TITANIC EXPLORATORY DATA ANALYSIS

Students: Álvaro Martín Ruiz and Pablo Simón Martín Sánchez

FIRST APPROACH

Before we start drawing any conclusion or looking for relationships, we are going to study and understand our variables first. To do so, we are going to start by using the functions *head(), tail(), summary()* and *str()* to have an idea of how our data is, and later we are going to look deeply in every variable making use of the functions table and prop.table for the categorical variables and summary for the numerical ones.

`head(titanic.train)`

```
> head(titanic.train)
   Survived Pclass     Sex Age SibSp Parch    Ticket    Fare Cabin Embarked
2         1      1  female  38     1     0 PC 17599 71.2833   C85        C
5         0      3    male  35     0     0    373450  8.0500              S
7         0      1    male  54     0     0     17463 51.8625   E46        S
10        1      2  female  14     1     0    237736 30.0708              C
11        1      3  female   4     1     1  PP 9549 16.7000    G6        S
14        0      3    male  39     1     5    347082 31.2750              S
```

`summary(titanic.train)`

```
> summary(titanic.train)
 Survived Pclass      Sex          Age            SibSp           Parch           Ticket       Fare                Cabin     Embarked
 0:412    1:167   female:243   Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   CA. 2343:  7   Min.   :  0.000             :515    :  0
 1:256    2:138   male  :425   1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.0000   1601    :  6   1st Qu.:  7.925   B96 B98   :  3   C:132
          3:363                Median :28.00   Median :0.0000   Median :0.0000   347082  :  6   Median : 15.246   C22 C26   :  3   Q: 57
                               Mean   :29.15   Mean   :0.5748   Mean   :0.4042   347088  :  6   Mean   : 34.066   C23 C25 C27:  3   S:479
                               3rd Qu.:35.00   3rd Qu.:1.0000   3rd Qu.:0.0000   382652  :  5   3rd Qu.: 34.109   E101      :  3
                               Max.   :74.00   Max.   :8.0000   Max.   :6.0000   CA 2144 :  5   Max.   :512.329   B18       :  2
                                                                                 (Other) :633                     (Other)   :139
```

`str(titanic.train)`

```
> str(titanic.train)
'data.frame':   668 obs. of  10 variables:
 $ Survived: Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 1 2 ...
 $ Pclass  : Factor w/ 3 levels "1","2","3": 1 3 1 2 3 3 3 2 3 2 ...
 $ Sex     : Factor w/ 2 levels "female","male": 1 2 2 1 1 2 1 1 2 2 ...
 $ Age     : num  38 35 54 14 4 39 14 55 2 28 ...
 $ SibSp   : int  1 0 0 1 1 1 0 0 4 0 ...
 $ Parch   : int  0 0 0 0 1 5 0 0 1 0 ...
 $ Ticket  : Factor w/ 681 levels "110152","110413",..: 597 473 86 133 617 334 414 154 481 152 ...
 $ Fare    : num  71.28 8.05 51.86 30.07 16.7 ...
 $ Cabin   : Factor w/ 148 levels "","A10","A14",..: 83 1 131 1 147 1 1 1 1 1 ...
 $ Embarked: Factor w/ 4 levels "","C","Q","S": 2 4 4 2 4 4 4 4 3 4 ...
```

Now, we are going to extract the main characteristics of each variable:

```
> summary(Survived)
  0   1
412 256
```

```
> summary(Pclass)
  1   2   3
167 138 363
```

```
> table_sex = table(Sex);table_sex
Sex
female   male
   243    425
```

```
> prop.table(table_sex)
Sex
   female      male
0.3637725 0.6362275
```

```
> summary(Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.42   22.00   28.00   29.15   35.00   74.00
```

```
> summary(SibSp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max
 0.0000  0.0000  0.0000  0.5748  1.0000  8.000
```

```
> summary(Parch)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4042  0.0000  6.0000
```

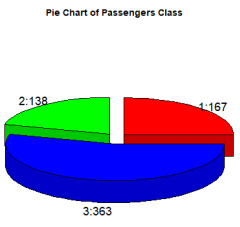
```
> summary(Fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   7.925  15.246  34.066  34.109 512.329
```

```
> table_embarked = table(Embarked); table_embarked
Embarked
      C   Q   S
  0 132  57 479
```

