LAPORAN TUGAS MATA KULIAH
PENGANTAR SAINS DATA
"EXPLORATORY DATA ANALYSIS (EDA) PADA DATA COVID-19 INDONESIA"

Disusun Oleh:

Benony Gabriel	**105222002**
Muhammad S. Wiwi	**105222012**
Amanda Weza Chania	**105222018**

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS SAINS DAN ILMU KOMPUTER
UNIVERSITAS PERTAMINA
2024

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pandemi COVID-19 telah membawa dampak yang sangat besar di seluruh dunia, termasuk Indonesia. Menurut data dari World Health Organization (WHO), sejak virus ini pertama kali terdeteksi pada akhir 2019, lebih dari 200 juta orang di seluruh dunia telah terinfeksi, dan jutaan lainnya meninggal dunia[1]. Di Indonesia sendiri, COVID-19 menyebabkan perubahan besar dalam kehidupan masyarakat, ekonomi, dan sistem kesehatan. Berdasarkan laporan dari Kementerian Kesehatan Republik Indonesia , sejak kasus pertama ditemukan pada Maret 2020, jumlah kasus terus meningkat hingga mencapai puncaknya pada beberapa gelombang, salah satunya pada pertengahan tahun 2021[2].

Data yang dikumpulkan terkait penyebaran COVID-19 di Indonesia sangat penting untuk dianalisis guna memahami tren penyebaran, dampak, serta efektivitas intervensi yang dilakukan pemerintah. Dengan memanfaatkan pendekatan Exploratory Data Analysis (EDA), kita dapat menggali insight yang bermanfaat dari data, seperti tren peningkatan kasus, faktor yang memengaruhi penyebaran, dan daerah dengan tingkat risiko tinggi. Analisis ini tidak hanya penting dalam mempelajari pola pandemi di masa lalu, tetapi juga membantu dalam pengambilan keputusan terkait kebijakan kesehatan publik yang lebih baik di masa depan.

1.2 Tujuan

Tujuan dari EDA ini adalah untuk:

1. Menganalisis tren kasus COVID-19 di Indonesia dari awal pandemi hingga titik waktu tertentu, sehingga dapat dilihat pola penyebaran dan lonjakan kasus yang terjadi.
2. Mengidentifikasi daerah-daerah dengan jumlah kasus tertinggi serta memahami distribusi kasus COVID-19 berdasarkan provinsi atau kabupaten.
3. Memvisualisasikan hubungan antara jumlah kasus aktif, sembuh, dan meninggal, serta menganalisis faktor-faktor yang mungkin berkontribusi terhadap perbedaan antar daerah.

4. Memberikan insight yang relevan bagi pengambil kebijakan, khususnya dalam perencanaan strategi mitigasi dan pengendalian penyebaran virus di masa depan.

1.3 Deskripsi Dataset

Dataset COVID-19 Indonesia ini berisi data harian tentang perkembangan pandemi di berbagai wilayah Indonesia, termasuk provinsi dan kota besar. Dataset mencakup tanggal pencatatan, jumlah kasus baru, kematian baru, dan pasien yang sembuh setiap hari, serta total akumulasi kasus, kematian, dan pemulihan sejak awal pandemi. Selain itu, variabel lain seperti populasi, kepadatan penduduk, dan luas wilayah juga disertakan, yang memungkinkan analisis lebih mendalam tentang dampak COVID-19 di setiap wilayah.

1.4 Tahapan Melakukan EDA

Berikut ini adalah tahapan-tahapan yang kami lakukan dalam melakukan **Exploratory Data Analysis (EDA)** dari dataset Covid-19 Indonesia:

1. Mengimport Library yang dibutuhkan
2. Memuat dan memeriksa dataset
3. Gambaran umum dataset
4. Pre-processing data
5. Melakukan analisis deskriptif
6. Visualisasi data
7. Mengambil insights

BAB II
TINJAUAN PUSTAKA

2.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan langkah penting dalam setiap analisis data. Tujuan utama dari EDA adalah untuk memahami karakteristik data, seperti distribusi, keberadaan outlier, dan anomali, yang dapat mengarahkan pengujian hipotesis lebih lanjut[3]. EDA juga berperan dalam proses pembangkitan hipotesis dengan memvisualisasikan dan memahami data, biasanya menggunakan representasi grafis[3]. Dengan EDA, analis dapat mengenali pola-pola alami dalam data, dan seringkali metode feature selections juga termasuk dalam proses ini[3].

