TO WORK ON ISSUE OF UNDER-PREDICTING HOUSE PRICES VIA LINEAR REGRESSION ABOVE $320,000

Mohammad Sufyan

PROJECT OVERVIEW

Exploratory Data Analysis

Identify key features , pairwise relationships

Linear regression and Feature engineering

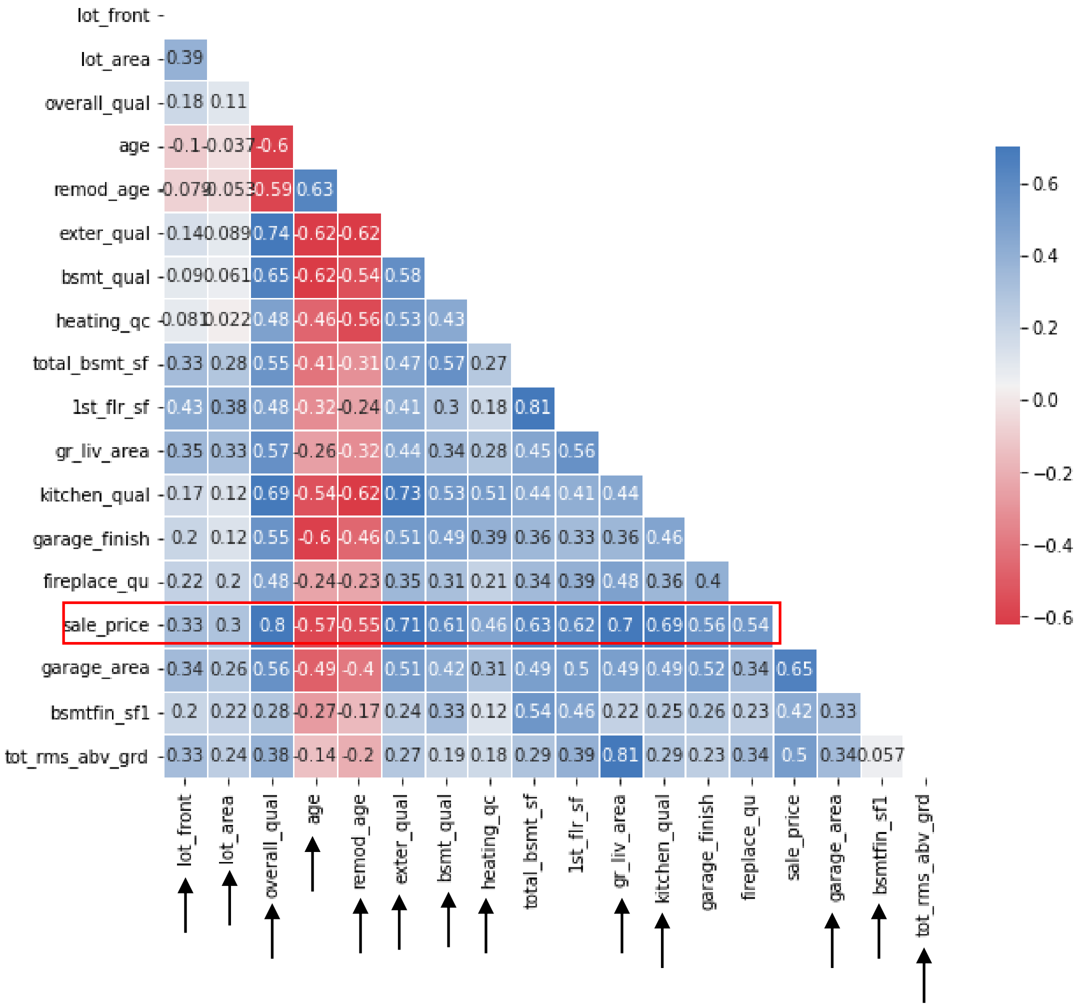
Interaction terms, Optimise ordinal features, Minimize Outliers

Lasso Regression

Optimise coefficients

Key Takeaways and Future work ahead

HEATMAP ON CORRELATION TO SALE PRICE



FEATURES IDENTIFIED

1) Generally the bigger the area the higher the price
2) Ordinal categories with strong correlation with Sale Price

   (Quality and conditions)
3) Notably strong correlations
   - Overall quality +ve correlation (+0.8)
   - Age of buildings –ve correlation generally with all features

     (-0.55 with sale price)

