# Report on Titanic Data Analysis

**1. Data analysis:**

The data set provided  represents the passengers on the Titanic, and some information about them. Our data consists from the following variables:

- PassengerId - A numerical id assigned to each passenger
- Survived - Whether the passenger survived (1), or didn't (0). This is going to be the dependent variable of our study
- Pclass - The class the passenger was in - first class (1), second class (2), or third class (3). Pclass is going to be one of the independent variables in our study
- Name - the name of the passenger
- Sex - The gender of the passenger - male or female. Sex is going to be one of the independent variables in our study
- Age - The age of the passenger. Fractional. Age (of age groups) is going to be one of the independent variables in our study
- SibSp - The number of siblings and spouses the passenger had on board
- Parch - The number of parents and children the passenger had on board
- Ticket - The ticket number of the passenger
- Fare - How much the passenger paid for the ticker
- Cabin - Which cabin the passenger was in
- Embarked - Where the passenger boarded the Titanic

The data provided has many missing details.
For the analysis, 'Name' and 'Passenger_id' are not considered.
For missing data in 'Age' column, it is filled with '0' just to keep track it.
Following columns provides can play a vital role in predicting the reasons of survival.
Factors:
-Age
-Gender
-Pclass
-Cabin
Code for the analysis is attached with name: titanic_final.py

**2. Following questions can be answered from the data available:**

i) Number of female survivals vs Number of male survival

This result can be seen in the code results which is tabulated below:

| | **Total** | **Survived** | **Dead** | **Survival %** |
|---|---|---|---|---|
| **Male** | 577 | 109 | 468 | 18.89% |
| **Female** | 314 | 233 | 81 | 74.20% |

Here, we can see the percentage of female survival is way higher than male survival. It might the reason that during any accidents females and children are saved first.

ii) Percentage of Children survival vs Adult

Here children are considered anyone who is below 20 years. For this the missing age data set are not taken.

| | **Total** | **Survived** | **Dead** | **Survival %** |
|---|---|---|---|---|
| **Children** | 164 | 79 | 85 | 48.17% |
| **Adult** | 550 | 211 | 339 | 38.36% |

Percentage of children surviving is more than adult. And it is obvious that children must have came with parents and parents will prefer to save their children first.

iii) Survival based on port of Embarkment

| | **Total** | **Survived** | **Dead** | **Survival %** |
|---|---|---|---|---|
| **S**(Southampton) | 644 | 217 | 427 | 33.69% |
| **C**(Cherbourg) | 168 | 93 | 75 | 55.35% |
| **Q**(Queenstown) | 77 | 30 | 47 | 38.96% |

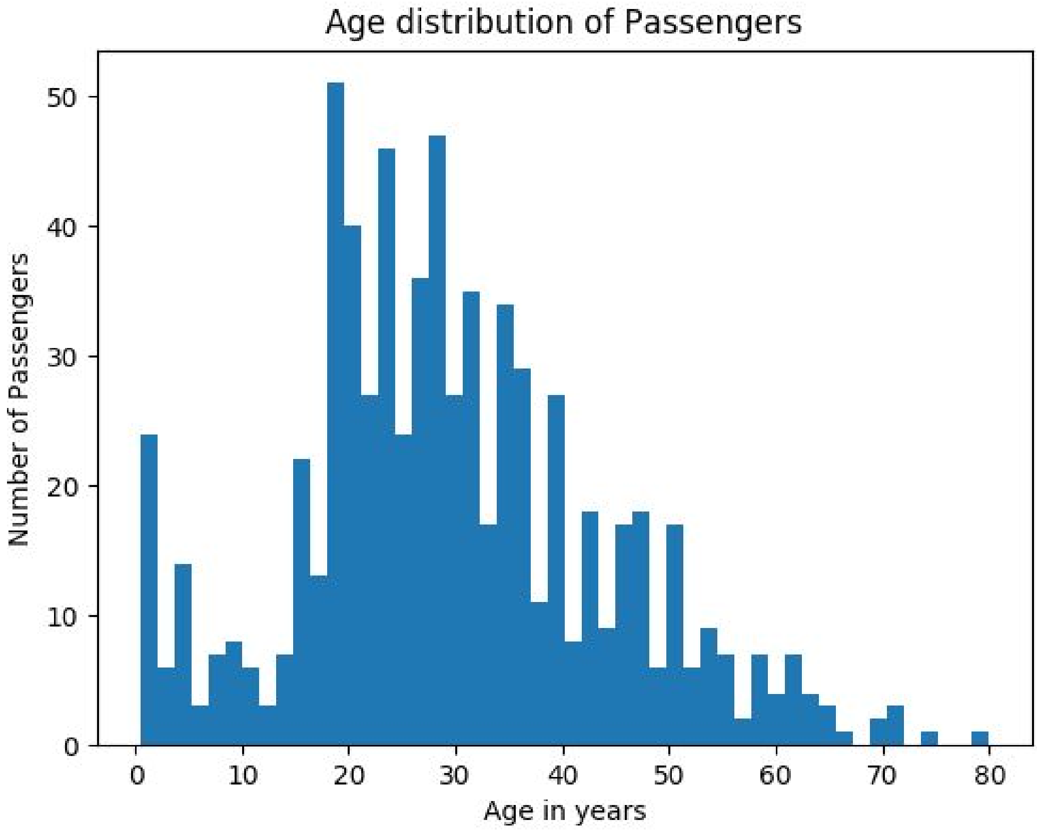
As we can see from the above results Passengers with different embarkation survival percentage. We can say that passengers from ‘Cherbourg’ travelled just to ‘Queenstown’ hence their survival percentage is high.

