# Chicago Housing Data

Collins Westnedge

# Motivation and Goals

Scrape usable data from the web

Clean, process and encode data

Train a model

Make housing prices predictions

Refine and repeat


# *What is a “Full Bathroom”?!*

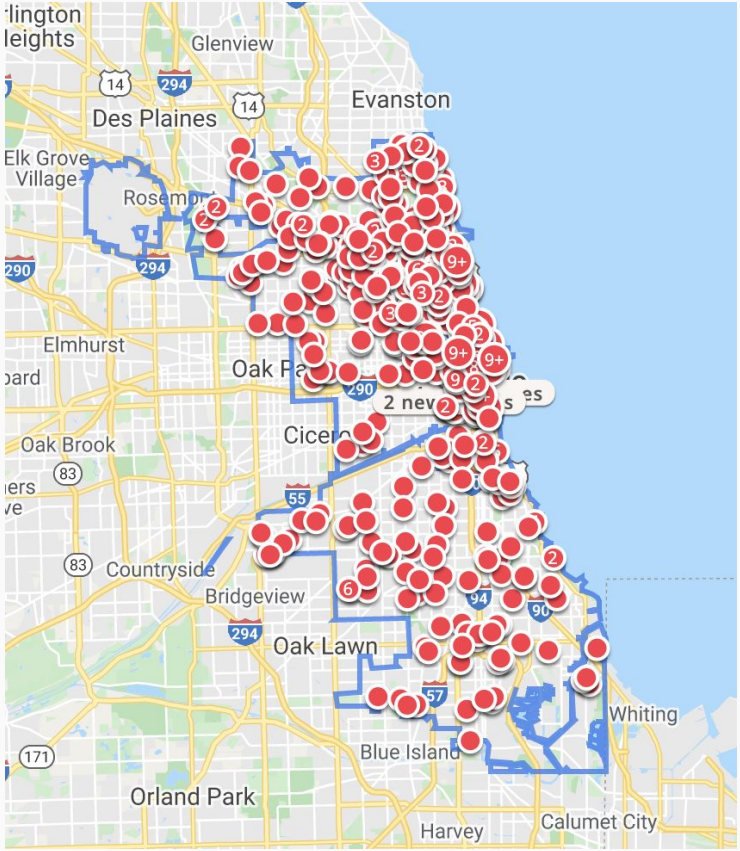
# Data Source

Zillow is an online real estate database that has data on approximately 110 million homes in the US.

Great for web scraping, but they don't show you the full picture so you have to be explicit with your searches.

"Showing 500 of 15000 results in this area" what about the other 14500 properties?!

And the CAPTCHA



# Collecting the Data

## Hurdles

- CAPTCHAS
- Speed
- Incomplete data
- Bad queries
- Mislabeled properties (Ohio, IL?)
- Flipped Properties

## Design Choices/Resolutions

- Request Headers and Random User Agents
- Beautiful Soup to deal with speed, who needs the javascript/selenium anyway
- Conditions to check if the property status matches your search query
- Letting Zillow do the work for you (grabbing all the text in each table)

