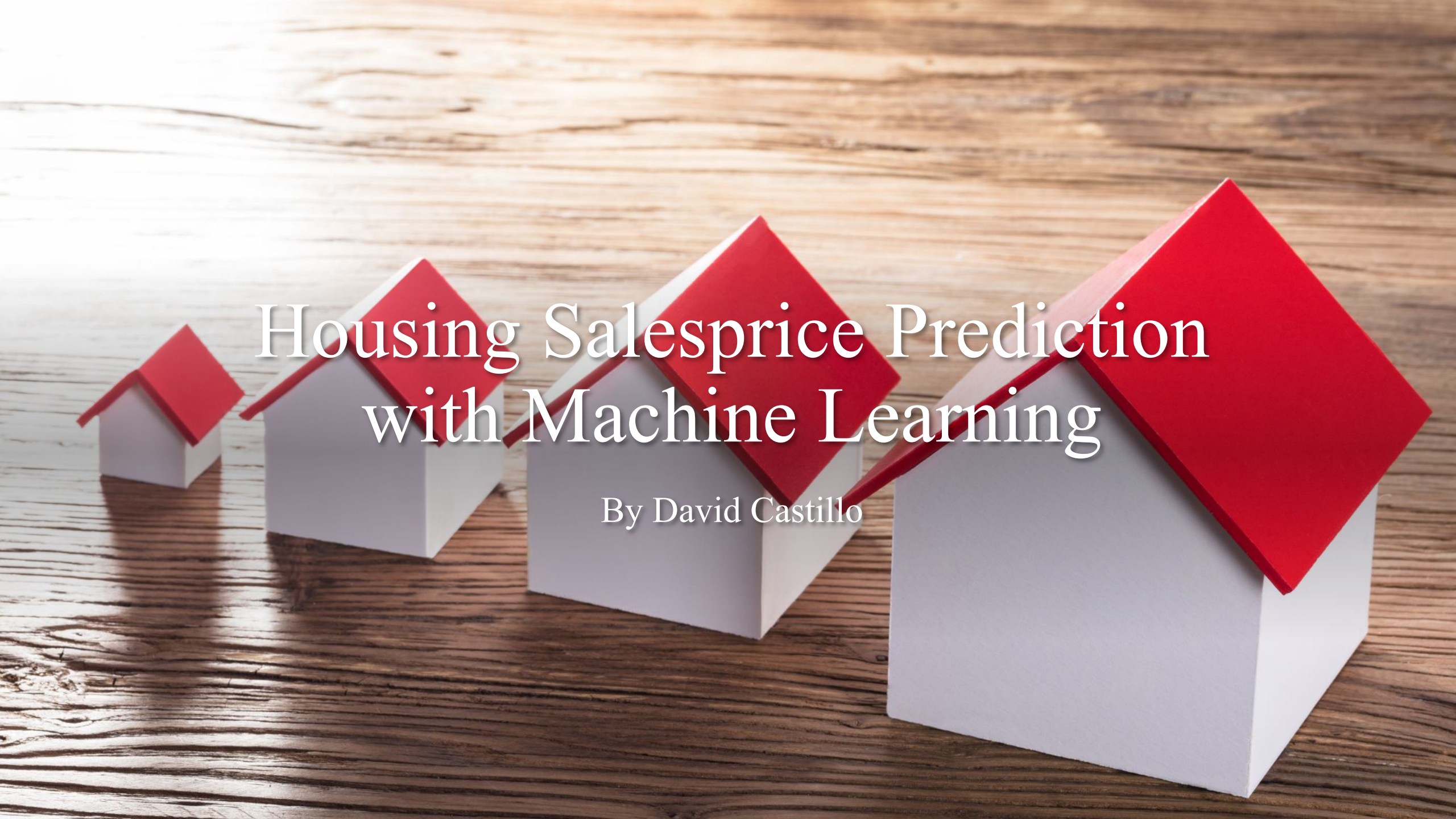


Housing Salesprice Prediction with Machine Learning

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 important 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 From Research

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).

Processing the data/Feature Engineering after revisiting the Data

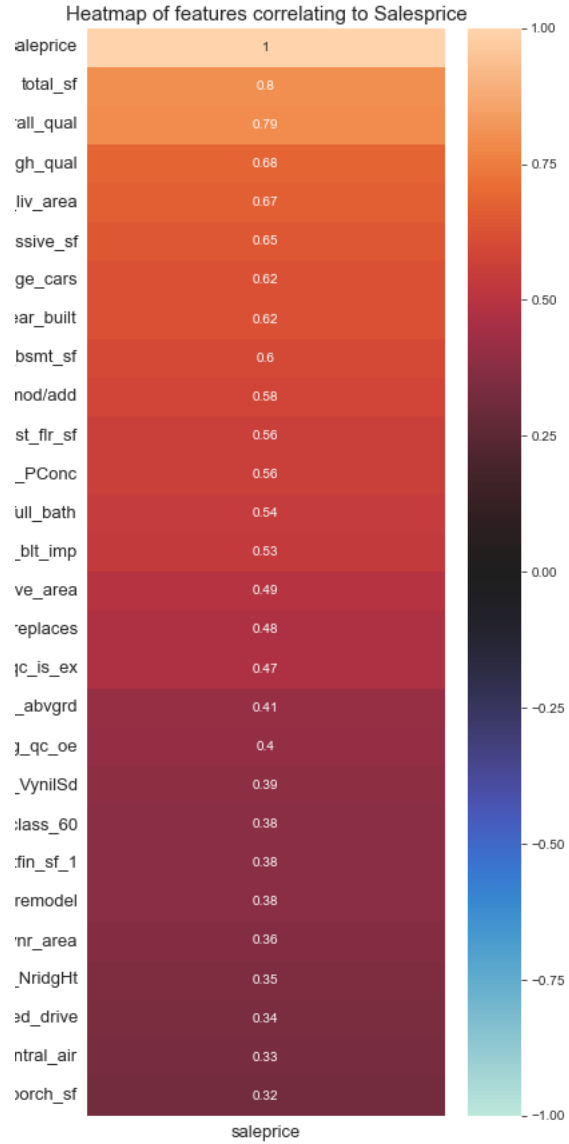
- Dealt with nulls by using iterative imputation and simple imputation from the Scikit-Learn library.
- Dealt with outliers by using statistical approach(1.5 Interquartile range to establish cutoff points)
- Ordinal columns were processed through ordinal encoding, get dummies for rest of categorical columns, and binarized remaining 2 columns
- Kept it DRY!

