# **REPORT ON THE TITANIC DATA SET**

This dataset consists of 1309 rows and 13 columns containing the names, passenger Id’s, sex, age, etc of passengers on board the titanic ship that sank on April 15,1912. Below is the first five rows of the dataset.

**Table 1**

| **S/N** | **PASSENGERID** | **SURVIVED** | **PCLASS** | **NAME** |
| --- | --- | --- | --- | --- |
| 0 | 1 | 0.0 | 3 | Braund, Mr. Owen Harris |
| 1 | 2 | 1.0 | 1 | Cumings, Mrs. John Bradley(Florence Briggs) |
| 2 | 3 | 1.0 | 3 | Heikkinen, Miss Laina |
| 3 | 4 | 1.0 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) |
| 4 | 5 | 0.0 | 3 | Allen, Mr. William Henry |

**Table 2**

| **S/N** | **SEX** | **AGE** | **SIBSP** | **PARCH** | **TICKET** | **FARE** | **CABIN** | **EMBARKED** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | Female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | Female | 26.0 | 0 | 0 | STON/O2 3101282 | 7.9250 | NaN | S |
| 3 | Female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | Male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

**EXPLORATORY DATA ANALYSIS**

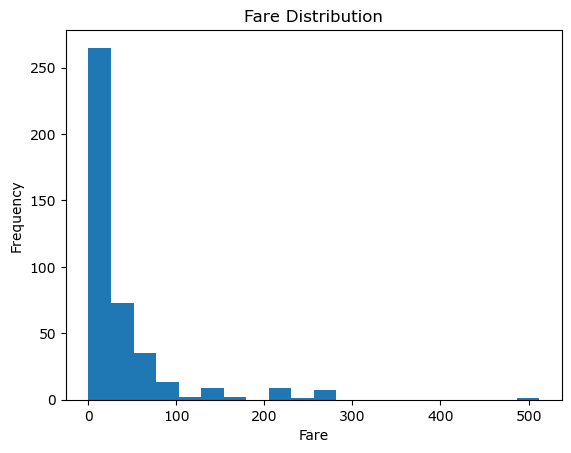
PASSENGER DEMOGRAPHICS:

In this dataset, the number of females on the ship are more than the number of males. Passengers within the age group 20 to 30 are more compared to other age groups with relatively low frequencies. The most common titles in passengers’ names are Countess, Lady, Mlle, Mme, Ms, Sir etc.

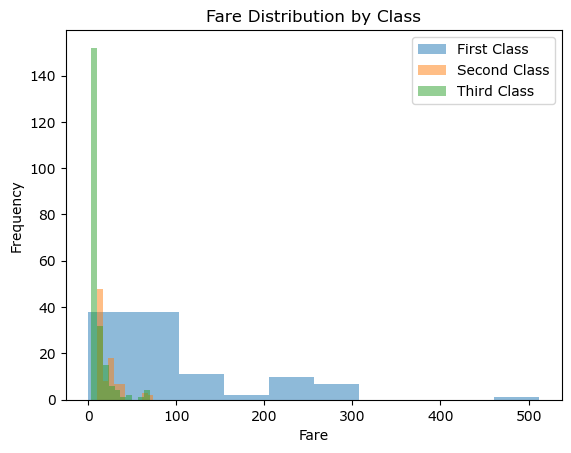
SURVIVAL ANALYSIS:

From the analysis carried out on this data set, the rate of those that survived the accident is 38%. The survival rate between genders shows all females survived and the males did not. Basically, women survived and men didn’t. For each class, we see that the survival rate of those in first class is high at 46.73 followed by those in third class at 33.03 and least in second class at 32.26. So basically, more people survived in first class than in other classes. Those that survived and age are negatively correlated with a value of -1.3e-05 which implies that the ages of passengers does not determine their survival. With family relationships, it is positively correlated (0.16) indicating that there’s a high survival chance of passengers with family.

