# Ames, Iowa
# Zillow Estimated Housing Prices

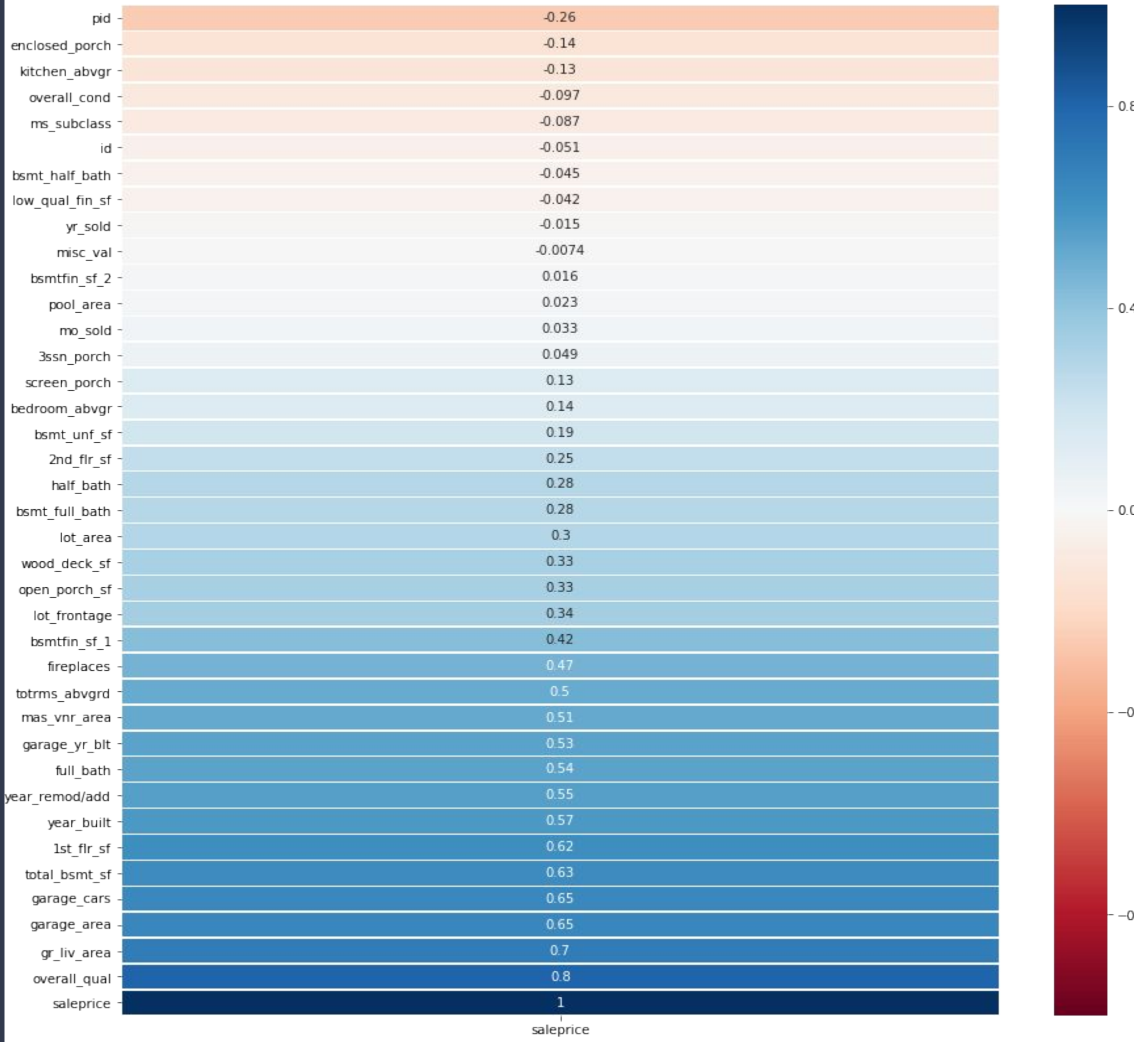
By Chris Shaw

# EDA

- 80 features
  - 20 continuous
  - 14 discrete
  - 46 categorical

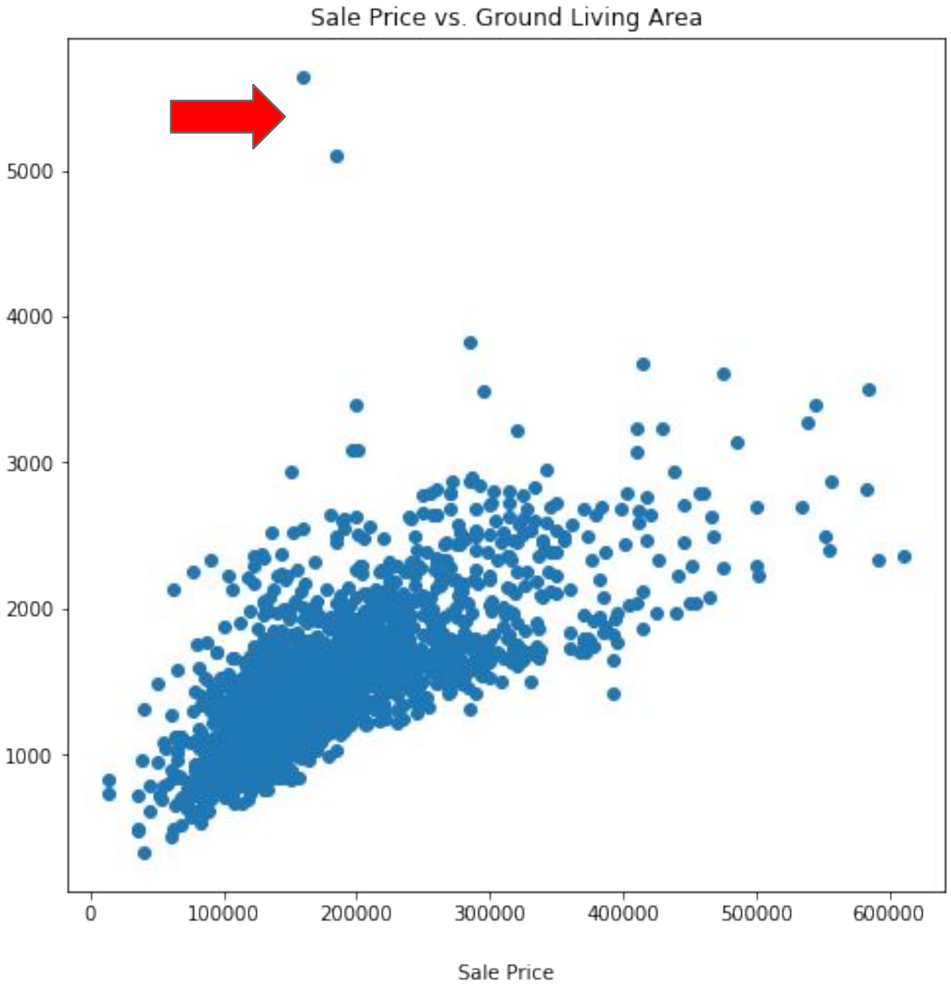
- Significant null values in alley, pool quality, fence and misc feature

- Strong correlations with overall quality, ground living area, and size and capacity of the garage



