Predictive Modelling and Statistical Analysis of Housing Prices in Lagos State, Nigeria

Odukoya E.A., Oyelakin O.P., Lawal O. J

Department of Statistics, Faculty of Science, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria

DOI: https://doi.org/10.51244/IJRSI.2025.120800153

Received: 12 July 2025; Accepted: 18 July 2025; Published: 16 September 2025

ABSTACT

This study delves into housing market analysis and price prediction, leveraging statistical modeling and machine learning techniques to uncover patterns and forecast property prices. The housing market is influenced by diverse factors such as location, property features, economic indicators, and market trends, necessitating a comprehensive analytical approach. Using a dataset comprising historical housing prices and relevant attributes, the study employs exploratory data analysis to identify key determinants of property values. The findings highlight significant predictors of housing prices and demonstrate the potential of predictive analytics in guiding buyers, sellers, and policymakers. This research offers valuable insights into market dynamics and contributes to data-driven decision-making in real estate

Keywords: Machine Learning Techniques, Housing Market, Price Prediction

INTRODUCTION

The housing market which is important to the economy affects individual wealth and overall economic health. It's almost vital for a range of people like potential buyers, investors, policymakers, and real estate experts to get a good grasp on housing market patterns and predict property prices accurately. So basically, looking at the housing market and trying to figure out future prices means you need to get what's going on in residential properties.

Different analytical methods are employed to study and predict market behaviour. These methods range from basic statistics, which provide simple metrics like average prices, to more sophisticated techniques such as time series analysis and comparative market analysis. Advanced approaches also include machine learning models that use large datasets to forecast future prices based on various factors. However, predicting housing prices is challenging due to the volatility of the market and the complexity of factors involved. Changes in the economy, government policies, and demographic shifts can all influence housing prices in ways that are not always predictable. Overall, housing market analysis and price prediction are vital for making informed decisions in real estate investment, policymaking, and valuation. They help stakeholders understand market trends, anticipate future conditions, and navigate the complexities of the housing market effectively.

The housing market plays a crucial role in a country's economy and significantly impacts the financial well-being of individuals and families. Housing prices are influenced by a multitude of factors, including local market demand, the economic environment, interest rates, and government policies (Paciorek, 2013). As such, understanding these influences and being able to predict housing prices is of great interest to economists, policymakers, investors, and home buyers. The price of housing is not only a reflection of property characteristics such as size, location, and amenities but also macroeconomic indicators like inflation, unemployment rates, and mortgage interest rates (Kain & Quigley, 1970). Numerous studies have shown that housing prices can be volatile, responding to sudden changes in these factors (Glaeser, Gyourko, & Saks, 2005). This volatility poses risks to stakeholders in the housing market, making accurate predictions of housing prices critical for informed decision-making.

With advancements in data science and machine learning, predictive models for housing prices have improved significantly. Machine learning algorithms can leverage vast amounts of historical data, property