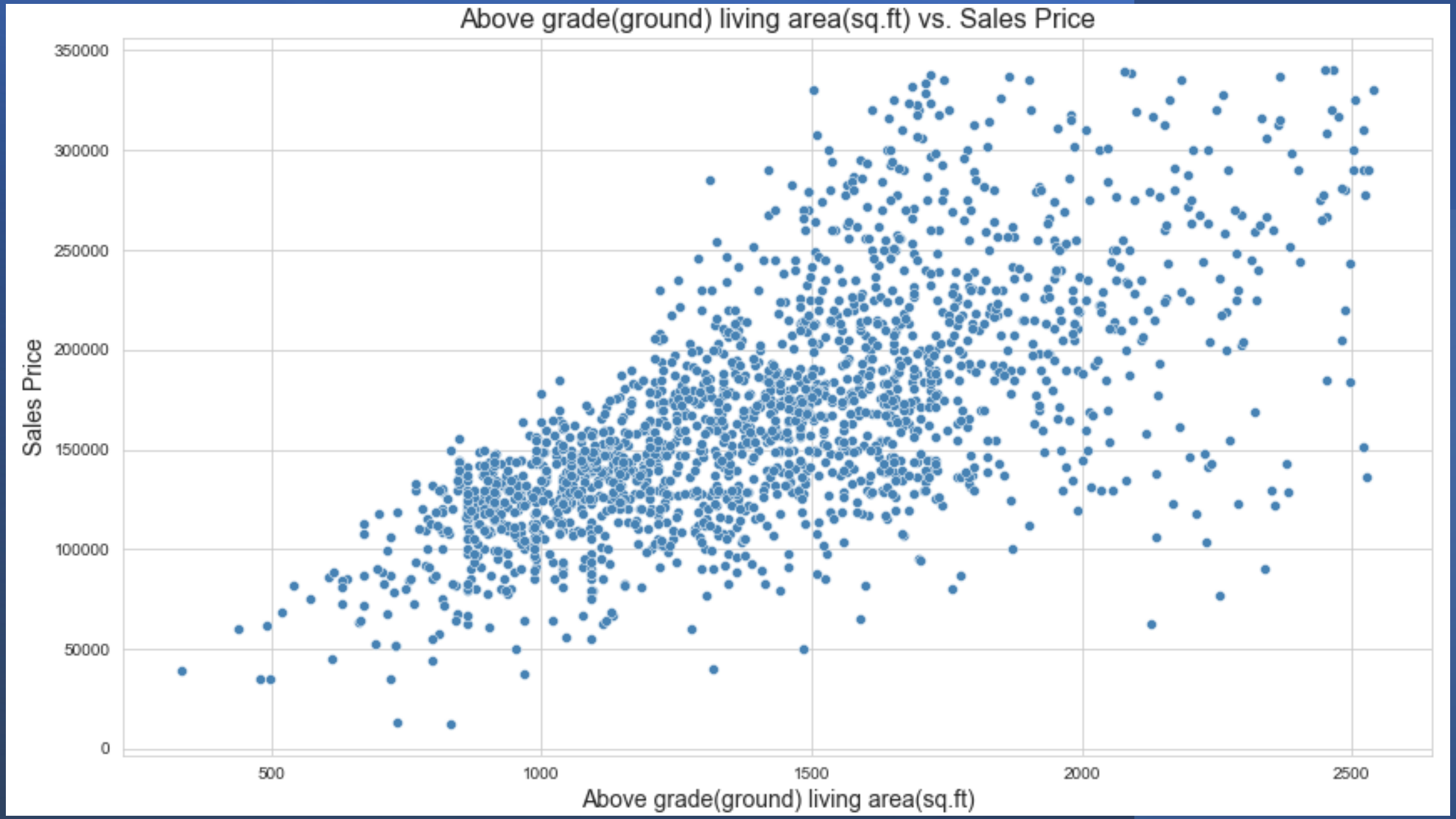
The Right Model

- R2 Score
- RMSE
- Adjusted R2
- Coefficients
- Bias-Variance Tradeoff

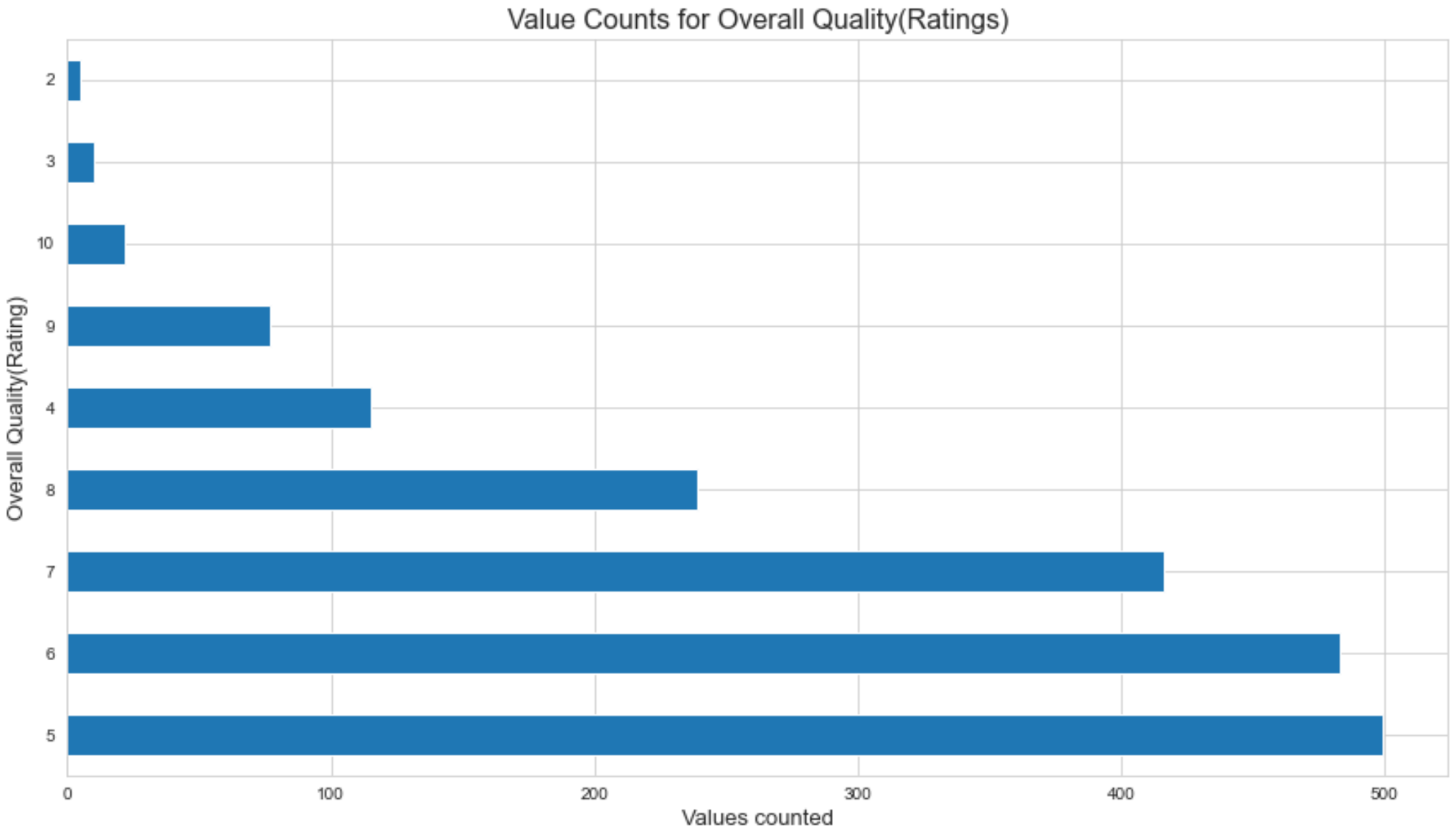
Top Features

- Total Square footage
- Overall Quality
- High Quality
- Above Grade Living Area
- Massive Square Footage
- Garage Cars
- Year Built

