House Price Prediction Using Linear Regression

Bobby Susanto Ali1, Kevin Wijaya2, Jalaludin Ar Raniry Wiliam Tchin 3, Johannes Simatupang 4

*School of Computer Science, Bina Nusantara University  
Jakarta, Indonesia 11480*

[1 bobby.ali@binus.ac.id](mailto:1first.author@first-third.edu)

2 kevin.wijaya014@binus.ac.id

3 jalaludin.tchin@binus.ac.id

4 johannes.simatupang@binus.ac.id

***Abstract*— House prices are increasing every year which has necessitated the modeling of house price prediction. These models are constructed, to help the customers to purchase a house suitable for their needs. The proposed work makes use of the attributes or features of the houses such as the number of bedrooms available in the house and its building area and land area. By making use of the house price prediction system, the seller would be able to decide what features the customers could add to the house so that the house can be sold for a higher price.**

**This paper will help to predict house prices based on various parameters. The customers will be able to input any type of house they desire to buy, with a machine learning algorithm the house price predictor will display the house's estimated price.**

***Keywords***— ***Machine Learning, House Price, South Jakarta area, Linear Regression, Prediction***

1. Introduction

House is one of the assets that is often traded in Indonesia. When selling a house, people often find it difficult to determine a suitable price, which is not too expensive but does not cause the seller to lose. Likewise for prospective buyers, buyers need to know the approximate price of the house so they are not deceived by the price given by the seller.

Therefore, this project intends to apply Multiple Linear Regression to the obtained house price data. The results can help people estimate house prices more accurately based on the features installed on a house.

1. Methodology
2. *Dataset*

The dataset used in this project is data on house prices in the South Jakarta Administrative Region areas. The dataset is obtained from Kaggle. Data is taken and collected from several house marketplace such as rumah123.com. According to the Kaggle website, this data was updated in 2020. [4]

There are 2 datasets available. The first dataset contains 1000 rows, and the second dataset contains 1010 rows. This project uses a combination of the two datasets due to the similarity in the regional category of South Jakarta. Therefore the dataset we are using total contains 2010 rows. House location factors also affect house prices [1]. According to Kaiming Cheng and Qing Xia, 2007, urbanization could influence the supply and demand in the housing market and affect the price of housing. The population also affects housing prices. [2]. Based on these theories, we limit the scope of the prediction model in our project to only houses in the South Jakarta area.

The limitation of this project is that we have not included the time factor in house prices that may affect house prices.

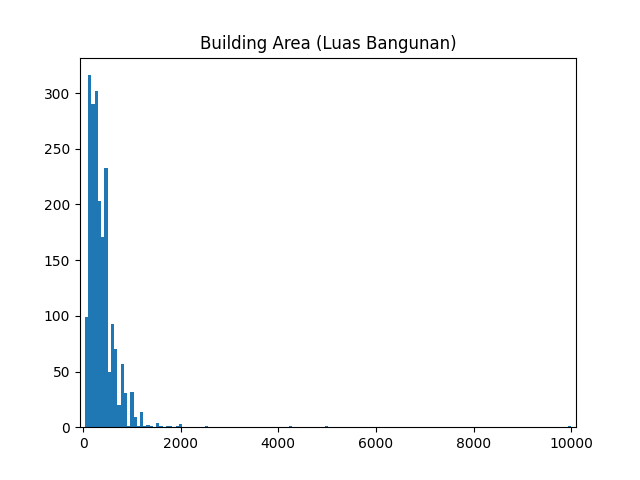
1. *Preprocessing*
2. First, we combine the house data using excel. We take the same columns from the two data, namely columns.