iv) Survival based on Passenger Class

| | **Total** | **Survived** | **Dead** | **Survival %** |
|---|---|---|---|---|
| **1st Class** | 216 | 136 | 80 | 62.96% |
| **2nd Class** | 184 | 87 | 97 | 47.28% |
| **3rd Class** | 491 | 119 | 372 | 24.23% |

Most of the Class 1 passenger survived. There can be many possibilities, one of it can be they are rich people they had the resources or they might be having advantage during the process of sinking of the ship.

v) Age distribution of the passengers



From this plot we can see the age distribution and it can be seen that average population are is around 30.

**3. Missing Data Analysis:**

i) Missing age data survival percentage

| | **Total** | **Survived** | **Dead** | **Survival %** |
|---|---|---|---|---|
| **Missing Age** | 177 | 52 | 125 | 29.37% |

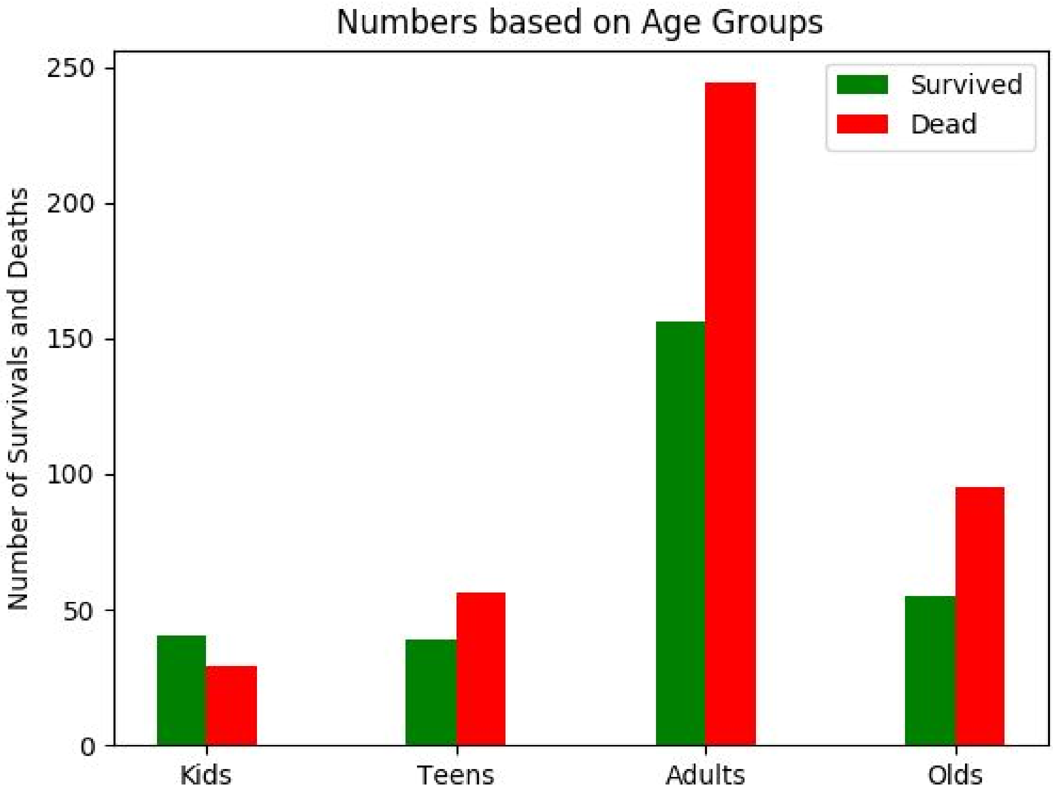
As it can be seen that most of the missing data refers to dead people and that is why it is mostly missing. The data for ‘Age’ column is filtered for the analysis. As the ages were categorised in age groups hence, the missing data is filled with ‘0’.

