Exploratory Data Analysis Report: Titanic Dataset

By: Abdul Wahab Aziz

Introduction

The Titanic dataset provides detailed information about passengers aboard the RMS Titanic, which sank on April 15, 1912. This dataset contains demographic, economic, and survival information for passengers. The goal of this analysis is to uncover meaningful insights and patterns using exploratory data analysis (EDA).

Objectives

1. Understand the distribution of passengers across various categories (e.g., class, gender, and embarkation port).
2. Analyze the age and fare distributions to identify trends and anomalies.
3. Examine the relationships between key features and survival outcomes.

Steps Performed

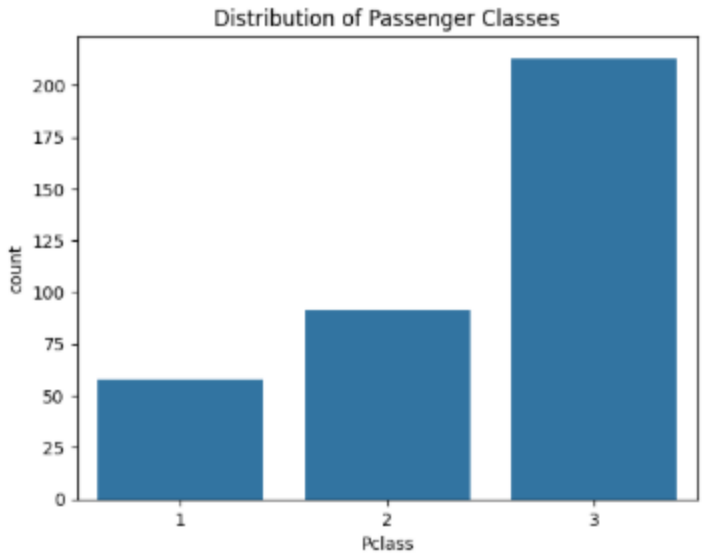
1. Data Cleaning

- **Missing Values**: Imputed missing Age values with the median, filled missing Embarked values with the mode, and dropped the Cabin column due to excessive missing data.
- **Duplicates**: Removed any duplicate rows to ensure data consistency.
- **Outlier Handling**: Used interquartile range (IQR) to filter out extreme outliers in numerical columns like Fare.