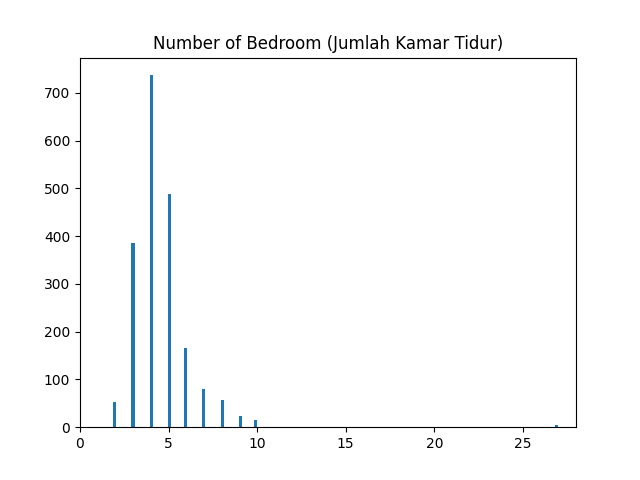
* LT (Land area)
* LB (Building area)
* KM (Bathroom)
* KT (Bedroom)
* GRS (Garage )

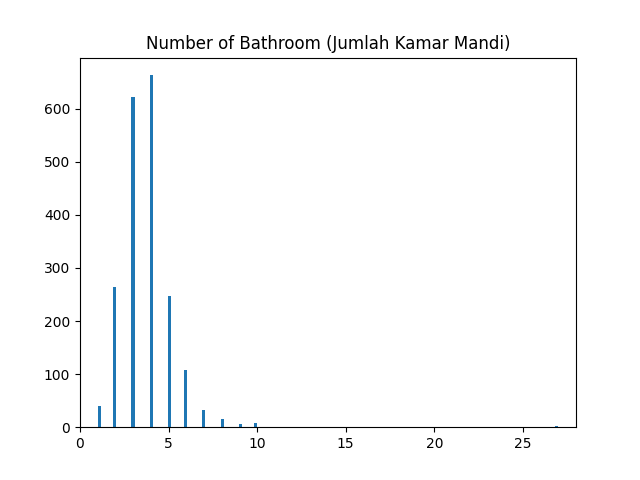
TABLE I

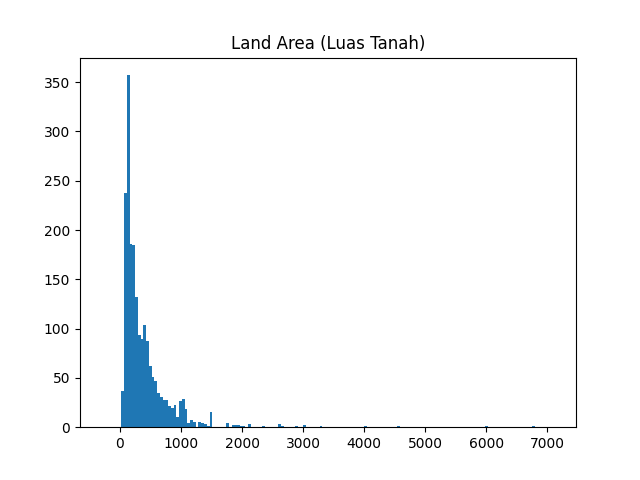
|  |  |  |
| --- | --- | --- |
| **Columns** | **Data 1** | **Data 2** |
| No | Data Available | Data Available |
| Price (rupiah) | Data Available | Data Available |
| House Description | Data Available | Data Not Available |
| Building Area (m2) | Data Available | Data Available |
| Land Area in (m2) | Data Available | Data Available |
| Number of Bedroom | Data Available | Data Available |
| Number of Bathroom | Data Available | Data Available |
| Capacity of Garage | Data Available | Data Not Available |
| Existence of Garage | Can be Inferred from Capacity of Garage. | Data Available |
| City | Data Not Available | Data Available |

1. There is a difference in the format of the garage in dataset 1 and dataset 2. The garage in dataset 1 contains detailed information about the capacity of the garage, while the garage column in dataset 2 only specifies whether there is a garage available in the house.
2. We combine the data, and we take the columns that intersect (exist in both dataset 1 and dataset 2).
3. Furthermore, the garage variable is still written in a different format. We standardized the format by converting all garage data into the string “ADA” or “TIDAK ADA” (available or unavailable).
4. Then the garage data is changed to code, “ADA” becomes 1, and “TIDAK ADA” becomes 0.

