Data Analytics 350 Titanic Lab

Matthew Rinker

***Abstract*— In this report I will chronicle the process I went through to complete the Titanic Lab. This will begin with a brief discussions of my preconceptions going into the lab. Next, I will discuss the data and discuss the characteristics and challenges that arose from this data. Finally, I will discuss the methods I used to analyze the data and what results and conclusions I drew from that Data.**

I. INTRODUCTION

Going into this lab I expected this to be a cursory examination of the passenger manifests of the Titanic. I expected discussion of mortality rates among the different classes and expected to be able to predict the survival and mortality rates almost purely on economic background. These preconceived notions came about from my general learning of the incident as in history classes it is often taught that the biggest tragedy of the titanic was that all the people who were not wealthy died. I find it to be a very significant finding that this is not entirely true. In fact, I found that there was a focus on saving the women and children first and then the men (starting with the wealthy). I also found it interesting that there did not seem to be that much of a correlation between survival and whether or not people were traveling alone. My personal assumption had always been that they had tried to keep families together.

II. DATA

The data itself contained 891 rows and 12 columns. Of those 12 columns only a few were of interest to me. The columns contained in this data set indicated a passenger ID number, whether or not they survived, the class of the passenger, Name, Sex, Age, Number of siblings and spouses on board, Number of parents or children on board, ticket number, fare, cabin, and the port where they embarked. For the most part I focused entirely on the numerical columns with the exception of the name and sex columns.

The data required some cleaning after acquiring it. In fact, this was the biggest issue with the data set. When looking through the data I found that there were many passengers for whom there was no age data. Since my analysis would revolve around age for a large part of my analysis I had to drop these rows. This ended up being a staggering 177 rows from the original data set which was comprised of 891 rows. I was left with 714 rows. When I had finished dropping incomplete rows I ended up having gotten rid of 19.7% of my data set.

The last bit of editing I did to the data set was to create a new column called "NotAlone" this column indicated whether or not the passenger was traveling alone on the ship by using the Parch and Sibsp columns.

Class	Passengers
C1	186
C2	173
C3	355

Fig. 1. Class Sizes

Class	Measurement	Age	Fare
C1	Mean	38.23	87.96
	Median	37	69.3
C2	Mean	29.88	21.47
	Median	29	15.05
C3	Mean	25.14	13.23
	Median	24	8.05

Fig. 2. Age, Fare Mean and Median by Class

III. RESULTS

To start, I conducted a cursory examination of the data. To do this I separated the data into 3 smaller subsets. From this I found that third class was indisputably the biggest class. This goes along with my preconceptions of the data as I had assumed that the first and second classes would be filled with the wealthy elite.The class sizes are depicted in Fig. 1: To define the splits I used the Pclass column which represented the class the passenger was traveling in. I then used pandas to calculate the mean and median values of the Age and Fare for each class. My findings are depicted in Fig. 2:

We can see from Fig. 2 that the average age is highest in first class as well as the average price. Unsurprisingly, the people in the lower classes were on average younger and were paying much less than everyone else.

Next, I split the original data set again into two subsets. For this split I used the "Survived" column to split my data. From this split I was able to see the number of passengers who survived the tragedy and the number of passengers who did not. My findings are depicted in Fig. 3 as well as a graphical depiction in Fig 4.:

From this data we can clearly see that the vast majority of passengers did not survive the incident. This clearly shows how tragic an incident we are exploring as over half of the passengers did not survive. For reference, only 40.6% of

Status	Passengers
Survived	290
Not Survived	424

Fig. 3. Passengers by Survival Status

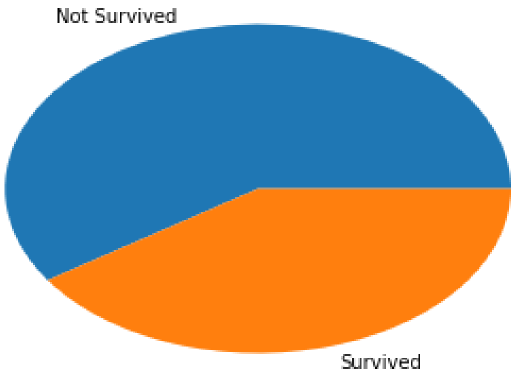

Fig. 4. Passengers by Survival Status

Class	Survived	Not Survived
C1	122	64
C2	83	90
C3	85	270

Fig. 5. Survival Numbers by Class

passengers survived the tragic accident.

To gain more insight on the figure shown above I once again broke the data set down. This time, I split on both the "Survived" column and the "Pclass" column. This gave me figures for how many people of each class survived and how many did not. My findings are depicted in Fig. 5:

From the table in Fig. 5 we can see that the vast majority of fatalities occurred in Third class. However, the mortality rate in first and second class were a lot higher than I had anticipated.

