Effect of Housing Factors on Price in Seattle During the COVID-19 Pandemic

Haoquan Hu *

Shenzhen College of International Education, Shenzhen, 518000, China

* Corresponding author: s22746.hu@stu.scie.com.cn

Abstract. Due to the influence of the COVID-19 pandemic, global economic trends have faced an immense change. The major difficulty experienced by workers, laborers, and households are the tremendous changes in housing prices in every part of the world which are both hard to bear and almost impossible to forecast from current factors. With a particular dataset based on the housing price in Seattle, Oregon, USA. This research will use a multiple linear regression model to predict the housing price through various housing factors and to compare the accuracy of the prediction and the relationship between variables by contrasting the train and test set whilst also constructing a residual graph. The result ends up with a moderate accuracy due to the influence of multicollinearity between the factors, this can be seen through the variance Inflation Factor. The process requires the standardization of the data in order to make the test legitimate.

Keywords: Housing Price, Linear regression, Multicollinearity.

1. Introduction

It is essential for everyone to keep a look out for the prices of properties. As the population of the globe increases, resources are increasingly scarcer due to their finite quantity and that real property is one of those scarce resources [1]. As scarcity rises, the demand for goods would also make increase, which will raise their price [1]. This has already made housing prices extremely nondeterministic [2]. In December 2019, the outbreak of the COVID-19 pandemic brought numerous uncertainties to the economy, one of which was the fluctuations in housing prices. Predicting housing price in this scenario will provide more assistance and answers to both households in terms of microeconomics and stakeholders and some governments in terms of policy and the macroeconomy [3].

Originally, the main obstacle which has perplexed the problem of housing price is overpopulation. However, as the pandemic broke out, not only did it bring the problem of more price fluctuations but a larger degree of uncertainty. Forecasting housing price using known statistics was nearly impossible due to too many undefined factors. The predicted correlation of housing factors such as the number of rooms with price is directly proportional [4,5]. But as the average income over the pandemic decreases, the purchasing power of the population will also fall, hence lowering the demand for housing and a decrease in housing price would present [6]. There are much more factors to be considered but negligible compared to this when forecasting housing patterns.

Therefore, using a dataset that was collected during the pandemic based on housing price and housing factors to build a model can analyze whether using factors to predict housing price and ignoring pandemic-related influences could draw a suitable conclusion [7]. It is important to first process the dataset and visualizes it to get a vague idea of its correlations. Then, there is the construction of a multiple linear regression model is to obtain the relationship of the variables and predict the housing price using this information. Finally, taking the residual of the predicted and actual prices can it determine whether the prediction is trustworthy.

2. Methodology

2.1. Description of dataset

The dataset which will be analyzing is named “HOUSE PRICE PREDICTION - SEATTLE” [7]. The dataset has two csv files, the trained csv is used for building the regression model and obtaining

the relationship. The test csv will be used to compare the difference in prediction and actual at the end.

There are 2016 rows of data collected in this dataset which are recorded from August 2022 to December 2022 in Washing, Seattle, Oregon, USA [7]. There are 8 columns in total, 7 of which can be used as independent variables and price brings the dependent variable. However, two of the columns are the units of other independent variables so it is not possible to build a model with these columns, but it will take into account later.

2.2. Preprocessing and visualizing the dataset

The programming language that is used in building the model in Python.

First of all, the libraries that will be used include Pandas, Matplotlib, Seaborn and various Sklearn models. Pandas will be used for data processing, and matplotlib being used as a graph plotter. Seaborn is used for data visualizing and Sklearn models and algorithms are used to put the data into practice and to draw a conclusion.

After importing the libraries, it is crucial to make adjustments to the dataset. The first step is to drop some irrelevant columns. Columns such as "Zip code" should be dropped although it represents the location and that location can be an influential factor affecting housing price, it is difficult to come up with a method to make a contrast between them so dropping the zip code column is the first step. Then, dropping the units' columns "Size units" and "Lot size units" should be considered as it does not have any effect on the dataset. The only problem which needs to be solved is to either change the units for all rows using the minority units' "Acres", or to just drop all the rows which use acres as their units.

