

**PFA HOUSING PROJECT**

Submitted by:

NIKHIL SINGH RANA

**ACKNOWLEDGMENT**

The satiation that accompanies the successful completion of the project would be incomplete without the mention of the people who made it possible

I would like to take the opportunity to thank and express my deep sense of gratitude to my data trained academy mentors for providing their valuable guidance at all stages of the study of my data scientist course, their advice and constructive suggestion through which I have gained this much skills that I can complete this project.

I have taken the help of my previous projects which I had done in my training phase with data trained academy and also refered google for some line of codes .

**INTRODUCTION**

* **Business Problem Framing**

The first step in not just an ML but any project is to simply define the problem at hand. You first need to understand the situation and the problem which needs to be solved.

This Problem is related to the Housing Prices which is dependent upon various independent variables in the dataset:

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. You are 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?

* **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 compan

* **Review of Literature**

The real estate market is one of the most competitive in terms of pricing and same tends to vary significantly based on numerous factors; forecasting property price is an important module in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features multiple regressions or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result.

Need for Real Estate Value Prediction:

• While our nation continues growth trend and the construction industry lags behind demand, prices will continue to rise while interest rates bump upward.

• Securing investment property now with thorough due diligence and watch investment pay off over the next few years.

Advantages of Machine Learning Over Real Estate Value Prediction :

• Trends and Patterns Are Identified with Ease

• Machine Learning Improves Over Time

• Machine Learning Lets You Adapt Without Human Intervention

• Enables Automation

* **Motivation for the Problem Undertaken**

The Objective behind to do this project is following:

We are required 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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

1. Prepare the problem

The first step in not just an ML but any project is to simply define the problem at hand. First understanding the situation and the problem which needs to be solved. Once you know the problem well, you then head on to solve it.

**Load libraries:**

The very first step is to load or import the all the libraries and the packages required to get the results you want. Some very primary and almost necessary packages for Machine Learning are — Numpy, Pandas, MatplotLib, sikit learn, Dtale etc.

**Load Dataset:**

Once the libraries are loaded, you need to get the data loaded. Pandas has a very straightforward function to perform this task — pandas.read csv.The read.csv function is not just limited to csv files, but also can read other text based files as well

2. Summarise problem

Okay, so the data is loaded and ready to be actioned upon. But first I need to check how the data looks and what all does it contain. To begin with, you would want to see how many rows and columns does the data have and what all are the data types of each column (which pandas thinks they are).

**Descriptive statistics:**

Descriptive statistics, as the name suggests, describes the data in terms of its statistics — mean,standard deviation, quantiles etc. The easiest way to get a complete description is by pandas.DataFrame.describe.

Data Visualization :

Data Visualizations are very important as they are the quickest way to know the data and the patterns — if they even exist or not. Your data may have thousands of features and even more instances.

Visualizations using Matplotlib, Seaborn and Dtale can be used to check the correlations  within the features and with the target, scatter plots of data, histograms and boxplots  for checking the spread and skewness and much more. Even pandas has its own built in visualization library — pandas.DataFrame.plot which has bar plot, scatter plot, histograms etc.

So I have performed barplot , histogram and distplot for the visualization purpose. Also I have used Dtale library which performs evry visualization possible from the data .

3. Prepare Data

Once you know what your data has and looks like you will have to transform it in order to make it suitable for algorithms to process and work more efficiently in order to give more accurate and precise results. This is essentially Data Pre-Processing which is the most important and the most time consuming stage of any ML project.

Also I have checked the correlation between the columns for better data cleaning purpose

Data cleaning:

Real life data is not arranged and presented to you nicely and in a dataframe with no abnormalities. All this needs to be handled manually which takes a lot of time and coding skills

Pandas has various functions to check for such abnormalities like pandas.DataFrame.isna  to check for values with NaNs etc.

drop irrelevant features using pandas.DataFrame.drop

Outlier Removal:

This is also the main step[ to improve the accuracy of the model, therefore I firstly check that how many outliers are present in each column by the method of box plot And then remove the outliers by the threshold method .

