SURVIVING THE TITANIC

W200 Section 2
Project 2: Data Analysis
Authors: Danish Iqbal, Matthew Holmes, Kalvin Kao

1. Introduction

This study seeks to investigate survival factors aboard the Titanic, which was the largest vessel afloat at the time of its construction, but also one of the most catastrophic shipwrecks of the 20th century. The Titanic began her maiden voyage from Southampton, England to New York in April of 1912, only to collide with an iceberg and sink four days later. Only 712 of the 2208 people aboard survived.

The event was well-documented, and our team obtained data surrounding survival of the passengers aboard. These passenger details help paint a picture of the identities and composition of people aboard, and also raise interesting questions about the possible factors that related to death or survival when the ship sank. These questions are as follows:

- How did survival rates differ between men and women?
- How do the survival rates compare to other shipwrecks for men and women?
- Does cabin class or name prefix (ie. social class) affect the survival rate?
- How did survival rates differ between children and adults?
- Who was in the lifeboats?
- What was survivorship like among different points of origin?
- Did families survive more effectively than individuals? Did families survive better than other groups?
- Did the group or family size affect survival rate?
- Was there less documentation for lower class passengers?

2. The Data

The source for the data used in this discussion is the Encyclopedia Titanica. We used the ‘titanic3’ dataset, which describes the survival status of individual passengers on the Titanic, excluding crew. The dataset includes the variables described in figure 1:

Variable Name	Description
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd), proxy for socio-economic status
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age (in years; fractional if Age less than one)
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (pre-1970 British Pounds)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Fig. 1: Variable descriptions

To provide further clarification, figure 2 shows how siblings, spouses, parents, or children were defined by the ‘sibsp’ and ‘parch’ variables:

Relation	Definition
Sibling	brother, sister, stepbrother, or stepsister of passenger aboard Titanic
Spouse	husband or wife of passenger aboard Titanic (not including mistresses and fiances)
Parent	mother or father of passenger aboard Titanic
Child	son, daughter, stepson, or stepdaughter of passenger aboard Titanic

Fig. 2: Definitions of familial relations

3. Exploratory Data Analysis (EDA)

'sibsp' and 'parch'

For both 'sibsp' and 'parch', there are no indications that there is an issue with data integrity and there are no missing values. Figures 3 through 6 show that most passengers did not have family relations onboard (siblings, spouses, parents, or children), and the ones that did generally just had 1 sibling/spouse, or 1-2 parents/children onboard:

sibsp =	0	1	2	3	4	5	8
Count:	891	319	42	20	22	6	9

Fig. 3: Value counts of the 'sibsp' variable

parch =	0	1	2	3	4	5	6	9
Count:	1002	170	113	8	6	6	2	2

Fig. 4: Value counts of the 'parch' variable

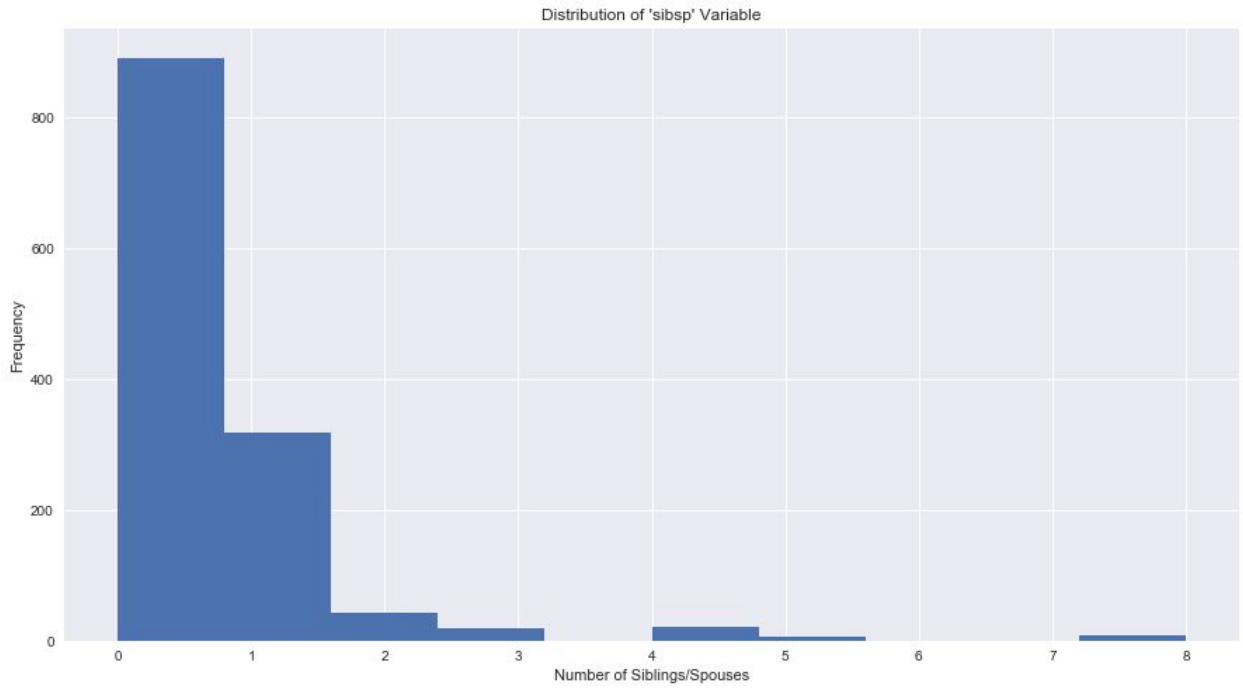


Fig. 5: Histogram showing the distribution of the 'sibsp' variable

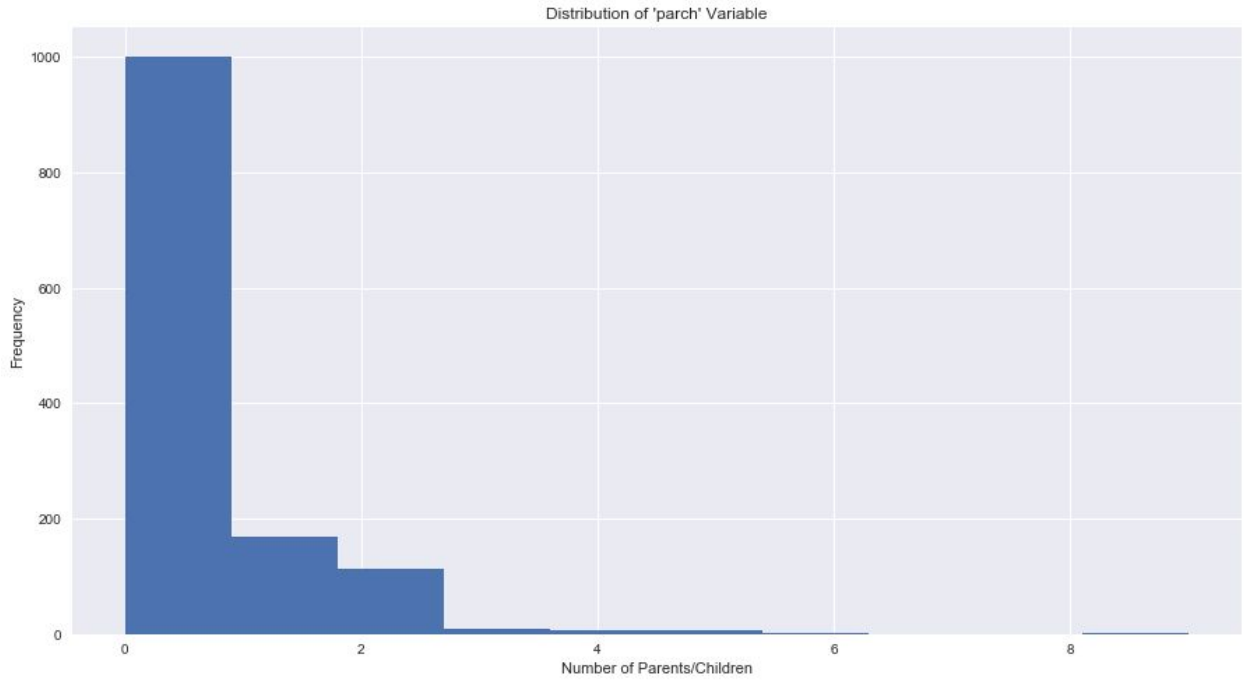


Fig. 6: Histogram showing the distribution of the ‘parch’ variable

The average and standard deviation of select other variables were aggregated for each ‘sibsp’ and ‘parch’ value, to gain a sense of how such family details relate to other metrics on the Titanic-- see figures 7 and 8.

