HOUSE PRICE PREDICTION

Submitted by:

Atul Mahajan

**INTRODUCTION**

**Business Problem Framing**

The objective was to model the price of houses with the available independent variables. This model can then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**Conceptual Background of the Domain Problem**

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. I was required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

* Which variables are important to predict the price of variable?
* How do these variables describe the price of the house?

**Technical Requirements:**

* Data contains 1460 entries each having 81 variables.
* Data contains Null values. You need to treat them using the domain knowledge and your own understanding.
* Extensive EDA has to be performed to gain relationships of important variable

and price.

* Data contains numerical as well as categorical variable. You need to handle them

accordingly.

* Need to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
* Need to find important features which affect the price positively or negatively.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This is a Regression problem, where our end goal is to predict the Prices of House based on given data. I will be dividing my data into **Training** and **Testing** parts. A Regression Model will be built and trained using the Training data and the Test data will be used to predict the outcomes. This will be compared with available test results to find how well the model has performed.

The ‘r2’ score will be used to determine the best model among,

* + Linear Regression with Lasso, Ridge
  + Random Forest Regression
  + XGBoost
* The best results were obtained using Lasso Regression. So, let’s understand a little

about it.

In a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x, where B0 —intercept

B1 —coefficient

x —independent variable

y —output or the dependent variable

In higher dimensions when we have more than one input (x),

The General equation for a Multiple linear regression with p — independent variables: Y=B0 + B1 \* X1 + B2 \* X2 + + Bp \* Xp + E(Random Error or Noise)

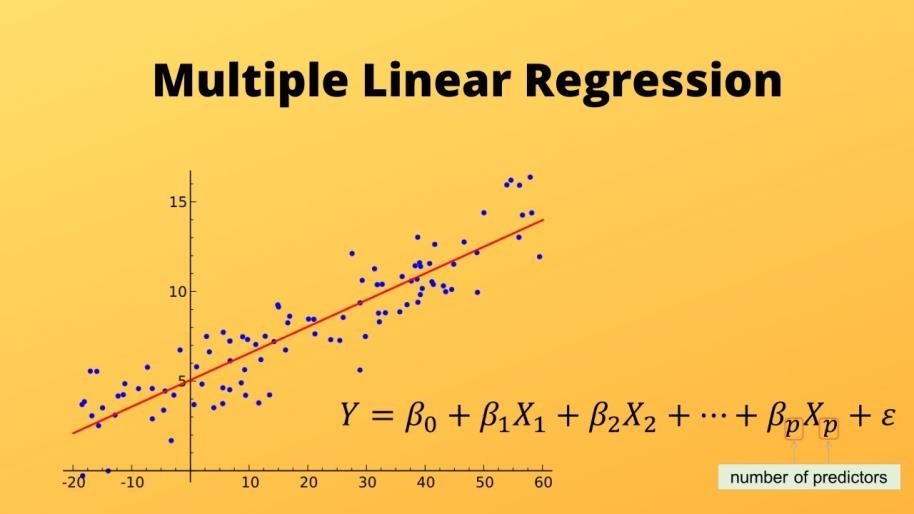


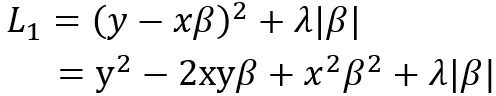
Image Source: https://morioh.com/p/0d9b2bedf683

#### Let’s consider a regression scenario where ‘y’ is the predicted vector and ‘x’ is the feature matrix. Basically in any regression problem, we try to minimize the squared error. Let ‘β’ be the vector of parameters (weights of importance of features) and ‘p’ be the number of features

Now, let’s discuss the case of **lasso regression**, which is also called L1 regression since it uses the L1 norm for regularization. In lasso regression, we try to solve the below minimization problem:

Minimized L1

For simplicity, let p=1 and βi = β. Now,



Example: Suppose we are building a linear model out of two features, we’ll have two coefficients (β1 and β2). For better understanding let β1 = 10 and β2 = 1000.

In lasso regression, the L1 penalty would look like,

**L1p = |β1| + |β2|**

#### Shrinking β1 to 8 and β2 to 100 would minimize the penalty to 108 from 1010, which means in this case the change is not so significant just by shrinking the larger quantity. So, in the case of the L1 penalty, both the coefficients have to be shrunk to extremely small values, in order to achieve regularization. And in this whole process, some coefficients may shrink to zero. **1** [Ref: URL for the above explanation in the foot note]

**Assumptions:**

##### There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.

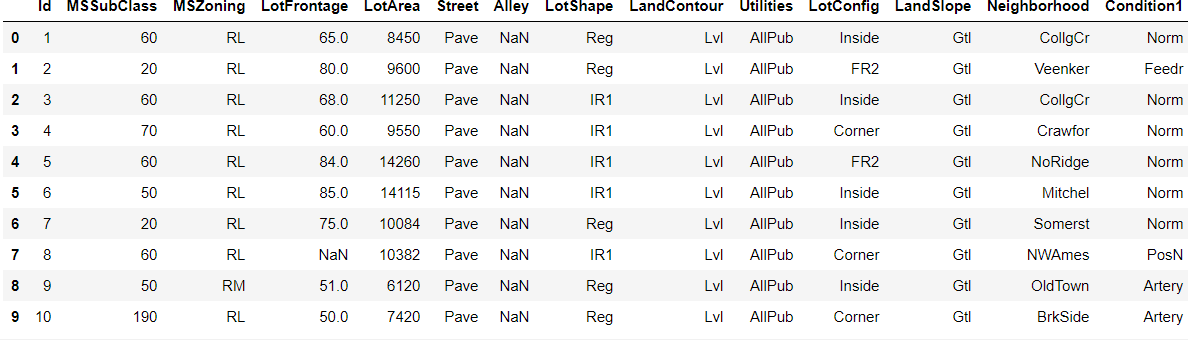
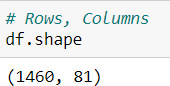
##### **Normality**: For any fixed value of X, Y is normally distributed.

