**Titanic Disaster Project: Data Analysis**

**About the Dataset:** The dataset consists of 891 rows with 12 columns. The summary of the dataset is shown below:

A screenshot of a cell phone

Description automatically generated

The first column is Passenger ID which gives us the primary key of each of the member in the Titanic dataset. The next column is survived, which tells if the person survived or not. If the value is ‘0’ then the person did not survive, otherwise he survived. ‘Pclass’ column tells you the class to which a particular person belongs. It is a categorical data with 3 levels. ‘SibSp’ column tells us the count of any sibling or spouse related to that person on board. ‘Parch’ column gives us the count of parent or children related to that person. ‘Cabin’ column will tell you the details of the cabin in which that particular person was staying. There were only few cabins in the dataset that was shown by the 1st letter of the ‘Cabin’ value. All the passengers in the Titanic embarked their journey from either of the 3 cities. This data is shown in the last column named ‘Embarked’.

While analysis the dataset, it was clear that the data had 891 passengers. However, the ‘Cabin’ column had a lot of NA values, that was needed to be handled. ‘Age’ column has 714 non-null objects, which might me be important for the future analysis.

A close up of text on a black background

Description automatically generated

**Data Cleaning:** There were not much of the cleaning required in the dataset, however in this kind of dataset it was important to separate adults from children as in any disaster situation, women and children are the ones who are saved first and this will eventually be a big difference in the survival rate of men, women and children. Therefore, a new column was created called ‘Person’ to differentiate between men, women and child.

A screenshot of a cell phone

Description automatically generated

**Univariate and Bivariate Analysis:** To provide the insights on who were the passengers on the Titanic, analysis using histograms and KDE plots were done. Some of the key points that were taken from the analysis were:

1. Number of males are more than the number of females in the dataset. The ratio looks like the number of males to number of females is 2:1.

**A screenshot of a cell phone

Description automatically generated**

1. Males belonging to Class 3 was very much when compared to males belonging to other classes whereas number of females were evenly distributed within all the classes. Also, in 1st and 2nd class, the number of males and females were almost same.

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

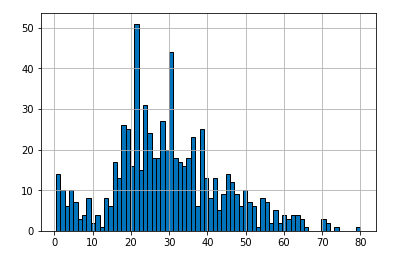
Description automatically generated**

1. Initially we created a new column to know the number of children in the dataset. The distribution of male, female and child tells us that there were a greater number of children in Class 3 than in any other class.

**A screenshot of a cell phone

Description automatically generated**

1. The age distribution tells us that the data is a bit right skewed and has a mean value of 29.69 years.

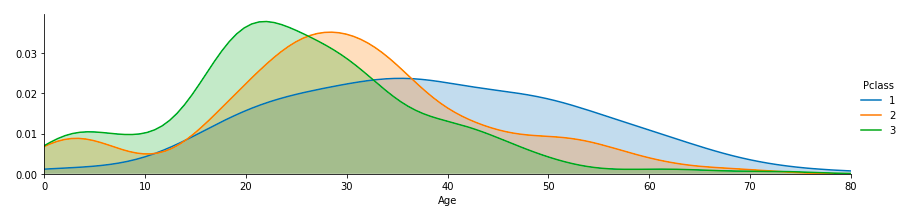


1. The age distribution with respect to male, female and children

**A picture containing furniture

Description automatically generated**

1. Age Range by class: It can be clearly seen that Class 1 has a normal distribution across the age where as Class 2 and 3 are right skewed.



**Handling null values from the Cabin column:** As stated initially, Cabin column had only 204 non-null values out of 891 rows. Therefore, in order to see the distribution of people with respect to cabins, we have to remove the null values and then count the number of people in each Cabin. Since, each cabin is designated by the 1st letter in the cabin column, we have to separate that 1st character in each of the row.

**Distribution on Cabin Column:**

**A screenshot of a cell phone

Description automatically generated**

**Distribution of Embarked Column:** Embarked column tells us the place of origin of the passenger. As can be seen from the below plot, most of the passengers came from ‘S’ town. Moreover, from the analysis, most of the passengers from ‘Q’ town are 3rd class, which is an interesting finding

A screenshot of a cell phone

Description automatically generated

**Key Analysis and conclusion:**

1. To get the insights on the number of passengers travelling alone, I created a new column. As per the analysis, the count of people who were travelling alone were more than those who were travelling with the family.

**A screenshot of a cell phone

Description automatically generated**

1. Number of males who survived were significantly less than the number of children and females who survived. This also proof that women and children were allowed to take the safety boat first and men were allowed to go at the end.

**A screenshot of a cell phone

Description automatically generated**

1. The survival rate of men was not dependent on their class.

**A close up of a map

Description automatically generated**