CLUSTERING DATA COVID-19 DENGAN K-MEANS

Kelompok 4

Table of contents

01 Pendahuluan

02 Metode

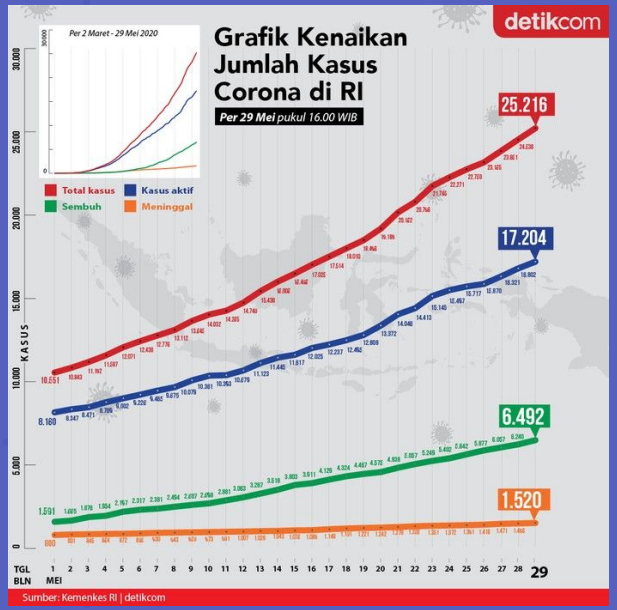
03 Hasil dan Pembahasan

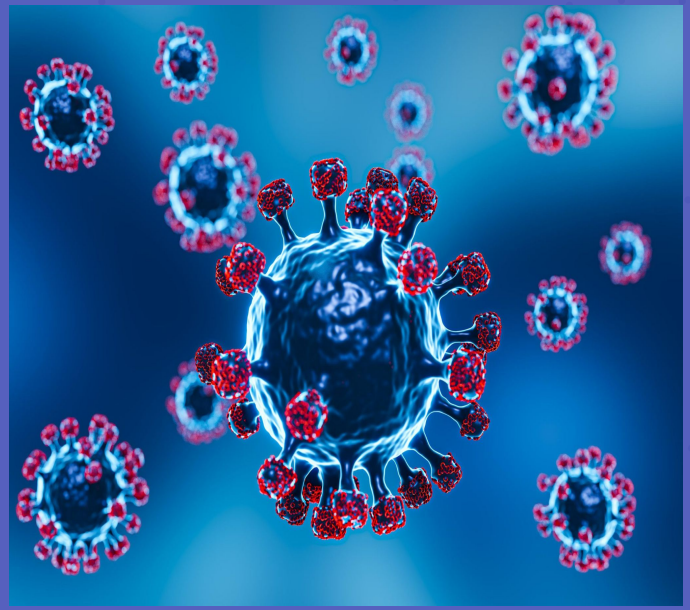
04 Kesimpulan

01

Pendahuluan

Latar Belakang





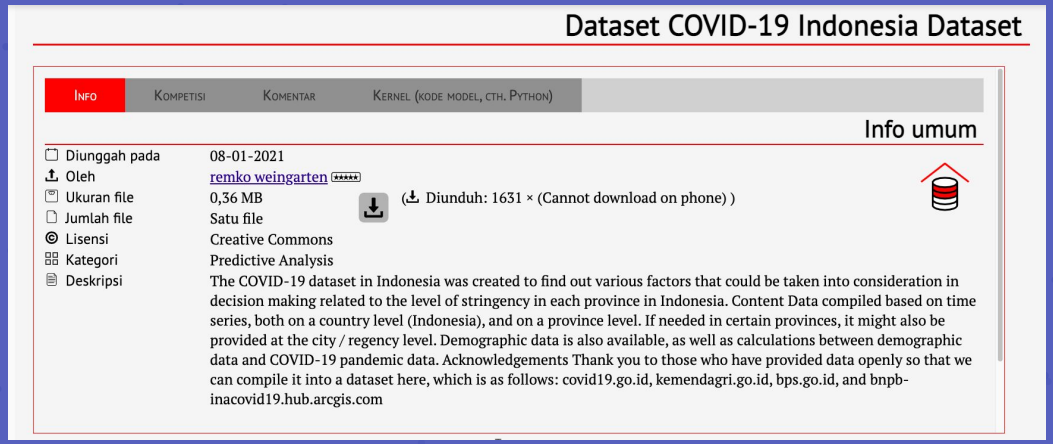
Rumusan Masalah

Bagaimana hasil dari Pengclusteran ?

Provinsi mana yang jumlah kasus tinggi

Provinsi mana yang jumlah kasus rendah?

