Titanic EDA — Report

Author: Krishna Shah

---

1. Objective

Perform exploratory data analysis (EDA) on the Titanic dataset to check data quality, explore variable distributions, relationships with survival, and derive key insights.

2. Dataset

- **Rows:** 1300
- **Columns:** 10
- **Target Variable:** survived (0 = did not survive, 1 = survived)
- Features include: Passenger class (pclass), Name, Gender, Age, Family, Fare, Embarked, etc.

3. Data Quality Check

- Missing values:
  - age: **258 missing**
  - fare: **3 missing**
  - gender: **1 missing**
  - family: **2 missing**
  - embarked: **5 missing**
- Data types: mixture of int, float, and object (categorical).
- Duplicates: some duplicate names were observed (possible repeat records).

4. Univariate Analysis

- **Survival:** Only **38.5% survived**, showing class imbalance.
- **Gender:** Slightly more males than females.
- **Passenger Class:** Majority were in **3rd class**, fewer in 1st and 2nd.
- **Age:** Ranges from 0.17 to 80 years, mean ~30 years. Distribution fairly normal, slight right skew.
- **Fare:** Highly skewed (**skewness = 4.3**), with most fares < 50 but some extreme outliers up to 512.

5. Bivariate Analysis

- **Gender vs Survival:** Females had **significantly higher survival rates** than males.
- **Class vs Survival:** 1st class passengers survived more than 2nd and 3rd class.
- **Age vs Survival:** Younger passengers had slightly higher survival rates.
- **Fare vs Survival:** Passengers who paid higher fares had a greater chance of survival.

6. Multivariate / Feature Engineering

- pclass negatively correlated with survival (lower class → lower survival).
- fare positively correlated with survival.
- age had weak correlation with survival.
- family showed some non-linear effects (very large families had lower survival).

7. Outliers & Skew

- **Age:** Slight skew (0.41), fairly normally distributed.
- **Fare:** Strong right skew (4.35) → transformation (e.g., log scale) recommended.
- Outliers present in fare (very high ticket prices).

8. Key Insights (Bullet Summary)

- Females were more likely to survive than males.
- Survival was strongly linked to passenger class → 1st > 2nd > 3rd.
- Higher fare = better survival chances.
- Children and younger passengers had slightly better outcomes.
- Data quality issues: missing values in age, fare, embarked.

9. Limitations

- Missing data (especially age).
- Some duplicate records.
- Survival distribution is imbalanced (only 38.5% survived).

10. Conclusion

The Titanic dataset shows clear survival patterns:

- **Gender and class are the strongest predictors.**
- **Fare also has predictive power** (wealthier passengers more likely to survive).
- **Age plays a secondary role.**

These findings form a strong foundation for feature engineering and predictive modelling.