House Price Prediction Using Multiple Linear Regression

Vernando Bayu Putra Pratama[1], Reyhan Agus Priyatna[1], Fikri Arif Rahman[1]
[1]Department of Software Engineering, Universitas Pendidikan Indonesia, Indonesia

ARTICLE INFO

Article history's:

Keywords:

House Prices
Predictions
Multiple Linear Regression
House Area
Cost

ABSTRACT

A house is very important for most people, because it is a place where someone can feel comfortable and safe. However, the existence of various types and models of houses makes it confusing for someone to determine their residence, especially in matching with their financials because the selling price of a house is always changing and increasing, so a system is needed that can predict the selling price of a house in the future. Therefore, in this research, we will create a prediction system for determining the price of a house using multiple linear regression method. Using a dataset containing 21,640 house data with 20 variables. Of the many variables, the area of the house is the main factor in the selling price of the house itself. In this modeling, only 5 variables are used which are determined based on the correlation value of the variables with the price, namely the area of the building, the quality of the house, the number of bathrooms, the area of the house above the ground, and the area of the house from the nearest 15 neighbors. Research using multiple linear regression method produces a model score value of 55%. With that, the expected results are able to help someone in predicting the price of a house, preparing the budget, and determining the criteria for the type of house they want to buy.

Vernando Bayu Putra Pratama,
Department of Software Engineering, Universitas Pendidikan Indonesia, Indonesia
Email: vbpp@upi.edu

1. INTRODUCTION

Basically, humans live by striving to fulfill their basic needs in order to achieve a sense of well-being. One of the most basic needs that humans need is a house. A house is a building that serves as a place to live and an asset for a family that is livable [1]. Houses have a role in helping humans fulfill physiological needs. According to Abraham Maslow, physiological needs are absolute human needs that are extremely necessary for survival [2].

Nowadays, more and more people are thinking of buying a house at a young age, the goal is to invest because from year to year the price of a house will increase due to many things [3]. In its implementation, buying a house is not as easy as it seems, there are many aspects to consider, especially the price, the price is an important thing that is noticed by consumers when they want to buy a house. In order to speed up the process of buying a house, we can plan to save at the right time by estimating or predicting the selling price of a house in the future. However, the selling price of a house is always changing, uncertain and cannot be predicted manually. Therefore, a system is needed that can predict the selling price of a house so that the results are accurate and can be used in saving plans to determine the amount of money and when to start saving in order to be able to buy a house because. Because, at present, prediction systems are used as a tool to consider a decision, one of which is in the economic business sector, especially in the process of purchasing or transaction [4].

Based on these problems, we are interested in conducting research to create a system that can calculate the selling price of a house using the multiple linear regression method because this method is considered to produce more accurate values compared to other methods [4]. Previous research using the same method has been carried out by Andi et al. (2020) resulting in an accuracy rate of 80% in the prediction analysis of house prices, however, this research uses a small amount of data [5]. Then, the research conducted by Mu'tashim et al. (2021) resulted in an accuracy rate of 66% in the prediction analysis of house prices with 5 variables against the price variable [6]. The purpose of this study is to analyze the prediction of house prices using the

multiple linear regression method by increasing the amount of data to help the process of determining or predicting the quick and accurate sale price of the house. The benefit of this study is to learn and understand how the multiple linear regression method is applied in predicting house prices.

