Ames Housing Project

by Casey Liu

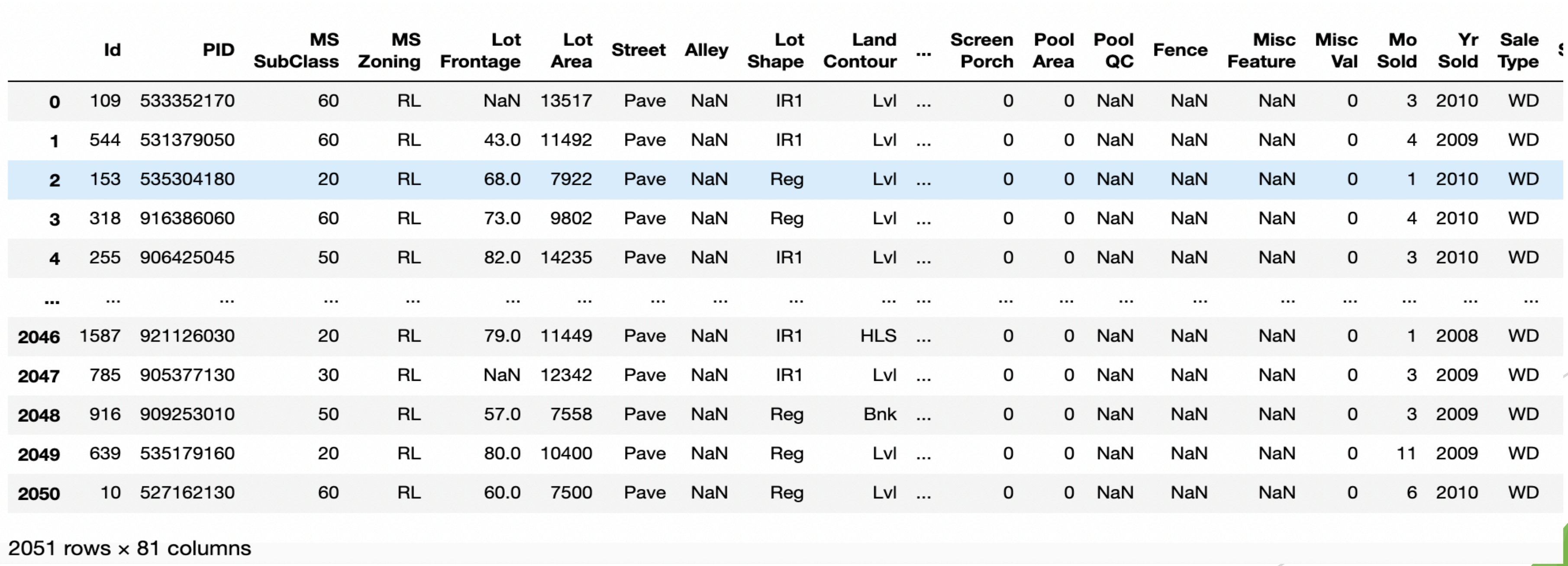
Problem statement

Ames, Iowa

- I have datasets of Ames housing. My goal is to use the datasets to predict the price of houses at sale and identify the important factors that have impact on to the value of houses. The Ames Housing Dataset is an exceptionally detailed and robust dataset with over 70 columns of different features relating to houses. I am going to process the data and create a regression model based on the Dataset.

Dataset

- Data set contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010
- 81 columns and 2051 rows of housing data
- Describe features of houses and the sale prices
- Include nominal, ordinal, discrete and continuous variables



	Id	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Screen Porch	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type
0	109	533352170	60	RL	NaN	13517	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	0	3	2010	WD
1	544	531379050	60	RL	43.0	11492	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	0	4	2009	WD
2	153	535304180	20	RL	68.0	7922	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	1	2010	WD
3	318	916386060	60	RL	73.0	9802	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	4	2010	WD
4	255	906425045	50	RL	82.0	14235	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	0	3	2010	WD
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2046	1587	921126030	20	RL	79.0	11449	Pave	NaN	IR1	HLS	...	0	0	NaN	NaN	NaN	0	1	2008	WD
2047	785	905377130	30	RL	NaN	12342	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	0	3	2009	WD
2048	916	909253010	50	RL	57.0	7558	Pave	NaN	Reg	Bnk	...	0	0	NaN	NaN	NaN	0	3	2009	WD
2049	639	535179160	20	RL	80.0	10400	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	11	2009	WD
2050	10	527162130	60	RL	60.0	7500	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	6	2010	WD

