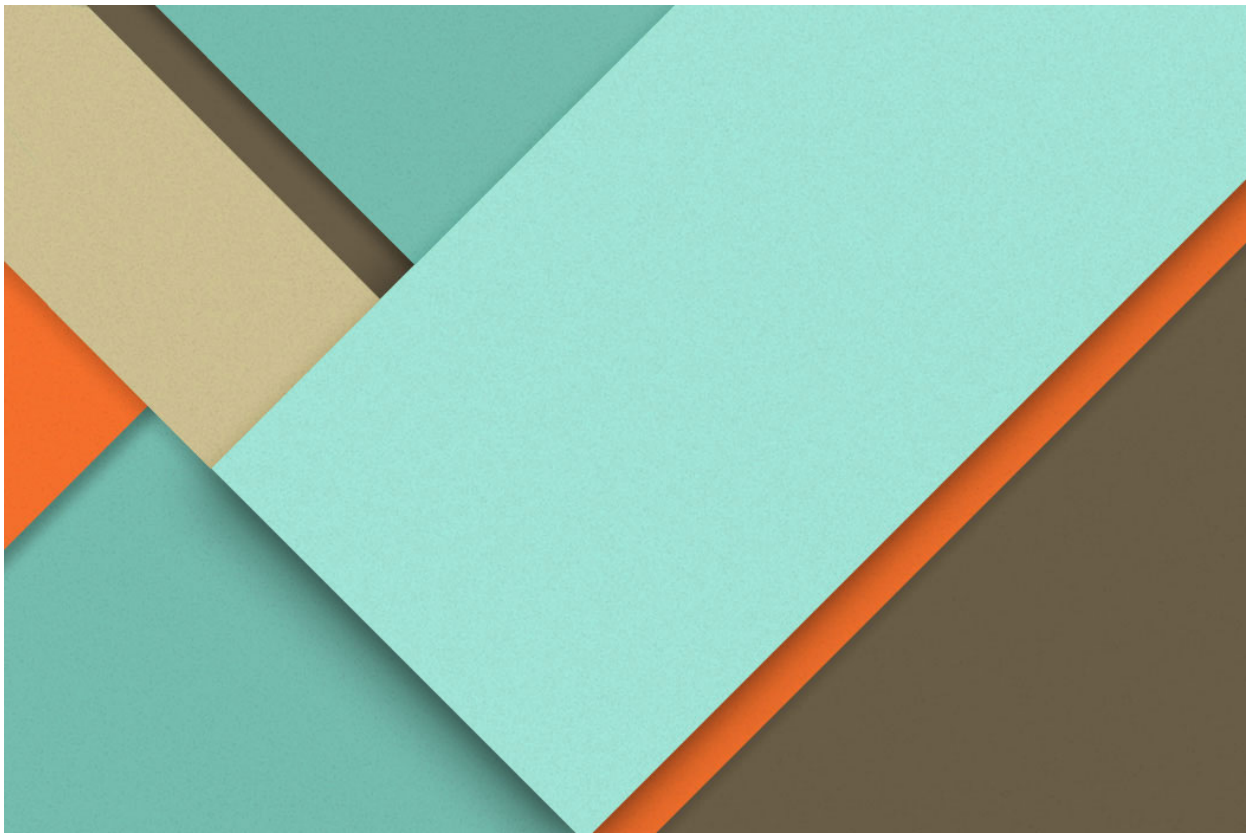


Predicting Connecticut Housing Prices: 2020-2021

Springboard Data Science Career Track Capstone

Steven Pereira

Introduction

This Capstone builds a predictive price model for residential single family homes. Data collected from the Connecticut Multiple Listing Service is used to identify characteristics which most affect listing closing prices. The data set used in this analysis consisted of 58,107 listings and their characteristics including acreage, square footage, bathrooms, and garage size. Any single family property which closed in Connecticut between 7/28/2020 - 7/29/2021 is included in this dataset, excluding private sales.

