

# Predicting Home Prices in Los Angeles



Elizabeth Naameh,  
VP of Holdings, Redfin

# Design

**Client:** Urban housing markets were particularly disrupted by the Covid pandemic. For investors, this presents an opportunity to capitalize in a new market environment. Redfin wants to understand the LA housing market so that it invest in properties.

**Objective:** Explore whether the sale price of a home can be modeled against other housing/geographic features.

**Goal:** Produce a regression model that can best interpret a relationship for sale price and a model that can best predict home sale price in Los Angeles.

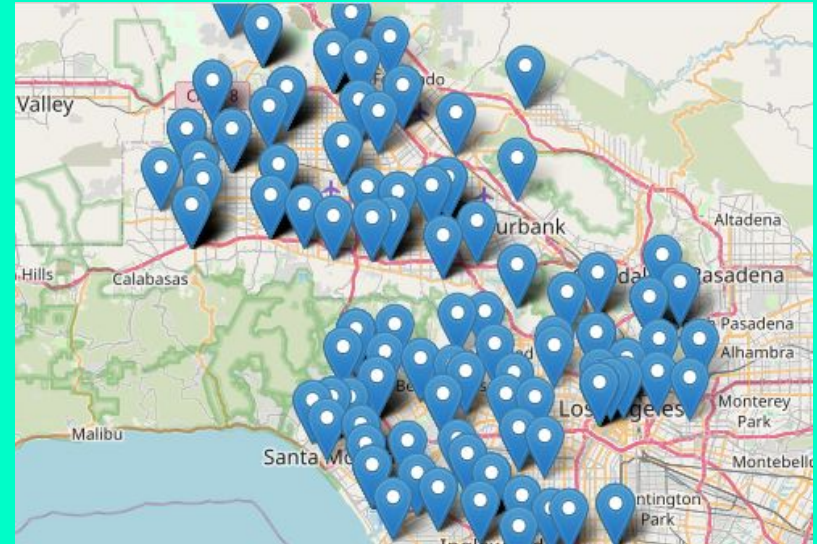
Home data was scraped from **Zillow** and geographic socio-economic data was scraped from **City-Data.com** by zip code.

Each row represents a 'Recently Sold' property in the Los Angeles area.

Of 600+ home sales scraped, 288 were used in our final analysis. We excluded manufactured homes, lots, multi-family residences.

- BeautifulSoup & Selenium
- Numpy & Pandas
- Scikit-learn & Statsmodels
- Matplotlib & Seaborn
- Folium, Geopandas & Geopy

