# Predicting Connecticut Housing Prices: 2020-2021


Springboard Data Science Career Track Capstone

# Problem

## Audience

- Mortgage Lenders
- Investors and Developers
- Market Brokerages
- Lead Generating Services

## Context

- Many businesses depend on accurately pricing the Real Estate Market.
- Real Estate Assets are complex with tangible and intangible characteristics.
- Buyer's preferences can be a formed on any number of these characteristics. Therefore pricing can be tied to these characteristics

## Problem Statement

Can the key factors which determine housing prices be identified to accurately model and predict Real estate closing prices?

# Data Wrangling

—

- Raw data from CT MLS
- Sold or ‘Closed’ listings
- 7/28/2020 - 7/29/2021
- Every available city in Connecticut
- Limited to single family homes, condos, co-operatives.
- Total of 58,107 single family dwellings.

# Features

- Property Type
  - 46,390 SF
  - 11,468 CO
  - 249 CP
- Listing/Closing Price
- City
- Acreage
- Total Square Feet
- Square Footage Heated above Grade
- Style
- Total Rooms
- Total Bedrooms
- Bathrooms
- Garage/Parking
- Year Built
- Days on the Market

# Features requiring splitting

Features that were split into 2 new features

- Listing/Closing Price
  - LP: $3,750,000 CP: $3,750,000
- Bathrooms
  - 2 Full & 1 Half
- Garage/Parking
  - 2 Car/Off Street Parking
  - 2 Car/Attached Garage

# Features requiring Numeric Conversion

Features that required stripping of non-numeric characters and conversion to numeric type

- Listing Price
- Closing Price
- Acreage
- Total Square Feet
- Square Footage Heated above Grade
- Full Bath
- Half Bath
- Garage

# Misc. Feature Work

- Style
  - [“Colonial”, “Contemporary”]
- Age
  - Created from year built
- County
  - Data merged with table of CT Cities and counties.  Resulted in dropping of 162 invalid listings.

# Outlier and Missing Value Removal

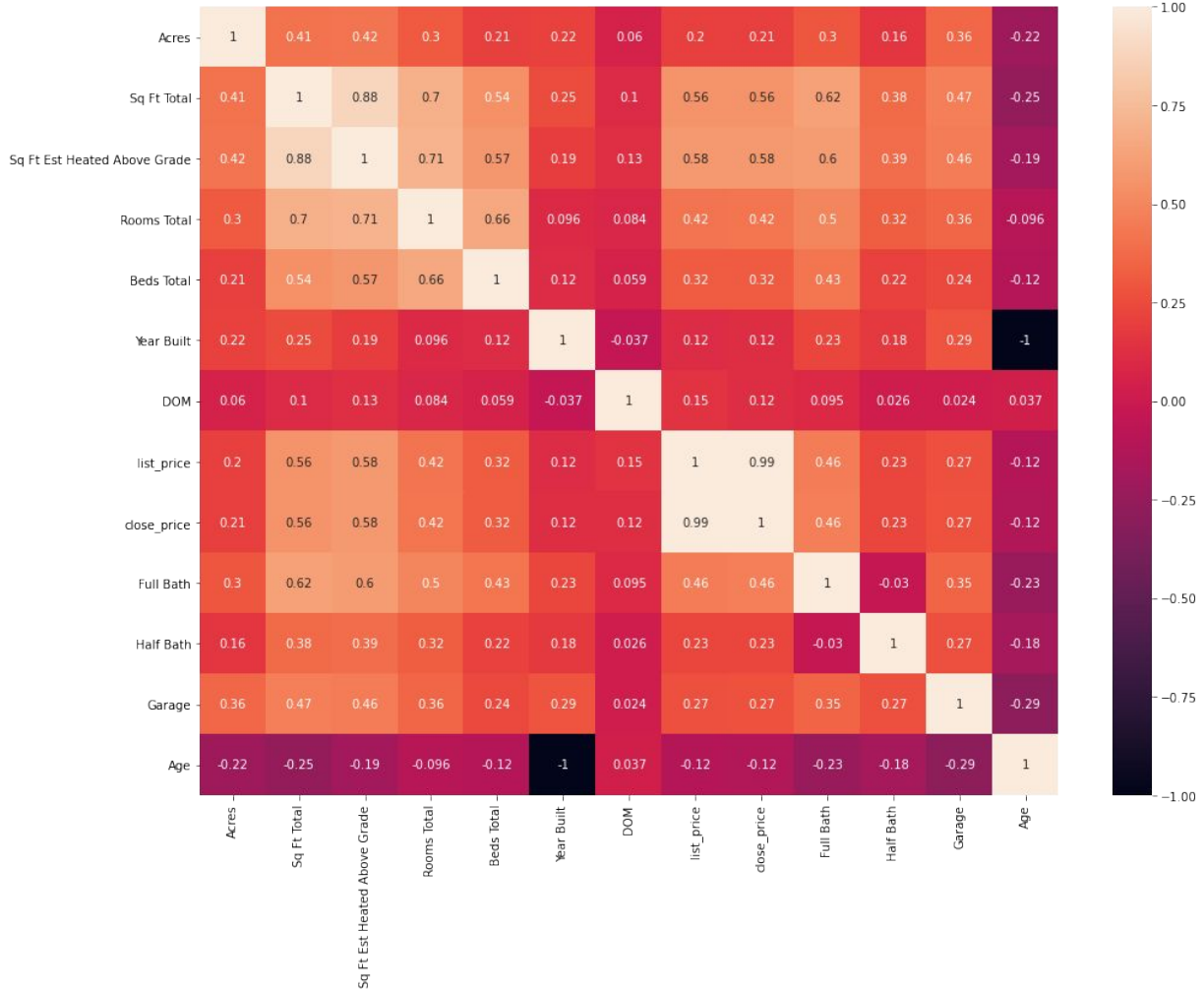
- Observations dropped by feature outliers
  - 1.5 * IQR
  - Acreage
  - Total Square Footage
  - Square Footage Heated above Grade
  - Total Rooms
  - Total Bedrooms
  - Full Baths
  - Half Baths.
- 9,871 total observations removed due to errors, missing values or outliers

