EDA: Titanic Data

Module: Data & IA

Presnted by:
SIDIBE Moussa
CULPIN Alexis

AGENDA

01
Description

02
Miss value treatment

03
Correlation Analysis

04
Relate graphics

05
Conclusion

Variables

VARIABLE QUANTITATIVES

- PassengerId
- Survived
- Pclass
- Age
- SibSp
- Parch
- Fare
- Ticket

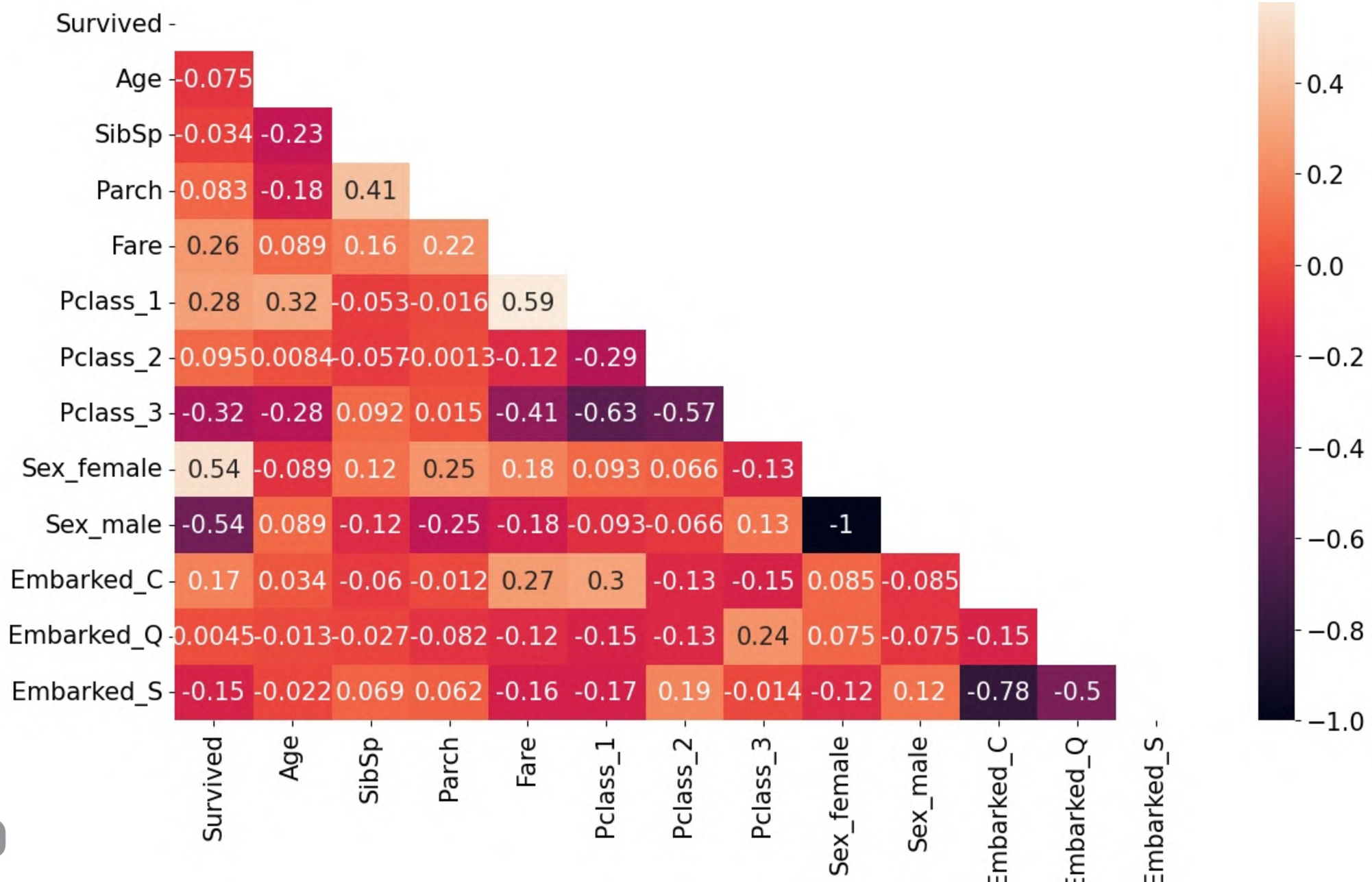
VARIABLE QUALITATIVES

- Name
- Embarked
- Sex

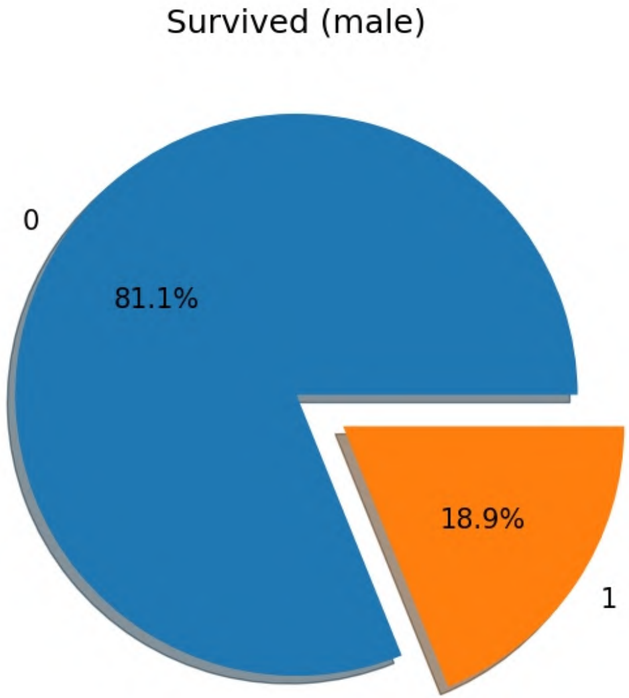
Missing Values

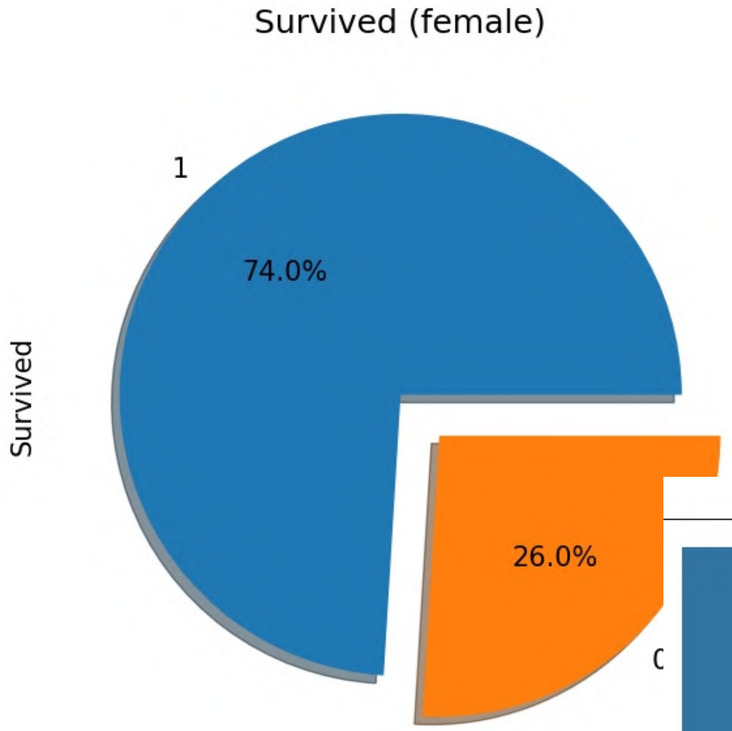
- Age : contains 177 NaNs → Filled by mean of Age
- Cabin : Contains 687 NaNs → Columns dropped
- Embarked: Contains just 2 NaNs → Lines dropped