Audience

Finding an accurate model to predict housing prices can be valuable to multiple businesses. Being able to accurately predict the sales price of a target house would help lending institutions lower risk of overvaluing properties. This is especially important for limiting losses due to borrower default. This tool can also be useful for real estate investors and developers. Being able to identify factors that may make a house more valuable can focus where and how to develop areas. Additionally predictive modeling can be used in sales and marketing. Real Estate brokers and lead generating websites like Realtor.com or Zillow may use predictive modeling to aid pricing houses in the listing process or offering process.

Data Wrangling

This data was pulled from the CT MLS system from 7/28 - 7/29/21. The listings are those which closed within the range of 0-365 days from the previous dates listed. All cities available were pulled from CT. Listings are limited to single family homes, condos, co-op. Data from a total of 58,107 single family dwellings were collected. The raw data contained the following features. Listed below is how each feature was used and prepared for analysis

MLS Number

Each listing has a unique identifier. This feature was checked for duplicates to ensure that no observation was represented more than once in error. MLS Number was ultimately dropped after this check.

Status

On the MLS, a listing's status indicates whether a listing is active, expired, under deposit, closed, etc. In this analysis only sold listings were examined so all listings had the status of "Closed." This feature was checked to confirm all listings were "Closed" and subsequently dropped.

Property Type

This data consisted of single family residential dwellings. This feature indicates what type of dwelling the listing is. Included are 46,390 single family homes, 11,468 condominiums, and 249 cooperative properties. Ultimately the data was limited to single family homes, and all observations which were condominiums or cooperatives were dropped. These observations were dropped at the end of the data wrangling step

Listing/Closing Price

The listing and closing price are the price at which the property originally entered the market and the price which the property ultimately sold for. This feature had the following format: "LP: $3,750,000 CP: $3,750,000". This feature had to be split into 2 separate features: Listing Price and Closing price. These features were then stripped of non-numeric characters and converted from string objects to numeric.

Address

This feature is the address of the observation listing. This feature was ultimately dropped.

City

This feature is the city in which the observation listing is located. This feature was checked for null values.

Acres

Acres is the area of the land in which the listing sits, measured in acreage. This feature was checked for null values, which corresponded to property type. Only single family homes had an acreage listed. The null values were associated with condominiums and cooperatives, which were ultimately dropped from this analysis. The remaining observations were mixed numeric and object types. All observations were stripped of non-numeric characters and converted to numeric type. 673 observations had zero acreage. 362 of these observations were potentially mislabelled condominiums or cooperatives. All 673 observations were dropped.

Total Square Feet

Total square feet is the area of the living space in the listed building, measured in square footage. This feature includes the square footage of any basement spaces converted to living space. This feature needed to be stripped of non-numeric characters and converted to numeric type. 2 observations had invalid square footage listed. These listings were removed.

Square Footage Heated above Grade

Square Footage Heated above Grade is the area of the living space, in the listed building, measured in square footage that is not finished basement space. This feature was stripped of non- numeric characters and converted to numeric type. 16 observations had null values for this feature which were filled with valid values from Total Square Feet.

Style

This feature indicates the style of the observation house. For example: Colonial, Raised Ranch, Cape Cod, etc. This feature contained single and compound responses. Compound responses like [“Colonial”, “Contemporary”] were limited to the first list response or in the example’s case, “Colonial”.

Total Rooms

Total Rooms represents the total rooms in an observation listing. This count will include dining rooms, bedrooms, family rooms etc. This feature was numeric and had no null values however 3 observations were dropped due to miscellaneous issues.

Total Bedrooms

Total bedrooms indicates the number of bedrooms in the house. Usually defined as a room with a door and closet. This feature required no modifications as it had no null values and was already a numeric type.

