TASK 3

EXPLORATORY DATA ANALYSIS ON TITANIC DATASET

Name : Kajal Prajapati

Date : 05-06-2025

Dataset : train.csv

OBJECTIVE:

TO ANALYZE THE TITANIC DATASET USING STATISTICAL AND VISUAL EXPLORATION, WITH THE GOAL OF IDENTIFYING IMPORTANT PATTERNS, TRENDS, AND RELATIONSHIPS—ESPECIALLY REGARDING THE SURVIVAL OF PASSENGERS.

TOOLS USED:

1. PYTHON
2. PANDAS
3. MATPLOTLIB
4. SEABORN

DATASET COLUMNS OVERVIEW

Column	Description
PassengerId	Unique ID for each passenger
Survived	0 = No, 1 = Yes (Target column)
Pclass	Ticket class (1 = Upper, 2 = Middle, 3 = Lower)
Name	Passenger's full name
Sex	Gender
Age	Age in years
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket fare
Cabin	Cabin number (many missing)
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

DATA CLEANING PERFORMED

- Handled missing values:
  1. Age: Missing values handled using median
  2. Embarked: Filled with most common value (‘S’)
  3. Cabin: Dropped due to too many nulls

- Converted data types as needed
- Verified data using .info() and .describe()

Statistical Summary (using describe()):

- Age ranged from ~0.4 to 80 years, average around 29.7
- Fare ranged widely (min = 0, max = 512), skewed distribution
- Pclass mostly 3rd class passengers
- SibSp and Parch mostly 0 (solo travelers)

```
df.describe()
```

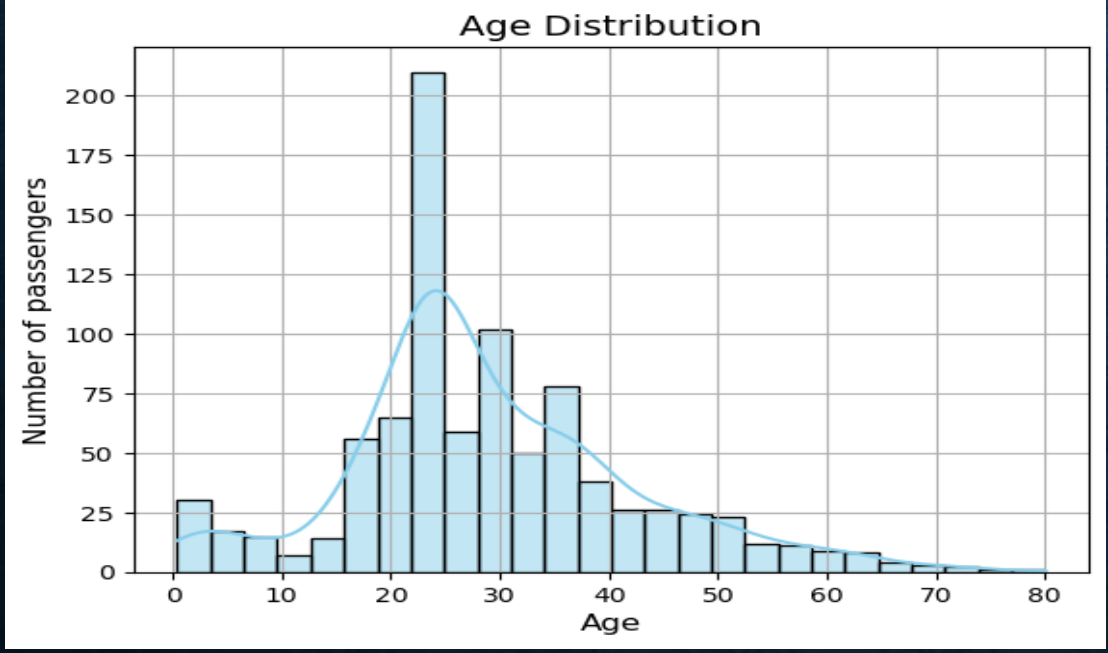
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.066409	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	13.244532	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	26.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	37.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

