

# EXPLORATORY DATA ANALYSIS



**An analytic approach to the Titanic tragedy**

Introduction to data science

Lucía Cordero Sánchez  
Alonso Madroñal de Mesa



# EXPLORATORY DATA ANALYSIS

Universidad Carlos III de Madrid   
Introduction to data science 



## INTRODUCTION: APPROACH TO THE DATA

Titanic sinking was a catastrophe which took place between April 14th and 15th, 1912. The casualties in Titanic were largely not arbitrary as we will show in the next pages, and that is the main reason why, based on our interests, we decided to focus our questions on inequality.

First of all, the boat was structured depending on the social position and it reflected the British society at that time, where money was not the only determinant fact, but also status.

From now on, we will see how this fact affected life expectancy in the tragedy.



## HOW WE APPLIED EXPLORATORY DATA ANALYSIS

To fulfill this task, we loaded libraries such as **ggplot2**, **ggthemes**, **gganimate** for animations and some libraries for design from GitHub developers.

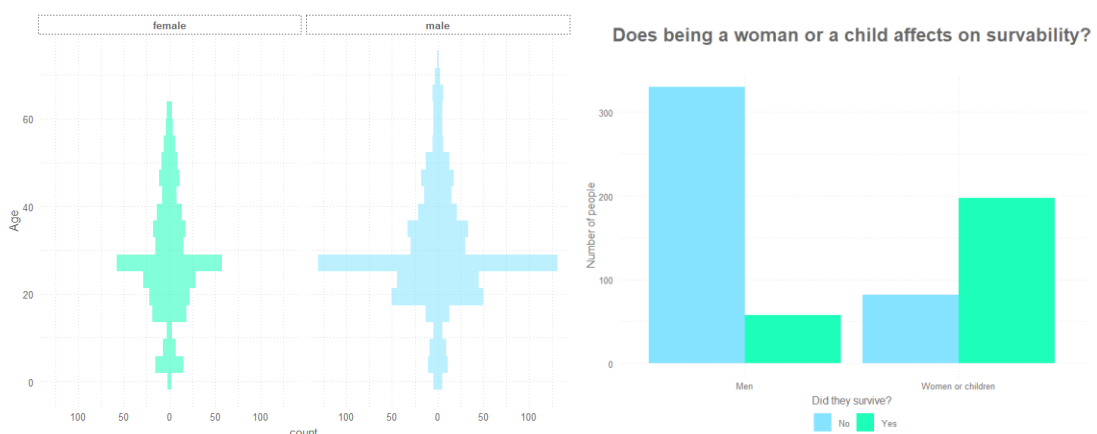
In order to get to know R Studio itself and to complete the task successfully, we made research across **CRAN Repository**, **R Graph Gallery**, **Datanovia**, **Sthda** or **Tidyverse**.

We have generated six questions and two extra questions as extensions from the main ones; later on, we visualised it and finally, new conclusions were drawn.



## Q1: DOES BEING A WOMAN AND A CHILD AFFECTS ON SURVABILITY?

It is well-known the evacuation protocol concerning women and children first; was it actually followed by the cabin staff during the crisis, though?



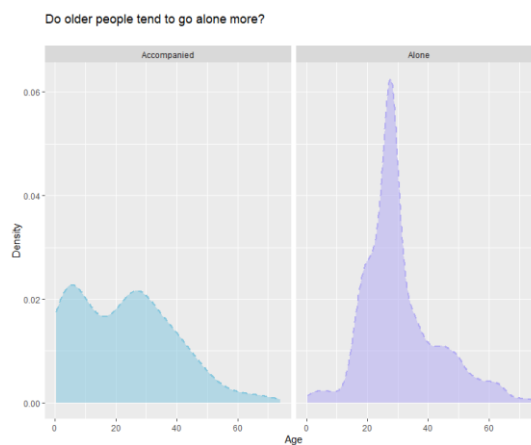
The illustration from the left side is a population pyramid (it reflects the same result on both side, like in a “mirror”) based on the passengers abroad; from it, we observe that there were a bigger amount of men compared to women in the transatlantic.