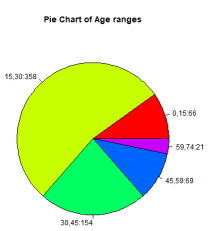
```
> prop.table(table_embarked)
Embarked
                  C          Q          S
0.00000000 0.19760479 0.08532934 0.71706587
```

We could also analyze the variables with graphs instead of using tables. Here there are some pie charts of different variables:

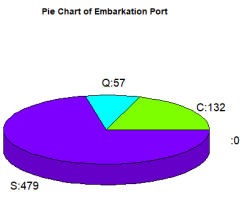
3D Pie chart of Pclass



Pie chart of age ranges



Pie chart of embarkation port



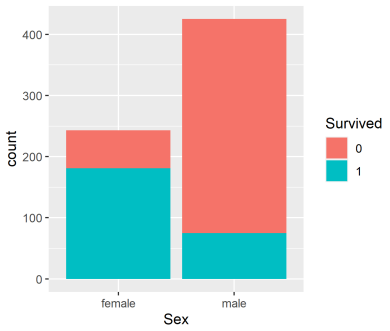
We can also check if there are any missing values on the data, using the function *is.na*(), which checks every value of a variable and gives us a TRUE/FALSE whether there is a value missing or not.

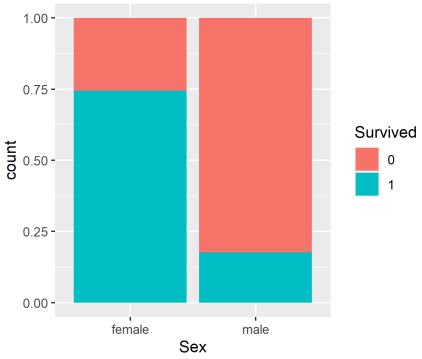
```
> table(is.na(Age))

FALSE
  668
```

If we do it with every variable, we'll see that there are no missing values in any variable, however, if we have a look at the column of "Cabin", we'll see that most of the spaces are empty. So, how is that possible? This is because to R, *NA* and empty "" are different. The reason for it is that "" is a blank, and NA is something that is truly missing.

QUESTION 1- BY VISUALIZING THE DATA, TRY TO SEE IF THERE IS ANY RELATIONSHIP BETWEEN SURVIVING AND SEX, AGE, FARE AND PCLASS

1) First, let's look at the relationship between **sex and survival**:

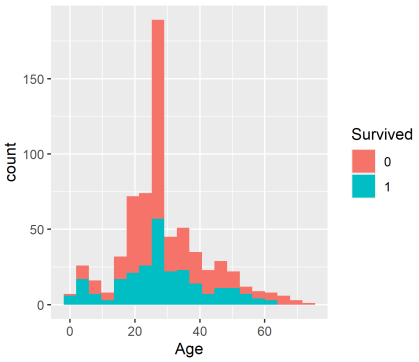


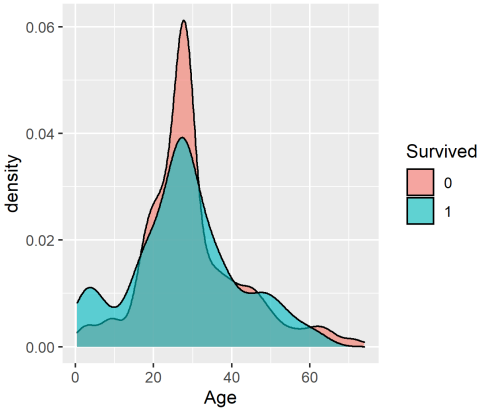


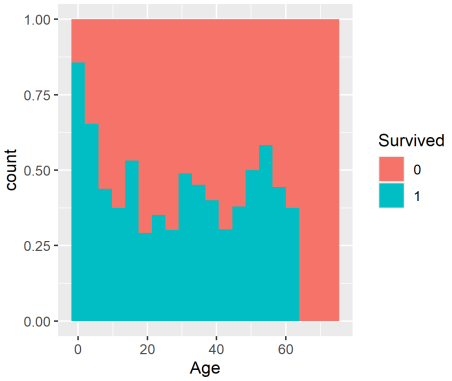
We can see that even though only one third (approximately) of the crew were women, only less than one third of the survivors (approximately) were men.

