ISSN 2959-6157

Research on the Factors Affecting House Price using Multiple Linear Regression Model

Jully Ye[1, *]

[1]Merchiston international school, Shenzhen, 518000, China
*Corresponding author: zhuohang.ye@merchiston.cn

Abstract:
This article aims to identify the factors that affect the change in house prices. In this paper, the Multiple Linear Regression Model is used to analyze the factors with 1000 random samples from the USA which was collected from the 2nd of May in 2014 to the 10th of July in 2014. Based on the datasets, 13 variables (bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft-above, sqft_basement, yr_built, yr_renovated, street, city) are selected as independent variables, and used in scatter diagram to show the correlation with the dependent variable price. Turns out that 3 variables have not shown a significant correlation with the price (yr_built, yr_renovate, city). To find which of the factors have the most significant impact on the change in house prices, VIF value is used to compare the collinearity of those 10 variables. Finally turns out that only 5 variables (bedrooms, view, condition, waterfront, sqft_lot) show the most significant effect on the price of housing. Overall, the main factors affecting the price of housing in the USA were the number of bedrooms and views next to the house, as well as the conditions needed for buying the house.

Keywords: Price of housing; multiple linear regression model; VIF value.

1. Introduction

Lots of literature illustrates that the demand and supply of housing is the main reason why the price of housing differed. From the demand side, as the population and disposable income increase, development in infrastructure and education will all lead to a rise in the price of housing. From the supply side, as the supply of land was limited the number of housings that could be built was also limited this led to a higher cost of production of the house, therefore the price of housing will increase.

Nowadays the price of housing is affected by different factors both internal and external [1]. Internally the area and the infrastructure are some of the factors which is affecting the price of the housing. Externally the house's location, government policy, or even the human population will affect the house prices [2]. Chang concluded that government policy will directly affect the price of housing, such as the housing purchase restriction policy, which is a good way to control the price of housing by lowering the demand for housing. Chang also concluded that when this policy was introduced the price of housing had a significant decrease in the short run, but in the long run the power of this policy would decrease [3]. Wang purposed that changing the interest rate or borrowing difficulties will also affect the price of housing by influencing the demand of householders [4]. Chen proposed that consumer's prediction of the price of housing in the future will bring a very significant impact on the price of housing now. If the consumer believes that the price of housing will increase in the future, their demand for this house will increase and the price of the housing will be affected [5]. Chen also concluded that the effect of price in housing, due to consumer's house price increase expectations will be stronger in first-tier cities than those second and third-tier cities [6]. Cui concluded that housing prices are also being affected by the uneven distribution of public resources such as hospitals and schools, usually, a house near these places will have a higher price because it is more convenient, and the transportation around the house will bring a huge impact on the price of housing as well [7]. Luo found out from her research that the southern part of Xi'an has more population, and the economic development was also better, therefore it has a more house price in the southern part. However, in the center of Xi'an, the price of housing was much lower because it was kind of like an old town in the center. She then concludes that house age is also a very important factor that affects the prices of houses, and another factor is the convenience of the basic public infrastructure [8]. Li's research found that the price of second-handed houses in Wuhan decreased due to the Covid-19 pandemic, from January to February 2020 there was no trade happening in the second-handed house market until October, but the price had decreased. After the pandemic people in Wuhan considered to be more likely to buy houses near the hospital [9]. In the past houses located in the urban usually had a higher price than the suburban, but as rural tourism became more popular, more bed and breakfast hotels were built, and the prices of those houses were usually quite high. Wang found out that the price of the bed and breakfast hotel was based on the type of scenery and the environment around the hotel [10].

Overall, there are lots of different factors that are affecting the change in the pricing of houses, such as populations, locations, government interventions, natural disasters, and transportation. However, those factors can change the price of houses due to the influence on the demand and supply of housing in the market. If the factors increase the demand for the housing the price will increase, if not the price will decrease. If the supply in the market is limited the price will increase, if the supply is unlimited then the price will decrease. This research aims to use a multiple linear regression method to find out which factors are affecting the demand and supply of housing the most.