Once the required rows and columns are filtered, the dataset will be ready for building the model and analysis.

The independent variable for this research is "Number of bedrooms", "Number of bathrooms", "Total floor area of housing" and "Total land area occupied" whilst the dependent variable is the housing price: "Price".

Now that the data is preprocessed, visualizing the relationship between the independent variables between each other and the dependent variable is important as it will give a brief understanding of the variables. n or another standard font to avoid font errors.

2.3. Method and procedure

From the regression plots, it is not hard to identify a direct relationship between all of the independent variables and price with "lot_size" being a weak exception of having a very small positive correlation with price. The independent variables are also positively related to each other.

However, the independent variables relating to each other are not a good thing as this will create multicollinearity. Multicollinearity is a factor that prevents datasets from being easily forecasted due to the independent variables being correlated with each other by too much [8]. The high similarity and relatedness of the independent variables create distortion in the dataset. This can be seen through common sense that these independent variables must be highly related to each other as they are highly complementary to each other [9].

This can be proved by considering a test that can be calculated simply which is the Variance inflation factor, the test for the severity of multicollinearity. The higher the number, the more severe multicollinearity is affecting the variable, and a variable with VIF less than 5 is optimum for testing [10]. The factor can be calculated as the reciprocal of the level of tolerance which is

$$\mathrm{VIF} = \frac{1}{(1-\mathrm{R_i})^2} \tag{1}$$

The method to reduce the effect of this is through centering or standardizing the variables by subtracting the mean and dividing it by the standard deviation, more like the standardization of normal distributions as it reduces the multicollinearity of the independent variables [9]. In this research, it is

difficult to do so as the units of some independent variables are not uniform so there would be other uncertainties produced.

Moving on to the model, linear regression is a statistical model which can be used to draw a relationship between various variables and produce an equation containing the variables [11]. Linear regression can be used to create models where forecasting relevant dependent variables is very useful. The result produces an equation in the form of the dependent variable equating to a constant added with multiple independent variables multiplied by a coefficient [12]. This research uses a multiple linear regression which uses more than one factor and obtains an unstandardized equation with a small bias due to explained reasons.

3. Results

3.1. Visualizing the dataset

From Figure. 1, it is simple to see that the independent variables "Number of bedrooms", "Number of bathrooms" and "Total floor area" all have a mutually positive relationship. There is a very weak correlation between the three factors and "Total land area occupied" due to the difference in units used in the dataset for these independent variables.

