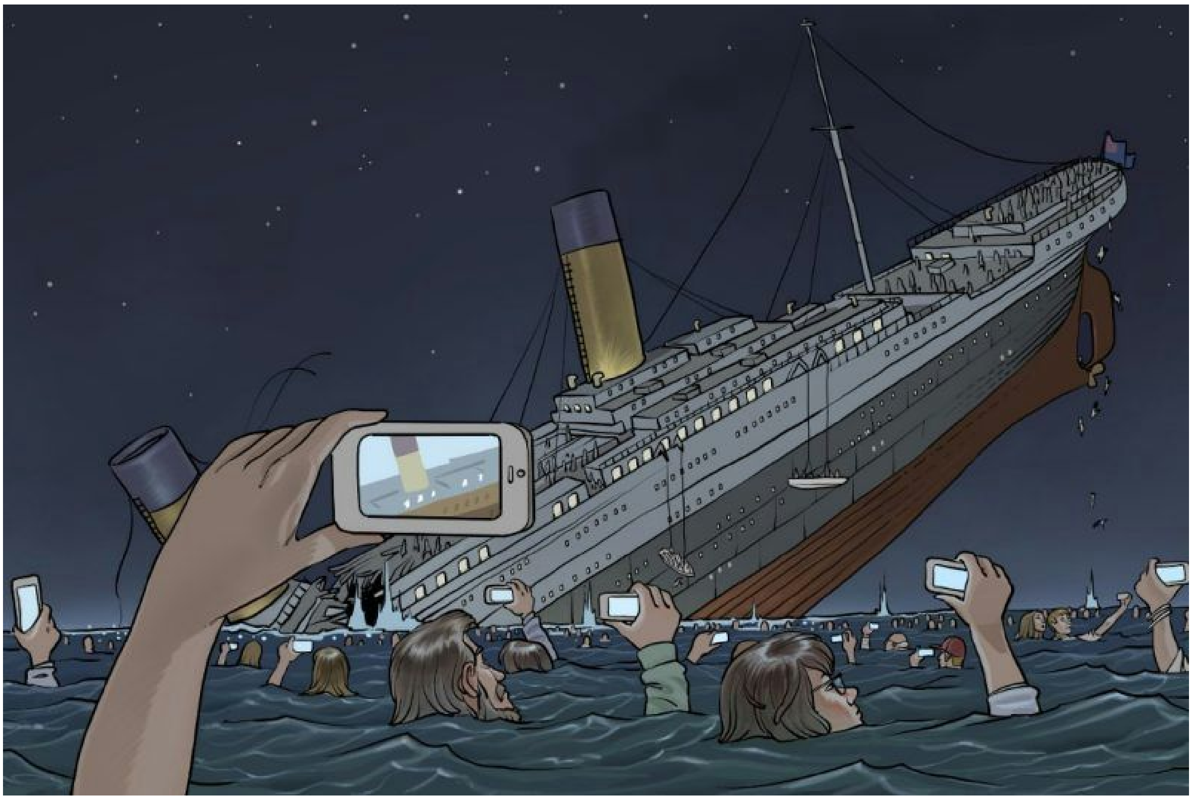
Titanic Data Analysis Report



Introduction:
The data set I chose to analyse is the Titanic Data from Kaggle website, following are the questions I am going to investigate:

1. What factors made people more likely to survive? Such as:
   - Gender (Sex)
   - Passenger Class (Pclass)
   - Age
   - Embarked

2. If I was on Titanic, how should I purchase the ticket in order to have higher survival rate? With following condition:
   - Male
   - Age 25-30
   - no children/no partner

Data Wrangling Phase:
<u>Create New Column</u>
Since the age column has been filled up with various numbers, it is not efficient to group the age column by itself. The function 'group_age_rank' is used to categorise the age numbers into different age rank/band and a new column [Age_rank] has also been created after applying the function to the original data set.

Missing Values

In this dataset, there are missing values in both Age and Cabin columns. Since we are more interesting in how the age factor would impact on the survivor rate, we are going to ignore the missing value in Cabin column.