Deskripsi Dataset



Dataset COVID-19 Indonesia Dataset

INFO | KOMPETISI | KOMENTAR | KERNEL (KODE MODEL, CTH. PYTHON)

Info umum

Diunggah pada	08-01-2021
Oleh	remko weingarten
Ukuran file	0,36 MB
Jumlah file	Satu file
Lisensi	Creative Commons
Kategori	Predictive Analysis
Deskripsi	The COVID-19 dataset in Indonesia was created to find out various factors that could be taken into consideration in decision making related to the level of stringency in each province in Indonesia. Content Data compiled based on time series, both on a country level (Indonesia), and on a province level. If needed in certain provinces, it might also be provided at the city / regency level. Demographic data is also available, as well as calculations between demographic data and COVID-19 pandemic data. Acknowledgements Thank you to those who have provided data openly so that we can compile it into a dataset here, which is as follows: covid19.go.id, kemendagri.go.id, bps.go.id, and bnpb-inacovid19.hub.arcgis.com

(Diunduh: 1631 × (Cannot download on phone))

Maret - Desember 2020

Deret Waktu

atapdata.ai

Sumber

9659 Baris

Jumlah Baris

35 Kolom

Jumlah Baris

Deskripsi Dataset

	Date	Location ISO Code	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	...	Longitude	Latitude	New Cases per Million	Total Cases per Million	New Deaths per Million	Total Deaths per Million	Case Fatality Rate	Case Recovered Rate
0	3/1/2020	ID-JK	DKI Jakarta	2	0	0	2	489	20	39	...	106.836118	-6.204699	0.18	45.09	0.0	1.84	4.09%	7.98%
1	3/2/2020	ID-JK	DKI Jakarta	2	0	0	2	491	20	39	...	106.836118	-6.204699	0.18	45.27	0.0	1.84	4.07%	7.94%
2	3/2/2020	IDN	Indonesia	2	0	0	2	2	0	0	...	113.921327	-0.789275	0.01	0.01	0.0	0.00	0.00%	0.00%
3	3/2/2020	ID-JB	Jawa Barat	3	0	0	3	12	5	120	...	107.603708	-6.920432	0.07	0.27	0.0	0.11	41.67%	1000.00%
4	3/2/2020	ID-RI	Riau	1	0	0	1	2	1	1	...	101.805109	0.511648	0.16	0.33	0.0	0.16	50.00%	50.00%

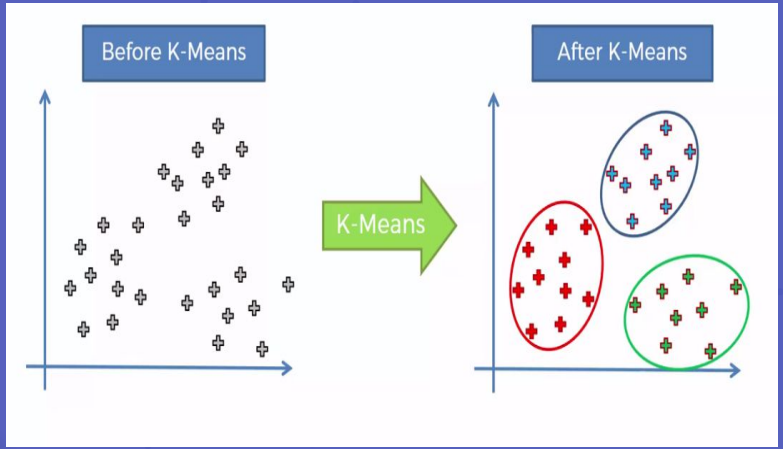
5 rows × 37 columns

02

Metode

Gambaran Umum

K-Means merupakan salah satu algoritma dalam data mining yang digunakan untuk melakukan pengelompokan/clustering suatu data. Tujuan dari algoritma K-Means adalah meminimalkan fungsi dengan meminimalkan variasi antar data dalam suatu cluster dan memaksimalkan variasi data pada cluster lainnya.

