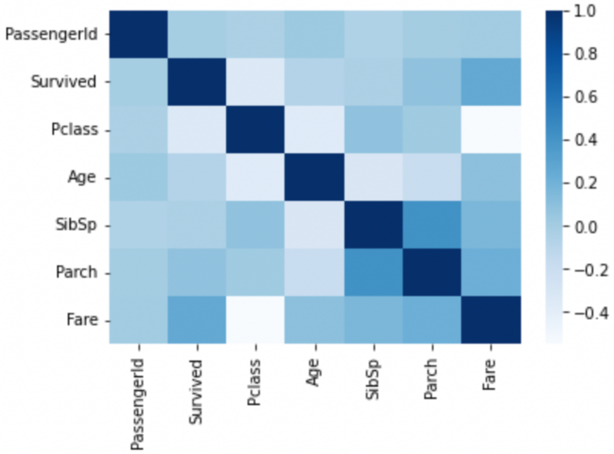
Daniela Idler
100375638
CSPC – 4800 – W01
March 10, 2022

Assessment 3 – Part 2

The focus of EDA is to understand your data. it's interesting to ask yourself some questions that should be answered before the actual work starts. Importing all the necessary libraries to Jupyter and uploading the dataset is a good start. Also, understand the number of observations and columns (891 and 12 respectively in Titanic's dataset). Checking the type of each variable and null values is important as well, I found 2 floats, 5 integers and 5 objects.

Creating a descriptive table is a simple step that brings a lot of information. Using the .describe() function I could discover that the surviving rate was 38%. In Titanic's dataset we have some missing values in two columns: age and cabin. Depending on your study, you will have to treat this missing data before start.

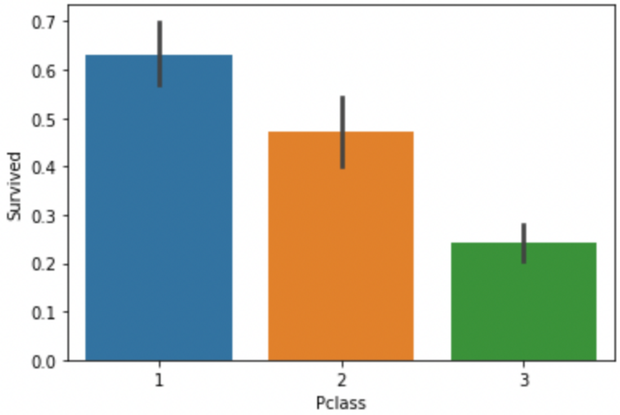
In our case, the objective of the study is to understand the correlation between some variables, to start I ran the correlation between all of them.



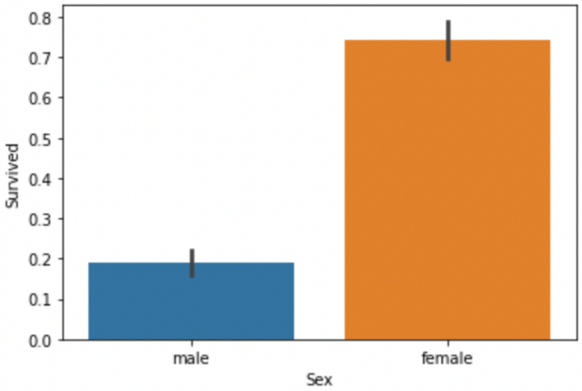
After that I decided to test 3 hypotheses:

- If the survival rate is associated to the class of passenger
- If the survival rate is associated to the gender
- If survival rate is associated to the age

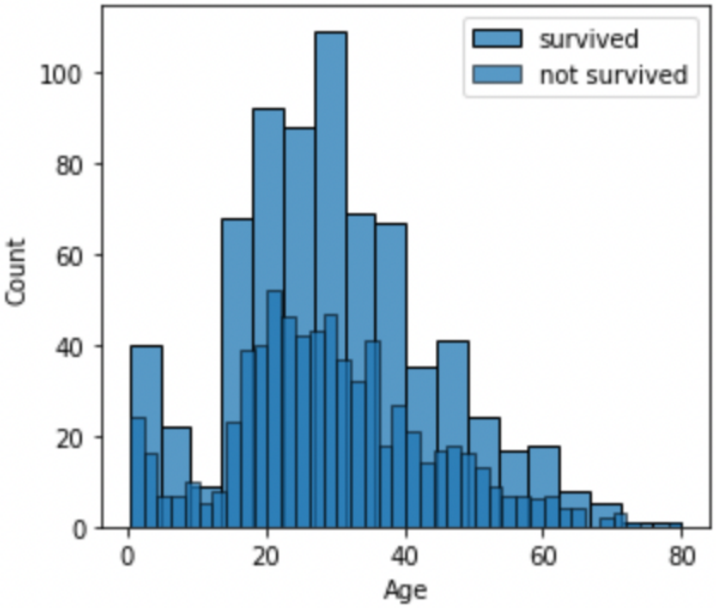
1. Class vs Survival: after creating a bar plot, it's clear that the higher the passenger class, higher their change of surviving.



2. Gender vs Survival: we can see that most people who survived from the disaster were woman.



3. Age vs Survival: you can see that the higher chances of survival are between 18 and 30 years old. Besides that, infants also have a higher probability of surviving.


survived
not survived
Count
Age