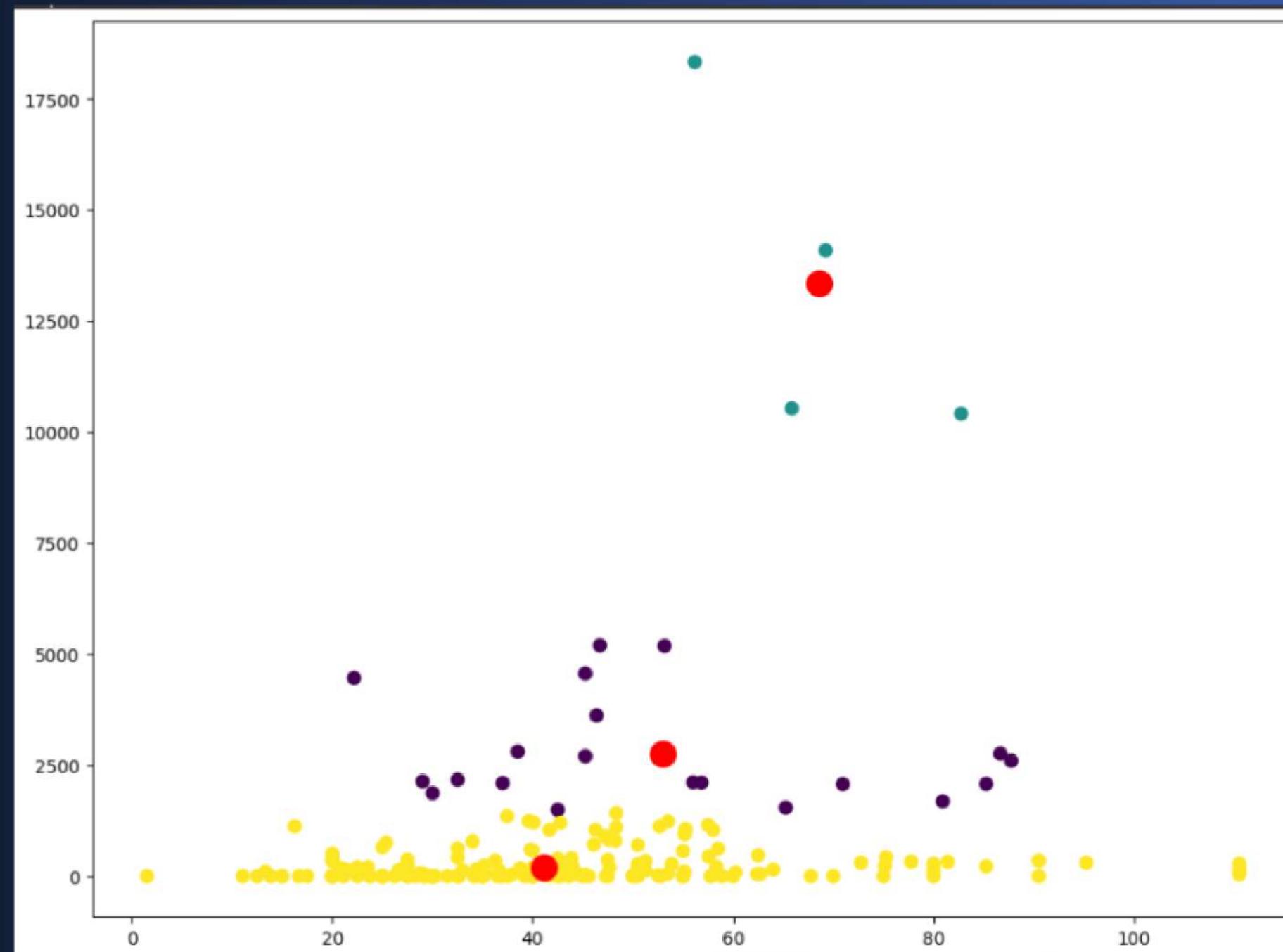
## Penentuan banyaknya cluster



Metode yang digunakan dalam menentukan banyak cluster adalah Metode Silhouette. Metode ini menghitung skor Silhouette untuk setiap titik data, yang merupakan ukuran seberapa dekat titik data dengan titik-titik lain dalam cluster yang sama dibandingkan dengan titik-titik di cluster lain.

