Statistical Analysis of the Titanic Dataset: Survival Patterns and Passenger Demographics

Research Study

July 31, 2025

Abstract

This study presents a comprehensive statistical analysis of the Titanic passenger dataset, examining survival patterns across different demographic groups and passenger classes. Through frequency analysis, probability calculations, and correlation studies, we investigate the relationships between passenger characteristics and survival outcomes. The analysis reveals significant disparities in survival rates based on gender and passenger class, with implications for understanding historical maritime disaster patterns and emergency evacuation protocols.

1 Introduction

The RMS Titanic disaster of April 15, 1912, remains one of the most studied maritime tragedies in history. This analysis examines passenger data to understand survival patterns and demographic distributions among the 891 passengers in our dataset. Using statistical methods including frequency analysis, probability calculations, and correlation analysis, we aim to identify key factors that influenced survival outcomes.

2 Methodology

2.1 Data Preparation and Libraries

The analysis was conducted using Python with specialized libraries for data manipulation and statistical analysis:

```
%pip install seaborn
```

```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

df = sns.load_dataset('titanic')
```

2.2 Dataset Overview

The Titanic dataset contains information about passenger demographics, ticket class, survival status, and other relevant variables. Table 1 shows a sample of the dataset structure:

```
df.head()
```

Table 1: Sample of Titanic Dataset Structure

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class
0	0	3	male	22.0	1	0	7.2500	S	Third
1	1	1	female	38.0	1	0	71.2833	C	First
2	1	3	female	26.0	0	0	7.9250	S	Third
3	1	1	female	35.0	1	0	53.1000	S	First
4	0	3	male	35.0	0	0	8.0500	S	Third

	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True

3 Statistical Analysis

3.1 Frequency Analysis of Passenger Classes

We begin our analysis by examining the distribution of passengers across different ticket classes:

```
freq_table = df['class'].value_counts()
print(freq_table)
```

```
class
Third     491
First     216
```

```
Second    184
Name: count, dtype: int64
```

3.2 Comprehensive Frequency Analysis

To provide a complete statistical picture, we calculated absolute, relative, and cumulative frequencies for passenger classes:

```
absolute_freq = df['class'].value_counts()
relative_freq = df['class'].value_counts(normalize=True)
cumulative_freq = df['class'].value_counts().cumsum()

print("Absolute Frequency:\n ", absolute_freq)
print("Relative Frequency:\n ", relative_freq)
print("Cumulative Frequency:\n ", cumulative_freq)
```

```
Absolute Frequency:
  class
Third     491
First     216
Second    184
Name: count, dtype: int64
Relative Frequency:
  class
Third     0.551066
First     0.242424
Second    0.206510
Name: proportion, dtype: float64
Cumulative Frequency:
  class
Third     491
First     707
Second    891
Name: count, dtype: int64
```

The analysis reveals that Third Class passengers comprised the majority (55.1%) of the dataset, followed by First Class (24.2%) and Second Class (20.7%) passengers.

3.3 Cross-Tabulation Analysis: Gender and Survival

We constructed a two-way contingency table to examine the relationship between gender and survival:

```
two_way_table = pd.crosstab(df['sex'], df['survived'])
two_way_table['total'] = two_way_table.sum(axis=1)
print(two_way_table)
```

Table 2: Cross-tabulation of Gender and Survival

survived	0	1	total
sex			
female	81	233	314
male	468	109	577

4 Probability Analysis

4.1 Joint Probability

The joint probability of being female and surviving was calculated as follows:

```
joint_prob = pd.crosstab(df['sex'], df['survived'], normalize=True)
print(joint_prob.loc['female', 1])
```

```
0.2615039281705948
```

The joint probability P(Female Survived) = 0.262, indicating that approximately 26.2% of all passengers were female survivors.

