The Sinking of the Titanic

Lana Botchorishvili, Christoph Sieve

18 10 2020

Disclaimer

The data for this project was taken https://knowledge.domo.com/Training/Self-Service_Training/Onboarding_Resources/Fun_Sample_Datasets. This website teaches its user how to perform advanced data analysis and write meaningful calculations. They in turn got their data from http://www.kaggle.com, a website where users upload their own datasets for competition.

For this reason the project at hand does not claim accuracy about the true numbers for the sinking of the Titanic. It is rather meant to showcase the ability to select suitable R tools for data analysis. Thus it might be possible that some conclusions don't align with reality.

About the data

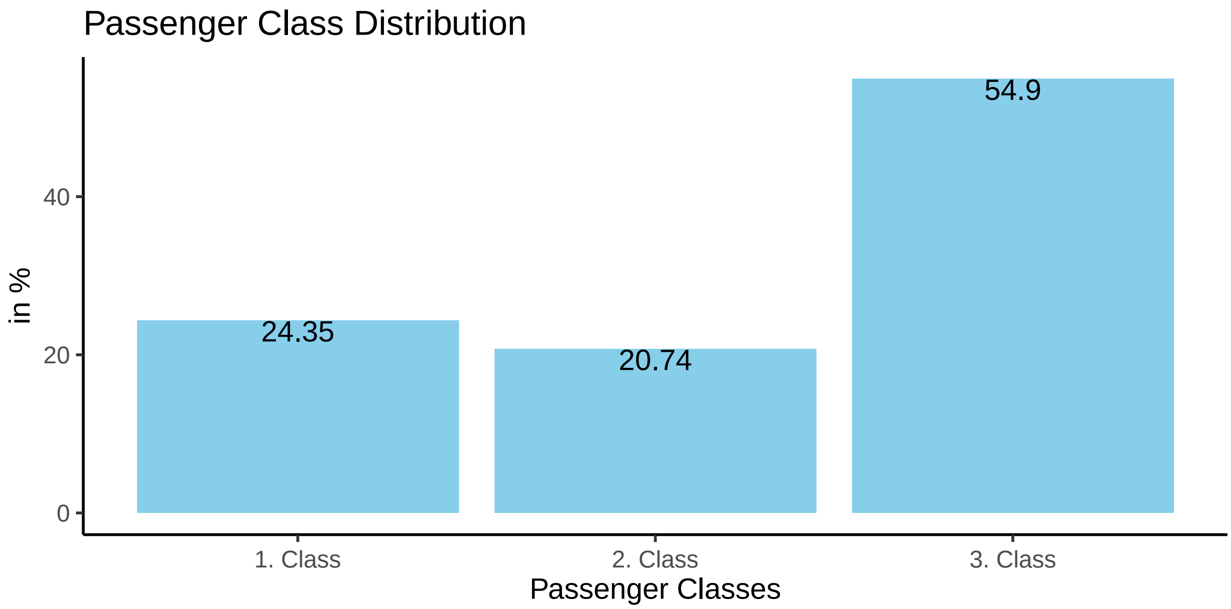
According to Wikipedia the Sinking of the Titanic is one of the deadliest peacetime maritime disasters in history killing more than 1,500 people. The CSV dataset at hand consists of basic information for 887 passengers aboard the HMS Titanic when it sank in 1912, including name, age, gender, passenger class, fare amount, number of family members aboard, and whether they survived the disaster. The columns in this dataset are as follows:

Column	Descritpion
Survived	Shows using Boolean values whether or not this passenger survived (0 is no and 1 is yes)
Pclass	The passenger class for this person, either 1, 2, or 3
Name	The name of this passenger
Sex	The gender of this passenger
Age	The age of this passenger
Siblings/Spouses Aboard	The number of siblings and/or spouses accompanying this passenger
Parents/Children Aboard	The number of parents and/or children accompanying this passenger
Fare	The fare paid by this passenger to board, in British pounds (£)

