RMS Titanic Analyst Project

Student: Alex Sewell | 6/17/2016

*Udacity Introduction to Programming*

*Choose Your Path – Data Analyst*

# Introduction

This report focused on deriving insights from RMS Titanic data provided by Kaggle/Udacity in order to exhibit my ability to use specific analytics tools such as;

* NumPy arrays, Pandas Series and DataFrames.
* Functions are optional but encouraged where repeatable code is used.
* Where possible, vectorized operations
* Graphical and multi-angle analysis with at least two kinds of plots should be used.
* Reasoning for each analysis decision, plot and statistical summary.

The RMS Titanic was an Olympic-class ocean liner spanning 883 feet or 269 meters in length with a crew and passenger count of 2,224. On its maiden voyage across the North Atlantic Ocean towards Canada’s eastern coast the ship collided with an iceberg at approximately 2:00 am the morning of April 15th, 1912. 1,502 passengers and crew died.

From this data, my analysis focused on three key questions;

* Can we determine survival counts by age?
* Can we determine survival counts by gender?
* Can we determine a male to female survival comparison?

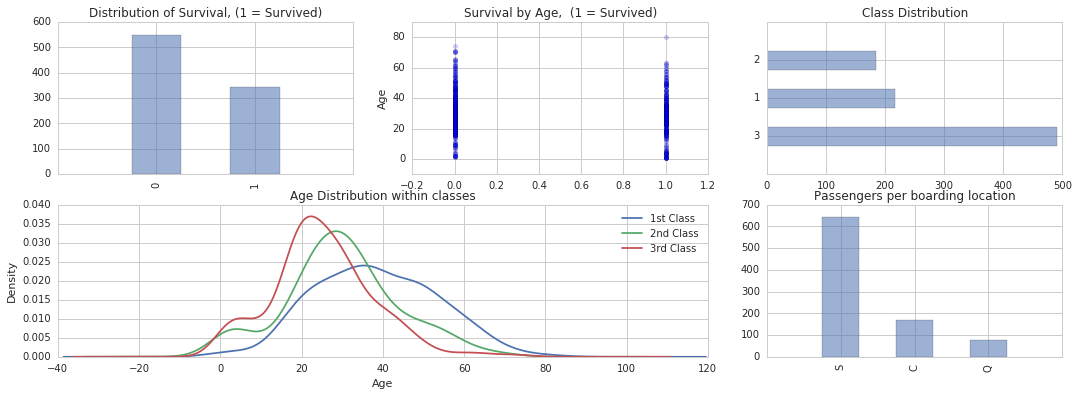
From what I’ve learned here, the majority of passengers were less than 30 years of age and that of the survivors, female to male survival rates were proportionately higher. Herein I’ve provided data cleaning methods, analysis work with visualizations and information about the variables and data logic as determined by the data providers.

# Analysis

## *Age Analysis*

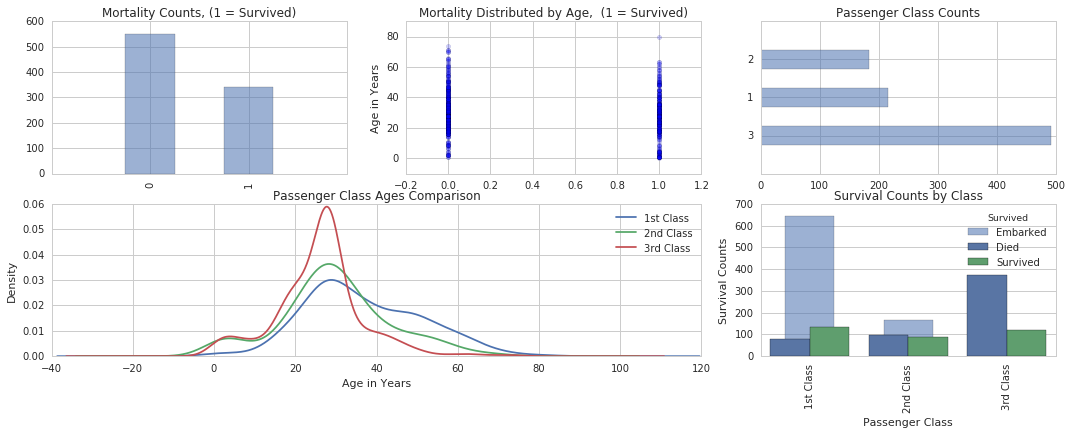
Can we determine survival counts by age?

Yes, the median age is 28 years and the mean age is approximately 29 years for each passenger and we can see this distribution among the classes as well.



