I have analyzed the Titanic dataset drawn from Kaggle website.

1. Questioning Phase

The question I would ask broadly would be:

"What factors made people more likely to survive?"

I initially thought that the following factors would affect the survival rate of passengers.

- Pclass
- Gender
- Age
- Fare
- Cabin

When observing each factor, I will narrow down the question.

2. Data Acquisition Phase

In this phase, we will acquire data to be analyzed. We have the csv file so we will simply import the data. We would like the current data to be in 2D data structure. Therefore, I believe that Pandas DataFrame data structure is appropriate.

3. Data Cleaning Phase

When cleaning data, it is important to see whether there is any problem with the given data. The criteria for judging the quality of the data would be:

- data type error (e.g. date in string format)
- missing data
- user entry errors
- different schemas (e.g. age in Western and Eastern calculation)
- no unique identifier(s)
- unnecessary data

In our data, 'Passenger' variable is the unique identifier. Furthermore, there are some missing data however it is not wise to fill that with anything. As such, when parsing data, I will make it 'None'. I believe that some data in this excel file are redundant. Thus, I will drop the data including name and ticket. At first, I was not quite sure about the usability of 'PassengerId' so I left it there. However, during the data exploration phase, I realised that this variable is also redundant so I have decided to drop it.

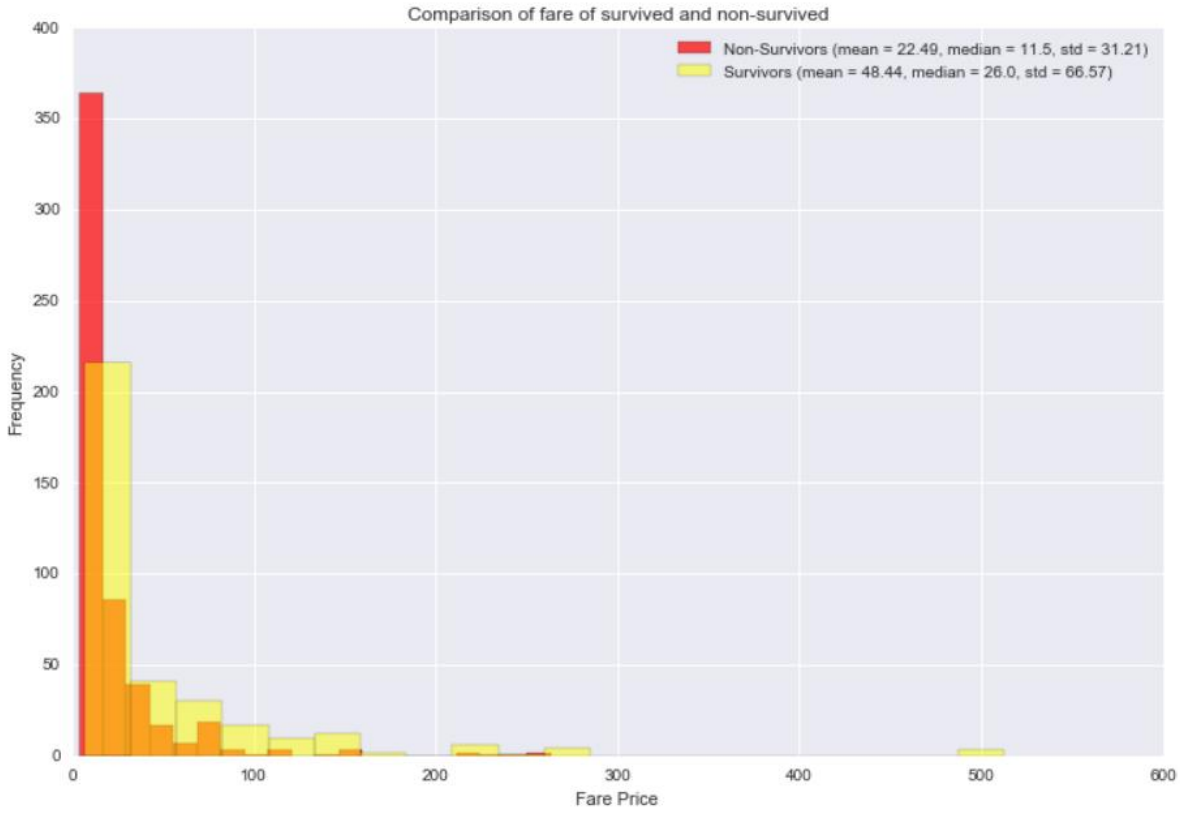
One intersting observation can be drawn from 'Fare' variable as there are some rows where the value for that variable is 0. I first speculated that those who have 'Fare' of 0 would be babies (i.e. as babies normally don't get charged!). However, in this case, there is a person, aged 38, who has the fare of 0. As the value of 0 does not make sense to me, to make it fair, I replaced the value to the median of the fare so that the replaced value does not heavily affect the overall analysis.