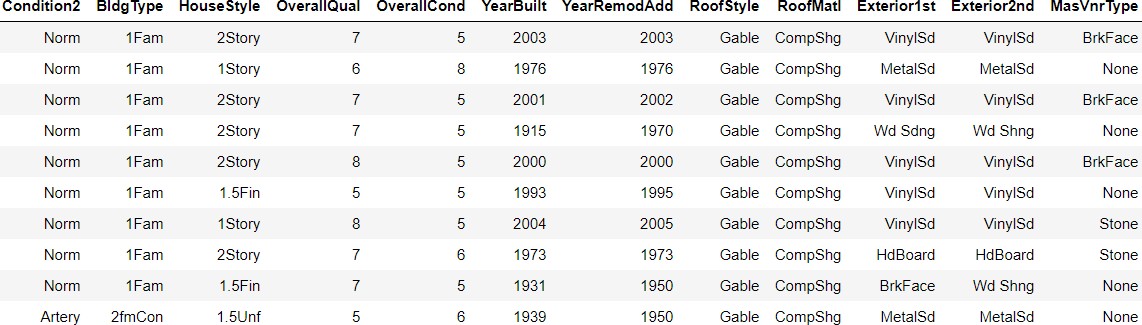
1 https:[//w](http://www.analyticsvidhya.com/blog/2020/11/lasso-regression-causes-sparsity-while-ridge-regression-)w[w.analyticsvidhya.com/blog/2020/11/lasso-regression-causes-sparsity-while-ridge-regression-](http://www.analyticsvidhya.com/blog/2020/11/lasso-regression-causes-sparsity-while-ridge-regression-) doesnt-unfolding-the-math/

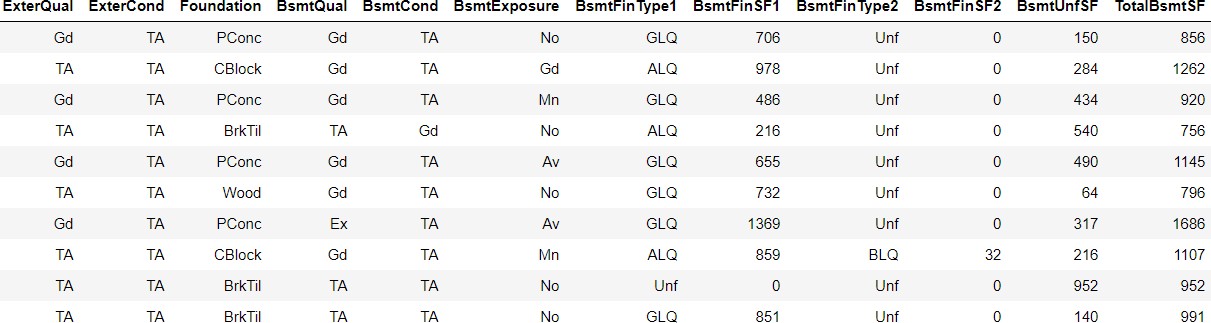
Data Sources and their formats

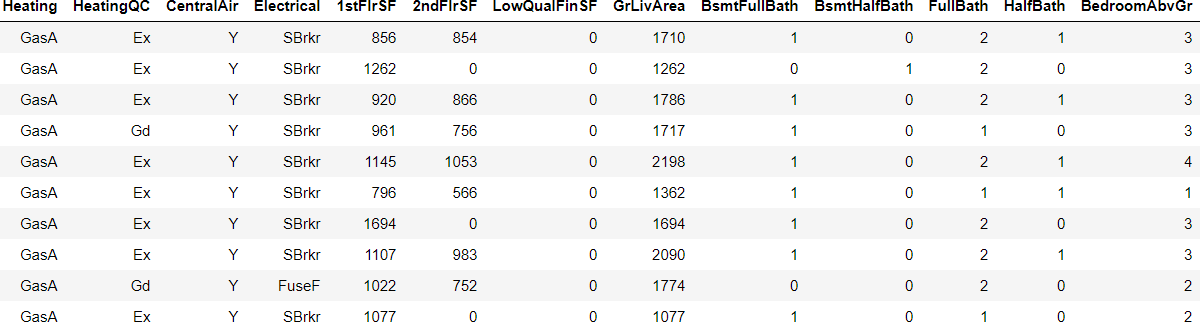
A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file

