AMES HOUSING PROJECT

Evonne, Pallavi, Sharmaine

Zillow®
Zestimates

WRONG

- ★ Popular
- ★ Zestimates
- ★ Check Price

The Property Watch Of AMES, IA

- *What determines the price of a property?*
- *What property features are most important in accurately predicting sale price*
- *What kind of approach yield the most accurate prediction*

Data Review

- 2 sets of Ames Housing datasets were provided dated from 2006-2010
- **Training data** : 2051 rows of observations with 81 columns
- **Test data**: 879 rows of observation with 80 columns
- The train dataset has 81 columns which includes 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers).

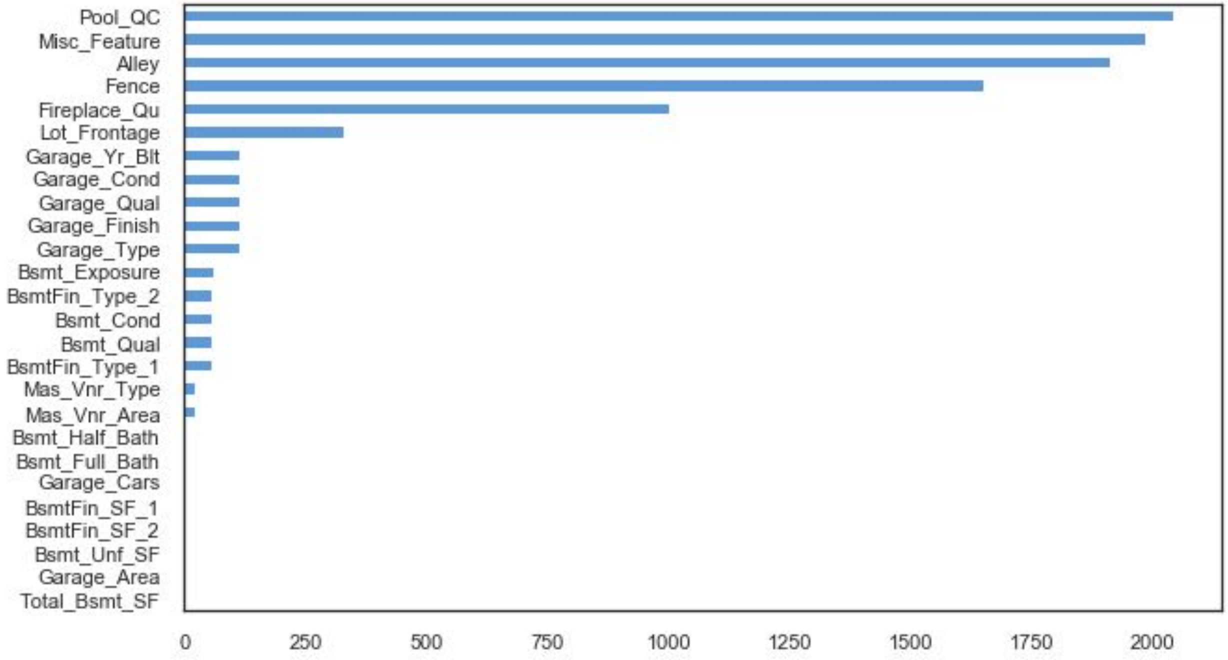
Procedure / Methodology

1. Data Cleaning
2. EDA
3. Feature Engineering
4. Model Prep
5. Cross Validation
6. Model Fitting (Linear Regression, Ridge, or Lasso)

