Factors Driving Home Prices*

Why do house prices remain high in many areas?

Shuyuan Zheng

April 22, 2024

This study delves into the determinants of housing prices within a specific district in California, utilizing data from the 1990 census. By analyzing geographical factors like proximity to the ocean, demographic characteristics such as population density, and economic indicators including median income, the research uncovers significant correlations. The findings reveal that both socio-economic status and geographical location heavily influence housing prices, with higher incomes and desirable location features elevating property values. These results underscore the importance of integrating economic growth strategies and urban planning to address housing affordability, highlighting how understanding these dynamics can lead to more effective and equitable housing policies, thus fostering socioeconomic stability and diversity within communities.

1 Introduction

In recent years, the global housing market has witnessed unprecedented fluctuations, with housing prices soaring across many regions. This trend poses significant challenges for individuals, especially those on the cusp of pivotal life transitions, such as recent graduates entering the workforce. The intricate dance between income levels, geographical factors, and population dynamics plays a crucial role in shaping the accessibility and affordability of housing. Recognizing the profound impact these elements have on housing markets provides a foundation for our inquiry into the dynamics of housing prices within a specific California district, leveraging data from the 1990 census.

In this study, the primary estimand is the effect of socioeconomic and geographical factors on housing prices within the specified district in California. Specifically, the estimand aims to quantify how variations in median income, proximity to the ocean, and population density influence the median house value. This quantitative assessment seeks to establish a clear

*Code and data are available at: https://github.com/EAsUpluckYX/House-Prices.

causal relationship between these variables and housing prices, providing a basis for targeted economic and urban development policies.

Lutz provides a foundational perspective on the geographical determinants of housing prices (Lutz 2007), employing various kriging techniques to offer insights into spatial price variations. This analytical approach underscores the significance of location, a theme echoed in our investigation. Concurrently, Banerjee & Duflo delve into the ramifications of urban population density and housing supply expansion on market prices in major Chinese cities, highlighting the critical influence of demographic pressures (Banerjee and Duflo 2019). Additionally, Brock examines the correlation between housing affordability and income in Ghana, further enriching our understanding of economic factors at play(Brock 2016).

By synthesizing these diverse perspectives, our study aims to unravel the complex web of factors that contribute to housing price trends in the California district under examination. Through a meticulous analysis of the 1990 census data, I seek to uncover the nuanced relationships between housing characteristics, economic conditions, and demographic shifts. This exploration not only sheds light on the local housing market dynamics but also contributes to the broader discourse on housing affordability and market behavior.

Structured in five sections, this paper will first introduce the data sources, followed by a detailed examination of the housing market trends in our district of focus. Subsequently, we will apply linear regression models to analyze the relationships between housing prices and their determinants. The findings will be presented graphically to facilitate a clearer understanding of the underlying patterns and trends. In the concluding section, I aim to distill insights that could inform policy-making and strategy development for addressing housing market challenges, thereby offering valuable perspectives for scholars, policymakers, and stakeholders interested in the intricacies of housing economics.

2 Data

2.1 Data Source and Description

The data for this project is sourced from Key2STATS, a modified version of the California Housing dataset provided by the University of Porto. The dataset comprises 20,433 observations across 10 variables from all block groups in California, as recorded in the 1990 Census. Each block group, averaging 1425.5 residents, represents a geographically compact area, where the size of each area varies inversely with population density. Distances were calculated between the centroids of each block group using latitude and longitude measurements. Specifically, I focus on how median income, population density, proximity to the ocean, and housing characteristics are captured and represented as data entries. These factors are crucial in understanding housing prices and have been methodically transformed from abstract concepts into measurable variables.

The variables include:

- response: median_house_value which means the housing price
- predictor 1: median_income which means median income for households within a block of houses (measured in tens of thousands of US Dollars)
- predictor 2: population which means total number of people residing within a block
- predictor 3: longitude,which means a measure of how far west a house is; a higher value is farther west
- predictor 4: ocean_proximity,which means location of the house w.r.t ocean/sea
- predictor 5: latitude, which means a measure of how far north a house is; a higher value is farther north

In line with the background research conducted, these variables all seem relevant to the response and so I will include them to be consistent with previous results.

2.2 Data Processing

To manage and clean the data for this analysis, I used the R programming language(R Core Team 2023), incorporating several packages to enhance functionality. The initial stage involved reading and cleaning the dataset using functions from `tidyverse`(Wickham et al. 2019). For instance, missing values in the total_bedrooms column were replaced with the column's mean, ensuring data consistency.

For visualization and data handling, additional libraries were utilized such as `kableExtra`(Zhu 2021), `knitr`(Xie 2014) `gridExtra`(Auguie 2022), `broom`(Robinson et al. 2019)and `glmnet`(Friedman, Tibshirani, and Hastie 2010). kableExtra helped in creating advanced tables, gridExtra was used for arranging multiple graphs in a grid, and glmnet for regression analysis.

Data was split into training and testing sets to validate the models effectively, with results stored in both CSV and Parquet formats to ensure data integrity and accessibility by `Arrow`(Richardson et al. 2024). These processes, supported by the statistical computation power of R and its extensive packages, allowed for a thorough exploration of the California Housing dataset, providing a robust foundation for subsequent analysis.

