Titanic Dataset : **EDA Report & Insight**

(Using Python)

Dania Amelia Ansyori | Assignment Dibimbing Digital Skillfair 38.0 | Faculty of Data

Introduction

Background

I am working on this Titanic EDA project as:

- Part of an assignment submission for Digital Skillfair 38.0 (Data Science & Analysis) by Dibimbing.id
- Part of my personal learning journey in data analysis, with a focus on deepening my understanding of Exploratory Data Analysis (EDA).

The structure of the Titanic dataset offers a solid foundation for applying EDA techniques to uncover meaningful insights — particularly around factors that may have influenced passenger survival during the disaster.

Objective

The objective of this project is to explore the Titanic dataset in order to gain a deeper understanding of passenger survival rates. The analysis focuses on identifying key patterns based on factors such as age, gender, and passenger class. Through data cleaning and exploration, this project also aims to strengthen my skills in applying EDA techniques to practical datasets

Introduction

Tools

Python

Google
Collaboratory

ChatGPT

Introduction

Dataset Description

The Titanic dataset contains information about passengers aboard the Titanic ship. The dataset consists of **500 Rows** and **4 Columns**, including features such as **Name**, **Sex**, **Age**, and **Survived**. Below is a brief overview of the columns :

Column	Description	Type
Survived	Survival Status (0 = No, 1 = Yes)	Binary
Name	Name of Passenger	Categorical
Sex	Sex of Passenger (Male or Female)	Categorical
Age	Age in Years	Numerical

The Steps

Step 1 - Data Understanding

Step 2 - Data Cleansing

Step 3 - Findings & Visualisation

Data Understanding

Display the First 5 Rows of Dataset

	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

Display the Last 5 Rows of Dataset

	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

Data Understanding

Data Sampling (5 Samples from Dataset)

	survived	name	sex	age
86	1	Daly, Mr. Peter Denis	male	51.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
445	0	Hickman, Mr. Stanley George	male	21.0000
163	1	Holverson, Mrs. Alexander Oskar (Mary Aline To...	female	35.0000
136	1	Gracie, Col. Archibald IV	male	53.0000

Observations :

1. Column `survived` and `age` are numeric
2. Column `name` and `sex` are categorical
3. `sex` column seems to have 2 distinct values (female OR male), but will confirm it later
4. `survived` is apparently a binary (1, 0), but will confirm it later
5. data in `name` includes (last name, title, and first name)
6. There are NULL values in `age` column.

Data Understanding

Dataset Summary Information

```
<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
```

Observations :

1. Data contains 4 Column with 500 rows.
2. Only `age` column has missing values (will be handled later)
3. All the data types seems OK (appropiate), given the corresponding column name.
4. Some values in `age` are under 1 (e.g. 0.6667). It was an 8 month baby, 0.6667 * 12 = 8 months y/o.

Data Cleansing

Duplicate Handling

```
[44] duplicate_counts = duplicates.groupby(list(df.columns)).size().reset_index(name='jumlah duplikat')

     sorted_duplicates = duplicate_counts.sort_values(by='jumlah duplikat', ascending=False)

     print("Baris duplikat yang sudah diurutkan berdasarkan jumlah kemunculannya : ")
     sorted_duplicates
```

```
Baris duplikat yang sudah diurutkan berdasarkan jumlah kemunculannya :
```

	survived	name	sex	age	jumlah duplikat
0	1	Eustis, Miss. Elizabeth Mussey	female	54.0	2

Check for fully duplicated rows based on all available columns. There is a row that appears 2x.

Duplicated data will be droped

```
[45] df = df.drop_duplicates()
```

Re-Check if The Duplicate Data Handled Correctly

```
[47] len(df.drop_duplicates()) / len(df)
```

```
1.0
```

**Result = 1.0 (no duplicated data)* ✅

From This ⬇

```
[32] #untuk cek ada berapa baris di dataframe kita
     len(df)
```

```
500
```

To This ⬇

```
len(df)
```

```
499
```

Data Cleansing

Handling Missing Value

```
df.isna().sum()
```

	0
survived	0
name	0
sex	0
age	49

dtype: int64

Check the Missing Value in Data Frame

Presentate the missing value

```
[55] #dipersentasekan
     total_rows = len(df)

     #hitung dan tampilkan persentase missing value di setiap kolom
     for column in df.columns:
       missing_count = df[column].isna().sum()
       missing_percentage = (missing_count / total_rows) * 100
       print(f"Column {column} has {missing_count} missing values ({missing_percentage:.2f}%)")

     Column survived has 0 missing values (0.00%)
     Column name has 0 missing values (0.00%)
     Column sex has 0 missing values (0.00%)
     Column age has 49 missing values (9.82%)
```

**The percentage of missing values below 20% so we handle numerically with median, categorical with mode. But the categorical data type does not have missing values, namely name and sex.*

Handle Missing Value

```
#handle missing value in EDA without splitting
for column in df.columns:
  if df[column].dtype == 'object':
    #jika data tipenya kolom, isi dengan modulus :
    df[column].fillna(df[column].mode()[0], inplace = True)
  else :
    #jika kolom tipenya numerik, isi dengan median
    df[column].fillna(df[column].median(), inplace = True)
```

