"The Titanic Disaster: A Data-Driven Exploration of Survival Factors"

INTRODUCTION

The sinking of the RMS Titanic on April 15, 1912, stands as one of the most infamous maritime tragedies in history. Known for its cutting-edge design and lavish features—including squash courts, a Turkish bath, a gymnasium, a barber shop, and the first-ever swimming pool on a ship—the Titanic set sail on its maiden voyage from Southampton, England, to New York City, USA. However, disaster struck when the ship collided with an iceberg in the North Atlantic, resulting in a swift sinking and the tragic loss of over 1,500 lives. Despite the high fatality rate, many passengers managed to survive, offering a compelling dataset to explore survival factors.

PROBLEM STATEMENT

This analysis aims to uncover the key factors that influenced the survival chances of passengers aboard the Titanic. By examining variables such as age, gender, ticket class, and embarkation points, we seek to identify patterns that determined who survived and who didn't.

Microsoft Excel was the primary tool used for conducting this analysis.

THE DATATSET

ThIs Titanic dataset comprises data on 891 passengers who were aboard the Titanic during its tragic maiden voyage. The dataset covers passengers only, not crew.

The dataset for this study was obtained from Kaggle.

Features

Passenger id – Uniquely Identify Passengers

Survival — Survival (0 = No; 1 = Yes)

Pclass — Passenger Class (1 = First Class, 2 = Second Class, 3 = Third Class)

Name — Name of Passengers (Full name)

Sex — Gender (Male or Female)

Age — Age

Sibsp — Number of Siblings/Spouses Aboard

Parch — Number of Parents/Children Aboard
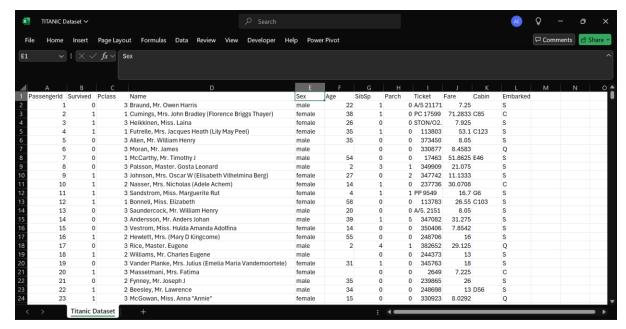
Ticket — Ticket Number

Fare — Passenger Fare

Cabin — Cabin

Embarked — Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
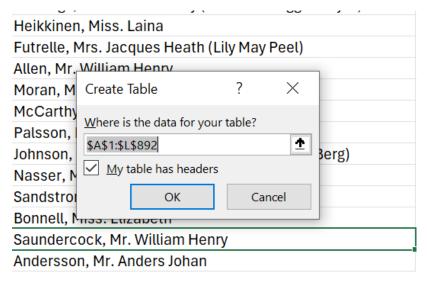
DATA CLEANING

The dataset was first imported into Microsoft Excel before beginning the cleaning process.
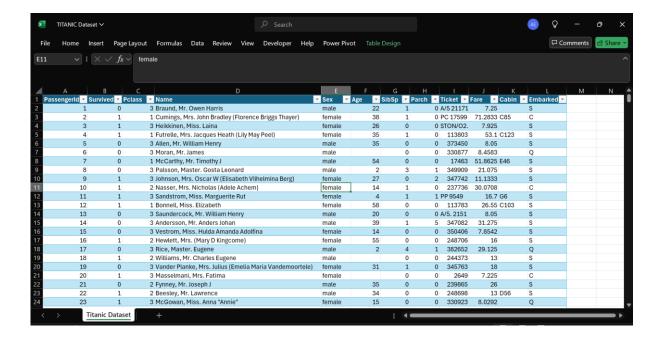
View after import

