

**HOUSING PRICE PREDICTION**



Submitted by:

DEEPTHI MALLU

**ACKNOWLEDGMENT**

I take this opportunity to acknowledge everyone who have helped me in every stage of this project.

Firstly, I am indebtedly grateful to my SME MR. Sajid Chowdary sir, who helped me from beginning of my Projects. Am also thankful to my Mentor Shankar Gowda Sir and my whole Data Trained team, where I have learnt Analysing the datasets and building the models using Machine learning and making the projects. Finally, am so thankful to my Flip Robo Technologies team, as they provided me the opportunity to work as intern in their company.

I feel pleasure, to make project report on “Housing Price Prediction”. It has been my privilege to have a team of project guide who have assisted me from the commencement of this project. The project is a result of my hard work, and determination put on by me with the help of Wikipedia, You Tube videos of Krish Naik, skikit-learn.org and referred some old projects on Kaggle.com.

**INTRODUCTION**

**Business Problem Framing**

Houses are one of the basic needs of every individual around the world, housing and real estate markets are the major contributors of world’s economy. House prices increase every year, so there is a need for a system to predict house prices. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

There are three factors that influence the price of a house which include physical conditions, concept and location. So, Machine Learning models are used to predict the actual value of the prospective properties and decide whether to invest in them or not and draw insights from the given variables and know which variables contribute much for our prediction.

**Conceptual Background of the Domain Problem**

Real estate is a very large market which has a huge demand, where selling and buying of houses takes place. These days buying a house involves huge amounts of money to be invested. 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 which is US-based housing company named Surprise Housing who 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.

Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results.

So, Machine Learning is used in order to predict the actual value of the prospective properties and decide whether to invest in them or not and draw insights from the given variables and know which variables contribute much for our prediction.

**Review of Literature**

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted on given explanatory variables that cover many aspects of residential houses.

As continuous house prices, they will be predicted with various regression techniques including Linear Regression, KNeighbors Regression, Decision Tree Regression, Lasso, Ridge, SVM regression, Extreme gradient boost Regression and Random Forest regression which are able to accurately estimate the price of the house based on the given features.

**Motivation for the Problem Undertaken**

Buying a home is a big decision, every individual before purchasing a house they look into many factors one among is price. The price of the houses will be will be incremented every year eventually. So, my motive to make this project is to predict the price of the house and make others know how much they can invest on buying.

So, Prediction of house prices are expected to help people who plan to buy a house based on their meeting standards. So, they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. The main objective of this project is to build machine learning models to get insights from the given data and build a model in terms of performance and reduced value of error.

**Analytical Problem Framing**

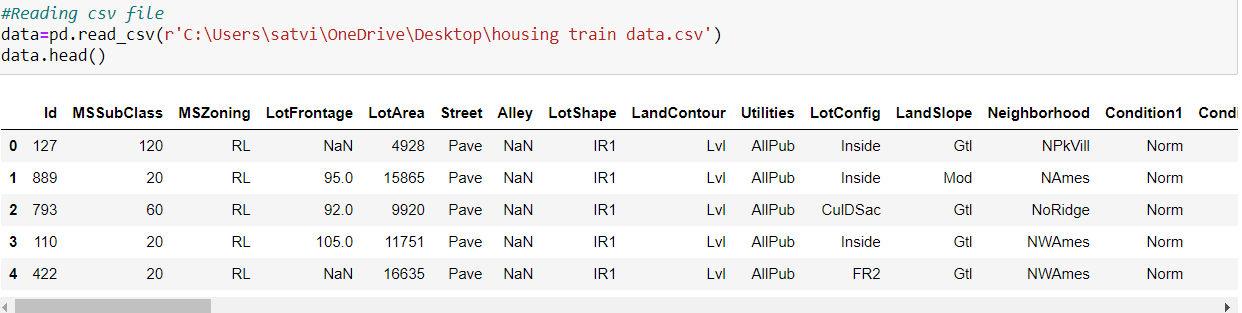
**Mathematical/ Analytical Modelling of the Problem**

A CSV file has given by a US-based housing company named Surprise Housing who decided to enter the Australian market. The company wants to use 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 dataset consists of different columns which we need to use and predict the label.

Our dataset consists of 1460 rows and 81 columns in both train data and test data. We are provided by two datasets of test.csv file, train.csv file. On which we will train on train.csv dataset and predict on test.csv file. I have checked for shape of the data then the train dataset consists of 1168 rows and 81 columns, the test dataset consists of 292 rows and 80 columns.

The data contains Null values which need to be treated using imputing techniques in this I have used mean and mode methods to fill the null values for continuous and categorical values respectively. The Data contains numerical as well as categorical variable which need to be treated using encoding techniques, I have used label encoder to encode the values.

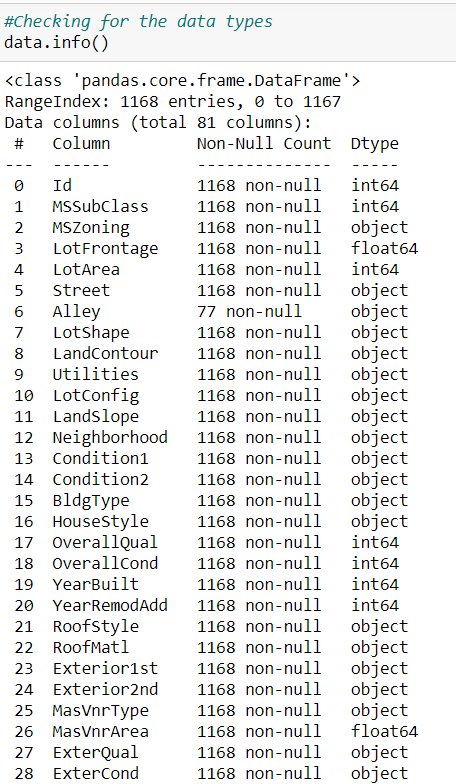
In order to find the relation between the features and label, I have used correlation matrix to check the correlation. Since our price prediction is a continous variable it comes under regression problem so I have used various regression algorithms for prediction of our label and tried achevieing the good performance of the model.

