Ames Housing Modeling and Predictions

Alex Harmon

The Process

1. Defining the problem/goal
2. Importing, cleaning, and organizing the data
3. Making inferences
4. Creating and testing the model
5. Generating predictions
6. The next step

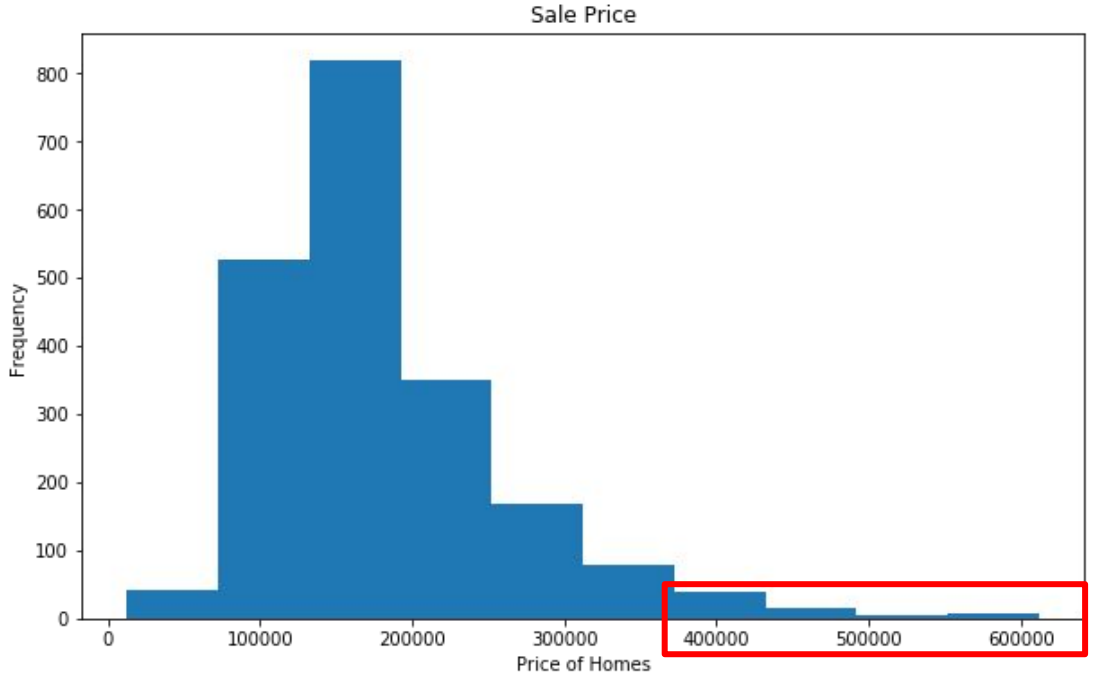
The Data Science Problem

- Using data gathered from housing prices in Ames, Iowa, can we predict home values?

- What effect do the variables have on the pricing of homes in Ames, Iowa?

Exploratory Data Analysis
&
Feature Engineering

Distribution of Sale Prices



Feature Engineering

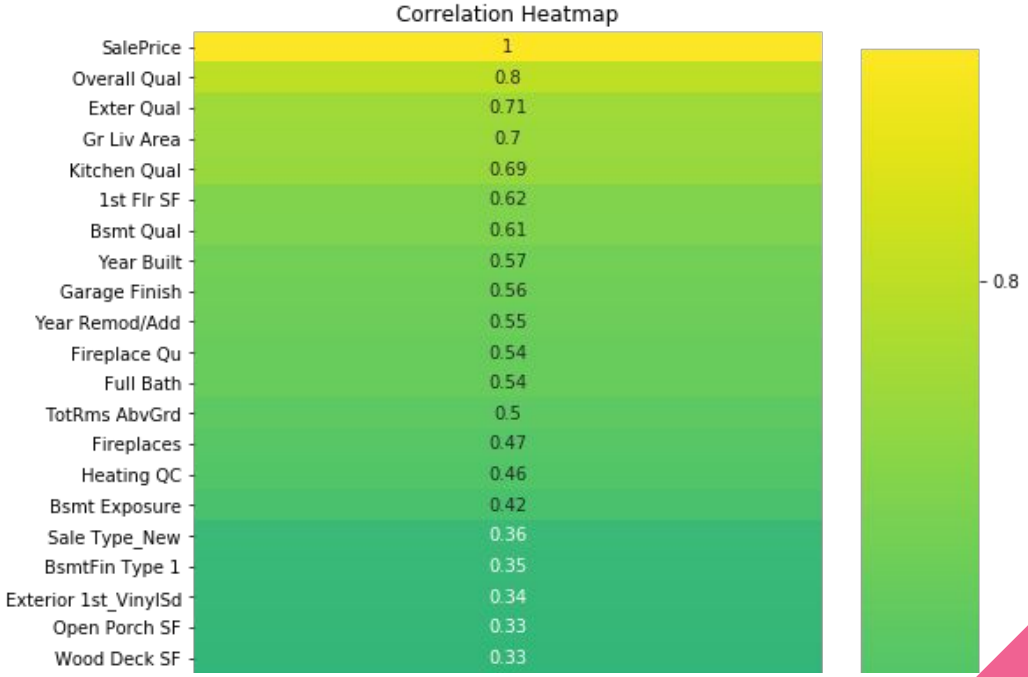
Ordinal Variables

- Lot Shape
- Land Slope
- Basement Exposure
- Basement Quality
- Basement Condition
- Bsmt Finish 1
- Bsmt Finish 2
- Heating Quality
- Utilities
- Exterior Quality
- Exterior Condition
- Kitchen Quality
- Electrical
- Functional
- Pool QC
- Paved Driveway

