Ames Iowa Housing Analysis

Group 1

Don, Grace, Jocelyn, Junyuan, Randy

TABLE OF CONTENTS

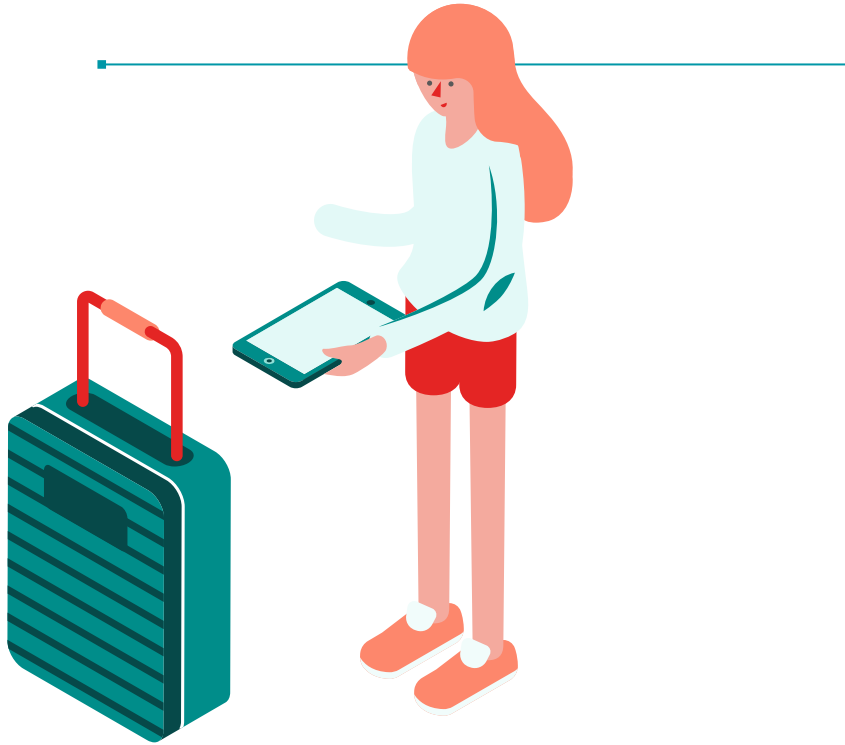
INTRODUCTION

In this Kaggle challenge,we are to use the well known Ames housing data to create a regression model that predicts the price of houses in Ames.

PROBLEM STATEMENT

Identify

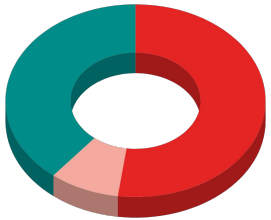
1. the best model to achieve the lowest RMSE scores and
2. the top 3 predictors of housing prices

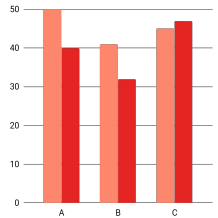


03

Exploratory Data Analysis

The data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers)

