

**HOUSING PRICE PREDICTION PROJECT**

Submitted by:

Durgesh Pandey

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to all the Mentors who have taught me Machine Learning because of the knowledge they had provided to me I am able to complete this project.

**INTRODUCTION**

* Business Problem Framing

Housing and Real Estate Markets are one of the major contributor’s in a country’s economy. It is very large market and various companies are working in this domain. Data Science can play a vital role in solving problems related to this domain and can help the countries in their overall revenue, profits and improving their marketing strategies. Machine learning techniques can be used for achieving business goals for this housing companies. Our Problem is related to one of such U.S based housing company named Surprise Housing which want to enter Australian Market. The company want to use Data Analytics to purchase houses at a price below their actual values and flip them at a higher price. The company has collected a dataset from the sale of houses in Australia. The company is looking at prospective properties to buy houses to enter the market. we will build a model using Machine Learning to predict the actual value of the prospective properties and it will help the company to decide whether to invest in property or not.

* Conceptual Background of the Domain Problem

Trends in housing prices indicate the current economic situation and also are a concern to the buyers and sellers. There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms. House price depends upon its location as well. A house with great accessibility to highways, schools, malls, employment opportunities, would have a greater price as compared to a house with no such accessibility. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction.

* Review of Literature

The world is shifting from manual to automated systems. The objective of our project is to reduce the problems faced by the customer. In the present situation, the customer visits a real estate agent so that he/she can suggest suitable showplaces for his investments. But the above method is risky as the agent may forecast wrong prices to the customer and that will lead to loss of customer’s investment. This manual technique which is currently used in the market is outdated and has a high risk. So as to overcome the drawback, there is a need for an updated and automated system. So we are using machine learning techniques where we will be using different algorithms for this project to get the accurate prediction for the price of the house.

Machine learning is a form of artificial intelligence which compose available computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadly classified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning.

Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data

* Undertaken

Growing unaffordability of housing has become one of the major challenge for countries around the world. In order to gain a better understanding of the commercialized housing market we are currently facing, we want to figure out what are the top influential factors of the housing price. Apart from the more obvious driving forces such as the inflation and the scarcity of land, there are also a number of variable that are worth looking into. Therefore, we choose to study the house price prediction., which enables us to dig into the variables in depth and to provide a model that could more accurately estimate house prices. In this way, people could make better decision when it comes to home investment.

Our objective is to discuss the major factors that affect housing price and make precise prediction for it. We use 80 explanatory variables including almost every aspect of residential homes in Australia. Methods of both statistical, regression models and machine learning models are applied and further compared according to their performance to better estimate the final price of each house. The model provides price prediction based on similar comparable of people’s dream house, which allow both buyers and sellers to better negotiate home prices according to market trend.

**Analytical Problem Framing**

* Data Sources and their formats

In this project we are given two CSV file containing train and test dataset of sales of houses in Australia.

There are 80 columns and our target variable is to predict Sale price. Below are the description of the columns

MSSubClass: Identifies the type of dwelling involved in the sale. MSZoning: Identifies the general zoning classification of the sale. LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Alley: Type of alley access to property LotShape: General shape of property LandContour: Flatness of the property Utilities: Type of utilities available LotConfig: Lot configuration

LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits Condition1: Proximity to various conditions

Condition2: Proximity to various conditions (if more than one is present) BldgType: Type of dwelling

HouseStyle: Style of dwelling

OverallQual: Rates the overall material and finish of the house OverallCond: Rates the overall condition of the house YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodelling or additions) RoofStyle: Type of roof

RoofMatl: Roof material

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material) MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior ExterCond: Evaluates the present condition of the material on the exterior Foundation: Type of foundation

BsmtQual: Evaluates the height of the basement

BsmtCond: Evaluates the general condition of the basement BsmtExposure: Refers to walkout or garden level walls BsmtFinType1: Rating of basement finished area

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types) BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area TotalBsmtSF: Total square feet of basement area Heating: Type of heating

HeatingQC: Heating quality and condition CentralAir: Central air conditioning Electrical: Electrical system

1stFlrSF: First Floor square feet 2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors) GrLivArea: Above grade (ground) living area square feet BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms FullBath: Full bathrooms above grade HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms) Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) Functional: Home functionality (Assume typical unless deductions are warranted) Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity GarageArea: Size of garage in square feet GarageQual: Garage quality

GarageCond: Garage condition PavedDrive: Paved driveway

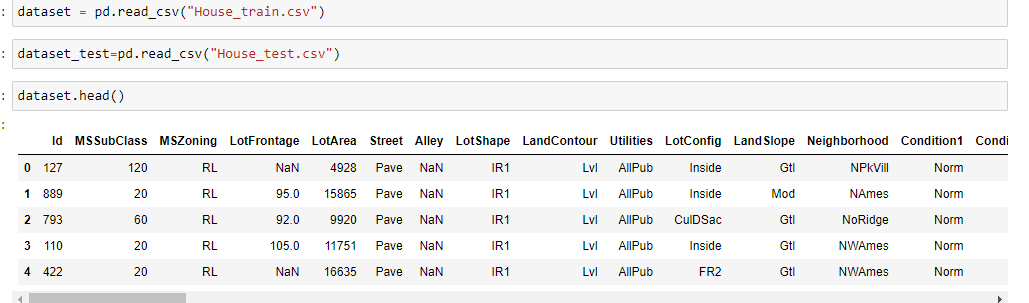
WoodDeckSF: Wood deck area in square feet OpenPorchSF: Open porch area in square feet EnclosedPorch: Enclosed porch area in square feet 3SsnPorch: Three season porch area in square feet ScreenPorch: Screen porch area in square feet PoolArea: Pool area in square feet

PoolQC: Pool quality Fence: Fence quality

MiscFeature: Miscellaneous feature not covered in other categories MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM) YrSold: Year Sold (YYYY) SaleType: Type of sale SaleCondition: Condition of sale

* Mathematical/ Analytical Modeling of the Problem



First of all we will load necessary libraries and then will load our HOUSING.csv file

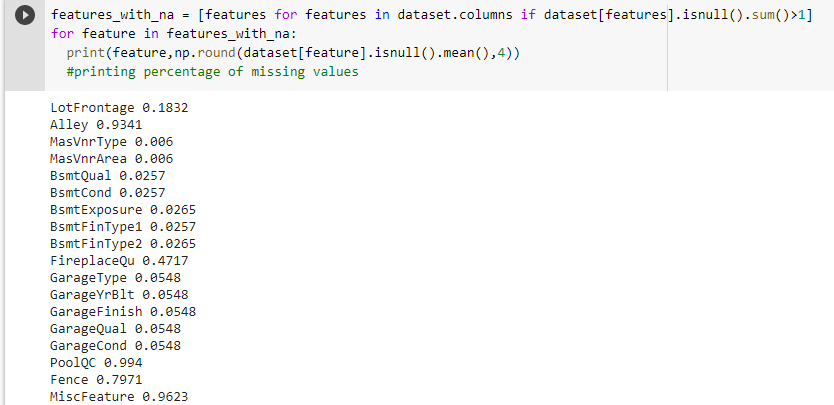
Now we will check the columns and shape of the dataset

Simultaneously with the same process we will check our test data set.

