AI Titanic project

Yuliia Nikolaenko
Yacine Belamiri
Charles MONOD-BROCA
Akshay Chaturvedi

Titanic: overview

891 passengers ('train' dataset)

38% survived

The majority of the passengers were in 3nd class (55%), 24% in 1st and 20% in 2nd

The mean of the passengers age is 30 years old

The majority of the passengers (72%) embarked at the Southampton, 19% at the Cherbourg, and just 1% at Queenstown
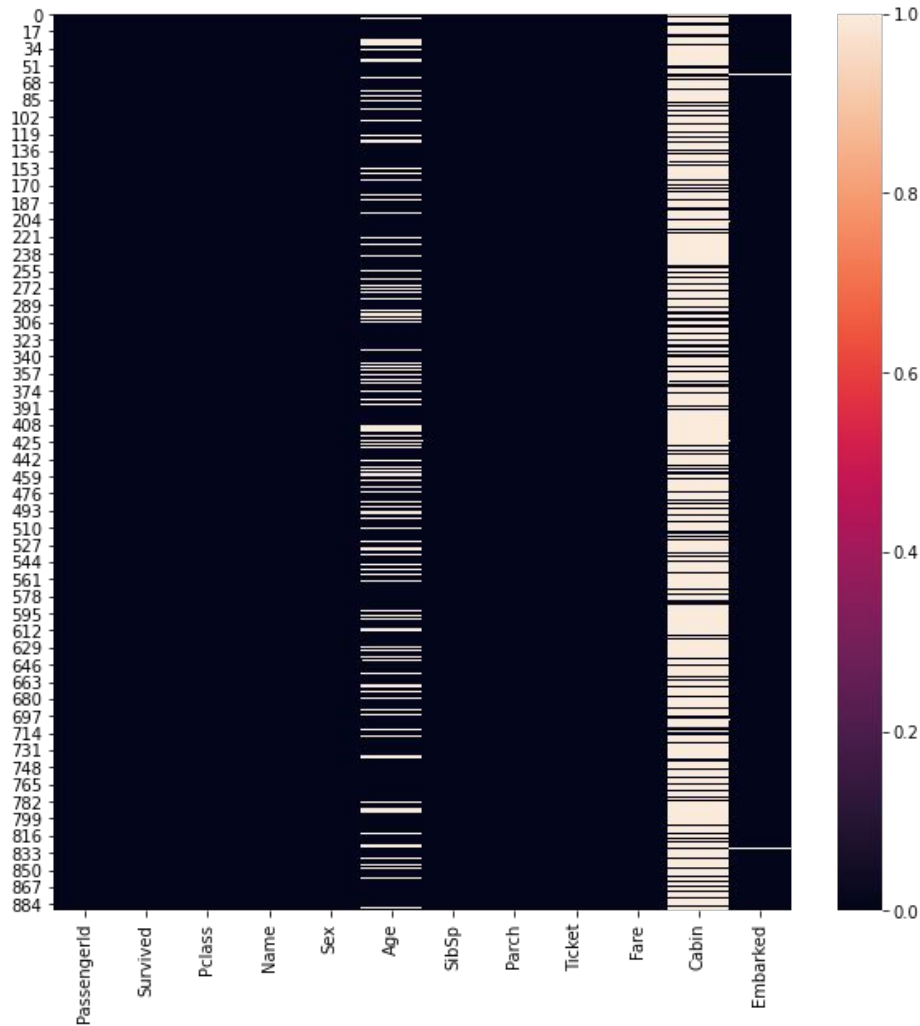
The 65% male and 35% female

