TASK-04: TITANIC DATASET – EXPLORATORY DATA ANALYSIS

Submitted By: Syeda Faizah

Course: Data Analytics

Submission: December 2025

Task 04 – Titanic Dataset: Exploratory Data Analysis (EDA)

1. Introduction

The sinking of the RMS Titanic in 1912 is one of the most historically significant maritime disasters. The Titanic dataset provides demographic, social, and travel-related data of the passengers, enabling analysis of survival patterns using exploratory data analysis techniques. The objective of this study is to uncover meaningful insights regarding survival outcomes based on features such as age, gender, socioeconomic status, and passenger class.

In this project, Python and data science libraries are used to perform data cleaning, visualization, and statistical interpretation. This project enhances analytical thinking and builds essential skills required for beginner-level data analytics roles.

2. Problem Statement

Data exploratory analysis aims to answer the following key questions:

- Which groups of passengers were more likely to survive?
- How did age, gender, and socio-economic status influence survival?
- What patterns can be observed using visual exploration?
- How can missing values be treated to improve the reliability of analysis?

Through this analysis, we attempt to identify trends and correlations that explain survival outcomes.

3. Dataset Description

The Titanic dataset was downloaded from Kaggle: *Titanic – Machine Learning From Disaster*. The dataset contains records of passengers, including demographic, survival, and ticket information.

Attributes include:

Feature	Description
PassengerId	Unique ID
Survived	0 = No, 1 = Yes
Pclass	Ticket class
Name	Passenger name
Sex	Gender
Age	Passenger age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of boarding

4. Data Cleaning and Missing Value Handling

The dataset contains missing values, especially in Age, Cabin, and Embarked. Handling missing data improves the robustness of EDA.

Cleaning steps:

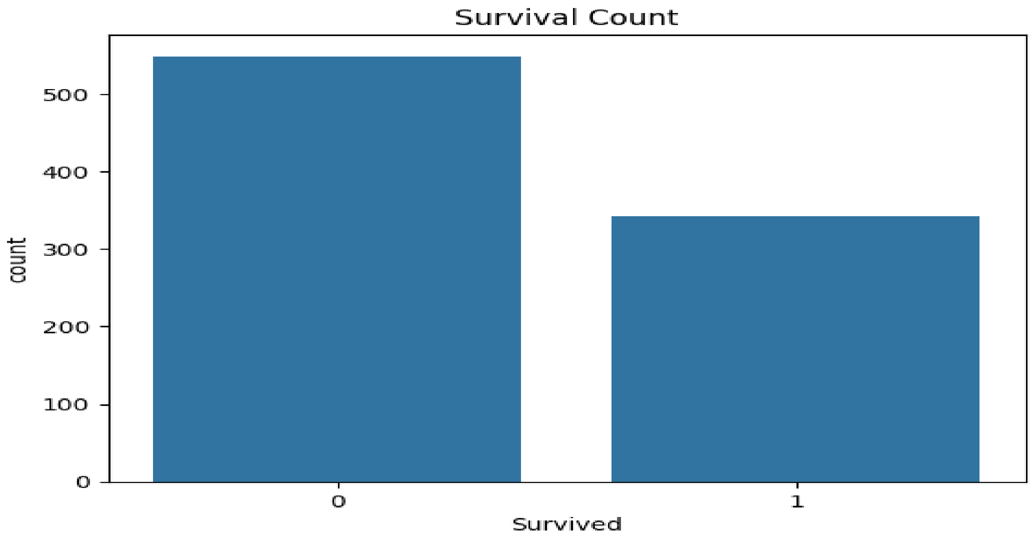
- Age missing values replaced using **median**
- Embarked filled using **mode**
- Cabin filled with **Unknown**

This allows complete visualization without removing essential rows.

