Titanic Data Investigation Report

Long Wan, May 20th, 2016

Question

Since it is a dataset regarding Titanic, I was curious about what kind of passengers could survive. Therefore I set my question as what factors affected passenger's survival. It is a big problem. After carefully studying the dataset, I divided the question into several smaller part.

- Did passenger gender affected survival and how?
- Did passenger class affected survival and how?
- Did parch affected survival and how?
- Did passenger age affected survival and how?

By solving these problems, I tried to restore what was happening on board and how people got survived.

Cleaning the data

When firstly getting the data, I saw the head of the data to know about its content. See table 1 as following.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Table 1 Head of data set

Obviously, name and ticket number had nothing to do with survival, so I dropped these two variables first. Then I found that there were NaNs in Cabin column. Knowing the cabin information might be helpful to study survival, because the location of cabin could affect people's escaping path and further affect people's survival rate. However, I was not able to learn from cabin, because I had no access to cabin location. Meanwhile, a large portion of cabin information had been lost. Involving cabin was of no use, thus I deleted the variable simply.

Sex was useful. It would be more convenient if converting string into pure number. So I created a new column called *gender,* filling it with 1 if for male and 0 for female. Then see the head of dataset and I got Table 2 as following, which is cleaner and more readable.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Embarked	Gender
0	1	0	3	22.0	1	0	7.2500	S	1
1	2	1	1	38.0	1	0	71.2833	C	0
2	3	1	3	26.0	0	0	7.9250	S	0
3	4	1	1	35.0	1	0	53.1000	S	0
4	5	0	3	35.0	0	0	8.0500	S	1

Table 2 More readable data

Describe the dataset as showed on Table 3 followed.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Gender
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208	0.647587
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429	0.477990
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400	0.000000
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200	1.000000
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000	1.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200	1.000000

Table 3 Dataset description

Everything seemed OK except Age. Only 714 ages were available and others were NA. Simply deleting those samples without age would cause estimation bias. So we should compared with-age group to without-age group to see if there were different. Their descriptions are showed as follows.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Gender
count	177.000000	177.000000	177.000000	0.0	177.000000	177.000000	177.000000	177.000000
mean	435.581921	0.293785	2.598870	NaN	0.564972	0.180791	22.158567	0.700565
std	250.552901	0.456787	0.763216	NaN	1.626316	0.534145	31.874608	0.459310
min	6.000000	0.000000	1.000000	NaN	0.000000	0.000000	0.000000	0.000000
25%	230.000000	0.000000	3.000000	NaN	0.000000	0.000000	7.750000	0.000000
50%	452.000000	0.000000	3.000000	NaN	0.000000	0.000000	8.050000	1.000000
75%	634.000000	1.000000	3.000000	NaN	0.000000	0.000000	24.150000	1.000000
max	889.000000	1.000000	3.000000	NaN	8.000000	2.000000	227.525000	1.000000

Table 4 Without-age group description

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Gender
count	714.000000	714.000000	714.000000	714.000000	714.000000	714.000000	714.000000	714.000000
mean	448.582633	0.406162	2.236695	29.699118	0.512605	0.431373	34.694514	0.634454
std	259.119524	0.491460	0.838250	14.526497	0.929783	0.853289	52.918930	0.481921
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	222.250000	0.000000	1.000000	20.125000	0.000000	0.000000	8.050000	0.000000
50%	445.000000	0.000000	2.000000	28.000000	0.000000	0.000000	15.741700	1.000000
75%	677.750000	1.000000	3.000000	38.000000	1.000000	1.000000	33.375000	1.000000
max	891.000000	1.000000	3.000000	80.000000	5.000000	6.000000	512.329200	1.000000

Table 5 With-age group description

It seems that means of *Pclass*, *Sibsip*, *Gender, Fare* and *Parch* were different between two groups.

Parch	0	1	2	3	4	5	6
Number	676	118	80	5	4	5	1

Table 6 Number of passenger with different parch

In order to simplify the issue, I explored the relationship between *Fare* and *Pclass*, and I guess that the higher the passenger class was, the more expensive the fare was so that I might consider one variable only instead of considering both. I made a boxplot to show their relationship. See Figure 1 below.

