Introduction:

The sinking of the Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her first voyage, the Titanic sank after colliding with an iceberg, killing over 67% 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. There was some element of luck involved in surviving the sinking. In this project, we would like to complete the analysis of what sorts of people were likely to survive.

Data set:

The dataset contains 891 observations and 12 variables. The variables include: survival (0=No, 1=Yes), Pclass (Ticket class,1st = Upper,2nd = Middle, 3rd = Lower), Age,   
Sibsp (siblings / spouses aboard the Titanic, the dataset defines family relations about brother, sister, husband and wife), Parch (parents / children aboard the Titanic), Ticket ( Ticket number), Fare (Passenger fare), Cabin (Cabin number), Embarked (Port of Embarkation, C = Cherbourg, Q = Queenstown, S = Southampton).

After analyzing the data, we can find that there are three variables have missing values. Cabin has the highest amount of missing values and the second one is Age and the last one is Embarked. Therefore, the next step we need to do is to replace NA in variables.