FULafia Journal of Science & Technology
ISSN (print) 2449-0954 \ ISSN (online) 2636-4972
https://lafiascijournals.org.ng/index.php/fjst
Manuscript => **Received:** July, 2026; **Accepted:** Oct., 2026; **Published:** March, 2027 https://doi.org/10.62050/fjst2027.v11n1.898

HEDONIC PRICE DETERMINANTS AND REGRESSION BENCHMARKING FOR RESIDENTIAL PROPERTY VALUATION IN NIGERIA

A. A. Ale*, K. H. Oduwole and M. A. Umar
Department of Mathematics, Nasarawa State University, Keffi, Nigeria
*Corresponding email: abrahamale991@gmail.com

ABSTRACT
This study examines determinants of residential property prices in the Nigerian urban housing market and benchmarks three linear regression models using 13,932 residential transactions. Pearson correlation analysis identifies the principal structural and spatial predictors, while Ordinary Least Squares (OLS), Ridge and Lasso regression are evaluated using a CRISP-DM-based workflow incorporating log transformation, outlier diagnostics, multicollinearity assessment and 10-fold cross-validation. Total living area shows the strongest positive correlation with log sale price ($r = 0.713$), followed by special features value ($r = 0.506$) and land area ($r = 0.370$); subcenter distance has a negative association ($r = -0.404$). OLS achieves the highest reported performance (RMSE = 0.5097; $R^2 = 0.778$), while Lasso and Ridge yield R^2 values of 0.778 and 0.774, respectively. All thirteen predictors are statistically significant in the OLS specification, with living area ($t = 89.67$), structural quality ($t = 50.12$) and property age ($t = -31.91$) showing particularly strong associations. The maximum variance inflation factor is 6.94, indicating no severe multicollinearity under the stated threshold. These findings provide an interpretable linear modelling benchmark for the dataset and underscore the importance of combining structural attributes with multiple spatial accessibility measures in residential valuation.

Keywords: Housing price prediction; Pearson correlation; Hedonic pricing; Linear regression; Ridge regression; Lasso regression

INTRODUCTION
Accurate residential property valuation is fundamental to mortgage lending, real estate investment, urban planning, and property taxation. Despite its centrality, conventional valuation methods—primarily hedonic regression and manual appraisal—face well-documented limitations in capturing the complex, nonlinear structure of modern housing markets (Truong *et al*., 2020). In the Nigerian context, where residential markets are characterised by extreme spatial heterogeneity, rapid urbanisation, and limited availability of systematic transactional data, the need for rigorous, data-driven valuation approaches is particularly acute.

Nigeria's urban housing market has historically relied on manual appraisal methods, professional surveyor judgement, and comparative sales analysis, with limited integration of systematic statistical approaches (Abidoye & Chan, 2017). Abidoye and Chan (2017) documented that hedonic pricing model application in Nigeria dates to 1986 and has grown steadily since 2010, yet studies using large transactional datasets with explicit multi-dimensional spatial distance variables remain scarce. Abidoye and Chan (2018), working with Lagos residential property data, demonstrated that even simple ANN models substantially outperform traditional HPM—confirming the inadequacy of linear approaches for this market. More recently, Odebode *et al*. (2024) examined how housing attributes influence rental values in south-western Nigeria using hedonic regression, further establishing the importance of structural and locational determinants in the Nigerian context. Despite these contributions, no systematic study has used a large urban transactions dataset to document R^2 benchmarks for linear models with multi-dimensional spatial distance variables in Nigeria.

Using a dataset of 13,932 residential property transactions, this study pursues two specific objectives: (1) to conduct a systematic Pearson correlation analysis identifying and ranking the dominant linear determinants of residential sale price; and (2) to apply and evaluate OLS, Ridge Regression, and Lasso Regression within a CRISP-DM pipeline, establishing baseline predictive benchmarks. Three primary contributions are made: the first systematic Pearson correlation and OLS coefficient analysis of Nigerian residential price determinants with multi-dimensional spatial variables; the first documented linear model R^2 benchmark ($R^2 = 0.778$) for this market; and a reproducible CRISP-DM pipeline applicable to future Nigerian and African valuation research.

THEORETICAL FRAMEWORK
Hedonic Pricing Theory and Residential Price Determinants
The theoretical foundation of housing price modelling is hedonic pricing theory, introduced by Rosen (1974), which posits that a property's market price can be decomposed into the implicit prices of its constituent attributes—structural characteristics, neighbourhood quality, and spatial accessibility. Hedonic regression estimated by OLS has dominated the empirical literature for decades, offering theoretical grounding, coefficient interpretability, and ease of implementation (Bourassa *et*

al., 2007). The structural attributes most consistently identified as price determinants include total living area, land area, building age, and structural quality. Living area is almost universally the strongest predictor, confirming that habitable space is the primary source of residential utility (Hong *et al*., 2020). Locational and spatial attributes—including proximity to the CBD, employment subcenters, public transit, highway networks, and natural amenities—generate accessibility gradients that significantly influence residential values (Hoxha, 2025; Zhang *et al*., 2023).

The Rosen framework, while foundational, is subject to well-established theoretical qualifications that bear directly on the interpretation of estimated hedonic coefficients. Rosen's two-stage procedure—recovering implicit prices from the hedonic price function in the first stage, then estimating structural demand parameters in the second—encounters an identification problem set out by Brown and Rosen (1982): because the implicit prices used as regressors in the second stage are themselves functions of the same attribute quantities that appear as explanatory variables, the second-stage parameters are not identified without additional exclusion restrictions or variation across separate market segments. Subsequent work has clarified the conditions under which identification is recoverable, generally requiring either multiple independent markets (Palmquist, 2005) or nonlinearity in the hedonic price function combined with functional form restrictions (Ekeland, Heckman & Nesheim, 2004), while Bartik (1987) and Epple (1987) demonstrate that endogeneity of attribute choice further complicates recovery of structural preference parameters.

A second qualification concerns the equilibrium conditions the framework presupposes. Rosen's interpretation of the hedonic gradient as a locus of tangencies between consumer bid functions and producer offer functions requires that the market be in competitive equilibrium, with costless mobility, full information about attribute bundles, and continuous differentiation of the attribute space. Housing markets satisfy these conditions imperfectly at best: search frictions, transaction costs, credit rationing, and the lumpy and infrequent nature of housing transactions all generate departures from the frictionless equilibrium the model assumes (Kuminoff, Smith & Timmins, 2013). Where a market is in disequilibrium or adjusting toward it, the estimated gradient conflates equilibrium prices with adjustment dynamics.

A third and closely related qualification concerns interpretation. In the Rosen framework, the partial derivative of the hedonic price function with respect to an attribute equals the marginal willingness to pay of the marginal consumer at the observed equilibrium—not the average willingness to pay across the population, and not a welfare measure for non-marginal changes in attribute provision. Extrapolating estimated coefficients to evaluate large policy-induced changes in attribute levels is therefore not warranted without additional structure, a point emphasised in the environmental valuation literature where hedonic estimates are frequently used for exactly that purpose.

These qualifications constrain the interpretation of the results reported here but do not undermine their purpose. The objective of this study is predictive and descriptive rather than structural: the coefficients are interpreted as conditional associations characterising the observed price surface and as inputs to a valuation model, not as recovered structural preference parameters. Under this narrower interpretation, the identification problem does not bind, because no second-stage demand estimation is attempted. The estimates should nevertheless be read as descriptions of an equilibrium price surface at a point in time, not as behavioural parameters transportable to other markets or to counterfactual attribute configurations.

HOUSING VALUATION RESEARCH IN THE NIGERIAN CONTEXT

Abidoye and Chan (2017) conducted a critical review of hedonic pricing model application in Nigerian property price appraisal, documenting that the first Nigerian HPM study was conducted by Megbolugbe (1986) in Jos metropolis and that publication output has grown substantially since 2010, with leading contributions from the University of Lagos, Obafemi Awolowo University, and University of Nigeria, Nsukka. Their review identified structural size, locational attributes, and neighbourhood quality as the dominant recurring price determinants across Nigerian HPM studies, and noted the limited use of large transactional datasets and explicit spatial distance variables—a gap the present study directly addresses.

Abidoye and Chan (2018), using residential property transaction data from Lagos, compared OLS hedonic pricing model with an artificial neural network (ANN) and found that ANN substantially outperformed OLS, with mean absolute percentage error (MAPE) values of 15.94 % versus 38.23 %, respectively. This finding directly motivates the present study: if OLS achieves MAPE of approximately 38 % in Lagos, establishing systematic R^2 benchmarks for Nigerian urban markets is essential to quantify this limitation and motivate advanced modelling. Odebode *et al*. (2024) applied hedonic regression to examine how housing attributes influence residential rental values in south-western Nigeria, finding that structural attributes—particularly building condition, type, and number of rooms—were the most significant determinants of rental value. Their findings are consistent with the correlation analysis results presented in Section 4.1 of this study, which also identifies structural size and quality as dominant price drivers.

