PROJECT

Predicting Price of House

Prepared by.
Shuvam Singh

Prepared for.
InsAnalaytics

Scope of Work

- **In this project our aim is to build a Machine learning model which will predict the price of houses.**
- **Pre-process the Data.**
    1. Treating Categorical Variables.
    2. Treat NAN values
- **Select Machine learning models.**
- **Create plots for Exploratory data analytics.**

The Features provided in the Dataset are.

1. Local Price:
2. Bathrooms Count
3. Area in Sqft
4. Living Space in Sqft:
5. Size of the living space in thousands of square feet;
6. Garage Count: Rooms Count: Bedrooms Count: Age:
7. Material:
8. Level:
9. Selling Price:
10. Number of garages
11. Number of rooms
12. Number of bedrooms
13. Age in years for the site
14. Brick, or Brick/Wood, or Aluminium/Wood, or Wood. Two story, or Split level, or Ranch
15. The final selling price

Algorithm Used- Linear Regression

In statistics, **linear regression** is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[2]

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, or forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify

which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares functions as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

n statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[2]

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.[3] Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, or forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares functions in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Data Preprocessing

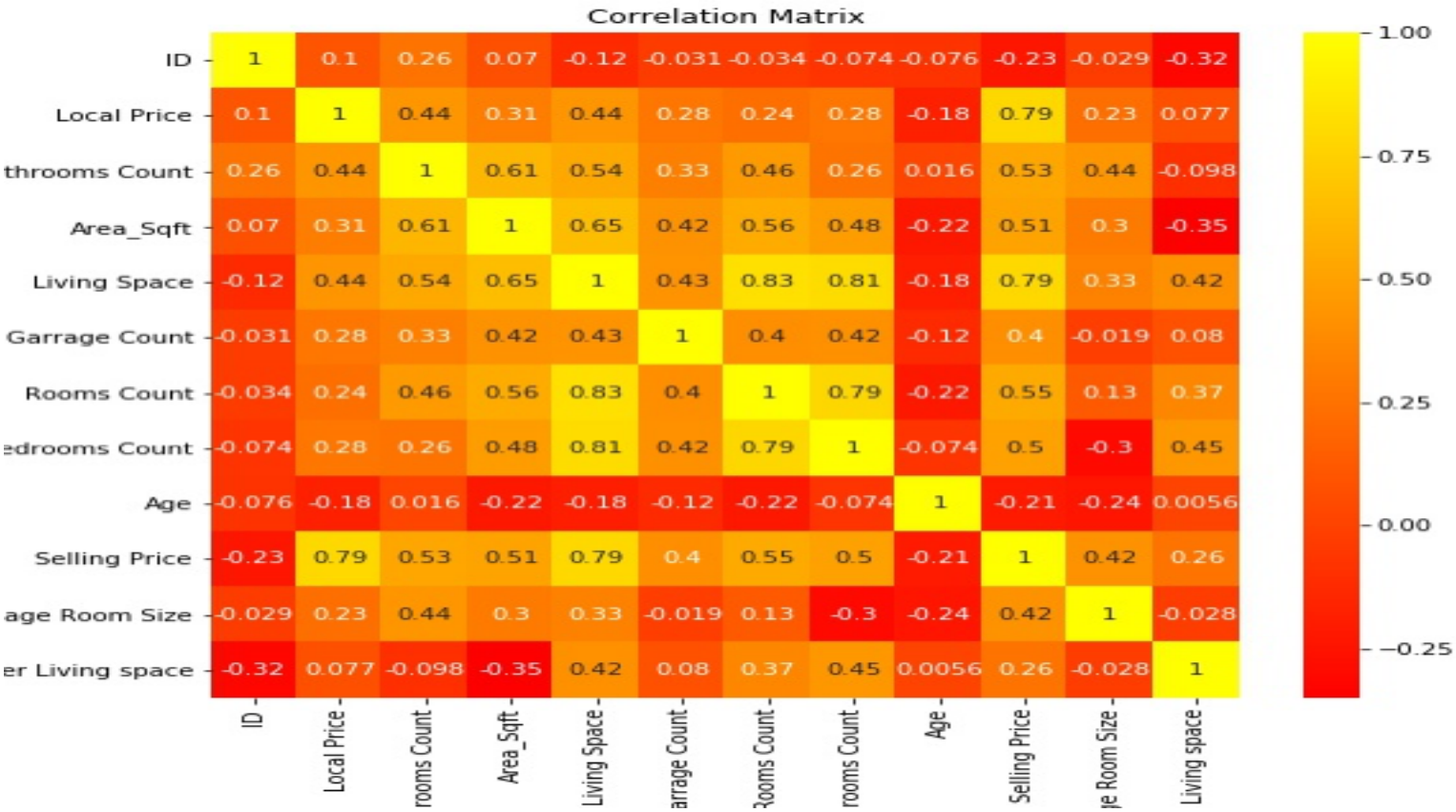
The two attributes In the Datasets were categorical DataTypes. So to input this variables into our Machine learning model, we need to encode it into dummy variables.