# Cleaning the Data

This was not easy

[['Type', 'Single Family'], ['Year built', '1999'], ['Heating', 'Forced air'], ['Cooling', 'Central'], ['Parking', '2 spaces'], ['HOA', 'None'], ['Lot', '6,873 sqft'], ['Price/sqft', '$183'], ['Total Price', '$374,999'], ['Address', '5241 W 108th Pl, Oak Lawn, IL 60453'], ['Square Feet', '2,048'], ['New construction', ' No'], ['Bedrooms', ' 4'], ['Bathrooms', ' 4'], ['Full bathrooms', ' 3'], ['Half bathrooms', ' 1'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will rise 1.3% (↑) in the next year.This home is valued 78.1% higher (↑) than the median home in 60453.The median Zestimate® for this neighborhood is $210,544.'], ['URL', 'https://www.zillow.com/homedetails/5241-W-108th-Pl-Oak-Lawn-IL-60453/4088199_zpid/']]
https://www.zillow.com/homedetails/9613-Oak-Park-Ave-Oak-Lawn-IL-60453/99358941_zpid/
[['Type', 'Single Family'], ['Year built', '2008'], ['Heating', 'Gas'], ['Cooling', 'Central'], ['Parking', '2 spaces'], ['Lot', '6,664 sqft'], ['Price/sqft', '$163'], ['Total Price', '$479,000'], ['Address', '9613 Oak Park Ave, Oak Lawn, IL 60453'], ['Square Feet', '2,946'], ['New construction', ' No'], ['Bedrooms', ' 4'], ['Bathrooms', ' 4'], ['Full bathrooms', ' 4'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will rise 1.3% (↑) in the next year.This home is valued 127.5% higher (↑) than the median home in 60453.The median Zestimate® for this neighborhood is $210,544.'], ['URL', 'https://www.zillow.com/homedetails/9613-Oak-Park-Ave-Oak-Lawn-IL-60453/99358941_zpid/']]
https://www.zillow.com/homedetails/10324-S-Pulaski-Rd-APT-309-Oak-Lawn-IL-60453/4085396_zpid/
[['Type', 'Condo'], ['Year built', '1969'], ['Heating', 'Other'], ['Cooling', 'Refrigeration'], ['Parking', '1 space'], ['HOA', '$278/month'], ['Price/sqft', 'No Data'], ['Total Price', '$118,500'], ['Address', '10324 S Pulaski Rd APT 309, Oak Lawn, IL 60453'], ['sqft', '--'], ['New construction', ' No'], ['Bedrooms', ' 2'], ['Bathrooms', ' 2'], ['Full bathrooms', ' 2'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will rise 1.3% (↑) in the next year.This home is valued 43.7% lower (↓) than the median home in 60453.The median Zestimate® for this neighborhood is $210,544.'], ['URL', 'https://www.zillow.com/homedetails/10324-S-Pulaski-Rd-APT-309-Oak-Lawn-IL-60453/4085396_zpid/']]
https://www.zillow.com/homedetails/10002-S-Pulaski-Rd-APT-307-Oak-Lawn-IL-60453/4072503_zpid/
[['Type', 'Condo'], ['Year built', '1974'], ['Heating', 'Forced air'], ['Cooling', 'Central'], ['Parking', '1 space'], ['HOA', '$269/month'], ['Price/sqft', '$150'], ['Total Price', '$134,900'], ['Address', '10002 S Pulaski Rd APT 307, Oak Lawn, IL 60453'], ['Square Feet', '900'], ['New construction', ' No'], ['Bedrooms', ' 2'], ['Bathrooms', ' 2'], ['Full bathrooms', ' 1'], ['Half bathrooms', ' 1'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will rise 1.3% (↑) in the next year.This home is valued 35.9% lower (↓) than the median home in 60453.The median Zestimate® for this neighborhood is $210,544.'], ['URL', 'https://www.zillow.com/homedetails/10002-S-Pulaski-Rd-APT-307-Oak-Lawn-IL-60453/4072503_zpid/']]
https://www.zillow.com/homedetails/10313-Long-Ave-Oak-Lawn-IL-60453/4087100_zpid/
[['Type', 'Single Family'], ['Year built', '1962'], ['Heating', 'Baseboard'], ['Cooling', 'Central'], ['Parking', '2 spaces'], ['HOA', 'None'], ['Lot', '7,370 sqft'], ['Price/sqft', '$166'], ['Total Price', '$279,000'], ['Address', '10313 Long Ave, Oak Lawn, IL 60453'], ['Square Feet', '1,677'], ['New construction', ' No'], ['Bedrooms', ' 3'], ['Bathrooms', ' 2'], ['Full bathrooms', ' 1'], ['Half bathrooms', ' 1'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will rise 1.3% (↑) in the next year.This home is valued 32.5% higher (↑) than the median home in 60453.The median Zestimate® for this neighborhood is $210,544.'], ['URL', 'https://www.zillow.com/homedetails/10313-Long-Ave-Oak-Lawn-IL-60453/4087100_zpid/']]
https://www.zillow.com/homedetails/5140-W-90th-St-Oak-Lawn-IL-60453/4057923_zpid/
https://www.zillow.com/homedetails/11009-Jodan-Dr-Oak-Lawn-IL-60453/4088966_zpid/
[['Type', 'Townhouse'], ['Year built', '1994'], ['Heating', 'Forced air'], ['Cooling', 'Central'], ['Parking', '1 space'], ['HOA', '$200/month'], ['Lot', '1,421 sqft'], ['Price/sqft', '$126'], ['Total Price', '$207,000'], ['Address', '11009 Jodan Dr, Oak Lawn, IL 60453'], ['Square Feet', '1,641'], ['New construction', ' No'], ['Bedrooms', ' 2'], ['Bathrooms', ' 3'], ['Full bathrooms', ' 2'], ['Half bathrooms', ' 1'], ['neighborhood_stats', 'Home values in 60453 have risen 1.3\u200a% (↑) over the past 12 months.Zillow predicts the home values in 60453 will increase 1.3% (↑) in th

