Laporan Analisis Covid-19 Berdasarkan Gender dan Usia

Disusun oleh:

Mokhamad Arfan Wicaksono



Magister Informatika

Institut Teknologi Bandung

2020

# Latar Belakang

Beberapa latar belakang yang mengarahkan kepada pelaksanaan eksplorasi ini dijelaskan pada poin-poin berikut:

* Kasus COVID-19 saat ini menjadi salah satu perhatian penting pada kalangan masyarakat di dunia karena peningkatan kasus yang terus bertambah seiring berjalannya waktu.
* COVID-19 telah terjangkit oleh berbagai kalangan pasien dengan macam-macam usia dan gender.
* Saat ini pemahaman atas karakteristik pada kasus COVID-19 menjadi penting dalam rangka menetapkan kebijakan sebagai upaya menyelesaikan pandemik ini.

# Problem Statement

Rumusan masalah dari eksplorasi yang dilakukan adalah sebagai berikut:

* Bagaimana pengaruh tingkat kesembuhan dan kematian pada kasus COVID-19 berdasarkan usia dan gender
* Bagaimana pengaruh peningkatan kasus COVID-19 berdasarkan usia dan gender
* Bagaimana prediksi peningkatan kasus COVID-19 berdasarkan usia dan gender

# Dataset

Beberapa dataset yang digunakan pada eksplorasi yang dilakukan adalah sebagai berikut:

1. novel-corona-virus-2019-dataset  
   <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>  
   **Day level information on covid-19 affected cases**

Data yang tersedia pada link tersebut terdiri atas beberapa tabel dengan format csv. Pada eksplorasi ini data yang digunakan adalah `COVID19\_open\_line\_list.csv`. Total data yang ada pada dataset tersebut adalah sebanyak 13479 dengan 32 kolom. Data tersebut sangat tidak bersih dengan banyaknya nilai `NaN` yang ada di dalam data dan juga inkonsistensi bentuk data dari setiap kolomnya sehingga pembersihan data perlu dilakukan. Beberapa langkah pembersihan data pada dataset ini adalah.

* Memilih kolom yang dianggap menarik
* Mengganti data `age` yang berbentuk range (seperti: 20-29) menjadi nilai median dari data tersebut.
* Melakukan penamaan ulang data `sex` sehingga didapat dua kategori yaitu female dan male
* Menghilangkan data dengan `age` dan `sex` yang bernilai NaN
* Melakukan pembersihan data tanggal
* Mengelompokan data `outcome` menjadi tiga kategori yaitu `died`, `discharged`, dan `isolated`
* Mengambah kolom `age\_group` yaitu usia dikelompokkan setiap 10 tahun

Setelah data dibersihkan, diperoleh data sebanyak 1176 dengan jumlah kolom sebanyak 5. Deskripsi dari data tersebut adalah sebagai berikut.

|  | **age** | **sex** | **date\_confirmation** | **outcome** | **age\_group** |
| --- | --- | --- | --- | --- | --- |
| **count** | 1176.000000 | 1176 | 1176 | 1176 | 1176 |
| **unique** | NaN | 2 | 45 | 3 | 10 |
| **top** | NaN | male | 2020-02-02 00:00:00 | confirmed | 40-49 |
| **freq** | NaN | 672 | 65 | 1099 | 238 |
| **first** | NaN | NaN | 2020-01-12 00:00:00 | NaN | NaN |
| **last** | NaN | NaN | 2020-02-28 00:00:00 | NaN | NaN |
| **mean** | 45.213506 | NaN | NaN | NaN | NaN |
| **std** | 17.502754 | NaN | NaN | NaN | NaN |
| **min** | 0.083330 | NaN | NaN | NaN | NaN |
| **25%** | 33.000000 | NaN | NaN | NaN | NaN |
| **50%** | 44.500000 | NaN | NaN | NaN | NaN |
| **75%** | 56.000000 | NaN | NaN | NaN | NaN |
| **max** | 96.000000 | NaN | NaN | NaN | NaN |

