King County House Price Prediction Report

By:

Yash Chaudhary

GitHub link for repository: https://github.com/yashchaudhary28/HousePricePrediction

Abstract

For this project I choose to predict house prices of King County region in United States.

The purpose of this project was to get an estimate on what factors do the price of house majorly depend and to try and predict price using machine learning algorithm linear regression without any bias to help both seller and buyer make decisions.

Introduction

Property Technology is the next big thing to disrupt the real estate market. It espouses the use of technology to facilitate the management and operations of real estate assets. The assets here mean either buildings or cities. Property Technology is a part of digital transformation in the real estate industry and it focuses on both the technology and mentality changes of the people involved in this industry. It can also lead to new functionalities such as more transparency, which were not possible earlier. Big data, data analytics, machine learning form a part of PropTech.

Being a civil engineer, I have always wondered how technology can be used in real estate and so I decided to work on one of its use case which is to predict the price of the houses in King County region of United States.

The relationship between house prices and the economy is an important motivating factor for predicting house prices. There is no accurate measure of house prices. A property's value is important in real estate transactions. House prices trends are not only the concerns for buyers and sellers, but they also indicate the current economic situations. Therefore, it is important to predict the house prices without bias to help both buyers and sellers make their decisions. Houses tend to have various competing features which makes it difficult to decide on the right price.

Objective

The goal of this project is to create a regression model that can accurately estimate the price of the house given the features.

Methodology

Dataset Description

The real estate housing data is used and it is taken from https://data.kingcounty.gov/ . This dataset includes homes that were sold between May 2014 and May 2015 in King County, Seattle. It has a total of 21613 observations with 20 features including the target variable ‘price’.

Feature Description

It has a total of 20 features out of which 7 are categorical features and 11 numerical features. One is date and other is id. We will describe each feature as:

id	Unique numeric number assigned to each house being sold.
date	Date on which the house was sold out.
price	Price of house, this is our target variable.
bedrooms	Number of bedrooms in a house.
bathrooms	Number of bathrooms in a bedroom of a house.
sqft_living	Total area of house in square foot.
sqft_lot	Total area of lot in square foot.
floors	It determines total floors means levels of house.
waterfront	Whether a house has a view to waterfront, 0 means no, 1 means yes.
view	Whether a house has been viewed or not, 0 means no, 1 means yes.
condition	Overall condition of a house on a scale of 1 to 5.
grade	Overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11
sqft_above	Square footage of house apart from basement.
sqft_basement	Square footage of the basement of the house.
yr_built	Date of building of the house.
yr_renovated	Year of renovation of house.
zipcode	Zip Code of the location of the house.
lat	Latitude of the location of the house.
long	Longitude of the location of the house.
sqft_living15	Living room area in 2015(implies-- some renovations)
sqft_lot15	Lot Size area in 2015(implies-- some renovations)

Dataset Division

The dataset has been split into train, validation, and test, with the Test data having 2217 observations, as against train and validation data having 9761 and 9635 observations, respectively. We train the model on the Train data. Evaluate several models on the Validation data. The final model is then used for predicting the target variable (Price) for the Test data.

Data Cleaning and Integration

Data cleaning is an iterative process, the first iterate is on detecting and correcting bad records or mistyped records. Before loading into the machine learning models the data should be corrected to get the high accuracy of prediction. In our data we found one record which specified 33 bedrooms with 1.75 bathrooms that is feasibly not possible for a house. So, we corrected it to 3. The dataset did not have any missing and null values.

Tools

- Python
- jupyter Notebook

Libraries

- NumPy
- Pandas
- Seaborn
- Matplotlib.pyplot
- Sklearn

Algorithms

- Regression: It is a data mining task of predicting the value of target (numerical variable) by building a model based on the one or more predictors the predictors can either be numerical or the categorical variables.
    - Simple Linear Regression
    - Multiple Linear Regression

Implementation

The main aim of this project is to accurately predict the price of the house of King County, Seattle without any bias. The below segment blankets will help you to know the implementation process in depth. Here step by step process involved is represented below.

- Scientific Environment.
- Source of Data.
- Loading data into Python using jupyter notebook.
- Data Cleaning
- Analysis and Visualization using python
- Feature Engineering
- Feature Scaling
- Model Building

Scientific Environment

There is a need for a technical environment for the mentioned processes:

- Python for creating scripts
- Using jupyter notebook for easy implementation of python.
- Essential libraries to be installed in notebook.