The mean of the ticket fare is 32, the most expensive ticket cost 512, 15 free tickets
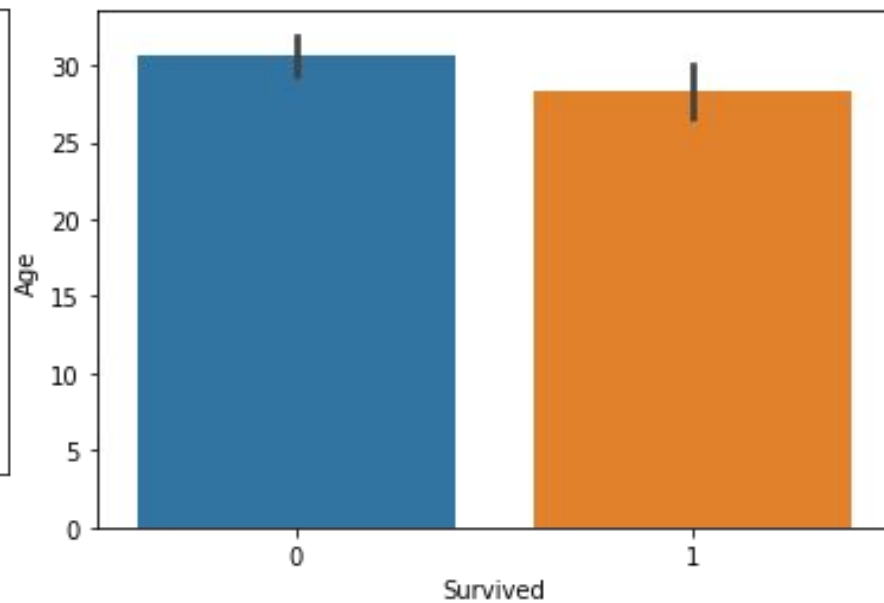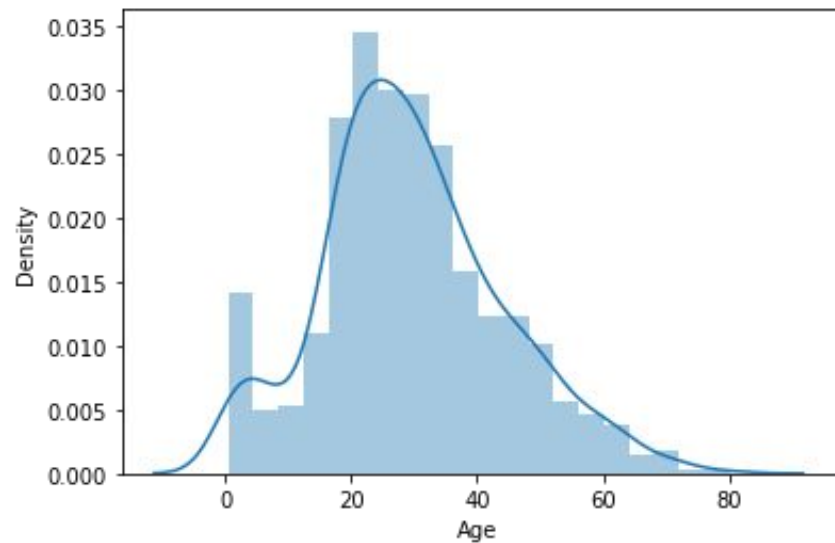
Missing values

