Problem Statement:

To predict the sale price of residential homes in Ames, Iowa

Data Collection

- Source:
  - Ames, Iowa Assessor’s Office
- Period:
  - 2006 to 2010
- Size:
  - 2930 observations, 82 variables

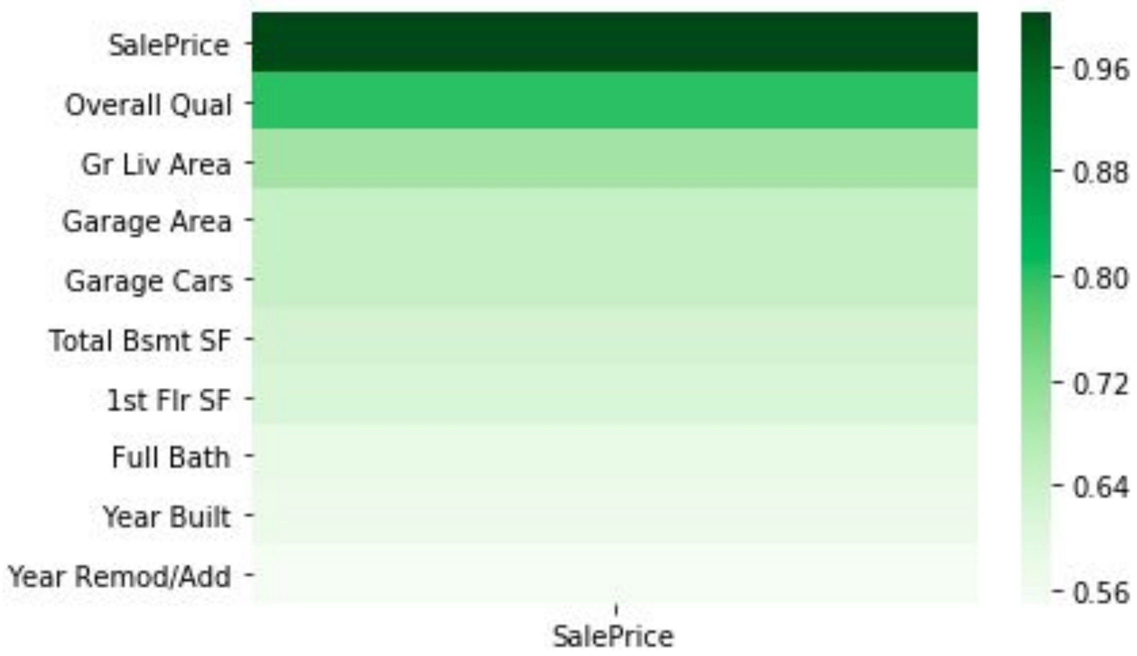
Data Cleaning

- Missing/Null values
    - Identify
        - Categorical
        - Numerical variables
    - Imputing
    - Combine features with interaction terms

DOES
THIS DATA
SPARK JOY?

