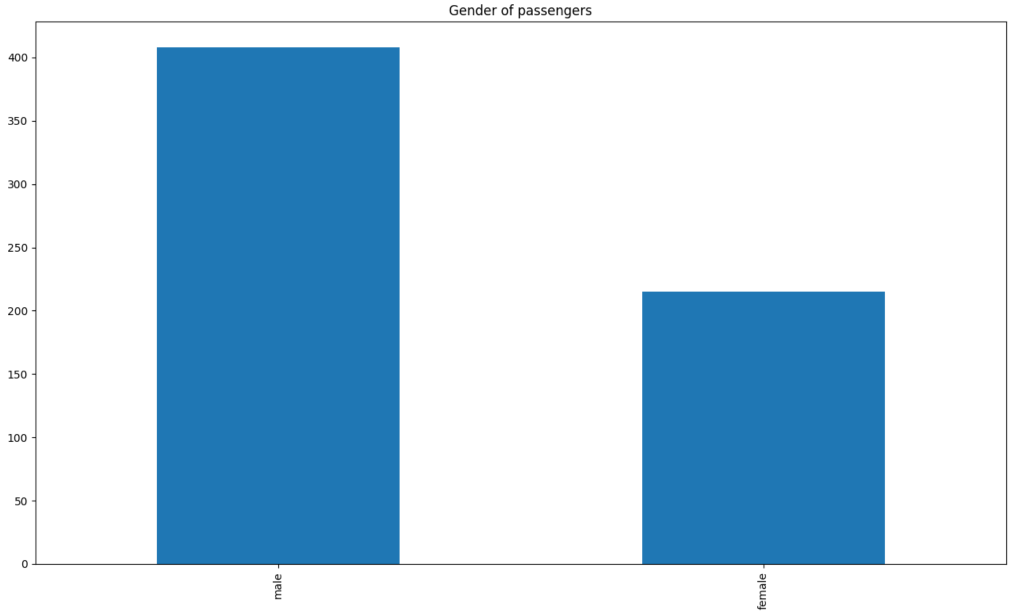
Titanic survival project - Exploratory Data Analysis

The DataFrame of the training data contains the columns:

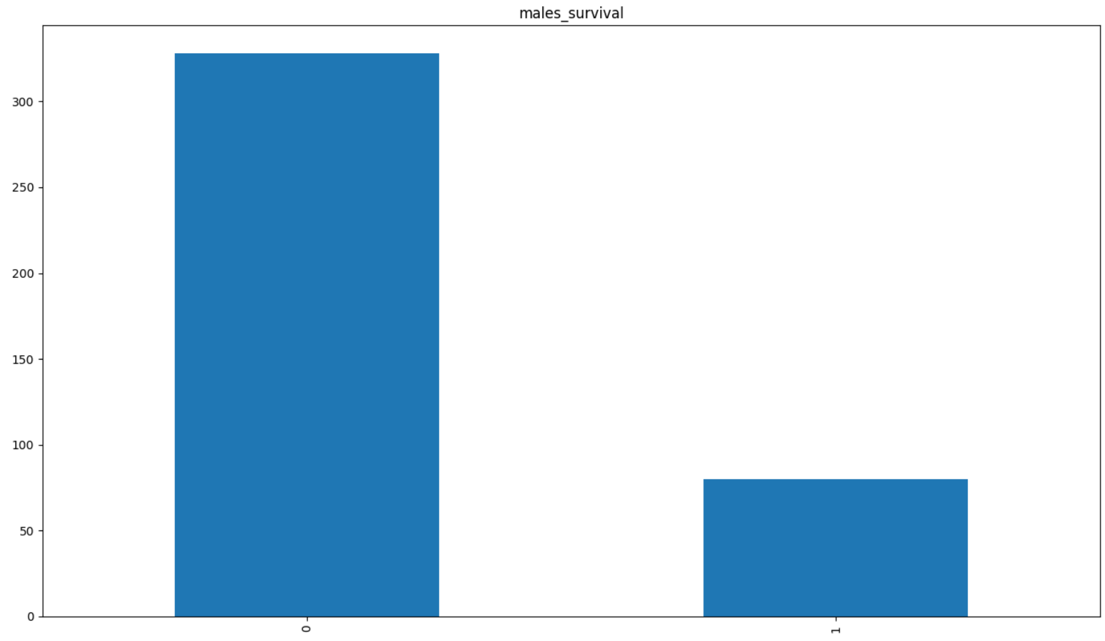
'PassengerId'- unique id number for each passenger in the data set.
'Survived'- 0 = No, 1 = Yes.
'Pclass'- 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
'Name'- Name of the passenger.
'Sex'- ‘male’ or ‘female’.
'Age'- Age in years, float.
'SibSp'- number of siblings/spouses of the passenger aboard the Titanic.
'Parch'- number of parents/children of the passenger aboard the Titanic.
'Ticket'- Ticket of the passenger, a string of letters and numbers.
'Fare'- Passenger fare.
'Cabin'- The cabin in which the passenger travels.
'Embarked'- Port of Embarkation. C = Cherbourg, Q = Queenstown, S = Southampton.

Data exploration

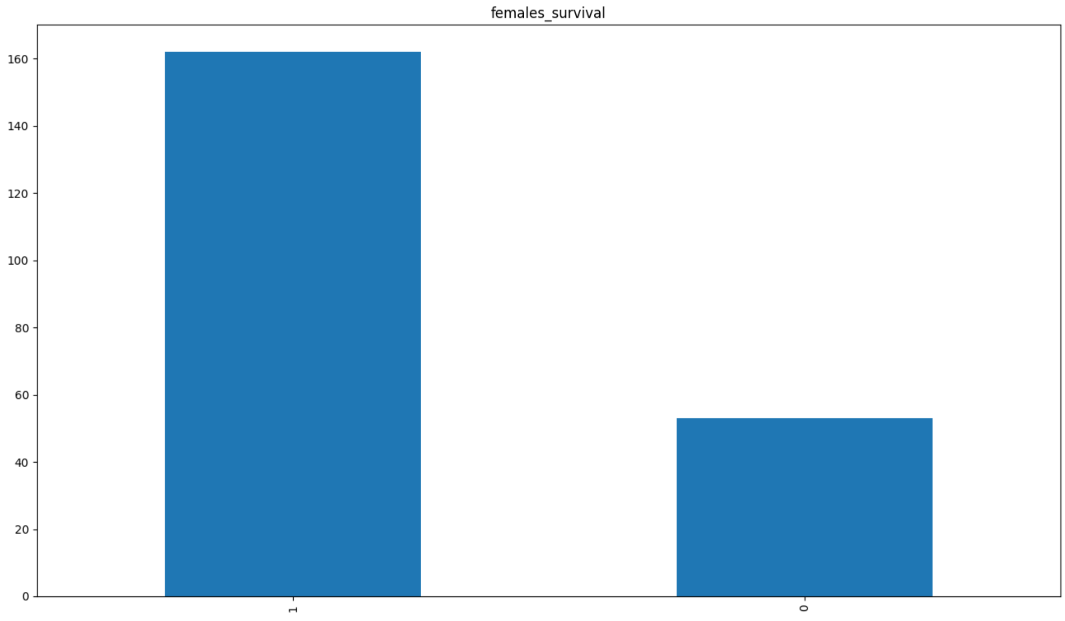
Made only using 0.7 of the data, to keep 0.3 of the data untouched.



As seen in the plot there are approximately 400 male passengers and approximately 225 female passengers in the training data. The majority of passengers are males.

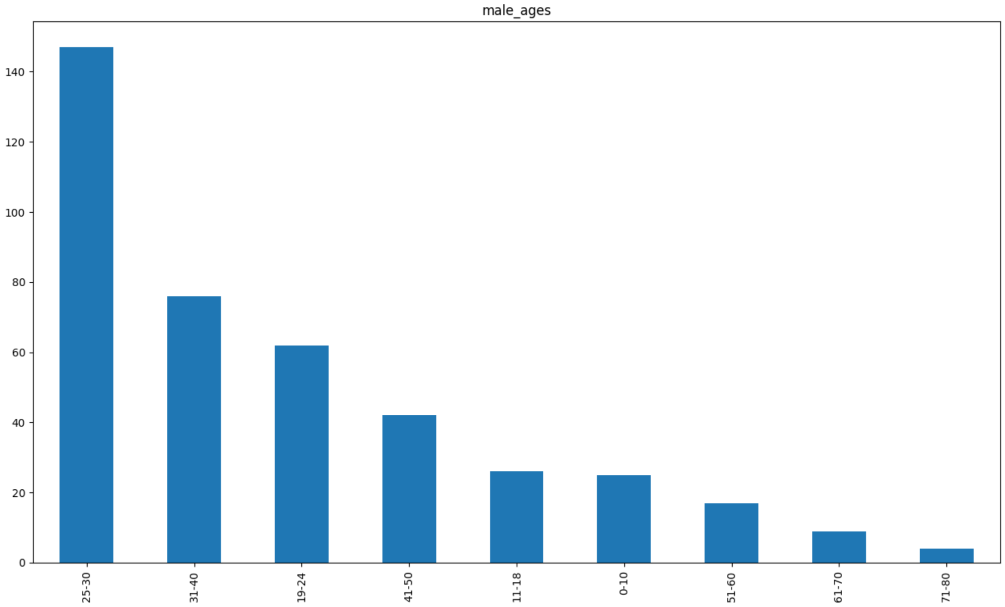


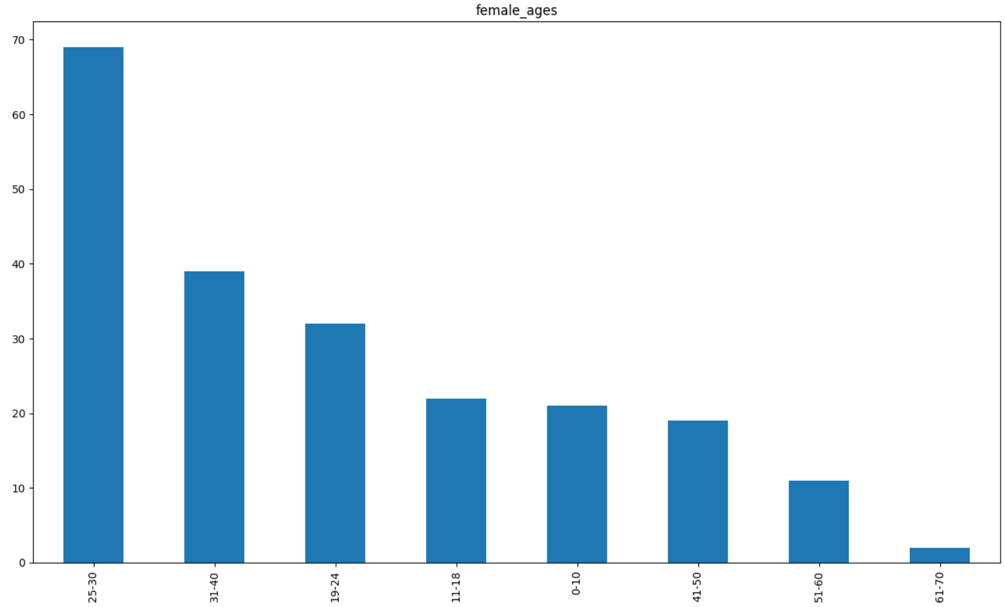
As seen in the plot approximately 325 of the male passengers survived and 75 didn't.
The survival percentage among males is close to 19.



