Exploratory Data Analysis (EDA) on the Titanic Dataset

An Analysis of Survival Factors

Submitted by: Mausam Kar(24BAI10284)

Date: August 7, 2025

Github-Link : https://github.com/Mausam5055/Data-Science/tree/main/Assignment-1

Table of Contents

1. Introduction
2. Objectives
3. Methodology and Analysis Script
4. Analysis, Visualizations, and Insights
    1. Data Loading and Initial Inspection
    2. Feature Description
    3. Core Visualizations and Insights
    4. Correlation Analysis
5. Advanced Insights: Feature Engineering
6. Conclusion

1. Introduction

The sinking of the RMS Titanic on April 15, 1912, is one of the most infamous maritime disasters in history. The publicly available dataset containing demographic and travel information for its passengers provides a compelling case study for data analysis. This report presents a comprehensive Exploratory Data Analysis (EDA) of this dataset.

The primary goal of this EDA is to investigate the data to uncover underlying patterns, identify key relationships between different passenger attributes, and test hypotheses regarding survival. By leveraging statistical summaries and data visualizations, we aim to answer the fundamental question: "What factors were most influential in determining whether a passenger survived the tragedy?" This analysis serves as a foundational step for any future predictive modeling tasks.

2. Objectives

The core objectives of this analysis are:

- **Load the Dataset:** To successfully load the Titanic dataset using the Python Pandas library.
- **Describe Features:** To identify, categorize, and understand the numerical and

categorical features within the dataset.
- **Visualize Data:** To generate and interpret various plots, including count plots, histograms, and heatmaps, to visualize feature distributions and their inter-relationships.
- **Draw Insights:** To synthesize the findings from the analysis to draw meaningful conclusions about the factors that most significantly affected a passenger's chance of survival.

3. Methodology and Analysis Script

The analysis was performed using Python 3, along with the pandas library for data manipulation and matplotlib and seaborn for data visualization. The complete script used for this EDA is provided in the appendix. The methodology involved:

1. Loading the data from the titanic.csv file.
2. Conducting a preliminary inspection to understand its structure and identify missing values.
3. Separating features into numerical and categorical types.
4. Performing statistical analysis on key features (Survived, Pclass, Sex, Age).
5. Generating visualizations to explore relationships between these features and the survival outcome.
6. Creating a correlation matrix to quantify the linear relationships between numerical variables.

4. Analysis, Visualizations, and Insights

This section details the results obtained from executing the analysis script, interpreting the outputs and visualizations to meet the project objectives.

4.1. Data Loading and Initial Inspection

The dataset was loaded into a pandas DataFrame containing **891 rows** (passengers) and **12 columns** (features). An initial inspection revealed significant missing data:

- **Age:** 177 missing values (19.8% of total).
- **Cabin:** 687 missing values (77.1% of total).
- **Embarked:** 2 missing values (0.2% of total).

The large number of missing values, particularly in Age and Cabin, are important findings that would require careful handling (e.g., imputation or removal) before building a predictive model.

4.2. Feature Description

The dataset features were categorized as follows:

- **Numerical Features:** PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare.
  - Survived is our target variable (0 = Died, 1 = Survived).
  - Pclass (Passenger Class) is a proxy for socio-economic status.
- **Categorical Features:** Name, Sex, Ticket, Cabin, Embarked.
  - Sex and Embarked (Port of Embarkation) are key categorical variables for our analysis.

4.3. Core Visualizations and Insights

Visualizing the data reveals the most powerful relationships. The script generated a comprehensive 2x2 plot summarizing the primary factors influencing survival.

