Titanic.train Assignment

Introduction to Data Science

Dual Bachelor in Data Science and Engineering and Telecommunications Engineering

Alejandro Barroso - 100499081
Juan José Rosales - 100499176
Group 196

uc3m | Universidad Carlos III de Madrid

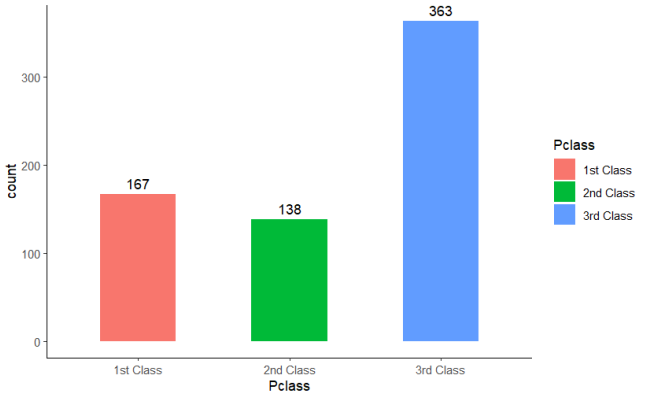
Index

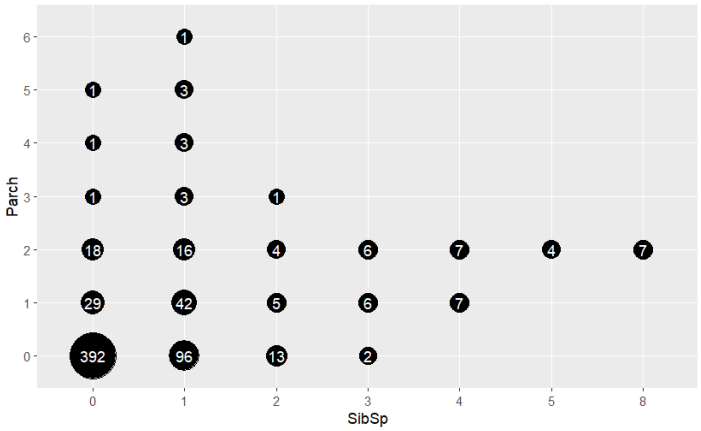
1. INTRODUCTION

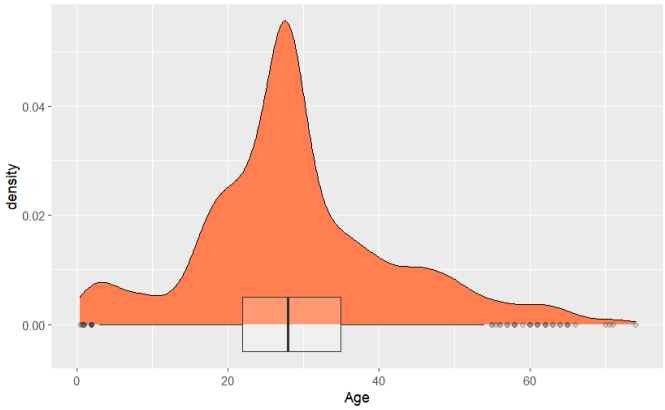
For the realization of this project we have used the exploratory data analysis techniques learned in class to obtain conclusions from the data frame “Titanic_train.Rdata”. Besides, the visualization of this results with the *ggplot2* package has allowed us to represent the data relationships in a clear way which lets other people understand easily our outcomes.

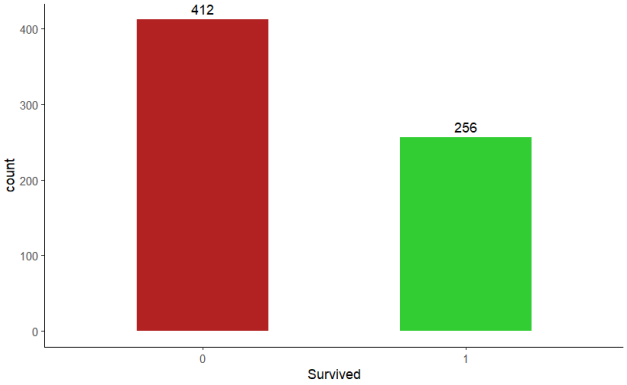
On our first approach to the data frame we needed to understand every variable and its values. We divided each variable in two groups: the categorical variables (“Survived”, “PClass”, “Sex”, “Embarked”, “Cabin”) and the numerical ones, where we can find “ SibSp“ and “Parch” (discrete variables) and “Fare” and “Age” (continuous variables).