1. coronavirusdataset-korea  
   <https://www.kaggle.com/kimjihoo/coronavirusdataset>  
   **DS4C: Data Science for COVID-19 in South Korea**

Terdapat beberapa file dengan format csv pada dataset ini. Data yang digunakan untuk eksplorasi adalah `PatientInfo.csv` dengan jumlah baris sebanyak 2771 dan kolom sebanyak 17. Pada data ini, kolom `age` telah menunjukkan nilai umur dengan range setiap 10 tahun. Data ini sudah cukup bersih sehingga hanya dilakukan pengambilan kolom yang menarik sehingga didapat deskripsi data sebagai berikut.

|  | **sex** | **age** | **confirmed\_date** | **state** |
| --- | --- | --- | --- | --- |
| **count** | 2765 | 2758 | 2771 | 2771 |
| **unique** | 2 | 12 | 57 | 3 |
| **top** | female | 20s | 2020-03-05 00:00:00 | isolated |
| **freq** | 1547 | 651 | 145 | 1839 |
| **first** | NaN | NaN | 2020-01-20 00:00:00 | NaN |
| **last** | NaN | NaN | 2020-03-30 00:00:00 | NaN |

1. indonesia-coronavirus-cases  
   <https://www.kaggle.com/ardisragen/indonesia-coronavirus-cases>  
   **Data Science for COVID-19 Indonesia (DSCI) Initiative**

Data yang digunakan pada dataset tersebut adalah `patient.csv` dengan jumlah baris sebanyak 893 dan kolom sebanyak 10. Pada data ini terdapat banyak sekali nilai null pada usia dan gender. Data gender dan usia yang tersedia hanya sebanyak 156. Pengelompokan usia dilakukan dengan range setiap 10 tahu. Kemudian dilakukan pengambilan kolom yang menarik pada data tersebut sehingga diperoleh data yang menarik sebanyak 5 kolom dengan deskripsi sebagai berikut.

|  | **gender** | **age** | **current\_state** | **confirmed\_date** | **age\_group** |
| --- | --- | --- | --- | --- | --- |
| **count** | 156 | 156.000000 | 166 | 893 | 156 |
| **unique** | 2 | NaN | 3 | 20 | 9 |
| **top** | male | NaN | isolated | 2020-03-24 00:00:00 | 50-59 |
| **freq** | 91 | NaN | 143 | 106 | 38 |
| **first** | NaN | NaN | NaN | 2020-03-02 00:00:00 | NaN |
| **last** | NaN | NaN | NaN | 2020-03-26 00:00:00 | NaN |
| **mean** | NaN | 48.108974 | NaN | NaN | NaN |
| **std** | NaN | 16.627833 | NaN | NaN | NaN |
| **min** | NaN | 2.000000 | NaN | NaN | NaN |
| **25%** | NaN | 36.000000 | NaN | NaN | NaN |
| **50%** | NaN | 49.000000 | NaN | NaN | NaN |
| **75%** | NaN | 59.000000 | NaN | NaN | NaN |
| **max** | NaN | 86.000000 | NaN | NaN | NaN |

1. covid19-in-india  
   <https://www.kaggle.com/sudalairajkumar/covid19-in-india>  
   **Dataset on Novel Corona Virus Disease 2019 in India**

Dataset yang digunakan pada sumber tersebut adalah `IndividualDetails.csv`. Jumlah data yang ada pada dataset tersebut adalah sebanyak 3671 dengan kolom sebanyak 11. Dilakukan pembersihan data pada dataset tersebut yaitu pada kolom `age` karena terdapat inkonsistensi tipe data pada kolom tersebut. Pembersihan yang dilakukan adalah mengumbah data range usia (seperti: 10-19) menjadi nilai tengahnya. Kemudian diambil kolom yang dianggap menarik sehingga diperoleh data dengan deskripsi sebagai berikut.