Age - 20% (177) of missing values

Cabin - 77% (687) of missing values
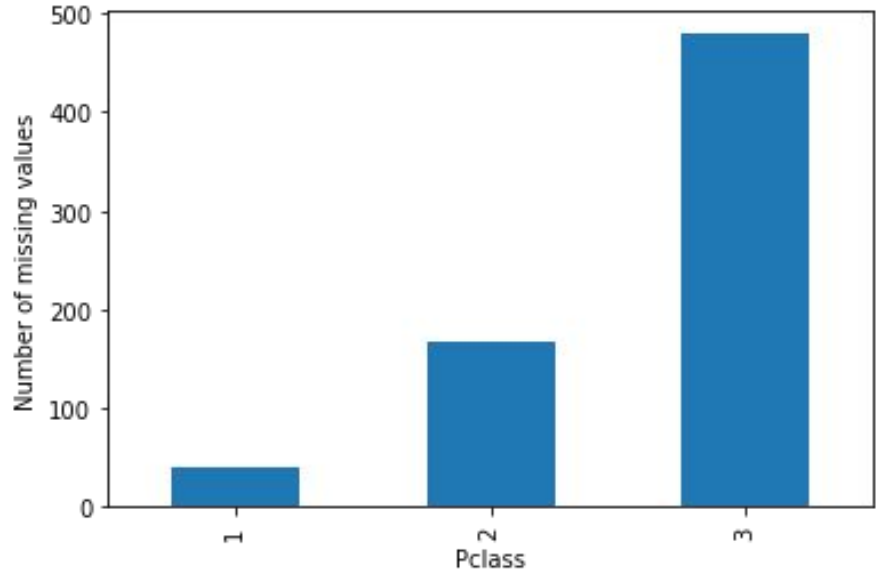
Embarked - 0.22% (2) of missing values

Missing values: age

Missing values: cabin and class

Hypothesis 1: the missing values could be connected with the passenger class. The passengers from the lower classes might not have fixed cabin number.
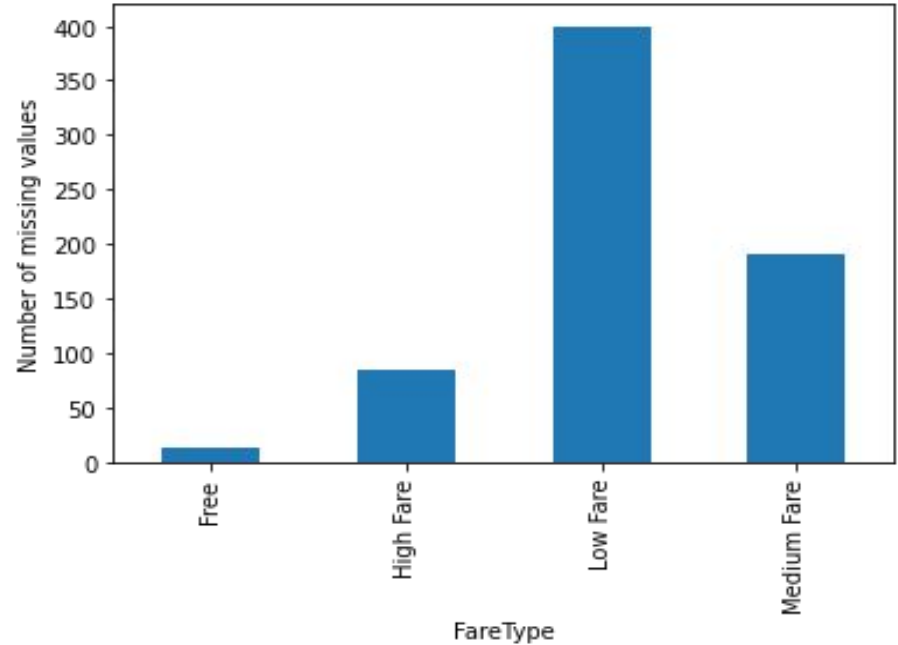
We can confirm Hypothesis 1 as we can see that 1st class passengers have less missing values
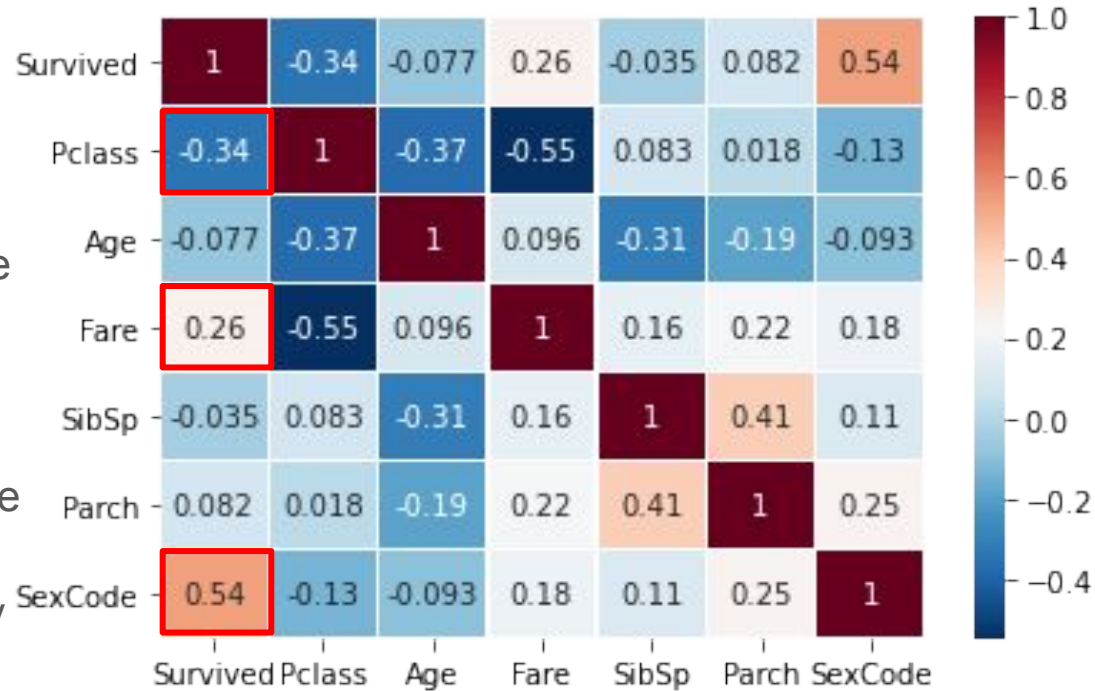
Missing values: cabin and fare

