Machine Learning - Predicting Survival on the Titanic

Sylvia Lee(sylvia19) and Patrick Tung(ptung)

23 Nov, 2018

Introduction

Who will survive through the Titanic disaster?

For most people, "Titanic" is both a classic movie and a beautiful love story. However, the infamous Titanic catastrophe had also been said to be a prime example of social stratification and status discriminations in the 1900s. In addition to the "women and children first" evacuation method, it had been rumored that the lives of the people with social prestige and high class standing were prioritized in the momment of danger. In this analysis, we used supervised machine learning (ML) to answer the question *"What are the 3 strongest predictors of people who survived on the Titanic?"*

We retrieved the data from Kaggle's Titanic:Machine Learning from Disaster and developed a decision-classification-tree machine learning model focusing on following features:

Feature	Type	Description
Pclass	Categorical	Passenger Class
Sex	Categorical	Sex of Passenger
Age	Continuous	Age of Passenger
SibSp	Discrete	Number of siblings/spouses onboard
Parch	Discrete	Number of parents/children onboard
Fare	Continuous	Fare price

In our project, we explored the dataset by generating graphs for distribution of each features in the population of passengers. Subsequently we developed the decision tree model using Python's scikit-learn package and applied the model to a test dataset to predict the survival of the passenger given the same list of features. Lastly, we summarized our analysis by calculating the accuracy of our ML model and ranking the predictive power of each feature.

Exploratory Analysis

The RMS Titanic carried enough life boats for only one third of the passengers, and our data was reflective of this situaton. The data showed disproportionately larger proportion of passengers that did not survive. Therefore, we compared the feature distributions within each desinated groups, the "survived"and the "did not survive", and plotted each feature according to the passenger's survival status (Appendix I). This exploratory analysis allowed us to gain a sense of the differential distribution of features depending on the passenger's survival. If all features were equally weighted dueing evacuation, we assumed that the "survived" distribution would have frequecies equal to 1/3 of the "did not survive". However, that was not the case.

In general we found the data reflective of the "women and children first" evacuation policy. There seemed to be larger proportion of women and children that survived than those that did not. Interestingly, we found that there were indeed larger proportion of survived passengers that had the features of "first class passenger" and "paid high fare price". On the other hand, family size (number of parent, children, siblings and spouse) did not appear to cause large differences.

Predictions and Evaluations

Decision Tree

We generated a decision classification tree model using scikit-learn package. In order to reduce overfitting, we ran a 10-folds cross-validation to find the best `max-depth` hyperparameter and developed the learning model accordingly.

Our decision tree model made the first split on the feature "Sex", meaning that the model evaluated gender as the best general feature for predicting survival. A graphic representation can be found in the `results` sub-repository.

Predictions

We ran our trained decision tree model on both the training and testing dateset to inspect its predictive capabilities (Table 1, Table 2). Qualitatively inspecting the target ("Survived") column and the "Prediction" column, we found our model did reasonably well predicting survivals in both datasets.

Table 1. Snippet of Predictions for both the Training set.

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Survived	Prediction
1	3	1	22.00000	1	0	7.2500	0	0
2	1	0	38.00000	1	0	71.2833	1	1
3	3	0	26.00000	0	0	7.9250	1	0
4	1	0	35.00000	1	0	53.1000	1	1
5	3	1	35.00000	0	0	8.0500	0	0
6	3	1	29.69912	0	0	8.4583	0	0
7	1	1	54.00000	0	0	51.8625	0	0
8	3	1	2.00000	3	1	21.0750	0	0
9	3	0	27.00000	0	2	11.1333	1	1
10	2	0	14.00000	1	0	30.0708	1	1

Pclass = Passenger Class, Sex = 0-Female, 1-Male, SibSp = #siblings/spouse onboard, Parch = #parents/children onboard, Survived = 0-Died, 1-Survived

Table 2. Snippet of Predictions for Testing set.