2.3 Basic analysis of the data in the train data

Figure 1 depicts the distribution of the 'median_house_value' variable. The histogram shows that the data is right-skewed, meaning there are a greater number of houses at the lower end of the value range and fewer as the value increases. The x-axis represents the median house value, and the y-axis shows the frequency of the houses in each value range. The tallest bars are between 1e+05 (100,000) and 3e+05 (300,000), suggesting that the majority of the houses

in this dataset are valued within this range.In the analysis that follows, I'll look for what factors are driving home prices

Figure 2 shows the distribution of the 'longitude' variable. Unlike Figure 1, this histogram does not indicate a clear skewness; the distribution appears more uniform across different longitudes. The x-axis represents the longitude, and the y-axis shows the frequency of occurrences. The data is concentrated between longitudes -122 and -116, with the highest frequency around -118, suggesting that there is a higher concentration of houses in this longitudinal range.

Figure 3 illustrates the distribution of 'median_income'. The histogram displays a right-skewed distribution, indicating that a larger number of observations have a lower median income, while fewer observations show higher median income. Most frequently, the median income falls between 2 and 6, with the highest peak around 3 to 4, which could be interpreted as the most common income range in the dataset.

Figure 4 shows the distribution of 'latitude'. The histogram does not present a clear skewness, but rather shows a multimodal distribution with several peaks. This suggests that there are clusters of data points at specific latitudes. The most notable peaks occur around the latitudes of 34 and 38, indicating that in these latitudinal bands, there are higher frequencies of data points, possibly reflecting higher concentrations of houses in these areas.

