EXPLORATORY DATA ANALYSIS (EDA)

Introduction to Data Science

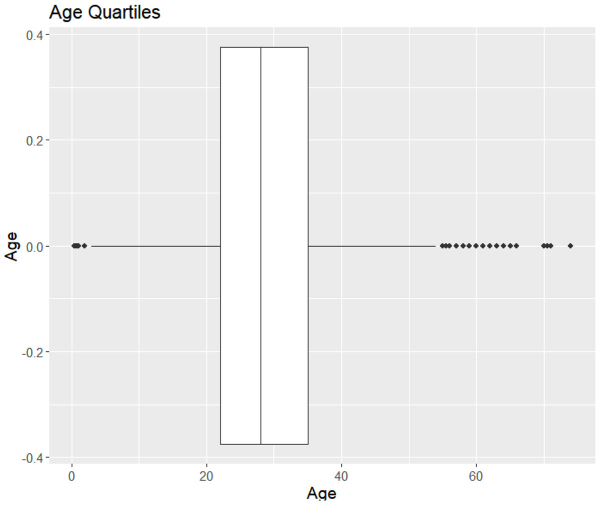
Universidad Carlos III de Madrid

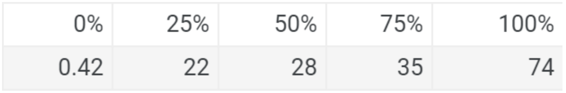
Jamison Biddle (100492595) & María Victoria Vivas Gutiérrez (100452684)

Group 96

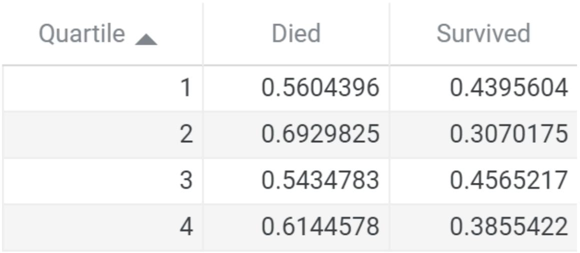
How does age quartile affect the survival rate?

In order to better understand how age impacts the rate of survival, the data is split into quartiles. This is shown in the first chart below, and with the box plot.

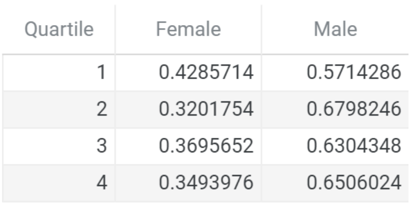




0%	25%	50%	75%	100%
0.42	22	28	35	74

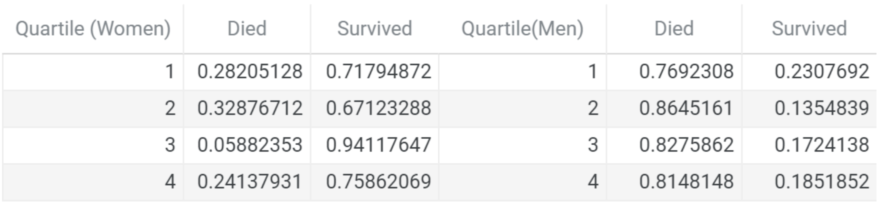


Quartile ▲	Died	Survived
1	0.5604396	0.4395604
2	0.6929825	0.3070175
3	0.5434783	0.4565217
4	0.6144578	0.3855422



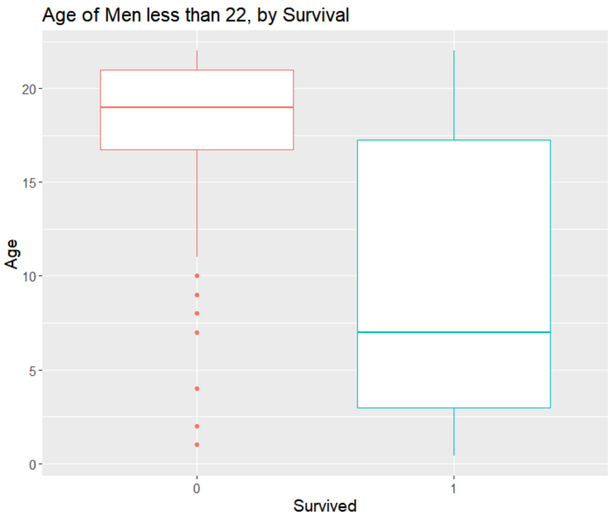
Quartile	Female	Male
1	0.4285714	0.5714286
2	0.3201754	0.6798246
3	0.3695652	0.6304348
4	0.3493976	0.6506024