2. Methods

2.1 Data Source

The dataset used in this paper is found form a post on the website called Kaggle. This dataset was posted in 2020, the data contains some samples of house prices in the USA in US dollars, there are a total of 4600 samples collected from the 2nd of May in 2014 to the 10th of July in 2014.

2.2 Variable Selection

There is the majority of samples in the original dataset that require lots of analysis, therefore in this literature, there will be only 1000 randomly selected samples. The data selected contain 13 variables (bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft-above, sqft_basement, yr_built, yr_renovated, street, city) and an independent variable (price). The variables will be shown in more detailed in Table 1 below.

Table 1. Variable introduction

Variable	Meaning	Range
Bathrooms	The number of bathrooms	0-8
Sqft_living	The living area of the home (square feet)	370-13.5k
Sqft_lot	Total area of a piece of land (square feet)	638-1.07m
floors	How many floors does the house has	1-3.5
waterfront	Is the house located near the waterfront	0-1
view	Good views around the house	0-4
condition	How many conditions needed for buying the house	1-5
Sqft_above	Total area above ground level (square feet)	370-9410
Sqft_basement	Total area below ground level (square feet)	0-4820
Yr_built	Year built	1900-2014
Yr_renovated	Year renovated	0-2014
City	Located in which city in USA	Seattle (34%), Renton (6%) other(59%)
Price	Prices of housing in USA	USA

2.3 Method Introduction

This literature used the multiple linear regression method to model the dataset and to find out which factors are affecting consumer demand the most. Multiple linear regression is a regression model that aims to estimate the relationship between variables in situations where there are several independent variables. The independent variables can either be continuous or qualitative, however, the dependent variable must be measured on a continuous scale.

3. Results and Discusion

3.1 Multiple Linear Regression

From Figure 1, the graph shows that the linear equation $price = 129284.552 + 124285.401 * bedrooms$, $R^2 = 0.040$. From this equation and the line, the price of housing increases by 124285.401 per bedroom.