Hypothesis 2: the missing values could be connected with the ticket fare. The passengers with low ticket price might not have fixed cabin number.

We can confirm Hypothesis 2 as we can see that passengers with higher ticket price have less missing values
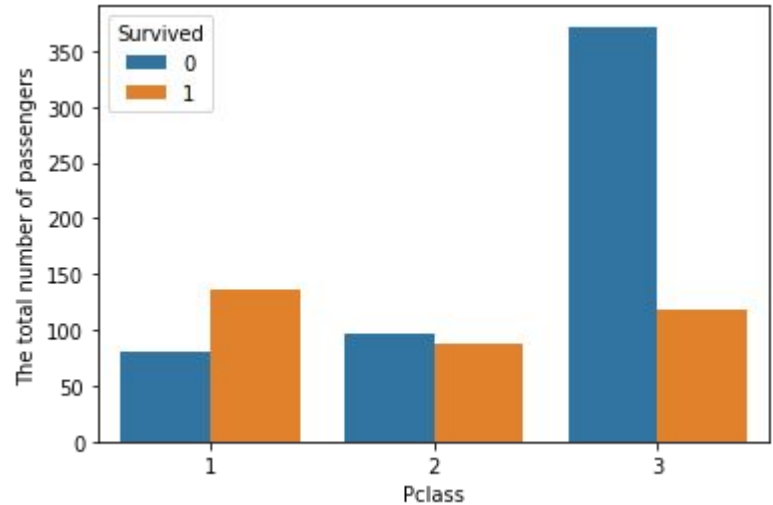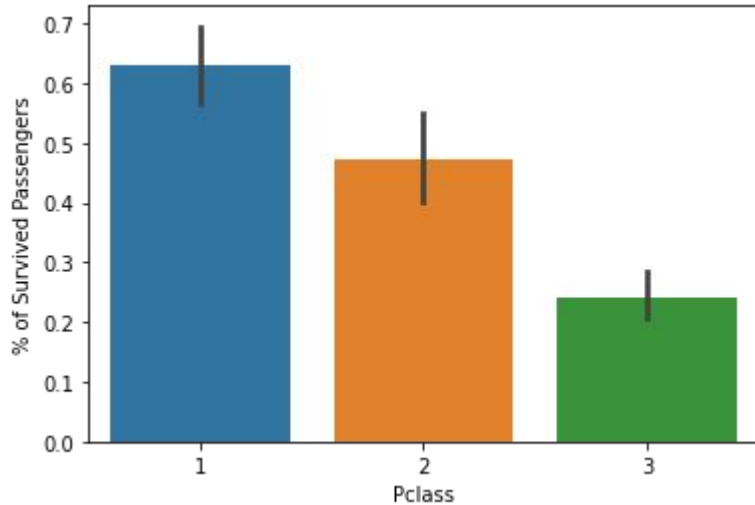
Correlation matrix