Data Source

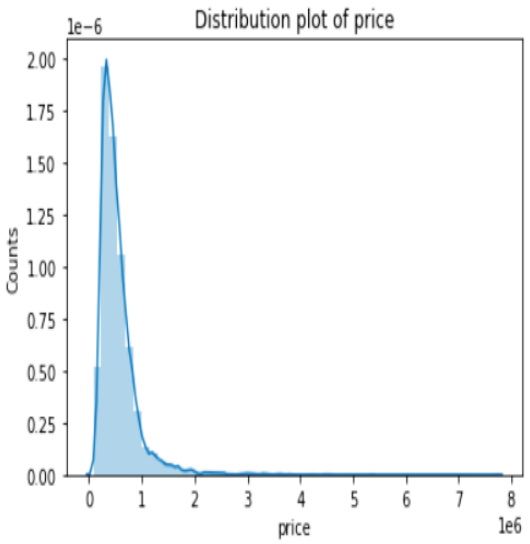
As I mentioned before, the main data set was collected from https://data.kingcounty.gov/ website which is open data resource for the data mining and for the predictive analytics purposes. The acquired data source was a CSV file which was further divided into 3 parts.

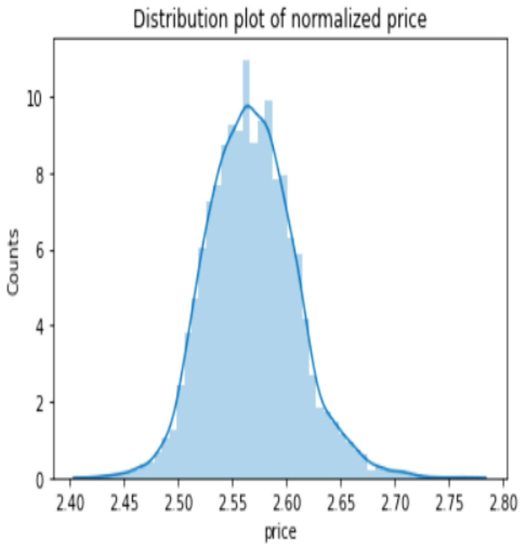
Data Cleaning

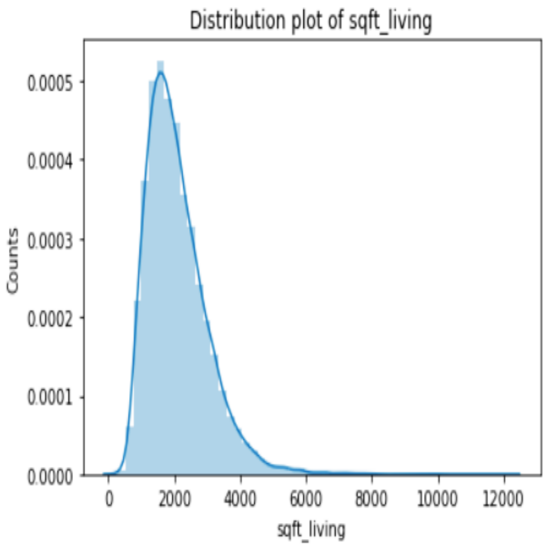
The data is first imported into the jupyter notebook and then using python we will check various attributes of data like shape, size of data. The data in the CSV files need to be checked whether it has any missing values, using the filters and null values are removed which helps to increase the accuracy level. It also includes detecting and correcting anomalies if possible.

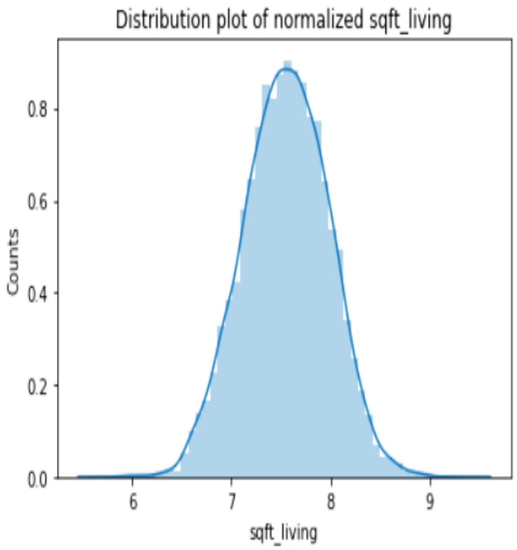
Analysis and Visualization using Python

After checking for missing values and correcting it we will try and visualize our data, by plotting bar charts for categorical data and distribution plots for numerical data. From plotting the distribution charts, we got to know that our continuous variable data I skewed so we normalize our data using log transformation.