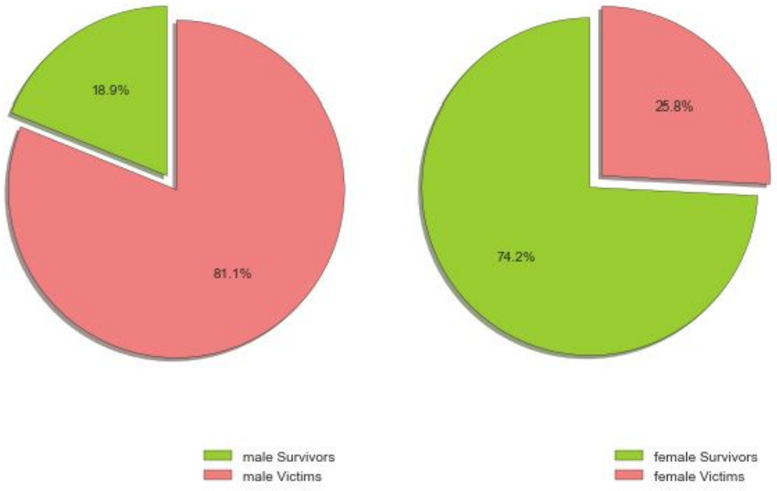
For the missing value in Age group, the function ‘group_age_rank’ is used to categorise them into a ‘nan’ age rank. The reason I didn’t exclude this ‘nan’ group in the further analysis is because I would like to see how many people with unknown age have survived, what’s the survival rate and where are these group of people came from (i.e. why they are marked with unknown age)

Findings:

According in the movie ‘Titanic’, women and children were first aboard the lifeboats when the Titanic Struck an iceberg and sank to the bottom of the North Atlantic.
This gives the first clue that the female and children might have higher survival rate (SR) than male passengers. In order to investigate these factors, the data were grouped by ‘Sex’ and ‘Age’.

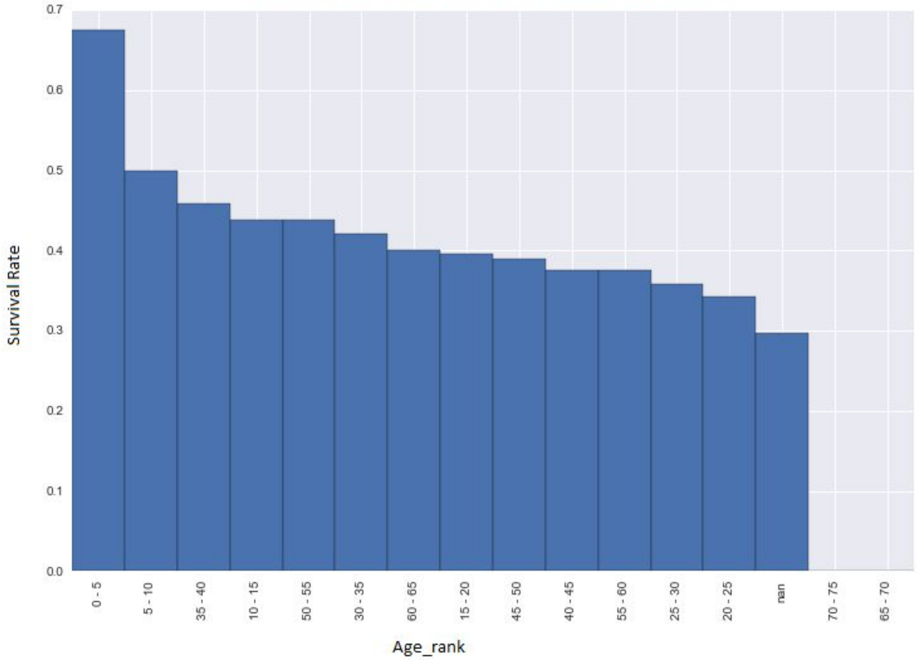
As calculated in script SC1, there are 314 female and 577 male passengers in the data set. Although the total male passengers are almost double the female passengers , female has 74.2% of SR and male has only 18.9% of SR (See figure 1). That was due to the women had been given the priority to board the lifeboats.

Figure 1, Survivors Ratio by Gender



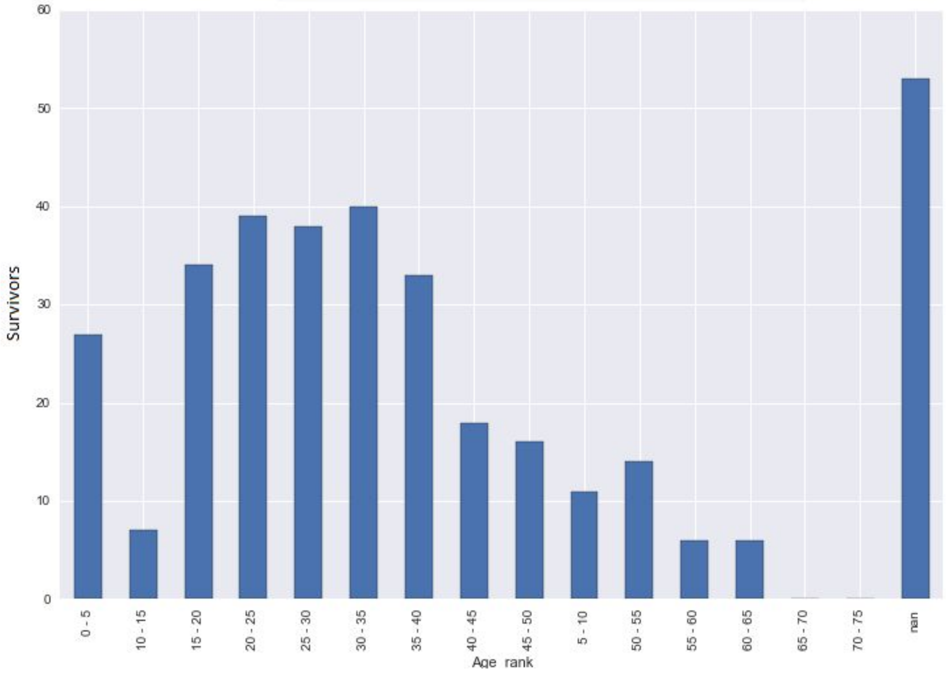
In figure 5, the bar chart shows the survivor who is in age 0 - 5 group has highest SR which is about 0.68, then the SR drops down to around 0.3~0.5 for the rest age ranges. Despite that, the group of age 5 - 15 still has high SR. This proved that children had also been given the priority to board the lifeboats.

Figure 5, Survivors rate at different age ranges



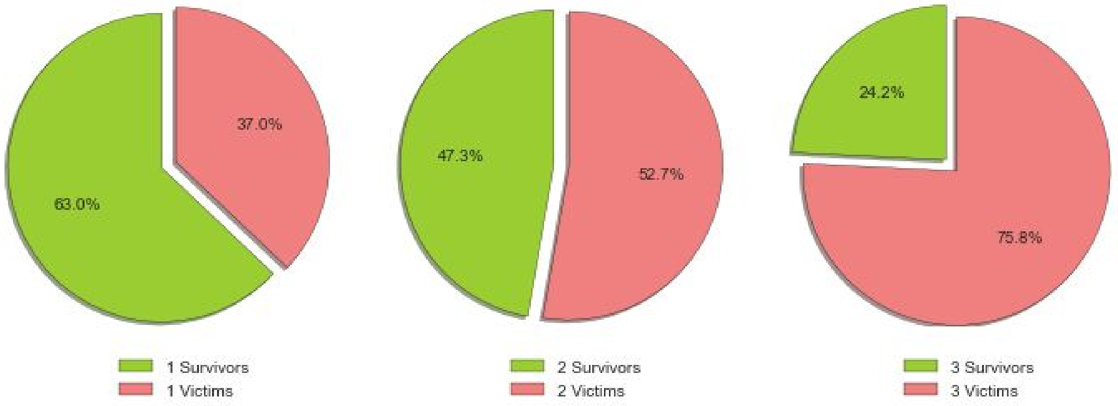
In figure 4, the survivors with unknown ages have been categorised as ‘nan’ and this group of people has the highest survived number in all age groups. However, by comparing to figure 5, its SR is sitting at the bottom. This is because 81.6 % of the victims with unknown age are from class 3 (See SC5) and people in 3rd class were primarily immigrants, hence there are no record of their birthday.