- The class of the passenger is associated with decrease (-0.338) of survival probability. Thus, the passengers of 1st had higher chance to survive compared to the 2nd and 3rd class passengers.
- The same mechanism for ticket fare (0.257), the higher the price the higher the probability to survive the crush.
- Women were twice (0.54) as likely to survive a disaster as men.
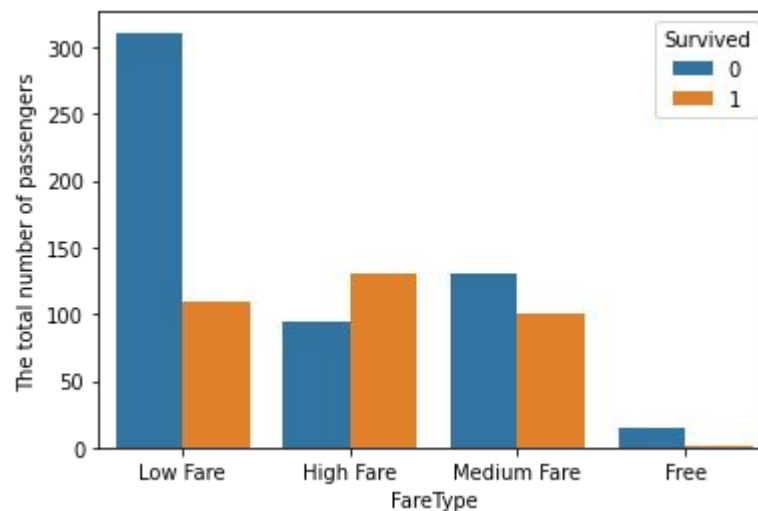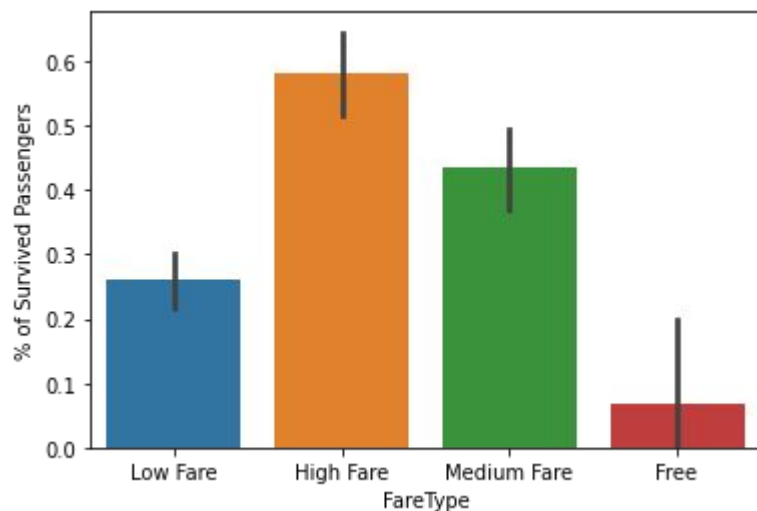
Visualization