Data Cleaning

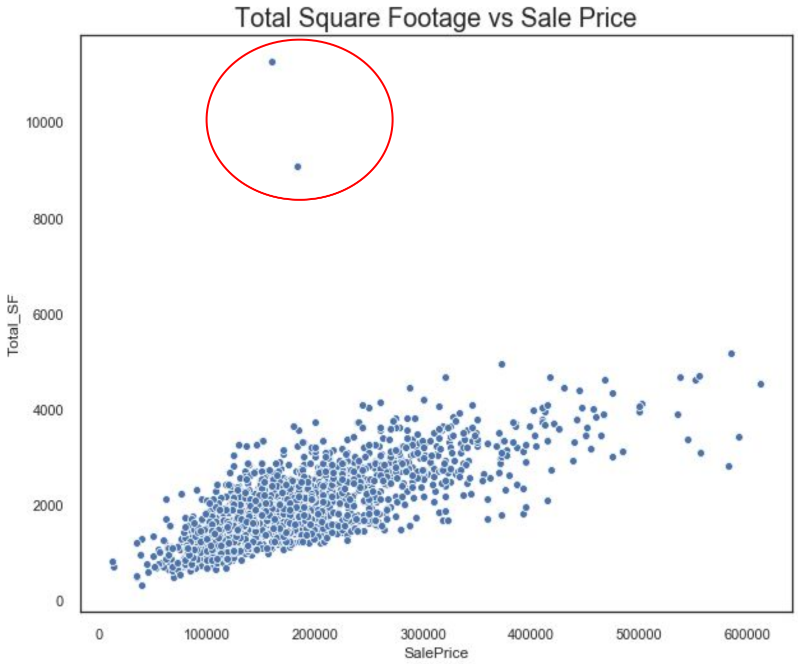
- Replacing the ‘null’ values with ‘None’
- Imputing Missing Values
- Converting all ordinal variables to numeric values
- Dummying the categorical variables
- Identifying and removing the outliers

Columns with the null values

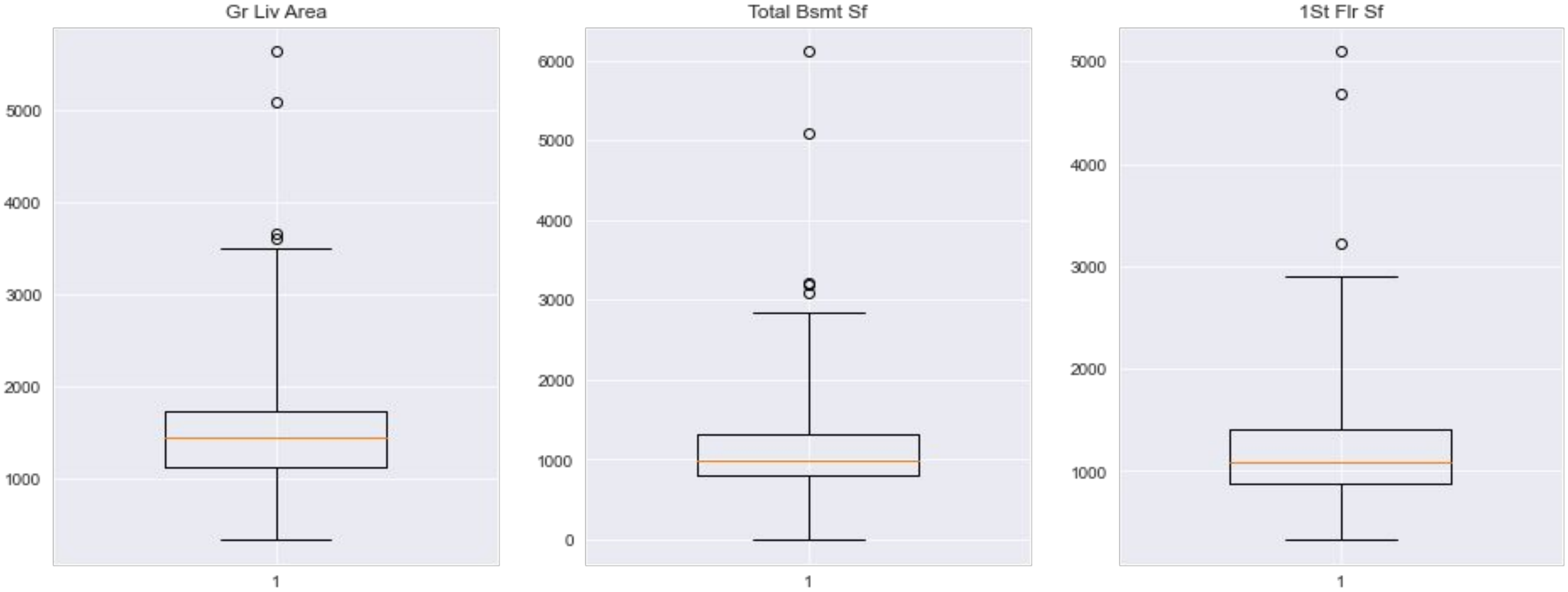


Outliers


Gr Living Area vs Sale Price
Gr_Liv_Area
5000
4000
3000
2000
1000
0
100000
200000
300000
400000
500000
600000
SalePrice

