Predicting Home Values in Los Angeles' South Bay

Springboard | Capstone 1 Final Report

by: Lauren Broussard

Summary:

This project sought to predict home prices in one area of Los Angeles County: South Bay. Home price information is helpful for a number of clients such as new home buyers/sellers, real estate agents, and new home builders.

Data Wrangling and Exploratory Data Analysis (EDA) were done primarily using the pandas library in Python. The original dataset was manually downloaded from Redfin.com into 84 separate files. It was merged, cleaned and reduced to 13,631 sales, with 18 individual features, ranging from 2018 and 2020.

Some insights from the project include:

- *More homes were sold in the summer, and there seems to be a price difference between homes sold in summer versus homes sold in other seasons.*
- *The top two predictors of home prices in this area were Square Feet and Longitude (i.e. proximity to the ocean).*
- *Other predictors of home price include: Year Built, Lot Size, and whether a home is in a specific neighborhood, with homes in Hermosa Price being a larger predictor than other neighborhoods.*

I was able to train a model using Random Forest Regression in scikit-learn to predict homes with an R^2 of 0.84. Corresponding code for the project and other write ups (including slides) can be found on Github.

Business Problem:

Home prices in Los Angeles County remain consistently high, which can make purchasing a home difficult for those unfamiliar with the market, the neighborhoods, or the popular home features in those neighborhoods (i.e. number of bedrooms, time of year, proximity to the ocean, etc). Being able to better predict home prices would be a benefit to first-time home buyers, new home builders, or real estate agents. For example, for a first-time home buyer, the decision to put a bid on a home is a big one, and knowing early on whether or not a particular home is over-(or under-)valued would save time, money, and unnecessary stress. Additionally, this kind of information may be useful for a new home builder, as they assess the features that would be important to get the most value for their new build. In this project, I looked at a specific set of neighborhoods in the southwest corner of Los Angeles County, called the South Bay. The project looks at home sales in that area for an approximate 2-year period.

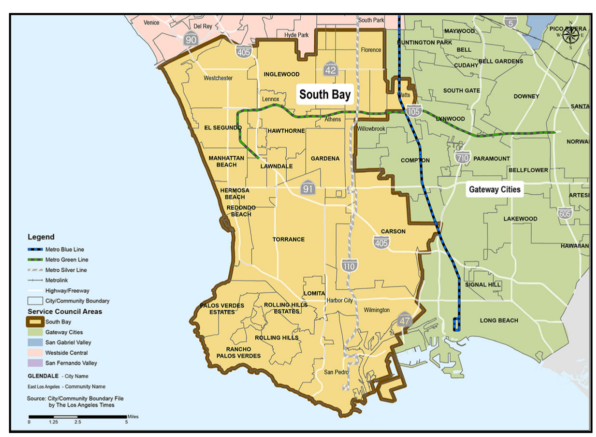


Image: South Bay Area; Source: Metro.net

Potential Client(s):

- *New Home Buyer/Investor:* For someone considering a home purchase, it would be useful to know whether or not a home is worth bidding on - i.e. if the home is undervalued in the market. Additionally, for an individual selling their home, it would be useful to know what might be a suitable asking price.
- *New Home Builders:* Being able to predict the price could help a new home builder determine what features may be important to add to a home to get the highest asking price for the area.
- *Real Estate Agents*: This information would help them to determine what a reasonable listing price could be for a client's home.

Data Collection and Wrangling:

The data was retrieved from Redfin.com. The site allows you to download sales, but limits downloads to 350 properties at a time. Files were downloaded manually from the site in groups of approximately 350 properties, and at 2 different points in time. The resulting .csv files were named according to neighborhood and filtered attributes (i.e. 3 bedroom properties in Redondo Beach).

The original dataset consisted of **19,527 rows and 29 columns in 84 separate files**, representing South Bay home sales for an approximate 2-year period.

Data was merged and cleaned in Python using pandas. The following cleaning steps were taken:

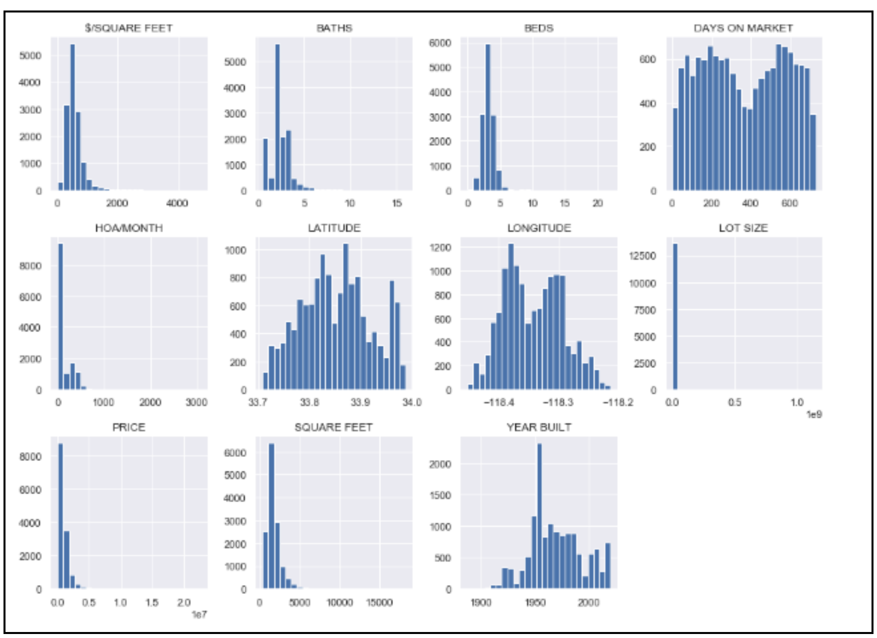
Merging Tables: The 84 files were imported as pandas data frames and merged using *pandas.concat()*. Since data was collected at two different times, the column "collection" was added (Collection 1: 51 csv files, and Collection 2: 33 csv files) to distinguish which sale dates were included, in case that information would be needed in the future. This column was eventually dropped once all data was merged and other cleaning steps were completed.