sibsp	**pclass**		**survived**		**age**		**parch**		**individual fare**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
0	2.35	0.82	0.35	0.48	30.92	13.06	0.18	0.58	14.15	13.45
1	2.03	0.86	0.51	0.50	31.06	15.97	0.72	1.23	17.77	13.94
2	2.33	0.79	0.45	0.50	23.57	14.91	0.57	0.83	13.65	12.39
3	2.55	0.83	0.30	0.47	16.31	11.82	1.25	0.72	13.30	15.76
4	3.00	0.00	0.14	0.35	8.77	8.01	1.55	0.51	5.14	1.00
5	3.00	0.00	0.00	0.00	10.17	5.19	2.00	0.00	5.86	0.00
8	3.00	0.00	0.00	0.00	14.50	NaN	2.00	0.00	6.32	0.00

Fig. 7: Aggregated statistics for each sibsp value

parch	pclass		survived		age		sibsp		individual fare	
	mean	std	mean	std	mean	std	mean	std	mean	std
0	2.31	0.84	0.34	0.47	31.97	12.42	0.24	0.50	14.66	12.47
1	2.16	0.85	0.59	0.49	24.97	18.41	1.03	1.05	16.71	19.30
2	2.30	0.83	0.50	0.50	18.98	14.56	1.90	2.38	13.34	12.12
3	2.13	0.83	0.63	0.52	38.88	16.30	0.88	0.64	15.13	13.95
4	2.33	1.03	0.17	0.41	47.60	14.43	0.67	0.52	17.71	20.24
5	3.00	0.00	0.17	0.41	39.33	1.03	0.67	0.52	4.74	0.48
6	3.00	0.00	0.00	0.00	41.50	2.12	1.00	0.00	5.86	0.00
9	3.00	0.00	0.00	0.00	NaN	NaN	1.00	0.00	6.32	0.00

Fig. 8: Aggregated statistics for the ‘parch’ variable

The 6 individuals for whom there were 5 siblings/spouses were a family unit (the Goodwin family). The 9 individuals for whom there were 8 siblings/spouses were also a family unit (the Sage family). This is also indicated in figure 7 by the value of 2 for 'parch' for those groups. Figure 8 similarly shows that there are 2 individuals for each of ‘parch’ values of 6 and 9 (indicated by ‘sibsp’ == 1), and these are the parents. These largest families were in the lowest class and did not survive. In figure 7, the reason that there is a NaN value for the standard deviation of age in the case of sibsp=8 is because the age is known for only one individual in that group. In figure 8, the reason why the mean and standard deviation of the age is NaN for 'parch'== 9 is because the ages of the two individuals in this group (the parents) are unknown.

The exploratory data analysis of the ‘sibsp’ and ‘parch’ variables provides a sense of the fraction of passengers who had familial relations aboard the Titanic, and raises further questions about how family relations, group size, and passenger class related to survival on the Titanic

‘ticket’ and ‘fare’

With regards to the ‘ticket’ variable, there are no indications that there is an issue with data integrity and there are no missing values. It was noted that the ticket IDs vary in length and sometimes include an additional string prefix. Interestingly, it was found that ticket IDs are not unique for each passenger-- a large number of passengers traveled in a group, under a single ticket ID. Figure 9 shows the 5 largest ticket ID groupings:

Ticket ID	Group Size	Standard Deviation of 'fare'
CA. 2343	11	0.0
CA 2144	8	0.0
1601	8	0.0
PC 17608	7	0.0
S.O.C. 14879	7	0.0

Fig. 9: Ticket numbers that had the highest number of passengers sharing that ticket ID, as well as the standard deviation of the ‘fare’ listed for each individual in each ticket group.

The Sage and Goodwin families mentioned previously were under ticket IDs ‘CA. 2343’ and ‘CA 2144’, respectively. The presence of other large group sizes that did not appear in the large values of ‘sibsp’ and ‘parch’ indicates that there were additional large groups onboard that were not families. The ticket ID grouping also shows that the ‘fare’ variable is the same for every individual under that ticket ID, since it represents the total fare for the entire ticket group. This finding prompted the engineering of an additional feature in the data that will be discussed later-- the ‘individual fare’ (‘fare’ / group size).

The ticket prefix that exists for some ticket IDs was also examined for possible patterns. Figure 10 shows aggregated statistics for the most common prefixes-- the count of tickets with that prefix, the percentage of passengers under that prefix that survived, and also the average passenger class for each ticket prefix:

Ticket Prefix	Count	Survival Rate	Average Passenger Class
PC	92	63.04%	1
CA	68	33.82%	2.51
A/5	25	8.00%	3
SOTON/OQ	24	16.67%	3
SC/PARIS	19	57.89%	2
W/C	15	20.00%	2.67
STON/O	14	35.71%	3

Fig. 10: Aggregated statistics for the most common ticket ID prefixes

The figure shows that it is likely that the ticket prefix is related to passenger class, but it is not immediately apparent that other patterns exist. Additional processing on the ticket prefix using regular expressions may be helpful-- for example, it is possible that the prefix 'SOTON' and 'STON' are the same, and owe their difference to a coding or input error.

A number of data integrity issues were found in the 'fare' variable and are discussed in section 4.6. Furthermore, a transformation was required for this variable in order to perform a more meaningful analysis (discussed in section 5.2).

4. Data Integrity and Missing Values

Some data integrity issues were discovered during the initial data exploration, and are discussed in this section.

4.1 Missing Ages

The ages contains 263 or 20% NaN's or missing data for the ages of the passengers. For the given data frame we don't have any features that are highly correlated enough to age to accurately try to predict the ages for the missing values. We did consider using a mean or median our grouping method to fill in the missing ages as well as a linear regression model, however all of these methods were problematic and created issues with our data analysis. For this reason we are removing the age data that is missing from our data frame.

	name	**age**
count	1309	1046.000000

Fig. 11: Missing values for age feature.

	pclass	survived	age	sibsp	parch	fare	body
pclass	1.000000	-0.320486	-0.408106	0.047221	0.017224	-0.565255	-0.034122
survived	-0.320486	1.000000	-0.055513	-0.012213	0.114438	0.249164	NaN
age	-0.408106	-0.055513	1.000000	-0.243699	-0.150917	0.178739	0.058809
sibsp	0.047221	-0.012213	-0.243699	1.000000	0.374456	0.141184	-0.100289
parch	0.017224	0.114438	-0.150917	0.374456	1.000000	0.216723	0.050902
fare	-0.565255	0.249164	0.178739	0.141184	0.216723	1.000000	-0.043514
body	-0.034122	NaN	0.058809	-0.100289	0.050902	-0.043514	1.000000

Fig. 12: Correlation Matrix for each numerical feature in our data frame.

