A Comprehensive Statistical Analysis of Residential Property Valuation Using Multiple Linear Regression Techniques

V. Munaiah[1], P. Maheswari[2], T. Gangaram[3], K. Sreenivasulu[4], K. Murali[5**]

1. Associate Professor of Statistics, S.V. A Govt. College(M), Srikalahasti.
2. Assistant Professor, Department of Statistics, Govt. Degree College(A), Nagari, Chittoor.
3. Associate Professor of Statistics, S.V.A Govt. College (M), Srikalahasti.
4. Lecturer in Statistics, Sri Padmavathi Women's Degree and PG College, TTD(A), Tirupati.
5. Department of Statistics, S V University, Tirupati.

[**]Corresponding Author: **dr.murali4stat@gmail.com**

Abstract

This research paper presents a comprehensive statistical analysis of residential property valuation in King County, Washington, employing Multiple Linear Regression (MLR) as the primary analytical framework. Utilizing a dataset comprising 21,613 residential property transactions recorded between May 2014 and May 2015, this study examines how structural, locational, and qualitative property attributes jointly determine sale prices. The dataset encompasses 21 variables including living area square footage, number of bedrooms and bathrooms, property grade, condition, waterfront access, view quality, and proximity-based neighborhood features.

Two regression models were developed and evaluated: a standard MLR model and a log-transformed variant for price normalization. The standard MLR model achieved an R^2 of 0.6526 and an adjusted R^2 of 0.6514, with a Root Mean Square Error (RMSE) of \$229,186, while the log-transformed model produced an R^2 of 0.6554 and adjusted R^2 of 0.6543. Key findings reveal that property grade ($r = 0.667$), living area ($r = 0.702$), and waterfront status ($\beta = \$554,704$) are the strongest determinants of residential value. Multicollinearity diagnostics using Variance Inflation Factors (VIF), Durbin-Watson statistics ($DW = 2.008$), and residual normality tests were conducted to validate model integrity. The study concludes that

while MLR provides robust interpretable predictions, log-transformation addresses price skewness (skew = 4.024) and improves distributional assumptions.

Keywords: Multiple Linear Regression, Residential Property Valuation, Hedonic Pricing, King County, Housing Market, VIF, Multicollinearity, Log-transformation, Predictive Modeling

1. Introduction

1.1 Background and Motivation

Accurate residential property valuation is fundamental to the efficient functioning of real estate markets, mortgage lending institutions, tax assessment authorities, and investment decision-making processes. The determination of fair market value for residential properties involves the complex interplay of hundreds of physical, locational, and socioeconomic factors, making it one of the most analytically challenging tasks in applied economics.

Traditional appraisal methodologies — including the sales comparison approach, the income capitalization approach, and the cost approach — rely heavily on expert judgment and small comparable samples. While these methods remain standard practice, they are susceptible to subjective bias, limited scalability, and inconsistency across appraisers. The increasing availability of large-scale property transaction datasets has created an unprecedented opportunity to apply statistical and machine learning methods to property valuation at scale.

Multiple Linear Regression (MLR) has long served as the workhorse of hedonic property pricing analysis. Originating from Rosen's (1974) seminal hedonic price theory, MLR decomposes the market price of a heterogeneous good into the implicit prices of its constituent attributes. In the residential housing context, each structural feature, location characteristic, and quality indicator contributes a marginal price premium or discount that, in aggregate, yields the predicted transaction price.

1.2 Research Objectives

This study is guided by the following primary objectives:

1. To identify and quantify the statistically significant determinants of residential property values in King County, Washington.
2. To develop and validate standard and log-transformed Multiple Linear Regression models for property price prediction.
3. To assess model performance using R^2, Adjusted R^2, RMSE, and MAE metrics.
4. To diagnose model assumptions including multicollinearity (VIF), heteroscedasticity, normality of residuals, and autocorrelation.
5. To compare the predictive efficacy of standard versus log-transformed regression specifications.

1.3 Significance of the Study

King County, encompassing the Seattle metropolitan area, represents one of the most dynamic and economically significant residential property markets in the United States. The housing market in this region has experienced substantial price appreciation driven by technology sector growth, population migration, and constrained housing supply. Consequently, developing accurate, data-driven valuation models for this market has both academic significance and substantial practical implications for real estate professionals, policymakers, and financial institutions.

