PREDICTING HOME PRICES IN THE SOUTH BAY

SPRINGBOARD DSC CAPSTONE PROJECT 1

by: Lauren Broussard

OUTLINE

- PROBLEM & POTENTIAL CLIENT(S)
- APPROACH AND DATA SET
- DATA WRANGLING STEPS
- EXPLORATORY DATA ANALYSIS
- IN-DEPTH ANALYSIS
- RESULTS
- RECOMMENDATIONS
- FURTHER CONSIDERATIONS

PROBLEM AND POTENTIAL CLIENT(S)

CAN WE CREATE A MODEL TO PREDICT HOME PRICES?

- Home prices in Los Angeles County remain consistently high
- Unfamiliarity with market can make purchasing difficult
- Better predictions can help purchasers or builders

POTENTIAL CLIENTS:

New Home Buyer/Investor

New Home Builder

Real Estate Agent

ABOUT SOUTH BAY

"The South Bay is a region of the Los Angeles metropolitan area, **located in the southwest corner of Los Angeles County.** The name stems from its geographic location stretching along the southern shore of Santa Monica Bay. The South Bay contains **fifteen cities plus portions of the City of Los Angeles and unincorporated portions of the county.** The area is bounded by the Pacific Ocean on the south and west and generally by the City of Los Angeles on the north and east."

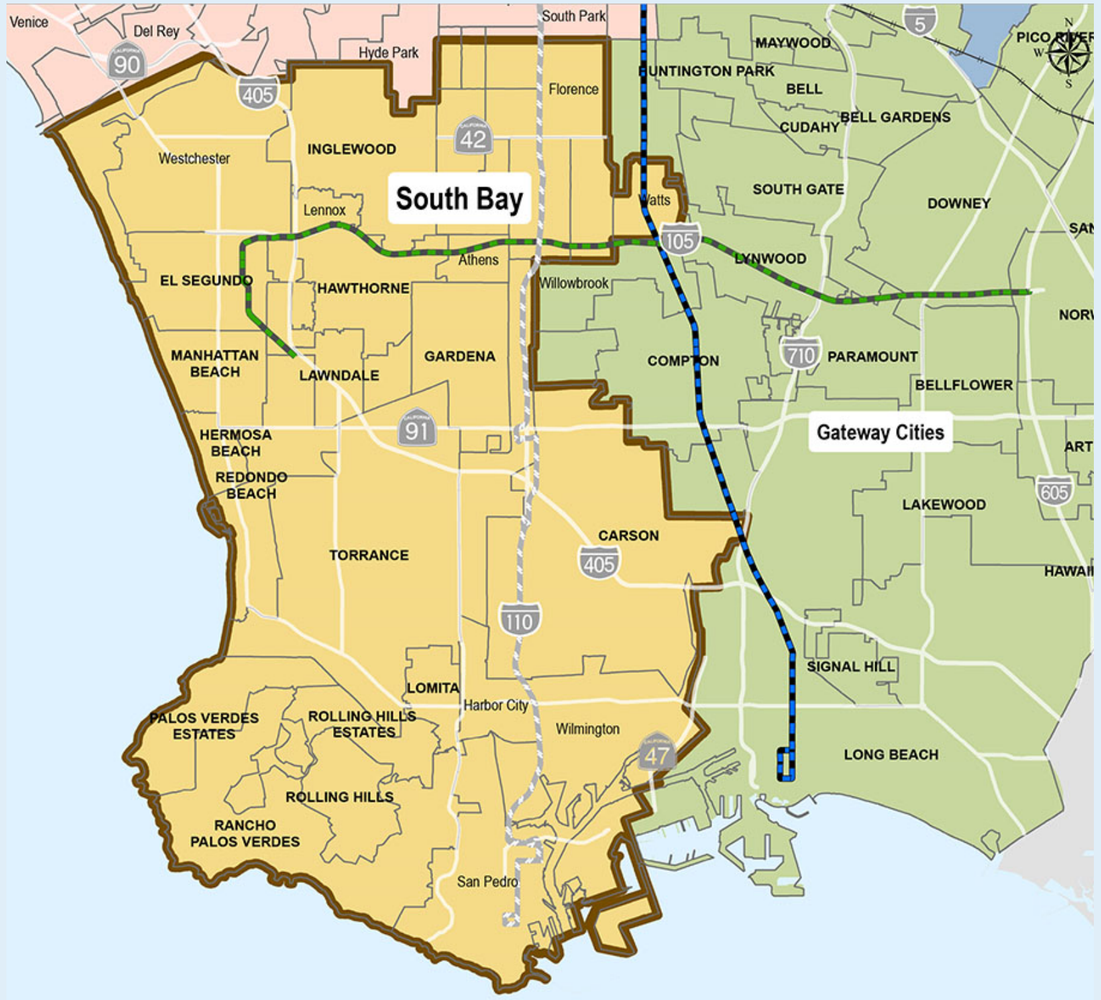


IMAGE SOURCE: METRO.NET

SOURCE: WIKIPEDIA, SOUTH BAY (LOS ANGELES C

APPROACH & ORIGINAL DATA SET

> DATA SOURCED FROM REDFIN.COM

- The data was retrieved from Redfin.com. Redfin provides property data for sold homes including: ***sale price, sale date, lot size, # of bedrooms, # of bathrooms, zip code, etc.***

> DOWNLOADED & MERGED

- Data was manually downloaded 350 properties at a time, per Redfin's downloading rules. Neighborhoods were chosen by those included in LA Metro's South Bay service map. **Eighty-four separate files were downloaded, merged, and cleaned using pandas for Python.**

> PRE-CLEANING DATA SET

- **19,527 rows and 29 columns,** representing South Bay home sales for an approximate 2-year period.

DATA WRANGLING

CLEANING: USING PANDAS

> DROPPING COLUMNS:
Superfluous columns, columns with empty data, or those not crucial to solving the problem, were dropped from the DataFrame (i.e. 'SALE TYPE', 'NEXT OPEN HOUSE START TIME', 'NEXT OPEN HOUSE END TIME', 'STATE')

> REMOVING DUPLICATES:
Duplicated data was largely due to the neighborhood column, or days on market (i.e. if a home was taken off the market and put back on). Neighborhood duplicates were filled with city name, and days on market were filled with the max value.

> MISSING/INCORRECT VALUES:
Missing values were validated on other housing sites where available, were filled with median values, or were dropped from the DataFrame. Properties with no sold date, or outside of certain property types, were removed.

