1. How many missing value in the dataset?

I wrote a nan_count function to see how many missing value in dataset.

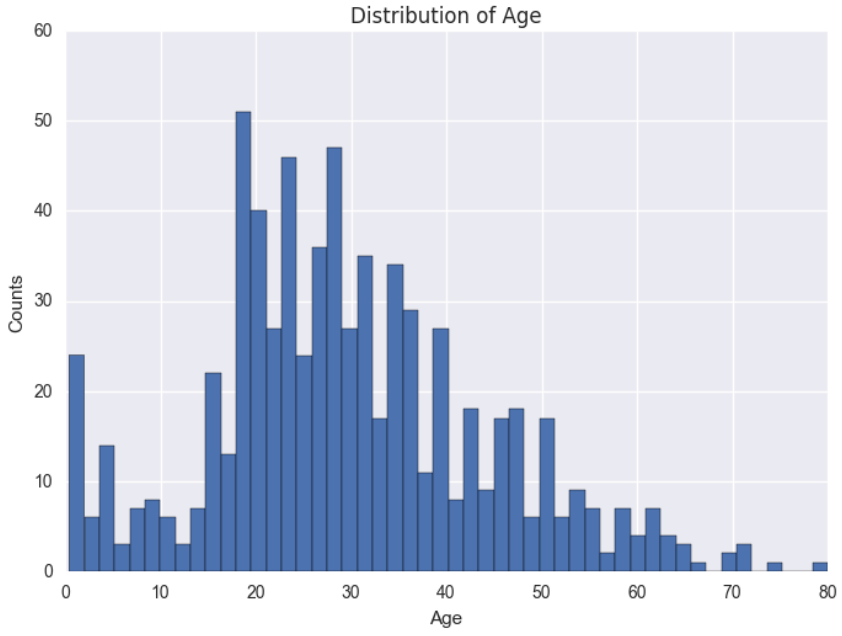
There are total 866 missing value in titanic dataset. 177 from “Age”, 687 from “Cabin”, 2 from “Embarked”.

2. How is the descriptive statistics about the dataset?

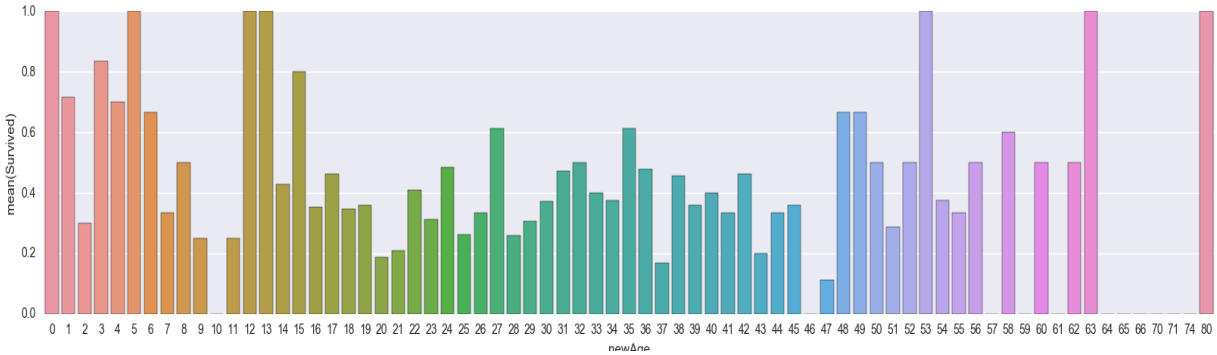
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

We have 891 passenger data to analyze. The total survive probabilities is 0.384. Age average is about 30.

3. <u>The distribution of “Age”, we want to know that how’s the average age of titanic passenger. After that, we can see that if age will affect the survived chance.</u>



It seems that the distribution of “Age” is approximately normal distribution. But we can see that we have more children than the elders. Ages above 50 are not counts over 10.



This graph is the mean (survived) group by “Age”, we can see that all adult’s chance to live are not over 0.6. Only children and elders have higher chance to live.

