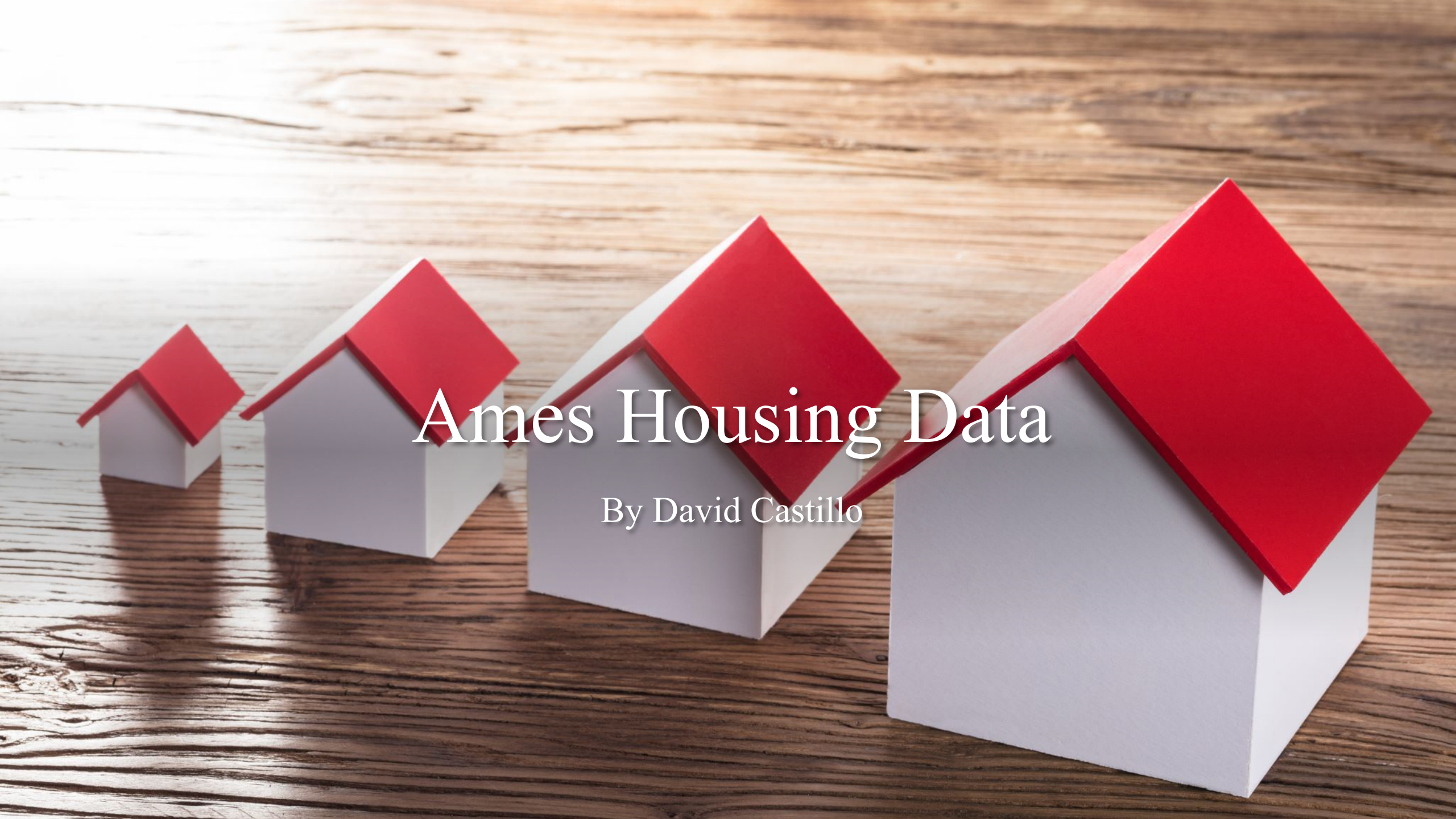


Ames Housing Data

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.