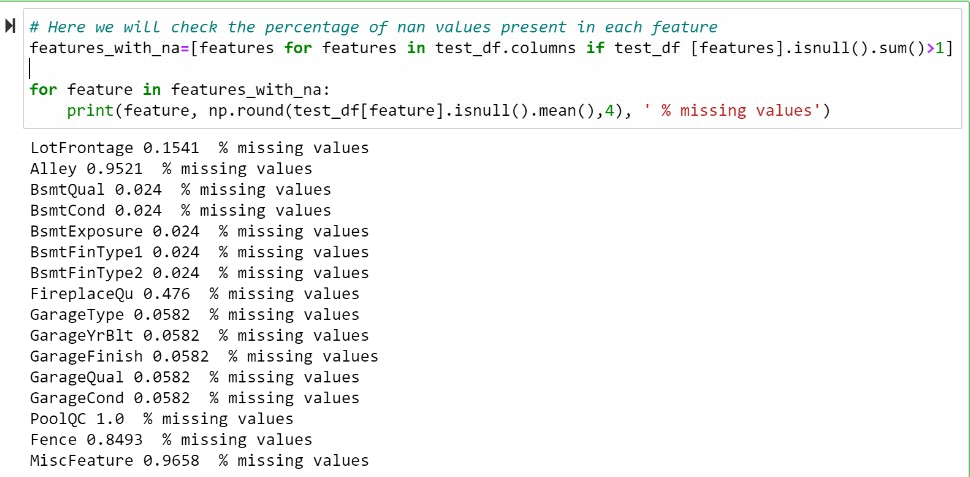


From the above visualization we can see that we have 1168 rows and 81 columns in train data set and 292 rows and 80 columns in test data set.

Let’s check null values in both the data set.



Here we can see that there are many null values present in train dataset. Some columns contains more that 80% of the null values.

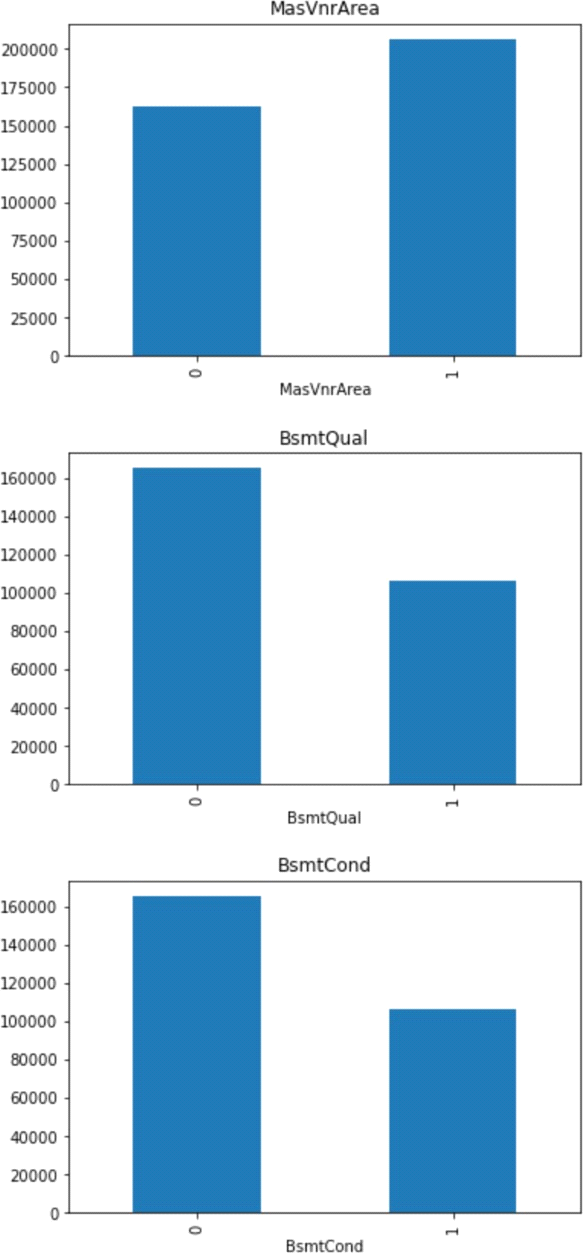
Let’s go with test dataset.

The same can be seen in test data set also.

* Visualization

Since there are many missing values, we need to find the relationship between missing values and Sales price

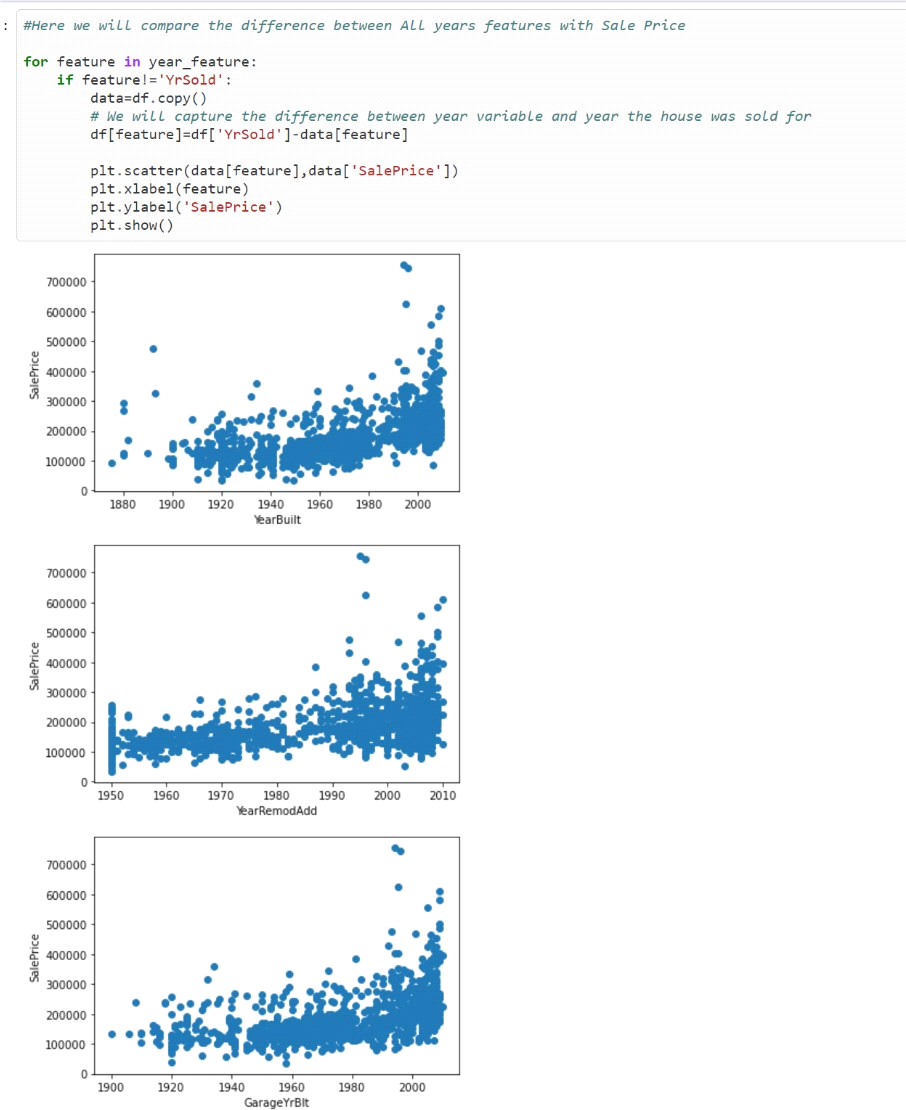




Here variable 1 indicates missing and 0 indicates its not

With the relation between the missing values and the dependent variable is clearly visible. So We need to replace these nan values with something meaningful.

we can see that features like lot frontage, alley has nan values because of this the median sale price is increasing so we will replace it with some meaningful later on.



