Data Analyst Nanodegree Assignment II—Investigating a Dataset

Introduction:

For this project, I chose the Titanic data set from Kaggle. It contains the names of the passengers, the number of any people they traveled with, the fare they paid, their compartment class, their gender, their ages, their original port, their cabin, and whether or not they survived. The passengers who survived have a value of 1, and the others have a value of 0.

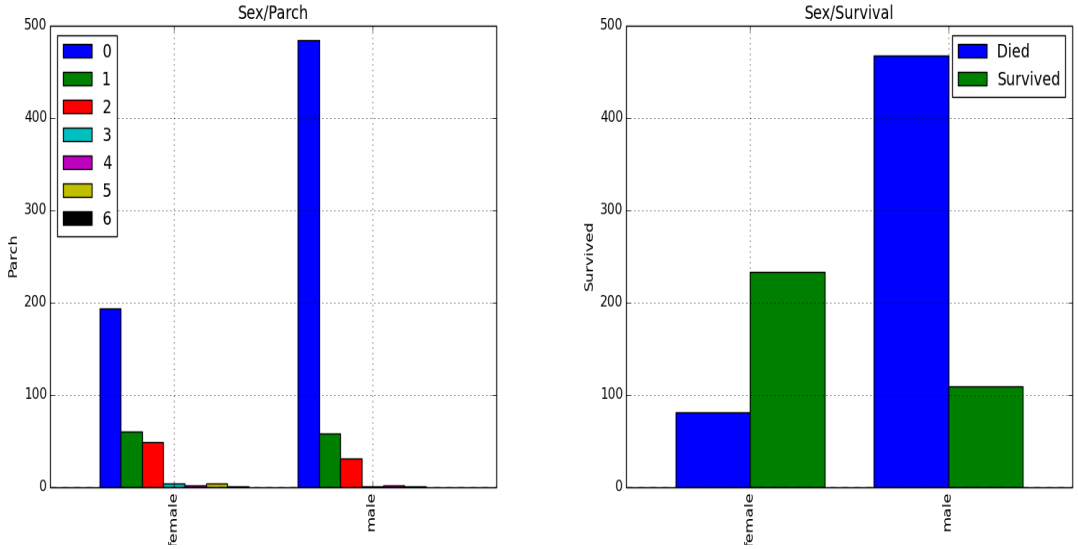
Questions:

1. Did the fare amount have anything to do with whether or not the passengers survived?
2. Did the passengers' cabin class have anything to do with survival rates?
3. Did the passengers' gender affect the people the number of people they traveled with?
4. Did the passengers' ages affect their fare rates?
5. Did the passengers' gender affect whether or not they survived?

Data Handling:

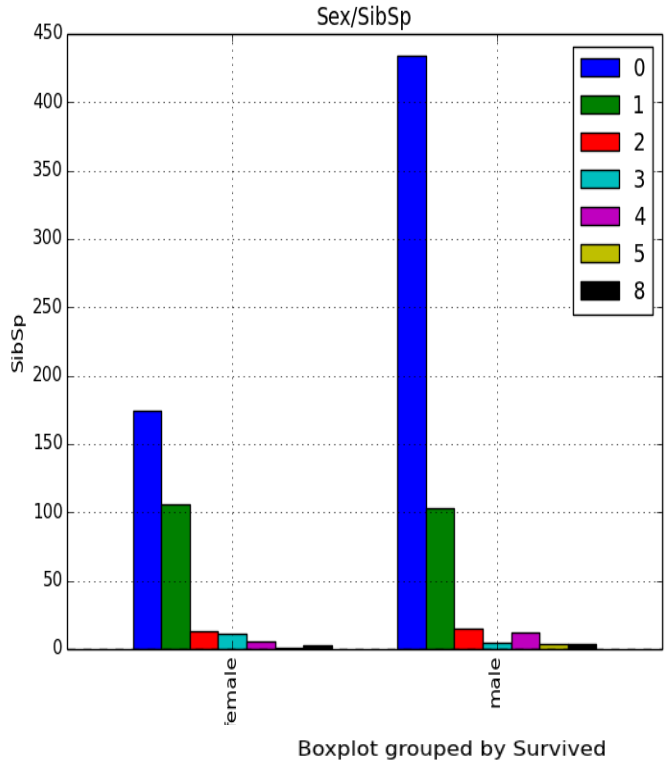
Several of the columns have missing values, namely, the Age column and the Cabin column. I replaced the missing values in the Age column with the string 'NaN.' The relationships I tried to determine were those between fare and survival rates, the passengers' ages and fare rates, the compartment class and survival rates, and the passengers' genders and the number of extra people (if any) that they were traveling with, and consequently, whether the passengers' gender affected their survival.