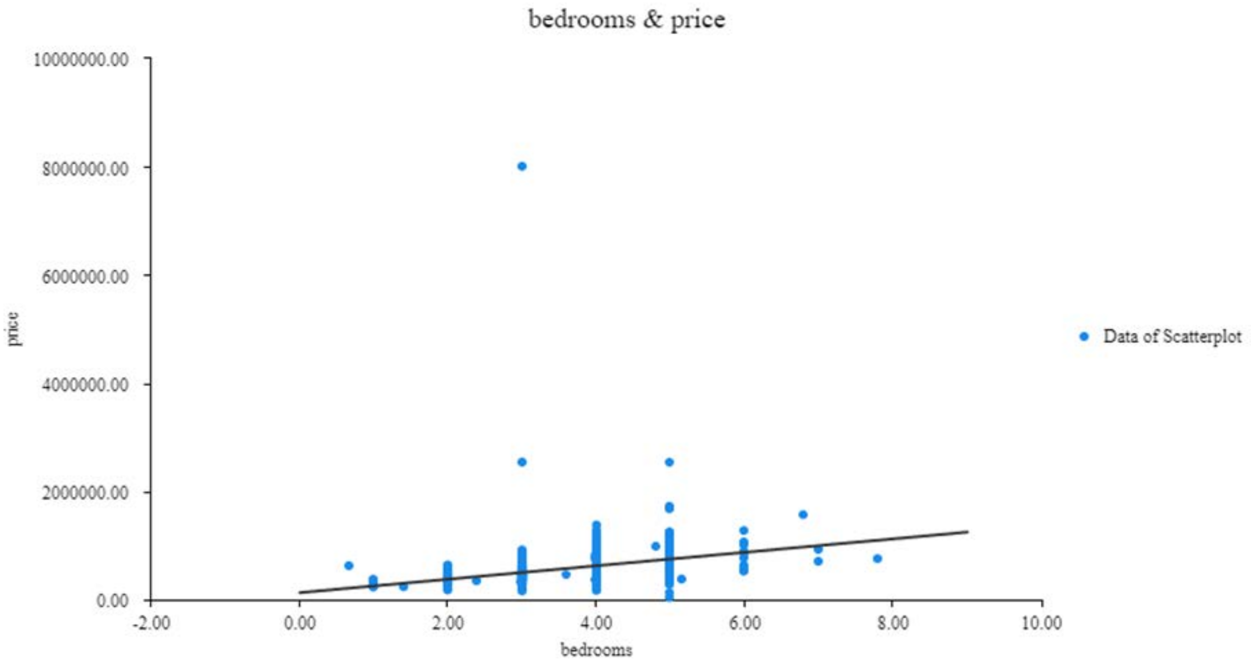

Fig. 1 Scatter of bedrooms & price

From Figure 2, the graph shows that the linear equation $price = 43489.433 + 235315.612 * bathrooms$, R^2 = 0.107. From this equation and the line, the price of housing increases by 235315.612 per bathroom.

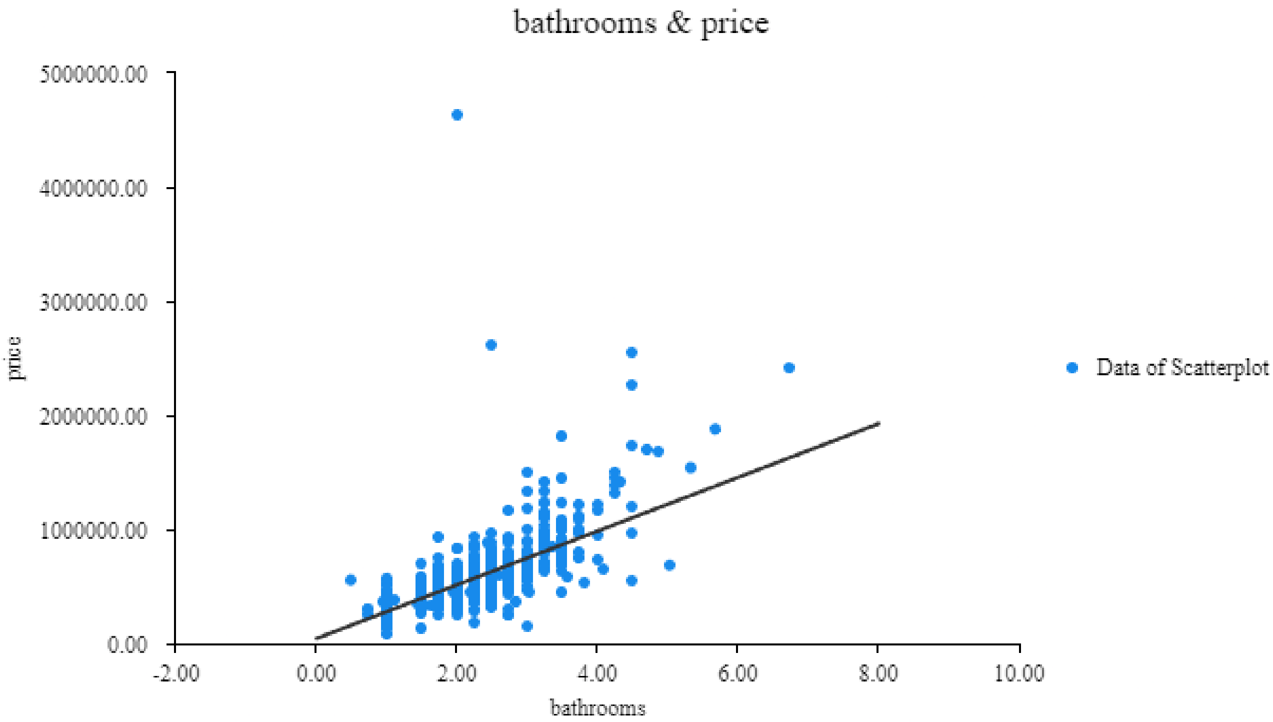

Fig. 2 Scatter of bathrooms & price

From Figure 3, the graph shows that the linear equation $price = 13099.748 + 251.911 * sqft\_living$, R^2 = 0.185.

From this equation and the line, the price of housing increases by 251.911 per sqft_living.