Here we can see that year built, The house which was built earlier has less price than the recent one.

Year of recodification, if the year is 60 the price is very less as compared to 0 to 10 years.

Same thing happening with garage built year.

MSSubClass

Overall@ual

350000

250000

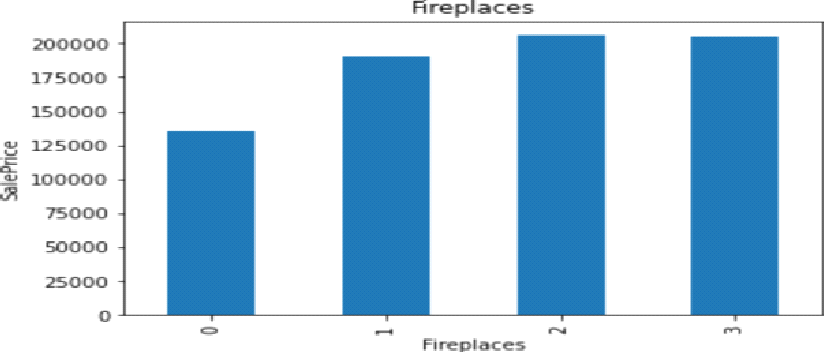
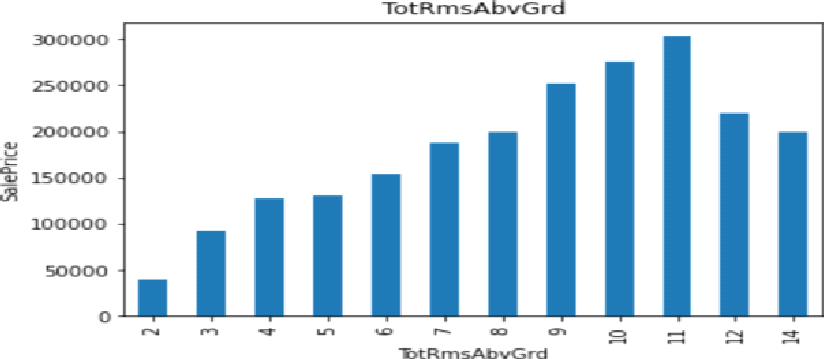
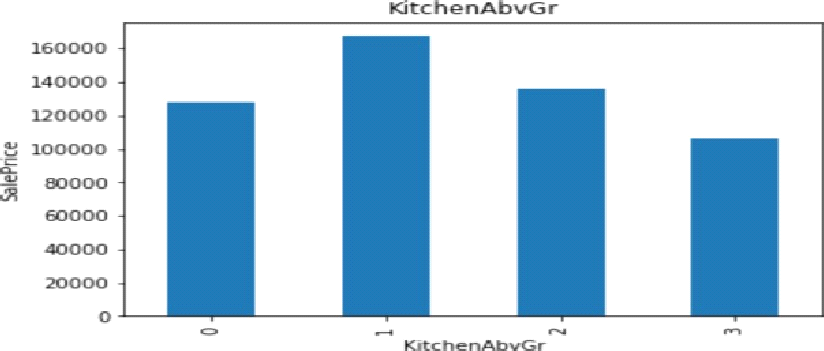
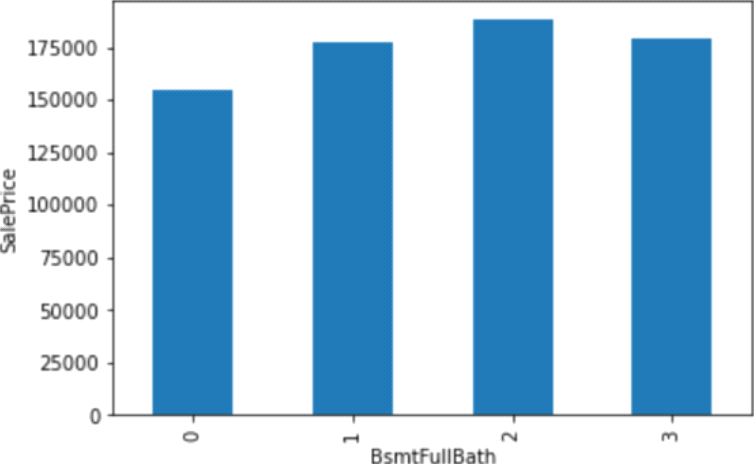
â

