Exploring the Determinants of Housing Prices: Insights from the U.S. Market

Sitong Wang *

Massey Institute, Nanjing University of Finance and Economics, Nanjing, China

* Corresponding Author Email: Sitong.Wang.1@uni.massey.ac.nz

Abstract. Housing prices significantly influence economic stability, household wealth, and urban development, making their determinants a crucial area of study. This study investigates the determinants of housing prices in the U.S. market, focusing on structural, locational, and temporal factors. Using a dataset of 4,600 housing samples from Kaggle, the research employs multiple regression analysis and interaction models to quantify the effects of variables such as living area, year built, and waterfront presence. Results show that additional living space, bathrooms, and waterfront houses have positive price effects, while construction year and additional bedrooms have adverse effects. Interaction effects also show that older houses have diminishing price effects from the living space. However, serious multicollinearity problems with the interaction model raise concerns about exercising caution when interpreting and using alternative methods. There are implications for policymakers looking for greater affordability, estate agents setting price policies, and researchers using new modelling techniques to research housing markets.

Keywords: Housing price determinants, Multiple regression analysis, Interaction effects, U.S. housing market.

1. Introduction

Property prices are among the strongest economic development and urbanization markers and reflect the general socio-economic condition [1]. Determinants of property prices are crucial to long-run market forecasting and decision-making [2].

The property market is complex because many factors, like construction characteristics, timing, and locational advantages, determine the value of residential property. Structural characteristics, including the size of the dwellings and proximity to amenities, are crucial to the price determination of housing. Huang et al. utilized the spatial quantile regression methodology in Changsha, China [3]. They concluded that housing near urban blue spaces, i.e., rivers and lakes, enjoyed a 5–15% price premium, with the effect being more pronounced in higher-priced housing segments [4]. Time factors, including renovation history, also influence housing prices. Bogin and Doerner investigated the impact of renovations on property house price indices. They concluded that recently renovated properties enjoyed a mean value appreciation of 10–12% in the year after renovation, depending on the extent and quality of renovation work undertaken [5].

Locational attributes remain significant determinants of the value of housing. Cheshire and Sheppard provided evidence of the effect of being close to high-demand amenities, such as high-performing schools, in increasing property value by an average of 12% [6]. Gyourko and Saiz further emphasized the influence of construction costs, with neighborhoods of more costly labour and materials reporting increases in property value of 20–30% compared to lower-cost neighborhoods [7]. Despite advances in the modelling of housing prices, Malpezzi emphasized the need to control for interaction effects, such as the combined effect of structural and locational determinants. He demonstrated that not controlling for these dynamics would create biases in valuations of up to 15% in hedonic price models [8]. Goodman and Thibodeau also emphasized the need to segment housing markets, demonstrating that controlling for segmentation in models enhanced predictive power by nearly 10% [9].

This study builds upon these bases based on a sample of 4,600 housing observations on Kaggle. Multiple regression analysis and interaction models are employed to examine the interaction effects of living area, year of construction, waterfront view, and the like on house prices. This study aims to bridge gaps in the literature and contribute to our understanding of the mechanisms of the housing market with policy-related recommendations to policymakers, property practitioners, and researchers.

2. Method

2.1. Data Sources and Indicator Selection

The data employed in this study were collected from Kaggle and provided by Shree [10]. The database has 4,600 samples of homes on the US housing market, spanning extensively to permit a broad analysis of housing price determinants. The independent variables include structural, temporal, and locational aspects to provide a balanced perspective of the house value determinants.

The key measures are the number of bedrooms, number of bathrooms, size of an above-ground living area ($\mathrm{Sqft_{living}}$), the total size of the lot ($\mathrm{Sqft_{lot}}$), the number of stories of the property, whether waterfront views exist (Waterfront), size of an above-ground living area ($\mathrm{Sqft_{above}}$), construction year of the property ($\mathrm{Yr_{built}}$), and year of the last renovation($\mathrm{Yr_{renovated}}$). These measures combine intrinsic property characteristics, including size and age, with contextual characteristics, such as location and refurbishments.

