

The Investigation and Prediction of Influencing Factors of Australian Housing Price Based on Machine Learning Models

Chongsen Ma

Department of Banking and Finance, Monash University, Melbourne, Australia

cmaa0049@student.monash.edu

Abstract. With the ever-changing economic environment in recent years, people pay more and more attention to the real estate market, and different house configurations affect house prices. In this paper, house features such as bedroom count and bathroom count are analyzed as factors influencing house prices. Analyze the relationship between influencing factors and real estate prices and predict the trend of housing prices. First, constant values and values that have no impact on the research results should be selected and deleted, and effective influencing factors such as bedroom count and bathroom count should be left. Then, Exploratory Data Analysis (EDA) analysis was carried out on these effective influencing factors, and range, mean, standard deviation and quartiles of the data were read. Then the scatter plot is made to observe whether there is a correlation between the influencing factors and real estate prices. In the part of linear regression, the least squares method was first used to establish Ordinary Least Squares (OLS) model for R square, F-statistic and other calculations, observe the values, and further understand the correlation between influencing factors and housing prices. Through the comprehensive analysis of the above EDA and OLS models, it is concluded that there is a correlation between the influencing factors and real estate prices. Finally, plot is used to plot the linear relationship between them. It can be seen from the image that with the increase of the number of influencing factors, the housing price also increases, and the two are in a positive correlation.

Keywords: Machine Learning, Linear Regression Analysis, House Price Forecast.

1. Introduction

According to the latest report from real estate analysis company SQM Research [1], if the Reserve Bank of Australia pauses to raise interest rates and inflation falls, Australian house prices will recover in 2023. It is worth noting that Australian property prices are expected to rise next year, and the forecast assumes a cash rate of no more than 4%, with Australia's big four banks cautious about a recovery in the housing market. For first home buyers, from July 1, Australian Prime Minister Morrison announced to increase the price cap of the first home loan scheme, the purchase of property only requires a 5% deposit. In this context, it is important to note that these forecasts are influenced by a variety of factors, including objective factors such as the cash rate, inflation rate and unemployment rate, as well as subjective factors such as location type, bedroom count and parking count. Before this, some scholars have conducted Exploratory Data Analysis of similar data and multiple linear regression analysis of housing prices, but they failed to give the trend analysis of influencing factors and real estate prices. For instance, in an earlier study, Graham charted a 230% increase in real house prices between 1991 and 2020 (the time of Australia's last two recessions). Over the same period, the Reserve Bank of Australia's (RBA) cash rate target fell by 5.2 percentage points [2]. Raju used Exploratory Data Analysis (EDA) and other machine learning methods to study the impact of variables such as house age, furniture state, floor area, room size and number of rooms on rental price, showing how machine learning models reveal hidden correlations and provide accurate rental estimates [3]. Truong et al. analyzed the correlation of House prices with other factors such as geographical location, size, and population, and predicted that information other than the House Price Index (HPI) was needed to predict individual house prices. Traditional and advanced machine learning methods are employed to investigate the differences between several advanced models. It will also comprehensively validate multiple techniques in the implementation of regression models and provide optimistic results for house price forecasts [4].

Therefore, this paper will further study the relationship between influencing factors and housing prices. On the premise of collecting relevant data, EDA correlation analysis will be carried out on the data, to obtain reference data such as Range, Mean, Quartiles and Standard Deviation of the research data. The independent variables are screened, and then the selected data is analyzed by linear regression and plotted to visualize the data. Through the observation results, this study will make predictions and give relevant suggestions for future house purchases.

2. Method

2.1. Dataset Preparation

This dataset used in this paper contains information on more than 1000 properties in Australia, including location, size, price and other details.

It can be used to predict house prices in Australia and find the relationship between house prices and characteristics such as home size, number of bedrooms and number of bathrooms. The number of features has 25 in total, such as property type, bedroom count, bathroom count and parking count and so on. Each feature corresponding to the number of data has 1000.

All panel and time series data are initially filtered, consolidated into a single file, and then read. After that, identifying highly similar columns and removing them helps reduce the complexity of the dataset, making the analysis more efficient and easier to understand. For example, since this study focuses on trends in house prices, the data set should retain "Price". In general, the location name is not the primary factor, so delete "location name". During the study, it was found that some data is a constant value, which means that the value has no impact on the study, so it should be deleted. Also, delete other information that has no impact on the study, such as "Phone", "address", "category name", etc. Use a regular expression to find a pattern that matches the price and ignore any text that follows, remove the space from the captured price and return, and then check the unique data type for each column in the data box. Finally, the maximum and minimum values are removed to prepare for the EDA analysis and Linear Regression analysis in the following. The main research object of this paper is the real estate Price, so "Price" is the target variable.

