The Study on the Influencing Factors of Housing Price

Zhiyi Chen
School of Beijing Haidian KaiWen Academy, Beijing, 100000, China

Keywords: House Price, Multiple Linear Regression, House Sales.

Abstract: This paper aims to find out the factors that affect the housing price. Multiple Linear Regression method was used to analyze the significant factors between May 2014 and May 2015 in King County. Through the graph test, 13 of the 15 variables selected are related to house prices. In the model analysis, the p values and the VIF values of these variables were compared, and the accuracy of the model was confirmed. The results showed that the number of bedrooms, number of bathrooms, living area, waterfront, view level, overall condition, building grade, above ground area, below ground area, year of renovation, there is a significant linear relationship between the living area and lot size of neighbors and the housing price, and there is no significance test between the number of floors of the building and the housing price. In general, the fluctuation of home prices in King County can be considered by the extent to which these factors affect home prices.

1. INTRODUCTION

House is an important part of people's life, and its price has always been a concern of people. Since 2002, house prices in the United States have continued to rise, which have a significant impact on people's lives. According to the research of Case and Shiller, the real estate market has obvious fluctuations, and the average house price will soar by about 20% from 2000 to 2020 (Mian and Amir, 2024). Nowadays, the house prices continue to change and rise, according to the latest S&P Case-Shiller Index, the average home price in the United States increased by 1% year over year in July 2023, hit a record high. Meanwhile, the cumulative 5.3% increase in the S&P Case-Shiller average home price index from January 2023 has offset the cumulative 5% decline in home prices from the peak last June to the slowdown in January this year (Feng, 2023). The fluctuation of housing price has a great impact on the consumption structure of urban residents (Xu and Liu, 2024). While choosing the house for a suitable price, it is also important to consider the conditions of the house that affect the price of the house. Therefore, this paper will analyze the structural direct factors that have the greatest impact on home prices and provides advice on what to expect to buy, based on the home price data of King County, Seattle.

The real estate market is a complex model with many factors, in addition to the general condition of the house and the grade of construction, there are many factors related to the interior of the house. By investigating the demand for housing in various regions, the main influencing factors are found and determined. Quitzau's study in 2008 showed that people are very concerned about the number of bedrooms and bathrooms in the house (Quitzau and Røpke, 2008). Khajehzadeh et al. also showed that since the concept of bedroom size may be different from region to region, living area should also be a priority for people (Khajehzadeh and Brenda, 2017). Jim's research found that the number of floors, floor height and window orientation (view level) of the house also deeply affect people's purchase intention, and thus affect the price (Jim and Chen, 2006). Water source is the most commonly used and typical housing characteristic. In North's research, the availability of high-quality and available water resources also plays a significant role in real estate valuation (North and Griffin, 1993). In the study of Zabel et al. and Cho et al., the effect of minimum lot size restrictions (MLRs) in the housing price model is also very large in the value of real estate, so this is also considered to be one of the main influencing factors of housing price (Jeffrey and Maurice, 2011 & Cho et al., 2009).

As mentioned in the research of Liu Wenjing et al., owners have the demand for the basement, thus, the balance between basement and above-ground floor area is also important for housing. This also

Chen, Z.
The Study on the Influencing Factors of Housing Price.
DOI: 10.5220/0013055900004601

In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)*, pages 258-263
ISBN: 978-989-758-722-1

provides a new way of thinking about the factors of housing prices, that is, the above-ground area and the underground area (Liu et al., 2022). At the same time, Cho's research model also showed that from the perspective of low-carbon green buildings, there is also a correlation between the property value caused by renovation and the house price (Cho et al., 2020).

This paper will mainly study and analyze the impact of 15 variables (house condition (overall), building grade, bedroom numbers, bathroom numbers, living area, number of floors, view level, whether there is water source, lot size, basement size, living area above the ground, year of construction, year of renovation, neighborhood living condition) on the housing price, and build a suitable housing price model to provide people with better purchase suggestions, the variables may adjust with the dataset selected.

2. METHODS

2.1 Data Source

The data used of this analysis is from a public data set of Kaggle.com, the title of the dataset is “House Sales in King County, USA”. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The original dataset is in .csv format.

2.2 Variable Introduction

The dataset includes 21613 records, each of them represent a sale of a house. There are 21 columns in the dataset, as table 1 shows:

Table 1. List of variables.

