
📊 Titanic Dataset – EDA Findings Report

This report summarizes key insights and observations from the Exploratory Data Analysis (EDA) performed on the Titanic dataset.

---

1. Dataset Overview

- The dataset contains **891** records and **12 columns**.
- Features include passenger demographics, ticket info, and survival outcome.
- Target variable: Survived (0 = No, 1 = Yes)

---

2. Missing Data

- **Age**: 177 missing values (~20%)
- **Cabin**: 687 missing values (~77%)
- **Embarked**: 2 missing values
- The **Cabin** column has extensive missing data and may not be reliable for modeling without heavy processing.

---

3. Univariate Analysis

- **Survival Distribution**:
  - Fewer passengers survived than perished.
- **Fare**:
  - Highly skewed distribution with significant outliers.
  - Most fares are clustered at the lower end.

---

4. Bivariate Analysis

- **Gender vs Survival**:
  - **Females** had a much higher survival rate compared to **males**.
- **Age vs Fare**:
  - No strong linear correlation between age and fare.
  - Younger and older passengers are scattered across all fare ranges.

---

5. Multivariate Analysis

- **Pairplot** of Survived, Age, SibSp, and Parch:
    - Survivors tended to have fewer siblings/spouses and parents/children aboard.
- **Correlation Heatmap**:
    - Fare shows moderate positive correlation with Survived (wealthier passengers had higher survival chances).
    - Family-related features (SibSp, Parch) also show weak positive correlation with survival.

---

6. Missing Value Treatment

- **Age**: Filled missing values using the **median**, a robust method that handles outliers effectively.
- **Embarked**: Can be filled using the mode or dropped depending on modeling needs.
- **Cabin**: Considered for removal or feature engineering due to high missing rate.

---

7. Skewness

- **Age** skewness: ~0.51 → Slight positive skew (right-tailed), but still close to normal distribution.

✅ **Summary**

- Gender and Fare are strong indicators of survival.
- Dataset is now mostly clean after handling missing values.
- Data is ready for feature engineering and predictive modeling.
