KC Real Estate

Predicting Home Prices

using linear/multiple regression

King County House Sales Data (2014-2015)

- 21,597 house sales

- includes 21 features from the King County Property Assessor

- description of features from the King County Assessor website here.

Modeling Technique:

- Linear/Multiple Regression

Model Success Metrics:

- R-Squared
- Root Mean Squared Error

Linear/Multiple Regression Modeling

- predictive framework for future house sales

- iterative approach to build an accurate model

Modeling Metrics:

- R-Squared: goodness-of-fit, range: 0-1 (bigger=better)
- Root Mean Squared Error: standard deviation of the model's error (smaller=better)

Baseline Model:

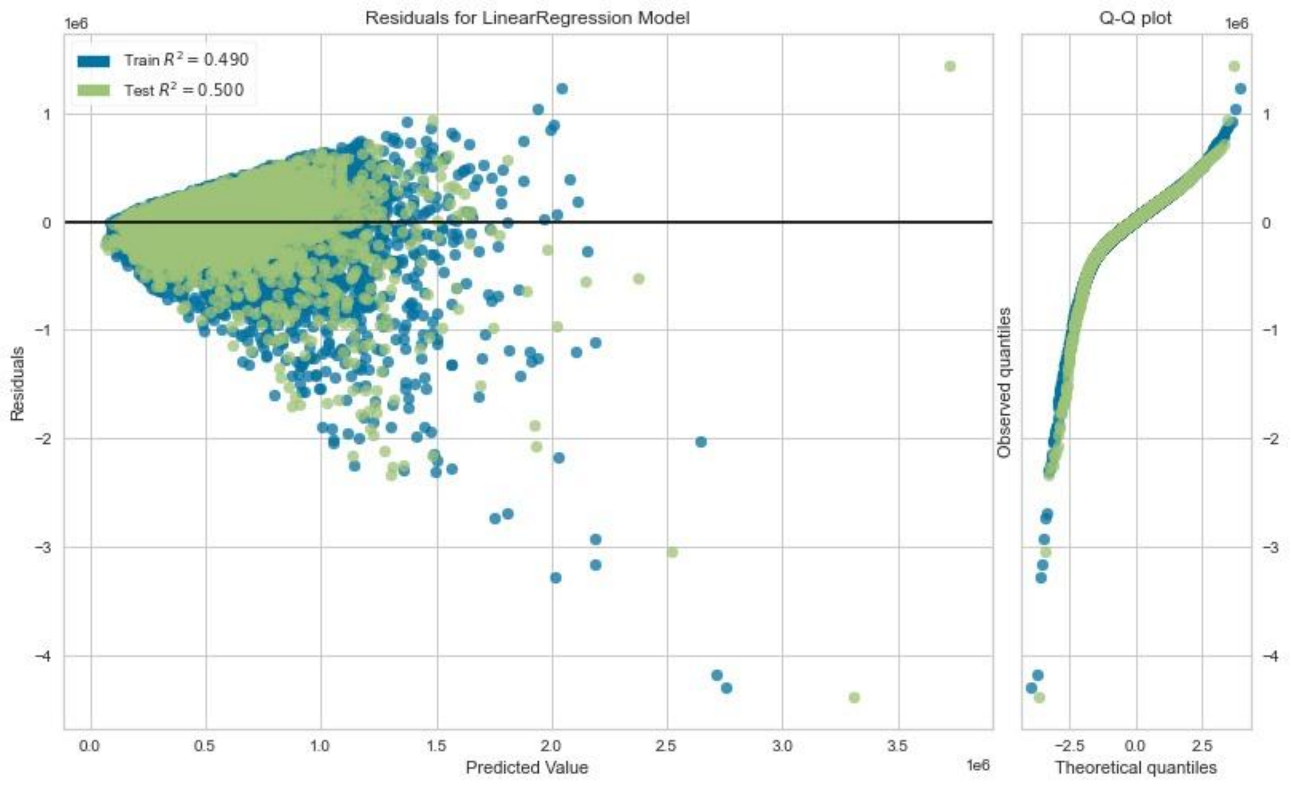
- began with a 1-variable model: Square footage of the living space.

