Credora Internship – Data Science

WEEK 2 -Task 02

[DATA CLEANING & EXPLORATORY DATA ANALYSIS]

Submitted by: Alamuru Venkata Harshitha

1. Objective

The goal of this task is to perform exploratory data analysis (EDA) on the Titanic passenger dataset to identify key patterns and factors that influenced survival outcomes during the disaster. This analysis serves as a foundational step in the data science workflow, focusing on data quality, preprocessing, feature insights, and visual storytelling.

2. *Dataset Overview*

The dataset originates from the Kaggle Titanic Competition and contains demographic and travel information of 891 passengers aboard the Titanic.

- Target Variable: Survived (1 = survived, 0 = did not survive)
- Key Features:
  - Demographics: Age, Sex
  - Class: Pclass, Fare
  - Boarding Info: Embarked
  - Identifiers: PassengerId, Name, Ticket, Cabin

3. *Data Cleaning & Preprocessing*

Missing Age values were filled with the **median** to retain a central value without being influenced by outliers.

Embarked column missing values were replaced with the **mode**, as it's categorical and the most frequent port made the best default.

Cabin column was dropped entirely due to having over 75% missing data, making it unreliable for analysis.

Sex column was converted to numeric format, mapping **male to 0** and **female to 1**, for compatibility with analysis tools.

Embarked column was encoded numerically (**S = 0, C = 1, Q = 2**) to allow correlation and visualization.

Name, Ticket, and Passenger ID columns were removed as they were either identifiers or text fields not useful for EDA.