For each variable we did a first visualization with “table” and “prop.table” for categorical variables and the “summary” function for numerical variables. Then, we decided to do simple plots to understand them in a more visual way selecting the most useful plot for each variable. Some examples:

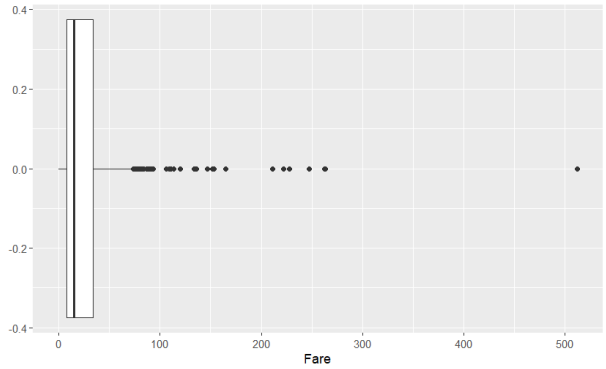








From this first insight we concluded that that Ticket variable was an arbitrary variable which won’t be useful in our investigation.
Besides, the boxplot that represented the Fare values indicated us the existance of a possible extreme outlier, since we can see that it is far away from the other values.



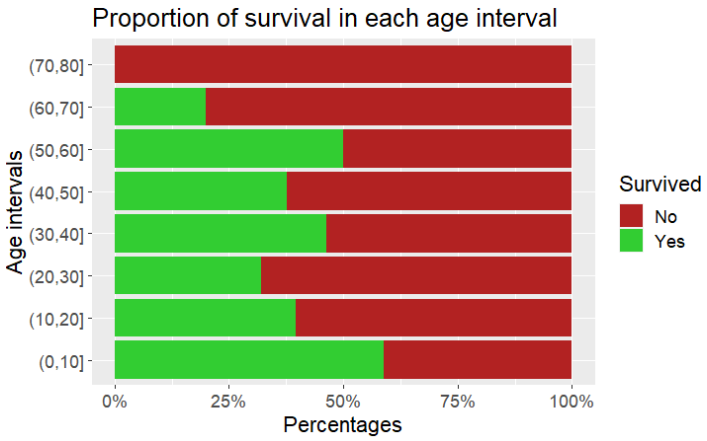
QUESTIONS

First Question: Did the passengers had a better chance of surviving dependig on their age or sex?

The first question we had about the data was how did the age and sex of a passenger affected its survival. Our theory before starting to get the values was that men and elderly people had lower chances of surviving.

First we are going to create a new variable in which we split the age discrete variable into different intervals with this script: `data$DiscretizedAge = cut(data$Age, c(0,12,18,30,40,50,60,70,80,100))`

Now, the intervals are the levels of a factor variable so we can plot them easily.



We can see how the graph changes as it goes up in the age intervals. We can affirm that the highest percentage of survivors is from 0 to 10 and from 50 to 60. On the other hand, the group with the least survival rate is from 60 to 80 and from 20 to 30, the latter having the highest number of samples.

In the second part of this question we are going to relate sex with the previous plot to find some more relation between the variables age, sex and survived.

This plot combines two histograms with male and female passengers. At first sight we can notice two things: there were much more males, and the female histogram is almost completely green. With this graph, we prove that although there were many more men, their mortality was much higher, except in the age range from 0 to 20, the children. So, these two plots conclude that women and children were much more likely to survive than men and old people.

Second Question: Did the price of the ticket have any relationship with the survival rate of the passengers?

In this question we want to check whether the fare paid by the passengers has any relation with their survival rate. For this purpouse we are going to create a new variable which will be the variable fare but in intervals: `data$DiscretizedFare = cut(data$Fare, c(-0.1,10,30,50,70,80,90,550))`