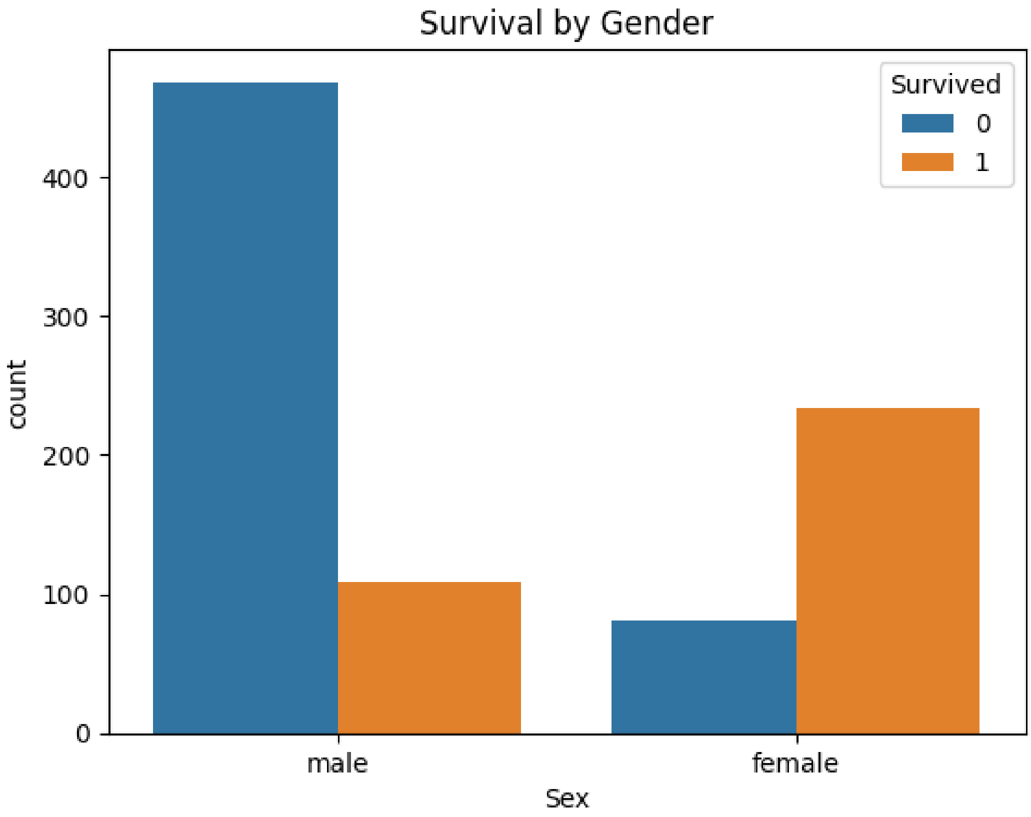
5. Univariate and Bivariate Analysis

Various visualizations were used to understand trends:

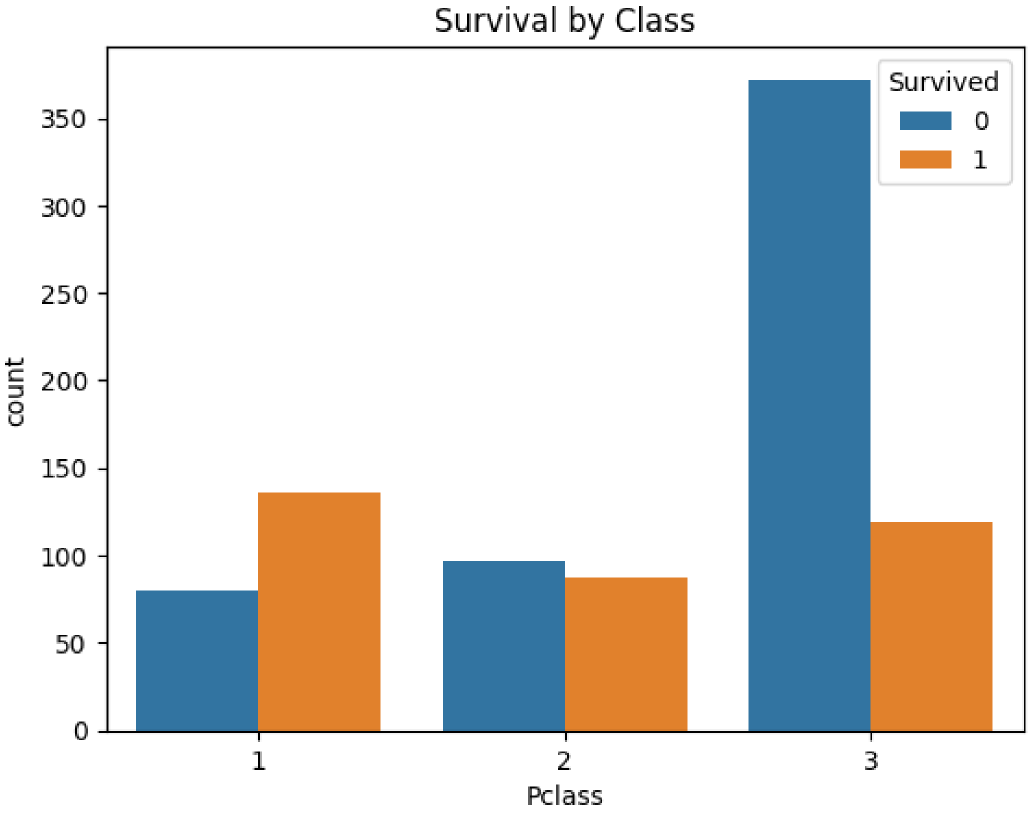
- **Survival Count:** Shows total survived vs not survived.



