MODELING SEATTLE HOUSING PRICES

by Matthew E. Parker

INPUTS

Starting dataset:

> 21,500 house sales
19 variables per sale

variables list:

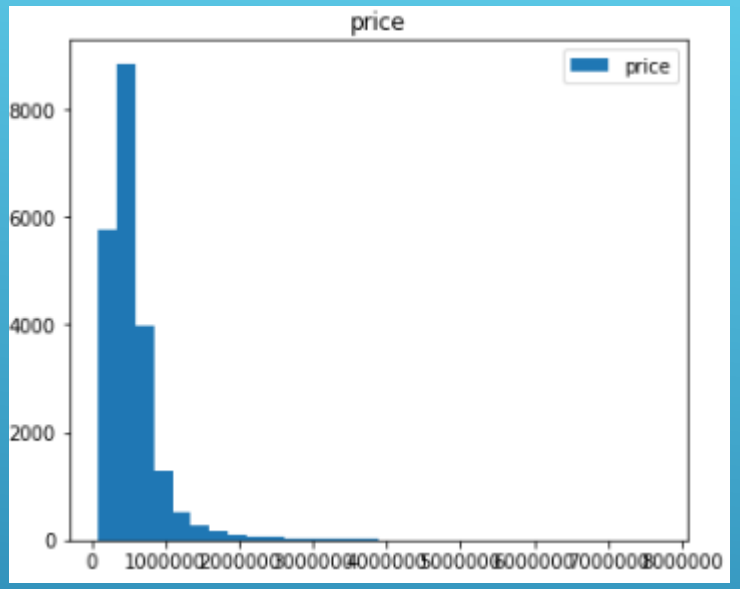
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15

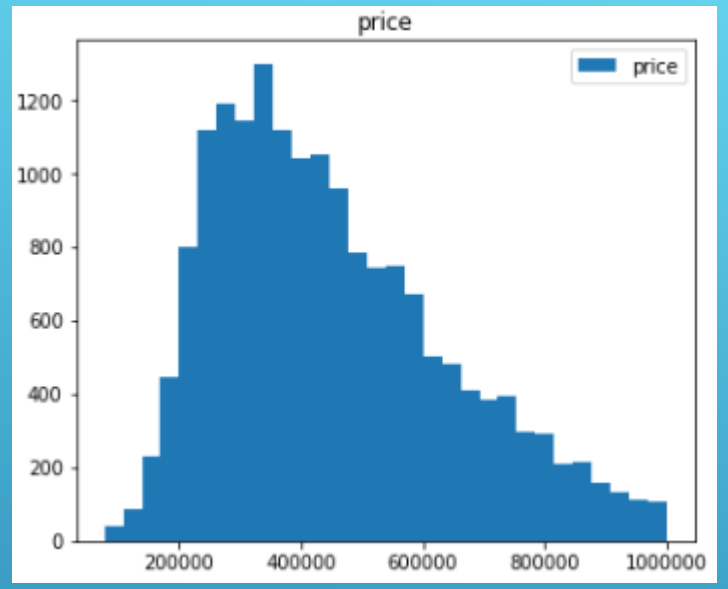
Our approach to constructing a model:

- Clean the data
- Explore and analyze the data
- Identify significant variables and build model around them
- Test and Validate model accuracy