Bathrooms

Bathrooms indicate the number of bathrooms in a listed house. Bathrooms can either be half or full indicating the presence of shower/bathtub. This feature had the following format, "2 Full & 1 Half". First 16 observations were dropped for null values in this feature. The feature was split into new features: Full Bath and Half Baths. Then the feature was stripped of non-numeric characters and converted into numeric type.

Garage/Parking

Garage and Parking indicates the number of spots for cars and the type of parking available at listed property. This feature was in the following format "2 Car/Attached Garage" or "2 Car/Off Street Parking". This feature was split on "/" to determine which observations contained "garage" parking. Observations which had garage parking used the spots to indicate the size of the garage. All non-numeric characters were stripped and the feature was converted to numeric type. All other observations, without garage parking had this feature set to 0.

Year Built/Age

Year built indicates the year in which the listing was built. 4 observations had null values which were manually changed using data from Zillow. The age of the property was calculated from the year built and added as a feature.

Day on the Market

Days on the market indicate the number of days in which the listing took to sell measured from the listing date.

Listing Agent/Office

Each observation had the information of the Listing Real Estate Agent and their respective offices. This feature was ultimately dropped

County

Using a table of Connecticut cities and their respective counties, a merge resulted in the dropping of 162 observations due to them being listed under states other than Connecticut.

Outlier Removal

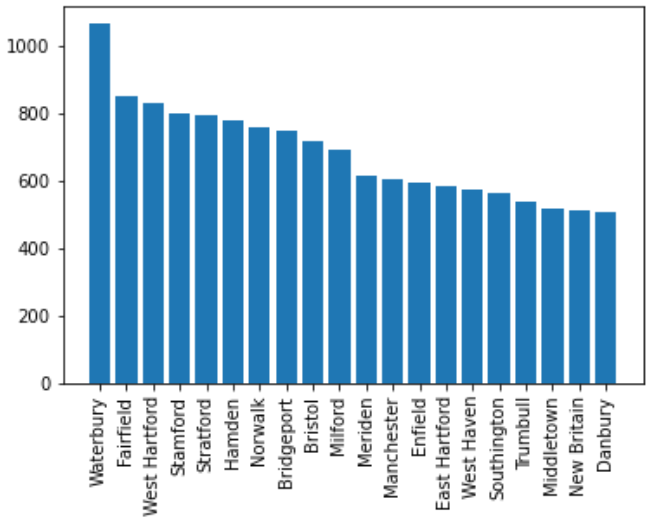
Using a standardized removal of outliers, observations were dropped by looking at Acreage, Total Square Footage, Square Footage Heated above Grade, Total Rooms, Total Bedrooms, Full Baths, and Half Baths. This was done by removing observations whose features were outside 1.5 times the interquartile range.

Exploratory Data Analysis

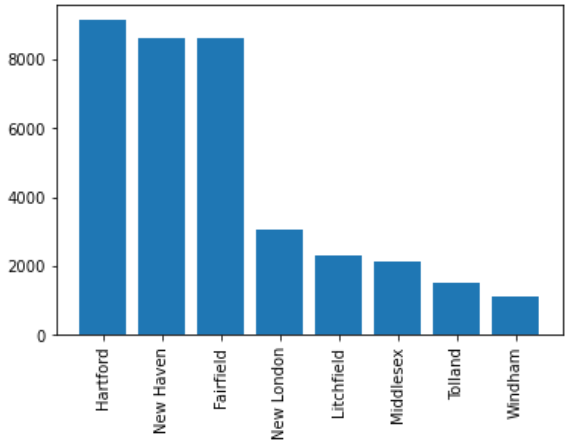
The final data set consisted of 36,519 single family houses after 11,468 condominium and cooperative listings were dropped. An additional 9,871 observations due to errors, missing values and outliers. Below are some of the feature distributions of the dataset.

