Exploratory Data Analysis (EDA) Report — Titanic Dataset

1. Introduction

The goal of this analysis is to explore the Titanic passenger dataset to extract insights related to survival patterns. The dataset contains features including age, sex, passenger class, fare, number of siblings/spouses aboard (sibsp), number of parents/children aboard (parch), embarkation port, and deck.

---

2. Data Overview

- **Number of rows and columns:** See dataset summary.
- **Data types:** Mix of categorical and numerical variables.
- **Missing values:** Notable missingness in age, deck, and embark_town.
- **Unique values:**
    - Sex: male, female
    - Class: First, Second, Third
    - Embarked: Cherbourg, Queenstown, Southampton

---

3. Univariate Analysis

Numerical Features:

- **Age:** Right-skewed distribution; median age around 28; some missing values.
- **Fare:** Skewed distribution with extreme outliers representing higher-paying passengers.
- **SibSp & Parch:** Most passengers had 0–1 family members aboard.

Categorical Features:

- **Sex:** Majority male passengers.
- **Class:** Highest counts in 3rd and 2nd class.
- **Embarked:** Most passengers boarded at Southampton.

Observation: Socioeconomic patterns are evident from class and fare distributions.

Figures:

- Histogram of Age
- Histogram of Fare
- Countplots for Sex, Class, Embarked

4. Bivariate Analysis

Survival vs Categorical Features:

- **Sex:** Female passengers had much higher survival rates (~74%) than males (~19%).
- **Class:** First-class passengers survived more often than third-class passengers.
- **Embarked:** Slight variations in survival rate based on port.

Survival vs Numerical Features:

- **Age:** Children and young adults had higher survival probabilities.
- **Fare:** Survivors tended to pay higher fares, reflecting correlation with class.

Figures:

- Barplots of survival rate by sex, class, embarked
- Boxplots of age and fare by survival

---

5. Correlation & Multivariate Analysis

- **Correlation heatmap:** Shows moderate correlation between fare and passenger class (pclass).
- **Pairplot (subset):** Shows relationships among age, fare, sibsp, parch, and survival.

Figures:

- Correlation heatmap
- Pairplot of numerical features colored by survival

---

6. Feature Engineering & Missing Values

- **Imputation:**
    - Age → median
    - Deck → 'Unknown'
    - Embark_town → mode
- **Derived features:**
    - family_size = sibsp + parch
    - is_alone = (family_size == 0)

Purpose: Improve model performance and handle missing or skewed data.

7. Key Findings

- **Sex & class are strongest survival predictors.**
- **Children (0–12) survived better than adults in some groups.**
- **Higher fare correlates weakly with survival.**
- Missing data in age, deck, and embark_town must be handled before modeling.
- Outliers in fare should be considered for transformation or capping.