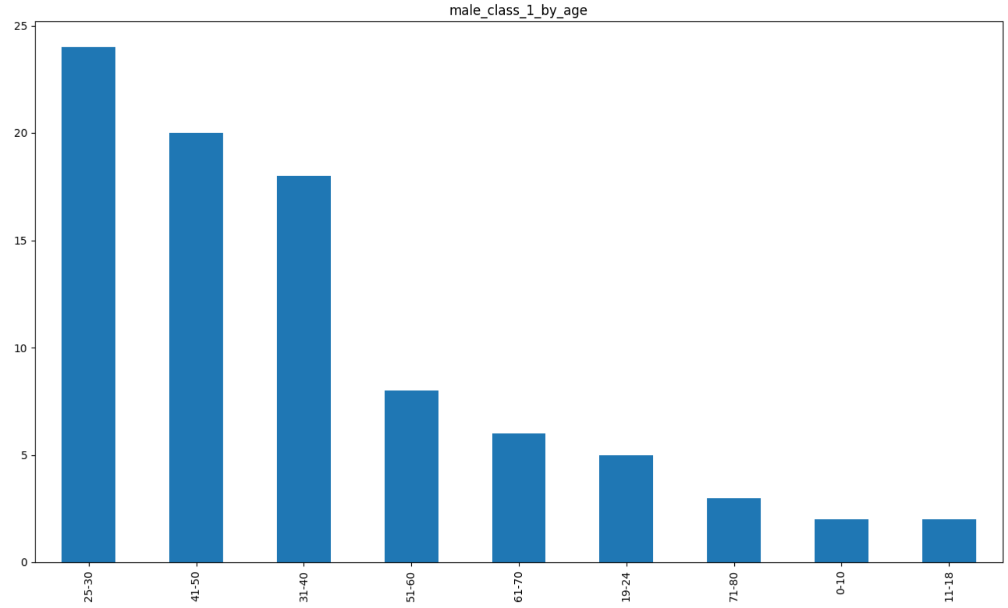
As seen in the plot approximately 160 of the female passengers survived and 50 didn't.

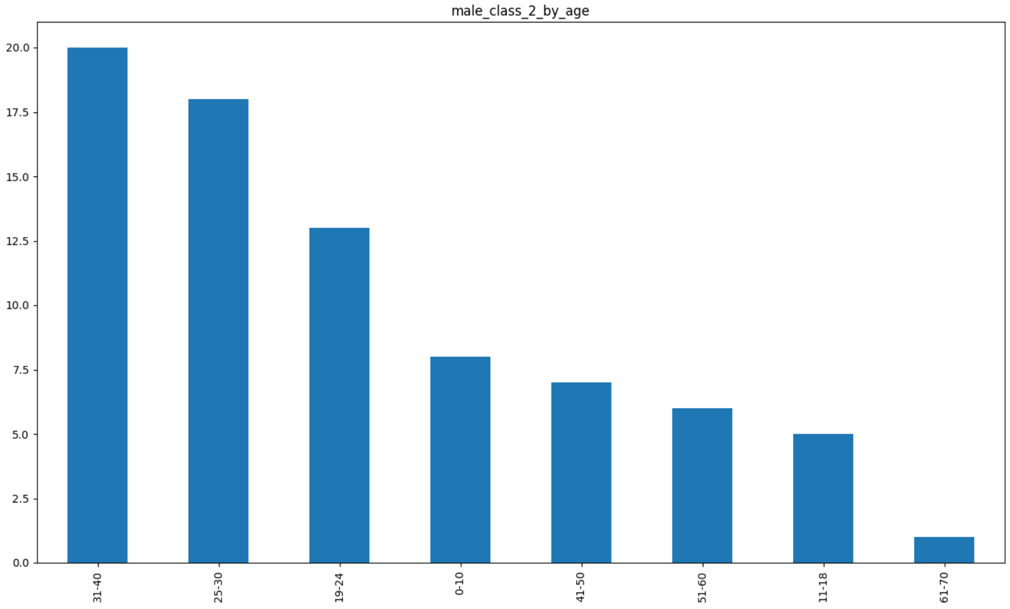
The survival percentage among females is close to 75.

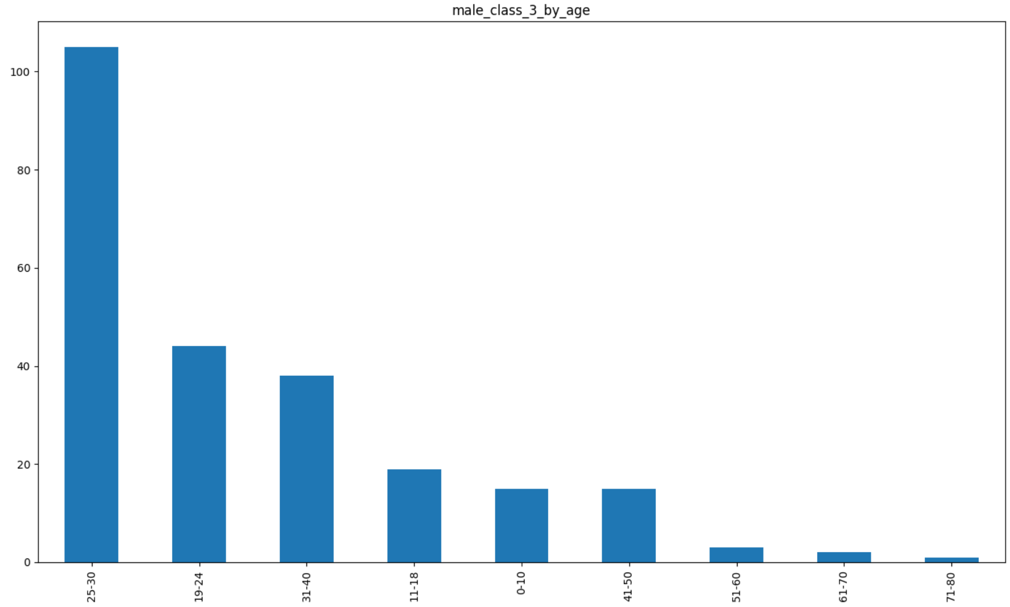


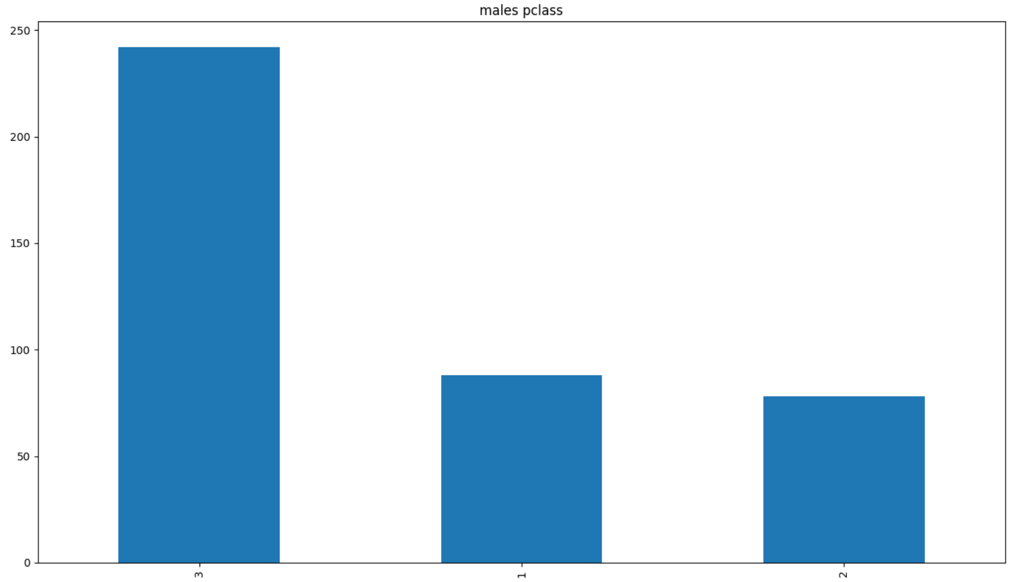


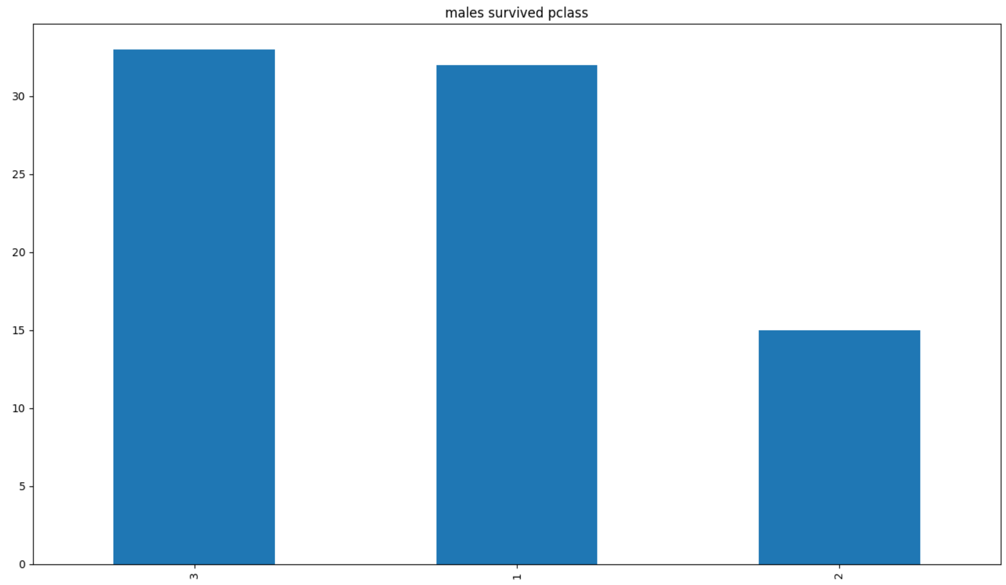
As seen in these plots, the majority of both men and women are between the ages of 25-30. Notice that among the female passengers there are no women between the ages of 71-80, in contrast to the male passengers.

