Titanic Survival Rate Analysis

Everyone knows the famous romantic movie 'Titanic", a fictionalized account of the sinking of the RMS Titanic, produced by James Cameron. RMS Titanic was a passenger liner that sank in the North Atlantic Ocean in the early morning of April 15, 1912, on its maiden voyage. It carried approximately 1317 people: 324 in First Class, 284 in Second Class, and 709 in Third Class. Of these, 869(66%) were male and 447(34%) were female. During the evacuation, the "women and children first" protocol was generally followed when loading the lifeboats. In this study, **the titanic dataset was selected and analyzed, to find answers for the following questions:**

1. What factors are contributing to higher chance of survival?
2. Based on those key factor combinations, which group got the highest survival rate?

A simple distribution analysis was conducted aiming at getting more familiar with the dataset. As shown, the largest age group is between 15-35, and there are very few people older than 65. Over 55% passengers are in 1st class, almost double times those for 2nd and 3rd classes. Among them, around two thirds are males passengers (65%).

Chart 1. Distribution of Age

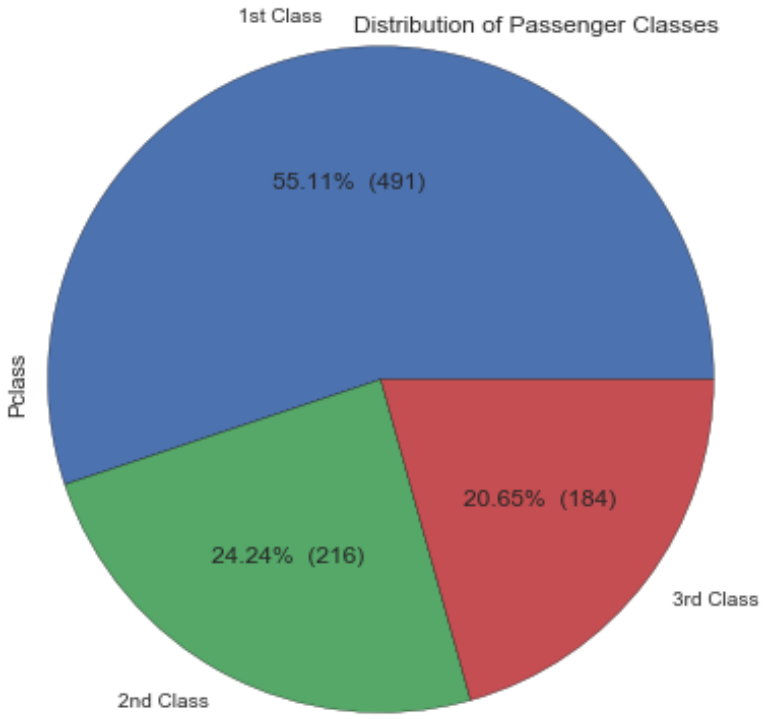

Chart 2. Percentage by Passenger Classes

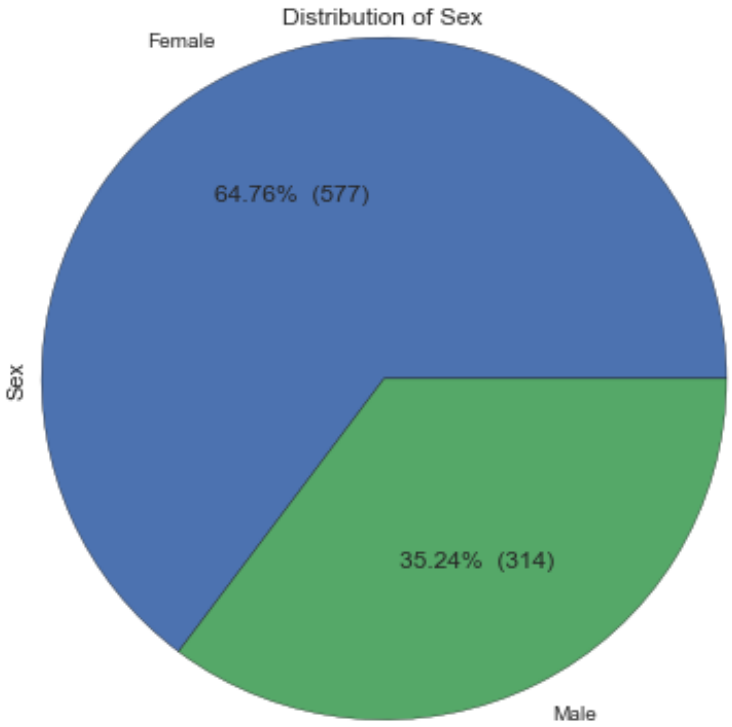

Chart 3. Percentage by Sex Group

The chart below is a joint distribution study of age and passenger classes, dot size indicates passenger counts in this age and class combination. It shows that passengers from 3rd class were more at younger age compared to that from 1st and 2nd classes, mainly under their middle 40s. The majority of 1st and 2nd classes passengers were between 15 and 60, except differences such as 2nd class had more people at young age (under 10) while 1st class has more senior people over 60.