Once the data was successfully imported and displayed in a sheet, I created a duplicate of the raw dataset in a new sheet for cleaning. After duplicating, I formatted the data as a table using the shortcut CTRL + T. This was to create a standard excel table.
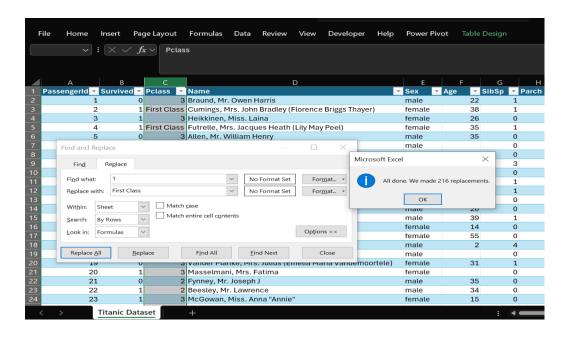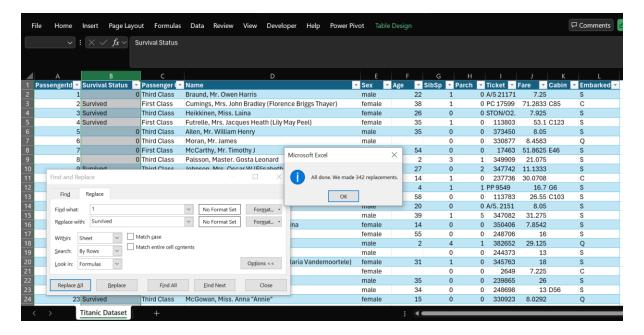

Data in a table format

PCLASS

In the dataset, the 'pclass' column originally used numerical values (1, 2, and 3) to represent passenger classes. To make the data more understandable, I converted these numbers into their respective class names: 'First Class,' 'Second Class,' and 'Third Class'. I used the Find and Replace function by selecting the 'pclass' column, pressing Ctrl + H, and entering '1' in the 'Find what' field and 'First Class' in the 'Replace with' field. After clicking 'Replace All,' I repeated the process to change '2' to 'Second Class' and '3' to 'Third Class'.

I also renamed the 'pclass' column to 'Passenger Class' for improved clarity and to better represent its contents.

Survived

The 'Survived' column in the Titanic dataset indicates whether a passenger survived the disaster, with 1 representing survival and 0 indicating death. I applied the same cleaning process used for the 'pclass' column, replacing 1 with 'Survived' and 0 with 'Dead.' Additionally, I renamed the column to 'Survival Status' for greater clarity and to better reflect its meaning.

Name

The 'Name' column in the Titanic dataset included the names of all passengers, but it was not formatted correctly.

To properly separate the names, I used the **"Split-to-Column"** function from the Data tab. This was necessary because the names contained various delimiters such as com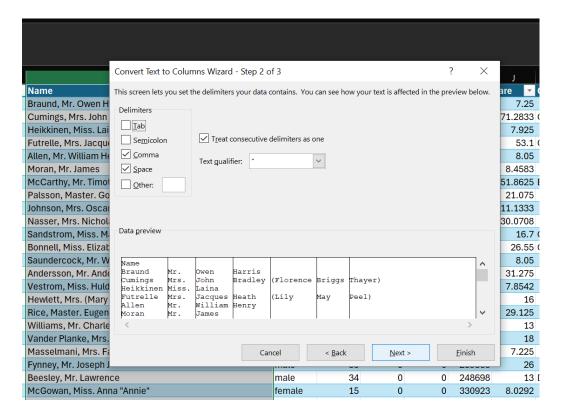mas, brackets, and spaces. By using this function, I was able to split the names into distinct components and ensure proper organization of the data.

Setting delimiters for the text to column function

To manage the split names, I added destination columns right after the 'Name' column. This approach ensured that the split names didn't overwrite data in other columns, preserving the dataset's integrity.

Although the names were successfully split, I still needed to combine the titles with the first and last names. I achieved this by using the CONCAT function.

```
=CONCAT([@Title]," ",[@Lastname]," ",[@Firstname])
```

