First Assignment
Exploratory Data Analysis
Titanic

Jaime Salafranca Pardo & Diego Stergar Vega

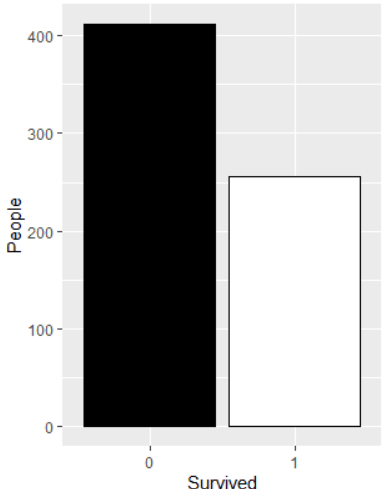
Wednesday, 20th of October 2021

Summary

1.First approach for each variable

1.1 Variable Survived

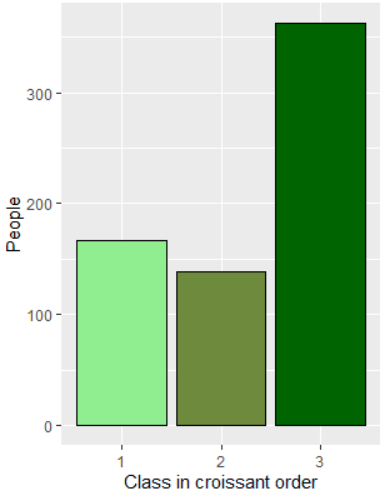
```
survivors      Freq
     Died 0.6167665
 Survived 0.3832335
```



We see that more than 61% of the analyzed passengers died, with the prop table but we created a bar chart to represent it better. The black bar represent people who died, which are significantly more than the ones who survived, which represent the white bar.

1.2. Variable class

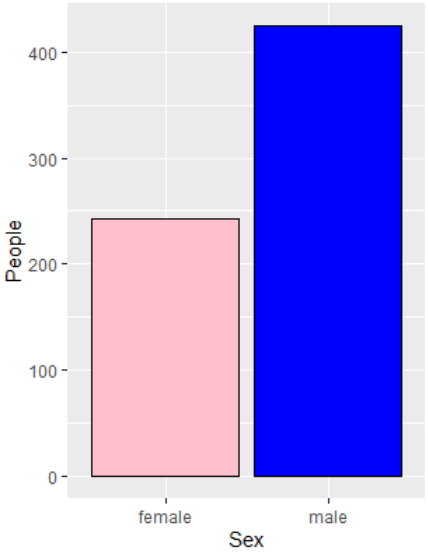
```
          1         2         3
0.2500000 0.2065868 0.5434132
```



In this proportion table, we can see that the third class had more than half of the passengers. The graph that shows the number of people by classes seems to correspond to the data. We can see the classes represented in croissant order. It is interesting that the second class has less people than the first. That means that people at that time could be very rich or, on the other hand, very poor. There were barely not people who belonged to the "middle class".

1.3. Variable Sex

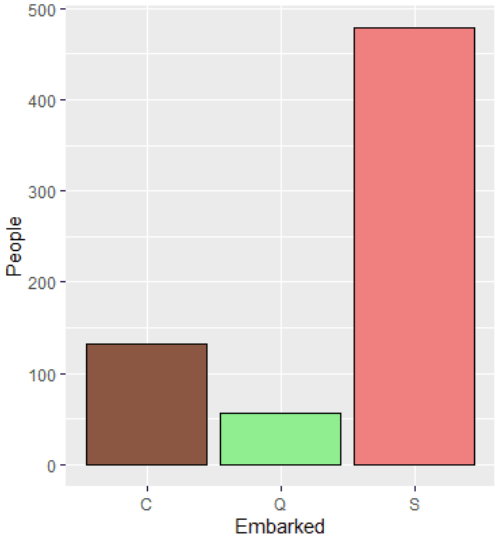
```
female 0.3637725
  male 0.6362275
```



In this table, we can see that there were more men than women in the titanic.
It is important to take that into account for the questions. We have represented that in a bar chart where the blue represents the proportion of men on the pink for women. This will be extremely important when we consider the relationships between sex and possibilities of surviving.

1.4. Variable Embarked

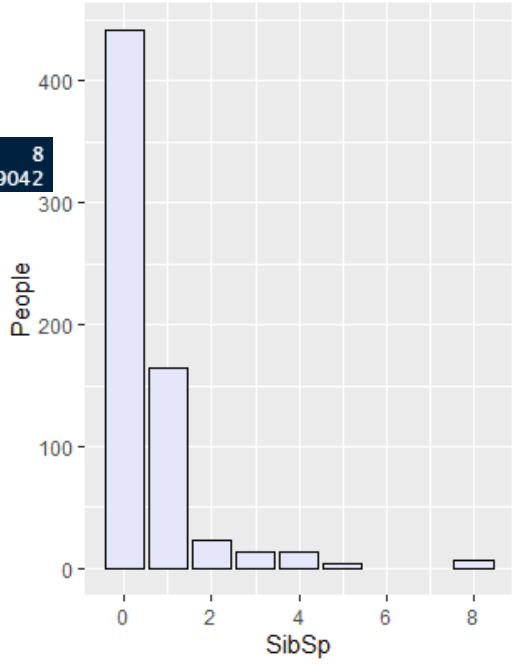
```
C 0.19760479
Q 0.08532934
S 0.71706587
```



This variable shows the port of embarkation for each passenger; there were three options: Southampton(S in Red), Queenstown(Q in green), and Cherbourg (C in brown). These same colors will remain representative of each port for the long of the work. This variable could be useful to see different economic classes in the different cities.