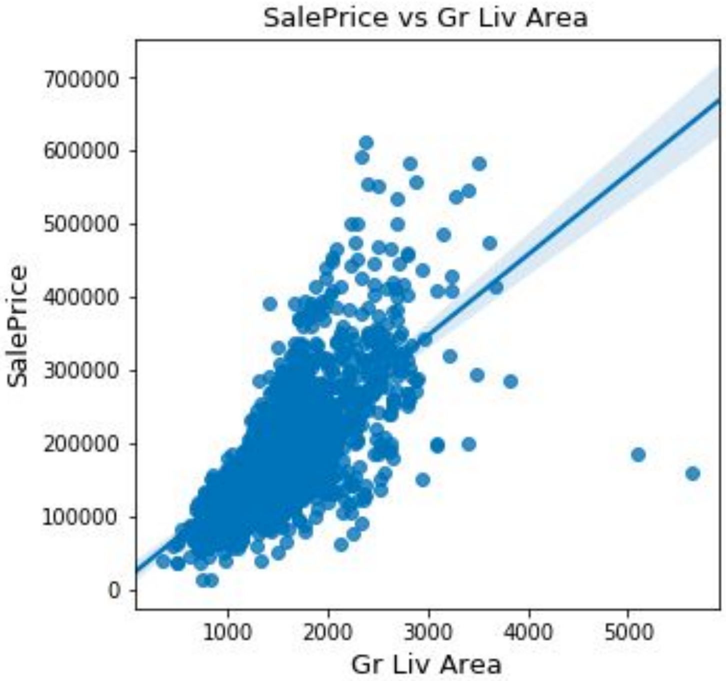
EDA

- Identify variables with highest correlations with Sale Price



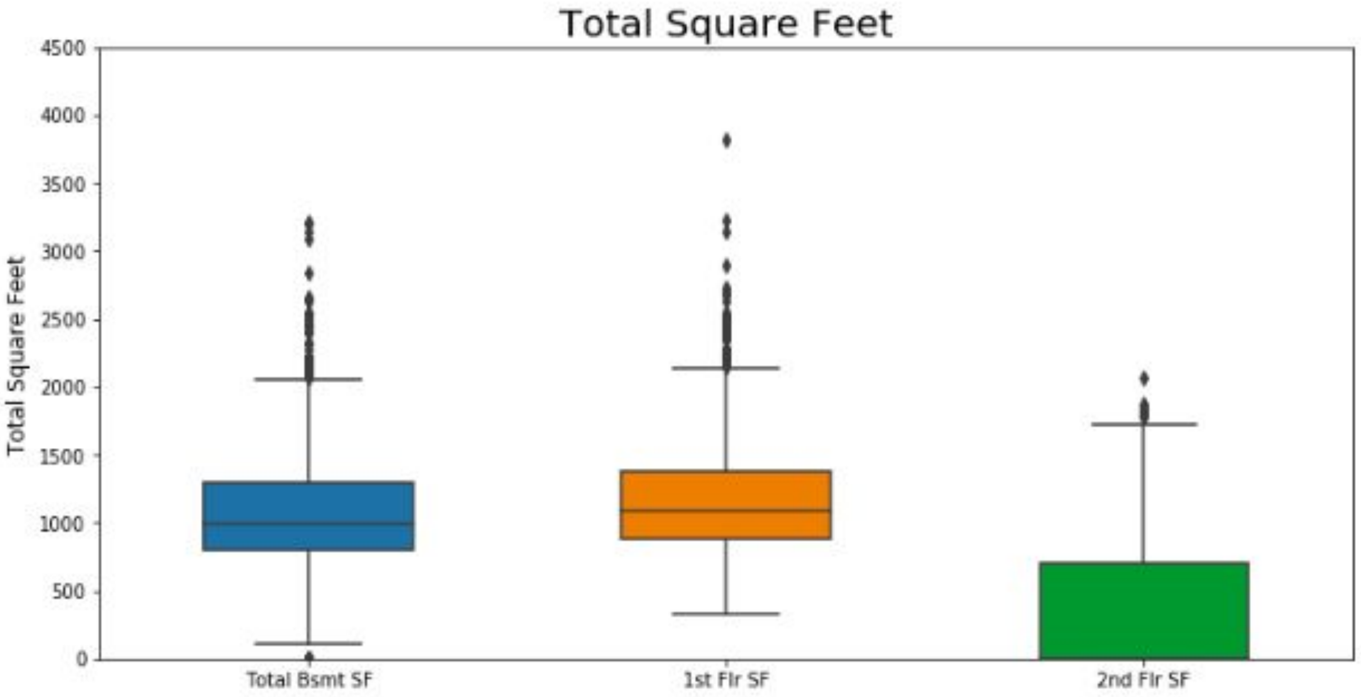
EDA

- Remove outliers

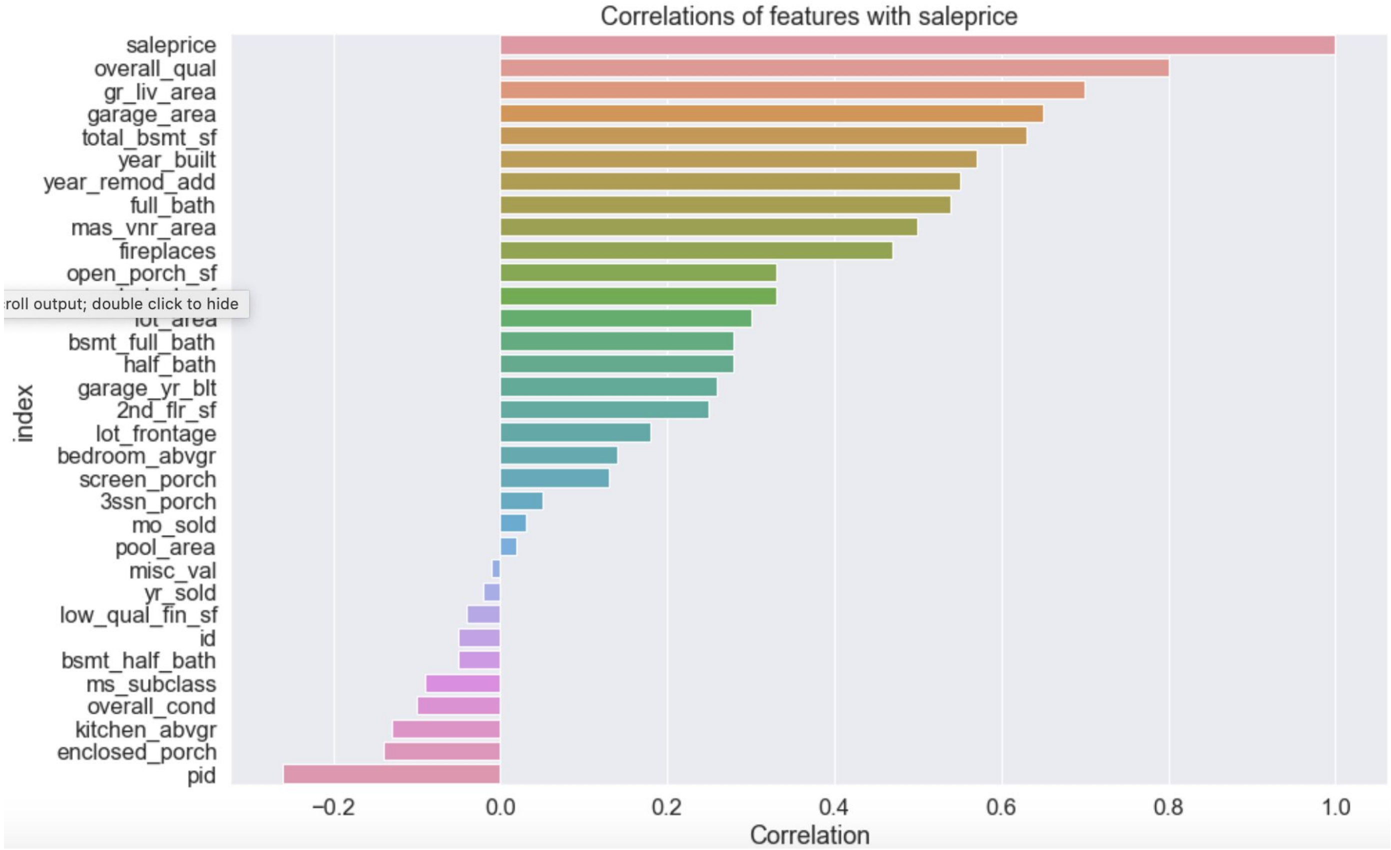


EDA

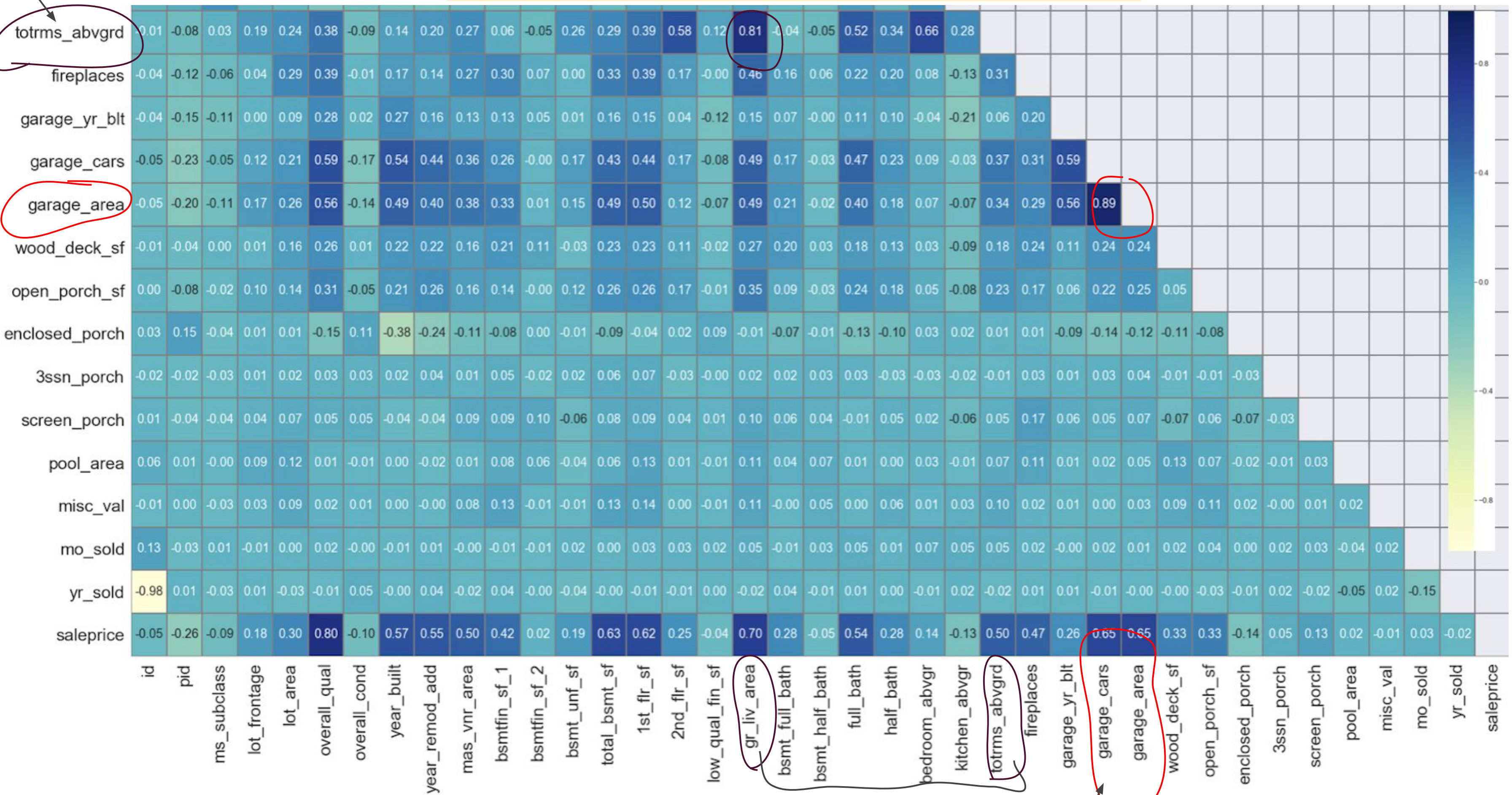
- Remove outliers - cont’



