Udacity Project: Investigate a Dataset

Usman Rizwan

1 Introduction

The titanic data is analyzed in this report. The main question investigated is what are the factors that made people more likely to survive? The question is approached from a variety of different angles. We look at the survival rate of men vs. women irregardless of the passenger class they were on, then we look at the survival rate of men and women taking into account the passenger class they were in. Finally we look at the relationship between the age of the passenger and the survival rate.

2 Results

Lets look at the rate of survival of men vs. women irregardless of the passenger class they fall in. To do this, the data was grouped by 'Sex':

```
#calculate the total number of males and females in the data set

total_number_of_males = (data_df['Sex']=='male').sum()
total_number_of_females = (data_df['Sex']=='female').sum()

#calculate the total number of males and females who survived
data_by_sex = data_df.groupby(['Sex']).sum()
total_number_of_males_survived = data_by_sex['Survived']['male']
total_number_of_females_survived = data_by_sex['Survived']['female']
```

As can be seen from Fig. 1, overall men had a much smaller survival rate than women.

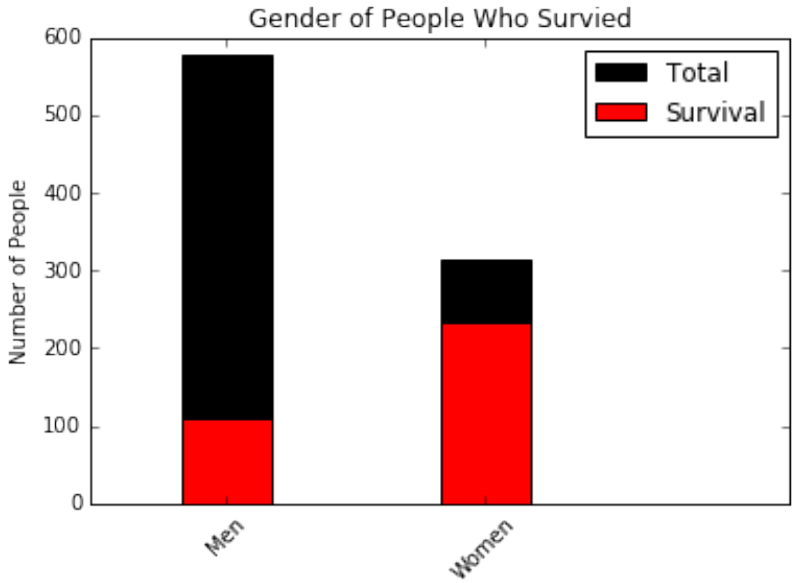

Figure 1: Total number of passengers and passengers who survived, break down by gender.

Now let us take the passenger class into account. As noted in variable description[1], passenger class or Pclass is a proxy for socio-economic status. We can verify this by looking at the average passenger fare for passengers in different classes. To do this the following code snippet was used:

```
data_df.groupby(['Pclass']).mean()
```

The results are presented in Table 1, the average price of a ticket in Pclass 1 was considerably greater than Pclass 2 and 3.

Table 1: A comparison of the average ticket prices for the three 3 passenger classes in the data set.

Pclass	Average price of ticket
1	84.15
2	20.66
3	13.68

To calculate the number of men and women by Pclass they were in, the following code snippet was used:

```
Men_in_first_class = (first_class_passengers['Sex']=='male').sum()
Women_in_first_class = (first_class_passengers['Sex']=='female').sum()

Men_in_second_class = (second_class_passengers['Sex']=='male').sum()
Women_in_second_class = (second_class_passengers['Sex']=='female').sum()

Men_in_third_class = (third_class_passengers['Sex']=='male').sum()
Women_in_third_class = (third_class_passengers['Sex']=='female').sum()

#class data by sex and passenger class

first_class_data_by_sex = first_class_passengers.groupby(['Sex']).sum()
second_class_data_by_sex = second_class_passengers.groupby(['Sex']).sum()
third_class_data_by_sex = third_class_passengers.groupby(['Sex']).sum()
```

If we look at Pclass breakdown we can clearly see that men who fall in the second and third passenger class had the lowest survival rate, 15.7% and 13.5% respectively, while women in the first class had the highest survival rate, 96.8%. See Fig. 2 for a detailed figure.

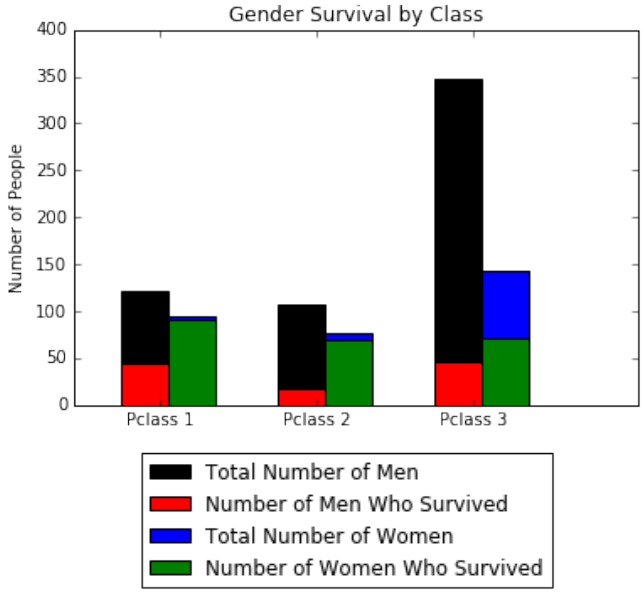

Figure 2: Total number of passengers and passengers who survived in each passenger class, break down by gender.

