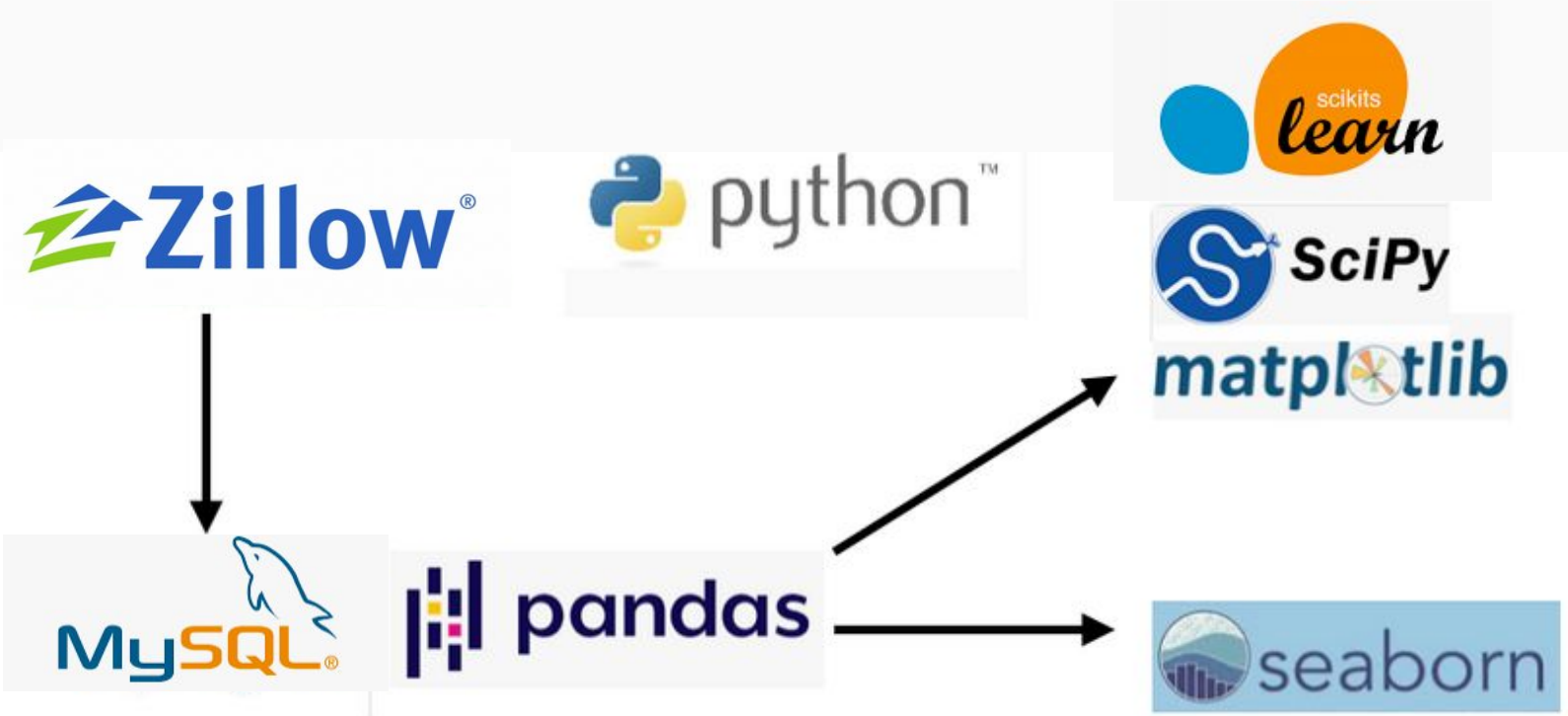
Estimating Home Values

September 15, 2021

A presentation by Steven J. Newton


Zillow
python
scikits
learn
SciPy
matplotlib
MySQL
pandas
seaborn

How outliers are defined has a large impact on evaluation stage

