Achieving a better house price

Top 10 factors affecting housing price that you may not know!

Thomas
Lucus
Mark

We'll be able to help you..

1. Identify the top 10 features that affect your home price.
2. Predict your selling price (>85% accuracy)

Introducing your friendly real estate team:

1. Deep market experience
   a. 2000 transactions
   b. US$366M in properties
2. Experience selling different asset classes and building types
3. Data driven (we built a model)

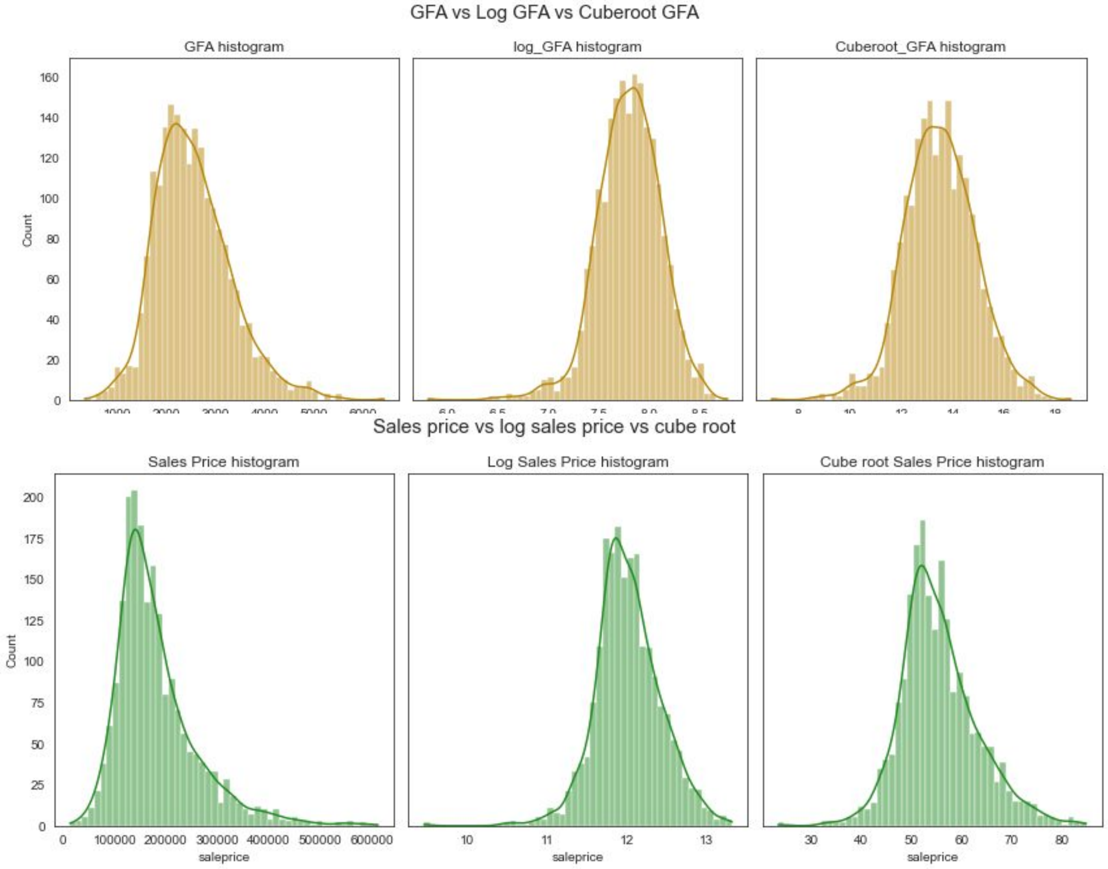
What does it mean to be data driven?

