Titanic Data Analysis Project

The following report is for the analysis of data on passengers of the Titanic.

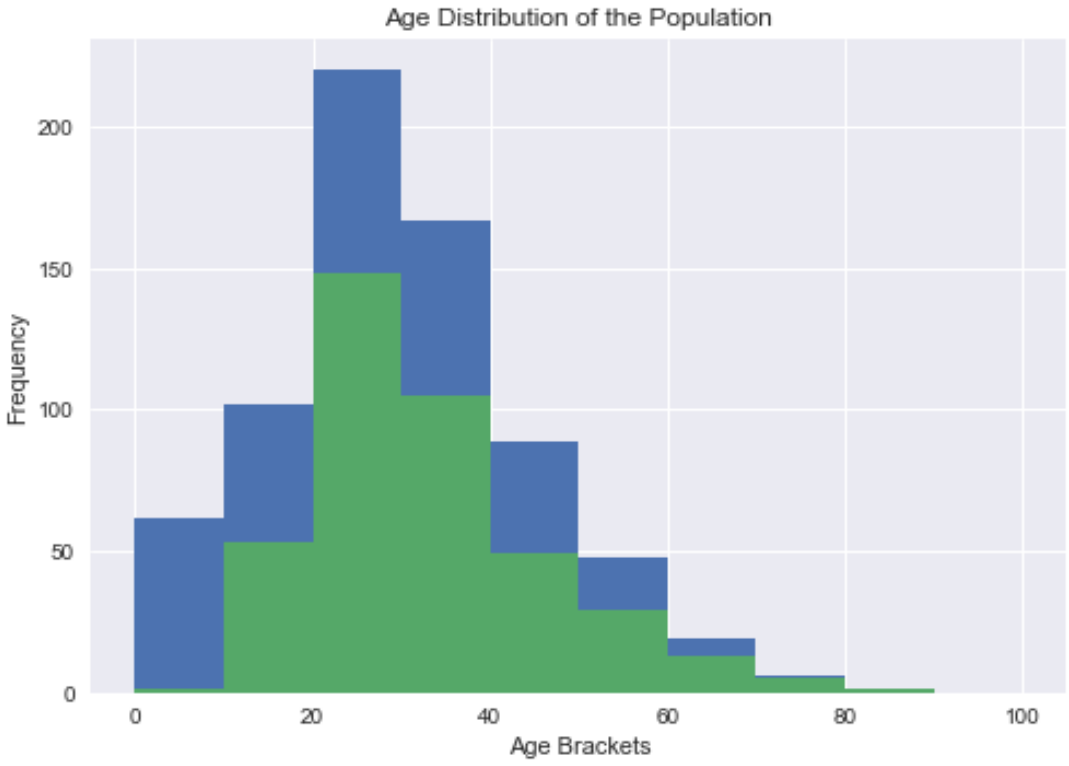
INTRODUCTION

The data had several missing entries for 'Age' and 'Cabin' variables. On reading the CSV file, the missing values were replaced with NaNs. Since the 'Cabin' variable was not relevant for this analysis, I replaced the NaNs with missing values for the Cabin data. However, some of the analysis uses the 'Age' data and some of it does not. Therefore, to avoid loss of any useful information, I captured the data with 'Age' missing in a separate list and used the restricted list or full data list based on whether the question requires 'Age' information or not. The following questions are posed and answered based on the data:

1) What is the age distribution of the passengers on the Titanic?
2) What is the survival rate of passengers on board the Titanic?

ANALYSIS

1. What is the age distribution of the passengers on the Titanic?



Mean Age of passengers:	29.7
Median Age of passengers:	28.0
Mean Age of lone passengers:	32.2
Median Age of lone passengers:	29.5

