Real Estate Valuation Prediction

Deep Narendrabhai Mer (04)

M.Sc. Applied Statistics, Sardar Patel University, Anand

Acknowledgement

- In the accomplishment of completion of my project on House price prediction, I would like to express my special thanks of gratitude to Mr. Das Sir as well as Dr. Jyoti Divecha HOD, Department of Statistics.
- It is because of the knowledge and skills acquired during the course work, along with his best style of teaching, that we are able understand the subject in a better way and are able to complete this project successfully.

Introduction

- The Real Estate valuation is affected by various factors. It is complex function of various geographical factors, economic factors, Real Estate condition etc.
- The given Real Estate data from New Taipei City, Taiwan. In this data consist transaction date**,** house age, distance to the nearest MRT station, geographic coordinate(latitude and longitude).
- We have analysed the effect of each of these factors on the prices of houses and have tried to construct best regression models based on these factors and included models that were found significant.

Problem Statement

Examining real estate valuation helps understand where people tend to live in a city. The higher the price, the greater the demand to live in the property. Predicting real estate valuation can help urban design and urban policies, as it could help identify what factors have the most impact on property prices. Our aim is to predict real estate value, based on several features.

Our objective is to:

1. Understand the data available
2. Test best regression models
3. Assess the best model and improve them
4. Present the results and address the results

Attribute Information:

The inputs are as follows

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (Unit: degree)

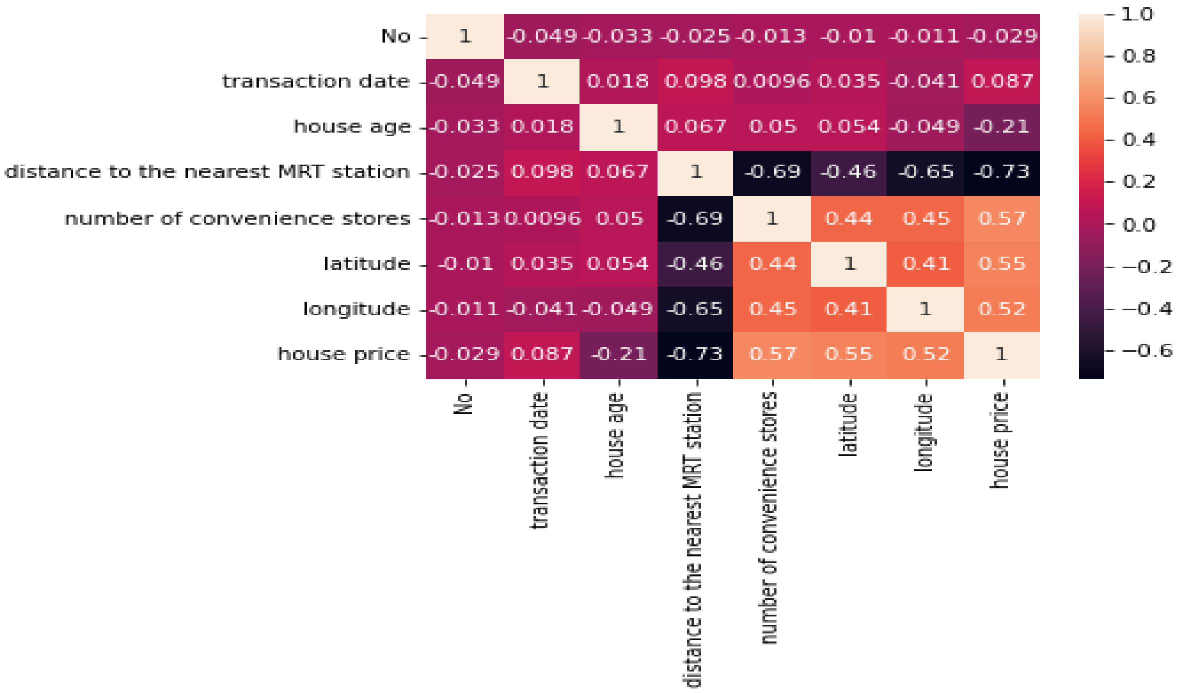
X6=the geographic coordinate, longitude. (Unit: degree)

The output is as follow

Y= house price of unit area

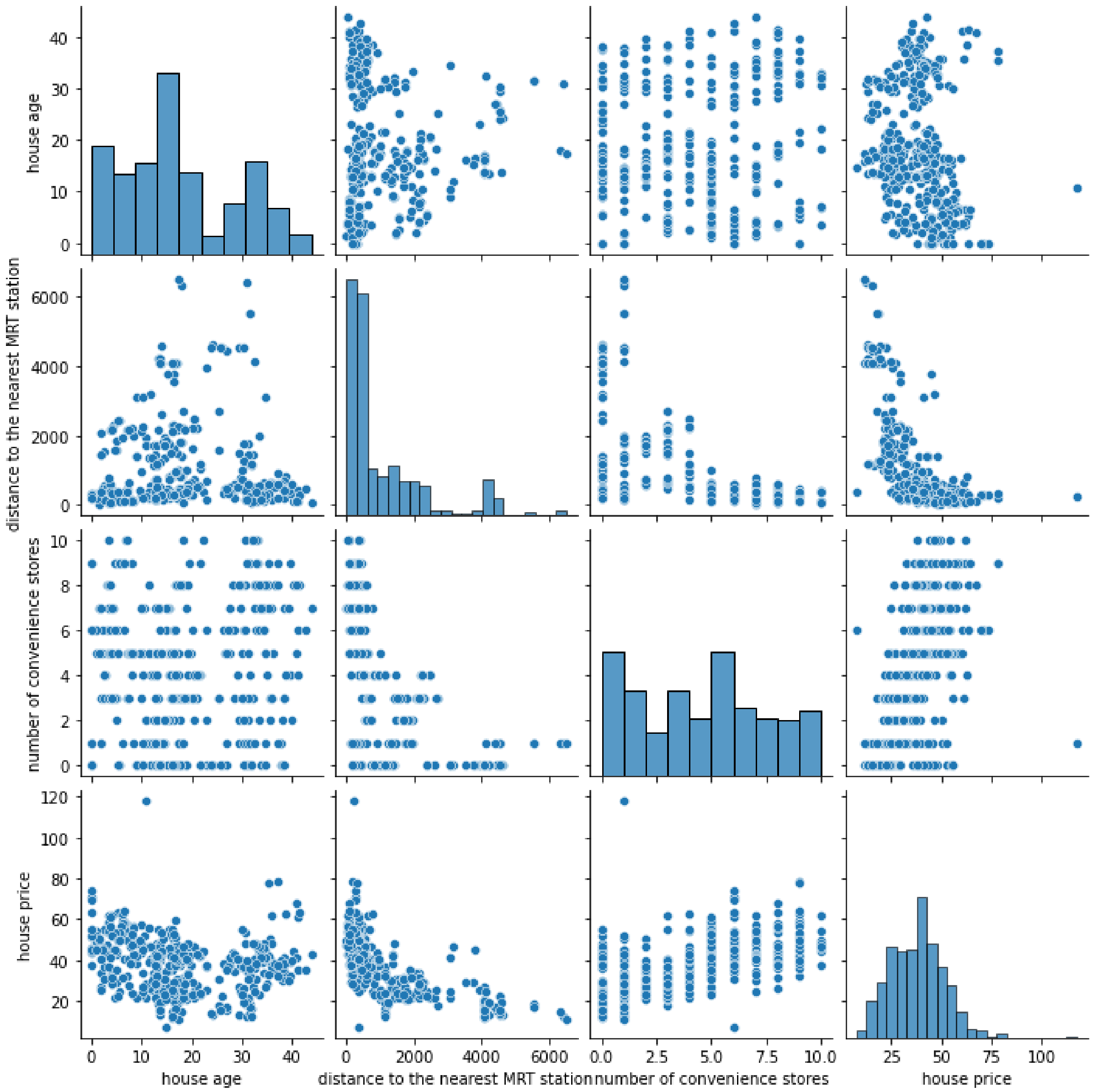
ANALYSIS

After loading the data, we performed this method by comparing our target variable (house price of unit area) that is decision to be taken with other independent variables. By performing heatmap of corelation first we have to use all variables



From this heatmap, it can be concluded that there is no perfect multicollinearity but there is imperfect multi collinearity, so we must drop the column 'transaction date’, 'latitude’, 'longitude' because these variables are not affected more so we drop.

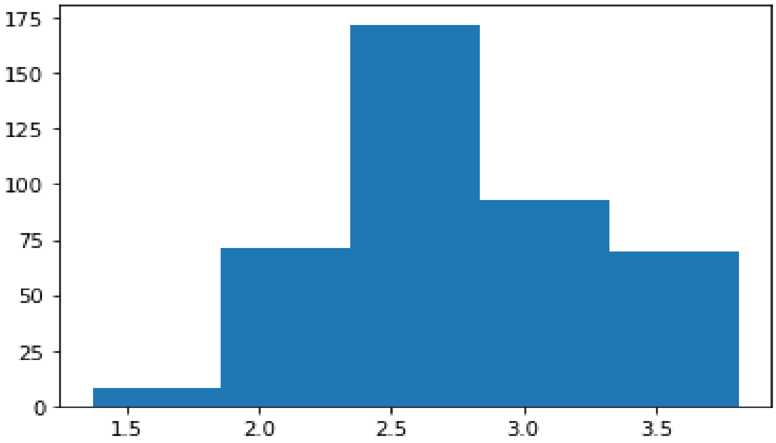
By using pair plots of 'house age', 'distance to the nearest MRT station ','number of convenience stores ','house price'



From the above charts, it can be it can be assumed that variables house price, house age and number of convenient stores are symmetric and i.e., normal. The variable distance to the nearest MRT station seems to be positively skewed.

So, we can use log transformation and make normal

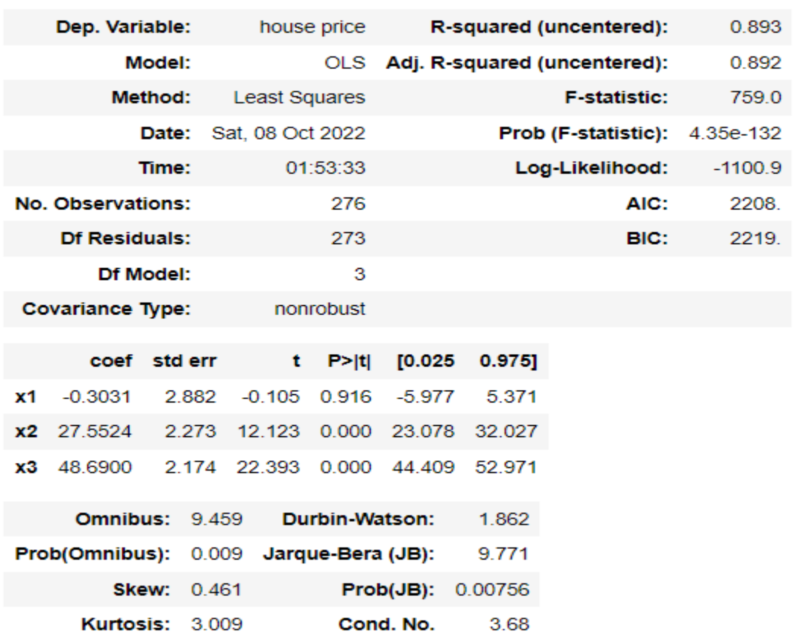
From the chart it can be drawn that the variable distance to the nearest MRT station to be symmetric in nature.



Fitting models:-

For model fitting use we will use Python's **statsmodels** module to implement **Ordinary Least Squares** (**OLS**) method of linear regression.

A linear regression model establishes the relation between a dependent variable(**y**) and at least one independent variable(**x**) as: In *OLS* method, we must choose the values of and such that, the total sum of squares of the difference between the calculated and observed values of y, is minimised.



```
   Dep. Variable:      house price      R-squared (uncentered):      0.893
          Model:               OLS  Adj. R-squared (uncentered):      0.892
         Method:     Least Squares                  F-statistic:      759.0
           Date:  Sat, 08 Oct 2022            Prob (F-statistic):  4.35e-132
           Time:          01:53:33               Log-Likelihood:    -1100.9
No. Observations:              276                          AIC:      2208.
    Df Residuals:              273                          BIC:      2219.
        Df Model:                3
 Covariance Type:        nonrobust

       coef    std err      t     P>|t|   [0.025   0.975]
x1   -0.3031    2.882   -0.105   0.916   -5.977    5.371
x2   27.5524    2.273   12.123   0.000   23.078   32.027
x3   48.6900    2.174   22.393   0.000   44.409   52.971

       Omnibus:  9.459     Durbin-Watson:    1.862
Prob(Omnibus):  0.009  Jarque-Bera (JB):    9.771
          Skew:  0.461          Prob(JB):  0.00756
      Kurtosis:  3.009         Cond. No.     3.68
```

Model can be given as,

House price = -0.3031*(House age) +27.5524*(distance to the nearest MRT station) + 48.69*(number of convenient stores)

From this summary output of regression model, it can be concluded that model is statistically significant. The R-squared for the model is 0.892 i.e., Around 89% of total variation in the response house price can be explained by the predictor variables.

From summary output, it can be drawn that house age is not significantly contributing in the prediction of house price.

Conclusion:

We get a best model of the R-squared for the model is 0.892 which can good for given dataset. This performance could be due to various reason of not enough relevant features.

GitHub repository

R:

https://github.com/TommyShelby05/ML_Assignments/blob/c8dc96b97a7a75b4599fa96a13697e18c09e1dbb/House-Price_Prediction.R

Python:

https://github.com/TommyShelby05/ML_Assignments/blob/c8dc96b97a7a75b4599fa96a13697e18c09e1dbb/House-Price_Prediction.ipynb