# Cleaning the Data

Sometimes Lot size is in square feet sometimes its in acres.

The heating column is basically a mash up of a word cloud and a relators blog.

Neighborhoods stats contains a lot of useful information but is basically one giant string.

Same goes for the url thankfully this is one of the most reliable and useful data points since it contains address zip and is unique to each property.

| | Type | Year built | Heating | Cooling | HOA | Total Price | Square Feet | New construction | Bedrooms | Bathrooms | Full bathrooms | neighborhood_stats | Half bathrooms | zip |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Condo | 1910.0 | Gas | Central | No | 249900.0 | 1000.0 | No | 2.0 | 2.0 | 2.0 | Home values in Hyde Park have risen 0.6% (↑) ... | 0 | 60615 |
| **1** | Single Family | 1957.0 | Forced air | Central | Yes | 399000.0 | 1224.0 | No | 2.0 | 2.0 | 2.0 | Home values in Edison Park have risen 0.6% (↑... | 0 | 60631 |
| **2** | Condo | 2007.0 | Forced air | Central | No | 169898.0 | 775.0 | No | 2.0 | 1.0 | 1.0 | Home values in Rogers Park have risen 0.6% (↑... | 0 | 60626 |
| **3** | Condo | 2008.0 | Gas | Central | No | 489900.0 | 1273.0 | No | 2.0 | 2.0 | 2.0 | Home values in South Loop have risen 0.6% (↑)... | 0 | 60605 |
| **4** | Single Family | 1971.0 | Forced air | Central | Yes | 389000.0 | 2179.0 | No | 3.0 | 3.0 | 2.0 | Home values in Beverly have risen 0.6% (↑) ov... | 1 | 60643 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2595** | Condo | 2005.0 | Forced air | Central | No | 168999.0 | 1052.0 | No | 1.0 | 1.0 | 1.0 | Home values in 60453 have risen 1.3 % (↑) over... | 0 | 60453 |
| **2596** | Single Family | 1973.0 | Gas | Central | Yes | 399900.0 | 3400.0 | No | 4.0 | 3.0 | 3.0 | Home values in 60453 have risen 1.3 % (↑) over... | 0 | 60453 |
| **2597** | Condo | 1973.0 | Other | Refrigeration | No | 94500.0 | 852.0 | No | 2.0 | 1.0 | 1.0 | Home values in 60453 have risen 1.3 % (↑) over... | 0 | 60453 |
| **2598** | Single Family | 1949.0 | Forced air, Gas | Central | Yes | 249900.0 | 1343.0 | No | 3.0 | 2.0 | 1.0 | Home values in 60453 have risen 1.3 % (↑) over... | 1 | 60453 |
| **2599** | Condo | 1983.0 | Forced air | Central | No | 124900.0 | 1000.0 | No | 2.0 | 2.0 | 2.0 | Home values in 60453 have risen 1.3 % (↑) over... | 0 | 60453 |