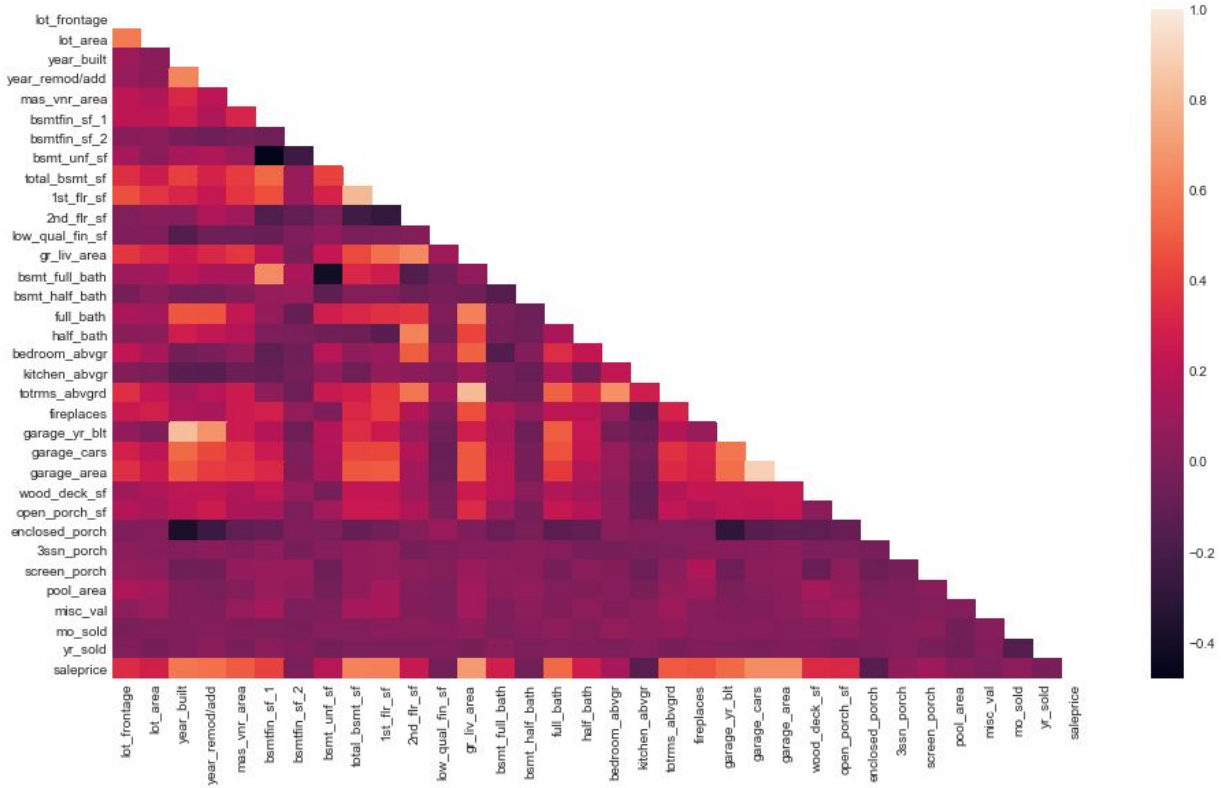




Mean: 181,469

Point at 5 standard deviations away

Heatmap of Numerical variables



Some variables are collinear

A number of variables have moderate and strong correlation to sale price

Collinear Variables

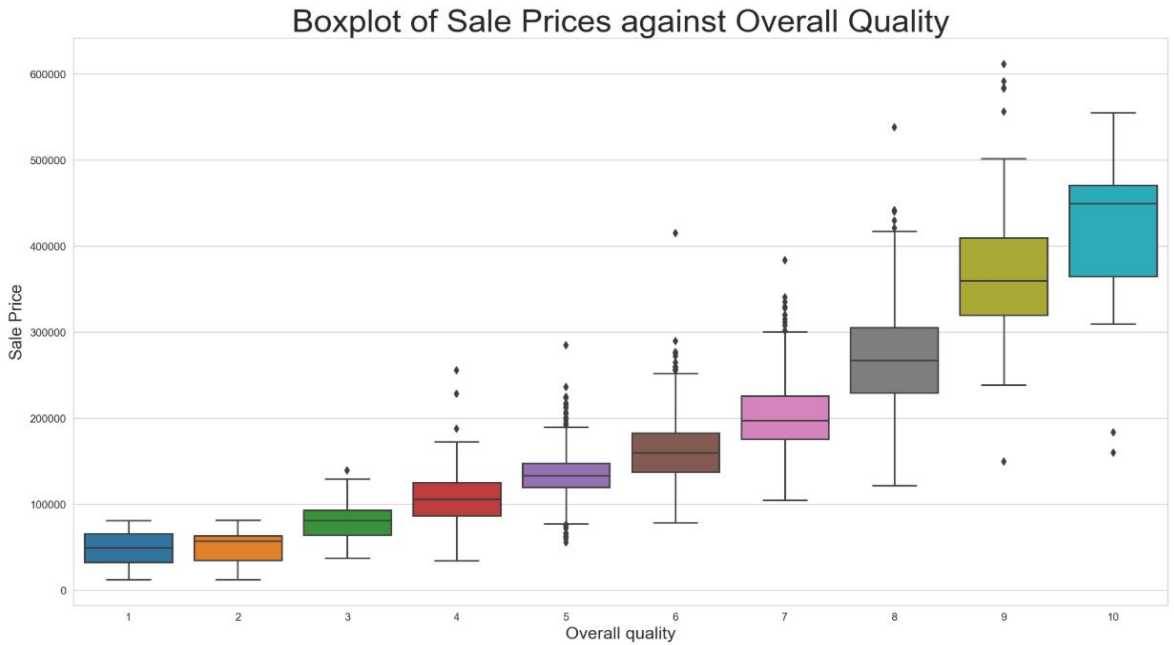
garage_cars	garage_area	0.892951
year_built	garage_yr_blt	0.824603
gr_liv_area	totrms_abvgrd	0.810453
total_bsmt_sf	1st_flr_sf	0.804157
year_remod/add	garage_yr_blt	0.672759
bedroom_abvgr	totrms_abvgrd	0.660571
bsmtfin_sf_1	bsmt_full_bath	0.646153
2nd_flr_sf	gr_liv_area	0.641396
year_built	year_remod/add	0.627660
2nd_flr_sf	half_bath	0.615997
gr_liv_area	full_bath	0.612622

