Exploratory Data Analysis (EDA) on Titanic Dataset

1. Introduction

The Titanic dataset provides information on the passengers who boarded the Titanic, including whether they survived or not. The objective of this EDA is to extract insights using statistical and visual exploration techniques.

2. Dataset Overview

The dataset contains the following columns:

Column	Description
PassengerId	Passenger ID
Survived	Survival Status (0 = No, 1 = Yes)
Pclass	Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Name of the Passenger
Sex	Gender
Age	Age in Years
SibSp	Number of Siblings/Spouses Aboard
Parch	Number of Parents/Children Aboard
Ticket	Ticket Number
Fare	Passenger Fare

3. Data Information

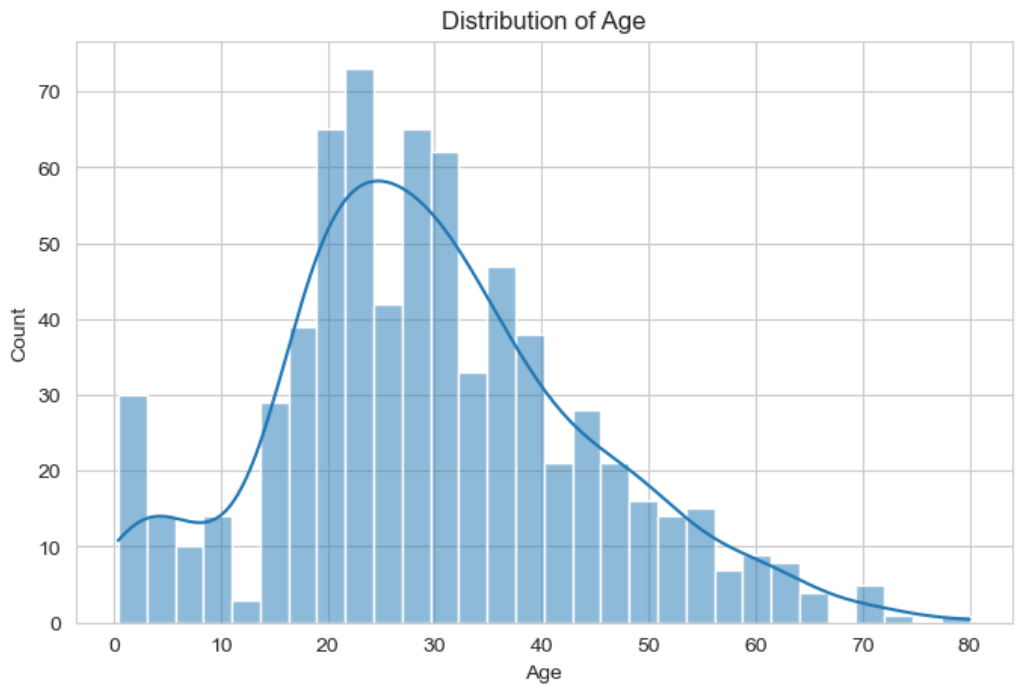
- Total Entries: 891
- Features: 10
- Missing Values:
  - Age: Some missing values
  - Cabin: Missing in many rows (if present)
  - Embarked: A few missing values
- Data Types: Both numerical and categorical features

4. Univariate Analysis

4.1 Age Distribution

- The Age distribution is right-skewed.
- Most passengers were young adults between 20 to 40 years.
- There are some missing values.

