Titanic Survivor Predictor 3000

CLOUP Félix & LESAFFRE Maxime

Introduction :

In this project, we have analysed the data from the disaster of the titanic which happened in the night of 14th April of 1912. Indeed, We have computed how people survived depend on different features.

Data Engineering :

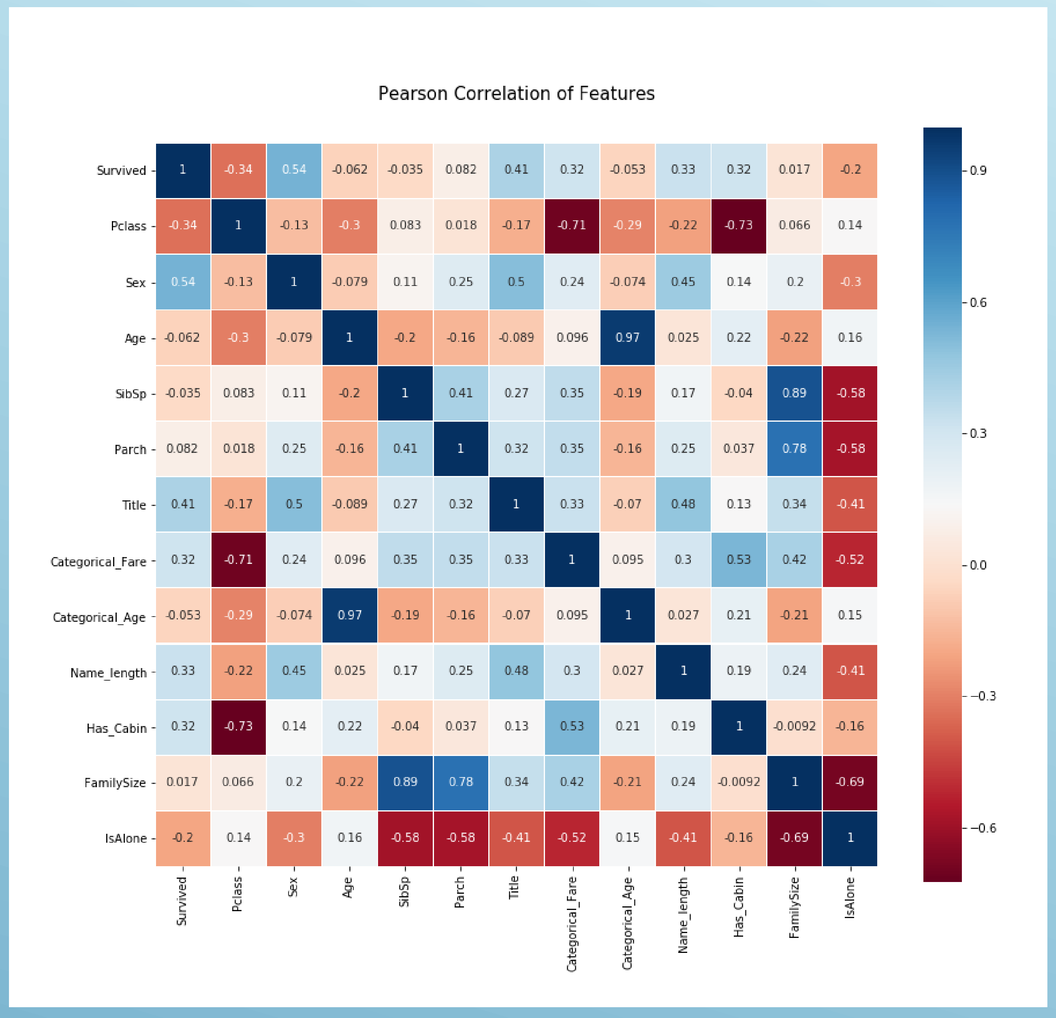
To be able to use this dataset, we had to make some modifications.

First, we cleaned the data. Indeed, if we take the example of the "age", we attributed random age following the age distribution for people without value in "age".

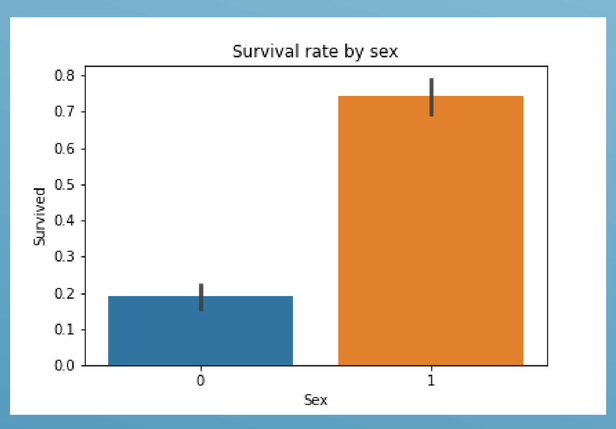
Then, we created features to get a better understanding of the data. For instance, we extracted the title of the person (e.g Mr, Miss, etc) to create a new feature.

We also divided the "fare" and "age" features into 5 and 6 categories to process them more easily.

Then, we made a Boolean is the passenger have a cabin or not, and we categorized passenger by the number of family number they had on board.