4. Key Insights & Visualizations

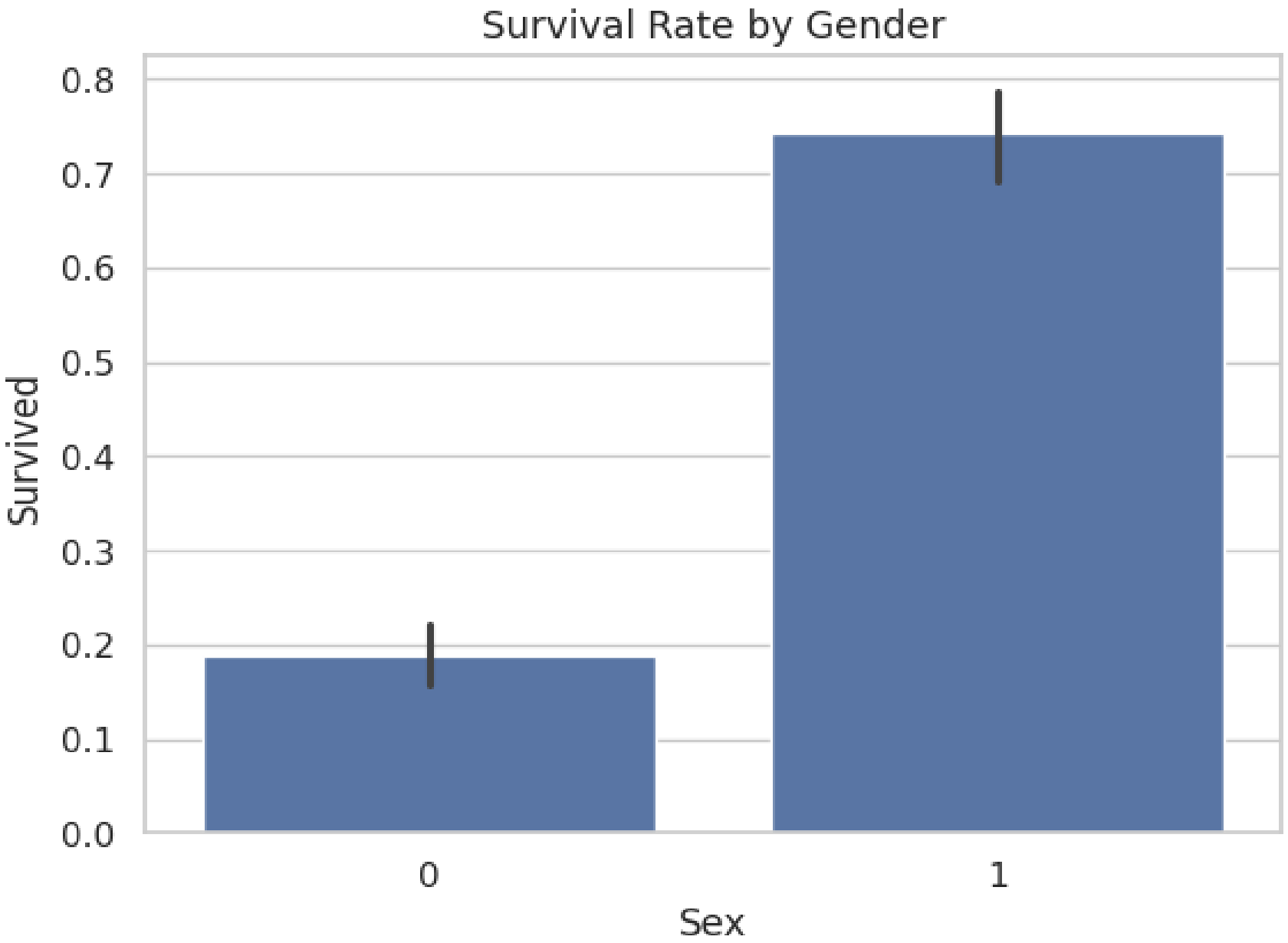
4.1 **Survival by Gender**

Females were 4x more likely to survive than males.

0 = male & 1 = female

Bar plot showed ~75% of females survived vs ~20% of males

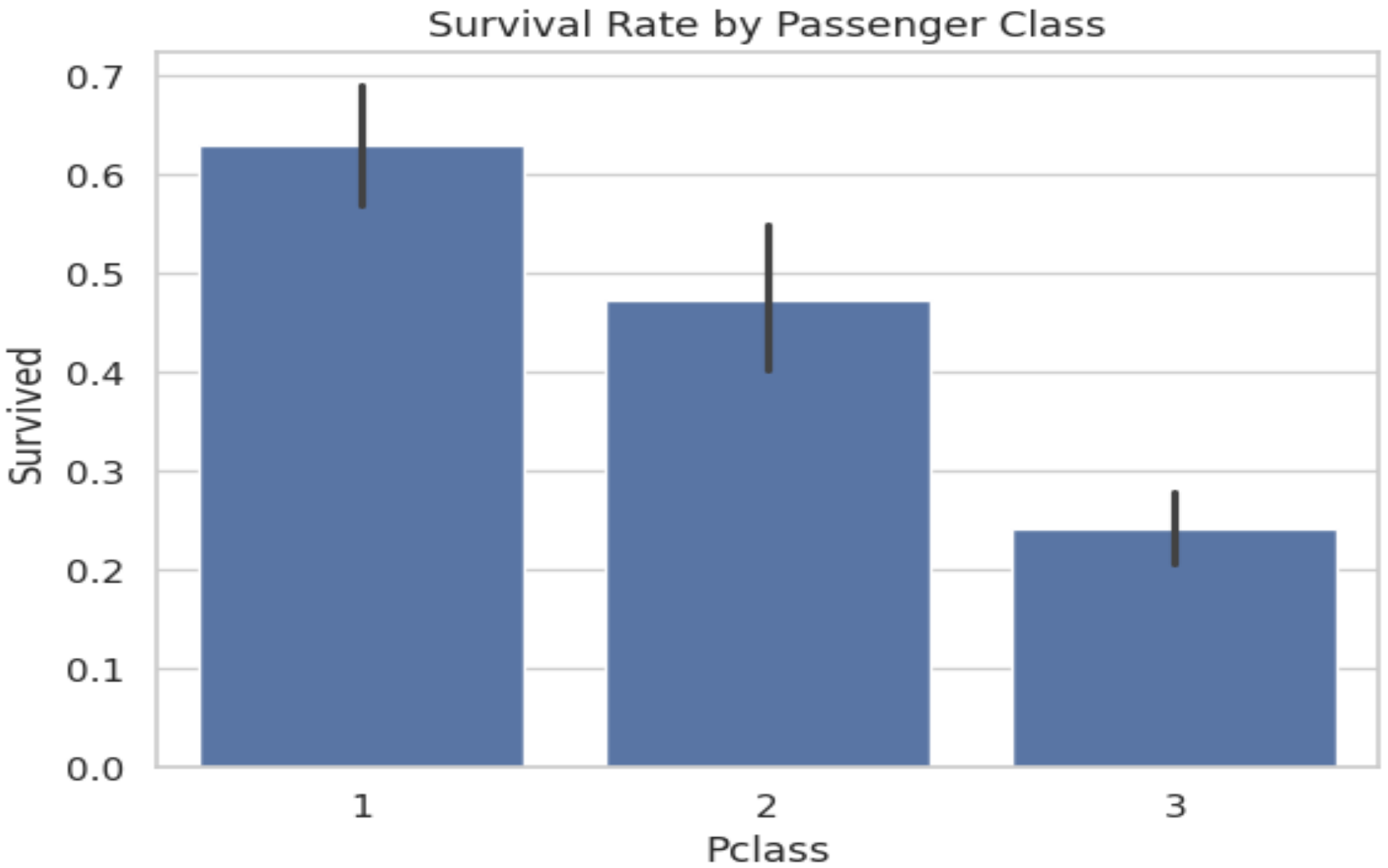
This was the strongest predictive feature in the dataset.