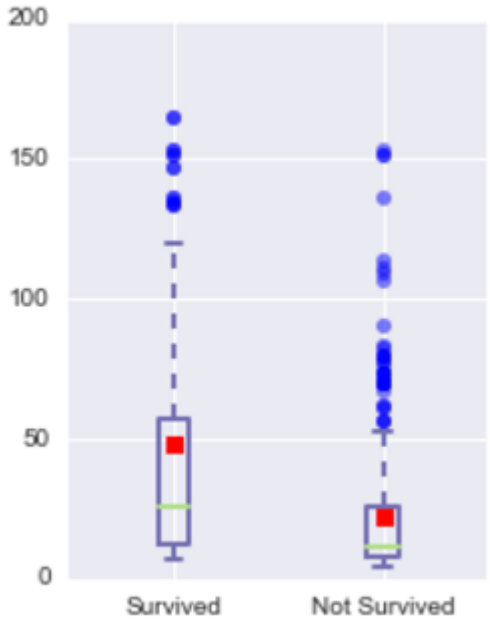
Regarding the 'Age' variable, there are certainly massive amount of missing data (about 177) and there is no way to deal with the missing data. Therefore, I have decided to leave them with NaN values.

4. Data Exploration, Communication and Conclusion Phase

Fare Data

I believe that those who survived must have paid the higher fare on average than those who did not survive (as a premium they must have prioritised those who paid the most).

