Ames Housing Data

Can we predict a price for a home with the optimal linear regression model that outperforms a baseline model?

The Problem Statement

Are inaccurate sales prices leading to negative outcomes for homeowners looking to sell? **This impacts homeowners pricing their homes way too high, which makes their homes unsellable even in the best markets. If we can predict a price for a home with the optimal linear regression model that can outperform a baseline model could we end up aiding homeowners in this situation**. As a data scientist working for Real Estate Company X, eliminating this pain point would also help real estate agents close more deals, in turn netting more revenue for the company via closing fees.

Research

Per Trulia.com's article "8 reasons Your House Isn't Selling", Determining the best asking price is one of the most important aspects of selling a home. **The bottom line - if you list the price of the home way above market value, you will miss out on prospective buyers**(trulia.com).

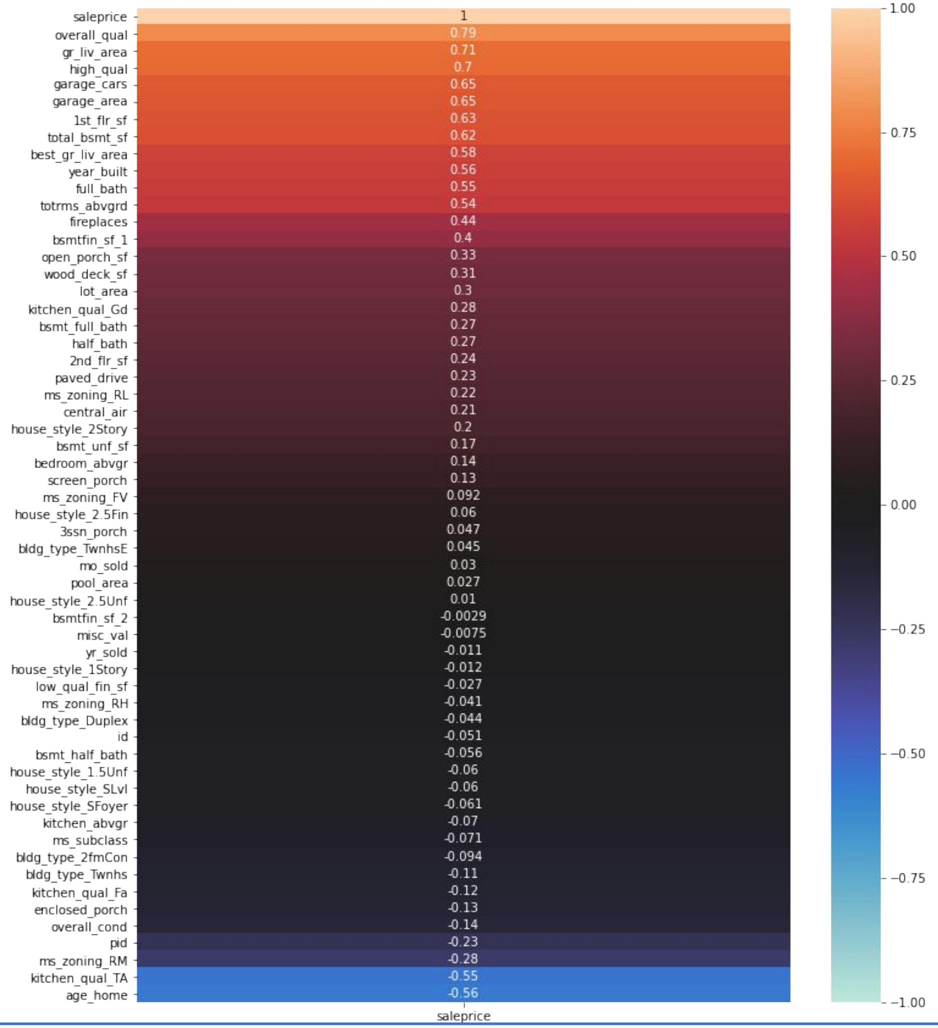
Feature Engineering

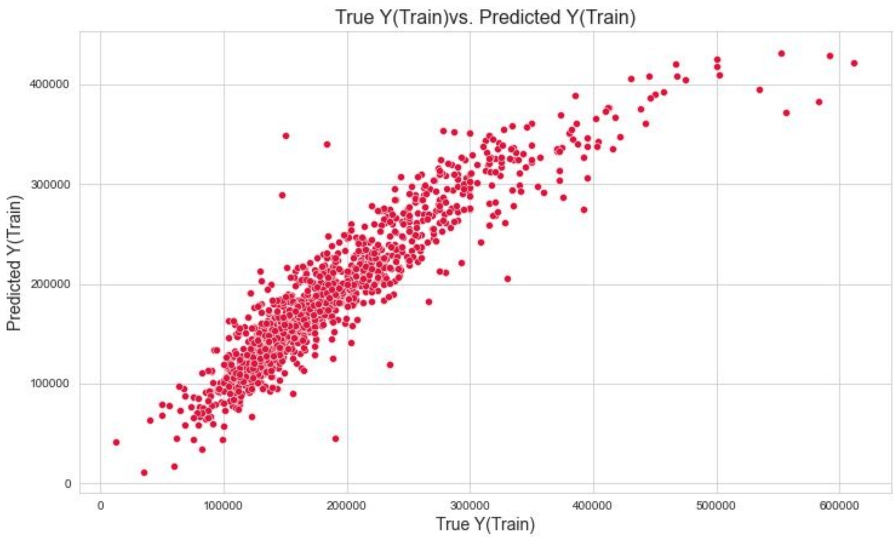
- Features such as number of fireplaces, 40% of home buyers are willing to cough up an extra 1400 dollars for one(Weigley, S.,2013).
- Another example of this is central air conditioning, "with nearly seven in 10 homeowners willing to pay extra"(Weigley, S., 2013).