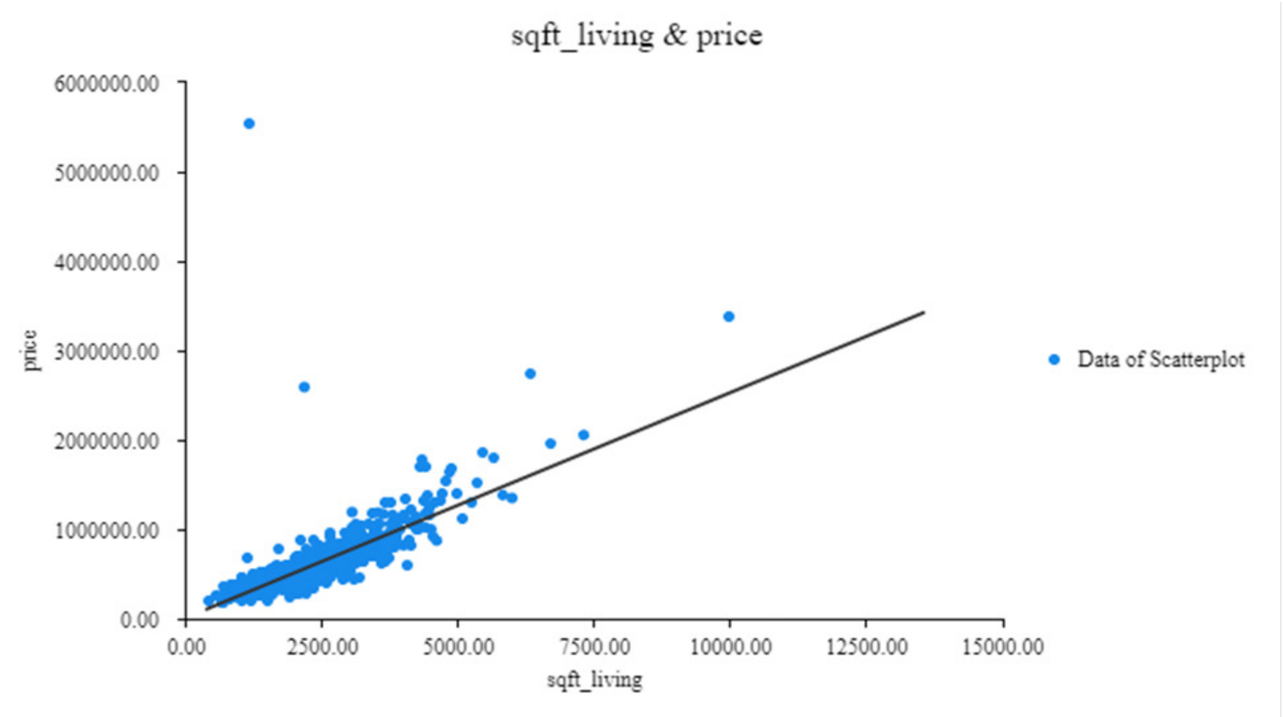

Fig. 3 Scatter of sqft_living & price

From Figure 4, the graph shows that the linear equation $price = -264913.903 + 414.493 * yr\_built$, $R^2 = 0.000$. From this equation and the line, the price of housing increases by 414.493 per yr_built. However, the R^2 shows that there is no correlation between yr_built and price at all.