- Correlation: 70.2%

Model Metrics:

- R-Squared: 50%

- RMSE: $269,556



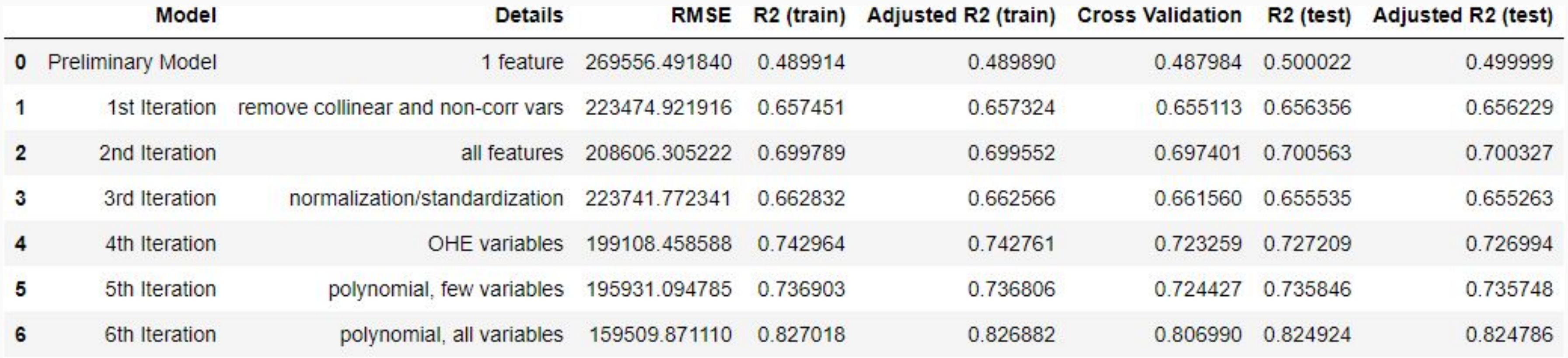
Model Progression:

Iterative Approaches:

- Increase variables
- Remove collinear variables
- One-Hot-Encoding
- Multiple Regression

Results:

- Increase in R-squared: 50% to 82%
- Decrease in RMSE: $269,556 to $159,510



	Model	Details	RMSE	R2 (train)	Adjusted R2 (train)	Cross Validation	R2 (test)	Adjusted R2 (test)
0	Preliminary Model	1 feature	269556.491840	0.489914	0.489890	0.487984	0.500022	0.499999
1	1st Iteration	remove collinear and non-corr vars	223474.921916	0.657451	0.657324	0.655113	0.656356	0.656229
2	2nd Iteration	all features	208606.305222	0.699789	0.699552	0.697401	0.700563	0.700327
3	3rd Iteration	normalization/standardization	223741.772341	0.662832	0.662566	0.661560	0.655535	0.655263
4	4th Iteration	OHE variables	199108.458588	0.742964	0.742761	0.723259	0.727209	0.726994
5	5th Iteration	polynomial, few variables	195931.094785	0.736903	0.736806	0.724427	0.735846	0.735748
6	6th Iteration	polynomial, all variables	159509.871110	0.827018	0.826882	0.806990	0.824924	0.824786

