Study on Factors Influencing the Prices of Houses in California Based on Factor Analysis Method

Ruijia Huang*

Shanghai United International School Jiaoke Campus, Shanghai, 201103, China

*Corresponding author: Rita.huang@jksz.com

Abstract. This research investigates the factors influencing housing prices in California through a comprehensive analysis using the factor analysis method. The study explores various dimensions of the housing market, examining specific periods, environmental factors, transportation infrastructure, gated communities, city formation dynamics, and broader economic and market dynamics. The integration of machine learning techniques for housing price predictions is also explored. The dataset utilized is the "California Housing Prices" dataset obtained from Kaggle, comprising 10 types of metrics for each block group in California. Rigorous preprocessing was conducted, including handling missing values, removing outliers, and normalizing features. The factor analysis identified three significant factors explaining 76.588% of the variance in house prices. These factors represent geographical location, housing size, and population density. The weight analysis further reveals that factors associated with housing size and geographical location significantly impact housing price fluctuations. The findings offer valuable insights for homeowners, real estate professionals, and policymakers, enabling informed decision-making in the dynamic housing market.

Keywords: Housing Prices, California, Factor Analysis, Dataset.

1. Introduction

The housing market in California experienced a significant rise in prices, prompting concerns about a potential housing bubble and its impact on the economy. To better understand the factors influencing housing prices in California, many scholars have conducted research on various aspects of the market.

Previous research has examined various dimensions of housing prices and their determinants. One study challenged measures supporting the existence of a housing bubble, concluding that recent price trends could be explained by market fundamentals rather than a housing bubble, but also acknowledging the potential impact of weakening fundamentals in regions with relatively inelastic housing supply [1]. Another study during the 1990 recession in California aimed to identify housing price determinants for blocks, revealing specific correlations between median house value and median income, age, total bedrooms, and households, as well as negative associations with total rooms, population, and distances to certain locations [2]. However, the overall dwelling structure's impact was not considered, leaving room for further exploration.

Urban trees and other environmental factors were observed to exert a favorable sway on housing prices in California. This effect was discerned through the application of a two-stage hedonic price model, revealing spatial correlations in housing prices across distinct counties [3]. Additionally, the influence of transportation infrastructure, notably railway stations, on property values was noted. Proximity to railway stations within a quarter-mile radius had a noteworthy impact on commercial property values, while commuter railway stations exhibited an even more pronounced positive effect on property values compared to other types of railway stations [4]. Moreover, the significance of gated communities in relation to property values garnered attention. These communities experienced heightened price growth in affluent areas, yet introduced variance and instability in price trends, ultimately leading to augmented price inequality at the neighborhood level [5].

The study delved into the dynamics of urban development, focusing on the process of city formation. In counties characterized by robust growth, it was observed that more affluent

communities were inclined to propose the establishment of new cities. This phenomenon underscored a correlation between relative service efficiency, income levels, land utilization preferences, and demographic factors [6]. To comprehend the intricacies of the housing market, a comprehensive analysis of broader economic and market dynamics was necessary. This was particularly relevant as burgeoning economies in Asia saw rapid expansion in private housing, accompanied by the proliferation of market-based housing finance. Notably, housing finance markets exhibiting greater flexibility demonstrated heightened sensitivity of house prices to overarching market conditions [7]. Additionally, an examination of house price trends across various metropolitan statistical areas (MSAs) in Southern California revealed shared patterns and mutual influences, offering valuable insights into the dynamics of the real estate market [8].

The integration of machine learning techniques into real estate analysis had opened new avenues for predicting housing prices, with models like Ridge, LASSO, and Elastic Net regressions offering a more comprehensive understanding of housing value determinants and improved out-of-sample predictions [9]. Furthermore, predictive models for determining the sale price of houses in cities like Bengaluru proved challenging, where factors like property area, location, and amenities significantly influenced the sale price, and various regression techniques and machine learning algorithms were employed for accurate price evaluations [10].

