DAFTAR ISI

BAGIAN 1

NON-DATASET: DATA SCIENCE FUNDAMENTALS

THINGAMAJIG

2021

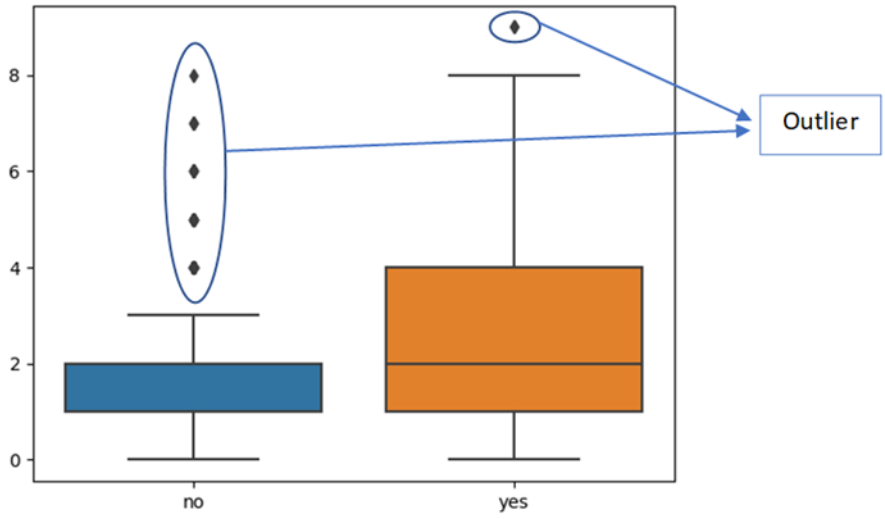
1.1 Outlier dalam Statistika

1.1.1 Problematika

"Jelaskan secara teori statistik mengenai outlier (pencilan), implikasinya dalam analisis data, serta bagaimana melakukan manajemen data terhadap kasus outlier".

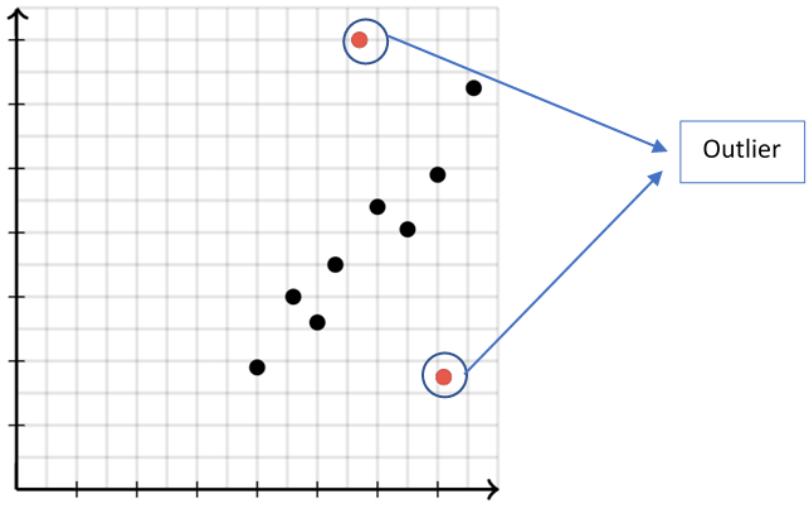
1.1.2 Solusi

Outlier merupakan observasi atau data poin yang nilainya berbeda atau jauh daripada observasi pada umumnya. Pendeteksian outlier bertujuan untuk menemukan pola tertentu dalam data yang sifatnya berupa anomali, pendeteksian bisa menggunakan visualisasi menggunakan *boxplot* sebagai berikut.



Gambar 1. Boxplot

Alternatif lain dalam melihat outlier adalah dengan menggunakan *scatter plot* sebagai berikut.



Gambar 2. Scatter Plot

Selain menggunakan visualisasi, outlier dapat dideteksi menggunakan IQR (*interquartile range*) atau rentang kuartil, yaitu menghitung jarak antara Q_3- Q_1 dimana

Q_3 : persentil yang ke-75 atau 75%

Q_1 : persentil yang ke-25 atau 25%

Sedemikian sehingga diperoleh persamaan

$$IQR = Q_3 - Q_1 \ (\mathbf{1})$$

Suatu data merupakan outlier jika nilainya dari data tersebut berada di bawah $Q_1 - 1{,}5\,IQR$ atau berada di atas $Q_3 + 1{,}5\,IQR$. Munculnya outlier pada kumpulan data disebabkan oleh beberapa kemungkinan yaitu sebagai berikut.

1. Adanya kesalahan prosedur dalam memasukkan data.
2. Kesalahan dalam pengukuran atau analisis.
3. Adanya keadaan yang benar-benar khusus, seperti pandangan responden terhadap sesuatu yang menyimpang dikarenakan adanya suatu alasan yang tidak diketahui oleh peneliti sendiri.

Selain itu, outlier juga dapat menyebabkan beberapa hal seperti berikut.

1. Memperburuk model yg diperoleh.
2. Varian pada data tersebut menjadi besar.
3. Taksiran interval memiliki interval yang lebar.

Sebagai contoh, misal diberikan dua buah kumpulan data dalam perhitungan *mean* yaitu data 1 = [2,3,5,6,100] dan memiliki *mean* sebesar 23,2 dan data 2 = [2,3,5,6,7] dan memiliki *mean* sebesar 4,6. Pada data 1, nilai 100 dianggap sebagai outlier karena nilai tersebut sangat berbeda dari nilai-nilai sebelumnya yang jenisnya satuan dan hal tersebut akan mempengaruhi nilai *mean* yang dibentuk. Terdapat beberapa solusi dalam mengatasi outlier yaitu sebagai berikut.

1. Mengganti metode yang kita gunakan dengan metode-metode yang lebih peka terhadap adanya outlier atau yang dikenal dengan non parametrik. Statistik parametrik merupakan ilmu statistik yang mempertimbangkan jenis sebaran atau distribusi data, yaitu apakah data menyebar secara normal atau tidak. Statistik parametrik digunakan untuk menguji hipotesis dan variabel yang terukur. Dengan kata lain, data yang akan dianalisis menggunakan statistik parametrik harus memenuhi asumsi normalitas. Statistik non parametrik merupakan tes yang modelnya tidak menetapkan syarat-syarat mengenai parameter-parameter populasi

atau dengan kata lain statistik non parametrik tidak menuntut pengukuran sekuat yang dituntut tes statistik parametrik.

2. Mengurangi jumlah data khususnya data-data yang memiliki nilai outlier.
3. Melakukan transformasi data dalam bentuk logaritma, kuadrat dan lain sebagainya.

1.2 Prinsip Korelasi

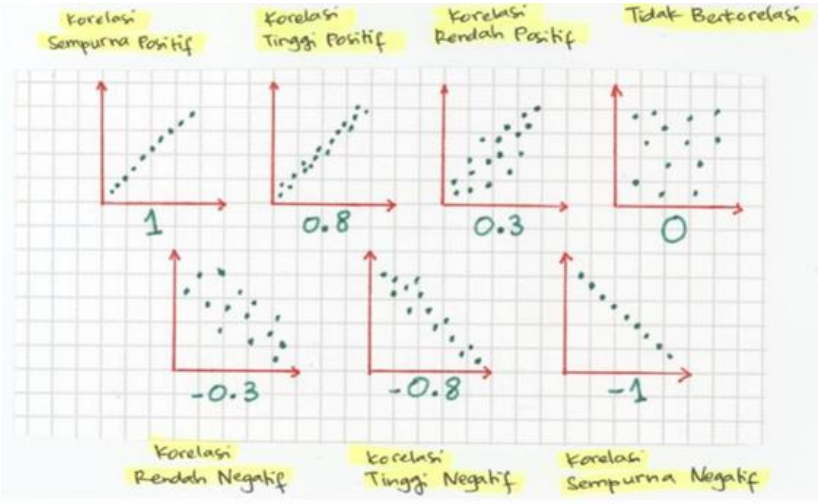
1.2.1 Problematika

"Jelaskan konsep dan prinsip korelasi, lalu sebisa mungkin kaitkan dengan dasar-dasar statistik serta implikasinya terhadap konsep/teori statistik lain".

1.2.2 Solusi

Korelasi adalah salah satu jenis hubungan yang tidak berimplikasi pada sebab akibat dengan prinsip mengukur keeratan hubungan linear dari 2 variabel. Sehingga korelasi hanya mengukur seberapa kuat dan arah hubungan. Berikut merupakan interpretasi dari nilai korelasi.

- Jika nilai korelasi bernilai 0 (nol) artinya kedua variabel kurang memiliki keterhubungan.
- Jika nilai korelasi bernilai 1, artinya jika satu nilai meningkat maka nilai yang lain secara linear akan meningkat.
- Jika nilai korelasi bernilai -1, artinya jika satu nilai meningkat maka nilai yang lain secara linear akan menurun.



Gambar 3. Jenis Korelasi

Korelasi memiliki keterkaitan dengan statistik inferensial yang fungsinya untuk menguji hipotesis dan melakukan generalisasi hasil analisis sampel ke populasi yang terbagi menjadi dua bagian, yaitu statistik parametrik dan non-parametrik. Statistik

parametrik merupakan statistika yang mengharuskan sebaran data normal sedangkan statistika non-parametrik merupakan statistika yang mengabaikan segala asumsi dari statistika parametrik terutama yang berkaitan dengan distribusi normal.

Pada statistika parametrik cocok jika menggunakan uji korelasi *pearson*. Korelasi *pearson* adalah korelasi yang digunakan untuk data diskrit dan kontinu yang memiliki jumlah data yang besar serta dengan skala pengukuran interval atau rasio. Korelasi *pearson* menghitung korelasi dengan menggunakan variansi data dengan rumus sebagai berikut.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)(n\sum y_i^2 - (\sum y_i)^2)}}$$

Keterangan :

$r_{xy} = korelasi\ antara\ x\ dan\ y$

$x_i = nilai\ x\ ke - i$

$y_i = nilai\ y\ ke - i$

$n = banyaknya\ nilai$

Pada statistika non-parametrik cocok jika menggunakan uji korelasi *spearman*. Uji korelasi *spearman* merupakan uji korelasi variabel dengan skala ordinal dan secara visual terlihat hubungan kedua variabel tidak linear dengan rumus sebagai berikut.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Keterangan :

$r_s = korelasi\ spearman$

$d = selisih\ antara\ X\ dan\ Y$

$n = jumlah\ pasangan\ (data)$

Hasil dari uji korelasi dapat diinterpretasikan sebagai berikut.

Tabel 1. Tabel Tingkat Hubungan Korelasi

Interval Koefisien	**Tingkat Hubungan**
0,8 - 1,0	Sangat Tinggi
0,6 - 0,8	Kuat
0,4 - 0,6	Cukup

0,2 - 0,4	Rendah
0,0 - 0,2	Sangat Rendah

1.3 Teori Dasar *Machine Learning*

1.3.1 Problematika

"Sebutkan teori dasar machine learning yang kalian ketahui, lalu jelaskan dalam bahasa sederhana mengenai teori tersebut dan implikasinya".

1.3.2 Solusi

Machine learning adalah suatu kecerdasan buatan yang membuat sistem bisa mengadaptasi kemampuan manusia untuk belajar. Mengutip dari laman Towards AI, Sederhananya algoritma *machine learning* belajar berdasarkan pengalaman, mirip dengan yang dilakukan manusia. Misalnya, setelah melihat beberapa contoh objek, algoritma *machine learning* yang menggunakan komputasi dapat mengenali objek itu dalam skenario baru yang sebelumnya tidak terlihat. Proses belajar dilakukan untuk menemukan pola dan fitur tertentu dalam jumlah data yang besar. Hal ini bertujuan untuk membuat suatu keputusan maupun prediksi berdasarkan data-data tersebut. Semakin bagus algoritmanya, akurasi keputusan dan prediksi sistem akan semakin baik.

