House Price Prediction

GA DSI Project 2: Sep 27, 2021

Basazin Belhu

Introduction

Problem Statements:

A real estate company has datasets that contain qualitative and quantitative criteria of the houses they have sold. They want to improve the sales and predict the price of their future new houses.

This project is focus on develop a machine learning model that able to extract the different criteria that determining the house price. It also optimize sales price of the house as accurate as possible depending on the information feed.

- Want to know which criteria mainly determine the house price
- Create a multi linear regression to predict the price of the house based on features
- Want to know how well the model predict the price.

Initial Approach

```
15]: train_data.shape

15]: (2051, 81)

16]: train_data.isnull().sum().sort_values(ascending = False)[:10]

16]: Pool QC          2042
     Misc Feature     1986
     Alley            1911
     Fence            1651
     Fireplace Qu     1000
     Lot Frontage      330
     Garage Finish     114
     Garage Qual       114
     Garage Yr Blt     114
     Garage Cond       114
     dtype: int64
```

```
Id                     int64
PID                    int64
MS SubClass            int64
MS Zoning             object
Lot Frontage         float64
                        ...
Misc Val               int64
Mo Sold                int64
Yr Sold                int64
Sale Type             object
SalePrice              int64
Length: 81, dtype: object
```

81 Variables:

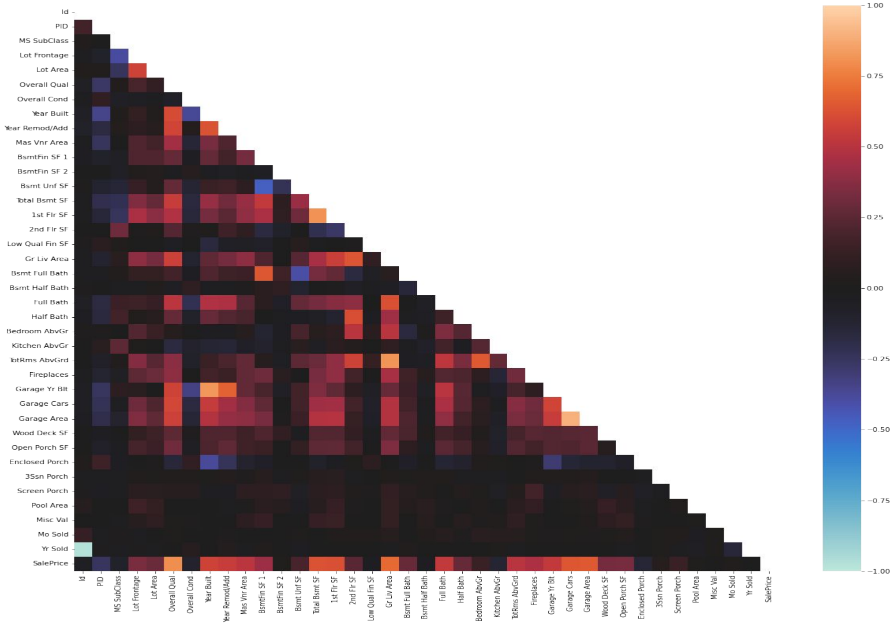
- Sales price: dependent variable and the price the house was sold
- 80 Independent variables