4.2 Marginal Probabilities

We calculated the marginal probabilities for gender and survival:

```
marginal_sex = df['sex'].value_counts(normalize=True)
marginal_survived = df['survived'].value_counts(normalize=True)

print(marginal_sex)
print(marginal_sex.loc['female'])
print()
print(marginal_survived)
print(marginal_survived.loc[1])
```

```
sex
male      0.647587
female    0.352413
Name: proportion, dtype: float64
0.35241301907968575
```

```
survived
0    0.616162
1    0.383838
Name: proportion, dtype: float64
0.3838383838383838
```

The marginal probability of being female P(Female) = 0.352, while the marginal probability of survival P(Survived) = 0.384.

4.3 Conditional Probabilities

We calculated conditional probabilities to understand survival patterns:

```
cond_prob= pd.crosstab(df['sex'], df['survived'], normalize='index')
print('Females who survived: ', cond_prob.loc['female', 1])
```

```
Females who survived:  0.7420382165605095
```

```
cond_prob_survived = pd.crosstab(df['survived'], df['sex'], normalize='index')
print("Survivors that were female: ", cond_prob_survived.loc[1, 'female'])
```

```
Survivors that were female:  0.6812865497076024
```

The conditional probability P(Survived|Female) = 0.742 indicates that 74.2% of female passengers survived, while P(Female|Survived) = 0.681 shows that 68.1% of survivors were female.

5 Correlation Analysis

5.1 Age and Fare Correlation

We investigated the relationship between passenger age and fare paid:

```
df_clean = df[['age', 'fare']].dropna()

corr = df_clean['age'].corr(df_clean['fare'])
print("Pearson correlation between age and fare: ", corr)

sns.heatmap(df_clean.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap: Age vs Fare')
plt.show()

sns.pairplot(df_clean)
plt.show()
```

```
plt.scatter(df_clean['age'], df_clean['fare'], alpha=0.5)
plt.xlabel('Age')
plt.ylabel('Fare')
plt.title('Scatter Plot: Age vs Fare')
plt.show()
```

```
Pearson correlation between age and fare:  0.09606669176903888
```