Correlation of features/predictors with target (saleprice)

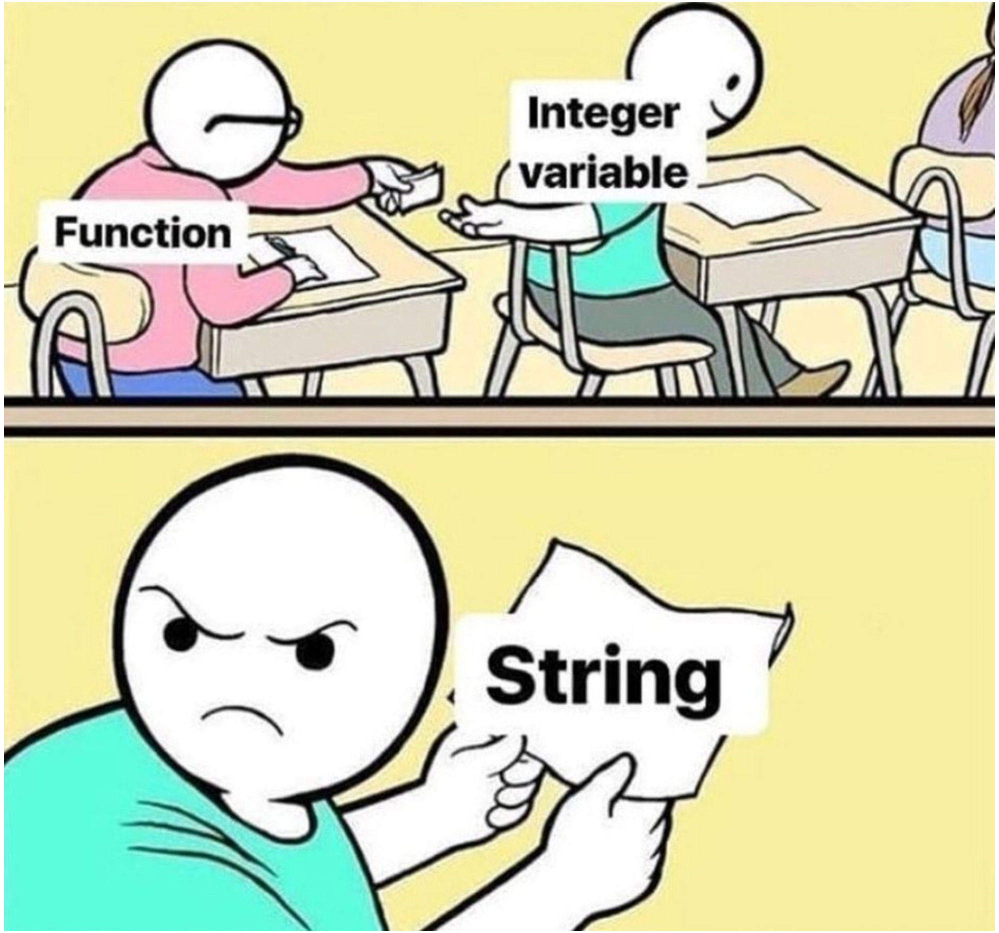


Heatmap & Feature Selection



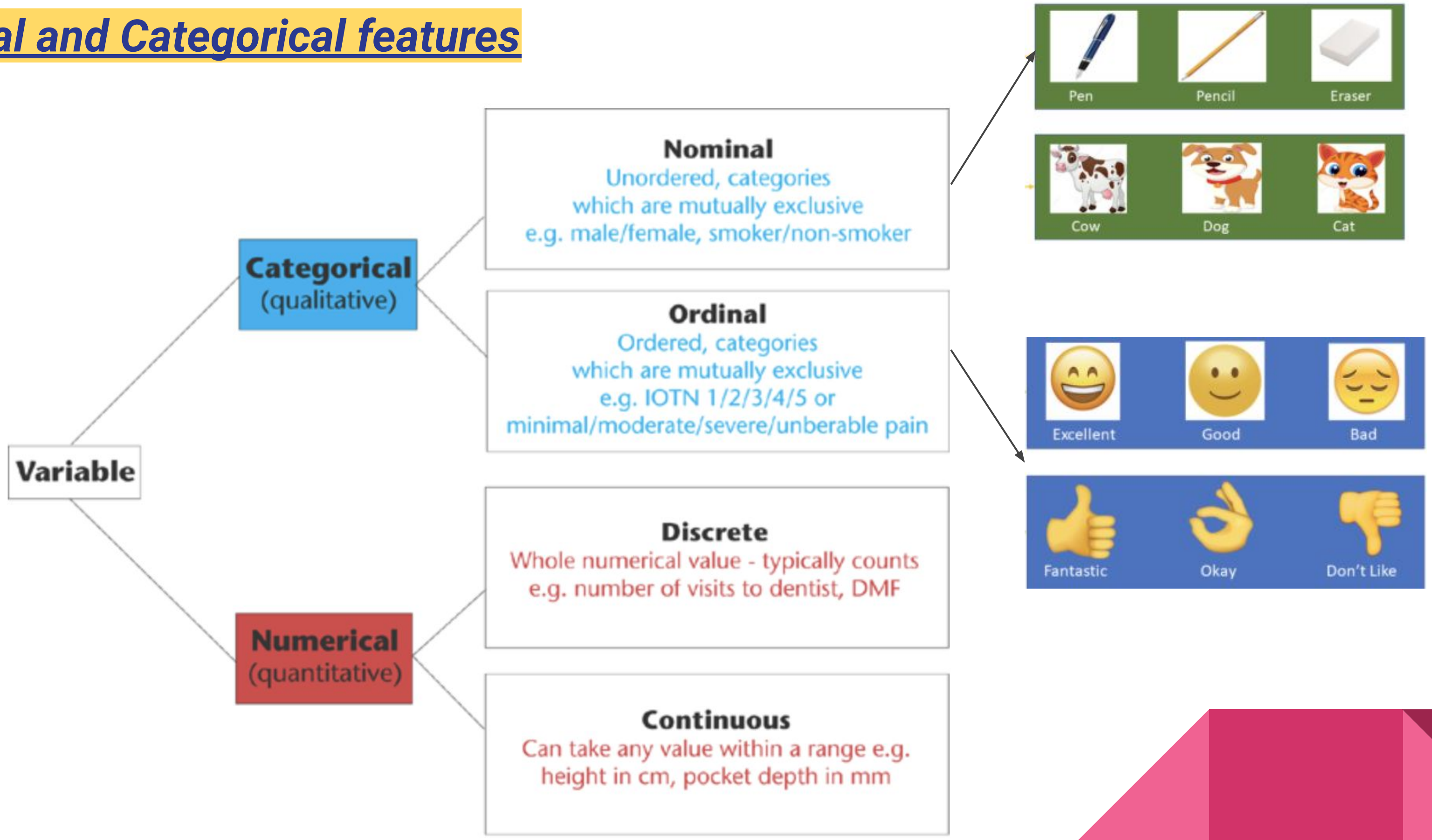
	id	pid	ms_subclass	lot_frontage	lot_area	overall_qual	overall_cond	year_built	year_remod_add	mas_vnr_area	bsmtfin_sf_1	bsmtfin_sf_2	bsmt_unf_sf	total_bsmt_sf	1st_flr_sf	2nd_flr_sf	low_qual_fin_sf	gr_liv_area	bsmt_full_bath	bsmt_half_bath	full_bath	half_bath	bedroom_abvgr	kitchen_abvgr	totrms_abvgrd	fireplaces	garage_yr_blt	garage_cars	garage_area	wood_deck_sf	open_porch_sf	enclosed_porch	3ssn_porch	screen_porch	pool_area	misc_val	mo_sold	yr_sold	saleprice
totrms_abvgrd	-0.01	-0.08	0.03	0.19	0.24	0.38	-0.09	0.14	0.20	0.27	0.06	-0.05	0.26	0.29	0.39	0.58	0.12	0.81	-0.04	-0.05	0.52	0.34	0.66	0.28															
fireplaces	-0.04	-0.12	-0.06	0.04	0.29	0.39	-0.01	0.17	0.14	0.27	0.30	0.07	0.00	0.33	0.39	0.17	-0.00	0.46	0.16	0.06	0.22	0.20	0.08	-0.13	0.31														
garage_yr_blt	-0.04	-0.15	-0.11	0.00	0.09	0.28	0.02	0.27	0.16	0.13	0.13	0.05	0.01	0.16	0.15	0.04	-0.12	0.15	0.07	-0.00	0.11	0.10	-0.04	-0.21	0.06	0.20													
garage_cars	-0.05	-0.23	-0.05	0.12	0.21	0.59	-0.17	0.54	0.44	0.36	0.26	-0.00	0.17	0.43	0.44	0.17	-0.08	0.49	0.17	-0.03	0.47	0.23	0.09	-0.03	0.37	0.31	0.59												
garage_area	-0.05	-0.20	-0.11	0.17	0.26	0.56	-0.14	0.49	0.40	0.38	0.33	0.01	0.15	0.49	0.50	0.12	-0.07	0.49	0.21	-0.02	0.40	0.18	0.07	-0.07	0.34	0.29	0.56	0.89											
wood_deck_sf	-0.01	-0.04	0.00	0.01	0.16	0.26	0.01	0.22	0.22	0.16	0.21	0.11	-0.03	0.23	0.23	0.11	-0.02	0.27	0.20	0.03	0.18	0.13	0.03	-0.09	0.18	0.24	0.11	0.24	0.24										
open_porch_sf	0.00	-0.08	-0.02	0.10	0.14	0.31	-0.05	0.21	0.26	0.16	0.14	-0.00	0.12	0.26	0.26	0.17	-0.01	0.35	0.09	-0.03	0.24	0.18	0.05	-0.08	0.23	0.17	0.06	0.22	0.25	0.05									