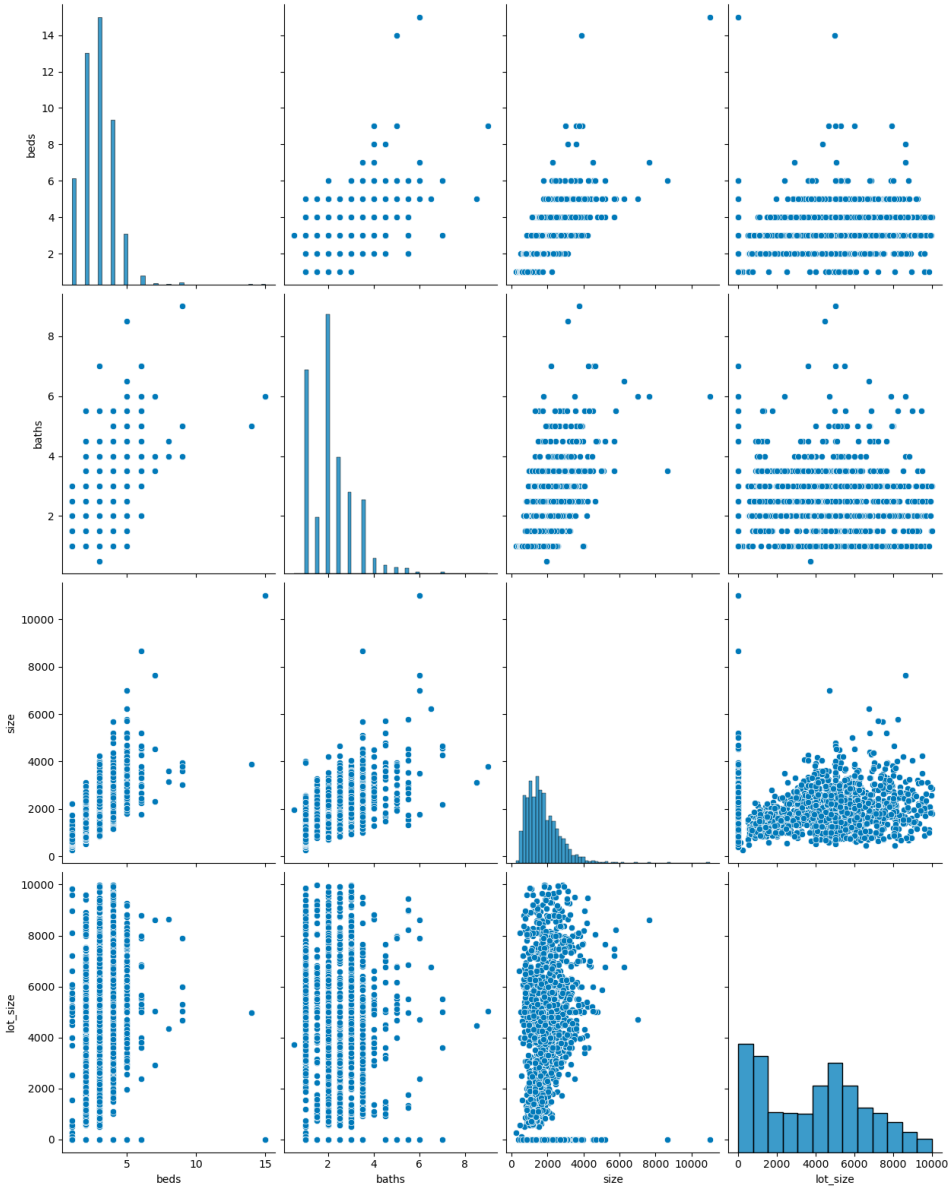

Figure 1. Scatter plot of the relationship between the independent variables (Photo credit: Original).

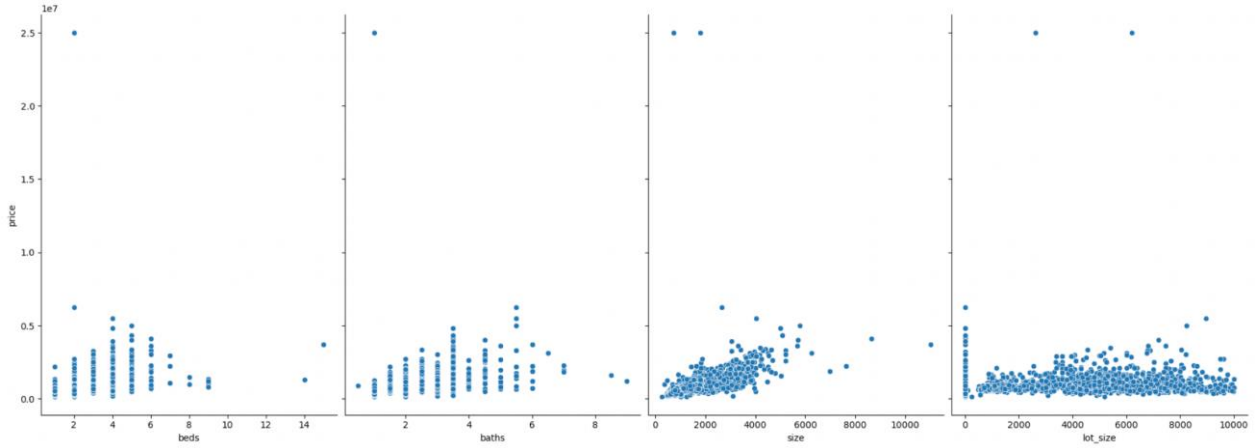

Figure 2. Scatter and regression plot of independent variables and the dependent variable (Photo credit: Original).

