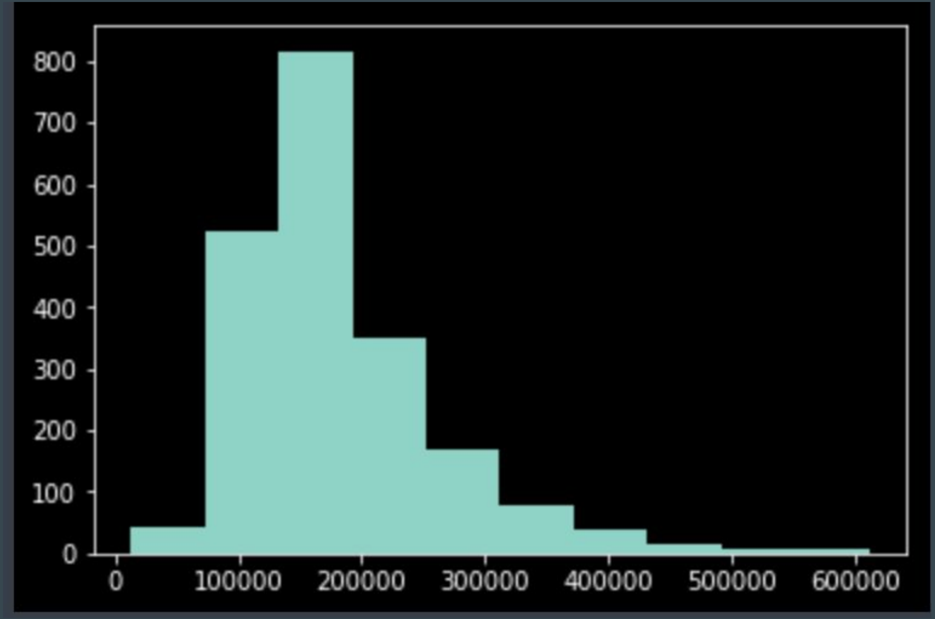
Ames Housing Market Analysis: 2006-2010

Kate Dowdy

Housing Prices



Sale Prices

From 2006-2010 (training data):

- **Mean price:** $181, 470
- **Median price:** $162,500

EDA & Feature Engineering Process

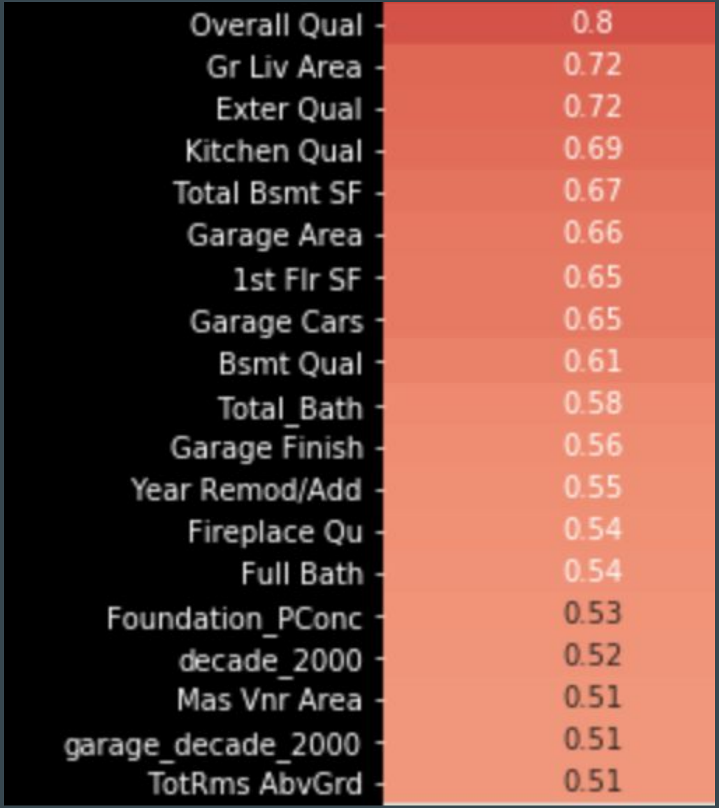
Munging/EDA:

- **Filling nulls:** 'NA' not missing data, but reading as nulls (changed to '0')
- **Object columns to numeric:** changed those with ordinal ratings
- **Numeric columns to categorical:** MS Subclass, year built (by decade), garage year built (by decade)

Feature Engineering:

- **Created new features:** mansion, all bathrooms (float), asbestos, sell time (yr + mo)
- **Dummied** categorical variables
- **Dropped features:** ID, PID, Year Built, Garage Year Built
- **Polynomial features** for all features
- **Scaled** all features