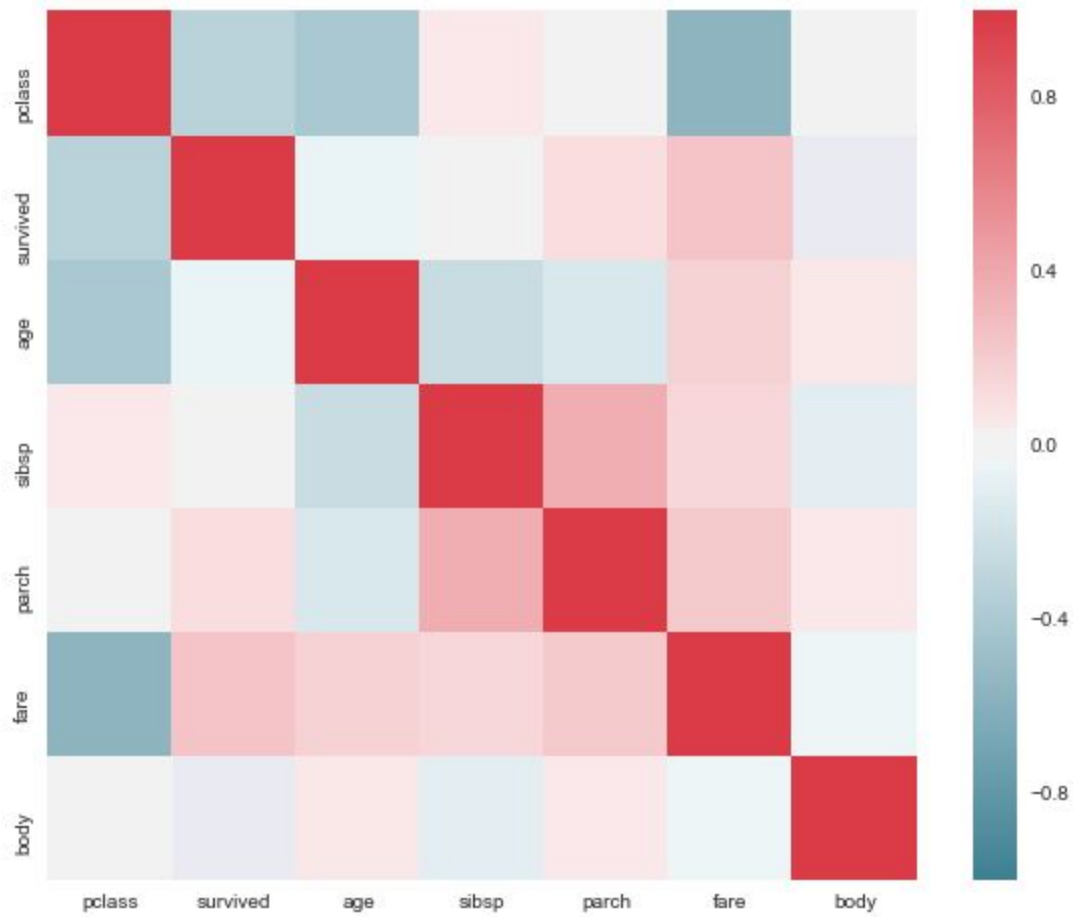


Fig. 13: Correlation heat map for each of the numerical features in our data frame.

4.2 Missing Cabins

There are 1014 missing cabin values. Missing values are overwhelmingly from 3rd and 2nd class passengers; 693 missing values vs 709 third class passengers and 254 missing 2nd class values vs 277 2nd class passenger. Missing values are also mostly males; 689 missing values vs 843 males

4.3 Missing Body Count

There are 1189 missing values body count values. This variable is only recorded for those who died. Most of missing values come from third class and second class; 654 third class, 246 from second class. It is hard to tell how this was variable was recorded as bodies are likely to have drowned without having been tabulated for after the fact.

4.4 Missing Home_Dest (Home variable)

This variable 564 missing values of which 411 come from those who didn't survive. Many of these missing values are also third class passengers; 514 missing values from third class passengers

4.5 Missing Boat

There are 824 missing values, which is close to the number of people who survived. Almost everyone who was in a boat survived. Only 9 people recorded in boats died. 23 people don't have a boat record but survived, who were possibly no-shows. The interactive plot let's you see the composition of each boat

4.6 Missing Fare

A number of data integrity issues were found in the 'fare' variable: there were 17 individuals for whom the fare was zero, and there was one missing value (Mr. Thomas Storey, ticket 3701). Additional, ticket ID '7534' has different fares listed for its two members. Fortunately, there is still enough data for this variable to investigate how ticket price related to survival.

5. Data Transformations

The exploratory data analysis revealed opportunities to engineer additional features that could also be useful for understanding the factors related to survival.

5.1 Names

The names feature is in the format: surname, prefix. firstname. We think that the prefix may hold useful information for determining the age or survivability of a passenger. For example, a

prefix of Miss means that a woman is unmarried and may indicate that she is younger or a widower and could be correlated to a younger or older age group. We also think that preference could have possibly been given to those with higher social class for lifeboats or exiting the ship which may map to the prefix of their name. Figure 14 shows the firstname, lastname, and prefix and the first 5 rows of the data frame.

	name	**firstname**	**lastname**	**prefix**
0	Allen, Miss. Elisabeth Walton	Elisabeth Walton	Allen	Miss
1	Allison, Master. Hudson Trevor	Hudson Trevor	Allison	Master
2	Allison, Miss. Helen Loraine	Helen Loraine	Allison	Miss
3	Allison, Mr. Hudson Joshua Creighton	Hudson Joshua Creighton	Allison	Mr
4	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	Hudson J C (Bessie Waldo Daniels)	Allison	Mrs

Fig. 14: Example of name split function to extract the firstname, lastname, and prefix.

5.2 Fare

'individual_fare'

The individual fare could not be calculated for the individual whose 'fare' was missing. For the remaining 1308 passengers, their individual fares were distributed as shown in figures 15-16.

Fare Range ($)	Count
0-20	980
20-40	244
40-60	56
60-80	4
80-100	3
100-120	0
120-140	4

Fig. 15: Counts for each range of individual fare or ticket value

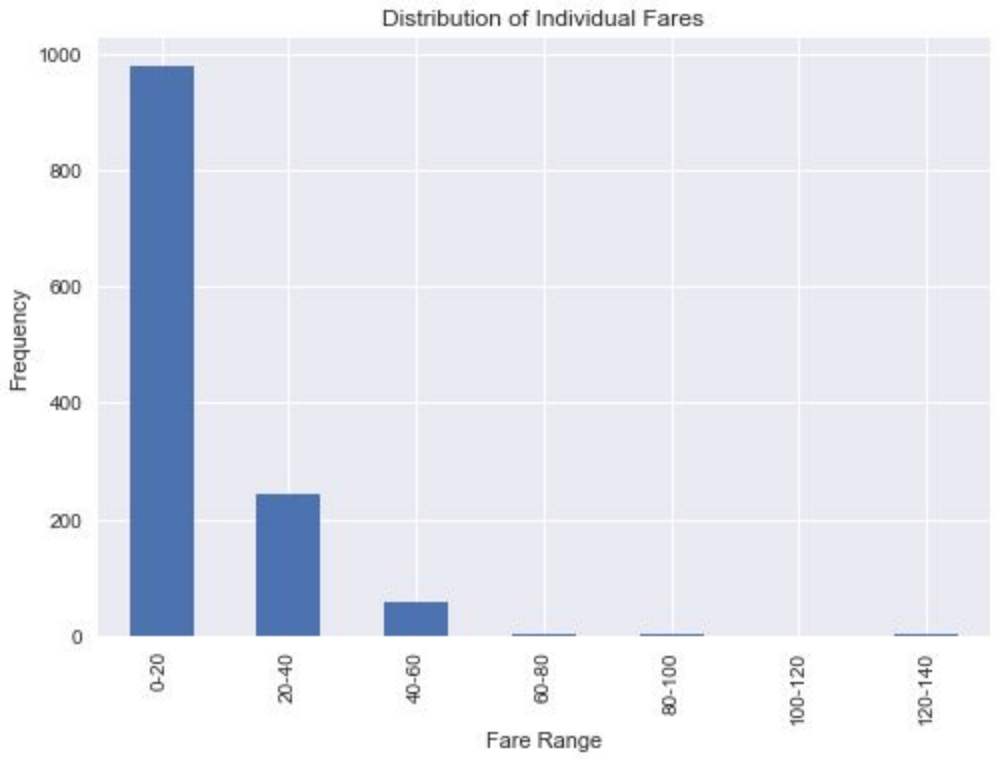


Fig. 16: Histogram of the individual fares or ticket values

The vast majority of the ticket values were under $20. The data showed a strong correlation (-0.77) between the individual fare and the passenger class, which makes sense-- 3rd class tickets are the cheapest while 1st class tickets are the most expensive.
It is interesting to see that a handful of individual fares ranged from $60 to a maximum of $128.08 and there may have been something special about this group of passengers-- only 2 of these 11 individuals did not survive-- Mr. Quigg Baxter and Mr. Walter Clark, who were both young adult males.