MOOO

£X/erallCond

LowOualFinSF

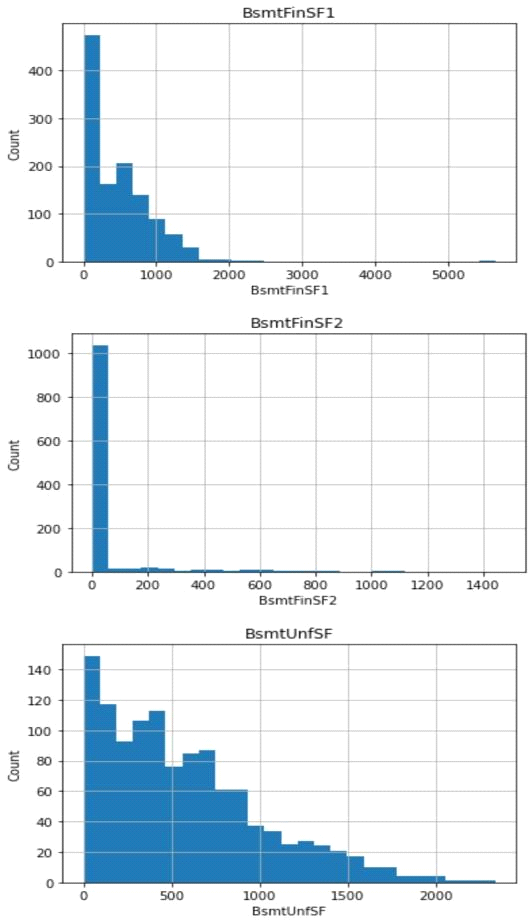
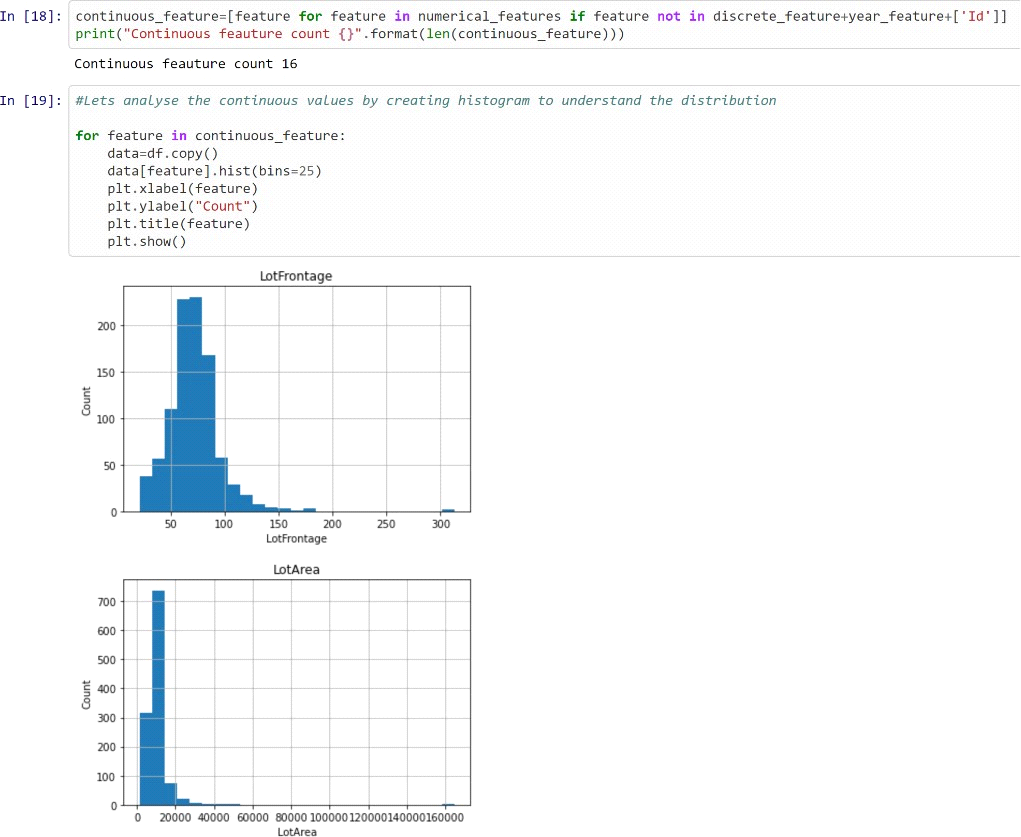
tm\*Q«aMnSF

%mtFuIIBath

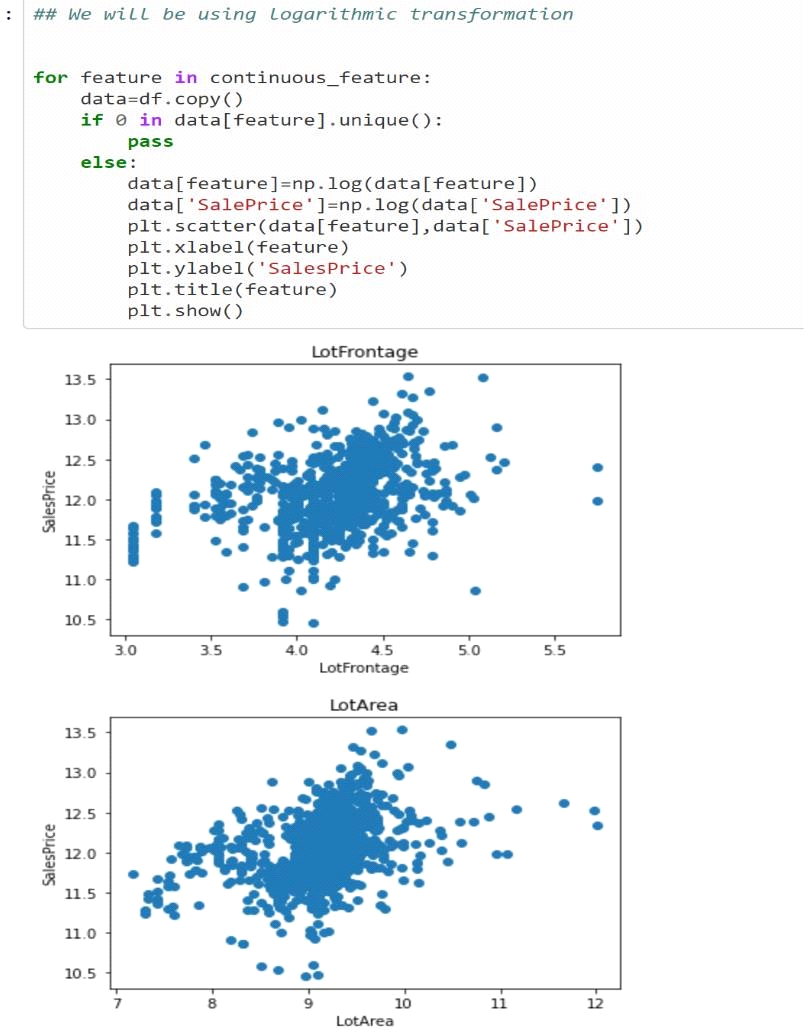
As we can see the overall quality is increasing the price is high.

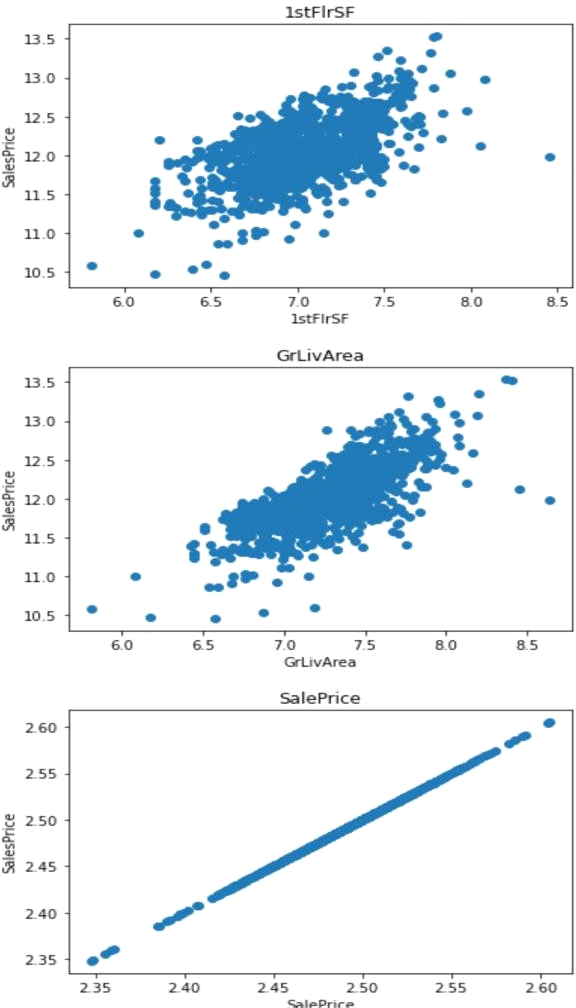
If the overall condition of the house is average the price is more that others. As there is increase in no of fire places the price of house is rising.

