



HOME

ABOUT US

SERVICES

CONTACT



EXPLORATORY DATA ANALYSIS

START



OBSERVASI DATA

Dataset Titanic ini memuat empat kolom utama yang menggambarkan karakteristik dasar tiap penumpang. Kolom survived menunjukkan status keselamatan. Nilai 1 berarti penumpang selamat, sedangkan 0 artinya tidak selamat. Kolom name berisi nama lengkap penumpang, termasuk gelar sosialnya seperti "Mr.", "Miss.", atau "Master.". Kolom sex mencatat jenis kelamin penumpang, yakni male untuk laki-laki dan female untuk perempuan. Terakhir, kolom age menggambarkan usia penumpang dalam satuan tahun, dengan pecahan desimal untuk bayi dan beberapa nilai kosong jika data usia tidak tersedia. Keempat kolom ini sangat berguna untuk analisis awal, misalnya melihat pola keselamatan berdasarkan usia, gender, maupun status sosial penumpang.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column      Non-Null Count Dtype  
 --- 
 0   survived    500 non-null   int64  
 1   name        500 non-null   object  
 2   sex         500 non-null   object  
 3   age         451 non-null   float64 
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```

	survived	name	sex	age
19	0	Beattie, Mr. Thomson	male	36.0
186	1	Lindstrom, Mrs. Carl Johan (Sigrid Posse)	female	55.0
18	1	Bazzani, Miss. Albina	female	32.0



THE DATASET

Untuk dapat mendapatkan data dari dataset dan menganalisisnya, kita harus mengimportnya terlebih dahulu.

```
# import data
df = pd.read_excel('titanic.xlsx')
```

Setelah itu, kita akan menampilkan bagian 5 data dari bagian awal (head) dataset, dan dari bagian bawah (tail) dataset terlebih dahulu.

```
# buat nampilih 5 data pertama
df.head()
```

```
# buat nampilih 5 data terakhir
df.tail()
```

HEAD

	survived	name	sex	age
0	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
2	0	Allison, Miss. Helen Loraine	female	2.0000
3	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
4	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000

TAIL

	survived	name	sex	age
495	1	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.0
496	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN
497	0	Matthews, Mr. William John	male	30.0
498	0	Maybery, Mr. Frank Hubert	male	40.0
499	0	McCrae, Mr. Arthur Gordon	male	32.0



HOME

ABOUT US

SERVICES

CONTACT



THE DATASET

Itu tadi adalah cara mendapatkan data dari head dan tail. Sekarang, kita akan mendapatkan data dengan 2 cara lagi.

Kurang lebih sama, tetapi method yang akan kita gunakan adalah sample (data acak), dan info (keseluruhan dijadikan kesimpulan).

```
# nampilih 5 data secara acak
df.sample(5)
```

```
# kurang lebih kesimpulan
df.info()
```

SAMPLE

	survived	name	sex	age
223	0	Parr, Mr. William Henry Marsh	male	NaN
19	0	Beattie, Mr. Thomson	male	36.0
186	1	Lindstrom, Mrs. Carl Johan (Sigrid Posse)	female	55.0
18	1	Bazzani, Miss. Albina	female	32.0
308	1	White, Mrs. John Stuart (Ella Holmes)	female	55.0

INFO

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   survived    500 non-null    int64  
 1   name        500 non-null    object  
 2   sex         500 non-null    object  
 3   age         451 non-null    float64 
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```



	survived	age
count	500.000000	451.000000
mean	0.540000	35.917775
std	0.498897	14.766454
min	0.000000	0.666700
25%	0.000000	24.000000
50%	1.000000	35.000000
75%	1.000000	47.000000
max	1.000000	80.000000

STATISTICAL SUMMARY

Kita bisa mengerti dan paham isi dataset ini dengan menampilkan ringkasan semua kolom. Caranya, kita run code ini.

`df.describe()`

Dan jika dilihat di samping, kita bisa mengerti bahwa:

- Penumpang yang selamat ada sekitar 54% dari berdasarkan mean di kolom 'survived'. ($0.54 = 54\%$)
- Demografi umur penumpang untuk yang paling muda berumur 0.66 tahun atau sekitar 7 bulan. Sedangkan yang paling tua berumur 80 tahun.
- Rata-rata umur penumpang adalah 35.9 tahun, dan sebagian besar umurnya rentang di kurang lebih 14.7 tahun dari rata-rata.

Dari ringkasan ini, kita bisa paham mengenai data umur dan berapa persen yang survive kejadian Titanic ini.



