Ana Javed

Titanic Data Analysis

The Titanic was a British ship that tragically sank in the Atlantic Ocean on its route to New York City on April 15, 1912. Approximately 2,224 passengers and crewmembers were aboard at the time. Data containing information of 891 passengers has been access from https://www.kaggle.com/c/titanic/data to conduct the following analysis on. The results and visualizations were produced from utilizing Python, NumPy, Pandas, and Matplotlib. In addition, the questions explored were:

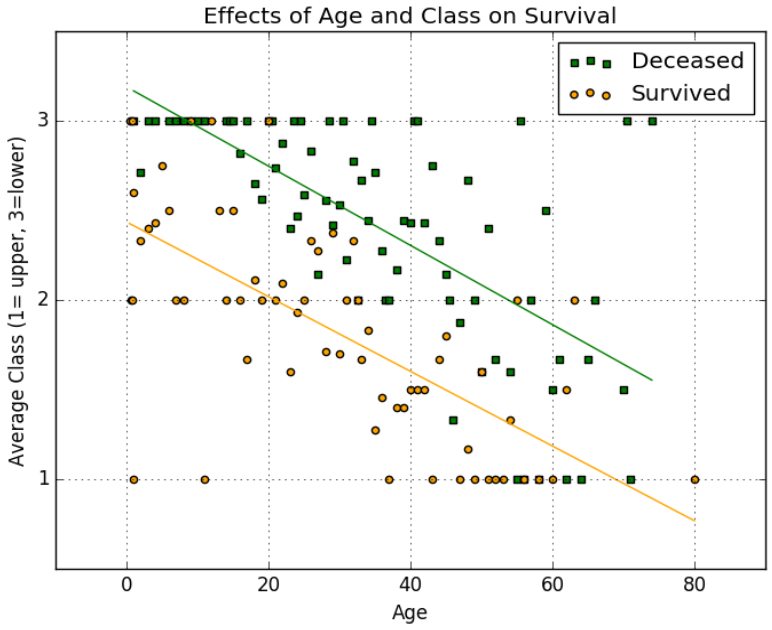
- Is there any relationship between a surviving or deceased passenger's age and class?
- Were younger passengers more likely to survive than older ones?
- Did more men die than women?

The data was examined by using Pandas DataFrames to group and isolate particular variables. Age was one of the columns that had 150 missing values from the set and they were omitted when examining the effect of age on survival.

Effects of Age and Class on a Passenger's Survival

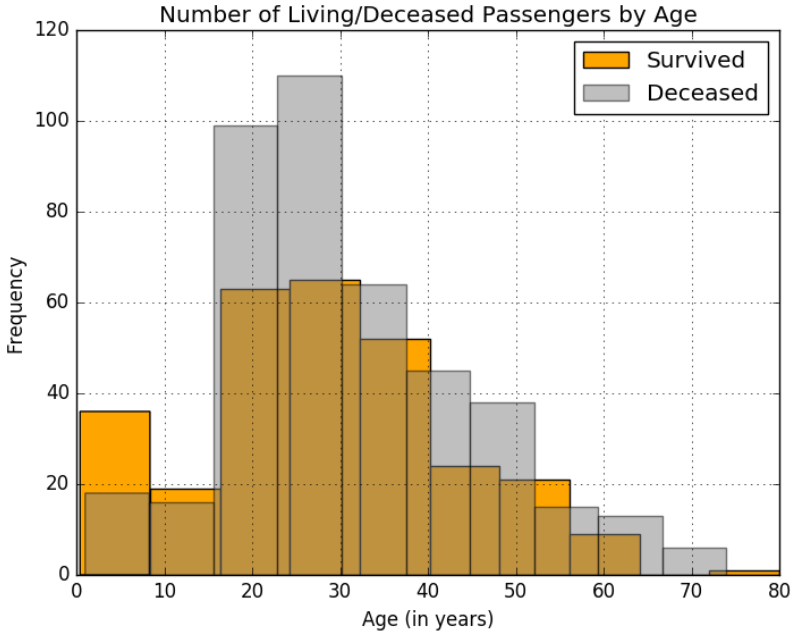
The Titanic hosted individuals of varying ages and class (upper, lower, and middle). This first graph was created to examine if there were any relationships between age and class for passengers who did and did not survive. The graph illustrates that there are more passengers who were aged between 40-70 years in the upper class. In addition, more of these data points are yellow, which implies that individuals in this group mostly survived. On the other hand, individuals aged between 0-40 years appear more in middle and lower classes. These individuals appear often as green on the graph, which could mean that many died after the Titanic sank.

Additional calculations also revealed that the probably of individuals in each class surviving was 0.629 or 69.2% for first class, 0.473 or 47.3% for middle class, and 0.242 or 24.2% for the lower class. The line of best fit was placed to model the trend of each scatterplot, which showed that as age increased, average class decreased (or became more upper class).