2. Literature Review

2.1 Hedonic Pricing Theory

The theoretical foundation of this research rests upon Lancaster's (1966) consumer demand theory and Rosen's (1974) hedonic pricing model. Rosen's model posits that differentiated products, such as residential properties, can be characterized as bundles of objectively measurable attributes and that equilibrium market prices implicitly reveal the marginal willingness to pay for each attribute. In the residential property context, the hedonic price function $P = f(S, L, N)$ models sale price (P) as a function of structural characteristics (S), locational attributes (L), and neighborhood quality variables (N).

2.2 Application of MLR in Property Valuation

Multiple Linear Regression has been extensively applied in hedonic property valuation studies since the late 1960s. Early contributions by Kain and Quigley (1970) demonstrated the feasibility of using regression techniques to explain approximately 70–80% of price variance using structural and neighborhood variables. Subsequent studies refined variable selection, transformations, and spatial controls.

Oladunni and Sharma (2018) achieved an adjusted R^2 of 0.928 in predicting King County housing prices using 80 features with the Ames dataset framework, demonstrating the high explanatory power achievable with comprehensive variable sets. Bourassa et al. (2003) found that spatial autoregressive models outperform standard OLS in capturing neighborhood spillover effects, while Malpezzi (2003) provided a comprehensive review of hedonic price indices and their applications in housing market analysis.

2.3 Key Determinants of Residential Property Value

The literature consistently identifies several primary categories of value determinants. Structural characteristics — particularly living area square footage, bathroom count, and construction quality grade — emerge as the most robust predictors across geographic markets. Sirmans et al. (2005) conducted a metaanalysis of 125 hedonic studies and found square footage to carry the highest average elasticity (0.44), followed by bathroom count (0.09) and property age (-0.02).

Location variables, including proximity to waterfronts, parks, schools, and employment centers, command significant price premiums. Waterfront properties in particular have been documented to carry premiums of 20–150% depending on water body type and accessibility. Condition and quality grade ratings, while subjective, have been shown to account for substantial price variation when standardized across assessment systems.

2.4 Methodological Considerations

A critical methodological concern in hedonic regression is multicollinearity among correlated predictors — particularly between measures of property size such as living area, above-ground square footage, and basement area. Variance Inflation Factor (VIF) diagnostics are standard practice, with thresholds of VIF > 10 indicating problematic collinearity. Price distribution skewness is another persistent challenge; sale prices typically exhibit strong positive skewness and leptokurtosis, violating the normality assumption of OLS residuals. The log-linear specification — regressing log(price) on predictors — is the most commonly employed remedy.

3. Data Description

3.1 Dataset Overview

The analysis employs the King County House Sales dataset, a publicly available repository of residential property transactions in King County, Washington, USA, spanning the period from May 2014 to May 2015. The dataset contains 21,613 individual property sale records across 21 variables, encompassing both the dependent variable (sale price) and 20 potential predictor variables covering structural, physical, locational, and temporal property characteristics.

Variable	Type	Description	Range / Categories
price	Continuous	Sale price (dependent variable)	$75,000 – $7,700,000
sqft_living	Continuous	Interior living space (sq ft)	290 – 13,540
sqft_lot	Continuous	Lot/land area (sq ft)	520 – 1,651,359
bedrooms	Discrete	Number of bedrooms	0 – 33
bathrooms	Continuous	Number of bathrooms (incl. half)	0 – 8
floors	Continuous	Number of floors	1.0 – 3.5
waterfront	Binary	Waterfront property indicator	0 = No, 1 = Yes

Variable	Type	Description	Range / Categories
view	Ordinal	Quality of property view	0 – 4 scale
condition	Ordinal	Overall physical condition	1 – 5 scale
grade	Ordinal	Construction & design quality	1 – 13 scale
sqft_above	Continuous	Above-ground square footage	290 – 9,410
sqft_basement	Continuous	Basement area (sq ft)	0 – 4,820
yr_built	Discrete	Year property was built	1900 – 2015
yr_renovated	Discrete	Year of last renovation (0 if none)	0, 1934 – 2015
zipcode	Categorical	ZIP code of the property	70 unique codes
lat / long	Continuous	Geographic coordinates	Lat: 47.16–47.78
sqft_living15	Continuous	Avg living area of 15 neighbors	399 – 6,210
sqft_lot15	Continuous	Avg lot area of 15 neighbors	651 – 871,200

Table 1: Variable Descriptions for King County House Sales Dataset (N = 21,613)

3.2 Descriptive Statistics

Table 2 presents the descriptive statistics for key continuous variables in the dataset. The mean sale price of $540,088 with a standard deviation of $367,127 indicates substantial price dispersion. The median price of $450,000 is considerably below the mean, confirming the strong positive skewness of the price distribution (skewness = 4.024, kurtosis = 34.586). This departure from normality motivates the log-transformation explored in Section 5.