Dropping Columns: The following columns were dropped from the data frame, because they contained information not useful to solving the problem, the columns were empty, or because the information included was identical down the entire data frame: *'SALE TYPE', 'NEXT OPEN HOUSE START TIME', 'NEXT OPEN HOUSE END TIME', 'STATUS', 'FAVORITE', 'INTERESTED', 'URL', 'SOURCE', 'LOCATION', and 'STATE.'*

Deleting Duplicates: After inspecting the duplicated data, I determined that most were either due to a difference in the 'NEIGHBORHOOD' (a column that was added based on the original file name of the downloads), or a difference in the number of days a house was on the market The neighborhood column was replaced in the duplicated rows with the city name, and the max value was used for days on market when those were the cause of the duplicates.

Missing/Incorrect Values:

- **SOLD DATE, PROPERTY TYPE:** I used the .dropna() method to remove properties with no sold date, and dropped all properties other than the following types: *Single Family Home, Townhouse, Condo/Co-op, Mobile/Manufactured Home.*
- **ADDRESS, CITY, ZIP/POSTAL CODES:** I manually looked up missing addresses, city, and zip/postal codes to verify information against other real estate sites and Google. I dropped any records I could not verify. Additionally, I updated any incorrect data, such as a property with a postal code outside of the CA range.
- **PRICE, HOA/MONTH:** I removed any rows in which the price was less than $10,000 (one was a property priced at $25). I filled empty HOA/MONTH values with 0, as having no HOA is not uncommon for some properties.
- **BEDS, BATHS:** I dropped rows with missing information on the number of bedrooms and/or bathrooms. I tried searching the information as well on Redfin and Zillow (another real estate site) to verify but could not find consistent information. I dropped 2 rows of single family homes with beds/baths higher than 10 that I could not verify between Redfin, Zillow, and the LA County Assessor's data.
- **LOCATION/NEIGHBORHOOD:** I dropped the original "Location" feature as the format of entries did not seem to be standardized. Instead, I created a column called "Neighborhood" based on the neighborhood I looked up on Redfin and the name of the created .csv file.
- **SQUARE FEET, LOT SIZE, $/SQUARE FEET:** I dropped 6 records (SFH & Townhomes) for which I could not verify the square feet information through another source. For Mobile Homes with no square feet information (56), I filled with the average square feet for other mobile homes in the area. Lot size was filled with the median size by property type. $/Square Feet was filled by calculating values from both the Price and Square Feet columns.
- **YEAR BUILT:** The earliest standing home built in Los Angeles was built around 1818. I looked for any properties sold before this date or after 2020 and dropped those records.



Bar Graphs of South Bay Home Sale Numeric Features

The final data frame consisted of **13,631 rows and 18 columns.** It includes home sales data in South Bay neighborhoods between **February 06, 2018 and January 24, 2020**.

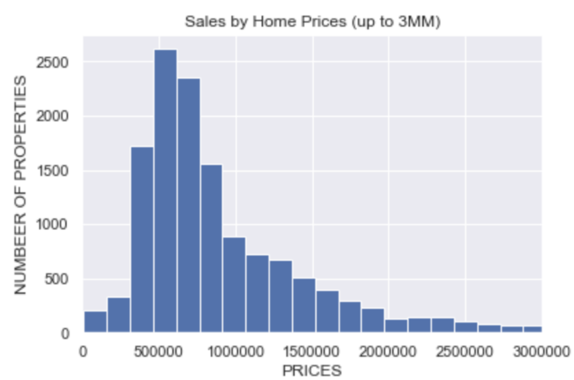
Initial Findings:

With the data cleaned, I turned my attention to looking further at some of the features of this dataset to see how the variables relate to each other, as well as how these variables relate to the price of a home.

Home Prices. To build a model to predict home prices, it would be good to have an understanding of the distribution of prices in the area. I created visualizations for prices in a number of ways: overall prices, then by property type, and finally by neighborhood. There is a large spread in home prices with a minimum price of $10,000, and a maximum at over $22 million, so I tend to display the median price.

