Kaggle has a whole bunch of dataset

Train to get highest prediction accuracy

Ames is a college town

Each obs is a house

Goal:

* Perform descriptive data analysis to gain business (i.e. housing market) insights
* Build descriptive machine learning models to understand local housing market
* Build predictive machine learning models for the local house price prediction

Close to house

Areas of house renovated

Role:

* You’re a housing market consultant
* Data sceintis for online real estate database sompnay which provides house price estimation

Want to understand the local housing market in a data driven way

Key idea to understand the local housing market

Tools you are expected to use:

* **Data analysis and missing value imputation**: pandas dyplyr and the accompanying visualization tools, some linear regression or logistic regression
* **Descriptive models:** feature selection, feature engineering, multiple linear regression, Python statsmodels, penalized linear regression, stepwise regression (AIC,BIC, etc.) some gridsearchCV, KNN, SVR, Random Forest, Gradient Boosting and potentially other models
* **Predictive models:** feature selection, feature engineering , gridsearchCV, multiple/penalized linear regression, gridsearch, Gradient Boosting and potentially other models.

You can use machine learning models beyond those we list above, but don’t make these optional models your first priority. Doing this would limit your time spend on the more fundamental part of the project. This is not wise!

The art of **feature engineering** includes generating new features, or selecting optimal features to feed into the machine learning models. If your goal is to improve accuracy, then the optimal features would mean the feature combination that offers the highest predictive accuracy. However, if your goal is to fit an optimal descriptive model for the purpose of improving business efficiency/profitability, the most accurate model might not give you the insights you want. As a modeler, you need to gauge which is most appropriate based on your objectives.

Clever ideas for feature generation are often inspired by high quality data analysis. Thus, we will suggest a few research questions to point you in the right direction. Chronologically, your team should work on data analysis and modeling simultaneously. Do not wait to start modeling until after you complete your data analysis phase, you will be pressed for time.

Focus on linear models

Use address to get long lat to calculate distances from various places use geopy.

Range of dates 2006 to 2010

Calculate price per square foot as a feature

Sale