# Project 1: Exploring Titanic Database

By Aiman Aisamuddin Ab Ghapar

## 1. Understanding the Business Context

One of the most infamous shipwrecks in the history of mankind is the sinking of the Titanic. The widely considered "unsinkable" RMS Titanic sank due to the iceberg collision during its maiden voyage on April 15th of 1912. Due to shortage of lifeboats for everyone onboard, only about a quarter of the passengers survived. Undeniably, there was some element of luck involved in surviving the shipwreck. However, it does seem that there are some groups of people were more likely to survive than others. Therefore, these datasets are analyzed to determine out what sorts of people were more likely to survive the shipwreck.

## 2. Understanding the Technical Context

These data are collected from the Kaggle website which is provided for the Titanic Machine Learning Competition [1]. The data provided only consist of 891 passengers out of 2224 passengers. The data might have been manually input by some data entry personnel without using any electronic system since during 1912 computer system were limited. Due to this, the data might be prone to fraud since there were not cross-checking between systems in checking the passengers' details. The data might also be incorrectly recorded by the officer since they were collected manually. The data is incomplete, since the data provided less than half of the actual number of passengers in the Titanic. Plus, some passengers' details were not recorded.

## 3. Understanding the Tables and Fields

The Titanic database provided only consist of one table called "passengers" which contains the passengers' details. The following are the fields contain in the table with definition:

- PassengerId – passengers' ID
- Survived – Survival status of the passengers, 0 = No, 1 = Yes
- Pclass – Ticket class, 1 = 1st, 2 = 2nd , 3 = 3rd
- Name – passengers' name
- Sex – passengers' gender
- Age – passengers' age
- SibSp – number of siblings / spouses aboard the Titanic
- Parch – number of parents / children aboard the Titanic
- Ticket – ticket number
- Fare – passengers' fare
- Cabin – Cabin number
- Embarked – port of embarkation, C = Cherbourg, Q = Queenstown, S = Southampton

*Variable notes:*

**Pclass**: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

**Age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

**SibSp**: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

**Parch**: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

The data well-structured and does not need extensive pre-processing except for the null value in the “Age”, “Cabin”, and “Embarked” fields. The number of null values can be determined using this query:

```
SELECT count(*) FROM passengers

WHERE field_x is NULL
```

We found that there are 177, 687 and 2 null values in “Age”, “Cabin” and “Embarked” respectively. In the data exploration we could disregard the null values in “Age” and “Embarked” to get the average age of the passengers and the embarkation distribution of the Titanic. The “Cabin” field could be disregard completely since the number of null values in the field is more than 75% of the number of passengers in the datasets.

## 4. Free Exploration

In this project we explored a few aspects that contribute to the survival of the people in the shipwreck.

### a) Are children and elderlies have a higher survival rate?

Before answering the question, we need to define the age group first. INTEGRIS Health classifies the age group as follows [2]:

- Infant = 0-1 year
- Toddler = 2-4 yrs
- Child = 5-12 yrs
- Teen = 13-19 yrs
- Adult = 20-39 yrs
- Middle-Aged Adult = 40-59 yrs
- Senior Adult = 60+

To simplify our analysis, we define the children age group as from infant to child/teen (0-15 years), adult as from teen to middle-aged adult (15-59 years) and elderlies individual age of 60 years and above. First let’s determined the number of children and elderlies in the datasets using the following queries:

```
SELECT count(*) FROM passengers

WHERE (Age > 59 OR Age < 16)
```

| count(*) |
|---:|
| 79 |

Then, let’s determined how many people survived the shipwreck using the following queries:

```
SELECT count(*) FROM passengers

WHERE Survived = 1
```

| count(*) |
|---:|
| 342 |

We found that there are 342 people survived the shipwreck. Then we determine the number of children and elderlies survived using the following queries:

```
SELECT count(*) FROM passengers

WHERE (Age > 59 OR Age < 16) AND Survived = 1
```

| count(*) |
|---:|
| 37 |

Using these results, we can determine the rate of survival between the age group as follows:

$$Rate\ of\ survival\ (children\ and\ elderlies) = \frac{37}{79} \times 100\% = 46\%$$

$$Rate\ of\ survival\ (adult) = \frac{342-37}{891-79} \times 100\% = 37.56\%$$

The survival rate of children and elderlies’ group are higher as depicted in the movie.

