Report on
Exploratory Data Analysis on Titanic Dataset

(Data Analyst Internship – Task 5)

Date: June 30th,2025
Submitted by: Akshitha Reddy Jajapuram
Dataset Source: Kaggle Titanic Dataset
Tools Used: Python, Pandas, Seaborn, Matplotlib, Jupyter Notebook

1. Introduction

The Titanic dataset is one of the most iconic datasets in the data science community. It contains information about passengers who were aboard the Titanic, which tragically sank after hitting an iceberg. The goal of this exploratory data analysis (EDA) is to uncover patterns and relationships within the dataset that could provide insight into survival factors.

2. Dataset Description

The dataset used contains 891 records of passengers with the following primary features:

- **PassengerId**: Unique identifier for each passenger
- **Survived**: Survival status (0 = No, 1 = Yes)
- **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Name**: Name of the passenger
- **Sex**: Gender
- **Age**: Age of the passenger
- **SibSp**: Number of siblings/spouses aboard
- **Parch**: Number of parents/children aboard
- **Ticket**: Ticket number

- **Fare**: Ticket fare
- **Cabin**: Cabin number
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3. Data Cleaning and Preprocessing

- Missing values were found in Age, Cabin, and Embarked.
    - Age: Filled using median value.
    - Cabin: A large number of missing values, dropped for correlation analysis.
    - Embarked: Filled with the mode.
- Converted categorical variables to numerical using encoding (e.g., Sex, Embarked) for deeper analysis.
- Dropped columns like Name, Ticket, and Cabin in numerical correlation analysis as they are not directly useful.

4. Univariate Analysis

4.1 Age

- The distribution of age is slightly right-skewed.
- Most passengers were between 20 to 40 years old.

4.2 Fare

- Highly skewed distribution with a few passengers paying extremely high fares.
- Most fares range below $100.

4.3 Survived

- About 38% of passengers survived.

4.4 Sex

- Majority of passengers were male (~65%).

4.5 Pclass

- Class 3 had the highest number of passengers, followed by Class 1 and Class 2.

5. Bivariate Analysis

5.1 Survived vs Sex

- Survival rate among females was significantly higher (~75%) compared to males (~20%).
- Gender played a crucial role in survival.

5.2 Survived vs Pclass

- First-class passengers had the highest survival rate.
- Third-class passengers had the lowest survival rate, indicating class was an important factor.

5.3 Survived vs Age

- Children (age < 10) had higher survival rates.
- Elderly passengers had lower survival probabilities.

5.4 Survived vs Fare

- Passengers who paid higher fares had higher survival rates.
- Indicates possible influence of social status and access to lifeboats.

6. Multivariate Analysis

6.1 Heatmap (Correlation Matrix)

- Fare and Pclass showed moderate correlation with Survived.
- SibSp and Parch showed weak correlation.
- Sex after encoding showed a strong positive correlation with Survived.

6.2 Pairplot

- Clear separation of survivors and non-survivors visible across features like Age, Fare, and Pclass.
- Helped visualize combined feature interactions.

7. Observations & Insights

- **Gender is the most influential feature**: Female passengers had significantly higher survival rates.
- **Class matters**: Passengers in 1st class had better survival odds than those in 2nd and 3rd classes.
- **Fare reflects status and access**: Those who paid more had a higher chance of survival.
- **Children were prioritized**: Higher survival among kids supports the "women and children first" protocol.

- **Embarked point had minimal impact**: No significant difference found in survival rate based on port of embarkation.

8. Summary

This EDA reveals that survival on the Titanic was not random. Gender, class, and fare were critical factors. The insights derived are not just useful for machine learning modeling but also highlight the historical and social contexts that influenced survival chances. Such analysis is essential in any real-world data science project to understand data distributions, relationships, and potential biases.

9. Recommendations for Further Analysis

- Impute missing values in Cabin and investigate its effect on survival.
- Use advanced visualizations like violin plots and swarm plots.
- Conduct feature engineering for family size, title extraction from names, etc.
- Build predictive models (Logistic Regression, Random Forest) based on these insights.

10. Tools Used

- **Pandas**: Data manipulation and statistics
- **Seaborn & Matplotlib**: Visualization
- **Jupyter Notebook**: Interactive analysis environment