4. We all have seen the movie “Titanic”. In the movie, we know that there are not enough life boats on titanic. So they decided to let the women and children have the priority to get on the life boats. We can analyze the dataset to check if women and children really have the higher probability to survive.

First, I wrote an Age_series function. Use the age column to create a series contain "adult", "children", "NaN", which could easily see the numbers of children and adult There are 113 children and 601 adult on titanic. Then I add a new (age_range) in dataframe in order to see the survived probabilities between adult and children. There are many factors may affect the probability to survived. Such as “Pclass”, ”Age”, “Sex”, “Embarked”.

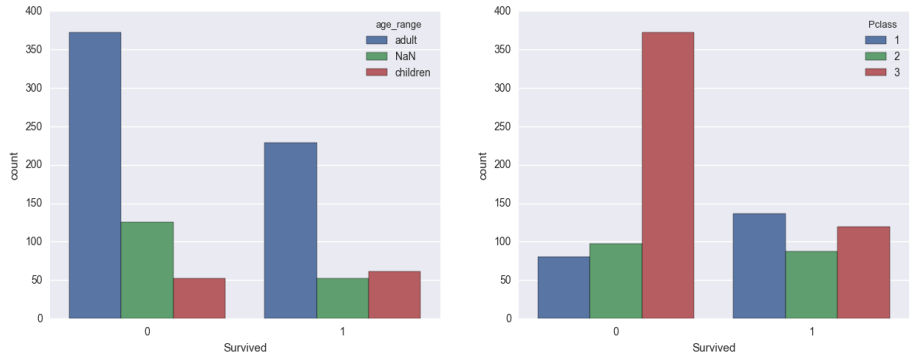
```
age_range
NaN         0.293785
adult        0.381032
children    0.539823
```

Based on age range. It seems that children has higher probability to survive.

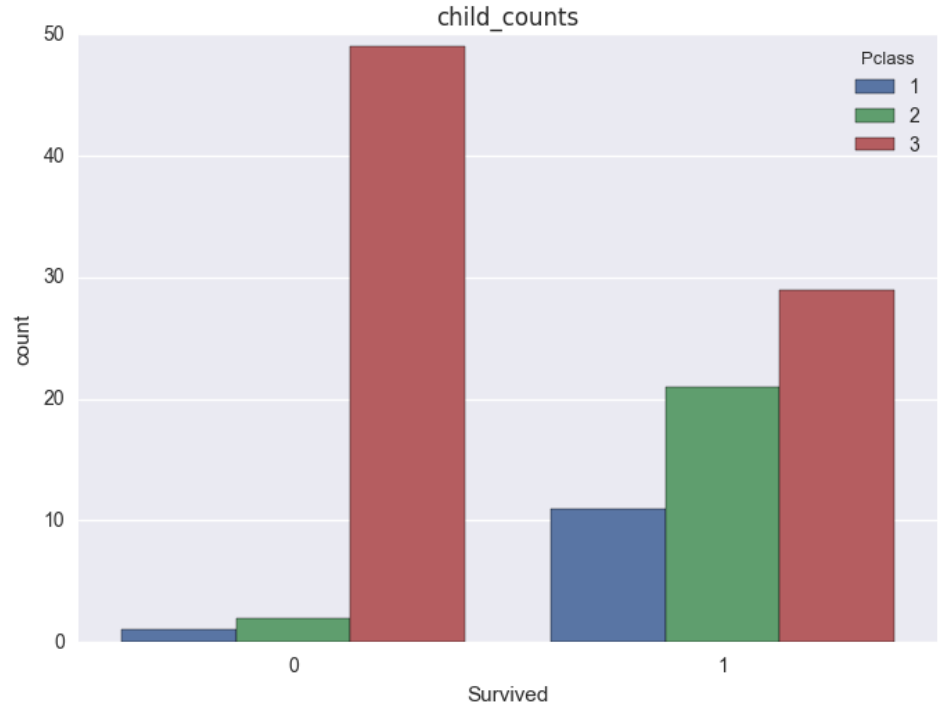
```
age_range  Pclass
NaN        1        0.466667 (14/30)
           2        0.363636 (4/11)
           3        0.250000 (36/134)
```

```
adult      1        0.637931 (111/174)
           2        0.413333 (62/150)
           3        0.202166 (56/277)
children   1        0.916667 (11/12)
           2        0.913043 (21/23)
           3        0.371795 (29/78)
```