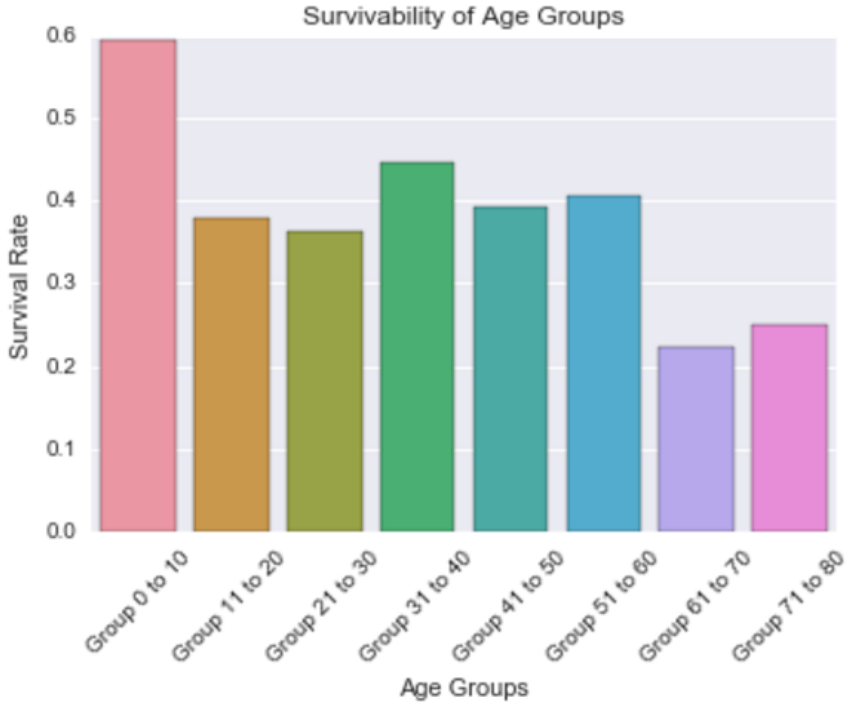




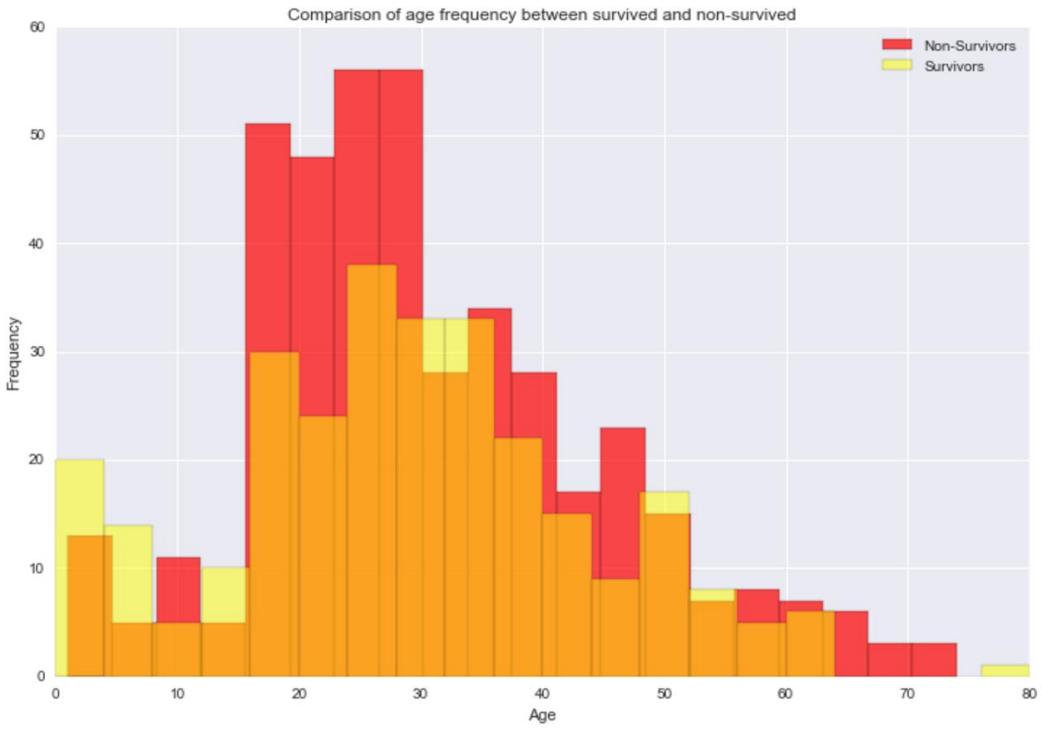
Certainly the majority of non-survived paid less than 100 and certainly nothing beyond 300. However, most survivors also paid less than 200 and surely the standard deviation for survivors are significantly higher than that of non-survivors, which tells that the distribution of fare is widely spread from the mean. Although the mean is twice as high as that of non-survivors, it is difficult to find correlation by looking at the analysis given.