EDA dilakukan setelah data dikumpulkan dan melalui tahap pre-processing, dengan tujuan memvisualisasikan dan memanipulasi data tanpa asumsi tertentu. Hal ini memungkinkan analisis terhadap kualitas data serta membantu dalam pembuatan model. Sebagian besar teknik EDA berfokus pada aspek grafis, meskipun ada beberapa teknik kuantitatif[3]. Penggunaan visualisasi menjadi sangat penting dalam EDA karena visualisasi memudahkan analisis data dan memberikan wawasan yang lebih dalam secara intuitif[3].

2.2 Pandemi COVID-19 dan Analisis Data

COVID-19 adalah penyakit menular yang disebabkan oleh virus SARS-CoV-2 [4]. Sebagian besar orang yang terinfeksi akan mengalami gejala ringan hingga sedang pada saluran pernapasan dan sembuh tanpa perawatan khusus. Namun, sebagian lainnya bisa mengalami kondisi yang lebih serius dan memerlukan penanganan medis. Lansia dan mereka yang memiliki kondisi kesehatan seperti penyakit jantung, diabetes, gangguan pernapasan kronis, atau kanker lebih berisiko mengalami sakit parah[4].

Di Indonesia, data COVID-19 yang dikumpulkan oleh pemerintah, seperti yang disajikan melalui platform https://dashboardcovid19.kemkes.go.id/, digunakan untuk menganalisis tren peningkatan kasus, lonjakan di berbagai provinsi, dan distribusi berdasarkan usia atau jenis

kelamin. EDA pada data COVID-19 di Indonesia sangat penting karena dapat memberikan gambaran terkait periode kritis lonjakan kasus, wilayah-wilayah yang membutuhkan perhatian khusus, serta efektivitas kebijakan pembatasan sosial yang diterapkan.

BAB III
HASIL DAN PEMBAHASAN

3.1 Hasil Analisis Statistik Deskriptif

Melakukan analisis statistik untuk menghitung ukuran statistik utama seperti mean, median, mode, standar deviasi, dan persentil untuk kolom-kolom numerik yang penting. Tujuannya adalah untuk mendapatkan gambaran umum tentang data:

- **Mean (Rata-rata)**: Nilai rata-rata dari data.
- **Median**: Nilai tengah dari data.
- **Mode (Modus)**: Nilai yang paling sering muncul.
- **Standar Deviasi**: Untuk melihat variasi atau penyebaran data.
- **Min dan Max**: Nilai minimum dan maksimum dalam data.
- **Persentil dan Kuartil**: Pembagian data ke dalam bagian yang lebih kecil.

Gambaran Hasil Analisis Statistik:

