

Titanic EDA - Summary of Findings

1. Missing Values

- **Age:** Approximately 20% missing. Instead of dropping the rows, values were imputed based on average age grouped by passenger class (Pclass), since older passengers tended to be in higher classes.
 - **Cabin:** Too many missing values, so the column was dropped from analysis.
 - **Embarked:** Only one missing value found; the corresponding row was removed.
-

2. Survival Distribution

- **Women** had significantly higher survival rates compared to **men**.
 - **Children** and **first-class** passengers had the highest survival chances.
 - **Third-class** passengers had the lowest survival rate overall.
-

3. Passenger Class Insights

- **First-class** passengers were typically older and wealthier, and thus had better access to lifeboats.
 - **Third-class** passengers were the largest group onboard and had the highest mortality rate.
-

4. Age Distribution

- **Younger passengers**, especially children, were more likely to survive.
 - Among older passengers, those in **first class** had notably higher survival rates than those in other classes.
-

5. Fare Analysis

- A **higher fare** was strongly associated with higher chances of survival.
 - This aligns with wealthier passengers being placed in more favorable positions on the ship.
-

6. Embarkation Port

- Most passengers boarded from **Southampton**, followed by **Cherbourg** and **Queenstown**.
- Passengers from **Cherbourg** had the highest survival rates, likely due to a higher proportion of first-class passengers.

7. Feature Engineering

- Categorical variables such as **Sex** and **Embarked** were converted into numerical dummy variables using one-hot encoding.
- This prepared the data for further machine learning modeling.