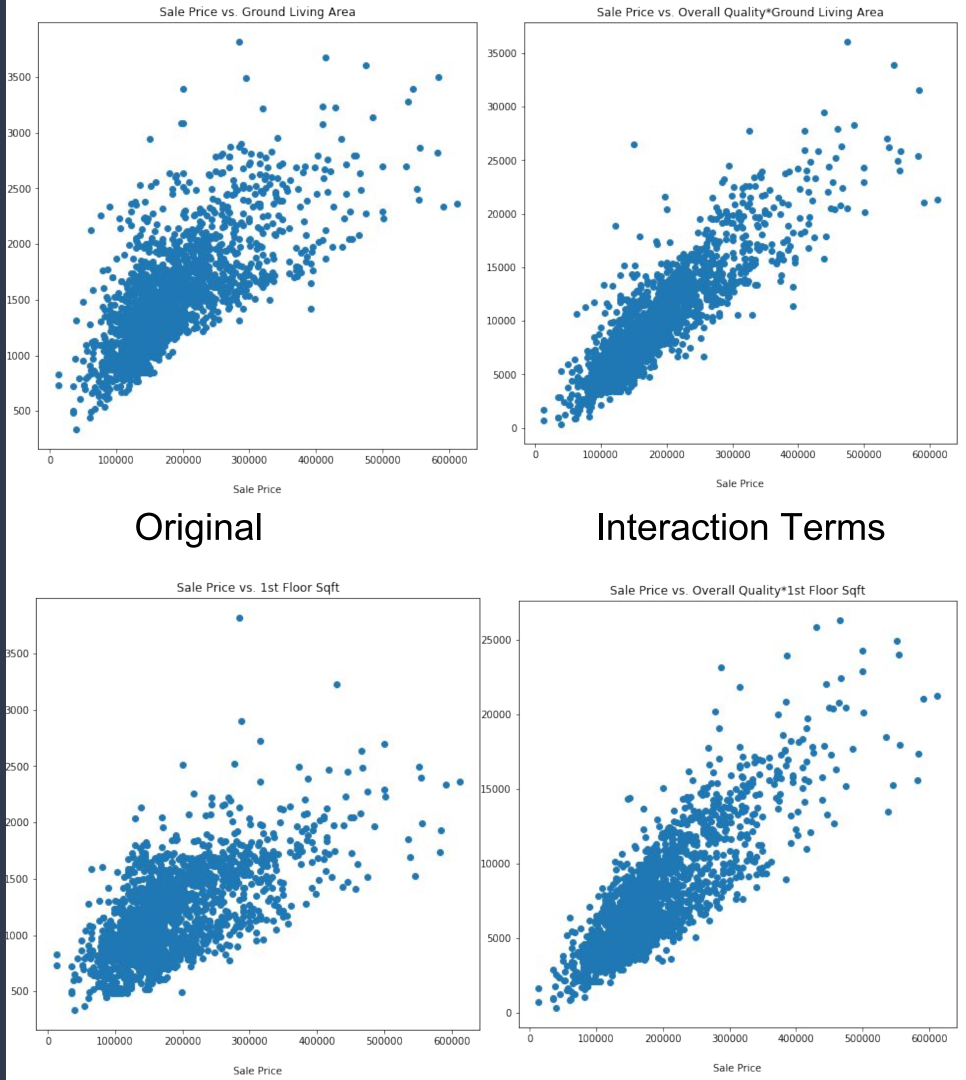
# EDA

- 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



Original

Interaction Terms

# 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_val score | -7.42e+19 | .915 | .916 |

# Recommendations

- 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