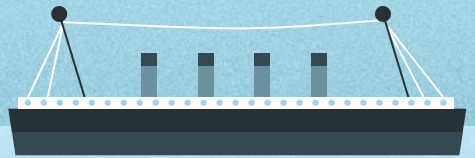
TITANIC DATA

— Exploration and Pre-processing Data with Python —

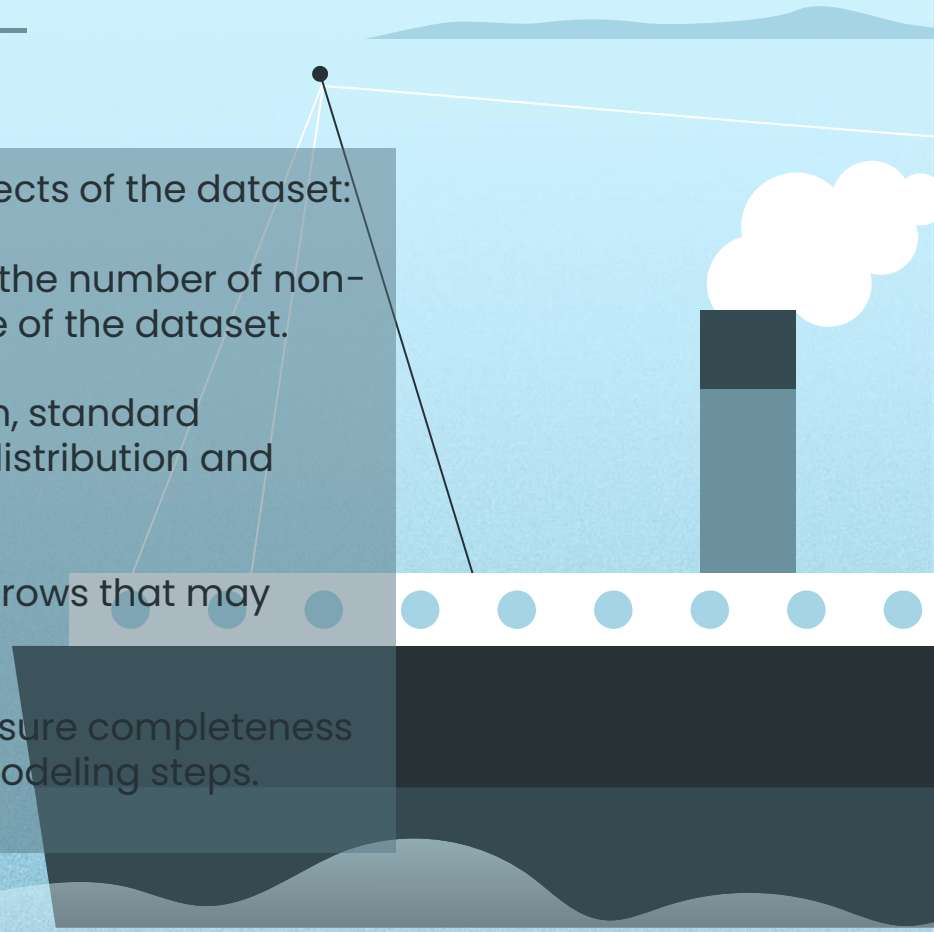
By: Panreshma Rizkha



— Titanic Data Exploration —

This discussion will cover the following key aspects of the dataset:

1. **Structure and Characteristics**
   Reviewing column names, data types, and the number of non-null entries to understand the overall shape of the dataset.
2. **Statistical Summary**
   Providing summary statistics such as mean, standard deviation, and percentiles to examine the distribution and central tendencies of numeric features.
3. **Duplicate Records**
   Identifying and addressing any duplicated rows that may affect the integrity of the analysis.
4. **Missing Values**
   Detecting and handling missing data to ensure completeness and reliability for subsequent analysis or modeling steps.



By: Panreshma Rizkha

— View Data —

To display data, there are several display options presented, here are some data displays along with examples and syntax:

no	survived	name	sex	age
1	1	Allen, Miss. Elisabeth Walton	female	29.0
2	1	Allison, Master. Hudson Trevor	male	0.9167
3	0	Allison, Miss. Helen Loraine	female	2.0
4	0	Allison, Mr. Hudson Joshua Creighton	male	30.0
5	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0

`df.head()`

=> Display the first 5 data

no	survived	name	sex	age
496	1	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.0
497	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN
498	0	Matthews, Mr. William John	male	30.0
499	0	Maybery, Mr. Frank Hubert	male	40.0
500	0	McCrae, Mr. Arthur Gordon	male	32.0