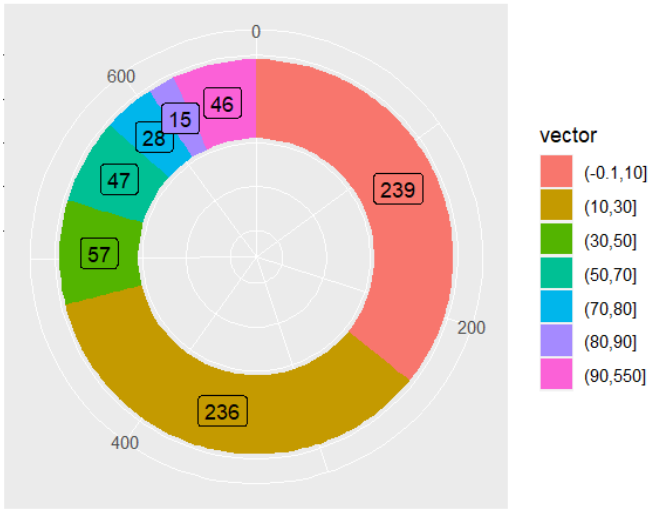
The first step is to visualy see how many passengers are in each interval. To do this we create a new data frame with the information of the number of people in each interval. Now with these data frame we can plot the donut graph.

```
vector = c()
vector2 = c()
for (i in levels(data$DiscretizedFare)){
  vector = c(vector,i)
  vector2 = c(vector2,length(which(data$DiscretizedFare == i)))
```


Juan José Rosales and Alejandro Barroso

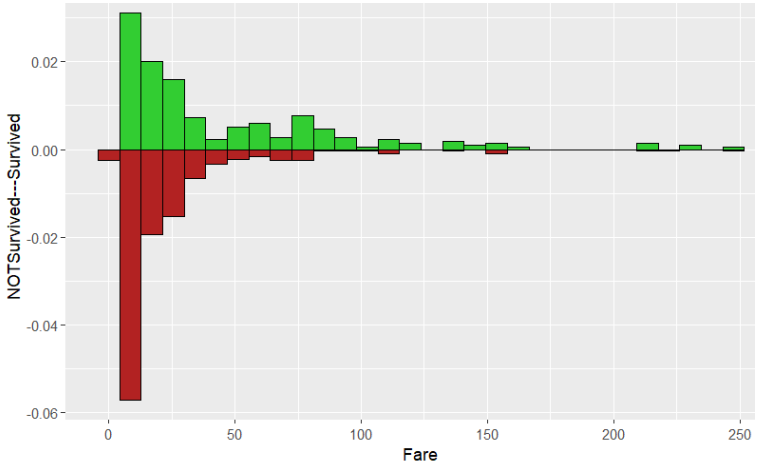
The plot show how almost 2/3 of the passengers has paid a fare between 0 and 30 (Cheap).

Number of passengers who paid the fare in these different intervals



The other 1/3 of the passengers had paid between 30 and 550. So, in conlusion, we can imagine two groups: the passengers that paid less than 30, which are the majority; and the ones who paid more. Knowing this, we created a new plot in which we combine the variables fare and survived and make conclusions about it. This plot is a dual histogram density plot in which we can visualize at the same time the passengers and their survival rate. To create this new plot we took out the outliers that are over 250, because they can cause a distorsion of the real plot.

This histogram tells us that passengers who had less expenses were more likely to die. We can see how the first bins are bigger in the red side. When the fare increases over 50, red bins decrease and green bins keep an elevate survival rate growing again around 75, meaning that more people survived in that positions. In the highest Fare positions, over 150, almost nobody died. With this information we can conclude that passengers who paid a higher fare may had some preference to the lifeboats.



Third Question: Is having a cabin related to the class of the person and the embarkation place?

To answer this question, the first thing we had to do was creating a variable that stored if the passenger had a cabin or not. We used the variable cabin to extract the data for the new variable, which we called *Room.*

```
aux = which(data$Cabin != "")
data$Room = rep("No Cabin",
                length(data$Cabin))
data$Room[aux] = "Cabin"
data$Room = factor(data$Room)
```