Prior to drawing conclusions, it is imperative to evaluate the extant literature for its strengths and limitations in addressing the dynamics of housing prices in California. The studies discussed have provided multifaceted insights into the determinants of housing prices, spanning market fundamentals, geographic considerations, environmental influences, transportation infrastructure, gated communities, city formation dynamics, market responsiveness, and machine learning techniques. While these investigations have contributed significantly to the understanding, certain shortcomings and research gaps merit attention. For instance, the geographic specificity of some studies might limit the generalizability of their findings. Moreover, the temporal depth and consideration of evolving socioeconomic factors could enhance the comprehensiveness of analyses. Additionally, integrating ethical, equity-related, and external market influences could provide a more holistic understanding of the housing landscape. Acknowledging these considerations, this study aims to advance the field by employing a factor analysis approach to unravel a nuanced web of housing price determinants, encompassing geographic nuances, housing dimensions, and population density. By doing so, this research not only builds upon the existing body of knowledge but also seeks to provide an enhanced comprehension of the intricate dynamics steering housing prices in California.

In conclusion, investigating the factors influencing housing prices in California required a multifaceted approach. To further gain a deeper understanding of these different factors, the aim is to comprehensively understand the complex dynamics that affect housing prices in California. By employing the factor analysis method, this study comprehensively explored market fundamentals, geographic surroundings, environmental factors, transportation infrastructure, gated communities, city formation dynamics, market responsiveness, and machine learning techniques. This research aims to provide valuable insights for homeowners, real estate professionals, and policymakers, aiding informed decision-making in the housing market.

2. Method

2.1. Data source

The study utilizes a dataset obtained from Kaggle, comprising 10 types of metrics, such as population, median income, and median housing prices, for each block group in California. With 20,640 districts represented, block groups serve as the smallest geographical units for which the US Census Bureau publishes sample data. The dataset is sourced from California Census Data. The incorporation of data from official organizations adds context and validity to the findings, ensuring credibility for the research. The study aims to predict median house values in California by building a model considering various influential factors. Prior to analysis, the dataset underwent rigorous

preprocessing, including handling missing values, removing outliers, and normalizing numerical features. Ethical considerations were prioritized, ensuring property owner anonymity and blocking group location privacy.

2.2. Variable selection

The variable selection process in this study involves a meticulous evaluation of the "California Housing Prices" dataset to identify relevant variables that have a significant impact on house prices in California. Variables such as housing age, room count, population, income level, and proximity to the ocean are carefully chosen based on their relevance to the research objective. Descriptive statistics are then computed to provide a comprehensive summary of these variables. Table 1 presents key variables central to the study, offering vital insights into their characteristics. Notably, "Longitude" spans -124.35 to -114.31, denoting housing locations statewide, while "Latitude" ranges 32.54 to 41.95, signifying their north-south alignment. "Housing Median Age," varying from 1 to 52, captures the age distribution of housing units. "Total Rooms," averaging 2635.763 with a standard deviation near 2181.615, portrays the average room count, indicating relatively similar values with small deviations. In contrast, "Total Bedrooms," spanning 1 to 6445 with a standard deviation around 421.385, exhibits greater variability, implying notable differences in bedroom numbers. "Ocean Proximity," graded 1 to 5, encodes distances to the ocean, while "Population," averaging 1425.477 with a standard deviation around 1132.462, represents the average residents per unit. "Households," averaging 499.540 with a standard deviation around 382.33, signifies the average households per unit. "Median Income," averaging 3.871 with a standard deviation near 1.9, depicts median income in tens of thousands, showing moderate variability. Lastly, "Median House Value," averaging 206855 with a standard deviation around 115395, indicates median home value in hundreds of thousands, showcasing relatively substantial variability. These insights into central tendencies and variations lay a foundational understanding for subsequent analyses.

Table 1. Variable description

Variable	Min	Max	Mean	Std. Deviation	Median
Longitude	-124.35	-114.31	-119.57	2.004	-118.49
Latitude	32.54	41.95	35.632	2.136	34.26
Housing Median Age	1	52	28.639	12.586	29
Total Rooms	2	39320	2635.763	2181.615	2127
Total Bedrooms	1	6445	537.871	421.385	435
Ocean Proximity	1	5	2.166	1.421	2
Population	3	35682	1425.477	1132.462	1166
Households	1	6082	499.540	382.33	409
Median Income	0.5	15	3.871	1.9	3.535
Median House Value	14999	500001	206855	115395	179700