2497 rows × 18 columns

# Creating a Model

Key Features

- Square Footage
- Bathrooms
- Bedrooms
- Median Neighborhood Price
- Single Family Home

Other Features

- Latitude longitude
- Zip code
- HOA
- Proximity to Public schools
- Year built

# The Process

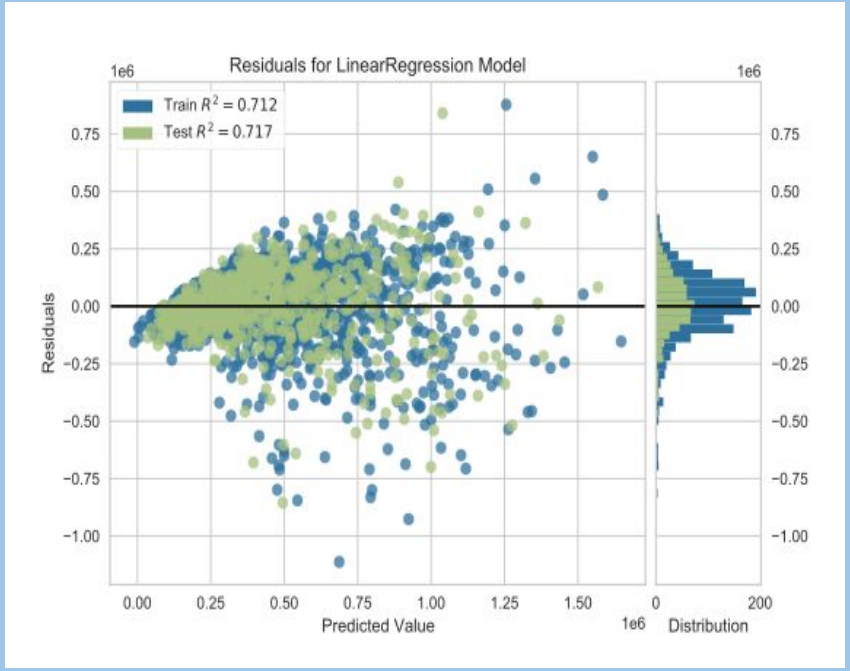
- Start with a baseline model only including continuous features.
- Keep it simple and interpretable… for now.
- Gradually increase the number of features see how that affects the p-values and R^2
- See how the model generalizes to the test set.
- Graph the Residuals look at their distribution.

