Patrick Coakley
2/16/2017

Project 2: The Titanic

What Am I Trying To Solve?

For this project, I chose to work with the Titanic data set because of the impact on history it had, as well as the mysteries surrounding it, some of which continue to remain unanswered. Using this dataset, I would like to find out: what kind of person was most likely to survive the Titanic? I plan to accomplish this by looking at different attributes of the survivors, including gender and age.

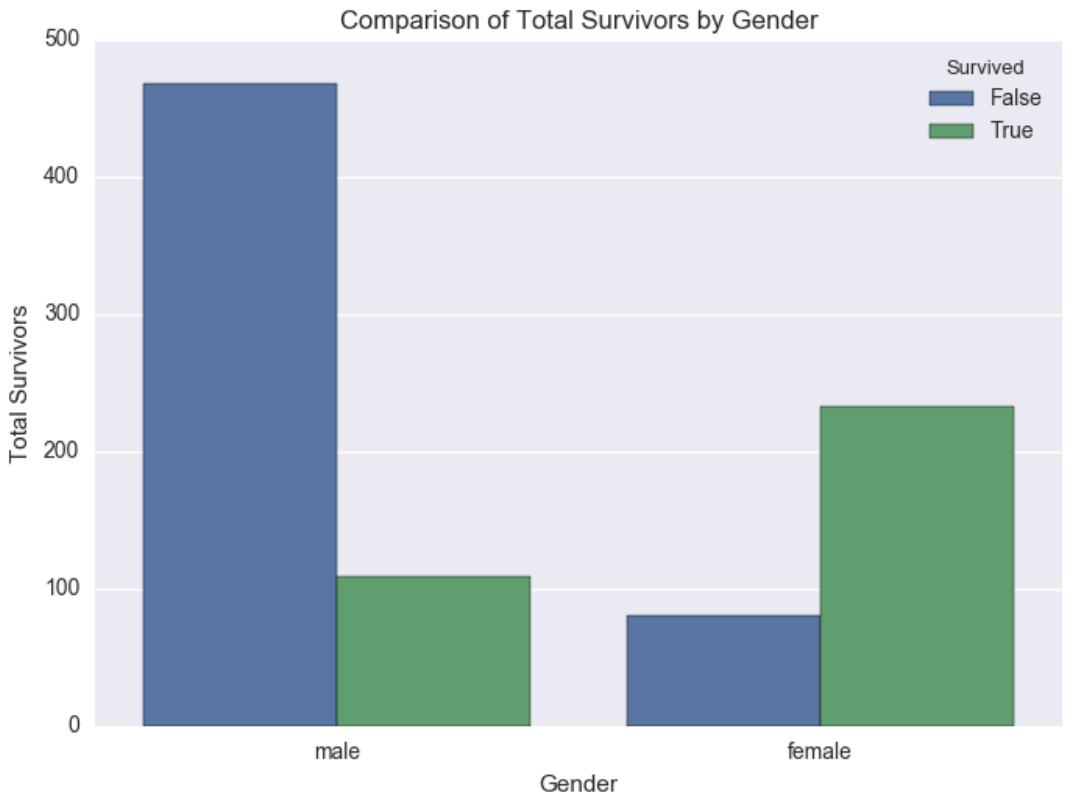
Cleaning The Data

Originally I spent some time figuring out the easiest way to clean some of the data, but eventually settled on keeping a simpler approach. For example, I originally planned to drop all unknown ages because any NaN rows in this analysis would not count towards looking at the ages of the survivors, and if they were, for instance, infants, then it 0 would be written for the age. However, it made more sense to keep the original data as is and simply drop them when it comes time to visualize or query the data.

This also applies to breaking down the genders and ages of the survivors; originally I wanted to use different variables assigned to each demographic, but again it's not necessary because you can simply filter it by adding comparators to your Python code. I think this approach is much better than creating named variables because that is simply using more memory than needed. The end result is that I now have much less code than I originally created.