The Categorical Variables are:-

- Material
- Level

Coorelation-Matrix

A **correlation matrix** is a table showing **correlation** coefficients between variables. Each cell in the table shows the **correlation** between two variables. A **correlation matrix** is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.



The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no relationship between the movement of the two variables

** The points in the blocks are correlation coefficient.*

Feature Selection

The Correlation matrix is used for Feature selection in our model like the Feature should be in high correlation with Target Variable which is price in our case. And the Feature (Independent Variables) should have less correlation between themselves.

So we will eliminate the those features which have high correlation between themselves and those features which have very low correlation with the target variable.

Splitting into train and test

The data has been splitted into train and test, train dataset is used to train the model and when the trained the test dataset is used to check the accuracy of the model.

Medium of Execution

The execution of algorithm has been done on Python programming using the the below Libraryof has been used for for Plotting Seaborn library is used for Data Manipulation Pandas has been used.

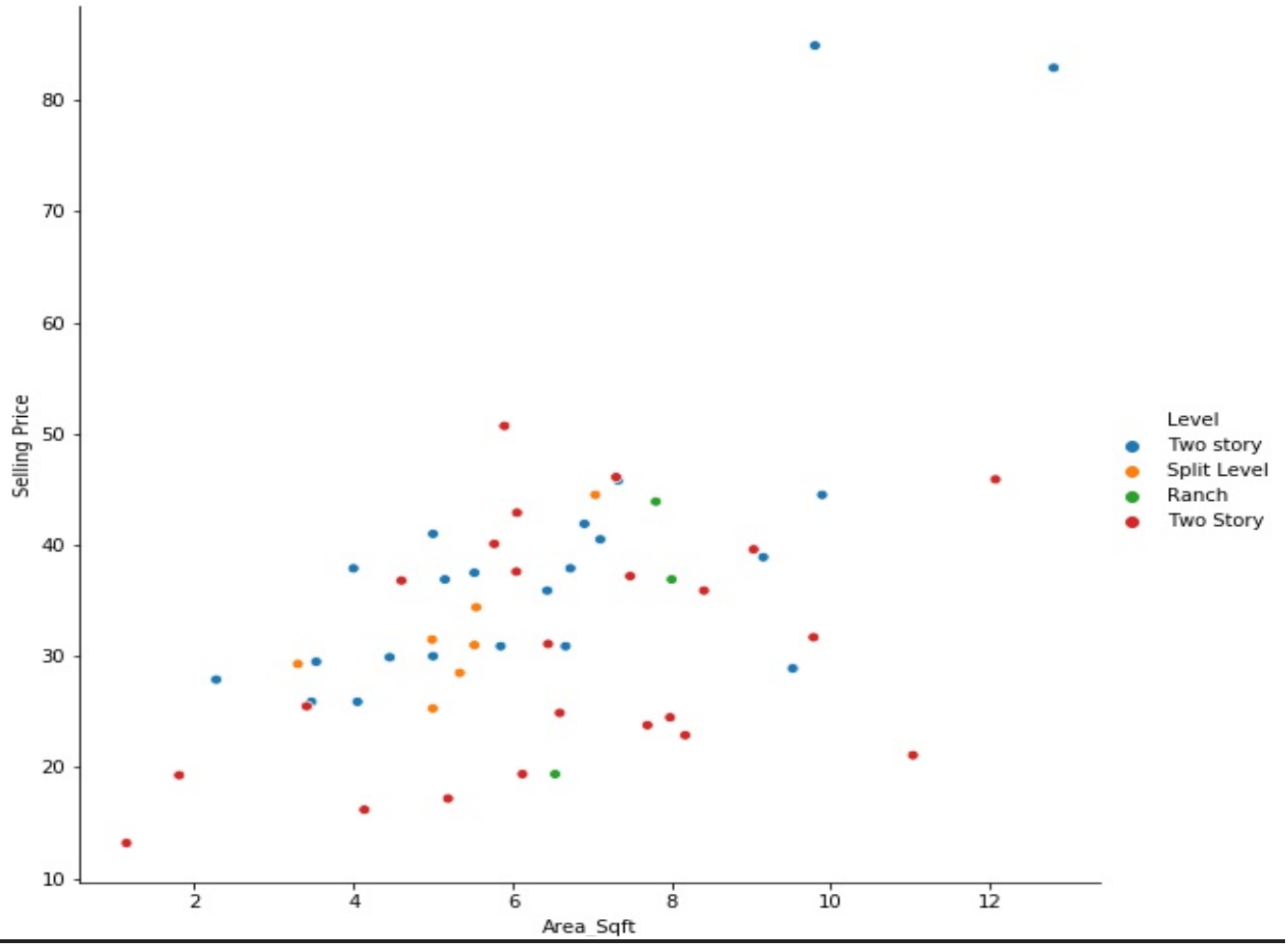
1. ***Scikit-learn-*** *For Algorithm implementation****.***
    - *From sklearn.linear_model- Used for importing Linear Regression Model*
2. ***Pandas****- For Data Manipulation*
3. ***Matplotlib-*** *for plotting the Data*
4. ***Seaborn-*** *For interactive Plots*
5. ***Scipy.stats*** *– For Correlation Coefficients*