As the house has a pool area of 555 sq feet the price of the house is more as compared to others.



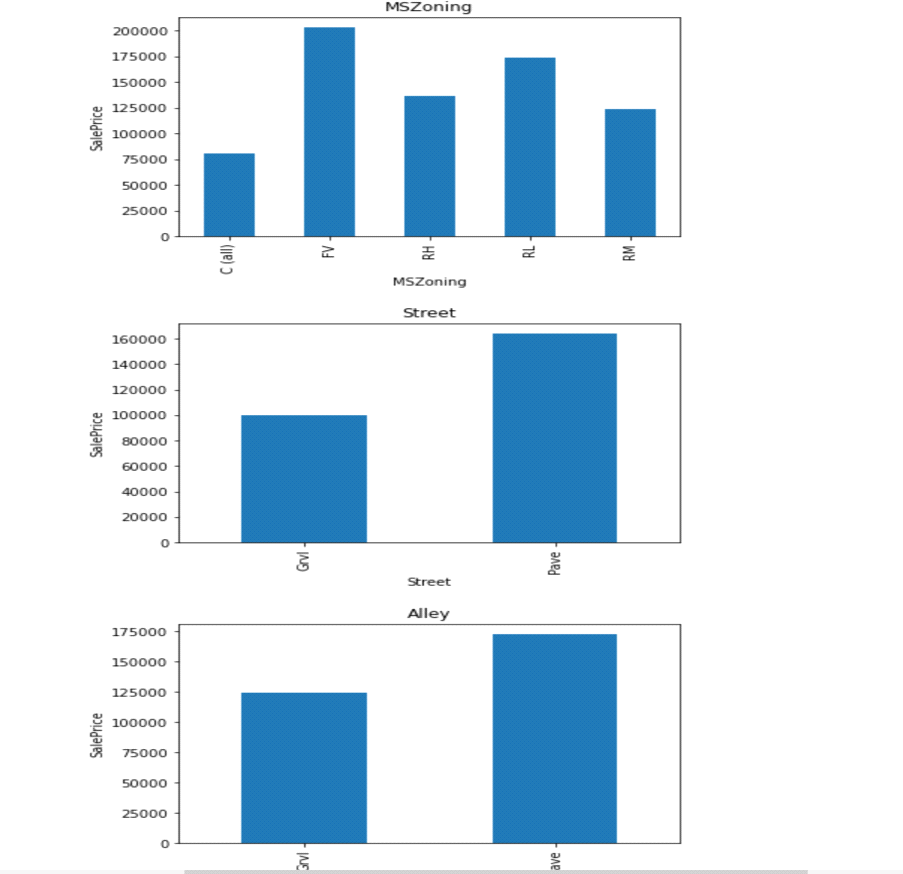
Here you can see that the data is skewed so we will perform log transformation to reduced skewness of the data.