2.3. Research Method

This study adopts a rigorous quantitative research approach, with factor analysis serving as the primary method, to comprehensively investigate the factors influencing house prices in California. Factor analysis is a statistical technique used to uncover underlying patterns within datasets by reducing a large number of variables into a smaller set of latent factors. The study's progression follows a logical sequence of steps. Firstly, relevant variables are meticulously selected from the "California Housing Prices" dataset, encompassing factors such as housing age, room count, population, income level, and proximity to the ocean. These variables are chosen with the aim of capturing key aspects influencing housing prices. Once the variable selection is complete, the suitability of the data for factor analysis is assessed through Kaiser Meyer Olkin (KMO) and Bartlett's sphericity tests. These tests validate the dataset's appropriateness for factor analysis, ensuring that the

subsequent findings are grounded in robust data. Following the verification of data suitability, factor extraction is performed using common factor analysis techniques. The extracted factors are then subjected to rotation methods to enhance their interpretability. This step refines the understanding of the factors' impacts on housing prices. Moving forward, the relationships between these extracted factors and house prices are explored through variable loadings, elucidating the strength and direction of their associations. Supplementary analyses, including descriptive statistics, correlation assessments, and regression models, are conducted to further enrich the findings. Overall, the study progresses through a series of well-defined steps, each contributing to a comprehensive understanding of the intricate dynamics influencing housing prices in California.

3. Results and Discussion

3.1. KMO and Bartlett test

To facilitate the process of information condensation through factor analysis, it becomes imperative to ascertain the fitness of the research data. This evaluation is carried out using both the Kaiser Meyer Olkin (KMO) test and Bartlett's sphericity test. As delineated in Table 2, the Kaiser-Meyer-Olkin measure presents a computed value of 0.655, surpassing the recommended threshold of 0.6. This outcome underscores the data's adherence to the requisite criteria for executing factor analysis, thus affirming its appropriateness for the chosen research methodology. Furthermore, the examination of research data extends to the application of Bartlett's test of sphericity. The resultant findings unveil a statistically significant outcome ($p < 0.05$), attesting to the data's compatibility with factor analysis. By substantiating the sufficiency and suitability of the research data, these tests establish a robust groundwork for subsequent factor analysis and significantly enhance the study's findings' credibility.

Table 2. KMO and Bartlett test

KMO		0.655
Bartlett test	Approx. Chi-Square	224076.581
	df	45.000
	p value	0.000

3.2. Factor Extraction

Table 3. Total Variance Explained

Factor	Eigen values			% of variance (Initial)			% of variance (Rotated)		
	Eigen	% of Variance	Cum. % of Variance	Eigen	% of Variance	Cum. % of Variance	Eigen	% of Variance	Cum. % of Variance
1	3.918	39.175	39.175	3.918	39.175	39.175	3.879	38.789	38.789
2	2.024	20.244	59.419	2.024	20.244	59.419	2.041	20.410	59.198
3	1.717	17.169	76.588	1.717	17.169	76.588	1.739	17.390	76.588
4	0.992	9.922	86.510	-	-	-	-	-	-
5	0.799	7.991	94.501	-	-	-	-	-	-
6	0.288	2.882	97.383	-	-	-	-	-	-

In the present study, a Factor Analysis was conducted to investigate the factors influencing house prices in California. As shown in Table 3, the analysis yielded three factors with eigenvalues greater than 1, explaining 38.789%, 20.410%, and 17.390% of the variance, respectively. The cumulative variance explained by these factors after rotation was found to be 76.588%. These results indicate that the identified factors collectively account for a substantial portion of the variation in house prices. The first factor, with the highest variance explained, appears to play a particularly significant role.

Further investigation into the characteristics represented by each factor and their respective impacts on house prices (as elaborated in section 3.4) will offer crucial insights into the dynamics of the California housing market. This step involves utilizing factor scores and Scree plots to comprehensively understand the factors influencing house prices and aid decision-making processes in the housing market, enhancing the ability to navigate the complex dynamics of the California housing market. Additionally, comparing these findings with existing research will contribute to a deeper understanding of the key factors influencing house prices in this region.