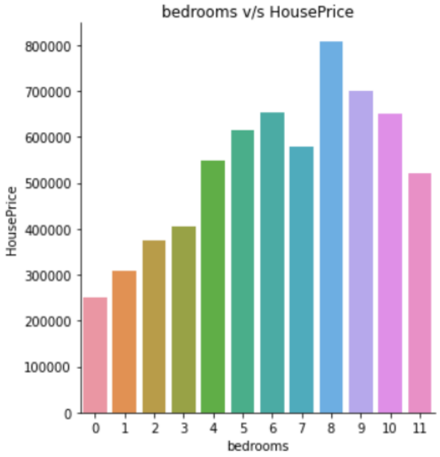


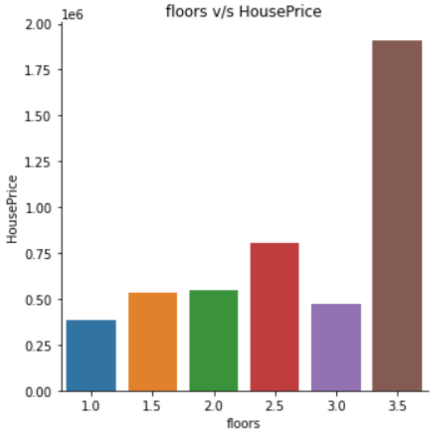


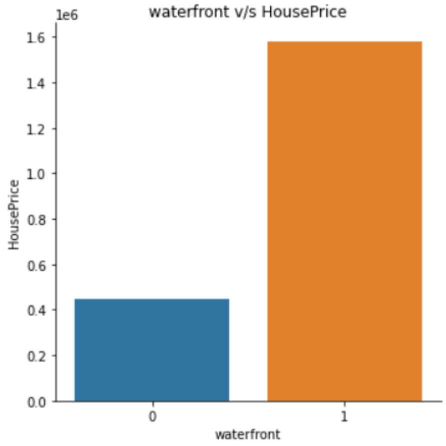


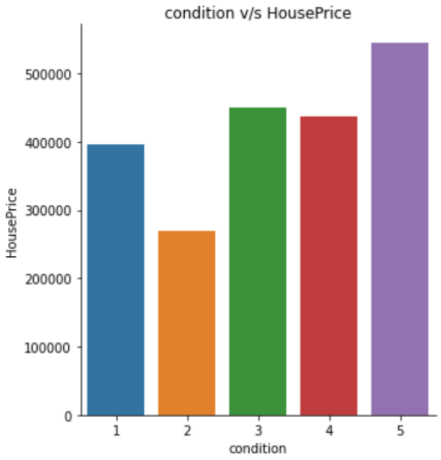


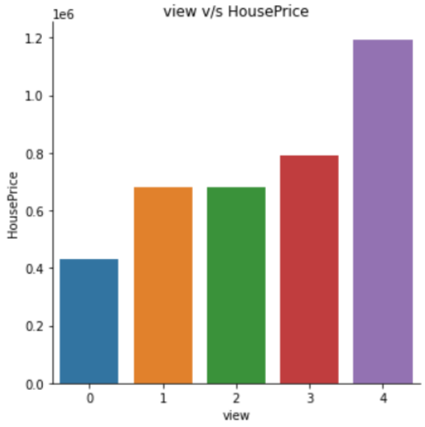
After that we will compare our features with the target variable by plotting bar plots for all our categorical variables.