https://memegenerator.net/instance/72837712/very-excited-dog-omg

Basic Summary Statistics

	count	mean	std	min	25%	50%	75%	max
PID	2051.0	7.135900e+08	1.886918e+08	526301100.0	528458140.0	5.354532e+08	907180080.0	924152030.0
MS SubClass	2051.0	5.700878e+01	4.282422e+01	20.0	20.0	5.000000e+01	70.0	190.0
Lot Frontage	2051.0	6.905520e+01	2.130636e+01	21.0	60.0	6.905520e+01	78.0	313.0
Lot Area	2051.0	1.006521e+04	6.742489e+03	1300.0	7500.0	9.430000e+03	11513.5	159000.0
Overall Qual	2051.0	6.112140e+00	1.426271e+00	1.0	5.0	6.000000e+00	7.0	10.0
Overall Cond	2051.0	5.562165e+00	1.104497e+00	1.0	5.0	5.000000e+00	6.0	9.0
Year Built	2051.0	1.971709e+03	3.017789e+01	1872.0	1953.5	1.974000e+03	2001.0	2010.0
Year Remod/Add	2051.0	1.984190e+03	2.103625e+01	1950.0	1964.5	1.993000e+03	2004.0	2010.0
Mas Vnr Area	2051.0	9.862652e+01	1.743247e+02	0.0	0.0	0.000000e+00	159.0	1600.0
BsmtFin SF 1	2051.0	4.420848e+02	4.611950e+02	0.0	0.0	3.680000e+02	733.5	5644.0
BsmtFin SF 2	2051.0	4.793564e+01	1.649641e+02	0.0	0.0	0.000000e+00	0.0	1474.0
Bsmt Unf SF	2051.0	5.674515e+02	4.450228e+02	0.0	220.0	4.740000e+02	811.0	2336.0
Total Bsmt SF	2051.0	1.057472e+03	4.499080e+02	0.0	793.0	9.940000e+02	1318.5	6110.0
1st Flr SF	2051.0	1.164488e+03	3.964469e+02	334.0	879.5	1.093000e+03	1405.0	5095.0
2nd Flr SF	2051.0	3.293291e+02	4.256710e+02	0.0	0.0	0.000000e+00	692.5	1862.0
Low Qual Fin SF	2051.0	5.512921e+00	5.106887e+01	0.0	0.0	0.000000e+00	0.0	1064.0
Gr Liv Area	2051.0	1.499330e+03	5.004478e+02	334.0	1129.0	1.444000e+03	1728.5	5642.0
Bsmt Full Bath	2051.0	4.271087e-01	5.225887e-01	0.0	0.0	0.000000e+00	1.0	3.0
Bsmt Half Bath	2051.0	6.338372e-02	2.515902e-01	0.0	0.0	0.000000e+00	0.0	2.0
Full Bath	2051.0	1.577279e+00	5.492794e-01	0.0	1.0	2.000000e+00	2.0	4.0
Half Bath	2051.0	3.710385e-01	5.010427e-01	0.0	0.0	0.000000e+00	1.0	2.0
Bedroom AbvGr	2051.0	2.843491e+00	8.266183e-01	0.0	2.0	3.000000e+00	3.0	8.0
Kitchen AbvGr	2051.0	1.042906e+00	2.097900e-01	0.0	1.0	1.000000e+00	1.0	3.0
TotRms AbvGrd	2051.0	6.435885e+00	1.560225e+00	2.0	5.0	6.000000e+00	7.0	15.0
Fireplaces	2051.0	5.909313e-01	6.385163e-01	0.0	0.0	1.000000e+00	1.0	4.0
Garage Yr Blt	2051.0	1.868726e+03	4.541337e+02	0.0	1957.0	1.978000e+03	2001.0	2207.0
Garage Cars	2051.0	1.775719e+00	7.653569e-01	0.0	1.0	2.000000e+00	2.0	5.0
Garage Area	2051.0	4.734408e+02	2.161351e+02	0.0	319.0	4.800000e+02	576.0	1418.0
Wood Deck SF	2051.0	9.383374e+01	1.285494e+02	0.0	0.0	0.000000e+00	168.0	1424.0
Open Porch SF	2051.0	4.755680e+01	6.674724e+01	0.0	0.0	2.700000e+01	70.0	547.0
Enclosed Porch	2051.0	2.257192e+01	5.984511e+01	0.0	0.0	0.000000e+00	0.0	432.0
3Ssn Porch	2051.0	2.591419e+00	2.522961e+01	0.0	0.0	0.000000e+00	0.0	508.0
Screen Porch	2051.0	1.651146e+01	5.737420e+01	0.0	0.0	0.000000e+00	0.0	490.0
Pool Area	2051.0	2.397855e+00	3.778257e+01	0.0	0.0	0.000000e+00	0.0	800.0
Misc Val	2051.0	5.157435e+01	5.733940e+02	0.0	0.0	0.000000e+00	0.0	17000.0
Mo Sold	2051.0	6.219893e+00	2.744736e+00	1.0	4.0	6.000000e+00	8.0	12.0
Yr Sold	2051.0	2.007776e+03	1.312014e+00	2006.0	2007.0	2.008000e+03	2009.0	2010.0
SalePrice	2051.0	1.814697e+05	7.925866e+04	12789.0	129825.0	1.625000e+05	214000.0	611657.0

