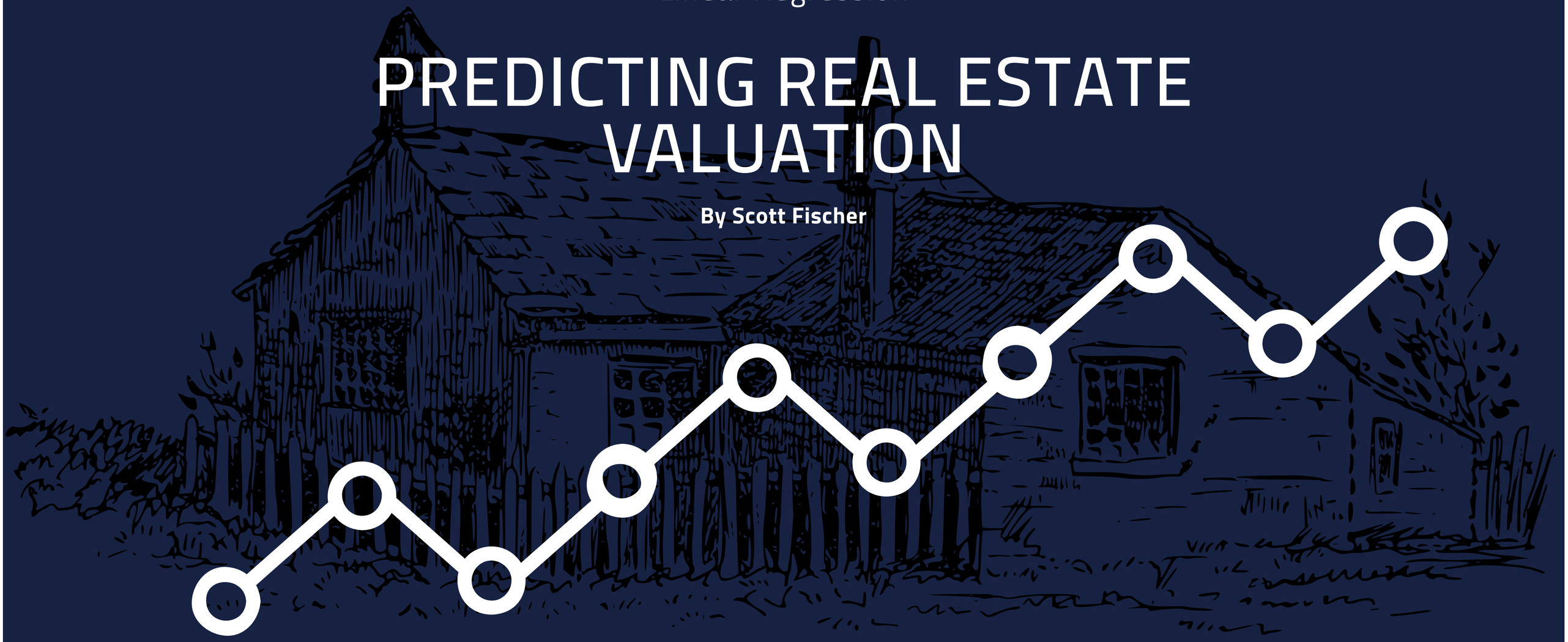
Linear Regression

PREDICTING REAL ESTATE VALUATION

By Scott Fischer



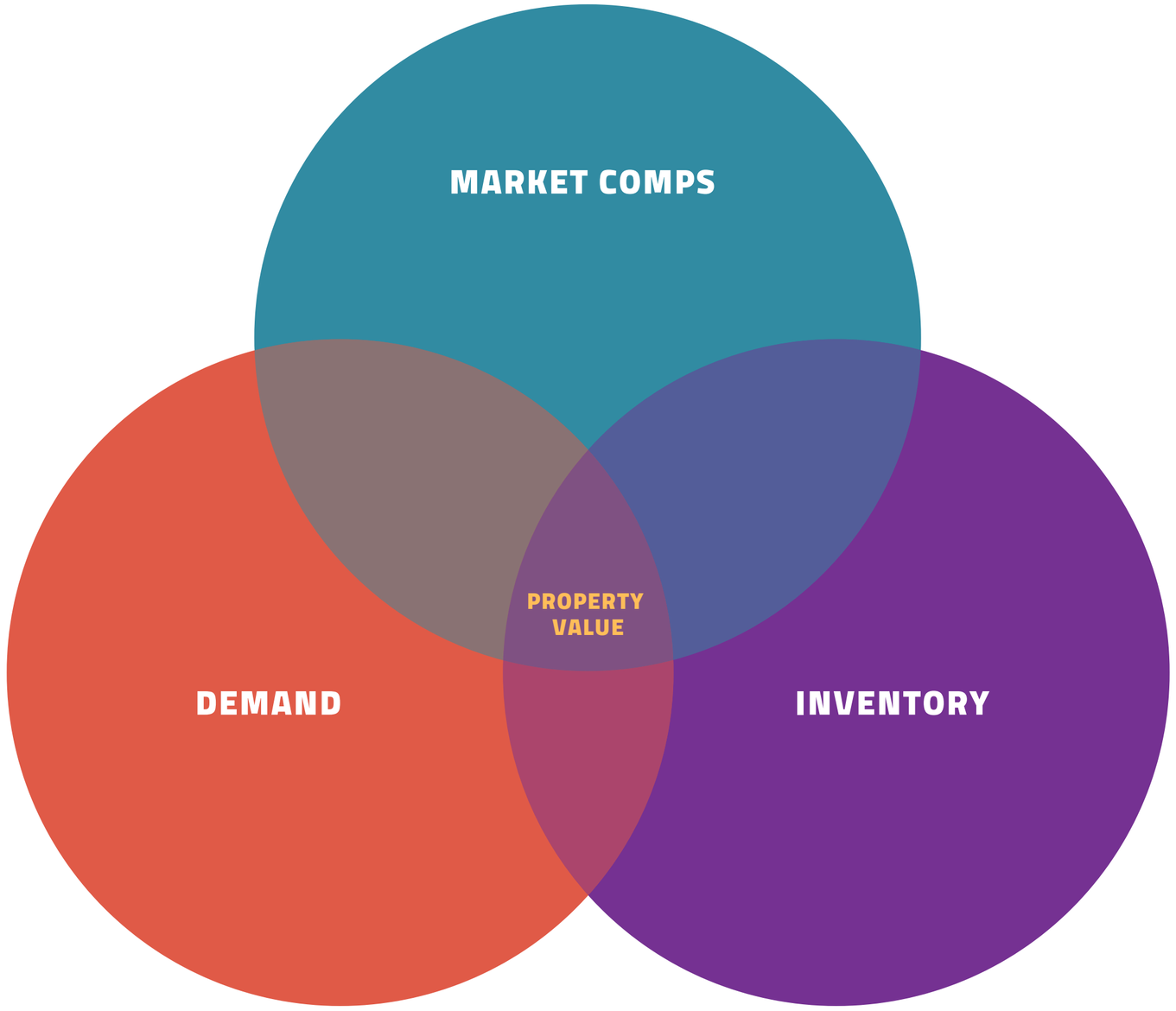
WHY WE CARE

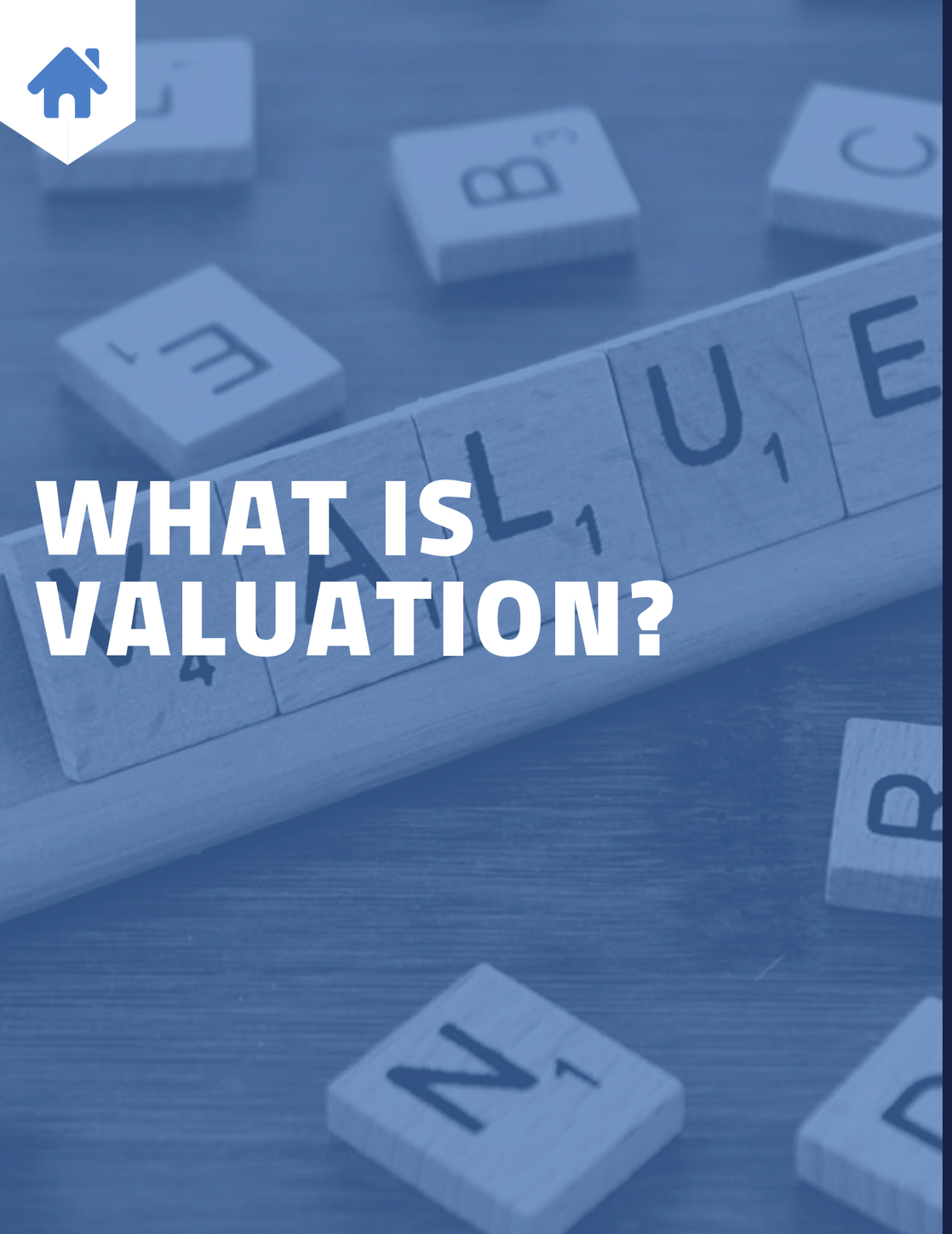
Accurate house prices valuations help create a more stable market.

Market Stability directly impacts home buyers, sellers, and renters.

WHAT IS VALUATION?

WHAT'S IT BASED ON?





WHAT IS VALUATION?

WHAT'S IT BASED ON?

Informed decisions are king! Being aware of any of the three elements of market valuation can make a huge difference in pricing on property.

Recall my prior seasonality analysis: while the number of transactions in spring in summer are double what they are in fall and winter, there isn't a statistically significant difference in cost.