It is easy to see a very strong and positive relationship in terms of the first three factors and the dependent variable which is "Price" from Figure. 2. The regression plot in the last graph shows a very weak positively correlated relationship between the independent variable "Total land area occupied" and "Price". This can be easily understood because there could be cramped but luxurious housing which has very expensive pricing. Reversely, spacious but unrefined real estate can have a low price.

3.2. Results for multicollinearity

From the output in Figure. 3, it can be seen that the first three independent variables are highly correlated with each other due to them being complementary to each other. "Lot_size", however, is not as much correlated due to it being random for properties.

	Factor	VIF
0	beds	16.346149
1	baths	10.179328
2	size	13.000589
3	lot_size	3.248462

Figure 3. Variance-inflation-Factor for the independent variables (Photo credit: Original).

3.3. Results of predictions and accuracy

From Figure. 4 and the equation, 292278.0 is the intercept, x1 represents the number of bedrooms in the house, x2 represents the number of bathrooms in the house, x3 represents the total floor area occupied by house, x4 represents the total area occupied by house.

```
292277.95998262265
              Coefficient
beds     -125889.210336
baths      71176.856757
size         510.015872
lot_size       2.693810
439128.1506850319
```

Figure 4. Regression coefficients of the relationships of the independent variables and the dependent variable (Photo credit: Original).

$$Y = 292278.0 - 125889.2x_1 + 71176.9x_2 + 510.0x_3 + 2.7x_4 \quad (2)$$

It can be clearly seen from the coefficients that most independent variables have a positive correlation with housing price with the exception of the number of bedrooms as it is the most affected

by multicollinearity and unstandardized. The total area occupied, as mentioned before has a very weak positive correlation with housing price. The Root Mean Squared Error is the last value in the outputs, it is a measure of how discrete the data is. It is high, again due to the fact that the unit is not the same.

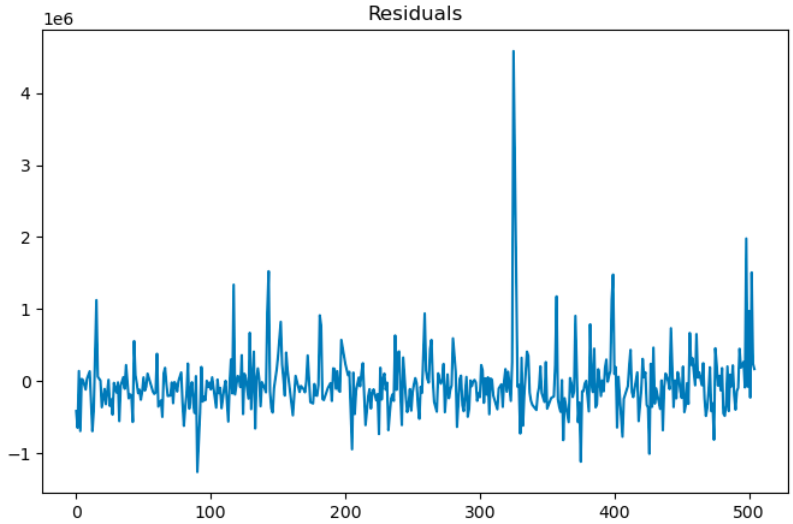

Figure 5. Graph of residuals between predictions and actual (Photo credit: Original).

