The background of the image features four small, white, three-dimensional paper houses with red roofs, arranged in a staggered line from left to right. They are placed on a dark brown, weathered wooden surface with visible grain and texture.

Ames Housing Data

By David Castillo



The Problem Statement

As a data scientist working for real estate company, how can we optimize predicting pricing outcomes for homeowners looking to sell? Whether it's the homeowner pricing too high, leading their homes unsellable even in good markets or homeowners not maximizing the profits by listing the prices too low. We can optimize the predictions through feature engineering and linear regression so that we can have an R² score that exceeds a baseline model. We can also gain insights how these features contribute to pricing as well.



Research

Determining the best asking price is one of the most import aspects of selling a home.

If you list the price of the home way above market value, you will miss out on prospective buyers(trulia.com).

Feature Engineering

- Look at correlation for relevant features and whittle down.
- 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).
- Does the data set match what the articles state?





The right model

R2 Score

Coefficients

Bias-Variance Tradeoff



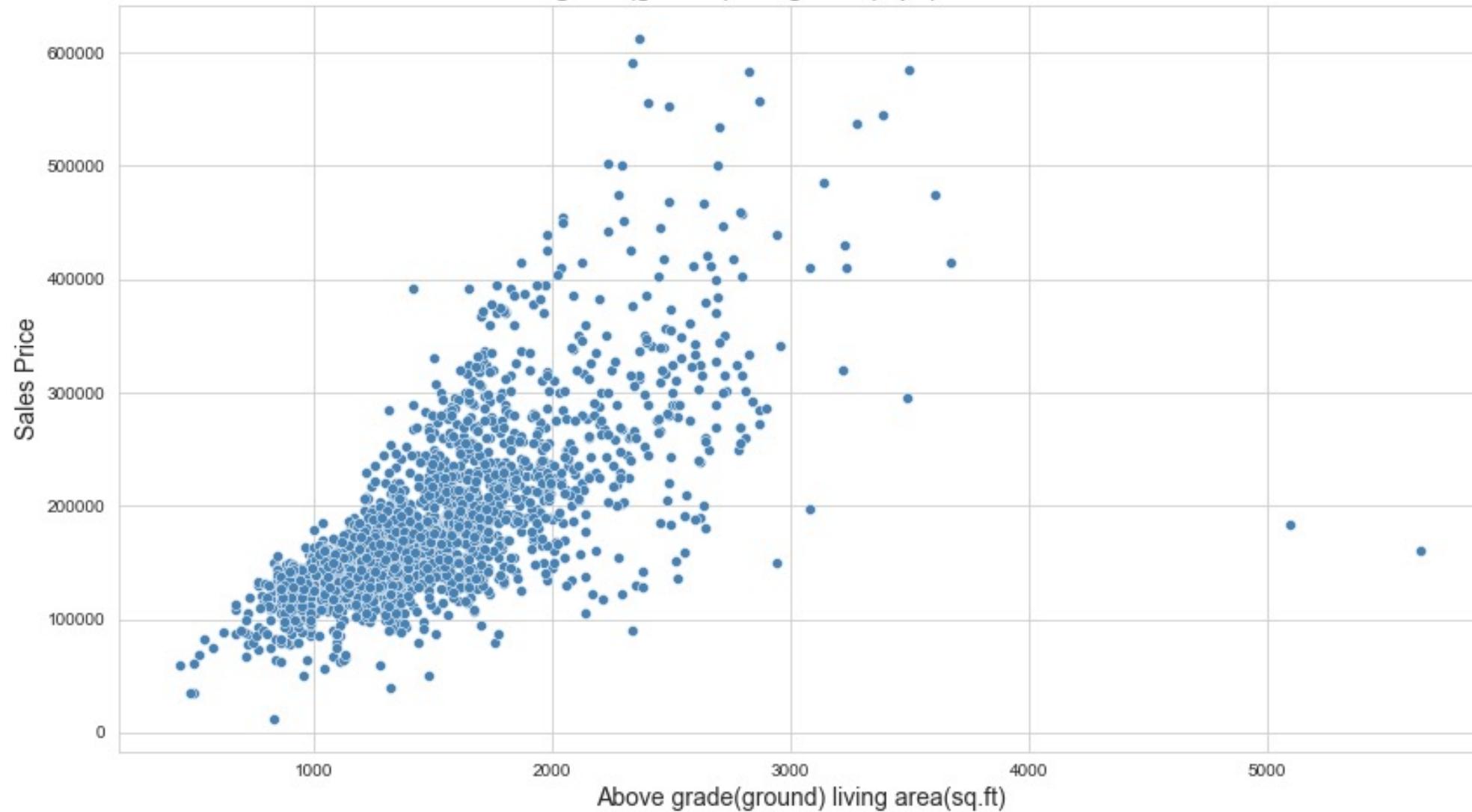
Top Features

- Overall Quality
- Above grade living area(sq ft)
- Garage area

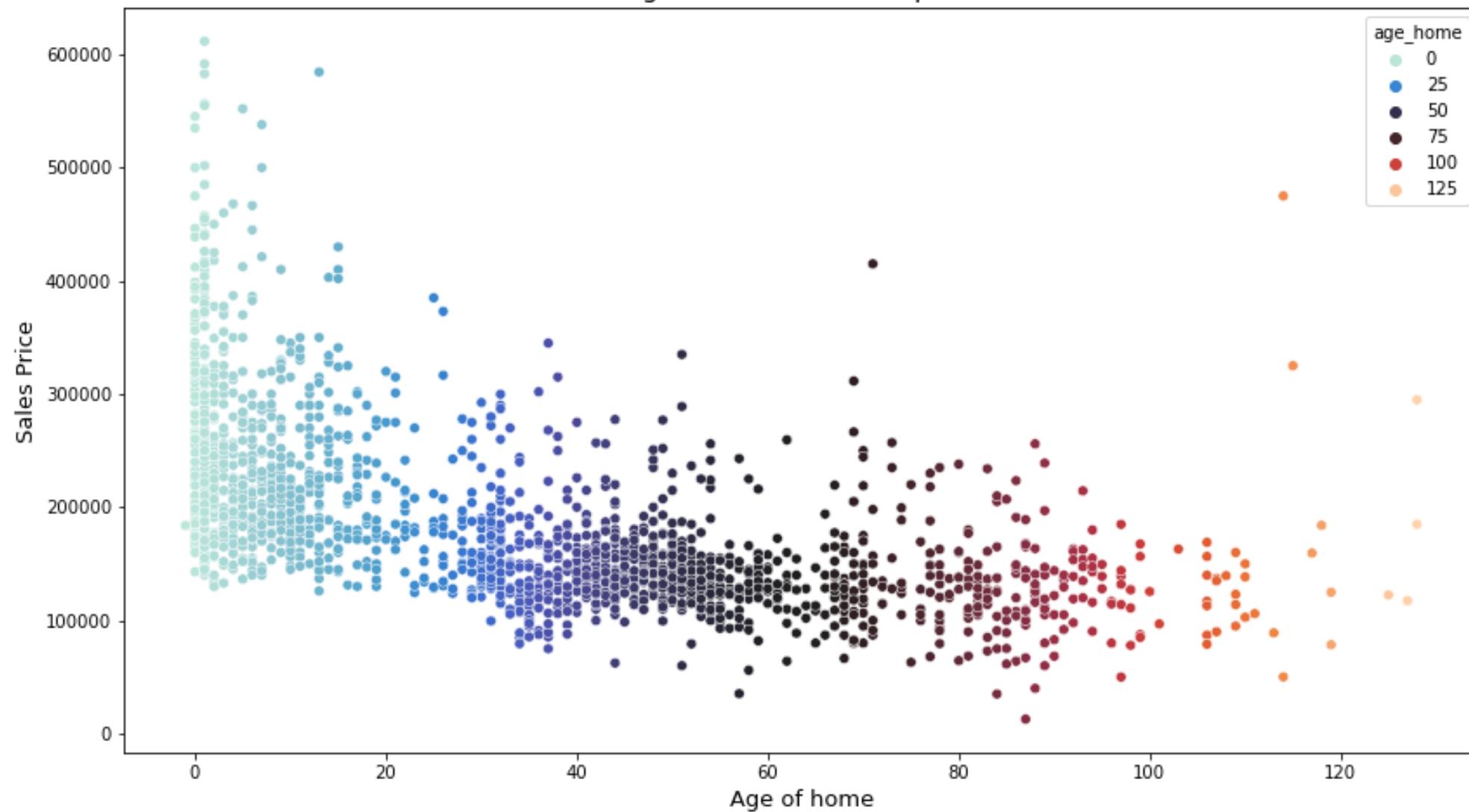
Bottom Features

- Age of the home(older ones lose out)
- Less than great Kitchen quality

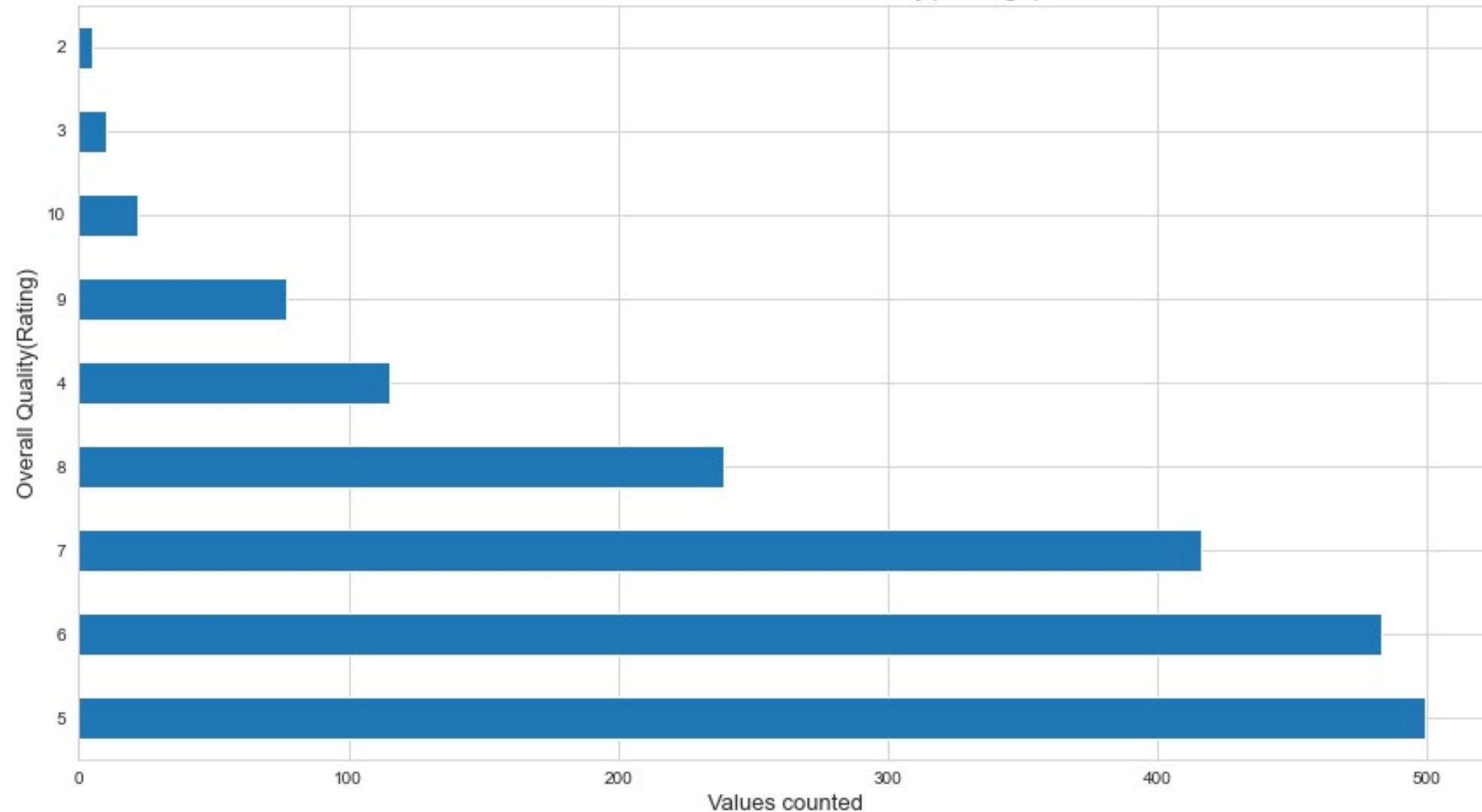
Above grade(ground) living area(sq.ft) vs. Sales Price