Figure 4, Survivors number at different age range



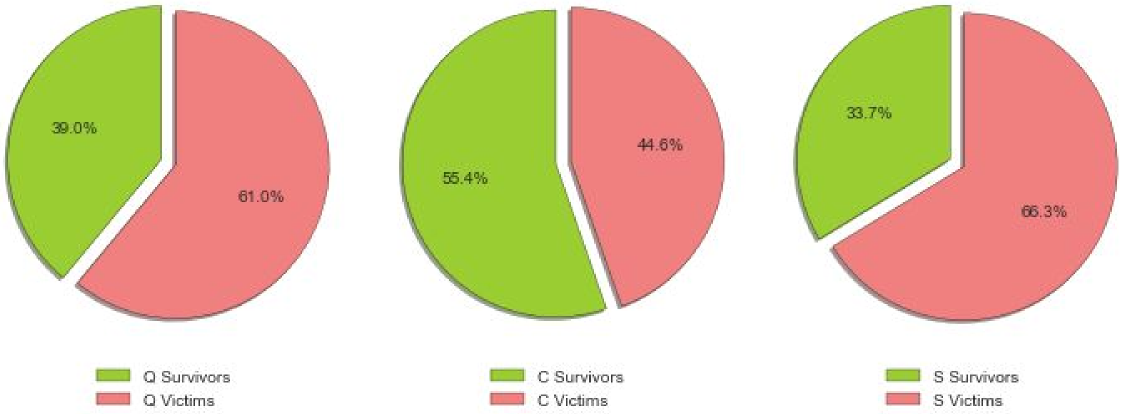
In term of SR in passengers class, the figure 2 shows that 1st class has the highest SR which is 63%, 2nd class has about 47.3% SR and the 3rd class only has 24.2% SR. This is because the the lifeboats were located at the upper deck where was very close to the 1st class cabin. As shown in SC6, 91 female survivors from 1st class, 70 female survivors from 2nd class and 72 female survivors from 3rd class.

Figure 2, Survivors Ratio by Passenage Class



In the record, Titanic departed from three ports, such as Cherbourg, Queenstown and Southampton. With the analysis in figure 3, it shows the passengers from Cherbourg has higher chance to survive since it has about 55.4% of SR, This might indicate the passengers from Cherbourg are either richer or female dominated. However, this shouldn't be considered as a significant factor to impact on the SR.

Figure 3, Survivors Ratio by boarding port



Imaging I was about to purchase a Titanic ticket, How should I purchase? Following is my personal condition:
I am a guy hence I would only have 18.9% to survive.
My age is in between 25-30 years old
I am going alone, hence no other partners with me
I want to survive
After a series of filtering on data set, the SC7 script shows there would be only 5 tickets from 1st class and 9 tickets from 3rd class available for me. 6 out of 9 tickets could be purchased from Southampton. Therefore, I should buy the 3rd class tickets and embarked at Southampton.

Conclusions:
This analysis explored the survival rate with different groups of passengers' gender, age and passenger class on the Titanic shipwreck. The results indicate female and children are most lightly to survive and passengers from 1st class also have the higher chance to survive.

The data set has limitations for the analysis, such as the missing value in the age group. It turns out the unknown age group has the most survivors in the 'age_rank' category (as shown in figure 4) and most of the unknown age passengers were allocated to 3rd class cabin. In addition, the new column could be added to determine if the passenger got on the lifeboat.

References:

1. Data Descriptions, Titanic: Machine Learning from Disaster, kaggle, https://www.kaggle.com/c/titanic/data
2. The Truth Behind "Women and Children First", Titanic, The Artifact Exhibition, http://www.premierexhibitions.com/exhibitions/3/3/titanic-artifact-exhibition/blog/truth-behind-women-and-children-first
3. Third Class Life on the Titanic, ALookThruRime, https://alookthrutime.wordpress.com/2012/04/13/third-class-life-on-the-titanic/