Aimes Iowa Housing Price Modeling

General Assembly: DSI-6

By: Dmitriy Pavlov

12/07/18

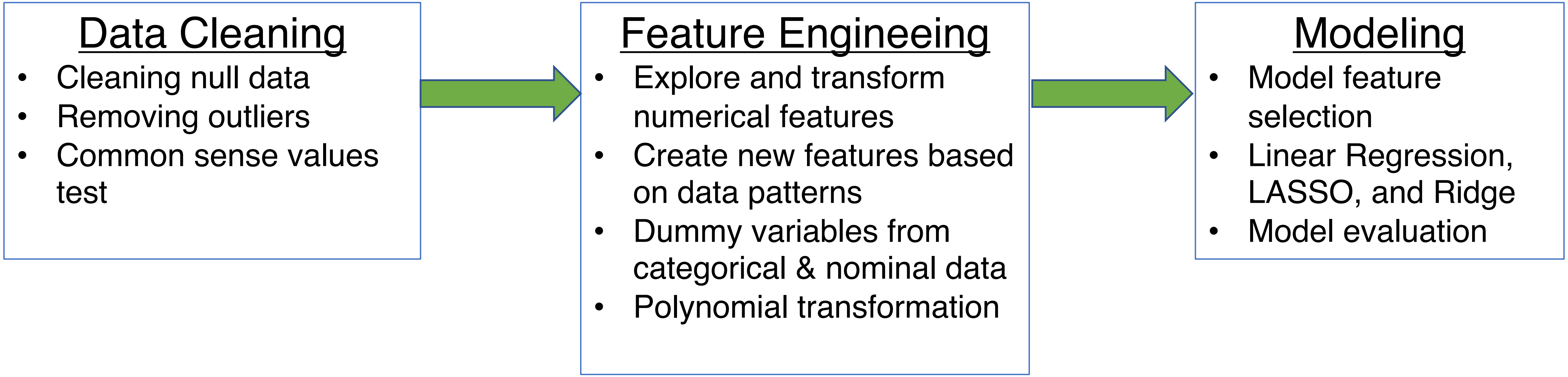
Overview

- Data Science Problem
- Modeling Procedure
  - Data Cleaning
  - Feature Engineering
  - Modeling
- Findings
- Recommendations

Data Science Problem:

Can we use data to predict housing prices? And if so what are those features and what is their impact?

Modeling Procedure Overview



Data Cleaning

- Cleaning null data
- Removing outliers
- Common sense values test

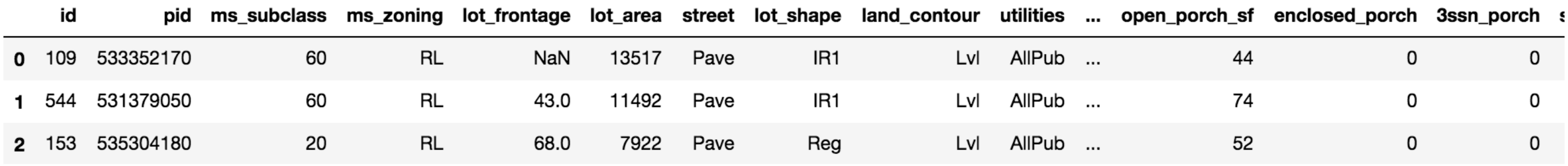
Feature Engineeing

- Explore and transform numerical features
- Create new features based on data patterns
- Dummy variables from categorical & nominal data
- Polynomial transformation

Modeling

- Model feature selection
- Linear Regression, LASSO, and Ridge
- Model evaluation

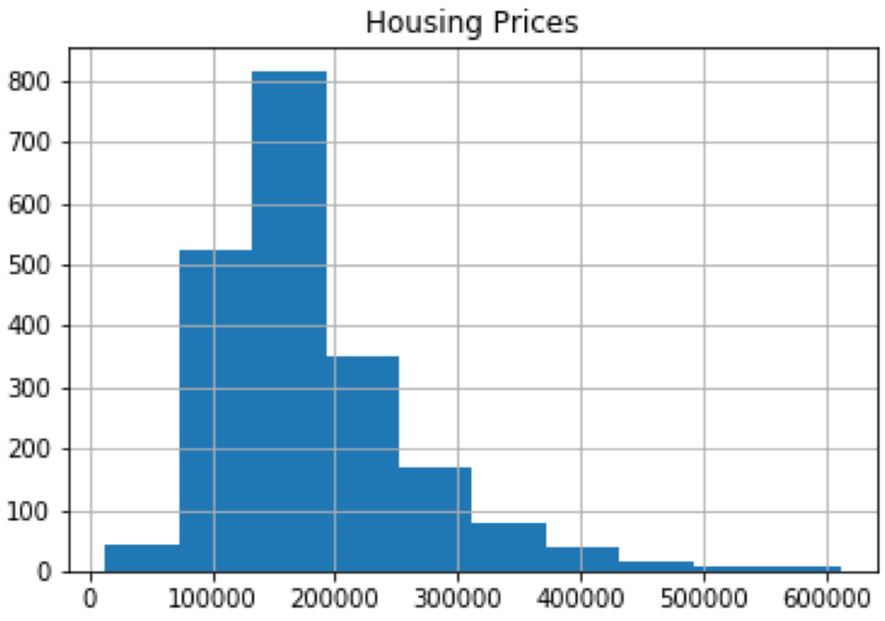
Data Cleaning: Goal is to clean our data and make sure we can trust it going forward

