PENUGASAN DATA SCIENCE FUNDAMENTAL (NON-DATASET)

SOAL

1. Jelaskan secara teori statistik mengenai *outlier* (pencilan), implikasinya dalam analisis data, serta bagaimana melakukan manajemen data terhadap kasus *outlier*.
2. Jelaskan konsep dan prinsip korelasi, lalu sebisa mungkin kaitkan dengan dasar-dasar statistik serta implikasinya terhadap konsep/teori statistik lain.
3. Sebutkan teori dasar *machine learning* yang kalian ketahui, lalu jelaskan dalam bahasa sederhana mengenai teori tersebut dan implikasinya.
4. Menggunakan bahasa kalian sendiri, jelaskan kaitan antara *artificial intelligence*, *machine learning*, dan *deep learning*.
5. Apakah yang kalian ketahui mengenai interpretasi data? Bagaimana signifikansi dan tantangannya? Bagaimana kaitan interpretasi data dengan *data story telling* dan *decision making*?

JAWABAN

1. Dalam ilmu statistika, *outlier* merupakan sebuah hasil observasi sampel dari sebuah populasi yang memiliki jarak yang abnormal dari nilai data lainnya. Status abnormal ini sendiri dapat didefinisikan sendiri oleh *data analyst* untuk menentukan tindakan yang akan dilakukan pada pencilan (*outlier*). Data yang termasuk dalam pencilan biasanya memiliki nilai data yang buruk karena dapat memperburuk kualitas dataset saat diinterpretasikan. Namun, dalam beberapa kasus, sebuah *outlier* juga bisa memuat nilai data yang penting meskipun jarak nilainya sangat jauh dari mayoritas data yang ada. Untuk mengatasi pencilan, seorang *data analyst* dapat mendeteksinya terlebih dahulu melalui beberapa metode, misalnya *boxplot*, sekaligus memahami apa yang menyebabkan data tersebut abnormal. Jika data tersebut memang dirasa memang tidak penting, maka dapat dilakukan eliminasi agar kualitas data di dalam dataset meningkat saat diinterpretasikan. Apabila data tersebut memang penting, misalkan pada kasus data kesehatan, maka data dapat disimpan dan diinterpretasikan.

2. Korelasi merupakan sebuah ukuran statistik yang menentukan apakah sebuah variabel berkaitan secara linear dengan variabel lain secara konstan. Penentuan ini biasanya dilakukan menggunakan scatterplot dengan range korelasi variabel x dan y berada di

antara 1 dan -1. Apabila variabel memiliki sebuah korelasi, maka akan terjadi perubahan pada arah tertentu (+ berarti nilai bertambah; - nilai berkurang) saat variabel yang berkorelasi mengalami perubahan nilai. Namun, apabila korelasi antar variabel bernilai 0, maka kedua variabel tersebut tidak ada korelasinya. Korelasi ini sangat berpengaruh saat data akan diinterpretasikan karena variabel yang tidak berkorelasi, maka data-datanya tidak perlu diinterpretasikan.

3. *Machine Learning* adalah salah satu cabang ilmu komputer yang mengimplementasikan sains data di dalamnya. *Machine Learning* atau pembelajaran mesin melatih sebuah mesin untuk menghasilkan suatu *output* tanpa diprogram secara eksplisit, inilah hal utama yang membedakan *machine learning* dengan program tradisional. *Machine learning* hanya membutuhkan data untuk dilatih dan menghasilkan model, dan akhirnya dapat secara otomatis menghasilkan keluaran, sedangkan program tradisional terlebih dahulu memerlukan aturan berupa program, barulah keluaran dapat dihasilkan. *Machine learning* mengimplementasikan sains data, statistik, dan algoritma untuk melakukan klasifikasi ataupun prediksi, serta menguak kunci informasi di dalam data mining.

4. *Artificial Intelligence* atau kecerdasan buatan merupakan sebuah bidang ilmu yang bertujuan meniru kecerdasan manusia untuk membantu mengerjakan pekerjaan manusia, sedangkan *machine learning* merupakan bagian kecil dari *artificial intelligence*, yang secara spesifik hanya diperuntukkan untuk mesin melatih dirinya sendiri. Lalu *Deep Learning* juga merupakan salah satu bagian dari *artificial intelligence*, dimana pada bidang merupakan pengembangan dari artificial intelligence dan machine learning yang terdiri dari banyak *layer* yang tiap *layer*-nya berisikan neuron-neuron untuk memberikan hasil seperti deteksi objek, pengenalan suara, terjemahan bahasa dan lain – lain.

5. Interpretasi data adalah proses melihat serta menganalisis data dengan berbagai proses statistik, dan akan membantu menemukan maksud maupun intisari dari data. Tantangannya dari interpretasi data adalah bagaimana data harus diolah, karena perbedaan pengolahan data dapat mempengaruhi interpretasi data itu sendiri. Interpretasi data sangat erat kaitannya dengan *data story telling*, karena interpretasi data digunakan untuk menyampaikan maksud maupun informasi penting mengenai data. *Data storytelling* merupakan sebuah metode komunikasi untuk menyampaikan

hasil analisis data berupa informasi yang penting untuk diketahui dari suatu data. Interpretasi data maupun *data story telling* juga erat kaitannya dengan *decision making*, karena dengan ditemukannya informasi penting dari suatu data, keputusan untuk melakukan suatu hal terhadap suatu data jadi lebih akurat dan baik.

SUMBER

Correlation. JMP. (n.d.). https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-correlation.html.

IBM Cloud Learn Hub. IBM. (n.d.). https://www.ibm.com/cloud/learn.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2012.

Taylor, C. (2018, February 20). *What Role Does Correlation Play in Statistics?* ThoughtCo. https://www.thoughtco.com/what-is-correlation-3126364.

Taylor, C. (2019, May 22). *How Do We Determine Outliers in Statistics?* ThoughtCo. https://www.thoughtco.com/what-is-an-outlier-3126227.

LAPORAN HASIL ANALISIS DATASET KASUS PENYAKIT COVID-19 DI DAERAH DKI JAKARTA SEBAGAI REKOMENDASI PENENTUAN KEBIJAKAN PEMERINTAH DAERAH

Laporan ini Ditujukan Sebagai Syarat Mengikuti *Data Science Academy* dalam Acara Compfest 15

Disusun oleh:

YannLecun

- **M. Nugraha Delta Revanza** **(Teknik Informatika Universitas Brawijaya)**
- **Muhammad Furqony Sabililhaq** **(Teknik Informatika Universitas Brawijaya)**
- **Erland Hilman Fuadi** **(Teknik Informatika Universitas Brawijaya)**

2021

LATAR BELAKANG

Penyakit COVID-19 (*Coronavirus Disease* 2019) yang ditemukan pertama kali di Hubei, Tiongkok, telah tersebar di berbagai benua secara global (Velavan, T. P., 2020). Virus yang menyebabkan penyakit ini tidak hanya menyebar melalui kontak fisik secara langsung, tetapi juga dapat menular melalui udara (Zhang, Renyi, 2020). Dengan media persebaran virus yang luas, WHO (*World Health Organization*) mendeklarasikan kondisi darurat secara global setelah kasus yang diakibatkannya terus meningkat secara internasional pada 30 Juni 2020, termasuk di Indonesia. Menurut situs covid19.go.id yang diakses pada 3 Juli 2021 pukul 22:39, terdapat 2.256.851 kasus terkonfirmasi selama setahun terakhir dengan Kota DKI Jakarta sebagai kota dengan sebaran kasus terbanyak (23,9%).

