**Task 5: Exploratory Data Analysis (EDA)**  
**Data Analyst Internship**  
**Dataset: Titanic Dataset**  
**Date:**28.4.2025  
**Name:**Saisruthi V

**1. INTRODUCTION**

**Objective:**  
The goal of this project is to perform exploratory data analysis on the Titanic dataset to uncover hidden patterns, insights, and relationships among variables that influence survival.

**2. DATASET DESCRIPTION**

**Dataset Features:**

* PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
* Target Variable: **Survived**

**3. EDA AND VISUALIZATIONS**

**3.1 Basic Information**

**Dataset Shape:**  
The Titanic dataset contains **891 rows** and **12 columns**.  
Each row represents a single passenger's information.

**Missing Values:**  
Several columns have missing data:

* **Age**: About 19.9% missing.
* **Cabin**: Around 77% missing.
* **Embarked**: 2 missing values.

Handling missing values carefully is crucial during preprocessing to avoid bias in the analysis.

**Data Types:**  
The dataset contains both **numerical** and **categorical** features:

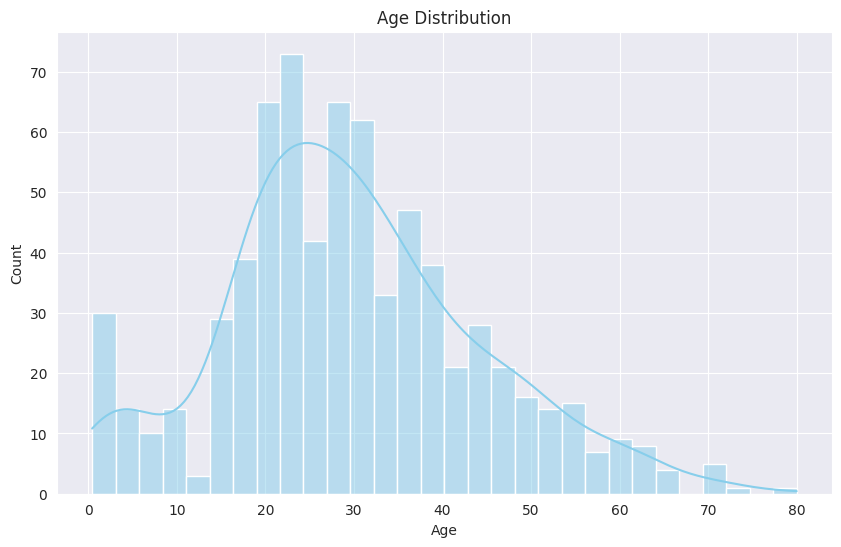
* **Numerical Columns**:
  + Age (float)
  + Fare (float)
  + SibSp (integer)
  + Parch (integer)
  + PassengerId (integer)
* **Categorical Columns**:
  + Survived (binary: 0 = No, 1 = Yes)
  + Pclass (ordinal)
  + Sex (nominal)
  + Ticket (nominal)
  + Cabin (nominal)
  + Embarked (nominal)
  + Name (text)

**Initial Observations:**

* A significant amount of missing data exists in the **Cabin** column.
* **Sex** and **Pclass** seem like important features for survival prediction.
* **Age** has a right-skewed distribution which could affect modeling later.