enclosed_porch	0.03	0.15	-0.04	0.01	0.01	-0.15	0.11	-0.38	-0.24	-0.11	-0.08	0.00	-0.01	-0.09	-0.04	0.02	0.09	-0.01	-0.07	-0.01	-0.13	-0.10	0.03	0.02	0.01	0.01	-0.09	-0.14	-0.12	-0.11	-0.08								
3ssn_porch	-0.02	-0.02	-0.03	0.01	0.02	0.03	0.03	0.02	0.04	0.01	0.05	-0.02	0.02	0.06	0.07	-0.03	-0.00	0.02	0.02	0.03	0.03	-0.03	-0.03	-0.02	-0.01	0.03	0.01	0.03	0.04	-0.01	-0.01	-0.03							
screen_porch	0.01	-0.04	-0.04	0.04	0.07	0.05	0.05	-0.04	-0.04	0.09	0.09	0.10	-0.06	0.08	0.09	0.04	0.01	0.10	0.06	0.04	-0.01	0.05	0.02	-0.06	0.05	0.17	0.06	0.05	0.07	-0.07	0.06	-0.07	-0.03						
pool_area	0.06	0.01	-0.00	0.09	0.12	0.01	-0.01	0.00	-0.02	0.01	0.08	0.06	-0.04	0.06	0.13	0.01	-0.01	0.11	0.04	0.07	0.01	0.00	0.03	-0.01	0.07	0.11	0.01	0.02	0.05	0.13	0.07	-0.02	-0.01	0.03					
misc_val	-0.01	0.00	-0.03	0.03	0.09	0.02	0.01	0.00	-0.00	0.08	0.13	-0.01	-0.01	0.13	0.14	0.00	-0.01	0.11	-0.00	0.05	0.00	0.06	0.01	0.03	0.10	0.02	0.01	0.00	0.03	0.09	0.11	0.02	-0.00	0.01	0.02				
mo_sold	0.13	-0.03	0.01	-0.01	0.00	0.02	-0.00	-0.01	0.01	-0.00	-0.01	-0.01	0.02	0.00	0.03	0.03	0.02	0.05	-0.01	0.03	0.05	0.01	0.07	0.05	0.05	0.02	-0.00	0.02	0.01	0.02	0.04	0.00	0.02	0.03	-0.04	0.02			
yr_sold	-0.98	0.01	-0.03	0.01	-0.03	-0.01	0.05	-0.00	0.04	-0.02	0.04	-0.00	-0.04	-0.00	-0.01	-0.01	0.00	-0.02	0.04	-0.01	0.01	0.00	-0.01	0.02	-0.02	0.01	0.01	-0.01	-0.00	-0.00	-0.03	-0.01	0.02	-0.02	-0.05	0.02	-0.15		
saleprice	-0.05	-0.26	-0.09	0.18	0.30	0.80	-0.10	0.57	0.55	0.50	0.42	0.02	0.19	0.63	0.62	0.25	-0.04	0.70	0.28	-0.05	0.54	0.28	0.14	-0.13	0.50	0.47	0.26	0.65	0.65	0.33	0.33	-0.14	0.05	0.13	0.02	-0.01	0.03	-0.02	

If computers
can
express????.....



