Project: Titanic Dataset – Exploratory Data Analysis (EDA)

---

1. Data Structure Overview

- **Dataset Size:**
  The Titanic dataset contains **891 rows** and **15 columns**.
- **Suitability:**
  This dataset is relatively small but **sufficient for learning basic Machine Learning (ML) classification algorithms** and understanding data preprocessing techniques.

---

2. Variable Identification

Based on df.head() output and manual inspection, the dataset features can be classified as follows:

a) Numerical Variables

- age
- fare
- sibsp (Number of siblings/spouses aboard)
- parch (Number of parents/children aboard)

b) Categorical Variables (Nominal)

- sex
- embarked
- who
- deck

c) Categorical Variables (Ordinal)

- class (First > Second > Third)
- pclass (1 > 2 > 3)

d) Binary Variables

- adult_male (True / False)
- alone (True / False)

e) Target Variable

- survived
  - 0 → Did not survive
  - 1 → Survived

Since the target variable is binary, this is a **Classification problem**.

---

3. Data Quality & Observations

a) Missing Values

- **Deck column:**
  Contains approximately **77% missing values**.
  ➤ Recommended to **drop this column**.
- **Age column:**
  Has around **20% missing values**.
  ➤ Requires **imputation** (mean/median/group-based) before modeling.

---

b) Class Imbalance

- About **61% passengers did not survive**
- About **38% passengers survived**

This is a **slight imbalance**, but it is acceptable for standard ML models.

---

4. ML Readiness Conclusion

- **Dataset Status:** Partially Ready
- **Actions Required Before Modeling:**
    1. Handle missing values (especially age)
    2. Convert categorical variables to numerical form
        - Example: male/female → 0/1
    3. Drop unnecessary or highly missing columns

After these steps, the dataset is **ready for ML model training**.

Summary of Commands Used

Command	Purpose
pd.read_csv('file.csv')	Load data into a DataFrame
df.head() / df.tail()	View initial and final rows
df.info()	Check data types and missing values
df.describe()	Statistical summary
df.columns	List column names
df['column'].unique()	View distinct category values