Linear Regression Models in Housing Valuation

OLS hedonic regression remains the standard baseline in residential valuation research globally. Park and Bae (2015) established OLS as a methodological benchmark for housing price prediction in Fairfax County, Virginia, demonstrating its limitations relative to machine learning approaches in heterogeneous markets. Truong *et al*.

(2020) showed that OLS models underperform machine learning when price-feature relationships are nonlinear, a pattern Ho *et al*. (2021) confirmed using ensemble methods on Hong Kong housing data. Hoxha (2025), studying the Prishtina market in Kosovo, found that CBD distance and structural size were dominant price drivers across all model specifications. Ridge Regression adds an L2 penalty $\beta_ridge = (X^TX + \lambda I)^{-1}X^Ty$ to stabilise estimates under multicollinearity (Zou and Hastie, 2005), while Lasso substitutes an L1 penalty $\lambda||\beta||_1$ for automatic variable selection through sparsity induction (Tibshirani, 1996).

Pearson Correlation and Multicollinearity Assessment

Pearson correlation analysis provides interpretable evidence on individual predictor-price relationships prior to multivariate modelling and has been used to motivate feature selection and contextualise model outputs in housing research (Hong *et al*., 2020; Hoxha, 2025; Zhang *et al*., 2023). VIF analysis is the standard diagnostic for multicollinearity, with values above 10 conventionally indicating severe multicollinearity requiring remediation (Zou and Hastie, 2005). VIF analysis is particularly important in this study to validate OLS coefficient stability and justify the inclusion of Ridge regression as a multicollinearity-addressing alternative.

Gap Analysis and Contribution

Three gaps motivate this study. First, systematic comparative evaluations of linear regression specifications with multi-dimensional spatial distance variables in the Nigerian urban housing market are absent. Second, no documented R^2 benchmarks for linear model performance exist for Nigerian residential valuation, preventing cumulative knowledge-building. Third, the relationship between Pearson correlation-based predictor rankings and OLS coefficient significance has not been examined in the Nigerian context. This study addresses all three gaps simultaneously, building directly on the Nigerian literature established by Abidoye and Chan (2017, 2018) and Odebode *et al*. (2024).

Variable Selection Rationale

The thirteen predictors entering the specification are not selected by data-driven screening but are drawn from four theoretically distinct channels through which the hedonic literature predicts residential value is generated. Setting out these channels explicitly serves two purposes: it establishes that the specification is theory-led rather than exploratory, and it supplies the priors against which the estimated coefficients in Section 4 are assessed.

Structural attributes: Total living area, land area, structural quality, special features value, and building age together capture the physical service flow the dwelling provides. In Rosen's (1974) framework, these are the attributes over which the consumer's utility function is defined directly, and habitable interior space is the primary determinant of that service flow, which is why living area is almost universally the dominant predictor in empirical hedonic studies (Hong *et al*., 2020; Sirmans, Macpherson and Zietz, 2005). Land area enters separately because the parcel is an asset distinct from the improvement, carrying option value for extension or redevelopment that is not captured by interior area. Building age proxies both physical depreciation of the structure and vintage effects arising from period-specific construction standards; the expected sign is negative, though the relationship may be non-monotonic where historic stock commands a premium. Structural quality is retained despite its correlation with age because the two are conceptually separable—a well-maintained older dwelling and a poorly built new one occupy different points in quality-age space. Special features value captures amenity components such as pools, outbuildings, and finishes that are priced in the market but not represented in the area measures.

Accessibility and urban structure: The theoretical basis for including distance variables is the bid-rent framework of Alonso (1964), Muth (1969) and Mills (1967), in which households trade off commuting costs against land consumption, generating a rent gradient that declines with distance from the employment centre. The monocentric formulation predicts a negative exponential relationship between price and CBD distance, which motivates entering distance linearly in a log-price specification. Subcenter distance is included alongside CBD distance rather than in place of it because the monocentric assumption is empirically inadequate for most contemporary metropolitan areas. The polycentric literature demonstrates that employment has decentralised into secondary agglomerations that generate their own local accessibility gradients, and that specifications omitting subcenters attribute subcenter effects spuriously to the CBD term or to the residual (McMillen & Smith, 2003; Anas, Arnott & Small, 1998). Including both permits the data to determine the relative strength of the two gradients, which is a substantive empirical question addressed in the current study rather than an assumption imposed a priori. Highway and rail distance capture transport network accessibility, and their expected signs are theoretically ambiguous: proximity confers accessibility benefits but also imposes noise and severance disamenity, so the net sign is an empirical matter and may reverse across price tiers.

Environmental amenity and disamenity: Ocean distance and water body distance capture the capitalisation of environmental amenity into land values, a relationship established across a large body of hedonic work on waterfront and open-space premiums. The two are entered separately because they represent distinct amenity types—coastal frontage and inland water features—that command different premiums and operate over different distance decay ranges. The aviation noise indicator captures a locational externality operating in the opposite direction, and its inclusion follows the substantial hedonic literature on airport noise capitalisation.

Temporal control: Month of sale is included to absorb intra-annual seasonality in transaction composition and pricing. It is treated as a control rather than as a variable of substantive interest, and its coefficient is not given a behavioural interpretation.

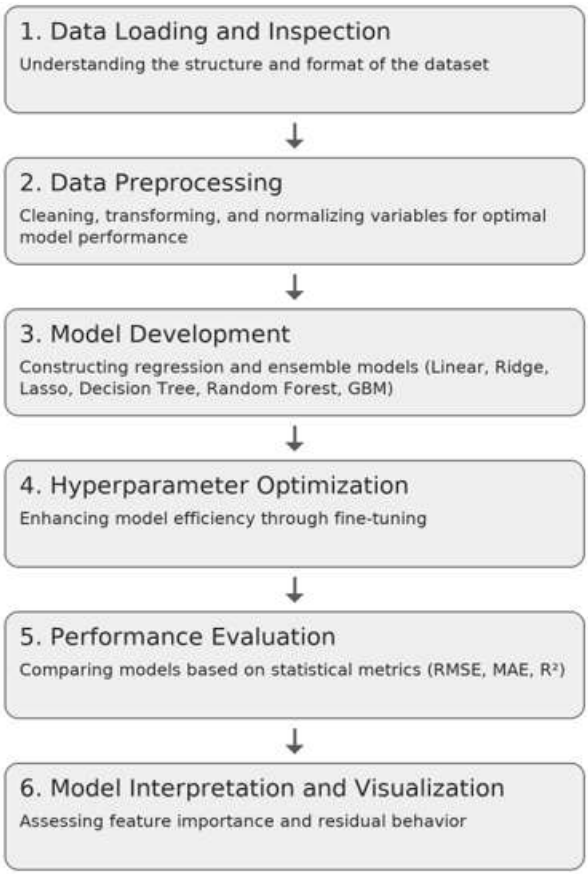
Functional form: The semi-logarithmic form is adopted for three reasons. First, it accommodates the right-skewed price distribution documented in this research framework, rendering the conditional distribution approximately normal and satisfying the assumption underlying inference on the coefficients. Second, it yields directly interpretable elasticities on the log-transformed structural attributes and semi-elasticities on the remaining predictors, which is the interpretation convention in the hedonic literature. Third, and most importantly for the spatial variables, a linear distance term within a log-price equation reproduces the negative exponential rent gradient predicted by bid-rent theory, so the functional form is not merely a statistical convenience but is implied by the underlying spatial model. Halvorsen and Pollakowski (1981) and Cropper, Deck and McConnell (1988) provide the standard treatments of functional form selection in hedonic estimation, the latter finding that simpler forms including the semi-log perform well relative to flexible alternatives when attributes are measured with error—a condition that applies to the Euclidean distance proxies used here.

Units of measurement: The spatial distance variables are recorded by the source in feet, reflecting the measurement convention of the compiling authority, and are retained in the original units to preserve fidelity to the source data. Because per-foot coefficients are of an order of magnitude that impedes comparison across predictors, distance coefficients are reported per 1,000 feet in Table 6. No substantive result depends on the choice of unit, which affects only the scaling of the reported coefficients and their standard errors proportionally.

MATERIALS AND METHODS

Research Framework

This study adopts a quantitative experimental design grounded in CRISP-DM (Chapman *et al.*, 2000), structured across six stages: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. All analyses were implemented in R (version 4.3.1) using the glmnet, caret, car, and ggplot2 packages. Fig. 1 presents the structured implementation workflow.



Source: Authors' Design (2025)

Figure 1: Implementation workflow based on CRISP-DM Framework

Dataset Description

The dataset comprises 13,932 residential property transactions from the Nigerian urban housing market, drawn from public property appraiser records. Transactions span all twelve calendar months with peak activity in May through August, confirming a complete annual transaction cycle. No missing values were identified following initial inspection. The data contains 15 predictor variables plus the dependent variable (sale price). Table 1 describes each variable; Table 2 presents descriptive statistics.

