RMS Titanic Analyst Project

Version: 2 | Student: Alex Sewell | [as643g@att.com](mailto:as643g@att.com) | 7/6/2016

*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 of April 15th, 1912 and 1,502 passengers and crew died.

For this analysis, we’ll focus on biological factors related to survivability using only age and sex, including;

* Univariate (1D) Analysis – Can we quantify age, gender and survival for further analysis?
* Bivariate (2D) Analysis – Can we determine male vs. female survival?
* Bivariate (2D) Analysis – Can we quantify adult men, women and child survival counts?
* Multivariate (3D) - Is survivability more likely if you’re a man, woman or child?

# Exploratory Analysis

## *Age - Univariate Distribution*

Univariate analysis of ‘Age’ of passengers revealed the mean age to be 29-30 for 714 passengers with 177 NaN values present *(Fig. 1)*. Because ‘Age’ is a key data point for this analysis, we chose not to leave out NaN values by replacing them with the median age of 28 *(Fig. 2)*. Before the changes, the middle 50% of passengers were age ~20 to ~40 years old, but with the adjustments to NaN values they now range between ~25 and ~35 years of age and we don’t have to leave out the 177 passengers from the study.

|  |  |
| --- | --- |
| *Fig 1. - 177 NaN values present* | *Fig. 2 - NaN values replaced by median age* |

## *Univariate Counts – ‘Sex’ and ‘Survived’*

‘Sex’ and ‘Survived’ contain no NaN or anomalous values. We’re able to account for 891 passengers with each variable. From the Passenger by Gender chart *(Fig. 3)*, we can see that the majority of passengers were male and the male population is proportionately larger than the female population. From the Passenger Mortality Count chart *(Fig. 4)*, we can see that of the 891 passengers provided, the majority (550) died and 342 lived.

|  |  |
| --- | --- |
| *Figure 3* | *Figure 4* |

Now that we have a univariate understanding of age, sex and survival through the raw counts, we’ll look to explore and work towards our ultimate question about survivability and combine these points to gain additional insights through bivariate analysis.

## *Bivariate – Exploring Survivability by ‘Sex’ and ‘Age’*

Using a Regression Plot to explore the relationship between the two variables ‘Survived’ and ‘Age’ *(Figure 5)* we find that survivability was greater for younger passengers than older – regression line trending down left to right. What if we looked at survivability by sex?

In Figure 6 we see that survivability was higher for females than males, but when we look back at Figure 3 we see there were proportionately more males than females aboard the ship. Proportional survival rates may actually paint a different picture of male to female survivability, so we’ll need to investigate that as well.

|  |  |
| --- | --- |
| *Figure 5* | *Fig. 6* |

## *Multivariate – Combining ‘Age’ and ‘Sex’*

We’ll begin this section by creating a new variable called ‘Person’, which we’ll use to identify a passenger as a ‘Man’, ‘Woman’ or ‘Child’. However this raises a new fundamental question;

What can we label a child in order to achieve this separation?

Through online research we’ve identified an official document titled ‘Titanic Certificate of Clearance (MT 9/920f)’ that points to children being anyone 12 or less years old and adults as being anyone 13 or more years of age. We create a function called ‘def man\_woman\_child(passenger)’ and apply it to come up with the following;

While we’re not

## *Multivariate*

With univariate analysis supporting the bivariate and multivariate analysis we’re working towards, we’ll being the process of combining more than one variable to begin our bivariate analysis.

Revisiting our question; can we determine male vs. female survival? We’ll use a regression plot for age and assign sex as the secondary variable to compare the two trends seen in Figure 7.

|  |  |
| --- | --- |
|  |  |

## *Multivariate Analysis of Survivors by Persons*

Multivariate (3D) Analysis – What are proportion of survivors were men, women and children?

In order to determine this we needed to evaluate what constitutes a child. Research

|  |  |
| --- | --- |
|  |  |

# Data Cleaning

## Cleaning

1. ‘Ticket’ and ‘Cabin’ are omitted 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.

