TITANIC PASSENGER DATA ANALYSIS PROJECT

Questions:
1. Did children (<=14Years old) survive more than adults?
2. Did females have a higher survival rate than males?
3. Did richer passengers (Pclass = 1) survive more?

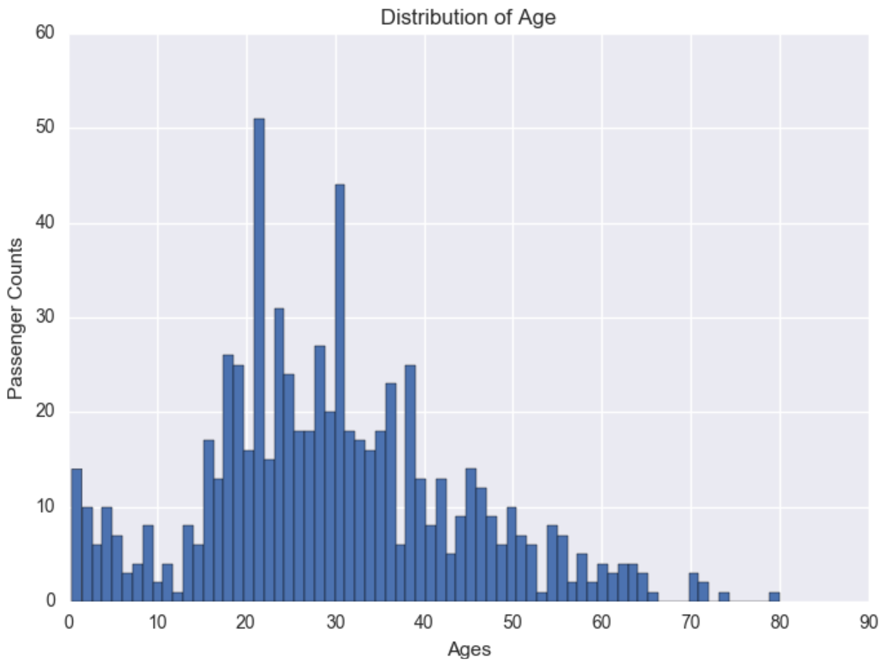
Missing Values Handling:

- Many of the data points did not have ages. A total of 177 passengers age data was missing. Since I could not guess their ages I decided to ignore these data points when I reviewed the age distributions. I did this my using the .groupby function which ignored data points which did not have age listed. On the other hand, while assessing Pclass and Sex data, I found no missing data.

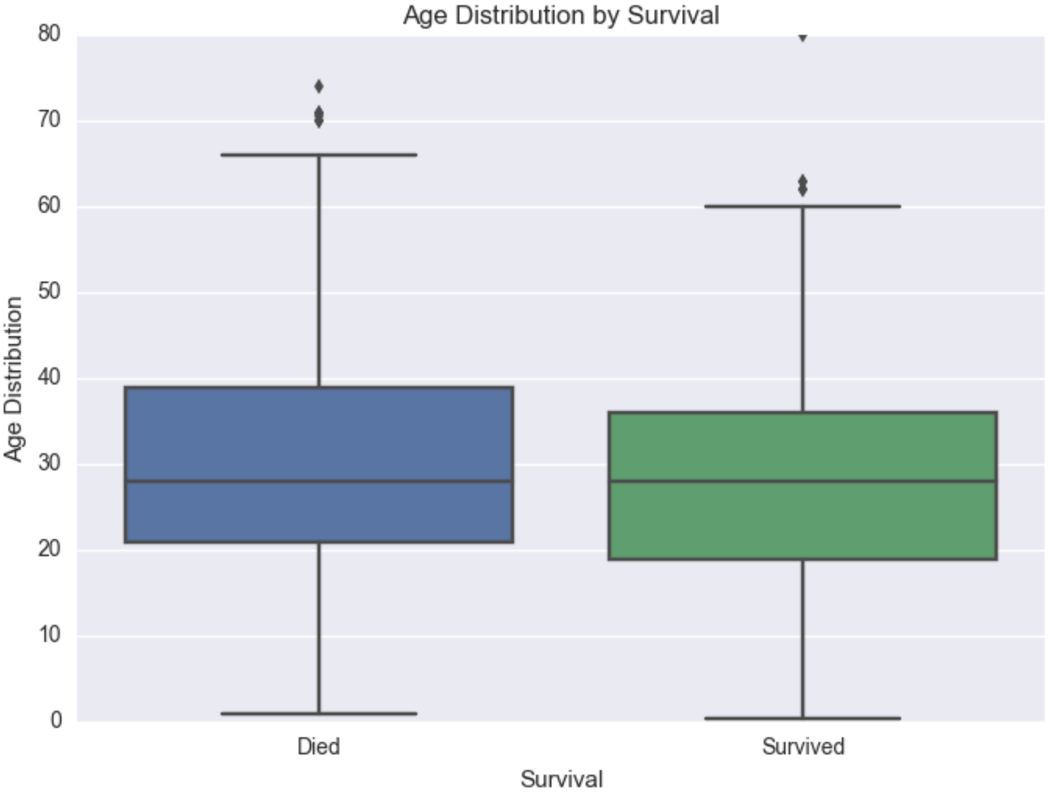
Column Name:	***Amount of Missing Data Points:***
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

