

# **LAPORAN PRAKTIKUM 1**

## **BIG DATA**



Disusun Oleh :

Wahyu Khairi

2311531009

Dosen Pengampu :

Luthfil Khairi S.Kom., M.Cs

**PROGRAM STUDI INFORMATIKA**  
**FAKULTAS TEKNOLOGI INFORMASI**  
**UNIVERSITAS ANDALAS**  
**2025/2026**

## A. Eksplorasi Dasar

1. Tampilkan 10 baris pertama dataset.

```
1. menampilkan 10 baris pertama dataset
```

```
[11] import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df = sns.load_dataset("titanic")
df.head(10)
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

Mengimport library pandas, seaborn dan matplotlib untuk membaca dataset dan juga melakukan visualisasi. Lalu buat kode dibawah untuk menampilkan 10 barus dataset teratas.

```
df = sns.load_dataset("titanic")
df.head(10)
```

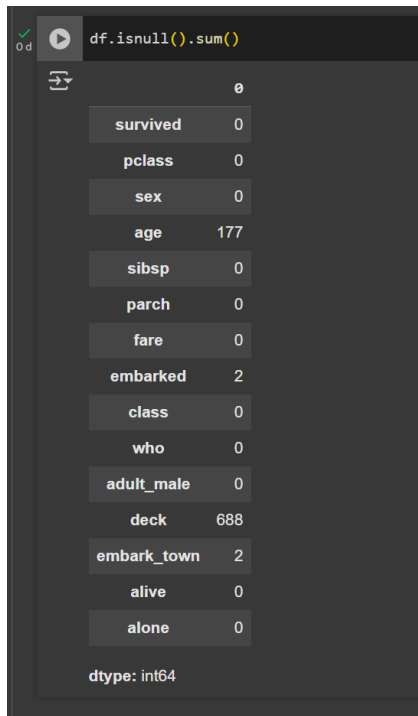
2. Hitung jumlah baris dan kolom pada dataset.

```
df.shape
```

```
(891, 15)
```

Dari dataset titanic ini terdapat 891 baris data dan ada 15 kolom fitur

3. Tampilkan jumlah missing values pada setiap kolom.



```
df.isnull().sum()
survived    0
pclass      0
sex         0
age        177
sibsp       0
parch       0
fare        0
embarked    2
class       0
who         0
adult_male  0
deck       688
embark_town 2
alive       0
alone       0
dtype: int64
```

Dari 15 kolom fitur terdapat data kosong (null) pada kolom age = 177 data, pada kolom embarked = 2, pada kolom deck = 688 data, dan pada kolom embark\_town = 2 data.

4. Deskripsikan secara singkat struktur dataset Titanic (fitur-fitur penting).

Dataset Titanic mengandung berbagai fitur penting yang mencakup data mengenai penumpang, status kelangsungan hidup, dan rincian perjalanan. Beberapa atribut utama meliputi:

- survived: Indikator kelangsungan hidup (0 = Tidak, 1 = Ya)
- pclass: Kategori kelas penumpang (1, 2, atau 3)
- sex: Gender penumpang (laki-laki atau perempuan)
- age: Umur penumpang
- sibsp: Total saudara atau pasangan di atas kapal
- parch: Jumlah orang tua atau anak di atas kapal
- fare: Biaya yang dibayar
- embarked: Pelabuhan tempat keberangkatan
- class: Sama dengan pclass, namun dalam bentuk kategori
- who: Kategori usia (pria, wanita, anak)
- adult\_male: Menentukan apakah penumpang adalah pria dewasa

- deck: Tingkat kabin
- embark\_town: Kota tempat keberangkatan
- alive: Status kelangsungan hidup dalam bentuk teks (ya, tidak)
- alone: Menentukan apakah penumpang melakukan perjalanan sendiri

## B. Analisis Statistik Sederhana

1. Hitung rata-rata umur (age) penumpang berdasarkan jenis kelamin.

```
[49] average_age_by_sex = df.groupby('sex')['age'].mean()
      print("Rata-rata usia penumpang berdasarkan jenis kelamin:")
      print(average_age_by_sex)
```

```
↗ Rata-rata usia penumpang berdasarkan jenis kelamin:
sex
female    28.216730
male      30.505824
Name: age, dtype: float64
```

Dari seluruh data penumpang berdasarkan jenis kelamin (sex) dan juga umur (age) didapatkan rata rata umur nya yaitu sekitar

- Pria = 30 an tahun
- Wanita = 28 an tahun

2. Hitung persentase penumpang yang selamat (survived) berdasarkan kelas (class).

