TITANIC DATASET EXPLORATORY DATA ANALYSIS (EDA)

MEINA LISA

DESCRIPTION

Background

This dataset provides passenger information from the Titanic. The initial exploration focuses on Survived, Name, Sex, and Age to understand how certain factors might have influenced survival chances, as the data contains many interesting details.

Goals

- Understand the distribution of Titanic Data (age, gender, etc.)
- Analyze patterns and relationships between variables.
- Find then handle any duplicate entries and missing data.
- Discover initial insights useful for further analysis.

WORKFLOW OVERVIEW

1. Data Understanding

Examine data structure.
Observe data distribution and statistics.

2. Data Cleaning

Remove duplicates.
Handle missing values.

3. Exploratory Data Analysis

Visualize distributions.
Analyze variable relationships.

4. Insight & Interpretation

Summarize data.
Identify influential factors.

DATASET OVERVIEW

Titanic dataset contains 500 rows and 4 columns: survived, name, sex, and age.

survived	name	sex	age
1	Allen, Miss. Elisabeth Walton	female	29.00
1	Allison, Master. Hudson Trevor	male	0.92
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

DATA SAMPLE

Showing Top 5 rows

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

Showing Bottom 5 rows

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

Showing Random 5 rows

```
survived                           name     sex   age
    1   Christy, Mrs. (Alice Frances)  female  45.0
    1   Duran y More, Miss. Asuncion  female  27.0
    0         Smith, Mr. Lucien Philip    male  24.0
    0               Keane, Mr. Daniel    male  35.0
    1          Ismay, Mr. Joseph Bruce    male  49.0
```

DATASET INFORMATION

```
Number of rows and columns: (500, 4)
List of columns:
['survived', 'name', 'sex', 'age']
Dataset 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)
```

DATA CLEANING

Remove Duplicate

Showing duplicate data:

Method:
duplicates = df.duplicated()

Result:

```
Number of duplicate rows: 1
      survived                          name     sex   age
104          1  Eustis, Miss. Elizabeth Mussey  female  54.0
349          1  Eustis, Miss. Elizabeth Mussey  female  54.0
```

Handling:
df = df.drop_duplicates
().reset_index(drop=True)

Handle Missing Value

Identifying missing value:

Method:
missing_count = df.isnull().sum()

Result:

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

Handling:
Missing values in the 'Age' column
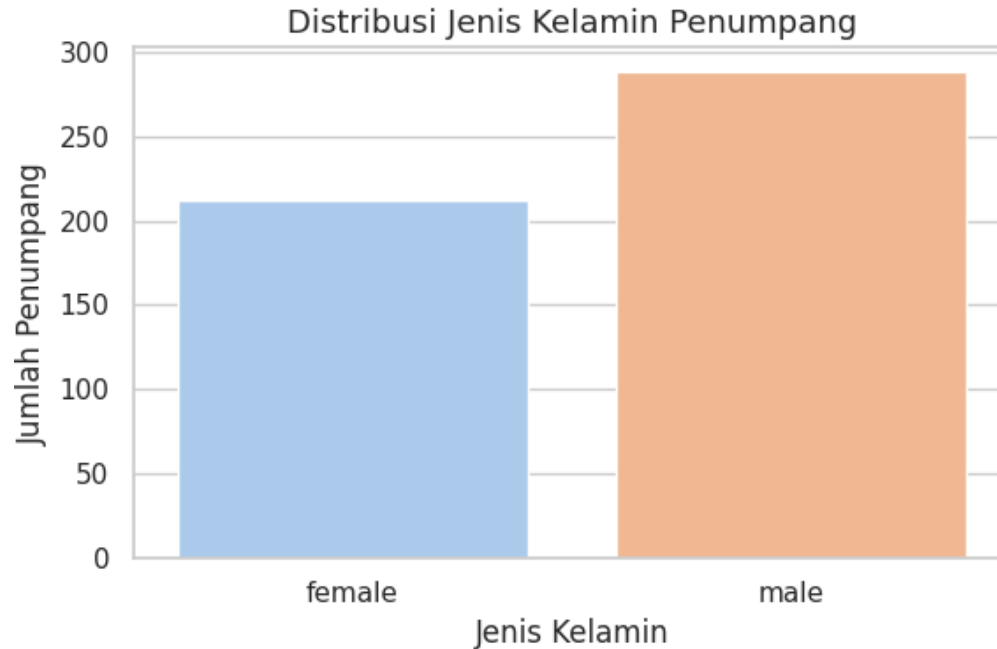were filled with the median: 35.0