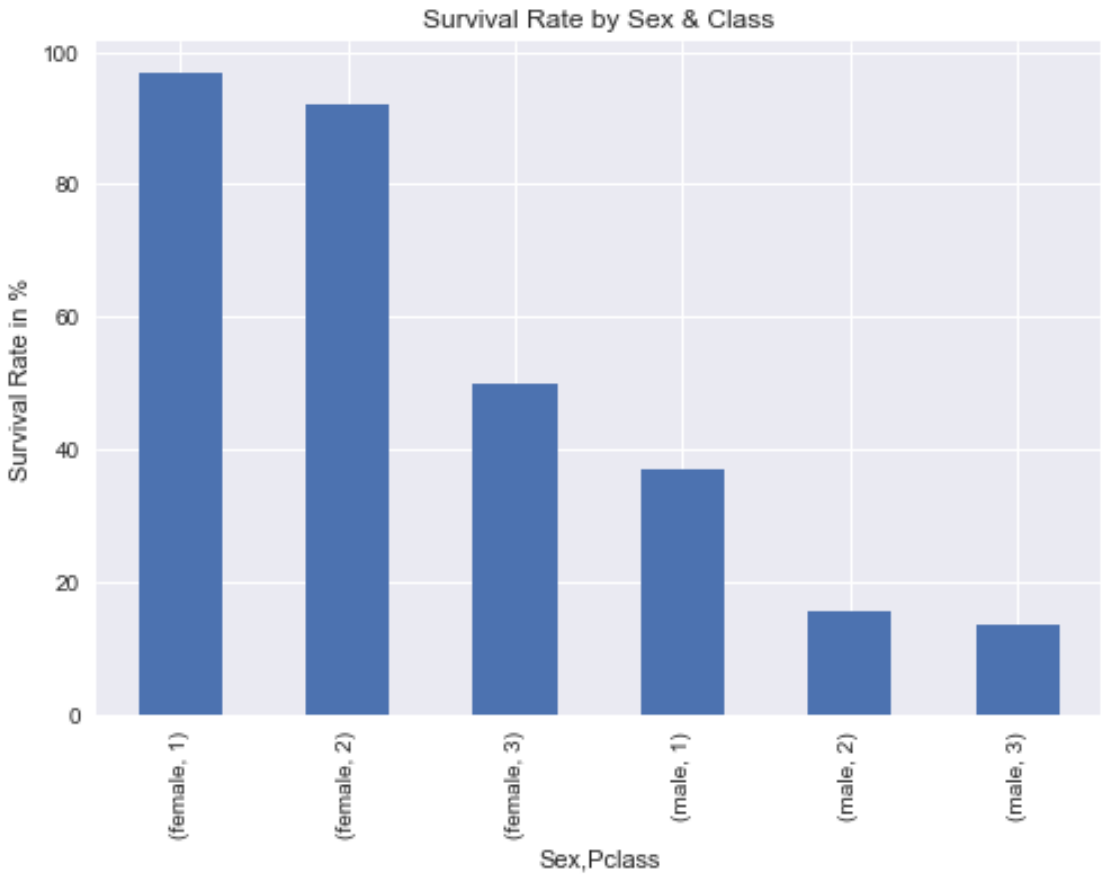
The chart above shows the distribution of passengers by age group. The following important points are to be noted here:

a) The Blue histogram represents the age distribution of the overall population. As can be seen from the histogram, there are several kids and several elderly people on board the Titanic.

b) The Green histogram represents the age distribution of those people who came alone on board the titanic, i.e. those without any siblings/spouses/parents/children