The results of that study could be due to several things, but one thing is for certain: with that number of transactions demand is high and inventory is low! A savvy seller would leverage this information and try to push for a higher price.

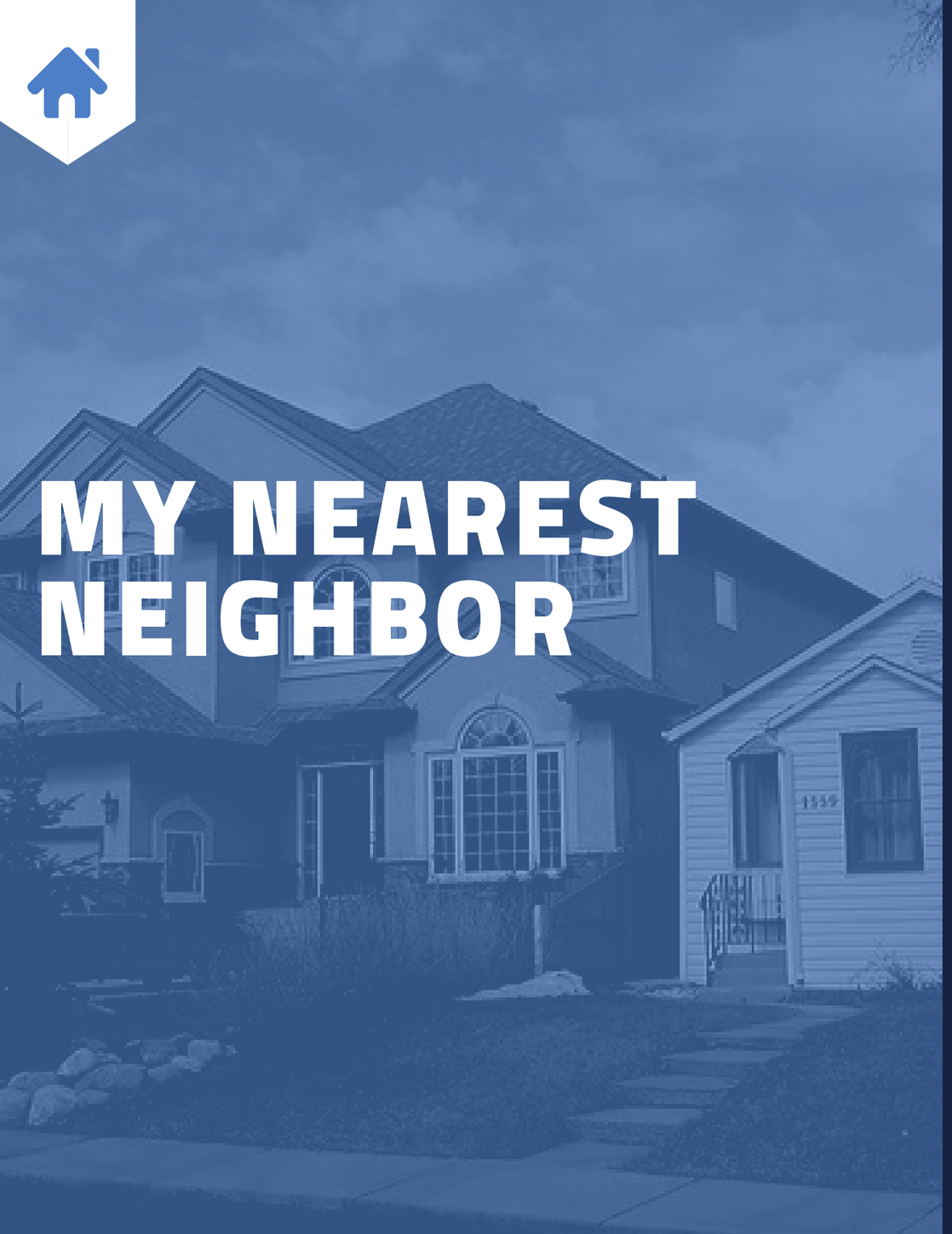
MY NEAREST NEIGHBOR

HOW CMA'S WORK

Comparative market analysis is a method for agents to give an estimated value on property. The steps usually look like the following:

Agent gives anecdotal evaluation of market based on subjective experience

These "evaluations" are usually reactive and based on feelings.

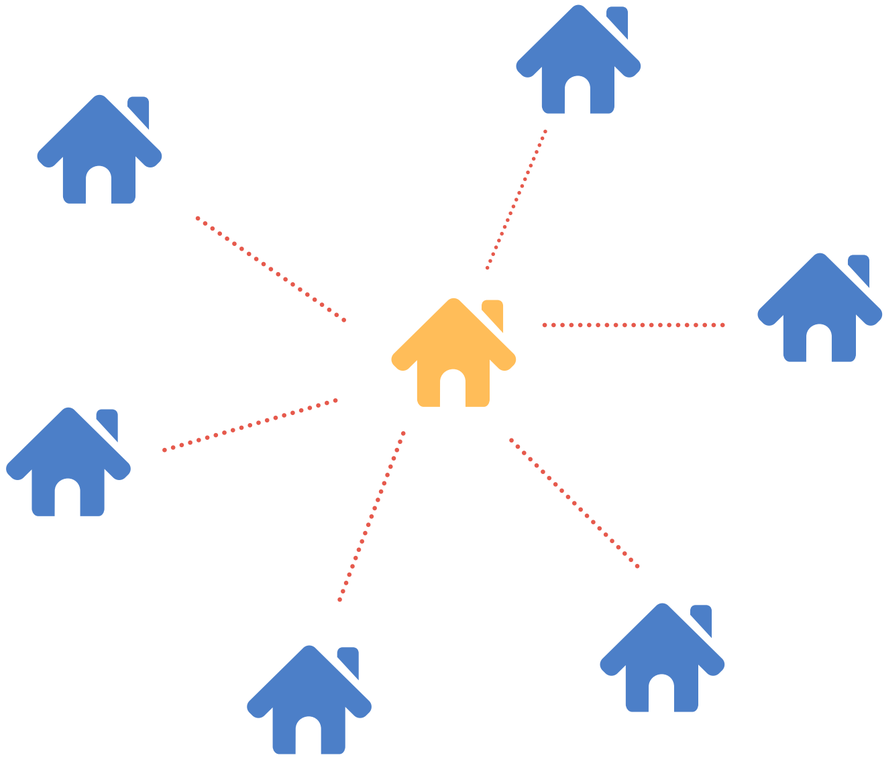


MY NEAREST NEIGHBOR

HOW CMA'S WORK

Comparative market analysis is a method for agents to give an estimated value on property. The steps usually look like the following:

Agent picks 5 - 10 houses within a 1-4 mile radius that have sold in the past 5 years. These houses "determine" the market value of your property.

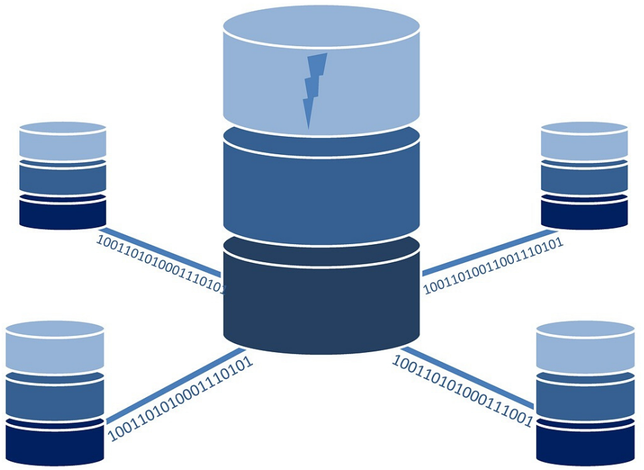


FEATURES & DATA SET

AREA OF FOCUS: GAINESVILLE, FL

FEATURES & DATA SET

DATA SOURCE: MLS



The local MLS is a shared listing service by all agents in an area.

It's a cross between a database and marketplace.

Data from the MLS is what gets pushed to Zillow, Realist, and other online listing platforms used by agents.

Very robust and (typically) accurate information about a property's details.

ANALYZING VALUATION

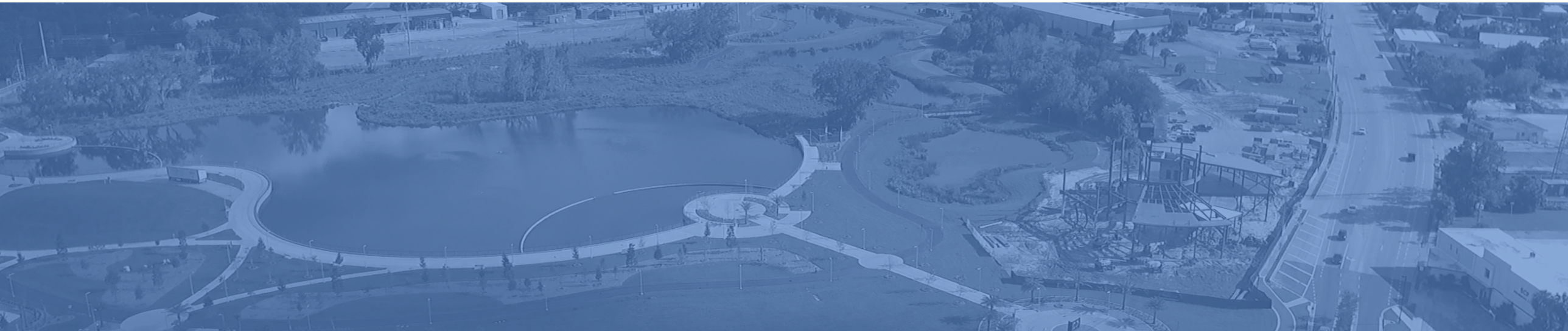
MACHINE LEARNING

TARGET VARIABLE

Sales Price

LEARNING MODEL:

Linear Regresson



EXPLORING THE DATA

GENERAL EDA

PROPERTIES THAT SKEW OUR DATA

Properties like bank owned, short sales, and auctions represent uniquely distressed properties that are going through foreclosure. These properties have their own share of issues and don't represent a regular transaction