As we can see in the second plot, 75% of the women survived, whereas just less than 25 % of the men did it.

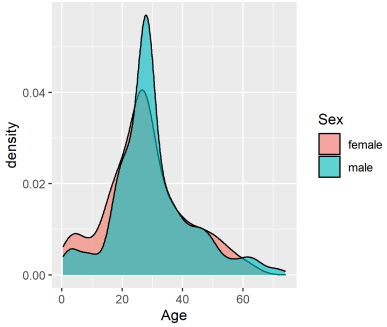
2) Age and survival

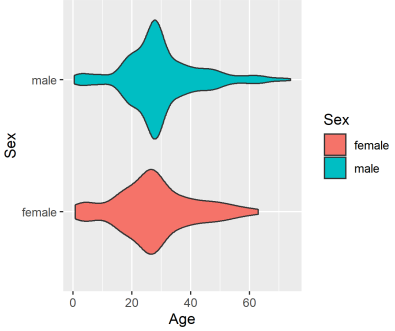






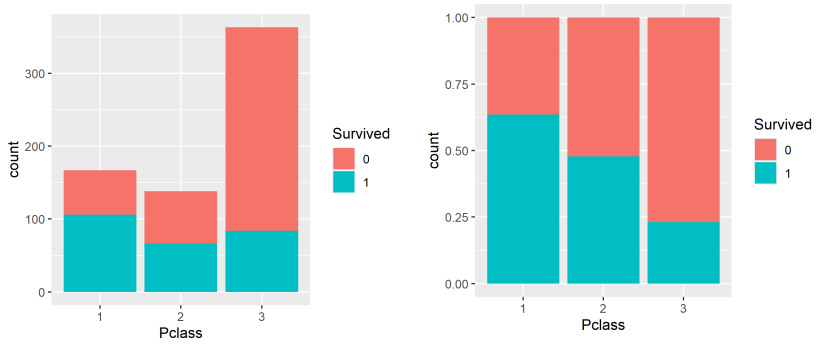
We can observe that children were probably a priority in the evacuation of the boat, since the young ages are the ones with higher survival rate. As could be expected, the survival rate among adults is lower than among children, nevertheless, we can see an upturn in the ages between 50-60. What might be the explanation? It is generally known that in the sink of the Titanic, the captain explicitly issued an order for women and children to be saved first (""Women and children first"), and we have already proved that women had a higher survival rate, so maybe, between 50-60 years old there were more women than men. Let's see:





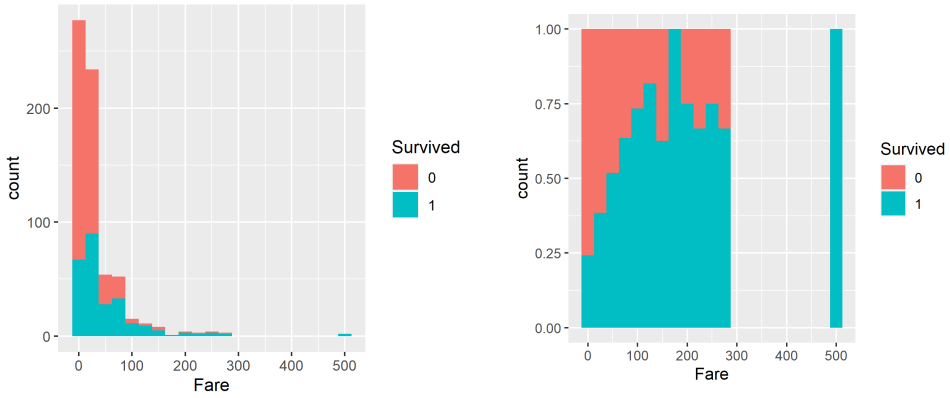
Exactly as we guessed, the reason why the survival rate increases among people about 55 years old is that in that age range there are more women than men. Linked to this but in the other way, we can observe that the age range with higher mortality coincides with the one with a higher percentage of men.

