Project 2

Titanic data set is used for this project

Questions:

1.Is fare related to passenger class? (In other words does the fare increase for higher passenger classes(first class)?

2.Is there any relationship between passenger class and the number of survivals for each gender?

(In other words were people in higher classes (first, second then third) more likely to be survived?)

3.Is there any relationship between passenger class and the number of survivals for each gender?

Data cleaning:

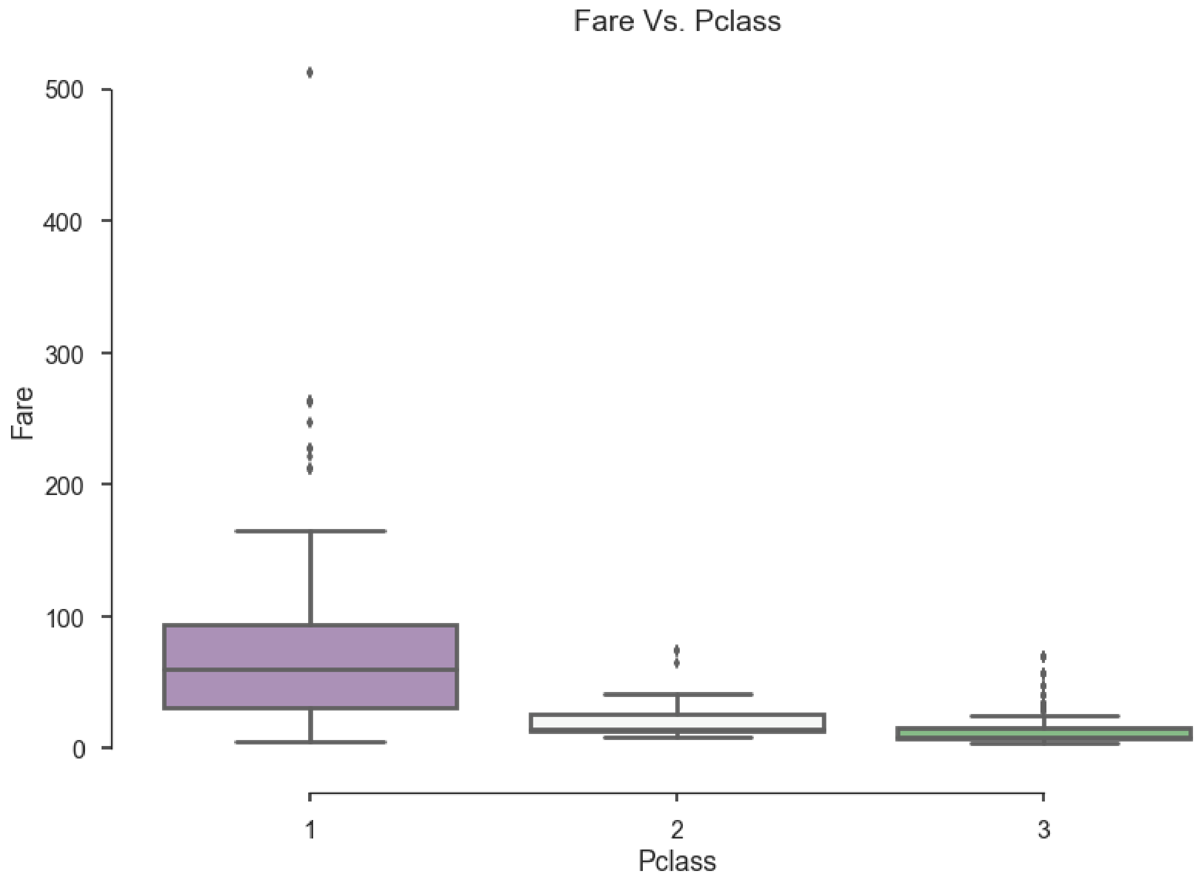
First I converted the type of data from string to convenient data types in some columns. The survival data is converted from string to Boolean type, the data in age and fare column is converted from string to float type.

As part of data cleaning, I got the prefixes of the people’s name or actually their titles and compared it with the sex column to see whether they are compatible or not. If the prefixes like Mr, Mrs, Miss, Master,… do not match the gender of the people, the code will throw an error. In this data set all the titles match the gender data, so there is no need for cleaning in this part.

