Titanic Dataset – Exploratory Data Analysis (EDA) Report

Objective

To explore the Titanic dataset and find patterns, trends, and relationships that help us understand what factors influenced passenger survival.

Step 1: Basic Information

- The dataset contains details of passengers like age, gender, ticket class, fare, and whether they survived.
- We used .info(), .describe(), and .value_counts() to understand the structure and key statistics.

Observations:

- There are missing values in Age, Cabin, and Embarked.
- Most passengers were in 3rd class.
- Around 38% of passengers survived.

Step 2: Data Cleaning

- Filled missing Age values with the median age.
- Filled missing Embarked values with the most common port.
- Dropped the Cabin column due to too many missing values.
- Removed duplicate rows.

Step 3: Visual Analysis

1. Survival Count

- More passengers died than survived.
- Survival rate was around 38%.

2. Survival by Gender

- Females had a much higher survival rate than males.

3. Survival by Class

- 1st class passengers had the highest survival rate.
- 3rd class had the lowest.

4. Age Distribution

- Most passengers were between 20 and 40 years old.
- Children and elderly were fewer in number.

5. Fare Distribution

- Most fares were low, but a few passengers paid very high fares.

6. Boxplot: Age vs Survival

- Younger passengers had slightly better chances of survival.

7. Scatterplot: Fare vs Age

- Passengers who paid higher fares were often older and had better survival chances.

8. Correlation Heatmap

- Survival is positively correlated with Fare and Pclass.
- Age has a weak negative correlation with survival.

Summary of Findings

- Gender played a major role in survival — females were more likely to survive.
- Class mattered — higher-class passengers had better chances.
- Fare and Age showed some influence, but not as strong.
- Embarked port and family size could be explored further in future analysis.