### 🚢 Titanic Dataset EDA Report

**Dataset Source:** [Kaggle Titanic - Machine Learning from Disaster](https://www.kaggle.com/c/titanic/data)

## ✅ Step 1: Load & Understand the Dataset

We use the train.csv file, which contains data of ~891 passengers.

### 📄 Dataset Columns

| **Column** | **Description** |
| --- | --- |
| PassengerId | ID of the passenger |
| Survived | 0 = No, 1 = Yes (Target) |
| Pclass | Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) |
| Name | Name of passenger |
| Sex | Gender |
| Age | Age in years |
| SibSp | # of siblings / spouses aboard |
| Parch | # of parents / children aboard |
| Ticket | Ticket number |
| Fare | Passenger fare |
| Cabin | Cabin number |
| Embarked | Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) |

## ✅ Step 2: Data Cleaning

### 🔍 Missing Values

| **Column** | **Missing Count** |
| --- | --- |
| Age | 177 |
| Cabin | 687 |
| Embarked | 2 |

### 🧹 Cleaning Strategy

**1.Age**: Impute with **median** age

**2.Cabin**: Drop column (too sparse)

**3.Embarked**: Impute with **mode** ('S')

import pandas as pd

df = pd.read\_csv("train.csv")

# Impute Age

df['Age'].fillna(df['Age'].median(), inplace=True)

# Impute Embarked

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# Drop Cabin

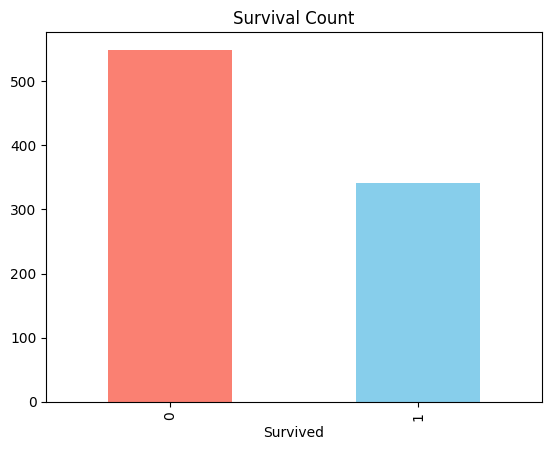
df.drop(columns=['Cabin'], inplace=True)

## ✅ Step 3: Exploratory Data Analysis (EDA)

### 🎯 1. Target Distribution: Survival

df['Survived'].value\_counts().plot(kind='bar', title='Survival Count', color=['salmon', 'skyblue'])

**Result**: ~38% Survived, ~62% Did Not

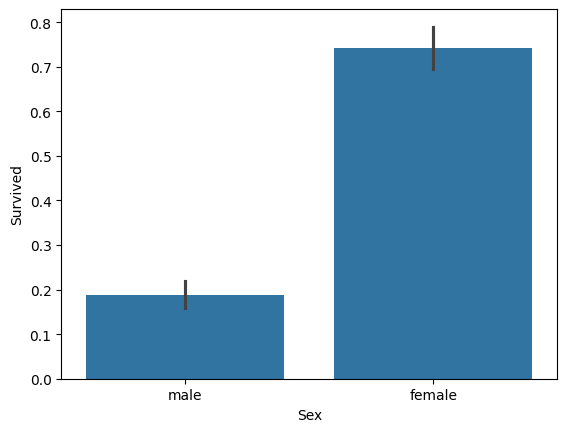


### 👥 2. Survival by Gender

import seaborn as sns

sns.barplot(x='Sex', y='Survived', data=df)

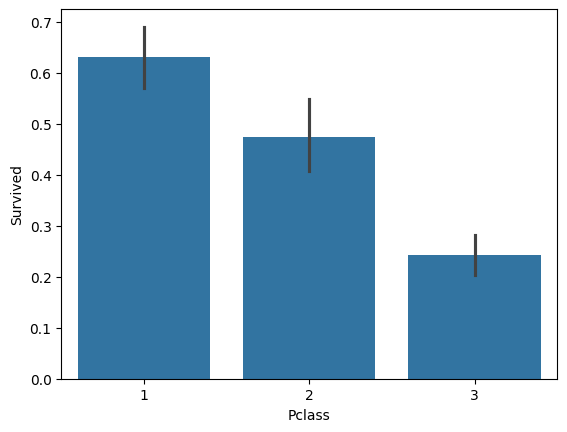
**Insight**: Females had a much higher survival rate (~74%) compared to males (~19%).



### 💼 3. Survival by Passenger Class (Pclass)

sns.barplot(x='Pclass', y='Survived', data=df)

**Insight**: 1st class passengers had highest survival (~63%), 3rd class lowest (~24%).

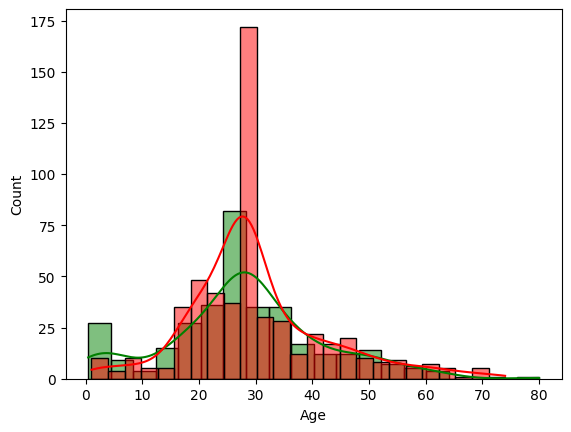


### 🧒 4. Age Distribution & Survival

sns.histplot(df[df['Survived']==1]['Age'], color='green', label='Survived', kde=True)

sns.histplot(df[df['Survived']==0]['Age'], color='red', label='Not Survived', kde=True)

**Insight**: Children had better survival rates; many elderly did not survive.



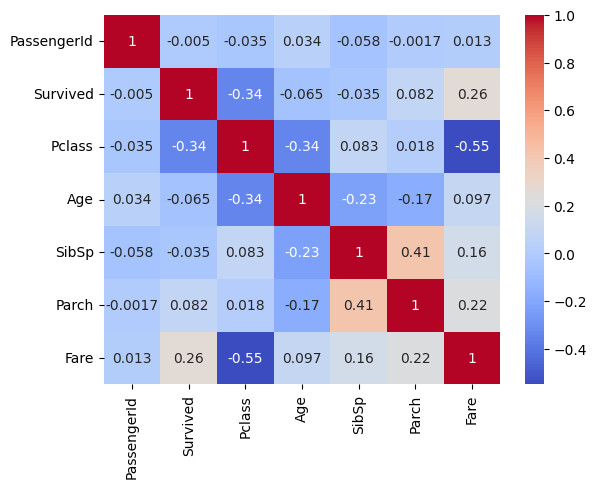
### 🧾 5. Correlation Heatmap

import matplotlib.pyplot as plt

import numpy as np

corr = df.corr(numeric\_only=True)

sns.heatmap(corr, annot=True, cmap='coolwarm')



**Insight**:

Fare and Pclass negatively correlated.

Pclass and Survived are moderately negatively correlated.

Age weakly correlated with Survived.

## ✅ Step 4: Key Business Insights (Internship Deliverable)

**1.Females & children in 1st class had highest survival rates** → could inform priority protocols.

**2.Ticket class (Pclass) and Fare had strong ties to survival** → shows socio-economic influence.

**3.Embarkation point** also had a subtle relationship with survival — most survivors boarded from **Cherbourg (C)**.

4.Columns like **Cabin and Ticket** hold little predictive power in raw form and need heavy feature engineering.