Titanic Dataset – Exploratory Data Analysis Report

Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous maritime disasters in history. Over 1,500 people lost their lives, while just over 700 survived. Beyond the human tragedy, the event offers a unique historical dataset, providing valuable insights into the socio-economic and demographic patterns that influenced survival chances during the disaster.

This report performs an Exploratory Data Analysis (EDA) of the Titanic dataset, aiming to:

1. Understand the composition of passengers in terms of demographics, socio-economic status, and travel conditions.
2. Identify statistical relationships between features and survival status.
3. Use visual analysis to uncover patterns not easily detectable from raw numbers alone.
4. Provide well-documented, reproducible findings with professional visual presentation.

The results are not intended to predict survival but to **understand historical patterns** and offer insights that reflect the inequalities of the time.

Dataset Overview

The Titanic dataset, provided by Kaggle, contains passenger details across 891 records with 12 key variables.

Feature Descriptions:

- **Survived:** Binary indicator of survival (0 = No, 1 = Yes)
- **Pclass:** Ticket class (1st, 2nd, or 3rd), representing socio-economic standing
- **Name:** Passenger's name (not used in numeric analysis)
- **Sex:** Gender of the passenger
- **Age:** Passenger's age in years
- **SibSp:** Number of siblings/spouses aboard
- **Parch:** Number of parents/children aboard
- **Ticket:** Ticket number (categorical)

Titanic Dataset – Exploratory Data Analysis Report

- **Fare:** Ticket fare paid
- **Cabin:** Cabin identifier (many missing values)
- **Embarked:** Port of embarkation — C (Cherbourg), Q (Queenstown), or S (Southampton)

Data Understanding

Initial Observations:

- The dataset has mixed data types — numerical (Age, Fare), categorical (Pclass, Sex), and textual (Name, Ticket).
- The Survived variable is the main target for understanding survival patterns.

Missing Data:

- Age has approximately **19.8% missing values**.
- Cabin is missing in over **77% of cases** — too incomplete for robust analysis.
- Embarked has two missing values.

Missing values in Age were handled through **median imputation** to preserve statistical balance without skewing distributions.

Data Visualization & Observations

Pair Plot

A Seaborn pairplot revealed clustering patterns:

- High fares are concentrated in first-class passengers.
- Females cluster more heavily in the survival group, especially in higher classes.

Correlation Heatmap

The correlation matrix shows:

- **Negative correlation** between Pclass and Fare (-0.55), meaning wealthier passengers tended to have higher-class tickets.
- **Positive correlation** between Fare and survival (0.26).

Titanic Dataset – Exploratory Data Analysis Report

- Very weak correlation between Age and survival (-0.08).

Age Distribution

- Most passengers were **20–40 years old**.
- Children under 10 show higher survival percentages.

Fare Analysis

- First-class passengers paid fares significantly higher than second and third class.
- Some extreme outliers in fare (above 500 units) indicate luxury travel suites.

Class and Survival Relationship

- First class: **63% survival rate**
- Second class: **47% survival rate**
- Third class: **24% survival rate**

Key Findings

1. **Gender Disparity in Survival**
    - Female passengers had a survival rate over **70%**, compared to **19%** for males.
    - Reflects the "women and children first" evacuation policy.
2. **Impact of Passenger Class**
    - Socio-economic status was a major determinant. First-class passengers had priority access to lifeboats.
3. **Children Had Higher Chances**
    - Particularly in first and second class.
4. **Fare and Survival Link**
    - Passengers who paid higher fares tended to survive, possibly due to location of cabins and faster evacuation.
5. **Port of Embarkation Effect**

- Passengers from Cherbourg had the highest survival rate (~55%), possibly reflecting wealthier demographics boarding there.

Ethical Considerations

- **Historical Context:**
  This dataset reflects early 20th-century societal norms, which included class and gender inequalities. The findings should be interpreted within that historical framework.
- **Privacy & Sensitivity:**
  Data contains no modern personally identifiable information.
- **Bias Awareness:**
  Results should not be generalized beyond the Titanic scenario — correlations observed here may not apply to modern-day maritime safety or disaster survival.

Conclusion

The Titanic disaster was shaped not just by the iceberg collision, but by human social structures — wealth, gender, and class played decisive roles in determining who lived and who died. This analysis reinforces the importance of considering socio-economic factors when designing emergency protocols and safety measures.

The dataset continues to serve as an important educational tool for statistical analysis, machine learning, and ethical discussions about data interpretation.

References

- Kaggle Titanic Dataset: https://www.kaggle.com/datasets/yasserh/titanic-dataset
- GitHub Repo Link: https://github.com/sivasakthi-15/Data_analyst_internship/tree/main/Task%205