House Prices: Advanced Regression Techniques

# Springboard Capstone Project #1

# Introduction and Objective

Capstone Project 1 Proposal

House Prices: Advanced Regression Techniques from Kaggle competitions is chosen to be first Capstone Project.

Typically, a home buyer would not describe their dream house beginning with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. The project is to explore 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa to predict the final price of each home.

Real estate sales agents, sellers and buyers will be interesting in predicting the sales price as it will help them to set their listing price, offering price and final sales price.

The dataset from Kaggle is modified from the Ames Housing dataset which was compiled by Dean De Cock for use in data science education. It's an incredible alternative for scientists looking for a modernized and expanded version of the often-cited Boston Housing dataset. The data set is provided in csv files.

The work would include data wrangling, exploratory data analysis and machine learning.

Creative feature engineering and advanced regression techniques like random forest and gradient boosting skills are needed to complete this project.

The result will be submitted to Kaggle. It will be evaluated on Root-Mean-Squared-Error (RMSE)

between the logarithm of the predicted value

# Data Wrangling

## 1. Data Set Variables

The data set is provided by Kaggle. The data is loaded in to Pandas data frame without difficulty.

In order to understand our data, we should look at each variable and try to understand their meaning and relevance to this problem. The detail description of each variable can be found in provided "data\_description.txt". We can create a spreadsheet with the following columns:

* **Variable** - Variable name.
* **Type** - Identification of the variables' type. There are two possible values for this field: 'numerical' or 'categorical'.
* **Expectation** - Our expectation about the variable influence in 'SalePrice'. We can use a categorical scale with 'High', 'Medium' and 'Low' as possible values.
* **Conclusion** - Our conclusions about the importance of the variable, after we give a quick look at the data. We can keep with the same categorical scale as in 'Expectation'.
* **Comments** - Any general comments that occurred to us.

## 2. Missing data

The Missing data can imply a reduction of the sample size. This can prevent us from proceeding with the analysis. We need to check prevalent of missing data and if missing data is random or have a pattern.

If more than 15% of the data is missing, we should exclude the corresponding variable in our analysis. Variables 'PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQu', 'LotFrontage' will be excluded.

'GarageCond', 'GarageType', 'GarageYrBlt', 'GarageFinish' and 'GarageQual' have the same percentage of missing data. These should be for the same set of obervation. Since the number of car space is the most important factor for garage, we will take 'GarageCars' into our analysis and exclude the other Garage\* with missing data.

'MasVnrArea' 'MasVnrType', 'BsmtExposure', 'BsmtFinType2', 'BsmtFinType1', 'BsmtCond' and 'BsmtQualwe’ are not essential variabls. Furthermore, they have a strong correlation with 'YearBuilt’, which we will considered. Thus, we will not lose information if we exclude these.

Finally, we have one missing observation in 'Electrical'. We'll delete this observation and keep the variable.

In summary, to handle missing data, we'll delete all the variables with missing data, except the variable 'Electrical'. In 'Electrical' we'll just delete the observation with missing data.

Outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviors. We'll just do a quick analysis through the standard deviation of 'SalePrice' and a set of scatter plots.

## 3.Outliers

### Univariate analysis

The primary concern here is to establish a threshold that defines an observation as an outlier. To do so, we'll standardize the data by converting sale price data values to have mean of 0 and a standard deviation of 1.

Low range values are similar and not too far from 0. High range values are far from 0 and two values larger than 7 are really out of range. For now, we'll not consider any of these values as an outlier but we should be careful with those two values larger than 7.

### Bivariate analysis

From ‘SalePrice’-'GrLivArea' scatter plot, the two values with bigger 'GrLivArea' seem strange and they are not following the crowd. Therefore, we'll define them as outliers and delete them. The two observations in the top of the plot look like two special cases, however they seem to be following the trend. We will keep them.

From ‘SalePrice’-' TotalBsmtSF ' scatter plot, The ones with TotalBsmtSF > 3000 look like outliers. But we can keep them for now.