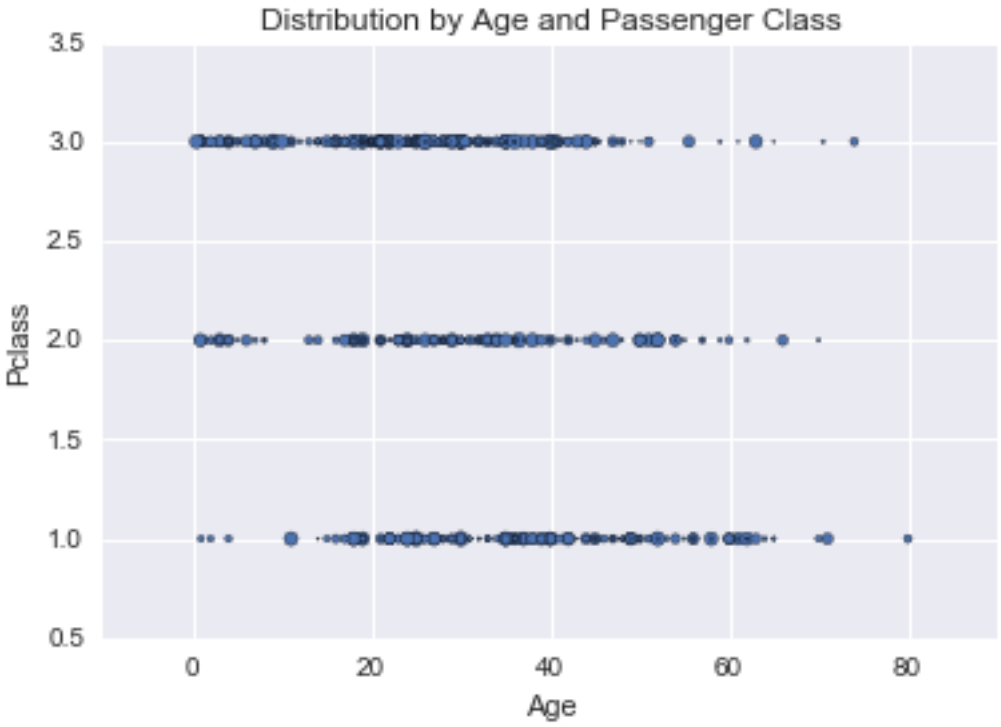

Chart 4. Distribution by Age and Passenger Classes

As observed, this dataset has missing data points in age columns. Those missing values were filled as 0 for future analysis. No missing values for other key parameters, such as passenger class, sex, age, number of siblings and number of parches. For the convenience of future analysis, a sub-dataset (Age > 0) was used for analysis related with variable Age. The whole dataset was employed for other analysis.

Question 1: Key factors for higher survival rate

A correlation analysis was carried out between survival rates and potential contributor factors: age, sex, passenger class, sibling and parch. Results showed that Sex and Passenger Class are highly correlated with survival rate. Generally, the higher the passenger class, the higher survival rate. Female passengers got significant higher survival chance compared to male passengers: 74% VS 19%, while the average survival rate is 38%. This could be explained by the "Women and Children First" rule.

Table 1. Correlation

Group	Correlation
Age	-0.078
Sex	-0.543
Class	-0.338
Sib	-0.035
Parch	0.082

Table 2. Survival Rate by Passenger Group

Group	Survival Rate
1st Class	63%
2nd Class	47%
3rd Class	24%
Average	38%

Table 3. Survival Rate by Sex Group

Group	Survival Rate
Female	74%
Male	19%
Average	38%

Considering the "Women and Children First" rule, age was studied by groups. According to the Wikipedia, 'In the 19th century, children younger than seven years old were believed incapable of crime'. According to the passenger demography, the oldest passenger is 80 years old, and the oungest is less than 1 year old. The whole dataset was divided into 9 age groups:

1. Children: 0-6 years
2. Teenager: 7-15 years
3. Adult 1: 1-20 years

4. Adult 2: 21-30 years
5. Adult 3:31-40 years
6. Adult 4: 41-50 years
7. Adult 5: 51-60 years
8. Senior 1: 60-69 years
9. Senior 2: 70-80 years