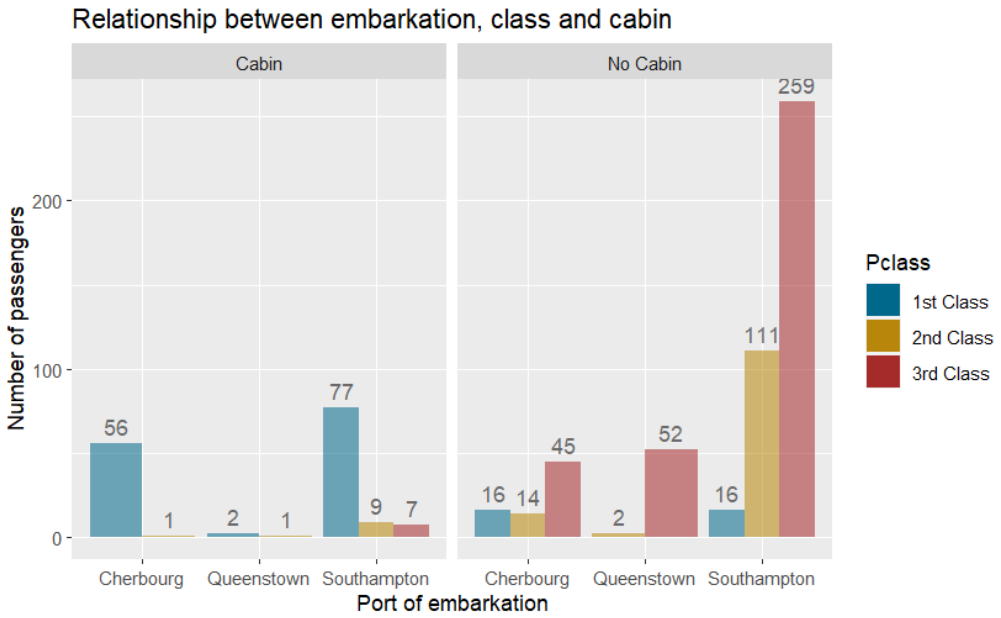
Once the new variable was ready, we searched the most efficient way of representing the variables Room, Pclass and Embarked, and we decided that a bar chart was the best option for this case.

Relationship between embarkation port, class and having a cabin or not

The first thing we can see is that the majority of people did not have a cabin. We can also conclude that almost all the people who had a cabin had a 1st Class ticket, what makes us suppose that having a cabin was very expensive and exclusive.

On the other hand, we can see that only 3 people from Queenstown had a Cabin. We think that the main reason of this is that in this port embarked just a few people, and only 2 were in the 1st Class.

In conclusion, having a cabin was strongly related to the person's class, but not much with the embarkation port.



Fourth Question: How does the age and the class of a person affected its chances of surviving?

The next question that we had was about the relation of a passenger's class with its age and how affected this combination to its chances of surviving. Since we knew that the older people were more likely to die in the sinking of the titanic thanks to the first question, we wanted to see whether their chances changed due to their ticket class.

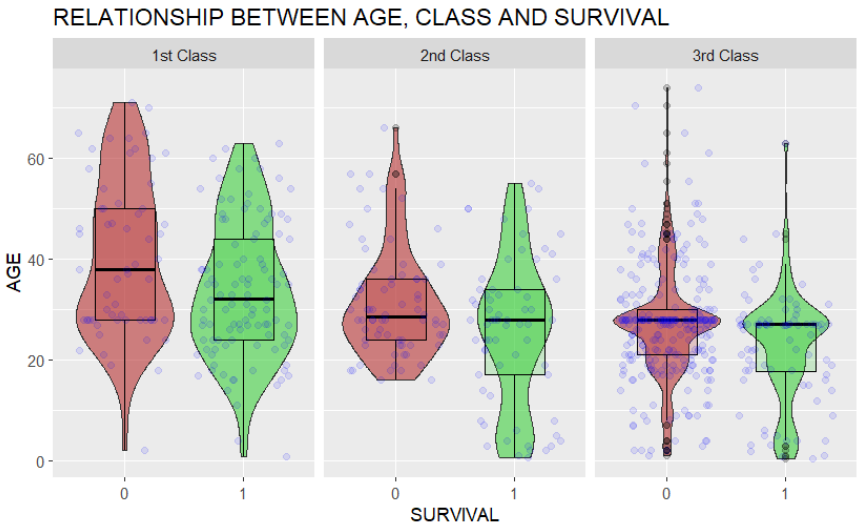
At first, we decided to use a violine-boxplot chart to see in a clear manner the relationship between these variables, but then we thought that adding a scatter plot over the other charts was a greater idea, since now we could also see the amount of people in each category represented as points.

Relationship between age, class and survival

The first thing that we can extract from this plot is that the median is higher between classes, which tells us that the first class passengers were on average a little bit older and the third class was composed of younger people.

Thanks to the points of the geom_jitter, we can clearly see that the third group was also the group with more not surviving passengers, which was not surprising due to the era in which the Titanic sinking happened.

Finally, the width of the violin plots in the range of children, from 0 to 15 years old approximately, expose that every children from the second class suvived.



Fifth question: How did travelling alone affect each interval of age in their survival?