1.5. Variable Siblings/Spouse

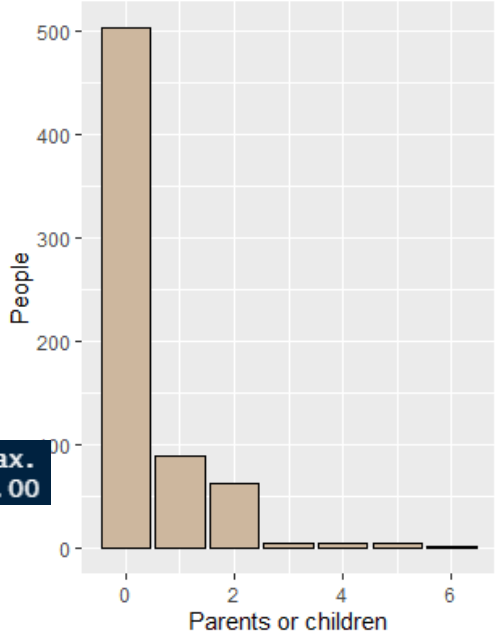
0	1	2	3	4	5	8
0.661676647	0.245508982	0.034431138	0.020958084	0.020958084	0.005988024	0.010479042



This is the proportion of the people for each number of siblings and spouses combined. We have also created a bar chart to visualize it better. In this case, we don't use a histogram even if the values are numerical because we interpret them as discrete values. We think that this point of view is more interesting for the analysis. We can take a look at the fact most of the people had 0 sib/spouses and some of them had 1. People with 2 or more spouses/sib can be considered as outliers.

1.6. Variable Parents/Children

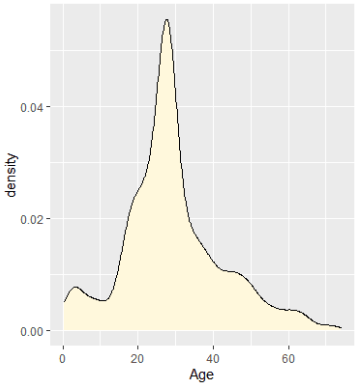
0	1	2	3	4	5	6
0.752994012	0.133233533	0.092814371	0.007485030	0.005988024	0.005988024	0.001497006



This table is also the percentage of people for the number of parents and children above combined. We have also done a graph to represent the variable. We can see that having more than 3 parents or children is not a very common thing. This will be useful when we are going to take a look only at children knowing how many parents they have on board.

1.7. Variable Age

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.42	22.00	28.00	29.15	35.00	74.00

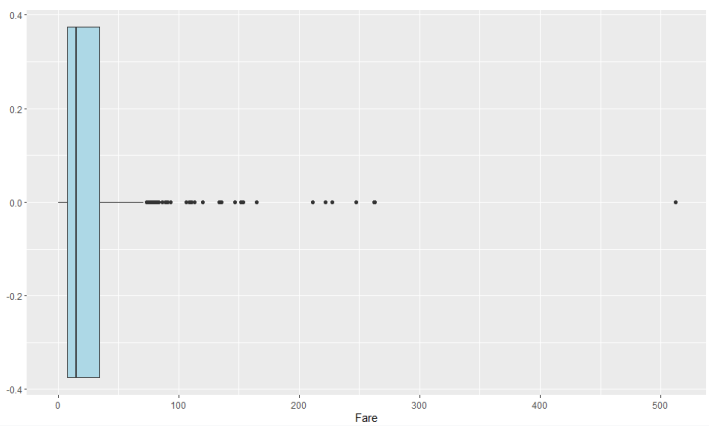


This variable shows us the age of each passenger from the age of 5 months to the age of 74 years. We see that the biggest part of the people is pretty young. We have done a density curve that shows us more in detail data. We see that there are a lot of people between 20 and 30 years old but also a lot of kids in comparison with older people. It makes sense because life expectancy at that time was quite less than now.

1.8. Variable fare

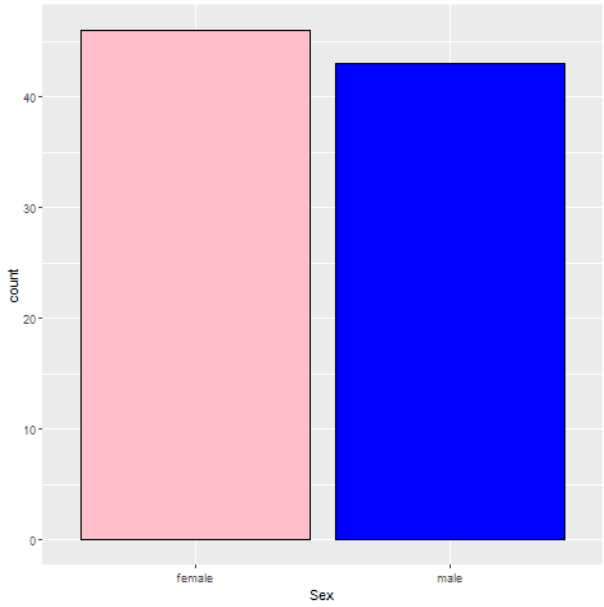
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	7.925	15.246	34.066	34.109	512.329

