Goal: Create a Regression Model over Spring Break and Evaluate its Effectiveness

Steps:

1. Import required packages and models
2. Import Data
3. Clean Data
   1. Turn categorical variables into numerical
   2. Remove rows with sales price missing or less than 100,000
   3. Remove rows with missing square feet values (for now)
4. Data Features to Use

* Borough
* Building Class Category
  + If this feature proves to be significant, use Building Class at Present instead
* Block
* Lot
* Residential Units
* Commercial Unit
* Total units
* Land Square Feet
* Gross Square Feet
* Age at time of sale (create from year built column)
* Tax Class at present
* TARGET VALUE (Y): Sales Price

1. Split Data: Test/train/holdout
2. Run Model: random forest regression (Pratyush), linear reg (Kathleen, ridge/lasso regression (Alex)
3. Tuning

* check for multicollinearity
* remove insignificant features

1. Evaluating

* Use evaluation method that SciKitLearn specifies in documentation
* precision/accuracy
* roc/auc

NICOLE

How to deal with missing columns?

Square footage

* Use the rows that are intact
  + Break out into test and train data
  + Look at accuracy score
  + If model is good at predicting the square footage use it, if not then only use the rows that are filled
* Don’t use the sales price in predicting the square footage