Menurut Glints dalam lamannya, mengutip dari Towards AI, *machine learning* adalah hal yang sangat penting sekarang ini. Dengan *machine learning*, kita bisa memproses dan menganalisis data yang lebih besar dan rumit dengan waktu yang lebih singkat. Pengaplikasian ilmu *machine learning* ini bisa diaplikasikan pada berbagai macam industri dan terus dikembangkan. Contoh penerapan *machine learning* ini yaitu pada Netflix yang bisa mengetahui preferensi film atau serial sesuai dengan apa yang selama ini telah kita tonton.

Dalam lamannya Glints juga menambahkan bahwa terdapat 4 jenis m*achine learning*, diantaranya sebagai berikut.

Tabel 2. Tabel Jenis *Machine Learning*

Supervised learning	Pada algoritma *machine learning* ini menggunakan data terlabel, contohnya input di mana output-nya diketahui. DQlab dalam lamannya menambahkan bahwa pada dasarnya algoritma dilatih agar dapat memilih fungsi-fungsi yang paling menggambarkan input dimana X tertentu membuat estimasi terbaik dari y. Algoritma ini membutuhkan data latih yang benar sehingga sistem dapat mempelajari polanya dan regresi, klasifikasi, *K-NN, Naive Bayes, Decision Trees,* regresi linier, *Support Vector Machine, dan neural network.*
Unsupervised learning	Pada metode *machine learning* ini, data yang diolah tidak memiliki label dan sistem tidak mengetahui jawaban atau output yang benar. Tujuan dari *machine learning* dengan metode ini adalah untuk mengeksplorasi data dan menemukan struktur di dalamnya. Algoritma ini digunakan untuk *clustering* dan *association rule.*
Semi-supervised learning	Algoritma ini cocok digunakan untuk sejumlah data berukuran besar yang dibagi menjadi dua bagian yang diberi label dan tidak diberi label. Contoh penggunaan *semi-supervised learning* adalah untuk proses identifikasi wajah seseorang pada webcam atau kamera smartphone.

Reinforcement Learning	Menurut Glints dalam lamannya, algoritma ini akan mampu menemukan aksi atau perlakuan yang menghasilkan output terbaik dari hasil uji coba berulang kali (*trial and error*). Ada tiga komponen utama untuk *reinforcement learning*, yaitu agen (pembuat keputusan), lingkungan (apa saja yang berinteraksi dengan agen), dan aksi (apa yang agen bisa lakukan). Tujuan utama reinforcement *machine learning* adalah bagi agen untuk menentukan aksi apa yang memaksimalkan hasil dalam waktu yang ditentukan. Penerapannya biasanya pada robotik, pembuatan game, dan navigasi.

Implikasi dari *machine learning* ini mampu membawa atau memberikan dampak pada berbagai bidang, terbukti bahwa *machine learning* ini mulai diterapkan pada berbagai bidang seperti contohnya yaitu memberi rekomendasi film di Netflix.

1.4 Kecerdasan Buatan dan Turunannya

1.4.1 Problematika

"Menggunakan bahasa kalian sendiri, jelaskan kaitan antara artificial intelligence, machine learning, dan deep learning".

1.4.2 Solusi

Artificial intelligence, machine learning, dan deep learning, sekilas istilah ketiga komponen ini merupakan istilah futuristik yang sangat banyak digunakan di masa depan. Namun terdapat perbedaan dari ketiga istilah tersebut. *Artificial intelligence* jika ditinjau dari namanya merupakan merupakan kerangka berpikir yang di desain menyerupai pola pikir manusia. Kemampuan ini pastinya mengandalkan alur berpikir matematis yang terstruktur untuk menyelesaikan *complex-problem* yang sulit di jangkau oleh pemikiran manusia yang terbatas oleh waktu dan tenaga.

Selanjutnya, jika meninjau istilah machine learning, istilah ini mengacu pada sekumpulan pola pikir dan kerangka berpikir penyusun *artificial intelligence*. Tidak hanya *machine learning*, *deep learning* pun juga demikian. *Deep learning* juga merupakan sekumpulan pola pikir penyusun *artificial intelligence*. Hal yang membedakan *machine learning* dan *deep learning* dapat dilihat dari istilah dan fungsinya.

Mengutip dari laman blog.udemy.com *"Deep learning and machine learning are both subsets of artificial intelligence and deep learning is a subset of machine learning. Machine learning is an AI technique, and deep learning is a machine learning technique".* Pada *deep learning*, istilah *deep* mengacu pada pembelajaran yang lebih dalam daripada machine learning. Hal ini mengasumsikan bahwa *deep learning* digunakan untuk mencari solusi yang lebih kompleks dibandingkan dengan *machine learning*.

Sistematika *artificial intelligence* dapat berupa perintah atau *command* sederhana. Sebagai contoh, seorang programmer akan membuat algoritma untuk menentukan apakah input yang dimasukkan merupakan bilangan bulat atau bilangan prima. Secara sederhana, tugas programmer hanya membuat code dan algoritma yang sesuai. Dalam konteks ini, *artificial intelligence* yang dibuat hanya akan memproses sesuai dengan yang konteks permasalahan yang ada tanpa dapat belajar kembali.

Berbeda dengan *machine learning* dan *deep learning*, mereka dapat belajar dan mengenali masalah kompleks dan dapat menjadi lebih pintar tanpa campur tangan manusia dari waktu ke waktu jika terus berlatih. Sebagai contoh, translator yang disediakan oleh Google atau biasa disebut Google Translate jika terdapat arti kata yang kurang tepat, pengguna dapat memberikan update terhadap sistem dan memungkinkan sistem mengenali arti kata yang lebih tepat. Contoh lain terdapat pada sistem rekomendasi YouTube. Mesin akan mengenali kegiatan pengguna selama menggunakan aplikasi YouTube terhadap apa yang dinikmatinya di YouTube dan terus belajar untuk memberikan rekomendasi video yang cocok untuk pengguna.

Machine learning dan *deep learning* dapat belajar dan meningkatkan kecerdasan secara otomatis tanpa di program secara eksplisit. Namun diantara keduanya, terdapat perbedaan yang signifikan. Secara umum, *deep learning* merupakan pengembangan dari *machine learning* untuk mengatasi permasalahan yang lebih kompleks atau permasalahan yang kurang cocok jika menggunakan *machine learning*. Pada *machine learning*, terdapat *feature extraction* yang harus dipikirkan seseorang sebelum

mengolah data. Informasi yang masih abstrak dan besar volumenya akan diringkas secara manual untuk mendapatkan informasi yang lebih efektif dan efisien dan kaya akan informasi yang dibutuhkan dari data tersebut. Sementara itu, *deep learning* sendiri tidak memerlukan *feature extraction* yang artinya *deep learning* dapat langsung memproses suatu data abstrak tersebut. *Deep learning* umumnya diterapkan pada sekumpulan data yang sangat besar (*big data*) dan lebih efektif digunakan dibandingkan dengan *machine learning*.

1.5 Interpretasi Data

1.5.1 Problematika

"Apakah yang kalian ketahui mengenai interpretasi data? Bagaimana signifikansi dan tantangannya? Bagaimana kaitan interpretasi data dengan data story telling dan decision making".

1.5.2 Solusi

Interpretasi data adalah proses meninjau serta menjelaskan data sesuai dengan tujuannya, tak hanya menjelaskan, tetapi memberikan gambaran, penafsiran, dan mampu memberi pemahaman kepada siapa yang dituju. Interpretasi ini diharapkan mampu meminimalisir kesalahpahaman dan kesalahan penafsiran. Interpretasi data memberikan makna pada informasi yang dianalisis dan menentukan signifikansi dan implikasinya.

Interpretasi data memiliki peranan yang sangat penting, terlebih untuk meminimalisir kesalahpahaman dan kesalahan penafsiran karena melalui interpretasi ini dapat terciptanya sebuah kesimpulan atau bahkan pengambilan keputusan. Mengutip dari lama Minera, interpretasi data sebagian besar digunakan untuk pengambilan keputusan yang terinformasi dan memprediksi tren dan perilaku yang akan datang. Sumber daya berharga lainnya yang dapat anda manfaatkan dengan interpretasi data adalah mengidentifikasi masalah dan solusi.

Interpretasi data tentu memiliki tantangan dalam praktiknya, menurut datapine dapat dikatakan bahwa masalah interpretasi data tertentu atau "perangkap" ada dan dapat terjadi saat menganalisis data. Hal ini juga tidak menutup kemungkinan kesalahan dalam menampilkan interpretasi dalam grafik, seperti salah memilih grafik.

Berdasarkan yang sudah dijelaskan sebelumnya, bahwa interpretasi data ini tentu memiliki kaitan dengan data *story telling* karena data *story telling* adalah proses bercerita menjelaskan hasil dari interpretasi yg kita punya dengan harapan tujuan dan

pesan tertentu dapat tersampaikan dengan baik, pun demikian dengan *decision making* karena melalui interpretasi data diharapkan tercipta sebuah kesimpulan yang dapat membantu dalam pengambilan keputusan yang baik.

DAFTAR PUSTAKA

Basalamah, Salsabila. (2020). Cara Mengidentifikasi dan Penanganan Data Outlier. https://salsabilabasalamah.medium.com/cara-mengidentifikasi-dan-penanganan-data-outlier-d2fe16c6d62c

Bertan, Cindy Viane dkk. (2016). Pengaruh pendayagunaan sumber daya manusia (tenaga kerja) terhadap hasil pekerjaan. *Jurnal sipil statistik*. Vol 4. No 1

DQlab. (2020). Jenis - Jenis Machine Learning yang Harus Kamu Pahami. https://www.dqlab.id/jenis-machine-learning-yang-perlu-diketahui

Iriondo, Roberto. (2020). *What is Machine Learning?*. https://pub.towardsai.net/what-is-machine-learning-ml-b58162f97ec7

Lebied, Mona. (2018). *A Guide To The Methods, Benefits & Problems of The Interpretation of Data.* https://www.datapine.com/blog/data-interpretation-methods-benefits-problems/

Minera. *The significance of data interpretation for your business.* *https://www.minerra.net/data-interpretation-methods/*

Rahmalia, Nadiyah. (2021). Kenalan dengan Machine Learning, Sebuah Cabang Ilmu Kecerdasan Buatan. https://glints.com/id/lowongan/machine-learning/#.YO1BSugzaMp

Soemartini. (2007). *Pencilan (Outlier)*. Bandung: UNPAD.

BAGIAN 2

DATASET: ANALISIS DATASET

THINGAMAJIG

2021

2.1 Latar Belakang

Pandemi Covid-19 di Indonesia telah membawa negara Indonesia dalam kesedihan hingga saat ini. Pemerintah dalam upaya membatasi penyebaran Covid-19 di Indonesia yaitu dengan menerapkan beberapa kegiatan yang membatasi pergerakan masyarakat seperti Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Selain itu, masyarakat juga telah dihimbau untuk selalu menjaga jarak, tidak berkerumun dan menggunakan *double-masker* ketika keluar rumah. Tujuannya adalah menghindari penyebaran dan membuat *cluster* baru Covid-19 di Indonesia.

