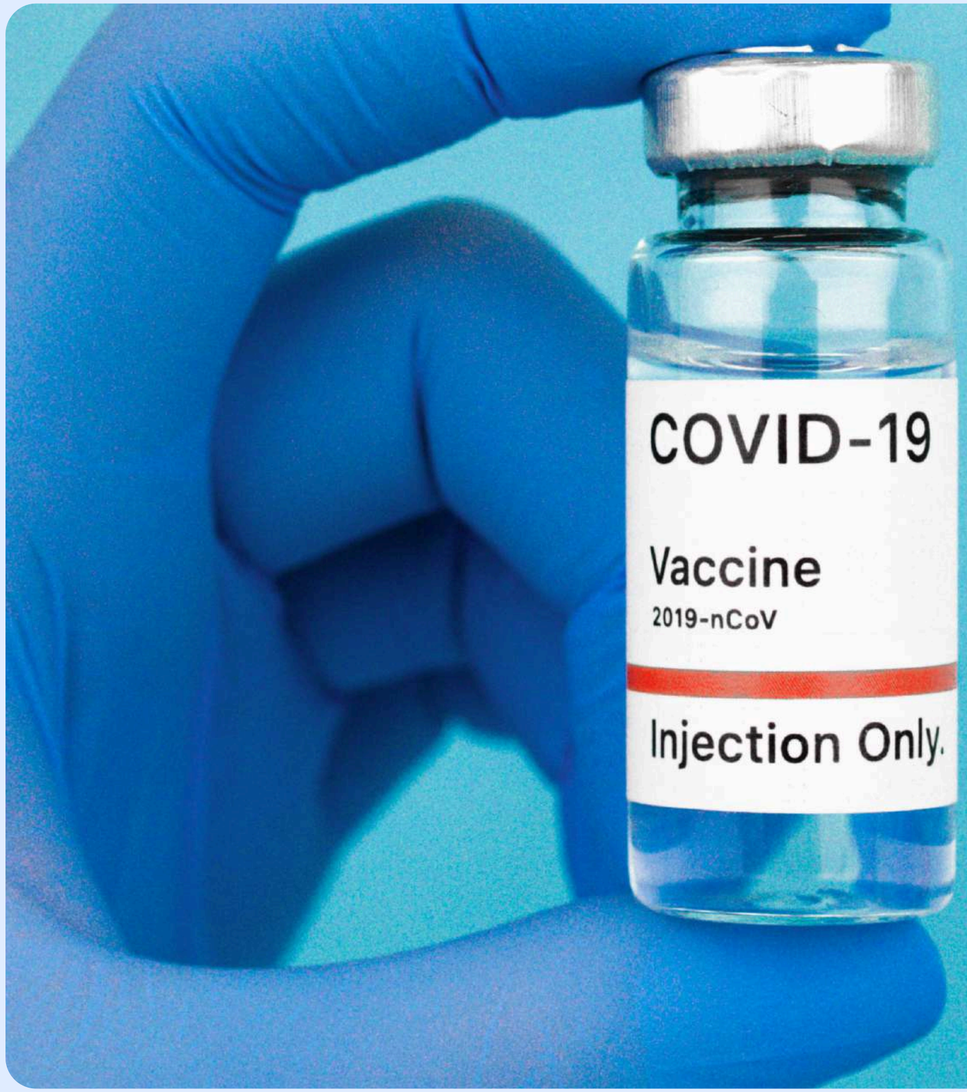
ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0



COVID -19

EXPLORATORY DATA ANALYSIS

HASBUL WAFI

Hello!

I'm Hasbul

Sebagai seorang data enthusiast, saya sedang mencoba menganalisis data pasien covid.
Hal ini akan mempermudah strategi penanganan di masa yang akan mendatang. Dalam proyek ini saya melakukan **Exploratory Data Analysis (EDA)** untuk memanipulasi data agar mudah dipahami strukturnya, diidentifikasi masalahnya serta mengvisualisasikan datanya.

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

INTRODUCTION

ANALISIS DATA PASIEN COVID-19 DIPERLUKAN UNTUK MEMAHAMI PENYEBARAN VIRUS DAN MENGENALI KELOMPOK YANG PALING RENTAN TERHADAP DAMPAK SERIUS.

BACKGROUND

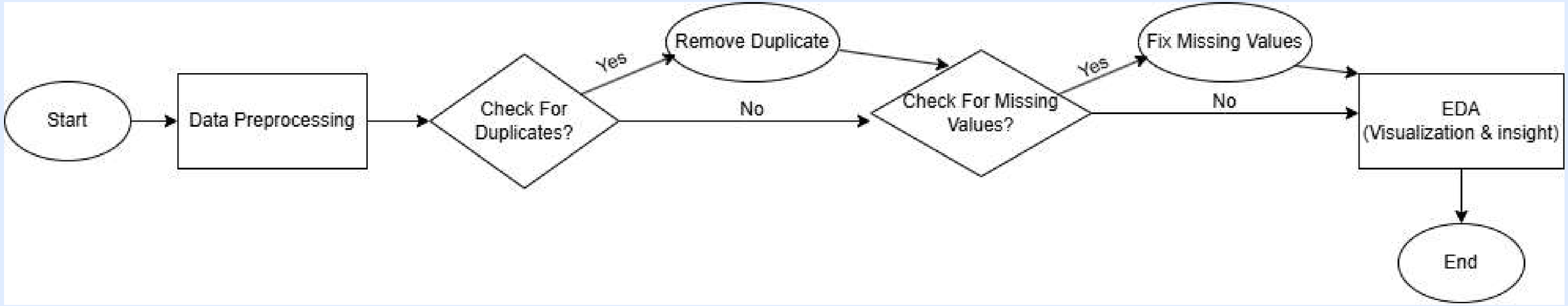
DATA YANG TIDAK LENGKAP DAN TIDAK TERSTRUKTUR MEMERLUKAN PEMBERSIHAN DAN VISUALISASI AGAR DAPAT MENGHASILKAN INSIGHT YANG AKURAT DAN BERMANFAAT.