`df.tail()`

=> Display the last 5 data

no	survived	name	sex	age
378	1	Collyer, Miss. Marjorie "Lottie"	female	8.0
281	1	Stengel, Mrs. Charles Emil Henry (Annie May Mo...	female	43.0
131	1	Gibson, Mrs. Leonard (Pauline C Boeson)	female	45.0
114	0	Fortune, Mr. Charles Alexander	male	19.0
107	1	Flegenheim, Mrs. Alfred (Antoinette)	female	NaN

`df.sample()`

=> Display 5 random data

By: Panreshma Rizkha

— Info Data —

To display data info, here is the syntax and display of observation results:

```
df.info()
```

```
RangeIndex: 500 entries
```

No	Column	Non-Null Count	Dtype
1	survived	500 non-null	int64
2	name	500 non-null	object
3	sex	500 non-null	object
4	age	451 non-null	float64

Column description:

- No = contains the sequence number or the number of variables in the data
- Column = The name of the column in the DataFrame
- Non-Null Count = The number of non-null data in the column.
- Dtype = The data type of the column. For example: int64 (integer), float64 (decimal number), object (usually string or text).

Observation:
1. The data contains 4 variables with 500 rows of data
2. There is empty data in the “age” column
3. The type of data type corresponds to its name

By: Panreshma Rizkha

— Statistical Summary —

An overview of the main characteristics of a dataset. datasets can be divided into numerical and categorical (non-numerical) data.

```
categoricals = ['name', 'sex']
numericals = ['survived', 'age']
```

	name	sex
count	500	500
unique	499	2
top	Eustis, Miss. Elizabeth Mussey	male
freq	2	288

```
df[categoricals].describe()
```

⇒ Count (total amount of data)

⇒ Unique (number of unique categories)

⇒ Top (most frequently occurring category)

⇒ Freq (frequency of occurrence of the most common category)

	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

```
df[numericals].describe()
```

⇒ Count (number of data)

⇒ Mean (average)

⇒ Std (standard deviation)

⇒ Min (minimum value)

⇒ 25%, 50%, 75% (quartiles)

⇒ Max (maximum value)

Observations on categories:
1. “name” may have duplicate data.
2. “sex” has female and male categories, with the remaining 288 males being female.

Observations on numeric:
1. “age” shows a symmetrical distribution based on similar mean and median values.
2. “survived” is a binary column as its value is 0 or 1.

By: Panreshma Rizkha

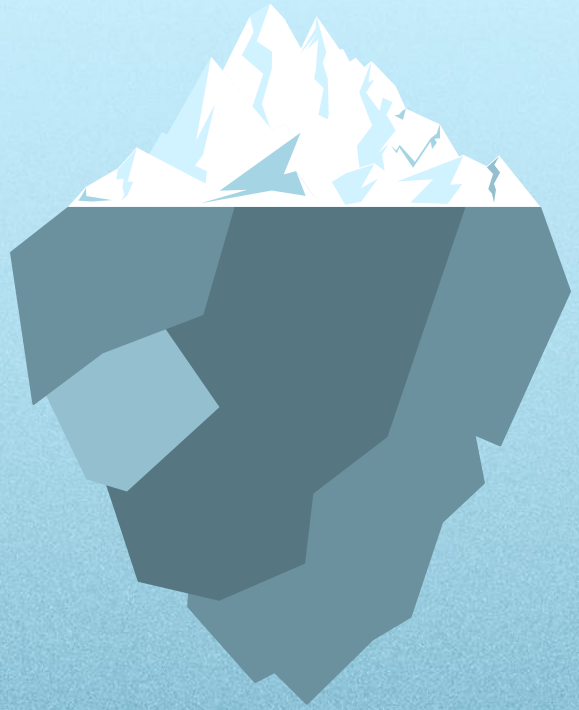
— Duplicate Data Check and Resolve —

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

no	survived	name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0
349	1	Eustis, Miss. Elizabeth Mussey	female	54.0

to solve it with this syntax:

⇒ `df = df.drop_duplicates()`



By: Panreshma Rizkha

— Missing Value Check and Resolve —

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

Variable	Missing Value
survived	0
name	0
sex	0
age	49

```
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%)
```

to solve it with this syntax:

- Jika kolom bertipe object, isi dengan mode:

⇒ `df[column].fillna(df[column].mode()[0], inplace=True)`

- Jika kolom bertipe numerik, isi dengan median:

⇒`df[column].fillna(df[column].median(), inplace=True)`

By: Panreshma Rizkha

— Conclusion —

Variable	Missing Value
survived	0
name	0
sex	0
age	0

No	Column	Non-Null Count	Dtype
1	survived	499 non-null	int64
2	name	499 non-null	object
3	sex	499 non-null	object
4	age	499 non-null	float64