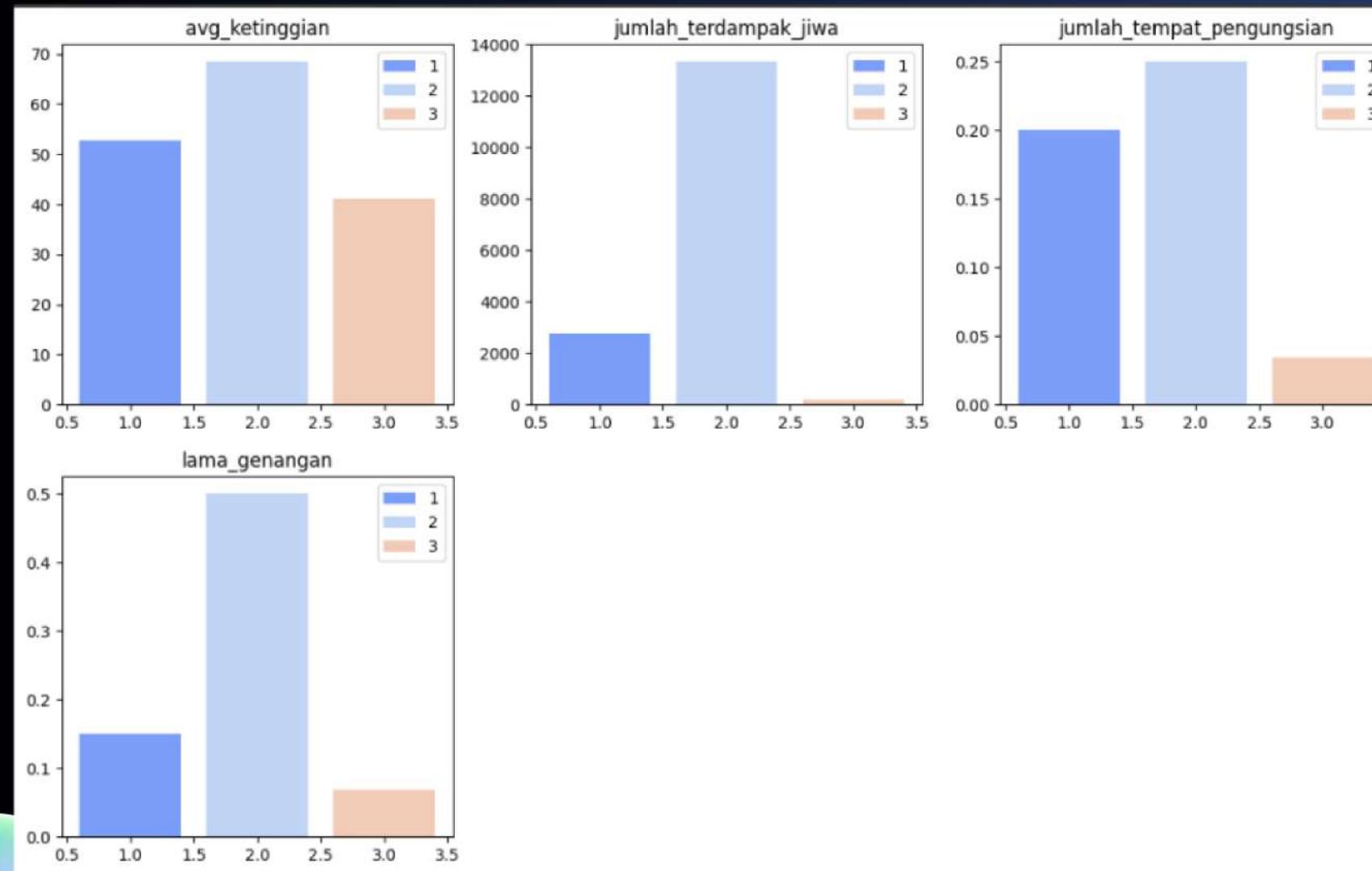
# K-means Clustering tanpa PCA

Melakukan Prediksi Cluster Terhadap Data



# K-means Clustering tanpa PCA

Melihat distribusi cluster berdasarkan masing-masing variabel



**Cluster 1 = Ringan.** Terjadi pada rata-rata ketinggian kurang dari 60 cm, jumlah terdampak jiwa kurang dari 3000 orang, rata-rata jumlah tempat pengungsian kurang dari 1 dan rata-rata lama genangan kurang dari 1 hari.

**Cluster 2 = Sedang.** Terjadi pada rata-rata ketinggian kurang dari 50 cm, jumlah terdampak jiwa kurang dari 1000 orang, rata-rata jumlah tempat pengungsian kurang dari 1 dan rata-rata lama genangan kurang dari 1 hari.

**Cluster 3 = Berat.** Terjadi pada rata-rata ketinggian kurang dari 70 cm, jumlah terdampak jiwa kurang dari 14000 orang, rata-rata jumlah tempat pengungsian kurang dari 1 dan rata-rata lama genangan kurang dari 1 hari.

# K-means Clustering dengan PCA

*K-means clustering* dengan PCA artinya kita akan mereduksi dimensi data dari variabel yang sudah didefinisikan sebelumnya yaitu avg\_ketinggian, jumlah\_terdampak\_jiwa, lama\_genangan dan jumlah\_tempat pengungsian sehingga diperoleh variabel yang signifikan saja dengan asumsi akan menghasilkan prediksi cluster yang lebih baik.

Data yang akan digunakan pertama dilakukan proses PCA terlebih dahulu:

```
X_scaled = preprocessing.scale(df3_drop)
pca = PCA(n_components=2)
df4 = pca.fit_transform(X_scaled)
df4 = pd.DataFrame(df4)
```