Initial Correlations

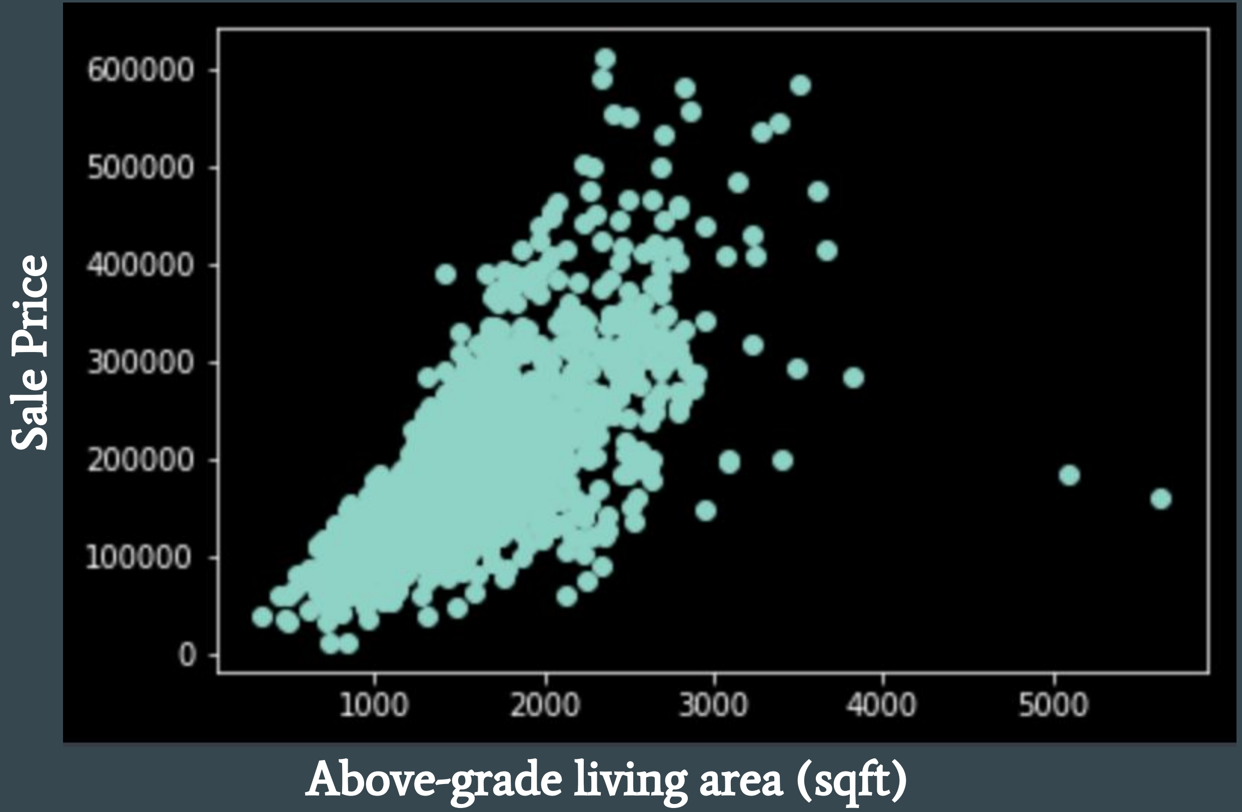


Overall Qual	0.8
Gr Liv Area	0.72
Exter Qual	0.72
Kitchen Qual	0.69
Total Bsmt SF	0.67
Garage Area	0.66
1st Flr SF	0.65
Garage Cars	0.65
Bsmt Qual	0.61
Total_Bath	0.58
Garage Finish	0.56
Year Remod/Add	0.55
Fireplace Qu	0.54
Full Bath	0.54
Foundation_PConc	0.53
decade_2000	0.52
Mas Vnr Area	0.51
garage_decade_2000	0.51
TotRms AbvGrd	0.51

Highest correlations with sale price (before polynomial features):

- **Quality ordinal ratings** (overall, kitchen, exterior, basement, fireplace)
- **Square footage** (Gr liv area, garage area, basement sqft, 1st floor sqft)
- **Bathrooms**
- **Year built/renovated**, year garage built
- **Building materials/finish**