5.3 Cabin

Cabin values are indicated by a letter, followed by a number (ie C55). Some records have multiple cabin values. We are interested in seeing cabins as grouped by their prefix letter. In creating this variable we take the first letter of each cabin record. We also assume the same cabin letter for those records which contain multiple cabin values. Only one of these records has cabin values spanning different letters.

An interactive plot in the python notebook shows the composition by class and sex for each cabin letter

5.4 Home_Dest

This variable should have both a home and destination place in this following format - home/destination. However, many destination values are missing. For this EDA, we assume that the recorded values are home values and will concern ourselves with this part of the variable. A new column, home, was created for this.

An interactive plot in the python notebook shows the composition by class and sex for each the top six home values.

6. Questions about Survival

Our investigations into the data led to the following discussion points related to survival.

6.1 How did survival rates differ between men, women, and children?

The figures 17 and 18 below in terms of count and percentage show that there were a significantly more males (577) on board the Titanic compared to the number of females (314). This may be for many reasons one plausible reason is that it was typical for males to immigrate first and then send for their families when they were established at their new locations. The second plot (% of survivors by gender) also verifies and shows that females had a higher proportion (74.2%) of survivors compared to males (18.9%). Both of these figures confirm that females had a higher rate of survival on the Titanic compared to males, and that the maritime rule of women and children first may have been followed.

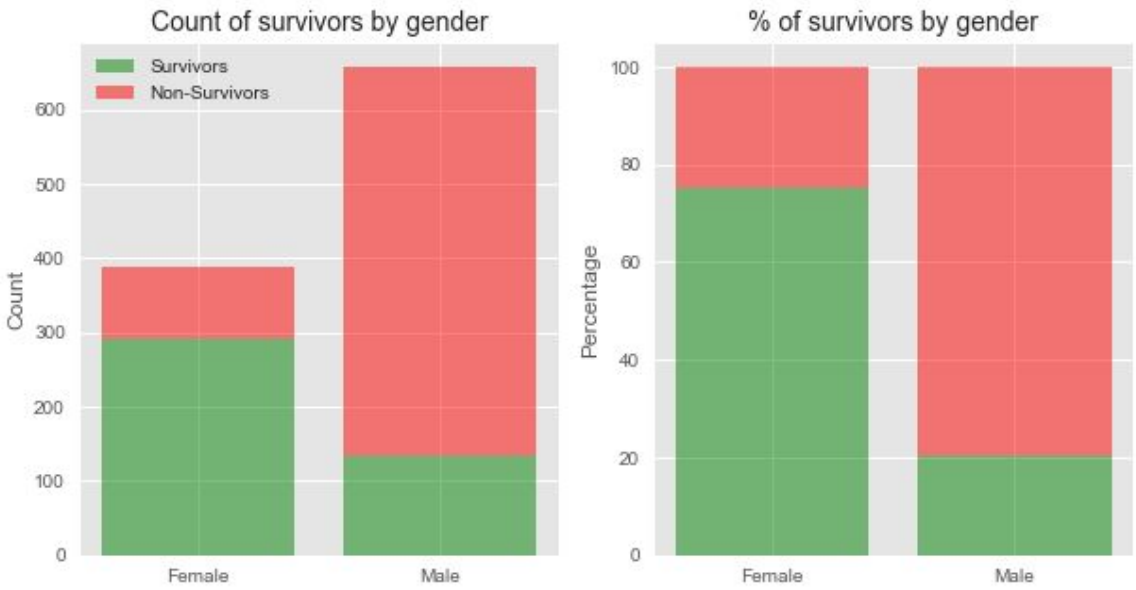


Fig. 17, 18: Count and percentage of male vs female survivors on the Titanic.

6.2 How do the survival rates compare to other shipwrecks for men and women?

We were also interested in determining whether the maritime rule of women and children first for the Titanic was generally true for other shipwrecks. We gathered data sets for 18 other shipwrecks spanning three centuries, 15,000 passengers, and 30 different nationalities. The figures 19 and 20 below in terms of count and percentage show that with the exception of the Titanic and the Birkenhead typically men have a higher survival rate than women. There are 2x the number of men on the ships and 5% more of the men survive compared to the women. There could be many reasons for this, including the time and duration of the evacuation, the captain/crews instructions, cabin demographics/proximity to lifeboats, etc. but at the very least it calls into question the validity of the adherence to the widely accepted women and children first maritime rule.

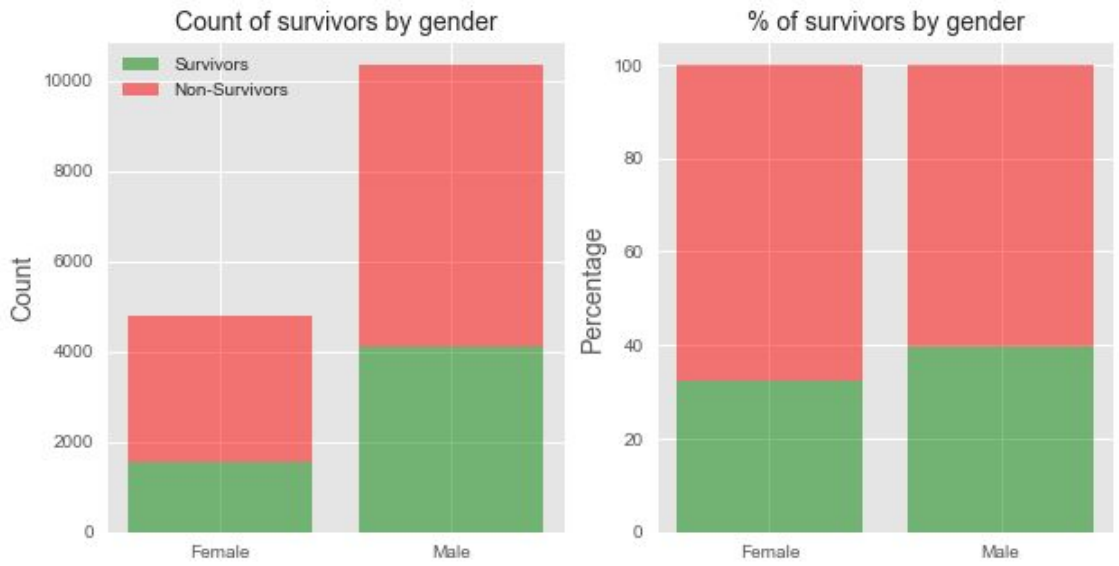


Fig. 19, 20: Count and percentage of male vs. female survivors for 18 different shipwrecks.

