



Predictive Modeling for Property Prices: Linear Regression Approach

2023

Anis Dela Desela
Jakarta, Indonesia

Overview

The real estate company seeks to optimize the sale prices of properties by leveraging data-driven insights. This project aims to identify the key variables affecting house prices and develop a predictive linear model to quantitatively relate these variables to property prices. This project specifically considers the area of the property and preferred area to choose. However, the real estate company is open to the possibility that other variables may also affect the price of a home. The accuracy of the model will be assessed to evaluate its predictive capabilities.

Goals

- Identify the variables influencing house prices.
- Create a linear regression model to predict property prices based on these variables.
- Evaluate the model's accuracy and predictive performance.

Hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$$

(There is no significant influence from at least one predictor variable)

$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \text{ or } \dots \text{ or } \beta_p \neq 0$$

(At least one predictor variable ($\beta_1, \beta_2, \dots, \beta_p$) has a significant influence on the response variable)

Methodology

I. Dataset

The dataset utilized in this project comprises property price data for Delhi, including various parameters. This data was sourced from Kaggle: [Housing Price Prediction](#)

The data have the following columns:

Variable	Definition
area	The total area of the housing property
bedrooms	The total of bedrooms in the house Numerical data that consists of: 1,2,3,4,5,6
bathrooms	The number of bathrooms or toilets in the house Numerical data that consists of: 1,2,3,4
stories	The number of floors or levels in the building Numerical data that consists of: 1,2,3,4
mainroad	This variable indicating whether the house has direct access to the main road or not Numerical data that consists of: 0,1
guestroom	This variable indicating whether the house has a guest

	room or not Numerical data that consists of: 0,1
basement	This variable indicate the presence of a basement (underground space) in the house Numerical data that consists of: 0,1
hotwaterheating	This variable indicate whether the house has a hot water heating system or not Numerical data that consists of: 0,1
airconditioning	This variable indicating whether the house is equipped with an air conditioning system or not Numerical data that consists of: 0,1
parking	This variable indicate the number of parking spaces available on the housing property Numerical data that consists of: 0,1,2,3
prefarea	This variable indicating whether the house is in a preferred area or not Numerical data that consists of: 0,1
furnishingstatus	This variable contain information about the furnishing status of the house Numerical data that consists of: 1,2,3

II. Statistical Test

A. Ordinary Least Square

Linear Regression is selected as the statistical analysis method for this project due to its suitability for identifying and quantifying the relationships between variables, such as how property prices are influenced by factors like property area and preferred location.

This method allows us to build a predictive model that can provide insights into how these variables affect property prices, making it an appropriate choice for achieving the project's goals

III. Tools

In conducting this analysis, I utilized Python. The python script for this purpose is accessible at : [House Prediction Code - Anis' Github](#)

Analysis

In this analysis, I utilized a linear regression model to build our project. However, in order to create the model, there are several essential steps that need to be executed. The following outlines the sequence of actions taken in the project's development.

I. Data Overview

A. Data Type

In this section, my goal is to obtain an overview of the data used, which in this case is housing data.

The housing data consists of 545 rows and 13 columns. The data types present in the housing data include 6 data columns with 'int64' data type, while the remaining 7 data columns are of 'object' data type. In order to analyze it using a model, I will transform the data in the object format into nominal format through encoding, as shown in picture 1.

<pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 545 entries, 0 to 544 Data columns (total 13 columns): # Column Non-Null Count Dtype --- - 0 price 545 non-null int64 1 area 545 non-null int64 2 bedrooms 545 non-null int64 3 bathrooms 545 non-null int64 4 stories 545 non-null int64 5 mainroad 545 non-null object 6 guestroom 545 non-null object 7 basement 545 non-null object 8 hotwaterheating 545 non-null object 9 airconditioning 545 non-null object 10 parking 545 non-null int64 11 prefarea 545 non-null object 12 furnishingstatus 545 non-null object dtypes: int64(6), object(7) memory usage: 55.5+ KB</pre>	<pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 545 entries, 0 to 544 Data columns (total 13 columns): # Column Non-Null Count Dtype --- - 0 price 545 non-null int64 1 area 545 non-null int64 2 bedrooms 545 non-null int64 3 bathrooms 545 non-null int64 4 stories 545 non-null int64 5 mainroad 545 non-null int64 6 guestroom 545 non-null int64 7 basement 545 non-null int64 8 hotwaterheating 545 non-null int64 9 airconditioning 545 non-null int64 10 parking 545 non-null int64 11 prefarea 545 non-null int64 12 furnishingstatus 545 non-null int64 dtypes: int64(13) memory usage: 55.5 KB</pre>
--	--

Pict 1. Encoding data type from object to int64

B. Missing Value

For the purpose of checking missing values, the housing data does not contain any missing values. Therefore, there is no need for any value adjustments.