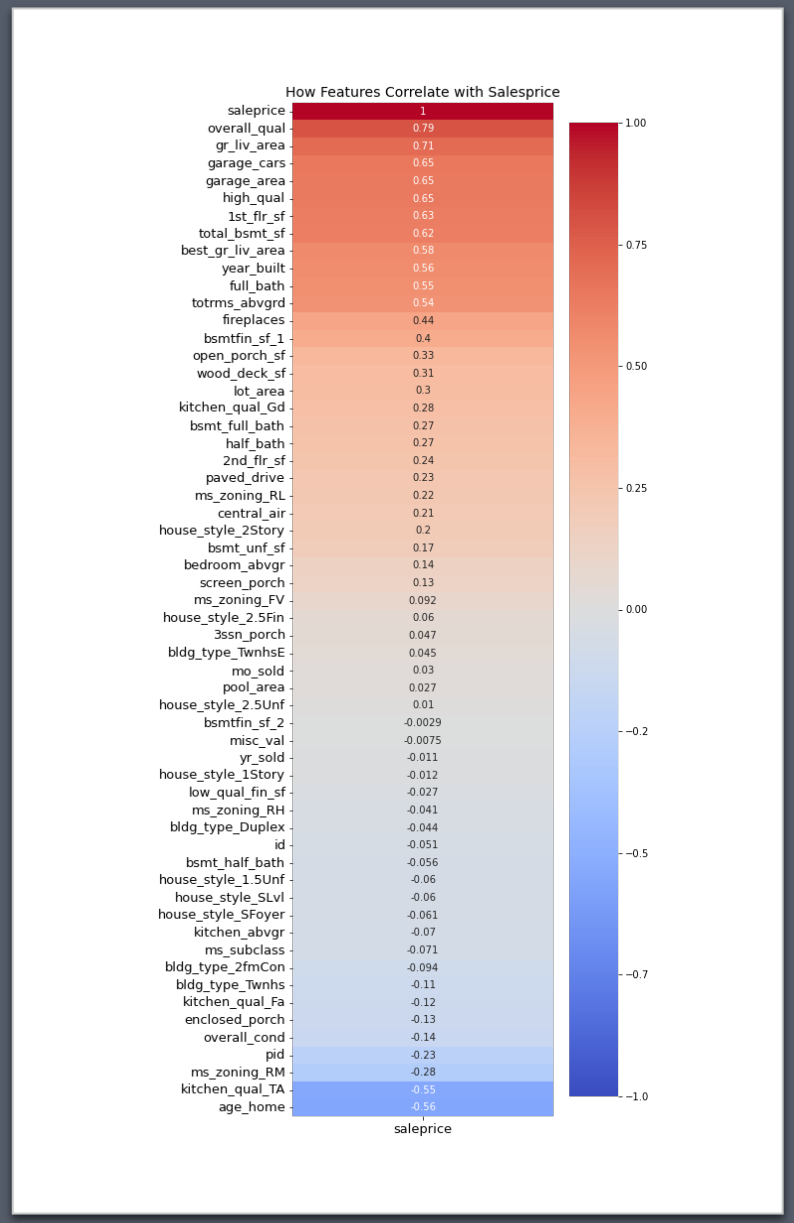
If you list the price of the home way above market value, you will miss out on prospective buyers(trulia.com).