4. Evaluate Algorithms :

Once the data is ready,I ll proceed to check the performance of the various classification algorithms .

Split-Out validation Dataset:

Once the model is trained, it needs to be validated as well to see if it really generalized the data or it over/under fitted. The data in hand can be split up beforehand as training set and validation set. This split-out has various techniques — Train Test Split, Shuffle split etc. You can also run Cross Validation on the entire data set for a more robust validation.

Test options and Evaluation

The models need to evaluated based on a certain set of evaluation metrics which need to defined.

5. Imporve Accuracy:

After you have the best performing algorithms with you, their parameters and the Hyperparameters can be tuned to give maximum results. Multiple algorithms can be chained as well.

Ensembles :

Multiple Machine Learning algorithms can be combined to make a more robust and optimal model that gives better predictions than the single algorithm. This is known as an ensemble.

* **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 below. The company is looking at prospective properties to buy houses to enter the market. You are 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?

Data Description:

• Data contains 1460 entries each having 81 variables.

• Data contains Null values. We treated them using the domain knowledge and our own understanding

• Data contains numerical as well as categorical variable. we handled them accordingly

• There are some important feature which i have mentioned using feature selection technique

* **Data Preprocessing Done**

Data usually has a lot of so called abnormalities like missing values, a lot of features with incorrect format, features on different scales etc.

Pandas has various functions to check for such abnormalities, following are some of the functions which I have performed for data cleaning purpose:

* Df.drop, to drop the columns which are not needed.
* Label Encoder, Feature Scaling to convert the object columns into the Integer columns.
* Outlier removal, to remove the outliers so that accuracy should be maintained.
* Knn Imputer to fill in the Nan Values is LotFrontage Column.
* For Visualization, I have used **Dtale** library.
* **Hardware and Software Requirements and Tools Used:**

The hardware and software requirements along with the tools, libraries and packages used are mentioned as follows:

* **Firstly you will need the python in your desktop or laptop**
* **Then import pandas, numpy, csv to read the csv files and work on it.**
* **Now I have imported matplotlib, Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python**
* **Imported Seaborn.sns, Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics**.
* **Imported Label Encoderand KNN Imputer, to encode the object columns into the integer columns.**
* **Imported Dtale to perform visualization, describe, etc.**

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

The approaches I followed, both statistical and analytical, for solving of this problem are as follows:

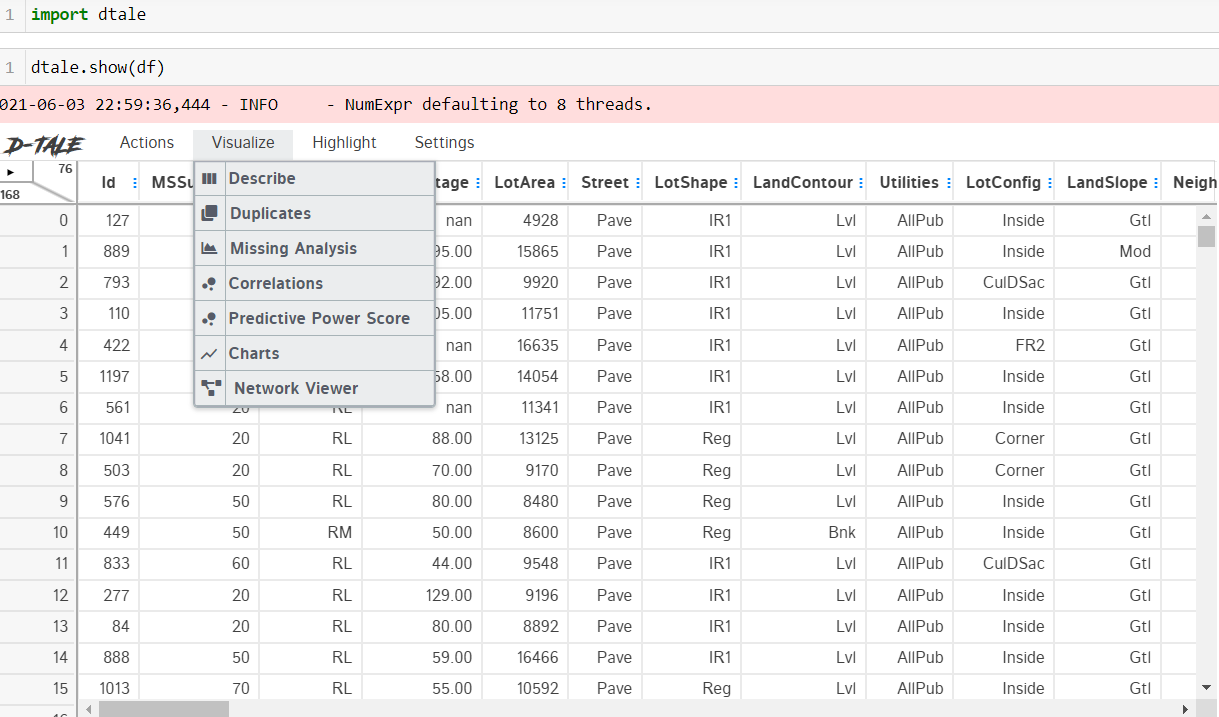
* Firstly in data cleaning, I drop irrelevant features using pandas.DataFrame.drop, checked if there’s any Nan values, removed the outliers, checked the skewness, checked the correlation and then performed the visualization
* Filled the Nan values by using label encoder and feature scaling.
* I have imported Dtale library for the visualization purposes.
* **Visualizations**

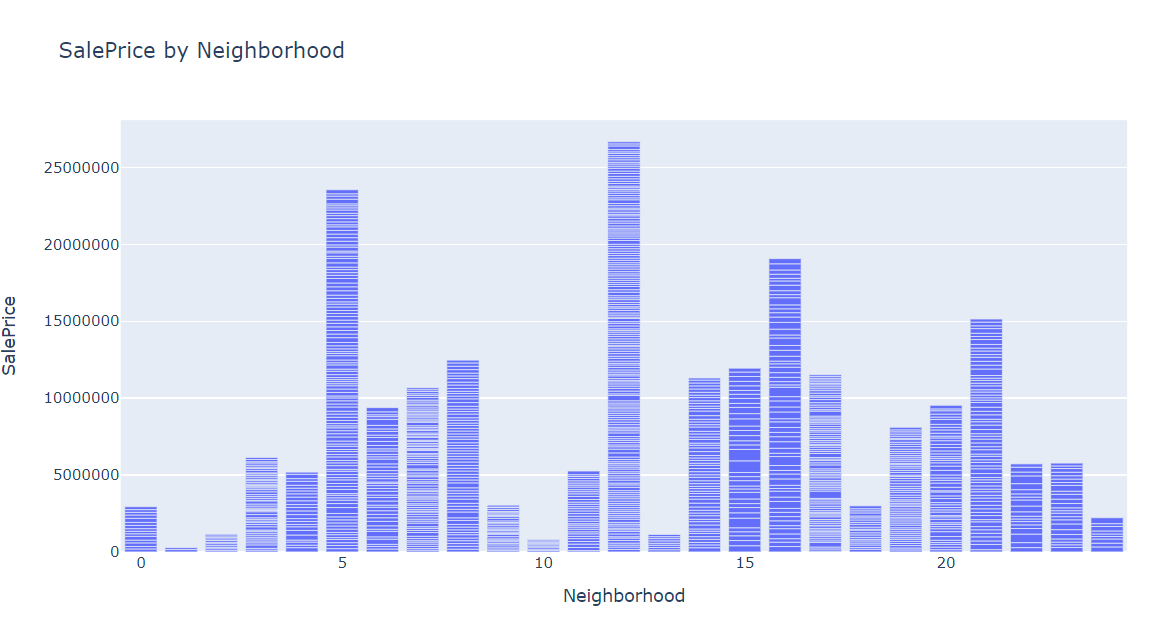
Data Visualizations are very important as they are the quickest way to know the data and the patterns — if they even exist or not. Your data may have thousands of features and even more instances. It is not possible to analyze the numeric data for all of them .