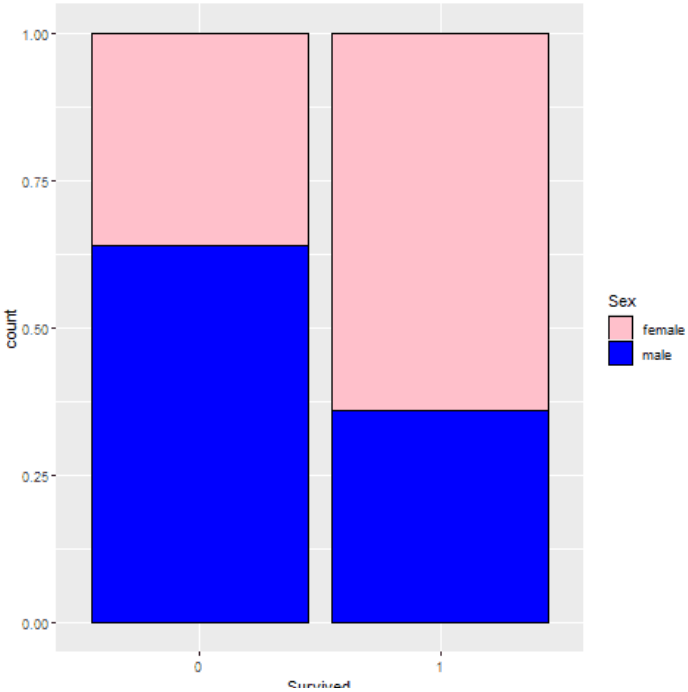
The variable Fare is the price the people paid for the trip, we think that is relevant to show that some people don't pay anything while others pay more than 500 thousand. The reason why we have used a boxplot is that some people have paid quantities that are very far from the mean. This means that there are a lot of outliers, which can be observed in the boxplot



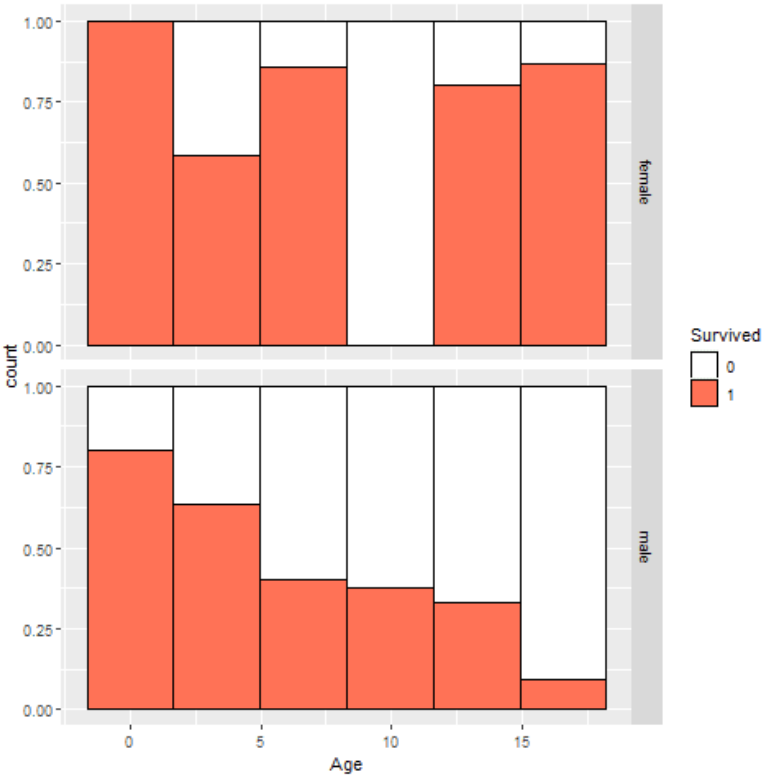
2.1. For minors was gender an important characteristic for survival?

One question we could ask is if for minors the gender was relevant for survival. But before verifying that condition we have to see the proportion of males and females for children. In this table(2.1.1.) male (blue) and female (pink) we can see that the proportion is very similar.
After that, we can look for the gender of the minors who survived.





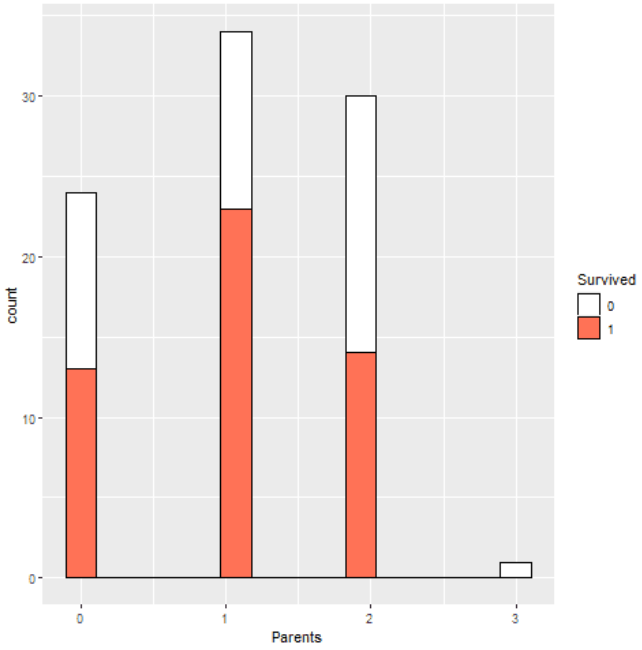
After seeing this graph(2.1.2.) we can see that girls are highly more likely to survive. But seeing this result the reason is not very clear. We would like to explain this situation: **Why are girls more likely to survive?** To find an answer, we divide the minors that survived by their age and by genre.