Distributions

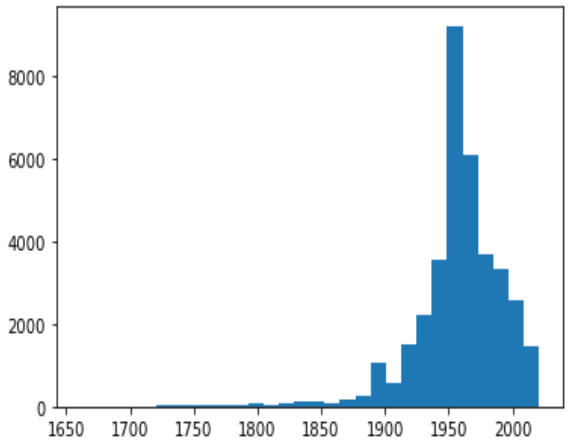
City Counts



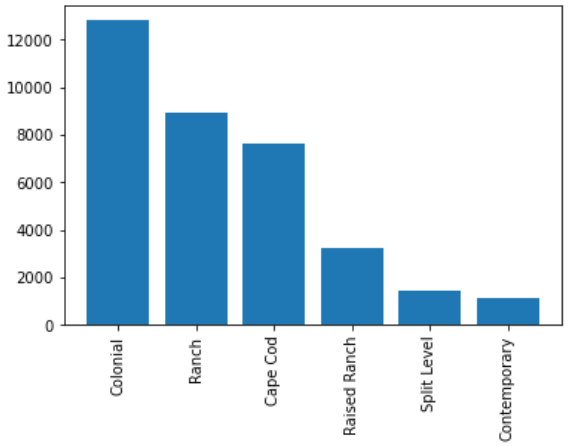
County



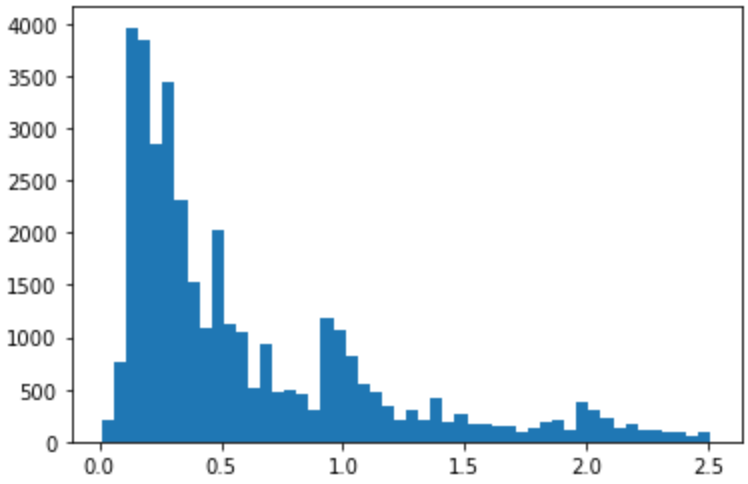
Year Built



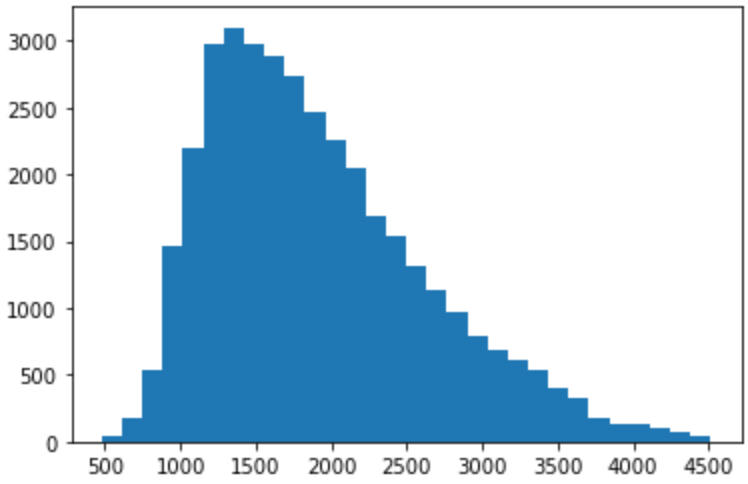
Style



Acreage



Total Square Footage



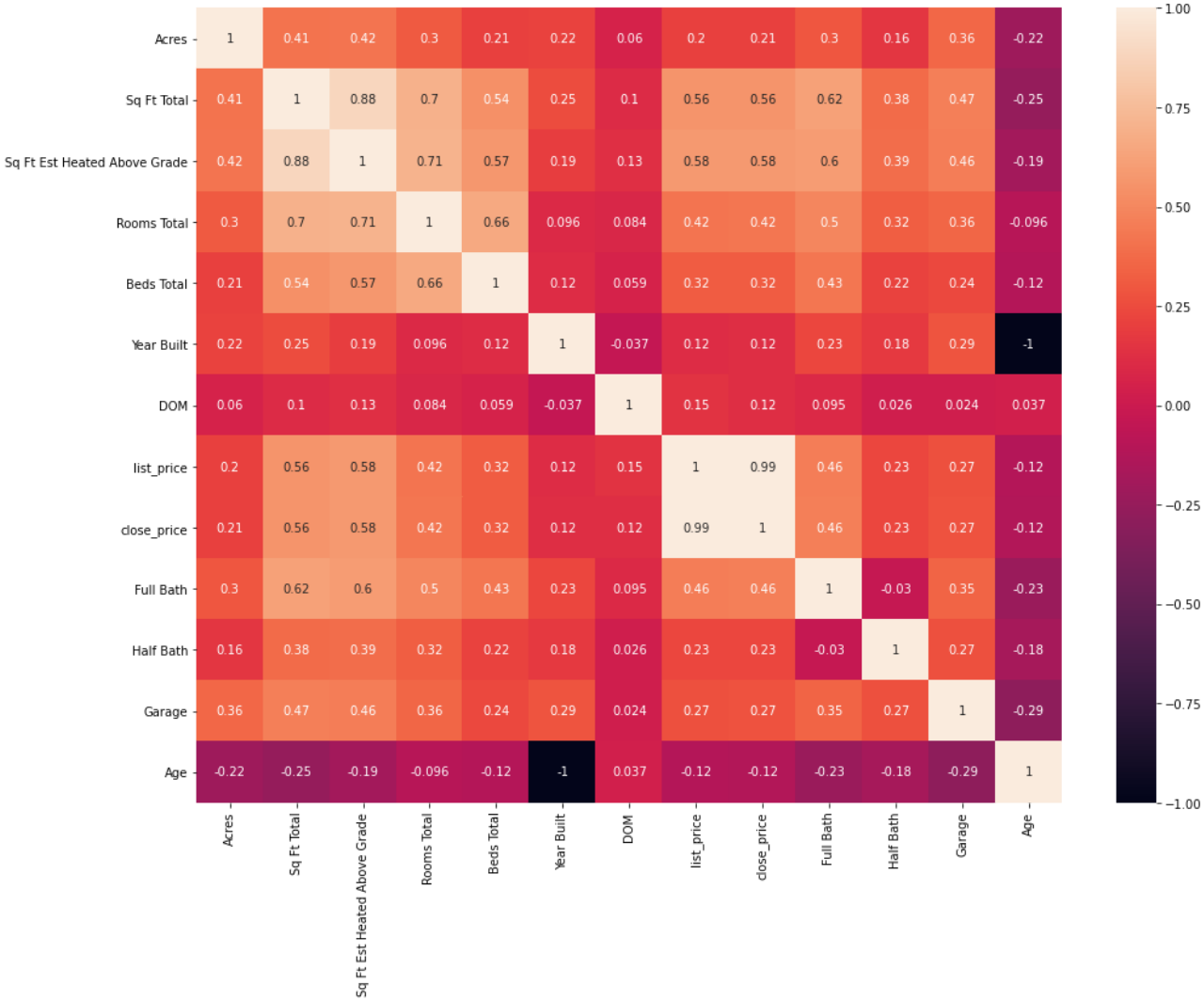
Distribution Summary

The average listing from this dataset is a 1900 square foot 3 bedroom, 2 bathroom house with a 1 car garage on .61 acre of land, built in 1957. The average listing was listed for $363,000 and sold for $365,000. Houses were most commonly in Hartford, New Haven, or Fairfield county. The mean and median number of days on the market was 43 days and 24 days. The most common house styles were Colonial, Ranch, or Cape Cod.

