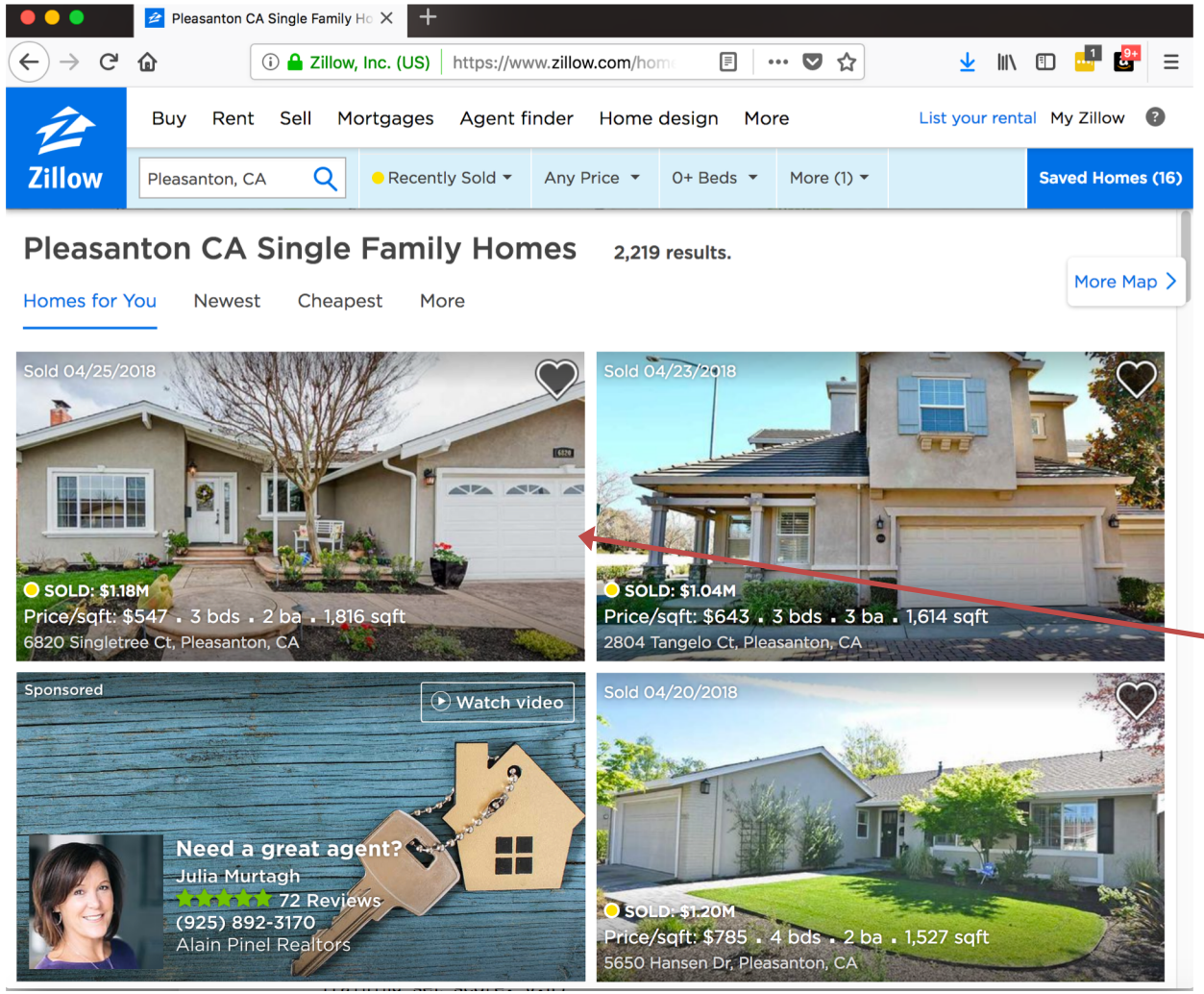
Project Luther: Predicting House Prices in Pleasanton, California

