**Data Project- Titanic Data set**

In this project we will be analyzing data imported from the kaggle website. We will try analysing and answering the following questions:

1.) Who were the passengers on the Titanic? (Ages,Gender,Class,..etc)

2.) What deck were the passengers on and how does that relate to their class?

3.) Where did the passengers come from?

4.) Who was alone and who was with family?

5.) What factors helped someone survive the sinking?

* The columns in the dataset are: · 'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'.
* Now we plot a factor plot on the sex of the passengers and this reveals the disproportionately high male population in the ship. In the factorplot, the argument hue is passed as the ‘Pclass’ and reveals the distribution of the male and female population over the 3 different passenger classes.
* A function is defined to differentiate between males, females and children.This function is then applied over the dataframe and the resulting column is named ‘person’.The same distribution is then plotted over a factor plot with ‘Sex’ as the argument.
* A histogram is also plotted over the various bins of ages.
* The counts of each of the person types were also found using the ‘.value\_counts()’ method.
* The kde plot of the distribution of the different ages is plotted. Multiple kde plots can be constructed in the same plot using the function ‘FactGrid()’. Here, multiple kde plots are created which are distinguished by the sex of the passengers.
* The resulting graph interpretation was that the male population was more or less more dominant than the female population, but the female population was more in the age group of 0-20 years. The male population was dominant in the age group of 20-40 years and also marginally in the elder age group.
* The cabin column from the dataframe is stored in another Series. The first letter of this series is then taken from the series and stored in a list using the for loop. Then it is converted to a dataframe.
* The information about the decks are then plotted using the factor plot. The resulting plot showed an increased count in the cabin ’C’ and last in cabin ‘G’.
* Then we eliminate the defaulting attribute ‘T’.
* A factorplot is plotted over the embarked destination, with the hue as ‘Pclass’.The result  is that irrespective of the embarked destination, the population of the people in the lower economic class namely the ones in the 3rd class dominate.
* A new attribute alone is created as the sum of the ‘SibSp’ and ‘Parch’ attributes, and then in all the locations of alone where it equals 0, we change it to ‘alone’ or else give the value ‘with family’.
* After constructing the factorplot, the resulting observations are:
* Most of the passengers travelled in titanic alone and also,most of the people who travelled alone were males while the sex ratio of the ones travelled with family were almost the comparable.A large number of people who travelled alone comprised of the 3rd class. The ratios were comparable on the ones with family.
* To find out who all survived, the ‘survivor’ column is created by mapping with the ‘Survived’ column. On plotting the ‘survivor’ column using factorplot and lmplot, we get the following observations.
* The 3rd class people were the ones that suffered maximum percentage of casualty, followed by 2nd class, followed by the 1st class.
* The males have a higher percentage casualty irrespective of the class, with the second and 3rd being the worst hit.
* The females have a better survival, but among them, the lower the class, the lesser is their survival.
* The lower aged ones have better survival as compared to the middle and older aged ones for the second and 3rd classes.The middle aged people have the lowest survival for the 2nd and 3rd class, but interestingly for the 1st class, the middle aged people had better survival.
* On interpreting the dependencies of the survival , age and sex it was found that overall males had a lesser percentage of survival than females. But, as the age increases, the survival for males decreased and the survival for females increased.
* The chances of survival in deck ‘B’ was the highest while it was more probable for a person of deck ‘A’ to die.
* All the other decks have comparable survivors and casualties
* A person travelling alone is far more likely to die as compared to with family.
* Families have equal percentages of survival and death.