As another task of data cleaning, I noticed that for some rows, the data in fare column was equal to zero. At first I supposed that maybe it is for crew members but after reading the Udacity forum I figured out that this dataset doesn’t contain information about crew members, so I decided to eliminate these data rows with fare value equal to zero to prepare my dataset for investigating the first question.

Question 1:

For analyzing the relationship between passenger class and fare we plotted boxplots for different passenger classes and compared them with each other. For these boxplots we removed the rows with fare values equal to zero to get meaningful results.



These boxplots show that according to our expectation, for lower passenger classes (third class and then second class)the median is smaller than the median for higher passenger classes(first class).

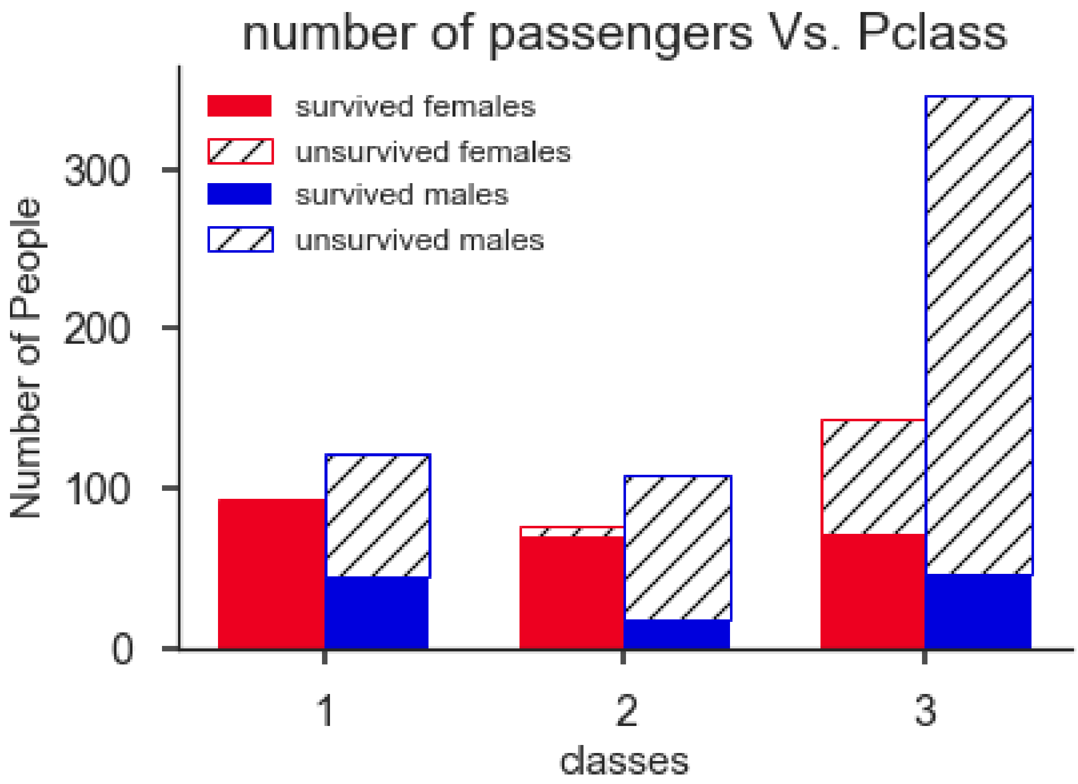
Conclusions:

In original dataset, for imputing those data rows in which fare value is equal to zero, we can take the median of the fare for each passenger class and replace zero value with those values. Because clearly the median of fare values is larger for higher passenger classes and vice versa.

Also these box plots shows that the variation of the fare value is much larger for first class, and of course there is an outlier in the data. (The fare value which is higher than 500)

Question 2:

Now for investigating the relationship between the passenger class in each gender and the survival rate, the stacked bar chart for different classes and different genders has been plotted.



This bar chart shows that the number of survived female passengers is higher in first class, which is what we expected. About male passengers, we calculated the percentage of survivals.