# Data

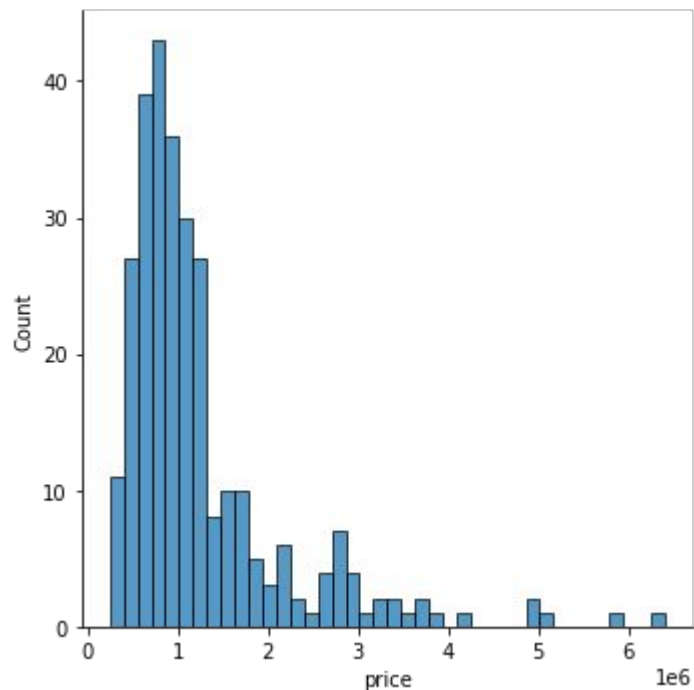


# Data Cleaning & EDA

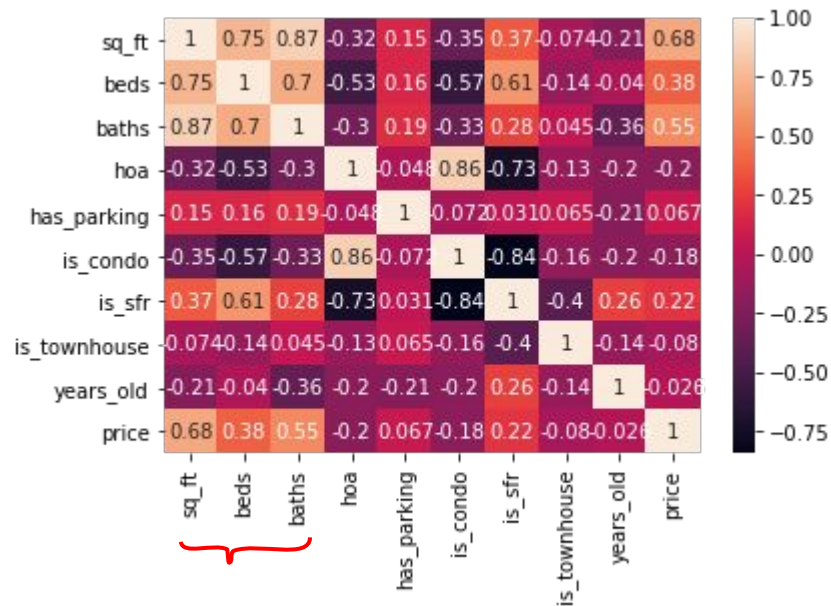
- Delete duplicates and drop rows with null *'prices'*, *'addresses'*
- Binarize categorical data
  - New columns include *'has\_parking'*, *'is\_condo'*, *'is\_sfr'*, *'is\_townhouse'*
- Restrict sold listings to prior three weeks
- Impute missing *'hoa'* (0) and *'beds'* (0, indicates studio)

	address	price	sq_ft	date_sold	beds	baths	year_built	hoa	parking	type_
0	1135 W 185th St, Gardena, CA 90248	800000.0	1866.0	07/07/21	3.0	3.0	1954.0	NaN	2 Attached Garage spaces	Single Family Residence
1	9631 Compton Ave, Los Angeles, CA 90002	610000.0	1522.0	07/07/21	4.0	3.0	2021.0	NaN	2 Attached Garage spaces	Single Family Residence
2	8701 Delgany Ave UNIT 304, Los Angeles, CA 90293	1125000.0	1785.0	07/07/21	3.0	3.0	1964.0	460.0	2 Garage spaces	Condominium
3	906 Parkman Ave, Los Angeles, CA 90026	1455000.0	NaN	07/07/21	NaN	NaN	1937.0	NaN	Carport	Multi Family