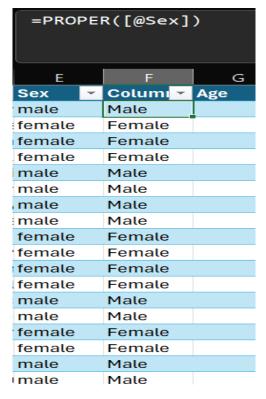| D |
| --- |
| **Fullname** |
| Mr. Braund Owen |
| Mrs. Cumings John |
| Miss. Heikkinen Laina |
| Mrs. Futrelle Jacques |
| Mr. Allen William |
| Mr. Moran James |
| Mr. McCarthy Timothy |
| Master. Palsson Gosta |
| Mrs. Johnson Oscar |
| Mrs. Nasser Nicholas |
| Miss. Sandstrom Marguerite |
| Miss. Bonnell Elizabeth |
| Mr. Saundercock William |
| Mr. Andersson Anders |
| Miss. Vestrom Hulda |
| Mrs. Hewlett (Mary |
| Master. Rice Eugene |
| Mr. Williams Charles |

Name column after cleaning

After this was successfully done, I copied the column and pasted it in a new column as values to get rid of the formula. Then I deleted the initial column and named the new column 'Full Name'.

SEX

The 'Sex' column in the dataset contained the gender of each passenger that boarded the ship. I needed to change the records to Proper Case since they were all in small letters and I did this using Excel's **Proper Function**.
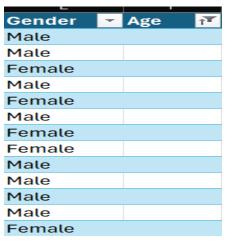
```
=PROPER([@Sex])
```

Sex	Column	Age
male	Male	
female	Female	
female	Female	
female	Female	
male	Male	
male	Male	
male	Male	
male	Male	
female	Female	
female	Female	
female	Female	
female	Female	
male	Male	
male	Male	
female	Female	
female	Female	
male	Male	
male	Male	

Sex column Before and After Cleaning

After cleaning the sex column, I renamed it 'Gender' for accurate analysis and interpretation of the data.
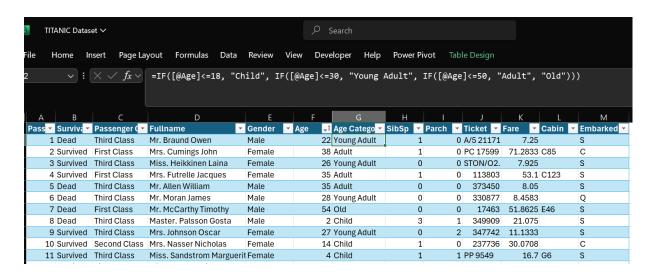
Age

This column recorded the age of each passenger on the ship. As the dataset was formatted as a table, I utilized the filter feature to identify any missing values (blanks) in the column, which I successfully located.

Gender	Age
Male	
Male	
Female	
Male	
Female	
Male	
Female	
Female	
Male	
Male	
Male	
Male	
Female	

blanks in the age column

I addressed the missing values by replacing them with the median age of the passengers. I chose to use the median rather than the mean because the median is less affected by outliers. I calculated the median age using Excel's MEDIAN function, then filtered out the blanks and filled them with the median value using Flash Fill.

Additionally, I created a new column called 'Age Group' to categorize passengers into distinct age groups. This made it easier to compare survival rates across different age groups rather than individual ages. I accomplished this by using the IF function.

Assigning Age Groups with the IF Function

Parch and SibSp

The 'Parch' and 'SibSp' columns contained information about the number of 'parents/children' and 'siblings/spouses' that each passenger had on the ship. So, I added these columns together using the **SUM** function to get a new column which I called 'Family Size'.

SibSp	Parch	Family	T
1	0	1	A
1	0	1	F
0	0	0	S
1	0	1	
0	0	0	
0	0	0	
0	0	0	
3	1	4	
0	2	2	
1	0	1	

Ticket

The 'Ticket' column contained information about the ticket number of each passenger onboard the ship. Since it was wasn't needed for the analysis, I deleted it.
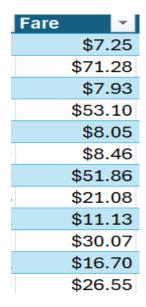
Same process was done for the cabin columns since they weren't needed for the analysis.

