Razel Sadaya
100381744
CPSC 4800-W01
March 11, 2022

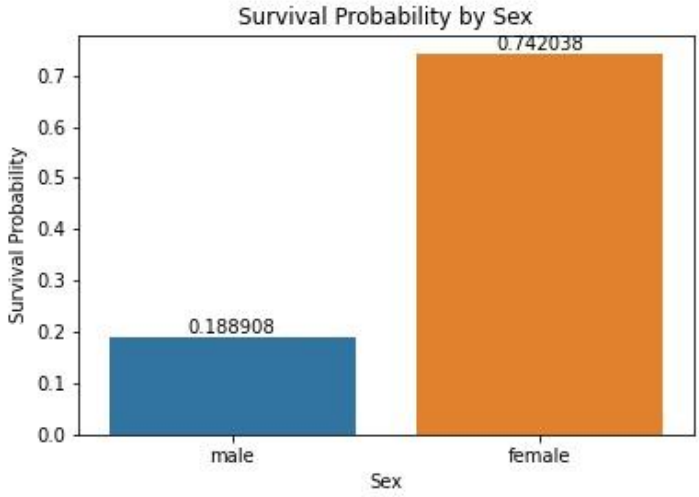
Analysis of the Titanic Dataset

1. Data

The dataset has a total of 891 entries. There are 12 variables – 5 variables are numeric, and 7 variables are categorical. There appear to be missing values in the dataset for the variables Age, Cabin, and Embarked. Since we want to use Age for this analysis, the missing values for Age are imputed using the *fillna()* function with the mean value of age.

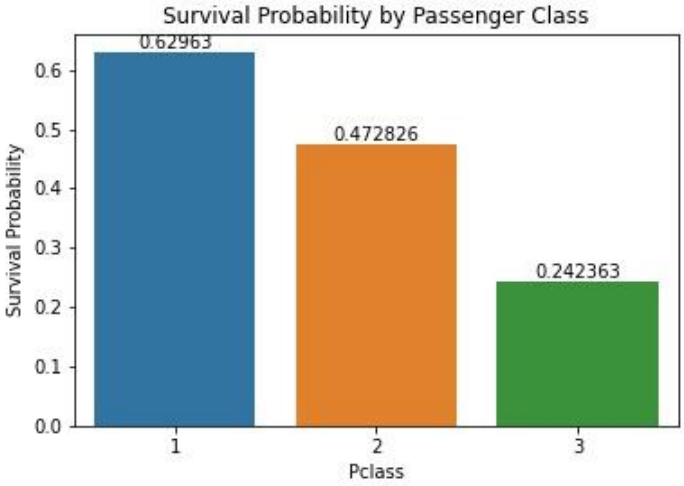
2. Analysis

2.1 Is survival probability associated with gender?



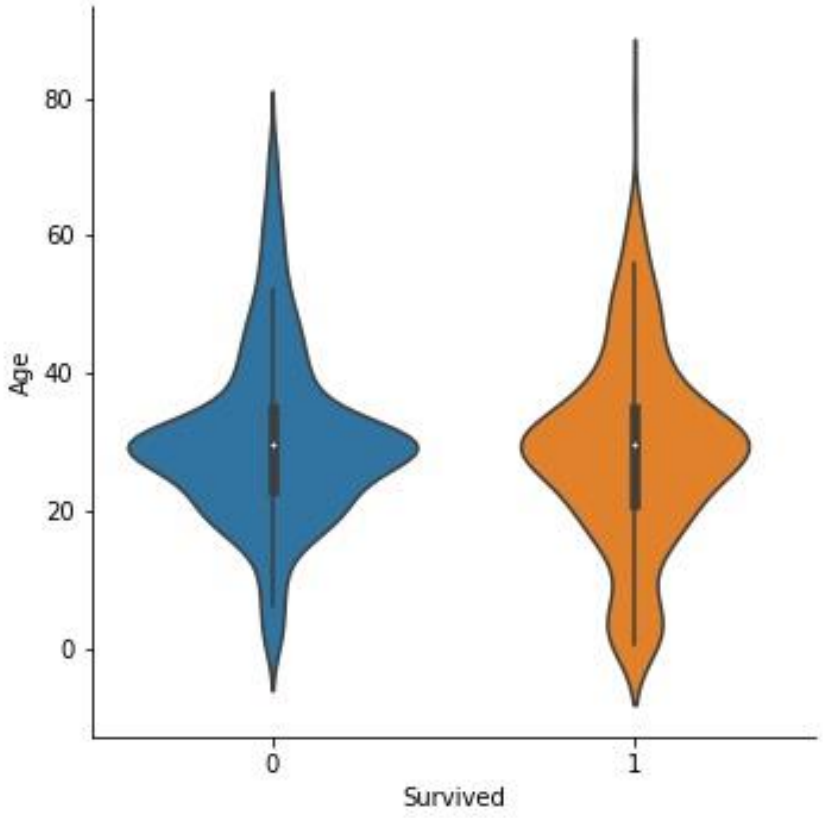
The bar graph above shows that women have a survival rate of about 74%, which is higher than men's survival rate at about 18%. However, the result of the chi-square test, where c = 260.72 and p = 1.2, suggests that there is no statistical evidence that survival probability is associated with sex.

2.2 Is survival probability associated with passenger class?



The bar graph above shows that the passengers in the higher class have a higher survival probability. The upper class (plcass = 1) has a survival rate of about 63% percent, the middle class (pclass=2) has a survival rate of about 47% and the lower class (pclass=3) has a survival rate of about 24%. However, the result of the chi-square test, where c = 102.89 and p = 4.55, suggests that there is no statistical evidence that survival probability is associated with passenger class.

2.3 Is survival probability associated with age?



Looking at the violin plot above, it seems that age is not associated with survival probability. Passengers who did not survive, the one represented by the blue 'violin', seem to have the same distribution as the passengers who survived as seen in the orange 'violin'. Both survivors and non-survivors have a median of around age 30. The point-biserial correlation also indicates there is no correlation between the survival probability and age since the correlation coefficient is -0.070.