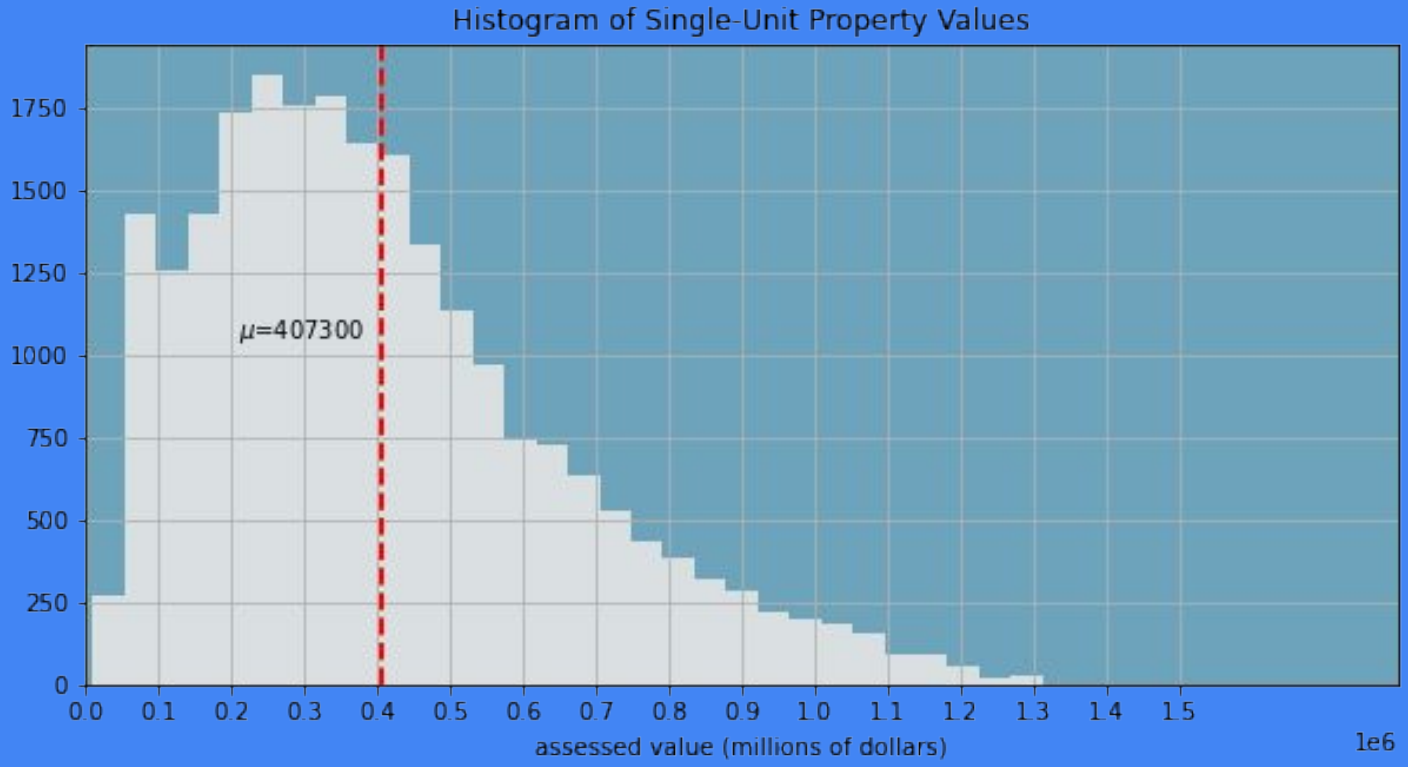
Incomplete cases: any property missing either f**inished sqft.** or **tax value** → **114**

Outlier: a property with **tax value** or **tax amount** at least **2*IQR** beyond Q1 or Q3 → **5,132**