---

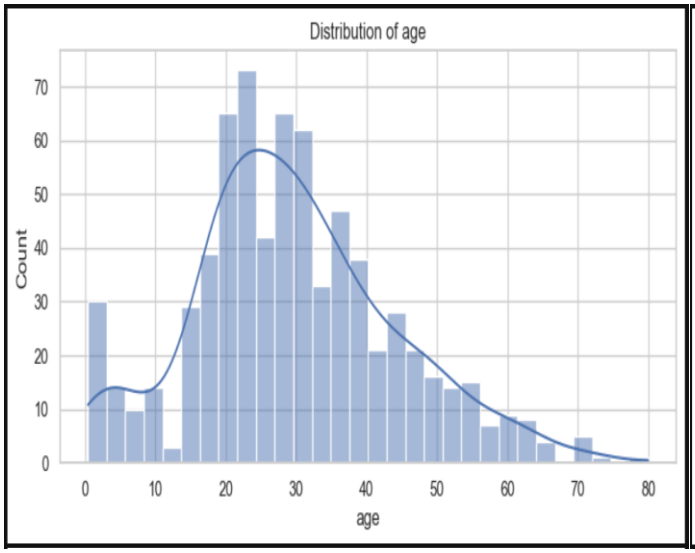
8. Recommendations

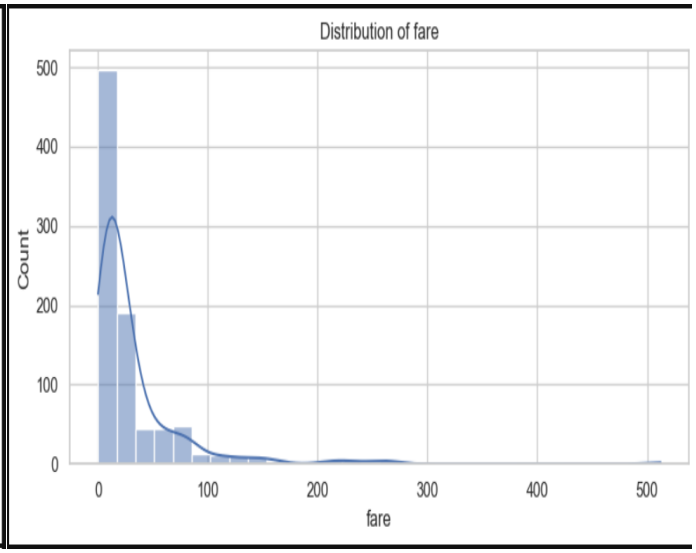
1. **Preprocessing:** Impute missing values, encode categorical variables, handle outliers.
2. **Feature engineering:** Include family_size, is_alone, and interaction terms (sex*class).
3. **Modeling:** Logistic Regression, Random Forest, or XGBoost after proper preprocessing.
4. **Visualization:** Include plots for presentations; visual summaries aid understanding.