PAIRWISE RELATIONSHIPS

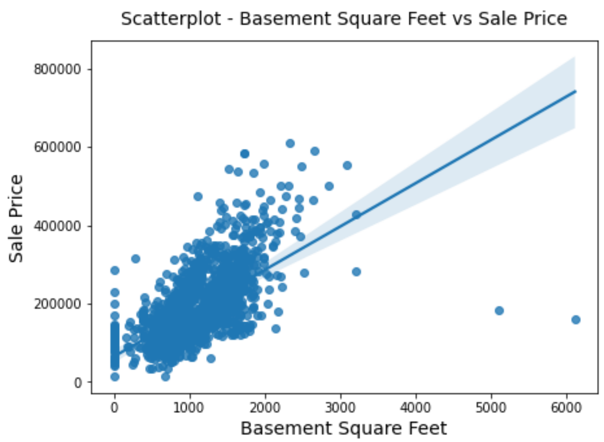
Feature 1	Feature 2	Pair Correlation	Feature 1 & Sales Correlation	Feature 2 & Sales Correlation
garag_cars	garage_area	0.889558	0.648128	0.650246
gr_liv_area	tot_rms_abv_grd	0.808174	0.697038	0.504014
total_bsmt_sf	1st_flr_sf	0.798801	0.629303	0.618486
bedroom_abv_gr	tot_rms_abv_grd	0.673442	0.137067	0.504014
2nd_flr_sf	gr_liv_area	0.654530	0.248452	0.697038
bsmtfin_sf1	bsmt_full_bath	0.640606	0.423856	0.283332
gr_liv_area	full_bath	0.629736	0.697038	0.537969
2nd_flr_sf	half_bath	0.611432	0.248452	0.283001
overall_qual	garag_cars	0.598912	0.800207	0.648128
2nd_flr_sf	tot_rms_abv_grd	0.584059	0.248452	0.504014

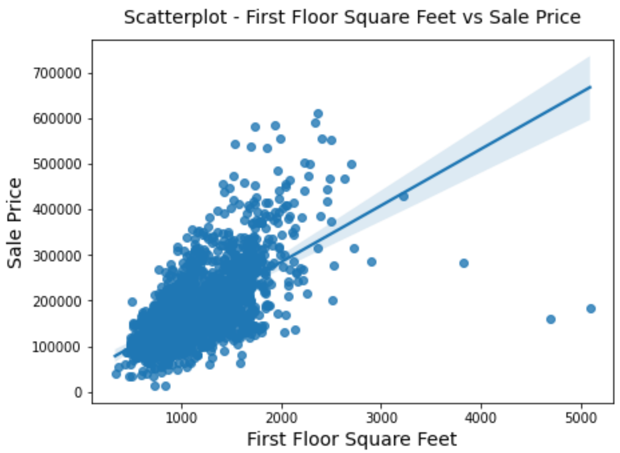
High correlation to each other

PAIRWISE RELATIONSHIPS

Feature 1	Feature 2	Pair Correlation	Feature 1 & Sales Correlation	Feature 2 & Sales Correlation
garag_cars	garage_area	0.889558	0.648128	0.650246
gr_liv_area	tot_rms_abv_grd	0.808174	0.697038	0.504014
total_bsmt_sf	1st_flr_sf	0.798801	0.629303	0.618486
bedroom_abv_gr	tot_rms_abv_grd	0.673442	0.137067	0.504014
2nd_flr_sf	gr_liv_area	0.654530	0.248452	0.697038
bsmtfin_sf1	bsmt_full_bath	0.640606	0.423856	0.283332
gr_liv_area	full_bath	0.629736	0.697038	0.537969
2nd_flr_sf	half_bath	0.611432	0.248452	0.283001
overall_qual	garag_cars	0.598912	0.800207	0.648128
2nd_flr_sf	tot_rms_abv_grd	0.584059	0.248452	0.504014

