Mata Kuliah Pilihan Data Science (3 SKS)

Covid-19 Case Data Science and Data Visualization

Mustakim, S.T., M.Kom

Program Studi Sistem Informasi

https://mustakim.predatech.org | mustakim@uin-suska.ac.id | +6285275359942

Visualisasi Data

Mustakim, S.T., M.Kom.

Case Data Visualization

Salah satu pekerjaan penting seorang data scientist adalah membuat visualisasi dari data yang efektif dalam menjawab tujuan yang disasar

Case Data Visualization

Tahap pembuatan visualisasi dari data

Case Data Visualization

Tahap pembuatan visualisasi dari data:

1. Pertama, merumuskan insights apa saja yang ingin "digali" dan disampaikan dari data yang dimiliki. Namun, sebelum dapat merumuskannya, semua elemen data harus dipelajari dengan seksama dan teliti dulu sehingga data benar-benar dipahami dan dikuasai.
2. Kedua, menentukan bentuk visualisasi, apakah itu grafik, text atau tabel. Bentuk perlu dipilih yang sesuai dan efektif untuk menyampaikan tiap informasi dan audiens atau pembaca yang ditarget.

Tahap pembuatan visualisasi dari data:

3. Ketiga, memilih tools, software atau perangkat lunak yang tepat untuk tiap bentuk visual yang akan dibuat.
4. Keempat, menyiapkan data dengan format sedemikian rupa, sehingga dapat ditangani atau diproses oleh tools yang dipilih. Jika penyiapan data tidak dapat dilakukan dengan tools itu, maka perlu merancang algoritma dan dilanjutkan dengan pembuatan program dengan Python atau bahasa pemrograman lainnya.
5. Kelima, membuat visualisasi dari data (dengan tools atau program) yang telah disiapkan. Ini biasanya tidak "sekali jadi". Setelah bentuk visual ada, harus dievaluasi apakah sudah jelas, bagus, dan informasi tersampaikan.

Case Data Visualization: Covid-19 Korea Selatan

Data yang tersedia merupakan hasil rekaman kasus-kasus mulai 20 Januari s/d 30 April 2020 (Kaggle, 2020). Setiap data berupa tabel, yang dapat dibuka dengan Excel.

DATA KASUS

- Case.csv (112 baris): Kasus-kasus terpapar COVID-19 dengan kolom case_id, province, city, group, infection_case, confirmed, latitude dan longitude

DATA PASIEN

- PatientInfo.csv (3.388 baris): Data epidemis pasien COVID-19 dengan kolom patient_id, global_num, sex, birth_year, age, country, province, city, disease, infection_case, infection_order, infected_by, contact_number, symptom_onset_date, confirmed_date, released_date,deceased_date, state dan confirm_released.

Case Data Visualization: Covid-19 Korea Selatan

DATA TIME SERIES

- Time.csv (102 baris): data untuk status COVID-19 dengan kolom date, time, test, negative, confirmed, released dan deceased.
- TimeAge.csv (540 baris): data untuk status COVID-19 berdasar umur dengan kolom date, time, age, confirmed, dan deceased.
- TimeGender.csv (120 baris): data untuk status COVID-19 berdasar gender dengan kolom date, time, sex, confirmed dan deceased.
- TimeProvince.csv (1.734 baris): data untuk status COVID-19 untuk tiap provinsi dengan kolom date, time, province, confirmed, released dan deceased.

Bentuk-bentuk Visualisasi Data

- Garis (line): Cocok untuk data "time-series" dan memberikan *trend*, misalnya harga satu atau lebih saham dari tanggal ke tanggal.
- Plot tersebar (*scatter plot*): Cocok untuk menunjukkan hubungan antara dua nilai variabel, misalnya berat terhadap tinggi badan dari para pemain sepakbola.
- Bar vertikal: Cocok digunakan ketika ingin ditunjukkan nilai-nilai beberapa variabel atau kategori agar terlihat perbandingaannya.
- Bar horisontal: Sama dengan bar vertikal, namun lebih cocok digunakan ketika nama variabel atau kategori dari data panjang (misalnya, nama provinsi).
- Teks sederhana: Jika terdapat satu atau dua angka penting yang akan dibagikan, visualisasi ini pas untuk digunakan.

Bentuk-bentuk Visualisasi Data

- Garis (line): Cocok untuk data "time-series" dan memberikan *trend*, misalnya harga satu atau lebih saham dari tanggal ke tanggal.
- Plot tersebar (*scatter plot*): Cocok untuk menunjukkan hubungan antara dua nilai variabel, misalnya berat terhadap tinggi badan dari para pemain sepakbola.
- Bar vertikal: Cocok digunakan ketika ingin ditunjukkan nilai-nilai beberapa variabel atau kategori agar terlihat perbandingaannya.
- Bar horisontal: Sama dengan bar vertikal, namun lebih cocok digunakan ketika nama variabel atau kategori dari data panjang (misalnya, nama provinsi).
- Teks sederhana: Jika terdapat satu atau dua angka penting yang akan dibagikan, visualisasi ini pas untuk digunakan.

