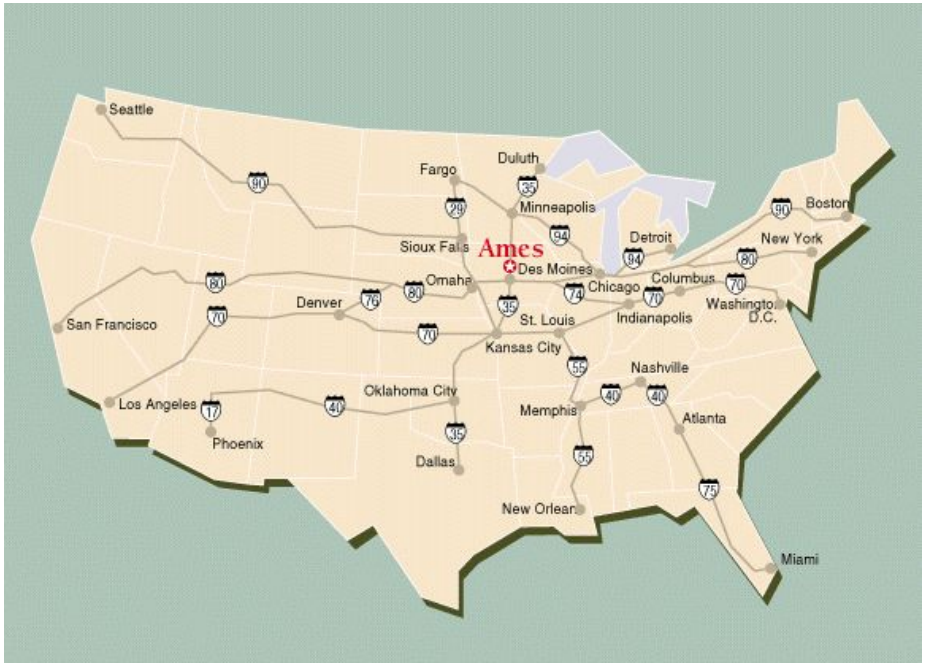
The Housing Market in Ames, IA

By Adriana Chacon
October 9, 2009

Project Goal

- To create a regression model based on the Ames Housing Dataset to predict the price of a house at sale



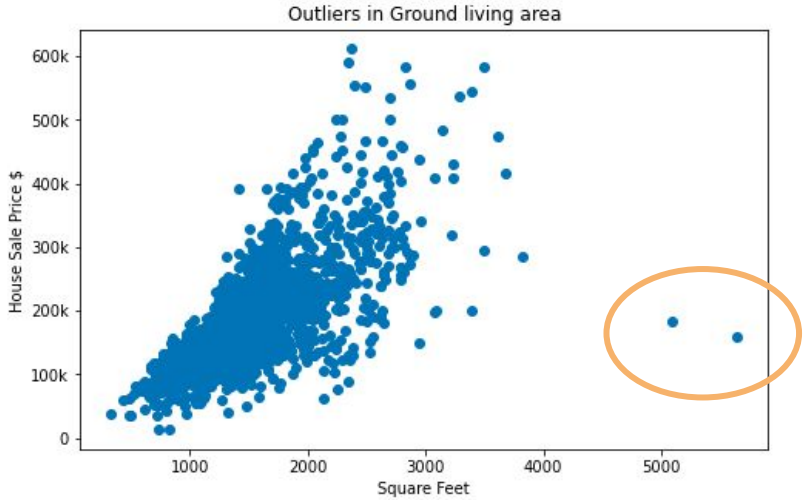
Dataset Information

- Data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA
- Data from 2006 to 2010
- 2930 observations
- 82 features: Categorical and numerical
- Some features were removed from the model because of low correlation to target, collinearity with other features, or because they were combined with other similar features.
- A few null values were removed from the dataset (less than 5)
- Two outliers were removed from the dataset

Cleaning Data

Removed features which content was included in another column

	BsmtFin SF 1	BsmtFin SF 2	Bsmt Unf SF	Total Bsmt SF
0	533.0	0.0	192.0	725.0
1	637.0	0.0	276.0	913.0
2	731.0	0.0	326.0	1057.0
3	0.0	0.0	384.0	384.0
4	0.0	0.0	676.0	676.0