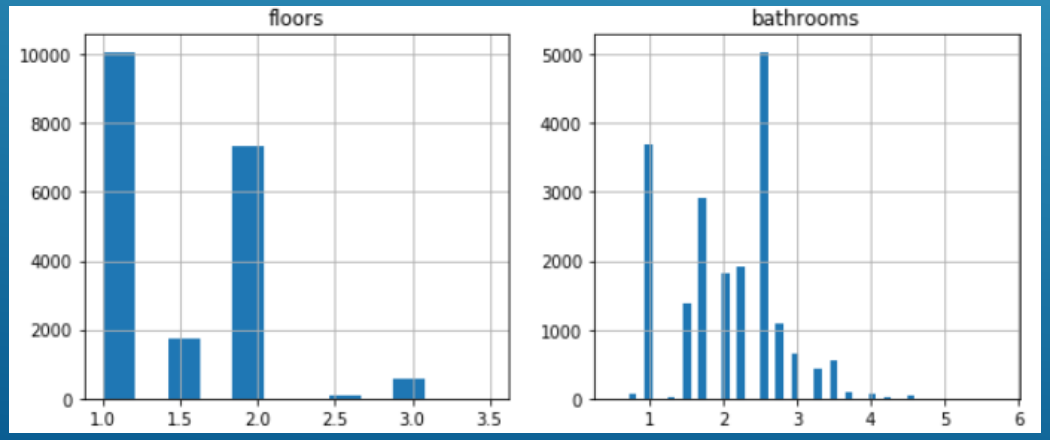
CLEANING THE DATA

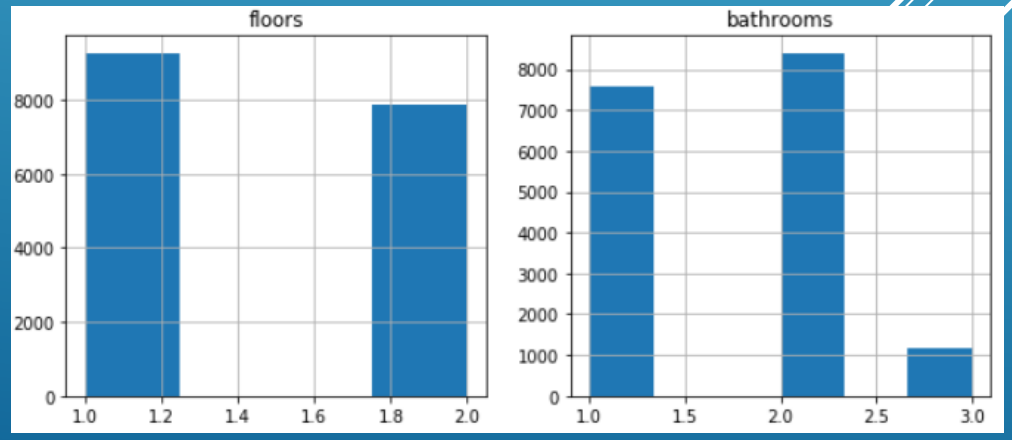
Removing outliers from the dataset