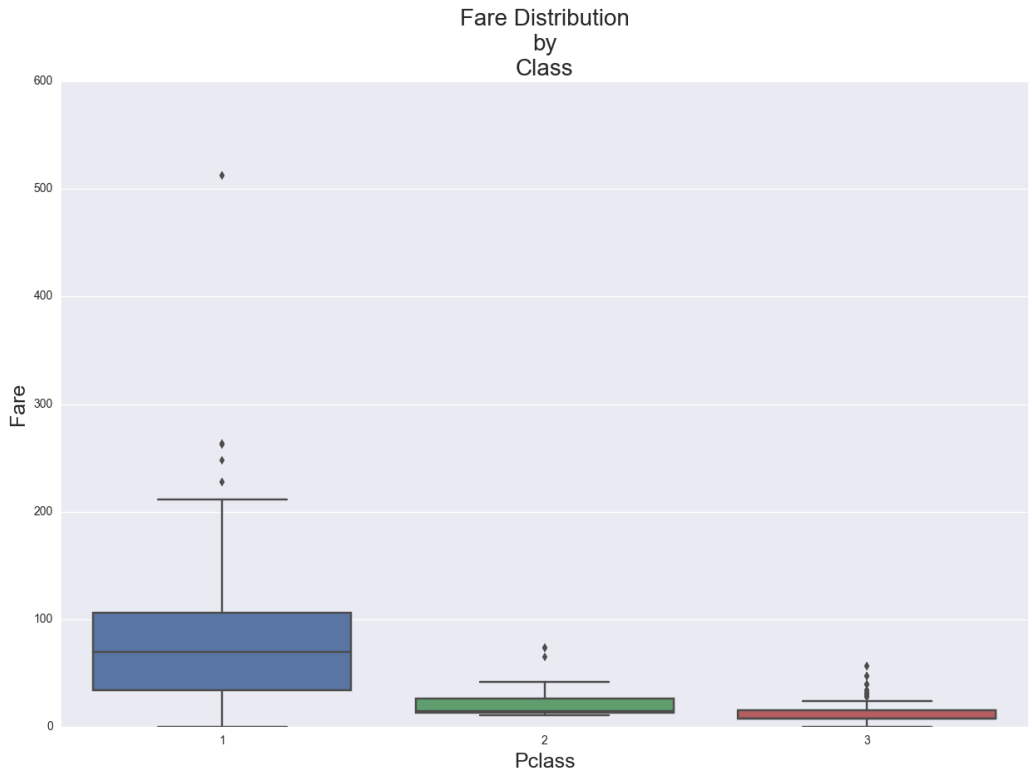

Figure 1 Boxplot of fare and Pclass

From the boxplot we can see that their relationship is almost like what I guessed. We didn't know population variance and assume that sample mean followed normal distribution, so I processed the ANOVA test on three samples for the purpose of knowing whether mean of fare were different. ANOVA results show that three groups have different means of fare at 1% level. It means that there was a different fare for each passenger class.

```
One-way ANOVA P = 1.03137632091e-84
```

Considering passenger class only is more convenient, I left out fare column. Besides, take a glance at the parch information, I found that 213 passengers had parents and children on board and the maximum number was 6 and most lied in 1 or 2. Comparing the survival rates grouped by the right parch number is not necessary, because the number of samples with 3,4,5,6 parch were limited and survival rates reflected by them were meaningless. See the table below showing number of people with different parch.

Thus, I created a new variable called *Parch Flag*, with 1 representing non-zero parch and 0 representing zero parch and take passengers with parents and children as a whole. Also, I created the *SibSp Flag* due to the same reason.

Now I could chi2 test Pclass, Parch Flag, SibSp Flag and Gender for with-age group and without-age grou p to see if there were different. Results are showed below.

	Pclass	SibSp Flag	Parch Flag	Gender
P-value	1.1727232089e-10 ***	0.00427775937203 ***	1.62561739747e-05 ***	0.129316115859

Table 7 Chi2-test results for with-age and without-age group

Two groups are significantly different in *Pclass*, *SibSp Flag* and *Parch* at 1% level. Thus, deleting samples without age information would cause biases since they are quite different and I decided to remain them temporarily.

