
Titanic Dataset Exploratory Data Analysis (EDA) — Summary Report

---

1. Dataset Overview

The dataset contains 891 passengers with 12 columns including features like Age, Sex, Fare, Pclass, and the target variable Survived.

Missing data is present mainly in Age (177 missing), Cabin (687 missing), and a few in Embarked (2 missing).

---

2. Survival Overview

Overall survival rate is approximately 38% (342 survived, 549 did not survive).

Females had a much higher survival rate (~74%) compared to males (~19%).

Survival rates decrease as passenger class decreases:

- Pclass 1: ~63% survival
- Pclass 2: ~47% survival
- Pclass 3: ~24% survival

---

3. Key Relationships & Trends

Sex and Pclass are strong predictors of survival; females and higher class passengers had better chances.

Embarkation port also influences survival rates; passengers boarding from C had slightly higher survival.

Age distribution shows survivors tended to be slightly younger on average, with a median age of about 28 years for survivors and similar median age for non-survivors, though distributions differ.

Fare is positively associated with survival: survivors paid higher median fares (~26) compared to non-survivors (~10.5), indicating wealth and class impact.

---

4. Visualizations & Insights

Correlation heatmaps confirm that Pclass (negatively) and Fare (positively) correlate with survival.

Pairplots reveal clusters where higher fare and lower class passengers are associated with survival.

Violin plots of age show wider spread among survivors, with some children having higher survival rates.

Histograms highlight a right-skewed fare distribution, recommending log transformation for modeling.

---

5. Missing Data & Data Quality

Large amount of missing data in Cabin suggests it may be less reliable or require special imputation techniques.

Age missingness is significant but manageable via median imputation or prediction models.

---

6. Next Steps

- Feature engineering: create features like FamilySize, IsAlone, Title extracted from Name, and CabinDeck from Cabin.
- Impute missing Age values using group medians based on Title and Pclass.
- Log-transform Fare to reduce skewness.
- Encode categorical features (Sex, Embarked, Title, CabinDeck).
- Build predictive models (e.g., Logistic Regression, Random Forest) to classify survival using these engineered features.
- Perform model validation and tuning.