Jumlah akumulasi pasien terkonfirmasi positif yang semakin bertambah membuat beberapa hal menjadi buruk seperti kurangnya ketersediaan kamar perawatan dan isolasi di rumah sakit dan berkurangnya lahan pemakaman Covid-19 di beberapa daerah di Indonesia. Jumlah akumulasi pasien sembuh dari Covid-19 di Indonesia per tanggal 18 Juli 2021 berdasarkan situs Satuan Gugus Tugas Penanganan Covid-19 di Indonesia sebesar 2.877.476 jiwa. Di sisi lain, akumulasi pasien sembuh dari covid-19 di Indonesia juga semakin meningkat berdasarkan data dari https://corona.jakarta.go.id/id. Meninjau data yang ada yakni data periode 29 Februari 2020 hingga 29 Juni 2021, nilai *positive rate cases* semakin meningkat. Nilai terbesar selama periode yaitu 48,05% yang berarti 48 dari 100 orang beresiko positif Covid-19 di Jakarta. Menurut laman https://megapolitan.kompas.com/, *World Health Organization* (WHO) memberikan batasan maksimal untuk nilai *positive rate cases* yaitu sebesar 5%. Dengan melihat data tersebut, dapat diketahui bahwa Indonesia sedang berada dalam keadaan yang kurang baik terlebih lagi dalam upaya menekan penyebaran virus Covid-19 diperlukan upaya dari berbagai pihak untuk membuat Indonesia normal kembali.

Pemerintah juga menerapkan program vaksinasi untuk menekan penyebaran Covid-19. Vaksinasi sudah dilakukan di beberapa daerah. Mengutip dari laman https://www.presidenri.go.id/ , kepala negara menyampaikan hingga saat ini Indonesia telah memesan kurang lebih sebanyak 329,5 juta dosis vaksin. Dalam upaya melakukan vaksinasi, tentu terdapat beberapa masalah di lapangan yang menyebabkan sebagian masyarakat enggan untuk di vaksin. Rumor yang beredar dan berita *hoax* merupakan salah satu penyebab masyarakat enggan mengikuti vaksinasi dan hal ini menyebabkan stigma di masyarakat bahwa vaksinasi tidaklah baik untuk tubuh.

Atas dasar hal tersebut, kami membuat laporan ini untuk menganalisis lebih lanjut mengenai perkembangan Covid-19 di Indonesia khususnya Jakarta. Data science merupakan cabang ilmu yang menggabungkan beberapa disiplin ilmu eksakta dengan tujuan mencari pengetahuan mendalam mengenai suatu data dari suatu permasalahan tertentu. Komputasi

numerik, statistika matematika, *business insight* merupakan salah satu hal yang wajib dikuasai oleh seorang data science. Dalam kesempatan ini, Thingamajig akan mencari dan menganalisis mengenai problematika dari dataset yang telah disediakan. Dataset berasal dari Compfest 2021 berupa data mengenai Covid-19. Secara inovatif, akan dicari beberapa hal yang merupakan kunci untuk mengetahui karakteristik suatu data yang diberikan serta mendapatkan kesimpulan dari data tersebut. Dalam laporan ini, Python merupakan bahasa pemrograman yang digunakan dengan IDE Jupyter Notebook. Analisis yang dilakukan merupakan analisis deskriptif dengan judul **"Analisis Informasi Data Agregat Covid-19 di Jakarta"**. Penulisan dilakukan merujuk pada pertanyaan yang diberikan sebelumnya dan batasan masalah digunakan untuk menghindari persepsi secara luas.

2.2 Jawaban Soal

1. *Dari dataset yang disediakan, temukan nilai mean, median, dan modus dari positif Covid-19 harian Jakarta.*

   Solusi :

   Berdasarkan data dari tanggal 1 Maret 2021 sampai 30 Juni 2021, dari data positif Covid-19 harian Jakarta dapat melihat ringkasan numerik menggunakan ukuran pusat data atau *measures central tendency* untuk menggambarkan posisi sentral dan distribusi frekuensi untuk sekelompok data yang dapat dideskripsikan menggunakan *mean* (rata-rata), *median* (nilai tengah) dan *modus* (nilai yang sering muncul).

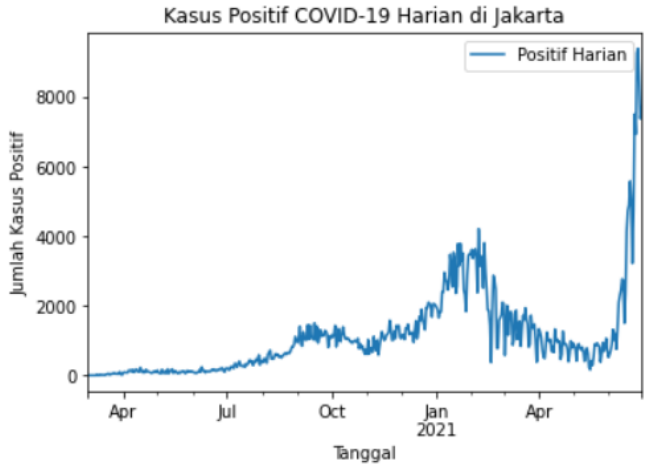
   Nilai *mean* didapat dari jumlah nilai observasi dalam sebuah data dibagi dengan jumlah observasi, sehingga nilai *mean* dari positif Covid-19 harian adalah 1115,95. Nilai *median* didapat dari nilai tengah dalam daftar yang sudah diurutkan, sehingga nilai *median* dari positif Covid-19 harian adalah 845. Nilai *modus* didapat dari nilai yang paling sering muncul dan frekuensinya tertinggi, sehingga nilai *modus* positif Covid-19 harian adalah 0.

   Jika diasumsikan pada tanggal yang sama dengan mengabaikan nilai 0, maka diperoleh nilai *mean* adalah 1132,22, nilai *median* adalah 864,5 dan nilai *modus* adalah 127. Dari hasil diatas terdapat perbedaan nilai *mean*, *median* dan *modus* jika terdapat nilai 0 dan jika diasumsikan mengabaikan nilai 0.

2. *Dari dataset yang disediakan, temukan nilai minimal dan maksimal dari positif COVID-19 harian Jakarta.*

   Solusi :

Berdasarkan data per 30 Juni 2021, ditemukan nilai minimal dan maksimal dari positif Covid-19 harian Jakarta sebesar 0 dan 9394. Trend kasus positif Covid-19 Jakarta dapat dilihat pada gambar di bawah.



Gambar 4. Grafik Kasus Positif COVID-19 Harian di Jakarta

Melihat grafik di atas, diketahui bahwa nilai minimal yang bernilai 0 ini terjadi di awal-awal 2020 saat virus Covid-19 ini baru mulai menyebar khususnya di Jakarta. Hal ini mengingat pada awal pandemi belum banyak tes Covid-19 yang dilakukan. Sedangkan untuk nilai maksimal yang bernilai 9394 terjadi baru-baru ini, lebih tepatnya pada bulan Juni 2021. Mengutip dari laman Kompas per 13 Juli 2021, wakil gubernur DKI Jakarta Ahmad Riza Patria mengatakan, tingginya kasus harian Covid-19 di ibu kota disebabkan oleh tingkat tes PCR di Jakarta yang cukup tinggi. Pak Riza juga menambahkan bahwa tes PCR di ibu kota sudah lebih tinggi 20 kali lipat dari standar yang ditetapkan WHO.

3. *Dari dataset yang disediakan, temukan nilai-nilai outlier yang ada (menggunakan variabel yang kalian tentukan).*

Solusi :

Berdasarkan pengertian outlier, outlier merupakan observasi atau data poin yang nilainya berbeda atau jauh daripada observasi pada umumnya. Berikut telah diperoleh beberapa outlier dalam dataset “Daily Update Data Agregat Covid-19 Jakarta”. Pendeteksian outlier diterapkan pada data harian “Daily Update Data Agregat Covid-19 Jakarta” dengan menggunakan rumus pencarian outlier pada persamaan sebelumnya.

a. Data Pemakaman periode 01-03-2020 sampai 29-06-2021

Dalam sheet pemakaman, diambil data harian pemakaman Covid-19 harian dan data pemakaman umum harian. Dengan menggunakan formula IQR (lampiran 3)

diperoleh jumlah outlier pemakaman Covid-19 harian sebanyak 22 dan pemakaman umum harian sebanyak 18.

b. Data Hasil Lab 29-02-2020 sampai 29-06-2021

Dalam sheet hasil lab, diambil data harian *positivity rate* kasus baru harian. Dengan menggunakan formula IQR (lampiran 4) diperoleh jumlah outlier *positivity rate* kasus baru harian sebanyak 28.

c. Data Suspek (17-07-2020 sampai 30-06-2021)

Dalam sheet data suspek, diambil sekumpulan data penjumlahan isolasi harian dari kolom "Isolasi di RS (Discarded)", "Isolasi di Rumah (Discarded)", "Isolasi di RS (Kontak Erat)", "Isolasi di Rumah (Kontak Erat)", "Isolasi di RS (Pelaku Perjalanan)","Isolasi di Rumah (Pelaku Perjalanan)", "Isolasi di RS (Probable)", "Isolasi di Rumah (Probable)", "Isolasi di RS (Suspek)", dan "Isolasi di Rumah (Suspek)". Dengan menggunakan formula IQR (lampiran 5) diperoleh jumlah outlier isolasi harian sebanyak 24.

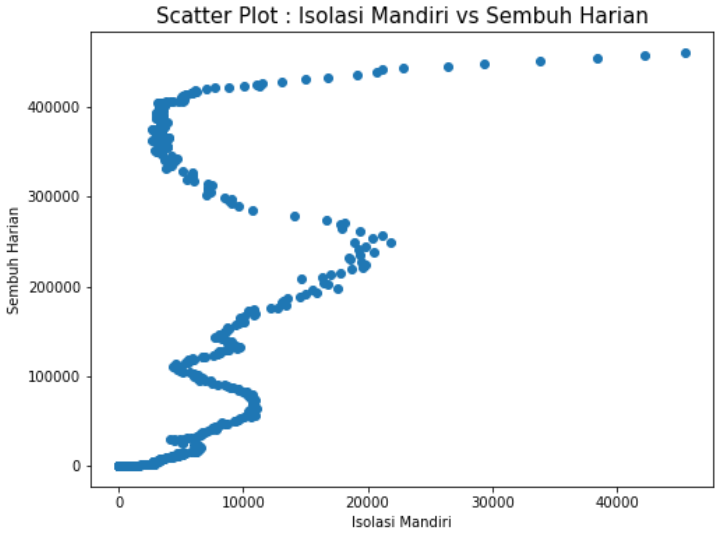
d. Data Jakarta (01-03-2020 sampai 30-06-2021)

Dalam sheet data Jakarta, diambil data positif harian, sembuh harian, tanpa gejala, dan bergejala. Dengan menggunakan formula IQR (lampiran 6) diperoleh jumlah outlier positif harian sebanyak 36, sembuh harian sebanyak 25, tanpa gejala sebanyak 31, dan bergejala sebanyak 7.

4. *Dari dataset yang disediakan, usulkan dua buah variabel dan berikan analisis korelasi antara kedua variabel tersebut. Jelaskan apa kesimpulan yang dapat diambil berdasarkan analisis kalian.*

Solusi :

Dari variabel yang kami analisis yakni isolasi mandiri dengan sembuh harian didapatkan bahwa Isolasi mandiri dengan sembuh harian memiliki nilai korelasi sebesar 0,983429 mengacu pada data per 30 Juni 2021. Hubungan dari dua variabel ini dapat dilihat pada grafik *scatter plot* dibawah ini.



Gambar 5. Grafik Scatter Plot Isolasi Mandiri vs Sembuh Harian

Hal ini berarti banyak penderita Covid-19 yang memilih untuk isolasi mandiri kemudian dinyatakan sembuh. Isolasi mandiri ini memang menjadi salah satu pilihan terbaik dalam upaya penyembuhan Covid-19, mengingat saat ini sulit mendapatkan perawatan di rumah sakit karena tingginya kasus yang terjadi.