Selection process to identify representative variable, e.g

- garage_area over garage_cars due to better granularity

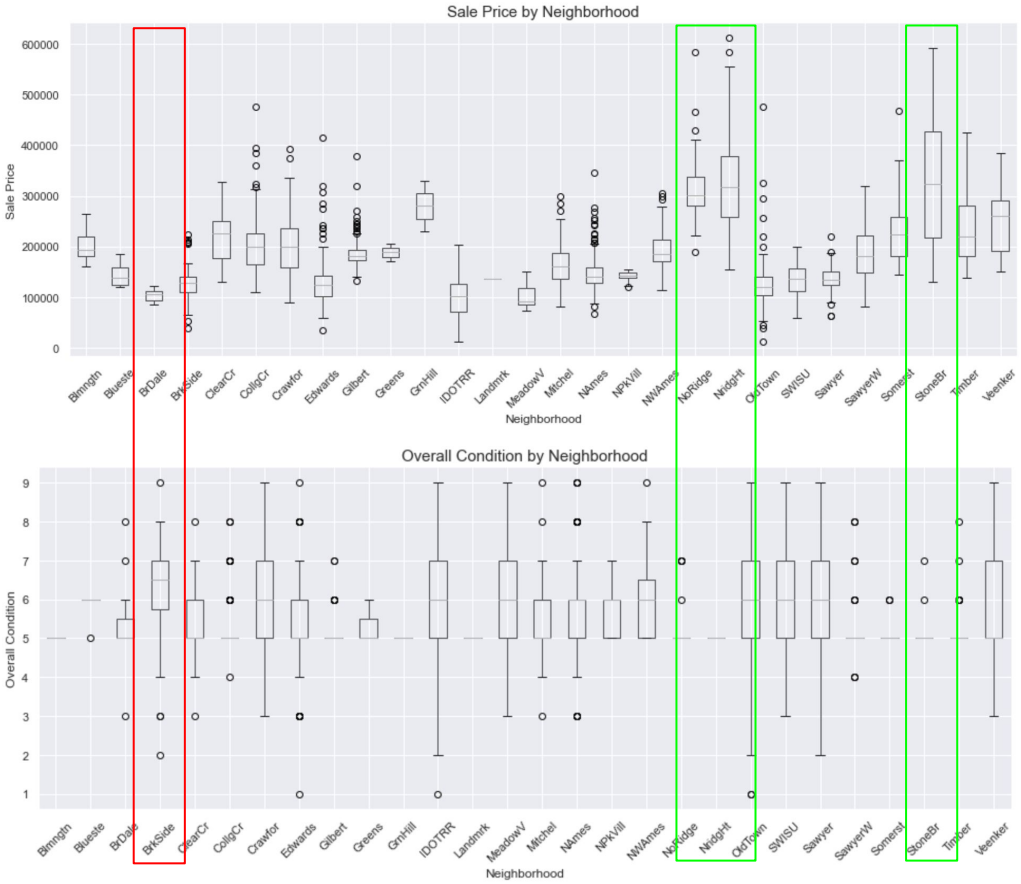
- Year_built over garage_yr_blt as logic would suggest garages are built at the same time as when the property is built