To avoid this.......

Ordinal and Categorical features



Variable

Categorical (qualitative)

Nominal
Unordered, categories which are mutually exclusive
e.g. male/female, smoker/non-smoker

Ordinal
Ordered, categories which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or minimal/moderate/severe/unberable pain

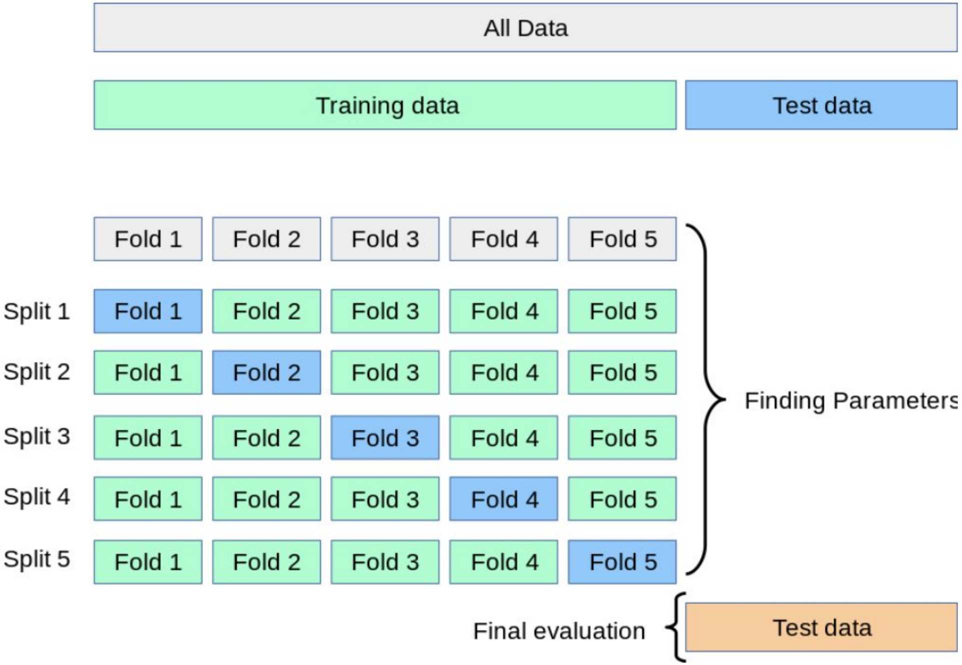
Numerical (quantitative)

Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF

Continuous
Can take any value within a range e.g. height in cm, pocket depth in mm

Pre-processing

Train-test-split



All Data

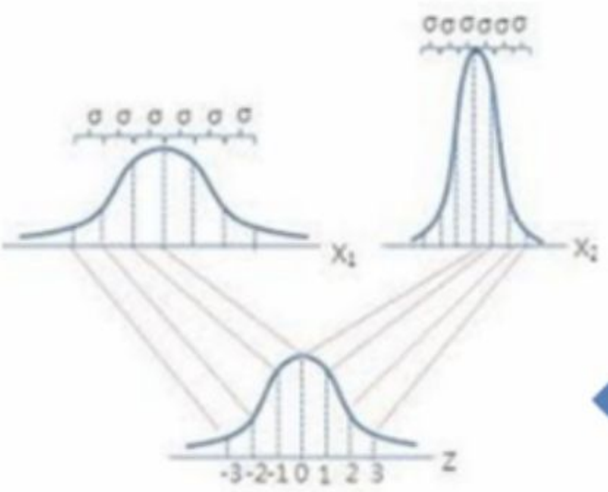
Training data	Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding Parameters