Age Data

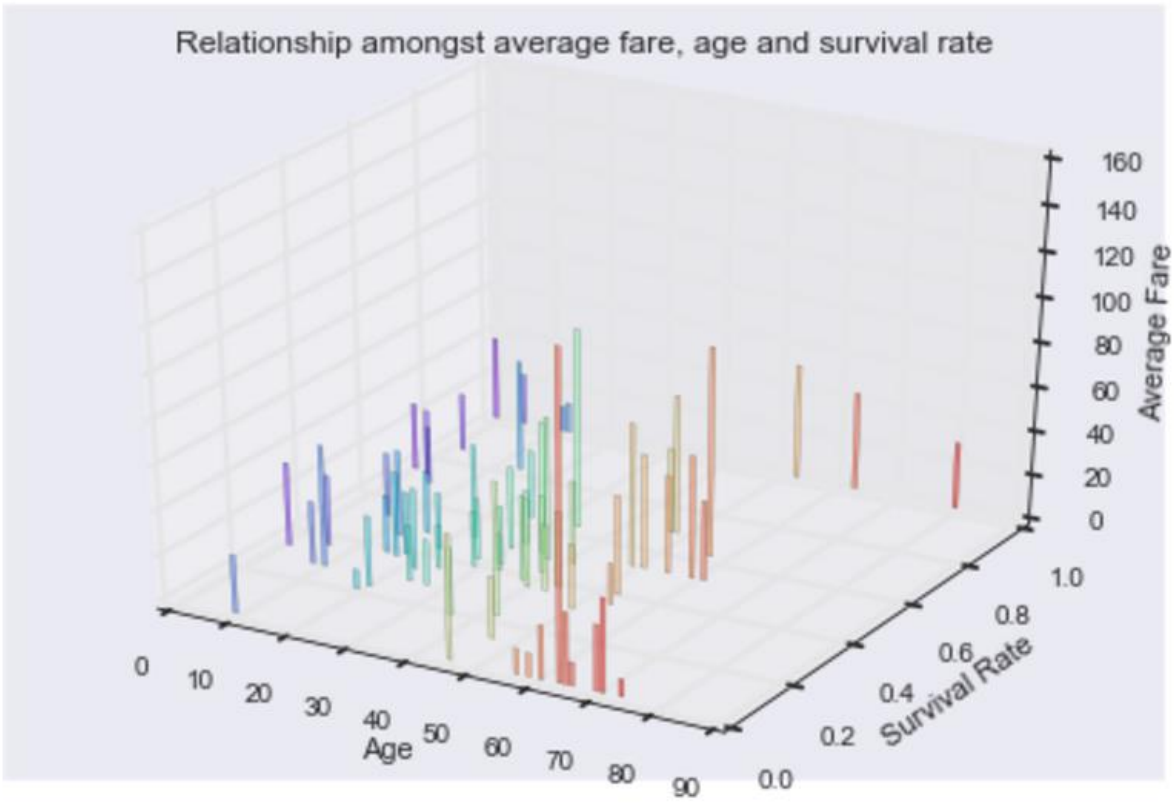
I would like to get the survival rate of each age group to see whether there is any correlation between age group and survival rate.



People between age 0 and 10 showed very high rate of survival (0.6). The age group that showed the lowest rate of survival was between 61 and 70 (slightly higher than 0.2). Furthermore, I would like to compare the age distribution of survived group and non-survived group.



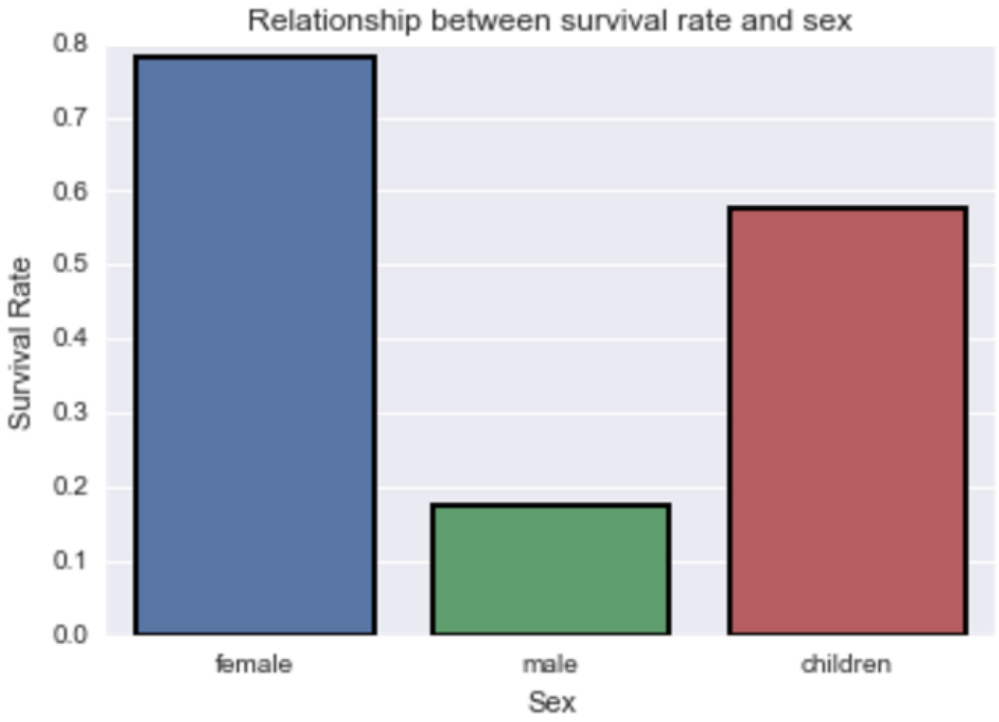
The distribution of age group is very similar. I am starting to realise that there may be some correlation between survival rate, age and fare price. So I want to see the correlation between Fare price, Age and Survival Rate.