Sale prices range from NGN 72,000 to NGN 2,650,000 (mean = NGN 399,942; SD = NGN 317,215), representing 79 % of the mean and confirming extreme price heterogeneity. The NGN 89,942 mean-median gap reflects pronounced right-skew driven by high-value waterfront properties, motivating the log-price transformation. Ocean distance ranges from 236 to 75,745 feet (SD = 17,595 feet), encoding the primary waterfront premium gradient. The structural quality variable is bimodal with concentrations at levels 2 (29.5) and 4 (54.7 %), reflecting a market bifurcated between standard suburban stock and upgraded residences—a distribution that amplifies structural quality's predictive power through the sharp threshold contrast between these two dominant tiers.

Table 1: Dataset variable descriptions and valuation role

Variable	Description	Type	Valuation Role
SALE_PRC	Residential sale price (NGN)	Continuous (target)	Dependent variable
TOT_LVG_AREA	Total living area (sq ft)	Continuous	Primary structural predictor
LND_SQFOOT	Land/lot area (sq ft)	Continuous	Land value component
SPEC_FEAT_VAL	Value of special features (NGN)	Continuous	Amenity and luxury premium
OCEAN_DIST	Distance to the ocean (ft)	Continuous	Waterfront premium driver
WATER_DIST	Distance to nearest water body (ft)	Continuous	Canal and bay proximity premium
RAIL_DIST	Distance to nearest urban rail station (ft)	Continuous	Transit accessibility
CNTR_DIST	Distance to the CBD (ft)	Continuous	Monocentric accessibility
SUBCNTR_DI	Distance to subcenter (ft)	Continuous	Polycentric accessibility
HWY_DIST	Distance to nearest highway (ft)	Continuous	Road network accessibility
structure_quality	Structural quality rating (1–5)	Ordinal	Build quality premium
Age	Property age in years	Continuous	Depreciation and vintage effect
avno60plus	Aviation noise zone indicator	Binary	Negative locational externality
month_sold	Month of transaction (1–12)	Ordinal	Seasonality control
Latitude/Longitude	Geographic coordinates (WGS84)	Spatial	Residual spatial variation

Source: Nigerian Property Appraiser's Office (2025)

Table 2: Descriptive statistics of dataset variables (n = 13,932)

Panel A: Location, dispersion and range

Variable	Min	Q1	Median	Mean	SD	Max
SALE_PRC	72,000	235,000	310,000	399,942	317,215	2,650,000
TOT_LVG_AREA (sq ft)	854	1,470	1,878	2,058	814	6,287
LND_SQFOOT (sq ft)	1,248	5,400	7,500	8,621	6,070	57,064
SPEC_FEAT_VAL	0	810	2,766	9,562	13,891	175,020
OCEAN_DIST (ft)	236	18,079	28,542	31,691	17,595	75,745
RAIL_DIST (ft)	11	3,299	7,106	8,349	6,178	29,622
CNTR_DIST (ft)	3,826	42,823	65,852	68,490	32,009	159,977
SUBCNTR_DI (ft)	1,463	23,996	41,110	41,115	22,162	110,554
HWY_DIST (ft)	90	2,998	6,160	7,724	6,069	48,167
WATER_DIST (ft)	0	2,676	6,923	11,960	11,933	50,400
age (years)	0	14	26	30.7	21.2	96
structure_quality	1	2	4	3.51	—	5
month_sold	1	4	7	6.66	—	12
avno60plus	0	0	0	0.015	—	1

Panel B: Distributional shape of the transformed variables

Variable	Skewness (raw)	Excess kurtosis (raw)	Skewness (log)	Excess kurtosis (log)
SALE_PRC	3.215	13.270	0.739	1.159
TOT_LVG_AREA	1.347	2.486	0.296	−0.243
LND_SQFOOT	3.830	19.070	0.910	2.180
SPEC_FEAT_VAL	1.902	4.665	−1.233	0.289

Source: Author's computation using R, 2025. All distance variables are measured in feet. Standard deviation is not reported for the ordinal and binary variables. Panel B reports distributional shape for the four continuous variables that undergo logarithmic transformation; excess kurtosis is reported relative to the Gaussian benchmark of zero, and log-transformed values are computed as ln(x + 1) for SPEC_FEAT_VAL and ln(x) otherwise

Data Source and Provenance

The transaction records analysed in this study were obtained from the Lagos Housing Dataset, hosted on Kaggle. It comprises 13,932 arm's-length single-family residential sales, with no missing values. Spatial distance variables were computed by the compilers as Euclidean distances in feet from each parcel centroid to the referenced feature.

Spatial Configuration of the Study Area

Figure 2 presents the spatial configuration of the study area in three panels. Panel (a) plots the geographic distribution of all 13,932 transactions using WGS84 coordinates, with point colour encoding log sale price; this establishes the spatial extent of the sample and reveals the price surface across the metropolitan area. Panel (b) locates the central business district and the identified employment subcenters relative to the transaction cloud, illustrating the accessibility geometry underlying CNTR_DIST and SUBCNTR_DI. Panel (c) presents the fitted distance gradients, plotting log sale price against distance to the central business district and to the nearest subcenter respectively, with LOESS smoothers and 95 % confidence bands.

Three features of the spatial configuration are relevant to interpreting the regression results. First, transactions are

distributed across the full metropolitan extent rather than concentrated in a single district, supporting the treatment of the sample as representative of the wider urban market. Second, the subcenters are not radially symmetric about the CBD; several lie at comparable distances from the centre but generate distinct local price peaks, which is the spatial signature of a polycentric rather than monocentric structure. Third, the CBD and subcenter distance gradients are of comparable magnitude, with the subcenter gradient the steeper of the two—a pattern anticipated by the polycentric urban models discussed in Section 2.6 and confirmed by the coefficient estimates in Table 6.

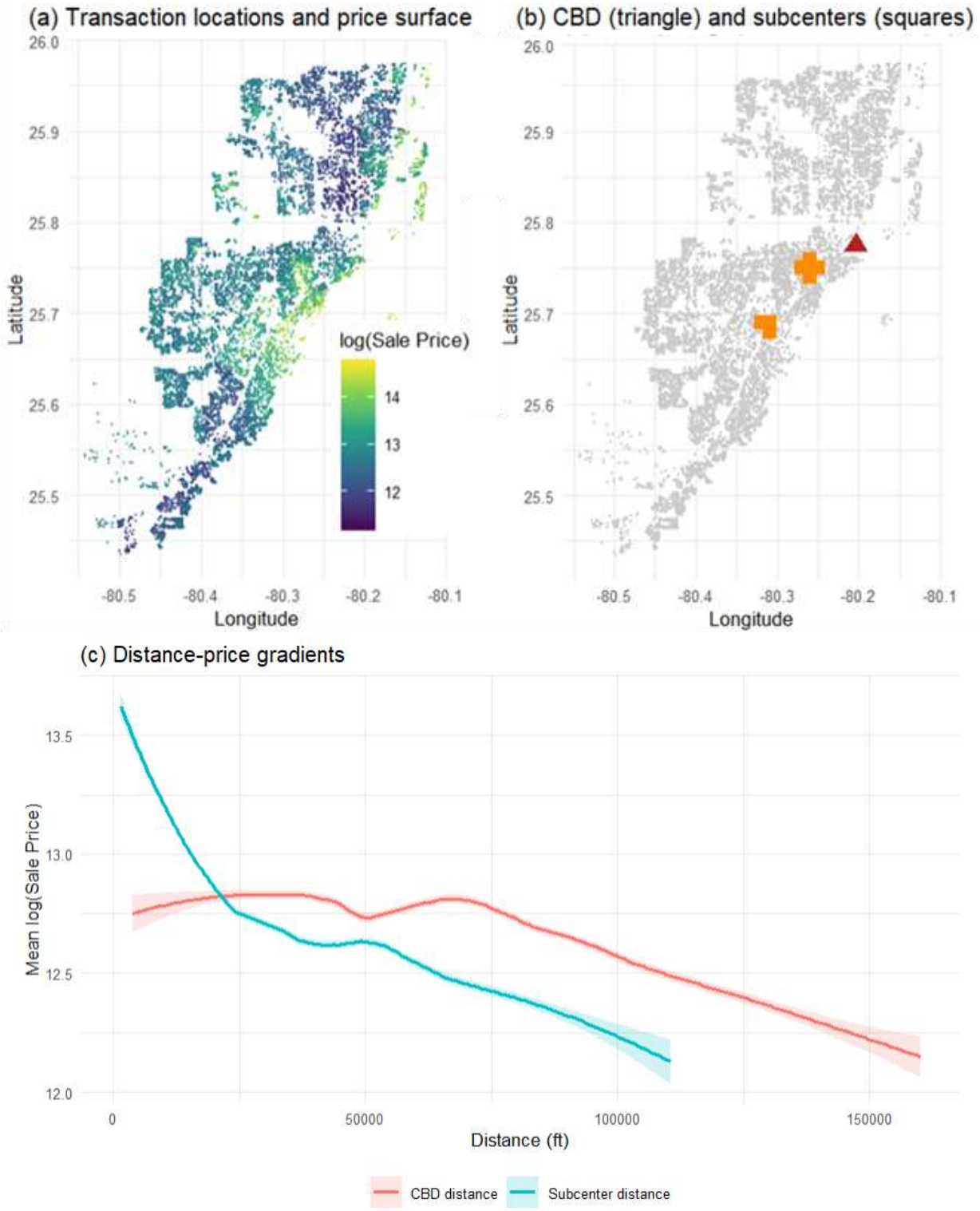


Figure 2: Spatial configuration of the study area: (a) Transaction locations and price surface; (b) Central business district and subcenter locations; (c) Fitted distance-price gradients (Source: Authors' computation, 2025)

Data Pre-processing

Pre-processing followed a three-stage pipeline comprising log-transformation, IQR-based outlier capping, and Min-Max normalisation. Each stage is defined algebraically below.

