Research on the Influencing Factors of House Prices in Sydney

Hengjia Xu *

Changjun High School, Changsha, 410000, China

* Corresponding Author Email: xhj060306@outlook.com

Abstract. Considering various factors that affect house prices, this study investigates whether housing prices in Sydney can maintain their value or even increase in the future. A multiple linear regression model was applied using a dataset obtained from Kaggle, which includes 11,161 real estate transactions in Sydney from 2016 to 2021. This paper uses it to explore the relationship between housing prices and 12 independent variables (originally 14 but some have to be abandoned), such as the number of bedrooms, property size, median suburban income, and distance to the central business district (CBD). The results indicate that some factors have a positive, negative, or no impact on house prices and that 39.6% of the real estate price is influenced by them, indicating a moderate level of predictability. These findings provide valuable insights for investors and policymakers interested in the dynamics of the Sydney real estate market, emphasizing the importance of geographic location and property attributes in determining housing prices.

Keywords: Housing Prices; Multiple Linear Regression; Influencing Factors.

1. Introduction

As a crucial sector in Australia for decades, the issue concerning house prices has been consolidating the status as the dominant topic among the public. Particularly in the past thirty years, house prices in metropolises have immensely soared, making property investment a common wealth-building strategy. However, the global pandemic has brought unprecedented and unpredictable challenges, causing a temporary and moderate decline in house prices. As the current situations of real estate markets fluctuate and variate, controversy continues to arise about the soundness of house prices as an investment or a means to protect property. Therefore, accurately predicting price volatility trends has become a focus of research in Australia, given its significant impact on the economy and personal wealth.

Australian house prices have risen a lot in the past 30 years, Hamish et al. found the purely financial approach has often dominated the real state market's real priorities [1]. House prices are affected by a number of factors. Huang stated that when there is a high expectation of yuan appreciation, it will attract capital inflows and cause housing prices to rise [2]. In addition, Jacob et al. believed that the expansion of financial leverage exists as an indispensable factor of the growth of housing prices [3]. Moreover, Sun and Zhao pointed out that the urban rail transit loop exerts a positive influence on the value of surrounding real estate [4].

The fundamental purpose of analyzing the correlation between many variables and the price is to make the predicted price more accurate and precise. In the research process of many scholars, prices are predicted by many mathematical models. Theurillat et al. used the principal component analysis-back propagation neural network (PCA-BPNN) algorithm to analyze and model prices and found that this model is more accurate than BPNN and this model (PCA-BPNN) can help support the decisions made [5]. Xia used the Gaussian Mixed Mod (GMM) method to apply it to short-term price prediction and the finding is that it is the proportion of population flow well-educated plays a pivotal role in the house prices [6]. At the same time, more and more scholars have found that a single model is not ideal for predicting house prices in most cases, and using more and complicated models can really be useful. Yu chose to use panel regression model to study the impact of low housing price on the proportion of population in each age group and found that there was a positive correlation between the proportion of population aged 65 and above and housing price [7]. Then Zhang et al. used multiple linear regression model to investigate the impact of population proportion on the development potential of housing price, and finally concluded that the increase of the proportion of youth and the

elderly would promote the climb of the house price [8]. Zhang employed fixed effects model and mediator model and summarize the results into a point that an increase in income levels promotes the value of housing in China [9]. In addition, Zhao et al. applied a combination model and value at risk (VAR) model to speculate whether the house bubble will occur [10]. Furthermore, Hamish suggested that women who expect to work more in the future purchase more expensive houses [1]. He explored the relationship between the labor and house price using log-linear model and multiple regression analysis.

To sum up, the research on the important data of house prices has attracted numerous scholars. This article mainly used multiple linear regression model to predict and analyze oil prices and provide corresponding suggestions to owners of houses and new investors based on the predicted results.

2. Methods

2.1. Data Source

The dataset used in this paper is fetched from the Kaggle website (Sydney house prices). It was from 2016 to 2021, collected on domain.com.au by Alex Lau. This dataset contains 11161 groups of data, and this research selected all of them as samples to make the data reliable. The original dataset remained in .csv format.

2.2. Variable Selection

The original dataset consists of 11,160 observations and 17 variables, capturing a wide range of information about the trades of Australia houses. However, due to the data in words instead of digits, several variables were excluded from the final analysis. Additionally, because of the existence of inconsistent and excessive values, these variables were also removed. A random sampling was conducted to select 1000 observations for the final analysis. The dataset used in the study contains 12 independent variables (property size, number of bedrooms, number of bathrooms, number of parking spaces, suburb population, suburb median income, suburb area, latitude, longitude, elevation, and distance from CBD) and one dependent variable (price). The dataset's summary is shown in Table 1.