Column	Variable Type	Dataset	Description
id	Nominal	train.csv & test.csv	Observation ID of the property
pid	Nominal	train.csv & test.csv	Parcel identification number - can be used with city web site for parcel review.
ms_subclass	Nominal	train.csv & test.csv	Identifies the type of dwelling involved in the sale.
ms_zoning	Nominal	train.csv & test.csv	Identifies the general zoning classification of the sale.
lot_frontage	Continuous	train.csv & test.csv	Linear feet of street connected to property
lot_area	Continuous	train.csv & test.csv	Lot size in square feet
street	Nominal	train.csv & test.csv	Type of road access to property
alley	Nominal	train.csv & test.csv	Type of alley access to property
lot_shape	Ordinal	train.csv & test.csv	General shape of property
land_contour	Nominal	train.csv & test.csv	Flatness of the property
utilities	Ordinal	train.csv & test.csv	Type of utilities available
lot_config	Nominal	train.csv & test.csv	Lot configuration
land_slope	Ordinal	train.csv & test.csv	Slope of property
neighborhood	Nominal	train.csv & test.csv	Physical locations within Ames city limits
condition_1	Nominal	train.csv & test.csv	Proximity to various conditions
condition_2	Nominal	train.csv & test.csv	Proximity to various conditions (if more than one is present)
bldg_type	Nominal	train.csv & test.csv	Type of dwelling
house_style	Nominal	train.csv & test.csv	Style of dwelling
overall_qual	Ordinal	train.csv & test.csv	Rates the overall material and finish of the house
overall_cond	Ordinal	train.csv & test.csv	Rates the overall condition of the house

Cleaning the data

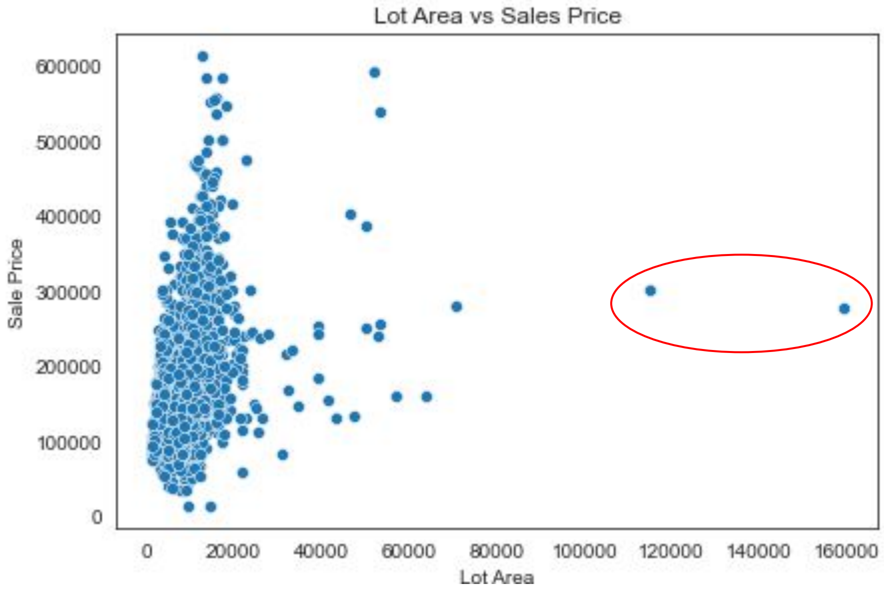
Column	Cleaning methodology
pool_qc	Column dropped
misc_feature	Column dropped
alley	Column dropped
fence	Column dropped
fireplace_qu	Column dropped
lot_frontage	Column dropped
garage_yr_blt	Column dropped
garage_finish	Column dropped
garage_qual	Column dropped
garage_cond	Column dropped
garage_type	Column dropped