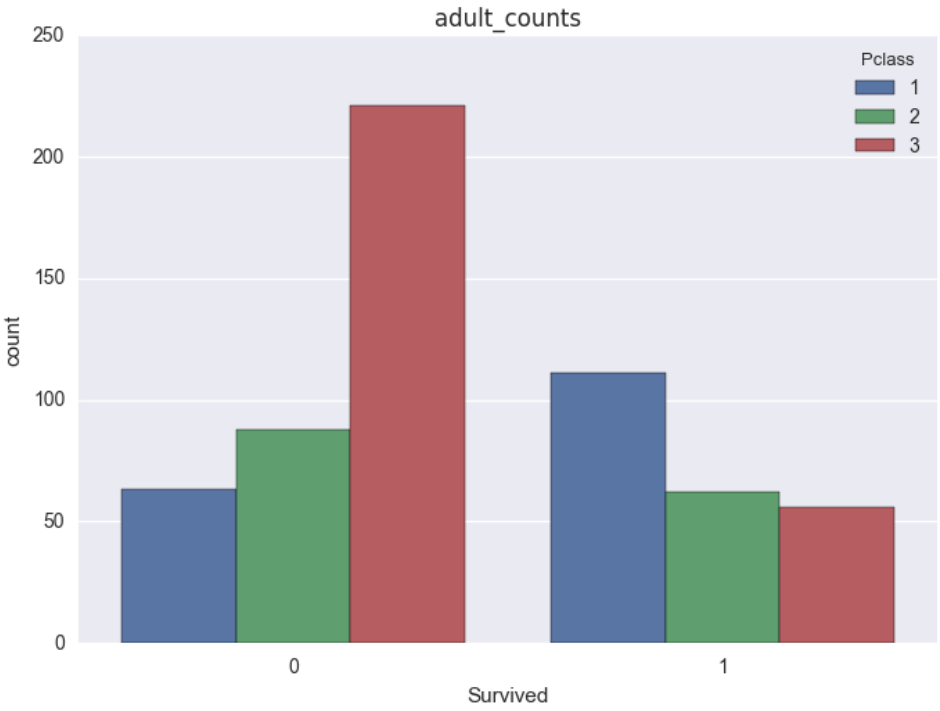
We can clearly see that in all age range, higher the Pclass you took, higher the survived probability you got. It's quite true, but I am so surprise that the survived probability of children in class 3 is much less than other class, even lower than the adult in class 2! I guess that maybe the room for class 3 is near the bottom of titanic, causing that they don't have much time to escape.



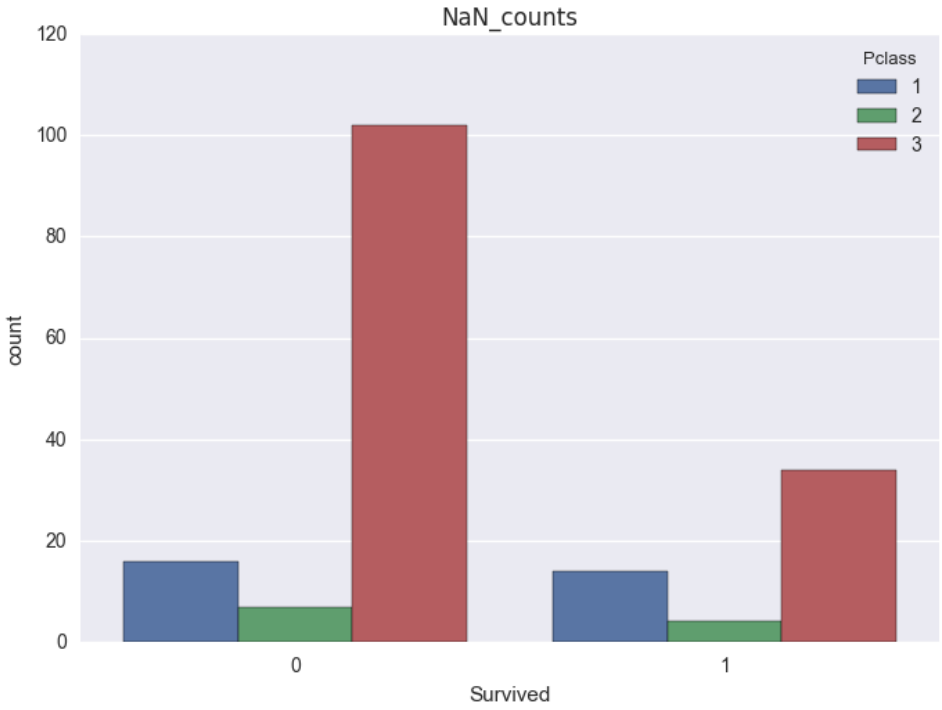
Form the graph above, I found that there is some mistake I've made. From previously graph about age and survived probability. It seems that children would have higher chance to live. But by the left survived counts by age range, we can see that the death counts of children and live counts of children are almost even.