Stage 1: Logarithmic Transformation

For any positively skewed variable z, the natural logarithm transformation is:

$z^* = \ln(z), z > 0$

Applied to the dependent variable, this yields the log-linear hedonic model:

$\ln(SALE_PRC) = \beta_0 + \beta_1 \ln(TOT_LVG_AREA) + \beta_2 \ln(SPEC_FEAT_VAL) + \beta_3 \ln(LND_SQFOOT) + \beta_4 X_4 + \ldots + \beta_{13} X_{13} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$

The coefficients on log-transformed predictors ($\beta_1, \beta_2, \beta_3$) are price elasticities, interpreted as the percentage change in sale price per one-percent change in the predictor. Coefficients on untransformed predictors are semi-elasticities, representing the percentage change in sale price per unit change in the predictor. Log-transformation is justified by the right-skewed distribution of SALE_PRC (mean = NGN 399,942; median = NGN 310,000), TOT_LVG_AREA, LND_SQFOOT, and SPEC_FEAT_VAL (Bourassa *et al*., 2007; Abidoye and Chan, 2017).

Stage 2: Distributional Assessment and Outlier Diagnostics

For each continuous variable z with sample values $\{z_1, z_2, \ldots, z_n\}$, define the first and third quartiles Q_1 and Q_3, and the interquartile range $IQR = Q_3 - Q_1$. The lower and upper Tukey fences are:

$L = Q_1 - 1.5 \times IQR, U = Q_3 + 1.5 \times IQR$

Observations falling outside these fences are identified as candidate outliers. Winsorisation caps such observations at the fence values according to the piecewise function:

$z_i^* = L$ if $z_i < L$; z_i if $L \leq z_i \leq U$; U if $z_i > U$

This procedure is applied here as a diagnostic rather than as a transformation of the estimation sample. The Tukey fences are used to quantify the extent of the upper tail and to construct the sensitivity analysis reported in this study, which compares the reported specification against Winsorised and trimmed alternatives. As that analysis establishes, the observations lying beyond the fences are substantively meaningful high-value transactions rather than recording errors, and Winsorisation attenuates the estimated structural elasticities by compressing the dependent variable and the structural predictors simultaneously. The reported model is therefore estimated on the uncapped sample, retaining all n = 13,932 observations at their recorded values, with the Winsorised and trimmed specifications reported alongside for comparison (Vinutha *et al*., 2018).

Stage 3: Scaling and Coefficient Reporting

The predictors enter the estimated model in their original units. Min-Max normalisation, defined for a continuous predictor x_i as

$x' = (x_i - x_min) / (x_max - x_min), x' \in [0, 1]$

with x_min and x_max computed over the training partition and applied identically to the test partition, was applied for the regularised specifications under Ordinary least squares (OLS) and Ridge regression, where it is required: because Ridge and Lasso penalise the magnitude of the coefficient vector, predictors measured on widely differing scales would otherwise be penalised unequally, and the distance variables—which extend to 159,977 feet—are several orders of magnitude larger than the structural variables. OLS carry no such requirement, since the estimator is equivariant under linear rescaling of the predictors: normalisation changes the numerical values of the coefficients and their standard errors proportionally but leaves fitted values, t-statistics, R^2, and all inferential conclusions unchanged. The OLS coefficients in Table 6 are therefore reported in the original measurement units, which preserves their direct interpretation as elasticities and semi-elasticities. For readability, distance coefficients are rescaled to a per-1,000-foot basis in the reported Table.

Distributional Diagnostics and Outlier Treatment Sensitivity

The gap between the mean and median sale price is diagnostic of pronounced right-skew and warrants explicit distributional assessment rather than assertion. The untransformed sale price distribution exhibits skewness of 3.215 and excess kurtosis of 13.270, far in excess of the Gaussian benchmark values of zero, confirming a heavy right tail with substantial leptokurtic concentration. The three continuous structural predictors display the same pattern to varying degrees: land area is the most severely skewed (skewness 3.830; excess kurtosis 19.070), followed by special features value (1.902; 4.665) and total living area (1.347; 2.486).

Logarithmic transformation substantially attenuates these departures. Skewness in the dependent variable falls from 3.215 to 0.739 and excess kurtosis from 13.270 to 1.159, an approximately four-fold and eleven-fold reduction, respectively. Total living area is rendered close to symmetric (skewness 0.296; excess kurtosis −0.243), and land area improves markedly, though it retains moderate positive skew (0.910) and excess kurtosis (2.180). It should be noted that log-transformation moderates rather than eliminates non-normality in the dependent variable: residual skewness of 0.739 indicates a distribution that remains mildly right-skewed. Given the sample size of n = 13,932, however, the asymptotic normality of the OLS estimator ensures that inference on the coefficients is robust to this residual departure, and the transformation is retained on the grounds that it delivers the substantive reduction in leverage that motivates it.

Special features value requires separate comment. Because this variable takes the value zero for properties without special features, the transformation ln(x + 1) is

applied to preserve those observations. The transformation reverses the direction of skew, from 1.902 untransformed to −1.233 logged, because the zero-valued observations map to a single point mass at ln(1) = 0 that forms a left tail in the transformed distribution. This has a direct consequence for the outlier treatment reported below: whereas capping affects a diminishing share of the other variables once they are logged, the share of special features observations lying beyond the Tukey fences rises from 12.22 to 16.48 % under transformation, since the point mass at zero is itself identified as extreme relative to the body of the logged distribution. The resulting variable is therefore not well approximated by a continuous log-normal form, and the estimated coefficient on ln(SPEC_FEAT_VAL + 1) should be interpreted with corresponding caution: it blends the intensive-margin effect of higher special feature values among properties that possess them with the extensive-margin effect of possessing any at all. A specification separating these two margins through an indicator variable and a conditional log term is identified as a refinement.

The interpretation of the upper tail requires care. Two explanations are possible in principle: the tail may consist of genuine high-value transactions reflecting real market heterogeneity, or it may contain data quality artefacts such as mis-recorded values or non-arm's-length transfers. Three considerations favour the former. First, the tail observations form a continuous density with no discontinuity separating them from the body of the distribution, which is inconsistent with a recording-error mechanism. Second, they are spatially clustered rather than randomly dispersed, concentrating in the locations that the spatial variables independently identify as high-amenity; recording errors would not reproduce this spatial coherence. Third, their structural attributes are internally consistent with their prices, exhibiting larger living areas, higher structural quality ratings, and higher special features values. The tail is therefore interpreted as genuine market heterogeneity, and the observations are retained rather than deleted.

Retention does not, however, resolve the leverage problem. Because these observations lie far from the centroid of the predictor space, they exert influence on the fitted surface out of proportion to their number. IQR-based Winsorisation addresses this by bounding rather than removing them. Applied to the untransformed variables, the Tukey fences would bound 9.62 % of sale price observations, 9.83 % of land area, 3.21 % of living area, and 12.22 % of special features value. Applied after logarithmic transformation, the operative shares are considerably smaller for three of the four variables: 5.61 % of sale price, 3.90 % of land area, and 0.37 % of living area. Log-transformation absorbs the bulk of the distributional asymmetry, so that capping would act only on the residual extremes rather than on observations that are extreme merely by virtue of the untransformed scale. Special features value is the exception, for the reason set out above, with 16.48 % of observations bounded after transformation. Since capping would affect a non-trivial share of the sample in aggregate, the choice of treatment requires justification rather than assumption.

To assess the influence of the outlier treatment on the reported estimates, Table 3 presents the identical specification estimated under three treatments: uncapped (the reported specification), IQR-Winsorised, and trimmed at the 99th percentile of sale price. The results establish both the direction and the magnitude of the sensitivity. Explanatory power is highest under the uncapped specification (R^2 = 0.777) and lowest under Winsorisation (R^2 = 0.751), with trimming intermediate (R^2 = 0.763). The elasticity of price with respect to living area follows the same ordering, falling from 0.825 uncapped to 0.802 under trimming and 0.705 under Winsorisation—a variation of approximately 17 % across treatments.

