

**Housing Project**

Submitted by:

Abhinav Verma

**ACKNOWLEDGMENT**

I, would like to express my gratitude to my SME Sapna Verma , who gave me the opportunity to work on project “**Housing Price Prediction**” and also guiding me in gaining in-depth knowledge of Data Science and Machine Learning in deriving insights for Organizational goals and meeting Business needs.

References:

https://stackoverflow.com/

<https://scikit-learn.org/stable/>

<https://seaborn.pydata.org>

**INTRODUCTION**

• **Business Problem Framing**

The main objective of this project is to model the price of houses with the available independent variables. This model will 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 needs 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.

• **Technical Requirements**

© Data contains 1460 entries each having 81 variables.

© Data contains Null values. We 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. We need to handle them accordingly.

© We have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.

© We need to find important features which affect the price positively or negatively.

© Two datasets are being provided to us (test.csv, train.csv).

**• Motivation for the Problem Undertaken**

1. The objective behind to take this project is to harness the required data science skills.

2. Improve the analytical thinking.

3. Get into the real-world problem-solving mechanics.

**Analytical Problem Framing**

**•** **Mathematical/ Analytical Modelling of the Problem**

This is a Regression problem, where our end goal is to predict the Prices of House based on given data provided in the dataset. We have divided the provided dataset 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 our model has performed. We are using Mean Absolute Error, Root Mean Square Error, and ‘R2\_Score’ to determine the best model among,

• Linear Regression

• Lasso

• Decision Tree Regression

• K Neighbors Regression

• Random Forest Regression

• SVR

• Ridge

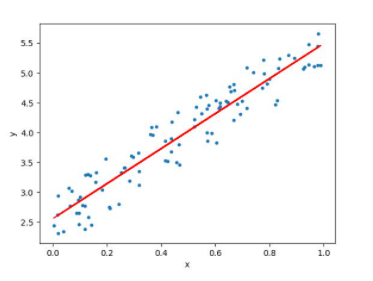
• AdaBoostRegressor

The best results were obtained using Lasso Regression. So, let’s discuss a little bit 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://www.datasciencecentral.com/>)

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.

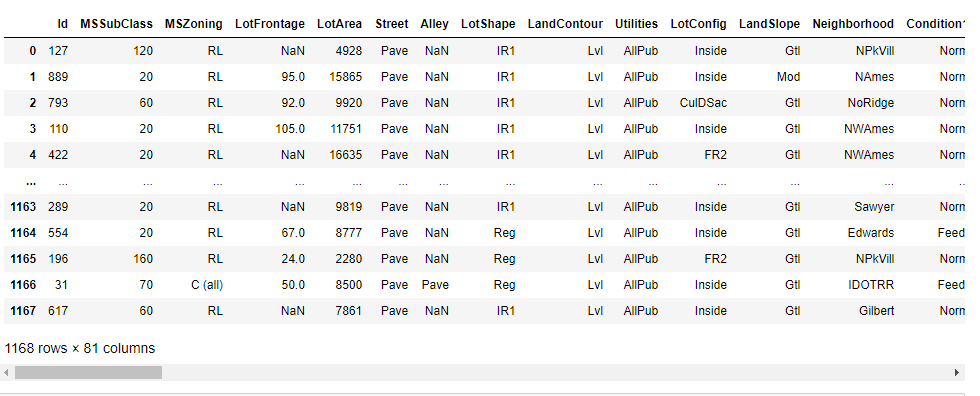
**• Data Sources and their formats**

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytic 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.

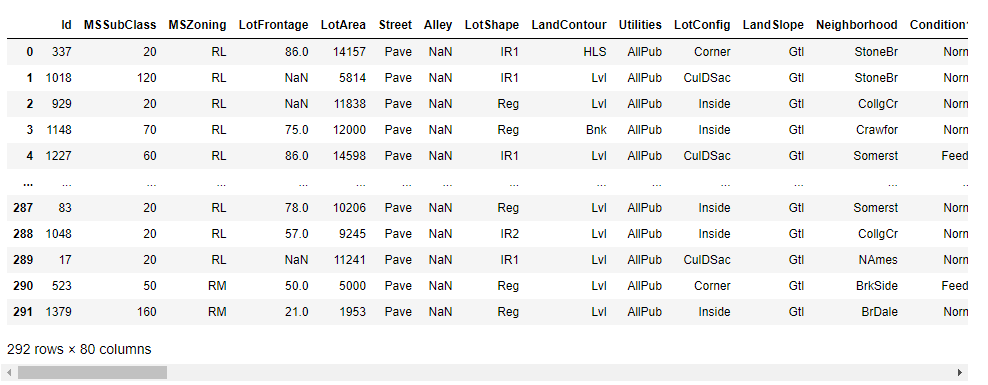
The dataset contains 1460 rows and 81 columns (including the train dataset and test dataset).

