Ames, Iowa Zillow Estimated Housing Prices

By Chris Shaw

<u>EDA</u>

- 80 features
 - 20 continuous
 - 14 discrete
 - 46 categorical
- Significant null values in alley, pool quality, fence and misc feature
- Strong correlations with overall quality, ground living area, and size and capacity of the garage

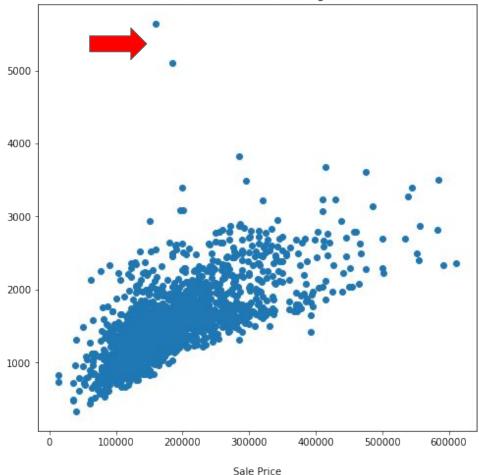
pid -	-0.26
enclosed_porch -	-0.14
kitchen_abvgr -	-0.13
overall_cond -	-0.097
ms_subclass ⁻	-0.087
id -	-0.051
bsmt_half_bath -	-0.045
low_qual_fin_sf -	-0.042
yr_sold -	-0.015
misc_val -	-0.0074
bsmtfin_sf_2 -	0.016
pool_area -	0.023
mo_sold -	0.033
3ssn_porch -	0.049
screen_porch -	0.13
bedroom_abvgr	0.14
bsmt_unf_sf	0.19
2nd_flr_sf -	0.25
half_bath -	0.28
bsmt_full_bath -	0.28
lot_area -	0.3
wood_deck_sf -	0.33
open_porch_sf -	0.33
lot_frontage -	0.34
bsmtfin_sf_1 -	0.42
fireplaces -	0.47
totrms_abvgrd -	0.5
mas_vnr_area -	0.51
garage_yr_blt -	0.53
full_bath -	0.54
year_remod/add -	0.55
year_built -	0.57
1st_flr_sf	0.62
total_bsmt_sf -	0.63
garage_cars -	0.65
garage_area -	0.65
gr_liv_area -	0.7
overall_qual =	0.8
saleprice -	1

saleprice

<u>EDA</u>

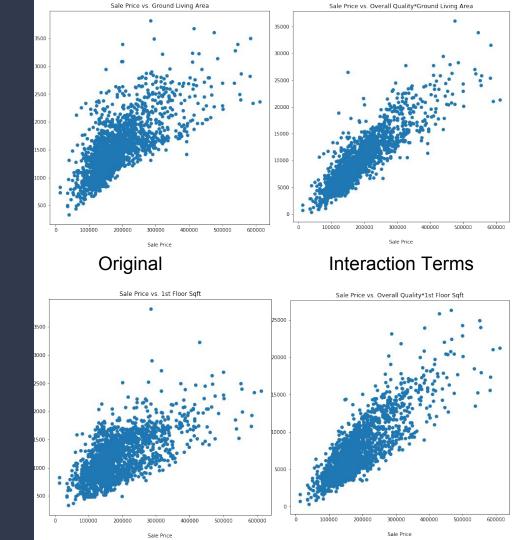
 Two outliers of large area houses that sold for low amounts





Feature Selection & Engineering

- Dummied out categorical features, then dropped the least correlated
- Doubled highly correlated features to increase signal
- Created polynomial features of the most highly correlated features



Modeling & Evaluation

- Used Linear, Ridge and Lasso Regression
- Scaled using Standardscaler
- Linear Regression performed poorly
- Ridge and Lasso regressions did much better
- Slightly overfit but not a lot

	Linear Regression	Ridge Regression	Lasso Regression	
X_train score	.937	.938	.934	
X_test score	-6.77e+20	.925	.924	
Cross_va Iscore	-7.42e+19	.915	.916	

<u>Recommendations</u>

- Use interaction terms to boost the significant features
- Binarize the categorical variables to look for other useful features
- Eliminate features that have close to no correlation
- Use Lasso or Ridge regression