Exploratory Data Analysis (EDA) on the Titanic Dataset

1. Introduction

This dataset contains various features such as passenger demographics, class, fare, age, survival status, etc. This analysis aimed to explore and visualize patterns in the data, specifically identifying factors that might have influenced passenger survival.

Dataset Overview

The dataset consists of the following columns:

- **Survived**: Whether the passenger survived (0 = No, 1 = Yes)
- **Pclass**: Passenger class (1 = First class, 2 = Second class, 3 = Third class)
- **Name**: Name of the passenger
- **Sex**: Gender of the passenger (Male/Female)
- **Age**: Age of the passenger
- **SibSp**: Number of siblings or spouses aboard
- **Parch**: Number of parents or children aboard
- **Fare**: Fare paid by the passenger
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

2. Data Preprocessing

Before performing the exploratory analysis, several steps were taken to clean and prepare the dataset:

Handling Missing Data

- **Age**: Missing values in the Age column were filled with the median value of the column.
- **Embarked**: The missing values in the `Embarked` column were filled with the most frequent port (`S` for Southampton).
- **Fare**: No missing values were found in the Fare column.
- **Sex** and **Pclass**: No missing values were found in these categorical columns.

Feature Encoding

- **Sex**: The `Sex` feature was encoded with 0 for male and 1 for female.
- **Embarked**: The `Embarked` column was encoded as:
    - C → 0
    - Q → 1
    - S → 2

Statistical Summary

A statistical summary of the numerical features was generated using the `.describe()` method, which includes key metrics such as mean, standard deviation, minimum, and maximum values for features like Age and Fare.

3. Key Findings and Insights

1. Survival Rate

- The overall survival rate was 38%.
- The **female passengers** had a much higher chance of survival compared to male passengers.

2. Passenger Class and Survival

- **First-class passengers** had a significantly higher survival rate compared to passengers in the other classes.

3. Age and Survival

- **Children** and young adults were more likely to survive compared to older adults.
- The survival rate decreased as the age of passengers increased.

4. Embarked Port

- Passengers who boarded at **Cherbourg** had a significantly higher survival rate compared to passengers who boarded in **Southampton**.

5. Correlation

- The **Fare** and **Pclass** variables were positively correlated, indicating that higher-class passengers paid higher fares.

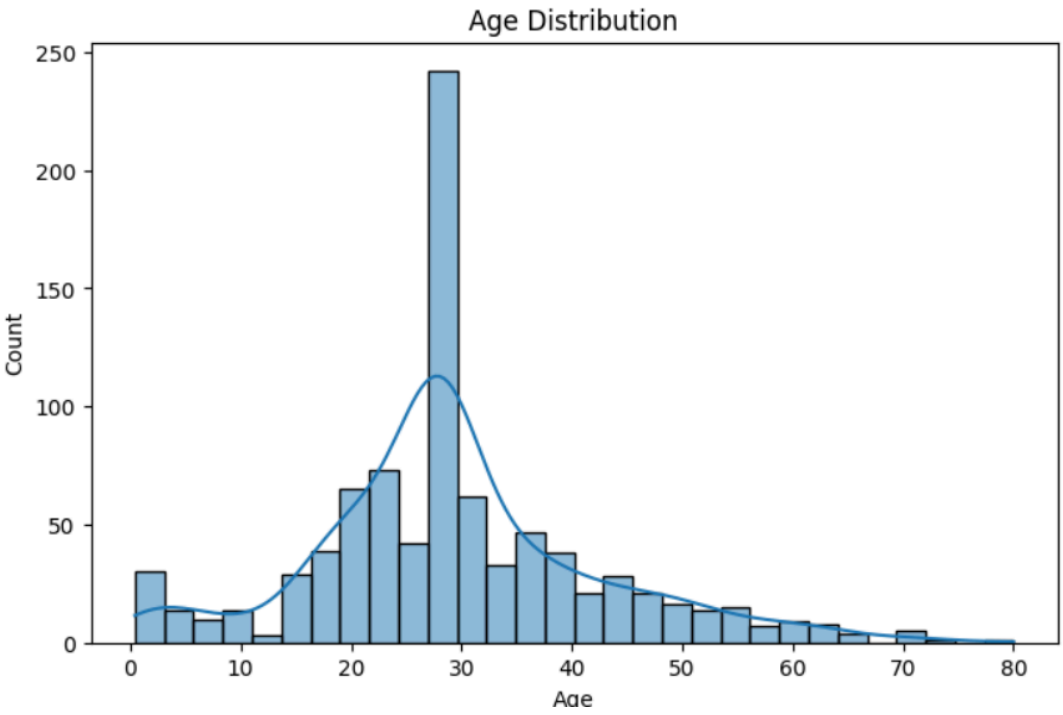
4. Conclusion

This **Exploratory Data Analysis** of the Titanic dataset provided valuable insights into the factors influencing passenger survival. The key takeaways include the significant impact of **gender**, **passenger class**, and **age** on survival rates.