|  | **diagnosed\_date** | **age** | **gender** | **current\_status** | **age\_group** |
| --- | --- | --- | --- | --- | --- |
| **count** | 3671 | 805.000000 | 1035 | 3671 | 805 |
| **unique** | 37 | NaN | 2 | 4 | 10 |
| **top** | 2020-04-04 00:00:00 | NaN | M | Hospitalized | 20-29 |
| **freq** | 563 | NaN | 782 | 3557 | 180 |
| **first** | 2020-01-30 00:00:00 | NaN | NaN | NaN | NaN |
| **last** | 2020-04-04 00:00:00 | NaN | NaN | NaN | NaN |
| **mean** | NaN | 40.777640 | NaN | NaN | NaN |
| **std** | NaN | 16.944165 | NaN | NaN | NaN |
| **min** | NaN | 1.000000 | NaN | NaN | NaN |
| **25%** | NaN | 27.000000 | NaN | NaN | NaN |
| **50%** | NaN | 39.000000 | NaN | NaN | NaN |
| **75%** | NaN | 54.000000 | NaN | NaN | NaN |
| **max** | NaN | 96.000000 | NaN | NaN | NaN |

# Metode dan Teknik Analisis

Analisis deskriptif dan prediktif atas kasus COVID-19 berdasarkan fitur usia dan gender.

Dianalisis melalui kerangka kerja analisis data yang terdiri dari:

* Pengambilan dan pembersihan data
* Eksplorasi dan visualisasi data (Eksplorasi umum dan menelusuri korelasi antar fitur)
* Training dan testing model prediktif (Model regresi logistic)

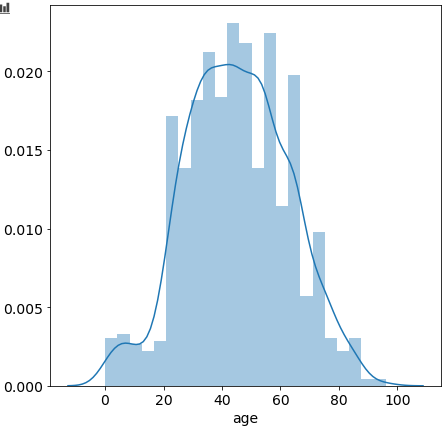
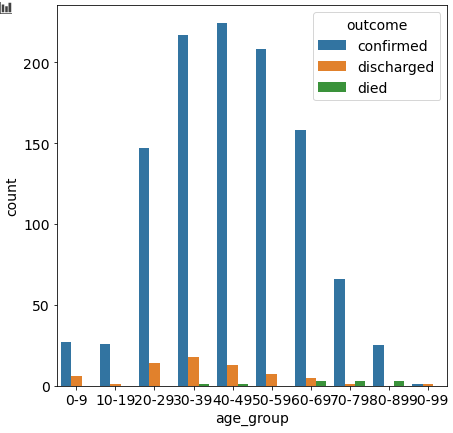
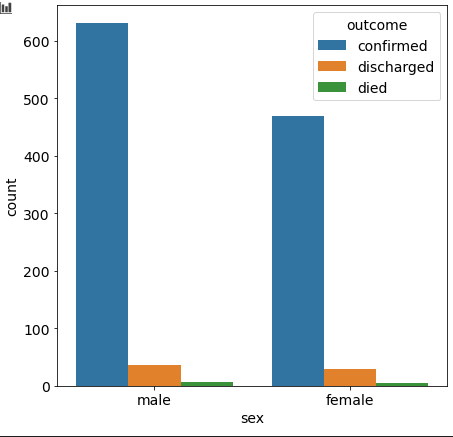
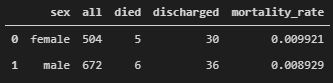
Secara teknis, analisis akan dilakukan pada bahasa pemrograman python.