Age of home vs Sales price



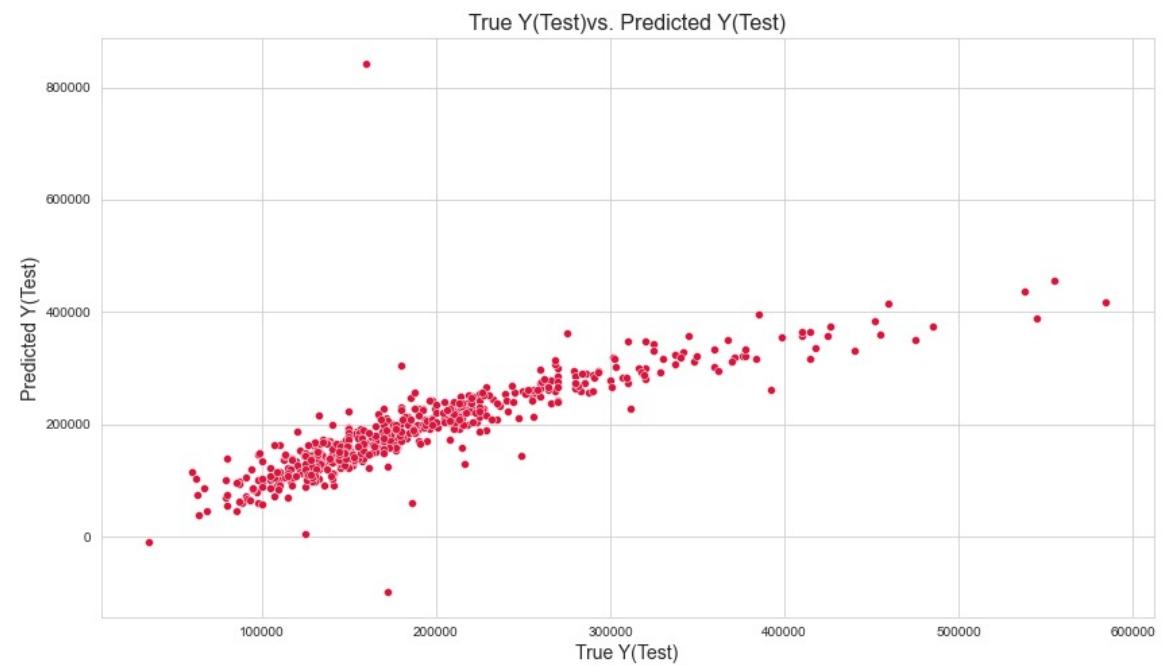
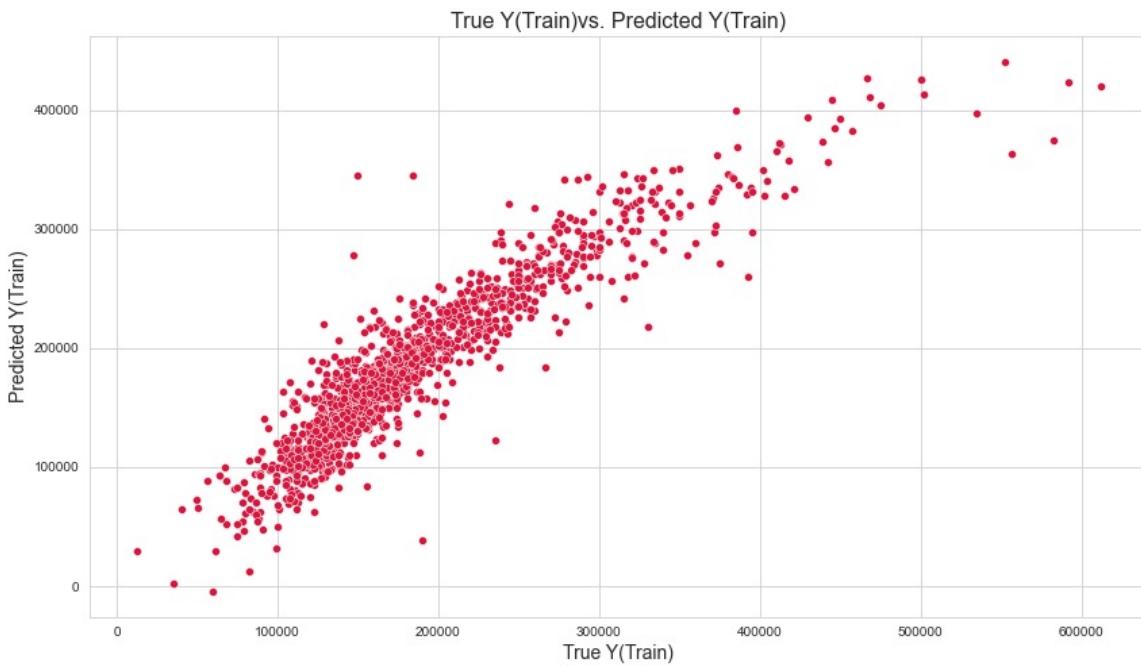
Value Counts for Overall Quality(Ratings)