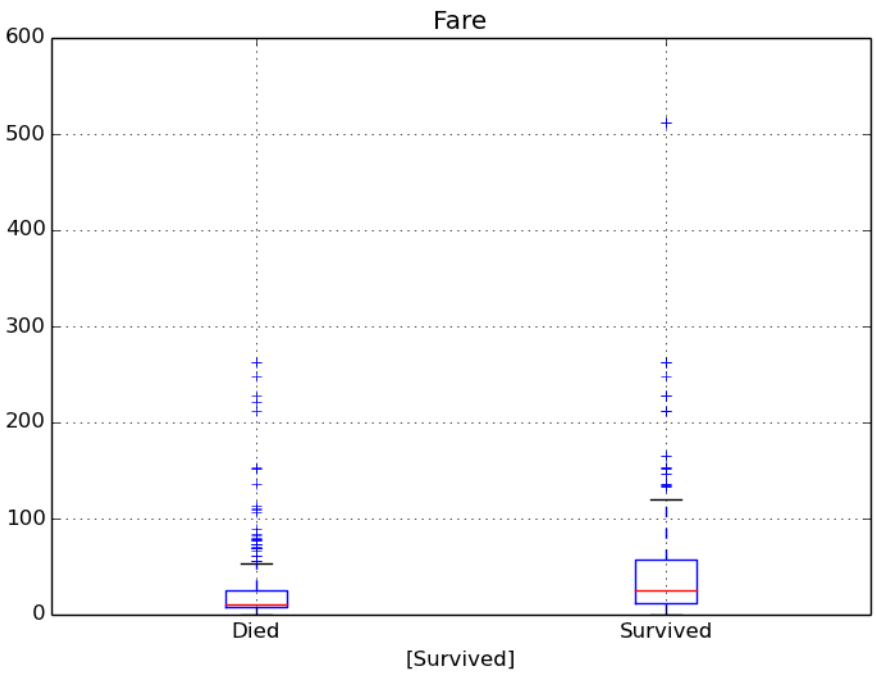
Exploration:



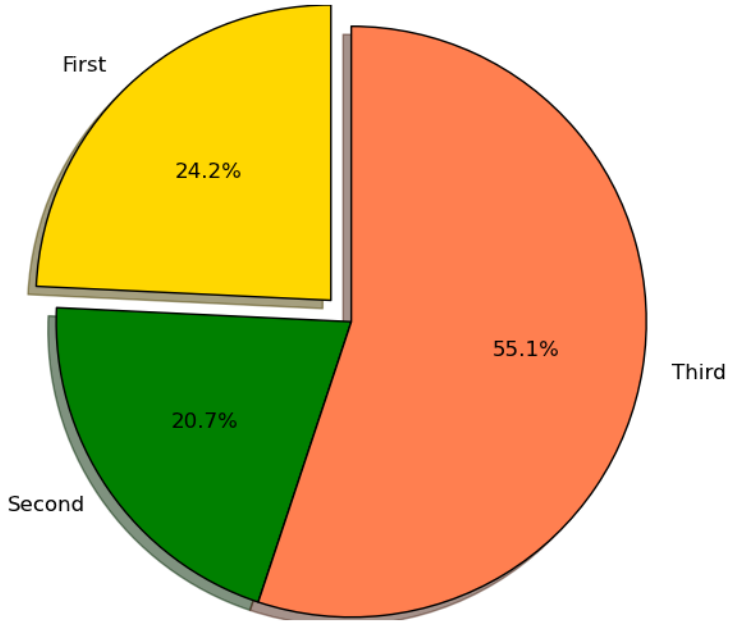
The histogram on the left displays the number of people that passengers of each gender traveled with and the histogram on the right displays how many passengers of each gender survived or died. As I expected, women were far more likely to travel with parents and children and men were far more likely to travel alone, which is why far more males died than women: the women and their children would have been evacuated first,

leaving less resources for the men. The same held for the relationships between siblings and spouses and the survival rates (the histogram for sex and siblings and spouses is shown below). As expected, a large majority of men traveled alone, so they would have gotten last priority when emergency resources were being allocated because the people with families would have been evacuated first.





Interestingly, the fare did not have a significant impact on the survival (the box-and-whisker plot is shown at the bottom of the previous page). I assumed that the people who paid higher fares would have had better accommodations, and therefore better access to the emergency provisions, but there was no correlation between the fare rates and survival rates. There is not a significant difference in quartile ranges for the people who died and the fare they paid. The difference is slightly more pronounced for the people who survived, but the people who paid higher fares are still considered outliers, which contradicted my initial assumption, even though almost half of the passengers traveled on first-class and second-class combined (see the pie chart below).



Results and Conclusions:

For this project, I used data frames and arrays from numpy. I used five arrays, one to store the values for fare and survival, one for passenger class and survival, one for age, one for cabin, and one for the original port of departure. The data frames extract the necessary data from each of the columns in the original csv and replaces empty values as necessary with the help of the 'replace' method. If passengers had any company traveling with them, the numbers of people in their company are either stored under Parch, which stands for "parent/child" or SibSp, which stands for "sibling/spouse."

Other factors that would have skewed the data were the number of people included in parents and children: parents and children included stepchildren of the passengers, which could have influenced how many people were given priority in the evacuation process and it would have influenced how much space there had been on the ship in order to be evacuated properly. The number of entries itself also hampered analysis as the assignment description itself mentions that only data from 891 of the passengers are included in the set, so we do not know how much the analysis would have

been affected with their inclusion. Another limiting factor is the exclusion of aunts, uncles, nannies, and grandparents. It is not clear how they would have been given priority, which can also affect research methods, thereby producing slanted results.