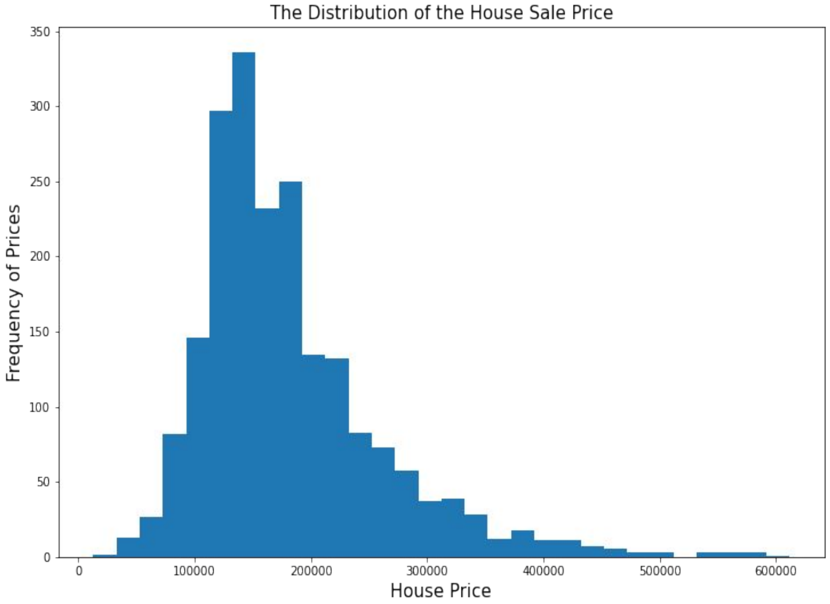
Check the correlation matrices



- Strong Correlation:
  - Overall quality, Gr live area,
  - Garage area, and garage cars

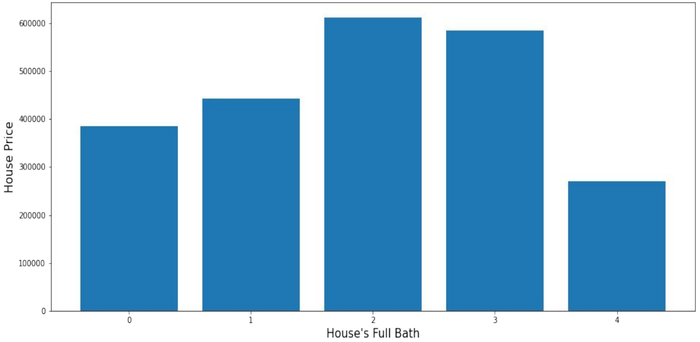
- Weak Correlation:
  - Year and Month of sold
- PID has a negative correlation

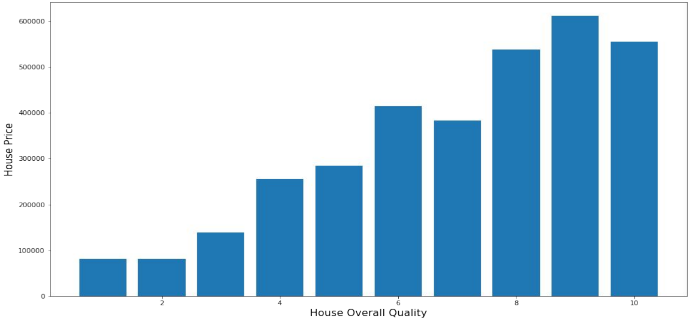
Target Distribution:

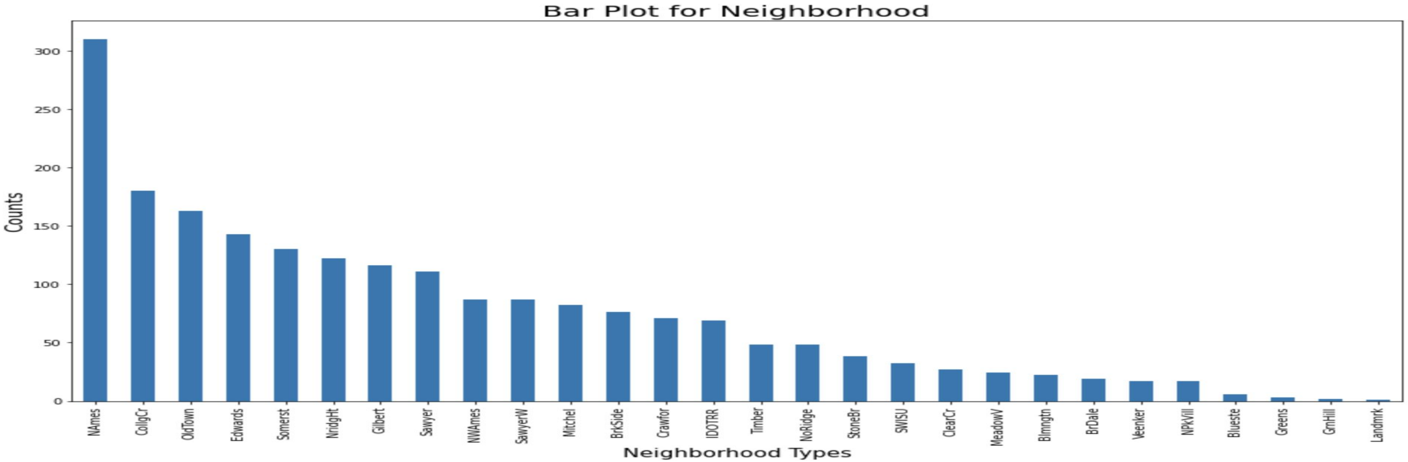


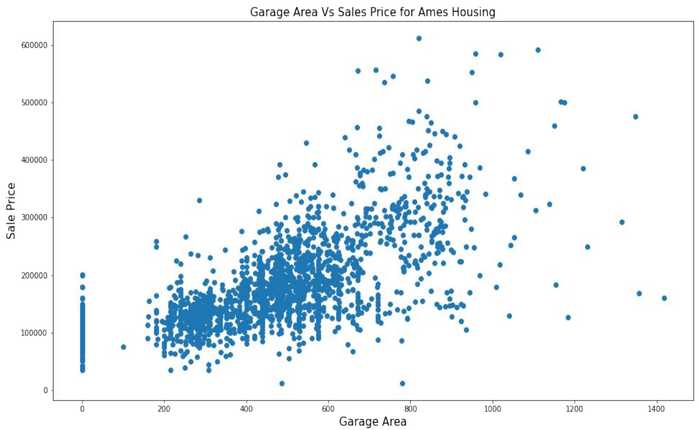
- Close to normal distribution
- No need to transform the target values