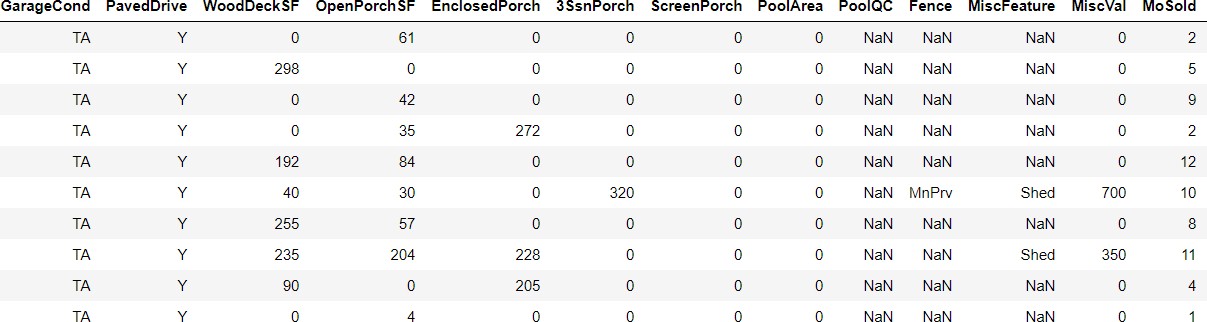
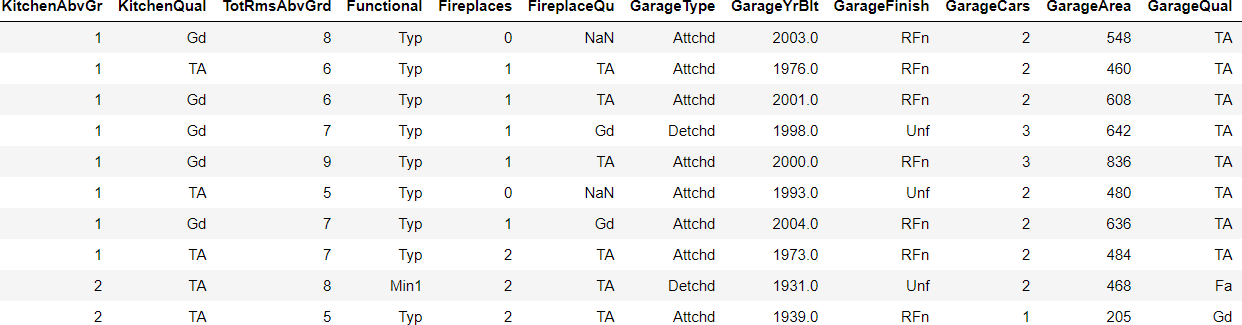
Here's how the top 10 rows of the data looks like:





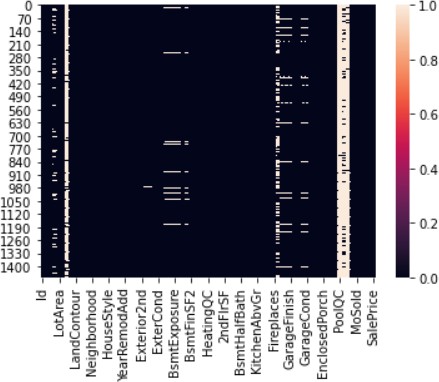






The last Feature: SalePrice is the target variable. The above Snapshots show all the features and the top 10 rows. As mentioned earlier, there are 1460 rows and 81 columns.

## Data Preprocessing



The above heatmap shows there are many Null Values, which can’t be processed. One Observation here is that a lot of variables have been labelled at NaN, but they are actually not null values and have certain meaning.

For Example,

 NA in feature 'Alley' means No\_Alley

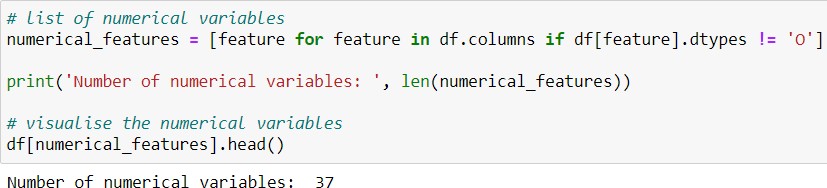
 in case of PoolQC, NA means 'No Pool' (\* Refer Data Description at the end of the notebook)

I've replaced them with actual variables before going further.

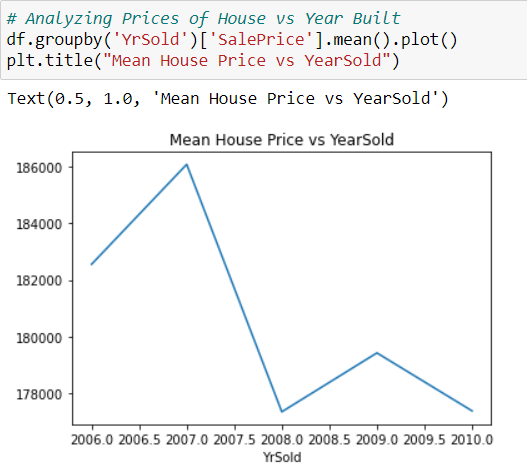
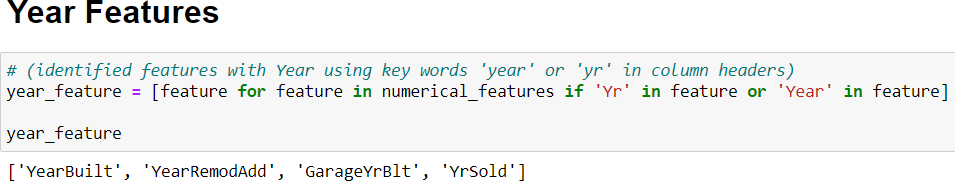
First let us handle Categorical features which are missing; based on domain knowledge and given explanation. The percentage of Null values in Categorical features:

Alley: 0.9377% missing values MasVnrType: 0.0055% missing values BsmtQual: 0.0253% missing values BsmtCond: 0.0253% missing values BsmtExposure: 0.026% missing values BsmtFinType1: 0.0253% missing values BsmtFinType2: 0.026% missing values FireplaceQu: 0.4726% missing values GarageType: 0.0555% missing values GarageFinish: 0.0555% missing values GarageQual: 0.0555% missing values GarageCond: 0.0555% missing values PoolQC: 0.9952% missing values Fence: 0.8075% missing values MiscFeature: 0.963% missing values

Then I replaced all other categorical missing values with a new label 'Missing'. The numerical missing values will be imputed during feature engineering.