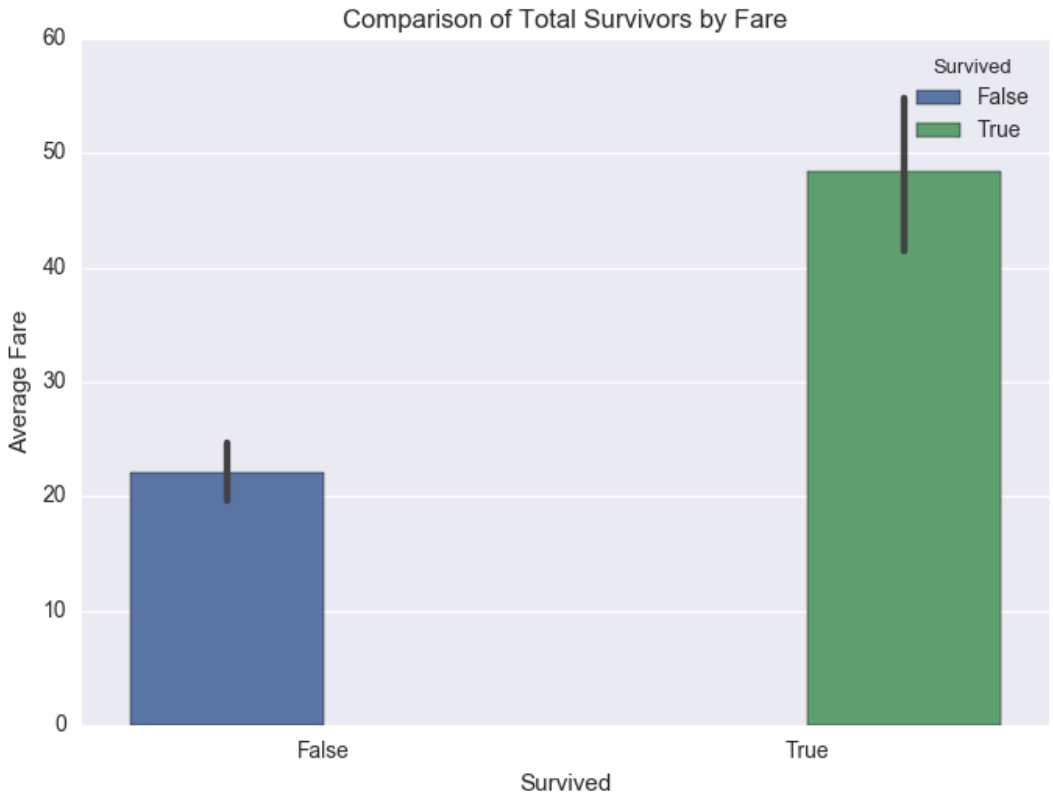
Exploring The Data

The first thing I tested was to see the total count for men and women who survived the Titanic by making a simple bar chart using Seaborn and Matplotlib:

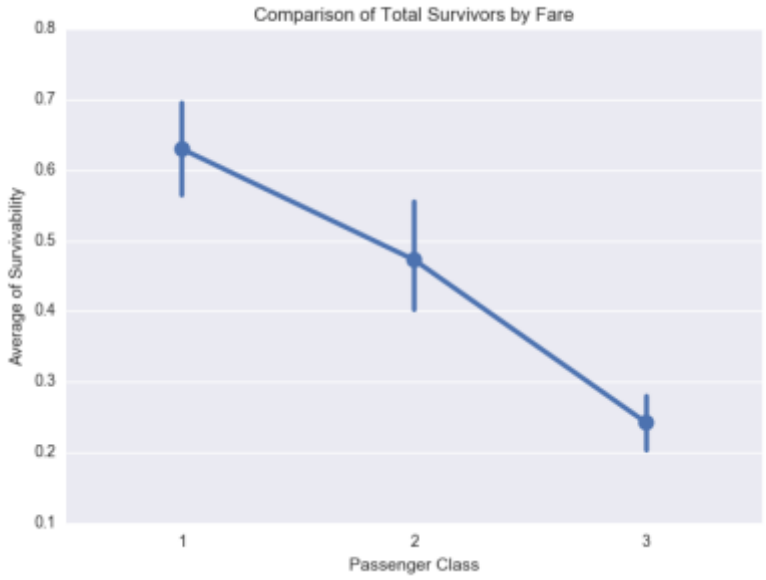


From this we can see that most of the survivors were female, and that there were far more men on the boat in total than women.

Next, I looked at the average fare of the passengers and found that the average fare of the survivors was much higher than those that died:



Then I used a point plot to compare the survivability of a passenger to the class they were assigned to:



I found that there is a higher average of survivability in the first class than the other two classes, though second class is also quite high in comparison to third class passengers.