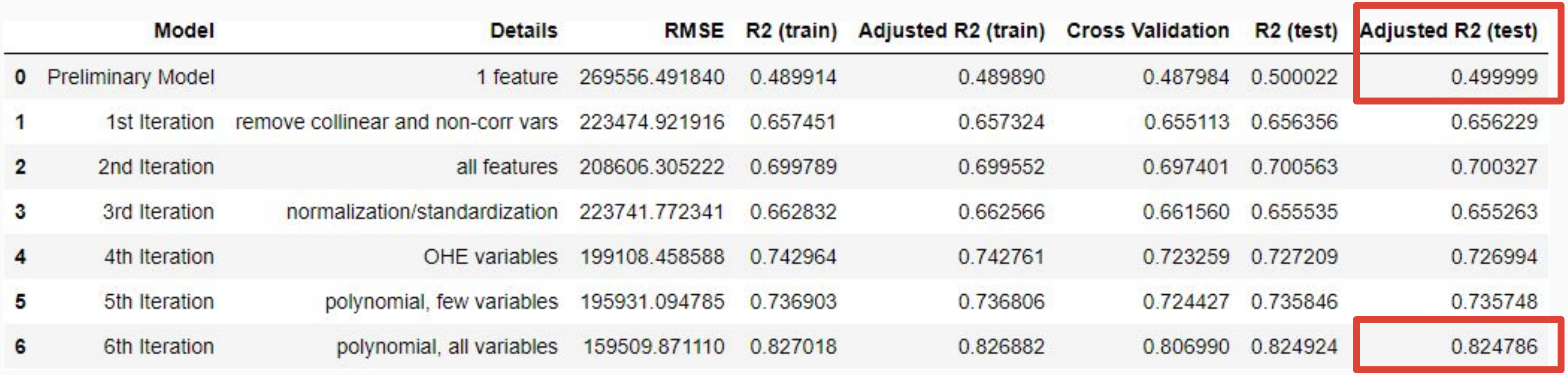
Model Progression:

Iterative Approaches:

- Increase variables
- Remove collinear variables
- One-Hot-Encoding
- Multiple Regression

Results:

- Increase in R-squared: 50% to 82%
- Decrease in RMSE: $269,556 to $159,510



	Model	Details	RMSE	R2 (train)	Adjusted R2 (train)	Cross Validation	R2 (test)	Adjusted R2 (test)
0	Preliminary Model	1 feature	269556.491840	0.489914	0.489890	0.487984	0.500022	0.499999
1	1st Iteration	remove collinear and non-corr vars	223474.921916	0.657451	0.657324	0.655113	0.656356	0.656229
2	2nd Iteration	all features	208606.305222	0.699789	0.699552	0.697401	0.700563	0.700327
3	3rd Iteration	normalization/standardization	223741.772341	0.662832	0.662566	0.661560	0.655535	0.655263
4	4th Iteration	OHE variables	199108.458588	0.742964	0.742761	0.723259	0.727209	0.726994
5	5th Iteration	polynomial, few variables	195931.094785	0.736903	0.736806	0.724427	0.735846	0.735748
6	6th Iteration	polynomial, all variables	159509.871110	0.827018	0.826882	0.806990	0.824924	0.824786

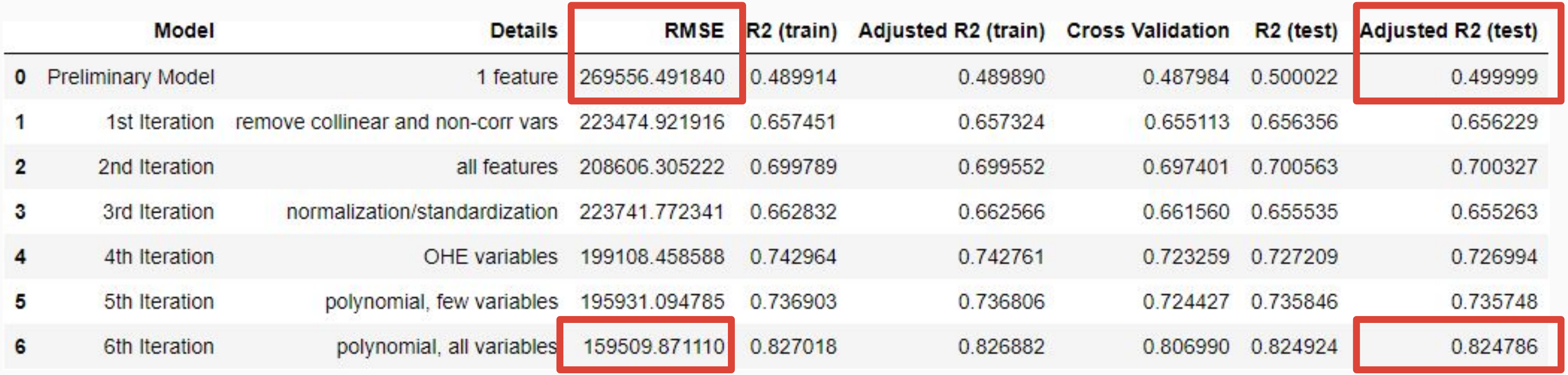
Model Progression:

Iterative Approaches:

- Increase variables
- Remove collinear variables
- One-Hot-Encoding
- Multiple Regression

Results:

- Increase in R-squared: 50% to 82%
- Decrease in RMSE: $269,556 to $159,510



	Model	Details	RMSE	R2 (train)	Adjusted R2 (train)	Cross Validation	R2 (test)	Adjusted R2 (test)
0	Preliminary Model	1 feature	269556.491840	0.489914	0.489890	0.487984	0.500022	0.499999
1	1st Iteration	remove collinear and non-corr vars	223474.921916	0.657451	0.657324	0.655113	0.656356	0.656229
2	2nd Iteration	all features	208606.305222	0.699789	0.699552	0.697401	0.700563	0.700327
3	3rd Iteration	normalization/standardization	223741.772341	0.662832	0.662566	0.661560	0.655535	0.655263
4	4th Iteration	OHE variables	199108.458588	0.742964	0.742761	0.723259	0.727209	0.726994
5	5th Iteration	polynomial, few variables	195931.094785	0.736903	0.736806	0.724427	0.735846	0.735748
6	6th Iteration	polynomial, all variables	159509.871110	0.827018	0.826882	0.806990	0.824924	0.824786

Final Model:

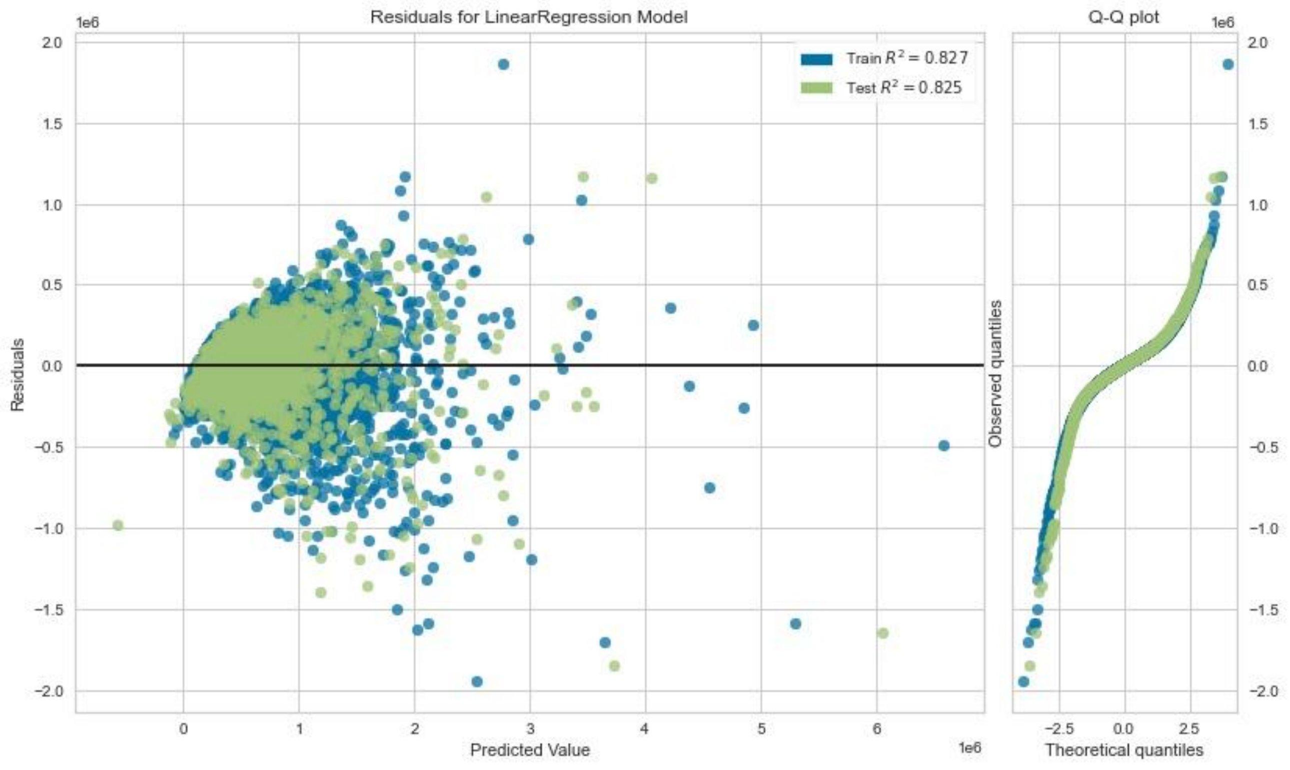
- 2nd degree polynomial regression model.

