Overall Observations & Findings – Titanic EDA

1. **Passenger Demographics**
    - Majority of passengers were between **20–40 years old**.
    - There was a significant number of children and elderly passengers, but they formed a smaller proportion.
    - More **male passengers** traveled compared to females.
2. **Survival Patterns**
    - Overall survival rate was **~38%**, meaning most passengers did not survive.
    - **Females had a much higher survival rate** than males.
    - **1st class passengers** had the highest survival rate, followed by 2nd class; 3rd class had the lowest.
3. **Economic Factors**
    - **Fare** was a strong indicator of survival — passengers who paid higher fares had better survival chances, often linked to being in higher classes.
    - **Pclass** and Fare are inversely related (1st class = lower Pclass number but higher fare).
4. **Family Influence**
    - Most passengers traveled alone (SibSp = 0, Parch = 0).
    - Those with a small number of family members (1–2) had slightly higher survival chances compared to those traveling alone or in very large groups.
5. **Missing Data**
    - **Cabin** had a large proportion of missing values (>75%); handled by creating a Cabin_Available flag.
    - **Age** missing values were imputed with median age based on Pclass and Sex.
    - **Embarked** missing values were filled with the most common value ('S').
6. **Outlier Handling**
    - Detected and capped extreme values in **Age, SibSp, Parch, and Fare** using the IQR method.
    - This reduced skew in visuals and made distributions cleaner without removing data points.
7. **Key Correlations**
    - **Fare** positively correlated with survival; **Pclass** negatively correlated with survival.
    - Most other numerical variables had weak correlations with survival.
8. **Actionable Insights**

  - Economic status (class and fare) and gender were the most influential factors in survival.
  - Data can be used to build predictive models for survival probability using features like Pclass, Sex, Age, Fare, and family size.

Observations from EDA

1. Pairplot – Relationships & Trends

- Passengers who paid higher fares generally had higher survival rates.
- 1st class passengers (low Pclass value) were more likely to survive.
- Age distribution overlaps for survivors and non-survivors, but children had better survival chances.

---

2. Heatmap – Correlation Analysis

- **Fare** has a strong negative correlation with **Pclass** (higher class = higher fare).
- **Survived** is positively correlated with **Fare** and negatively correlated with **Pclass**.
- Other numeric variables show weak correlations with survival.

---

3. Histograms – Numeric Distributions

- Majority of passengers were aged between 20–40 years.
- Most passengers paid lower fares; very high fares were rare.
- Most passengers traveled without many siblings/spouses or parents/children.

---

4. Boxplots – Outlier Check

- Outliers in Age, SibSp, Parch, and Fare have been capped, resulting in cleaner distributions.
- Fare still shows skewness due to a few very expensive tickets in 1st class.

---

5. Scatterplot – Age vs Fare by Survival

- Higher survival rates among high-fare passengers, regardless of age.
- Low-fare passengers had significantly lower survival chances.

---

6. Countplots – Categorical Insights

- Female passengers had a much higher survival rate than males.

- 1st class passengers had the highest survival rate, followed by 2nd class; 3rd class had the lowest.