VISUAL EXPLORATIONS

1. Age Distribution
Plot: Histogram with KDE

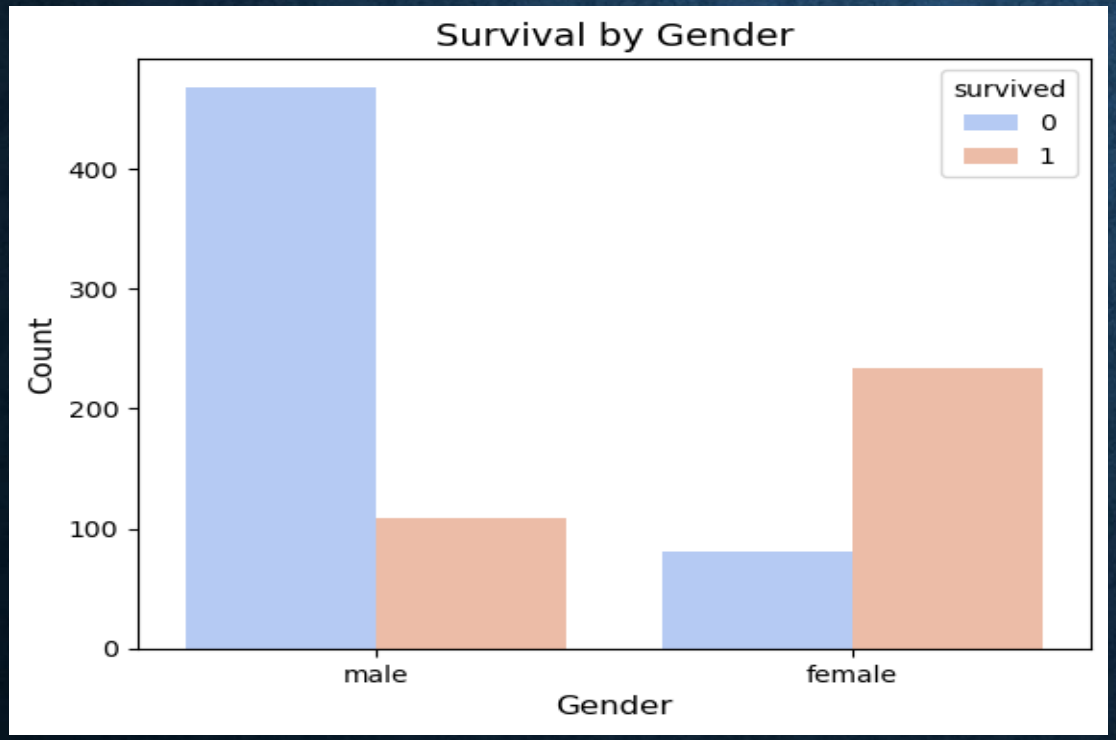
Observation:

Most passengers were aged between 20–40.
Very few were elderly (60+), some children were also present.



2. SURVIVAL BY GENDER(BAR PLOT)

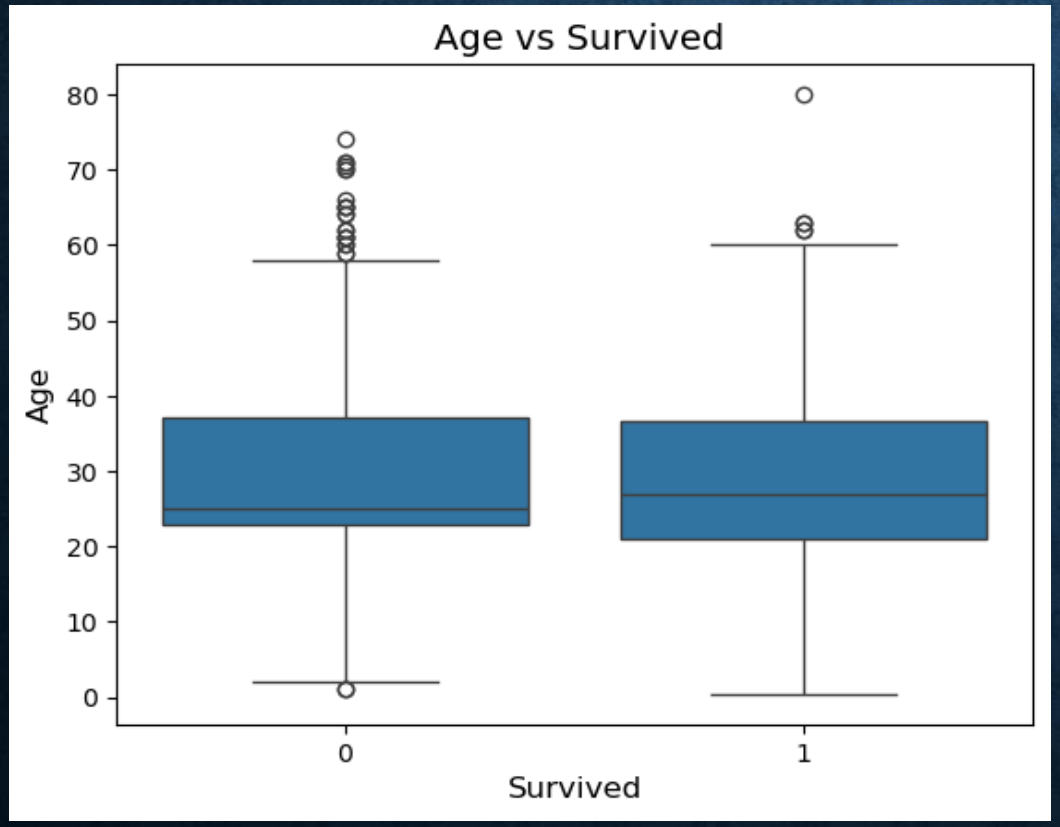
Observation:
Survival rate was higher among females.