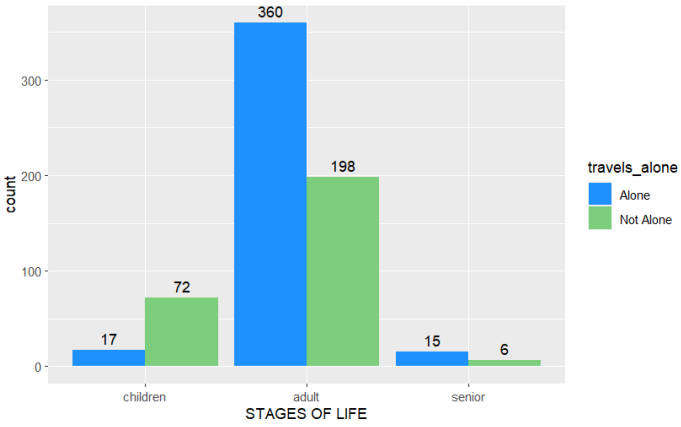
The answer to this question required the creation of two new variables, one that divided the passengers in two groups: the ones who travelled alone and the ones who traveled with a member of their family (spouse, siblings, parents or children). The other variable was the one for age-range groups, which we called *stages.* This new variable is a factor of the 3 stages of life: “children”, from 0 to 17 years old; “adults”, from 18 to 59 years old; and “seniors”, from 60 years old.

```
children = which(data$Age < 18)
adults = which(data$Age >= 18 & data$Age < 60)
seniors = which(data$Age >= 60)
stages = rep("children", length(data$Age))
stages[adults] = "adult"
stages[seniors] = "senior"
data$stages = stages
data$stages = factor(data$stages, levels = c("children", "adult", "senior"))
```

Since the passengers are divided in intervals, which aren’t of the same size, their content should be represented using percentages, so it becomes easier to visualize the relation between each interval.

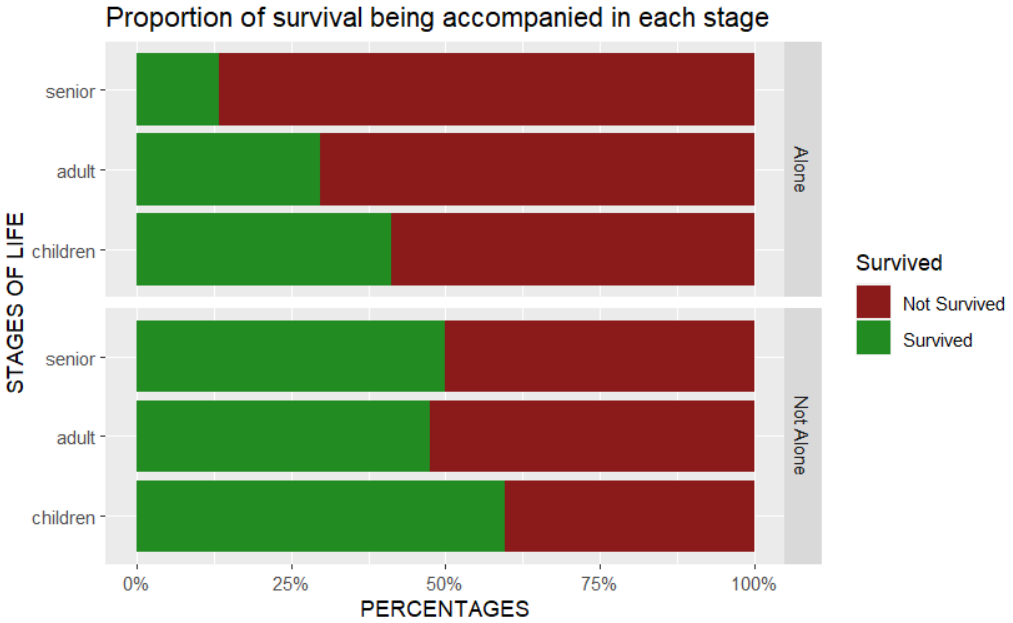
Proportion of survival being accompanied or not in each stage of life.

From the bar chart we can easily conclude that adults travelled in the majority with no familiar company, whereas most children travelled with a familiar, which was expected.



Besides, with the proportions of surviving or not we can clearly see that the passengers who were travelling alone, no matters the age, had lower chances of surviving than the ones travelling accompanied.

We can also conclude that children were the stage group who survived the most, even though they travelled accompanied or not. This result is coherent since people always prioritize the life of children over any other age group.



In conclusion, passengers who travelled alone had less chances of surviving no matters their age-group.