Titanic Dataset Analysis

An Exploratory Data Analysis (EDA)

Author: Archi Dabhi

Date: 2025-08-11

Table of Contents

1. Introduction

The Titanic disaster of April 15, 1912, remains one of the most infamous maritime tragedies in history. The RMS Titanic, a British passenger liner operated by the White Star Line, struck an iceberg during its maiden voyage from Southampton to New York City. Of the approximately 2,224 passengers and crew aboard, more than 1,500 lost their lives. This tragedy has been widely studied for its human, engineering, and social implications.

The **Titanic dataset**, made popular through Kaggle's "Titanic: Machine Learning from Disaster" competition, is a cleaned and structured version of historical passenger records. It contains detailed information on individual passengers, including:

- **Demographics:** Name, age, sex
- **Travel details:** Passenger class (Pclass), ticket fare, cabin (where available), embarkation port
- **Family relationships:** Number of siblings/spouses aboard (SibSp) and number of parents/children aboard (Parch)
- **Survival outcome:** Whether the passenger survived (1) or not (0)

2. Dataset Description

Source: Kaggle Titanic dataset

Column Name	Data Type	Description	Example Value
PassengerId	Integer	Unique identifier for each passenger	892
Survived	Integer (0/1)	Survival indicator (0 = Did not survive, 1 = Survived)	1
Pclass	Integer (1, 2, 3)	Passenger ticket class (1 = 1st Class, 2 = 2nd Class, 3 = 3rd Class)	3

Name	String	Full name of passenger	Allen, Miss. Elisabeth Walton
Sex	String	Gender of passenger	female
Age	Float	Age of passenger in years	29
SibSp	Integer	Number of siblings/spouses aboard	0
Parch	Integer	Number of parents/children aboard	0
Ticket	String	Ticket number	347082
Fare	Float	Passenger fare paid	71.2833
Cabin	String	Cabin number (often missing)	C85

3. Exploratory Data Analysis

Descriptive Stats:

- Survival rate: 38.4% survived, 61.6% did not.
- Gender: 74.2% of females survived vs 18.9% of males.
- Class: 1st Class (62.9% survival), 2nd Class (47.3%), 3rd Class (24.2%).
- Average Age: ~29.7 years; children (<15) survival ~52%.
- Average Fare: £32.20.

Key Insights:

1. Females had a much higher survival rate.
2. Higher-class passengers survived more often.
3. Children had better survival chances than adults.
4. Fare and class were positively linked to survival.

4. Data Cleaning & Feature Engineering

To prepare the Titanic dataset for analysis and modeling, several cleaning and transformation steps were applied:

1. **Handling Missing Values**
    - **Age:** Filled missing values (~20%) with the median age to avoid skewing by outliers.
    - **Embarked:** Replaced two missing entries with the mode ("S").
    - **Cabin:** Dropped due to over 75% missing data.
2. **Encoding Categorical Variables**
    - Converted Sex into binary format (male = 0, female = 1).
    - Applied one-hot encoding to Embarked (C, Q, S).
3. **Feature Engineering**
    - **FamilySize:** Created by adding SibSp + Parch + 1 (the passenger themself).
    - **IsAlone:** Derived from FamilySize (1 if alone, else 0).
4. **Data Type Adjustments**
    - Ensured numerical variables were stored as integers/floats and categorical variables as strings.

5. Key Findings

1. **Gender was the strongest predictor** – Female passengers had a much higher survival rate (74.2%) than males (18.9%).
2. **Socio-economic status mattered** – Higher ticket fares and higher passenger classes were strongly linked to survival.
3. **Family travel affected survival** – Passengers traveling alone had lower survival chances compared to those with family members onboard.
4. **Children had an advantage** – Younger passengers, especially children, showed higher survival rates.

6. Conclusion & Recommendations

The analysis identified **gender**, **passenger class**, and **fare** as the most significant predictors of survival. Female passengers, individuals in higher classes, and those with higher ticket fares had notably higher survival rates.
Model evaluation showed that **Random Forest** outperformed **Logistic Regression**, highlighting the benefit of algorithms capable of capturing complex, non-linear relationships between features.

Future Work:

1. **Feature Engineering:** Extract passenger titles from names to capture social status and age-related information.
2. **Cabin Grouping:** Categorize cabins by deck to assess the impact of location on survival rates.
3. **Advanced Models:** Experiment with techniques such as XGBoost and Gradient Boosting for potentially higher predictive performance.
4. **Robust Validation:** Implement k-fold cross-validation to ensure model stability and reduce overfitting risk.