2. RESEARCH METHOD

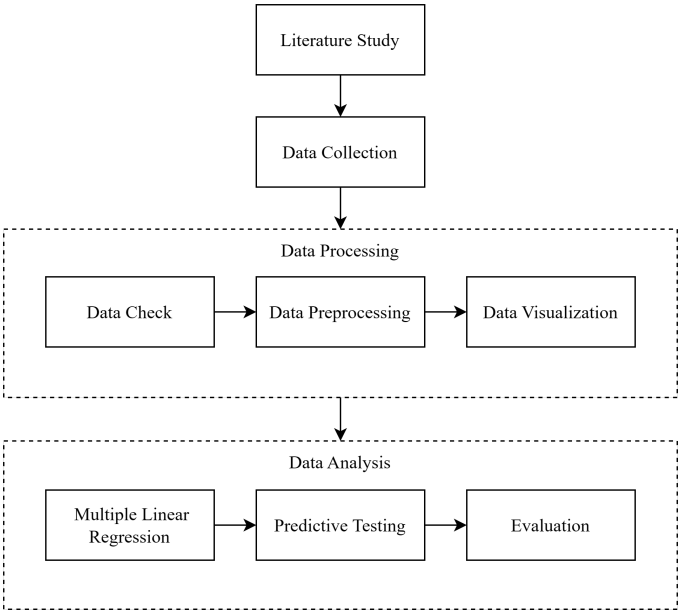


image 1. Research Procedure

2.1. Literature study

Literature study is the first step in research by doing literature related to theories related to all research activities. Theories related to this research include house price analysis and multiple linear regression. In this stage, research references are obtained from reading scientific journals, books, or the internet (with reliable sources).

2.2. Data Collecting

The next step is the data collection phase. In this phase, data is collected as supporting data to determine the price of a house. The data is obtained from the www.kaggle.com website, which contains a collection of datasets ready for processing, as many as 21,640 houses in King County, USA with 20 variables as specifications of a house obtained to be used in the next phase [7].

2.3. Data Processing

This stage is a stage that aims to represent collected data into data ready for analysis. This stage is divided into 3 stages, namely data check, data preprocessing, and data visualization. In the data check stage, the minimum, maximum, average value, and the number of data for each data attribute will be calculated to help the process of analysis. Then in the data preprocessing stage, the aim is to clean or improve irrelevant data so that the data will be processed effectively. Finally, the last stage of data processing is data visualization, data that has passed the data check and data preprocessing stage is then displayed in the form of a graph.

2.4. Implementation Of Multiple Linear Regression

The next step is to analyze the cleaned data using the multiple linear regression method. Regression is a method to study or analyze the relationship between one variable and one or more other variables. If the comparator or independent variable is one, then the analysis is called simple linear regression. Whereas if the comparator or independent variables are two or more, then the analysis is called multiple linear regression. The independent variables are variables that can affect or are known as independent variables and the non-free variables are variables that are affected or known as dependent variables.[8]

The goal of multiple linear regression is to make predictions on the value of the dependent variable (Y) if the values of the independent variables (X) are already known. It can also be used to determine the

direction/influence of the dependent variable with the independent variables. The formula for multiple linear regression is expressed in (1).

$$Y = a + b_1X_1 + b_2X_2 + \ldots + bnXn \quad (1)$$

Description:
Y = Dependent Variable
a = Konstanta
$X_1, X_2..Xn$ = Independent Variable
$b_1, b_2..bn$ = Koefisien regresi

The next step is to find the value of the constant a using Formula (2) and the value of the regression coefficient from $b_1, b_2..bn$ with formulas (3) and (4)

$$a = \frac{(\Sigma Y) - (b_1 \Sigma x_1) - (b_1 \Sigma x_2)}{n} \quad (2)$$

$$b_1 = \frac{[(\Sigma(x_2)^2 \Sigma x_1 y) - (\Sigma x_2 y \Sigma x_1 x_2)]}{[(\Sigma(x_1)^2 \Sigma(x_2)^2) - (\Sigma(x_1 x_2)^2)]} \quad (3)$$

$$b_2 = \frac{[(\Sigma(x_1)^2 \Sigma x_2 y) - (\Sigma x_1 y \Sigma x_1 x_2)]}{[(\Sigma(x_1)^2 \Sigma(x_2)^2) - (\Sigma(x_1 x_2)^2)]} \quad (4)$$