A Note on the Outliers

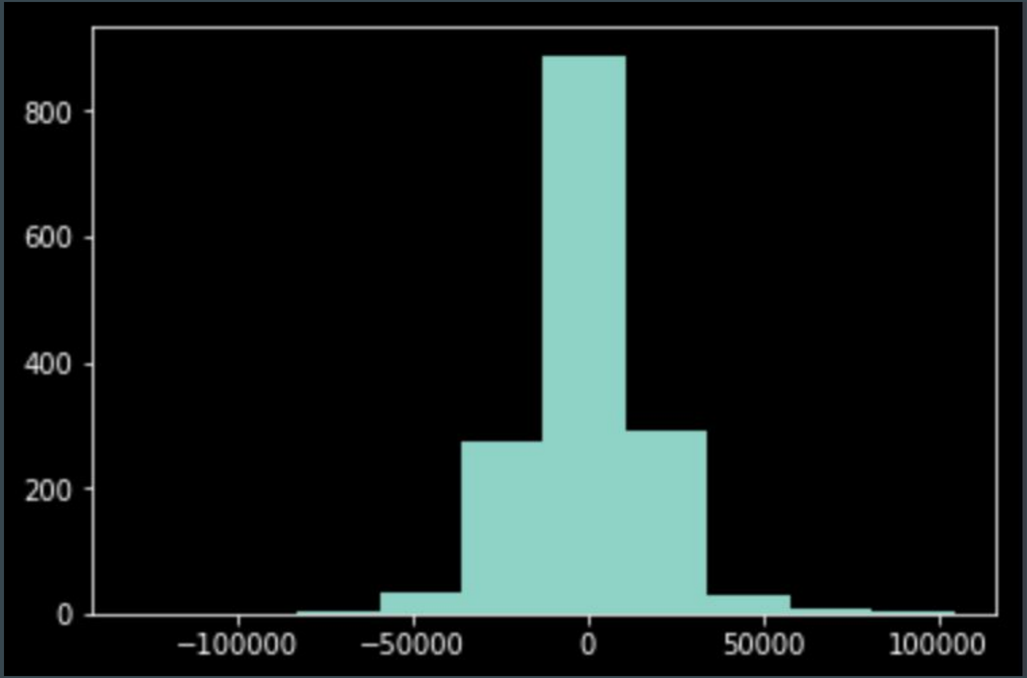


Data dictionary indicates the two outliers in current data may be due to unusual partial sales of large properties.

Modelling

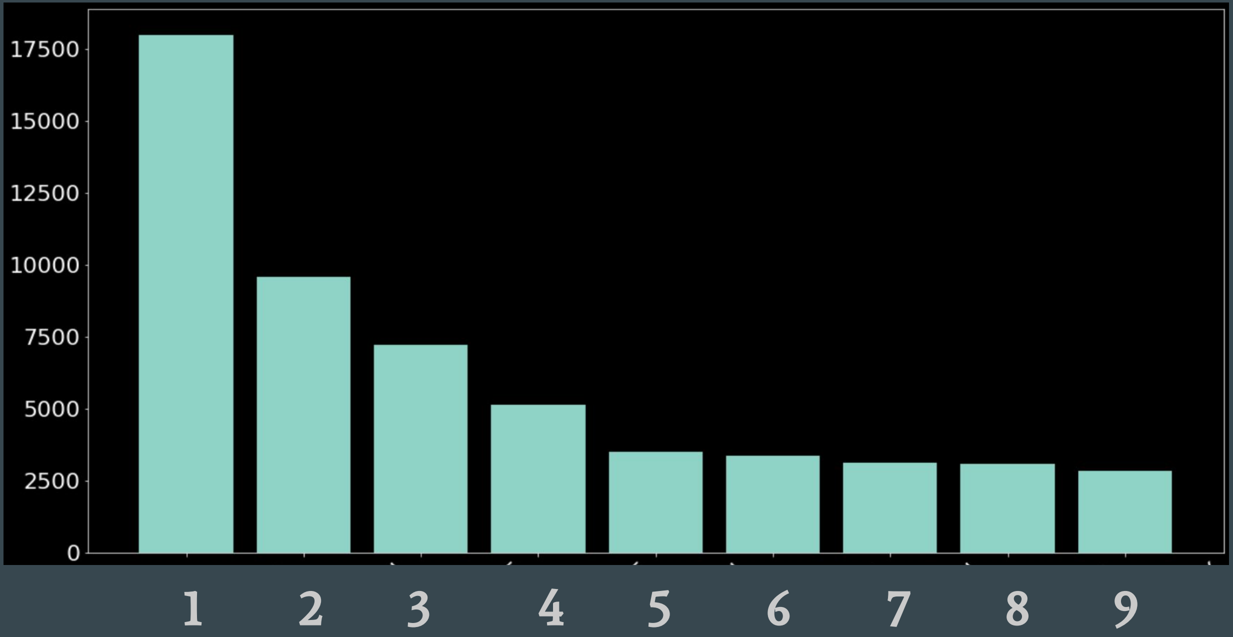
Used Lasso; ran variations (taking the natural log scored higher)

Residuals Distribution



Lasso optimal alpha: 40.708
CV score mean (train): 0.924
CV score mean (test): 0.918

Bottom Line: Quality, Space, and Basements Matter



1. Overall Qual / Gr Liv Area
2. Gr Liv Area / Kitchen Qual
3. Bsmt Qual / Bsmt Sqft
4. Overall Qual / Bsmt Sqft
5. Lot Area / Overall Qual
6. Bsmt Exposure / Gr Liv Area
7. Exter Qual / Functionality
8. Overall Qual / Garage Area
9. Lot Area / Paved Drive

Next steps: GridSearch, Random Forest, deeper EDA