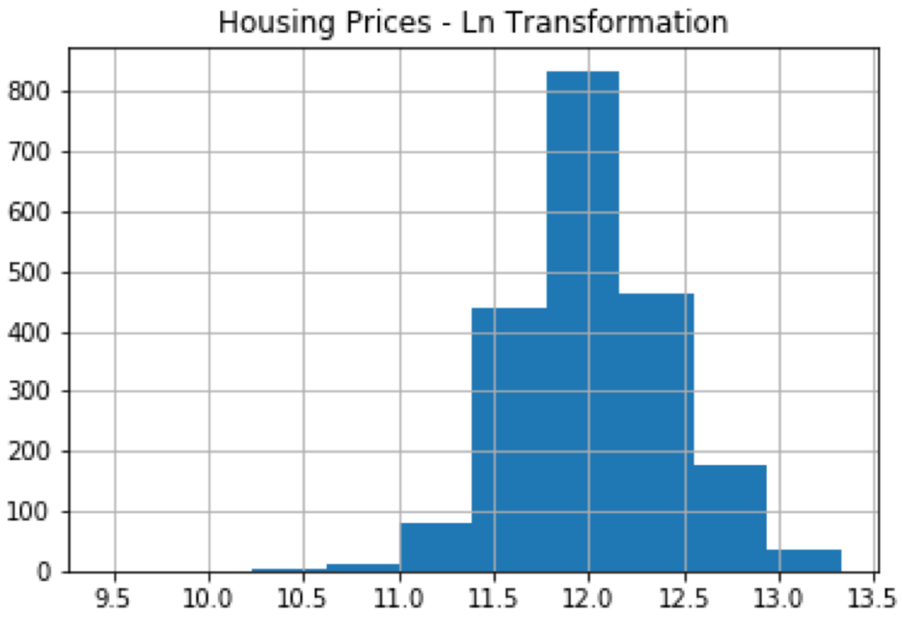


	id	pid	ms_subclass	ms_zoning	lot_frontage	lot_area	street	lot_shape	land_contour	utilities	...	open_porch_sf	enclosed_porch	3ssn_porch	s
0	109	533352170	60	RL	NaN	13517	Pave	IR1	Lvl	AllPub	...	44	0	0	
1	544	531379050	60	RL	43.0	11492	Pave	IR1	Lvl	AllPub	...	74	0	0	
2	153	535304180	20	RL	68.0	7922	Pave	Reg	Lvl	AllPub	...	52	0	0	

- Address null data value and possibly drop the features
- Common sense test – do our feature values make sense
- Address outliers

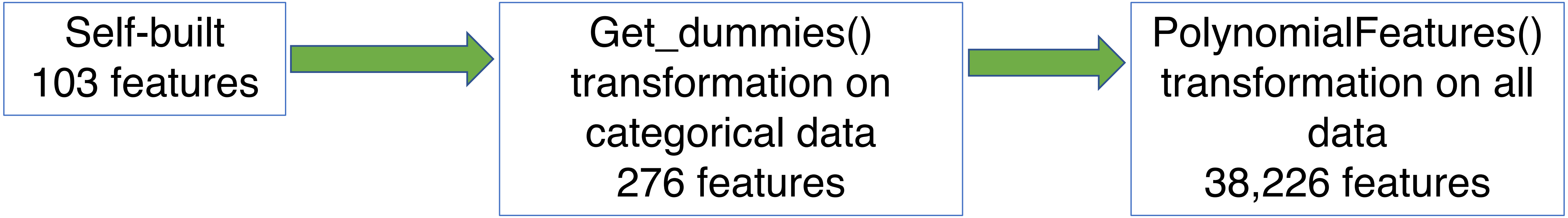
Feature Engineering – Created 28 of my own features via transformations, true/false, and setting level thresholds





Check through histograms and scatter plots vs housing prices to create additional features to help normalize the distributions and better separate the data. Going from 75 to 103 total features.

