Project 2 - Ames housing project

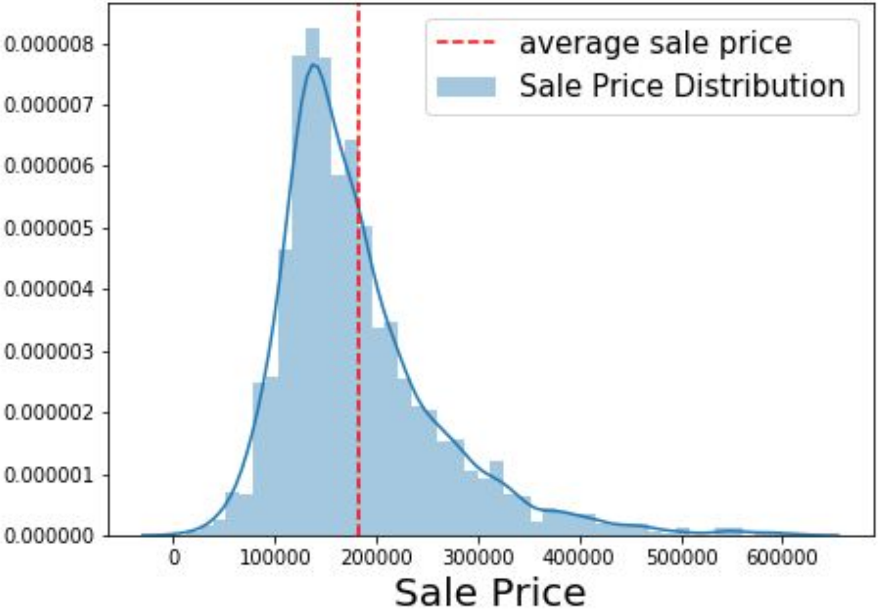
Chris Lee

EDA (Exploratory Data Analysis)

The data has 2051 cases (row) and 82 categories (column)

- 23 nominal (Sale Condition variable missing) - *object type - dummy*
- 23 ordinal (qualitative - quality / condition) *- object type - numerize*
- 14 discrete (quantitative - year / # of rooms) *- numeric*
- 20 continuous variables (area or size (square feet)) *- numeric*
- 2 additional observation identifiers (ID, PID)

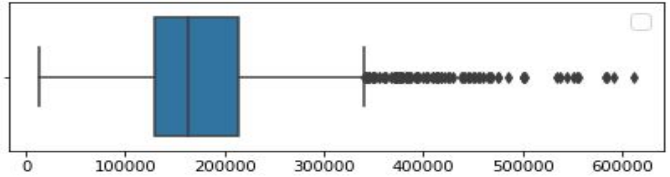
EDA (Exploratory Data Analysis)



Minimum sale price : $12,789

Average sale price : $181,470

Maximum sale price : $611,657

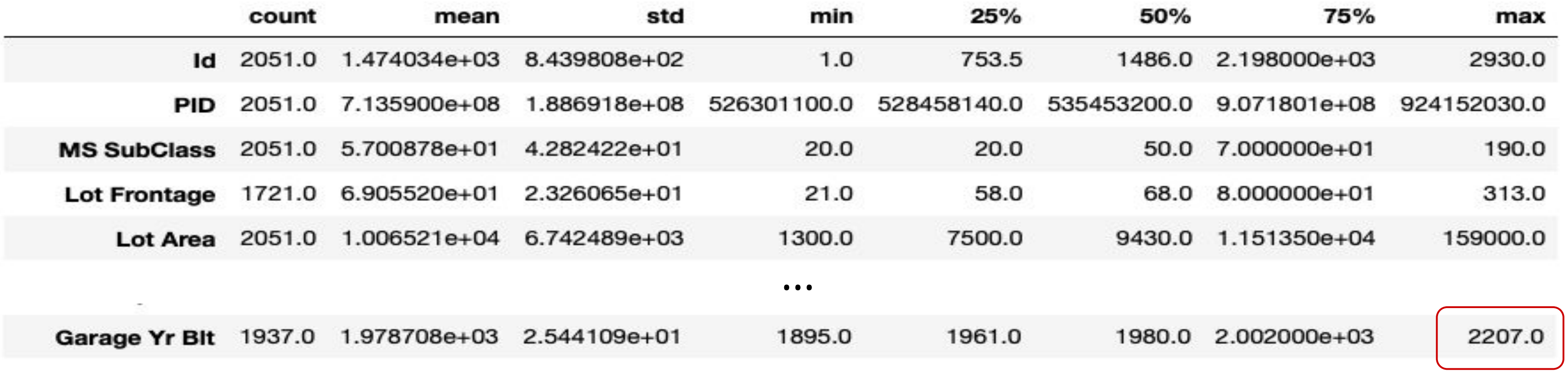


Data Cleaning

- 23 nominal - *object type (get dummy)*
- 23 ordinal - *object type (numeritize data)* * →
- 14 discrete - *numeric*
- 20 continuous variables - *numeric*
- 2 additional observation identifiers(ID, PID)