2.3 Hasil Analisis Tambahan

2.3.1 PROBLEM STATEMENT

1. Apakah terdapat perbedaan nilai rataan vaksin 1 dan 2 di tiap kota/kabupaten di DKI Jakarta?
2. Kota/kabupaten mana yang memiliki nilai jumlah vaksin tertinggi dan terendah di DKI Jakarta?
3. Bagaimana perbandingan dari vaksin ke-1 yang telah dilakukan di tiap-tiap kota/kabupaten di DKI Jakarta?
4. Bagaimana perbandingan dari vaksin ke-2 yang telah dilakukan di tiap-tiap kota/kabupaten di DKI Jakarta?
5. Bagaimana clustering dari vaksin di tiap kota/kabupaten di DKI Jakarta?
6. Apakah jumlah akumulasi kasus positif pasien Covid-19, kasus pasien sembuh dari Covid-19, dan pasien isolasi mempengaruhi tingkat kematian akibat Covid-19 di DKI Jakarta ?
7. Bagaimana model yang dibentuk oleh variabel jumlah akumulasi kasus positif pasien Covid-19, pasien sembuh dari Covid-19, dan pasien isolasi jika faktor-faktor tersebut mempengaruhi tingkat kematian akibat Covid-19 di DKI Jakarta ?

2.3.2 HYPOTHESIS

- Uji hipotesis untuk rata-rata vaksin

  H_0: Rata-rata total vaksin di Jakarta Barat = Jakarta Pusat = Jakarta Selatan = Jakarta Timur = Jakarta Utara = Kepulauan Seribu

  H_a: Setidaknya ada satu pasang daerah yang memiliki rata-rata total vaksin yang tidak sama

- Uji hipotesis untuk regresi

  H_0: Tidak terdapat pengaruh antara jumlah akumulasi kasus positif pasien Covid-19, pasien sembuh dari Covid-19, dan pasien isolasi dengan tingkat kematian akibat Covid-19 di DKI Jakarta

  H_a: Terdapat pengaruh antara jumlah akumulasi kasus positif pasien Covid-19, pasien sembuh dari Covid-19, dan pasien isolasi dengan tingkat kematian akibat Covid-19 di DKI Jakarta

2.3.3 EXPLORATORY DATA ANALYSIS

1. ***Data cleansing* dan *data transformation* pada data vaksin DKI Jakarta**

   Sumber data yang digunakan pada penelitian ini berasal dari https://tiny.cc/Datacovidjakarta. Data yang digunakan adalah data pada sheet vaksinasi wilayah yang berisikan data tentang vaksinasi wilayah khususnya di tiap kota/kabupaten di DKI Jakarta. Data yang digunakan hanya data pada tanggal 12 Juni 2021 hingga 30 Juni 2021. Alasan utama mengapa hanya dipilih tanggal 12 Juni 2021 hingga 30 Juni 2021 ialah untuk menghindari outlier, karena sebelum tanggal 12 Juni 2021 merupakan data akumulasi yang memiliki nilai sangat besar dan akan menimbulkan outlier.

   Pada sumber data, data memiliki multi index dan multi kolom sehingga akan lebih rumit untuk mengolahnya. Oleh karena itu diputuskan untuk mengubah bentuk tabel data atau *data transformation* menjadi index tunggal dan kolom tunggal untuk masing-masing index dan kolom yang dipilih. Index merupakan runtun waktu dari 12 Juni 2021 hingga 30 Juni 2021, sedangkan kolom yang dipilih merupakan nilai dari jumlah vaksin ke-1 dan vaksin ke-2 untuk tiap-tiap kota/kabupaten di DKI Jakarta. Kami juga menambahkan kolom baru yang berisi nilai total dari penjumlahan vaksin ke-1 dan vaksin ke-2. Untuk tipe data, semua kolom bertipe *int64* dan tidak ada data yang *missing*.

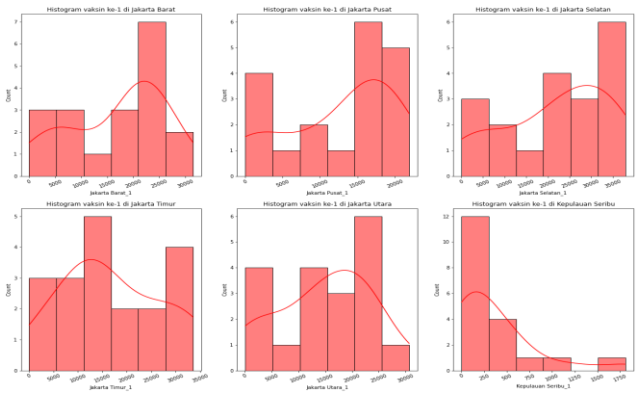
2. **Mencari statistik deskriptif pada data vaksin DKI Jakarta**

Dalam data tentu dapat dicari statistik deskriptif dari data tersebut, biasanya statistik deskriptif ini berisi nilai rata-rata (*mean*), nilai tengah (*median*), nilai terkecil, nilai terbesar, dan lainnya. Pada statistik deskriptif data vaksin wilayah DKI Jakarta didapatkan untuk nilai rata-rata (*mean*) terbesar pada vaksin ke-1 adalah wilayah Jakarta Selatan dengan rataan sebesar 12393,913921 atau jika dibulatkan menjadi 12394, sedangkan untuk nilai rata-rata (*mean*) terkecilnya adalah wilayah Kepulauan Seribu sebesar 460,692772 atau 461. Untuk vaksin ke-2 wilayah Jakarta Timur memiliki nilai rata-rata (*mean*) terbesar yakni 719,212939 atau jika dibulatkan menjadi 719. Sedangkan untuk wilayah yang memiliki rata-rata (*mean*) terkecil untuk vaksin ke-2 kembali didapatkan oleh wilayah Kepulauan Seribu yakni sebesar 21,598543. Untuk lebih lengkap dapat dilihat pada lampiran.

3. Melihat distribusi sebaran pada data vaksin DKI Jakarta

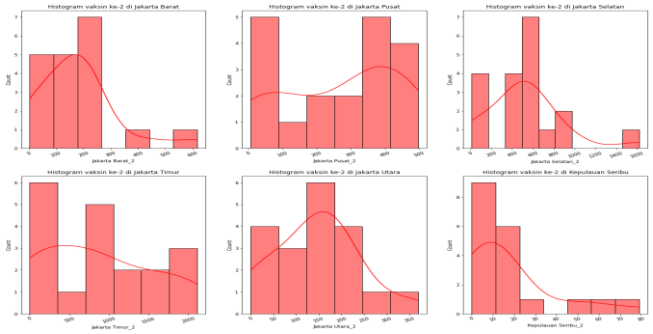
Jika melihat data yang ada dapat diketahui bahwa data berupa numerik. Salah satu cara untuk mengetahui distribusi dari data yaitu dengan melihat histogram data tersebut. Pada dasarnya histogram sangat membantu untuk mengetahui apakah data tersebut berdistribusi normal atau tidak, sehingga kita bisa memikirkan langkah selanjutnya (Andreas Chandra, 2019).

Dengan menggunakan histogram pada data vaksin wilayah DKI Jakarta didapatkan hasil untuk vaksin ke-1 sebagai berikut.



Gambar 6. Histogram Vaksin ke-1 di tiap Kota/Kabupaten di DKI Jakarta

Pada Gambar 6 diatas dapat diketahui bahwa data untuk vaksin ke-1 sebagian besar tidak berdistribusi normal untuk tiap kota/kabupaten di DKI Jakarta. Distribusinya juga terlihat tidak seimbang, beberapa kota/kabupaten ada yang cenderung miring ke kanan dan ada juga yang cenderung miring ke kiri. Untuk sebaran vaksin ke-2 yakni sebagai berikut.



Gambar 7. Histogram Vaksin ke-2 di tiap Kota/Kabupaten di DKI Jakarta

Seperti sebelumnya, pada Gambar 7 diatas dapat diketahui bahwa data untuk vaksin ke-2 juga sebagian besar tidak berdistribusi normal untuk tiap kota/kabupaten di DKI Jakarta. Distribusinya juga terlihat tidak seimbang seperti pada vaksin ke-1, beberapa kota/kabupaten ada yang cenderung miring ke kanan dan ada juga yang cenderung miring ke kiri. Karena data memiliki skala yang tidak jauh berbeda jadi data tidak diubah menjadi berdistribusi normal atau tidak juga di standardisasi.

4. Melihat hubungan atau korelasi masing-masing variabel pada data vaksin DKI Jakarta

Analisis korelasi menjelaskan ada atau tidaknya hubungan antar dua variabel. Nilai korelasi bisa positif atau negatif atau lemah. Korelasi positif yang artinya jika penambahan pada nilai X maka bertambah juga nilai Y. Korelasi negatif menjelaskan hubungan setiap kenaikan nilai X maka ada penurunan pada nilai Y. Korelasi yang lemah menjelaskan dua variabel ini tidak ada hubungannya sama sekali.

Dengan menggunakan data yang disediakan dan telah di transformasi sebelumnya, dapat diketahui beberapa hubungan yang ada di tiap masing-masing variabel wilayah dengan variabel wilayah lainnya. Hasil korelasi yang didapatkan yakni sebagai berikut.

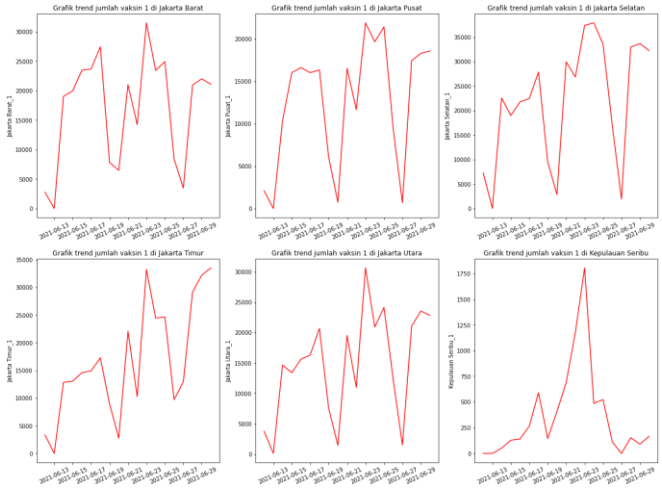
A. Untuk setiap variabel wilayah cakupan vaksin ke-1 dengan wilayah vaksin ke-1 lainnya terlihat memiliki nilai korelasi diatas 0,7 atau bisa dikatakan korelasi nya kuat, kecuali untuk wilayah Kepulauan Seribu dengan wilayah lainnya dalam cakupan vaksin ke-1 memiliki korelasi di bawah 0,5. Hal ini sama persis untuk bagian korelasi vaksin ke-1 dengan vaksin total, yakni hanya Kepulauan Seribu yang memiliki nilai korelasi dibawah 0,5.
B. Untuk setiap variabel wilayah cakupan vaksin ke-2 dengan wilayah vaksin ke-2 lainnya terlihat dominan nilai korelasi dibawah 0,5 atau bisa dikatakan korelasi

nya cukup bahkan lemah, namun beberapa wilayah seperti Jakarta Barat dengan Jakarta Pusat, Jakarta Barat dengan Jakarta Timur, dan Jakarta Pusat dengan Jakarta Timur mendapatkan nilai korelasi diatas 0,7 atau kuat.

Tingkat kuat lemahnya korelasi mengacu pada tabel korelasi pada soal non dataset sebelumnya. Untuk korelasi lebih jelasnya dapat dilihat pada lampiran.

