Exploratory Data Analysis on Titanic Dataset

Introduction

The Titanic dataset is one of the most popular datasets for practicing data analysis and machine learning. It contains detailed information about passengers on the Titanic, including demographics, ticket information, and survival status. The main objective of this analysis is to explore the dataset, understand its structure, identify patterns, detect anomalies, and generate meaningful insights that may help in building predictive models.

Dataset Overview

The dataset consists of passenger details with attributes such as *PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.*
It contains both numerical and categorical variables, making it suitable for a variety of analyses.

Dataset Structure

From the .info() output

- **Total Entries:** 891
- **Number of Columns:** 12
- **Data types:** Integer (5), Float (2), Object/String (5)

Statistical Summary

From the .describe() output for numerical variables:

- **Age:** Mean ~29.7 years, Min = 0.42, Max = 80
- **Fare:** Mean ~32.2, Min = 0, Max = 512.33
- **SibSp:** Most passengers traveled with 0 siblings/spouses
- **Parch:** Most passengers traveled with 0 parents/children

Missing Values

The missing value analysis shows:

- **Age:** 177 missing (~19.86%)
- **Cabin:** 687 missing (~77.10%)
- **Embarked:** 2 missing (~0.22%)
  Other columns have complete data.

Reasons for missing data:

- ***Age:*** Passenger ages were not always recorded in historical logs.
- ***Cabin:*** Cabin numbers were not assigned to all passengers, especially lower-class passengers.
- ***Embarked:*** A small number of records lacked port-of-embarkation information.

Categorical Data Distribution

Sex

Males represent about **64.8%** of passengers, females about **35.2%**.

Pclass

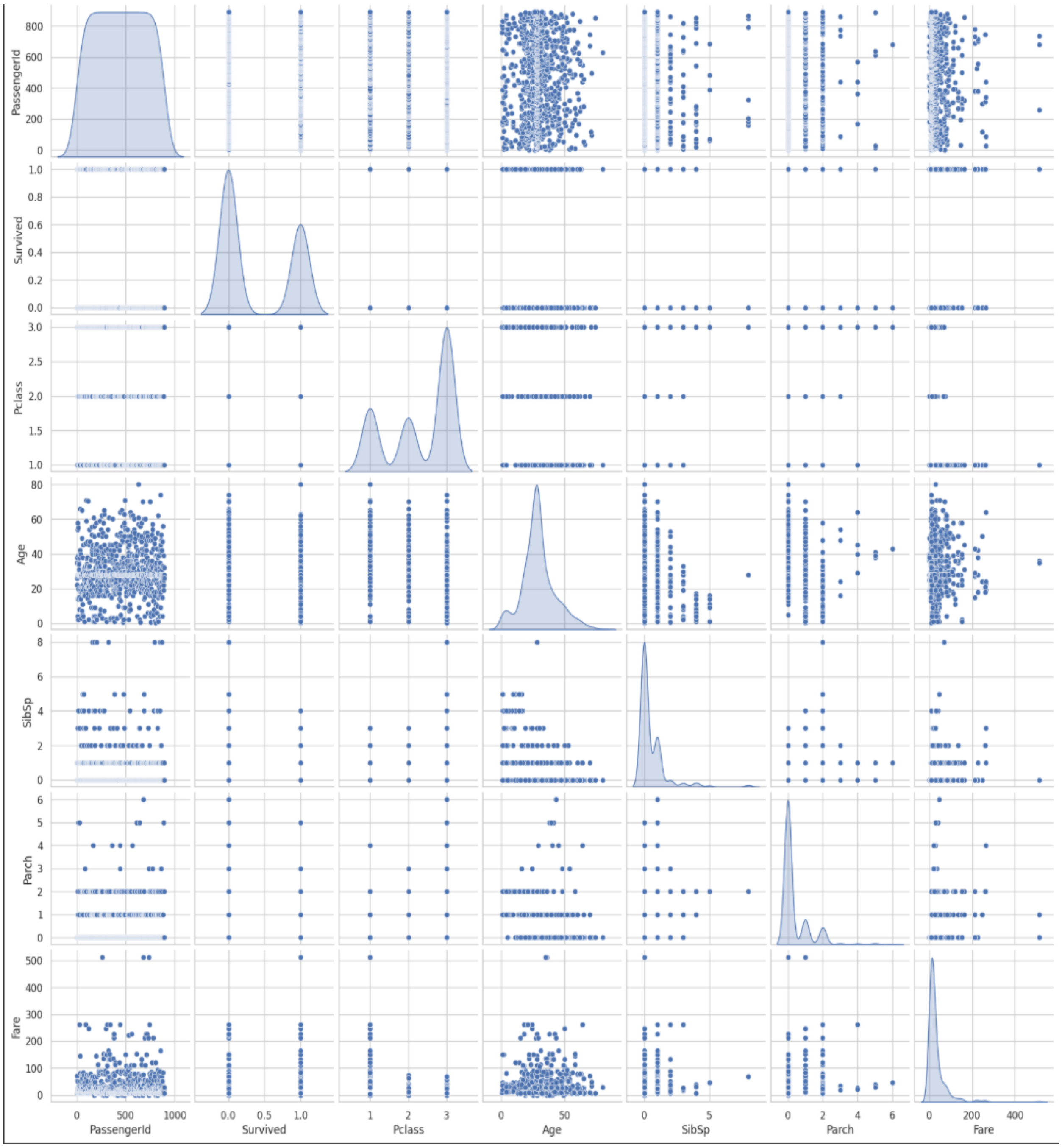
3rd class has the highest number of passengers (~55%), followed by 1st class (~24%), and 2nd class (~21%).

