How do geographical locations, housing characteristics, demographics, and income levels affect the housing value in California?

Xingyu (Megan) Wang

June 17, 2024

Table of Contents

Introduction

California's real estate market is one of the most dynamic and diverse in the United States, characterized by wide regional variations and a wide range of influencing factors. Also, the value of a house is a crucial consideration for both buyers and sellers due to its significant financial implications. Therefore, understanding the determinants of housing value is critical for homeowners, potential buyers, investors, and policymakers. In California, several key factors play a key role in shaping real estate values, such as location, housing characteristics, demographics and income levels. Each of these factors interacts in complex ways to influence the housing market and drive changes in house prices in different regions. Furthermore, there are about 60,000 academic papers on Google Scholar discussing this topic, which demonstrates how important and interesting the question is.

Sirmans et al. (2006) use meta-regression to discuss that the age and bathroom coefficients are affected by geographic location, but not income. While the bedroom coefficient is not sensitive by either of them. This result tells us some variables such as house age, number of bathrooms, and location are not completely independent variables. Moreover, it is reasonable to think that the closer the beach is, the higher the house price will be; since seaside houses are far from urban pollution and have higher entertainment value. Conroy and Milosch (2011) suggest that proximity to the coast has a great positive effect on the value of houses. However, this effect is not linear in distance from the coast, which means the further a house is from the coast, the smaller the percentage of the decline in value. In addition, Gyourko et al. (2010) mention that there is a correlation between house price growth and income growth. Their distributions look the same with wide distribution and right skewed.

Overall, in this project, I am going to research how geographic locations, housing characteristics, demographics, and income levels affect the housing value in California. Especially study how they interplay with each other. The outcome will be the median house value, and there are five variables, which are housing median age, total bedrooms, population, median income, and ocean proximity.

Methods

Data

This basic and original dataset reference is extracted from Kaggle, which contains median house prices for California districts. It has 20640 observations and 10 columns, which include 7 predictors that I am interested in, housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity. It is worth noting that ocean proximity is a dummy variable, and it is divided into inland, <1h ocean, near bay, near ocean, and island. After dropping missing values and useless columns, we have 20433 observations left. In this project, I randomly select 2000 observations from the dataset. Also, the number of median incomes is measured in tens of thousands of US Dollars in the original dataset. To be consistent with the median house value units, I multiplied the median income by 10000.

Model Validation

To validate whether a model is effective or not, it is essential to randomly split a dataset into two parts: a training set and a testing set (70-30). The training set is used to build and train the model, while the testing set is used to evaluate the model's performance on unseen data. After splitting, the training set has 1400 observations, and the testing set has 600 observations.

Model Diagnostics

After cleaning and setting up data, I will create a linear model of all predictors without any transformations first. Then, there are four assumptions that we need to check for the linear model. The first one is linearity, by plotting scatter plots to see whether pairs of predictors and responses look in a linear shape. The second one is constant variance, which means the residual versus fitted plots should not have any pattern, such as fanning and clustering, but spread equally. The last two assumptions are errors that should be uncorrelated and normally distributed, by checking with the Q-Q plot and scale-location plot.

Furthermore, there are two conditions we need to follow as well. One is that the conditional mean response is a single function of a linear combination of the predictors; another is that the conditional mean of each predictor is a linear function with another predictor. For the first condition, I will plot response versus fitted values to check. Condition 2 will be checked by plotting all pairs of predictors and inspecting whether they follow linear relations. If the model violates one or more assumptions above, I will correct it through transformation for either the response, predictors, or both, by using the Box-Cox or power transform.

In addition, combining Residuals vs Leverage plot and a criterion of $[-4,\ 4]$ for the standardized residual to identify outliers and remove them. Afterwards, using the Variance Inflation Factor (VIF) to analyze the degree of collinearity between independent variables is essential since there exists correlations between predictors. Collinearity results in unreliable model results. If VIF is greater than 5, the predictors are considered serious collinearity problems, so I need to remove one or more of them and compare the models. For those not significant, using ANOVA to see if I should remove those variables.