The distribution of the survived passenger in each passenger class
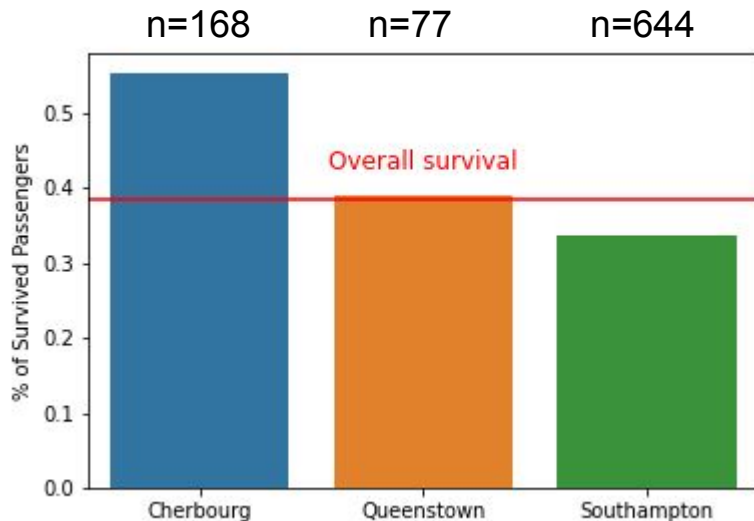
Visualization

The distribution of the survived passenger in each fare category.
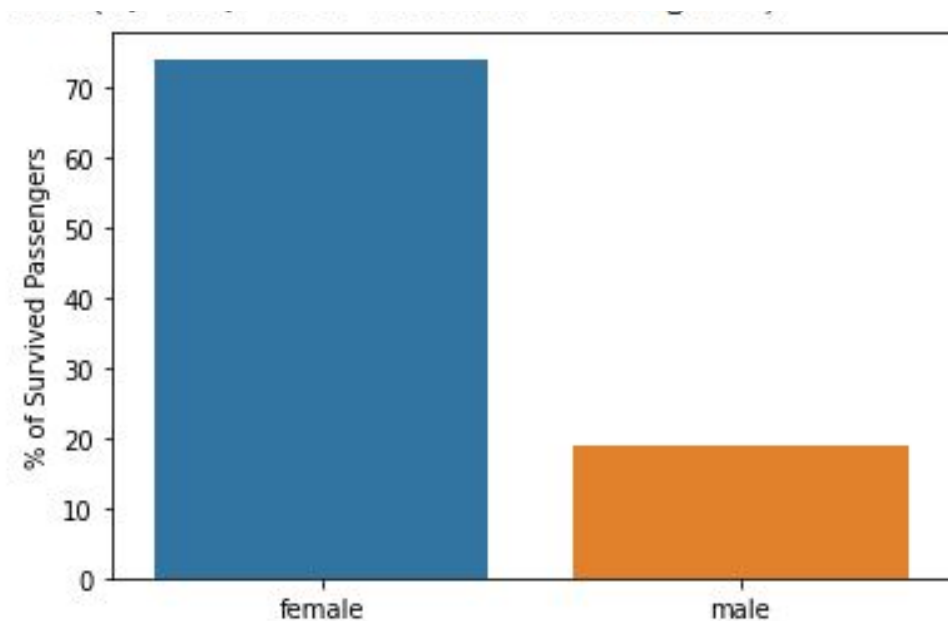
15 free tickets (=0)