Furthermore, to determine the percentage of influence of independent variables(X_1, $X_2..Xn$) of the independent variable/dependent (Y) can be calculated using the formula coefficient of determination (R2). The coefficient of determination (R2) only produces a number between 0 and 1. If the result of R2 = 0, then it can be said that the variable used in modeling is wrong which makes the prediction result is not good and vice versa the further the result of R^2 from 0, then it can be said that the variable used is good so as to produce accurate predictions. The formula used to find R2 is (5).

$$R^2 = \frac{(b_1 \Sigma x_1 y) - (b_2 \Sigma x_2 y)}{\Sigma y^2} \quad (5)$$

Then, since the coefficient of determination is expressed in percent, the result is recalculated by Formula (5).

$$\text{Koefisien Determinasi} = R^2 \times 100\% \quad (6)$$

2.5. Data Analysis

The data analysis stage is the stage of analyzing refined data. The data analysis process is divided into 3 stages: multiple linear regression, testing and evaluation. In this stage, the data is divided into training data and test data that are used for different functions. The first stage is the implementation of multiple linear regression, which is done by calculating the correlation between variables in the independent variables to the dependent variables that will affect the final value. The second stage is the accuracy test by performing normality test, t-statistic test and hypothesis test. The hypothesis test is conducted to find the influence between the independent variables (X) and the dependent variables (Y). In this research, the influence of the house building area (X1) and house quality (X2) as independent variables on the price (Y) is sought. The hypothesis test is done with a partial test, which is a test to find out whether the independent variables (X) can still have an effect on the dependent variables (Y) separately. Finally, the evaluation is done by calculating the coefficient of determination, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE) as the performance evaluation of the model with the prediction calculation process used."

3. RESULTS AND DISCUSSION

3.1. Data Processing

3.1.1. Data Check

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21608	263000018	20140521T000000	360000.0	3	2.50	1530	1131	3.0	0	0	...	8	1530	0	2009	0	98103	47.6993	-122.346	1530	1509
21609	6600060120	20150223T000000	400000.0	4	2.50	2310	5813	2.0	0	0	...	8	2310	0	2014	0	98146	47.5107	-122.362	1830	7200
21610	1523300141	20140623T000000	402101.0	2	0.75	1020	1350	2.0	0	0	...	7	1020	0	2009	0	98144	47.5944	-122.299	1020	2007
21611	291310100	20150116T000000	400000.0	3	2.50	1600	2388	2.0	0	0	...	8	1600	0	2004	0	98027	47.5345	-122.069	1410	1287
21612	1523300157	20141015T000000	325000.0	2	0.75	1020	1076	2.0	0	0	...	7	1020	0	2008	0	98144	47.5941	-122.299	1020	1357

21613 rows × 21 columns

Image 2. Amount Of Data

Image 2. shows data that has 21,612 data rows and columns as parameters namely id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yt_built, yt_renovated, zipcode, lat, long, sqft_living15 and sqft_lot15. The Data is obtained from the website www.kaggle.com which is a website containing a set of datasets that are ready to be processed.