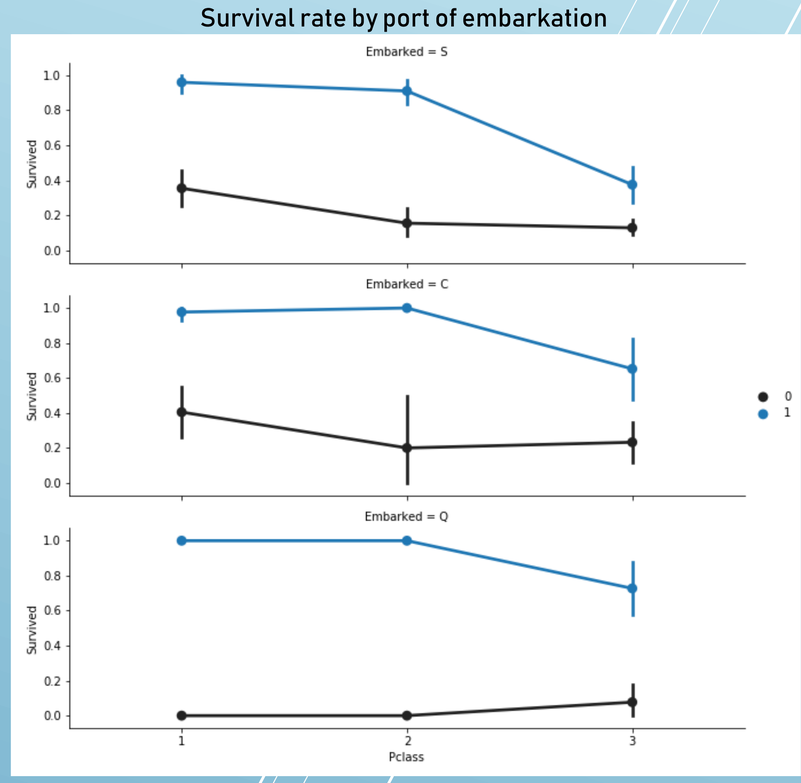


The Pearson Correlation Heatmap shows us the correlation between features.

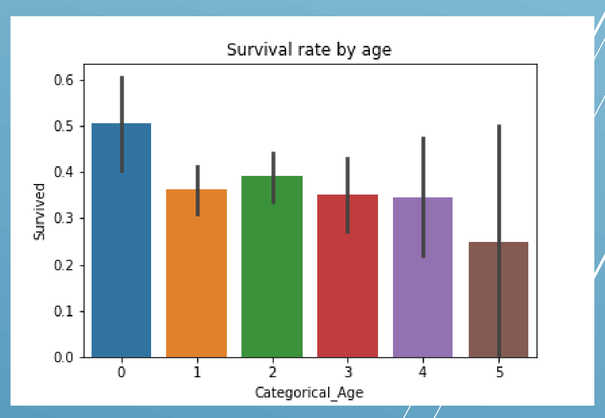


0. Man
1. Women

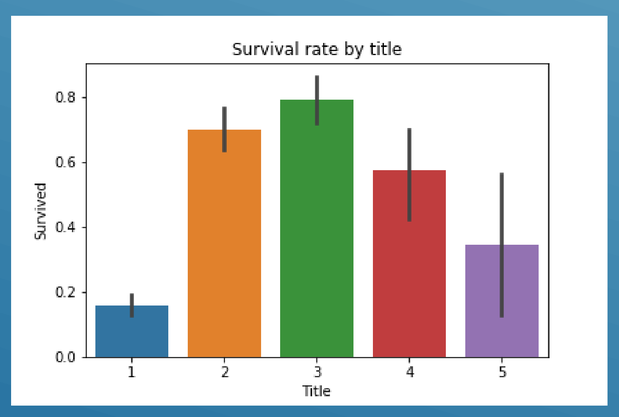
We can see that the proportion of surviving women is much higher than for surviving men.



0. Man
1. Women

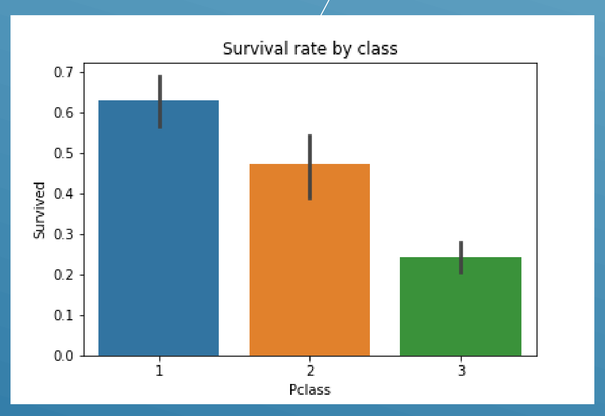


We used a function to make ranges of age and we see that the younger people survived more.

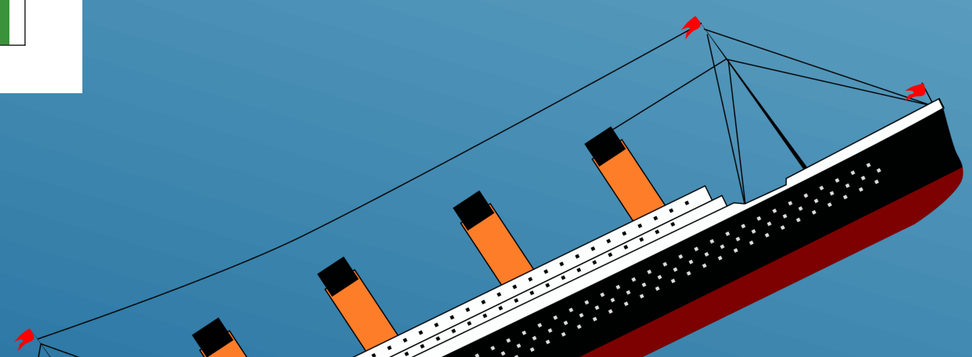


1. Mr
2. Miss
3. Mrs
4. Master
5. Others

As we show above, the proportion of Miss and Mrs is higher than other title for surviving passenger, so the gender is a good feature to guess if a passenger survived or not.



First and second class passenger had the chance to survive more.



Source: https://www.kaggle.com/c/titanic/data