Capstone Project Proposal

# Zillow Prize: Zillow’s Home Value Prediction

Link: <https://www.kaggle.com/c/zillow-prize-1#description>

1. What is the problem you want to solve?
   1. Predict the log-error between Zillow’s estimates and the actual home sale prices, using all the features of these homes and
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?
   1. My client is Zillow and this project will help them better minimize the discrepancies between their estimates and the actual prices, which help the company provide more accurate predictions for its users.
   2. My client will adjust it estimates based on my analysis
3. What data are you going to use for this? How will you acquire this data?
   1. Zillow provides several datasets:
      1. properties\_2016.csv - all the properties with their home features for 2016.
      2. properties\_2017.csv - all the properties with their home features for 2017
      3. train\_2016.csv - the training set with transactions from 1/1/2016 to 12/31/2016
      4. train\_2017.csv - the training set with transactions from 1/1/2017 to 9/15/2017
   2. I may use other data based on geo information
4. In brief, outline your approach to solving this problem (knowing that this might change later).
   1. Import data into SQLite to make it easier to manipulate in Python (the datasets are as big as 600+ megabytes. Hope SQLite will help me deal with it)
   2. Doing some data wrangling as there are many blanks (will learn data wrangling next week)
   3. Do linear regressions on each of the numerical parameters and correlation analysis on categorical parameters
   4. After identifying the ones with strong predicting capability, eliminate the ones with limited contributions to the price estimates
   5. More works to be done (should get more idea as I progress in my data science courses)
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.
   1. The format of deliverables are well defined by Zillows, which are the errors between their estimates and the actual home prices for each of the properties.