The dependent variable, ln (Price), is the natural log of house prices and is of interest. Housing prices are highly variable, with a mean of 13.065 and a standard deviation of 0.543, indicating a wide variation among houses. Data preprocessing was done to handle extreme outliers and missing values appropriately to ensure data quality and reliability.

Table 1 presents the summary statistics of the key variables, their types, and their ranges, means, and standard deviations to provide an idea about the structure and variability of the dataset.

Table 1. Basic indicators

Variable	minimum	maximum	Average	Sd	median
ln(price)	8.962	17.096	13.065	0.543	13.050
bedrooms	0.000	9.000	3.401	0.909	3.000
bathrooms	0.000	8.000	2.161	0.784	2.250
$\mathrm{Sqft_{living}}$	370.000	13540.000	2139.347	963.207	1980.000
$\mathrm{Sqft_{lot}}$	638.000	1074218.000	14852.516	35884.436	7683.000
floors	1.000	3.500	1.512	0.538	1.500
waterfront	0.000	1.000	0.007	0.084	0.000
$\mathrm{Sqft_{above}}$	370.000	9410.000	1827.265	862.169	1590.000
$\mathrm{Yr_{built}}$	1900.000	2014.000	1970.786	29.732	1976.000
$\mathrm{Yr_{renovated}}$	0.000	2014.000	808.608	979.415	0.000
$\mathrm{Sqft_{living}}*\mathrm{Yr_{built}}$	711510.000	27066460.000	4224435.175	1922591.228	3891654.000

The selection of these indicators ensures a balanced representation of the factors impacting ln (Price), encompassing both the intrinsic attributes of the properties and the external influences of location and age.

2.2. Methodology Introduction

To analyze the relationship between these variables and housing prices, this study employs multiple linear regression analysis. The primary objective is to quantify how various property features influence housing prices while controlling for other factors. The regression model 1 is expressed as follows:

$$\begin{aligned}\mathrm{Ln(price)} = \beta_0 + \beta_1 \mathrm{Bedrooms} + \beta_2 Bathrooms + \beta_3 \mathrm{Sqft_{living}} + \beta_4 \mathrm{Sqft_{lot}} + \\ \beta_5 \mathrm{Yr_{built}} + \beta_6 Waterfront + \beta_7 Yr_{renovated} + \varepsilon\end{aligned} \tag{1}$$

In this equation, Price represents the housing price (dependent variable), while the independent variables (such as Bedrooms, Bathrooms, $\mathrm{Sqft}_{\mathrm{living}}$, and so on) represent the property features. The coefficients $\beta_1, \beta_2, \ldots.. \beta_6$ indicate the magnitude of the effect each feature has on the housing price, and ϵ is the error term, capturing unexplained variation.

Furthermore, to explore potential interaction effects between different variables, the model was extended by including interaction terms. For example, the joint effect of living area ($\mathrm{Sqft}_{\mathrm{living}}$) and year built ($\mathrm{Yr}_{\mathrm{built}}$) was considered, as these two variables may have a compounded influence on property prices. The enhanced model 2 is as follows:

$$\mathrm{Ln(price)} = \beta_0 + \beta_1 \mathrm{Bedrooms} + \beta_2 Bathrooms + \beta_3 \mathrm{Sqft}_{\mathrm{living}} + \beta_4 \mathrm{Sqft}_{\mathrm{lot}} + \beta_5 \mathrm{Yr}_{\mathrm{built}} + \beta_6 Waterfront + \beta_7 Yr_{renovated} + \beta_8 \left(\mathrm{Sqrt}_{\mathrm{living}} * \mathrm{Yr}_{\mathrm{built}}\right) + \varepsilon \quad (2)$$

By including interaction terms, this model can capture the combined effects of multiple variables on housing prices, offering a more nuanced understanding of their relationships.

