Exploratory Data Analysis Report - Titanic Dataset

Basic Dataset Overview

- Total Rows: 891

- Columns: 12

- Age has missing values (714/891 present)

- Cabin column has high missing values (204/891)

- Embarked has 889 non-null entries

Columns:

- Numerical: Age, SibSp, Parch, Fare

- Categorical: Survived, Pclass, Sex, Embarked, Ticket, Cabin

Descriptive Statistics

- Most passengers are aged between 20-40.

- Fare varies widely; skewed with high outliers.

- 577 males, 314 females.

- Class distribution: 491 in 3rd, 216 in 1st, 184 in 2nd.

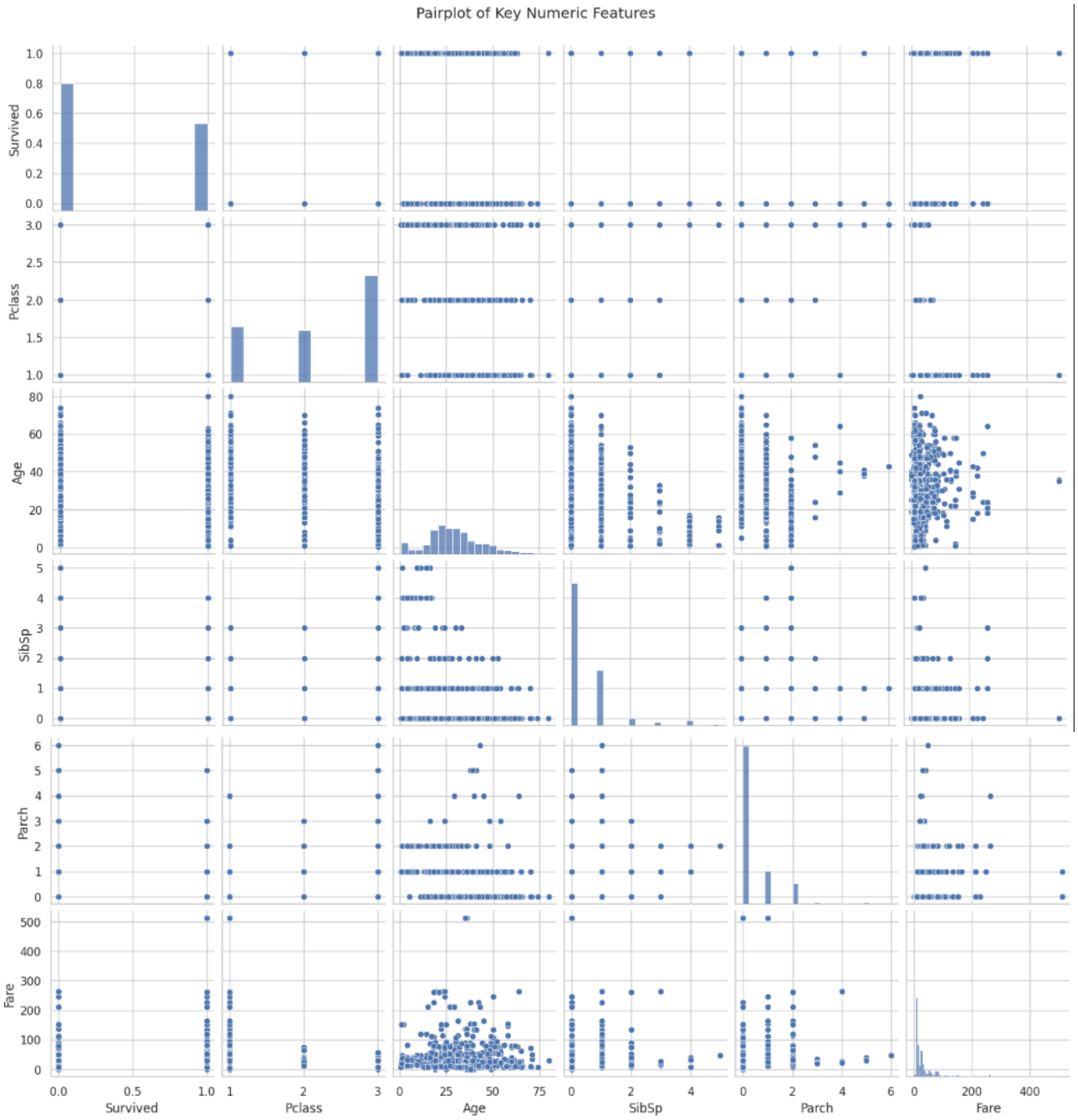
- Embarked: Most from 'S' (644), then 'C' (168), 'Q' (77).

Visual Insights

1. Pair plot:

- Age and Fare are right-skewed.