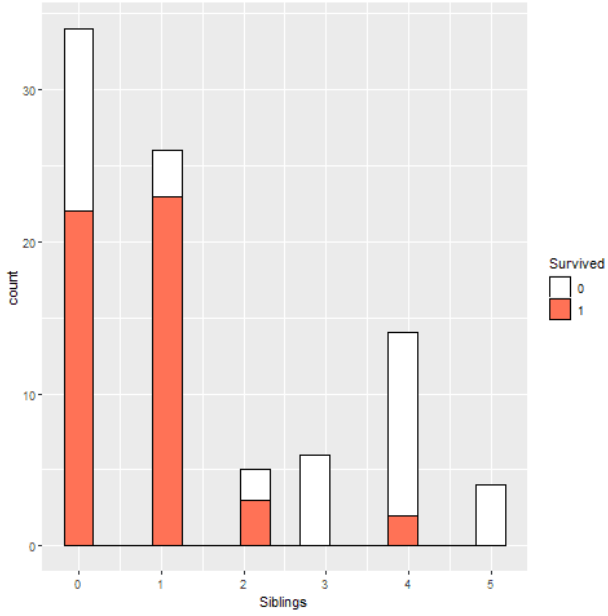


Here we find something very interesting. For girls in the top and boys in the low part of the graph (2.1.3.). Even if the chances of survival of the children are pretty similar when they are young, the more they grow the more the boys seem to have fewer chances to survive than the girls. A hypothesis is that the boys approaching 18 years are treated more like adults by the society of that time while they consider the girls more vulnerable. This explains the difference in survival between boys and girls. It is not the only fact of gender. Whenbeing young, sex fact doesn't matter but when they grow boys tend to survive less and that is the difference.

2.2. Were the children with parents more likely to survive?

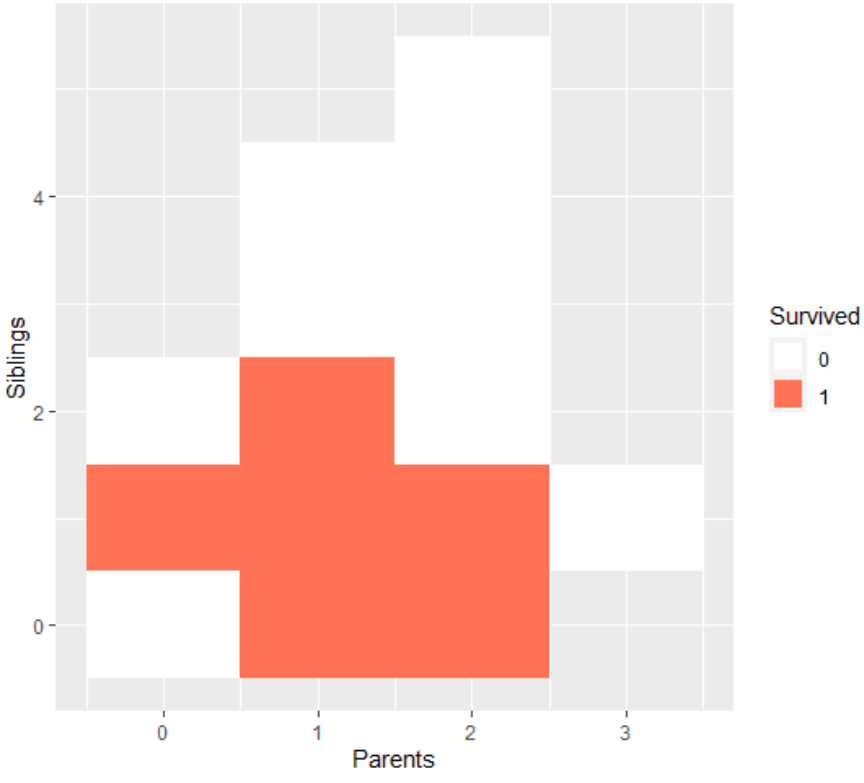
The variable parents & children are very interesting when we consider that for minors the value of children is 0 and it only represents the parents. Considering that we took a look at the chart describing the chances of survival knowing the number of parents. And we did that graph (2.2.1) where we plot the number of children for every number of parents, and we show the ones who survived in orange. We see that there is a child with three parents, it could be an error or a real value or even a child with a son, But concerning survival, we see that children with one parent are more likely to survive. A hypothesis is that when there is more than one parent there are also siblings. That's the reason why we look for the siblings of children.





We do the same thing thinking that for the minors the variable Siblings/Spouse only shows the siblings. In this graph (2.2.2.), the results are even more interesting, we see that the children without siblings, with one or with two are extremely more likely to survive than the ones in big families.

