Udacity Data Analyst Nanodegree Project 2

Titanic Data Analyze

Zhongwei Teng

5/18/2016

1.Question

This data set is used for machine learning from disaster. It contains information of passagers on the titanic. Main features are describe below:

PassengerId: Unique Id for each passagers.
Survived: 1 means survived and 0 means not.
Pclass: Pclass is a proxy for socio-economic status (SES) .1st Upper; 2nd Middle; 3rd Lower .
Name: Name of passagers
Sex: Gender
Age: Age
SibSp: Number of Siblings/Spouses Aboard

Parch: Number of Parents/Children Aboard

Ticket: Ticket Number

Fare: Passenger Fare

Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

From the dataset, we can find there are several features may influence survivability which include: Pclass, Sex, Age, Parch, SibSp, Fare. More specifically, I want to analyze the 'Age' and 'Fare'. This project try to figure out the relationship between these features and survivability. Which age or fare have higher chance to be rescue?

On the other hand, we also need to explore the relationship between these features to see if other features affect our result.

2.Clean Data

Our dataset include some null value which can not be used in analyze. The number of null value is shown as belown:

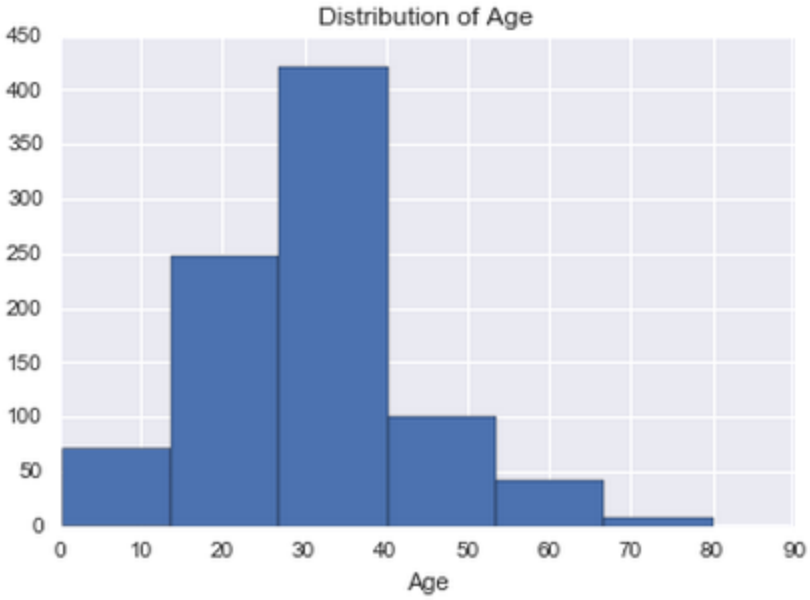
```
There is 891 data
PassengerId exist 0 null values
Survived exist 0 null values
Pclass exist 0 null values
Name exist 0 null values
Sex exist 0 null values
Age exist 177 null values
SibSp exist 0 null values
Parch exist 0 null values
Ticket exist 0 null values
Fare exist 0 null values
Cabin exist 687 null values
PEmbarked exist 2 null values
```

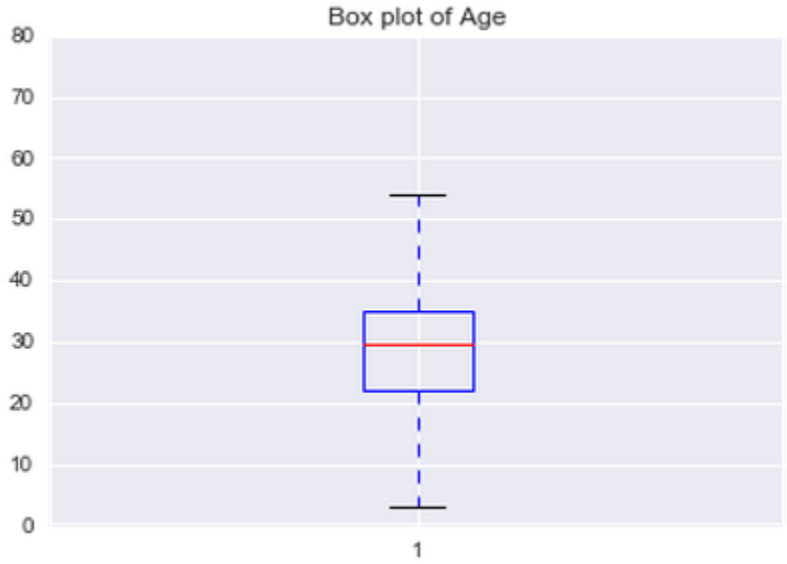
We can see most features are clean data. For ‘Age’, we will replace the null value with the mean of others. For Cabin, we will replace the null with zero. For PEmbarked, since there are only 2 null values, we decide to drop the row directly.