2051 rows × 81 columns

Data Cleaning

- Deal with null values (replace missing value as NA or 0)
- Transform objects to numbers
- Transform categorical variables to ordinal

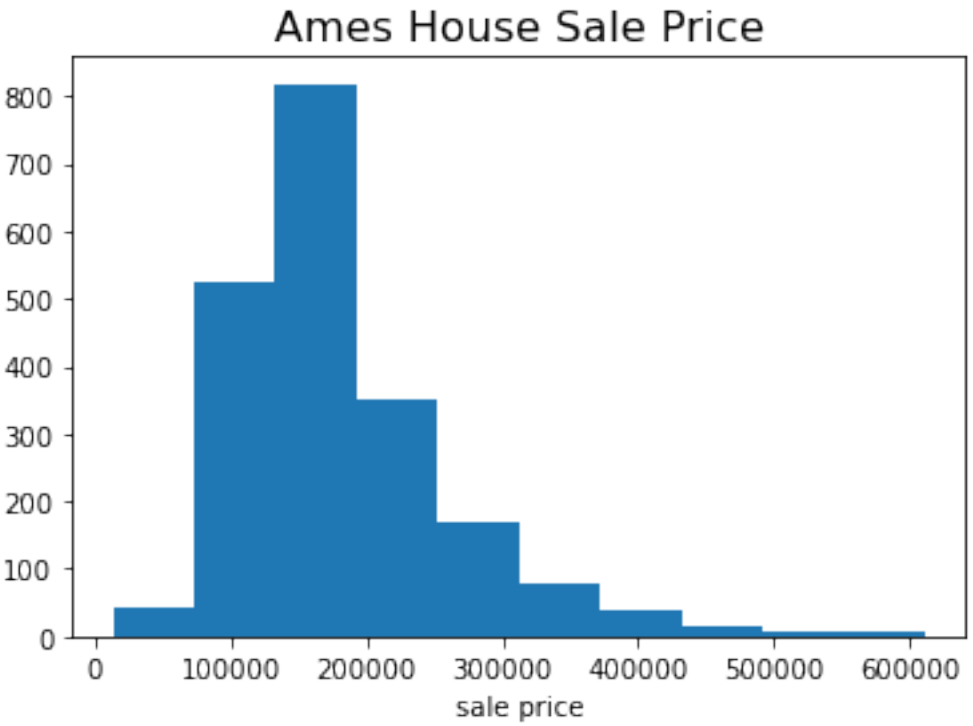
  For example : 'NA':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4, 'Ex':5
- Combine some columns

  For example: get “Total Square Feet”, “Total bathrooms”
- Identify if the house has certain feature

  For example: if the house has Central Air, Garage, Fence
- Create new column based on the date from existing columns