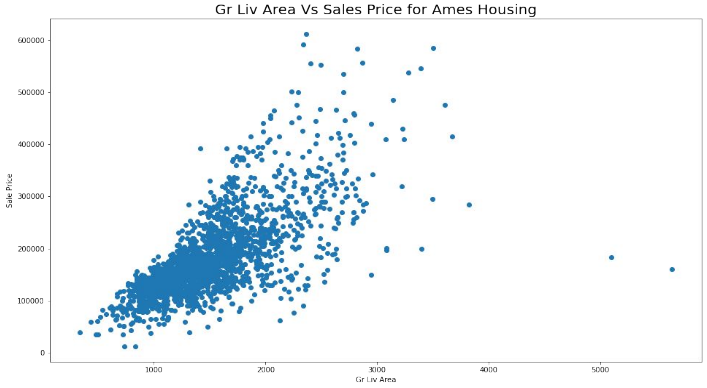
Bar Plots Analysis

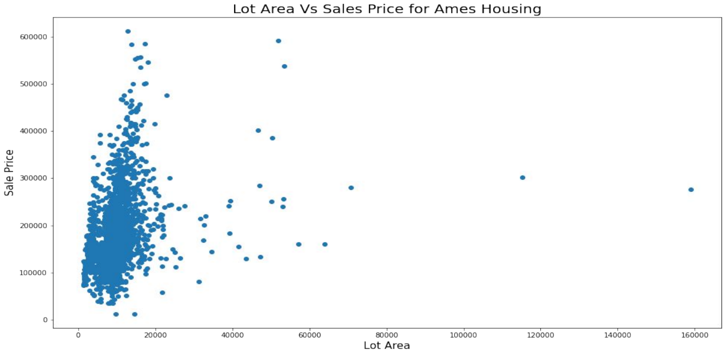


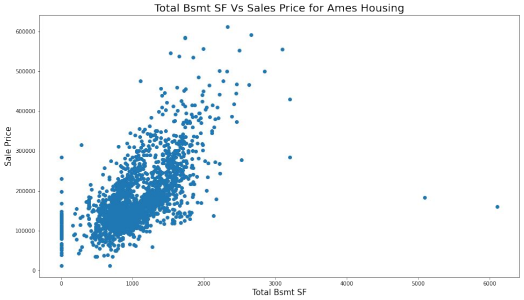




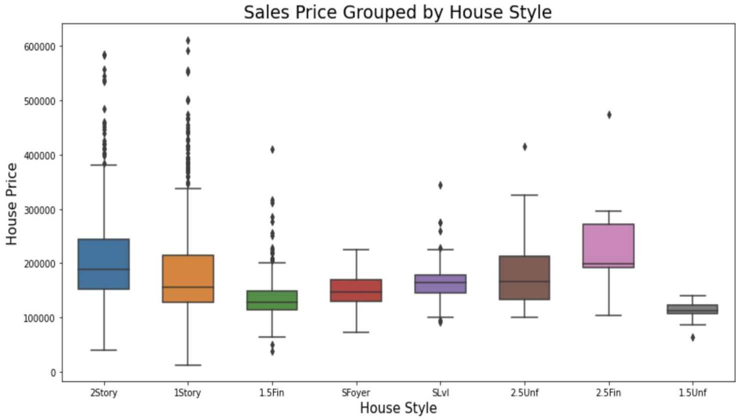

Garage Area Vs Sales Price for Ames Housing
Sale Price
Garage Area

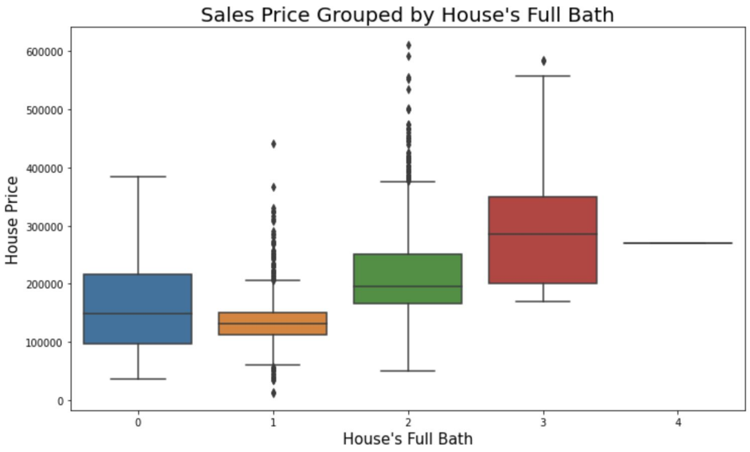

Gr Liv Area Vs Sales Price for Ames Housing
Sale Price
Gr Liv Area

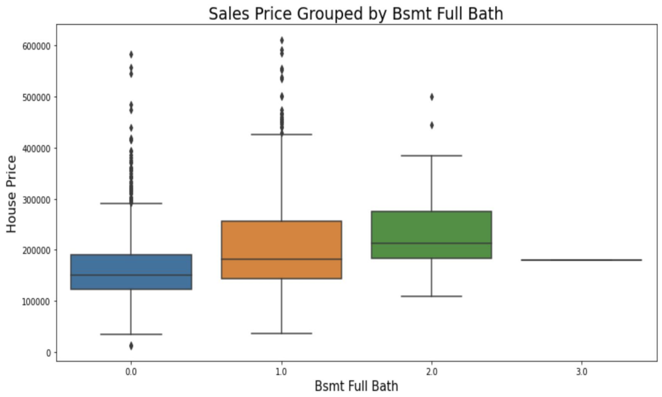

Lot Area Vs Sales Price for Ames Housing
Sale Price
Lot Area

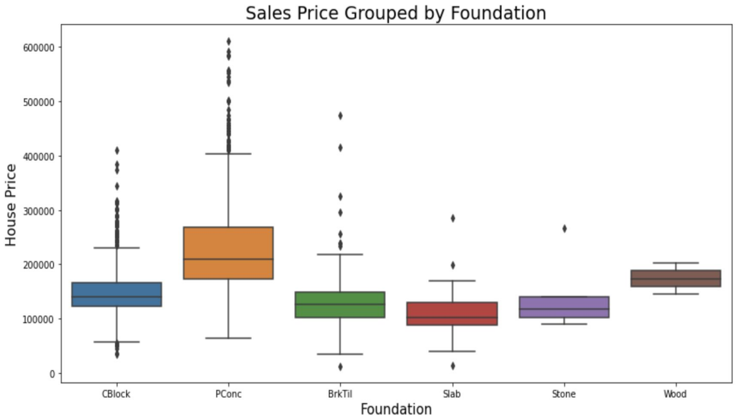

Total Bsmt SF Vs Sales Price for Ames Housing
Sale Price
Total Bsmt SF

Box Plot Analysis









Data Processing:

Training Datasets:

- 75% of our loaded training data
- The model training based on this datasets

Testing Datasets:

- It is a 25% of the loaded training data
- Used to test the model

Feature Engineering and Selection

- Drop columns that are less correlation with the sales price
- Drop columns that has mostly filled with 0
- Add a columns from year of sold - year of built
- Select columns that has varies in observations
- One hot encoded some categorical features

Modeling

Baseline Model

- Based on the mean of the training datasets
- Baseline RMSE Train: 78434.16
- Baseline RMSE Test: 81625.43

Sklearn Linear Regression:

Formula

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable
f = function
X_i = independent variable
β = unknown parameters
e_i = error terms

Assumptions:

❖ Linearity features
❖ Independence of observations
❖ Normality error distribution
❖ Homoscedasticity equal variance
❖ Multicollinearity features not linearly correlated

Metrics:

- Determine the Coefficient of the features
- Metrics RMSE and R2

Linear Reg Con't

The model training on 75% of our training data:

Significance of features:

- R2 training: 0.883
- R2 testing: 0.89

Accuracy of the model:

- Train RMSE: 26801.87
- Test RMSE: 26886.049

	columns	coef
51	Neighborhood_GrnHill	100930.502485
66	Neighborhood_StoneBr	55003.826436
59	Neighborhood_NoRidge	42510.996120
60	Neighborhood_NridgHt	40662.772054
30	Bsmt Qual_Ex	27905.119636
28	Electrical_Mix	22093.281609
73	Bsmt Exposure_Gd	18371.625990
65	Neighborhood_Somerst	14923.136764
86	Functional_Typ	14872.846750
81	Functional_Min1	13727.117475
57	Neighborhood_NPkVill	13458.819397
69	Lot Config_CulDSac	13135.274010
83	Functional_Mod	12643.435978
17	House Style_1Story	12530.062904
34	Bsmt Qual_TA	11409.569378

- R2 and RMSE is acceptable
- Some coefficients are very unlikely high and Scale the data to try different model

Lasso and GridSearch

Significance:

- R2 for training with lasso:0.88
- R2 for testing lass: 0.89

Accuracy:

- Train RMSE: 26801.89
- Test RMSE: 26884.53

Strong R2 values: the 88% variability in the house prices are due to the feature we selected in our model.

```
grid_coef_df.head()
```

	columns	coef
7	Gr Liv Area	25244.525336
6	Overall Qual	11252.732672
60	Neighborhood_NridgHt	9921.375242
30	Bsmt Qual_Ex	8051.237313
66	Neighborhood_StoneBr	7749.943942

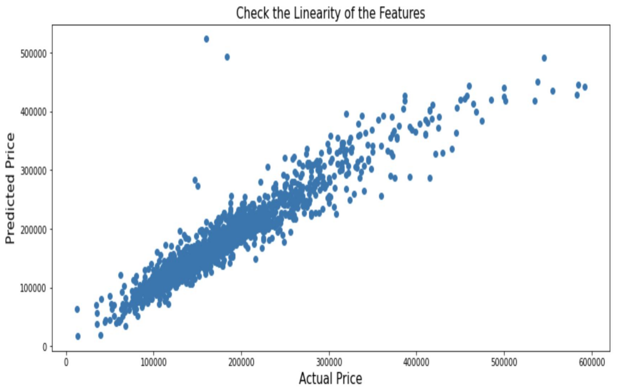
```
grid_coef_df.tail()
```

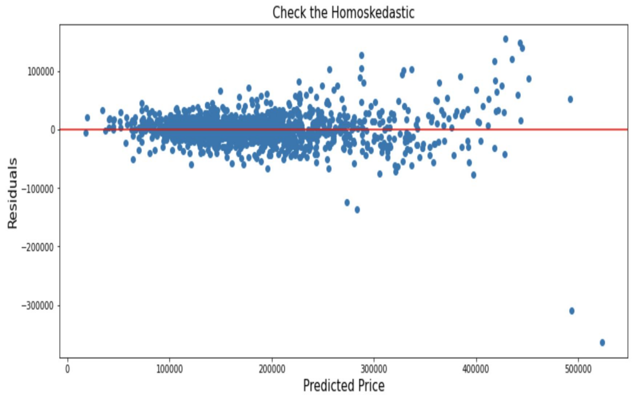
	columns	coef
8	Age Sold	-9724.117723
36	Exter Qual_Gd	-11686.177651
37	Exter Qual_TA	-13275.173063
24	Kitchen Qual_Gd	-13691.197681
25	Kitchen Qual_TA	-16063.092559