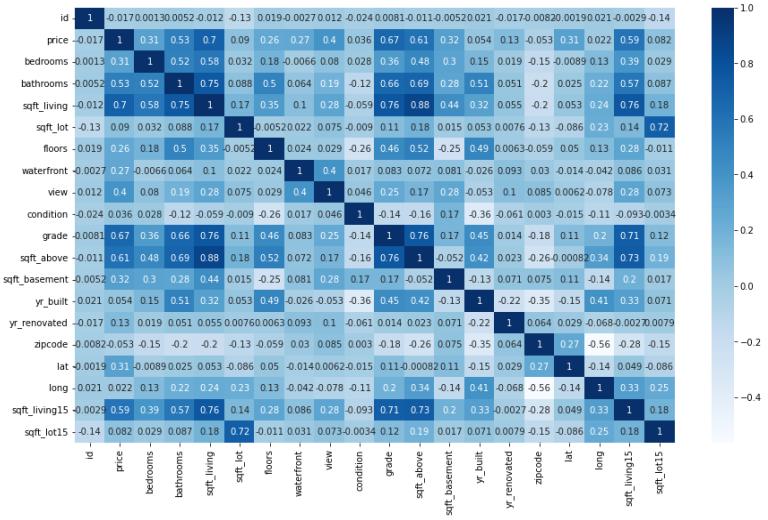
3.1.1. Pre-Process Data

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
count	2.161300e+04	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000
mean	4.580302e+09	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04	1.494309	0.007542	0.234303	3.409430
std	2.876566e+09	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04	0.539989	0.086517	0.766318	0.650743
min	1.000102e+06	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000
25%	2.123049e+09	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000
50%	3.904930e+09	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000	3.000000
75%	7.308900e+09	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000	4.000000
max	9.900000e+09	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000	5.000000

grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
7.656873	1788.390691	291.509045	1971.005136	84.402258	98077.939805	47.560053	-122.213896	1986.552492	12768.455652
1.175459	828.090978	442.575043	29.373411	401.679240	53.505026	0.138564	0.140828	685.391304	27304.179631
1.000000	290.000000	0.000000	1900.000000	0.000000	98001.000000	47.155900	-122.519000	399.000000	651.000000
7.000000	1190.000000	0.000000	1951.000000	0.000000	98033.000000	47.471000	-122.328000	1490.000000	5100.000000
7.000000	1560.000000	0.000000	1975.000000	0.000000	98065.000000	47.571800	-122.230000	1840.000000	7620.000000
8.000000	2210.000000	560.000000	1997.000000	0.000000	98118.000000	47.678000	-122.125000	2360.000000	10083.000000
13.000000	9410.000000	4820.000000	2015.000000	2015.000000	98199.000000	47.777600	-121.315000	6210.000000	871200.000000

Image 3. Descriptive Statistics Data

In the data pre-processing stage, data description is done by calculating descriptive statistics from a data frame, this function will calculate statistics such as mean, median, mode, min, and max from each numeric column in dataframe [5]. The results will be displayed in the form of a table that shows information about the distribution of data in the data frame. The purpose of this stage is to determine the information dissemination or distribution of data that can help in the next stage of data analysis .




```
price            1.000000
sqft_living      0.702035
grade            0.667434
sqft_above       0.605567
sqft_living15    0.585379
bathrooms        0.525138
view             0.397293
sqft_basement    0.323816
bedrooms         0.308350
lat              0.307003
waterfront       0.266369
floors           0.256794
yr_renovated     0.126434
sqft_lot         0.089661
sqft_lot15       0.082447
yr_built         0.054012
condition        0.036362
long             0.021626
id              -0.016762
zipcode         -0.053203
Name: price, dtype: float64
```

Image 4 Korelasi Antar Variabel

The image above illustrates the correlation relationship of all the house price variables in the dataset. This is intended to identify the interrelationships among the variables in the dataset.

3.1.3. Data Visualization

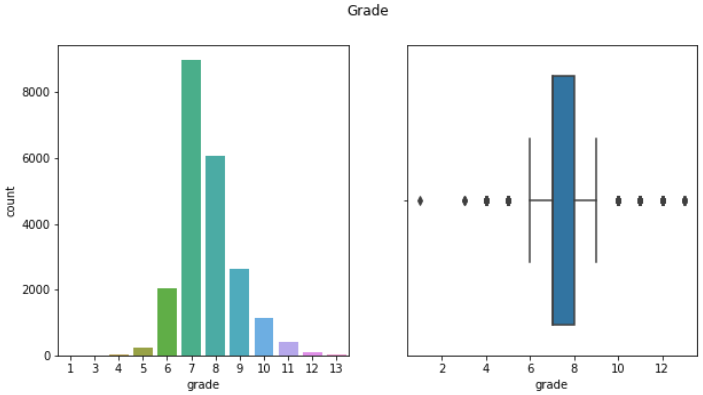


Image 5. Data Visualization Variabel Grade

In Image X, it explains the distribution of the grade column data, which ranges from 1 to 13. The most common grade is grade 7, with a data count of more than 8000.