To get a final model, compared with Adjusted R^2, Residual standard error, F-statistic, and p-value are essential. Also, checking each model's Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare the models' fitness. The best-fitted linear model should be a higher Adjusted R^2, lower Residual standard error, higher F-statistic, significant p-value, and lower AIC and BIC values. Especially, it should still follow the criteria before.

Finally, using test set to ensure that the model generalizes well to unseen data.

Results

The sample contains 2000 observations and 7 variables randomly selected from the original cleaning dataset and split into a training set (1400) and a testing set (600). According to the histogram and scatter plots from the training set below, we found that the initial histogram of the response variable, median house value, indicated a right-skewed distribution, suggesting potential issues for subsequent analyses that assume normality. To address this, I will apply a logarithmic transformation to the median house value later, resulting in a distribution that is

more symmetric and closer to normality. Additionally, scatter plots of the predictor variables (housing median age, total rooms, total bedrooms, population, households, and median income) and a boxplot of ocean proximity against the transformed median house value demonstrated linear relationships, confirming the appropriateness of linear regression models for our analysis.

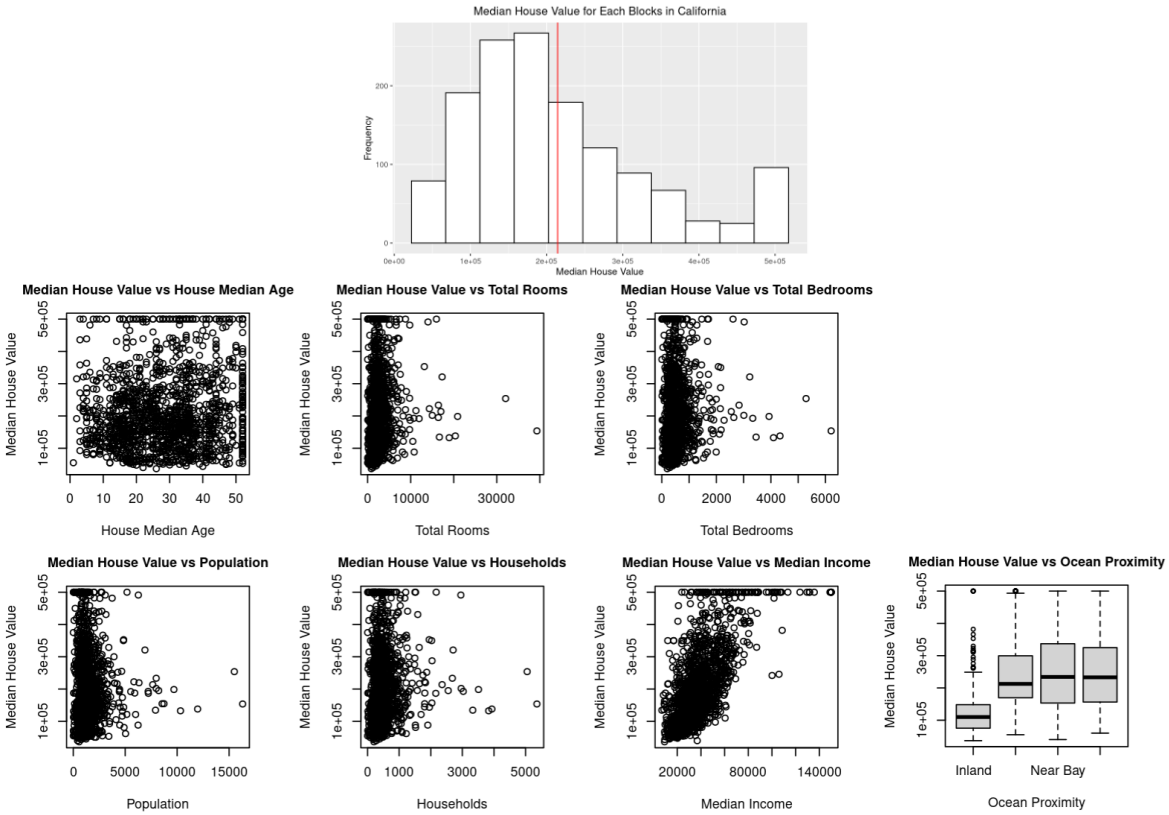


Figure 1. Response Histogram and Predictors Plots

