RMS Titanic Analyst Project

Student: Alex Sewell | as643g@att.com | Version: 11 | 8/04/2016

Udacity Introduction to Programming

Choose Your Path – Data Analyst

Introduction

Objective: Derive insights from the data provided by Kaggle/Udacity for the RMS Titanic. 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.

Approach: For this analysis, we'll focus on biological factors and environmental factors;

- Univariate (1D) – Quantitative review of variables used in this analysis.
- Bivariate (2D) – Survival rates between male and females.
- Bivariate (2D) – Survival rates and distribution among passenger ages.
- Multivariate (3D) – Is survivability more likely if you're a man, woman?
- Multivariate (3D) – Is survivability more likely if you're a higher class passenger?

The tools we'll be using include;

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

Initial Data Staging

Data Specs: titanic.csv file containing Passenger data for 891 passengers. From initial review we see the data is mostly equal in passenger count across each variable with exception to 'Age' and 'Cabin'. All changes to the data are described in line with each analysis of this report.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Rows	891	891	891	891	891	714	891	891	891	891	204	889

Cleaning & Transformation

1. 'Name', 'Ticket' and 'Cabin' – Excluded. Cabin was missing 77% of its values and PClass could be used in about the same manner as 'Ticket' so we've left that out.
2. 'Age' has 177 NaN values – Significant to our study so the median age 28 was used in place of NaN values.
3. 'Survived'- We assigns a new variable 'Mortality' to represent 'Survived' with '0' and '1' renamed 'Died' and 'Lived' respectively.

Exploratory Analysis

We'll begin our univariate analysis with passenger 'Age'. The mean age using the raw data is 29-30. However, this represents only 714 of the 891 passengers given 177 (~20%) NaN values present as seen in Figure 1. Because 'Age' is a key data point for this analysis, we cannot leave out NaN values. Instead, we'll replace them with the median age of 28 and return a new visual under Figure 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.

Figure 1

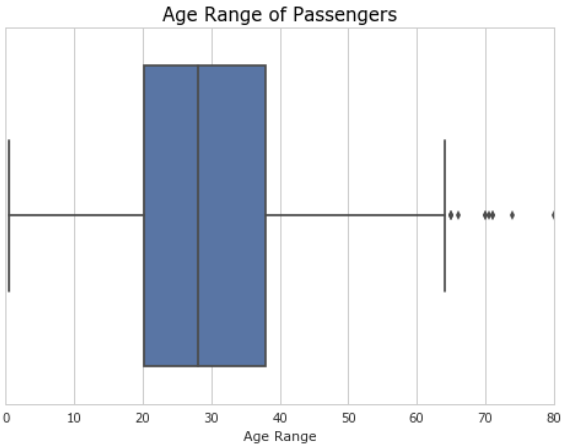
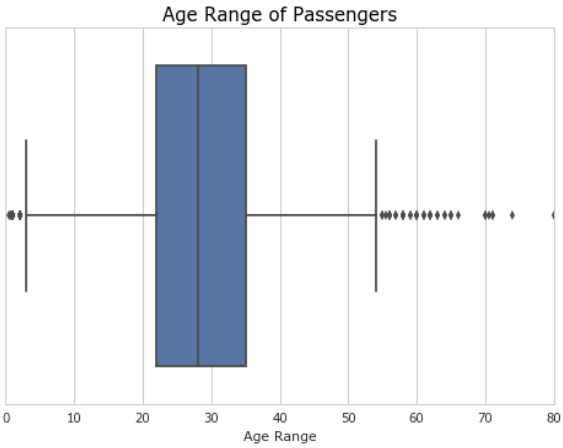


Figure 2



From this we can see that most passengers were at the mid-point of their life with some outliers spanning less than 1 year old and exceeding 60 years of age. With 'Age' cleaned up, we can move on to analyze the univariate variables of 'Sex' and 'Survived'. There are any NaN or anomalous values present and we're able to account for 891 passengers with each variable.

The majority of passengers were male ~50% and Figure 4 tells us that the majority of passengers (550) died and 342 or 38% lived. Now we'll move forward with reviewing 'Pclass' as our socioeconomic status as it may be a a potential influence on survivability.