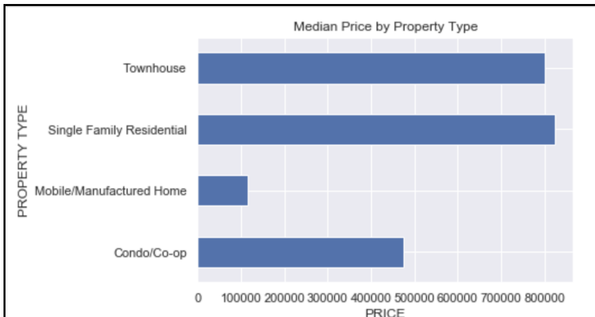
Overall. The histograms below look at the overall distribution of home prices. The histogram on the left displays all the data, while the one on the right zooms in to ignore outlier properties priced higher than $3 million. **It appears that the majority of properties are clustered between about $500,000 and $900,000,** and then begin to taper off around the $1 million mark.

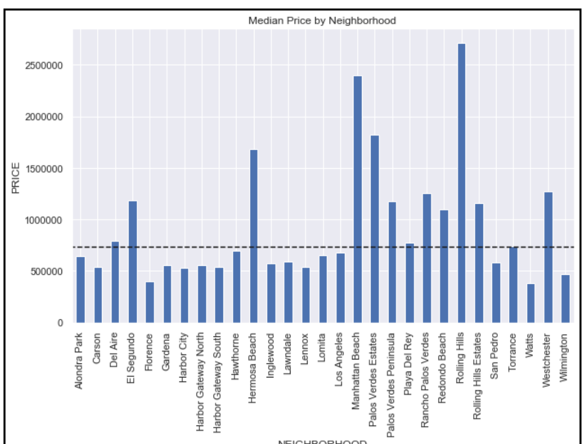
South Bay Home Prices: Overall Distribution



South Bay Home Prices: Capped at $3M

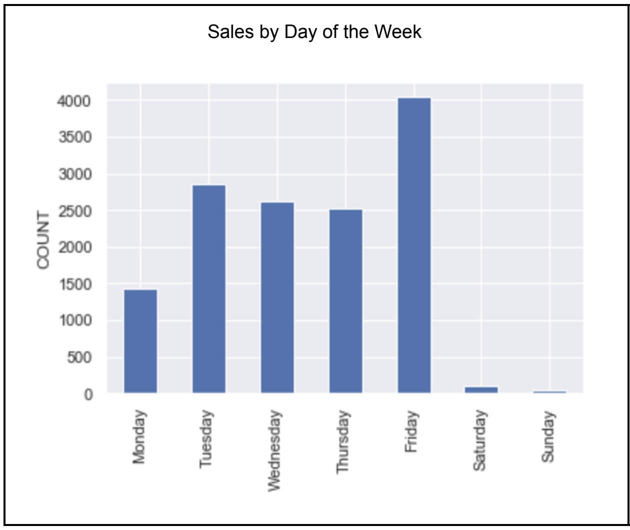
Property Type. When looking at median prices by home type, **single family homes and townhomes appear to have sold for a median price of about $800,000**, while the median price of a Condo/Co-op, or Mobile/Manufactured home is significantly less. This may be due to the difference in square feet between these home types, or some other factors. We may explore this further in the future.

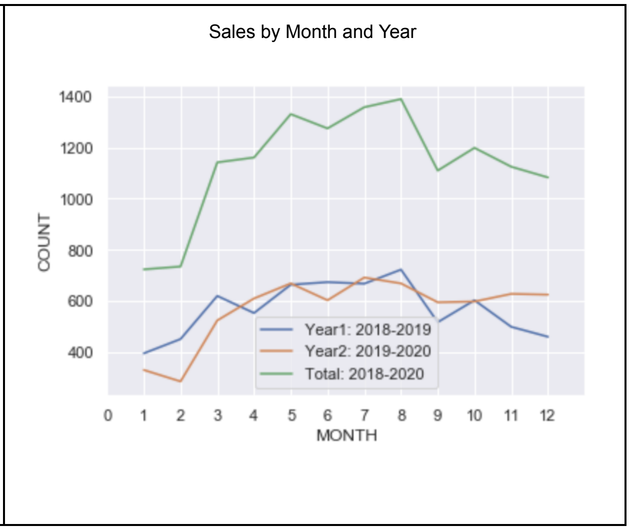




Neighborhood. **Four neighborhoods appear to have the highest median home prices: Rolling Hills, Manhattan Beach, Palos Verdes Estates, and Hermosa Beach,** three of which are near the ocean**.** We may look further in the future at what similarities exist between these three neighborhoods that affect home prices.

Timing of Home Sales. When are the most frequent time periods for home sales? Analyzing the data by the “SOLD DATE” column, I was able to group the data first by days of the week, and then by months of the year. **More homes were counted as “Sold” on Friday than any other day of the week**, though it is unclear what constitutes a home as being considered “Sold” (close date, contract date, etc.). Further, in looking at sales by month of the year**, it appears that overall, more homes were sold in the summer months in this data set.** Home sales were highest in July and August, and lowest in the months of January and February.