Finally, I studied the *Embark* Variable. There were two missing. I deleted the samples without valid *Embark* value as two samples would not affect the result. Up to now, I had finished cleaning the data and finally got 889 valid samples and 11 columns, which is clean and quite readable without redundancy.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	Parch Flag	Age Flag	SibSp Flag
0	1	0	3	22.0	1	0	S	1	0	1	1
1	2	1	1	38.0	1	0	C	0	0	1	1
2	3	1	3	26.0	0	0	S	0	0	1	0
3	4	1	1	35.0	1	0	S	0	0	1	1
4	5	0	3	35.0	0	0	S	1	0	1	0

Table 8 Final clean data

Analyzing the data

Survival and Gender

Made the plot to see the number of survived people and not survived people with respect to gender(Figure 2). In fact, the mean of *Survived* column is the right survival rate, so t-tests made followed examine exact survival rate.

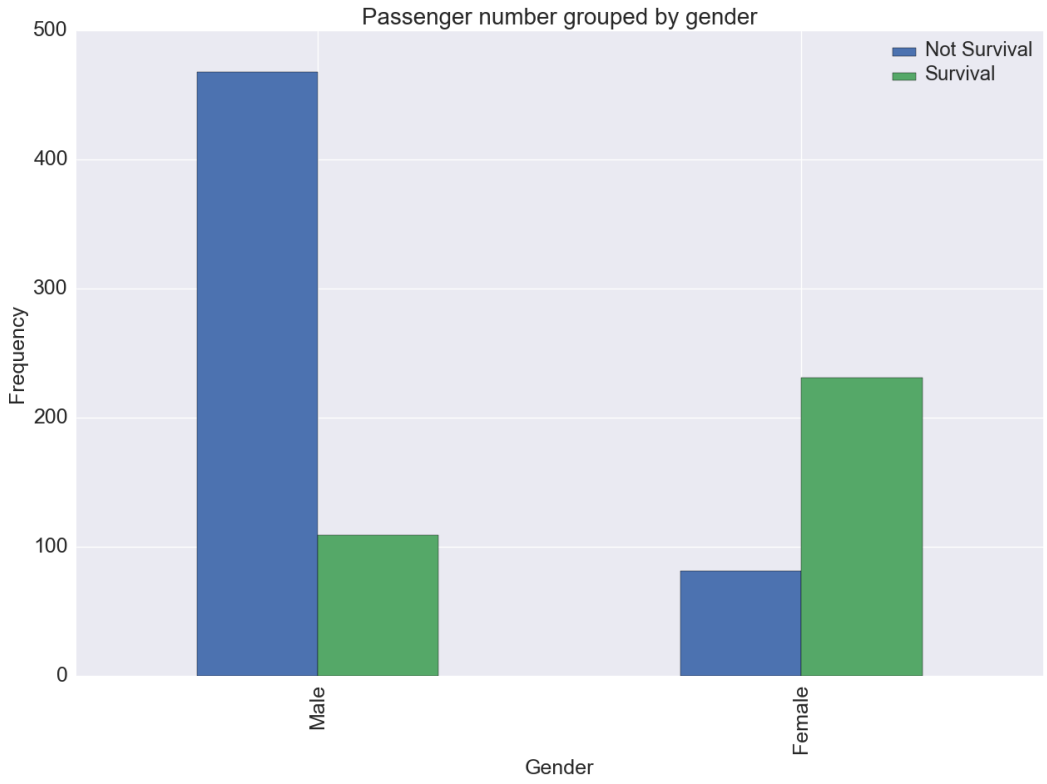

Figure 2 Number of male and female passenger who were survived or not survived

This plot shows that female passengers had a much higher survival rate than male passengers. 74.04% female passengers survived while only 18.89% male passengers survived. Implemented chi2 test to see if the survival rate for two groups are different and the result said yes at a strongly significant level.

```
Results of Chi-Squared test on Gender to Survival.
Does Gender have a significant effect on Survival?
Chi-Squared Score = 260.717020167
Pvalue = 1.19735706278e-58
```

So we can conclude that gender highly affected passengers' survival rate and female passengers were m ore likely to survive.

Survival and Passenger Class

Made the plot to see passengers' survival grouped by their class.