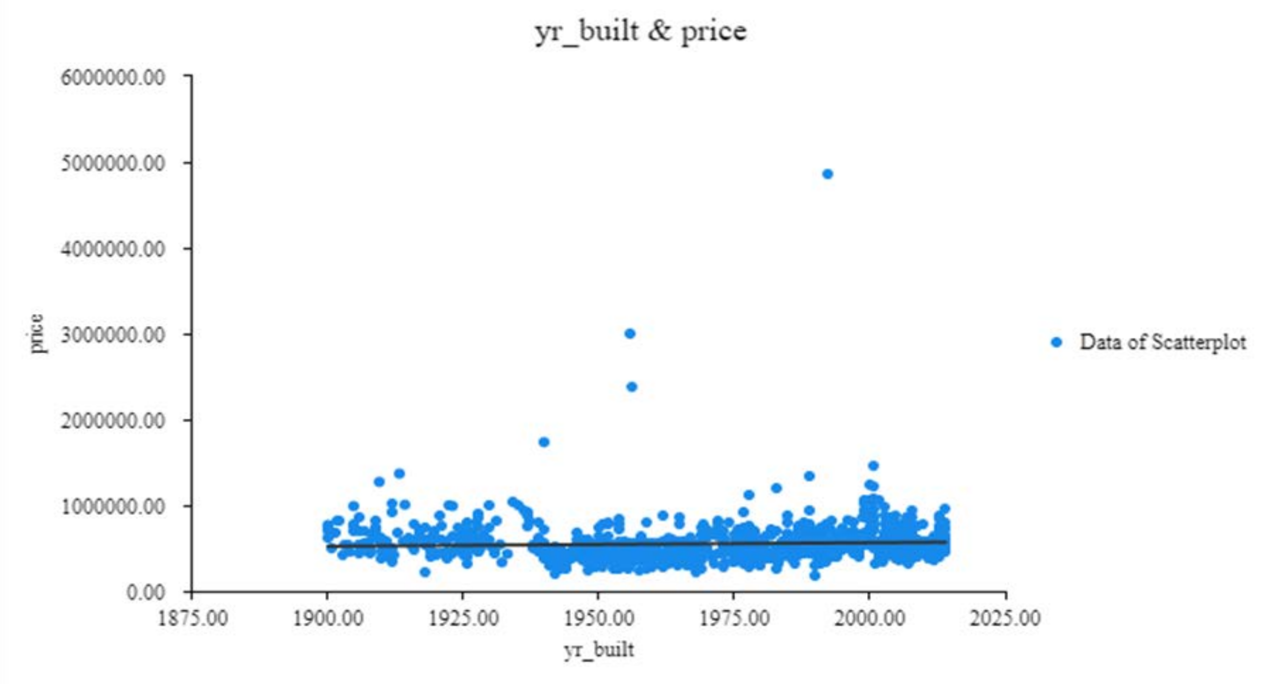

Fig. 4 Scatter of yr_built & price

Figure 5, the y axis is the price of the house, and the x axis are the 13 variables that the author believes was affecting the housing price.

