Exploratory Data Analysis Report: Titanic Dataset

Executive Summary

This report explores the Titanic dataset containing details of passengers aboard the ill-fated RMS Titanic.

The goal is to identify factors that influenced survival outcomes.

- Dataset Shape: 891 rows × 12 columns

- Target Variable: Survived (0 = Died, 1 = Survived)

- Important Features: Sex, Pclass, Age, Fare, Embarked, etc.

- Missing Values: Found in Age, Cabin, Embarked

Data Overview

Feature Description

PassengerId - Unique ID of each passenger

Survived - Survival status (0 = No, 1 = Yes)

Pclass - Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)

Name - Name of the passenger

Sex - Gender of the passenger

Age - Age in years

SibSp - # Siblings / spouses aboard

Parch - # Parents / children aboard

Ticket - Ticket number

Fare - Ticket fare

Cabin - Cabin number

Embarked - Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Univariate Analysis

- Survival Count: ~38% survived, ~62% did not.

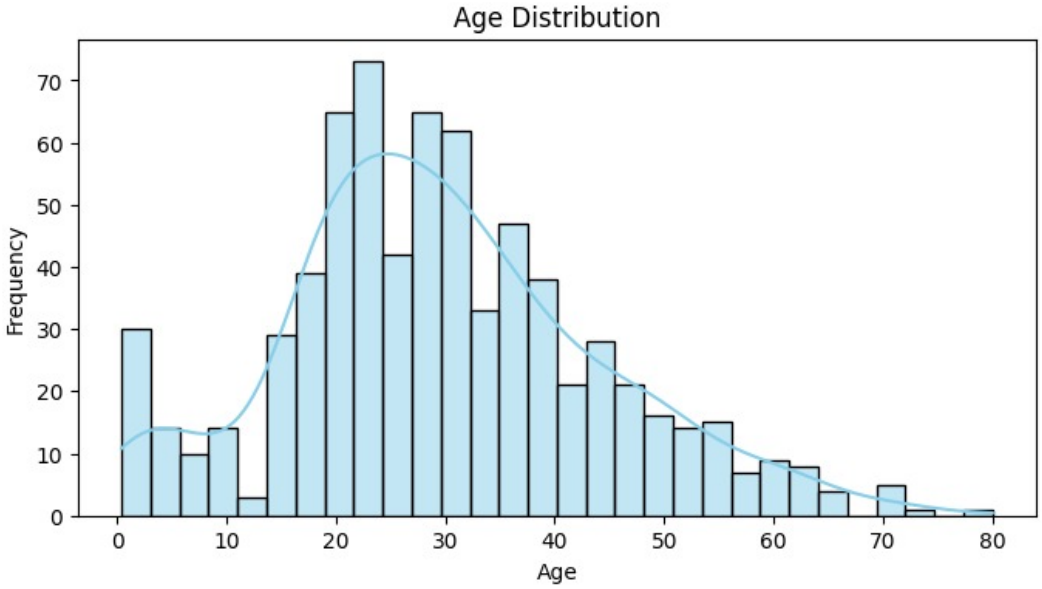
- Gender: More males than females.

- Passenger Class: Majority belonged to 3rd class.

- Embarked: Most passengers boarded at Southampton (S).

- Age Distribution: Majority between 20-40 years.

- Fare: Right-skewed distribution; a few paid very high fares.