Results were shown in chart below:

Chart 5. Survival Rate by Age Groups

The total survival rate is 41% among all passengers with age record (age > 0). For children younger than 7, the survival rate was significantly increased to 70%, almost doubled. Children between 7 and 15 got survived 39%, which is quite close to the average survival rate. Adult between 16 and 60 also has similar survival rate as average, ranging from 35% to 44%. Senior passengers in their 60s was much lower than the average survival rate, around 32%. Senior group older than 70 has got significantly lower survival rate 14%. The could be explained by that all passenger in the senior group elder than 70 were male, when the sank happened, the 'Women and Children First' rule applied to them, resulting in much lower survival rate. Actually only one gentleman, Mr. Algernon Henry Wilson Barkworth, survived from all 7 passengers in this group. Passenger classes' analysis was conducted for the same group, however, no obvious conclusion could be draw from it.

The correlation between survival rate and age groups is 0.162, based on three age groups: <7, 7-69, > 69

Table 4. Demographic Analysis for Senior (Age >= 70) by Sex groups

Group	Survival Rate
Female	0%
Male	100%

Table 5. Demographic Analysis for Senior (Age >= 70) by Passenger Classes

Group	Survival Rate
Class 1	57%
Class 2	14%
Class 3	29%

To summarize, **sex, passenger classes and age groups are key factors significantly influencing the survival rate. The correlation with survival rate from high to low: sex, passenger class, age group.**

Q2: Groups with highest and lowest survival rate

Scatter plots were draw to examine the demograph of whole passenger group as well as survived passenger groups. In charts below, x-axis represents age, y-axis represents passenger classes: 1- first class, 2-second class, 3-third class., and z-axis represents sex groups: 1-female, 2-male. From those plots, it could be indicated that female passengers in 1^{st} and 2^{nd} classes have comparatively higher survived population. Male passengers in the 2^{nd} class show much lower survival population. To evaluate which group has the highest and lowest survival rate, a percentage analysis was conducted based on the sex, passenger class and age groups.