Sebagai ibukota dari Indonesia, DKI Jakarta menjadi daerah dengan sebaran kasus COVID-19 terbanyak di Indonesia. Melonjaknya kasus di kota metropolitan ini disebabkan dari berbagai macam faktor, seperti cuaca yang mendukung perkembangan virus, tingginya mobilitas penduduk di dalamnya, dan padatnya penduduk Jakarta (Tosepu, R., 2020). Kasus penyakit COVID-19 sendiri menyangkut pasien yang positif, pasien yang meninggal, pasien yang sudah sembuh, pasien positif aktif dan isolasi mandiri. Salah satu upaya dalam penanganan kasus COVID-19 adalah transparansi data dari hasil penanganan suatu negara atau daerah (Djalante, R, 2020). Data yang dihimpun ini nantinya dapat diolah sesuai dengan kasus tertentu, sehingga dapat mendukung suatu instansi terkait dalam pengambilan keputusan atau pembuatan kebijakan untuk mengendalikan kasus COVID-19 di daerahnya.

Dalam laporan ini, pembahasan terhadap data-data terkait hasil penanganan pandemi COVID-19 dibatasi pada daerah DKI Jakarta melalui laman web corona.jakarta.go.id hingga tanggal 30 Juni 2021. Hasil analisis data tersebut diharapkan dapat membantu berbagai pihak, baik dari instansi terkait maupun masyarakat umum, dalam memahami kondisi yang diakibatkan pandemi COVID-19 di daerah Kota Jakarta sebagai pusat klaster kasus COVID-19 di Indonesia.

JAWABAN SOAL

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

   Jawaban:

Agregasi	**Positif COVID-19 Harian Jakarta**
Mean (rata-rata)	1115.95
Median (nilai tengah)	854
Modus	0

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

   Jawaban:

Min/Max	**Positif COVID-19 Harian Jakarta**
Minimal	0
Maksimal	9394

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

   Jawaban :

   Untuk mencari nilai-nilai outlier, kami menggunakan aturan $1.5\,x\,IQR$. IQR atau *Interquartile Range* itu sendiri merupakan hasil perhitungan $Q3 - Q1$ atau kuartil 3 dari data dikurang dengan kuartil 1. Aturan ini merupakan aturan yang umum digunakan untuk mencari outlier pada data. Karena IQR sendiri digunakan untuk mencari persebaran data terhadap median. Sehingga data-data yang memiliki nilai lebih dari kuartil 3 ditambah 1.5 x IQR atau kurang dari kuartil 1 dikurang 1.5 x IQR akan dianggap outlier karena persebaran datanya yang sudah cukup jauh dari median.

$$Outlier = Q1 - 1.5\,x\,IQR > data \vee Q3 + 1.5\,x\,IQR < data$$

Berdasarkan perhitungan diatas, didapatkan bahwa outlier dari positif COVID-19 harian Jakarta adalah :

$$Outlier = -1553.75 > data \vee 3116.25 < data$$

Sehingga, data-data yang termasuk data outlier adalah data positif COVID-19 harian jakarta yang bernilai [3144, 3151, 3165, 3221, 3285, 3309, 3340, 3362, 3395, 3437, 3448, 3474, 3476, 3491, 3512, 3536, 3567, 3614, 3632, 3786, 3792, 3810, 4144, 4213, 4693, 4737, 4895, 5014, 5582, 6934, 7379, 7505, 7680, 8348, 9271, 9394]

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.

   Jawaban :

   Disini kami akan mengamati fitur ‘Positif Harian’ dan fitur ‘Sembuh Harian’. Metode yang akan kami gunakan untuk menghitung korelasi antar dua fitur adalah *Pearson Correlation*. Metode ini akan menghitung ketergantungan linear antara 2 fitur berdasarkan persebaran datanya. Oleh karena itu, metode ini juga disebut sebagai *Parametric Correlation*.

$$r = \frac{\Sigma(x - m_x)(y - m_y)}{\sqrt{\Sigma(x - m_x^{\,2})\Sigma(y - m_y^{\,2})}}$$

Berdasarkan rumus diatas, didapat bahwa koefisien korelasi antara fitur ‘Positif Harian’ dan fitur ‘Sembuh Harian’ adalah :

$$r = 0.78$$

Koefisien yang didapat adalah 0.78. Artinya, fitur ‘Positif Harian’ dan fitur ‘Sembuh Harian’ memiliki korelasi positif yang kuat. Bila nilai pada fitur ‘Positif Harian’

meningkat, maka nilai fitur ‘Sembuh Harian’ juga cenderung meningkat. Hal ini dapat dibenarkan karena apabila orang yang berstatus positif COVID-19 meningkat, maka jumlah orang yang dapat sembuh dari virus COVID-19 juga akan meningkat. Sehingga wajar apabila kedua fitur ini saling berkorelasi.

5. Dari Dataset Yang Disediakan,buatlah analisis dengan runtunan berikut:
    A. Problem Statement,
    B. Hypothesis,
    C. Exploratory Data Analysis,
    D. Initial Findings,
    E. Deep Dive Analysis, dan
    F. Conclusion and Recommendation

Jawaban:

A. Masyarakat yang terpapar virus COVID-19 semakin meningkat di wilayah DKI Jakarta. Sehingga perlu dilakukan kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM), namun harus tetap mempertimbangkan aspek ekonomi daerah. Oleh karena itu, kami akan mencari informasi terkait tren COVID-19 pada dataset COVID-19 DKI Jakarta yang dapat membantu pemerintah daerah dalam pembuatan regulasi.
B. Tren akan meningkat ketika akhir pekan dan akan menurun ketika hari kerja. Pemerintah dapat membuat regulasi untuk memperketat PPKM pada akhir pekan
C. Untuk melakukan Exploratory Data Analysis (EDA). Kita terlebih dahulu mempersiapkan data. Persiapan data disini yang dimaksud adalah pemilihan fitur, ekstraksi fitur, dan deteksi outlier. Kemudian, barulah kami melakukan EDA pada dataset.