2.2. EDA Correlation Analysis

EDA is a method of analyzing a dataset to summarize its main features, often using statistical graphics and other data visualization methods.

After completing the above data processing, this paper carries out the EDA Correlation process, and obtains the results including "price", "bedroom count", The Range composed of the maximum and minimum values of "bathroom count" and other independent variables, and their Mean, Standard Deviation, Quartiles.

In this paper, the "plot" function is used to make a scatter plot to observe whether there is a Linear relationship between "price" and "bedroom count", to prepare for the following Linear regression analysis.

2.3. Linear Regression

Linear regression is a statistical model that estimates the linear relationship between a scalar response and one or more explanatory variables, also known as dependent and independent variables [5-7]. In linear regression, the relationship is modeled using a linear prediction function whose unknown model parameters are estimated from the data. Such a model is called a linear model. Ordinary Least Squares (OLS) are a linear least squares method for selecting unknown parameters in a linear regression model linear functions affecting a set of explanatory variables.

In the Multiple Linear Regression (MR) Model, the dependent variable - the observed house price - regresses against a set of predictors (or independent variables). The variables selected are considered to contribute to the house price, including bedroom counts, bathroom counts and parking counts, among others [8].

The MR model is a function of variables specified in Equation 1:

$$\text{Sale Price} = f(\text{location number}, \text{zip code}, \text{bedroom count}, \text{bathroom count}, \text{parking count}) \quad (1)$$

In abbreviated form the MR variables are: 1) location number: The number that corresponds to the property's location (Numeric). 2) zip code: The zip code that the property is in (String). 3) bedroom count: The number of bedrooms on the property (Numeric). 4) bathroom count: The number of bathrooms on the property (Numeric). 5) parking count: The number of parking spaces on the property (Numeric). 6) The data is then subjected to OLS regression and linear regression is graphically processed using the "plot" function.

3. Results and Discussion

3.1. EDA Correlation Analysis Results

Fig. 1 shows the correlation analysis for different features based on the EDA method.

	price	location_number	zip_code	bedroom_count	bathroom_count	parking_count
count	734.00	734.00	734.00	734.00	734.00	734.00
mean	524235.35	141522784.92	816.24	2.76	1.69	2.07
std	225427.34	41668389.19	13.67	1.09	0.60	1.47
min	99950.00	108530451.00	800.00	0.00	1.00	0.00
25%	381250.00	138611055.00	800.00	2.00	1.00	1.00
50%	489500.00	138997529.00	812.00	3.00	2.00	2.00
75%	610000.00	139276203.00	830.00	3.00	2.00	2.00
max	1950000.00	700199623.00	839.00	8.00	5.00	12.00

Figure 1. EDA correlation analysis (Photo/Picture credit: Original)

More details about them can be found as below:

1) Price:

Range: The minimum price is \$99950, and the maximum is \$1,950,000. Such a large range shows a huge difference in real estate prices.

Mean: The average price is approximately \$524235.35, Indicating a concentration tendency toward this value.

Quartiles: 50% of the properties are priced between \$381,250 (25th percentile) and \$610,000 (75th percentile).

Standard Deviation: A standard deviation of \$225427.34 indicates the spread or variability in the prices.

2) Location Number:

Range and Mean: These values may correspond to specific location codes or identifiers. A deeper understanding of what this number represents would provide more insights.

3) Zip Code:

Range: Zip codes range from 800 to 839, representing the geographic coverage of the dataset.

Average: The average of zip codes may not provide important insights because zip codes are categorical variables, not continuous variables.

4) Bedroom Count:

Range: From 0 to 8 bedrooms, indicating a variety of property sizes.

Mean: The average number of bedrooms is approximately 2.76.

Standard Deviation: A standard deviation of 1.09 suggests moderate variability in bedroom counts.

5) Bathroom Count:

Range: 1 to 5 bathrooms, indicating diversity in amenities.

Mean: The average number of bathrooms is around 1.70.

Standard Deviation: A standard deviation of 0.60 suggests most properties have 1 or 2 bathrooms.

6) Parking Count:

Range: From 0 to 12 parking spaces, highlighting varied parking facilities.

Mean: The average number of parking spaces is approximately 2.10.

Standard Deviation: A standard deviation of 1.50 indicates a wider spread in parking availability.

Through EDA analysis, the above data analysis results are obtained, which is helpful to understand the structure of the data.