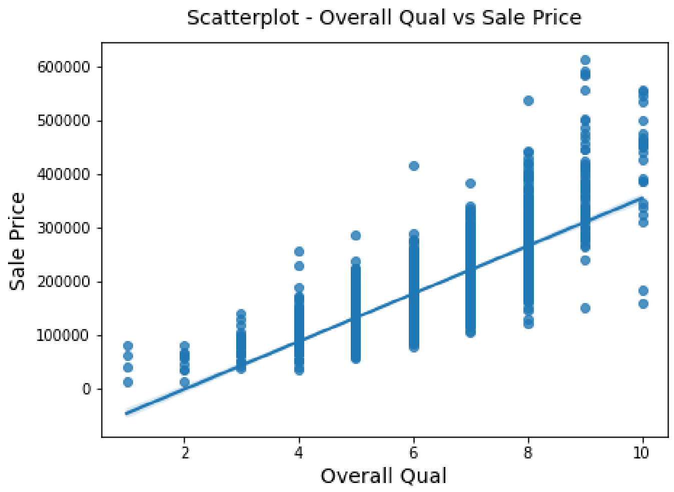
High correlation to each other



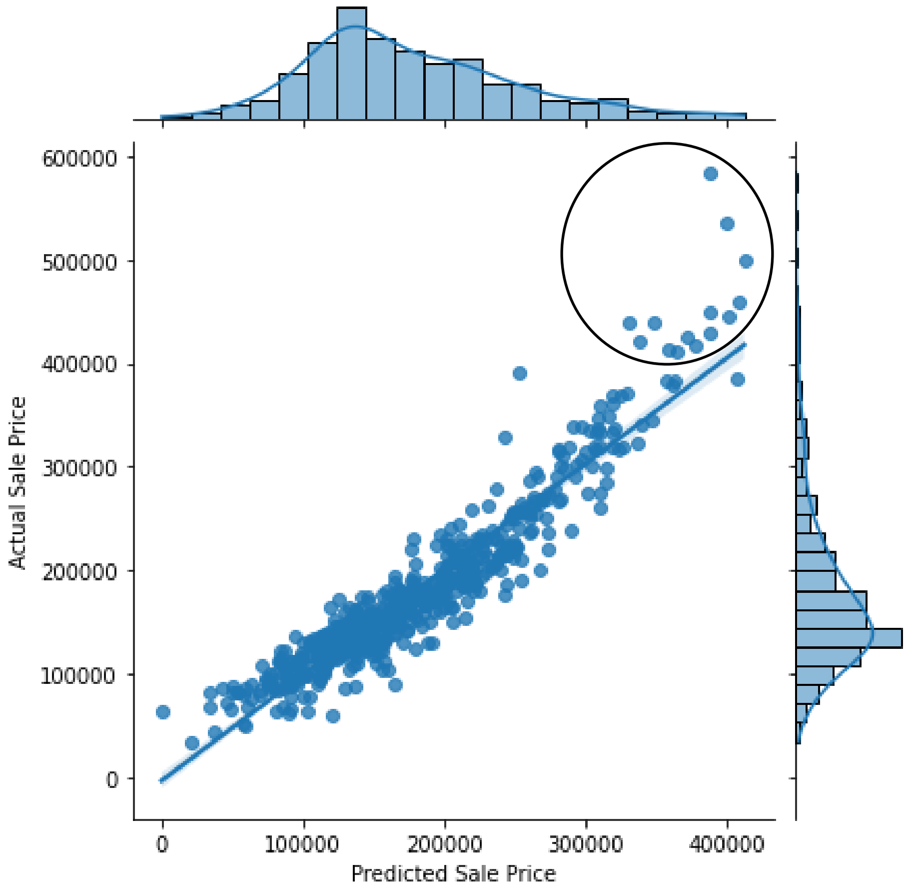


PAIRWISE RELATIONSHIPS

Feature 1	Feature 2	Pair Correlation	Feature 1 & Sales Correlation	Feature 2 & Sales Correlation
garag_cars	garage_area	0.889558	0.648128	0.650246
gr_liv_area	tot_rms_abv_grd	0.808174	0.697038	0.504014
total_bsmt_sf	1st_flr_sf	0.798801	0.629303	0.618486
bedroom_abv_gr	tot_rms_abv_grd	0.673442	0.137067	0.504014
2nd_flr_sf	gr_liv_area	0.654530	0.248452	0.697038
bsmtfin_sf1	bsmt_full_bath	0.640606	0.423856	0.283332
gr_liv_area	full_bath	0.629736	0.697038	0.537969
2nd_flr_sf	half_bath	0.611432	0.248452	0.283001
overall_qual	garag_cars	0.598912	0.800207	0.648128
2nd_flr_sf	tot_rms_abv_grd	0.584059	0.248452	0.504014

