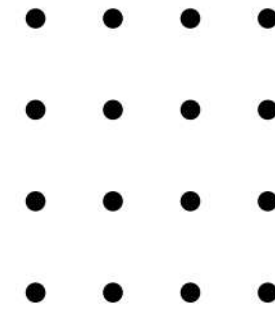


Ada apa Dibalik Titanic?

Dalam proyek ini, saya melakukan Explanatory Data Analysis (EDA) terhadap dataset legendaris Titanic untuk memahami fakta atau informasi apa saja yang bisa kita ungkap dari data penumpang Titanic, melalui visualisasi data, pembersihan data, serta analisis distribusi dan korelasi antar variabel.



Introduction

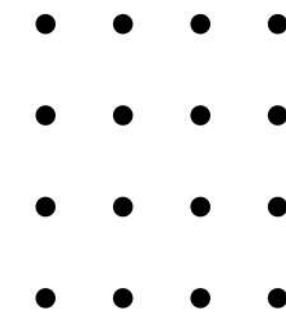
Apakah terdapat pola tersembunyi ?

Tragedi tenggelamnya kapal RMS Titanic pada tahun 1912 merupakan salah satu peristiwa paling ikonik dalam sejarah maritim. Di balik kisah tragis tersebut, tersembunyi berbagai pola dan informasi penting yang bisa kita gali dari data penumpangnya.

Dengan bantuan data, kita mencoba mengungkap:

Siapa yang memiliki kemungkinan lebih besar untuk selamat ?

Serta faktor-faktor apa saja yang mungkin memengaruhi keputusan penyelamatan dalam tragedi Titanic ?.



Tujuan Analisis

1. Melakukan investigasi awal terhadap data penumpang Titanic.
2. Mengetahui distribusi keselamatan berdasarkan jenis kelamin dan usia.
3. Mengidentifikasi keberadaan data duplikat dan data hilang (missing value), serta bagaimana cara menanganinya.
4. Menyajikan visualisasi yang mendukung pemahaman terhadap pola keselamatan penumpang.

Metodologi

Analisis dilakukan melalui pendekatan **Exploratory Data Analysis (EDA)**, dengan langkah-langkah sebagai berikut:

- Inspeksi Awal Data: Menggunakan **head()**, **tail()**, **sample()**, dan **info()** untuk memahami struktur dan karakteristik dataset.

head()

```
print(df.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()

```
print(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

sample()

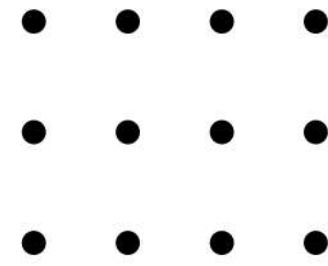
```
[33] print(df.sample(5, random_state=42))
```

	survived	name	sex	age
361	1	Caldwell, Mr. Albert Francis	male	26.0
73	1	Cleaver, Miss. Alice	female	22.0
374	1	Clarke, Mrs. Charles V (Ada Maria Winfield)	female	28.0
155	1	Hays, Mrs. Charles Melville (Clara Jennings Gr...	female	52.0
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0

