Exploratory Data Analysis Report – Titanic Dataset

Introduction

This report presents findings from exploratory data analysis (EDA) of the Titanic dataset. The goal was to uncover patterns, relationships, and trends affecting passenger survival, which will help in building predictive models.

Dataset Overview

Source: Kaggle Titanic
Dataset Rows: 891
Columns: 12
Target Variable: Survived (0 = Did not survive, 1 = Survived)
Key Features: Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

Data Quality

Missing values:
Age: 177
Cabin: 687
Embarked: 2
Outliers in Fare (skewness □ 4.79)

Univariate Analysis

Survival Rate: 38.38%
Passenger Class: 3rd Class = 491, 1st Class = 216, 2nd Class = 184
Gender Distribution: Male = 577, Female = 314
Median Age: 28.0 years
Fare: Highly right-skewed

Bivariate Analysis

- Females had much higher survival rates than males across all classes.
- 1st Class passengers had the highest survival rate, 3rd Class the lowest.
- Higher fares strongly correlated with better survival chances.
- Children (age < 10) had significantly higher survival odds than adults and elderly.
- Passengers embarking from port 'C' had the highest survival rate, while those from 'S' had the lowest.
- Very small families (FamilySize 2–4) survived better than solo travelers or large families.
- Within each class, women consistently survived at higher rates than men.
- Fare distribution showed large variation in 1st Class, moderate in 2nd Class, and small in 3rd Class.

Correlation Analysis

Top correlations with survival:
Fare: +0.25

Parch: +0.082
Pclass: -0.338
Age: weak negative correlation
Strong negative correlation between Pclass and Survived.

Summary of Insights

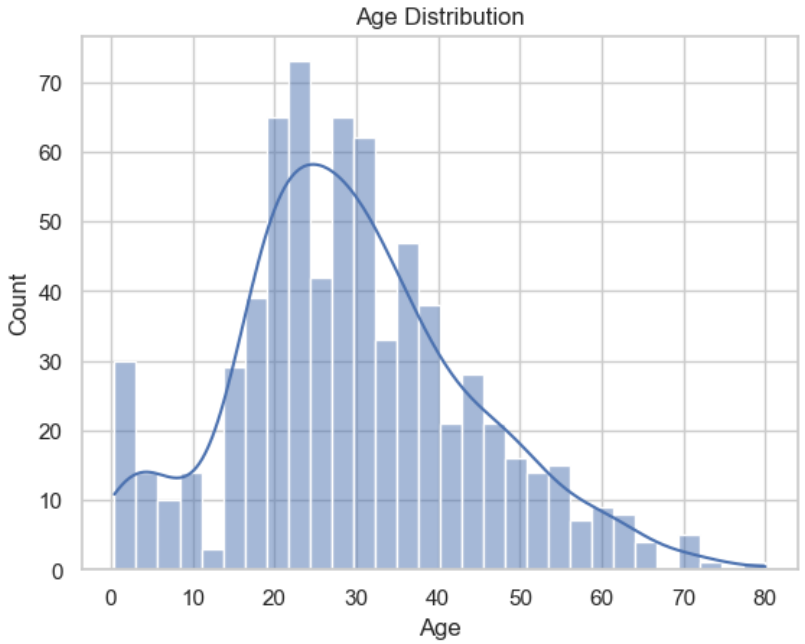
- Gender, class, and fare are the strongest predictors of survival.
- Age plays a significant role for children — they survived more often than adults.
- Port of embarkation impacted survival rates, with 'C' being most favorable.
- Medium-sized families survived best; very large families and solo travelers fared worse.
- Fares vary widely by class, influencing survival odds.
- Cabin data is highly incomplete and unsuitable for direct modeling.
- Outliers in Fare should be handled before modeling.

Recommendations

- Impute missing ages using median or predictive methods.
- Drop or categorize Cabin due to high missingness.
- Normalize or log-transform Fare to handle skewness.
- Encode categorical variables (Sex, Embarked) for modeling.
- Create engineered features such as Family Size and Title from passenger names.