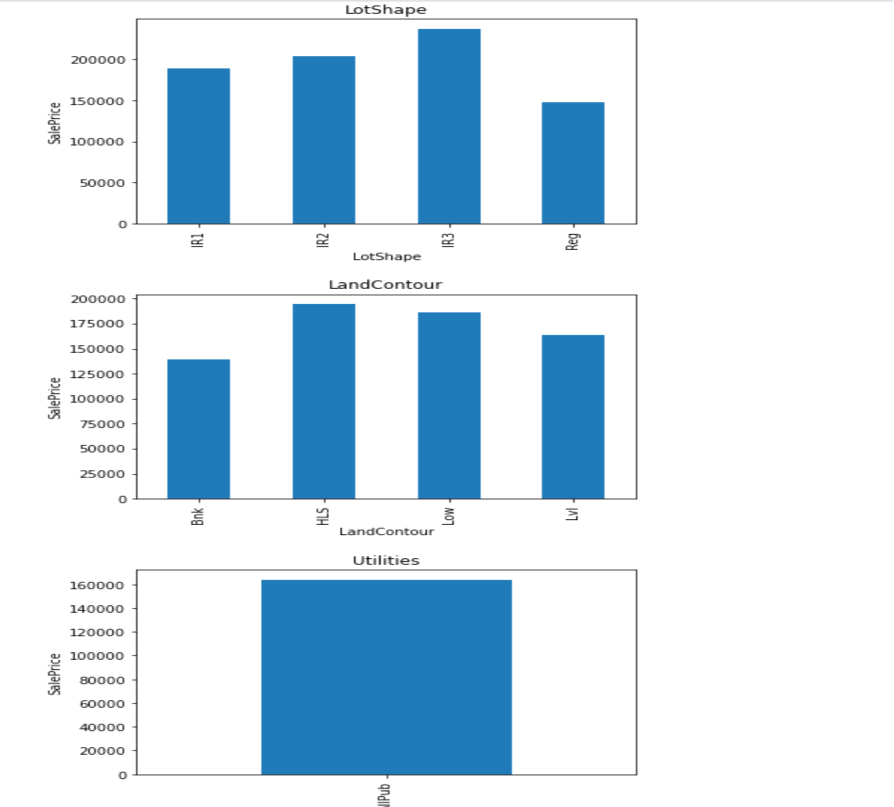


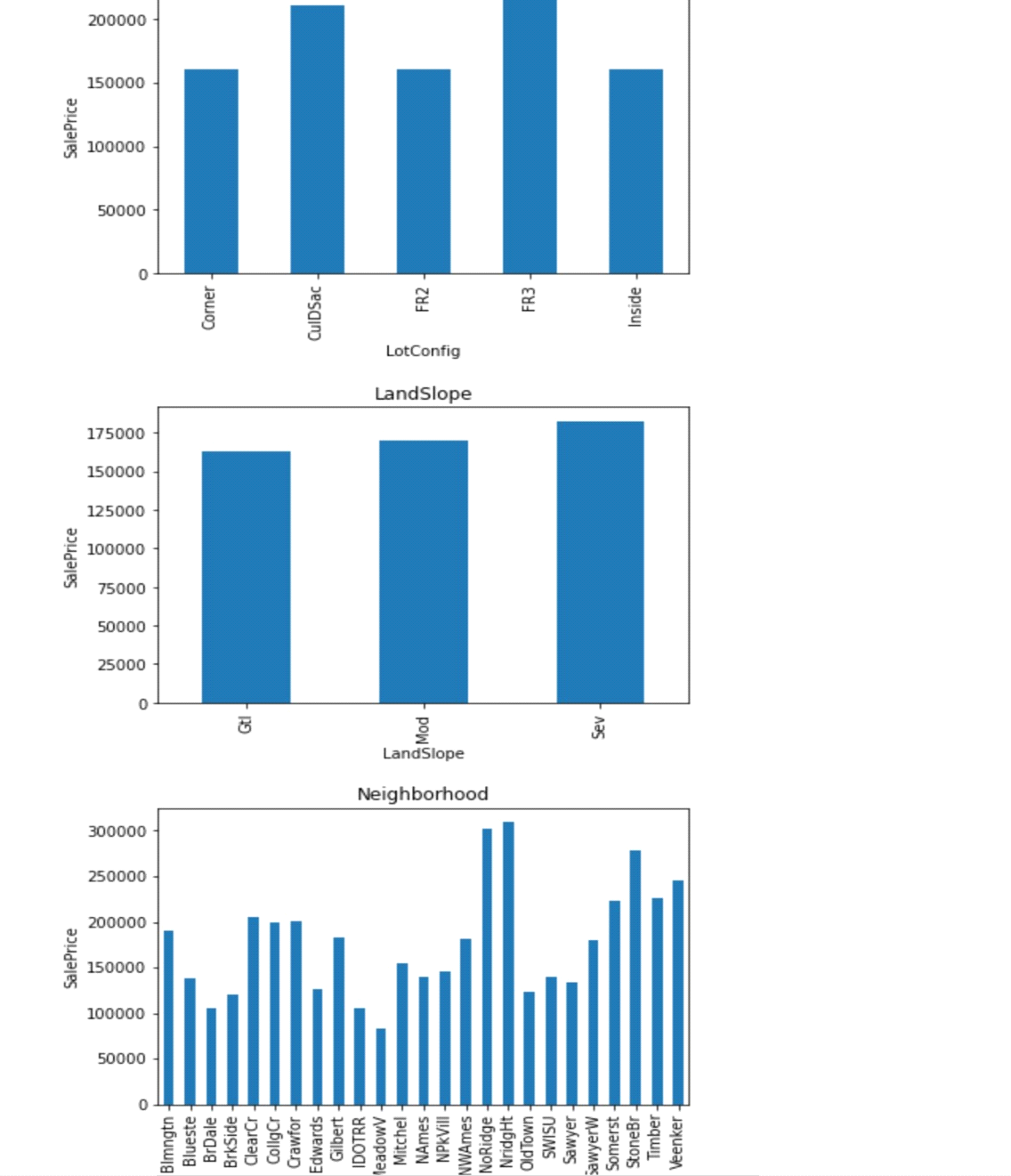


So after applying log transformation its giving monotonic relationship, as lot frontage, lotarea, Grlivarea is increasing the price is increasing.

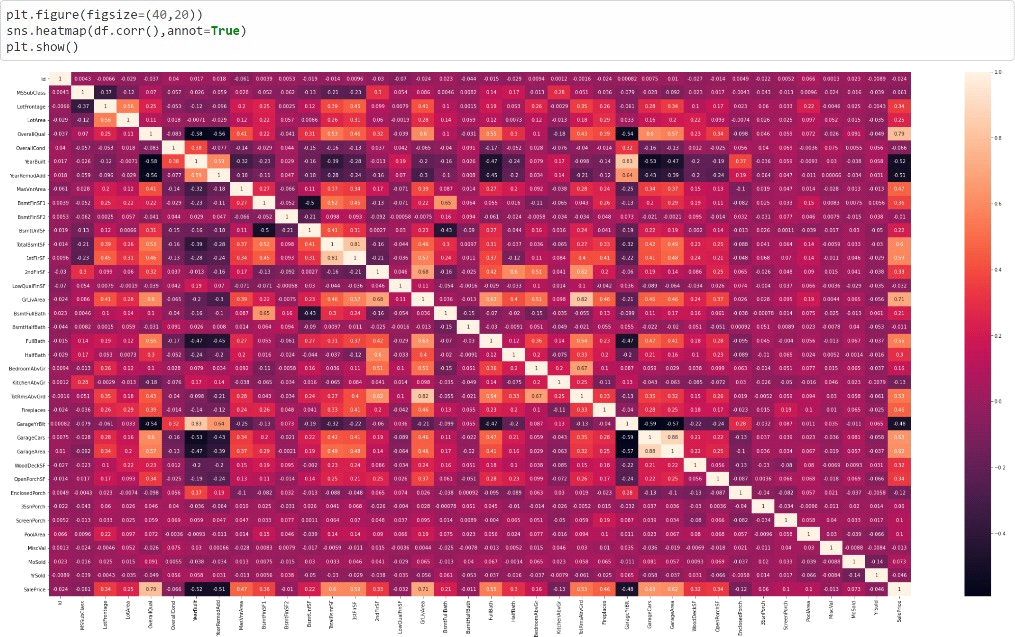
Let’s see the relationship between categorical feature and dependent variable.







Let’s do some statistical analysis



In this we found that Variables like OverallQual (overall material and finish of the house ) , Year Built , TotRmsAbvGrd ( Total rooms above grade (does not include bathrooms) , GarageCars (Size of garage in car capacity ) ,GarageArea ( Size of garage in square feet ) , GrLivArea ( Above grade (ground) living area square feet ) , FullBath ( Full bathrooms above grade ) have positive relationship with the sales Price.

YearBuilt, YearRemodAdd, GarageYrBuilt are negatively related with sale price.

* Data Preprocessing Done

As above we have seen there is lot of missing data so we will deal with these missing values .

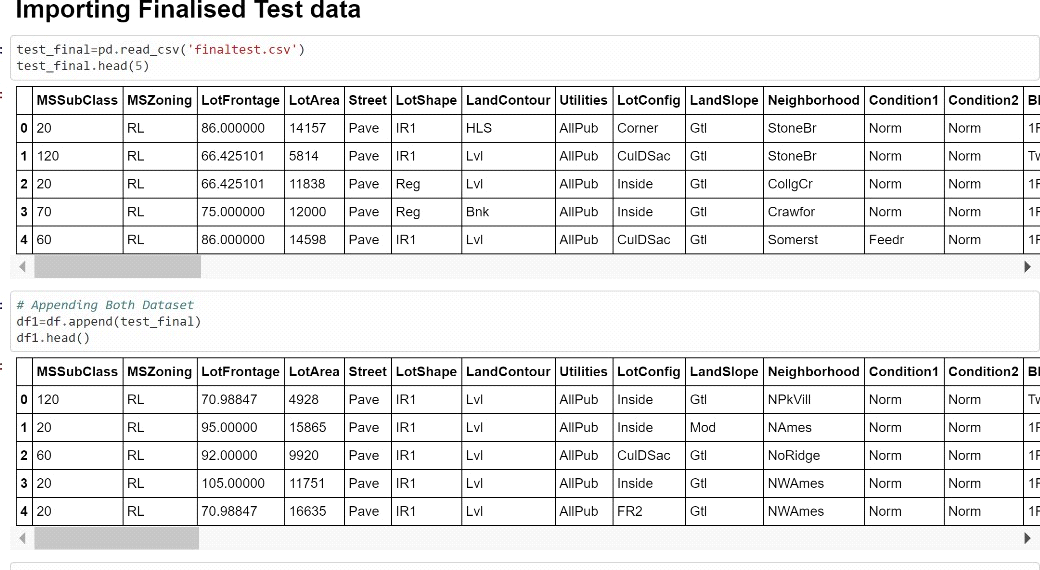
First of all we will drop columns Alley , MiscFeature , PoolQC , Fence and GarageYrBlt because more than 80 % data in these columns are missing if we replace these missing data with some data it can give us wrong prediction in the final model thus making our model less effective so better to drop these columns.