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

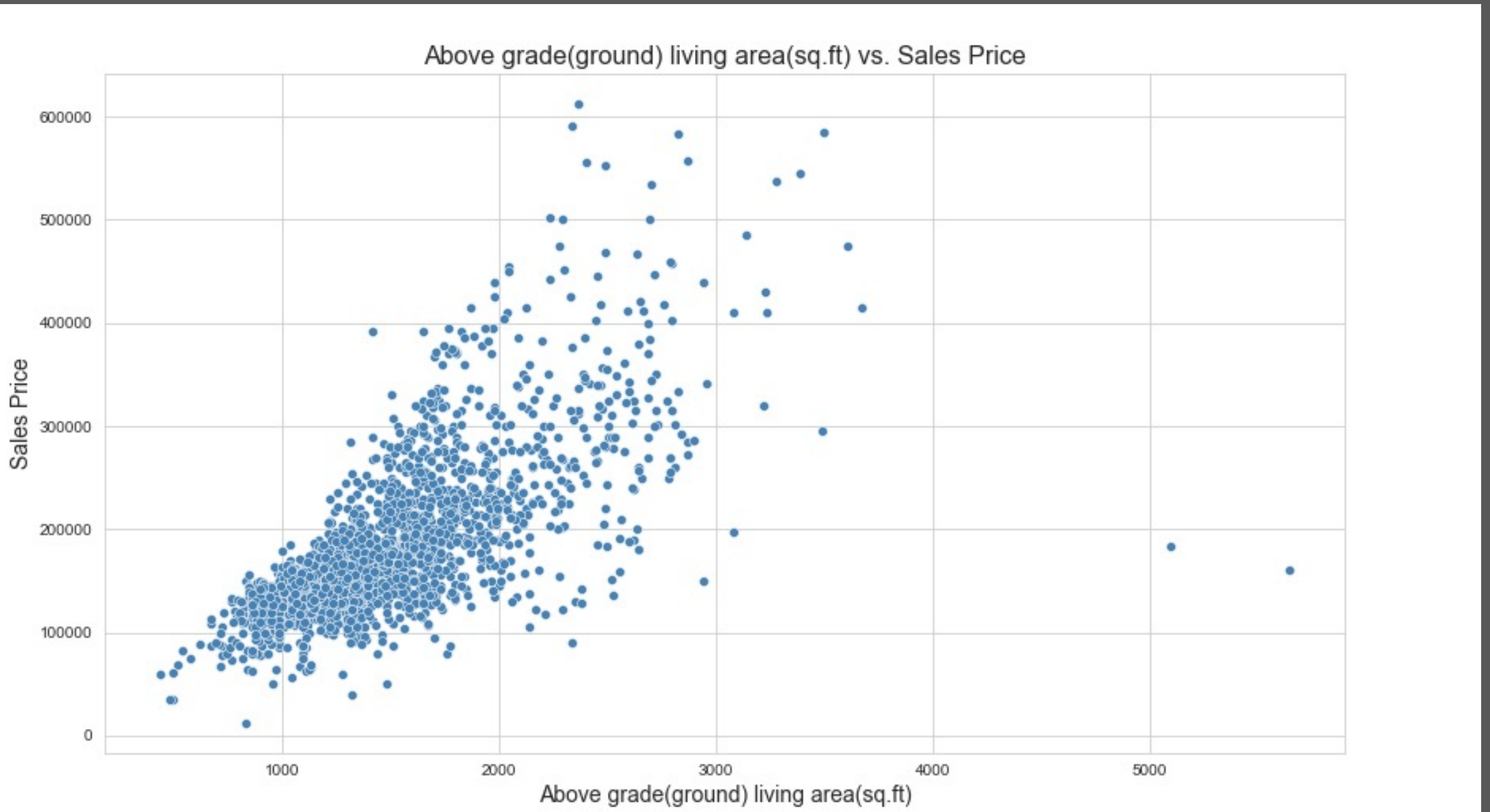


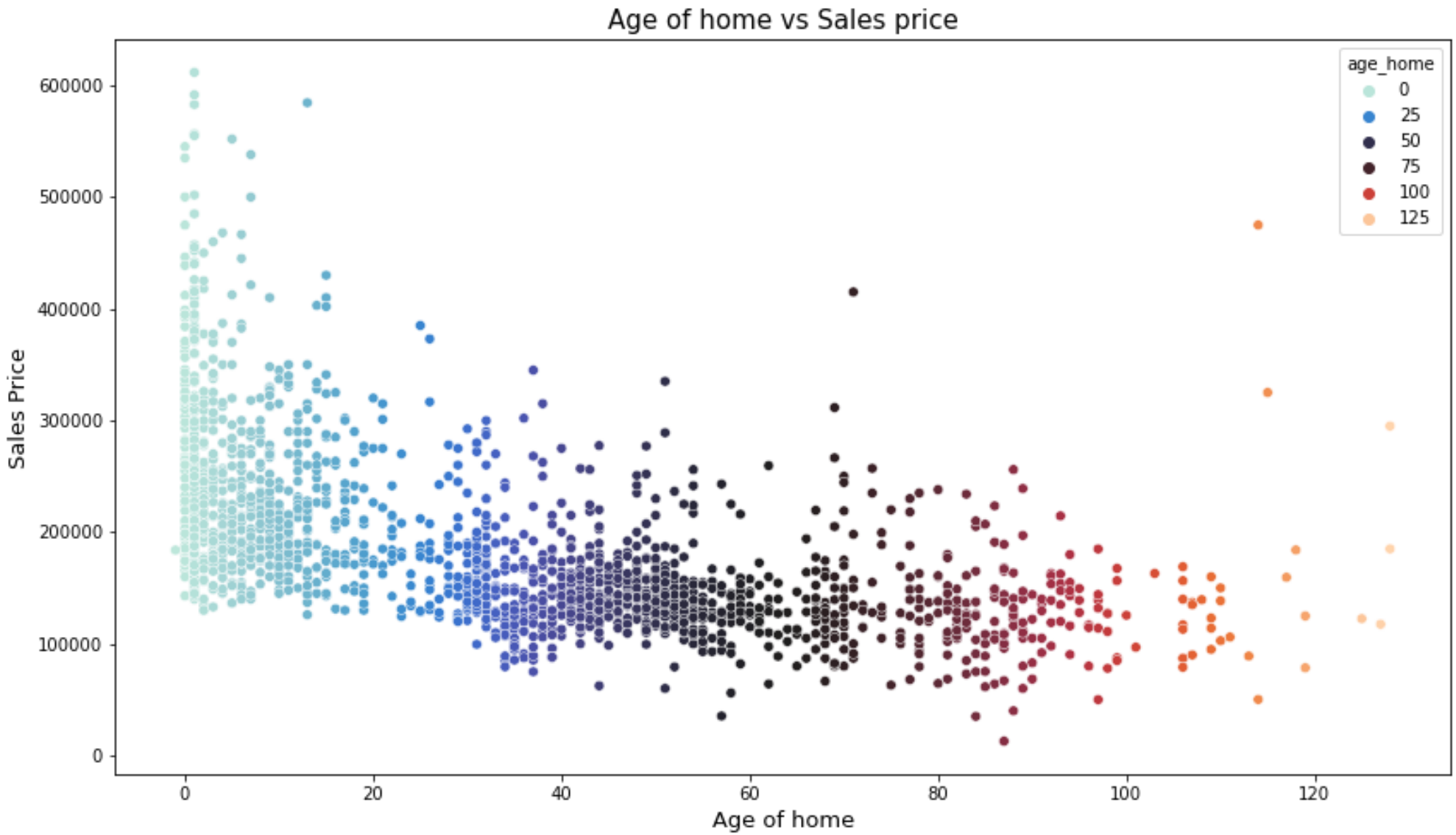
Top Features

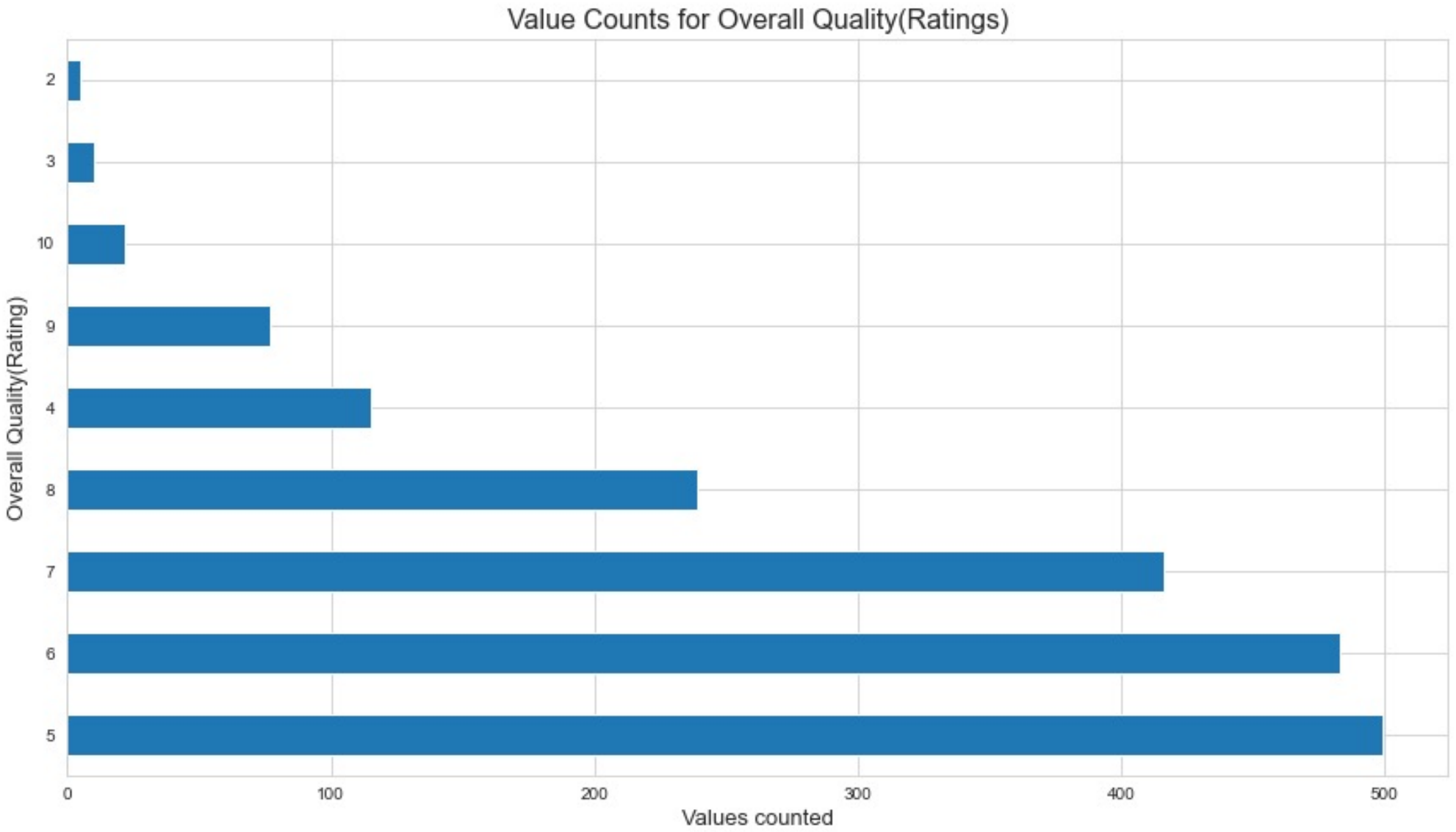
- Overall Quality
- Above grade living area(sq ft)
- Garage area

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
Sales Price
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Above grade(ground) living area(sq.ft)