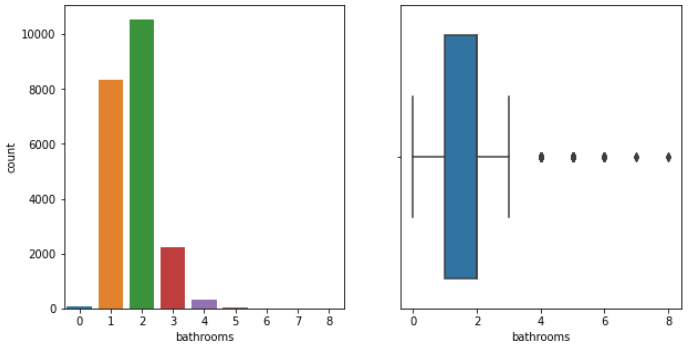


Image 6. Data Visualization Variabel Bathrooms

In Image X, it explains the distribution of the bathrooms column data, which ranges from 0 to 8. The data processed is dominated by the number of bathrooms being 2, followed by 1.

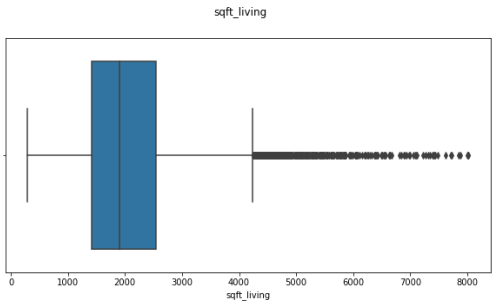


Image 7. Data Visualization Variabel Sqft_living

In Image X, it explains the distribution of the sqft_living column data, which ranges from 290 to 8500.

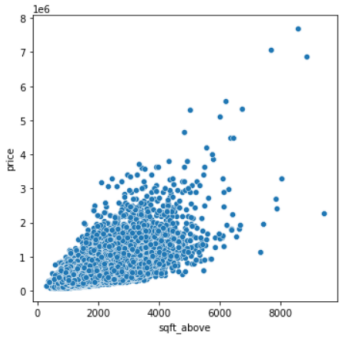


Image 8. Data Visualization Variabel Sqft_above

The boxplot above explains the distribution of the sqft_living15 column data, showing that the larger the value of sqft_living15, the more it is linearly correlated with the house price.

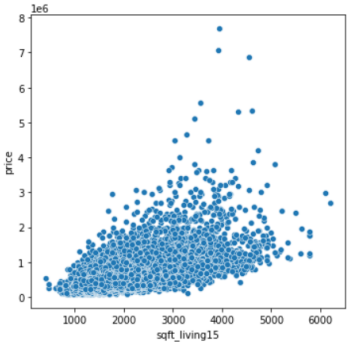


Image 9. Data Visualization Variable Sqft_living

The boxplot above describes the distribution of sqft_living15 column data which illustrates that the greater the value of sqft_living15 will linier with house prices.

3.2. Data Analysis

3.2.1. Implementation of MLR

a. Test the correlation between variables and independent variables

	price	sqft_living	grade	sqft_above	sqft_living15	bathrooms
price	1.00	0.70	0.67	0.60	0.59	0.50
sqft_living	0.70	1.00	0.76	0.88	0.76	0.69
grade	0.67	0.76	1.00	0.76	0.71	0.60
sqft_above	0.60	0.88	0.76	1.00	0.73	0.64
sqft_living15	0.59	0.76	0.71	0.73	1.00	0.51
bathrooms	0.50	0.69	0.60	0.64	0.51	1.00

