An Statistical Walk Aboard the Titanic

Marshall McQuillen

Guiding Question

What characteristics separate those who survived from those who died?

Secondary Questions

1. Does socioeconomic status have an affect on survival?
2. How do Gender and Age affect rurvival rates?
3. How does family size affect surival rates?

1 Data Overview

Looking at the training data froma bird-eye view, there are 891 observations representing passengers and 12 variables. Since some of the variable name are a little cryptic, an description for each is provided below.

Variable Name	Description
PassengerId	Unique identifier for each passenger
Survived	Binary; 1 = Survied & 0 = Died
Pclass	Socio-economic status; 1 = Upper, 2 = Middle & 3 = Lower
Name	Passenger Name
Sex	Male or Female
Age	Passenger Age
SibSp	Number of siblings + spouse aboard ship
Parch	Number of parents + children aboard ship
Ticket	Ticket Number
Fare	Amount paid for ticket
Cabin	Cabin number
Embarked	The town from which the passenger boarded the ship; C = Cherbourg, Q = Queenstown & S = Southhampton

First and foremost, by running `str(training)` on the data, it is apparent that the first entries in the Cabin and Embarked columns are empty strings, indicating that the data is probably not perfectly clean (no surprises there). Checking to see where any Null's might be, it becomes clear that there are in fact no nulls, and that these spaces were intentionally left empty. In addition to null values, all the NA's are in the Age, accounting for roughly 20% of the values in that column. Both of these will need to be imputed intelligently when the time to create a predictive model comes around.

In addition to the missing values, it is important to note that some of the discrete attributes have been read in as continous variabes such as Pclass, Sibsp and Parch. Since these variables actually represent discrete characteristics of each passenger, changing them to be non-continuous will allow a more representative analysis.

Table 2: Attribute Null & NA Counts

	PassengerId	Survived	Pclass	Name	Sex	Age
Null Count	0	0	0	0	0	0
NA Count	0	0	0	0	0	177

Table 3: Attribute Null & NA Counts (continued)

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Null Count	0	0	0	0	0	0
NA Count	0	0	0	0	0	0

2 Does Socioeconomic Status have an affect on Survival?

2.1 Does Money Sink or Swim?

By creating a table with the Pclass (which refers to the socioeconomic status (SES) of the passenger) and Survived variables, I can get a good sense of the number of passengers that lived and died, based on their SES. Simple summation and division returns the probabilites of a passenger living given their respective SES

