**Data Science Project Report**

**Survival of Passengers in Titanic Dataset**

###### **Under the guidance**

**of**

###### **Mr. Sumit Kumar Shukla**

###### 

###### **Submitted by: Sharath C Nair**

###### **Rajadhani Institute of Engineering and Technology,**

###### **Trivandrum Kerala**

###### 

###### EITHICAL EDUFABRICA (2020)

Contents

#### Introduction to Titanic dataset

1. Data Visualization
2. Analysis of Dataset
3. Building Model
4. Evaluating Model
5. Predicting with new data
6. Conclusion

Acknowledgement

I would like to express my sincere gratitude to my supervisor Mr. Sumit Kumar Shukla for providing their invaluable guidance, comments and suggestions throughout the course of the project. I would specially thank Mr. Sumit Kumar Sir for getting me the iris dataset for my exploration and successful completion of my project.

# About Titanic Dataset

It is one of the most popular datasets used for understanding machine learning basics.

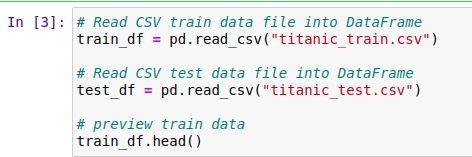
It contains information of all the passengers aboard the RMS Titanic, which

unfortunately, was shipwrecked. This dataset can be used to predict whether a given passenger survived or not.

**Importing libraries and loading the file.**



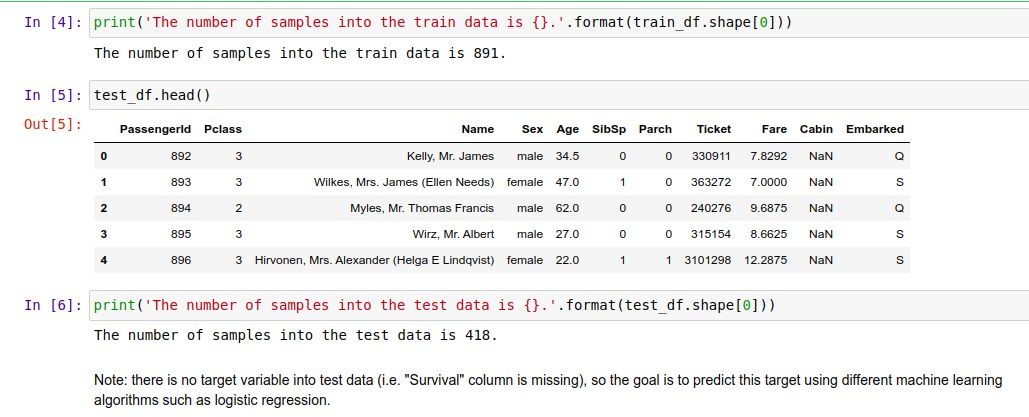
### Let’s load the data in a data frame and check how data looks like..