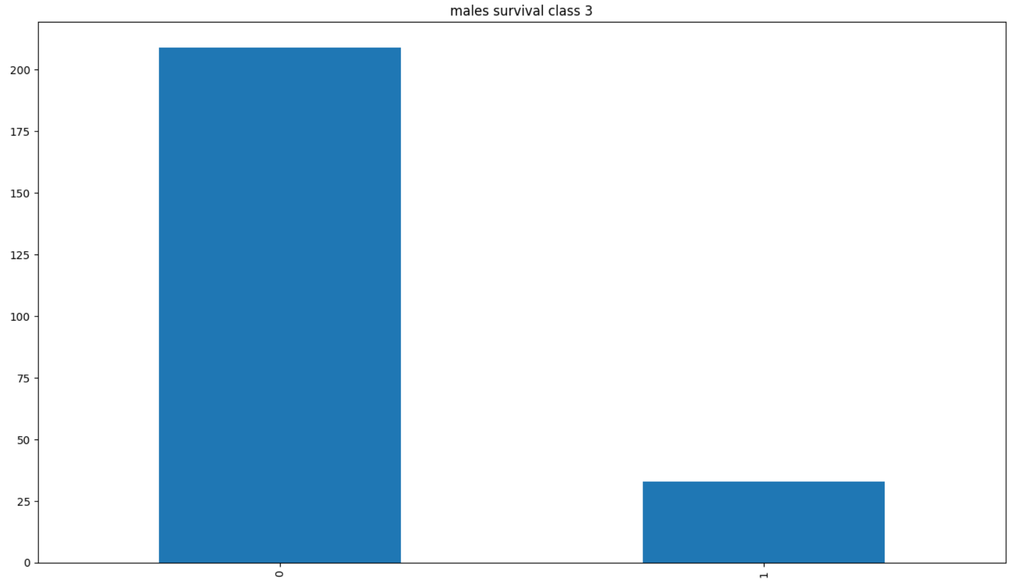

male_class_1_by_age
25
20
15
10
5
0
25-30
41-50
31-40
51-60
61-70
19-24
71-80
0-10
11-18


male_class_2_by_age
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
31-40
25-30
19-24
0-10
41-50
51-60
11-18
61-70

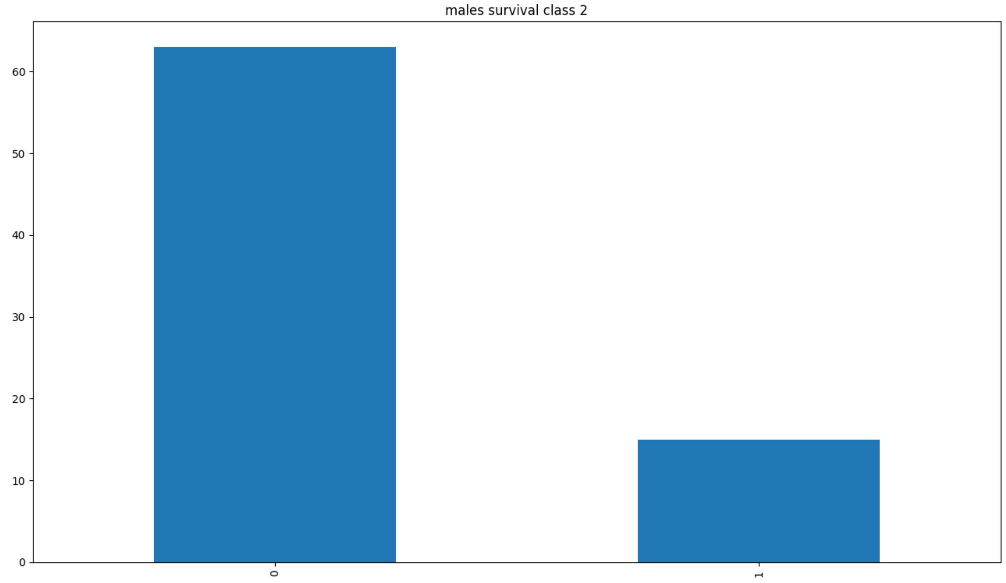

male_class_3_by_age
100
80
60
40
20
0
25-30
19-24
31-40
11-18
0-10
41-50
51-60
61-70
71-80

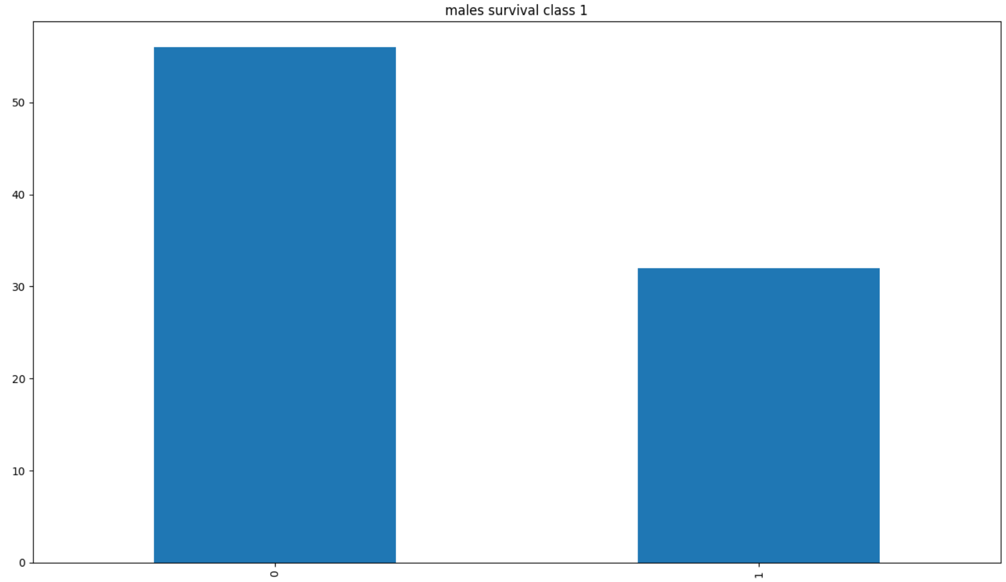

males pclass
250
200
150
100
50
0
3
1
2



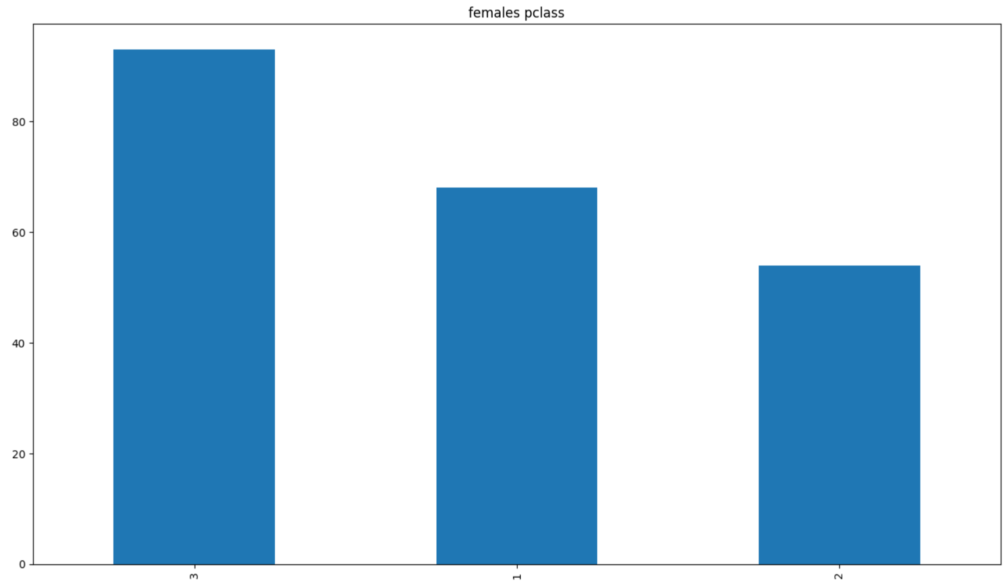


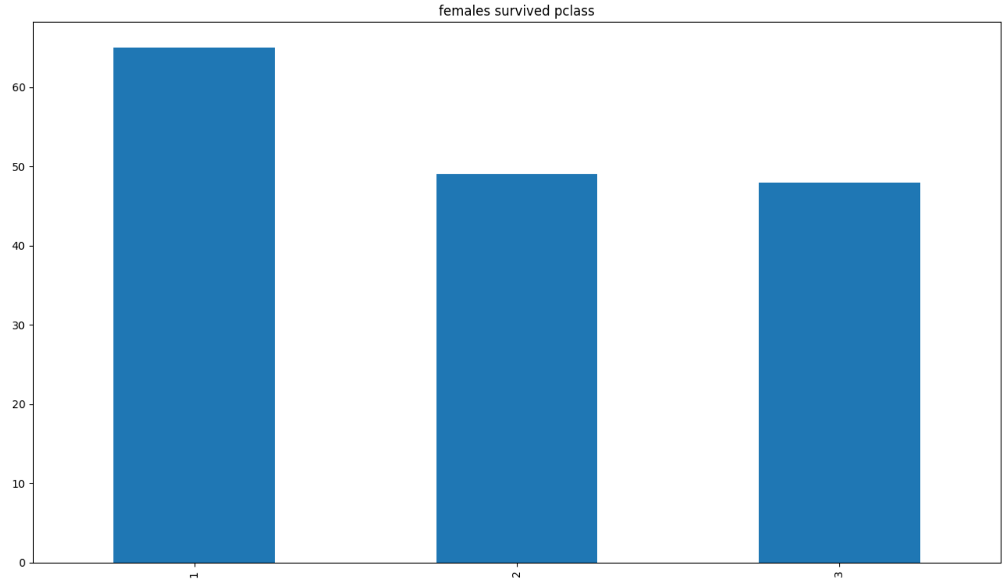
In conclusion, the majority of male passengers were in class 3. Most of the male survivors are in class 3, but most of the males in class 3 didn't survive. The survival percentage in this class is 13%.