Penggalian Insights

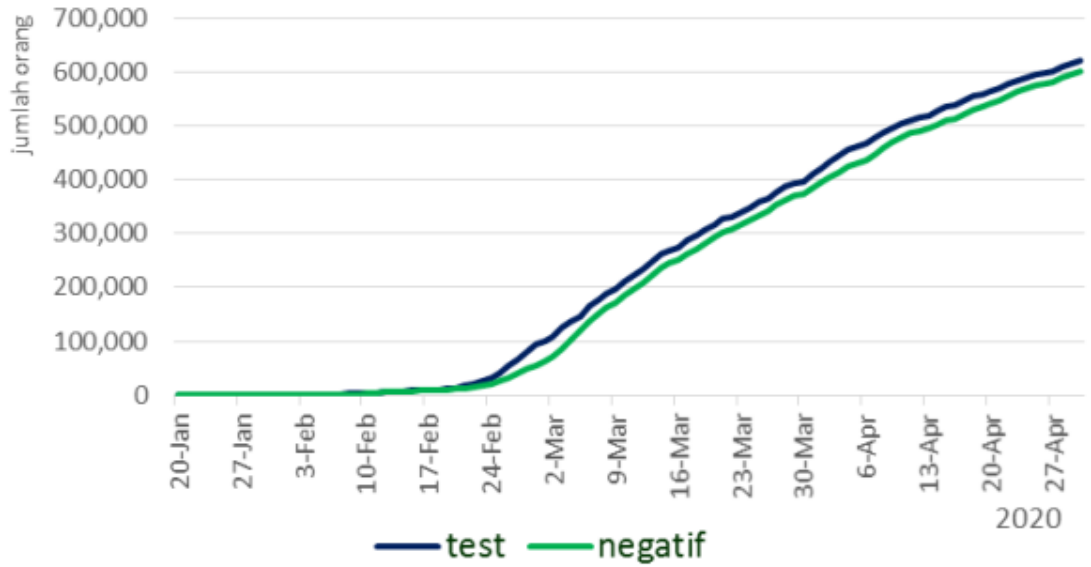
Case Data Visualization: Covid-19 Korea Selatan

Diberikan 14 contoh pertanyaan yang timbul yang didasari karena adanya rasa ingin tau terhadap data, apa yang dilakukan untuk menjawab pertanyaan, sampai mendapatkan visualisasi dan insights yang tersampaikan melalui visualisasi itu

Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-1:
Bagaimana trend test COVID-19 dilakukan di Korsel dari waktu ke waktu? Apakah banyak orang yang "terbebas"?

Bentuk visual yang cocok adalah garis, yang merepresentasikan jumlah (test dan yang negatif) terhadap waktu.



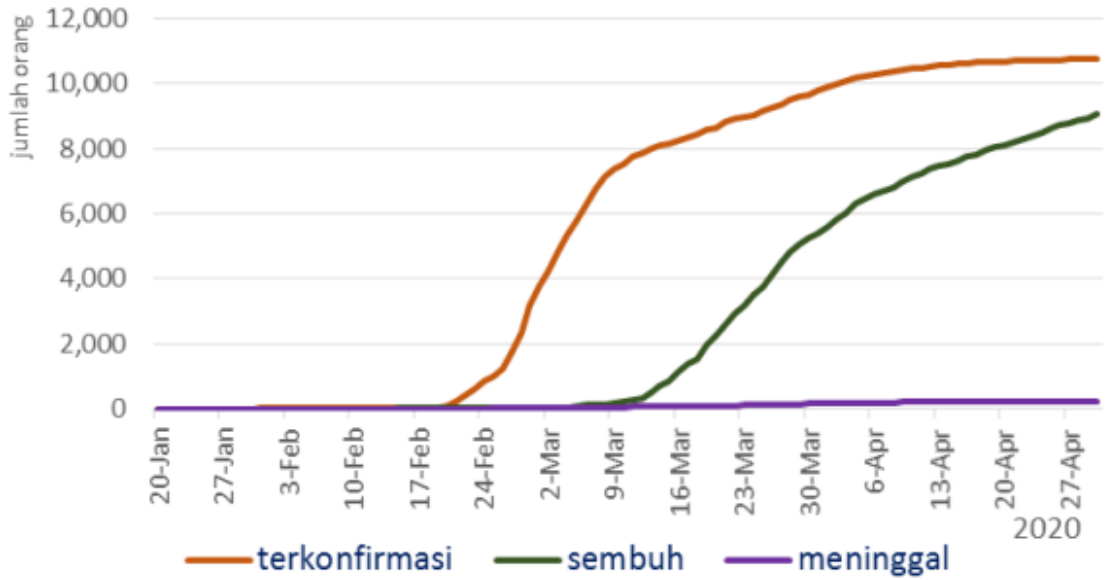
Gambar 6.3. Trend jumlah test dan hasil yag negatif.

Insights dari data: Test dilakukan dengan cepat (grafik naik eksponensial dari Februari ke akhir April) dan dari waktu ke waktu, hasilnya sebagian besar negatif

Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-2:
Bagaimana trend akumulasi terkonfirmasi (positif), yang sembuh dan meninggal dari waktu ke waktu?

Sama dengan trend test, visualisasi yang cocok adalah grafik garis. Data tersedia pada file Time.csv, kolom date, confirmed, released dan deceased.



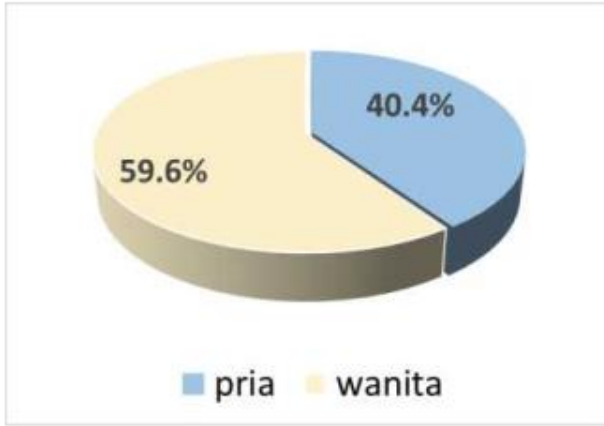
Gambar 6.4. Akumulasi terkonfirmasi, sembuh dan meninggal.

Insights dari data: Penyebaran COVID-19 di Korsel segera terkendali (grafik naik dari pertengahan Februari sampai akhir Maret, selanjutnya landai). Bagi yang terpapar, proses penyembuhan juga relatif cepat (grafik naik secara tajam dari 9 Maret sampai akhir April).

Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-3:
Jika di banyak negara, pria lebih banyak yang terinfeksi COVID-19, bagaimana dengan di Korsel?

Untuk menjawab pertanyaan tersebut, data dapat diperoleh dari file TimeGender.csv pada dua baris terakhir, yang berisi jumlah wanita dan pria yang terkonfirmasi terpapar COVID-19 dan yang meninggal pada tanggal 30 April 2020.



Gambar 6.5. Persentase terinfeksi COVID-19 berdasar gender.

Insights dari data: Di Korsel lebih banyak wanita, sekitar 2/3 dari total, yang terinfeksi.

Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-4:
Bagaimana tingkat kematian dari yang terinfeksi? Apakah wanita, yang lebih banyak terinfeksi, memiliki resiko kematian yang lebih tinggi pula?

Untuk menjawabnya, digunakan data dua baris terakhir dari file TimeGender.csv. Persentase meninggal wanita dan pria dihitung dari jumlah per gender dan dari total yang terinfeksi dari kedua gender.

Gambar 6.6. Persentase meninggal berdasar gender.

Insights dari data: Dibanding banyak negara lain (misalnya USA, Itali, UK dan Perancis, dimana resiko kematian mencapai lebih dari 5%15), tingkat kematian akibat COVID-19 di Korsel lebih rendah. Pria memiliki resiko hampir dua kali dibanding wanita.

Case Data Visualization: Covid-19 Korea Selatan

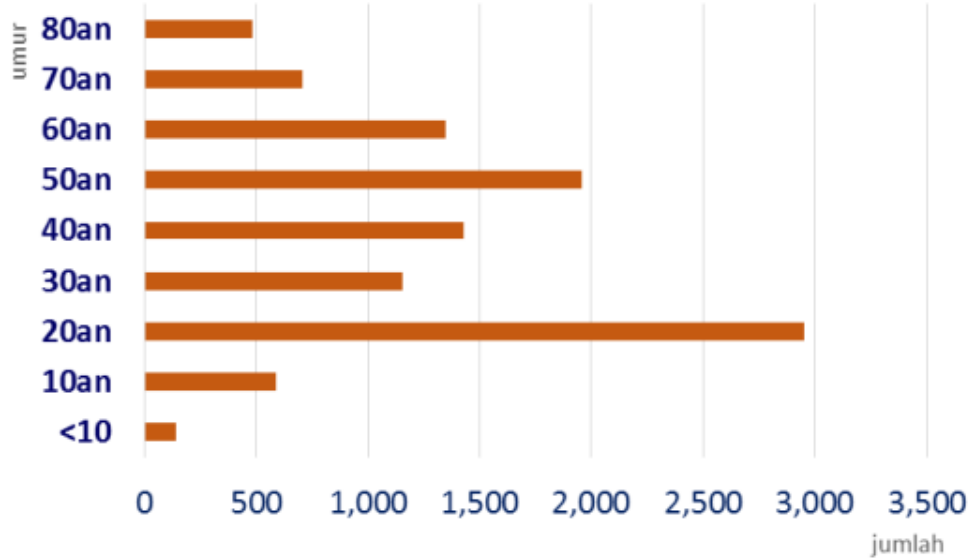
Pertanyaan-5:

Berbagai hasil analisis data COVID-19 berdasar umur menunjukkan hasil bahwa dari satu negara ke negara lain, distribusi orang yang terserang COVID-19 berbeda-beda. Ada orang-orang yang mengira bahwa COVID-19 lebih banyak "menyerang kaum tua". Bagaimana dengan di Korsel? Bagaimana persentase tiap kelompok umur?

Untuk menjawabnya, data tersedia di file TimeAge.csv. Namun harus dipilih jumlah per kelompok umur pada tanggal terakhir, yaitu 30 April 2020. Untuk mevisualisasikan jumlah terinfeksi pada tiap kelompok umur, dipilih grafik bar horisontal agar perbandingan terlihat jelas.

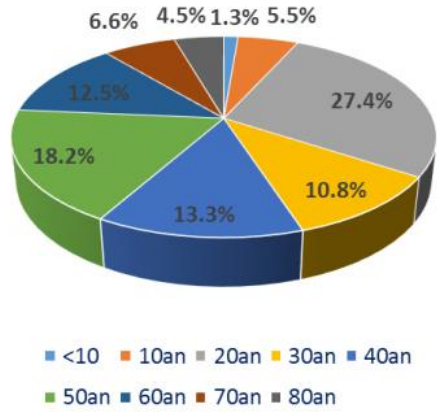
Insights dari data: Dibanding banyak negara lain (misalnya USA, Itali, UK dan Perancis, dimana resiko kematian mencapai lebih dari 5%15), tingkat kematian akibat COVID-19 di Korsel lebih rendah. Pria memiliki resiko hampir dua kali dibanding wanita.

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.7. Distribusi terkonfirmasi COVID-19 berdasar kelompok umur.

Setelah mendapatkan jumlah terinfeksi per kelompok umur, dapat dihitung persentasenya. Tiap jumlah dibagi dengan total terinfeksi (10.765). Untuk menunjukkan "porsi kue" (dari total 100%) per kelompok umur



Gambar 6.8. Persentase terkonfirmasi COVID-19 berdasar umur.

Insights dari data: Yang terpapar COVID-19, terbanyak di umur 20-an, kedua di 50-an, ketiga di 40-an. Jadi, berbeda dengan anggapan banyak orang, di Korsel ternyata umur 20-an memiliki resiko tertinggi terinfeksi COVID-19.