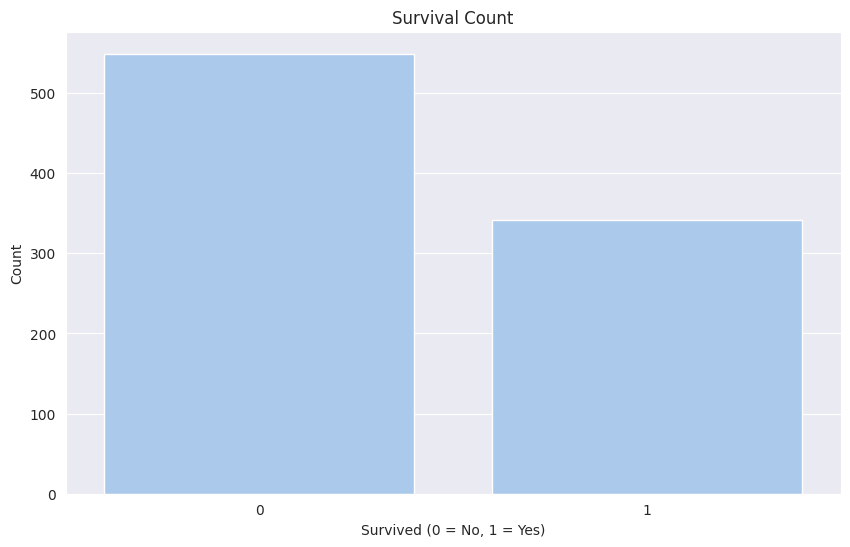
**3.2 Univariate Analysis**

**a) Age Distribution**

* 

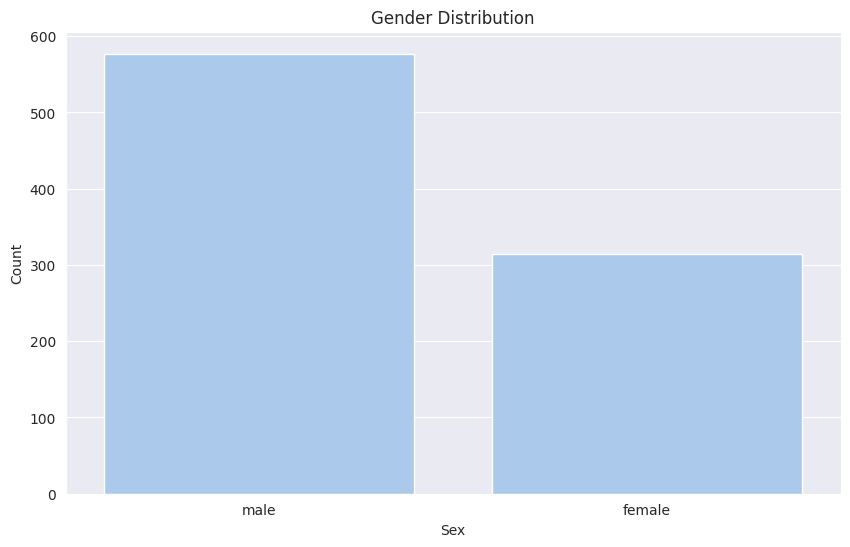
**Observation:**  
Most passengers were aged between 20–40 years.

**b) Survival Count**

* 

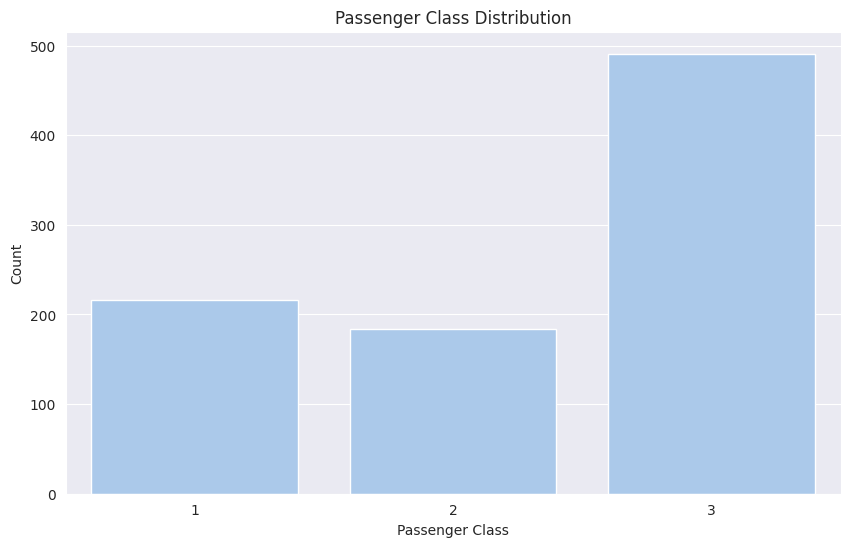
**Observation:**  
Majority (62%) did not survive.