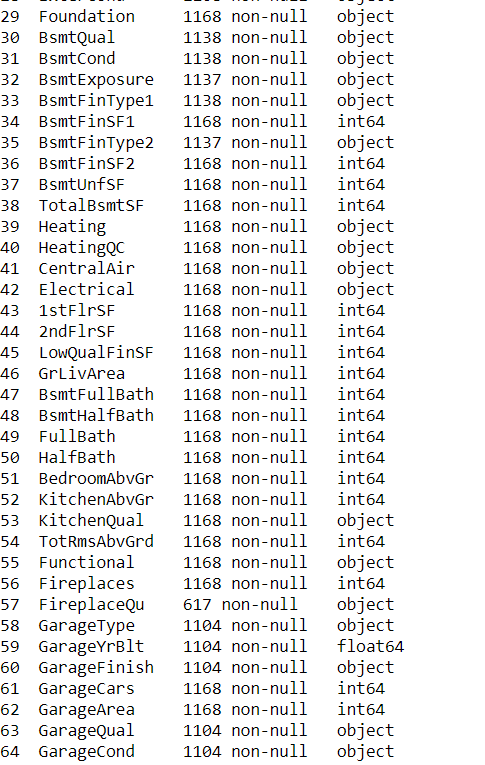


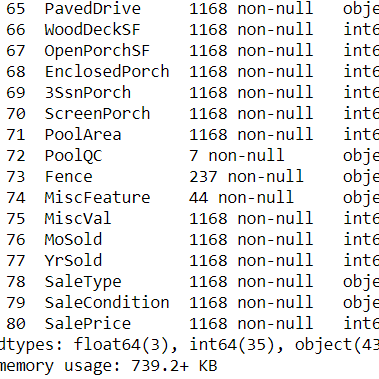
**Data Sources and their formats**

By using pandas, I have first imported the csv file and it consists of different columns which includes data in it. Our dataset consists of Features and label. After importing I have checked for shape of the dataset and which consists of rows and columns. Then I checked for null values and need to be treated and then I checked for info () method for knowing the type of the data then I checked for stats using describe method.

Our label is price prediction which is a continuous variable based on the values of independent variables our dependent variable depends.

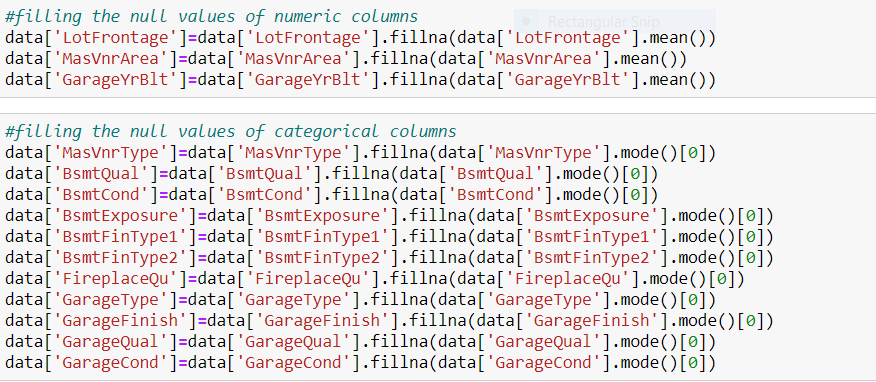






**Data Pre-processing Done**

There are many null values present in the dataset by using some of the imputing techniques we need to fill the NaN values using mean or model based on the type of the data in the columns. For the categorical variables I have use mode method to fill and for continuous columns I used mean to fill the NaN’s.