Methods to Solve Question:

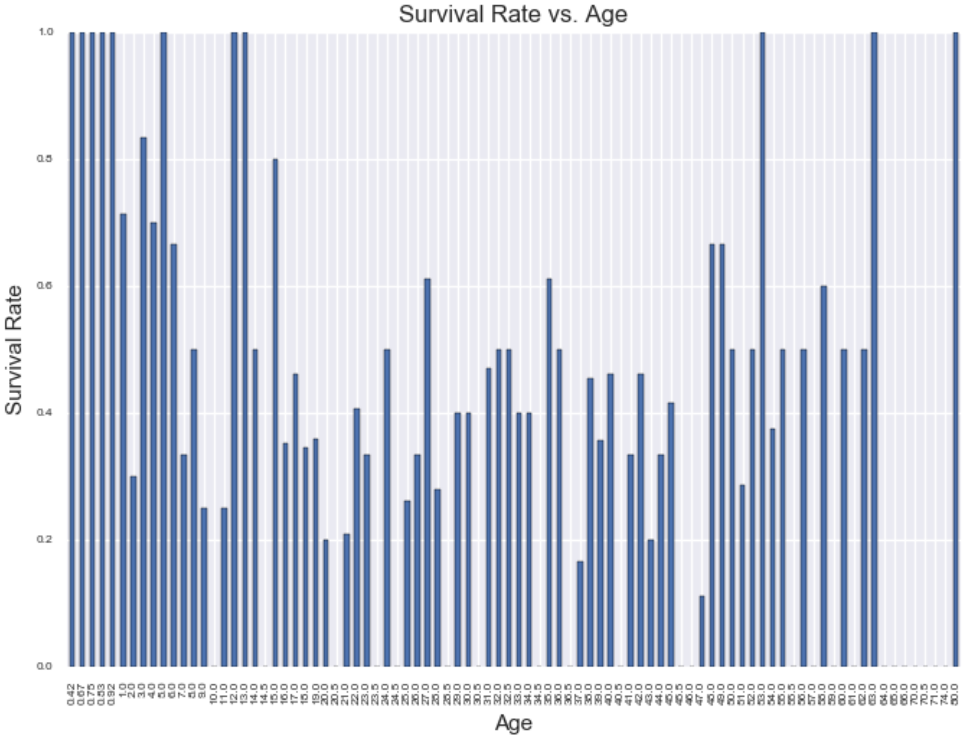
1. I first reviewed the distribution of age data to get a feeling for the age distributed passenger counts alone. I found that there was a lot is more data for passengers ages 18-37.