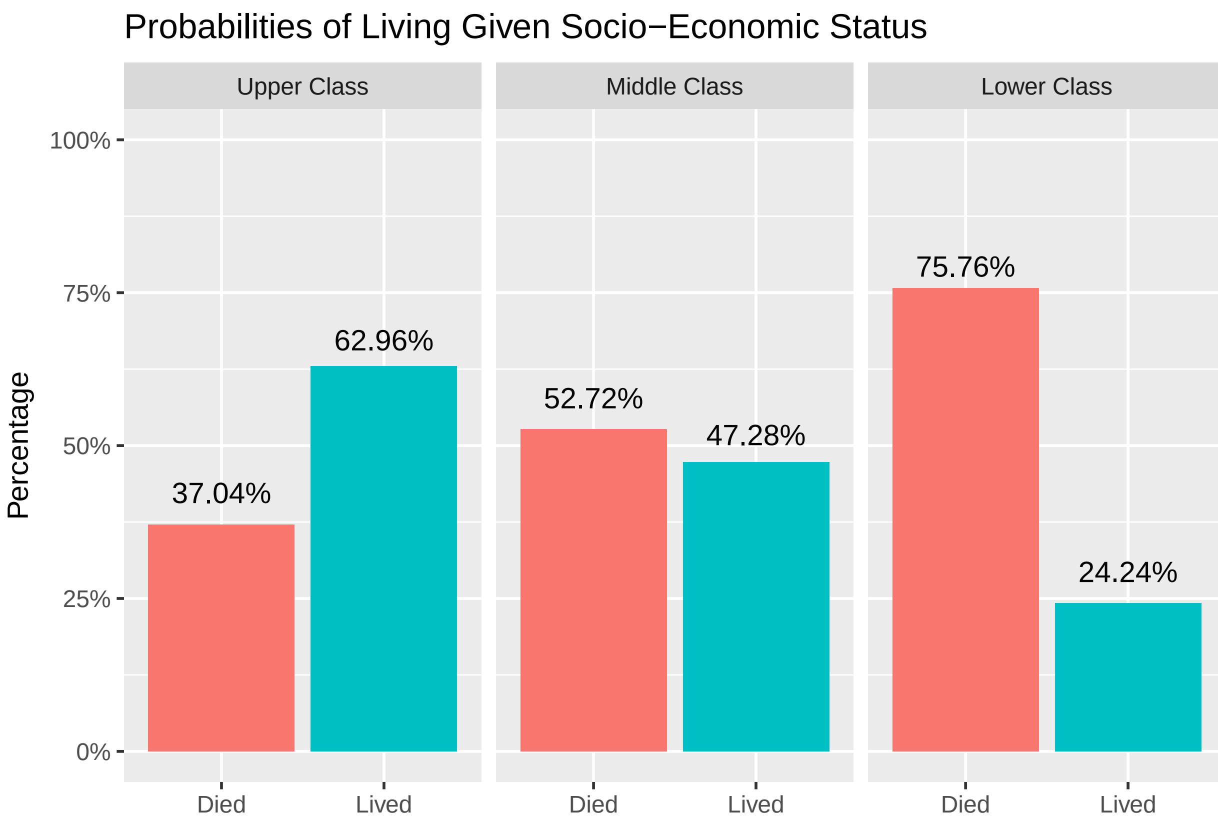
Table 4: Survival Counts by SES

	Upper	Middle	Lower
Died	80	97	372
Survived	136	87	119

Table 5: Survival Rates by SES

	Probability of Living
Upper Class	62.96%
Middle Class	47.28%
Lower Class	24.24%

This same information is displayed visually below.



2.1.1 Illustrating Bayes Theorem with Survival Rates and Socioeconomic Status

This type of classification problem creates a great opportunity to illustrate Bayes' Theorem. Recall that Bayes Theorem is defined as:

$$P(A|B) \ = \ \frac{P(B|A)P(A)}{P(B)}$$

where:

- $P(A|B)$ = Posterior
- $P(B|A)$ = Likelihood
- $P(A)$ = Prior
- $P(B)$ = Normalizing Constant.

The equation above can be rewritten to better match the problem context as:

$$P(\ "X\ class\ citizen"\ |\ "Lived"\ )\ =\ \frac{P(\ "Lived"\ |\ "X\ class\ citizen"\ )\ P(\ "X\ class\ citizen"\ )}{P(\ "Lived"\ )}$$

where:

- $P(\ "X\ class\ citizen"\ |\ "Lived"\ )$ = Posterior
- $P(\ "Lived"\ |\ "X\ class\ citizen"\ )$ = Likelihood
- $P(\ "X\ class\ citizen"\ )$ = Prior
- $P(\ "Lived"\ )$ = Normalizing Constant.

$P($ "*Lived*" $)$, the Normalizing Constant, will be the probability of living, *regardless of SES.* This could be broken out into three terms,

$$P(\text{"}Lived\text{"} \mid \text{"}Upper\ class\ citizen\text{"}) + P(\text{"}Lived\text{"} \mid \text{"}Middle\ class\ citizen\text{"}) + P(\text{"}Lived\text{"} \mid \text{"}Lower\ class\ citizen\text{"})$$

however it is far easier to calculate the proportion of those that lived over everyone that was aboard the ship. This comes out to be 38.38%.

The final term needed to complete the right hand side of the equation, the Prior, is simply the proportion of those on board that were Upper, Middle or Lower class. These come out to be 24.24%, 20.65% and 55.11%, respectively, shown in the table below.

Table 6: Socioeconomic Status Proportions Aboard the Titanic

	Probability of Being X Class
Upper Class	24.24%
Middle Class	20.65%
Lower Class	55.11%