NOTE: I was working on the duplicated data and not the original data. Hence, I could comfortably delete from the duplicated one since I had the original data safely stored.
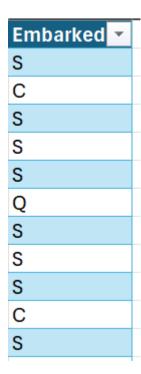
Fare

The 'Fare' column contained information about the cost of each passenger's ticket. I formatted the column as Currency ($) and rounded up the values to 2 decimal places for accuracy.

Fare
$7.25
$71.28
$7.93
$53.10
$8.05
$8.46
$51.86
$21.08
$11.13
$30.07
$16.70
$26.55

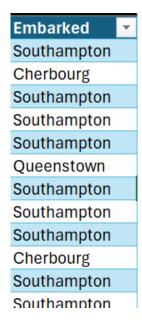Fare column after cleaning

Embarked

The 'Embarked' column contained information about the port where each passenger embarked. There were three ports which were abbreviated S, C, and Q.

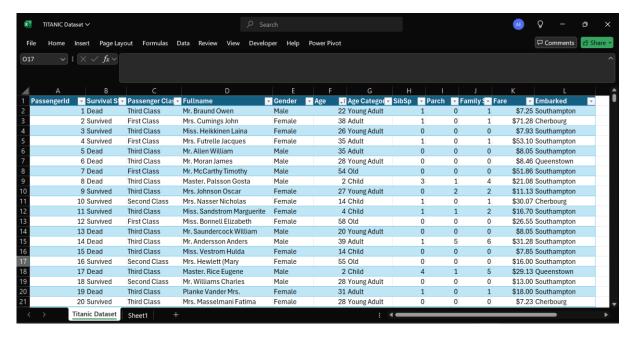| Embarked |
|---|
| S |
| C |
| S |
| S |
| S |
| Q |
| S |
| S |
| S |
| C |
| S |

Embarked column before cleaning

For clarity, I replaced S, C, Q in the embarked column with their corresponding names which are Southampton, Cherbourg, and Queenstown respectively by highlighting the column and using Find and Replace.
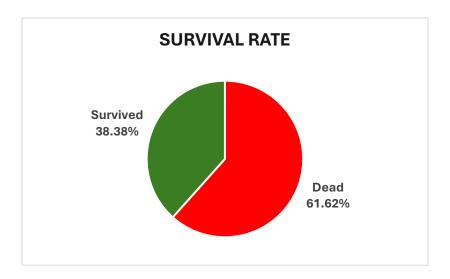

Embarked column after cleaning

titanic dataset after general cleaning

ANALYSIS AND VISUALISATION

Of the 891 passengers aboard the Titanic, only 38.3% (342) survived, while a tragic 62.6% (549) lost their lives. This underscores the devastating impact of the disaster.

SURVIVAL RATE

Survived
38.38%

Dead
61.62%

GENDER ANALYSIS

Although women made up only about 35.2% of the total passengers, their survival rate was much higher compared to men, who accounted for 64.8%. Notably, 26.2% of female passengers survived, while just 12.2% of male passengers did.

SURVIVAL RATE BY GENDER

	Female		Male	
	Dead 9.09%	Survived 26.15%	Dead 52.53%	Survived 12.23%

This difference in survival rates illustrates the "women and children first" policy commonly followed during maritime disasters, including the Titanic tragedy. The prioritization of women for lifeboat access significantly increased their chances of survival, even though they were outnumbered by male passengers.

As a result, gender was a key factor in determining survival during the Titanic disaster. Women benefited from the protective measures in place during evacuation, which ultimately contributed to their higher survival rates. This also reflects the sacrifices made by many men who likely gave up their spots on lifeboats to ensure the safety of women and children.
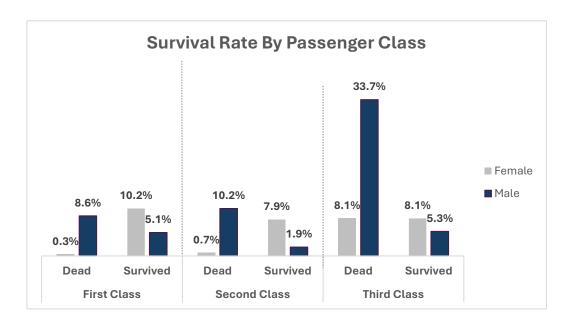
Additionally, this analysis highlights the influence of societal norms and gender roles at the time. The adherence to these customs, even in a life-or-death situation, reveals the deeply ingrained expectations of chivalry and selflessness among men, leading to their lower survival rates. This gender dynamic provides a broader understanding of how social structures affected individual outcomes in the tragedy.