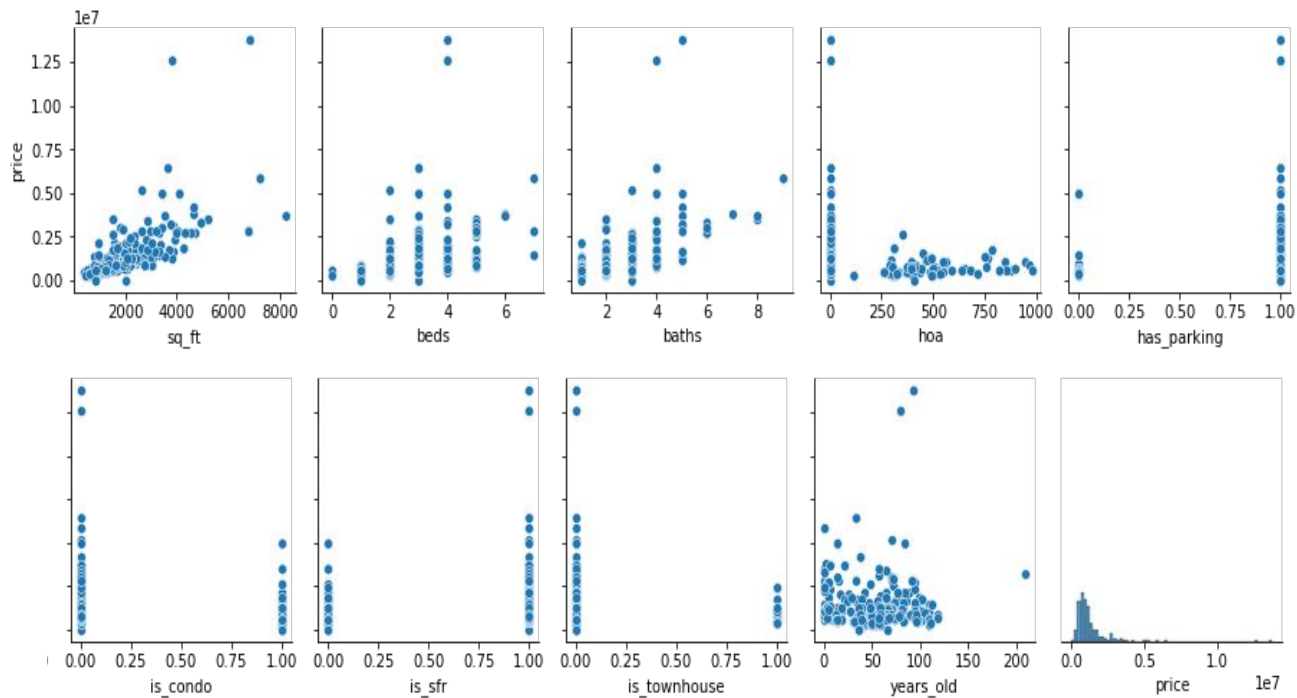
## Distribution of Home Sale Price



## Correlation Matrix for Home Features



# Baseline Model



OLS

Train: 0.579

Val: 0.235

# Import Geographic Data

- Scrape socio-economic data by Zip Code
  - *'median\_resident\_age'*
  - *'avg\_household\_size'*
  - *'avg\_household\_income'*
  - *'pct\_poverty'*
  - *'pct\_bachelors'*
- Merge datasets on zip\_code



	zip_code	median_age	avg_household	median_income	pct_poverty	pct_bachelors	url
0	90248	43.5	3.0	64253.0	9.7	27.9	<a href="http://www.city-data.com/zips/90248.html">http://www.city-data.com/zips/90248.html</a>
1	90002	28.3	4.1	38987.0	28.8	6.1	<a href="http://www.city-data.com/zips/90002.html">http://www.city-data.com/zips/90002.html</a>
2	90293	38.3	1.9	110698.0	5.8	74.7	<a href="http://www.city-data.com/zips/90293.html">http://www.city-data.com/zips/90293.html</a>

# Feature Selection

Drop features with high multicollinearity and low correlation.

- *'hoa'*
- *'median\_age'*
- *'has\_parking'*
- *'is\_townhouse'*

Improvement on baseline:

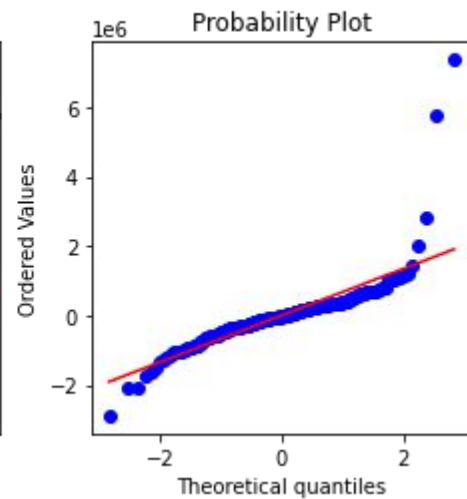
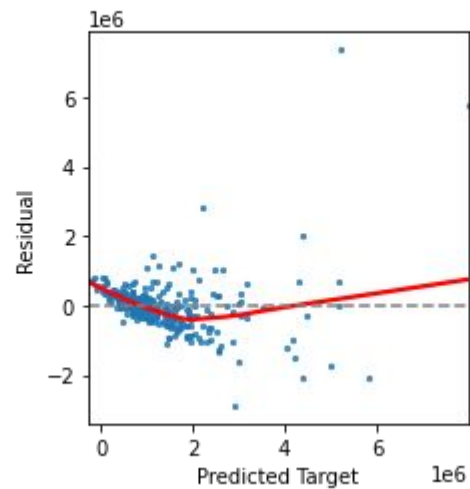
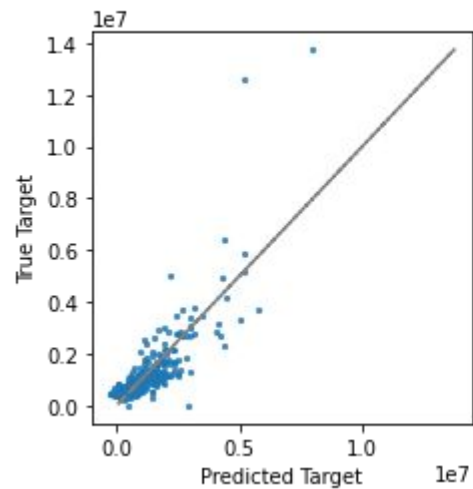
```
OLS
Train:  0.693
Val:    0.356
```

## Variance Inflation Factor

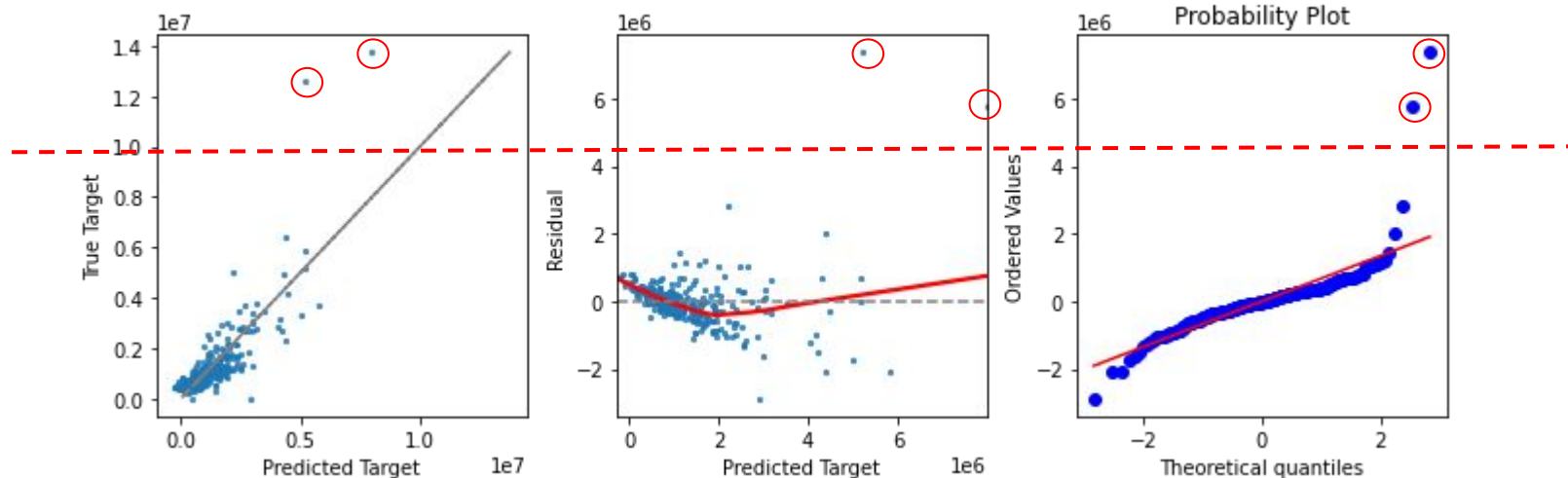
sq_ft	20.85
beds	28.54
baths	27.67
is_condo	4.95
is_sfr	13.49
years_old	5.68
avg_household	37.07
median_income	47.67
pct_poverty	10.10
pct_bachelors	34.73
dtype:	float64



