EDA for Titanic Dataset

1. Dataset Overview

- The dataset contains data on passengers aboard the Titanic.
- Key features include: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked.

2. Libraries Used

- pandas for data manipulation
- matplotlib.pyplot and seaborn for visualization

3. Data Cleaning Techniques & Their Impact

a. Handling Missing Values

- Age: Imputed using median or group-based medians (e.g., by Pclass and Sex).
    - *Impact*: Prevented loss of data but could suppress age variability; important to preserve survival analysis by age groups.
- Embarked: Missing values filled with the mode ('S').
    - *Impact*: Simplifies analysis but may obscure real embarkation-survival trends if missing values weren't random.
- Cabin: Heavily missing; either dropped or used to extract deck information.
    - *Impact*: Dropping loses potential insight (deck could indicate survival probability); extracting deck enables deeper spatial analysis**.**

b. Outlier Handling

- Fare: Outliers capped or log-transformed for normalization.
    - *Impact*: Prevents skew in plots and models, revealing better correlations in lower fare ranges.

c. Row Removal vs. Imputation

- *Row removal*: Ensures clean data but may bias results (e.g., toward younger passengers with known ages).
- *Imputation*: Retains dataset size but risks masking true data variance.

4. Univariate Analysis

- **Survival Rate**: Checked distribution of survivors vs non-survivors.
- **Pclass**: Majority of passengers were in 3rd class.
- **Sex**: Notable survival bias toward females.
- **Age**: Broad age distribution, with many passengers between 20-40 years old.
- **Fare**: Skewed fare distribution; some passengers paid very high fares.

5. Bivariate Analysis

- **Sex vs Survival**: Females had a much higher survival rate.
- **Pclass vs Survival**: Passengers in 1st class were more likely to survive.
- **Age vs Survival**: Children had slightly better survival odds.
- **Embarked vs Survival**: Point of embarkation influenced survival, with 'C' embarkation showing higher survival.

6. Visualizations

- Count plots, bar plots, histograms, and box plots used extensively.
- Correlation heatmaps showed relationships between numeric features.

7. Key Insights

- Gender and class were the strongest predictors of survival.
- Age and fare played secondary roles.
- Most survivors were women and children in higher classes.
- Data required preprocessing for missing values and normalization.

8. Next Steps

- Apply feature engineering for model training.
- Build classification models (e.g., logistic regression, decision trees).