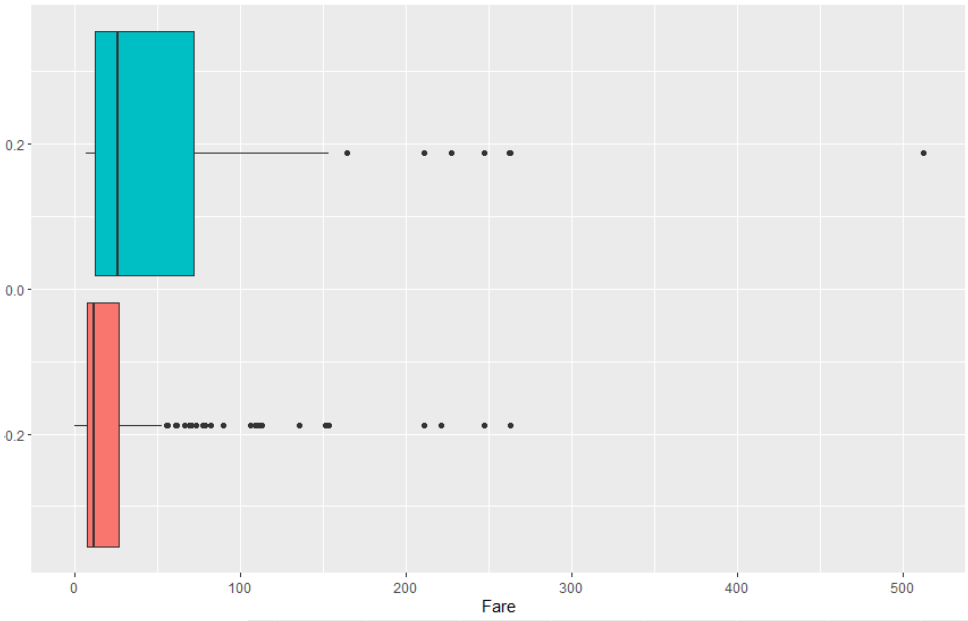
That brings us to one last question: **What is the influence of the size of the family on survival?**

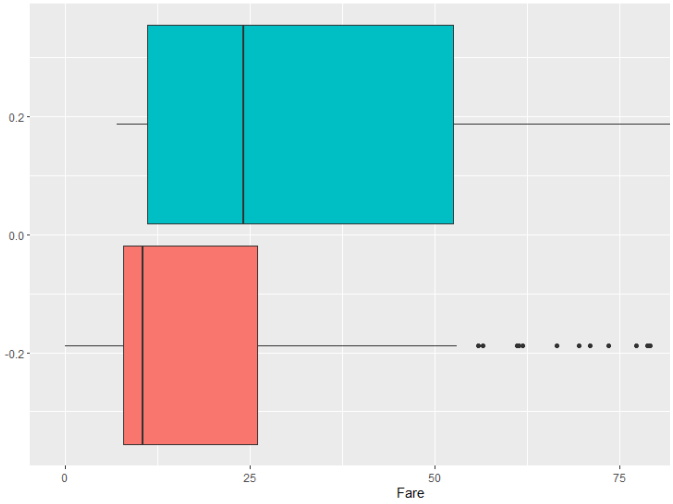


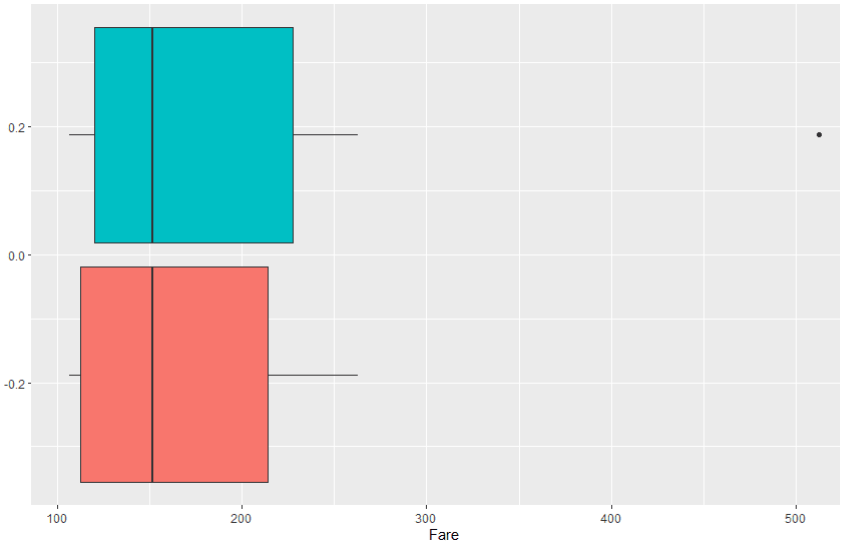
This last Scatterplot is very interesting. It shows us the number of parents and siblings of children and if they were more likely to survive or to die. We can draw several conclusions from this graph. First, the children that were traveling alone were more likely to die which is logical. But, on the other hand, children from big families with a lot of parents and a lot of children usually die. We can try to draw some hypotheses like that these families preferred to die together than to let someone sink with the boat. Or that the parents alone couldn't take care of a big number of children in a fast and strategic evacuation. In conclusion, we can say that the more the families were big or little the more they did instead in medium-size families the children were more likely to survive.