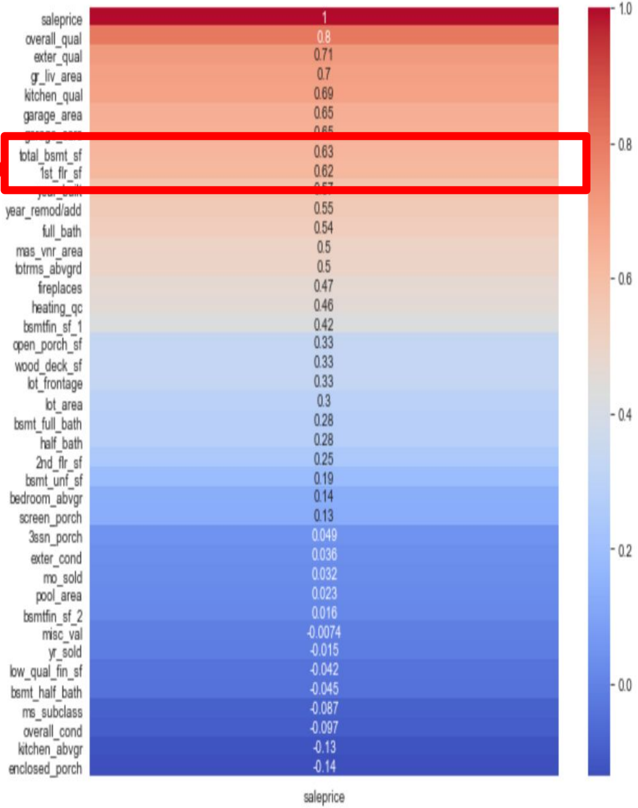

Total Square Footage vs Sale Price
Total_SF
10000
8000
6000
4000
2000
0
0
100000
200000
300000
400000
500000
600000
SalePrice