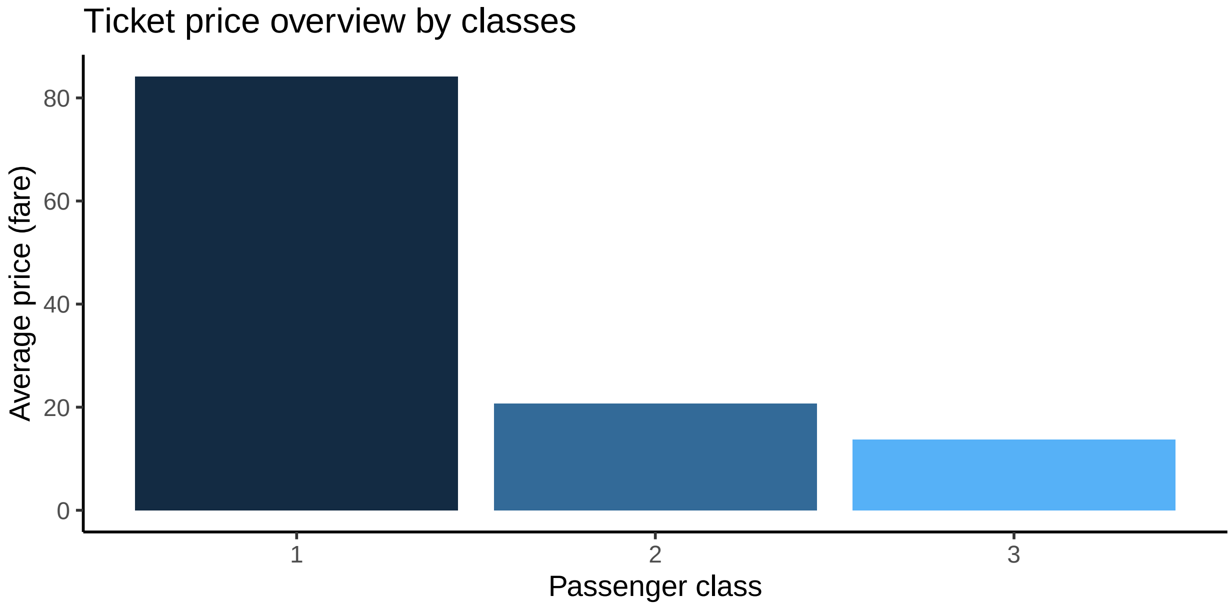
Analysis of the data

A simple overview

While the content of the columns *Survived*, *Sex* and *Name* are easy to imagine for all 887 passengers, the same cannot be said about the remaining 5 columns. In order to get familiar with the data set we will have a look at a simple distribution for the number of tickets sold in each class.