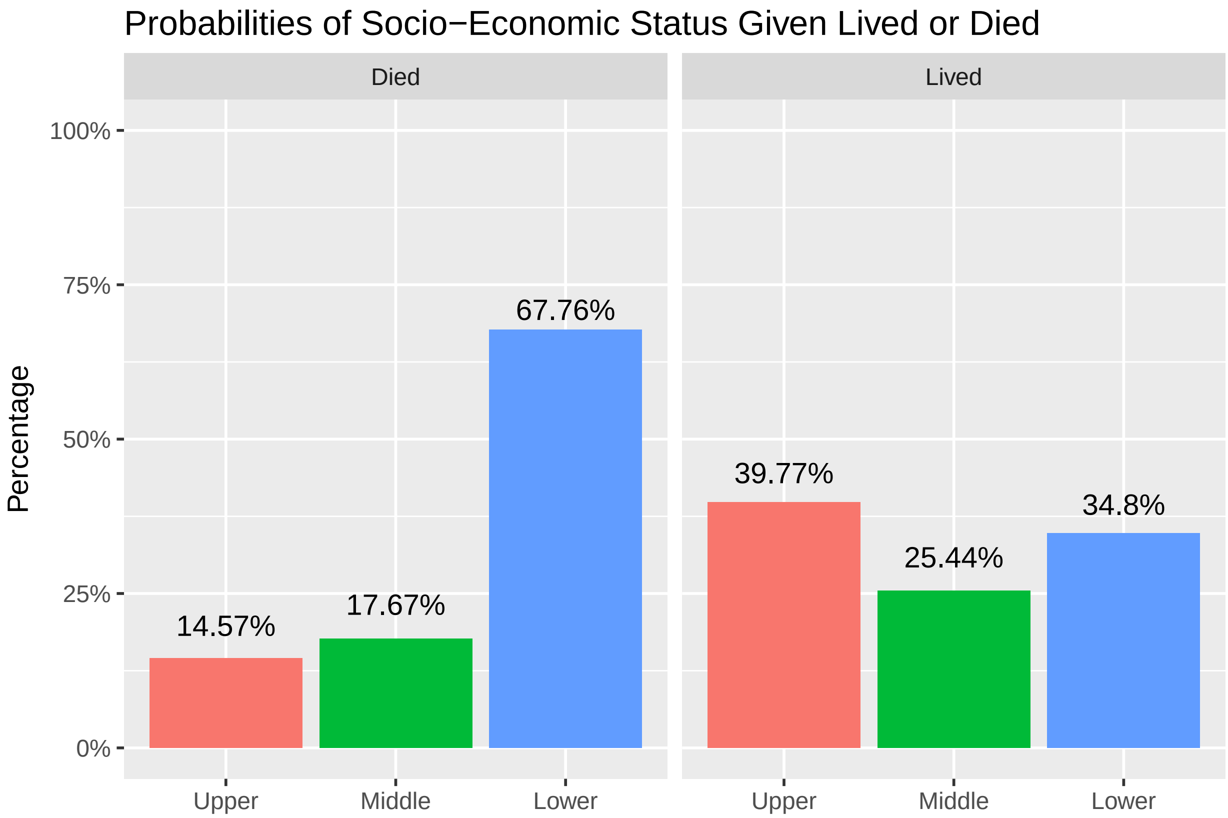
Now it is simply a matter of defining three different equations for each of the three possible socioeconomic status', and substituting in the corresponding numbers (Note that in the above percentages I rounded to two decimal places, however when calculating the final probability it is paramount that the entire number is used).

$$P(\text{"}Upper\ class\ citizen\text{"} \mid \text{"}Lived\text{"}) = \frac{0.6296296 \cdot 0.2424242}{0.3838384} = 0.3976608 = 39.77\%$$

$$P(\text{"}Middle\ class\ citizen\text{"} \mid \text{"}Lived\text{"}) = \frac{0.4728261 \cdot 0.2065095}{0.3838384} = 0.254386 = 25.44\%$$

$$P(\text{"}Lower\ class\ citizen\text{"} \mid \text{"}Lived\text{"}) = \frac{0.2423625 \cdot 0.5510662}{0.3838384} = 0.3479532 = 34.8\%$$

This can be double checked visually by dividing the passengers into those that lived and died, and then, for each of those groups, plotting the percentage that were Upper, Middle and Lower class. Low and behold, Bayes was right.