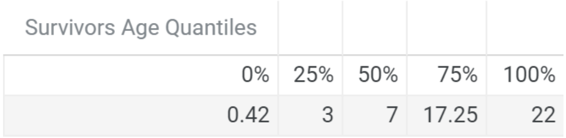
The quartiles were charted alongside the survival rate. There's not a huge difference between any quartiles, except that the second quartile (22-28) is more likely to die than expected. This sample includes a large group of men at age 28, so perhaps the quartile is overrepresented with men. This is supported by the table of Sex breakdown by quartile. The portion of survivors is almost identical as the ratio of women for the first quartile. This is surprising as a 10-year-old boy should survive, theoretically, as often as a 10-year-old boy. This leads to the question: How do the survival metrics differ for the men and women in each quartile?



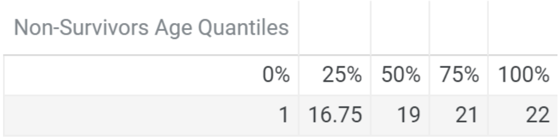
Quartile (Women)	Died	Survived	Quartile(Men)	Died	Survived
1	0.28205128	0.71794872	1	0.7692308	0.2307692
2	0.32876712	0.67123288	2	0.8645161	0.1354839
3	0.05882353	0.94117647	3	0.8275862	0.1724138
4	0.24137931	0.75862069	4	0.8148148	0.1851852

It's incredibly notable that almost every woman, 94%, between the ages of 28 & 35 survived. Also notable: Being under 22 doesn't actually dramatically improve the chances of survival; it is only about double or 1.5x the chance. So, at what age are men no longer "children"?





Survivors Age Quantiles				
0%	25%	50%	75%	100%
0.42	3	7	17.25	22

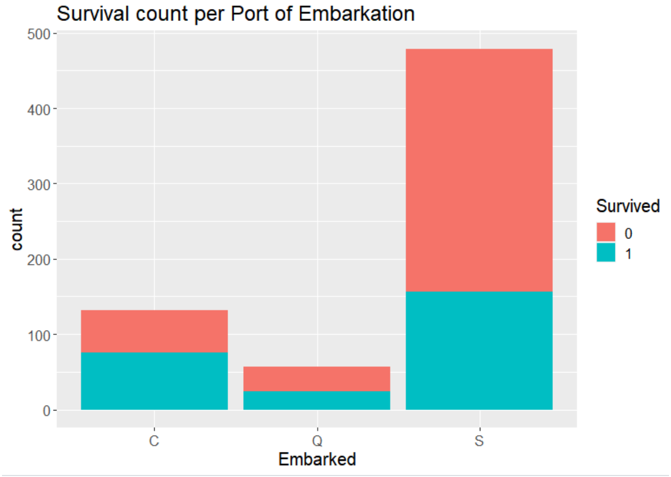


Non-Survivors Age Quantiles				
0%	25%	50%	75%	100%
1	16.75	19	21	22

Most of the survivors who are young and male are very young, with 75% of them being under 17.25, and a full 50% being under 7. Among those who died, 75% were above 16.75. So somewhere around the age of 17, boys begin to be considered "men" and thus wouldn't make it.

How does port of embarkation affect survival rates?