We can see much clearly by this graph, the deaths counts of class 1 and 2 are much less than class three. After those graph, we can conclude that children will have higher chance to live, but you have to pay for higher class first.

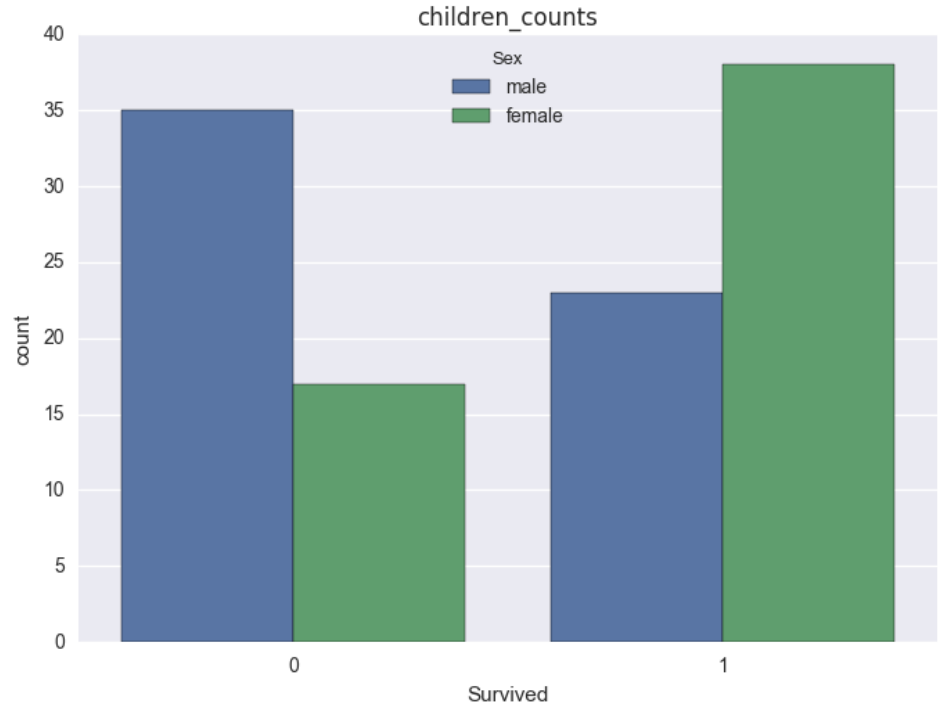


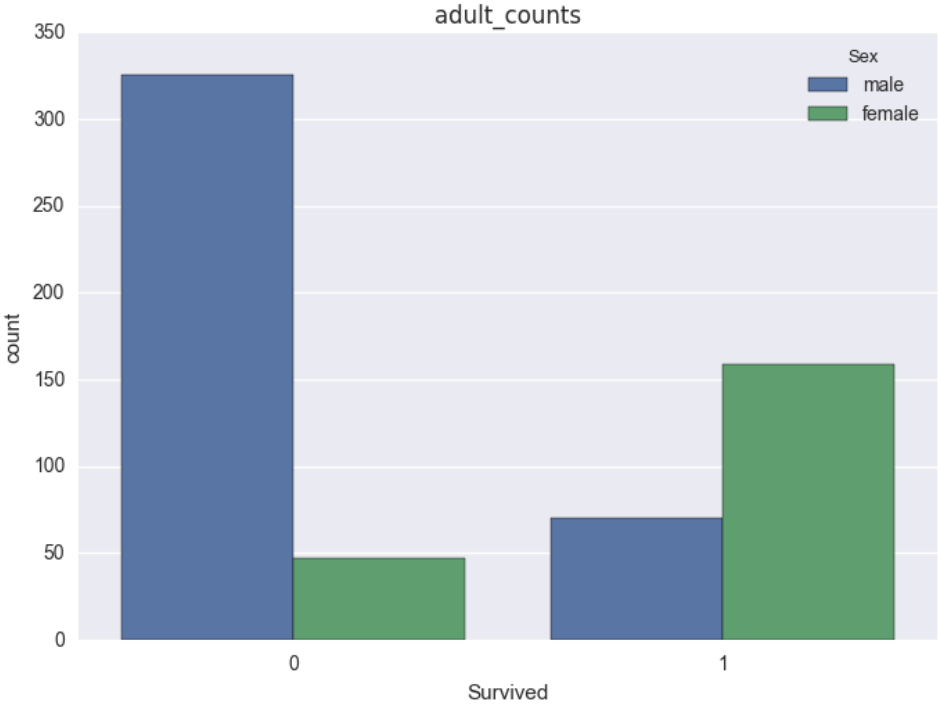
Most adult in class 3 have lower chance to live too. But compare to children, adults in class 1 or 2 are not have that higher chance to live.

