**ZILLOW PROBLEM STATEMENT**

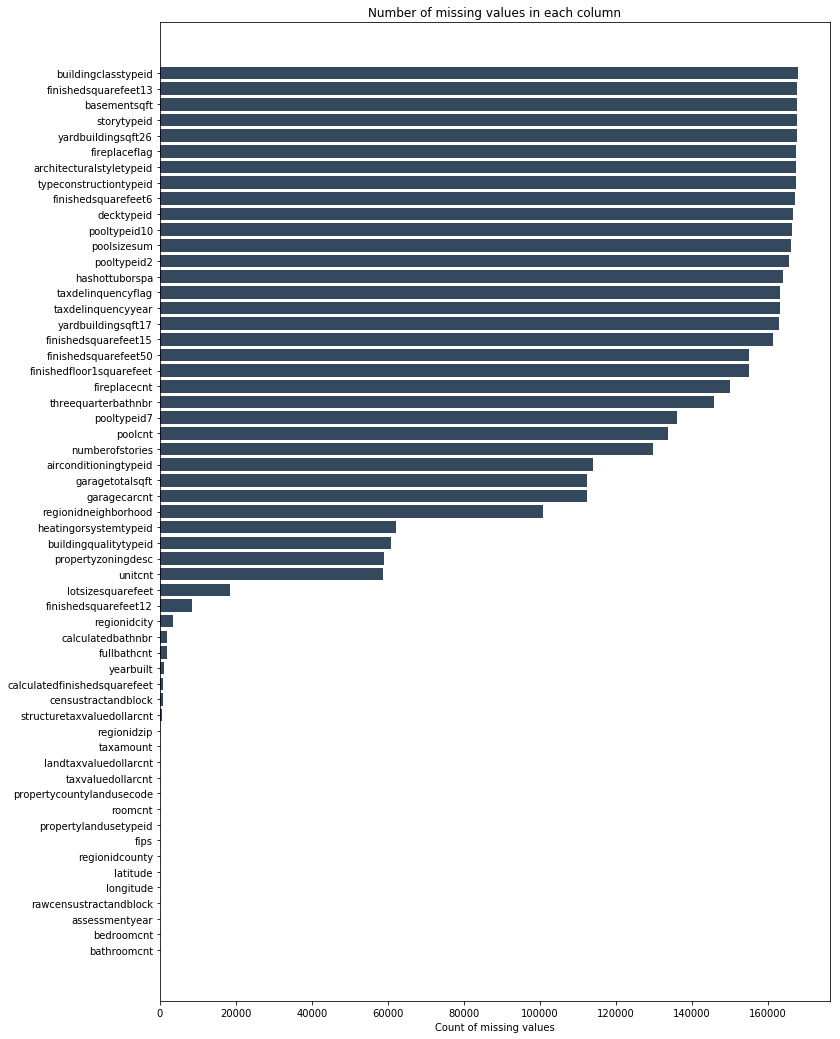
# The Problem Statement

We have to make predictions on logerror defined as difference of log(Zestimate) and log(actual sale price).

We have to make predictions on 6 different time lapses – i.e., 201610(October 2016), 201611, 201612, 201710, 201711, 201712.

# Approach

Upon inspection of features of the properties in both 2016 and 2017 data, we see that a huge chunk of data is missing.



For the training of this data, I have considered only the data which have less missing values than unitcnt (which has about 35% missing data) and replaced missing values with the respective mean of each column.

The regressor used is XGB regressor as evident by most popular kernels. The data is split 75% and 25% into train and validation sets, and MAE on our validation set comes around 0.069.