[1]https://www.kaggle.com/c/titanic/data

What affect does age have on survival? To answer this question we first need to drop people from the dataset whose age is not known.

```
#drop the data where the age of the passenger is not known
data_df_mod = data_df.dropna(subset = ['Age'])
```

A histogram of ages and the number of people in the data set and the number of who survived is shown in Fig. 3. It can clearly be seen that if you were below 15 years of age than you had a much greater chance of survivng than if you were older. Most people above the age of 60 died.

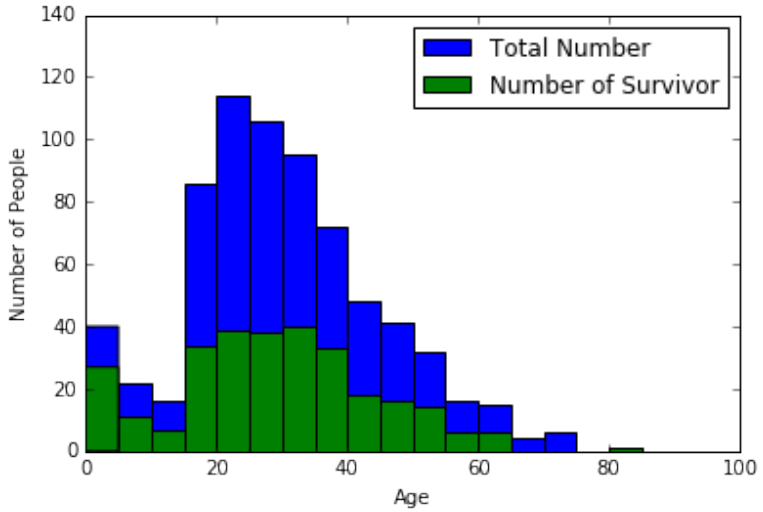

Figure 3: Total number of passengers and passengers who survived, break down by age. Bin size = 5 years.

Another way of looking at the relationship between age and survival rate is to plot the survival rate as a function of age. Since some of the people ages are given as fractions (for example 14.5), python's built-in round function is used to round all the ages to integers. For the purpose of this analysis a person aged 14.5 is assumed to be aged just 14. Survival rate is defined as:

$$\text{Survival rate(age=}n\text{)} = \frac{\text{People aged } n \text{ who survived}}{\text{Total number of people aged } n \text{ in the data set}} \tag{1}$$

It can be seen from Fig. 4 that most older passengers did not survive while the most of the younger children survived. For people who fell in the 15-60 age range, survival rate varied considerably.

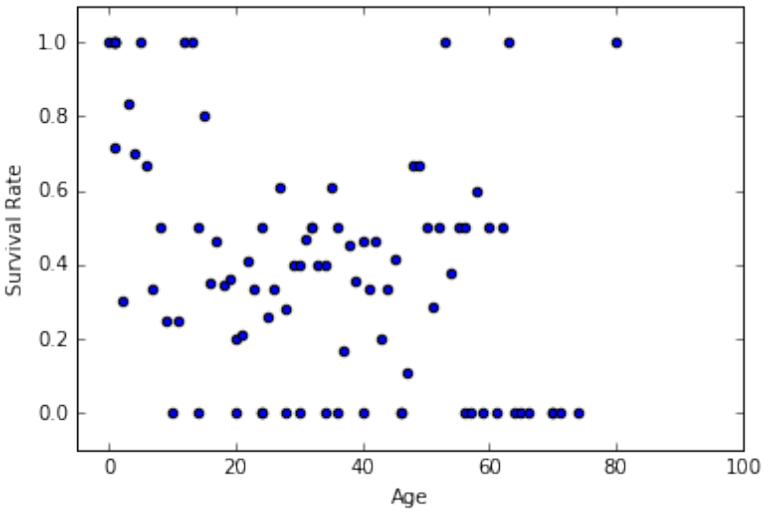

Figure 4: Survival rate as a function of age.

3 Conclusion

The data analysis presented in this document suggests that there was "sex discrimination" in the rescue effort, this resulted in women having a considerably higher overall survival rate then men. The data also suggests that there was class based discrimination, people who fell in the second and third passenger class had a lower rate of survival than people in the first passenger class. Tha data analysis also suggests that there was "age discrimination", this ensured that youger people had a higher survival rate than older people.

It is important to note some of the shortcomings present in the data set and the analysis. The data set used for this analysis is incomplete. The data contains only contains information on 891 passengers of Titanic. According to http://www.encyclopedia-titanica.org/ there were 2,208 paasengers and crew present on Titanic when it set sail. This limitation in the data set limits the application of the analyses performed. It is unclear if the data is biased in any way, this also means that it is unclear if the results presented here can be generalized to the whole Titanic population. In addition the data set given is also missing the ages of some of the passengers. The passengers whose age is given is only 714.

It is important to stress that the analysis is meant to provide some general features and trends present in the data set. No statistical tests have been performed to check the significance of the results.

4 References

- I used a variety of pages on http://stackoverflow.com/ for help.