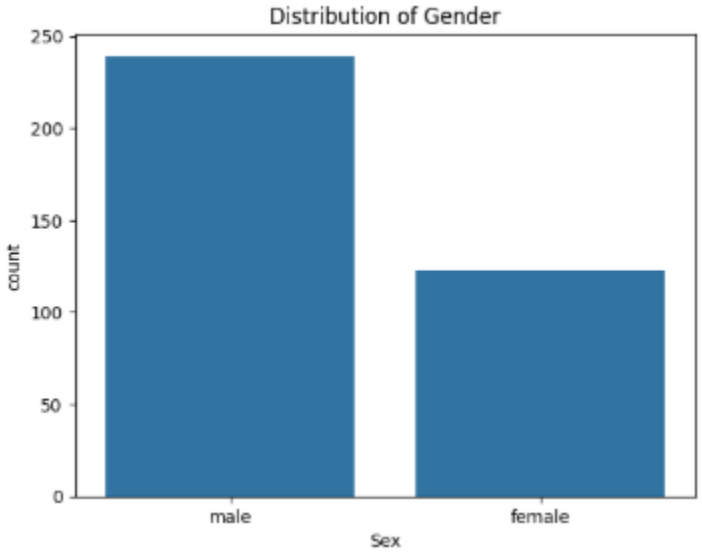
2. Visualizations and Insights

A. Bar Charts for Categorical Variables

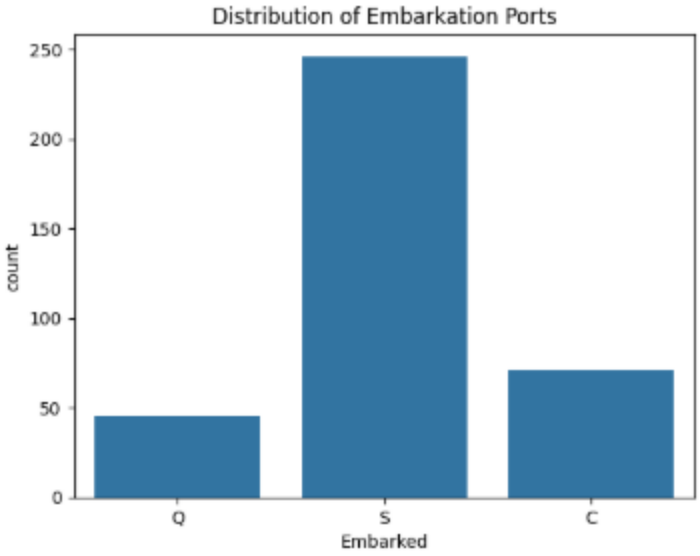
1. **Passenger Class Distribution**:
    - Most passengers belonged to the third class (Pclass = 3).
    - First-class passengers were the fewest, suggesting a socioeconomic divide.



2. **Gender Distribution**:
    - There were significantly more male passengers than females.

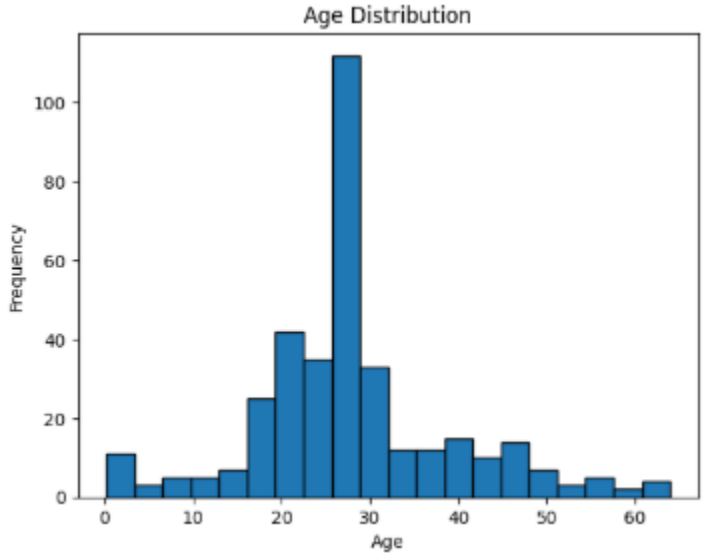


3. **Embarkation Port Distribution**:
    - Most passengers boarded at Southampton (S), followed by Cherbourg (C) and Queenstown (Q).

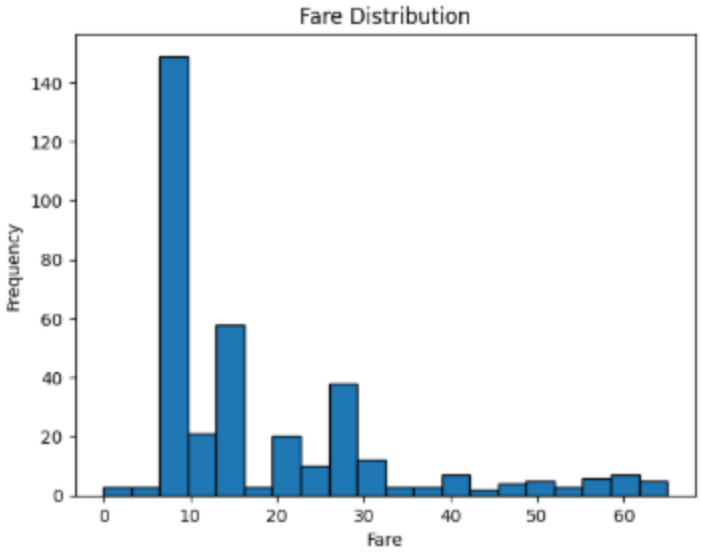


B. Histograms for Numerical Variables

1. **Age Distribution**:
    - The majority of passengers were in their 20s and 30s.
    - Very young children and older adults were less frequent.