C.1. Pemilihan Fitur

Pada dataset, terdapat 18 fitur yang tersedia. Fitur-fitur tersebut beserta penjelasannya adalah :

a. Tanggal = Tanggal data
b. Jam = Jam data
c. Tanggal Jam = Gabungan antara fitur tanggal dengan fitur jam

d. Total Pasien = Jumlah pasien baik yang masih positif, sembuh, maupun meninggal dalam bentuk akumulasi
e. Sembuh = Jumlah pasien yang sembuh dalam bentuk akumulasi
f. Meninggal = Jumlah pasien yang meninggal dalam bentuk akumulasi
g. Self Isolation = Jumlah pasien yang isolasi mandiri pada hari tersebut
h. Masih Perawatan = Jumlah pasien yang berada dalam perawatan pada hari tersebut
i. Belum Diketahui (masih verifikasi) = -
j. Menunggu Hasil = Jumlah pasien yang telah uji infeksi namun sedang menunggu hasil
k. Tenaga Kesehatan Terinfeksi = Jumlah tenaga kesehatan yang terinfeksi COVID-19
l. Positif Harian = Jumlah pasien baru terinfeksi COVID-19 pada hari tersebut
m. Positif Aktif = Jumlah pasien terinfeksi COVID-19 dalam bentuk akumulasi
n. Sembuh Harian = Jumlah pasien baru sembuh pada hari tersebut
o. Tanpa Gejala = Jumlah pasien terinfeksi COVID-19 yang tidak memiliki gejala pada hari tersebut
p. Bergejala = Jumlah pasien terinfeksi COVID-19 yang memiliki gejala pada hari tersebut
q. Belum Ada Data = Jumlah pasien terinfeksi COVID-19 yang tidak diketahui bergejala atau tidak bergejala pada hari tersebut

Fitur ‘Jam’ berisikan nilai antara 8 atau 18. Fitur ini bukan berisikan kapan pasien tertentu telah terinfeksi COVID-19, melainkan berisikan kapan data tersebut diinputkan ke dataset. Oleh karena itu fitur ‘Jam’ tidak akan kami gunakan. Begitu juga fitur ‘Tanggal Jam’ yang hanya merupakan gabungan dari fitur ‘Tanggal’ dan fitur ‘Jam’.

Selanjutnya kami akan mengambil fitur-fitur yang berupa harian atau fitur akumulasi yang dapat diubah kedalam fitur harian dengan cara mengurangi data ke-n dikurangi data ke-(n-1). Fitur ‘Total Pasien’ merupakan fitur akumulasi yang apabila diubah menjadi fitur harian maka akan sama dengan fitur ‘Positif

Harian’. Oleh karena itu fitur ini tidak akan kami pakai. Fitur ‘Sembuh’ juga sudah memiliki fitur harian yang bernama fitur ‘Sembuh Harian’ sehingga tidak akan kami pakai juga. Fitur ‘Positif Aktif’ juga telah memiliki fitur harian yaitu fitur ‘Positif Harian’ sehingga tidak akan kami pakai.

Fitur ‘Belum Diketahui (Masih Verifikasi)’ memiliki data null yang jumlahnya mencapai lebih dari 95% data sehingga tidak akan kami pakai. Begitu juga dengan fitur ‘Tenaga Kesehatan Terinfeksi’ dan fitur ‘Menunggu Hasil’ yang juga memiliki data null yang jumlahnya mencapai lebih dari 95% sehingga kedua fitur tersebut tidak akan kami pakai. Dan fitur ‘Bergejala’, ‘Tanpa Gejala’, dan ‘Belum Ada Data’ merupakan fitur bagian dari fitur ‘Positif Harian’ dan tidak ada informasi yang bisa kita dapatkan pada fitur tersebut sehingga fitur tersebut tidak akan kami pakai. Sehingga, fitur yang akan kami pakai adalah fitur ‘Tanggal’ yang menjadi index dari dataset, ‘Meninggal’, ‘Self Isolation’, ‘Masih Perawatan’, ‘Positif Harian’, dan ‘Sembuh Harian’.

C.2. Ekstraksi Fitur

Disini kami menambah fitur serta mengurangi berbagai fitur akumulasi. Untuk fitur-fitur akumulasi yang ada seperti ‘Meninggal’, ‘Self Isolation’, dan ‘Masih Perawatan’ kami mengubahnya menjadi fitur harian. Caranya yaitu tiap data akan diberlakukan rumus sebagai berikut :

$$data(n) = data(n) - data(n-1)$$

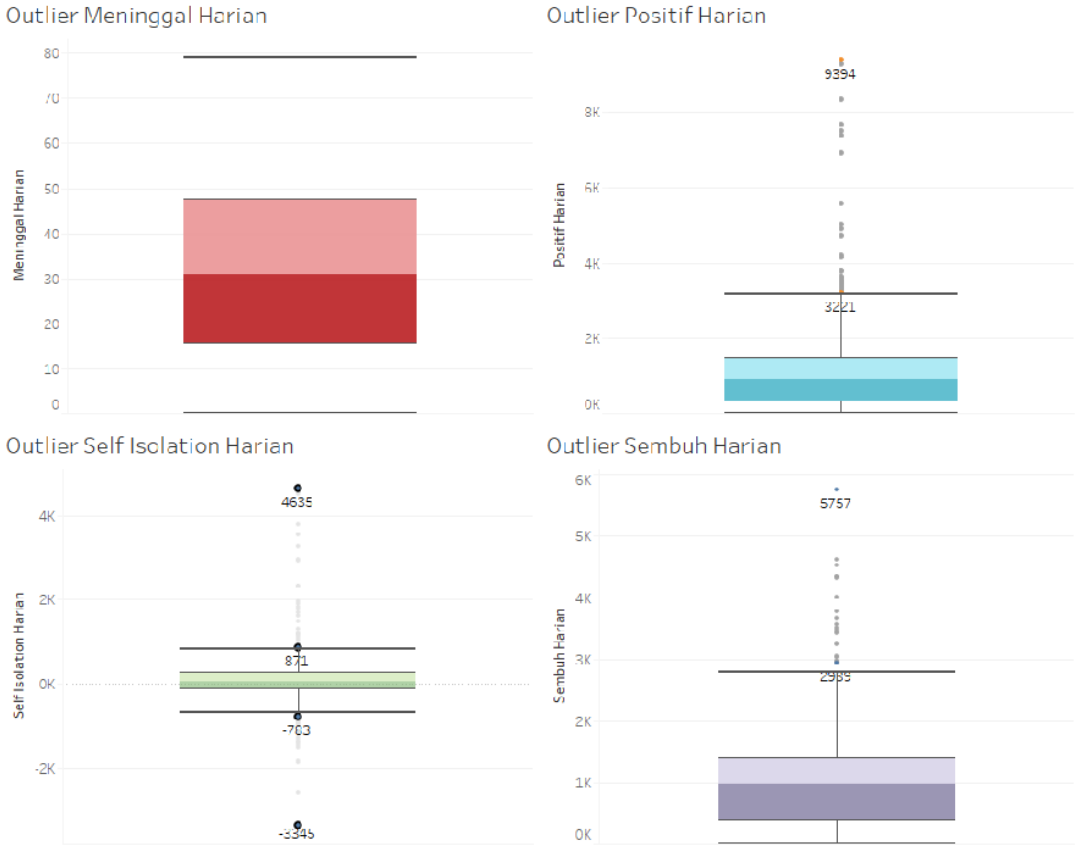
Sehingga, akan ada fitur baru yaitu fitur ‘Meninggal Harian’, ‘Self Isolation Harian’ dan ‘Masih Perawatan Harian’. Dan fitur ‘Meninggal’, ‘Self Isolation’, dan ‘Masih Perawatan’ akan dihapus.