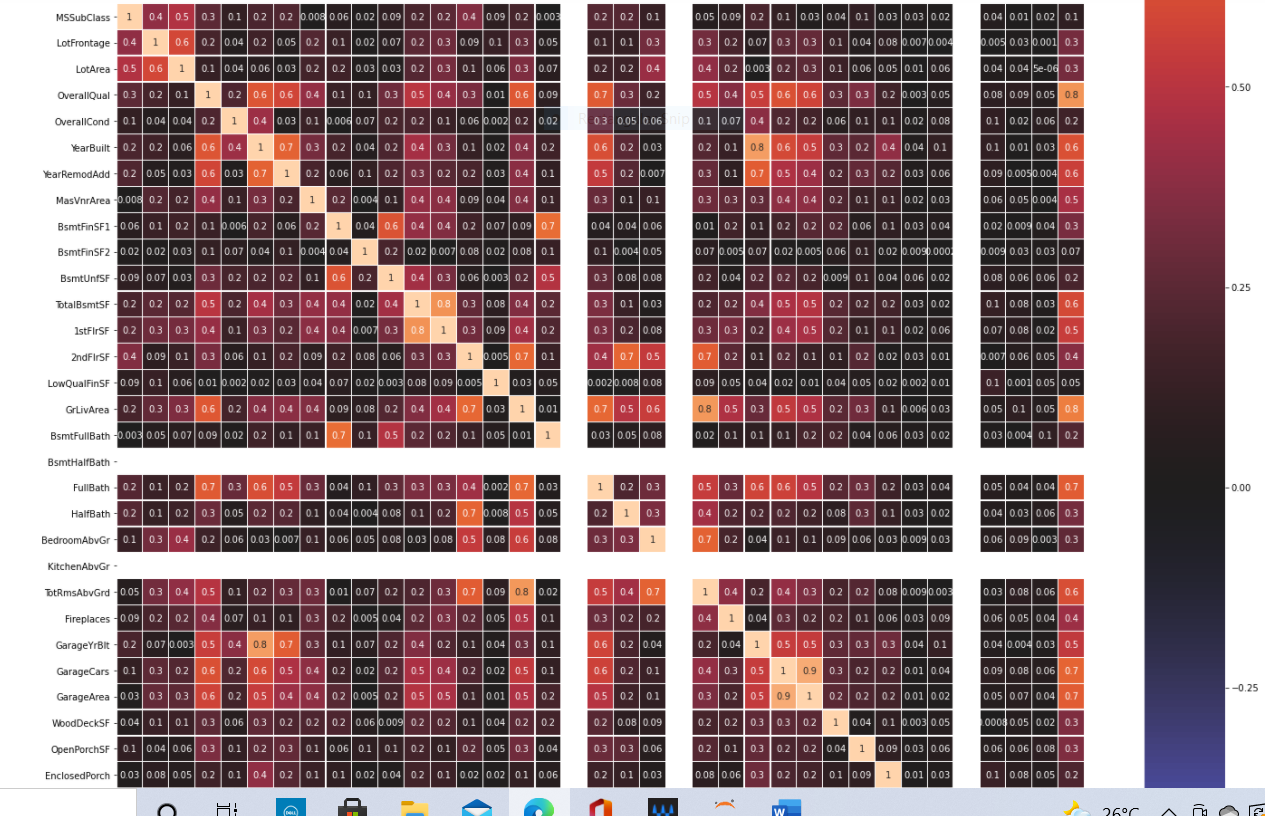
**Data Inputs- Logic- Output Relationships**

For checking the relation between the columns, I have used correlation matrix to find the relation and plotted heat map to visualise the percentage of the correlation. The below are the observations from the heatmap.

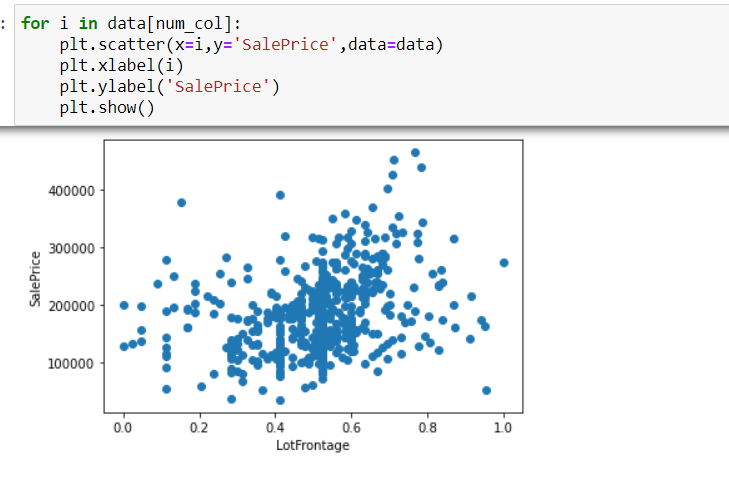
1.In the heat map the lightly coloured are high correlated and the dark coloured are less correlated.

2.Our label is highly correlated with overallqual, yearbuilt, yearremodass, masvnarea, etc. and very less correlated with screenporch.

3.There is also multi collinearity exists with some of the features between yearbuilt and garageyearbuilt, gargecars and garage area and some more

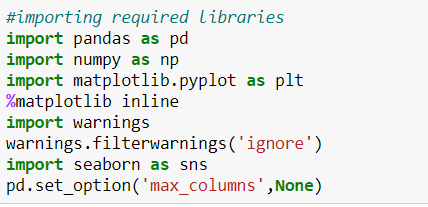


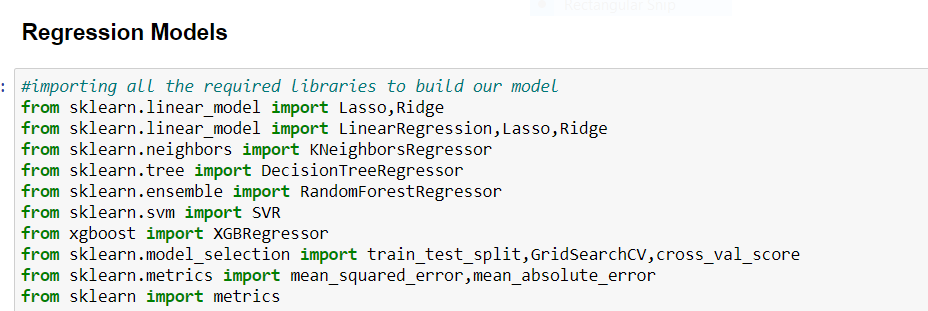
I have plotted scatter plots and checked the correlation between the features and the label.



**Hardware and Software Requirements and Tools Used**

I have used my laptop, Jupiter Notebook which is having GUI interface. Imported necessary libraries from python such as pandas, NumPy, seaborn, matplotlib, then imported the required model libraries from Scikit learn to import our algorithms.

