Final Summary – Titanic3 Dataset EDA

➢ Dataset Overview

The dataset contains passenger-level data from the Titanic, including details such as:

sex, age, pclass, fare, embarked, home.dest, survived, etc.

Objective: To clean the data and explore factors influencing survival.

➢ Missing Value Summary (Before Cleaning)

Column	Missing Values
age	263
fare	1
embarked	2
home.dest	153
cabin	1014
boat	823
body	1188

➢ Data Cleaning Steps

Dropped:

cabin, boat, body due to excessive missing data and irrelevance.

Filled:

age → with median (around 28–30)

embarked → with mode (likely "S")

Verified that the final DataFrame has more less missing values.

➢ EDA Visualizations & Insights

1. Survival Count

- About 38% survived, 62% did not.
- Sets the context for further breakdown by gender, class, etc.

2. Survival by Gender

- Female survival rate ~70%, male survival ~20%.
- Strong indicator that gender significantly influenced survival.

3. Survival by Passenger Class

Survival rate:

- 1st class: Highest
- 2nd class: Medium
- 3rd class: Lowest

Indicates socioeconomic bias in survival.

4. Age Distribution

- Most passengers were between 20–40 years.
- Children and elderly passengers were fewer.
- No direct correlation with survival alone.

5. Fare by Class (Boxplot)

- Higher classes paid more — clear in fare distribution.
- Some high outliers in 1st class.

6. Correlation Heatmap

- Strong correlation: fare ↔ pclass
- Moderate: pclass ↔ survived
- Weak/none: age ↔ survived

7. Pair Plot

- Visualizes relationships between features like fare, age, pclass, etc.
- Survival distribution shown with blue (0 = not survived) and orange (1 = survived).
- Higher fares seem correlated with higher survival.

8. Age & Gender vs Survival (Violin Plot)

- Young women had best survival density.
- Males had low survival across all ages.

➢ Conclusion

- Strongest survival predictors: Gender, Class, and Fare
- Women, children, and 1st-class passengers had the highest survival chances.
- The cleaned dataset allowed clear visual insights and pattern discovery.