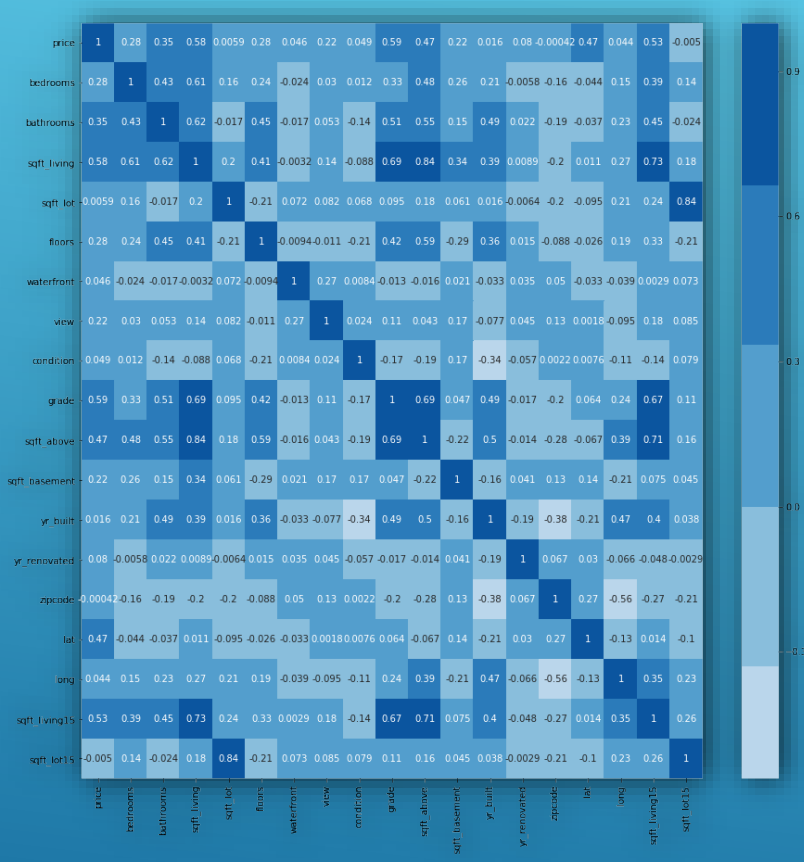


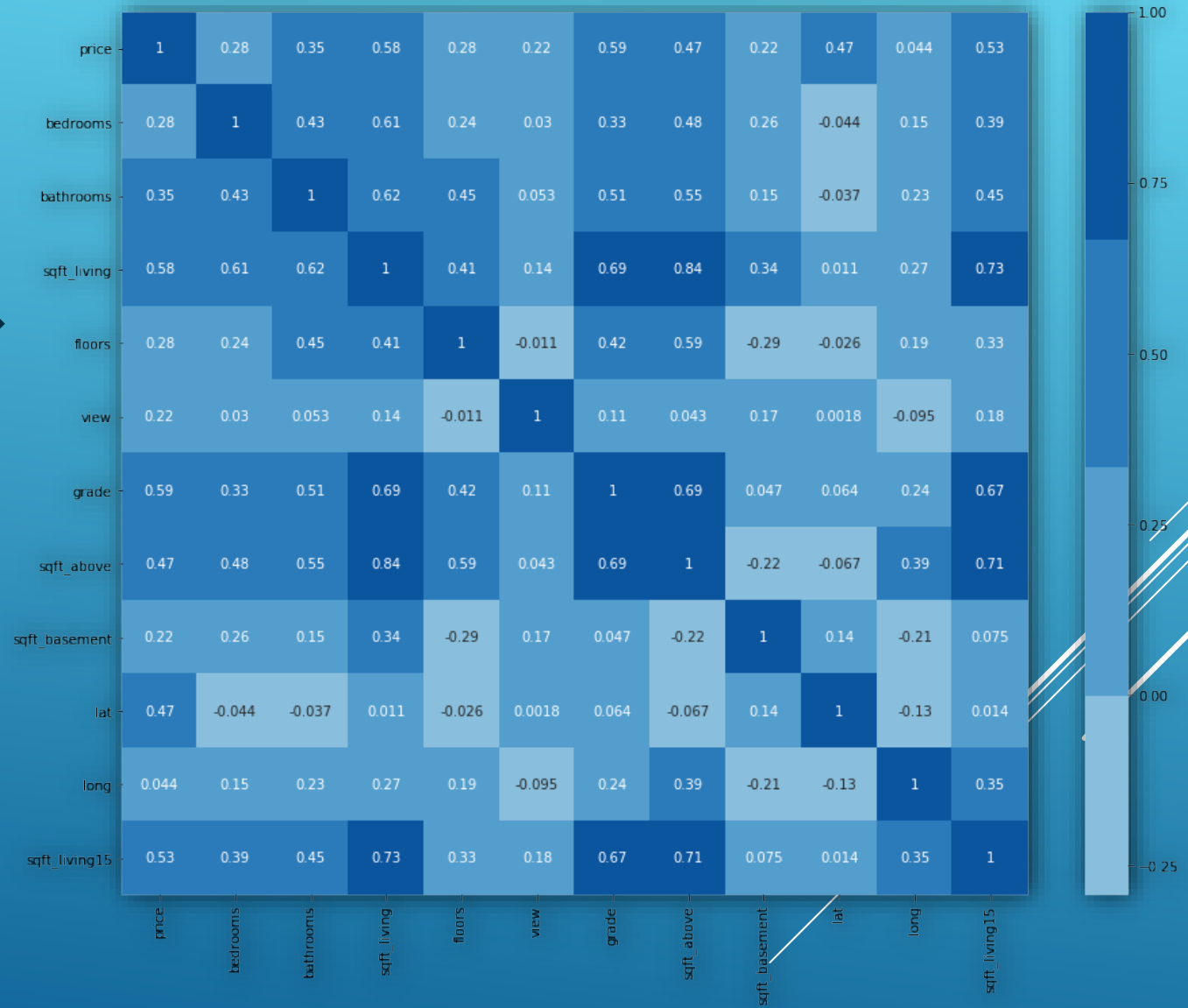
Improving irregular data through grouping