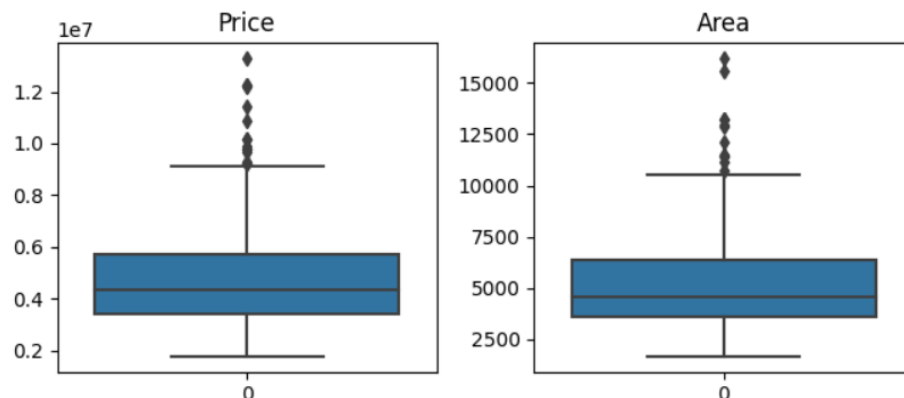
```
#Missing Value
missing_value = housing.isnull().sum()
print(missing_value)
```

price	0
area	0
bedrooms	0
bathrooms	0
stories	0
mainroad	0
guestroom	0
basement	0
hotwaterheating	0
airconditioning	0
parking	0
prefarea	0
furnishingstatus	0
dtype:	int64

Pict 2. Detecting missing values

C. Outliers

There is a notable presence of outliers in the 'Price' and 'Area' measurements. This is understandable, as various properties of houses can naturally lead to outliers in other metrics as well. As a next step, I intend to remove these outliers specifically from the 'Price' and 'Area' data, considering the potential impact on the overall analysis or model performance.

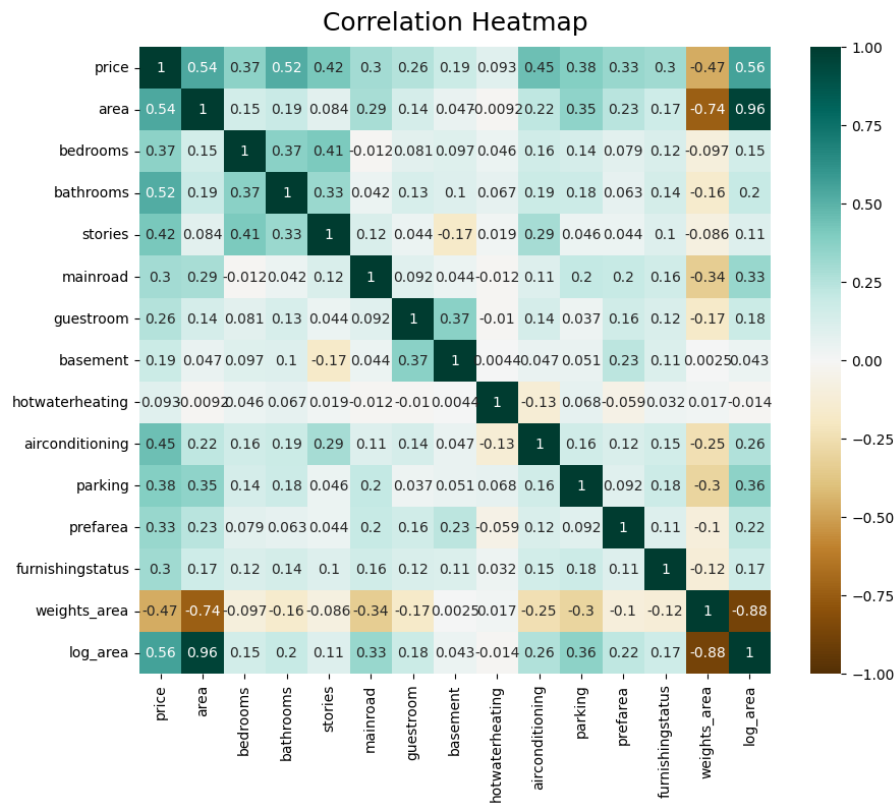


Pict 3. Detecting outliers

II. Visualization

A. Correlation

When examining the correlation, I observe that the **variables 'area' and 'bathrooms' show a relatively strong correlation with 'price'**, while **the remaining variables exhibit a correlation weaker than |0.5| with 'price.'** This suggests that 'price' and 'bathroom' could have a significant influence on the price, but I am still interested in determining which variables contribute the most to building the best model.