3. BOXPLOT - AGE VS SURVIVED

Observation:

Younger passengers had slightly higher chances of survival.

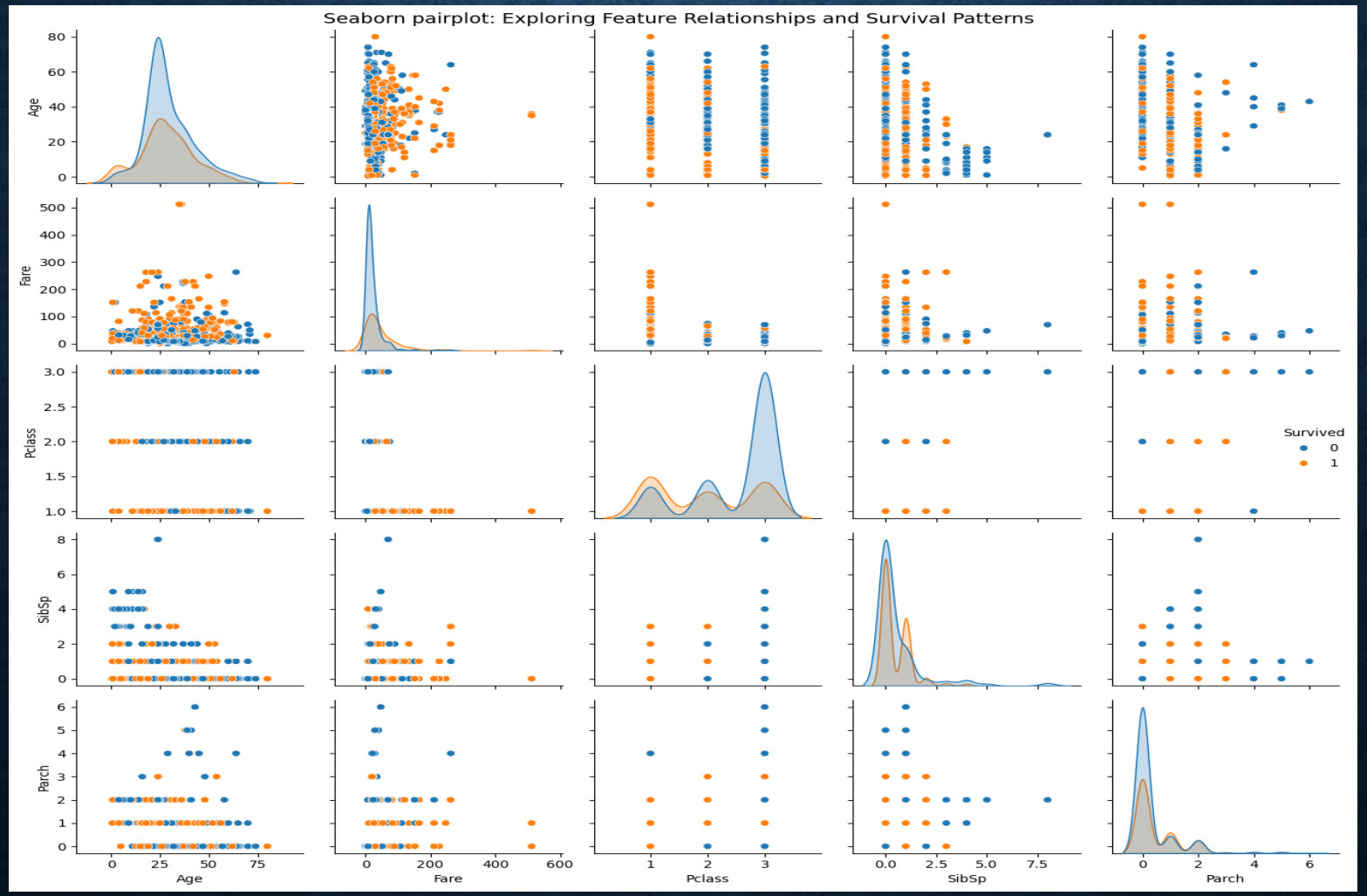


4. PAIRPLOT - FEATURE RELATIONSHIPS

Observation:

- Higher Fare → more likely to survive.
- Most survivors were in Pclass 1.
- Some outliers in Fare and Age.
- Plot – On 10th slide

PAIRPLOT - FEATURE RELATIONSHIPS



SUMMARY OF INSIGHTS

- **Females, 1st class passengers, and high fare payers** had higher chances of survival.
- **Age group 20–40** was the largest group onboard.
- Most passengers traveled alone (SibSp=0, Parch=0).
- Embarked location ‘S’ had the most passengers, but survival was higher in ‘C’.