PROBLEM STATEMENT

HASIL ANALISIS INI PENTING UNTUK MEMBANTU TENAGA MEDIS DAN PEMBUAT KEBIJAKAN MENGAMBIL KEPUTUSAN YANG TEPAT DALAM MENGHADAPI PANDEMI SAAT INI DAN DI MASA DEPAN.

URGENCY

FLOWCHART


Start
Data Preprocessing
Check For Duplicates?
Yes
Remove Duplicate
No
Check For Missing Values?
Yes
Fix Missing Values
No
EDA (Visualization & insight)
End

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

KEY STEPS

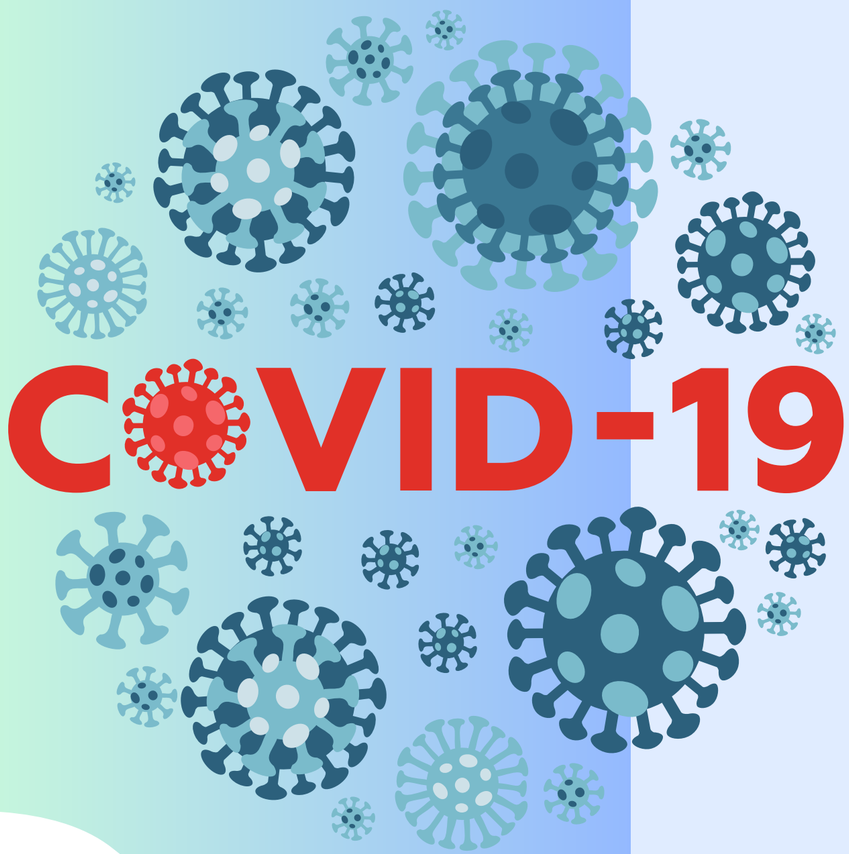
What Will i Do?

OBSERVATION DATA

CLEANSING DATA

DATA VISUALIZATION

ibimbing



PROJECT GOALS

01 **BERSIHKAN DAN PAHAMI STRUKTUR SET DATA COVID** SEBELUM DIPINDAHKAN KE ANALISIS LEBIH LANJUT.

02 **MELAKUKAN ANALISIS DATA EKSPLORASI (EDA)** UNTUK MENGEKSTRAK PANDANGAN AWAL DARI STATISTIK DESKRIPTIF, KHUSUSNYA PADA DEMOGRAFI PASIEN.

03 **PASTIKAN KUALITAS DATA** DENGAN MENANGANI DUPLIKASI MASUK DAN NILAI YANG HILANG SECARA TEPAT.

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

Data Overview

```
1  # Menampilkan 5 data teratas
2  data.head()
```

	name	sex	age	tested_positive	hospitalized	symptom_onset_date	comorbidity
0	Patricia Soto	male	34.0	1	yes	2021-04-04	none
1	Jessica Burke	female	33.0	1	yes	2021-09-19	none
2	Juan Martinez	male	26.0	1	yes	2021-09-29	asthma
3	Jessica Ward	male	47.0	0	no	2021-05-14	none
4	Kimberly Walker	male	45.0	1	no	2021-09-20	diabetes

```
1  # Menampilkan 5 data terbawah
2  data.tail()
```

	name	sex	age	tested_positive	hospitalized	symptom_onset_date	comorbidity
505	Darrell Wilson	male	15.0	0	no	2021-08-16	none
506	Heather Douglas	male	37.0	1	no	2022-03-15	none
507	Craig Woods	male	58.0	1	yes	2022-02-11	cancer
508	Patrick Barry	female	41.0	0	yes	2021-05-13	none
509	Alyssa Burke	female	31.0	0	no	2022-01-29	cancer

Dataset ini memuat informasi mengenai pasien COVID-19, termasuk atribut demografis, status kesehatan, dan hasil tes. Fitur utama yang dianalisis meliputi usia, jenis kelamin, status rawat inap, komorbiditas, serta hasil tes COVID-19. Variabel-variabel ini dipilih untuk mengeksplorasi bagaimana faktor kesehatan dan demografis individu dapat memengaruhi kemungkinan terinfeksi dan tingkat keparahan penyakit selama pandemi.