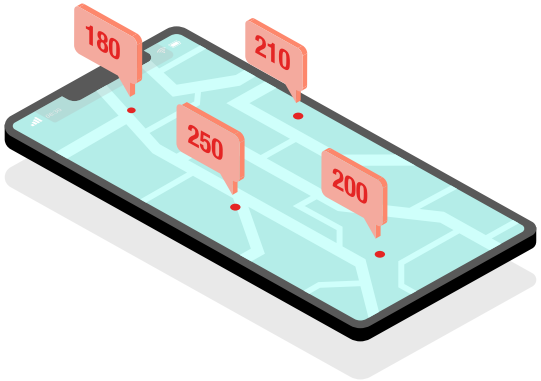
SALE PRICE AGAINST OVERALL QUALITY



A house with a better overall quality finish always results in higher prices

SALE PRICE AGAINST NEIGHBOURHOOD

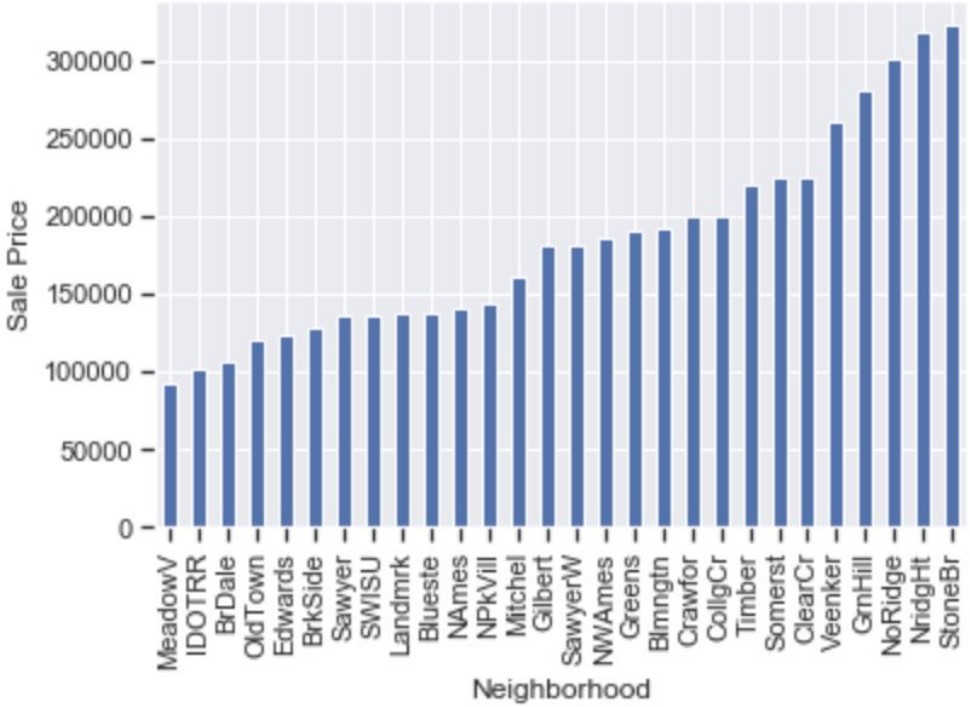




From this graph, we see that despite of the overall condition of the property, people were willing to pay a premium for locations. Eg. Properties in North Ridge Heights, North Ridge and Stone Brook.

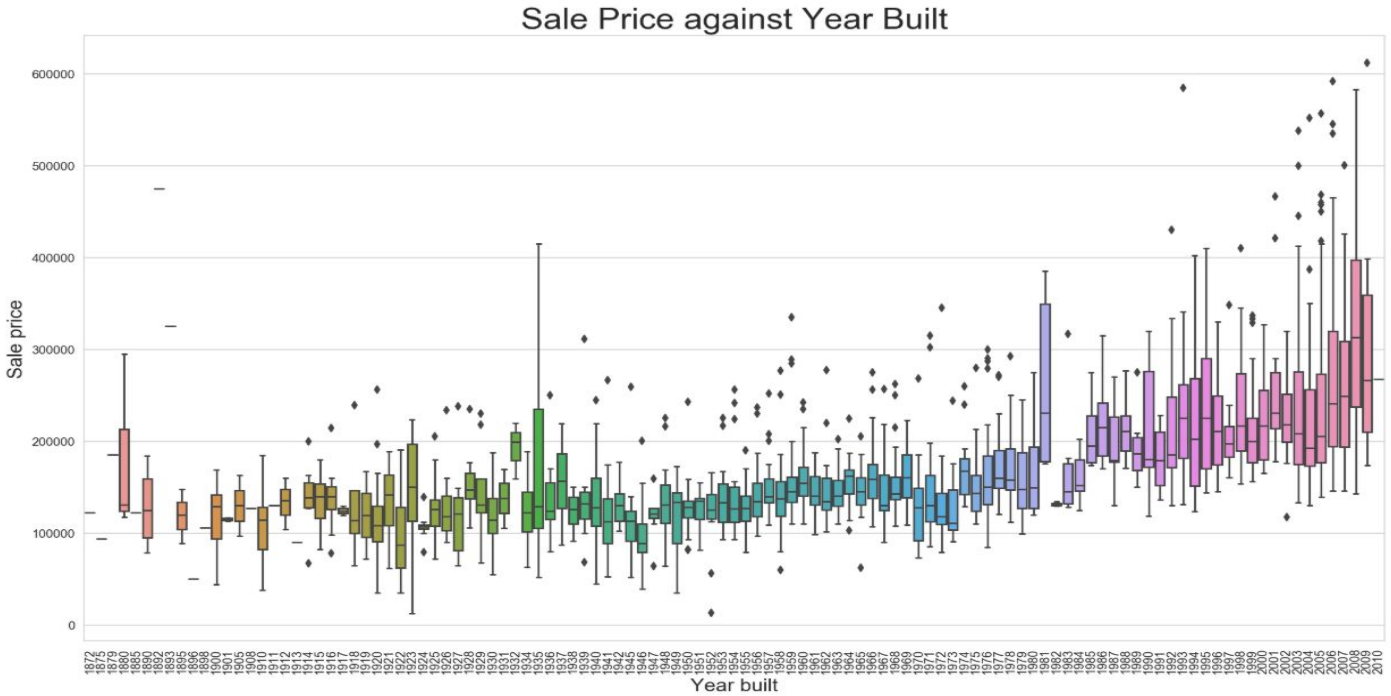
Whereas locations such as Brookside, despite of better overall condition commanded a low sale price.