The Right Model

- R2 Score
- Bias-Variance Tradeoff

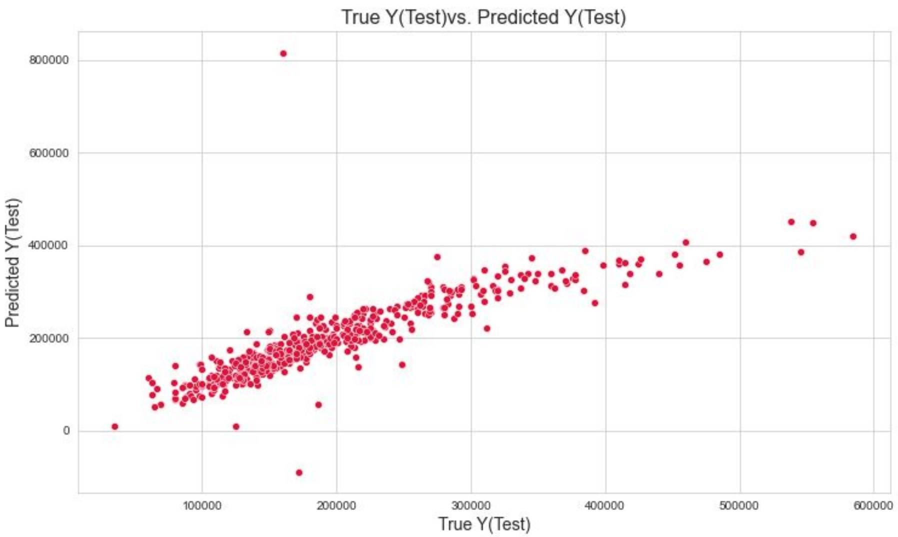
How Features Correlate with Sales Price

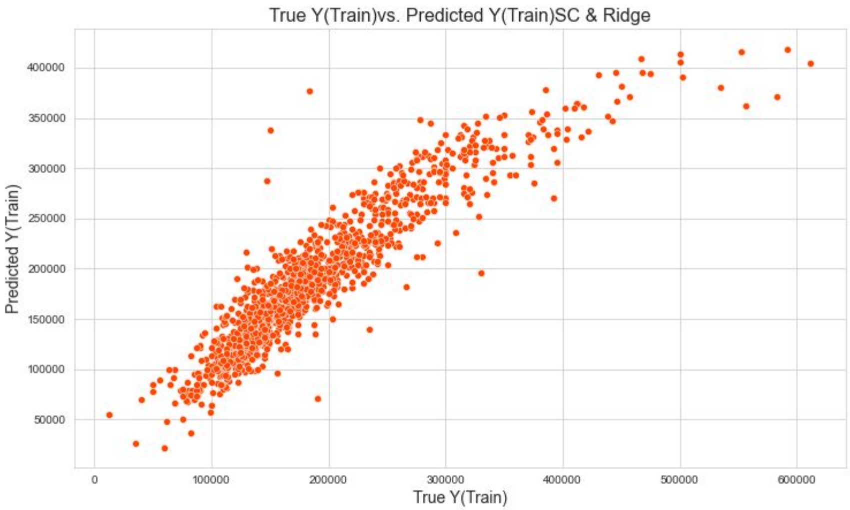




Train
86.28%

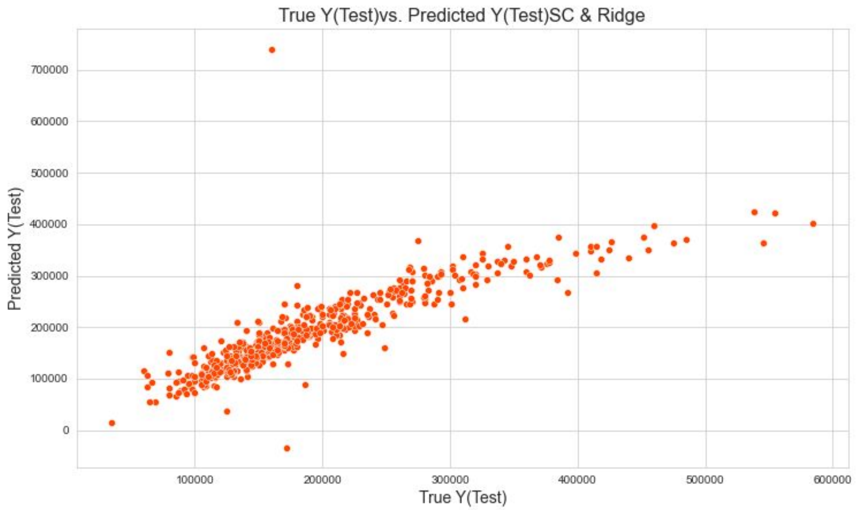
Test
73.46%

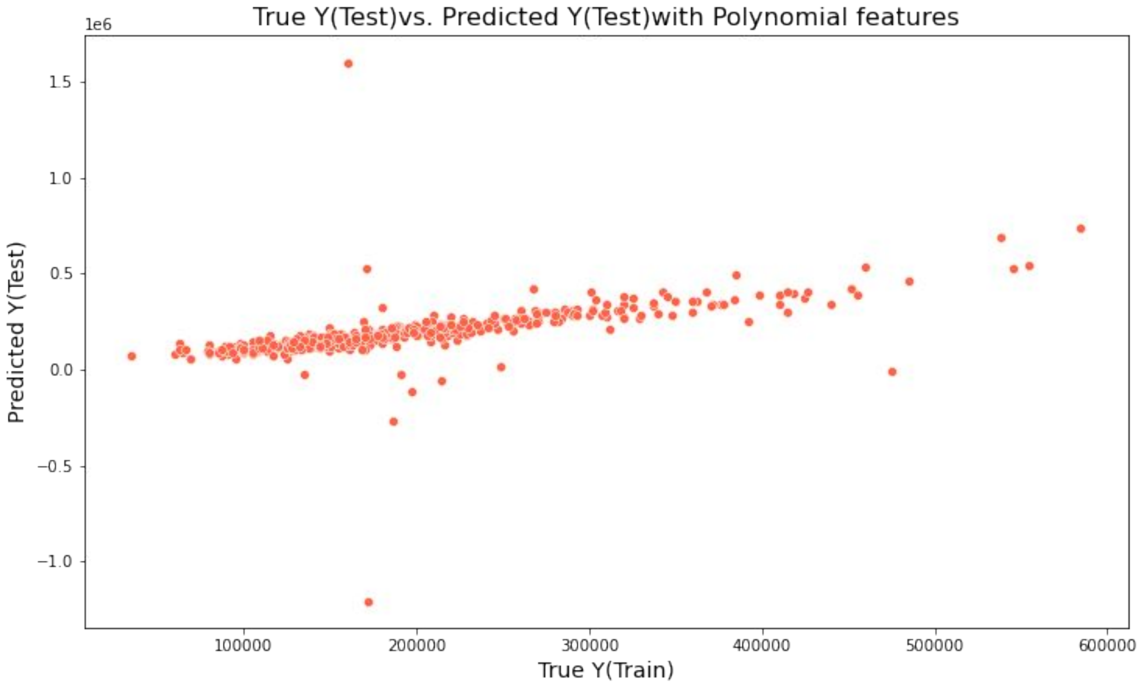




Train
85.80%

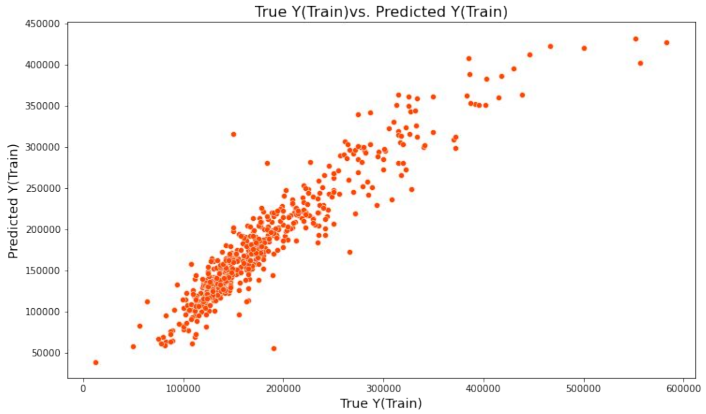
Test
76.02%





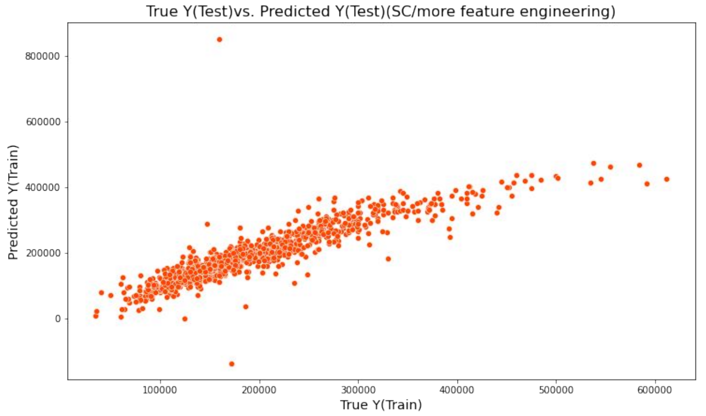
Train
94%

Test
-42%