2.3. Did money influence survival? If yes, was it the same on a large scale as in a smaller one?

We as humans always talk about a recurrent debate: does money buy life? Is it ethical to save people because of their money? This study of cases is a very good option to study this debate in an important episode of humanity. For that reason, we can ask ourselves if money had a big influence on survival. The results of these boxplots were pretty obvious: the median of the people who survived is richer. But what happens if we make the same for people with more than 100.00 and less than 100.00, dividing his question into 2 cases.



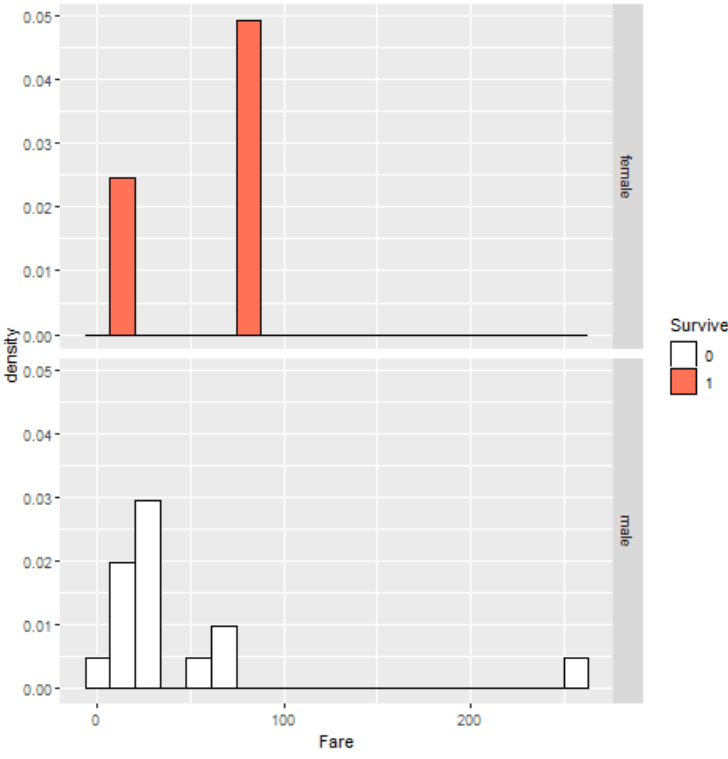




For people with less than 10,.000(1rst graph 2.3.2.) we see that the amount of money is very relevant. On average the survivors are richer and 3 quartiles of the survivors are richer than half the people who died. On the other hand, after 100,00(second graph 2.3.3. The money doesn't matter. The 2 boxplots are very similar which means that there is no relation between survival and money. So we can say that money is relevant to survival. It even seems to be a point where you are rich enough and that's when money doesn't really make the difference.

2.4.Were the aged people more likely to spend more money on the trip?

Taking into account that the average age of the passengers is 29 years old. We have created a variable with people with more than 50 years considering them old. We have created a histogram in the function of the fare and we have seen that older people don't spend as much money as we could think. But if we divide these variables into males and females and we look for the survivors the results are surprising.



Men of more than 60 years

Survived	Pclass	Sex	Age	SibSp	Parch
0	2	male	66.0	0	0
0	1	male	65.0	0	1
0	3	male	70.5	0	0
0	1	male	61.0	0	0
0	1	male	62.0	0	0
0	3	male	65.0	0	0
0	3	male	61.0	0	0
0	1	male	64.0	1	4
0	1	male	65.0	0	0
0	1	male	71.0	0	0
0	1	male	64.0	0	0
0	1	male	62.0	0	0
0	1	male	61.0	0	0
0	1	male	70.0	1	1
0	3	male	74.0	0	0

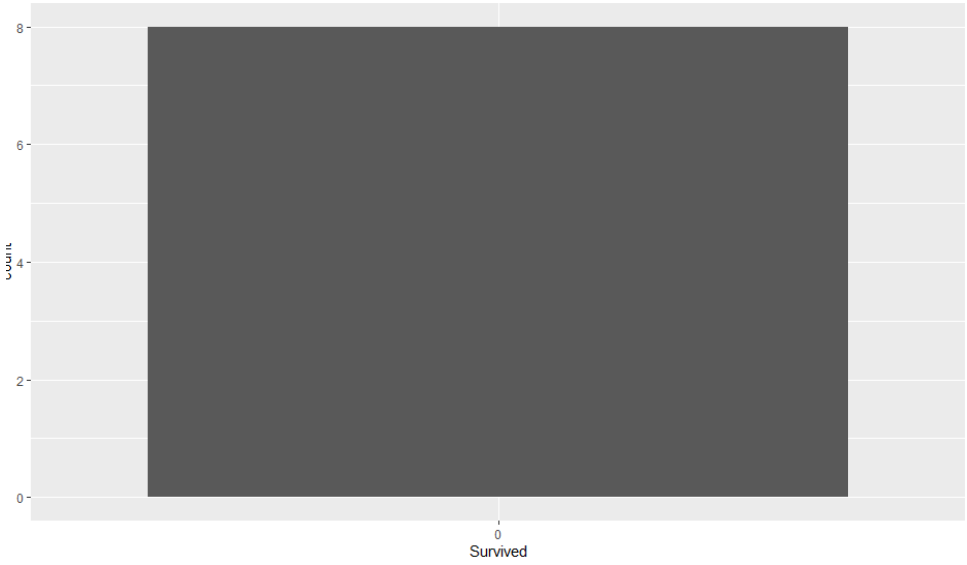
Women of more than 60 years

Survived	Pclass	Sex	Age	SibSp	Parch
1	1	female	63	1	0
1	3	female	63	0	0
1	1	female	62	0	0

Only women survived in between the older people of the boat and none of them died which is more surprising. Even if it wasn't the answer we were searching for, that seemed more interesting to us. So we continued researching for an explanation. The only thing we found creating data frames of people of more than 60 years is that maybe the first woman and the 8th man were engaged. For the rest, it seems that the women survived by luck. Or maybe we can explain this by saying that they were only three women of that age.