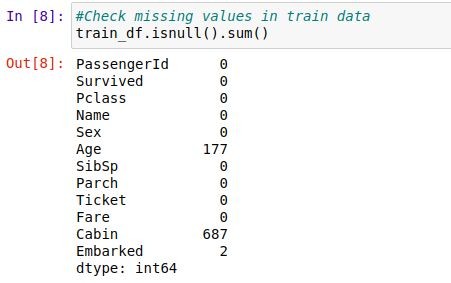


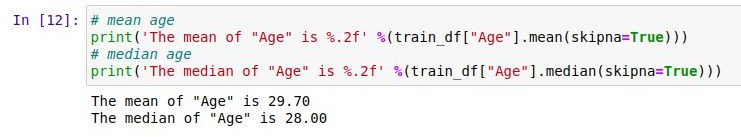
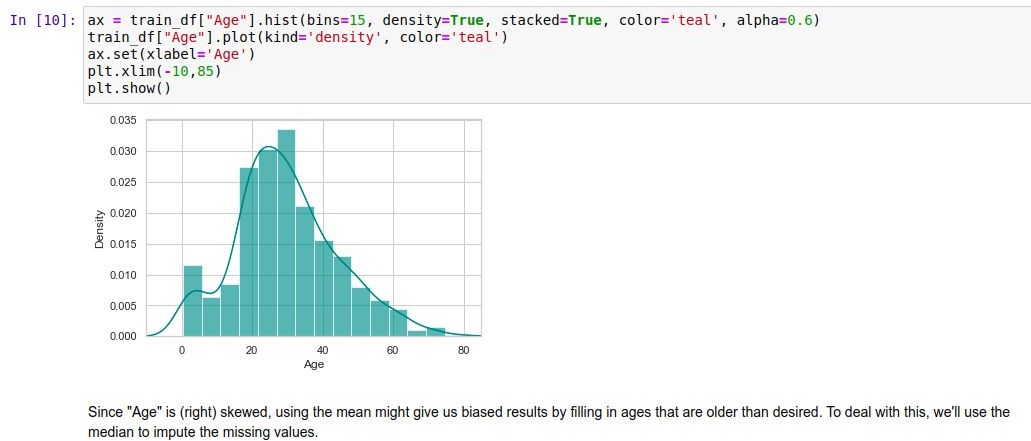
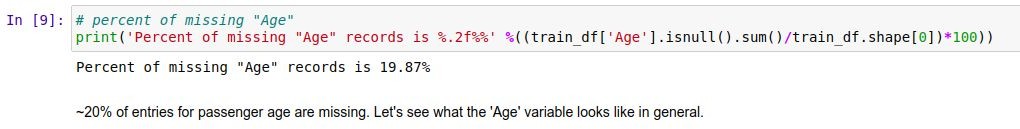
**Observations-**

1. As we can see above the data set has 12 columns or features. We already discussed what these features are all about earlier.
2. Here the Survived column is the target variable or class label. Target variable is the feature which needs to be predicted by our models.
3. We have numerical , categorical Type of features.

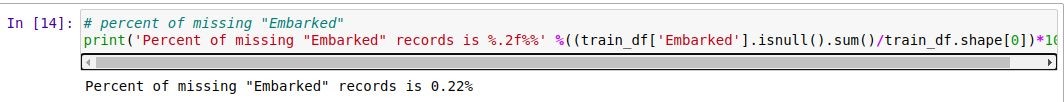
***Fetch some info about data***

**Data Quality & Missing Value Assessment¶**

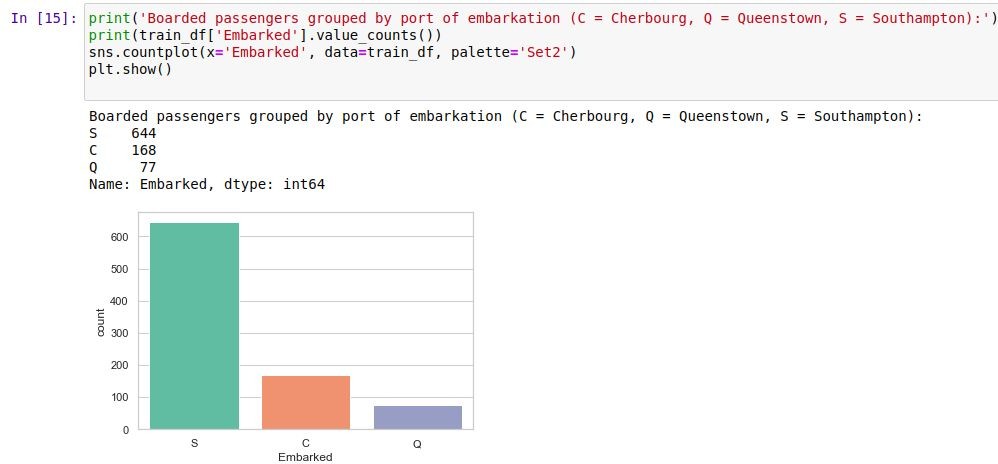


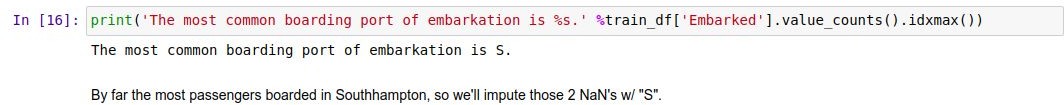
**Percentage of Missing age record:**

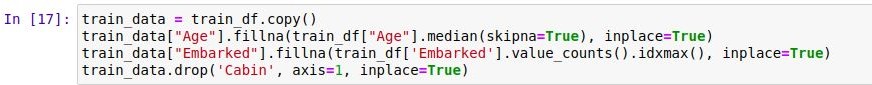
**77% of records are missing, which means that imputing information and using this variable for prediction is probably not wise. We'll ignore this variable in our model.**