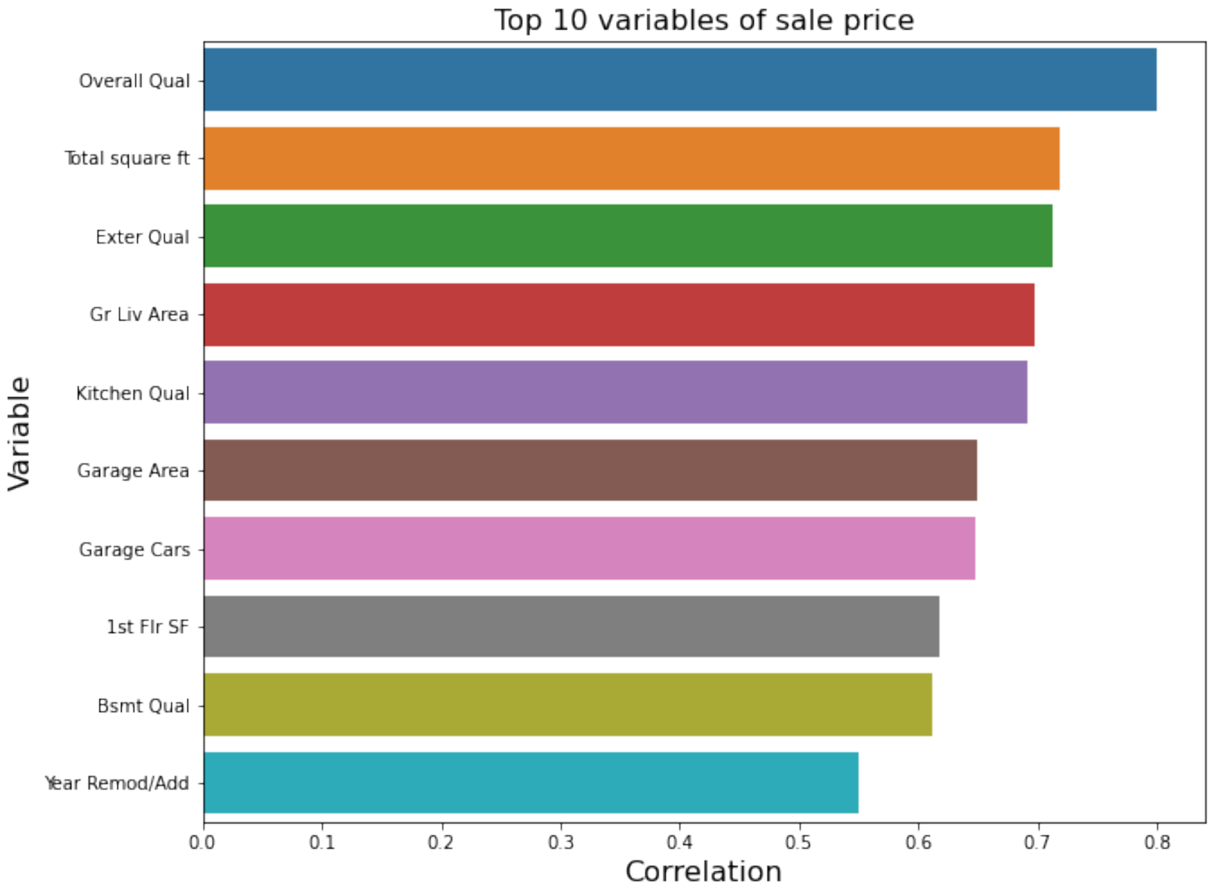
  For example: Calculate the age of the house based on “Year Built”

Exploratory Data Analysis



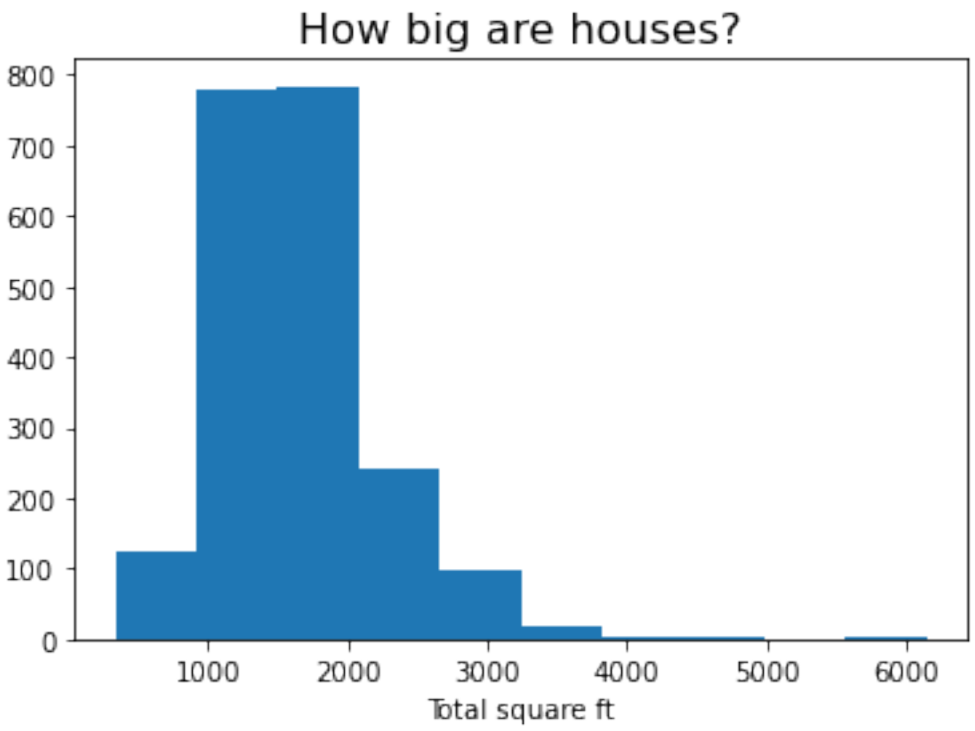
- count 2051.000000
- mean 181469.701609
- std 79258.659352
- min 12789.000000
- max 611657.000000
- Name: SalePrice, dtype: float64