info()

```
[34] df.info()
```

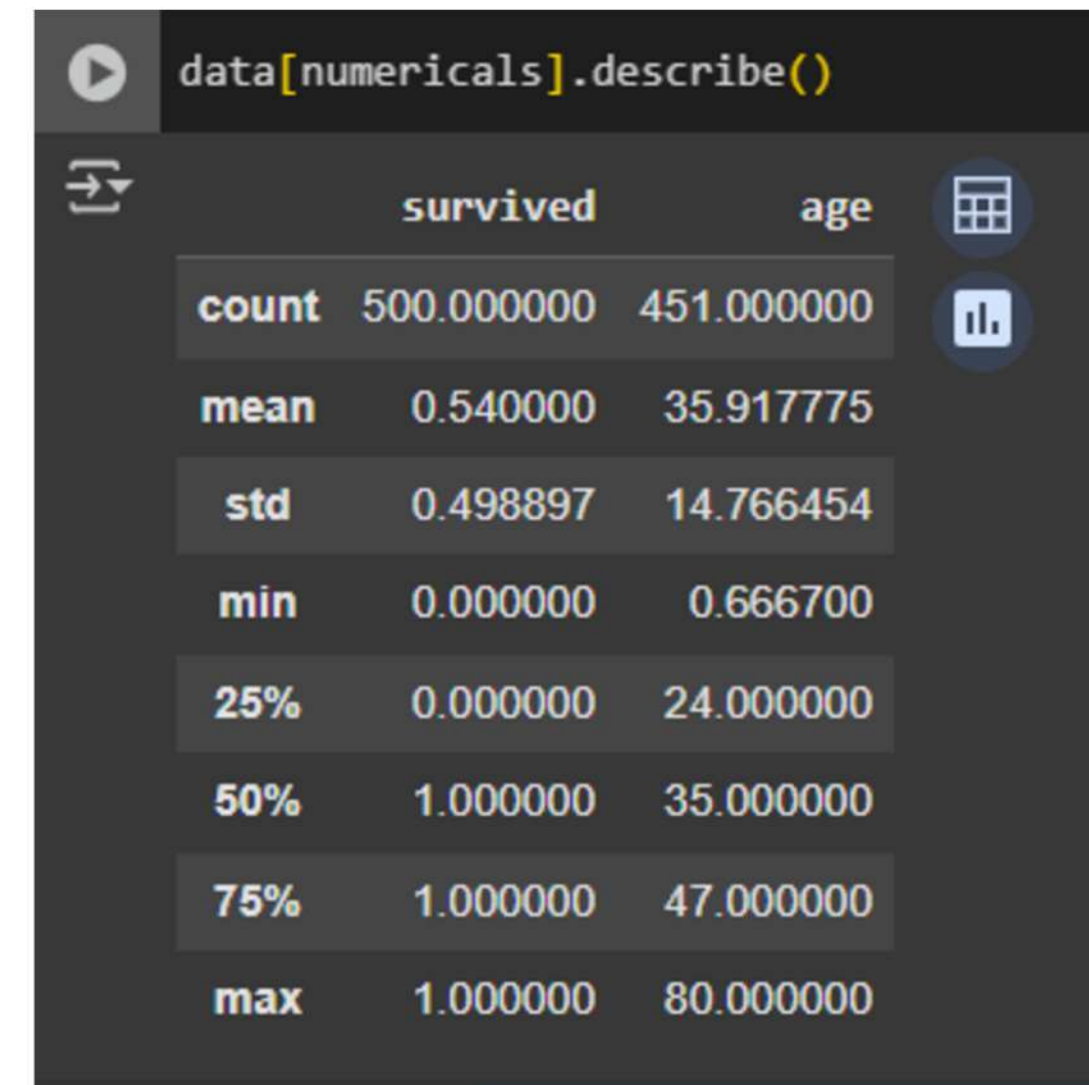
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  -
0   survived    500 non-null    int64
1   name        500 non-null    object
2   sex         500 non-null    object
3   age         451 non-null   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

Statistical Summary

Melihat statistik deskriptif untuk **kolom numerik** seperti kolom **survived** dan **age**

- Berdasarkan analisis kolom **survived**, diketahui bahwa dari **total 500** penumpang, sekitar **54% dinyatakan selamat**, sementara sisanya tidak selamat. Data ini menunjukkan bahwa tingkat keselamatan penumpang cenderung seimbang, meskipun sedikit lebih banyak yang selamat. Nilai median (50%) untuk kolom ini adalah 1, yang berarti sebagian besar penumpang memang selamat. Karena kolom ini bersifat biner (**0 = tidak selamat, 1 = selamat**), penyebarannya relatif tinggi, dengan standar deviasi sekitar 0.49.
- pada kolom **age**, tercatat bahwa hanya **451 penumpang** yang memiliki data usia, sehingga terdapat **49 data yang hilang** atau tidak terisi. Usia penumpang berkisar antara 0.67 tahun (sekitar 8 bulan) hingga 80 tahun, dengan **rata-rata usia sekitar 35.9 tahun**. Distribusi usia menunjukkan bahwa seperempat penumpang berusia di bawah 24 tahun, dan tiga perempat lainnya di bawah 47 tahun. Hal ini mencerminkan adanya **variasi usia yang cukup luas**, dari bayi hingga lansia, di antara penumpang Titanic.



```
data[numericals].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



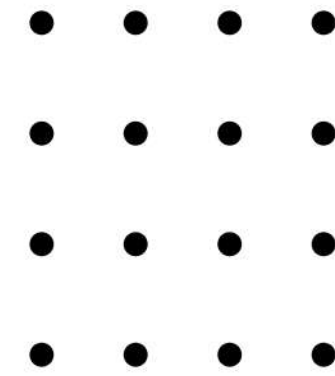
Statistical Summary

Berdasarkan hasil **analisis deskriptif**, dataset Titanic terdiri dari 500 entri dengan informasi lengkap pada kolom **name** dan **sex**, namun terdapat 49 data hilang pada kolom age. Mayoritas penumpang berjenis kelamin laki-laki (288 orang), dan sekitar 54% penumpang tercatat selamat. Kolom name hampir seluruhnya unik, hanya **satu nama** yang muncul **dua kali**. Data ini menunjukkan **adanya missing value** yang perlu ditangani serta distribusi demografis yang dapat dianalisis lebih lanjut untuk memahami pola keselamatan penumpang.