Variable name	meaning
Id	The unique identifier for the house
Date	the date when the house was sold
Price	the final sale price of the house
Bedroom	Number of bedrooms
Bathrooms	Number of bathrooms
sqft_living	Living area (in square feet)
sqft_lot	Lot area (in square feet)
floors	Number of floors
waterfront	Whether there is water (0 means no, 1 means yes)
view	The rate of view of the house
condition	the living area (square feet) above ground level
grade	The area of the basement (in square feet)
sqft_above	year the house was built
sqft_basement	The area of the basement (in square feet)
yr_built	year the house was built
yr_renovated	The year in which the house was last renovated
zipcode	zipcode
lat	Latitude coordinates of the house
long	the longitude coordinates of the house
sqft_living15	Living area (square feet) of 15 neighbors in the vicinity of the house in 2015.
sqft_lot15	sqft_lot15: Lot size (square feet) of 15 neighbors near the house in 2015.

In the data set, abnormal values are identified and found based on the box plot, in this scenario, the author finds the data with extremely high price and extremely low price. Figure 1 shows that there is some extremely high price data recorded. Now this paper back to the dataset and try to find the reason for those abnormal value. By observing those abnormal value, the author finds that extremely high housing prices are generally accompanied by extremely high building levels, living area, lot area, condition level, or view level. These variables are considered in the dataset, those are not outliers due to factors that are not considered, so this paper chooses to keep these data.

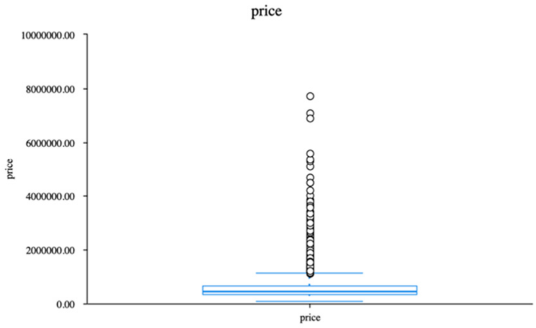

Fig. 1 Box plot of the price of the data.

2.3 Method Introduction

A multiple linear regression model is used to analysis and predict house price based on the selected independent variables. The regression model equation includes the intercept term and coefficients of each variable. The multiple linear regression model is a model with multiple explanatory variables, and it can be used to explain the relationship between dependent variable and multiple independent variables. The principle is to apply least squares on the data and minimize the sum of the squares of the residuals between the dependent variables and the independent variables.

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis

The dataset contains a large amount of data, which there is some of the variables in of the column has no effect, or not a direct effect, on the house prices that. Using scatter diagram of each independent variable to see how they related to the dependent variable.

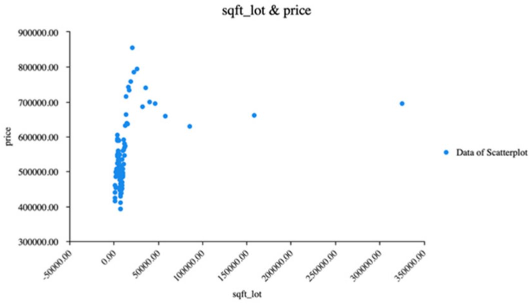

Fig. 2 Scatter Digram of sqft_lot & price.

Figure 2 shows that there is no linear relationship between the lot size and the housing price. In order to obtain better model results, it has been decided to remove this variable.

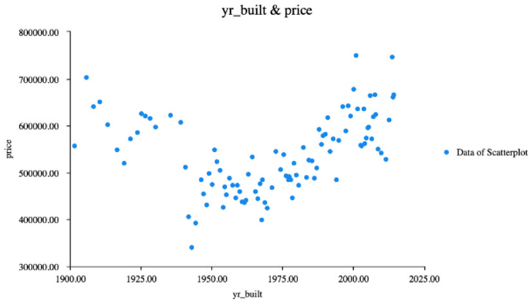

Fig. 3 Scatter Digram of yr_built & price.

Figure 3 shows that there is no linear relationship between the year the house was built and the housing price, it has been decided to remove this variable.

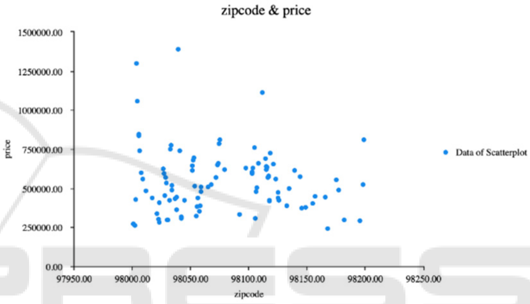

Fig. 4 Scatter Digram of zipcode & price.

Figure 4 shows that there is no linear relationship between the zipcode of the house and the housing price, it has been decided to remove this variable.