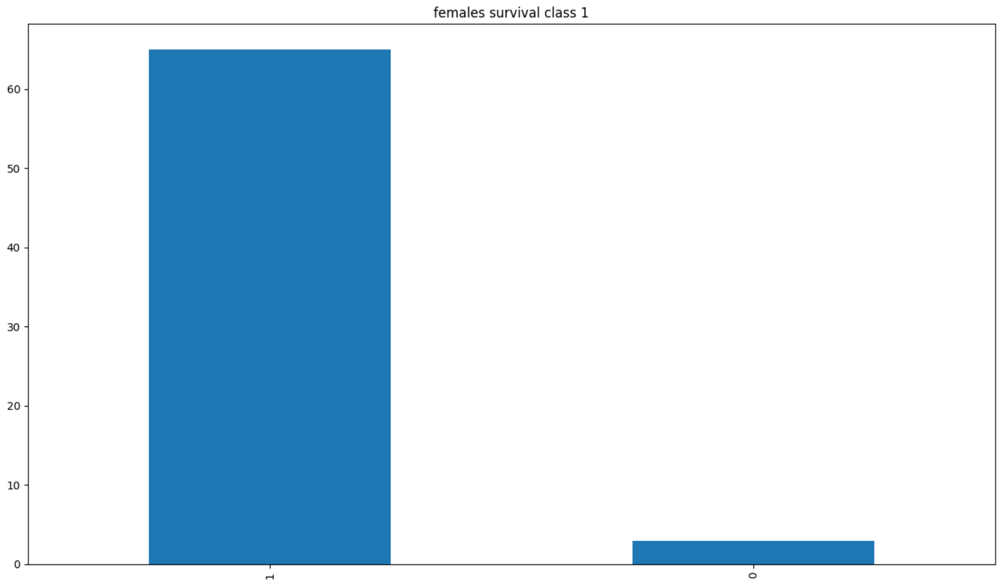


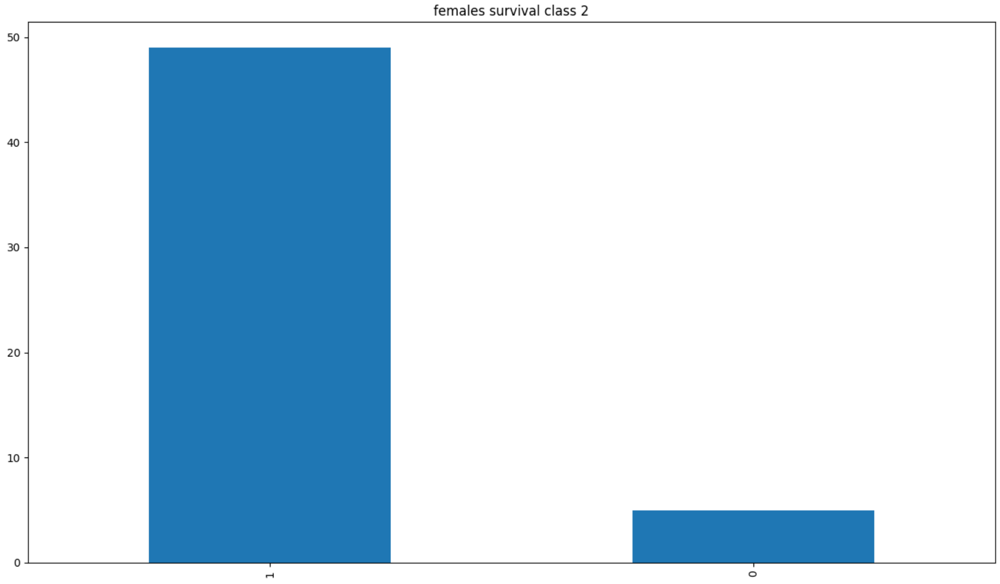
The survival percentage in class 2 is 19% and in class 1 is 36%.

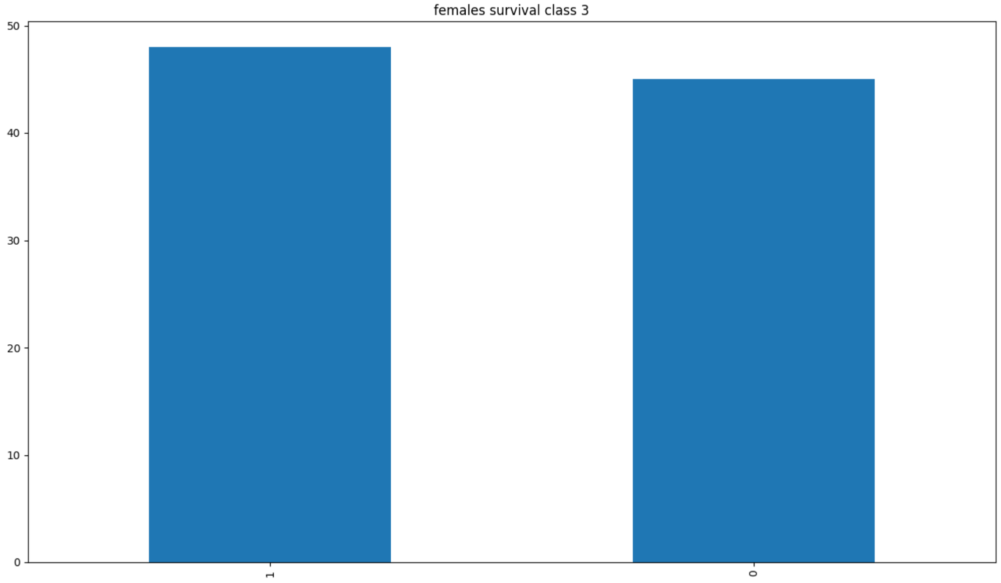




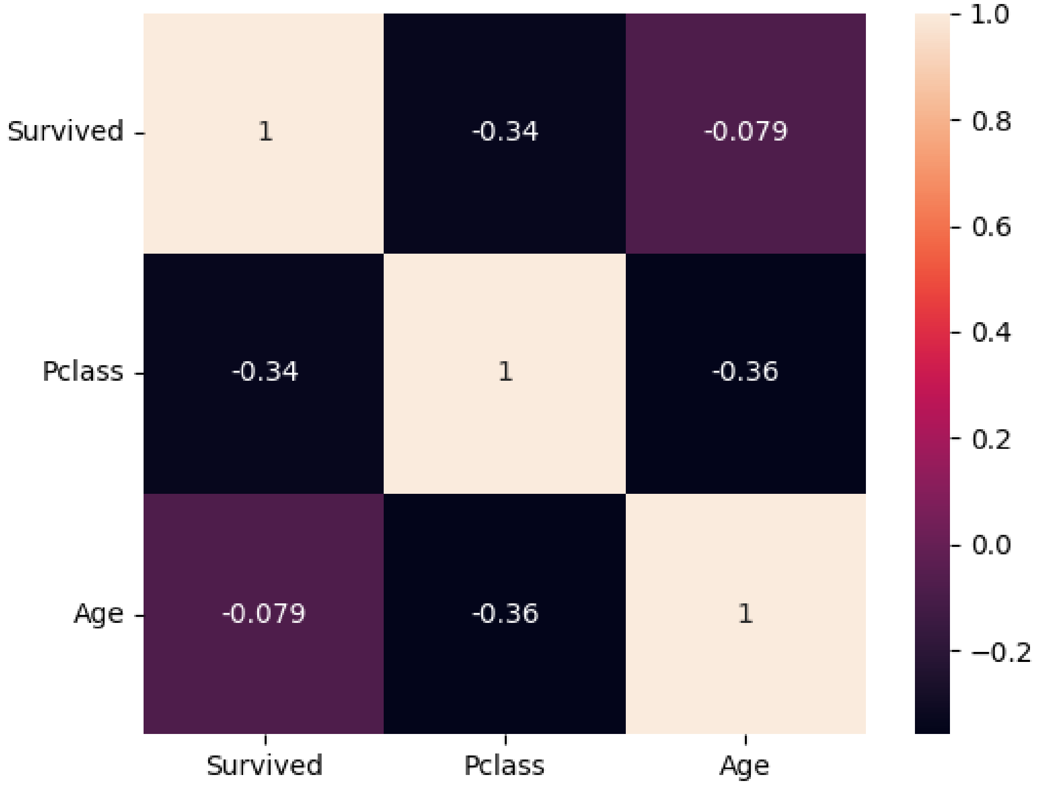
As seen, most of the female passengers are in class 3, but the majority of survivors are in class 1. This is different from what we saw for male passengers.