Figure 3

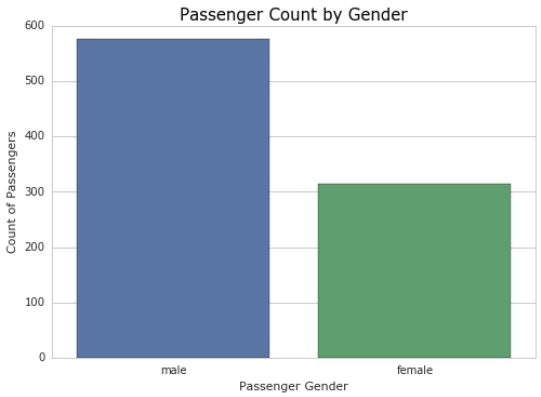
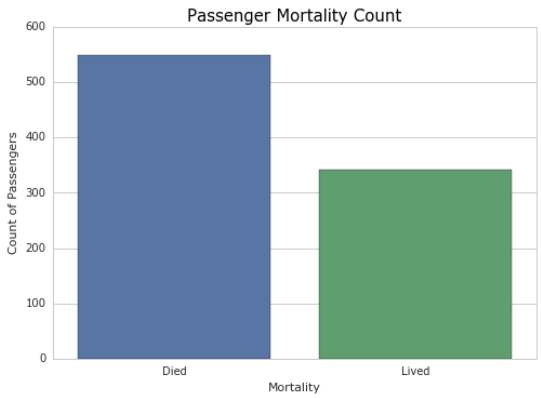
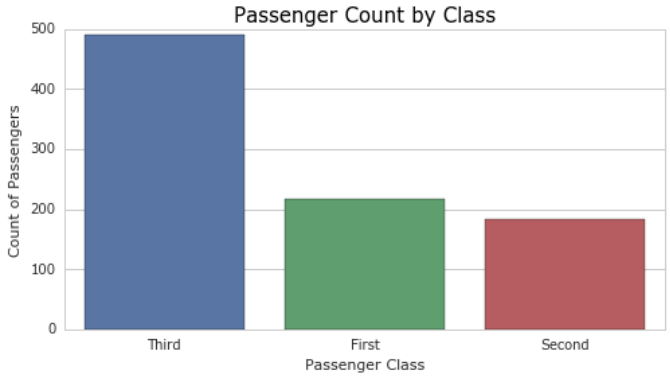


Figure 4



Upon inspection of we find the Third Class has the highest passenger count, being more than twice that of First Class or Second Class as seen in Figure 5.

Figure 5



We can already begin to predict certain insights from the univariate data. For example with the majority of passengers being male, one could expect the majority of deaths to be male. With the passenger count of 3rd Class being greater than that of 1st and 2nd Class passengers combined, one could reasonably expect the majority of those deaths to be among 3rd Class passengers. Let's go on to explore these relationships.

Age, Class and Survivability (Bivariate and Multivariate)

Using a Regression Plot let's explore the relationship between the two variables 'Survived' and 'Age' in Figure 6. Here we find that survivability was greater for younger passengers than older – regression line trending down left to right.

Figure 7 we use violin plot and see a greater density of passengers dying between 20 and 30 years of age. We also see more passengers lived 18 and younger, and 60 and older, than those that died. Could we say this is a product of human tendencies to preserve youths, women and elderly? Yes, but that remains an assumption.

Figure 6

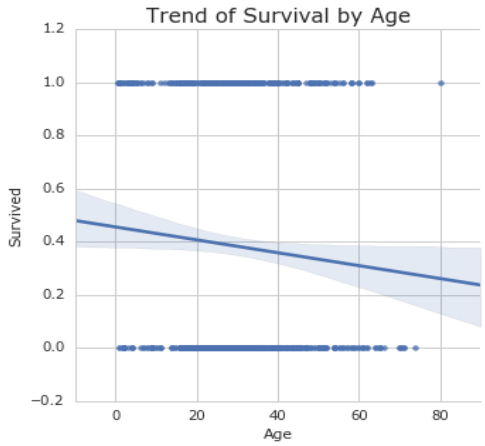
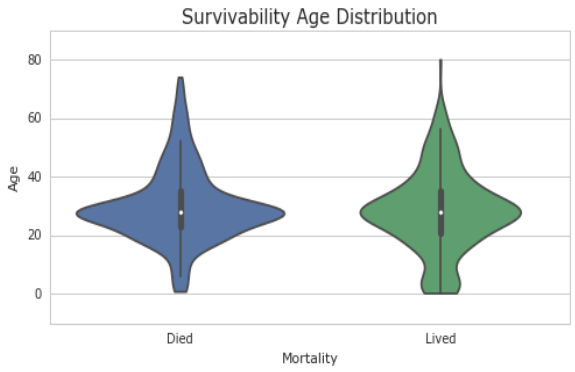


