Titanic Dataset Analysis – EDA Summary

Introduction

This project explores the Titanic dataset (df), which contains details of 891 passengers. The goal was to understand survival patterns using exploratory data analysis (EDA). We focused only on the training data and did not use any modeling or prediction.

Data Cleaning

- Missing values in the Age column were replaced with "Unknown" to preserve the structure.
- Null values in the Embarked column were filled using the mode, which was 'S' (Southampton).
- A new column called FamilySize was created by adding SibSp + Parch + 1 to represent total family members aboard.

Visualizations & Insights

1. Survival Distribution
- Out of 891 passengers, 342 survived and 549 did not.
- The survival rate was around 38%.

2. Survival by Gender
- Females had a much higher survival rate than males.
- Most women survived, while most men did not.

3. Survival by Passenger Class
- 1st class passengers had the highest survival rate.
- 3rd class passengers had the lowest survival rate.
- This shows that class played a big role in survival chances.

4. Fare Distribution
- Most passengers paid low fares (under ₹100).
- A few passengers paid very high fares, up to ₹550.
- These high-paying passengers were likely in 1st class.

5. Family Size Distribution
- Most people traveled alone (FamilySize = 1).
- Very large families (5 or more) were rare and had lower survival.
- Small families had slightly better survival chances.

6. Embarkation Port Distribution
- Most passengers boarded from Southampton (72.5%).
- Cherbourg had 18.9%, and Queenstown had 8.5%.
- The port of embarkation is related to class and survival indirectly.

Conclusion

This EDA helped me understand how gender, class, fare, family size, and embarkation point influenced survival on the Titanic. These insights are useful for building models later, but for now, they show how different groups were affected during the disaster.