### Diagnostic Plots



Diagnostic Plots



	address	y_pred	y_actual	residual
181	1650 Amalfi Dr, Pacific Palisades, CA 90272, USA	7.983059e+06	13750000.0	5.766941e+06
157	992 Napoli Dr, Pacific Palisades, CA 90272, USA	5.225908e+06	12600000.0	7.374092e+06
243	1002 Alta Ave, Santa Monica, CA 90402, USA	4.368953e+06	6402725.0	2.033772e+06



[Save](#) [Share](#) [More](#)

4 bd | 4 ba | 3,802 sqft

992 Napoli Dr, Pacific Palisades, CA 90272

● **Sold: \$12,600,000** | Sold on 06/30/21 | Zestimate®: \$12,666,900

Est. refi payment: \$54,449/mo [\\$ Refinance your loan](#)

[Home value](#) [Owner tools](#) [Home details](#) [Neighborhood details](#) [Similar homes](#)



[Save](#) [Share](#) [More](#)

4 bd | 5 ba | 6,800 sqft

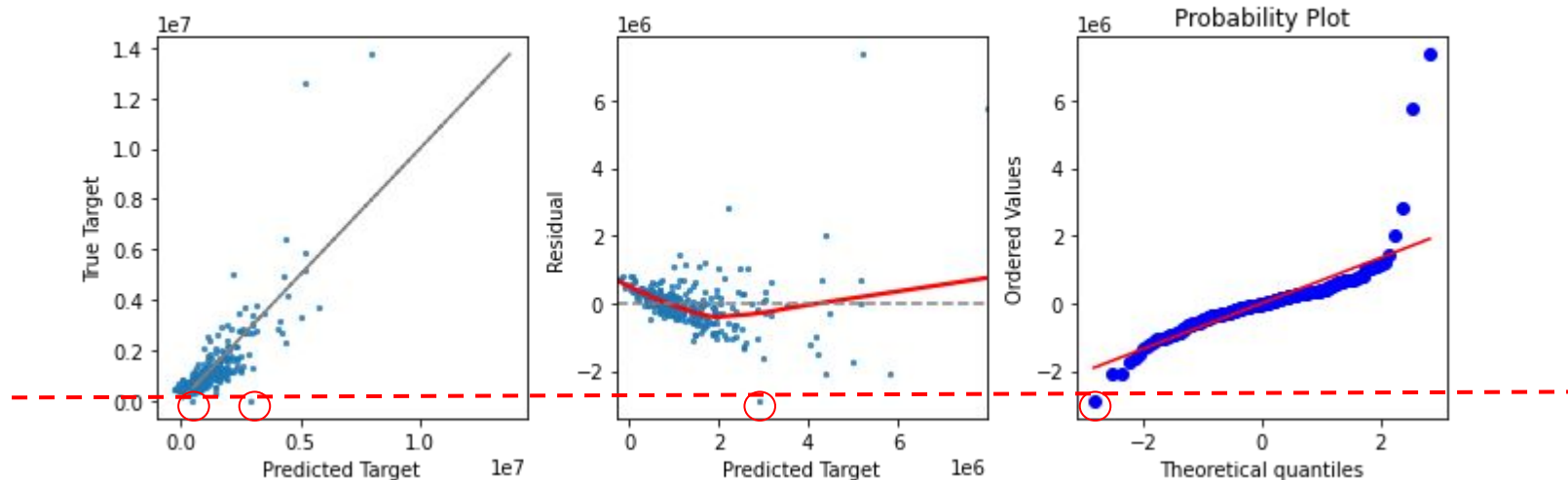
1650 Amalfi Dr, Pacific Palisades, CA 90272

● **Sold: \$13,750,000** | Sold on 06/29/21 | Zestimate®: \$13,832,200

Est. refi payment: \$59,391/mo [\\$ Refinance your loan](#)

[Home value](#) [Owner tools](#) [Home details](#) [Neighborhood details](#) [Similar homes](#)