Benjamin Sturm

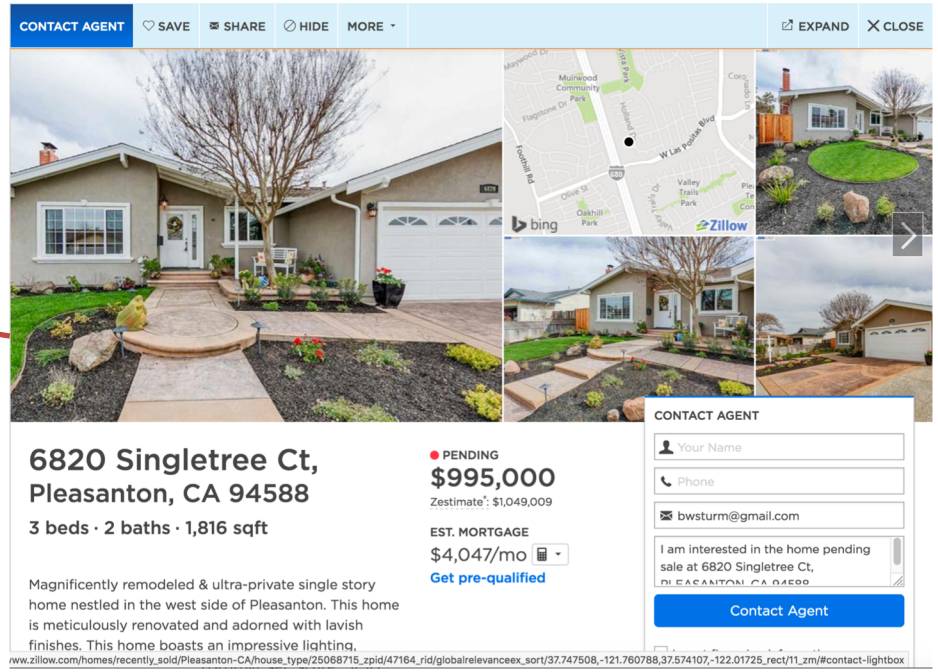
April 27, 2018

Zillow house price data was scraped using Selenium and Beautiful Soup



Data that was scraped:

Address, City, Zip, Bedrooms, Bathrooms,
Floor Size, Lot Size, Year Built, Sale Date, Sale Price



How accurately can I predict house prices for recently sold homes?

Possible uses for this model:

To help the buyer determine if they are getting a fair price on a particular home