Removed these outliers

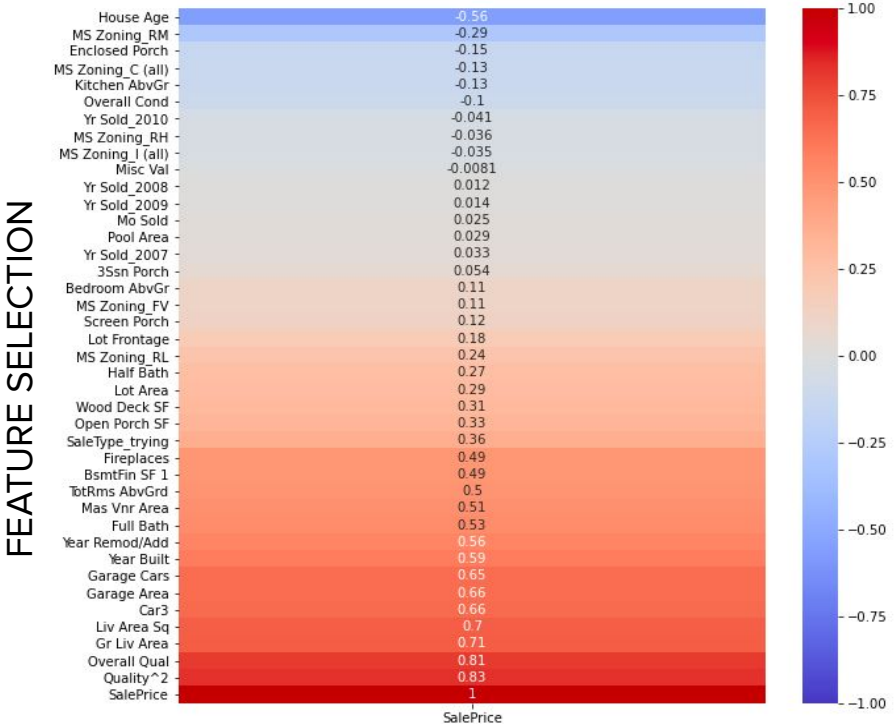
	Utilities	Condition 2	Roof Matl	Heating
count	2051	2051	2051	2051
unique	3	8	6	5
top	AllPub	Norm	CompShg	GasA
freq	2049	2025	2025	2018

Removed features where all info was the same

Data Analysis - Summary

- The following analysis was conducted using multilinear regression
- Categorical features were hot coded
- Polynomial features were explored but it was determined not to improve the model metrics therefore, it was not used in the final model.
- Regularization was used to shrink the data values
- Models used: linear regression, Lasso, Ridge, Lasso CV, Ridge CV.
- The Lasso regression model was selected to help with the high levels of multicollinearity in the data, and to automate the feature selection.
- The Lasso CV linear regression model was used to iterate over the alphas. Cross-validation was used to select the best model.

First Model - Linear Regression



Metrics:

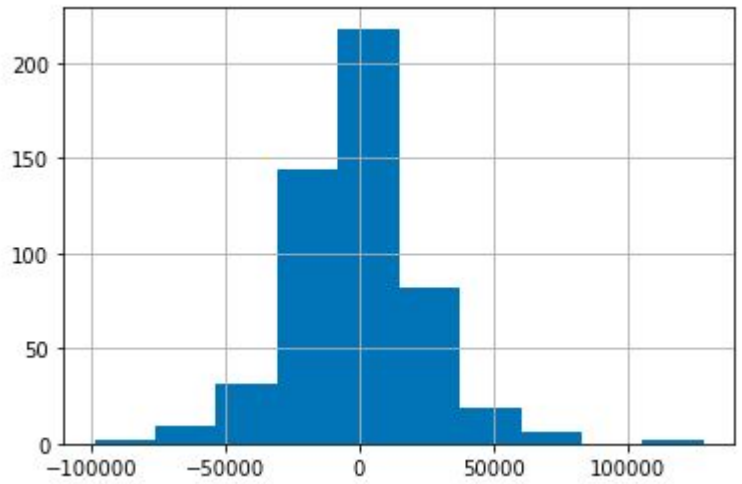
R2 Score: 0.8117859588898988

MSE: 1197036605.3354926

RMSE: 34598.21679415707

Evaluation of Regression Model

Residuals Plot for Lasso CV



Baseline:

Baseline: Avg Price = $181, 534
R2 Score: -0.0001046358443927975
MSE: 6360640530.379699
RMSE: 79753.62393258189

Simple Linear Regression:

R2 Score: 0.8117859588898988
MSE: 1197036605.3354926
RMSE: 34598.21679415707

Lasso CV Linear Model

R2 Score: 0.9071028931402941
MSE: 590823281.753886
RMSE: 24306.85668188888

Good Predictors for House Price Sale

- Ground Living Area
- Total Basement Square Foot
- Exterior Quality
- Kitchen Quality
- Overall Quality
- Neighborhood
- Garage Area

