TITANIC DATASET – EXPLORATORY DATA ANALYSIS REPORT

1. Introduction

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics using statistical methods and visualizations. The objective of this project is to analyze the Titanic dataset and extract meaningful insights regarding passenger survival patterns.

The analysis helps in identifying relationships, trends, missing values, and key influencing factors in the dataset.

2. Objective

The main objective of this project is:

- To explore and understand the Titanic dataset.
- To identify patterns and trends affecting passenger survival.
- To visualize relationships between different variables.
- To summarize key insights using statistical and graphical techniques.

3. Tools and Technologies Used

- Python
- Pandas (Data Analysis)
- Matplotlib (Visualization)
- Seaborn (Statistical Visualization)
- VS Code

4. Dataset Description

The Titanic dataset contains information about 891 passengers and 12 features:

- PassengerId
- Survived
- Pclass (Passenger Class)
- Name
- Sex
- Age
- SibSp (Number of siblings/spouses aboard)
- Parch (Number of parents/children aboard)
- Ticket
- Fare
- Cabin
- Embarked

The dataset provides details about passenger demographics, ticket class, and survival outcome.

5. Data Overview

Total Records: 891
Total Features: 12

Missing Values Identified:

- Age → 177 missing values
- Cabin → 687 missing values
- Embarked → 2 missing values

This indicates that Cabin has a significant number of missing entries and Age also requires handling before modeling.

6. Statistical Summary

From the statistical analysis:

- Average Age ≈ 29 years
- Maximum Age = 80 years
- Average Fare ≈ 32
- Majority passengers belonged to 3rd class

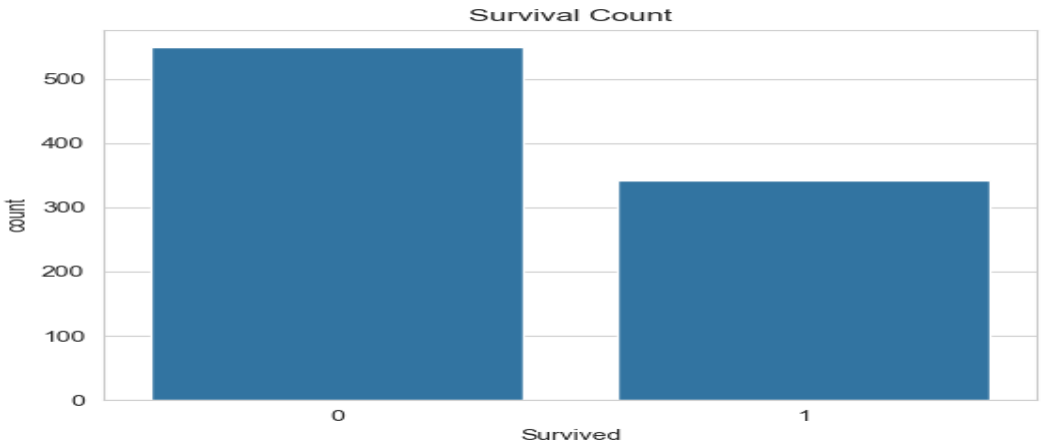
The survival count showed:

- 549 passengers did not survive
- 342 passengers survived

Thus, survival rate was lower than death rate.