```
data.describe(include='all')
```

	survived	name	sex	age
count	500.000000	500	500	451.000000
unique	NaN	499	2	NaN
top	NaN	Eustis, Miss. Elizabeth Mussey	male	NaN
freq	NaN	2	288	NaN
mean	0.540000	NaN	NaN	35.917775
std	0.498897	NaN	NaN	14.766454
min	0.000000	NaN	NaN	0.666700
25%	0.000000	NaN	NaN	24.000000
50%	1.000000	NaN	NaN	35.000000
75%	1.000000	NaN	NaN	47.000000
max	1.000000	NaN	NaN	80.000000

DataCleaning



Mendeteksi dan menghapus data duplikat.

Terdapat data duplikat untuk penumpang bernama **Eustis, Miss. Elizabeth Mussey**. Duplikasi seperti ini kemungkinan merupakan kesalahan pencatatan dan perlu dihapus salah satunya

```
data = data.drop_duplicates()

len(data.drop_duplicates()) / len(data)

1.0
```

```
duplicates = data[data.duplicated(keep=False)]
```

duplicates

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

Menghandle duplikat data dengan syntax disamping ini

Data Cleaning

Menghitung dan menangani missing value

```
data.isnull().sum()
```

```
0
survived 0
name 0
sex 0
age 49
```

Kolom **age** memiliki **451** dari 500 nilai artinya terdapat **49 missing value**.

```
data.isnull().sum()
(data.isnull().sum() / len(df)) * 100
```

```
0
survived 0.0
name 0.0
sex 0.0
age 9.8
```

Kolom **age** memiliki nilai **9.8** artinya terdapat **9.8%** presentase **missing value**.

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

Show hidden output