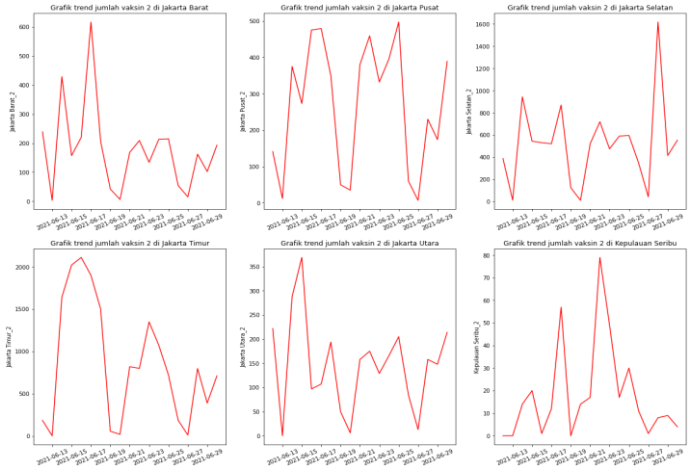
5. **Melihat tren data vaksin Covid-19 melalui grafik trennya dan melihat proporsi masing-masing wilayah di DKI Jakarta**

Salah satu cara sederhana untuk melihat tren data yakni dengan menggunakan grafik tren dari data tersebut. Dalam kasus vaksin di DKI Jakarta dapat dilihat tren sebagai berikut.



Gambar 8. Tren Vaksin ke-1 di tiap Kota/Kabupaten di DKI Jakarta

Berdasarkan gambar diatas tren vaksin ke-1 cenderung mirip. Untuk vaksin ke-2 yakni sebagai berikut.



Gambar 9. Tren Vaksin ke-2 di tiap Kota/Kabupaten di DKI Jakarta

Untuk vaksin ke-2 ini cenderung berbeda-beda tidak seperti vaksin ke-1 yang cenderung mirip. Selanjutnya untuk mengetahui berapa proporsi masing-masing wilayah pada setiap cakupan vaksin maka digunakan pie chart sebagai berikut.

Gambar 10. Tren Vaksin ke-2 di tiap Kota/Kabupaten di DKI Jakarta

Proporsi dapat diartikan sebagai perbandingan tiap bagian dengan bagian lain atau bagian keseluruhan.

6. ***Data insights* untuk mengetahui pengaruh jumlah pasien positif Covid-19, pasien sembuh dari Covid-19, pasien isolasi terhadap jumlah kematian akibat Covid-19 di DKI Jakarta**

Untuk mengetahui hubungan antar variabel ini, digunakan data yang berasal dari *problem statement* yang diberikan oleh pihak penyelenggara (COMPFEST). Data diambil dari *sheet* Indonesia dan Jakarta periode 1 Maret 2020 sampai dengan 30 Juni 2021 untuk kolom "Tanggal", "Sembuh (Jakarta)", "Positif (Jakarta)", "Self-Isolation (Jakarta)", dan "Meninggal (Jakarta)".

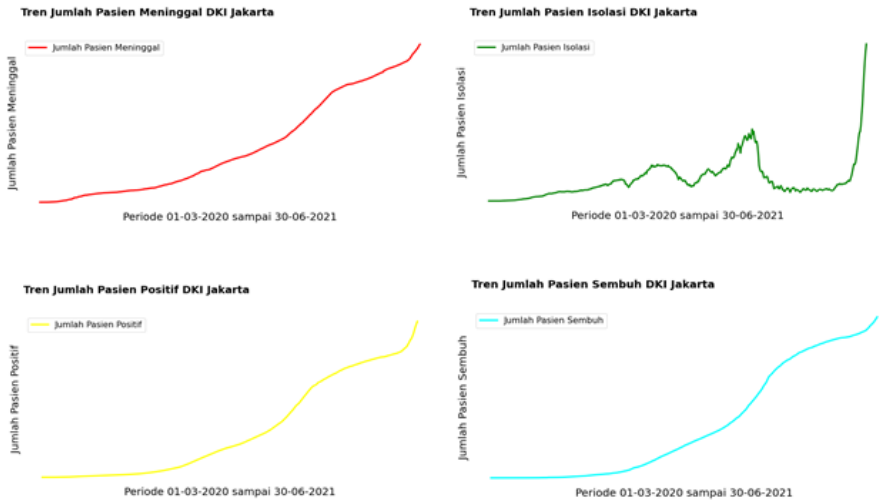
Dalam dataset ini, tipe data berjenis *int64* dan secara deskriptif dengan menggunakan korelasi diperoleh kekuatan hubungan antar variabel sebagai berikut.

	Sembuh (Jakarta)	Positif (Jakarta)	Self-Isolation (Jakarta)	Meninggal (Jakarta)
Sembuh (Jakarta)	1.000000	0.998956	0.346783	0.996246
Positif (Jakarta)	0.998956	1.000000	0.388611	0.997303
Self-Isolation (Jakarta)	0.346783	0.388611	1.000000	0.385783
Meninggal (Jakarta)	0.996246	0.997303	0.385783	1.000000

Gambar 11. Korelasi Antar Variabel

Berdasarkan hasil korelasi pada gambar 11, dapat dilihat bahwa tingkat korelasi antara variabel positif Jakarta - sembuh Jakarta, variabel positif Jakarta - meninggal Jakarta, dan sembuh Jakarta - meninggal Jakarta memiliki korelasi positif yang sangat tinggi yaitu lebih dari 0,9. Berbeda dengan variabel *self-isolation* dengan variabel lain yang berada dalam gambar memiliki tingkat korelasi yang tidak mencapai 0,4. Berdasarkan tabel hubungan korelasi pada pembahasan sebelumnya,

hal ini menunjukkan bahwa variabel *self-isolation* berpengaruh rendah terhadap variabel kematian akibat Covid-19 di Jakarta.



Gambar 12. Tren Perbandingan Antar Variabel

Selain itu dapat dilihat juga bahwa tren yang terjadi pada variabel pasien meninggal, pasien positif, dan pasien sembuh secara visualisasi cenderung serupa mengalami kenaikan tanpa terjadi penurunan. Berbeda dengan variabel *self-isolation*, ini membuktikan bahwa variabel ini tidak linear dengan ketiga variabel lainnya.

7. **Menguji hipotesis rata-rata total vaksin di daerah Jakarta Barat, Jakarta Pusat, Jakarta Selatan, Jakarta Timur, Jakarta Utara dan Kepulauan Seribu**

Seperti yang sudah diketahui sebelumnya, bahwa data vaksin di DKI Jakarta menunjukan data yang tidak normal. Sehingga untuk menguji hipotesis rata-rata membutuhkan uji untuk statistik non-parametrik. Salah satu uji hipotesis rata-rata yang dapat digunakan untuk statistik non parametrik adalah uji *kruskal wallis*. Uji *kruskal wallis* adalah uji yang digunakan untuk menguji hipotesis rata-rata lebih dari 2 populasi. Dikatakan tolak H_0 ketika P-Value ≤ alpha, nilai alpha yang digunakan adalah 0,05. Nilai rata-rata total suntik vaksin berdasarkan daerah sebagai berikut:

Tabel 3. Tabel Rata-Rata Total Vaksin di DKI Jakarta

Kota/Kabupaten	Rata-Rata Total Vaksin
Jakarta Barat	17101,631579
Jakarta Pusat	12907,052632

Jakarta Selatan	22455,052632
Jakarta Timur	17698,105263
Jakarta Utara	14939,947368
Kepulauan Seribu	383,105263

Dari hasil uji rata-rata diatas menghasilkan nilai P-Value menggunakan uji *kruskal wallis* sebesar 0,415, karena nilai P-Value < 0,05 maka untuk uji hipotesis ini tolak H_0.

2.3.4 INITIAL FINDINGS

1. **Hubungan wilayah dengan wilayah lainnya pada data vaksin di DKI Jakarta**

Seperti yang sudah dibahas sebelumnya pada proses EDA, diketahui bahwa korelasi pada cakupan vaksin ke-1 hampir semua wilayah memiliki korelasi kuat, kecuali jika dipasangkan dengan Kepulauan Seribu. Hal ini dapat disebabkan oleh keadaan geografis atau kedekatan antara wilayah satu dengan wilayah lainnya, sebagaimana kita tahu dalam program percepatan vaksinasi ini sedang marak program vaksinasi massal. Vaksinasi massal sendiri tentu akan sangat ramai dikunjungi, dengan memanfaatkan kedekatan wilayah ini diduga masyarakat lebih memilih mencari alternatif wilayah di luar wilayah masing-masing atau bisa dikatakan vaksin di wilayah lain. Keadaan ini juga diperkuat mengingat saat ini sudah banyak tempat-tempat yang mengadakan vaksin tanpa surat domisili seperti contohnya bandara (laman https://www.suara.com/ , 2021).

2. **Pola tren data vaksin di DKI Jakarta cenderung mirip**

Sebelumnya dapat terlihat jelas untuk tren data dari setiap cakupan vaksin di masing-masing wilayah. Jika diamati lebih cermat khususnya untuk vaksin ke-1, dapat diketahui bahwa pola tren untuk vaksin ke-1 di setiap wilayah kota/kabupaten di DKI Jakarta cenderung mirip, kecuali untuk Kepulauan Seribu. Secara garis besar pola tren akan naik lalu turun sangat jauh sebanyak dua kali. Penurunan besar ini terjadi pada rentang tanggal 19 Juni 2021-20 Juni 2021 dan rentang tanggal 26 Juni 2021 - 27 Juni 2021. Kecenderungan kemiripan pada pola tren ini dapat disebabkan salah satunya karena program vaksinasi massal serentak di beberapa

wilayah Jakarta. Dimana kita ketahui bahwa program vaksinasi sedang marak-maraknya digencarkan mengingat bahwa pemerintah ingin agar vaksinasi dipercepat dengan target 1 juta dosis perhari secara nasional (laman https://nasional.kompas.com/, 2021). Dugaan lainnya yakni jenis vaksin mempengaruhi tren.

3. Proporsi wilayah pada data vaksin di DKI Jakarta

Pada data terdapat 6 wilayah dengan masing-masing proporsi. Jika melihat pada *pie chart* sebelumnya, pada vaksin ke-1 diketahui yang memiliki proporsi terbesar adalah Jakarta Selatan dengan proporsi sebesar 26,3%, kemudian Jakarta Barat dengan 20,3%, Jakarta Timur dengan 20,2%, Jakarta Utara 17,7%, Jakarta Pusat dengan 15,1%, dan terakhir yakni Kepulauan seribu dengan 0,4%. Dapat dilihat bahwa Jakarta Barat dan Jakarta Timur memiliki Proporsi yang hanya berbeda 0,1% artinya bisa dikatakan sangat kecil perbedaannya dan Kepulauan Seribu memiliki proporsi terkecil dan sangat jauh jika dibandingkan dengan 5 wilayah lainnya.

Kemudian untuk vaksin ke-2, dapat dilihat yang memiliki proporsi terbesar kali ini ialah Jakarta Timur dengan 43,2%, lalu Jakarta Selatan dengan 26%, Jakarta Pusat dengan 13,6 %, Jakarta Barat dengan 9%, Jakarta Utara 7,4%, dan Kepulauan Seribu dengan 0,9%. Pada vaksin ke-2 ini masing-masing wilayah memiliki perbedaan yang lumayan besar sebagai contoh Jakarta Timur dan Jakarta Selatan memiliki perbedaan sebesar 17,2%. Hal ini juga menjelaskan antusias atau keikutsertaan masyarakat pada vaksin ke-2 tidak merata seperti vaksin ke-1.