# Pelaksanaan analisis

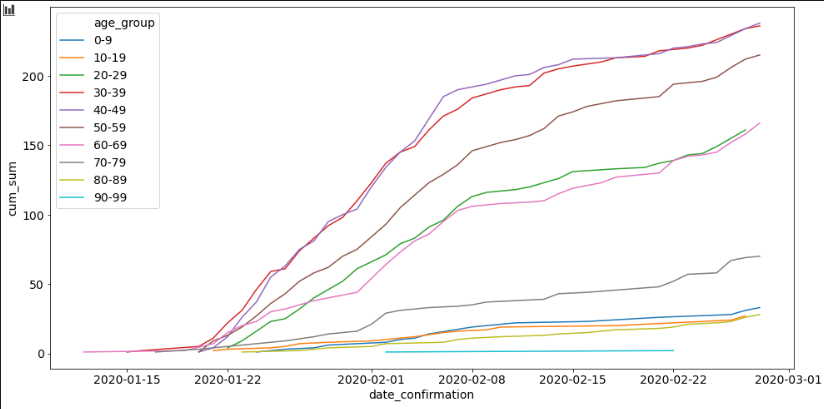
Dilakukan analisis terhadap setiap dataset.

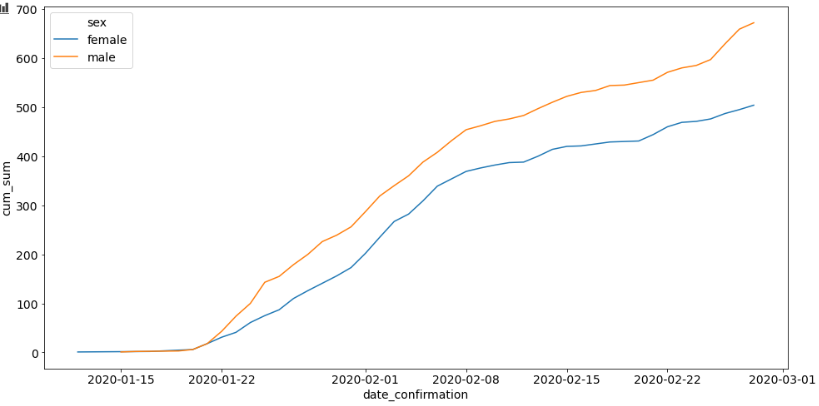
## novel-corona-virus-2019-dataset <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset> **Day level information on covid-19 affected cases**

### Distribusi data dan mortality rate

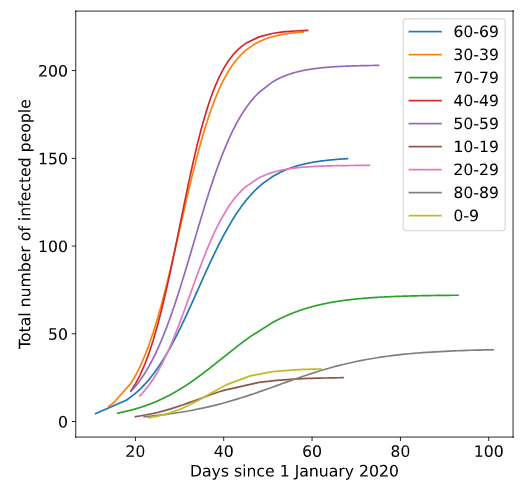
   

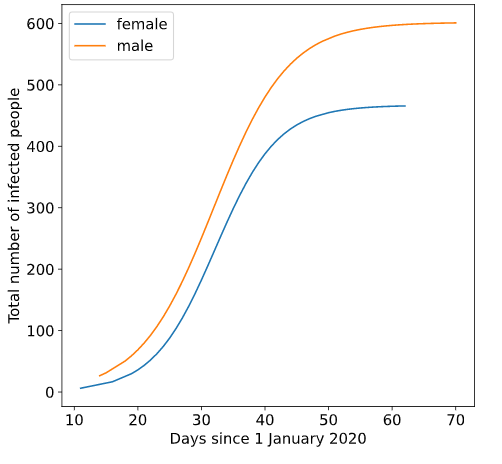
### Tren kasus berdasarkan usia dan gender