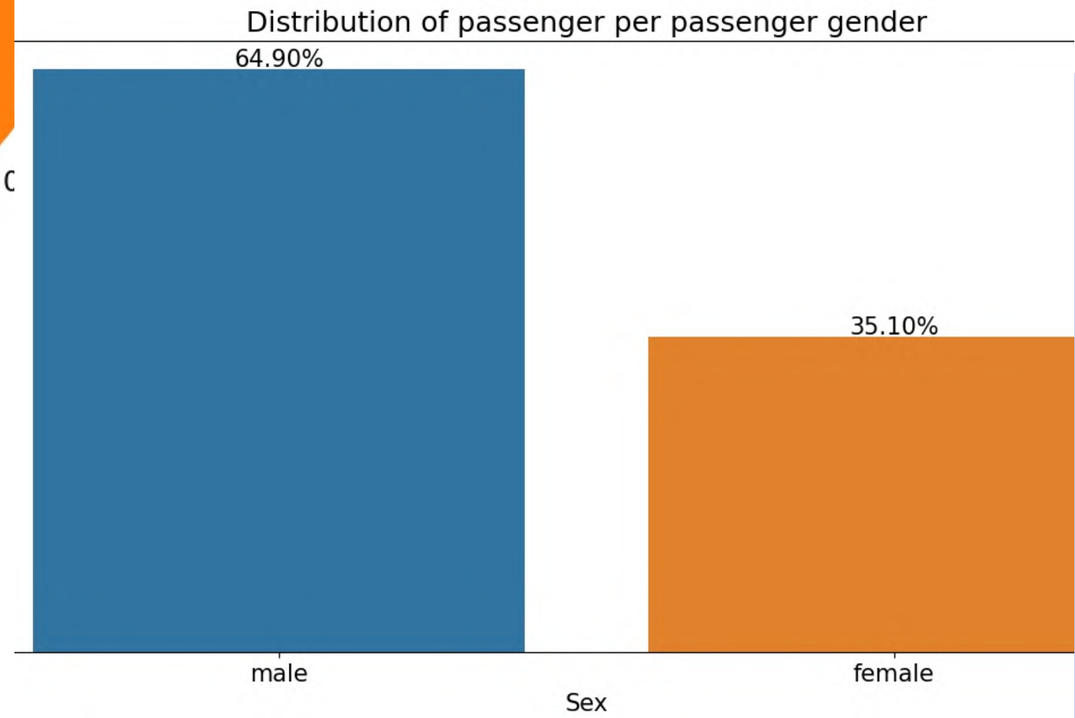
Correlation analysis

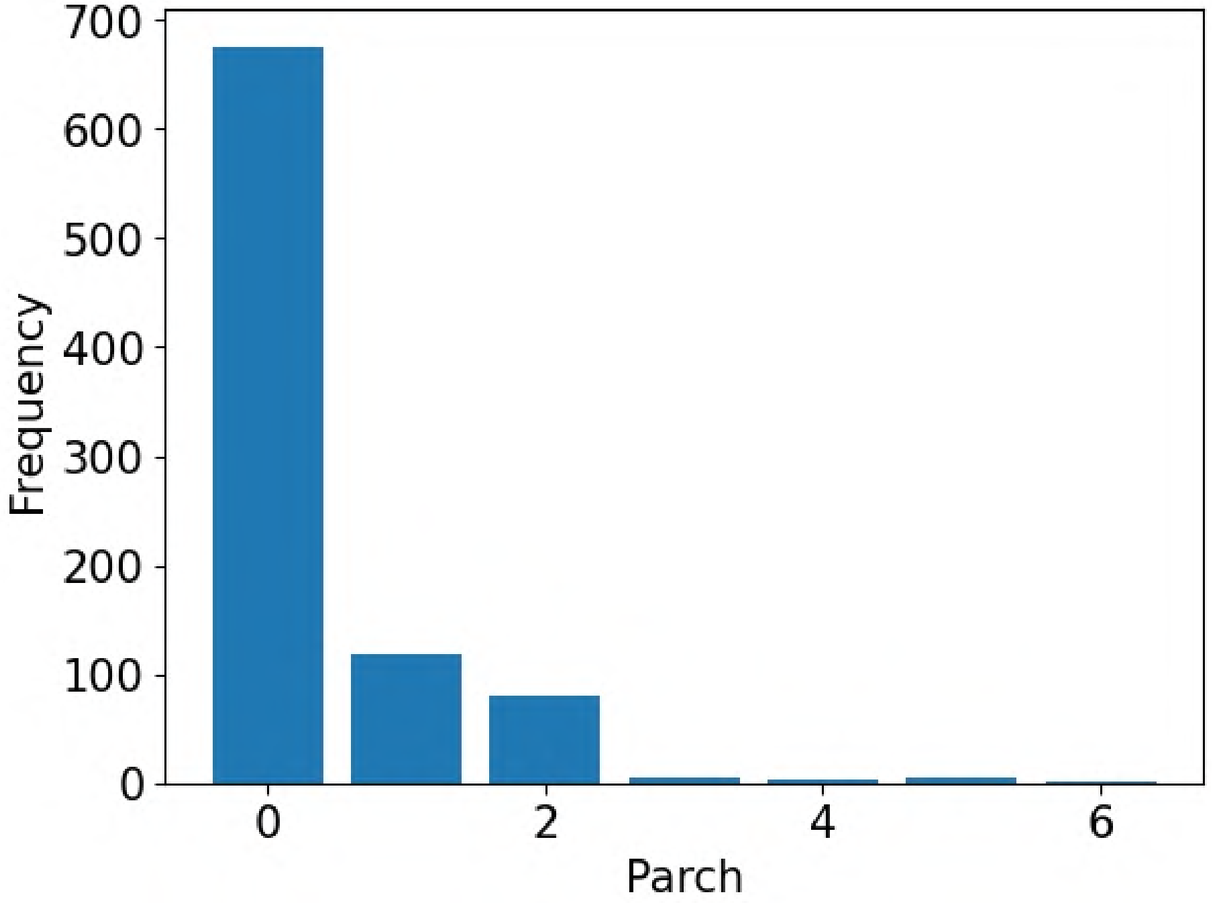


Some relates graphs

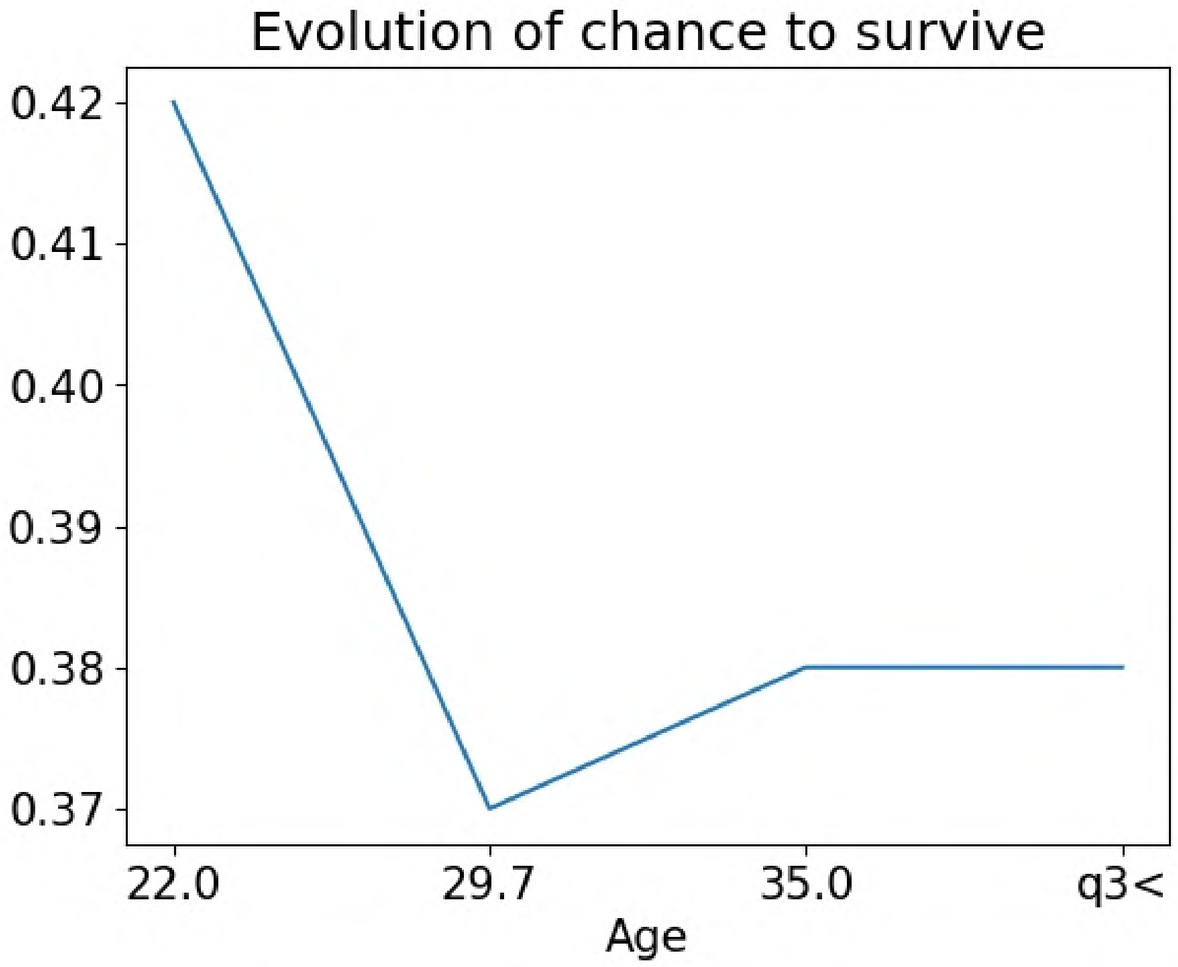


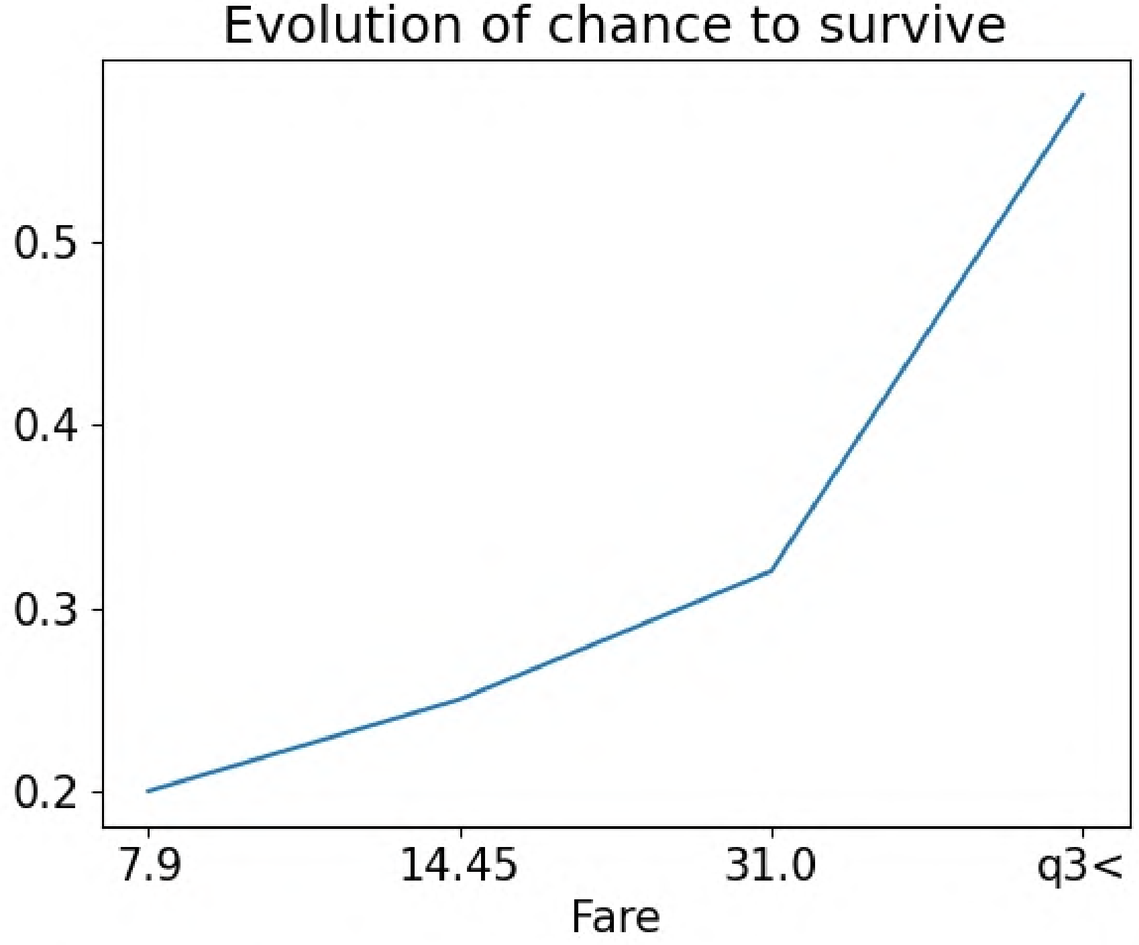


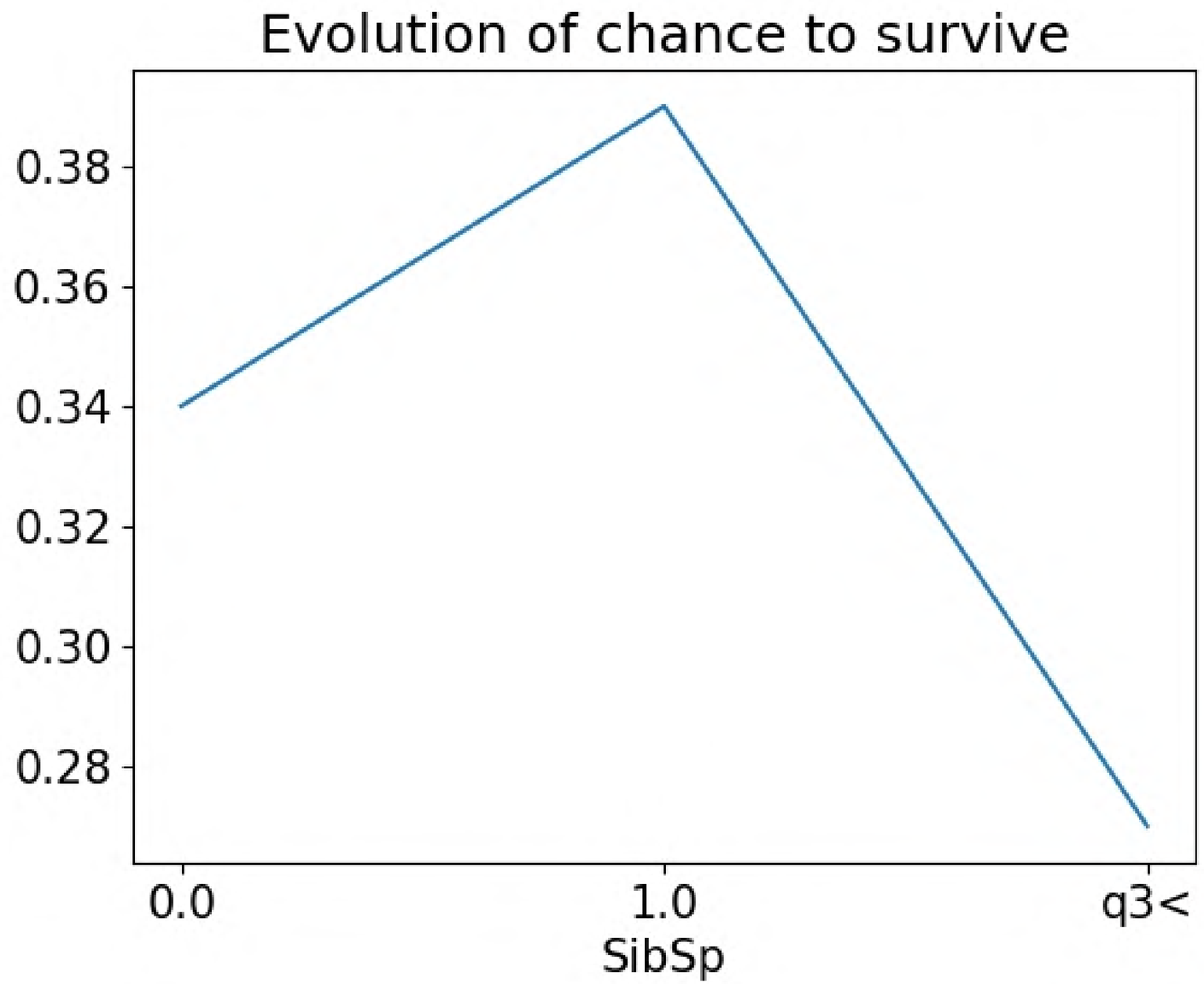



700
600
500
400
300
200
100
0
Frequency
0
2
4
6
Parch


Evolution of chance to survive
0.500
0.475
0.450
0.425
0.400
0.375
0.350
0.0
q3<
Parch


Evolution of chance to survive
0.42
0.41
0.40
0.39
0.38
0.37
22.0
29.7
35.0
q3<
Age