Visualizations using Matplotlib, Seaborn can be used to check the correlations within the features and with the target, scatter plots of data, histograms and boxplots for checking the spread and skewness and much more.

Following are the visualization I have used:

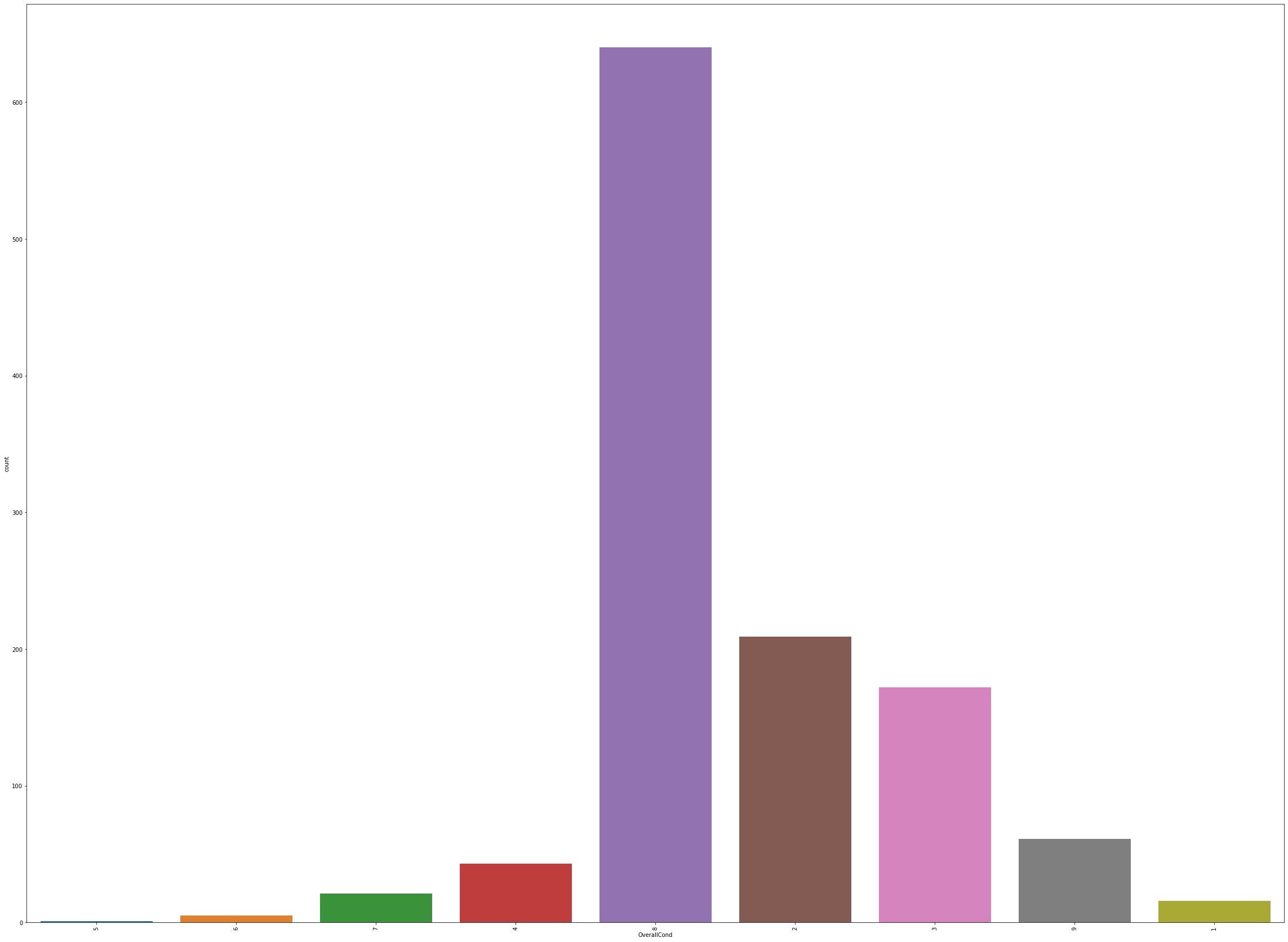
* Dtale Visual Library: With the help of this we can describe the data, find the duplicate entries, find correlation, find the information about each column and try many visual representation.





