Uncovering key determinants of survival in the Titanic tragedy

Analysis by

ERIC TEI

+233 (0) 556918064| Email |LinkedIn|Portfolio

OVERVIEW

The RMS Titanic, widely known simply as Titanic, is synonymous with one of the most tragic maritime disasters in history. Commissioned by the White Star Line, this British passenger liner was heralded as the epitome of luxury and technological advancement of its time. At the time of its completion in 1912, Titanic was the largest ship ever built, measuring approximately 270 meters (882 feet) in length and 28 meters (92 feet) in width. It was designed to offer unparalleled comfort and safety, with features such as watertight compartments and advanced safety mechanisms that led many to believe it was "unsinkable."

On April 10, 1912, Titanic embarked on its highly anticipated maiden voyage from Southampton, England, to New York City. Carrying over 2,200 passengers and crew, the ship represented a cross-section of early 20th-century society, including wealthy industrialists, immigrants seeking new opportunities, and a dedicated crew. However, just four days into the journey, on the night of April 14, 1912, tragedy struck. The Titanic collided with an iceberg on its starboard (right) side. The collision caused catastrophic damage to the ship's hull, flooding multiple watertight compartments and sealing its fate. In the early hours of April 15, the Titanic sank into the icy waters of the North Atlantic, resulting in the loss of more than 1,500 lives.

About the Dataset:

This dataset provides a comprehensive look into the demographics, socioeconomic status, and survival outcomes of Titanic's passengers. It aims to analyze factors that may have influenced survival rates, such as age, gender, ticket class, and family size. By examining these variables, the dataset enables the exploration of patterns and insights into one of history's most infamous events.

The data was sourced from Kaggle and made publicly available by user **vinicius150987**. The dataset serves as a foundation for numerous machine learning and statistical modeling projects. You can access a copy of the dataset here.

The Titanic dataset is a rich and widely used resource for data analysis and machine learning tasks. It contains detailed records of 1,309 passengers aboard the RMS Titanic, which tragically sank on April 15, 1912. This dataset provides information about passenger demographics, socioeconomic status, and survival outcomes.

Key Features of the Dataset:

- **Passenger Details**: Includes personal information such as name, gender, and age.
- **Socioeconomic Indicators**: Captures ticket class (pclass), fare paid, and cabin details.
- **Family Connections**: Tracks the number of siblings/spouses (sibsp) and parents/children (parch) aboard.
- **Survival Information**: Indicates whether a passenger survived (1) or perished (0).
- **Embarkation Details**: Notes the port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).
- **Additional Attributes**: Information on lifeboats used (boat), body identification numbers (body), and the passenger's home destination (home.dest).

Overview:

- The dataset has 14 columns and 1,309 rows, although some columns like age, cabin, and boat have missing values, which is typical in real-world datasets.
- It serves as a practical example for tasks such as data cleaning, exploratory data analysis, feature engineering, and predictive modeling, particularly for binary classification problems like survival prediction.

The Titanic dataset continues to be a benchmark for learning and experimenting with data science techniques, offering valuable insights into the passengers' stories and the factors that influenced survival.

OBJECTIVE

To identify and analyze the key factors influencing passenger survival during the Titanic disaster, including demographics, class, and other attributes, in order to derive meaningful insights about survival probabilities and patterns.

DATA CLEANING AND TRANSFORMATION

To ensure accurate and meaningful analysis, data cleaning and transformation are crucial steps to prepare the dataset for insights. For the Titanic dataset, this process was conducted using Power Query, focusing on making the data more descriptive and suitable for statistical analysis. Below are the key steps undertaken:

1. **Survival Column Transformation**: The survived column was made more descriptive by replacing 0 with "Died" and 1 with "Survived."
2. **Embarked Column Transformation**: The embarked column was enhanced by changing codes (C, Q, S) to their corresponding port names: "Cherbourg," "Queenstown," and "Southampton."
3. **Handling Missing Age Values**: Missing values in the age column were replaced with the median age of 28.
4. **Renaming and Describing Pclass**: The pclass column was renamed to passenger_class and made more descriptive by replacing 1, 2, and 3 with "First Class," "Second Class," and "Third Class," respectively.
5. **Creating Family Size**: The sibsp (siblings and spouse) and parch (parents and children) columns were combined into a new column called family_size to better represent family units aboard the ship.

6. **Removing Insignificant Columns**: Columns with more than 30% missing values—cabin, home.dest, and body—were deleted as they were deemed insignificant for statistical analysis.
7. **Dropping the Boat Column**: The boat column was removed as it was not relevant for statistical analysis.
8. **Categorizing Age**: The age column was categorized into three groups: "0–17" for minors, "18–59" for adults, and "60+" for seniors.