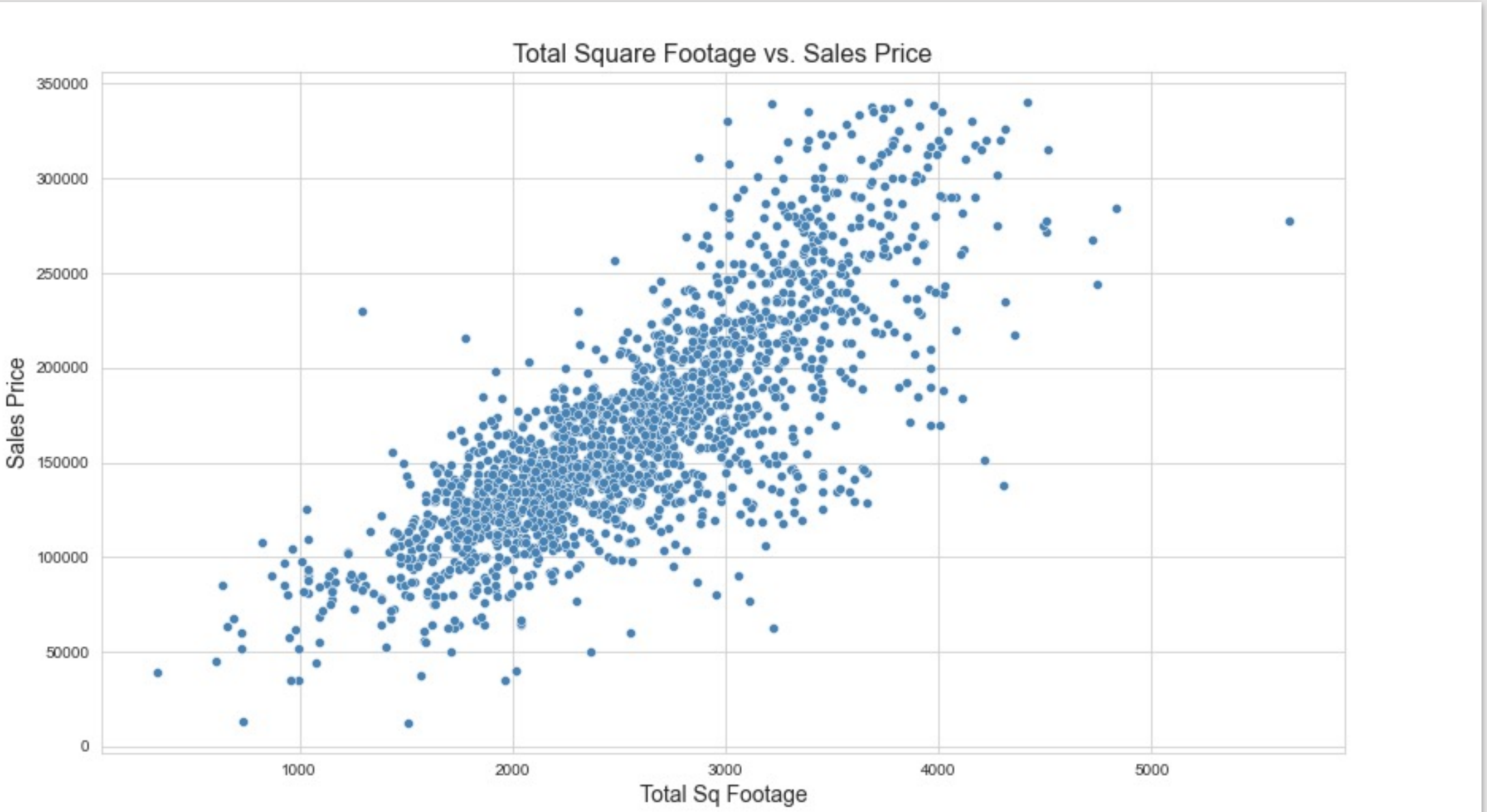



Above grade(ground) living area(sq.ft) vs. Sales Price
Sales Price
350000
300000
250000
200000
150000
100000
50000
0
500
1000
1500
2000
2500
Above grade(ground) living area(sq.ft)


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


Value counts for Houses Built By Year
Count of houses built
0
20
40
60
80
Year
1885 2010 1913 1875 1872 1908 1879 1898 1911 1896 1880 1901 1905 1919 1917 1895 1918 1982 1912 1932 1931 1981 1934 1890 1985 1989 1991 1914 1983 1942 1938 1986 1937 1946 1987 1928 1935 1927 1929 1947 1921 1916 1923 1936 2009 1922 1984 1952 1945 1924 1974 1988 1949 1926 1951 1973 1990 1979 1939 1969 1980 1915 1992 1930 1941 1997 2008 1948 1953 1995 1900 1925 1975 2001 1972 1957 1993 1994 1960 1961 1965 1955 1964 1963 1996 1962 1966 1971 1910 1967 1970 1950 1956 1954 1940 2002 1959 1978 1968 1998 2000 1977 1958 2007 2004 2006 2005


Total Square Footage vs. Sales Price
Sales Price
350000
300000
250000
200000
150000
100000
50000
0
1000
2000
3000
4000
5000
Total Sq Footage