```
data.isnull().sum()
(data.isnull().sum() / len(df)) * 100
```

```
0
survived 0.0
name 0.0
sex 0.0
age 0.0
```

Data Titanic dibersihkan dengan **menghapus baris duplikat** agar **analisis tidak bias**. Nilai kosong pada kolom **age** diisi dengan **rata-rata usia** menggunakan metode **imputasi**. Setelah itu, dicek kembali missing value dan hasilnya menunjukkan semua kolom sudah bersih dari nilai kosong, sehingga data siap untuk dianalisis.



Data Visualisasi

Dalam analisis ini, digunakan empat jenis visualisasi untuk membantu memahami karakteristik penumpang Titanic dan faktor-faktor yang memengaruhi keselamatan mereka. Visualisasi pertama berbentuk diagram lingkaran (pie chart) yang menggambarkan proporsi penumpang yang selamat dan tidak selamat. Dari visualisasi ini, tampak bahwa sebagian besar penumpang Titanic tidak selamat dari tragedi tersebut, menunjukkan betapa besar dampak dari bencana ini terhadap penumpang secara keseluruhan.

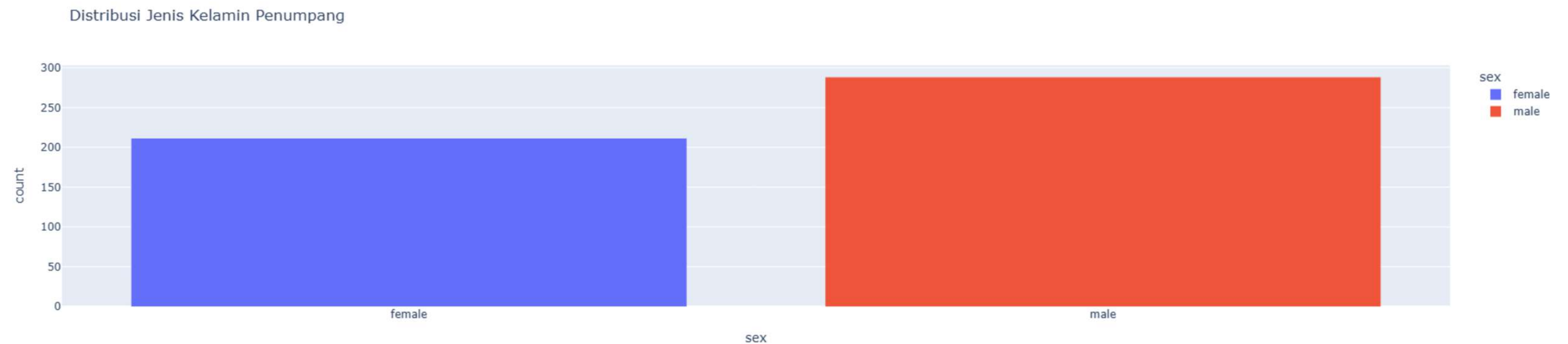
Distribusi Penumpang: Selamat vs Tidak Selamat





Data Visualisasi

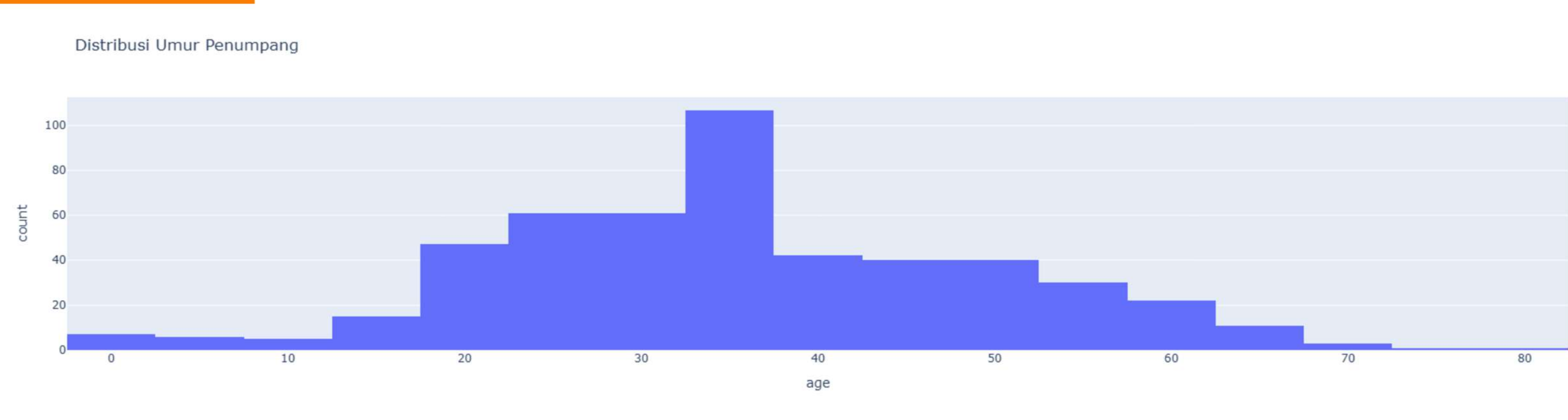
Visualisasi kedua menampilkan distribusi jenis kelamin penumpang dalam bentuk diagram batang. Hasilnya menunjukkan bahwa penumpang laki-laki jumlahnya lebih banyak dibandingkan perempuan. Informasi ini menjadi penting saat dikaitkan dengan tingkat keselamatan, karena dapat memberikan gambaran awal mengapa distribusi keselamatan mungkin tidak merata.





Data Visualisasi

Distribusi usia penumpang dianalisis melalui histogram. Rentang usia penumpang sangat bervariasi, mulai dari anak-anak hingga lansia. Sebagian besar penumpang berada pada usia produktif. Visualisasi ini membantu mengidentifikasi apakah ada kelompok umur tertentu yang lebih dominan dan mungkin mendapat prioritas dalam proses evakuasi.





Data Visualisasi

melakukan perbandingan antara jenis kelamin dan status keselamatan menggunakan grouped bar chart. Visualisasi ini mengungkapkan bahwa perempuan memiliki kemungkinan selamat yang jauh lebih tinggi dibandingkan laki-laki. Temuan ini memperkuat asumsi bahwa kebijakan “women and children first” diterapkan saat proses penyelamatan berlangsung, memberikan wawasan berharga tentang prioritas keselamatan dalam tragedi Titanic.



TerimaKasih