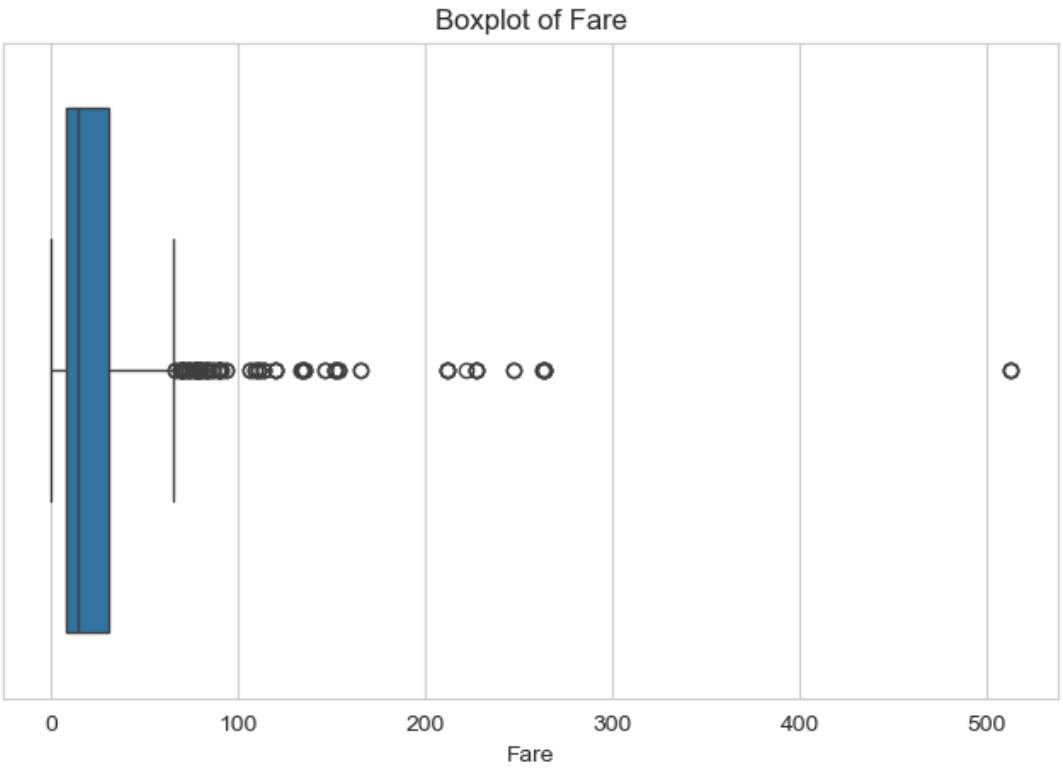
Visualization Used: Histogram



4.2 Fare Distribution

- Fare shows a highly right-skewed distribution.
- Some passengers paid significantly high fares (outliers observed).
- Majority of passengers paid fares less than 100.

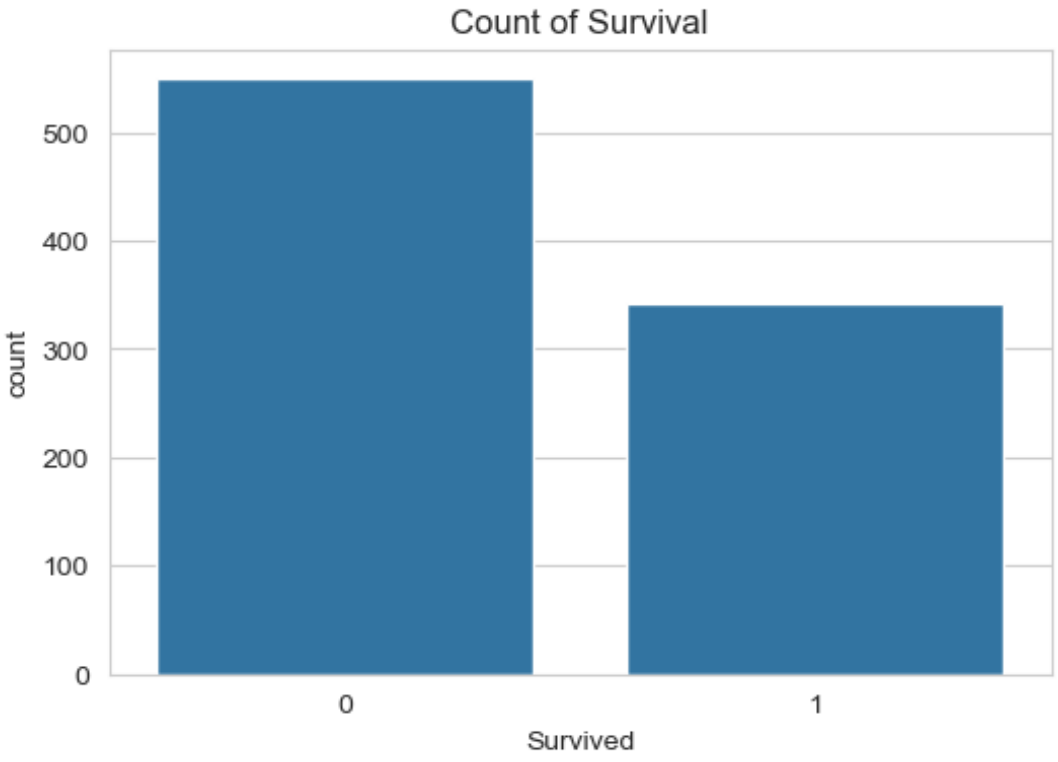
Visualization Used: Boxplot



4.3 Survival Distribution

- Around 62% of passengers did not survive.
- Approximately 38% survived.