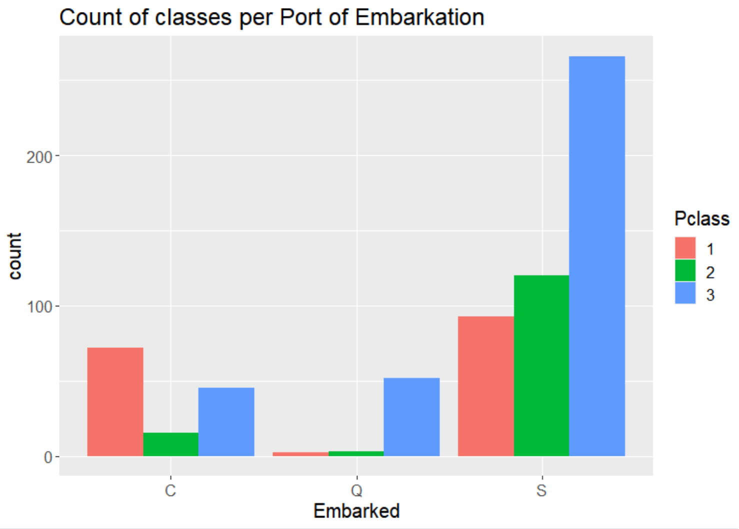
In an initial review, wherein all variables were graphed or put in a table comparing them to survival, there was an unexpected result. One would expect that, given roughly similar demographics embarking at each port, the survival rates would be similar.



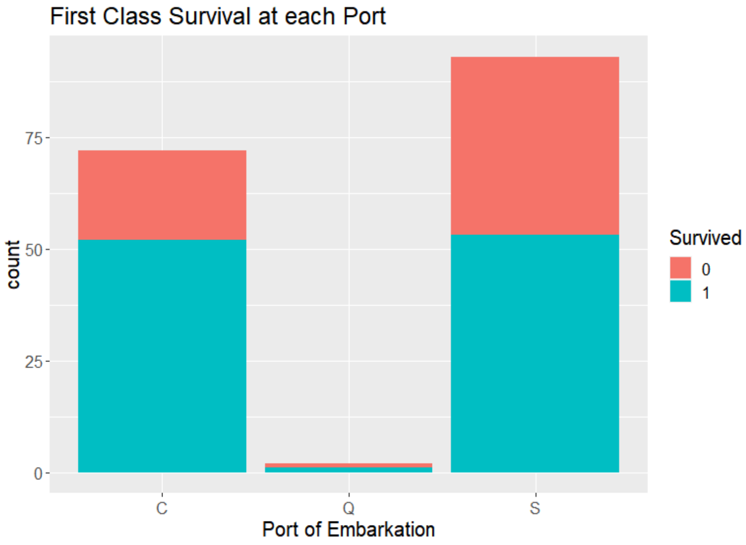
Those who embarked at Cherbourg had a higher survival rate than those who embarked at both Queenstown and Southampton; in fact, the only group with a better chance of survival than death is those who embarked at Cherbourg. The first consideration we reviewed was the class of those embarking.

This is consistent with the results, as people in third class were far more likely to die than people in first class. As a check, we'll compare the rate of survival for the first-class passengers at each port- this will be particularly notable for the Cherbourg/Southampton distinction, as there are similar amounts of first class for each.

This only provides more questions: Cherbourg's survival rate is still higher, even within first class. Thus, we'll check the other big culprit: Sex.

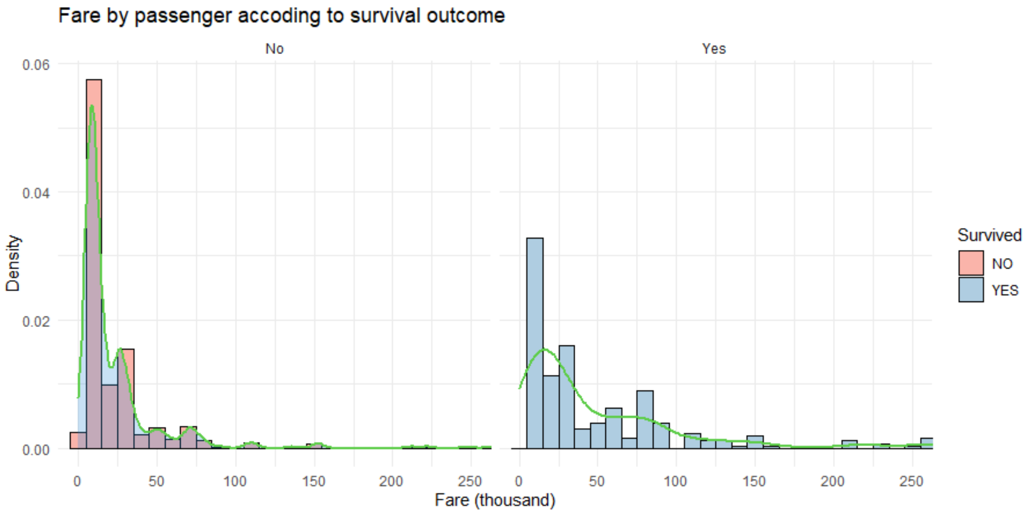


This seems to be notable: Cherbourg had a notably higher proportion of women among the first-class passengers, whereas Southampton had more men than women. This seems to be resolved, as with a much higher proportion of women from Cherbourg, the survival rate can be expected to be much higher too.