Table 1. List of Variables

Variable	Algebra	Meaning
date_sold	X1	The date was sold.
num-bath	X2	The number of bathrooms
num_bed	X3	The number of bedrooms
num_parking	X4	The number of parking lot
property size	X5	The size in square metres
population	X6	The population of the surburb
median_income	X7	The median income in surburb
square kilometre	X8	The size of surburb
latitude	X9	The latitude of the suburb
longitude	X10	The longitude oft he suburb
elevation	X11	The elavation oft he suburb
km_from_cbd	X12	The distance from CBD
price	Y	The price

2.3. Method Introduction

This article mainly uses multiple linear regression models to analyse the significance of the factors that can affect house prices to predict the future years in Australia. Then this paper uses the predicted results to compare to the actual results to conclude that whether this model is accurate.

Multiple linear regression is a statistical method used to analyze the linear relationship between two or more independent variables and one dependent variable. Such models can help us understand how different factors collectively affect the outcome variable and can be used to predict or control the value of the outcome variable

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_{12} + e \quad (1)$$

Y is the dependent variable and x_1, $x_2 \cdots x_n$ are independent variables. β_0 is the intercept term that represents the expected value of the dependent variable when all the independent variables are zero. $\beta_1, \beta_2, \cdots \beta_n$ is the regression coefficient, which represents the expected change of the dependent variable for each change of one unit of the corresponding independent variable. ϵ is the error term, representing the effect of factors other than the independent variable on the dependent variable, which is assumed to be random and follows a normal distribution with zero mean.

3. Results and Discussion

3.1. Descriptive Analysis

Scatter diagram is used to show the relationship between a single independent variable and a dependent variable (Figure 1). By observing the distribution pattern of scatter points, it can initially determine whether there is a linear relationship to decide whether to use this method. Set x_7 and x_{12} as two examples.

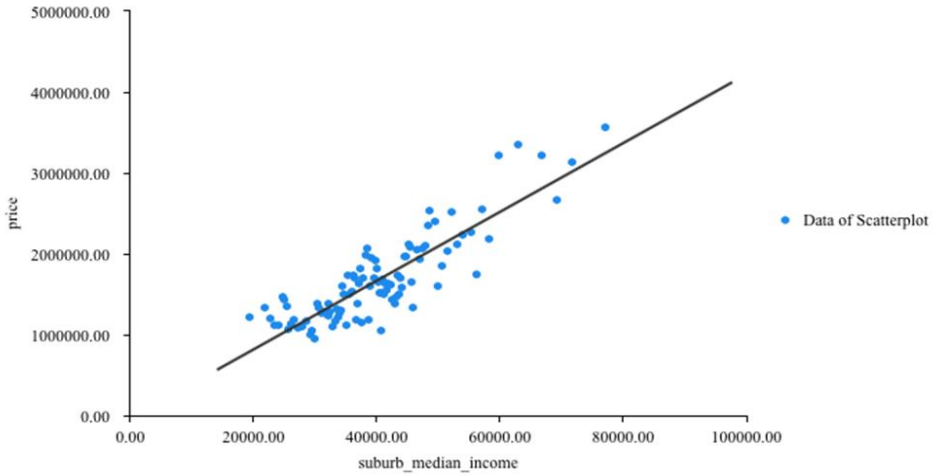

Fig. 1 The scatter diagram and the best-fit line of x_7

$$E(Y) = 3.26 \times 10^4 + 42.5 x_7 \quad (2)$$

The x-axis represents "the median income of the suburb", then this graph gives information about that as the median income of the suburb increases, housing prices tend to increase. This is a common pattern in urban economics there is a positive correlation between the median income of the suburb and the house price.