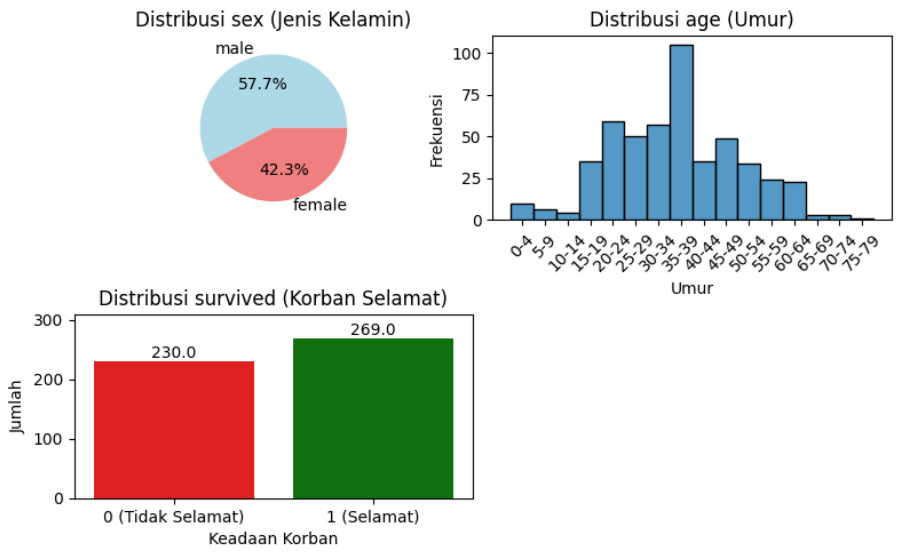
In conclusion, the initial exploration of the dataset has provided a clear understanding of its structure and key characteristics. The statistical summary has helped us understand the distribution and central tendencies of the numeric features. Additionally, duplicate records have been identified and removed, and missing values have been appropriately handled. These preprocessing steps ensure the dataset is clean and ready for the next stages of analysis or modeling. Proper data preparation is essential to ensure the quality, accuracy, and reliability of any insights or predictive results derived from the dataset.

By: Panreshma Rizkha

ADDITIONAL PART

DATA VISUALIZATION

— Graphics By Variable —



The graph on the side visualizes the variables of the net data. The following is the explanation:

1. Sex Graph

Shows that there are more male passengers than female passengers.

2. Age Range Graph

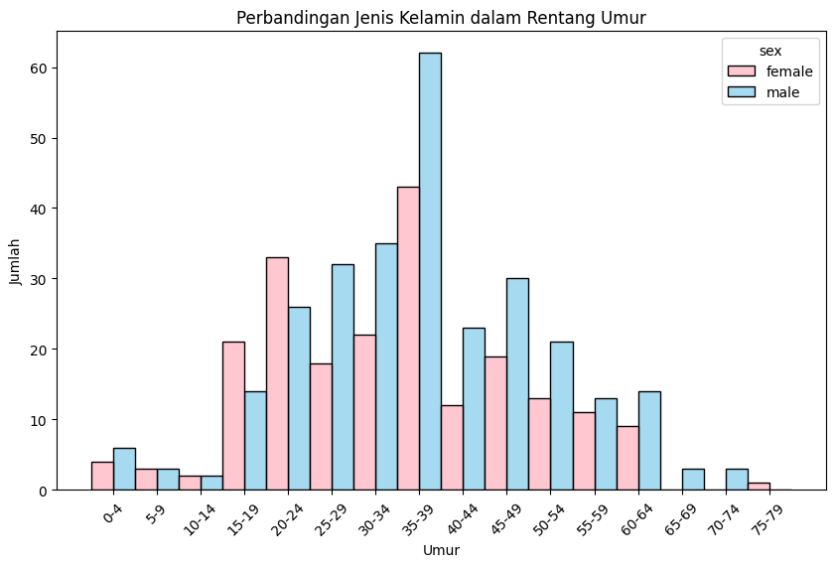
Shows that the dominant passengers are in the age range of 35 to 39 years old, with the number reaching around 100 people.