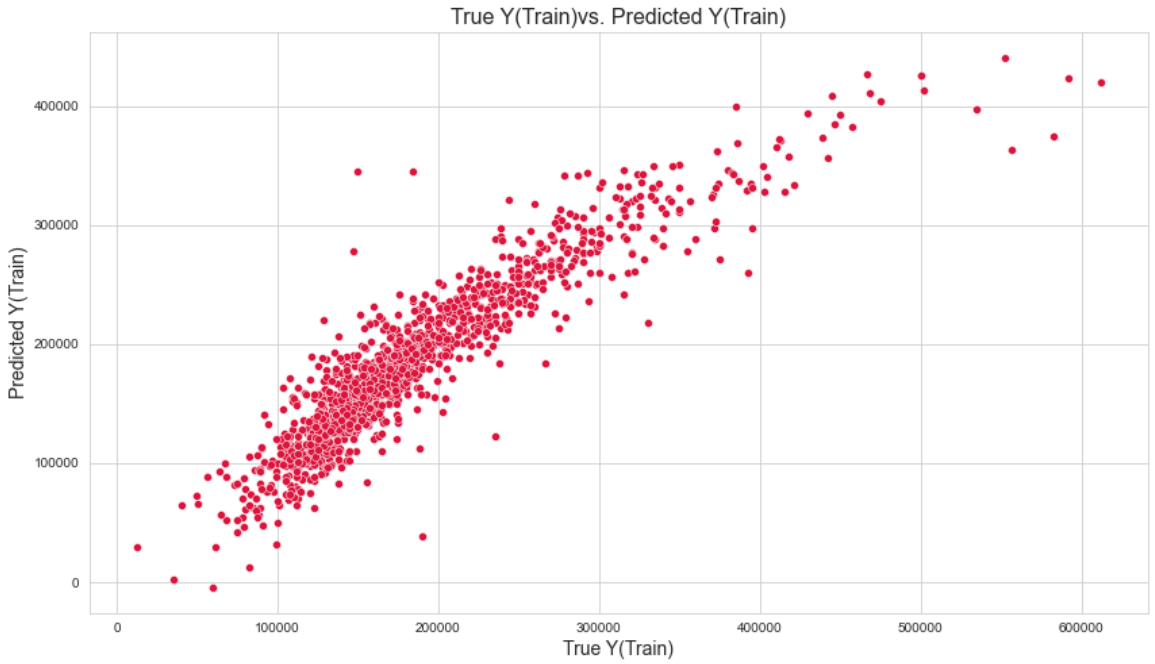
Initial Model R2 and RMSE Scores

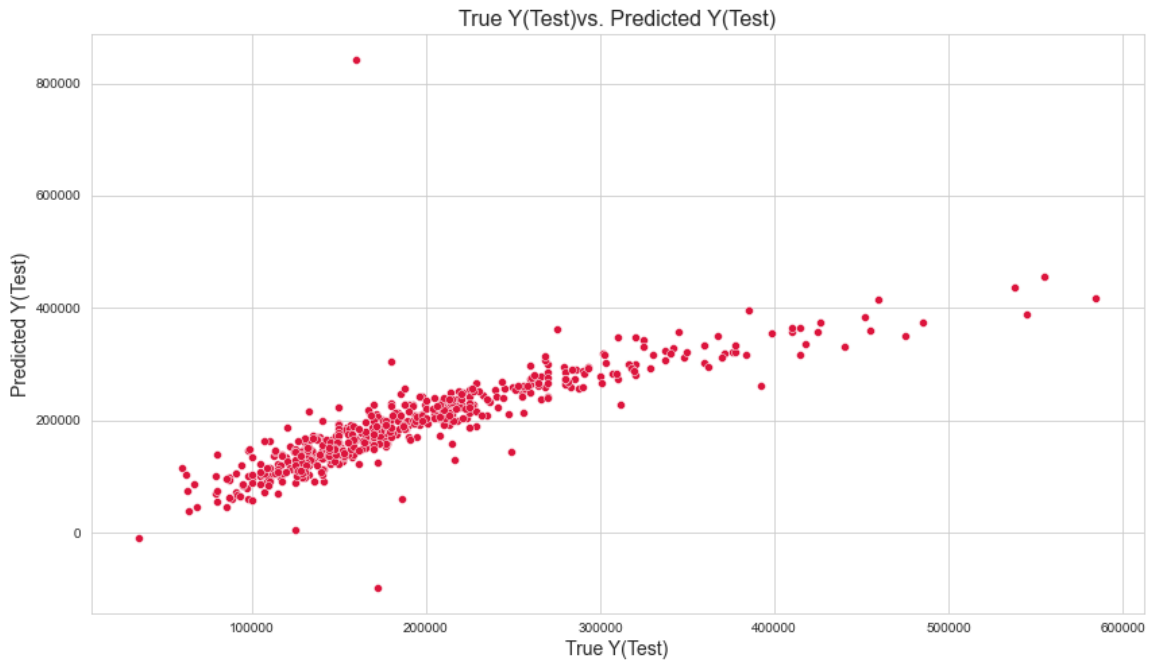
Model	Train RMSE	Test RMSE	Train R2	Test R2
Base_line	NA	NA	.80	.65
Model_1	29587.55	43739.88	.85	.72
Model_2	30098.74	42362.35	.85	.73
Model_3	18850.60	109767.34	.94	-.79
Model_4	27087.66	41485.18	.88	.74
Model_5(Lasso)	27087.68	41483.23	.88	.74
Model_6(ridge)	27270.32	40967.73	.87	.75
Model_7	31694.52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

MSE, RMSE, R2, Adjusted R2
For latest iteration of Models

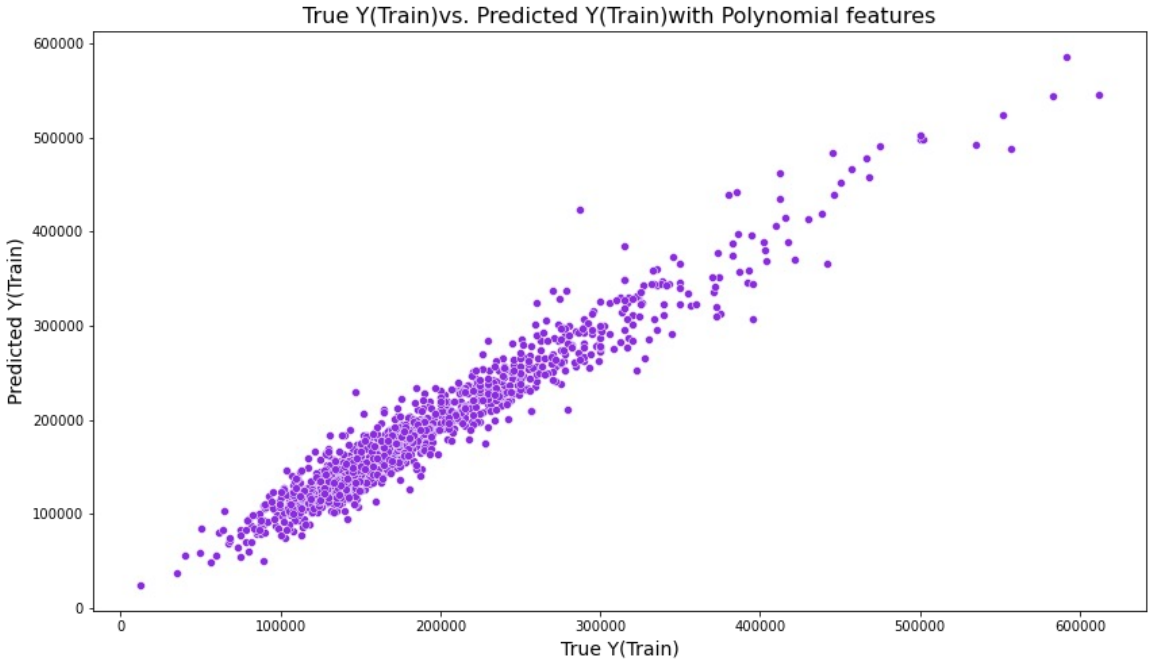
Model	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Adjusted R2
Base_line Cross validation mean	NA	NA	NA	NA	.80	.65	
Model_8 Correlated features > .3	454556326.22	426113318.24	21320.33	20642.51	.87	.87	.92
Model_9 Correlated features > .25	408688775.06	397568963.95	20216.05	19939.13	.88	.88	.92
Model_10 Correlated features > .20	385480212.23	395821216.32	19633.65	19895.26	.89	.88	.92
Model_11 Correlated features > .15	375816013.37	368520569.34	19385.97	19196.89	.89	.89	.92
Model_12 Correlated features > .10	349035246.53	337097661.46	18682.49	18360.22	.9	.9	.92
Model_13 Correlated features > .05	321993966.47	338455267.02	17944.19	18397.15	.9	.9	.92
Model_14 Correlated features >-10	279402217.91	**323432630.71**	16715.33	**17984.23**	.92	**.9**	.92
Model_15 Correlated features >-25	247087410.31	325002715.87	15719.01	18027.83	.93	.9	.92