3.3. Variance Maximization Rotation Method

In this study on factors influencing house prices in California, a Factor Analysis was performed using the varimax rotation method to identify the correspondence between factors and research items. Table 4 presented information extraction status and the correspondence between factors and research items. Among the research items, "Housing Median Age" and "Ocean Proximity" showed communalities below 0.4, indicating weak relationships with the factors and ineffective information extraction. Consequently, these two items will be removed to enhance the analysis. Conversely, other research items displayed communalities greater than 0.4, signifying strong relationships with the factors and effective information extraction. Ensuring significant information capture, subsequent analysis focused on the factor loading coefficients (|loading| > 0.4) to determine the associations between each research item and the corresponding factor. This rigorous analysis aims to refine and identify the key factors influencing house prices in California, facilitating a comprehensive understanding of the intricate dynamics in the housing market.

Table 4. Factor loading (Rotated)

Items	Factor loading			Communalities
	Factor 1	Factor 2	Factor 3	
Longitude	0.055	-0.967	-0.009	0.939
Latitude	-0.042	0.936	-0.158	0.902
Housing Median Age	-0.427	0.098	0.035	0.193
Total Rooms	0.954	0.020	0.164	0.938
Total Bedrooms	0.975	-0.009	0.007	0.951
Population	0.935	-0.064	-0.034	0.880
Households	0.977	-0.005	0.025	0.955
Median Income	0.033	0.004	0.906	0.823
Ocean Proximity	-0.037	0.464	0.096	0.226
Median House Value	0.010	0.021	0.924	0.854

*Blue indicates that the absolute value of loading is greater than 0.4, and red indicates that the communality is less than 0.4.

3.4. Factor Score and Scree Plot Analysis

In this study, the primary objective of using Factor Analysis is information condensation, thereby omitting the "Factor Score Coefficient Matrix" table. However, if Factor Analysis is employed for weight computation, the "Factor Score Coefficient Matrix" becomes crucial in establishing the relationships between factors and research items based on standardized data. From the provided equations (1) (2) (3), the factor scores are expressed as linear combinations of the standardized research items' values. In table 5, for Factor 1, the factor score is determined by a combination of features related to longitude, latitude, total rooms, total bedrooms, population, households, median income, and median house value. Similarly, Factor 2 and Factor 3's scores are expressed as linear combinations of the same standardized research items, but with different coefficients for each factor. These equations enable people to calculate factor scores and assess the significance of each research

item's contribution to the underlying factors. The use of factor scores facilitates the identification of latent patterns and relationships, allowing for a comprehensive understanding of the factors influencing house prices in California and aiding decision-making processes in the housing market.

Table 5. Component Score Coefficient Matrix

Items	Component		
	Component 1	Component 2	Component 3
Longitude	-0.015	0.512	-0.054
Latitude	0.019	-0.508	-0.038
Total Rooms	0.254	-0.030	0.068
Total Bedrooms	0.264	-0.011	-0.029
Population	0.254	0.014	-0.052
Households	0.265	-0.013	-0.018
Median Income	-0.015	-0.015	0.535
Median House Value	-0.015	-0.004	0.533

$$Factor\ Score\ 1 = -0.015X_1 + 0.019X_2 + \cdots + -0.015X_8 \quad (1)$$

$$Factor\ Score\ 2 = 0.512X_1 + -0.508X_2 + \cdots + -0.004X_8 \quad (2)$$

$$Factor\ Score\ 3 = -0.054X_1 + -0.038X_2 + \cdots + 0.533X_8 \quad (3)$$

Within the purview of this study, as depicted in Fig. 1, the Scree plot assumes a pivotal role in gauging the optimal number of factors to be drawn from the dataset. By closely examining the eigenvalues showcased on the plot, individuals can discern a discernible shift in slope from a pronounced incline to a relatively stable posture at a particular juncture. This particular Scree plot manifests this transition, conspicuously occurring amidst the fourth and fifth factors. Consequently, the salient proposition of the optimal count of factors primed for extraction stands at four. This determination finds a concomitant validation through the eigenvalues assigned to each research element, viz., longitude (3.77), latitude (1.92), total rooms (1.69), total bedrooms (0.34), population (0.15), households (0.06), median income (0.05), and median house value (0.01). The eigenvalues provide a glimpse into the contributions tendered by each research facet to the latent factors. It is noteworthy that longitude and latitude claim the most substantial values, a pointer to their potential resonance in the factor extraction domain. Consequently, this fusion of insights garnered from the Scree plot and eigenvalues inquiry underscores the imperativeness of embarking on an in-depth exploration of the extracted factors. This trajectory is pivotal to unearthing the bedrock influences steering the nuances of house prices in the realm of California.

