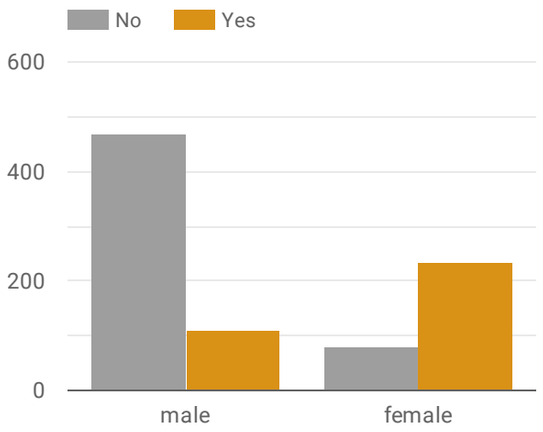
48hours.ai

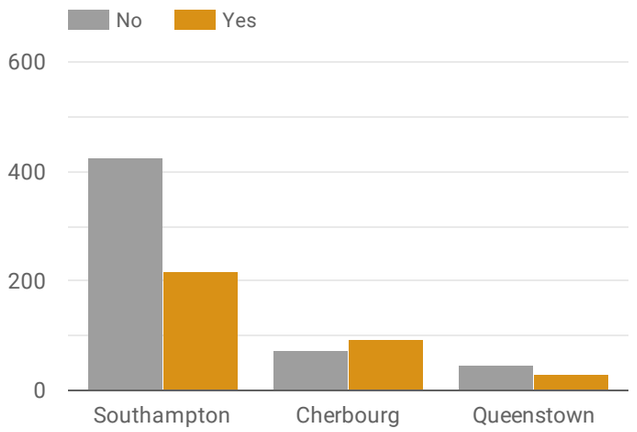
Data set name: Titanic
Date received: 2019/08/01
Process time: 38/48 hours
Engineer: G van Eeden
E-mail: info@itinnovate.co.za
Data files: test.csv & train.csv
General report notes:
This data set has good predictive capabilities.

48hours Data Exploration



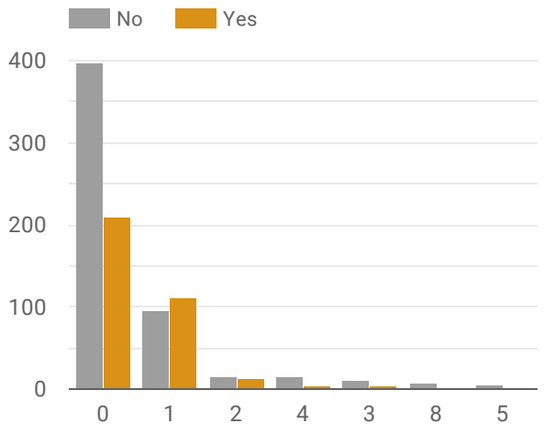
Survival by Gender

From the above chart it is clear that the survival rate of females was greater than the survival rate of males.



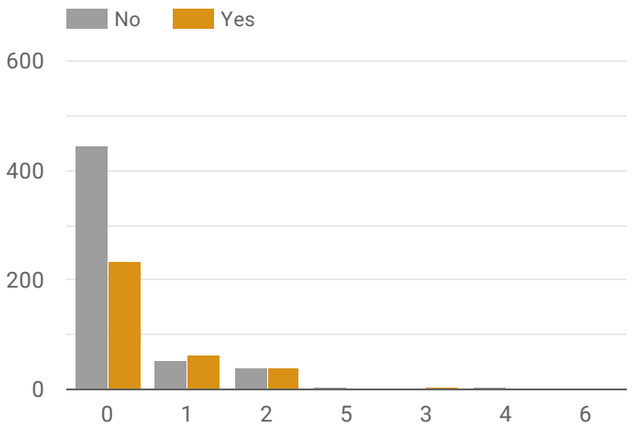
Survival by boarding port

From the above chart it is clear that the lowest survival rates were amongst people from Southampton. Passengers who had had boarded at Cherbourg had a greater than 50% chance of survival.