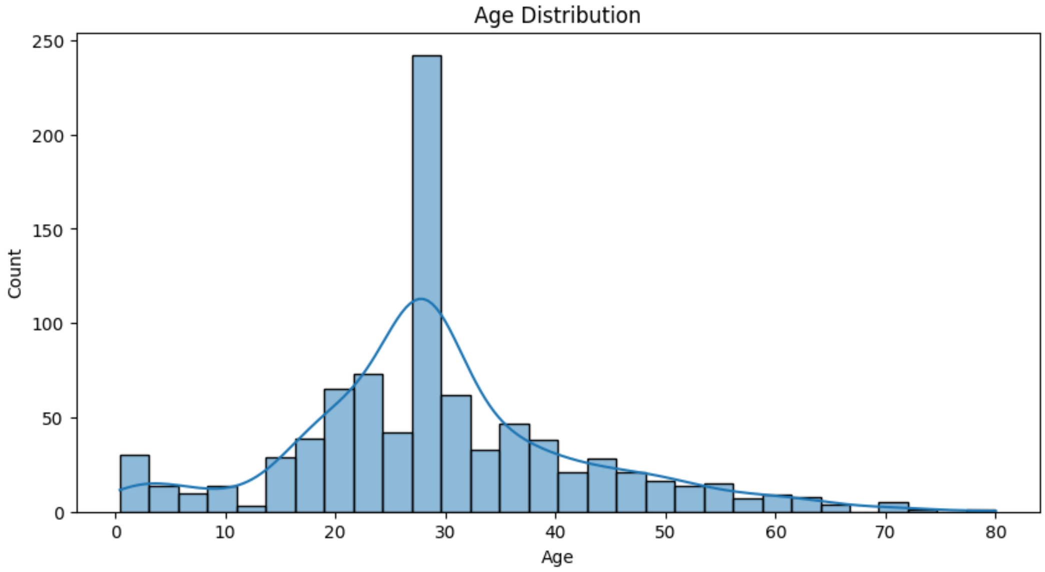
- **Survival by Gender:** Reveals that **female passengers had higher survival rates**.