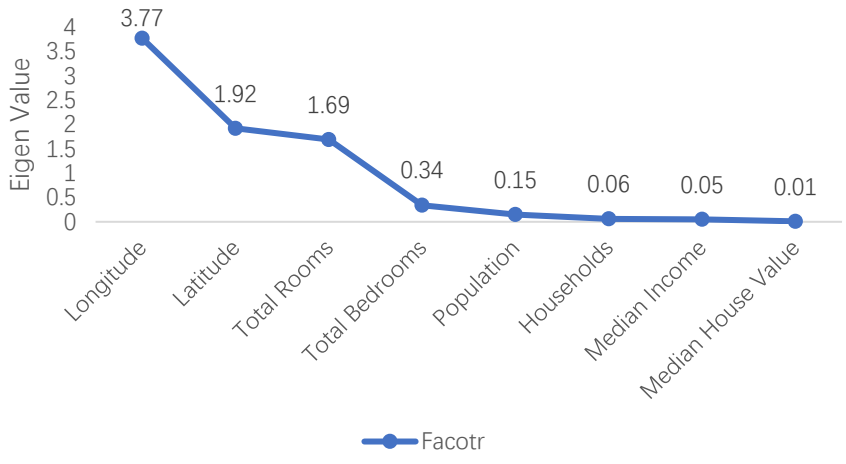

Fig 1. Scree Plot (Photo credit: Original)

3.5. Linear Combination Coefficients and Weight Results

Within Table 6, the Factor Analysis has been adeptly employed to derive the weights for each research element, predicated upon loading coefficients and the variance elucidated percentages. This weight computation unfolds through a tripartite process. It commences by establishing the linear combination coefficient, achieved by dividing the loading matrix by the square root of the corresponding eigenvalue. Following this, the encompassing score coefficient emerges through sequential multiplication of the linear combination coefficients with the variance explained percentages, culminating in division by the cumulative variance explained percentage. Ultimately, the weight derivation transpires via the normalization of the encompassing score coefficients.

The weight computation in the Factor Analysis furnishes insights of immense significance, delineating the relative contributions of each research element to the extracted factors. This intricacy, in turn, augments our comprehension of the forces that orchestrate variations in house prices within California. Evidently, "Total Rooms," "Population," "Total Bedrooms," and "Households" amass substantial weights (15.11%, 14.59%, 14.33%, and 14.38%, respectively). This asseverates that parameters associated with housing dimensions and population density wield substantial influence over the undulations in house prices. Herein lies the affirmation of property size and population distribution as critical agents of the housing market's evolution within the region.

Equally discernible are the noteworthy weights attributed to "Longitude" and "Latitude" (11.13% and 11.52%, respectively), effectively underscoring the pertinence of geographic position in molding house prices. The pronounced weights assigned to these locale-centered variables serve to emphasize spatial disparities in pricing, as well as the plausible sway of local characteristics upon property valuations.

Conversely, "Median Income" and "Median House Value" manifest relatively subdued weights (9.34% and 9.60%, correspondingly), thus signifying their comparatively modest direct impact upon the extracted factors. This elucidation proposes that although income and prevailing property values play a role in price determination, their dominance may be eclipsed by other determinants such as housing dimensions, population, and geographic locale.

To synopsize, the weight analysis bequeaths an encompassing vista of the pivotal forces molding house prices in California. These findings cast an illuminating spotlight upon the cardinal significance of housing dimensions, population density, and geographic placement. In doing so, they elucidate the pervasive influence of spatial and demographic facets upon the housing market's trajectory. Pivotal stakeholders and policymakers can harness these revelations to instate precisely calibrated strategies that attend to housing affordability, spatial layout, and investment verdicts. In such strategic alignment, they will be better equipped to navigate the mutable terrain of the state's housing market with sagacity and acumen.