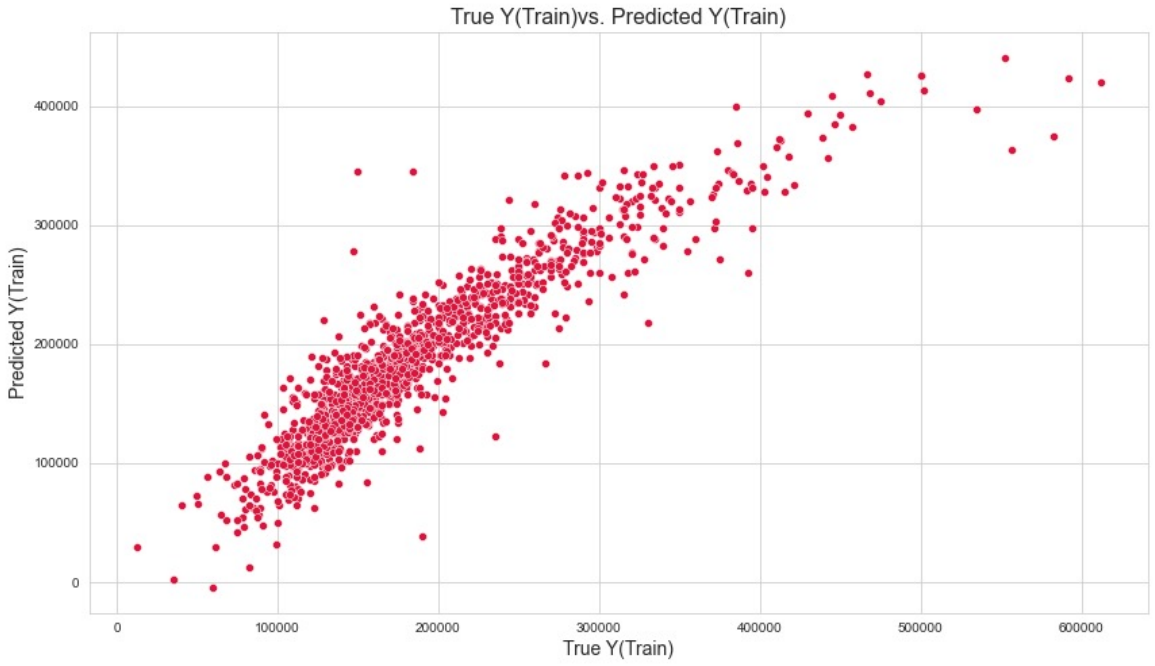

Age of home vs Sales price
Sales Price
Age of home
age_home
0
25
50
75
100
125
600000
500000
400000
300000
200000
100000
0
0
20
40
60
80
100
120

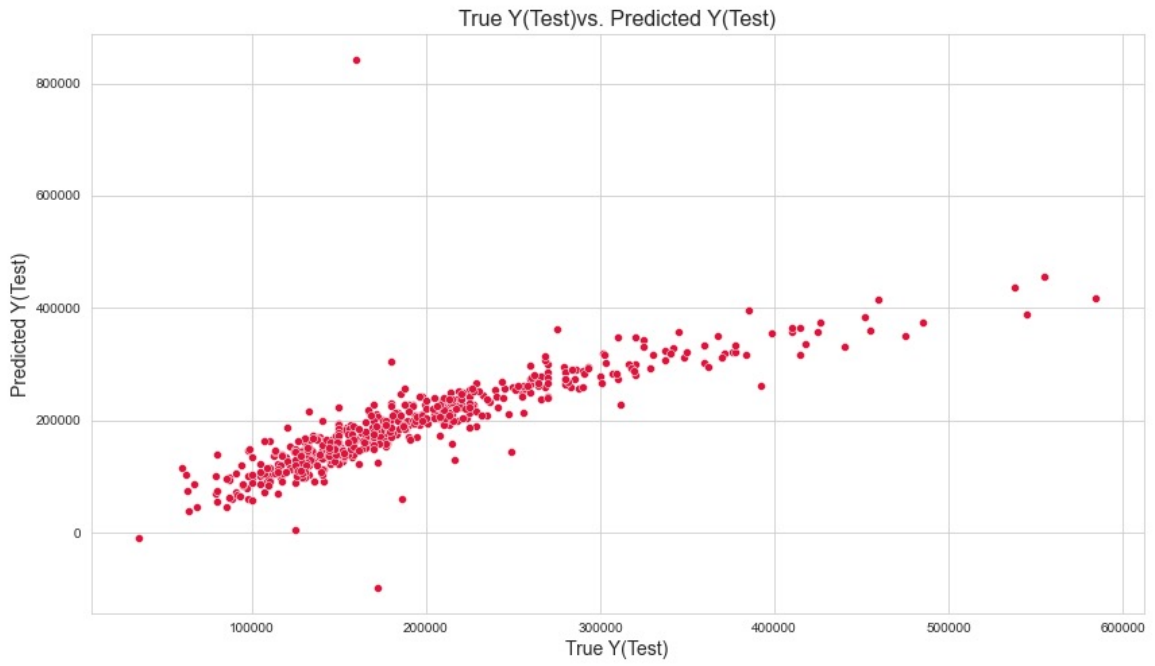

Value Counts for Overall Quality(Ratings)
Overall Quality(Rating)
2
3
10
9
4
8
7
6
5
0
100
200
300
400
500
Values counted