Image 10. Correlation Data of each variable

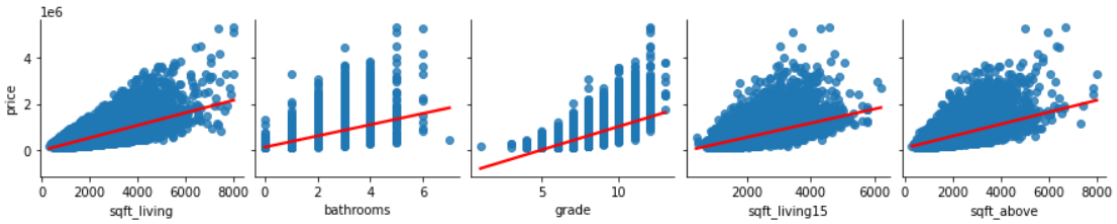


Image 11. Results Of Linear Regression Analysis

In Image 10. describes the correlation of variables used in modeling and can be seen from the correlation results that all variables have more than 60% attachment. Then in Image11 pairplot above describes the distribution of the price column of 5 variables that have been determined to do the modeling. This is done to determine the location of data on price variables.

3.2.1 Predictive Testing

A. Normality Test

	coef	std err	t	P>|t|
const	-6.554e+05	1.31e+04	-49.938	0.000
sqft_living	209.0195	4.347	48.084	0.000
grade	1.085e+05	2372.168	45.720	0.000
sqft_above	-82.1924	4.333	-18.971	0.000
sqft_living15	37.0154	3.930	9.420	0.000
bathrooms	1606.5724	3170.393	0.507	0.612

Image 12. Normality Test Results

Normalcy test is a statistical test used to determine whether a data distribution has a normal distribution or not. A normal distribution is a data distribution that has the same symmetry around the mean, with the mean as the center of the distribution. The normal distribution is also known as the Gaussian distribution or bell curve. The image above shows that there are variables whose distribution is not normal (indicated by P>0.05), namely the bathrooms variable.

B. Test statistic-t

	coef	std err	t	P>|t|
const	-6.554e+05	1.31e+04	-49.938	0.000
sqft_living	209.0195	4.347	48.084	0.000
grade	1.085e+05	2372.168	45.720	0.000
sqft_above	-82.1924	4.333	-18.971	0.000
sqft_living15	37.0154	3.930	9.420	0.000
bathrooms	1606.5724	3170.393	0.507	0.612

Image 13. Statistic t Test Result

The t-test is a test that compares the sample mean of each group by calculating the difference in sample means. A variable is said to be poor or incompatible when the t value of a variable is less than 0.05. Therefore, in this case, the sqft_above variable does not pass the t-test.

C. Testing Hipotesis

In this research, the hypothesis test is partially based on the t value of the variables used to determine whether variables X_1 and X_2, which are the size of the house and the quality of the house, respectively, have a

partial effect on the determination of variable Y, which is the price. Based on the t-test that has been conducted, the initial hypothesis and its alternative hypothesis are formulated. The hypotheses are as follows:
H_0: There is no significant partial influence between the size of the house and the quality of the house on the price.
H_1: There is a significant partial influence between the size of the house and the quality of the house on the price.
Based on Figure 13. Results of the t-test on independent variables, in the t column it can be seen that the significant t value for sqft_living (building size) is 48.084 and grade (quality) is 45.72. Then, the degree of freedom (df) is 21607 with an alpha (significance level) of 5% or 0.05, so the hypothesis results are as follows:

- X_1, which is 48.084 > 0.05, so H_0 is rejected. Thus it is proven that the size of the building has a significant partial effect on the price (Y).
- X_2, which is 45.72 > 0.05, so H_0 is rejected. Thus it is proven that the quality of the house has a significant partial effect on the price (Y).