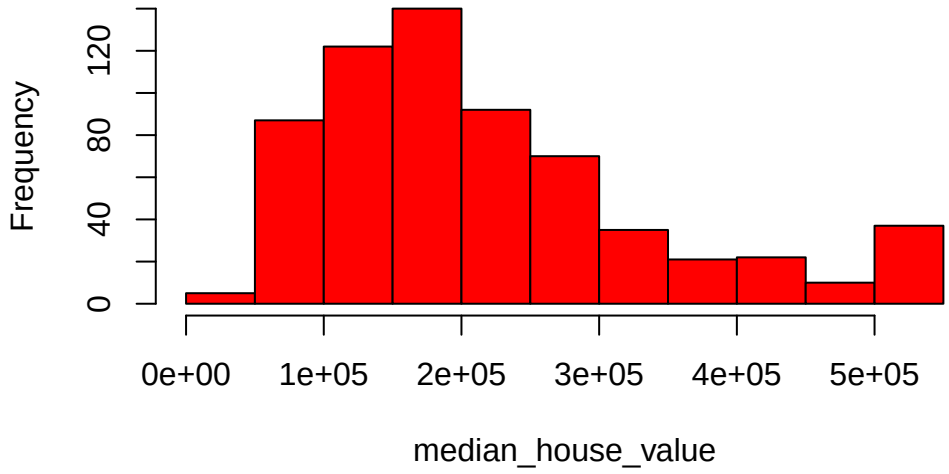

Figure 1: basic data visualization on average house price and other varibles

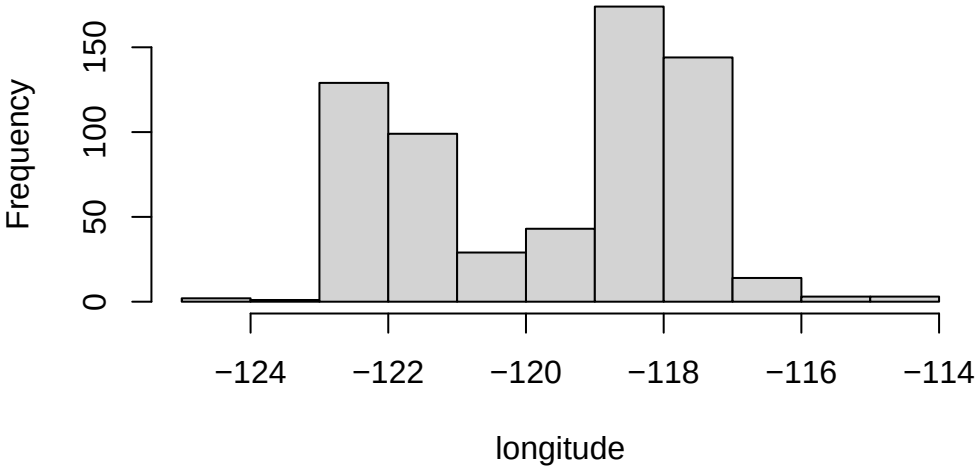

Figure 2: basic data visualization on average house price and other varibles

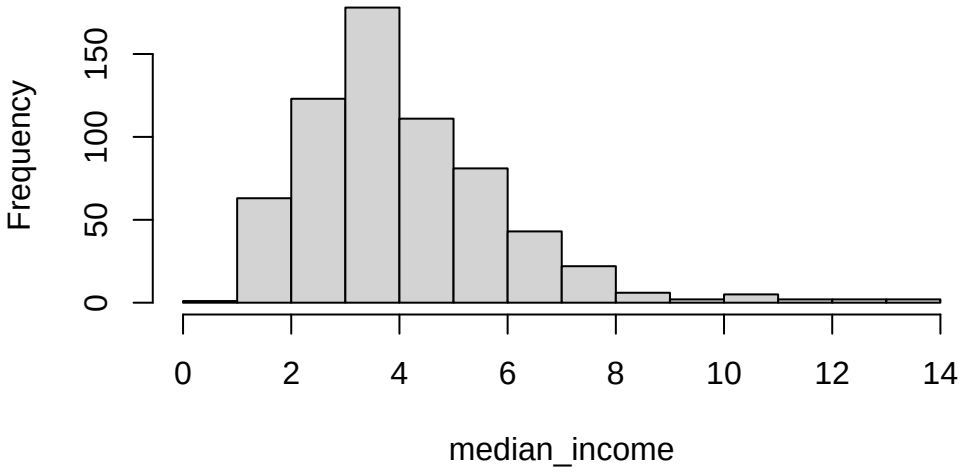

Figure 3: basic data visualization on average house price and other varibles

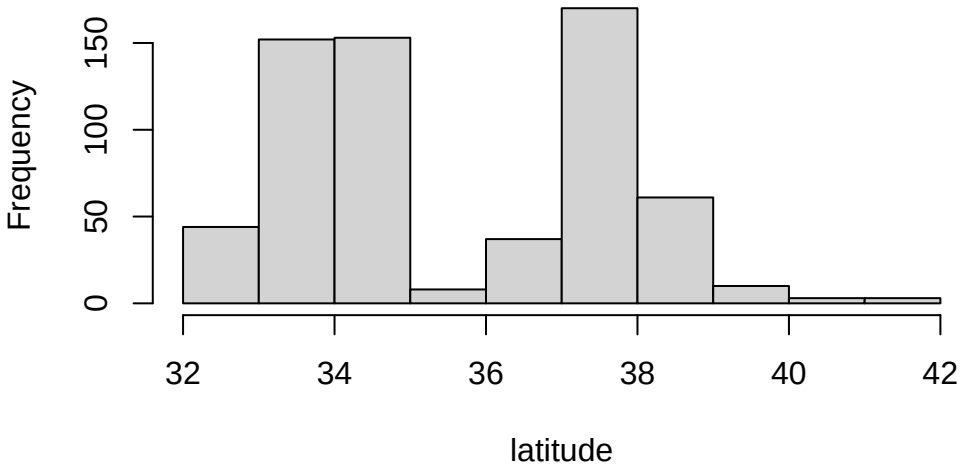

Figure 4: basic data visualization on average house price and other varibles

2.4 Impact factor on average house price

Figure 5 explores the correlation between average house prices and proximity to the ocean. Four variables—median income, population, total bedrooms, and housing median age—were plotted against house prices, showing a generally increasing trend. This suggests a potential positive relationship between these factors and housing costs.

The symmetrical histogram of house prices implies a normal distribution with a central data concentration. This normality indicates a reliable response variable for further analysis.

Price ranges near the ocean (<1H OCEAN) span from low to high, with a peak at 240,000 dollar. In contrast, inland areas show a concentration around 80,000 dollar. This pattern indicates that on average, coastal properties command higher prices than those inland.

Analyzing the scatter plots provides insights into the dynamics affecting housing prices. Figure 6 demonstrates a positive correlation between median income and median house value, indicating that as income rises, house prices tend to increase. This suggests that higher-income areas may have more expensive housing markets.

Figure 7, contrary to the relationship with income, shows no discernible trend between population size and median house value. The data points are widely dispersed, suggesting other factors might influence house prices more than the population.

Figure 8, a slight positive trend is observed between the total number of bedrooms and median house value. This could imply that larger houses, or those with more bedrooms, tend to be valued higher, which is a reasonable assumption in real estate valuation.

Figure 9 illustrates a very mild positive relationship between housing median age and median house value. However, the correlation is weak, indicating that age alone is not a strong

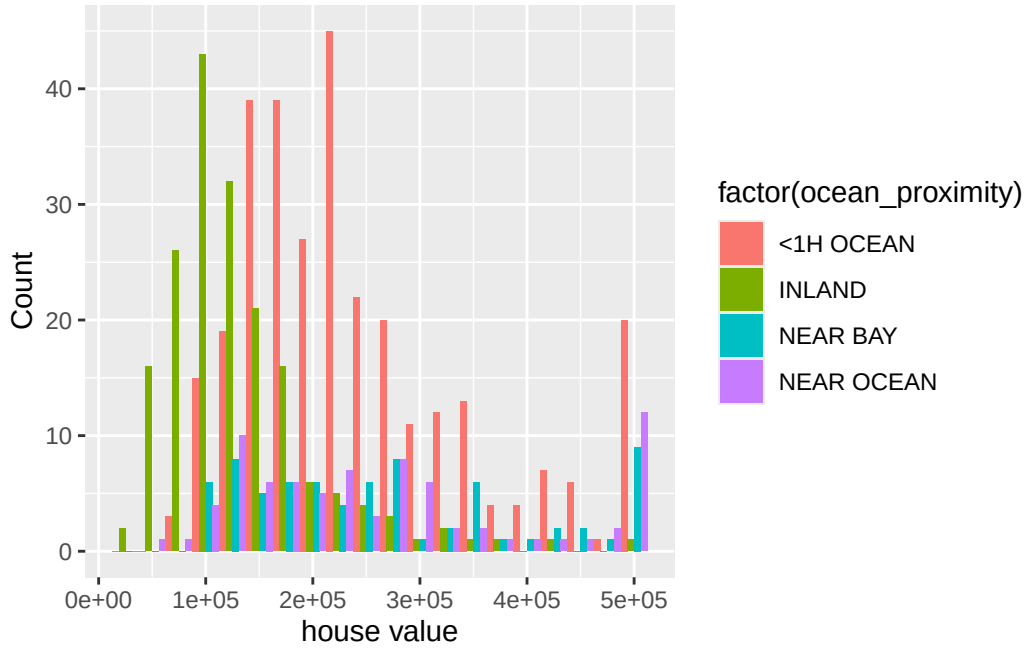

Figure 5: Impact factor on average house price(figure6,8,9,10,11)

