Ames Housing Prices

Problem Statement

To predict housing prices in Ames, Iowa to help in decision-making for:

- Home-owners to buy a house
- Policy-makers to better regulate housing prices
- Developers to know how and where to build.

Description of the data - 80 variables; categorical, numerical, ordinal

- Overall quality
- Square feet of houses

Data Cleaning

Issue: Many missing values across variables

<insert bar graph>

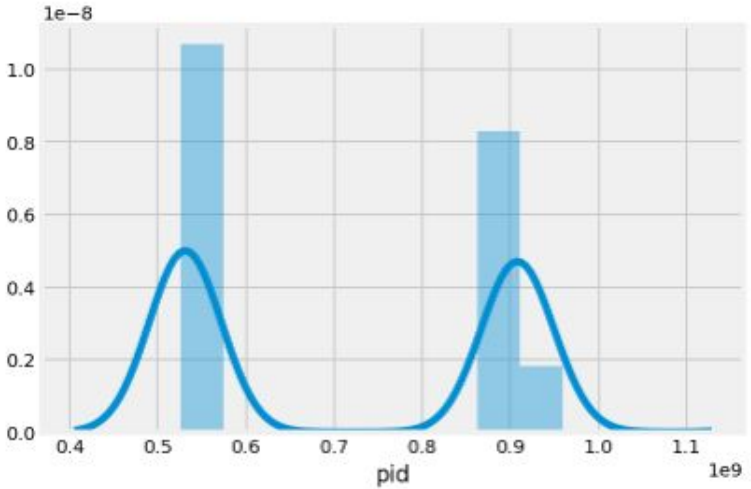
Data Cleaning

Issue: Variables with wrong data type

Feature	DataType	Type	Description
pid	int64	nominal	Parcel identification number - can be used with city web site for parcel review.
ms_subclass	int64	nominal	Identifies the type of dwelling involved in the sale.
ms_zoning	object	nominal	Identifies the general zoning classification of the sale.
street	object	nominal	Type of road access to property.
alley	object	nominal	Type of alley access to property.
land_contour	object	nominal	Flatness of the property.

Data Cleaning

Issue: Variables with wrong data type



pid

```
#renaming the categories in ms_subclass

feature.replace({20: 'twenty',
                 30: 'thirty',
                 40: 'forty',
                 45: 'fortyfive',
                 50: 'fifty',
                 60: 'sixty',
                 70: 'seventy',
                 75: 'seventyfive',
                 80: 'eighty',
                 85: 'eightyfive',
                 90: 'ninety',
                 120: 'hundredtwenty',
                 150: 'hundredfifty',
                 160: 'hundredsixty',
                 180: 'hundredeighty',
                 190: 'hundredninety'}, inplace=True)
```

ms_subclass

Data Cleaning

Issue: Impossible values e.g. garage was built in year 2207

Choice of Imputation becomes critical

Imputations

How we filled in missing values:

- Categorical Variables
- Continuous Variables
- Assumptions

Feature Selection

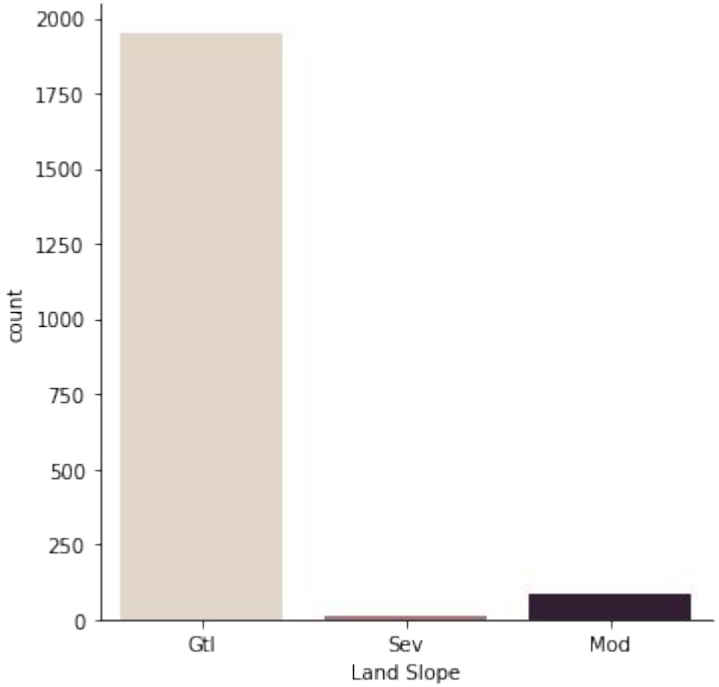
Feature Selection

- **Ordinal Variables**

```
Exter Qual (Ordinal): Evaluates the quality

       Ex       Excellent          5
       Gd       Good               4
       TA       Average/Typical    3
       Fa       Fair               2
       Po       Poor               1
```

- **Categorical Variables**



Feature Selection

a. Combination:

i) Total SF = Total Bsmt SF + 1st Flr SF + 2nd Flr SF

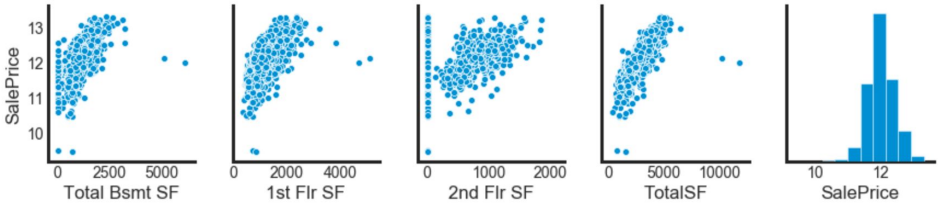
Apply the same for:

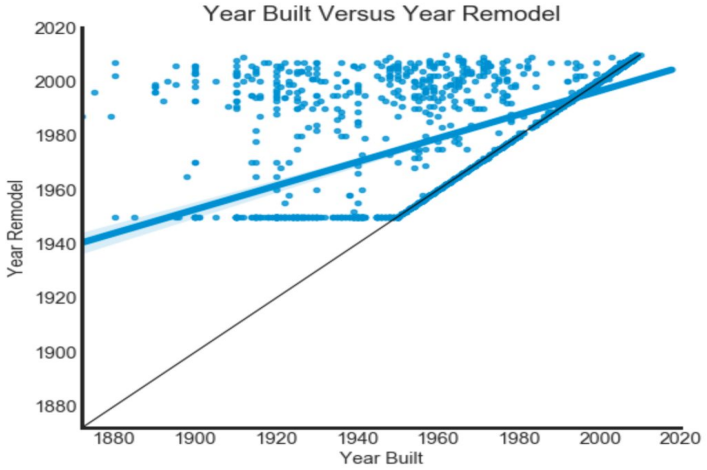
Total Full Bathroom = Full Bath + Bsmt Full Bath

Total Half Bathroom = Half Bath + Half Bathroom

ii) Remodel = 1 (if yes)

0 (if no)



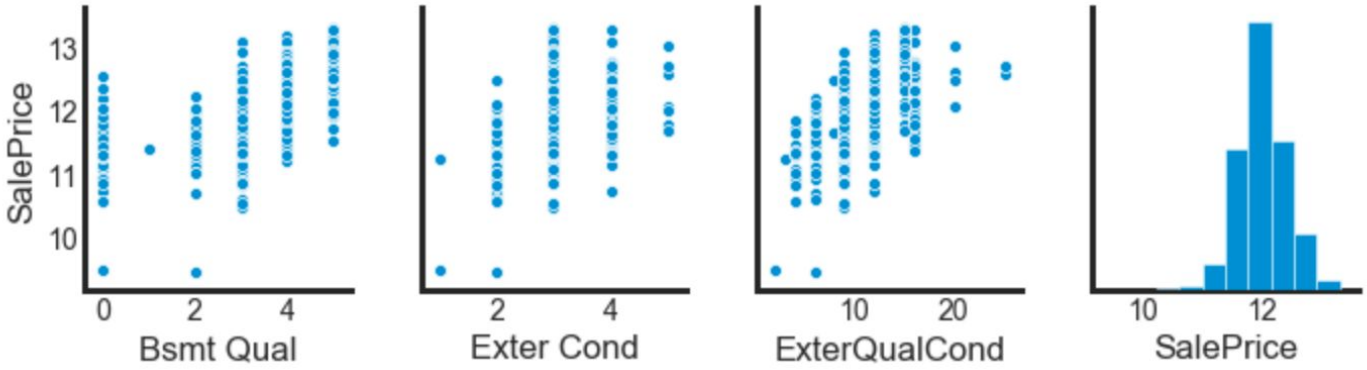


Feature Selection

b. Interaction:

i) Interaction [External Quality Condition] = External Quality * External Condition