Embarked

Most passengers embarked from **Southampton (S)**, followed by Cherbourg (C) and Queenstown (Q).

Visualizations and Observations

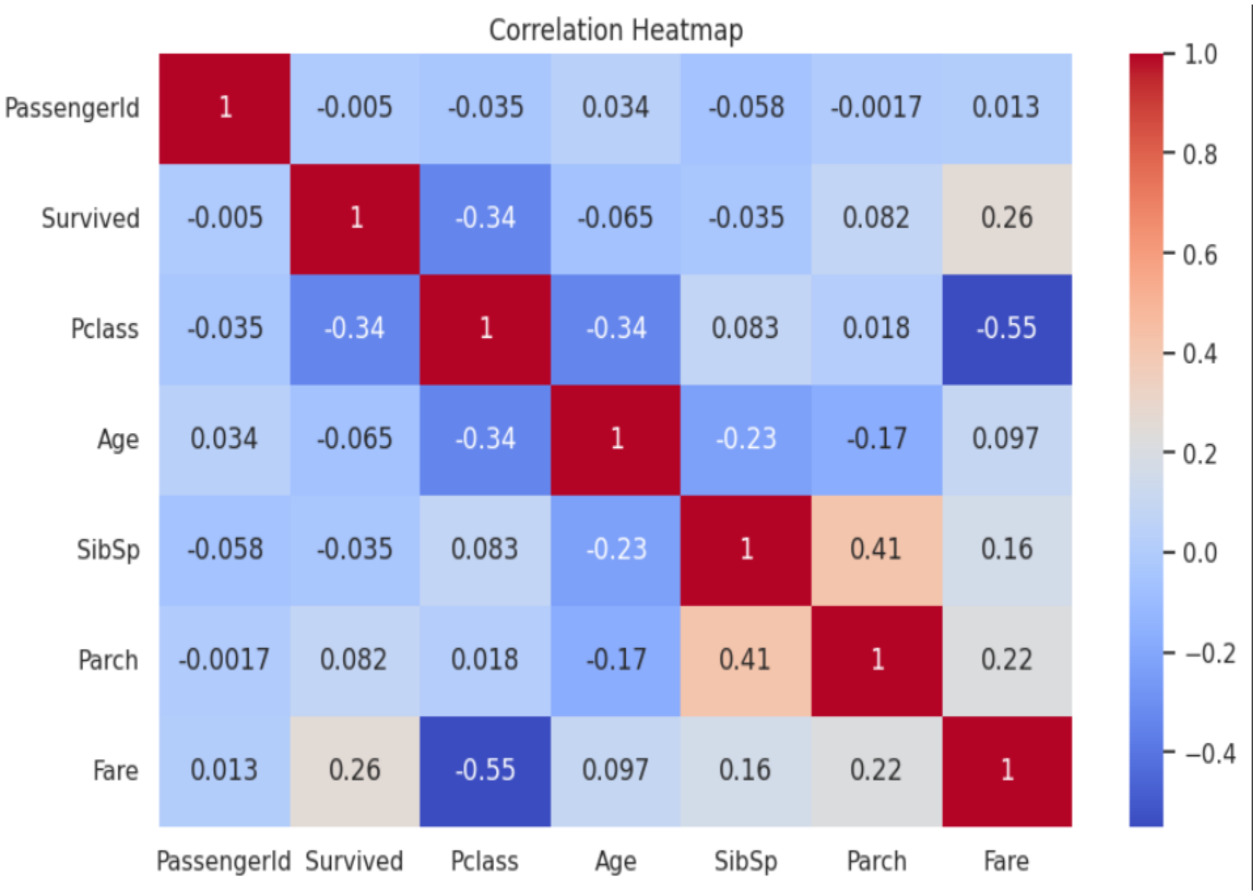
Pair Plot



Observation:

- Clear distinction in survival patterns between male and female passengers.
- 1st class passengers dominate higher fare ranges.
- Most 3rd class passengers paid low fares and had lower survival chances.