```
survive_class = df.groupby(['class', 'survived']).size().unstack()
persentase_survive_class = survive_class.apply(lambda x: x / x.sum() * 100, axis=1)
print("Persentase penumpang yang selamat berdasarkan kelas:")
print(persentase_survive_class)
```

```
↗ Persentase penumpang yang selamat berdasarkan kelas:
      Tidak Selamat (%)  Selamat (%)
class
First                37.04        62.96
Second               52.72        47.28
Third                75.76        24.24
```

Dari data penumpang selamat berdasarkan kelas didapatkan persentase seperti gambar diatas :

Class	Tidak selamat (%)	Selamat (%)
First	37.04%	62.96%
Second	52.72%	47.28%
third	75.76%	24.24%

3. Cari jumlah penumpang yang tidak diketahui umurnya, lalu isi dengan rata-rata umur.

```
jumlah_usia_tidak_diketahui = df['age'].isnull().sum()
print(f"Jumlah penumpang dengan usia tidak diketahui: {jumlah_usia_tidak_diketahui}")

mean_age = df['age'].mean()
df['age'].fillna(mean_age, inplace=True)

print("\nJumlah missing values setelah pengisian:")
print(df['age'].isnull().sum())
```

Jumlah penumpang dengan usia tidak diketahui: 177

Jumlah missing values setelah pengisian:  
0

Cari data jumlah penumpang yang usianya tidak diketahui (null) dengan `df['age'].isnull().sum()` dan didapatkan datanya sejumlah 177 data kosong yaitu usia yang tidak diketahui.

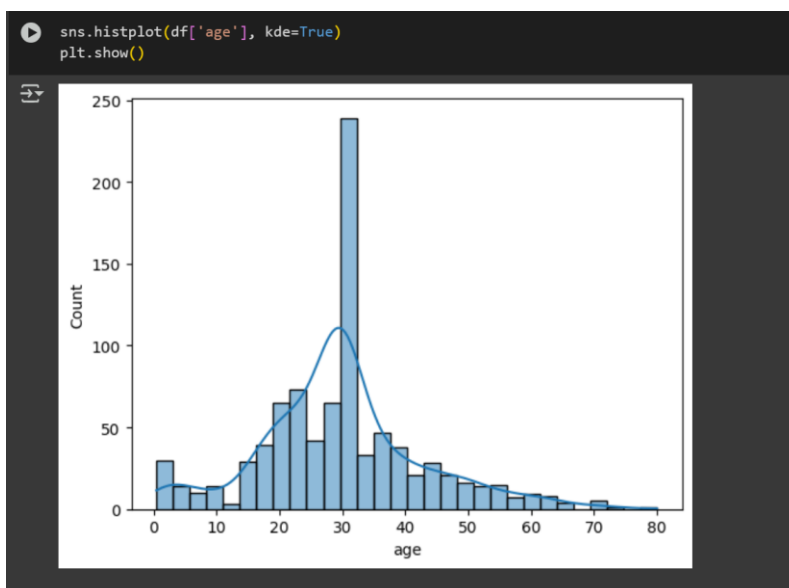
Lalu buat kode :