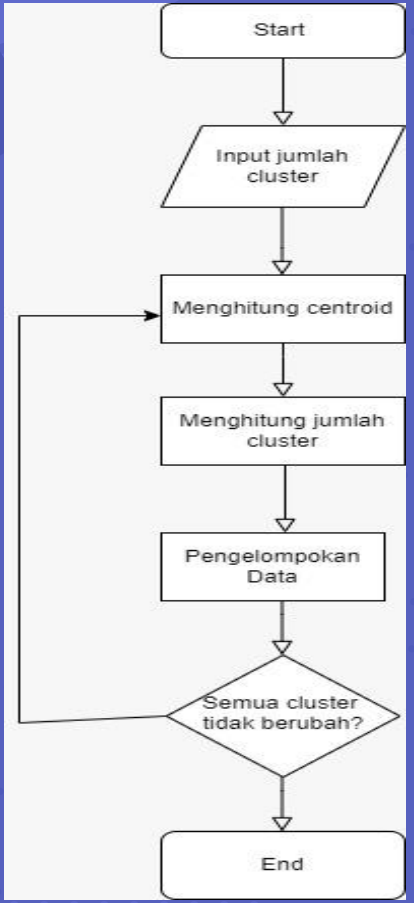


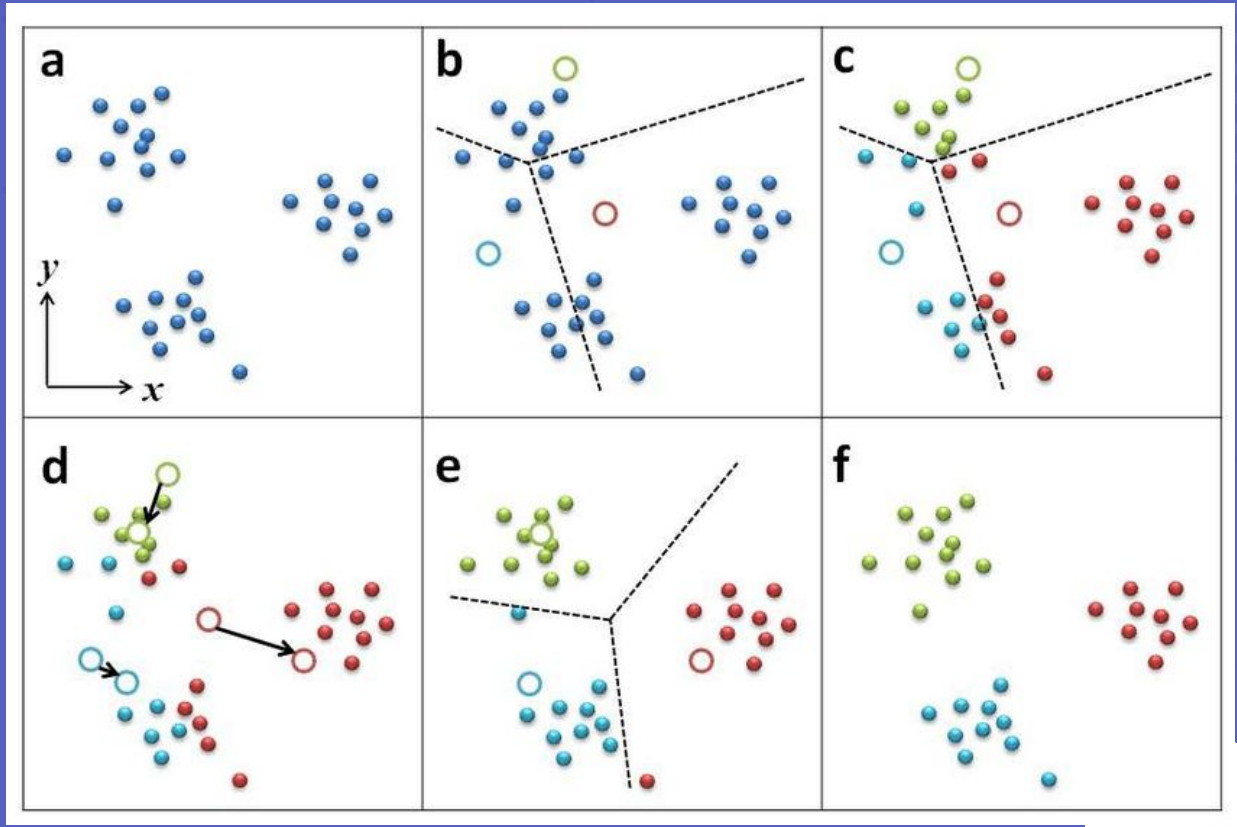
Kmeans

Clustering

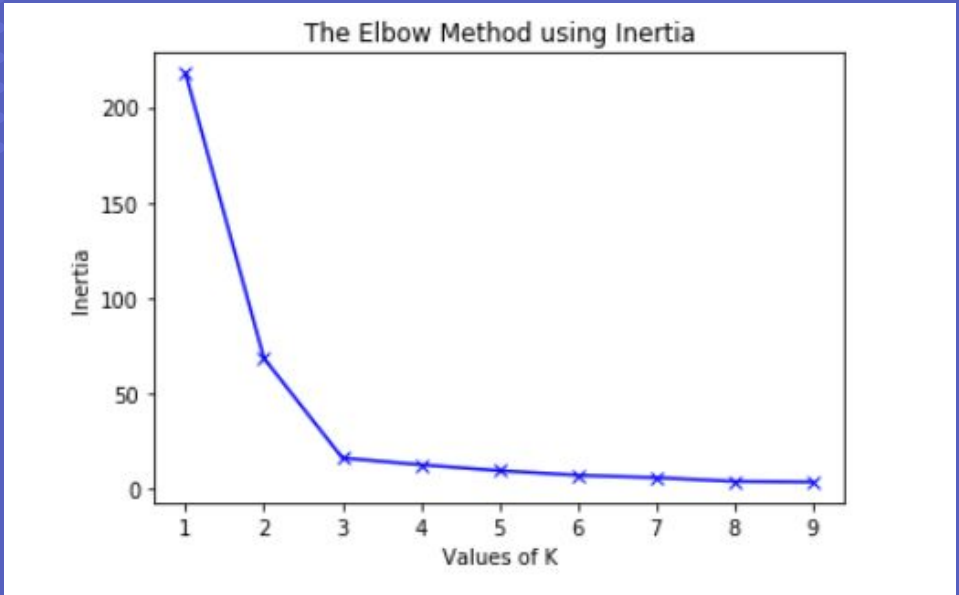
Unsupervised Learning

Flowchart Sistem





Metrics Evaluation



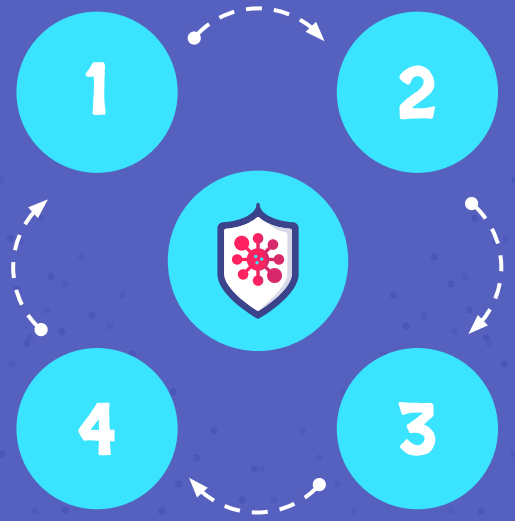
Metode Elbow

03

Hasil dan Pembahasan

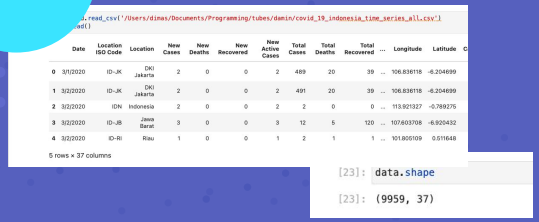
Tahapan yang Dilakukan

1. Pengumpulan Data
2. Processing Data
3. Visualisasi Data
4. Modeling, Pengujian, Analisis



Memproses Data

1



```
[23]: data.shape
[23]: (9959, 37)
```

Load Dataset

2

```
[3]: class ColumnData:
         date = 'Date'
         province = 'Province'
         island = 'Island'
         cases = 'Total Cases'
         deaths = 'Total Deaths'
         recovered = 'Total Recovered'
         actives_cases = 'Total Active Cases'
         population = 'Population'
         area = 'Area (km2)'
         mortality = 'Mortality'
         density = 'Population Density'
```

Pendefinisian Nama Kolom

3

Memilih kolom yang akan digunakan

```
[34]: data = data[[
          ColumnData.date,
          ColumnData.province,
          ColumnData.island,
          ColumnData.cases,
          ColumnData.deaths,
          ColumnData.recovered,
          ColumnData.actives_cases,
          ColumnData.population,
          ColumnData.area,
          ColumnData.density
      ]]
```

```
[38]: data.shape
[38]: (9959, 10)
```

```
[35]: data
```

[35]:

	Date	Province	Island	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Population	Area (km2)	Population Density
0	3/1/2020	DKI Jakarta	Jawa	489	20	39	430	10846145	664	16334.31
1	3/2/2020	DKI Jakarta	Jawa	491	20	39	432	10846145	664	16334.31
2	3/2/2020	NaN	NaN	2	0	0	2	265185520	1916907	138.34
3	3/2/2020	Jawa Barat	Jawa	12	5	120	-113	45161325	35378	1276.55