### b) Are females more likely to survive in this incident?

First, we need to determine the number of females in the datasets as follows:

```
SELECT count(*) FROM passengers

WHERE Sex = "female"
```

| count(*) |
|---:|
| 314 |

Then, we need to determine how many females survived the shipwreck as follows:

```
SELECT count(*) FROM passengers

WHERE Sex = "female" AND Survived = 1
```

| count(*) |
|---:|
| 233 |

Calculating the survival rate as below, we found that female passengers are more likely to survive the shipwreck.

$$Rate\ of\ survival\ (female) = \frac{233}{314} \times 100\% = 74.2\%$$

$$Rate\ of\ survival\ (female) = \frac{342 - 233}{891 - 314} \times 100\% = 18.89\%$$

### c) Are rich people have a higher survival rate?

Let's assume the rich people are the passengers with 1st class ticket while the 2nd and 3rd class ticket are the lesser one. First, we need to determine the total rich people in the datasets as follows:

```
SELECT count(*) FROM passengers

WHERE Pclass = 1
```

| count(*) |
|---:|
| 216 |

Then, we need to determine how many of the rich people survived the shipwreck as follows:

```
SELECT count(*) FROM passengers

WHERE Pclass = 1 AND Survived = 1
```

| count(*) |
|---:|
| 136 |

Then, we can calculate the survival rate of the rich people and compared it to the lesser one as below:

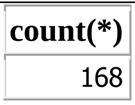
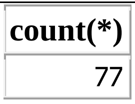
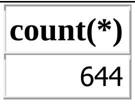
$$Rate\ of\ survival\ (rich) = \frac{136}{216} \times 100\% = 63\%$$

$$Rate\ of\ survival\ (lesser) = \frac{342 - 136}{891 - 216} \times 100\% = 30.5\%$$

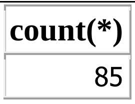
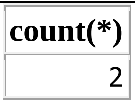
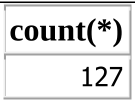
Definitely, we can see that the rich people are more likely to survive the shipwreck.

### d) Is there any relation between embarkation port with the socio-economic status of the passengers?

First, let's determine the total passengers in each embarkation ports as follows:

| | | |
|---|---|---|
| Cherbourg | SELECT count(*) FROM passengers<br>WHERE Embarked = "C" |  count(*)<br>168 |
| Queenstown | SELECT count(*) FROM passengers<br>WHERE Embarked = "Q" |  count(*)<br>77 |
| Southampton | SELECT count(*) FROM passengers<br>WHERE Embarked = "S" |  count(*)<br>644 |

Then, we determined the total rich passengers (Pclass = 1) in each port:

| | | |
|---|---|---|
| Cherbourg | SELECT count(*) FROM passengers<br>WHERE Embarked = "C" AND Pclass = 1 |  count(*)<br>85 |
| Queenstown | SELECT count(*) FROM passengers<br>WHERE Embarked = "Q" AND Pclass = 1 |  count(*)<br>2 |
| Southampton | SELECT count(*) FROM passengers<br>WHERE Embarked = "S" AND Pclass = 1 |  count(*)<br>127 |

Then we determine the percentage of rich people from each port as follows:

$$Percentage\ of\ rich\ people\ in\ Cherbourg = \frac{85}{168} \times 100\% = 50.6\%$$

$$Percentage\ of\ rich\ people\ in\ Queenstown = \frac{2}{77} \times 100\% = 2.6\%$$

$$Percentage\ of\ rich\ people\ in\ Cherbourg = \frac{127}{644} \times 100\% = 19.7\%$$

From this analysis, we can safely estimate that most rich people lives close to Cherbourg.

## Bibliography

[1] kaggle, “Titanic - Machine Learning for Disaster”.

[2] I. H. Oklohama, “Stage of Life: Health for Every Age,” 9 October 2015. [Online]. Available: https://integrisok.com/resources/on-your-health/2015/october/stages-of-life-health-for-every-age. [Accessed 18 July 2022].