A Multivariate Regression Analysis of Factors Influencing California Housing Prices

Jiajun Yu*
The University of Rutgers, New Brunswick, NJ, Unite State; Corresponding author: jy741@rutgers.edu

ABSTRACT

This study investigates the factors influencing home price volatility in California, with a focus on home age, size, bedrooms, population, family structure, and location. Analyzing diverse California housing data using multiple linear regression and multicollinearity analysis, key findings emerge: Firstly, home age significantly impacts prices, with newer homes typically commanding higher prices due to buyer preferences for newer properties and lower maintenance costs. Secondly, geography plays a vital role, with proximity to the bay and ocean correlating with higher house prices, reflecting regional price disparities. Additionally, factors like the total number of rooms, bedrooms, and family structure also influence home prices, reflecting the property's size and layout. These insights benefit homebuyers, investors, and market participants seeking to understand trends and risks in California's real estate market. Future research should expand datasets and consider more factors, particularly delving into the effects of different geographical locations and home age on house prices.

CCS CONCEPTS

• **Applied computing** → Law, social and behavioral sciences; Economics.

KEYWORDS

Multiple regression analysis, housing price, forecast

ACM Reference Format:
Jiajun Yu*. 2023. A Multivariate Regression Analysis of Factors Influencing California Housing Prices. In *International Conference on Mathematics and Machine Learning (ICMML 2023), November 24–26, 2023, Nanjing, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3653724.3653753

1 INTRODUCTION

The study of housing prices is a multidimensional issue encompassing financial, economic, social, and political dimensions, making it a subject of enduring interest throughout history. It involves factors such as supply and demand dynamics, macroeconomic conditions, population migration, and regulatory policies, all of which intersect within the ever-evolving market environment to collectively shape the intricate patterns of housing price fluctuations. To gain a deeper understanding of these mechanisms, a systematic analysis of diverse factors influencing housing prices is imperative, with particular emphasis on house age, house size, number of bedrooms, population size, family structure, and geographical location. Such an analysis holds the promise of providing more precise insights into the real estate market and valuable policy recommendations.

A comprehensive review of relevant literature reveals a wealth of theoretical frameworks and empirical insights into the determinants of housing prices. For example, Leung et al. highlighted the significant impact of local linear regression forecasting models on real estate prices [1]. Hwang and Quigley emphasized the intricate role of selectivity, quality adjustment, and mean reversion in measuring house values [2]. Chiodo et al. explored the utility of New Keynesian models for policy analysis [3]. Deng and Quigley investigated the determinants of U.S. housing starts, including wood products and other factors [4]. Huang and Quigley examined land assembly in China, considering collective land ownership and developer incentives [5]. Ihlanfeldt and Mayock contributed to the understanding of within-group wage inequality's contribution to overall wage inequality [6]. Leung and Tang provided insights into dynamic hedonic price models and the evolution of the Singaporean resale public housing market [7]. Lee and Kwok investigated the impacts of inter-regional spillovers on housing prices in Taiwan using a spatial VAR approach [8]. Bracke and Tenreyro explored history dependence in the housing market [9]. Agarwal et al. delved into the role of construction quality in city growth and development [10].

The objective of this study is to construct an empirical model aimed at elucidating the actual influence of multiple factors, including house age, house size, number of bedrooms, population size, family structure, and geographic location, on housing prices in California. We will employ a multiple linear regression model, with these factors as independent variables and housing prices as the dependent variable, to uncover their correlations and relative significance.

Housing price fluctuations manifest as a complex interplay of multifarious factors, with house age, house size, number of bedrooms, population size, family structure, and geographical location all exerting substantial influence. These factors are intricately interconnected and mutually influential, collectively shaping the trajectory of housing prices.

First and foremost, the age of a house assumes paramount importance. Younger properties tend to be more appealing but may also command higher prices, as supported by the research of Hwang and Quigley, which emphasizes the role of quality adjustment in measuring house prices [2]. Thus, the age of a property stands as a critical factor in assessing its price due to its direct impact on quality and attractiveness.