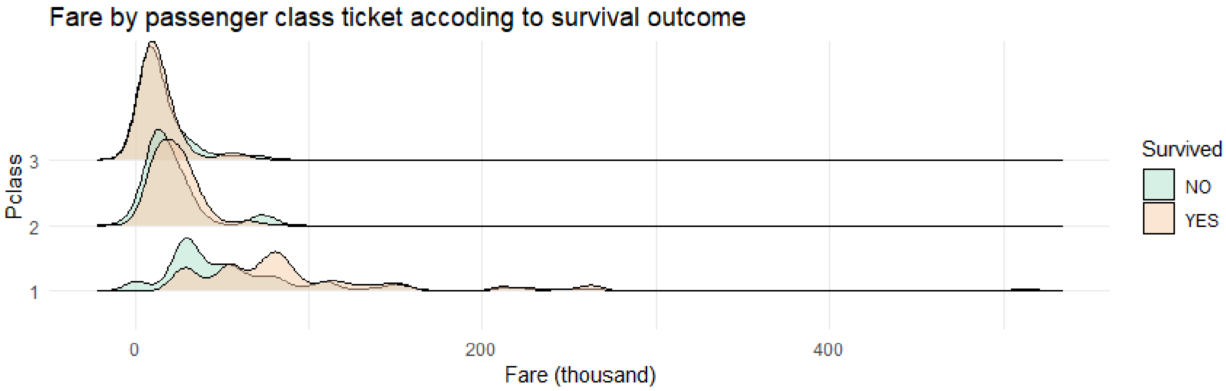
When the fare gets higher, the proportion of survivors is higher too?

After analyzing the effect of the port of embarkation in the survival outcome, it could be interesting to assess how the fare paid per passenger influenced, if it did it, the survival rate. One first possible approach leads to the expectation that when the fare gets higher, the proportion of survivors is higher too, which is related to the class. However, almost every passenger paid a similar fare. Thus, as it can be shown on the following figure, it is true that when the fare is incremented, apparently the proportion of passengers who did not survive is decreased.



Therefore, the survivor’s distribution is more like a leptokurtic distribution which is positively skewed and bimodal.

In the section above it was proved that people who bought the third-class ticket were more likely to die in relation to those who bought the first one. There were a small number of people who bought the second-class ticket for being relevant.