Dummy Variables

- Street
- Sale Type
- Building Type
- House Style
- Exterior 1st
- Central Air

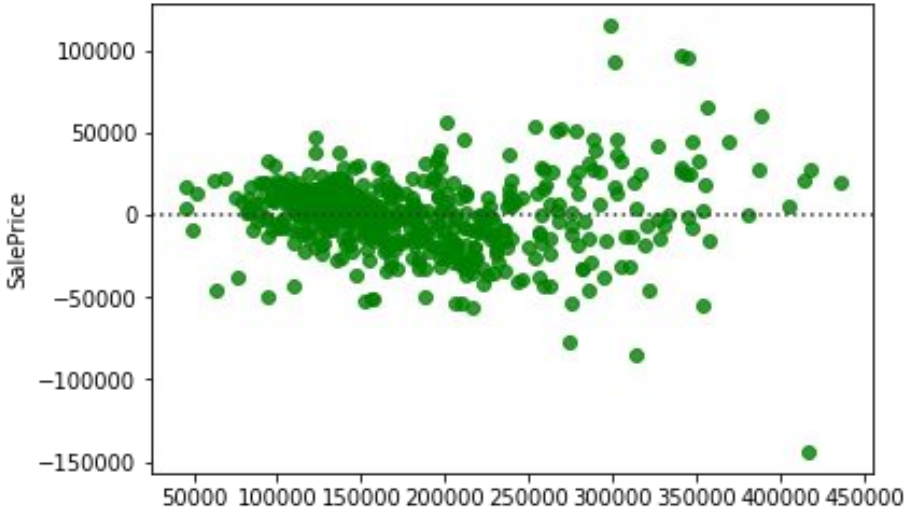
Variable Correlations to Sale Price



Modeling
&
Results

Modeling Choices

- Polynomial Features
- Standard Scaling
- LassoCV
- Pipeline
- Gridsearch



Regression Model

Most Powerful Features

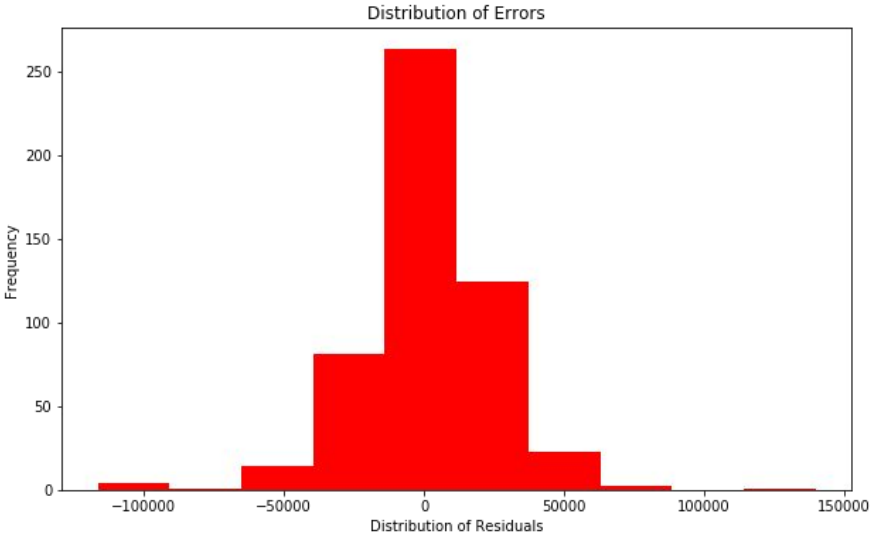
	coefs	vals
463	Overall Qual Gr Liv Area	19313.117096
2347	Pool Area Sale Type_New	-14707.160753
2382	Misc Val Sale Type_New	-13646.172885
468	Overall Qual Kitchen Qual	9609.956578
734	Exter Qual 1st Flr SF	5846.950939
1178	Heating QC Gr Liv Area	5317.651143
861	Bsmt Qual BsmtFin Type 1	4624.698129
460	Overall Qual 1st Flr SF	4458.809914
865	Bsmt Qual 1st Flr SF	3755.505073
1468	Gr Liv Area Kitchen Qual	3479.362137
1004	Bsmt Exposure Fireplace Qu	3173.379277
1054	BsmtFin Type 1 1st Flr SF	3143.738184

- Polynomial features created several powerful interaction features

- “Overall Quality * Gr Living Area”
  - For a one unit increase in this feature, we expect an increase in housing price of approximately $19,313, given all other features held constant
- “Pool Area * Sale Type_New”
  - Single result in this category

Results

- R2 score = 0.91
  - This means approximately 91% of our total variance can be explained by the model in comparison to the null



- Root Mean Squared Error = 23,774.03
  - This means that on average our predictions varied by 23,774 from the true values (punished for outliers)

What’s Next

- Continue collecting data to see how the model performs against future/unseen data
- Collect additional data to iterate and improve prediction and generalization power of the model
- Attempt non-linear regression models