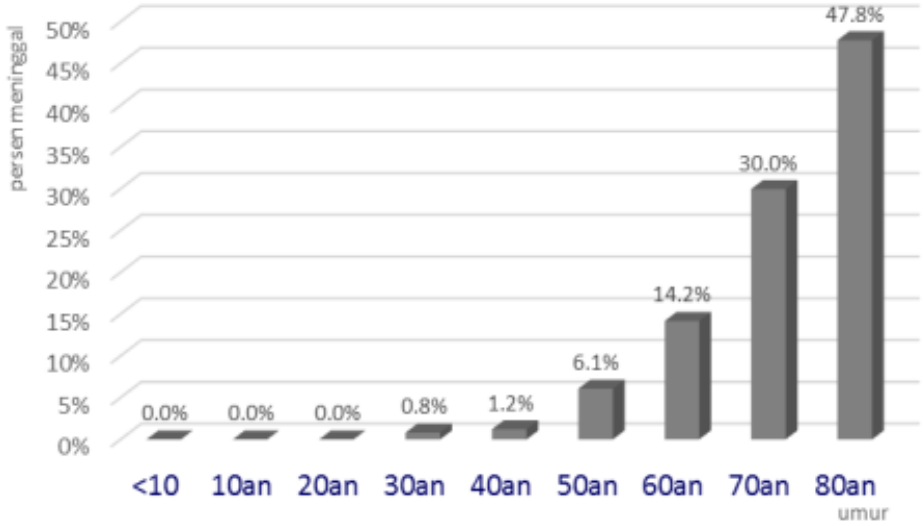
Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-6:

Hasil analisis dari berbagai negara mengindikasikan bahwa semakin tua pasien, resiko kematian semakin tinggi. Untuk Indonesia, berdasar informasi pada website Peta Sebaran, mulai umur 45 persentase meninggal di atas 40%. Bagaimana dengan pasien di Korsel?

Untuk menjawabnya, data harus disiapkan dari file TimeAge.csv. Data jumlah orang meninggal dipilih per kelompok umur pada tanggal terakhir, yaitu 30 April 2020. Lalu persentase dihitung untuk tiap kelompok umur dengan membaginya dengan jumlah total meninggal. Di sini, dipilih grafik bar vertikal agar kenaikan dari umur <10 sampai 80-an terlihat jelas. Dengan

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.9. Persentase meninggal karena COVID-19 berdasar umur.

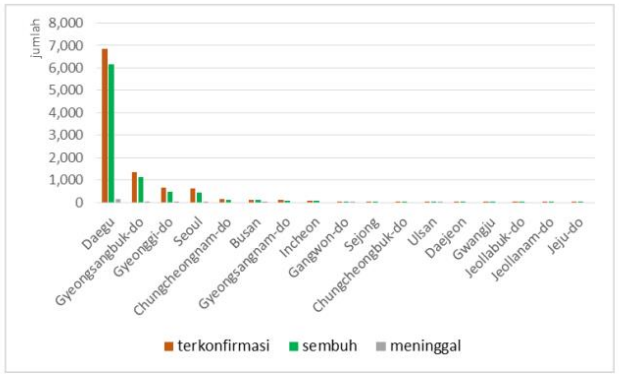
Insights dari data: Makin tua umur orang yang terinfeksi COVID-19 makin besar resiko kematiannya. Resiko meningkat tajam sejak umur 50-an.

Case Data Visualization: Covid-19 Korea Selatan

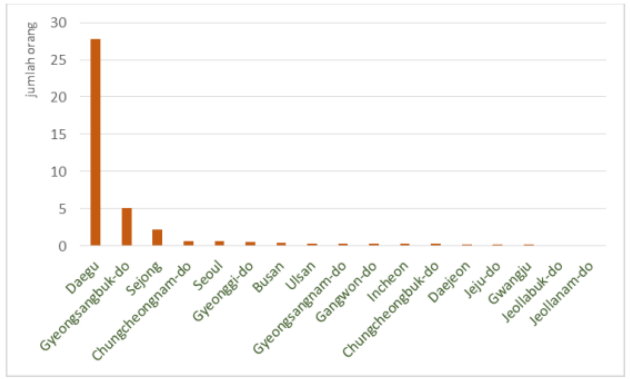
Pertanyaan-7:
Korsel memiliki 17 provinsi. Apakah seluruh provinsi sudah terpapar? Bagaimana tingkat paparan terhadap jumlah penduduk? Bagaimana perbandingan terinfeksi (terkonfirmasi), sembuh dan meninggal di tiap provinsi?

Untuk menjawabnya, data diambil dari 17 baris terakhir dari file TimeProvince.csv. Hasilnya lalu diurutkan dari terbesar ke lebih kecil dan digunakan untuk membuat grafik bar vertikal pada Gambar 6.10, sedangkan perbandingan jumlah terkonfirmasi per 10.000 penduduk diberikan pada Gambar 6.11.

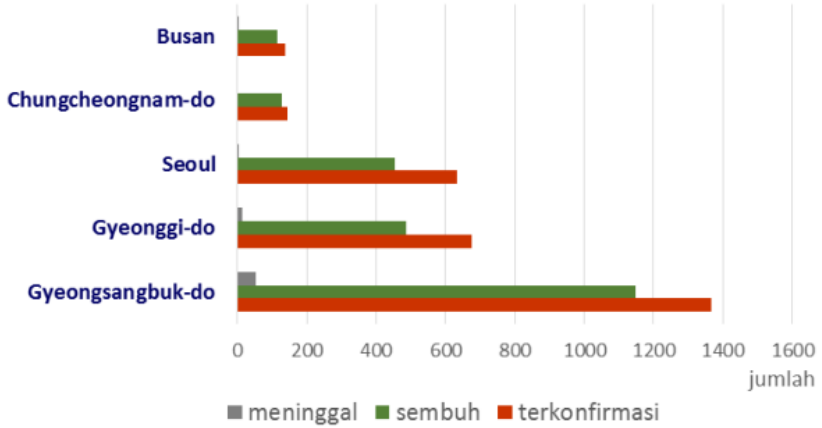
Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.10. Perbandingan jumlah terkonfirmasi, sembuh dan meninggal di seluruh provinsi.