The sample dataset are:

**Train data.**



**Test Data**

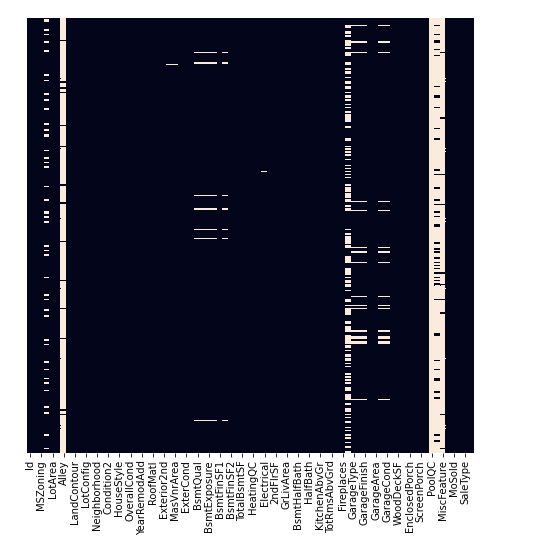


• The column ‘SalePrice’ is the target column. We need to predict the sale price of the houses.

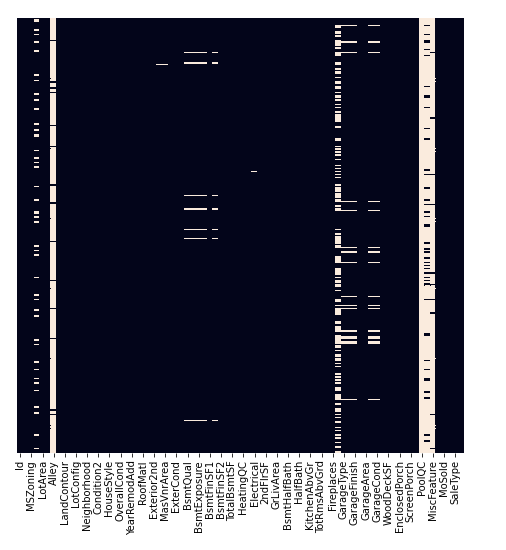
**• Data Pre-processing Done**

As our dataset contains null values (missing values) so we have replaced the missing values with the required values. Details are mentioned below:

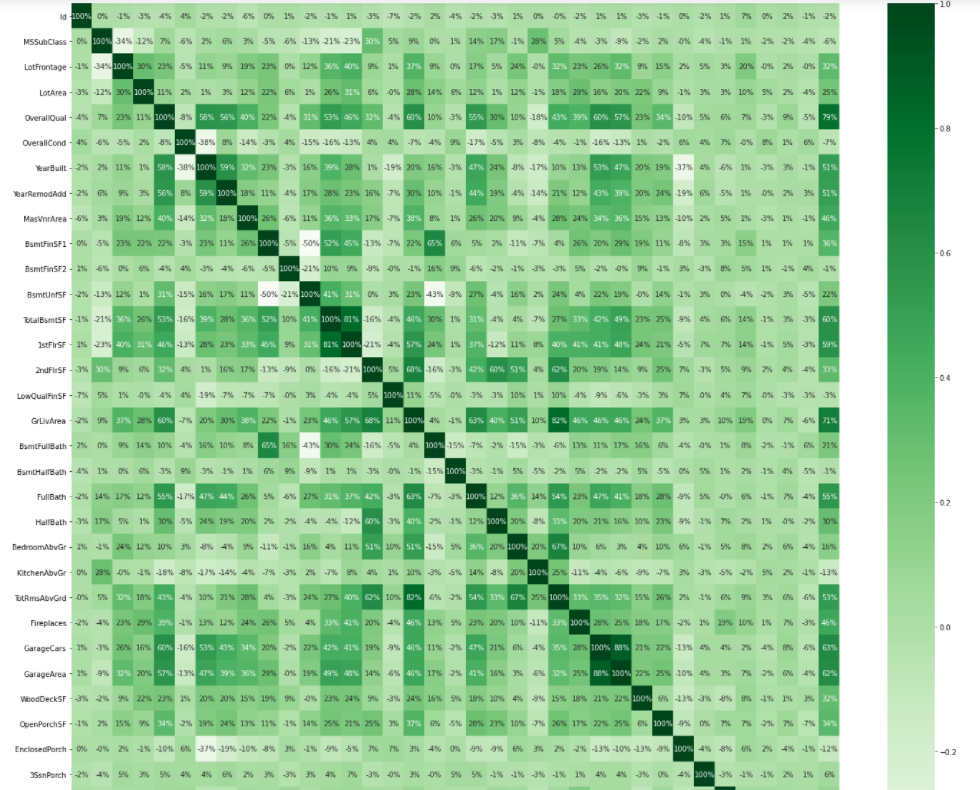
**Train data**



**Test data**



**Checking the Correlation by heat map.**



Here we can see that:

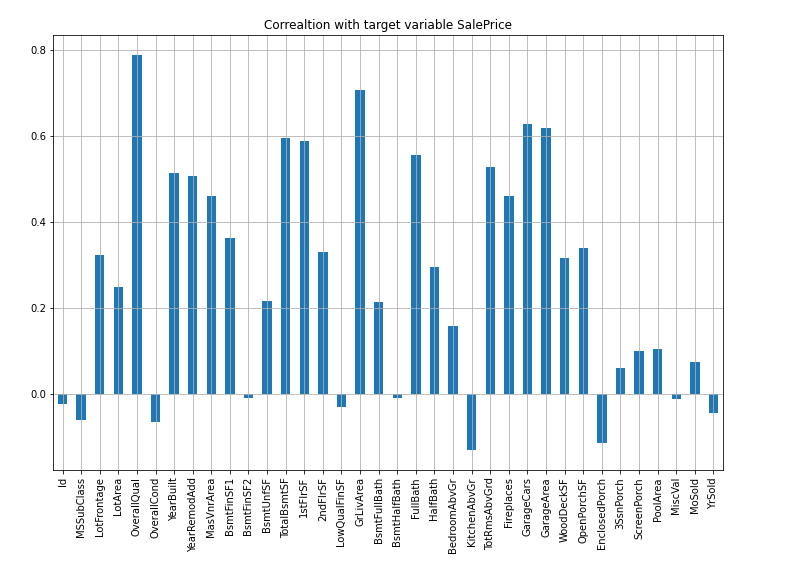
1- from the above plot we can say that SalesPrice is highly positively correlated with the columns OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, 1stFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, GarageCars, GarageArea.

2- from the above plot we can say that SalePrice is negatively correlated with OverallCond, KitchenAbvGr, Encloseporch, YrSold.

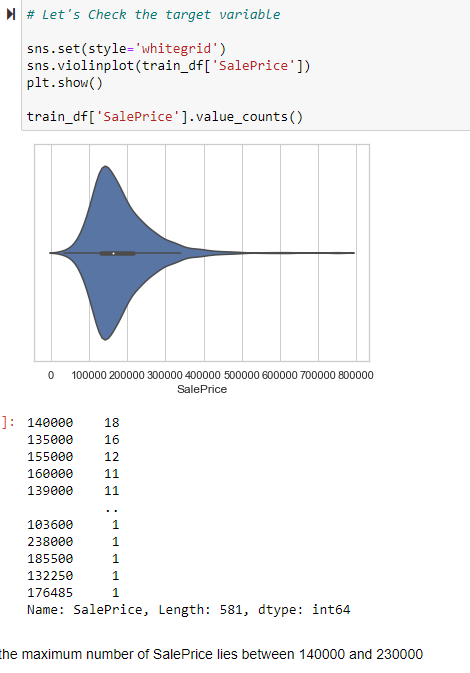
3- from the above plot we can observe that multicollinearity in between columns so we will be using Principal Component Analysis(PCA).

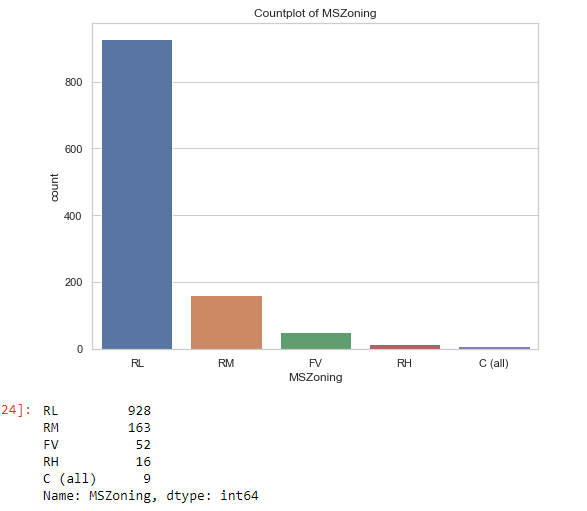
4- from the above plot we can say that No correlation has been observed between the column Id and other columns so we will be dropping this column.