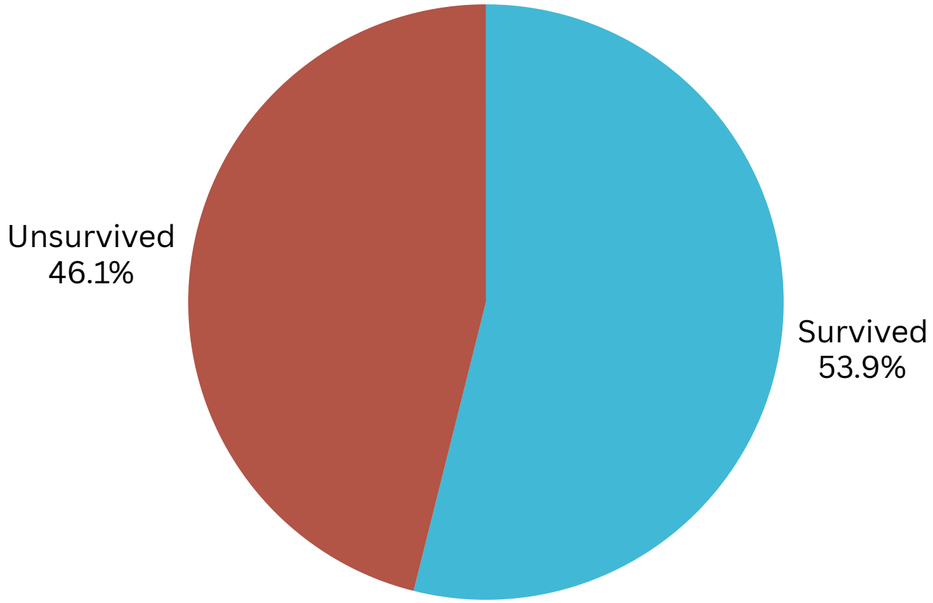
Re-Check if The Missing Value Handled Correctly