	count	mean	std	min	25%	50%	75%	max	mode
New Cases	21759.0	3.912936e+02	2.074551e+03	0.00	7.000	41.00	151.00	5.675700e+04	0.00
New Deaths	21759.0	1.322041e+01	7.648262e+01	0.00	0.000	1.00	5.00	2.069000e+03	0.00
New Recovered	21759.0	3.773110e+02	1.999063e+03	0.00	4.000	31.00	143.00	4.883200e+04	0.00
New Active Cases	21759.0	7.621674e-01	9.372135e+02	-25725.00	-16.000	0.00	27.00	3.672600e+04	0.00
Total Cases	21759.0	8.525997e+04	3.685133e+05	1.00	1822.500	10780.00	36464.50	4.257243e+06	2.00
Total Deaths	21759.0	2.648289e+03	1.177601e+04	0.00	50.000	283.00	1050.00	1.438580e+05	1.00
Total Recovered	21759.0	7.671260e+04	3.403957e+05	0.00	1038.500	8745.00	32932.50	4.105680e+06	0.00
Total Active Cases	21759.0	5.899079e+03	2.751810e+04	-2306.00	182.000	919.00	2607.50	5.741350e+05	1.00
Population	21759.0	1.547817e+07	4.483574e+07	648407.00	1999539.000	4216171.00	9095591.00	2.651855e+08	10846145.00
Population Density	21759.0	7.449898e+02	2.743210e+03	8.59	47.790	103.84	262.70	1.633431e+04	138.34
Area (km2)	21759.0	1.112418e+05	3.203746e+05	664.00	16787.000	42013.00	75468.00	1.916907e+06	664.00
New Cases per Million	21759.0	2.939303e+01	6.994109e+01	0.00	1.750	8.16	26.26	1.348130e+03	0.00
Total Cases per Million	21759.0	6.183651e+03	1.028812e+04	0.01	415.915	2727.46	7374.12	7.966379e+04	0.33
New Deaths per Million	21759.0	8.527267e-01	2.277310e+00	0.00	0.000	0.18	0.75	6.380000e+01	0.00
Total Deaths per Million	21759.0	1.635422e+02	2.541958e+02	0.00	10.880	73.55	195.85	1.533980e+03	0.00

- **Kasus Baru**: Rata-rata jumlah kasus baru per hari adalah sekitar 391, tetapi terdapat variabilitas yang tinggi, sebagaimana ditunjukkan oleh standar deviasi sebesar 2074. Modusnya adalah 0, yang berarti pada banyak hari tidak ada kasus baru yang dilaporkan.
- **Kematian Baru**: Rata-rata dilaporkan 13 kematian baru per hari, dengan standar deviasi sebesar 76. Nilai median adalah 1.
- **Sembuh Baru**: Rata-rata jumlah sembuh per hari adalah sekitar 377, dengan standar deviasi sebesar 1999.
- **Total Kasus**: Rata-rata total kasus per lokasi adalah 85.259, dengan penyebaran yang besar, sebagaimana ditunjukkan oleh standar deviasi sebesar 368.513.
- **Populasi**: Ukuran populasi rata-rata di lokasi-lokasi dalam dataset adalah sekitar 15,5 juta, namun terdapat variasi yang signifikan dengan standar deviasi sebesar 44,8 juta.
- **Kepadatan Populasi**: Rata-rata kepadatan populasi adalah 744,99 orang per km², dengan beberapa wilayah sangat padat (hingga 16.334 orang per km²).

Analisis statistik ini menyoroti variabilitas yang tinggi dalam dataset, dengan sejumlah besar hari dan lokasi melaporkan nol kasus baru atau kematian. Standar deviasi yang besar mencerminkan penyebaran COVID-19 yang bervariasi di berbagai wilayah dan waktu.