Model R2 and RMSE Scores

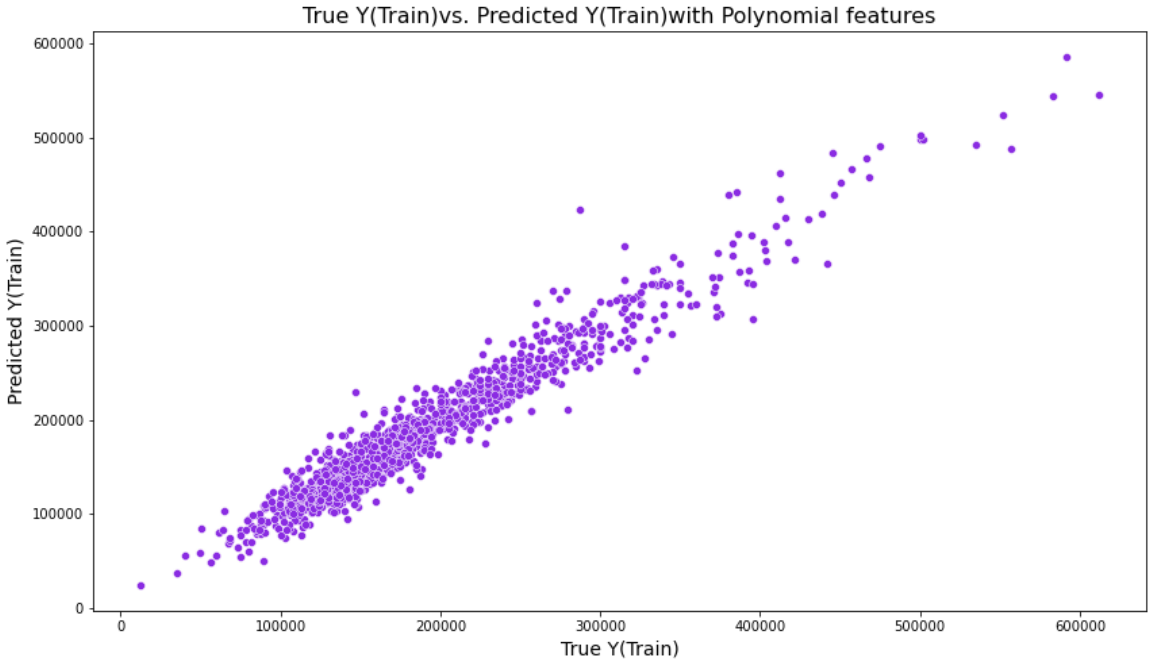
Model	Train RMSE	Test RMSE	Train R2	Test R2
Base_line	NA	NA	.80	.65
Model_1	29587.55	43739.88	.85	.72
Model_2	30098.74	42362.35	.85	.73
Model_3	18850.60	109767.34	.94	-.79
Model_4	27087.66	41485.18	.88	.74
Model_5(Lasso)	27087.68	41483.23	.88	.74
Model_6(ridge)	27270.32	40967.73	.87	.75
Model_7	31694.52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