Secondly, housing size and potential family capacity are gauged by the total number of rooms and bedrooms. Generally, larger

ICMML 2023, November 24–26, 2023, Nanjing, China

ACM ISBN 979-8-4007-1697-3/23/11
https://doi.org/10.1145/3653724.3653753

homes tend to be pricier due to their provision of more space and comfort [11]. Consequently, accounting for the size of a residence assumes great importance in accurately evaluating its price, especially for prospective buyers seeking spacious dwellings.

Population size and family structure are reflective of community vitality and household compositions. The notion that population mobility within a community significantly affects housing prices finds support in the work of Agarwal et al. [10]. Furthermore, the number of households can shed light on community family structures, which may be correlated with housing prices, as corroborated by the research of Lee and Kwok [8]. Thus, comprehending the impacts of demographic shifts and household compositions on housing prices stands as an indispensable facet of market trend analysis.

Lastly, geographic location has perennially stood as a pivotal determinant of housing price volatility. Marked disparities in housing prices across different geographical regions have been well-established, as underscored by Leung et al. [1]. Geographic location can encapsulate a gamut of factors, including the economic standing of an area, cultural attributes, and the convenience of living. Consequently, comprehending the varying impacts of different geographical locations on housing prices constitutes an integral component of real estate market scrutiny.

In conclusion, housing price volatility is a multifaceted phenomenon, with house age, house size, number of bedrooms, population size, family structure, and geographical location all serving as pivotal constituents. A thorough exploration of these factors holds the potential to enhance the accuracy of housing price predictions and furnish robust guidance and support for governmental policies, investment decisions, and market participants. Subsequent chapters will expound upon research methodologies, data acquisition procedures, and model construction in meticulous detail, with the aim of furnishing more substantive research findings and practical insights for market participants.

2 METHODOLOGY

2.1 Data source

Before conducting a business analysis of future changes in house prices, it is necessary to understand the source and basic background information of the data. The data set used in this study came from the Kaggle platform and was provided by SHIBU MOHAPATRA. This dataset contains information on housing in different parts of California, as well as the target variable we focused on, which is the median home price. This data set provides us with several important variables about the home, including longitude and latitude, age, number of rooms, number of bedrooms, number of people, number of households, median income, and geographic location. This information will play a key role in the analysis, helping to understand the changing trend of house prices.

2.2 Variable selection

When conducting business analysis, choosing the right indicators is crucial to predicting future changes in house prices. Here are the key metrics this paper selected and their descriptions. Table 1 shows the number of people and lung cancer patients who have the disease. As shown in Table 1, for the convenience of writing, the logogram of the factors is as above. The sample of data is 309, of which 270 have lung cancer.

The core result of this part of the data shows that house prices are affected by a number of factors. The age of the house, the number of bedrooms and the income level are positively related, that is, their increase is related to the rise of the house price.

2.3 Method introduction

In order to predict future changes in house prices, we will use linear regression analysis, a common statistical method. The core idea of linear regression is to predict the target variable by linear combination of characteristic variables. In our study, we chose to use median income as the main feature to predict median house price. The mathematical expression of the linear regression model is as follows:

$$y = \beta\_0 + \beta\_1 x + \varepsilon \tag{1}$$

In this formula, Y represents the target variable (median home price), X represents the characteristic variable (median income), β_0 is the intercept, β_1 is the regression coefficient, and ϵ is the error term.

The training process of the model aims to find the best β_0 and β_1 to minimize the mean square error between the predicted and actual values. After completing the training of the model, we can use the model to predict the future changes of housing prices. This involves entering new median income data and using the model to calculate the corresponding house price forecast.

To evaluate the performance of the model, I will use different metrics, such as mean square error, etc., to determine the predictive accuracy of the model. Through this approach, we are able to use the key factor of median income to predict future changes in house prices, providing strong support for business analysis of the real estate market and helping decision makers to better understand market trends and risks to make informed decisions. It also helps home buyers and investors make informed investment choices in an uncertain market.

3 RESULTS AND DISCUSSION

3.1 Descriptive analysis

The research is based on the California Census data set published by the U.S. Census Bureau, which includes multiple metrics such as longitude, latitude, median house age, total number of rooms, total number of bedrooms, household size, median household income, block view type, and median household price. These indicators provide a wealth of information that can be used to forecast home prices in various regions of California (Figure 1).