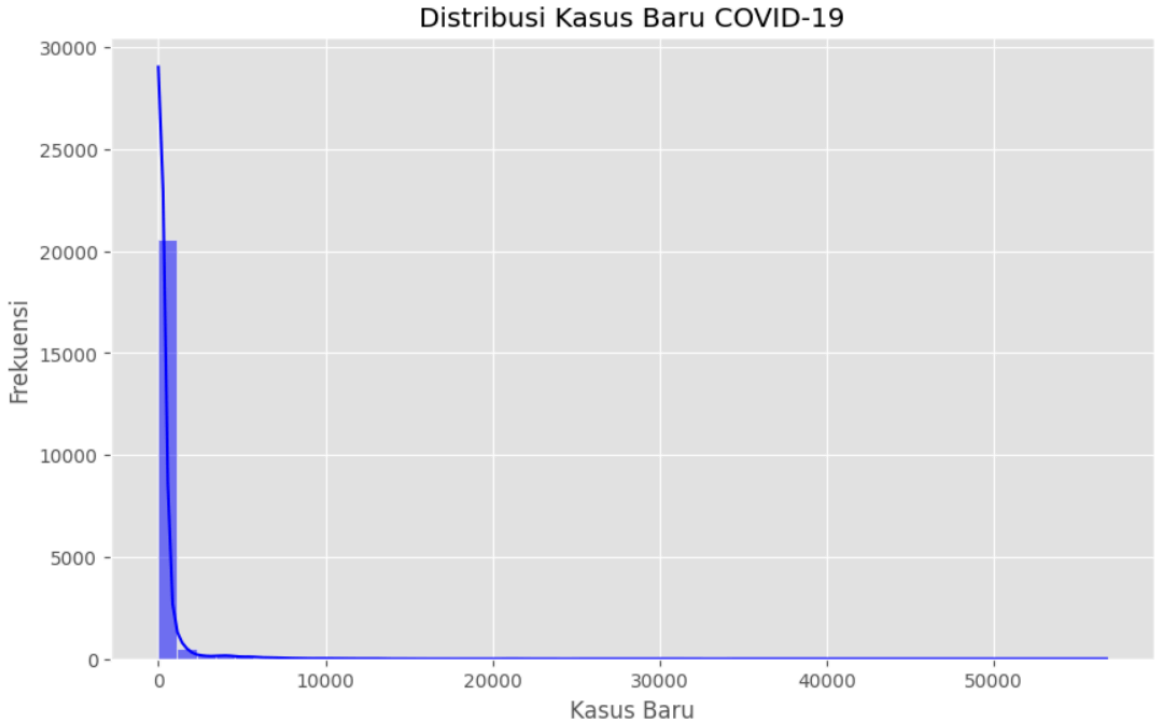
3.2 Visualisasi Data

Visualisasi sangat penting untuk memahami pola dalam data. Beberapa visualisasi dasar yang sering digunakan dalam EDA adalah:

- **Histogram**: Menampilkan distribusi frekuensi data.
- **Boxplot**: Menampilkan penyebaran data dan mengidentifikasi outliers (nilai pencilan).
- **Scatter Plot**: Memvisualisasikan hubungan antara dua variabel.
- **Heatmap**: Untuk melihat korelasi antar variabel.
- **Line Plot**: Untuk memvisualisasikan tren data dari waktu ke waktu.

Visualisasi 1: Histogram Kasus Baru COVID-19

Untuk memahami distribusi kasus baru.