We can check all the relation between different columns with the help of dtale

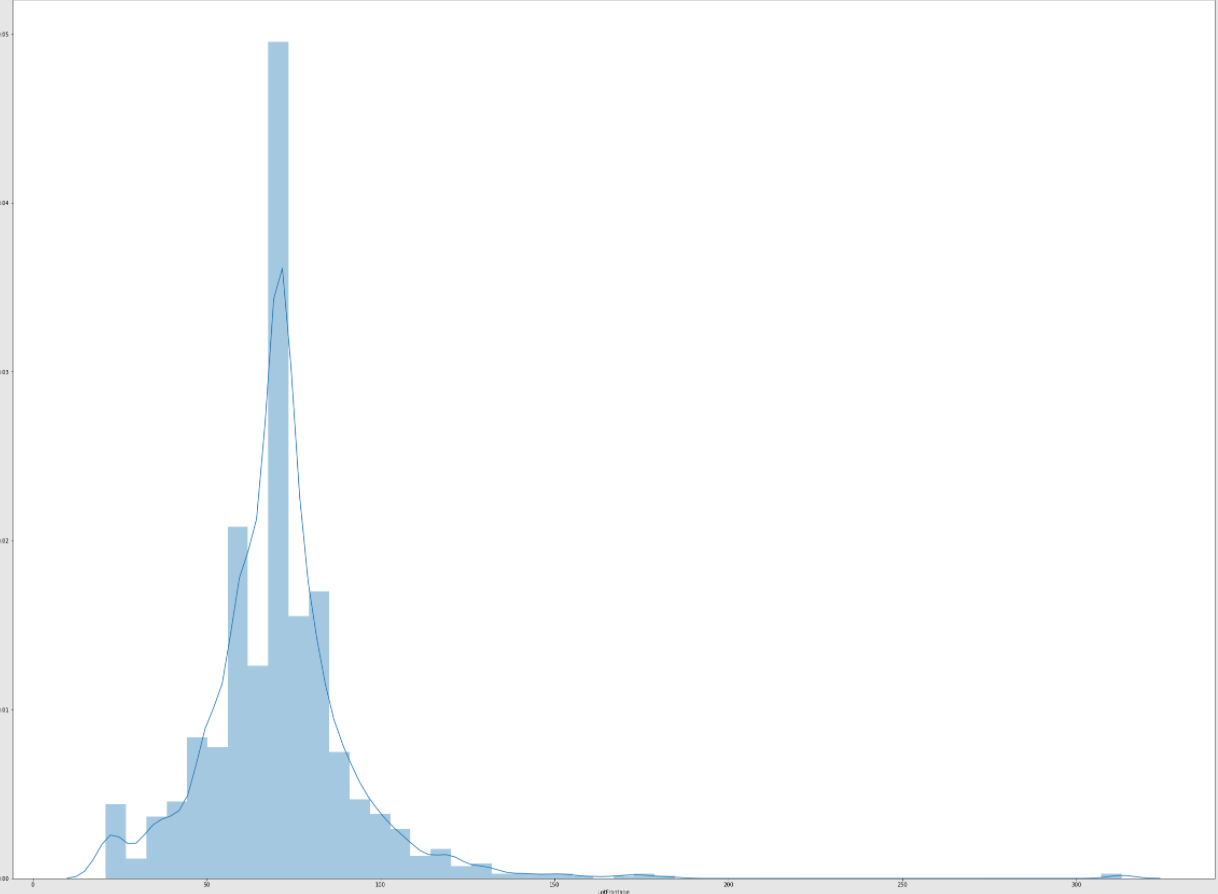
- Count Plot: With the help of it I can see the count of each column For example We can see that how many houses are in what condition:



Here we can see that most of the house have very good condition followed by the poor condition Houses.

