Titanic Dataset - Exploratory Data Analysis Report

Executive Summary

This comprehensive exploratory data analysis (EDA) of the Titanic dataset reveals critical insights into passenger survival patterns during the historic maritime disaster. The analysis examined 891 passengers across 12 variables, uncovering significant relationships between demographic factors, socioeconomic status, and survival outcomes.

Key Findings:

- Overall survival rate: 38.4% (342 survivors out of 891 passengers)
- Gender was the strongest predictor: Women had 74.2% survival rate vs. 18.9% for men
- Passenger class showed clear survival hierarchy: 1st class 63.0%, 2nd class 47.3%, 3rd class 24.2%
- Family composition mattered: Passengers with family had 50.6% survival rate vs. 30.4% for solo travelers

Dataset Overview

The Titanic dataset contains information about 891 passengers with the following characteristics:

Dataset Structure:

- Training set: 891 passengers × 12 features
- Test set: 418 passengers × 11 features (no survival labels)
- Submission template: 418 passenger IDs with survival predictions

Feature Types:

- Numerical: Age, Fare, SibSp (siblings/spouses), Parch (parents/children)
- Categorical: Sex, Pclass (passenger class), Embarked (port), Cabin
- Text: Name, Ticket
- Target: Survived (0 = No, 1 = Yes)

Missing Data Analysis

Missing data presented significant challenges for analysis:

Variable	Missing Count	Percentage
Cabin	687	77.1%
Age	177	19.9%

Variable	Missing Count	Percentage
Embarked	2	0.2%

The high percentage of missing cabin data (77.1%) severely limited deck-based survival analysis, while missing age data (19.9%) required careful handling in age-related investigations.

Survival Analysis by Demographics

Gender Analysis

The most striking survival pattern emerged from gender differences:

- **Female passengers**: 233 survivors out of 314 total (74.2% survival rate)
- **Male passengers**: 109 survivors out of 577 total (18.9% survival rate)

This dramatic difference reflects the "women and children first" maritime evacuation protocol.

Age Group Analysis

Age-based survival patterns revealed interesting insights:

Age Group	Total	Survivors	Survival Rate
Children (0-12)	69	40	57.9%
Teens (13-18)	70	30	42.9%
Adults (19-35)	358	137	38.3%
Middle Age (36-60)	195	78	40.0%
Seniors (60+)	22	5	22.7%

Children had notably higher survival rates than other age groups, while seniors faced the lowest survival rates.

Socioeconomic Factors

Passenger Class Impact

Passenger class demonstrated a clear survival hierarchy:

Class	Total	Survivors	Survival Rate
1st Class	216	136	63.0%
2nd Class	184	87	47.3%
3rd Class	491	119	24.2%

First-class passengers had nearly three times the survival rate of third-class passengers, indicating significant socioeconomic bias in rescue operations.

Fare Analysis

Ticket fare served as a continuous measure of socioeconomic status:

Fare Group	Range	Survival Rate
Low	£0-7.9	19.7%
Medium-Low	£7.9-14.45	30.9%
Medium-High	£14.45-31	44.5%
High	£31+	58.1%

Higher fares correlated positively with survival (correlation: +0.257), reflecting the class-based survival patterns.

Embarkation Port Analysis

Port of embarkation showed modest survival differences:

- **Cherbourg (C)**: 55.4% survival rate
- **Queenstown (Q)**: 39.0% survival rate
- **Southampton (S)**: 33.7% survival rate

Family Structure Analysis

Family Size Impact

Family composition significantly influenced survival outcomes:

Family Size	Count	Survival Rate
1 (Alone)	537	30.4%
2	161	55.3%
3	102	57.8%
4	29	72.4%
5+	62	20.0%

Medium-sized families (2-4 members) had the highest survival rates, while solo travelers and very large families faced greater risks.

Title Analysis

Passenger titles extracted from names revealed social status patterns:

Title	Count	Survival Rate
Mrs	126	79.4%

Title	Count	Survival Rate
Miss	185	70.3%
Master	40	57.5%
Mr	517	15.7%
Rare	23	34.8%

Married women (Mrs) had the highest survival rate, followed by unmarried women (Miss) and young boys (Master).

Statistical Correlations

The correlation analysis revealed key relationships:

Strongest Correlations with Survival:

- Passenger Class: -0.338 (negative - lower class numbers = higher survival)
- Fare: +0.257 (positive - higher fares = higher survival)
- Being Alone: -0.203 (negative - traveling alone reduced survival)
- Age: -0.077 (weak negative - slight decrease with age)

Inter-feature Correlations:

- Family Size components (SibSp and Parch): +0.415
- Passenger Class and Fare: -0.549 (higher class = higher fare)
- Age and Passenger Class: -0.369 (younger passengers in lower classes)

Key Insights and Patterns

Primary Survival Factors

1. **Gender dominance**: The most powerful predictor with women having 4× higher survival odds
2. **Socioeconomic privilege**: First-class passengers had 2.6× higher survival than third-class
3. **Family protection**: Traveling with family improved survival odds by 66%
4. **Age vulnerability**: Children prioritized, seniors disadvantaged

Secondary Observations

1. **Optimal family size**: Families of 2-4 members had highest survival rates
2. **Port patterns**: Cherbourg passengers had better outcomes, possibly reflecting class composition
3. **Fare gradients**: Clear linear relationship between ticket price and survival probability
4. **Title significance**: Social titles effectively captured gender and status simultaneously

Data Quality Considerations

1. **Missing cabin data**: 77% missing values limited spatial analysis capabilities
2. **Age imputation needs**: 20% missing age values require careful handling for predictive modeling
3. **Ticket complexity**: Ticket codes showed high variability and unclear patterns
4. **Survival bias**: Analysis limited to recorded passengers, excluding potential unrecorded victims

Conclusions and Implications

This exploratory data analysis reveals that Titanic survival was far from random, instead following clear patterns based on demographic and socioeconomic factors. The "women and children first" protocol was evident but unevenly applied across passenger classes.

Primary Determinants:

1. Gender (74.2% female vs. 18.9% male survival)
2. Passenger class (63.0% first vs. 24.2% third class)
3. Family structure (50.6% with family vs. 30.4% alone)
4. Economic status (fare correlation +0.257)

Implications for Predictive Modeling:

- Gender and passenger class should be primary features
- Family-derived features (family size, traveling alone) add predictive value
- Age groups may be more useful than continuous age
- Missing data strategies crucial for cabin and age variables

Historical Context:
The survival patterns reflect the social structures and evacuation protocols of early 20th century maritime travel, where class distinctions and gender norms significantly influenced life-and-death outcomes during emergencies.

This analysis provides a solid foundation for subsequent predictive modeling efforts and demonstrates the power of thorough exploratory data analysis in understanding complex datasets with human survival outcomes.

Technical Appendix

Analysis Tools: Python (Pandas, NumPy, Matplotlib, Seaborn, SciPy)
Dataset Source: Kaggle Titanic Competition
Analysis Date: October 2025
Missing Data Handling: Analysis of patterns, no imputation performed
Statistical Tests: Descriptive statistics, correlation analysis, cross-tabulation
Visualizations: Histograms, box plots, bar charts, correlation heatmap, scatter plots