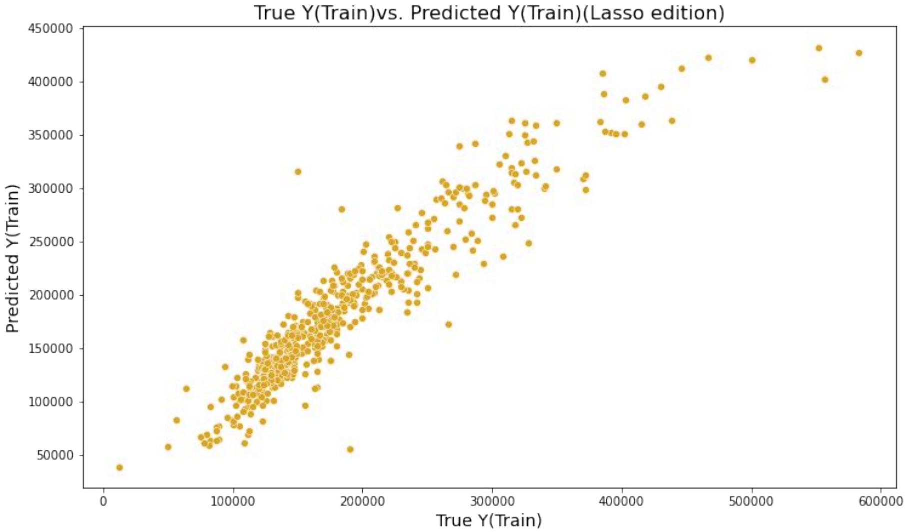


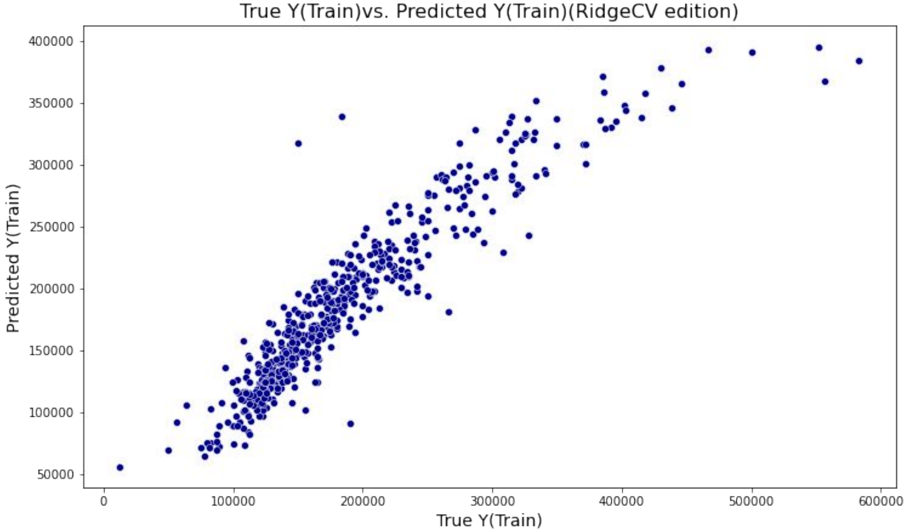
Train
88.24%

Test
80.64%



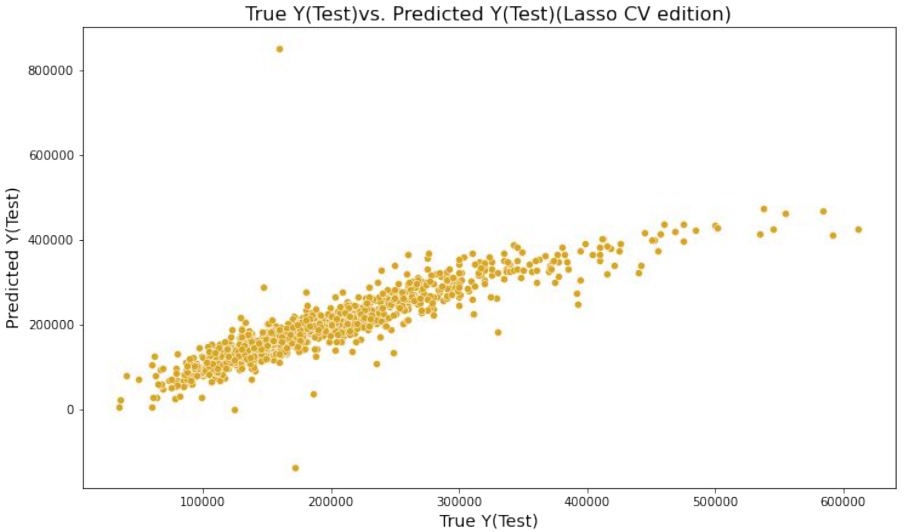
Train
88.24%

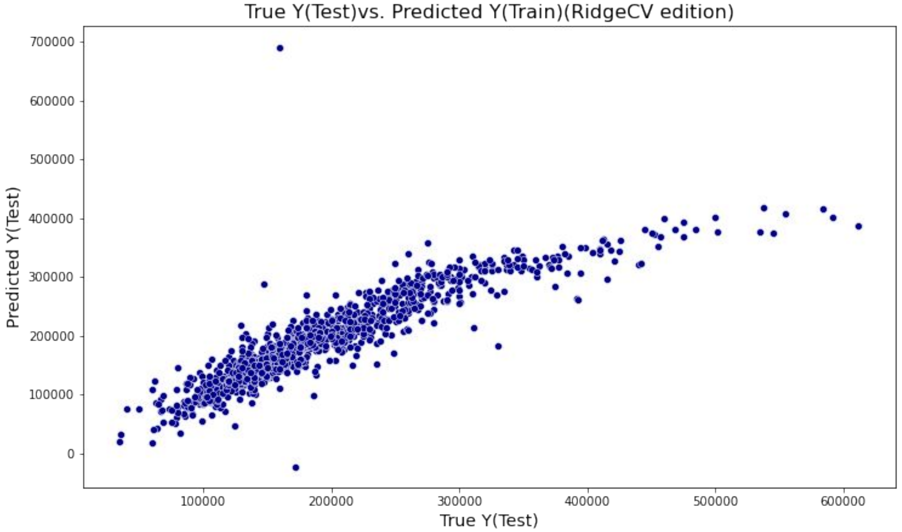




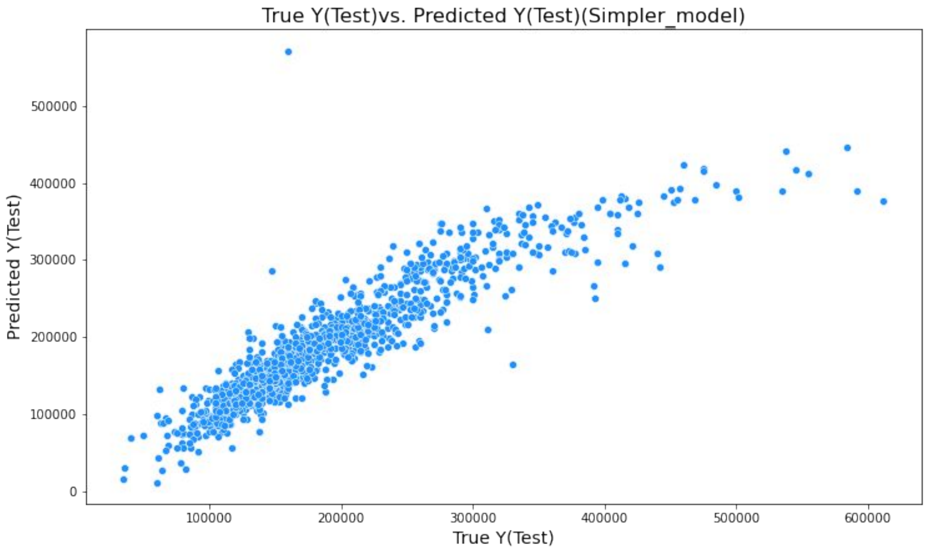
Train
86.30%

Test
80.65%



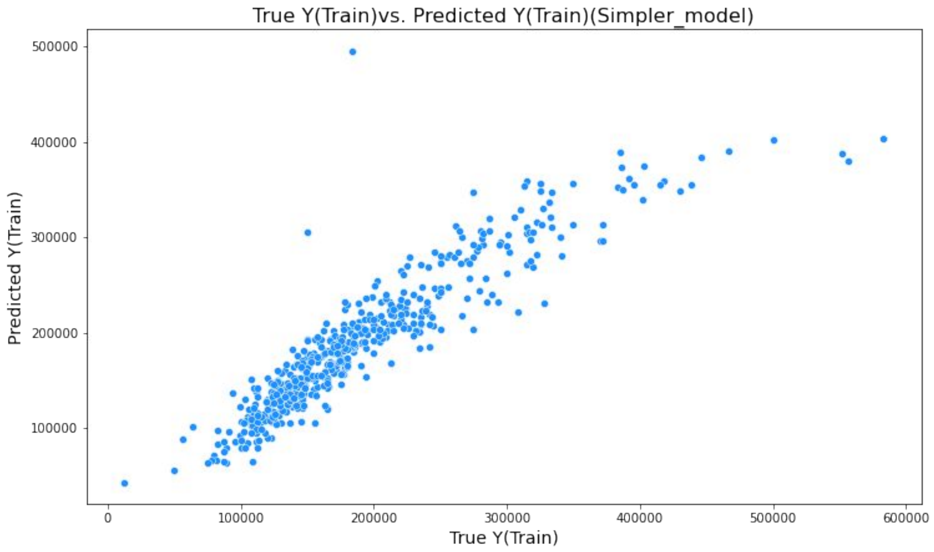


Test
82.54%



Train
84.17%

Test
84.44%



Conclusion

The initial premise here was to predict a price for a home using a linear regression model that outperforms a baseline model. After iterating through several models, I selected a Linear Regression model that whittled down features that had poor correlation. From there the polynomial features transformation and the standard scaler were applied. This ended up having the best R2 Score for testing portion of the train-test-split. It had an R2 score of 84.44% which was an improvement over the baseline of the original model many iterations ago, that baseline had a score of 67.59%. There was a 17-point improvement in score. Improvement as 84.44% of the variability in our sales price that could be explained by the features in our model. This is something that could help benefit homeowners and the real estate agents that work with them as far as being able to predict what an adequate price is that would enable them to sell the home. As we know from the research portion of this being overpriced will not allow for you to sell the home.

References

- Weigley, S.(2019). 11 home features buyers will pay extra for. Retrieved from [USA Today](https://www.usatoday.com/story/money/personalfinance/2013/04/28/24-7-home-features/2106203/).
- Web Editor.(nd). 8 Reasons Your House Isn't Selling. Retrieved from [Trulia Blog](https://www.trulia.com/blog/how-to-sell-a-house-8-reasons/).