Investigate a Dataset Project

Dataset Used:

I used the Titanic dataset. The file was titanic_data.csv.

Limitations to Data:

After performing a count on the data, I noticed that our dataset only contains 891 records. In real life, the number of passengers on the Titanic, including crew, was closer to 3,300. Therefore we do not have the complete passenger list. In addition, we do not know if this population of 891 is a biased sampled because it was provided to us. For this analysis, we will assume this sample is representative of the entire population on the Titanic.

In addition, of those 891 records, 177 of those records had a null value in the Age field. We did not remove these records, I replaced the age with the average age for the sex and class relating to that sex. Example, the average age for male in the third class was 26.5. If there was a record entry where the passenger was male and the class was third class and the age was null, then I replaced the null value with the average age for that sex and class which was 26.5

Questions:

I wanted to assess which variables indicated if a passenger was more or less likely to survive and therefore I asked:

- How many people survived and how many people died in the dataset?
- How many people survived and died from each class?
- How many people survived and died from each age group?
- How many people survived and died from each sex?

Process:

My overall process was to create frequency charts for each of the variables mentioned and compare that variable in the survived and died populations to look for significant differences to see which variables might indicate if a passenger was likely to survive.

I also created a histogram for age to assess if the distribution of age was different.

Data Wrangling:

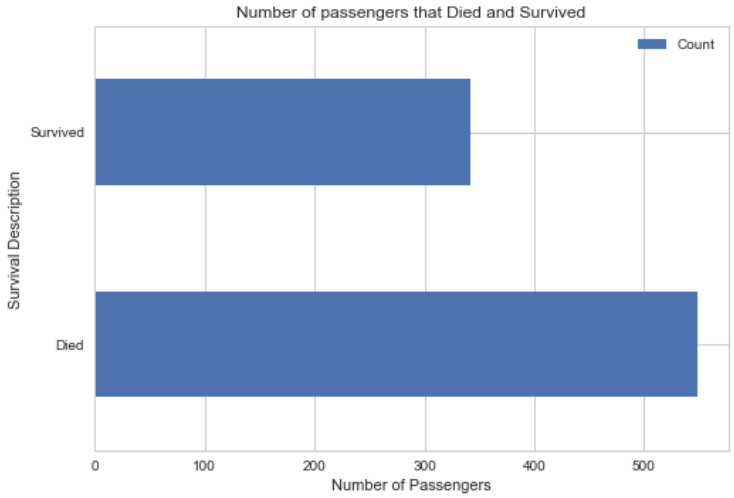
1. I created a count field that I could use to count records.
2. I found the average age by sex and class.
3. I replaced the null age values in the age field with the appropriate age by Sex and Pclass.
4. I dropped fields that I was not going to use in my analysis, such as Name.
5. I created labels for fields to make them more descriptive and to use in charts for later. I did this for Survived and Pclass.
6. I created age group bins for the age variable.
7. Then for each question, I took the data elements that I needed and grouped the data if they survived and the variable that I was assessing and found the sum of the records.
8. I then pivoted the variable so that those values were now the column headings.

9. I finally visualized the results using plot().

Results:

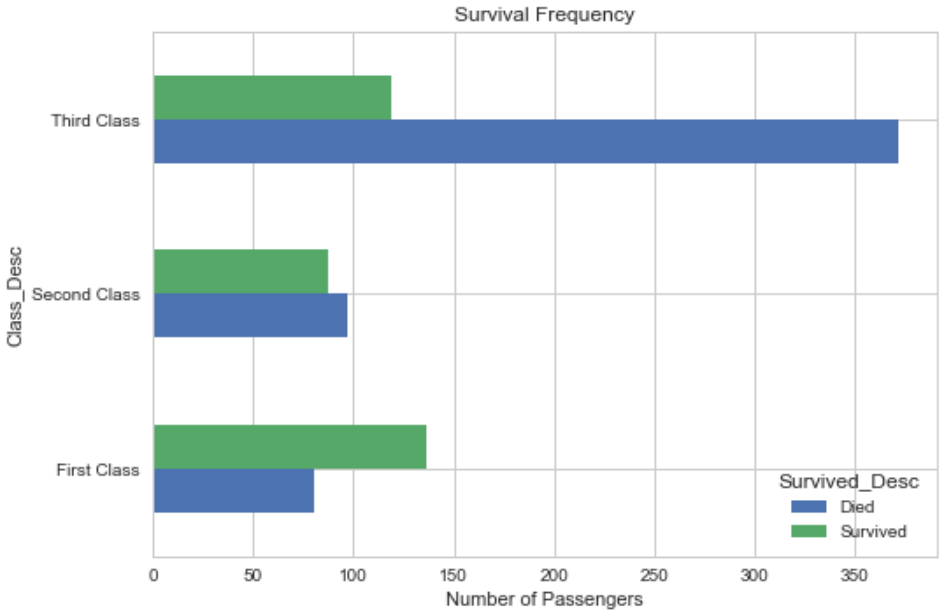
For Question- *How many people survived and how many people died in the dataset?*

There were 891 passengers in our records. 549 died and 342 survived. See chart below.