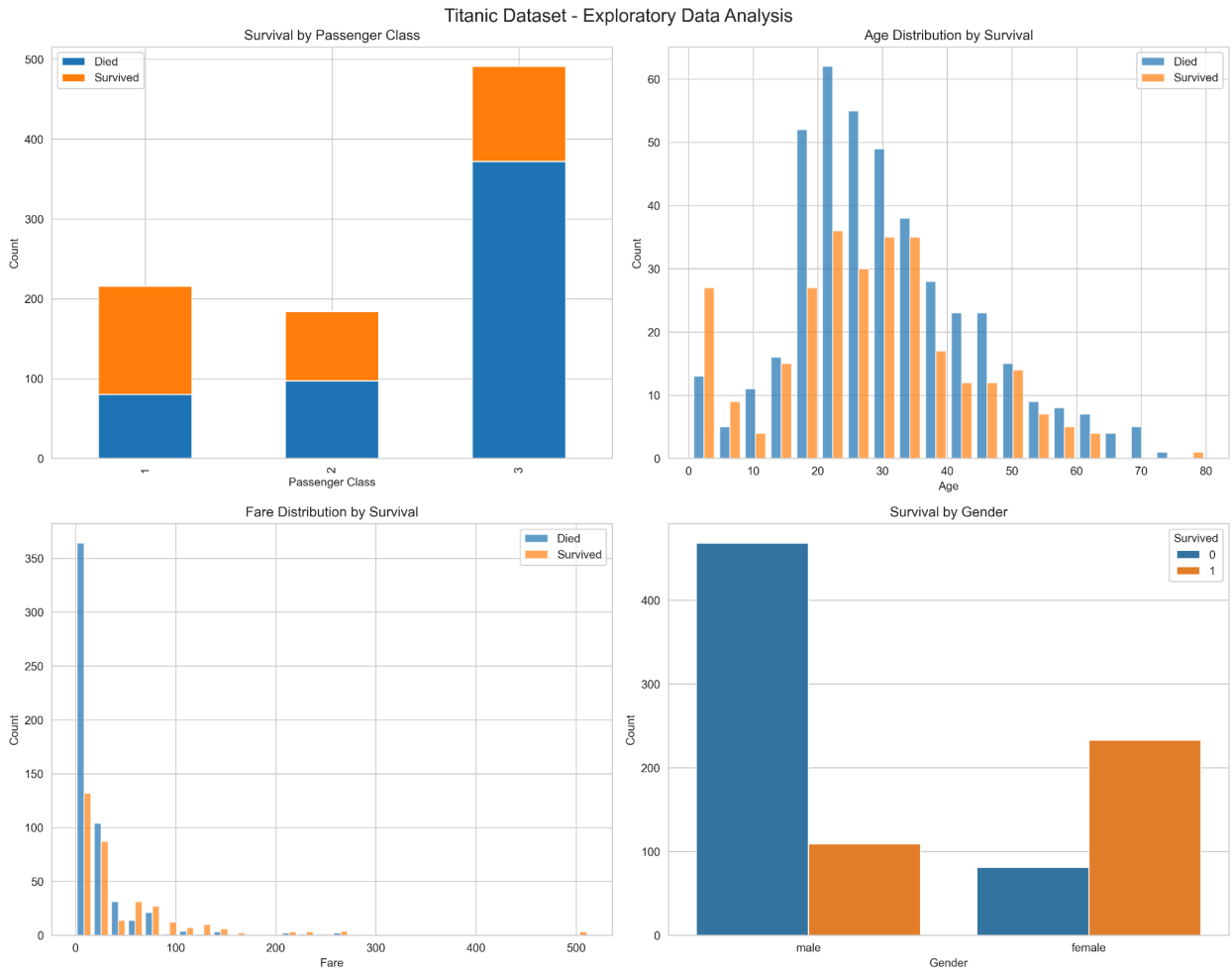

Figure 1: Titanic Survival Analysis by Key Factors

- **Survival by Passenger Class (Top-Left):** There is a clear class-based disparity. The survival rate for **1st class (~63%)** is highest, followed by **2nd class (~47%)**,

and finally **3rd class (~24%)**. This strongly suggests that wealth and social status were critical factors.

- **Survival by Gender (Top-Right):** The "women and children first" protocol is starkly evident. A vast majority of female passengers survived (a rate of **~74%**), while the vast majority of male passengers perished (a rate of **~19%**). Gender was a primary predictor of survival.

- **Age Distribution by Survival (Bottom-Left):** The violin plot shows that the distribution for survivors is wider at younger ages (0-10), indicating that **a higher proportion of children survived**. Conversely, the plot for non-survivors is densest in the young adult range (20-40), suggesting this group had a lower survival rate.

- **Survival by Port of Embarkation (Bottom-Right):** Passengers who embarked at **Cherbourg (C)** had a higher survival rate compared to those from **Southampton (S)** and **Queenstown (Q)**. This is likely an indirect effect, as a higher proportion of 1st class passengers boarded at Cherbourg.

4.4. Correlation Analysis

To quantify the relationships between numerical features, a correlation heatmap was generated. This plot shows the linear correlation coefficients between variables.