TICKET FARE AND CLASS:



The above chart shows a right-skew pattern with the highest fare prices ranging from 0 to 40 and we see, the lowest fare prices ranging from about 490 to 501.

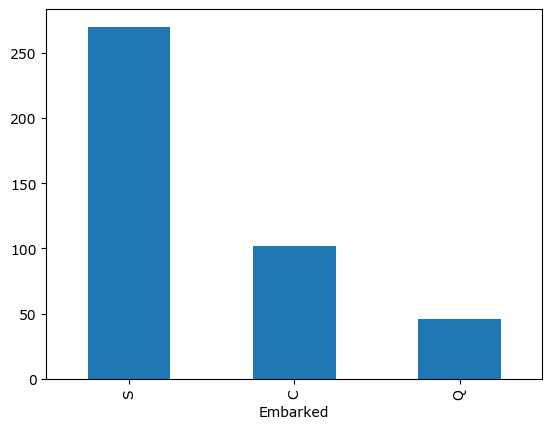


The fare prices also affect the number of passengers in each class. As usual, the price of seats in the first class are expensive, slightly expensive in the second class and cheap in the third class. So as we can see in the chart above, the price for first class is more expensive compared to the other classes. There’s a positive relationship between fares and those that survived at 0.19 indicating that the higher the fares, the higher the chance of survival.

FAMILY RELATIONSHIPS:

The rate of passengers with family members (composed of siblings, spouses, parents and children) is 39.47%; meaning that not a lot of passengers travelled with their family. Having family aboard the ship affects survival rate because the rate of those that survived with family (50.91%) is higher than those that are without family (26.88%) which means that more passengers survived with their families.

EMBARKATION PORT:



A maximum number of passengers boarded at

Southampton port (270), 102 passengers at Cherbourg port and 46 passengers at Queenstown. The correlation of Southampton, Cherbourg and Queenstown ports, and survival rates are positive with 0.33, 0.39 and 0.52 respectively. This simply means that those who embarked at Queenstown have more chances at survival than the others.

**EXPLANATORY DATA ANALYSIS**

FEATURE CORRELATION:



There are strong correlations between Family and SibSp, and Parch at 0.79 and 0.83 respectively, indicating that there are more passengers with children. Between Age and Fare, the relationship is positive (0.34) which means that the age of passengers determines the price of the tickets. Parch and SibSp are also positively correlated (0.31). At 0.25 between Family and Fare, 0.23 between Parch and Fare, 0.19 between Survived and Fare, 0.17 between SibSp and Fare, 0.16 between Survived and Parch, and Family; and 0.1 with SibSp, 0.019 between Parch and Pclass, 0.013 between Family and Pclass, and 0.0011 between SibSp and Pclass. Negatively correlated relationships are between Survived and Pclass, and Age at -0.11 and -1.3e-05 respectively, between Pclass and Fare, and Age at -0.58 and -0.49 respectively, between Age and Family, and Parch, and SibSp at -0.094, -0.061 and -0.092 respectively.

NAME ANALYSIS:

The passengers with the titles Dona, Miss, Mrs and Ms have high survival rates compared to others. Passengers with titles Col, Dr, Master, Mr and Rev all had low chances in survival.

MISSING DATA HANDLING:

There are high number of missing values in Age and Cabin columns (86 and 327 respectively). There are no missing values in the other columns. For Age, imputation using mean/median age for predictive models. For Cabin, let’s consider dropping this feature due to high number of missing values or use it as a categorical variable. For Embarked, imputation with the mode value or other meaningful approaches.

CABIN CLASS AND SURVIVAL:

There were no survivors each in 35 cabins but in 32 cabins, there were a 100 survivors in each. Likewise 50 survivors each in 8 cabins and 1 survivor each in 33 cabins. It seems that passengers in 32 specific cabins had higher chances of survival compared to other cabins.

**CONCLUSION**

Although the incident was tragic, there were some factors that affected the rate of survival of the passengers like Pclass, embarkation, etc. Even with all these, the survival rate was low at 36.36%.