SALE PRICE AGAINST NEIGHBOURHOOD



The plot above shows that housing in certain neighborhoods are sold at comparatively higher prices than others. StoneBr, NridgHt and NoRidge are the top 3 neighborhoods.

SALE PRICE AGAINST YEAR BUILT



There is a general trend of increases house prices over the years. 2 reasons to explain this trend:

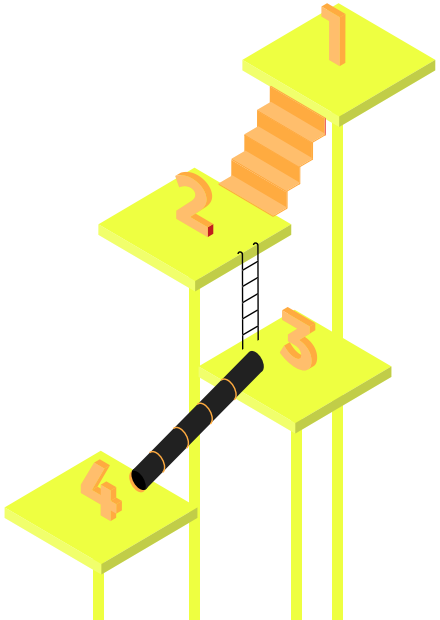
- Inflation of Prices
- Newly Built houses will cost more; less depreciation compared to older houses.

SALE PRICE AGAINST GROUND LIVING AREA & OVERALL QUALITY

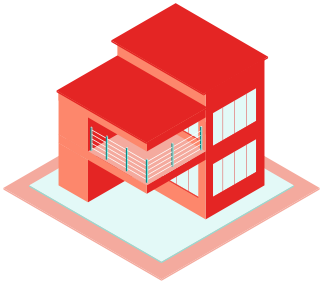
There is a strong positive correlation between ground living area and sale price. More living space would generally result in higher prices.