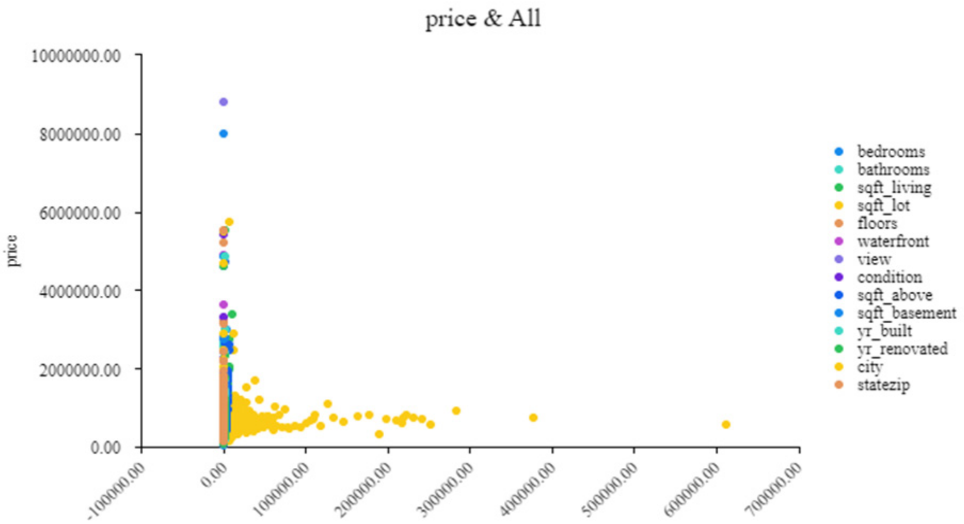

Fig. 5 Scatter of all & price

Table 2 illustrates the correlation between the price and the other 13 variables, this paper uses the Pearson correlation coefficient to show how strong the relationship between price and the 13 variables is. The correlation coefficient shows that 12 variables have a positive correlation with the price but only 1 variable the yr_renovated shows a negative correlation with the price. Even though 12 variables have a positive correlation 3 variables bathroom number, sqft_living, and sqft_above show the most significant positive correlation with the price. This means those 3 variables have more influence on the housing price. The coefficient between price and yr_built is 0.022, which is close to 0 and the value of p is 0.138 > 0.05, therefore price and yr_built do not correlate. The coefficient between price and yr_renovated is -0.029, which is close to 0 and the value of p is 0.051 > 0.05, therefore price and yr_renovated do not correlate. The coefficient between price and city is 0.019, which is close to 0 and has a p-value of 0.207 > 0.05, therefore there is no correlation between price and city. From the scatter and the Pearson correlation coefficient, this paper can conclude that in 13 variables, there are 3 which do not correlate with the price, so these 3 variables will be removed from the calculation of the multiple linear regression method. The multiple linear regression will be carried out by 1 dependent variable (price as Y) and 10 dependent variables (bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement as X).

Table 2. Pearson correlation coefficient between Price & 13 variables

Variable	Price
Bedrooms	0.200**
Bathrooms	0.327**
Sqft_living	0.430**
Sqft_lot	0.050**
Floors	0.151**
Waterfront	0.136**
View	0.229**
Condition	0.035*
Sqft_above	0.368**
Sqft_basement	0.210**
Yr_built	0.022
Yr_renovated	-0.029
City	0.019

3.2 Model Results

F is a variable use for Significance Testing, this variable can be used to check if a multiple linear regression formula exists. From Table 3, Variable F = 120.347, and the p-value is smaller than 0.05. Therefore, this multiple linear regression passes the Significance Test, and this means the regression model is meaningful, concluding that there will be at least one X that will affect the dependent variable Y.

Regression coefficient significance testing is to test the significance of the effect of the dependent variable with different independent variables and is recorded by the variable t. All t variables have a p value respectively, for bedrooms, conditions, sqft_lot, waterfront, and view have a p value smaller than 0.05 therefore these 5 X variables show significance on the dependent Y variable.

Table 3. Parameter estimate

	Unstandardized coefficients	Unstandardized coefficients	Standardized coefficients			
	B	Std. Error	Beta	t	p	VIF
Constant	-86999.716	54996.395	-	-1.582	0.114	-
Sqft_living	7.672	346.733	0.013	0.022	0.982	2032.111
Bedrooms	-51913.966	10524.536	-0.084	-4.933	0.000**	1.667
Bathrooms	14063.601	16153.095	0.020	0.871	0.384	2.921
Conditions	56765.807	11558.935	0.068	4.911	0.000**	1.117
Floors	18856.947	18514.967	0.018	1.018	0.309	1.810
Sqft_basement	248.659	346.364	0.205	0.718	0.473	470.889
Sqft_lot	-0.723	0.214	-0.046	-3.379	0.001**	1.074
Waterfront	368590.439	94345.309	0.055	3.907	0.000**	1.155
View	54338.731	10976.504	0.075	4.950	0.000**	1.330
Sqft_above	252.660	347.071	0.386	0.728	0.467	1631.479

3.3 Model Evaluation

VIF value is used to judge collinearity, which refers to the phenomenon that independent variables are correlated with each other in linear regression analysis. The reason for collinearity may be that there is a strong correlation between multiple independent variables, the sample size is insufficient, or the incorrect use of virtual variables in regression analysis may lead to the occurrence of collinearity problems. If the VIF value is greater than 5 this means that there has been strong collinearity. Sqft_living, sqft_above, and sqft_basement all have a VIF value much larger than 5 which means there is very strong collinearity.

R^2 is used to analyze the goodness of fit of the model, also known as the coefficient of determination. R^2 lies between 0-1. Adj R^2 has no real meaning but is usually used in model adjustments. In this model, R^2 is 0.208 which means that the independent variables can explain 20.8% of the price change (Table 4).