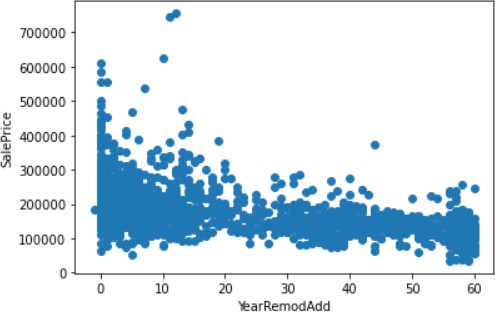
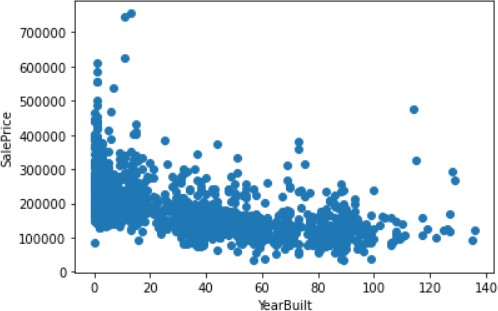
**Numerical variables**

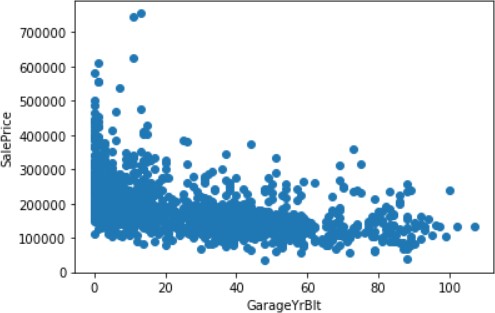
Identified all features that were numerical



There seems to be a peak in House Prices, but a sharp drop in between 2007 to 2008. This can be due to Economic Crash. "Economies worldwide slowed during this period since credit tightened and international trade declined. Housing markets suffered and unemployment soared, resulting in evictions and foreclosures."

# Let's see the scatterplot between All years features with SalePrice



 Obs 1: The Houses built recently have Higher Sales Price.

 Obs 2: The Houses remodelled recently have Higher Sales Price.

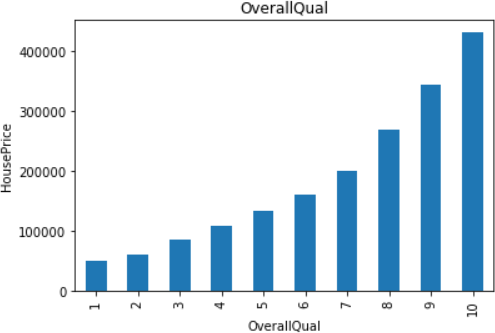
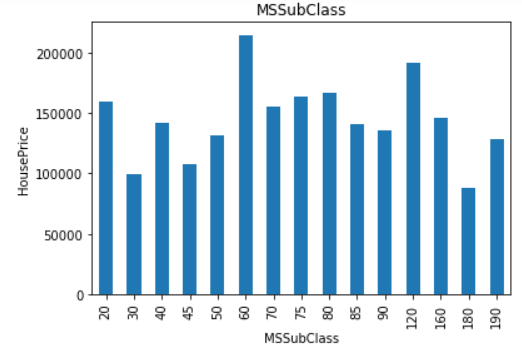
 Obs 3: The Houses whose Garages were built recently have Higher Sales Price.

**Identifying Discrete Variables**

The following 17 features were identified as discrete variables:

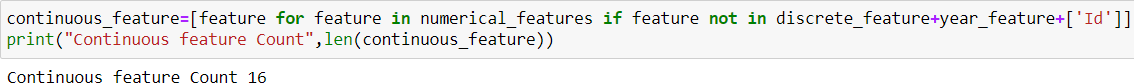
['MSSubClass', 'OverallQual', 'OverallCond', 'LowQualFinSF', 'BsmtFullB ath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'BedroomAbvGr', 'KitchenA bvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageCars', '3SsnPorch', 'PoolAr ea', 'MiscVal', 'MoSold']

Plotted Bar Plots like these to understand relations with Sale Price

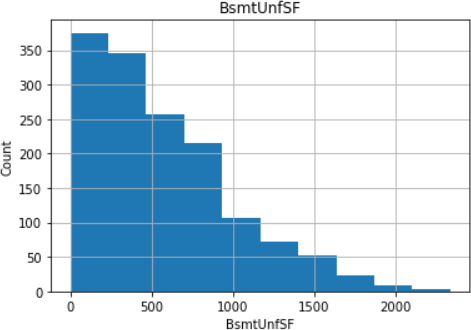
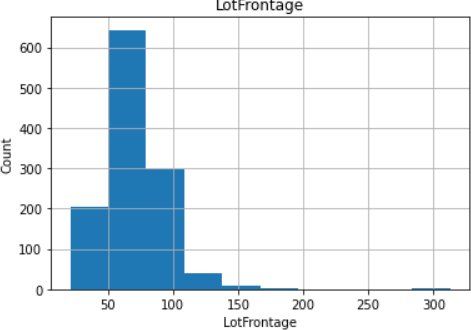


Similarly, plotted for all discrete values, and observed features.