Survival by number of siblings / spouses on board

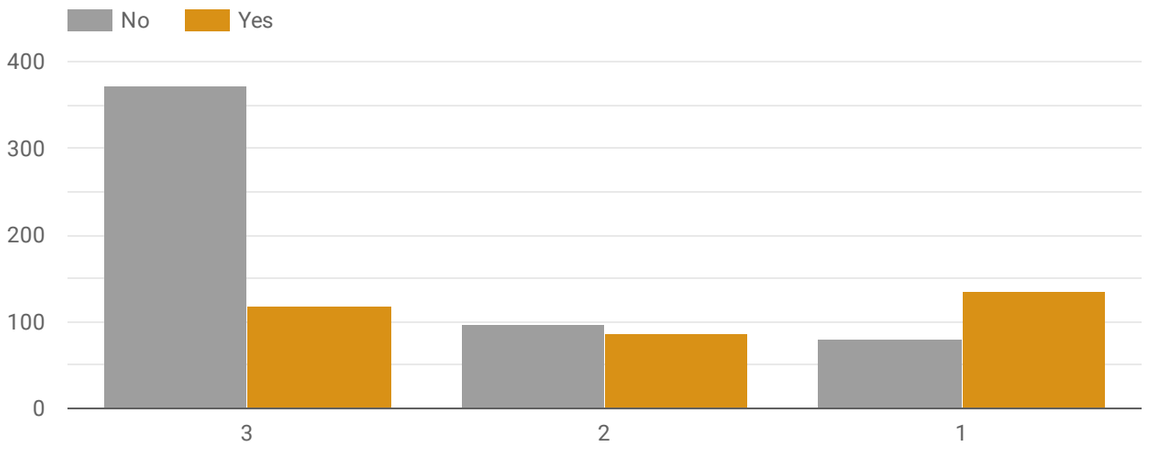
Passengers who travelled without siblings or spouses had a lower survival rate.



Survival by parents / children on board

Parents and children travelling together had higher survival rates than those travelling solo.

48hours Data Exploration



Survival by Passenger class

Remember the movie. This graph clearly shows that chances of survival for first class passengers were much higher than for third class passengers.

Next steps
Following the initial data exploration, we are now going to try to predict survivals on the Titanic by using machine learning.

48hours Machine Learning Models

Can future survivals be predicted from the data? To make the predictions, we follow this methodology:

1. Run the training file through a few statistical models.
2. Test the models against unseen data.
3. Select the best model and make predictions.

Is data science that simple? **Yes!**

Models used

The following models were used:
- AdaBoost
- Logistic Regression
- Naive Bayes
- Neural Network
- Random Forest
- Support-vector Machine
- k-nearest Neighbors

And the winner is: **Logistic Regression**!

Below is 'n summary of the performance of different models.

AUC score

Model	AUC
Logistic Regression	0.984
Neural Network	0.964
SVM	0.956
Naive Bayes	0.903
Random Forest	0.9
AdaBoost	0.843
kNN	0.711

What is AUC?

In essence AUC is a model evaluation metric. If AUC equals 0.5, a model has zero value. If the AUC is close to 1, the model is a good one and has sound predictive capabilities. To read more about AUC download the AUC.pdf from the 48hours.ai website. (https://www.48hours.ai/files/AUC.pdf)

48hours Machine Learning Models

Confusion Matrix

Another way to evaluate models is to use a confusion matrix. The confusion matrix below clearly indicates that incorrect predictions are statistically very low: 7.7% for non-survivals and 3.1% for survivals.

		Predicted		
		0	**1**	**∑**
Actual	**0**	96.9 %	7.7 %	**266**
	1	3.1 %	92.3 %	**152**
	∑	**262**	**156**	**418**

Model predictions

To view the predictions of the model, download it from the following URL:
(https://www.48hours.ai/files/test_predictions.csv)