🧾 Exploratory Data Analysis (EDA) Report

🔢 Column Analysis

Numerical Columns

- **Age**: Investigated for distribution, central tendency, and missing values. Possibly skewed with some outliers.
- **Fare**: Analyzed for skewness and outliers. Likely shows right-skew due to a few passengers with expensive fares.
- **PassengerId**: Treated as an identifier; not used in analysis.

Categorical Columns

- **Survived**: Target variable (0 = No, 1 = Yes). Class imbalance may exist.
- **Pclass**: Passenger class (1, 2, 3). Shows correlation with survival—higher classes have better survival rates.
- **Sex**: Strong indicator of survival. Females had higher survival probabilities.
- **SibSp / Parch**: Number of siblings/spouses or parents/children aboard. Grouped for family size analysis.
- **Embarked**: Port of embarkation (C, Q, S). Distribution examined; some missing values noted.

Mixed Columns

- **Name, Ticket, Cabin**: Used for feature engineering, such as extracting titles or cabin prefixes. Many missing values in **Cabin**.

🔗 Relationships and Correlations

- **Survival vs Pclass**: Clear trend—1st class had higher survival rates.
- **Survival vs Sex**: Female passengers significantly more likely to survive.
- **Fare vs Survival**: Passengers who paid more were likelier to survive (likely due to class correlation).
- **Age vs Survival**: Children had higher survival rates; elderly showed lower.
- **Family Size (SibSp + Parch)**: Moderate family sizes (2-4) had better survival odds compared to solo travelers or large families.

📈 Trends and Patterns

- **Skewness** observed in `Fare`, handled possibly via log transformation.
- **Missing Values**:
  - **Age**: Imputed using median or based on similar passengers (e.g., by class or sex).
  - **Cabin**: Mostly missing—potentially dropped or used as a binary indicator.
  - **Embarked**: A few missing; filled with mode (most common value).
- **Outliers** detected in `Fare` and `Age` via boxplots and distribution analysis.
- **Distribution Shapes** analyzed (normal, skewed, bimodal) to guide modeling choices.

📊 Visualization Techniques Used

- Histograms, Boxplots, KDE plots
- Count plots for categorical data
- Correlation heatmaps for numerical features
- Bar plots comparing survival rates across categories