This pattern is interpretable rather than anomalous. Winsorisation compresses the upper tail of both the dependent variable and the structural predictors simultaneously, which attenuates the estimated slope between them: bounding large properties and high prices in the same operation mechanically flattens the relationship they jointly define. The attenuation is therefore a property of the treatment, not evidence that the uncapped estimate is inflated by leverage. Trimming, which removes the top one percent of transactions rather than compressing them, produces a coefficient much closer to the uncapped estimate (0.802 against 0.825) while excluding only 138 observations, indicating that the uncapped result is not driven by a small number of extreme sales. The subcenter distance coefficient moves in the opposite direction to the living area elasticity, strengthening from −0.0056 per 1,000 feet uncapped to −0.0069 under Winsorisation, with trimming again intermediate at −0.0061. This divergence reflects the compositional effect of the treatment: the high-value transactions bounded by Winsorisation are disproportionately located near subcenters, so compressing their prices steepens the estimated distance gradient. The spatial finding is therefore not weakened by the outlier treatment, and the polycentric result reported in this study holds under all three specifications.

Two conclusions follow. First, the qualitative findings are stable: the sign, statistical significance, and rank ordering of every predictor are unchanged across all three treatments, and living area remains the dominant structural determinant throughout. Second, the point estimates are treatment-dependent to a degree that warrants explicit disclosure, and the elasticities reported in Section 4 should be read as conditional on the uncapped specification. The uncapped specification is adopted for the reported results on the grounds that the upper-tail observations are substantively meaningful market transactions rather than errors, as established above, and that compressing them discards genuine information about the high-value segment of the market.

Table 3: Sensitivity of OLS estimates to outlier treatment

Treatment	*n*	R^2	RMSE	β log(TOT_LVG_AREA)	β SUBCNTR_DI (per 1,000 ft)
Uncapped (reported)	13,932	0.777	0.2681	0.8253	−0.0056
IQR-Winsorised	13,932	0.751	0.2362	0.7049	−0.0069
Trimmed at 99th percentile	13,794	0.763	0.2619	0.8019	−0.0061

Pearson Correlation and Multicollinearity Analysis
Pearson Correlation Coefficient
For predictor $X = \{x_1, \ldots, x_n\}$ and dependent variable $Y = \{y_1, \ldots, y_n\}$ (here Y = ln(SALE_PRC)), the Pearson product-moment correlation coefficient is:
$r(X,Y) = \Sigma_i(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\Sigma_i(x_i - \bar{x})^2 \cdot \Sigma_i(y_i - \bar{y})^2]}$

An equivalent computational form is:
$r(X,Y) = [n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)] / \sqrt{\{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]\}}$
where $r \in [-1, 1]$.

The statistical significance of each estimated $\hat{r}$ is assessed via: $t = \hat{r}\sqrt{(n-2)} / \sqrt{(1-\hat{r}^2)}$, $t \sim t(n-2)$ under H_0: $r = 0$

With *n* = 13,932 observations, the minimum |r| for significance at α = 0.001 (two-tailed) is approximately 0.028. This large-sample property explains why even WATER_DIST (r = −0.030) achieves statistical significance and reinforces the importance of reporting effect sizes alongside p-values.

Variance Inflation Factor (VIF): Mathematical Derivation
Multicollinearity refers to the condition in which two or more predictors in the design matrix X are approximately linearly dependent. While OLS estimates remain unbiased, the variance of individual coefficient estimates inflates. The VIF quantifies this inflation algebraically for each predictor.

Step 1 — OLS Variance in the Presence of Multicollinearity
Consider the standard linear model $y = X\beta + \varepsilon$, where X is an n × p design matrix and $\varepsilon \sim N(0, \sigma^2 I)$. The OLS estimator $\hat{\beta} = (X^TX)^{-1}X^Ty$ has covariance matrix:
$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

The variance of the j-th coefficient is the j-th diagonal element: $Var(\hat{\beta}_j) = \sigma^2[(X^TX)^{-1}]_{jj}$. When near-linear dependence exists among predictors, the determinant $|X^TX| \to 0$, $(X^TX)^{-1}$ becomes large, and the variance of $\hat{\beta}_j$ inflates.

Step 2 — Auxiliary Regression
To quantify the multicollinearity specific to predictor X_j, an auxiliary OLS regression of X_j on all remaining predictors is estimated:
$X_j = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{j-1}X_{j-1} + \gamma_{j+1}X_{j+1} + \ldots + \gamma_p X_p + \nu$

Let R_j^2 denote the coefficient of determination from this auxiliary regression. A high $R_j^2 \to 1$ signals that X_j is nearly a linear combination of the other predictors — the defining condition of multicollinearity.

Step 3 — The VIF Formula
The Variance Inflation Factor for predictor X_j is defined as: $\mathbf{VIF_j = 1 / (1 - R_j^2), j = 1, 2, \ldots, p}$

Step 4—Link Between VIF and Coefficient Variance
It can be shown algebraically that VIF_j is exactly the factor by which the variance of $\hat{\beta}_j$ is inflated relative to the orthogonal case. For standardised predictors:
$Var(\hat{\beta}_j) = [\sigma^2 / (n-1)S_j^2] \times VIF_j = [\sigma^2 / (n-1)S_j^2] \times [1 / (1 - R_j^2)]$
where S_j^2 is the sample variance of X_j. The term $[\sigma^2 / (n-1)S_j^2]$ is the baseline variance under orthogonality; multiplying by $VIF_j = 1/(1-R_j^2)$ gives the actual inflated variance. Thus:

- $VIF_j = 1 \to R_j^2 = 0$: zero multicollinearity (X_j completely uncorrelated with all other predictors)
- $VIF_j = 5 \to R_j^2 = 0.80$: 80% of variance in X_j explained by other predictors; $SE(\hat{\beta}_j)$ is $\sqrt{5} \approx 2.24\times$ larger
- $VIF_j = 10 \to R_j^2 = 0.90$: 90% explained; $SE(\hat{\beta}_j)$ is $\sqrt{10} \approx 3.16\times$ larger — the conventional remediation threshold

Step 5 — The Tolerance Statistic
The reciprocal of VIF provides an equivalent diagnostic:
$TOL_j = 1 / VIF_j = 1 - R_j^2 \in (0, 1]$
$TOL_j \to 0$ indicates severe multicollinearity; $TOL_j = 1$ indicates complete orthogonality

Model Specification
The hedonic specification estimated in this study is a semi-logarithmic function in which the natural logarithm of sale price is regressed on thirteen predictors representing the structural, locational, spatial, and temporal dimensions of residential value. Writing the model in full for property i:
$\ln(SALE_PRC_i) = \beta_0 + \beta_1 \ln(TOT_LVG_AREA_i) + \beta_2 \ln(LND_SQFOOT_i) + \beta_3 \ln(SPEC_FEAT_VAL_i + 1) + \beta_4 SUBCNTR_DI_i + \beta_5 CNTR_DIST_i + \beta_6 OCEAN_DIST_i + \beta_7 WATER_DIST_i + \beta_8 HWY_DIST_i + \beta_9 RAIL_DIST_i + \beta_{10} structure_quality_i + \beta_{11} age_i + \beta_{12} avno60plus_i + \beta_{13} month_sold_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ and i = 1, …, n (1)

In compact matrix form, this is $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$ is the vector of log sale prices, $X \in \mathbb{R}^{n \times 14}$ is the design matrix whose first column is a vector of ones, $\beta \in \mathbb{R}^{14}$ is the coefficient vector, and ε is the disturbance vector. Equation (1) is the specification estimated under all three linear methods in subsequent parts of this study; the methods differ only in the penalty imposed on β, not in the set of predictors.

Three features of the specification require comment. First, a constant of one is added to SPEC_FEAT_VAL before logging because that variable takes the value zero for properties without special features; the transformation ln(x + 1) preserves these observations while retaining the concavity of the logarithmic form. The distributional consequences of this choice, and the resulting caution required in interpreting β_3, are set out in this study. Second, the coefficients β_1, β_2 and β_3 on log-transformed predictors are constant elasticities, interpretable as the percentage change in sale price associated with a one-percent change in the predictor. The remaining coefficients are semi-elasticities: for a continuous predictor, β_j is the approximate proportional change in sale price per one-unit increase in X_j, and for the binary predictor avno60plus the exact proportional effect is $\exp(\beta_{12}) - 1$. Third, the distance variables enter linearly rather than in logarithmic form because the theoretical bid-rent gradient is exponential in distance rather than in log distance, so that a linear distance term in a log-price equation reproduces the negative exponential rent surface predicted by the monocentric and polycentric urban models discussed previously.

