Titanic Survival Analysis: Exploratory Data Analysis Report

Title: Titanic Survival Analysis

Name: Daniel Muthama

Program: Data Science Certification

Date: 30 May 2025

GitHub: Link

Notebook Link: **Notebook**

[ALL FIGURES, DATASET AND IMAGES ATTACHED IN THE GITHUB]

1. Introduction

This report presents a comprehensive exploratory data analysis (EDA) of the Titanic dataset, which contains information about 1,309 passengers aboard the RMS Titanic. The primary objective is to identify factors that influenced survival rates during this historic maritime disaster. The analysis follows a structured approach including data cleaning, univariate analysis, bivariate analysis, multivariate analysis, and target variable examination.

Key Questions Explored:

What was the overall survival rate?

How did gender and passenger class affect survival?

What was the relationship between age, fare, and survival?

How did embarkation port influence survival rates?

What interactions existed between class, gender, and survival?

2. Task Completion

2.1 Data Cleaning and Preparation

Renamed columns for clarity (e.g., '2urvived' → 'Survived')

Mapped numerical codes to categorical labels:

Gender: 0 → Male, 1 → Female

EmbarkPort: 0 → Cherbourg, 1 → Queenstown, 2 → Southampton

Removed 18 redundant zero-value columns

Handled missing values in EmbarkPort (2 missing values filled with mode)

Capped fare outliers using IQR method (Upper limit: $112.75)

Cleaned Dataset Features:

Passengerid, Survived, Age, Fare, Gender, SiblingsSpouses,
ParentsChildren, PassengerClass, EmbarkPort
Evidence of Cleaning:
Missing Values Heatmap
Figure 1: Initial missing values visualization

Outlier Treatment
Figure 2: Fare distribution before and after outlier treatment

2.2 Univariate Analysis
Categorical Distributions:
Categorical Distributions
Figure 3: Distributions of key categorical features

Key Findings:

Only 26.1% of passengers survived

62.5% were 3rd class passengers

Male passengers outnumbered females 2:1 (65.4% vs 34.6%)

Southampton was the most common embarkation port (72.5%)

Numerical Distributions:
Numerical Distributions
Figure 4: Distributions of age and fare

Key Findings:

Age distribution is right-skewed (median: 28 years)

Fare distribution is highly right-skewed (median: $14.45)

Most passengers were young adults (20-40 years)

Most fares were under $50, with a few high-value outliers

2.3 Bivariate Analysis
Survival vs Categorical Features:
Survival by Category
Figure 5: Survival rates by passenger class, gender, and embarkation port

Key Findings:

Gender: 74.3% of females survived vs 18.9% of males

Passenger Class:

1st class: 63.0% survival

2nd class: 47.3% survival

3rd class: 24.2% survival

Embarkation Port:

Cherbourg: 55.4% survival

Queenstown: 39.0% survival

Southampton: 33.7% survival

Survival vs Numerical Features:
Survival by Numerical
Figure 6: Age and fare distributions by survival status

Key Findings:

Survivors tended to be younger (median age 28 vs 30)

Survivors paid higher fares (median $26 vs $11)

Children (0-10 years) had highest survival rates

Higher fare classes correlated with better survival

2.4 Multivariate Analysis
Class-Age-Survival Interaction:

Class-Age-Survival
Figure 7: Survival distribution by class and age

Key Findings:

1st class children had near-perfect survival rates

3rd class passengers had lowest survival across all age groups

Survival advantage for young adults in 1st/2nd class

Embarkation Port-Class-Survival:
Embarkation-Class-Survival
Figure 8: Survival rates by embarkation port and class

Key Findings:

Cherbourg had highest proportion of 1st class passengers

Queenstown passengers were predominantly 3rd class

Southampton had most 3rd class passengers and lowest survival

Feature Correlation:
Correlation Matrix
Figure 9: Correlation between key features

Key Findings:

Strongest correlations:

PassengerClass ↔ Fare (-0.55)

PassengerClass ↔ Survived (-0.34)

Fare ↔ Survived (0.26)

Age shows weak correlation with survival (-0.07)

Family size features show minimal correlation

2.5 Target Variable (Survived) Analysis

Survival Distribution:

Survival Distribution

Figure 10: Overall survival distribution (0 = Died, 1 = Survived)

Overall Survival Rate: 26.1%

Key Survival Factors:

Key Survival Factors

Figure 11: Survival rates by gender and passenger class

Gender Impact:

Female survival rate: 74.3%

Male survival rate: 18.9%

Class Impact:

1st class survival: 63.0%

2nd class survival: 47.3%

3rd class survival: 24.2%

Class-Gender Interaction:

Class-Gender Interaction

Figure 12: Survival rate interaction between class and gender

Key Finding:

1st class females had 96.8% survival rate

3rd class males had only 13.5% survival rate

The "women and children first" protocol was clearly followed, especially in higher classes

3. Conclusion

3.1 Key Findings Summary

Overall Survival: Only 26.1% of passengers survived, highlighting the severity of the disaster

Gender Disparity: Females had 3.9× higher survival rate than males (74.3% vs 18.9%)

Class Privilege: 1st class passengers had 2.6× higher survival rate than 3rd class passengers

Age Advantage: Children (0-10 years) had the highest survival rate (59.0%)

Embarkation Effect: Cherbourg passengers had highest survival (55.4%), likely due to higher proportion of 1st class passengers

Fare Correlation: Higher fares correlated with better survival, though this is confounded by passenger class

Interaction Effect: 1st class females had near-perfect survival (96.8%) while 3rd class males had the lowest (13.5%)

3.2 Implications

The analysis reveals significant social stratification in survival outcomes:

The "women and children first" protocol was followed, but with class-based discrimination

1st class passengers had priority access to lifeboats regardless of gender

3rd class passengers faced structural barriers to survival regardless of age or gender

3.3 Recommendations for Further Analysis

Investigate family group survival patterns

Analyze ticket information for grouping effects

Build predictive models to quantify feature importance

Compare survival rates by nationality

Study crew member survival patterns separately

Appendix: Technical Implementation

Python Libraries Used: Pandas, NumPy, Seaborn, Matplotlib

Data Cleaning Approach:

Renamed columns and mapped categorical variables

Removed redundant features

Handled missing values and outliers

Visualization Strategy:

Used count plots for categorical distributions

Employed histograms and box plots for numerical features

Created violin plots for multivariate relationships

Utilized heatmaps for correlation and survival rates

Ethical Consideration: This analysis respects the memory of Titanic victims and presents findings objectively without sensationalism. The dataset has been treated with appropriate historical sensitivity.

Submitted By: Daniel Muthama
Date: 30 May 2025
Kaggle Dataset: [Public Kaggle Notebook]