Dibimbing.id

HOME SERVICE ABOUT US CONTACT US

EXPLANATORY DATA ANALYSIS TITANIC

Presentation by Dimas Faza Julian Putra

Presentasi ini membahas analisis data Titanic menggunakan metode Exploratory Data Analysis (EDA). Tujuannya untuk menemukan pola dan insight terkait faktor keselamatan penumpang berdasarkan usia, jenis kelamin, dan status sosial. Dataset ini populer di dunia data science karena sederhana dan cocok untuk latihan analisis data.

DIGITAL SKILL FAIR 38 - DATA SCIENCE

BACKGROUND

Setelah tragedi tenggelamnya kapal Titanic pada tahun 1912, banyak cerita tentang siapa yang selamat dan siapa yang tidak. Kita dapat melihat bagaimana variabel seperti usia, jenis kelamin, dan status sosial memengaruhi peluang hidup seseorang dengan menganalisis data ini. Karena data set Titanic lengkap, menarik, dan cocok untuk latihan analisis data dasar, mereka sering digunakan dalam bidang data science.

Dibimbing.id

HOME SERVICE ABOUT US CONTACT US

DATASET TITANIC

Dataset Titanic terdiri dari beberapa kolom penting yang merepresentasikan informasi dasar tiap penumpang. Kolom survived menunjukkan status keselamatan penumpang, di mana nilai 1 berarti selamat dan 0 berarti tidak selamat. Kolom name berisi nama lengkap penumpang lengkap dengan gelar sosial. Kolom sex mencatat jenis kelamin penumpang, yaitu male (laki-laki) atau female (perempuan). Sementara itu, kolom age mencerminkan usia penumpang dalam satuan tahun, termasuk pecahan desimal untuk usia bayi. Keempat kolom ini sangat penting dalam analisis awal untuk melihat pola keselamatan berdasarkan usia, jenis kelamin, dan status sosial.

urvived	name	sex	a
1	Allen, Miss. Elisabeth Walton	female	29.00
1	Allison, Master. Hudson Trevor	male	0.91
0	Allison, Miss. Helen Loraine	female	2.00
0	Allison, Mr. Hudson Joshua Creighton	male	30.00
0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00

```
<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
```

DATA OBSERVATION #1

Codingan tersebut adalah tiga cara menampilkan dataset Titanic dengan kode Python yang digunakan di Google Colab, tapi yang pertama yang harus dilakukan adalah mengimpor data dari file titanic.xlsx menggunakan kode pd.read_excel().

Berikut penjelasannya:

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

df.head() # menampilkan 5 teratas
```

	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

pd.read_excel berfungsi untuk menampilkan dataset titanic, dan df.head() nampilkan 5 baris pertama dari dataset, untuk melihat data bagian awal.

DATA OBSERVATION #2

Codingan tersebut adalah tiga cara menampilkan dataset Titanic dengan kode Python yang digunakan di Google Colab, tapi yang pertama yang harus dilakukan adalah mengimpor data dari file titanic.xlsx menggunakan kode pd.read_excel().

Berikut penjelasannya:

```
#menampilkan 5 terbawah
df.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

Meirip seperti df.head(), tapi df.tail() menambilkan sebaliknya, menampilkan 5 baris terakhir, untuk mengecek bagian akhir dataset.

DATA OBSERVATION #3

Codingan tersebut adalah tiga cara menampilkan dataset Titanic dengan kode Python yang digunakan di Google Colab, tapi yang pertama yang harus dilakukan adalah mengimpor data dari file titanic.xlsx menggunakan kode pd.read_excel().

Berikut penjelasannya:

```
# mengambil data dari 5 baris acak
df.sample(5)
```

	survived	name	sex	age
294	0	Thayer, Mr. John Borland	male	49.0
194	0	Maguire, Mr. John Edward	male	30.0
483	1	Lehmann, Miss. Bertha	female	17.0
79	1	Cornell, Mrs. Robert Clifford (Malvina Helen L...	female	55.0
496	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN

Hampir sama seperti sebelumnya, df.sample(5) akan menampilkan 5 baris data secara acak, sebagai sampel variasi isi dataset.

DATA OBSERVATION #4

Codingan tersebut adalah tiga cara menampilkan dataset Titanic dengan kode Python yang digunakan di Google Colab, tapi yang pertama yang harus dilakukan adalah mengimpor data dari file titanic.xlsx menggunakan kode pd.read_excel().

Berikut penjelasannya:

```
# cek struktur info dataset titanic
df.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
```

df.info() digunakan untuk melihat struktur data jumlah baris, kolom, tipe data, dan data yang kosong. Contohnya, kolom "age" punya data kosong karena cuma 451 dari 500 yang terisi.