EXPLORING THE DATA

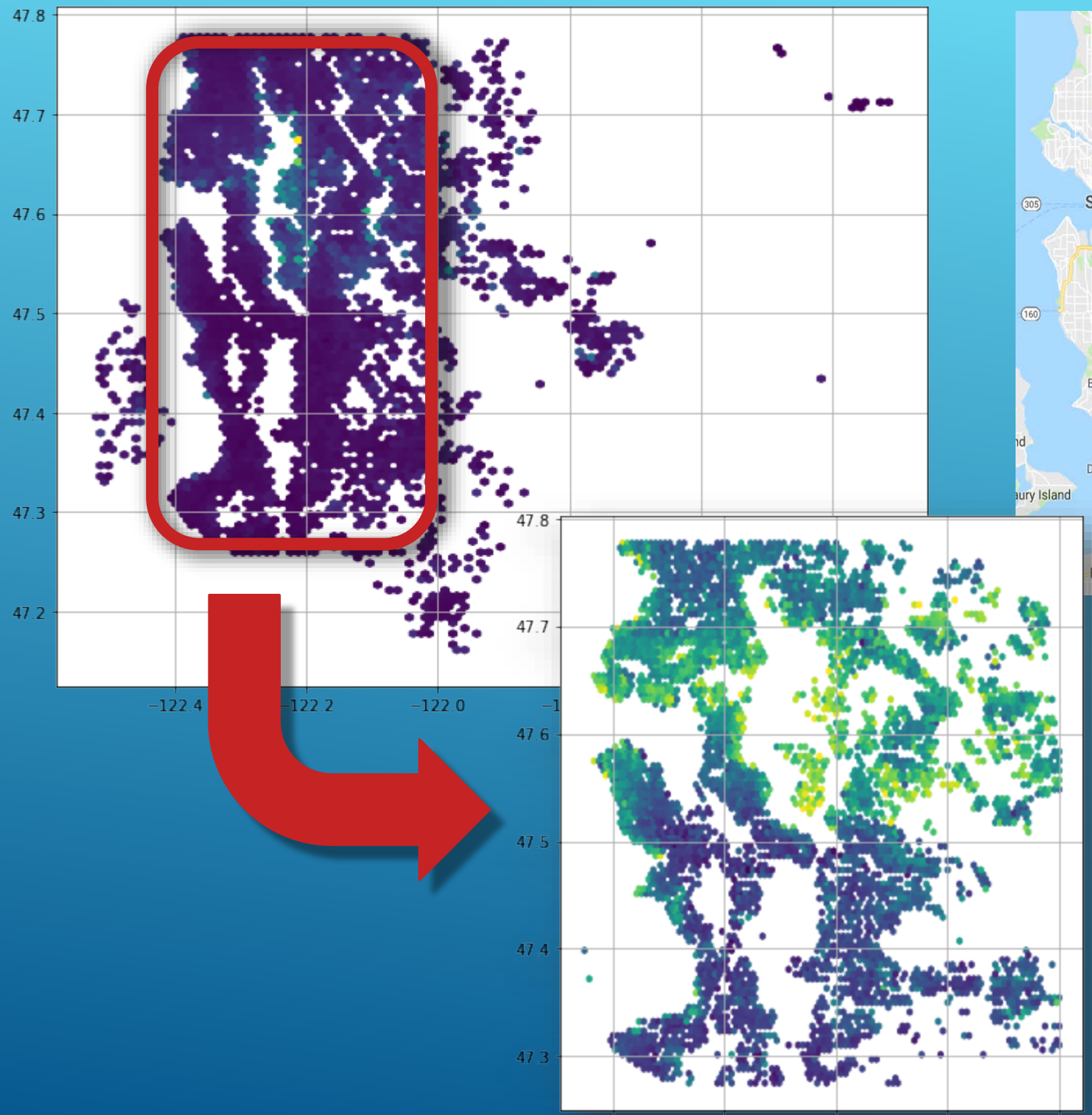


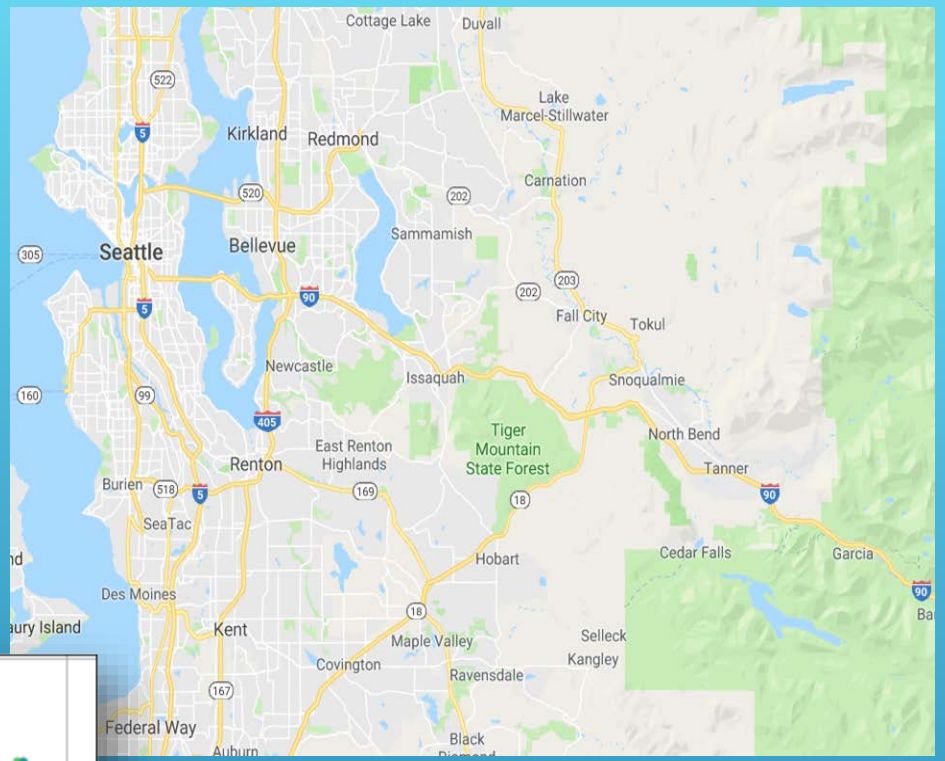


Correlation Heatmap

dark blue = high correlation

light blue = low correlation





HANDLING GEOGRAPHY

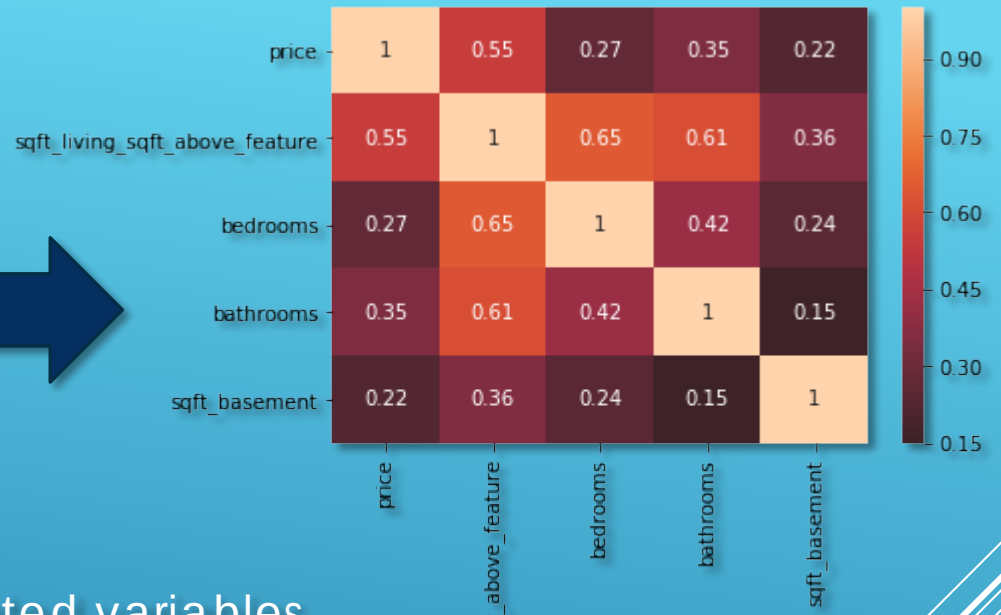
When several variables are collinear, changing one has a change in the others.

This is bad for modeling since changes get multiplied in the model output.

For instance, you can't change a house's total living space without *also* changing it's total above-ground living space as well.