# K-means Clustering dengan PCA

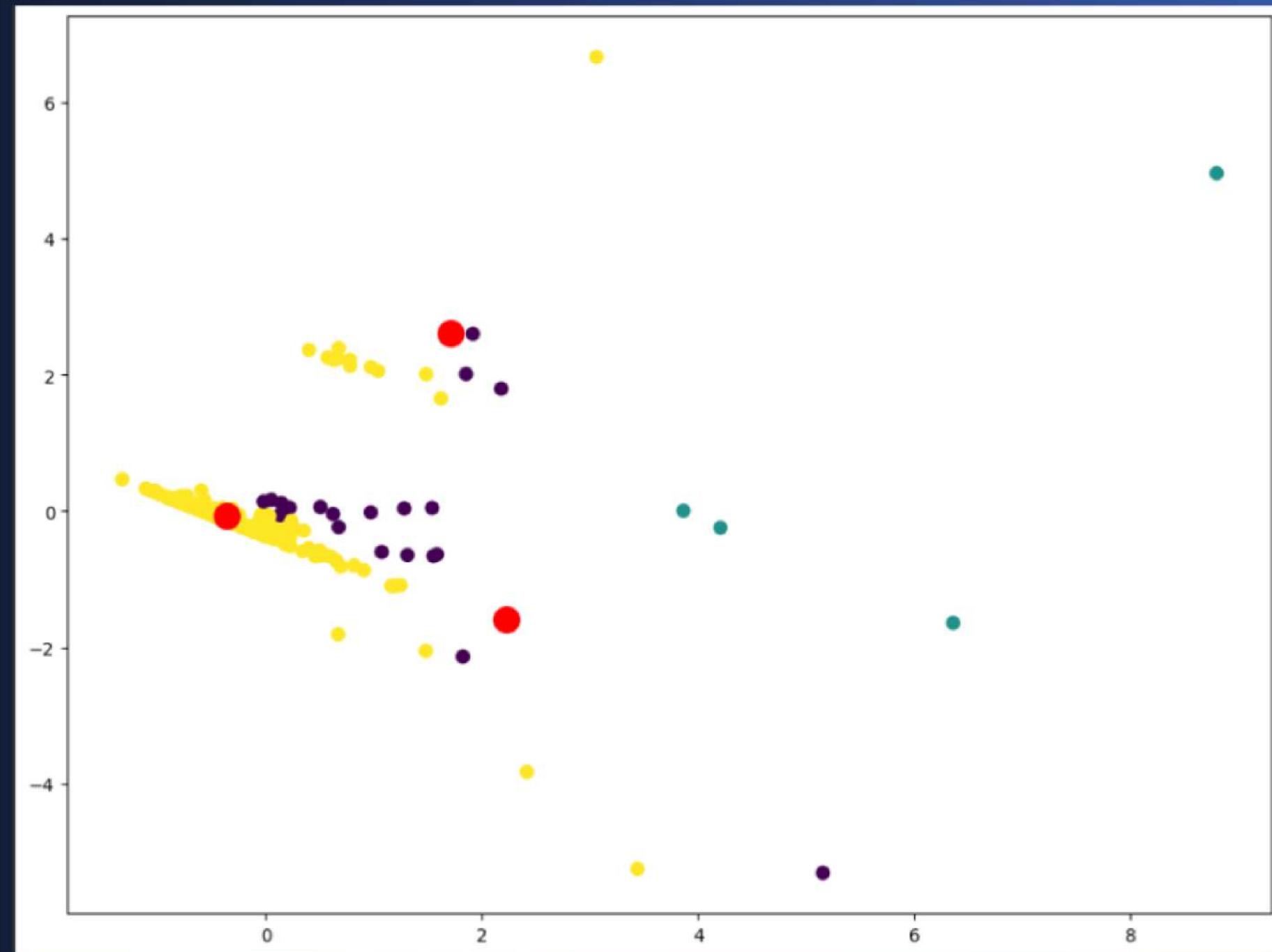
## Penentuan banyaknya cluster



Metode yang digunakan sama seperti sebelumnya yaitu Silhouette. Metode ini menghitung skor Silhouette untuk setiap titik data, yang merupakan ukuran seberapa dekat titik data dengan titik-titik lain dalam cluster yang sama dibandingkan dengan titik-titik di cluster lain.