4	3/2/2020	Riau	Sumatera	2	1	1	0	6074100	87024	69.80
...	...	...	...	...	...	...	...	...	...	...
9954	12/31/2020	Sulawesi Tenggara	Sulawesi	7907	147	6696	1064	2635461	38068	69.23
9955	12/31/2020	Sulawesi Utara	Sulawesi	9671	310	7003	2358	2641884	13892	190.17
9956	12/31/2020	Sumatera Barat	Sumatera	23464	522	18030	4912	5519245	42013	131.37
9957	12/31/2020	Sumatera Selatan	Sumatera	11826	599	9364	1863	8217551	91592	89.72
9958	12/31/2020	Sumatera Utara	Sumatera	18149	679	15403	2067	14874889	72981	203.82

9959 rows × 10 columns

Memproses Data

4

```
]: data.isnull().sum()

51]: Date                    0
     Province              305
     Island                305
     Total Cases             0
     Total Deaths            0
     Total Recovered         0
     Total Active Cases      0
     Population              0
     Area (km2)              0
     Population Density      0
     dtype: int64
```

Menghapus jika ada data yang kosong

```
[9]: data = data.dropna(axis=0, how="any")
```

Memeriksa dan Menghapus Null Value

5

```
]: data
```

	Date	Province	Island	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Population	Area
0	3/1/2020	DKI Jakarta	Jawa	489	20	39	430	10846145	
1	3/2/2020	DKI Jakarta	Jawa	491	20	39	432	10846145	
3	3/2/2020	Jawa Barat	Jawa	12	5	120	-113	45161325	
4	3/2/2020	Riau	Sumatera	2	1	1	0	6074100	
5	3/3/2020	DKI Jakarta	Jawa	493	20	39	434	10846145	
...	...	...	...	...	...	...	...	...	...
9954	12/31/2020	Sulawesi Tenggara	Sulawesi	7907	147	6696	1064	2635461	
9955	12/31/2020	Sulawesi Utara	Sulawesi	9671	310	7003	2358	2641884	
9956	12/31/2020	Sumatera Barat	Sumatera	23464	522	180			
9957	12/31/2020	Sumatera Selatan	Sumatera	11826	599	93			
9958	12/31/2020	Sumatera Utara	Sumatera	18149	679	154			