Next, I created a correlation Matrix (Fig. 6) to show the correlation (if any) between my variables. I found that there were some obvious correlations in the data set. There are some obvious correlations in this dataset: The number of siblings on board is negatively correlated with age. The number of parents and children on board is also negatively correlated by age. This seems to mean that there were a larger amount of children on board than parents, leading to the assumption that of the families on board there were most likely more children than parents. Also, survival rate seems to be negatively correlated with age suggesting that although it was a weak correlation that the survival of children was prioritized over adults. Additionally, there is a weak correlation between fare and survival. Also, there is an obvious correlation between NotAlone and both siblings on board and Parents on board. Finally, there is a weak correlation between NotAlone and survival.For the most part however, there correlations are all relatively weak. Even the correlations between SibSp, Parch, and NotAlone are weaker than I would have thought.

Next, I calculated the Standard Deviation of the Age and Fare Columns. My findings are in Fig 8.

These figures were unsurprising to me given the wide spread of data and the outliers generated by the extremes in First and Third classes on the high and low side of the spectrum respectively.

Next, I calculated the interquartile ranges of all of my variables however the only two variables which this value had any meaning for was Age and Fare. The values are displayed in Fig. 9.

From these results we can see that there is a greater spread in Fares paid than there is in Age. The maximum difference in Age for the middle 50% of the data is 17.875 while the maximum difference in Fares is 25.325. We can see that since the mean age is 29.70 and given that the first quantile is 20.125 and the third quantile is 38 that the mean lies within the third quantile. Thus, the majority of data lies underneath the mean and the data is is grouped towards the lower end of the range. For the mean Fate we see the value is 34.69. Q1 for fare is 8.050 and q3 is 33.375. Again, we see that there are some much higher outliers in the fare column whick skew the average, and as such the majority of the data is again below the mean, in fact the entirety of the data within our interquartile range is below the overall mean.

Next, I moved on to calculating the conditional probabilities of survival given a specific class and gender. In this section, I found the most interesting findings of my exploration of the data. My findings are located in Fig. 10.

From this I found that men survived disproportionally less than women regardless of their class. This suggests that women's survival was prioritized over men regardless of their wealth. This is something that is contradicted in most history classes. Next I calculated the probability of a child of 10 years or younger surviving the tragedy. I found that 59% of children of this age range survived the tragedy. This is also significantly more than the survival rate of even the first class men.

Moving forward I wanted to calculated the expected value of a fare given the class of the passenger. My findings from this calculation is recorded in Fig. 12.

The findings from this calculation represent a huge disparity in wealth between the passengers of First class and everyone else on board the ship. The Expected value for a first class passenger is 66.29 higher than the expected value for a second class passenger. This coincides with my preconceived notions surrounding the wealth of the first class passengers.

Next, I looked at the title of each passenger and calculated the survival rates of each passenger given their respective title. My findings are in Fig. 13. In Fig. 14 I have included a bar chart which shows the top survival rates by titles. In the creation of this chart I have eliminated those titles which only had a single passenger as I felt this skewed my data.

As we can see there is a large disparity between the noble titles and the titles of the common people. Most women survived (as supported by the last section's data) however, we can see that those titles that indicate noble birth all survived. Additionally, we can see that those who were of religious titles all perished presumably due to carrying out the last rites of the people still on board and leading prayers. Additionally,

	Survived	Age	SibSp	Parch	Fare	NotAlone
Survived	x	-0.077221	-0.017358	0.093317	0.268189	0.196140
Age	-0.077221	x	-0.308247	-0.189119	0.096067	-0.192870
SibSp	-0.017358	-0.308247	x	0.383820	0.138329	0.629818
Parch	0.093317	-0.189119	0.383820	x	0.205119	0.577524
Fare	0.268189	0.096067	0.138329	0.205119	x	0.260139
NotAlone	0.196140	-0.198270	0.629818	0.577524	0.260136	x

Fig. 6. Correlation Matrix

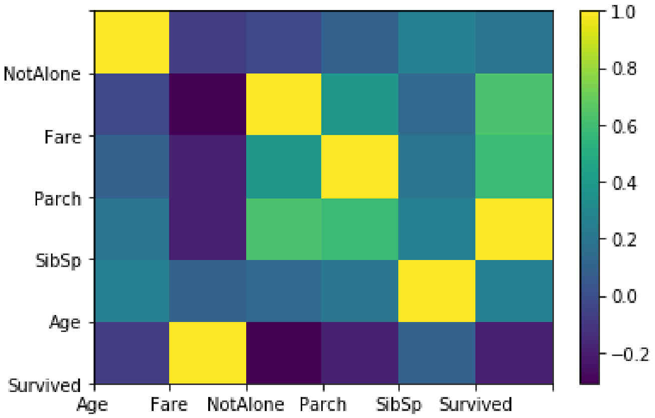

Fig. 7. Heatmap of Correlation Matrix

Measure	Value
Age	14.53
Fare	52.92

Fig. 8. Standard Deviation of Age and Fare

all those with nautical titles perished on board seemingly correlated with the maritime laws that the Captain may not leave the ship while there are still passengers on board.