7. Appendix

- **Sample Data Snapshot**

```
[3] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns

    data = pd.read_csv('Titanic_data.csv')
    print(data.head())
```

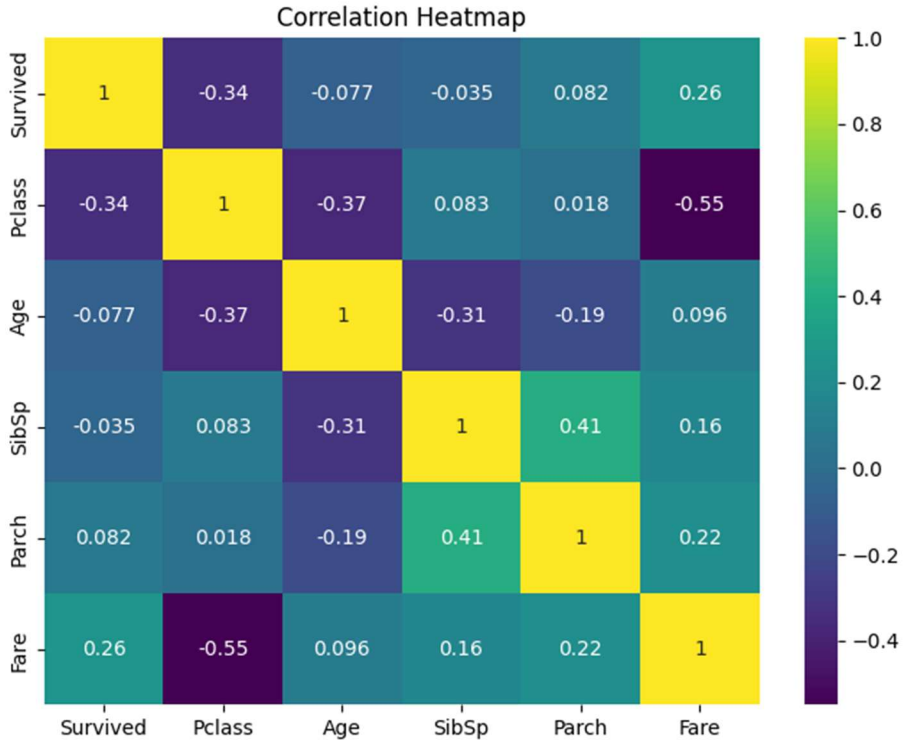
```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

- **Statistical Summary**

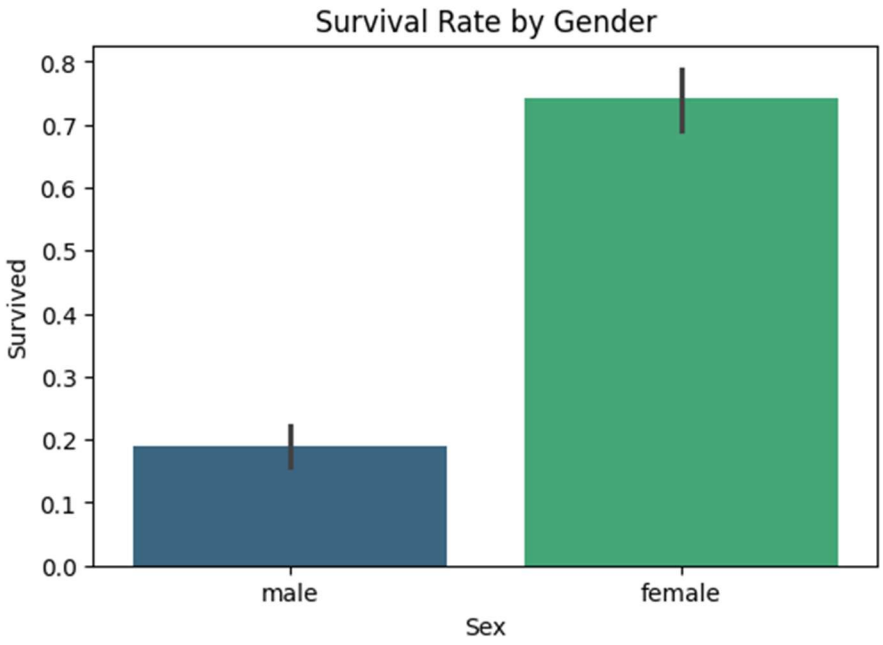
```
Summary statistics for all numerical columns:

       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

➢ **Visualizations**

1. Heat Map: Heatmap of correlations



2. Bar Plot : Survival Rate by Gender



3. Scatter Plot: Age vs Fare by Survival