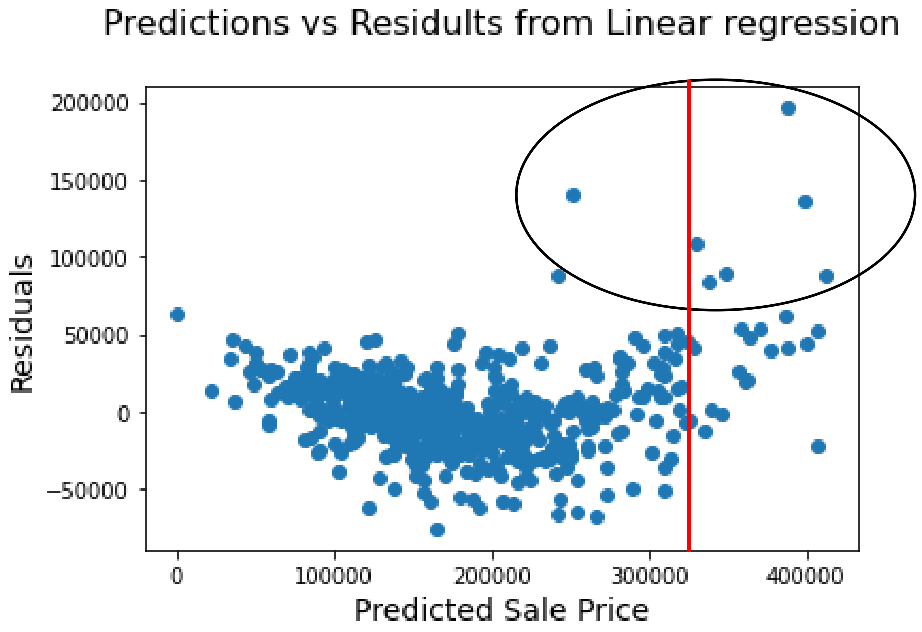


LINEAR REGRESSION MODEL

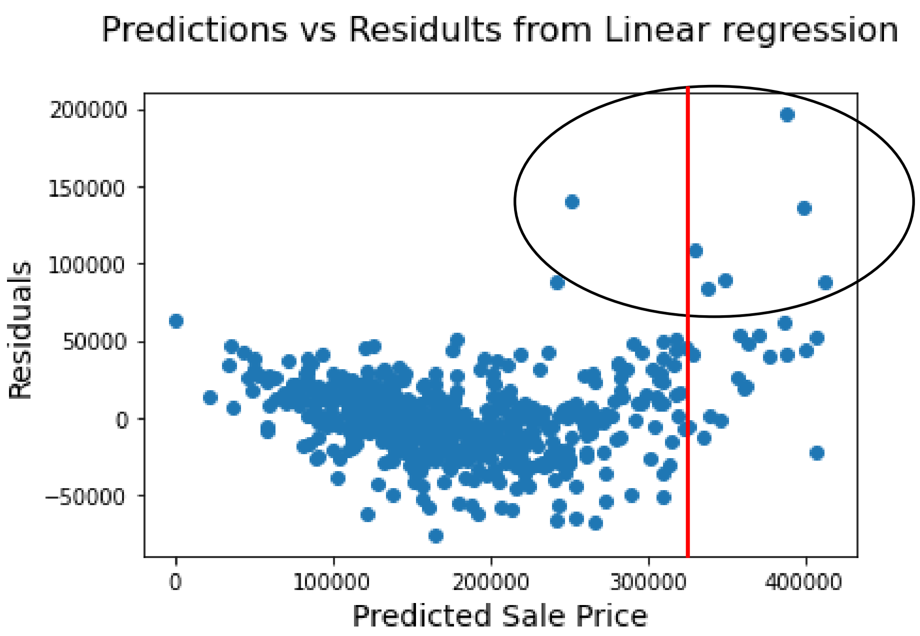


<u>Current RMSE</u>: 26149

- ✓ Underfitted, with higher variance at the top above
- ✓ Noted after $350,000, higher residuals observed



LINEAR REGRESSION MODEL



- ✓ Observation:
  - ✓ This indicates under-prediction of housing prices (Residual of $200,000!)
- ✓ Hypothesis:
  - ✓ We are missing the X-factor that helps bring up these assets to its true potential

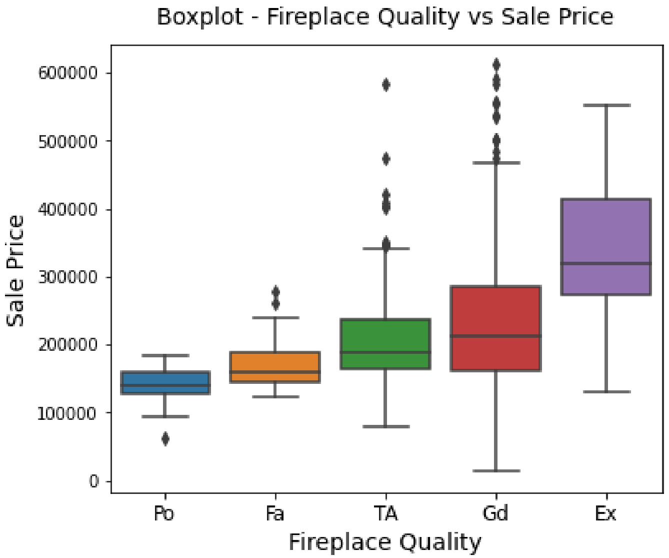
IMAGINE VIEWING...