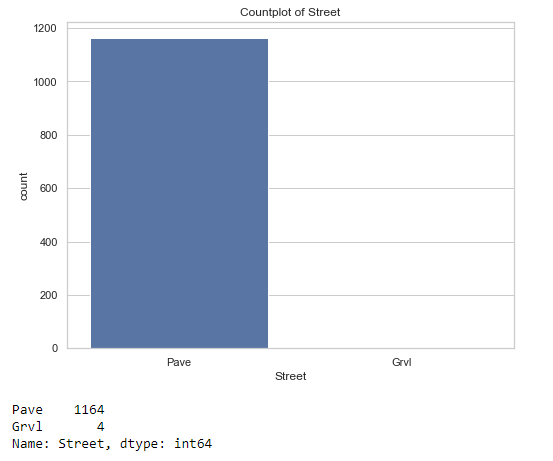
**• Correlation with target variable 'SalePrice'**

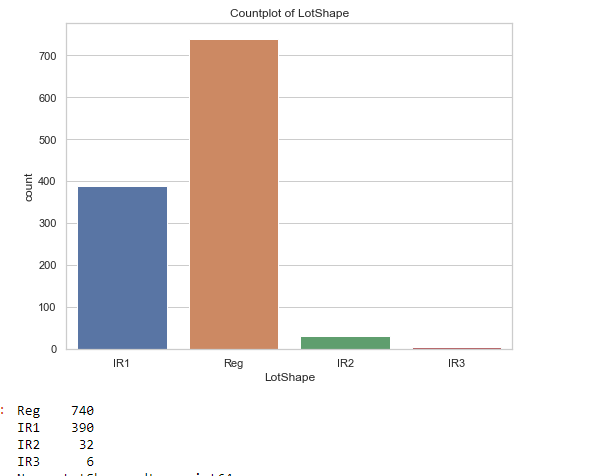


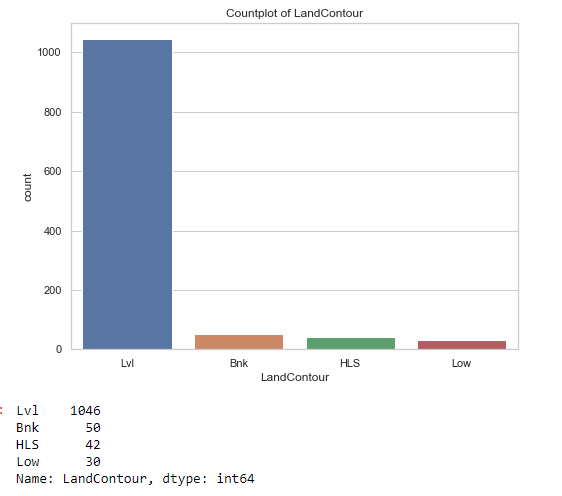
from the above plot we can say that the column OverallQual is most positively correlated with SalePrice and the column KitchenAbvGrd is most negatively correlated with SalePrice.