The initial model includes all predictors without any transformations. We could observe this model did not adequately meet the assumptions of linear regression, particularly linearity and homoscedasticity. Also, after calculating the Box-Cox power transformations (refer to Appendix Figure 1.), I decided to apply logarithmic transformations to the median house value, total rooms, total bedrooms, population, households, and median income, then fit a model again. Furthermore, some points have large leverage or higher standardized residuals that need to be removed (refer to Appendix Figure 2.). After fitting the transformed model, the improvement was evident.

Referring to Figure 2, which illustrates the comparison between the initial and transformed models, we can obviously see that the second model is better than the first one. The first two graphs in Figure 2 assess the linearity condition. The right-hand side plot for the transformed model displays a more linear relationship, indicating a better fit. In addition, the second residuals vs fitted plot, Q-Q residuals plot, and scale-location plot all demonstrate that the transformed model corrects the violations before.

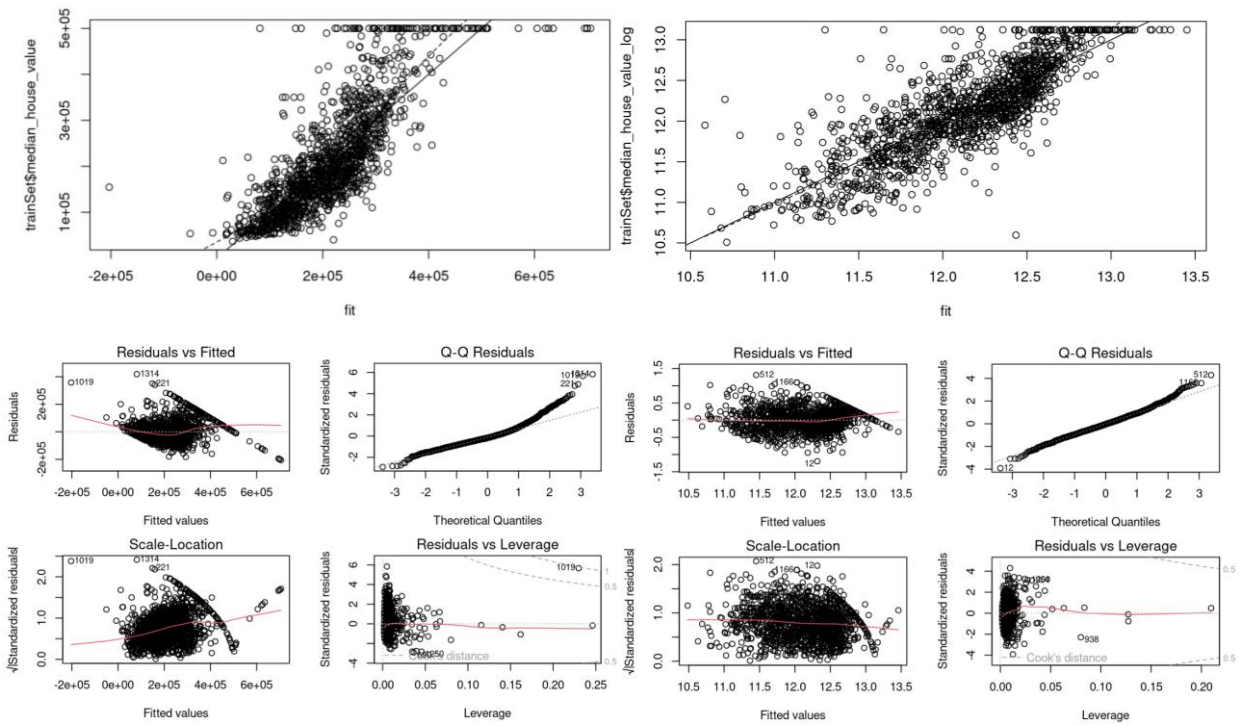


Figure 2. Comparison of the 1st and 2nd Models

However, the Variance Inflation Factor (VIF) values for the set of predictors indicate significant multicollinearity among total rooms, total bedrooms, and households. Hence, systematically removing one or more of the problematic variables and comparing the performance of each resulting model could help us find the best-fitting model.