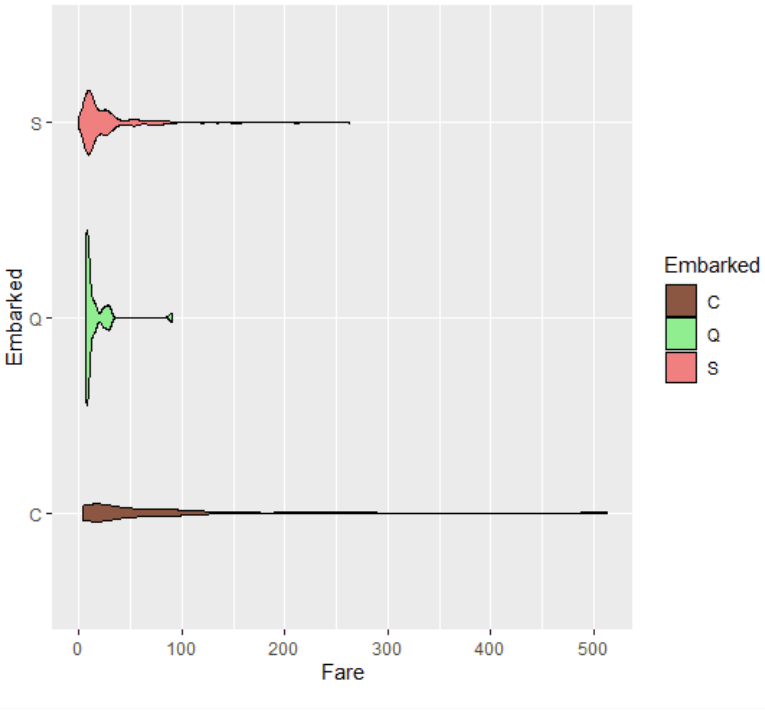
2.5. Do the people that worked on the Titanic were more likely to survive?

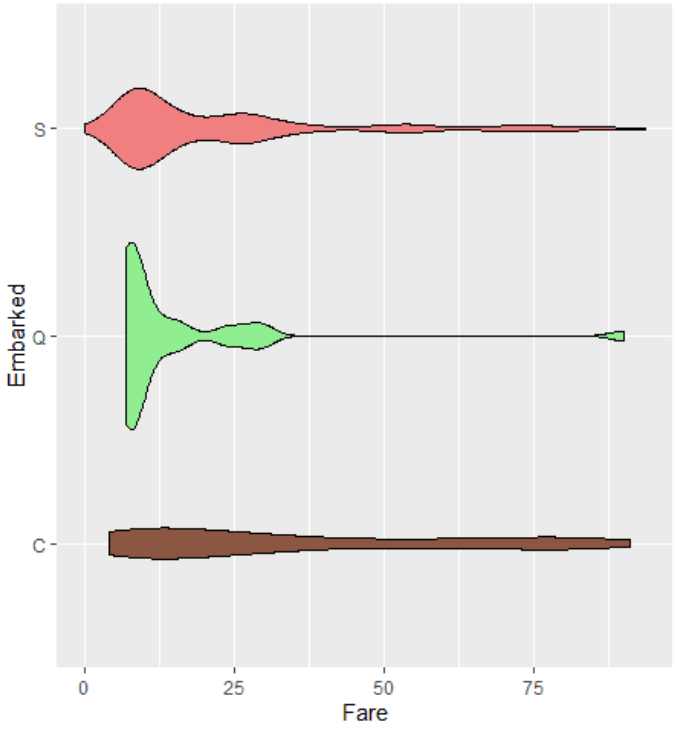
We suppose that the people that worked at the titanic were the ones who didn't pay anything (fare = 0). We have created the data frame for these individuals and everything works. They all travel alone and they all are young enough to work but old enough to be of age to work. We have created a bar chart that shows us the amount of them that survived and none of them survived. We can think that the crew worked to save the maximum number of passengers and clients even if that was a synonym of losing their own life.