3) **Pclass and Survival**



It seems like people of a higher social-class had some preference when leaving the boat.

4) **Survival/Fare**

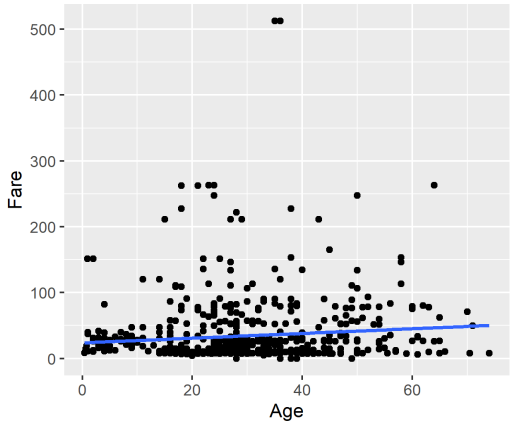


We could imagine from the previous graphs (Pclass-Survival), that in this one will happen something similar, the higher the ticket price (which is representative of the social class), the higher chances of surviving.

Q2- IS OLDER PEOPLE WEALTHIER? HOW WAS THE MAJORITY OF THE PASSENGERS IN TERMS OF AGE AND WEALTH?

Prediction: Older people are richer than young people.

How can we answer the question? Since the ticket fare can be highly representative of how wealthy a person is, we are going to see if there is a positive linear relationship between the variables age and fare.

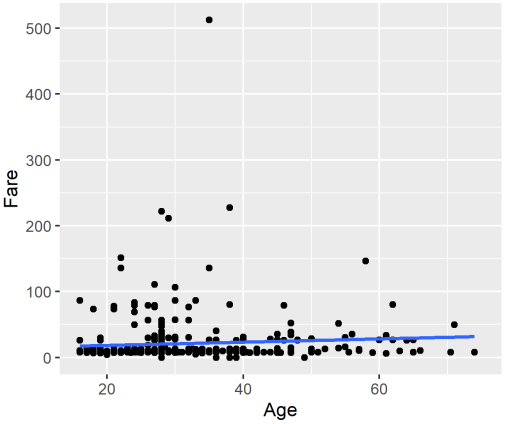


By looking at the graph we can state that, although there is a positive linear relationship, it isn't strong enough to draw any conclusions.

What we can do now is to “clean” the data of the plot. If we think about the variables for a second, we can assure that the fare of a baby of 1 year doesn’t make any sense with what we want to analyze. Children don’t have their own “fortune”, the parents are the ones who probably paid their tickets, hence we can remove all the values of age lower than 16.

Another important thing to bear in mind is that if a member of a family is really rich, he or she might have paid the ticket of the rest of family members, so the whole family would have the cabins next to each other (we are supposing that cabins were organised by class, which is the normal thing).

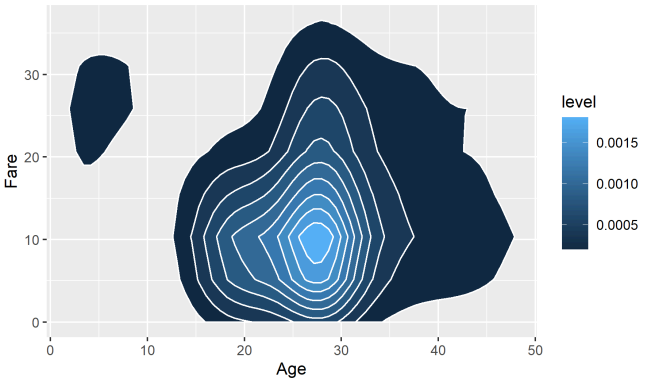
In conclusion, it would be interesting to repeat this graphic but just taking into account the people older than 16 years old who traveled alone. To determine the range of age, we can use the function *xlim(),* and to select the people who traveled alone, we can create a new data set with just the rows of the people who didn’t have parents, siblings, spouses or children abroad, and use it as the data frame where the function ggplot searches the information.