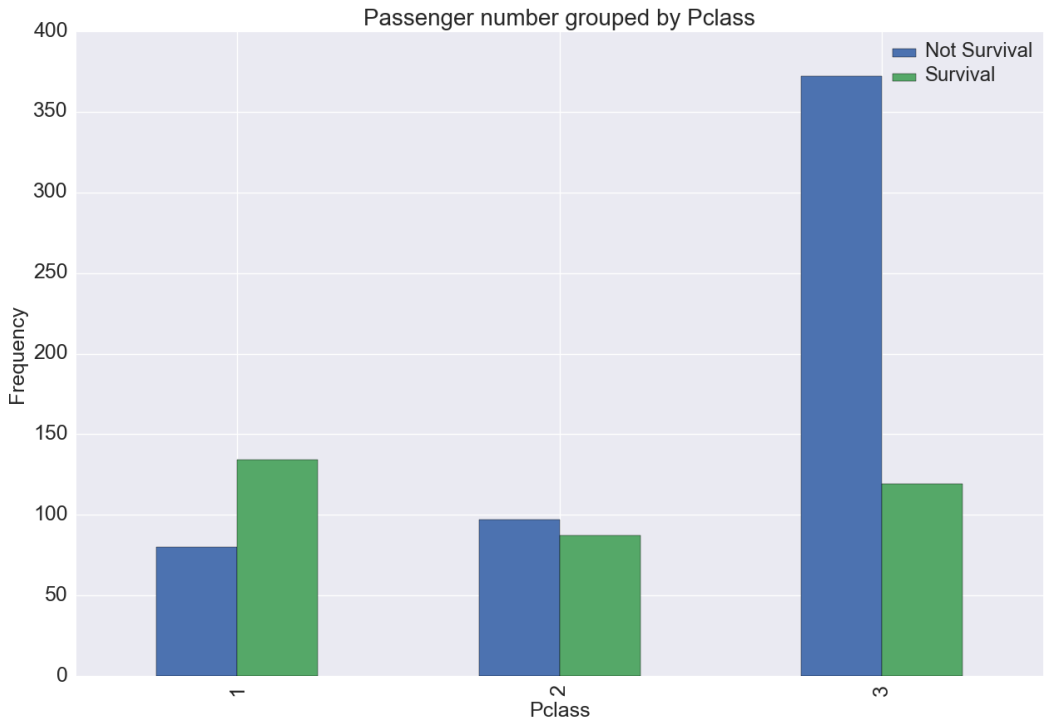

Figure 3 Number of passengers with different class who were survived and not survived

Class 1 was the highest priority class while class 3 was the lowest priority class. See that class 1 had the highest survival rate, 62.62%, followed by class 2 with 47.28% and class 3 with 24.24% survival rate. Implemented chi2 test to see if the survival rate for three groups are different. Chi2 results reject null hypothesis and show that three group have significant different survival rate at a very low percentage level.

```
Results of Chi-Squared test on Pclass to Survival.
Does Pclass have a significant effect on Survival?
Chi-Squared Score = 102.888988757
Pvalue = 4.5492517113e-23
```

So we can conclude that passenger class did affect people's survival rate. The higher their class were, the higher probability they could have to survive.

Survival and Parch

Then plotted as below.