These cleaning and transformation steps were essential in preparing the Titanic dataset for effective analysis and deriving meaningful insights.

DATA ANALYSIS AND VISUALIZATION

The data was analyzed using Pivot Tables and visualized using pivot charts and below are some of my findings

1. Passenger: of the 1309 passengers recorded, 843 were male and 466 were female
2. Survivors: of the 1309 passengers 500 of them survived. Giving a survival rate of 38%.
3. Distribution of Survivors by gender: of the 466 females aboard the ship a whooping 339 survived, this amounts to 73% of the female population surviving. For the male gender only 161 of the 843 survived giving a survival rate of 19%

Row Labels	Died	Survived	Grand Total	Survival %
Female	127	339	466	73%
Male	682	161	843	19%
Grand Total	809	500	1309	38%

FINDINGS: Despite having less females aboard the ship, more of them survived as compared with the males.

4. Comparing survivors by gender: Trying to dig deeper into the survivor data by gender, we found that of the total 500 survivors, 339 female amounts to 68% of the total survivors and the male amounts to 32%

Row Labels	Survived	%
Female	339	68%
Male	161	32%
Grand Total	**500**	**100%**

Findings: Our analysis revealed that out of every 100 survivors, approximately 68 were female. Given that there were more male passengers onboard and considering the general perception of physical strength, this outcome is unlikely to be random. Instead, it suggests a structured evacuation strategy that prioritized female passengers.

5. Analysis by passenger class: The ship has 3 different classes denoted by 1,2 and 3 in the dataset each of them implying First class, Middle Class and Lower Class respectively.

Row Labels	Survived	Grand Total	% of Survival
First Class	200	323	62%
Second Class	119	277	43%
Third Class	181	709	26%
Grand Total	**500**	**1309**	**38%**

The analysis revealed that 62% of first-class ticket holders survived, compared to 43% of middle-class passengers and only 26% of those in the lower class.

Verdict: Ticket class significantly influenced passenger survival rates. Those with higher social status appeared to have better access to resources and opportunities that increased their chances of survival.

6. Analysis by passenger class and gender

Row Labels	Died	Survived	Grand Total	Survival %
First Class	**123**	**200**	**323**	**62%**
Female	5	139	144	97%
Male	118	61	179	34%
Second Class	**158**	**119**	**277**	**43%**
Female	12	94	106	89%
Male	146	25	171	15%
Third Class	**528**	**181**	**709**	**26%**
Female	110	106	216	49%
Male	418	75	493	15%
Grand Total	**809**	**500**	**1309**	**38%**

Building on the established finding that passenger class influenced survival rates, a deeper analysis revealed that gender also played a significant role within each class. Among first-class passengers, 97% of females survived, with only 5 out of 144 not making it. In contrast, male survival rates dropped significantly, with only 34% (61 out of 179) surviving. A similar pattern was observed in the middle class, where 89% of women survived compared to just 15% of men. The trend continued in the lower class, where 49% of women survived, while only 15% of men did.

Verdict: Given the consistency of these survival patterns across all passenger classes, this cannot be attributed to random chance. It strongly suggests that deliberate efforts were made to prioritize female passengers during evacuation.

7. Survival by Age Groups: For this analysis, the passengers were grouped into 3 age categories. Age 0-17 were grouped as 0-17years, age 18-59, as 18-59years and those above 60 as 60+ years and above.

Row Labels	Died	Survived	Grand Total	Survival %
0-17 yrs	73	81	154	53%
18-59 yrs	708	407	1115	37%
60+ yrs	28	12	40	30%
Grand Total	**809**	**500**	**1309**	**38%**

A total of 154 passengers were between the ages of 0 and 17, with a survival rate of 53%. Among the 1,115 passengers aged 18 to 59, 407 survived, resulting in a 37% survival rate. For the elderly (60 and above), only 30% survived.

Verdict: There appears to have been a deliberate effort to prioritize the survival of children, followed by the working-age population (18-59 years).

CHALLENGES

Several challenges were encountered during the data analysis. While the actual number of passengers aboard the Titanic exceeded 2,200, the dataset contained records for only 1,309 passengers.

Handling missing values was a key issue. Some columns were entirely removed due to excessive missing data, while others were excluded as they had no significant impact on the analysis. Specifically, the age column contained missing values, which were addressed

by replacing them with the median age of 28. Notably, the proportion of missing values accounted for approximately 20% of the total dataset, which falls within the acceptable threshold of less than 30%.

CONCLUSION

The analysis reveals that despite the chaos and crisis unfolding, the ship's captain and crew made a conscious effort to prioritize the safety of children and young adults. Additionally, ticket class significantly influenced passengers' chances of survival. Finally, the evacuation process also prioritized females over males.