However, it happens the same as before. Even though our graph is more realistic now, there isn't a correlation strong enough to say that older people were wealthier.

Q3-How were the majority of the passengers in terms of age and wealth?

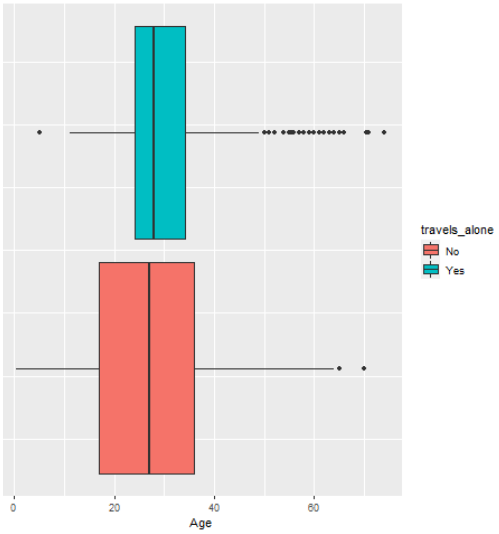
With the 2d density graph, we can also see clearly which are the most repeated values. We can observe that the majority of the crew was between 18-40 years old, and paid a “low cost” fare (lower than 40).



Q4- Is there any relation between age and the passenger's class and company?

Let’s start analysing the passenger’s company in terms of their age. Just by looking at the summaries, we can see that, as it could have been guessed, people traveling alone´s mean is higher than the other, mainly because really young children, which could be considered as outliers, affect the mean heavily. In contrast, we see that the median is similar, so we could guess that, for middle-aged people, there is not a clear trend. In order to visualize the data in a better way, we could use some plots:

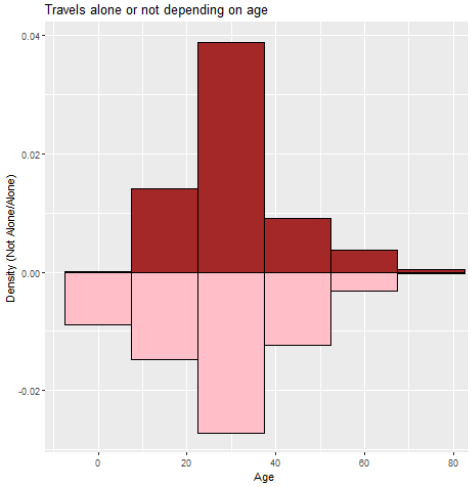
```
summary(Age[travels_alone == "No"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.42   17.00   27.00   26.83   36.00   70.00
summary(Age[travels_alone == "Yes"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 5.00   24.00   28.00   30.78   34.25   74.00
```



As we can see in the boxplots, the IQR for people traveling alone is smaller than for people accompanied, because the data is heavily condensed in the range of 24 and 34.25 years. This forces the appearance of many outliers, mainly on the right side of the plot. The data is also clearly right-skewed. These happens because the mean is higher than the median

On the other hand, people who were accompanied have a bigger IQR, thus limiting the existence of outliers, even though there are some. In this case, the data is slightly left skewed because the median is a bit higher than the mean

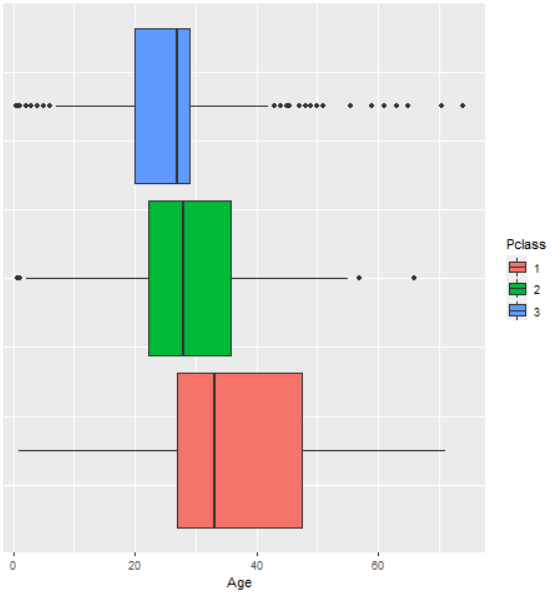
The way the boxplots look makes sense because childrens mostly travel accompanied. As a result, as in the boxplot above there’s almost no children, all data is more condensed and tends to be higher. In contrast, in the boxplot below, the existence of children forces the data to be more evenly distributed.