DUPLICATE HANDLING

Kita dapat menemukan data mana yang double atau duplikat dengan menjalani kode ini

```
dups_all = df[df.duplicated(keep=False)]  
print(dups_all)
```

	survived		name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0	
349	1	Eustis, Miss. Elizabeth Mussey	female	54.0	

Jika dilihat, data Miss. Elizabeth Mussey ada data duplikatnya. Namun secara keseluruhan, dataset ini hanya mempunyai 1 data duplikat.

Terus cara handlingnya? Kita tinggal drop aja pake code ini.

```
df.drop_duplicates(inplace=True)
```

Jika sudah, maka data akan terhapus. Jika kita cek untuk data duplikat lagi, kita bisa lihat bahwa data duplikat tadi telah terhapus.

```
dups_all = df[df.duplicated(keep=False)]  
print(dups_all)
```

```
Empty DataFrame  
Columns: [survived, name, sex, age]  
Index: []
```



MISSING VALUE HANDLING

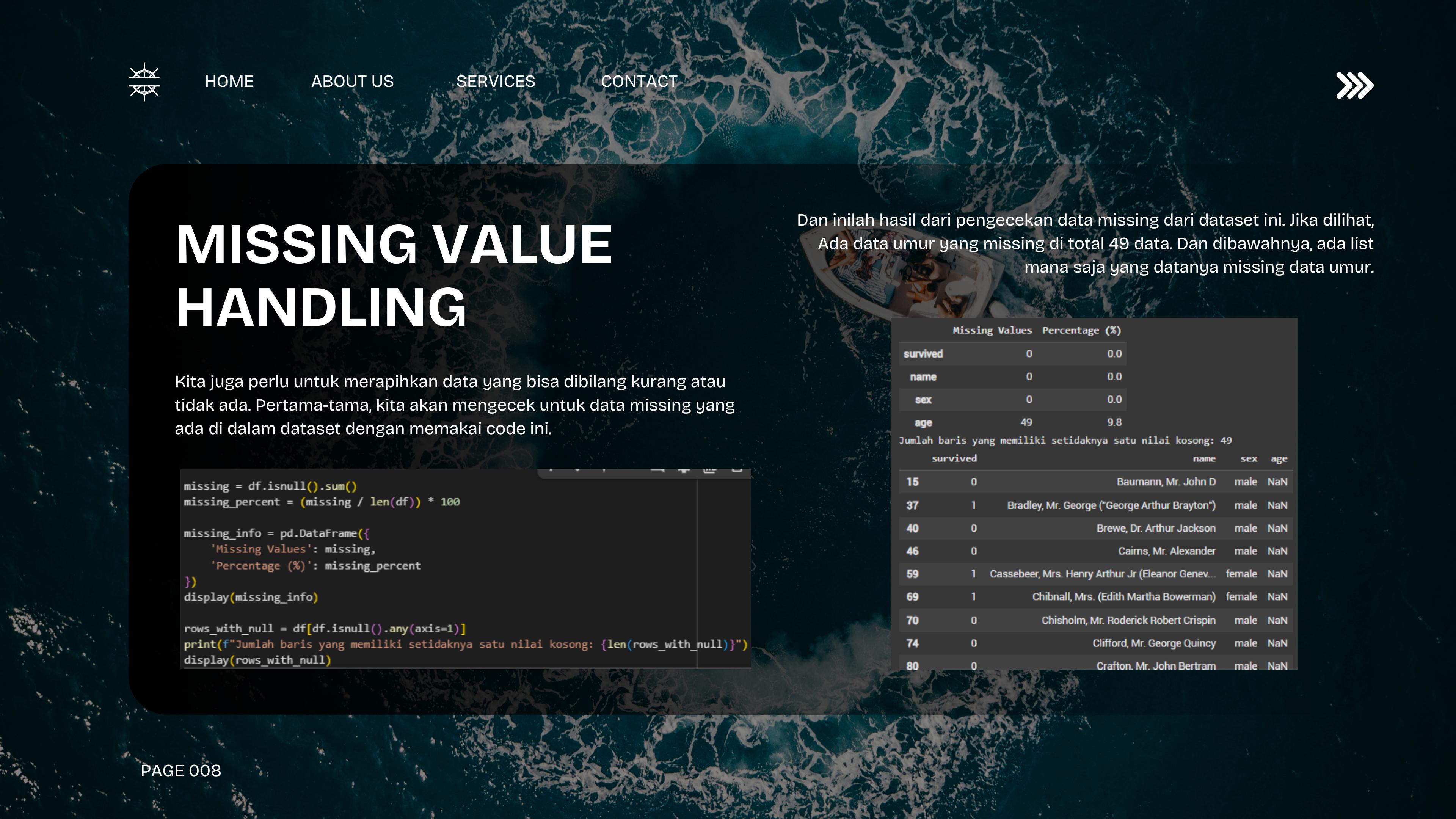
Kita juga perlu untuk merapihkan data yang bisa dibilang kurang atau tidak ada. Pertama-tama, kita akan mengecek untuk data missing yang ada di dalam dataset dengan memakai code ini.

```
missing = df.isnull().sum()
missing_percent = (missing / len(df)) * 100

missing_info = pd.DataFrame({
    'Missing Values': missing,
    'Percentage (%)': missing_percent
})
display(missing_info)

rows_with_null = df[df.isnull().any(axis=1)]
print(f"Jumlah baris yang memiliki setidaknya satu nilai kosong: {len(rows_with_null)}")
display(rows_with_null)
```