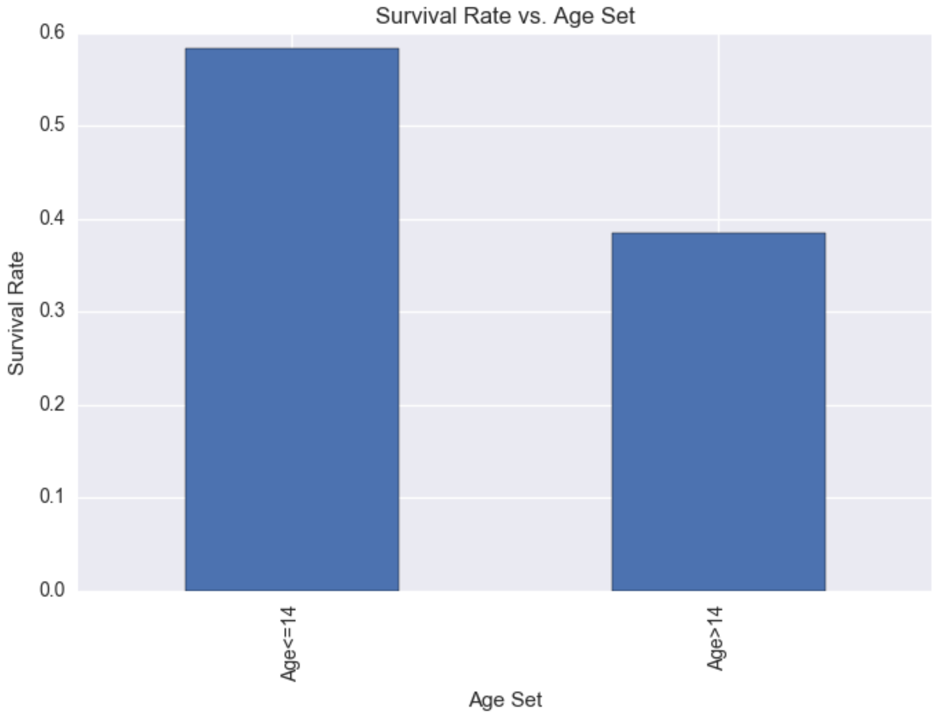
2. I then reviewed the boxplot of age distribution by survival and found the overall boxplots to be similar, with those who survived to overall be a bit younger than those who died.



3. Next, I looked at Survival Rate vs. Age in a bar distribution plot. This gave me an overall view of survival rates throughout all ages. It seems that younger children survived more but there were also less child data points than adult, which could screw the data.