ii) The columns ‘Cabin’, ‘Parch’, and ‘SibSp’ are dropped in the analysis:

Mostly ‘Cabin’ column is empty that can be because one cabin is allotted to many hence it is not sure how the data is recorded.
The ‘Parch’ and ‘SibSp’ columns are mostly zero hence dropped these two for making the analysis focused on the other important factors.

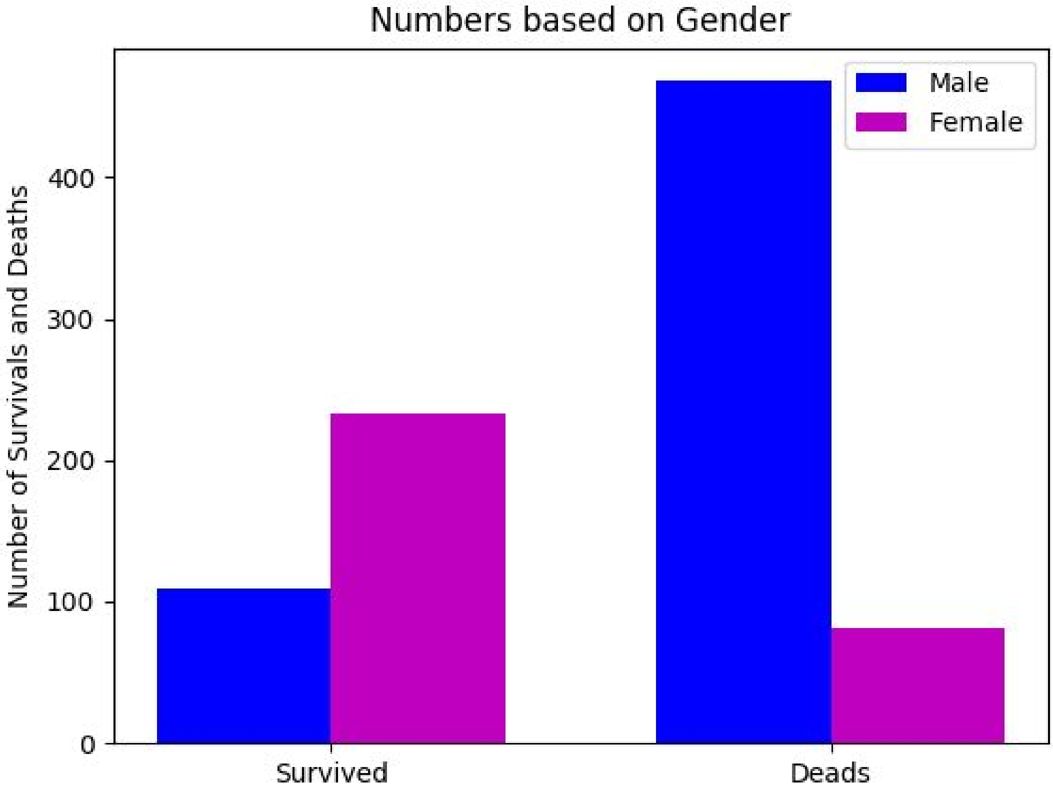
**4. Data plot of survivals and deaths for Age groups, Gender, and P class:**

i) Plot based on age groups. In the following plot, the ages are grouped into four groups. First, ‘kids’ where age is less than 12 years, second ‘teens’ where age is between 13 to 19 years, third ‘adults’ where age is between 20 to 40 years, and last ‘olds’ where age is more than 40 years.