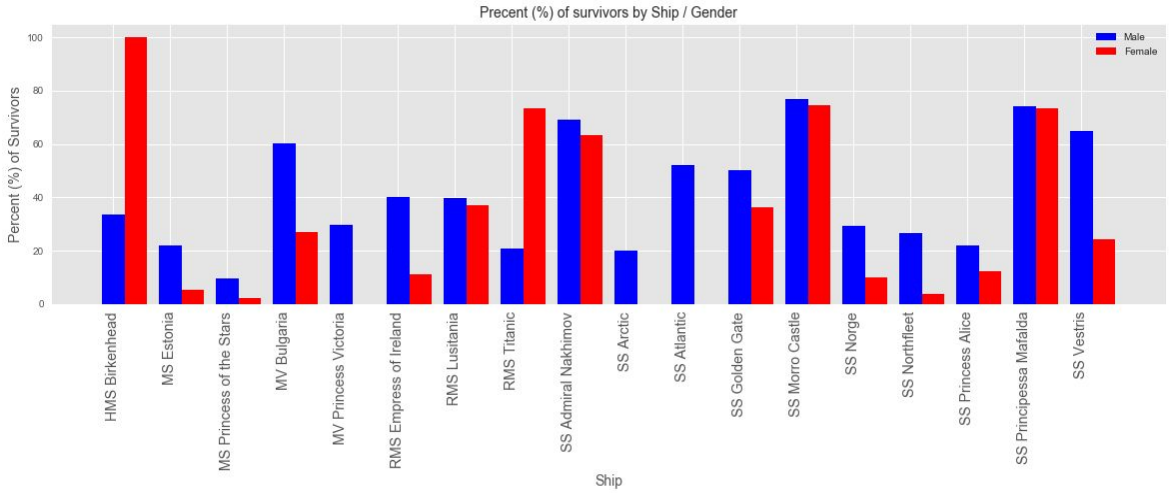
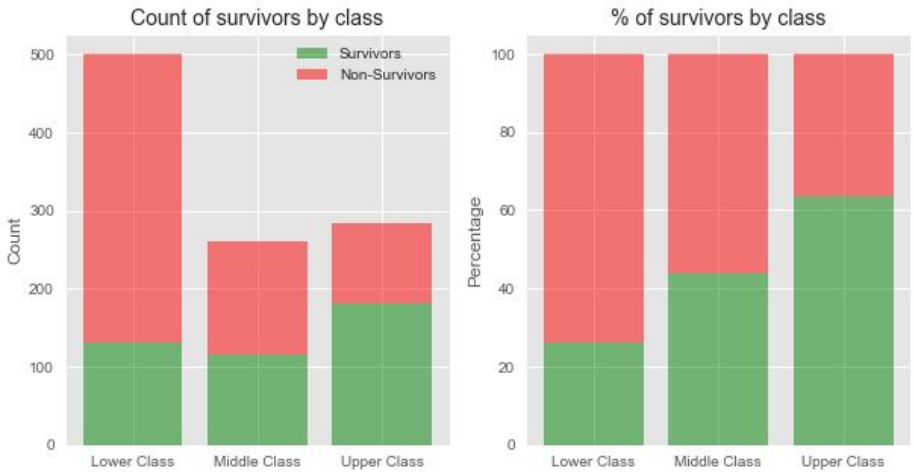


Fig. 21: Percent survival for male vs. female by ship for 18 different shipwrecks.

6.3 Does passenger class or name prefix (ie. social class) affect the survival rate?

We used passenger class and name prefix as indicators to determine if social class affects survival rates. The figures 22 and 23 below show that passengers in lower class had 2-3x more deaths by count than passengers in middle and upper class which were more comparable. In terms of percentages lower class lower class class had the highest percentage of non-survivors at almost 63% while almost 250% more passengers survived in upper class than lower class. It should be noted that the reasons for this may not be singularly due to social class and may depend more on proximity to lifeboats since upper class passengers were located on higher decks of the Titanic as shown in figure 24. Figures 25 and 26 show the count and percentage of survivors by name prefixes. The he Mr.'s had the worst survival rates while the Miss's and Mrs.'s, Masters, Mlle, Mme, Dona, and the Countess had high survival rates. This may indicate preference towards certain individuals based on their social class and again reconfirms that the males had a lower survival rate than females.



.

Fig. 22, 23: Counts and percentages of survivors by passenger class.

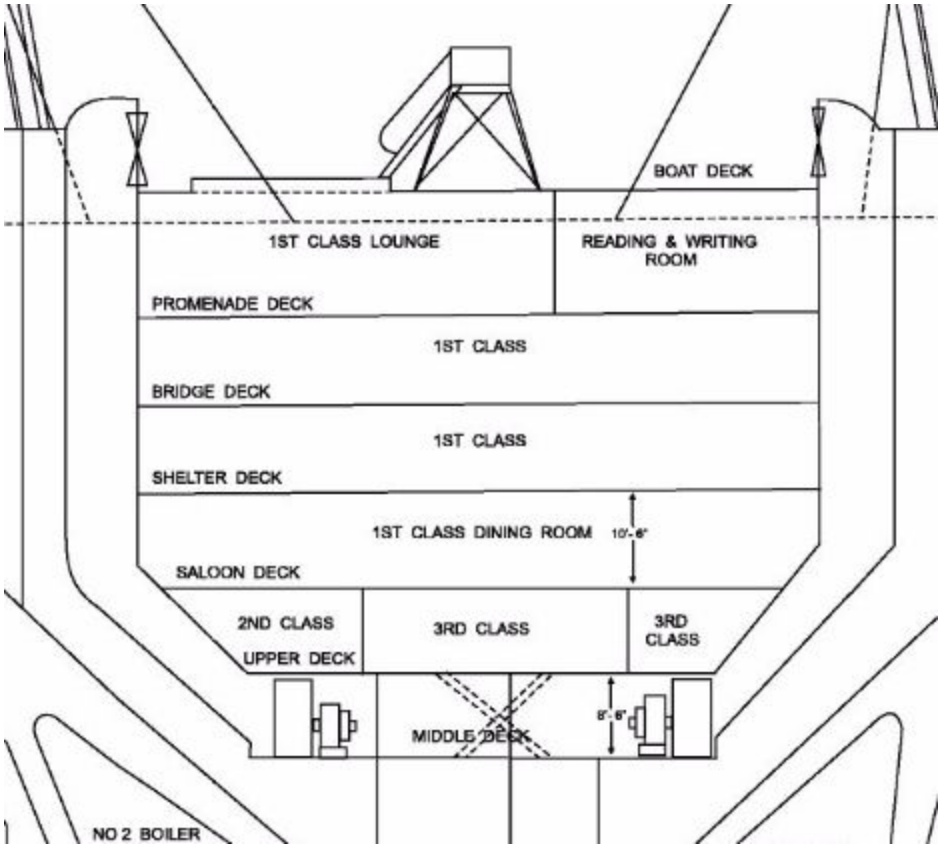


Fig. 24: Representative Titanic deck/class cross section.

Fig. 25, 26: Counts and percentages for survivors by name prefix.

6.4 How did survival rates differ between children, adults, and elderly?

We were also interested in survival rates for different age groups. One difficult question is how you classify different age groups as children/adult/elderly so we picked age groupings bins of 10 years to give enough granularity while still providing a functional classification of groups. Figure 27 below shows that by count the passengers from 20-39yrs had the least survivors, while other groups varied. By percentage the 0-9yrs group had the highest survival, while the 10-59yrs group had equal percentages, and the > 60yrs had diminishing survival percentages. Drilling down further, Figure 29 shows that male and females in the 0-9yrs group which could be classified as children had much higher survival rates than all other age groups which could be classified as adults and elderly. Furthermore, we see again in figure 29 that females across age groups > 10yrs had a much higher survival rate than men again supporting the women and children first model on the Titanic.