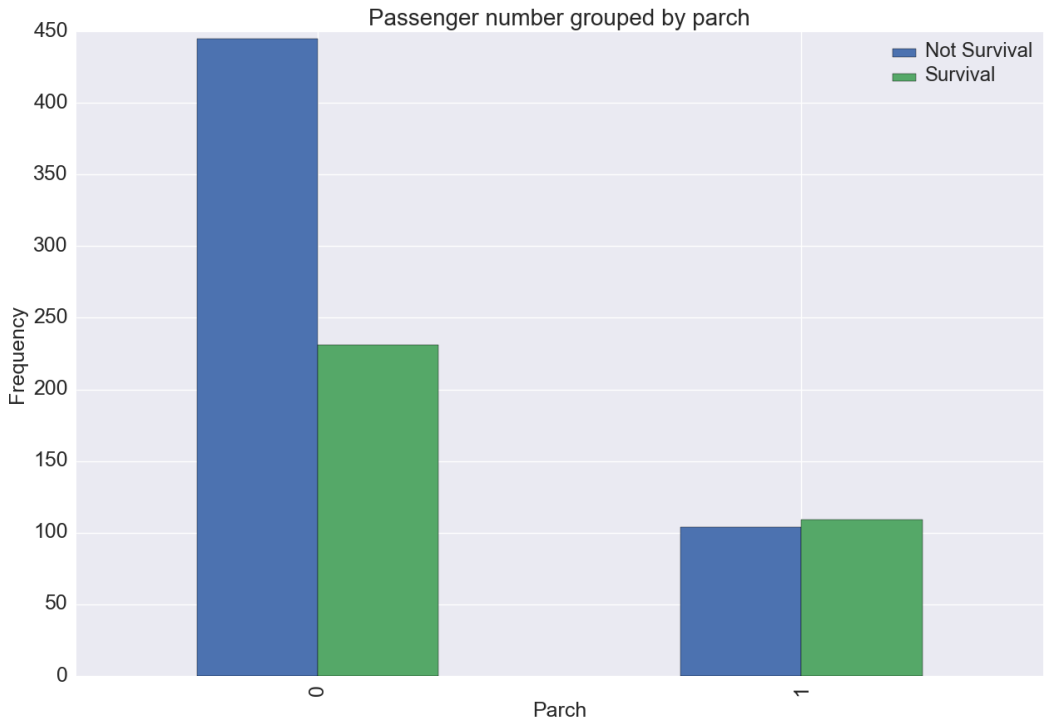

Figure 4 Survival condition between zero-parch group and non-zero-parch group

Passengers with non-zero-parch had a higher survival rate of 51.17% and those without parch had a survival rate of 34.17%.Processed chi2 test to see if two groups have different survival rate. The result rejects null hypothesis and shows that they did have different survival rate at a pretty low percentage level.

```
Results of Chi-Squared test on Parch to Survival.
Does Parch have a significant effect on Survival?
Chi-Squared Score = 19.1102627858
Pvalue = 1.23379653344e-05
```

So we can conclude that having parch or not on board affected people's survival rate on Titanic. Passengers having parch on board had a higher survival rate than those without parch.

Survival and Age

Take a first glance at the number of passengers survived or not survived at different age stages. I grouped passengers every 10 years old and got 8 age intervals. In order to achieve this, I created a new column called *Age Interval*, 0-10 representing older than 0 year old but younger than 10 years old and deduce the rest like this.

The sample size reduced to 714 since some ages were missing. I analyzed in previous part that leaving out samples without age would cause bias on estimation of survival rate. So the survival rate I calculated

below does not reflect the estimation of population. I just care about the trend with respect to age interval and tried to extract some non-quantitative information from them.

