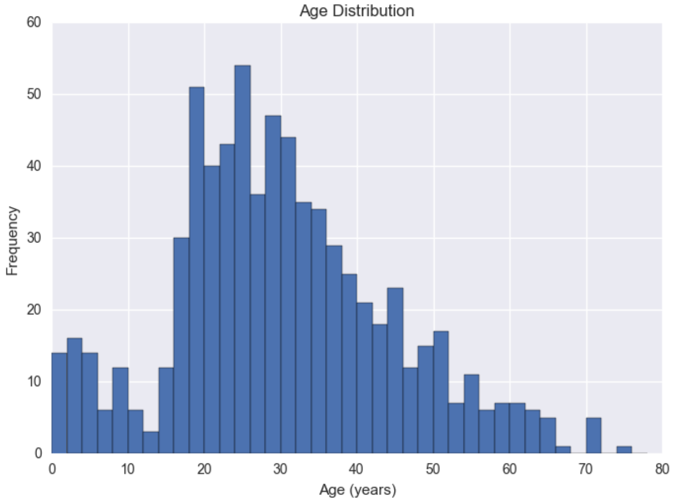
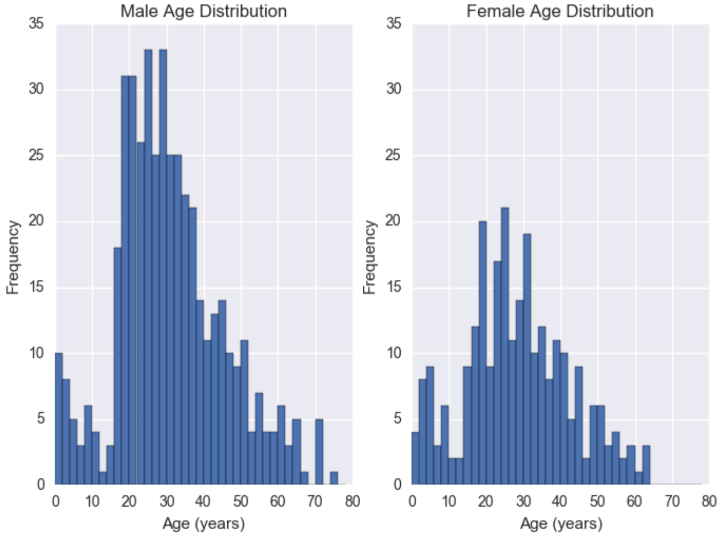
Question 1: What does the age distribution of the passengers look like for our Titanic sample?



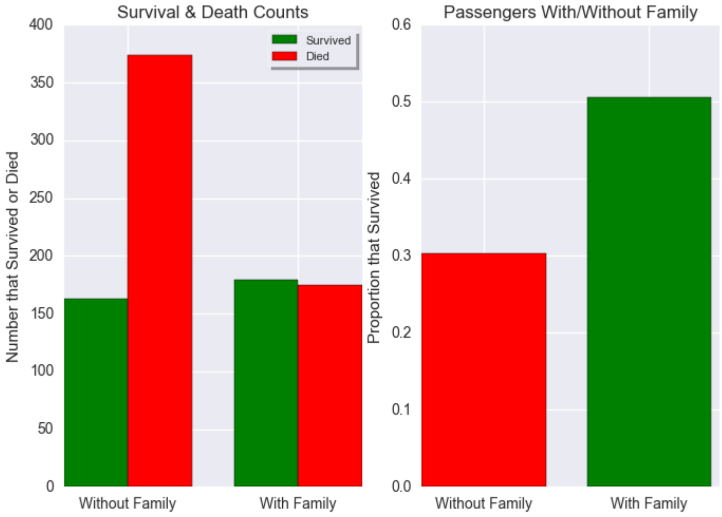
From the total sample of 891 passengers only 714 records included an age. The average of this sample is 29.69 years old, the median is 28 years old, and the standard deviation is 14.52 years. A quasi bimodal distribution is apparent with a cluster of passengers between 0 and 10 years old and then a large number of passengers between the ages of 16 and 50 years old. It is important to note that these are statistics of the sample and not of the entire population. We are not given the population mean (µ) or standard deviation (σ). I infer the population mean from our large random sample of passengers and provide a 95% confidence interval for which the true population mean probably lies within [28.47 30.91]. The sample mean is a point estimator for the population mean. For confidence intervals of µ when we do not know σ we will use s (standard deviation of the sample) to approximate σ and we find the margin of error by multiplying the SE (standard error) by a t-score which comes from the T-distributions. In this case the standard error is 0.54 years, t= 2.2461 (2.500% of a t distribution has values greater than 2.246 or less than -2.2461), and the degrees of freedom = 713. I use a t-distribution as a more conservative measure to find my error bounds of the true population mean since I am not given *any* population parameters.