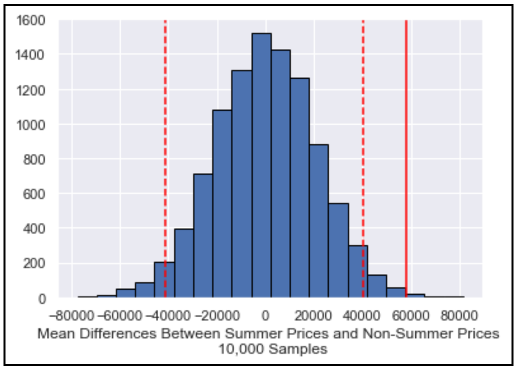




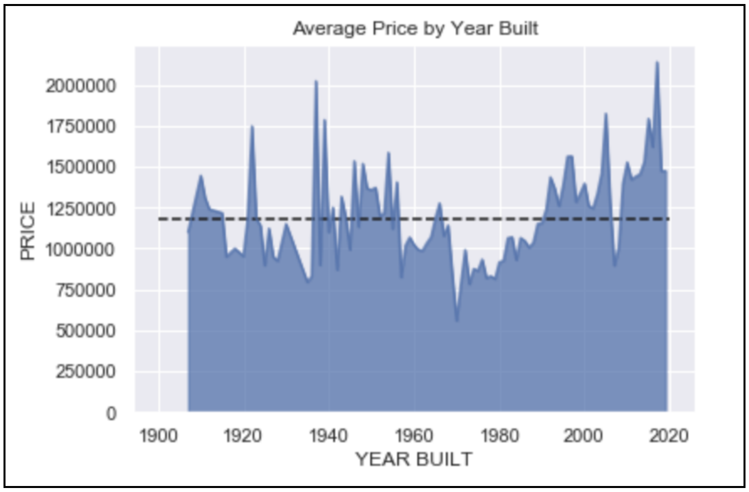
Summer vs. Non-Summer. Is there a “hot time” to buy a home? From initial data, it looked like the *number* of home sales in the summer months was greater than non-summer months. Could the increase in sales be due to a difference in price in the summer months vs. other months? I sought to answer the question: **Is there a difference in average home price for homes that sell in the summer vs. non-summer months?** I defined “summer” months as June, July, and August, and “non-summer” months as the other 9 months in the year. Then, I got values for the average price in each group, as well as the difference in means.

Mean Price, Summer: $1,034,713.47
<u>Mean Price, Not Summer: $976,924.60</u>
Difference in Mean Prices: $57,788.87

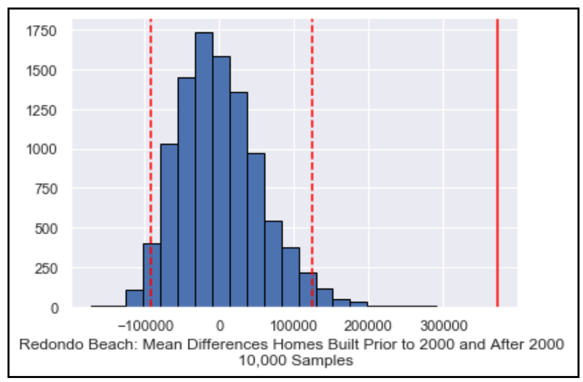
If the prices of the two groups were the same, we would expect the difference in mean to be closer to 0. We set the null hypothesis to be that there is no difference in the two groups, and chose a significance level of 0.05. I used Bootstrap Inference to simulate resampling our data, shifted the means of the groups so that the average prices were equal in both groups, and ran our experiment 10,000 times to see how likely it would be to get the mean difference observed above. **The p-value from this test was 0.0019, so we would reject the null hypothesis. There may, then, be a difference in price between summer and non-summer prices.**

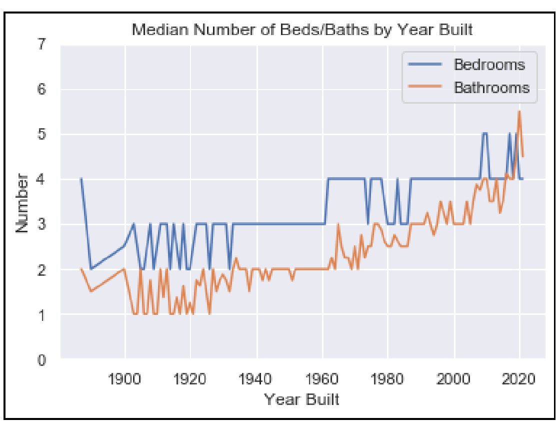


Newer vs. Older Homes. Are newer homes really all the rage? To control for the fluctuation in prices between neighborhoods, I chose one neighborhood to look at in more depth - Redondo Beach. I split the homes then into those that were built prior to 1970 and those built after 1970. I chose this year as it seemed that this was about when prices began to steadily increase.