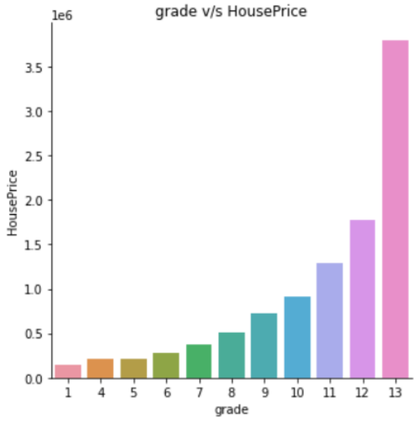












Conclusions made from these plots:

→Up to 6 bedrooms we can say that price of house increases as the number of bedroom increases.

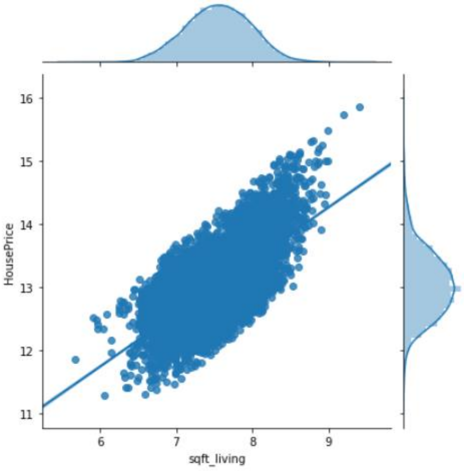
→ Houses with floor 3.5 are the most expensive.

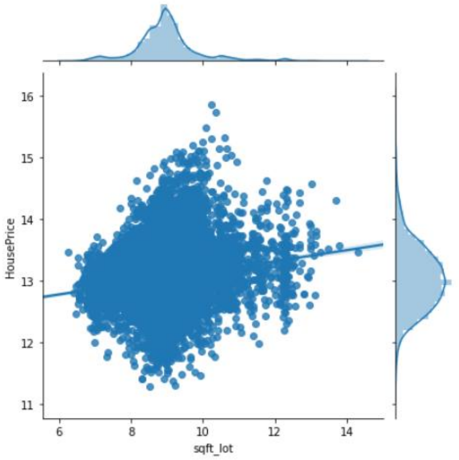
→House with a waterfront are costlier than without waterfront.

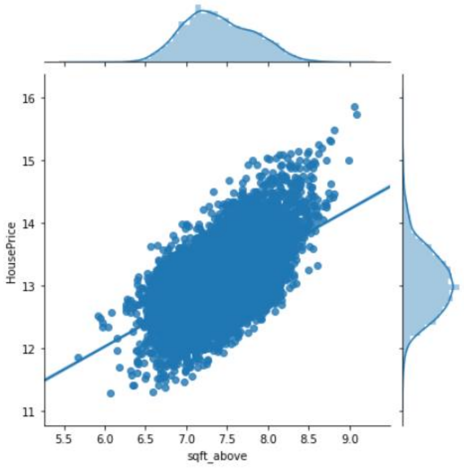
→As the number of views of the house increase its cost increases. The house with 4 views is most expensive in its category.

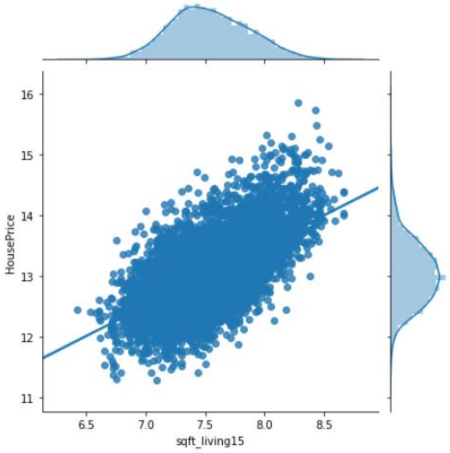
→We can see a clear increase in the price of the houses as the grade of the house increases.

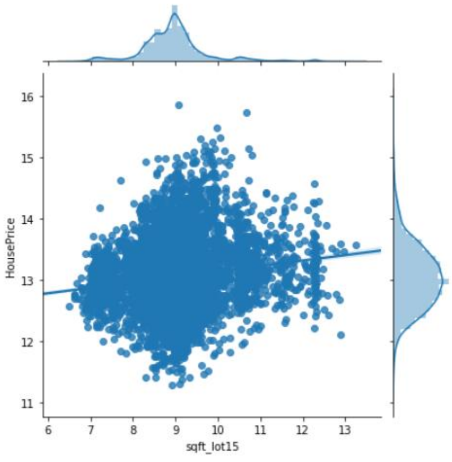
Now we will compare our continuous numerical variable with the target variable using scatter plots.