To evaluate the effectiveness of the regression model, performance metrics such as adjusted R^2 were used, which indicates the proportion of variance in housing prices explained by the independent variables. Diagnostic tests for model robustness were performed, including residual analysis and variance Inflation Factor (VIF) to check multicollinearity and heteroscedasticity tests.

In this way, the research will better explain residential price determinants, with the findings contributing significantly to the decision-making of policymakers, real estate professionals, and investors.

3. Results and Discussion

3.1. Key Findings from Model 1 (No Interaction Term)

The specific regression coefficients and covariance diagnostics are given in Table 2 below:

Table 2. Results of linear regression analysis

	Unstandardized Coefficients		Unstandardized Coefficients	*t*	*p*	Collinearity Diagnosis	
	B	Std. Error	*Beta*			VIF	Tolerance
constant	20.947	0.463	-	45.285	0.000**	-	-
bedrooms	-0.060	0.008	-0.101	-7.530	0.000**	1.664	0.601
bathrooms	0.128	0.013	0.184	9.815	0.000**	3.257	0.307
sqft_{living}	0.000	0.000	0.630	22.435	0.000**	7.344	0.136
sqft_{lot}	-0.000	0.000	-0.048	-4.491	0.000**	1.075	0.930
floors	0.138	0.014	0.137	9.732	0.000**	1.840	0.544
waterfront	0.358	0.070	0.053	5.089	0.000**	1.026	0.974
sqft_{above}	-0.000	0.000	-0.012	-0.468	0.640	6.119	0.163
$\mathrm{yr}_{\mathrm{built}}$	-0.005	0.000	-0.248	18.951	0.000**	1.589	0.629
$\mathrm{yr}_{renovated}$	-0.000	0.000	-0.004	-0.382	0.703	1.134	0.882
R^2	0.512						
Adjusted R^2	0.511						
F	*F* (9, 4541) =529.273, *p*=0.000						
D-W value	1.999						
Remarks: dependent variable = ln(price)							
* p<0.05 ** p<0.01							

Model 1 identified several significant predictors of housing prices. Bedroom count had a surprising adverse effect, perhaps reflecting demand on the part of buyers for fewer but bigger or better-quality bedrooms instead of smaller bedrooms. Bathrooms and living area, as expected, positively influenced property values, emphasizing the importance of functional and spacious features in determining price. Additionally, older homes were associated with lower prices, likely due to depreciation or outdated characteristics.

Interestingly, variables such as $\mathrm{sqft}_{\mathrm{above}}$ and $\mathrm{yr}_{renovated}$ were statistically insignificant, indicating limited influence on housing prices in this dataset. In contrast, waterfront properties exhibited a strong positive effect, highlighting the premium that buyers place on such desirable locations.

3.2. Key Findings from Model 2 (With Interaction Term)

Model 2 introduced an interaction term between sqft_{living} and $\mathrm{yr}_{\mathrm{built}}$ ($\mathrm{Sqft}_{\mathrm{living}}$*$\mathrm{Yr}_{\mathrm{built}}$) to explore whether the effect of living area on price varies with the age of the property.

The specific regression results and covariance diagnostics for the model including the interaction term are shown in Table 3.