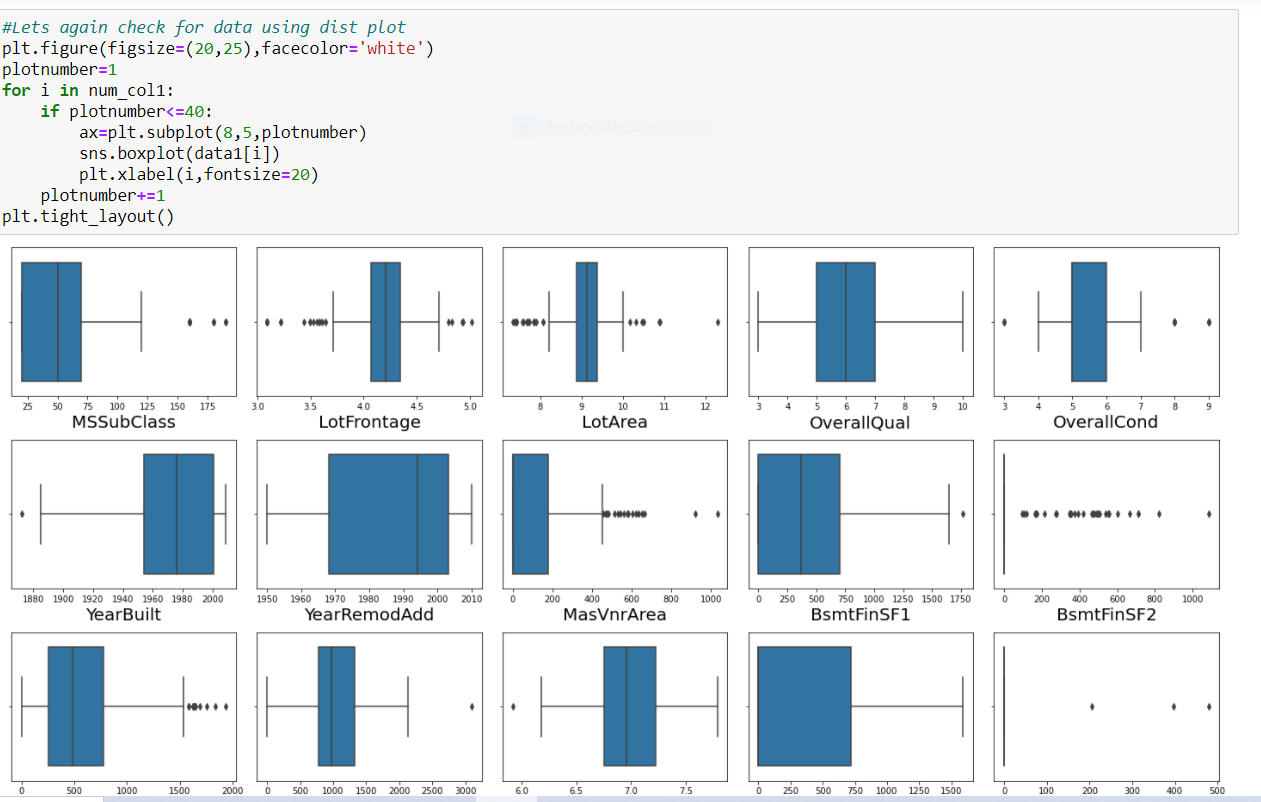




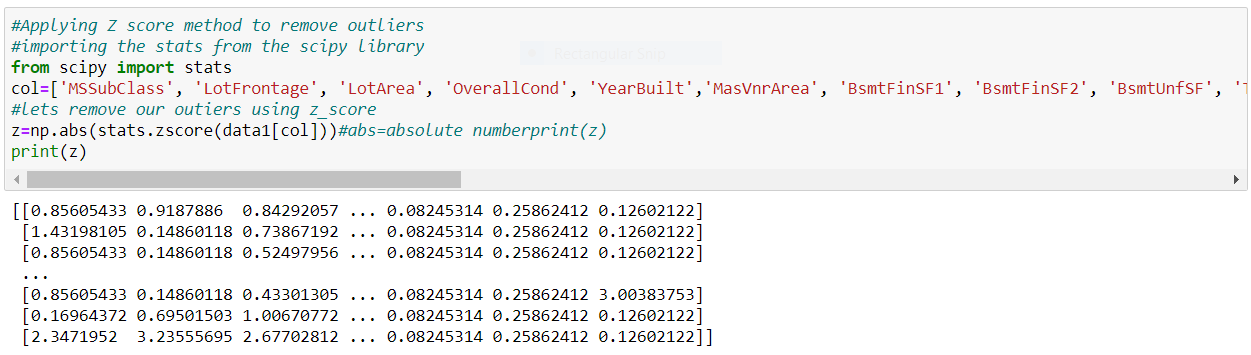
**Model/s Development and Evaluation**

I have plotted box plots to check for outliers and distribution plots to check the skewness so I found the presence of outliers and skewness in the continuous data. So, I have used Z-Score method for removing the outliers and log transformation method the remove the skewness.