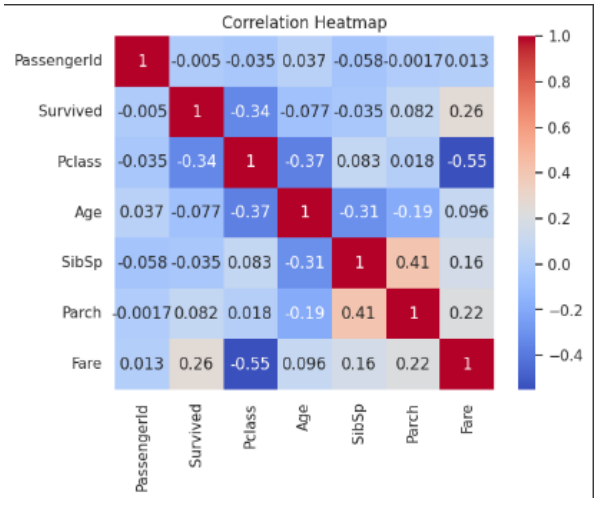
- Survived is a binary variable.



2. Heat map Correlation:

- Fare and Pclass are negatively correlated.

- Survived positively correlates with Fare, negatively with Pclass.

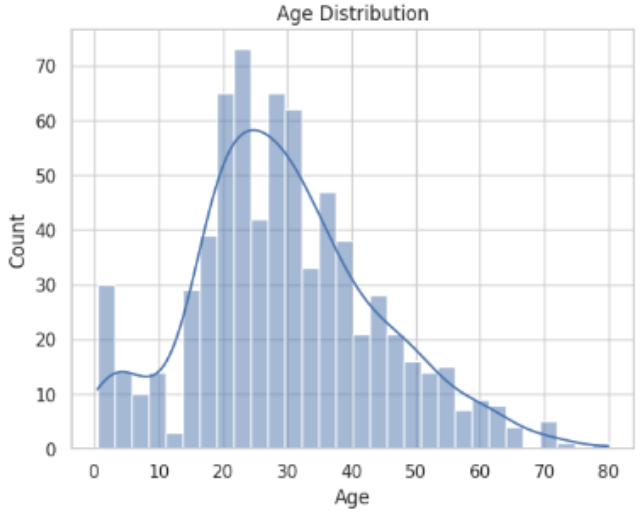


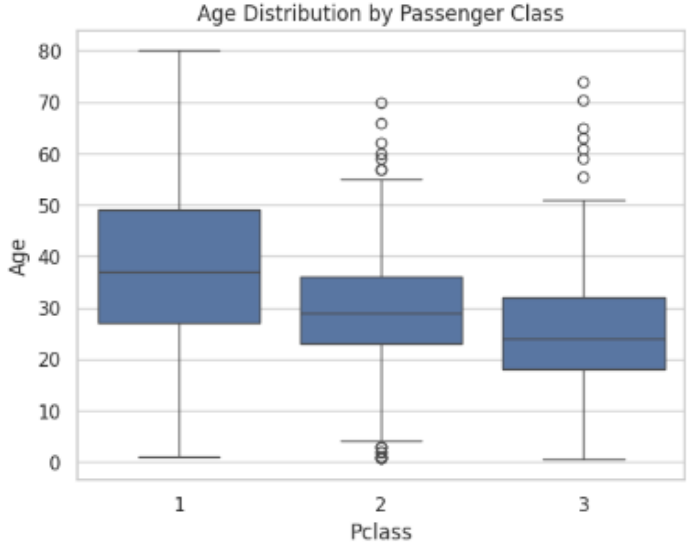
3. Histograms and Box plots:

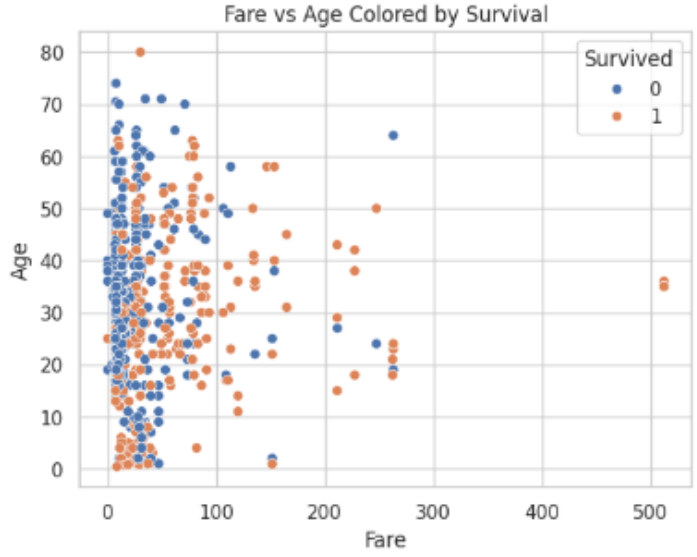
- Age: Peaks at 20-30 years.

- Class vs Age: 1st class older on average.

- Fare vs Age scatterplot shows higher fare linked to better survival.







4. Grouped Survival Rate:

- Female survival: 74%

- Male survival: 19%

- Pclass survival: 1st (63%), 2nd (47%), 3rd (24%)

- Embarked 'C' has highest survival rate (55%)

Summary of Findings

1. Females had a significantly higher survival rate than males.

2. 1st class passengers were more likely to survive than 2nd or 3rd class.

3. Higher fare was positively associated with survival.

4. Younger children had a better chance of survival.

5. Most passengers were in the 20-40 age range.

6. Age and Fare are not normally distributed - they are skewed.

7. Embarked location shows some influence on survival, especially from port 'C'.