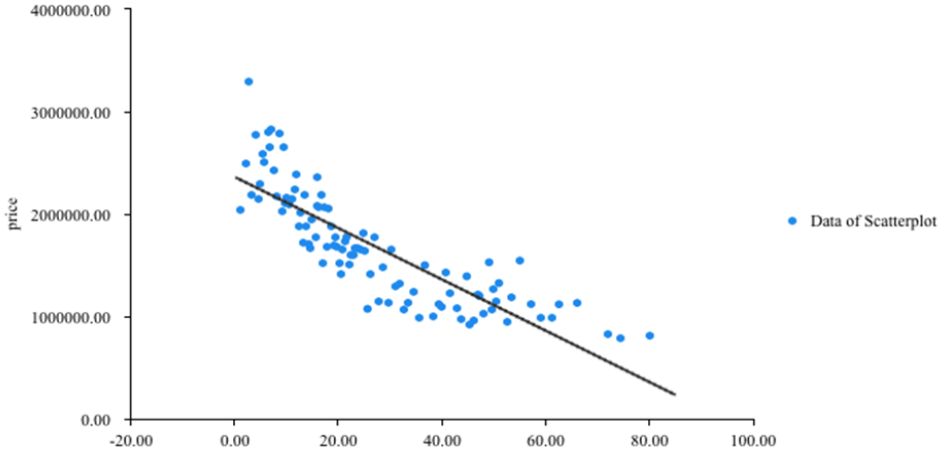

Fig. 2 The scatter diagram and the best-fit line of x_{12}

$$E(Y) = 2.36 \times 10^6 - 2.5 \times 10^4 x_{12} \quad (3)$$

The x-axis represents "distance from CBD", then figure 2 gives information about that as the distance from the CBD increases, housing prices tend to decrease. This is a common pattern in urban economics where proximity to central areas is often associated with higher property values.

Figure 1 and 2 illustrate the comparison and variation of x_7 and x_{12}. Overall, these two examples fit the linear relationship, it reflects this methodology is suitable. Moreover, the data is evenly distributed around the best-fit line, and the data occupy more than half of the area, which give information that the best-fit lines are proper. Furthermore, the value of coefficient demonstrates the significance of the variable to the dependent variables, e.g., the coefficient of x_{12} is 25000, while that of x_7 is 42.5, meaning x_{12} affects the price more immensely than x_7.

3.2. Multiple Linear Regression

This table 2 presents the results of multiple linear regression analysis, including the independent Variable name and regression coefficient, which reflects how the expected value changes. Standard error, which measures the accuracy of the regression coefficient estimate.

Table 2. Analysis results

Variable	B	Std.Error	Beta	T Value	P Value	VIF	Tolerance
onstant	-1,78×108	1.45×107	-	-12.23	0.00	-	-
Date_sold	-0.000	-0.000	-0.01	-1.52	0.130	1.00	0.998
Num_bath	2.34×105	1.36×104	0.215	17.2	0.00	2.89	0.346
Num_bath	1.19×105	1.04×104	0.144	11.4	0.00	2.94	0.340
Num_parking	6.72×104	8.25×103	0.076	8.14	0.00	1.60	0.626
Property_size	155.7	9.6	0.127	16.3	0.00	1.12	0.895
Suburb_population	-0.753	1.46	-0.004	-0.52	0.606	1.35	0.742
Median_income	29.6	1.00	0.254	29.4	0.00	1.38	0.727
Suburb_sqkm	3.46×103	1.97×103	-0.016	-1.76	0.078	1.46	0.687
Suburb_lat	2.44×104	9.75×104	0.004	0.25	0.803	4.32	0.231
Suburb_Ing	1.16×106	7.64×104	0.191	15.1	0.00	2.95	0.339
Suburb_elevation	376	207	0.015	1.82	0.069	1.32	0.756
Km_from_cbd	-2.09×104	929	-0.3	-22.5	0.00	3.27	0.306

R2:0.396
Adjusted R2:0.396
F value:F(12,11147)=610.086,p=0.00
D-W value:1.570

The value will be accurate if this is low. A standardized regression coefficient (Beta), allowing comparison of the relative size of the influence of different independent variables on the dependent variable. T value, which tests whether the regression coefficient is significantly different from zero. P Value, used to determine statistical significance, usually less than 0.05 indicates that the variable has a significant effect on the dependent variable. Variance inflation factor (VIF) is used to detect multicollinearity between independent variables. VIF values greater than 10 may stand for serious collinearity problems. Tolerance, the reciprocal of VIF, is used to measure the severity of collinearity, if the value is small, the problem is serious. Together, these statistical indicators help us understand the influence of each independent variable in the model on housing price and its statistical significance, and evaluate the collinearity problems that may exist in the model. After the discussion finishes, Skew-Normal Conditions was taken into account. Zhang et al. pointed out it is proved that the parameter maximum likelihood estimation is unique with probability 1 under the condition n≥p+1 [9]. Fortunately, it is finally tested these data conform to a normal distribution. The following steps will apply the data from this table.

β indicates Unstandardized coefficients, whereas Beta denotes standardized coefficients. The procedure will be done as follows. The first step is to set an original model in the form:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_{12} + e \quad (4)$$

Then referring to the coefficients peivided in the Table 1, gaining

$$E(Y) = -1.78 \times 10^8 + 0.00 x_1 + 2.34 \times 10^5 x_2 + \cdots + -2.19 \times 10^4 x_{12} \quad (5)$$

3.3. Collinearity Diagnostics

Then next thing is to taking all the factors which may influence each other into account. But the VIP of all the factors is less than 5, meaning that the factors have no significant impact on each other. So, the equation of this model remains the same.