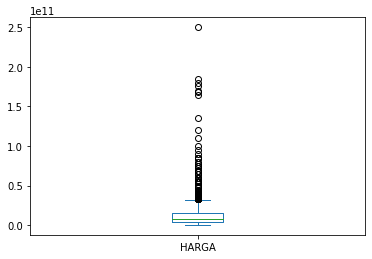
Histogram

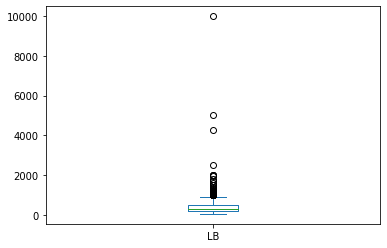


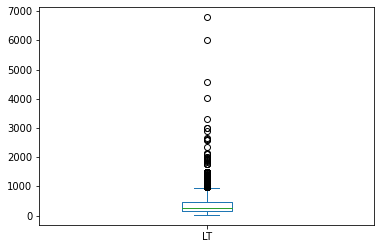


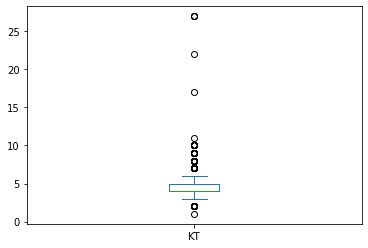


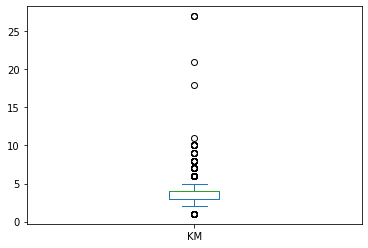
Box Whisker Plot











Next, standardization is carried out. After looking at the exploratory data analysis, we see that the variable data on building area and land area are normally distributed and the distribution is uneven, containing some outliers. So based on our observation of data visualization, the Z-Score normalization method will be more suitable to use for normalizing this dataset instead of the maximum absolute scaler or min-max scaler.

TABLE II

|  |  |  |
| --- | --- | --- |
| **Variable** | **Mean** | **Standard Deviation** |
| LT | 381.4361014420686 | 359.20056658411033 |
| LB | 383.3127797115863 | 421.96660419457726 |
| KM | 4.56340129288911 | 1.80335043836757 |
| KT | 3.773247140726007 | 1.685849903933497 |
| Harga | 12529821426.753855 | 16323248921.26816 |

1. *Feature Selection*

After we looked at the covariance and correlation of the data, we decided to select the features of the building area, land area, bedroom, and bathroom. We did not choose the garage variable, because after looking at the low correlation and covariance, it showed that the garage had no effect and was not closely related to house prices.

TABLE III

|  |  |
| --- | --- |
| **Covariance of** | **Harga** |
| LT | 5327089668205.516 |
| LB | 4061933468486.8057 |
| KM | 7502007360.009331 |
| KT | 5870632885.77855 |
| GRS | -285111922.2110003 |

Pearson Correlation



1. *Regression*

We use linear regression to predict home prices based on those features. In this project, we set test\_size for 0.6 to get the best R2 Score Result. The linear regression model generated from training data has coefficients below.

|  |  |
| --- | --- |
| LB | 0.37210847216041837 |
| LT | 0.6509273591933307 |
| KT | -0.041073854049557594 |
| KM | 0.047985319817969156 |