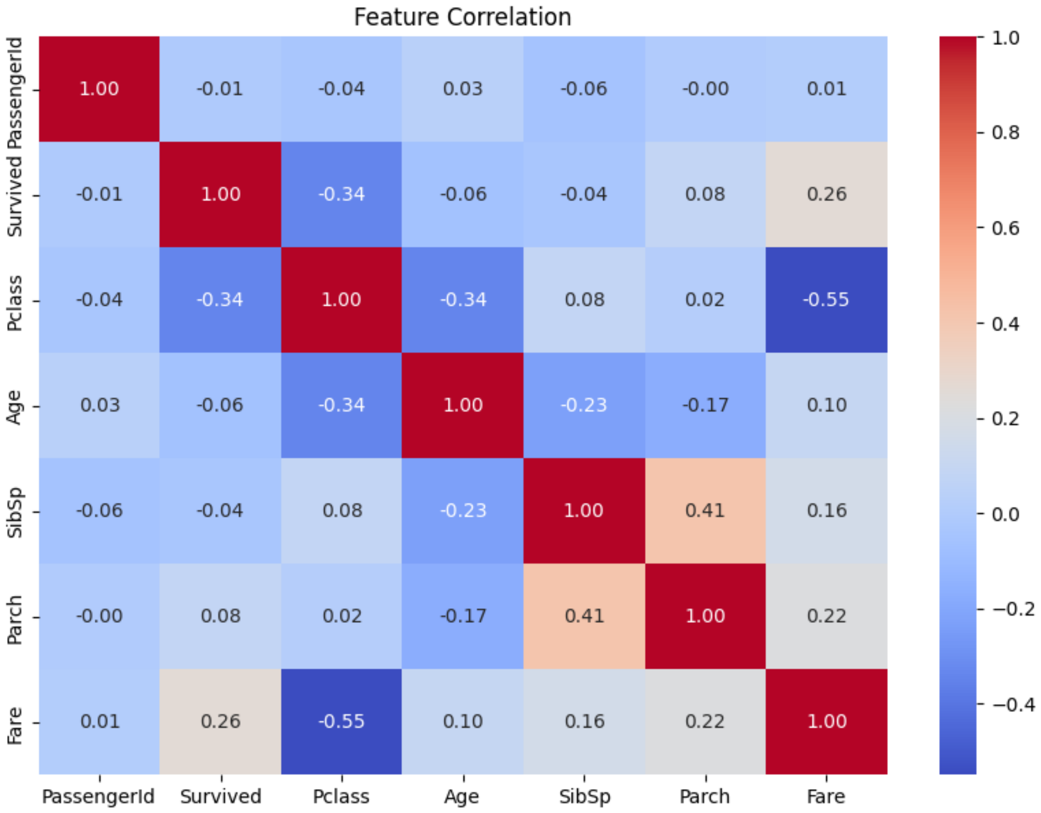
- **Survival by Passenger Class:** Passengers in higher classes (especially Pclass = 1) had greater chances of survival.



- **Age Distribution:** Most passengers onboard were young adults.



- **Heatmap Correlation:** Shows relationships between numerical variables and survival probability.



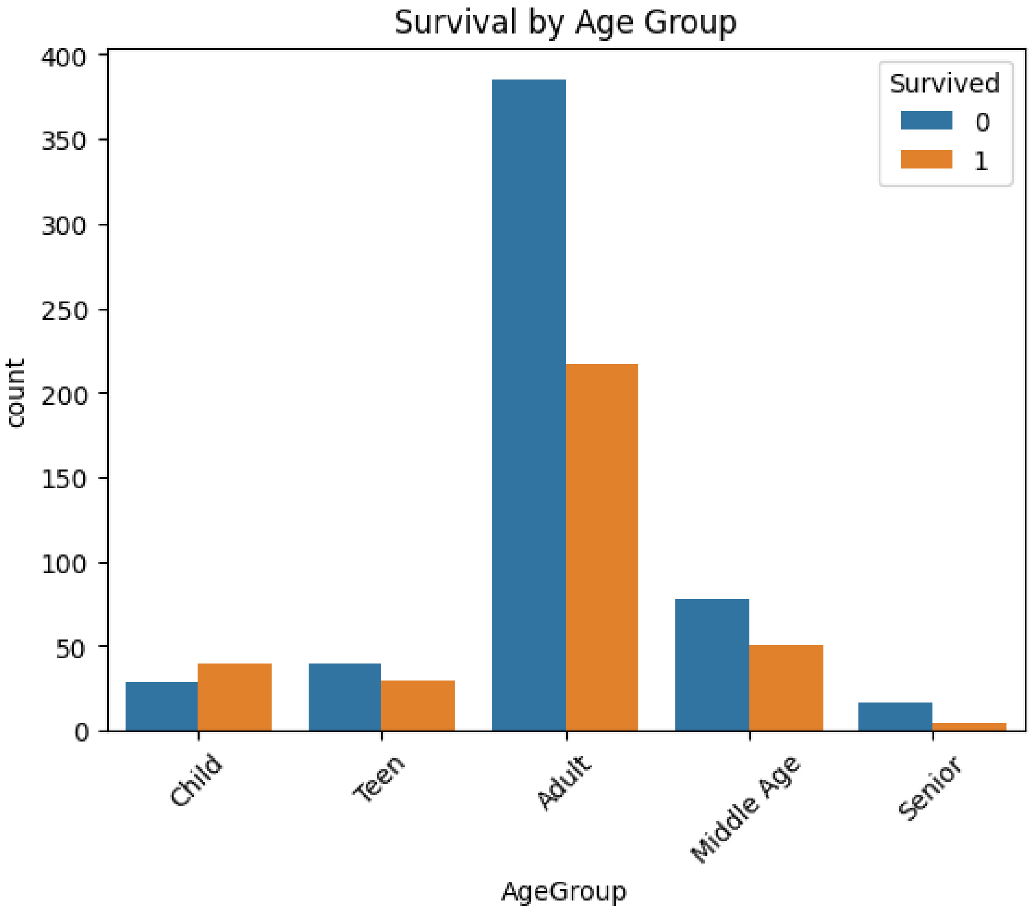
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.00	-0.01	-0.04	0.03	-0.06	-0.00	0.01
Survived	-0.01	1.00	-0.34	-0.06	-0.04	0.08	0.26
Pclass	-0.04	-0.34	1.00	-0.34	0.08	0.02	-0.55
Age	0.03	-0.06	-0.34	1.00	-0.23	-0.17	0.10
SibSp	-0.06	-0.04	0.08	-0.23	1.00	0.41	0.16
Parch	-0.00	0.08	0.02	-0.17	0.41	1.00	0.22
Fare	0.01	0.26	-0.55	0.10	0.16	0.22	1.00

6. Age Group Segmentation

A new categorical feature, AgeGroup, was created using age bins:

- Child (0–12)
- Teen (12–18)
- Adult (18–40)
- Middle Age (40–60)
- Senior (60+)

Analysis shows children and women had significantly higher survival probability.

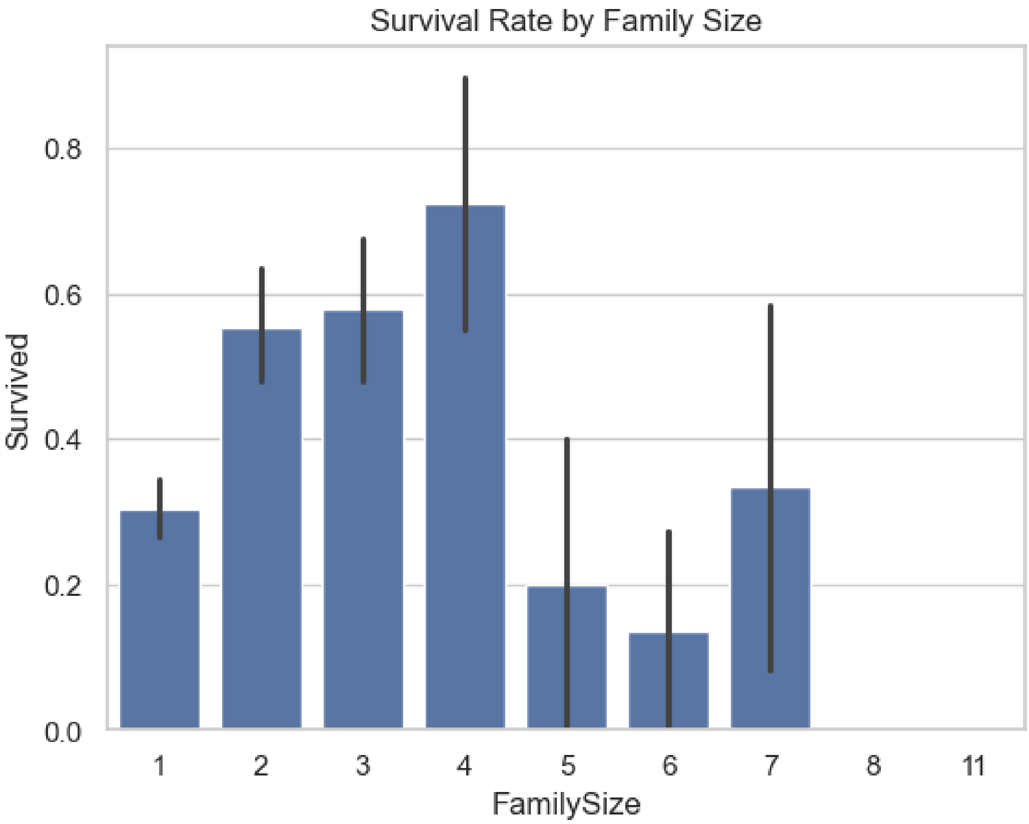


7. Advanced Feature Engineering

To uncover deeper survival patterns, three new features were engineered:

a) FamilySize

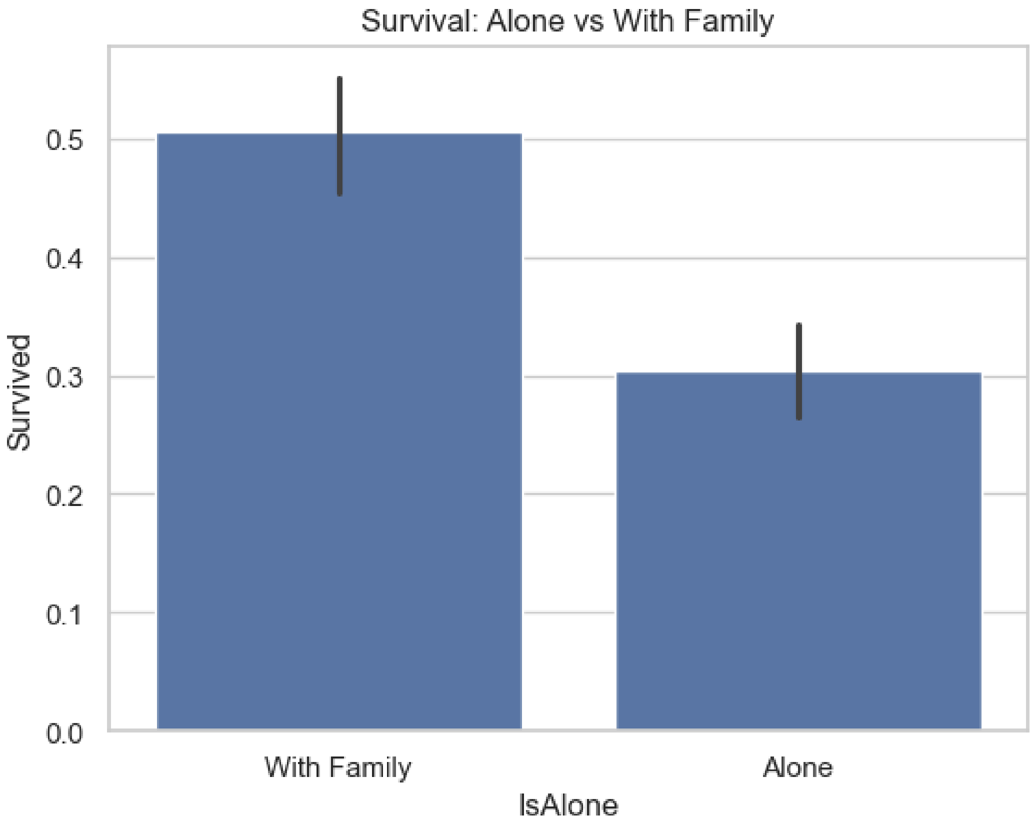
Calculated as **SibSp + Parch + 1**, representing how many family members a passenger traveled with.
Insight: Passengers with a **small family size (2–4)** had better survival, while those **alone** or with **very large families** had lower survival.



b) IsAlone

A binary feature:

- 1 = Passenger travelling alone
- 0 = Passenger with family
  Insight: Passengers traveling **alone had lower survival**, proving social support influenced survival.