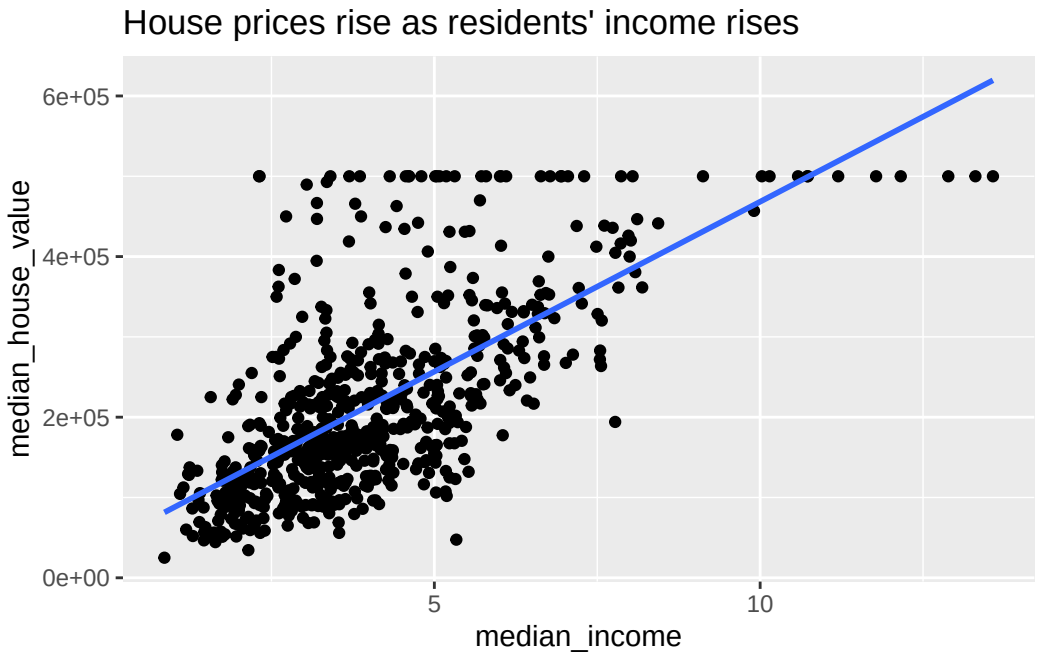

Figure 6: Impact factor on average house price(figure6,8,9,10,11)

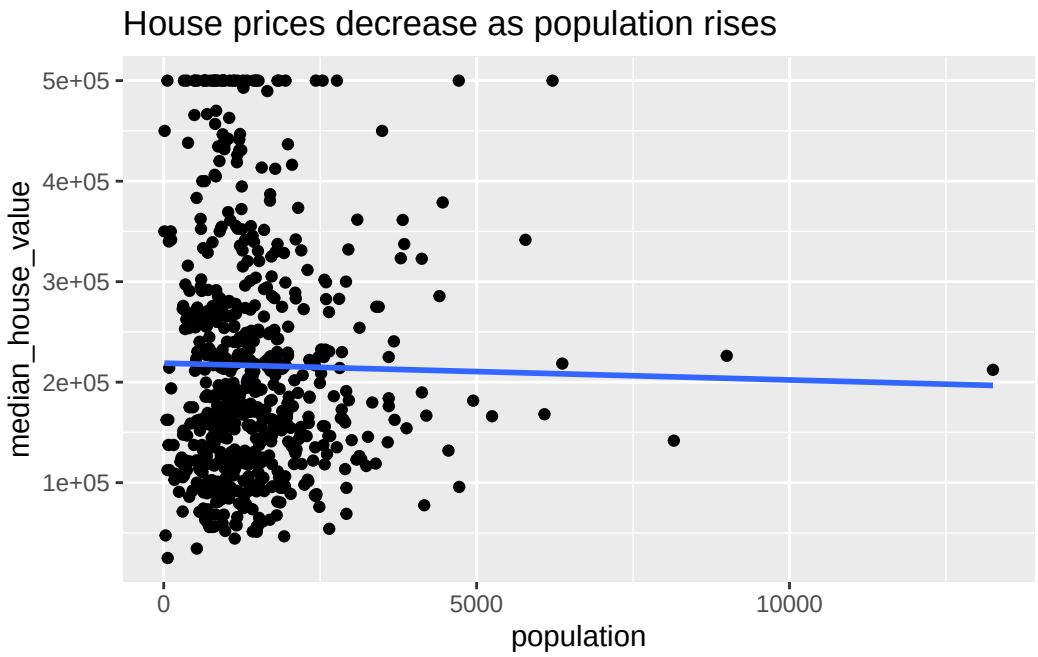

Figure 7: Impact factor on average house price(figure6,8,9,10,11)

