

Exploratory Data Analysis Report: Titanic Dataset

1. Dataset Overview

- Total passengers: 891
- Features analyzed: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- Missing values detected in: Age, Cabin, Embarked
- Mix of categorical and numerical variables

2. Key Findings

Survival Analysis:

- Overall survival rate: ~38.4%
- Class influence: 1st class passengers had higher survival rates compared to 2nd and 3rd class.
- Gender influence: Female passengers (~74.2% survival) vs Male passengers (~18.9% survival).
- Women and children were more likely to survive.

Age distribution:

- Children under 10 years had better survival rates than adults.

Fare & Class Insights:

- Higher fares correlate with higher survival chances.
- Class 3 passengers paid the lowest fares but had the lowest survival rate.

Family Influence:

- Passengers traveling alone had mixed survival chances, depending on class.

- Larger families (>4 members) tended to have lower survival rates.

3. Data Visualizations

- Histogram of Age Distribution: Showed normal distribution with missing values.
- Boxplots: Identified outliers in Fare and Age columns.
- Heatmap: Correlations between Fare, Pclass, and Survived.
- Scatterplots: Showed relationships between Fare, Age, and survival rates.

4. Missing Data Considerations

- Age imputation: Fill missing values using median or mean based on Pclass.
- Cabin missing data: Majority of values missing; might be excluded or used for feature engineering.
- Embarked missing data: Imputation with most frequent value (S).

5. Conclusions & Actionable Insights

Business implications:

- Marketing for higher-class bookings may emphasize survival rates.
- Data can improve ship evacuation planning and future safety measures.

Further steps:

- Feature engineering for Age and Family size variables.
- Optimizing predictive modeling based on correlation analysis.