**DATASET – ADVANCED REGRESSION TECHNIQUES**

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

The dataset is based on residential homes in Ames, Iowa and has 79 Explanatory variables. This dataset lets you predict the final price of the home. I used a variety of plots to get an overview of data and used Random forest regressor for Analyze some data points. The target to predict is SalePrice which is property’s sale price in dollars. There are a lot of features like LotArea, Street, Alley, Utilities, Building Type, GarageType etc. I used python’s jupyter notebook to execute the code

**PART A – EXPLORATORY DATA ANALYSIS**

Columns which contribute to EDA are considered. A description of the dataset is given by calculating count, mean, standard deviation, different quartiles and maximum value. The line of code is dataset.describe()

Information on the dataset like the number of entries under one column and datatype of each column. Based on this we can figure out whether data is continuous or categorical.

**Value\_counts** function gives the number of entries under one category in a categorical column.

**Plots**

Histograms and density plots are used to show the distribution of a single variable. I plotted the density distribution for SalePrice. Box plot gives a distribution of data based on 5 basic mathematical number: minimum, 1st quartile, median, third quartile and maximum. Box plot is depicted for Number of bedrooms vs SalePrice. Strip plots is done for GarageType vs SalePrice. From this plot, we can conclude that attached garages have better selling prices than detached ones. As the dataset is focused on Iowa, garages are given more importance and hence have better saleprice. Another strip plot is done for Fireplace vs SalePrice from which we can conclude that houses with fireplaces are preferred. Since the place is in northern United states. It is seen that houses with central air conditioning are sold with better prices than houses that don’t have. Linear plots between number of bedrooms and saleprice shows that average saleprice is around 150000 when number of bedrooms feature is considered. And houses with 4bedrooms are sold at higher prices. Pair plots are done for a combination of variables. Features such as saleprice, yearbuilt, lotarea, garagearea are considered for pairplots.

**PART B – RANDOMFOREST REGRESSOR**

Both train and test datasets are read. An outlier function for calculating median absolute deviation is defined. Distplot for SalePrice and Natural log of SalePrice are plotted. I evaluated ranges and missing values. Differentiate categorical and continuous data. Missing data in continuous and categorical data are eliminated. These separated continuous and categorical data are integrated into train and test sets. Cross Validation score is calculated and graph is plotted. Feature significance is done to fit model with output features. The predicted calculated prices from regression and actual prices are plotted on a graph. Shows the deviation rate.