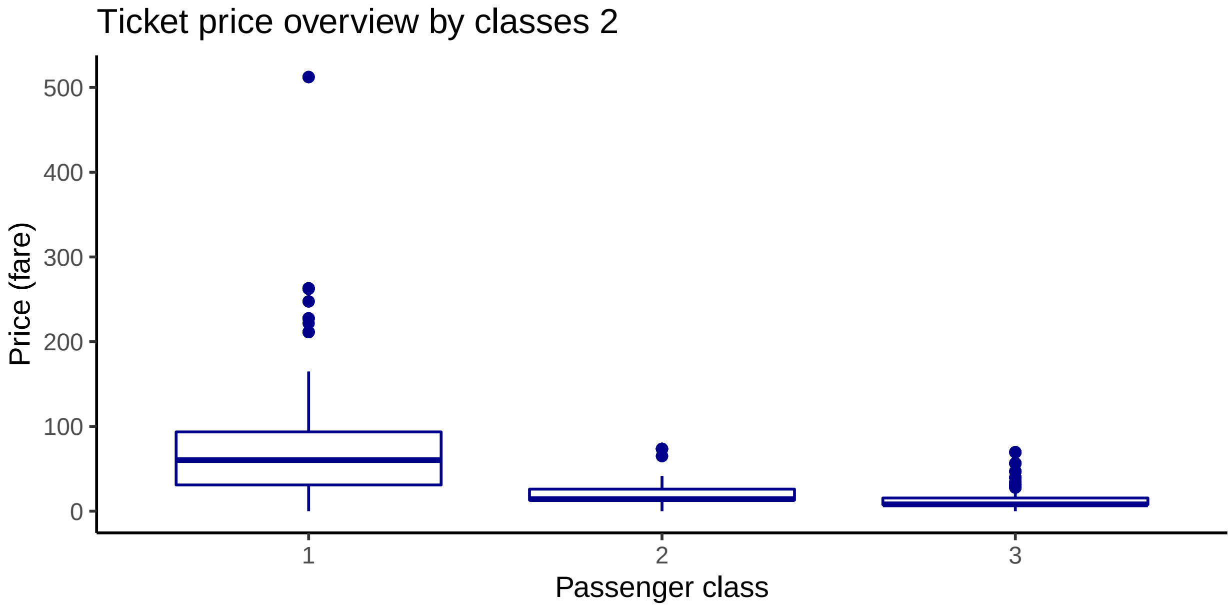
The following graph helps to get an idea about the average price in each passenger class.



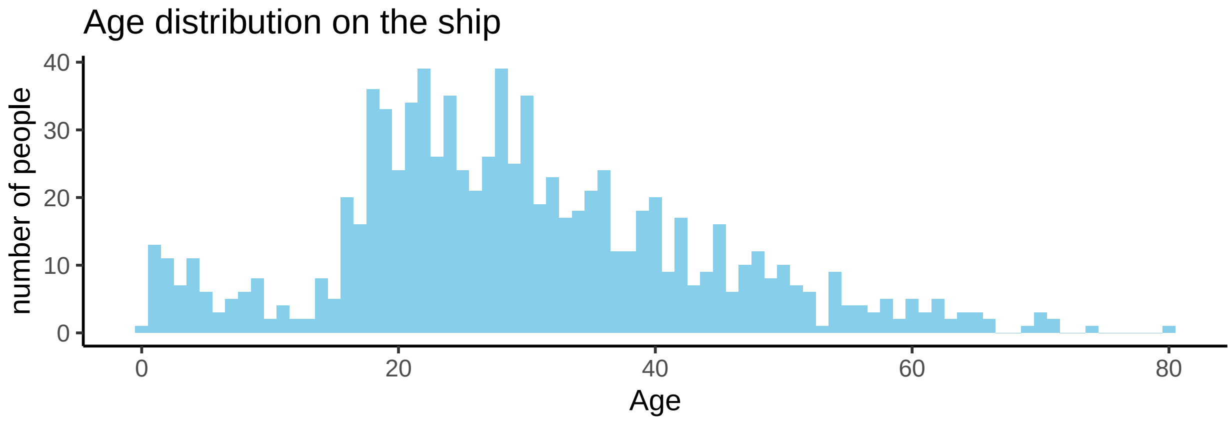
As one would expect, the first class is the most expensive. The second and third class follow with a surprisingly huge margin to the first class.

Although this helps to develop a rough understanding of the price, it would be wise to go a little bit more into detail as averages tend to be near meaningless without some context. For that matter the following table adds to the plot from above more detail. It becomes clear that that huge jump from 2. class to 1.