9654 rows × 10 columns

```
[58]: data.shape

[58]: (9654, 10)
```

Melihat Data yang Sekarang

6

Membuat format data pada tanggal menjadi date.

```
[39]: data[ColumnData.date] = pd.to_datetime(data.Date, infer_datetime_format=True).dt.date
```

Membuat kolom baru yaitu 'Mortality'

```
[41]: data[ColumnData.mortality] = data[ColumnData.deaths] / data[ColumnData.cases]
```

Membuat Kolom Baru, Mengubah Format Data

Memproses Data

7

```
[ ]: data
```

[2]:

	Date	Province	Island	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Population	Area (km2)	Population Density	Mortality
0	2020-03-01	DKI Jakarta	Jawa	489	20	39	430	10846145	664	16334.31	0.040900
1	2020-03-02	DKI Jakarta	Jawa	491	20	39	432	10846145	664	16334.31	0.040733
2	2020-03-02	NaN	NaN	2	0	0	2	265185520	1916907	138.34	0.000000
3	2020-03-02	Jawa Barat	Jawa	12	5	120	-113	45161325	35378	1276.55	0.416667
4	2020-03-02	Riau	Sumatera	2	1	1	0	6074100	87024	69.80	0.500000
...	...	...	...	...	...	...	...	...	...	...	...
9954	2020-12-31	Sulawesi Tenggara	Sulawesi	7907	147	6696	1064	2635461	38068	69.23	0.018591
9955	2020-12-31	Sulawesi Utara	Sulawesi	9671	310	7003	2358	2641884	13892	190.17	0.032055
9956	2020-12-31	Sumatera Barat	Sumatera	23464	522	18030	4912	5519245	42013	131.37	0.022247
9957	2020-12-31	Sumatera Selatan	Sumatera	11826	599	9364	1863	8217551	91592	89.72	0.050651
9958	2020-12-31	Sumatera Utara	Sumatera	18149	679	15403	2067	14874889	72981	203.82	0.037413

9959 rows × 11 columns

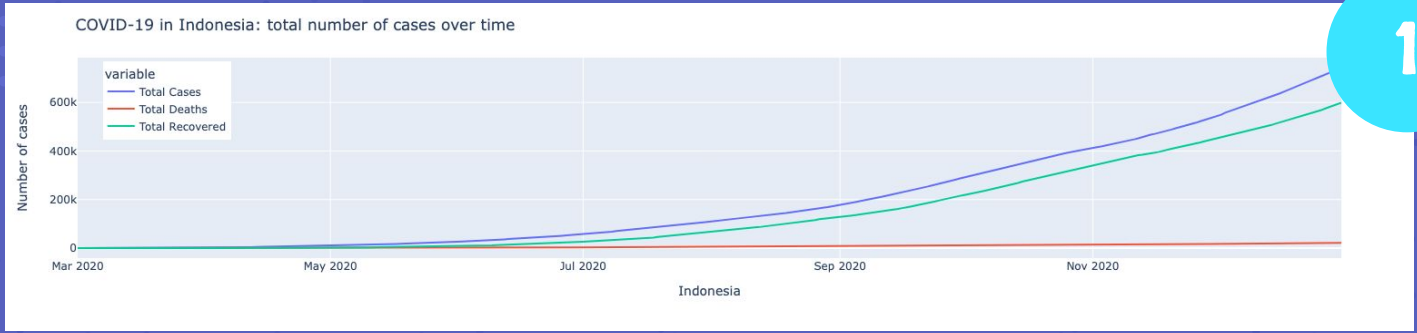
Data Final yang Digunakan

[99]:

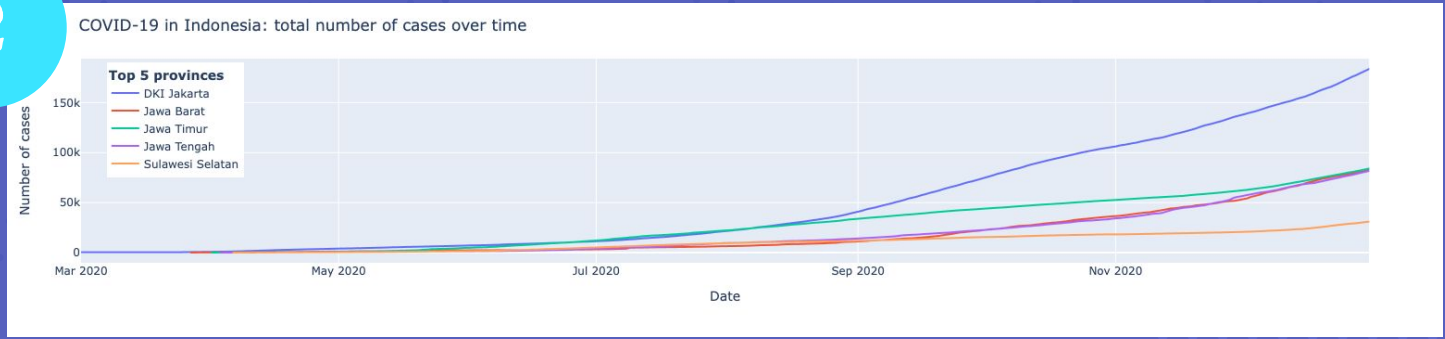
Province	Total Cases	Total Recovered	Population Density	Total Deaths	Total Active Cases	Population	Mortality
DKI Jakarta	183735	164776	16334.31	3270	15689	10846145	0.017797
Jawa Timur	84152	72135	846.78	5827	6190	40479023	0.069244
Jawa Barat	83579	70896	1276.55	1172	11511	45161325	0.014023
Jawa Tengah	81716	54212	1108.64	3361	24143	36364072	0.041130
Sulawesi Selatan	31047	25861	201.78	580	4606	9426885	0.018681

Visualisasi Data

1

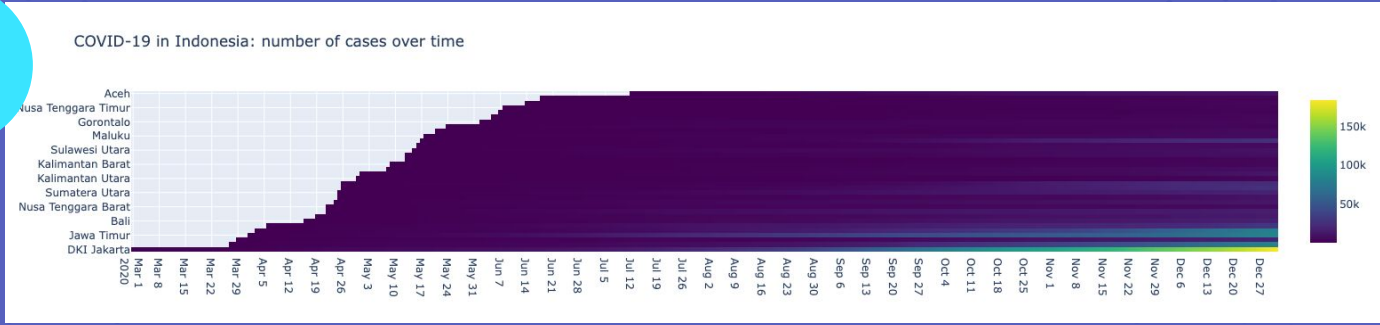


2

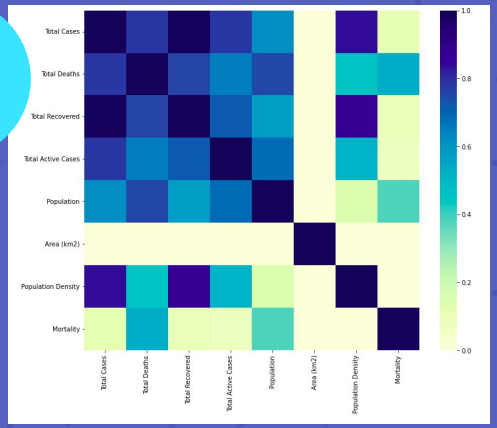


Visualisasi Data

3



4



1. Total Cases
2. Total Recovered
3. Total Active Cases
4. Population Density
5. Total Deaths
6. Population
7. Mortality

Akan digunakan untuk pengolahan data Selanjutnya

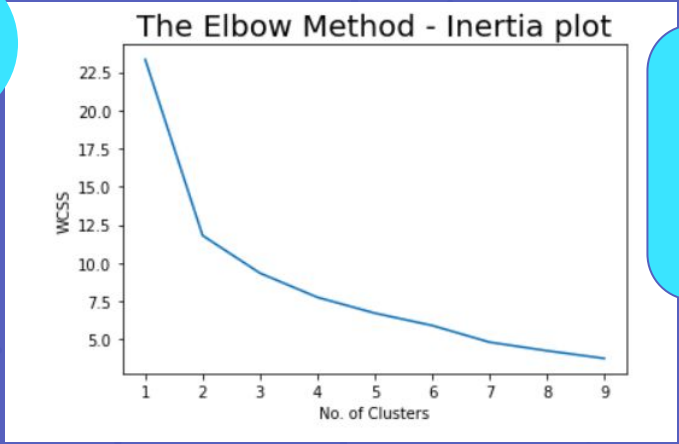
Modelling Data

1

```
[18]: #selecting features
X = d[['Mortality', 'Total Cases','Total Active Cases', 'Population Density', 'Population', 'Total Deaths']]
```