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

characteristics, and economic variables to generate more accurate price forecasts (Bajari, Chu, & Park, 2012). Traditional models, such as hedonic pricing models, use regression analysis to estimate the influence of various factors on housing prices. However, machine learning techniques, like random forests, gradient boosting, and neural networks, have been shown to outperform traditional models due to their ability to capture nonlinear relationships between features (Galster & Tatian, 2009). Studies by Bajari et al. (2012) and Wen, Yang, and Song (2019) highlight the effectiveness of machine learning models in predicting real estate prices. These models consider an array of variables, including structural features (e.g., square footage, age of property), location features (e.g., proximity to schools, transportation), and economic indicators (e.g., interest rates, inflation). The integration of such variables into predictive models offers greater flexibility in adjusting for changing market conditions and helps stakeholders make better-informed decisions. Even with these improvements, there are still hurdles in creating dependable forecasting systems. Real estate markets tend to be diverse, so the things that impact home prices in one area can be quite different from those in another. In addition, the availability and quality of data might vary a lot, which could affect how well the forecasting systems work, the housing data was analysed from townhouses in Fairfax Country and compared the classification accuracy performance of various algorithms. To help a real estate agent, he then developed a better prediction model for enhanced decisions based on house price assessment. Jafari and Akhavian (2019) stated that the square footage of a unit of a house is the most important variable in predicting the price of a house, followed by the number of bathrooms and number of bedrooms. Raga Madhuri, Anuradha, & Vani Pujitha, (2019) discussed diverse regression techniques such as Gradient boosting and AdaBoost Regression, Ridge, Elastic Net, Multiple linear, and Lasso to locate the most excellent. The performance measures used are Mean Square Error (MSE) and Root Mean Square Error (RMSE). Predicting the failure of industrial equipment's mechanical parts Regression is a supervised learning technique that aims to find the relationships between the dependent and independent variables. Ridge regression is a method of estimating the coefficients of multiple regression models in scenarios where the independent variables are highly correlated (Hilt, & Seegrist, 1977). It has been used in many fields including econometrics, chemistry, and engineering (Gruber, & Schucany, 2020). Uzoma, & Jeremiah, (2016) developed outlier detection and optimal variable selection techniques in regression analysis and other fascinating papers by the research include (Anabike *et al.*, 2023; Innocent *et al*., 2023; Abuh, Onyeagu, & Obulezi, 2023a; Abuh, Onyeagu, & Obulezi, 2023b; Obulezi *et al*., 2022; Onyekwere, & Obulezi, 2022; Onyekwere *et a*l., 2022). This section summarizes the concept of relevant work on Prediction of House Prices in Lagos Nigeria using a machine learning model. Here, the house price prediction can be divided into two categories (Zulkifley *et al*., 2020), first by focusing on house characteristics, and secondly by focusing on the model used in house price prediction. Many researchers have produced a house price prediction model, including Temur, Akgün, & Temur, (2019), Jafari, & Akhavian, (2019), Gao *et al*.

METHOD

Data Collection Method

This study use secondary data obtained from several real estate's company platforms specialized in the purchasing and selling of land, building of houses (bungalow, apartments, shops, etc) in Lagos State, Nigeria.

Real Estate Platforms:
Extract data from popular property listing websites in Lagos (e.g., PropertyPro.ng, PrivateProperty.com.ng) to obtain information on housing prices, features, and availability.

Government and Agency Reports:
Use reports and datasets from relevant Nigerian agencies like the Lagos State Ministry of Housing, National Bureau of Statistics (NBS), and Central Bank of Nigeria (CBN) for economic and demographic data.

Historical Market Data:
Collect historical property transaction records, where available, from real estate firms or public registries

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Data Analysis

The analysis of the data followed a multi-step process, utilizing both quantitative and qualitative methods to interpret the information collected. The analytical approach used in this study included the following techniques:

Simple Linear Regression Analysis

Linear regression assumes a linear relationship between the dependent variable γ (e.g., housing price) and independent variables X (e.g., size, number of bedrooms, location). The goal is to find the best-fitting line (or hyperplane in multiple dimensions) that predicts Y based on the X variables.

Model Representation

The relationship is typically expressed with the following equation:

$$\boldsymbol{\gamma = \beta_0 + \beta_1 X + \in}$$

Where

γ : The **dependent variable** (the outcome being studied, such as hosing price).

β_0: The **intercept** (the value of γ when all independent variables are zero).

β_1: The **coefficient** of the independent variables. These represent the change in the dependent variable for a one-unit change in the corresponding independent variable.

X: The **independent variables** (predictors), which may include property size, lot size or the dependent variable.

$\in$: The **error term**, accounting for the variance in γ not explained by the independent variables. It captures the factors that affect γ but are not included in the model

Model Form:

$$\textbf{Price} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ \textbf{(Size)} + \epsilon$$

Multiple linear regressions Analysis

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between one **dependent variable** (in this case, housing price) and **multiple independent variables** (like square footage, number of bedrooms, location factors, etc.). It is a common approach for predicting housing prices because it can handle the complexity of multiple factors influencing a single outcome.