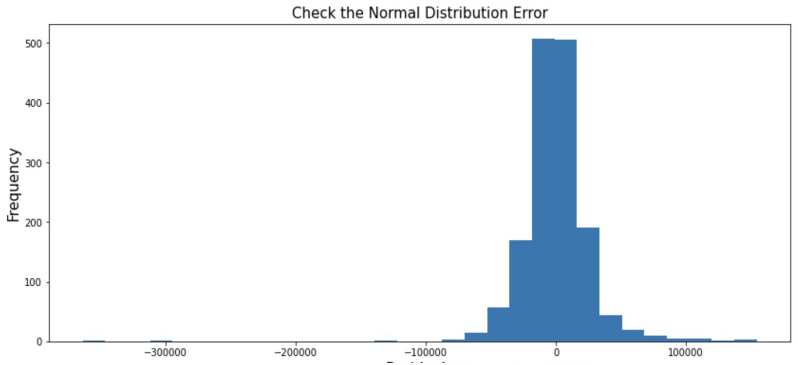
Interprete top features:

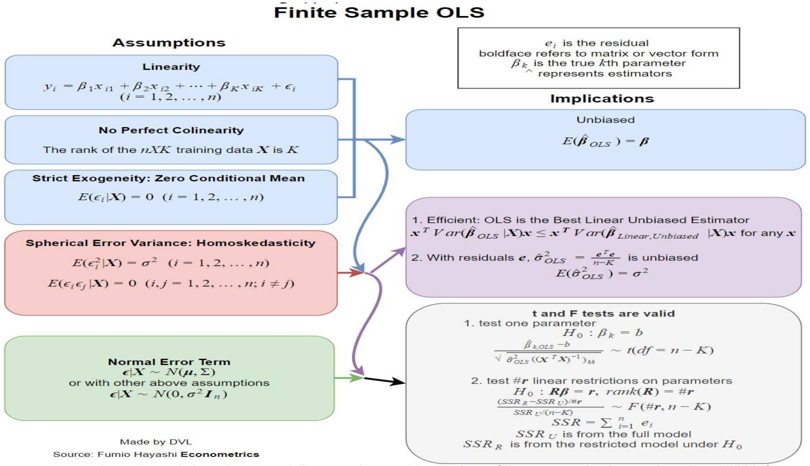
- Gr Liv Area: A one unit increase in Gr Liv Area of the house we expect the house price increased by $25244, all else held constant
- Overall Qual: A one unit increase in Overall Qual of the house we expect the house price increased by $11252, all else held constant
- Neighborhood_GrnHil: A 1 unit increase in Neighborhood_GrnHil the house price predicted to increase by $9921., all else hold constant or keep equal.
- Neighborhood_StoneB: A 1 unit increase in Neighborhood_StoneB of a house, we expect the price increased by $7749, all else keep constant or equal.
- Age Sold: A 1 unit older of a house, we expect the price decreased by $9274., all else keep constant or equal.
- Kitchen Qual_TA: A one unit decreased in Kitchen Qual_TA, we expect the house price will decreased by $16000.
- Kitchen Qual_Gd: A one unit decreased in Kitchen Qual_Gd, we expect the house price will decreased by $13690.

Does Assumption Fulfilled?









https://towardsdatascience.com/all-assumptions-and-implications-of-linear-regression-in-one-chart-5674c060025f

Conclusion

We solved the problems that we want to discover

- Want to know which criteria has main impact on determine the house price
    a. Features with high negative and positive coefficients as strong impact in house price
- Create a multi linear regression to predict the price of the hase based on the import
    a. The model we create has features that contributes 88% of variability in house price
- They want to know the performance of this model, how well the model predict the price.
    a. The model predict the house price in unseen data and it has the smaller RMSE and high R2 values

Future Work:

Add feature interaction and develop an app with this model, so the company just need to put the features and the app will tell them the estimated prices