Kemudian, kami ingin melihat trend dari COVID-19 di Jakarta. Oleh karena itu kami melakukan ekstraksi fitur ‘Tanggal’ dengan cara mengambil angka tanggal menjadi fitur ‘Tanggal Angka, mengambil angka tanggal kemudian diubah dengan rumus $((angka\ tanggal)\ div\ 7) + 1$ menjadi fitur ‘Minggu ke-’, mengambil fitur ‘Hari’, mengambil fitur ‘Bulan’, dan mengambil fitur ‘Tahun’. Sehingga, fitur yang akan kami pakai adalah fitur ‘Tanggal’ yang menjadi index

dari dataset, ‘Meninggal’, ‘Self Isolation’, ‘Masih Perawatan’, ‘Positif Harian’, ‘Sembuh Harian’, ‘Tanggal Angka’, ‘Minggu ke-’, ‘Hari’, ‘Bulan’ dan ‘Tahun’.

C.3. Deteksi *Outlier*

Tahap selanjutnya yaitu mendeteksi *outlier* pada fitur ‘Positif Harian’, ‘Sembuh Harian’, ‘Meninggal Harian’, dan ‘Self Isolation Harian’. Kami menggunakan metode Boxplot dimana Boxplot itu sendiri menggunakan metode $1.5\,x\,IQR$ untuk mendeteksi outlier.



Gambar 1. Outlier data

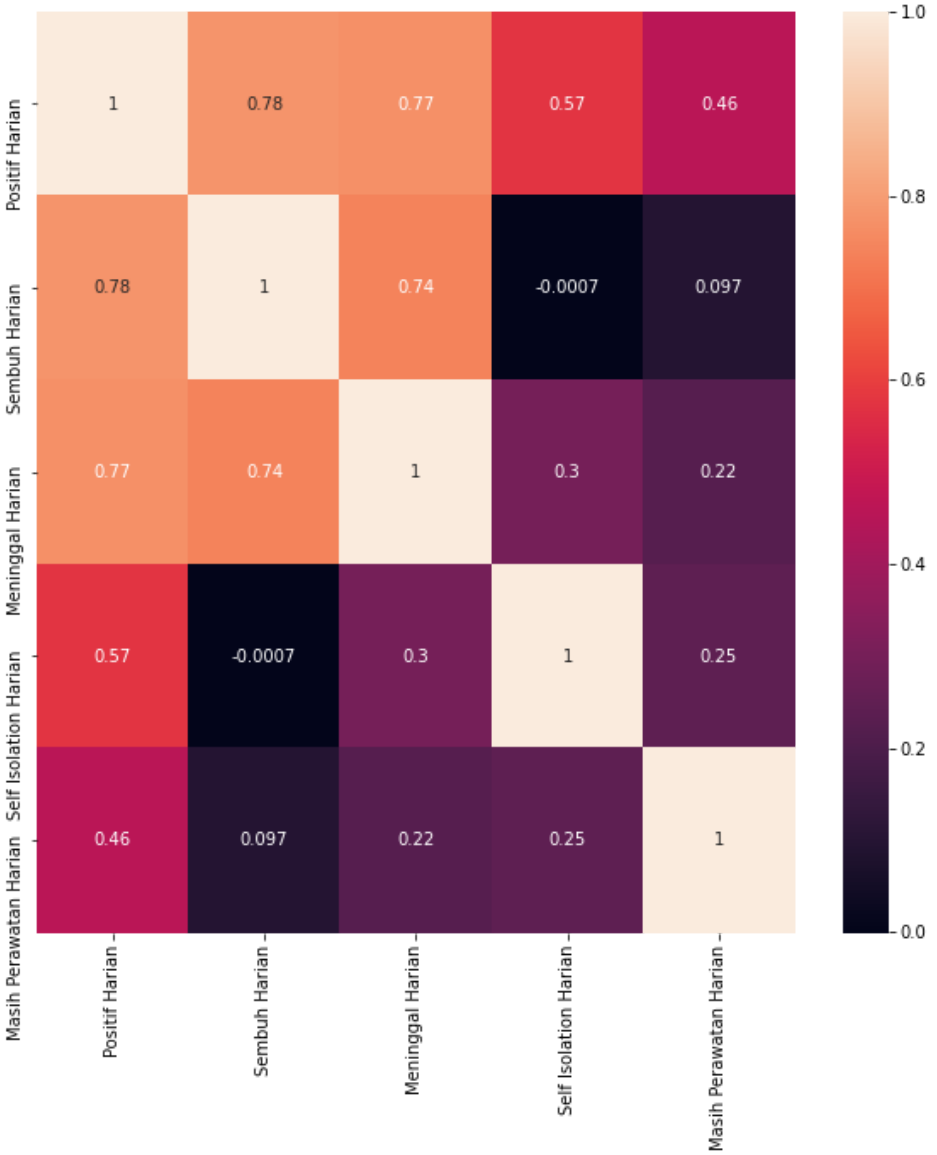
Terlihat untuk fitur ‘Meninggal Harian’ tidak ada *outlier* pada datanya. Pada fitur ‘Positif Harian’, outlier adalah data yang bernilai 3221 sampai 9394. Pada fitur ‘Self Isolation Harian’, outlier adalah data yang bernilai 871 sampai 4635 dan -783 sampai -3345. Dan yang terakhir adalah pada fitur ‘Sembuh Harian’ dimana outliernya adalah data yang bernilai 2989 sampai 5757.

Untuk *outlier* tidak akan kami tangani karena data COVID-19 ini tergolong ke dalam data kesehatan. Data outlier tidak boleh sembarangan dihapus layaknya

data *outlier* pada dataset lainnya. Justru outlier pada data COVID-19 dapat menjadi analisa lebih lanjut terkait lonjakan-lonjakan yang terjadi pada outlier.

C.4. *Exploratory Data Analysis* (EDA)

Hal yang dapat kita lakukan ketika melakukan EDA adalah dengan melihat korelasi antar fitur kontinu. Cara mudah untuk melihat korelasi antar fitur kontinu adalah dengan menggunakan Heatmap.



Gambar 2. *Heatmap* data

Seperti yang terlihat bahwa antara fitur ‘Positif Harian’, ‘Sembuh Harian’, dan ’Meninggal Harian’ memiliki korelasi positif yang kuat dengan kisaran 0.7. Hal ini memang dapat dibenarkan karena apabila jumlah pasien yang terkena virus COVID-19 meningkat, maka pasien yang dapat sembuh atau meninggal karena

terkena virus COVID-19 juga meningkat. Namun, yang menarik adalah tidak ada korelasi yang kuat antara fitur 'Self Isolation Harian' atau 'Masih Perawatan Harian'. Padahal, seharusnya kasus kedua fitur tersebut sama dengan fitur 'Sembuh Harian' dan 'Meninggal Harian' karena semakin banyak jumlah pasien yang terkena virus COVID-19, maka semakin banyak juga pasien yang akan menjalani isolasi mandiri ataupun perawatan medis.

Selanjutnya kami ingin melihat jumlah, rata-rata, persebaran, minimal, kuartil 1, median, kuartil 2, dan maksimum dari fitur-fitur kontinu.