We confirm that being a woman or a child was a determinant factor in order to survive: there were more men in proportion, but clearly the proportion of survivors is mostly compound by the last ones.



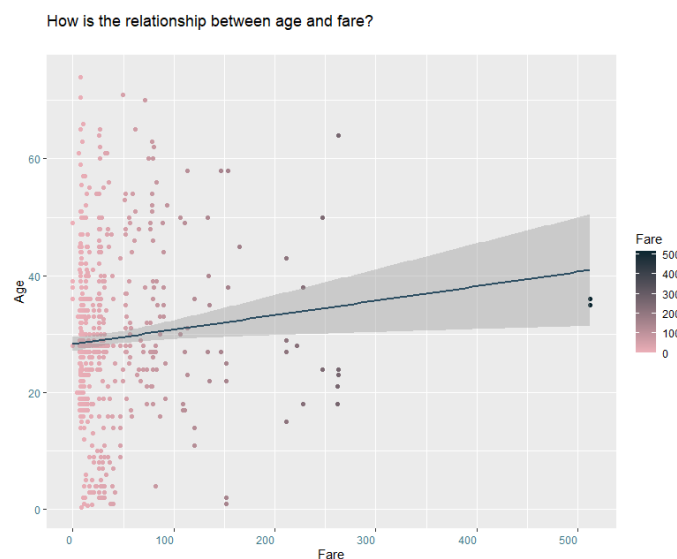
## Q2: DID PEOPLE OFTEN TRAVEL ALONE?

As we expected from the very beginning and later we contrasted graphically, children and teenagers tend to go accompanied (probably by other family members) and the peak of people going alone was reached when they were in their 30s. Moreover, there is another maximum in the people going accompanied in that interval, which may be thanks to the passengers who embarked with their partners.



## Q3: AGE DISPERSION COMPARED TO FARE?

The expected answer we thought we would get was that in the underage interval of people, their fare would be very similar to the adults one since most of their fares were paid by their parents, and from there, the expectation was a linear relationship; the older, the cheapest fare they paid. This is the scatterplot we obtained:



The navy blue line is the expected relationship for the data if it was a linear relation (which is not, clearly).

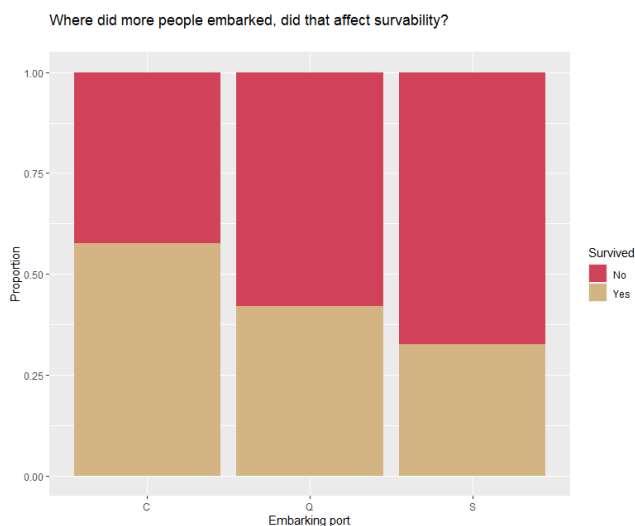
We cannot get to clear facts given that apparently the relationship is not linear with this data (but neither weak), so the only conclusion we can assume is that most of the passengers paid between zero and fifty pounds from the time.



## Q4: WHERE DID PEOPLE EMBARK AFFECTED SURVIVABILITY?

Since the harbors in which passengers embarked were in different countries, we supposed their profiles would be different too depending on it. Finally, making research we discovered that there were considerably more people who left from Southampton than in the other two (in fact, most of the foreigners embarked there).

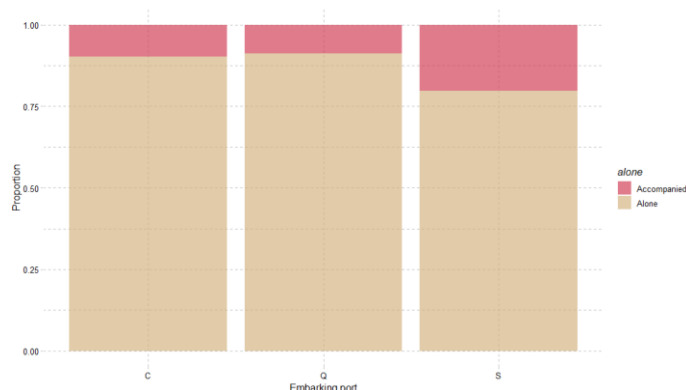
Our theory is that due to the strict immigration laws that did not allow immigrants and Americans to be in the same place, many foreigners were in third class and all together behind bars not because of their fare but because of their nationality. This is the visualisation of these data:



As most of them embarked in Southampton, it would be reasonable to think that given that they were in third class, most people who died in proportion (not considering the amount of people from each harbor but their survival rate) were from there.



## Q5: IN WHICH HARBOR FAMILIES EMBARKED MORE?

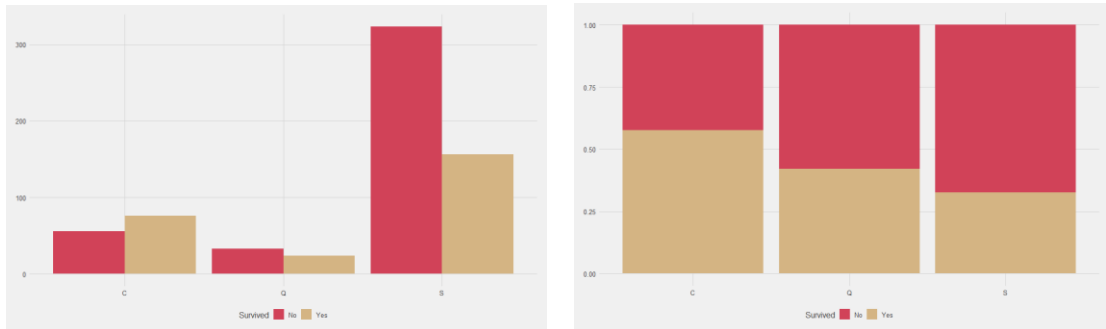


From this barchart we draw some conclusions:

- ✚ People who travelled both alone and accompanied does barely differ in Queenstown (Ireland) and Cherbourg (France), passengers were predominantly alone in both cases.
- ✚ There is a slight difference in Southampton (England), where more families and couples embarked.

At this point it is almost ineluctable to ask ourselves in which port the survival rate was bigger; if in Southampton, where more families embarked, the death rate differed a lot from the other two.

### EXTRA QUESTION 1: IN WHICH PORT THEY SURVIVED THE MOST?



The right image shows in absolute terms the survivals in terms of the port, while the one in the left shows the proportion with the three bars having the same area.

Thanks to these illustrations, in them we observe that the survival rate was considerably more important in Southampton, where most people embarked (taking this aspect into account, it is logical that there were more passengers surviving).

In spite of this difference of passengers, drawing on the second image we know that in proportion, less people survived in Southampton and there are slightly more survivals in Cherbourg rather than in Queenstown (even though we knew they were accompanied in the same proportion).

From this, we extract a conclusion: probably more people who travelled alone died if they were travelling from Southampton (compared to the other ports) since the families were more likely to survive together and in this port the tendency to embark with company was greater.

Finally, we subtly get to the same conclusion from question 4 concerning immigrants: if the mortality rate is greater in Southampton, statistically more immigrants passed away.

