Titanic Exploratory Data Analysis

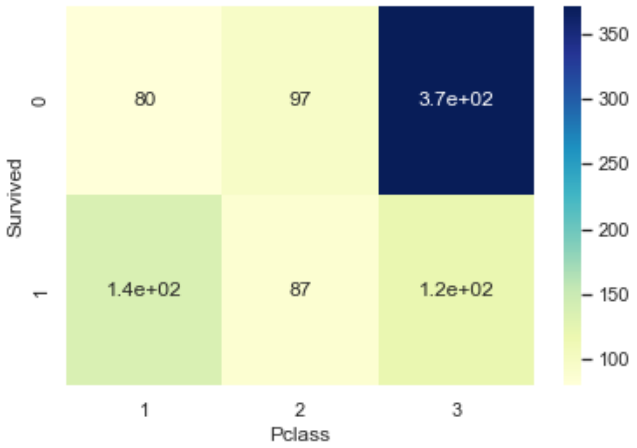
In this report, the Titanic dataset from Kaggle is explored and the following three hypotheses are tested:

- Determine if the survival rate is associated with the class of passenger
- Determine if the survival rate is associated with the gender
- Determine the survival rate is associated with the age

Survival of a passenger is specified in the variable Survived, with 0 for didn't survive and 1 for survived.

Passenger Class

To check if this variable is associated with the survival of the passengers, the interaction between the variables is visualized in a heatmap, and a chi-square test is performed. The class of the passenger is binned in 1 for first class, 2 for second class, and 3 for third class.



A quick glance at the table shows that there is a negative (or positive, depending on how you interpret the pclass value) relationship; most first-class passengers survived, while most lower-class passengers didn't.

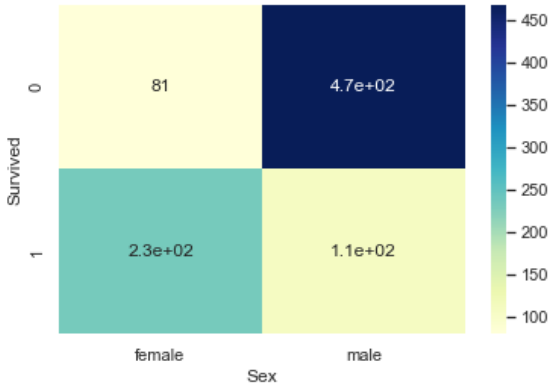
The chi-square test results in a coefficient of 102.9 and a p-value lower than the significance level of 0.05 (4.549251711298793e-23), corroborating the association between the variables.

Gender

For this variable, the same procedure is followed as with the Class variable, a visualization of the interactions followed by a chi-square test.

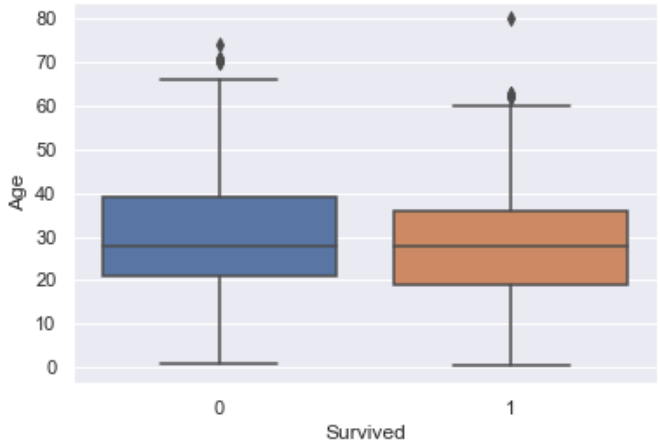
A similar result is obtained; The gender of a passenger is associated with the survival rate; three-quarters of female passengers survived, while only a fifth of male passengers did.

The chi-square tests further reaffirm the association, with a coefficient value of 260.71702016732104 and a p-value lower than the significance level (1.1973570627755645e-58)



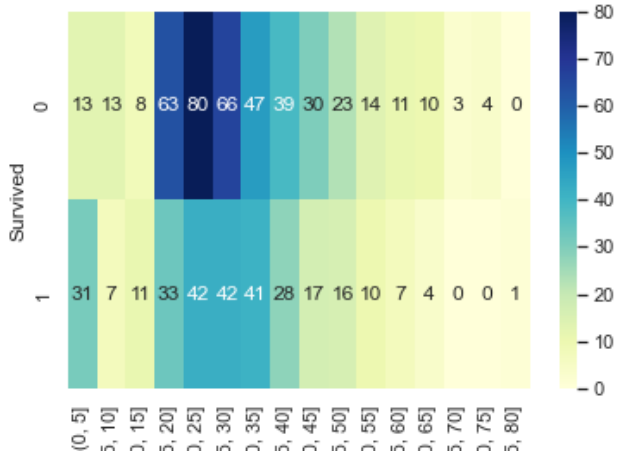
Age

For this variable, as we are comparing a categorical (Survived) with a discrete (Age) variable, a box plot is used as a first visualization to see if there is any association between them. Also, a chi-square test is performed.



The box plot does not show a clear association between the variables, as no matter the Age, passengers are not guaranteed to survive. The chi-square results also support the lack ofo association, with a p-value of 0.10.

But, a second association test is performed for this variable, this time, binning the values of Age in 5 years intervals. The results change drastically.



After binning the Age, we get somewhat different results. There are some age groups that see a higher or lower survival rate. For example, the group of 0 to 5 ages has a good survival rate (~70%) while the 20 to 25 group has a bad survival rate (~33%). But, this is not definitely a strong association.

The chi-square result this time supports the association, with a coefficient of 31.330998516525035 and a p-value (0.008) lower than the significance level, but various interactions have frequencies lower than 5, which tells us that the chi-square results might not be valid.