Variable	Mean	Median	Std Dev	Min	Max
Price ($)	540,088	450,000	367,127	75,000	7,700,000
Sqft Living (ft²)	2,079.9	1,910	918.4	290	13,540
Bedrooms	3.37	3	0.93	0	33
Bathrooms	2.11	2.25	0.77	0	8
Grade	7.66	7	1.18	1	13
Property Age (yrs)	43.99	40	29.37	0	115
Sqft Above (ft²)	1,788.4	1,560	828.1	290	9,410
Sqft Basement (ft²)	291.5	0	442.6	0	4,820

Table 2: Descriptive Statistics for Key Variables (N = 21,613)

3.3 Feature Engineering

Two derived features were constructed prior to regression modeling:

- Property Age: Computed as 2015 − yr_built, yielding values ranging from 0 to 115 years (mean = 43.99 years, SD = 29.37 years).
- Renovation Indicator: A binary variable coded 1 if yr_renovated > 0 (i.e., property has been renovated at some point) and 0 otherwise.

4. Methodology

4.1 Theoretical Framework: Multiple Linear Regression

Multiple Linear Regression (MLR) is a supervised statistical technique that models the linear relationship between a continuous dependent variable and two or more independent predictor variables. The general MLR equation is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \varepsilon$$

where **Y** is the dependent variable (sale price), $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{\beta_1 ... \beta_n}$ are the regression coefficients for each predictor, $\mathbf{X_1 ... X_n}$ are the independent predictor variables, and $\boldsymbol{\varepsilon}$ is the random error term assumed to be normally distributed with mean zero and constant variance ($\varepsilon \sim N(0, \sigma^2)$).

For the applied model, the specific formulation is:

$$\text{Price} = \beta_0 + \beta_1(\text{sqft\_living}) + \beta_2(\text{grade}) + \beta_3(\text{bathrooms}) + \beta_4(\text{waterfront}) + \beta_5(\text{view})$$
$$+ \beta_6(\text{condition}) + \beta_7(\text{sqft\_basement}) + \beta_8(\text{floors}) + \beta_9(\text{age}) + \beta_{10}(\text{renovated}) + \varepsilon$$

4.2 Log-Transformed Model Specification

Given the significant positive skewness of the price distribution (skewness = 4.024), a log-linear model specification was also estimated. The log-transformation stabilizes variance, reduces the influence of extreme outliers, and often produces residuals more consistent with the normality assumption:

$$\ln(\text{Price}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

In this specification, regression coefficients represent percentage changes in price associated with one-unit increases in each predictor, offering intuitive proportional interpretations.

4.3 Variable Selection

The final set of 14 predictor variables was selected based on: (1) theoretical justification from the hedonic pricing literature, (2) correlation screening ($|r| > 0.05$ with price), and (3) practical availability and data quality. The selected features span five categories: size-related metrics (sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15), quality indicators (grade, condition), locational factors (waterfront, view), structural attributes (bedrooms, bathrooms, floors), and temporal variables (age, renovated).

4.4 Model Assumptions and Diagnostic Tests

The validity of MLR rests upon the Gauss-Markov assumptions. The following diagnostic tests were applied:

Assumption	Diagnostic Test	Threshold / Criterion
No Multicollinearity	Variance Inflation Factor (VIF)	VIF > 10 indicates problematic collinearity
Homoscedasticity	Breusch-Pagan Test, Residual Plots	$p < 0.05$ indicates heteroscedasticity
Normality of Residuals	Shapiro-Wilk Test, Q-Q Plot	$p > 0.05$ indicates normality
No Autocorrelation	Durbin-Watson Statistic	Values near 2.0 indicate no autocorrelation
Linearity	Partial Regression Plots	Visual inspection of linear trends
Outliers	Cook's Distance	Cook's D > 4/n indicates influential point

Table 3: Model Assumption Diagnostic Tests

4.5 Model Evaluation Metrics

Model performance was assessed using four standard metrics applied to an 80/20 train-test split (17,290 training observations, 4,323 test observations):