Boxplot

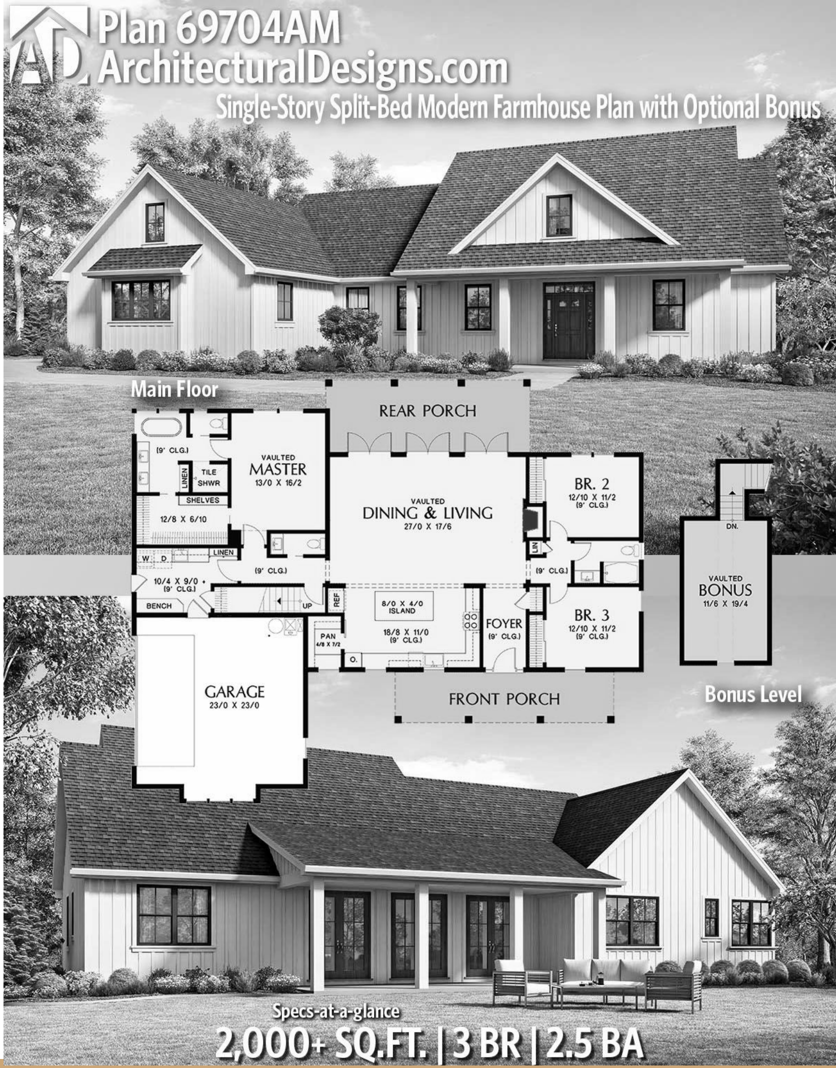

Gr Liv Area
5000
4000
3000
2000
1000
1
Total Bsmt Sf
6000
5000
4000
3000
2000
1000
0
1
1St Flr Sf
5000
4000
3000
2000
1000
1

Feature Engineering

With the Use Heatmap



- **Total Basement Sqft (Corr .63)**
- **1st Floor Sqft (Corr. 62)**



Feature Engineering

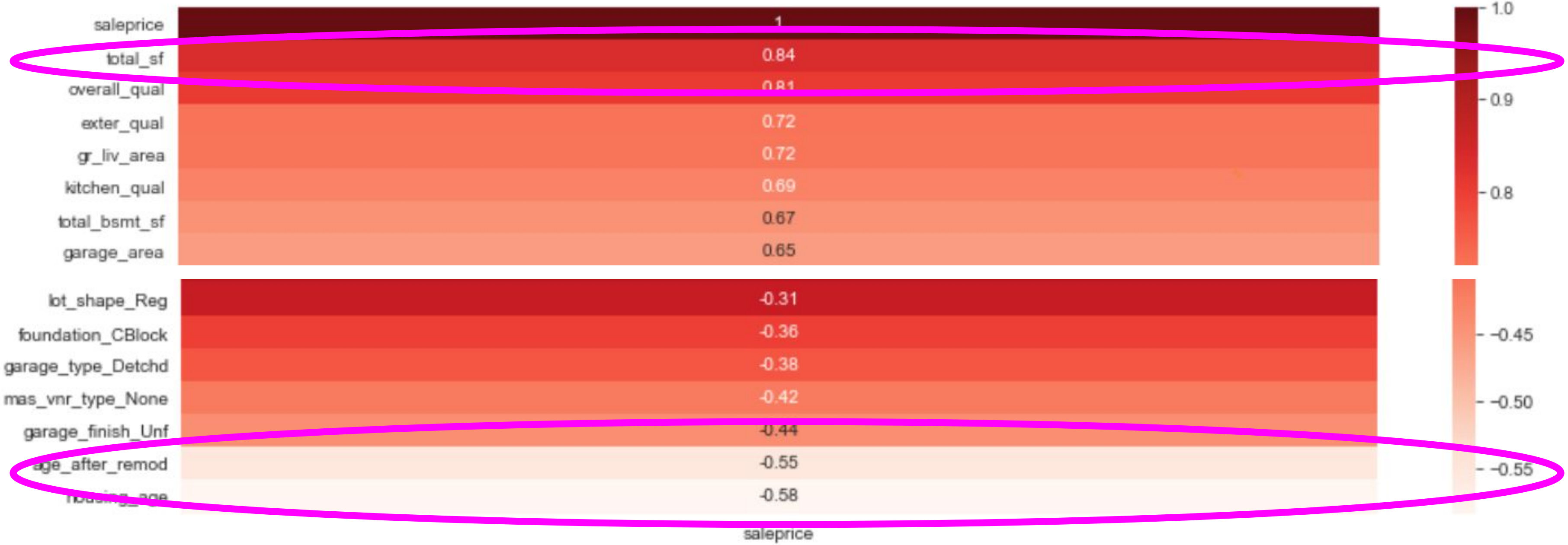
OLD Vs NEW

- Year Sold
- Year Built
- Year of Remodeling

Dummies Variables

- Categorical
- Nominal

Feature Engineering (NEWLY IMPROVED)



Problem | Methodology | Key Visualisations | Primary Findings | Recommendations

Model Selection

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Model	R2 score	Selection
Linear regression	-1.3232	
Ridge	0.8994	
Lasso	0.9114	

Model Selection

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Model	R2 score
Train set	0.9310
Test set	0.9129

Goal: Get R2 as **close to 1** as possible.