7. Visual Analysis and Observations

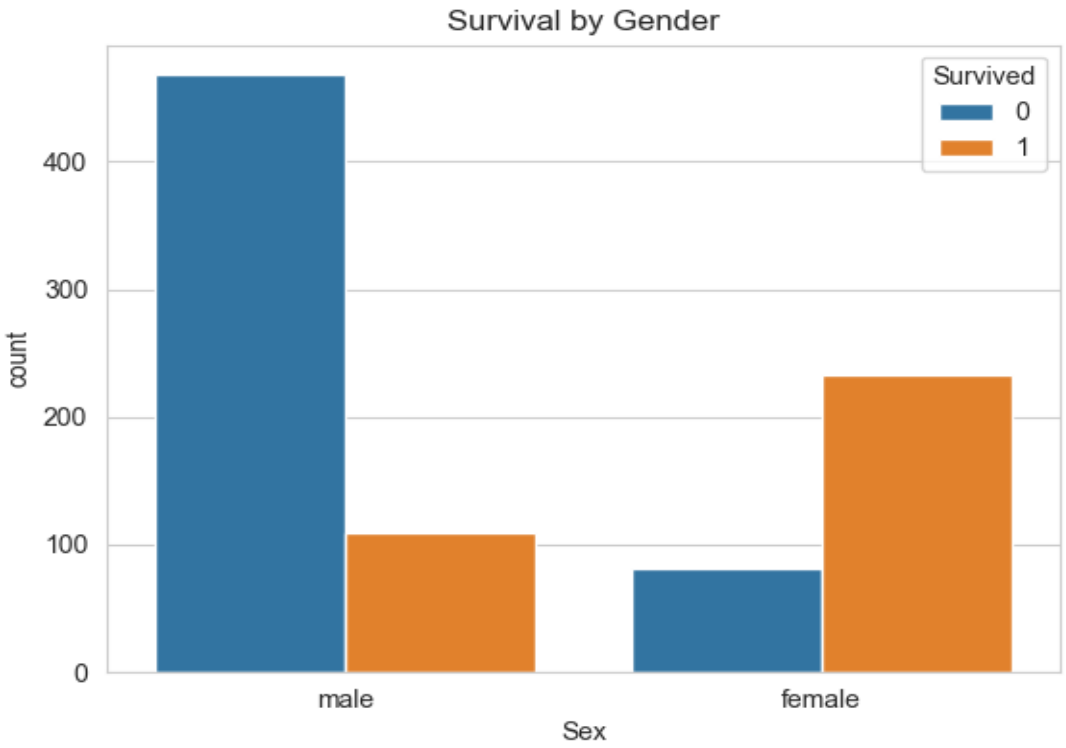
7.1 Survival Count



(survival_count.png)

Observation:
More passengers died than survived, indicating the severity of the disaster.

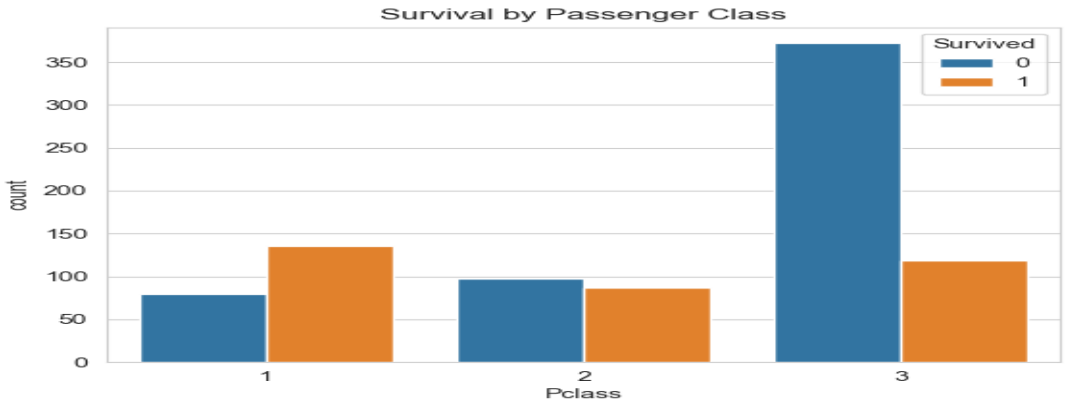
7.2 Survival by Gender



(survival_by_gender.png)

Observation:
Female passengers had a significantly higher survival rate compared to male passengers. This suggests priority rescue policy for women.

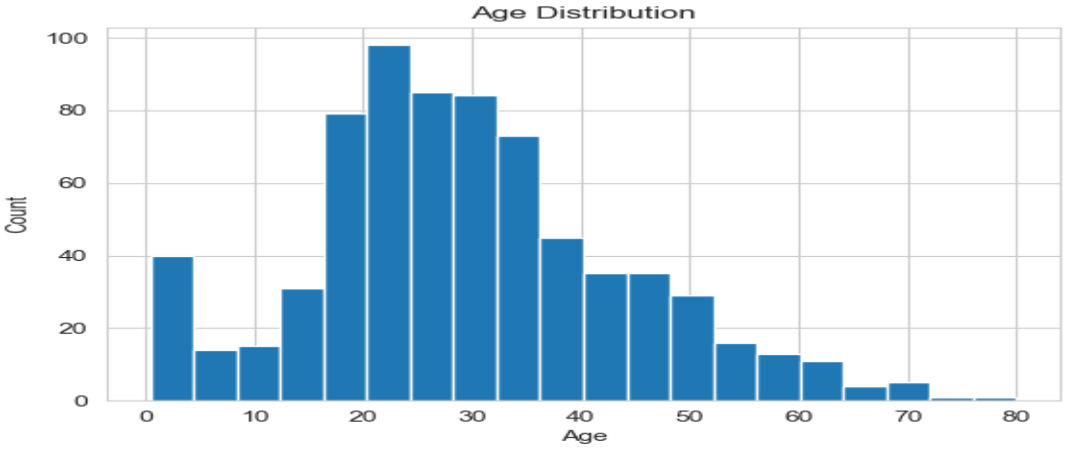
7.3 Survival by Passenger Class



(survival_by_class.png)

Observation:
First-class passengers had a higher survival rate compared to second and third class passengers. Socioeconomic status influenced survival probability.