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Survived	Prediction
892	3	1	34.5	0	0	7.8292	0	0
893	3	0	47.0	1	0	7.0000	1	0
894	2	1	62.0	0	0	9.6875	0	0
895	3	1	27.0	0	0	8.6625	0	0
896	3	0	22.0	1	1	12.2875	1	1
897	3	1	14.0	0	0	9.2250	0	0
898	3	0	30.0	0	0	7.6292	1	1
899	2	1	26.0	1	1	29.0000	0	0
900	3	0	18.0	0	0	7.2292	1	1
901	3	1	21.0	2	0	24.1500	0	0

Pclass = Passenger Class, Sex = 0-Female, 1-Male, SibSp = #siblings/spouse onboard, Parch = #parents/children onboard, Survived = 0-Died, 1-Survived

Model Performance

To quanititatively evaluate the accuracy of the model, we calculated both the training and testing accuracies by taking the proportion of correct predictions in both the training and testing datasets (Table 3.). What we were trying to inquire was whether or not the accuracy of our testing model would decrease in comparison to our training model. This was done to address possible overfitting problems.

Table 3. Prediction accuracy scores of ML model on the training and testing sets.

Dataset	#Total Samples	#Correct predictions	#Incorrect predictions	Accuracy Score
train	342	266	76	0.7778
test	152	129	23	0.8487

Our model predicted the training dataset with an accuracy of 0.7778, and predicted the testing dataset with an accuracy of 0.8421. Interestingly, we found higher accuracy in our test dataset than the training dataset, which suggested that our model was adequetely generalizable for data outside of the training dataset.

Feature Importance Ranking

The ultimate goal of our research was to determine which three features were the most important among others. In order to achieve this goal, we took our classification tree model and generated an importance score using the `sci-kit` learn package. The importance score was evaluated based on "gini importance", which was also known as the "mean decrease in impurity". Essentially, the higher the importance value, the more important that feature was.

Table 4. Ranks of each feature based on Gini Importance

```
## Warning in read.table(file = file, header = header, sep = sep, quote =
## quote, : cols = 4 != length(data) = 6
```

Rank	Feature	Importance
1	Sex	0.4787907
2	Pclass	0.1664133
3	Age	0.1455608
4	Fare	0.1435919
5	SibSp	0.0496373
6	Parch	0.0160059

Pclass = Passenger Class, SibSp = #siblings/spouse onboard, Parch = #parents/children onboard

From our results, we determined that the three most important features in our model were: 1) Sex, 2) Passenger Class, 3) Age. The gini importances were 0.4788, 0.1703, and 0.1555 respectively.

Limitations and Assumptions

First of all, the biggest limitation to our project was that we chose to explore only one type of model, the decision tree. Given more time and resources, we would test out different models to find the best predictive model for our problem. However, because we had not yet learned other ML models, we were wary of conducting an analysis with unfamiliar methods. In order to compensate for the lack of model exploration, we used cross validation to pick the best `max_depth` hyperparameter for our decision tree

Cross-validation assumed that all features were iid variables. However, two of our features might had been correlated. We had #siblings/spouse and #parent/children as two different features, but these two features could had been analyzed as one feature of "family size". Logically, the two features would have influenced each other, thus undermined the effectiveness of our cross-validation. However, our model's high testing accuracy suggested that this would not cause major downfalls in the machine learning.

Another limitation that we encountered was that we used means and medians for the imputation of missing values. As an alternative, we could use regressors to make predictions on the best value to replace the missing values. However, this was beyond our current knowledge, so we resorted to means and medians as sufficient imputation methods for our predictions.

Lastly, for our prediction, we decided to subset the dataset to only the relevant features that we were looking for in our research question. The entire data set that we originally started with had many more features such as where the passenger embarked, however, we decided to use only a subset of the features to predict survival rates for simplicity sake. Despite using less features, we believe that we still performed quite well with our predictions.