In the context of multiple regression analysis, statistical significance is a tool to identify how the independent variables have their impacts on the dependent variables, it helps make a decision and verify. The results work as follows: The variable representing "date_sold" (x_1) has a regression coefficient of -0.000, which is not statistically significant (t-value = -1.520, p-value = 0.129), indicating that there is no significant relationship between the time of sale and the property price.

On the contrary, the number of bathrooms ('num_bath', x_2) and the number of bedrooms ('num_bed', x_3) are directly proportional to the housing price, and their regression coefficients are 2343300.596 and 119062.385, respectively, with a p value significantly less than 0.01. This highlights the value that buyers place on these amenities.

The availability of parking Spaces, represented by "num_parking" (x_4), also considerably positively affects real estate prices. And the regression coefficient is 67,204.967, indicating the importance that convenience in the form of parking facilities is a highly valued attribute in the real estate market. The size of the property, as shown by "property_size" (x_5), is another important factor of price, with a regression coefficient of 155.698, further confirming the principle that larger living space is associated with higher market value.

While demographic factors such as' suburban population '(x_6) and' suburban area '(x_9) do not appear to affect property prices, as shown by their non-significant P-values, economic indicators such as' suburban median income' (x_7) and geographical coordinates such as' suburban LNG '(x_{10}) show a significant positive correlation with property prices. The regression coefficients are 29.567 and 1155036.319 respectively, and the p value is much lower than the threshold value of 0.01, highlighting the economic and location premium of the housing market.

The distance from CBD, represented by "km_from_cbd" (x_{12}), has an immensely negative impact on the housing price, with a regression coefficient of -20947.237 and a p value of 0.000, indicating that the distance from CBD is the key factor playing a pivotal role in the housing price.

Regression coefficients for "suburban square kilometres" (x_8) and "suburban elevation" (x_{11}) did not reach statistical significance, suggesting that suburban size and elevation do not play a key role in property pricing.

4. Conclusion

In conclusion, the study provides a comprehensive analysis of the factors influencing house prices in Sydney, employing a multiple linear regression model to examine the relationship between property value and a suite of independent variables. The findings reveal that certain factors, such as the number of bathrooms, bedrooms, parking spaces, property size, median suburban income, and longitude coordinates, significantly and positively affect house prices. Conversely, the distance from the CBD exhibits a substantial negative impact. The model's moderate R-square value of 0.396 indicates a reasonable fit for the data, suggesting that while the model captures key market dynamics, other variables may also influence house prices. This research offers valuable insights for investors and policymakers, emphasizing the complex interplay of factors that shape the Sydney real estate market.

References

[1] Hamish L, et al. Labour of supply and cost of house price booms and busts. Labour Economics, 2024, 102591.

[2] Huang Lina. Analysis of the Correlation between RMB Exchange Rate and Real Estate Prices. Science and Technology Economy Market, 2020, 9: 29-30.

[3] Jacob Hansen, Stig Møller, et al. Schütte House price bubbles under the COVID-19 pandemic. Journal of Empirical Finance, 2023, 101462.

[4] Sun Wei, Zhao Yu. Research on the Impact of Income Level and Population Scale on Real Estate Prices. Journal of Yuncheng University, 2023, 41(02): 1-9.

[5] Theurillat Thierry, et al. The Real Estate Markets: Players, Institutions and Territories. Urban Studies, 2015, 52(8): 1414-1433.

[6] Xia Yikai. Study on the Mechanism and Effect of Financial Leverage on Real Estate Prices. Southeast University, 2020.

[7] Yu Hanjie. Research on the Impact of Macroeconomic Policy on Real Estate Prices. Jilin University, 2022.

[8] Zhang Lulu, Ma Xiaomin, Wang Xingyue, et al. Application Research on Housing Price Forecasting Based on PCA-BPNN Algorithm. Journal of Changchun University of Science and Technology (Natural Science Edition), 2024, 25(02): 114-118.

[9] Zhangyu. Loop line of urban rail transit impact on real estate prices research. Chongqing university, 2020.

[10] Zhao Weikai, Yang Lanjun, Dai Lin, et al. Statistical Inference and Application of Multivariate Linear Regression Model under Skew-Normal Conditions. Applied Mathematics, 2024, 37(02): 519-529.