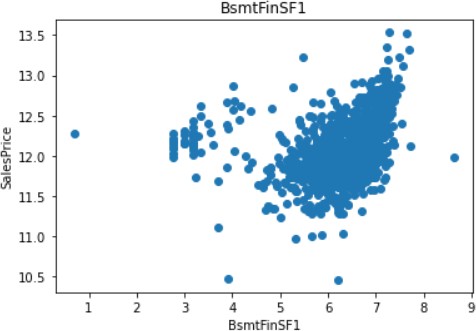
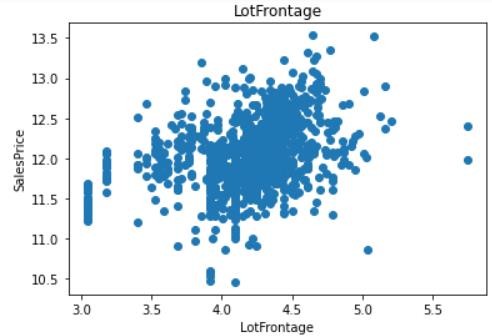
**Identifying Continuous Features**



I’ve plotted Histograms for all 16 features like the following



# As clear from above a lot of features were not normally distributed. Let's I did log transformation, plotted the scatterplots to see the trends.



**Categorical Features**



Identified total unique categories in each feature:

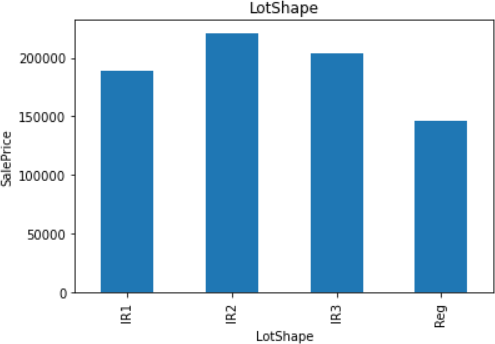
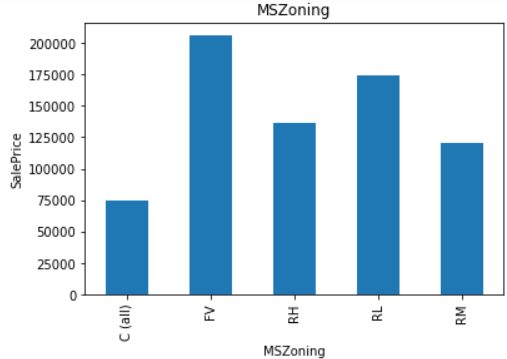
MSZoning has 5 categories Street has 2 categories Alley has 3 categories LotShape has 4 categories

LandContour has 4 categories Utilities has 2 categories LotConfig has 5 categories LandSlope has 3 categories Neighborhood has 25 categories Condition1 has 9 categories Condition2 has 8 categories BldgType has 5 categories HouseStyle has 8 categories RoofStyle has 6 categories RoofMatl has 8 categories Exterior1st has 15 categories Exterior2nd has 16 categories

MasVnrType has 5 categories ExterQual has 4 categories ExterCond has 5 categories Foundation has 6 categories BsmtQual has 5 categories BsmtCond has 5 categories BsmtExposure has 5 categories BsmtFinType1 has 7 categories BsmtFinType2 has 7 categories Heating has 6 categories HeatingQC has 5 categories CentralAir has 2 categories Electrical has 6 categories KitchenQual has 4 categories Functional has 7 categories FireplaceQu has 6 categories GarageType has 7 categories GarageFinish has 4 categories GarageQual has 6 categories GarageCond has 6 categories PavedDrive has 3 categories PoolQC has 4 categories

Fence has 5 categories MiscFeature has 5 categories SaleType has 9 categories SaleCondition has 6 categories

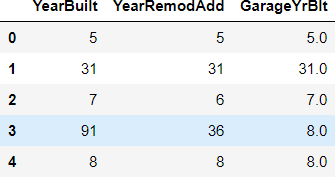
Plotted all Categorical variables vs SalesPrice as shown below

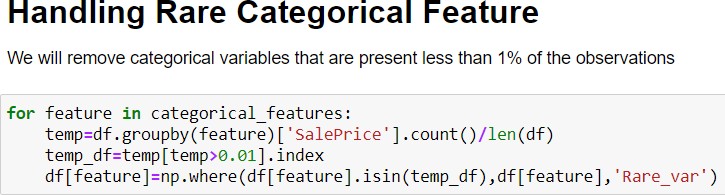


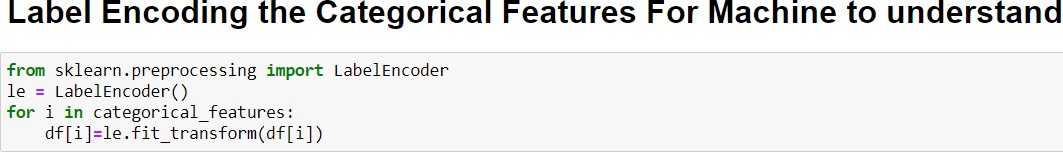
**Feature Engineering**

# I had already treated all Null Values in categorical Features, Now I will check for numerical variables. Imputed the numerical null values with medians.