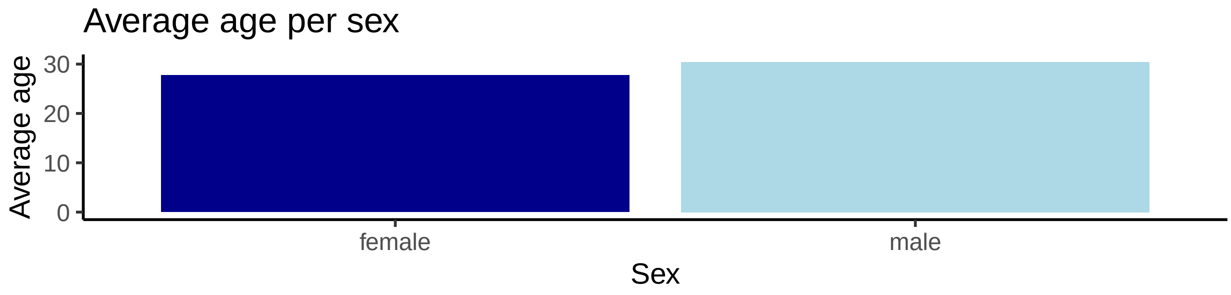
class above is mainly due to the relatively high standard deviation for the 1. class fare.



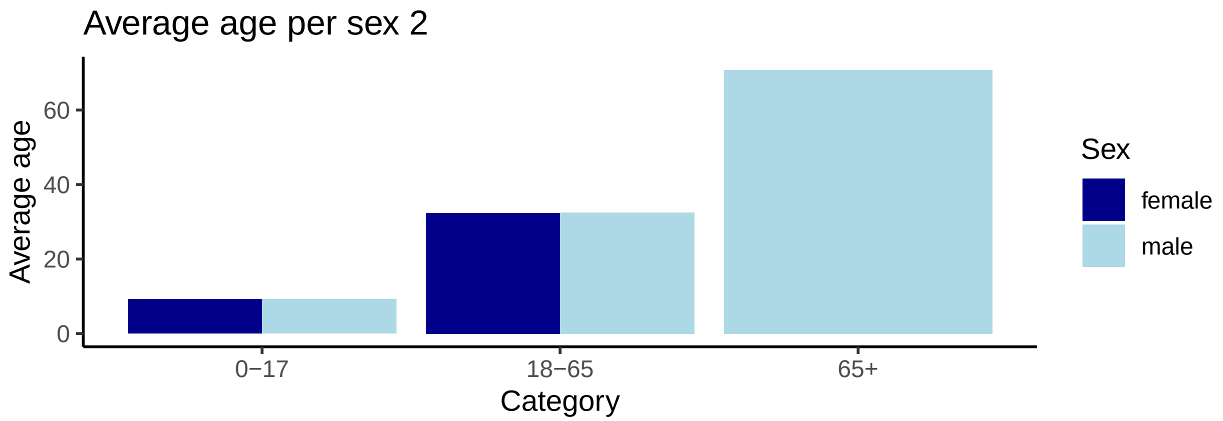
We are going to investigate the age structure on the Titanic with the help of the next two plots. The ship has children of any age, even as low as multiple months as well as adults that are well over 60 years old, as illustrated in the histogram below.



If we were interested in the average age of a person on the ship and just took the average, we would get the following plot.



As discovered in the histogram though, we have a huge chunk of people between 20 and 40 as well as a moderate amount of people below 10 and some above 60. In order to get a better overview of the average age on the ship, it might be advisable to put them into age groups. If we set the interval for "being an adult" to 18-65 years and "being a pensioner" to 65 and older, we should get a much more refined picture.



It turns out that the average age of an adult on the ship is a little over 30 years and, even more surprisingly, that the pensioner group consists exclusively out of men. Furthermore, in both groups, 0-17 and 18-65, men and women are represented. We are left wondering, however, how many of each sex are in the age group 18-65.

One thing to consider is that the Titanic incident happened at in beginning of the 20th century, meaning we can expect some unbalanced distribution with regards to sexes on the ship. We can make this fact visible in the data by just looking at the names. In that time period naming conventions were strict. All females in England were either "Miss" if not married or "Mrs" if married. For males the etiquette dictated that men be addressed as Mister, and boys as Master. Let's have a closer look.