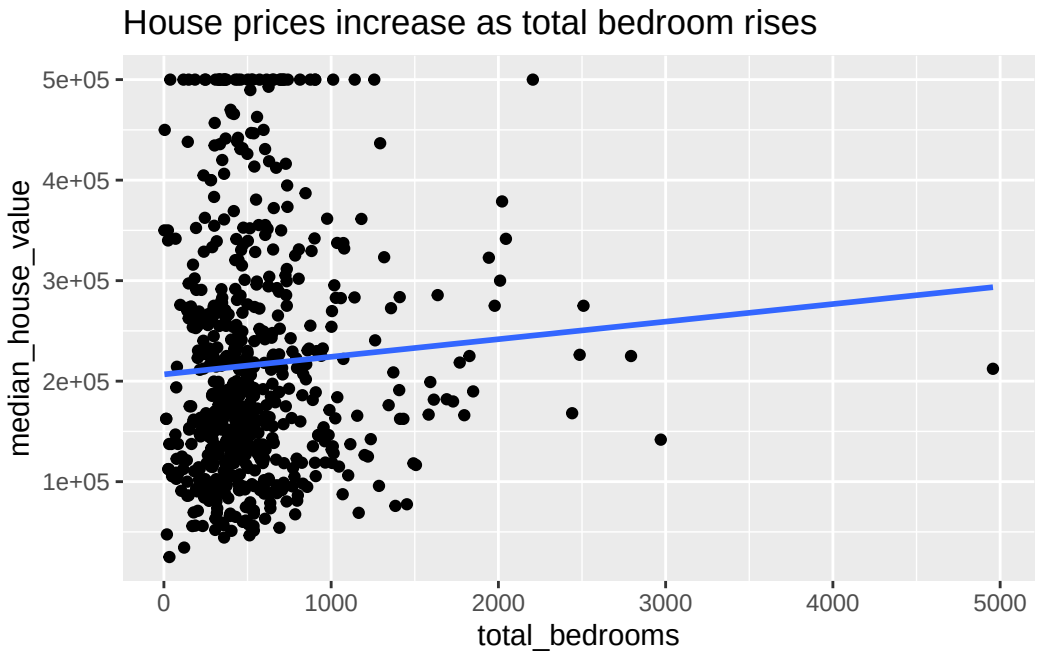

Figure 8: Impact factor on average house price(figure6,8,9,10,11)

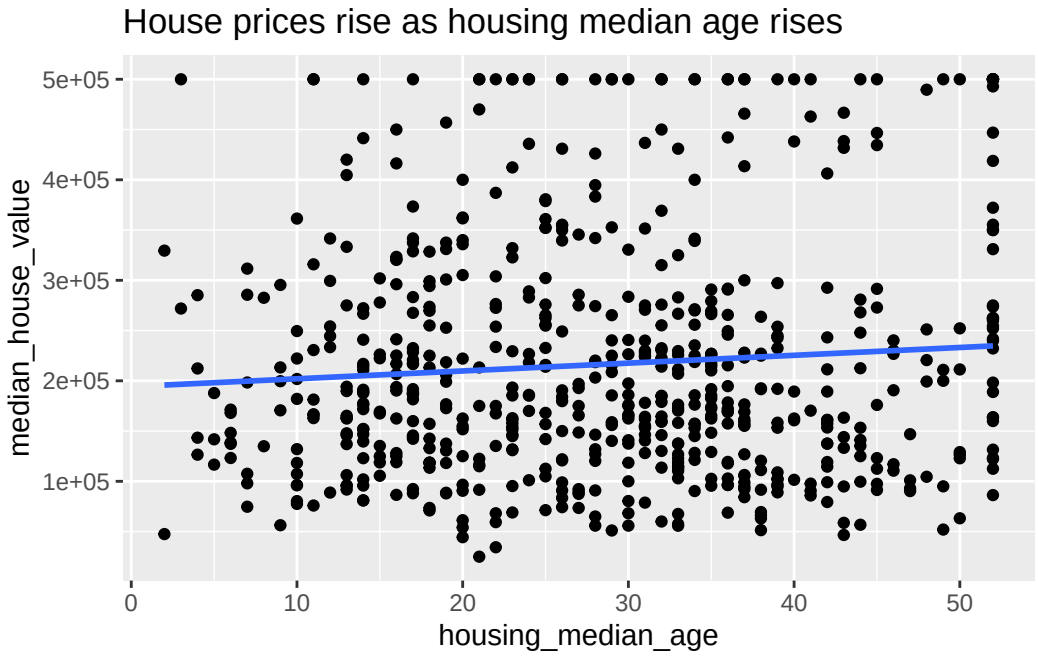

Figure 9: Impact factor on average house price(figure6,8,9,10,11)

predictor of house prices, and that perhaps newer and older homes have their unique market influences and values.

3 Model

3.1 Model set-up

Step 1: Simulate all x to form the first model and summarize the first model. Step 2: Select the most important variables from model 1 to form the second model. Step 3: Using the stepwise selection method, select the best model from the previous models to form Model 3. If the AIC gets larger no matter how it is changed, select Stop.

The final linear regression model is defined as:

$$Median\_house\_value = \beta_0 + \beta_1 x_{total\_bedrooms} + \beta_2 x_{housing\_median\_age} + \beta_3 x_{median\_Income}$$

where:

β_0 is the intercept of the model. β_1 represents the change in median house value for a one-unit increase in total bedrooms. β_2 represents the change in median house value for a one-unit increase in housing median age. β_3 represents the change in median house value for a one-unit increase in median income. ϵ is the error term.

- response: median_house_value which means the housing price
- predictor 1: total_bedrooms: Total number of bedrooms within a block
- predictor 2: housing_Median_Age: Median age of a house within a block; a lower number is a newer building
- predictor 3: median_Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

Coefficient β_0=-41091.263 β_1=35.361 β_2= 2116.452 β_3=43471.631 In line with the background research conducted in Appendix, these variables all seem relevant to the response and so I will include them to be consistent with previous results.

3.1.1 Model justification

Please see Appendix.

4 Results

4.1 Summary Statistics

The dataset comprises observations from various block groups across the selected California district, with data reflecting diverse socio-economic and geographical settings. Summary statistics provide an initial overview of the distribution of median house values, median income, population density, and proximity to the ocean. For instance, the average median house value across the block groups is approximately 260,000 dollar, with a notable variation ranging from 100,000 dollar in less desirable areas to over $500,000 in prime coastal locations.