# Now, as there were some features(Temporal) which contained year values. Differences:

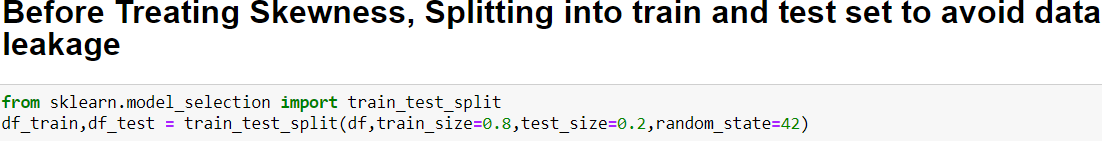




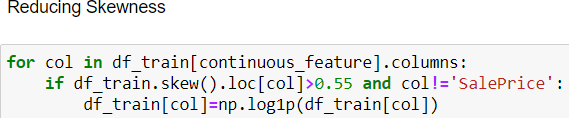


**Skewness in some Continuous Variables**

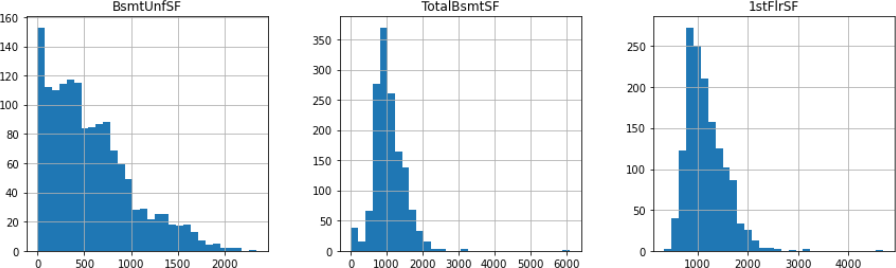
**There are a lot of skewed variables. I have treated them with log1 transformation.**



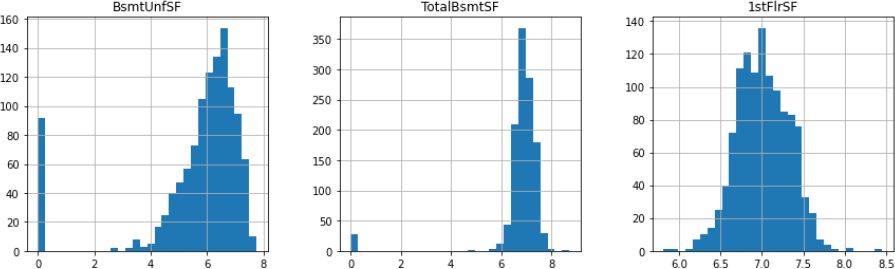
80% data will be used for training and 20% for Testing



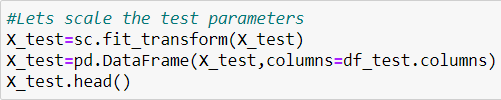
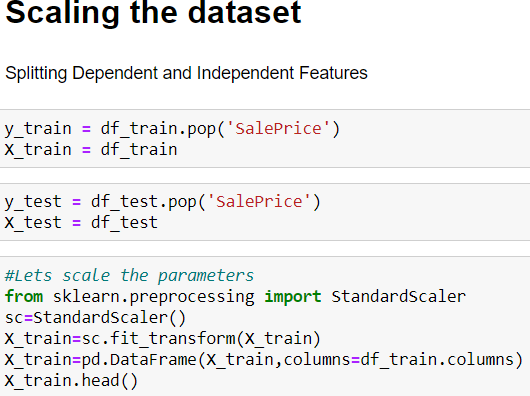
As seen in the below examples, I’ve treated all the features.



Before Treating for Skewness

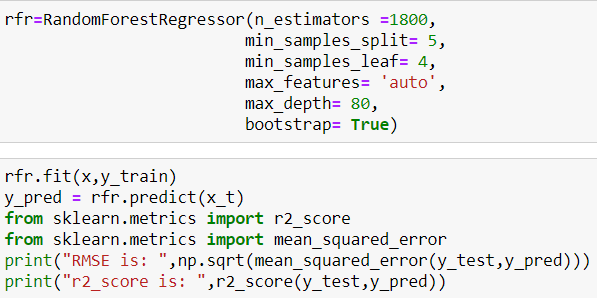
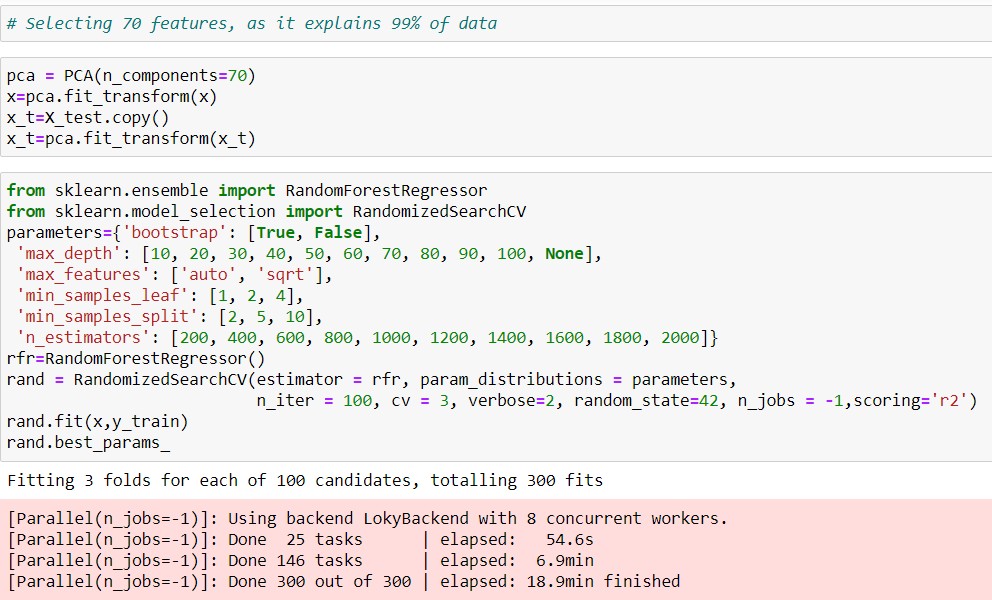