4.2 **Survival by Passenger Class**

1st class passengers had the highest survival rate (~63%).

Reflects socioeconomic privilege — access to lifeboats, location proximity.

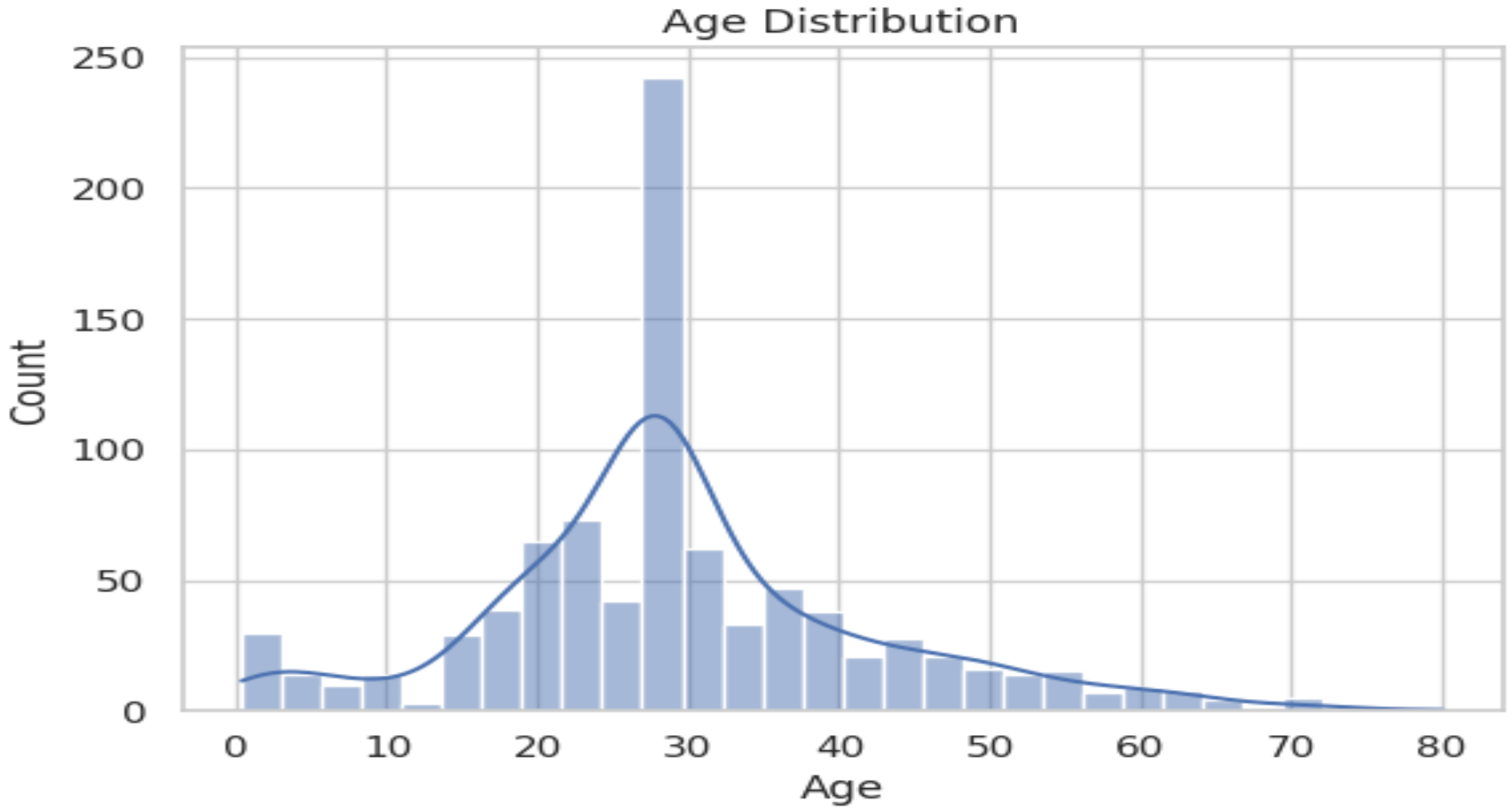


4.3 **Age Distribution**

Young passengers (especially children) had better chances of survival.