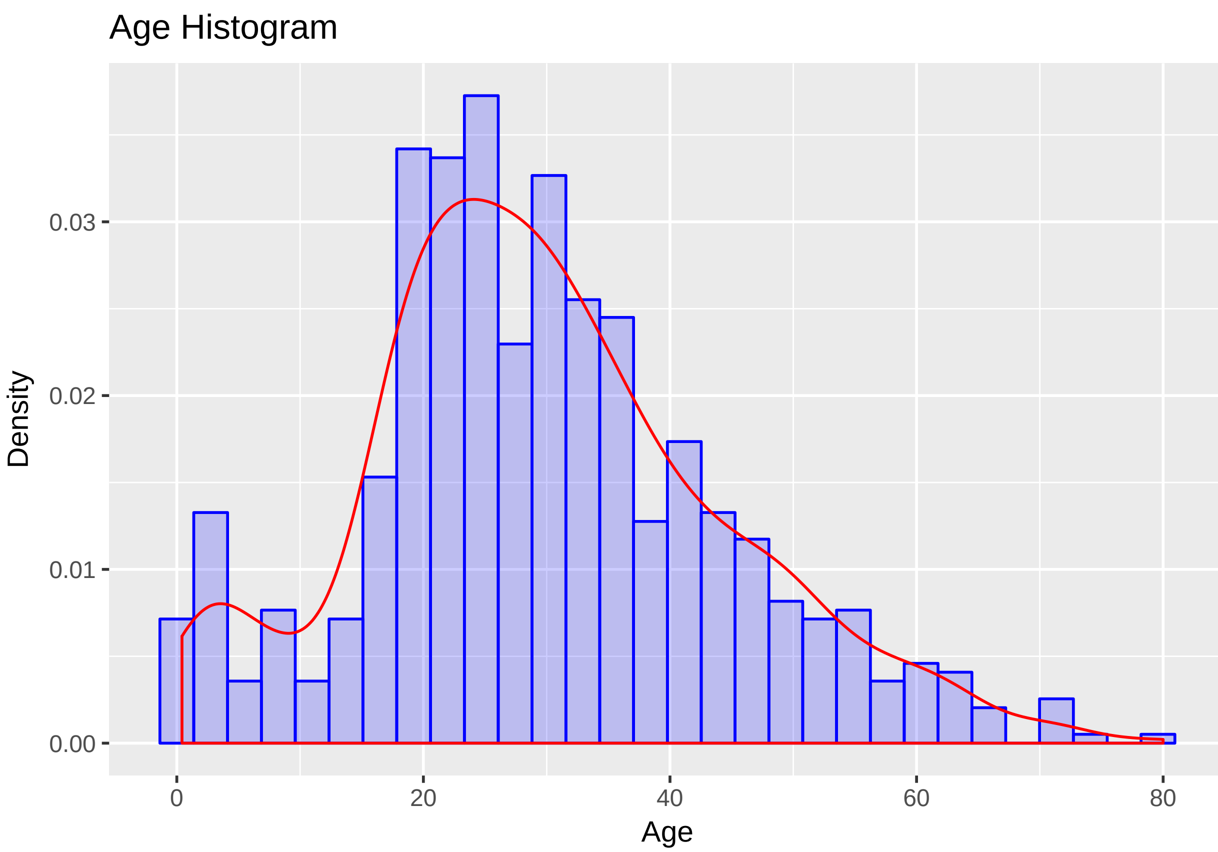


3 How do Gender and Age affect Survival Rates?

A quick overview of the Gender and Age variables are shown below, demonstrating that most people aboard were men and between 20 - 40 years old.