After Treating for Skewness



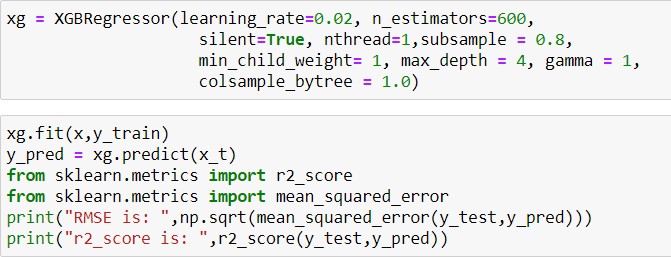
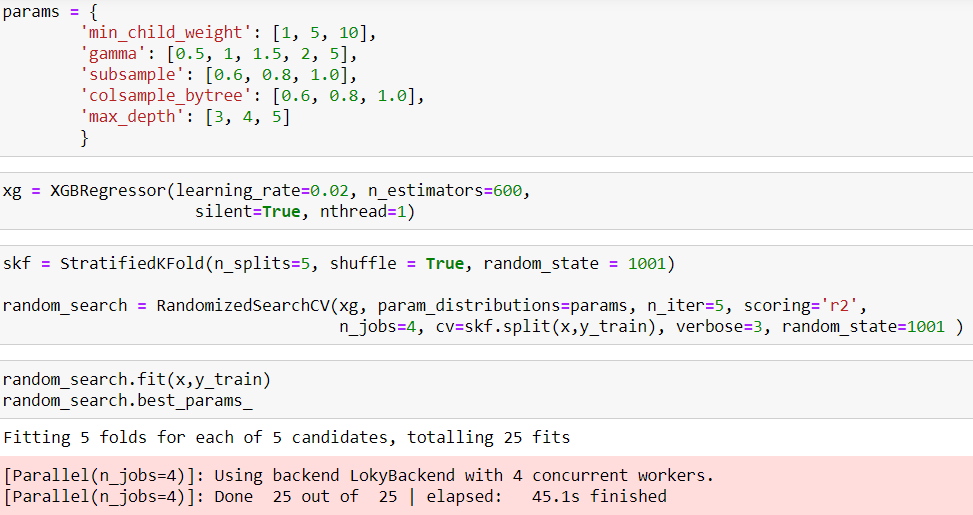
I’ve used Standard Scalar to make all the data comparable.

**Modelling**

1. **Random Forest Regressor with PCA**

**Results: Top 10 Features and R2 Score**



1. **XGBoost Regressor with PCA**

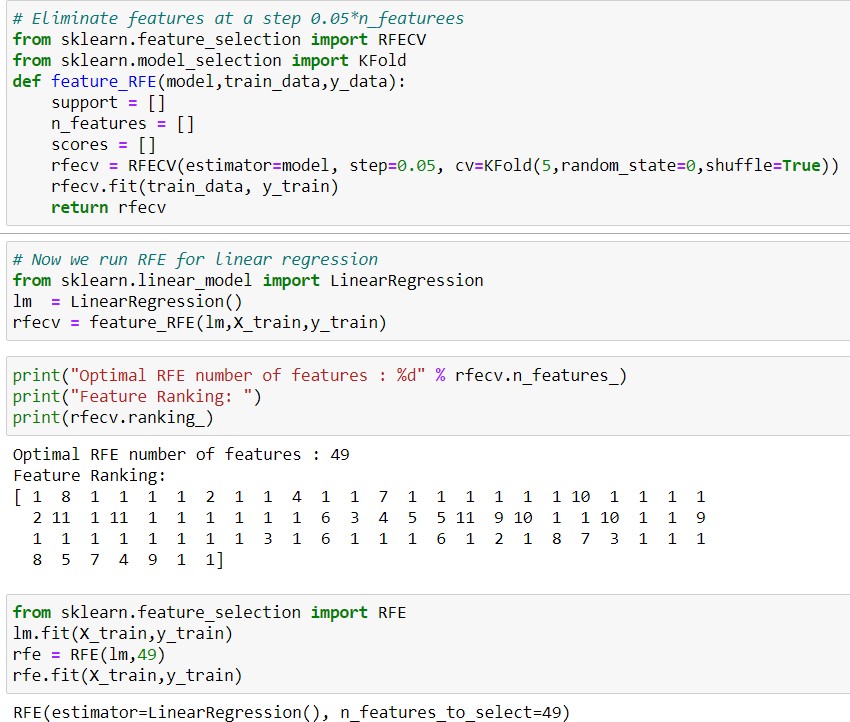
**Results:**

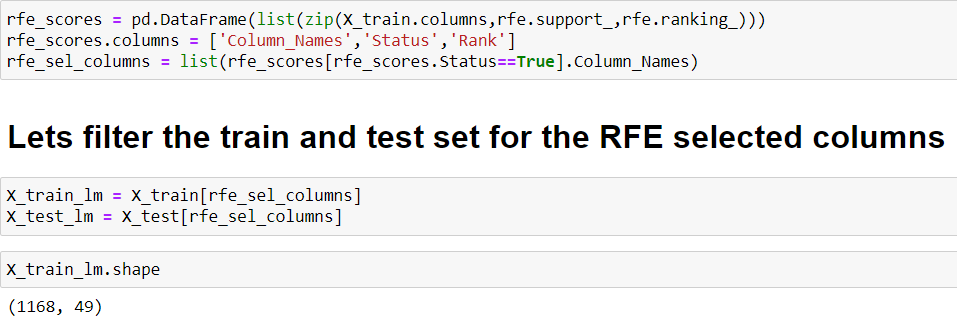


The score was way less than Random Forest, so I’ve rejected this model. Then I checked with the following models