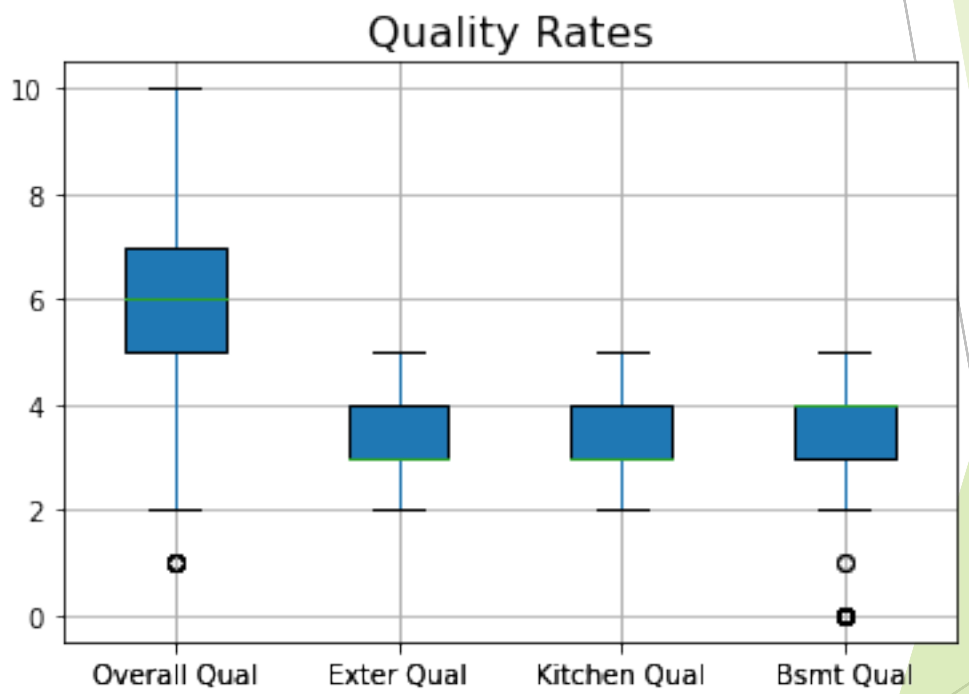
Exploratory Data Analysis



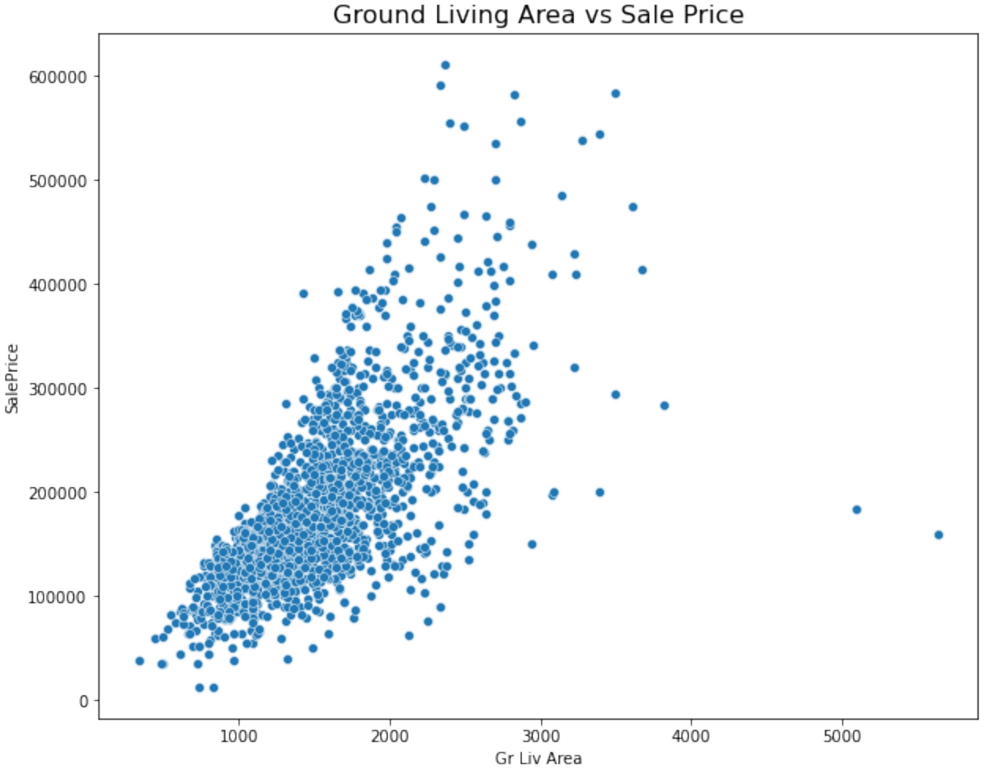
- Overall quality
- Total square feet
- Exterior quality
- Ground living area
- Kitchen quality
- Garage area
- Garage cars
- 1st floor square feet
- Basement quality
- Remodel year