Features that increase the value of houses

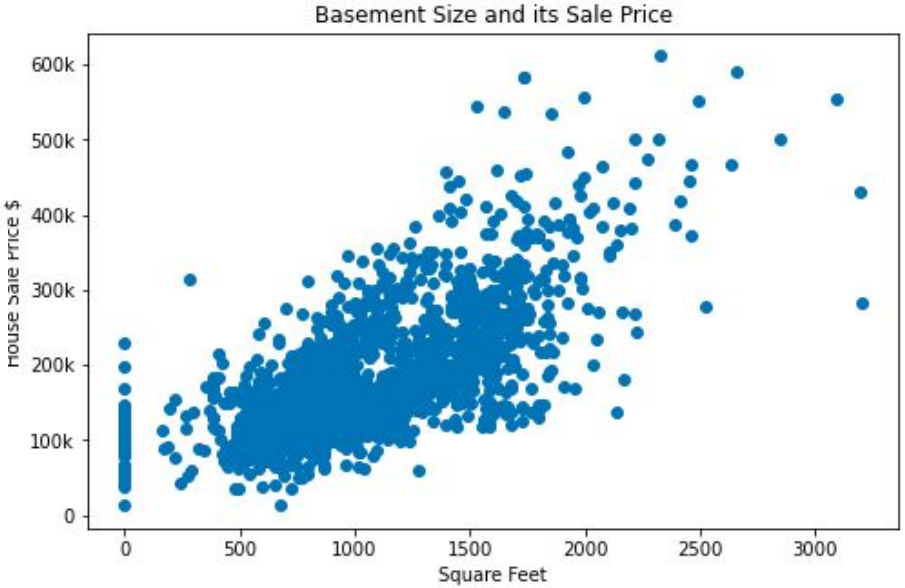
Interpreting the Coefficients

Feature	coef_value
Gr Liv Area	21849.259288
Overall Qual	13883.368978
Neighborhood_NridgHt	8132.269235
Total Bsmt SF	7776.775467
Year Built	6370.704445
Land Contour_HLS	6141.973801
Neighborhood_StoneBr	5823.167662
Misc Feature_Gar2	5549.186055
Garage Cars	5047.728593
Overall Cond	4996.140897

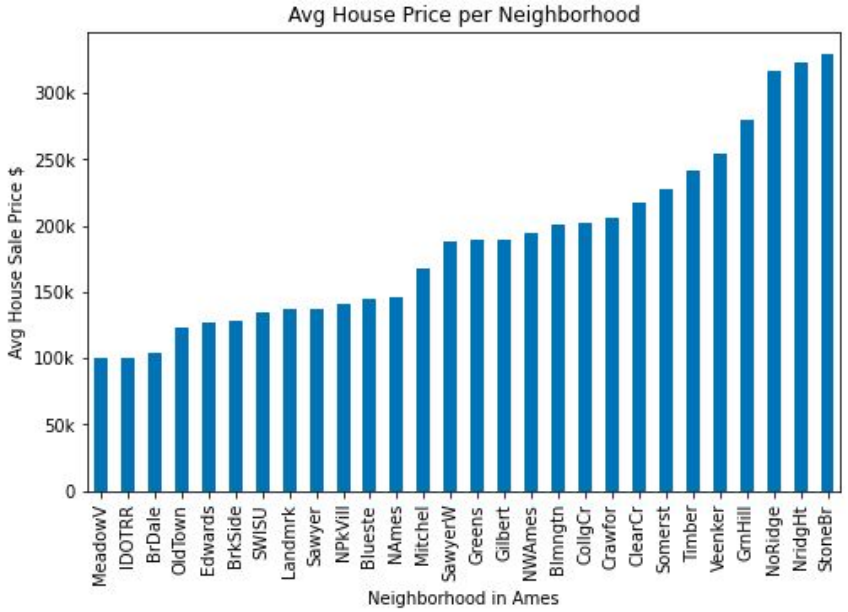
Feature	coef_value
Exter Qual_TA	-13460.273961
Kitchen Qual_TA	-12618.635764
Kitchen Qual_Gd	-11099.625903
Misc Val	-10664.719542
Exter Qual_Gd	-10006.036263
Bsmt Qual_Gd	-8279.455459
Bsmt Qual_TA	-6680.433253
Neighborhood_OldTown	-5248.145886

