**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.

**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:**

# [Outbrain Click Prediction](https://www.kaggle.com/c/outbrain-click-prediction)

In Outbrain click prediction we are expected to predict which adds the users are more likely to click on. A better recommendation algorithm is needed.

**Data Analysis:**

**Dataset:** Because the dataset is huge, we are trying to figure out a way on how and where to use the dataset. Also we are trying to understand different attributes in different files.

**Questions:**

1. For huge datasets like this one, what will be the best way to store and analyze it – we are considering postgresql, SQL Server are there any other better options?