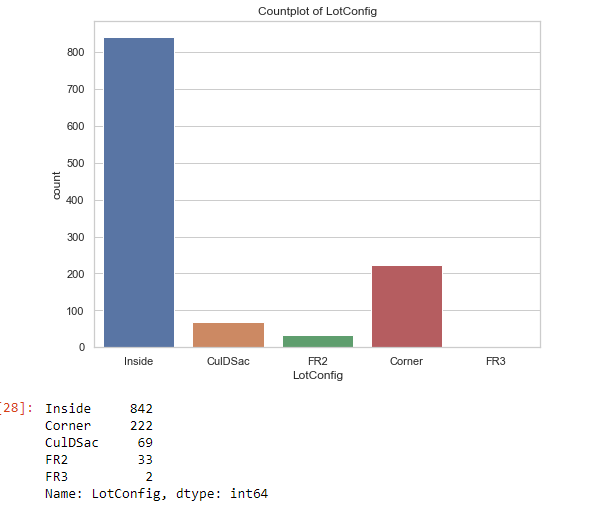


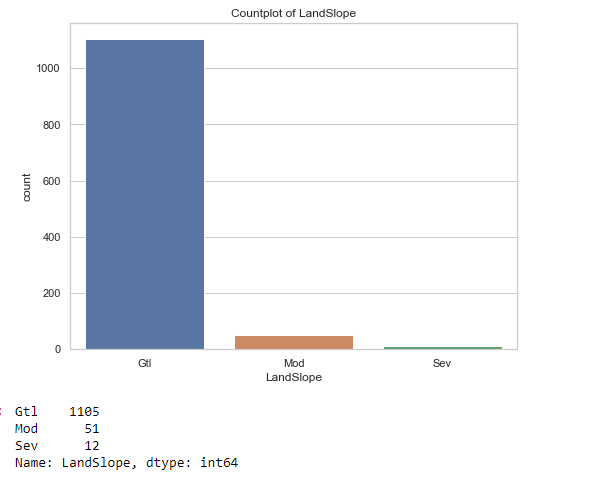


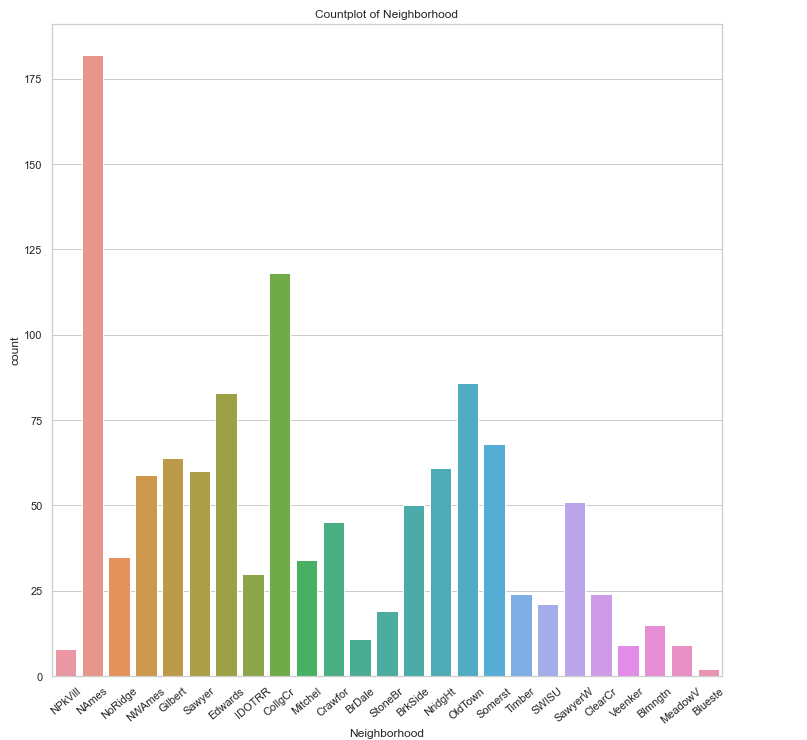






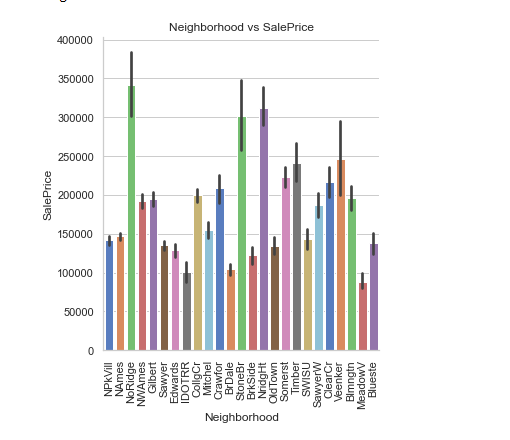


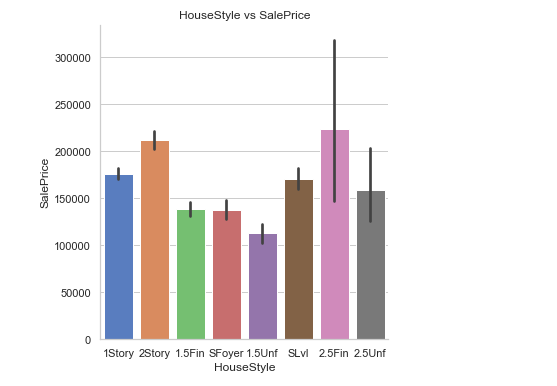