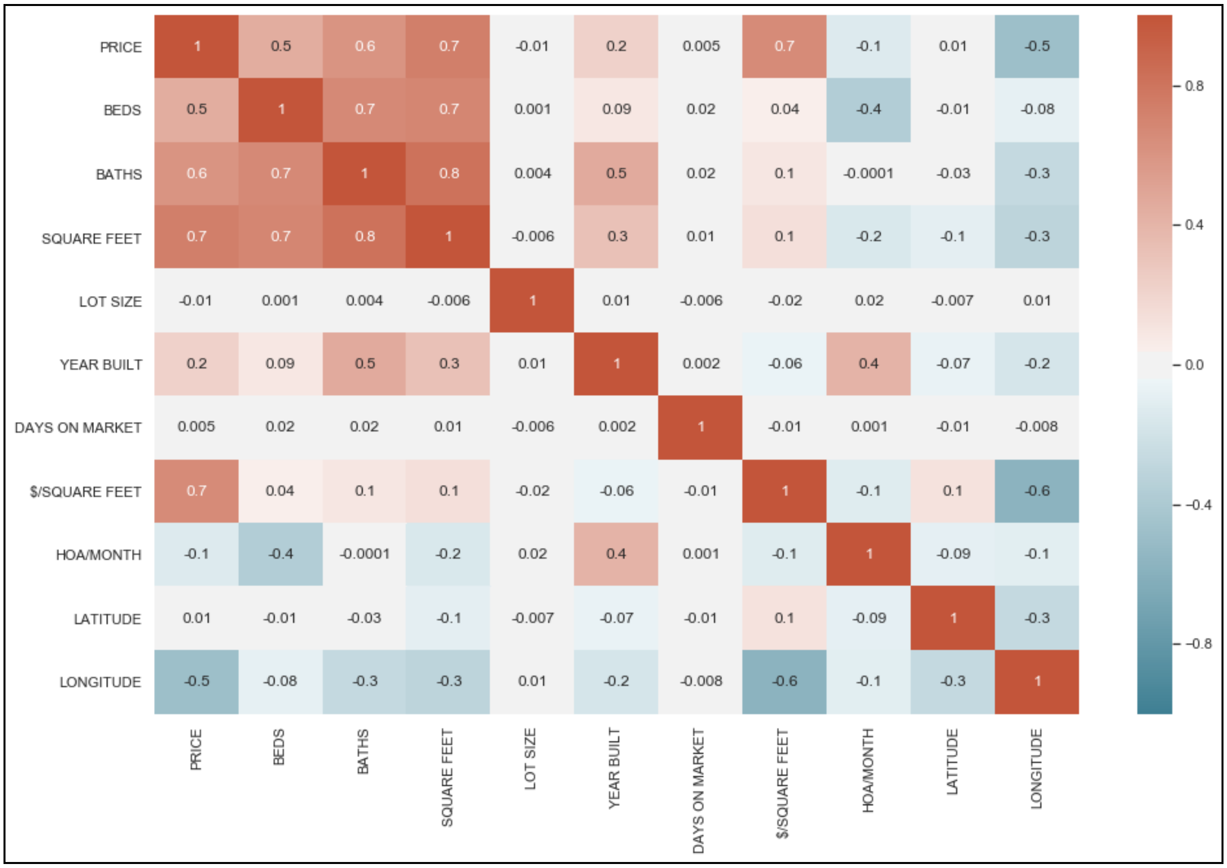
Again, I chose a Bootstrap test like the one above to compare the differences between these two groups. Running this test 10,000 times, **we ended up with a p-value of 0.0 for a significance level of 0.05. With this information, we would *reject* the null hypothesis that there is no difference in average price between homes built before 2000 and those built after in this particular neighborhood of Redondo Beach.**



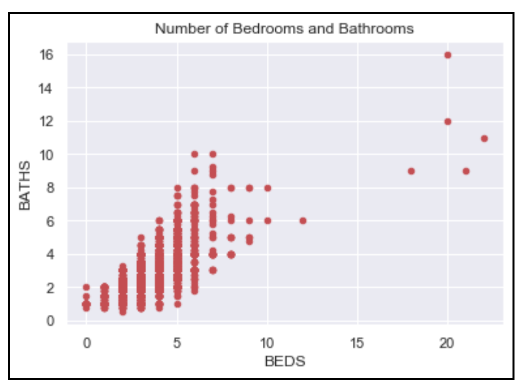


Bedrooms and Bathrooms. I also wanted to see whether or not features like bedrooms or bathrooms are increasing with the year the home is built. This was plotted looking at single family homes. The median number of bedrooms and bathrooms both seem to be increasing over time, although it appears that the number of bathrooms has more of an increasing trend (albeit a slight one).

Correlation Matrix. As we look further into the factors that affect price, let's look at a correlation matrix of all of the numerical features. We would expect that features like bedrooms, bathrooms, and square feet of a home are all positively correlated with price, and they indeed are. Also of note is the negative correlation between price and longitude. It may stand to reason that as homes get closer to the beach, the price increases. This is something we can look at further in the future.



As we saw earlier, the year a home is built is also slightly positively correlated with both the number of bedrooms and bathrooms. Looking at a scatter plot, we can also see that the number of bedrooms and bathrooms are positively correlated as we would expect - so the more bedrooms a home has, the more bathrooms it would also likely have.