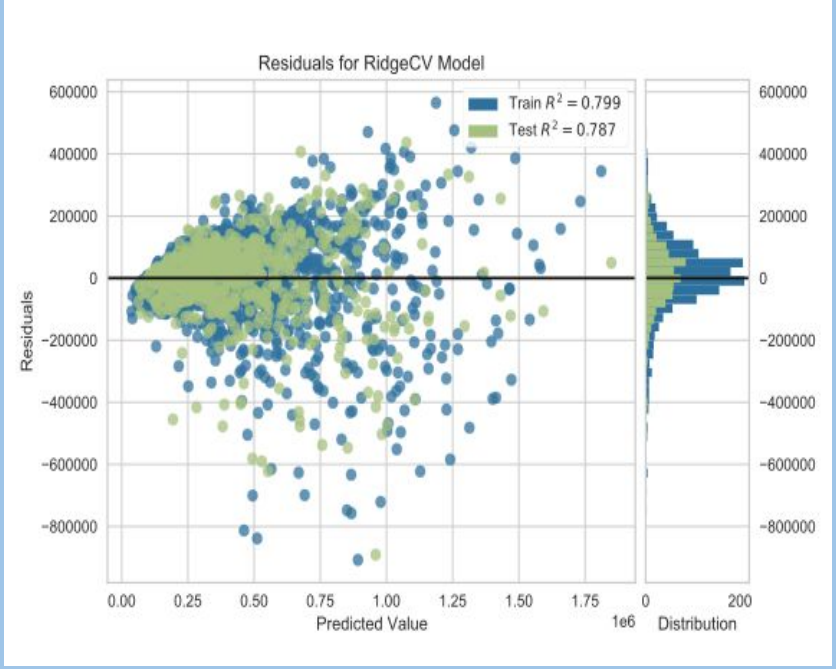
OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Total Price | **R-squared:** | 0.739 |
| **Model:** | OLS | **Adj. R-squared:** | 0.738 |
| **Method:** | Least Squares | **F-statistic:** | 866.8 |
| **Date:** | Fri, 09 Oct 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 00:14:33 | **Log-Likelihood:** | -28797. |
| **No. Observations:** | 2149 | **AIC:** | 5.761e+04 |
| **Df Residuals:** | 2141 | **BIC:** | 5.765e+04 |
| **Df Model:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -4.951e+05 | 2.07e+05 | -2.393 | 0.017 | -9.01e+05 | -8.93e+04 |
| **Year built** | 198.6060 | 105.200 | 1.888 | 0.059 | -7.699 | 404.911 |
| **Square Feet** | 92.8956 | 4.667 | 19.907 | 0.000 | 83.744 | 102.047 |
| **Bedrooms** | -1.711e+04 | 3953.266 | -4.328 | 0.000 | -2.49e+04 | -9358.256 |
| **Bathrooms** | 5.652e+04 | 4128.929 | 13.689 | 0.000 | 4.84e+04 | 6.46e+04 |
| **Full bathrooms** | 4.068e+04 | 2834.505 | 14.351 | 0.000 | 3.51e+04 | 4.62e+04 |
| **Half bathrooms** | -2.78e+04 | 8909.926 | -3.120 | 0.002 | -4.53e+04 | -1.03e+04 |
| **median_nhv** | 0.4814 | 0.018 | 26.300 | 0.000 | 0.445 | 0.517 |
| **Bathrooms_adj** | 1.584e+04 | 4239.629 | 3.737 | 0.000 | 7528.476 | 2.42e+04 |