Bivariate Analysis

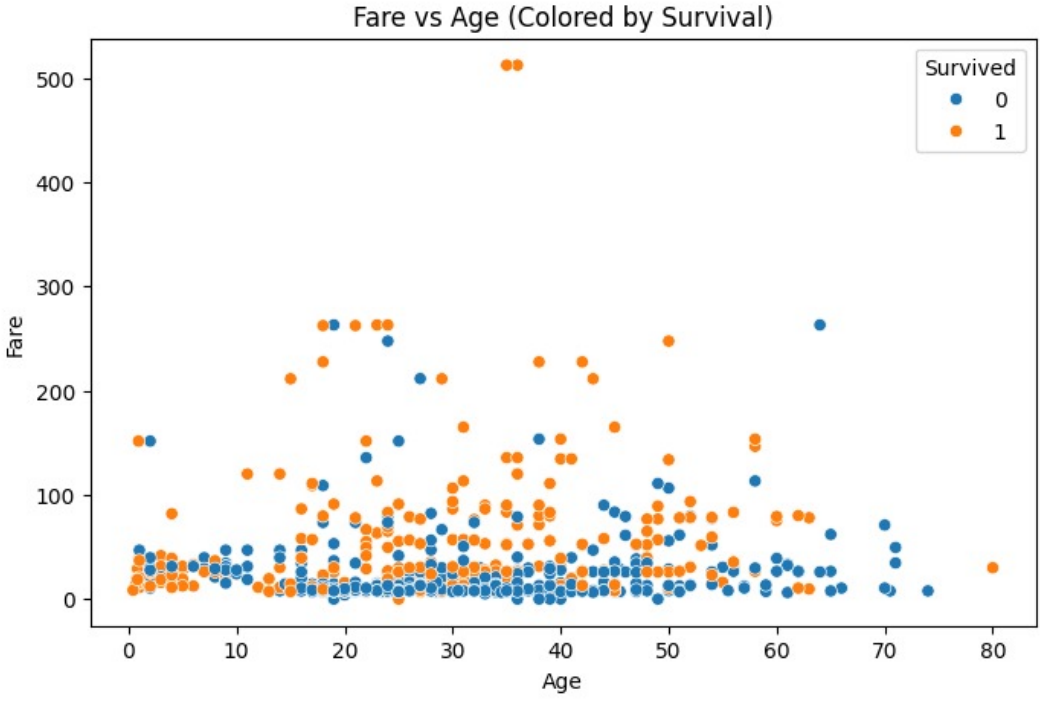
- Survival by Gender: Females had much higher survival rates.

- Survival by Class: 1st class passengers had the highest survival probability.

- Survival by Embarkation: Cherbourg (C) passengers had better survival rates.

- Age vs Survival: Younger passengers, especially under age 10, were more likely to survive.

- Fare vs Survival: Those who paid higher fares were more likely to survive.