Single-unit properties sold May-Aug. 2017 in dataset → **33,376**

Range of tax values (**target variable**) in training set → **$11k - $1.31M**

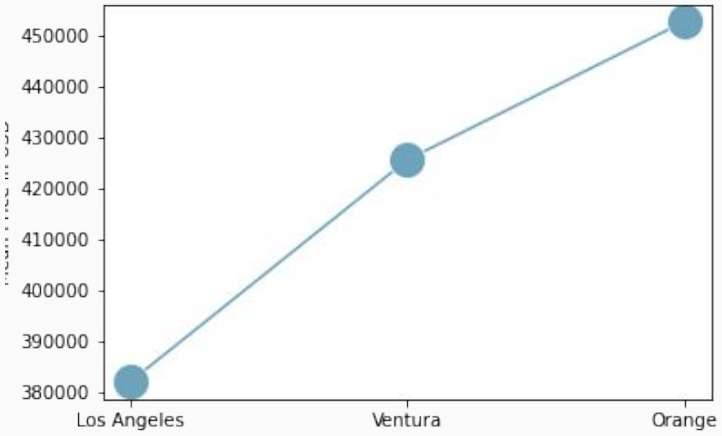
A Right-Skewed Target Variable



Mann-Whitney U and Pearson Correlation Tests

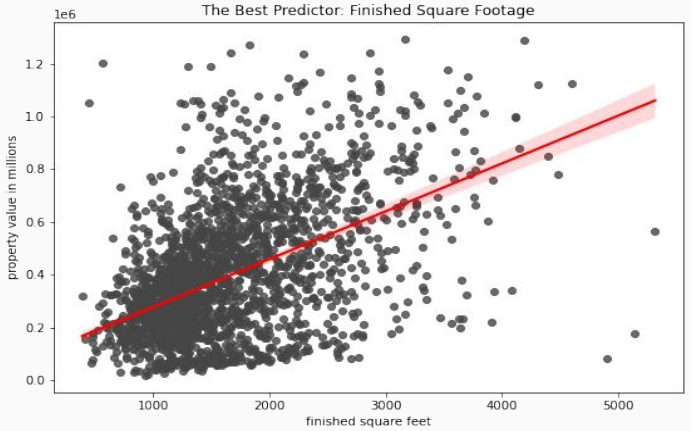
Is value related to county?

Yes, across all counties, p < .01.



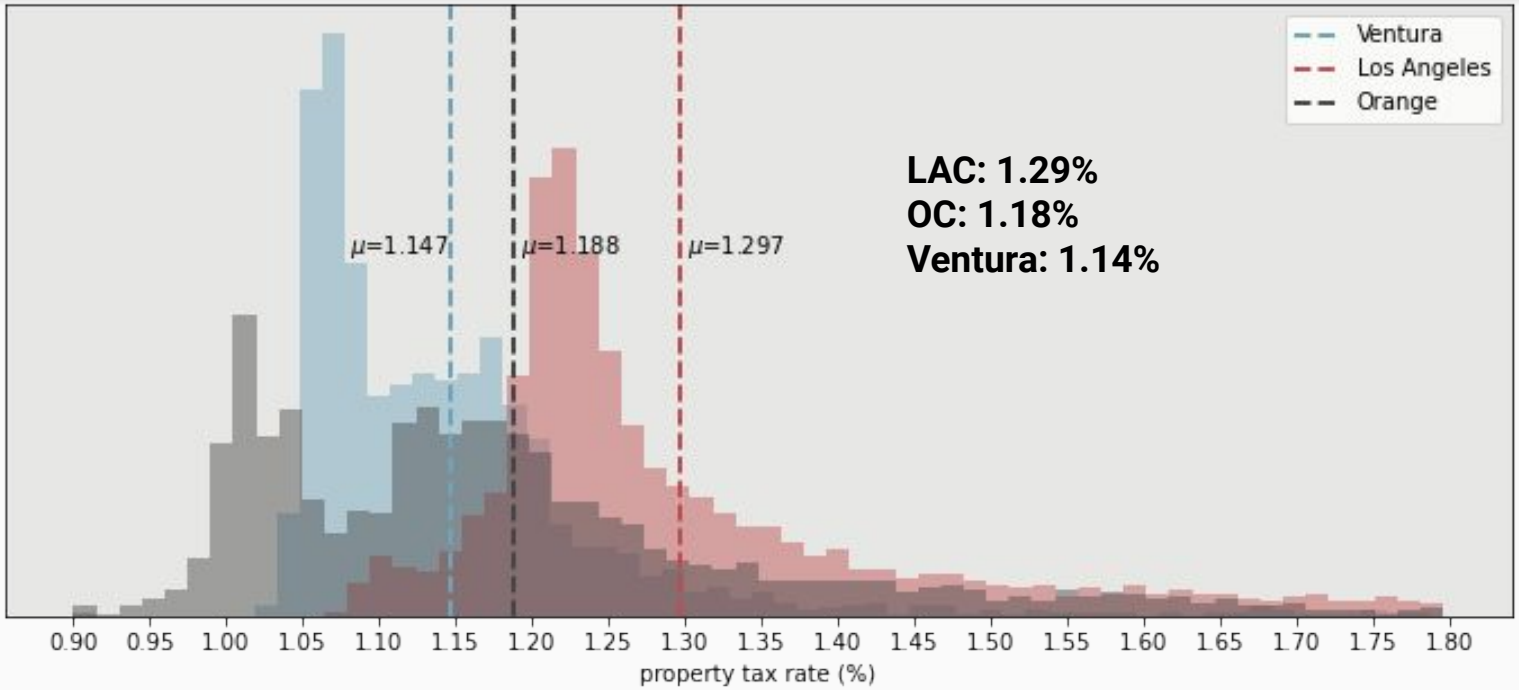
What numerical variables are strong predictors of the target?

