# Titanic Data Analysis

## Introduction

This is my first data science project. In this project, I have studied and analysed the data of the people who were present on RMS Titanic, and tried to approximate some of the missing values.

## Source

https://www.kaggle.com/c/titanic/data

## Variable Descriptions

| | |
|---|---|
| Survival | Survival |
| | (0 = No; 1 = Yes) |
| pclass | Passenger Class |
| | (1 = 1st; 2 = 2nd; 3 = 3rd) |
| name | Name |
| sex | Sex |
| 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) |

SPECIAL NOTES:

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than 1;

If the Age is estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used:

For sibsp and parch.

| | |
|---|---|
| Sibling: | Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic |
| Spouse: | Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored) |
| Parent: | Mother or Father of Passenger Aboard Titanic |
| Child: | Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic |

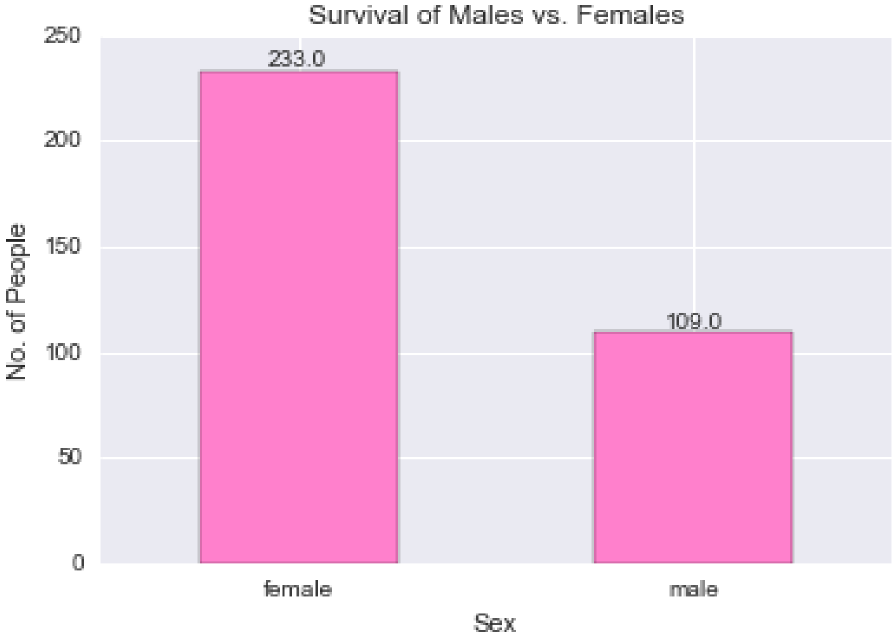
Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbours in a village, however, the definitions do not support such relations.

## Analysing the Sex component

Following table shows us the comparison between the numbers of males and females that survived.

| **Sex** | **Number of People** | **Survived** | **Proportion Survived** |
|---|---|---|---|
| **female** | 314 | 233 | 0.74 |
| **male** | 577 | 109 | 0.19 |

The graph below shows the number of females that survived vs. the number of men that survived.

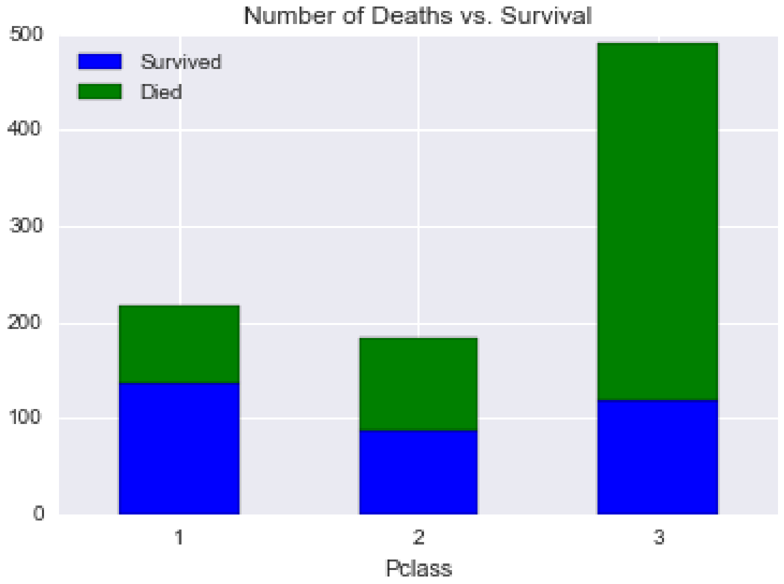


From the above graph and table, we conclude that evacuating females was given a much greater priority over evacuating men.