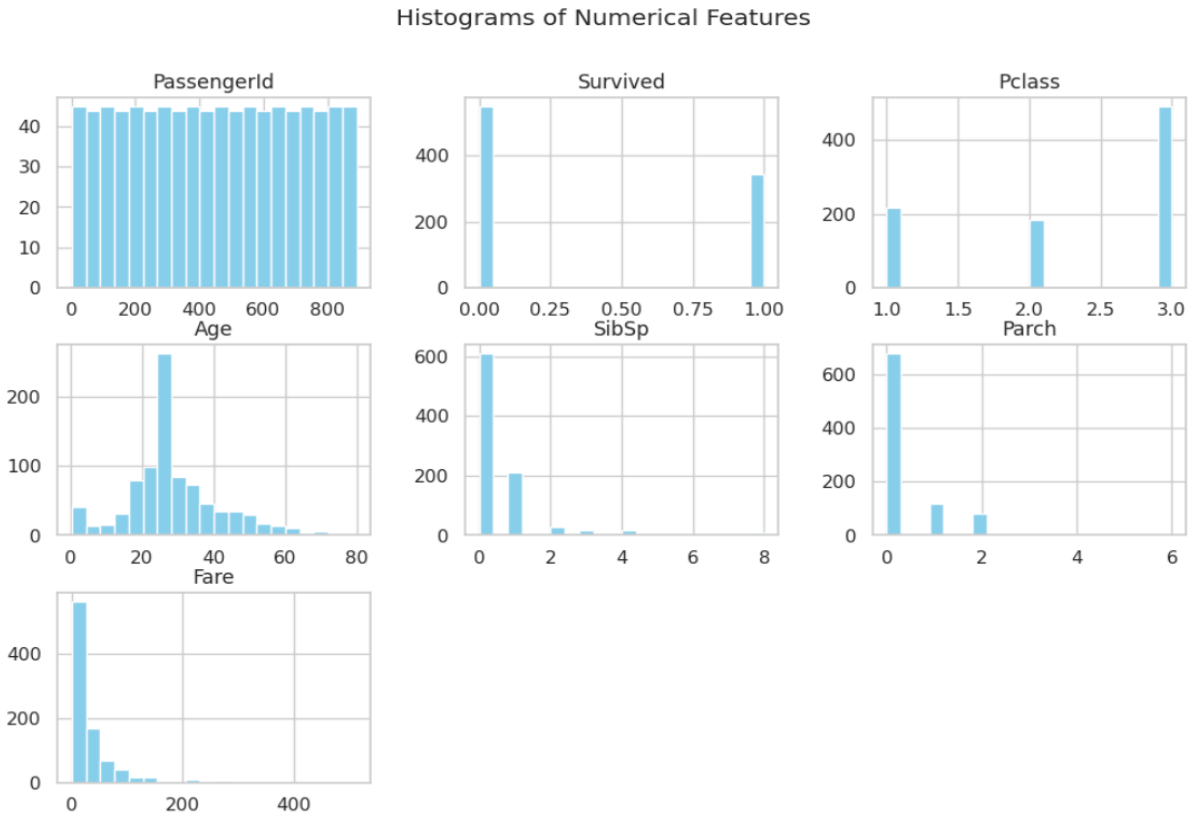
Correlation Heatmap



Observation:

- **Survived** is moderately positively correlated with **Fare** (0.26) and moderately negatively correlated with Pclass(-0.34).
- **Pclass** is negatively correlated with **Fare** (~-0.55), showing higher-class passengers paid higher fares.
- Age shows weak correlation with survival.