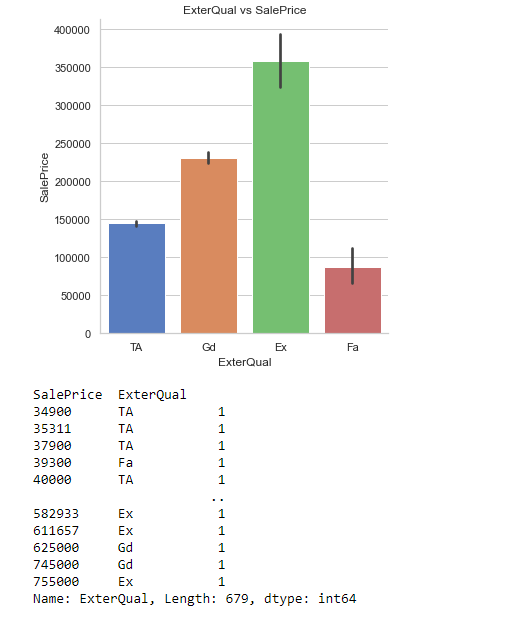


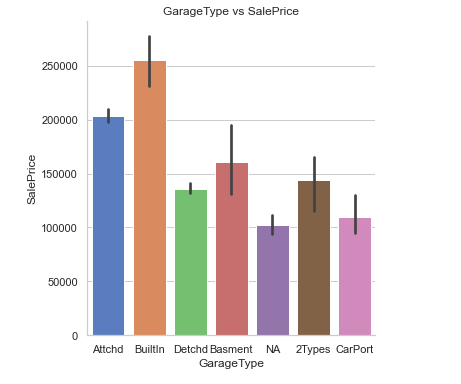


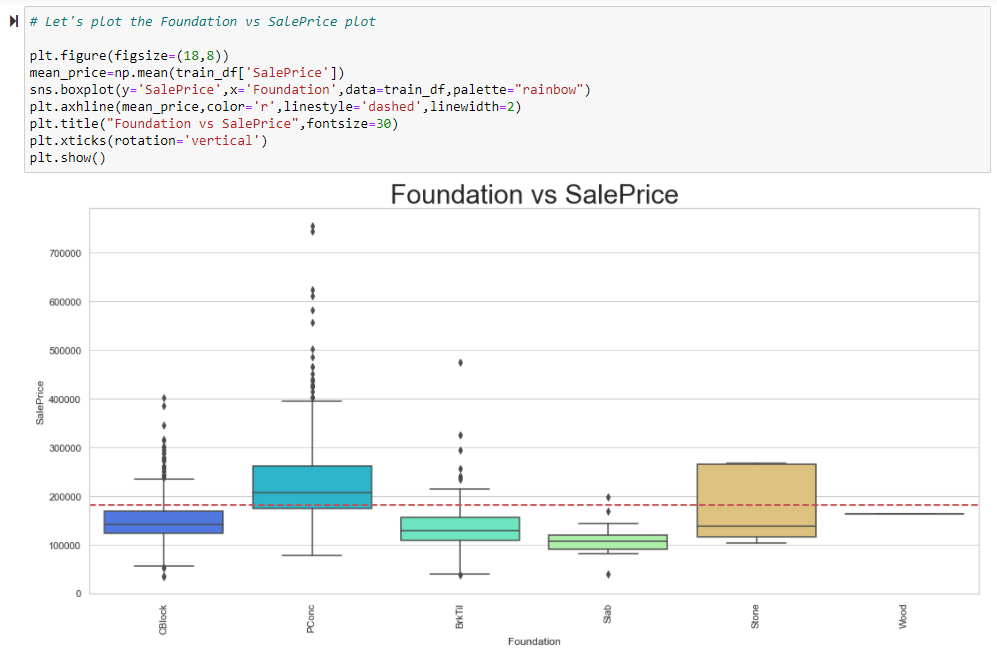
Plotting Factor plot of Neighborhood vs SalePrice.