Load Dataset

2



Menggunakan WCSS dengan Metode Elbow Graph untuk menentukan nilai cluster(k)

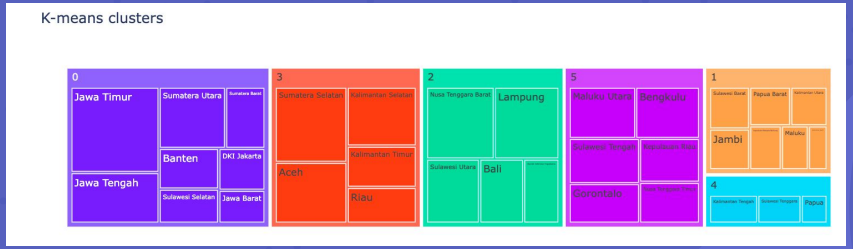
Digunakan k = 6.

Modelling Data

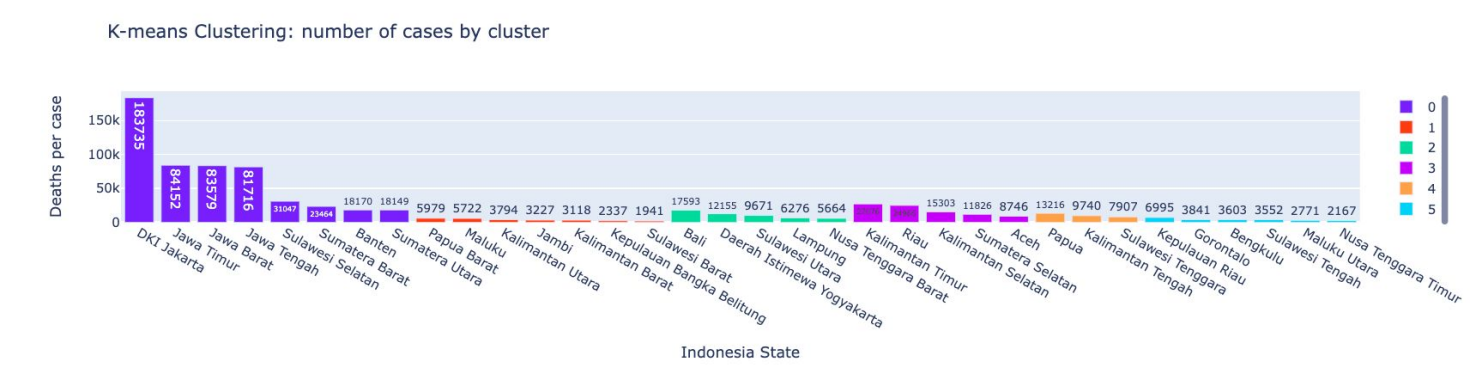
3

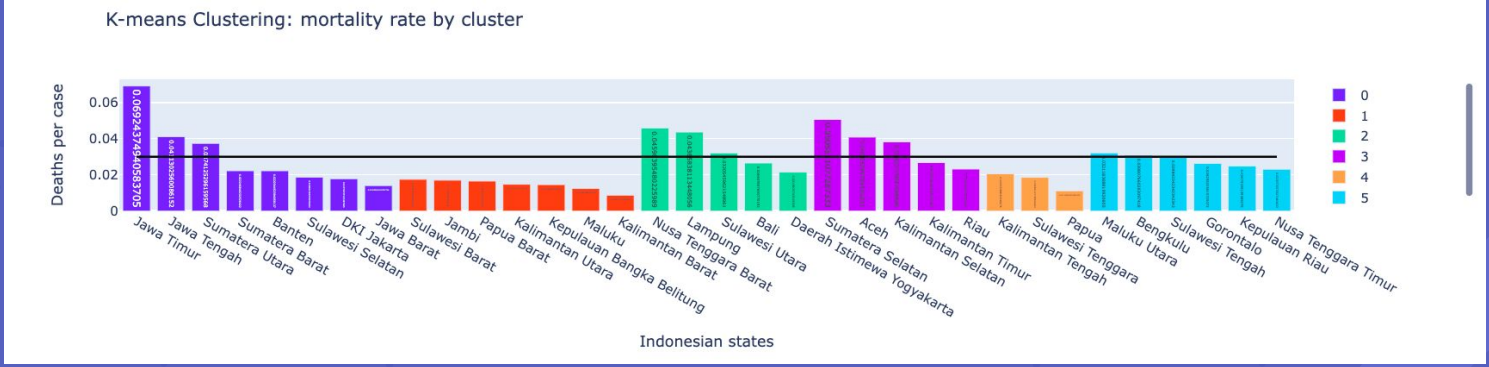
[20]:

Province	Total Cases	Total Recovered	Population Density	Total Deaths	Total Active Cases	Population	Mortality	K-means
Maluku Utara	0.000000	0.166667	0.166667	0.166667	0.000000	0.000000	0.833333	5
Sulawesi Tengah	0.166667	0.000000	0.166667	0.333333	0.500000	0.333333	0.666667	5
Bengkulu	0.166667	0.166667	0.500000	0.333333	0.333333	0.166667	0.666667	5
Gorontalo	0.166667	0.166667	0.500000	0.166667	0.000000	0.000000	0.666667	5
Kepulauan Riau	0.333333	0.500000	0.833333	0.333333	0.166667	0.166667	0.500000	5
Nusa Tenggara Timur	0.000000	0.000000	0.666667	0.000000	0.166667	0.666667	0.500000	5
Kalimantan Tengah	0.500000	0.333333	0.000000	0.500000	0.833333	0.166667	0.333333	4
Sulawesi Tenggara	0.500000	0.500000	0.166667	0.333333	0.333333	0.333333	0.333333	4
Papua	0.666667	0.500000	0.000000	0.333333	0.833333	0.500000	0.000000	4
Sumatera Selatan	0.666667	0.666667	0.333333	0.833333	0.666667	0.833333	1.000000	3
Kalimantan Selatan	0.666667	0.833333	0.500000	0.833333	0.166667	0.500000	0.833333	3
Aceh	0.500000	0.666667	0.333333	0.666667	0.333333	0.666667	0.833333	3
Kalimantan Timur	0.833333	0.833333	0.000000	1.000000	0.666667	0.333333	0.666667	3
Riau	0.833333	0.833333	0.333333	0.833333	0.333333	0.833333	0.500000	3
Lampung	0.333333	0.333333	0.833333	0.500000	0.500000	0.833333	1.000000	2
Nusa Tenggara Barat	0.333333	0.333333	0.833333	0.500000	0.500000	0.666667	1.000000	2
Sulawesi Utara	0.500000	0.500000	0.666667	0.666667	0.666667	0.333333	0.833333	2
Bali	0.666667	0.666667	0.833333	0.666667	0.833333	0.500000	0.666667	2
Daerah Istimewa Yogyakarta	0.666667	0.666667	1.000000	0.500000	0.666667	0.500000	0.333333	2
Papua Barat	0.333333	0.333333	0.000000	0.166667	0.000000	0.000000	0.166667	1
Kalimantan Utara	0.166667	0.000000	0.000000	0.166667	0.500000	0.000000	0.166667	1
Jambi	0.166667	0.166667	0.333333	0.000000	0.166667	0.333333	0.166667	1
Sulawesi Barat	0.000000	0.000000	0.500000	0.000000	0.000000	0.166667	0.166667	1
Maluku	0.333333	0.333333	0.166667	0.166667	0.333333	0.166667	0.000000	1
Kepulauan Bangka Belitung	0.000000	0.000000	0.333333	0.000000	0.166667	0.000000	0.000000	1
Kalimantan Barat	0.000000	0.166667	0.166667	0.000000	0.000000	0.666667	0.000000	1
Jawa Tengah	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0
Jawa Timur	1.000000	1.000000	0.833333	1.000000	1.000000	1.000000	1.000000	0
Sumatera Utara	0.833333	0.833333	0.666667	0.833333	0.666667	1.000000	0.833333	0
Sumatera Barat	0.833333	0.833333	0.666667	0.666667	0.833333	0.666667	0.500000	0
Sulawesi Selatan	1.000000	1.000000	0.666667	0.833333	0.833333	0.833333	0.333333	0
Banten	0.833333	0.666667	1.000000	0.666667	1.000000	0.833333	0.333333	0
DKI Jakarta	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.166667	0
Jawa Barat	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0



Modelling Data



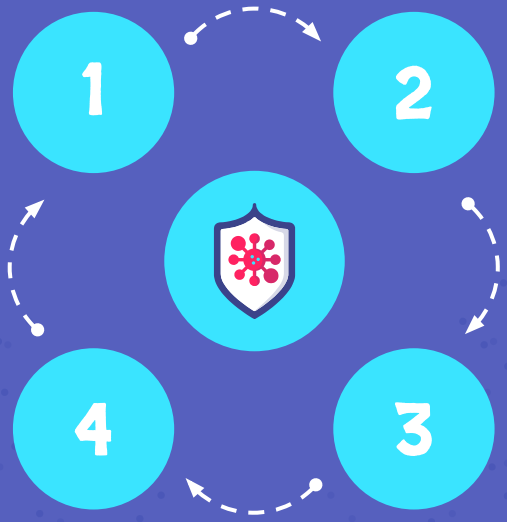


04

Kesimpulan

Jambi menempati posisi pertama dengan jumlah kasus terendah yakni sebanyak 4.687 **kasus**

Pembagian Klasterisasi penanganan COVID-19 dipengaruhi tingkat **kemungkinan kematian, jumlah kasus yang masih aktif, kepadatan penduduk dan luas area provinsi tersebut.**



DKI Jakarta menempati posisi pertama dengan jumlah kasus tertinggi yakni sebanyak **86.963 kasus**

Tingkat **kematian** paling tinggi akibat COVID-19 yaitu provinsi **Jawa Timur**

Thank u!