## Analysing the Pclass component

Below we can see the distribution of people among different Pclasses:

| Pclass | Survived | Died | Total | Proportion Survived |
|---|---|---|---|---|
| 1 | 136 | 80 | 216 | 0.63 |
| 2 | 87 | 97 | 184 | 0.47 |
| 3 | 119 | 372 | 491 | 0.24 |



We can see that there was a lot of bias towards evacuating people on the basis of their Pclass.

## Analysing the Age component

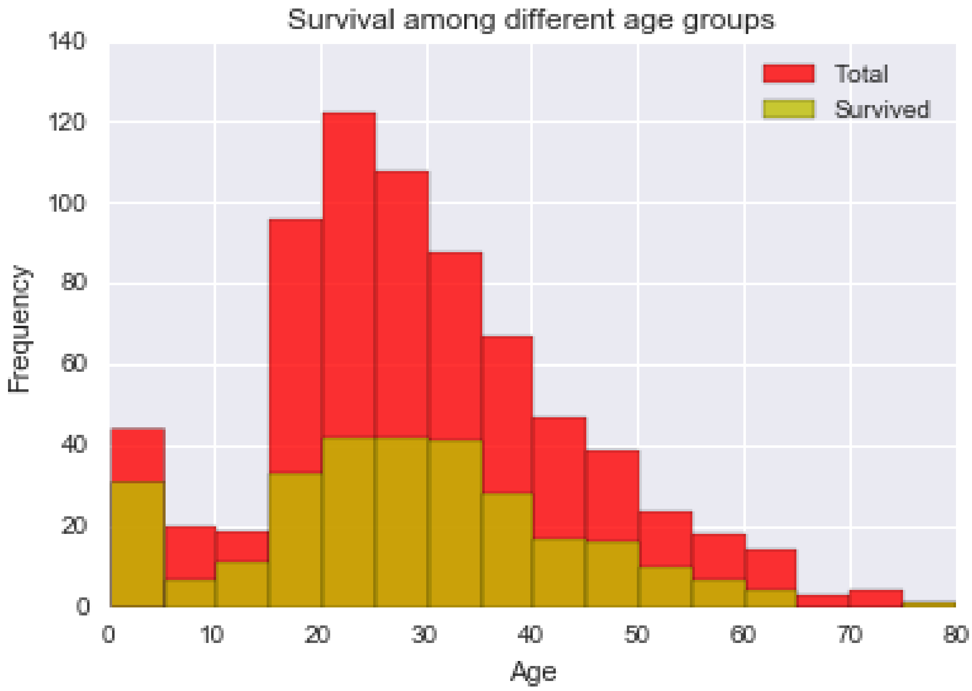
In the original data that was given to us, the ages of 177 people were missing.

In the following table, the survival data of the people has been given in the corresponding age-ranges (excluding the missing ages).

| Age Range | Number of People | Survived | Proportion Survived |
|---|---|---|---|
| 0-5 | 44 | 31 | 0.7 |
| 5-10 | 20 | 7 | 0.35 |
| 10-15 | 19 | 11 | 0.58 |
| 15-20 | 96 | 33 | 0.34 |
| 20-25 | 122 | 42 | 0.34 |
| 25-30 | 108 | 42 | 0.39 |
| 30-35 | 88 | 41 | 0.47 |
| 35-40 | 67 | 28 | 0.42 |
| 40-45 | 47 | 17 | 0.36 |
| 45-50 | 39 | 16 | 0.41 |
| 50-55 | 24 | 10 | 0.42 |
| 55-60 | 18 | 7 | 0.39 |
| 60-65 | 14 | 4 | 0.29 |
| 65-70 | 3 | 0 | 0 |
| 70-75 | 4 | 0 | 0 |
| 75-80 | 1 | 1 | 1 |

It can be seen that evacuation of children aged 0-5 years was given a very high priority.