To solve this, we can build features that weight the variables proportional to their influence.

Reduced collinearity

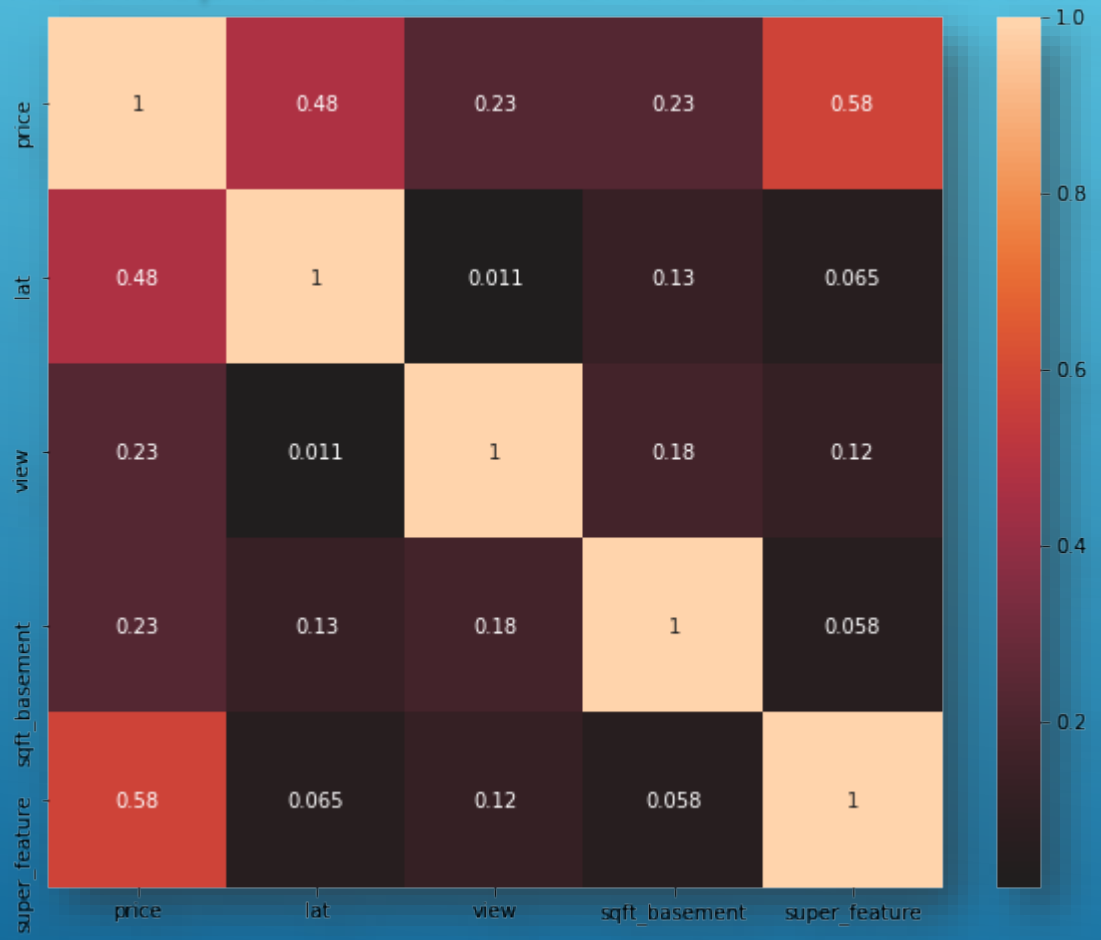


Heatmap of highly co-related variables

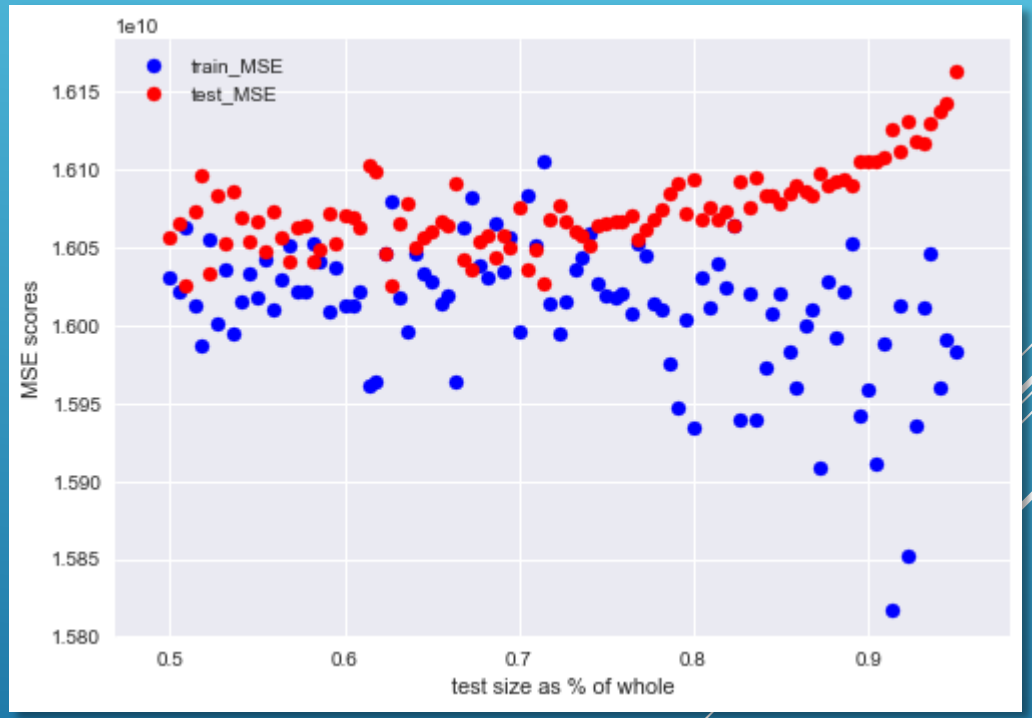
BUILDING CUSTOM FEATURES

TRAINING & VALIDATING THE MODEL

Heatmap of correlation of final model features



Model validation involved averaging 100 randomly sampled iterations per 100 different sample sizes for a total of 10,000 tests.



The data was then also validated using the K-folds cross-validation method to run 50 cross-validation routines, resulting in an average model **margin of error of $126,700.00.**

$$p = (1296579.6 \times f_L) + (60401.32 \times f_V) + (110.79 \times f_B) + (59144.57 \times f_G) + (111362.3 \times log_e(f_S)) - 62564620.4$$

p = House price (in USD)

f_L = latitude

f_V = times property has been viewed

f_B = square footage of basement

f_G = grade given to the housing unit, based on King County grading system

f_S = square footage of living space

MODEL SUMMARY

If you know a house's latitude, basement ft^2, living space ft^2, King County grade, and the number of times it has been viewed, then you can estimate it's sale price within a margin of $126,700.^{00}.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.885
Model:	OLS	**Adj. R-squared:**	0.885