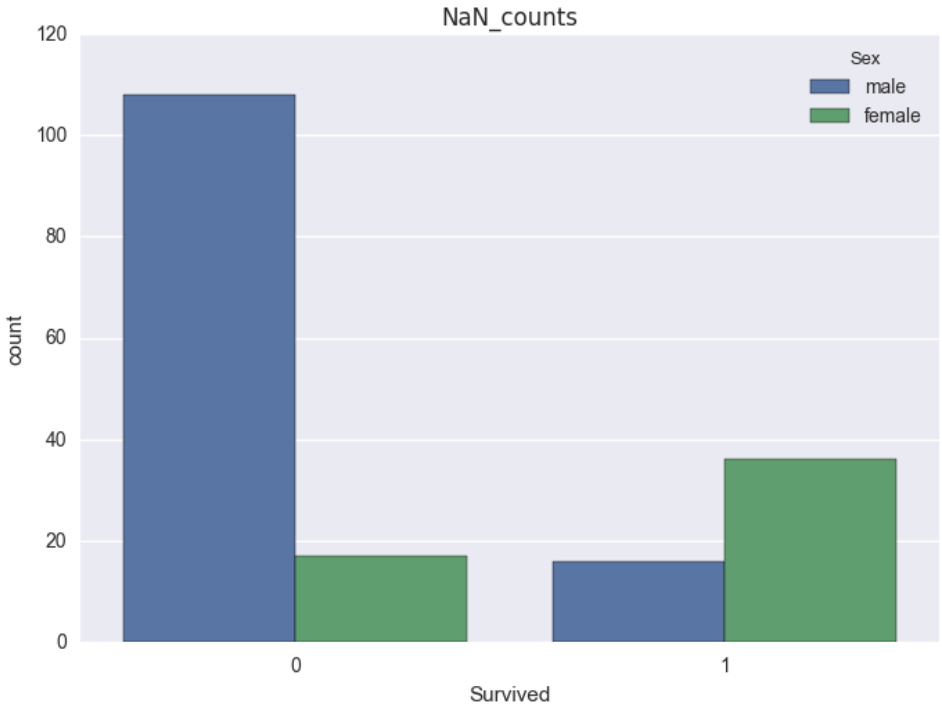


We can’t know for sure that age affect the chance to live since this graph is about missing value in age. But we still can see that class 3 really have a lower chance to live than other class.

```
age_range  Sex
NaN        female    0.679245 (53/36)
           male      0.129032 (16/124)
adult      female    0.771845 (159/206)
           male      0.177215 (70/395)
children   female    0.690909 (38/55)
           male      0.396552 (23/58)
```


