🗒️ Titanic Dataset - Exploratory Data Analysis (EDA) Report

🔢 1. Dataset Overview

- **Total Passengers**: 891
- **Features of Interest**: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked
- **Missing Values**:
  - Age: ~19.8% missing
  - Cabin: ~77% missing (significant)
  - Embarked: 2 missing entries

📊 VISUAL ANALYSIS

🟦 2. Histograms: Age and Fare Distribution

Visual: sns.histplot for Age and Fare
Relationships/Trends:

- Age follows a **bell-like curve**, peaking at 20–30 years.
- Fare is **right-skewed** with outliers (some paid > $200).

Observations:

- **Younger adults** were the most common travelers.
- Most passengers paid **low fares**, suggesting many were lower-class.

🟩 3. Count Plots: Survived and Pclass

Visual: sns.countplot
Relationships/Trends:

- Most passengers in **Pclass 3** (lowest class).
- More people **died (Survived = 0)** than survived.

Observations:

- 3rd class made up the **largest portion**, but had **low survival**.
- Indicates **wealth/class was a survival factor**.

🟧 4. Boxplots: Age and Fare by Pclass

Visual: sns.boxplot for Pclass vs Age and Fare
Relationships/Trends:

- 1st class passengers were **older** on average.
- Fare increases sharply from Pclass 3 to 1.

Observations:

- Higher-paying passengers were **older and wealthier**.
- Fare has **outliers** in all classes but extreme in 1st.

🟥 5. Heatmap: Correlation Matrix

Visual: sns.heatmap()
Key Correlations:

- Survived vs Pclass: **-0.34**
- Survived vs Fare: **+0.26**
- Age, SibSp, and Parch have weak correlation with survival

Observations:

- Passengers in **higher class had better survival rates**.
- Fare and class are the most predictive among numerics.

🟪 6. Pairplot: Survival by Key Features

Visual: sns.pairplot() with Survived as hue
Trends:

- Survivors cluster in **low Pclass and high Fare**.

- Many **non-survivors** were in **Pclass 3 and paid low fares**.

Observations:

- Strong visual evidence that **wealth and class helped survival**.
- Survivors also had a **slightly wider age range**.

📌 **Summary of Findings**

1. **Survival Patterns**:
    - Higher **class and fare** → higher chance of survival.
    - Most **non-survivors** were in **3rd class**.
2. **Feature Importance**:
    - **Pclass**, **Fare**, and (not shown but known) **Sex** strongly influence survival.
3. **Missing Data**:
    - Age needs **imputation** (mean/median or model-based).
    - Cabin is too incomplete to use directly.
4. **Outliers**:
    - Fare has **extreme values**—consider transformation (e.g., log scale).
5. **Next Steps**:
    - Handle missing data and encode categoricals.
    - Consider feature engineering (e.g., FamilySize, Title from Name).

📌 Key Insights from EDA:

- Majority of passengers were in 3rd class; most of them did not survive.

- Females had higher survival rates (not shown here but evident if analyzed).

- Passengers in 1st class were generally older and paid higher fares.

- Survival was positively related to higher class and fare.

- Age and Fare distributions are skewed; handling missing values and outliers is key for modeling.