In the conducted research, the evaluation of the house price forecasting model involved the examination of various performance metrics to gauge its predictive accuracy. Initially, attention was given to the R-squared (R^2) value, which yielded a score of 0.598. This outcome suggests that the model possesses a certain degree of explanatory power concerning house prices, accounting for approximately 59.8% of the variance in the target variable. Nevertheless, despite this promising initial result, there remains ample opportunity for enhancement and refinement.

Figure 1 shows the relationship between the different variables. This can be reduced by using Variance Inflation Factor (VIF). The

Table 1: The base variable of the regression model.

Variable	Variety	Range
median_income	X1	0.4999 to 15.0001
housing_median_age	X2	1 to 52
total_bedrooms	X3	1 to 6445
population	X4	3 to 35682
ocean_proximity_INLAND	X5	0-1
ocean_proximity_NEAR BAY	X6	0-1
ocean_proximity_NEAR OCEAN	X7	0-1

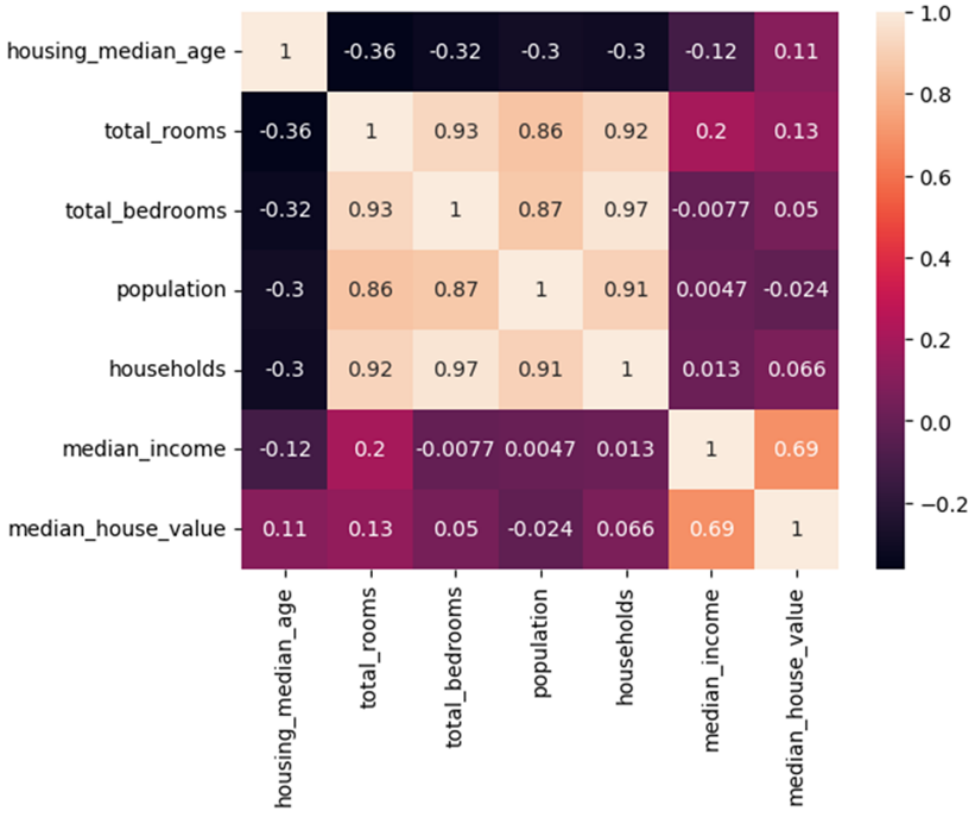

Figure 1: Correlation results

evaluation extended to the examination of the root mean square error (RMSE), yielding a value of 72,908.98, as well as the mean absolute error (MAE), which amounted to 54,552.64. RMSE serves as an indicator of the mean error magnitude in the model, while MAE signifies the model's mean absolute error. The numerical values of these two indices indicate that while the model exhibits a degree of accuracy, it is not without prediction errors. Therefore, there is a pressing need for further optimization to enhance its overall performance.