Figure 7



Let's now turn to socioeconomic status using 'Pclass' to derive insights around survivability. Consider ages within each passenger class using Figure 8. The age group densities among 2nd and 3rd Class are closest to the median of 28 than 1st Class passengers. 1st Class passengers show a higher density of passengers between 30 and 40 years of age and more 65 and older than 2nd or 3rd Class but, fewer passengers 18 and younger. One could expect this to reflect when we consider survivability/mortality.

Figure 8

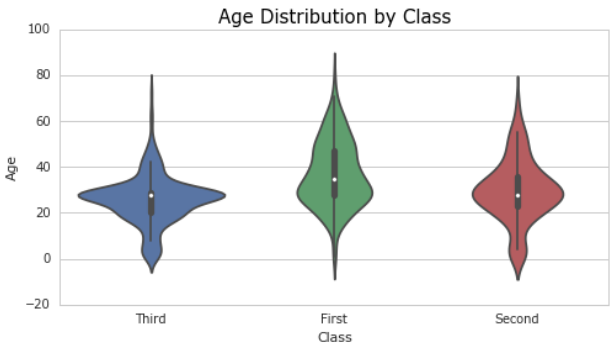


Figure 9

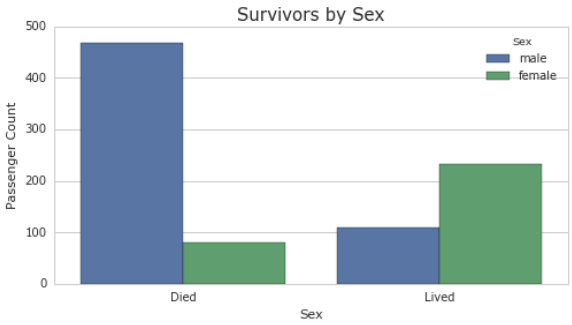
Figure 9 illustrates our expectations are accurate, but also shows a proportional difference in mortality age density with 3rd and 2nd Class passengers, with 2nd Class visibly showing the greatest disproportion. 1st Class mortality ages appear more balanced.

This balance of 1st Class age densities vs. the imbalance of 3rd and 2nd Class densities could point to the influence of class on survivability. Let's consider 'Sex' and work to put those in combination with other variables.

Adding in Sex with Age, Class and Survivability

Beginning with raw count comparisons in Figure 10, we can see a vast majority of those that died were men and that more females lived than men. However, we need to consider proportionality to really understand survivability of women vs. men.

Figure 10



For male to female survival, there were proportionately more male deaths than females as seen in Figures 11 and 12.

Figure 11

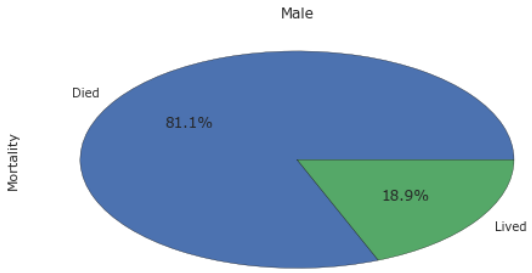
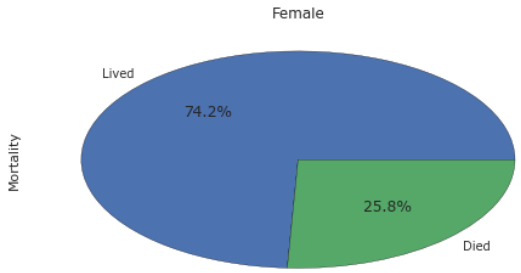


Figure 12



You could attribute this to female resilience or male role in their protection. You could infer natural human tendency to preserve the life of children and women. However, we cannot conclude that with the data given, we're relying on subjective human experience to speculate about that point.

Let's now consider class with the gender of passengers. Figure 13 reveals the majority of passengers were male with the highest proportion being 3rd Class passengers. We also see that there were more males than females across each class.