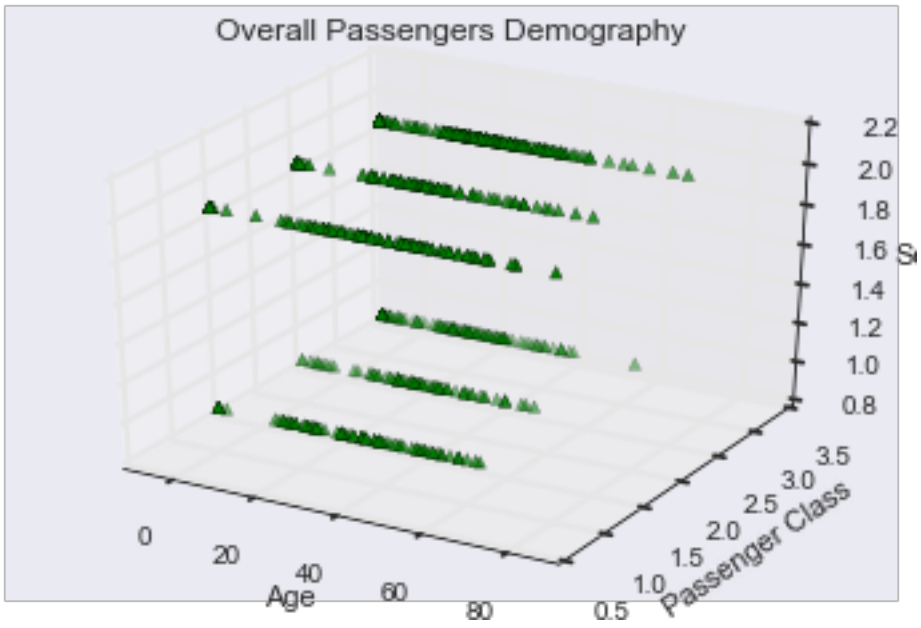

Chart 6. Overall Passenger Demography

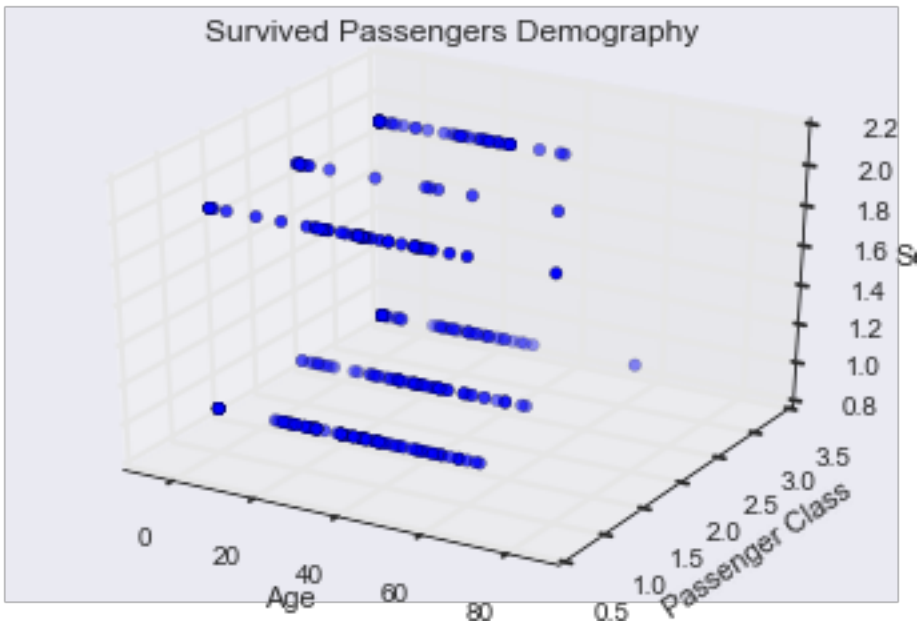

Chart 7. Survived Passenger Demography

Detailed analysis was carried out. The table and chart below show the survival rate by different key factior combination groups. Despise those groups lack of enough data points (less than 10 passengers, shaded), **results indicate female passenger from 1st between age 7-59 got the largest chance of survival, comparatively, male passengers from 2nd class between age 7-59 got lowest survival rate**.

Table 6. Survival Rate (left) and Passenger Count (right) by three independent variables: Sex, Age Group and Passenger Classes

Sex	Age	Pclass	
1	7-59	1	98%
		2	91%
		3	41%
	Older than 60	1	100%
		3	100%
	Younger than 7	1	0%
		2	100%
		3	69%
2	7-59	1	42%
		2	7%
		3	14%
	Older than 60	1	14%
		2	25%
		3	0%
	Younger than 7	1	100%
		2	100%
		3	43%

Sex	Age	Pclass	
1	7-59	1	81
		2	68
		3	85
	Older than 60	1	3
		3	1
	Younger than 7	1	1
		2	6
		3	16
2	7-59	1	85
		2	87
		3	235
	Older than 60	1	14
		2	4
		3	4
	Younger than 7	1	2
		2	8
		3	14