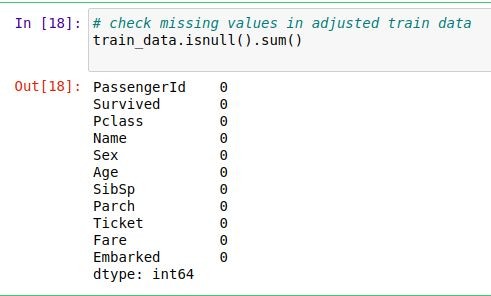
**There are only 2 (0.22%) missing values for "Embarked", so we can just impute with the port where most people boarded.**

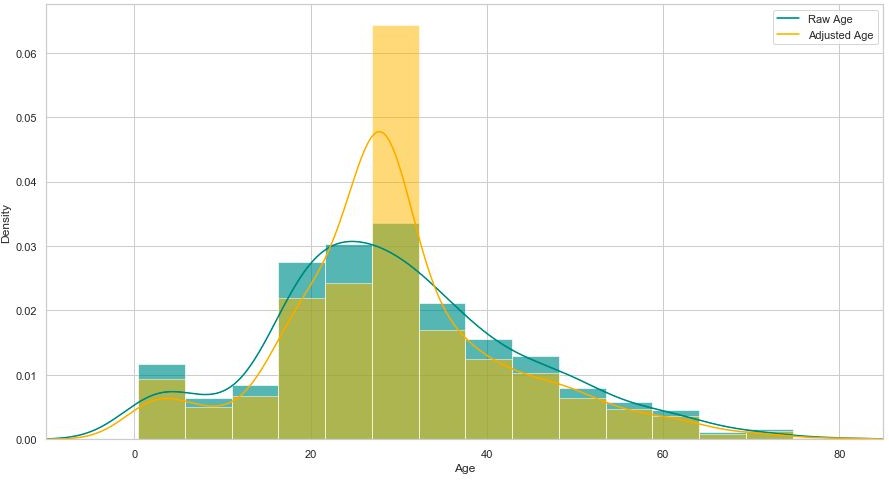
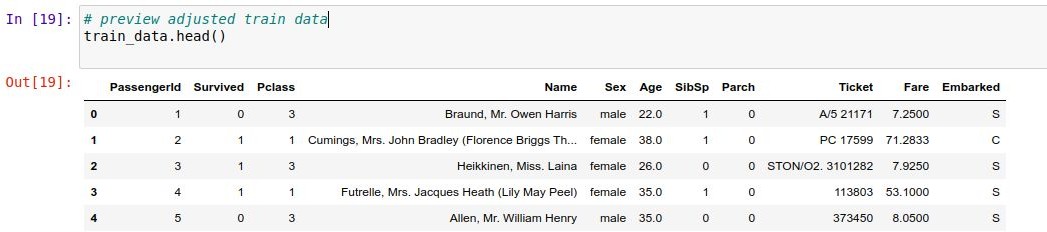




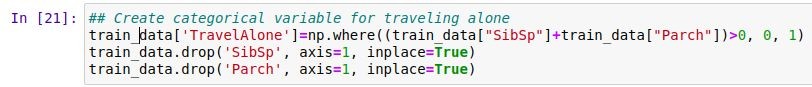
**Final Adjustments to Data (Train & Test)**

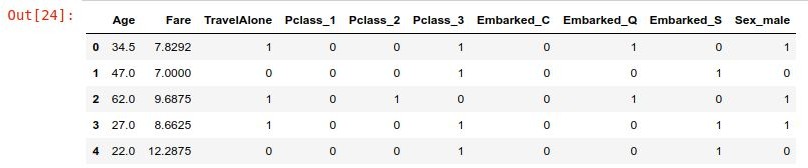
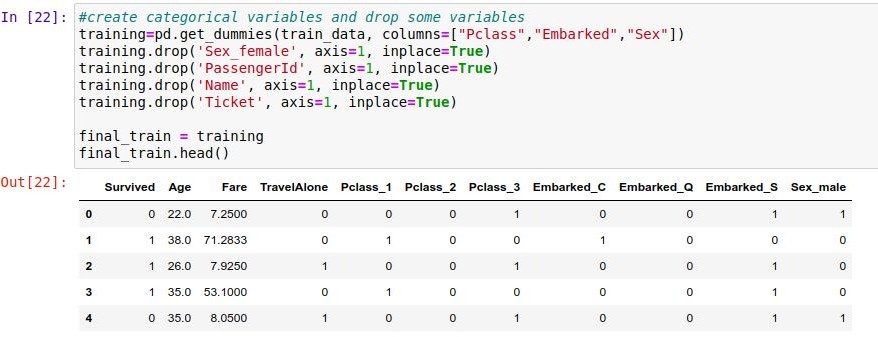
**check missing values in adjusted train data**