Most passengers were aged 20–40.

Children under 10 had a higher survival rate than any other age group.

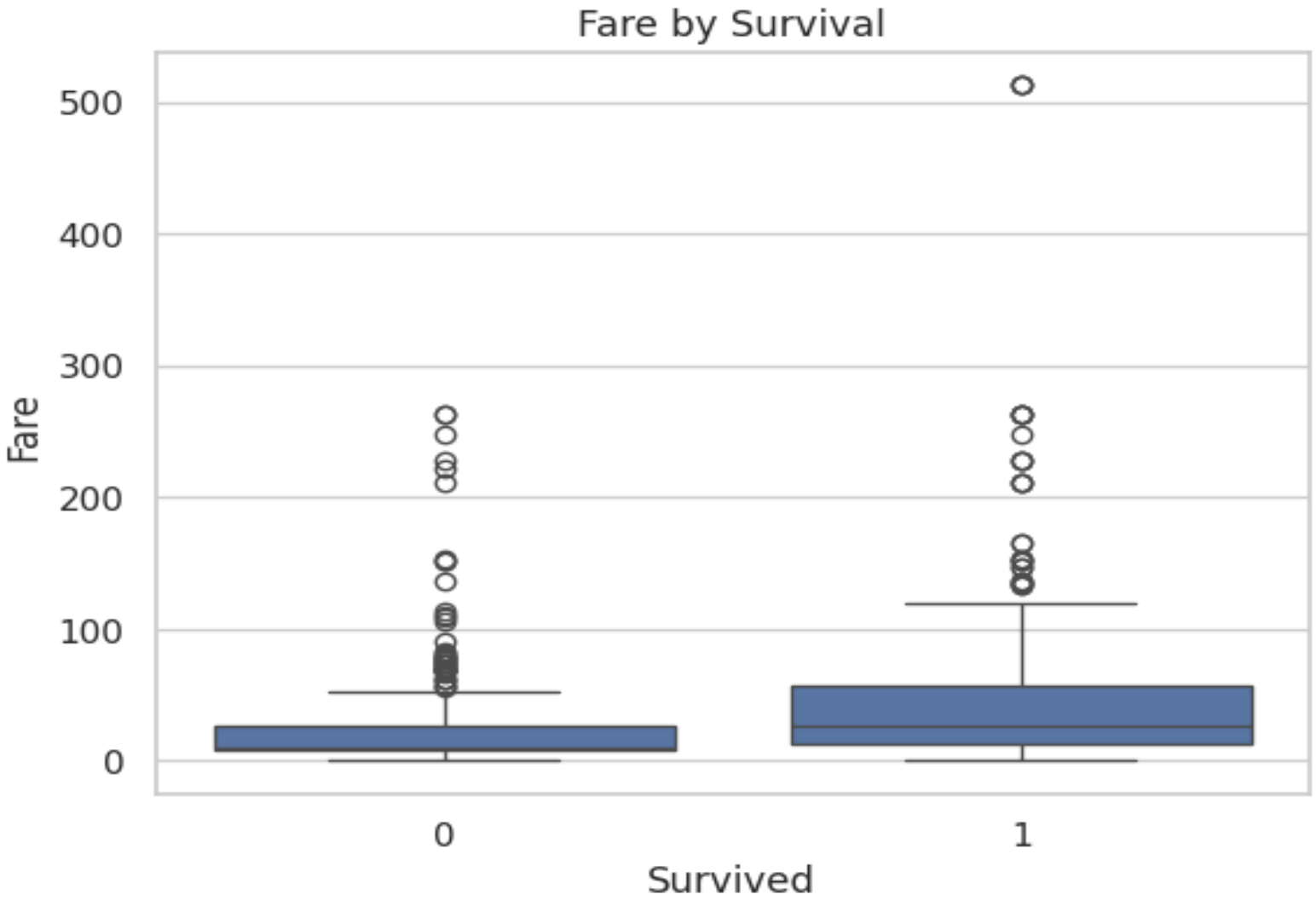


4.4 **Fare vs Survival**

Higher fare correlated with higher survival.

Survivors had a higher average fare than non-survivors.