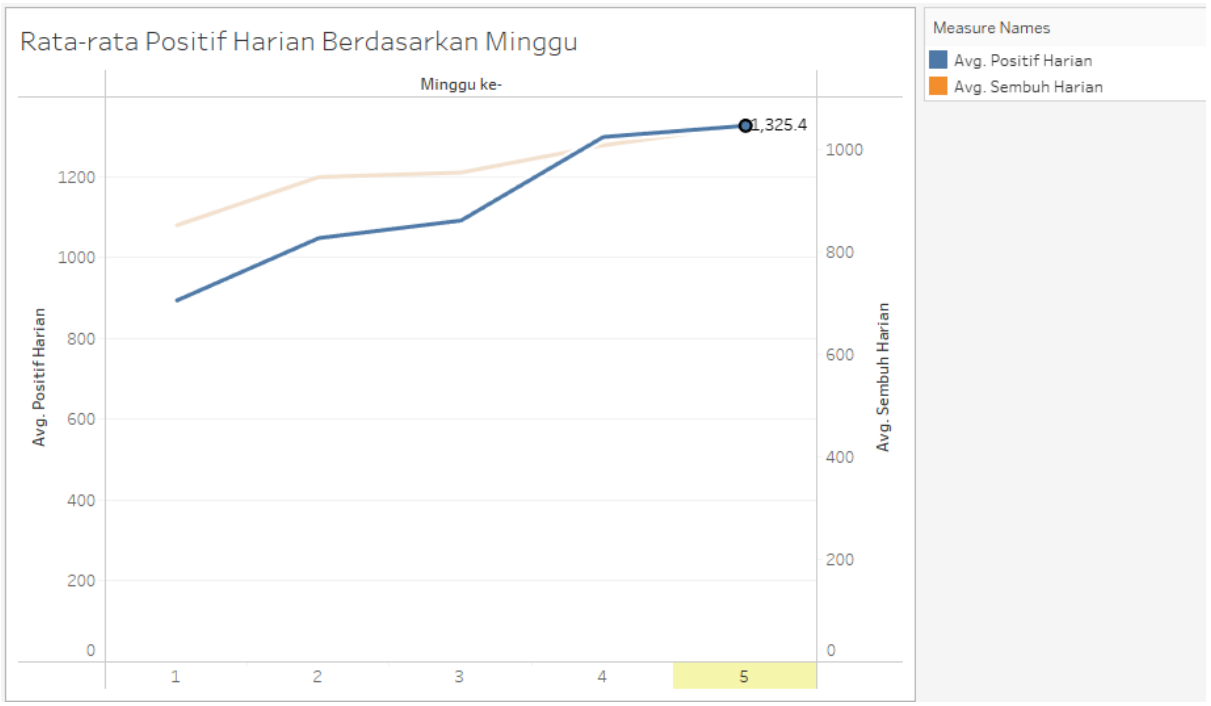
	Positif Harian	Sembuh Harian	Meninggal Harian	Self Isolation Harian	Masih Perawatan Harian
count	487.000000	487.000000	487.000000	487.000000	487.000000
mean	1115.950719	954.708419	17.425051	98.211499	45.605749
std	1293.180643	963.252516	14.607133	655.783274	338.367602
min	0.000000	0.000000	0.000000	-3345.000000	-1271.000000
25%	197.500000	150.000000	6.000000	-83.500000	-66.500000
50%	845.000000	835.000000	15.000000	29.000000	20.000000
75%	1365.000000	1233.000000	22.000000	194.000000	121.000000
max	9394.000000	5757.000000	79.000000	4635.000000	1999.000000

Gambar 3. Deskripsi data

Disini terlihat bahwa yang pertama, persebaran data pada tiap-tiap fitur sangat tinggi. Hal ini menandakan bahwa rentang data cukup jauh. Dapat dilihat dari misalnya fitur 'Positif Harian' dimana minimal data tersebut adalah 0 sedangkan maksimalnya adalah 9394. Karena dataset ini merupakan tipe dataset *Time-Series* yang artinya data ini merupakan data yang berurutan. Maka dapat diartikan juga bahwa persebaran data yang tinggi menunjukan peningkatan pasien terinfeksi COVID-19 yang tinggi pula setiap harinya.

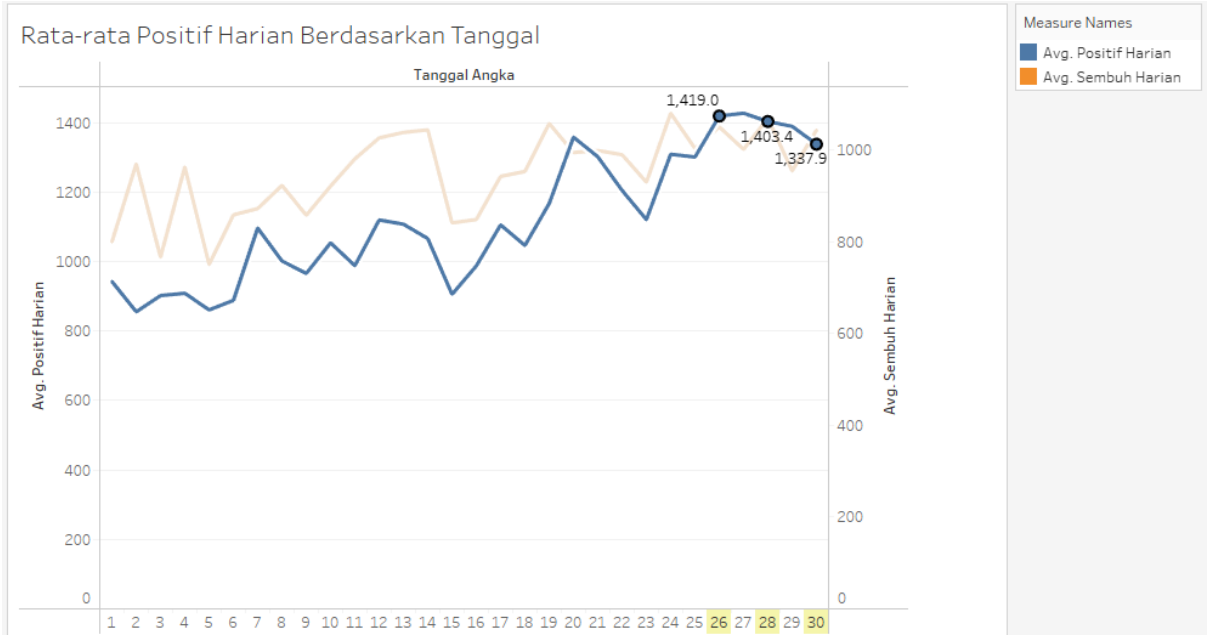
Hal kedua yang terlihat bahwa pada fitur 'Self Isolation Harian' dan 'Masih Perawatan Harian' adalah adanya data negatif pada fitur tersebut. Terlihat dari nilai minimal dan kuartil 1 dari fitur tersebut. Data negatif disini berarti bahwa ada pasien yang sudah sembuh atau meninggal atau berpindah (dari *Self Isolation* ke Masih Perawatan dan sebaliknya) sehingga data pada fitur '*Self Isolation* Harian' dan 'Masih Perawatan Harian' berupa nilai negatif.

D. Untuk menemukan informasi lebih dalam mengenai dataset ini, kami melakukan beberapa visualisasi guna mempermudah kami untuk melakukan analisis sebagai berikut.