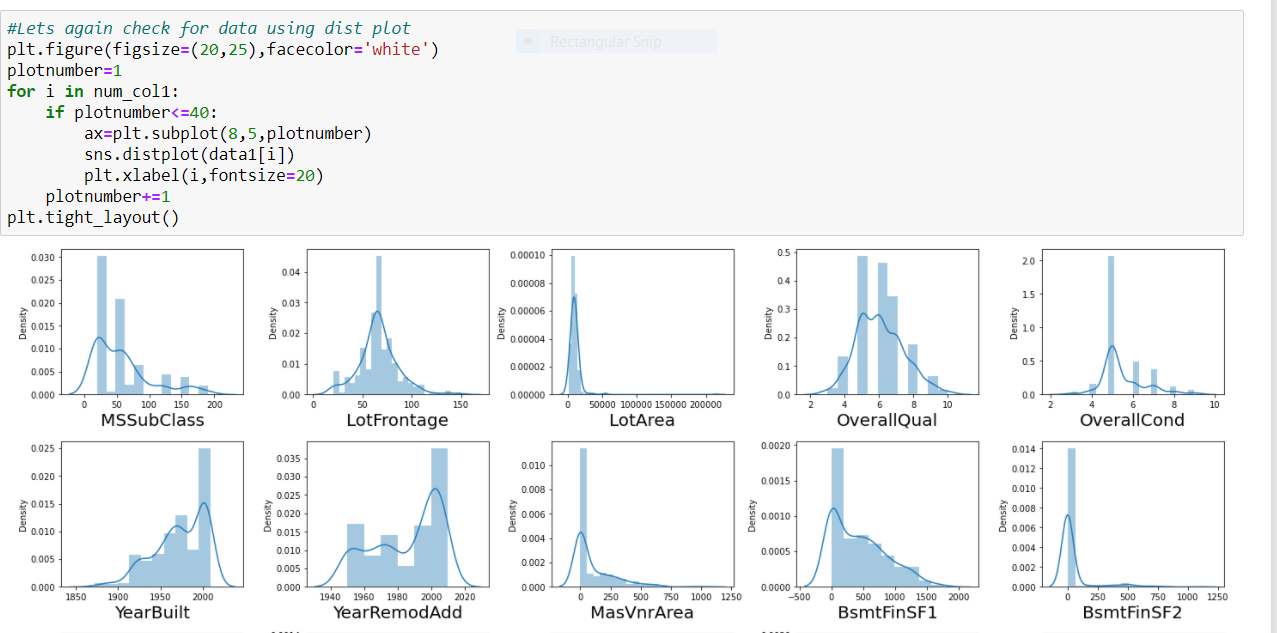
**Box Plots**:



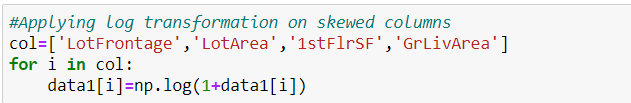
**Z-Score**



**Distribution Plots** are used to check the flow of the data in the columns.



**Log Transformation** is used to remove the skewness from the columns

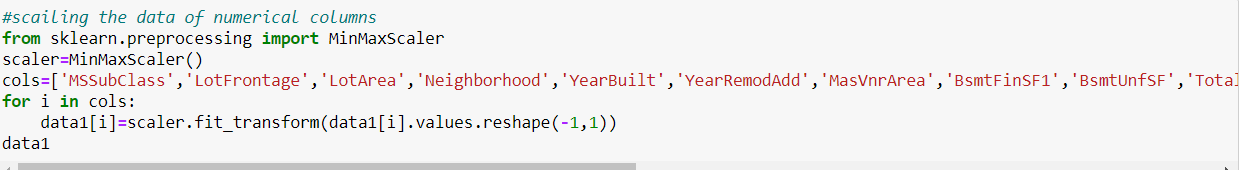


**Testing of Identified Approaches (Algorithms)**

The price of the house prediction is a numerical variable so it comes under regression problem, So I have used 8 different algorithms to check the model patterns. In order to final a model we have to check on different evaluation metrics like finding the score of the training data and testing data and finding the errors like mean absolute error (MAE), mean squared error (MSE) and Root mean squared error (RMSE). In order to tell a model is good their RMSE value should be as less as possible then we can say the model is efficiently working on the given data.

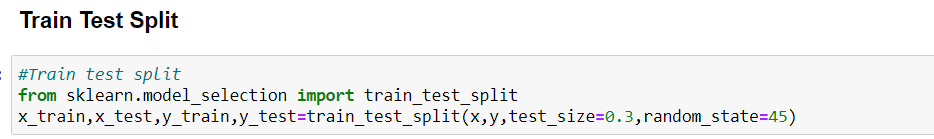
**Scaling the data**

I have used Min Max Scaler to scale the data**.**



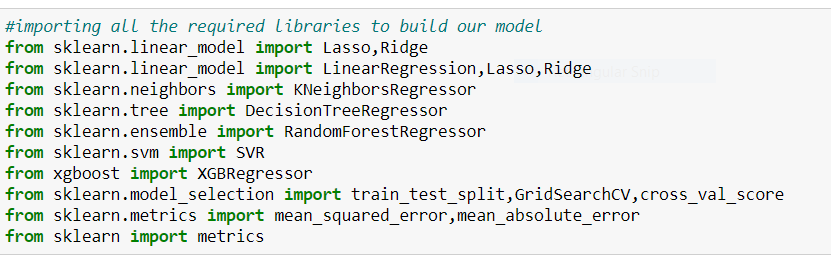
**Train Test Split**

I have imported the train\_test\_split from the module sklearn from model\_selection. And used 70% of the data for training and 30% of the data for testing and splitted the data into x-train, x-test, y-train, y-test.



