Predicting Real Estate Prices based on linear regression

Group – A6

Team Members:

Saksham Bansal

Ashley Lazarus

Paramjeet Kour Na

CONTENTS

Problem Statement

The old way of predicting the price of the house was based on cost and sales price comparison lacking an overall accepted standard process. With so much money involved in buying and selling a house for personal use, business or investing in one, one must have some serious model to calculate the price of the house. Therefore, it can be an important decision for both households and enterprises. How to build a realistic model of accurately predicting the price of real estate has become a challenging topic with great potential for further research. Scholars often believe that predicting the exact cost of a particular property is impossible because it involves several factors that ultimately affect costs.

The goal of this statistical analysis is basically to find the relationship between house features and how these variables help us to predict the price of the house

Introduction

House pricing is a complex thing to calculate as there are so many variables affecting them. In this study, we first analyze the major factors affecting housing prices with linear regression, selects significant factors influencing general housing prices, and conducts a combined analysis algorithm. Then, the project establishes a linear regression model for housing price prediction and applies the data set of real estate prices in Sindian district, New Taipai City, Taiwan to test the method. Through the data analysis and test in this project, it can be summarized that the multiple linear regression model can effectively predict and analyze the housing price to some extent, while the algorithm can still be improved through more advanced machine learning methods.

Data Set

As a successful businessman in the property industry, what makes him attach so much significance to some specific factors like location when appraising a property is crucial. To what extent a particular factor like location plays an essential role in pricing a property is worth exploring by adopting a statistical model in real estate economics research. There are six major factors that determine the price of a property.

1. Response variable and predictor variables:

   $Y=B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6$

   Where B's denote the regression coefficients.

2. The response (Y) variable:

House price of unit area: This is the final price of a house at which it can be bought or sold.

3. The predictor variables X:

   Transaction Date (X1): The last date of transaction.

   House Age (X2): The age of the house i.e., when the house was built in years.

   Distance to the nearest MRT station (X3): The distance to the nearest MRT station.

   Number of convenience Stores (X4): The number of convenience stores in the area.

   Latitude (X5): Latitude of the house.

   Longitude (X6): Longitude of the house.

Data Source

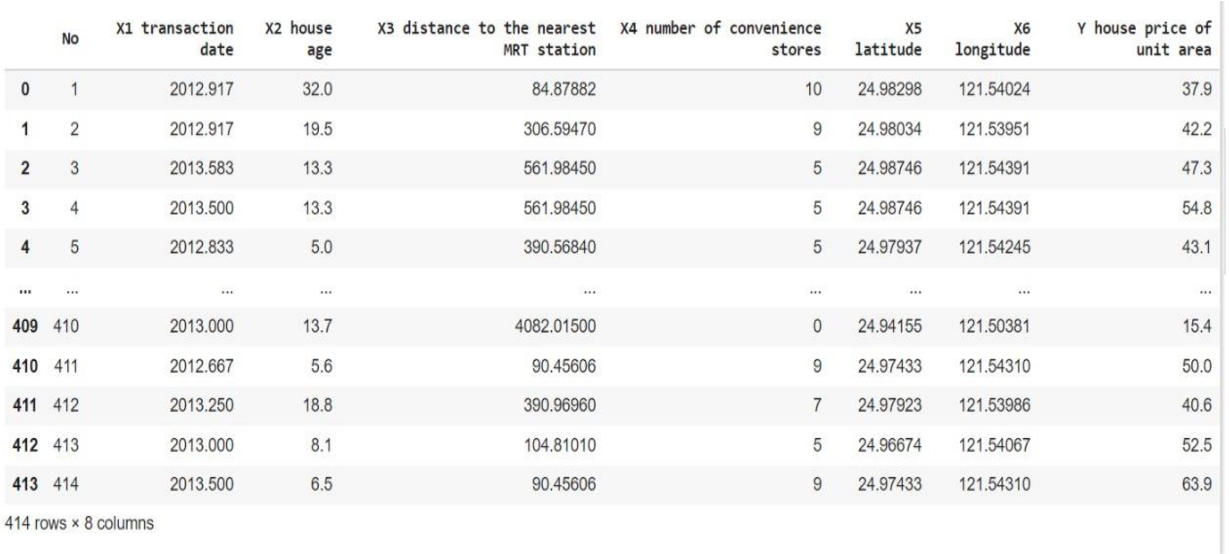
We used the data set from the website named Kaggle https://www.kaggle.com/quantbruce/real-estate-price-prediction