EXPLORING THE DATA

GENERAL EDA

PROPERTIES THAT SKEW OUR DATA

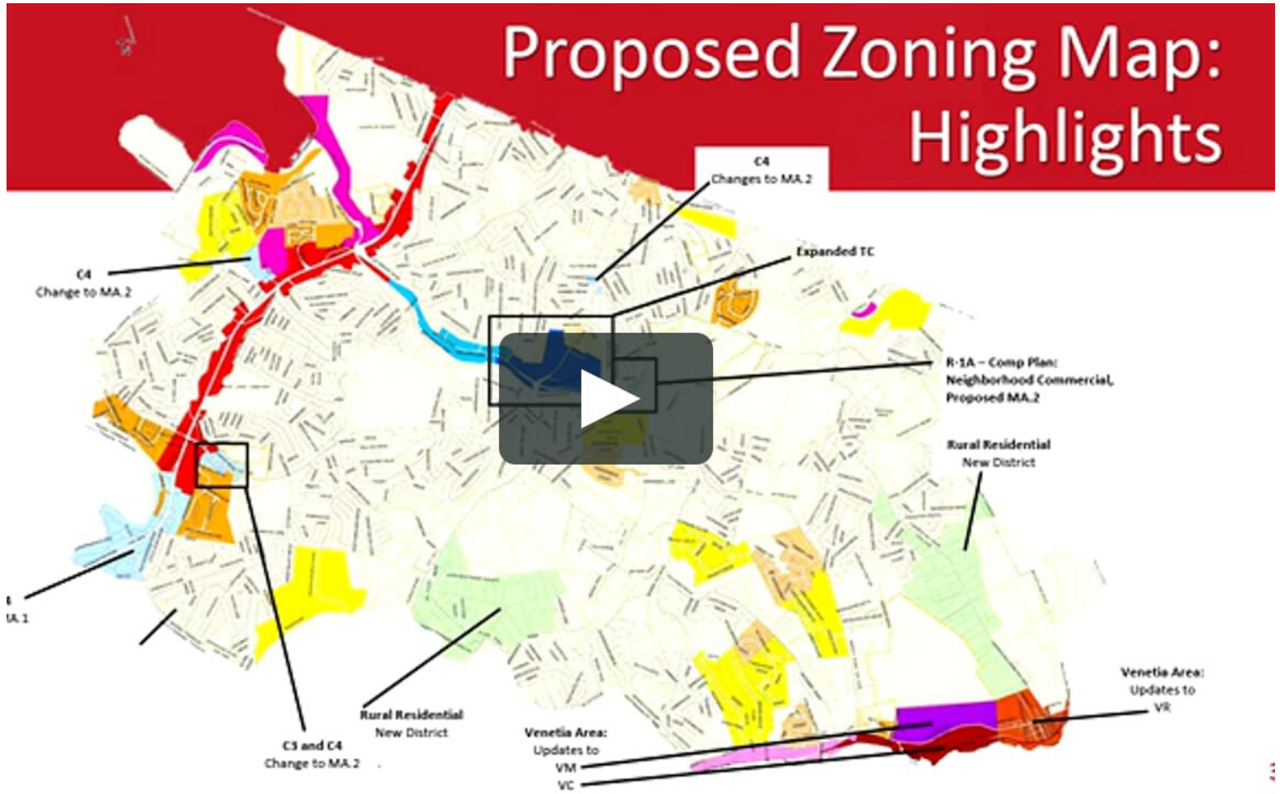
Waterfront properties also make up the minority as Gainesville is fairly land locked. They represent a small portion of the market and there is concern they could skew the weight of other metrics.

EXPLORING THE DATA

GENERAL EDA

PROPERTIES THAT SKEW OUR DATA

Zoning also impacts property value in a way that's inconsistent. Properties that are zoned for commercial usages are sold for redevelopment purposes.



EXPLORING THE DATA

GENERAL EDA

CLEANING USER INPUT ERRORS

Agents are human, and it's common for input errors to occur with properties. These are usually easy to manually impute (if there aren't many) by cross referencing the tax Id with the county database.

Some interesting input errors:

- **The house from nightmares:** 33 bedrooms and 2 baths
- **From the distant past and future:** year built 839 and 2340
- **Lots of Land:** 6000 acre cottage

EARLY INSIGHTS FROM THE DATA

WHAT OUTLIERS TELL US

Expensive Properties

Properties worth $1.2 million plus are mostly new construction or under construction.

The outliers of these outliers? custom homes built by famous architects.

Property Size

Most properties are on less than an acre of property.

The biggest? A two story contemporary house on 160 acres slightly outside of the general town space.