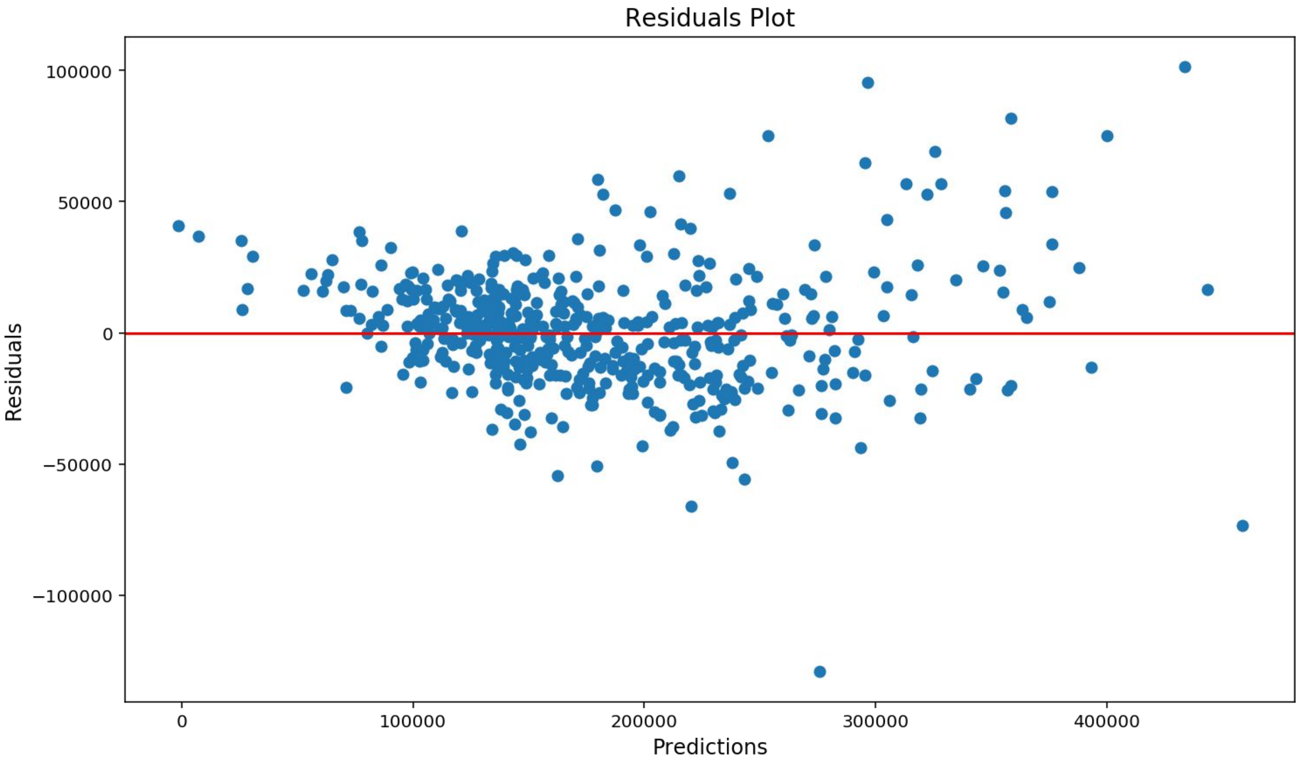
Model Selection

$$RMSE(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Model	RMSE
Baseline	78869.43
Test set	22298.09

Goal: Get RMSE as **close to 0** as possible.