To help the seller when listing a home

After scraping, I had data from 500 houses

```
houses_df.head()
```

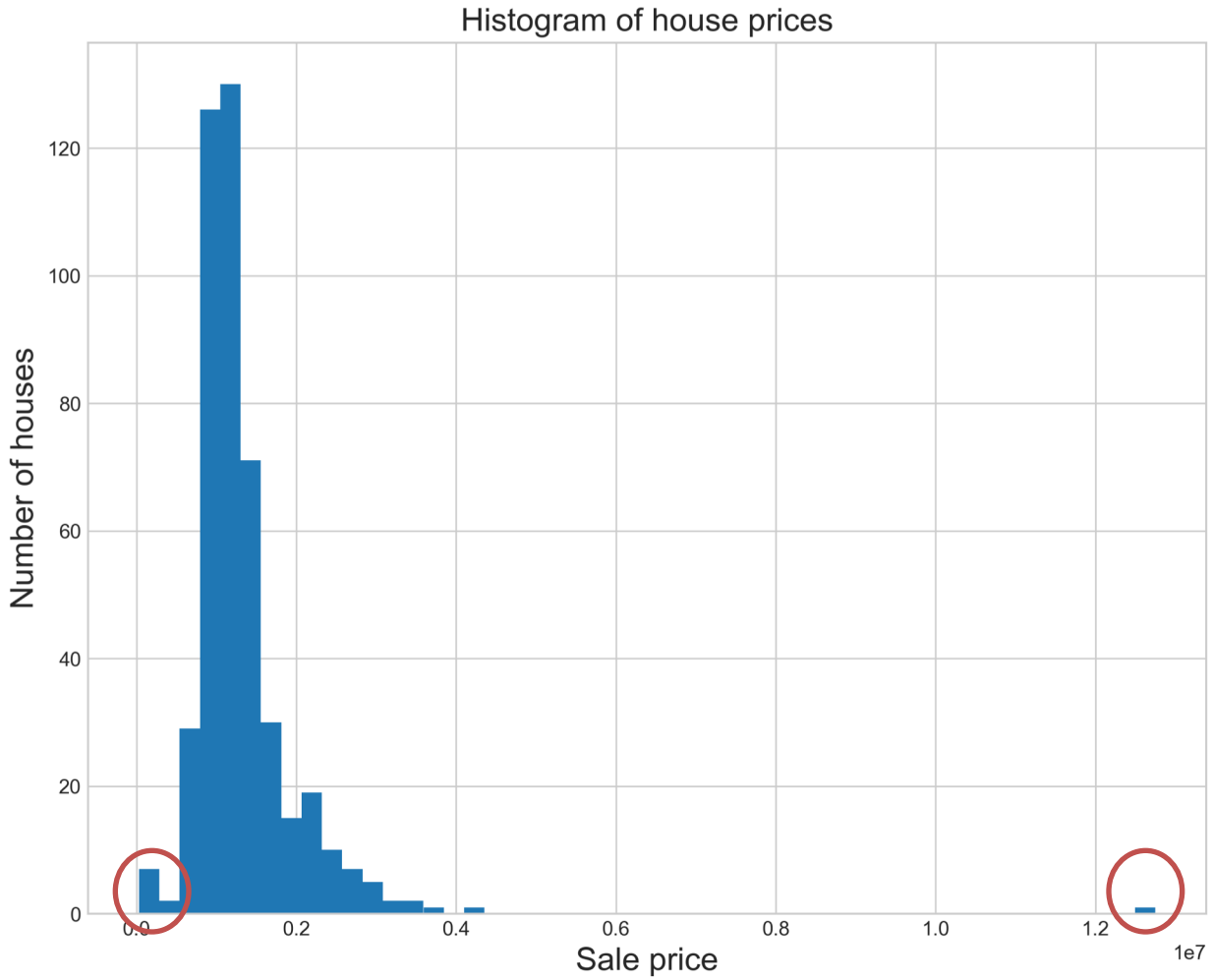
	address	city	zip	bedrooms	bathrooms	floor_size	lot_size	year_built	sale_date	sale_price
0	2804 Tangelo Ct	Pleasanton	94588	3	3	1614	2,526 sqft	Built in 1998	04/23/18	1039000.0
1	5650 Hansen Dr	Pleasanton	94566	4	2	1527	6,699 sqft	Built in 1973	04/20/18	1200000.0
2	592 Tawny Dr	Pleasanton	94566	3	2	1956	0.28 acres	Built in 1977	04/19/18	1150000.0
3	2668 Calle Morelia	Pleasanton	94566	5	3	2422	6,500 sqft	Built in 1984	04/18/18	1430000.0
4	6048 Inglewood Dr	Pleasanton	94588	4	2	1733	6,499 sqft	Built in 1968	04/18/18	1125000.0

Data cleaning process:

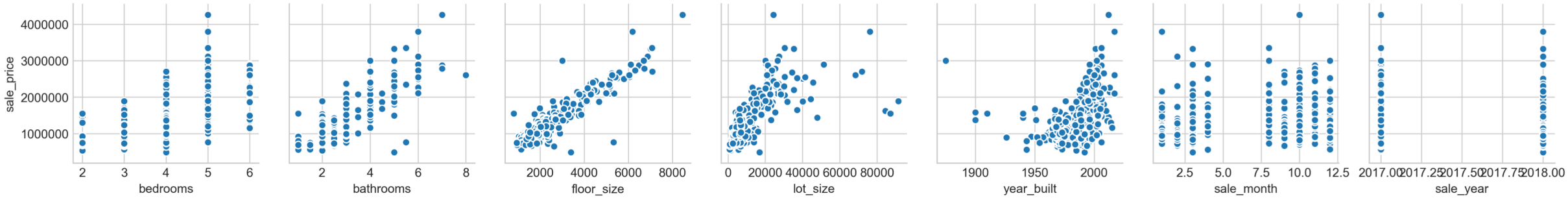
- convert all strings to numerical values
- drop any rows that had NaNs
- convert acres to sqft
- convert sale date to datetime format
- create new columns corresponding to sale month and sale year