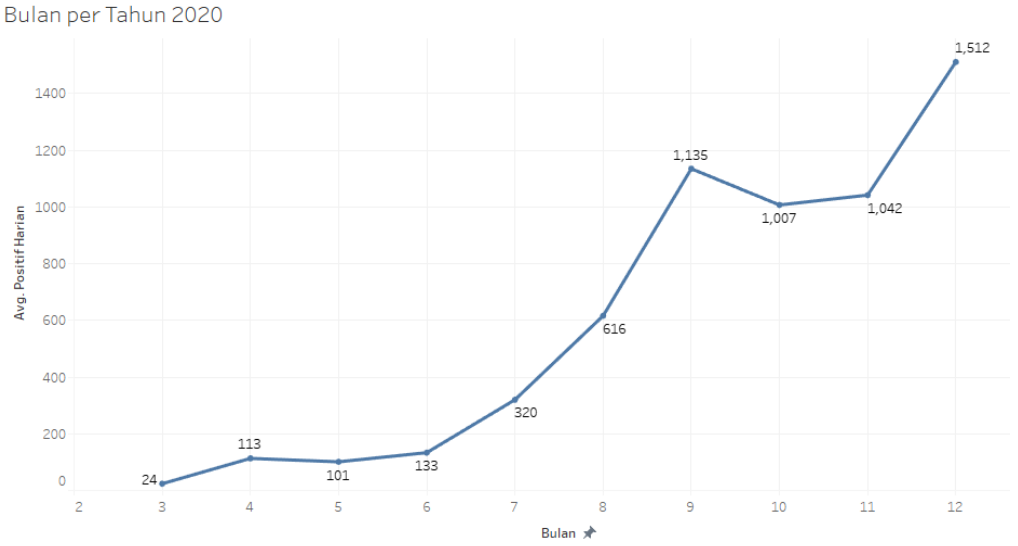
Gambar 4. Grafik Rata-rata Positif Harian Berdasarkan Minggu

Berdasarkan grafik pada gambar 4, kami menemukan bahwa rata-rata positif harian minggu kelima atau dari tanggal 29 sampai tanggal 31 adalah yang tertinggi daripada minggu-minggu lainnya. Yang artinya bahwa semakin dekat dengan akhir bulan, maka kecenderungan kasus pasien positif juga meningkat.



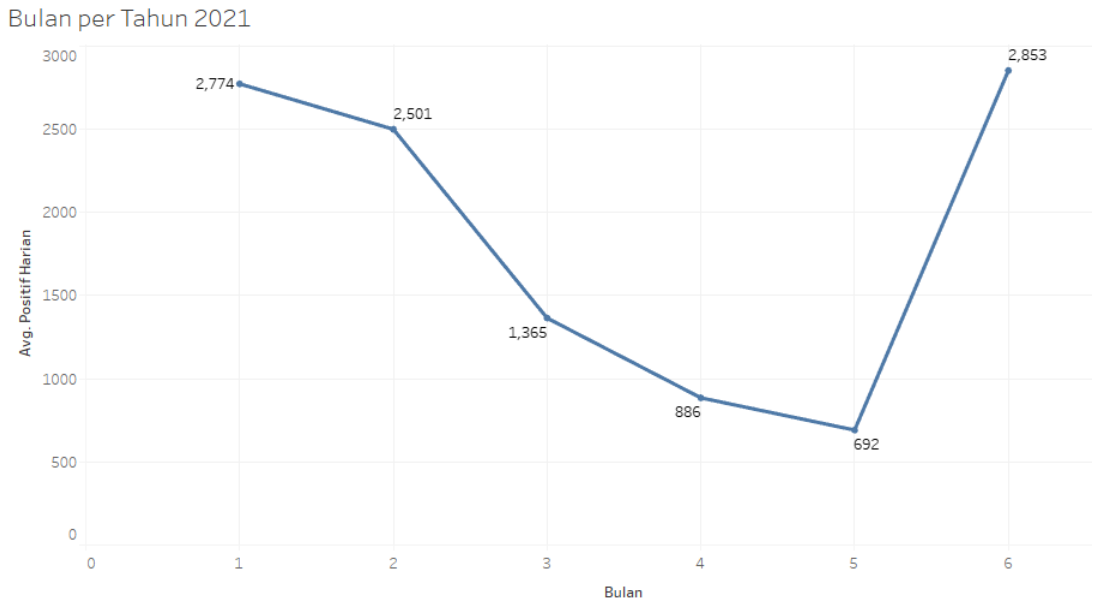
Gambar 5. Grafik Rata-rata Positif Harian Berdasarkan Tanggal

Berdasarkan grafik pada gambar 5, kami menemukan bahwa rata-rata positif harian juga cenderung meningkat pada tanggal-tanggal di akhir bulan.



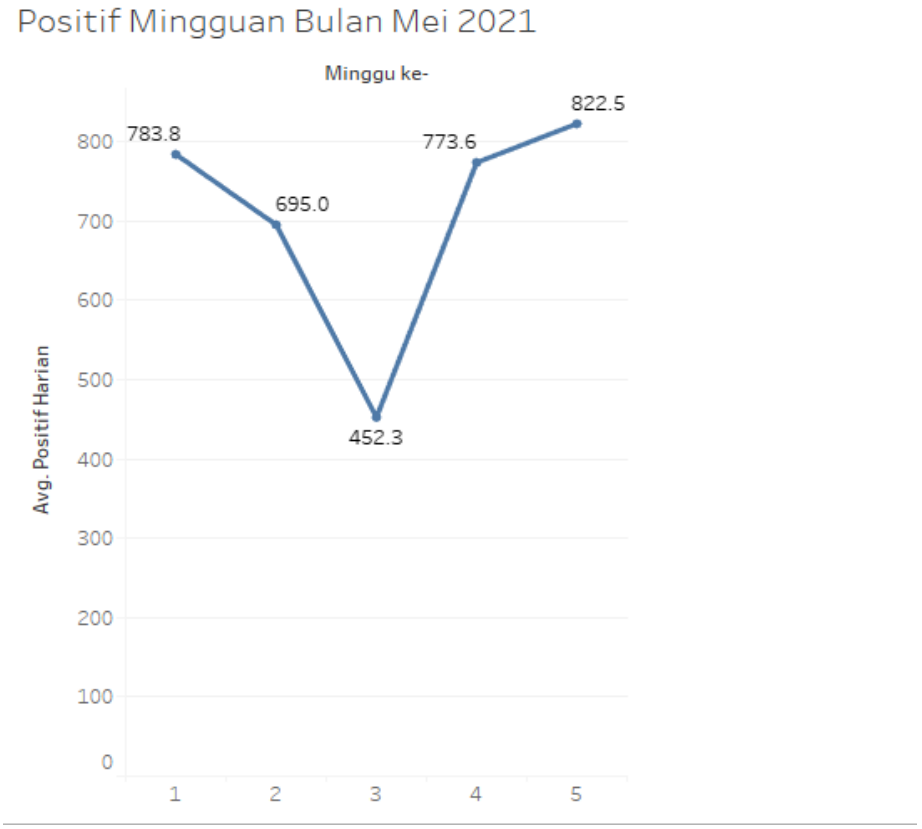
Gambar 6. Grafik Bulan per Tahun 2020

Berdasarkan grafik pada gambar 6, kami menemukan bahwa rata-rata positif harian mengalami kenaikan secara bertahap dari bulan Mei hingga bulan Desember.