Method:	Least Squares	**F-statistic:**	3.067e+04
Date:	Sun, 02 Jun 2019	**Prob (F-statistic):**	0.00
Time:	17:25:32	**Log-Likelihood:**	-2.1350e+05
No. Observations:	15877	**AIC:**	4.270e+05
Df Residuals:	15873	**BIC:**	4.270e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
lat	6.517e+05	2557.555	254.798	0.000	6.47e+05	6.57e+05
view	6.04e+04	2360.498	25.588	0.000	5.58e+04	6.5e+04
sqft_basement	1.651e+05	5214.250	31.659	0.000	1.55e+05	1.75e+05
super_feature	6.91e+05	9893.180	69.845	0.000	6.72e+05	7.1e+05

Omnibus:	198.032	**Durbin-Watson:**	1.841
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	205.268
Skew:	-0.278	**Prob(JB):**	2.67e-45
Kurtosis:	2.958	**Cond. No.**	5.25

RECOMMENDATIONS

If you can purchase a house for $126,700 less than the price predicted by our model, then you should definitely do so. When trying to sell it, you could begin by listing it for $126,700 above the model prediction and then have room to negotiate down to the model's predicted price if need be, still making a hefty profit.

FURTHER INVESTIGATION

The current model could likely be enhanced by the addition of more variables.

In particular, information on crime rates, transportation accessibility, school district ratings, etc. would be useful as these factors have in the past been shown to influence real estate pricing.

THANK YOU