Gambar 6.11. Jumlah terkonfirmasi per 10.000 penduduk di semua provi



Gambar 6.12. Top-5 provinsi (di bawah Daegu).

Karena bar Daegu terlalu tinggi, perbandingan terkonfirmasi – sembuh – meninggal di provinsi lainnya tidak jelas. Maka, dibuat juga grafik bar untuk top-5 provinsi di bawah Daegu

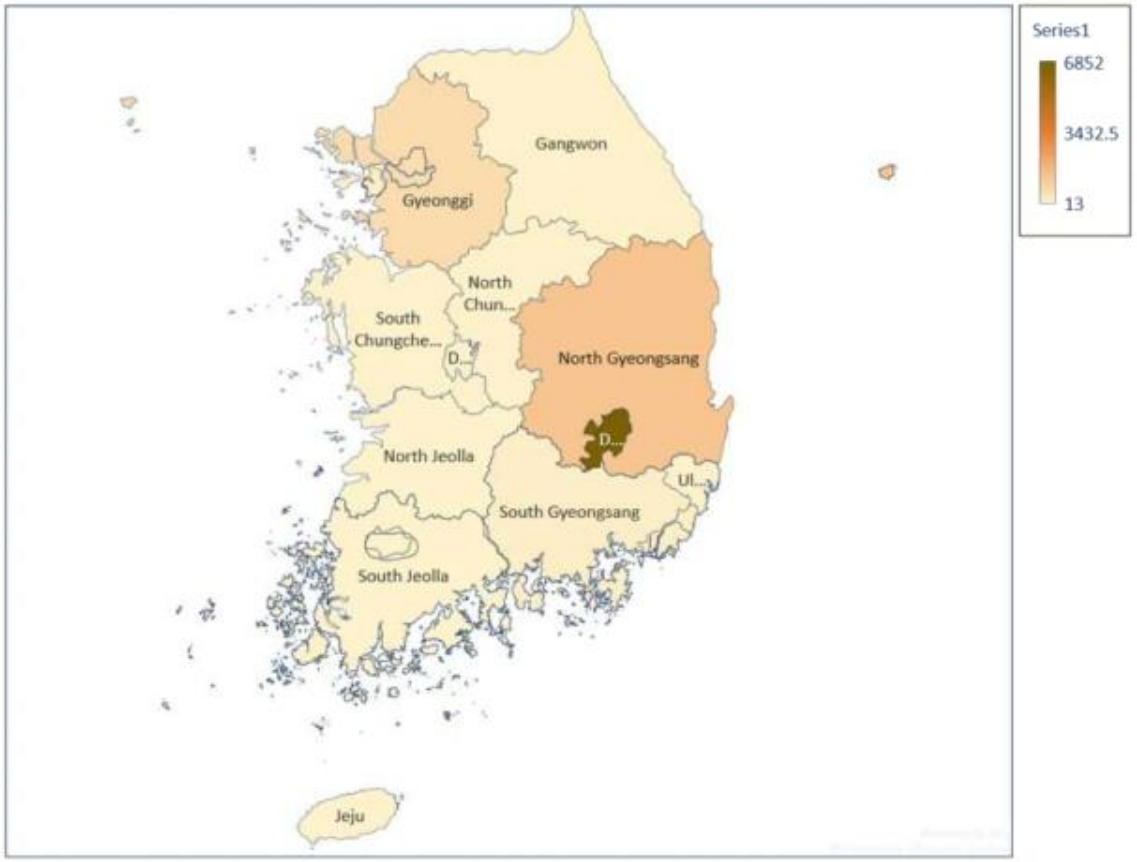
Insights dari data: Jumlah terinfeksi di provinsi Daegu, jauh melampaui yang lain, disusul Gyeongsakbuk-do, Gyeonggi-do, dan Seuol. Setelah itu, jumlah relatif sedikit

Pertanyaan-8:

Jika pada Gambar 6.12 ditunjukkan bahwa pada beberapa provinsi memiliki angka paparan yang tinggi, apakah lokasi mereka berdekatan?

Untuk menjawab pertanyaan itu, perlu dicari tools yang dapat memaparkan peta distribusi per provinsi. Excel versi 2016 ke atas sudah memiliki kemampuan untuk membuat visualisasi distribusi pada pada peta. Namun, pada saat membuatanya harus terkoneksi ke Internet untuk mendapatkan dengan peta. Pada Gambar 6.13 diberikan hasil visualisasi yang dibuat dengan Excel. Opsi lain adalah membuat program dengan Python dengan menggunakan library Geopandas yang instalasinya tidak mudah karena membutuhkan kecocokan berbagai library. Program lalu dibuat dengan masukan data paparan tiap provinsi di atas dan peta Korsel.

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.13. Tingkat paparan pada tiap provinsi di Korsel.