females survival class 1
0
10
20
30
40
50
60
1
0

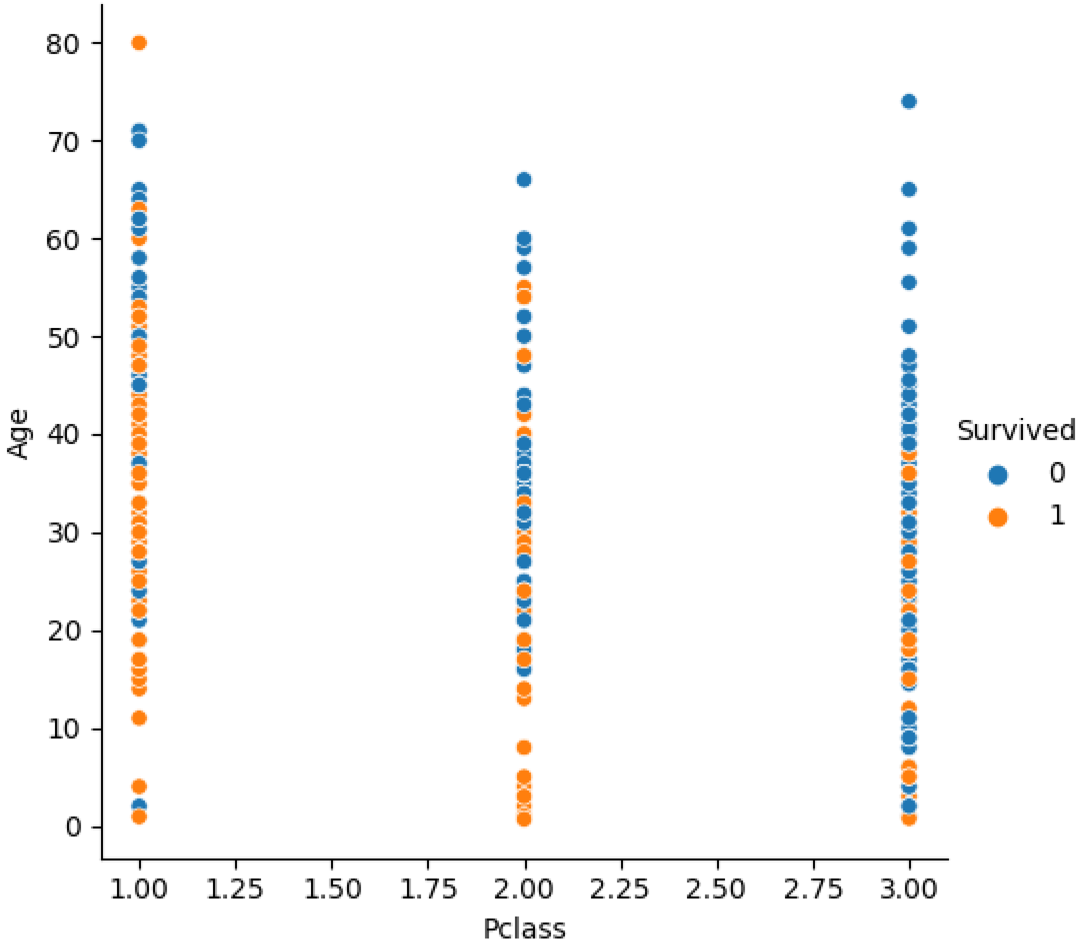

females survival class 2
0
10
20
30
40
50
1
0



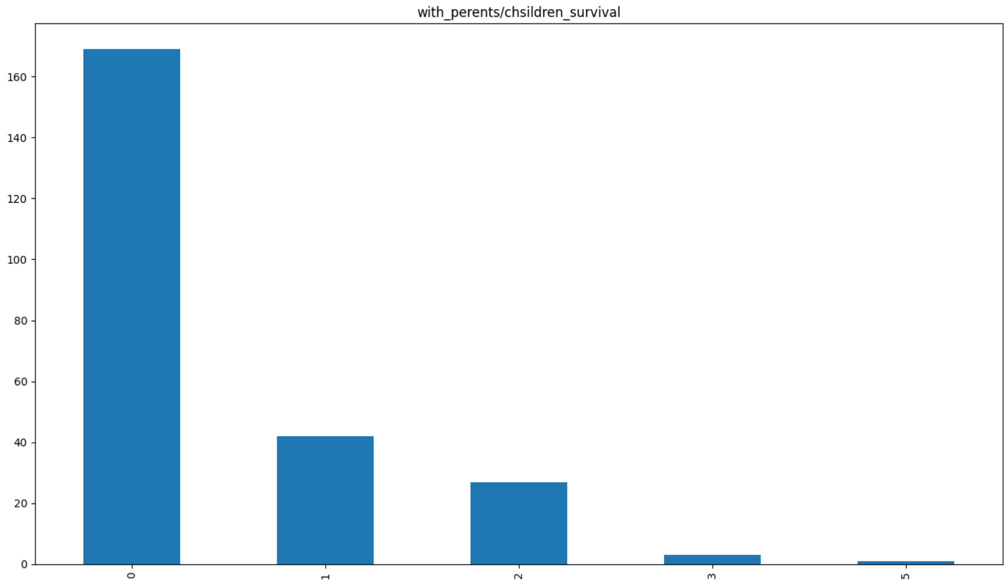
The survival percentage of females in class 1 is 95%, in class 2 90%, and in class 3 is 51%.



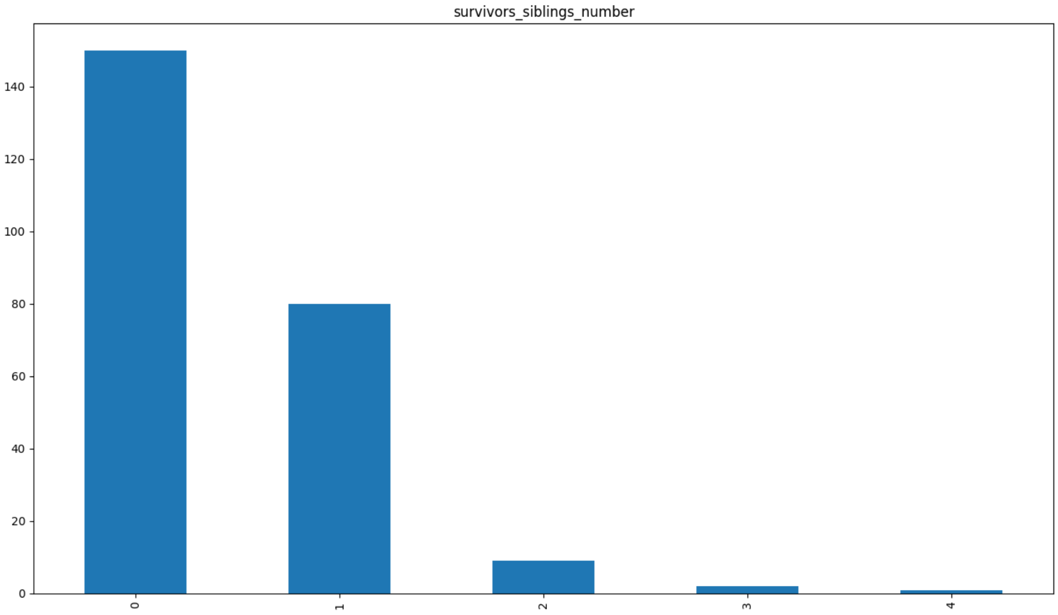
The correlation between survival and class is -0.34, The correlation between survival and age is more weak and also negative. As The correlation is negative between every two variables it holds that as the value of one variable increases, the other decreases. This matches what we saw in the previous plots.



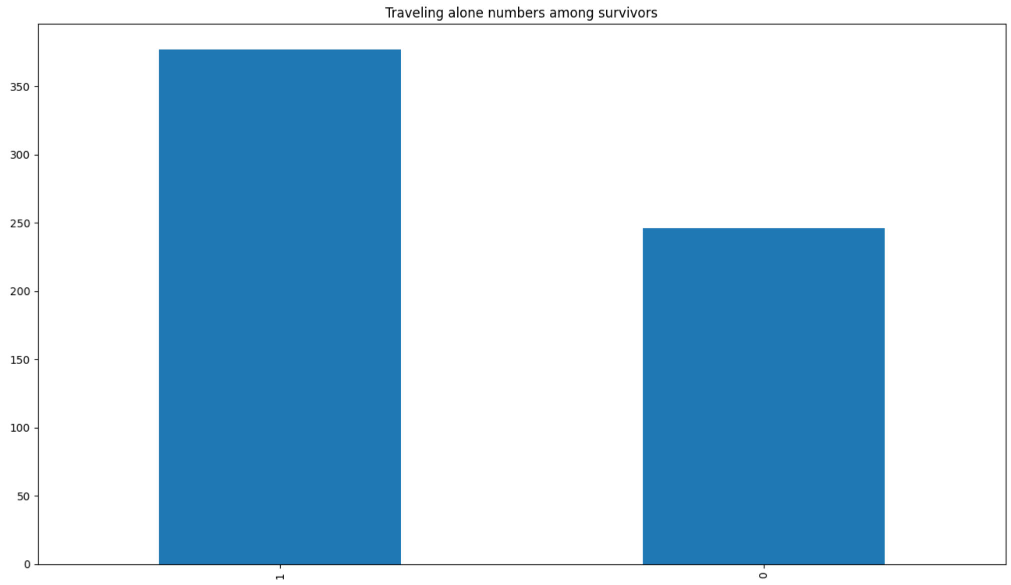
This plot shows that the higher the class, The lower the age of the oldest survivors.