DATA WRANGLING

FINAL SOUTH BAY DATASET

13,631 ROWS AND 18 COLUMNS

```
south_bay.head(3)
```

	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	PRICE	ZIP OR POSTAL CODE	BEDS	BATHS	SQUARE FEET	LOT SIZE	YEAR BUILT	DAYS ON MARKET	$/SQUARE FEET	HOA/MC
0	2019-02-01	Single Family Residential	1641 Bay View Ave	Wilmington	730000	90744	7.0	5.0	3401.0	6651.0	2008	358.0	215.0	
1	2018-05-31	Single Family Residential	1410 W Sandison St	Wilmington	547000	90744	4.0	2.0	1948.0	5399.0	1962	604.0	281.0	
2	2019-10-31	Single Family Residential	1703 N Marine Ave	Wilmington	774000	90744	5.0	3.5	2900.0	5857.0	1940	86.0	267.0	

PROPERTIES SOLD BETWEEN:
FEBRUARY 06, 2018 AND JANUARY 24, 2020

EXPLORATORY DATA ANALYSIS

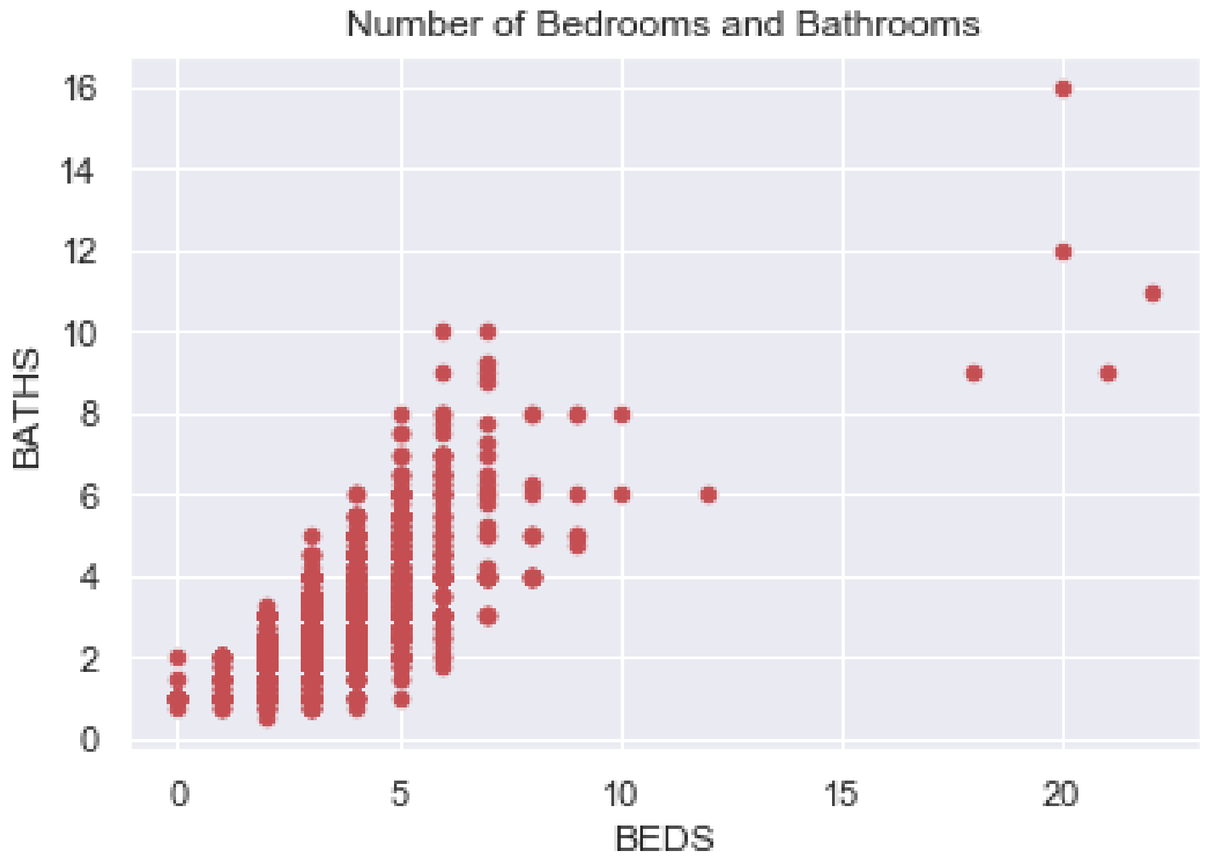
HOME PRICES & FEATURES

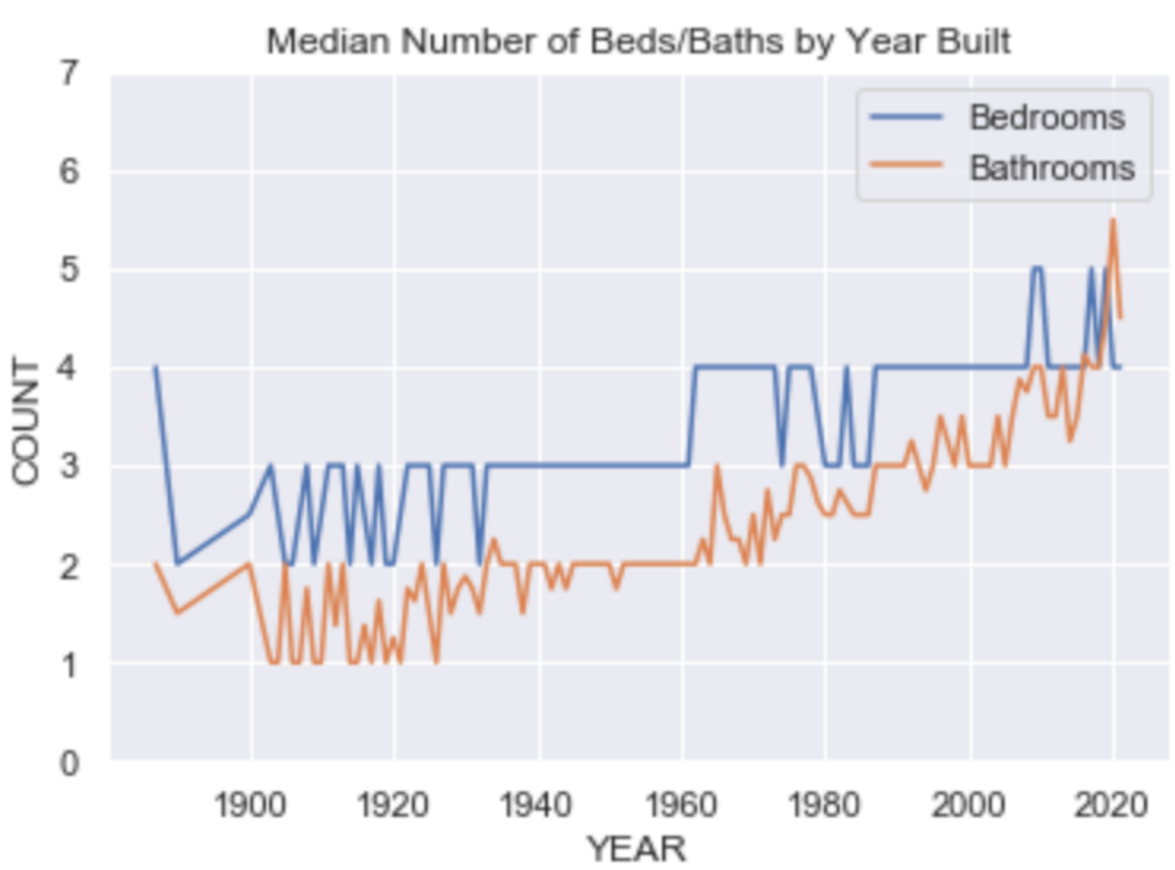
MIN PRICE: $10K

MAX PRICE : $22MM

BEDROOMS AND BATHROOMS

The median number of bedrooms and bathrooms both seem to be increasing together over time, although it appears that the number of bathrooms has more of an increasing trend (albeit a slight one).