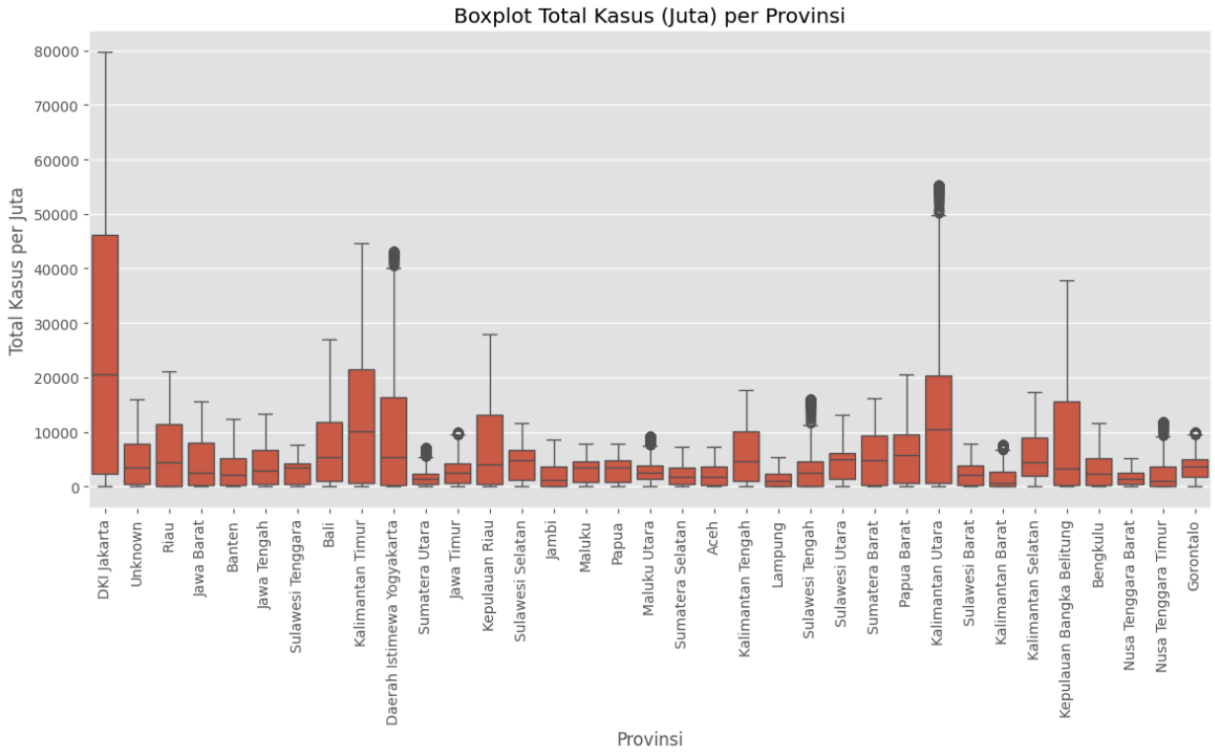


Insights yang diperoleh:

- Distribusi frekuensi kasus baru menunjukkan bahwa sebagian besar hari memiliki jumlah kasus baru yang rendah.
- Ada beberapa hari dengan jumlah kasus yang sangat tinggi, menghasilkan distribusi yang mencakup ekor panjang ke kanan (right-skewed). Ini menunjukkan bahwa meskipun sebagian besar hari relatif lebih aman, ada beberapa hari lonjakan besar dalam kasus baru.

Visualisasi 2: Boxplot Total Kasus per Juta per Provinsi

Untuk memvisualisasikan variasi jumlah kasus di berbagai provinsi.

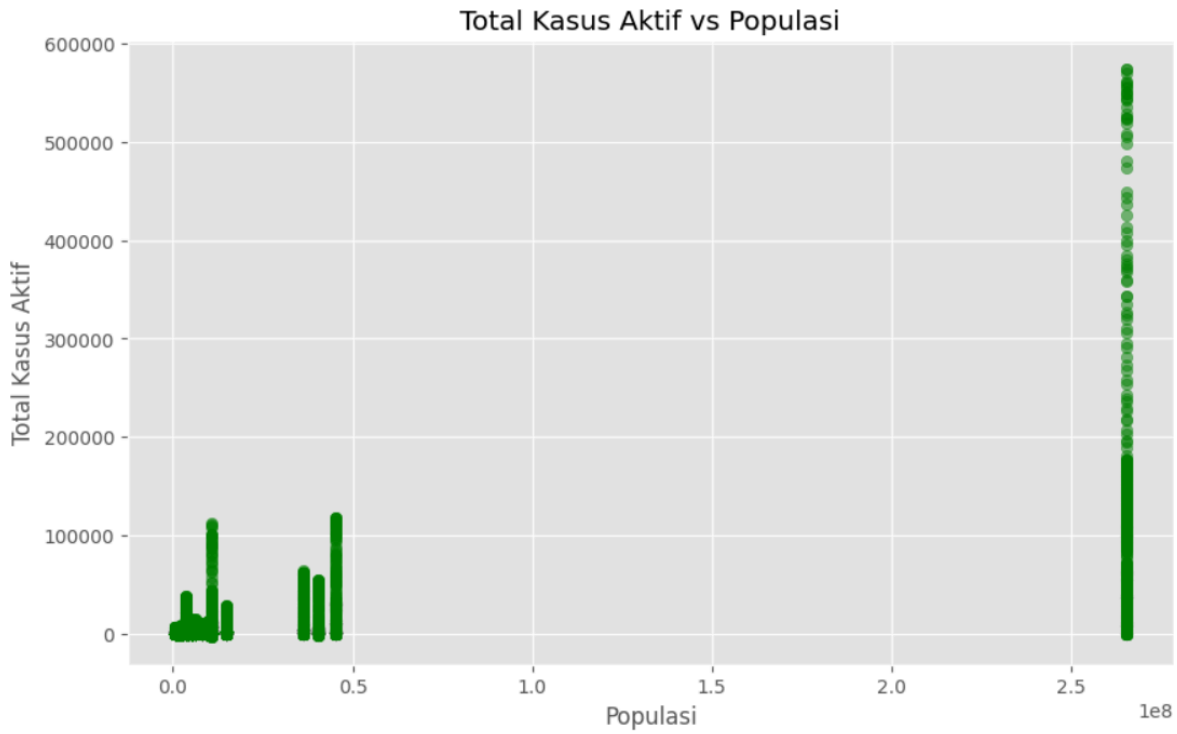


Insights yang diperoleh:

- Boxplot memperlihatkan variasi yang signifikan dalam jumlah kasus per juta penduduk di berbagai provinsi.
- Beberapa provinsi memiliki penyebaran data yang lebih lebar dan terdapat outliers (nilai pencilan) yang menunjukkan bahwa di provinsi-provinsi tertentu ada hari-hari dengan lonjakan yang sangat tinggi dalam jumlah kasus.
- Provinsi dengan median yang lebih tinggi menunjukkan dampak COVID-19 yang lebih besar pada populasi mereka.

Visualisasi 3: Scatter Plot Total Kasus Aktif vs Populasi

Untuk melihat apakah ada hubungan antara ukuran populasi dan jumlah kasus aktif.

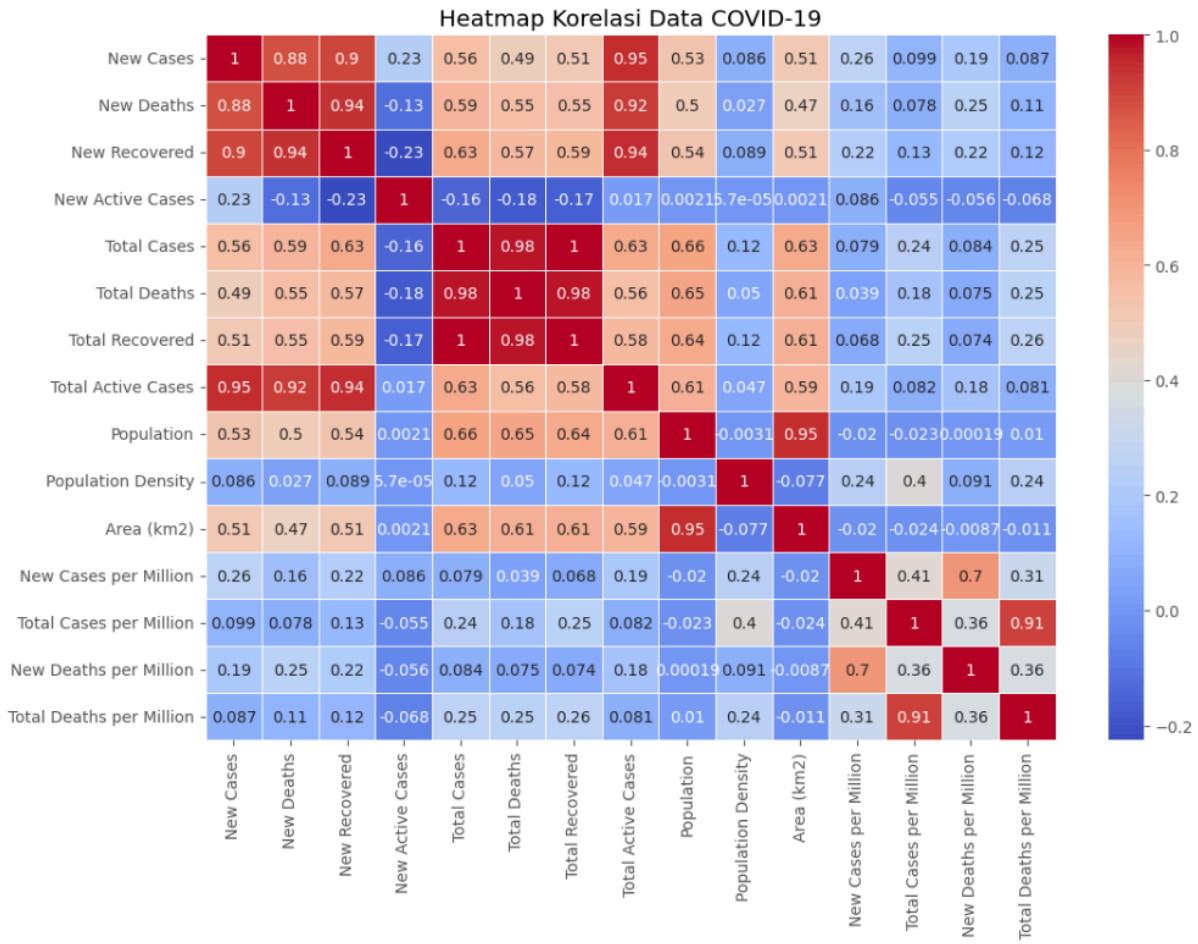


Insights yang diperoleh:

- Tidak ada korelasi yang sangat jelas antara ukuran populasi dan jumlah kasus aktif. Ini menunjukkan bahwa jumlah kasus aktif di sebuah wilayah tidak selalu berbanding lurus dengan jumlah penduduk.
- Beberapa provinsi dengan populasi lebih kecil memiliki jumlah kasus aktif yang tinggi, yang mungkin mengindikasikan penanganan atau penyebaran virus yang lebih cepat di wilayah tersebut.
- Hal ini juga bisa disebabkan oleh kepadatan penduduk atau faktor lain seperti akses ke layanan kesehatan dan tingkat kesadaran masyarakat.

Visualisasi 4: Heatmap Korelasi Antara Variabel

Untuk mengidentifikasi korelasi antar variabel.

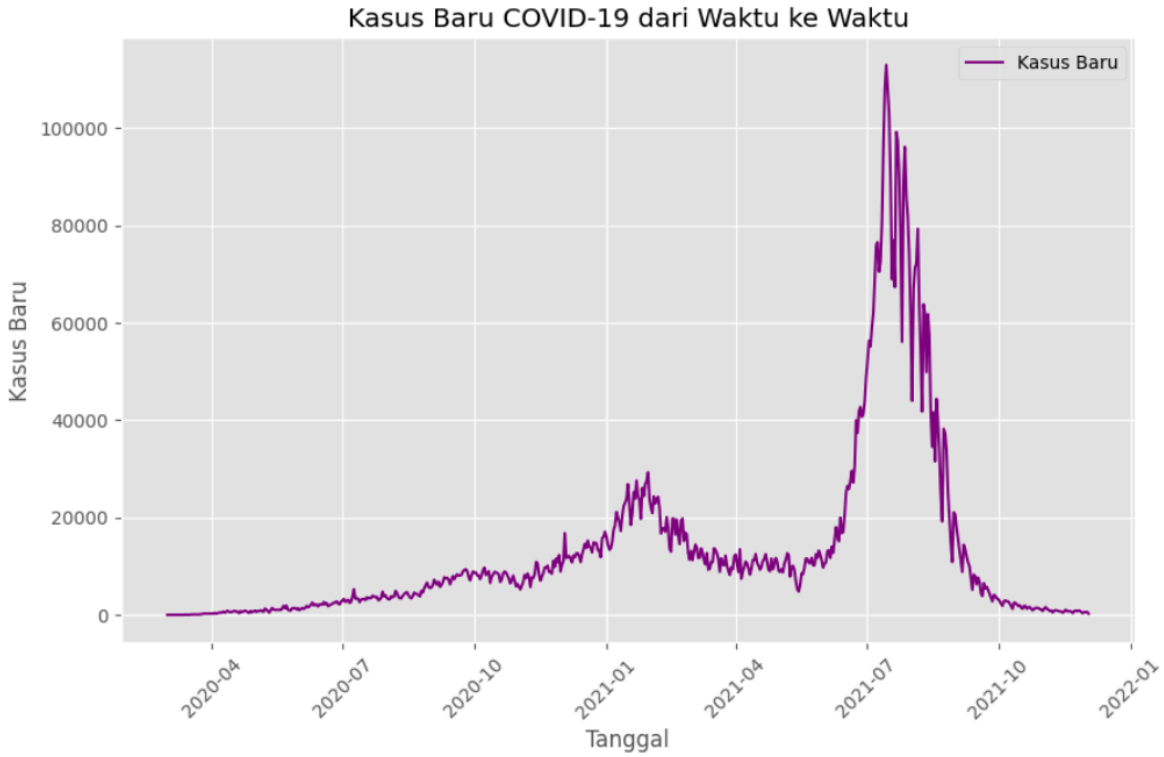


Insights yang diperoleh:

- Korelasi positif yang kuat terlihat antara variabel-variabel seperti **Total Kasus** dan **Total Kematian** atau **Total Kasus** dan **Total Sembuh**, yang diharapkan karena semakin banyak kasus, semakin banyak juga kematian dan pasien sembuh.
- Korelasi lemah atau hampir tidak ada antara **Total Kasus** dan beberapa faktor seperti **Populasi**, yang menunjukkan bahwa jumlah kasus di suatu wilayah tidak selalu terkait langsung dengan ukuran populasi.
- Korelasi ini membantu dalam memahami hubungan antar variabel dan juga bisa menjadi dasar untuk analisis yang lebih lanjut.

Visualisasi 5: Line Plot Kasus Baru dari Waktu ke Waktu

Untuk mengamati tren kasus baru COVID-19 dari waktu ke waktu.



Insights yang diperoleh:

- Plot ini menunjukkan tren fluktuasi jumlah kasus baru dari waktu ke waktu, di mana puncak dan lembah dari kasus baru dapat dengan jelas terlihat.
- Wawasan yang bisa diperoleh adalah kapan puncak tertinggi dalam jumlah kasus terjadi, yang mungkin berhubungan dengan kebijakan kesehatan masyarakat, pembatasan sosial, atau pengabaian terhadap protokol kesehatan.
- Tren ini juga bisa digunakan untuk memprediksi atau mengantisipasi pola lonjakan kasus di masa mendatang.

BAB IV
PENUTUP

4.1 Kesimpulan

Dari seluruh visualisasi, kita dapat memahami bahwa:

- Kasus COVID-19 di Indonesia tidak terdistribusi secara merata di seluruh provinsi. Beberapa provinsi lebih terdampak daripada yang lain.
- Lonjakan kasus terjadi pada waktu tertentu, dan hal ini mungkin terkait dengan faktor-faktor sosial, geografis, atau kebijakan pemerintah.
- Variabel-variabel seperti jumlah kasus dan kematian memiliki korelasi yang jelas, tetapi faktor-faktor lain seperti populasi dan kepadatan penduduk memerlukan analisis lebih lanjut untuk memahami pengaruhnya.
- Data ini sangat bermanfaat untuk memahami pola penyebaran virus dan untuk merancang strategi yang lebih baik dalam menangani pandemi di masa depan.

4.2 Saran

- Disarankan untuk melakukan analisis lebih lanjut dengan membandingkan data kasus COVID-19 antar provinsi dan kota. Ini dapat membantu dalam memahami perbedaan respons dan efektivitas kebijakan di berbagai daerah.
- Disarankan untuk melakukan pemantauan berkala terhadap tren kasus, kematian, dan vaksinasi agar dapat mengambil keputusan yang cepat dan tepat dalam penanganan pandemi.
- Mempertimbangkan penelitian lebih lanjut untuk mengeksplorasi dampak psikologis dan sosial dari pandemi terhadap masyarakat, serta efektivitas kebijakan yang diterapkan.

DAFTAR PUSTAKA

[1] World Health Organization. (2021). Coronavirus Disease (COVID-19) Dashboard. Retrieved October 9, 2024, from https://covid19.who.int

[2] Kementerian Kesehatan Republik Indonesia. (2021). Data COVID-19 Indonesia. Retrieved October 9, 2024, from https://dashboardcovid19.kemkes.go.id/

[3] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis. 10.1007/978-3-319-43742-2_15.

[4] World Health Organization. (n.d.). Coronavirus disease (COVID-19). Retrieved October 9, 2024, from https://www.who.int/health-topics/coronavirus#tab=tab_1