# **Project Proposal**

Team members: Yuqiao Han, Shang Liu, Jiayu Yao

Team name: House Price Team

**Motivation**

Investment is a business activity that most people interested in, and in particular property investment has been increased significantly since 2011. The relation between house price and the market is an important motivating factor for house price prediction, while there is no accurate measurement of house prices. Real estate and house price is always a hot topic in the market, which related to many other sectors in economics and pushes the market trend. For a developer, a fundamental problem is to determine the selling price of the house. Developers must calculate carefully and determine the appropriate method, so a mechanism or a model to evaluate and predict house price is needed. For buyers, as house price increases, it will be great if people can predict future house prices, stepping ahead of the overall market to obtain valuable and cost-effective houses. On the other hand, while other people piling their money into real estate, a good house price predictor will allow many to sell their assets before the market crash. Therefore, an indicator of house price can not only be an assist for investment, but also a protector for future loss.

**Hypothesis**

1. Linear correlation between input attributes and output prices.

2. Attributes such as building class, general zoning, available utilities and types of sale have more weights than attributes such as the number of kitchens, the number of fireplaces and screen porch have.

3. Attributes are correlated. Different combinations of attributes can yield more indications of results than treating attributes independently.

**Technical approaches**

First, we would do preprocessing, to deal with the missing data and outliers. The dataset includes 79 attributes of a property, to avoid “the curse of dimensions”, we would use Principal Components Analysis method to reduce the dimensions.

After we choose the principal components, we would apply 3 models to predict the sale price of a property and compare their performance. These models include the linear regression model, the SVR model, and the Random Forest model.

**Dataset**

The dataset includes a train.csv, a test.csv, a sample submission.csv and a data\_description.txt. In the training data, there are 81 columns and 1460 rows. The columns include 79 attributes, describing multiple aspects of the house, a column of house ID, and another column of the final sold prices. The test data has 80 columns, the prices column is excluded.

These attributes of houses describe the conditions of houses in various aspects: from the building class, size, type of road access, the general shape of property, types of available utilities, conditions of dwellings, roofs, basements, kitchens, exterior materials etc. These attributes include both categorical ones and numerical ones.

The sample submission file is a benchmark solution from a linear regression on year and month of sale, lot square footage, and nuMSSubClass: Identifies the type of dwelling involved in the sale. It includes 2 columns which are the house ID and the predicted sale price.

The data\_description.txt file explains the meaning of attributes and their possible values.

**Risks**

Missing Data: impute

Too many features: reduce dimensions

Training sample not enough: cross-validation

**Goals**

* Discover which kind of features influence the price of a house, by choosing from 79 different attributes.
* Predict the price of a new house when giving some of the value of the features.
* Give the business intelligence advice to a house selling company and investor of which kind of house has higher value.

**Deliverables**

* Model Architecture & Code
* Report: Model results & Comparison & Business intelligence advice

**Data Source**

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