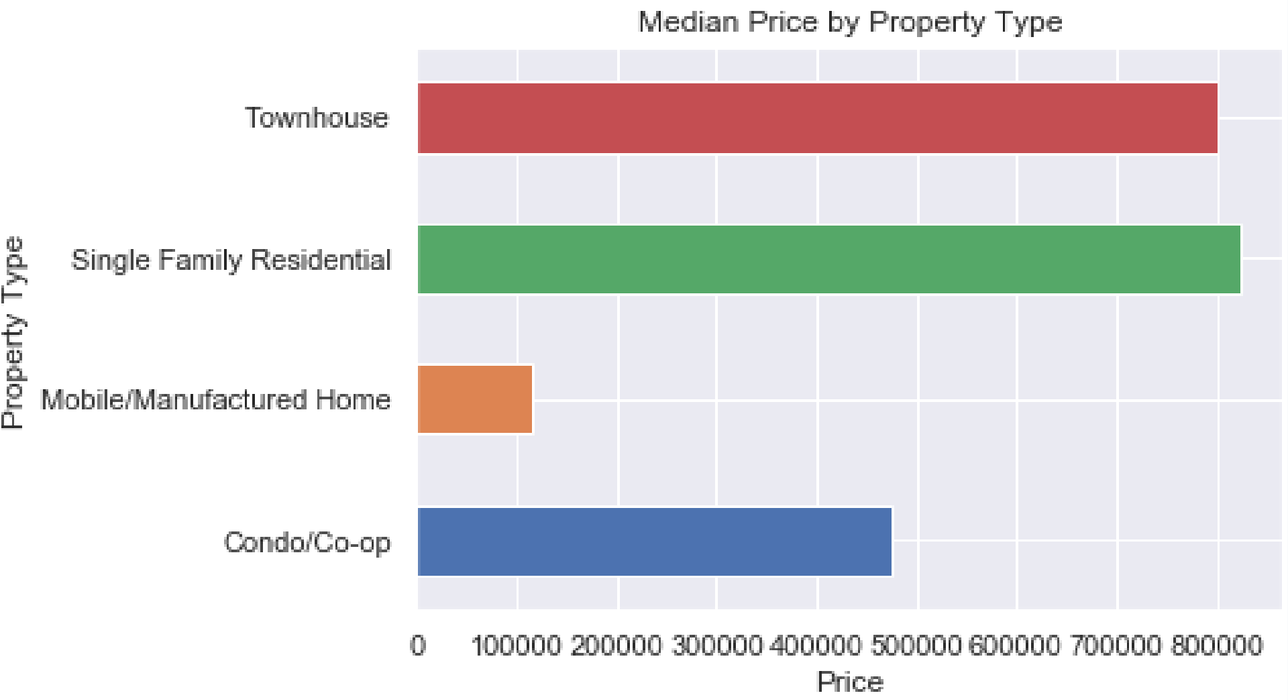


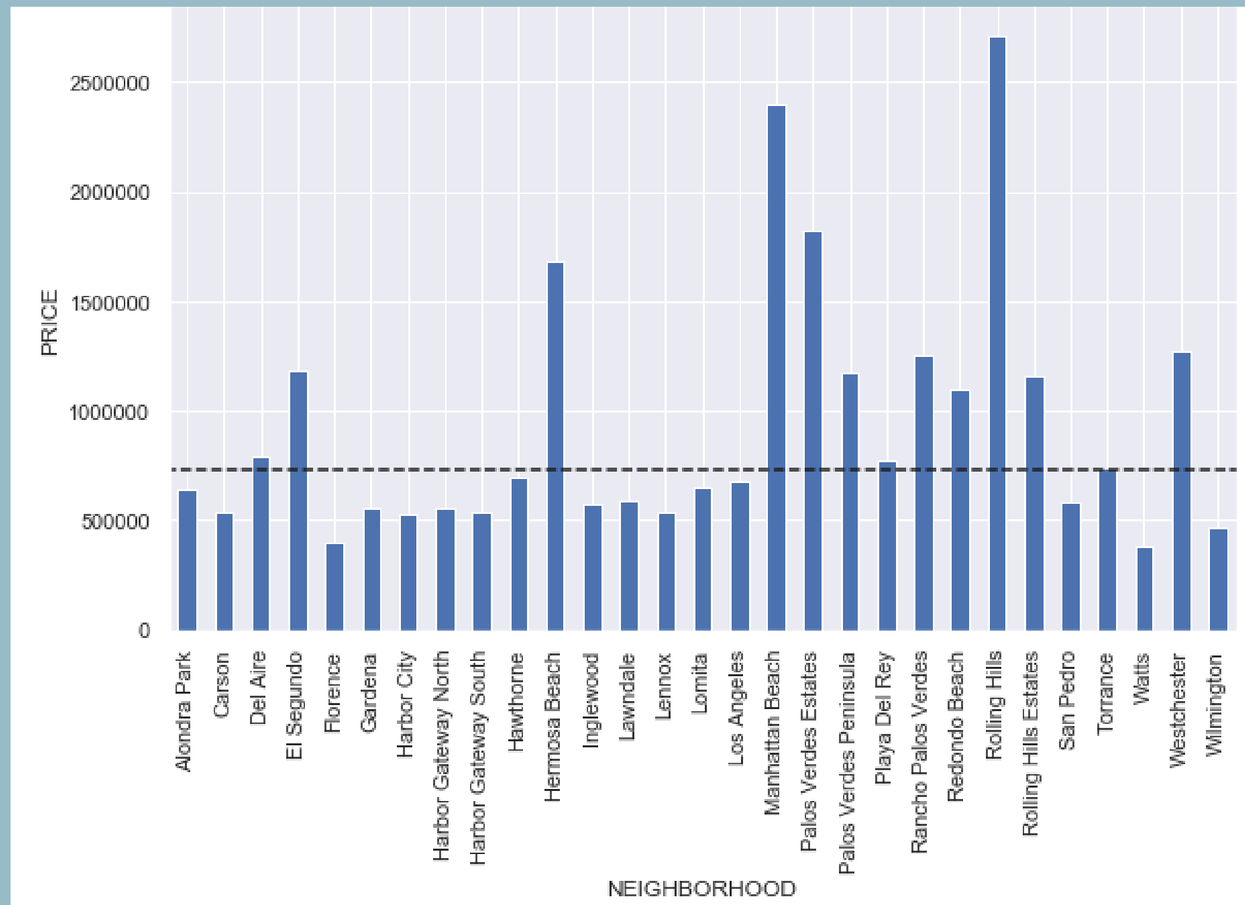


PROPERTY TYPE & NEIGHBORHOOD

- **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.

- Four neighborhoods appear to have the highest median home prices: **Rolling Hills, Manhattan Beach, Palos Verdes Estates, and Hermosa Beach,** three of which are right next to the beach**.**