04

Feature Engineering

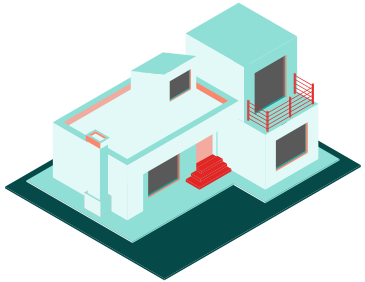


INTERACTION TERMS



Total number of bathrooms

Full Bath + Bsmt Full Bath + 0.5*Half Bath + 0.5*Bsmt Half Bath



Total Sq feet

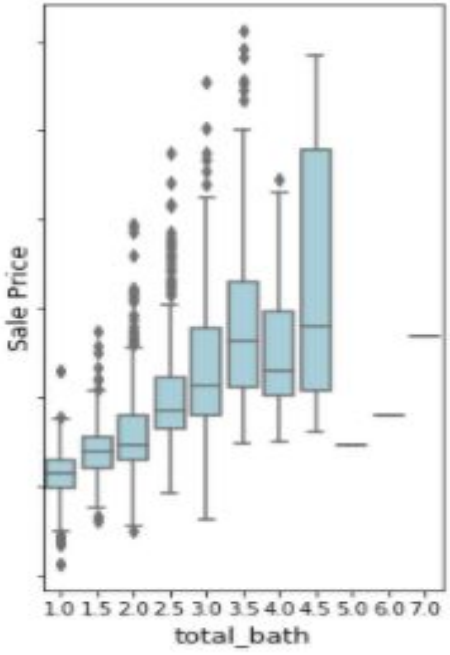
Ground Living Area
+
Total Basement

Neighborhood North cluster

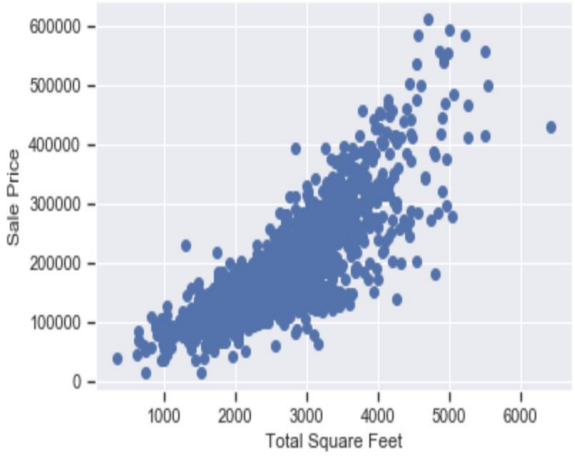
Located near golf courses and lakes

Stone Brook
+
North Ridge
+
North Ridge Heights

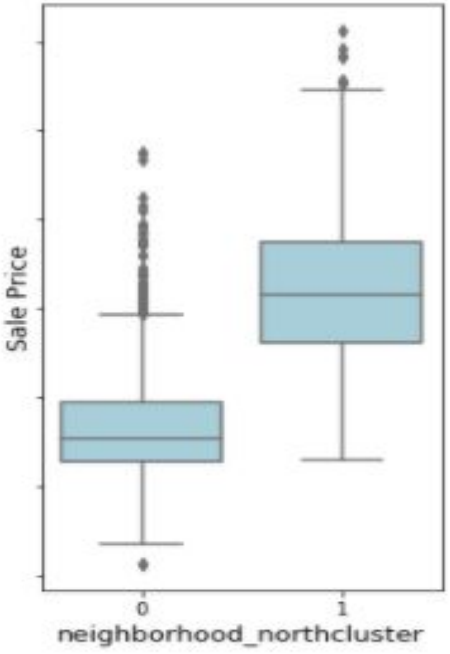
INTERACTION TERMS



Total number of bathrooms

