Analysis of Titanic DataSet

By Jombee Tupas

DATASET VARIABLES

Variables in this data set include age, ticketed class, gender, name of the passenger, the cabin, whether they survived, fare, etc. It has 12 columns (variables) and 891 rows. The focus of this analysis is to examine the relationship of each variable to whether the passenger survived or not, thus answering the following hypothesis.

FINDINGS

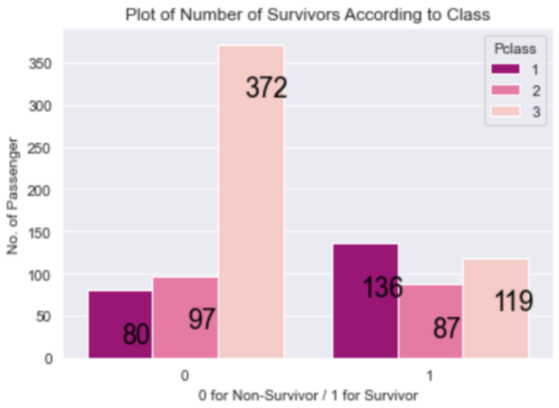
Hypothesis:

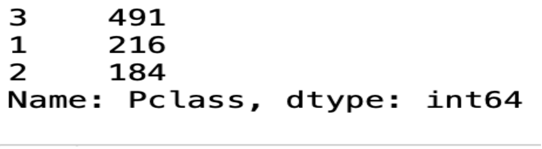
The survival rate is associated with the class of the passenger

Results:

The total number of passengers in this data set is 891 of which 216 were from Class 1, 184 from Class 2 and 491 from Class 3. Out of the 342 survivors, 136 passengers were from Class 1, 119 were from Class 2 and 87 were from Class 3. On the other hand, most of the passengers who died was coming from Class 3 in the count of 372 persons, followed by Class 2 (97 persons) and then by 80 passengers from Class 1.

This means that the survival rate is associated with Class since the number of deaths in Class C is high and most of the survivors were from first class. This is true because Class 1 passengers in Titanic had extensive facilities.





```
3     491
1     216
2     184
Name: Pclass, dtype: int64
```

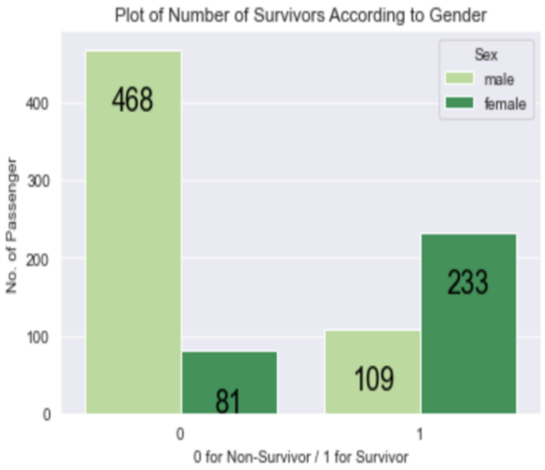
Hypothesis:

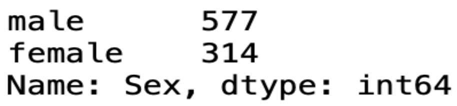
The survival rate is associated with the gender

Results:

There are 577 males and 314 females in this data set. According to the count plot, out of 342 survivors, 233 passengers were female and only 109 were male. Comparing it to the non-survivors count, 468 passengers were male and 81 were female.

The statistics show that the survival rate is associated with gender because it seems that the 'women first' attitude was practiced in the situation. I believe that most of the husbands had saved their family or wife first before themselves.





```
male      577
female    314
Name: Sex, dtype: int64
```

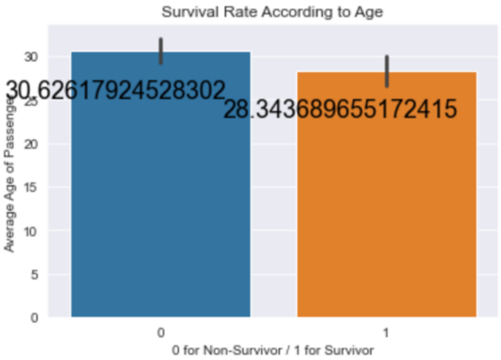
Hypothesis:

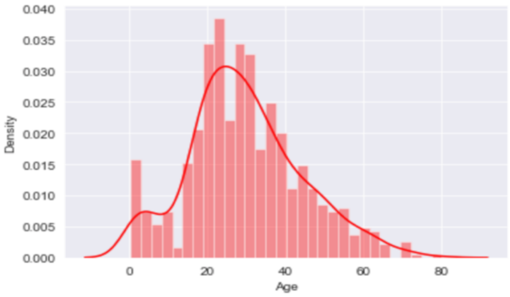
The survival rate is associated with the age

Results:

The histogram shows the distribution of data when it comes to the age of passengers. It can be noted that the passenger's age in this dataset ranges from 1 year to around 80 years old with a mean of 29.70 and which mostly were from ages 20-40 years old. Using the barplot to analyze the association of survival rate to age, it shows that the average age of survivors was 28 years old but since most of the passengers were adults, the average age of non-survivors was 31 years old.