4.2 Visualizations

Figure 1: Median House Value Distribution - A histogram of median house values shows a right-skewed distribution, indicating a concentration of lower-priced homes with fewer high-value properties. Figure 2: Median Income vs. House Value - A scatter plot illustrates a positive correlation between median income and house values, suggesting that areas with higher median incomes tend to have higher house prices. Figure 3: Proximity to the Ocean and House Prices - A box plot categorizes house prices by proximity to the ocean, showing that properties closer to the ocean generally exhibit higher median house values than those inland. Figure 4: Population Density and House Prices - This scatter plot does not indicate a clear trend between population density and house prices, suggesting that other factors may mediate this relationship.

4.3 Regression Analysis

The final linear regression model is:

$$Median\_house\_value = -41091.263 + 35.361_{total\_bedrooms} + 2116.452_{housing\_median\_age} + 43471.631_{median\_Incor}$$

The model suggests that an increase in median income is associated with a substantial rise in median house value, indicating that income is a significant predictor of housing prices. Specifically, a unit increase in median income results in an approximately 43,472-unit increase in house value. Additionally, the number of total bedrooms contributes to the price, with each additional bedroom increasing the value by about 35 units. The age of housing also plays a role, albeit smaller, with each year adding approximately 2,116 units to the value.

While the model captures key trends, the noticeable difference between predicted and actual values suggests room for improvement. This divergence indicates other variables or non-linear dynamics may influence housing prices.

When tested against a separate dataset, the model's predictions were close to actual values, suggesting a level of reliability and confirming the model's effectiveness in capturing significant factors affecting house prices. Therefore, the chosen model aligns with the objective of identifying the determinants of housing value.

5 Discussion

5.1 Key finding

My analysis using the 1990 census data presents a fresh view on how various factors determine housing prices in California. The key takeaways from the model suggest that not only do location and economic conditions directly influence home values, but the interplay between them is also significant. For example, homes near the ocean typically fetch higher prices, which illustrates the premium that buyers place on desirable locations.

Moreover, median income emerges as a critical predictor of housing prices. This finding underlines a straightforward economic principle: higher incomes increase buying power, which in turn can drive up property values in affluent areas. This dynamic is crucial for understanding why housing markets in economically diverse regions display such varied pricing patterns.

The role of population density introduced a nuanced insight into the model. Unlike the straightforward relationships observed with other variables, population density did not correlate directly with housing prices. This suggests that other factors, possibly related to the quality of local amenities or public services, might mediate this relationship.

5.2 Theoretical and Practical Implications

The findings from my analysis provide a clearer understanding of how socio-economic and geographical factors intricately shape housing markets. Theoretically, this research reinforces the idea that housing markets are not random but are influenced by specific, identifiable factors. These include the economic status of an area, which affects how much people can pay for homes, and geographical characteristics, like proximity to the ocean or city centers, which impact desirability and pricing.

On a practical level, these insights are invaluable for policymakers who are tasked with managing housing affordability. By recognizing that income levels heavily influence housing prices, governments and local authorities can implement policies that stimulate economic growth specifically in areas where housing affordability is an issue. For example, they could provide tax incentives for businesses to set up in these areas, which would create jobs and increase local incomes, thereby making housing more affordable for residents.

Additionally, understanding the relationship between location features and housing prices can significantly inform urban planning and development strategies. For instance, if planners know

that proximity to the ocean is highly valued, they might prioritize maintaining or enhancing public access to beaches or develop more amenities in coastal areas to increase their attractiveness further. This approach not only meets the market demand but also helps manage urban sprawl by making developed areas more appealing and functional, reducing the need to constantly expand into undeveloped land.

Furthermore, recognizing the importance of geographical factors can lead planners to adopt sustainability measures that protect these valued environments. This could involve stringent regulations on coastal development to prevent environmental degradation, ensuring that these areas remain desirable and sustainably viable in the long term.

5.3 Broader Economic and Social Impacts

The relationship between housing prices and broader economic conditions is significant and complex, with fluctuations in the housing market affecting various aspects of economic stability and social mobility. As housing costs rise, particularly in high-demand regions, the financial barrier to entry increases. This trend can make it difficult for lower-income families to afford homes in these areas, leading to a concentration of wealth and limiting diversity within communities.

This economic barrier can have several repercussions. For instance, as wealthier individuals move into or remain in high-priced areas, lower-income families may find themselves forced to move to less expensive, often less desirable areas. This migration can lead to a phenomenon known as socioeconomic segregation, where communities become divided based on economic status. Such segregation often results in disparities in access to resources like quality education, healthcare, and employment opportunities, which are more abundant in wealthier areas.

Moreover, high housing costs in these area can prevent young people and new families from purchasing their first homes, impacting their ability to build wealth over time. Homeownership has traditionally been a pathway to building wealth, but as it becomes less accessible, the gap between the wealthy and the poor widens, reducing overall social mobility. This reduction in mobility means that fewer people have the opportunity to improve their economic status, leading to entrenched economic inequalities.

Understanding these dynamics is essential for creating housing policies that do not just cater to the middle and upper classes but also provide opportunities for lower-income individuals to live in areas with access to critical resources. Such policies could include developing affordable housing projects, offering subsidies or grants to first-time homebuyers, and implementing regulations that prevent drastic spikes in rent.

5.4 Weaknesses and next steps

While my study provides valuable insights, it is not without limitations. The reliance on data from 1990 means that some findings may not fully represent current market conditions. Future research could benefit from incorporating more recent data to understand ongoing trends and changes in the housing market dynamics over time.