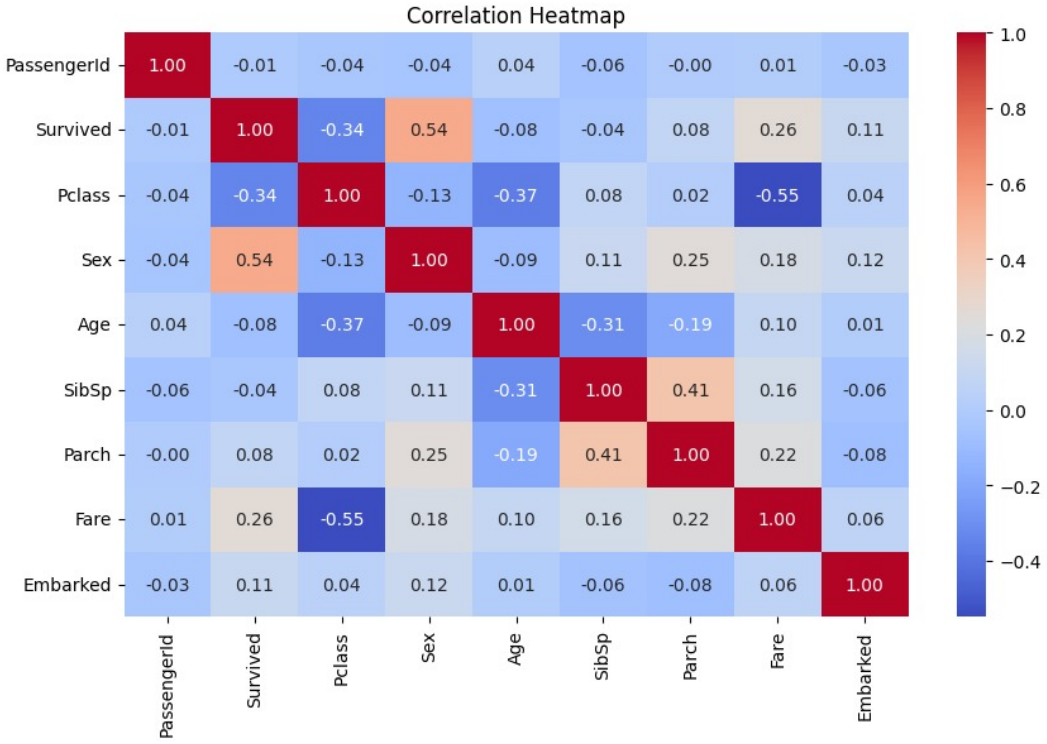
Multivariate Insights

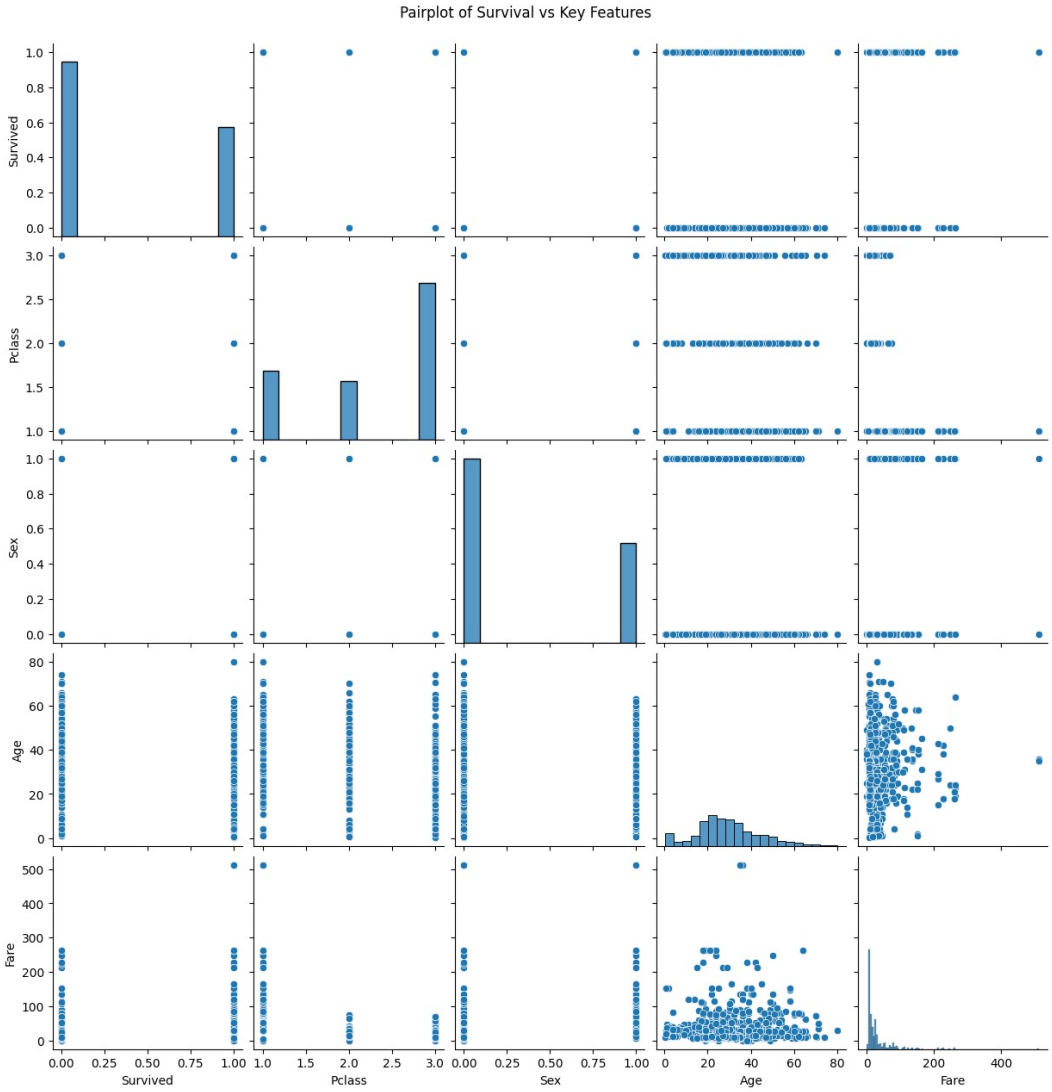
- Pairplot: Clear distinction in survival based on Fare and Pclass.