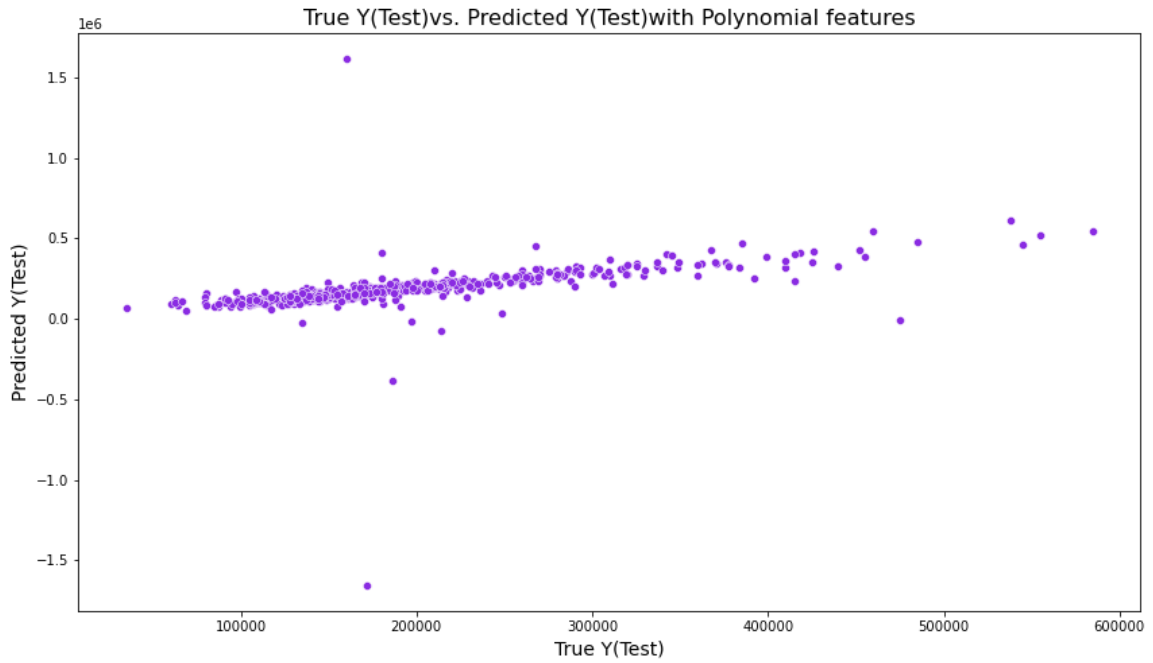
- Model 1 Linear Regression



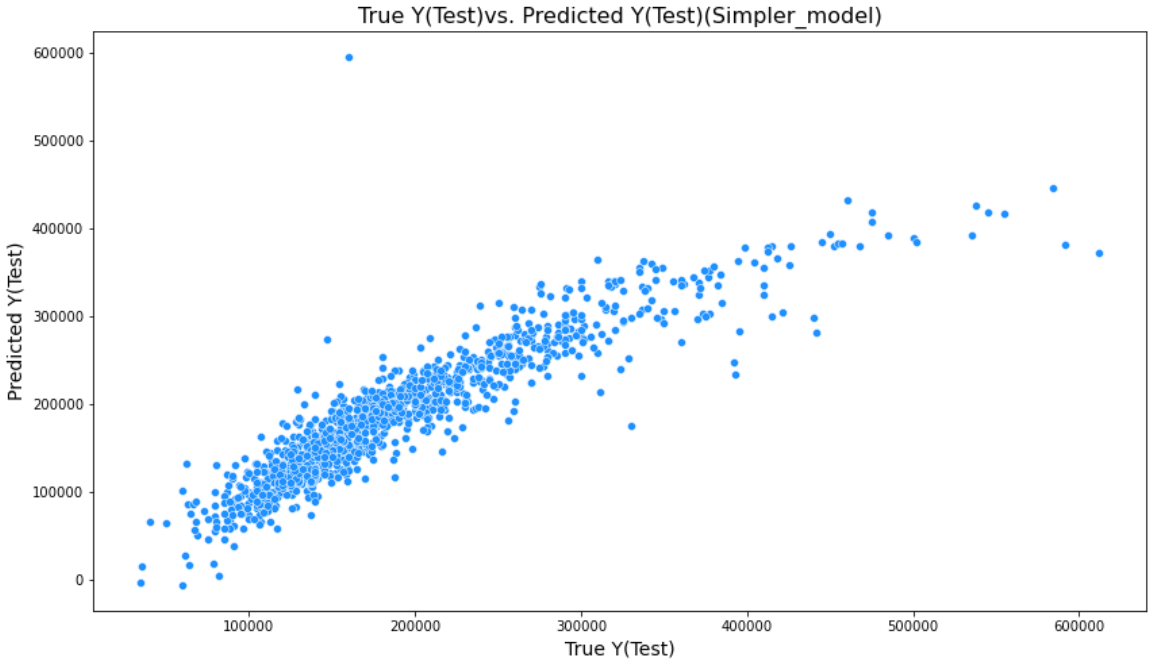


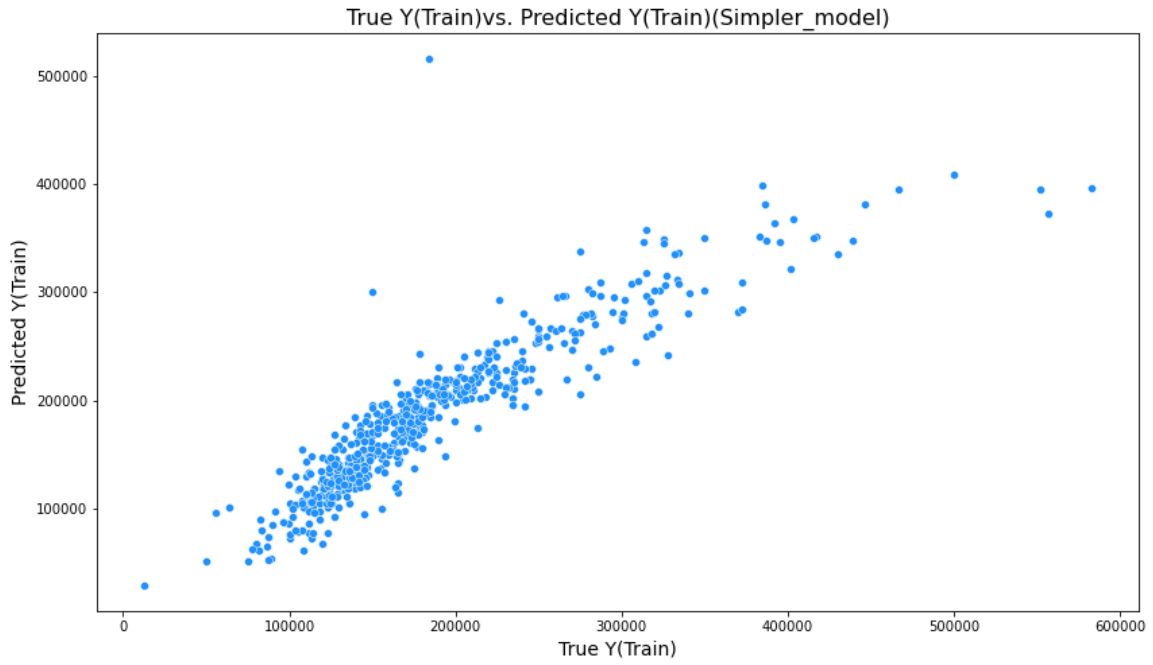
- Model 3- Polynomial Features(no scaler)