Exploratory data analysis.

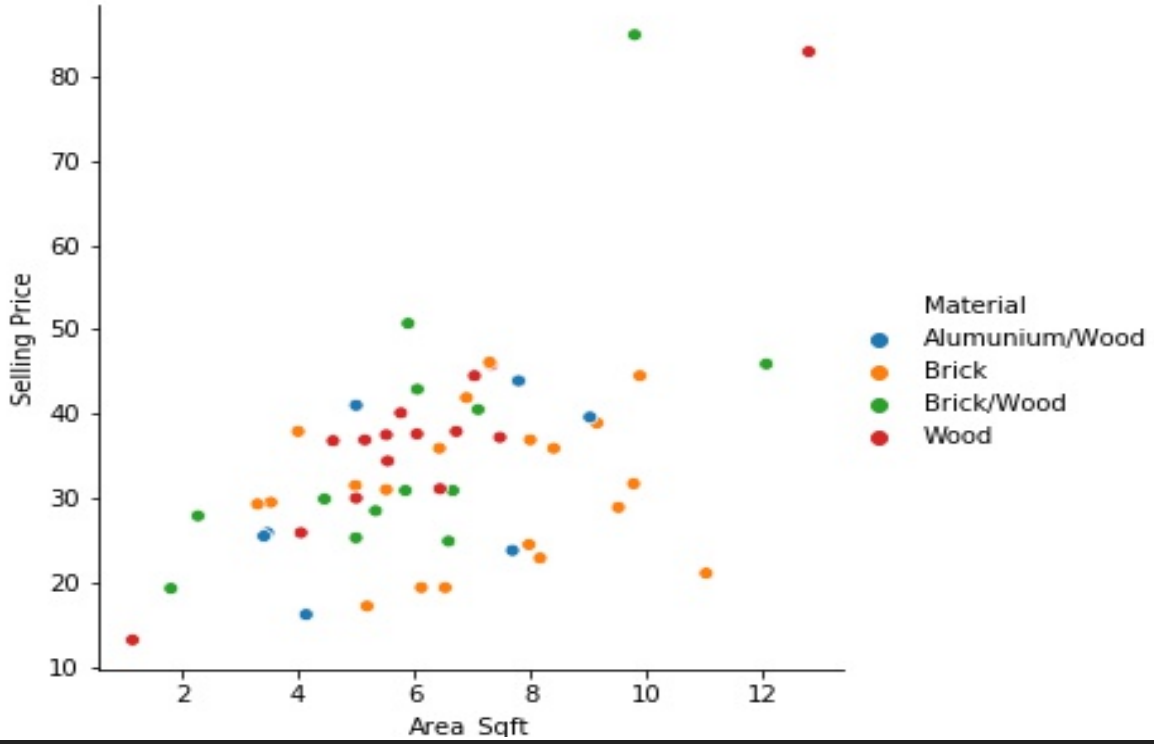
In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task