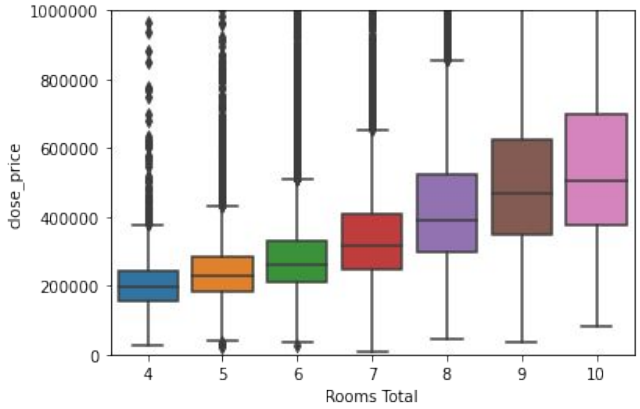
# Exploratory Data Analysis

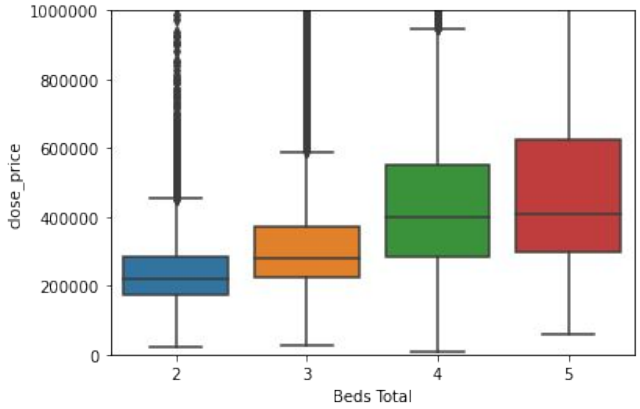
- **The final data set consisted of 36,519 single family houses**
  - 11,468 condominium and cooperative listings were dropped.
  - 9,871 observations dropped due to errors, missing values and outliers.
- **Average listing**
  - 1900 square foot
  - A colonial, ranch, or cape cod
  - 3 bedroom
  - 2 bathroom house
  - 1 car garage
  - .61 acre of land
  - Built in 1957.
  - Listed for $363,000
  - Sold for $365,000
  - Most likely in Hartford, New Haven, or Fairfield county.
  - Had a median DOM of 24 days and mean of 43 day