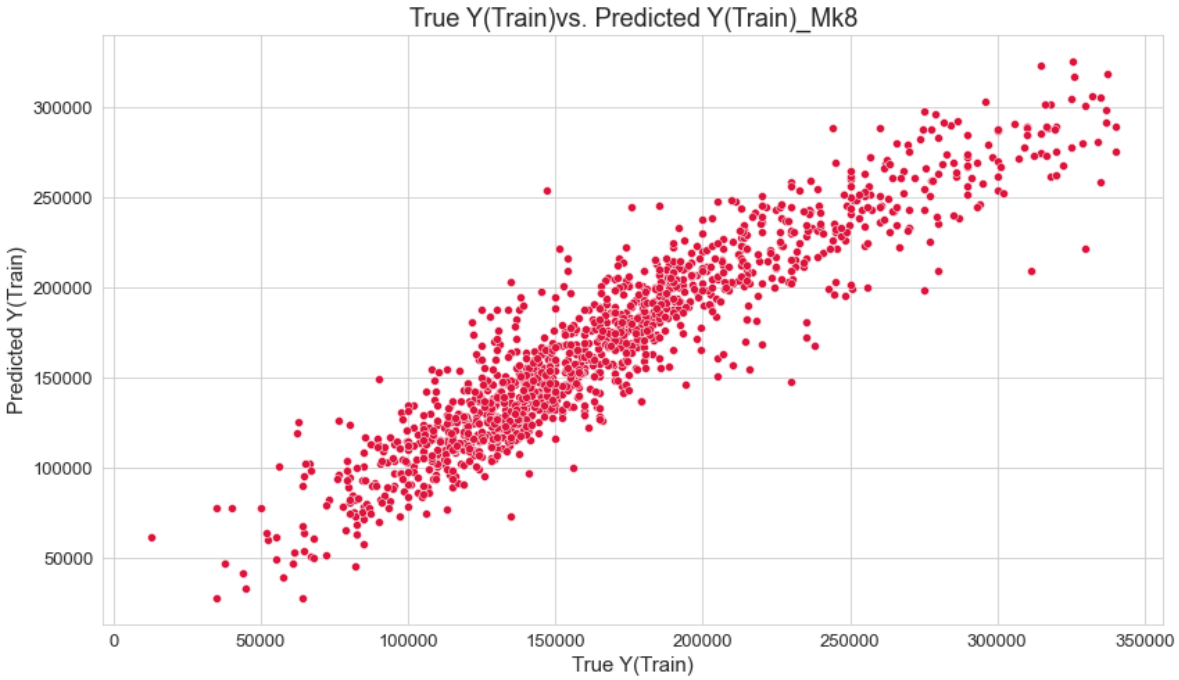


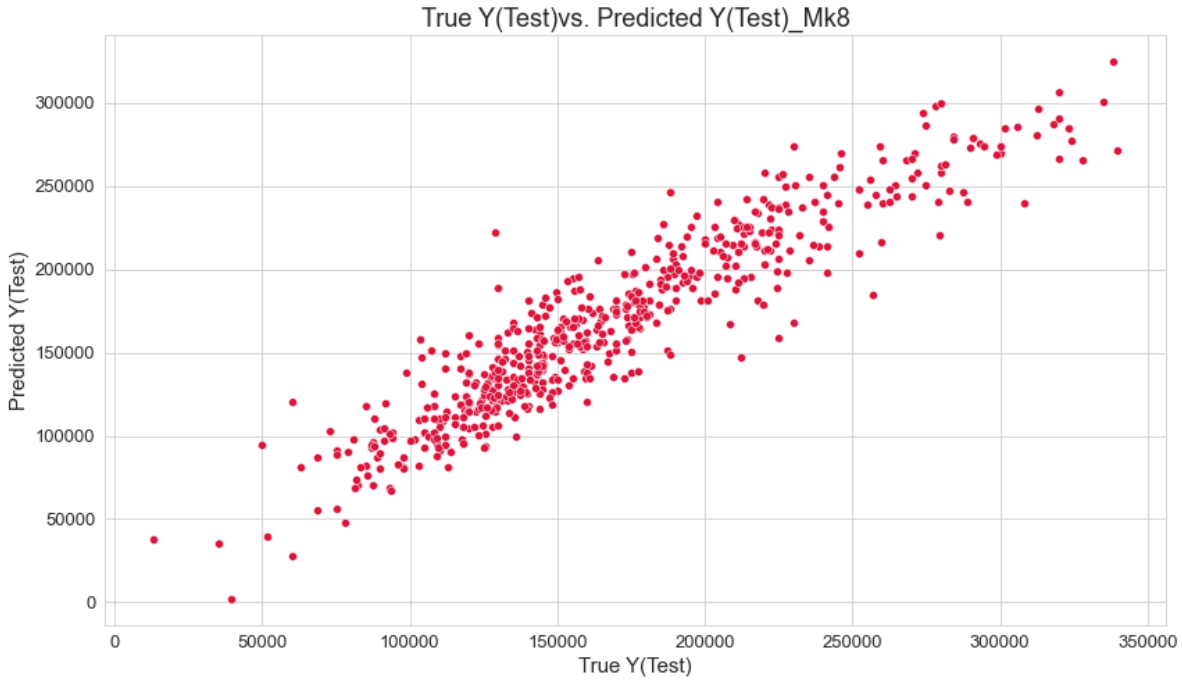
- Model 7-(high correlated features/poly/scaled)



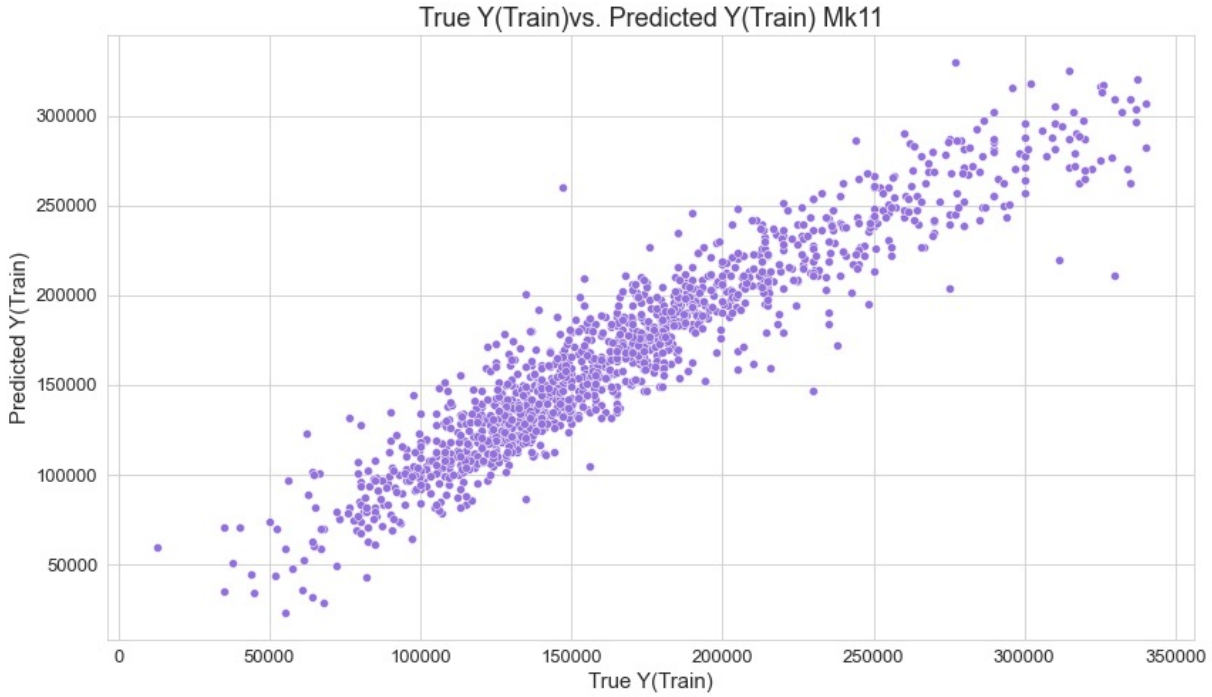


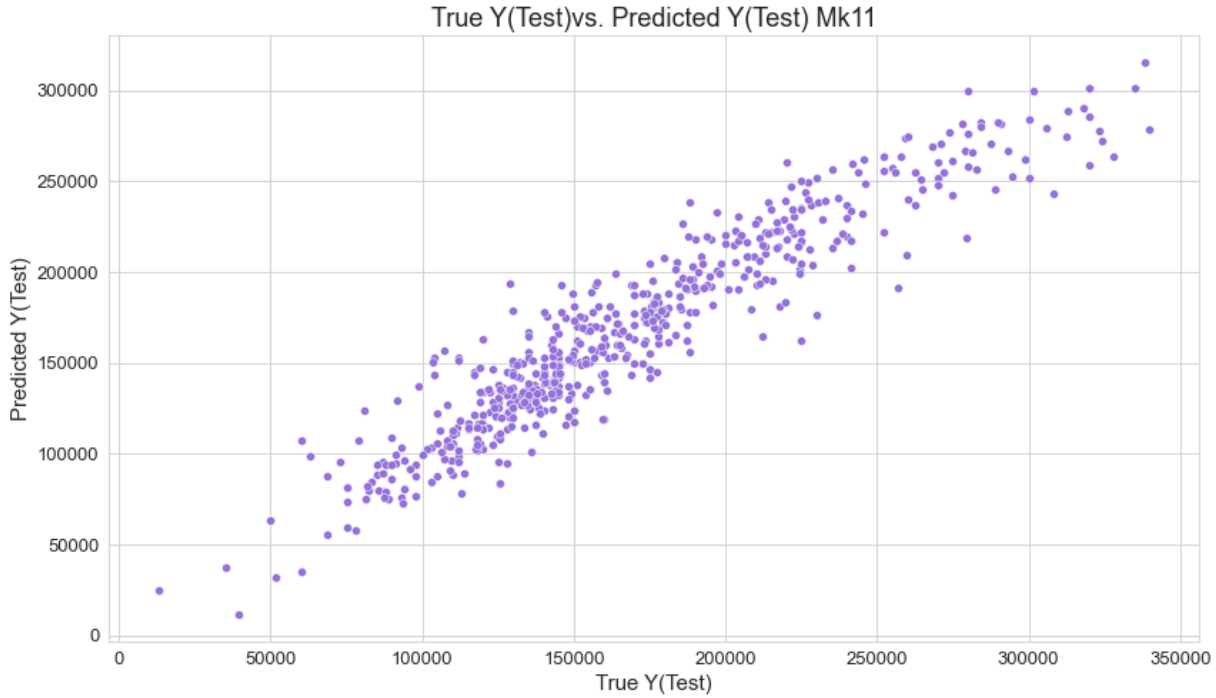
Model 8 (1st model post revisiting project) Correlated Features >.3