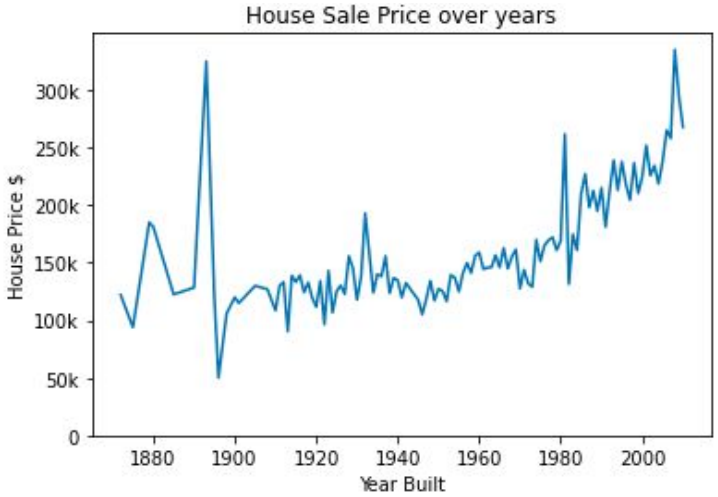
An increase in one standard deviation of 'Feature', means an increase by ‘coef_value’ of the sale price (holding the rest of the features constant.)

Property Size and Price

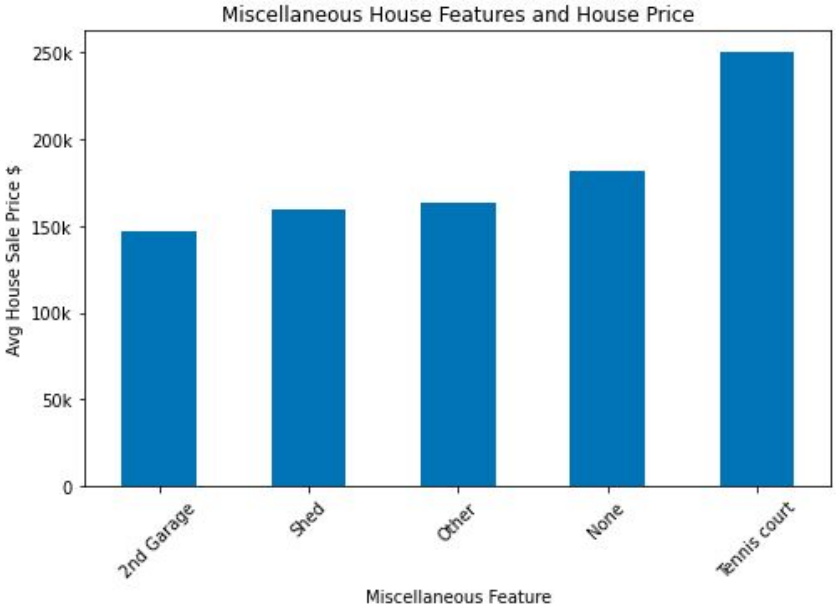


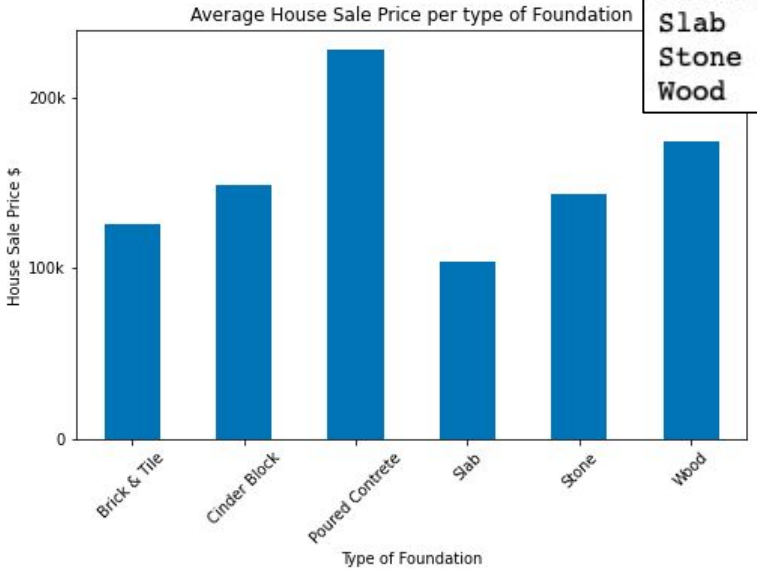
Neighborhoods and Year Built





Observations





PConc	923
CBlock	863
BrkTil	218
Slab	33
Stone	5
Wood	2

Results

- Developed a Lasso CV regression model to predict 90% of the housing price variance.
- Used Lasso CV to automate the feature selection process.
- Determined the features that have a strong negative and positive correlation with Sale Price

Recommendations and Next Steps

- Homeowners can increase the values of their properties by improving the quality of their kitchen and/or exterior covering of the house
- Houses in these neighborhoods would be a good investment: Stone Brook, Northridge Heights, Northridge, Green Hills
- This model is automated to clean and process data, and choose the best predictors for house pricing, therefore it can be used in other cities.
- The next step for this model would be to break down the analysis by type of house (no. of bedrooms, neighborhood, type of home) to provide more insights to specific homeowners.

Questions?