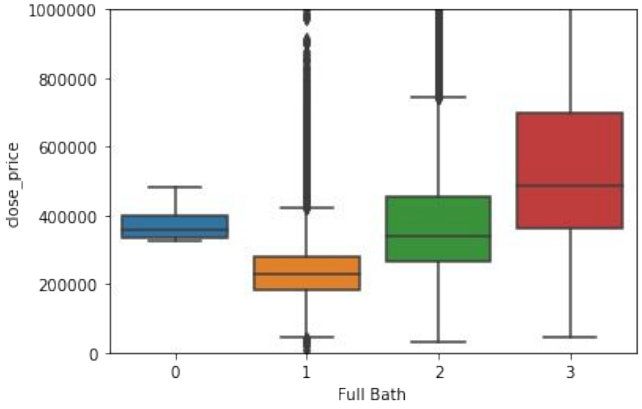
# Correlations

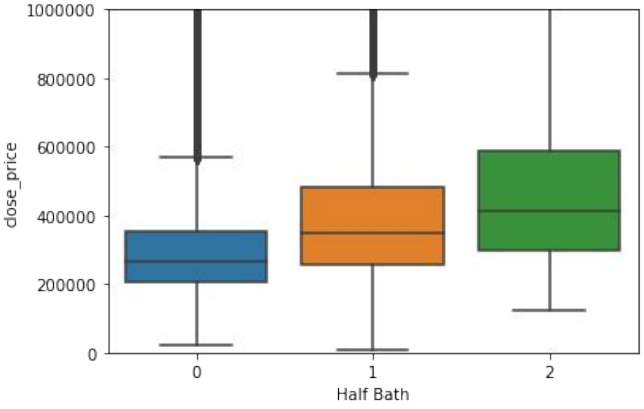
- Between features
  - Rooms, Bedrooms, Bathrooms with Square Footage.
- Between the target (closing price)
  - Square footage
  - Number of rooms
  - Number of bathrooms
  - County plays a big part in determining price