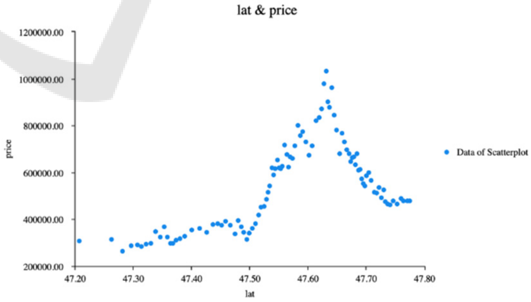

Fig. 5 Scatter Digram of lat & price.

Figure 5 shows that there is no linear relationship between the latitude of the house and the housing price, it has been decided to remove this variable.

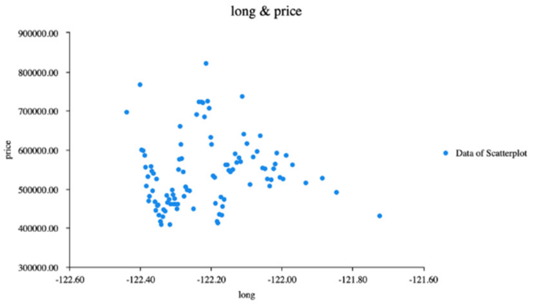

Fig. 6 Scatter Digram of long & price.

Figure 6 shows that there is no linear relationship between the longitude of the house and the housing price, it has been decided to remove this variable. After removing those variables, the features remaining that might influence house prices as explanatory variables including (Table 2):

3.2 Model Results

When applying multiple linear regression with the tool, the author finds that some independent variables have multicollinearity, which will adversely affect the stability and interpretability of the model and may lead to inaccurate coefficient estimation and increased standard error. First of all, a collinearity analysis of the data is carried out:

On the table 3 shown, the author found that the absolute value of correlation coefficient between variable sqft_above and sqft_living is very strong (>0.8), this means there may multicollinearity between them. Thus, this paper chooses to remove sqft_living in the linear regression, since it has high absolute value of correlation coefficient with some other variables. Finally, the multiple linear regression model is applied.

Table 2. Explanatory variables.

Variable name	Meaning	Range
Bedrooms	Number of bedrooms	no
Bathrooms	Number of bathrooms	no
sqft_living	Living area	no
Floors	Number of floors	no
Waterfront	Whether there is water	1 or 0
View	The rate of view of the house	0~4
Condition	The overall rate of condition of the house	1~5
Grade	The building grade of the house	1~13
Sqft_above	The living area above ground level	no
Sqft_basement	The area of the basement	no
yr_renovated	The year the house was last renovated	no
Sqft_living15	Living area (square feet) of 15 neighbors in the vicinity of the house in 2015	no
Spft_lot15	Lot size (square feet) of 15 neighbors near the house in 2015	no

Table 3. Correlation results of part of the variables.

	bedrooms	bathrooms	sqft_living	floors	waterfront	sqft_lot15	view	condition	grade
bedrooms	1								
bathrooms	0.516**	1							
sqft_living	0.577**	0.755**	1						
floors	0.175**	0.501**	0.354**	1					
waterfront	-0.007	0.064**	0.104**	0.024**	1				
sqft_lot15	0.029**	0.087**	0.183**	-0.011	0.031**	1			
view	0.080**	0.188**	0.285**	0.029**	0.402**	0.073**	1		
condition	0.028**	-0.125**	-0.059**	-0.264**	0.017*	-0.003	0.046**	1	
grade	0.357**	0.665**	0.763**	0.458**	0.083**	0.119**	0.251**	-0.145**	1

3.3 Model Results

The table 4 shows the result of the linear regression. The model formula is:

$$price = -717503.880 - 34136.190 * bedrooms - \cdots + 565696.394 * waterfront \quad (1)$$

Where R^2 of the model is 0.613. this means the variables selected can be used to explain 61.3% of the house price variation. The VIF of the variables in this model are all less than 5, means that there is no collinearity problem.

The p value of the independent variables: bedrooms, bathrooms, sqft_living, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_renovated, sqft_living15, sqft_lot15 are 0.00, so it can be considered that all variables except floors have a significant effect on the dependent variable price. Moreover, D-W value of the model is close to number 2, which indicates that there is no autocorrelation in the model, and there is no correlation between the sample data, and the model is good.

The result shows that sqft_above, yr_renovated, sqft_basement, grade, condition, sqft_living15, view, waterfront has a significant positive impact on price. And bedrooms, bathrooms, sqft_lot15 have significant negative influence on price. But floors have no effect on price.