Fireplace

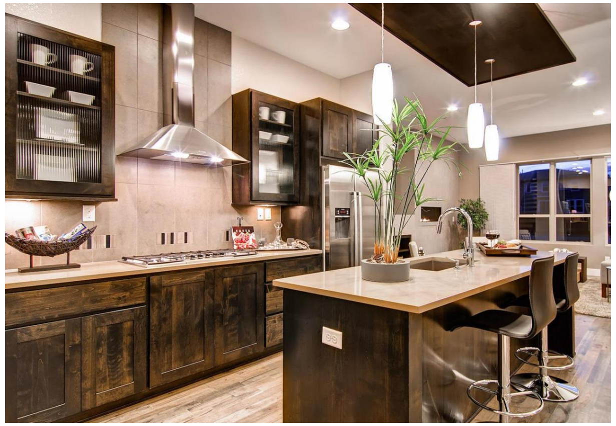
Kitchen Quality

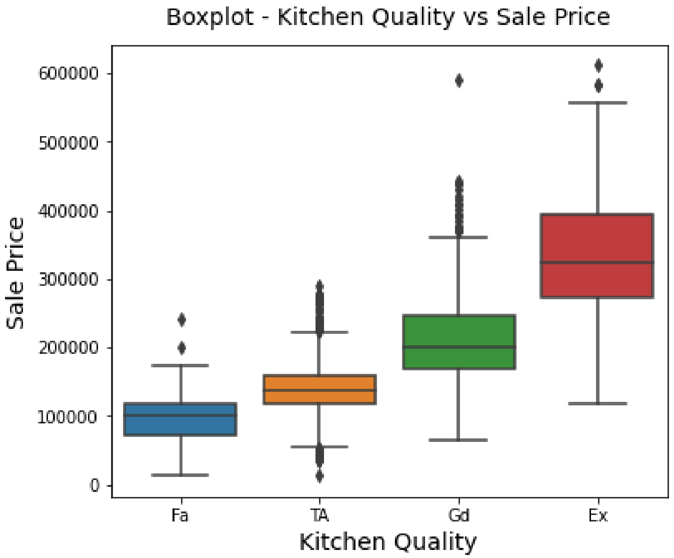
LOOKING AT THEIR CORRELATION INDEPENDENTLY