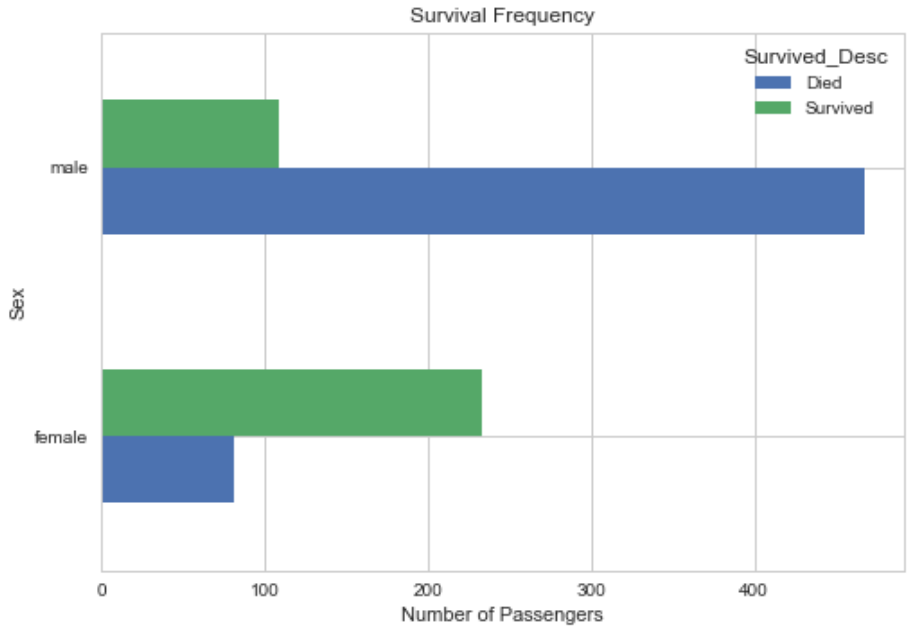
For Question- *How many people survived and died from each class?*

It appears below that you were more likely to die if you were in third class. 372 of third class died while only 119 survived and that is about 24% survival rate. Meanwhile, if you were in first class you were more likely to survive. In first class, 136 survived while 80 died and that is about a 62% survival rate. (See chart below)



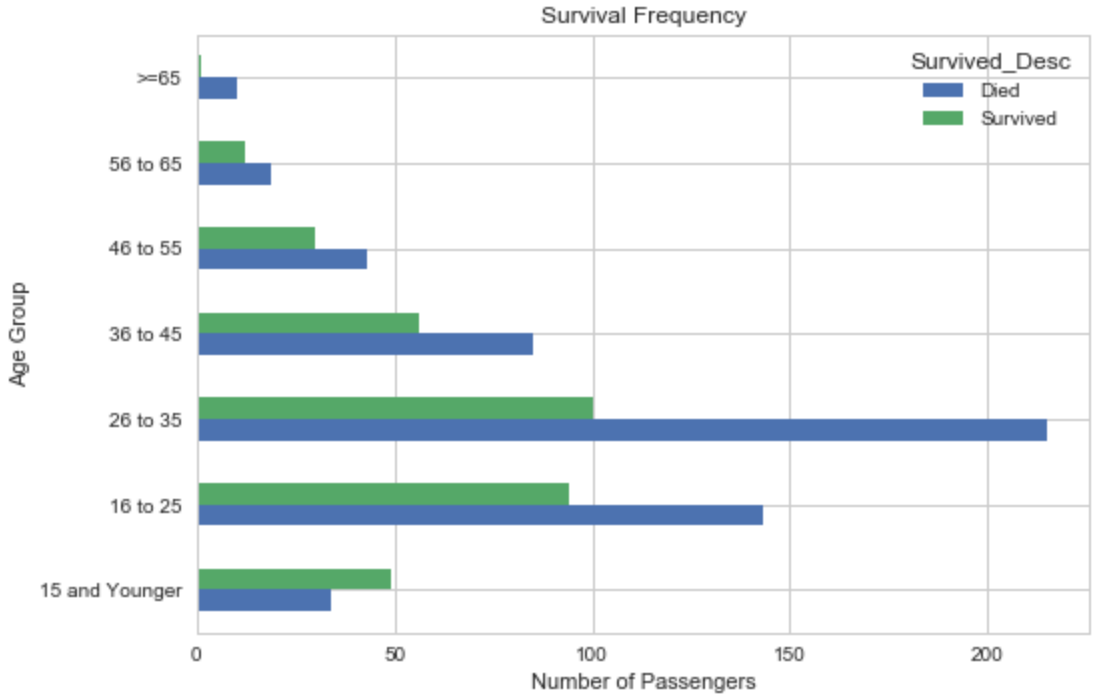
For Question- *How many people survived and died from each sex?*