Ordinary Least Squares (OLS)

Let $y \in \mathbb{R}^n$ denote the vector of observed log-prices and $X \in \mathbb{R}^{nxp}$ the design matrix with rows $x_i^T = (1, x_{i1}, \ldots, x_{ip})$. OLS minimises the residual sum of squares:

$\hat{\beta}_OLS = \text{argmin}_\beta\ L(\beta) = \text{argmin}_\beta\ \|y - X\beta\|^2 = \text{argmin}_\beta\ \Sigma_i(y_i - x_i^T\beta)^2$

Setting the gradient $\partial L/\partial\beta = -2X^T(y - X\beta) = 0$ yields the normal equations $X^TX\beta = X^Ty$, whose unique solution (assuming rank(X) = p) is: $\hat{\beta}_OLS = (X^TX)^{-1}X^Ty$

The Gauss–Markov theorem guarantees $\hat{\beta}$_OLS is the Best Linear Unbiased Estimator (BLUE). Standard errors are the square roots of the diagonal elements of $\hat{\sigma}^2(X^TX)^{-1}$, where $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2 / (n - p - 1)$. The t-statistic for each coefficient is $t_j = \hat{\beta}_j / SE(\hat{\beta}_j) \sim t(n-p-1)$ under H_0: $\beta_j = 0$.

Ridge Regression (L2 Regularisation)

Ridge augments the OLS loss with an L2 penalty:

$\hat{\beta}_ridge = \text{argmin}_\beta\ \{ \|y - X\beta\|^2 + \lambda\|\beta\|^2 \} = \text{argmin}_\beta\ \{ \Sigma_i(y_i - x_i^T\beta)^2 + \lambda\Sigma_j\beta_j^2 \}, \lambda \geq 0$

Setting the gradient $-2X^T(y - X\beta) + 2\lambda\beta = 0$ and solving yields the closed form: $\hat{\beta}_ridge = (X^TX + \lambda I)^{-1}X^Ty$

The addition of λI ensures $(X^TX + \lambda I)$ is positive definite and invertible for all $\lambda > 0$, even under multicollinearity. Via singular value decomposition $X = UDV^T$, the Ridge fitted values are $\hat{y}_ridge = \Sigma_j[d_j^2/(d_j^2+\lambda)]u_ju_j^Ty$, where $d_j^2/(d_j^2+\lambda) \in (0,1)$ are the shrinkage factors. As $\lambda \to 0$, $\hat{\beta}$_ridge $\to$ $\hat{\beta}$_OLS; as $\lambda \to \infty$, $\hat{\beta}$_ridge $\to 0$. Ridge introduces bias but reduces variance. The penalty λ was selected by 10-fold cross-validated cv.glmnet(alpha=0) in R (Zou & Hastie, 2005).

Lasso Regression (L1 Regularisation)

The Lasso substitutes an L1 penalty:

$\hat{\beta}_lasso = \text{argmin}_\beta\ \{ \|y - X\beta\|^2 + \lambda\|\beta\|_1 \} = \text{argmin}_\beta\ \{ \Sigma_i(y_i - x_i^T\beta)^2 + \lambda\Sigma_j|\beta_j| \}, \lambda \geq 0$

The L1 penalty is non-differentiable at $\beta_j = 0$, so no closed form exists. The solution is obtained via coordinate descent using the soft-thresholding update:

$\hat{\beta}_j = S(\rho_j, \lambda) = \text{sign}(\rho_j)(|\rho_j| - \lambda/2)_+$

where $\rho_j = X_j^T(y - X_{-j}\beta_{-j})$ is the partial residual inner product (the OLS estimate of β_j holding all other coefficients fixed), and $(a)_+ = \max(0, a)$. The soft-thresholding operator shrinks $|\rho_j|$ by $\lambda/2$ and sets the result to **exactly zero** whenever $|\rho_j| \leq \lambda/2$ — producing automatic variable selection.

In this study, Lasso retained all p = 13 predictors at its cross-validated optimal λ, confirming no predictor is a genuine noise variable. The penalty λ was selected using cv.glmnet(alpha=1) in R (Tibshirani, 1996)

Evaluation Protocol

The dataset was partitioned into training (80 %; n = 11,145) and held-out test (20 %; n = 2,787) sets using stratified random sampling with set.seed (123). All models were trained using 10-fold cross-validation (Park & Bae, 2015; Kuhn & Johnson, 2013). Cross-validation fold RMSE values were inspected for consistency to confirm training stability. Performance metrics are RMSE and R^2.

RESULTS AND DISCUSSION

Pearson Correlation Results

Table 4 presents Pearson correlation coefficients between each predictor and log(SALE_PRC), ranked by absolute magnitude. Figs 3–5 provide visual corroboration of three key predictor relationships.

Table 4: Pearson correlation coefficients — Predictors vs log(sale price)

Variable	*r* with log(SALE_PRC)	Interpretation
TOT_LVG_AREA (Living area)	0.713	Strong positive — dominant predictor
SPEC_FEAT_VAL (Special features)	0.506	Moderate positive — luxury premium
LND_SQFOOT (Land area)	0.370	Moderate positive — land value
HWY_DIST (Highway distance)	0.295	Positive — outer suburban premium effect
SUBCNTR_DI (Subcenter distance)	−0.404	Moderate inverse — subcenter accessibility premium
CNTR_DIST (CBD distance)	−0.233	Inverse — CBD accessibility premium
OCEAN_DIST (Ocean distance)	−0.209	Inverse — waterfront proximity premium
age (Property age)	−0.193	Inverse — depreciation effect
RAIL_DIST (Rail distance)	−0.073	Weak inverse — transit accessibility
WATER_DIST (Water body distance)	−0.030	Negligible — subsumed by OCEAN_DIST

Source: Author's computation using R cor() function, 2025. Negative values indicate that proximity (shorter distance) is associated with higher prices.

Total living area (r = 0.713) is the strongest linear predictor by a substantial margin, consistent with hedonic theory (Rosen, 1974; Hong et al., 2020) and with Nigerian valuation literature (Abidoye & Chan, 2017; Odebode et al., 2024). Fig. 3 presents the scatter plot of total living area against sale price, visually confirming the strong positive monotonic relationship through the clear positive regression slope and the colour gradient from deep purple at lower prices to bright yellow at higher prices.



Figure 3: Relationship between total living area and sale price (Source: R Output, Author's Computation, 2025)



Figure 4: Relationship between land area and sale price (Source: R Output, Authors' Computation, 2025)

Special features value (r = 0.506) ranks second and land area (r = 0.370) third. Fig. 4 presents the scatter plot of land area against sale price, showing a moderate positive relationship with considerably more scatter than the living area plot—consistent with land area being an incomplete proxy for property value in urban contexts where building improvements dominate value (Abidoye & Chan, 2018).

Among spatial distance variables, subcenter distance (r = −0.404) shows the strongest linear association—stronger than CBD distance (r = −0.233)—suggesting a polycentric urban structure. This extends Hoxha's (2025) finding of CBD dominance in monocentric Prishtina to a polycentric Nigerian context. Ocean distance (r = −0.209) confirms the waterfront premium; age (r = −0.193) reflects depreciation. Highway distance shows a counterintuitive positive correlation (r = 0.295)—Fig. 5 illustrates that higher-value properties concentrate at low-to-moderate distances, but the fitted curve rises initially before declining, reflecting buyer preference for residential tranquillity over direct highway adjacency at the upper price tier. Water body distance has negligible correlation (r = −0.030), subsumed by ocean distance.

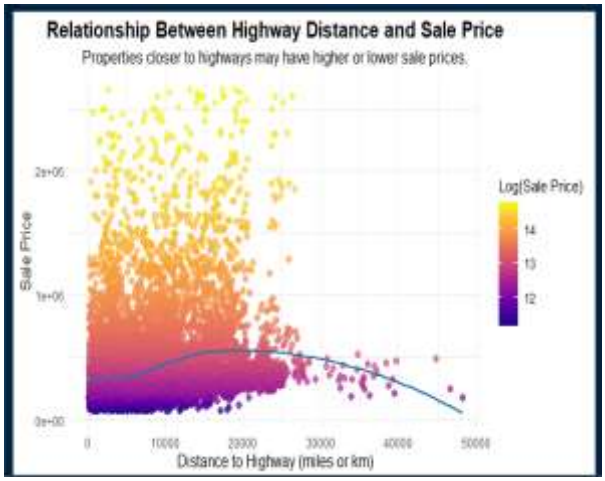


Figure 5: Relationship between highway distance and sale price (Source: R Output, Author's Computation, 2025)

Multicollinearity Results

Table 5 presents VIF estimates from the OLS model. All values are below 10, confirming the absence of severe multicollinearity. The highest values are CNTR_DIST (VIF = 6.94) and SUBCNTR_DI (VIF = 5.53)—predictable given their spatial correlation as both measure distances from urban centres that are geographically proximate. Structural predictors (living area, land area, special features, structural quality) all show very low VIF values between 1.18 and 1.75, confirming negligible multicollinearity among structural attributes. OLS coefficient estimates are therefore stable and reliable for interpretive use. The moderate spatial variable VIFs (5–7) provide theoretical justification for Ridge regression's inclusion as a stability-enhancing alternative.