Visualization Used: Countplot

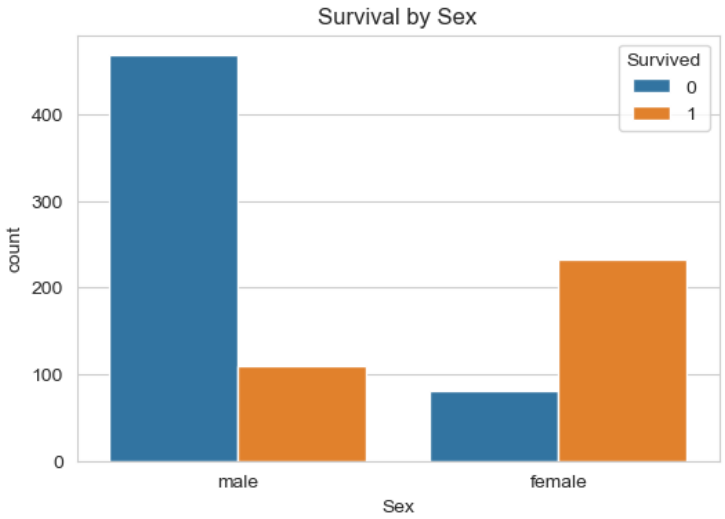


5. Bivariate Analysis

5.1 Survival by Gender

- A significantly higher proportion of females survived compared to males.
- Female passengers had a much better survival rate.

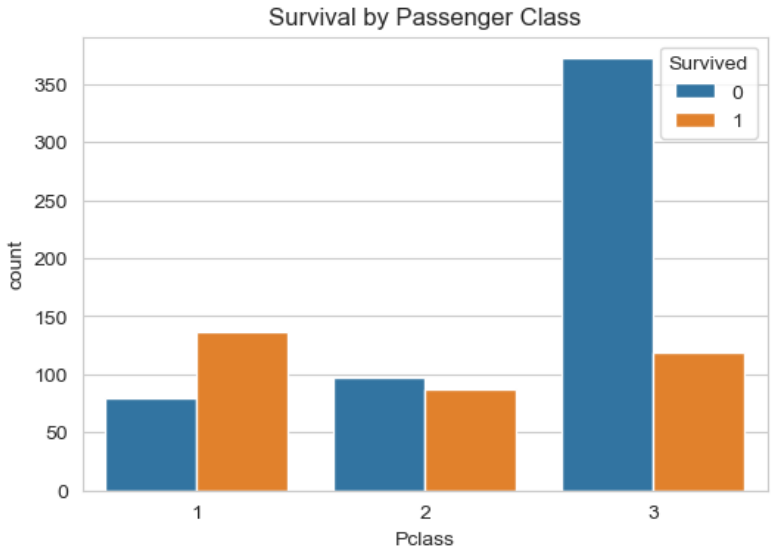
Visualization Used: Countplot (Sex vs Survived)



5.2 Survival by Passenger Class

- First-class passengers had a higher survival rate.
- Third-class passengers had the lowest survival rate.

Visualization Used: Countplot (Pclass vs Survived)

