Report

I. Missing data

```
Out[92]: PassengerId      0
         Survived         0
         Pclass           0
         Name             0
         Sex              0
         Age            177
         SibSp            0
         Parch            0
         Ticket           0
         Fare             0
         Cabin          687
         Embarked         2
         dtype: int64
```

Before doing any analysis on the data, we need to find out any missing data.Based on the table Age, Cabin and Embark have null values so we take a further look at the columns

we start with Embark which has 2 missing values. we look for the value that occurs the most often for that column and we find out that it is "S" so we replace the missing values with "S".

Then, we look at the cabin column. There are 687 missing values. The values that occur the most often in Cabin are 3 ('G6', 'C23 C25 C27', 'B96 B98') so it is better not to replace the missing Cabin values with any other value as there is no clear dominant Cabin. Also the column will not be needed for the purposes of our analysis

Finally, we look at the third and last column with missing values which is Age. there are 177 rows with missing ages. this is more than 10% of the data so deleting the rows is not a good solution. therefore, we replace the null values with the mean of the ages that have the same survival, PClass and gender values.

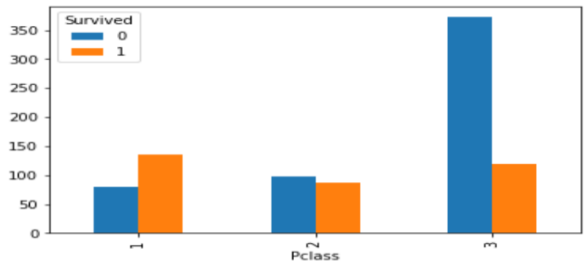
II. Survival rate's association with the class of passenger

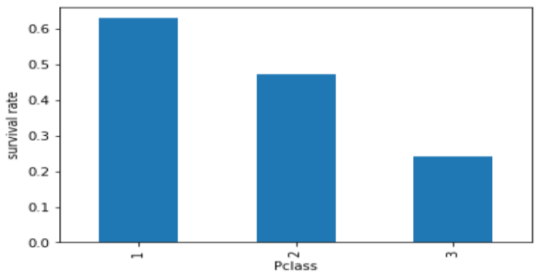
To find out the relationship between the survival rate and the class of the passengers and since the two columns are categorical, I used the cross table and the calculated survival rate by pclass then used the 2 new tables to plot the bar graphs and below are the results:

Out[25]:

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

```
Out[27]: Pclass
         1    0.629630
         2    0.472826
         3    0.242363
         Name: Survived, dtype: float64
```





based on the results above, we can see that class 1 passengers have a better rate of survival than any of the other 2 classes and class 2 passengers' survival rate is d**ou**ble the survival rates of class 3 passengers. both bar graphs also confirm the same finding: there were more survivals among class 1 passengers than deaths and the number started decreased in classes 2 and 3 where the number of deaths is 3 times more than the number survivors.

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

The chi-square value is another way to prove the association between the 2 categorical variables and the above method shows that degree of freedom is 2 so the decision point is 5.99. then, we compare the chi-square value (102.88) with the DP (5.99) and find out that it is a lot greater than the decision point, we can conclude that there is a strong relationship between the pclass and survival rate which confirms my findings.

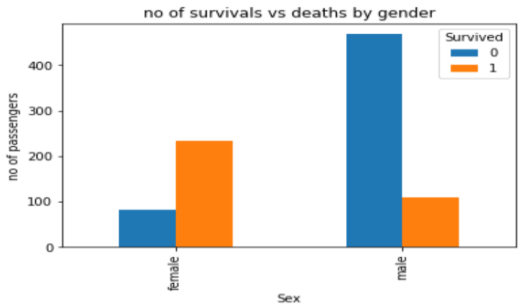
III. Survival rate's association with the gender of passenger:

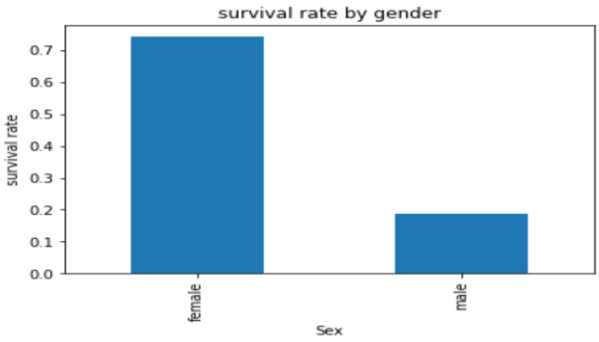
The above procedure was done also to find out the association between the gender and the survival rate.

