Prediction of Ames Housing saleprice

By Irene Izere

Problem Statement

The goal of this project is creating a regression model based on the Ames Housing Dataset. This model will predict the price of a house at sale. As a data scientist , I was hired by a real estate agency to help them understand the market .With this prediction it will help them understand if it is a good time to sell or buy more houses for the business.

Dataset and Background

In this dataset , the collect data of the housing .This dataset is comprise of different characteristics of the house already sold such as their location, , their sizes, types, finishing materials ,utilities they have , the time they were sold and how much they were sold.

Using this dataset we need to build a model that will predict how much the house will sell given those data.

Dealing with missing Values

- The dataset has **2051** observations and **81** attributes .
- First step, dropping all columns with more than half values missing .
- Replacing with missing values with the mean of the columns or mode depending of it has categories or it is the size.
- Made an assumption also to match values of different columns for example the Garage year built was missing and was replaced by the building year making an assumption that there were built at the same time.

Modeling : Linear Regression

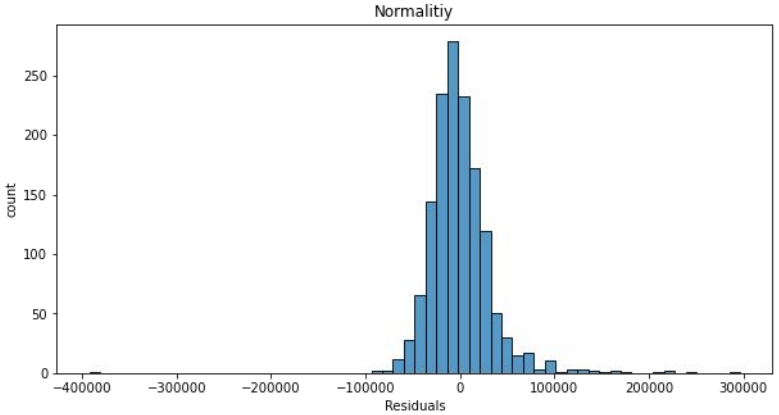
1. **First model**

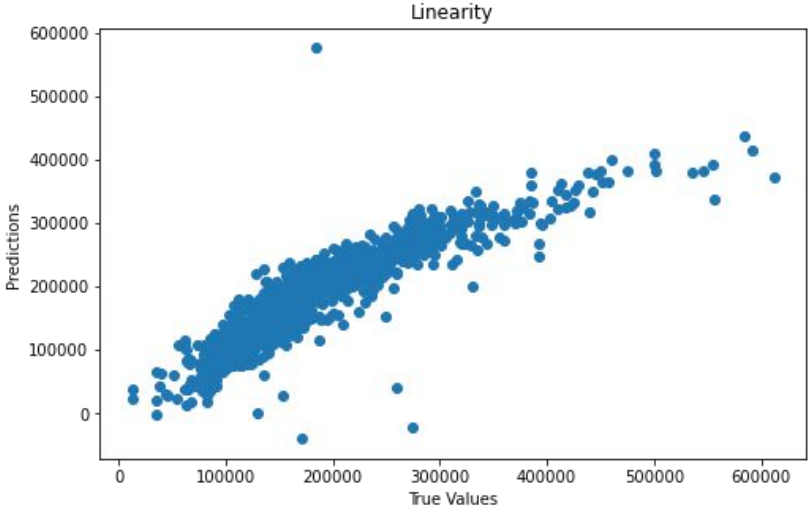
In this model , we are using only numerical features of the datasets.

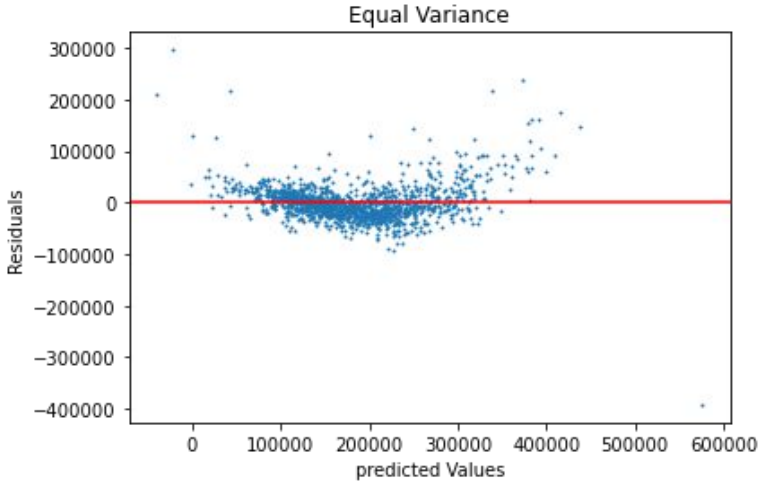
After choosing our features , we split our dataframe into training -split which allows to test the model and see how it is working.

The LINE Assumptions

- On the Equal variance plot we can see a parabolic pattern which means that the line assumptions was violated.







