**Team Members: Vandana Kolli, Anusha Ramamurthy**

**Project 1:** [**House Prices: Advanced Regression Techniques**](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

The main idea of this project is to predict the final price of each home With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

**Data Analysis:**

The housing data set has 1460 rows and 81 features with the target feature Sale Price. As part of data analysis we made the following observation on categorical and non-categorical data.

# Missing values in the data:

* The categorical variables with the largest number of missing values are: Alley, FirePlaceQu, PoolQC, Fence, and MiscFeature.
* The numeric variables do not have as many missing values but there are still some present. There are 259 values for the LotFrontage, 8 missing values for MasVnrArea and 81 missing values for GarageYrBlt.
* We applied a technique where missing values are replaced by its mean.

**Future Implementation Ideas:**

We are trying to identify a better method to replace missing values. Replacing by mean doesn’t make sense, since the values like LotFrontage may depend on the street where the house is located.

Apply KNN to identify how to replace the missing values.

Add categorical data as a separate column to be 0 for absent, 1 for presence and check how the linear regression values change.

**Feature Extraction:** Identify features that strongly influence the regression line, and eliminate features that don’t influence the Sale Price. Thus reducing the feature set.

Apply other algorithms like SVM or Random forests instead of Linear Regression to check if the sale price improves.

**Histograms:**

We tried to plot different histograms to get an idea of how the sale price is impacted by various attributes.

**Algorithm:**

We used some part of train data to train the system and rest of the train data (last 20 records) to test the actual sale price to predicted sale price. We applied linear regression on the clean data to predict sale price.

**Questions:**

We are confused about which algorithm to use for predicting the house price?

How do we use Categorical Data in linear regression? We are confused how they will influence the regression line.

Is clustering a better method for this dataset?

# Project 2: [Out brain Click Prediction](https://www.kaggle.com/c/outbrain-click-prediction)

# Questions:

# How to interpret confidence values?

# There are duplicate values, which values should we retain?

# Can we consider this to be a clustering or Estimation Maximization problem?