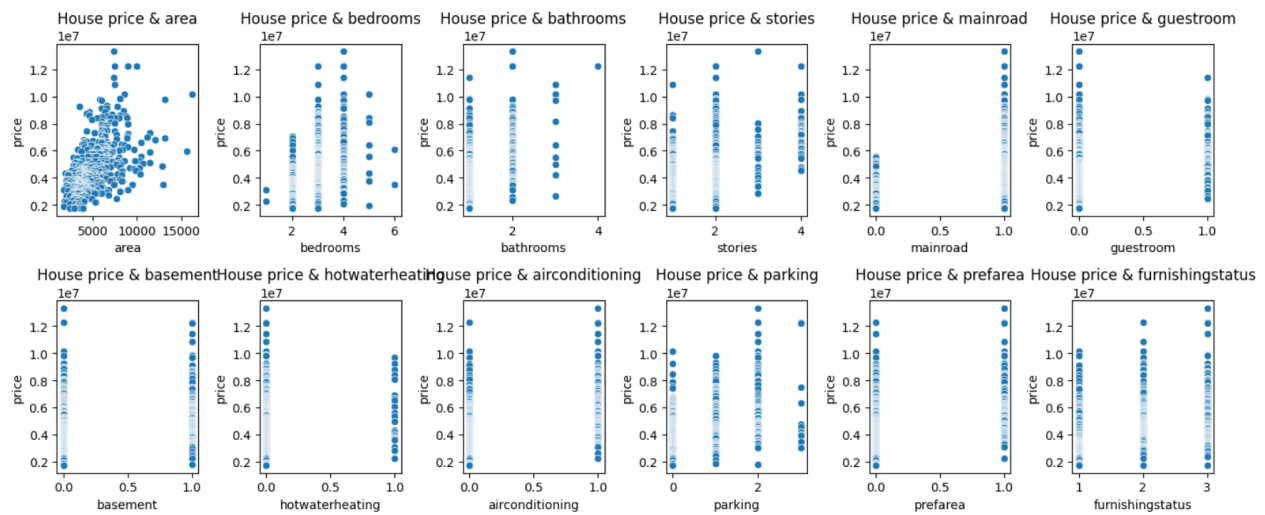
Pict 4. Checking correlations

B. Scatterplot

The scatterplot displayed below illustrates the relationship of items in each variable and price. Below provided the summarize explanations from the scatterplot:

The scatter plots for variables such as **area, bedrooms, bathrooms, stories, mainroad, airconditioning, parking, prefarea, and furnishing status** exhibit a positive slope towards higher values, indicating that **higher values in these variables are associated with higher house prices**.

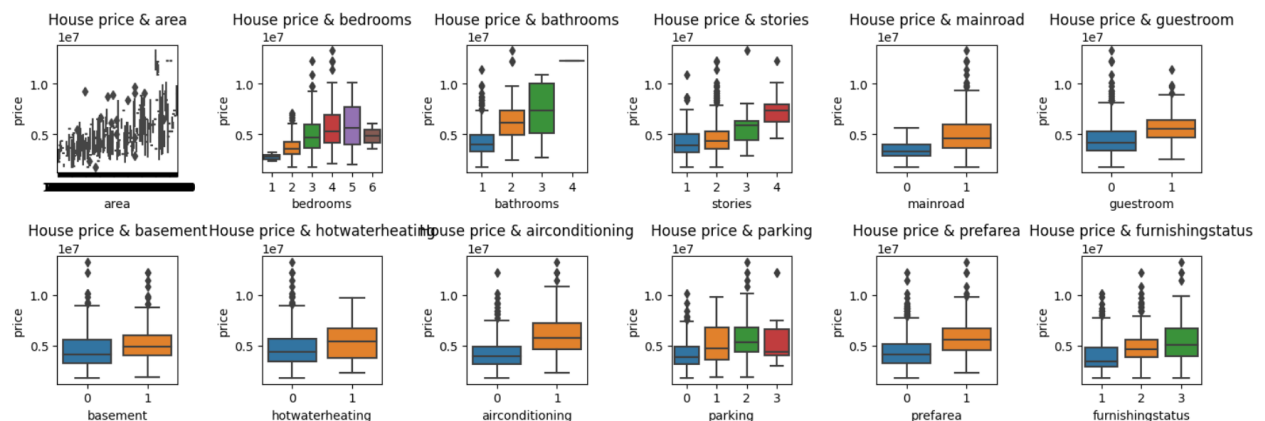
However, the variables **basement, hotwaterheating, and guestroom** do not follow this trend. It's important to note that this variation may arise due to the inherent data variability in house prices. Therefore, in the following section, I will explore the central values of each item within these variables using boxplots in relation to house prices.