<u>In-Depth Analysis -- *Random Forest Regression:*</u>

Finally, I used machine learning to see how well I could predict housing prices in the South Bay area. As the data is in the form of a continuous random variable (as opposed to a discrete variable), I looked at this as a Regression problem. Further, since we already have labeled data (features and housing prices), I took a Supervised Learning approach.

For this problem, I used Random Forest Regression, an ensemble method that expands on the Decision Tree approach.

Data Pre-Processing & Encoding. To prepare the data to work with the chosen model, I did data transformation of certain columns and data types. For instance, I dropped columns like 'Address', 'MLS#, and 'City." that were either redundant or could not easily be changed into a numerical value. Further, I broke 'SOLD DATE' out into its component parts (year, month, day) as separate columns since the model would not take the entire DateTime object. Finally, I create a column called "SEASON" in order to further see whether the time of year of a home sale affected the price of the home.

*One Hot Encoding***.** One Hot encoding was used with the *get_dummies* method in pandas to change categorical variables to binary values before putting them into the model.

	PROPERTY TYPE_Condo/Co-op	PROPERTY TYPE_Mobile/Manufactured Home	PROPERTY TYPE_Single Family Residential	PROPERTY TYPE_Townhouse	NEIGHBORHOOD_Alondra Park	NEIGHBORHOOD_Carson	NEIGHBORHOOD_Del Aire
0	0	0	1	0	0	0	0
1	0	0	1	0	0	0	0
2	0	0	1	0	0	0	0
3	0	0	0	1	0	0	0
4	0	0	0	1	0	0	0

Separate Data. From there, I segmented the data with X being all columns (features) of our DataFrame except for price, and y being the price column. Training and testing data were then separated out, with a 70/30 train/test split. This reserved 9541 homes in the training data and 4090 homes in the testing data. After preprocessing using one hot encoding, I ended up with 51 feature columns.

```python
# split data into training and testing
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3, random_state=42)
```

```python
print("Shape of training data:", X_train.shape)
print("Shape of test data:", X_test.shape)
```

```
Shape of training data: (9541, 51)
Shape of test data: (4090, 51)
```

Random Forest Regression. After running pre-processing steps, I ran a Random Forest Regressor model with the default parameters and all features included.To establish a baseline metric, I considered what might be a reasonable guess for the price of a home. Without any other knowledge, one might guess the median home price for the area. I set this as the baseline to see if our model could outperform that. The baseline Mean Absolute Error (MAE) using this logic was 485739.85, suggesting that just guessing the median area price our average error on home prices would be $485,739.

Running the initial prediction model yielded the following preliminary results:

	Original Model
Mean Absolute Error (MAE)	120915
Root Mean Squared Error (RMSE)	363501
R^2 Score	0.840562

The mean absolute error indicates that on average, the model predicts homes within about $120,915. While this is better than our baseline, I wanted to see if I could improve it further.

Additionally, I ran training and testing accuracy scores as found below:

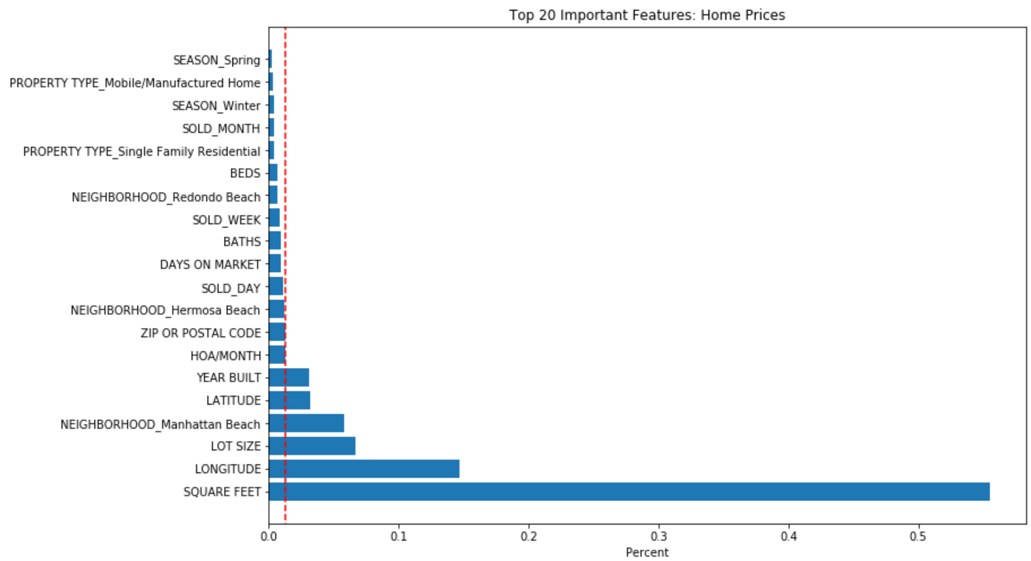
- **Mean Accuracy(Training):** 0.97
- **Mean Accuracy(Testing):** 0.84

The training data has an accuracy of 97%, but the accuracy goes down to 0.841 with the testing data. This could suggest some overfitting of the training data.

