# EDA Titanic cleaning data process report

**1. Title & Objective**

**Exploratory Data Analysis of Titanic Dataset**
**Objective:** The goal is to explore the dataset to identify survival trends and relationships between passenger attributes and survival probability.

---

**2. Dataset Overview**

- **Source:** train.csv (Titanic dataset)
- **Number of rows:** 891
- **Number of columns:** 12
- **Key Features:**
    - **PassengerId**: Unique passenger identifier
    - **Survived**: Survival status (0 = No, 1 = Yes)
    - **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
    - **Name**: Passenger's name
    - **Sex**: Gender
    - **Age**: Age in years
    - **SibSp**: Number of siblings/spouses aboard
    - **Parch**: Number of parents/children aboard
    - **Ticket**: Ticket number
    - **Fare**: Passenger fare
    - **Cabin**: Cabin number
    - **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

---

**3. Data Cleaning Summary**

- Missing Values:
    - Age: 177 missing values
    - Cabin: 687 missing values
    - Embarked: 2 missing values

- Fixed data types for analysis (Age → float, Fare → float)
- No duplicate rows found

---

**4. Statistical Summary**

From .describe() and .value_counts() analysis:

- **Age**: Mean = 29.7, Min = 0.42, Max = 80
- **Fare**: Mean = 32.2, Min = 0, Max = 512.3
- **Survival rate**: 38.38% survived (342/891)
- **Gender count**: 577 males, 314 females
- **Class distribution**: 1st (216), 2nd (184), 3rd (491)

---

**5. Visual Insights**

- **Bar Chart**: Higher survival rate for females than males
- **Histogram (Age)**: Most passengers between 20–40 years old
- **Boxplot (Fare vs Class)**: Higher ticket class → Higher fare
- **Heatmap**: Survival is positively correlated with Fare and Pclass (inverse correlation)
- **Pairplot**: Clear separation between classes and survival probability

---

**6. Key Findings**

- Females had a much higher survival rate than males
- 1st class passengers had the highest survival probability
- Younger passengers, especially children, survived more often
- Higher fares were generally linked to better survival rates
- Port of embarkation "C" passengers had slightly higher survival rates than others

---

**7. Conclusion & Next Steps**

- Gender, class, and age are strong predictors of survival
- Missing age and cabin data could be imputed for better modeling
- Next step: Build a predictive model (e.g., logistic regression, decision tree) using these insights