PASSENGER CLASS ANALYSIS

 By distribution, third-class passengers constituted the majority, accounting for 55.2% of the total passengers. First-class passengers comprised 24.3%, while second-class passengers were the smallest group, representing 20.7%.

Among first-class passengers, approximately 15.3% survived, while 9% did not. Second-class passengers had a survival rate of 9.8%, while 10.9%

died. Meanwhile, third-class passengers experienced the highest fatality rate, with only 13.4% surviving, while 41.8% did not.

Survival Rate By Passenger Class

		First Class		Second Class		Third Class	
Female		0.3%	10.2%	0.7%	7.9%	8.1%	8.1%
Male		8.6%	5.1%	10.2%	1.9%	33.7%	5.3%
		Dead	Survived	Dead	Survived	Dead	Survived

the survival rates aboard the Titanic varied significantly across social classes, underscoring the impact of wealth and status on passengers' chances of survival.

First Class passengers made up 24.2% of the total, with 15.3% surviving and 9.0% perishing. Their higher survival rate reflects their greater access to lifeboats and prioritization during evacuation.

Second Class passengers constituted 20.7%, with a survival rate of 9.8% and a death rate of 10.9%. While their chances of survival were lower than First Class, they still fared better than Third Class passengers.

Third Class passengers, who comprised 55.1% of the total, faced the highest fatality rate, with 41.8% losing their lives and only 13.4% surviving. This highlights the severe disadvantages they faced, such as being located on the lower decks and having limited access to lifeboats.
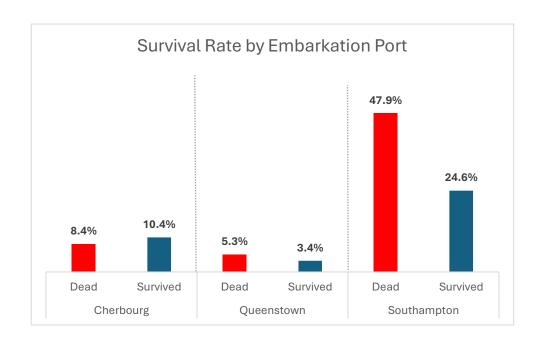
PASSENGER CLASS AND GENDER ANALYSIS

Among First Class survivors, 10.2% were women and 5.1% were men. In Second Class, 7.9% of the survivors were women and 1.9% were men, while in Third Class, 8.1% were women and 5.3% were men. This shows that women had higher survival rates across all classes compared to men. It also suggests that the "women and children first" policy was more strictly applied in higher classes, as the survival rate of women decreases progressively in the lower passenger classes.

Count of Passenger Class	Column Labels		
Row Labels	Female	Male	Grand Total
First Class	**10.5%**	**13.7%**	**24.2%**
Dead	0.3%	8.6%	9.0%
Survived	10.2%	5.1%	15.3%
Second Class	**8.5%**	**12.1%**	**20.7%**
Dead	0.7%	10.2%	10.9%
Survived	7.9%	1.9%	9.8%
Third Class	**16.2%**	**38.9%**	**55.1%**
Dead	8.1%	33.7%	41.8%
Survived	8.1%	5.3%	13.4%

EMBARKATION PORT ANALYSIS

 Of the passengers who boarded in Cherbourg, 10.4% survived and 8.4% perished. Those embarking from Queenstown had a lower survival rate, with just 3.4% surviving and 5.3% dying. Southampton, which had the largest share of passengers, also experienced the highest fatality rate, with 47.9% of its passengers dying and 24.6% surviving.