Total Rows 510

7 **Total Columns**

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

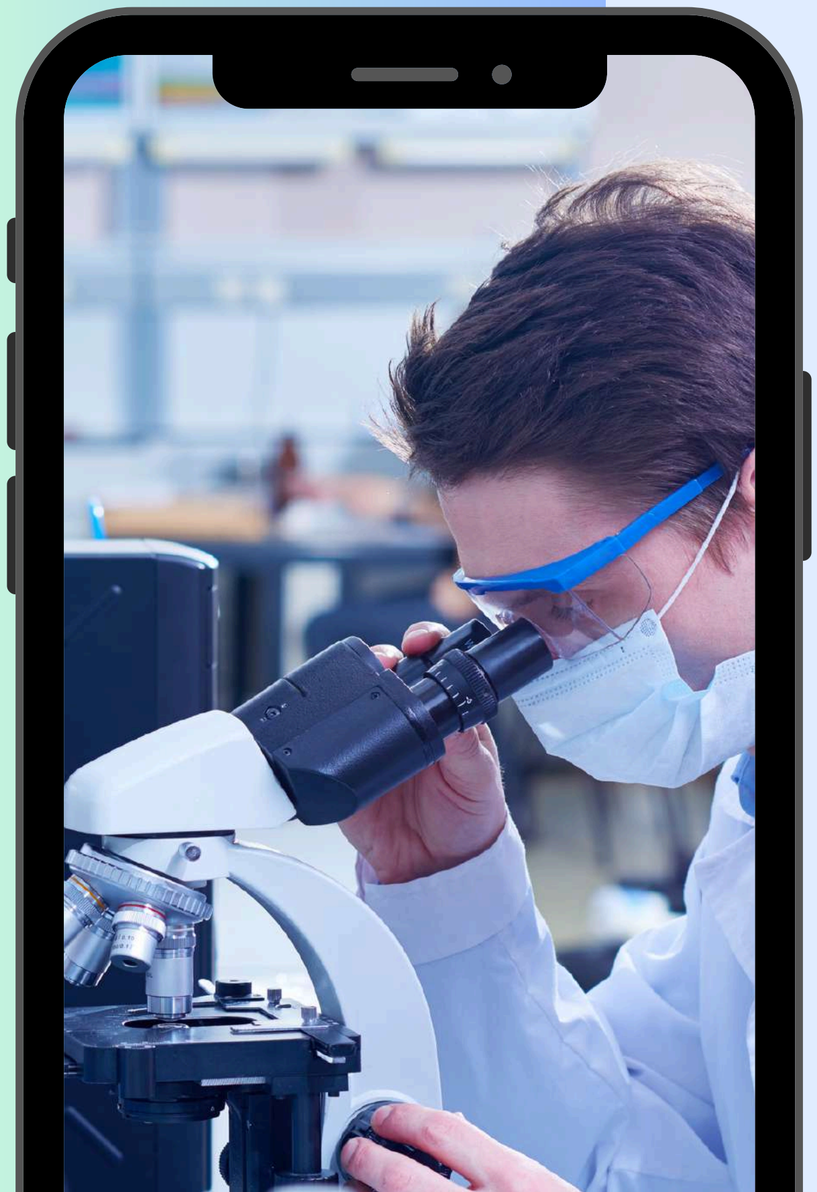
About Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 510 entries, 0 to 509
Data columns (total 7 columns):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  name            510 non-null    object
 1  sex             510 non-null    object
 2  age             489 non-null    float64
 3  tested_positive 510 non-null    int64
 4  hospitalized    510 non-null    object
 5  symptom_onset_date 510 non-null datetime64[ns]
 6  comorbidity     510 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(1), object(4)
memory usage: 28.0+ KB
None
```

1. **Jumlah Data:**
   - Dataset terdiri dari 510 entri (baris) dengan indeks dari 0 hingga 509.
2. **Kolom dan Tipe Data:**
   - Terdapat 7 kolom dengan rincian sebagai berikut:
     - name: 510 non-null, tipe object (string/teks).
     - sex: 510 non-null, tipe object.
     - age: 489 non-null, tipe float64 (terdapat missing values, karena hanya 489 dari 510 yang terisi).
     - tested_positive: 510 non-null, tipe int64 (bilangan bulat).
     - hospitalized: 510 non-null, tipe object.
     - symptom_onset_date: 510 non-null, tipe datetime64[ns] (tanggal dan waktu).
     - comorbidity: 510 non-null, tipe object.
3. **Missing Values:**
   - Hanya kolom age yang memiliki missing values, dengan 21 entri yang kosong dari total 510.
4. **Penggunaan Memori:**
   - Dataset menggunakan sekitar 28.0+ KB memori.
5. **Kesimpulan Umum:**
   - Dataset ini berisi informasi tentang individu (nama, jenis kelamin, usia, dll.) terkait dengan status kesehatan mereka (positif tes, rawat inap, tanggal gejala, komorbiditas).
   - Kolom age perlu diperhatikan karena memiliki missing values yang mungkin perlu diatasi sebelum analisis lebih lanjut.
   - Tipe data untuk setiap kolom sudah sesuai, termasuk kolom tanggal yang sudah dalam format datetime64[ns].

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0



01 Data Preprocessing

• Import Libraries

```
1  import numpy as np
2  import pandas as pd
3  import seaborn as sns
4  import matplotlib.pyplot as plt
5  pd.set_option("display.max_columns", None)
6  pd.set_option("display.max_rows", None)
```

- **numpy**: Library untuk operasi numerik pada array dan matriks.
- **pandas**: Library untuk manipulasi dan analisis data menggunakan struktur DataFrame dan Series.
- **seaborn**: Library visualisasi data statistik yang dibangun di atas matplotlib.
- **matplotlib.pyplot**: Modul untuk membuat visualisasi seperti grafik dan plot.
- **pd.set_option()**: Metode untuk mengatur tampilan Pandas, di sini digunakan untuk menampilkan semua kolom dan baris dalam DataFrame.

• Menampilkan 5 baris teratas & terbawah data

data.head()

