**Project Proposal**

**Project Title**: Prediction House Price by Multiple Data Analytic Techniques

**Research Objectives**:

This project explores the relationship between house price and a lot of features, using the raw Ames house price data. This project aims at constructing a reliable model which can be applicable to real estate industry for house price predictions.

**Research Question**:

1. What are the most important factors, and what’s the relationship between those factors with house price?
2. How strong is the relationship between those factors and the house price?
3. What’s the most efficient model among Multiple Linear Regression, Lasso, and Neural Network?
4. Are those three methods both give a good explanation of the first 2 problems?

**The Dataset:**

The initial dataset is approximately 942KB, and comes from <https://www.kaggle.com/prevek18/ames-housing-dataset>, which has 2930 observations and 81 variables including 38 numerical variables and 43 categorical variables.

**Tools:**

1. Exploration the dataset using scatter plot and box plot, and Feature Engineering
2. Feature selection using forward section
3. Build the model using multiple linear regression (MLR), and Lasso
4. Experiment further with Deep Learning method neural network, and compare the accuracy with MLR and Lasso

**Reference:**

The literature on house price prediction is limited, though there’s significant commercial meaning to do house price prediction. However, the fundamental prediction methods are very similar in different industries, so I have surveyed two papers in forecasting in power industry to get an insight to complete the project.

(1) A Naïve Multiple Linear Regression Benchmark for Short Term Load Forecasting

Hong Tao (2011) implemented a Multiple Linear Regression method for short term load forecasting at a medium-sized US utility, which involved feature engineering, 3 order polynomial and interaction between variables.

(2) Weather station selection for electric load forecasting

Hong Tao (2014) offered Greedy algorithm to select weather station and to determine how many and which weather stations to use for a territory of interest, which can be used in feature selection.