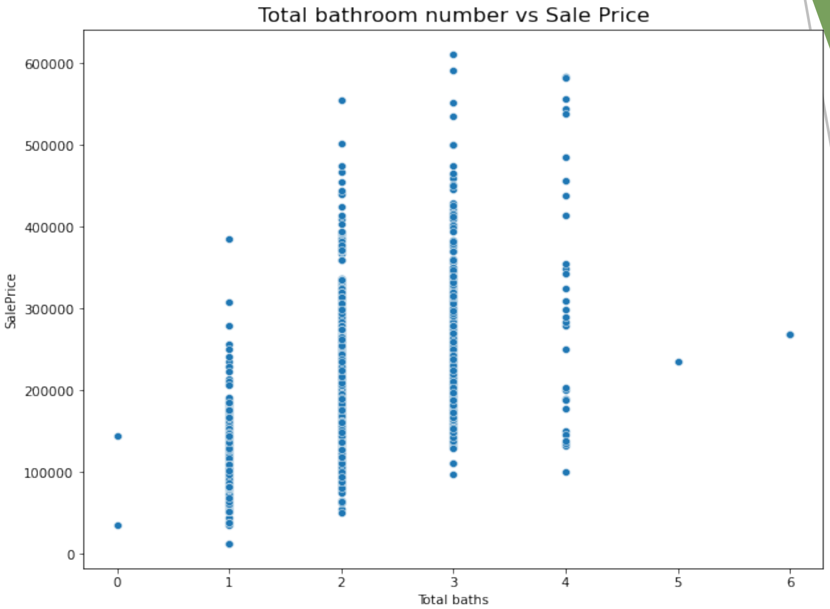
Exploratory Data Analysis

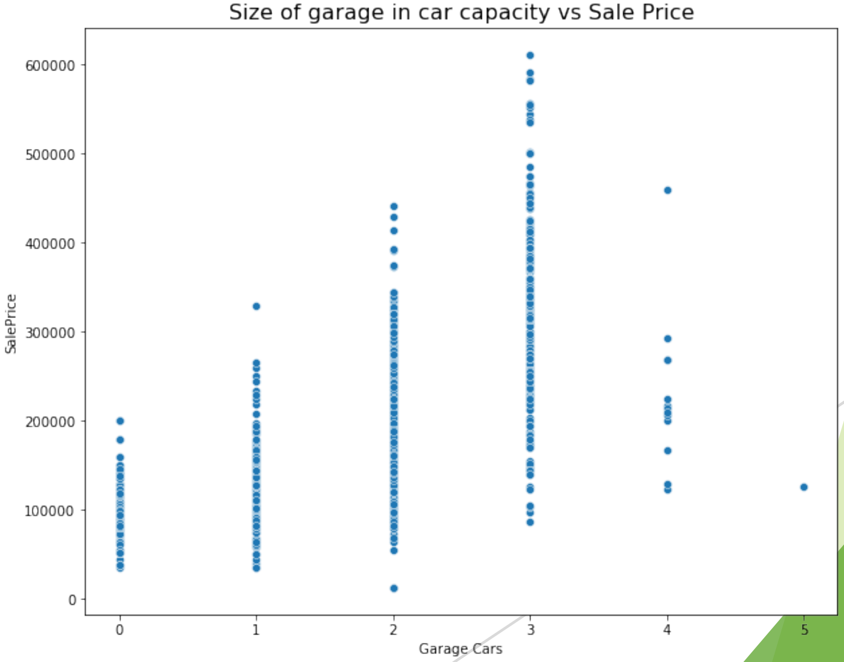




Exploratory Data Analysis