Model

$$\boldsymbol{\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots\ldots + \beta_n X_n + \in}$$

Where

γ : The **dependent variable** (the outcome being studied, such as maternal mortality ratio).

β_0: The **intercept** (the value of γ when all independent variables are zero).

$\beta_1, \beta_2, \ldots \beta_n$: The **coefficients** of the independent variables. These represent the change in the dependent variable for a one-unit change in the corresponding independent variable.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

$X_1, X_2, \ldots X_n$: The **independent variables** (predictors), which may include socio-economic, healthcare access, or demographic factors

$\in$: The **error term**, accounting for the variance in γ not explained by the independent variables. It captures the factors that affect γ but are not included in the model.

Fit the Model: Use statistical software (e.g., Python's scikit-learn or R) to fit the MLR model to the training data, estimating coefficients for each independent variable.

Interpret Coefficients: Each coefficient β_1 indicates the effect of that variable Xi on Y, holding other variables constant. For example, in a housing study, the coefficient for **square footage** shows how much price is expected to increase per additional square foot.

Residual Analysis: Examine the residuals to ensure assumptions (like homoscedasticity and normality) are met.

Model Form:

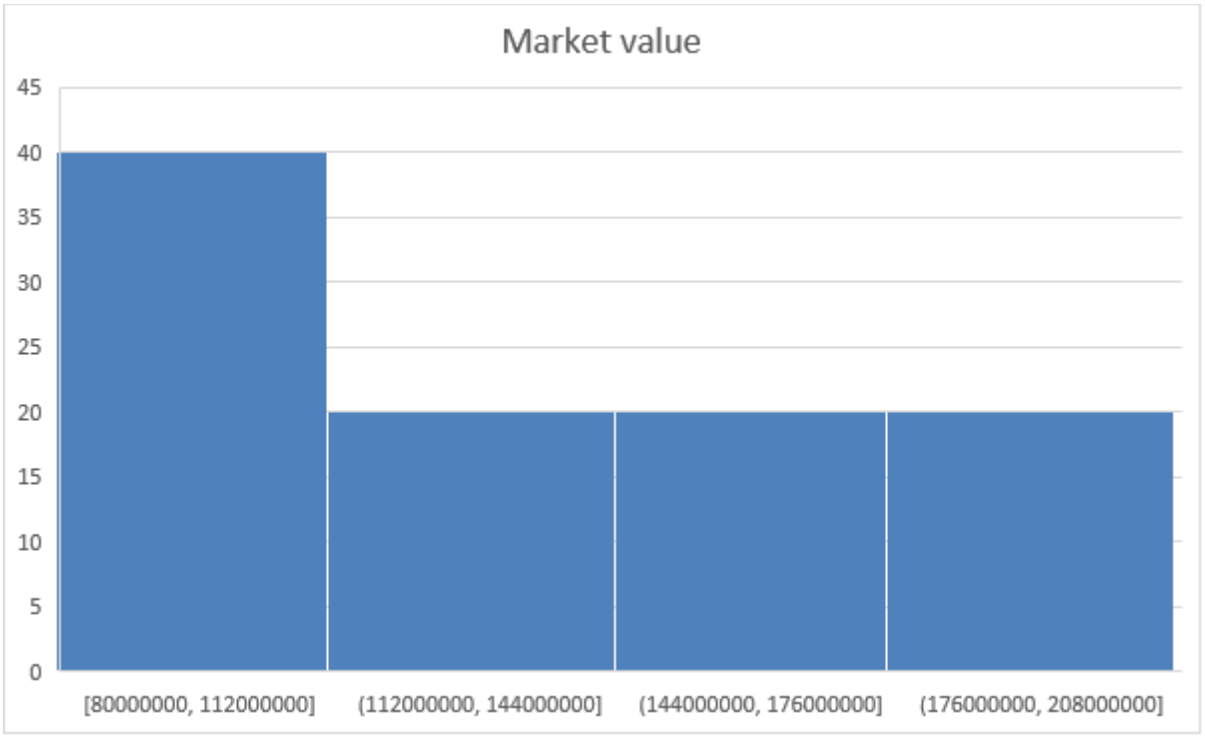
$$\textbf{Price} = \beta_0 + \beta_1(\textbf{Size}) + \beta_2(\textbf{Bedrooms}) + \ldots +$$