This purpose of this graphic is to visualize the age distribution of males and females in our sample. The mean female age is 27.9 years and the median is 27 years. For men the mean age is 30.7 years and the median age is 29 years. The minimum female age is .75 years while the maximum age is 63 years. The minimum male age is 0.42 years while the maximum age is 80 years.

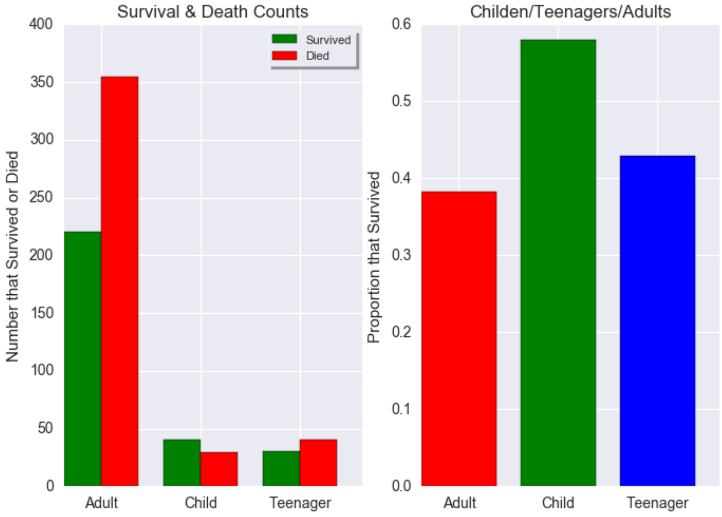
<u>Question 2:</u> <u>What factors influence survival?</u>

-Did a higher proportion of those with family survive?



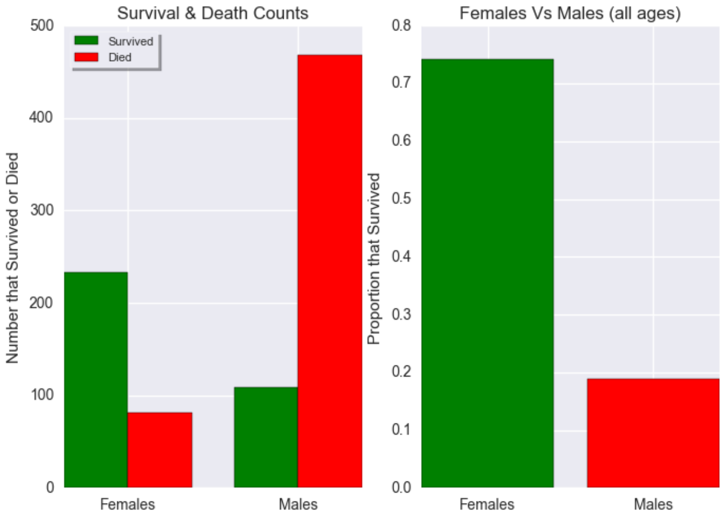
Based on our sample a higher proportion of those with family survived (179/354=.505) versus those without (163/537=.303). I defined family as the case where the parameters “Parch” (number of parents and children aboard) and “SibSp” number of siblings and spouses aboard) was greater than or equal to 1. A limitation of this analysis is in the way family is defined. For example, extended family is not considered and could alter these proportions.

-How does survival proportion compare for adults, children, and teenagers?



Based on our sample a higher proportion of children (40/69=.579) (ages 0-12 years) survived compared with teenagers (30/70=.428) or adults (220/575=.382). These differences are not necessary indicative of the entire population onboard the Titanic. A limitation of this analysis is in the definitions of the groups by age. This is a somewhat fuzzy metric and changing the ranges will undoubtedly affect the survival proportion for each group.