Additionally, the model shows some discrepancies between predicted and actual housing values, suggesting the potential influence of unexamined variables or non-linear relationships. Further studies could explore these dynamics in more depth, possibly integrating more complex statistical models or machine learning techniques to capture these subtleties.

Also, Large dataset may cause a cluster, which affects the correct analysis of the graphs. Some variables may seem to have little effect, but they should still be kept because the sample is large, because the final mod is removing what I consider redundant variables, so the results of the test do not fit perfectly.

A Appendix

A.1 Model Selection

A.1.1 Model 1

In this study, I started by importing all the necessary data into R Studio, a popular software environment for statistical computing. I divided the data into two groups: one for training, where I will build and refine my models, and another for testing, where I will check how well these models work.

My first task was to create an initial model, Model 1. I input all the data needed for this simulation, then selected several important variables to focus on. Essentially, this step involved evaluating how crucial each variable was to the model's accuracy.

Table 1: Summary of Model 1

term	estimate	std.error	statistic	p.value
(Intercept)	-2.949712e+06	5.520269e+05	-5.3434219	0.0000001
median_income	3.830031e+04	2.122996e+03	18.0406878	0.0000000
ocean_proximityINLAND	-3.425034e+04	1.103430e+04	-3.1039900	0.0019950
ocean_proximityNEAR BAY	2.529785e+02	1.103672e+04	0.0229215	0.9817201
ocean_proximityNEAR OCEAN	1.617996e+04	9.611068e+03	1.6834717	0.0927800
total_bedrooms	5.598007e+01	3.619386e+01	1.5466734	0.1224450
housing_median_age	1.091200e+03	2.752337e+02	3.9646320	0.0000819
population	-4.479346e+01	7.330620e+00	-6.1104608	0.0000000
longitude	-3.468939e+04	6.514864e+03	-5.3246526	0.0000001
latitude	-3.279626e+04	6.658184e+03	-4.9257069	0.0000011
total_rooms	1.641314e+00	4.511525e+00	0.3638047	0.7161261
households	7.105203e+01	4.096127e+01	1.7346149	0.0832989

A.1.2 Model 2

I followed this by constructing a second model, Model 2, using a criterion called the Akaike Information Criterion (AIC). AIC helps in selecting a simpler, yet effective model. Here, I chose three key variables: 'population', 'median income', and their average values.

Table 2: Summary of Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-2.822476e+04	62072.816239	-0.4547040	0.6494788
latitude	1.497337e+03	1724.739611	0.8681526	0.3856414
housing_median_age	1.272312e+03	278.989846	4.5604237	0.0000061
total_rooms	-4.160244e-01	4.591489	-0.0906077	0.9278331
total_bedrooms	5.011335e+01	36.954065	1.3560984	0.1755537
population	-4.274427e+01	7.477735	-5.7162056	0.0000000
households	8.467292e+01	41.759329	2.0276407	0.0430173
median_income	4.058890e+04	2123.684266	19.1124916	0.0000000
ocean_proximityINLAND	-7.344286e+04	8397.113908	-8.7462023	0.0000000
ocean_proximityNEAR BAY	5.110261e+03	11235.183575	0.4548444	0.6493778
ocean_proximityNEAR OCEAN	2.969675e+04	9468.852143	3.1362561	0.0017909

Table 3: Significant Factors Determined by AIC

Significant_Factors_Using_AIC
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
ocean_proximity

A.1.3 Model 3

Next, I developed a third model, Model 3, using another criterion called the Bayesian Information Criterion (BIC). BIC is similar to AIC but usually favors simpler models when comparing two models with a similar level of accuracy. Model 3 was selected because it had the best (or nearly the best) AIC score from my previous models, indicating it was likely the most reliable.

Table 4: Summary of Model 3

term	estimate	std.error	statistic	p.value
(Intercept)	-2.822476e+04	62072.816239	-0.4547040	0.6494788
latitude	1.497337e+03	1724.739611	0.8681526	0.3856414
housing_median_age	1.272312e+03	278.989846	4.5604237	0.0000061
total_rooms	-4.160244e-01	4.591489	-0.0906077	0.9278331
total_bedrooms	5.011335e+01	36.954065	1.3560984	0.1755537
population	-4.274427e+01	7.477735	-5.7162056	0.0000000
households	8.467292e+01	41.759329	2.0276407	0.0430173
median_income	4.058890e+04	2123.684266	19.1124916	0.0000000
ocean_proximityINLAND	-7.344286e+04	8397.113908	-8.7462023	0.0000000
ocean_proximityNEAR BAY	5.110261e+03	11235.183575	0.4548444	0.6493778
ocean_proximityNEAR OCEAN	2.969675e+04	9468.852143	3.1362561	0.0017909

Table 5: Significant Factors Determined by BIC

Significant_Factors_Using_BIC
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
ocean_proximity

A.1.4 Model 4

After establishing these models, I focused on Model 3 to further refine it into Model 4. This involved validating Model 3's predictions and ensuring it was the best choice before proceeding. Validation here means testing Model 3 to make sure it accurately predicts outcomes based on new data not used in the model's creation.