Table 7: Gender Proportions

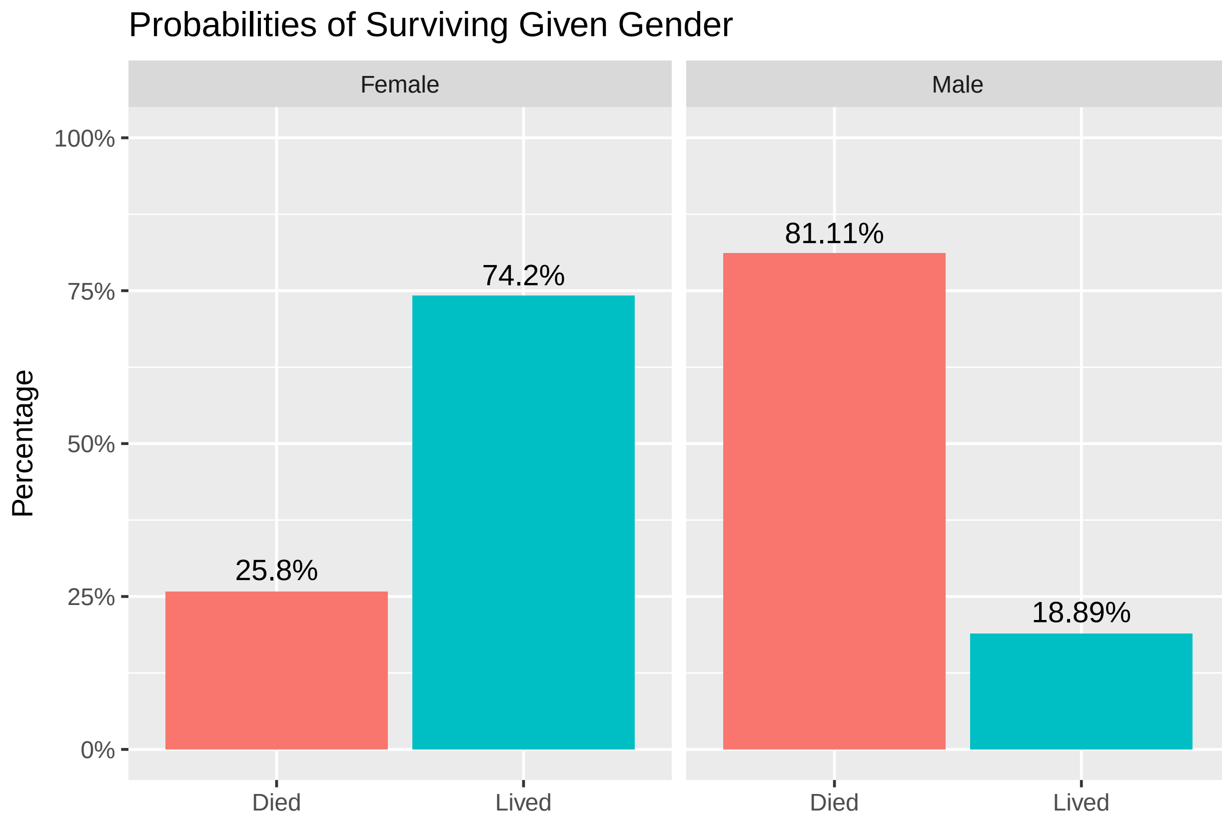
	Proportion
Female	35.24%
Male	64.76%

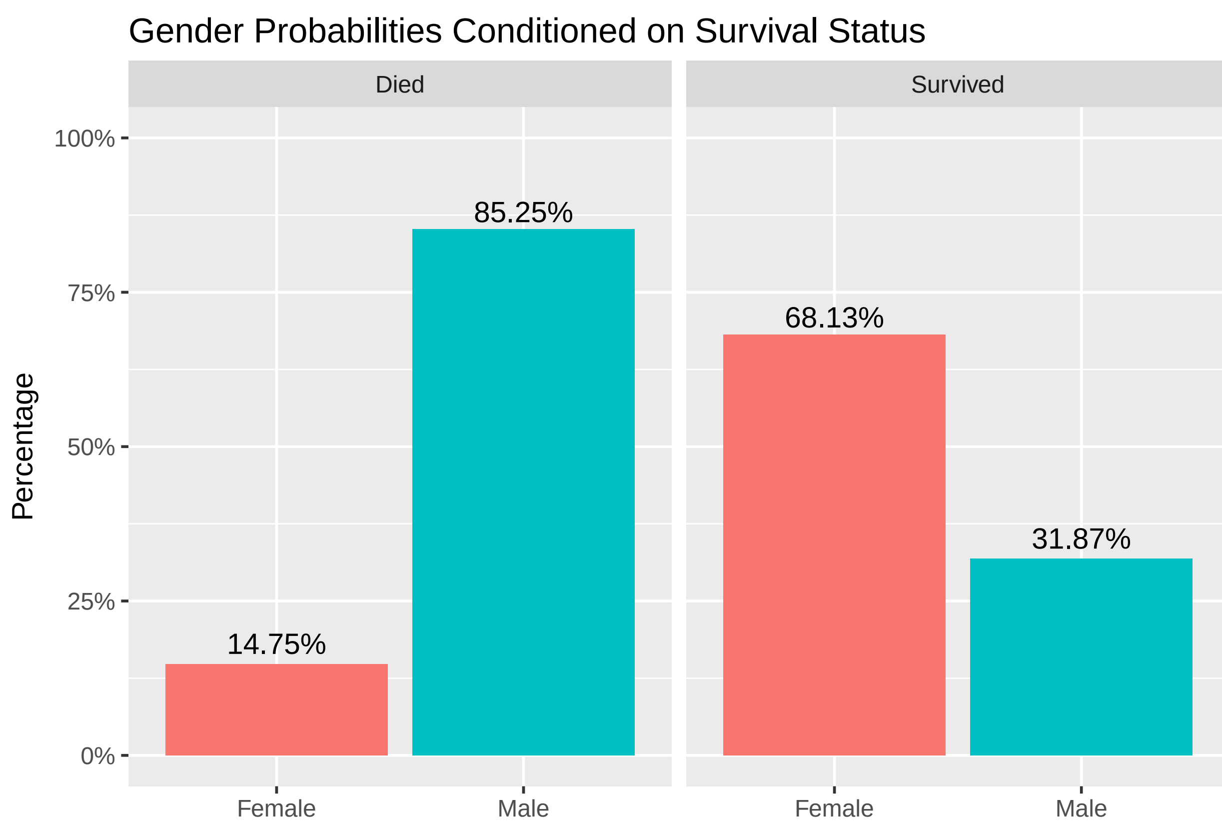


Gender

Looking at the two plots on the following page, it is apparent that given that a passenber was Female, her probability of Surviving was 74.2%. On the other hand, if a passenger was Male, he had over an 80% chance of dying.

The second plot shows the probabilities of a passenger being Male or Female, conditioned on whether they survived the sinking of the ship or not. Once again, there is a vast gender divide of the passengers that died, with 85% of all passenger that died being Male. Given that a passenger survived, there is more than double the chance that she was Female than Male.