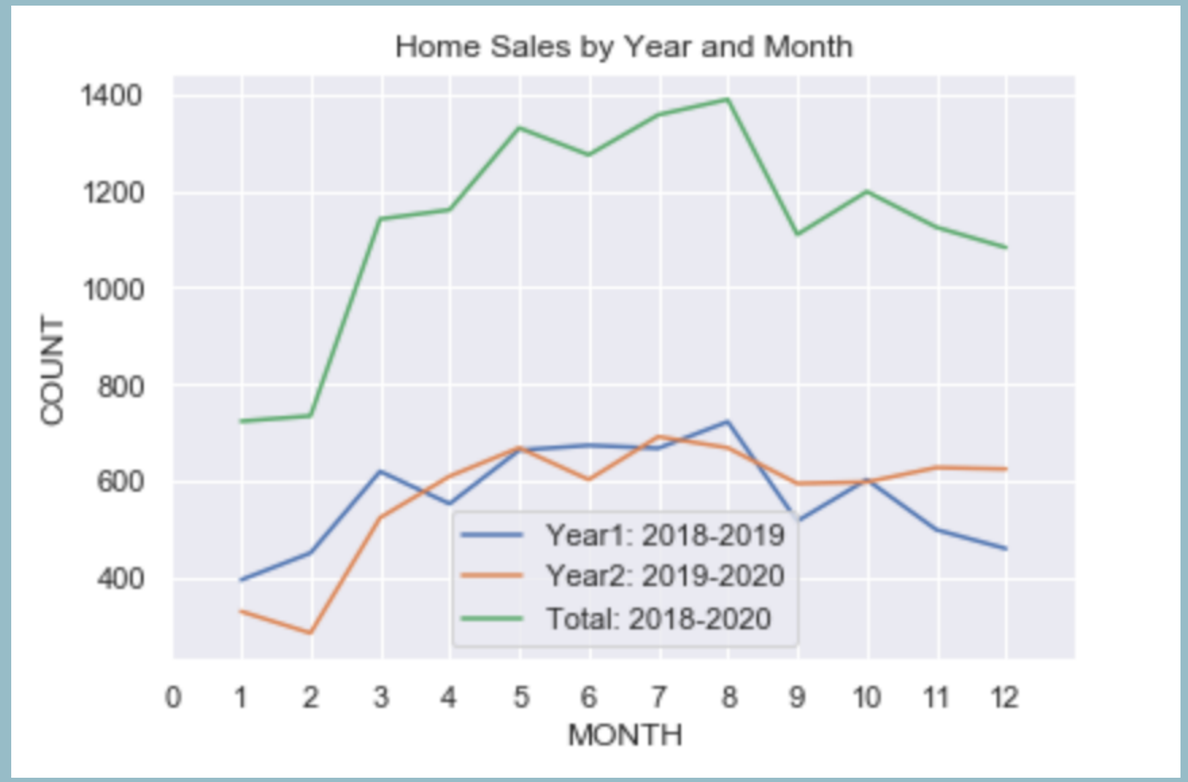




EXPLORATORY DATA ANALYSIS

TIMING OF SALES: SUMMER?

IS SUMMER A "HOT TIME" FOR HOME SALES?



BOOTSTRAP ANALYSIS
TIMING OF SALES:

IS THERE A DIFFERENCE IN AVERAGE HOME PRICE FOR SUMMER HOME SALES VS. NON-SUMMER SALES?*

Short Answer: YES

(We reject the null hypothesis)