Fireplace



Kitchen Quality

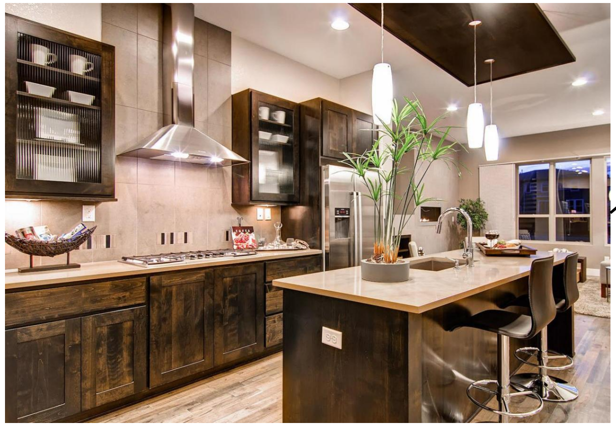




IN THE REAL WORLD, WE SEE IT AS A WHOLE

Fireplace

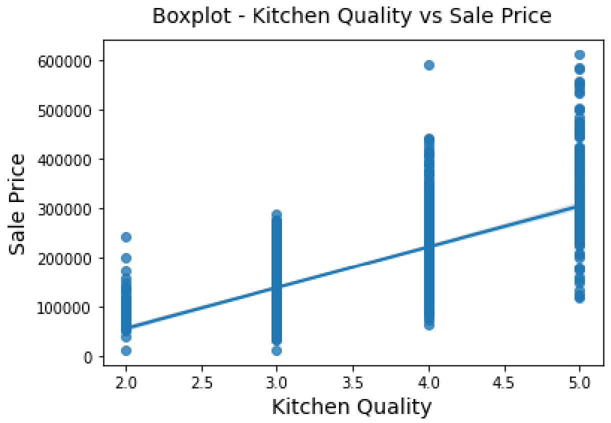
Kitchen Quality

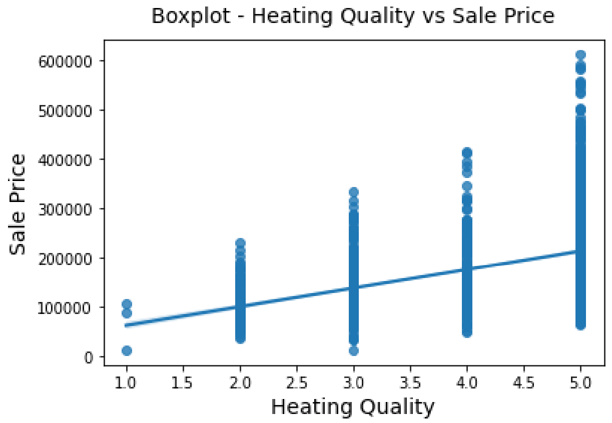


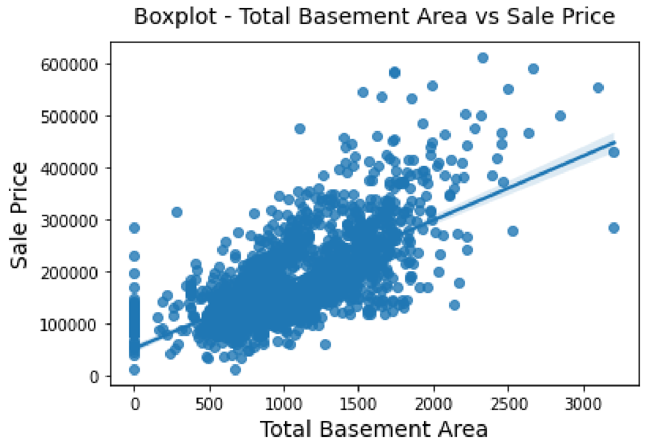
Interacting with one another

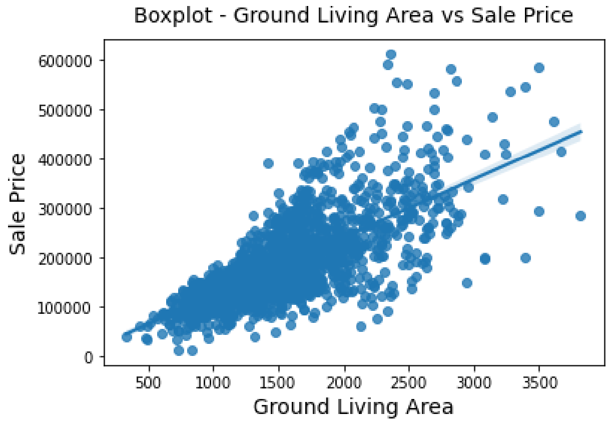
- ✓ How big the living space is?
- ✓ How grand it looks overall
- ✓ Hence, we multiply these together to have an amplified effect of these features together