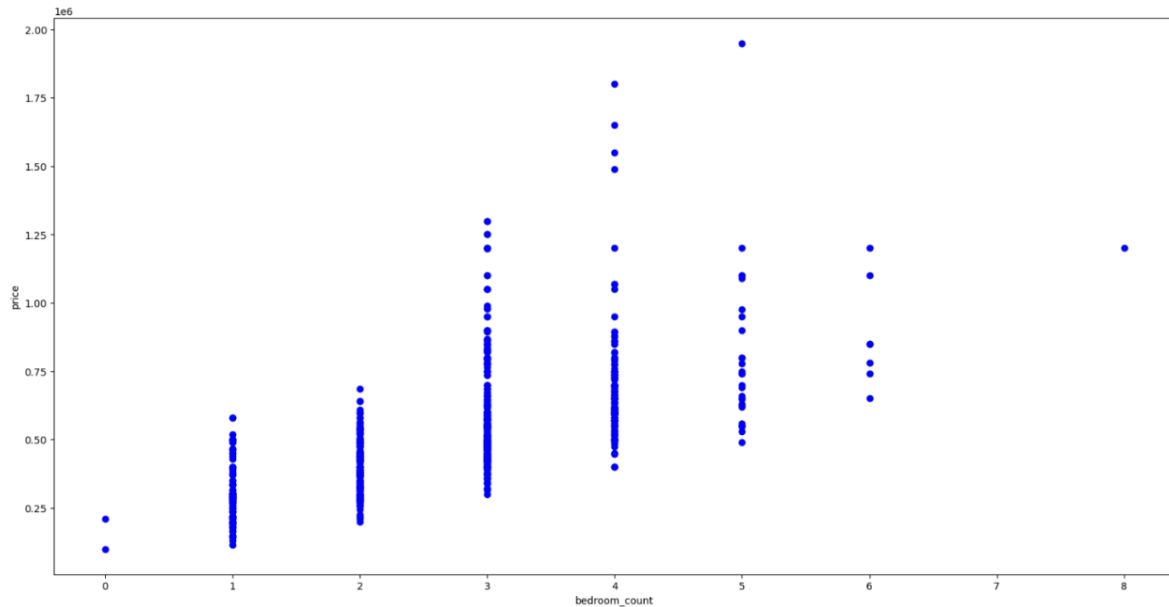


Figure 2. Scatter plot of price and bedroom count (Picture credit: Original)

The horizontal coordinate in Fig. 2 is the independent variable of bedroom count, and the vertical coordinate is the dependent variable of price. As the bedroom count increases, the scatter in the image continues to rise, that is, the price continues to rise. This means that there is some linear relationship between the two variables, and it is a positive correlation. In general, in this data set, for houses with more bedroom counts, people are more inclined to buy even though the house price has gone up.

3.2. Linear Regression Results

After scatterplot analysis shown in Fig. 3, it is concluded that there is a Linear relationship between bedroom count and price. Therefore, the Linear Regression of the two is further analyzed, OLS model is introduced, and least squares method is used. In figure, R-squared is shown.

Dep. Variable:	price		R-squared:	0.404	
Model:	OLS		Adj. R-squared:	0.403	
Method:	Least Squares		F-statistic:	496.6	
Date:	Sat, 24 Feb 2024		Prob (F-statistic):	2.24e-84	
Time:	22:13:50		Log-Likelihood:	-9898.0	
No. Observations:	734		AIC:	1.980e+04	
Df Residuals:	732		BIC:	1.981e+04	
Df Model:	1				
Covariance Type:	nonrobust				
	coef	std err	t	P> t	[0.025 0.975]
const	1.624e+05	1.75e+04	9.304	0.000	1.28e+05 1.97e+05
bedroom_count	1.31e+05	5878.703	22.285	0.000	1.19e+05 1.43e+05
Omnibus:	382.919	Durbin-Watson:	1.831		
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2856.773		
Skew:	2.252	Prob(JB):	0.00		
Kurtosis:	11.551	Cond. No.	8.87		

Figure 3. OLS regression results (Picture credit: Original)