Survival Rate by Embarkation Port

8.4% (Dead) — Cherbourg
10.4% (Survived) — Cherbourg
5.3% (Dead) — Queenstown
3.4% (Survived) — Queenstown
47.9% (Dead) — Southampton
24.6% (Survived) — Southampton

AGE CATEGORY ANALYSIS

The passenger distribution on the Titanic was as follows: young adults (ages 19–30) made up the largest group at 50.2%, followed by adults (ages 31–50) at 27%, children (age ≤18) at 15.6%, and older adults (age >50) at just 7.2%.
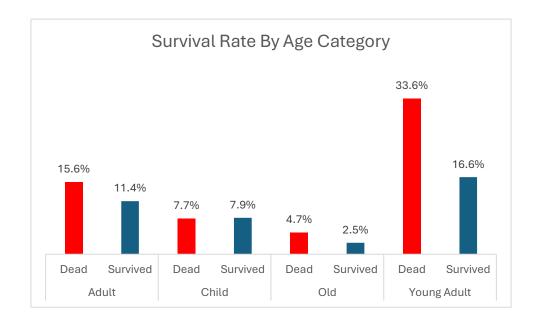
This distribution suggests:

In the early 20th century, young adults in their 20s were often at a stage of life marked by career exploration, travel, and adventure. This demographic might have been particularly attracted to the opportunities presented by a voyage on the Titanic, whether for adventure, work, or immigration.

Adults aged 31 to 50 were likely more established in their careers and personal lives, leading to a more selective approach to travel. They may have chosen the Titanic for specific reasons, such as relocating for professional opportunities or seeking a new life in New York, their intended destination.

Children on board were likely part of families traveling together or unaccompanied minors moving for various reasons. Families might have chosen to emigrate together to seize new opportunities, making the presence of children more common in this context.

Older adults over 50 years old were less represented and might have included retirees, individuals traveling with family, or those seeking a leisurely journey. Their smaller representation could be attributed to their preference for more comfortable and leisurely travel experiences offered by the Titanic's luxury and prestige.

Survival Rate By Age Category

Category	Dead	Survived
Adult	15.6%	11.4%
Child	7.7%	7.9%
Old	4.7%	2.5%
Young Adult	33.6%	16.6%

Unexpectedly, young adults aged 19–30 years experienced the highest fatality rate among all age groups, despite their substantial representation. This could be due to their possible misjudgement of the severity of the situation, leading to delayed or ineffective evacuation. Additionally, they might have been involved in riskier behaviours or failed to follow evacuation instructions effectively.

In contrast, children (≤18 years) and older adults (>50 years) had relatively lower survival rates compared to adults aged 31–50 years. This discrepancy might be attributed to their heightened vulnerability during the evacuation. Children likely depended on adults for help, which could have hindered their chances of survival. Older adults may have faced difficulties with mobility or health issues that made it harder for them to escape.

Adults aged 31–50 years had a moderate survival rate, reflecting their physical capability and responsibility to help others during the evacuation. This age group was likely in a position to balance their own safety with the need to assist family members or fellow passengers, which might have influenced their survival outcomes.

RECOMMENDATIONS

1. Strengthen Safety Measures and Regulations: Enhance maritime safety protocols to prevent future tragedies. The Titanic disaster underscores the need for robust safety measures, including adequate lifeboats, regular emergency drills, and advanced communication systems.

2. Ensure Equitable Access to Safety Resources: Distribute life-saving resources, such as lifeboats and emergency supplies, fairly among all passengers, regardless of their class or financial status. This approach helps mitigate the disproportionate impact of disasters on vulnerable groups.

3. Implement Comprehensive Training Programs: Provide thorough training for both passengers and crew on emergency procedures and

evacuation protocols. Educating individuals on effective crisis response can significantly improve safety outcomes and boost survival rates.

4. Increase Public Awareness: Raise awareness about maritime safety and disaster preparedness through educational campaigns, outreach programs, and media coverage. Empowering the public with knowledge on emergency preparedness can help individuals protect themselves and potentially save lives during crises.

CONCLUSION

In summary, this analysis of Titanic data has provided insights into the factors that influenced survival during the ship's sinking. We examined variables such as age, passenger location etc. Revealing that certain groups, such as women and children, had higher survival rates. This underscores the critical importance of effective safety measures and equal access to life-saving resources for all passengers. By leveraging data to inform safety protocols and increasing awareness about emergency preparedness, we can work towards preventing future tragedies like the Titanic disaster.