Visualization

The percentage of survival depending on the port of embarkment and overall survival level
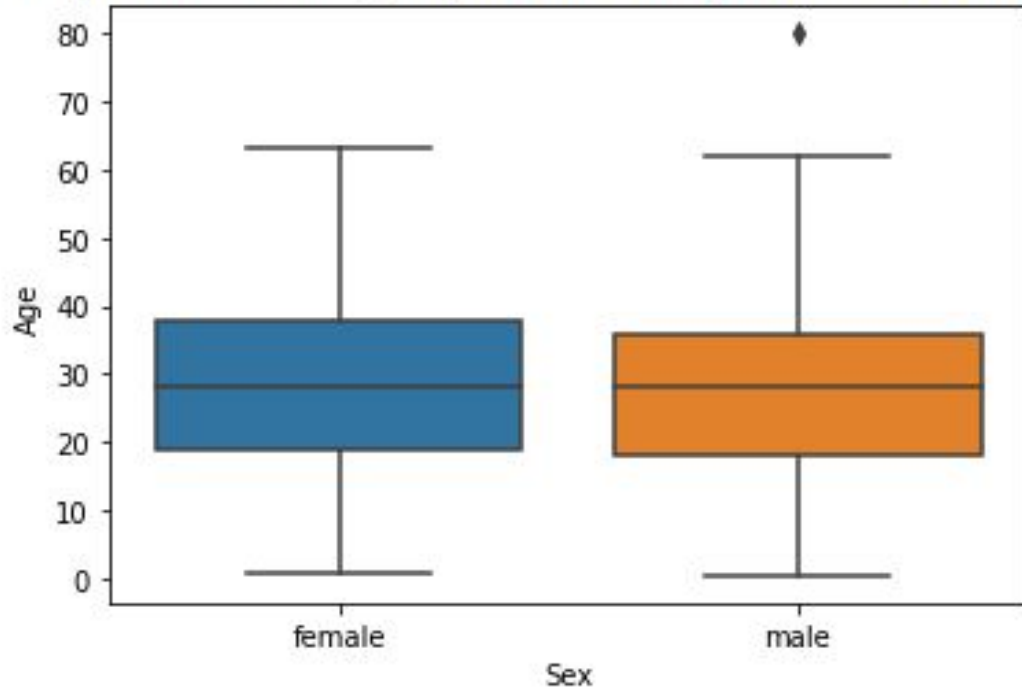
Visualization

The percentage of females and males that have survived on the overall number passengers on our dataset, t-test significant results (pvalue=1.4060661308802594e-69)
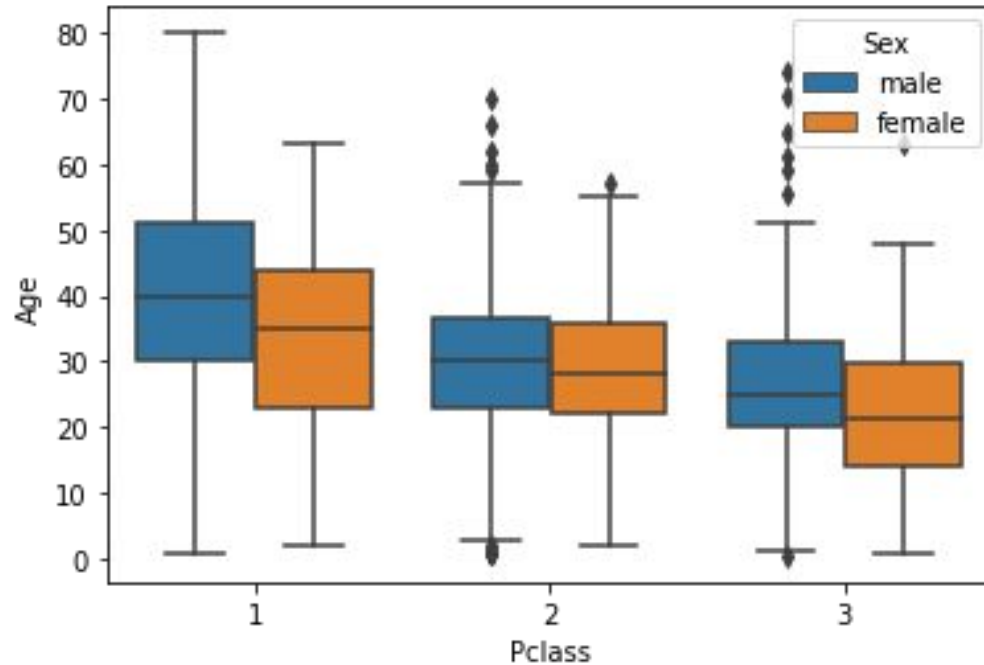
Visualization