Adjusting for skew



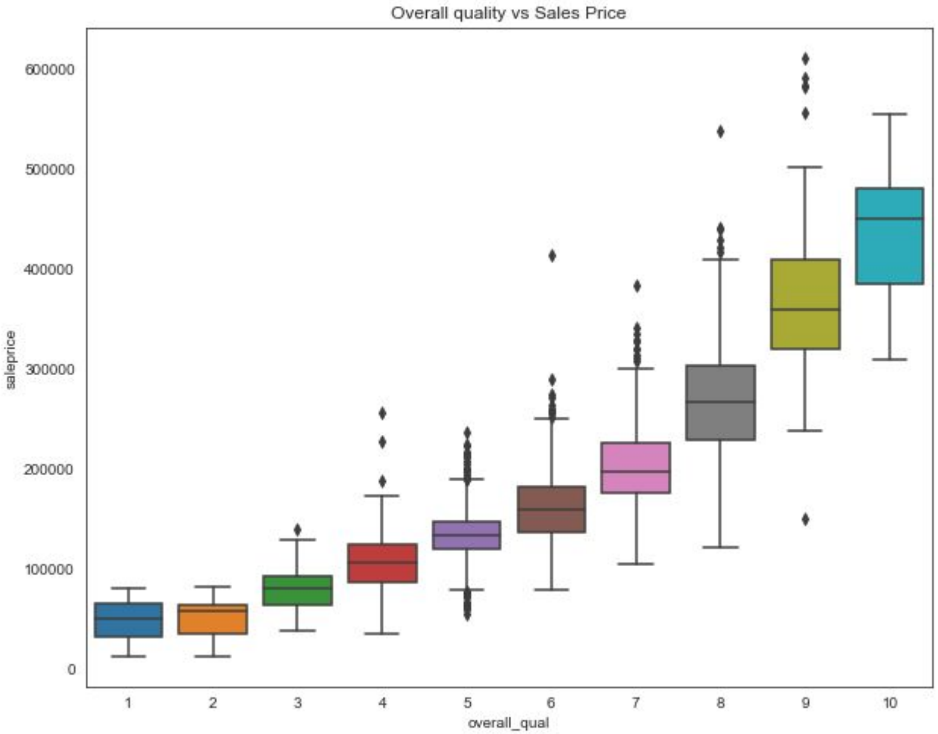
1. Apply either logarithmic or cube root functions to data to adjust for right skew.

Identification of outliers



Types of data (ordinal variables)

Score	Quality
10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor



Types of data: Categorical variables

Foundation	Brick & Tile Cinder Block Poured Contrete Slab Stone Wood
Central Air	Yes No
Garage Type	2 Types Attached Basement Built in Car Port Detached

EDA

Categorical Variables

Ordinal Variables

- that have a natural ordering while maintaining their class of values

Mapping for Ordinal Variables

Column	Description	Mapped values
lot_shape	General shape of property	'IR3'(Irregular): 1, ‘IR2'(Moderately Irregular): 2, ‘IR1'(Slightly irregular): 3, 'Reg'(Regular): 4
utilities	Type of utilities available	'ELO'(Electricity only): 1, 'NoSeWa'(Electricity and Gas Only): 2, 'NoSewr'(Electricity, Gas, and Water (Septic Tank) : 3, AllPub'(All public Utilities (E,G,W,& S): 4
land_slope	Slope of property	'Sev'(Severe Slope): 1, 'Mod'(Moderate Slope): 2, ’Gtl'(Gentle slope): 3

Categorical Variables

Ordinal Variables

- that have a natural ordering while maintaining their class of values

Nominal Variables:

- does not possess a natural order

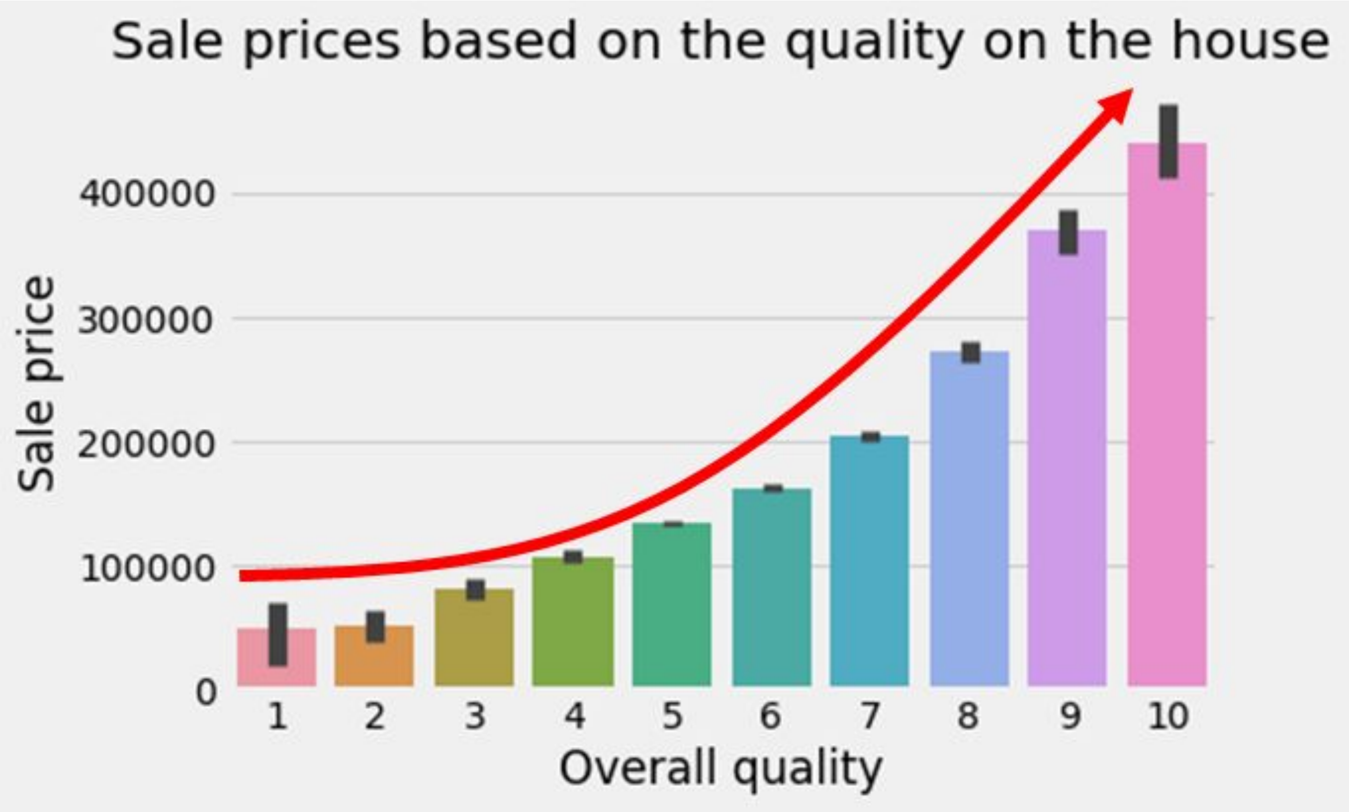
Dummify for Nominal Variables

Columns	Description
ms_subclass	Identifies the type of dwelling involved in the sale
ms_zoning	Identifies the general zoning classification of the sale
street	Type of road access to property

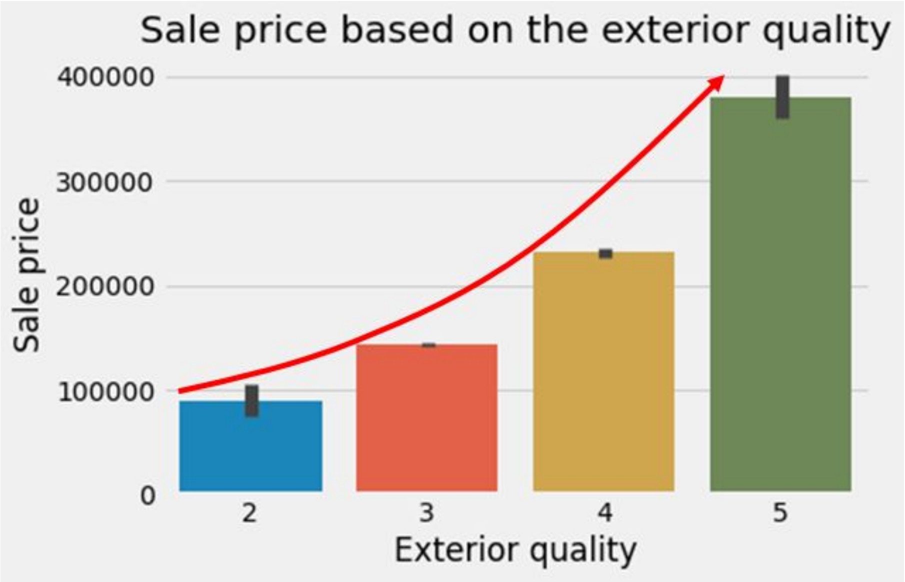
Highly correlated variables with sale price