Results

Using this model to calculate **R-Squared** we can see that on trained data the model can have a score of 82% but on unseen data 80% which can be interpreted as having a high variance .

The model will not generalize well on unseen data .

How can we improve it ?

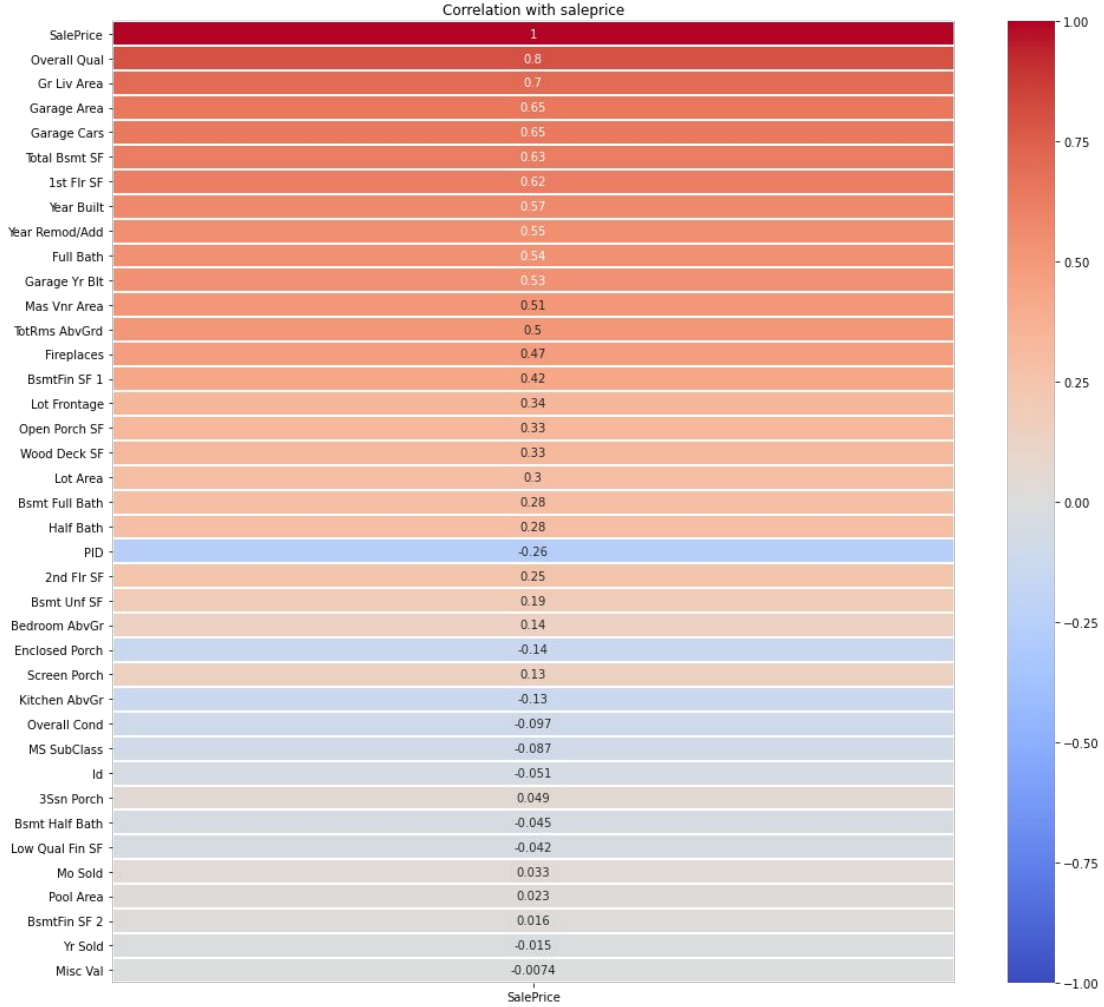
- One way to improve the model , we can use **log regression** to fix the equal variance
- We can also use use the Ridge and Lasso to regularize the model .

2. Second model

In this model , we are using numerical features with more than 0.25 correlation with the salepreice and adding some categoricals features to the datasets.

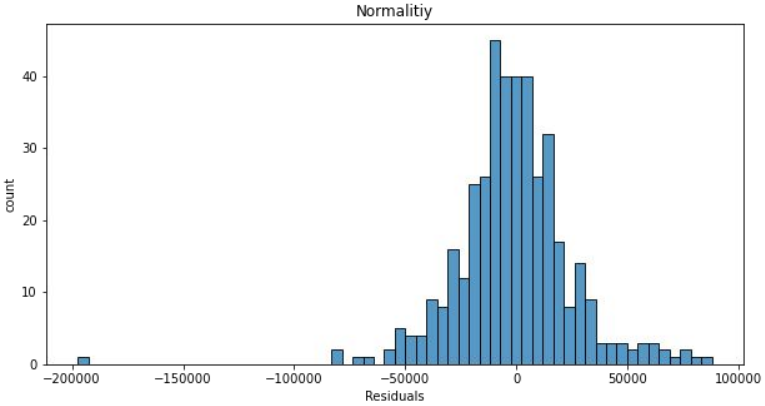
After choosing our features , we split our dataframe into training -split which allows to test the model and see how it is working. We will do that after dummify the categorical which is a way to give a value to a characteristics 1 if it has it and 0 if it doesn't.