children_counts
Sex
male
female
40
35
30
25
20
15
10
5
0
count
0
1
Survived


adult_counts
Sex
male
female
350
300
250
200
150
100
50
0
count
0
1
Survived



By those graphs about counts of each age range, We can see that female really has higher survived probability than male in all range. Especially in adult, survived probability of man is only 0.177. So the movie was right, there were so many gentleman on titanic!

5. From the results above, we can see that the higher class you took, the higher chance you live. So let's group by pclass and see the result.

Pclass	Fare	Age	Survived
1	84.154687	38.233441	0.629630
2	20.662183	29.877630	0.472826
3	13.675550	25.140620	0.242363

As my expectation, higher class you took, the higher chance you live And the age is quite reasonable, older people are much more affordable to higher class.

6. Strange about the "Fare"

When analyzing the descriptive statistics, I am curious about the fare because the minimum value of fare is zero. We know that the leading man "Jack" in the titanic movie, winning the titanic ticket by gambling! Let's see the how many zero "fare" ticket in the dataset.

	PassengerId	Survived	Pclass	Name	Sex	Fare	Embarked	age_range
179	180	0	3	Leonard, Mr. Lionel	male	0.0	S	adult
263	264	0	1	Harrison, Mr. William	male	0.0	S	adult
271	272	1	3	Tornquist, Mr. William Henry	male	0.0	S	adult
277	278	0	2	Parkes, Mr. Francis "Frank"	male	0.0	S	NaN
302	303	0	3	Johnson, Mr. William Cahoone Jr	male	0.0	S	adult
413	414	0	2	Cunningham, Mr. Alfred Fleming	male	0.0	S	NaN
466	467	0	2	Campbell, Mr. William	male	0.0	S	NaN
481	482	0	2	Frost, Mr. Anthony Wood "Archie"	male	0.0	S	NaN
597	598	0	3	Johnson, Mr. Alfred	male	0.0	S	adult
633	634	0	1	Parr, Mr. William Henry Marsh	male	0.0	S	NaN
674	675	0	2	Watson, Mr. Ennis Hastings	male	0.0	S	NaN
732	733	0	2	Knight, Mr. Robert J	male	0.0	S	NaN
806	807	0	1	Andrews, Mr. Thomas Jr	male	0.0	S	adult
815	816	0	1	Fry, Mr. Richard	male	0.0	S	NaN
822	823	0	1	Reuchlin, Jonkheer. John George	male	0.0	S	adult

There are 15 people who paid nothing for the fare. They are all male and all from Embarked "S". Sadly, only one survived. They might use their all luck when winning the tickets.

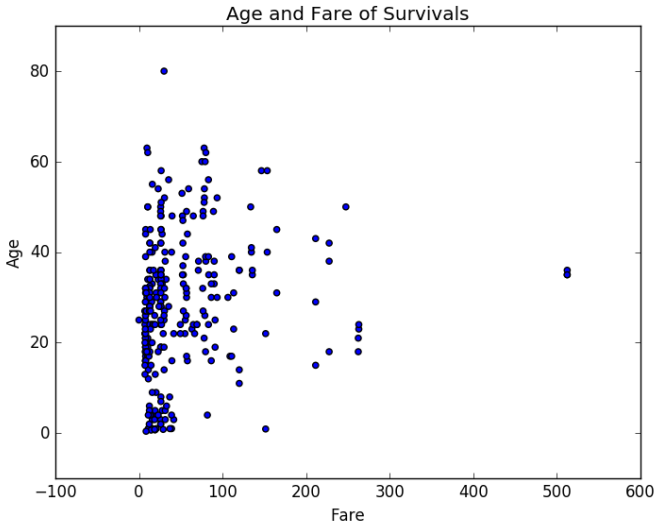
7. After the disaster, still many people survived. To check if the leading woman in Titanic “Rose”, were survived. I wrote a function to do it!