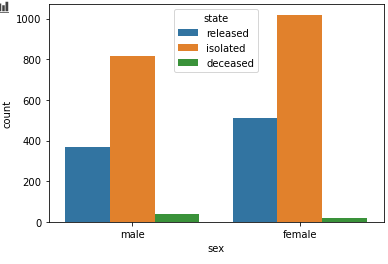
### Regresi logistik untuk analisa prediktif berdasarkan usia dan gender

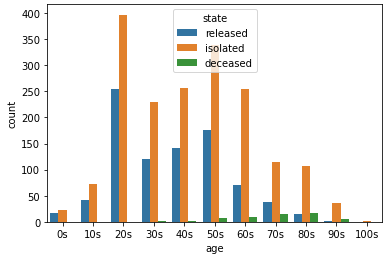
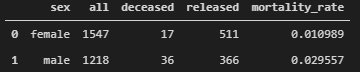


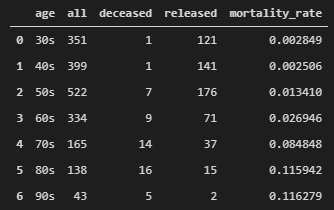


## coronavirusdataset-korea <https://www.kaggle.com/kimjihoo/coronavirusdataset> **DS4C: Data Science for COVID-19 in South Korea**

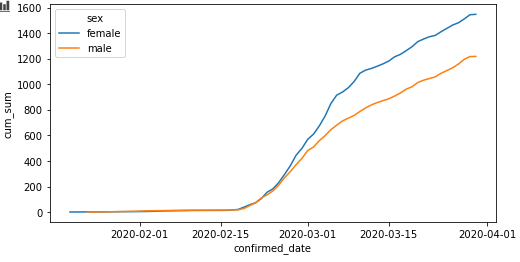
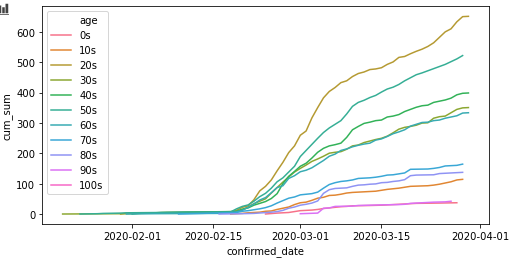
### Distribusi data dan mortality rate



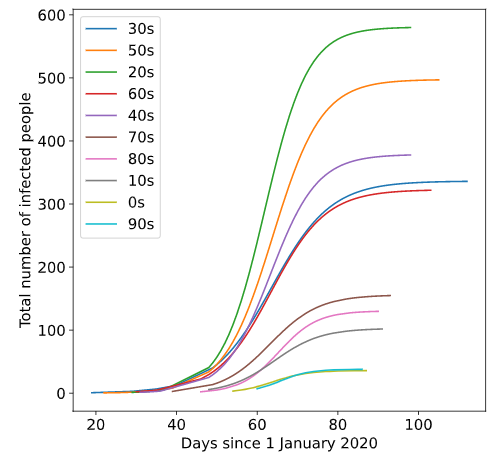


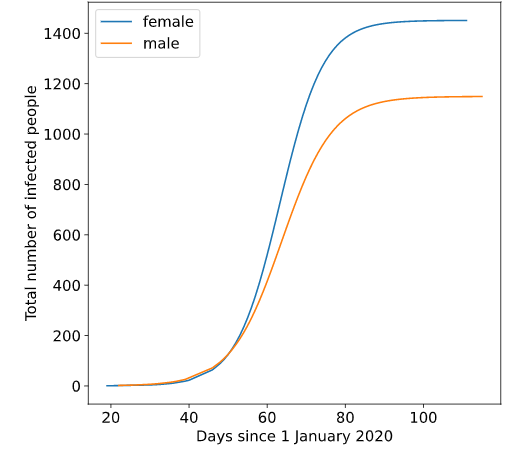


### Tren kasus berdasarkan usia dan gender



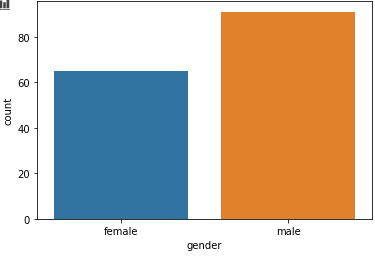
### Regresi logistik untuk analisa prediktif berdasarkan usia dan gender