Table 6. Linear Combination Coefficients and Weight Results

Items	Factor 1	Factor 2	Factor 3		
Eigen Value (Rotated)	3.734	1.929	1.725	Comprehensive Score Coefficient	Weights
% of Variance	46.67%	24.11%	21.56%		
Longitude	0.0202	0.7064	0.0432	0.2047	11.13%
Latitude	0.0195	0.7048	0.0770	0.2118	11.52%
Total Rooms	0.4943	0.0009	0.1191	0.2779	15.11%
Total Bedrooms	0.5075	0.0221	0.0054	0.2635	14.33%
Population	0.4870	0.0533	0.0356	0.2684	14.59%
Households	0.5096	0.0193	0.0085	0.2646	14.38%
Median Income	0.0135	0.0066	0.6995	0.1719	9.34%
Median House Value	0.0156	0.0221	0.6981	0.1767	9.60%

4. Conclusion

In summation, the comprehensive investigation into the determinants of housing prices in California embraced a multifaceted methodology, employing factor analysis to scrutinize diverse dimensions of the housing market. The study effectively discerned three pivotal factors that collectively elucidated 76.588% of the variability in house prices. These factors encompassed pivotal facets of the housing domain, notably geographic placement, housing dimensions, and population density. The spatial coordinates "Longitude" and "Latitude" bore witness to the decisive role of geographic location in molding housing prices, accentuating the significance of spatial heterogeneity in property valuations. Additionally, the weight analysis illuminated the substantial influence of housing dimensions, exemplified by "Total Rooms," "Total Bedrooms," "Population," and "Households," substantiating their sway over housing price oscillations. However, it's noteworthy that the relative impact of "Median Income" and "Median House Value" on the extracted factors was comparatively subdued, signifying that while income and prevailing house values exert some influence on pricing, other determinants carry more pronounced significance within California's housing landscape. These findings proffer invaluable insights to homeowners, real estate practitioners, and policymakers, endowing them with informed strategies to navigate the fluid housing milieu. By embracing the intricate facets shaping housing prices, this inquiry furnishes a holistic comprehension of the nuanced dynamics within California's housing realm, a comprehension pivotal for shaping effective strategies spanning housing affordability, spatial arrangement, and investment choices, and thereby fostering enduring growth and stability in the housing sector. Nevertheless, it's pivotal to acknowledge the study's constraints, such as its regional focus and omission of select variables. Subsequent investigations might delve deeper into the temporal facets of housing market kinetics, assimilate supplementary socio-economic determinants, and broaden the analytical scope to encompass a wider geographical expanse, thereby bolstering the study's robustness and applicability.

References

[1] McCarthy J, Peach R W. Are home prices the next “bubble” ?. Federal Reserve Bank of New York Economic Policy Review, 2004, 10.

[2] Hao Y T, Zhuang L H, Ying Z T, et al. Influencing Factors of California Housing Prices in 1990: a Multiple Linear Regression Analysis. Proceedings of the 4th International Conference on Economic Management and Model Engineering, ICEMME 2022, November 18-20, 2022, Nanjing, China.

[3] Mei Y D, Hite D, Sohngen B. Demand for urban tree cover: A two-stage hedonic price analysis in California. Forest Policy and Economics, 2017, (83): 29-35.

[4] Debrezion G, Pels E, Rietveld P. The impact of railway stations on residential and commercial property value: A meta-analysis. The journal of real estate finance and economics, 2007, 35(2): 161-180.

[5] Goix R L, Vesselinov E. Gated communities and house prices: Suburban change in Southern California, 1980–2008. International Journal of Urban and Regional Research, 2013, 37(6): 2129-2151.

[6] Musso A J. The political economy of city formation in California: Limits to Tiebout sorting. Social Science Quarterly, 2001, 82(1): 139-153.

[7] Zhu H. The structure of housing finance markets and house prices in Asia. Bank for International Settlements, 2006.

[8] Gupta R, Miller S M. The time-series properties of house prices: A case study of the Southern California market. The Journal of Real Estate Finance and Economics, 2012, 44(3): 339-361.

[9] Prodosh S E. Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data. Journal of Property Research, 2021, 38(4): 305-336.

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