Parameter Tuning & Feature Selection.

Parameter Tuning. To see if I could improve the model, I first looked to tune the parameters **n_estimators** and **max_depth**. The parameter n_estimators is the number of trees to be used in the forest, and max_depth tells the model how far down the tree to go. Tuning these parameters suggested that the best values parameters are **n_estimators = 450** and **max_depth = 20.**

Feature Selection. Using sci-kit learn's Feature Importance method, I was able to list out the top 20 most important features. Looking further at this information, only 8 of those features explained 92.5% of the model. **By far, the two most important features in the model turned out to be the Square Feet of the property and the Longitude (seemingly, how close or far a home is to the ocean or inland).** The square feet of a property determined about 0.55 of the price prediction. Also of note are two of the neighborhoods (Manhattan Beach and Hermosa Beach) that appeared earlier as the most expensive neighborhoods in the area.

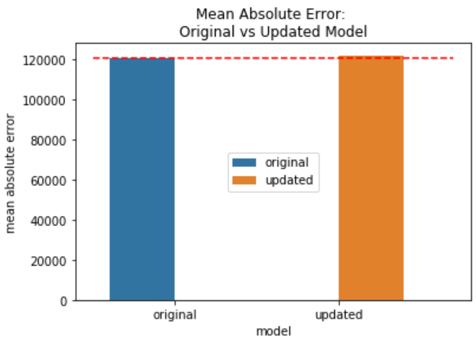


Evaluation of Updated Model. After implementing the newly tuned parameters and a subset of the original features, I retrained the model and reran the appropriate metrics to see how well it fit the data now.

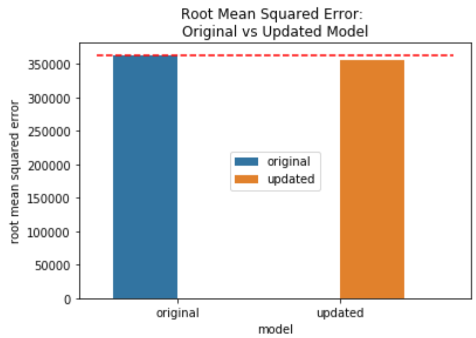
	Original Model	Updated Model	Difference
Mean Absolute Error (MAE)	120915	121040	124.574
Root Mean Squared Error (RMSE)	363501	356956	-6544.9
R-Squared Score	0.840562	0.846252	0.00568972

The final tuned model gave only slightly better results than the original model for some of the metrics, and slightly worse results for others.