*Summer = June, July, August; Non-Summer = All Other Months

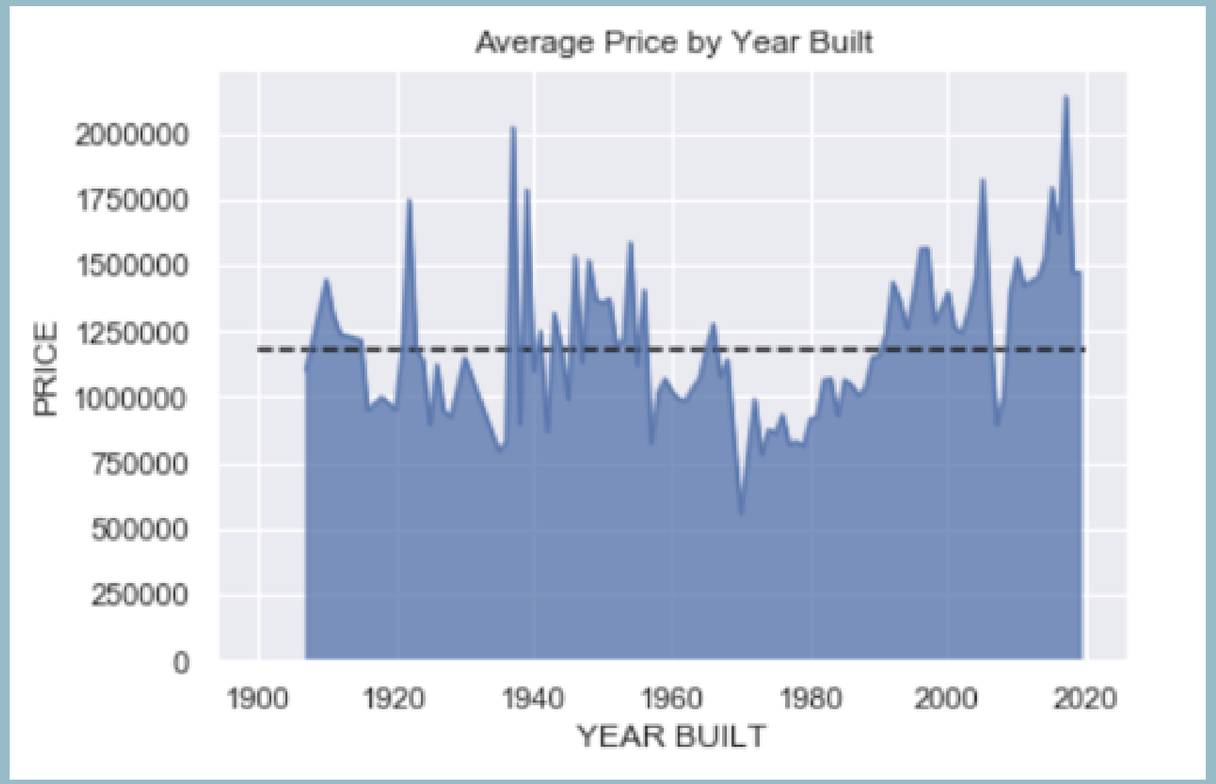
MEAN PRICE, SUMMER: $1,034,713.47
MEAN PRICE, NOT SUMMER: $976,924.60
DIFFERENCE IN MEAN PRICES: $57,788.87

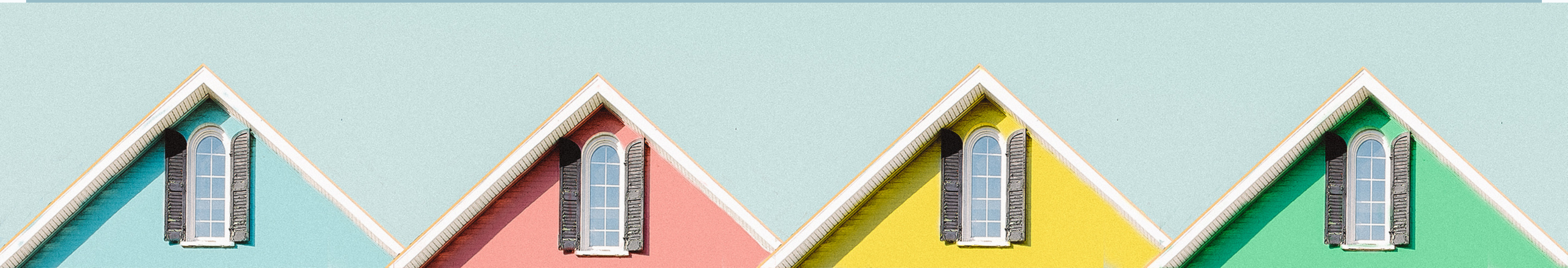
P-VALUE OF 0.002

EXPLORATORY DATA ANALYSIS

NEWER VS OLDER HOMES

ARE "NEWER" HOMES SELLING FOR HIGHER PRICES?





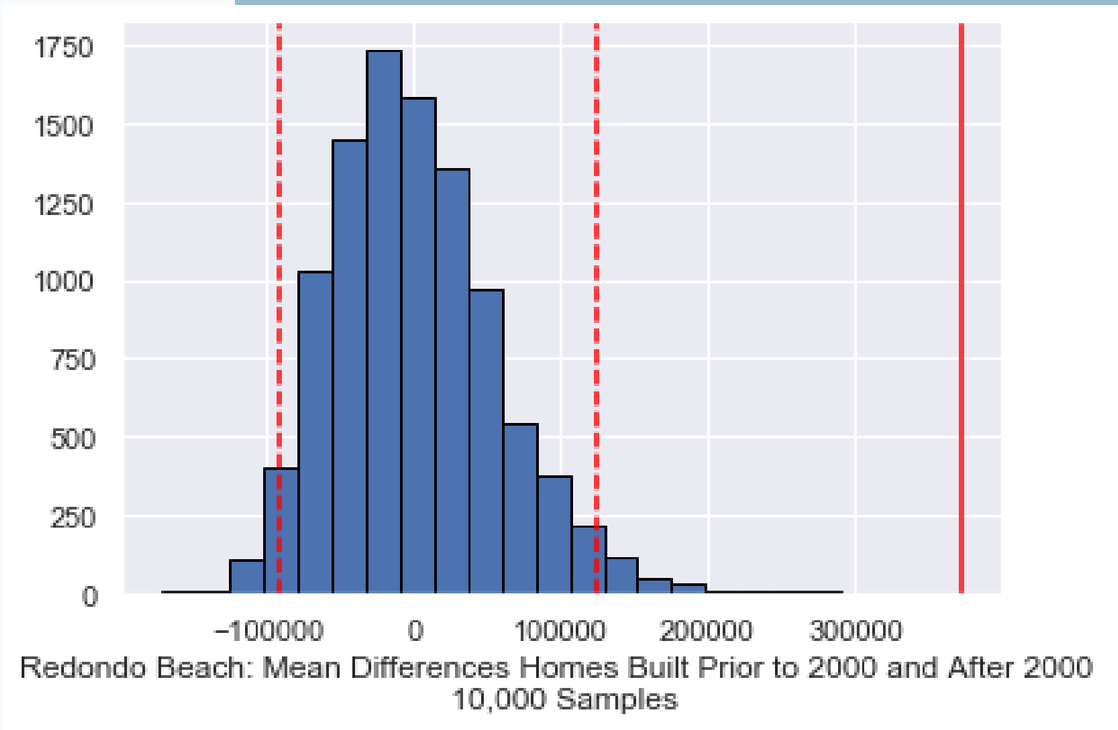
BOOTSTRAP ANALYSIS
NEWER VS OLDER HOMES:

IS THERE A DIFFERENCE IN AVERAGE HOME PRICE BETWEEN HOMES IN REDONDO BEACH* *BUILT PRIOR TO 2000 AND THOSE BUILT AFTER?*

Short Answer: ALSO YES

(We reject the null hypothesis)

*Redondo Beach is a neighborhood in South Bay with a large amount of homes sold in our dataset (n~1800). One neighborhood was chosen to control for the variability in neighborhoods.



