Summary for Titanic Dataset –

The titanic dataset consists of 7 columns. The sample size for dataset is 891. Following the variables used in dataset –

1. Passenger
2. Survived
3. PClass
4. Name
5. Sex
6. Age
7. SibSp
8. Parch
9. Ticket
10. Fare
11. Cabin
12. Embarked

Among then, the below mentioned are categorical variables –

- Nominal Categorical: Survived, Sex, and Embarked
- Ordinal: PClass

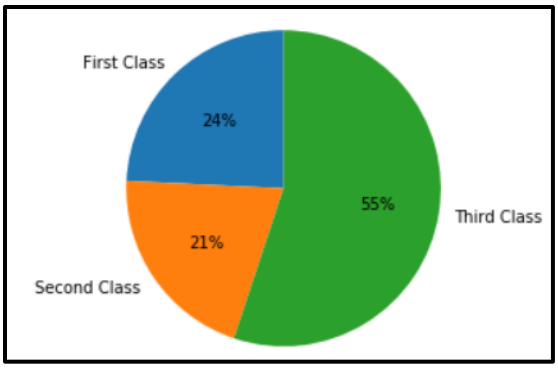
And following are the numerical variables –

- Continuous: Age, Fare.
- Discrete: SibSp, Parch.

The aim is to find out the survival rate of passengers based on different variables. So, by performing exploratory data analysis I have answered following three hypotheses based on the dataset.
- Whether the survival rate is associated to the class of passenger
- Whether the survival rate is associated to the gender
- Whether the survival rate is associated to the age

Hypothesis 1 –

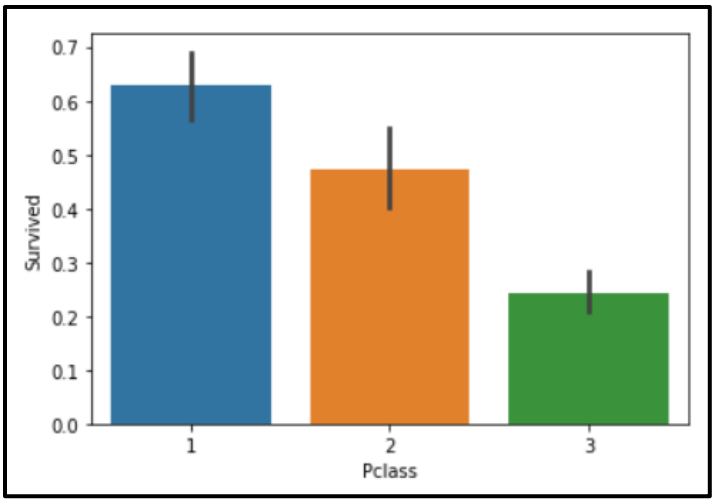
Firstly, let's check whether the survival rate is associated to the class of passenger or not. The following pie chart shows the class wise proportion of passengers travelling in the titanic.



Therefore, among all the passengers, 24% were travelling in first class, 21% were travelling in second class and 55% were travelling in third class.

Let's consider the bar graph below –

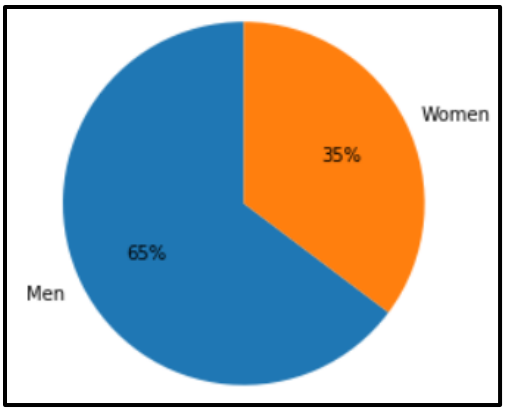
For the following bar graph, X-axis has 'PClass' variable and Y-axis has 'Survived' variable.



Hypothesis Result – The bar graph clearly indicates that the passengers in class one has less more chances of survival as compared to other classes. The passengers in class three has less chances of survival and the Passengers with class two tickets had equal chances of survival.