```
mean_age = df['age'].mean()
df['age'].fillna(mean_age, inplace=True)
```

kode ini berguna untuk memasukkan data usia yang kosong atau yang tidak diketahui dengan mean (rata-rata) value di kolom usia sehingga missing value nya menjadi 0

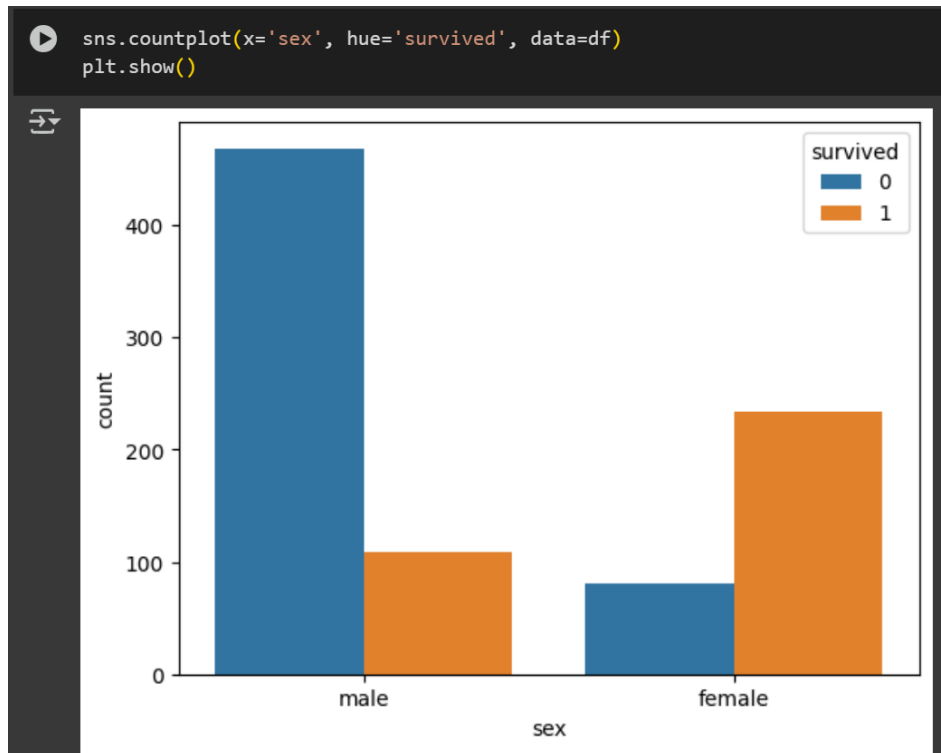
## C. Visualisasi Data

1. Buat histogram distribusi umur penumpang.



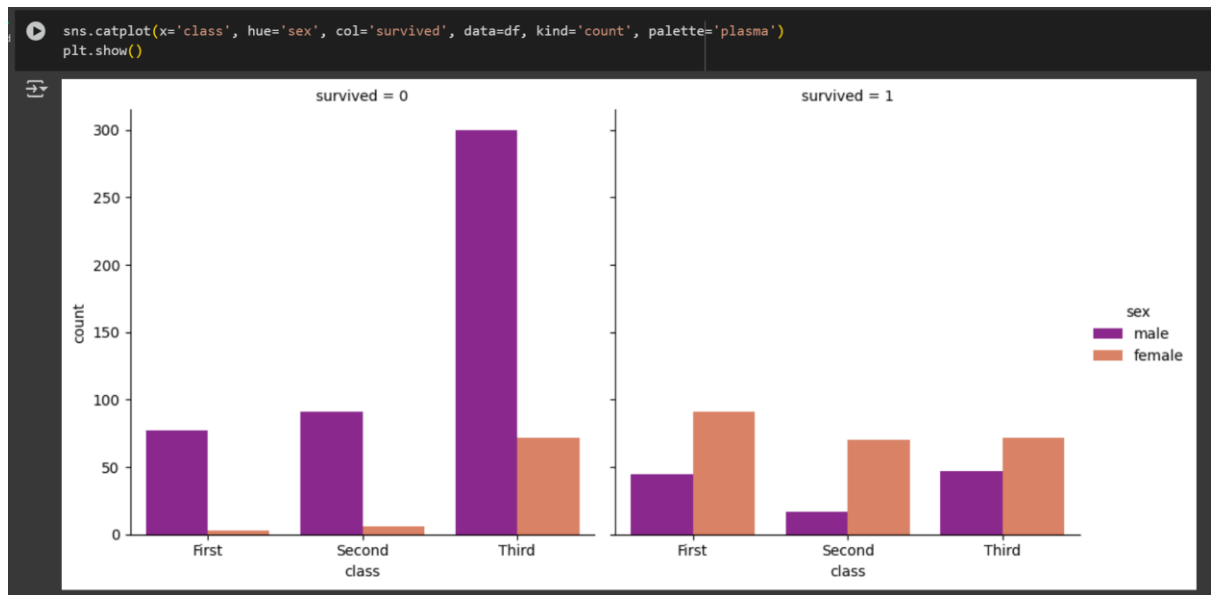
Dari visualisasi histogram ini dapat disimpulkan bahwa distribusi usia penumpang didominasi oleh usia antara 20 – 30 tahun.

2. Buat barplot survival berdasarkan jenis kelamin.



Dari visualisasi barplot ini dapat dilihat bahwa penumpang paling banyak tidak selamat yaitu pria dengan estimasi diatas 400 korban dan penumpang paling banyak selamat Adalah Wanita dengan estimasi diatas 200 nyawa.

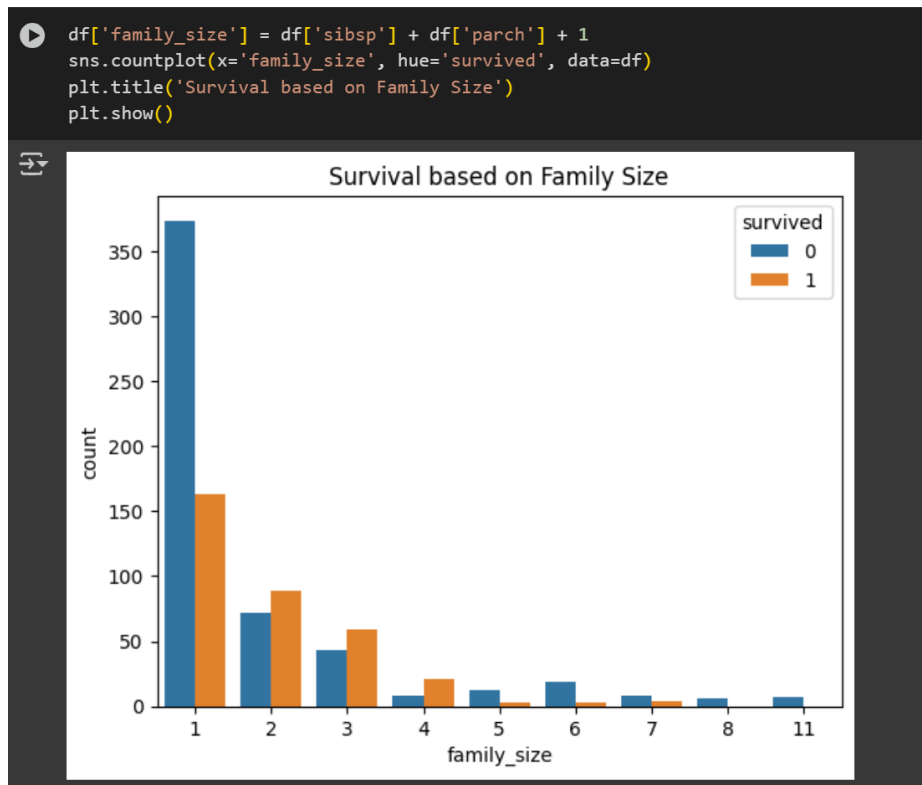
3. Buat barplot survival berdasarkan kelas dan jenis kelamin (gabungan / grouped barplot).



Dari visualiasai grup barplott ini dapat disimpulkan yaitu penumpang paling banyak tidak selamat Adalah pria dan Wanita dari kelas 3 disusul dengan penumpang kelas 2 dan 1

Dan penumpang yang selamat berada pada kelas 1 , disusul oleh penumpang kelas 3 dan 2

4. Buat visualisasi hubungan family\_size (dari sibsp + parch + 1) dengan survival.



kesimpulan dari visualisasi tersebut:

- Penumpang yang bepergian sendiri (family size = 1) jumlahnya paling banyak, tetapi mayoritas tidak selamat.
- Penumpang dengan keluarga kecil (2–4 orang) memiliki peluang selamat lebih tinggi, terlihat dari bar oranye yang relatif lebih besar dibanding bar biru.
- Penumpang dengan keluarga besar (>4 orang) jarang jumlahnya, dan mayoritas tidak selamat.

## D. Insight

1. Pengaruh Kelas Penumpang terhadap Kelangsungan Hidup: Penumpang di kelas yang lebih tinggi (Kelas 1) memiliki tingkat kelangsungan hidup yang jauh lebih tinggi dibandingkan dengan penumpang di kelas yang lebih rendah (Kelas 3). Ini terlihat jelas dari barplot survival berdasarkan kelas.
2. Perbedaan Tingkat Kelangsungan Hidup antara Jenis Kelamin: Wanita memiliki kemungkinan yang jauh lebih tinggi untuk selamat dibandingkan dengan pria. Barplot survival berdasarkan jenis kelamin menunjukkan perbedaan yang signifikan ini.



3. Pengaruh Ukuran Keluarga terhadap Kelangsungan Hidup: Ukuran keluarga tampaknya memiliki pengaruh terhadap kelangsungan hidup. Penumpang yang bepergian sendirian atau dalam keluarga kecil (2-4 orang) memiliki tingkat kelangsungan hidup yang berbeda dibandingkan dengan penumpang dalam keluarga besar.

Link source code tugas :

[https://github.com/WahyuKhairi06/BigData\\_2311531009\\_Wahyu-Khairi/blob/main/Praktikum%201/praktikum\\_bigdata\\_pertemuan1.ipynb](https://github.com/WahyuKhairi06/BigData_2311531009_Wahyu-Khairi/blob/main/Praktikum%201/praktikum_bigdata_pertemuan1.ipynb)