Table 5: Variance inflation factor (VIF) analysis — OLS model

Variable	VIF	Interpretation
log(TOT_LVG_AREA)	1.75	Acceptable — low multicollinearity
log(SPEC_FEAT_VAL)	1.18	Acceptable — low multicollinearity
log(LND_SQFOOT)	1.61	Acceptable — low multicollinearity
SUBCNTR_DI	5.53	Moderate — monitor carefully (spatial correlation)
CNTR_DIST	6.94	Moderate — monitor carefully (spatial correlation)
OCEAN_DIST	2.76	Acceptable — low multicollinearity
structure_quality	1.48	Acceptable — low multicollinearity
Age	1.84	Acceptable — low multicollinearity
HWY_DIST	1.37	Acceptable — low multicollinearity
RAIL_DIST	1.38	Acceptable — low multicollinearity
WATER_DIST	3.72	Acceptable — low multicollinearity
avno60plus	1.04	Acceptable — low multicollinearity
month_sold	1.00	Acceptable — negligible multicollinearity

Source: Authors' computation using R car::vif(), 2025. VIF < 5: acceptable; 5–10: moderate; > 10: severe multicollinearity.

Table 6: OLS regression coefficient estimates

Variable	Coefficient (β)	Std. Error	t-value	p-value
(Intercept)	5.3810	0.0683	78.84	< 0.001 ***
log(TOT_LVG_AREA)	0.8216	0.0092	89.67	< 0.001 ***
log(LND_SQFOOT)	0.1292	0.0066	19.52	< 0.001 ***
log(SPEC_FEAT_VAL + 1)	0.0079	0.0008	9.82	< 0.001 ***
SUBCNTR_DI (per 1,000 ft)	−0.0056	0.0003	−20.83	< 0.001 ***
CNTR_DIST (per 1,000 ft)	−0.0029	0.0002	−14.09	< 0.001 ***
OCEAN_DIST (per 1,000 ft)	−0.0045	0.0002	−18.82	< 0.001 ***
WATER_DIST (per 1,000 ft)	−0.0014	0.0004	−3.42	< 0.001 ***
HWY_DIST (per 1,000 ft)	0.0095	0.0005	19.54	< 0.001 ***
RAIL_DIST (per 1,000 ft)	0.0079	0.0005	16.22	< 0.001 ***
structure_quality	0.1413	0.0028	50.12	< 0.001 ***
Age	−0.0052	0.0002	−31.91	< 0.001 ***
avno60plus	−0.0541	0.0216	−2.51	0.012 *
month_sold	0.0036	0.0008	4.65	< 0.001 ***

OLS Coefficient Results

Table 6 presents the OLS regression coefficient estimates. A particularly notable finding is that all 13 predictors are statistically significant: twelve at the $p < 0.001$ level (***) and avno60plus at $p < 0.05$ (*). This universal significance distinguishes the Nigerian market from sparser international contexts where several variables typically prove non-significant, confirming that the full multi-dimensional predictor set—structural, locational, and spatial—contributes meaningfully to residential price formation.

Authors' computation using *R* lm(), 2025. Significance codes: *** $p < 0.001$; * $p < 0.05$. Distance coefficients are reported per 1,000 feet to permit direct comparison of magnitudes across predictors; the underlying variables are measured in feet. Coefficients on log-transformed predictors are elasticities; all others are semi-elasticities. n = 11,145; R^2 = 0.778; Adjusted R^2 = 0.778; F(13, 11,131) = 3,005; $p < 0.001$.

The coefficient on log(TOT_LVG_AREA) ($\beta = 0.8216$, $t = 89.67$, $p < 0.001$) is the largest in magnitude and statistical strength, confirming that a 1 % increase in living area is associated with approximately 0.82 % increase in sale price—consistent with the international hedonic literature (Hong *et al.*, 2020; Rosen, 1974). Structure_quality ($\beta = 0.1413$, $t = 50.12$, $p < 0.001$) shows a strong premium, amplified by the bimodal quality distribution. Age shows the expected depreciation effect ($\beta = -0.0052$, $t = -31.91$, $p < 0.001$). Among spatial predictors, SUBCNTR_DI ($\beta = -0.0056$ per 1,000 ft, $t = -20.83$) and OCEAN_DIST ($\beta = -0.0045$ per 1,000 ft, $t = -18.82$) are the strongest, with CNTR_DIST ($\beta = -0.0029$ per 1,000 ft, $t = -14.09$) also highly significant. The positive and significant coefficients on HWY_DIST ($\beta = 0.0095$ per 1,000 ft, $t = 19.54$) and RAIL_DIST ($\beta = 0.0079$ per 1,000 ft, $t = 16.22$) confirm the counterintuitive suburban tranquillity premium identified in the correlation analysis. Even WATER_DIST, avno60plus, and month_sold are significant, confirming that the full predictor set is necessary for reliable residential valuation — a finding with immediate practical implications for property valuers who may be tempted to use simplified models (Abidoye & Chan, 2018).

Model Performance Results

Table 7 presents performance results for all linear model specifications. Cross-validation fold RMSE values were consistent across all 10 folds for each model, confirming training stability.

Table 7: Linear model performance — Normalised scale (10-Fold CV, held-out test set)

Model	RMSE	R^2	% Error	Rank
Linear Regression (OLS)	0.5097	0.778	66.5	1st
Lasso Regression	0.5098	0.778	66.6	2nd
Ridge Regression	0.5112	0.774	66.7	3rd

Source: Authors' computation, *R* (version 4.3.1), 2025.
% Error ≈ (exp(RMSE) − 1) × 100

As shown in Table 7, OLS achieves $R^2 = 0.778$ (RMSE = 0.5097), explaining 77.8 % of log-price variance and establishing the first documented linear model benchmark for the Nigerian urban residential market. Lasso and Ridge achieve comparable performance ($R^2 = 0.778$ and $R^2 = 0.774$ respectively), with negligible differences from OLS. This near-equivalence is explained by the VIF results: since multicollinearity is absent or only moderate, regularisation provides no material benefit over OLS for this dataset. Lasso retains all 13 predictors at its optimal λ rather than zeroing any out, consistent with the OLS finding that all variables are statistically significant—confirming there are no true noise predictors in this dataset. The OLS $R^2 = 0.778$ therefore establishes a credible linear benchmark for this market—one that future ensemble and nonlinear specifications must surpass to demonstrate methodological value. Because the objective of this study is to document linear benchmarks rather than to conduct a full algorithmic comparison, nonlinear learners are deliberately excluded from the present evaluation and reserved for the extension described in the last part of this work.

Discussion of Results

The Pearson correlation and OLS coefficient results collectively confirm that residential property price formation in the Nigerian urban market is driven by a comprehensive set of structural, locational, and spatial determinants. Total living area ($r = 0.713$; $\beta = 0.8216$, $t = 89.67$) emerges as the overwhelmingly dominant predictor, consistent with hedonic theory that habitable space is the primary source of residential utility (Rosen, 1974; Hong *et al*., 2020) and with Nigerian valuation literature documenting structural size as the foremost price driver (Abidoye and Chan, 2017; Odebode *et al*., 2024). The OLS elasticity of 0.82 implies that a 10 % increase in living area is associated with approximately 8.2 % increase in sale price—a practically significant relationship with direct implications for property development decisions.

The significance of subcenter distance (β = −0.0056 per 1,000 ft, t = −20.83) over CBD distance (β = −0.0029 per 1,000 ft, t = −14.09) reveals a polycentric urban structure in which secondary employment and commercial centres generate accessibility gradients at least as influential as the central business district. This extends Hoxha's (2025) finding of CBD dominance in monocentric Prishtina to a polycentric setting and has immediate practical implications for urban planning — investment in subcenter infrastructure and accessibility directly translates into residential property value uplift across the urban area.

The statistically significant and positive coefficients on HWY_DIST (β = 0.0095 per 1,000 ft, t = 19.54) and RAIL_DIST (β = 0.0079 per 1,000 ft, t = 16.22) confirm the counterintuitive finding that higher-value properties tend to locate away from highway and rail corridors. This reflects a residential tranquillity premium at upper price tiers — buyers in this market segment are willing to sacrifice direct transit access for quieter residential environments. This finding is practically important for mortgage lenders: properties located near major transport corridors should not be assumed to command premiums in this market without controlling for property tier.

The near-equivalent performance of OLS ($R^2 = 0.778$), Lasso ($R^2 = 0.778$), and Ridge ($R^2 = 0.774$) is a theoretically meaningful result. It confirms that when all relevant predictors are included, and multicollinearity is absent or moderate—as confirmed by the VIF analysis—regularisation provides no material improvement over standard OLS. Lasso retaining all 13 predictors at its optimal λ rather than zeroing any out is particularly telling: it corroborates the OLS significance finding that there are no genuine noise predictors in this dataset. Every structural, locational, and spatial variable contributes meaningfully to Nigerian residential valuation. This has an important practical implication—simplified valuation models that omit any dimension of the predictor set will produce systematically biased estimates.

