### **Titanic Overview**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, the idea was to conduct an analysis of what sorts of people were likely to survive. In particular, to apply the tools of machine learning to predict which passengers survived the tragedy. In class up until this project, I had been learning all about Pandas and plotting from DataCamp. I have thoroughly enjoyed this project!

Brief Overview of the Workflow:

* Reading in the data
* Cleaning the data
* Basic feature extraction
* Exploratory analysis
* Visualizing findings

Step 1: Reading the data

* Data drawn from <https://www.kaggle.com/c/titanic/data>
* An available data dictionary explains each of the columns. It's important to get familiarized with how the csv is structured.
* Downloaded the train.csv file into this project repo
* Created an iPython notebook and read the csv in using pandas.

Step 2: Data Cleaning

* Investigated and dealt with null values
* Searched for any discrepancies with data types and dealt with them

Step 3: Feature extraction

1. There are two columns that pertain to how many family members are on the boat for a given person. Created a new column called FamilyCount which is the sum of those two columns.
2. In order to feed the training data into a classification algorithm, I needed to convert the categories into 1's and 0's using pd.get\_dummies

* Created 3 columns: Embarked\_C, Embarked\_Q and Embarked\_S. These columns have 1's and 0's that correspond to the C, Q and S values in the Embarked column
* Did the same thing for Sex

Step 4: Exploratory analysis

* Asking all of the important questions to see what factors influenced survival rates on the Titanic.
* Most importantly, would *I* have survived?! (Spoiler alert: not even close).

Step 5: Visualizing findings

Using Matplotlib and Seaborn, I created several charts showing the survival rates of different groups of people. A handful of the charts are basic (Gender, Age, etc), but I really tried looking for something beneath the surface.

Project Feedback + Evaluation

For this project, I wanted to make sure I was adhering to the following principles throughout:

* Organization: Clearly commented, annotated and sectioned Jupyter notebook. Comments and annotations add clarity, explanation and intent to the work. Notebook is well-structured with title, author and sections. Assumptions are stated and justified.
* Data Structures: Python data structures including lists, dictionaries and imported structures (e.g. DataFrames), are created and used correctly. The appropriate data structures are used in context. Data structures are created and accessed using appropriate mechanisms such as comprehensions, slices, filters and copies.
* Python Syntax and Control Flow: Python code is written correctly and follows standard style guidelines and best practices. There are no runtime errors.
* Charting: Charts (from Pandas, Matplotlib, and Seaborn) are created and used correctly. The appropriate charts are used in context. Charts are created and accessed using appropriate mechanisms.