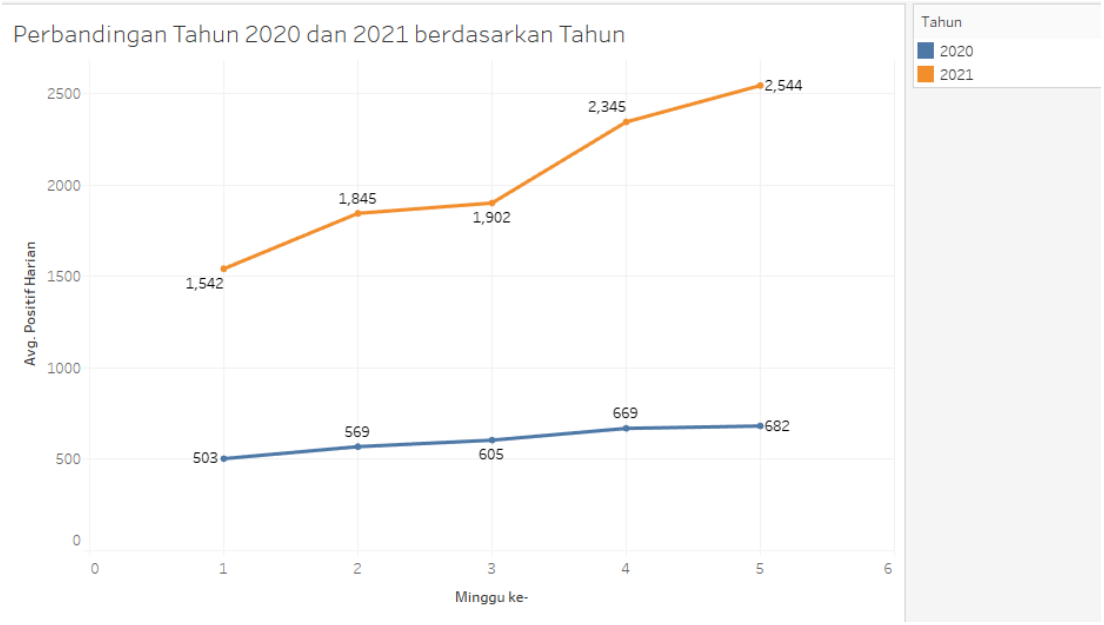
Gambar 7. Grafik Bulan per Tahun 2021

Berdasarkan grafik pada gambar 7, kami menemukan dari bulan Januari hingga bulan Mei mengalami penurunan rata-rata pasien per bulannya. Namun, tiba-tiba terjadi lonjakan besar pada bulan Juni.



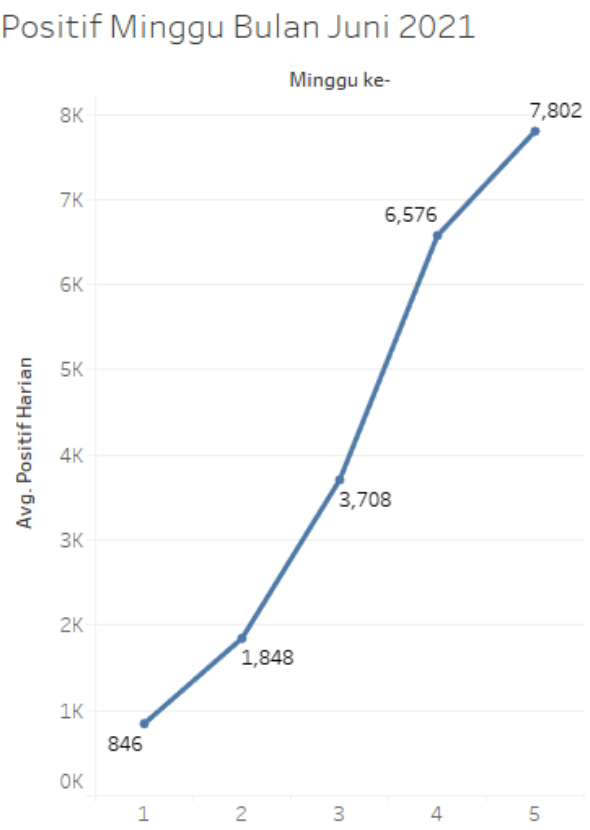
Gambar 8. Grafik Rata-rata Positif Harian Mingguan Bulan Mei 2021

Berdasarkan grafik pada gambar 8, kami menemukan dari awal bulan sampai minggu ke-3 terjadi penurunan rata-rata pasien COVID-19. Namun, dari minggu ke-4 sampai akhir bulan terjadi kenaikan kembali.



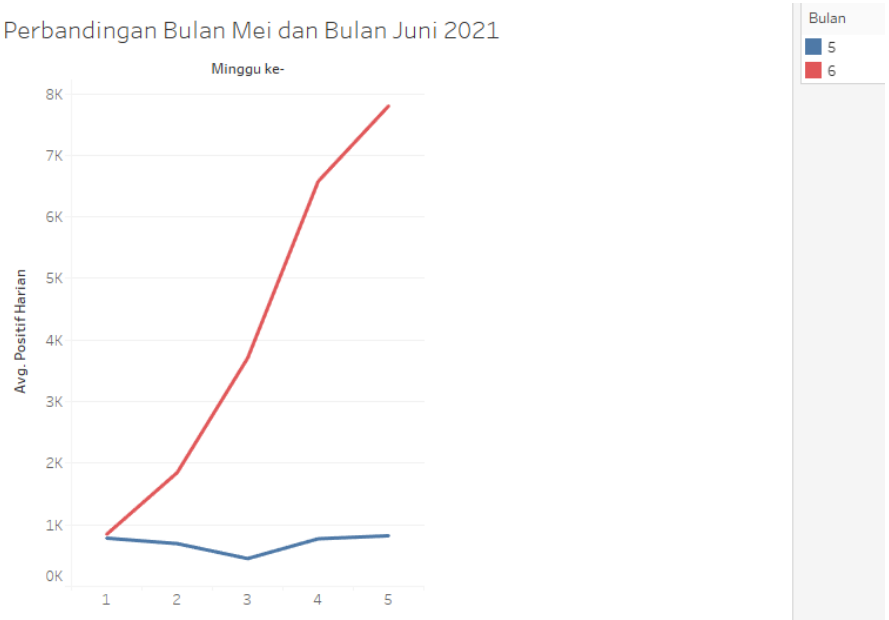
Gambar 9. Perbandingan Tahun 2020 dan 2021 berdasarkan Tahun.

Berdasarkan grafik pada gambar 9, kami menemukan bahwa rata-rata positif harian pada tahun 2021 lebih tinggi daripada tahun 2020. Dan kedua tahun juga menunjukan bahwa minggu keempat merupakan minggu tertinggi daripada minggu-minggu sebelumnya.



Gambar 10. Grafik Rata-rata Positif Harian Mingguan Bulan Juni 2021

Berdasarkan grafik pada gambar 10, kami menemukan bahwa rata-rata pasien positif harian mengalami kenaikan yang cukup signifikan dan terus naik.



Gambar 11. Grafik Perbandingan Rata-rata Positif Harian bulan Mei dan Juni 2021

Berdasarkan grafik pada gambar 11, kami menemukan bahwa perbandingan rata-rata pasien COVID-19 antara bulan Mei dan bulan Juni memiliki perbandingan yang sangat jauh kenaikannya.