Jika dijumlah vaksin ke-1 dan vaksin ke-2 maka proporsi yang didapatkan yakni ditempati oleh Jakarta Selatan dengan 26,3% pada posisi pertama, lalu Jakarta Timur dengan 20,7%, Jakarta Barat dengan 20%, Jakarta Utara dengan 17,5%, Jakarta Pusat dengan 15,1%, dan Kepulauan Seribu dengan 0,4%. Menarik untuk mengetahui bagaimana pengelompokan atas proporsi ini. Proporsi ini juga dipengaruhi salah satunya oleh jumlah penduduk di wilayah masing-masing.

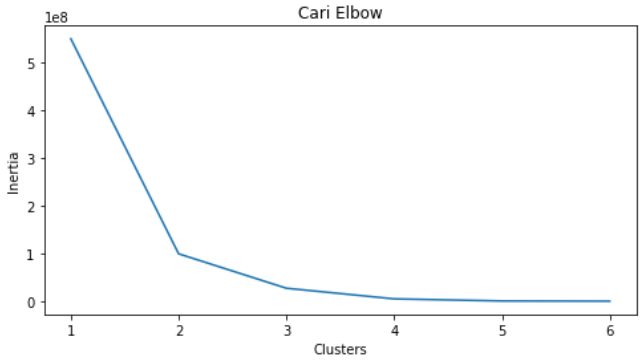
2.3.5 DEEP DIVE ANALYSIS

1. *Clustering* wilayah pada data vaksin di DKI Jakarta

Setelah sebelumnya membahas proporsi wilayah di DKI Jakarta, maka selanjutnya dilakukan analisis lanjutan mengenai clustering atau pengelompokkan berdasarkan kemiripan atribut. *Clustering* dianggap sebagai *unsupervised learning method* karena kami tidak memiliki kebenaran dasar untuk membandingkan output

dari algoritma pengelompokan dengan label yang sebenarnya untuk mengevaluasi kinerjanya. *Clustering* menggunakan algoritma K-Means karena lebih sederhana. K-Means adalah sebuah metode yang dikembangkan oleh Stuart Lloyd dari Bell Labs pada tahun 1957. K-Means mencoba membuat titik data *intra-cluster* semirip mungkin sambil juga menjaga *cluster* sejauh mungkin berbeda. Semakin sedikit variasi yang kita miliki dalam *cluster*, semakin homogen (mirip) titik data berada dalam *cluster* yang sama (Imad Dabbura, 2018).

Pada data vaksin di DKI Jakarta *clustering* dilakukan dengan atribut yang merupakan nilai rataan dari masing-masing wilayah. Proses pertama yakni menentukan jumlah K atau banyaknya *cluster* pada K-Means. Metode Elbow dapat digunakan dan berpotensi membantu dalam mencari jumlah K *cluster*. Metode Elbow memberi kita gambaran tentang berapa jumlah K *cluster* yang baik akan didasarkan pada jumlah jarak kuadrat antara titik data dan centroid *cluster* yang ditetapkan. Nilai K yang didapatkan pada data vaksin di DKI Jakarta yakni bernilai 3, untuk grafiknya dapat dilihat sebagai berikut.



Gambar 13. Grafik Elbow

Dapat dilihat pada gambar grafik diatas, garis terus turun atau berubah jauh namun ketika mencapai K=3 perubahannya menjadi perlahan, oleh karena itu dipilih K=3 untuk data kali ini. Proses K-Means dilanjutkan sampai terbentuk *cluster-cluster* dengan masing-masing label. Hasilnya dapat dilihat pada gambar dibawah ini.

Kota / Kabupaten	Rataan Vaksin 1	Rataan Vaksin 2	Rataan Vaksin Total	Labels
Jakarta Barat	16923.526316	178.105263	17101.631579	0
Jakarta Pusat	12626.578947	269.105263	12895.684211	0
Jakarta Selatan	21938.052632	517.000000	22455.052632	2
Jakarta Timur	16840.947368	857.157895	17698.105263	0
Jakarta Utara	14793.421053	146.526316	14939.947368	0
Kepulauan Seribu	365.052632	18.052632	383.105263	1

Gambar 14. Gambar Tabel beserta Label *Cluster*

Karena K=3 maka *cluster* yang terbentuk sebanyak 3 cluster masing-masing dengan label 0, 1, dan 2. Masing-masing *cluster* yaitu sebagai berikut,

A. *Cluste*r label 0 atau wilayah dengan nilai rataan vaksin total sedang berisi Jakarta Barat, Jakarta Pusat, Jakarta Timur, dan Jakarta Utara.
B. *Cluster* label 1 atau wilayah dengan nilai rataan vaksin total rendah berisi hanya Kepulauan Seribu.
C. *Cluster* label 2 atau wilayah dengan nilai rataan vaksin total tinggi berisi hanya Jakarta Selatan.

Untuk lebih jelasnya dapat melihat pada lampiran. *Clustering* juga sudah dilakukan dengan atribut lain yakni data jumlah vaksin ke-1, vaksin ke-2, dan vaksin total namun memberikan hasil yang sama seperti vaksin dengan atribut rataan.

2. Regresi dan *Machine Learning* untuk mengetahui pengaruh jumlah pasien positif Covid-19, pasien sembuh dari Covid-19, pasien isolasi terhadap jumlah kematian akibat Covid-19 di DKI Jakarta

Setelah sebelumnya ditelaah mengenai variabel pasien positif Covid-19, pasien sembuh dari Covid-19, pasien isolasi, dan kasus meninggal di DKI Jakarta, selanjutnya akan di analisis hubungan antar variabel tersebut dengan menggunakan *Machine Learning Regression* sebagai berikut.

a. *Linear Regression* pada variabel pasien positif dengan pasien meninggal akibat Covid-19 di DKI Jakarta

Variabel pasien positif dijadikan sebagai variabel bebas dan variabel meninggal dijadikan sebagai variabel terikat. Dengan melakukan transformasi *value reshape*(-1,1) dan *test-size* dengan proporsi 0,3 yang berarti komposisi data latih sebesar 70% dan data uji sebesar 30%, diperoleh nilai konstanta dan nilai *intercept* berturut-turut

sebesar 0,0155637 dan 391,43649625 dengan skor akurasi prediksi mencapai 0.9947197907345313.

b. *Linear Regression* pada variabel pasien sembuh dengan pasien meninggal akibat Covid-19 di DKI Jakarta

Variabel pasien sembuh dijadikan sebagai variabel bebas dan variabel meninggal dijadikan sebagai variabel terikat. Dengan melakukan transformasi *value reshape*(-1,1) dan *test-size* dengan proporsi 0,3 yang berarti komposisi data latih sebesar 70% dan data uji sebesar 30%, diperoleh nilai konstanta dan nilai *intercept* berturut-turut sebesar 0,01627064 dan 494,39108989 dengan skor akurasi prediksi mencapai 0,9931244918092648.

c. *Linear Regression* pada variabel pasien isolasi dengan pasien meninggal akibat Covid-19 di DKI Jakarta

Variabel pasien isolasi dijadikan sebagai variabel bebas dan variabel meninggal dijadikan sebagai variabel terikat. Dengan melakukan transformasi *value reshape*(-1,1) dan *test-size* dengan proporsi 0,3 yang berarti komposisi data latih sebesar 70% dan data uji sebesar 30%, diperoleh nilai konstanta dan nilai *intercept* berturut-turut sebesar 0,16233583 dan 1925,40504925 dengan skor akurasi prediksi mencapai 0,12977008939233958.

Jika meninjau dari variabel bebas berganda, variabel pasien positif, variabel pasien sembuh, dan variabel *self-isolation* dijadikan variabel bebas serta variabel pasien meninggal dijadikan variabel terikat, maka diperoleh nilai konstanta sebesar 384,1434, nilai koefisien pasien sembuh sebesar -0,0257, nilai koefisien pasien positif sebesar 0,0407, dan nilai koefisien pasien isolasi sebesar -0,0335 untuk regresi berganda.

2.3.6 CONCLUSION AND RECOMMENDATION

1. Untuk uji hipotesis rata-rata total vaksin di daerah Jakarta Barat, Jakarta Pusat, Jakarta Selatan, Jakarta Timur, Jakarta Utara dan Kepulauan Seribu. Hasil dari uji hipotesis, rata-rata total vaksin di setiap daerah ada yang sama namun tidak diketahui daerah tepatnya.
2. Untuk vaksin ke-1 Jakarta Selatan menjadi wilayah dengan jumlah vaksin terbesar, dan Kepulauan Seribu menjadi wilayah dengan jumlah vaksin terkecil. Sedangkan untuk vaksin ke-2 Jakarta Timur menjadi wilayah dengan jumlah vaksin terbesar, dan Kepulauan Seribu menjadi wilayah dengan jumlah vaksin terkecil.

3. Pola vaksin ke-1 cenderung mirip dilihat dari grafik tren yang ada, hanya berbeda untuk wilayah Kepulauan Seribu
4. Pola vaksin ke-2 bisa dibilang unik karena memiliki pola berbeda masing-masing wilayah.
5. Hasil *clustering* yang terbentuk sebanyak 3 *cluster*. *Cluster* dengan label 0 berisi 4 wilayah, *cluster* dengan label 1 berisi 1 wilayah, dan *cluster* dengan label 2 berisi 1 wilayah. Rekomendasinya adalah perlunya pemerataan vaksin agar tidak terlalu jauh perbedaannya
6. Berdasarkan hasil EDA dan *deep dive analysis*, variabel yang paling berpengaruh terhadap kematian akibat Covid-19 di DKI Jakarta adalah variabel pasien positif dan pasien sembuh dari Covid-19 dengan tingkat korelasi lebih dari 0,9 dan termasuk ke dalam kategori sangat kuat. Sementara itu, variabel *self-isolation* berpengaruh rendah terhadap kematian yang artinya jika pasien isolasi secara mandiri dan efektif, maka akan mengurangi persentase kematian seseorang dari Covid-19.
7. Berdasarkan hasil EDA dan *deep dive analysis*, diperoleh model sebagai berikut.
    a. Model pasien positif dengan pasien meninggal akibat Covid-19 di DKI Jakarta yaitu $Y_{meninggal} = 0{,}0155637 + 391{,}43649625\, X_{positif}$
    b. Model pasien sembuh dengan pasien meninggal akibat Covid-19 di DKI Jakarta yaitu $Y_{meninggal} = 0{,}01627064 + 494{,}39108989\ X_{sembuh}$
    c. Model pasien isolasi dengan pasien meninggal akibat Covid-19 di DKI Jakarta yaitu $Y_{meninggal} = 0{,}16233583 + 1925{,}40504925\ X_{isolasi}$
    d. Model pasien positif, pasien sembuh, pasien isolasi dengan pasien meninggal akibat Covid-19 di DKI Jakarta yaitu

$$Y_{meninggal} = 384{,}1434 + 0{,}0407\, X_{positif} - 0{,}0257\, X_{sembuh} - 0{,}0335\, X_{isolasi}$$

ALASAN PENGGUNAAN TEKNIK EDA YANG DIPILIH

1. *Data cleansing* dan *data transformation*

Proses ini merupakan proses awal pada setiap pengolahan data. Pada proses ini data diamati dan dipilih kemudian di cek untuk data missing, data shape, dan lainnya. Proses ini pada pengolahan data vaksin tujuan utamanya adalah

menghindari bentuk data yang multi index dan multi kolom, karena khawatir akan sulit untuk mengolahnya.

2. **Mencari statistik deskriptif**

   Proses ini bertujuan untuk mengetahui statistik deskriptif seperti nilai rata-rata (*mean*), nilai tengah (*median*), nilai terkecil, nilai terbesar, dan lainnya. Statistik deskriptif juga dapat berguna pada pengolahan data seperti mencari nilai outlier.

3. **Melihat distribusi sebaran data**

   Proses ini bertujuan untuk melihat apakah data memiliki sebaran normal atau tidak. Kurva lonceng menandakan bahwa data memiliki sebaran normal, meskipun ini bukan cara satu-satunya namun cara ini cukup sederhana yakni hanya melihat grafik histogram nya.