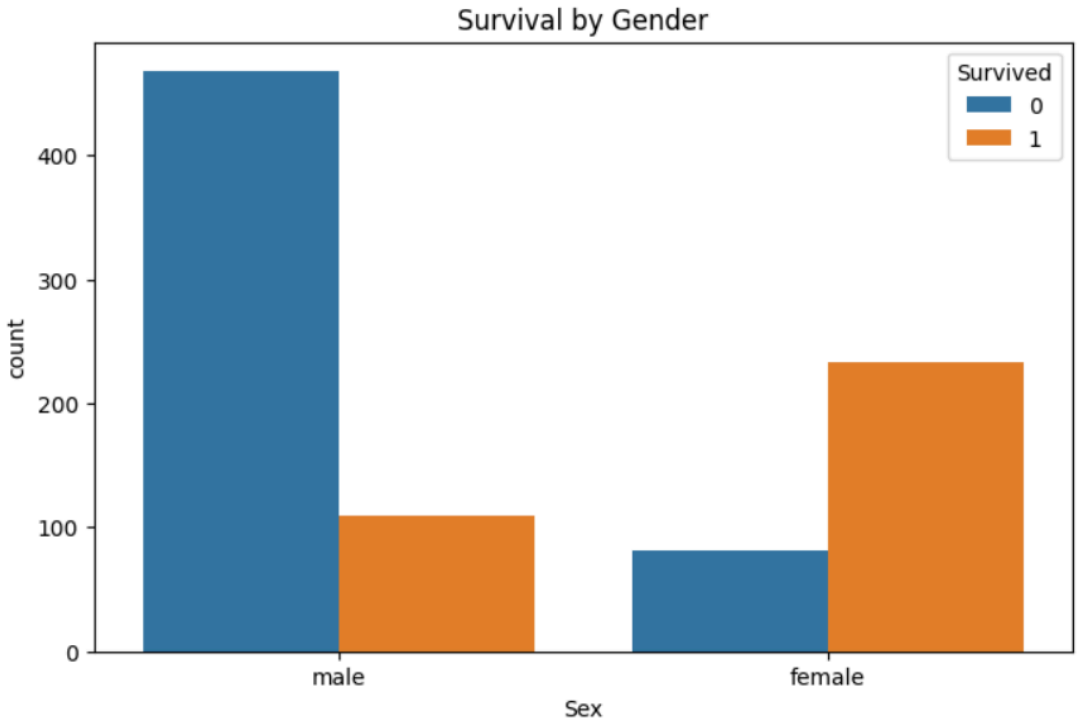
5. Visualizations

Here are the key visualizations from the EDA:

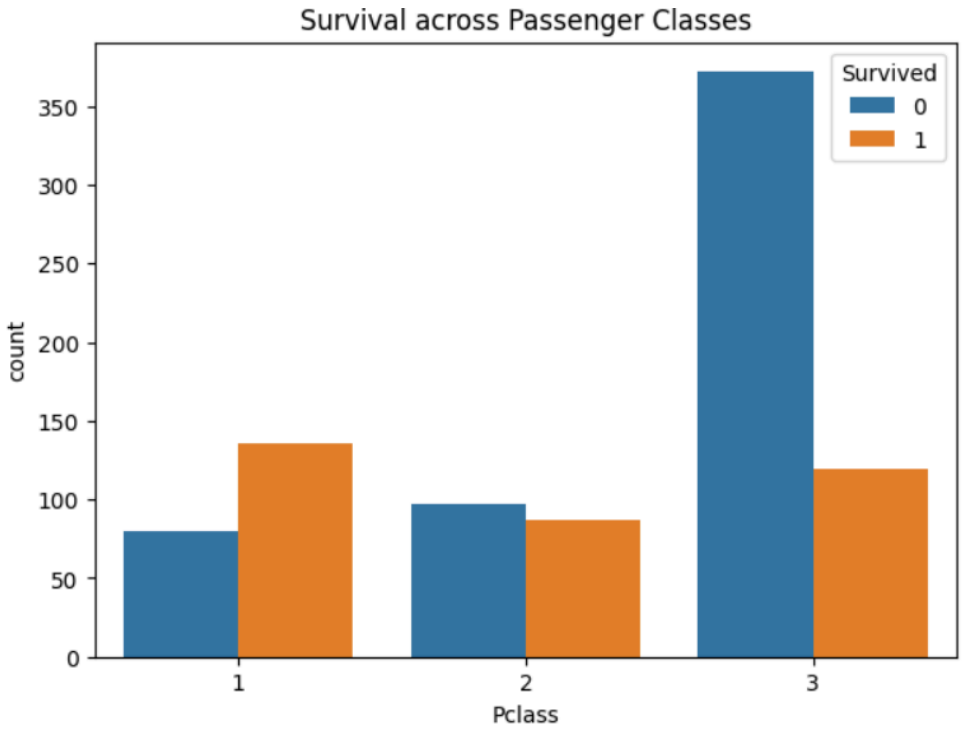
1. **Age Distribution**



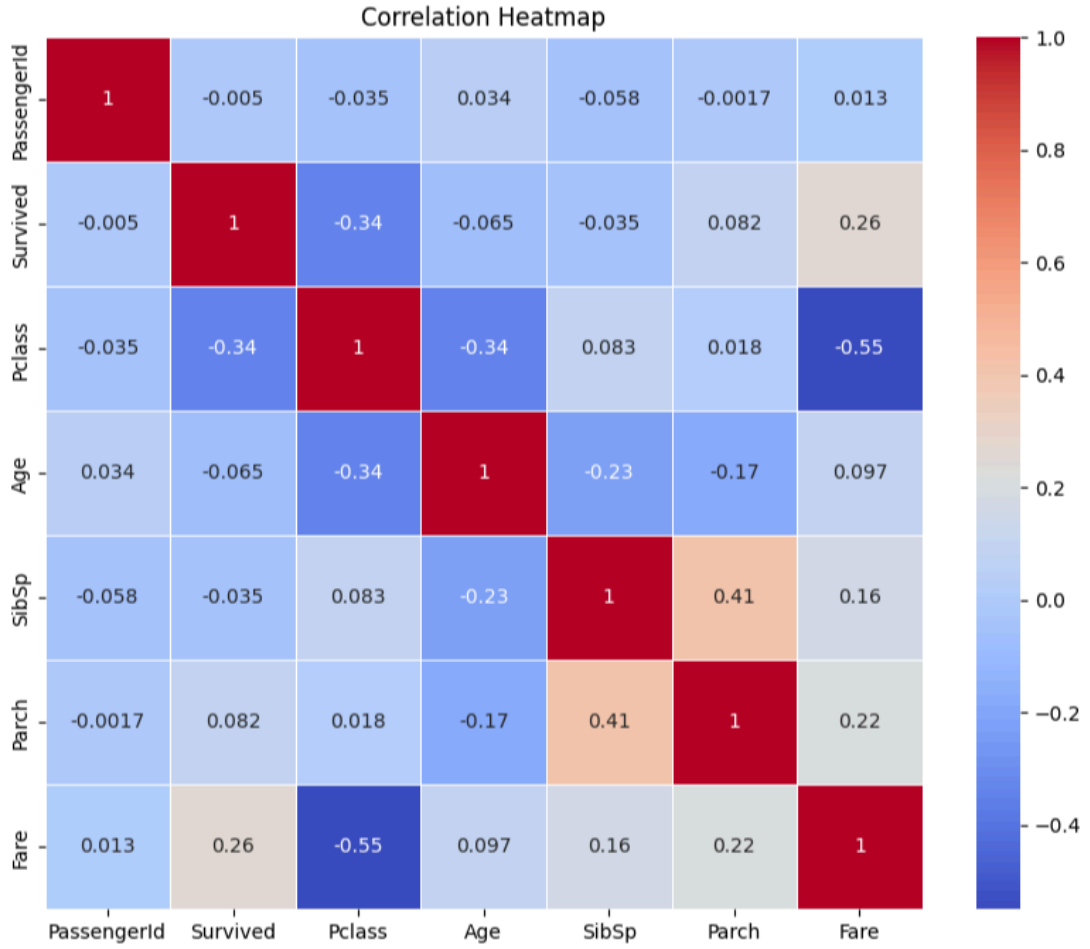
2. **Survival by Gender**



3. **Survival by Passenger Class**



4. **Correlation Heatmap**



Tools Used:

- **Python Libraries:** Pandas, NumPy, Matplotlib, Seaborn
- **Data Source:** Titanic dataset from Kaggle