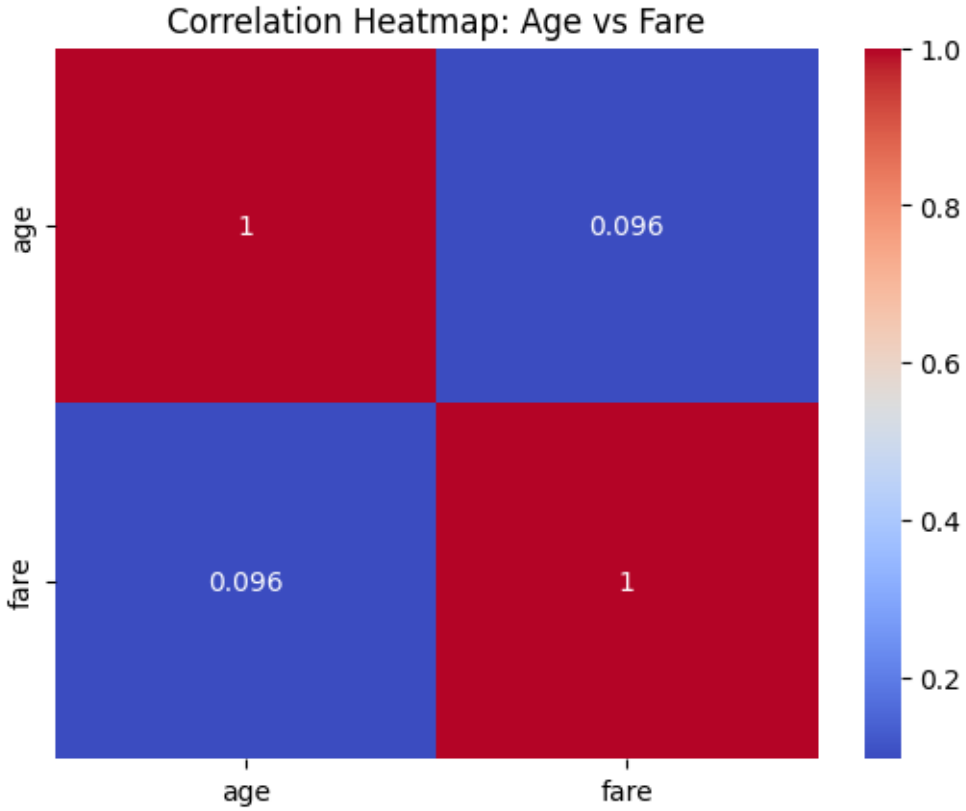

Figure 1: Correlation Heatmap: Age vs Fare

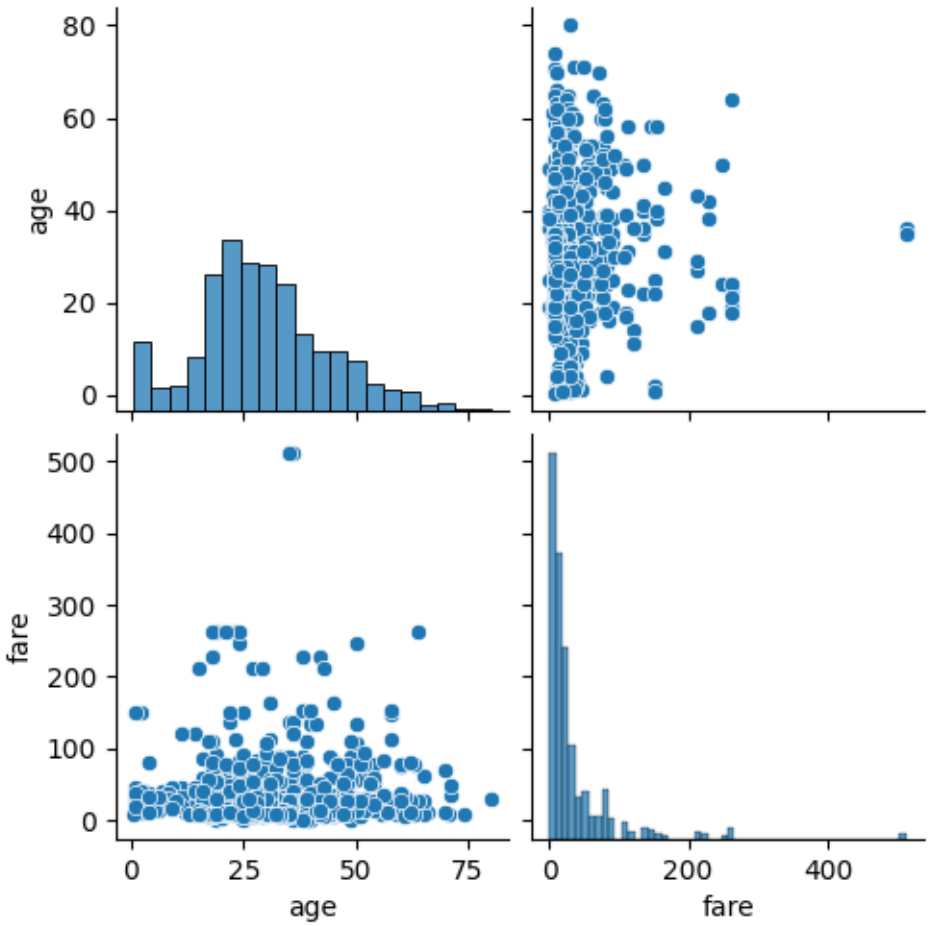

Figure 2: Pairplot: Age vs Fare Distribution

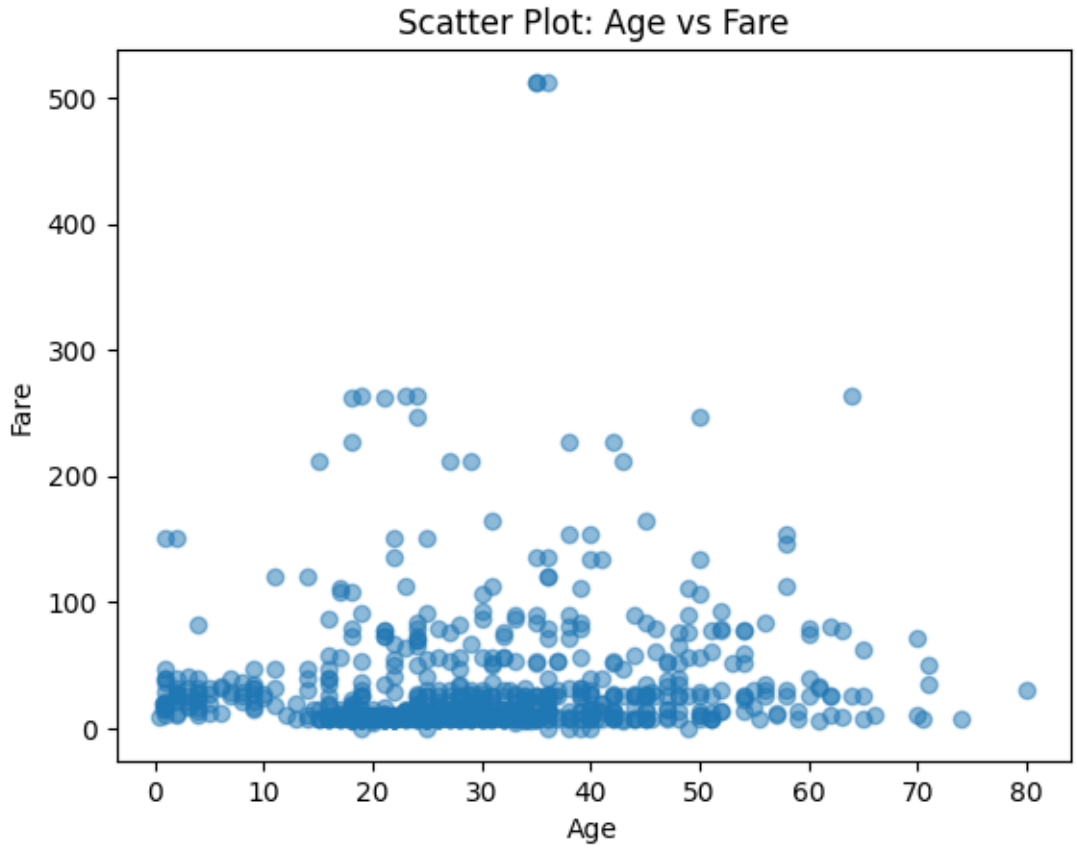

Figure 3: Scatter Plot: Age vs Fare Relationship

The Pearson correlation coefficient between age and fare is 0.096, indicating a weak positive correlation. This suggests that while there is a slight tendency for older passengers to pay higher fares, the relationship is not strong. The correlation coefficient being positive indicates that age and fare tend to increase together, though the effect size is minimal.

6 Survival Analysis by Passenger Class

6.1 Class-Based Survival Patterns

We examined survival rates across different passenger classes using a stacked bar chart:

```
survival_by_class = pd.crosstab(df['class'], df['survived'])
survived_by_class_plot = survival_by_class.plot(kind='bar', stacked=True, color=['salmon','s
plt.xlabel('Class')
plt.ylabel('Number of Passengers')
plt.title('Survival by Class (Stacked Bar Chart)')
plt.legend(['Did Not Survive', 'Survived'])
plt.show()
```