MEAN PRICE, BEFORE 2000: $1,078,491.73
MEAN PRICE, AFTER 2000: $1,450,324.14
DIFFERENCE IN MEAN PRICES: $371,832.41

P-VALUE OF 0

IN-DEPTH ANALYSIS

MACHINE LEARNING APPROACH

> APPROACH: SUPERVISED LEARNING PROBLEM

- Labeled data (we have both features and already labeled prices corresponding to those features)

> METHOD: **RANDOM FOREST REGRESSION**

- Ensemble method expanding on a Decision Tree
- Predicting Prices Lends to a Regression Analysis (vs Classification)

> STEP: PREPROCESSING

- Dropping additional columns and One Hot Encoding
- Separate data into training and testing (70/30 split)

> STEP: RUN INITIAL MODEL

- Fit to training data -> Predict on Testing Data
- Mean Accuracy(Training): **0.97** | Mean Accuracy(Testing): **0.84**
- Possible overfitting

> STEP: PARAMETER TUNING & FEATURE SELECTION

- Tuned parameters: n_estimators & max_depth
- Selected features accounting for 92.5% cumulative importance

> STEP: RE-RUN MODEL

PARAMETER TUNING & FEATURE SELECTION

We attempt to tune certain parameters of the model and select features most important to attempt to get a better fit in our model.

1

N_ESTIMATORS

The parameter **n_estimators** is the number of trees to be used in the forest.

2

MAX_DEPTH

The parameter **max_depth** tells the model how far down the tree to go.

1

SQUARE FEET (#1)

The feature square feet was a large predictor, accounting for more than 50% of the prediction in the model.

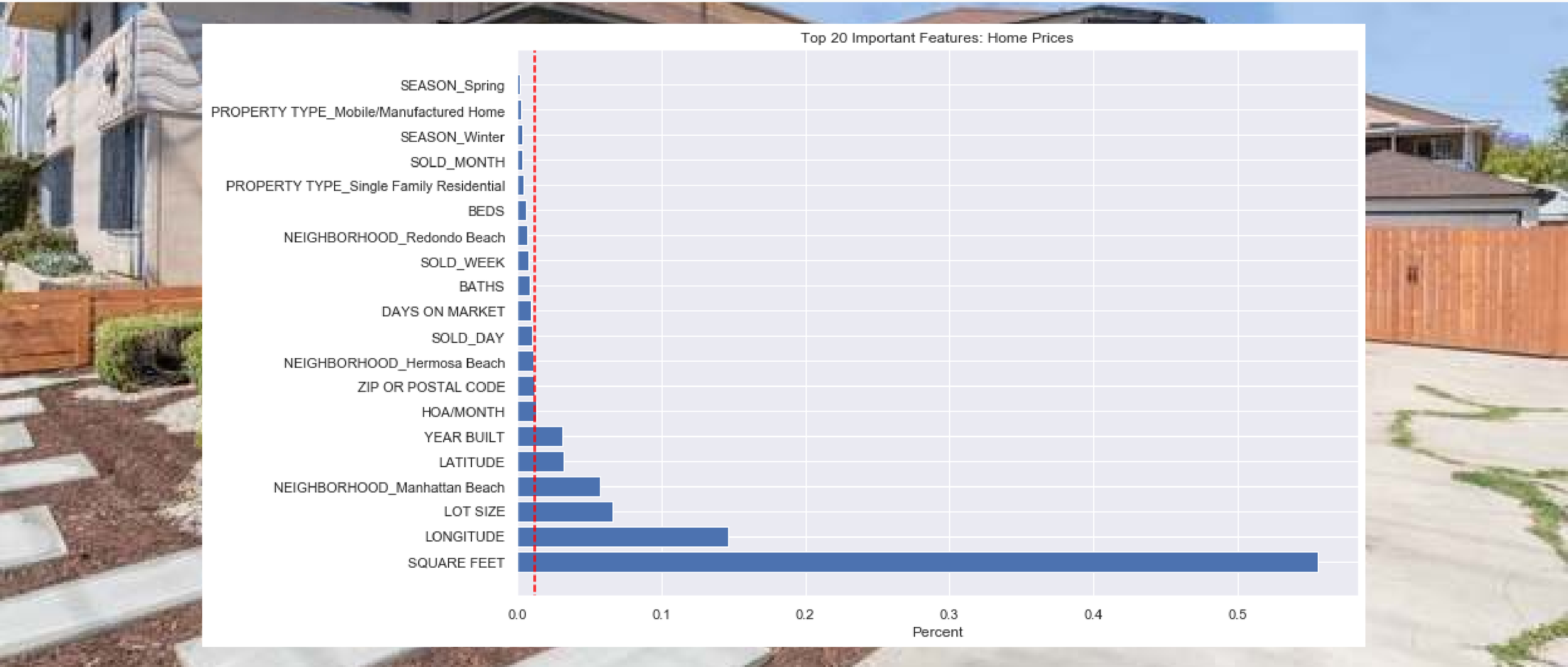
2

LONGITUDE (#2)

This feature accounted for 15% of the model. Longitude could account for how close (or far) a property is from the ocean.

TOP 20 IMPORTANT FEATURES

Of the 51 features that went into the model, listed below are the top 20. **The two most important features by far in predicting prices in our model are: square feet and longitude.** Other features include the year built, sold day, and whether the property is in one of a handful of high priced neighborhoods.