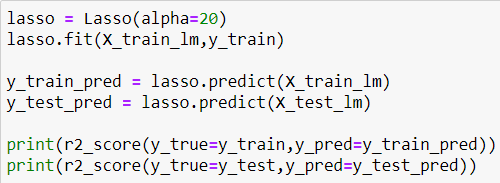
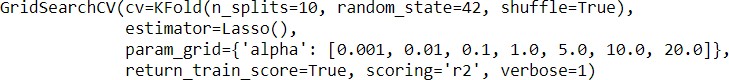
1. **Linear Regression with RFE**
   1. **Lasso b. Ridge**

**Preparing the Data by reducing features using RFE**



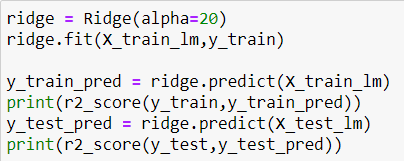
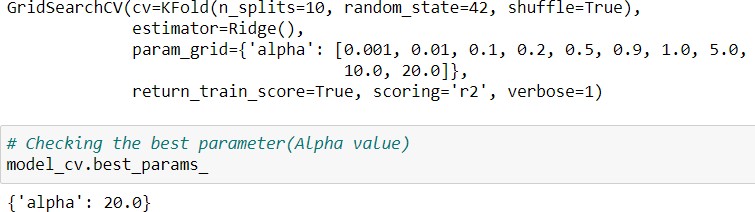






### R2 Scores for Train and Test Data





R2 Scores for Train and Test Data



Finally, after all the model testing, I’ve found Lasso Ridge to be the best

performing model. Building final Model.

Final Model

lasso = Lasso(alpha=20)

las so. lit ( X\_t ra1 n\_l m, y\_t *ra* in)

y train pred : lasso.predict(X train lm)

y te st p red = Las so . predlet (X te st La)

print(r2 score(y true:y train,y pred:y train pred)) print(r2 score(y true=y test,y pred=y test pred))

0.8413407167403752

0. 81154576 30494485

The R2 score is almost equal for both training and test data.

p ri nt ( "Rf‘\S E is : " , np . sqrt (mean squa red *e r* ror (y test,y *pred ) ) )*

RMSE is: 4Z960.40768938855

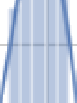
sns.distplot(y test-y test pred)

‹AxesSubplot:xlabel='SalePrice', ylabel='Density'>

1e—5

2.0

1.5



1.0

05

0.0

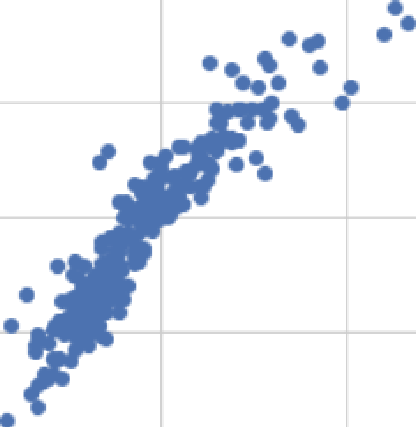
0 200000 400000

SalePrice



plt.scatter(y test,y test pred)

<matplotlib.collections.PathCollection at 0x1a95d102130›

400000 . • •

300000

200000

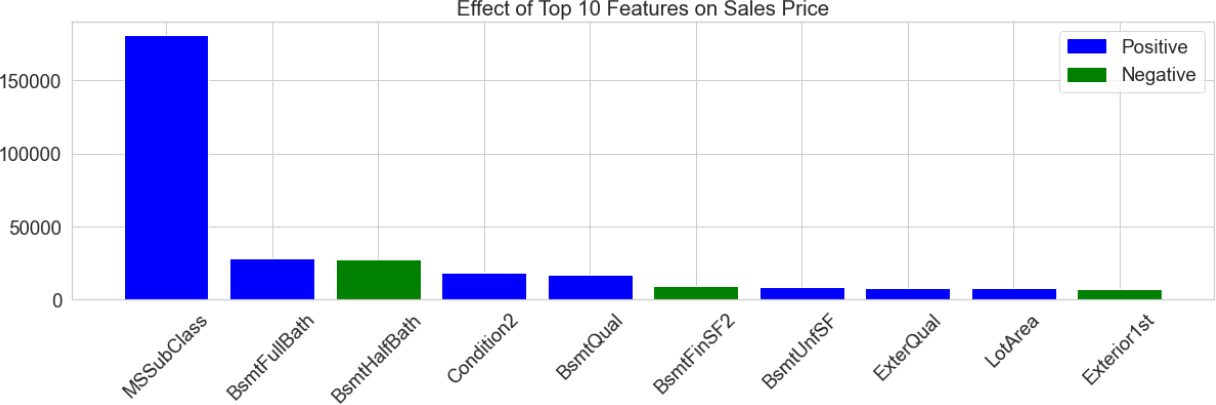
100000

0 200000 400000 600000



**Top 10 Features Based on effect on Sales Price of House**





**CONCLUSION**

* Key Findings and Conclusions of the Study:
  + MS Sub Class seems to have the biggest impact on House Prices, followed by Basement Full Bath and Basement Half Bath
  + Other than the Basement related features, Condition 2, Exterior Quality and Lot Area are some of the other important features.
* Learning Outcomes of the Study in respect of Data Science
  + Got to understand about the concept of Data Leakage. All transformation must be done after splitting the data to test and train, otherwise the parameters are affected.
  + Used RFE for the first time. It is a great technique for Feature Selection.
  + Learned about the usage of Lasso and Ridge Regression.