FEATURING ENGINEERING – INTERACTION TERMS





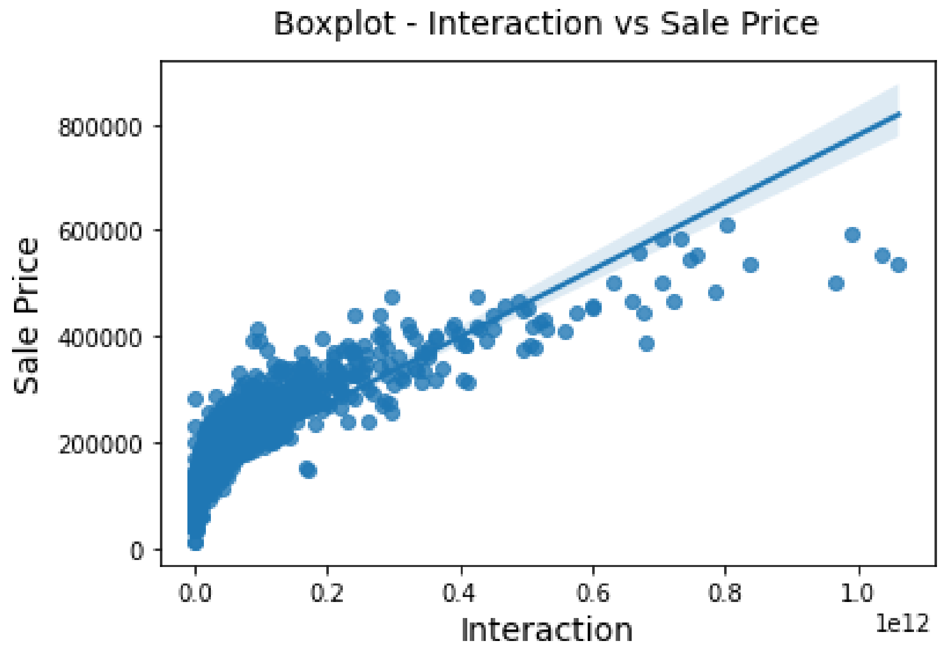


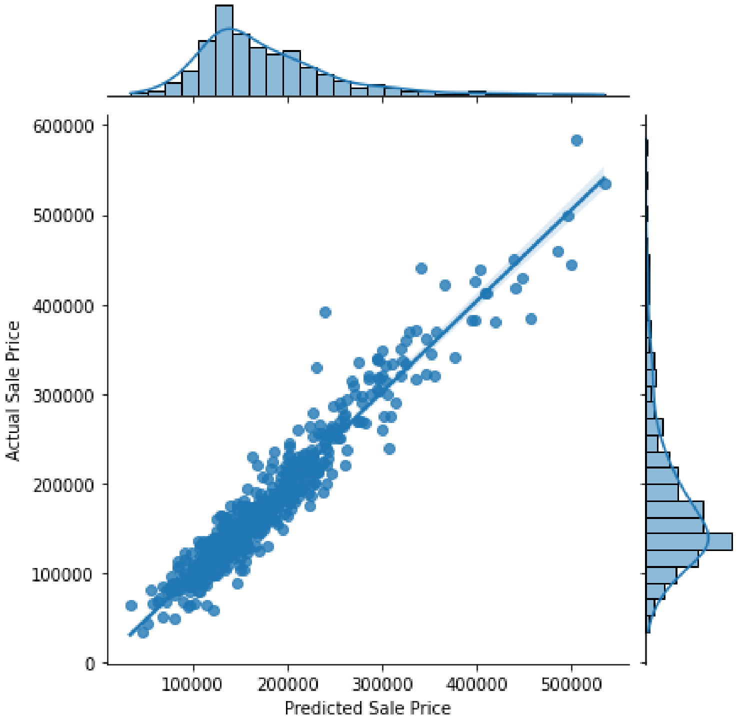


Model takes in interaction terms

- ✓ Quality
- ✓ Square Feet Area
- ✓ Condition

INTERACTION TERMS





Current RMSE: 21376, -18.76%

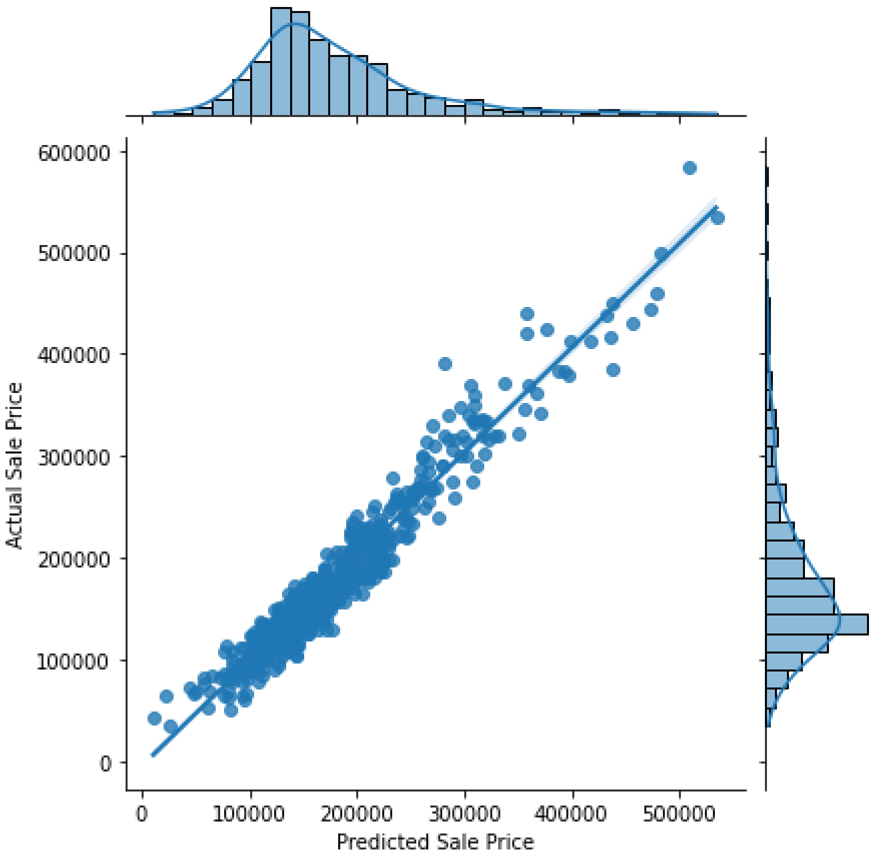
- ✓ Highly correlated to price
- ✓ High in magnitude
- ✓ Improved RMSE by 18.7%

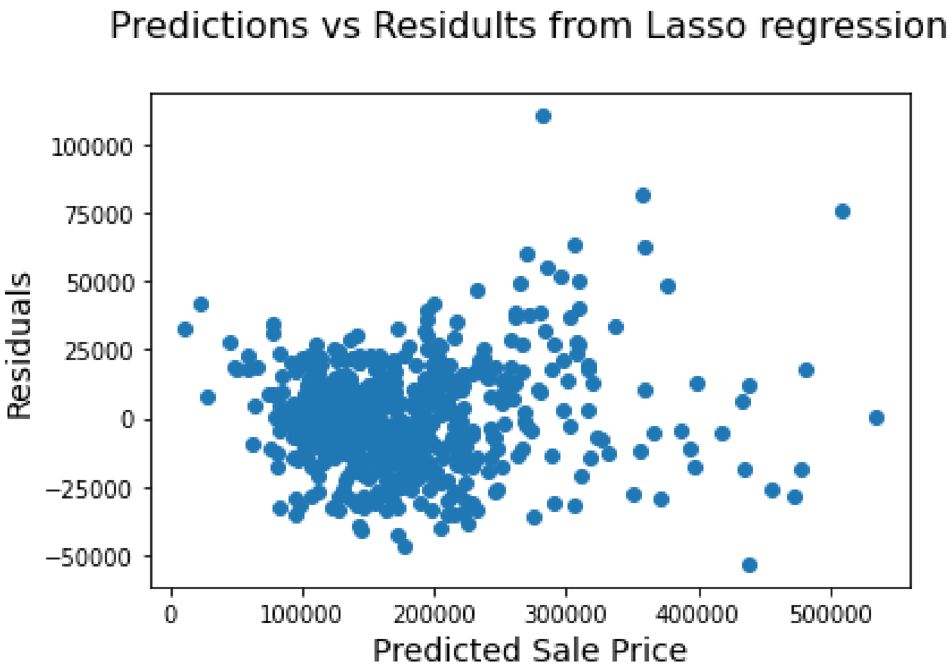
- ✓ No more skewing upwards from higher sale price
- ✓ Noted still high bias
- ✓ Need more variables