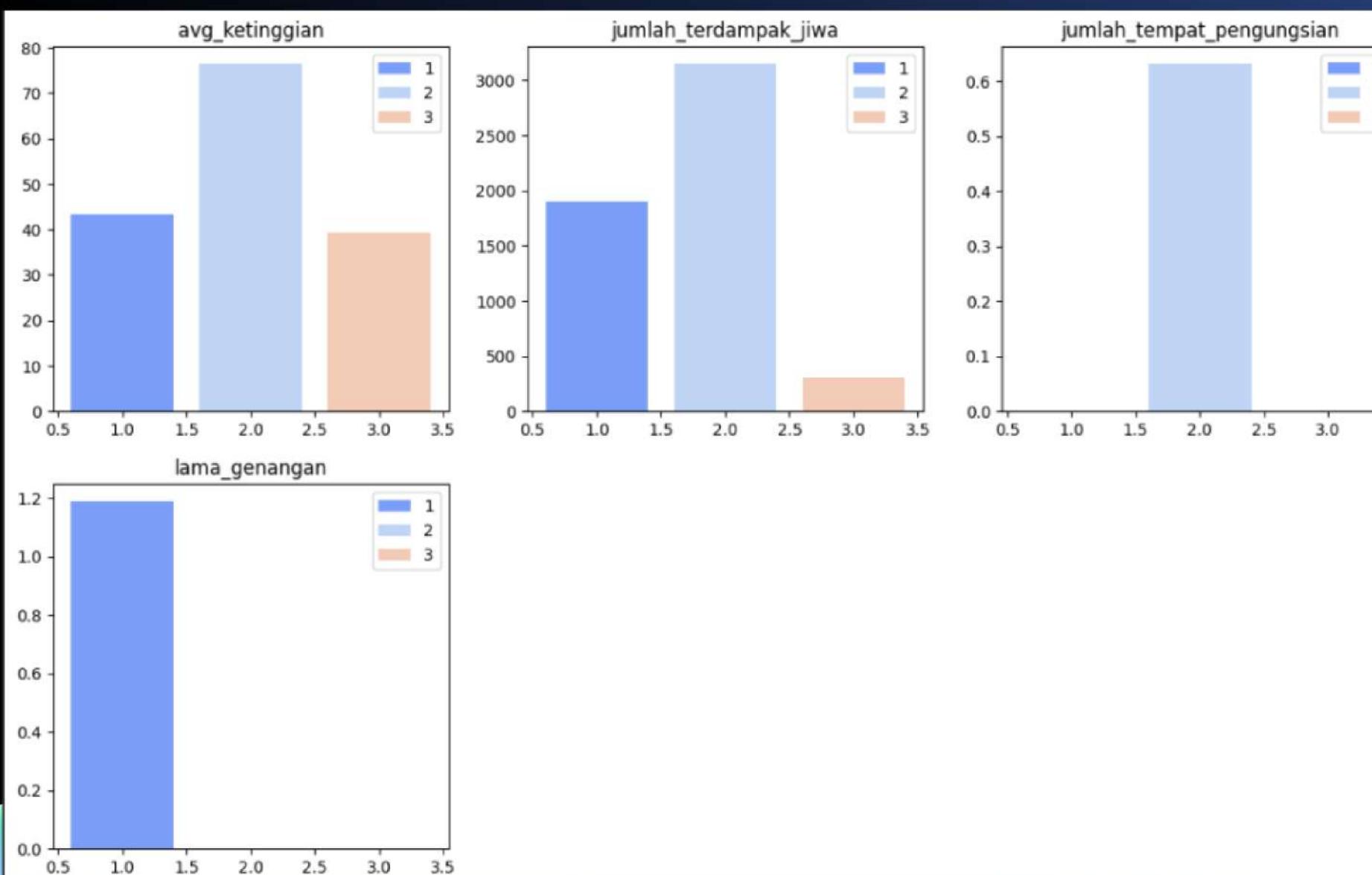
# K-means Clustering dengan PCA

Melakukan Prediksi Cluster Terhadap Data



# K-means Clustering dengan PCA

Melihat distribusi cluster berdasarkan masing-masing variabel



**Cluster 1 = Ringan.** Terjadi pada rata-rata ketinggian kurang dari 50 cm, jumlah terdampak jiwa kurang dari 2000 orang, rata-rata jumlah tempat pengungsian kurang dari 1 dan rata-rata lama genangan kurang dari 2 hari.

**Cluster 2 = Sedang.** Terjadi pada rata-rata ketinggian kurang dari 50 cm, jumlah terdampak jiwa kurang dari 1000 orang, rata-rata jumlah tempat pengungsian kurang dari 1 dan rata-rata lama genangan kurang dari 1 hari.

**Cluster 3 = Berat.** Terjadi pada rata-rata ketinggian lebih dari 60 cm, jumlah terdampak jiwa lebih dari 5000 orang, rata-rata jumlah tempat pengungsian kurang dari 2 dan rata-rata lama genangan kurang dari 1 hari.