Table 6: Best Lambda Value from LASSO Model

Best_Lambda
3644.423

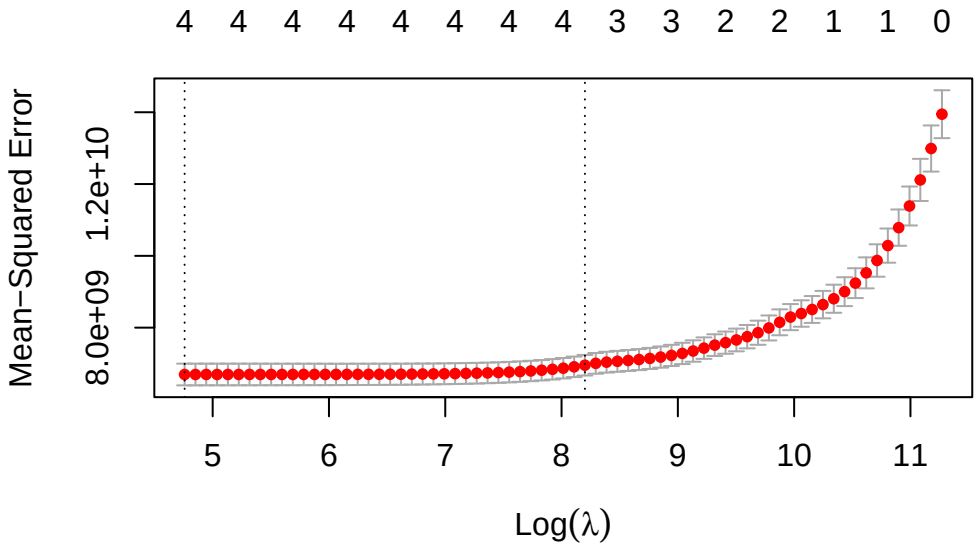

Table 7: Selected Variables

x
housing_median_age
total_bedrooms
population
median_income

In the final stages, I took some variables from lasso selection and used them to build a final model which I save it in my script.

Table 8: Significant Factors Determined by Lasso selection

Significant_Factors_Using_Lasso_selection
housing_median_age
total_bedrooms
median_income

A.2 Model justification

From the figure 10, it can be observed that Residual plots have no tendency to bend, which means that it is conformal to the linear.Although there are a large number of points aggregated together into a cluster, this is due to the large size of the data set, and the data as a whole is still independent. For QQ-plots, it obeys linearity.

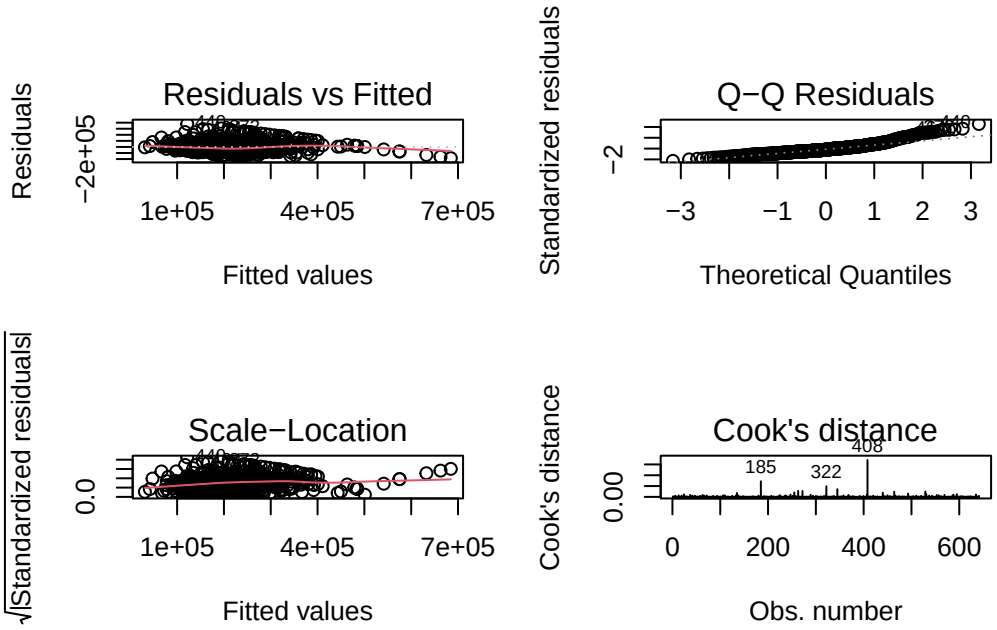

Figure 10: data visualization on 4 model for house prices

I then performed a statistical test known as the partial F-test on this model. This test checks if adding or removing variables significantly changes the model's predictions. The test yielded a p-value of 2.2e-16, which is much smaller than the typical threshold of 0.05. A small p-value like this strongly suggests that the model is statistically significant, affirming our choice of Model 3 based on its variables.

A.3 Final Model Details

Table 9: P value by Using Partial F test

P_value_by_Using_Partial_F_test
2.2e-16

Table 10: Coefficients for Final Model

	Factor	Coefficient
Intercept	Intercept	-41899.52
total_bedrooms	total_bedrooms	35.71
housing_median_age	housing_median_age	2146.32
median_income	median_income	43447.54

References

Auguie, Baptiste. 2022. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://cindyfang70.github.io/gridExtra/.

Banerjee, Abhijit V, and Esther Duflo. 2019. "The Experimental Approach to Development Economics." *Journal of Economic Perspectives* 33 (4): 3–28. https://doi.org/10.1257/jep.33.4.3.

Brock, G. W. 2016. "The First American Women Architects." *Journal of Architectural Education* 70 (2): 261–71. https://doi.org/10.1080/10464883.2016.1151253.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Lutz, C. 2007. "The New American Poverty." *The Atlantic* 299 (3): 43–53.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, Simon Couch, and Max Kuhn. 2019. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.