C. Boxplot

Using boxplots, we can detect outliers for each variable in relation to the dependent variable (price) in the dataset. We can also visualize the median and data distribution. It's worth noting that I have already eliminated outliers associated with price and area variables. Therefore, in this section, I will **focus on visualizing how the central values (medians) are distributed for each item within every variable.**

In a broader context, the boxplot visualizations presented below suggest that the central tendencies of each variable show a positive trend for each item with higher values. **Indicating that higher values of each variable tend to be associated with higher house prices.**



III. Model Creation

In this analysis, I conducted linear regression by exploring various combinations of variables. However, what I present below is the one that produced the best model,

where I used **Ordinary Least Squares (OLS) with all variables**, applied a Box-Cox transformation for the dependent variable, and then eliminated the independent variables that did not have a significant impact.

A. OLS Regression Model

Hypothesis:

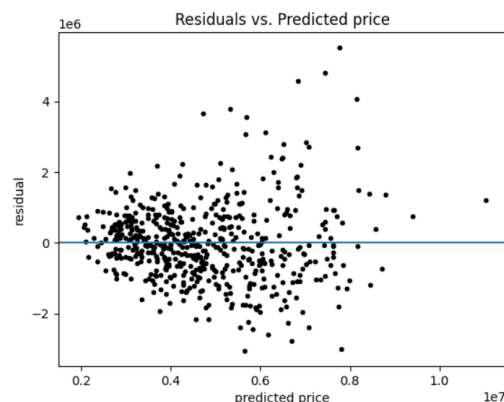
$$\alpha = 0.05$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$$

$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \text{ or } \dots \text{ or } \beta_p \neq 0$$

Analysis

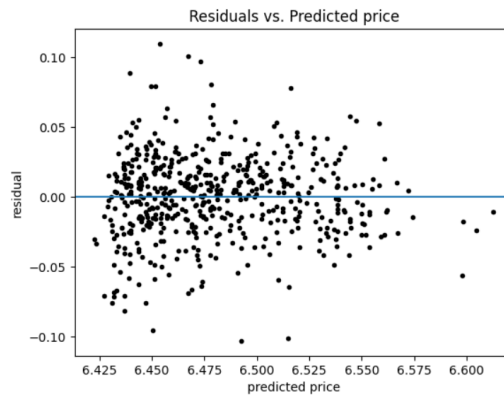
This analysis was conducted using OLS regression. First, I tried including all variables in the model. Then, I examined the residual plot with the dependent variable. It was found that **using this model, the residuals exhibited a plot that tends to form a pattern, which has an increasing trend. This suggests that the error variance increases with the dependent variable.** Additionally, the data distribution also showed a right-skewed tendency. Therefore, I performed a **transformation using the Box-Cox method on the dependent variable ('price') to transform the non-normal data into a normal distribution and maintain data homoscedasticity.** Box-Cox is used because with it, we can adapt to the most suitable transformation for the data (using the optimal value of λ).



After performing the transformation using the Box-Cox method on the "price" variable, I attempted to **re-create the model using the transformed "price" data**. Following that, I **conducted variable elimination with the assistance of feature selection**. The number of variables present in the model was too many, making it redundant to include all of them as reference variables in building the model. **The purpose of using feature selection is**

to avoid overfitting, remove less significant features, and make the model easier to interpret. By employing feature selection, it was determined that the variables significantly influencing the model are "area," "bathrooms," "stories," "guestroom", "airconditioning", and "prefarea".

The plot below depicts the relationship between residuals and predicted prices after the Box-Cox transformation, from the model with independent variables that have been selected using feature selection. The plot below shows a different residual overview compared to the plot before the Box Cox transformation & feature selection, which seem to be more random and do not display a trend.