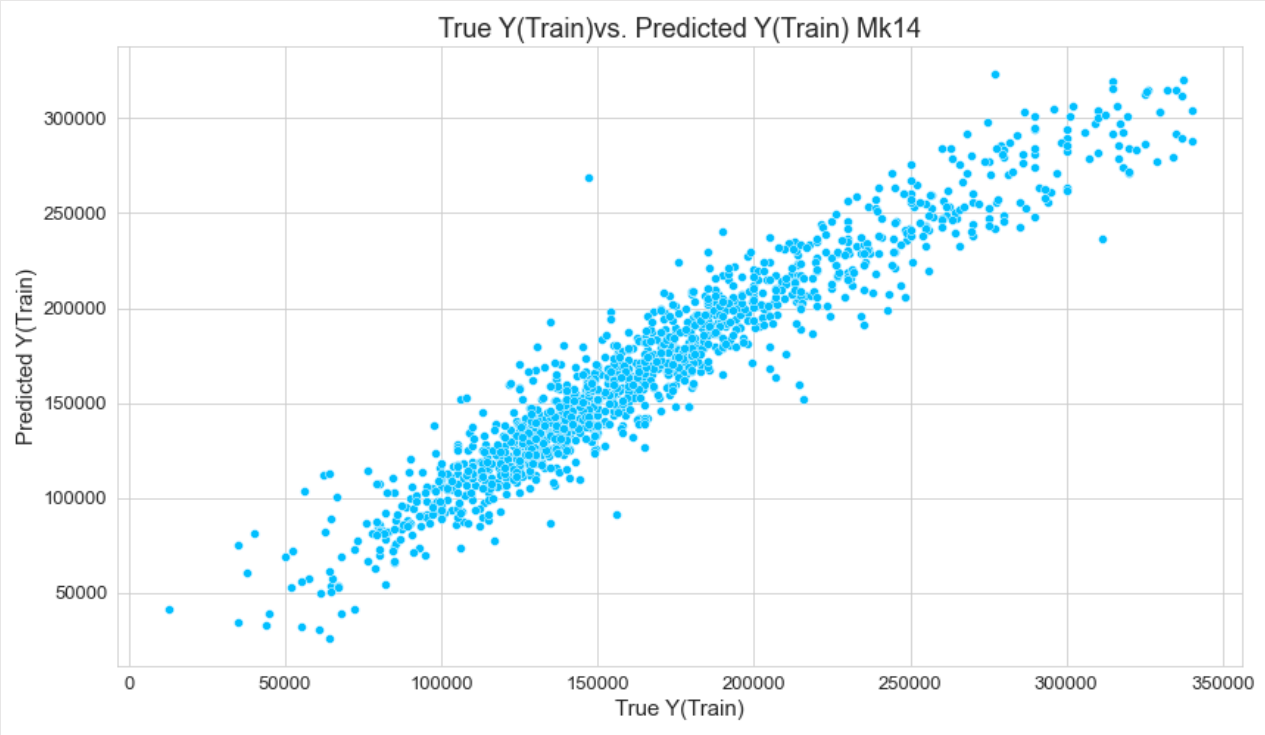


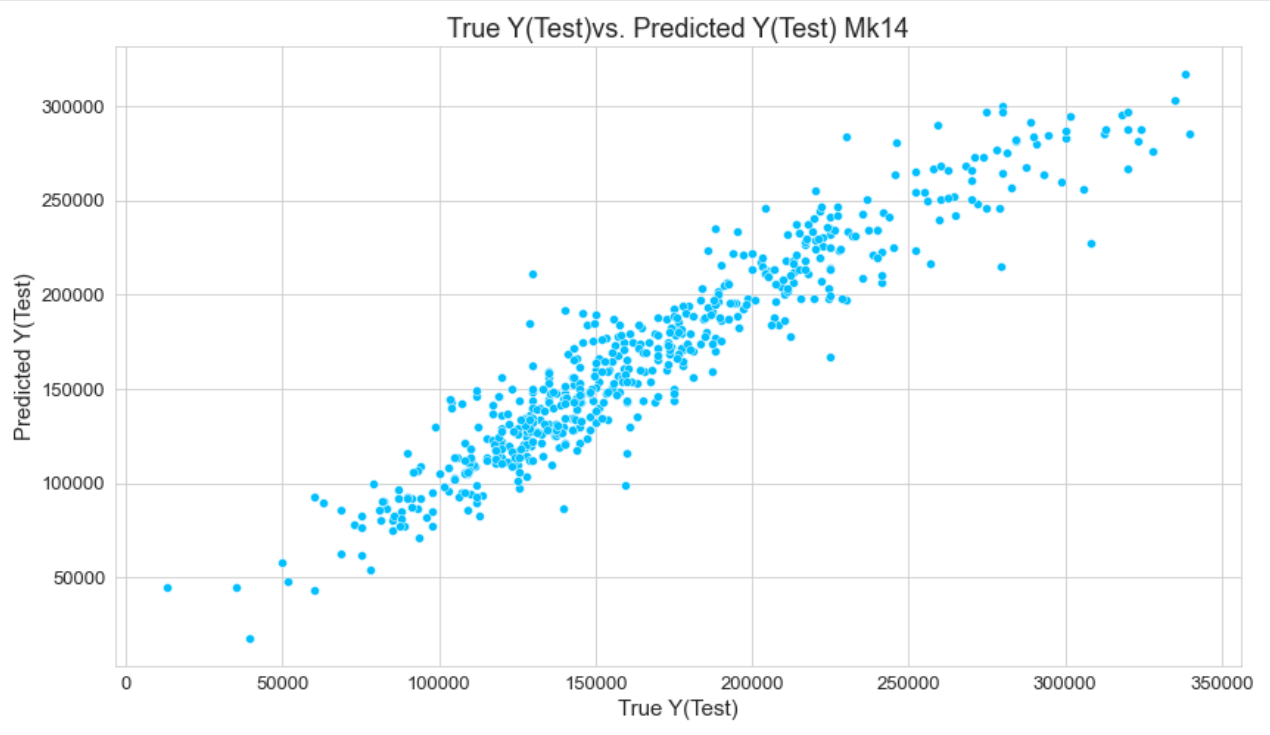
Model 11 Correlated Features > .15





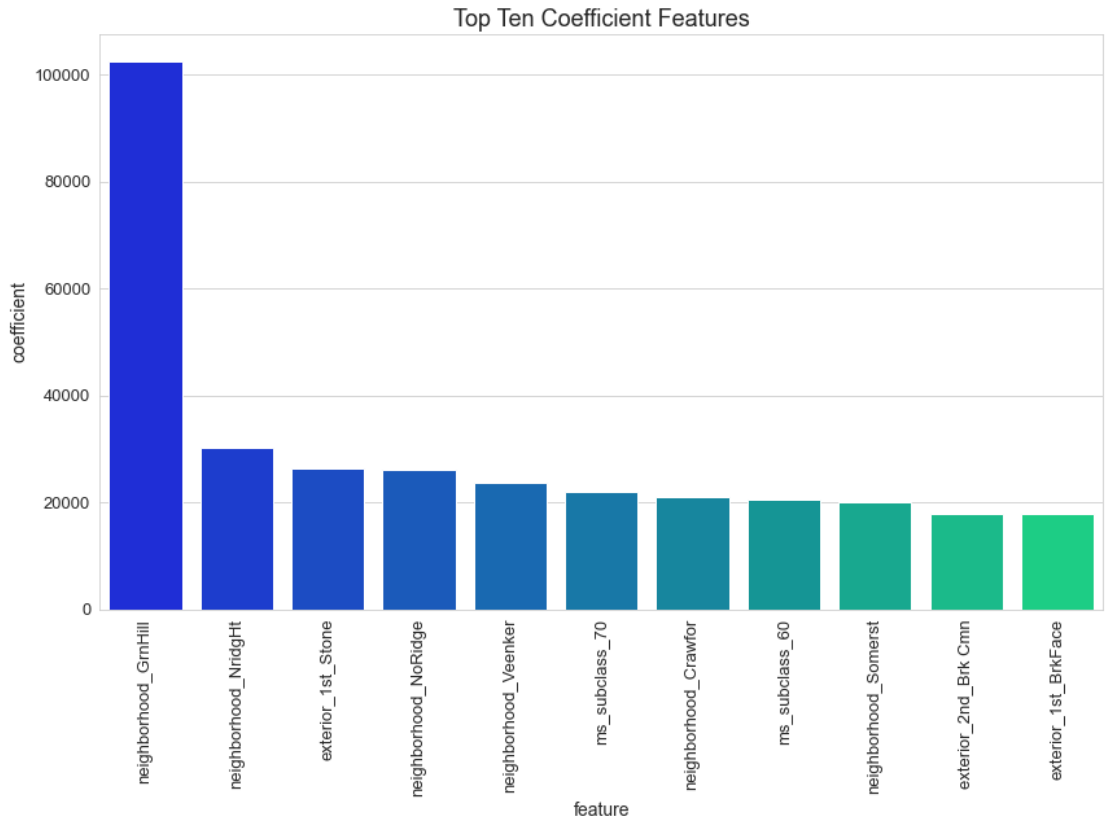
Model 14 Correlated features > -.10

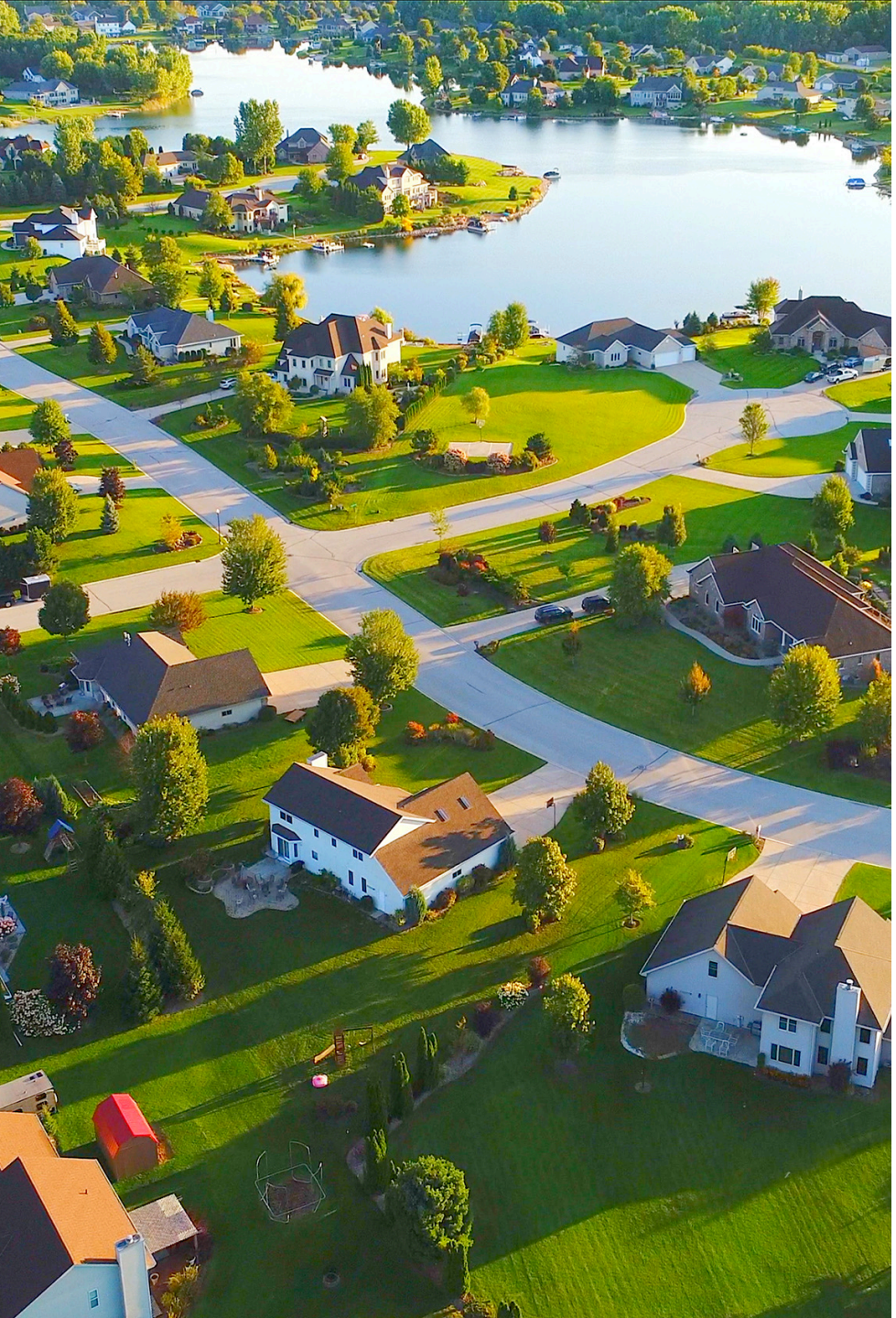




Top Coefficients Derived from Linear Regression Model 14

feature	coefficient
neighborhood_GrnHill	102503.320378
neighborhood_NridgHt	30243.253877
exterior_1st_Stone	26330.171572
neighborhood_NoRidge	26121.814429
neighborhood_Veenker	23702.573793
ms_subclass_70	21992.115718
neighborhood_Crawfor	21173.306295
ms_subclass_60	20511.615767
neighborhood_Somerst	20116.074366
exterior_2nd_Brk Cmn	18002.224645
exterior_1st_BrkFace	17896.376162





Initial Phase

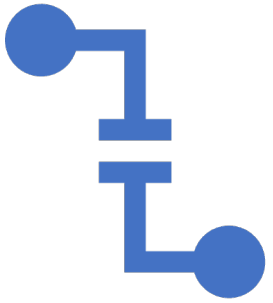
- The linear regression model number 7 outperformed the baseline model by whittling down any features that had a poor correlation with sales price. Used techniques such as polynomial features and standard scaler.
- R2 score improved from baseline of 65.59% to 83.41%(17.82-point improvement),in other words 84.44% of the variability in our sales price could be explained by the features in our model.

Conclusion

- Revisiting this project meant re-tackling dealing with outliers, dealing with Nulls, processing the data , feature selection and having a DRY approach.
- Some key take aways were an improvement to 87% for the testing R2 Score right out of the gate with model 8. Also, being able to halve my testing RMSE from model 1(43,739.88 to 20,642.51).
- An additional metric I wanted to evaluate with was the adjusted R2, this is because the fact that the Adjust R2 value was higher than my testing R2 indicated that my models would work well by adding additional features.
- This dictated the rest of my approach which peaks with model 14 which not only achieves the highest testing R2 score of .9 but also has the lowest RMSE of 17,984.23.

Next Steps

After improving the model by 25 points to a .90 R2 Score and minimizing the RMSE 17984.23, revisiting the project would entail applying different types of models



For instance, boosting models could lead to even better performance, will allow for grid-searching of hyper-parameters but may not be as useful for inference.

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/).