There is no obvious correlation between these three features.

Sex Data

We need to differentiate children with adults. I define child to be those 14 years or under. Therefore, now the sex will tentatively have children, male and female.



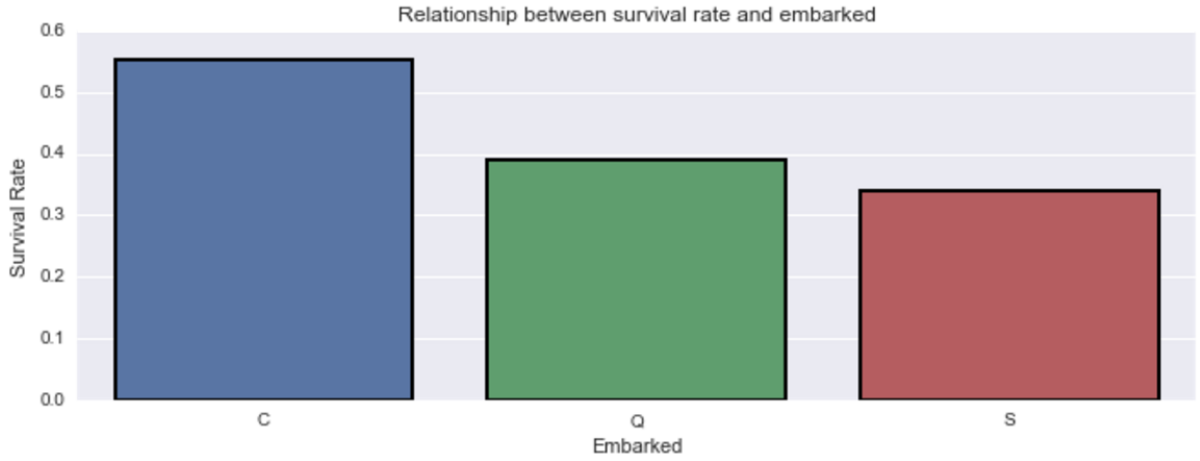
From this analysis, we can conclude that more than half of all children survived (in fact almost 80% of women survived). However, only 19% of men could survive. Although we cannot find the causation for this observation, we can definitely see the correlation between sex and survival rate.

Cabin Data

We have 687 missing data for cabin details. With this variable, I believe that it would be pretty difficult to conduct the exploration. It is better not to explore with this data.