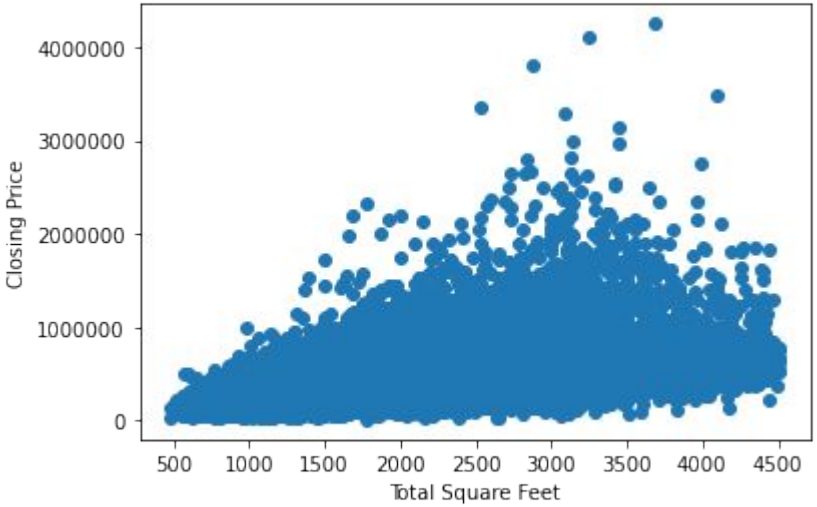

Acres 1 0.41 0.42 0.3 0.21 0.22 0.06 0.2 0.21 0.3 0.16 0.36 -0.22
Sq Ft Total 0.41 1 0.88 0.7 0.54 0.25 0.1 0.56 0.56 0.62 0.38 0.47 -0.25
Sq Ft Est Heated Above Grade 0.42 0.88 1 0.71 0.57 0.19 0.13 0.58 0.58 0.6 0.39 0.46 -0.19
Rooms Total 0.3 0.7 0.71 1 0.66 0.096 0.084 0.42 0.42 0.5 0.32 0.36 -0.096
Beds Total 0.21 0.54 0.57 0.66 1 0.12 0.059 0.32 0.32 0.43 0.22 0.24 -0.12
Year Built 0.22 0.25 0.19 0.096 0.12 1 -0.037 0.12 0.12 0.23 0.18 0.29 -1
DOM 0.06 0.1 0.13 0.084 0.059 -0.037 1 0.15 0.12 0.095 0.026 0.024 0.037
list_price 0.2 0.56 0.58 0.42 0.32 0.12 0.15 1 0.99 0.46 0.23 0.27 -0.12
close_price 0.21 0.56 0.58 0.42 0.32 0.12 0.12 0.99 1 0.46 0.23 0.27 -0.12
Full Bath 0.3 0.62 0.6 0.5 0.43 0.23 0.095 0.46 0.46 1 -0.03 0.35 -0.23
Half Bath 0.16 0.38 0.39 0.32 0.22 0.18 0.026 0.23 0.23 -0.03 1 0.27 -0.18
Garage 0.36 0.47 0.46 0.36 0.24 0.29 0.024 0.27 0.27 0.35 0.27 1 -0.29
Age -0.22 -0.25 -0.19 -0.096 -0.12 -1 0.037 -0.12 -0.12 -0.23 -0.18 -0.29 1
Acres
Sq Ft Total
Sq Ft Est Heated Above Grade
Rooms Total
Beds Total
Year Built
DOM
list_price
close_price
Full Bath
Half Bath
Garage
Age
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

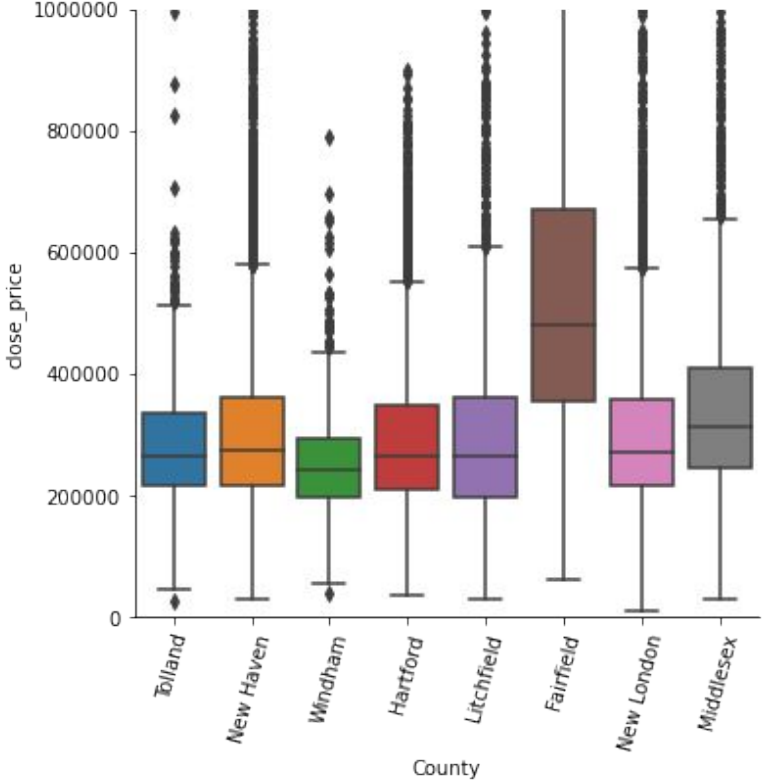

close_price
1000000
800000
600000
400000
200000
0
4
5
6
7
8
9
10
Rooms Total


close_price
1000000
800000
600000
400000
200000
0
2
3
4
5
Beds Total


close_price
1000000
800000
600000
400000
200000
0
0
1
2
3
Full Bath


close_price
1000000
800000
600000
400000
200000
0
0
1
2
Half Bath


Closing Price
4000000
3000000
2000000
1000000
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Total Square Feet


close_price
1000000
800000
600000
400000
200000
0
Tolland
New Haven
Windham
Hartford
Litchfield
Fairfield
New London
Middlesex
County