This means that the survival rate is not associated with age because of the distribution of age in this data set although young adults have a higher chance of survival compared to older people and children.



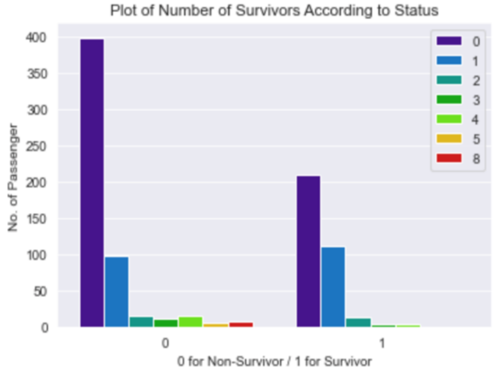


Hypothesis:

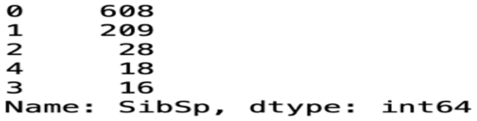
The survival rate is associated with the status (with siblings or spouse)

Results:

The plot shows that 210 passengers who survived were single or without a companion followed by 112 passengers with 1 companion. On the other hand, it can also be noted that 398 passengers who have no companion were dead and followed by passengers with 1 companion. However, it is interesting to know that most of the passengers with a companion of more than 1 didn't survive.



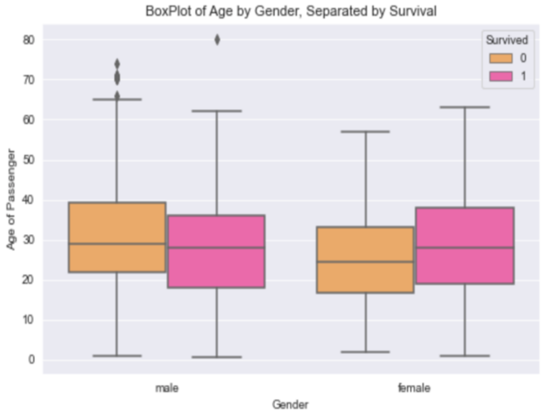
This means that the survival rate is associated with the companion status because the lesser the companion a passenger had, the higher the chance of surviving. Although the highest number of survivors were no companions, there's not enough information available to say that it is associated with survival rate since 69% of the passengers were no companions but only 4% had survived.



```
0      608
1      209
2       28
4       18
3       16
Name: SibSp, dtype: int64
```

Hypothesis:

The survival rate is associated with age and gender



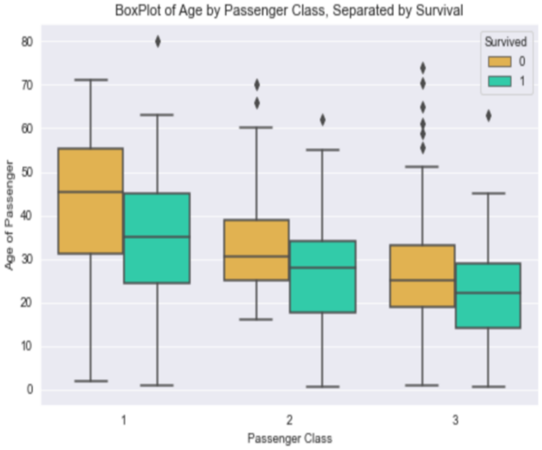
Results:

Although there were some outliers in this box plot due to null values in the age variable, it can be noted that female survivors were older than male survivors. On other hand, male non-survivors were older than the female.

It seems that the age and gender of passengers were not correlated with each other and the survival rate was not affected by considering the two factors.

Hypothesis:

The survival rate is associated with the age and Passenger Class



Results:

It can be noted in this boxplot that some outliers have existed because of null values in age. Though it can be cleaned, I opted not to since it didn't bear much in proving the hypothesis. The box plot shows that data were not normally distributed according to P-class. It also shows that there is a relationship between P-class and age because the passenger class level goes down, age goes as well. It can be true because young adults may not have the capacity to pay for Class 1 tickets due to financial stability as compared to older adults. When it comes to survival rate, most of the survivors were from Class 1 regardless of age (with 80-year-old survivors).

This means that P-class and Age have a direct relationship with each other but only Pclass is associated with survival rate.

CONCLUSIONS

I, therefore, conclude the following;

1. Passenger Class is associated with survival rate and also has a relationship with the age of passengers.
2. Age is not associated with survival rate although might be a factor in survival.
3. Gender is associated with survival rate because of the male “emotional” factor.
4. Age and gender are not related to each other and being together didn't affect the survival rate.
5. Age and Passenger Class although related, only Passenger Class is associated with survival rate.