Conclusion

We analysed passengers from the RMC Titanic and developed a classification-tree machine learning model that would allow us to predict which passenger was more likely to survive based on certain features. Our machine learning model achieved a fairly high accuracy of 84% in our testing model. Additionally, we found that the most predictive features were gender, passenger class, and age, which cohered with our expectation that in addition to the "women and children first" evacuation policy, passengers with higher social standing were prioritized as well.

References

Documentation of scikit-learn 0.20.1. (n.d.). Retrieved from https://scikit-learn.org/stable/documentation.html

On the importance of the i.i.d. assumption in statistical learning. (n.d.). Retrieved November 20, 2018, from https://stats.stackexchange.com/questions/213464/on-the-importance-of-the-i-i-d-assumption-in-statistical-learning

Titanic: Machine Learning from Disaster. (n.d.). Retrieved November 20, 2018, from https://www.kaggle.com/c/titanic/data

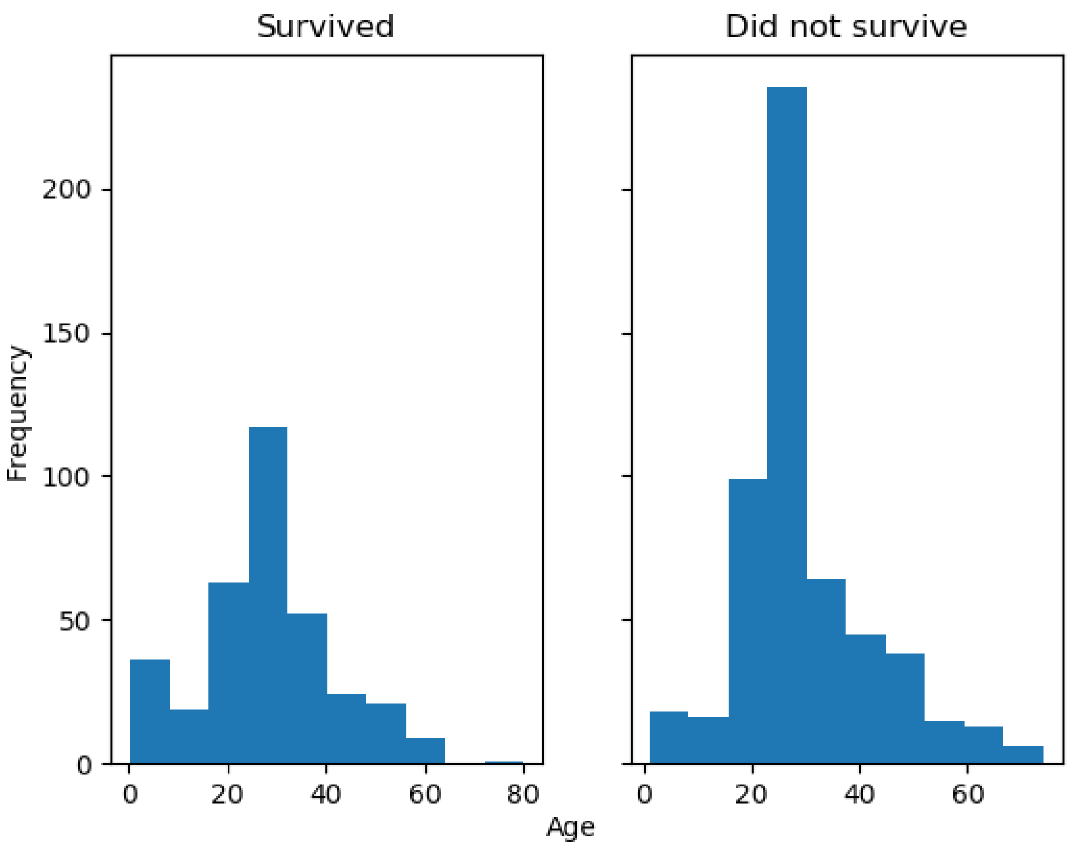
Swalin, A. (2018, January 31). How to Handle Missing Data - Towards Data Science. Retrieved from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Wikipedia - RMS Titanic. (n.d.). Retrieved from https://en.wikipedia.org/wiki/RMS_Titanic