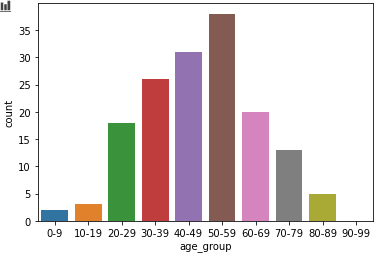


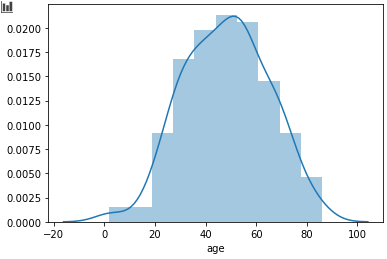


## indonesia-coronavirus-cases <https://www.kaggle.com/ardisragen/indonesia-coronavirus-cases> **Data Science for COVID-19 Indonesia (DSCI) Initiative**

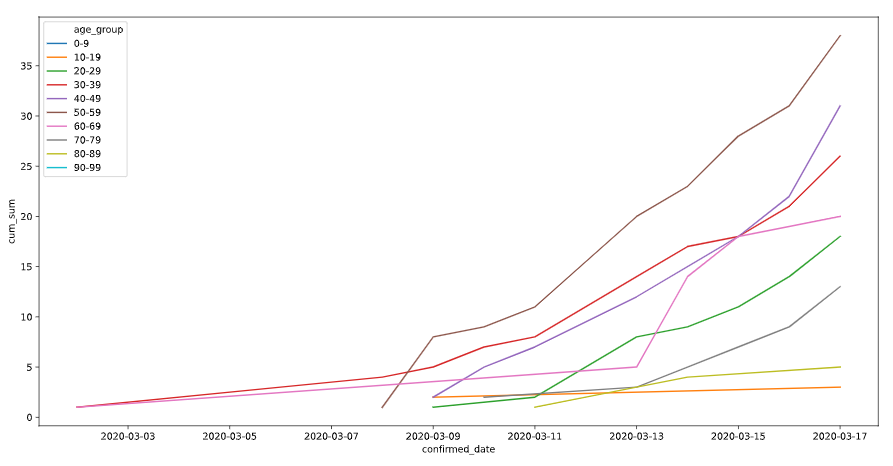
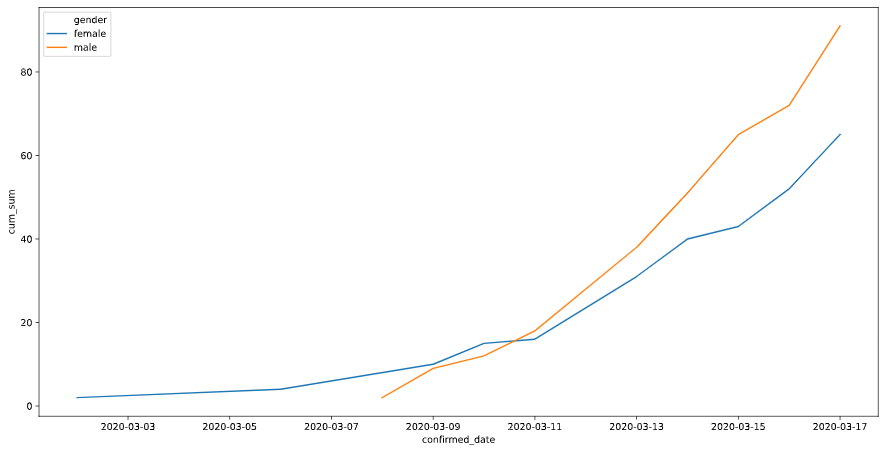
### Distribusi data