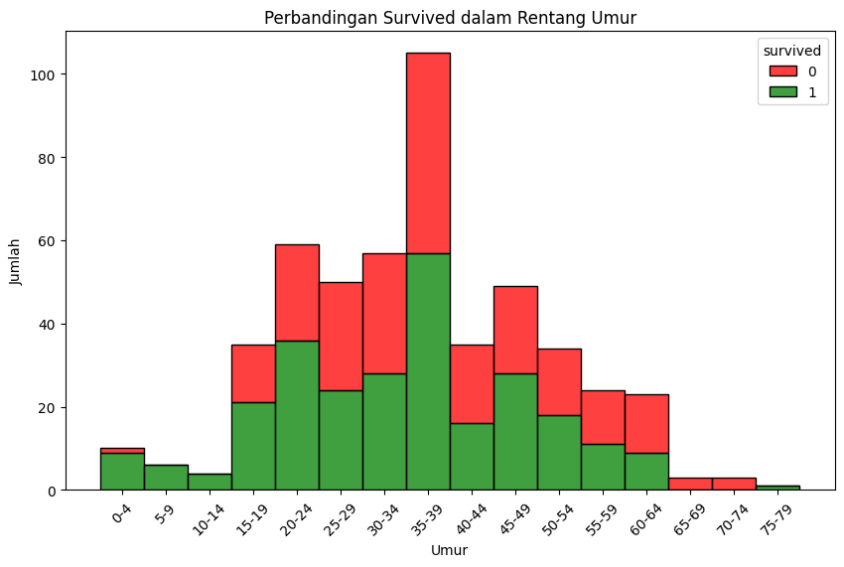
3. Survived Graph

Shows that the survivors did not survive almost half of the total.

By: Panreshma Rizkha

— Correlation Age With Sex and Survived —



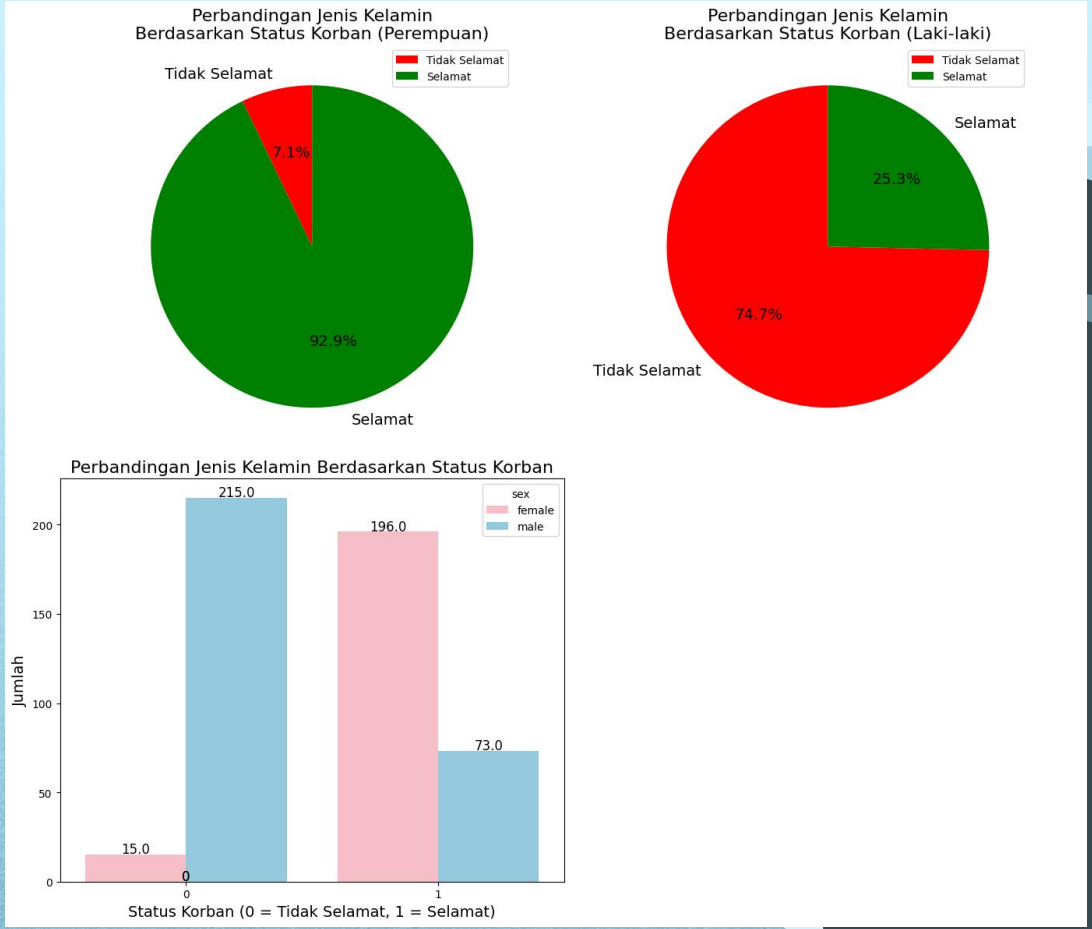


The graph shows the correlation of passenger age with gender and survived. In gender, ages 35-39 are dominated by male passengers. Next, in survived almost all age ranges can survive, only in the 65-74 range that cannot survive.

By: Panreshma Rizkha

— Correlation Survived with Sex —

The graph shows the correlation between survived and sex. There is a clear difference. In female passengers, only 7.1% did not survive, but in male passengers almost three out of four percent did not survive.



By: Panreshma Rizkha

Get More Information
View Repository on:

https://github.com/panreshma/Titanic-Data-Exploration-with-Python.git

THANK YOU!