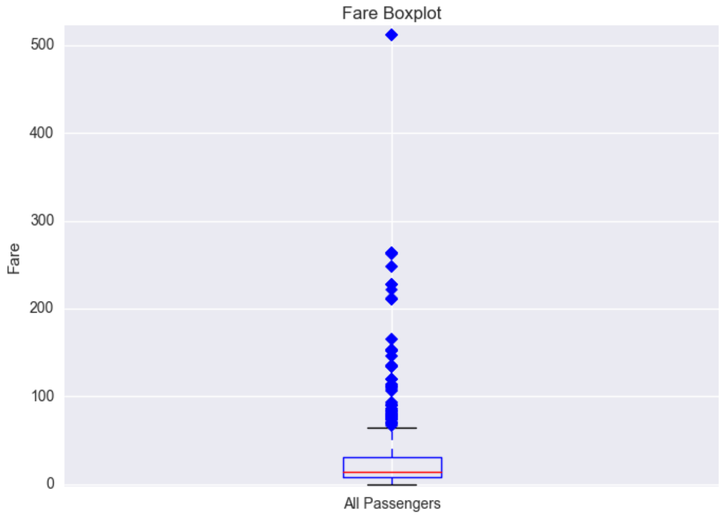
-Is there a difference between the survival proportions of females versus males?



From our sample it is clear to see that a higher proportion of females survived (233/314=.742) than men (109/577=.188). This could illustrate the socially accepted practice to save women first. What this graphic doesn't show was that there was a much larger number of men (577) than women (314) in our sample. A limitation of comparing these proportions using a 2 sample z-test is that the samples are not necessarily independent. For example a man and woman could be married and the survival of one could influence the survival of the other. The same argument could extend to entire families where the survival of one family member could influence the survival of the rest regardless of gender. Another condition of the 2 sample z-test for proportions that is not met is that that each population (male and female) is at least 20 times as big as its sample.

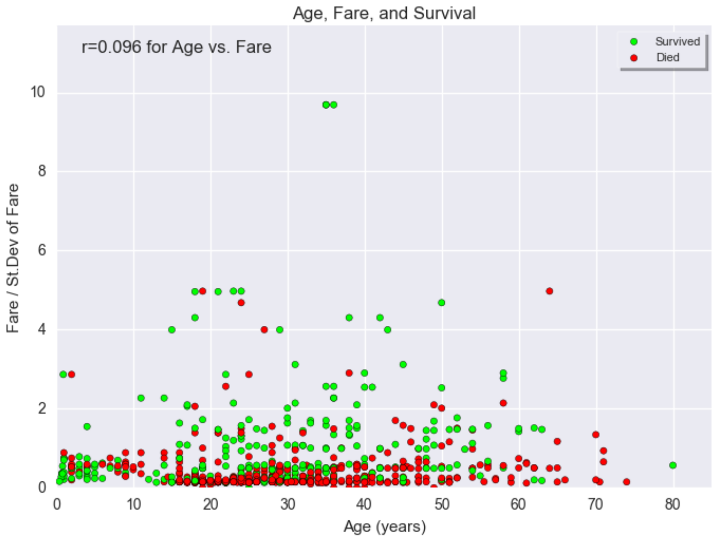
Question 3: What factors influence ticket fare?

-Before investigating I provide some summary statistics about ticket fare below:



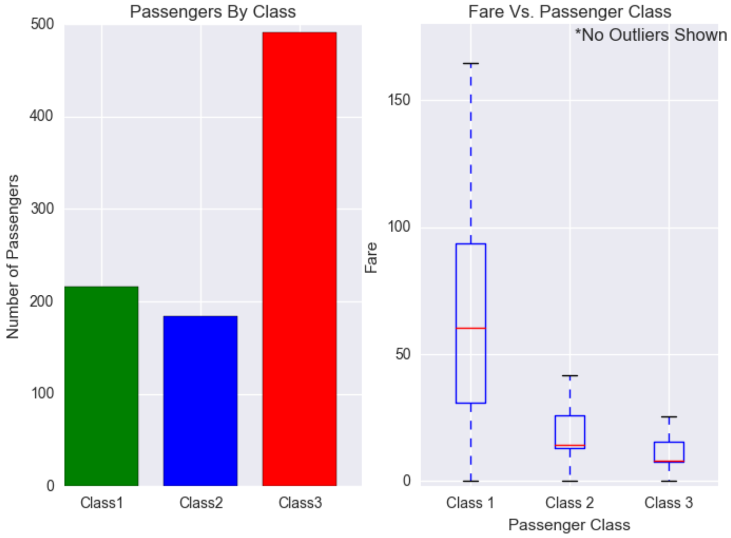
This boxplot helps to visualize the distribution of fares that passengers paid. The average ticket fare is 32.20 while the median is 14.45. The maximum fare paid was 512.32. The outliers (blue diamonds in above graphic) bring the average up. The 75% quartile is 31.0 while the 25% quartile is 7.91.

