# **REPORT ON THE TITANIC DATA SET**

The Titanic ship that sank in April 15, 1912 Consist of 1309 rows and 13 columns containing the names, Age, Passengers Id, Sibsp, Parch, Survived, Embarked, Cabin, Name, Fare, Pclass and Sex of Passengers.

In the dataset, we have three Passengers’ Class, ‘Pclass’ namely: 1,2 and 3.

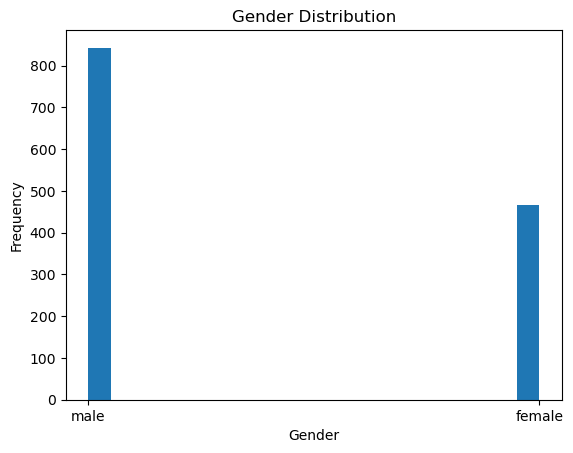
The 1309 passengers were made up of 466 females and 842 Males who boarded using 3 ports (Embarked) namely: Southampton, Cherbourg and Queenstown (S, C, Q) Represented by 914, 270 and 46 passengers respectively. Two of the columns in this dataset has missing/null values, which is Age column having 263 missing values and Cabin column with 1014 missing values).

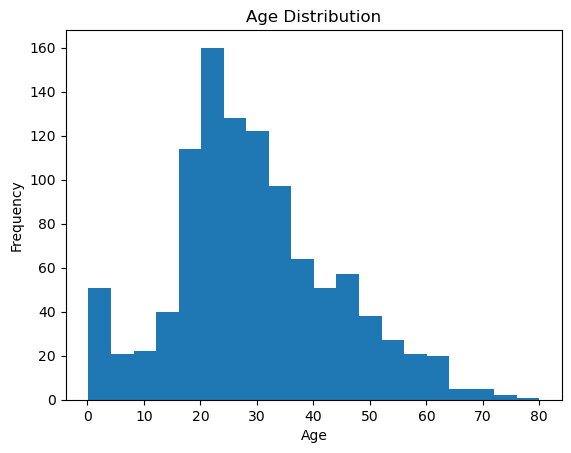
The people that constituted family members were siblings and spouses (SibSp), and parents with children (Parch) .

**EXPLORATORY DATA ANALYSIS**

**PASSENGER’S DEMOGRAPICS:**

The dataset shows that in the ship they were more Male than female. In comparison to other age groups with relatively low frequencies, passengers in the 15- 23 age bracket are numerous and they travelled more. The first bar chart below shows that they were 843 males and 466 females. This indicates that males who boarded the ship were more than females.





**For Age Distribution**

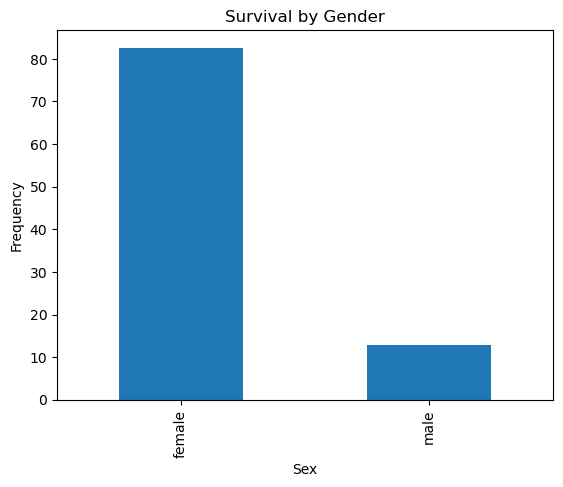
In the chart above, Age of passengers range from 1 to 80 years, with high frequency at 20 t0 30 years. This shows that most of the passengers were young aged people.