3.2.3. Evaluation

a. RMSE evaluation (root mean square error)

Table 1. RMSE (Root Mean Squared Error)

RMSE
255306.84273207048

Root Mean Squared Error (RMSE) is a measure of error used in prediction models to measure how close the prediction result is to the actual value. RMSE is calculated by calculating the difference between the prediction value and the actual value, then multiplying the difference by the square (squared error), and calculating the average of the squared errors. The square root of the average squared error is then calculated as the RMSE. In the modeling that has been done, the RMSE result is 255306.84.

b. MAE (Mean Absolute Error) evaluation

Table 2. MAE (*Mean Absolute Error*)

Model Score (MAE)
160033.32078485965

MAE (Mean Absolute Error) is a measure of error used in prediction models to measure how close the predicted result is to the actual value. MAE is calculated by calculating the difference between the predicted value and the actual value, then multiplying the difference by the absolute (absolute error), and calculating the average of the absolute error. In modeling that has been done RMSEnya result of 160033.32

c. MSE (Mean Squared Error) evaluation

Table 3. MSE (*Mean Squared Error*)

Model Score (MSE)
57843060914.67174

MSE (Mean Squared Error) is a measure of error used in prediction models to measure how close the predicted result is to the actual value. MSE is calculated by calculating the difference between the predicted value and the actual value, then multiplying the difference by the square (the square of the error), and calculating the average of the squares of the error. In the modeling that has been done produces a value of 57843060914

d. Test Coefficient Of Determination (R2)

Tabel 4. Test Results Coefficient Of Determination

R^2	Model Score
0.542	*0.5568829909967401*

The coefficient of determination is a measure used to determine how well a model explains the variation in the data to be predicted. The coefficient of determination is known as the R-square or R^2. R-square is calculated by dividing the variation explained by the model by the total variation of the data to be predicted. In this modeling, the R-squared is at a value of 0.542 or 54.2%.

4. CONCLUSION

It can be concluded from this research that the size of the house and the quality of the house significantly affect the price of a house. Similarly, the number of bathrooms, the size of the house above the ground, and the size of the house from the nearest 15 neighbors also significantly affect the price of a house in King County. Among the many variables, the size of the house is the main factor of the selling price of the house itself. In this modeling, the writer only uses 5 variables determined based on the correlation value of the variables with the price. Because only a quarter of the total variables are used, only 55% of the model score is obtained although it can be said to meet the standard because it is more than 50%. It is hoped that in future research, additional variables with a high correlation value with the price will be added to improve the model score.

REFERENCES

[1] M. Sari, et al, Kesehatan Lingkungan Perumahan. Yayasan Kita Menulis, Desember 2020. [*Online*]. Tersedia: https://books.google.co.id/books?id=TDgNEAAAQBAJ&newbks=1&newbks_redir=0&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false [Diakses pada 13 Desember 2022]

[2] J. H. Sada, "Kebutuhan Dasar Manusia dalam Perspektif Pendidikan Islam," Jurnal Pendidikan Islam., vol. 8, pp. 213-226, 2017.

[3] F. Azkia, "Kaum Milenial Makin 'Pede' Beli Rumah | Riset Konsumen Properti | Rumah.com," Rumah, Feb. 27, 2018. https://www.rumah.com/berita-properti/2018/2/169635/kaum-milenial-makin-pede-beli-rumah (accessed Des. 12, 2022).

[4] G. N. Ayuni, D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ", J. Telemat., Vol. 14, No. 2, hal. 79–86, 2019.

[5] A. Saiful, *et al*, "Prediksi Harga Rumah Menggunakan Web Scraping Dan Machine Learning Dengan Algoritma Linear Regression", *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 8, pp. 41-50, 2021.

[6] M. L. Mu'tashim, et al, "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression," Jurnal Informatik, 17th ed, pp. 238-245, 2021.

[7] Kaggle. "House Price", [*Online*]. www.kaggle.com. [Diakses pada 10 Desember 2022].

[8] I. M. Yuliara, Modul Regresi Linear Berganda, 2016.

[9] K. Puteri, A. Silvanie. "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda", Jurnal Nasional informatika, vol. 1, no. 2, pp. 82-94, 2020.

[10] C. M. Z. Huzaen, "Pengaruh Konsep Perumahan, Lokasi, Dan Penyesuaian Harga Terhadap Keputusan Pembelian Pada Properti Perumahan Pesona Prima Griya Makassar (Studi Kasus Pt Prima Karya Bentala Permai),"pp. 1-15, 2019.