Appendix

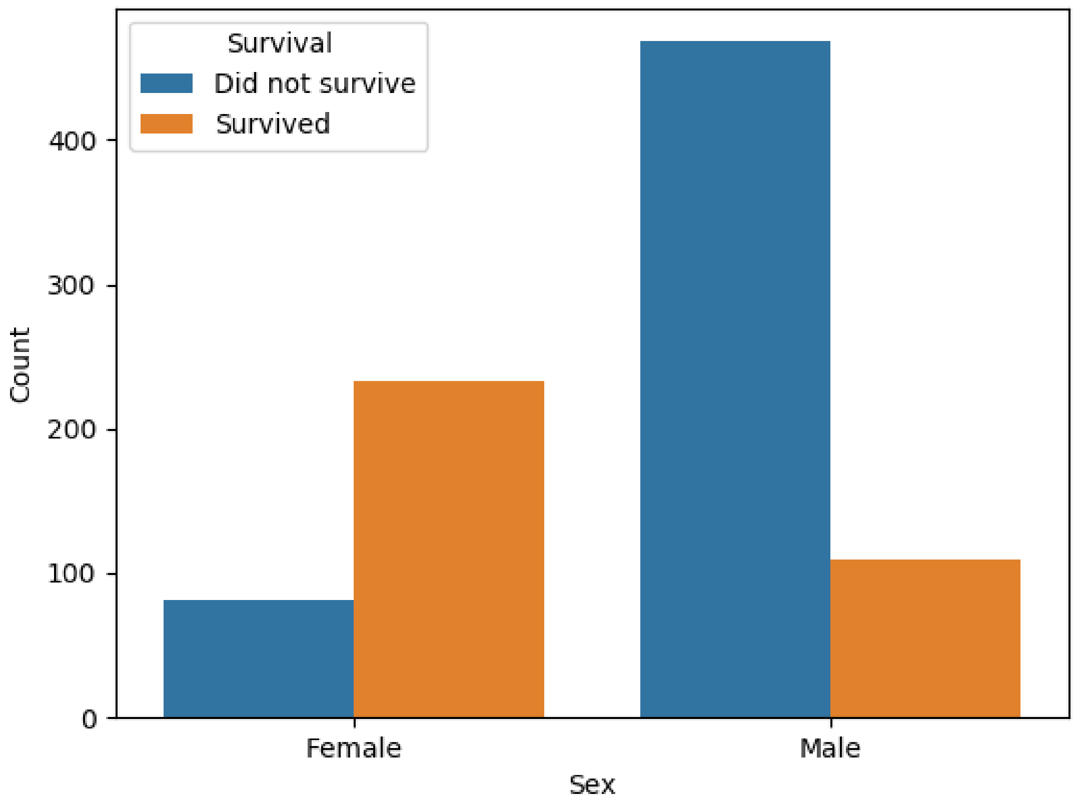
Appendix I: EDA Figures

Age



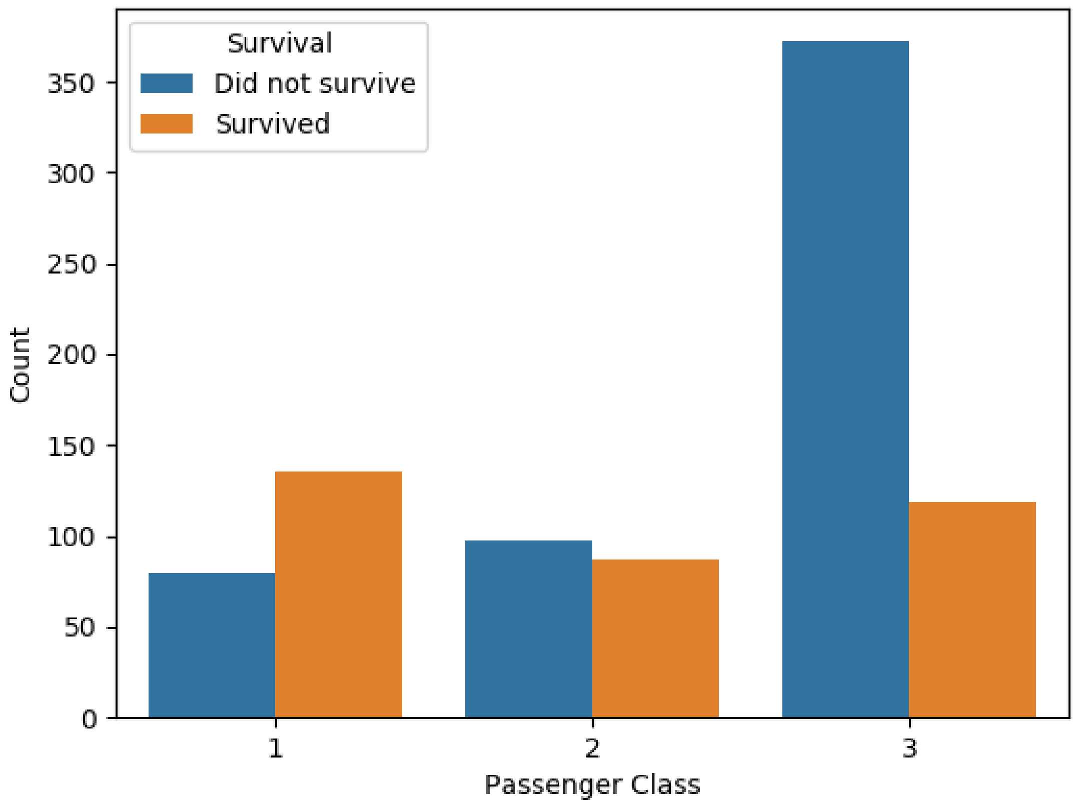
Append 1. Histograms of ages among the passengers that survived (left) and did not survive (right).

Sex



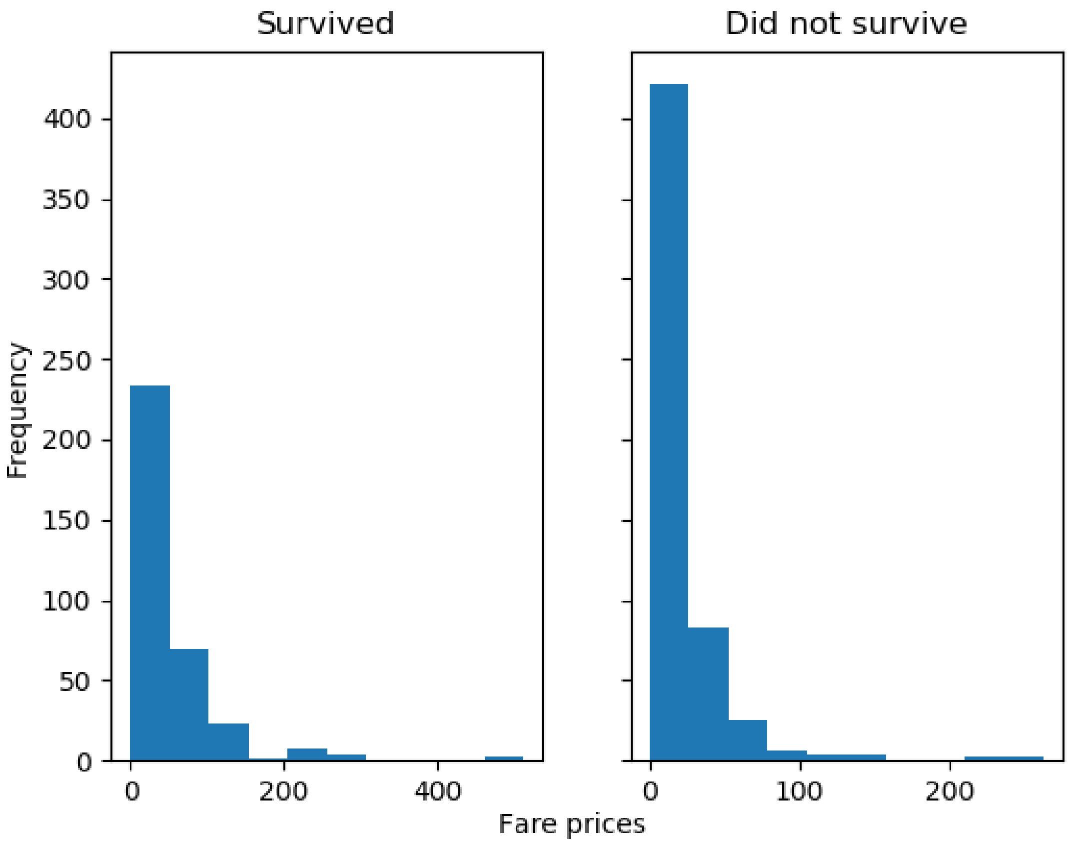
Append 2. Bar plot of sex distribution among the passengers that survived versus those did not survive.