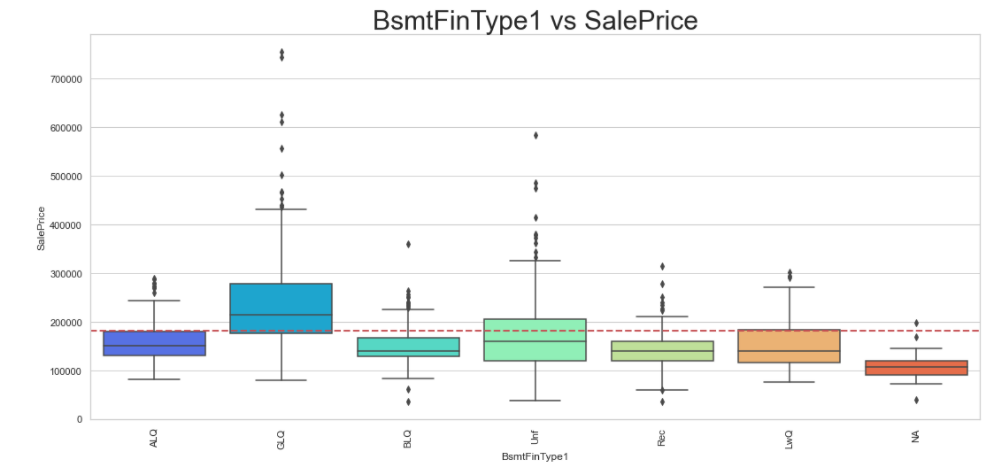


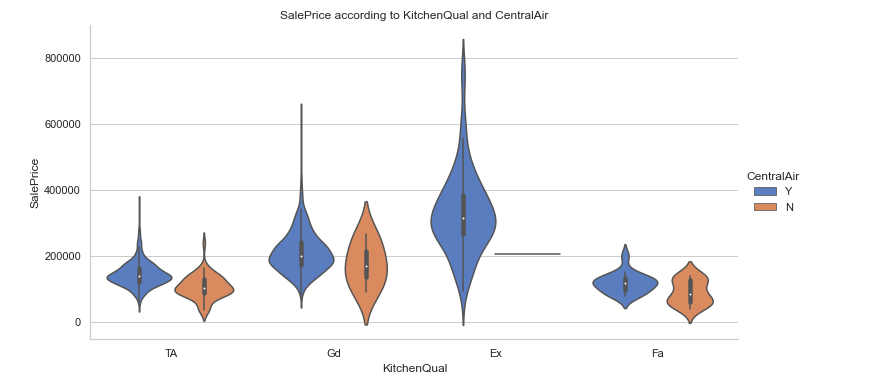




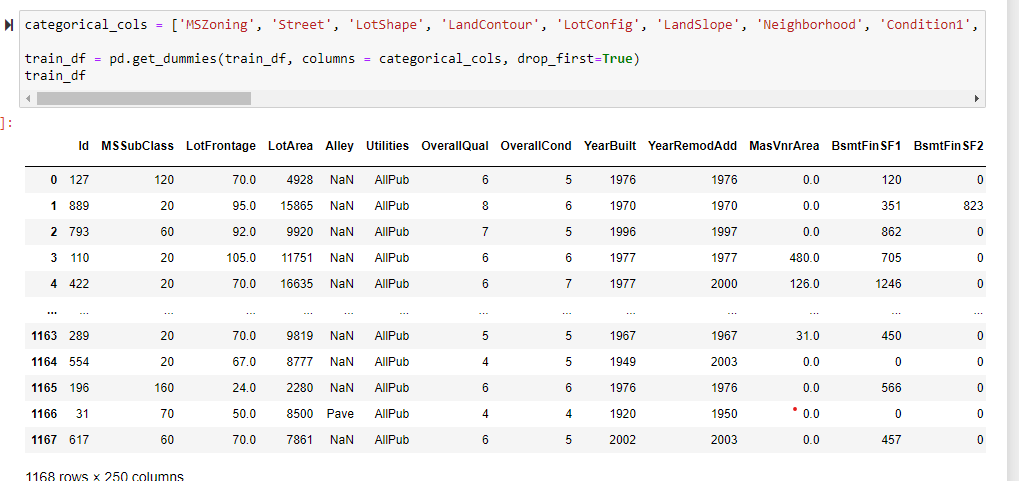




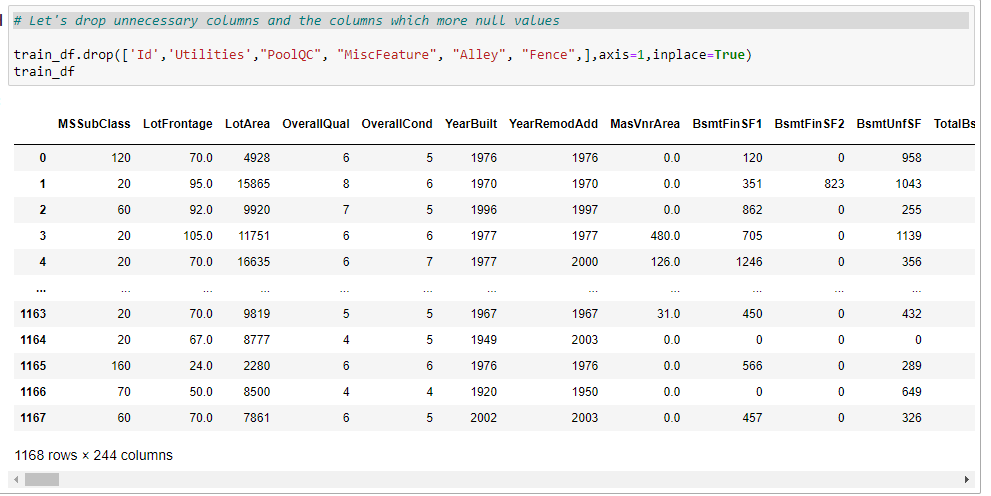




Now we will make dummies for categorical columns.



Let's drop unnecessary columns and the columns which more null values.