3.Analyze Data

3.1 Age

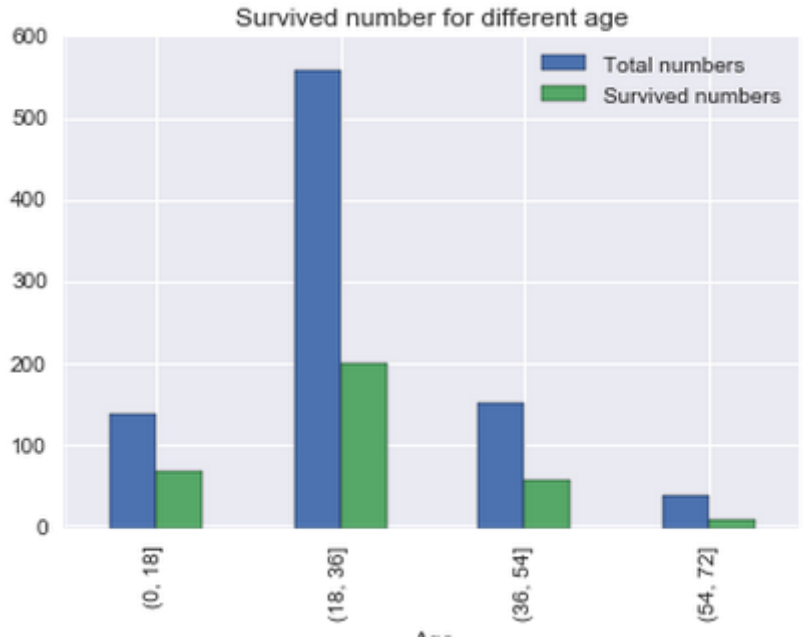
Let’s plot the distribution of age first:





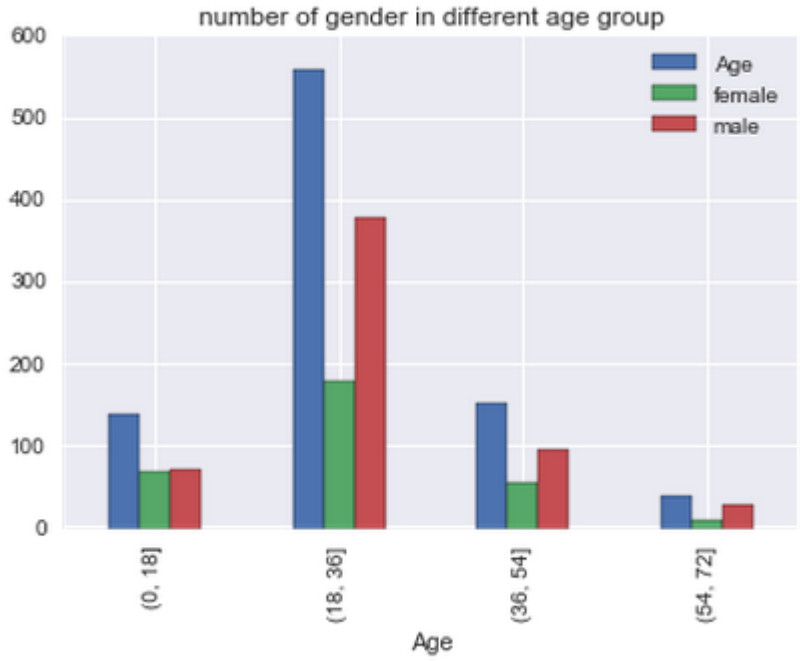
We can observe from the histogram that the distribution of age obey norm distribution, thus we can use t test later. In the box plot, we find our data mainly locate between 22 and 35.