Apply the same for: Basement Quality * Basement Condition, Garage Quality * Garage Condition



Model of Choice

Regression Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$

Linear Regression, Ridge, Lasso

$$\text{minimize: } MSE(\beta_0, \beta_1, \ldots) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \right)^2$$

Use cross_val_score to evaluate all the three models

Models chosen: Lasso

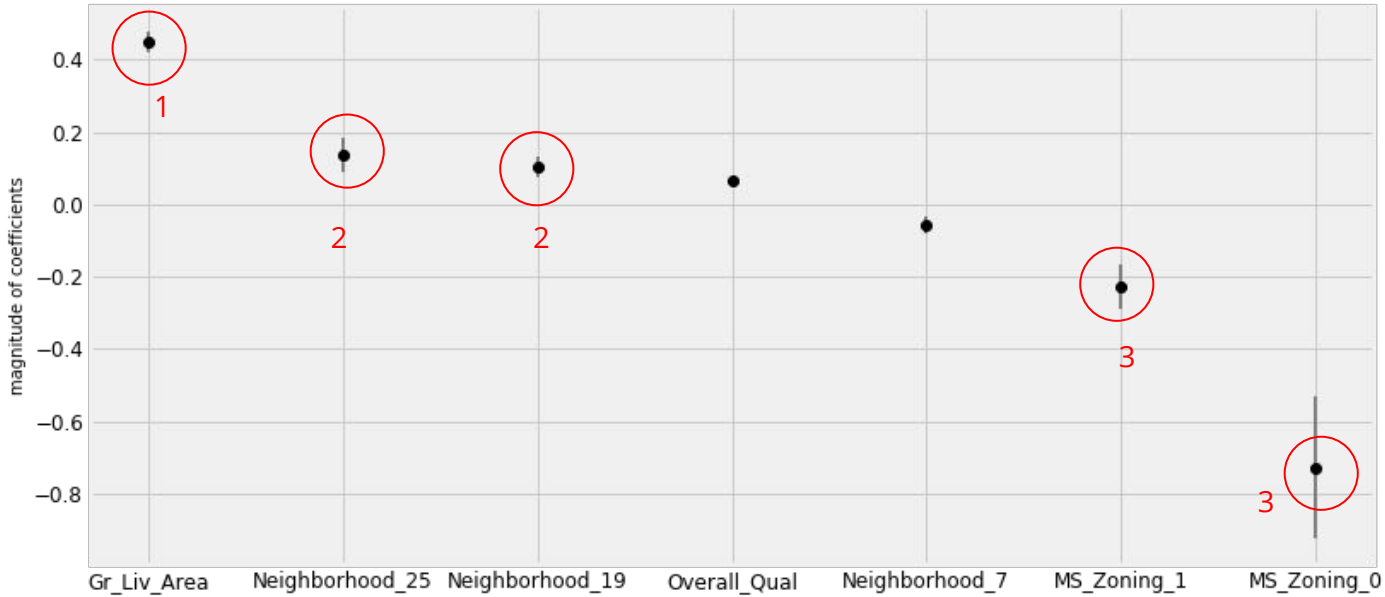
Reason: Lasso is able to do feature selection by zeroing out the insignificant features

Performance measure: R-square

It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable).

R-square : 89% (test set), 91% (train set)

Conclusion & Recommendation



1. Developers are encouraged to build housing with larger living room.
2. Certain Neighbourhoods (North-ridge Height,Somerset) seems to be popular
3. Avoid developing in agricultural/commercial area

Improvements

- More Data
  - Crime-rate in the region
  - Buyer info
- Considerations on business standpoint
  - Sale-price or Sale-price per sq feet

Note: Will not generalise well to other cities