Table 1 Descriptive Analysis

	Sample Size	Max	Min	Mean	Standard dev	Kurtosis	Skewness
Market value	100	200000000	80000000	130000000	42163702.14	-0.971445666	0.544524852
Average monthly rent	100	1000000	400000	680000	214617.348	-1.283177763	0.140134002
Property age	100	14	2	7.25	4.680445028	-1.440113222	0.33644709
Power supply reliability hrs per day	100	20	12	16.25	3.046358979	-1.423817641	-0.187987711
Distance to closest major road km	100	2	0.5	1.25	0.56183322	-1.368045445	0
Number of bedrooms	100	5	1	3	1.42133811	-1.304933726	0
Number of bathrooms	100	5	1	3	1.42133811	-1.304933726	0
Parking spaces	100	3	1	1.99	0.822597512	-1.522682477	0.018698722
Size sqm	100	300	100	200	71.06691	-1.30493	0

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Interpretation



Market Value

The market value of properties ranges from 8,000,000 to 20,000,000, with an average value of 13,000,000. The standard deviation of 42,163,702.14 indicates significant variability, reflecting a wide range of property values. The positive skewness (0.5445) suggests that slightly more properties have market values below the average. The kurtosis value (0.9714) is close to zero, indicating a near-normal distribution.

Average Monthly Rent

The monthly rent for properties varies between 400,000 and 1,000,000, with a mean rent of 680,000. The standard deviation of 214,617.35 suggests moderate variability in rent prices. The skewness (0.1401) is close to zero, indicating a nearly symmetric distribution, while the kurtosis (1.2832) shows a slightly peaked distribution compared to a normal curve.

Property Age

The age of properties ranges from 2 to 14 years, with an average age of 7.25 years. A standard deviation of 4.68 years reflects moderate variability. The positive skewness (0.3364) indicates a slight tendency for properties to be younger than the average. The kurtosis (1.4401) suggests a slightly more peaked distribution than normal.

Power Supply Reliability (Hours per Day)

Power supply reliability ranges from 12 to 20 hours daily, with a mean of 16.25 hours. The standard deviation of 3.05 hours indicates moderate variability. The skewness (-0.1880) is slightly negative, suggesting a small tendency for properties to have power supply reliability above the average. The kurtosis (1.4238) reflects a slightly peaked distribution.

Distance to Closest Major Road (km)

The distance to the nearest major road ranges from 0.5 km to 2 km, with an average of 1.25 km. A standard deviation of 0.56 km shows low variability. The skewness is zero, indicating a perfectly symmetric distribution, while the kurtosis (1.3680) suggests a slightly peaked distribution.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Number of Bedrooms

Properties have between 1 and 5 bedrooms, with an average of 3 bedrooms. The standard deviation of 1.42 indicates moderate variability. Both skewness and kurtosis are zero, suggesting a symmetric and normal distribution.

Number of Bathrooms

The number of bathrooms also ranges from 1 to 5, with an average of 3 bathrooms. The standard deviation (1.42), skewness (0), and kurtosis (0) are identical to the number of bedrooms, indicating similar distributional characteristics.

Parking Spaces

The number of parking spaces ranges from 1 to 3, with a mean of 1.99 spaces. The standard deviation of 0.82 reflects low variability. The skewness (0.0187) is close to zero, indicating an almost perfectly symmetric distribution, while the kurtosis (1.5227) suggests a slightly peaked distribution.

Size (Square Meters)

Property sizes range from 100 to 300 square meters, with an average size of 200 square meters. The standard deviation of 71.07 reflects moderate variability. The skewness is zero, indicating a symmetric distribution, while the negative kurtosis (-1.3049) suggests a flatter (platykurtic) distribution compared to normal.