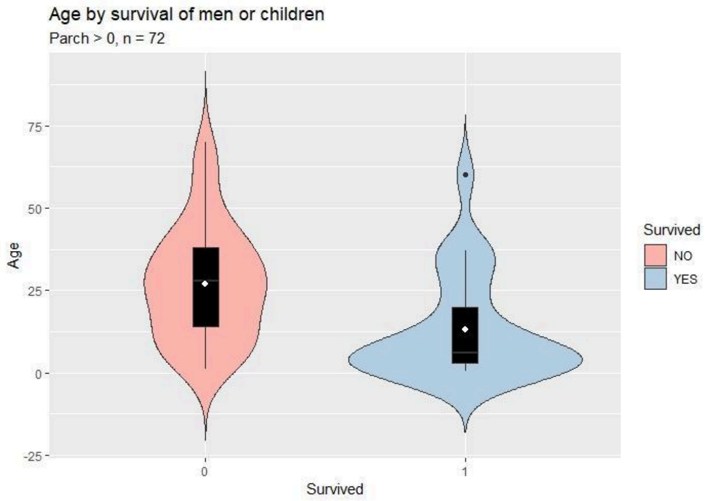


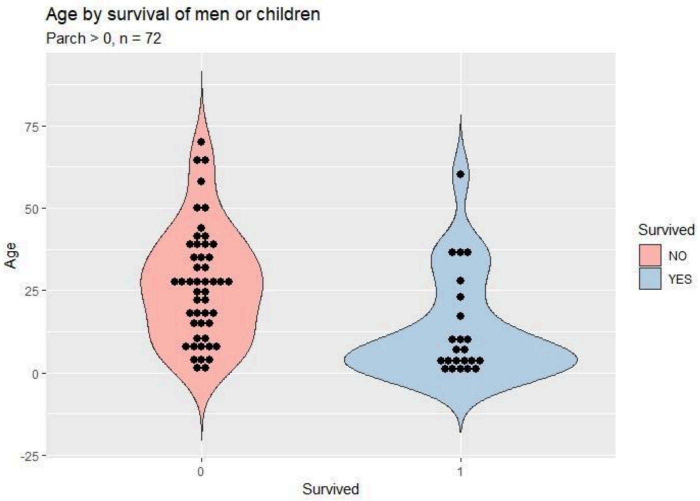
The figure does confirm what was suspected. As a result, passengers who paid less for the ticket, which are concentrated on the third class, have similar distributions of survivors and people who die. This could imply that for this group the variable “Survived” was a lucky question, but it did not was for the first-class cohort. Thus, the upper class could have been the priority when people contributed to solve passengers.

Has traveling with children had an impact on survival rates of men by age?

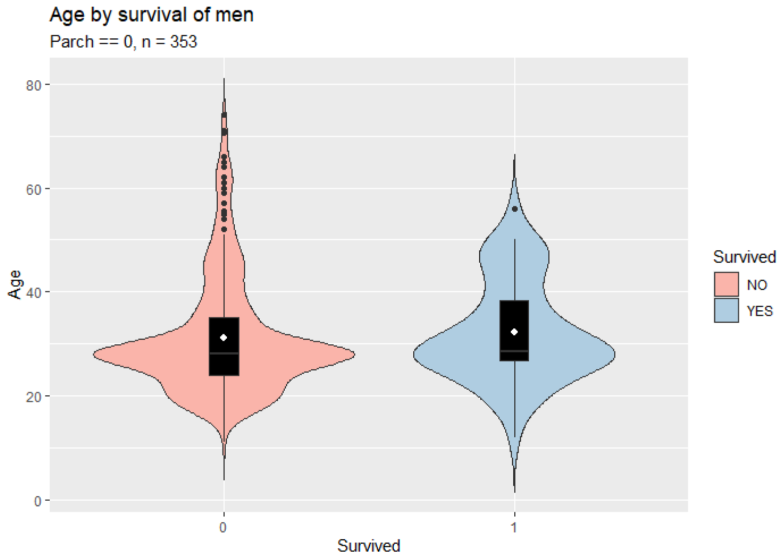
It has been shown beyond doubt in the first question that around the age of 17, male passengers could start to be considered as adults. The first quartile of the non survivors box plot of men under 22 (1st section remission) is 16.75. In this paragraph, the aim is to explore a potential explanation related to the age variable. Thus, it is likely that children survived the accident because of the help of their families. Or even that people in general tried to save the families of the catastrophe for humanitarian motivation. Therefore, the goal of answering this question is to deepen the first one and to assess whether the outliers of male passengers who are older than 17 did survive thanks to traveling accompanied by their children.

Firstly, the data set has been subsetted resulting in 72 observations only of those male passengers whose “Parch” variable scored higher than 0. In both figures, the white dot represents the mean (13.26 years for survivors and 26.91 not survivors) , while the line is the median (6 years for survivors and 18 for those who die); being the black box the interquartile ranges. The Age mean is higher among those who died than among the passengers who survived. The interquartile range is different in this data set for passengers who died and those who did not. First quartile is 3 years and the third one is 20 for the survivors, being most of them still considered as children. Nonetheless, the first quartile among the deceased is 14 and the third one is 38. The conclusion is that classifying by “Parch variable” remarks a difference in the age of the two levels of Survived: the 0 scored are most of them in their 20-30 years, while the majority of 1 scored passenger are children (according to the classification of the first section).





What the probability density of the graph is telling is actually the same as it was suspected. The wider parts of the graphs are located between different ages in Y axis, meaning that the probability of an observation of being one or another age is related to the group it belongs to, i.e., the most part of observations of survivors are in the lower part.



