Exploratory Data Analysis on Titanic Dataset

Objective: The goal of this analysis is to explore the Titanic dataset, understand the underlying patterns and relationships between passenger features (such as age, gender, class, and fare) and their survival status. This helps in identifying key factors influencing survival chances on the Titanic.

Tools Used:

Python

Pandas – For data manipulation and preprocessing

Seaborn – For data visualization

Matplotlib – For plotting graphs

Data Overview:

The Titanic dataset contains information about passengers aboard the Titanic ship that sank in 1912.

It includes features like Passenger ID, Survival status, Gender, Age, Class, Fare, and more.

Feature	Description
PassengerId	Unique ID for each passenger
Survived	0 = Not Survived, 1 = Survived
Pclass	Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Passenger's name
Sex	Gender
Age	Age of the passenger
SibSp	No. of siblings/spouses aboard
Parch	No. of parents/children aboard
Ticket	Ticket number
Fare	Amount paid for ticket
Cabin	Cabin number
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Shape of Data: 891 rows × 12 columns

Null Value Summary:

Column	Missing Values
Age	177
Cabin	687
Embarked	2

Observation:

1. Age and Cabin have significant missing values.
2. Cabin is mostly empty and can be dropped.
3. Age can be filled using the median.
4. Embarked can be filled using the mode ('S').

Univariate Analysis:

Survival Count:

- Around 38% of passengers survived.
- Class imbalance exists (more non-survivors).

Gender Distribution:

- More males were on board than females.
- However, females had a higher survival rate.

Passenger Class:

- Majority of passengers belonged to 3rd class.

Fare Distribution:

- Fare is right-skewed with a few high-paying passengers.

Bivariate Analysis:

Survival by Gender:

- Females had a much higher survival rate than males.

Survival by Class:

- 1st class passengers had better survival chances than 2nd and 3rd.

Survival by Age:

- Younger passengers had a slightly higher survival rate.

Fare vs Survival:

- Passengers who paid higher fares were more likely to survive.

Correlation Analysis:

Feature 1	Feature 2	Correlation
Survived	Fare	+0.26
Survived	Pclass	-0.34
Pclass	Fare	-0.55

Observation:

- Higher fares are positively correlated with survival.
- Pclass has a negative correlation (1st class passengers survived more).
- Other correlations are weak.

Key Findings & Conclusion:

Key Insights:

1. Females and 1st class passengers had the highest survival rate.
2. Passengers who paid higher fares had better chances of survival.
3. Age shows a weak relation with survival — younger passengers slightly favored.
4. Cabin feature had too many nulls, hence dropped.
5. Embarked feature was mostly ‘S’ (Southampton).

Conclusion:

The EDA revealed strong links between gender, class, and fare with survival outcomes. These insights can help in building predictive models to estimate survival probability.