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 One (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

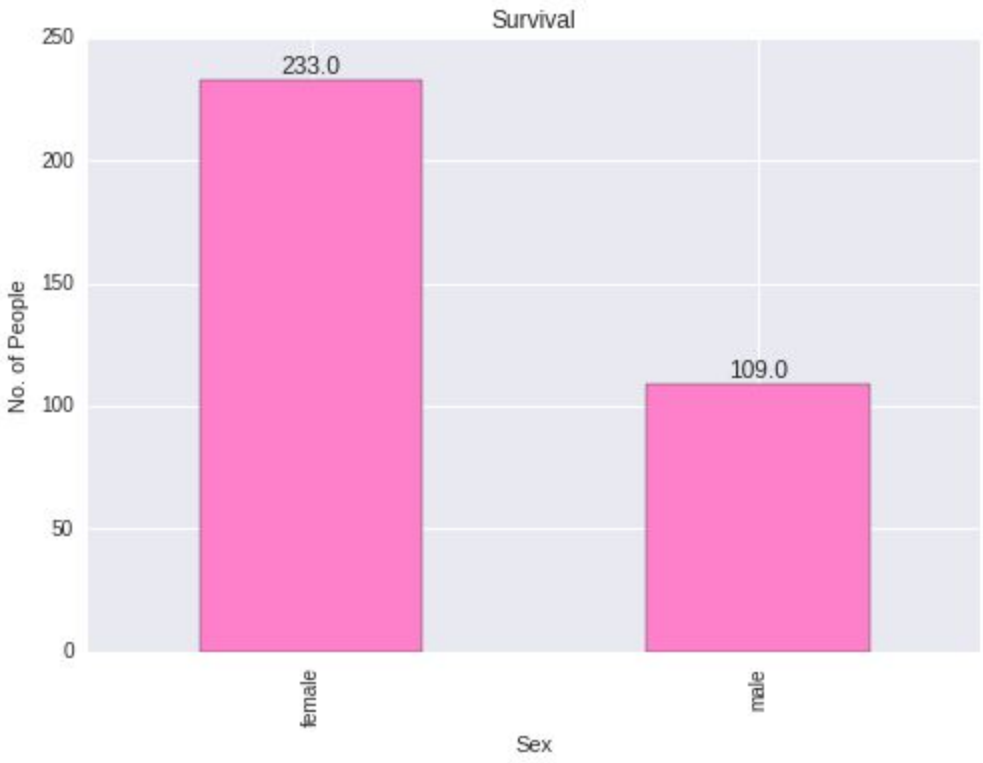
Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

Analysing the Sex component

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

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

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



From the above table and graph, it is very clear that probability of survival of females was much higher than that of men.

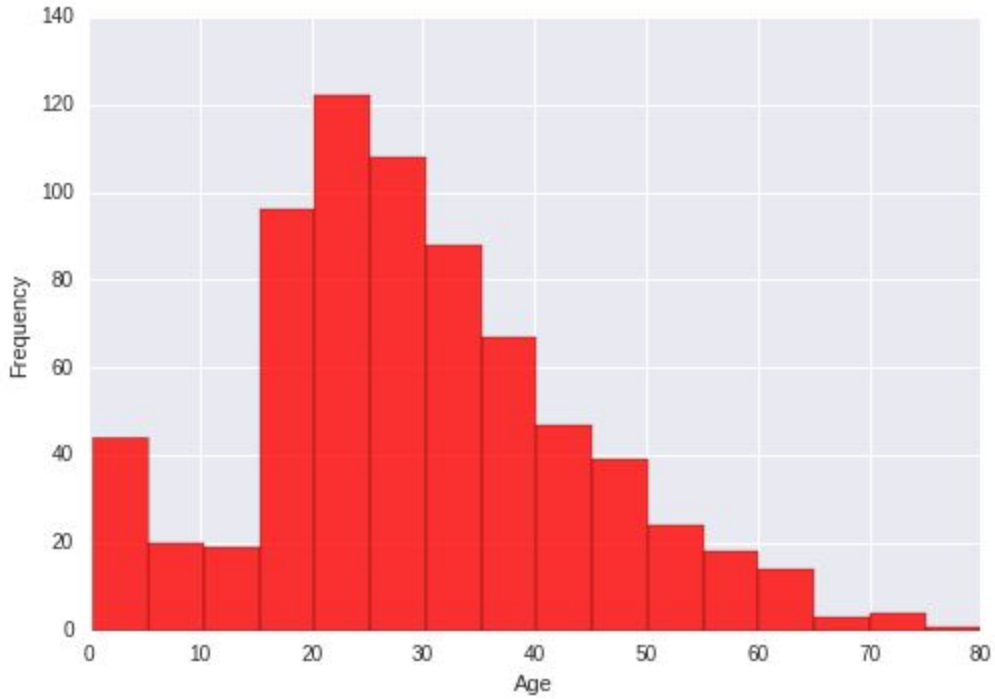
Analysing the Age component

In the data given to us, the ages of 177 people are 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	0.70
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

From the above table, we see that children in the age group of 0-5 years had a very high probability of survival compared to other age groups.

Following histogram gives us the age distribution of the people (excluding the missing ages). We can see that the distribution is a little rightward skewed, i.e., maximum number of people on the Titanic were in the age group 20-30 years.



Analysing Port of Embarkation Data

Following table contains information about different ports of embarkation.

Embarked	Mean Fare	Number of People	Survived	Proportion Survived
C	59.95	168	93	0.55
Q	13.28	77	30	0.39
S	27.08	644	217	0.34

We can see from the table that people who had C as their port of embarkation have a significantly greater probability of survival as compared to people from ports of embarkation Q and S.

Also the Mean Fare of port C is significantly higher than the mean fares of ports Q and S.

This suggests that people who had paid more Fare might have had a greater 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/Titanic-Data-Set

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/Titanic-Data-Set