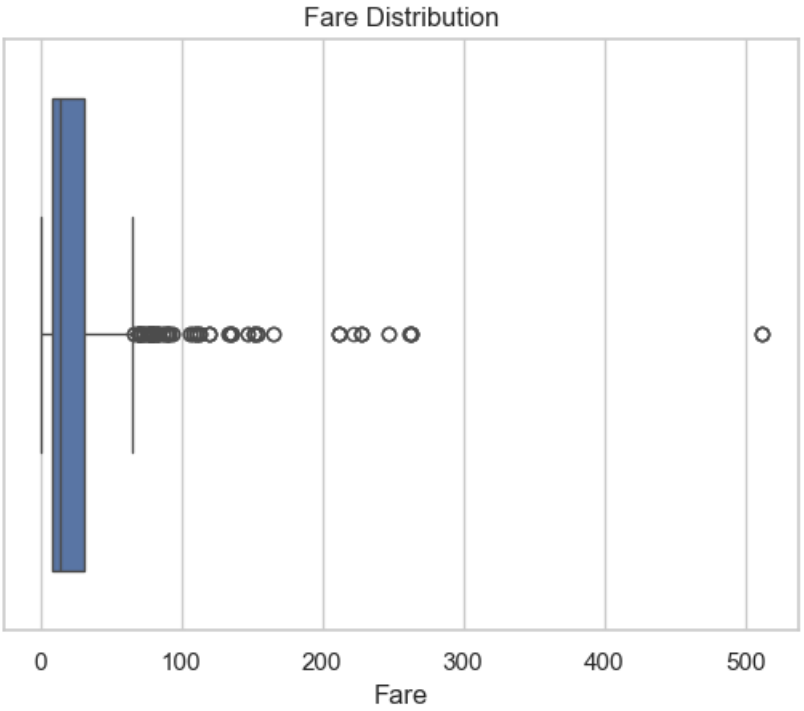
Key Visuals

Age Distribution



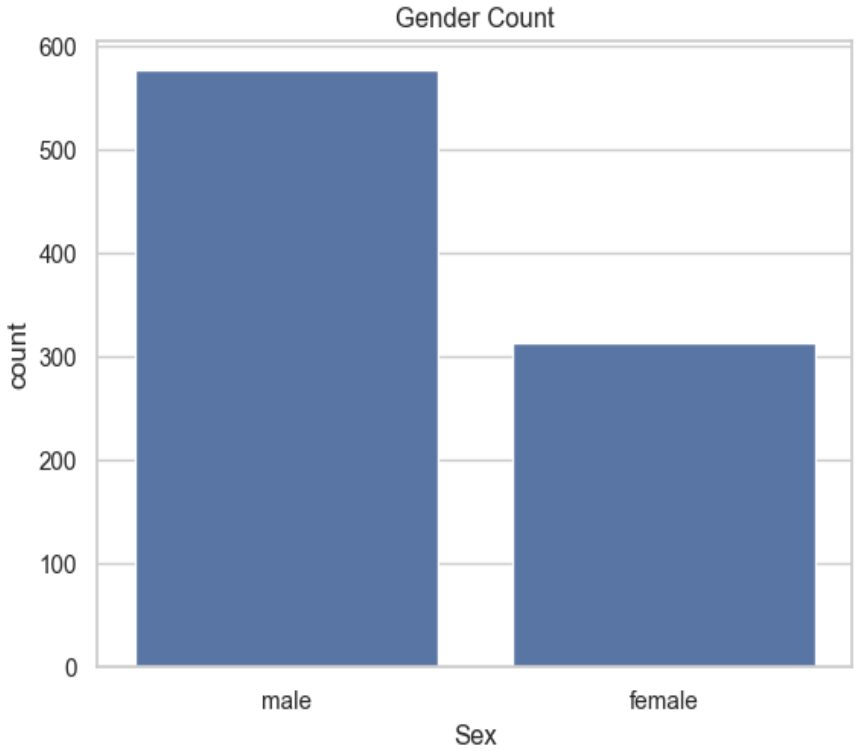
Most passengers were between 20–40 years old, with a noticeable number of children. The distribution is slightly right-skewed due to elderly passengers

