Data Analysis of Titanic data

1. Data description

It ("titanic-data.csv") contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic. Three columns (variables) are selected for further analysis: Pclass, Sex and Age.

2. Brief summary of data under investigation

Table 1 summarizes the data under investigation, including the number of observations, the number of missing values (NaN), minimum value (min), maximum value (max), mean and standard deviation (std) if applicable.

	Survived	Pclass	Sex	Age
count	891	891	891	714
mean	-	-	-	29.70
std	-	-	-	14.53
min	-	-	-	0.42
max	-	-	-	80
# of NaN	0	0	0	177

Table 1: The summary of data under investigation. "Survived" refers to the survival in Titanic disaster, in which 0 corresponds to no survival and 1 to survival. "Pclass" refers to the ticket class, 1, 2 and 3 correspond to 1st, 2nd and 3rd class, respectively. "Age" refers to the passengers' age in years.

There is a considerable percentage of passengers (19.8%) missing the age in this data set. If these 177 observations were removed, the result may not be representative. Thus, it was decided to keep all observations for the following data analysis.

3. Data exploration

In this section, it attempts to answer the following question: What factors made people more likely to survive?

3.1 Sex

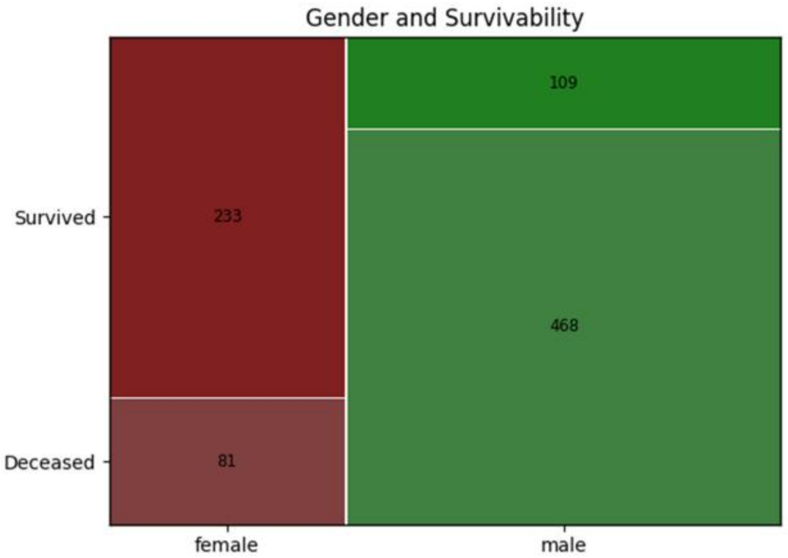

Figure 1. The mosaic plot of the survival versus sex.

As shown in Figure 1, there were more males than females among 891 passengers and as for the survival percentage, it is visible that higher percentage of female survived, possibly three times higher that of male.

3.2 Age

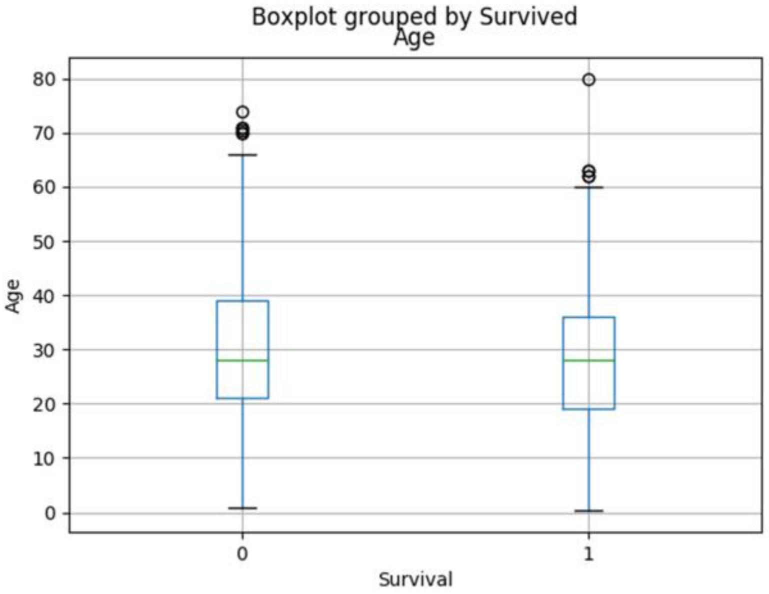

Figure 2. The boxplot of age grouped according to survival (0, 1)

It is clearly seen in Figure 2 that the average age for both the people survived and the people who did not survived is between 25-30. In addition, the people survivied is slightly younger than the people who did not survived because the IQR of the people survived located at a lower region than that of the people who did not survived. Overall, there is no significant difference in survival in terms of age.

3.3 Ticket class

Shown in Figure 3, there are more passengers in class 3, and then class 2, and the class 1 has the least number of people. Approximately more than half of the passengers are in class 3, which represents lower socio-economic status. (Note that the ticket class is a proxy for socio-economic status: 1st class, 2nd class, 3rd class corresponds to the upper class, middle class, lower class, respectively, based on the data description provided.) The survival rate among the people class 1 is higher than that of class 2, and the class 3 has the lowest survival rate.

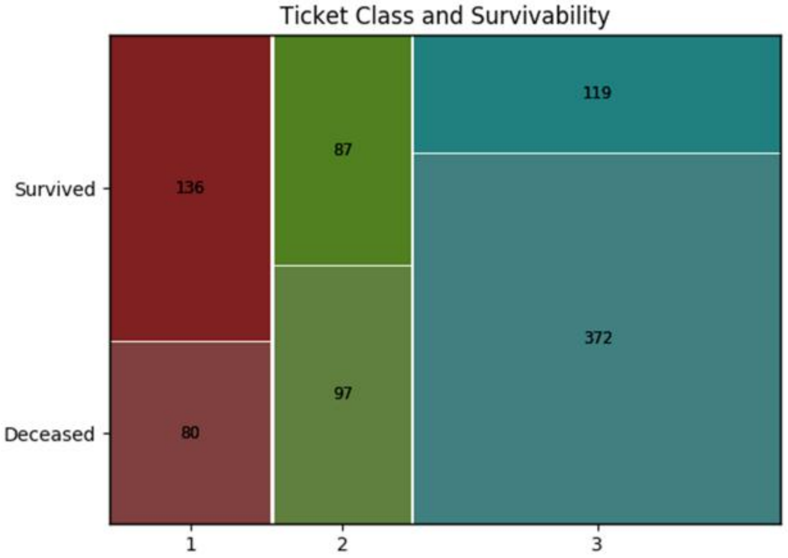

Figure 3. The mosaic plot of survival versus the ticket class (1, 2 and 3).

4. Conclusion

It is found that the passengers who are female and those who have 1st ticket class would have higher survival rate, indicated by the above statistical analysis. It should be noted that this data set contains only the information for 891 passengers out of all 2224 passengers. It is possible that majority of the passengers not in this data set did not survive, and thus their information got lost and did not remain in the record. In light of this, it would not be possible to draw any solid conclusion with statistical significance.