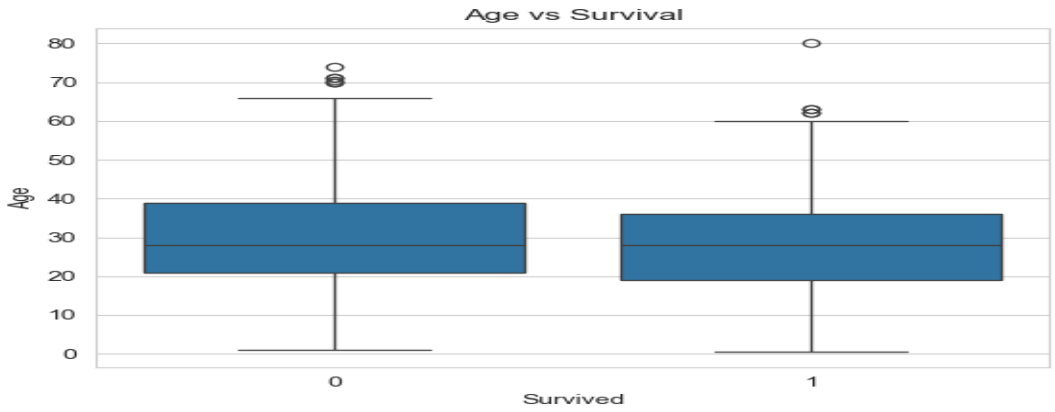
7.4 Age Distribution



(age_distribution.png)

Observation:
Most passengers were between 20–40 years of age. The distribution shows a concentration of young adults.

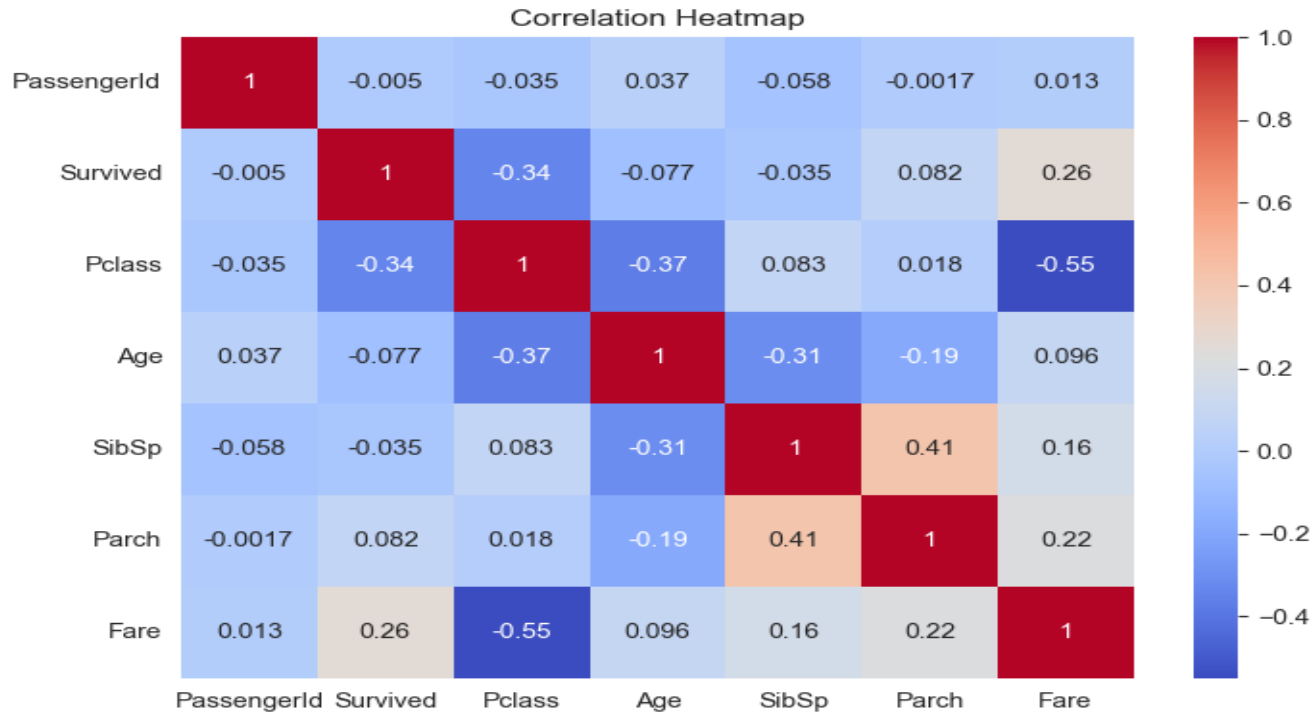
7.5 Age vs Survival (Boxplot)



(age_vs_survival.png)

Observation:
Younger passengers showed slightly higher survival chances compared to older passengers.

7.6 Correlation Heatmap



(correlation_heatmap.png)

Observation:
Passenger Class and Fare show correlation with survival. Higher fare and better class slightly increased survival probability.

8. Key Findings

- Female passengers had higher survival rates.
- First-class passengers were more likely to survive.
- Majority passengers were from third class.
- Age and Cabin columns contain missing values.
- Fare and Passenger Class influence survival outcomes.

9. Conclusion

The exploratory analysis of the Titanic dataset revealed meaningful patterns related to passenger survival. Gender and passenger class played a significant role in survival probability. Additionally, the dataset contains missing values that must be handled before applying predictive models.

EDA helped in understanding data structure, trends, and relationships, forming a strong foundation for further machine learning tasks.