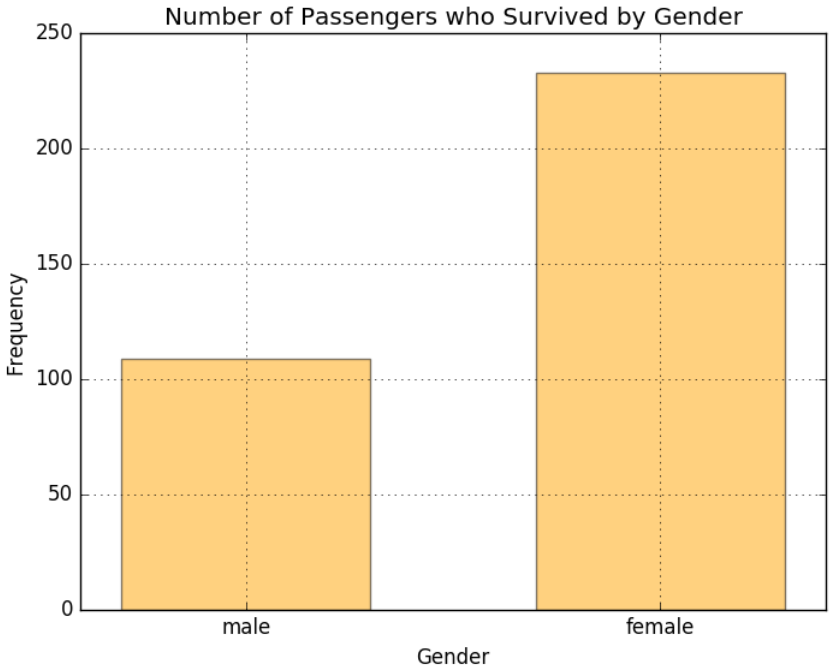
Effect of Age on a Passenger's Survival

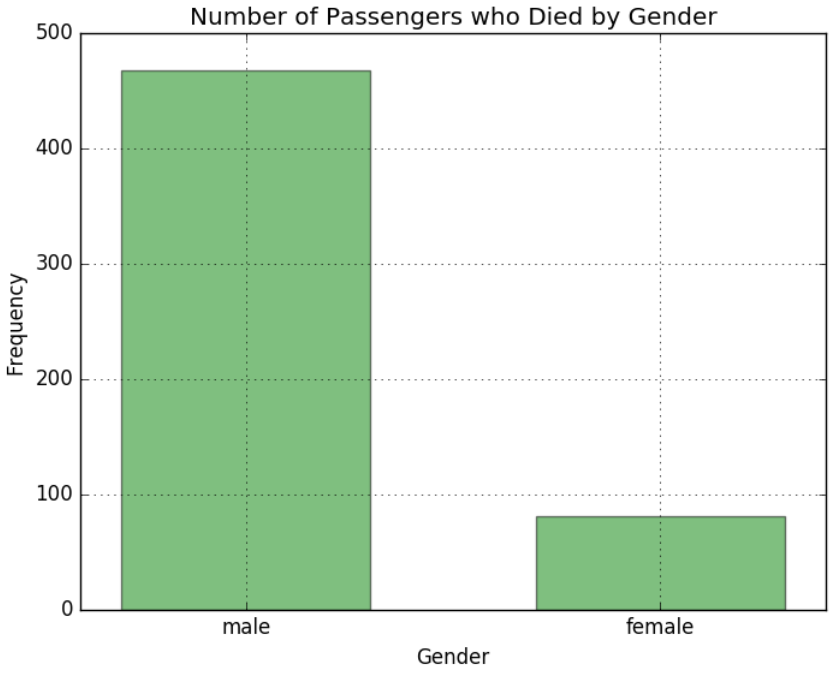
The graph below was created to further analyze the relationship between a passenger's age and whether they survived the Titanic sinking or not. The histogram demonstrates that approximately more 15-40 year old passengers died than survived in the aftermath, since the gray bars have higher frequencies than the orange bars for that age range. In addition, the graph also answers the initial question about survival of children, in that more children between ages 0-15 survived than died in the accident. The mean age of passengers who survived was 28.34, with ages that ranged between 0.42-80 years. In contrast, the mean age of those who perished was 30.62, with a range of ages between 1-74 years.



Effect of Gender on a Passenger's Survival

Gender was the final attribute that was examined to investigate its relationship with a passenger's survival. Most often in these tragic situations, women and children are given access to life save measures before men (i.e. row boats), and I want to see if that held true for the Titanic. The two bar charts below show a stark difference between the number of men and women who survived. From my calculations, out of the 549 individuals who died, 85.25% were men and 14.75% were women. Out of the 342 individuals who survived, 31.87% were men and 68.13% were women. In summary, women had a 0.742 or 74.2% chance of surviving, where as men only had a .1889 or 18.89% chance. It is indicative from both the calculations and the bar charts that more women survived, and had a higher probability of surviving, than men overall after the Titanic sank.





Conclusion

From wrangling and exploring the Titanic data in this report, it was deduced that passengers aged 40-70 years were more likely to be in first class and survive than the passengers aged 0-40 years who were in middle or lower classes. When focusing on age, passengers between 15-40 years died more than survived. In addition, younger children (between ages 0 -15) were more likely to survive the aftermath of the Titanic. Finally, when comparing genders, women had a higher chance of surviving.

It is necessary to also mention that the graphs created and analysis conducted in this report show correlation between variables (i.e. age, gender, class) and not causation. For example, being a woman passenger did not cause you to survive the Titanic sinking, just as how being in the lower class did not cause you to perish from the crash. Even though relationships between these factors do exist, the findings in this report are tentative and require further statistical testing (such as z-tests or t-tests) to accurately determine the true relationship and strength between two variables.

It is also possible there are other factors influencing this data that have been overlooked in the process. Some variables that could have been useful to include in the study include a passenger's average salary, or the highest level of education a passenger received. Another limitation of this study was that out of 891 passengers, only 741 had an age documented for them. I omitted these passengers from the dataset when investigating the relationship between age and survivorship to ensure the analysis was not influence by these values. The disadvantage in doing so, however, is that the sample size of the data was reduced by 150 and may not accurately represent the whole group of Titanic passengers.