Out of 314 females and 573 males in total, we found that by filtering names for either "Miss" or "Master" would result in:

Sex	Average age	Maximum Age	Number of people
female	21.99	63	182
male	4.64	12	40

We can see that "Miss" is not restricted to age while "Master" is. Taking the age 12 as a lower bar, we can see that about the same number of boys and girls are on the ship.

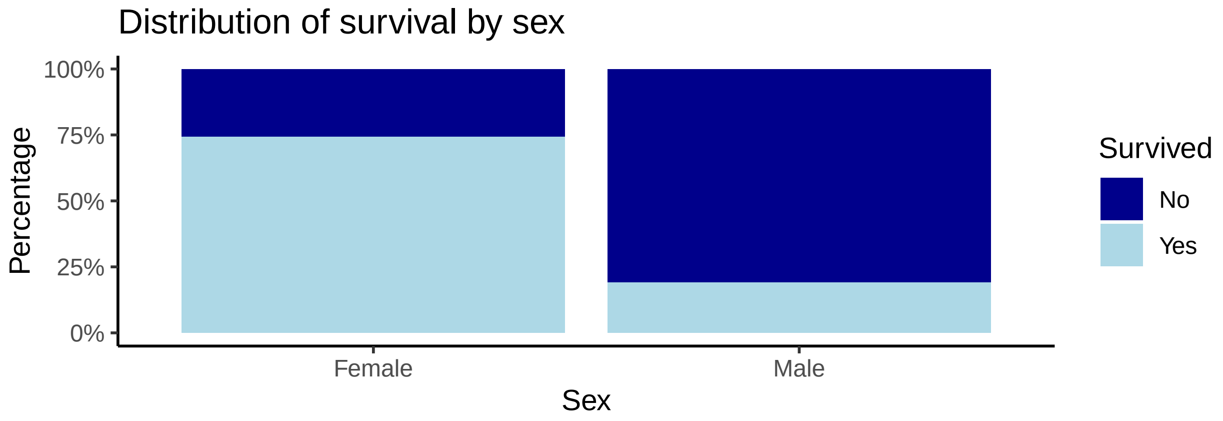
Sex	Average age	Maximum Age	Number of people
female	5.09	12	38

If we now look at all females having the title of "Miss" or "Mrs" in their names and are over 12 with those men that have the title "Mrs" we can see that there are almost twice as many men than women on the Titanic. The average age here seems to confirm our results from above, but differs slightly since we filtered for all people of ages 13+. Another thing to note is, that we have in total 781 people, while the full number of people aboard is 887. There are about 100 people with titles like Duke, Lord, Capt. or Dr. that were not considered here.

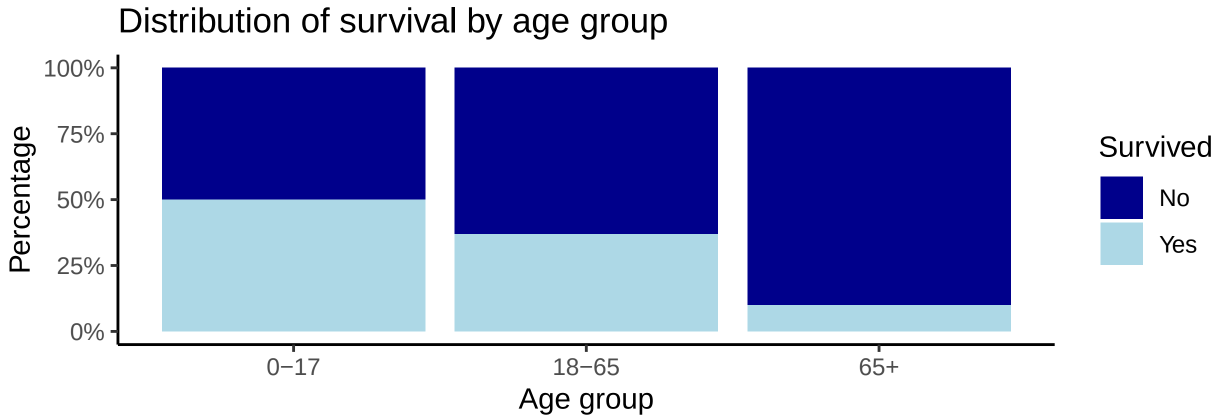
Sex	Average age	Maximum Age	Median	SD	Number of people
female	30.78	63	29	11.69	269
male	31.88	80	29	12.45	512