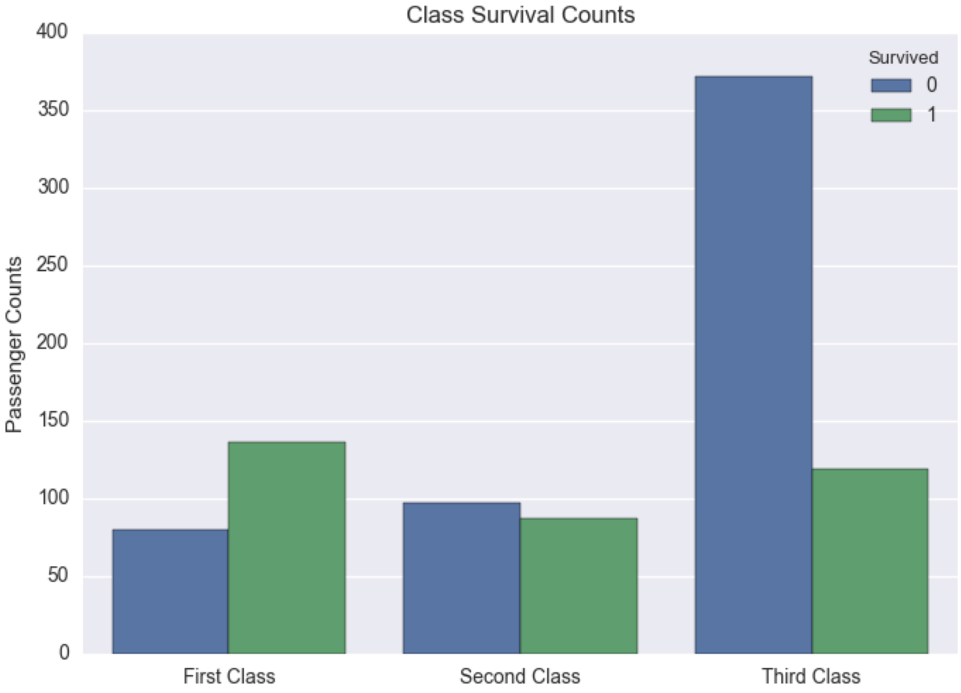


4. I then split the data between children (<=14 years old) and adults (>14 years old) and created a graph of survival rates. It shows that children were more likely to survive than adults.



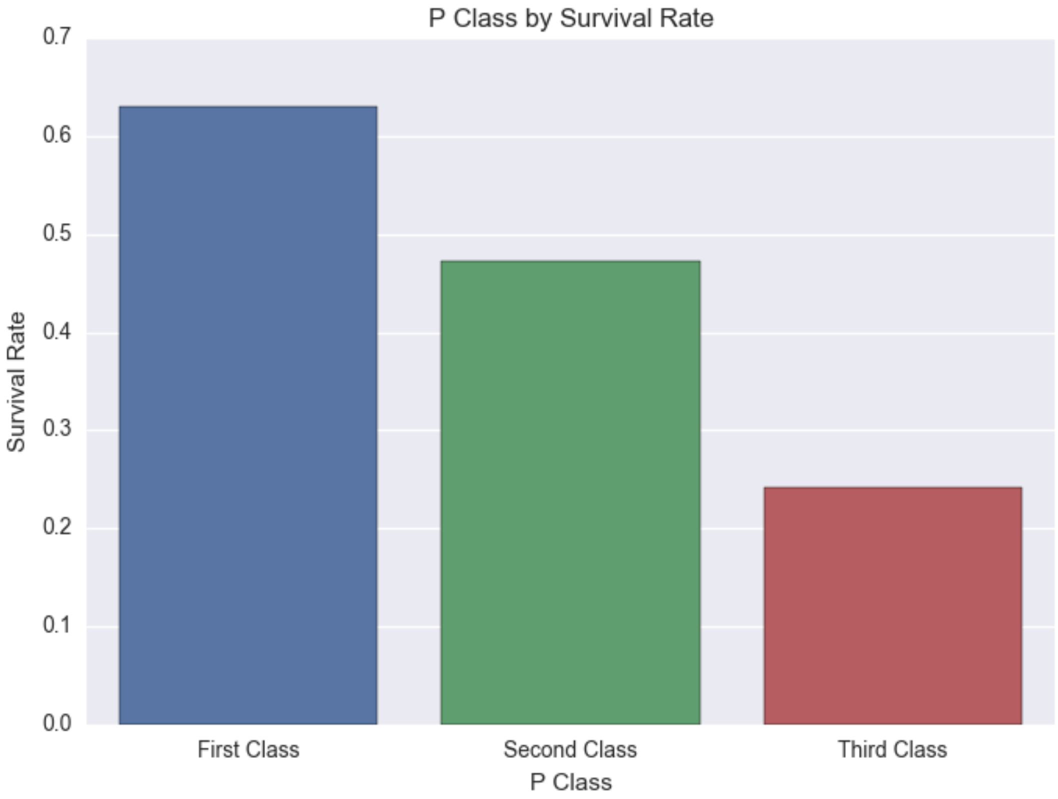
Ages:	***Survival Rate:***
Age<=14	0.584416
Age>14	0.384615