Feature Engineering – Leverage Python libraries' functions to explore more possible features

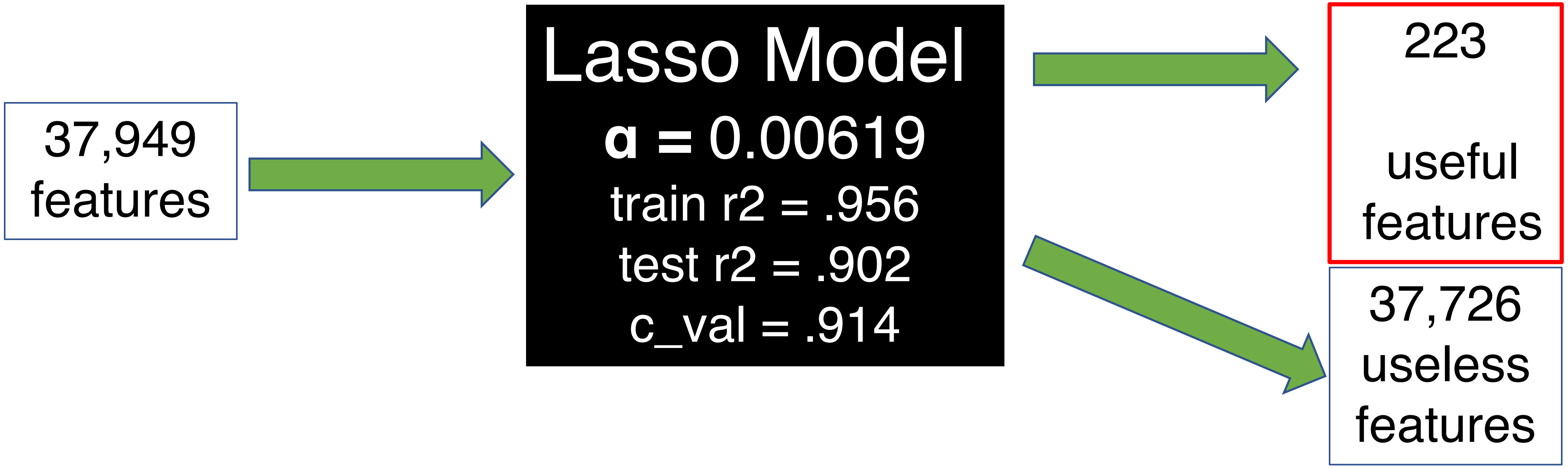


Self-built
103 features

Get_dummies()
transformation on categorical data
276 features

PolynomialFeatures()
transformation on all data
38,226 features

Modeling – Lasso model was able to eliminate 98.9% of the features, leaving us with only 400



37,949 features

Lasso Model

ɑ = 0.00619

train r2 = .956

test r2 = .902

c_val = .914

223

useful features

37,726 useless features

Modeling – We were able to increase model performance when modeling using ‘useful’ features