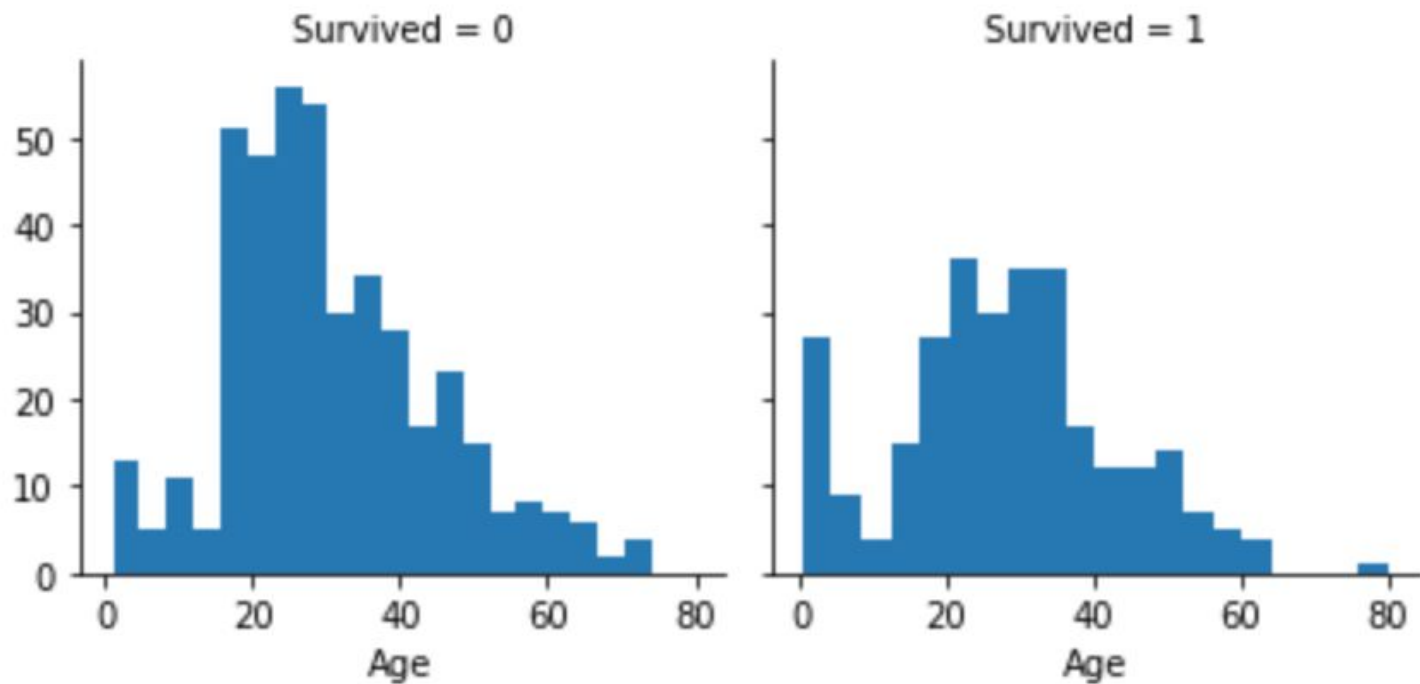The age distribution of male and female survivors

Visualization

Distribution of the male and female of different age according to the passenger class

Visualization

Visualization

Features: "Pclass","Age","SibSp",

"Parch","Fare","SexCode"

Target: "Survived"

2 components PCA