E. Dengan adanya beberapa penemuan kami pada tahapan *Exploratory Data Analysis* serta *Initial Finding*, beberapa hal yang ingin kami ketahui lebih dalam adalah :

1. Pada gambar 4, terlihat bahwa minggu ke-4 dan minggu ke-5 merupakan minggu yang paling banyak memiliki rata-rata positif COVID-19. Dugaan kami, hal ini disebabkan karena orang bepergian ketika akhir bulan. Sehingga terjadi kerumunan ketika akhir bulan tersebut yang menyebabkan jumlah pasien COVID-19 pada minggu ke-4 lebih tinggi.
2. Pada gambar 5, terlihat bahwa dari tanggal 26 sampai akhir bulan merupakan puncak rata-rata pasien positif COVID-19. Sama seperti dugaan kami pada nomor sebelumnya, hal ini disebabkan karena orang bepergian ketika akhir bulan. Lebih tepatnya, berdasarkan grafik terlihat bahwa positif COVID-19 terbanyak cenderung berada di 7 hari terakhir setiap bulannya.
3. Pada gambar 6, kami menemukan bahwa jumlah positif rata-rata harian cenderung meningkat setiap bulannya. Dugaan kami dikarenakan Jakarta merupakan kota metropolitan yang menjadi pusat perekonomian nasional, dimana hal ini menimbulkan banyak klaster di berbagai sektor, salah satunya adalah perkantoran. Kemudian, mayoritas masyarakat umum juga belum mendapatkan vaksinasi sebagai bentuk proteksi terhadap penyakit COVID-19. Terakhir, banyaknya masyarakat yang belum percaya akan bahaya yang disebabkan oleh pandemi ini.
4. Pada gambar 7, kami menemukan bahwa awal tahun sampai bulan Mei terjadi penurunan yang cukup stabil terhadap kasus pasien positif harian. Namun, terjadi kenaikan yang signifikan pada kasus yang sama pada bulan Juni 2021. Dugaan kami, fenomena terjadi akibat masyarakat Jakarta, dimana 40% penduduknya merupakan pendatang, merayakan Idul Fitri di kampung halamannya. Tradisi yang biasa disebut pulang kampung ini mengakibatkan penurunan rata-rata pasien COVID-19 di wilayah DKI Jakarta. Namun, kenaikan kasus terjadi semenjak

kembalinya masyarakat untuk bekerja, ditambah lagi dengan liburan saat hari raya yang menimbulkan kerumunan di berbagai daerah yang rawan keramaian.

5. Pada gambar 8, kami menemukan pola yang janggal, dimana pada minggu ke-3 rata-rata positif harian menurun. Ternyata hal ini sesuai dengan dugaan kami pada nomor sebelumnya, dimana pada saat hari raya tersebut tingkat positif sedang menurun karena masyarakat pendatang Jakarta sedang pulang kampung. Kemudian, minggu-minggu selanjutnya merupakan momen arus balik sehingga positif harian kembali naik.
6. Pada gambar 9, terlihat perbandingan rata-rata kasus positif harian tiap minggunya pada tahun 2020 dengan 2021 menunjukkan perbedaan yang cukup signifikan. Padahal, rata-rata pasien COVID-19 pada tahun 2021 mulai mereda sejak bulan Januari sampai bulan Mei. Ini membuktikan bahwa bulan Juni sangat mempengaruhi grafik pada tahun 2021, yang menandakan bahwa kasus penyakit COVID-19 bulan ini perlu perhatian ekstra.
7. Pada gambar 10, kami menemukan bahwa rata-rata pasien COVID-19 pada bulan Juni meningkat drastis. Menurut kami, hal ini disebabkan karena adanya varian baru yang masuk ke Indonesia yang disebut Varian Delta dan adanya efek lanjut dari perayaan Idul Fitri. Akibatnya, virus penyebab COVID-19 terus menyebar dengan cepat sehingga terjadi peningkatan rata-rata pasien COVID-19 sangat tinggi di bulan Juni.
8. Pada gambar 11, terlihat perbandingan rata-rata pasien COVID-19 pada bulan Juni menunjukkan peningkatan, sesuai dengan pernyataan-pernyataan pada soal sebelumnya. Hal ini menunjukkan bahwa bulan Juni harus mendapatkan perhatian lebih dari berbagai pihak agar pandemi dapat ditangani dengan tepat.

F. Jakarta merupakan pusat perekonomian negara yang menggerakkan roda perekonomian masyarakatnya. Dengan adanya pandemi COVID-19 disaat mobilitas kota yang sangat tinggi, Jakarta menjadi daerah dengan kasus positif pasien COVID-19 terbanyak di Indonesia. Berdasarkan hasil analisis kami, dapat disimpulkan bahwa terdapat trend pada rata-rata pasien COVID-19 yang

cenderung mengalami kenaikan setiap akhir bulan. Selain itu, dapat diketahui juga bahwa terdapat trend dimana bulan Juni memiliki kenaikan rata-rata pasien COVID-19 yang sangat tinggi. Berdasarkan itu, kami merekomendasikan pemerintah yang berwenang untuk membuat regulasi dengan lebih memperketat protokol kesehatan setiap akhir bulan. Selain itu, peningkatan terhadap pengawasan terhadap pelanggaran pada pelaksanaan protokol kesehatan juga harus dilakukan untuk meminimalisir penyebaran virus. Kemudian, kami juga menyarankan agar pemerintah yang berwenang untuk segera melakukan karantina wilayah Ibukota Jakarta karena grafik menunjukkan bahwa bulan Juni 2021 terjadi peningkatan kasus positif yang sangat signifikan dibandingkan bulan-bulan sebelumnya. Terakhir, Kami juga menyarankan kepada masyarakat umum untuk tetap mematuhi protokol kesehatan mengingat perlawanan terhadap pandemi merupakan kolaborasi antara berbagai pihak, termasuk masyarakat Indonesia secara umum.

6. Terkait pertanyaan sebelumnya, jelaskan dengan lebih detail usaha exploratory data analysis (EDA) yang kalian lakukan dan mengapa kalian melakukan teknik EDA tersebut.

   Jawaban:

   Kami menggunakan teknik EDA yaitu *multivariate graphical*. Kami menggunakan *multivariate graphical* pada Heatmap dan Boxplot. Kami menggunakan *multivariate graphical* karena ingin mencari informasi terhadap dua atau lebih variabel terhadap variabel tertentu yaitu korelasi setiap variabel dan outlier pada tiap-tiap variabel dengan grafik agar mudah terlihat informasinya dan dapat dengan mudah dipahami.

DAFTAR PUSTAKA

Djalante, R., Lassa, J., Setiamarga, D., Sudjatma, A., Indrawan, M., Haryanto, B., ... & Warsilah, H. (2020). Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020. *Progress in Disaster Science*, *6*, 100091.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., & Asfian, P. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the total environment*, *725*, 138436.

Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical medicine & international health : TM & IH*, *25*(3), 278–280. https://doi.org/10.1111/tmi.13383

ZHANG, RENYI, et al. (2020). Identifying airborne transmission as the dominant route for the spread of COVID-19. Proceedings of the National Academy of Sciences.

LAMPIRAN

Erland366/compfest-dsa-yannlecun (github.com/Erland366/compfest-dsa-yannlecun)