**Working Notes for ML Project**

General Workflow:

* Data Work
  + EDA
  + Feature Engineering
  + Impute missing values using random forest
  + Create as automated pipeline to ensure exact same code is run later on test data
  + Do we want to bring in data such as quality of schools, crime statistics, proximity of registered child molesters, census demographics, etc., based on neighborhood? As a former Realtor, these are things I cover w/my clients when buying. Links to these resources is required by AAR (Arizona Association of Realtors) in the Buyer’s Guide which we go over w/each buyer and have them sign prior to making an offer on any home. I specifically counsel each buyer that it is up to them to do the due diligence in EVERY area which matters to them… the buyer’s guide provides all links for them to research these ‘features’ before making an offer.
  + ?
* Modeling Work
  + MLR
  + Random Forest
  + Boosting
  + Testing for underfit vs overfit
  + Use MSE for comparing models
  + ***Is it necessary to try all different modeling techniques or does it make sense to fast forward to XGBoost (based on experience)?***
* ?

Questions:

* Should I use R at all… maybe for some of the EDA?

Things to mention in Overview:

* Data used covers most volatile time for housing prices in recent history… first time you see housing prices trending down due to 2006 mortgage crisis. In a way, the entire data set is an ‘outlier’ as to the way housing prices perform over time.
  + We should also see more sale types of ‘foreclosure’ as well due to this.
  + Except for… using this data for learning purposes, nobody would choose this particular data for modeling purposes.
* Ames, Iowa specifics… cold, etc.

Feature Engineering:

* A non-value in some features has meaning… no pool sq footage = no pool, but is this reconciled by the model anyway?