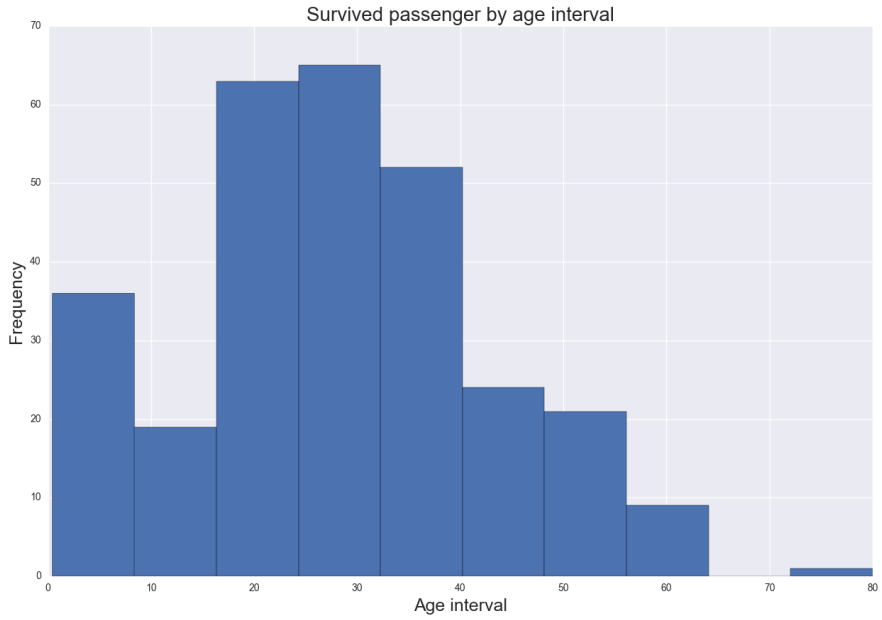

Figure 5 Number of survived people at different age stage

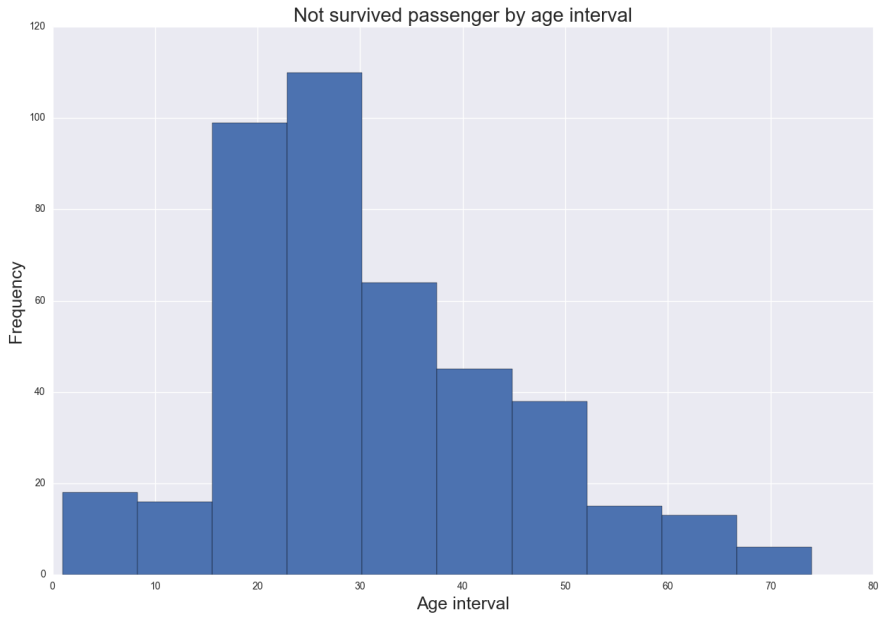

Figure 6 Number of not survived people at different age stage

See that passengers under 10 years old had an obvious higher survival rate than others ages. And then I calculated sample survival rate at each age interval and plotted them as below.

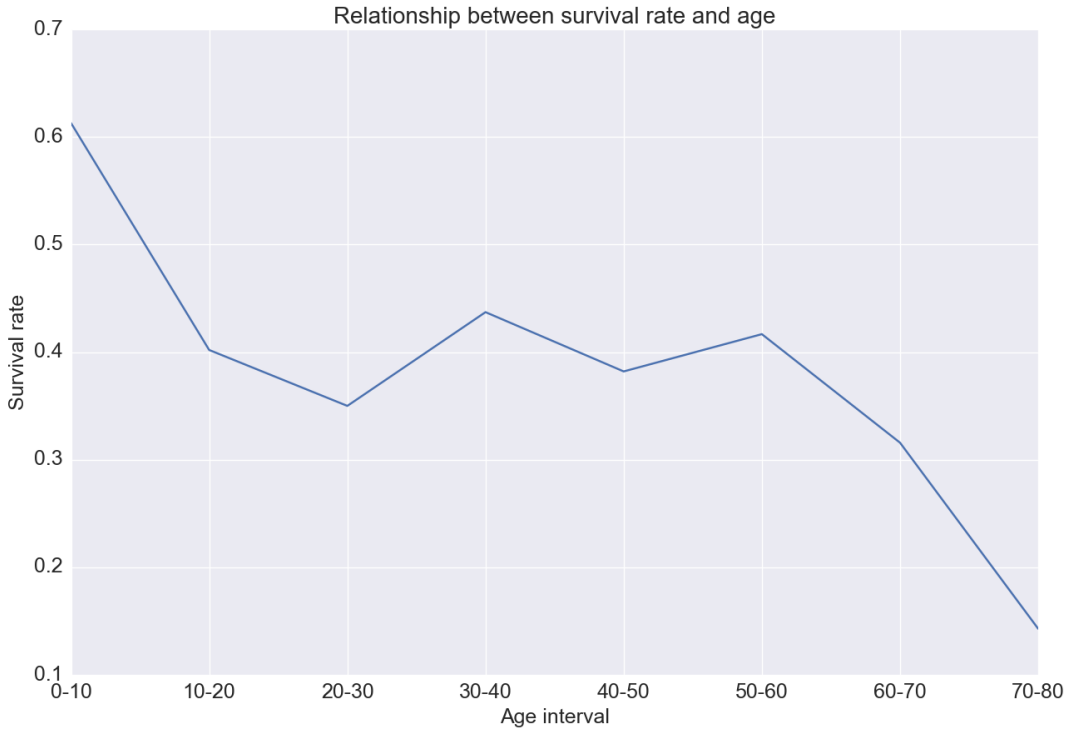

Figure 7 Relationship between survival rate and age interval

See that survival rate of passengers under 10 years old were higher than 60%, in comparison to other groups with lower survival rates. For further digging into it, I plotted a 2D scatter plot with *Age* on X axis and *Pclass* on Y axis. Red color stands for survival, while green color stands for death. See Figure 8 below.