```
# Menampilkan 5 data teratas
data.head()
```

	name	sex	age	tested_positive	hospitalized	symptom_onset_date	comorbidity
0	Patricia Soto	male	34.0	1	yes	2021-04-04	none
1	Jessica Burke	female	33.0	1	yes	2021-09-19	none
2	Juan Martinez	male	26.0	1	yes	2021-09-29	asthma
3	Jessica Ward	male	47.0	0	no	2021-05-14	none
4	Kimberly Walker	male	45.0	1	no	2021-09-20	diabetes

data.head() menunjukkan 5 baris pertama dari kumpulan data untuk memberikan gambaran umum data awal.

data.tail()

```
# Menampilkan 5 data terbawah
data.tail()
```

	name	sex	age	tested_positive	hospitalized	symptom_onset_date	comorbidity
505	Darrell Wilson	male	15.0	0	no	2021-08-16	none
506	Heather Douglas	male	37.0	1	no	2022-03-15	none
507	Craig Woods	male	58.0	1	yes	2022-02-11	cancer
508	Patrick Barry	female	41.0	0	yes	2021-05-13	none
509	Alyssa Burke	female	31.0	0	no	2022-01-29	cancer

data.tail() menunjukkan 5 baris terakhir dari kumpulan data untuk memahami data di akhir.

• Informasi tentang Data Covid

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 510 entries, 0 to 509
Data columns (total 7 columns):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  name            510 non-null    object
 1  sex             510 non-null    object
 2  age             489 non-null    float64
 3  tested_positive 510 non-null    int64
 4  hospitalized    510 non-null    object
 5  symptom_onset_date 510 non-null datetime64[ns]
 6  comorbidity     510 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(1), object(4)
memory usage: 28.0+ KB
None
```

Berdasarkan data, informasi berikut diamati:

- 510 entri data
- 7 kolom
- 4 kolom dengan tipe data objek (name, sex, hospitalized, comorbidity)
- 2 kolom dengan tipe data numerik (tested_positive sebagai int64 dan age sebagai float64)
- Juga terbukti bahwa kolom usia memiliki 21 nilai yang hilang (karena hanya ada 489 nilai bukan nol dari 510 entri)

• Mengatasi Duplikat

```
[581]   1   len(data)
        510
```

Hasil **len(data)** menunjukkan bahwa dataset berisi total 510 baris.

```
[582]   1   # Memeriksa apakah ada duplikasi
        2   print(df.duplicated().sum())
        8
```

Hasil **print(df.duplicated().sum())** menunjukkan bahwa dataset terdapat nilai duplikat sebanyak 8 baris.

```
    1   len(data.drop_duplicates())/ len(data)
0.984313725490196
```

Lebih jauh, rasio **len(data.drop_duplicates()) / len(data)** menghasilkan **nilai 0,984.**
Karena nilai ini kurang dari 1, maka hal ini mengonfirmasi keberadaan data duplikat dalam himpunan data.

• Mengatasi Duplikat

```
1  duplicates = data[data.duplicated(keep=False)]
2  duplicates = duplicates.sort_values(by=list(data.columns))
3  display(duplicates)
```

	name	sex	age	tested_positive	hospitalized	symptom_onset_date	comorbidity
73	Chelsea Kramer	female	NaN	0	no	2022-02-28	asthma
503	Chelsea Kramer	female	NaN	0	no	2022-02-28	asthma
124	Craig Woods	male	58.0	1	yes	2022-02-11	cancer
509	Craig Woods	male	58.0	1	yes	2022-02-11	cancer
394	Darrell Wilson	male	15.0	0	no	2021-08-16	none
507	Darrell Wilson	male	15.0	0	no	2021-08-16	none
377	Heather Douglas	male	37.0	1	no	2022-03-15	none
508	Heather Douglas	male	37.0	1	no	2022-03-15	none

Perintah ini mengembalikan semua baris duplikat, termasuk entri asli. Output menunjukkan bahwa ada **8 baris terindikasi duplikat** yang ditemukan dalam set data:

- **Baris 124** dan **baris 509**, keduanya berisi informasi yang sama: **Cralg Woods, male, 58, 1, yes, 2022-02-11, cancer**.
- **Baris 394** dan **baris 507**, keduanya berisi informasi yang sama: **Darrell Wilson, male, 15.0, 0, no, 2021-08-16, none.**

• Mengatasi Duplikat

```
[784]   1  # Menghapus baris duplikasi
        2  df = df.drop_duplicates()
```

```
        1  # Memastikan duplikasi telah dihapus
        2  print(df.duplicated().sum())

   0
```

```
1037]   1  len(df.drop_duplicates())/ len(data)

   1.0
```

Selanjutnya, entri duplikat dalam kumpulan data dihapus.

Lalu, perintah tersebut digunakan untuk memastikan apakah data duplikat telah dihapus.

Hasilnya menghasilkan **nilai 1.0**, yang menunjukkan tidak ada data duplikat yang tersisa setelah proses pembersihan. Oleh karena itu, dapat disimpulkan bahwa proses penanganan duplikat berhasil diselesaikan.