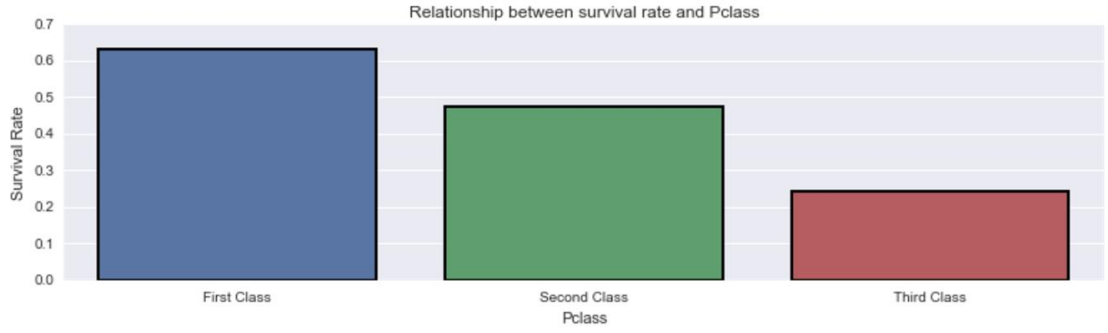
Embarked Data

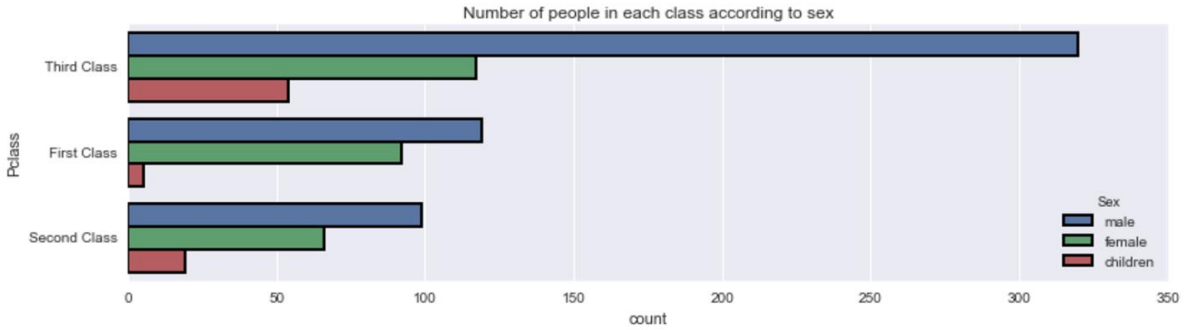
We need to explore whether the embarked and the survival rate is correlated or not. Before doing this, we need to determine whether there are any missing data.
As there are only 2 missing data, we will just fill them with most frequent value in this series (i.e. 'S').



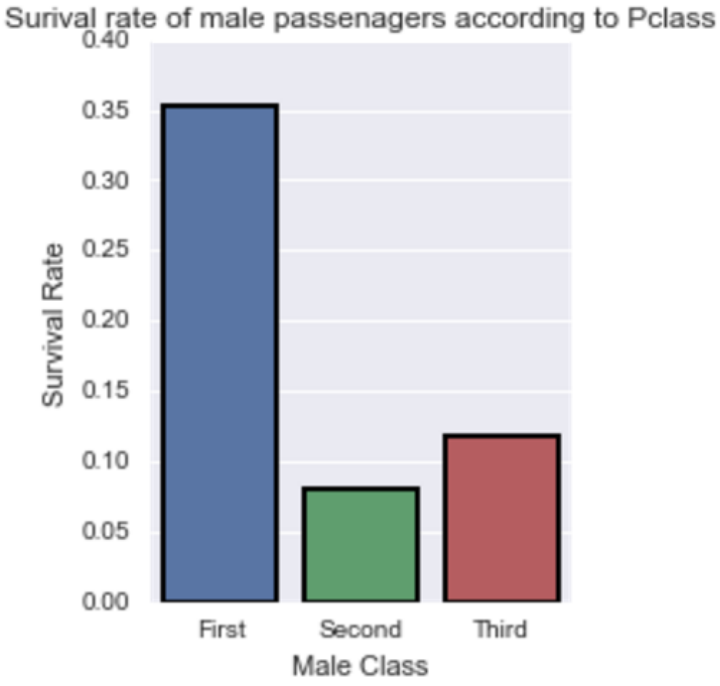
This analysis tells us that more than half of the people who survived have embarked data 'c' (and the lowest being 's'). There is a weak correlation between embarked and survival rate.

Pclass (Passenger Class) Data





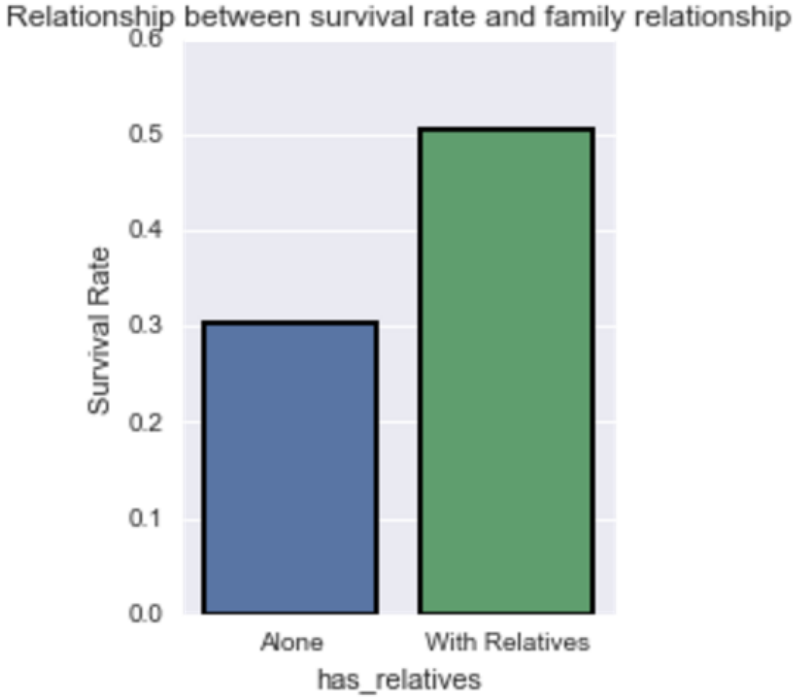
We can see that those who were classified as first class survived more than those who were classified as a third class passenger. One interesting thing to observe is that we have more male than female in the first class. However, we have seen previously that more females have survived than males. This probably suggests that within males who failed to survive, probably most of them belonged to third class.