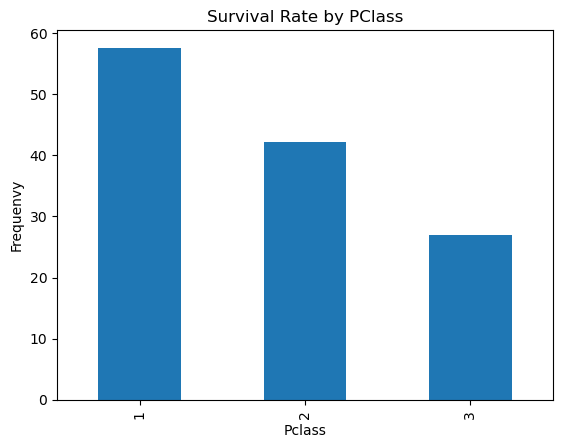
**SURVIVAL RATE**

In the dataset, analysis shows that the overall survival rate of People who survived is 37.74% while those that did not survive is 62.26%.

**Survival rate by gender**

shows that 13% male and 83% female survived. it shows that that the female survived more.



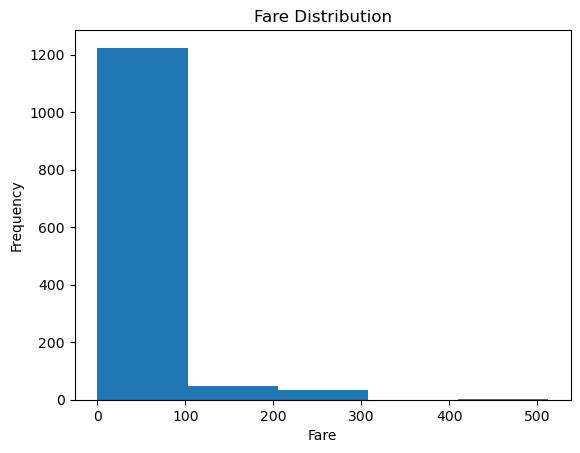


**For Survival Rate between classes: In the first** class, the Survival rate is high at 57% followed by the

People in Second class at 42% and THIRD class at 26% had the lowest survival rate.

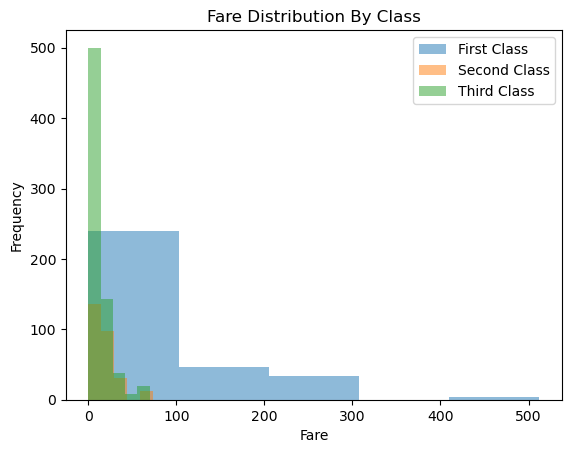
Therefore, more people Survived in first class than Other classes.

**TICKET FARE AND CLASS:**



**Fare Distribution**

The fare ranges from 0 to 515 naira. The Above chart displays a right Skewed Pattern with the highest price ranging from 0 to 100 naira. The lowest fare paid by Passengers Ranges from about 410 to 515 naira.

****

**Fare distribution by class**

In the chart above, the fare for first class is expensive followed by second class and the Third class has a cheaper fare. The fare also affected the number of passengers in each class

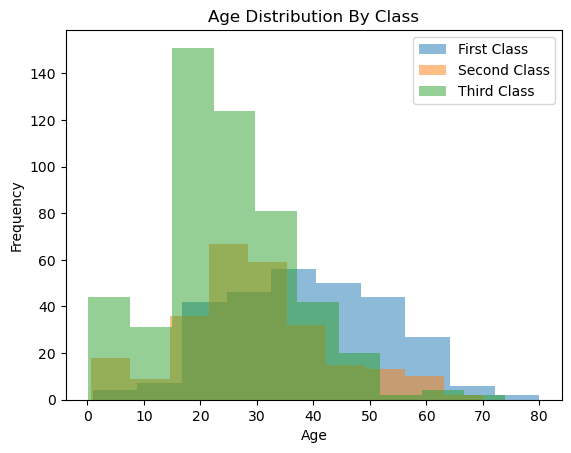
It is clear that the fare and survival has a significant positive correlation at 15. A passenger odd of surviving would increase with payments/fare.