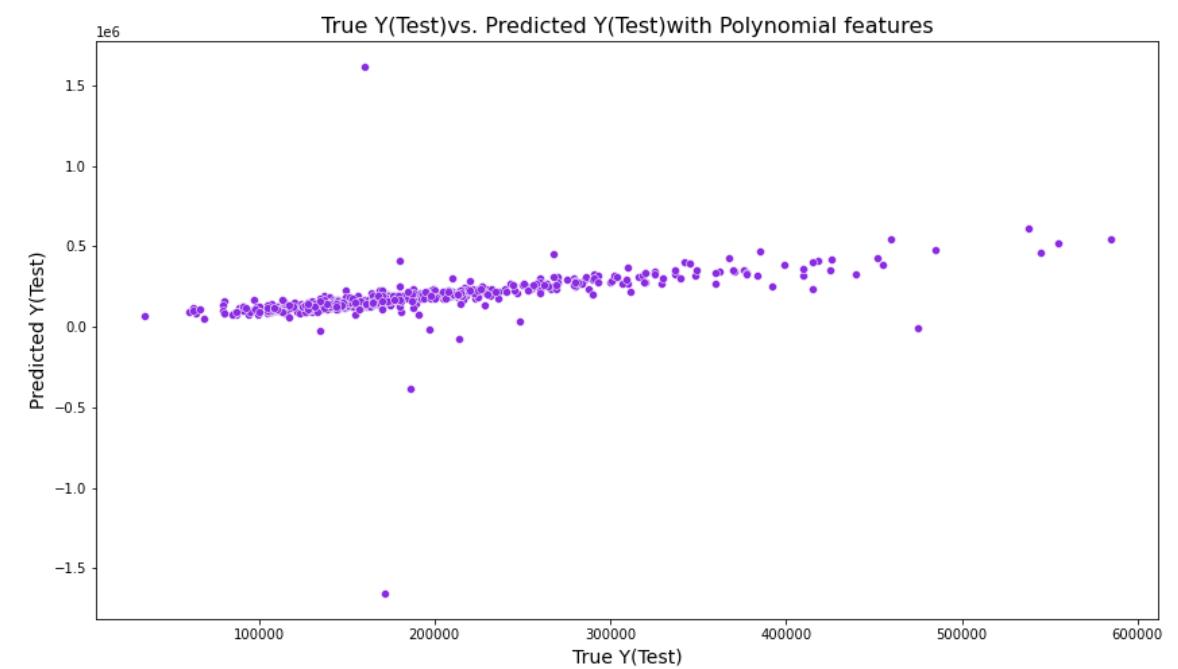
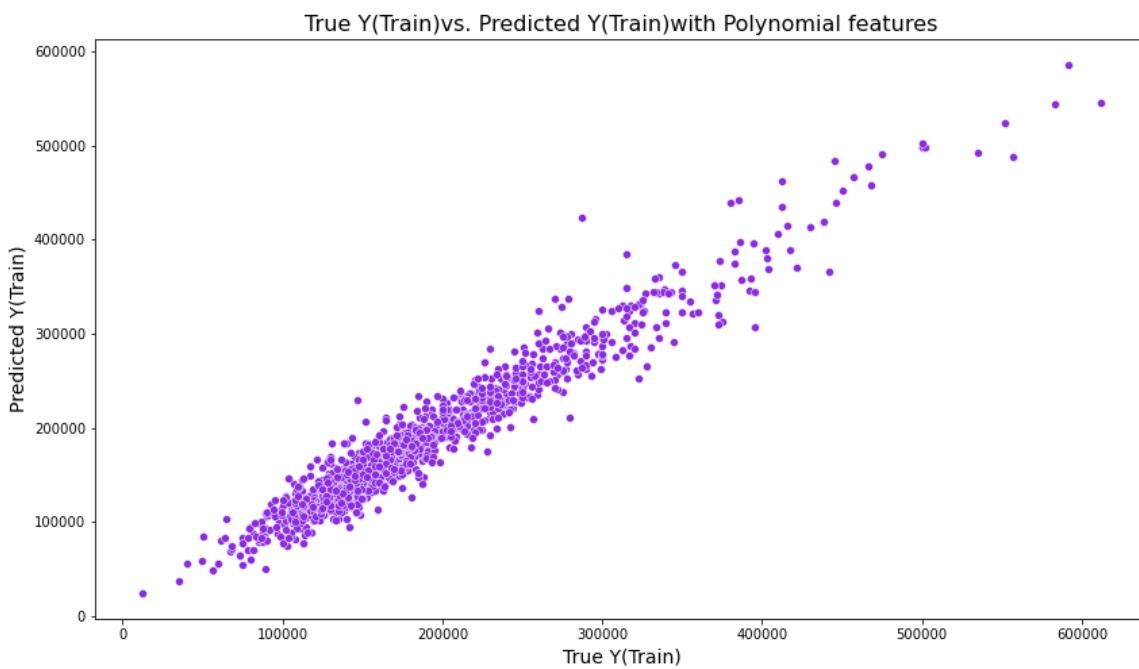
Model R2 and RMSE Scores