Out[34]:

Survived	0	1
Sex		
female	81	233
male	468	109

```
Out[58]: Sex
         female    0.742038
         male      0.188908
         Name: Survived, dtype: float64
```





The survival rate among females (74%) is way higher than the survival rate (18%) among males and the bar graphs show that. In fact, the number of females who survived are 3 times the number of females that died. However, the number of males who died was 4 times the number of males who survived.

```
Out[50]: (260.71702016732104,
          1.1973570627755645e-58,
          1,
          array([[193.47474747, 120.52525253],
                 [355.52525253, 221.47474747]]))
```

the above method shows that degree of freedom is 1 so the decision point is 3.84. then, we compare the chi-square value (260.71) with the DP (3.84) and find out that it is a lot greater than the decision point, we can conclude that there is a very strong relationship between the gender and survival rate which confirms my findings.

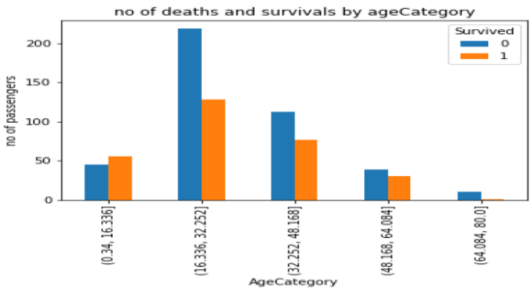
IV. Survival rate's association with the age of passenger

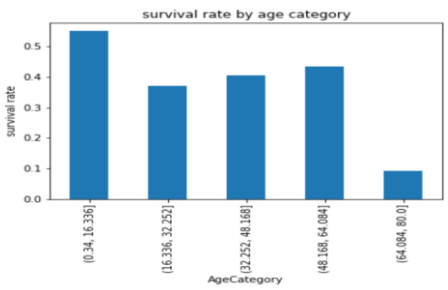
Since there are 88 unique values in the age column we cannot and both columns are considered categorical, we divide the ages into 5 equal groups, that way the results can be better interpreted. After binning, we analyze the association between the new agecategory column and the survival rate.

Out[43]:

Survived	0	1
AgeCategory		
(0.34, 16.336]	45	55
(16.336, 32.252]	218	128
(32.252, 48.168]	112	76
(48.168, 64.084]	39	30
(64.084, 80.0]	10	1

```
Out[41]: AgeCategory
         (0.34, 16.336]      0.550000
         (16.336, 32.252]    0.369942
         (32.252, 48.168]    0.404255
         (48.168, 64.084]    0.434783
         (64.084, 80.0]      0.090909
         Name: Survived, dtype: float64
```





The survival rate is highest among the youngest group (between the age of 3 months and 16 years and it is the worst in the oldest group (between the age of 60 and 80), there does not seem to be a pattern in the other 3 groups. For example, the second youngest group has a lower survival rate than the third and fourth youngest group.

```
(15.229524960081086,
 0.0042480926672647905,
 4,
 array([[ 59.3837535 ,  40.6162465 ],
        [205.46778711, 140.53221289],
        [111.64145658,  76.35854342],
        [ 40.97478992,  28.02521008],
        [  6.53221289,   4.46778711]]))
```

The chi-square method shows that degree of freedom is 4 so the decision point (9.49) and when we find out that the chi-square value (15.22) is a bit greater than the decision point, we can conclude that there could be a relationship between the age and survival rate but it is not a strong one which confirms my findings.

V. Conclusion

Based on the above analysis, we conclude that the survival is mostly affected by the gender of the passengers. Females had the largest rate of survival (74%) compared to a survival rate of only 18% for males. The second most important factor in the survival rate out of the 3 that we explored is the class of the passengers. In fact, survivals were the highest among 1st class passengers followed by 2nd and 3rd class passengers respectively. Lastly, age was the factor with the least influence compared to the other 2 discussed. Because while the youngest age category had the highest survival rates and the eldest people had the lowest rates, the trend was not followed in the 3 middle age categories where the survival rate remained constant. However, this does not dismiss the idea that if the passenger is older than 60, his/her chances of survival were very slim. While all people with ages between 16 and 59 had almost the same chance of survival and people younger 16 and younger had the highest chance of survival.