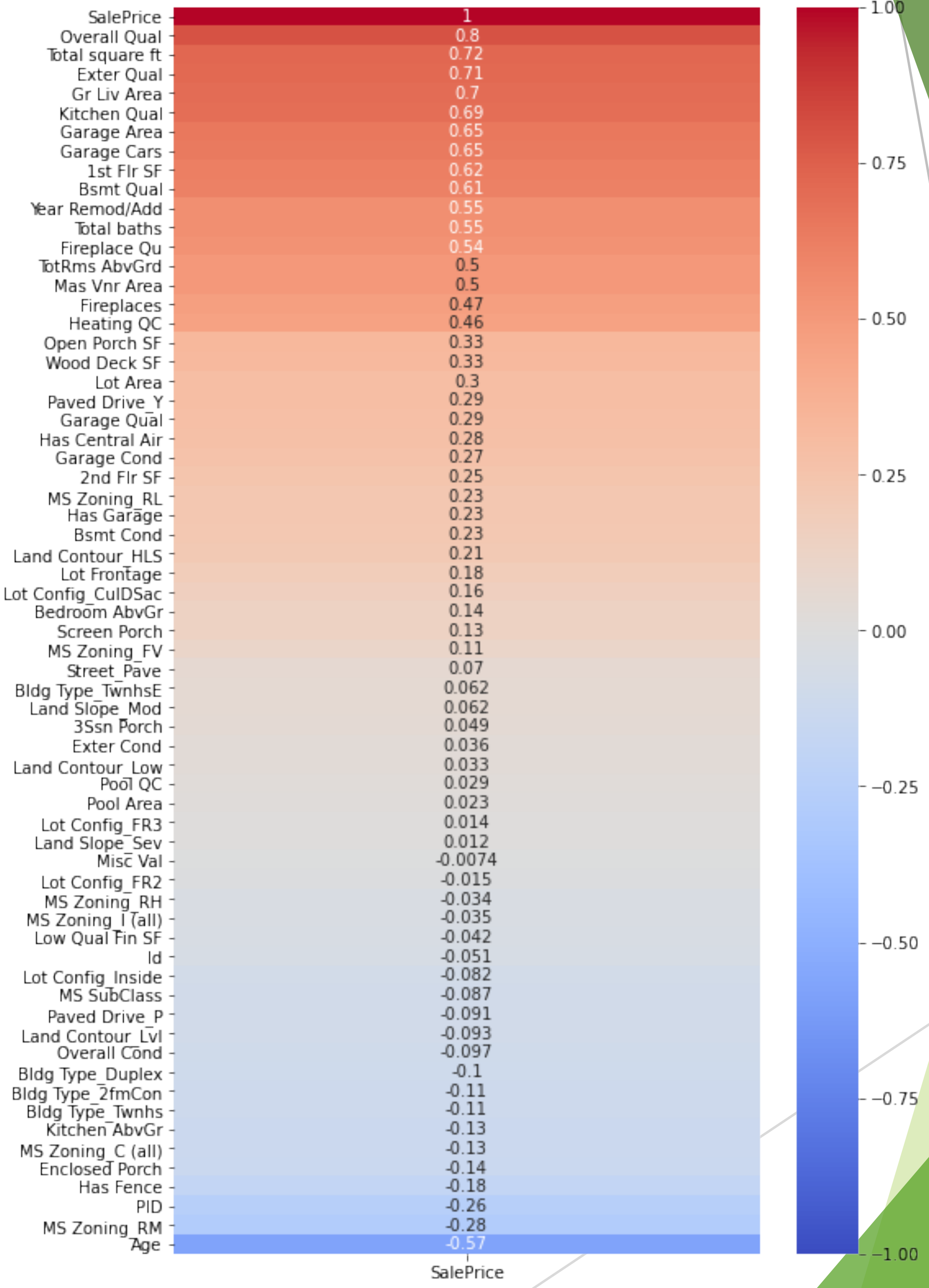






Variable Selection

- Select variables whose correlation with Sale Price is positive



Modeling