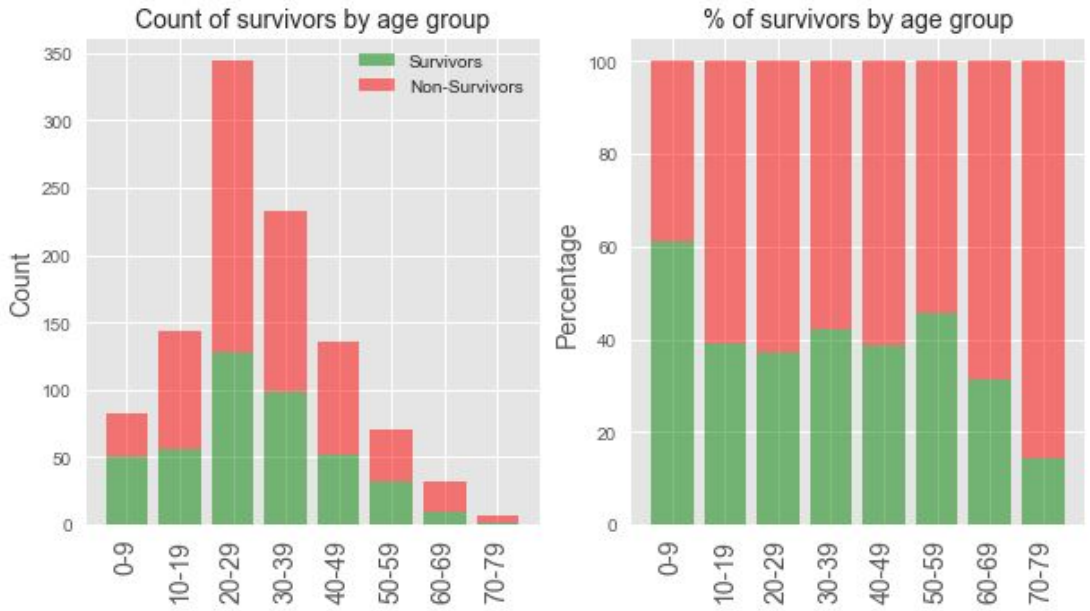


Fig. 27, 28: Counts for each range of individual fare or ticket value

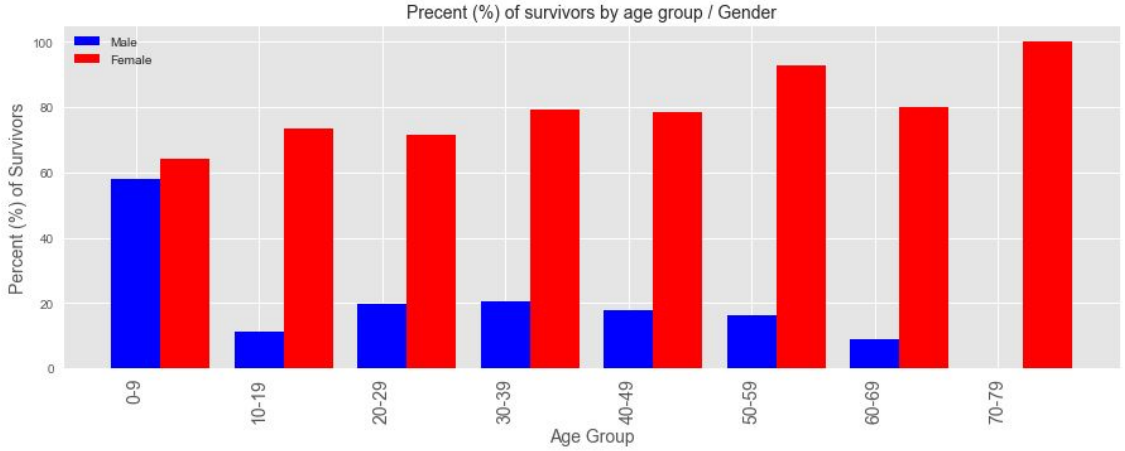


Fig. 29: Counts for each range of individual fare or ticket value

6.5 Who was in the lifeboats?

Everyone, except for 9 people recorded in lifeboats survived. There are 27 passengers who don't have a boat record, but survived. These are probably no-shows. First class passengers were more likely to make it onto boats. 201 out of 323 first class passengers made it to the boats. 112 out of 277 second class passengers made it to the boats. 173 out of 709 third class passengers went on to the boats

The table below shows percentages taken onto boats from each class

	pclass	percent taken
pclass		
1.00	201	0.62
2.00	112	0.40
3.00	173	0.24

Fig. 30: Percentages of each class taken onto boats

The notebook shows an interactive plot broken down by class and sex per lifeboat

6.6 What was survivorship like among different points of origin?

The bulk of passengers embarked from Southampton (S in the data), followed by Cherbourg, and Queenstown. Passengers from Cherbourg have the highest survival rate, at 56% whereas Queenstown and Southampton have survival rates of 36% and 33%.
About half of the passengers from Cherbourgh are in 1st class, whereas 2% and 19% of the passengers were first class from Queenstown and Southampton. 40%, 91%, and 60% of passengers from Cherbourgh, Queenstown, and Southampton were in third class. Within each class, however, survival rates of passengers from Cherbourgh are higher than average.

The table below shows the percentage of survival from each point of embarkment

Port	People Boarded	% Survived
C	270	56%
Q	123	36%
S	914	33%

Fig. 31: Survival Rates of Each Point of Embarkment

The table below further extrapolates on the above table, giving count as the total people boarded, sum as those that survived and mean as percentage of survivors.

		survived		
		count	sum	mean
pclass	embarked			
1.00	C	141	97.00	0.69
	Q	3	2.00	0.67
	S	177	99.00	0.56
2.00	C	28	16.00	0.57
	Q	7	2.00	0.29
	S	242	101.00	0.42
3.00	C	101	37.00	0.37
	Q	113	40.00	0.35
	S	495	104.00	0.21

Fig. 32: Count and Survival Rate of Each Class, from Each Point of Embarkment

The graph below shows survival rates for each class across the differing points of embarkment

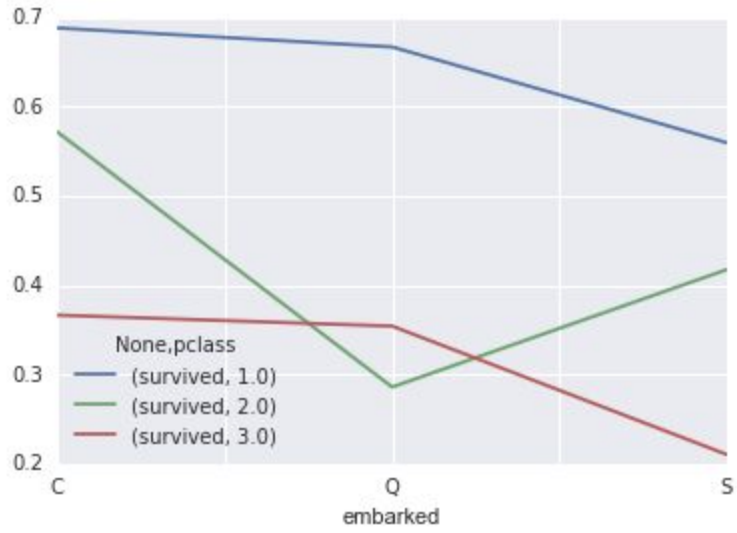


Fig. 33: Survival Rates for Each Class from Different Points of Embarkment

6.7 Did families survive more effectively than individuals? Did families survive better than other groups?

Examination of the ticket ID groupings shows that most passengers traveled as individuals, but many traveled in a group-- Figures 34-35 show the counts of each group size, which is defined as the number of individuals under a single ticket ID.

