# Description of how the Model Works¶

I have cleaned the data to first remove the columns which had more than 2.9 Million null values.In the reamaining dataset I had removed the rows in which all the values were non in the entire row.I had removed the columns which had more than 2Million of data as null .After doing these steps, the dataset was significantly reduced to halftone number of columns.I have studied each and every property in the columns and safely filled the na values with mean, mode and median where ever required  respectively.I have not filled any na value with the zero values because I think that is not good for the illustration.

In the second question for five informative plots I have made stripplot,line plot,regplot,countplot and a scatter plot retieving informations on outliers of some data and other useful information regarding categorical data distribution.

I have then used the linear regression method from sklearn and applied it on the cleaned and neatly filled data.
Intially for the simpler model I have used less variables and tried the regression.TO enhance I have added some more columns so that the prediction is more accurate and that worked.My new model worked well with other more variables.I have used the trained dataset to train the model and then found out the predicted values in the data set .I have made copies of the dataset at some check points so that I can directly use those variables if I screw the dataset with some redundant function call anywhere.

At last I have added the predicted values in the submission.csv and submitted in Kaggle website.

# Evaluation of how the Model Works¶

The evaluation metric of the model is mainly based on the number of rows that are processed and the number of useless data that is removed from the data set.As far as the significant data is concerned I have maintained a good number of signification data with around 2.9 Million rows and 30 columns which was filled with safe metrics to avoid any wrong assumptions.The model was trained on the trained data set which had diverse values in the dataset which also included some outliers.

# Experiences/Surprises¶

```
This was my first exposure to data science and working
on an assignment.It took me nearly all the days of time
period allotted for the assignment to gather the syntax
and get to know what I actually have to do.I had the
following experiences and some surprises :
1.Dealing with the vast 3 million data.
2.Exposure to nan values and get to know how to fill
them
3.The syntax to calculate mean, mode and median and the
decision to take to fill the null values with one the
three.
4.Data Munging was the hardest part and took most of my
time.
5.I also took time to find small syntax mistakes such
as index = False , where I could easily screw up with
the data.
6.Re starting the Kernel sometimes when it gets stuck
had made me do the exercise a couple of time,
I feel I had a good exposure to learn many things in
this  assignment and thrive to do better in the
upcoming more assignments with lots of enthusiasm.
```