Finished sqft. [corr.=.51] & **bathrooms** [corr.=.39]



Note heteroskedasticity

Property Tax Rates by County



Model Evaluation

metric/model	Mean	MVP	LinearReg	**LassoLars**
RMSE	245000	211000	206000	190000
MAPE	.90	.76	.73	.66
R^2	0	.26	.29	.40

RMSE: root mean squared error; units are same as target variable (USD)

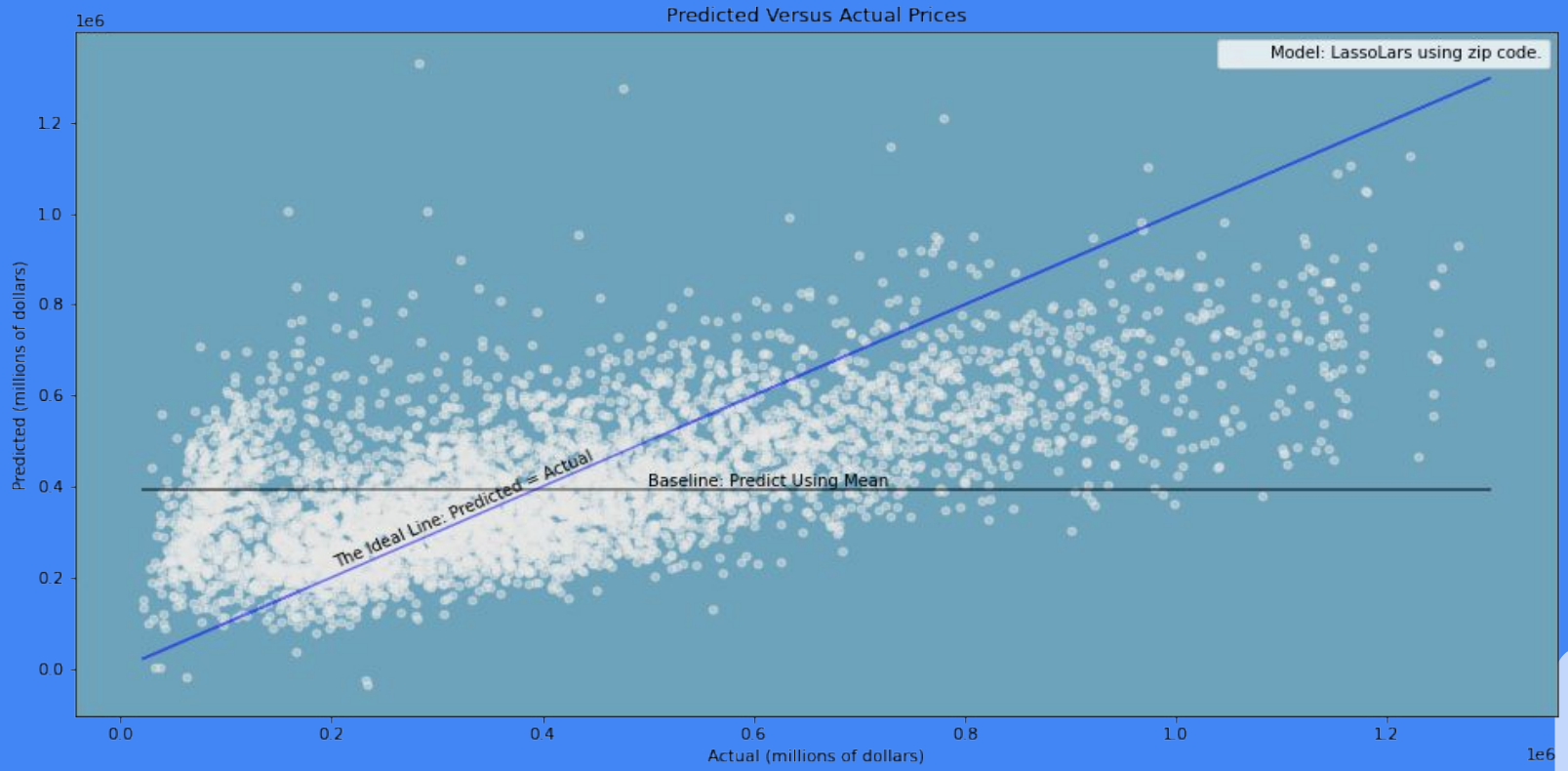
MAPE: mean absolute percentage error; from the Zillow Prize FAQ:

> "Home sale prices have a right-skewed distribution and are also strongly heteroscedastic so we need to use a relative error metric instead of an absolute metric to ensure valuation models are not biased towards expensive homes." Additional usage note: MAPE puts a heavier penalty on negative errors and is therefore biased towards a method whose forecasts are too low.