ex) Ex -> 5
Gd -> 4
TA -> 3
Fa -> 2
Po -> 1
NA -> 0

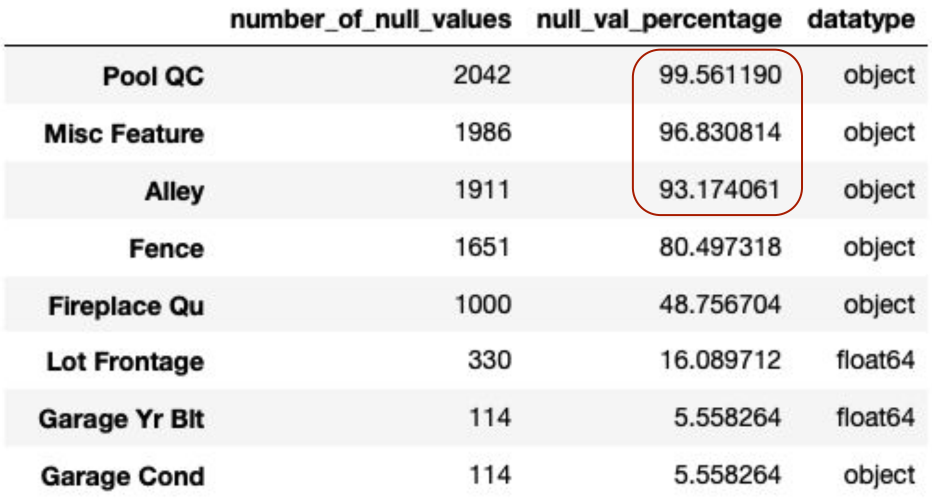
Data Cleaning



	count	mean	std	min	25%	50%	75%	max
Id	2051.0	1.474034e+03	8.439808e+02	1.0	753.5	1486.0	2.198000e+03	2930.0
PID	2051.0	7.135900e+08	1.886918e+08	526301100.0	528458140.0	535453200.0	9.071801e+08	924152030.0
MS SubClass	2051.0	5.700878e+01	4.282422e+01	20.0	20.0	50.0	7.000000e+01	190.0
Lot Frontage	1721.0	6.905520e+01	2.326065e+01	21.0	58.0	68.0	8.000000e+01	313.0
Lot Area	2051.0	1.006521e+04	6.742489e+03	1300.0	7500.0	9430.0	1.151350e+04	159000.0
				...				
Garage Yr Blt	1937.0	1.978708e+03	2.544109e+01	1895.0	1961.0	1980.0	2.002000e+03	2207.0

- Year built = 2006
- Year Remod/Add = 2007
- Garage Year built = 2207 ? -> adjusted to 2007

Data Cleaning



	number_of_null_values	null_val_percentage	datatype
Pool QC	2042	99.561190	object
Misc Feature	1986	96.830814	object
Alley	1911	93.174061	object
Fence	1651	80.497318	object
Fireplace Qu	1000	48.756704	object
Lot Frontage	330	16.089712	float64
Garage Yr Blt	114	5.558264	float64
Garage Cond	114	5.558264	object

.

.

.

- Pool QC
- Misc Feature
- Alley

more than 90% of null values

Remove three column since it won’t significantly impact our result

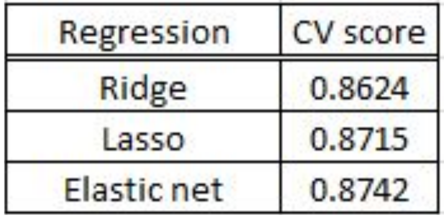
Rest of the columns are filled with 0 or ‘NA’ according to the data description.

Regression model selection

Linear regression

- Ridge regression - reduce variance by shrinking parameters
- Lasso regression - vanish useless parameters
- Elastic net regression (combination of Ridge and Lasso regression) *

- **Cross Validation Score**



Regression	CV score
Ridge	0.8624
Lasso	0.8715
Elastic net	0.8742

*

# Train score = 0.92652

# Test score   = 0.90255

Business recommendations

- Top 3 features from direct correlation was overall quality, above grade (ground) living area square feet, and garage area.
- Top 3 features from polynomial (interaction) correlation was
- There is almost no feature that has negative effect on the sale price.

To generalize the sale price prediction to other city,

We might need some more feature for the better prediction such as temperature, altitude, population, population density etc.