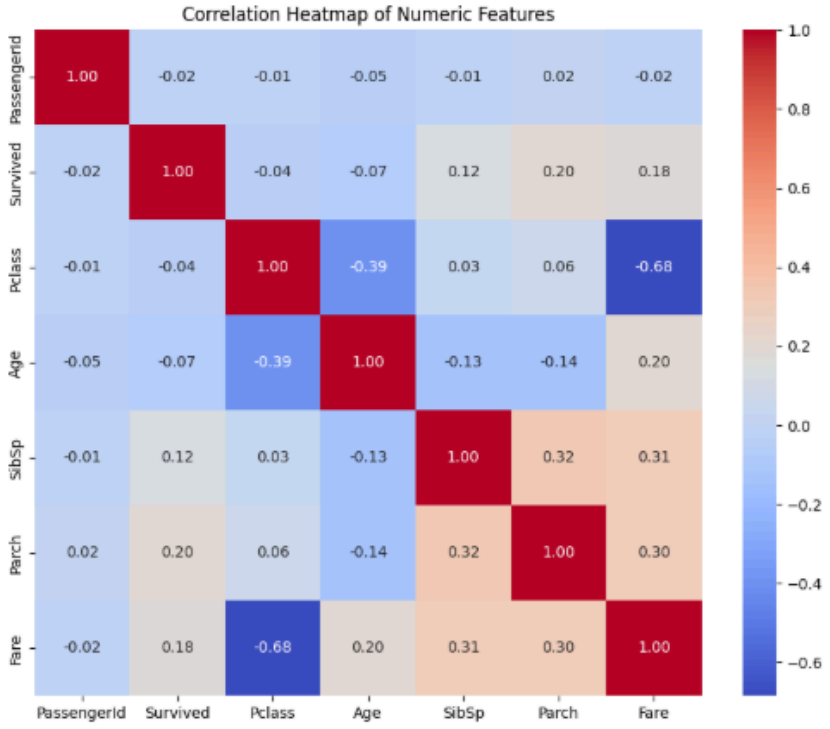


2. **Fare Distribution**:
    - Most fares were concentrated in the lower range, indicating affordability for third-class passengers.
    - A few outliers reflect the high cost of first-class tickets.