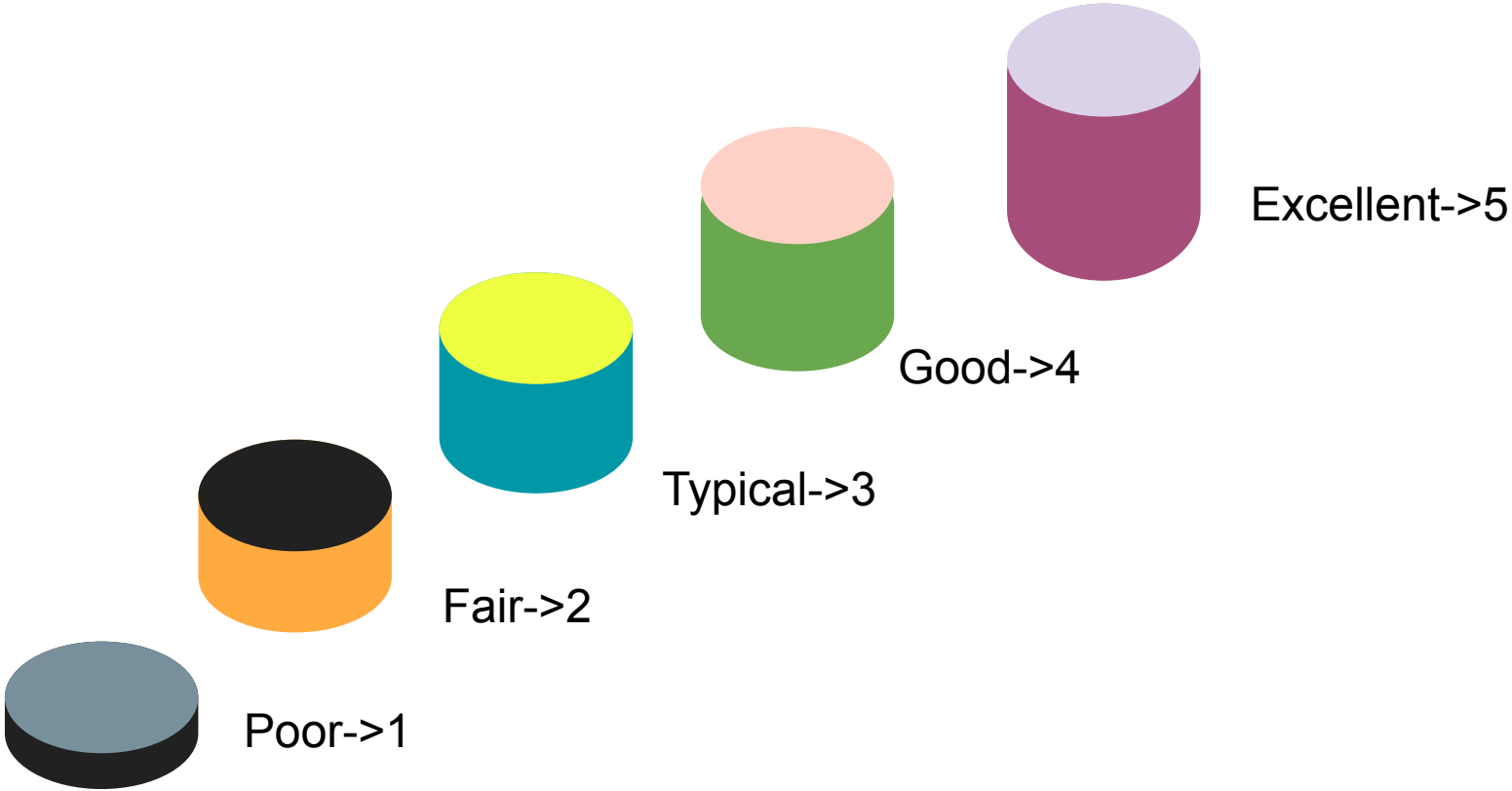


Total Sq feet

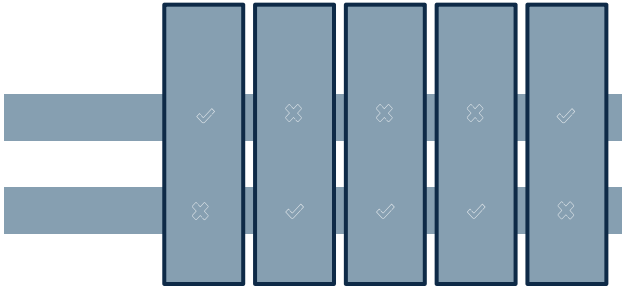


Neighborhood North cluster

Mapping for Ordinal Variables



OHE for Categorical Variables

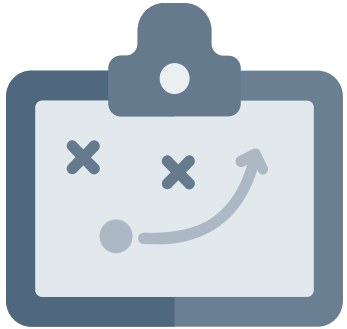


```
final_features =
pd.get_dummies(features).reset_index(drop=True)
```

W	Neighborhood_Somerst	Neighborhood_StoneBr	Neighborhood_Timber	Neighborhood_Veenker
0	0	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	1	0
1	0	0	0	0
...	...	...	...	...
0	0	0	1	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