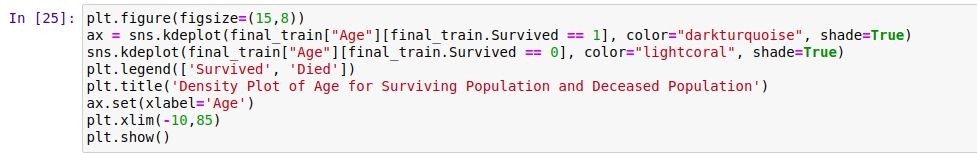
**preview adjusted train data**

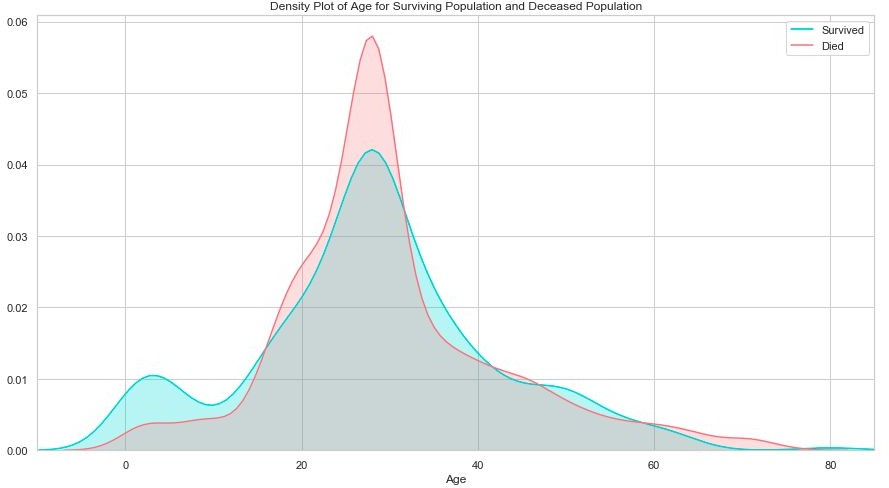
**Create categorical variable for traveling alone**

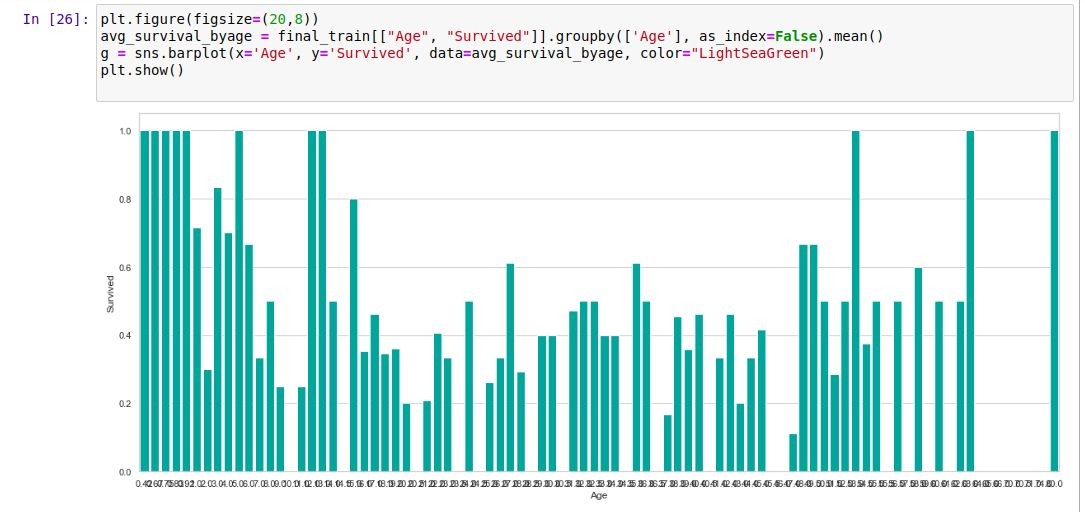


**I'll also create categorical variables for Passenger Class ("Pclass"), Gender ("Sex"), and Port Embarked ("Embarked").**