Using these other two graphs, we can see our previous guesses more clearly. Most middle-aged people travel on their own and almost every child is accompanied. As an extra information, we get that really old people (over 67 years old or so) were mainly travelling on their own.

```
summary(Age[Pclass == "1"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.92   27.00   33.00   36.07   47.50   71.00
summary(Age[Pclass == "2"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.67   22.25   28.00   28.73   35.75   66.00
summary(Age[Pclass == "3"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.42   20.00   27.00   26.12   29.00   74.00
```

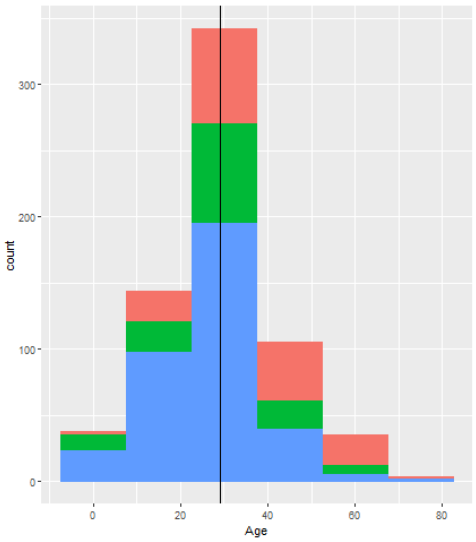
Now, let's look at the relation between the passenger's class and their age. Just by looking at the summaries, we can see there's a lot of differences.



To see these differences more clearly, we can graph the boxplots

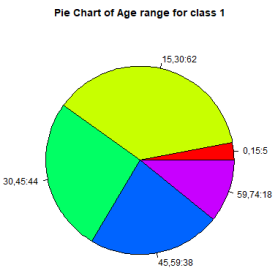
As we can see in the boxplots, the median for passengers of class 2 and class 3 are very similar. However, for class 2, the IQR is bigger and the data is a bit right skewed, while for class 3, the IQR is smaller and the data is really left skewed. Because of the slight differences on the IQR, the first boxplot has many more outliers, both in the right and in the left. These means that most passengers of class 3 were people between 20 and 29 years, but there were still some small kids and some people older than 45 years that could be considered as outliers

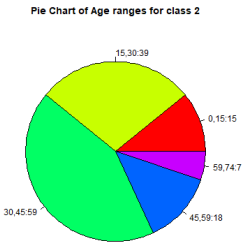
Finally in the boxplot for passengers of class 3, we see that the data is evenly distributed across all ages, as the IQR is very big, thus limiting the existence of outliers. In comparison to the others, the median is higher and the data is really right-skewed, which could be translated as passengers of class 3 being a bit older in general.

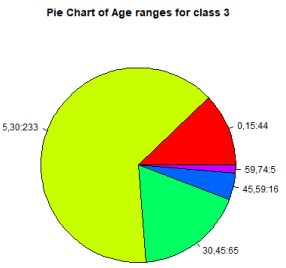


In the histogram, we can see how most young people belonged to class 3, while older people were more likely to belong to class 3.