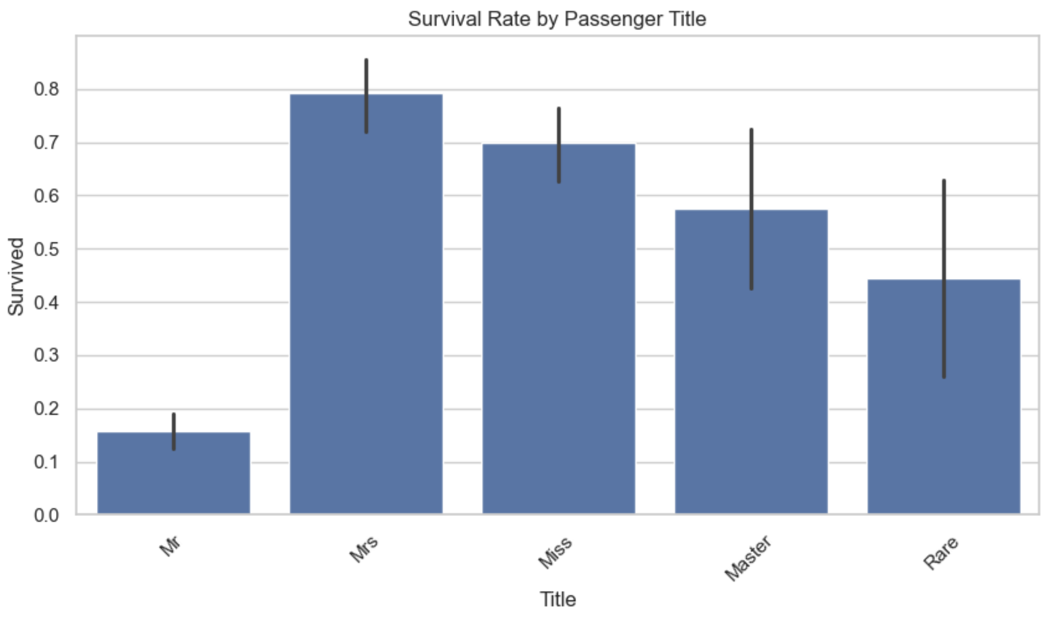
c) Title Extraction

Titles such as Mr, Mrs, Miss, Master, etc. were extracted from the Name column.
Rare titles were grouped into a “Rare” category.
Insight:

- **“Master” and “Mrs”** had the highest survival rates.
- **“Mr”** had the lowest.