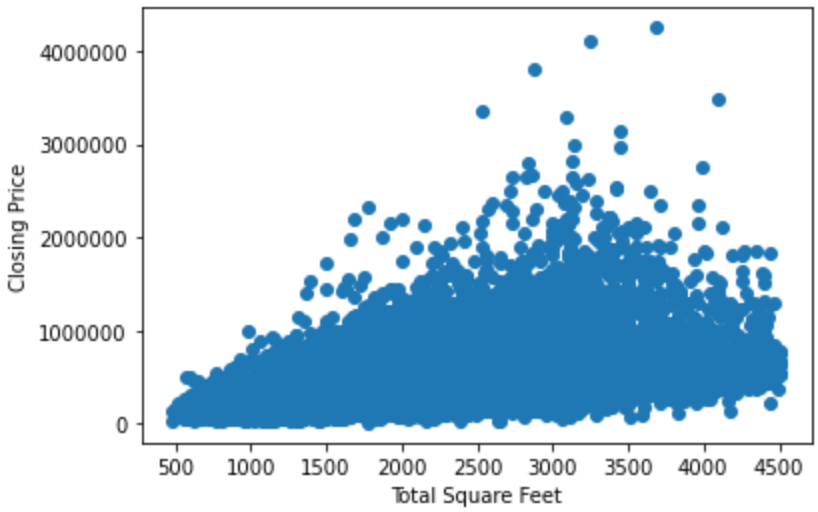
Interactions and Correlations

The following section shows some of the correlations between features. Additionally this section shows the correlation between features and the target variable, closing price.