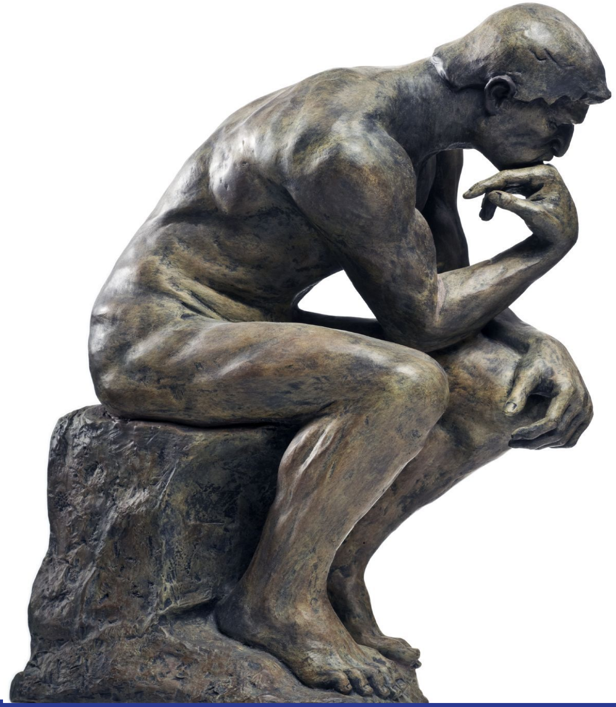
Final evaluation { Test data

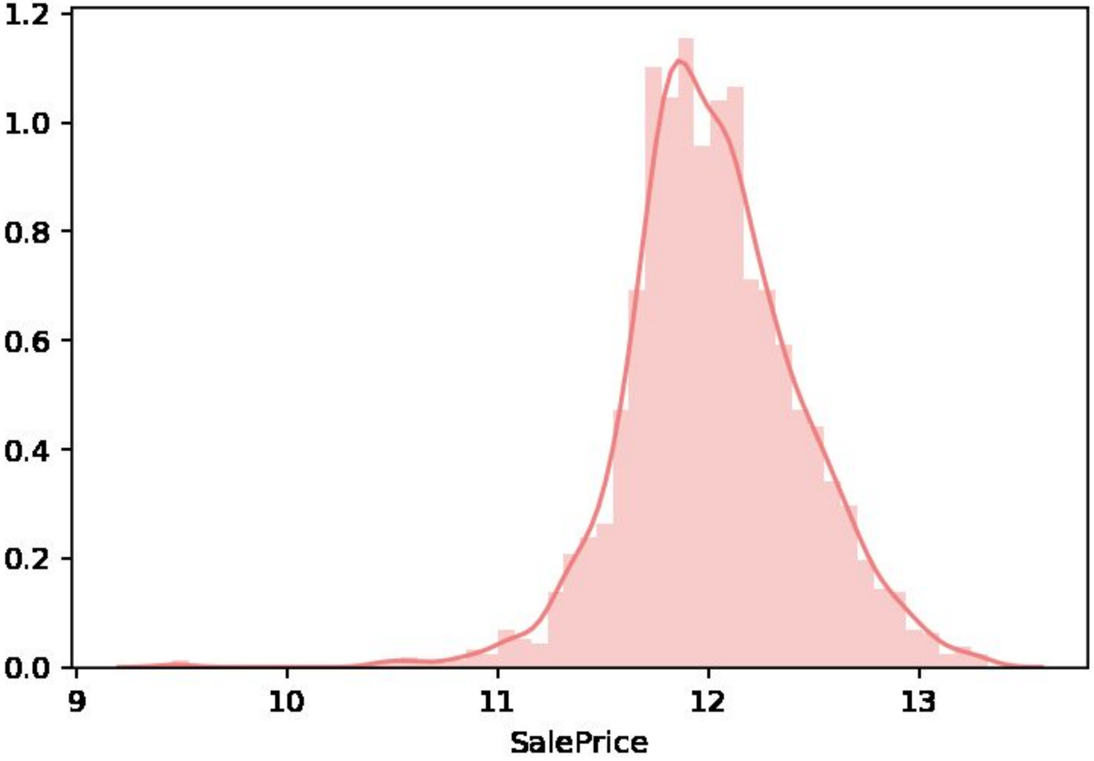
Scaling

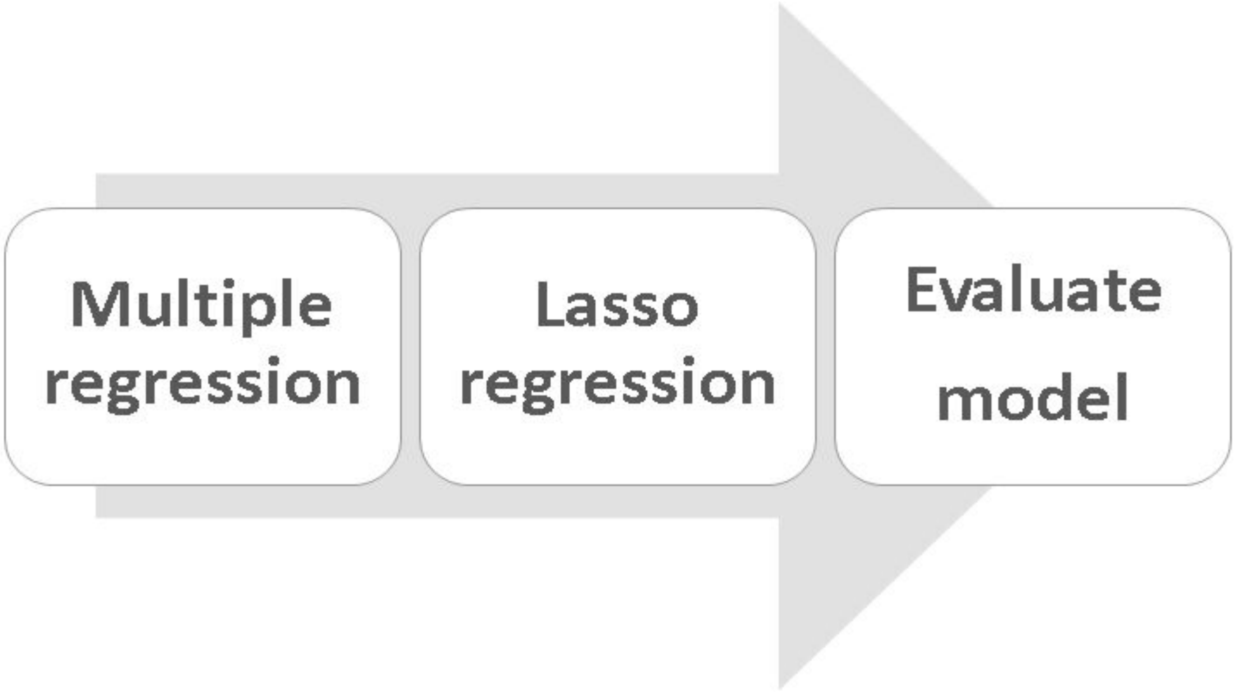


Modeling & Evaluation

What would be a naive guess of the sale price if we did not do any analysis?




1.2
1.0
0.8
0.6
0.4
0.2
0.0
9
10
11
12
13
SalePrice



Multiple regression

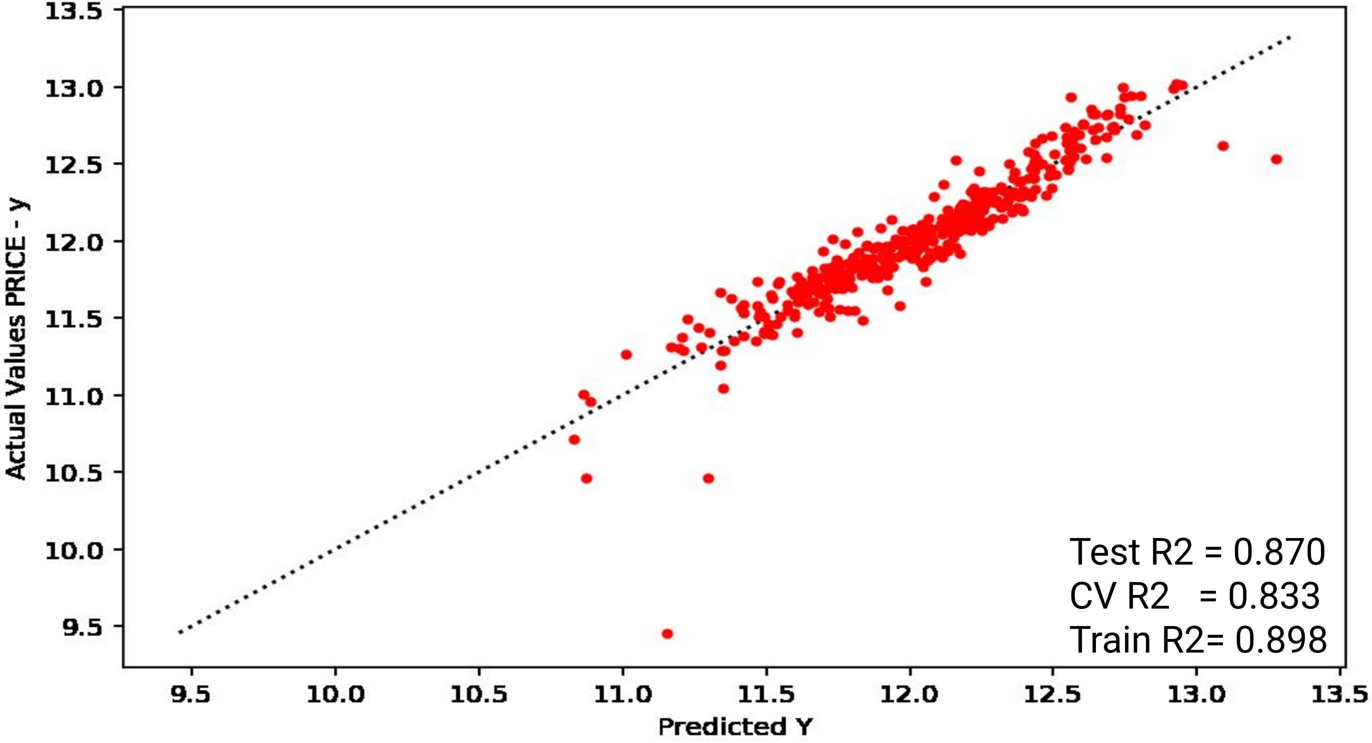
Lasso regression

Evaluate model

So how? Was it better or not?