Passenger Class



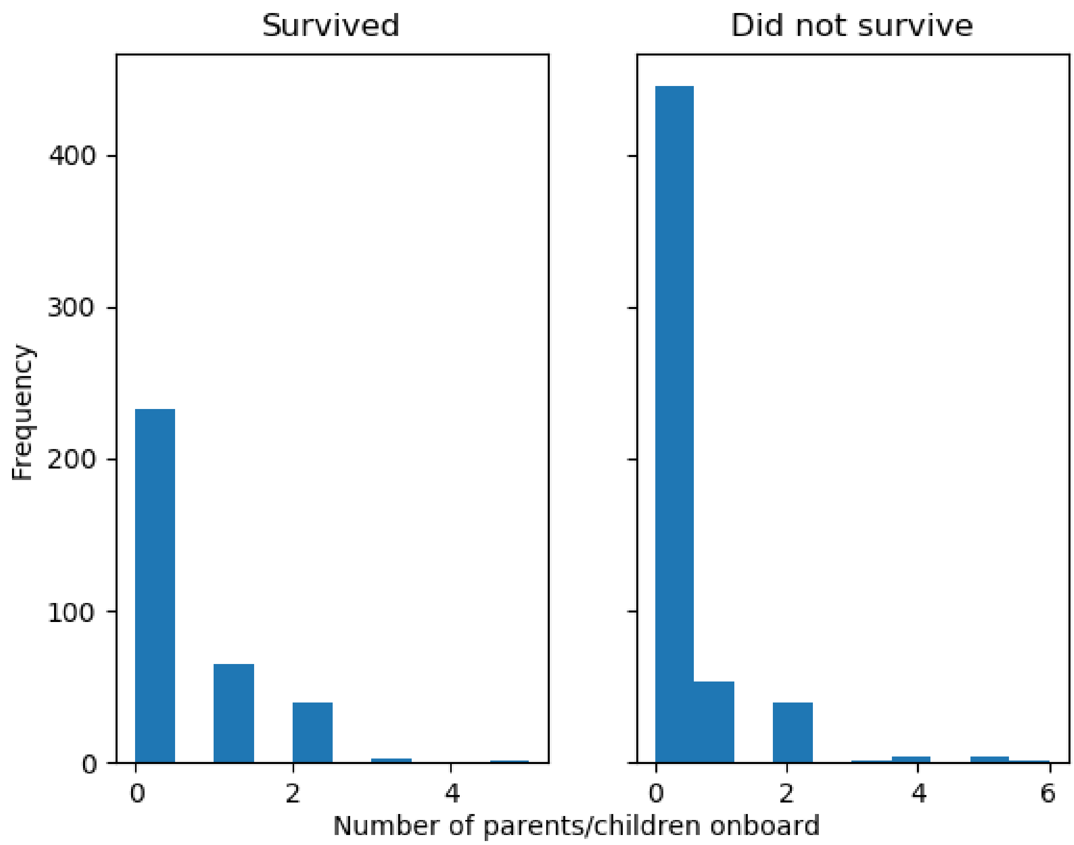
Append 3. Bar plot of passenger class distribution among the passengers that survived versus those did not survive.