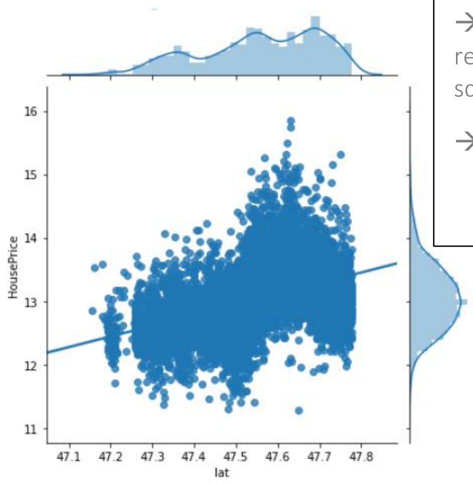












Conclusions made from these plots:

→There is a strong positive relationship between sqft_living and price.

→ There is a weak positive relationship between sqft_lot and price.

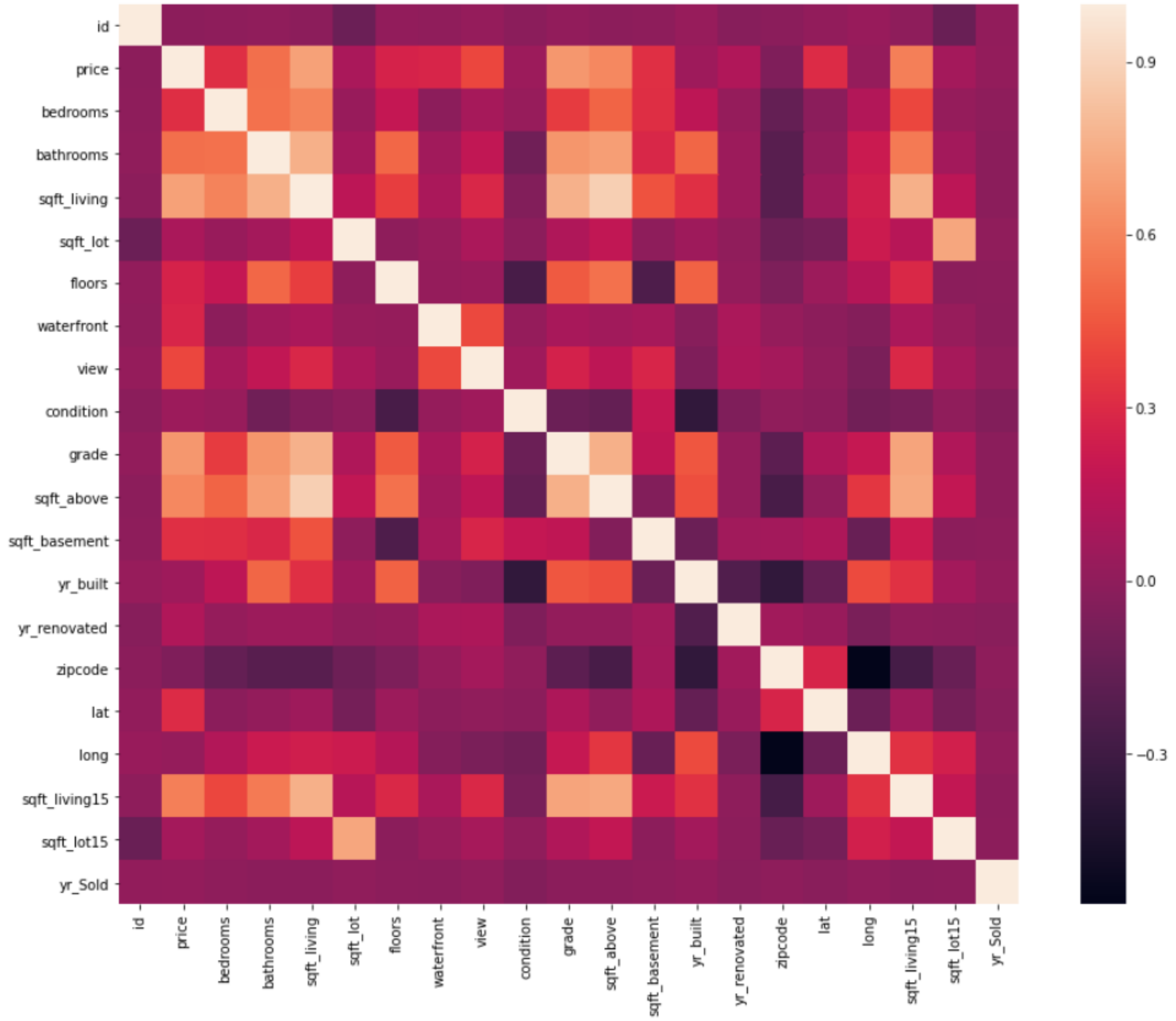
→There is a positive relationship between sqft_above and price.