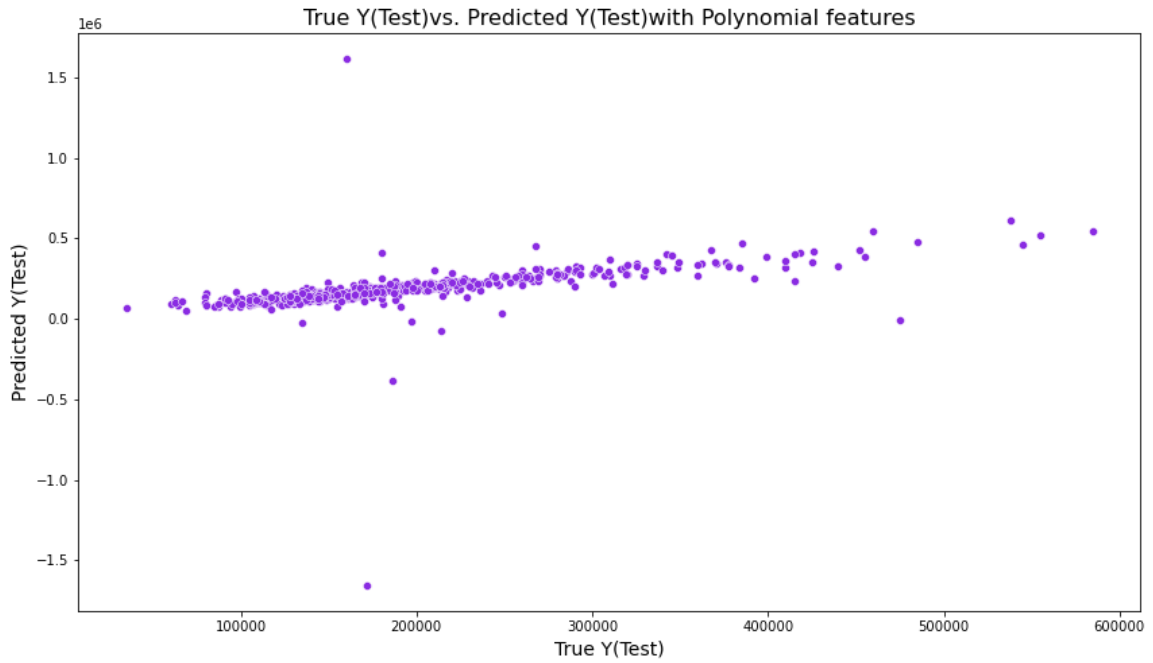
- Model 1 Linear Regression



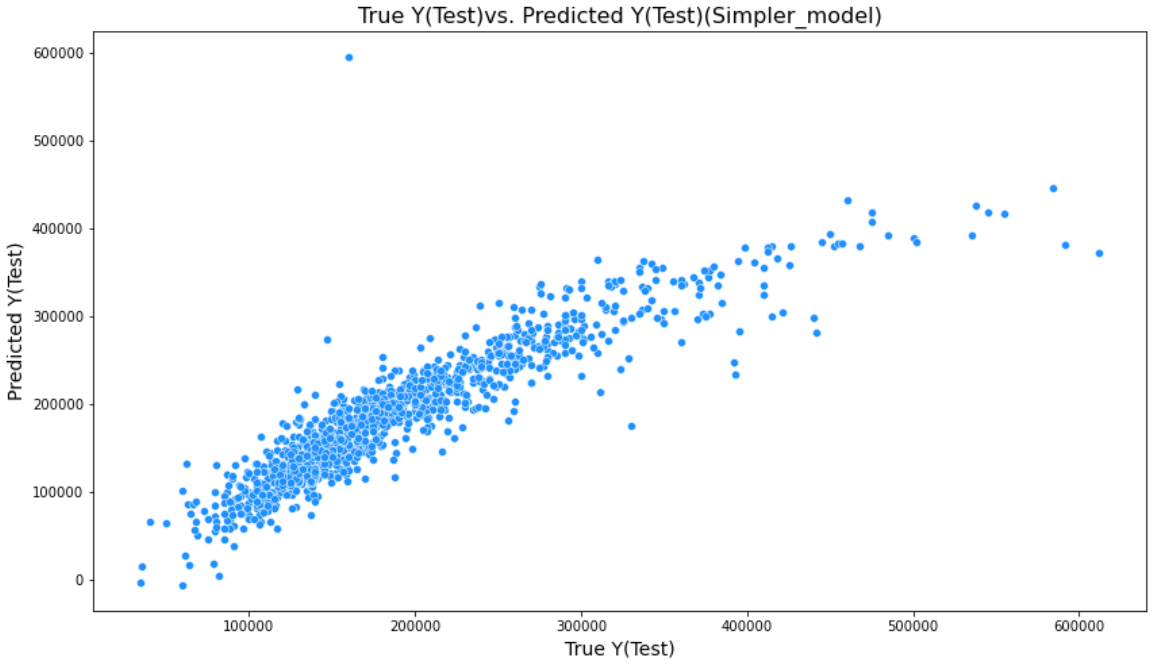


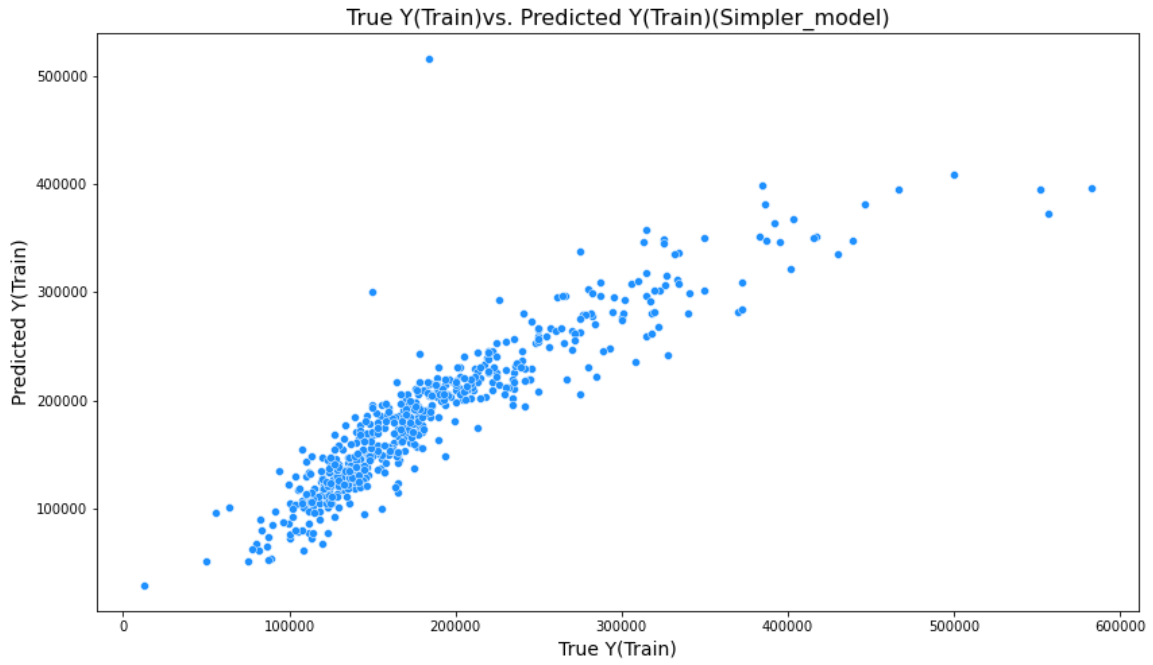
- Model 3- Polynomial Features(no scaler)