Intercept: 0.004877398351060949

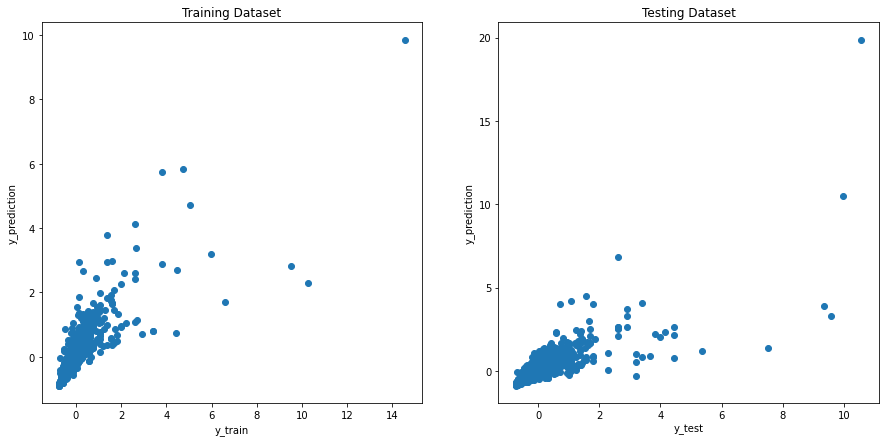
The coefficient of the linear regression model for LT and LB variables is larger than the coefficient of KM and KT because the correlation and covariance with the price variable are greater than the correlation and covariance of the price variable with KT and KM. LT (land area) and LB (building area) are closely related and greatly affect house prices.

1. *Evaluation*

|  |  |
| --- | --- |
| **MAE, MSE, RMSE, R2 Score for Training Data** | |
|
| MAE (Mean Absolute Error) | 0.2720203084072337 |
| MSE (Mean Squared Error) | 0.3922558215549916 |
| RMSE (Root of Mean Squared Error) | 0.6263032983746706 |
| R2 (R2 Score) | 0.6609139901712071 |

|  |  |
| --- | --- |
| **MAE, MSE, RMSE, R2 Score for Testing Data** | |
|
| MAE (Mean Absolute Error) | 0.2675510563148831 |
| MSE (Mean Squared Error) | 0.38440775732771065 |
| RMSE (Root of Mean Squared Error) | 0.6200062558778828 |
| R2 (R2 Score) | 0.5701117601538825 |

Plot Actual Data vs Predicted Result



1. Result

The results of this project we deploy in the form of a desktop application. We use the tkinter library in python and pyinstaller to convert the python file into an executable file. The user only needs to enter the area of the building, the area of the land, the number of bedrooms, and the number of bathrooms, and the results of the house price prediction will appear (in rupiah).

IV. Conclusion

In this project, we succeeded in making an application to predict the price of houses in the area of South Jakarta by entering several features such as building area, land area, number of bedrooms, and number of bathrooms. Here we use a combination of 2 different datasets that we get from Kaggle with the similarity of the regional category, namely South Jakarta which we then process using the Linear Regression Algorithm with a Test Size of 60% and obtain the value of training and testing R2 Score successively as follows 66% and 57%. Further research will be needed in the areas of the sample and optimize features such as the location of the house and the number of floors.

References

1. A. N. Hasanah and M. H. Yudhistira, “Landscape view, height preferences, and apartment prices: evidence from major urban areas in Indonesia,” International Journal of Housing Markets and Analysis, vol. 11, no. 4, pp. 701–715, Jul. 2018, doi: 10.1108/IJHMA-09-2017-0082.
2. Y. Wang and Y. Jiang, “An Empirical Analysis of Factors Affecting the Housing Price in Shanghai,” Asian Journal of Economic Modelling, vol. 4, no. 2, pp. 104–111, 2016, doi: 10.18488/journal.8/2016.4.2/8.2.104.111.
3. “Create Executable from Python Script using Pyinstaller.” https://datatofish.com/executable-pyinstaller/ (accessed Jun. 17, 2022).
4. W. Anggara, “Daftar Harga Rumah.” https://www.kaggle.com/datasets/wisnuanggara/daftar-harga-rumah?resource=download (accessed Apr. 29, 2022).
5. A. Primananda, “FAKTOR-FAKTOR YANG MEMPENGARUHI KONSUMEN DALAM MEMBELI RUMAH,” Semarang, 2010.