However, review of the data revealed 177 NaN values exist and could impact the median and mean values. Instead of omitting them, I’ve amended the NaN values with the median age 28.

The result shows a higher age density for each passenger class falling between the age of 20 and 30 years of age.



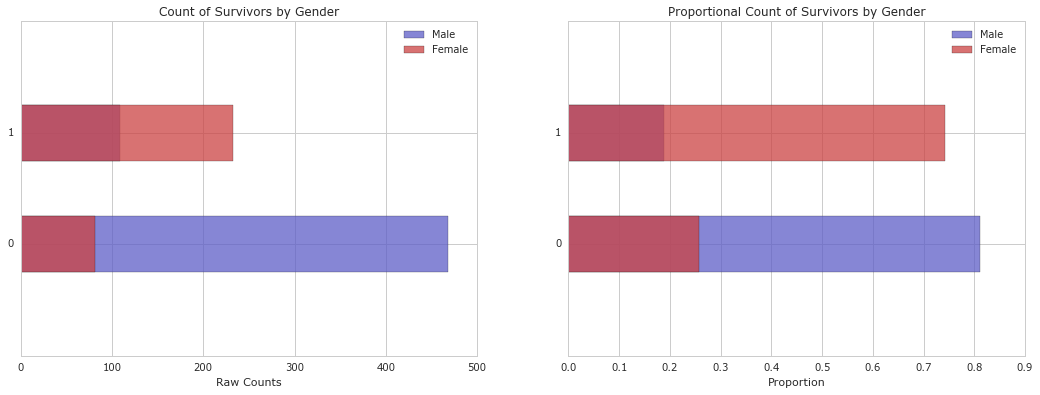
## *Sex Analysis*

Can we determine survival counts by gender? Can we determine a proportional survival comparison?

Yes, given the ‘Sex’ variable is not missing any values and ‘Survived’ is not missing any values, we’re able to both determine both counts of survivors by sex and represent survival rates proportionally for both male and female survivors.

‘Count of Survivors by Gender’ tells us that more males died males survived than died and that more females than males survived based on the raw counts. We can also see that male deaths were approximately 4x greater than females. However, this can be deceiving without proportionality of female to male counts.

‘Proportional Count of Survivors by Gender’ reveals that females were more likely to survive (~25%) than males (~20%) proportionally.



# Data Cleaning

## Cleaning

1. ‘Ticket’ and ‘Cabin’ as we’re omitted from the exercise as not meaningful.
2. ‘Age’ has 177 NaN values, so the median age 28 was used in place of NaN values.
3. ‘Embarked’ was missing values for two passengers traveling together out of South Hampton.

* Amelia Icard

<https://www.encyclopedia-titanica.org/titanic-survivor/amelia-icard.html>

* Mrs. George Nelson Stone

<https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html>

## Graphical Representations

1. Pandas and Matplotlib were used to create visual depictions of the data through the exploration and analysis phase.

# Data Summary

Captured from the data provider Kaggle & Udacity.

Data Descriptions:

Our data consists of passenger data stored in a CSV file type with 891 rows across 12 columns. This is example data for the first 5 rows by passenger ID. Dtypes includes categorical, object, 64 bit float and integer data points.

Built in functions for acquiring this include .head() and .info().



There are 890 rows of data with 712 non-null entries with exception to the following data columns or categories;

* PassengerID
* Age
* SibSp
* Parch
* Fare

These include *NaN* or null values that could be addressed by amending the mean values in place of zero’s, by removing/omitting zero values or by not including those categories in our analysis.

Variable Descriptions:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Variable** | **Description** |
| **survival** | Survival  (0 = No; 1 = Yes) | **sibsp** | Number of siblings/Spouses Aboard |
| **pclass** | Passenger Class  (1 = 1st; 2 = 2nd; 3 = 3rd) | **parch** | Number of Parents/Children Aboard |
| **name** | Name | **ticket** | Ticket Number |
| **sex** | Sex | **fare** | Passenger Fare |
| **age** | Age | **cabin** | Cabin |
|  |  | **embarked** | Port of Embarkation  C = Cherbourg; Q = Queenstown; S = Southampton |

Notes and Comments:

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)

If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored.

Definitions used for Sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger aboard Titanic

Spouse: Husband or Wife of Passenger aboard Titanic (Mistresses & Fiancés Ignored)

Parent: Mother or Father of Passenger aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

# Appendix

The data for this assignment was provided by Udacity but originated from Kaggle an online interactive Data Science community.

"Titanic: Machine Learning from Disaster." *train.csv*. N.p., 2 Feb. 2012. Web. 17 June 2016.

< <https://www.kaggle.com/c/titanic/data>>