5. I then looked at the survival passenger count for each class separately to get a feel for any trends I might see. I definitely found that third class passengers had many more deaths than second and first class passengers. (0 = Dead, 1 = Alive)



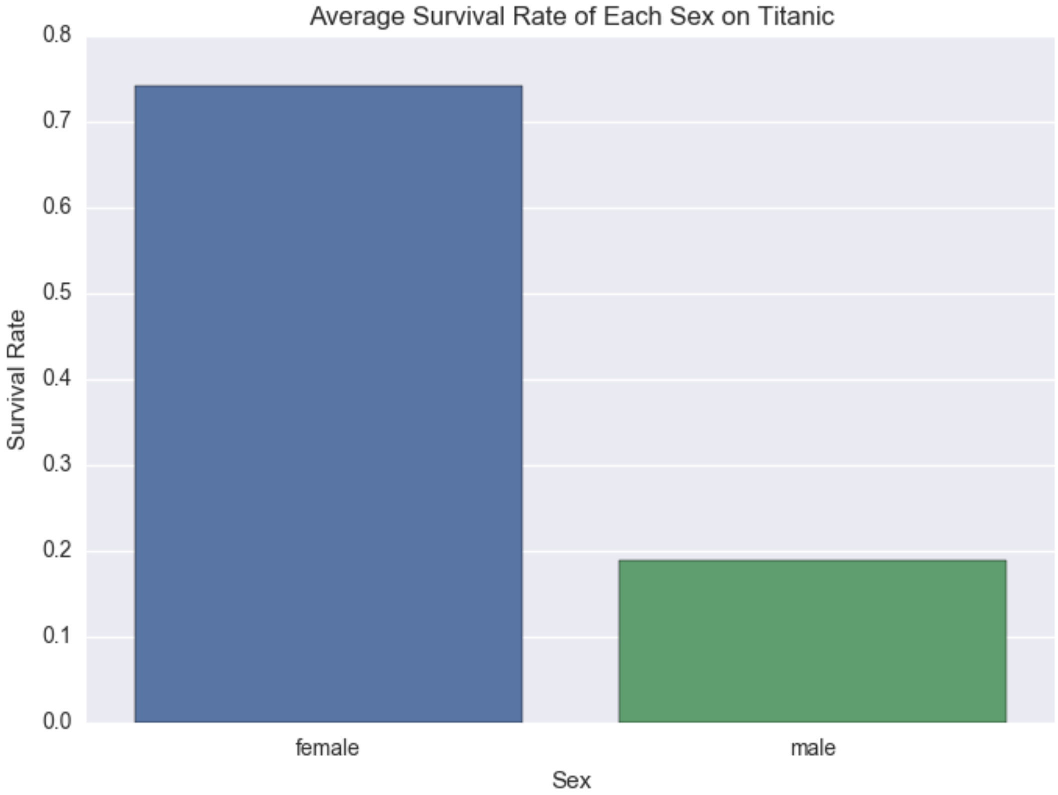
Pclass:	***Survived:***	***Died:***
1	136	80
2	87	97
3	119	372