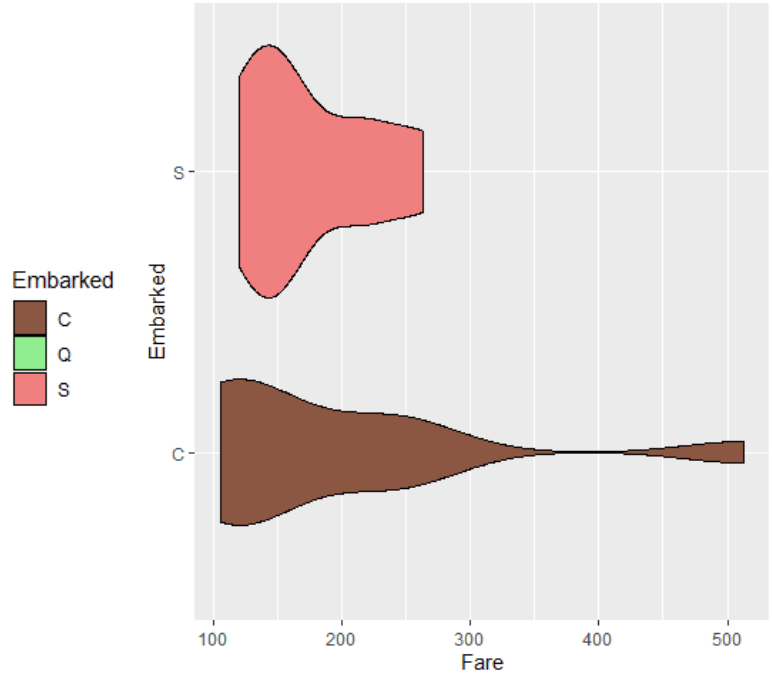


2.6. The people embarked in different ports for reasons of money?

We want to know if there is a link between the port of embarkation and the fare. We have created a graph with 3 curves (2.6.1.) in the function of the fare and we have shown the port of embarkation. If we look at this result, it is truly not very clear even if the people who have paid less seem to come from Queensbourg. But if we take a close look dividing the graph into one for people with who paid more than 100. and less, things change. We can say that Southampton is the biggest city because as we showed in the introduction 71% of people come from Southampton.



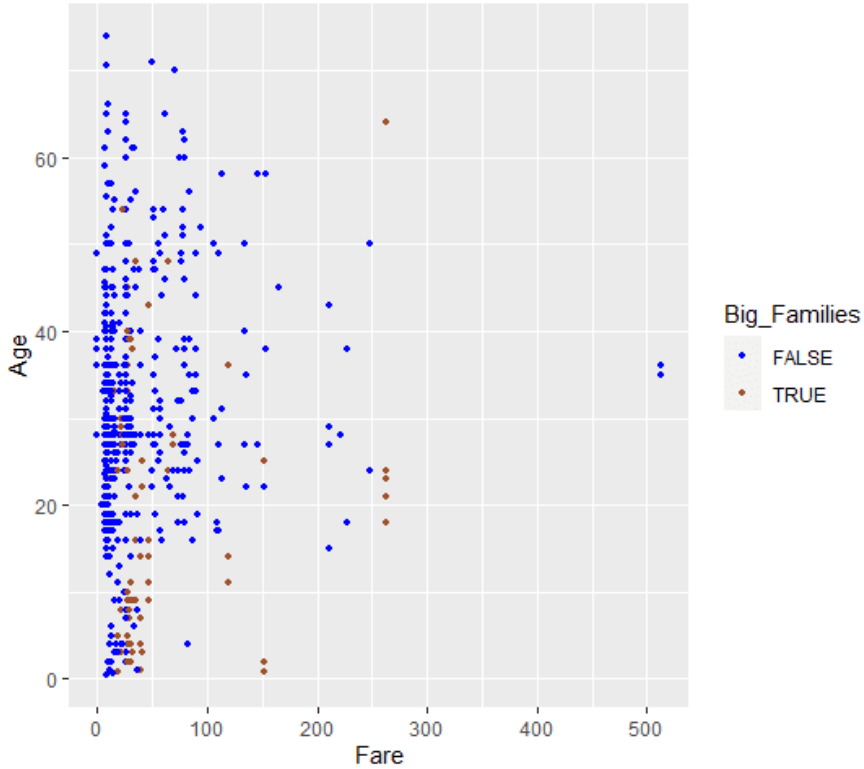




If we look at the first graph 2.6.2. We see that people who paid the least come from Queenstown, there is always the same amount of people from Cheesbourg and the number of people decreases when the fare grows. On the other hand, the second graph shows that the people who paid more than 100 nobodies came from Queensbourg and the ones who paid the most came from Cherbourg. If we only take into account what the data says we can say that Queenstown is a more industrial city that relies on the middle class and Cherbourg is a city where there are a lot of rich people, in addition, this city is in France and the tariff is more expensive because they travel more time. Southampton seems to be a middle point between the 2 cities and definitively the bigger one.

2.7. Was there a discount for big families in titanic or age ?

This graph(2.7.) is very interesting to search for discounts. We have the Fare the price people pay. In function of the age of people and the ones that are in blue represent members of big families. If there was some age discount we will be able to see some non-linear or linear relation this is not the case. And for the discount in big families, we can observe that people in blue don't have a lower fare than the ones that are not in big families. But this doesn't surprise us. The economic benefits of having a big family is a very recent cultural and political measure in developed countries to prevent demographic diminution. In this period, this measure was not necessary so they didn't exist.



Conclusion:

In conclusion, we can say that there are a lot of different types to visualize the data, but some of them are more appropriate to find solutions to our questions. For example curves, histograms and boxplots are more useful to display continuous numeric data. On the other hand, a bar chart or even pie can be more useful to visualize discrete numerical data or categorical data. And scatterplot to find the correlation between variables. In all these figures we can use different methods to visualize many variables at the same time. That will be most of the time the most important thing to answer difficult questions.