Hypothesis 2 –

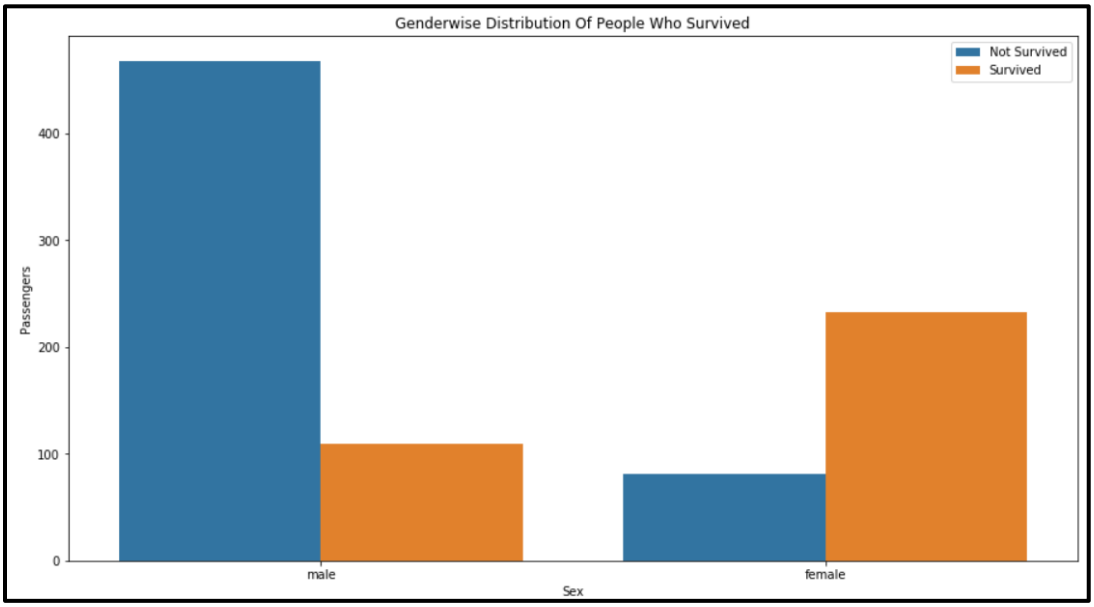
Now, let's check whether the survival rate is associated to the gender. The following pie chart shows the gender wise proportion of passengers travelling in the titanic.



Therefore, among all the passengers travelling, 35% were Women and 65% were Men.

Let's consider the side by side bar graph below –

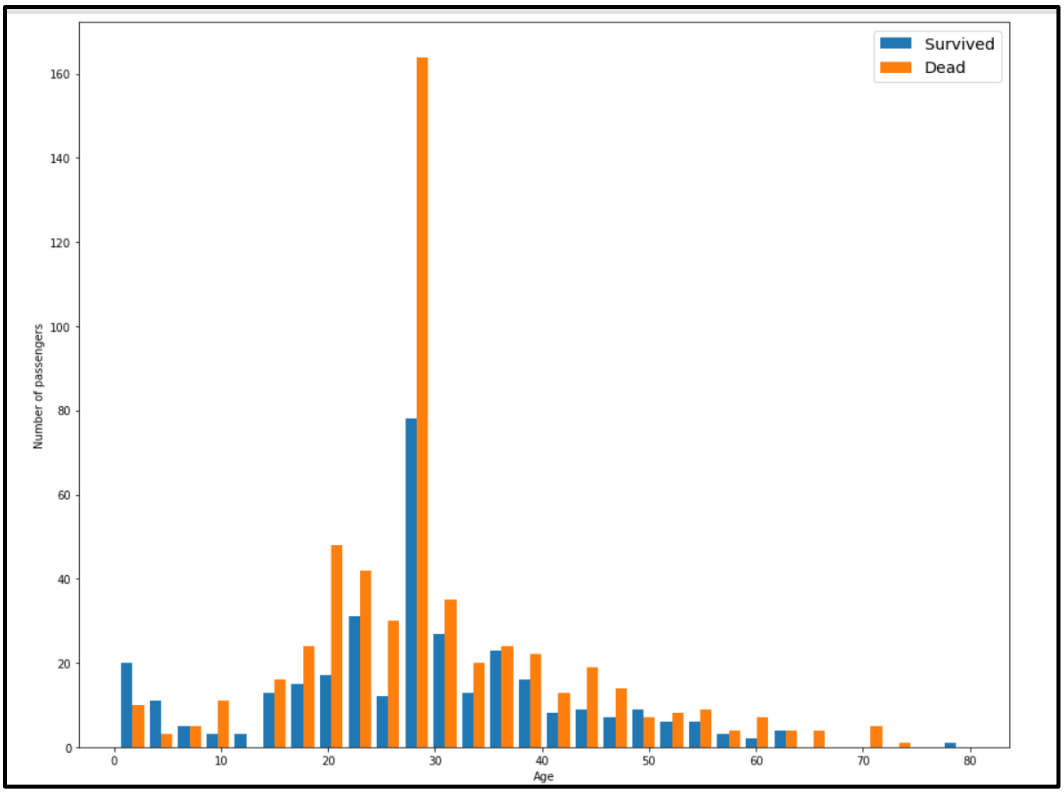
For the following bar graph, X-axis has 'Sex' variable and Y-axis has 'Passengers' variable.