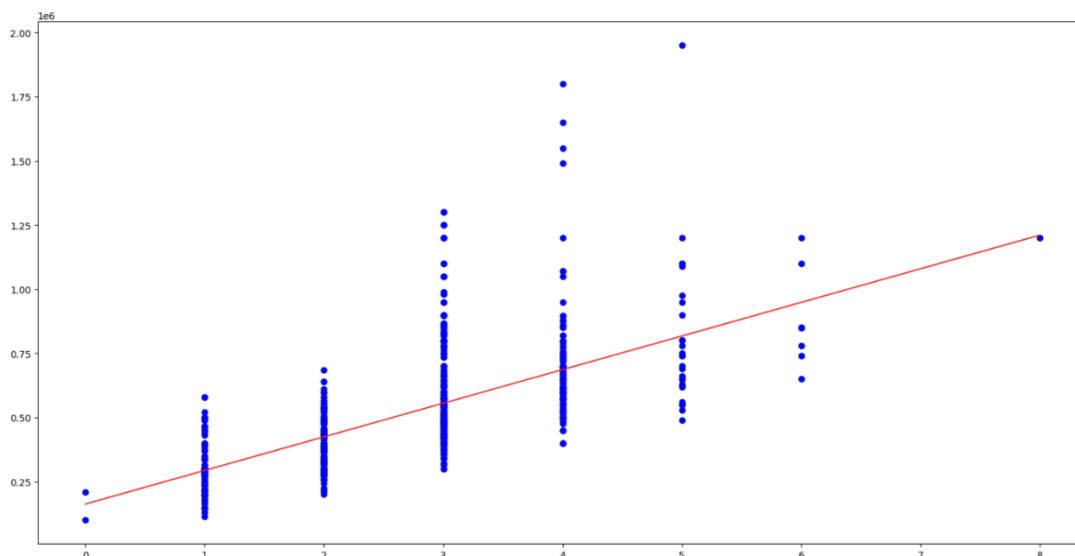


Figure 4. Linear regression analysis of price and bedroom count (Picture credit: Original)

Finally, a Linear correlation graph shown in Fig. 4 of bedroom count and price is obtained through Linear Regression analysis, and it is obvious that there is a positive correlation between the two. In this research, the linear model is mainly used to predict housing prices. However, the R square is 0.404, which may cause the problem of poor fit of the model. Therefore, non-linear models such as isotonic regression and polynomial regression may be considered to predict the housing price again in the future. In addition, more advanced artificial intelligence models e.g. neural networks could be also considered due to their excellent performance in numerous tasks [9-12].

4. Conclusion

This paper is mainly about the relationship between influencing factors and housing prices and the forecast of housing price trends. The data set comes from the Kaggle website, EDA analysis was used to conduct a comprehensive analysis of the data and screen the data, and then OLS model and Linear Regression analysis were used to analyze the filtered data. Finally, linear correlation graphs were made. The research results show that influencing factors such as bedroom count, bathroom count and parking count are positively correlated with price, that is, in the future trend, with the increase of count, price will also increase. The limitations of this study may be the completion of data collection and variable analysis, but in general, the paper does a good job of predicting the trends of influencing factors and prices, as well as the linear correlation between the two variables.

References

- [1] ABC <https://www.abc.net.au/>, 2024.
- [2] Graham J, Read A. House Prices, Monetary Policy and Commodities: Evidence from Australia. *Economic Record*, 2023, 99 (324): 1-31.
- [3] Raju R et al. Machine Learning for Rental Price Prediction: Regression Techniques and Random Forest Model. Available at SSRN 4587725, 2023.
- [4] Truong Q et al. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 2020, 174: 433-442.
- [5] Weisberg S. *Applied linear regression*. John Wiley & Sons; 2005 Apr 1.
- [6] Montgomery DC, Peck EA, Vining GG. *Introduction to linear regression analysis*. John Wiley & Sons; 2021 Feb 24.
- [7] Seber GA, Lee AJ. *Linear regression analysis*. John Wiley & Sons; 2012 Jan 20.

- [8] Amri S, Tularam G A Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. *Journal of Mathematics and Statistics*, 2012, 8 (4): 419-434.
- [9] Qiu Y, Wang J, Jin Z, Chen H, Zhang M, Guo L. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*. 2022 Feb 1; 72: 103323.
- [10] Veeramani N, Jayaraman P, Krishankumar R, Ravichandran KS, Gandomi AH. DDCNN-F: double decker convolutional neural network'F'feature fusion as a medical image classification framework. *Scientific Reports*. 2024 Jan 5; 14 (1): 676.
- [11] Qiu Y, Hui Y, Zhao P, Cai CH, Dai B, Dou J, Bhattacharya S, Yu J. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. *Energy*. 2024 Mar 7: 130866.
- [12] Wu B, Chao KM, Li Y. Heterogeneous graph neural networks for fraud detection and explanation in supply chain finance. *Information Systems*. 2024 Mar 1; 121: 102335.