Titanic Data Analysis Report

Chen Qingqing[1]

[1] *Udacity Data Analyst Nanodegree Program Student, Singapore*

Feb 20, 2018

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. In this project, I am going to create a visualization that shows the demographics or passenger information between those passengers who survived and those who died.

1 Choose a Data Set

A data set is choose from the Data Set Options document (**Titanic Data**). In this dataset, it contains demographics and passenger information from a subset of the 2224 passengers and crew on board the Titanic.

2 Exploratory of Data Analysis

2.1 Data description

The variables on the dataset are Passenger ID, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked.

1. Pclass: Passenger Class (1 = 1st (Upper); 2 = 2nd (Middle); 3 = 3rd (Lower))
2. Survived: 0 = No; 1 = Yes
3. Age: age is in Years; Fractional if Age less than One (1); If the Age is estimated, it is in the form xx.5
4. SibSp: Number of Siblings/Spouses Aboard
5. Parch: Number of Parents/Children Aboard
6. Fare: Passenger Fare (British pound), is in Pre-1970 British Pounds; Conversion Factors: 1 = 12s = 240d and 1s = 20d
7. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

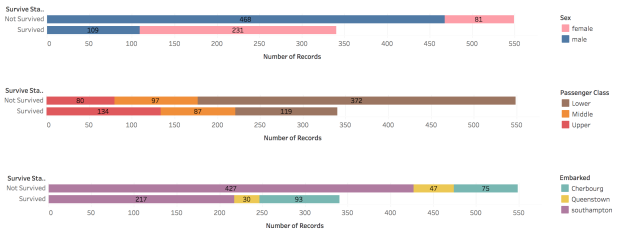
2.2 Data Wrangling

After visually looking at the dataset, I started data wrangling. First, I checked whether there was any null or duplicated passenger IDs, and dropped them. Then I modified columns of the dataset with the following steps and finally saved the modified dataset to a new csv file named "**clean.csv**".

1. Pclass column: Replace number "1, 2, 3" with "Upper, Middle, Lower" separately;
2. Survived column: Replace number "0 and 1" with "Not survived, survived";
3. Age column: Divide age to different ranges (0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80)
4. Embarked column: Replace capital characters "C, Q, S" to "Cherbourg, Queenstown, Southampton" separately;

3 Create Visualization

In this section, a visualizaition was created using **Tableau** to explain and help lead a reader to identify some key insights into the dataset. And there were two versions of the visualizaiton. One was initial version which could be found through this link: **First Version**, and another one was the final version which had some



Survive Status	Sex	Number of Records	% of Total Number of Records along Pane (Down)
Not Survived	female	81.0	14.75%
	male	468.0	85.25%
Survived	female	231.0	67.94%
	male	109.0	32.06%

Survive Status	Passenger Class	Number of Records	% of Total Number of Records along Pane (Down)
Not Survived	Lower	372.0	67.76%
	Middle	97.0	17.67%
	Upper	80.0	14.57%
Survived	Lower	119.0	35.00%
	Middle	87.0	25.59%
	Upper	134.0	39.41%

Survive Status	Embarked	Number of Records	% of Total Number of Records along Pane (Down)
Not Survived	Cherbourg	75.0	13.66%
	Queenstown	47.0	8.56%
	southampton	427.0	77.78%
Survived	Cherbourg	93.0	27.35%
	Queenstown	30.0	8.82%
	southampton	217.0	63.82%

Figure 1: *Sex, class level and embarked place information for survived and not survived passengers*

modifications after getting the feedback from readers and it could be found through this link: **Final Version**.

Firstly, I put passengers basic information in my story line and added filters like "Survival status, sex, ticket number, and embarked place. This was in order to help reader to find the passenger easier. Then I used bar plot to show detailed difference between survived and not survived passengers in sex. classes, and embarked place. For total 340 survived passengers, 67.94% is female and 32.06% is male; and for total 549 died passengers, 85.25% is male and 14.75% is female. The survived passengers percentage of upper class is higher than middle and lower class passengers. Mostly, the passengers were embarked at Southampton.[Figure1]

Sex	Age Range	Avg. Age	Number of Records	% of Total Number of Records along ..
female	0-10	4.6	31.0	9.94%
	11-20	16.8	46.0	14.74%
	21-30	25.4	81.0	25.96%
	31-40	35.3	54.0	17.31%
	41-50	45.5	31.0	9.94%
	51-60	55.1	14.0	4.49%
	61-70	63.0	2.0	0.64%
	None		53.0	16.99%
male	0-10	4.0	33.0	5.72%
	11-20	17.7	69.0	11.96%
	21-30	25.4	149.0	25.82%
	31-40	34.9	100.0	17.33%
	41-50	45.3	55.0	9.53%
	51-60	54.8	28.0	4.85%
	61-70	64.1	14.0	2.43%
	71-80	73.3	5.0	0.87%
	None		124.0	21.49%

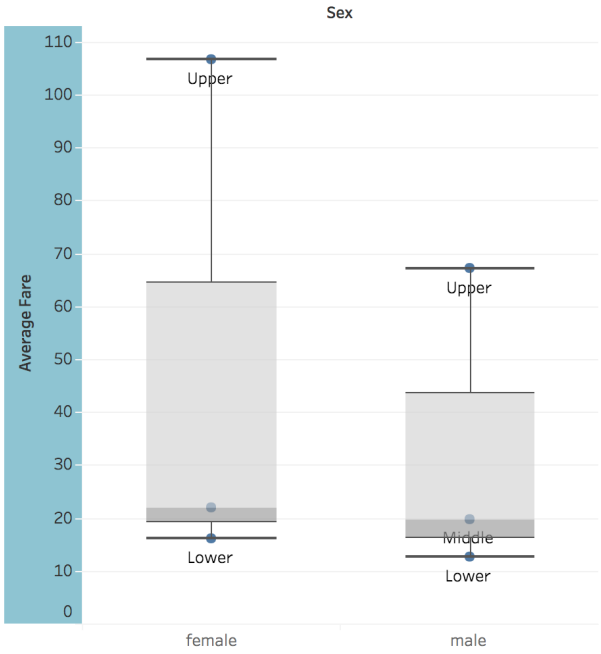
Figure 2: *Age range information for survived and not survived passengers*

For female, mostly the age is between 11 - 40 years old, the age in range of 21 to 30 years old has the highest percentage which is 25.96%; the age under 10 years old and above 60 years old is relatively less than other age ranges. For male, mostly the age is between 11-40 years old, the age in range of 21 to 30 years old has the highest percentage which is 25.82%; the age under 10 years old and above 60 years old is relatively less than other age ranges. [figure 2]

Then I used box plot and histogram to show the average fare between different genders and among different class levels. Because it is easier and more obvious for readers to see the highest, median and lowest average fares and it is clear for readers to compare the fares in different classes. The average fare of female in different class is higher than that of male, and the average fare for upper class passengers is higher than other two classes passengers in both male and female.

Mostly passengers bring the children together with them where there is 22.10% and 32.08% passengers is under 10 years old for not survived and survived passengers.

There are 27.3% and 21.60% siblings is under 10 years old for not survived and survived passengers.



4 Get Feedback and Improve the Visualization

4.1 Get Feedback

1. Advantage: There is a clear story line shown in the Tableau that I can understand all of the contents in the slides and all the calculation and plots are correct;

2. Disadvantage: It would be better if there will be some comments or conclusion for each slide;

4.2 Improve the Visualization

Based on the feedback, I add the conclusion in the slides to make it more understandable and I also add one picture in the title slide in order to catch readers' eyes.