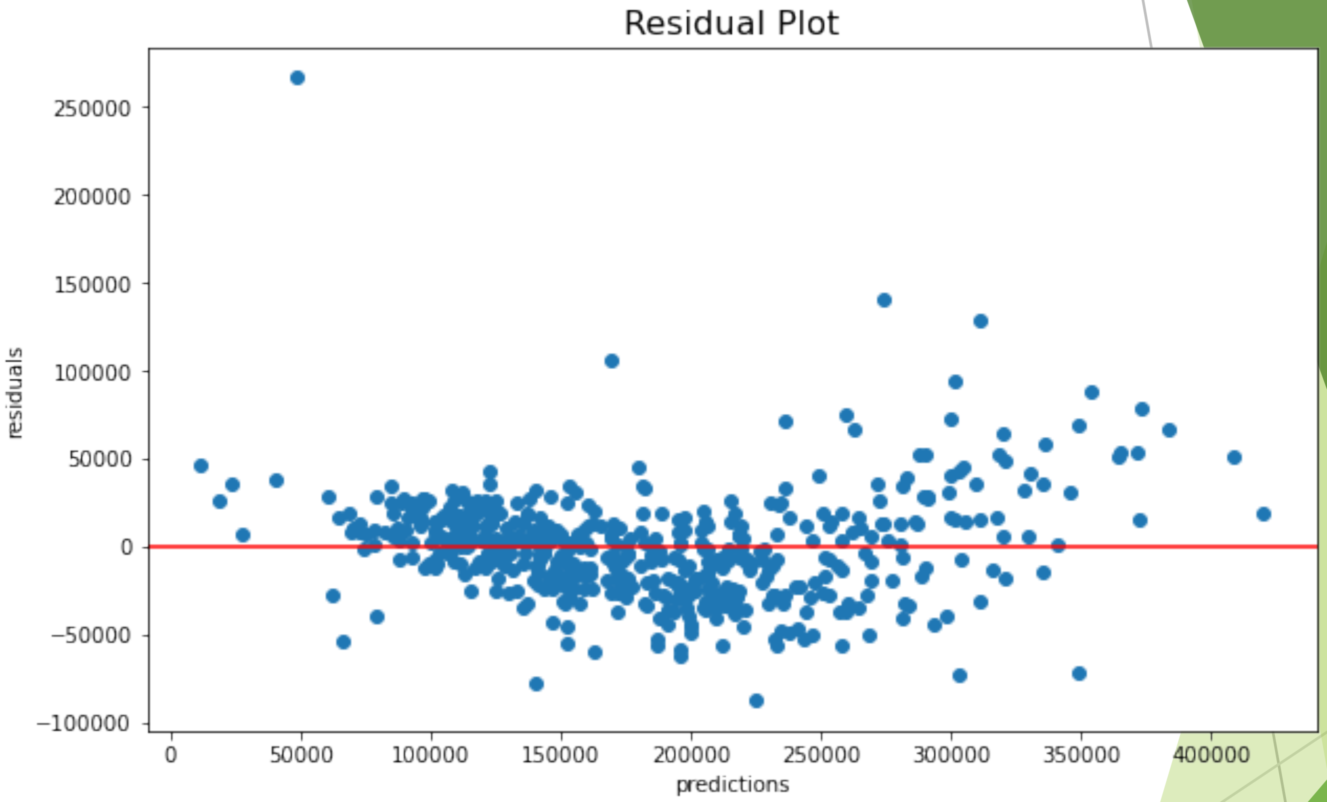
- Train-test split
- Scale variable
- Run cross validations
- Fit Linear Regression, RidgeCV and LassonCV