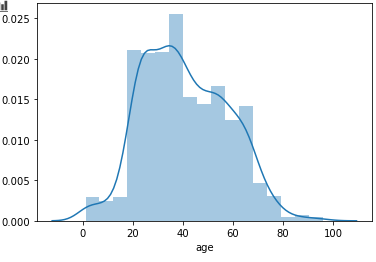


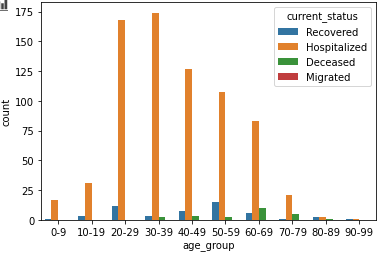
### Tren kasus berdasarkan usia dan gender

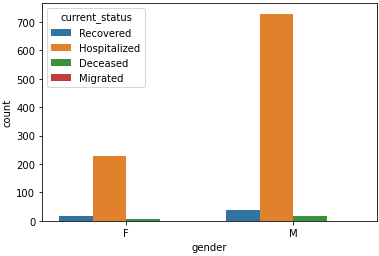
 

## covid19-in-india <https://www.kaggle.com/sudalairajkumar/covid19-in-india> **Dataset on Novel Corona Virus Disease 2019 in India**

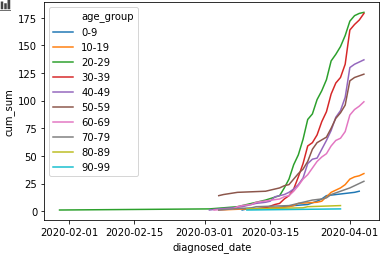
### Distribusi data

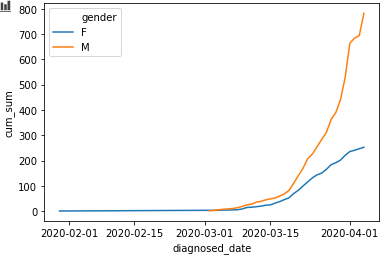






### Tren kasus berdasarkan usia dan gender





# Kesimpulan Analisis

(Seluruh rangkuman data dan pekerjaan dapat dilihat pada repositori sebagai berikut: <https://github.com/marfanwicaksono/covid19-exploration-analysis> , notebook eksplorasi data tersimpan pada setiap folder data pada repositori tersebut)

* Kasus kematian biasanya terjadi pada pengidap dengan usia tua dengan nilai mortality rate yang lebih besar daripada usia lainnya. Dalam hal ini perlu dianalisis lebih lanjut apakah ada lurking variable atau tidak (seperti penyakit kronis yang dimiliki pasien)
* Peningkatan kasus seringkali terlebih dahulu dialami pada pengidap usia menengah.
* Belum ditemukan hubungan yang signifikan antar kasus COVID-19 dengan gender.
* Dari hasil regresi logistik pada dataset 1, kemungkinan peningkatan kasus tertinggi terjadi pada usia menengah yaitu usia 30-49
* Dari hasil regresi logistik pada dataset 2, kemungkinan peningkatan kasus tertinggi terjadi pada usia 11-20 dan 49-50

# Diskusi

* Beberapa kendala yang ditemui pada analisis data adalah sebagai berikut
  + Terdapat data yang beragam dengan struktur data yang bermacam-macam
  + Seringkali ditemui data yang tidak lengkap, sehingga butuh waktu untuk data cleaning
  + Ketidaklengkapan data, ditambah data yang kurang banyak, menyebabkan penelusuran lebih lanjut harus dilakukan untuk melakukan penyimpulan
  + Kurangnya data sehingga pemodelan regresi tidak dapat dilakukan oleh beberapa dataset
* Hal-hal yang dapat dieksplorasi lebih lanjut
  + Eksplorasi terkait usia dapat dilakukan dengan mempertimbangkan berbagai range usia
  + Dapat dilakukan eksplorasi waktu pemulihan berdasarkan usia dan gender, hal ini juga dapat digunakan untuk melakukan pembersihan data terhadap waktu isolasi yang dianggap tidak logis