R^2: proportion of variance in target variable explained by model

	LassoLars on test set
RMSE	**188000**
MAPE	**.63**
R^2	**.40**

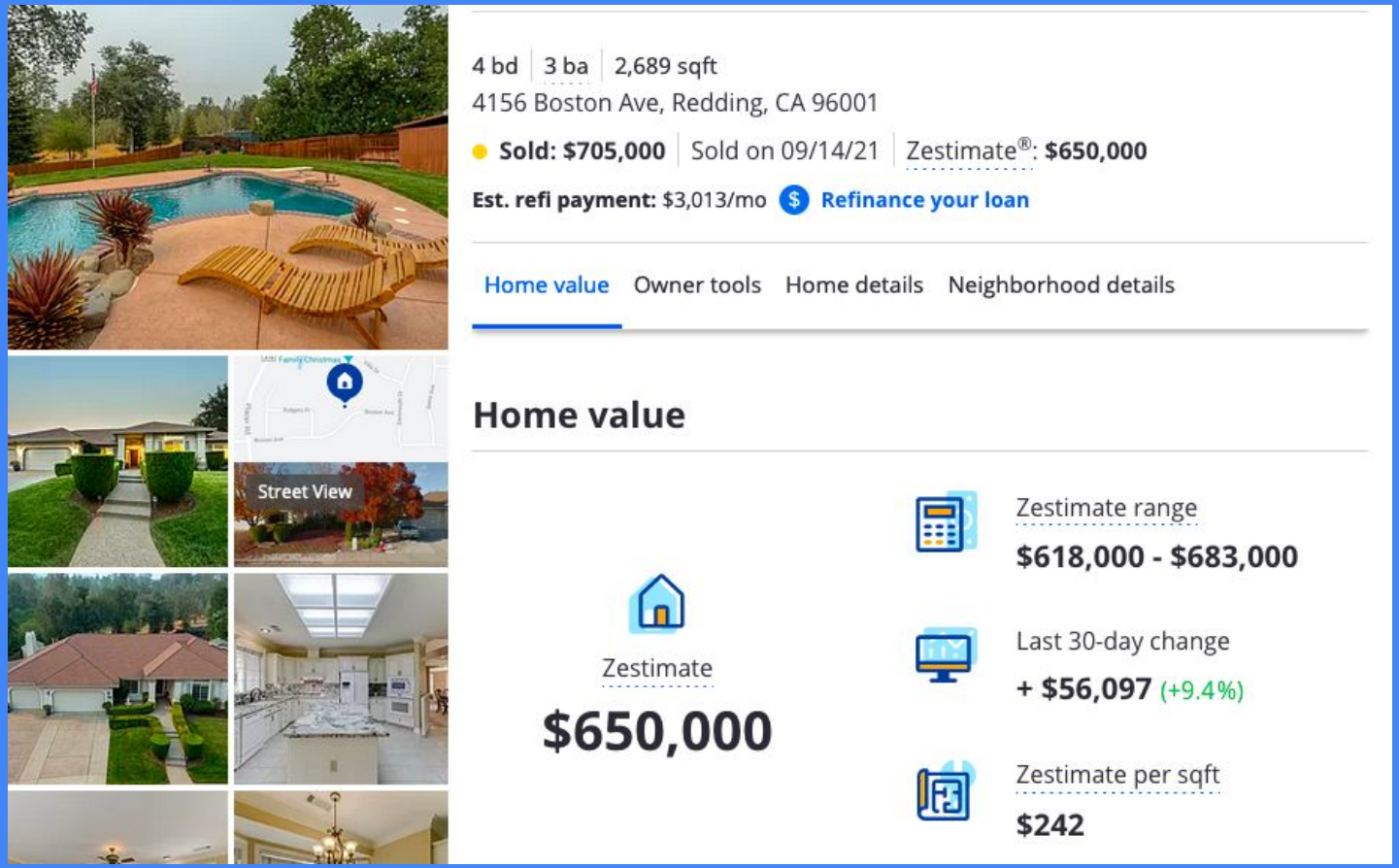
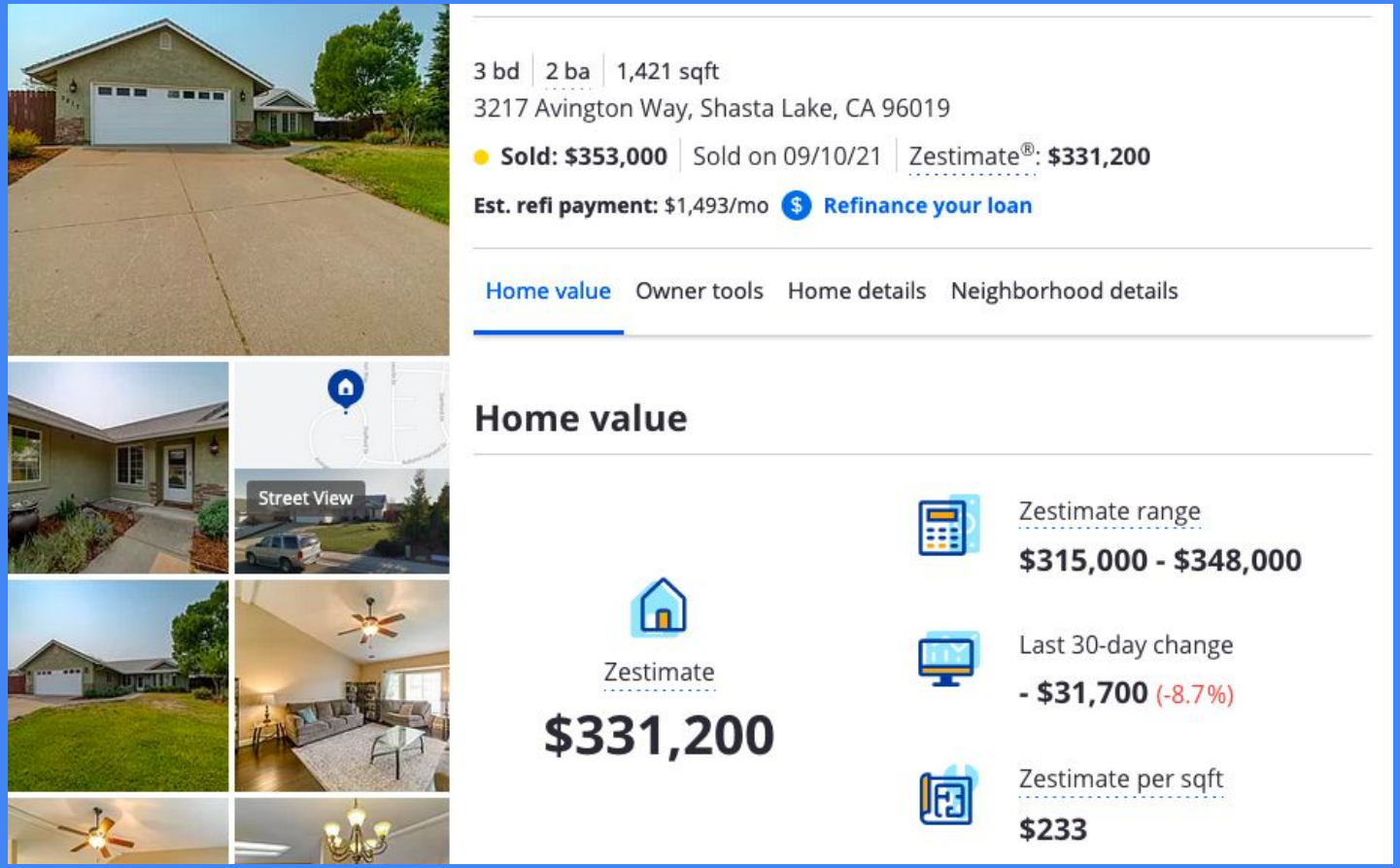
Residuals from LASSO+LARS: Is this model valid?



Where are these houses anyway?

Latitude-longitude and FIPS data placed the houses in three counties: **Orange, Ventura,** and **Los Angeles**. But the same houses' zip codes placed them in **Northern California** or elsewhere. Such a discrepancy becomes a show-stopper when testing on real data.

Flask App Example 1



Flask App Example 2