-Is there a relationship between age, ticket fare, and survival?



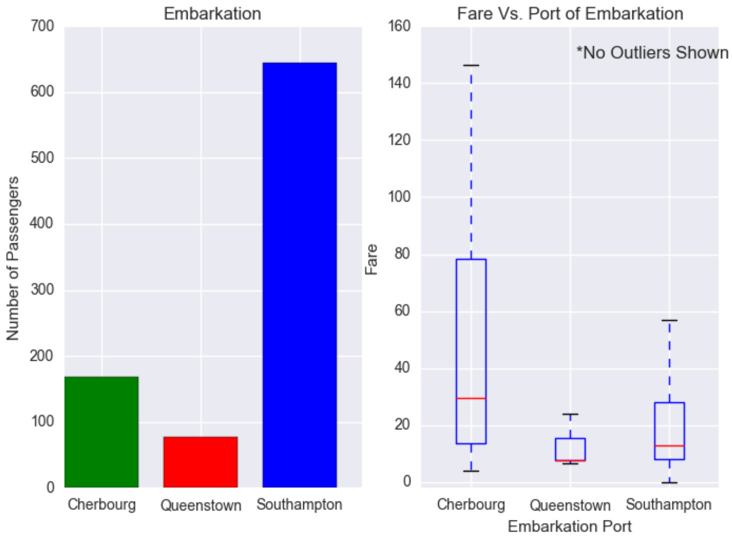
I found a very weak correlation between ticket fare and age. Pearson's r=.096 with 95% CI [0.0228, 0.1698]. It is important to remember that correlation does not mean causation. In other words, being older does not necessarily mean a higher fare. To compress the y-axis limits in the graphic, I divided each fare by the standard deviation of all the fares. I also colorized the points using the survival attribute as an additional visual method for observing trends in age and survival. From this it is possible to see a cluster of young children who survived and also a cluster of adults who perished between the ages of 20 and 40 who had a fare/st.dev fare less than 1.

-Is the median fare higher for 1st class passengers than 2nd or 3rd class?



In the sample, there were a total of 216, 184, and 491 passengers in Class 1, Class 2, and Class 3 respectively. I found that the median fare for first class (60.29) was higher than for second (14.25) or third class (8.05). This finding was not surprising. Better amenities typically mean higher fares. The interquartile range for Class 1 (30.9 to 93.5) was also wider than Class 2 (13 to 26) or Class 3 (7.75 to 15.5). In all three passenger classes some individuals went on the cruise for free.

-Is there a relationship between port of embarkation and ticket fare?



In our sample, a total of 168, 77, and 644 passengers embarked from Cherbourg, Queenstown, and Southampton, respectively. The median fare from Cherbourg (29.7) was higher than for Queenstown (7.75) or Southampton (13). The interquartile fare range for passengers embarking from Cherbourg (13.7 to 78.5) and overall fare range was wider for this port as well. This sample may not be representative of the entire population. An ANOVA test is necessary to determine if the mean fares between the embarkment ports are significantly different.

<u>REFERENCES</u>

https://www.kaggle.com/omarelgabry/titanic/a-journey-through-titanic/notebook

http://mathesaurus.sourceforge.net/matlab-python-xref.pdf

http://stackoverflow.com/questions/14016247/python-find-integer-index-of-rows-with-nan-in-pandas

http://pandas.pydata.org/pandas-docs/stable/indexing.html

http://stattrek.com/hypothesis-test/difference-in-proportions.aspx?Tutorial=AP

http://stackoverflow.com/questions/30390476/equivalent-of-rs-of-cor-test-in-python

http://www.graphpad.com/quickcalcs/statRatio2/

http://home.southernct.edu/~mugnor1/past/mat107/Lessons/chap8.doc