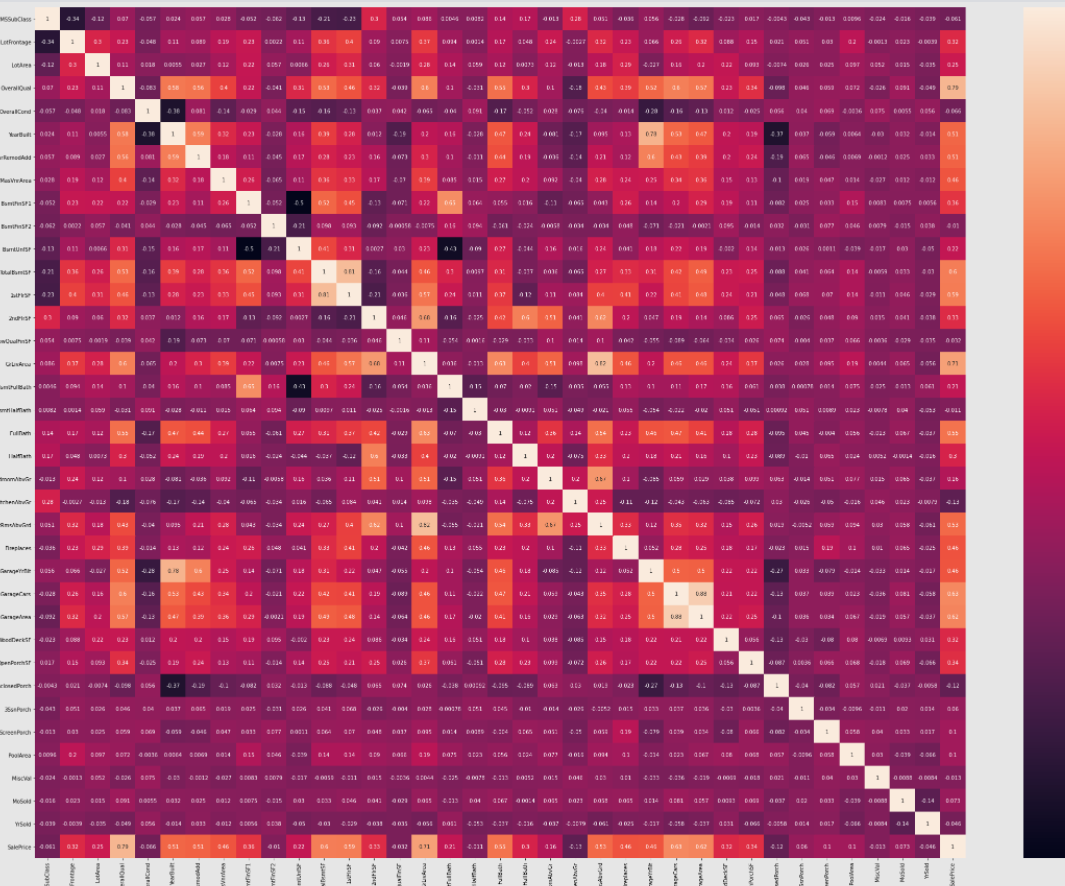
* Dist Plot: Seaborn **distplot** lets you show a histogram with a line on it. This can be shown in all kinds of variations. We use seaborn in combination with matplotlib, the **Python** plotting module. ... The **distplot**() function combines the matplotlib hist function with the seaborn kdeplot() and rugplot() functions.