We can clearly see that the survival rate for first class male passengers are significantly higher than that of second and third class. Therefore, there are correlation between Pclass and survival rate and again this does not tell us any causation.

Exploring with Family Relationship Data (SibSp and Parch)

Any specific relationship does not really matter in this case. What we want to see is whether being alone or being together with family and the survival rate are correlated or not.



With this data, we can see that passengers who travelled with families has higher survival rate than those who travelled alone. It is not to imply that this factor caused people to survive but surely we can see the correlation.

Limitations regarding the current findings

The statistical findings presented in this report have few limitations. This report only used statistical calculations and did not conduct any statistical test that would draw rigorous conclusions. For example, currently, the data only contains information from 891 of the 2224 passengers and crew on board the Titanic. As such, the current analysis does not reflect the true population's tendencies (e.g. survival rate). To draw more accurate statistical conclusion, I would had to come up with reasonable null hypothesis (and even alternative hypothesis) and reject it. Therefore, there is clearly a room to improve the analysis and this rigorous statistical test can be performed in future.

Furthermore, this analysis did not use cabin data for further analysis. As mentioned previously, I have decided not to use the cabin data as there are 687 missing data for that variable. As almost 77% of data is missing, I decided to not to use it. However, if there were sufficient data available for the analysis, this variable could have been really useful in answering the question: 'Was the location or positioning of cabin related to the survival rate?' and in overall better answering the main question "What factors made people more likely to survive?"

Regarding the unused data, I still do not have any idea how to use them. For example, the ticket data is pretty random and the format is different as well (i.e. it does not conform to one schema). I could have used this data if there was further information regarding the 'pattern' of this data.

If possible, it might been better if these information were available:

- ***Each passenger's education level:*** It may be possible that passenger's education level is correlated to the survival rate.
- ***Each passenger's ethnic group / race (or country of origin):*** It may be possible that passenger's ethnic group may be related to the survival rate.

Overall Conclusion (Most interesting feature)

I have explored 6 features contained in this dataset. Amongst them, Pclass and Sex features showed the most interesting insights. Females and children had a high rate of survival (approx. 0.8 and 0.6 respectively). Furthermore, most male passengers who survived belonged to the first class (more than 300) and their survival rate was higher than second and third class male passengers (about 0.35). Therefore, male passengers who bought the second or third class ticket had a pretty low rate of survival (about 0.08 and 0.11). This finding was very interesting for me at least. What was more intriguing was the fact that although those who bought the first class ticket had a relatively high rate of survival (more than 0.6), the fare price did not tell us much about its correlation with the survival rate. Again, a correlation does not show any causation.

References

- http://stanford.edu/~mwaskom/software/seaborn/index.html (Seaborn)
- https://www.kaggle.com/c/titanic/data (Titanic Data Explanation)
- http://matplotlib.org/users/pyplot_tutorial.html (Pyplot Tutorial)
- https://www.youtube.com/watch?v=afITiFR6vfw (Tutorial on Subplot - Really Concise and Easy-to-Understand)
- http://stackoverflow.com/questions/15315452/selecting-with-complex-criteria-from-pandas-dataframe (Good Answer regarding selecting with complex criteria from pandas dataframe)
- http://blog.bharatbhole.com/creating-boxplots-with-matplotlib/ (Boxplot)
- http://stackoverflow.com/questions/5159065/need-to-add-space-between-subplots-for-x-axis-label-maybe-remove-labelling-of-a (Adding margin between subplots)
- Udacity Reviews