Using a violin plot, I was able to see the distribution of age and class for each sex of survivors:

The distribution heavily favors women and men over 20 but under 40 years of age. Seeing this also led me to create a simple print statement to show survivors over 60 and non-survivors over 60 to compare the remaining numbers just to reinforce an idea relating to the previous visualization:

Survived	Pclass	Sex	Age	Fare
True	1	female	63.0	77.9583
True	1	female	60.0	75.2500
True	3	female	63.0	9.5875
True	2	male	62.0	10.5000
True	1	male	60.0	79.2000
True	1	male	80.0	30.0000
True	1	female	62.0	80.0000
Survived	Pclass	Sex	Age	Fare
False	2	male	66.0	10.5000
False	1	male	65.0	61.9792
False	1	male	71.0	34.6542
False	3	male	70.5	7.7500
False	1	male	61.0	33.5000
False	1	male	62.0	26.5500
False	3	male	65.0	7.7500
False	3	male	61.0	6.2375
False	1	male	64.0	263.0000
False	1	male	65.0	26.5500
False	1	male	71.0	49.5042
False	1	male	64.0	26.0000
False	1	male	62.0	26.5500
False	1	male	61.0	32.3208
False	2	male	70.0	10.5000
False	2	male	60.0	39.0000
False	1	male	60.0	26.5500
False	1	male	70.0	71.0000
False	3	male	74.0	7.7750

This shows that if you were over the age of 60, you were less likely to survive , first class or not. It also shows that there weren't many women on-board in this age bracket, and all but one of them were in first class.

Finally, I made one more print statement to confirm some more information about the survivors' ages and how they compared to non-survivors:

```
Total survivors with numeric ages: 290
Total survivors (including unknown ages): 342

Total non-survivors with numeric ages: 424
Total non-survivors (including unknown ages): 549

Total number of female survivors: 197
Total number of male survivors: 93

Total number of female non-survivors: 64
Total number of male non-survivors: 360

The average age of a female survivor: 28.847715736
The average age of a male survivor: 27.2760215054

The average age of a female non-survivor: 25.046875
The average age of a male non-survivor: 31.6180555556

The average age of a female survivor in first class: 34.9390243902
The average age of a male survivor in first class: 36.248

The average age of a female survivor in non-first class: 24.5043478261
The average age of a male survivor in non-first class: 20.5047169811
```

Conclusions Drawn From The Data

Based on the data exploration I did, I believe that the primary surviving demographic were upper-class adult women. Generally speaking, in emergency situations where evacuations are necessary, women and children would be the ones to take precedence over anyone else, though we could also consider the elderly and sick as well in this case. Of the passengers that are over 60, all of the ones that died were all male, while 4 women and 3 men survived. All but two of the survivors in this group were also in first class

One of the problems with the data is that there are a large number of unknowns in the ages. If we look at the total number of survivors where age is greater than zero, the total number is 290; the total of all including unknown is 342. That leaves 52 unaccounted for depending on how we examine the data. This problem also effects the total non-survivors, where the total with ages greater than zero are 424, and with all ages it is 549.

In any case, the average age of a survivor with an age greater than zero is around 29 years old for a female and 27 for a male. This is different than the average ages of the non-survivors, where women were at 25 years old and men were at around 32 years old. It's hard to really draw too much from these differences, but it does seem like if you look at them in conjunction with the other pieces of data, the older men were less likely to survive, but older women were more likely to survive.

This also carries over to first class average ages, where women were around 35 years old and men were around 36 years old compared to the lower classes, where women were around 25 years old and men were around 20 years old. From this you can surmise that the second and third class passengers were younger folks who had less money, where as the first class passengers had more money on average.

If I were to use this information in a prediction model, I would heavily favor first class female passengers, and females in general over the age of 25. I also would favor first class men with ages over 35. It seems like if you were less wealthy and were a younger male, your survivability went down drastically. It's hard to say what directly influenced the outcome of these numbers, but it would seem the wealthier passengers simply had more influence over the situation, or perhaps they were favored over others by the crew. Knowing the actual reasons would require more context of the data than is given.

Sources

Kaggle. (n.d.). *Kaggle*. Retrieved from Titanic: Machine Learning from Disaster : https://www.kaggle.com/c/titanic/data

Pandas. (n.d.). *pandas 0.19.2 documentation*. Retrieved from pandas: powerful Python data analysis toolkit: http://pandas.pydata.org/pandas-docs/stable/

Seaborn. (n.d.). *Seaborn: statistical data visualization*. Retrieved from Example gallery: https://seaborn.pydata.org/examples/index.html