• Penanganan “Missing Value”

```
1 #Melihat Jumlah Value null untuk setiap kolom
2 data.isna().sum()
```

	0
name	0
sex	0
age	20
tested_positive	0
hospitalized	0
symptom_onset_date	0
comorbidity	0

dtype: int64

```
1 #Melihat Jumlah Value null untuk setiap kolom
2 data.isnull().sum()
```

	0
name	0
sex	0
age	20
tested_positive	0
hospitalized	0
symptom_onset_date	0
comorbidity	0

dtype: int64

Berdasarkan output:

- **Kolom name, sex, tested_positive, hospitalized, symptom, comorbidity** tidak memiliki nilai yang hilang (jumlah yang hilang = 0).
- **Kolom age** berisi 20 nilai yang hilang, yang harus ditangani selama tahap praproses data.

• Penanganan “Missing Value”

```
print(data['age'].dtype)
print(data['age'].median())
```

```
float64
41.0
```

```
[1076]  # Mengisi nilai yang hilang dengan median

for column in data.select_dtypes(include=['number']).columns:
    data[column] = data[column].fillna(data[column].median())
```

```
#Melihat Jumlah Value null untuk setiap kolom
data.isna().sum()
```

	0
name	0
sex	0
age	0
tested_positive	0
hospitalized	0
symptom_onset_date	0
comorbidity	0

dtype: int64

```
#Menampilkan info data setelah proses handling duplicates
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 502 entries, 0 to 501
Data columns (total 7 columns):
 #  Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0  name             502 non-null    object
 1  sex              502 non-null    object
 2  age              502 non-null    float64
 3  tested_positive  502 non-null    int64
 4  hospitalized     502 non-null    object
 5  symptom_onset_date 502 non-null  datetime64[ns]
 6  comorbidity      502 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(1), object(4)
memory usage: 31.4+ KB
```

Penanganan Data yang Hilang:

- Kolom dengan tipe data kategoris (objek) diisi dengan modus, yang merupakan nilai yang paling sering muncul di kolom tersebut.
- Kolom dengan tipe data numerik diisi dengan median, yang merupakan nilai tengah dari distribusi data.

Karena **kolom age** bersifat **numerik**, nilai yang hilang di kolom ini diganti dengan median, dan proses penghapusan nilai NULL telah berhasil diselesaikan!

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

02

Exploratory Data Analysis

• Statistical Summary

	age	tested_positive
count	489.000000	510.000000
mean	40.443763	0.607843
std	14.698939	0.488711
min	0.000000	0.000000
25%	31.000000	0.000000
50%	41.000000	1.000000
75%	50.000000	1.000000
max	78.000000	1.000000

Statistik Deskriptif (age, tested_positive)

Jumlah observasi (count):

- Kolom tested_positive memiliki 510 data (tidak ada data yang hilang).
- Kolom age hanya memiliki 489 data, yang berarti terdapat 21 data usia yang hilang.

Rata-rata (mean):

- Usia rata-rata ≈ 40,44 tahun.
- tested_positive = 0,607843 → sekitar 60,8% individu terdeteksi positif (dengan kode "1").

Distribusi usia:

- Usia minimum = 0 tahun; usia maksimum = 78 tahun.
- Median (50%) = 41 tahun;
- Kuartil pertama (25%) = 31 tahun;
- Kuartil ketiga (75%) = 50 tahun.
- Standar deviasi (std) usia ≈ 14,7 tahun, menunjukkan adanya variasi usia yang cukup besar di dalam data.

• Categorical Summary

	name	sex	hospitalized	comorbidity
count	510	510	510	510
unique	500	2	2	5
top	Michael Perez	female	no	cancer
freq	3	260	358	111

Statistik Kategorikal (name, sex, hospitalized, comorbidity)

Jumlah observasi (count):

Semua kolom (name, sex, hospitalized, dan comorbidity) memiliki 510 data, artinya tidak ada data yang hilang (missing).

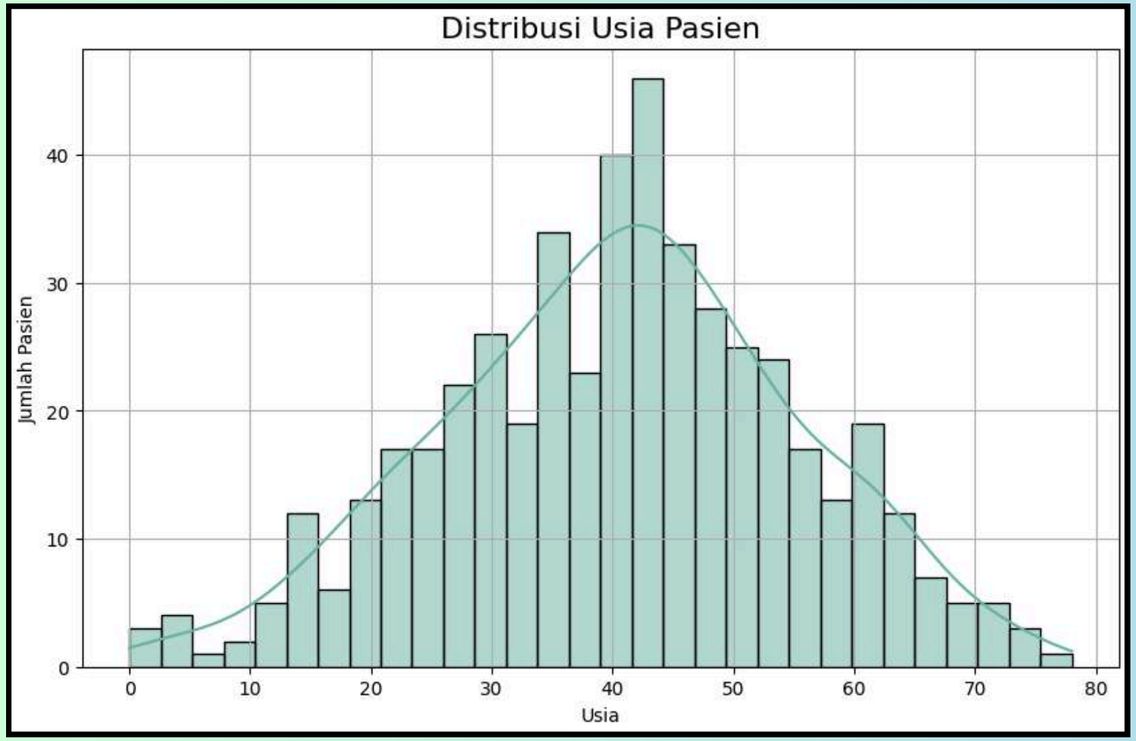
Jumlah nilai unik (unique):