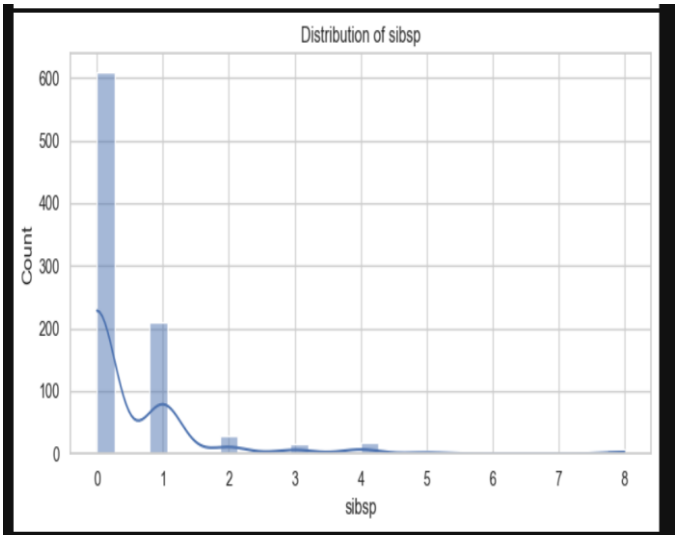
---

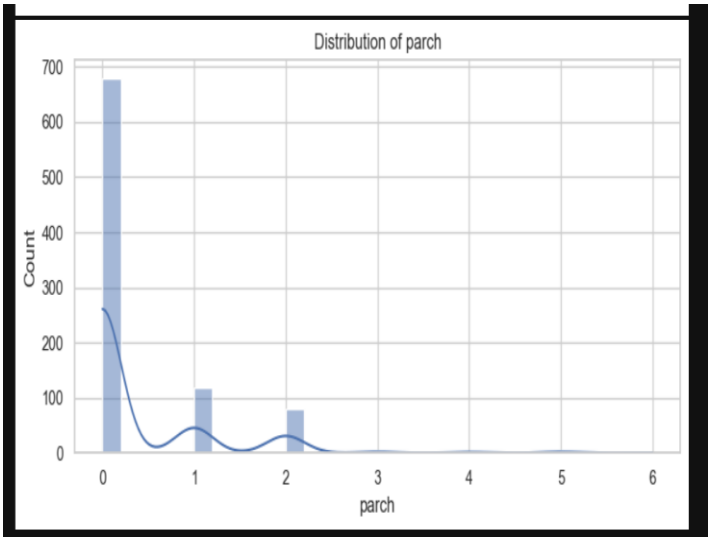
9. Appendix — Figures

- ## Histograms: Age, Fare

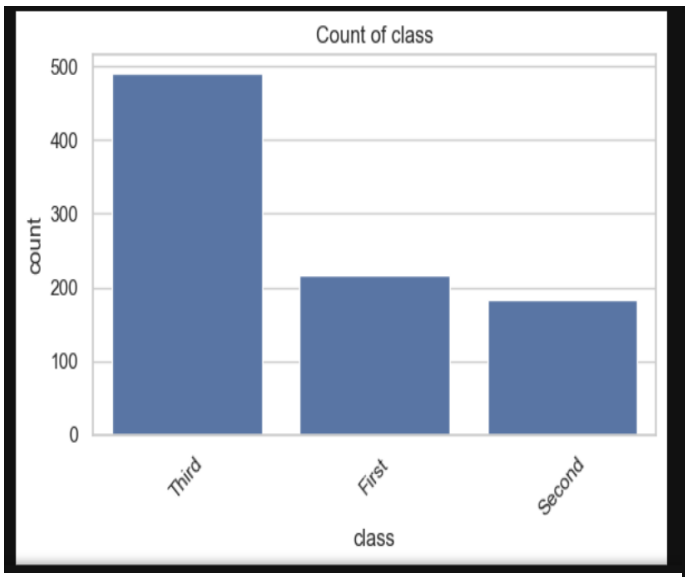


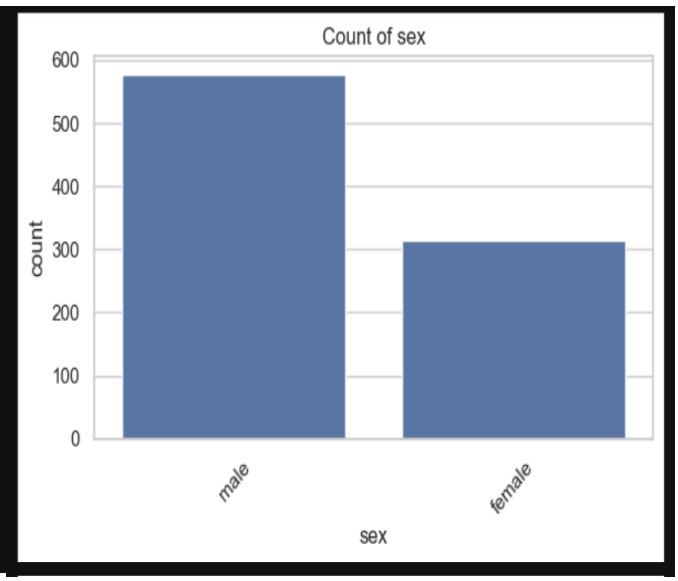