Also, we can see a sudden drop in the Proportion Survived in the Age-Range of 5-10 years.
On careful examination, it was found that the reason for this is that this age-group consisted of 80% people (16 children) form Pclass 3, and the remaining 20% people (4 children) from Pclass 2, with not even a single person from Pclass 1. Hence the low rate of Proportion Survived. This conforms to our analysis based on Pclasses earlier. Not only this, the 7 people who survived comprise of all of the 4 people from Pclass 2.

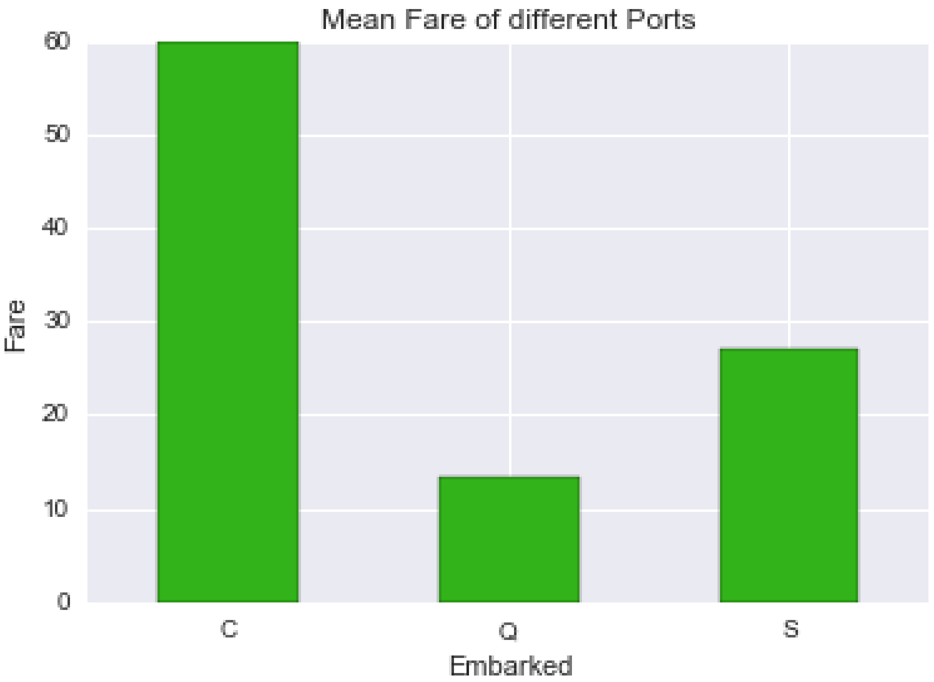


## Analysing Port of Embarkation Data

Following table shows us the distribution of people according to their ports of embarkation:

| Embarked | Number of People | Survived | Proportion Survived |
|---|---|---|---|
| C | 168 | 93 | 0.55 |
| Q | 77 | 30 | 0.39 |
| S | 644 | 217 | 0.34 |

From the table, we can see that the proportion of people survived whose port of embarkation was C is the highest.
To find the reason behind this, the mean fares of the 3 ports were calculated. It was found that the mean fare of port C was the highest.



This suggests that Fare might be directly related with the probability of survival.

## Approximating Missing Ages

There were 177 people in the original data whose ages were missing.
To approximate the missing ages, the ages of all the people were grouped together on the basis of their honorifics (Mr., Mrs., Miss, Dr., etc.) and their median was calculated.
Accordingly, the people with missing ages were assigned the median of the honorific age corresponding to them.
The dataset with the complete values for ages can be found in the following link:
https://github.com/ketangupta96/TitanicDataSet

## Approximating Missing Ports of Embarkation

There were 2 people in the original data whose port of embarkation was missing.
To calculate the missing ports of embarkation, the fares of ports of embarkation for each Pclass were grouped together (i.e. 3 groups for each port) and the medians of the fares were calculated.
Now for the people with the missing port values, their Pclass and Fare was noted.
They were assigned the port for which the difference between the Fare and the median fare of the corresponding group was the minimum.
The dataset with the complete values for ports of embarkation can be found in the following link:
https://github.com/ketangupta96/TitanicDataSet