In the chart, it shows that the first-class passengers paid 0 to 512-naira, second class paid 0 to 766 naira and third class paid 0 to 75 naira.

**AGE DISTRIBUTION**

**For Age Distribution by Class:**

The people that boarded the ship aged from zero (0) to about 80 years it has a normal distribution. The highest age frequency ranged from 15 years to about 33 years and the lowest at about 52 to 75 years’ old



Age distribution between class from the charts shows:

First class from 0.1 to 80 years old

Second class from 0.1 to 70 years old

Third class from 0 to 75 years old

The highest age frequency was being found in third class and second class at age range 15 to 33 and at 22 to 28 years.

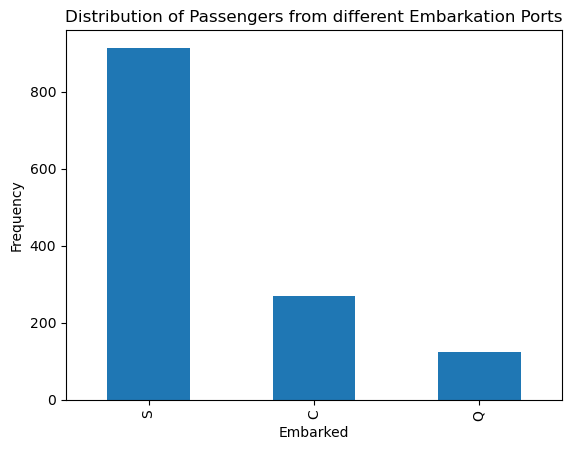
**FAMILY RELATIONSHIP**

There were relatively few passengers that travelled with their families as evidenced by the 39.64% proportion of passengers who were accompanied by siblings, Parents, children and spouses.

The presence of family members on board had an Impact on the survival rate since more passengers (50.67%) Survived with their families than without families (29.24%)

This indicates that having Family on board has a positive impact on survival Rates.

**EMBARKATION PORT**

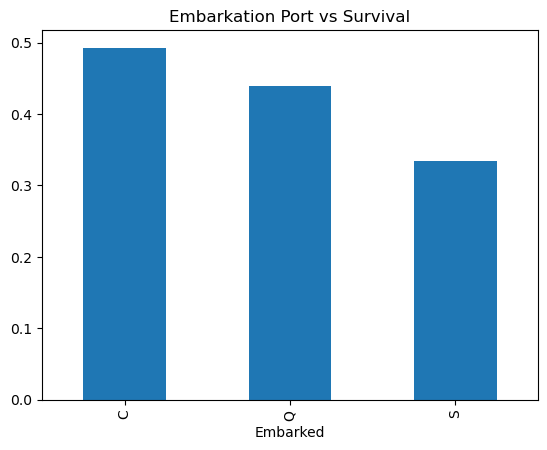


**DISTRIBUTION OF PASSENGERS FROM DIFFERENT EMBARKATION PORTS**

A minimum of 123 passengers embarked at Queenstown, maximum of

914 passengers boarded at the port in Southampton, 270 at the port in Cherbourg.

With Positive correlations at 0.492, 0.439 and 0.333 Is seen between the ports of Cherbourg, Queenstown and Southampton, and their survival rates observed respectively. This merely Indicates that there are greater Possibilities of survival for those who sailed from Cherbourg than for other ports. More passengers embarked at Southampton.



**FOR EMBARKATION PORT VS SURVIVAL**

The chart shows that there is 0.492593 correlation between those that embarked at Cherbourg and survival, there is 0.439024 correlation between those that embarked at Queenstown and survival and there is 0.333698 correlation between those that embarked at Southampton and survival. All the Embarkation ports have positive correlation with survival, however, Cherbourg has more positive correlation with survival

**EXPLANATORY DATA ANALYSIS**

**FEATURE CORRELATION:**



From the heatmap above the Highest correlation (0.86) is seen between Family and siblings with spouse (SibSp).