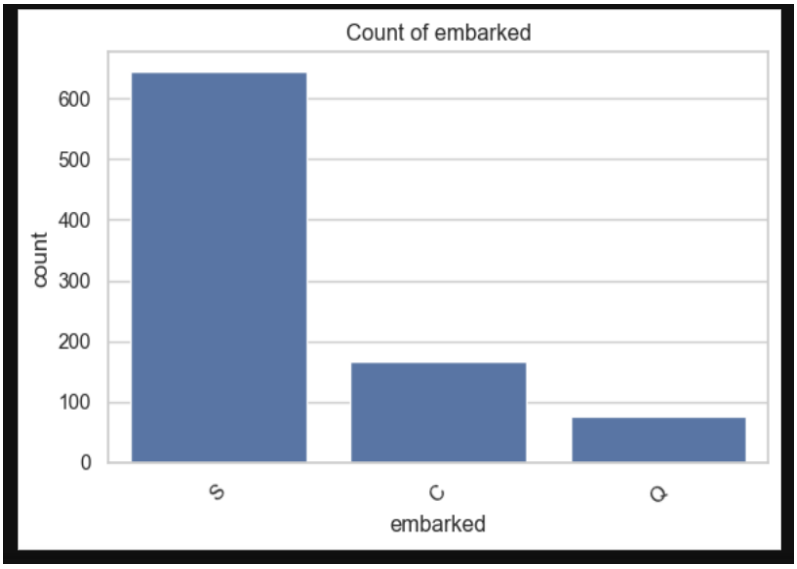




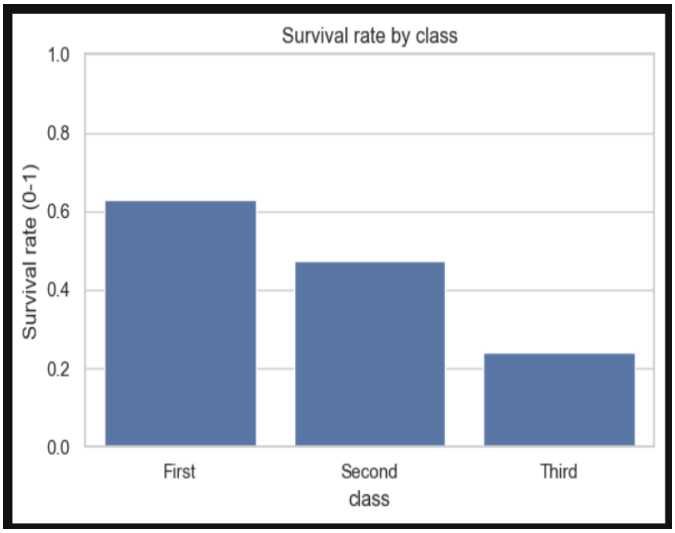
- **Countplots: Sex, Class, Embarked**

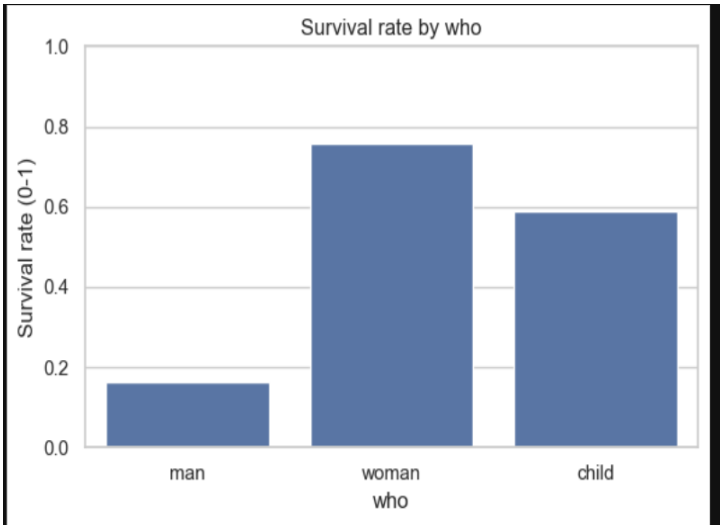


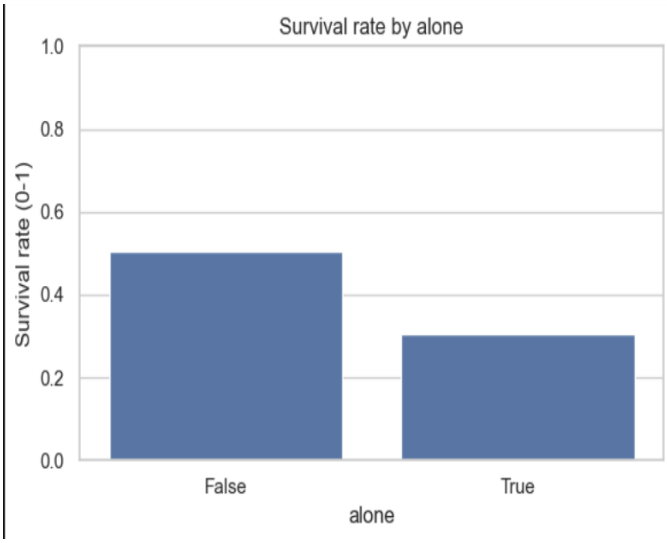


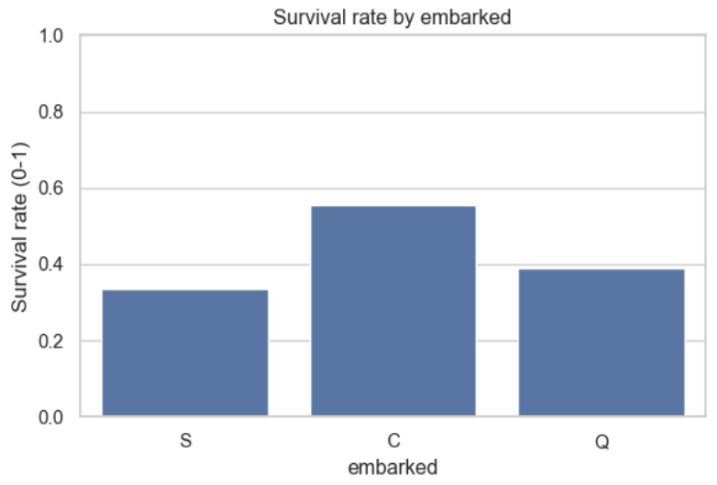