# Model Performance



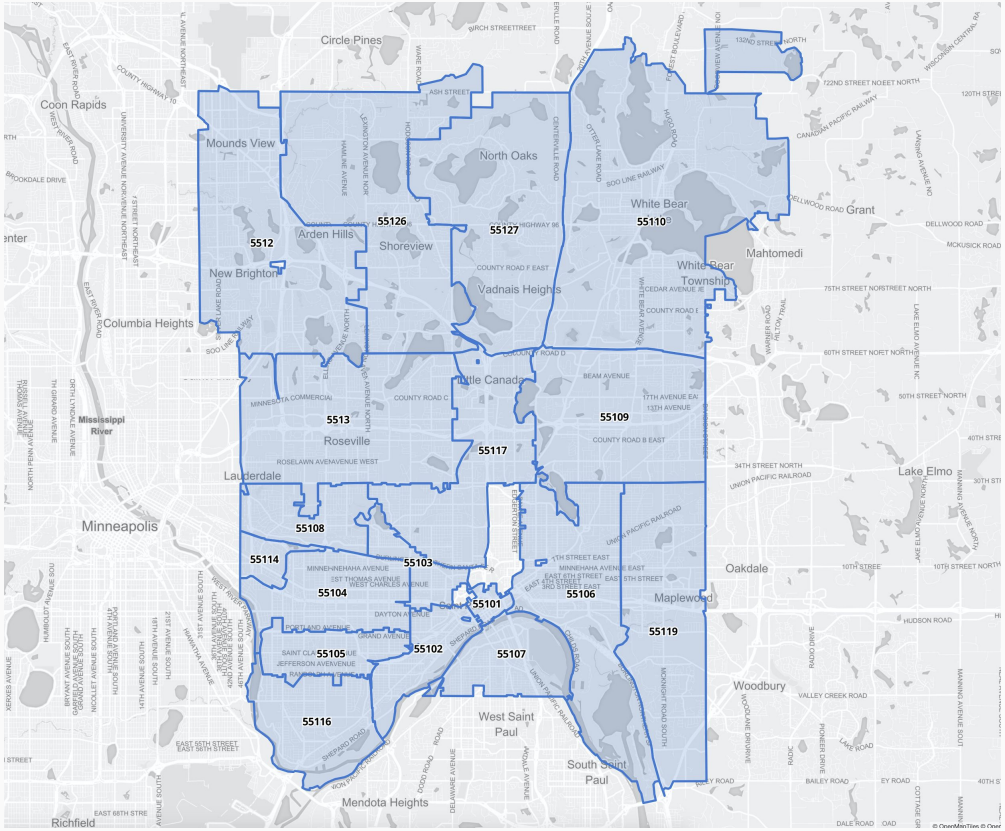



# Next steps

There's a lot of features like gps coordinates that we aren't using which could be incredibly useful or better handled by different algorithms

Try different algorithms with different features.

Incorporate new data especially since have all this location data we are not using.



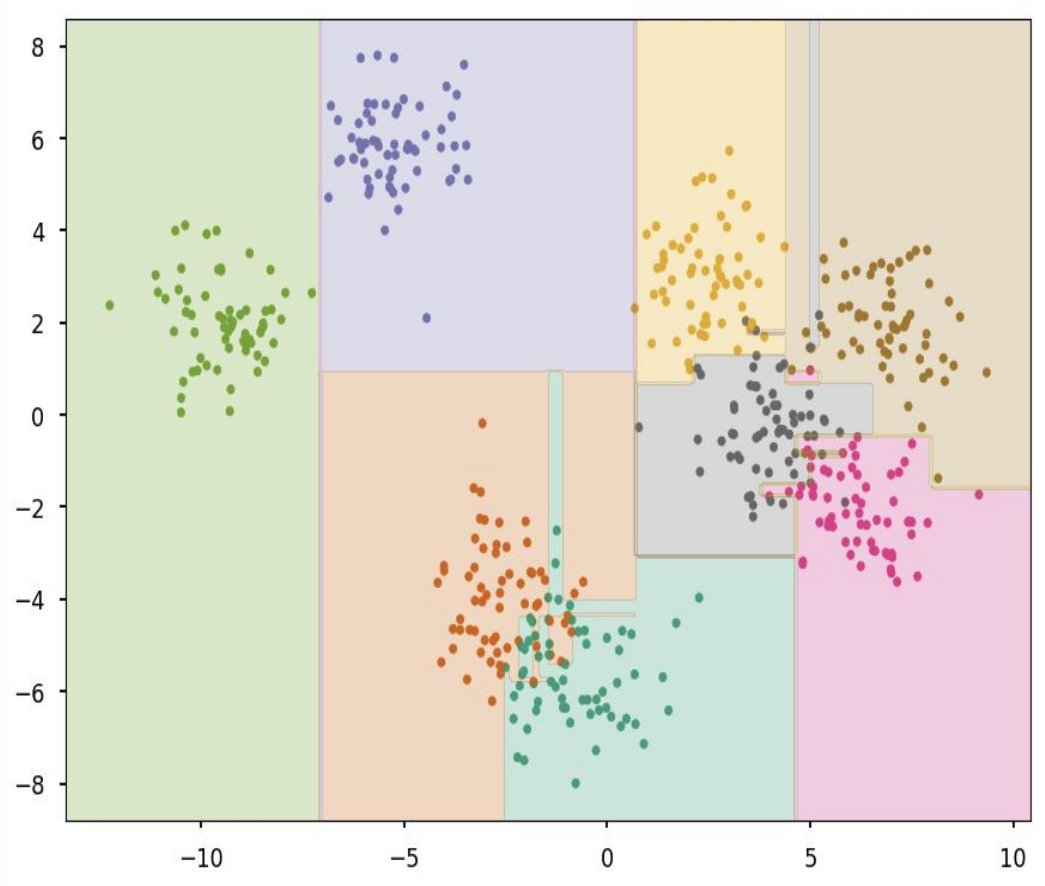
# Next steps

There's a lot of features like gps coordinates that we aren't using which could be incredibly useful or better handled by different algorithms

Try different algorithms with different features.

