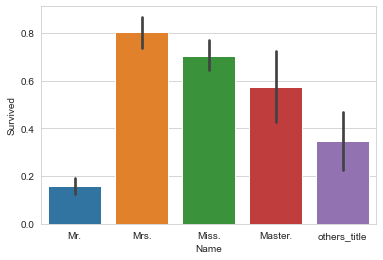
**Titanic - Machine Learning from Disaster**

Problem Statement: Build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data i.e name, age, gender, socio-economic class, etc.

There is train and test data set given,so first thing that we do is importing different python libraries for loading and cleaning/munging the data.

So, we clean each column for train and test data set and then apply the different model.

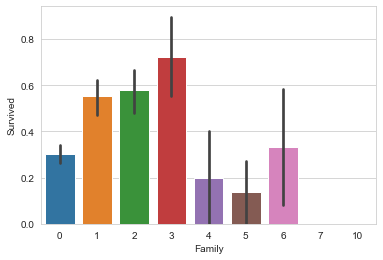
Different functions created for cleaning different columns .As **passengerId**  is having Id of passengers so there is nothing to do with that. 2nd column is **Name**  so replaced the larger name into the title as Mr.,Miss and so on given in data.



Above graph show Mrs. title have survived more.

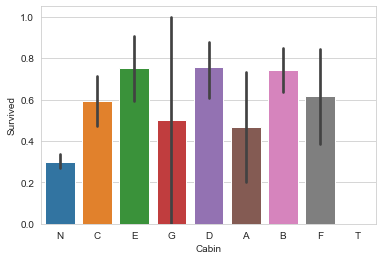
3rd column is **Sex** so no need to be clean as no null values is present.4th column is **Age** and it contain null values so we can replace the null values with the average of age but as avg. age can be find with reference from **Pclass** i.e Passenger class and it will clearly tell us what is average of people in each **Pclass**.

Sibling spouse (Sibsp) and Parent Children (Parch) joining together as a family to determine which family group has survived more.



As above graph shows that family with members 3 survived more.

Dropped the **Ticket** and **Fare** column from the data set as there is nothing to do with that. **Cabin** column contain more than 80% null values as we may drop it but replaced that null values with the **‘N’** and replaced each name in **Cabin** column with first letter to determine which **Cabin** have survived more.



Above graph shows that people survived less who are in unknown **Cabin** and more in B42.

Why this algorithm/model?

As I start with logistic regression. I split the data into a train set and test set. The train set is used to fit the model while test set is to check the performance of the model. Then I use K-Nearest neighbours classification for the dataset. But accuracy was lesser than than logistic regression and then applied Random forests and the accuracy was much higher.

Result/Conclusion: -

The accuracy rate for the best model i.e random forests model is 80.22% which indicates that this model correctly predicted the survival roughly 80%.