# Evaluasi Model

## Silhouette Score

Evaluasi model menjadi tantangan karena tidak ada label sebenarnya yang bisa kita bandingkan dengan hasil clustering. Namun, kita bisa menggunakan metrik seperti Silhouette Score untuk mengevaluasi seberapa baik hasil clustering.

```
from sklearn.metrics import silhouette_score

# Evaluasi Silhouette Coefficient tanpa PCA
silhouette_coefficient = silhouette_score(df2_drop, y_kmeans)
print("Silhouette Coefficient tanpa PCA:", silhouette_coefficient)

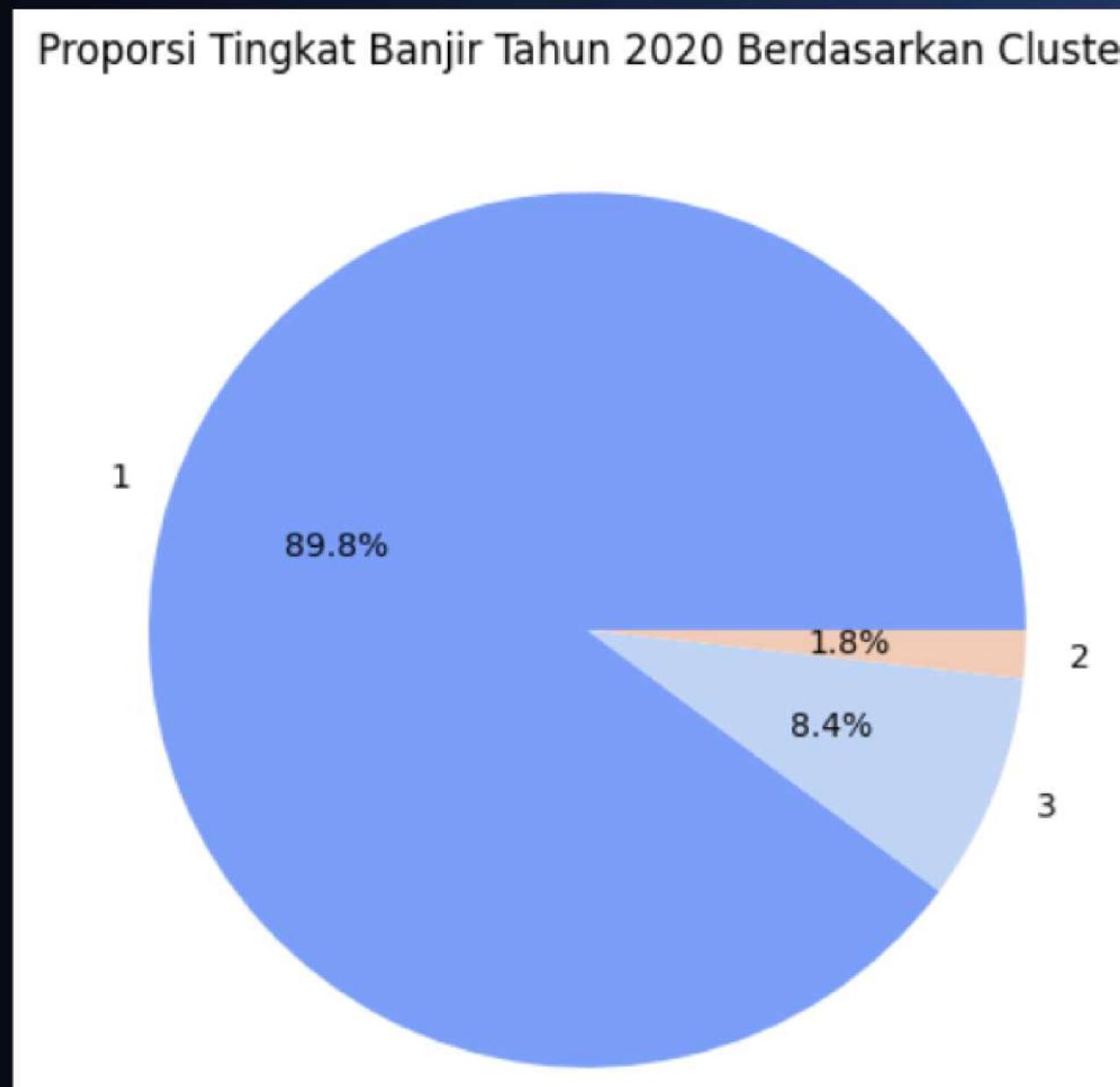
# Evaluasi Silhouette Coefficient dengan PCA
silhouette_coefficient = silhouette_score(df4, y_kmeans_pca)
print("Silhouette Coefficient dengan PCA:", silhouette_coefficient)

Silhouette Coefficient tanpa PCA: 0.8217032836689135
Silhouette Coefficient dengan PCA: 0.7140750249540224
```