Group Size	Count
1	713
2	132
3	49
4	16
5	7
6	4
7	5
8	2
11	1

Fig. 34: Number of groups of each size

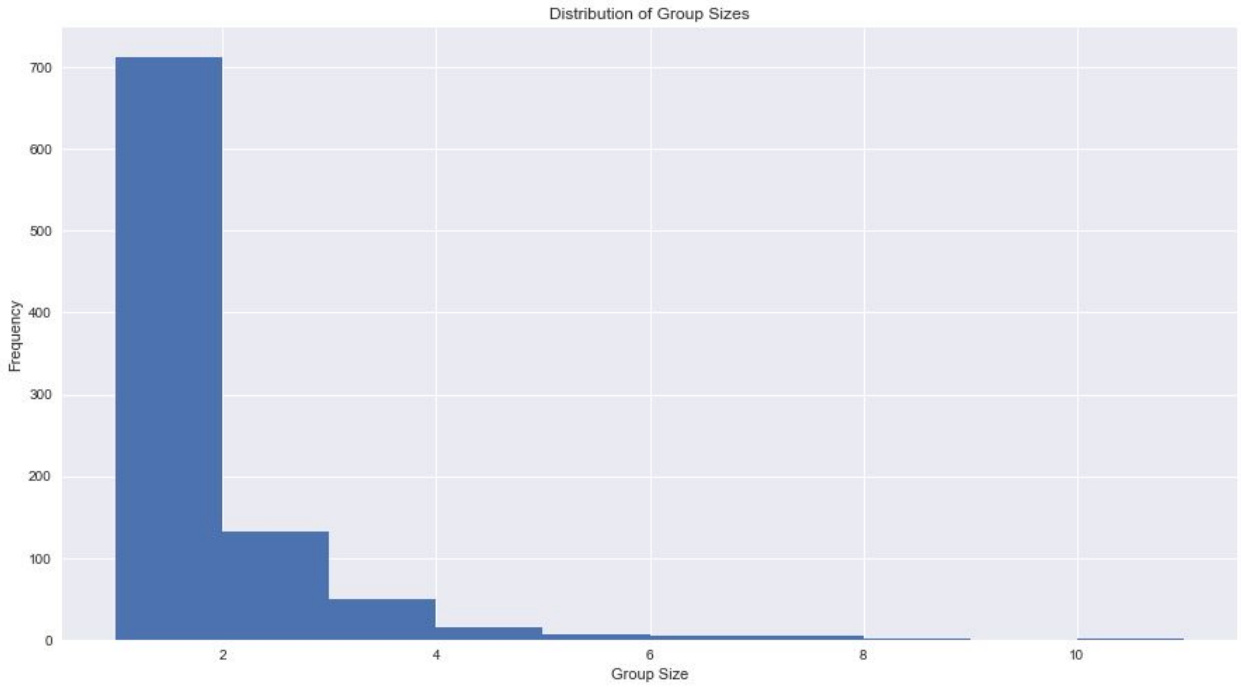


Fig. 35: Histogram of group size

Some of these groups had a single last name and were clearly family units (confirmed by counting the ‘sibsp’ and ‘parch’ variables), while others were not. Figures 36-37 show passenger information for two ticket IDs-- one is family and the other is not.

Data ID	Name	Ticket
1170	Sage, Master. Thomas Henry	CA. 2343
1171	Sage, Master. William Henry	CA. 2343
1172	Sage, Miss. Ada	CA. 2343
1173	Sage, Miss. Constance Gladys	CA. 2343
1174	Sage, Miss. Dorothy Edith "Dolly"	CA. 2343
1175	Sage, Miss. Stella Anna	CA. 2343
1176	Sage, Mr. Douglas Bullen	CA. 2343
1177	Sage, Mr. Frederick	CA. 2343
1178	Sage, Mr. George John Jr	CA. 2343
1179	Sage, Mr. John George	CA. 2343
1180	Sage, Mrs. John (Annie Bullen)	CA. 2343

Fig. 36: Names of members of the Sage family, traveling under ticket number ‘CA. 2343’.

Data ID	Name	Ticket
0	Allen, Miss. Elisabeth Walton	24160
180	Kreuchen, Miss. Emilie	24160
193	Madill, Miss. Georgette Alexandra	24160
238	Robert, Mrs. Edward Scott (Elisabeth Walton Mc...	24160

Fig. 37: Names of individuals traveling under ticket number ‘24160’.

To investigate details surrounding the group sizes and group types further, each ticket ID was categorized as an ‘individual’, a group of ‘friends’, or a group of ‘family,’ and the survival rates of the individuals belonging to each type of ticket ID were compared, see figure 38.

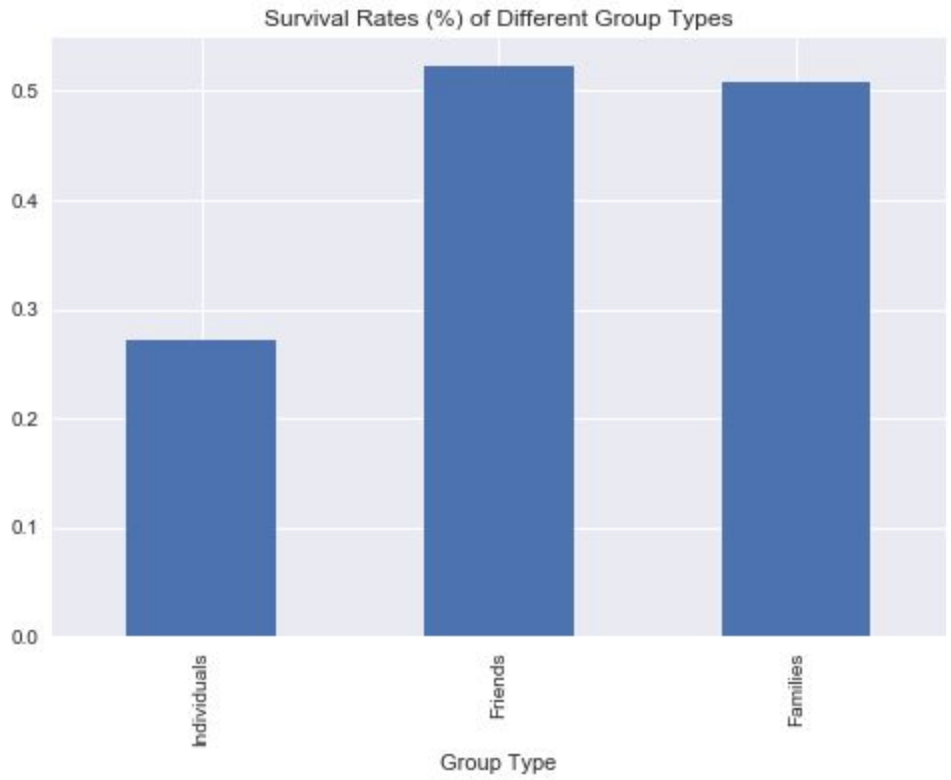


Fig. 38: Survival rates of passengers who traveled as individuals, with family, and with non-family.

While the data indicates that a family group, as defined by the analysis in this report, did not have greater survivability than other types of groups, the data does suggest that traveling in at least some sort of group did lead to a higher survival rate. In the case of ticket number '24160' (whose members are shown in figure 37), each passenger has a different last name but further investigation into their details reveals that they are actually related (just not a nuclear family, like the Sages). Strictly speaking, it is not completely clear how many of the groups were families.

6.8 Did the group or family size affect survival rate?

Another potential pattern was detected in how the group size related to survival, shown in figure 39.

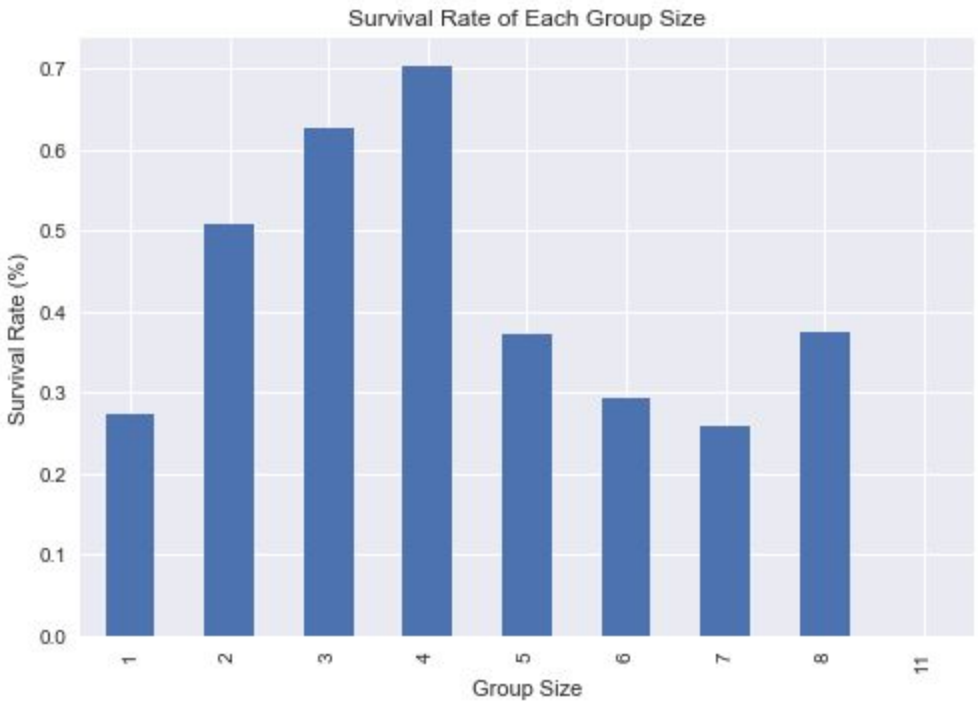


Fig. 39: The survival rate of the passengers in each group size.