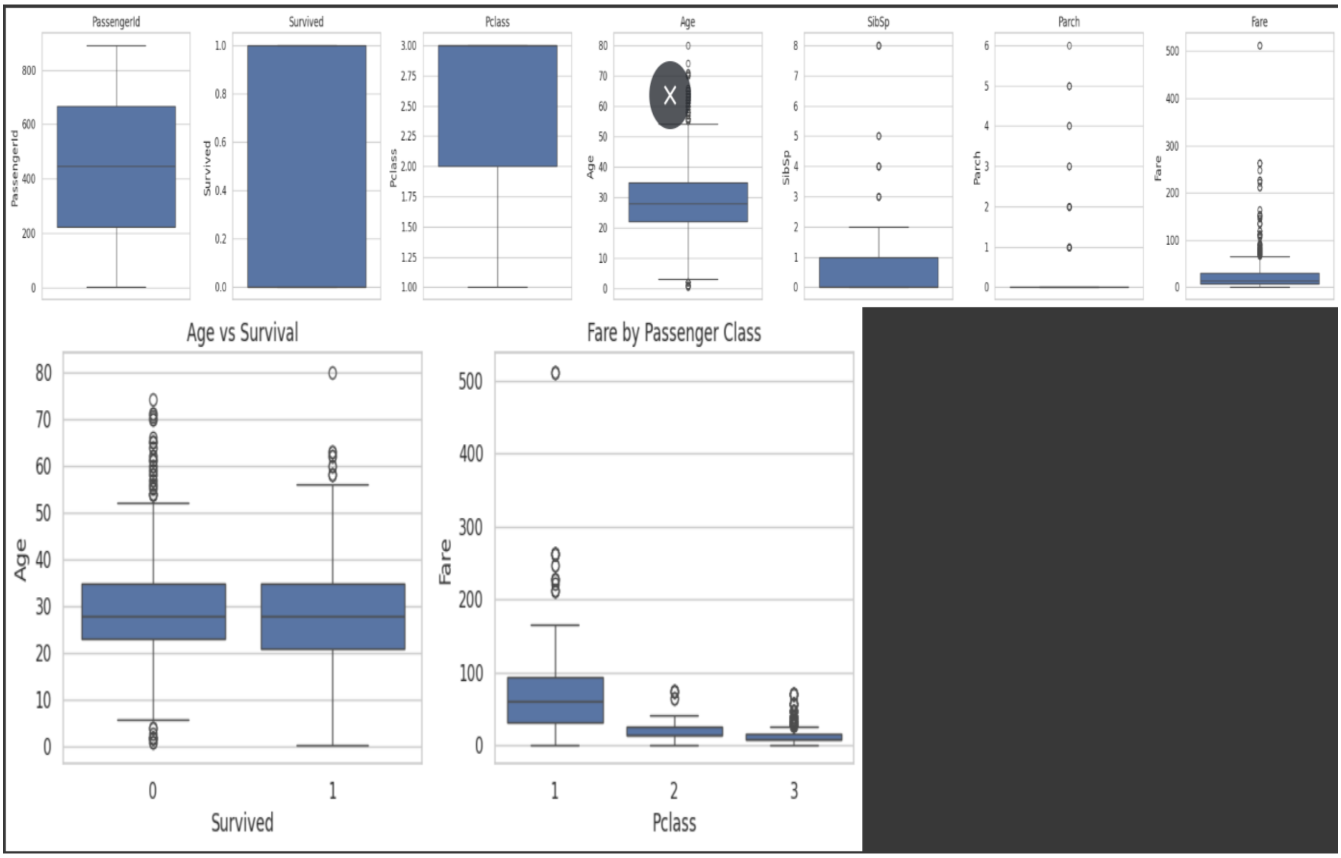
Histogram



Observation:

- **Age:** Most passengers were between 20–40 years old, with a slight peak around children aged 0–10.
- **Fare:** Positively skewed, with most fares below 100 and a few extreme outliers above 400.
- **SibSp & Parch:** Majority of passengers traveled alone (SibSp = 0, Parch= 0).
- **Survival:** Fewer passengers survived than perished.