To contrast the gap between the passengers who travelled with children or parents and passengers who did not, the statistics of another subsetted group has been analyzed (remission to script). The mean age of male survivors who were aboard the Titanic with Parch == 0 were 32.17, and the dead passengers 31.21. The difference is smaller, and the survivors' mean overcome the other group one. The median and the interquartile ranges are quite similar. Furthermore, the wider part of the two groups in the density of violin plots is similar, i.e., the probability of being of one age in the two levels of "Survived" is almost the same.

Finally, answering this question leads you to conclude that the effect of "Parch" among age is remarkable, since the interaction of this variable with age one definitely influences the survival rate. Most of the male parents died, while their guys survived.

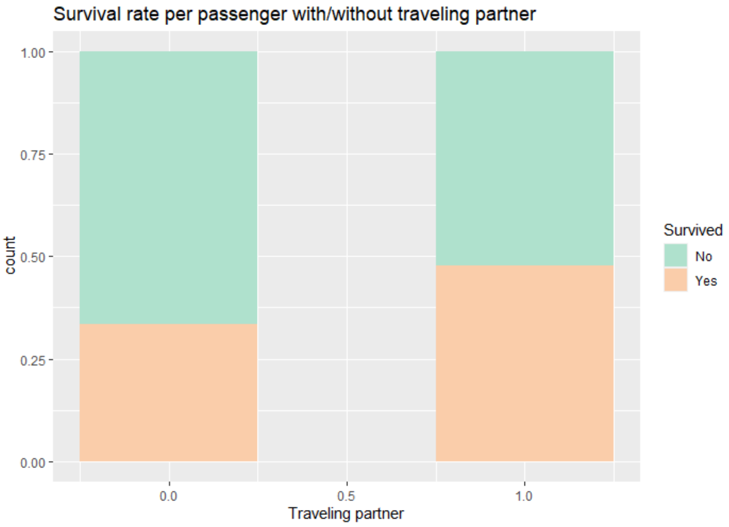
How does traveling with spouses or siblings affect the survival rate by sex?

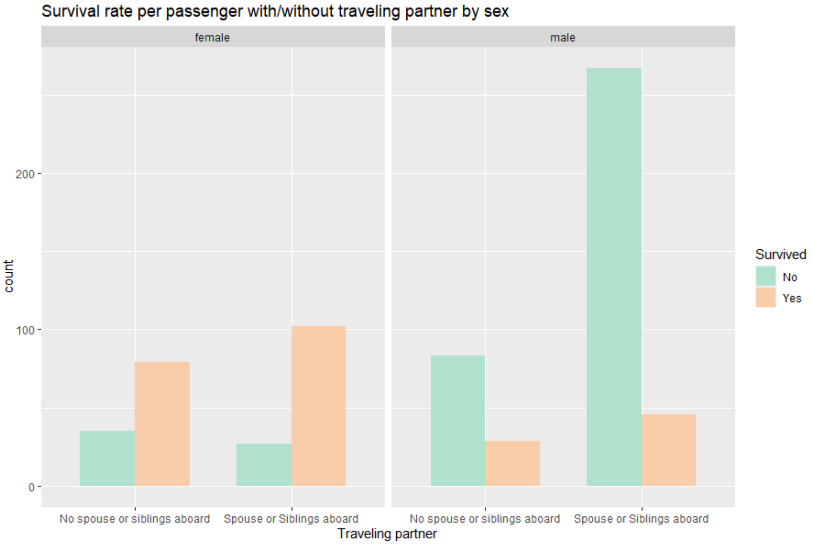
In an attempt to understand the relationship of having a familiar aboard and the survival rate, it is interesting to create a new dichotomous variable called "Traveling partner" with only two values: 0 for those who did embark "alone"; i.e, without spouses or siblings, and 1 for those who were accompanied by their spouses or their siblings. It could be seen the labels of the dummy on the script.

Here is a graph bar representing the total number of persons who survived or not, grouped by "Traveling partner". The difference between the survivors on both levels does not seem relevant. The count of passengers who were accompanied by their siblings or spouses is 442, and the count who were not is 226 (remission to the script). As more men did embark on the Titanic, it is possible to separate the effect of traveling with partner by sex, assessing if this categorical variable behaves on a different way. The behaviour of the levels is different for men and for women as it is displayed here. While for women those who scored 1 had a higher survival rate, for men there is a huge difference between those who were accompanied and those who did not. Apparently being a man with a spouse or sibling aboard increased the possibility of dying on Titanic, but on the other level in male passengers' graphs bar more people in comparison did not survive neither.

