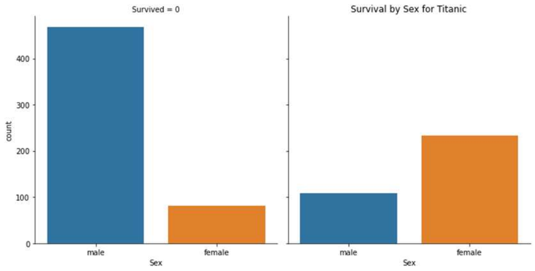
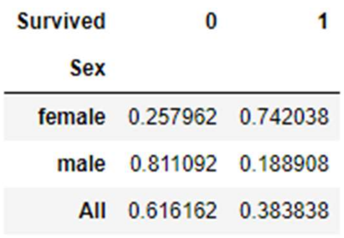
Assignment 3 Part 2 CPSC 4800 Gurbir Bhangu

Using the Titanic dataset from Kaggle, I conducted some exploratory data analysis. From this analysis I created 3 hypotheses that I tested using a variety of statistical graphs and some statistical calculations. For the data, survival was indicated by a 1 and did not survive was indicated by 0.

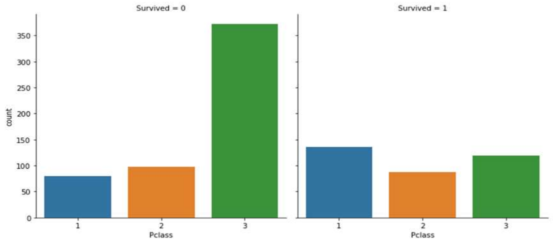
My first hypothesis was that there would be an association between survival rate and the gender of the passenger. I created a side-by-side bar plot, as well as a cross-tabulation table. As evidenced by both, we can see clearly, not only did more females total survive the titanic, but a much larger percentage of females survived the titanic. Therefore, I believe there is significant evidence to believe that survival and gender have some statistical association.

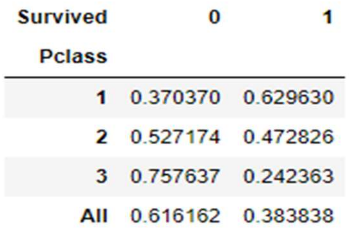




| Survived | 0 | 1 |
|---|---|---|
| Sex | | |
| female | 0.257962 | 0.742038 |
| male | 0.811092 | 0.188908 |
| All | 0.616162 | 0.383838 |

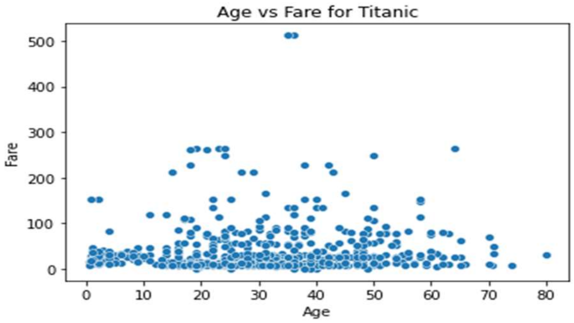
Next, I looked at if there was an association between survival rate the socioeconomic class (called Pclass by the data set). My hypothesis was that the higher the socioeconomic class of a passenger, the more likely they were to survive the titanic. Similar to above, I did a 2 categorical variable analysis with side-by-side bar plot, and a cross tabulation table. There seems to be sufficient evidence of a significant statistical association between survival and the socioeconomic class of the passenger. (Pclass 1 = upper class, Pclass 2 = middle class, Pclass 3 = lower class).

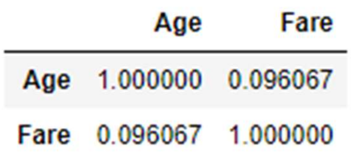




| Survived | 0 | 1 |
|---|---|---|
| Pclass | | |
| 1 | 0.370370 | 0.629630 |
| 2 | 0.527174 | 0.472826 |
| 3 | 0.757637 | 0.242363 |
| All | 0.616162 | 0.383838 |

For my final analysis, I wanted to see if there was any significant correlation between the age of a passenger and the fare that the passenger had paid. I produced a scatter plot of age vs fare and also did a correlation calculation. From below we can see there is no significant trend that can be discerned from the scatter plot, nor is the correlation coefficient high enough to say there is any linear trend in the data.





| | Age | Fare |
|---|---|---|
| Age | 1.000000 | 0.096067 |
| Fare | 0.096067 | 1.000000 |