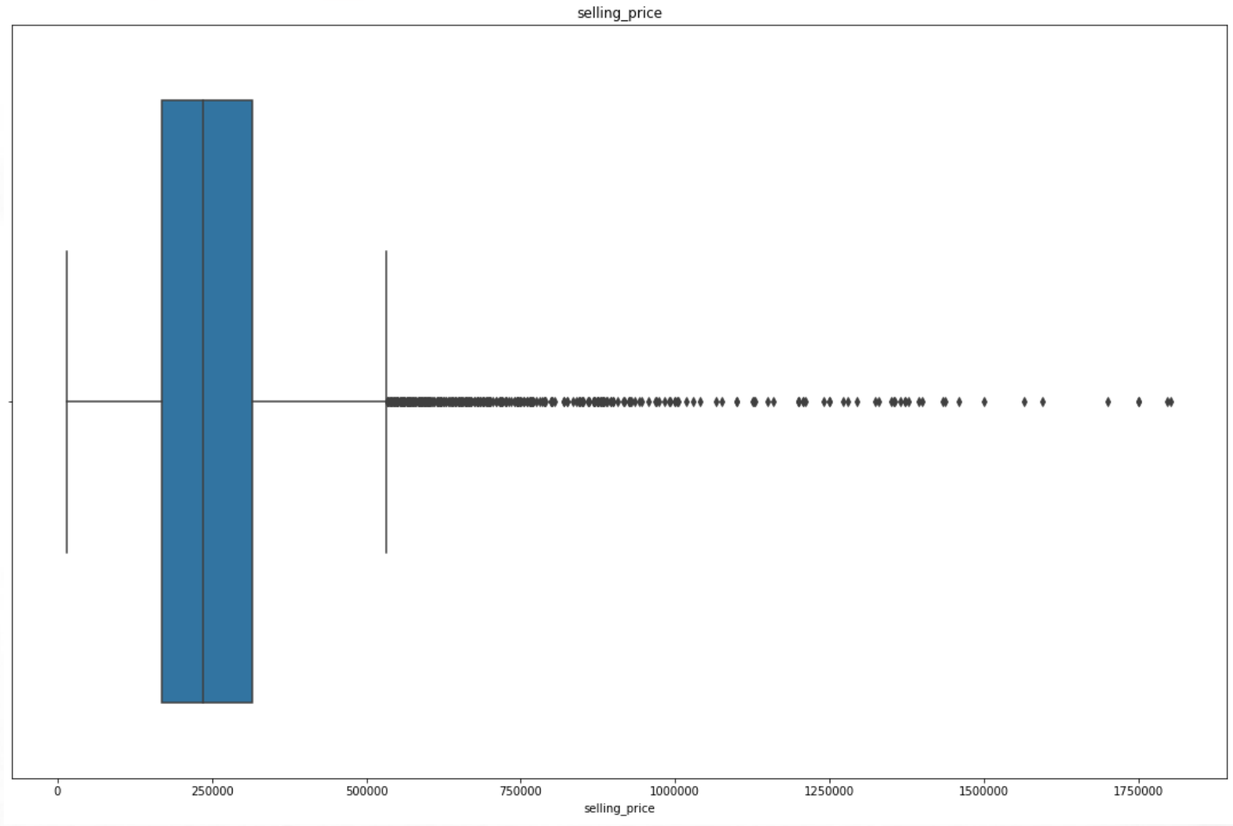
Square footage

Under 6000 square feet is the general threshold here.

Largest property is on 3 acres and is over 10,000+ square feet of air conditioned space.