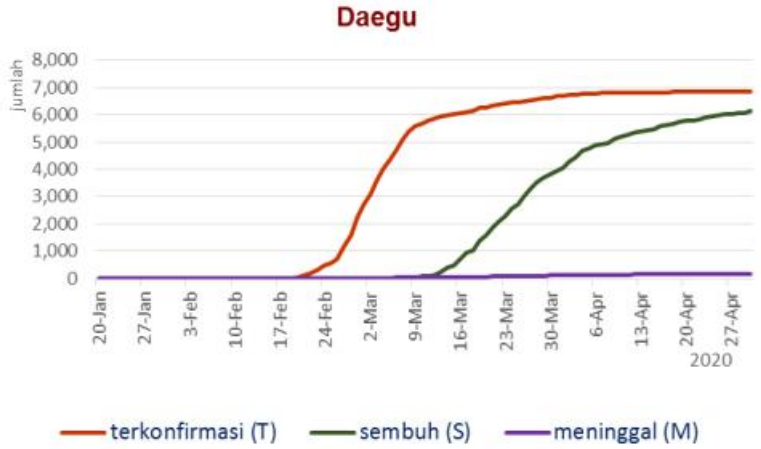
Insights dari data: Di sekitar provinsi Daegu, paparan cukup tinggi. Jadi, Daegu menjadi provinsi episentrum COVID-19. Episentrum lainnya terletak di sebelah utara, povinsi Gyeonggi dan Seoul yang berdekatan.

Pertanyaan-9:

Bagaimana trend atau pola terkonfirmasi dan sembuh di tiap propinsi berdasar waktu?

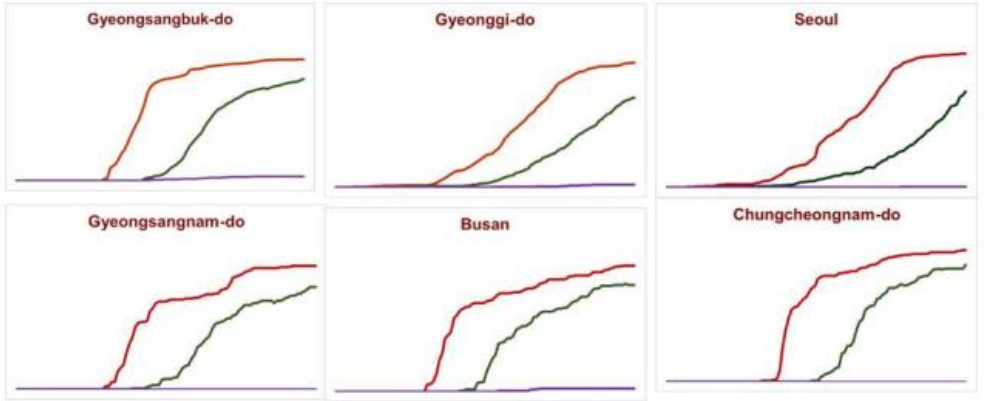
Data tersedia di file TimeProvince.csv, namun harus dipilih dulu. Pemilihan data untuk tiap provinsi dapat dengan mudah dilakukan dengan Excel (fitur filter). Tanggal perlu diubah, lalu dibuat grafik garis yang menunjukkan trend. Untuk menghemat tempat di buku ini, grafik tunggl dibuat untuk provinsi Daegu yang memiliki kasus terkonfirmasi/terinfeksi terbanyak (Gambar 6.14), sedangkan provinsi-provinsi lain digabung dalam satu gambar dengan hanya menunjukkan garis trend (Gambar 6.15).

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.14. Grafik akumulasi di provinsi Daegu yang memiliki jumlah terinfeksi terbanyak.

Insights dari data: Di semua provinsi, menjelang akhir April jumlah penambahan terinfeksi sudah mendekati nol. Penyebaran berhasil ditangani dengan baik. Selain itu, trend kesembuhan juga bagus, meningkat cepat dari Maret sampai akhir April.



Gambar 6.15. Trend akumulasi terkonfirmasi, sembuh dan meninggal di 6 provinsi terbanyak (selain Daegu).

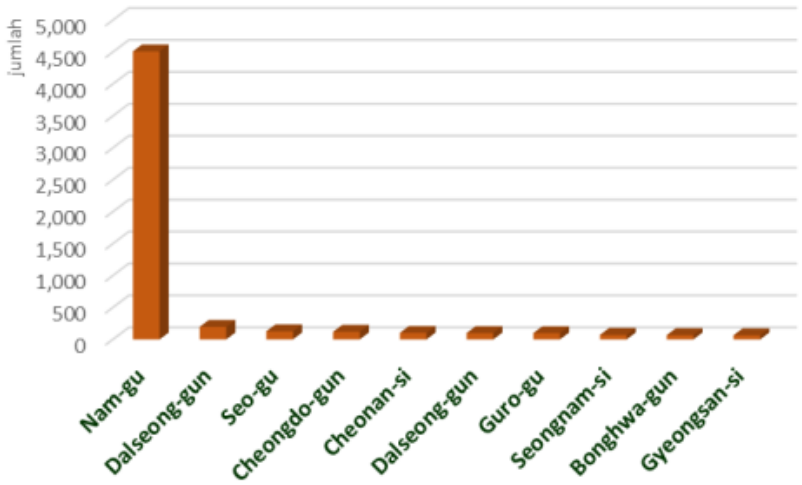
Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-10:

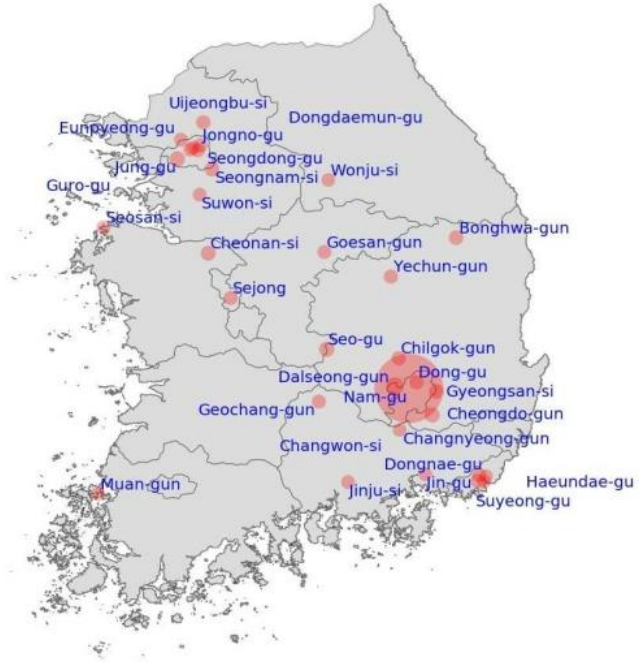
Bagaimana sebaran terinfeksi di kota-kota Korsel? Apakah terpusat di kotakota tertentu dan terdapat episentrum?