→There is a strong positive relationship between sqft_living15 and price.

→There is a weak positive relationship between sqft_lot15 and price.

→There is a weak positive relationship between latitude and price.

Finding relation between each feature using correlation matrix which is visualized using seaborn heat plot.



Feature Engineering

In this step we will try and make some new features which may have better impact on our model. We will find the age of the house by subtracting year built with present year. We will also convert the feature year renovated to 2 outputs such as 1 house was renovated and 0 as it was not. Also, how much area of house and area of lot changed from 2015 and the year built. Then we will convert these features to binary 0 and 1.

Feature Scaling

In this step we will do normalization or standardization of our features. As we know our features are on various scales like sqft_living in square foot and bedrooms in numbers. So, it is important to convert all these features on one scale. We will use MinMaxScaler from sklearn library to normalize our features. Normalization helps in better accuracy with the model.

Model Building for Simple Linear Regression

First, we will apply simple linear regression and predict the price by using only one feature in one model. So, we will make as number of models as our features and with each one of them we will predict price using single variable.

Using single feature, we do not get good accuracy in our model and the maximum accuracy that we got was from sqft_living it had an r-squared score of 0.34 which not low.

Model Building for Multiple Linear Regression

We removed sqft_above, yr_built and zipcode before this process as sqft_above had high correlation with sqft_living and for yr_built we created a new feature as age. Now using all other features, we will build a model and see how it performs on the validation data.

Now to further improve our accuracy of our model we did best subset selection method and by using for loop we looped through all the available combinations of our features and calculated r-squared for each one of them and stored them in a list. Now we found out which of these models had the best r-squared score. The model selected through this was the subset of all the possible combinations possible between our features.

Then we build a model of multiple linear regression using features selected by our best model selection algorithm.

Evaluation

The idea of a regression is to predict a real value which means number in regression model we can compute the several values the most common terms are explained below

Coefficient of determination – R^2

The coefficient of determination R square summarizes the explanatory power of the regression model and is computed from the sum of squares terms. The R square describes the proportion of variance of the dependent variable explained by the regression model and the equation is given below

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Adjusted R-squares

It measures the proportion of variance explained by only those independent features that really help in explaining the dependent variable. It penalizes you for adding independent variables that do not help in predicting dependent variable. The equation of adjusted r squared is given below

$$\mathit{Adjusted}\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Evaluation of result

Model	R^2 score	Adjusted R^2 score
Simple Linear Regression	0.3304	0.3304-
Multiple Linear Regression with all features	0.5516	0.5508
Multiple Linear Regression with best features	0.6982	0.6978
Multiple Linear Regression without outliers	0.6988	0.6984
Multiple Linear Regression without High leverage point	0.6982	0.6978

There is not much improvement in our score after removing outliers or high leverage point.

Now, using this model we will predict the value for our test data.

The result for our test data is as follows:

R^2 score	0.6820
Adjusted R^2 score	0.6970

Conclusion

The main goal of this project is to predict the prices, which we have successfully done using machine learning algorithms like multiple regression. So, the multiple regression using best subset of features without outliers gave the maximum accuracy prediction when compared to the others. So, I would believe this work will be helpful for both the buyer and the seller to make their decision.

Future Works

It would also be a good experiment to try joining the data to images. None of the given feature describe how the house looks to potential buyers and many more attributes could be used that affect the price of house.

Price prediction can be improved by adding many attributes like surroundings, marketplaces, and many other related variables to the houses. The predicted data can be stored in the databases and an app can be created for the people so they can invest by checking all the data and feel safer.