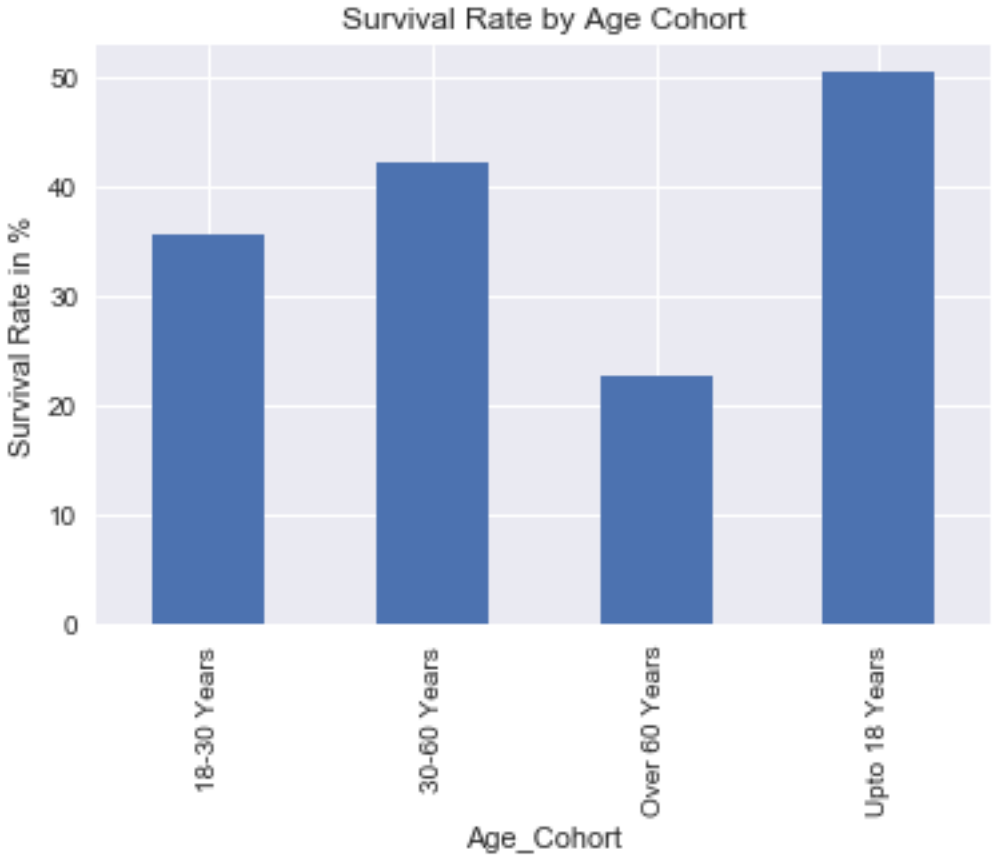
c) Interestingly even among the single people on board, there appear to be a few children as well a few very elderly people. This could possibly be explained by the fact that the original data classifies children coming on board with a 'nanny' with Parch=0, i.e they are classified as coming without parents. Additionally, mistresses/Fiancé's were also ignored in a similar manner

2. What is the survival rate of the passengers on board the Titanic?



The bar chart above shows the survival rate in % based on gender and ticket class. The following points can be observed from the chart:

a) Survival rates among women are higher than among men. This could possibly be explained by the fact that standard rescue/evacuation procedures generally dictate that women and children should be rescued first.
b) Since these figures are in % terms, any distortion due to more women than men being present on board is negated.
c) Survival rates among lower classes are much lower than the upper classes. This maybe because lower class living facilities were more unsafe and were the first to get flooded by water or it could be due to discrimination in allowing access to lifeboats or due to some other reason not directly understandable from the data



The chart above shows the survival rate in % terms by age cohort. The chart again shows an interesting trend:

a) In line with the general expectation that women and children are to be rescued first, the survival rate among children upto 18 years of age is the highest
b) Survival rate among aged passengers over 60 is the lowest. Not many could survive a disaster of that size

c) Interestingly, survival rates among passengers of 30-60 years of age is the highest. It is not clear why this is so. One possibility could be:
    a. There was a large segment of people who came alone on board the Titanic in that age group
    b. Unlike families which have the responsibility to save their children/spouses/parents, these individuals only have to save themselves
    c. This is possibly why they have a higher survival rate

CONCLUSIONS & LIMITATIONS

From the age distribution histograms we can see that majority of passengers on the Titanic are in the age group of 20-40. Even among the passengers who are alone, majority of passengers are in the age group of 20-40. Furthermore, the distribution is positively skewed, with more passengers below the mean than above the mean. This can be understood from the fact that the median for both data samples is below the mean. It must be noted however, that this conclusion is limited by the fact that the passenger information is available for 891 people which does not constitute the full population on board the Titanic. Additionally, information on the age of 177 passengers of the sample of 891 is missing. Inclusion of this information could change the results significantly. This data can be further investigated by taking various slices of the data such as age distribution by gender, by class etc.

From the survival rate bar charts, we can conclude that women and children had a higher survival rate than men. This could potentially be due to the fact that in emergency situations women and children are generally rescued first. We can also observe a clear fall in survival rate for lower classes when compared against upper classes and this trend is regardless of gender. This may be because lower class living facilities were less safe and may have been the earliest to get submerged or may be because lower class passengers did not have the same level of access to lifeboats as upper class passengers. If we had cabin information for all the passengers and had an idea on which level the cabins were located, we may have been able to establish whether in fact the lower class cabins were more likely to be submerged. Survival rates among elderly passengers aged above 60 is low. Older passengers may have found it difficult to survive a disaster of that magnitude.

REFERENCES

1. Data source: Udacity
2. Data Description: Kaggle (https://www.kaggle.com/c/titanic/data)
3. Adding histograms with bins: Matplotlib documentation (https://matplotlib.org/examples/statistics/histogram_demo_histtypes.html)