**Regression Algorithms used for our prediction**

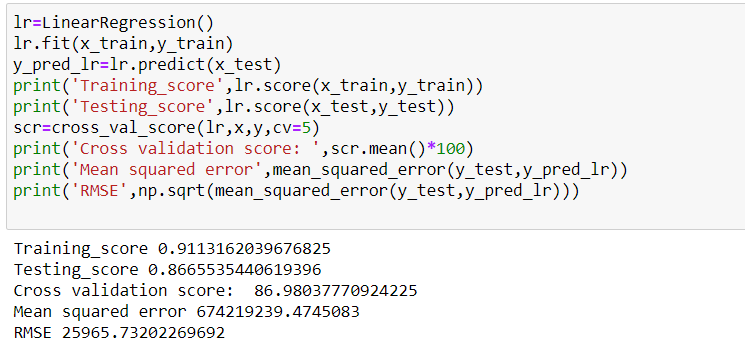
Since our price prediction is a continuous variable then this comes under Regression problem so I have used different Regression algorithms for predicting our label.

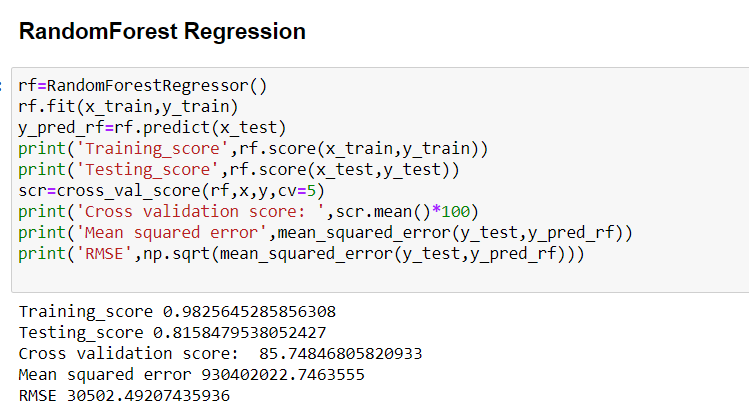


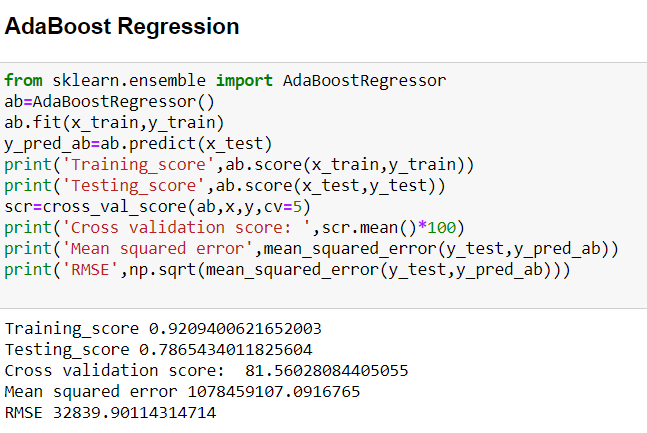
**Run and Evaluate selected models**

I have used various algorithms for predicting our label like Linear Regression, KNeighbors Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Extreme Gradient Boost Regression, Lasso Regression, Ridge Regression. For evaluating the model I have used Mean Squared Error (MSE), training score, testing score and root mean squared root (RMSE).

**Linear Regression**

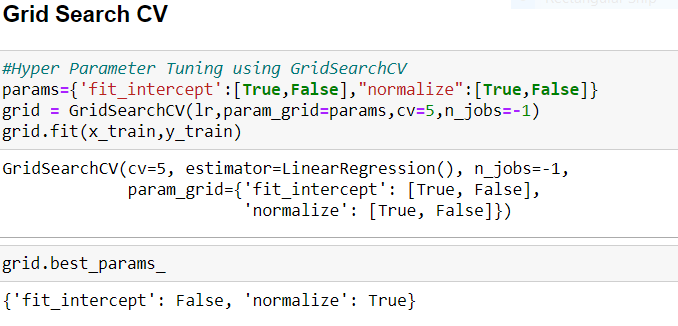






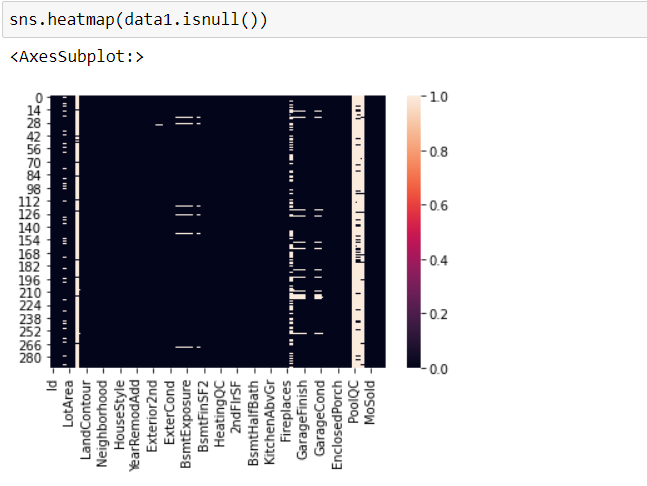
**Key Metrics for success in solving problem under consideration**