Factors contributing to survival

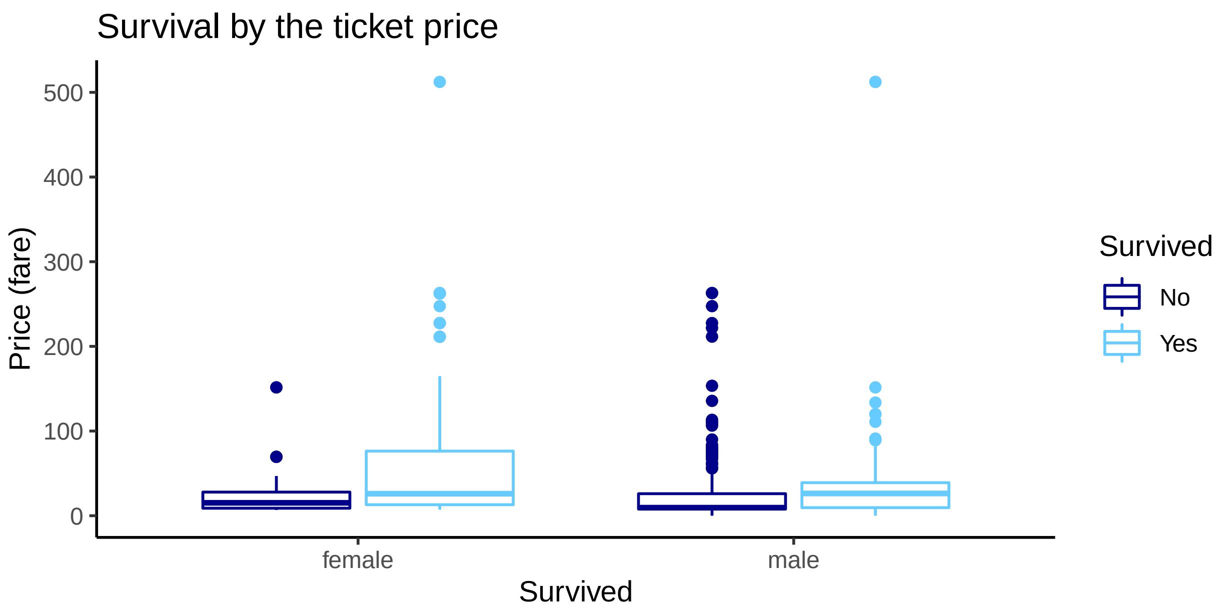
Surviving a ship crash may be influenced by a myriad of different factors, although, one, in particular, is most noticeable when making predictions. Gender plays a big role in what is socially acceptable to do in situations like these, and it usually dictates to cater to women and children first, which might skew the chances of survival quite a bit. The following graph shows the distribution of survivors between males and females. As expected, most females survived, while most males did not. As expected, being a woman is rather important in staying alive during a crash, since females have approximately 75% chance of survival, but men only have 25%.