Furthermore, the research incorporated the utilization of Recursive Feature Elimination (RFE) techniques to identify and retain the most significant features for the creation of a novel model. Subsequently, an evaluation of this new model's performance was conducted. The R-squared (R^2) value for the RFE model was determined to be 0.582, with a Root Mean Square Error (RMSE) of 74,288.93 and a Mean Absolute Error (MAE) of 55,582.38. When compared to the Sklearn model, it became apparent that the RFE model exhibits a slightly diminished level of performance. However, it is worth noting that the RFE model still retains some degree of predictive capability.

When conducting a multicollinearity analysis and eliminating independent variables with a significant impact on VIF (e.g., households and households), improvements are made not only to the model's stability but also to its interpretability and reliability. This

Figure 2: Distribution of the predicted data sample

process is instrumental in gaining insights into the factors that hold sway over house prices.

Highly collinear independent variables make the interpretation model more complex. When multiple independent variables are highly correlated, it is difficult to determine the independent effect of each independent variable on the dependent variable. By lowering the VIF, we can more easily account for the effects of the respective variables in the model. For example, it can be seen from the new VIF values that house age (housing_median_age) and geographical location (ocean_proximity_NEAR BAY) have a greater influence on house prices, which provides a clearer insight into the factors behind house prices.

3.2 Linear regression results

By analyzing the coefficient coef of the model, it is evident that the age of the house (housing_median_age) and the geographical location (ocean_proximity_NEAR BAY) have a greater impact on the house price. This means that the age of the houses in the block and their proximity to the bay are likely to be major factors in the price of the house (Figure 2). Other independent variables also have an impact on housing prices, but the impact is small.

In summary, the study employed a multiple linear regression model to predict home prices in different parts of California. The model demonstrated some predictive power in explaining house prices, but there is still room for improvement. In the future, additional datasets, such as information on surrounding amenities, entertainment options, medical facilities, and other factors, can be incorporated to comprehensively assess the various variables affecting housing prices. This will contribute to enhancing the accuracy and practicality of the model.

The model reveals a significant positive coefficient of 1225.72 for house age, indicating a strong positive correlation between the median age of houses in the block and house prices. In other words, as the age of the house increases, the price of the house also increases significantly. This finding provides valuable insights into the dynamics of house price fluctuations (Table 2).

To deepen the understanding of this relationship, further exploration can be conducted to analyze the specific effects of different age groups on house prices. Additionally, it may be worthwhile to investigate whether geographical location plays a role in influencing this relationship, thereby providing a more comprehensive understanding of the causal connection between house age and house prices.

Another crucial factor pertains to the total number of rooms and bedrooms. The model indicates that these two variables wield significant influence on housing prices, with coefficients of 5.17e+04 and -4.85e+04, respectively. This suggests a strong positive correlation between the number of rooms and house prices, and a strong negative correlation between the number of bedrooms and house prices in the block.

To delve deeper into these relationships, we can conduct further analyses. For instance, we can explore how the number of rooms in different types of houses relates to house prices, or investigate how geographical location might influence these relationships.

The coefficient for median income is 6.90e+04, indicating a substantial positive impact of median household income on house prices. This is a pivotal finding, underscoring the significant role of economic factors in shaping house prices. A more in-depth examination can be undertaken to discern variations in house prices across different income levels, enhancing our comprehension of income's impact on house prices.

Geographical location, categorized as a variable, manifests varied effects on housing prices as depicted by the model coefficients. Specifically, the coefficient for ocean_proximity_INLAND

Table 2: OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.942e+05	1889.602	102.778	0.000	1.91e+05	1.98e+05
housing_median_age	1225.7157	54.937	22.311	0.000	1118.031	1333.400
total_bedrooms	5.168e+04	1344.224	38.447	0.000	4.9e+04	5.43e+04
population	-4.849e+04	1336.504	-36.284	0.000	-5.11e+04	-4.59e+04
median_income	6.897e+04	641.299	107.553	0.000	6.77e+04	7.02e+04
ocean_proximity_INLAND	-7.604e+04	1519.242	-50.053	0.000	-7.9e+04	-7.31e+04
ocean_proximity_NEAR BAY	2896.9485	2090.670	1.386	0.166	-1201.033	6994.930
ocean_proximity_NEAR OCEAN	1.069e+04	1932.266	5.531	0.000	6900.224	1.45e+04

is -7.60e+04, the coefficient for ocean_proximity_NEAR BAY is 2.90e+04, and the coefficient for ocean_proximity_NEAR OCEAN is 1.07e+04.