Model	**Adjusted R^2**	**Residual Standard Error**	**AIC**	**BIC**	**Multicollinearity**
Transform model	0.7026	0.3074	679.2637	736.9109	Yes
Remove households	0.6922	0.3127	726.241	778.6475	Yes
Remove total bedrooms	0.6997	0.3089	691.8762	744.2827	No (high)
Remove total rooms	0.6979	0.3098	700.2582	752.6647	Yes
Remove total bedrooms + households	0.6676	0.3249	832.5172	879.683	No
Remove total rooms + households	0.6879	0.3149	744.5278	791.6936	No
Remove total bedrooms + total rooms	0.6979	0.3098	699.1125	746.2783	No
Remove total bedrooms + total rooms + households	0.6484	0.3342	909.6935	951.6187	No

Table 1. Summary Information of Each Model

According to Table 1, the best model is removing total bedrooms and total rooms with higher Adjusted R^2 and lower residual standard error, AIC, and BIC. Again, we need to re-check whether it follows the assumptions and conditions.

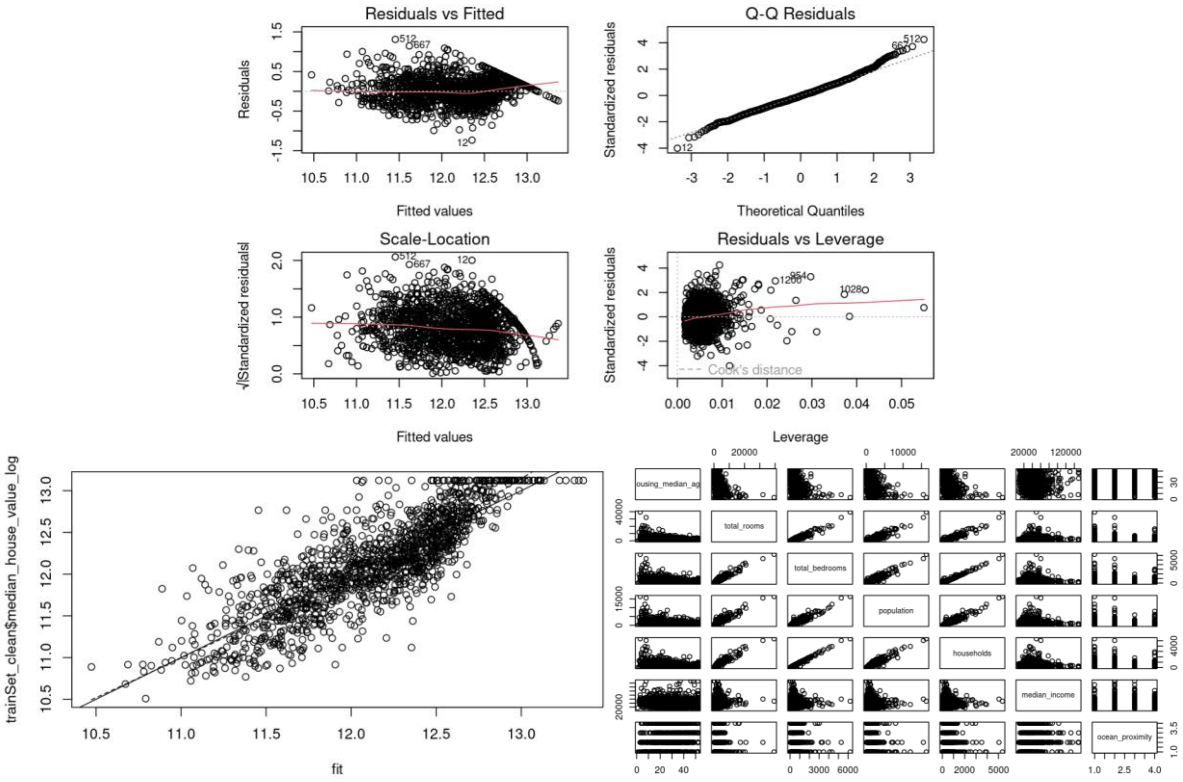


Figure 3. Check Assumptions and Conditions for Model 6

Figure 3 above presents several diagnostic plots that assess the validity of the regression model assumptions. The residuals vs fitted plot shows no discernible patterns, indicating that the residuals are randomly distributed. The Q-Q plot demonstrates linearity, suggesting that the residuals follow a normal distribution. The scale-location plot shows constant variance (homoscedasticity), as the residuals appear evenly spread across all levels of the fitted values. And there are no outliers in the leverage plot. Moreover, the last two plots display pairwise relationships between each predictor seem linear, and the points are evenly scattered around the identity line. Hence, this model satisfies all conditions.

Model Validation

To validate the final model, I used the testing set, which contains 600 observations. The model's performance was evaluated using several metrics.

Model	**Adjusted R^2**	**Residual Standard Error**	**AIC**	**BIC**
Train	0.6979	0.3098	699.1125	746.2783
Test	0.6331	0.3519	459.5144	499.0867

Table 2. Compare Results from Training Data and Test Data

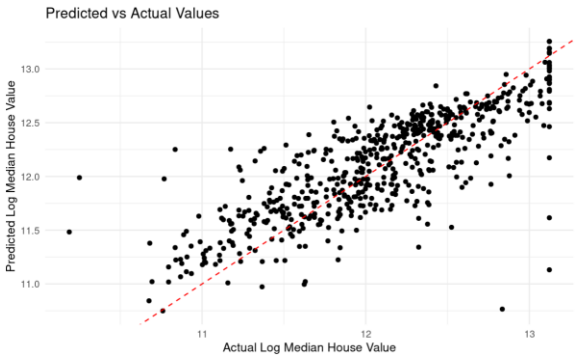


Figure 4. Predicted vs Actual Values