Table 4. Model results.

	Unstandardized Coe.		Standardized Coe.	t	p	collinearity diagnostics	
	B	SE	Beta			VIF	tolerability
Constant	-717503.880	17314.851	-	-41.439	0.000**	-	-
bedrooms	-34136.190	2138.522	-0.086	-15.963	0.000**	1.639	0.610
bathrooms	-16254.218	3453.362	-0.034	-4.707	0.000**	2.931	0.341
floors	-2324.943	3948.933	-0.003	-0.589	0.556	1.884	0.531
sqft_above	182.790	4.086	0.412	44.740	0.000**	4.742	0.211
yr_renovated	72.552	3.923	0.079	18.494	0.000**	1.029	0.972
sqft_basement	210.271	4.848	0.253	43.370	0.000**	1.907	0.524
grade	102110.393	2348.118	0.327	43.486	0.000**	3.156	0.317
condition	57357.875	2515.927	0.102	22.798	0.000**	1.111	0.900
sqft_lot15	-0.722	0.059	-0.054	-12.245	0.000**	1.073	0.932
sqft_living15	15.315	3.795	0.029	4.036	0.000**	2.802	0.357
view	56802.133	2384.400	0.119	23.822	0.000**	1.383	0.723
waterfront	565696.394	19691.576	0.133	28.728	0.000**	1.202	0.832
R2	0.613						
Adj R2	0.613						
F	F (12,21600)=2853.269,p=0.000						
D-W value	1.983						

Dependent Variable: price

* p<0.05 ** p<0.01

4. CONCLUSION

This study selected the dataset of home prices sold in King County which between May 2014 and May 2015. The study focused on 15 variables, two of which were not linearly related to the dependent variables in the selected data set and were removed from the analysis for better modeling. The Multiple Linear Regression accurately and effectively analyzes the correlation coefficient of multiple factors to the dependent variable.

This study analyzes the how the selected 13 variables related to housing price through the method of multiple linear regression and finds that the main factors affecting housing price are waterfront, building grade, overall condition, and view level. The number of floors in the variable has little effect on prices. Above and below ground living area, year of renovation, building grade, overall condition, neighbors living area, view level, waterfront has a significant positive impact on price, and then the number of bedrooms, the number of bathrooms, the lot size of the neighbors have significant negative influence on price.

This research can provide reference value for people are considering buying a house and help them to determine the initial budget from different factors perspective. However, this model still has shortcomings, such as other variables that are not considered, the relationship between some variables may not be represented, and the analysis sample scope is limited, and so on. To improve these problems, a larger range of data and a more optimized approach are needed.

REFERENCES

Mian Atif, Amir Sufi. HOUSE PRICE GAINS AND U.S. HOUSEHOLD SPENDIN. Working Paper, 2024, 2.

Feng Difan. American House Prices Have Hit Record Highs. China Academic Journal Electronic Publishing House, 2023, 1.

Xu Miao, Liu Xiaoying. The Impact of Housing Price Fluctuation on the Upgrading of Consumption Structure of Urban Residents: An Analysis Based on the Central Region. School of Management, Zhengzhou University, 2024, 44.

Quitzau M B, Røpke I. The construction of normal expectations: Consumption drivers for the Danish bathroom boom. Journal of Industrial Ecology, 2008, 186-206.

Khajehzadeh Iman, Brenda Vale. Estimating the Floor Area of a House Knowing Its Number of Rooms and How These Are Named, 2017, 315-323.

Jim C Y, Wendy Y. Chen, Impacts of urban environmental elements on residential housing prices in Guangzhou (China), Landscape and Urban Planning, 2006, 78: 422-434.

North J H, Griffin C. Water Sources as a Housing Charactersitic: Hedonic Valuation and Willingness to Pay for Water. Water Resources Research, 1993, 29: 1923-1929.

Jeffrey Zabel, Maurice Dalton, The impact of minimum lot size regulations on house prices in Eastern Massachusetts, Regional Science and Urban Economics, 2011, 41: 571-583.

Cho Seong Hoon, et al. Spatial and temporal variation in the housing market values of lot size and open space. Land Economics, 2009, 51-73.

Liu Wenjing, et al. The Above-Ground and Underground Areas of Residential EPC Projects Whose Unit Price Exceeds the Limit Are Balanced, 2022. Urban Architecture Space. 2022,29 (S2): 780-782.

Cho K, et al. Model for predicting price change patterns in multi-family houses post renovation work in South Korea. Journal of Asian Architecture and Building Engineering, 2020, 19(3): 230-241.