- name memiliki 500 nilai unik, menunjukkan sebagian besar pasien memiliki nama yang berbeda.
- sex terdiri dari 2 kategori, yaitu female dan male.
- hospitalized memiliki 2 kategori, kemungkinan yes(1) dan no(0).
- comorbidity memiliki 5 kategori penyakit penyerta yang berbeda.

Nilai yang paling sering muncul (top) dan frekuensinya (freq):

- name: Nama paling umum adalah Michael Perez, muncul sebanyak 3 kali.
- sex: Kategori terbanyak adalah female, sebanyak 260 orang.
- hospitalized: Mayoritas pasien tidak dirawat di rumah sakit (no), sebanyak 358 orang.
- comorbidity: Komorbid paling umum adalah cancer, dengan 111 kasus.

• Distribusi Usia Pasien Covid

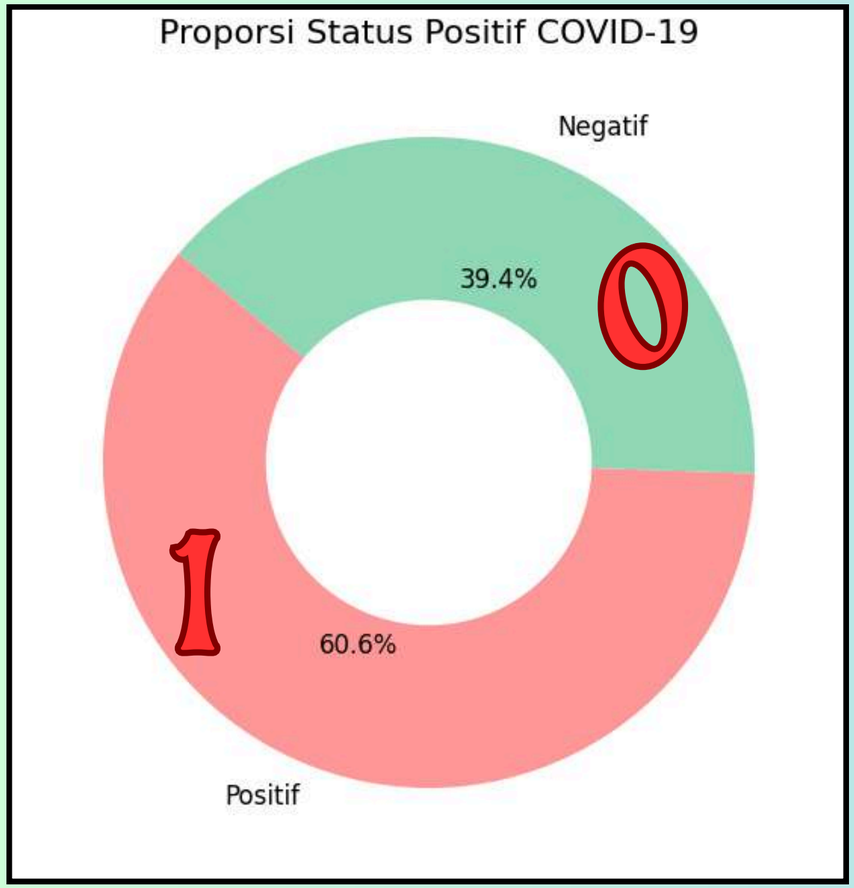


Dari diagram di atas, kita dapat melihat distribusi usia pasien dalam dataset ini:

- Sebagian besar pasien berada dalam rentang usia 30 hingga 50 tahun, dengan frekuensi tertinggi sekitar usia 40 tahun.
- Ada juga sejumlah pasien yang sangat muda, bahkan di bawah usia 10 tahun, meskipun jumlahnya lebih sedikit dibandingkan kelompok usia paruh baya.
- Frekuensi pasien menurun secara bertahap setelah usia 50 tahun, dan hanya sedikit pasien yang berusia di atas 70 tahun.
- Kurva distribusi tampak condong ke kanan, menunjukkan bahwa pasien berusia paruh baya lebih umum dibandingkan yang lebih tua.
- Distribusi ini mengindikasikan bahwa mayoritas pasien berada dalam usia produktif atau paruh baya.

Hal ini memberikan gambaran bahwa pasien dalam dataset ini didominasi oleh individu berusia produktif, dengan jumlah pasien yang lebih sedikit pada usia yang lebih tua.