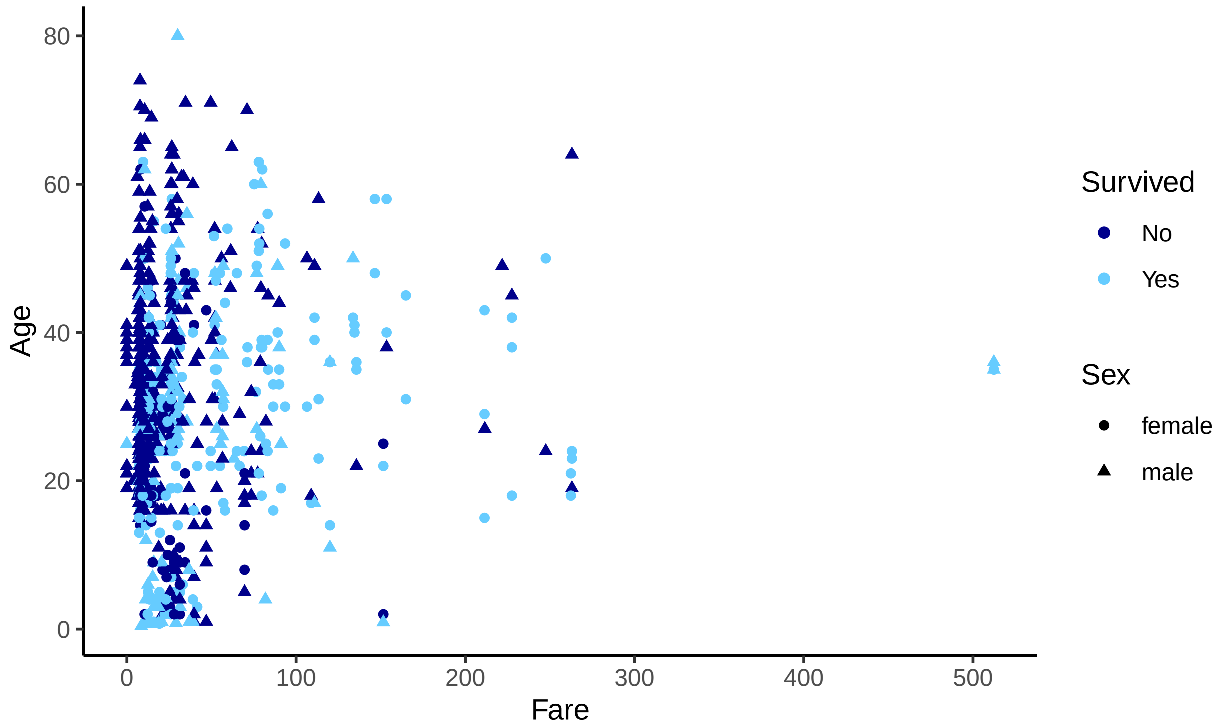
As expected, age also contributes to the likelihood of survival: as the age increases the chance of survival goes down, creating a negative relationship between the variables and making young people (females, usually) more likely to survive. Additionally, as observed before, all the people over 65 are men, which also contributes to the low survival rate in the older generation.



Now we will demonstrate whether ticket price affects the chances of survival. As seen from the boxplot, average ticket fare and quantile values of those who survived is slightly higher in both men and women, although there are more outliers in the men who didn't survive and women that did, leading us to believe that ticket fare plays a lesser role in the survival of women than men. Overall, those more financially capable have a bigger chance of surviving a ship crash, but only marginally.