Model	Train RMSE	Test RMSE	Train R2	Test R2
Base_line	NA	NA	.80	.65
Model_1	29587.55	43739.88	.85	.72
Model_2	30098.74	42362.35	.85	.73
Model_3	18850.60	109767.34	.94	-.79
Model_4	27087.66	41485.18	.88	.74
Model_5(Lasso)	27087.68	41483.23	.88	.74
Model_6(ridge)	27270.32	40967.73	.87	.75
Model_7	31694.52	32307.37	.83	.83

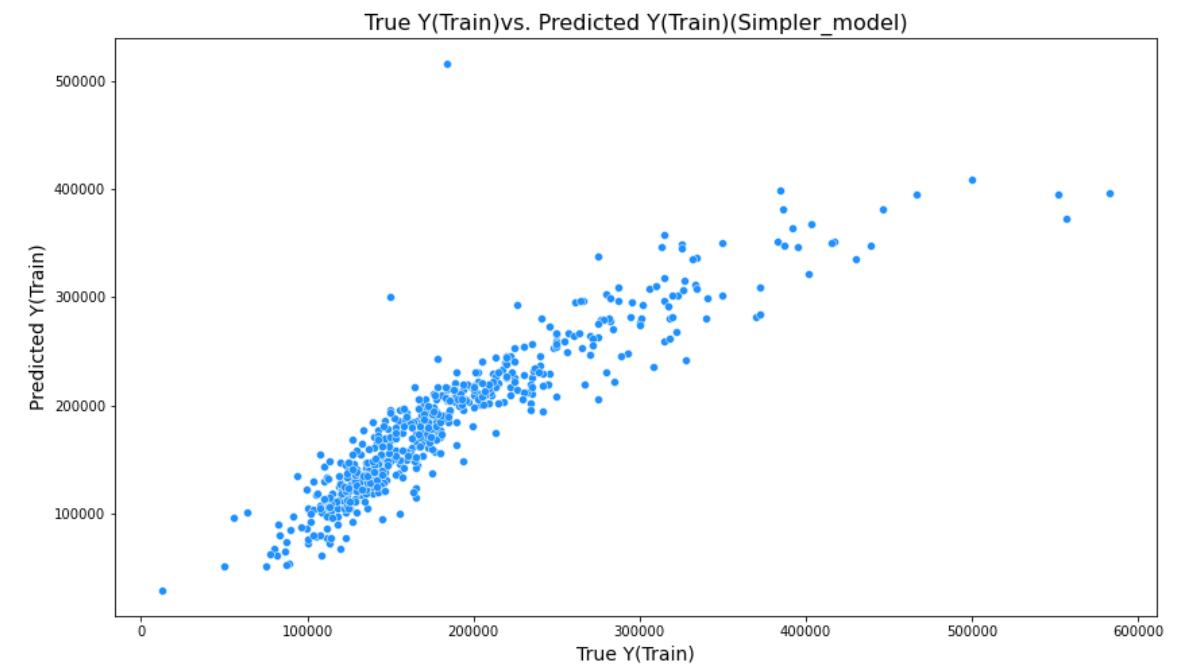
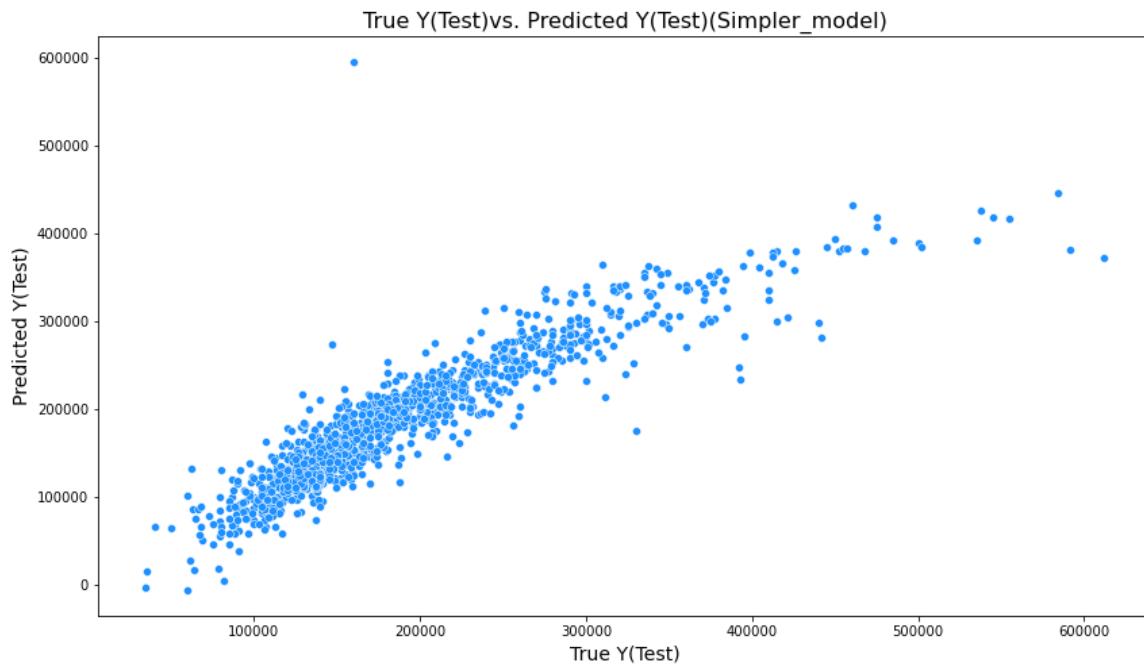
- Model 1 Linear Regression



- Model 3- Polynomial Features(no scaler)



- Model 7-(high correlated features/poly/scaled)





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 any features that had a poor correlation with sales price.
- It had an R² score of 83.41% which was an improvement over the baseline of the original model many iterations ago, that baseline had a score of 65.59%. That is a shade under 18-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.
- We also found which features impact pricing in positive and negative ways, which can help homeowners maximize their profits and appeal while minimizing their losses and time on the market.

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/>).