- variables: 171

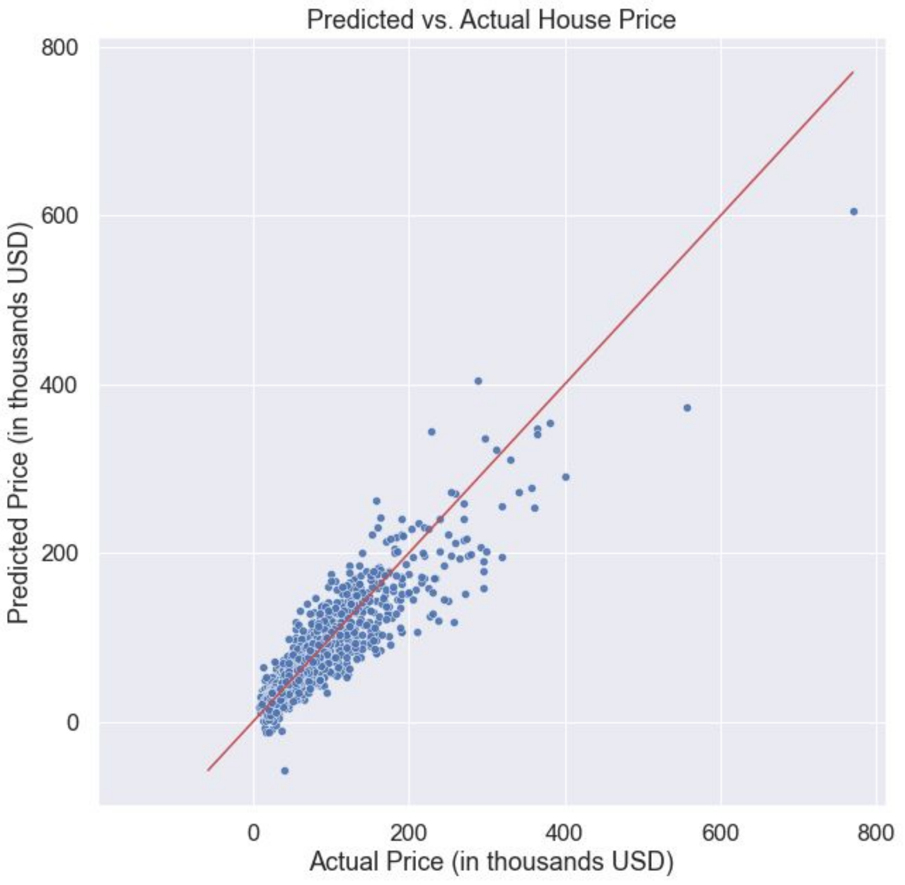
Model Metrics:

- R-Squared: 82%

- RMSE: $159,510



Model Validation/Recommendation:



Recommendation:

- The model is useful, but requires oversight. Especially for negative predictions and predictions over $200,000

Next Steps:

Follow-Up Analysis:

- Possible additional features
- Alternative modeling techniques
- Model updates as new data becomes available

Other Considerations:

- Market forces and impact
- Temporal changes

Questions?

Luke DiPerna

LinkedIn

GitHub