	train r2 score	test r2 score	cross_val_score
Linear Regression	0.8367	0.8439	0.7772
RidgeCV	0.8365	0.8492	0.7853
LassonCV	0.8306	0.8683	0.7837

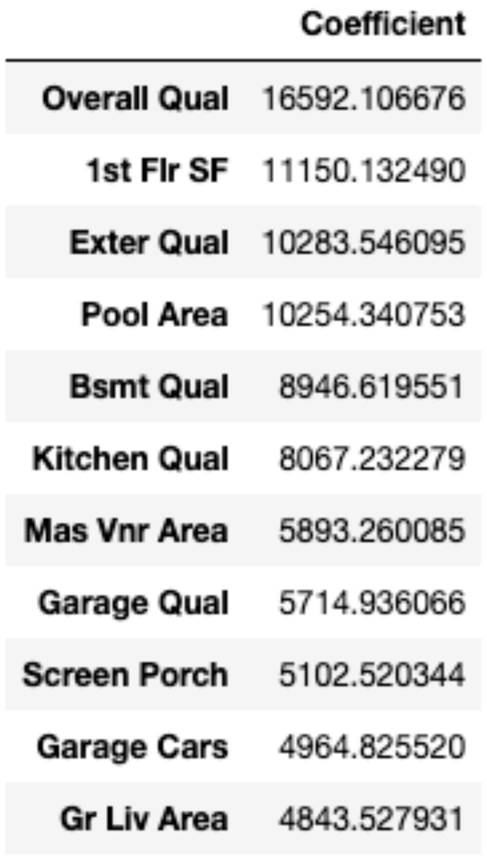
From the r2 scoring above, I find that RidgeCV has the highest score on training and cross validation while the difference between training and testing is lower than LASSO. I decide to run the Ridge model on the unseen data.

Residual Plot

- Residual plot shows the errors corresponding to the predicted values is randomly distributed. It looks normal, so I can go ahead and use the model.



Conclusions



	Coefficient
Overall Qual	16592.106676
1st Flr SF	11150.132490
Exter Qual	10283.546095
Pool Area	10254.340753
Bsmt Qual	8946.619551
Kitchen Qual	8067.232279
Mas Vnr Area	5893.260085
Garage Qual	5714.936066
Screen Porch	5102.520344
Garage Cars	4964.825520
Gr Liv Area	4843.527931

Based on the coefficient, the top 10 variables that can best predict Ames House Sale Price are:

- Overall quality
- First floor in square feet
- Exterior quality
- Pool area in square feet
- Basement quality
- Kitchen quality
- Masonry veneer area in square feet
- Garage quality
- Screen porch area
- Size of garage in car capacity

In conclusion, the quality of overall, exterior, basement, kitchen and garage is very important on a house value. The area size of first floor, pool, masonry veneer, screen porch and garage (in car capacity) would also impact the sale price. The real estate developers should pay attention on these factors to get higher house value.