**The RMS Titanic**

RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster.

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Overview

The data has been split into two groups:

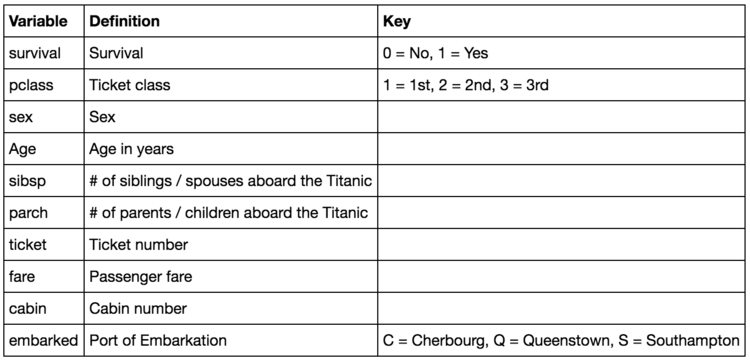
training set (train.csv)

test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use feature engineering to create new features.

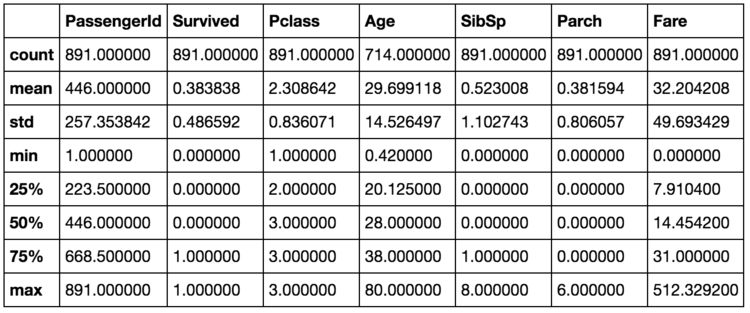
The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

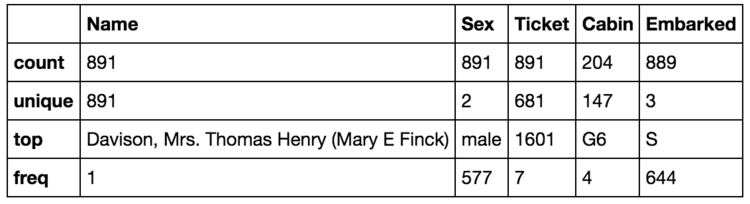
We also include gender\_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.



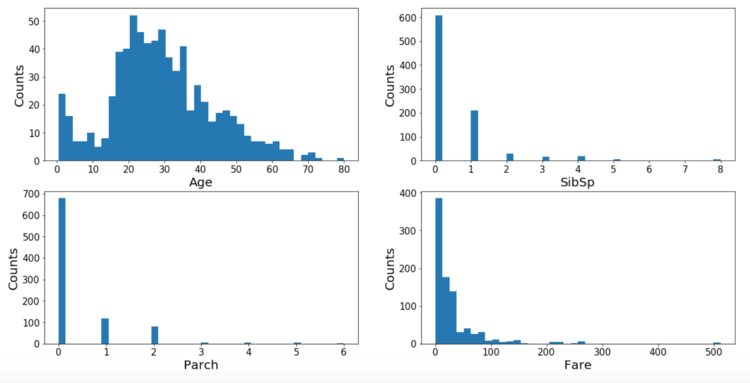
Tasks:

1. Getting the Data in Python environment
2. Take a look at the first 5 rows of our data set.
3. Data Exploration/Analysis with descriptive Statistics
   1. What do the variables look like?

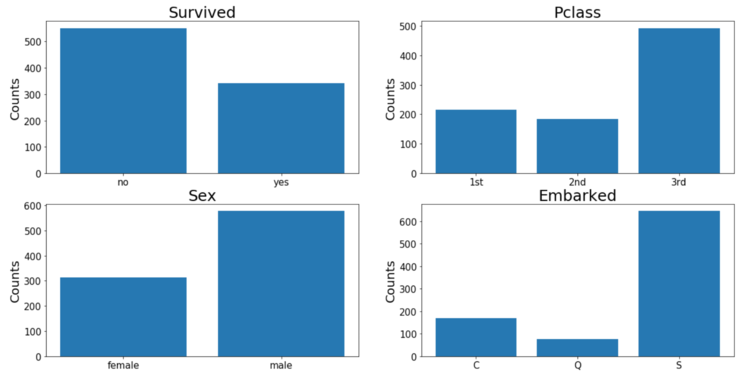




* 1. Are the numerical variables correlated?



* 1. Are the distributions of numerical variables the same or different among survived and not survived?



* 1. Are there different survival rates in different categories?
  2. Dropping Features We Will Not Use

We will delete the features "PassengerId", "Name", "Ticket" and "Cabin" from our model. The reasons are as follows:

* "PassengerId": just a series of numbers from 1 - 891 which is used to label each person.
* "Name": the names of all the passengers, which might give some information like if there are some people are related based on the last names. But to simplify things up at this stage, I will pass this feature.
* "Ticket" and "Cabin": too many levels with unknown information.

1. Filling In Missing Values

From EDA, we know there are some missing value in "Age", "Cabin" and "Embarked" variables. Since we are not going to use "Cabin" feature, we will just fill in "Age" and "Embarked." I will fill the missing values in "Age" using the median age and fill the missing value in "Embarked" with "S" since there are only 2 values missing and "S" is the most represent in the dataset.

1. Run Linear Regression to predict the fare of ticket based on best fit parameters.
2. Run Logistic Regression to predict the survival of passengers based on best fit parameters.