Data Features

The data includes various factors which will be used to predict the pricing of a real estate. The factors are written in the columns in the following picture-

- X1 Transaction date- The transaction date is a date upon which a trade takes place for a security or other financial instrument. The transaction date represents the time at which ownership of the house officially transfers.
- X2 House age- This is the time measure since the house has been made.
- X3 Distance to the nearest MRT station- MASS rapid transit (MRT) system is a rail system which is used for transporting passengers in urban areas. It is known by various other names such as mass transit, subway, underground railway, or metro. It is the distance between the house and the MRT station.
- X4 Number of convenience stores- A convenience store, convenience shop, corner store, or corner shop is a small retail business that stocks a range of everyday items such as coffee, groceries, snack foods, confectionery, soft drinks, tobacco products, lottery tickets, over-the-counter drugs, toiletries, newspapers, and magazines. This is the number of convenience store near the house.
- X5 Latitude- Latitude of the house

- X6 Longitude- Longitude of the house



	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1
...	...	...	...	...	...	...	...	...
409	410	2013.000	13.7	4082.01500	0	24.94155	121.50381	15.4
410	411	2012.667	5.6	90.45606	9	24.97433	121.54310	50.0
411	412	2013.250	18.8	390.96960	7	24.97923	121.53986	40.6
412	413	2013.000	8.1	104.81010	5	24.96674	121.54067	52.5
413	414	2013.500	6.5	90.45606	9	24.97433	121.54310	63.9