Post data cleaning resulted in 458 house listings

The distribution of the sale price data showed some outliers



The pair plot data indicates a few features are quit predictive of price

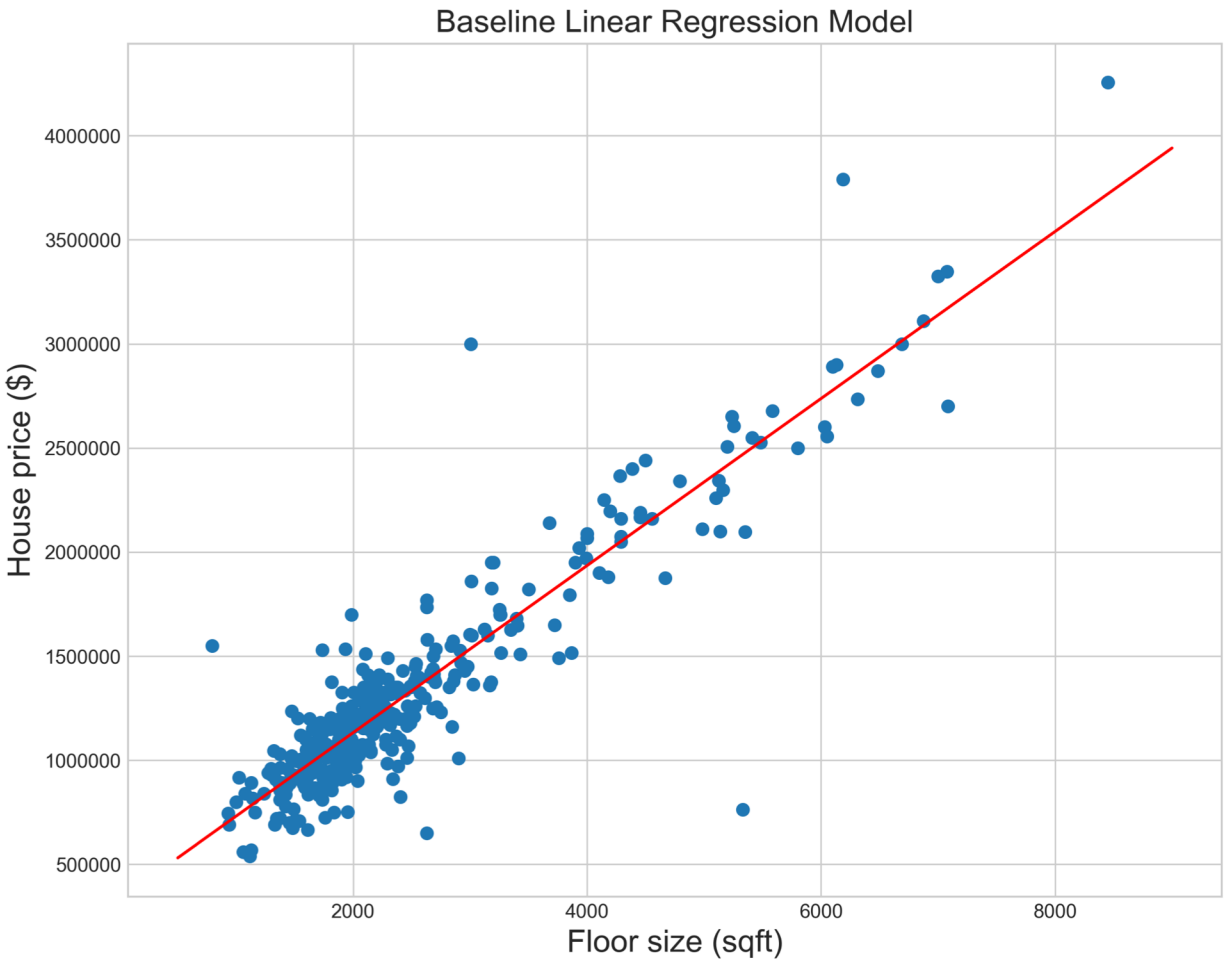


The correlation data shows I may be able to get a reasonable model just using floor_size

```
houses_df5.corr()['sale_price'].sort_values(ascending=False)
```

```
sale_price     1.000000
floor_size     0.915636
bathrooms      0.792113
lot_size       0.663012
bedrooms       0.592761
year_built     0.369638
sale_year      0.035768
sale_month    -0.022336
Name: sale_price, dtype: float64
```

The baseline model using one feature does a good job at predicting price



R^2 scores

	Baseline LR model
# features	1
Training set	0.85
Cross validation set	0.84

Linear regression using all features does slightly better

OLS Regression Results

Dep. Variable:	sale_price	**R-squared:**	0.866
Model:	OLS	**Adj. R-squared:**	0.864
Method:	Least Squares	**F-statistic:**	405.0
Date:	Wed, 25 Apr 2018	**Prob (F-statistic):**	4.34e-187
Time:	22:05:37	**Log-Likelihood:**	-6070.9
No. Observations:	447	**AIC:**	1.216e+04
Df Residuals:	439	**BIC:**	1.219e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.748e+08	1.14e+08	-3.293	0.001	-5.99e+08	-1.51e+08
bedrooms	1814.5525	1.46e+04	0.125	0.901	-2.68e+04	3.04e+04
bathrooms	-3367.6203	1.9e+04	-0.177	0.859	-4.07e+04	3.4e+04
floor_size	373.2247	19.102	19.539	0.000	335.682	410.767
lot_size	7.1670	1.062	6.747	0.000	5.079	9.255
year_built	-1586.0611	655.517	-2.420	0.016	-2874.402	-297.720
sale_month	1.41e+04	7203.931	1.958	0.051	-54.616	2.83e+04
sale_year	1.875e+05	5.63e+04	3.330	0.001	7.68e+04	2.98e+05

R^2 Results

	Baseline LR model	Complete LR model
# features	1	7
Training set	0.85	0.87
Cross validation set	0.84	0.84