Probabilities of Surviving Given Gender
Female
Male
Percentage
100%
75%
50%
25%
0%
25.8%
74.2%
81.11%
18.89%
Died
Lived
Died
Lived


Gender Probabilities Conditioned on Survival Status
Died
Survived
Percentage
100%
75%
50%
25%
0%
14.75%
85.25%
68.13%
31.87%
Female
Male
Female
Male

Interpretting Logistic Regression Using Passenger Age

If the goal of a model is a low error rate, using only one variable is rarely a good idea. However, if interpretability and inference are the goal, using linear or logistic regression can provide unique insight into our data. In order to see how a passenger's age affects their chance of surviving the sinking of the ship, I decided to create a new column that puts passengers into certain bins, based on their age. I created 7 different bins, outlined in the table below, which separates each age group into those who lived and those who died.

Table 8: Surival Counts by Age Group

	0-10	11-20	21-30	31-40	41-50	50-60	Over 60
Died	26	71	146	86	53	25	17
Survived	38	44	84	69	33	17	5

When a general linear model is fit using only this binned column, R one-hot-encodes the column, effectively creating a new column for each age group.The reason for the separation of the continuous age variable into bins becomes clear when the equation, with problem context accounted for, is expressed below.

$$\textit{Log Odds}(\textit{Surviving}) \;=\; \beta_1\;(\textit{Age}\;0-10)\;+\;\beta_2\;(\textit{Age}\;11-20)\;+\;\beta_3\;(\textit{Age}\;21-30)\;...\;\textit{etc.}$$

Since each coefficient represents a change in the log odds of survival with a one unit change in its associated variable, **and only one variable will be non-zero (if a passenger is in the age group 0 to 10, they aren't going to be in any other age group)**, the log odds of a passenger in age group j will be equal to coefficient β_j.

$$\textit{Log Odds}(\textit{Surviving}) \;=\; \beta_1\;(1)\;+\;\beta_2\;(0)\;+\;\beta_3\;(0)\;+\;\beta_4\;(0)\;...\;\textit{etc.}$$

Removing all 0 terms from the equation, the formula simplifies to:

$$\textit{Log Odds}(\textit{Surviving})_j \;=\; \beta_j$$

A little math will show us the probability that a passenger survives, given a specific age group. By exponentiating the log odds, the odds are returned, which can the be divided by one plus itself (equation below) to return the probability.

$$P \;=\; \frac{\textit{Odds}}{1\;+\;\textit{Odds}}$$

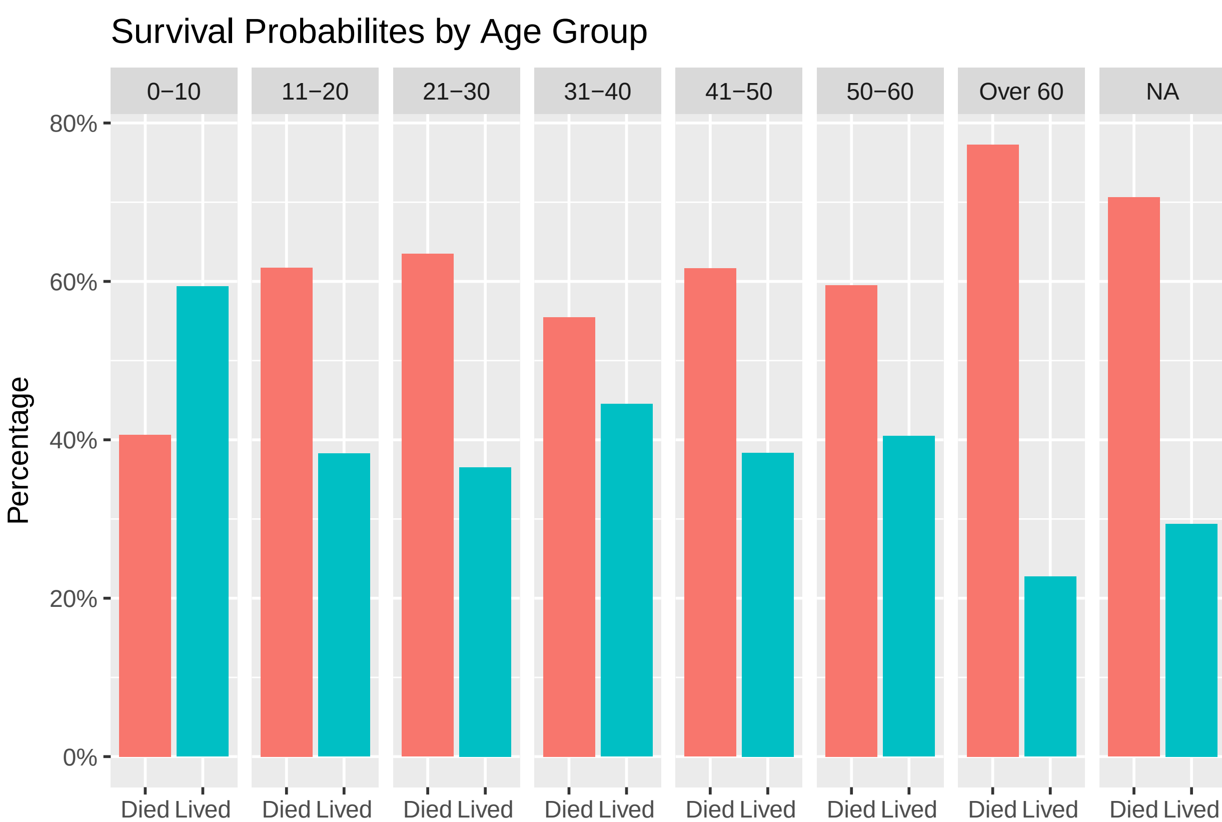
This is visually confirmed with a multi-faceted plot shown on the following page, in addition to taking the number of Survivors for each age group and dividing by the total number of passengers in that age group from the Survival Counts by Age Group Table.