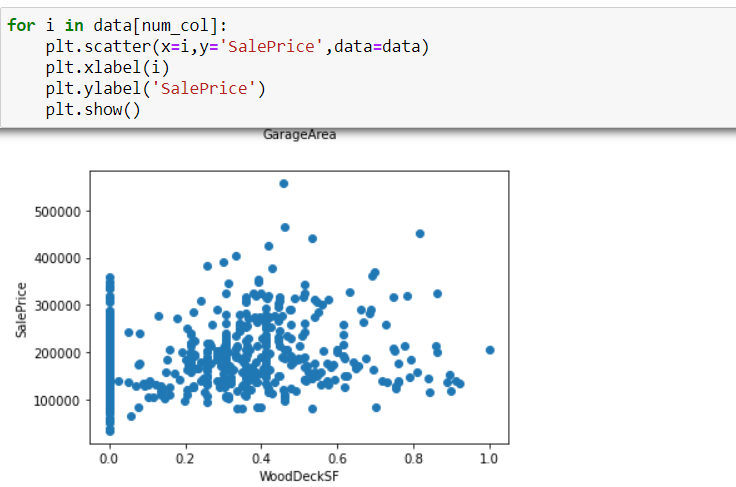
After Checking the model evaluation, I have considered Linear Regression, Lasso Regression and Ridge Regression as my final models and I have applied hyperparameter tuning for improving the scores. By using Grid search cv we are going to pass different parameters for the final algorithm which improves the evaluation of model.



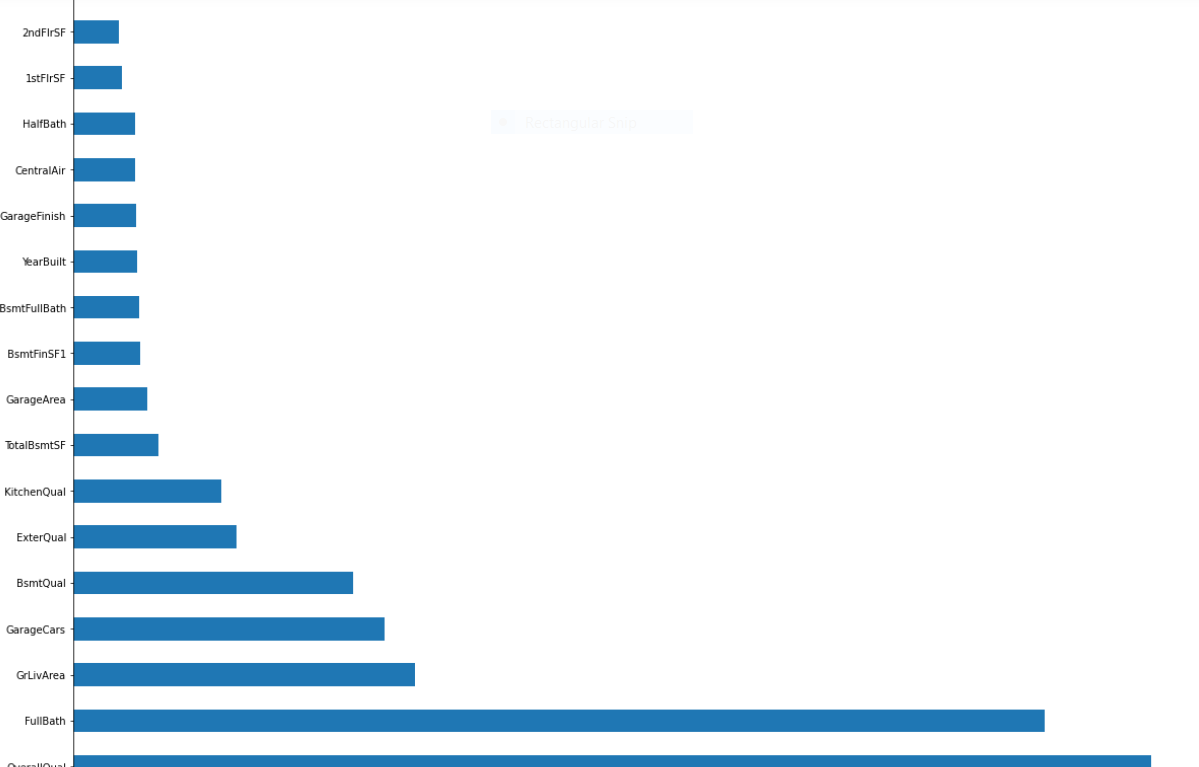
**Visualizations**

I have plotted heatmap for checking the null values in the columns using the is-null method and found the columns with null values the column which has more null values will be seen as shown in the below figure.

-

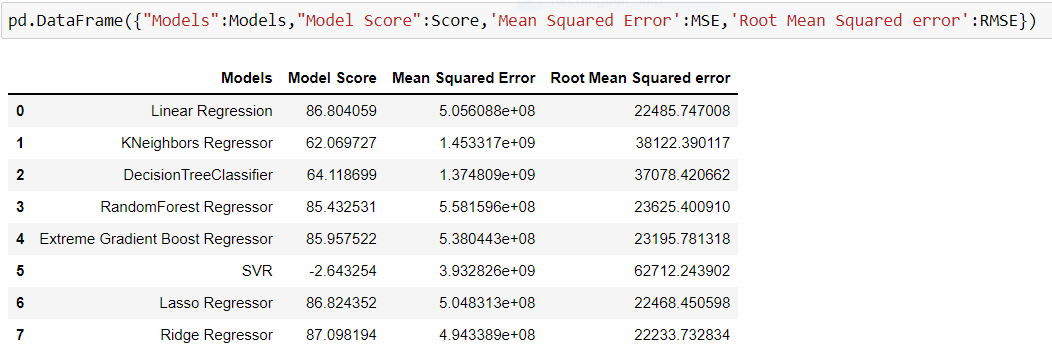
I have used scatter plots to check the relation between features and label.