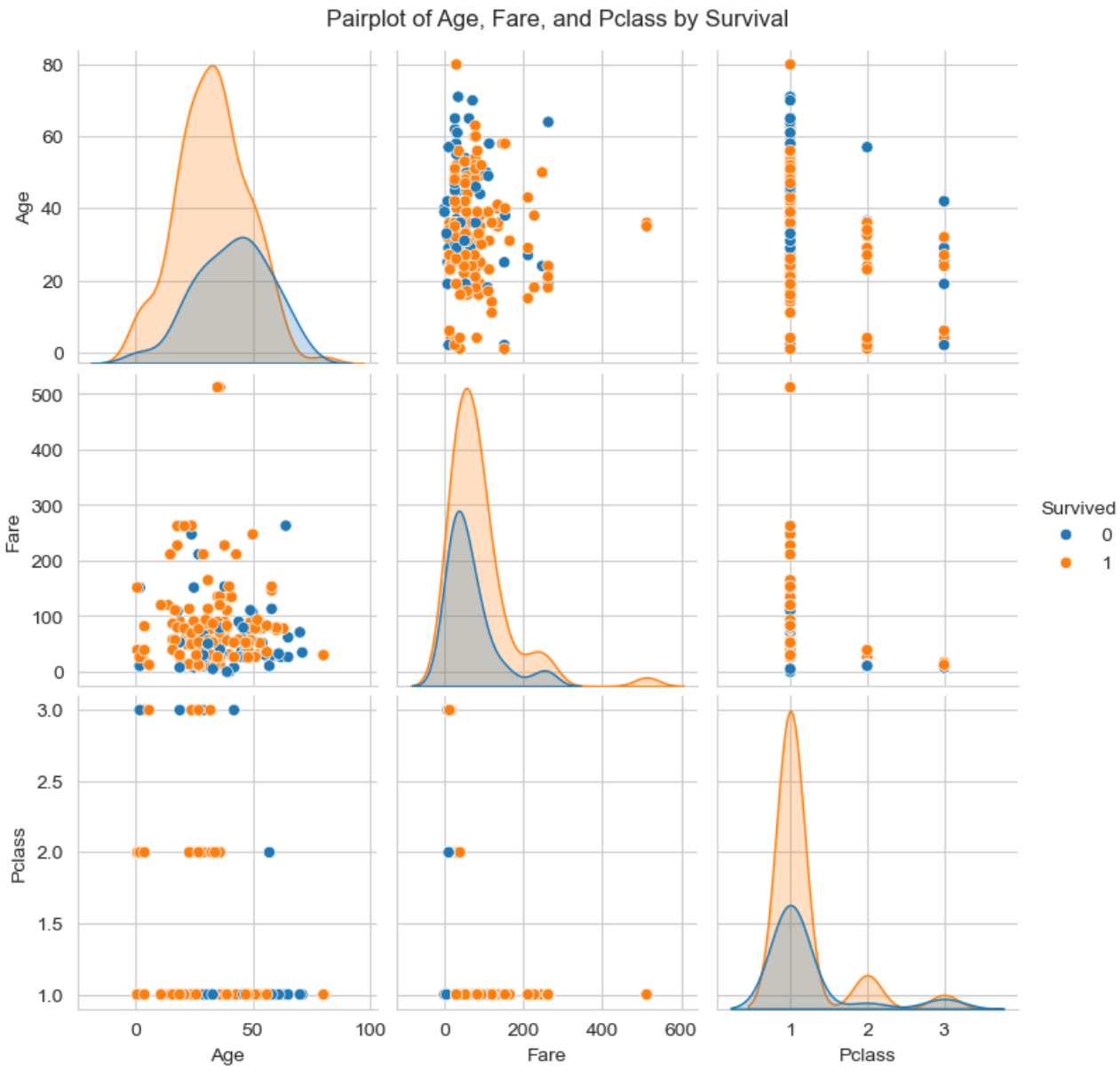


6. Multivariate Analysis

6.1 Pairplot

- Pairplots between Age, Fare, and Pclass show that survivors generally had higher fares and were often from first-class cabins.
- Younger passengers also showed better survival patterns compared to very old passengers.