Diagnostic Plots



	address	y_pred	y_actual	residual
<b>287</b>	222 S Central Ave APT 238, Los Angeles, CA 900...	4.794929e+05	2000.0	-4.774929e+05
<b>91</b>	2201 Coldwater Canyon Dr, Beverly Hills, CA 90...	2.890890e+06	7100.0	-2.883790e+06
<b>201</b>	5460 White Oak Ave UNIT A227, Encino, CA 91316...	3.074047e+05	242000.0	-6.540468e+04

# Improving on Baseline Model

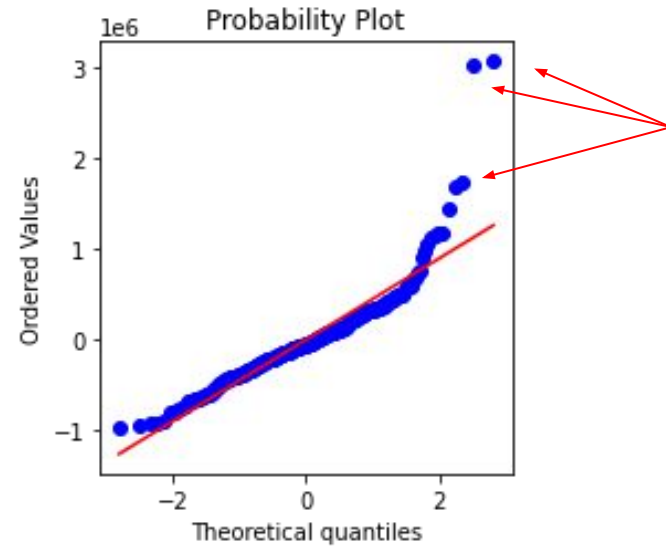
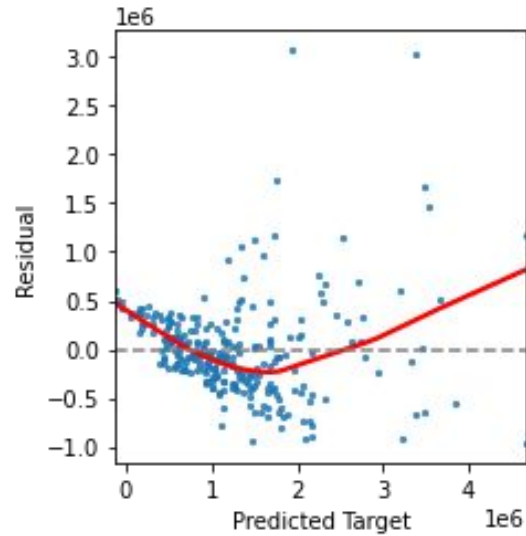
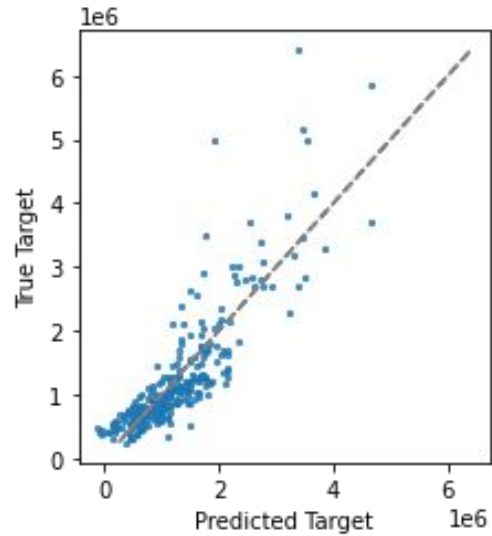
Drop homes with sale prices above \$10 million or below \$10,000.