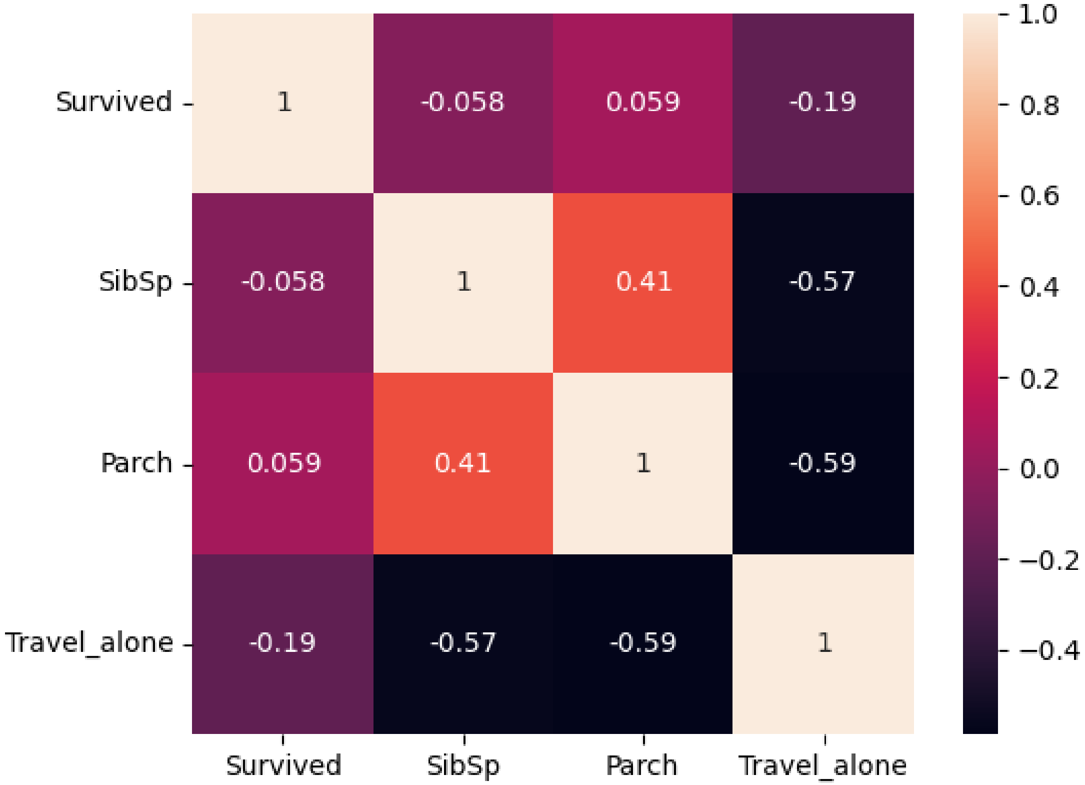
with_perents/chsildren_survival



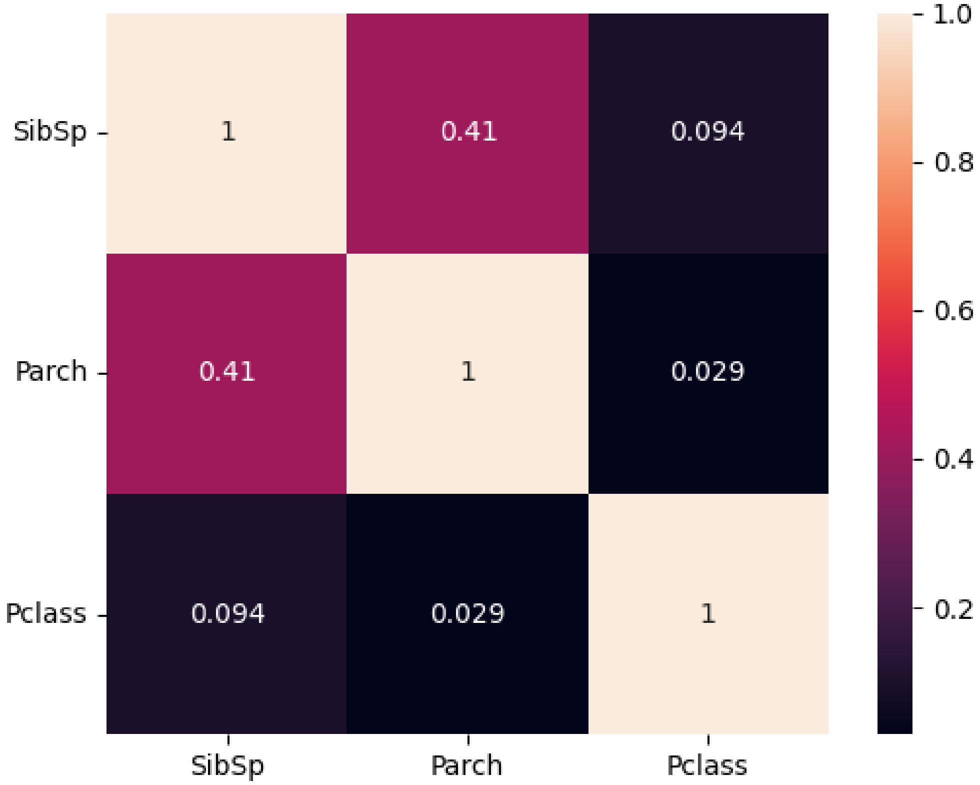
survivors_siblings_number



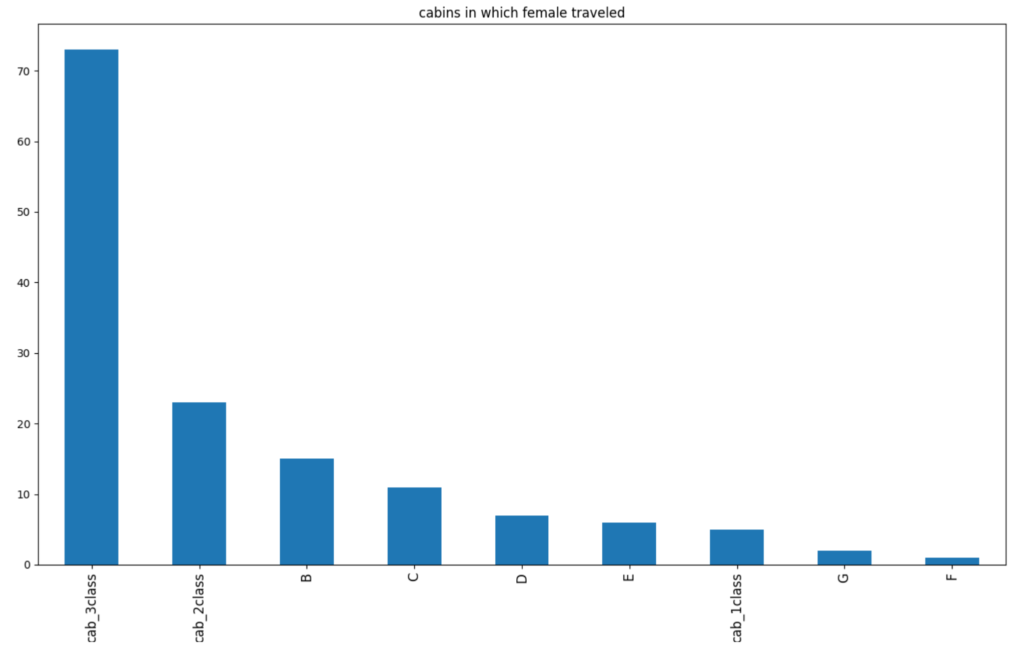
Most of the survivors had no family members onboard with them. The percentage of survivors among those who traveled alone is 31%. The percentage of survivors
among those who didn't travel alone is 50%.

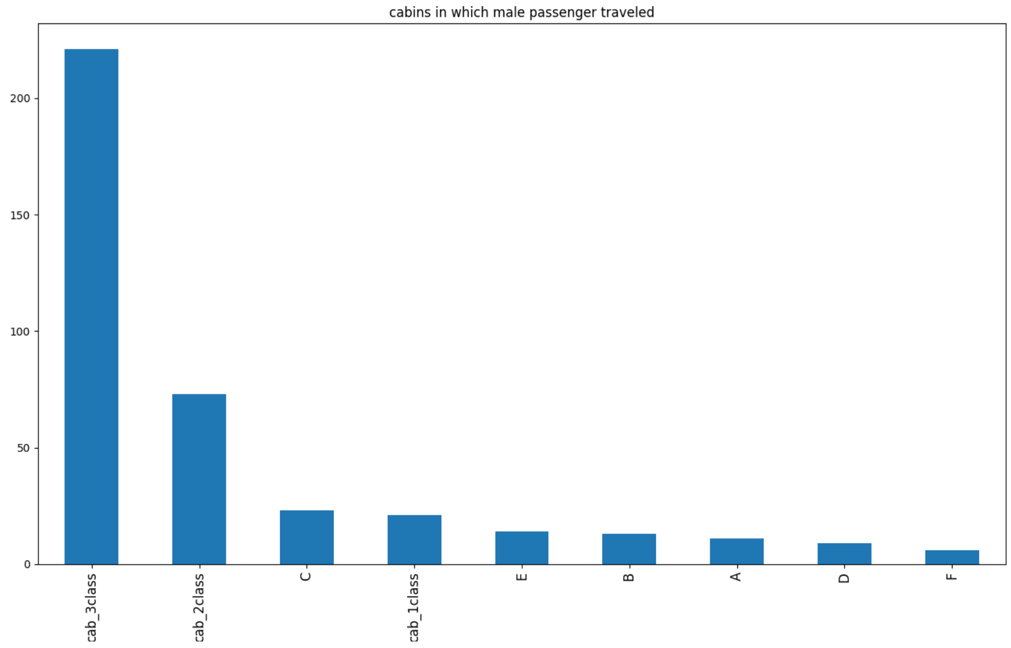


This shows a weak correlation between every 2 of those features (some of the correlations are obvious). Notice, that while the correlation between survival and having siblings is negative, the correlation between survival and having parents/children is positive. This can be explained by The next plot:

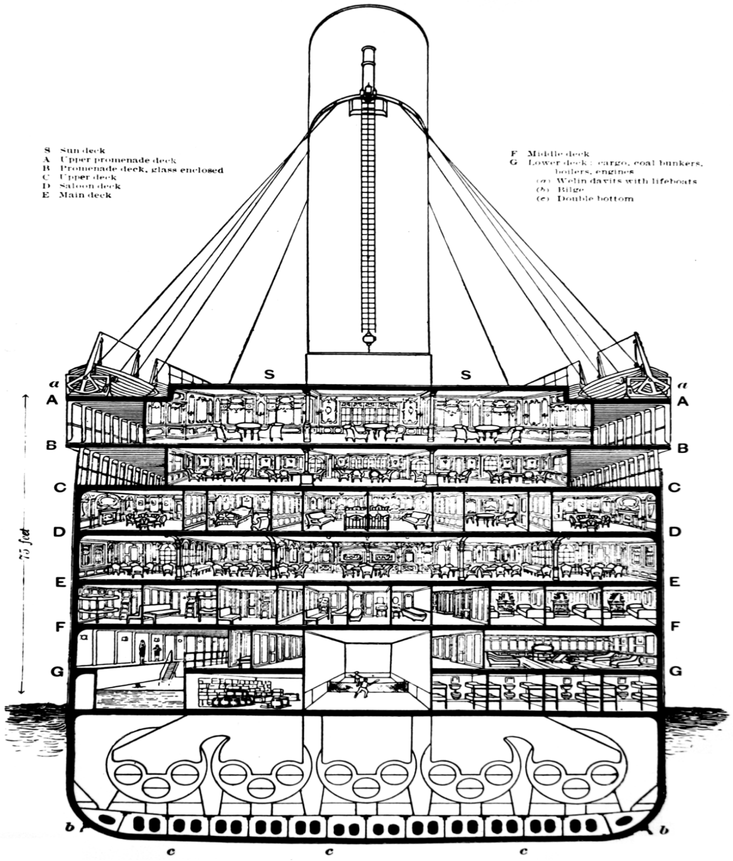


The correlation between having Parch and class is lower than the correlation between SibSp and class. As seen, higher classes had lower survival rates, and the higher correlation between SibSp and class suggests that people in higher classes were more likely to have Siblings than parents/children. In a similar way, people with siblings were more likely to be in a higher class, rather than people with parents/children.