• Proporsi Status Positif Pasien



Dari diagram yang ditampilkan, kita dapat melihat distribusi status COVID-19 dalam dataset ini:

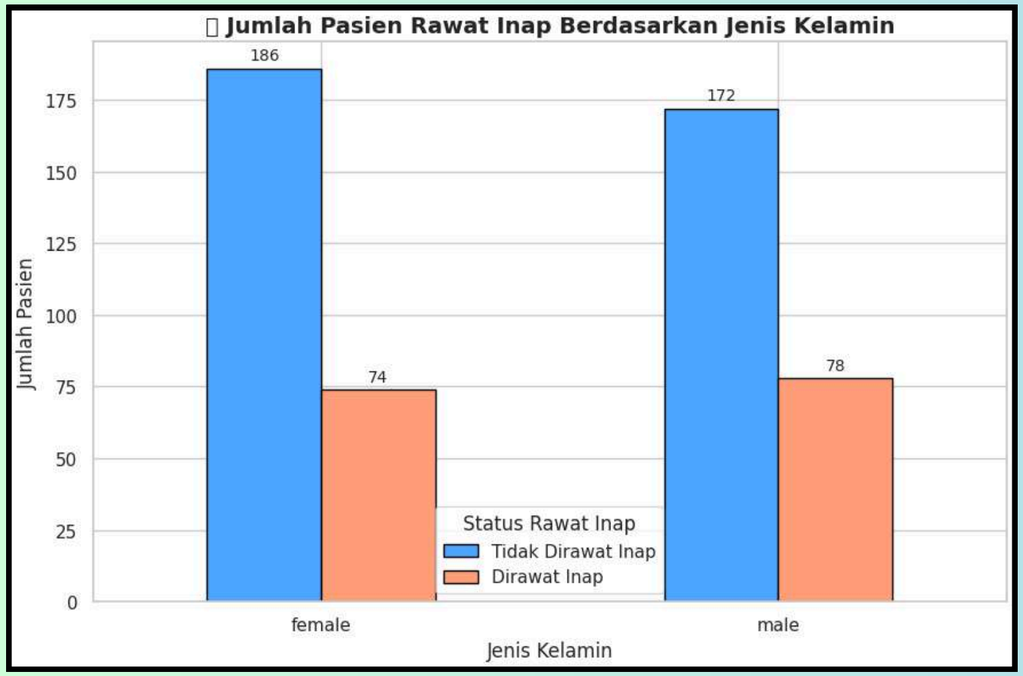
- Warna hijau merepresentasikan individu yang terdeteksi positif COVID-19.
- Warna merah muda merepresentasikan individu yang terdeteksi negatif COVID-19.

Dari diagram ini, kita dapat menyimpulkan bahwa:

- Jumlah individu yang terdeteksi positif COVID-19 (1) lebih tinggi dibandingkan dengan yang terdeteksi negatif (0).
- Persentase individu yang positif COVID-19 (1) adalah sekitar 60,6%, sedangkan yang negatif COVID-19 sekitar 39,4%.
- Mayoritas pasien dalam dataset ini terkonfirmasi positif COVID-19, yaitu sebanyak 310 orang (60,6%) dibandingkan 200 orang (39,4%) yang negatif.

Hal ini menunjukkan bahwa mayoritas individu dalam dataset terinfeksi, karena jumlah kasus positif yang signifikan dan menjadi perhatian.

• Jumlah Pasien Rawat Inap

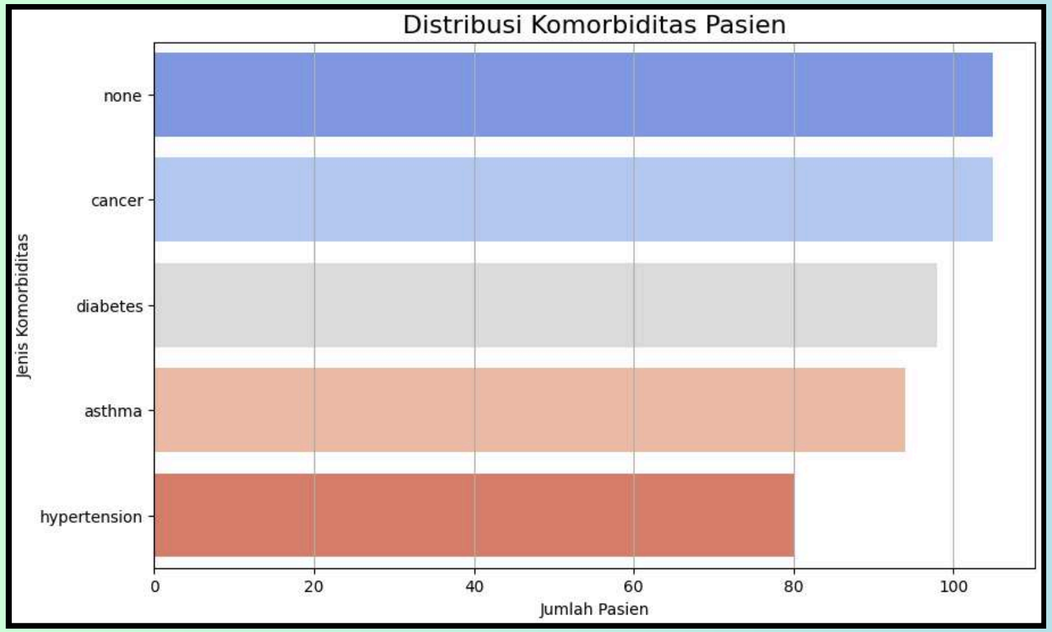


Berdasarkan diagram batang di atas, dapat disimpulkan bahwa:

- Mayoritas pasien, baik laki-laki maupun perempuan, tidak menjalani rawat inap.
- Namun, jumlah pasien **laki-laki** yang dirawat inap **(78 orang)** sedikit lebih tinggi dibandingkan **perempuan (74 orang).**

Meskipun jumlah pasien perempuan lebih banyak dalam dataset ini, proporsi laki-laki yang dirawat inap sedikit lebih tinggi, yang bisa menjadi indikasi bahwa laki-laki cenderung mengalami gejala yang lebih serius atau membutuhkan perawatan intensif lebih sering dibandingkan perempuan.