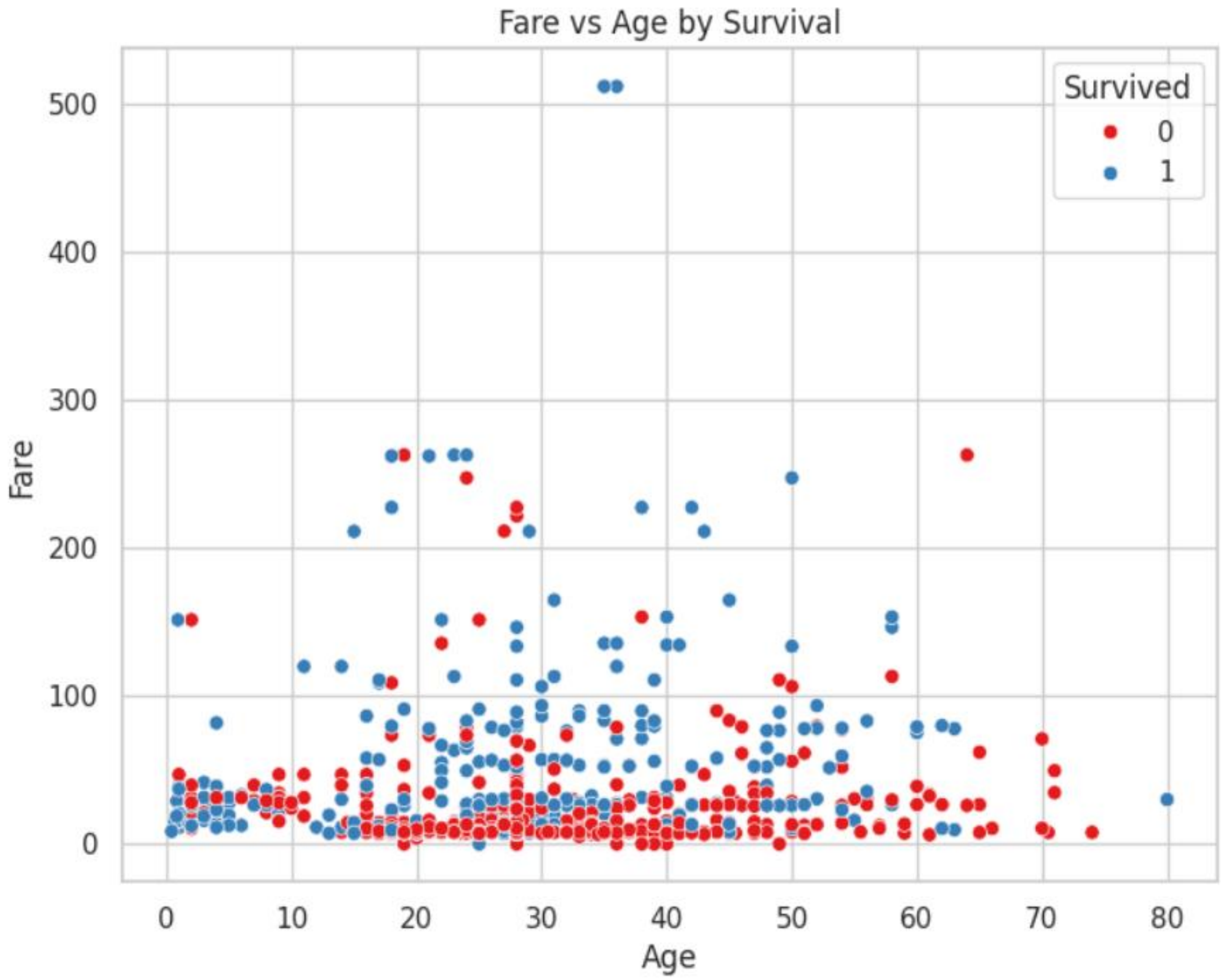
Box Plots



Observation:

- **Fare vs Pclass:** 1st class passengers paid significantly higher fares; 3rd class had lower fares with fewer outliers.
- **Age vs Pclass:** Age distribution is similar across classes, though 1st class had more elderly passengers.
- **Survival vs Fare:** Higher fares were generally associated with better survival chances.
- **Survival vs Age:** Younger children show higher survival rates.

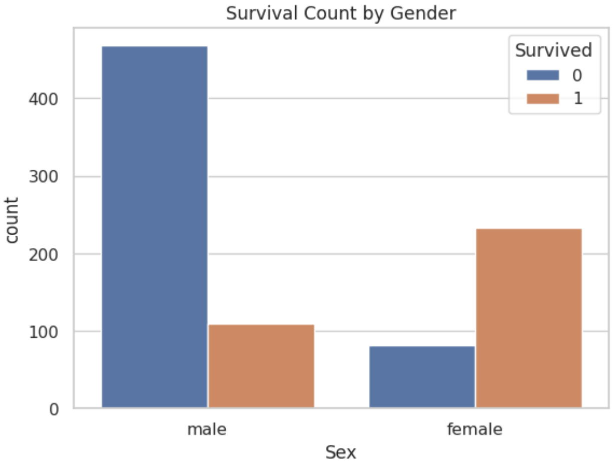
Scatter Plots

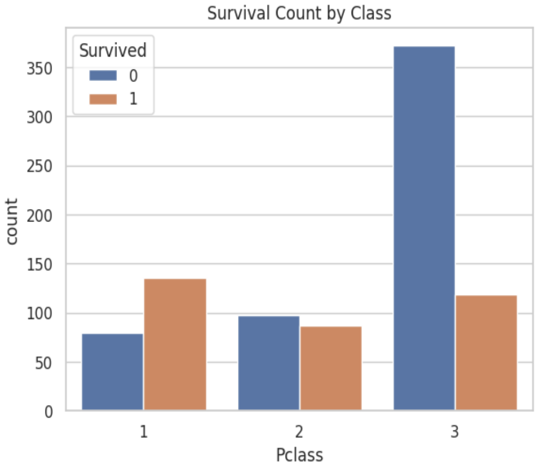


Observation:

- **Fare vs Age:** No clear correlation; high fares are scattered across ages, though most high-fare passengers were adults.
- **Fare vs Pclass:** Strong negative relationship — higher classes paid higher fares.
- **Survival vs Age/Fare:** Higher survival rates are concentrated among high-fare and younger passengers.

Count Plots





Observation:

- **Sex vs Survival:** Females had a much higher survival rate than males.
- **Pclass vs Survival:** 1st class had the highest survival rate, 3rd class the lowest.

Summary of Findings

1. The dataset contains both numerical and categorical variables, with some missing values that require handling (notably Age and Cabin).
2. Most passengers were male, in 3rd class, and embarked from Southampton.
3. Higher fares and higher classes are associated with better survival chances.
4. Females and younger passengers had a higher likelihood of survival.
5. Cabin information is missing for most passengers, which may limit certain analyses.
6. Correlation analysis shows Fare and Pclass are strong indicators of socio-economic status, which influenced survival rates.