Variables	Description	Correlation against Sale price
overall_qual	Rates the overall material and finish of the house	0.803462
gr_liv_area	Above grade (ground) living area square feet	0.719463
exter_qual	Evaluates the quality of the material on the exterior	0.715048
kitchen_qual	Kitchen quality	0.694295
total_bsmt_sf	Total square feet of basement area	0.665116
garage_area	Size of garage in square feet	0.655097
garage_cars	Size of garage in car capacity	0.648227
year_remod/add	Remodel date (same as construction date if no remodeling or additions)	0.572405
totrms_abvgrd	Total rooms above grade (does not include bathrooms)	0.509775

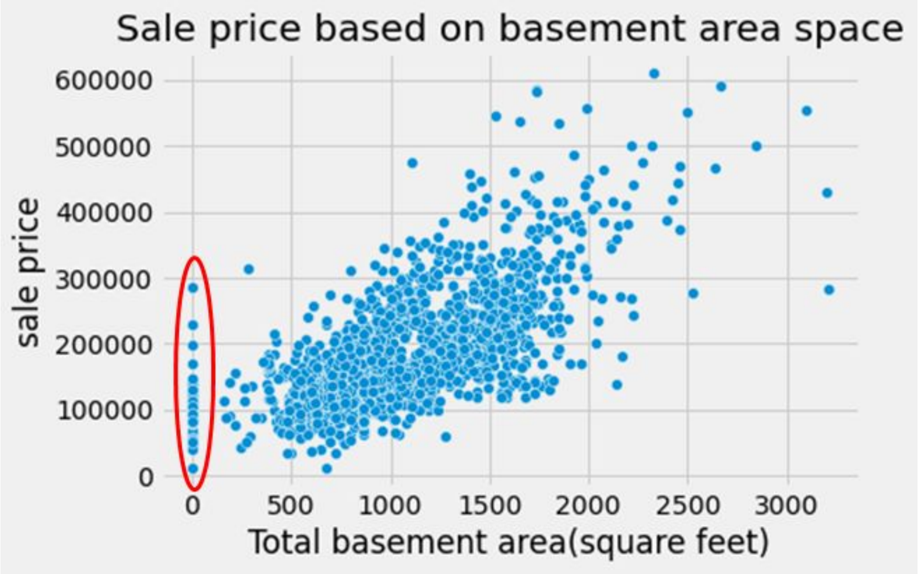
Quality variables



Quality variables

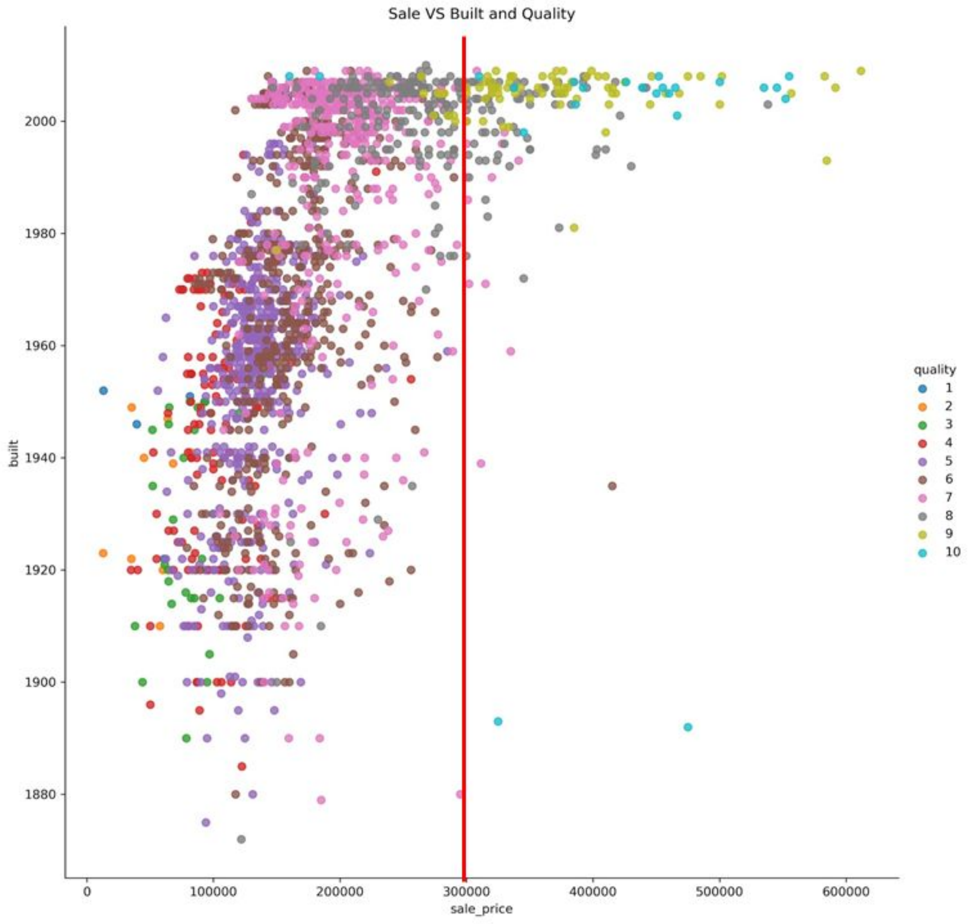


Space area variables



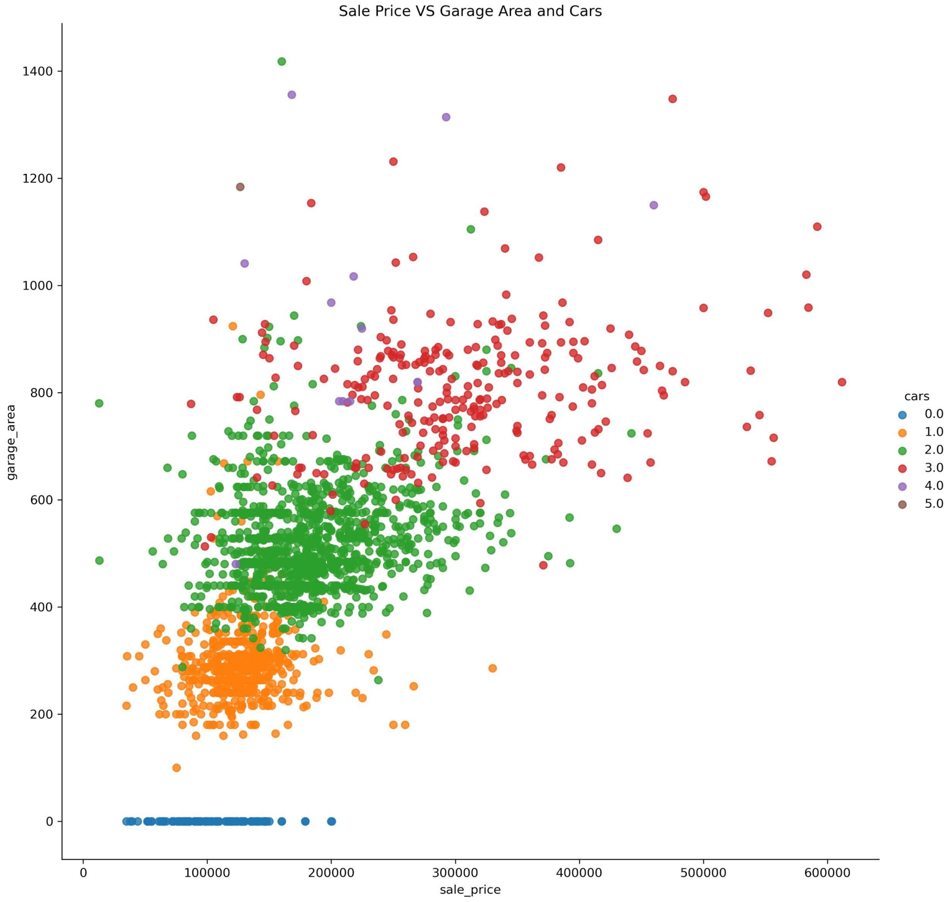
3 variables plot

- Sale price against years remodel and quality



3 variables plots

- Sale price against garage area and cars



Modeling : Train-Split-Test, OLS Summary

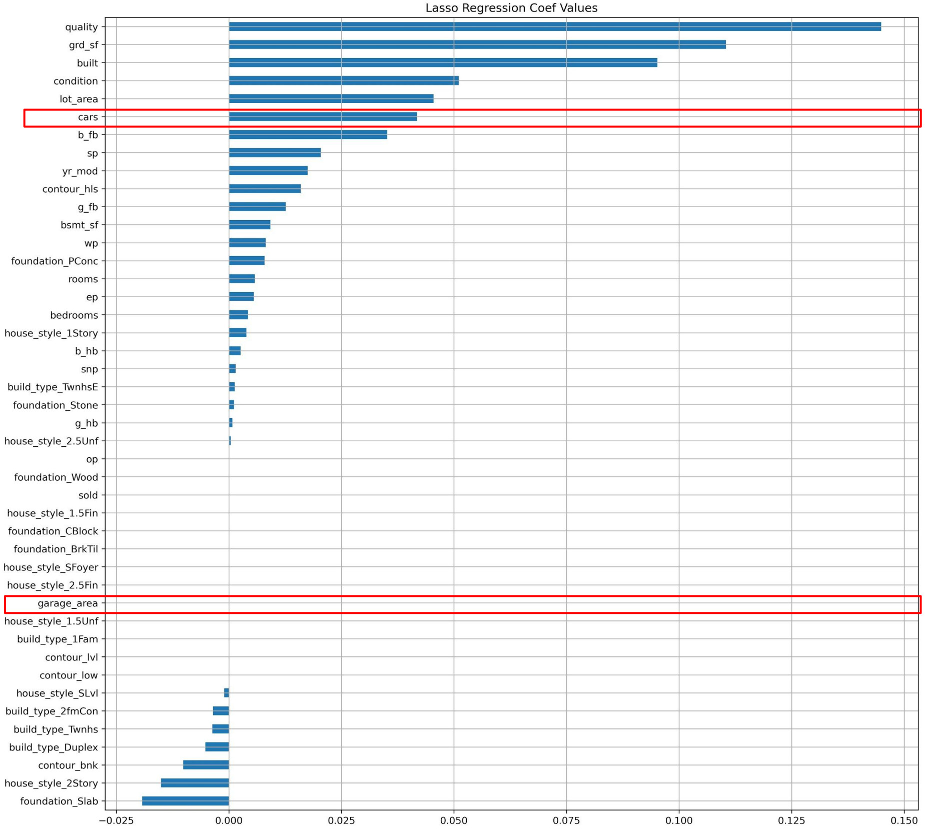
Dep. Variable:	sale_price	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	231.8