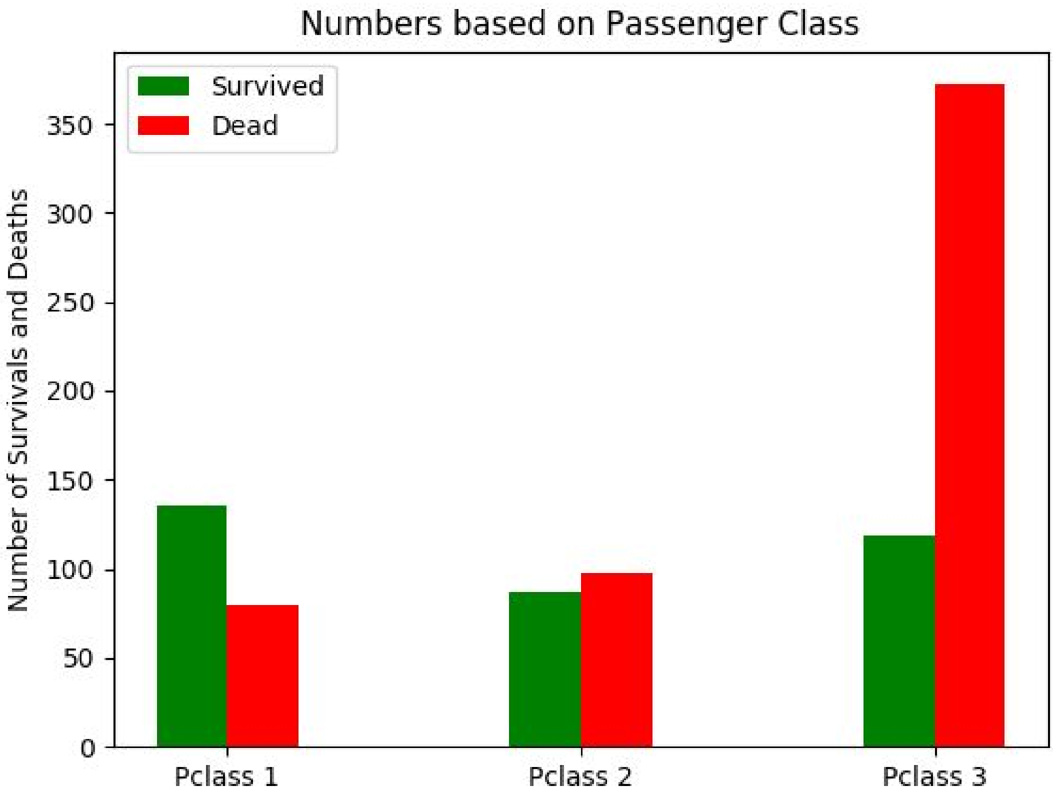
From the plot, we can see the number of kids surviving is more than the deads. It might be because kids are first saved. Adults are dead in high number because they must be helping kids and old people for the rescue.

ii) Plot based on Gender.



It is clearly evident from the plot that large fraction of female passenger survived over the male.

iii) Plot based on Passenger Class.



It can be seen that passenger class 3 are mostly dead. From this, we can say that passenger class 3 must be in the lower section of the ship which got sunk first hence a large portion of class 3 are dead. Where as in Class 1 that should be on the Top of the ship from where rescue is easy.

**5. Reasons for survival conclusion based on above factors:**

From the above analysis we cannot say anything concrete but still, we can correlate many factors which can tell what made passenger survive the disaster.

i) Based on Age Group, it can be observed that the survival rate is higher in case the passenger falls in age group 1, i.e. kids(ages 0 to 12), survival rate for age group 3(ages 20 to 40 years) is very less.

ii) Based on Passenger Class, it can be seen that if a passenger falls in the category of passenger class 1 they are most probable to survive the disaster.

iii) Based on Gender, it is clearly visible that the survival percentage of female passenger is way too high than male passenger.

iv) Based on Port of Embarkation, passenger with embarkation ‘Cherbourg’ are most probable to surviving the disaster.

## 6. Limitation of data set:

The dataset is not complete. There is much missing information like age, Cabin etc. We cannot be sure of our predictions but still we can come out with some result which can be useful enough to predict the factors of survival.