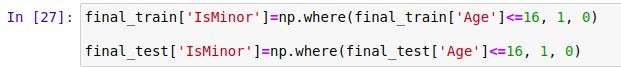
**Exploration of Age:-**

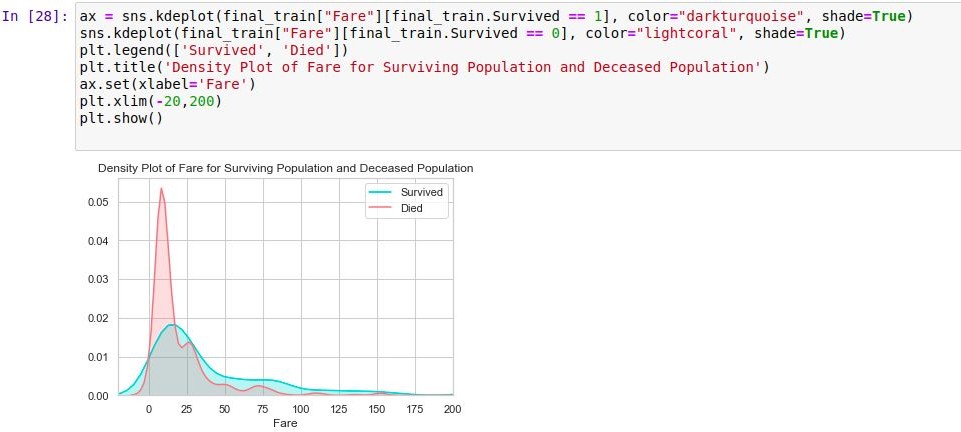




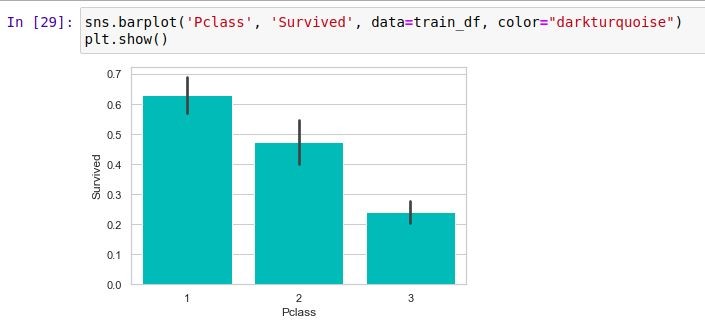
The age distribution for survivors and deceased is actually very similar. One notable difference is that, of the survivors, a larger proportion were children. The passengers evidently made an attempt to save children by giving them a place on the life rafts.

Considering the survival rate of passengers under 16, I'll also include another categorical variable in

my dataset: "Minor"

**Exploration of Fare**

As the distributions are clearly different for the fares of survivors vs. deceased, it's likely that this would be a significant predictor in our final model. Passengers who paid lower fare appear to have been less likely to survive. This is probably strongly correlated with Passenger Class, which we'll look at next.

**Exploration of Passenger Class**

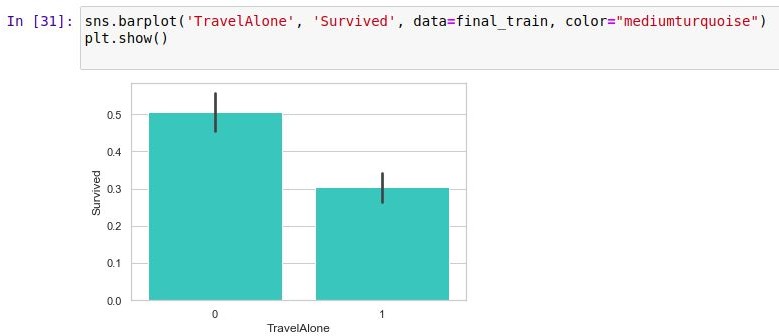
Unsurprisingly, being a first class passenger was safest.