05

Evaluation of Regression Models



Baseline model

Set all predictions to be the sale price mean from the the training dataset

RMSE:
$79,435

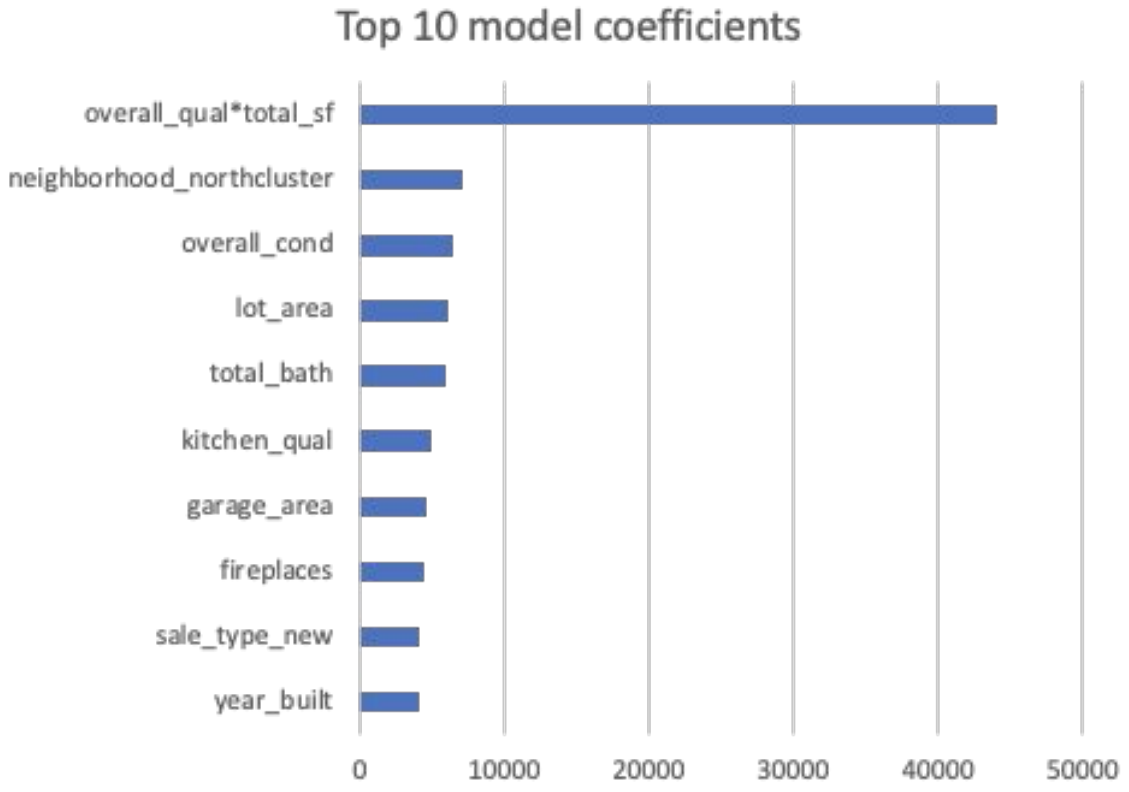
CROSS VALIDATION AND REGULARIZATION

Linear Regression	R2: 0.90	RMSE: 24,800
Lasso	R2: 0.90	RMSE: 24,700
Ridge	R2: 0.90	RMSE: 24,759
Elastic Net	R2: 0.86	RMSE: 29,462

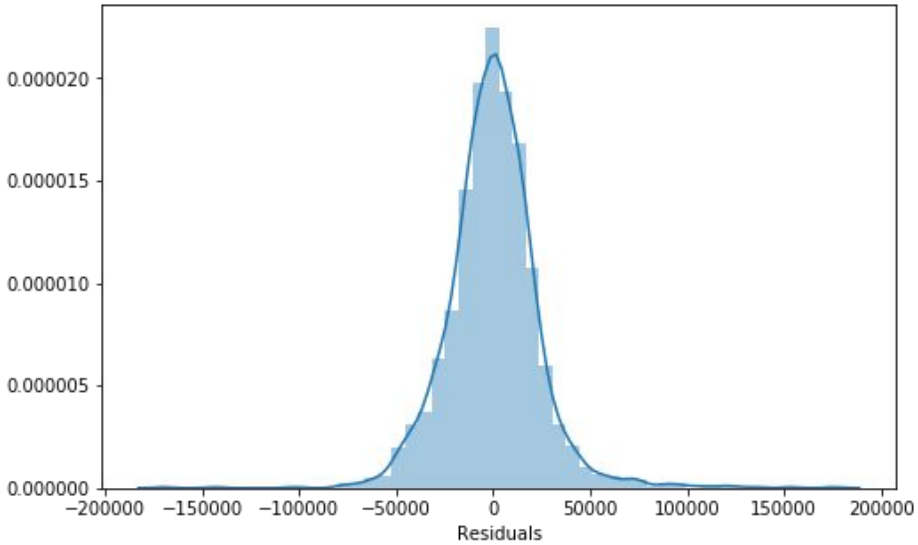
Best parameters: {'alpha': 3.2, 'l1_ratio': 0.8}

Hyperparameters selected using GridSearch

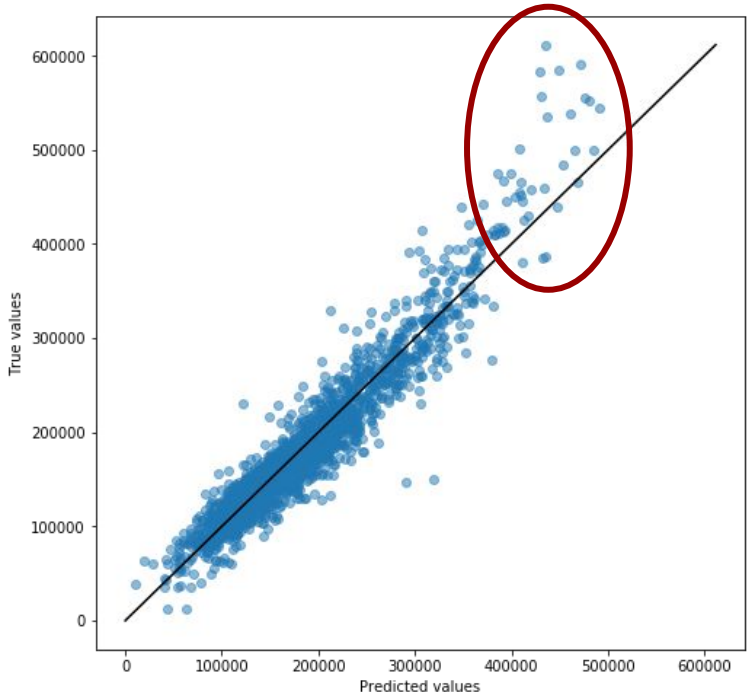
TOP 10 LASSO MODEL COEFFICIENTS



INFERENTIAL VISUALIZATIONS



Residuals are normally distributed with mean 0



Model does well in predicting houses that were priced within the **low to moderate range**. However, the model tends to **underestimate** house at higher prices.

06

Conclusion

IN CONCLUSION

01

Lasso is the model most able to meet our objectives

02

The top 3 predictors, including interaction features: overall_qual*total_sf, neighborhood_northcluster and overall_cond