Metric	Formula	Interpretation
R^2 (Coefficient of Determination)	$1 - (SS\_res / SS\_tot)$	Proportion of variance explained by the model
Adjusted R^2	$1 - [(1-R^2)(n-1)/(n-k-1)]$	R^2 penalized for number of predictors (k)
RMSE (Root Mean Square Error)	$\sqrt{[\Sigma(y_i - \hat{y}_i)^2 / n]}$	Average prediction error in original units
MAE (Mean Absolute Error)	$\Sigma\|y_i - \hat{y}_i\| / n$	Average absolute deviation of predictions

Table 4: Model Evaluation Metrics

5. Empirical Results and Analysis

5.1 Correlation Analysis

Prior to regression modeling, Pearson correlation coefficients were computed between all predictor variables and sale price. Table 5 presents the correlation results, revealing the relative strength of linear associations with sale price. Living area (sqft_living) exhibits the strongest positive correlation (r = 0.702), followed closely by construction grade (r = 0.667) and above-ground square footage (r = 0.606). These findings are consistent with the hedonic pricing literature's emphasis on size and quality as primary value drivers.

Notably, property condition exhibits a surprisingly low correlation with price (r = 0.036), possibly reflecting the confounding influence of neighborhood and location factors on assessed condition ratings. Property age shows a slightly negative correlation (r = −0.054), consistent with depreciation theory, though the relationship is weak given the presence of renovated properties in the sample.

Variable	Pearson r	Variable	Pearson r
sqft_living	0.702 ***	waterfront	0.266 ***
grade	0.667 ***	floors	0.257 ***
sqft_above	0.606 ***	renovated	0.126 ***
sqft_living15	0.585 ***	sqft_lot	0.090 ***
bathrooms	0.525 ***	condition	0.036 ***
view	0.397 ***	age	-0.054 ***
sqft_basement	0.324 ***	bedrooms	0.308 ***

*Table 5: Pearson Correlations with Sale Price (*** $p < 0.001$)*

5.2 Regression Model Results

Table 6 presents the full regression coefficient estimates, standard interpretations, and statistical significance for Model 1 (Standard MLR). All reported coefficients are based on the training dataset, while performance metrics reflect out-of-sample test set evaluation.

Variable	Coefficient (β)	Interpretation	Correlation r
Intercept (β_0)	—	Baseline price when all predictors = 0	—
waterfront	+$554,704	Waterfront property premium over non-waterfront	0.266
grade	+$118,596	Price increase per quality grade increment	0.667

Variable	Coefficient (β)	Interpretation	Correlation r
bathrooms	+$49,519	Additional value per bathroom unit	0.525
view	+$43,705	Price premium per view quality level (0–4)	0.397
bedrooms	−$38,163	Negative effect (size-adjusted; more rooms = smaller rooms)	0.308
floors	+$29,139	Price increase per additional floor	0.257
renovated	+$20,400	Premium for having undergone renovation	0.126
condition	+$16,847	Price per condition rating increment (1–5)	0.036
age	+$3,644	Per-year age effect (older homes may have larger lots)	−0.054
sqft_living	+$107.68	Price per additional square foot of living space	0.702
sqft_basement	+$57.71	Price per additional sq ft of basement	0.324
sqft_above	+$49.97	Price per additional sq ft above-ground	0.606
sqft_living15	+$24.20	Value per sq ft of avg neighborhood living area	0.585
sqft_lot	−$0.28	Marginal lot size effect (near zero)	0.090

Table 6: MLR Coefficient Estimates — Model 1 (Standard MLR)

The waterfront premium of $554,704 is the largest single coefficient, reflecting the exceptional scarcity value of waterfront access in the King County market. Each additional grade point on the construction quality scale (1–13) is associated with a $118,596 price increment, consistent with the strong correlation between grade and luxury amenities, material quality, and architectural design. The negative coefficient on bedrooms (−$38,163) is an artifact of size-control: holding living area constant, additional bedrooms imply smaller individual room sizes, which buyers may discount.