**Exploration of Embarked Port:-**

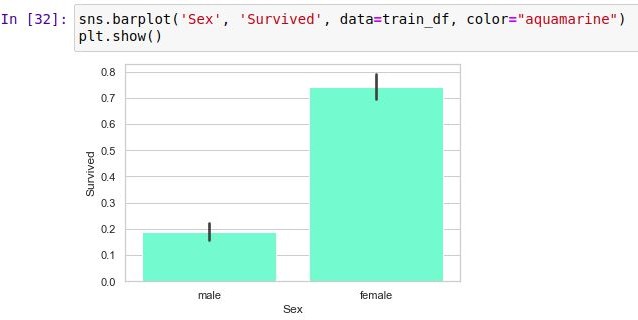


Passengers who boarded in Cherbourg, France, appear to have the highest survival rate. Passengers who boarded in Southampton were marginally less likely to survive than those who boarded in Queenstown. This is probably related to passenger class, or maybe even the order of room assignments (e.g. maybe earlier passengers were more likely to have rooms closer to deck). It's also worth noting the size of the whiskers in these plots. Because the number of passengers who boarded at Southampton was highest, the confidence around the survival rate is the highest. The whisker of the Queenstown plot includes the Southhampton average, as well

as the lower bound of its whisker. It's possible that Queenstown passengers were equally, or even more, ill-fated than their Southampton counterparts.

**Exploration of Traveling Alone vs. With Family:-**

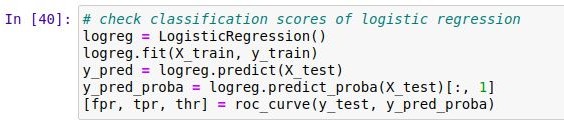
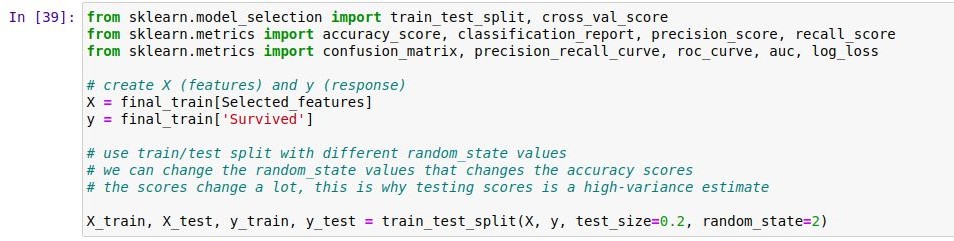
Individuals traveling without family were more likely to die in the disaster than those with family aboard. Given the era, it's likely that individuals traveling alone were likely male.

**Exploration of Gender Variable:-**

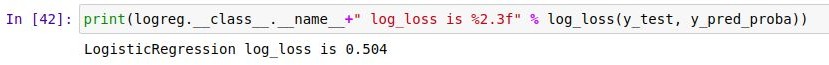
**This is a very obvious difference. Clearly being female greatly increased your chances of survival.**

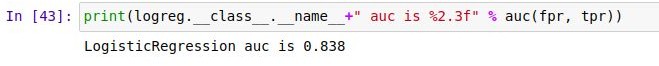
## Logistic Regression and Results

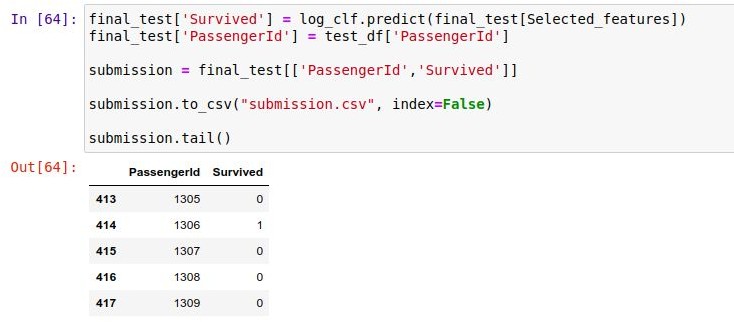


**Review of model evaluation procedures**

**Train/Test split results:**

**LogisticRegression accuracy is 0.782**





**Conclusion**

In this project we formulated the task of Survival of Passengers in Titanic Dataset.

Findings from EDA - If you were on "the Titanic", your chances to survive would be the

highest if you are a young female (or a child), have enough money to buy high fared tickets to get into a 1st class cabin, travelling in small family and getting aboard at the Port of Cherbourg.