Following we can distplot of a particular column and know the if it is right skewed or left :



Here we can see that the lotFrontage column is left skewed and similarly we can check for trhw other columns.

* Correlation: Here we can see the correlation between all the columns with each other, and remove the highly related columns:



* Feature Importance:

We can see the features which are most important in relation to the target variable.

Followiing are the top 10 features which we have interpreted:

Specs Score

3 LotArea 5.746267e+06

63 MiscVal 5.669726e+06

39 2ndFlrSF 3.837320e+05

32 BsmtFinSF2 3.407783e+05

30 BsmtFinSF1 3.338184e+05

62 PoolArea 3.241779e+05

22 MasVnrArea 2.458125e+05

33 BsmtUnfSF 2.402027e+05

40 LowQualFinSF 2.103716e+05

34 TotalBsmtSF 1.395232e+05

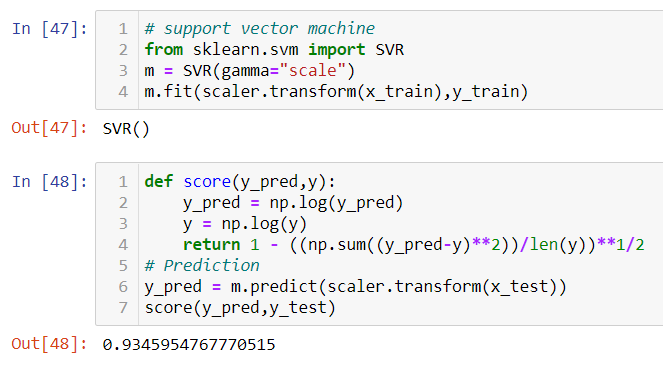
* **Interpretation of the Results:**

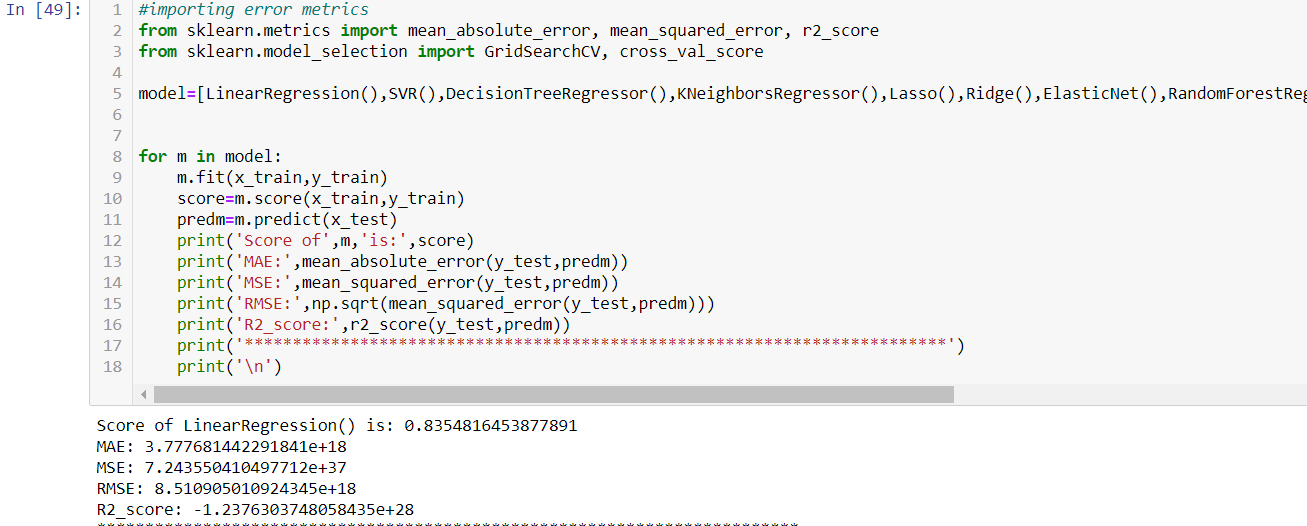
Summary of what results were interpreted from the visualizations, preprocessing and modeling:

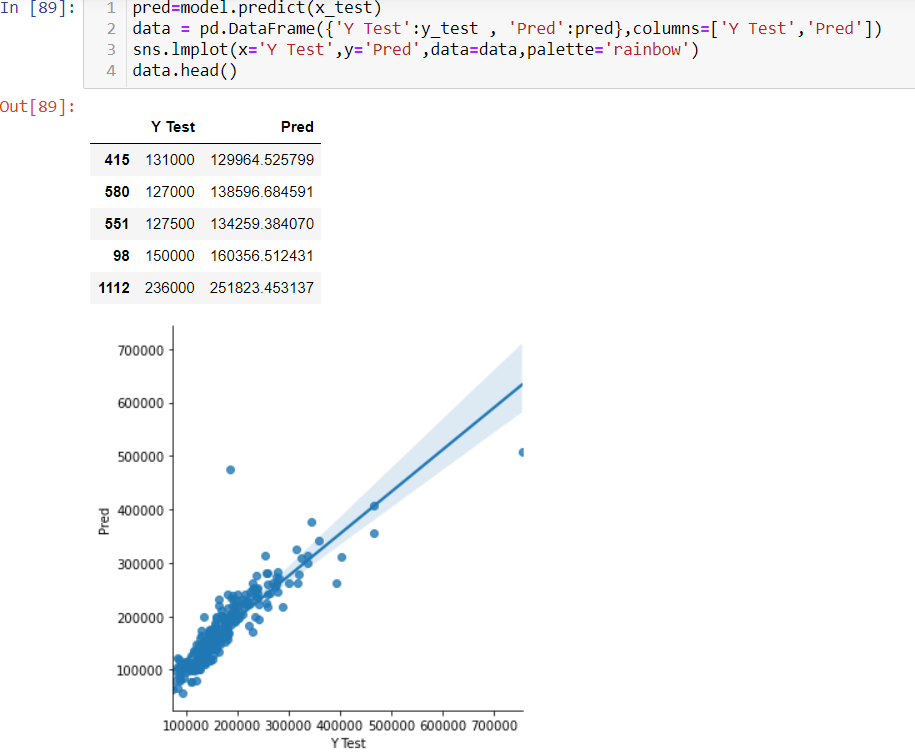
* There are No duplicates entries.
* There are some columns which have more than 50% of Nan values.
* There are small No.of Outliers which we can detect through Dtale library.
* variables TotalBsmtSF and 1stFlrSF are highly correlated with a correlation of 0.81 and
* variables GrLivArea and TotRmsAbvGrd are highly correlated with a correlation of 0.82
* variables GarageArea and GarageCars are highly correlated with a correlation of 0.88Most of the Customers Strongly agree that they have friendly interface of the website.
* Most of the column have categorical variables.
* House Style of one story dwelling style has greater sales price.
* Most of the data is skewed but didn’t effect the accuracy of the model
* We can detect the top 10 important features in the dataset
* Linear Regressor model has the good accuracy score.

* Run and Evaluate selected models

Following are the algorithms I have used:







**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Key findings are that, the EDA process helps us to understand the problem well through visualizations and data cleaning, this will further help us in model building purpose and gaining the maximum accuracy score.

Through EDA process we can see that there are some columns which are not needed and there are some outliers to be removed so that the accuracy should be maintained.

Also there are high correlation between some columns . Therefore we can visualize and check the relationship between that columns.

.

* Learning Outcomes of the Study in respect of Data Science

Firstly visualization helps us in various forms as follows:

* **Communicate Findings in Constructive Ways**

## Understand Connections Between Operations and Results

## Interacting With Data

## Create New Discussion

## Secondly, after applying algorithms we can see which algorithm performs better and gives better prediction.

## We can also use ensembling techniques to improve the accuracy of the models build.

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

We can hypertune the models to future work properly