The analysis suggests that there may have been an optimal group size (4) that perhaps enabled passengers to assist each other while also fitting all members into a lifeboat. While it is not likely that all members of a large group would survive (due to sheer probability) the difference in survival rate of individuals in small groups (2-4 members) from that of individuals in large groups (>4 members) is notable.

6.9 Was there less documentation for lower class passengers?

Due to the data integrity issues found in the Titanic dataset, the amount of missing data was investigated further. Figure 40 shows the percentage of missing values in each variable, grouped by passenger class.

class	age	fare	cabin	embarked	home.dest	boat (if survived)	body (if dead)
1	12.07 %	0.00%	20.74%	0.62%	10.53%	0.50%	71.54%
2	5.78%	0.00%	91.70%	0.00%	5.78%	6.72%	80.38%
3	29.34 %	0.14%	97.74%	0.00%	72.50%	7.73%	89.58%

Fig. 40: Percentage of missing data in each variable, for each passenger class. For the boat variable, the figure provides the percentage of surviving passengers whose boat number is missing. For the body variable, the figure provides the percentage of dead passengers whose body ID is missing. As mentioned previously, the other variables not shown did not have integrity issues or missing values.

The data shows that in many cases, i.e. ‘cabin’, ‘home.dest’, ‘boat’, and ‘body’, there may be a trend in the amount of documentation gathered, with respect to passenger class-- lower class passengers tended to have more missing information-- this is shown again in figures 41-43. While not conclusive, this does raise some questions about whether or not information about lower class individuals was considered important. For example, did search teams at the wreckage site try harder to find or identify the bodies of higher class passengers? These findings also raise questions about biases in the accuracy of data in general surrounding lower class individuals from the early 1900s time period.

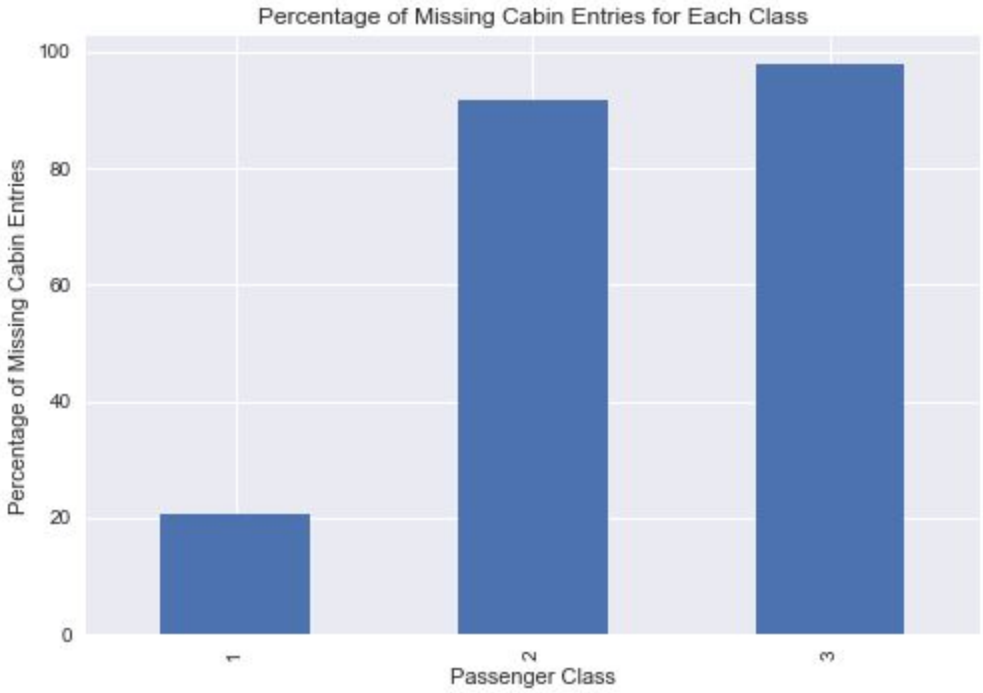


Fig. 41: Percentage of missing cabin entries for each passenger class.

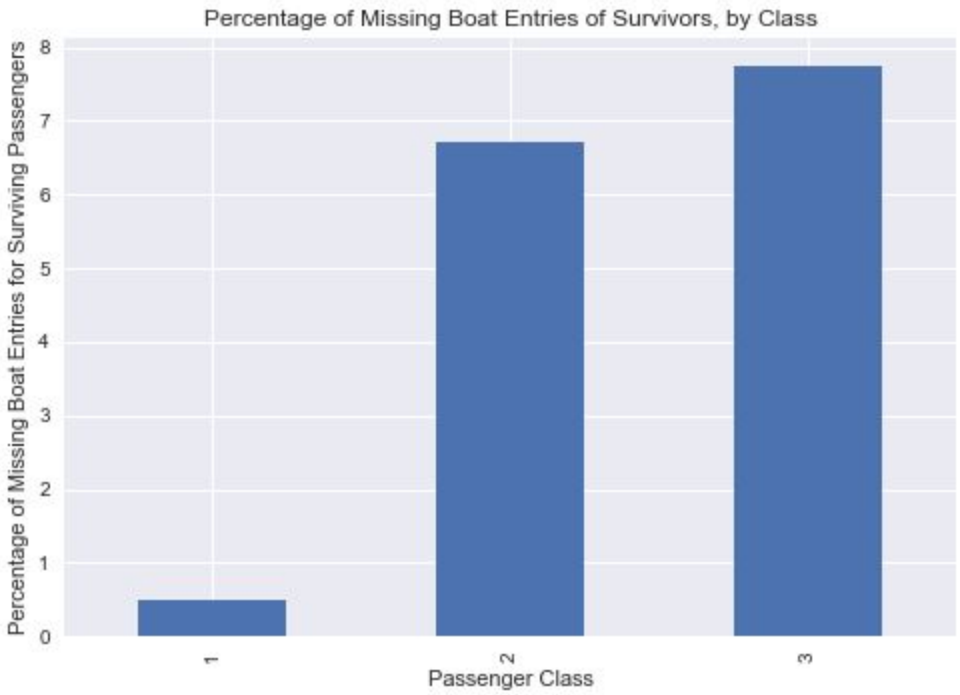


Fig. 42: Percentage of missing boat number entries for surviving passengers, by class.

Fig. 43: Percentage of missing body IDs for passengers who died on the Titanic, by class.

7. Conclusion

Women had much higher survival rate than men on the Titanic, though this is generally not true for shipwrecks looking at 18 different shipwrecks spanning 3 centuries, 15,000 passengers, and 30 nationalities men usually have a higher survival rate Children both male and female had a higher survival rate than adults and elderly. Higher social class, represented by boat class and name prefix does correlate to a higher survival rate.

While there isn't a clear difference in survival rate between groups of 'friends' and family groups, there is evidence that being in some sort of group helped. Additionally, there is some evidence that a mid-size group (~4 people) seemed to be optimal for surviving the Titanic wreck. With regards to missing data, we also noticed that lower class passengers tended to have more missing data, across many of our variables. Finally, an oddity of the data was that passengers who embarked from Cherbourgh happen to have higher survival rates across all classes.

Our advice to possible passengers on a would-be Titanic is, be a female who buys a first class ticket.

8. Works Cited

Elinder, Mikael, and Oscar Erixson. “Gender, Social Norms, and Survival in Maritime Disasters.” *Proceedings of the National Academy of Sciences of the United States of America* 109.33 (2012): 13220–13224. *PMC*. Web. 17 Aug. 2017.

Harrell, Frank E. “Titanic Data.” http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html. Vanderbilt University Department of Biostatistics, 27 Dec. 2002. Web. 3 Aug. 2017.

Hind, Philip. "Encyclopedia Titanica." Online. Internet.
n.p. 02 Aug 1999.