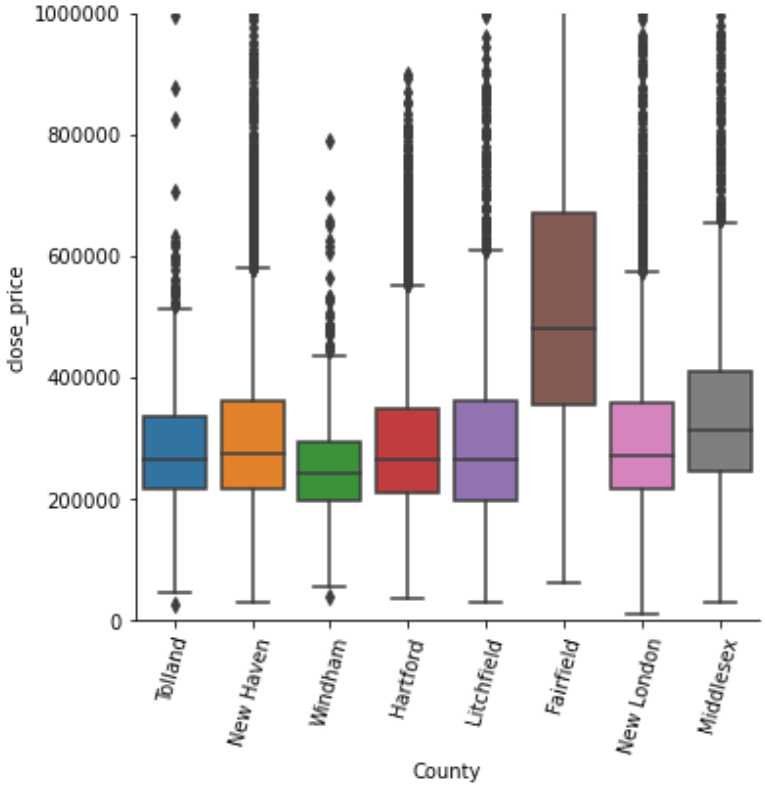
Feature Correlation Heatmap



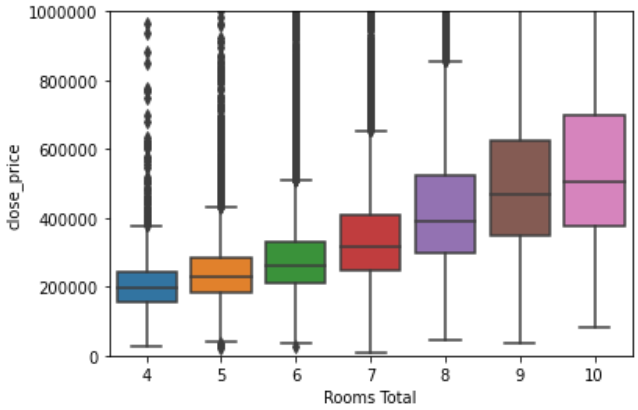
Closing Price By Total Square Feet

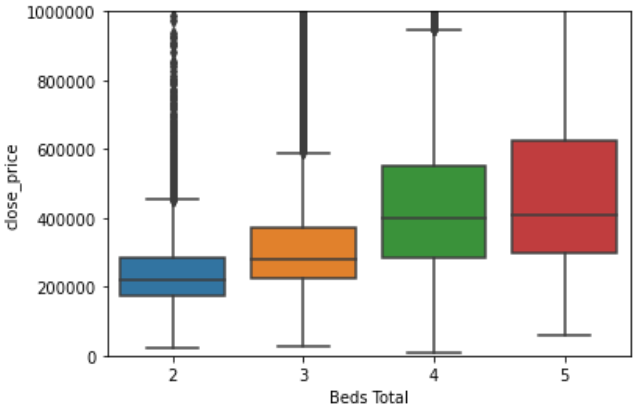


Closing Price by County

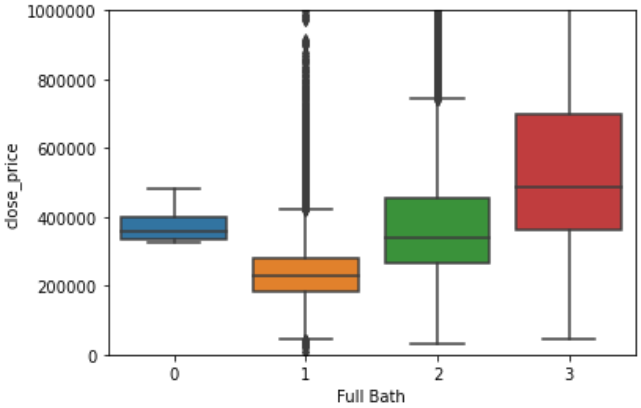


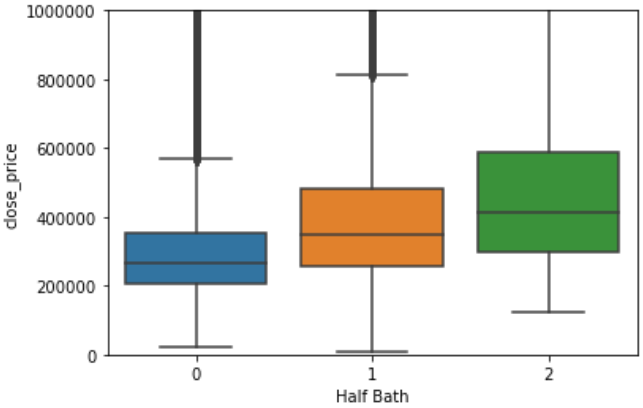
Rooms and Bedrooms by Closing Price



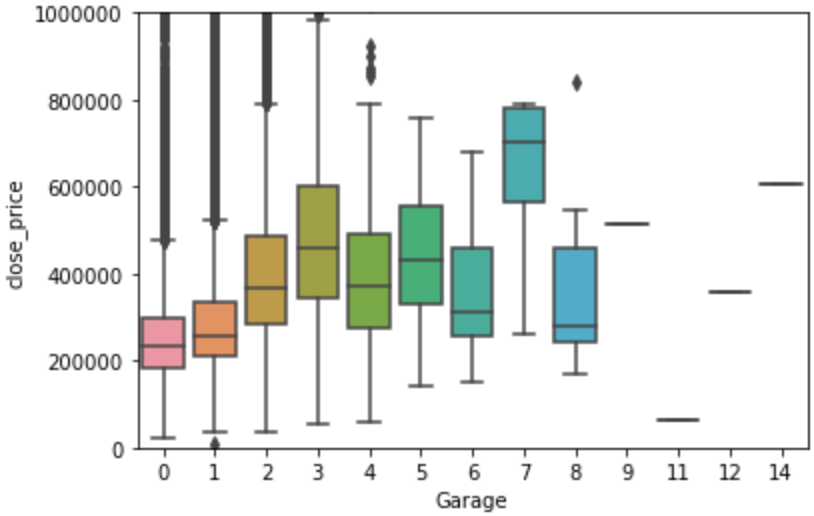


Full Baths and Half Baths by Price





Garage size by Closing Price



Correlation and Interactions Summaries

Between features there is a fairly strong correlation between some obvious pairs. The number of rooms, bedrooms, bathrooms are strongly correlated with the total square footage. Between the target, closing price, and the features there are a few strong correlations as well. Square footage, number of rooms, number of bathrooms are positively correlated with closing price. It is also apparent that the county plays a big part in determining the closing price of the listing.

Preprocessing

Feature Set to be Processed

```
 #   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 non-binary features were scaled to mean 0 and standard deviation 1. The scaled features included acreage, total square feet, square footage heated above grade, days on the market, age, total rooms, total bedrooms, full bathrooms, and half bathrooms.

One-Hot Encoding

Cities and counties were one hot encoded as features. This resulted in a total feature count of 188 features.

Train Test Split

The data was finally split into a training set and a test set. The test set was 25% of the samples. This resulted in a training set of 27,389 observations and a testing set of 9,130 observations.

Modeling