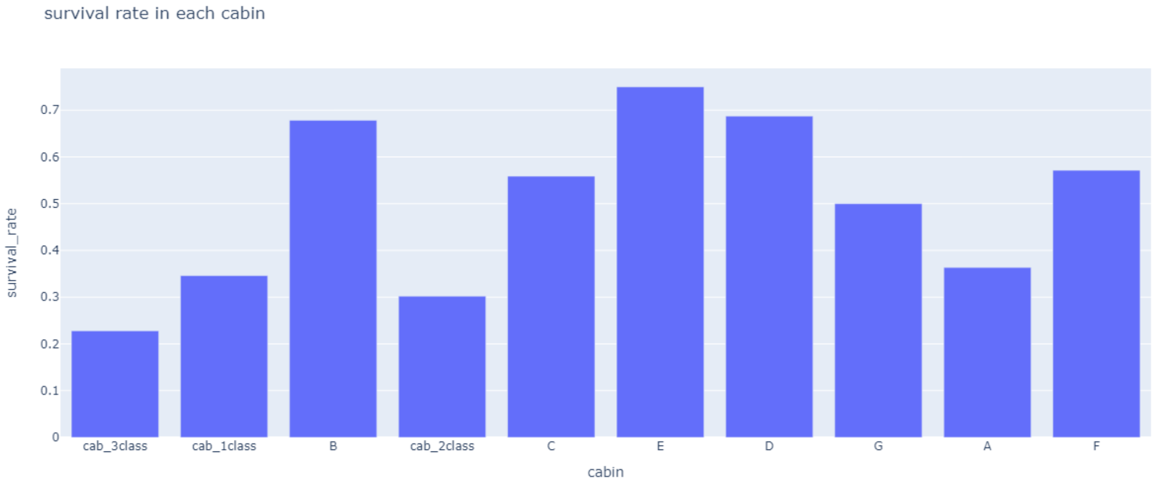




The cabins on the Titanic were divided by the floors of the ship as shown in the picture below:

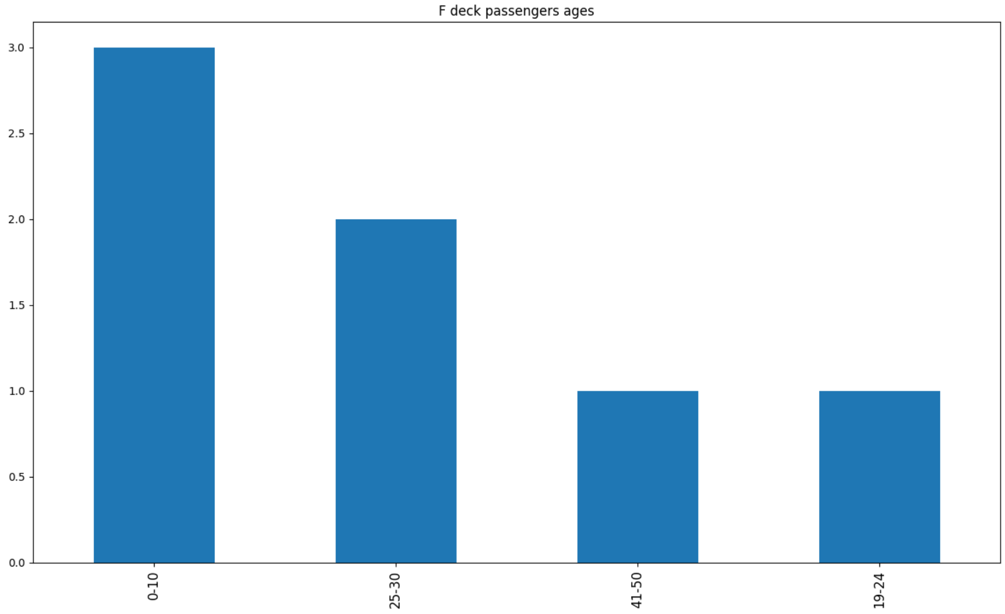


The decks A, B, C contained accommodations of first class passengers. Decks D&E contained all three classes accommodations. decks F&G contained both second and third class accommodations.

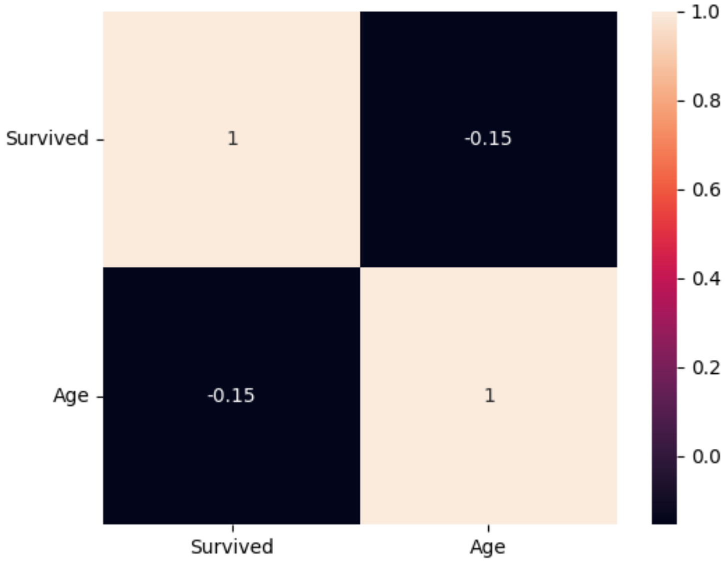


This plot shows the percentage of survival on each deck.

‘cabin_1class’, ‘cabin_2class’, ‘cabin_3class’ are replaced nan values according to the class of the passenger.
This is a little surprising as A deck had only first class passengers, nevertheless its survival percentage isnt relatively high. Althow deck G was second and third class deck, its relatively high survival percentage can be explained by the fact that G deck passengers were only women. We saw on earlier plots that females had greater percentage of survival compared to males. However this doesnt explain the relatively high survival percentage of deck F.

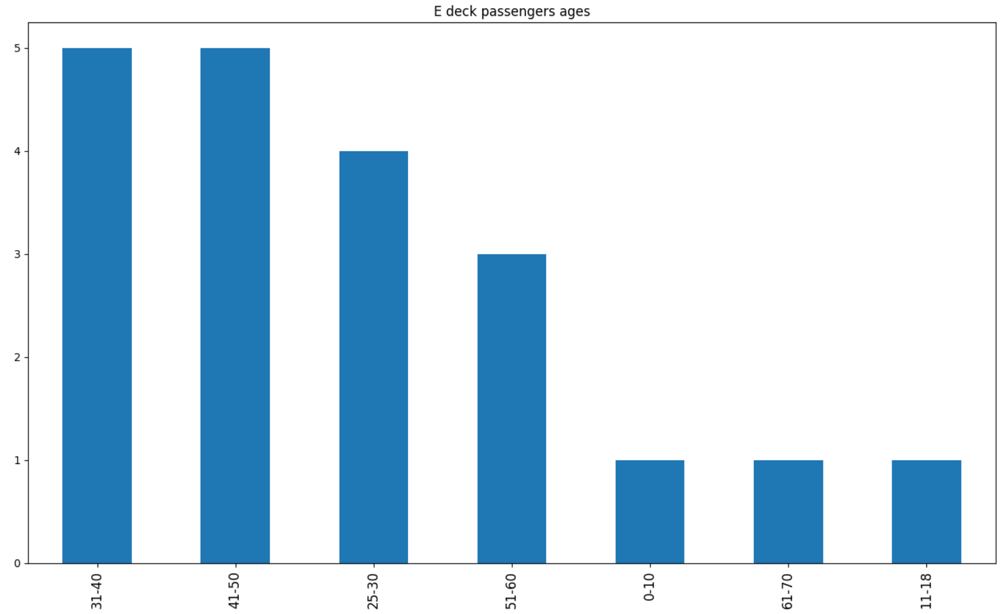


Most of passengers of deck F are children, The majority of other passengers are 19-30 years old.



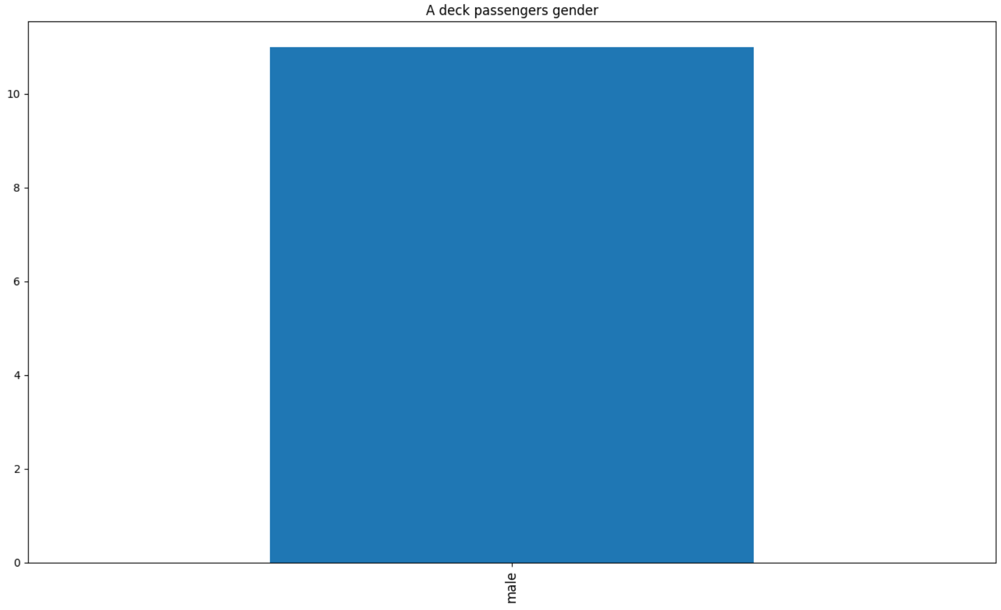
There is weak negative correlation between age and survival.