Table 2 Linear Regression

	coefficients	Standard error	T stat
Intercept	38000000	8029283.141	4.732676546
Size sqm	460000	37850.40371	12.15310683

Interpretation

The regression results indicate a significant relationship between the size of an asset (measured in square meters) and its estimated value. The model suggests that for every additional square meter of size, the value increases by **460,000 currency units** on average. This relationship is supported by a strong t-statistic of **12.15**, indicating high statistical significance.

The intercept of the model is **38,000,000**, which represents the estimated value of the asset when its size is zero. While this value may not have a practical interpretation in real-world terms, it serves as a baseline for the regression equation.

Overall, the findings demonstrate that size is a crucial factor in determining the asset's value, and the model provides a robust fit, as evidenced by the statistically significant coefficients for both the intercept and the size variable.

Table 3 Multiple Regression

	coefficients	Standard error	T stat
Intercept	-4031937.86	20372886.52	-0.19790705
Average monthly rent	174.7739625	4.934564662	35.41831437

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Property age	1276.842391	1366661.045	0.000934279
Power supply reliability hrs per day	-5.97447e-10	392576.0752	-1.5219e-15
Parking space	31921.05978	882721.7619	0.036162085
Size sqm	75500.47462	14893.22067	5.069452491

The regression analysis provides insights into the factors influencing the market value of properties. Here is a detailed interpretation of the results:

The intercept, with a coefficient of -40,319,387.86, represents the predicted market value when all independent variables are zero. However, this value lacks practical meaning in this context, as it is unrealistic for properties to have negative market values. The very low t-statistic (-0.1979) further suggests that the interception is not statistically significant.

Among the independent variables, the average monthly rent stands out as a highly significant predictor of market value. With a coefficient of 174.77, it indicates that for every 1-unit increase in rent, the market value is expected to rise by 174.77 units. The small standard error (4.93) and the very high t-statistic (35.42) confirm its strong predictive power.

Property age has a coefficient of 1,276.84, suggesting that for each additional year of age, the market value increases by this amount. However, the large standard error (1,366,661.05) and extremely low t-statistic (0.00093) indicate that property age is not a statistically significant factor in determining market value.

Power supply reliability, measured in hours per day, has a negligible coefficient (-5.97e-10), meaning it has no meaningful impact on market value. The extremely low t-statistics (-1.52e-15) further supports its lack of significance.

Parking spaces have a coefficient of 31,921.06, indicating a positive relationship with market value. However, the high standard error (882,721.76) and low t-statistic (0.036) suggest that the effect of parking spaces is not statistically significant in this model.

The size of the property, measured in square meters, is another significant predictor of market value. With a coefficient of 75,500.47, it indicates that for every additional square meter, the market value increases by this amount. The relatively low standard error (14,893.22) and high t-statistic (5.07) confirm its importance as a predictor.

Correlation

	Market value	Average monthly rent	Property age	Power supply reliability hrs per day	Distance to closest major road km+	Number of bedrooms	Number of bathrooms	Parking spaces	Size sqm
Market value	1								
Average monthly rent	0.98230	1							
Property age	0	0	1						
Power supply reliability hrs per day	0	0	-0.199	1					

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Distance to closest major road km	0	0	-0.984	0.33197	1				
Number of bedrooms	0.7753	0.7285	0	0	0	1			
Number of bathrooms	0	0	-0.984	0.33197	1	0	1		
Parking spaces	-0.0320	-0.0354	-0.012	-0.02721	0.00546	-0.00864	0.00546	1	
Size sqm	0.7753	0.7285	0	0	0	1	0	-0.00864	1

Interpretation

The correlation matrix reveals several significant relationships between the property-related variables. A very strong positive correlation exists between market value and average monthly rent (0.983), indicating that properties with higher market values tend to command higher rents. Similarly, market value is strongly associated with the number of bedrooms (0.775) and size in square meters (0.775), suggesting that larger properties and those with more bedrooms are generally more valuable.

The relationship between the average monthly rent and the number of bedrooms is also strong (0.729), showing that properties with more bedrooms typically have higher rental prices. Additionally, size in square meters correlates strongly with the number of bedrooms (0.775), highlighting that larger properties tend to accommodate more bedrooms.