5.3 Model Performance Comparison

Metric	Model 1: Standard MLR	Model 2: Log-Transformed MLR
R^2 (Test Set)	0.6526	0.6554
Adjusted R^2	0.6514	0.6543
RMSE	$229,186	$273,541 (back-transformed)
MAE	$143,500	$139,114 (back-transformed)
Price Specification	Level-Level	Log-Level
Interpretation of β	Unit dollar change	Approximate % change

Table 7: Model Performance Comparison (80/20 Train-Test Split, n_test = 4,323)

Both models explain approximately 65.3–65.5% of sale price variance, which is consistent with benchmark MLR studies on residential data that typically achieve R^2 values between 0.60 and 0.75 without spatial fixed effects. The log-transformed model achieves marginally higher R^2 (0.6554 vs. 0.6526) and lower MAE ($139,114 vs. $143,500), suggesting slightly improved prediction of typical-price properties. However, its higher back-transformed RMSE ($273,541 vs. $229,186) indicates less accuracy for high-value outliers, which are penalized more by the log retransformation.

6. Model Diagnostic Results

6.1 Multicollinearity Analysis (VIF)

Variance Inflation Factor (VIF) values were computed for all 14 predictor variables to assess multicollinearity. Table 8 reports the results.

Variable	VIF Value	Interpretation
sqft_living	∞ (collinear)	Perfect multicollinearity with sqft_above + sqft_basement
sqft_above	∞ (collinear)	sqft_living = sqft_above + sqft_basement (definitional)
sqft_basement	∞ (collinear)	Part of definitional relationship with sqft_living
bathrooms	3.34	Moderate — acceptable range
grade	3.23	Moderate — acceptable range
sqft_living15	2.80	Moderate — acceptable range
floors	1.93	Low — no concern
age	2.01	Low — no concern
bedrooms	1.64	Low — no concern
waterfront	1.20	Very low
view	1.40	Very low
condition	1.22	Very low
renovated	1.14	Very low
sqft_lot	1.06	Very low

Table 8: Variance Inflation Factor (VIF) Results

The infinite VIF values for sqft_living, sqft_above, and sqft_basement arise from a definitional identity: sqft_living = sqft_above + sqft_basement. This is a known structural multicollinearity issue in housing datasets. In practice, this is resolved by either dropping sqft_basement or sqft_above from the model specification. For the purposes of interpretability and completeness in this analysis, all three are retained with the caveat noted. All remaining predictors exhibit VIF values well below the critical threshold of 10, confirming that problematic multicollinearity is confined to the size-composition variables.

6.2 Autocorrelation (Durbin-Watson)

The Durbin-Watson statistic was computed on model residuals to test for first-order serial autocorrelation. The obtained value of DW = 2.008 falls squarely within the acceptable range of 1.5–2.5, indicating no significant positive or negative autocorrelation in the residuals. This result validates the independence assumption of OLS and confirms that the model residuals do not exhibit systematic temporal patterns.

6.3 Normality of Residuals (Shapiro-Wilk)

The Shapiro-Wilk test applied to the residuals yielded a test statistic of W = 0.8036 ($p < 0.0001$), indicating a statistically significant departure from normality. This rejection of the normality assumption is common in large housing datasets where extreme luxury properties create heavy-tailed residual distributions. The price distribution's skewness of 4.024 and kurtosis of 34.586 confirm substantial non-normality. While this technically violates an OLS assumption, the Central Limit Theorem ensures that coefficient estimates remain asymptotically consistent and unbiased in large samples (n = 17,290 training observations), though standard error estimates may be affected.

The log-transformed model partially addresses this concern by reducing price skewness and producing residuals more closely approximating normality. In practice, robust standard errors or bootstrapped confidence intervals are recommended for inference from either specification.

6.4 Heteroscedasticity

Visual inspection of the residuals-versus-fitted-values plot reveals a characteristic fan-shaped pattern, indicating the presence of heteroscedasticity — specifically, that prediction error variance increases with fitted price values. This is expected in property data where high-value homes are more idiosyncratic and harder to predict uniformly. Formal Breusch-Pagan testing confirms statistically significant heteroscedasticity ($p < 0.001$). To address this, weighted least squares (WLS) or heteroscedasticity-consistent (HC) standard errors would be appropriate extensions of the current analysis.

7. Discussion

7.1 Interpretation of Key Findings

The empirical results yield several substantively important findings about the King County residential property market. The dominance of construction grade as a price determinant (β = \$118,596 per grade point) underscores the premium buyers place on material quality, architectural design, and craftsmanship. Grade encompasses the type and quality of materials used in construction and finish level, representing accumulated investment decisions at the time of construction that are difficult to reverse without major renovation.

The living area square footage coefficient of \$107.68 per square foot is economically meaningful and consistent with market-level estimates for the Seattle metropolitan area during 2014–2015. This implies that a 1,000 square foot expansion of living space would add approximately \$107,680 to property value, all else equal — a figure that can inform cost-benefit analyses for home additions and renovation projects.