Untuk menjawabnya, data belum tersedia. Namun, jumlah terinfeksi di tiap kota dapat dihitung dari file Case.csv. Pada tiap kota, dilakukan penjumlahan (sum) dari kolom confirmed pada semua baris untuk kota tersebut. Komputasi dilakukan dengan melakukan group-by berdasar kota untuk menjumlah nilai kolom confirmed. Ini dapat dilakukan di Excel, dengan membuat program menggunakan library Pandas pada Python, atau SQL pada basisdata relasional.

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.16. Sepuluh kota dengan jumlah terinfeksi terbanyak di Korsel.



Gambar 6.17. Peta sebaran paparan COVID-19 di kota-kota Korsel (makin besar lingkaran, makin banyak yang terpapar).

Insights dari data: Penyebaran COVID-19 di Korsel hanya terjadi di beberapa kota dengan episentrum di Nam-gu, provinsi Daegu. Untuk provinsi dengan paparan terbanyak lainnya, hanya Seoul yang memiliki kota episentrum. Di Gyeonggi-do dan Gyeongsangbuk-do, kasus terbanyak berasal dari kota lain.

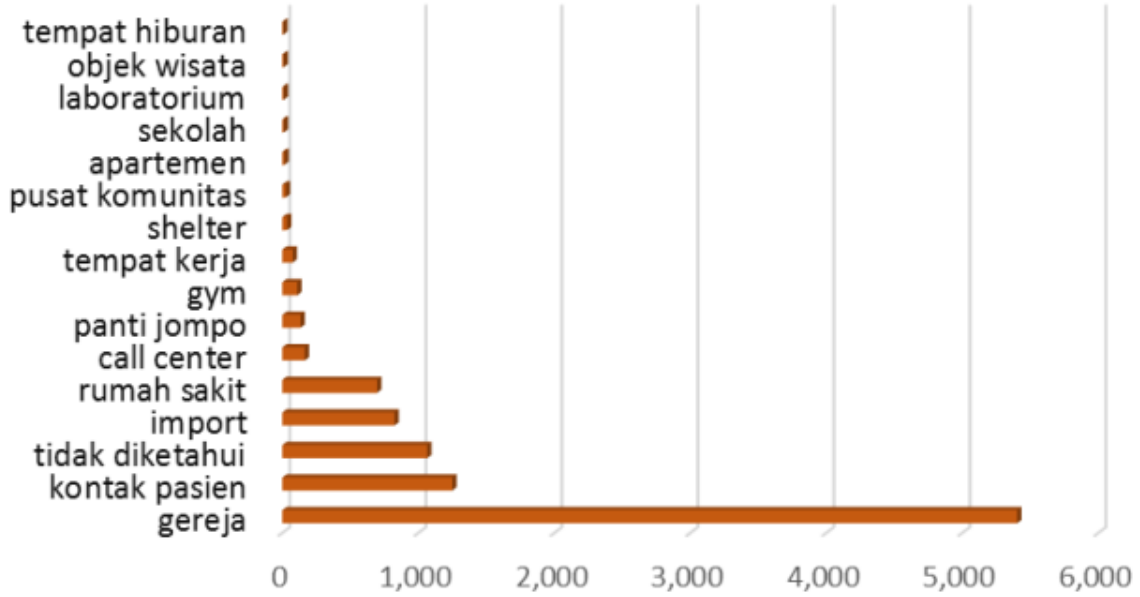
Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-11:

Bagaimana dengan asal paparan? Tempat-tempat mana saja yang paling banyak menjadi ajang penularan COVID-19?

Untuk menjawab, data belum tersedia namun dapat disiapkan dari file Case.csv dengan memanfaatkan kolom infection_case dan confirmed. Di sini perlu dibuat sebuah kolom baru, place_group, yang diisi dengan kategori tempat (sekolah, gereja, gym, dll.). Nilai kolom place_group ditentukan berdasar isi kolom infection_case. Perhitungan dengan group-by dilakukan untuk menjumlahkan nilai-nilai confirmed untuk tiap nilai di place_group. Hasilnya lalu diurutkan dari terbesar ke terkecil dan digunakan untuk membuat grafik bar horisontal pada Gambar 6.19.

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.19. Distribusi asal penularan COVID-19 di Korsel.

Insights dari data: Gereja dan rumah sakit merupakan tempat-tempat dimana mayoritas orang terpapar. Selain itu, orang dapat terpapar dari kontak dengan pasien dan berasal dari luar Korsel (import). Namun, terdapat lebih dari 1000 kasus yang tidak dapat diketahui darimana mereka tertular.