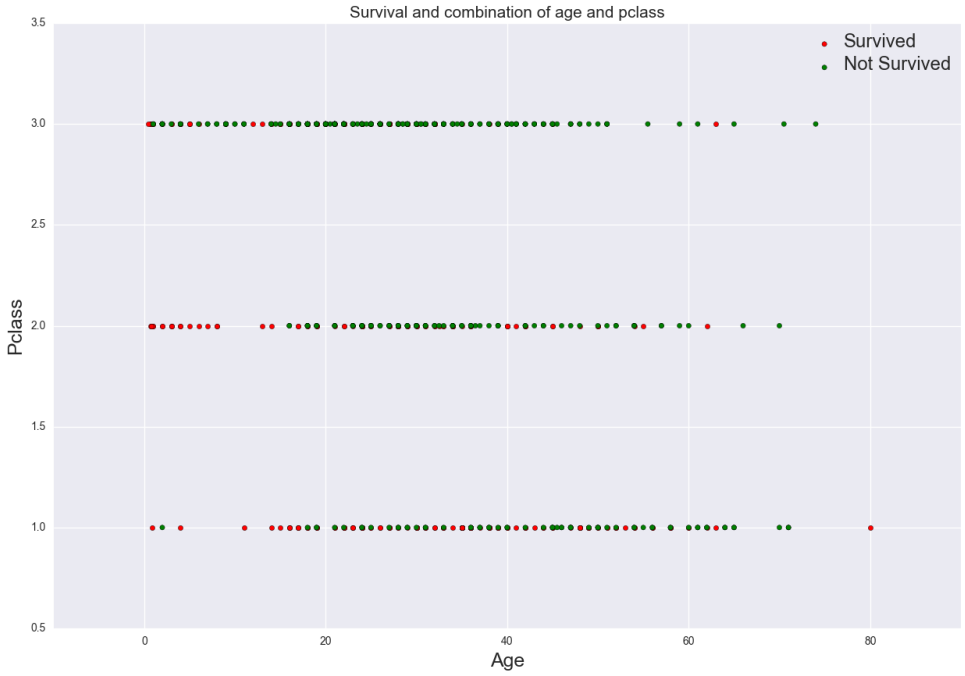

Figure 8 Scatterplot of survival and combination of age and pclass

Most children under 10 years old at class 1 and 2 were survived. However, children at class 3 were not so lucky. Meanwhile, survival rate of people older than 60 years old was not meaningful due to limited samples.

From the data above, we could guess that while Titanic was sinking, children and female passengers had priority over other passengers to get on lifeboats and those female passengers who survived were highly likely to be their mothers. That could explain why female and passengers with parch had relatively higher survival rate than their comparison groups. However, I was unable to prove so unless more information could be offered. In addition, the rescue rule seemed more friendly to passengers with higher class. Perhaps people with higher class got the message that the cruise was sinking much earlier than others knew and they had the priority to get on the lifeboat.

Conclusion

This report answers questions proposed at the beginning. Gender, passenger class and parch did affect passengers survival rate. Female, higher class passengers and those who have parents and children on board had a higher survival rate. There were some differences of survival rate among people at different age interval, but only the trend that children under 10 years old had higher survival rates was meaningful, showing that children had priority to get on lifeboat. Absolute data is not meaningful in this case.

There are some limitations. First of all, whether the sample could represent the population and was randomly selected from population is doubtable. Original data description does not explain this problem. Secondly, factors might have interactive effects with each other. For example, it is probable that most passengers with non-zero parch took their children under 10 year old with them on board. Children had

the priority of getting on lifeboat so that it increase the survival rate of non-zero parch group. More data and further study are needed to suppress potential drawbacks.