Fare Price



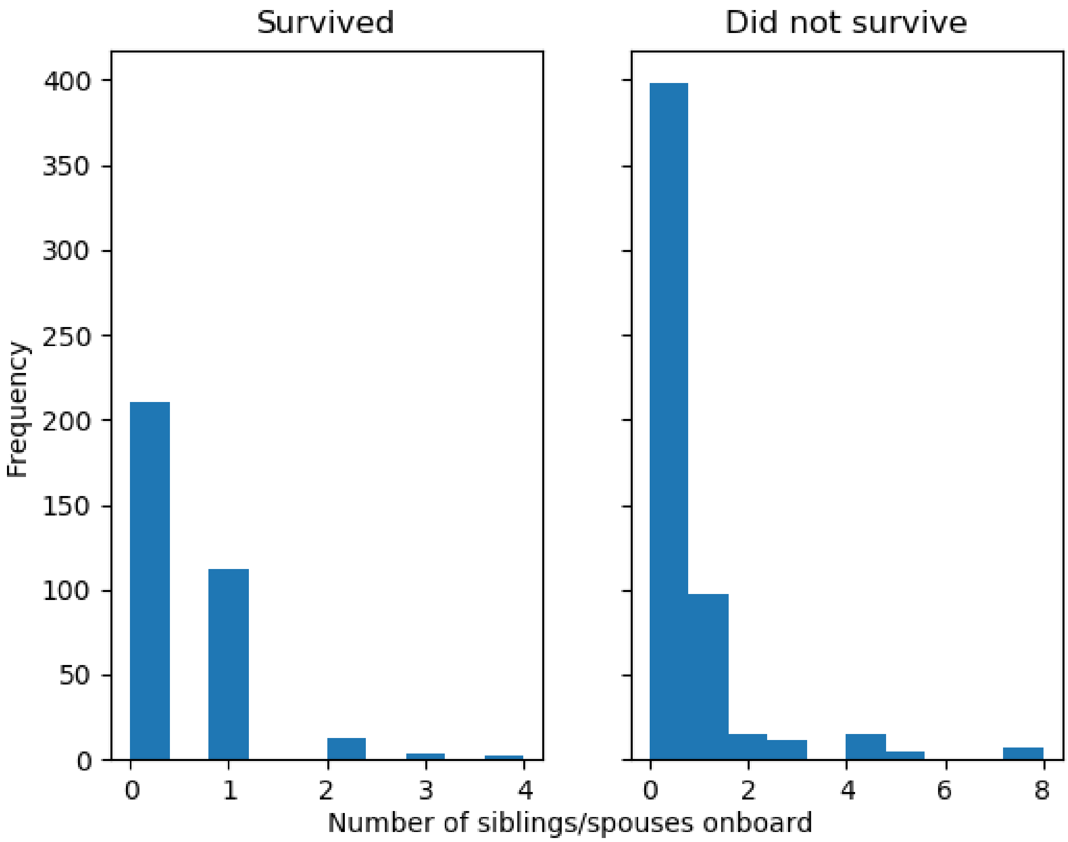
Append 4. Histograms of fare prices paid by the passengers that survived (left) and did not survive (right).

Number of Parents/Children Onboard



Append 5. Histograms of number of parent or children that was onboard with the passengers that did survive (left) and did not survive (right).

Number of Siblings/Spouses Onboard



Append 6. Histograms of number of siblings or spouse that was onboard with the passengers that did survive (left) and did not survive (right).