Interestingly, there is a weak positive correlation between power supply reliability and distance to the closest major road (0.332), which suggests that properties farther from major roads might have slightly more reliable power supply. However, power supply reliability shows a weak negative correlation with property age (-0.199), indicating that older properties may experience slightly less reliable power.

The number of bedrooms and number of bathrooms are perfectly correlated (1.000), implying that properties with more bedrooms invariably have an equal number of bathrooms. On the other hand, the number of parking spaces shows little to no correlation with other variables, indicating that it is an independent feature of the property.

Overall, the strongest relationships involve market value, rent, size, and the number of bedrooms, while variables like parking spaces and power supply reliability show weaker or negligible associations with most other factors.

CONCLUSION

The study concludes that housing prices are primarily driven by economic and physical attributes, such as rental income potential and property size. These factors reflect the utility and desirability of properties in the market. On the other hand, aspects like property age, proximity to roads, and parking spaces may hold less relevance in certain contexts or regions. The results underscore the importance of focusing on key variables when analysing housing markets or predicting property prices.

REFERENCE

1. Abuh, J., Onyeagu, A., & Obulezi, M. (2023a). Regression modelling in predictive analytics.
2. Abuh, J., Onyeagu, A., & Obulezi, M. (2023b). Machine learning applications in econometrics.
3. Adams, Z., & Fuss, R. (2010). Housing prices and macroeconomic fundamentals: An empirical analysis across countries. Journal of Housing Economics, 19(2), 84-103.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

4. Anabike, E., Innocent, I., & Jeremiah, A. (2023). Advanced statistical techniques in regression analysis.
5. Bajari, P., Chu, C., & Park, M. (2012). "An Empirical Model of Subprime Mortgage Default from 2000 to 2007." Examines the use of machine learning models for predicting real estate market trends.
6. Byeonghwa, C. (2015). Comparative analysis of machine learning algorithms for housing price prediction.
7. Galster, G., & Tatian, P. (2009). "Modeling Housing Appreciation Dynamics." Highlights the application of statistical and machine learning techniques in housing market analysis.
8. Gao, Y., Chen, F., & Liu, Z. (2022). Enhancing prediction accuracy in housing price models with advanced regression techniques.
9. Glaeser, E. L., Gyourko, J., & Saks, R. E. (2005). "Why Have Housing Prices Gone Up Analyzes the drivers of housing price increases, focusing on supply and demand dynamics.
10. Gruber, M. H. J., & Schucany, W. R. (2020). Applications of ridge regression in econometrics, chemistry, and engineering.
11. Hilt, D. E., & Seegrist, R. R. (1977). Ridge regression: A method of estimating the coefficients of multiple regression models in scenarios where independent variables are highly correlated.
12. Ho, L. S., & Wong, G. (2008). Housing and macroeconomics: The role of asset prices in monetary policy. Pacific Economic Review, *13(2), 223-239.*
13. Hossain, M., & Latif, E. (2016). Housing price determinants in the context of macroeconomic variables. Economic Modelling, *59, 174-183.*
14. Jafari, M., & Akhavian, R. (2019). Square footage and its impact on house price prediction models.
15. Kain, J. F., & Quigley, J. M. (1970). "Measuring the Value of Housing Quality." Explores the relationship between housing characteristics and prices using hedonic pricing models.
16. Khoo, K. L., Lee, C. H., & Goh, K. L. (2019). Housing prices and macroeconomic variables:
17. Paciorek, A. (2013). "Supply Constraints and Housing Market Dynamics." Discusses how supply constraints influence housing prices and market behavior.
18. Raga Madhuri, G., Anuradha, G., & Vani Pujitha, K. (2019). Regression techniques for house price prediction: A comparative study.
19. Temur, A., Akgün, A., & Temur, S. (2019). Gradient boosting and AdaBoost Regression for house price prediction.
20. Yu, J., Liu, H., & Zhang, W. (2016). The influence of structural features on housing price prediction.