Exploratory Data Analysis Report: Titanic Dataset

1. Executive Summary

This analysis examines survival patterns among Titanic passengers using demographic and ticket-class data. Key findings reveal strong correlations between survival rates, passenger class, and fare prices, with notable gender disparities in rescue priority.

2. Dataset Overview

- **Records**: 891 passengers (post-cleaning: 889 after removing embark_town NAs)
- **Features**: 14 columns including survived, pclass, sex, age, fare
- **Target Variable**: survived (Binary: 0=No, 1=Yes)

3. Data Cleaning

- **Missing Values**:
    - deck column dropped (70%+ missing)
    - age imputed with mean (28.7 years)
    - embark_town NAs removed (2 records)
- **Final Features**: 13 columns retained for analysis

4. Key Visual Insights

Demographic Distributions

- **Age**: Right-skewed (Skewness ~0.5), peak at 20-40 years[1]
- **Fare**: Extreme right-skew (Most fares < $100, max $512)[2]

Survival Patterns

- **Gender Disparity**:
    - Female survival rate ~75% vs Male ~20%
- **Class Advantage**:
    - 1st class survival rate: 62.9% vs 3rd class: 24.2%

Economic Factors

- **Fare by Class**:
    - 1st class median fare ($60) 6x higher than 3rd class ($10)
    - Significant outliers in 1st class fares[3]

5. Correlation Analysis

Feature	Survival Correlation
pclass	-0.34
fare	+0.26
age	-0.08
sex (female)	+0.54

6. Critical Findings

1. **Class Privilege**: 1st class passengers had 2.6x higher survival rate than 3rd class
2. **Gender Priority**: "Women and children first" policy clearly reflected in data
3. **Fare Influence**: Higher-paying passengers had better survival odds, likely due to cabin location
4. **Age Paradox**: No strong linear correlation, but survival patterns exist in specific age groups

8. Technical Appendix

- **Tools Used**: Python, Pandas, Seaborn, Matplotlib
- **Data Source**: Seaborn built-in Titanic dataset

9. Conclusion

- The **EDA** demonstrates that *passenger class* and *fare* are the most significant predictors of survival in the Titanic dataset, while *age* and *sex* also play roles but with less direct correlation.