Table 3. Results of linear regression analysis

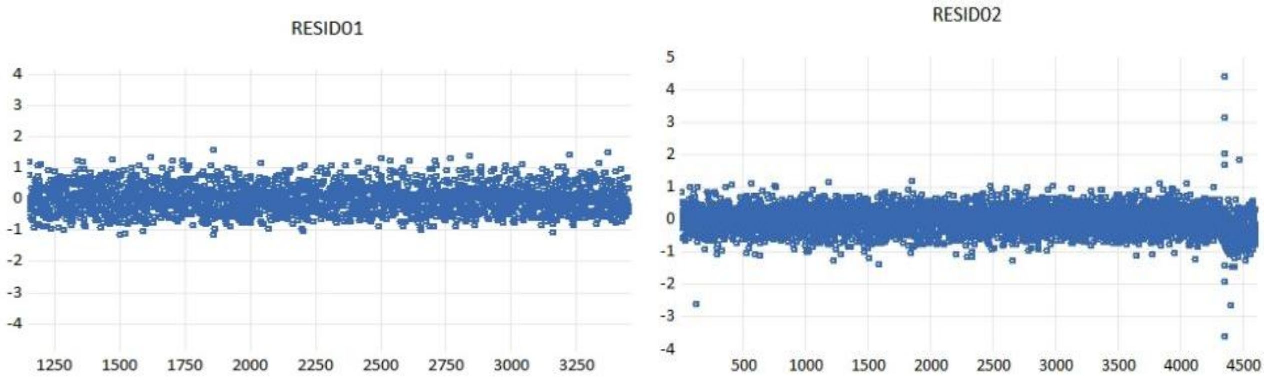
	Unstandardized Coefficients		Unstandardized Coefficients	*t*	*p*	Collinearity Diagnosis	
	B	Std. Error	*Beta*			VIF	Tolerance
constant	17.352	1.006	-	17.253	0.000**	-	-
bedrooms	-0.065	0.008	-0.109	-8.061	0.000**	1.702	0.588
bathrooms	0.124	0.013	0.177	9.451	0.000**	3.281	0.305
sqft_{living}	0.002	0.000	3.720	4.843	0.000**	5507.373	0.000
sqft_{lot}	-0.000	0.000	-0.048	-4.503	0.000**	1.075	0.930
floors	0.126	0.014	0.125	8.724	0.000**	1.920	0.521
waterfront	0.342	0.070	0.051	4.858	0.000**	1.030	0.971
sqft_{above}	0.000	0.000	0.018	0.659	0.510	6.623	0.151
$\mathrm{yr}_{\mathrm{built}}$	-0.003	0.001	-0.147	-5.213	0.000**	7.425	0.135
$\mathrm{yr}_{renovated}$	-0.000	0.000	-0.004	-0.387	0.699	1.134	0.882
$\mathrm{Sqft}_{\mathrm{living}}$*$\mathrm{Yr}_{\mathrm{built}}$	-0.000	0.000	-3.131	-4.025	0.000**	5651.530	0.000
R^2	0.514						
Adjusted R^2	0.513						
F	*F* (10, 4540) =479.559, *p*=0.000						
D-W value	2.002						
Remarks: dependent variable = ln (price)							
* *p*<0.05 ** *p*<0.01							

The interaction term was statistically significant, indicating that the positive impact of living area on price diminishes for older homes. This nuanced finding suggests that while larger living spaces generally increase value, this effect is less pronounced in older properties, possibly due to wear and tear or outdated designs.

However, the inclusion of the interaction term introduced severe multicollinearity, as evidenced by extremely high Variance Inflation Factor (VIF) values for sqft_{living} (5507.373) and the interaction term (5651.530). Such multicollinearity undermines the reliability of the coefficient estimates for these variables, signaling the need for alternative modeling approaches, such as ridge regression or the removal of highly correlated predictors.

3.3. Model Diagnostics and Robustness

Diagnostic tests were conducted to assess the validity of both models. The Durbin-Watson statistics (Model 1: DW = 1.999; Model 2: DW = 2.002) indicated no significant autocorrelation in the residuals, supporting the assumption of independent errors. The residual plots for Model 1 (without the interaction term) revealed a random scatter of residuals around zero, with no discernible patterns (see Fig. 1(a)). This suggests that the model meets the homoscedasticity assumption, as the variance of residuals remains constant across predicted values. Similarly, the residual plots for Model 2 (with the interaction term) exhibited no systematic trends, further confirming the absence of heteroscedasticity (see Fig. 1(b)). However, the multicollinearity issue in Model 2 necessitates caution in interpreting the interaction effect.



(a) (b)

Figure 1. Residual plots for Model 1 (a) and Model 2 (b) (Photo credit: Original).