The OLS $R^2 = 0.778$ benchmark established here represents a stronger linear baseline than comparable international studies. Park and Bae (2015) reported substantially lower OLS performance for Fairfax County, Virginia, likely because the present study includes a richer multi-dimensional spatial feature set—particularly the subcenter distance, ocean distance, and rail distance variables that are rarely available in Nigerian housing datasets. The 22.2 % unexplained variance represents the irreducible limitation of linear models in capturing the threshold effects, spatial interactions, and nonlinear price responses that characterise urban residential markets (Abidoye & Chan, 2018; Truong *et al*., 2020). Future research should extend this pipeline with ensemble and kernel-based machine learning approaches to quantify the performance gain over this established linear benchmark.

CONCLUSION

This study conducted a systematic Pearson correlation analysis, VIF-based multicollinearity assessment, OLS coefficient analysis, and evaluation of three linear regression specifications—OLS, Ridge, and Lasso—applied to 13,932 residential transactions from the Nigerian urban housing market within a reproducible CRISP-DM pipeline. Total living area ($r = 0.713$) emerged as the dominant linear predictor, followed by special features value ($r = 0.506$), land area ($r = 0.370$),

and subcenter distance ($r = -0.404$). OLS achieved $R^2 = 0.778$ (RMSE = 0.5097), establishing the first documented linear model benchmark for this market. All 13 predictors were statistically significant under OLS, and VIF analysis confirmed the absence of severe multicollinearity.

Three principal contributions are made to the Nigerian housing market literature. First, this study provides the first systematic Pearson correlation and OLS coefficient analysis of residential price determinants using a large transactional dataset with explicit multi-dimensional spatial distance variables, building on the Nigerian HPM literature established by Abidoye and Chan (2017, 2018) and Odebode *et al*. (2024). Second, it establishes $R^2 = 0.778$ as the first documented OLS benchmark for Nigerian urban residential valuation. Third, it demonstrates that all 13 structural, locational, and spatial predictors are statistically significant in this market—a finding that distinguishes Nigerian urban property markets from sparser contexts and confirms that simplified valuation models omitting any dimension of the predictor set will produce systematically biased estimates.

For practitioners, the full thirteen-variable OLS specification identified in this study is the minimum feature set for a reliable linear hedonic model in this market. Valuers and automated valuation model designers should include all structural, locational, and spatial distance variables—particularly subcenter distance, ocean distance, and CBD distance—to avoid systematic mispricing. For mortgage lenders, the $R^2 = 0.778$ benchmark means approximately 22.2 % of log-price variance remains unexplained by the linear model, representing irreducible uncertainty that should inform loan-to-value assessment buffers.

Limitations and Future Research Directions

Three limitations open productive research directions. First, the cross-sectional design prevents analysis of price dynamics and temporal market cycles. Future work should incorporate temporal dimensions through panel data or time-series-aware approaches (Zhang *et al*., 2023). Second, the spatial distance variables are Euclidean proxies for multidimensional locational quality; integrating socioeconomic indicators, school quality data, and flood zone classification would enrich the feature space (Bourassa *et al*., 2007). Third, while OLS $R^2 = 0.778$ establishes a strong linear baseline, the 22.2 % unexplained variance and the known nonlinear structure of housing markets motivate ensemble and kernel-based machine learning approaches --- Random Forest, Gradient Boosting, and Support Vector Regression --- capable of capturing threshold effects and spatial interactions that linear models cannot represent. Ho *et al*. (2021) demonstrated this advantage directly, finding that random forest and gradient boosting outperformed support vector machines in predicting Hong Kong property prices, while Tekouabou *et al*. (2023) confirm across a systematic review of the wider literature that ensemble methods now dominate the urban real estate prediction space. Extending this pipeline with such methods, as done by Abidoye and Chan (2018) for ANN in Lagos, represents the direct next research step and is expected to yield substantially higher predictive accuracy.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

Abidoye, R. B. and Chan, A. P. C. (2017). Critical review of hedonic pricing model application in property price appraisal: A case of Nigeria. *International Journal of Sustainable Built Environment*, 6(1), 250-259. https://doi.org/10.1016/j.ijsbe.2017.02.007

Abidoye, R. B. and Chan, A. P. C. (2018). Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal*, 24(1), 71-83. https://doi.org/10.1080/14445921.2018.1436306

Alonso, W. (1964), *Location and Land Use: Toward a General Theory of Land Rent*. Harvard University Press, Cambridge, MA.

Anas, A., Arnott, R. and Small, K. A. (1998). Urban spatial structure. *Journal of Economic Literature*, 36(3), 1426-1464. https://www.jstor.org/stable/2564805

Bartik, T. J. (1987). The estimation of demand parameters in hedonic price models. *Journal of Political Economy*, 95(1), 81-88. https://doi.org/10.1086/261442

Bourassa, S. C., Cantoni, E. and Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, 35(2), 143-160. DOI: 10.1007/s11146-007-9036-8

Brown, J. N. and Rosen, H. S. (1982). On the estimation of structural hedonic price models. *Econometrica*, 50(3), 765-768. https://doi.org/10.2307/1912614

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS Inc., Chicago, IL.

Cropper, M. L., Deck, L. B. and McConnell, K. E. (1988). On the choice of functional form for hedonic price functions. *The Review of Economics and Statistics*, 70(4), 668-675. DOI: 10.2307/1935831

Ekeland, I., Heckman, J. J. and Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, 112(S1), S60-S109. DOI: 10.3386/w9910

Epple, D. (1987). Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products. *Journal of Political Economy*, 95(1), 59-80. DOI: https://doi.org/10.1086/261441

Halvorsen, R. and Pollakowski, H. O. (1981). Choice of functional form for hedonic price equations.

Journal of Urban Economics, 10(1), 37-49. DOI: 10.1016/0094-1190(81)90021-8

Ho, W. K. O., Tang, B. S. and Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48-70. DOI: 10.1080/09599916.2020.1832558

Hong, J., Choi, H. and Kim, W. S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, 24(3), 140-152. DOI: 10.3846/ijspm.2020.11544

Hoxha, V. (2025). Comparative analysis of machine learning models in predicting housing prices: A case study of Prishtina's real estate market. *International Journal of Housing Markets and Analysis*, 18(3), 694-711.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Kuminoff, N. V., Smith, V. K. and Timmins, C. (2013). The new economics of equilibrium sorting and policy evaluation using housing markets. *Journal of Economic Literature*, 51(4), 1007-1062. DOI: 10.1257/jel.51.4.1007

McMillen, D. P. and Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics*, 53(3), 321-338. DOI: 10.1016/S0094-1190(03)00026-3

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, 57(2), 197-210. https://www.jstor.org/stable/1821621

Muth, R. F. (1969). *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*. University of Chicago Press, Chicago, IL. urn:oclc:record:1148205073

Odebode, A. A., Oladimeji, D. E. and Ogunbayo, O. T. (2024). An examination of the influence of housing attributes on residential property rental value in south-western Nigeria using the hedonic pricing model. *International Journal of Real Estate Studies*, 18(2), 120-129. DOI: 10.11113/intrest.v18n2.385

Palmquist, R. B. (2005). Property value models. In Mäler, K.-G. and Vincent, J. R. (Eds), *Handbook of Environmental Economics* (2, pp. 763-819). Elsevier, Amsterdam. 10.1016/S1574-0099(05)02016-4

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934. DOI: 10.1016/j.eswa.2014.11.040

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55. https://doi.org/10.1086/260169

Sirmans, G. S., Macpherson, D. A. and Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1), 3-43. https://www.jstor.org/stable/44103506

Tekouabou, S. C. K., Gherghina, S. C., Kameni, E. D., Filali, Y. and Idrissi Gartoumi, K. (2023), "AI-based on machine learning methods for urban real estate prediction: A systematic survey. *Archives of Computational Methods in Engineering*, 31, 1079-1095.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Truong, Q., Nguyen, M., Dang, H. and Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433-442. DOI: 10.1016/j.procs.2020.06.111

Vinutha, H. P., Poornima, B. and Sagar, B. M. (2018). Detection of outliers using interquartile range technique from intrusion dataset. In Satapathy, S. C., Tavares, J. M. R. S., Bhateja, V. and Mohanty, J. R. (Eds), *Information and Decision Sciences, Advances in Intelligent Systems and Computing* (701, pp. 511-518). Springer, Singapore. DOI: 10.1007/978-981-10-7563-6_53

Zhang, Y., Rahman, A. and Miller, E. (2023). Longitudinal modelling of housing prices with machine learning and temporal regression. *International Journal of Housing Markets and Analysis*, 16(4), 693-715. DOI: 10.1108/IJHMA-02-2022-0033

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x