**• Hardware and Software Requirements and Tools Used**

Hardware Configuration:

Operating System: Windows 10 System Type: 64-bit operating system, x64-based processor Processor: Intel® Core™ i3-5005U @ 2.00 GHz 2.00 GHz RAM: 4GB

Software & Tools:

a) Jupyter Notebook (used as a notebook to code)

b) Python (used for scientific computation)

c) Pandas (used for scientific computation)

d) Numpy (used for scientific computation)

e) Matplotlib (used for visualization)

f) Seaborn (used for visualization) g) Scikit-learn (used as algorithmic libraries)

**Model/s Development and Evaluation**

• Identification of possible problem-solving approaches (methods)

• Performed EDA (Exploratory Data Analysis).

• Data Cleaning and dropping the columns which were not contributing to the dataset.

• Handled the missing values.

• Checked for the outliers and tried to remove the outliers of the dataset.

• Checked for the skewness in the dataset and removed the skewness for better model building.

• Train- Test the dataset into independent and dependent variables. • Model Building.

**• Testing of Identified Approaches (Algorithms)**

Below are the algorithms used for the training and testing:

1) Linear Regression.

2) Lasso

3) Decision Tree Regression.

4) K Neighbour Regression.

5) Random Forest Regression.

6) DecisionTreeRegressor

7) SVR

8) Ridge

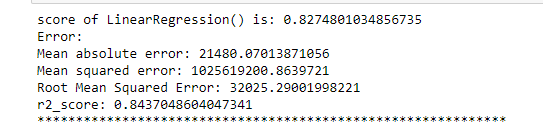
9) ElasticNet

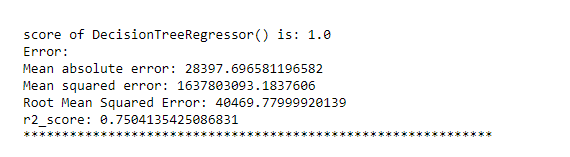
10) AdaBoostRegressor(),

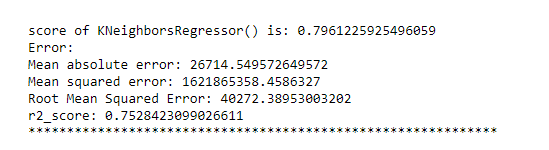
11) GradientBoostingRegressor()

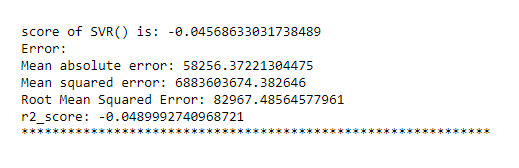
**•** Run and Evaluate selected models

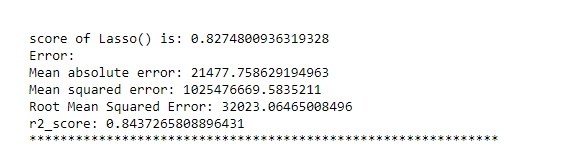


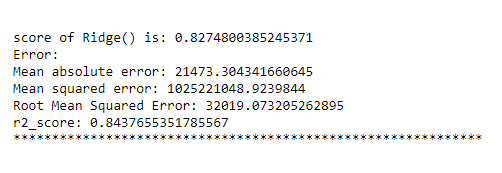


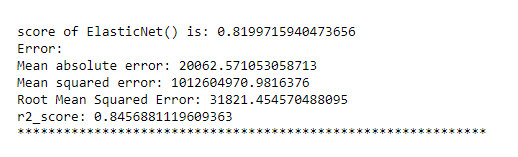


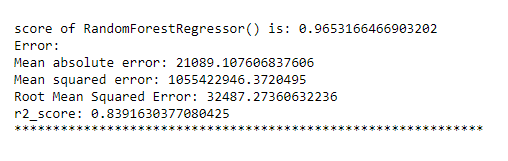


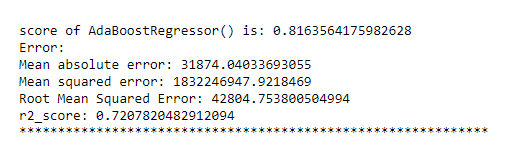


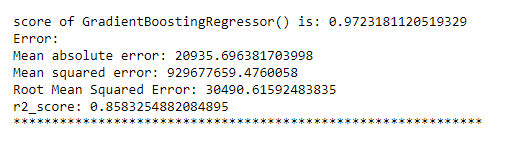




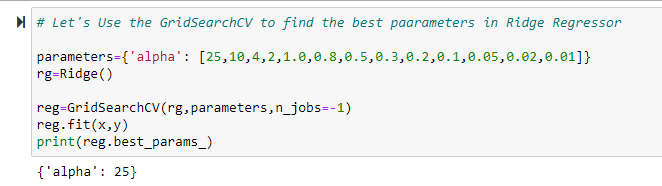








Finding the best parameter in Ridge Regressor.



**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.