Now in the other columns GarageType, GarageFinish , GarageQual , GarageCond , BsmtFinType2 , BsmtExposure, Bsmtclond which have missing data of 10-20 % we will replace the missing data with the mode value .

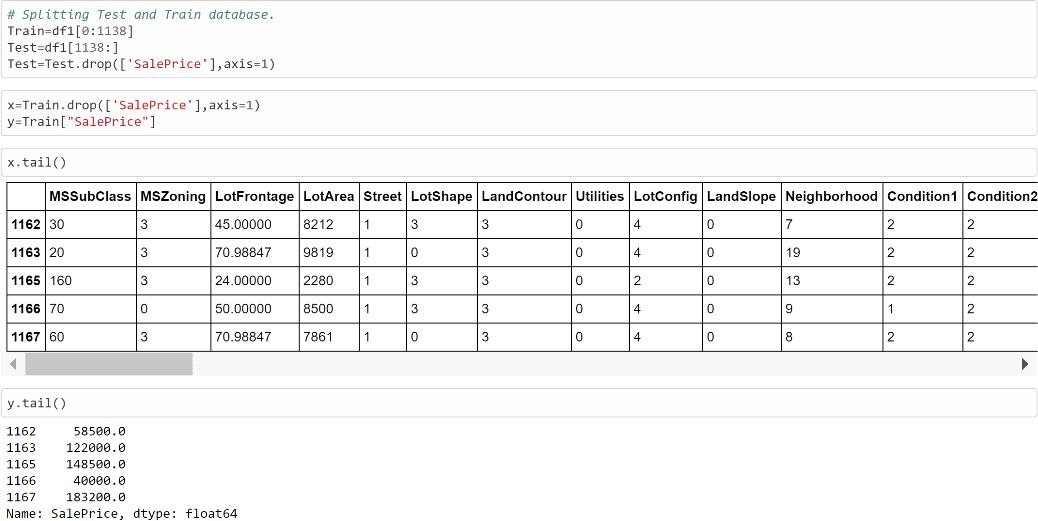
Same we will do with our test data set.

Now we will call our Finalised Test data set and append train and test data set .



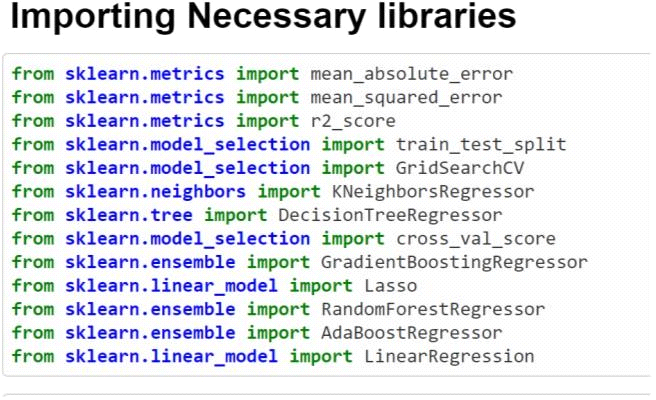
* Data Inputs- Logic- Output Relationships

Now we will divide the data into input and output , the output will be ‘SalePrice’ and all the other remaining columns will be input.



* Hardware and Software Requirements and Tools Used

Hardware: 8GB RAM, 64-bit, 7th gen i7 processor. Software: MS-Excel, Jupyter Notebook, python 3.6



**Model/s Development and Evaluation**

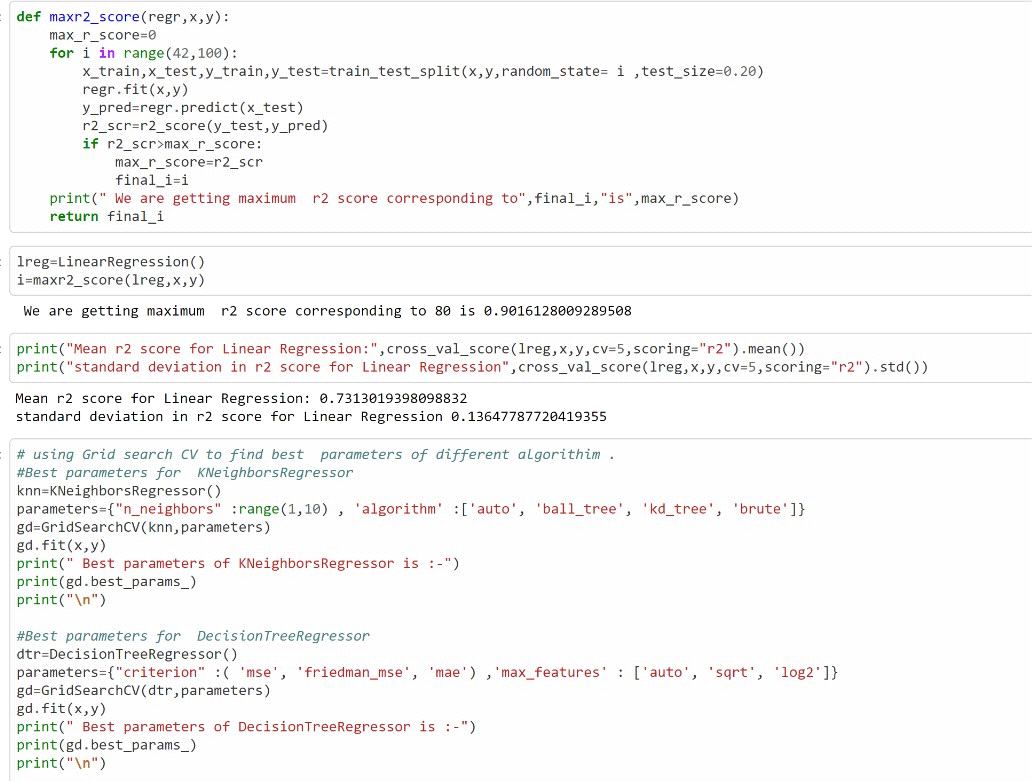
* Identification of possible problem-solving approaches (methods)

As we know that it is a Regression problem in which our output is ‘Sale Price’ so we will use the regression model like linear Regression , KNeighbors Regressor , Decision Tree Regressor, Gradient Boosting Regressor , Random Forest Regressor , Ada Boost Regressor etc we will train our training data using these algorithms and then we will test on the finalised test data set for final house price prediction . The algorithm which is giving better accuracy and Cross value score will be chosen as final model.