Incorporate new data especially since have all this location data we are not using.

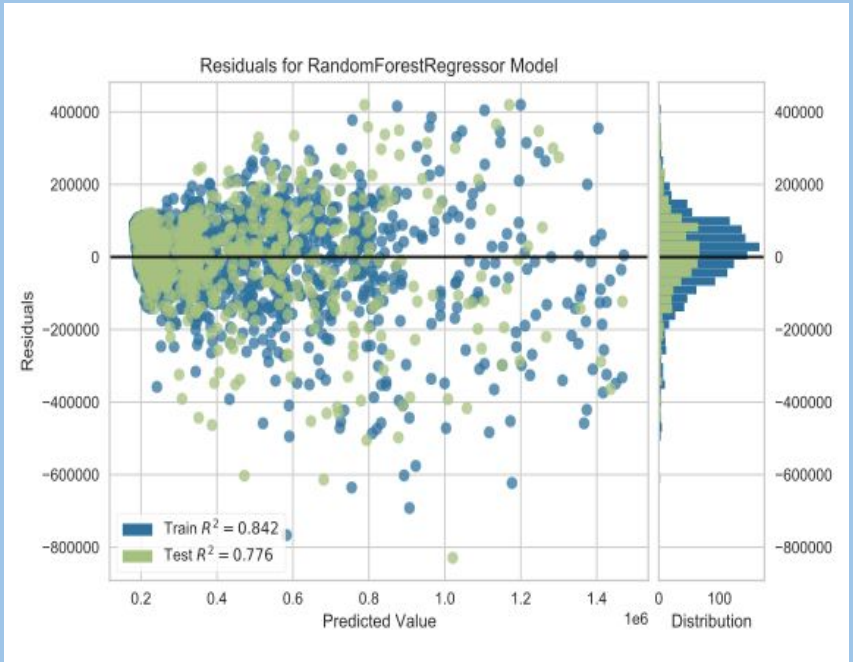
Random Forest Regressors could handle geo spatial data really well for our purposes.



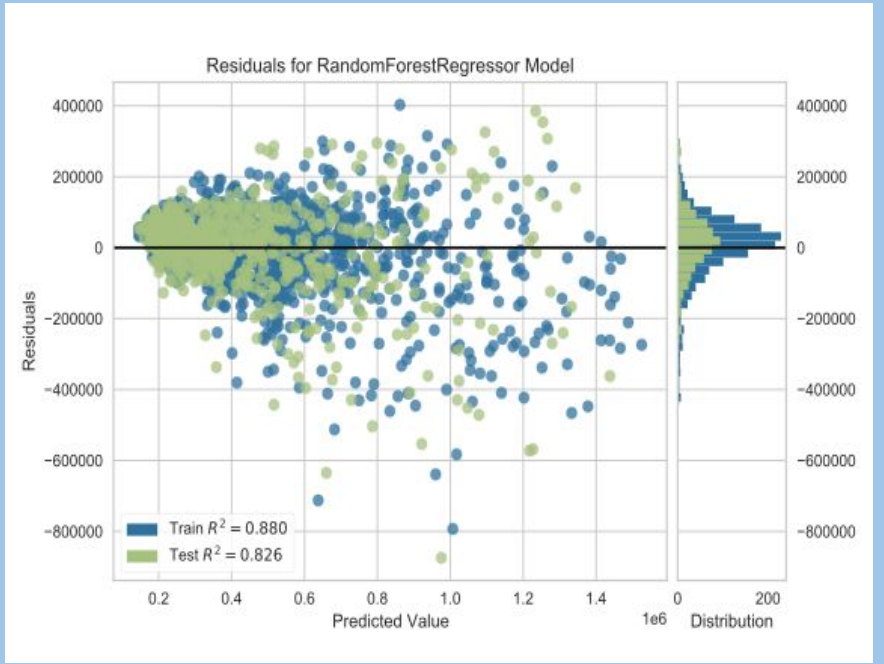
# Random Forest Performance

(starting to overfit)

## Default hyperparameters




## Optimized hyperparameters



# Takeaways

Training a model is an iterative process and there’s always room for improvement.