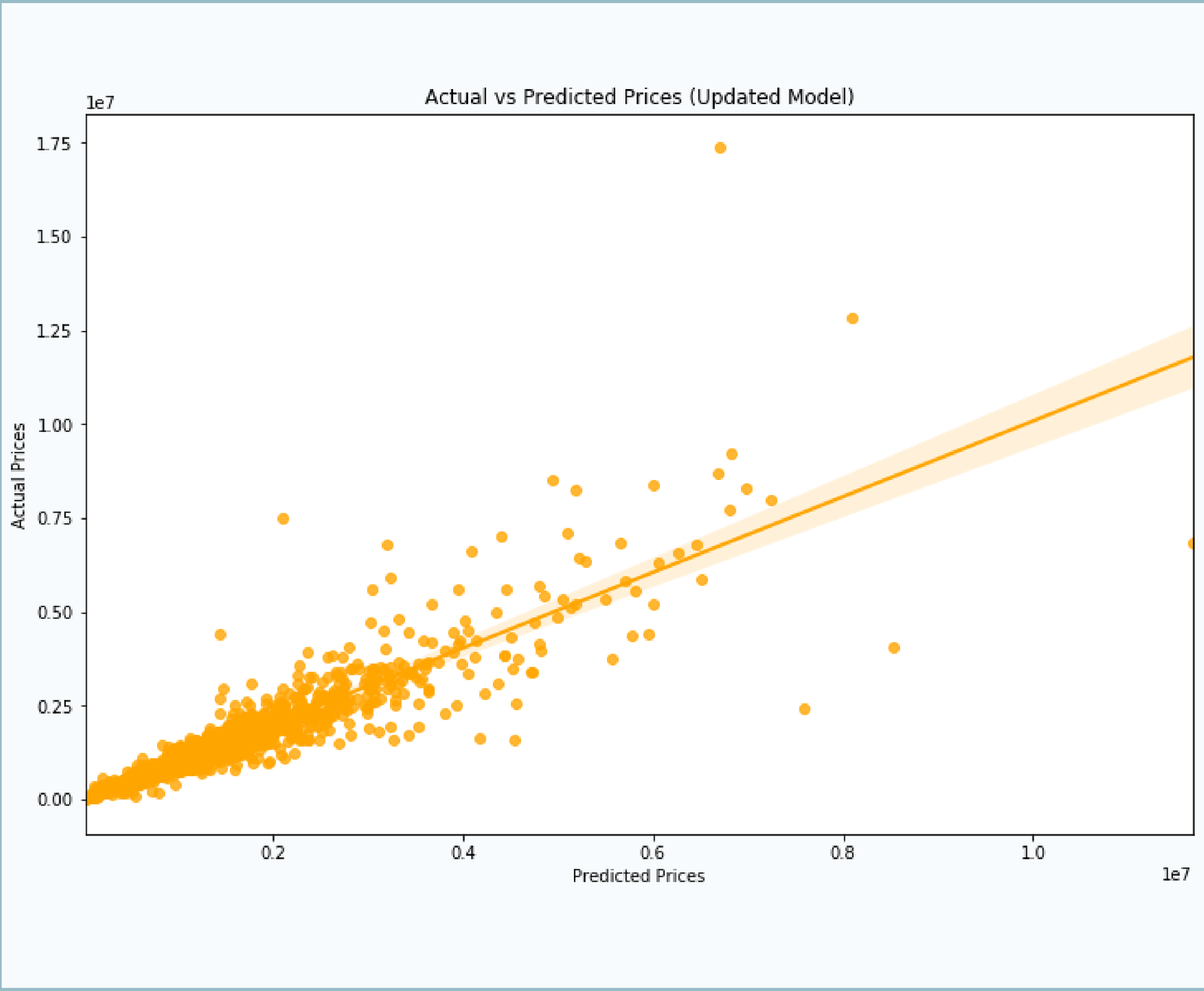


RESULTS

UPDATED MODEL

The Random Forest Regression model was rerun with the two tuned parameters, and the top 8 most important features (cumulative feature importance of 92.5%.).

R-SQUARED = 0.846252 (+ 0.006)

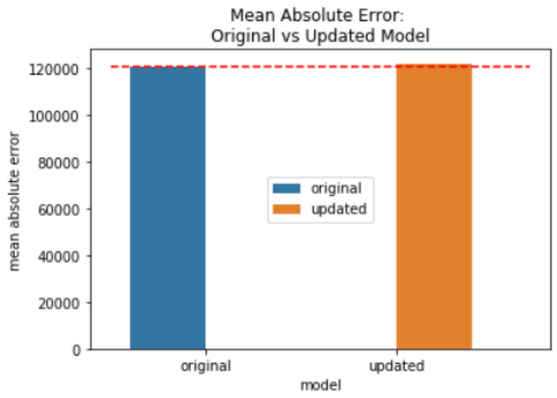


RESULTS

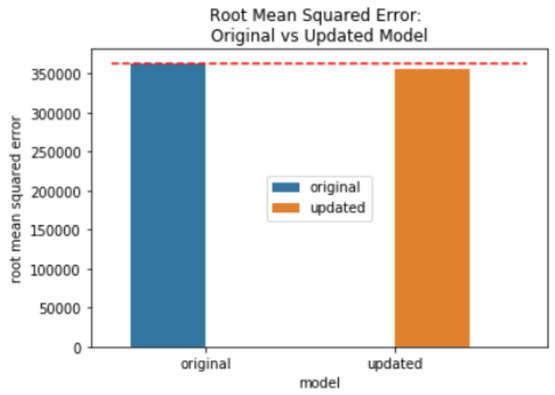
The updated model performs nearly identically to the original model.

	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

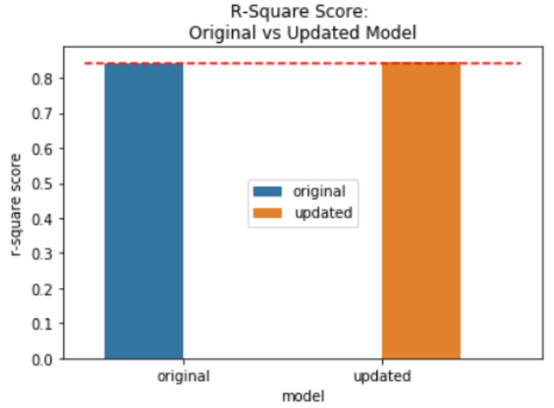
MAE



RMSE



R^2 Score



RESULTS

EXAMPLE PREDICTIONS

RANCHO PALOS VERDES: 3BR, 1B, 1084 SQ.FT.

PREDICTED: $753,800 | ACTUAL: $745,000

DIFFERENCE: $8,800

PLAYA DEL REY: 5BR, 4B, 4,590 SQ.FT.