- **Survival rate barplots**

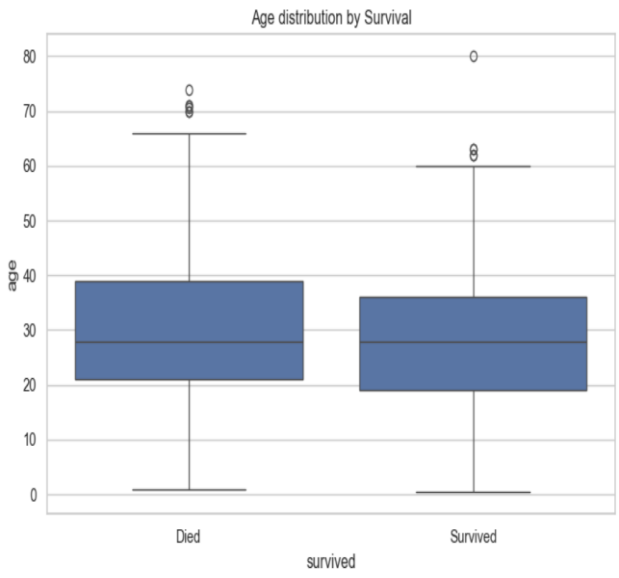


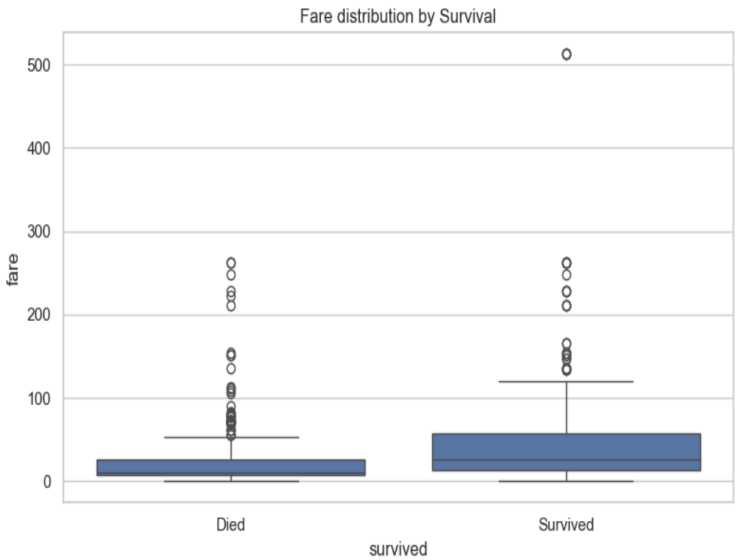




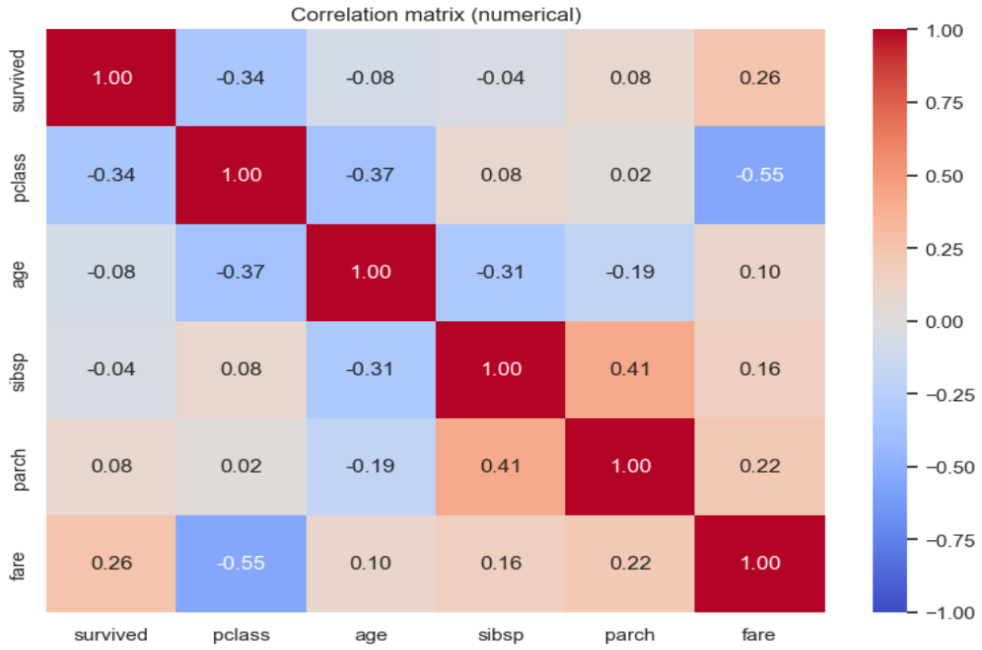


- **Boxplots: Age vs Survival, Fare vs Survival**





- **Correlation heatmap**



- **Pairplot**