From Table 2 above, we can see some results from training data and testing data. The comparison between the training and testing sets' metrics shows that the model maintains good predictive performance, with only a slight decrease in Adjusted R^2 and a reasonable increase in Residual Standard Error. In addition, Figure 4 presents the scatter plot of predicted versus actual log median house values. The points roughly lie close to the red dashed identity line, indicating a good prediction. Overall, these results indicate that the final model generalizes well to the testing data.

Discussion

The final model was developed to predict the log-transformed median house value in California using various predictors, including housing median age, log-transformed total rooms, log-transformed total bedrooms, log-transformed population, log-transformed households, log-transformed median income, and categorical ocean proximity. The results from both the training and testing sets demonstrate that the model performs well, with an Adjusted R^2 value of 0.6979 for the training set and 0.6331 for the testing set, indicating a good fit and reasonable generalization to new data.

The research question aimed to understand how each factor influences the median house value in California. The final model effectively answers this question by identifying significant predictors and quantifying their relationships with the median house value. We can conclude that the housing characteristics are not so significant based on this dataset since we have a higher Adjusted R^2 and lower AIC and BIC values when removing them. Also, refer to Appendix Figure 3, we can observe that median income has the most substantial positive impact on median house values. However, the negative coefficient (-0.47) indicates that higher population densities are associated with lower median house values. Other predictors have a more or less positive effect on the median house values.

This project is useful for the government or stakeholders. Understanding the factors that have an impact on housing prices can be better planned and formulated. In addition, developers can better develop according to demand, and buyers and sellers can also make effective investments. However, this model still has some limitations. I only use one dataset to do this research, but there are many other variables that might influence the median house value, such as school quality and crime rate. Moreover, even though I removed the total rooms and total bedrooms to reduce the multicollinearity issue, it does not mean they do not impact the house value.