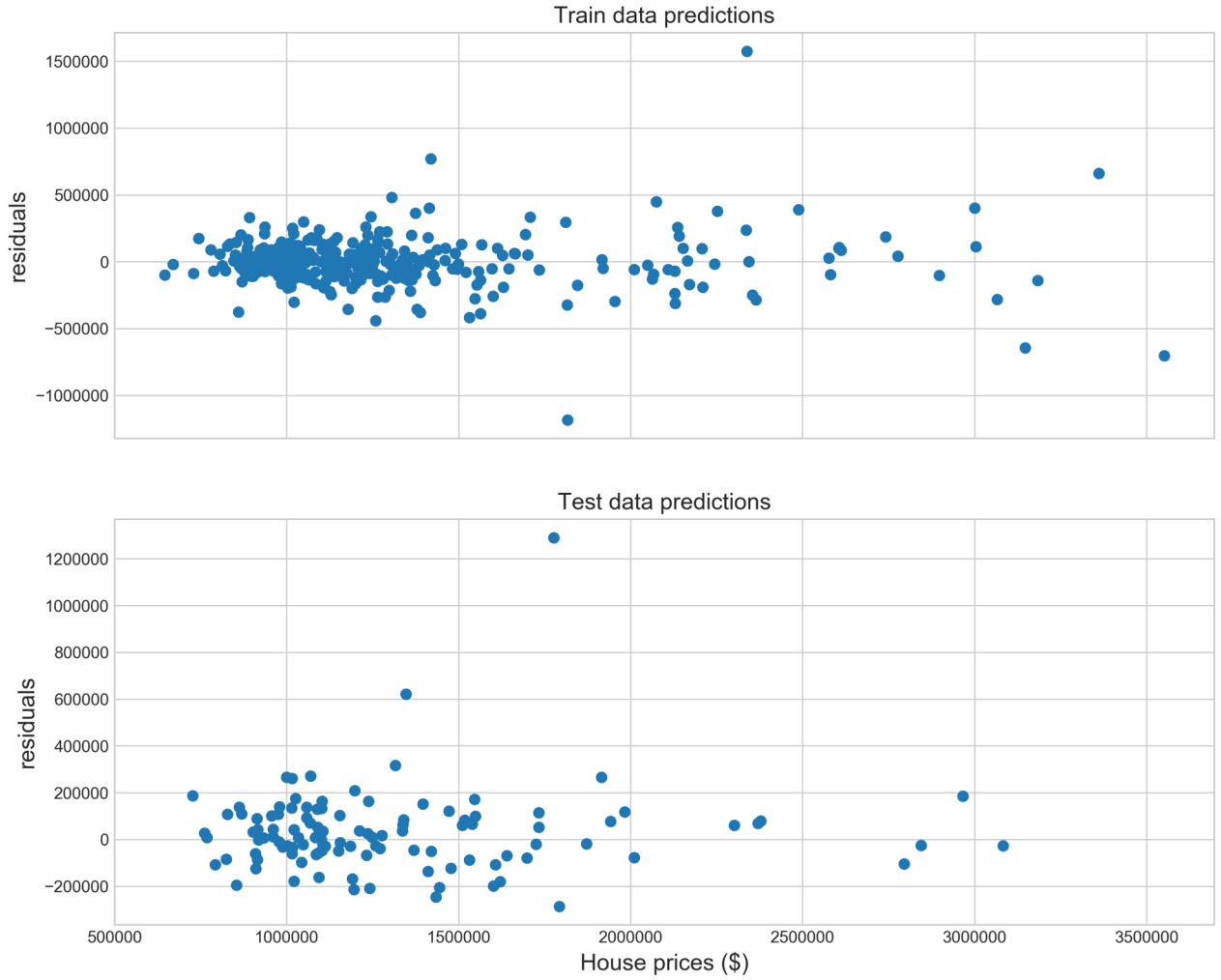
The P-values indicate that bedrooms and bathrooms are not significant for predicting price

Also tried Lasso and Ridge regression, and Lasso gave way to slightly better scores

	Baseline LR model	Complete LR model	Lasso Regression	Ridge Regression
# features	1	7	6	7
Training set	0.85	0.87	0.86	0.87
Cross validation set	0.84	0.84	0.85	0.84

Bedrooms feature ignored with Lasso

The residual plots for the Lasso regression model suggest that the linear model performs well across all house prices



The model is 86.5% accurate at predicting house prices in Pleasanton

	Lasso Regression
# features	6
Training set	0.86
Cross validation set	0.85
Test set score	0.865

Next steps

Scrape more features and do it for more homes

- Other good data to scrape
  - Key phrases: “Newly updated”, “Updated kitchen”, “Remodeled kitchen”
  - Indoor features: Fireplace, Central Air
  - Outdoor features: Patio, Pool

Incorporate geolocation data in model

Do log transform on the sale data to reduce the skew