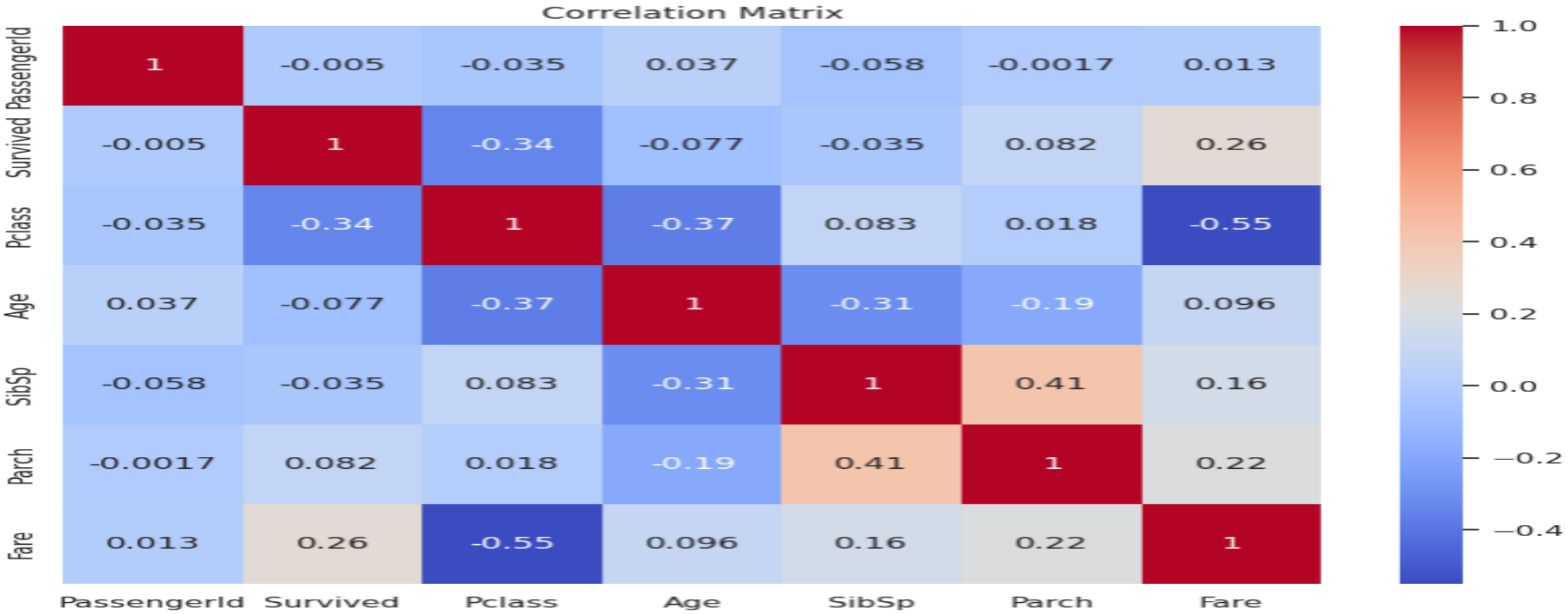
Reinforces the class gap in survival.



4.5 **Correlation Matrix**

Strongest correlations:

- Sex (female) and Survived: +0.54
- Pclass and Survived: –0.34
- Fare and Survived: +0.26



5. *Final Summary*

- **Females had a significantly higher survival rate** compared to males, highlighting gender-based evacuation priorities.
- **1st class passengers** were more likely to survive, indicating a clear advantage for wealthier individuals during the disaster
- **Children and younger passengers** had better survival chances, supporting the "women and children first" policy.
- **Passengers who paid higher fares** showed higher survival rates, as fare closely aligned with travel class
- **Class, gender, and fare** emerged as the most influential factors associated with survival.

6. Challenges Faced & Solutions

Missing Data: `Age` and `Embarked` had null values.
→ Filled `Age` with median and `Embarked` with mode for consistency.

High Null Column: `Cabin` had too many missing entries.
→ Dropped the column to avoid introducing noise.

Categorical Encoding: `Sex` and `Embarked` were non-numeric.
→ Mapped them to numeric values for analysis and visualization.

Unfocused Visuals: Initial plots lacked clarity and insights.
→ Refined visualizations to highlight survival patterns effectively

7. Deliverables

- Colab_task_2: Interactive analysis notebook
- README.md: Project overview and instructions
- data/: Contains train.csv, test.csv, Gender_submission.csv

8. Links

- 🗁 GitHub Repo: [GITHUB_TASK_2]
- 🖇 Google Colab Notebook: [colab_link]

9. Contact

[**Alamuru venkata Harshitha**]
Data Science Intern @ Credora
✉ Email: [chinnialamuru98@gmail.com]
🖇 LinkedIn: [LINKDIN_Harshitha Alamuru]
🐙 GitHub: [GITHUB_chessman and smiley]