Figure 13

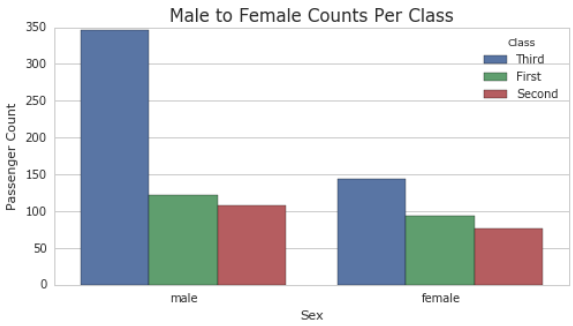


Figure 14

In Figure 14 we see that our previous predictions were correct. The majority of passenger deaths were 3rd Class members. We can also recognize that there is a noticeable disproportion of 3rd Class passenger deaths.

- 37% of 1st Class Passengers died while 63% survived
- 53% of 2nd Class Passengers died while 47% survived
- **76%** of 3rd Class Passengers died while 24% survived

Clearly this displays a disproportionate survival count. Comparing the classes, age and survivability Figure 15 a clear trend in favor of younger passengers, of higher class held a higher rate of survival. Figure 16 takes sex and age into account and reflects an increasing trend for female over male survival with a decreasing trend of male survival, with outliers. For example, there are females that were of age 50 to 60 that died and male(s) age 80 or older that survived, but they are not standard when compared with the rest of the data.

Figure 15

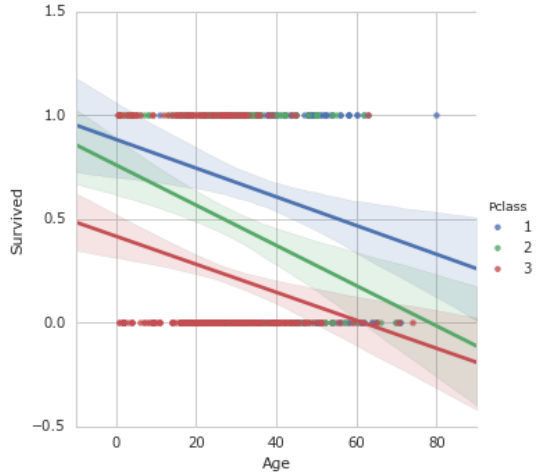
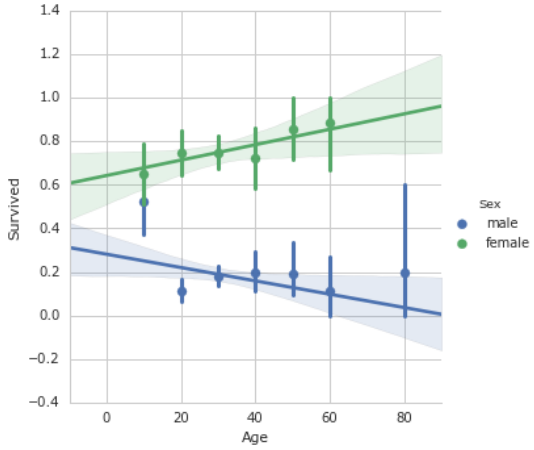


Figure 16



Conclusion & Additional Questions

The highest survivability of passengers rests with 1st Class females and that their survivability increased with age. Proportionately more females (74.2%) survived than men (18.9%). Among passenger classes 76% of 3rd Class passengers died while 63% of 1st Class passengers survived. Comparatively, 39% more passengers died in 3rd Class than 1st Class. Among youths, ages 18 and younger and elderly, ages 60 and older they were also more likely to survive based on *Figure 6*, but do not represent a large portion of the population. Regardless of sex or age, you were more likely to survive in 1st Class than you were of 3rd Class and that increases among females over males.

For predictability modeling we would need a more complete set of data. For example, this only contains 891 of the 2,222 total passenger/crew count. Also, having all of the cabin ID's would help to know whether passenger cabin location contributed to survival. Knowing the location of lifeboats relative to passenger cabins could have played a role in survivability. According to 'Titanic Facts', the ship left port with only 33% life boat capacity and of the lifeboats used, most departed under capacity. How close were the life boats to passenger cabin locations? Where the passengers in their cabins while the life boats were being deployed? We could guess no, but without knowing that we wouldn't be able to decide whether cabin or deck locations had anything to do with survivability.

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().

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florenc	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily Ma	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

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
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
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>

- Titanic's Certificate of Clearance (MT 9/920f) establishes adult status upon reaching the age of 13.