Fare Distribution



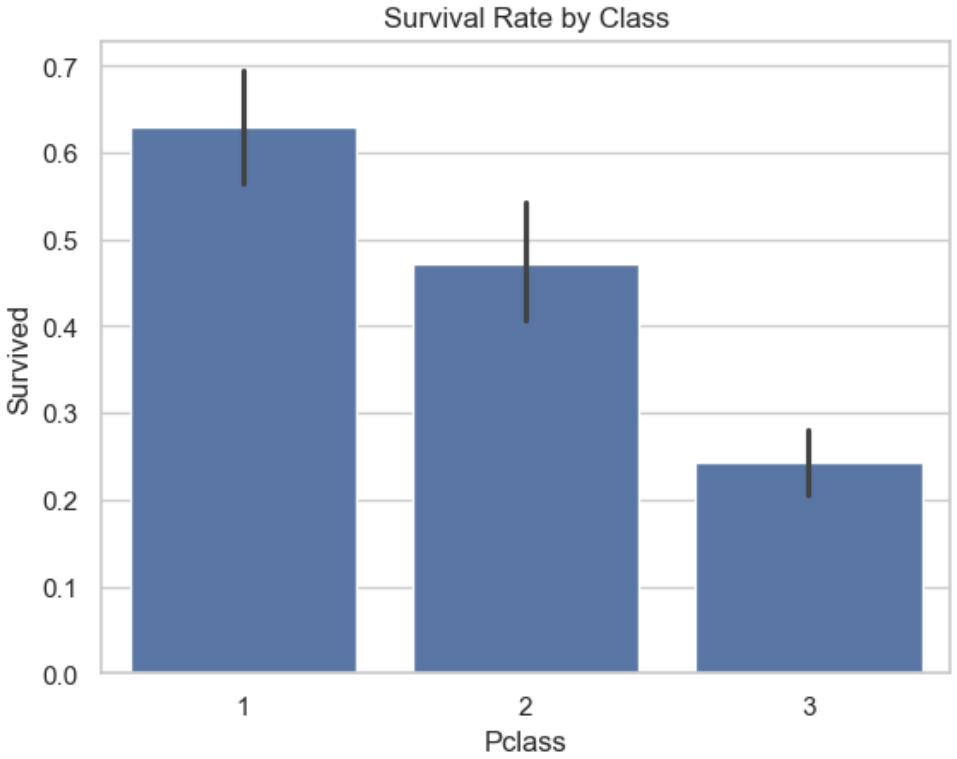
Fares are highly skewed to the right, with most passengers paying less than $50. A few high outliers represent wealthier passengers, often in 1st Class.

Gender Count



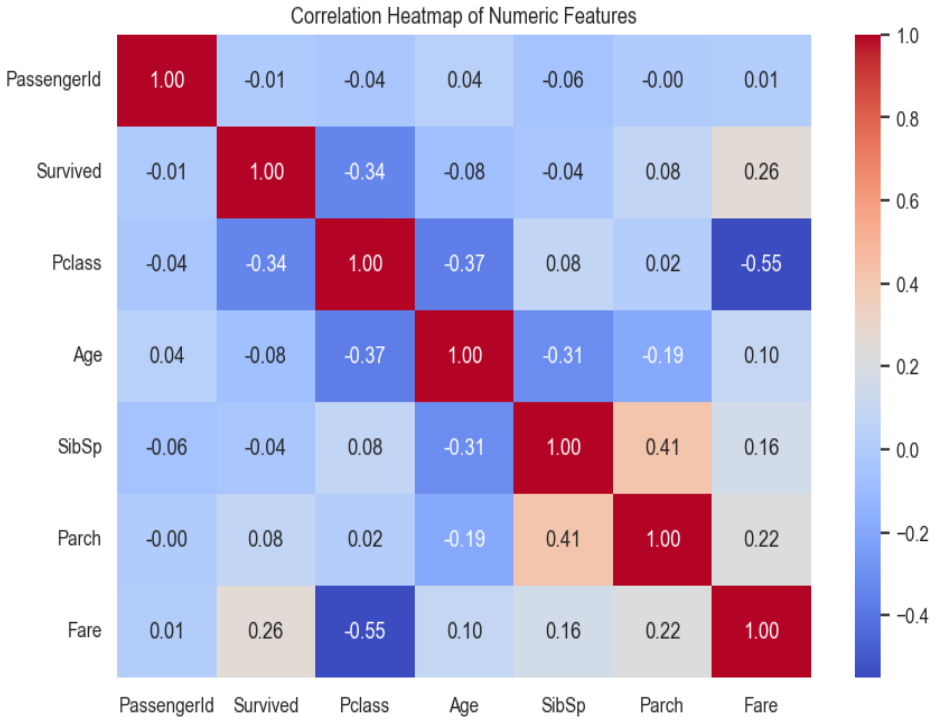
There were significantly more male passengers (577) than female passengers (314) on board.

Survival by Passenger Class



Survival rates were highest in 1st Class and lowest in 3rd Class, showing a clear link between social class and survival odds.

Correlation Heatmap



Fare shows a positive correlation with survival, while Pclass shows a strong negative correlation. Other features like Age have weaker relationships with survival.