Lasso Model

α = 0.00198

train r2 = .959

test r2 = .917

c_val = .941

test: +1.5 ppts

c_val: +2.7 ppts

Ridge Model

α = 0.004348

train r2 = .961

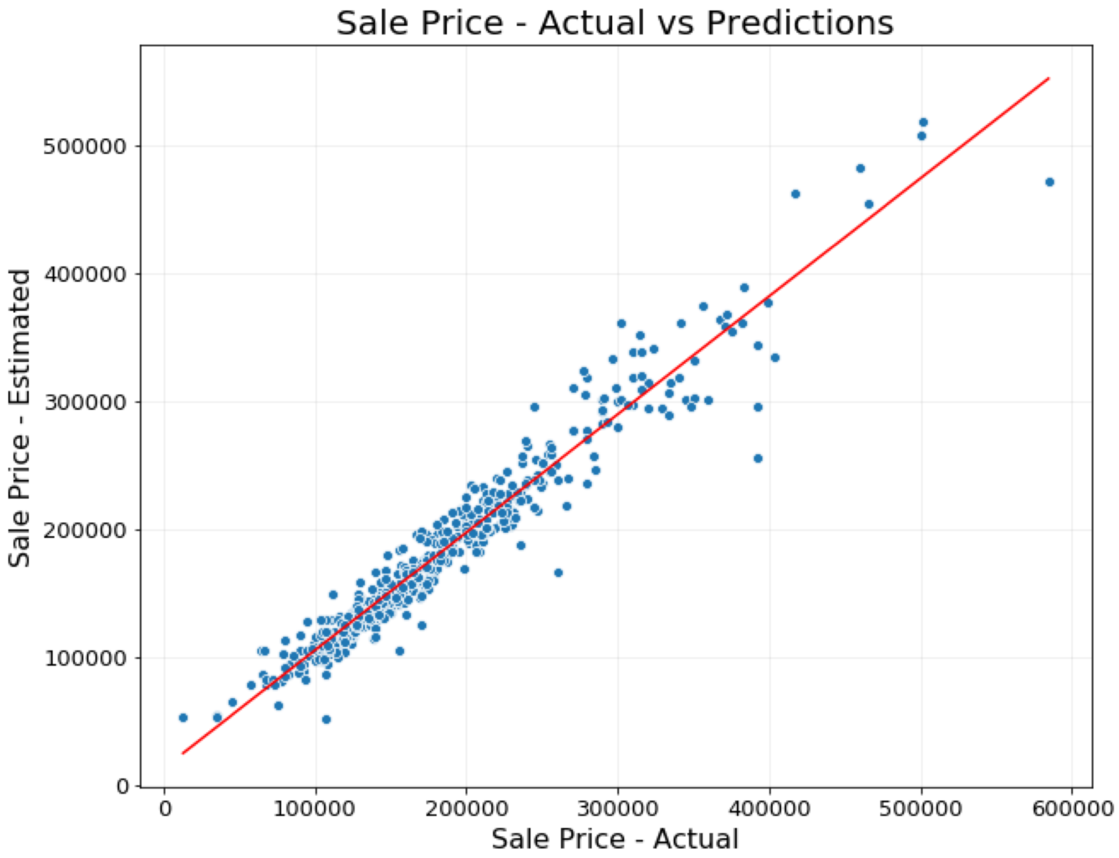
test r2 = .912

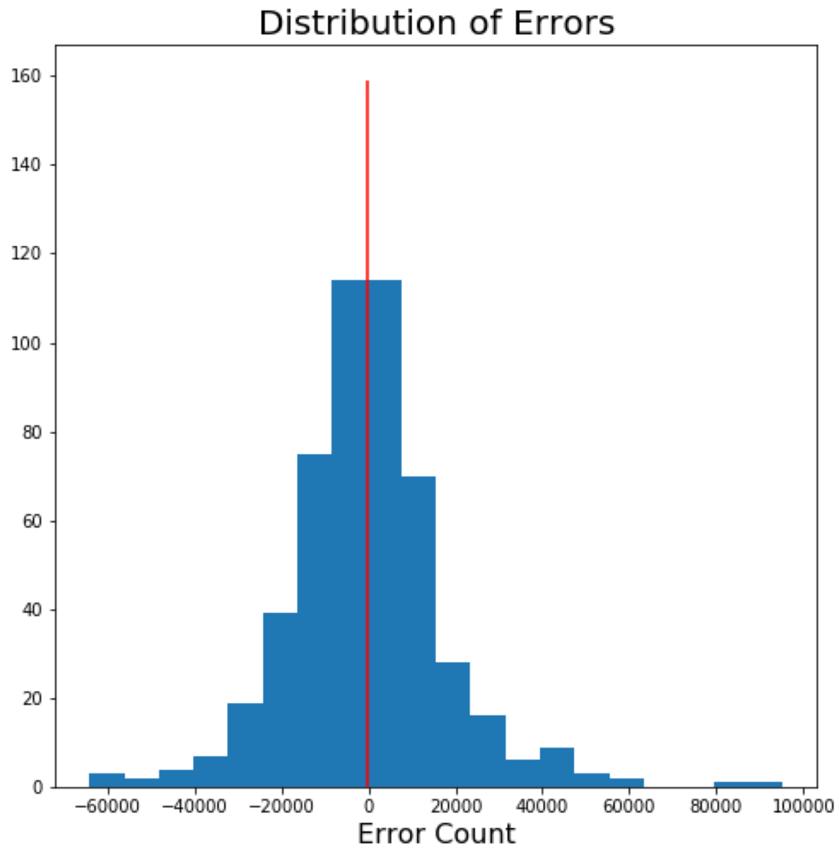
c_val = .941

test: +1.0 ppts

c_val: +2.7 ppts

Modeling – The the errors of our model are normally distributed and there are no patterns throughout the graph





Findings – List of top features with positive influence on the sale price of the house

Feature	Explanation	Occurrences
Gr liv area	Above grade (ground) living area square feet	5
Overall cond	Rates the overall material and finish of the house	3
Lot Area Log	Lot size in square feet – ln transformed	3
Year Built	Original construction date	2
Year Remod/add	Remodel date (same as construction date if no remodeling or additions)	2
Functional_Typ	Typical Functionality	2

Findings – List of top features with negative influence on the sale price of the house

Feature	**Explanation**	**Occurrences**
Ms Zone C	Commercial zoning classification	3
Garage Cond Fa	Fair garage condition	3
Overall Cond	Rates the overall condition of the home	2
Garage Yr Build	Year garage was built	2