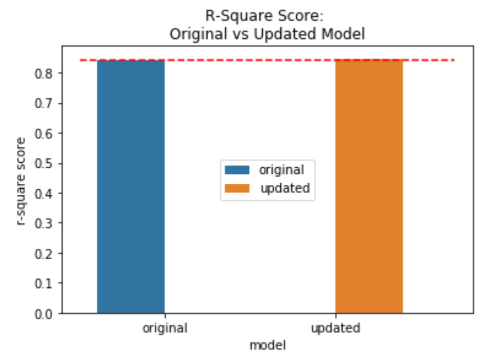
MAE



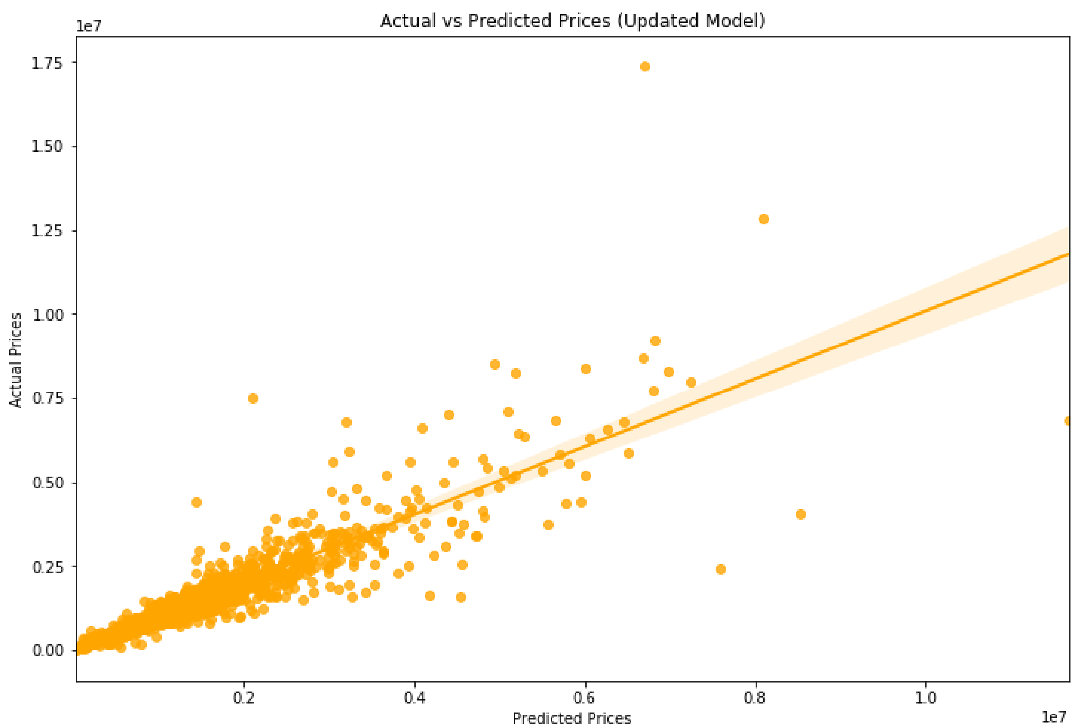
RMSE



R^2 Score

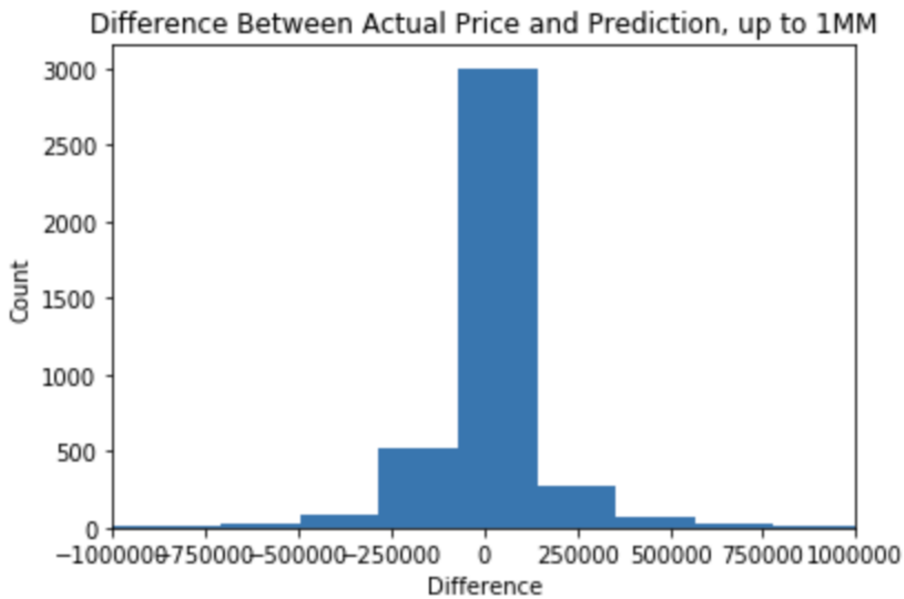


While I was able to beat the baseline prediction by $363,577 with the final model, the original model and the tuned model did not seem to perform significantly differently. The model seems to make fairly good predictions up to a certain price point, but does not predict well for others. This may be due to other outliers in the model or other factors. In the future it would be important to look for other parameters to tune, or work to normalize some of the other data.



Results & Recommendations:

The best prediction was off by less than $1,220, while the least accurate prediction was off by more than 5 million dollars (the actual price was approximately 17 million, and the predicted price was around 11 million). Of the 4090 test homes provided, the model was able to predict more than 50% of the homes within $50,000. From the model above, it appears that predictions were more accurate and precise up to about $400,000 - $600,000, and were less accurate for homes priced in the millions.



Client Recommendations.

- ***More Square Feet:*** *The most important predictor for home price in our model was Square Feet. Although "Tiny Homes" are getting more prominence in pop culture, homes providing more space are still selling for higher prices.*

- ***Location Matters:*** *As the old adage goes, "Location, Location, Location." The second largest predictor for home price in the model was longitude. This measures the location of the home east or west, or in other words -- how low close or far a property is from the ocean.*

- ***Summer Sales:*** *More homes were sold in the summer; and, there seems to be a price difference between homes sold in summer and homes sold in other seasons, with summer sales having a higher average price than other months. It may be worth it for someone looking for a home to try shopping outside of the summer months.*

- ***Other Features****: Other predictors of home price include: Year Built, Lot Size, and whether a home is in a specific neighborhood, with homes in Hermosa Price being a larger predictive feature than other neighborhoods.*

Other Considerations and Future Discovery:

For both of the Bootstrap tests done during EDA, there are other factors that could be at play to explain the difference in pricing data. For instance, the recession in 2008 could have helped to bring down the average prices in the group. It may make sense to look at homes built pre-recession and post recession, which may have yielded a different result. Also, home prices tend to go up over time in general, so the differences found could simply be due to prices rising over time. In a future project, I would need to control for other factors at play.

Considerations for future projects:

- **Other features may be interesting to look at, such as proximity to certain amenities** - beaches, gyms, grocery stores, etc.
- Additionally, I would want to look further into **the timing of a sale, specifically what constitutes a "Sale."** Are more homes selling on Friday for a reason, and is there a difference in price between a sale on a certain day of the week vs. another?
- While the data suggested that there was a difference in price between newer homes and older homes, in the future **it would be interesting to see how quickly newer homes sold versus older homes.**

Also of note to me was how much weight the overall square footage of a home had on its price. I assumed the neighborhood or another feature like the number of bathrooms would be strongly predictive. Knowing this may be important for home builders to consider in giving future homeowners the space and features that they really desire.