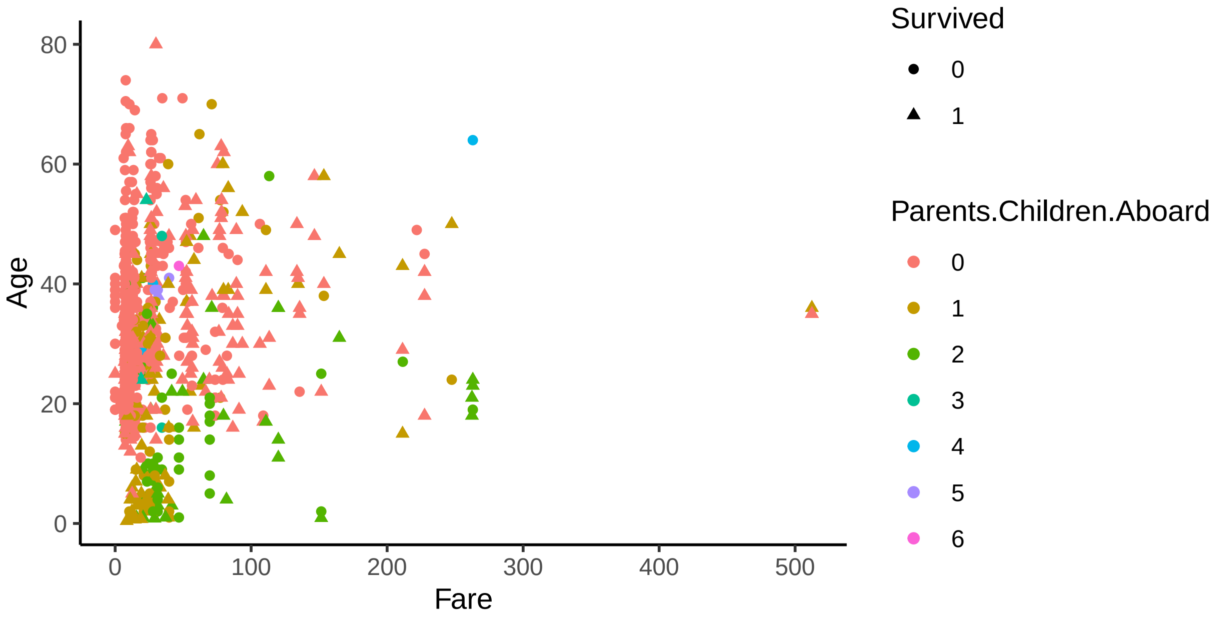
Relationship between the fare and age can also be observed in the following graph. The figure shows the clustering of people who didn't survive is concentrated at around 20 to 30 pounds in middle ages, which is also evident from the following table showing the mean fares for different age groups, with average ticket price being approximately 24. It is important to note the presence of male gender in this cluster, as well as lack thereof in mid-ranged fare prices and survived group.



It is interesting to note that there is not much disparity between the mean fare prices of people who did not survive, but there is a rather big gap between the age category 0-17 and 65+, and 18-65, the latter being almost 10 pounds in the lead. Perhaps this could be attributed to the 18-65 age group being working class.

Category	Did not survive	Survived
0-17	28.47	33.65
18-65	21.31	51.94
65+	23.8	30

Looking at the number of relatives on the ship, it is not clear whether this has any relation with the survival. Siblings/spouses aboard and parents/children aboard have a very similar distribution. Those who have 1 or less relatives are unexpectedly over-represented in the non-survivors: approximately half as many people with 0 relatives survived the wreck, as those who passed. However, interestingly, we see an increase in the survivors in the 1 relative class, which could be attributed to not having to search the entire ship for one's family, possible proximity of the relative, since couples are known to stick together, which saves time. There is also a likely trend of people with 4 or more relatives being exclusively in the non-survivors group.



Conclusion

Overall, the demographics of Titanic passengers were interesting to look into. it turns out that there are over twice as many men on the ship as women. Over half of all the people are third class passengers, with average ticket price being below 20, while average fare for quarter of the population is over 80 pounds. Looking at the age demographics, it is evident that average age of a female was a bit lower than that of a male. This difference can also be magnified by the fact that the group over 65 is exclusively populated by men. Although the amount of young boys and girls is nearly identical.

We take note of some trends: there is a disproportional representation of females in the survivor group, which, while expected, is surprising in its bias, as approximately 75% of women and only 25% of men survived. Both genders exhibit a higher average fare price for those survived, however, women survivors have more high priced outliers, while, in males, this attribute is evident in the non-survivors. In age groups, the highest survival rate can be attributed to the people aged 0-17, with 18-65 age coming next and 65+ group being the last. The lack of relatives during the wreck does not seem to offer too big of an advantage, although there are almost no people with more than 6 relatives that survived.