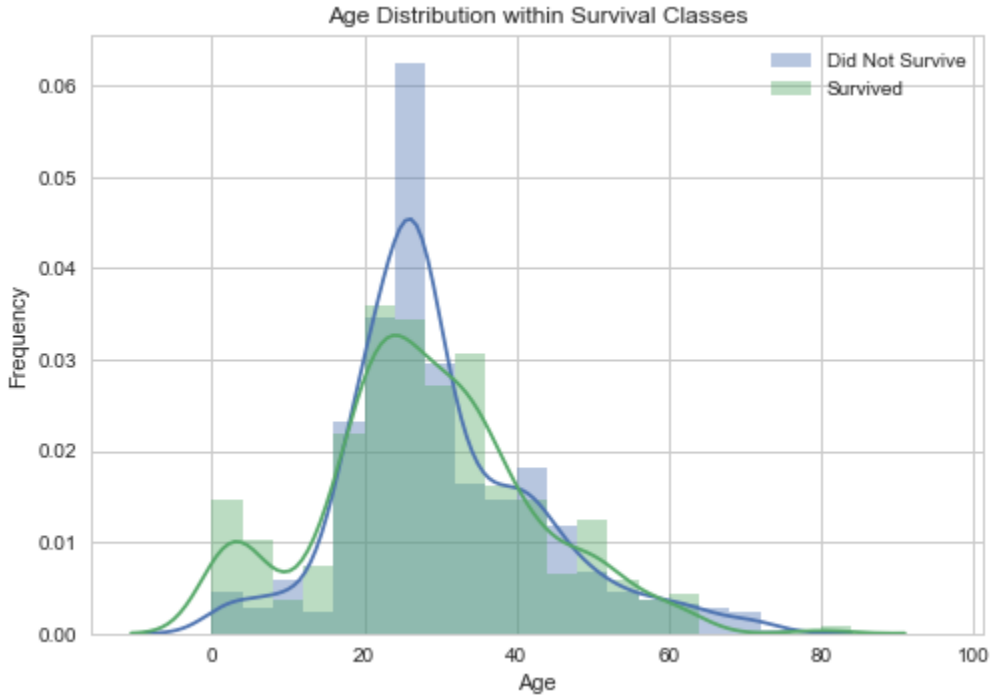
It appears below that you were more likely to die if you were a male. 468 of males died while only 109 survived and that is about 18% survival rate. Meanwhile, if you were a female you were more likely to survive. For females, 233 survived while 81 died and that is about a 74% survival rate. (See chart below)



For Question- *How many people survived and died from each age group?*

It appears below that you were more likely to die if you were older than 65. Of the elderly, 1 survived and 10 died and that is about 9% survival rate. Meanwhile, if you were younger than 15, you were more likely to survive. For the children, 49 survived while 34 died and that is about a 59% survival rate. In general, it appears that the younger you were the more likely you were to survive. (See chart below)





For Question- *Which variables indicated if a passenger was more or less likely to survive?*

Of the variables that I examined, age, sex and class, all of the variables appeared to be significant in regards to if a passenger was more or less likely to survive.

Limitations to this analysis are 1) I did not examine all of the possible variables in regards to the passengers, I only examined three. There could be additional variables or other overlaying variables that this variable correlates with. 2) We had only 891 passengers in our data and this could be a biased sample. 3) I did not actually perform a statistical significance test. I only looked for significance visually and through proportions.

Next Steps:

The next steps to this analysis would be to either see the frequency of combined variables, such as what would be the frequency if you were young and in first class versus young and in third class. This could then ultimately lead to a logistic regression analysis where you figure out the probability of someone surviving and thus we could predict the rest of the passengers on the Titanic that fateful night.

Resources Used:

https://stackoverflow.com
https://pandas.pydata.org