Evolution of chance to survive
0.5
0.4
0.3
0.2
7.9
14.45
31.0
q3<
Fare


Evolution of chance to survive
0.38
0.36
0.34
0.32
0.30
0.28
0.0
1.0
q3<
SibSp

Summary

This correlation matrix details more the chance of survive of passenger based on the allowed attributes. First of all, female were lucky, they have a high probability to survive than male in the accident. After gender lucky, coming class that the passenger was sat, more you get top class more your chance of survive is high, specifically for those in the first class. As we know that fare and pclass are positively associated, more richer or you pay more you have more chance to survive. Place of embarkation also influences the chance of survive, as we can read on the above matrix, passenger of Cherbourg get more chance than others.

Otherwise according to the correlation analysis, more you get older, more you have siblings on the board your chance of dying growth with. For places, peoples in class 3, whom from southampton was quite defavrorised for saving. Finally males were the most impacted that can be explained by many factors like priority when saving.

This analysis can be biaised by overreprensation for some modalities like sex, pclass 3

To go more deeper in our analyse, by this pivot table we can clearly notice that in average, `Fare` and shance of surviving strongly associated. Among female or male, for those who survived more than double of fare which ones who died. The last thing is that females paid more in average than males that can explain what we notice above but we have to check if it's not due to biais sample.

Survived	Sex	Age	Fare	Parch	SibSp
0	female	26.023272	23.024385	1.037037	1.209877
	male	31.175224	21.960993	0.207265	0.440171
1	female	28.797265	51.695617	0.519481	0.519481
	male	27.631705	40.821484	0.357798	0.385321

LEARNING PLANET INSTITUTE

Let's solve the challenges of our times

Thank you

As always, we're open for questions