6. I then looked at survival rates of each Pclass. I found the higher your Pclass the more likely you were to survive.



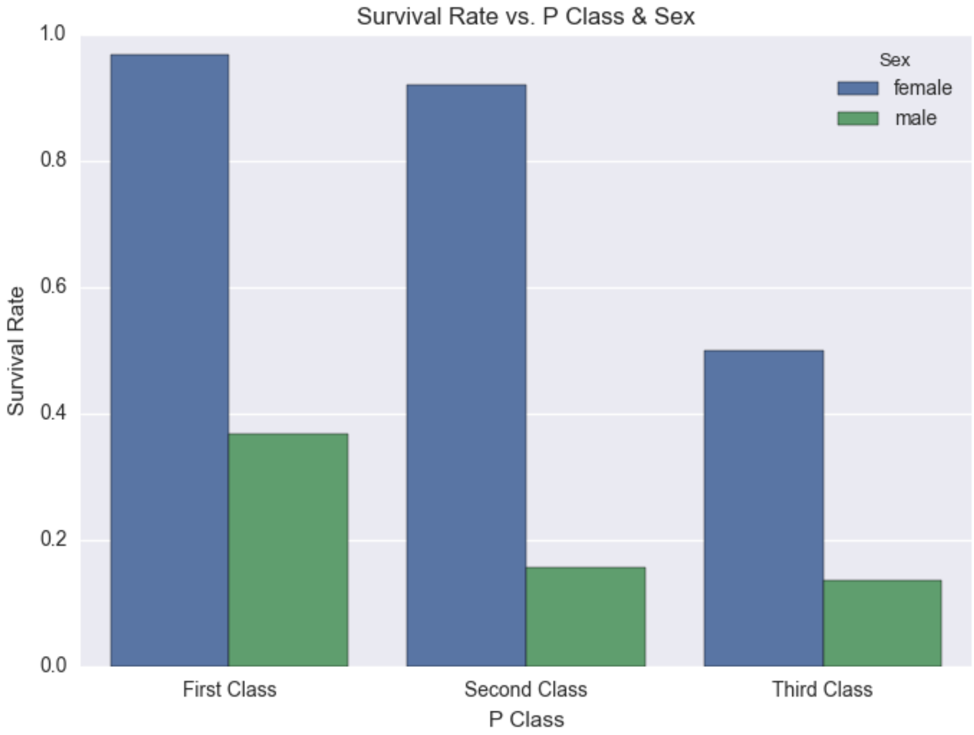
Pclass:	***Survival Rate:***
1	0.629630
2	0.472826
3	0.242363

7. I also looked at the average survival rate of females compared to males. I found females were about 4 times more likely to survive than males.



Sex:	***Survival Rate:***
Female	0.742038
male	0.188908

8. I then pushed a little deeper, with my first project reviewer’s encouragement, to see if there was a difference between Pclass vs. female and male survival rates. I found that female survival rate was still a lot larger than male survival rate in all Pclass levels.



Pclass:	***Sex:***	***Survival Rate:***
1	female	0.968085
	male	0.368852
2	female	0.921053
	male	0.157407
3	female	0.500000
	male	0.135447

9. Finally, with the encouragement of my first reviewer, I calculated total passenger count of survival vs. Pclass and Sex. It is very interesting that a female in third class survived based upon the probability of a flip of a coin while in first class only 3 females out of 94 females died.

Pclass:	***Sex:***	***Survived:***	***Died:***
1	female	91	3
	male	45	77
2	female	70	6
	male	17	91
3	female	72	72
	male	47	300

Conclusions:

1. Females on average survived more than males especially in Upper Pclasses.

2. The higher your Pclass the more likely you were to survive.

3. Children (<=14 Years Old) were 1.5 times more likely to survive.

Limitations of the Data:

- Since I had to omit 177 data points from the data it could have skewed the age data analysis. For example, if all the data I was missing were young children that died, it could lead to the conclusion that you were more likely to die if you were <= 14 years old.
- I cannot prove causation from the correlations in the data. I would like to state that females, higher class passengers and children survived more since they were loaded into the life rafts first. I could potentially prove this if I had the loading time of each passenger into life-rafts or if they never got into a life raft at all.
- I did not use any other statistical test, which would have shown the true significance of my findings. I could have used a z-test to show the probability of a female of a certain age surviving or dying.
- I would have loved to have information about the crew of the titanic to understand their survival rate as compared to passengers.
- All of my conclusions could have been incorrect since I only have 891 passenger data points while there were 1,317 passengers on the actual titanic when it sunk. If I had the information of the rest of the passenger, my conclusion would be more exact and valid. (ref: http://www.titanicfacts.net/titanic-passengers.html)

Questions for grader:

- I am trying to create a new column in the titanic_data_df dataframe named survival, which is the string “Survived” when the column Survived is 1 and “Died” when the column Survived is 0. Do you know a technique to do this?