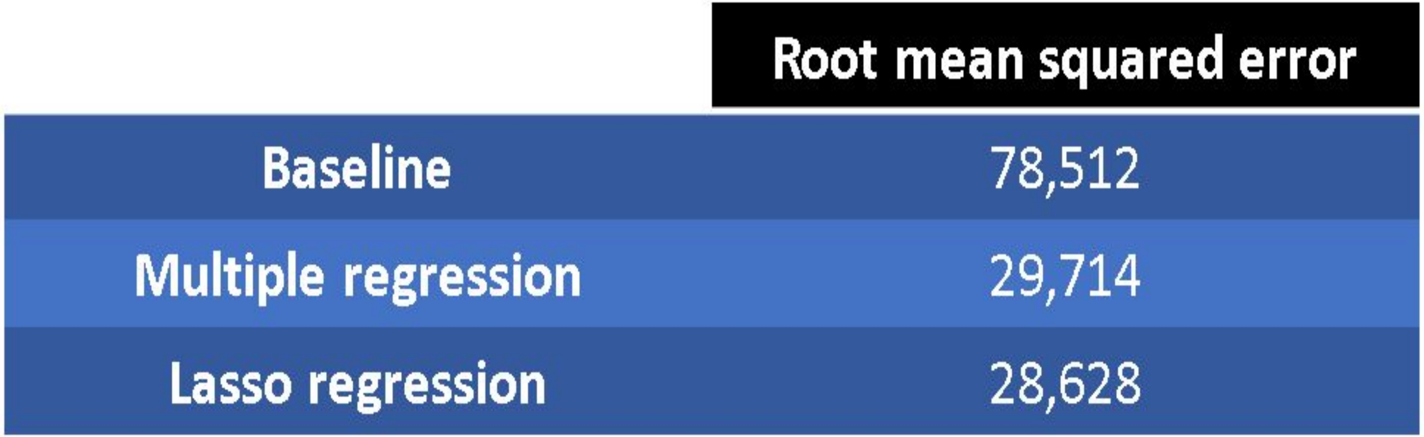
Model Accuracy 99.7%
Validation?
Training...

Train-test split comparison

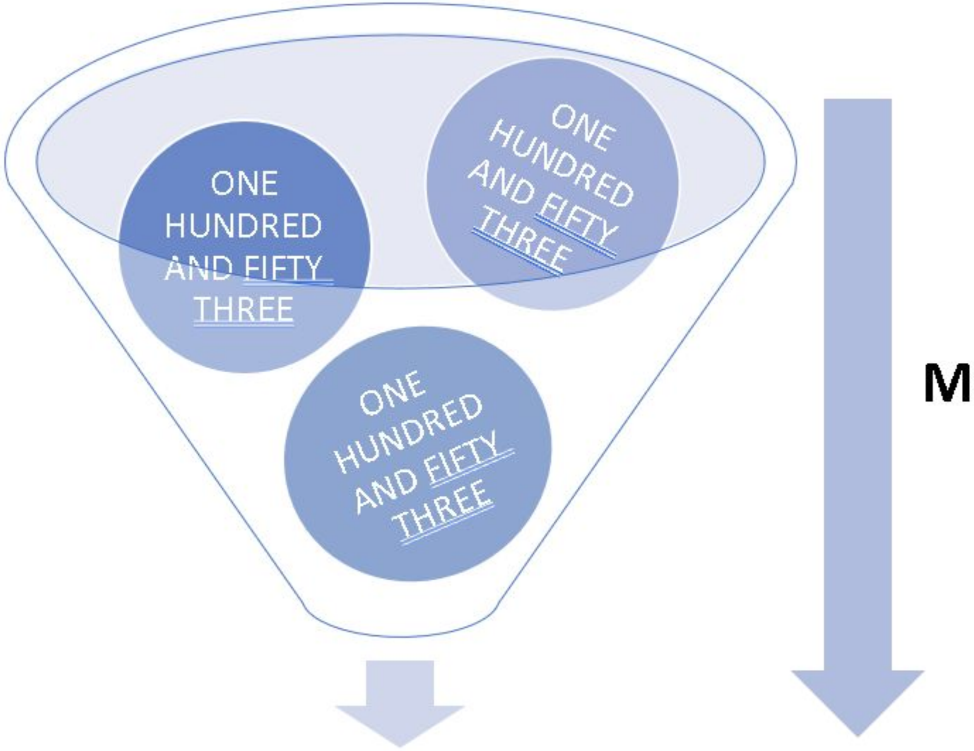


Test R2 = 0.870
CV R2   = 0.833
Train R2= 0.898

Better than a naive prediction



	Root mean squared error
Baseline	78,512
Multiple regression	29,714
Lasso regression	28,628