This indicates that there’s a greater number of passengers with Children and spouses as evidenced by the strong correlation between family and Parch and Sibsp which are 0.86 and 0.79 respectively

The association between family and fare is Positive (0.23), Indicating that the fare is determined by numbers of the people in the family

At 0.79 there is a Positive Correlation between parch and Family, 0.061 Positive Correlation between family and survived, 0.37 between SibSp and Parch, 0.22 Positive Correlation between parch and fare, 0.18 Positive Correlation between age and fare, 0.23 positive correlation between fare and survival and 0.05 between Pclass and family.

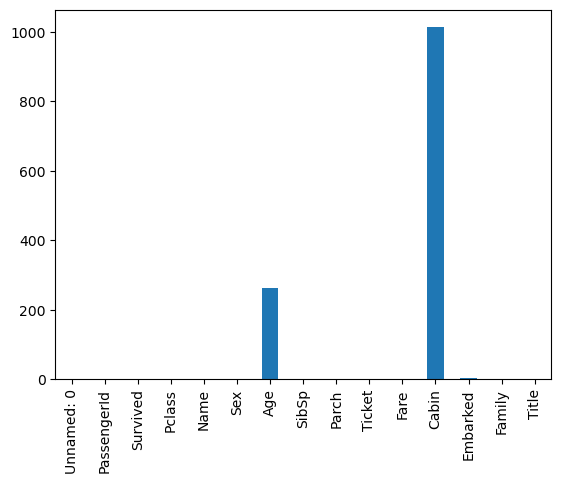
There exist negative correlations too between Survived and Pclass (-0.26), and between age and survival (-0.054), between Pclass and age (-0.41), between Pclass and Fare (-0.56), between

**NAME ANALYSIS**

Compared to other passengers, those titled Dona, Miss, Ms have higher survival rates.

Individuals bearing the titles Col, Dr, Master, Mr, and Rev. Had minimal prospects of survival.

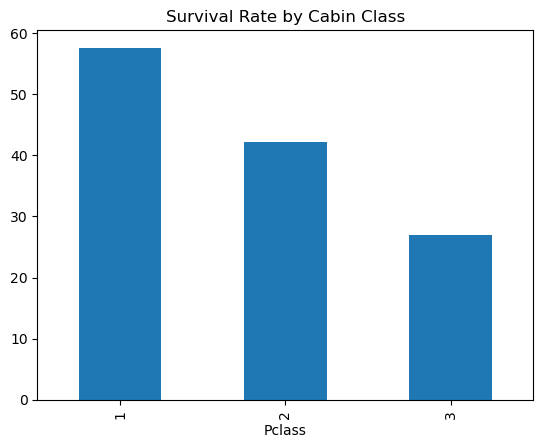
**MISSING DATA HANDLING**



From this chart, the values in the other columns are complete while Their Age and cabin have a large number of missing values which is 263 and 1014 respectively

Given large numbers of missing values for cabin, I consider either eliminating this feature or Using it as a categorical variable. Imputation using the mode values of other significant methods for Embarked. For Age in imputation in predictive models, use Mean/median Age.

**CABIN AND CLASS SURVIVAL**

****

Survived as a function of Pclass.

It is observed in the chart above, that First class have survival rate of 57.5%, second class 42.2% and third-class passengers 26.9%.

**CONCLUSION**

This is the Analysis on Titanic dataset in which I Predicted 815 died out of 1309 people. Survival rate by gender shows that 86.5 females survived and 13.5% male survived. The survival rate according to class shows that in the first, second and third class 186, 117, 191 persons survived.

This was an ugly incident but nonetheless Factors like Embarkation, Pclass affected the

Rate of Survival. People in first class had the Highest rate of survival and they paid more fare.

Factor that had the highest correlation (0.86) Was between family and SibSp followed by Family and parch (0.79)

There was no fare between 310 and 410. The fare ranged from zero to 512. The fare that was paid at the lowest frequency fell between 412 and 512. The range of fare with the highest frequency was zero to 100.

-