3.4. Comparative Insights

Comparing the two models, the inclusion of the interaction term in Model 2 provided only a slight improvement in explanatory power (adjusted R^2 increased from 0.511 to 0.513). This suggests that while the interaction between living area and year built is statistically significant, its practical impact on price prediction is minimal. The severe multicollinearity further complicates the interpretation of Model 2, highlighting the trade-off between model complexity and interpretability.

3.5. Limitations and Future Research

This study identifies two primary limitations. First, multicollinearity in Model 2 poses significant challenges to interpreting interaction effects. As noted by Weeks, overlooking such issues can lead to biased estimates and reduce the model's reliability [11]. Future research could address this limitation by employing regularization techniques, such as ridge regression, to enhance model stability. Second, the study's reliance on linear modeling may oversimplify the inherently complex relationships in housing price determinants. Case and Shiller emphasize the importance of capturing nonlinear dynamics, which are often pivotal in accurately modeling housing markets [12]. Adopting advanced methods, such as machine learning algorithms, could uncover more intricate patterns and improve predictive accuracy.

4. Conclusion

This study provides a comprehensive analysis of the factors influencing housing prices in the U.S. market, utilizing a large dataset of 4,600 samples and advanced regression techniques. The key findings reveal that structural, locational, and temporal features significantly impact housing prices. Fascinating results are the positive impact of bathrooms, living areas, and waterfront homes and the negative impact of older years of construction and additional bedrooms. The introduction of an interaction term between living area and construction year highlights a nuanced relationship, where the value-added by larger living spaces diminishes in older homes. However, the observed multicollinearity in Model 2 underscores the trade-off between capturing complex interactions and maintaining model robustness.

These insights offer practical implications for various stakeholders. Buyers and sellers can use the findings to inform pricing strategies by understanding the premium placed on specific features such as bathrooms and waterfront status. Policymakers are encouraged to consider sustainable management of coastal properties and develop strategies to improve affordability. For researchers, the study emphasizes the importance of addressing multicollinearity and exploring nonlinear effects in future housing market analyses.

In conclusion, this research advances the understanding of housing price determinants by leveraging large-scale data and interaction models, providing actionable insights for various market participants. Future studies could build on these findings by incorporating regularization techniques and exploring alternative model structures to address the observed limitations.

References

[1] Oluku U, Cheng S. A regional analysis of the relationship between housing affordability and business growth. Economic Development Quarterly, 2021, 35 (4): 269 – 286.

[2] Zhou S. Investigation of influential factors of housing price. Advances in Economics, Management and Political Sciences, 2024, 105: 55 – 62.

[3] Jiang Y, Qiu L. Empirical study on the influencing factors of housing price based on cross-section data of 31 provinces and cities in China. Procedia Computer Science, 2022, 199: 1498 – 1504.

[4] Huang T, He Q, Yang D, Ouyang X. Evaluating the impact of urban blue space accessibility on housing price: A spatial quantile regression approach applied in Changsha, China. Frontiers in Environmental Science, 2021, 9.

[5] Bogin A N, Doerner W M. Property renovations and their impact on house price index construction. Journal of Real Estate Research, 2019, 41 (2): 249 – 284.

[6] Gyourko J, Saiz A. Construction costs and the supply of housing structure. Journal of Regional Science, 2006, 46 (4): 661 – 680.

[7] Cheshire P, Sheppard S. On the price of land and the value of amenities. Economica, 1995, 62 (246): 247 – 267.

[8] Malpezzi S. Hedonic pricing models: A selective and applied review. In: O’Sullivan T, Gibb K, eds. Housing Economics and Public Policy. 2003: 67 – 89.

[9] Goodman A C, Thibodeau T G. Housing market segmentation and hedonic prediction accuracy. Journal of Housing Economics, 2003, 12 (3): 181.

[10] [Shree. House price prediction. Kaggle, 2018. Available at: https://www.kaggle.com/datasets/shree1992/housedata/data [Accessed June 5, 2025].

[11] Weeks M. Introductory Econometrics: A Modern Approach. Vol. 17, 2002.

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