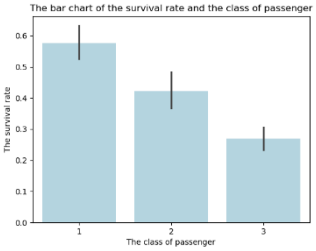
Exercise 2

Titanic Machine Learning Competition is collected from the sinking of Titanic which is known as an unsinkable ship in 1912. There are 1309 observations and 12 variables, and it divided by two categories of variables which are numerical variables (PassengerID, Survived, Pclass, Age, Sibsp, Parch, and Fare) and categorical variables (Name, Sex, Ticket, Cabin, and Embarked).

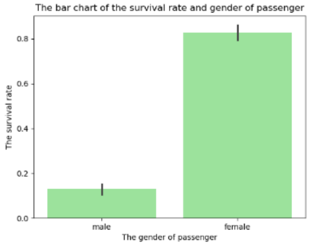
To summarize the datasets, 0 and 1 in the survived means the survival passengers, 0 is no and 1 is yes. Of course, the number of survivors is less than the number of shipwreck victims. The age of passengers is between 0 and 80 years, and the average is approximately 30 years old. Moreover, most titanic passengers traveled alone because they do not have any siblings, spouses, parents, and children aboard the titanic. Furthermore, the fares of passenger are very various from 0 to 512.329, and the means of the passenger fares is 33.295. In addition, I find that there are 263 missing values in age, and I impute them by using mean.

Hypothesis 1: The survival rate is associated to the class of passenger



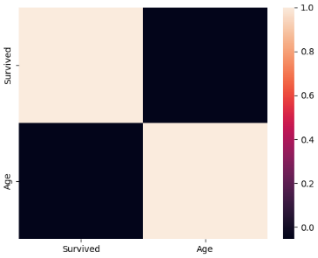
This bar chart shows the survival rate and three classes of passenger. The highest survival rate is the first class which is 0.576, followed by second class (0.422) and third class (0.269) respectively.

Hypothesis 2: The survival rate is associated to the gender



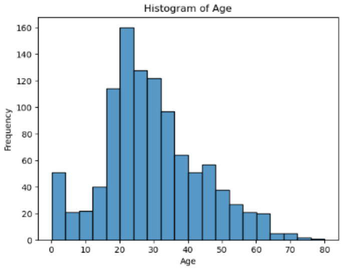
The bar chart of the survival rate and gender illustrates that the survival rate of female is 6 times greater than the survival rate of male from the Titanic shipwreck.

Hypothesis 3: The survival rate is associated to the age



The heat map provides the correlation between the survival rate and the age, which is -0.053695.

The histogram of age shows the right-skewed distribution because most data is on the left, with a long tail on the right.


Histogram of Age
Frequency
Age