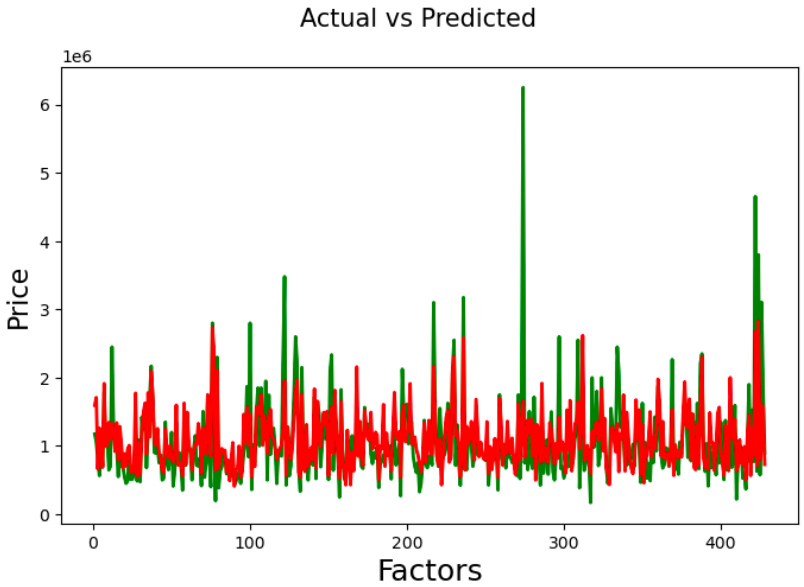

Figure 6. Graph of contrast between predicted and actual pricing (Photo credit: Original).

In Figure. 5 and Figure. 6, the graph plotted by matplotlib is the residual of the actual data, which is from the test csv and the predicted values. It can be seen that there is a large difference between the predicted and actual data. The right plot has a green plot of actual and a red plot of predicted which can see that there is a fairly large difference.

4. Conclusion

From the overall research and analysis, it can be concluded that it is still legitimate to use a property's characteristics to forecast housing price. The main source of error was caused by the units not being the same for various independent variables and a lack of standardization on the dataset. Multiple linear regression may not be the optimum model to predict housing price but is the simplest. Though the COVID-19 pandemic has had a large effect on the patterns of demand in markets, it is still necessary to forecast housing price through the characteristics of the house as there is still

insufficient evidence to reject this. For this research to be more accurate and helpful, it is possible that other models such as logistic regressions or random forest models can be used to improve the credibility and accuracy of the process of forecasting and to draw a firm conclusion on whether housing factors can still be used to predict housing price during other periods of the COVID-19 pandemic.

References

[1] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara, House Price Prediction using Random Forest Machine Learning Technique, Procedia Computer Science, Volume 199, 2022, Pages 806-813, ISSN 1877 - 0509.

[2] Ghosalkar, Nehal N., and Sudhir N. Dhage. "Real estate value prediction using linear regression." 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018.

[3] Piao, Yong, Ansheng Chen, and Zhendong Shang. "Housing price prediction based on CNN." 2019 9th international conference on information science and technology (ICIST). IEEE, 2019.

[4] G. Naga Satish, Ch. V. Raghavendran, M. D. Sugnana Rao, Ch. Srinivasulu, House Price Prediction Using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-9, July 2019, ISSN: 2278 - 3075.

[5] Manasa, J., Radha Gupta, and N. S. Narahari. "Machine learning based predicting house prices using regression techniques." 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020.

[6] Byeonghwa Park, Jae Kwon Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, 2015, Pages 2928 - 2934, ISSN 0957 - 4174.

[7] Samuel Cortinhas, HOUSE PRICE PREDICTION - SEATTLE, 2022, https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle?select=test.csv,https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle?select=train.csv.

[8] Shrestha, Noora. "Detecting multicollinearity in regression analysis." American Journal of Applied Mathematics and Statistics 8.2 (2020): 39 - 42.

[9] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 2017, pp. 2998 - 3000.

[10] Tay, Richard. "Correlation, variance inflation and multicollinearity in regression model." Journal of the Eastern Asia Society for Transportation Studies 12 (2017): 2006 - 2015.

[11] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[12] Diskin, M. H. "Definition and uses of the linear regression model." Water Resources Research 6.6 (1970): 1668 - 1673.