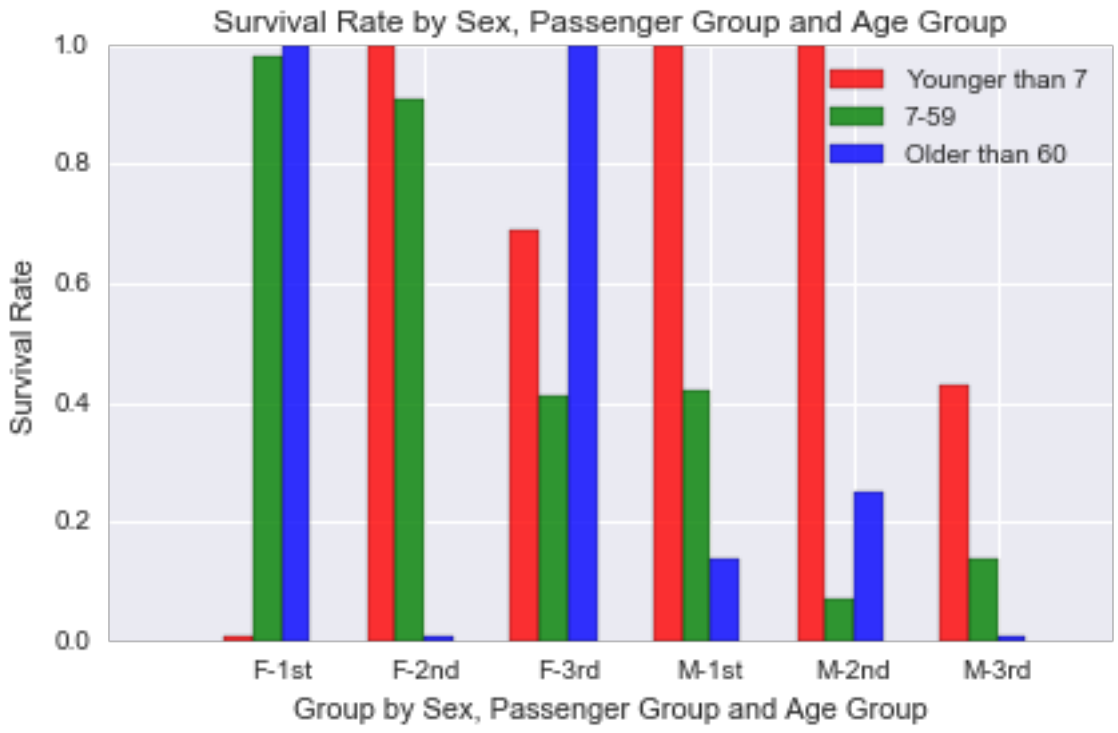

Chart 8. Survived Rate by Sex, Passenger Classes and Age Groups

Conclusion and Limitation Discussion

To conclude, sex, passenger classes and age groups are key factors significantly influencing the survival rate, and female passenger from 1st between age 7-59 got the largest chance of survival, comparatively, male passengers from 2nd class between age 7-59 got lowest survival rate.

However, this analysis has a few limitations, such as limited data points and missing data, there are over 2000 people on-board, but only has 891 available passenger information. Among them, only 714 people had age information (177 missing points). In the age related analysis, all missing age data were replace by 0 and got omitted in the analysis, which means, the conclusions drawn for entire group were only based on 80% passenger information.

To further dig into this, a parch analysis was carried out for passengers with missing age data. Based on common sense and social rules, children at young age are unlikely to travel without any parches. The chart showed that the majority (88%) passengers in this group travel without a parch, indicating that those 177 passengers were more likely at their elder ages, in another word, older than 7 years. Omitting data for this group would possibly result in the average survival rate and age range with highest and lowest survival rate.

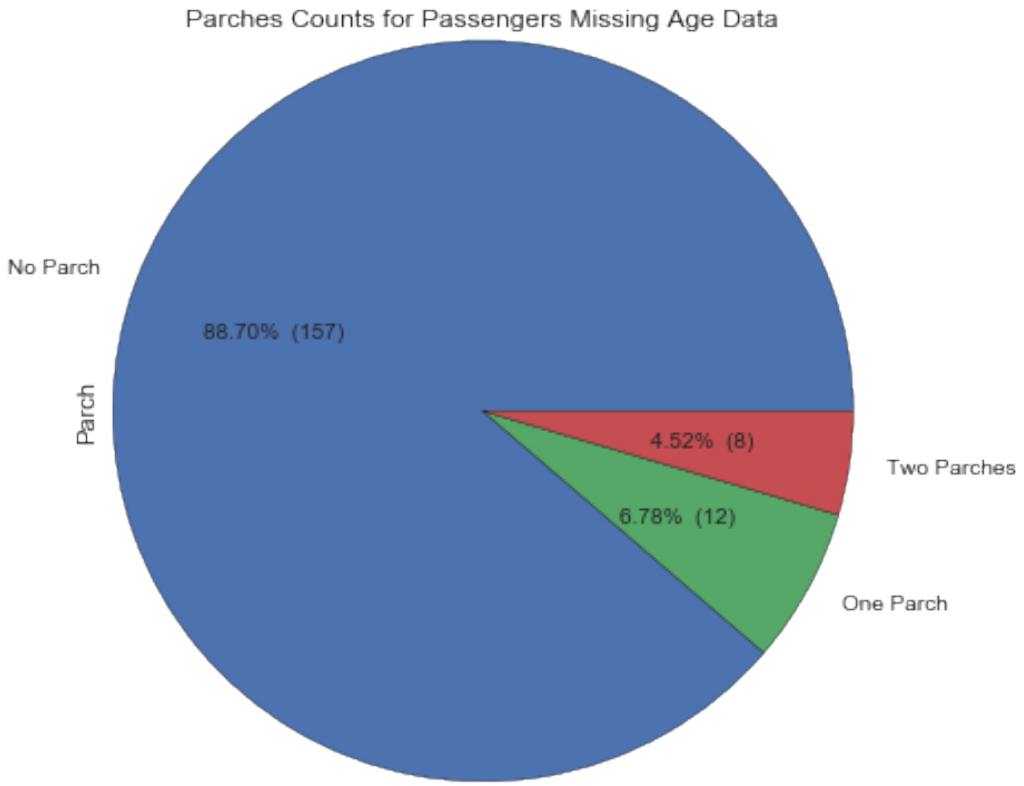

Chart 9. Parches Counts for Passengers without Age Information