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 499 entries, 0 to 499
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  499 non-null    int64
 1   name      499 non-null    object
 2   sex       499 non-null    object
 3   age       499 non-null    float64
dtypes: float64(1), int64(1), object(2)
memory usage: 19.5+ KB
```

Findings

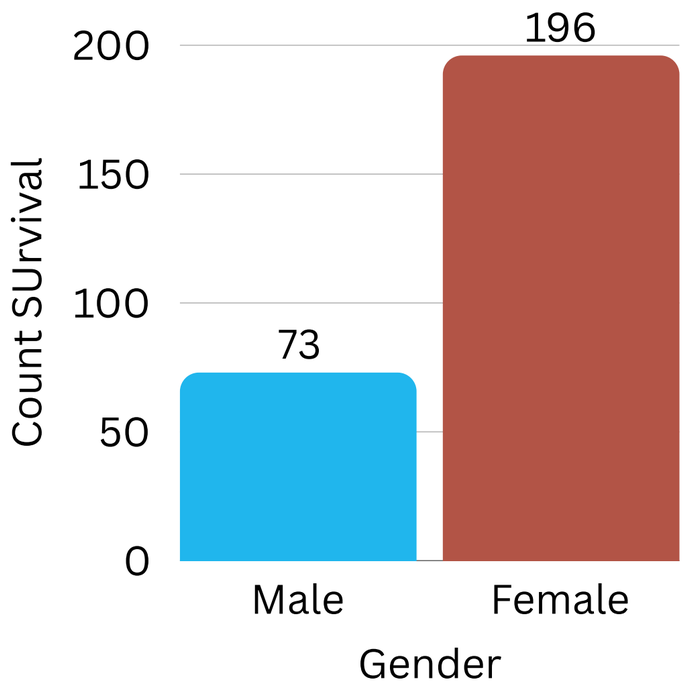
Survival Rate Overall



From the Total of 499 Passengers,
Survival Rate on Titanic :
53.9 % was Survived
46.1% was Unsurvived

Findings

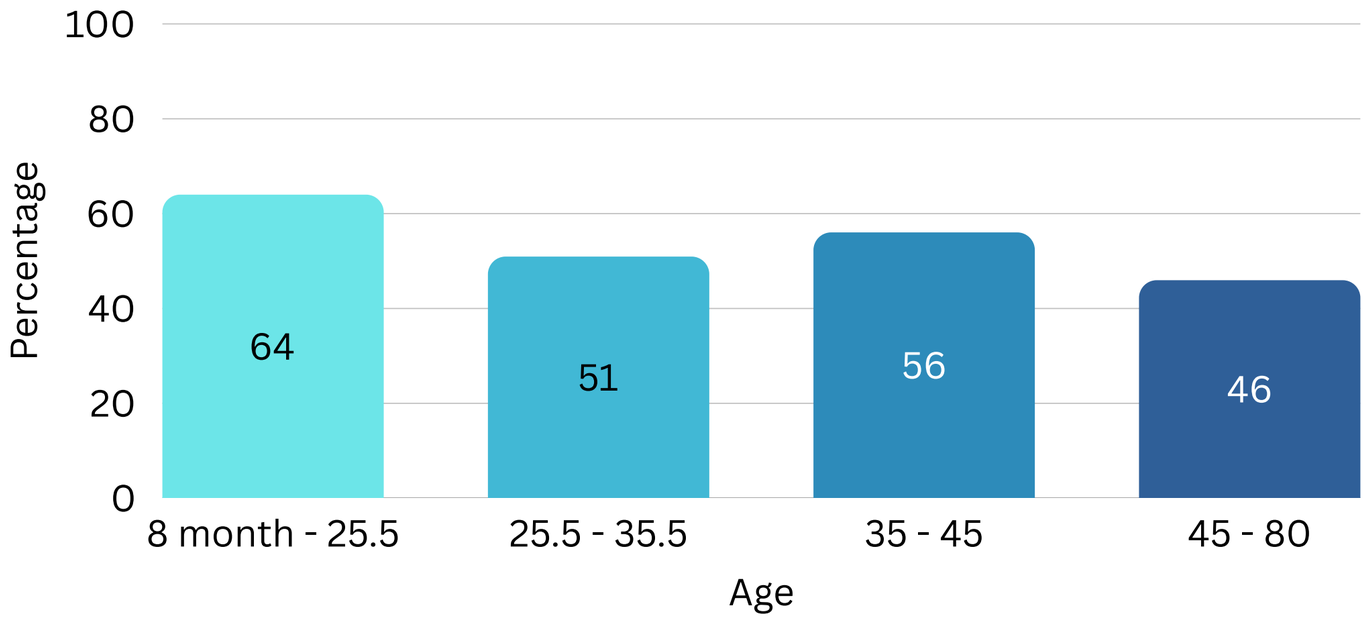
Survival Rate by Gender



Most Gender Survived is Female, with :
72.8 % survived Female
27.13 % survived Male

Findings

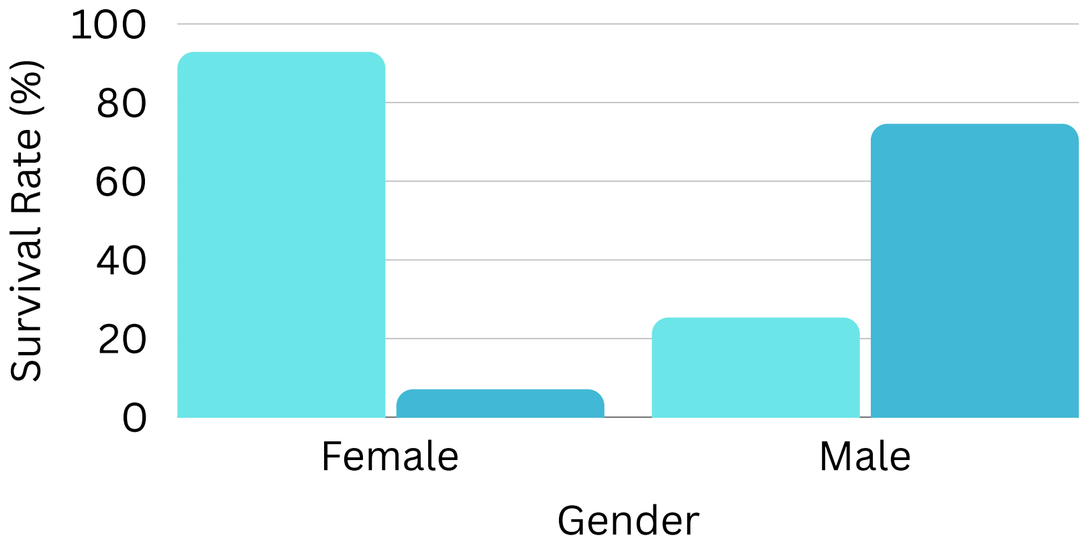
Survival Rate by Age Group



Most Age Group Survived are Passenger around 8 Month - 25.5 Years Old, with 64 %

Findings

Cross Analysis by Gender & Survival



Conclusion

Overall Survival Rate:
Only about 53.9% of the passengers survived the tragedy, reflecting the limited capacity of lifeboats and the chaotic nature of the evacuation process.

Survival by Gender:
There was a clear gender disparity in survival rates. Approximately 72.8% of female passengers survived, compared to only 27.13% of male passengers. This strongly supports the "women and children first" policy during the evacuation.

Survival by Age Quartiles:
When the passenger ages were divided into quartiles, it was found that the youngest age group (below ~20 years) had the highest survival rate. The survival rate declined with age, indicating that younger passengers were more likely to be saved.

Correlation Between Gender and Survival:
The data highlights a strong correlation between gender and survival likelihood. Being female significantly increased the chances of survival, while being male was associated with a much lower probability of survival.

Let’s Connect!

Dania Amelia (Amel)

dania amelia ansyori

ameliansyori