Then we want to know the relationship between age and survival rate. We group our dataset according to different age and plot how many people survived for different group:



	Total numbers	Survived numbers
Age		
(0, 18]	139	70
(18, 36]	557	200
(36, 54]	152	58
(54, 72]	39	11

From this plot, we can see children (age<18) have higher survivability than others when meet such a disaster. But we also curious about if the gender affect our conclusion since gender also can be a very important factor. To solve this problem, we plot a histogram as below:



:

	total	female	male
Age			
(0, 18]	139	68	71
(18, 36]	557	180	377
(36, 54]	152	55	97
(54, 72]	39	9	30

We find for children group, the number of male are almost equal to the number of female. Thus we can say the higher survivability of children is not caused by gender.

To Prove of conclusion that when children have higher survivability that others, we need to proceed p-test

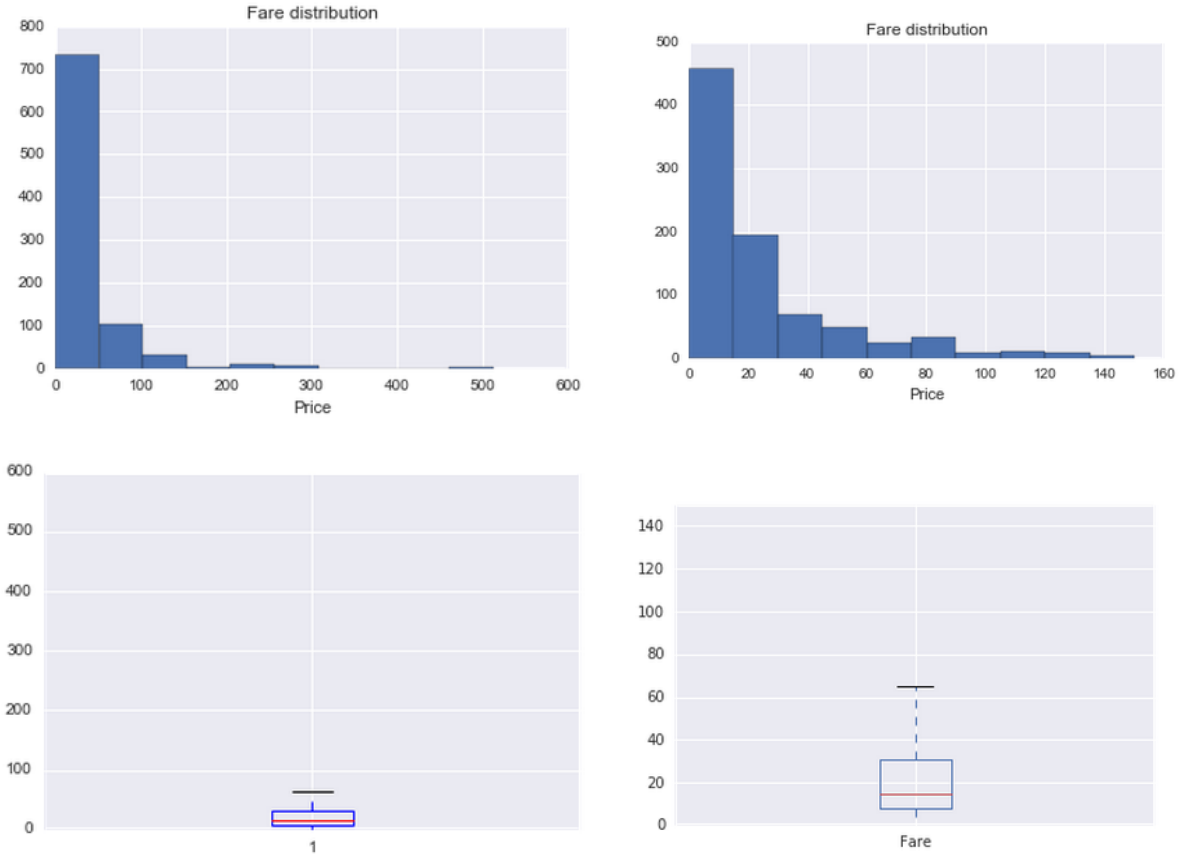
	(0, 18]	(18, 36]	(36, 54]	(54, 72]
Non-survived	69	357	94	28
Survived	70	200	58	11

```
Results of Chi-Squared test on Pclass to Survival.
Does Pclass have a significant effect on Survival?
Chi-Squared Score = 11.5948774937
Pvalue = 0.00890798586255
```