**c) Gender Distribution**

* 

**Observation:**  
More males were onboard than females.

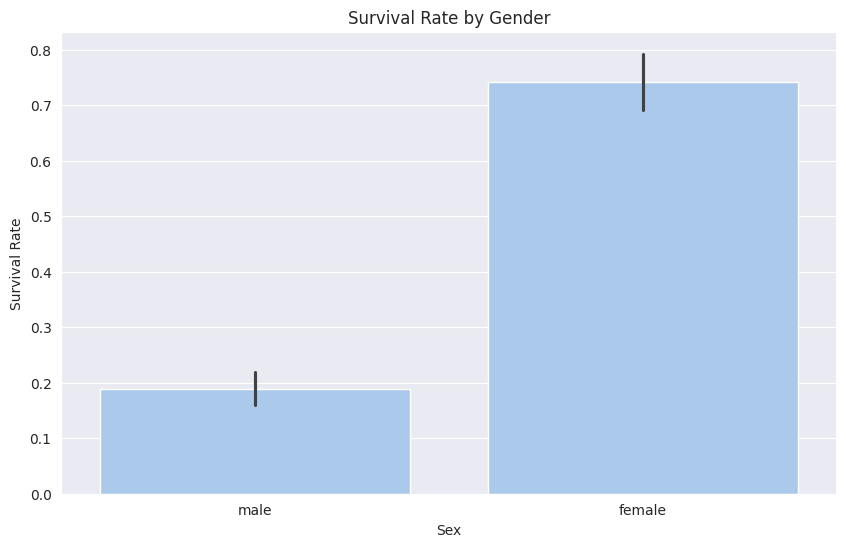
**d) Passenger Class Distribution**

* 

**Observation:**  
Most passengers belonged to 3rd class.

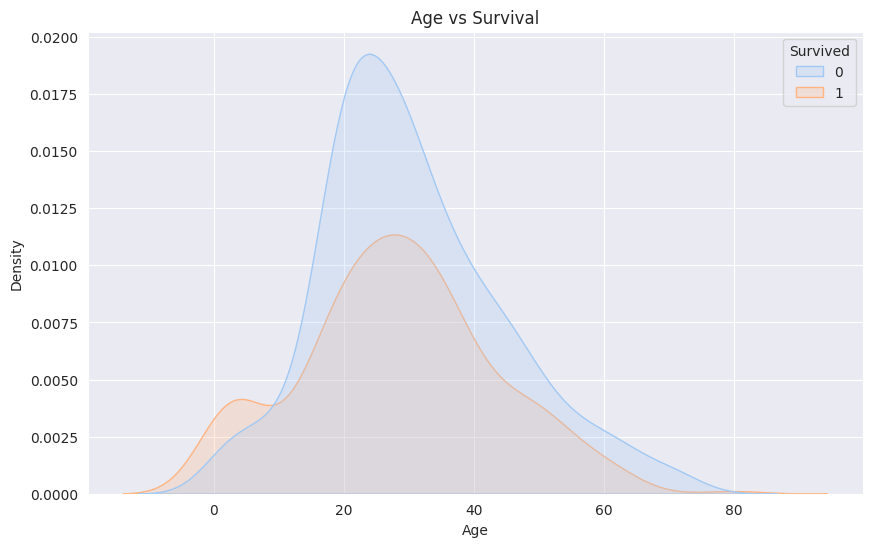
**3.3 Bivariate Analysis**

**a) Survival Rate by Gender**

* 

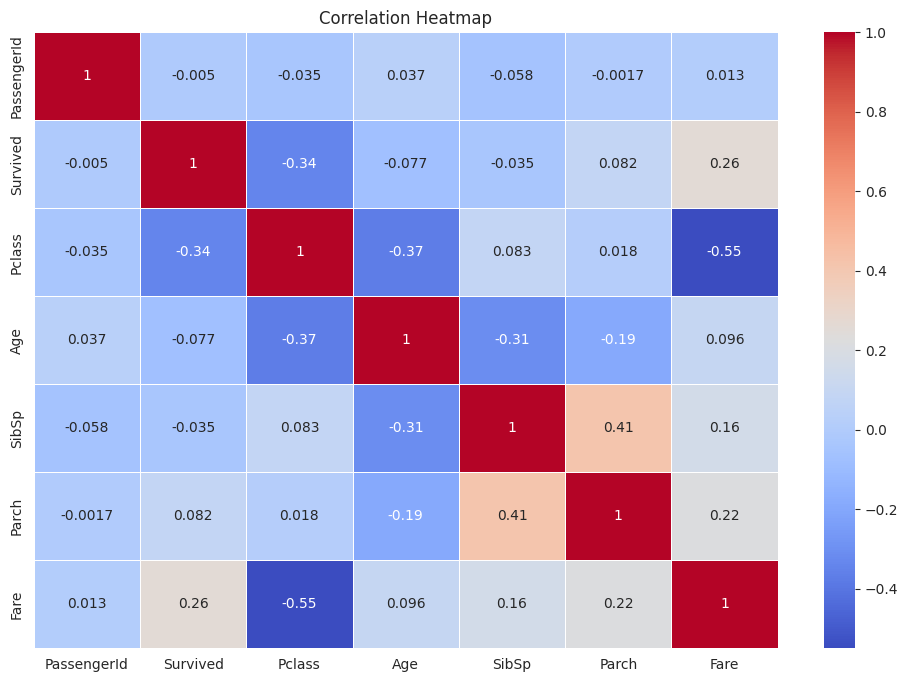
**Observation:**  
Females had a much higher survival rate.

**b) Age vs Survival**

* 

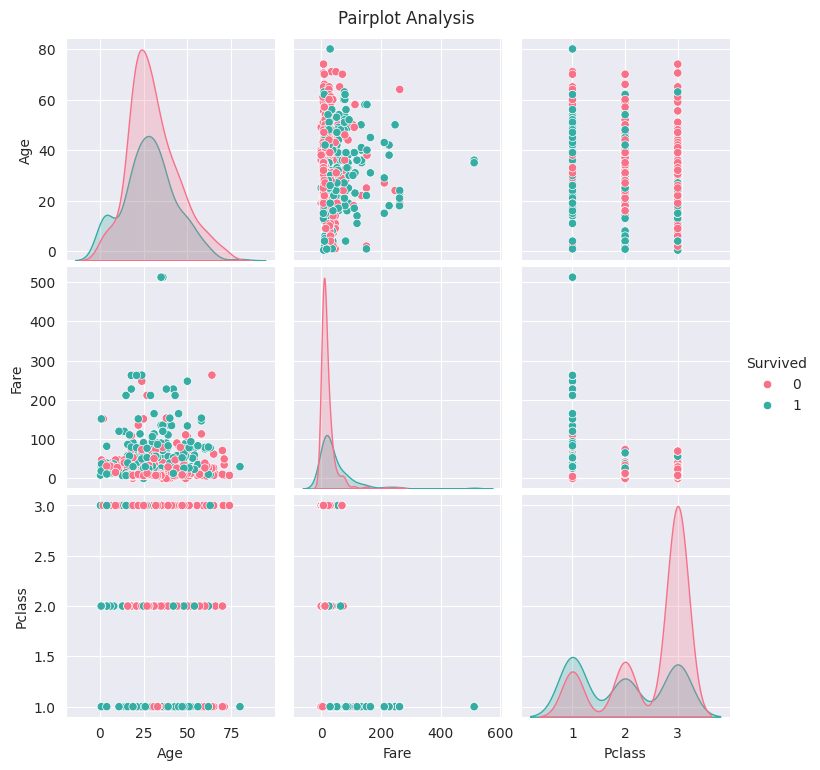
**Observation:**  
Young children had better survival odds.

**c) Correlation Heatmap**

* 

**Observation:**  
Fare and Sex are highly correlated with survival.

**d) Pairplot**

* 

**Observation:**  
1st Class passengers and higher Fare values are linked to higher survival rates.

**4. FINAL SUMMARY OF FINDINGS**

**Key Insights:**

* Young adults (20–40) were the majority of passengers.
* Survival favored females, young children, and 1st class passengers.
* Higher Fare linked to higher survival probability.
* 3rd class passengers had the lowest survival rates.