EARLY INSIGHTS FROM THE DATA

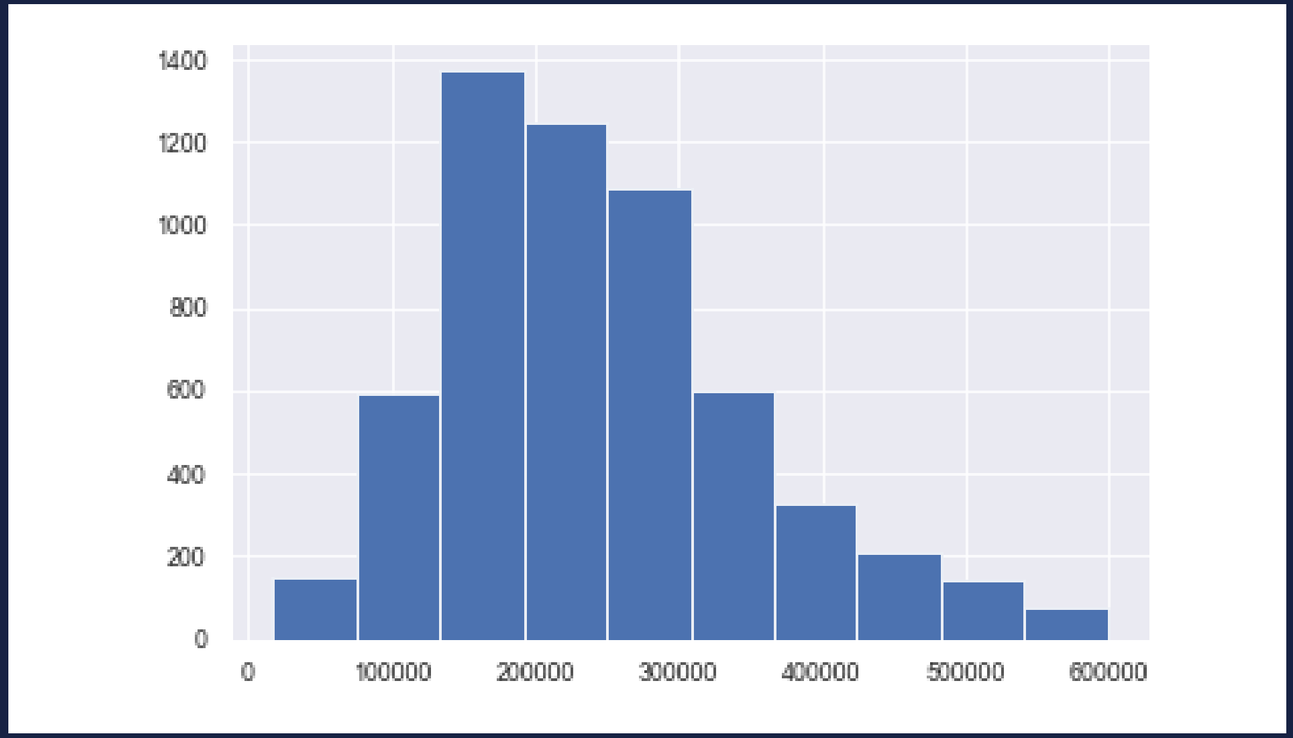
WHAT OUTLIERS TELL US



Linear Regression

Meeting the assumptons

Adjusting for Normality in Sales Price



Linear Regression

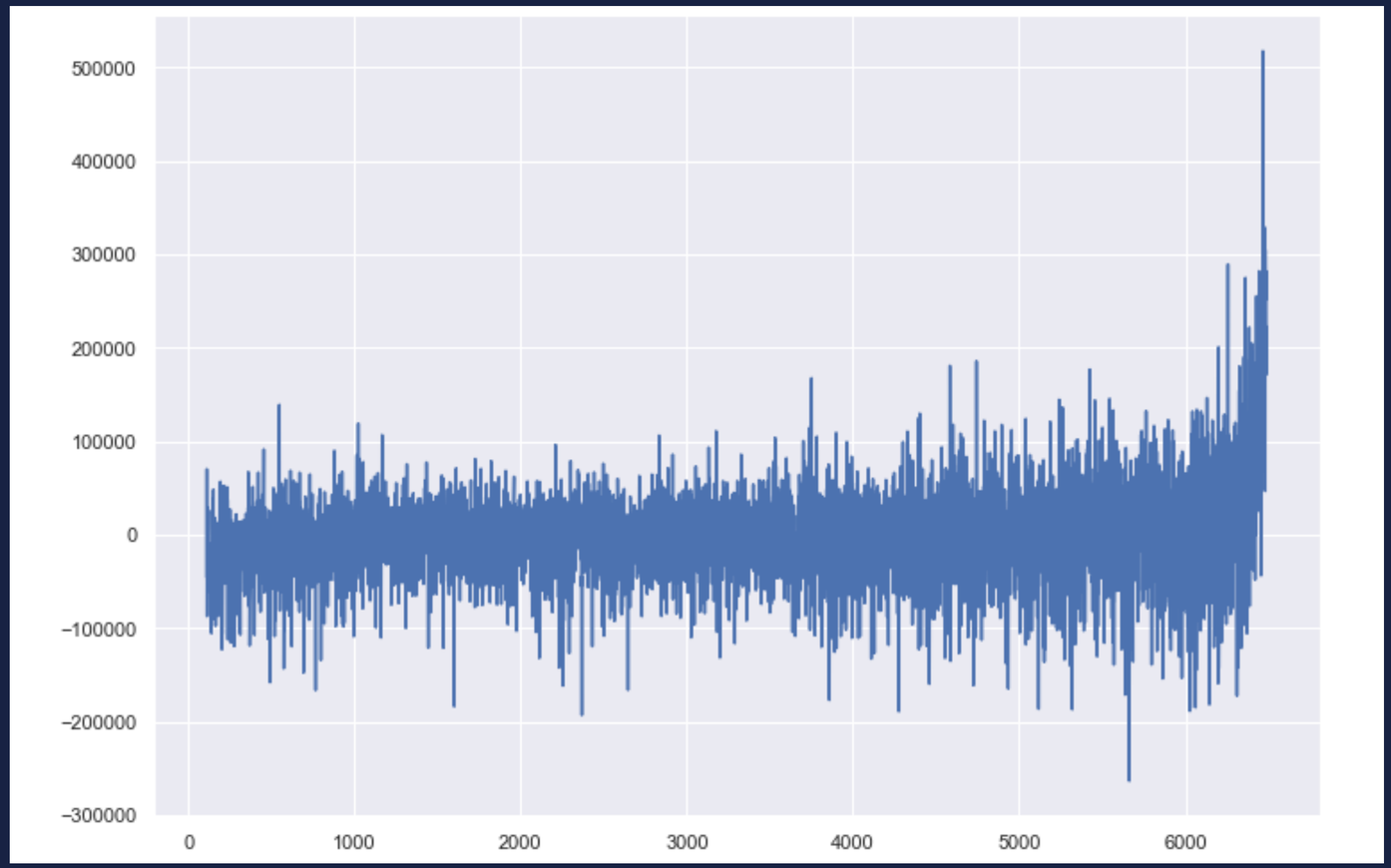
Meeting the assumptons

Homoscedasticity

Linear Regression

Meeting the assumptons

Errors not correlated



PREDICTIVE RESULTS

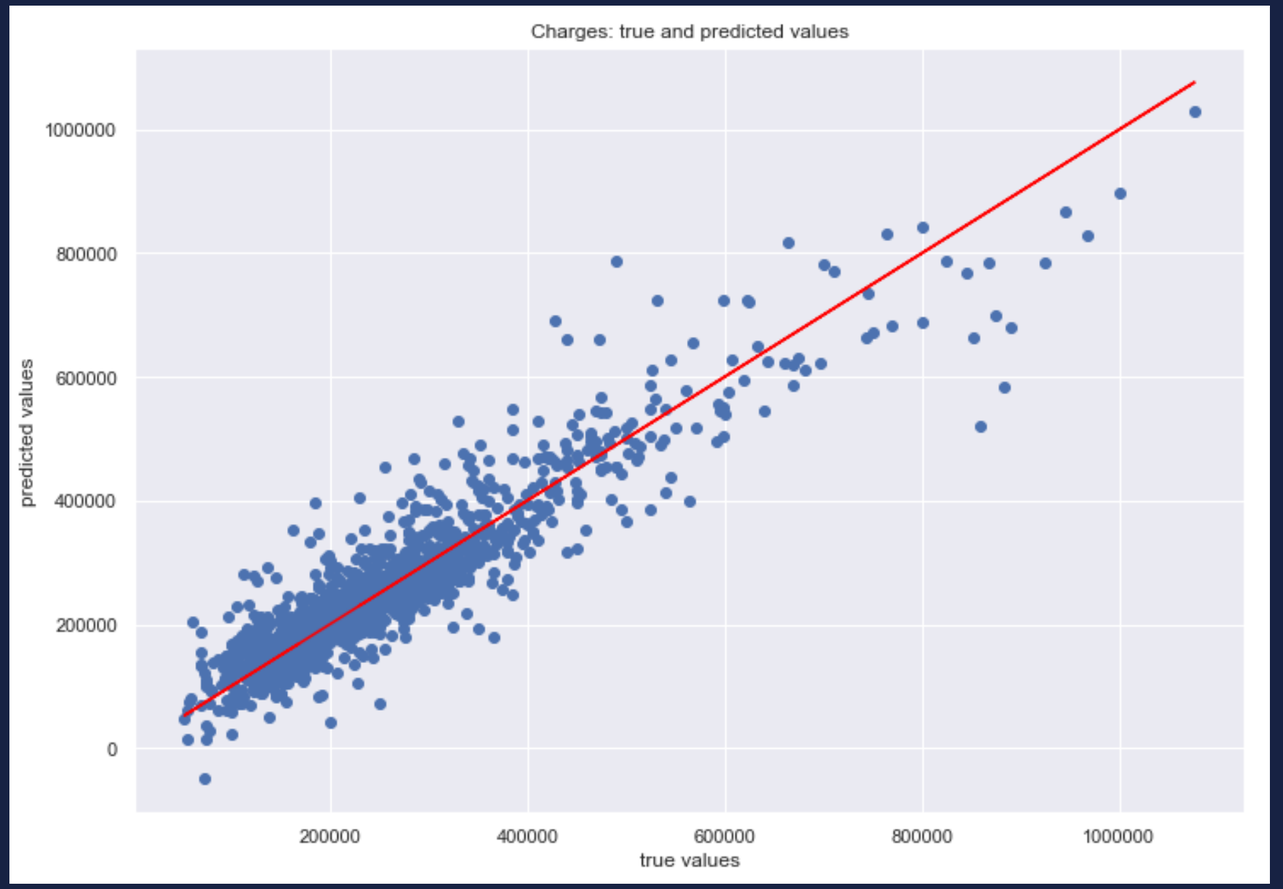
LINEAR REGRESSION

PREDICTIVE RESULTS

WHAT HAPPENED?