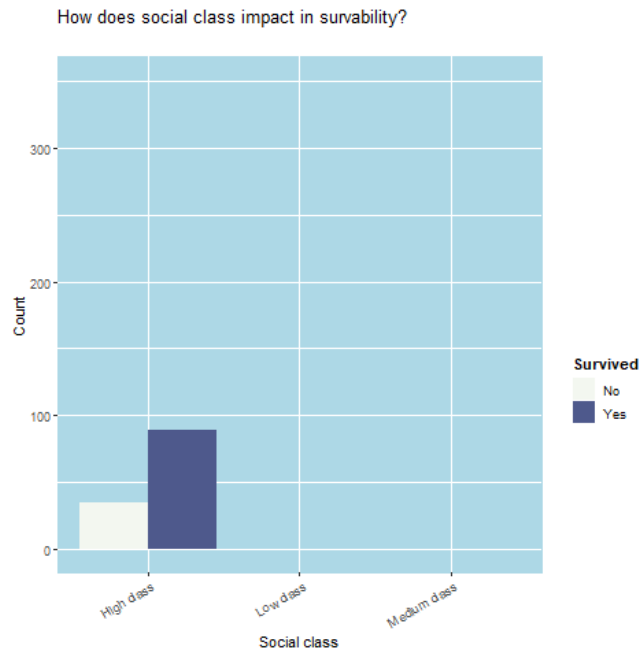


### Q6: HOW DID POSSIBILITIES OF SURVIVING CHANGE BEING THE RICHEST AND THE POOREST?

To answer this question, first we cleaned our data and create a new data frame with the data we needed: we classified the original data into different social classes (high, medium or low) depending on the cost of the tickets and the percentile in which each passenger was. After this, we defined what being in the top richest and poorest was (percentile 10 and percentile 90 in the fare and belonging to the first or the third class, respectively).

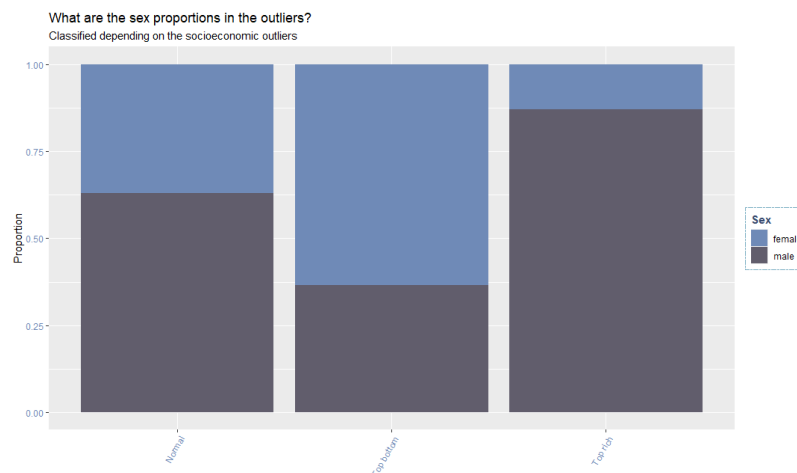
Once we had all our desired data classified, we created a new data frame with these new columns to start building the visualization.

[Author note](#): keep in mind that this animation will not be available in PDF format, but in HTML. Both of them are annexed in Aula Global.



As expected, the social class was quite determinant in order to choose who should live and stay in the boat; less than half of the high class passengers died, while approximately only one out of each four low class passengers survived. In the middle class, both values are quite adjusted but there were a little bit more survivals than people passing away.

## EXTRA QUESTION 2: IS SURVABILITY INFLUENCED BY SEX?



Wondering if the results were influenced by sex or not in the outliers, we discover that they clearly are.

When we pertain to the “top bottom” or “top rich” we refer to the percentiles 10 and 90 and the outliers (very extreme values, the poorest of the poor and the richest of the rich).

In this case, it is observed that women have a leading role at the top bottom, nevertheless when the graph reaches the top rich, men take greater importance and the percentage of females being in the rich outlier is minimum.