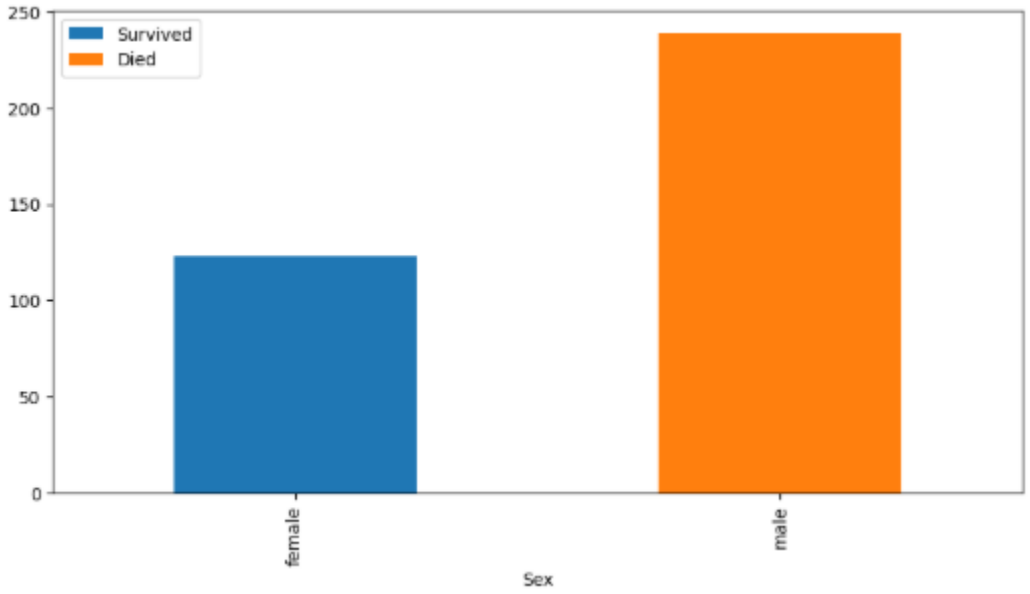
C. Correlation Heatmap

- **Key Observations**:
  - Pclass is negatively correlated with Fare, as higher-class passengers paid more.
  - Age has a weak correlation with survival, suggesting no strong linear relationship.
  - Fare has a moderate positive correlation with survival, indicating wealthier passengers had higher survival rates.

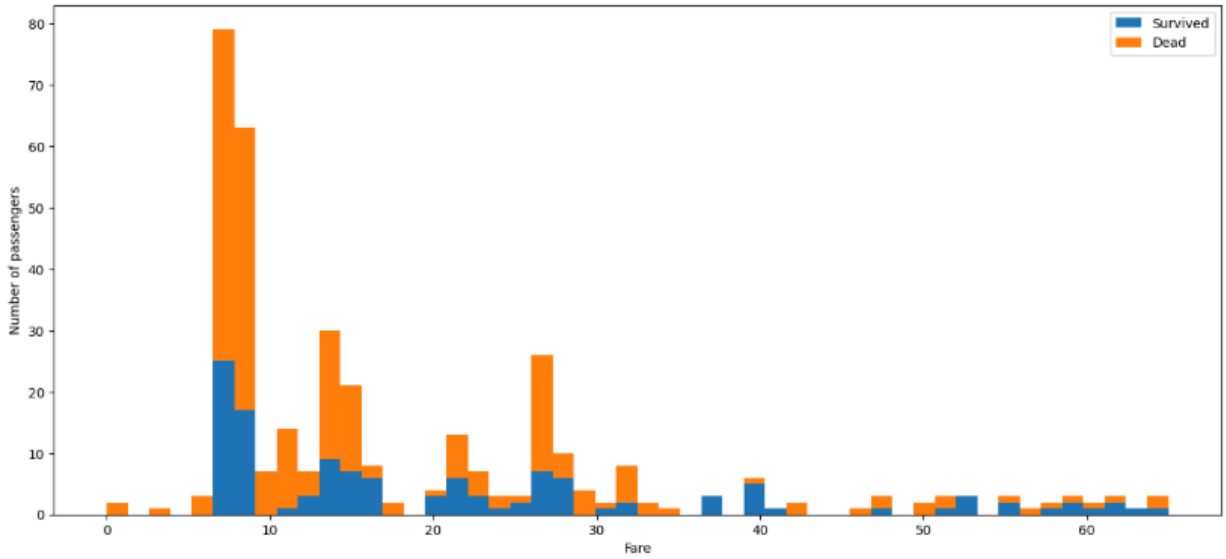


D. Survival Analysis

1. **Survival by Gender**:
    - Females had a much higher survival rate compared to males.
    - This aligns with the "women and children first" policy during the evacuation.



2. **Survival by Fare**:
    - Passengers who paid higher fares had better chances of survival.
    - Indicates a strong link between socioeconomic status and survival.



Conclusions and Insights

1. **Passenger Demographics**:
    - The dataset is dominated by third-class passengers and male passengers.
    - Most passengers were young adults.
2. **Key Factors Influencing Survival**:
    - **Gender**: Females had a significantly higher survival rate.
    - **Fare**: Higher fares correlated with better survival outcomes, suggesting that wealthier passengers had access to lifeboats.
    - **Class**: First-class passengers were more likely to survive compared to those in second or third class.
3. **General Observations**:
    - The embarkation port (Southampton) had the highest number of passengers.
    - The dataset reveals strong socioeconomic disparities in survival.