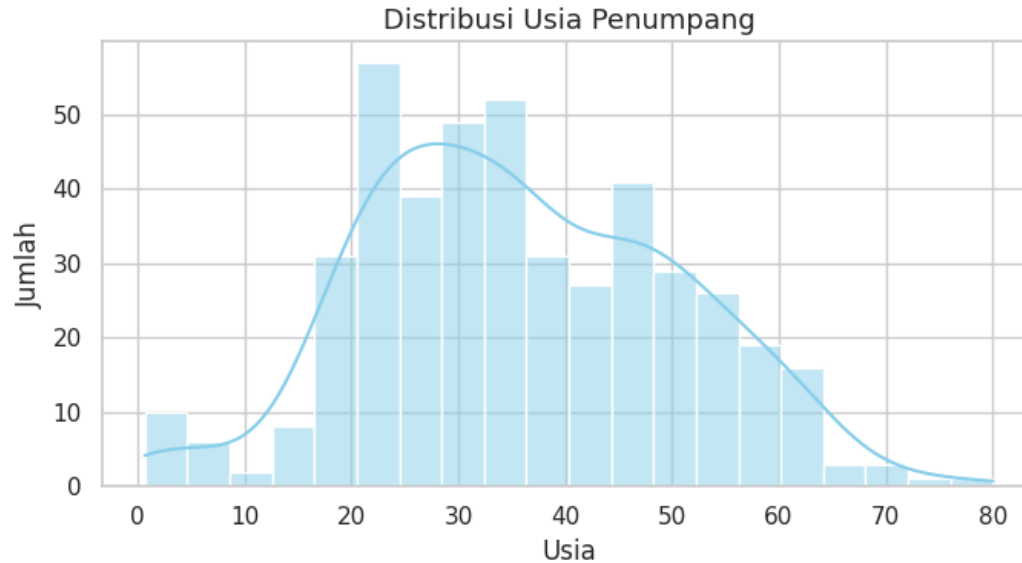# EXPLORATORY DATA ANALYSIS

SEX DISTRIBUTION

Distribusi Jenis Kelamin Penumpang

Analysis:

- 288 male passengers

- 211 female passengers

- Dominated by male (57.7%)

AGE DISTRIBUTION

Distribusi Usia Penumpang

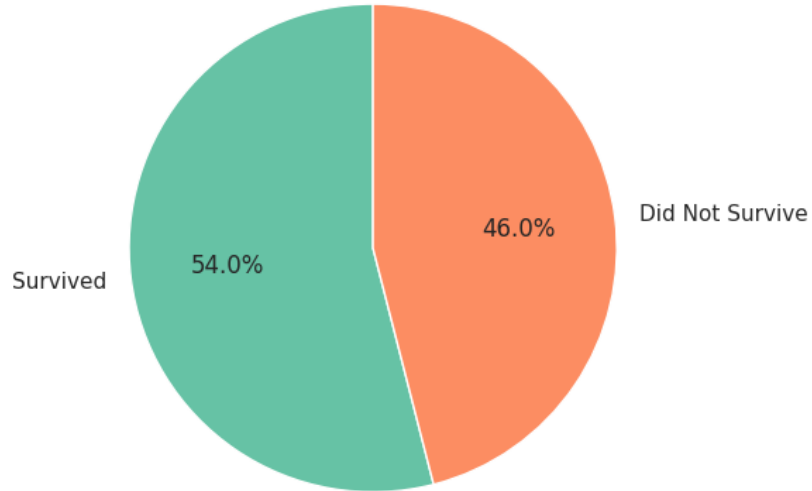**Analysis:**

- Dominated by adults
  (20 – 59 years), with 413
  people out of 499 total

SURVIVAL RATE

Titanic Passenger Survival Distribution

Analysis:

- The distribution of survivors vs non-survivors is relatively balanced.

CONCLUSIONS

Summary of Findings:
- The dataset contains 500 rows and 4 columns: survived, name, sex, and age.
- The age column had 49 missing values (9.8%), which were addressed using median imputation.
- There was one duplicate data, so 1 row were removed.
- Average passenger age: approximately 29 years, ranging from 0.75 to 80 years old.
- Passenger composition: 288 males and 212 females.
- Survival count: 270 passengers survived, and 230 did not survive.

THANKS!

meinalisa02@gmail.com
+(62) 852 6049 8159