In addition, we can also use some pie charts to visualize the data

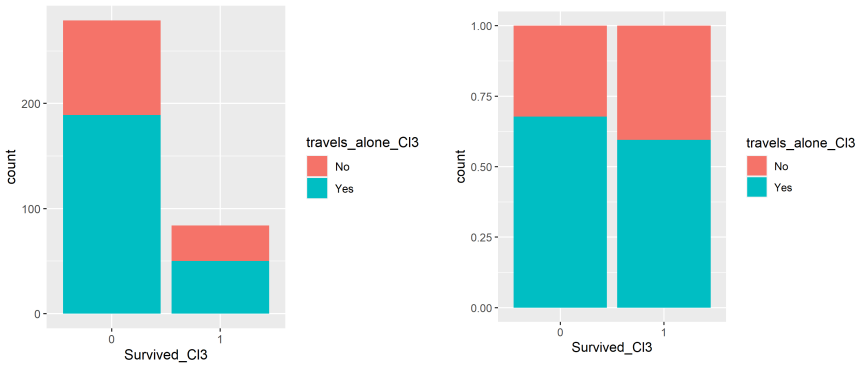






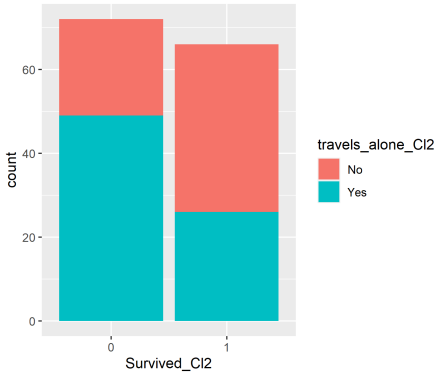
Q5 - Does the passenger's class affect the survival rate and does the relation between Traveling alone and surviving change depending on the class?

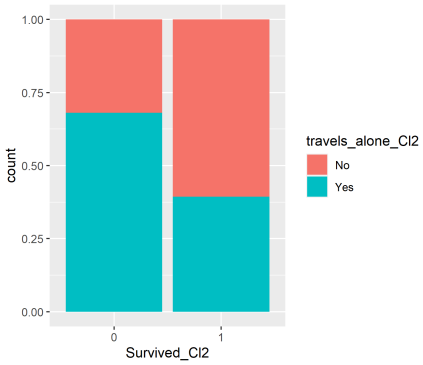
In the first question, we saw how people of higher social-class were more likely to survive. In addition, the example shown in class proved that most people traveling alone perished, mainly because most people traveling alone were men. Now, we can try to see if these two are related



Class 3

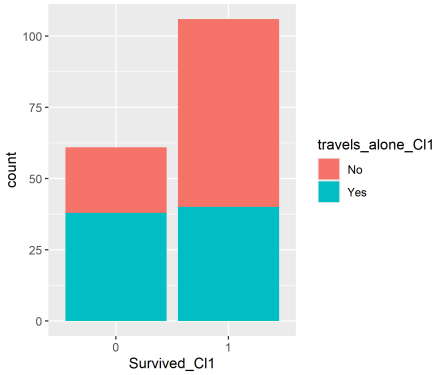
In the first graph we can see how most passengers of class 3 couldn't survive. The second graph confirms that, out of those who died, most were traveling alone, a situation that also happened with those who survived.

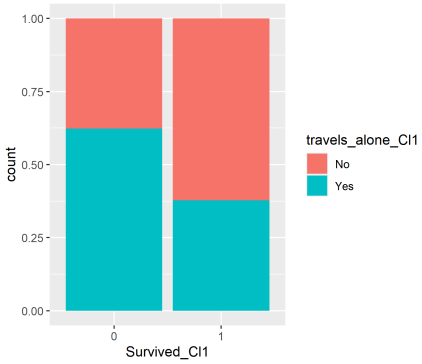




Class 2

In the first graph we see that there were slightly more people dying than surviving in class 2, but the difference is infimal. The second graph shows that most people who perished were travelling alone, but, on the contrary, most people who survived were accompanied.





Class 1

The first graph shows that most passengers of class 1 survived. Taking a look at the second graph, we can see that most passengers of class 1 who died were traveling alone, but most passengers who got to live were accompanied.

In conclusion, after exploring and analyzing the different graphs, we can see some similarities between some of them. For example, the second graph of class 2 and class 1 looks almost the same, because in both cases, people that perished tended to travel alone. In other terms, families had some sort of priority to leave the boat earlier. We can also see how the first graph of class 3 is almost the opposite to the first graph of class 1, because a majority of class 1 passengers survived, but the opposite happened to most passengers of class 3