Date:	Thu, 08 Apr 2021	Prob (F-statistic):	0.00
Time:	15:16:27	Log-Likelihood:	695.93
No. Observations:	1434	AIC:	-1310.
Df Residuals:	1393	BIC:	-1094.
Df Model:	40		
Covariance Type:	nonrobust		

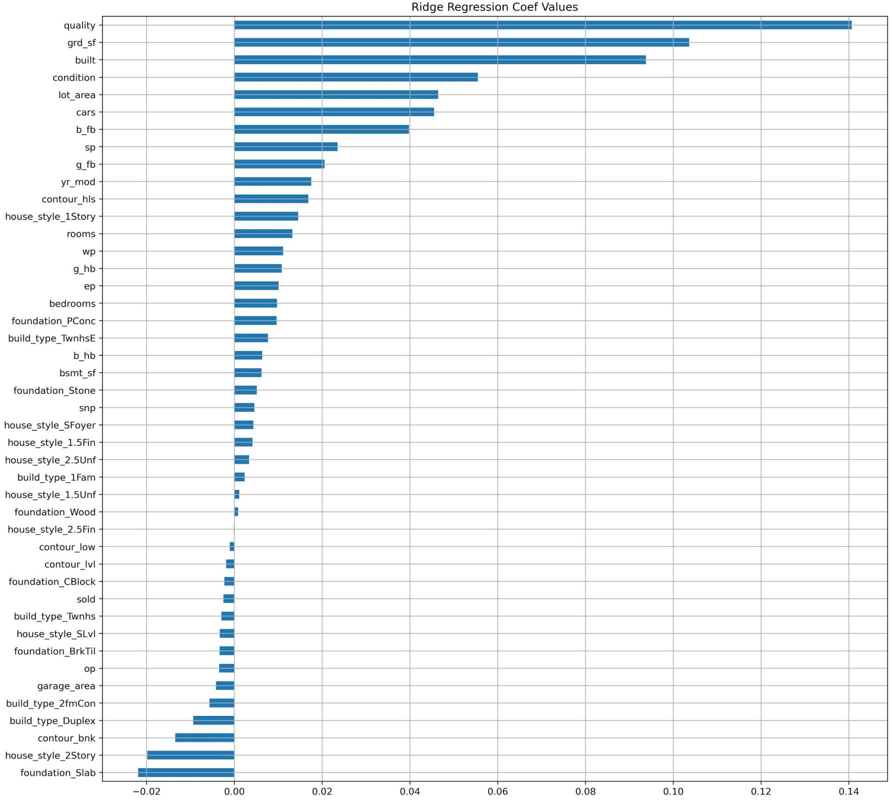
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3531	3.589	1.213	0.225	-2.686	11.393
lot_area	0.0988	0.014	7.214	0.000	0.072	0.126
quality	0.1008	0.005	20.049	0.000	0.091	0.111
condition	0.0501	0.005	10.609	0.000	0.041	0.059
bedrooms	0.0060	0.007	0.818	0.413	-0.008	0.020

