Titanic Lab

Jasper Lemberg

***Abstract*— By looking at the titanic data set, one could tell what groups were more likely to survive the sinking ship. Ultimately, higher class passengers, women, and children survived most often, whereas third class men were least likely to survive.**

I. INTRODUCTION

This lab used Python and the titanic data set in order to uncover key insights and information about a number of passengers sailing on the R.M.S. Titanic's only transatlantic voyage. To do this, I first cleaned the data set by first removing a number of NA values, most of which came from the "Age" feature and initialized a new feature called "NotAlone". After, I calculated a number of uni-variate and bi-variate statistics, including the mean and median of one feature and correlation of two separate features. This soon led to more in-depth analysis of conditional probabilities and conditional survival rates. Finally, a number of charts were created to better visualize and explain the large data set.

When I started working with this data set, I already had a number of assumptions about the data. To start, I assumed that first and possibly second class would have a higher survival rate than third class. This is because they are smaller classes and had better access to lifeboats because they were higher up on the Titanic itself. Also, I assumed that ages and genders of the survivors would be fairly mixed. The escape off of the sinking ship must have been chaotic and hectic, so it is doubtful that there would have been any attention paid to who got on the lifeboats.

Despite these assumptions, there were a number of findings that surprised me while I completed the lab. Firstly, although it was not surprising that the first class' mean and median fares were the highest amongst the classes and that the third class' mean and median fares were the lowest, it was interesting to note that there was a fairly noticeable difference between the mean and median fares for each class. For example, the mean fare for first class passengers was 87.96158225806447, while the median was 69.3. Outliers on the higher end of each class would most likely explain this trend, as they influence the mean much more than they influence the median. Also, while I assumed that more first class passengers would survive compared to the other classes (which was a correct assumption), I was not expecting more first class passengers to survive (120) versus not survive (only 64). Again, the first class passengers' location on the ship relative to the lifeboats and their "priority" as important passengers played a large role in this. Despite this, most of the passengers paid had ticket fares below 100. While this seems noteworthy at first, it is worth noting that most of the fares, including all of the second and third class fares, were below 100, so only a select number of first class passengers would be featured in the higher echelon of that histogram. Therefore, the histogram is skewed right.

II. DATA

The Titanic data set consists of initially 891 observations across 12 features, although this was eventually changed to 714 observations across 13 features by the end of data prepossessing. In total, five features consisted of string types and the rest were some sort of numeric type (either integers or, most likely, floats).

The first feature is "PassengerId", which is essentially just a count of what row in the data set the observation is. Beyond that, this feature has no real value, especially because Pandas data frames already keep track of this.

"Survived" describes whether a passenger survived or not and consists of an integer value that is either 1 or 0. If the passenger survived, the feature shows a 1. If the passenger did not survive, the feature shows a 0.

"Pclass" describes a passenger's class and consists of an integer value that is either 1, 2, or 3 (first class, second class, or third class respectively). If the passenger was in first class, the feature shows a 1. If the passenger was in second class, the feature shows a 2. If the passenger was in third class, the feature shows a 3.

"Name" is just a string consisting of the passenger's name. The format for this is "[Last Name], [Title]. [First Name and Other Names]".

"Sex" is also a string consisting of the passenger's sex, either "male" or "female".

"Age" consists of the passenger's age at the time of the Titanic's departure as an integer. This feature had the most NaN values, any observation with these values was removed.

Both "SibSp" and "Parch" consist of integer values. It is either 0 or 1, which is equivalent to True and False like for the "Survived" feature. If these two features both equal 0, then the integer-type feature "NotAlone" also equals 0. Otherwise, "NotAlone" equals 1.

"Ticket" is a string consisting of the passenger's ticket ID. They vary from just numbers to numbers, letters, and some punctuation.

"Cabin" consists of a string that represents the passenger's cabin number. This also featured a lot of NaN values, but these were not removed.

Lastly, "Embarked" consists of a single letter representing where the passenger departed from. This feature has only Cs, Qs, and Ss.

III. RESULTS

A. What helped passengers survive?

One of the most important factors for survival was class. Of the three classes, first class had the most survivors (122 for first class, 83 for second class, and 85 for third class). It also had the best ratio of surviving passengers to total passengers, with over half of the first class passengers surviving. Fare was also a key factor here, but mostly because it was higher for first class (surviving) passengers. For example, the mean and median fares for each class were 87.96158225806447 and 69.3 for first class, 21.47155606936416 and 15.0458 for second class, and 13.229435211267623 and 8.05 for third class respectively.

Certain titles were found to be more helpful for surviving than others. Several titles had a perfect survival rate, including Ms., Lady, Mlle. Mme., Countess, and Sir. The same can be said for perfect death rate (Rev., Jonkheer, Don., and Capt.). Interestingly, most of the "perfect" roles were women, whereas the worst ones were mostly men. That being said, these roles also had a very small number of people as well.

B. Did women and children first apply?

Yes, "women and children first" did apply when exiting the ship. Women survived at a 96.47058823529412 percent rate for first class, 91.8918918918919 percent rate for second class, and 46.078431372549017 percent for third class respectively. On the other hand, men survived at a 39.603960396039606 percent rate for first class, 15.151515151515152 percent rate for second class, and 15.019762845849802 percent rate for third class. Thus, women survived at a significantly greater rate than men.

Children also benefited from the "women and children first" policy. Third class children under the age of 10 survived at a 43.18181818181818 percent rate, which is just below the 46.078431372549017 rate for women (2.89661319073 less) and well above the 15.019762845849802 percent rate for men (28.162055336 greater).

C. Was the correlation analysis satisfactory?

Unfortunately, the correlation analysis was not satisfactory. Only one sets of features had medium correlation. NotAlone and SibSp have 0.63 correlation, but every other correlation score is well below 0.3. If Pclass was added to the correlation analysis, maybe there would be better results, especially between that and Fare.