The Below Plots are made in Python:-

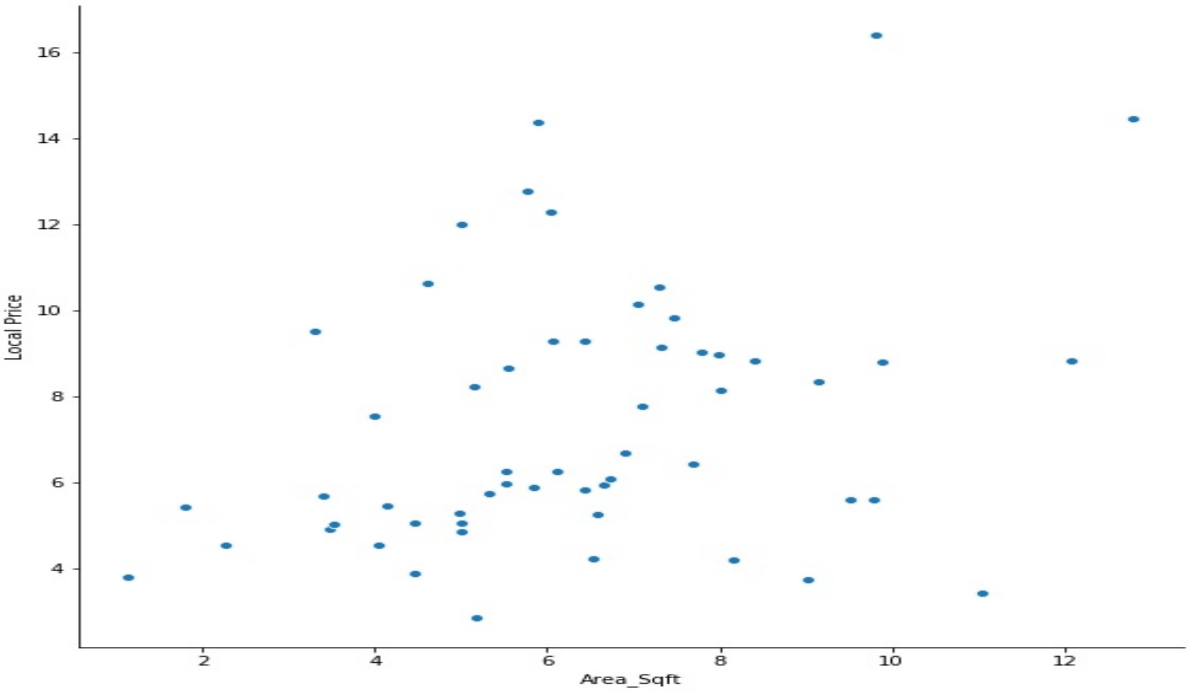
Graphs of Price Vs Area:-



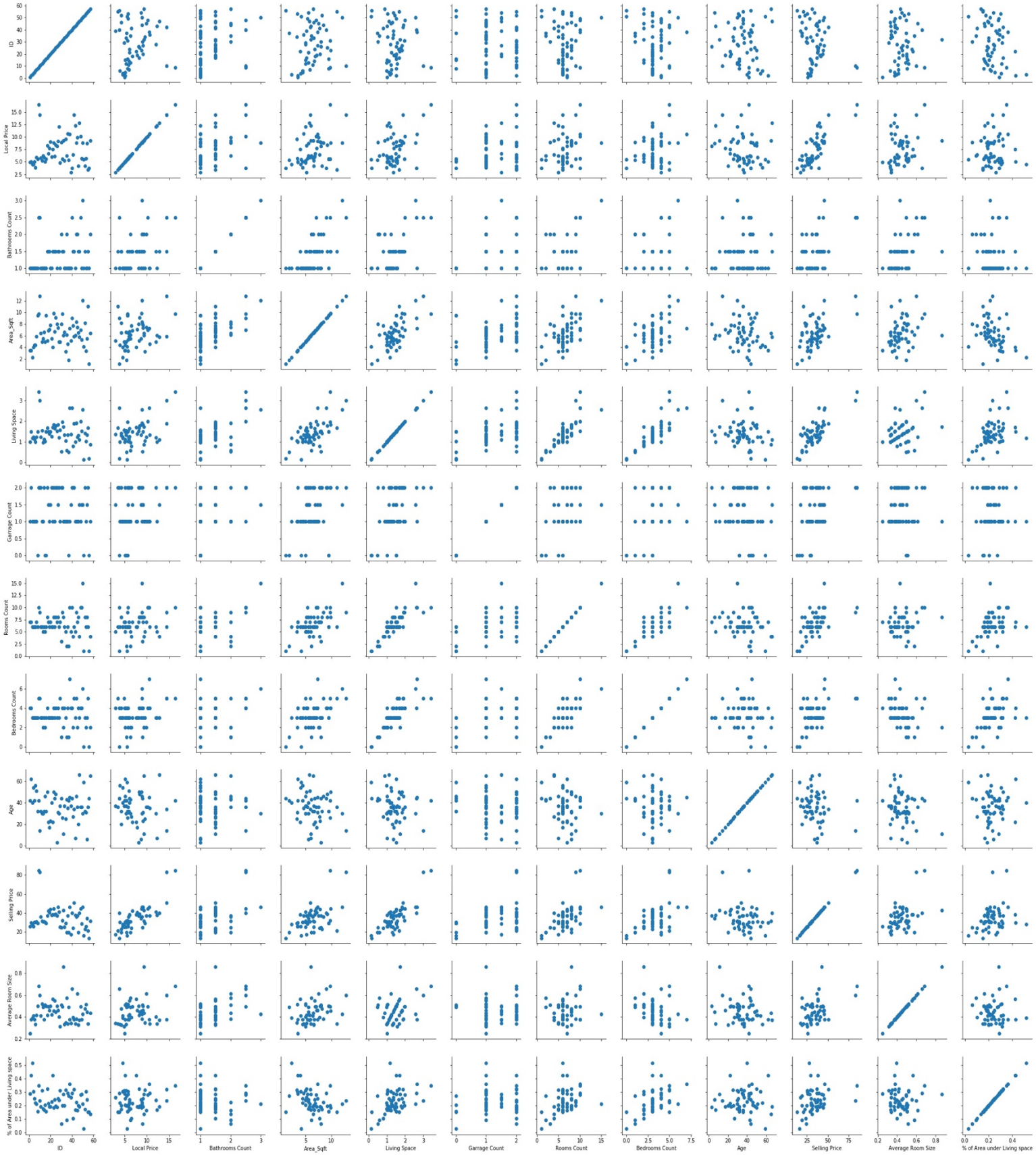
Graphs of Price Vs Area:-



Graphs of Local Price Vs Area Sq.ft:-




Dashboard
Selling Price
80
60
40
20
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70
Age
Material
(All)
Alumunium/Wood
Brick
Brick/Wood
Wood
Area_Sqft
3.301
12.800
Area_Sqft
3.301
12.800
Local Price Vs Level and Material
Level / Material
Ranch
Split Level
Two Story
Avg. Local Price
8
6
4
2
0
Alumuniu..
Brick
Brick
Brick/Wood
Wood
Alumuniu..
Brick
Brick/Wood
Wood


ID
Local Price
Bathrooms Count
Area_Sqft
Living Space
Garrage Count
Rooms Count
Bedrooms Count
Age
Selling Price
Average Room Size
% of Area under Living space

R Sqaured:-

- R Sqaured value of model **0.9028**

Conclusion

- The Avg. Local Price of Split level wood is higher than the all Local price of material and levels.
- Age does not have any influencing power on selling price as per the data.
- R Sqaured value of model **0.9028**
- MSE- 3.067