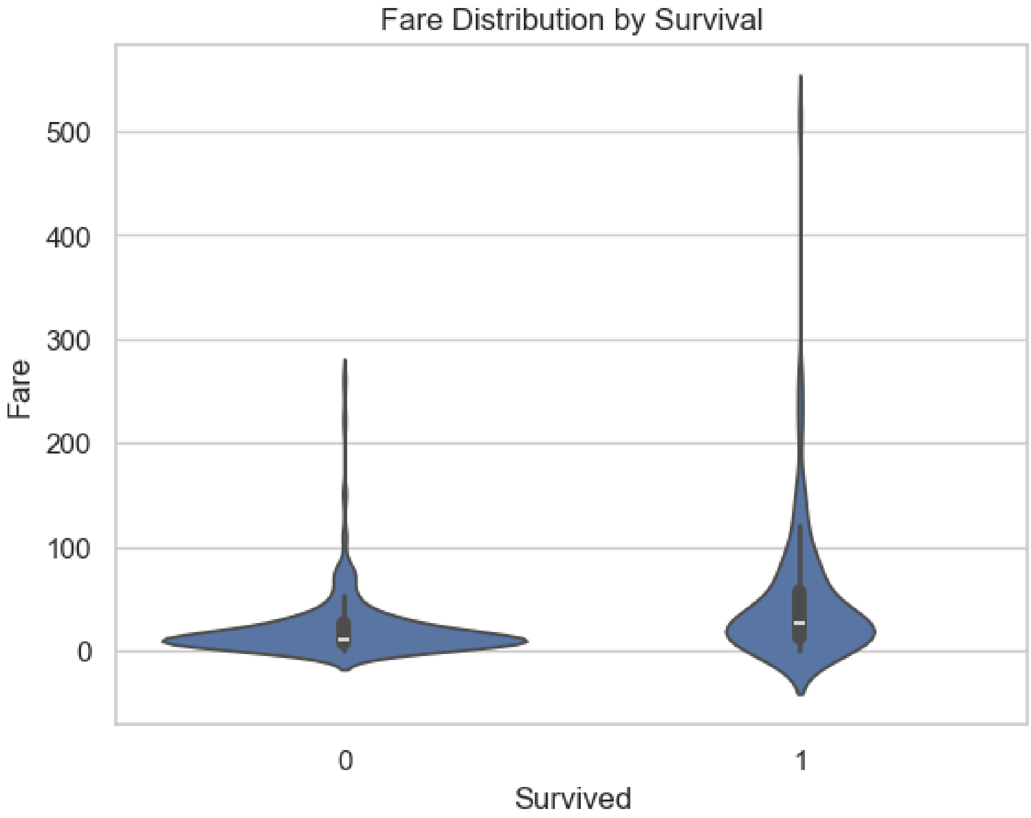
Highest and Lowest Title Survival Rates

- Master → High
- Mrs → High
- Miss → Moderate
- Mr → Very low
- Rare → Low or mixed



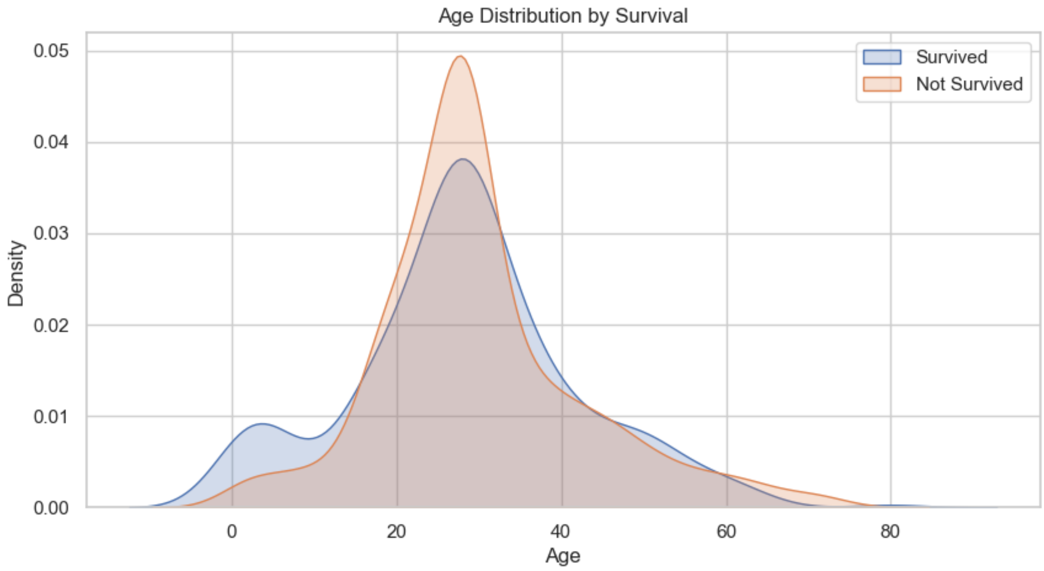
d) Fare Distribution (Violin Plot)

Higher fare passengers (1st class) survived more frequently.



e) KDE Age Distribution

KDE plots showed that survivors had a noticeably different age distribution than non-survivors.



8. Insights and Interpretation

Major Findings:

- Females had significantly higher survival probability.
- First-class passengers survived more than lower classes.
- Younger passengers had better survival outcomes.
- Fare price indirectly relates to socio-economic status and survival.
- Passengers traveling alone had lower survival rates.
- Family size played a crucial role in survival outcomes.
- Titles extracted from passenger names strongly predicted survival.
- Higher fare correlated with higher survival probability.
- KDE plots revealed clear age-related survival patterns.

9. Technologies Used

- Python
- Pandas

- Numpy
- Matplotlib
- Seaborn
- Jupyter Notebook

10. Skills Learned

- Data Cleaning
- Handling missing values
- Feature creation using age grouping
- Exploratory visualization
- Statistical summary analysis
- Heatmap correlation analysis

11. Conclusion

The Titanic dataset provides a strong foundation for applying data analytics and visualization. The impact of gender, age, and passenger class strongly influenced survival. Through EDA, meaningful insights were gained which builds beginner-level analytical thinking, making this project highly suitable for internship, academic portfolio, and GitHub documentation.

12. Reference

Dataset Source:
Kaggle – Titanic: Machine Learning from Disaster
Video Reference:
Titanic EDA project (for conceptual understanding only)