The result of p-test shows that the conclusion is trustful.

3.2 Fare

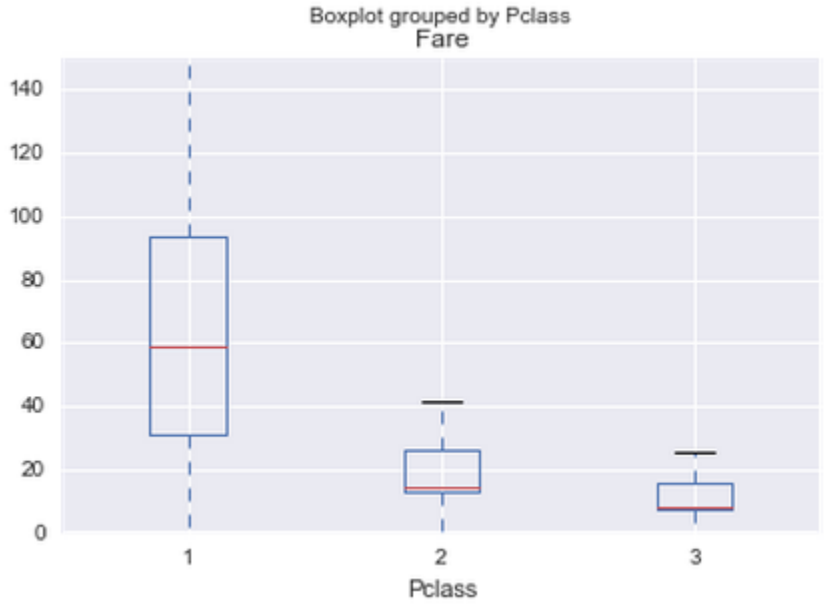
Let's plot the distribution of Fare first.



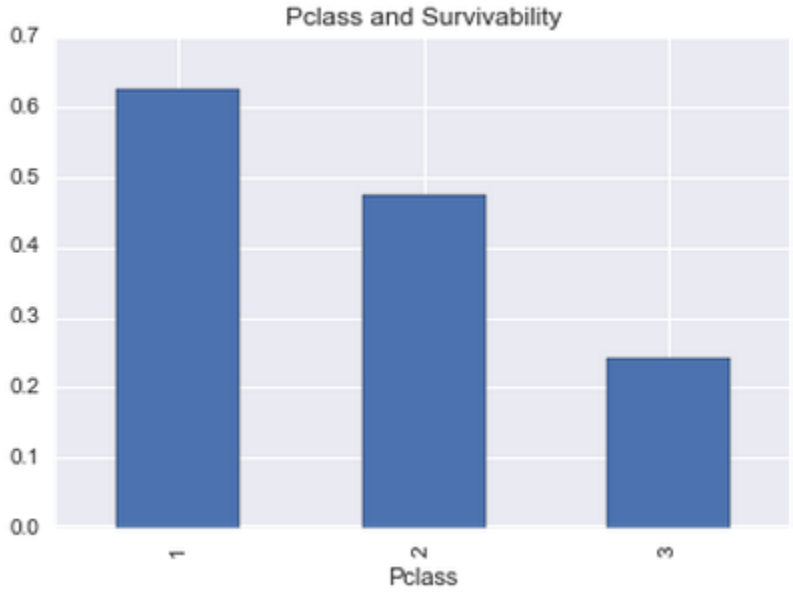
Since there exist some extremely large value of 'Fare', if we plot directly (left side), we can't see it very clear. Thus, I cut some data (greater than 150) to plot a more clear figure(right side.)

