**CS 6350 Big Data Project   
House Price Prediction**

**1. Introduction**

The Kaggle house price dataset provides feature variables that describe different aspect of residential houses. My aim is to predict the final price of each house based on the features.  
  
**2. Problem Definition and Algorithm**   
  
2.1 Task Definition

The goal is to predict sales prices of the houses using advanced regression models.

The task involves data exploration to analyze the structure and form of data followed by necessary feature engineering and pre-processing to create a trainable data set to train the models.

This is followed by evaluation to verify the accuracy of the model on the validation data set.

The model with the highest accuracy with an appropriate tradeoff between bias and variance to overcome underfitting or overfitting of data respectively is to be used to predict the sales price of the houses.  
  
2.2 Algorithm Definition

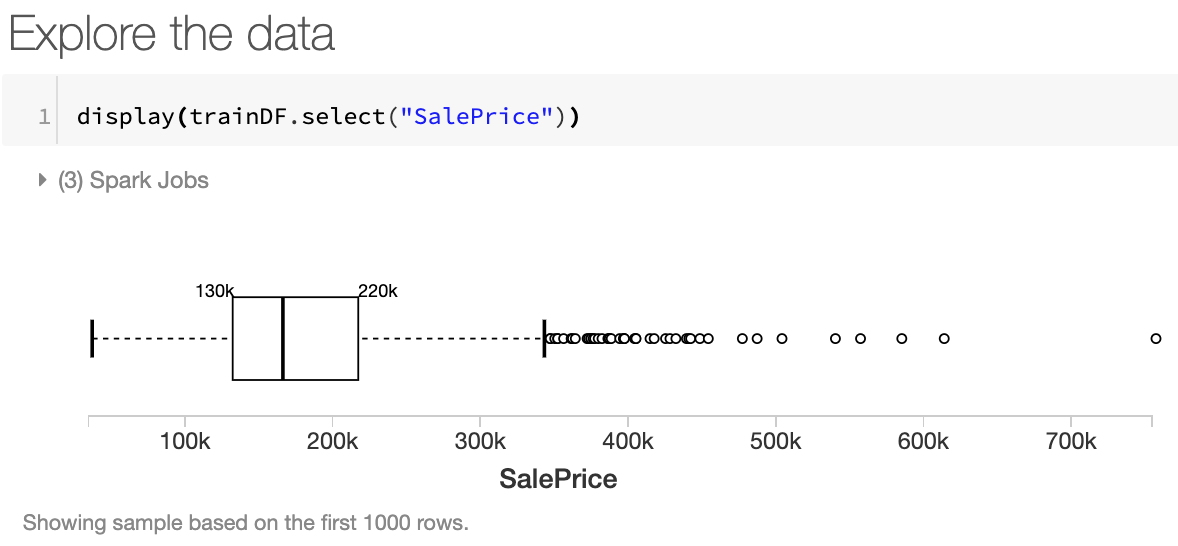
The different regression algorithms used are Linear Regression, Decision Tree Regression, Random Forest Regressor, Gradient Boosting Tree Regressor.  
  
**3. Experimental Evaluation**   
  
3.1 Data Exploration

The training data has 1460 rows, 79 feature columns, an Id column, and a SalePrice column, which is the label. The testing data has 1459 rows and does not have the label column.

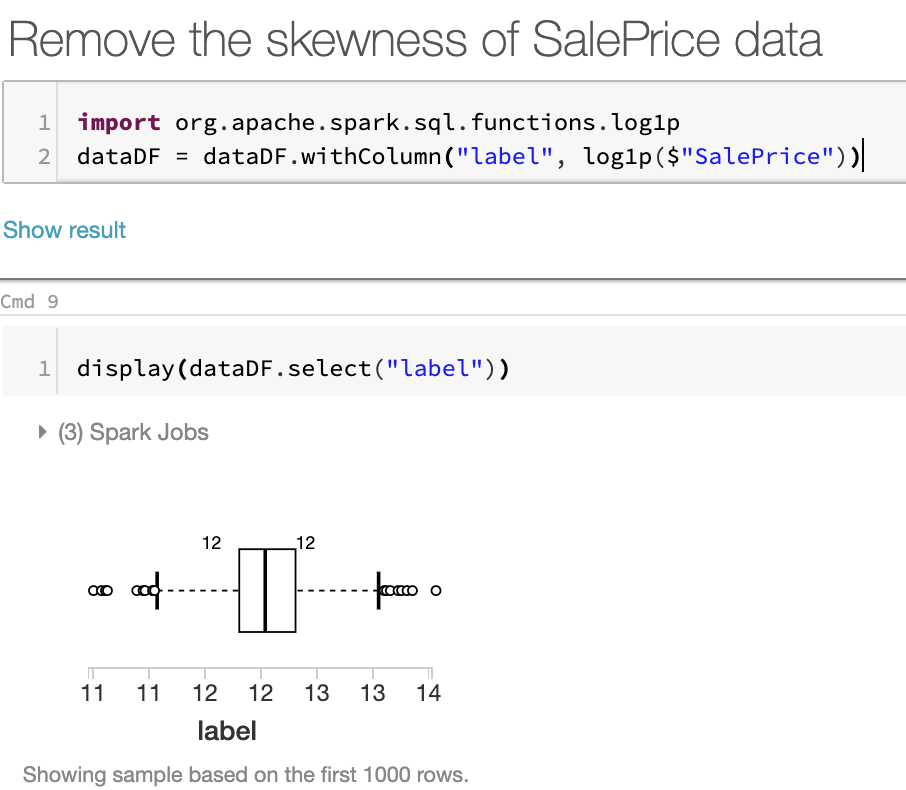
<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

3.1.1 Label

The plotted histogram of the SalePrice column shows that it is skewed to the left.



So, it is modified with log to get it uniformly distributed. Log operation is used to remove skewness as it is useful in cases where only few points are much larger than the bulk of the data.



The label data is now more balanced and suitable for training and prediction purposes

3.1.2 Important features

Possible important numerical features identified and how SalePrice holds its relationship

* OverallQual => Linear
* YearBuilt: Original construction date
* TotalBsmtSF: Total square feet of basement area => Sometimes Exponential
* GrLivArea: Above ground living area square feet => Linear

Multivariate study. Understand how the dependent variable and independent variables relate.

Basic cleaning.

Clean the dataset and handle the missing data, outliers and categorical variables.

Test assumptions. Check if the data meets the assumptions required by most multivariate techniques.

Used indexer to transform categorical features into numerical form.

Used assembler to form a single trainable feature taking all the features together.

Created a pipeline comprising of the pre-processing (indexing and assembling) stages and the regression stage with the various models.

Next steps:

Use parameter grid for the hyperparameter tuning of the various models, out of which the best model will be chosen using k-fold Cross-validation.

The models are to be evaluated based on their accuracy along with other evaluation metrics.

The best model will be used to predict the labels for the test set