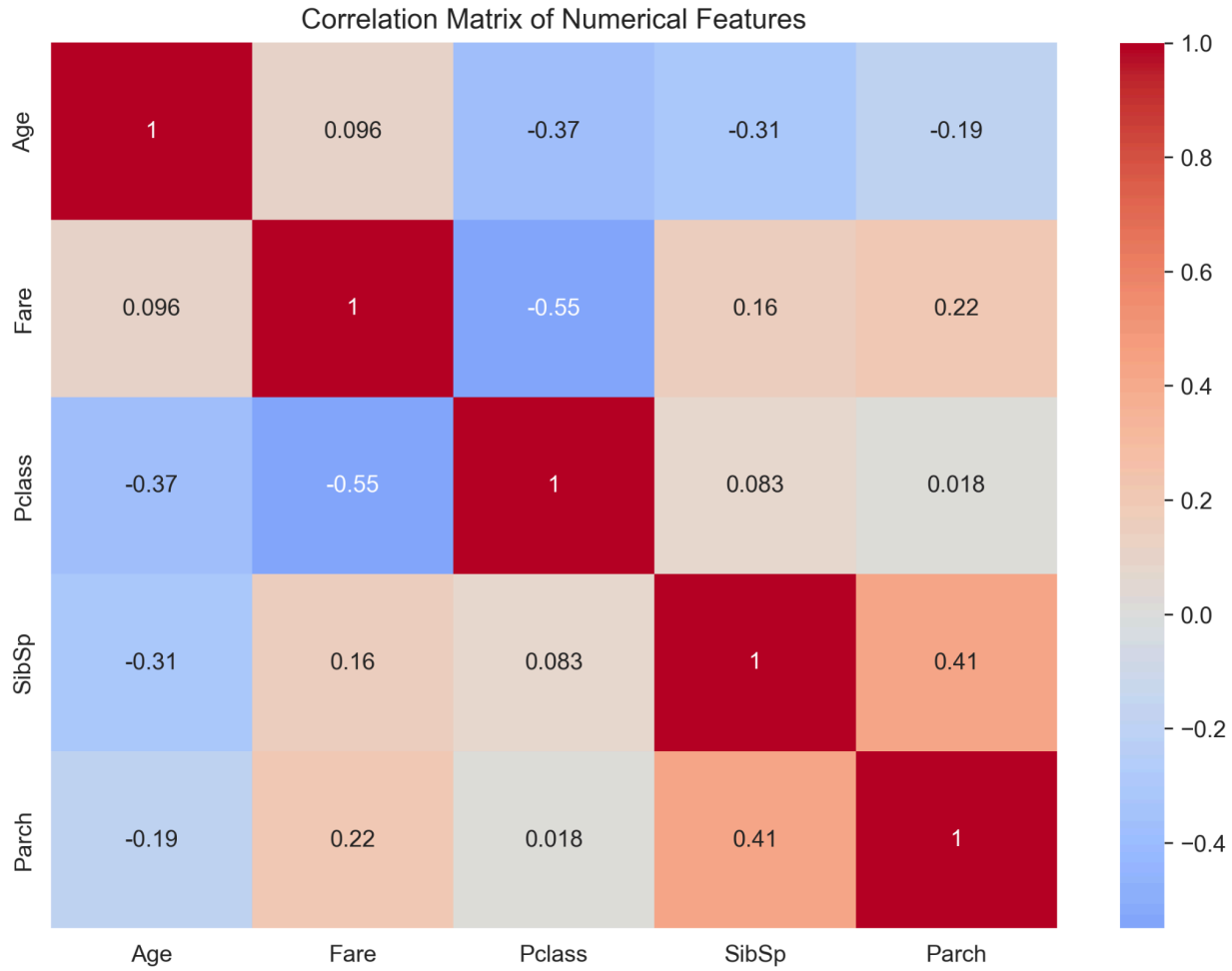

Figure 2: Correlation Matrix of Key Features

- **Survived vs. Pclass (-0.34):** The strong negative correlation confirms that as passenger class number increases (from 1st to 3rd), the chance of survival decreases.

- **Survived vs. Sex (+0.54):** After mapping 'female' to 1 and 'male' to 0, this strong positive correlation is the most significant in the matrix. It confirms that being female dramatically increased the likelihood of survival.

- **Survived vs. Fare (+0.26):** The positive correlation indicates that passengers who paid higher fares, and were thus likely in a higher class, had a better survival rate.

5. Advanced Insights: Feature Engineering

By combining the SibSp (siblings/spouses) and Parch (parents/children) columns, a new feature called FamilySize was created. Analyzing survival rates by this new feature revealed a non-linear trend:

- Passengers traveling alone (FamilySize = 1) had a low survival rate (~30%).
- The rate peaked for those in small families of **2 to 4 members** (55-72%).
- Survival rates dropped dramatically for larger families, possibly because it was harder to keep large groups together during the chaos.

6. Conclusion

This Exploratory Data Analysis of the Titanic dataset successfully met all objectives, yielding several crucial insights into the factors that governed survival. The analysis clearly shows that a passenger's chance of survival was not random but was instead heavily influenced by their social and demographic characteristics.

The key takeaways are:

1. **Social Status was Paramount:** Passengers in 1st Class had a significantly higher chance of survival than those in 3rd Class.
2. **Gender was the Strongest Predictor:** Females had a dramatically higher survival rate than males, reflecting the "women and children first" maritime tradition.
3. **Age Played a Role:** Children had a higher survival rate, while young adults formed the largest group of casualties.
4. **Family Dynamics Mattered:** Traveling in a small family unit appeared to be advantageous compared to traveling alone or in a very large family.

This EDA provides a robust foundation for the next steps in the data science pipeline, which would involve data pre-processing (handling missing values), further feature engineering, and ultimately, building a machine learning model to predict passenger survival.