We can see unfortunately that the distribution of Fare don't obey norm distribution.

We want to group our fare data to explore its survivability and we also notice that ‘Pclass’ may have strong connection with ‘Fare.’



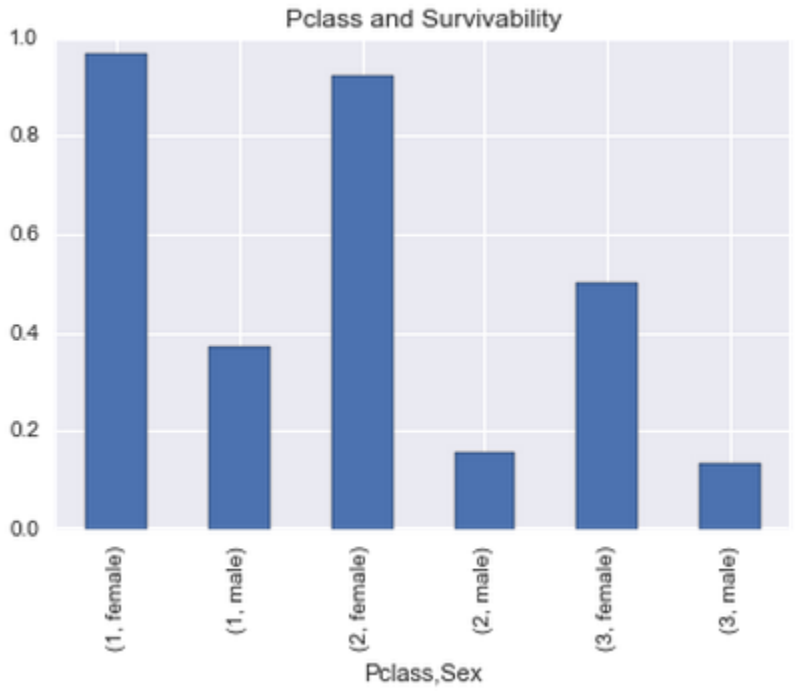
By plotting the box plot, we can see that Pclass and fare has similar tendency. The survival rate for different Fare group may also apply for Pclass. We can use Pclass to group ‘Fare’.



,

This figure shows that higher class will acompany with higher survivability. Combine with figure we plot before, we find higher fare will accompany with higher survivability.

We are also interested in if this difference is caused by gender. So we plot such figure:



		survivability
Pclass	**Sex**	
1	**female**	0.967391
	male	0.368852
2	**female**	0.921053
	male	0.157407
3	**female**	0.500000
	male	0.135447

It shows that higher class always has higher chance to be rescue no matter which gender.

Then we need to proceed p-test to our result.

Pclass	1	2	3
Survived			
0	80	97	372
1	134	87	119

```
Results of Chi-Squared test on Pclass to Survival.
Does Pclass have a significant effect on Survival?
Chi-Squared Score = 100.980407261
Pvalue = 1.18136247855e-22
```

The high score of P-test prove our result is trustful.

4.Conclusion

This project shows that children and rich people have higher chance to be rescue. But it still has some problem.

First, when deal with missing value in ‘age’, I choose to fill it with mean value of others which can cause inaccurate conclusion later.

Second, we can plot 3-dimensional figure to show the relationship of 3 different features and survivability. Analyzing only one or two features is not enough,

Finally, we can more plotting skill rather than just histogram and boxplot. It will makes our conclusion more clear.