Dan inilah hasil dari pengecekan data missing dari dataset ini. Jika dilihat, Ada data umur yang missing di total 49 data. Dan dibawahnya, ada list mana saja yang datanya missing data umur.



	Missing Values	Percentage (%)
survived	0	0.0
name	0	0.0
sex	0	0.0
age	49	9.8

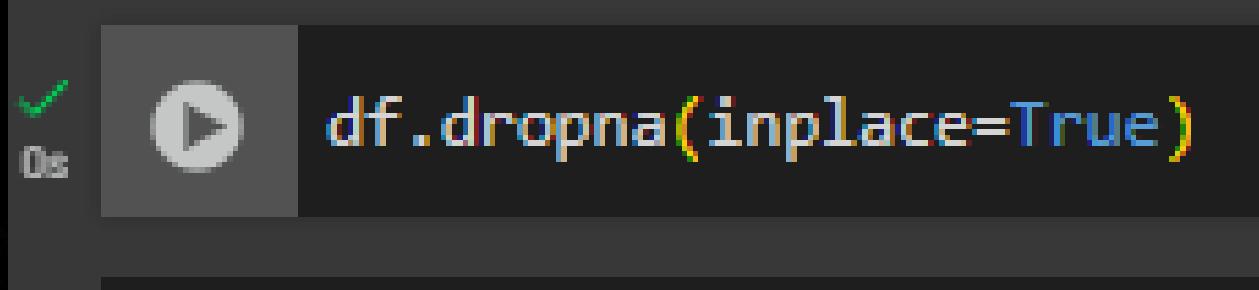
Jumlah baris yang memiliki setidaknya satu nilai kosong: 49

	survived	name	sex	age
15	0	Baumann, Mr. John D	male	NaN
37	1	Bradley, Mr. George ("George Arthur Brayton")	male	NaN
40	0	Brewe, Dr. Arthur Jackson	male	NaN
46	0	Cairns, Mr. Alexander	male	NaN
59	1	Cassebeer, Mrs. Henry Arthur Jr (Eleanor Genevieve)	female	NaN
69	1	Chibnall, Mrs. (Edith Martha Bowerman)	female	NaN
70	0	Chisholm, Mr. Roderick Robert Crispin	male	NaN
74	0	Clifford, Mr. George Quincy	male	NaN
80	0	Crafton, Mr. John Bertram	male	NaN

[HOME](#)[ABOUT US](#)[SERVICES](#)[CONTACT](#)

MISSING VALUE HANDLING

Yang udah pasti, kita itu juga mau dataset kita bersih dari berbagai data yang defect atau kurang bagus. Tahap terakhir dari ini adalah membersihkan data yang missing, dengan cara run code ini.



Jika sudah, maka saat kita cek lagi untuk data yang missing, kita bisa lihat bahwa sekarang tidak ada data yang missing sama sekali.

```
missing = df.isnull().sum()
missing_percent = (missing / len(df)) * 100

missing_info = pd.DataFrame({
    'Missing Values': missing,
    'Percentage (%)': missing_percent
})
display(missing_info)

rows_with_null = df[df.isnull().any(axis=1)]
print(f"Jumlah baris yang memiliki setidaknya satu nilai kosong: {len(rows_with_null)}")
display(rows_with_null)
```

	Missing Values	Percentage (%)	
survived	0	0.0	
name	0	0.0	
sex	0	0.0	
age	0	0.0	

Jumlah baris yang memiliki setidaknya satu nilai kosong: 0

survived name sex age



HOME

ABOUT US

SERVICES

CONTACT



THANK YOU!

END