Hypothesis Result – From the above bar graph, we can conclude that more male were likely to be dead than women

Hypothesis 3 –

Now, let's check whether the survival rate is associated to the age. The following side by side bar graph shows the age wise survival of passengers travelling in the titanic.



For the following bar graph, X-axis has ‘Age’ variable and Y-axis has ‘Passengers’ variable.

Now, consider the following code -

In [63]:
```
ss = pd.crosstab(train_df.Survived,train_df.Age)
from scipy import stats
stats.chi2_contingency(ss)
```

Out[63]:
```
(113.77670967050014,
 0.02858849255229426,
 87,
 array([[  0.61616162,   0.61616162,   1.23232323,   1.23232323,
           0.61616162,   4.31313131,   6.16161616,   3.6969697 ,
           6.16161616,   2.46464646,   1.84848485,   1.84848485,
           2.46464646,   4.92929293,   1.23232323,   2.46464646,
           0.61616162,   1.23232323,   3.6969697 ,   0.61616162,
           3.08080808,  10.47474747,   8.01010101,  16.02020202,
          15.4040404 ,   9.24242424,   0.61616162,  14.78787879,
          16.63636364,   9.24242424,   0.61616162,  18.48484848,
           0.61616162,  14.17171717,  11.09090909,  11.09090909,
         124.46464646,   1.23232323,  12.32323232,  15.4040404 ,
           1.23232323,  10.47474747,  11.09090909,   1.23232323,
           9.24242424,   9.24242424,   0.61616162,  11.09090909,
          13.55555556,   0.61616162,   3.6969697 ,   6.77777778,
           8.62626263,   8.01010101,   1.23232323,   3.6969697 ,
           8.01010101,   3.08080808,   5.54545455,   7.39393939,
           1.23232323,   1.84848485,   5.54545455,   5.54545455,
           3.6969697 ,   6.16161616,   4.31313131,   3.6969697 ,
           0.61616162,   4.92929293,   1.23232323,   0.61616162,
           2.46464646,   1.23232323,   3.08080808,   1.23232323,
           2.46464646,   1.84848485,   2.46464646,   1.23232323,
           1.23232323,   1.84848485,   0.61616162,   1.23232323,
           0.61616162,   1.23232323,   0.61616162,   0.61616162],
         [  0.38383838,   0.38383838,   0.76767677,   0.76767677,
           0.38383838,   2.68686869,   3.83838384,   2.3030303 ,
           3.83838384,   1.53535354,   1.15151515,   1.15151515,
           1.53535354,   3.07070707,   0.76767677,   1.53535354,
           0.38383838,   0.76767677,   2.3030303 ,   0.38383838,
           1.91919192,   6.52525253,   4.98989899,   9.97979798,
           9.5959596 ,   5.75757576,   0.38383838,   9.21212121,
          10.36363636,   5.75757576,   0.38383838,  11.51515152,
           0.38383838,   8.82828283,   6.90909091,   6.90909091,
          77.53535354,   0.76767677,   7.67676768,   9.5959596 ,
           0.76767677,   6.52525253,   6.90909091,   0.76767677,
           5.75757576,   5.75757576,   0.38383838,   6.90909091,
           8.44444444,   0.38383838,   2.3030303 ,   4.22222222,
           5.37373737,   4.98989899,   0.76767677,   2.3030303 ,
           4.98989899,   1.91919192,   3.45454545,   4.60606061,
           0.76767677,   1.15151515,   3.45454545,   3.45454545,
           2.3030303 ,   3.83838384,   2.68686869,   2.3030303 ,
           0.38383838,   3.07070707,   0.76767677,   0.38383838,
           1.53535354,   0.76767677,   1.91919192,   0.76767677,
           1.53535354,   1.15151515,   1.53535354,   0.76767677,
           0.76767677,   1.15151515,   0.38383838,   0.76767677,
           0.38383838,   0.76767677,   0.38383838,   0.38383838]]))
```

The following code displays the value for chi-square analysis.

H0: Survived people and Age are independent to each other.

Ha: Survived people and Age are not independent to each other

$\chi 2 = \sum((f - e)2 / e$

$\chi 2 = 113.7767$

DF = (r-1) (c-1)

DF = 87

And therefore DP = 113.00

Hypothesis Result – Since, $\chi 2$ is bigger than the decision point, we have enough evidence to reject the null hypothesis.

We conclude that, Survived people and Age are not significantly independent to each other.