Exploratory Data Analysis (EDA) Report - Titanic Dataset

1. Dataset Overview

The Titanic dataset contains 891 rows and 12 columns. Important variables include Survived (target), Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.

Key Findings:

- Age has 177 missing values, Cabin has ~77% missing, and Embarked has 2 missing values.
- Average Age ≈ 30 years; Fare is highly skewed with extreme outliers.
- Majority passengers: Male (577), 3rd class (491), embarked from Southampton (644).

2. Relationships and Trends

- Pairplot analysis shows Survival strongly linked with Sex, Pclass, and Fare.
- Heatmap reveals positive correlation between Fare and Survival, and negative correlation between Pclass and Survival.

3. Visual Observations

- Age distribution: Most passengers were between 20–40 years; slightly right-skewed.
- Fare distribution: Strong right skew with high-value outliers.
- Scatterplot (Age vs Fare): Higher Fare passengers, usually in 1st class, had higher survival rates.
- Boxplot (Pclass vs Age): 1st class passengers tended to be older, while 3rd class had more young adults and children.

4. Summary of Findings

- Dataset has 891 rows and 12 columns; Age, Cabin, and Embarked have missing values.
- Overall survival rate ≈ 38%.
- Higher survival likelihood for females, 1st class passengers, and those who paid higher fares.
- Strong correlation between Survival and Sex, Pclass, and Fare.
- Weak correlation with SibSp & Parch.
- Cabin data too sparse for reliable insights.