Looking at the OLS Regression Results table below, the accuracy obtained using the adjusted R-squared is 60.5%, with the model as follow:

$$y(\lambda) = 1381.7 + 12.6x_1 + 9.3x_2 + 7.3x_3 + 4.5x_4 + 6.5x_5 + 7.6x_6 + \varepsilon$$

$y(\lambda)$ = price (after the Box-Cox transformation)

x_1 = area, x_2 = bathrooms, x_3 = stories, x_4 = guestroom, x_5 = prefarea, x_6 = airconditioning

OLS Regression Results						
=====						
Dep. Variable:	price	R-squared:	0.610			
Model:	OLS	Adj. R-squared:	0.605			
Method:	Least Squares	F-statistic:	140.1			
Date:	Wed, 27 Sep 2023	Prob (F-statistic):	1.67e-106			
Time:	10:13:03	Log-Likelihood:	1146.5			
No. Observations:	545	AIC:	-2279.			
Df Residuals:	538	BIC:	-2249.			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	6.3710	0.005	1381.693	0.000	6.362	6.380
area	7.88e-06	6.26e-07	12.591	0.000	6.65e-06	9.11e-06
bathrooms	0.0254	0.003	9.277	0.000	0.020	0.031
stories	0.0117	0.002	7.307	0.000	0.009	0.015
guestroom	0.0153	0.003	4.463	0.000	0.009	0.022
prefarea	0.0204	0.003	6.527	0.000	0.014	0.027
airconditioning	0.0224	0.003	7.571	0.000	0.017	0.028

Because we are using Box-Cox with lambda (λ) \approx -0.13 (the lambda value is obtained through iteration via the stats.boxcox function). Therefore, the interpretation should be adjusted accordingly to the Box-Cox interpretation. It would be easier if we illustrate with a practical example

Box-Cox transformation formula:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda}$$

Therefore if we want to revert to the original scale for easier interpretation, the formula is as follows:

$$y = (y(\lambda) \cdot \lambda + 1)^{\frac{1}{\lambda}}$$

Sample Case:

Someone wants to buy a house with the following criteria:

- area = 5000, 1 bathroom, 2 stories, 0 guestroom, house in preferred area, and with air conditioning

How much is the price that he might spend for a house that meets these criteria?

- Using the above model, the result is $y(\lambda) = 6.502$
- As for transforming $y(\lambda)$ into y using the formula above, we obtain y or the price that the person might pay, which amounts to 5,239,881.

B. Statistical Significance Test

The interpretation of the statistical tests within the model are as follows:

Omnibus Test

Omnibus:	14.074
Prob(Omnibus):	0.001

$\alpha = 0.05$

$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$

$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \text{ or } \dots \text{ or } \beta_p \neq 0$

The null hypothesis (H_0) is rejected if the p-value is less than $\alpha = 0.05$. Since the p-value of the Omnibus test is 0.001, which is less than 0.05, we can reject H_0 . This implies that at least one regression coefficient is not equal to zero, indicating that the overall regression model has statistical significance.

Skewness

Skew:	-0.062
-------	--------

The skewness value is -0.062, which is close to 0. This indicates that the data tends to be symmetric, but the distribution tail elongates to the left.

Durbin-Watson

Durbin-Watson:	1.036
----------------	-------

Durbin-Watson is utilized to test autocorrelation. It is used to check whether there are any unexplained residual patterns in the model. In this analysis, the Durbin-Watson statistic is 1.036. This can be interpreted as indicating the absence of autocorrelation in the constructed model.

Conclusion

- The model created is:


$$y(\lambda) = 1381.7 + 12.6x_1 + 9.3x_2 + 7.3x_3 + 4.5x_4 + 6.5x_5 + 7.6x_6 + \varepsilon$$

$y(\lambda)$ = price (after the Box-Cox transformation)

x_1 = area, x_2 = bathrooms, x_3 = stories, x_4 = guestroom, x_5 = prefarea,

x_6 = aircond

- Given the model coefficients, it appears that "area" has the most significant impact on the predicted price.
- The number of bathrooms ("bathrooms") also contributes significantly to the predicted price. Buyers looking for properties with multiple bathrooms may expect a higher price.
- The number of stories ("stories") is another influential factor. Buyers interested in multi-story properties may anticipate a higher predicted price compared to single-story options.

- 
- The presence of a guestroom ("guestroom") is a positive contributor to the predicted price. Buyers desiring this feature should be prepared for a potentially higher cost.
 - Properties located in preferred areas ("prefarea") tend to have a positive impact on the predicted price. Buyers seeking homes in specific preferred locations may see a premium associated with such choices.
 - The presence of air conditioning ("airconditioning") is another factor positively influencing the predicted price. Buyers who prioritize this feature should be aware that it contributes to the overall cost.