4. **Melihat korelasi data**

   Proses ini bertujuan untuk melihat tingkat korelasi antar variabel. Dalam data vaksin variabel ialah wilayah serta cakupan vaksin. Proses ini membantu untuk mendapatkan informasi lanjutan untuk olah data selanjutnya.

5. **Melihat tren data**

   Proses ini bertujuan untuk mengetahui tren yang tercipta oleh data yang ada. Kita juga dapat membandingkan pola dari trend suatu variabel dengan variabel lainnya.

2.4 Kesimpulan

Pandemi covid-19 di Indonesia telah membawa Indonesia dalam kesedihan. Pemerintah berupaya menerapkan program vaksinasi dan telah memesan sebanyak 328,5 juta dosis vaksin. Atas dasar itu, kami memcoba menganalisis lebih lanjut terkait dengan vaksin dan pengaruh kematian dengan pasien positif dan sembuh. Analisis ini didapat dari dataset yang disediakan yg merupakan data yang dinamis, yakni data terus di*update* perhari atau persatuan waktu tertentu. Kami membatasi untuk hanya menganalisis pada periode hingga tanggal 30 Juni 2021. Dataset yang kami peroleh pertama-tama kami cleansing dan transformasi, lalu proses EDA untuk melihat insight pada data tersebut dan menjawab beberapa pertanyaan hingga menyelesaikan hipotesis yang ada, hingga akhirnya kami analisis lebih dalam pada *deep dive analysis*. Metode yg digunakan pada *deep dive analysis* yakni *clustering* untuk data vaksin dan regresi untuk data Covid-19. Hasil *clustering* menemukan cluster yang terbentuk yakni 3 *cluster* sedangkan untuk regresi terbentuk beberapa model. Rekomendasi utama yakni dapat menggunakan data terbaru yang lebih relevan dan memperdalam pengetahuan baik dalam data science maupun pengetahuan sosial tentang kejadian lapangan.

DAFTAR PUSTAKA

Assegaf, Alwi dkk. (2019). Analisis Kesehatan Bank Menggunakan Local Mean K-Nearest Neighbor dan Multi Local Means K-Harmonic Nearest Neighbor. Jurnal gaussian. Vol 8. No 3

Dabbura, Imad. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Data Science, Towards. (2020). Machine Learning with Python: Regression (complete tutorial).

https://towardsdatascience.com/machine-learning-with-python-regression-complete-tutorial-47268e546cea

Kompas. (2020). Positivity Rate Covid-19 di Bawah Batas Ideal WHO, Anies Ingatkan Warga untuk Waspada.

https://megapolitan.kompas.com/read/2020/07/16/20444351/positivity-rate-covid-19-di-bawah-batas-ideal-who-anies-ingatkan-warga

Kompas. (2021). Optimalisasi Vaksinasi Covid-19, Kemenkes Instruksikan Vaksinasi Tak Lagi Pandang Domisili.

https://nasional.kompas.com/read/2021/06/25/11325261/optimalisasi-vaksinasi-covid-19-kemenkes-instruksikan-vaksinasi-tak-lagi?page=all

Presiden RI. (2021). Presiden Jokowi Instruksikan Jajarannya Bersiap Jalankan Program Vaksinasi Covid-19.

https://www.presidenri.go.id/siaran-pers/presiden-jokowi-instruksikan-jajarannya-bersiap-jalankan-program-vaksinasi-covid-19/

Suara. (2021). Tanpa Surat Domisili, Sekarang Masyarakat Bisa Dapat Vaksin Covid-19 di Bandara Soetta.

https://www.suara.com/health/2021/07/12/160031/tanpa-surat-domisili-sekarang-masyarakat-bisa-dapat-vaksin-covid-19-di-bandara-soeta?page=all

LAMPIRAN

Lampiran 1: Menghitung nilai mean, median dan modus positif Covid-19 harian

```
mean = ambil_data['Positif Harian'].mean()
median = ambil_data['Positif Harian'].median()
modus = ambil_data['Positif Harian'].mode()

print('Nilai Mean: ', mean)
print('Nilai Median: ', median)
print('Nilai Modus: ', modus)
```

```
Nilai Mean:  1115.9507186858316
Nilai Median:  845.0
Nilai Modus:  0    0
dtype: int64
```

```
#jika diasumsikan nilai 0 tidak dianggap
mean2 = data_lain['Positif Harian'].mean()
median2 = data_lain['Positif Harian'].median()
modus2 = data_lain['Positif Harian'].mode()

print('Nilai Mean: ', mean2)
print('Nilai Median: ', median2)
print('Nilai Modus: ', modus2)
```

```
Nilai Mean:  1132.225
Nilai Median:  864.5
Nilai Modus:  0    127
dtype: int64
```

Lampiran 2: Nilai minimal dan maksimal positif harian Jakarta

```
maxmin = df.agg({'Positif Harian' : ['mean','median','max', 'min']})
print(maxmin)
```

```
        Positif Harian
mean       1115.950719
median      845.000000
max        9394.000000
min           0.000000
```

Lampiran 3: Nilai outlier pemakaman

```
print("Jumlah outlier Pemakaman_COVID19_Harian sebanyak {} \n"
      .format(pemakaman[pemakaman["Pemakaman_COVID19_Harian"]>pem5b].shape[0]))
print("Jumlah outlier Pemakaman_Umum_Harian sebanyak {} \n"
      .format(pemakaman[pemakaman["Pemakaman_Umum_Harian"]>pemu5b].shape[0]))
```

```
Jumlah outlier Pemakaman_COVID19_Harian sebanyak 22

Jumlah outlier Pemakaman_Umum_Harian sebanyak 18
```

Lampiran 4: Nilai outlier hasil lab

```
print("Jumlah outlier Positivity Rate Kasus Baru Harian sebanyak {} \n"
      .format(hl[hl["Positivity Rate Kasus Baru Harian"]>pos5b].shape[0]))
```

```
Jumlah outlier Positivity Rate Kasus Baru Harian sebanyak 28
```

Lampiran 5: Nilai outlier data suspek

```
print("Jumlah outlier Isolasi Harian sebanyak {} \n"
      .format(isolated0[isolated0["Isolated Patients"]>iso5b].shape[0]))
```

```
Jumlah outlier Isolasi Harian sebanyak 24
```

Lampiran 6: Nilai outlier data Jakarta

```
print("Jumlah outlier Positif Harian sebanyak {} \n"
      .format(dj2[dj2["Positif Harian"]>poz5b].shape[0]))
print("Jumlah outlier Sembuh Harian sebanyak {} \n"
      .format(dj2[dj2["Sembuh Harian"]>sem5b].shape[0]))
print("Jumlah outlier Tanpa Gejala sebanyak {} \n"
      .format(dj2[dj2["Tanpa Gejala"]>tg5b].shape[0]))
print("Jumlah outlier Bergejala sebanyak {} \n"
      .format(dj2[dj2["Bergejala"]>g5b].shape[0]))
```

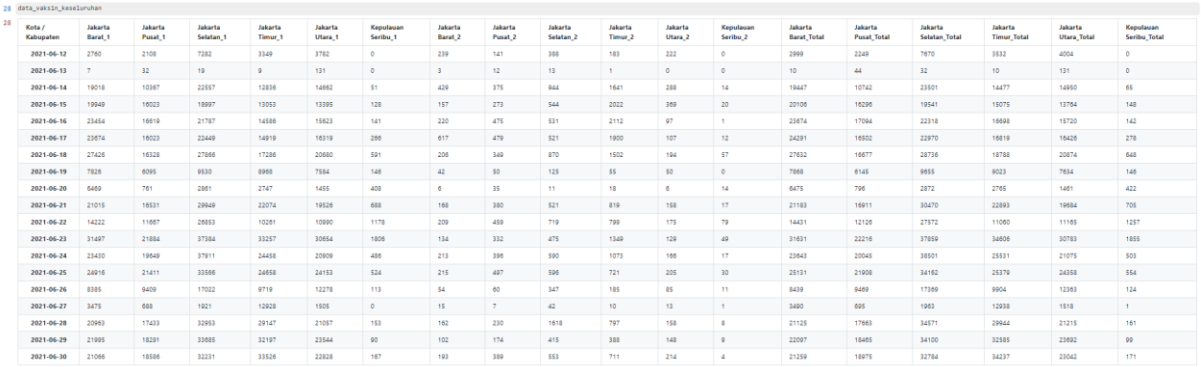
```
Jumlah outlier Positif Harian sebanyak 36

Jumlah outlier Sembuh Harian sebanyak 25

Jumlah outlier Tanpa Gejala sebanyak 31

Jumlah outlier Bergejala sebanyak 7
```

Lampiran 7: Nilai korelasi

```
1  nilai_korelasi = gabung['Self Isolation'].corr(gabung['Sembuh'], method = 'pearson')
2  nilai_korelasi
```

```
0.9872460736547924
```

Lampiran 8: Data vaksin setelah *cleansing* dan *transformation*



Kota / Kabupaten	Jakarta Barat_1	Jakarta Pusat_1	Jakarta Selatan_1	Jakarta Timur_1	Jakarta Utara_1	Kepulauan Seribu_1	Jakarta Barat_2	Jakarta Pusat_2	Jakarta Selatan_2	Jakarta Timur_2	Jakarta Utara_2	Kepulauan Seribu_2	Jakarta Barat_Total	Jakarta Pusat_Total	Jakarta Selatan_Total	Jakarta Timur_Total	Jakarta Utara_Total	Kepulauan Seribu_Total
2021-06-12	2760	2108	7282	3349	3782	0	239	141	388	183	222	0	2999	2249	7670	3532	4004	0
2021-06-13	7	32	19	9	131	0	3	12	13	1	0	0	10	44	32	10	131	0
2021-06-14	19018	10367	22557	12836	14662	51	429	375	944	1641	288	14	19447	10742	23501	14477	14950	65
2021-06-15	19949	16023	18997	13053	13395	128	157	273	544	2022	369	20	20106	16296	19541	15075	13764	148
2021-06-16	23454	16619	21787	14586	15823	141	220	475	531	2112	97	1	23674	17094	22318	16698	15720	142
2021-06-17	23674	16023	22449	14919	16319	266	617	479	521	1900	107	12	24291	16502	22970	16819	16426	278
2021-06-18	27426	16328	27866	17286	20680	591	206	349	870	1502	194	57	27632	16677	28736	18788	20874	648
2021-06-19	7826	6095	9530	8966	7564	148	42	80	125	55	80	0	7868	6145	9655	9023	7634	148
2021-06-20	6469	761	2861	2747	1455	408	6	35	11	18	6	14	6475	796	2872	2765	1461	422
2021-06-21	21015	16531	29949	22074	19526	686	168	380	521	619	158	17	21183	16911	30470	22693	19684	703
2021-06-22	14222	11667	26853	10261	10990	1178	209	459	719	799	175	79	14431	12126	27572	11060	11165	1257
2021-06-23	31497	21884	37384	33257	30654	1806	134	332	475	1349	129	49	31631	22216	37859	34606	30783	1855
2021-06-24	23430	19649	37911	24458	20909	486	213	396	590	1073	166	17	23643	20045	38501	25531	21075	503
2021-06-25	24916	21411	33566	24658	24153	524	215	497	596	721	205	30	25131	21908	34162	25379	24358	554
2021-06-26	8385	9409	17022	9719	12278	113	54	60	347	185	85	11	8439	9469	17369	9904	12363	124
2021-06-27	3475	688	1921	12928	1505	0	15	7	42	10	13	1	3490	695	1963	12938	1518	1
2021-06-28	20963	17433	32953	29147	21057	153	162	230	1618	797	158	8	21125	17663	34571	29944	21215	161
2021-06-29	21995	18291	33685	32197	23544	90	102	174	415	388	148	9	22097	18465	34100	32585	23692	99
2021-06-30	21066	18586	32231	33526	22828	167	193	389	553	711	214	4	21259	18975	32784	34237	23042	171