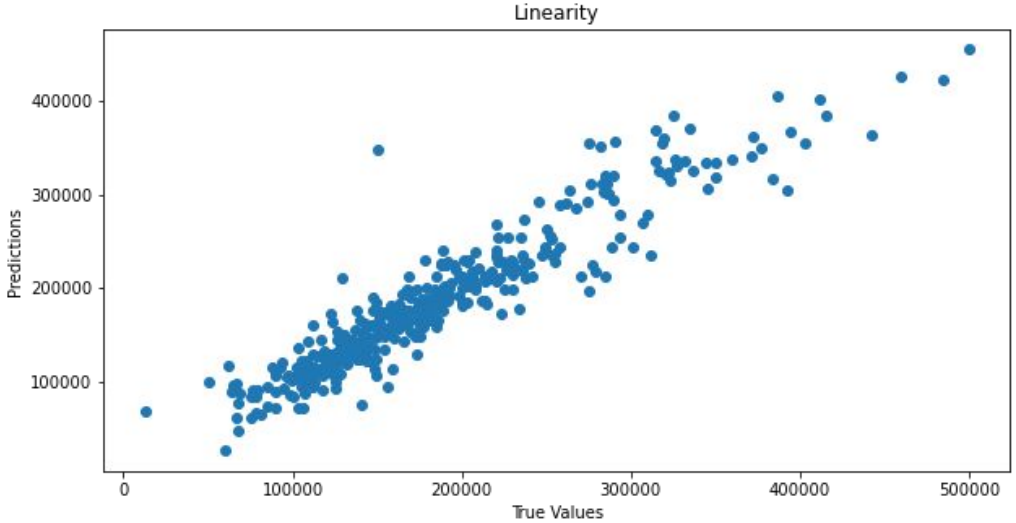
Doing that we can now compare apples and apples.

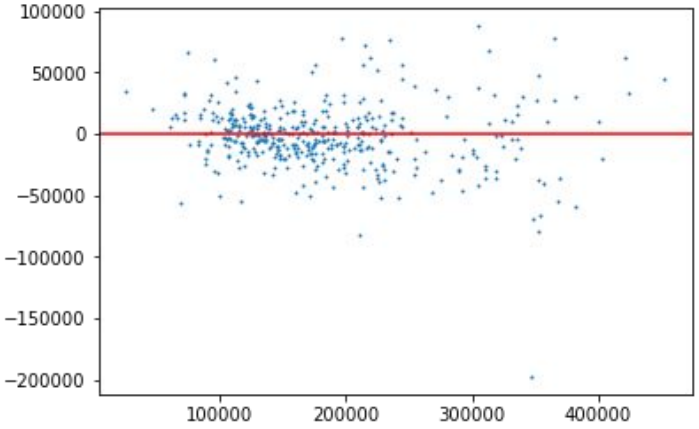

Correlation with saleprice
SalePrice 1
Overall Qual 0.8
Gr Liv Area 0.7
Garage Area 0.65
Garage Cars 0.65
Total Bsmt SF 0.63
1st Flr SF 0.62
Year Built 0.57
Year Remod/Add 0.55
Full Bath 0.54
Garage Yr Blt 0.53
Mas Vnr Area 0.51
TotRms AbvGrd 0.5
Fireplaces 0.47
BsmtFin SF 1 0.42
Lot Frontage 0.34
Open Porch SF 0.33
Wood Deck SF 0.33
Lot Area 0.3
Bsmt Full Bath 0.28
Half Bath 0.28
PID -0.26
2nd Flr SF 0.25
Bsmt Unf SF 0.19
Bedroom AbvGr 0.14
Enclosed Porch -0.14
Screen Porch 0.13
Kitchen AbvGr -0.13
Overall Cond -0.097
MS SubClass -0.087
Id -0.051
3Ssn Porch 0.049
Bsmt Half Bath -0.045
Low Qual Fin SF -0.042
Mo Sold 0.033
Pool Area 0.023
BsmtFin SF 2 0.016
Yr Sold -0.015
Misc Val -0.0074
SalePrice
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

The LINE Assumptions

- On this model we see that the line assumptions is not violated .







Results

- Using this linear regression model to calculate **R-Squared** we can see that on trained data the model can have a score of 86.7% but on unseen data 88.6% .
- Using this Ridge regression model to calculate **R-Squared** we can see that on trained data the model can have a score of 86.7% but on unseen data 88.6% .
- Using this linear regression model to calculate **R-Squared** we can see that on trained data the model can have a score of 86.6% but on unseen data 88.9% .

- Using this linear regression model to calculate **R-Squared** we can see that on trained data the model can have a score of 85.2% but on unseen data 88.7% .