References

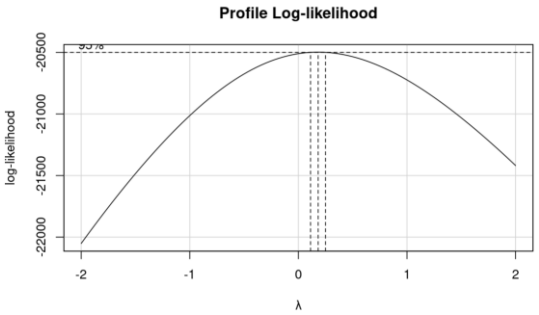
Conroy, S.J., Milosch, J.L. (2011). An Estimation of the Coastal Premium for Residential Housing Prices in San Diego County. J Real Estate Finan Econ 42, 211–228. https://doi.org/10.1007/s11146-009-9195-x

Gyourko, J., Mayer, C., & Sinai, T. (2010). Dispersion in house price and income growth across markets: Facts and theories. In Agglomeration economics (pp. 67-104). University of Chicago Press.

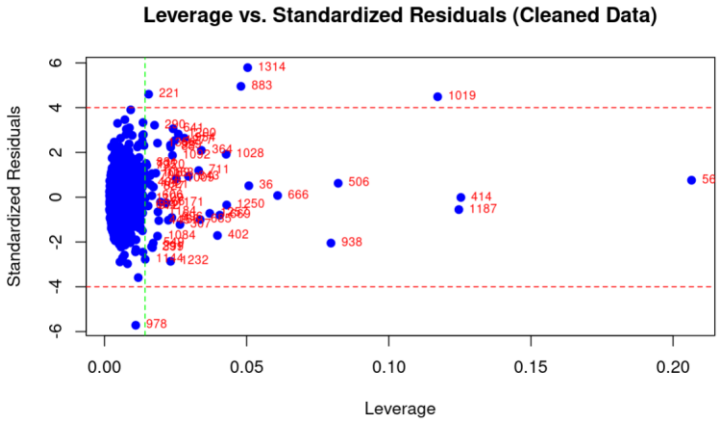
Sirmans, G.S., MacDonald, L., Macpherson, D.A. et al. (2006). The Value of Housing Characteristics: A Meta Analysis. J Real Estate Finan Econ 33, 215–240. https://doi.org/10.1007/s11146-006-9983-5

Appendix

```
bcPower Transformations to Multinormality
                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
housing_median_age     0.7858        0.79       0.6984       0.8731
total_rooms            0.2239        0.22       0.2016       0.2462
total_bedrooms         0.2351        0.24       0.2120       0.2583
population             0.2121        0.21       0.1839       0.2402
households             0.2410        0.24       0.2168       0.2653
median_income          0.2353        0.24       0.1787       0.2919
median_house_value     0.1870        0.19       0.1137       0.2604

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
bcPower Transformations to Multinormality
                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
housing_median_age     0.7854        0.79       0.6969       0.8740
total_rooms            0.2252        0.23       0.2025       0.2479
total_bedrooms         0.2363        0.24       0.2123       0.2604
population             0.2150        0.22       0.1865       0.2435
households             0.2424        0.24       0.2171       0.2676
median_income          0.1410        0.14       0.0713       0.2107

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```



Appendix Figure 1. Box-Cox Power Transformation and Log-likelihood



Appendix Figure 2. Outliers for Cleaning Dataset

```
Call:
lm(formula = median_house_value_log ~ housing_median_age + population_log +
    households_log + median_income_log + ocean_proximity, data = trainSet_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23481 -0.19525 -0.01278  0.19360  1.31068

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.5219406  0.2225172  20.322  < 2e-16 ***
housing_median_age        0.0042801  0.0007356   5.818 7.38e-09 ***
population_log           -0.4699723  0.0334762 -14.039  < 2e-16 ***
households_log            0.5170184  0.0342166  15.110  < 2e-16 ***
median_income_log         0.6992923  0.0187779  37.240  < 2e-16 ***
ocean_proximity<1H OCEAN  0.5238637  0.0211212  24.803  < 2e-16 ***
ocean_proximityNEAR BAY   0.4319352  0.0312431  13.825  < 2e-16 ***
ocean_proximityNEAR OCEAN 0.5201079  0.0280503  18.542  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3098 on 1387 degrees of freedom
Multiple R-squared:  0.6994,    Adjusted R-squared:  0.6979
F-statistic: 461.1 on 7 and 1387 DF,  p-value: < 2.2e-16
```

Appendix Figure 3. Summary for Final Model