Model Selection



Model Coefficients

	variable	coef	abs_coef
29	Gr Liv Area	16492.114	16492.114
58	total_sf	13025.572	13025.572
8	Overall Qual	12160.959	12160.959
19	BsmtFin SF 1	8910.810	8910.810
94	Neighborhood_NridgHt	7727.090	7727.090
10	Year Built	6695.917	6695.917
13	Exter Qual	6461.551	6461.551
12	Mas Vnr Area	5853.158	5853.158
100	Neighborhood_StoneBr	5481.017	5481.017
9	Overall Cond	5106.784	5106.784
17	Bsmt Exposure	4762.552	4762.552
36	Kitchen Qual	4192.917	4192.917
4	Lot Area	3975.733	3975.733
43	Garage Area	3610.511	3610.511

Recommendations

	variable	coef	abs_coef
29	Gr Liv Area	16492.114	16492.114
58	total_sf	13025.572	13025.572
8	Overall Qual	12160.959	12160.959
19	BsmtFin SF 1	8910.810	8910.810
94	Neighborhood_NridgHt	7727.090	7727.090
10	Year Built	6695.917	6695.917
13	Exter Qual	6461.551	6461.551
12	Mas Vnr Area	5853.158	5853.158
100	Neighborhood_StoneBr	5481.017	5481.017
9	Overall Cond	5106.784	5106.784
17	Bsmt Exposure	4762.552	4762.552
36	Kitchen Qual	4192.917	4192.917
4	Lot Area	3975.733	3975.733
43	Garage Area	3610.511	3610.511

Features that appear to add the Most Value to a Home :

Living Area, Total Sq feet, Overall Quality

Recommendations

	variable	coef	abs_coef
29	Gr Liv Area	16492.114	16492.114
58	total_sf	13025.572	13025.572
8	Overall Qual	12160.959	12160.959
19	BsmtFin SF 1	8910.810	8910.810
94	Neighborhood_NridgHt	7727.090	7727.090
10	Year Built	6695.917	6695.917
13	Exter Qual	6461.551	6461.551
12	$\boxed{\text{Mas Vnr Area}}$	5853.158	5853.158
100	Neighborhood_StoneBr	5481.017	5481.017
9	Overall Cond	5106.784	5106.784
17	$\boxed{\text{Bsmt Exposure}}$	4762.552	4762.552
36	$\boxed{\text{Kitchen Qual}}$	4192.917	4192.917
4	Lot Area	3975.733	3975.733
43	Garage Area	3610.511	3610.511

Features that could increase the value of the house:

Overall Quality, Overall Condition

Masonry veneer area, Basement Exposure, Kitchen Quality

Recommendations

	variable	coef	abs_coef
29	Gr Liv Area	16492.114	16492.114
58	total_sf	13025.572	13025.572
8	Overall Qual	12160.959	12160.959
19	BsmtFin SF 1	8910.810	8910.810
94	$\boxed{\text{Neighborhood\_NridgHt}}$	7727.090	7727.090
10	Year Built	6695.917	6695.917
13	Exter Qual	6461.551	6461.551
12	Mas Vnr Area	5853.158	5853.158
100	$\boxed{\text{Neighborhood\_StoneBr}}$	5481.017	5481.017
9	Overall Cond	5106.784	5106.784
17	Bsmt Exposure	4762.552	4762.552
36	Kitchen Qual	4192.917	4192.917
4	Lot Area	3975.733	3975.733
43	Garage Area	3610.511	3610.511

Neighbourhoods Seem Like a Good Investment:

Northridge Heights, Stone Brooke

Recommendations

	variable	coef	abs_coef
2	MS SubClass	-2823.937	2823.937
121	Bldg Type_TwnhsE	-1870.274	1870.274
132	Roof Style_Mansard	-1605.834	1605.834
120	Bldg Type_Twnhs	-1419.234	1419.234
16	Bsmt Cond	-1260.746	1260.746
34	Bedroom AbvGr	-1108.283	1108.283
119	Bldg Type_Duplex	-1107.471	1107.471
59	age	-1035.782	1035.782
166	Mas Vnr Type_BrkFace	-946.817	946.817
176	Heating_OthW	-778.322	778.322

Features that appears to hurt the home prices the most:

Roof Style - Mansard, Unfinished Basement

THANK YOU