PREDICTED: $2,857,484 | ACTUAL: $1,700,000

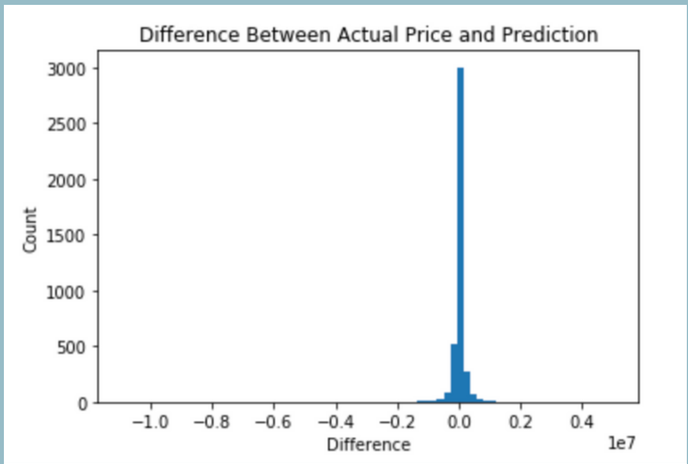
DIFFERENCE: $1,157,484

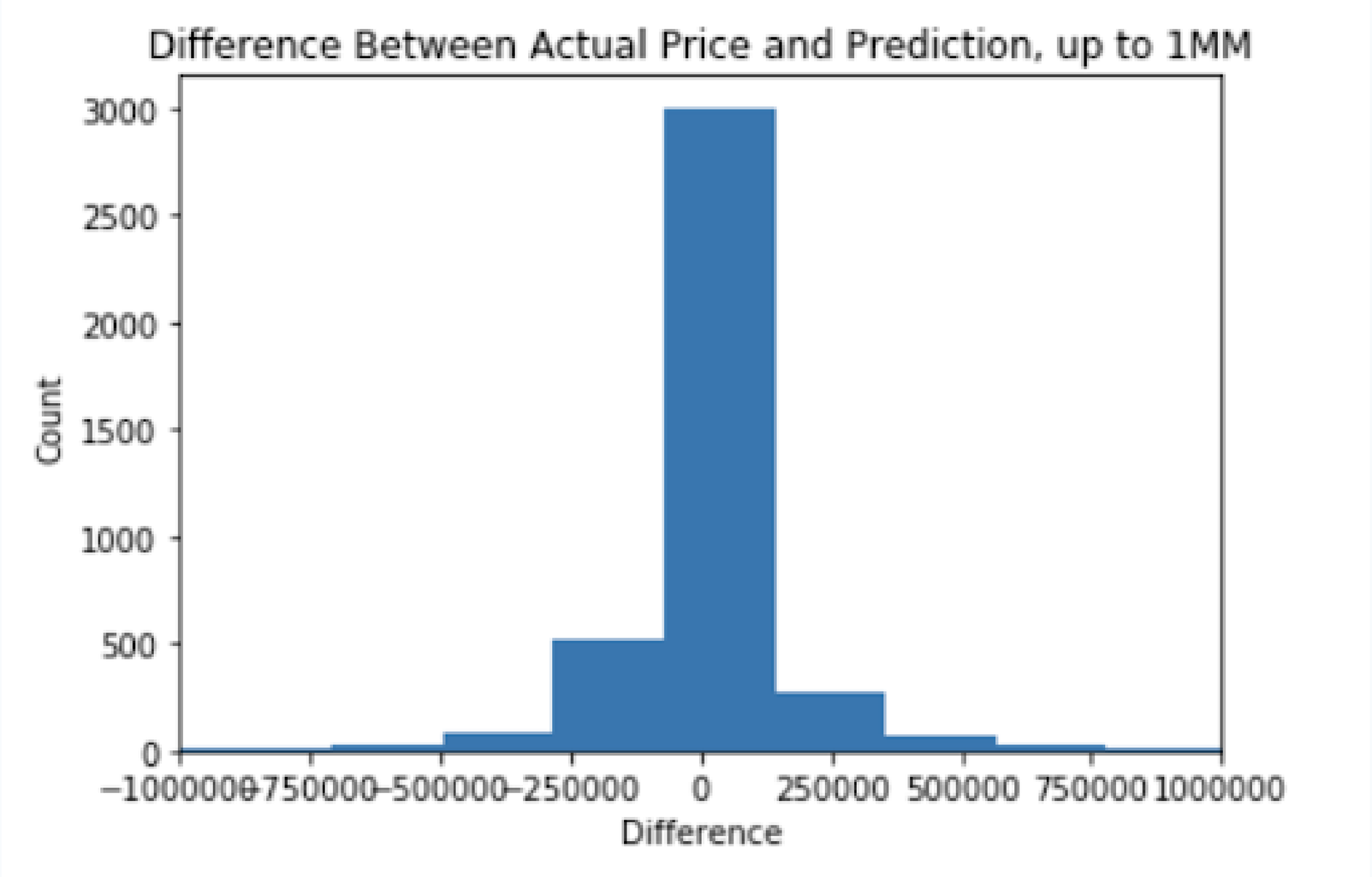
Images from Redfin.com Listings

RESULTS

HISTOGRAM OF DIFFERENCES

The model performs well at predicting more than half of the testing data within $50,000, but the performance deteriorates for higher priced homes. The image on the right is a zoomed in view of the histogram below.





RECOMMENDATIONS

The model achieved an R-squared score of 84.6%. We are also able to glean other insights about home prices in South Bay.

SPACE MATTERS

The most important predictor for home price in the model was Square Feet. Although “Tiny Homes” are getting more prominence in pop culture, homes providing more space still seem to be selling for higher prices. Builders may consider this when planning new builds.

LOCATION MATTERS

The second largest predictor in the model was longitude. In other words, this might be indicative of how close or far a property is from the beach. Home buyers may consider the proximity to the ocean when pricing out a potential home.

TIMING MATTERS

Although some say it's summer all year long in CA, sellers may want to consider the timing of their sale. There appeared to be a difference between homes sold in summer and homes sold in other seasons, with summer sales having a higher average price than other months.

FUTURE CONSIDERATIONS

Adding **different features** to the model: i.e. proximity to certain amenities (stores, parks, etc.), or air conditioning.

Looking further into **what constitutes a "Sale."** From the data, more homes tend to sell on Fridays. Why?

Controlling for **confounding variables**, **distribution, and outliers** may help improve the results of the model.

Looking at **tuning additional parameters**, or possibly looking at a **different algorithm** to get a better fit.

THANK YOU

GET IN TOUCH

LAUREN BROUSSARD

linkedin.com/in/laurenbroussard