• Distribusi Komorbiditas Pasien

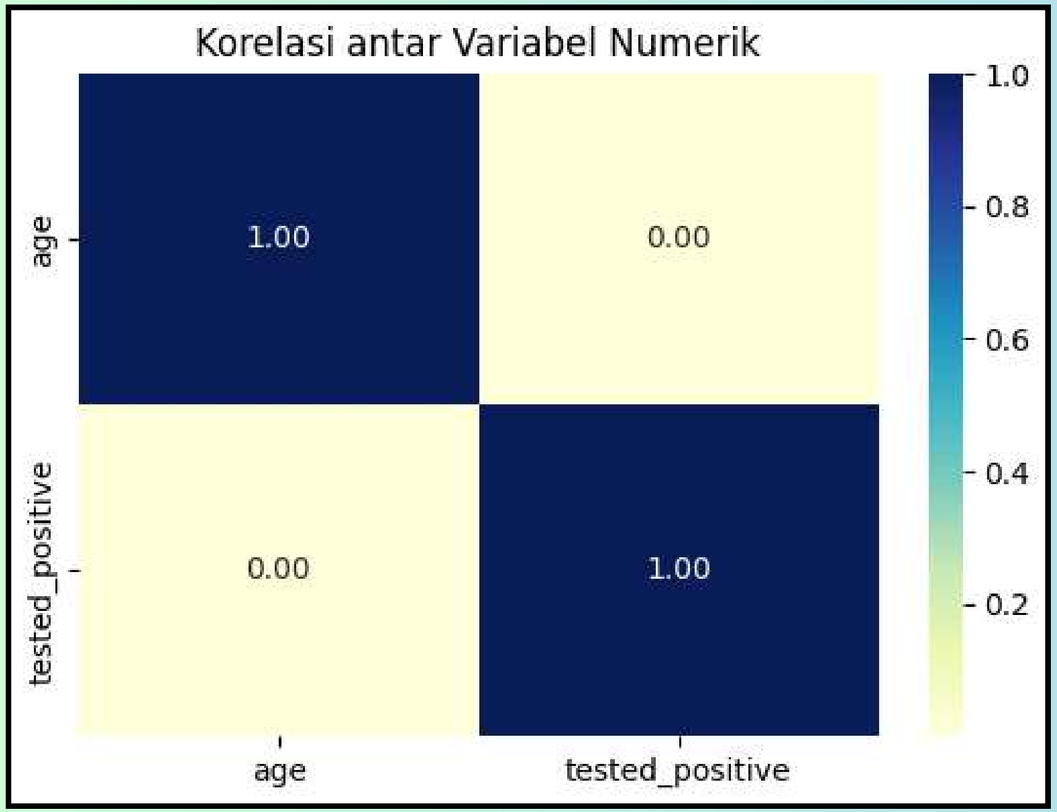


Dari diagram batang yang ditampilkan, kita dapat menyimpulkan bahwa:

- Jenis komorbiditas yang paling umum: Kanker dan pasien tanpa komorbiditas memiliki jumlah tertinggi, masing-masing sekitar 100 pasien.
- Distribusi penyakit lainnya: Pasien dengan diabetes dan asma berjumlah sekitar 90.
- Komorbiditas dengan jumlah lebih sedikit: hipertensi memiliki jumlah pasien paling sedikit, sekitar 80 orang.

Kesimpulan ini menunjukkan bahwa **kanker dan pasien tanpa komorbiditas mendominasi dataset**, sementara penyakit lain seperti hipertensi, diabetes, dan asma juga memiliki jumlah yang signifikan.

• Matriks Korelasi antar Variabel Numerik



Visualisasi di atas menunjukkan hubungan antar dua variabel numerik: age (usia) dan tested_positive (hasil tes COVID-19). **Hasil korelasi** adalah **0.00**, yang **artinya tidak ada hubungan linier antara usia pasien dan hasil tes positif COVID-19.**

Usia pasien bukan faktor penentu apakah seseorang akan terinfeksi COVID-19 atau tidak berdasarkan data ini. Baik pasien muda maupun tua memiliki peluang positif COVID-19 yang serupa.

Meskipun tidak berkorelasi secara linier, ini tidak berarti usia tidak relevan secara klinis. Faktor risiko mungkin muncul dalam bentuk hubungan non-linear atau berinteraksi dengan variabel lain seperti komorbiditas.

KESIMPULAN

Pembersihan Data

- 8 data duplikat dihapus (total data: 502).
- Nilai kosong di kolom usia diisi dengan median (41 tahun).

Temuan Utama

- Usia: Rata-rata 40 tahun, sebagian besar pasien berusia 30-50 tahun.
- Status COVID-19: 60,6% positif, 39,4% negatif.
- Rawat Inap: Mayoritas tidak dirawat (358 orang), tetapi lebih banyak pasien laki-laki yang dirawat.
- Komorbiditas: Kanker dan "tanpa komorbiditas" paling umum.

Korelasi

Tidak ada korelasi linier antara usia (age) dan hasil tes positif COVID-19 (tested_positive), menunjukkan bahwa usia bukan faktor penentu infeksi dalam dataset ini.

Implikasi

- Kesehatan Masyarakat: Kelompok usia produktif (30-50 tahun) mendominasi kasus, sehingga perlu menjadi fokus edukasi pencegahan.
- Kebijakan Medis: Tingkat rawat inap yang lebih tinggi pada laki-laki mungkin memerlukan penelitian lebih lanjut tentang faktor risiko gender.
- Kualitas Data: Kolom age perlu pemantauan ketat untuk menghindari missing values di masa depan.

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

GET MORE INFORMATION

View Repository on GitHub:

https://github.com/Sebul1306/DSF39-EDA-Covid

ibimbing

DATA SCIENCE | DIGITAL SKILL FAIR 39.0

Thank You :)

LET'S WORK TOGETHER!

@has.fie_

sebul1306

hasbulwafi1306@gmail.com