The negative bedroom coefficient (−\$38,163), controlling for total living area, reflects a well-documented empirical regularity in hedonic housing studies. Holding total square footage constant, additional bedrooms reduce the average room size, which many buyers discount relative to open floor plans. This finding highlights the importance of variable interaction and the need to control for size when interpreting room count effects.

7.2 Limitations

Several limitations of the current analysis warrant acknowledgment. First, the absence of spatial fixed effects (e.g., ZIP code dummy variables or latitude-longitude controls) likely understates the model's explanatory power. Location is consistently the most powerful determinant of residential values, and the model's omission of fine-grained spatial controls contributes to the unexplained 34.7% of price variance.

Second, the cross-sectional nature of the data (a single 12-month window) precludes analysis of temporal price dynamics or seasonality effects. Third, the identified multicollinearity among size variables constrains the precise separate estimation of sqft_above and sqft_basement contributions. Finally, unobserved characteristics — interior finishes, landscaping quality, proximity to specific amenities — that are not captured in the administrative dataset represent omitted variable bias that cannot be remedied without richer data collection.

7.3 Practical Implications

From a practical standpoint, the model provides a useful automated valuation mechanism for pre-screening property values. An RMSE of \$229,186 on a mean price of \$540,088 implies a coefficient of variation for prediction error of approximately 42%, which is typical for MLR-based automated valuation models (AVMs) without spatial adjustment. Real estate professionals and lenders can use such models as initial screening tools, with human appraisal reserved for properties flagged as outliers relative to model predictions.

The finding that waterfront status alone accounts for a premium exceeding \$554,000 has direct implications for infrastructure investment decisions, land use policy, and waterfront property taxation. Similarly, the strong grade effect suggests that building quality improvements yield substantial returns in the King County market, with implications for renovation investment strategy.

8. Conclusion

This study has presented a rigorous, data-driven analysis of residential property valuation in King County, Washington, employing Multiple Linear Regression techniques on a comprehensive dataset of 21,613 property transactions. The principal findings demonstrate that living area square footage, construction grade, waterfront access, and view quality are the dominant price determinants in this market, collectively explaining approximately 65% of sale price variance in both standard and log-transformed MLR specifications.

The model diagnostics reveal that while the Durbin-Watson statistic (2.008) confirms residual independence and most predictors are free from multicollinearity, the data exhibits significant positive price skewness (4.024) and heteroscedasticity — features that are endemic to residential transaction data and that motivate the use of log-transformation, robust standard errors, and spatial extensions.

The log-transformed MLR model offers marginal improvements in adjusted R^2 (0.6543 vs. 0.6514) and mean absolute error (\$139,114 vs. \$143,500), suggesting better prediction of typical-priced properties, while the standard MLR model demonstrates lower RMSE (\$229,186 vs. \$273,541 back-transformed) for point estimates across the full price range.

Future research directions include: (1) incorporating ZIP code or neighborhood fixed effects to capture spatial heterogeneity; (2) applying spatial autoregressive models (SAR) or geographically weighted regression (GWR) to account for spatial autocorrelation; (3) comparing MLR performance with ensemble machine learning methods (Random Forest, Gradient Boosting) to evaluate non-linearity gains; and (4) extending the temporal window to analyze market dynamics and seasonality effects on hedonic coefficients.

9. References

[1] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55.

[2] Lancaster, K. J. (1966). A new approach to consumer theory. Journal of Political Economy, 74(2), 132–157.

[3] Kain, J. F., & Quigley, J. M. (1970). Measuring the value of housing quality. Journal of the American Statistical Association, 65(330), 532–548.

[4] Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. Journal of Real Estate Literature, 13(1), 3–44.

[5] Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. Journal of Real Estate Finance and Economics, 35(2), 143–160.

[6] Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. O'Sullivan & K. Gibb (Eds.), Housing Economics and Public Policy (pp. 67–89). Blackwell Science.

[7] Oladunni, T., & Sharma, S. (2018). Hedonic housing theory: A machine learning investigation. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 471–477). IEEE.

[8] King County Department of Assessments. (2015). King County House Sales Dataset [Data file]. Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction

[9] Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics (5th ed.). McGraw-Hill Education.

[10] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

[11] Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. American Economic Review, 79(1), 125–137.

[12] Anselin, L. (1988). Spatial Econometrics: Methods and Models. Kluwer Academic Publishers.