Finally, I also created a figure which shows the correlation between Fare price and survival rate. I believe that this supports the notion that those who paid more survived more. This may most likely be attributed to the benefit of having a cabin that was closer to the decks and thus safety.

Variable	Value
Age	17.875
Fare	25.325

Fig. 9. Interquartile Ranges

Measure	Rate
P[Survival	Female and C1]	0.96
P[Survival	Female and C2]	0.91
P[Survival	Female and C3]	0.46
P[Survival	Male and C1]	0.40
P[Survival	Male and C2]	0.15
P[Survival	Male and C3]	0.15

Fig. 10. Conditional Probabilities of Survival

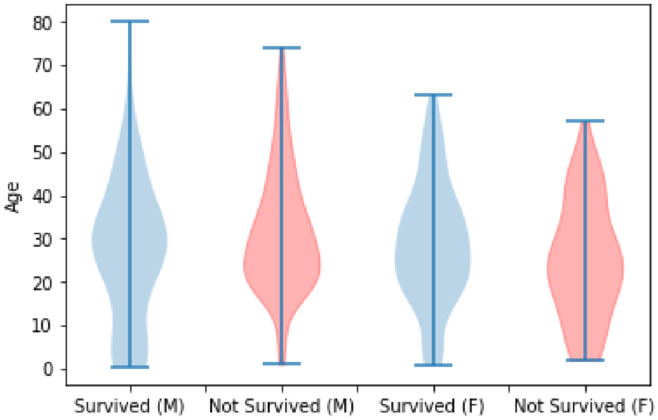

Fig. 11. Violin Chart showing the distribution of Survival by Age

Class	Expected Fare
C1	87.55
C2	21.26
C3	12.73

Fig. 12. Expected Fare

Title	Survival Rate
Mrs	0.787
Miss	0.719
Mr	0.168
Master	0.583
Mme	1.0
Ms	1.0
Major	0.5
Lady	1,0
Sir	1.0
Dr	0.5
Mlle	1.0
Col	0.5
Countess	1.0
Don	0
Rev	0
Capt	0
Jonkheer	0

Fig. 13. Survival Rates by Title

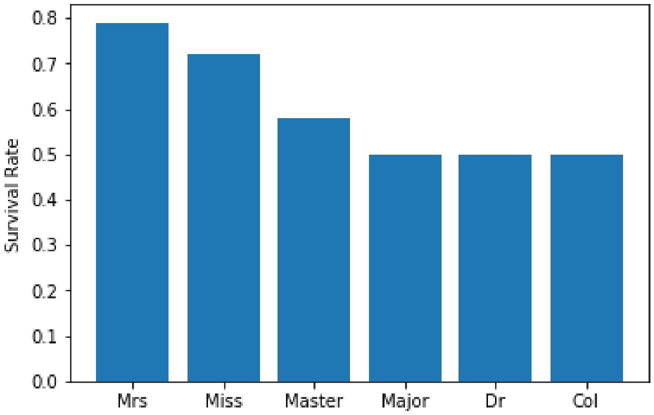

Fig. 14. Bar chart showing top survival rates

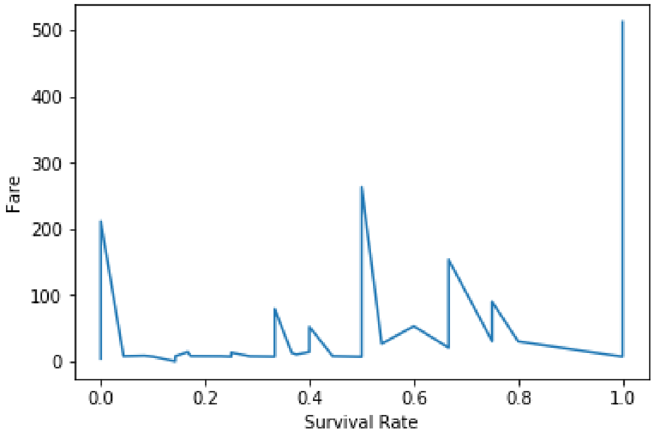

Fig. 15. Relationship between Fare and Survival Rate

IV. CONCLUSION

From my exploration of the data related to this tragedy I believe that some of my preconceived notions were correct. I do think that this data supports the notion that "Women and Children First" was practiced during this tragedy. Specifically, because of the high survival rate of women and children. Additionally, I believe that the wealth of each passenger and therefore the higher location of their cabins allowed for more wealthy passengers to reach the lifeboats first and get off the ship. I do feel that the methods I used could be improved on, specifically during the calculation of the correlation matrix. I believe that the methods I used must not be optimal as I did not get the result I expected. I do believe that there are some elements of selection bias in my analysis, specifically during the title portion of my analysis. I believe that the survival rates of specific titles could be attributed to the number of people who had each title on the ship and I corrected for that with my bar chart.