Dari hasil evaluasi menggunakan Silhouette Coefficient, ditemukan bahwa pengelompokan tanpa menggunakan PCA menghasilkan nilai yang lebih tinggi (0.8217) daripada pengelompokan dengan menggunakan PCA (0.7455). Hal ini menunjukkan bahwa pengelompokan langsung pada dataset asli, tanpa mengurangi dimensi menggunakan PCA, menghasilkan hasil clustering yang lebih baik dalam hal kualitas pengelompokan. Meskipun nilai Silhouette Coefficient dengan PCA masih cukup tinggi, perbedaan yang signifikan antara dua metode menunjukkan keunggulan pengelompokan tanpa menggunakan PCA dalam kasus ini.

# Evaluasi Model

Proporsi cluster tanpa PCA



Proporsi cluster dengan PCA



# Kesimpulan dan Rekomendasi

Didapatkan hasil temuan menarik dari analisis pola dan pengelompokan tingkat banjir di Jakarta pada tahun 2020 diantaranya:

- **Jakarta Timur, Jakarta Selatan dan Jakarta Barat menjadi Kota yang paling sering terjadi bencana banjir di Provinsi DKI Jakarta.**
- **Kelurahan Kobon Baru, Cawang dan Kampung Melayu menjadi kelurahan yang paling sering terjadi bencana banjir ditahun 2020 dengan total 9 kejadian untuk setiap kelurahan.**
- **Kejadian banjir di Jakarta puncaknya yaitu pada bulan Januari - Februari.**
- **Ketinggian air banjir maksimum mencapai 350 cm dan ketinggian minimum sekitar 3 cm**
- **Cluster optimal tanpa PCA dan dengan PCA yaitu  $k = 3$ . Diperoleh nilai Silhouette Score tanpa PCA > Silhouette Score dengan PCA.**  
Hal ini menunjukkan bahwa pengelompokan langsung pada dataset asli, tanpa mengurangi dimensi menggunakan PCA, menghasilkan hasil clustering yang lebih baik dalam hal kualitas pengelompokan.
- **Proporsi hasil cluster menunjukan cluster 1 (Ringan) lebih dominan terjadi diantara dua cluster lainnya.**

Perhatian khusus pada wilayah Jakarta Timur, Jakarta Selatan dan Jakarta Barat terutama di kelurahan Kobon Baru, Cawang, dan Kampung Melayu dalam menghadapi banjir. Pemerintah daerah dapat meningkatkan pemeliharaan sistem dan infrastruktur yang ada di wilayah-wilayah ini, dikarenakan merupakan wilayah yang paling sering terjadi banjir di Jakarta pada tahun 2020.