Table 9: Logistic Regression Age Group Coefficients

	Coefficient
0 - 10	0.3794896
11 - 20	-0.4784902
21 - 30	-0.5527898
31 - 40	-0.2202408
41 - 50	-0.4737844
50 - 60	-0.3856625
Over 60	-1.2237754

Table 10: Probability of Surviving given Age Group

Age Group	Probability of Surviving
0 - 10	59.37%
11 - 20	38.26%
21 - 30	36.52%
31 - 40	44.52%
41 - 50	38.37%
50 - 60	40.48%
Over 60	22.73%



4 Family First

Carrying the theme of the analysis into the familial realm, the two plots below show the probability of surviving based on the value in the *Parch* column (the number of parents + children aboard per passenger) and the value in the *SibSp* column (the spouse + number of siblings aboard per passenger).

One important thing to note is the sample sizes for each facet within each plot, provided in a table below the associated plot. Since the goal of this analysis (and statistics in general) is to make assumptions about a population based on a sample, I would only be willing to take the survival rates of the passengers with 1, 2 or 3 in the *Parch* column at face value; I would be hesitant to make any generalizations on the other possible values (4, 5, 6 and 7), due to such small sample sizes per group.

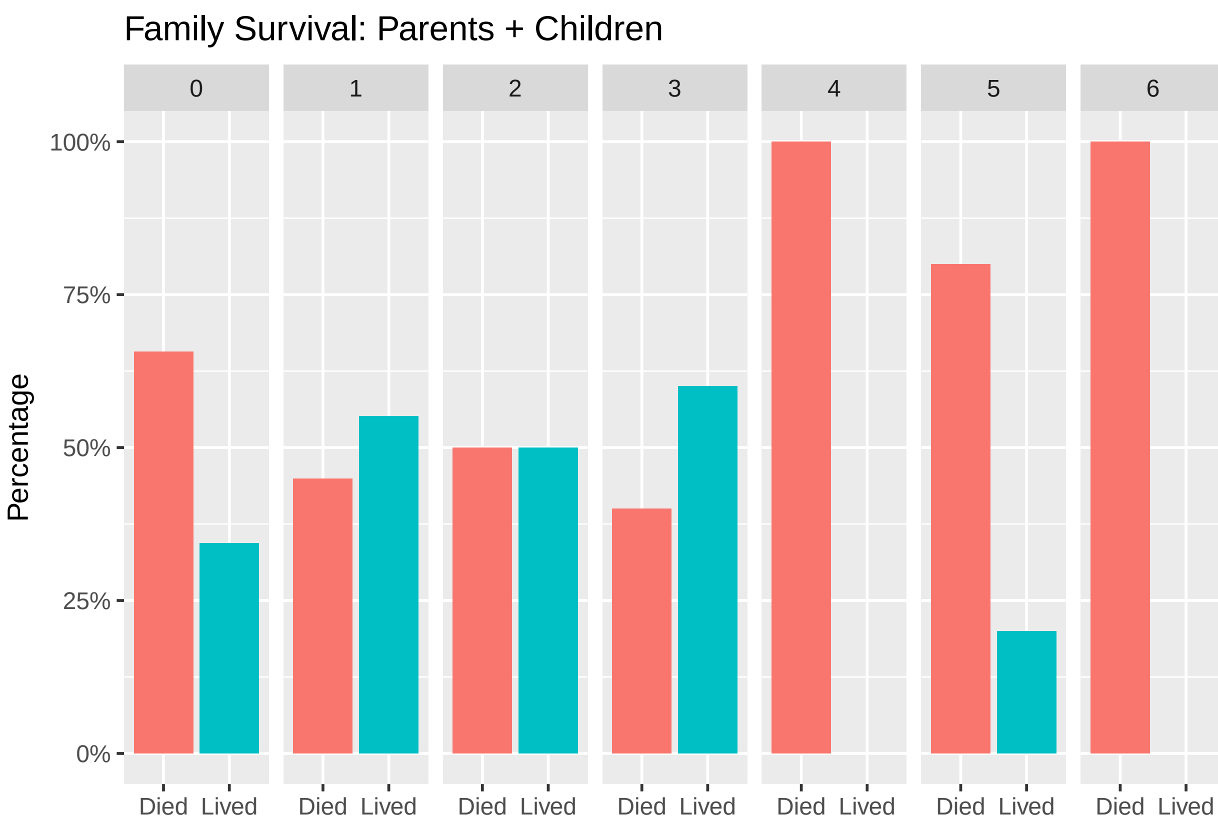


Table 11: Death & Surival Counts by Number of Parents or Children Aboard

	0	1	2	3	4	5	6
0	445	53	40	2	4	4	1
1	233	65	40	3	0	1	0

With regard to the previous paragraph, note that the sample sizes per possible value in the *SibSp* column decrease dramatically after 2. In the same vein as above, making assumptions about the population (test set) using *SibSp* with values greater than two would be statistically irresponsible.

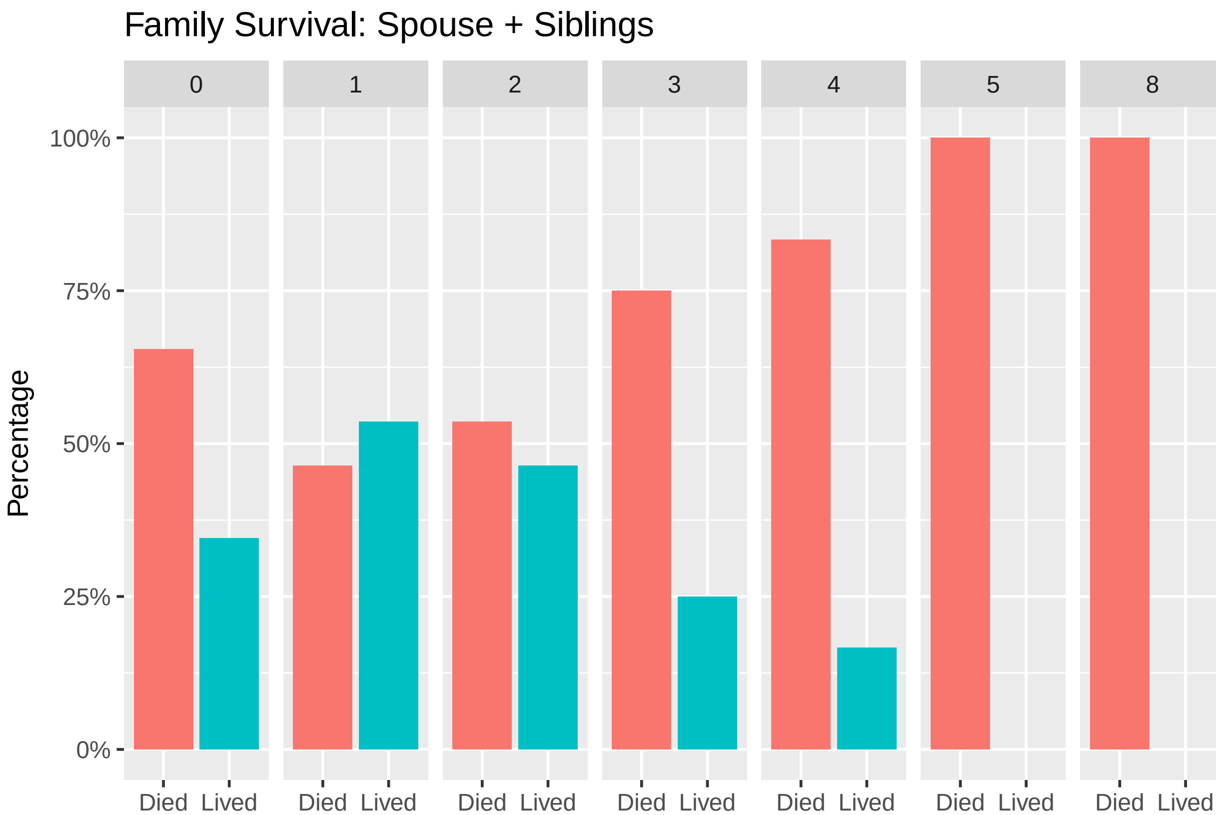


Table 12: Death & Surival Counts by Number of Parents or Children Aboard

	0	1	2	3	4	5	8
0	398	97	15	12	15	5	7
1	210	112	13	4	3	0	0