A total of 7 different models were explored. For linear models Ordinary Least Squares, Elasticnet, Lasso, and Ridge regressions were estimated. In an attempt to improve the performance of the linear models, each model was hyperparameter tuned. This however had little effect on model performance. All the linear models performed similarly well with a mean absolute error of $71,456, a root mean squared error of $130,391, and a R squared of 72%.

Random forest , gradient boosting, and k-nearest neighbors models were estimated to see if non-linear models would perform better at predicting house prices. Random forest performed the best out of all models thus far with a mean absolute error of $63,952, a root mean squared error of $122,584, and a R-Squared of 75%.

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

Final Model Selection

Random Forest will be the final model as it performed the best. This model was then hyperparameter tuned to improve performance.

Hyperparameter Tuning

```
param_grid ={'n_estimators':[50,100, 150],
             'max_depth' : [None, 20, 25, 30],
             'min_samples_split' : [2, 5, 10, 15, 20],
             'min_samples_leaf' : [1, 3, 4]}

RF = RandomForestRegressor(random_state = 343, n_jobs = -1)

RF_RCV = RandomizedSearchCV(RF,
                            param_grid,
                            scoring = 'neg_root_mean_squared_error',
                            random_state= 343,
                            n_jobs = -1)
```

Using a randomized grid search the number of trees, tree depth, minimum sample split, and minimum sample leaf were tuned. The optimal values resulted in a modest improvement in model performance metrics. The features which had the highest effect on predictions were Square Footage Heated above Grade and whether it was in Fairfield county.

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