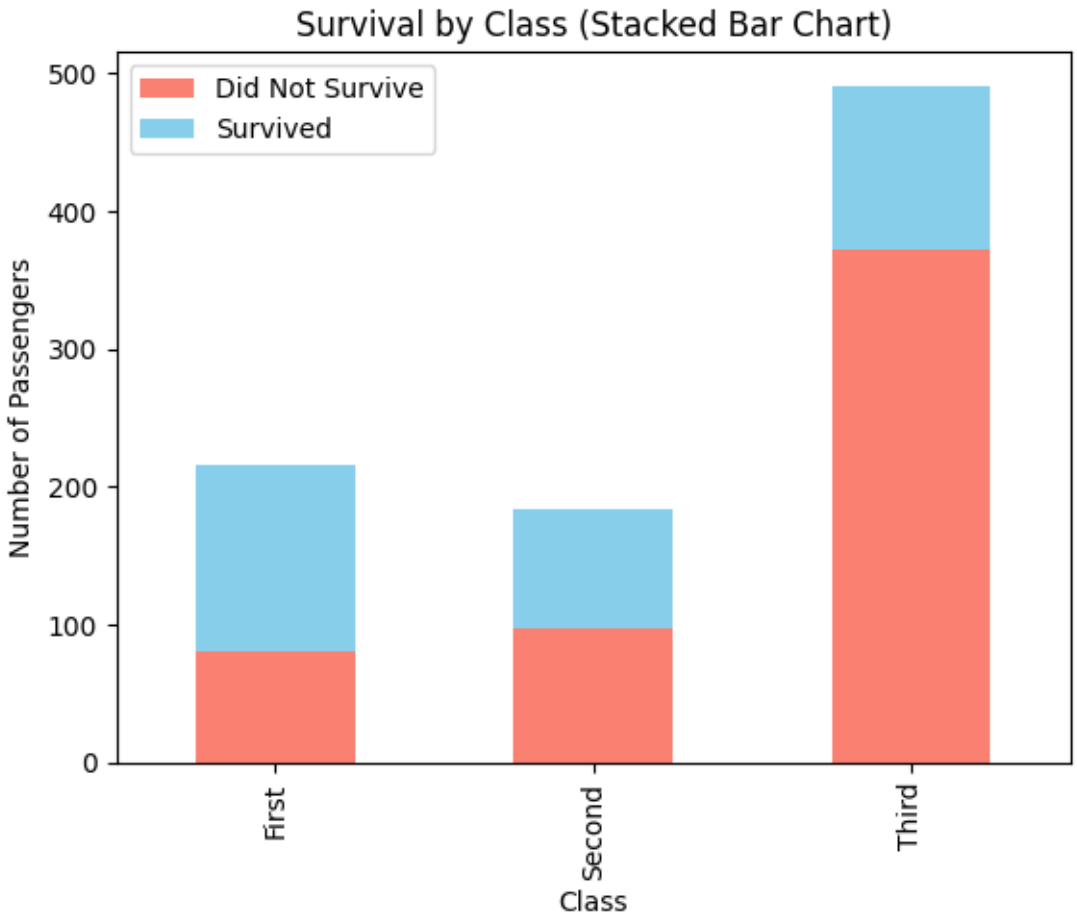

Figure 4: Survival Rates by Passenger Class

The analysis clearly demonstrates that First Class passengers had the highest survival rate. The proportion of survivors in First Class is visibly greater compared to Second and Third Class passengers, highlighting the significant impact of socioeconomic status on survival outcomes during the disaster.

7 Discussion and Conclusions

This comprehensive statistical analysis of the Titanic dataset reveals several critical findings:

1. **Gender Disparity**: Female passengers had a significantly higher survival rate (74.2%) compared to the overall survival rate (38.4%), demonstrating the "women and children first" evacuation protocol.

2. **Class-Based Survival**: First Class passengers experienced the highest survival rates, indicating that socioeconomic status played a crucial role in survival outcomes.

3. **Passenger Distribution**: The majority of passengers (55.1%) traveled in Third Class, yet this group experienced proportionally lower survival rates.

4. **Age-Fare Correlation**: The weak positive correlation (r = 0.096) between age and fare suggests minimal relationship between passenger age and ticket price.

These findings provide valuable insights into the social dynamics and emergency response patterns during one of history's most significant maritime disasters. The statistical evidence supports historical accounts of prioritized evacuation procedures and highlights the intersection of gender and class in survival outcomes.

8 Limitations and Future Research

While this analysis provides comprehensive insights into survival patterns, future research could explore additional variables such as family size, embarkation port, and cabin location to develop more sophisticated predictive models for survival outcomes.