Drop non-significant features:  
'avg\_household', 'is\_condo'.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.731	
	coef	std err	t	P> t
const	-1.865e+06	2.47e+05	-7.550	0.000
sq_ft	500.4162	62.459	8.012	0.000
beds	-1.742e+05	4.76e+04	-3.661	0.000
baths	1.403e+05	5.13e+04	2.735	0.007
is_sfr	5.299e+05	9.4e+04	5.640	0.000
years_old	2525.3727	1122.636	2.250	0.025
median_income	8.7311	1.965	4.444	0.000
pct_poverty	3.677e+04	7373.027	4.988	0.000
pct_bachelors	1.449e+04	2743.429	5.280	0.000

### Diagnostic Plots

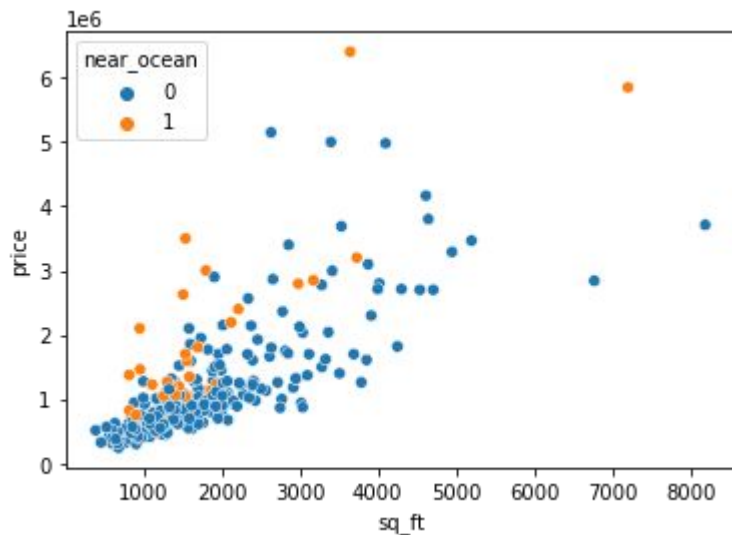


We can do better!

# Feature Engineering

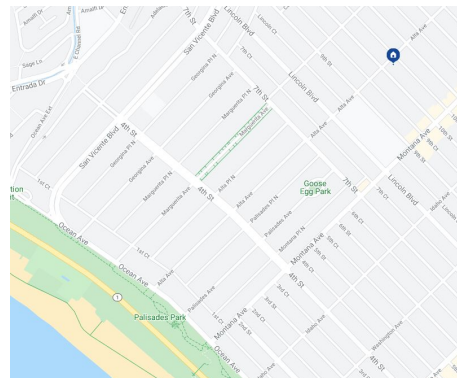
We continue to underperform for unique properties, e.g. sweeping downtown views, beachfront locations.