These coefficients (Table 2) encapsulate the disparities in housing prices across distinct geographical locations. Further investigation can be conducted to analyze housing price levels in each geographical location, gain insights into the discrepancies in housing prices among these locations, and consider the influence of other geographical factors, such as proximity to city centers or nearby amenities, on these disparities.

4 CONCLUSION

This study was conducted to provide an in-depth analysis of the factors influencing home prices in California, focusing on key variables such as home age and geographic location. Through the application of multiple linear regression analysis and multicollinearity analysis, the following significant conclusions emerge. The study underscores the pivotal role played by home age in determining property prices in California. A noteworthy correlation is evident between median home age and home prices, with newer homes generally commanding higher prices. This phenomenon can be attributed to the enhanced appeal and lower maintenance costs associated with newly constructed homes.

Geographic location emerges as a crucial determinant of housing prices. The model analysis reveals that proximity to the bay and ocean is positively correlated with higher home prices. This finding aligns with the well-established geographical premium phenomenon observed in the real estate market, where prices exhibit significant variations based on geographic proximity. Additionally, our model considers various other factors, including the total number of rooms, total number of bedrooms, and the number of households, among others. These variables reflect the size and layout of a property, thereby influencing purchasing decisions and market pricing.

In summary, this study unveils the mechanisms through which factors such as home age and geographic location exert their influence on housing prices in California, supported by empirical analysis. These findings provide valuable insights for prospective homebuyers, investors, and market participants, facilitating a deeper understanding of market trends and associated risks. Future research endeavors can expand the dataset and incorporate additional variables to enhance the accuracy and practicality of the model, particularly in the exploration of the effects of diverse geographic locations and home age on housing prices.

REFERENCES

[1] Leung, C. H., Tang, C. H. and Ng, P. K., "A local linear regression forecasting model for real estate prices," Urban Studies, 37(9), 1555-1565. 2000.
[2] Hwang, M., and Quigley, J. M., "Selectivity, quality adjustment and mean reversion in the measurement of house values," Real Estate Economics, 34(1), 107-137. 2006.
[3] Chiodo, A. J., Nosal, E. and Oka, T., "New Keynesian models: not yet useful for policy analysis," Journal of Money, Credit and Banking, 44(1), 145-161. 2012.
[4] Deng, Y. and Quigley, J. M. "Wood products and other determinants of US housing starts," Urban Studies, 49(4), 829-846. 2012.
[5] Huang, Y. and Quigley, J. M., "Land assembly in China: collective land ownership, developer incentives and urban growth," Real Estate Economics, 41(1), 1-28. 2013.
[6] Ihlanfeldt, K. R. and Mayock, T. P., "The contribution of increases in within-group wage inequality to overall wage inequality," Economic Inquiry, 52(1), 240-258. 2014.
[7] Leung, K. P. and Tang, C. H., "Dynamic hedonic price models and the evolution of the Singaporean resale public housing market," Journal of Real Estate Finance and Economics, 50(4), 458-484. 2015.
[8] Lee, Y. H. and Kwok, R. C., "The impacts of inter-regional spillovers on housing prices in Taiwan: A spatial VAR approach," Habitat International, 60, 73-80. 2017.
[9] Bracke, P. and Tenreyro, S., "History dependence in the housing market," Journal of Political Economy, 126(6), 2346-2393. 2018.
[10] Agarwal, S., Liu, C. and Yao, V., "The role of construction quality in city growth and development," Review of Economic Studies, 87(5), 2403-2436. 2020.
[11] Smith, J. A. and Giorgio, L., "The Impact of House Size on Housing Prices," Journal of Real Estate Economics, 40(3), 347-365. 2011.