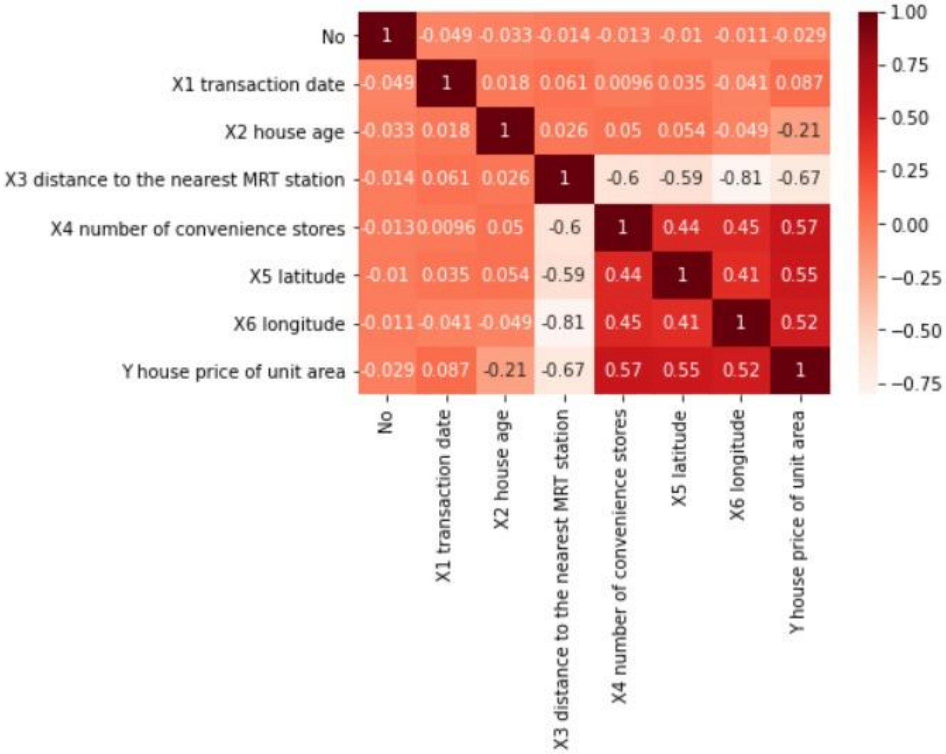
414 rows × 8 columns

Correlation

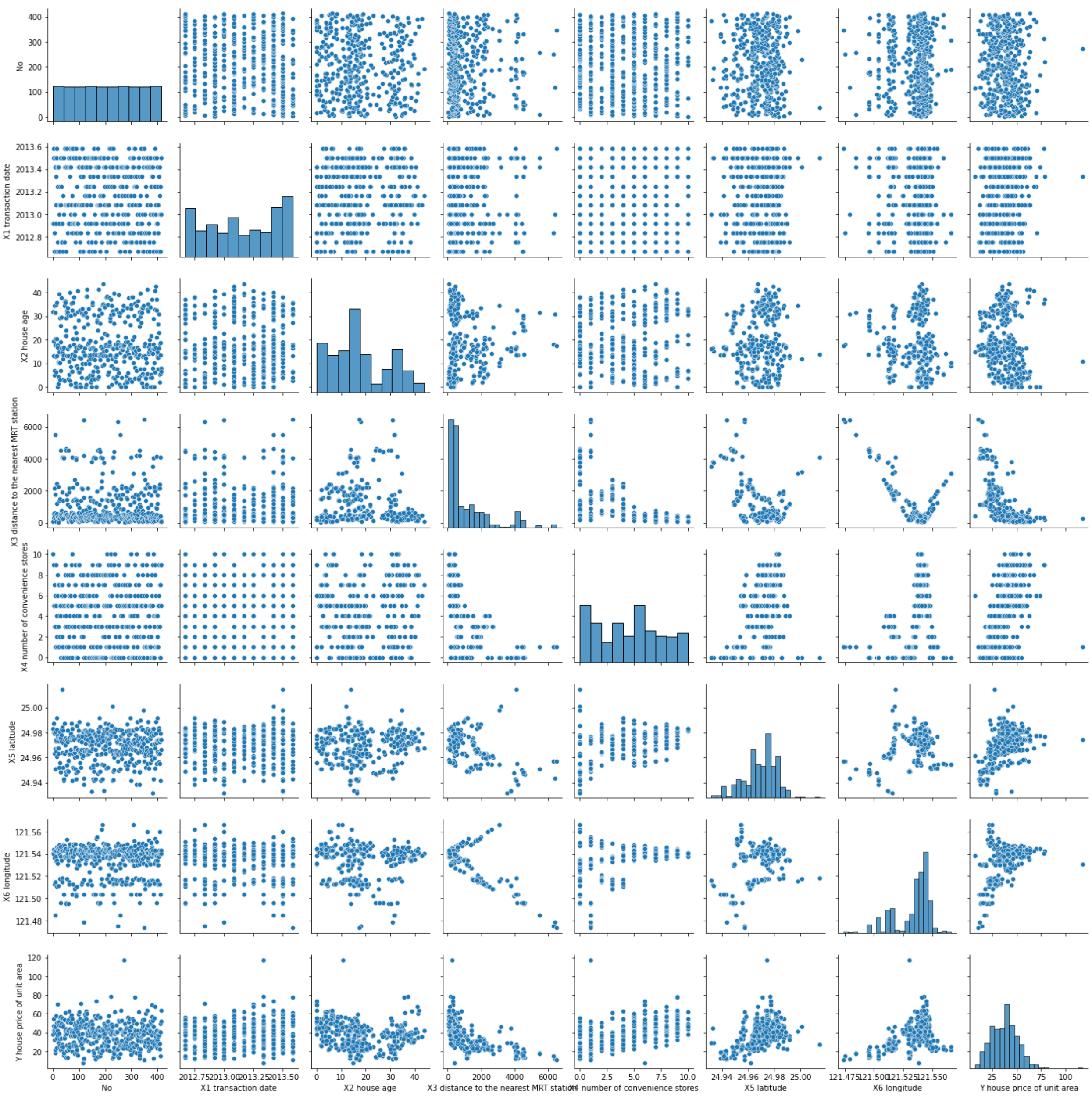
	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
No	**1.000000**	**-0.048658**	**-0.032808**	**-0.013573**	**-0.012699**	**-0.010110**	**-0.011059**	**-0.028587**
X1 transaction date	**-0.048658**	**1.000000**	**0.017549**	**0.060880**	**0.009635**	**0.035058**	**-0.041082**	**0.087491**
X2 house age	**-0.032808**	**0.017549**	**1.000000**	**0.025622**	**0.049593**	**0.054420**	**-0.048520**	**-0.210567**

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
X3 distance to the nearest MRT station	-0.013573	0.060880	0.025622	1.000000	-0.602519	-0.591067	-0.806317	-0.673613
X4 number of convenience stores	-0.012699	0.009635	0.049593	-0.602519	1.000000	0.444143	0.449099	0.571005
X5 latitude	-0.010110	0.035058	0.054420	-0.591067	0.444143	1.000000	0.412924	0.546307
X6 longitude	-0.011059	-0.041082	-0.048520	-0.806317	0.449099	0.412924	1.000000	0.523287
Y house price of unit area	-0.028587	0.087491	-0.210567	-0.673613	0.571005	0.546307	0.523287	1.000000

Correlation Matrix



Exploratory Data Analysis



Model

Regression analysis is a statistical method that helps us to analyze and understand the relationship between two or more variables of interest. It is a method of modeling a target value based on independent predictors. This method is used for forecasting and finding out cause and effect relationship between variables.