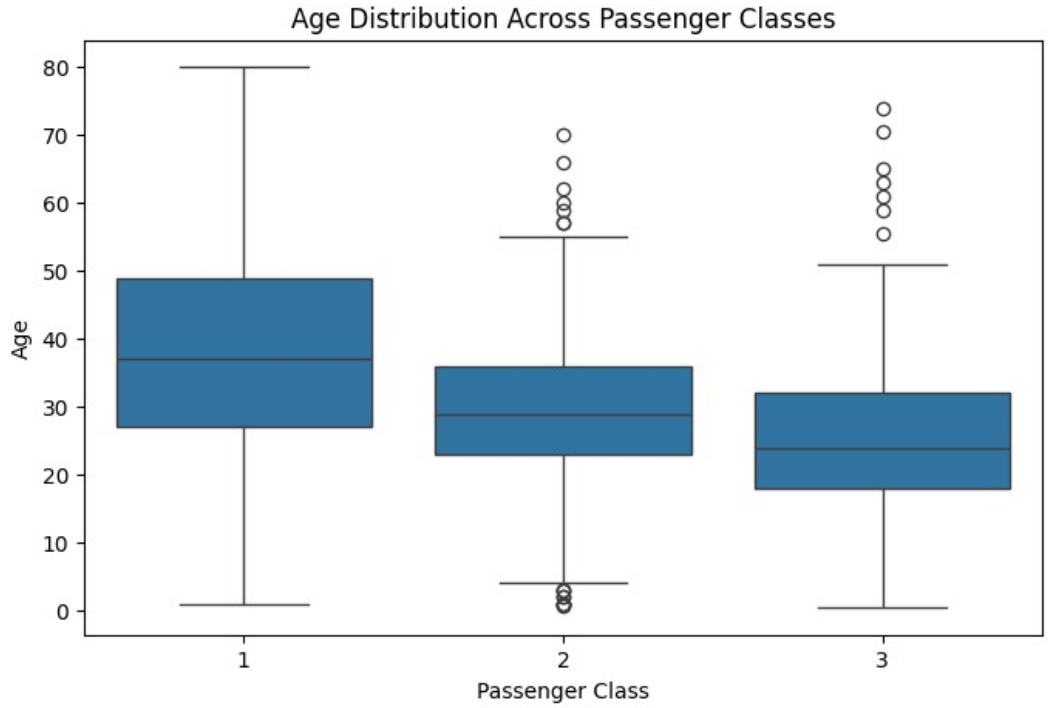
- Heatmap Correlation:

  - Moderate positive correlation between Fare and Survived

  - Weak correlation between SibSp, Parch and survival







Recommendations for Further Analysis

- Create new features like FamilySize, IsAlone, Title from Name

- Handle missing Age values with imputation

- Explore model building: Logistic Regression, Decision Trees, Random Forest

- Try dimensionality reduction or feature selection for model efficiency

Key Takeaways

Gender - Women had higher survival rates

Class - Higher class passengers had better survival odds

Fare - Passengers who paid more were prioritized

Children - Young children had higher survival

Embarkation Port - Cherbourg passengers had better survival rates