Create new variable, ``near_ocean``



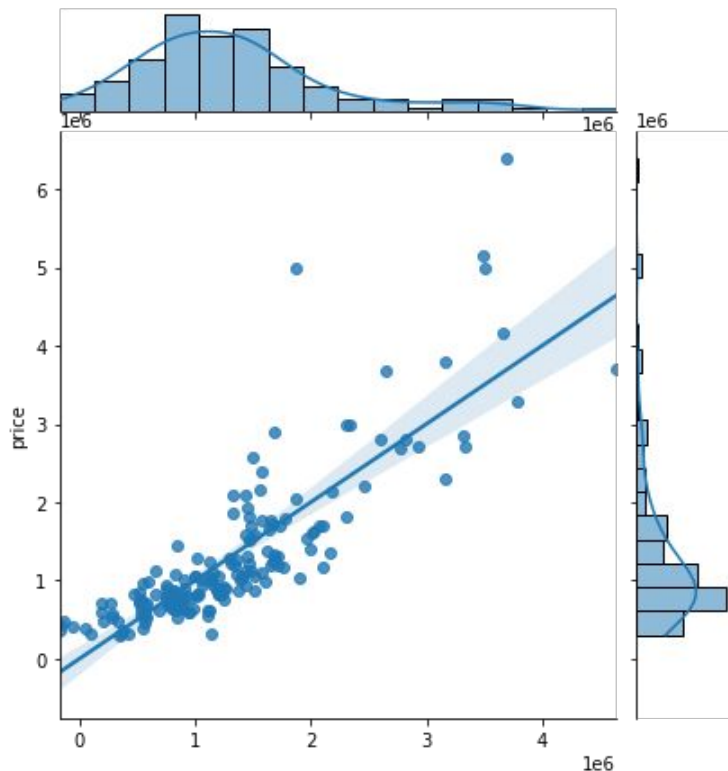
OLS  
Train: 0.709  
Val: 0.857

RMSE: 530052.544  
MAE: 355690.896





# Final Interpretive Model



## Repeated Cross Validation Results:

Simple mean cv  $r^2$ :  $-0.776 \pm 0.033$  ←

Ridge mean cv  $r^2$ :  $-0.776 \pm 0.033$

	coef	P> t
<b>const</b>	$-1.738e+06$	0.000
<b>sq_ft</b>	528.1412	0.000
<b>beds</b>	$-1.771e+05$	0.000
<b>baths</b>	$1.205e+05$	0.019
<b>is_sfr</b>	$5.685e+05$	0.000
<b>years_old</b>	2082.2522	0.063
<b>median_income</b>	7.7889	0.000
<b>pct_poverty</b>	$3.491e+04$	0.000
<b>pct_bachelors</b>	$1.32e+04$	0.000
<b>near_ocean</b>	$2.955e+05$	0.003



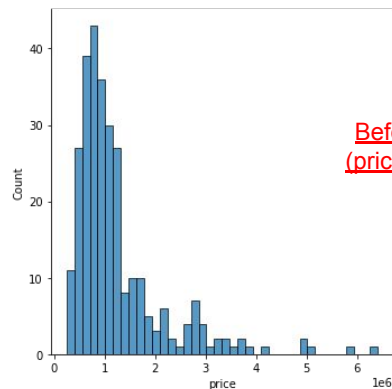
# Final Predictive Model

We apply a  $\log_2$  transformation to our target, home sale price.

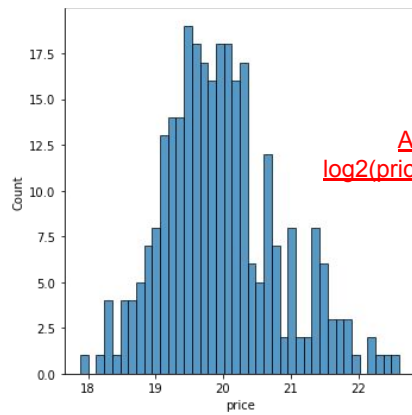
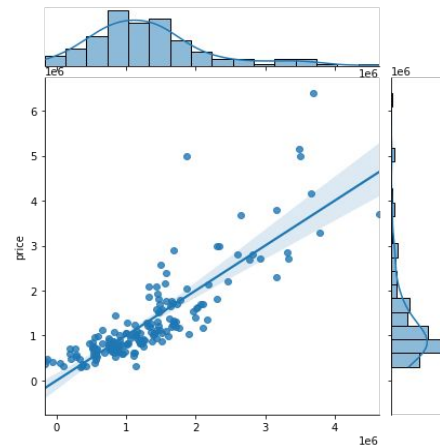
- + Residuals display constant variance
- Model loses interpretability

Linear Regression val  $R^2$ : 0.828  
Ridge Regression val  $R^2$ : 0.829  
Degree 2 polynomial regression val  $R^2$ : 0.697

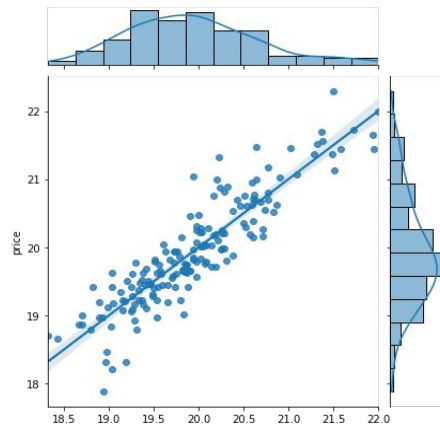
OLS with  $\log_2(\text{price})$   
Train: 0.815  
Val: 0.828  
Test: 0.77



Before  
(prices)



After  
log2(prices)



# Future Work

To improve our model, include additional features:

- Lot size
- Has view or Floor Number (if condo)
- Proximity to ocean (miles)

To improve business relevance:

- Predict rate of return for investment properties
- Model sale price vs. predicted monthly rent