# Preprocessing

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   City                          36519 non-null  object
 1   Acres                         36519 non-null  float64
 2   Sq Ft Total                   36519 non-null  int64
 3   Sq Ft Est Heated Above Grade  36519 non-null  float64
 4   Rooms Total                   36519 non-null  int64
 5   Beds Total                    36519 non-null  int64
 6   DOM                           36519 non-null  float64
 7   close_price                   36519 non-null  int64
 8   Full Bath                     36519 non-null  int64
 9   Half Bath                     36519 non-null  int64
 10  Garage                        36519 non-null  int64
 11  Age                           36519 non-null  float64
 12  County                        36519 non-null  object
dtypes: float64(4), int64(7), object(2)
```

- Standard Scaling
  - All numerical features scaled to mean 0 and STD 1
- One Hot Encoding
  - Results in 175 additional features
- Test Train Split
  - 25% Testing set
  - 27,389 observation training set
  - 9,130 observation test set

# Modeling

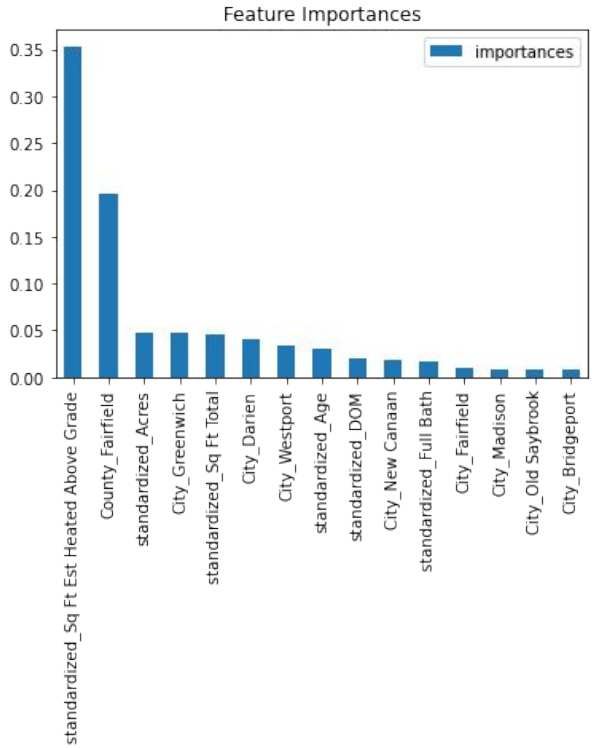
# Preliminary Modeling

| Model | MAE | MSE | RMSE | R-Squared |
|---|---|---|---|---|
| LinearRegression() | 71,527.51 | 17012550093 | 130,432.17 | 0.717 |
| ElasticNet(random_state=343) | 101,605.62 | 34222246761 | 184,992.56 | 0.430 |
| ElasticNet(l1_ratio=1.0, random_state=343, tol=0.05) | 71,456.87 | 17001968119 | 130,391.60 | 0.717 |
| Lasso(random_state=343, tol=0.05) | 71,456.87 | 17001968119 | 130,391.60 | 0.717 |
| Lasso(alpha=10.0, random_state=343, tol=0.05) | 71,312.37 | 16991498443 | 130,351.44 | 0.717 |
| Ridge(random_state=343, tol=0.05) | 71,456.38 | 17001575748 | 130,390.09 | 0.717 |
| Ridge(alpha=2.0, random_state=343, tol=0.05) | 71,443.16 | 17001458476 | 130,389.64 | 0.717 |
| RandomForestRegressor(n_jobs=-1, random_state=343) | 63,952.22 | 15026955463 | 122,584.48 | 0.750 |
| GradientBoostingRegressor(random_state=343) | 70,963.11 | 16706409287 | 129,253.28 | 0.722 |
| KNeighborsRegressor() | 75,326.12 | 21235487602 | 145,724.01 | 0.647 |

# Hyperparameter Tuned Final Model

| Model | MAE | MSE | RMSE | R-Squared |
|---|---|---|---|---|
| RandomForestRegressor(n_jobs=-1, random_state=343) | 63,952.22 | 15026955463 | 122,584.48 | 0.750 |
| RandomForestRegressor(max_depth=30, min_samples_split=20, n_estimators=150, n_jobs=-1, random_state=343) | 63,744.66 | 14902922330 | 122,077.53 | 0.752 |

# Feature Importances



# Changes for Improvement

- Explore additional internal features
  - Parking other than garage
  - SQFT / Room
- Explore additional external features
  - Mortgage Rates
  - Season
- Limit lose of observations
  - Removing less “outliers”
  - Imputing more null values