STATISTICAL SUMMARRY

```
# menampilkan statistik dasar dari kolom numerik dalam dataset
df.describe()
```

	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 membantu kita memahami distribusi data numerik dalam dataset. Menggunakan perintah df.describe(), kita bisa mendapatkan informasi seperti jumlah data valid, rata-rata, standar deviasi, dan nilai ekstrim seperti minimum dan maksimum. Dari hasil ini, kita bisa mengobservasi beberapa hal penting:

1. Usia Penumpang Titanic: Usia penumpang bervariasi dari 0.42 hingga 80 tahun, dengan rata-rata sekitar 36 tahun. Ini menunjukkan bahwa penumpang datang dari berbagai kelompok usia.

2. Prosentase Penumpang yang Selamat: Berdasarkan kolom ‘survived’, sekitar 54% penumpang selamat. Ini memberi gambaran awal mengenai tingkat keselamatan di kapal Titanic.

3. Sebaran Data: Standar deviasi yang cukup besar pada kolom ‘age’ (sekitar 14 tahun) menunjukkan bahwa usia penumpang tersebar cukup luas, dan ini penting untuk analisis lebih lanjut.

Dari ringkasan ini, kita bisa melihat pola utama dalam data dan menemukan nilai–nilai yang terlalu ekstrem untuk kemudian kita perbaiki, misalnya dengan mengisi data yang kosong atau menindaklanjuti outlier.

DUPLICATE HANDLING

Menghitung Jumlah Data Duplikat:

```
# Hitung total baris yang duplikat
# (kecuali baris pertama tiap grup)
n_dups = df.duplicated().sum()
print(f"Jumlah baris duplikat: {n_dups}")
```

```
Jumlah baris duplikat: 1
```

Menghapus Baris Duplikat:

```
# hapus data duplikat
df.drop_duplicates(inplace=True)
```

Proses mengidentifikasi dan menghapus baris data yang tercatat lebih dari satu kali dalam kumpulan data dikenal sebagai pengendalian duplikat data. Duplikasi data dapat berdampak pada hasil analisis karena satu data dihitung lebih dari sekali, sehingga hasilnya bias.

MISSING VALUE HANDLING #1

Kode disamping ini menghitung jumlah data kosong di setiap kolom, lalu menghitung persentasenya berdasarkan total jumlah baris. Hasilnya disajikan dalam bentuk tabel missing_info yang menampilkan kolom, jumlah missing value, dan persentasenya.

```python
# mengecek jumlah missing value di tiap kolom
missing = df.isnull().sum()
missing_percent = (missing / len(df)) * 100

# menggabungkan informasi jadi satu tabel
missing_info = pd.DataFrame({
    'Missing Values': missing,
    'Percentage (%)': missing_percent
})
display(missing_info)

# menampilkan baris yang punya nilai kosong di kolom mana pun
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	0.000000
name	0	0.000000
sex	0	0.000000
age	49	9.819639

```
Jumlah baris yang memiliki setidaknya satu nilai kosong: 49
```

Dibimbing.id

HOME SERVICE ABOUT US CONTACT US

MISSING VALUE HANDLING #2

	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 Genev...	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
106	0	Farthing, Mr. John	male	NaN
107	1	Flegenheim, Mrs. Alfred (Antoinette)	female	NaN
108	1	Fleming, Miss. Margaret	female	NaN
118	0	Franklin, Mr. Thomas Parham	male	NaN

108	1	Fleming, Miss. Margaret	female	NaN
118	0	Franklin, Mr. Thomas Parham	male	NaN
121	1	Frauenthal, Mrs. Henry William (Clara Heinshei...	female	NaN
125	0	Fry, Mr. Richard	male	NaN
134	1	Goldenberg, Mrs. Samuel L (Edwiga Grabowska)	female	NaN
147	0	Harrington, Mr. Charles H	male	NaN
152	1	Hawksford, Mr. Walter James	male	NaN
157	0	Hilliard, Mr. Herbert Henry	male	NaN
166	0	Hoyt, Mr. William Fisher	male	NaN
176	1	Kenyon, Mrs. Frederick R (Marion)	female	NaN
179	0	Klaber, Mr. Herman	male	NaN
184	0	Lewy, Mr. Ervin G	male	NaN
196	1	Marechal, Mr. Pierre	male	NaN
204	1	Meyer, Mrs. Edgar Joseph (Leila Saks)	female	NaN

Gambar di samping merupakan hasil output yg menampilkan baris atau data yg berisi kosong di kolom manapun, berikut kodenya:

```
# menampilkan baris yang punya nilai kosong di kolom mana pun
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)
```

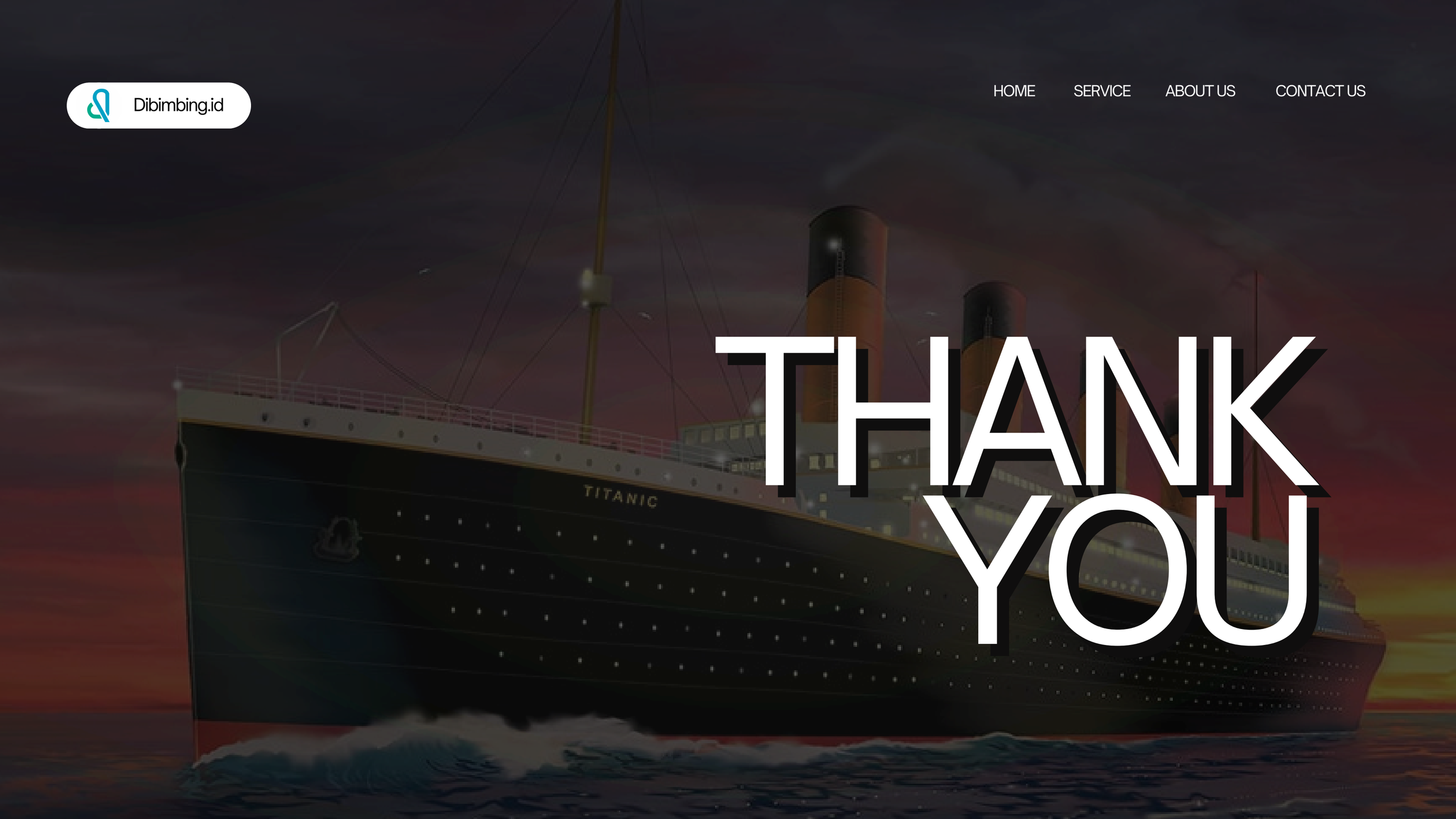
MISSING VALUE HANDLING #3

```
# buat hapus data yg null
df = df.dropna()
```

Kode pada gambar disamping ini berfungsi untuk menghapus seluruh baris dalam dataset yang memiliki nilai kosong di kolom manapun. Setelah kode ini dijalankan, dataset df hanya akan berisi data yang lengkap tanpa missing value.

Dibimbing.id

HOME SERVICE ABOUT US CONTACT US



THANK YOU