Lots of coefficients are in my model, so likely something is throwing off my test predictions.

Luckily there is an easy solution: Lasso Regression



TRAIN SCORE: 0.8971

TEST SCORE: -218778380266693.32

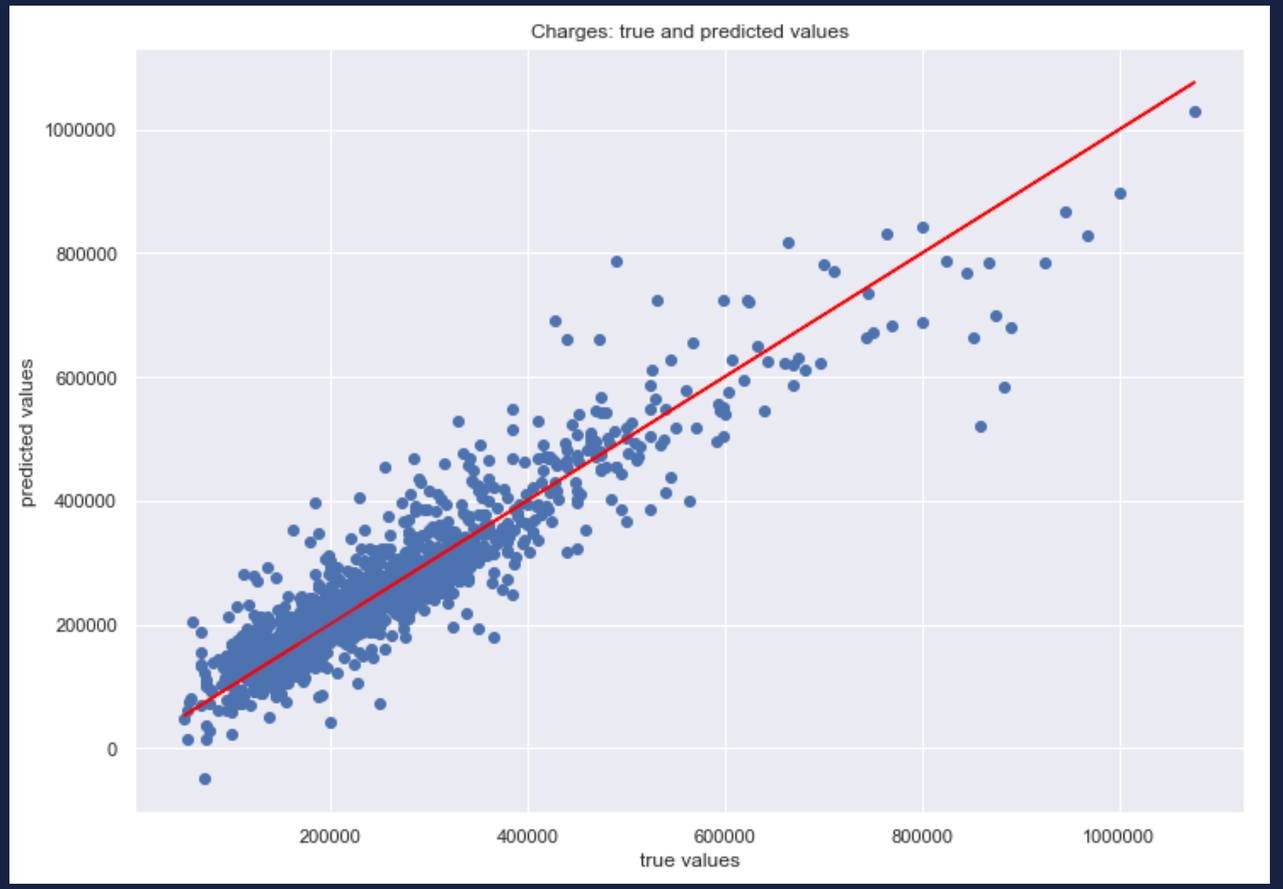
PREDICTIVE RESULETS

LASSO REGRESSION

Alpha = 100

Lasso helps "lasso" in the coefficients by scaling and forcing some coefficients to equal zero.

It helps select the best features that make for a simpler model.



LASSO TRAIN SCORE: 0.886

LASSO TEST SCORE: 0.859

WHICH FEATURES MATTER?

	Weight
Water Closet/Priv Toilet	69198.586248
Urinal	61592.196272
Oven - Double	54080.498995
Tall Countertops	45965.202229
Compactor	41213.661402
Board (Horse)	35692.481540
Aluminum	34646.590824
CB/Frame Front	33596.460093

The Top Five

Our top five are features that are fairly exclusive to luxury homes and make a lot of sense in terms of interpretability

Disclosure

These weights are unique to this iteration of modeling, and represent elements to houses in certain areas.

The most consistent coefficient for all property varieties is likely the Aluminum (roof).

NEXT STEPS TO EXPAND:

Revisiting CMA

Subjective Geofencing based on latitude and longitude would help provide better predictions due to regionality

Days on Market

This would be a great prediction to go along with house price for prospective sellers.

Revisiting some Variable Thresholds

Some variables could use more EDA than what the time for this project allowed. Especially ones that show non-normal distributions such as acreage.