ONE HUNDRED AND FIFTY THREE

ONE HUNDRED AND FIFTY THREE

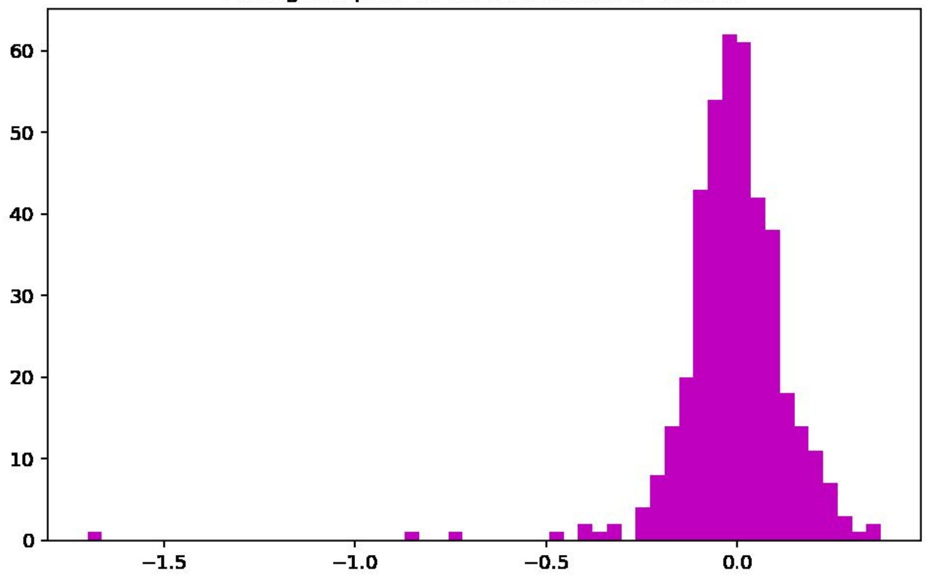
ONE HUNDRED AND FIFTY THREE

MORE THAN
5x

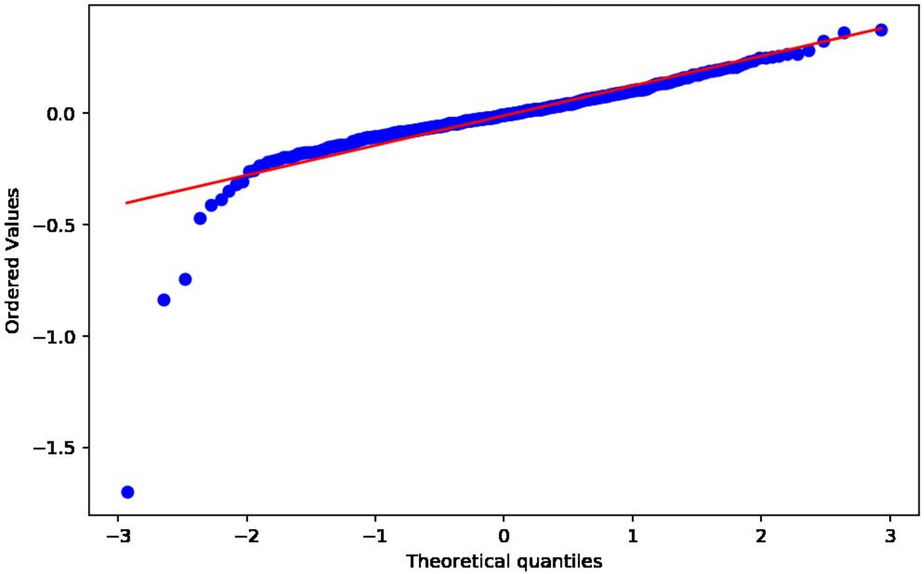
25

Assumptions of regression

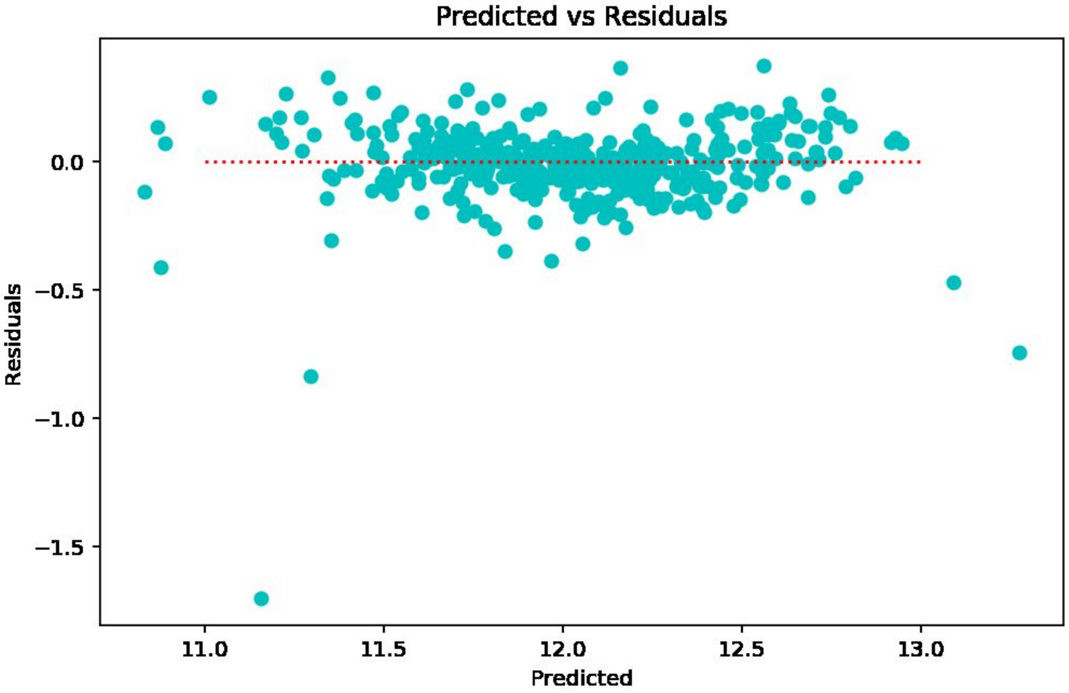
1. Predictors must have an approximately linear relationship with the dependent variable
2. Errors are normally distributed
3. Independence of error
4. Constant error variance
5. Independence of Predictors

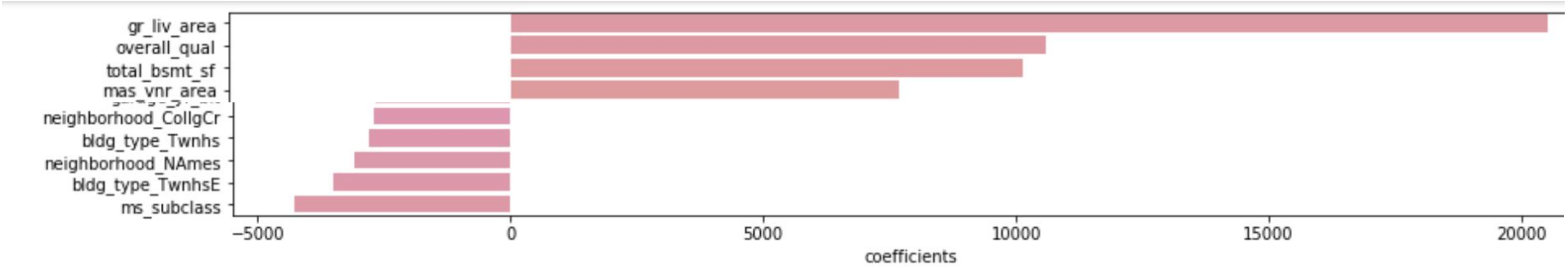


Histogram plots of unstandardized residuals



QQ plots of unstandardized residuals


Predicted vs Residuals
0.0
−0.5
−1.0
−1.5
Residuals
11.0
11.5
12.0
12.5
13.0
Predicted


gr_liv_area
overall_qual
total_bsmt_sf
mas_vnr_area
neighborhood_CollgCr
bldg_type_Twnhs
neighborhood_NAmes
bldg_type_TwnhsE
ms_subclass
−5000
0
5000
10000
15000
20000
coefficients

- Which features appear to add the most value to a home?

Total Living Space (TOTAL BSMT SF + GR LIV AREA) appears to add the most value to a home;

the more space, rooms and land that the house contains, the higher the valuation

Which features hurt the value of a home the most?

By the same token, the factors that add the most value to a home also hurt the value of a home the most.

A smaller Total Living Space and poor Overall quality of the property is detrimental to home value.

- What are things that homeowners could improve in their homes to increase the value?

Homeowners can improve the value of their homes by renovation using a dual approach to both increase the Total Living Space and Overall quality of the property.

Overall Qual is the best determinant providing the strongest correlations with SalePrice.

Gr Liv Area is the second best determinant providing a strong correlations with SalePrice.

Total Bsmt SF is the third best determinant providing a strong correlations with SalePrice.

Limitations

- High number of outliers means
- Not all independent variable are normally distributed
- Specific to Ames, Iowa
- 5 years of data from 2006 - 2010

Conclusion

This project addressed the non-fungibility of real estate property by building a model and make an accurate prediction for property sale price(target variable).

The prospective buyer(s), seller(s) and stakeholder(s) have the need to appraisal a fair market value of a property in order to facilitate transaction, taxation, appraisal, etc.