Table 4. Parameter estimate 2

Critial	Value
R^2	0.208
Adj R^2	0.206
F	F (10,4589) =120.347, p=0.000
D-W value	1.968

One of the basic assumptions of the multiple linear regression model is that the random interference terms of the model are independent or uncorrelated. The random interference term is the error caused by the uncertainty of the data itself. If the random interference terms of the model violate the basic assumption of mutual independence, it is called autocorrelation. Autocorrelation can be analyzed using the D-W test. Normally think that if the D-W value is close to 2 (lie between 1.7-2.3), means there is no autocorrelation. If not close to 2 means, there is an autocorrelation. The D-W value is 1.968 which is very close to 2, which means that there is barely any autocorrelation.

The comparison of the influence of independent variables on dependent variables is done through standardized regression coefficients. The larger the absolute value of the standardized regression coefficient, the greater the influence of the independent variable on the dependent variable. The standardized regression coefficient is the regression coefficient obtained after standardizing the independent variable and the dependent variable at the same time. After the data is standardized, the influence of differences in dimension and order of magnitude is eliminated, making different variables comparable. Therefore, the standardized regression coefficient is used to compare the influence of different independent variables on the dependent variable.

Bedrooms influence the price most, the view comes second, the condition comes third, the waterfront comes next, and the sqft_lot influences the price the least. Condition, waterfront, and view will have a significant positive impact on price. As well as bedrooms, sqft_lot will have a significant negative impact on price. However, sqft_living, bathrooms, floors, sqft_basement, and sqft_above will not affect the price.

4. Conclusion

The study selected 1000 random samples from the dataset, which have 10 independent variables and 1 dependent variable. The method (Multiple linear regression analysis) is effective and useful. Because the D-W value is very close to 2 which is 1.968 this means that the model is meaningful.

At the first stage of analysis, the study found that 3 variables have p-values greater than 0.05 this means these 3 variables do not show a correlation with the dependent variable (price), therefore the study is not considering these 3 variables. The study concludes from the model that in the 10 variables there are 5 variables not showing significance on the influence of house prices, only bedrooms, view, condition, waterfront, and sqft_lot show significance on the change in house price.

Overall, bedrooms influence the price most, the view comes second, the condition comes third, the waterfront comes next, and the sqft_lot influences the price the least. From the research, the main factors affecting the price of housing were the number of bedrooms and views next to the house, as well as the conditions needed for buying the house. The number of bathrooms and floors, sqft_basement, sqft_living, and sqft_above have nearly no effect on the house pricing. This means the bedrooms and views are what affect consumers' demand the most, and the other factors do not affect their demand at all.

References

[1] Zhenjiang Shen. Research on the Fluctuation Mechanism and System Simulation of Commodity Housing Market. Times Finance, 2009.

[2] Yao Zhang, Jintao Li. Analysis on the impact of housing purchase restriction policy on housing prices: Taking Wuhan urban circle as an example. Advances in Applied Mathematics, 2023, 2-6.

[3] Yu Chang. Analysis on the impact of housing purchase restriction policy on housing prices in Hunan Province. Journal of Banking and Finance, 2021.

[4] Weihua Wang. Analysis of the impact of mortgage interest rate multiples on housing prices. Journal of Banking and Finance, 2020, 14-18.

[5] Xiaoliang Chen, Kan Chen, Zhaorui Wang, Zhengyan Xiao. Research on the factors affecting the differentiation of housing prices among cities. Advances in Applied Mathematics, 2024.

[6] Xiaoliang Chen, Shuo Cheng, Kan Chen, Zhengyan Xiao. Research on factors affecting housing prices in first-tier cities based on machine learning methods. Journal of Banking and Finance, 2023, 22-24.

[7] Qinyu Cui, Yiting Huang, Chang Liu, Yu Chen. Research on the influencing factors of housing prices in Shenzhen based on multi-scale geographically weighted regression. Times Finance, 2023.

[8] Lin Luo, Xiping Yang, Jiguo Li. Influencing factors and spatial heterogeneity of second-hand housing prices in Xi'an. Advances in Applied Mathematics, 2023, 6.

[9] Yaqin Li. Analysis of the impact of the COVID-19 pandemic on second-hand housing prices in Wuhan. Journal of Banking and Finance, 2022, 9.

[10] Chaohui Wang, Simeng Li, Haohao Qiao, Yang Gao. Spatial heterogeneity characteristics of rural homestay prices and its influencing factors. Times Finance, 2023.