Linear Regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent (x) and the dependent (y) variable. Then based on the data points, we plot a line that models the points the best. In this paper, we use linear regression to predict real estate prices.

First, we will split the data into two parts, X and Y. Where X would be the data frame without the house price of unit area and Y is Y house price of unit area. The we use train test split to divide the data into training data and test data in 90:10 ratio. Then we will use linear regression model to train the data after that we predict house price of unit area using X test. At last, we compare Y test and Y pred.

Analyzed the data

Using Linear Regression model

	Y_Test	Y_Pred	Residuals
176	19.2	16.382266	2.817734
347	11.2	3.929628	7.270372
307	24.7	17.438121	7.261879
299	46.1	47.146088	-1.046088
391	31.3	27.340962	3.959038

Evaluating the model

	Metrics
MAE	4.490907
MSE	32.402898
RMSE	5.692354

Using polynomial Regression model

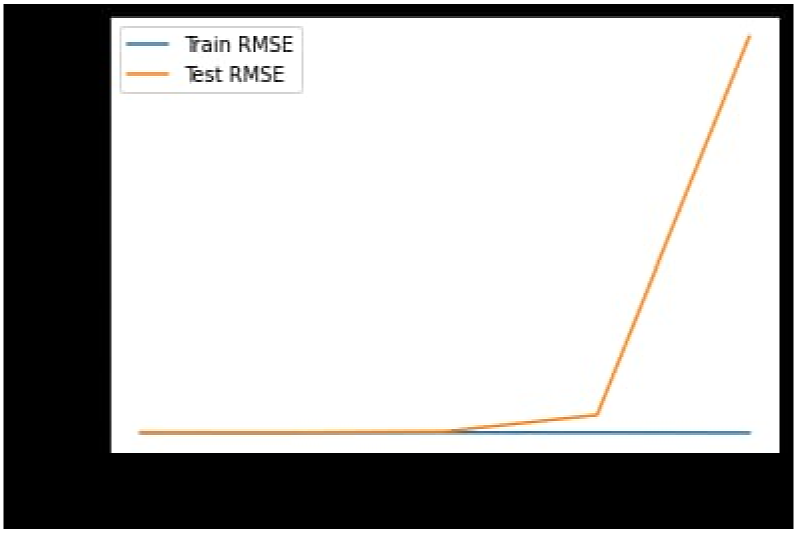
	Poly Metrics	Simple Metrics
MAE	4.490907	5.373025
MSE	32.402898	45.880307
RMSE	5.692354	6.773500

Adjusted model parameters to get best results

	Train RMSE List
0	9.537107
1	8.037433
2	7.100082
3	5.684359
4	1.812064
5	0.516267
6	0.725855
7	0.378090
8	0.435372

	Ttest RMSE List
0	6.773500e-00
1	5.692354e-00
2	2.378945e-01
3	2.872671e-02
4	6.306493e-03
5	3.366243e-04
6	2.020910e-05
7	1.588560e-06
8	1.501139e-07

Polynomial Degree Vs RMSE



Conclusion

Linear Regression performed well but when its results were compared with polynomial regression, it turns out polynomial regression performs better. And when in polynomial regression different degrees were used, they all got different results, while degree 9 got the best results with the least error. Both, linear regression and polynomial regression were highly efficient in predicting the house prices based on different factors that affect the prices of a house in a locality.

References

1. S. Borde, A. Rane, G. Shende, and S. Shetty, "Real estate investment advising using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 3, p. 1821, 2017.View at: Google Scholar
2. B. Trawinski, Z. Telec, J. Krasnoborski et al., "Comparison of expert algorithms with machine learning models for real estate appraisal," in *Proceedings of the 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Gdynia, Poland, July 2017.View at: Publisher Site | Google Scholar
3. V. Kontrimas and A. Verikas, "The mass appraisal of the real estate by computational intelligence," *Applied Soft Computing*, vol. 11, no. 1, pp. 443–448, 2011.View at: Publisher Site | Google Scholar
4. https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154

5. https://www.kaggle.com/quantbruce/real-estate-price-prediction
6. https://www.kaggle.com/aminizahra/polynomial-regression
7. A. S. Temür, M. Akgün, and G. Temür, “Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models,” J. Bus.