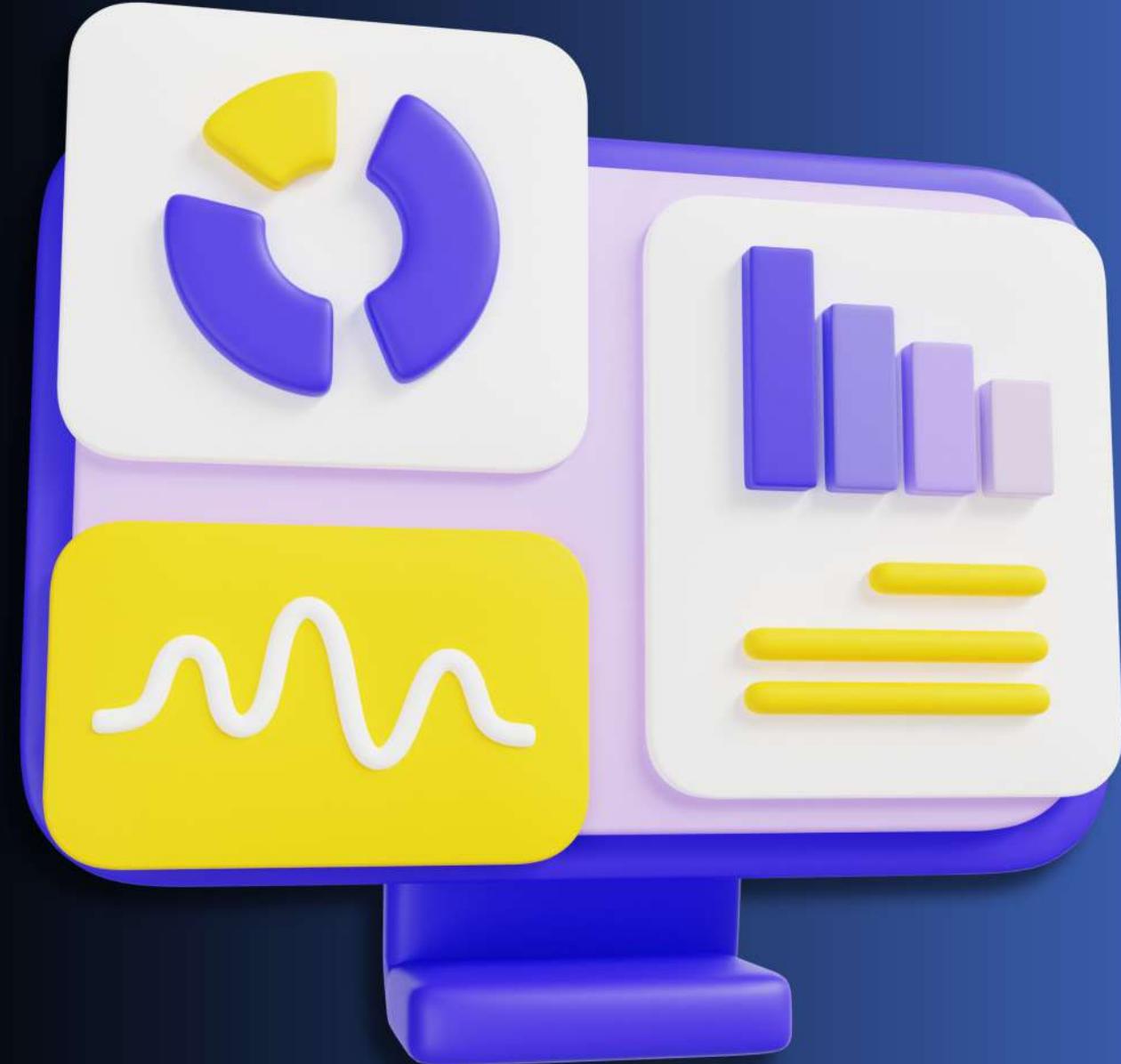
Dalam mempersiapkan diri menghadapi banjir, perlu dilakukan perencanaan yang matang, terutama pada bulan-bulan di mana puncak banjir terjadi yaitu di awal tahun (Januari-Februari). Pemerintah daerah dapat meningkatkan sistem peringatan dini, evakuasi, dan penyaluran bantuan pada periode ini. Pendidikan masyarakat tentang tindakan pencegahan banjir juga perlu ditingkatkan menjelang periode ini. Walaupun proporsi cluster 1 (Ringan) banjir di Jakarta paling dominan, hal ini tetap diperlukan kewaspadaan terhadap risiko banjir dan melakuakan mitigasi serta langkah stategis berdasarkan data hasil clustering tingkat banjir di Jakarta.



## Google Collaboratory



<http://bit.ly/INFINITY-IPYNB>



Link Live Dashboard Tableau Data Banjir DKI Jakarta  
2018-2020



<https://bit.ly/INFINITY-TABLEAU>

# **Daftar Pustaka**

**Agusta, Y. 2007. K-Means-Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol.3 , 47- 60.**

**Dunteman, H. George ,(1989). Principal Component Analysis. Sage Publications., Newbury Park London New Delhi. (Reseach Triangle Institute).**

**Agarwal, V. (2015). Research on data preprocessing and categorization technique for smartphone review analysis. International Journal of Computer Applications, 131(4), 30-36.**

**J. H. Maindonald and W. J. Braun, Data Analysis and Graphics Using R: An Example-Based Approach. Cambridge: Cambridge University Press, 2020.**