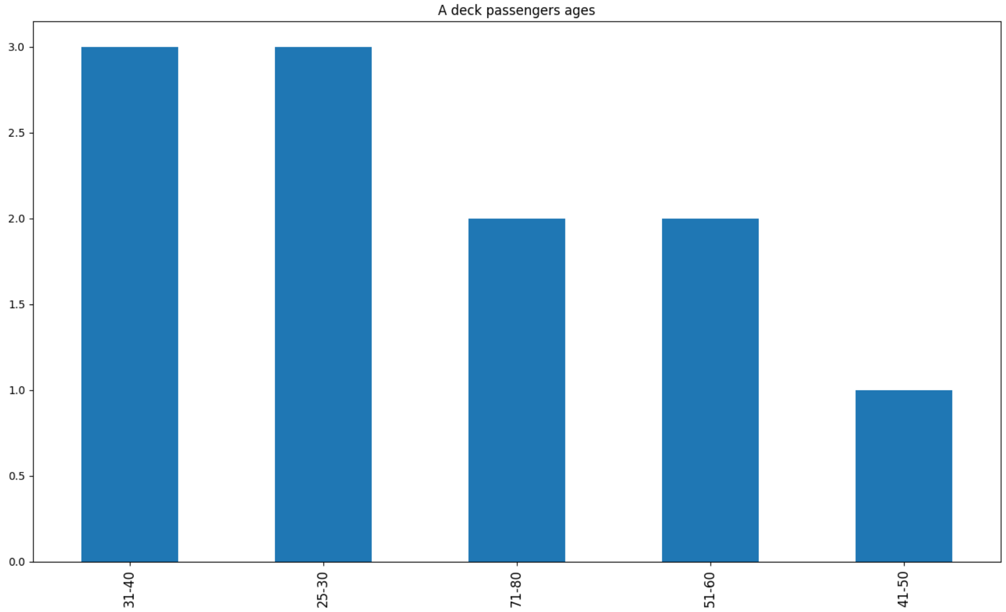
Deck E had the highest survival rate, althow this deck was intended for all three classes.



E deck passengers are of varios ages, and as seen in previous plots both males and females stayed on this deck.

Back to deck A:





Passengers of deck A where males only, and among them there were a lot of relatively older passengers. Those factors could have led to relatively low survival rate of this deck.

Conclusions:

It is clear that the factors age, gender, and class had an effect on Titanic passengers survival. It is still unclear whether the deck affected the survival directly, and not as a result of the previus factors. Female passengers had higher survival rates(75%) than male passengers(19%). Although most of the survivors didnt have any family members traviling with them, the survival percentage among those who didnt travel alone is higher (50%). Most of the passengers on the Titanic were males and the largest age group among them was 25-30. The largest age group among females was also 25-30, but in contrast to males, there were no females between the ages of 71-80. Most of the passengers were in class 3. Among males the class with most survivors was calss 3, and among females it was class 1. For males the class with the higher percentage of survival was 1 (36%), this was also the class with the higher percentage of surviva for female (95%). Notice that survival percentage of females in class 3 is significantly lower (51%) than in classes 1 and 2 (95% and 90%). There is negative correlation between every two factors in {survival, age, class}. There is negative correlation between survival and having siblings, and positive correlation between survival and having perents/children. There is positive correlation between every two factors in {class, perents/children/siblings}. The decks in descending order of survival rate:
E, D, B, F, C, G, A,‘cabin_1class’, ‘cabin_2class’, ‘cabin_3class’.

Data preprocessing

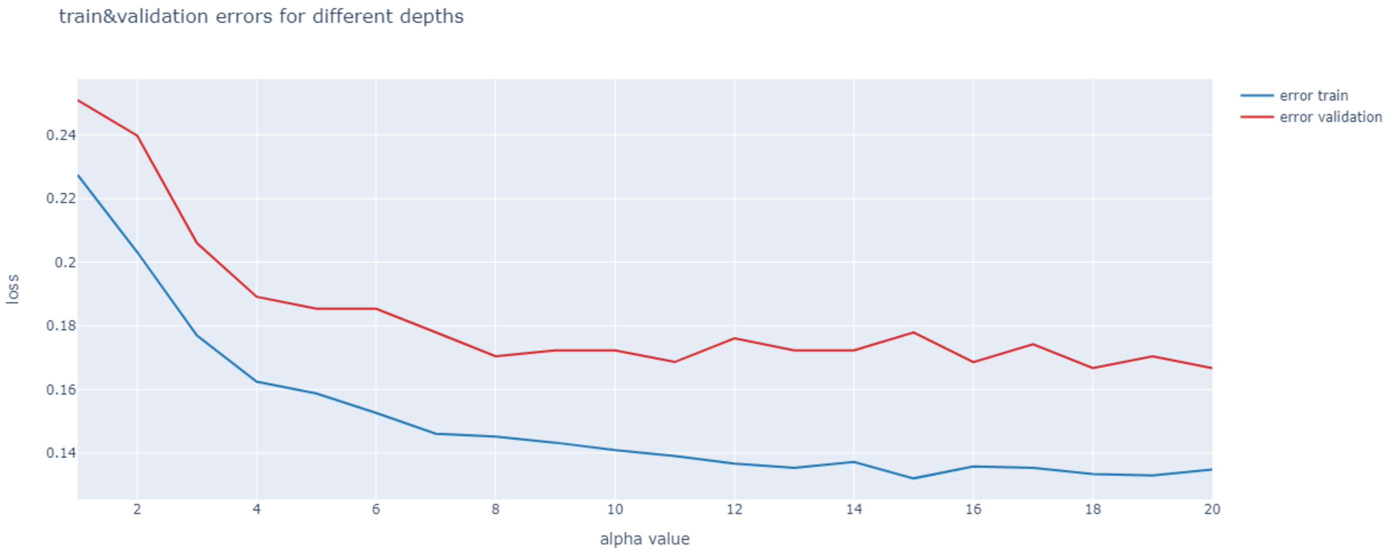
The training data contained missing fields, so I performed an imputation as follows:
Missing values of the column 'Pclass' were replaced with the mean value of the column.
The replacement was made under the consideration of the genders, meaning the mean
Value was measured separately for females and males, and was replaced according to
passenger's gender. In a similar way, missing fields of the columns ‘Age’, ‘SibSp’, ‘Parch’,
'Pclass' And ‘Fare’ was replaced by the mean column value, also considering gender.

As ‘Name’ is a categorical variable on one side but cannot be replaced with dummy values in its original format, I defined a group of the most commonly appearing titles and transformed the column to contain only titles instead of full names. Next, I checked the data integrity, I deleted all rows of data in which the passenger's title didn't match its gender, for example, a male passenger with the title ‘miss’. With this new format, I created dummy variables.

Many of the rows were missing the cabin value, So I relied on the class of the passenger
And filled the missing fields with the value ‘i_class_cabin’, where i stands for the passengers class. Also, because there were a lot of cabin values representing cabins on the same floor of the ship, I replaced all the values with the floor Letter. In addition, I checked the data integrity and deleted every row with an invalid floor. As the titanic had passenger cabins
on 7 floors, creating dummy values for the column ‘Cabin’ added 9 new columns to the DataFrame.
I dropped the column ‘passengerId’ from the DataFrame as it is just the row number
of the passenger in the DataFrame and doesn't hold any possible information about
The passenger survival. In addition, I decided to drop the ‘Ticket’ column, as there was no visible pattern to its values and every value was unique.
Lastly, I added a new feature to the data set, ‘Traveling_alone’, which is 1 if a passenger has no siblings, parents, or children on the ship with him, and 0 otherwise.

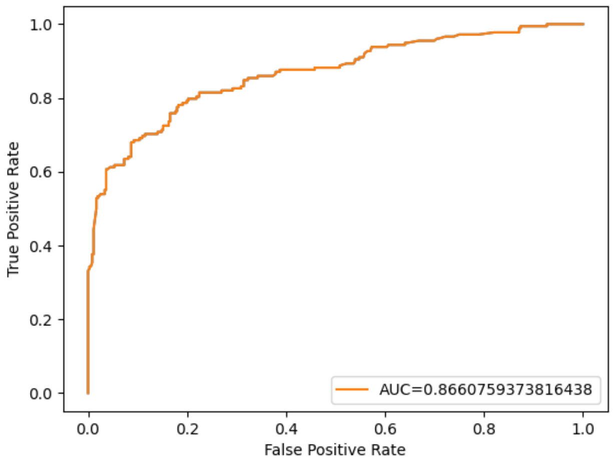
Model and prediction

I chose to use a random forest classifier for the task of predicting survival of Titanic passengers.
I wanted to avoid overfit and thus I set the min semples split parameter to 20.
To choose the depth of each tree in the forest I used cross validation:



As seen in the plot, the error value of the validation set decreases until depth=5, and from this depth value and on the validation error encreases and decreases.
I chose depth=5 by the scree plot method.

Next, I plotted the ROC curve:



The area under the curve is 0.866

I found those results satisfying and evaluated the model on the untouched test set, to estimate how it will succeed in the ‘real world’ (the unlabeled data received).
I predicted the labels for the test data with the fitted model, and measured the loss between the real and predicted labels. The value of the loss was 0.165, so I can expect the loss on the unlabeled data to be around 0.165.
I trained the model on the whole data set, preprocessed the unlabeled data and matched the columns of both labeled and unlabeled data.
The score I received for my predictions was 0.78708. As I dont know the scoring method Used by Kaggle I coudnt conclude whether my evaluation of loss in the ‘real world’ was accurate.