CPSC Assignment 3 Part 2 Summary

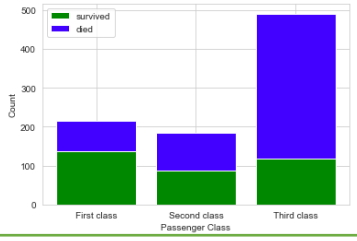
The purpose of this project is to analyse the Titanic data set and provide statistical interpretations. The Titanic event is considered as the biggest cruise ship disaster in history and the data for this has been obtained from the official Kaggle website. The variable of interest from the analysis is the survival rate. The survival rate has been analysed against the class of passengers, gender and age. Upon analysing survival rate against class of the passenger using the cross tabulation and bar graph, the first impression is that the deaths have gradually increased as the class of the passenger fell. On conducting the chi square results it is noticed that the frequencies are greater than 5, hence the test can be trusted. The statistic value has been recorded at 102.888, the p-value at 4.549251711298793e-23 and degree of freedom 2. From these results we can infer that the statistic value is significantly greater than the DP value of 5.99(DF =2) and the p-value is less than the alpha value of 0.05. Hence, we can reject null hypothesis and assert that there is a strong association between the survival rate and the class of the passenger, that is higher the class (1^{st} class) more the chances of surviving. Further the survival rate has also been analysed against the gender and the first impression from the cross tabulation and bar graph tells us that the the number of deaths is very high among the males compared to the females. The chi square test obtained a statistic value of 260.717, p-value of 1.1973570627755645e-58 and degree of freedom 1. The expected frequencies are all greater than 5, hence the test can be trusted. We can understand from the above results that the statistic value is significantly greater than the DP value of 3.84 (DF =1) and the p-value is less than the alpha value of 0.05, hence null hypothesis can be rejected and assert that there is a strong association between the survival rate and the gender, that is females have very high chances of surviving the catastrophe in comparison to the males. The survival rate has also been checked against the age. From the box plot plotted it is clear that there is no much significant relation between the two variables as the means of both the groups almost stand at the same level and the spread of the graph is identical too. Although, there is one particular age group which is of interest and that is 60-80. Passengers belonging to this age group have a very high risk of death from the event as it can be observed from the box plot comparison.

CPSC Assignment 3 Part 2 Summary

Link to Github repo:

https://github.com/seshkv/titanic.git

bar graph and cross tabulation for survival vs Pclass:

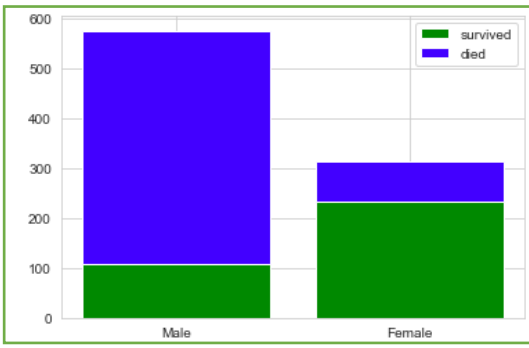


Pclass	1	2	3
Survived			
0	80	97	372
1	136	87	119

chi square test for survival and Pclass:

```
(102.88898875696056,
 4.549251711298793e-23,
 2,
 array([[133.09090909, 113.37373737, 302.53535354],
        [ 82.90909091,  70.62626263, 188.46464646]]))
```

bar graph and cross tabulation for survival vs gender:



Sex	female	male
Survived		
0	81	468
1	233	109

chi square test for survival and gender:

```
(260.71702016732104,
 1.1973570627755645e-58,
 1,
 array([[193.47474747, 355.52525253],
        [120.52525253, 221.47474747]]))
```

box plot for survival and Age: