Titanic Dataset – Exploratory Data Analysis (EDA) Report

📌 Objective:

To analyze the Titanic dataset to identify patterns and insights that influenced passenger survival during the Titanic disaster.

📂 Dataset Description:

- **Source:** train.csv from the Titanic Machine Learning dataset
- **Features analyzed:**
  - Survived (target)
  - Age, Sex, Pclass, Fare, Embarked, SibSp, Parch

🔍 Steps Performed:

1. Data Loading & Inspection

- Loaded using pandas
- Initial inspection with:
  - .info() for data types & null values
  - .describe() for summary statistics
  - .isnull().sum() to check missing data

2. Handling Missing Values

- No imputation steps were shown, but missing data was quantified, especially in columns like Age and Cabin.

📊 Visual Analysis:

✅ Plot 1: Survival Count

- Used seaborn.countplot() on the Survived column
- Insight: More people died than survived (class imbalance).

✅ Plot 2: Age Distribution

- Histogram with KDE for Age
- Insight: Most passengers were in the 20–40 age group.

✅ Plot 3: Survival by Gender

- Count plot comparing Sex vs Survived
- Insight: Women had a significantly higher survival rate than men.

✅ Plot 4: Survival by Passenger Class (Pclass)

- Count plot of Pclass vs Survived
- Insight: First-class passengers had higher survival rates, indicating class-based disparity.

📈 Potential Further Steps (not in notebook but recommended):

- Impute missing values (e.g., median age).
- Correlation heatmaps for numeric features.
- Feature engineering (e.g., creating FamilySize, encoding Sex and Embarked).
- Predictive modeling using logistic regression or decision trees.

📝 Conclusion:

This EDA provided a foundational understanding of factors affecting survival aboard the Titanic. Notably:

- Gender and class were strong indicators of survival.
- Age distribution offers demographic insights into passengers.