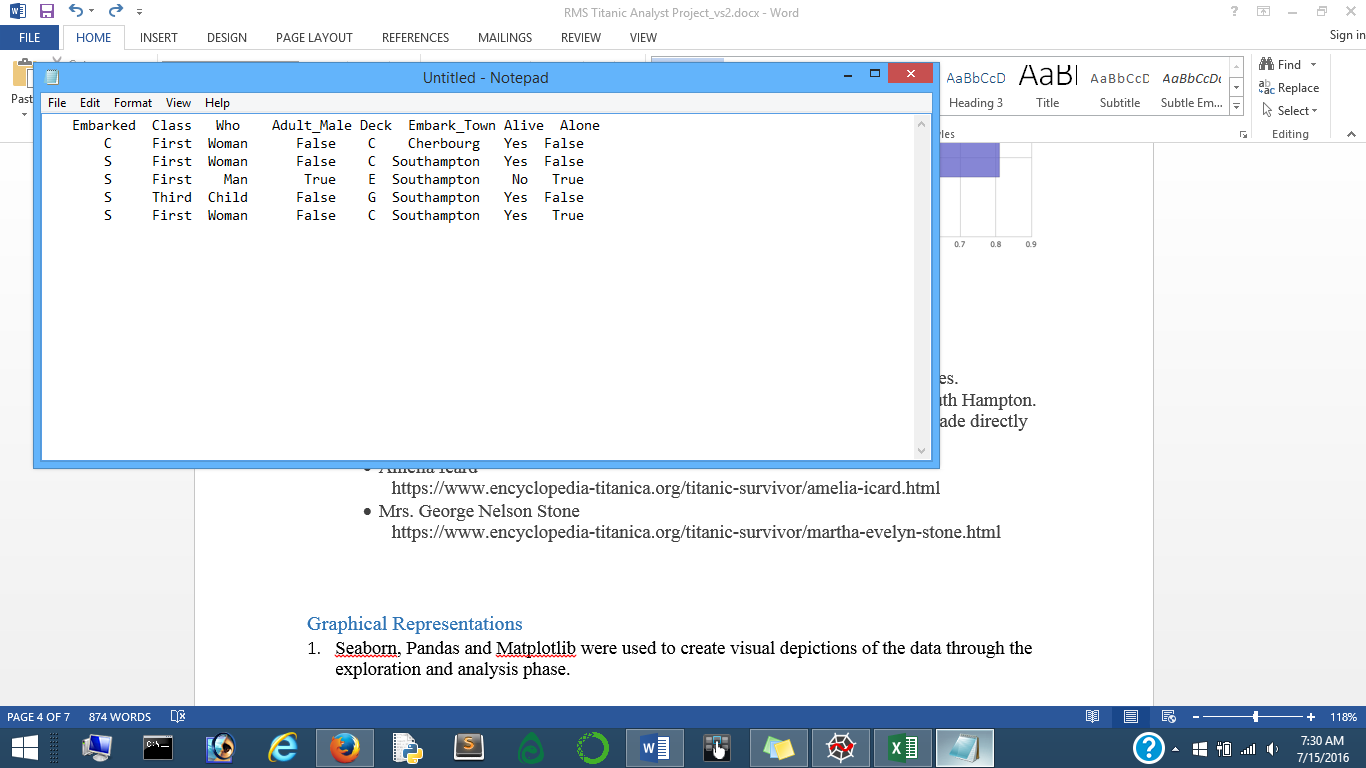
Given this only represented two missing values for the variable, updates were made directly to the file, without affecting the rest of the data.

* 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. Seaborn, 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. 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 family relation variables (i.e. sibsp & 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>>

Is there a relationship between the age of a passenger and passenger class. It’s possible with age that a passenger has had longer to accumulate wealth than that of younder passengers.

The age distribution for 3rd class is narrower (between 20 and 30 years of age) than with 1st class. This is true with and without the median age assignment to NaN values.

The median age for 1st class passengers traveling alone is \_\_\_\_, while the median age for 3rd class passengers traveling alone is \_\_\_\_. The average value for 1st class tickets vs. 3rd class tickets is \_\_\_\_\_

|  |  |
| --- | --- |
| *177 NaN values present* | *NaN values replaced by median age* |