Visualization Used: Seaborn Pairplot

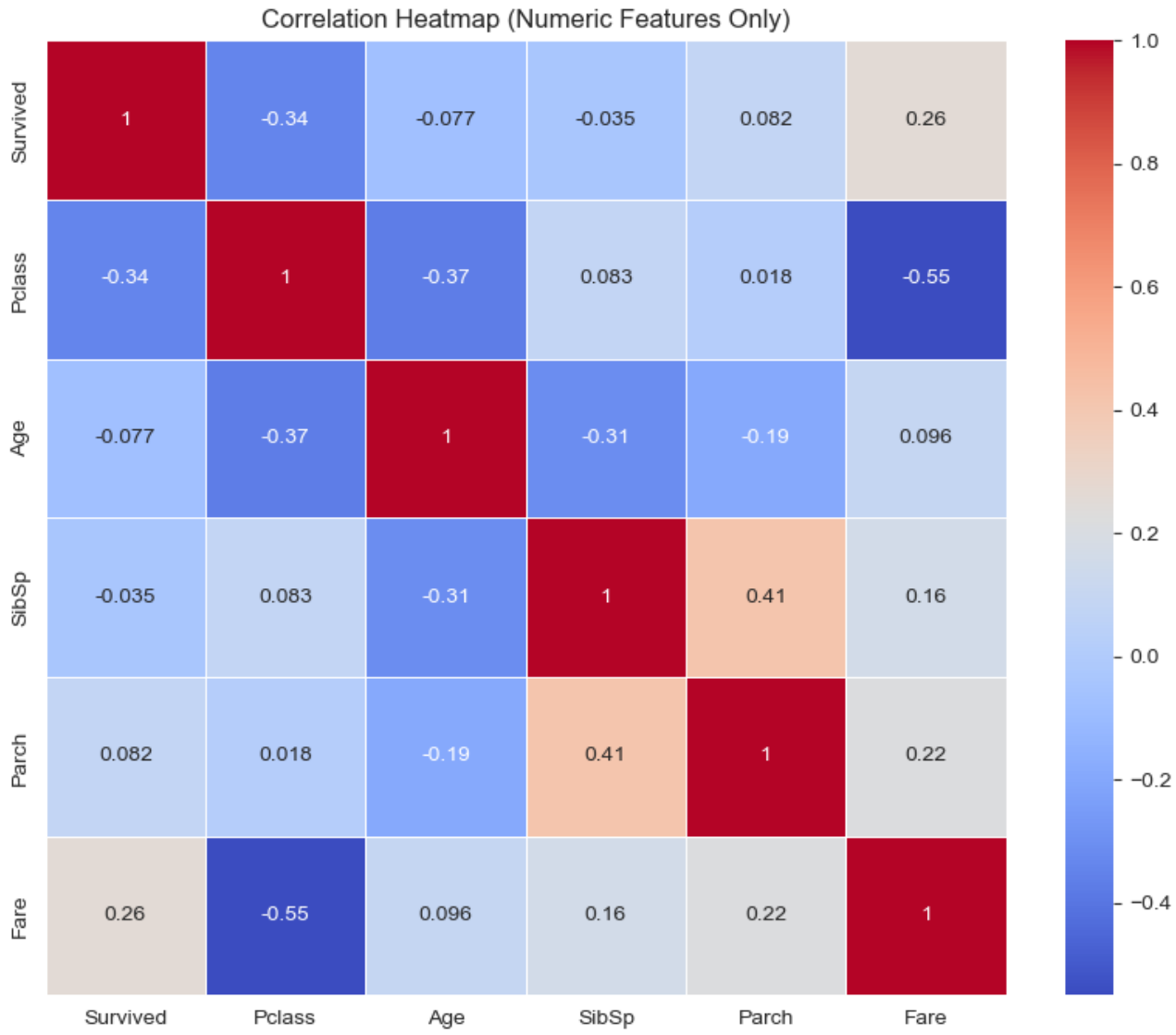


7. Correlation Analysis

7.1 Correlation Heatmap (Numeric Features)

- Positive correlation observed between Fare and Survived.
- Negative correlation observed between Pclass and Survived (lower Pclass number = higher survival chance).
- Sex when numerically encoded (female=1, male=0) shows a positive correlation with survival, confirming females had higher survival rates.

Visualization Used: Correlation Heatmap



8. Observations Summary

- Gender Impact: Female passengers had a much higher survival rate compared to males.
- Class Impact: First-class passengers had a much better survival rate compared to second and third-class passengers.
- Fare: Passengers who paid higher fares generally had better survival chances.
- Age: Younger passengers, especially children, had higher survival rates. Elderly passengers had lower survival chances.
- Family Aboard (SibSp, Parch): Having family members aboard showed mixed impacts, but generally, small family groups had a slightly better survival rate compared to solo travelers.

9. Conclusion

The EDA on the Titanic dataset revealed important patterns and relationships between passenger characteristics and survival rates. Key factors influencing survival were gender, passenger class, and fare paid. These insights provide a solid foundation for building predictive models in future analyses.