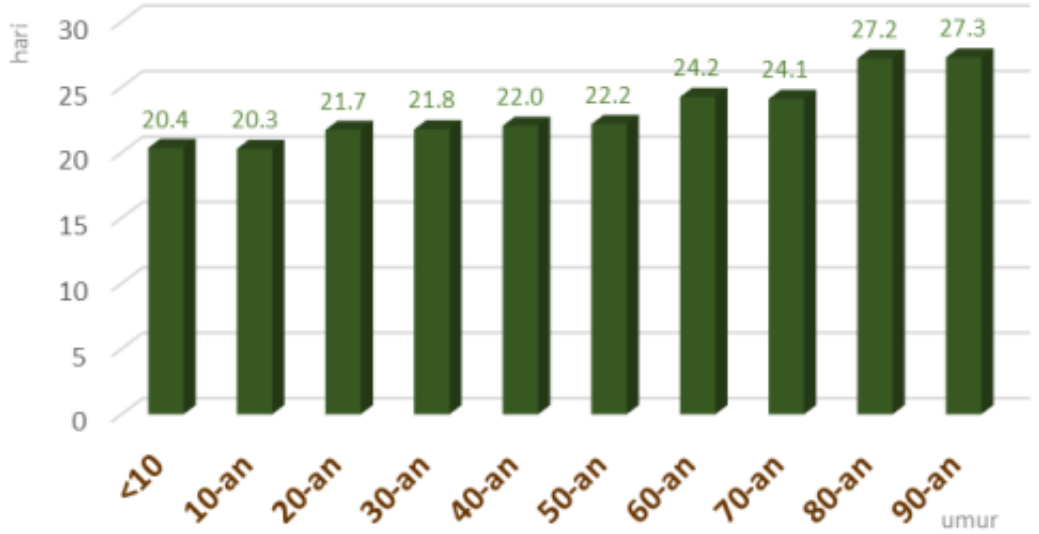
Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-12:

Berapa lama orang terinfeksi COVID-19 akan sembuh? Apakah umur berpengaruh terhadap lama sakit (dan dirawat di rumah sakit)?

Data belum tersedia, namun lama kesembuhan dapat dihitung dari file PatientInfo.csv (yang berisi data cukup detil dari 3.388 sampel pasien). Lama pasien sembuh dihitung dengan cara mengurangi nilai released_date dengan confirmed_date menggunakan Excel. Setelah itu, dengan group-by dihitung rata-rata kesembuhan tiap kelompok umur. Hasilnya digunakan untuk membuat grafik bar horisotal pada Gambar 6.20.

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.20. Rata-rata lama sembuh berdasar umur.

Insights dari data: Rata-rata lama pasien sembuh lebih dari 20 hari dan secara umum naik berdasar umur. Peningkatan secara signifikan terjadi mulai umur 60.

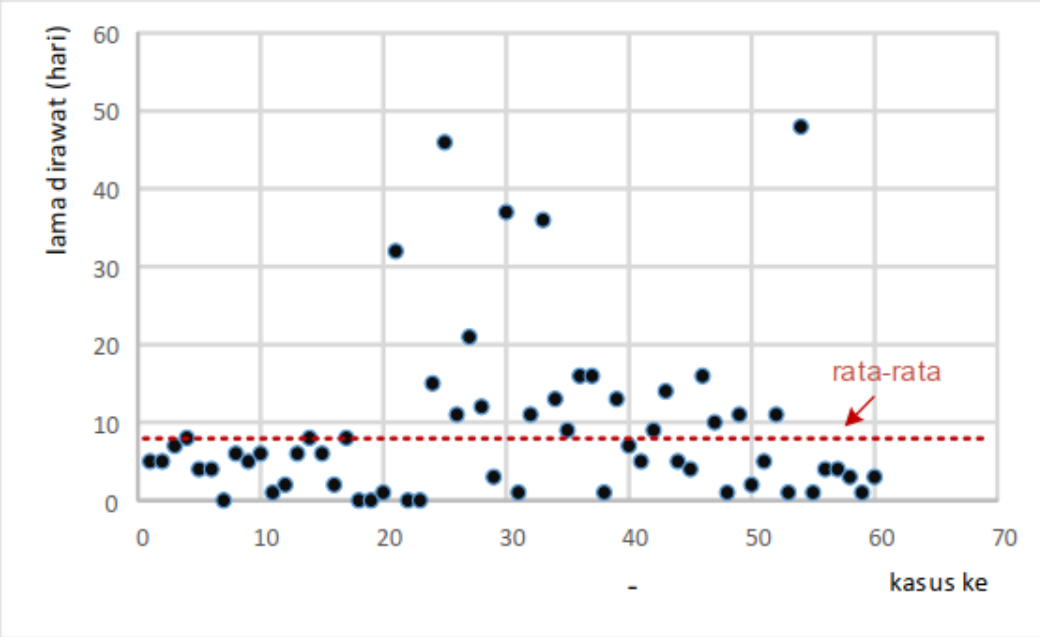
Case Data Visualization: Covid-19 Korea Selatan

Pertanyaan-13:

Untuk pasien yang meninggal, berapa lama pasien dirawat?

Data belum tersedia, namun lama kesembuhan dapat dihitung dari file PatientInfo.csv. Sebagaimana ditunjukkan pada Gambar 6.6, jumlah pasien meninggal di Korea relatif rendah. Kasus-kasus pada PatientInfo.csv harus dipilih dulu untuk mendapatkan kasus-kasus meninggal. Pemilihan dilakukan dengan filter dimana kolom state bernilai deceased (meninggal). Dari sini, hanya ditemukan 60 kasus. Kemudian, lama pasien dirawat (sampai meninggal) dihitung dengan mengurangi nilai deceased_date dengan confirmed_date. Setelah dilihat, ternyata jumlah hari pada 60 kasus bervariasi. Untuk menunjukkan variasi tersebut dibuat visualisasi dengan menggunakan scatter-plot pada tiap kasus (Gambar 6.21).

Case Data Visualization: Covid-19 Korea Selatan



Gambar 6.21. Distribusi lama pasien dirawat untuk 60 pasien yang meninggal.

Insights dari data: Lama pasien dirawat sebelum meninggal bervariasi, terbanyak berada di rentang 2 sampai 11 hari, dengan median (nilai tengah) 5.5 hari. Angka 0 (nol) mengindikasikan bahwa kasus tersebut terkonfirmasi pada tanggal yang bersamaan dengan terkonfirmasi terinfeksi.

Terima Kasih