I have used Extra Tree Regressor to select the top 30 best features.



**Interpretation of the Results**

I have created a data frame using all the models used for prediction, their scores and errors given by the models.

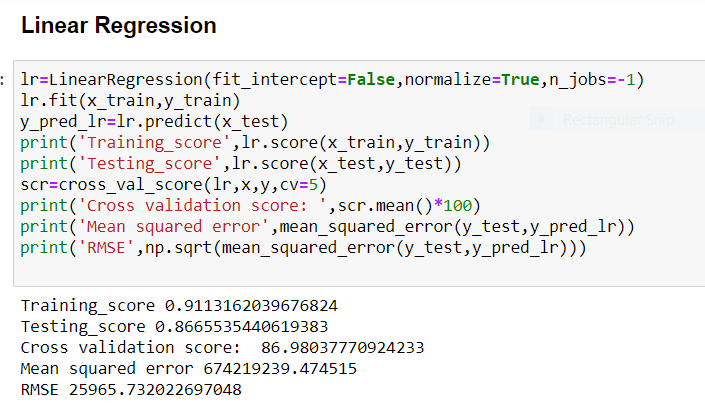


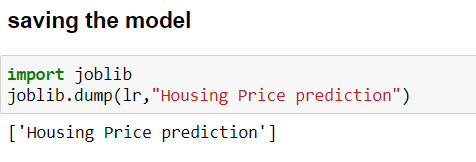
**CONCLUSION**

**Key Findings and Conclusions of the Study**

Housing is a major driving force of the economy and often an individual’s biggest asset. The situation within the housing sector is of high significance for a society’s social and economic development and there is a need to openly recognize problems within the sector in order to develop and implement feasible policy options. The machine learning models developed in this research will be tested using several methods like such as Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) RMSE is used to calculate predicted performance by considering the prediction error of each data.

So, after making many predictions using different models linear Regression and Lasso Regressions are best fit models for our prediction in terms of accuracy and the error. So, I have concluded the linear Regression Model as my best fit model and saved the model using job lib.





**Conclusions on our model building**

We got our best model as Linear Regression with the score of 86% and both the training and testing scores are balanced and the RMSE value is also less compared to all other models. So, we can go further build our model as Linear Regression.

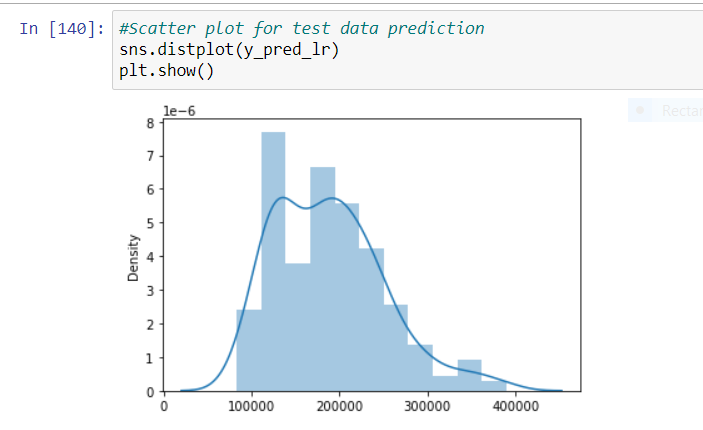
**Learning Outcomes of the Study in respect of Data Science**

First, we have filled the null values, then I have checked for the outliers and skewness in the dataset using the box plots and distribution plots respectively then I have tried removing the outliers using the z-score and skewness using log transformation. Then I have scaled the data. The same method I have followed the same procedure for the test dataset also and checked for the important features using extra tree regressor.

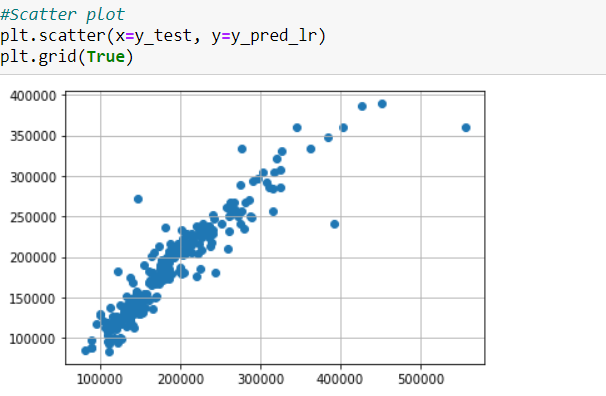
Since our prediction is a continuous type, it comes under regression problem so used various regression algorithms for label prediction.

Compared to all the other models used, Linear regression and Lasso Regression models are performing well and the score obtained is all good compared to the other algorithms and the error is also less. So, finally I saved my model as Linear Regression.

I have plotted distribution plot on predicted values and checked the flow of the data.



I even plotted scatter plot and compared the predicted values and the actual values and got to know that the predicted values are linearly varying with the actual values.



**Limitations of this work and Scope for Future Work**

In the housing project, I have performed many Regression algorithms to predict the house price particle swarm optimization methods to perform house price prediction. The RMSE errors calculated for all the algorithms are very high. Statistical methods work better, on large set data. But the length of the dataset is very less so using different methods that match the time-series data will be used in the future research to obtain smaller error prediction values (RMSE) and using more data to get the better result.