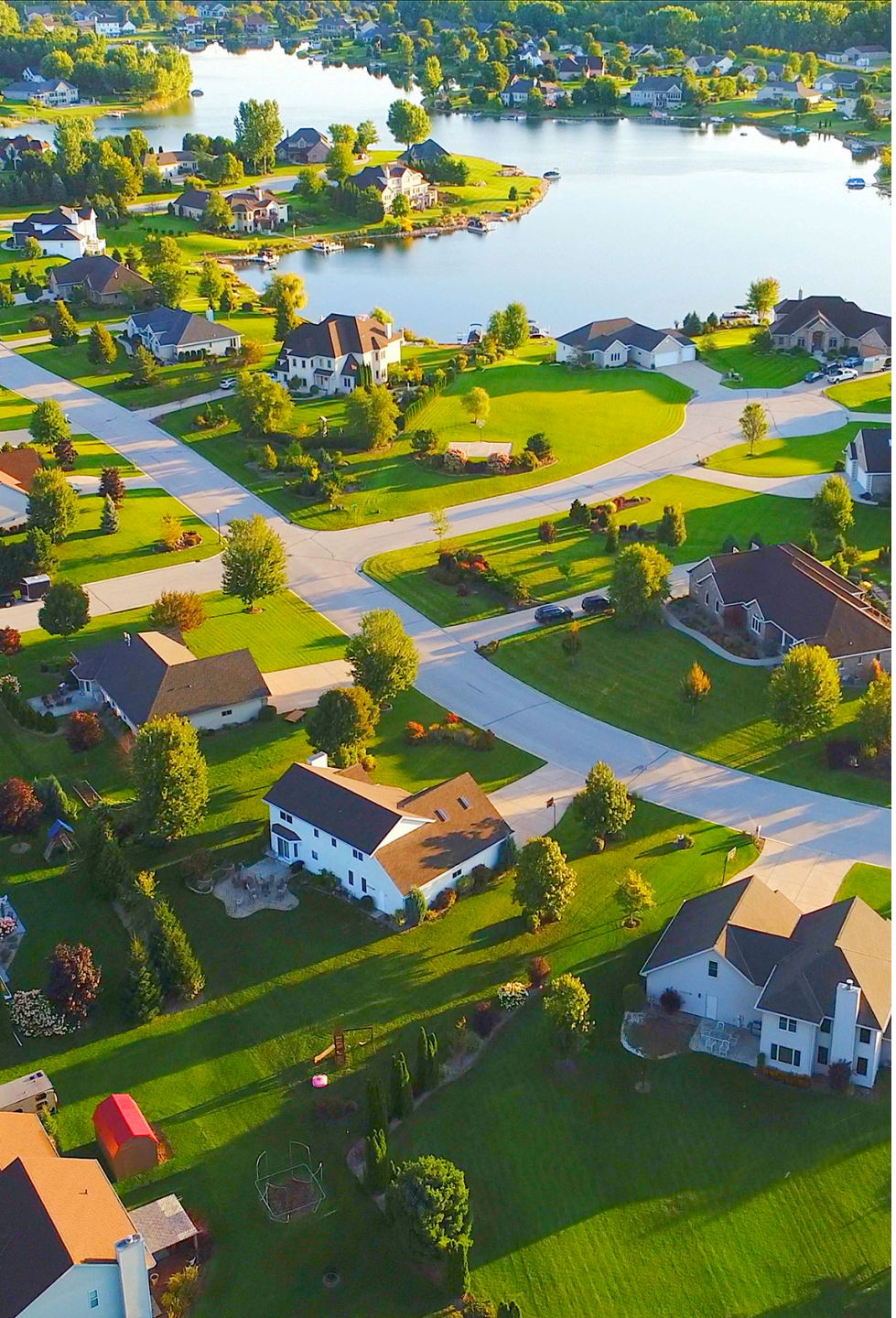




- Model 7-(high correlated features/poly/scaled)







Conclusion

- The linear regression model outperformed the baseline model by whittling down any features that had a poor correlation with sales price. Used techniques such as polynomial features and standard scaler.
- R2 score improved from baseline of 65.59% to 83.41%(17.82-point improvement),in other words 84.44% of the variability in our sales price 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/).