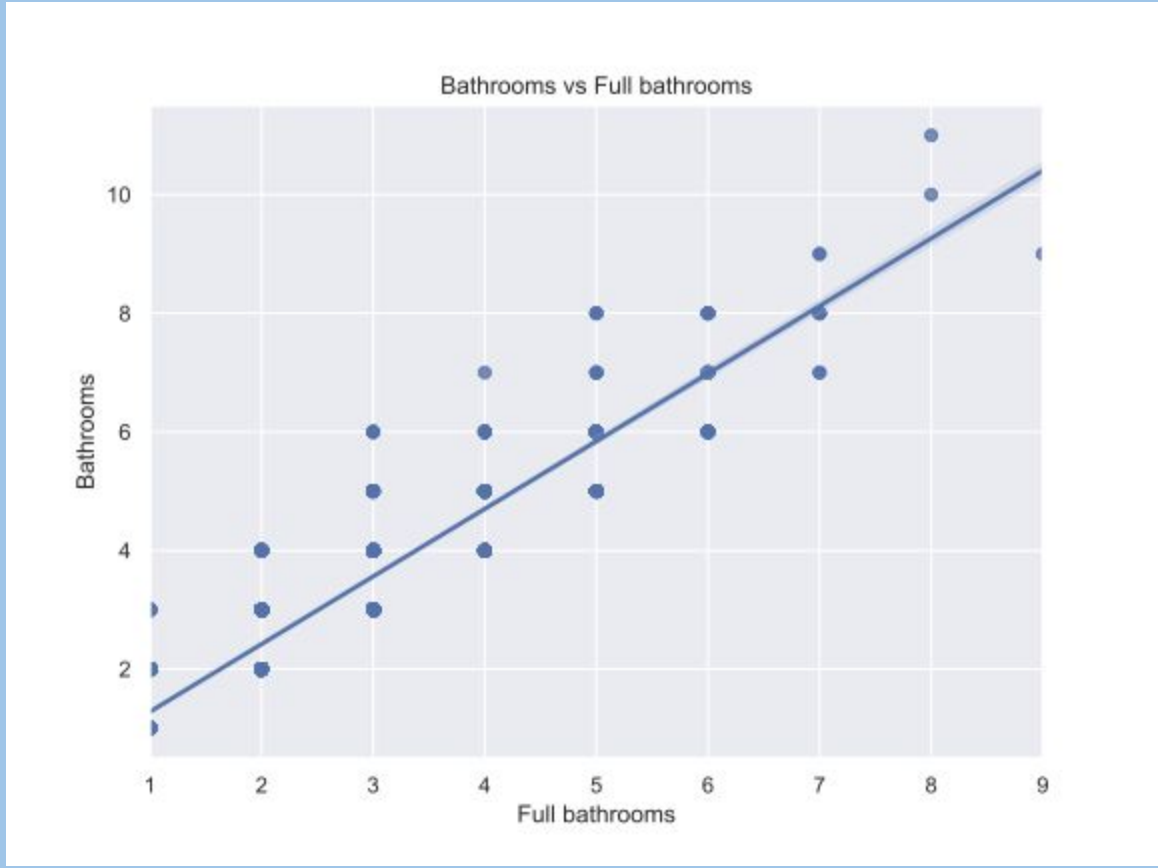
The relationship between price and our features is not always linear.

Bias-variance tradeoff is real.

Clean data goes a long way but is hard to collect.

# A full bathroom is basically a bathroom

- From an expert


Bathrooms vs Full bathrooms
Bathrooms
Full bathrooms
10
8
6
4
2
1
2
3
4
5
6
7
8
9

# Thanks!