

1. Introduction

The purpose of this Exploratory Data Analysis (EDA) is to extract meaningful insights from the Titanic passengers dataset. Using Python libraries such as Pandas, Matplotlib, and Seaborn, we explore statistical summaries, detect missing values, understand feature distributions, and discover hidden patterns and correlations.

2. Dataset Overview

- **Rows:** 891
- **Columns:** 12

Key Features:

- PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

3. Initial Observations

- The dataset contains both numerical and categorical features.
- The target variable is '**Survived**' (1 = Survived, 0 = Not Survived).

4. Statistical Insights

- The average age of passengers is around **29.7 years**.
- **38.4%** passengers survived.
- Majority passengers belonged to **3rd class (Pclass = 3)**.
- Median fare paid is approximately **14.45 units**.

5. Missing Values Analysis

- **Age:** 177 missing values.
- **Cabin:** 687 missing values (majority missing).
- **Embarked:** 2 missing values.

Observation: Cabin feature has too many missing values, making it less reliable without imputation.

6. Correlation Analysis

- Strong positive correlation between **SibSp** (siblings/spouses aboard) and **Parch** (parents/children aboard).
- Negative correlation between **Pclass** and **Fare** (lower class, lower fare).
- Heatmap and pairplot show some mild relationships but no multicollinearity issues.

7. Visual Insights

- **Survival by Gender:**
 - Females had a significantly higher survival rate compared to males.
- **Age Distribution:**
 - Most passengers are between 20-40 years.
 - Children under 10 also had notable survival rates.
- **Fare Distribution:**
 - Highly right-skewed; a few passengers paid very high fares.
- **Survival by Class:**
 - Higher-class passengers (1st class) had better chances of survival.
- **Embarked Port:**
 - Majority passengers embarked from port 'S' (Southampton).

8. Summary of Findings

- **Gender** is a strong predictor of survival; females had better survival rates.
- **Passenger class** is an important factor; 1st class had higher survival chances.
- **Younger passengers** had slightly better survival probabilities.
- **Fare** can be a good indicator, but it is skewed and needs transformation if used for modeling.
- Significant missing data in "Cabin" requires either removal or proper imputation.

9. Conclusion

This EDA provides a detailed understanding of the Titanic dataset's structure and key trends. The survival chances vary significantly across gender, passenger class, and age groups. These insights are crucial for further predictive modeling.

End of Report