- Provide us with a general view of the best line fit
- To measure the amount of error produced
- Can be used to determine features that are not significant to the sale prices.

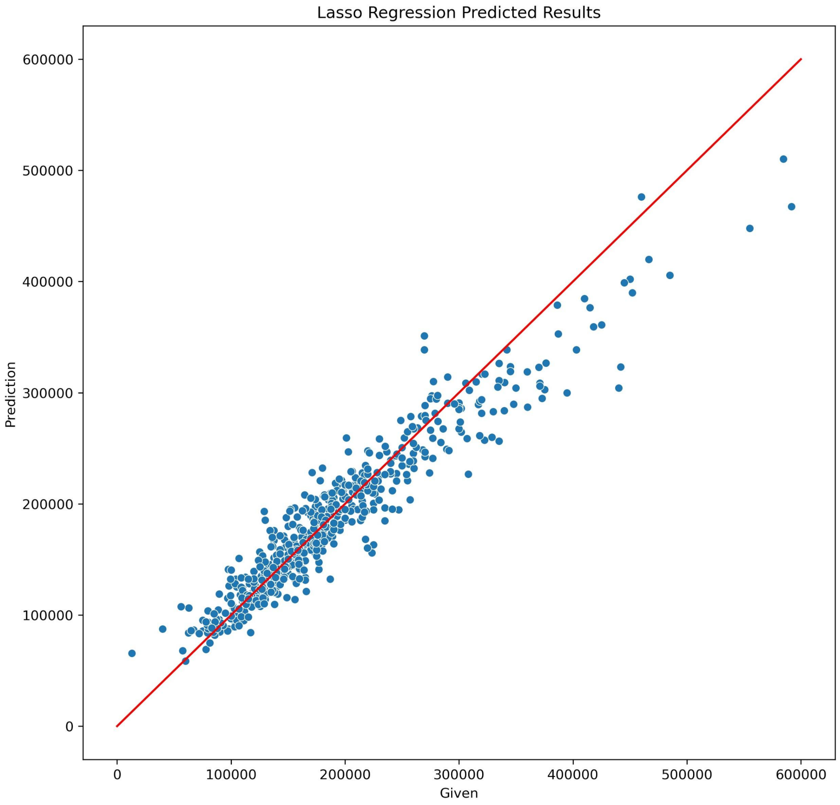
	coef	std err	t	P>|t|	[0.025	0.975]
cars	0.0633	0.013	4.746	0.000	0.037	0.090
garage_area	-3.723e-05	4.58e-05	-0.813	0.416	-0.000	5.26e-05

Modeling : Lasso Or Ridge Regression?





Predictions : How well can we estimate?



Score Measurements:

- RMSE : 25,379
- R^2 : 0.8821

Evaluation and Recommendation

- Our predictive model works well.
- Able to provide a reliable estimate base on actual data provided.
- Identify Features that are contributing to the sale price.
- Contribute to the company workflow. Agents able to quickly come up with a figure base on facts rather than guesses.