LASSO REGULARIZATION

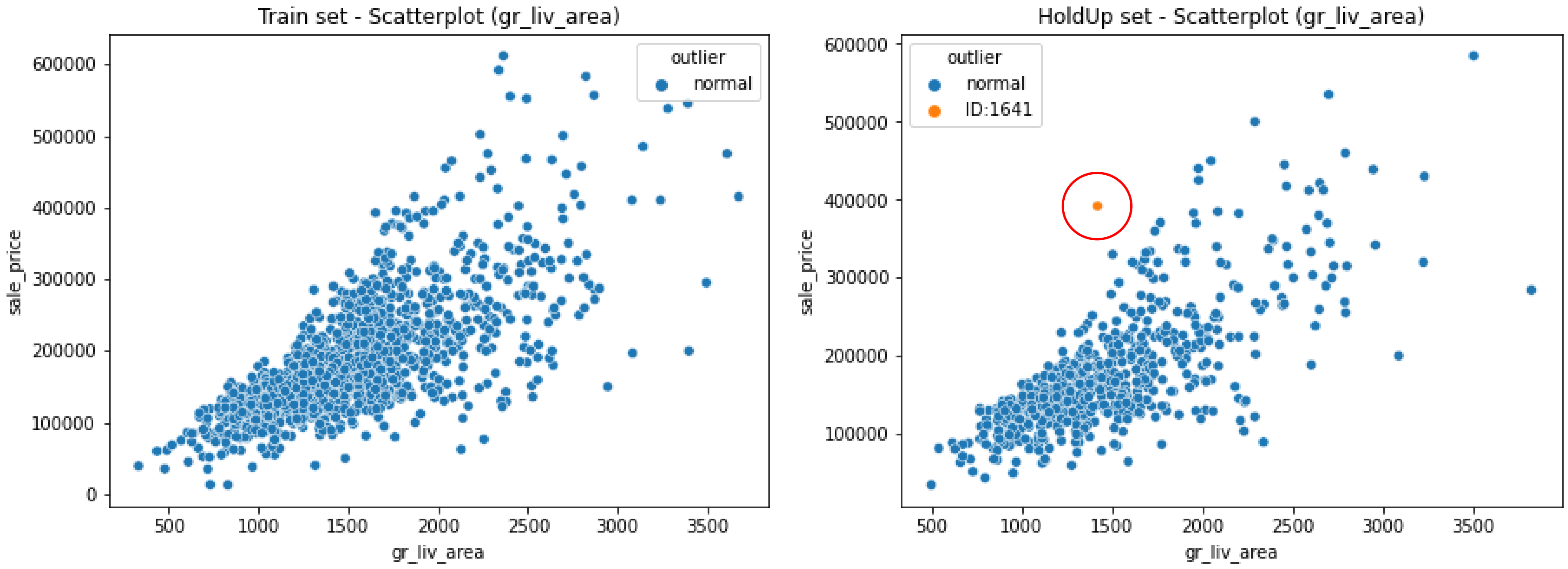
Model	Changes done	Comments	RMSE	RMSE % Change
Linear regression	Added more features, including ✓ All Ordinal ranked sequentially ✓ Nominal features	Variance and bias too large for large coefficients and number of variables Need loss function to penalize each variable	272751033127647.7	NA
Lasso Regression	Alpha used: 577	Most coefficients are zeroed out	19517	-7.87%
Lasso Regression	Remove zeroed features: 183 → 79 Alpha used: 171	Most coefficients are zeroed out	19497	-0.1%
Lasso Regression	Optimized ordinal features As not all ordinal features are strongly related and should not be ranked. Changed to nominal instead	Outliers removed More linear with lesser variance	18519	-5%

FINAL - LASSO REGRESSION





OUTLIERS REMOVED



✓ Outlier found in holdup set is not found in Train set hence removed

THANK YOU

Key takeaways

- ✓ Not all ordinal categories should be given sequential scoring, else unnecessary amplication can ruin predictions
- ✓ Understanding interaction terms in consumer's POV and putting into Data science concepts

Future work ahead

- ✓ Explore deeper EDA on each ordinal categories and considering doing interaction or putting a score to it instead of nominal categories that kills it with 1s and 0s
- ✓ Can explore rare variables and filter it off from model