Sadly, it return

```
Aks, Mrs. Sam (Leah Rosen)
```

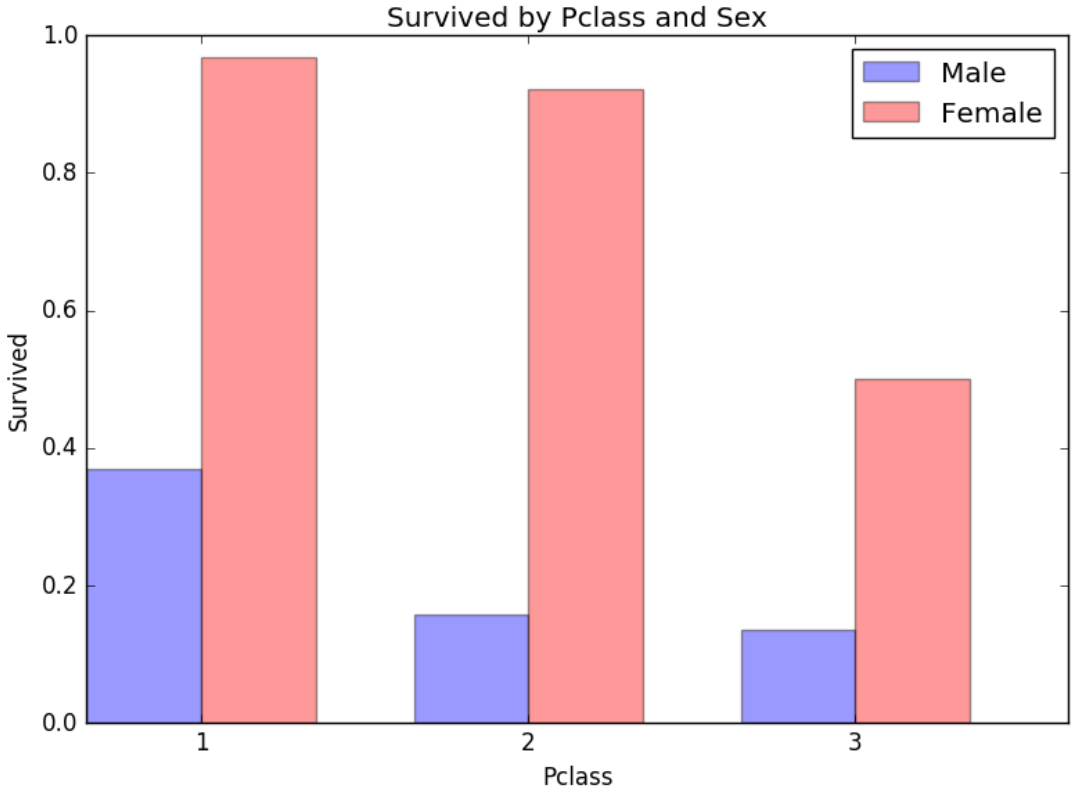
So we can make sure that “Rose” only show in the movie.

8. Graph about age and fare:



This graph is age and fare of survivals. The correlation between age and fare is 0.096, which is pretty low, indicating that age and fare have little correlation.

9. Survived by Pclass and Sex



By this figure, we can easily see that female has much higher chance to survive than male. And people in the class 1 have higher chance to survive than other class.
This figure is reviewed by [matplotlib bar_chart.demo.]

10. Conclusion by statistics!

As we keep saying that the higher class you took, the higher chance you live. So we need to use statistic way to prove our hypothesis. Here's our hypothesis

Null hypothesis: pclass and chance to survive are independent
Alternative hypothesis: pclass and chance to survive are dependent

We can use scipy to conduct the chi-square test. To check if the pclass and chance to survive is significant dependent, which indicate that the higher class you took, the higher chance you live.
Results shows that statistic=inf, pvalue=0.0. We can reject our null hypothesis with α level = 0.05.
Finally, we can conclude that, if you were in the titanic, you should buy the first class ticket!