Lampiran 9: Uji hipotesis data vaksin

```
from scipy.stats import kruskal
```

```
stat,pvalue = kruskal(a['Jakarta Barat'],a['Jakarta Pusat'],a['Jakarta Selatan'],a['Jakarta Timur'
```

```
print('Nilai Stat: ',stat)
print('Nilai P-Value: ',pvalue)

#conclusion
alpha = 0.05
if pvalue < alpha:
    print('Tolak Ho')
else:
    print('Terima Ha')
```

```
Nilai Stat:  5.0
Nilai P-Value:  0.4158801869955079
Terima Ha
```

Lampiran 10: Uji hipotesis hasil regresi

OLS Regression Results

Dep. Variable:	Meninggal (Jakarta)	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	3.067e+04
Date:	Mon, 19 Jul 2021	Prob (F-statistic):	0.00
Time:	13:54:01	Log-Likelihood:	-3237.9
No. Observations:	487	AIC:	6484.
Df Residuals:	483	BIC:	6501.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	384.1434	14.422	26.635	0.000	355.805	412.482
Sembuh (Jakarta)	-0.0257	0.007	-3.871	0.000	-0.039	-0.013
Positif (Jakarta)	0.0407	0.006	6.276	0.000	0.028	0.053
Self-Isolation (Jakarta)	-0.0335	0.009	-3.914	0.000	-0.050	-0.017

Omnibus:	45.399	Durbin-Watson:	0.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56.038
Skew:	-0.822	Prob(JB):	6.79e-13
Kurtosis:	2.753	Cond. No.	5.47e+05

Lampiran 11 : Statistik deskriptif data vaksin

```
data_vaksin_keseluruhan.describe()
```

Kota / Kabupaten	Jakarta Barat_1	Jakarta Pusat_1	Jakarta Selatan_1	Jakarta Timur_1	Jakarta Utara_1	Kepulauan Seribu_1	Jakarta Barat_2	Jakarta Pusat_2	Jakarta Selatan_2	Jakarta Timur_2	Jakarta Utara_2	Kepulauan Seribu_2	Jakarta Barat_Total	Jakarta Pusat_Total	Jakarta Selatan_Total	Jakarta Timur_Total	Jakarta Utara_Total	Kepulauan Seribu_Total
count	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000
mean	16923.526316	12626.578947	21938.052632	16840.947368	14793.421053	365.052632	178.105263	269.105263	517.000000	857.157895	146.526316	18.052632	17101.631579	12895.684211	22455.052632	17698.105263	14939.947368	383.105263
std	9269.123286	7387.920030	12393.913921	10425.128107	8763.750347	460.692772	148.123408	174.341904	374.226313	719.212939	95.265342	21.598543	9340.435638	7523.650450	12638.472541	10640.127792	8806.172037	477.994758
min	7.000000	32.000000	19.000000	9.000000	131.000000	0.000000	3.000000	7.000000	11.000000	1.000000	0.000000	0.000000	10.000000	44.000000	32.000000	10.000000	131.000000	0.000000
25%	8105.500000	7752.000000	13276.000000	9990.000000	9287.000000	101.500000	78.000000	100.500000	367.500000	184.000000	91.000000	2.500000	8153.500000	7807.000000	13512.000000	10482.000000	9399.500000	111.500000
50%	20963.000000	16023.000000	22657.000000	14586.000000	15623.000000	153.000000	168.000000	332.000000	821.000000	797.000000	158.000000	12.000000	21125.000000	16502.000000	23501.000000	16698.000000	15720.000000	161.000000
75%	23442.000000	17862.000000	32592.000000	24558.000000	20983.000000	505.000000	214.000000	392.500000	593.000000	1425.500000	199.500000	18.500000	23658.500000	18064.000000	33442.000000	25455.000000	21145.000000	528.500000
max	31497.000000	21884.000000	37911.000000	33526.000000	30654.000000	1806.000000	617.000000	497.000000	1618.000000	2112.000000	369.000000	79.000000	31631.000000	22216.000000	38501.000000	34606.000000	30783.000000	1855.000000

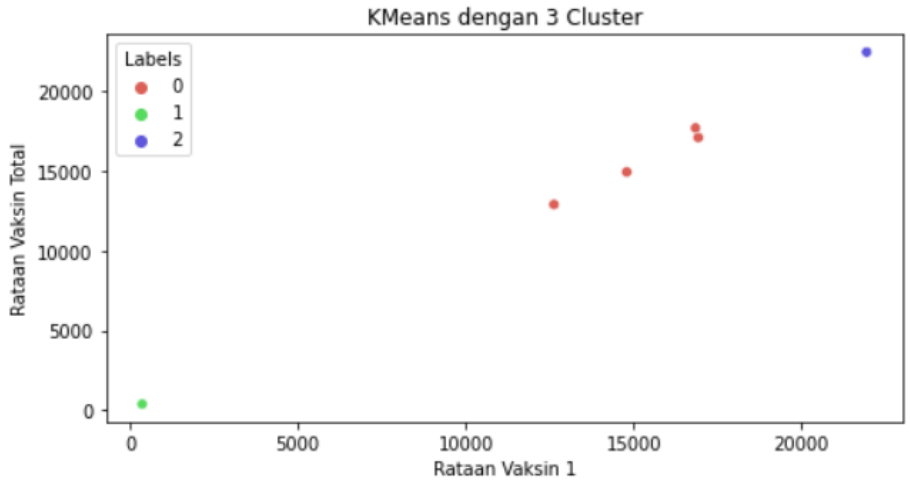
Lampiran 12: korelasi antar wilayah data vaksin

Kota / Kabupaten	Jakarta Barat_1	Jakarta Pusat_1	Jakarta Selatan_1	Jakarta Timur_1	Jakarta Utara_1	Kepulauan Seribu_1	Jakarta Barat_2	Jakarta Pusat_2	Jakarta Selatan_2	Jakarta Timur_2	Jakarta Utara_2	Kepulauan Seribu_2	Jakarta Barat_Total	Jakarta Pusat_Total	Jakarta Selatan_Total	Jakarta Timur_Total	Jakarta Utara_Total	Kepulauan Seribu_Total
Jakarta Barat_1	1.000000	0.948510	0.893277	0.777597	0.930525	0.493439	0.475301	0.794481	0.574894	0.738893	0.478017	0.426636	0.999903	0.949834	0.893014	0.811829	0.931213	0.494855
Jakarta Pusat_1	0.948510	1.000000	0.951837	0.844610	0.961100	0.430042	0.387714	0.772714	0.573848	0.621405	0.504590	0.354756	0.947417	0.999892	0.950410	0.869547	0.961929	0.430506
Jakarta Selatan_1	0.893277	0.951837	1.000000	0.868803	0.956380	0.492689	0.358642	0.738732	0.644855	0.481692	0.491779	0.433597	0.892144	0.951809	0.999744	0.883807	0.957093	0.494447
Jakarta Timur_1	0.777597	0.844610	0.868803	1.000000	0.905473	0.351486	0.134922	0.473608	0.478118	0.267526	0.303674	0.155336	0.773799	0.840370	0.866149	0.997877	0.904396	0.345783
Jakarta Utara_1	0.930525	0.961100	0.956380	0.905473	1.000000	0.487404	0.323765	0.670090	0.562231	0.500886	0.440943	0.352226	0.928555	0.959314	0.954521	0.921033	0.999953	0.485677
Kepulauan Seribu_1	0.493439	0.430042	0.492689	0.351486	0.487404	1.000000	0.019842	0.406020	0.116335	0.207959	0.039134	0.792673	0.489986	0.431704	0.486600	0.358441	0.485479	0.999621
Jakarta Barat_2	0.475301	0.387714	0.358642	0.134922	0.323765	0.019842	1.000000	0.709840	0.443230	0.677049	0.451869	0.123378	0.487530	0.397178	0.364826	0.177960	0.327094	0.024699
Jakarta Pusat_2	0.794481	0.772714	0.738732	0.473608	0.670090	0.406020	0.709840	1.000000	0.529759	0.752800	0.534983	0.461391	0.799672	0.781967	0.740123	0.514923	0.672650	0.412172
Jakarta Selatan_2	0.574894	0.573848	0.644855	0.478118	0.562231	0.116335	0.443230	0.529759	1.000000	0.482699	0.555934	0.294764	0.577533	0.575787	0.661987	0.501084	0.565536	0.125444
Jakarta Timur_2	0.738893	0.621405	0.481692	0.267526	0.500886	0.207959	0.677049	0.752800	0.482699	1.000000	0.564316	0.278815	0.743988	0.627656	0.486664	0.329714	0.504578	0.213030
Jakarta Utara_2	0.478017	0.504590	0.491779	0.303674	0.440943	0.039134	0.451869	0.534983	0.555934	0.564316	1.000000	0.266775	0.481534	0.507897	0.498724	0.335682	0.449637	0.049772
Kepulauan Seribu_2	0.426636	0.354756	0.433597	0.155336	0.352226	0.792673	0.123378	0.461391	0.294764	0.278815	0.266775	1.000000	0.425335	0.359057	0.433935	0.171043	0.353415	0.809167
Jakarta Barat_Total	0.999903	0.947417	0.892144	0.773799	0.928555	0.489986	0.487530	0.799672	0.577533	0.743988	0.481534	0.425335	1.000000	0.948880	0.891982	0.808453	0.929291	0.491469
Jakarta Pusat_Total	0.949834	0.999892	0.951809	0.840370	0.959314	0.431704	0.397178	0.781967	0.575787	0.627656	0.507897	0.359057	0.948880	1.000000	0.950440	0.865815	0.960187	0.432302
Jakarta Selatan_Total	0.893014	0.950410	0.999744	0.866149	0.954521	0.486600	0.364826	0.740123	0.661987	0.486664	0.498724	0.433935	0.891982	0.950440	1.000000	0.881543	0.955318	0.488594
Jakarta Timur_Total	0.811829	0.869547	0.883807	0.997877	0.921033	0.358441	0.177960	0.514923	0.501084	0.329714	0.335682	0.171043	0.808453	0.865815	0.881543	1.000000	0.920228	0.353195
Jakarta Utara_Total	0.931213	0.961929	0.957093	0.904396	0.999953	0.485479	0.327094	0.672650	0.565536	0.504578	0.449637	0.353415	0.929291	0.960187	0.955318	0.920228	1.000000	0.483876
Kepulauan Seribu_Total	0.494855	0.430506	0.494447	0.345783	0.485677	0.999621	0.024699	0.412172	0.125444	0.213030	0.049772	0.809167	0.491469	0.432302	0.488594	0.353195	0.483876	1.000000

Lampiran 13: Scatter plot clustering

```
# membuat plot KMeans dengan 3 klaster
plt.figure(figsize=(8,4))
sns.scatterplot(xx['Rataan Vaksin 1'], xx['Rataan Vaksin Total'], hue=x['Labels'],
                palette=sns.color_palette('hls', 3))
plt.title('KMeans dengan 3 Cluster')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass
  FutureWarning
```



```
# membuat plot KMeans dengan 3 klaster
plt.figure(figsize=(8,4))
sns.scatterplot(xx['Rataan Vaksin 2'], xx['Rataan Vaksin Total'], hue=x['Labels'],
                palette=sns.color_palette('hls', 3))
plt.title('KMeans dengan 3 Cluster')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
  FutureWarning
```