Exploratory Data Analysis (EDA)

Titanic Dataset Analysis

Generated on: June 25, 2025

Executive Summary

This report presents a comprehensive exploratory data analysis of the Titanic passenger dataset. The analysis reveals key patterns in passenger survival rates, demographic characteristics, and factors that influenced survival outcomes during the tragic sinking of the RMS Titanic on April 15, 1912.

Key Findings:
• Overall survival rate of 38.38% (342 out of 891 passengers)
• Significant gender disparity in survival rates (74.2% for women vs 18.9% for men)
• Clear class-based survival differences (63.0% for 1st class vs 24.2% for 3rd class)
• Age and family size also played important roles in survival outcomes
• Passengers who paid higher fares had better survival chances

1. Dataset Overview

The Titanic dataset contains information about **891 passengers** across **12 variables**.

Dataset Structure:
• PassengerId: Unique identifier for each passenger
• Survived: Binary indicator (0 = No, 1 = Yes)
• Pclass: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
• Name: Passenger name
• Sex: Gender (male/female)
• Age: Age in years
• SibSp: Number of siblings/spouses aboard
• Parch: Number of parents/children aboard
• Ticket: Ticket number
• Fare: Passenger fare
• Cabin: Cabin number
• Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Data Quality Issues:
• Age: 177 missing values (19.9%)
• Cabin: 687 missing values (77.1%)
• Embarked: 2 missing values (0.2%)

2. Summary Statistics

Variable	Count	Mean	Std Dev	Min	Max
Age	714	29.70	14.53	0.42	80.00
Fare	891	32.20	49.69	0.00	512.33
SibSp	891	0.52	1.10	0	8
Parch	891	0.38	0.81	0	6

3. Survival Analysis

Overall Survival Statistics:
• Total Passengers: 891
• Survivors: 342 (38.38%)
• Non-survivors: 549 (61.62%)

3.1 Survival by Gender

Gender	Total Passengers	Survivors	Survival Rate
Female	314	233	74.2%
Male	577	109	18.9%

3.2 Survival by Passenger Class

Class	Total Passengers	Survivors	Survival Rate

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

3.3 Survival by Age Group

Age groups were created to better understand survival patterns across different life stages. The analysis shows that children had higher survival rates, likely due to the "women and children first" protocol.

Age Group	Total Passengers	Survivors	Survival Rate
Child (0-12)	69	40	58.0%
Teen (13-18)	70	30	42.9%
Young Adult (19-35)	535	189	35.3%
Adult (36-60)	195	78	40.0%
Senior (60+)	22	5	22.7%

3.4 Survival by Embarkation Port

Passengers embarked from three different ports. Interestingly, passengers who boarded at Cherbourg (C) had higher survival rates, possibly due to higher proportions of first-class passengers from this port.

Port	Code	Total Passengers	Survivors	Survival Rate
Southampton	S	646	219	33.9%
Cherbourg	C	168	93	55.4%
Queenstown	Q	77	30	39.0%

4. Correlation Analysis

Key Correlations with Survival:
• Fare (0.257): Higher fares correlated with better survival
• Passenger Class (-0.338): Lower class numbers (higher class) correlated with better survival
• Age (-0.065): Slight negative correlation with survival
• SibSp (-0.035): Having siblings/spouses slightly reduced survival chances
• Parch (0.082): Having parents/children slightly improved survival chances

Other Notable Correlations:
• Fare and Passenger Class (-0.549): Strong negative correlation (higher class = higher fare)
• Family Size and SibSp/Parch (0.890/0.783): Expected strong correlations
• Age and Passenger Class (-0.369): Older passengers tended to be in higher classes

5. Key Insights and Patterns

1. Gender was the strongest predictor of survival
The survival rate for women (74.2%) was nearly four times higher than for men (18.9%). This reflects the "women and children first" evacuation protocol.

2. Socioeconomic class significantly impacted survival
First-class passengers had the highest survival rate (63.0%), followed by second-class (47.3%) and third-class (24.2%). This suggests that wealth and social status provided advantages during the evacuation.

3. Age played a role, especially for children
Children (0-12 years) had a relatively high survival rate of 58.0%, consistent with the "women and children first" protocol. However, seniors (60+) had the lowest survival rate at 22.7%.

4. Family size had complex effects
Small to medium family sizes (2-4 people) showed better survival rates than traveling alone or in very large families. This suggests optimal group sizes for evacuation assistance.

5. Economic factors mattered
Higher fares were positively correlated with survival, reflecting the class-based differences in cabin locations, evacuation access, and priority treatment.

6. Embarkation port showed surprising differences
Passengers from Cherbourg had significantly higher survival rates (55.4%) compared to Southampton (33.9%) and Queenstown (39.0%), likely due to the class composition of passengers from each port.

6. Methodology

Data Preparation:
• Missing age values were filled with the median age (28 years)
• Missing embarkation values were filled with the mode (Southampton)
• New features were created: Age Groups and Family Size
• Cabin data was not used due to extensive missing values (77.1%)

Analysis Techniques:
• Descriptive statistics for numerical variables
• Cross-tabulation analysis for categorical variables
• Correlation analysis using Pearson correlation coefficient
• Survival rate calculations by different demographic groups
• Data visualization using histograms, bar charts, and correlation heatmaps

Tools Used:
• Python with Pandas for data manipulation
• Matplotlib and Seaborn for data visualization
• Statistical analysis using NumPy
• ReportLab for PDF report generation

7. Conclusions

The Titanic disaster analysis reveals clear patterns of survival that were heavily influenced by demographic and socioeconomic factors. The most significant finding is the stark gender disparity in survival rates, with women having nearly four times better survival chances than men.

Social class played a crucial role, with first-class passengers having more than double the survival rate of third-class passengers. This highlights the tragic reality that wealth and social status provided significant advantages during the emergency evacuation.

The data also supports the historical accounts of "women and children first" protocols, as evidenced by higher survival rates among women and children. However, the analysis also reveals more nuanced patterns, such as the impact of family size and embarkation port on survival outcomes.

These findings provide quantitative support for historical accounts of the disaster and demonstrate how social inequalities persisted even in life-or-death situations. The analysis serves as both a statistical exercise and a sobering reminder of the human cost of the Titanic tragedy.

Report generated on June 25, 2025 at 02:27 PM
Total passengers analyzed: 891
Analysis completed using Python and statistical methods as specified in the task requirements.