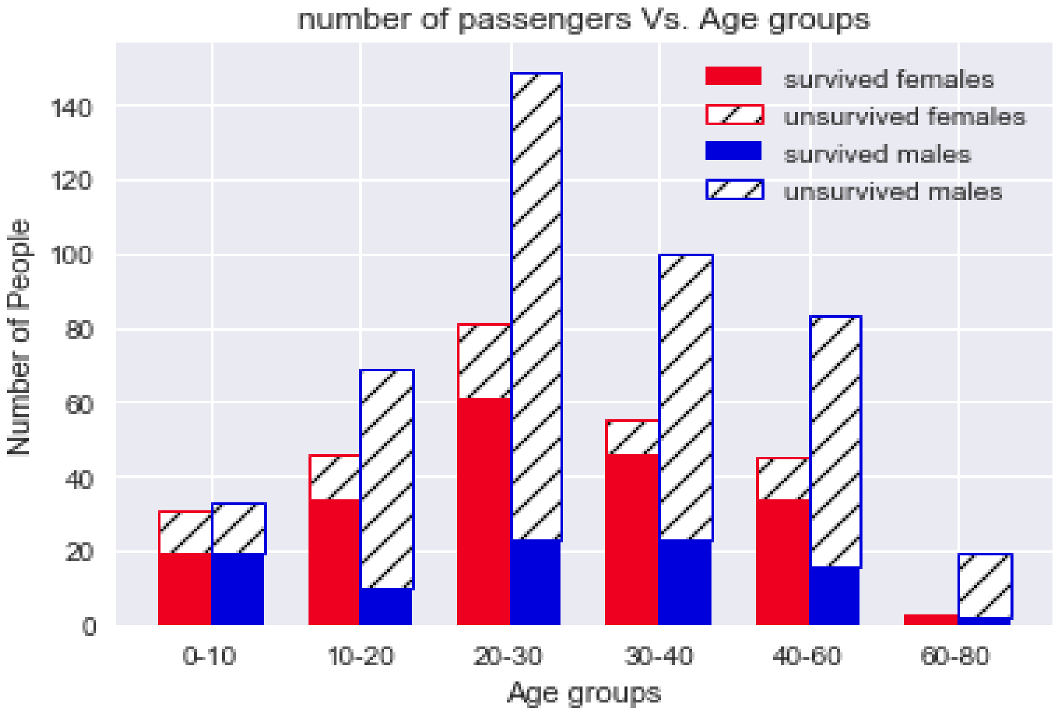
The following table shows the percentage of survival (It is calculated in cell 14)

classes	female	male
1	0.968	0.369
2	0.921	0.157
3	0.5	0.135

The result of this table and the bar chart shows that for each gender by increasing the passenger class(Actually lower passenger classes) the percentage of survival would decrease. This conclusion is logical because those passengers with higher socio-economic class were probably the first passengers who use the rescue boats. Also if we compare different genders we can see that the rate of survival among female passengers were much higher. This conclusion is logical as well, since female passengers had priority in using the rescue boats.

Question 3:

For realizing the relationship between the age and survival we plotted the stacked bar chart to show the number of survived and unsurvived people for different age groups for each gender.



With this bar chart we cannot make a good conclusion about the relationship between age and survival for each gender type. The only conclusion that is apparent is that male children under the age ten had the most priority among the other age groups in male passengers.

So the percentage of survived passengers for each age group and each gender type is calculated and presented in the following table (It is calculated in the last cell of the code)

Age groups	female	male
0-10	0.613	0.576
10-20	0.739	0.145
20-30	0.753	0.154
30-40	0.836	0.23
40-60	0.755	0.193
60-80	1.0	0.102

According to this table it appears that all the females in the age range of 60-80 were survived. But if we consider the bar chart, the height of the chart is shorter for the people in this group. We printed out the number of passengers in this group and only three people were in this group. So we cannot make a definite conclusion about whether old females had priority to be survived. Because three people in this group is not a large number to draw any conclusion about their survival.

```
Sex     Age_group  survived
female  0-10         19
        10-20        34
        20-30        61
        30-40        46
        40-60        34
        60-80         3
male    0-10         19
        10-20        10
        20-30        23
        30-40        23
        40-60        16
        60-80         2
```

If we consider male passengers, we can see that the male children had priority to other age groups of male passenger. Another conclusion about the table is that in both genders, passengers between the age 30-40 had the second ranking in survival rate.

Also in male passengers, the passengers with the age ranging from 60 to 80 had the least rate of survival. This can be due to the fact that the old males were the last passengers who used rescue boats.