.

In conclusion, being a man aboard the Titanic boosts the survival rate but being a man with a spouse or sibling there increases it more. The proportion of men who scored 1 in "Traveling partner" is 0.708 (remission to table on script), i.e., most of them were aboard with familiar members. Obviously, one could conclude that the impact of traveling on titanic with your parent or your children /previous section) is different, making the survival rate increase below the 17 years of male passengers than the impact of traveling with spouses or children, with seems to increase the possibility of die aboard. Age and Sex are hidden behind the results again.

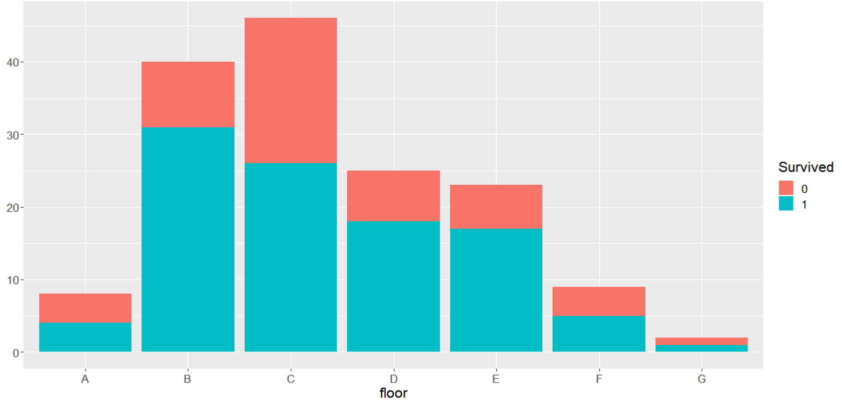

Survival rate per passenger with/without traveling partner

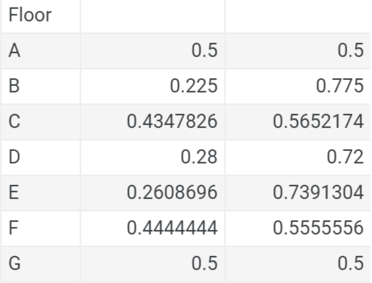

Survival rate per passenger with/without traveling partner by sex

How did the cabin floor affect survival rates?

The strike happened at night; thus many passengers were in their cabins. This raises the question of whether the cabin position in the ship affected survival. A preliminary graph of cabins was hard to interpret and didn't provide any clear trends. Thus, the cabins were graphed by floor. When reviewing the deck plans for the titanic, (https://www.encyclopedia-titanica.org/titanic-deckplans/), you can see that the letter at the start of the cabin number refers to the floor that the cabin is on, with A being the highest cabin and G being the lowest. Thus, maybe the cabin floor would correspond to survival.

There are no overall trends as the floor gets lower, but it is worth exploring a bit further in case. Overall, the survival rates are higher than for the average population; this is likely attributable to there being cabin numbers primarily for the first-class cabins/floors, with very few third class passengers having their cabins listed in this data source. Of those floors with higher numbers of data points (B, C, D, E) C has a significantly lower survival rate; the other three are all 70%+ surviving, and C is a mere 57%. This may be attributable to a few things; first, deck C on the Titanic had a large number of cabins, more than any of the other upper decks. This may have meant there were fewer stewards per passenger to help. It could also have meant there were more crowds/there was greater difficulty getting upstairs from this floor. It could also be attributable to the varied locations within the ship; a passenger much further from the strike may have assumed the vibrations were insignificant, but the cabins at the front of C would have felt the strike more strongly and more quickly believe that they had to leave.





Floor		
A	0.5	0.5
B	0.225	0.775
C	0.4347826	0.5652174
D	0.28	0.72
E	0.2608696	0.7391304
F	0.4444444	0.5555556
G	0.5	0.5

Overall, it was somewhat expected that the cabin would not be highly significant- with only 153 of the 668 passengers having a cabin listed, there would naturally be some floors with very few observations, making it harder to show trends.