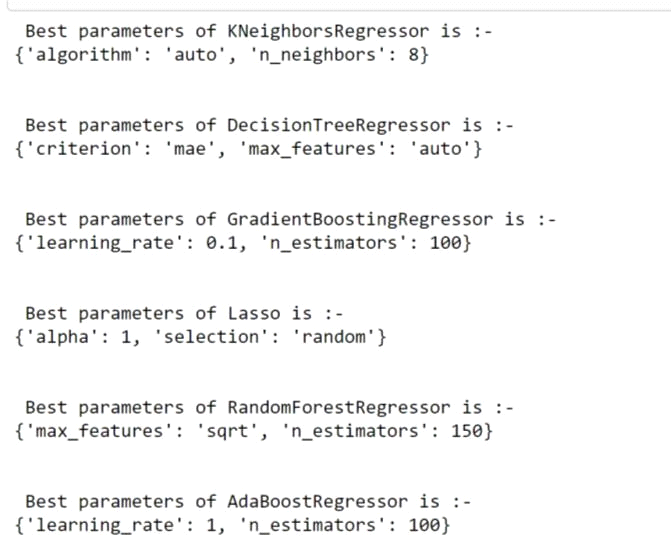
* Testing of Identified Approaches (Algorithms)

Here we will use here KNeighborsRegressor , Decision Tree Regressor , Gradient Boosting Regressor , Lasso , Random Forest Regressor ,Ada Boost Regressor And Linear Regression algorithms.

* Run and Evaluate selected models

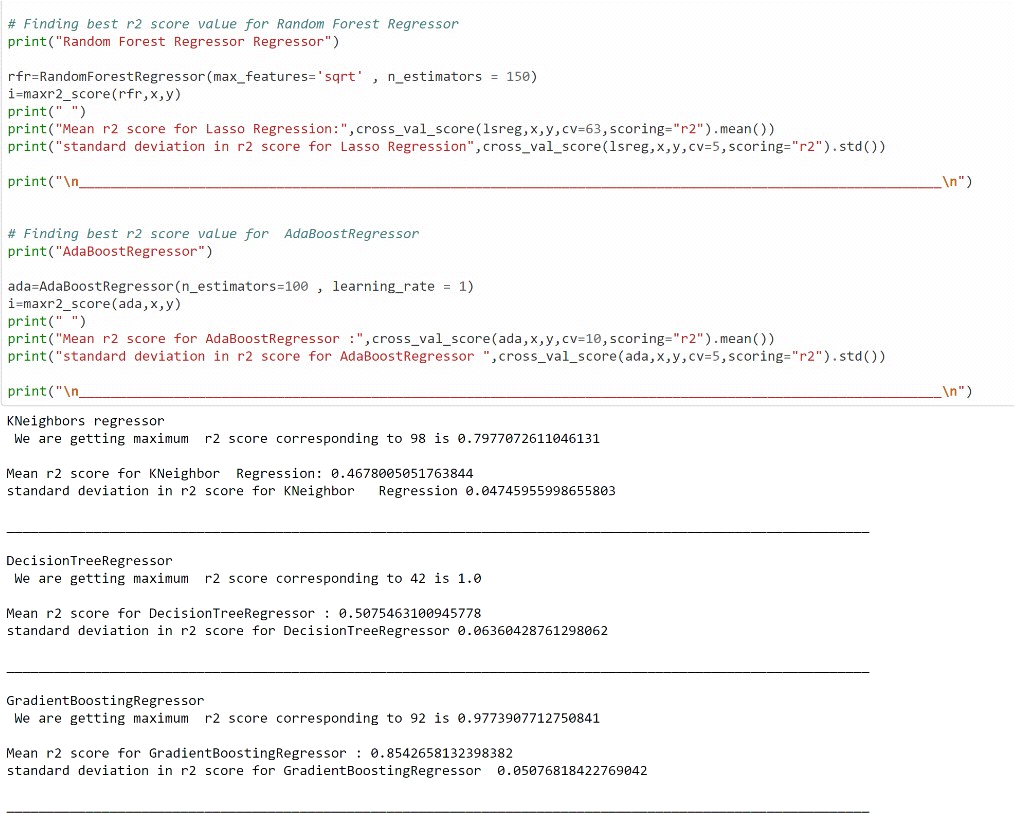
Now we will find best parameters of different algorithms.

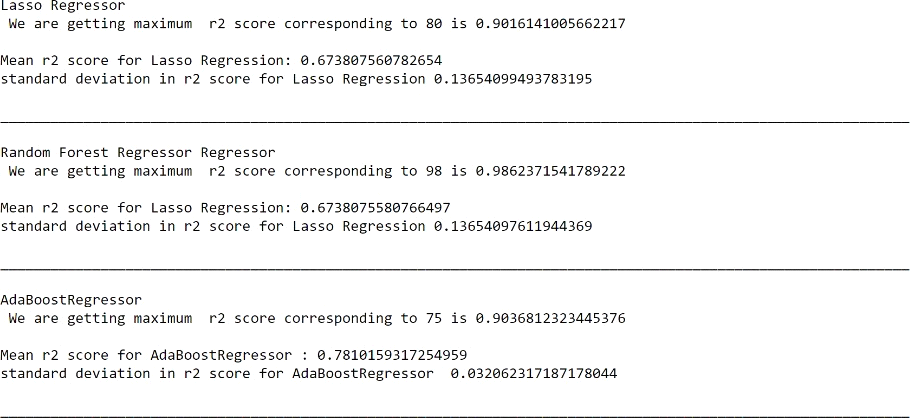




Now we will find the r2 score , cross value score and standard deviation of different algorithm.







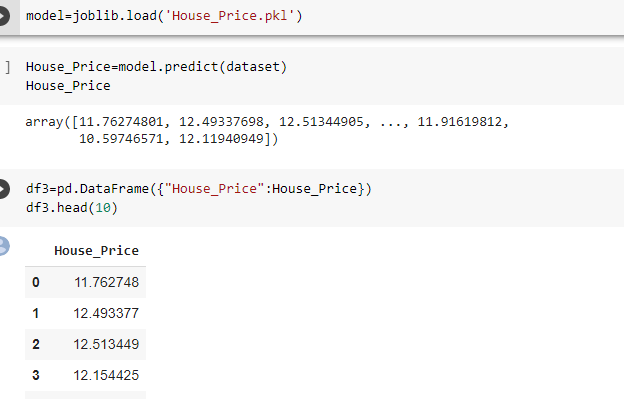
* Interpretation of the Results

From the above running different algorithms we found that Decision tree Regressor is giving better value of r2 score , cross value score and standard deviation so we will choose it as final model.



Now we will predict the sale price for our Test data.

**Using Model to predict House Price**



**CONCLUSION**

* Key Findings and Conclusions of the Study

In this we found that Variables like OverallQual (overall material and finish of the house ) , Year Built , TotRmsAbvGrd ( Total rooms above grade (does not include bathrooms) , GarageCars (Size of garage in car capacity ) ,GarageArea ( Size of garage in square feet ) , GrLivArea ( Above grade (ground) living area square feet ) , FullBath ( Full bathrooms above grade ) have positive relationship with the sales Price and they effect the sales price hence these factors should be considered

* Learning Outcomes of the Study in respect of Data Science

The goal is to achieve the system which will reduce the human effort to find a house having reasonable price. The

proposed system. House Price Prediction model approximately try to achieve the same one. Proposed system focused on predict the house price according to the area for that image processing and machine learning methods are used. The experimental results showed that this technique that are used while developing system will give accurate prediction of house price

* Limitations of this work and Scope for Future Work

There are some more algorithms which can be used and checked if they are giving better results , Using Deep Learning for predicting house price may get some good results.