final project

2024-03-01

Introduction

The dataset chosen for this research project is the California Housing Prices dataset, sourced from Kaggle.This data is from around 20000 blocks of housing in California in 1997, providing a detailed snapshot of the housing market in various Californian neighborhoods. These variables include geographical coordinates (longitude and latitude), housing median age, total rooms, total bedrooms, population, households, median income, median House value and ocean proximity. This comprehensive dataset forms the foundation of the study, containing the multifaceted factors influencing housing prices in California.

This research project aims to explore how a combination of socioeconomic and demographic factors influence house values in California, and whether a predictive model can be developed for accurate housing price forecasts. The goal is to develop a linear regression model that not only predicts house values within a block with high accuracy but also provides a holistic understanding of the complex interplay of various factors shaping California's real estate market.

Theory and hypothesis

The core of this research revolves around two key questions: "What factors influence housing prices in California?" and "Is it feasible to construct a linear model for this purpose?" These overarching inquiries lay the groundwork for a more specific exploration: "Which factors exert the most significant impact on housing values in California?" Pursuing answers to these questions is imperative for gaining a comprehensive understanding of the intricate dynamics of the economics and real estate market.

The hypothesis is that areas with higher median incomes, lower population densities, and closer ocean proximity will have higher median house values, reflecting the purchasing power, the preferences of residents and the desirability of ocean or big city. To test this, the study will analyze variables such as median income, population, housing median age, distance to city and ocean proximity. The primary outcome variable is the median house value in each neighborhood / block.

The research expects to find positive correlations between house values and factors like median income, newer housing, and ocean proximity, and negative correlations with higher population densities.

Based on this theoretical framework, two hypotheses are considered:

H_1: There is a positive correlation between median house values and (housing conditions and socioeconomic and demographic factors in the California housing market).

H_2: There is a negative correlation between median house values and (housing conditions and socioeconomic and demographic factors in the California housing market).

These hypotheses will be examined through a linear regression model, which aims to predict house values within a block based on these variables. This model is not just a predictive tool but also a means to understand the interplay of various factors in shaping the real estate market in California.

Data and visual analysis

Data

The dataset initially comprises 20,640 rows and 17 columns. Following cleaning and transformation, its dimensions are refined to 19,675 rows and 18 columns. This denotes a loss of approximately 4.5% of the dataset during the cleaning process. The meticulous cleaning and transformation procedures are executed to optimize the dataset for precise analysis.

Elimination of Outliers: To enhance the robustness of the predictive model and ensure more accurate prediction plots, housing blocks with median housing values exceeding $500,000 were excluded. This step was necessary to avoid truncation issues in the actual vs prediction plots.

Variable Aggregation: To reduce the complexity of the model and avoid overfitting, distances to major Californian cities - San Diego, San Francisco, Los Angeles, and San Jose - were combined into a single variable. This combination helps in simplifying the model without significantly compromising its accuracy.

Variable Conversion: The 'ocean proximity' and 'City' columns were transformed into numeric values. This conversion was crucial for calculations in the linear regression model.

The research design being observational underscores the importance of identifying and controlling for potential confounding factors. In this context, factors such as economic conditions, local policies, and demographic shifts may serve as confounders. Effectively controlling for these variables is essential to minimize their influence on the outcomes.

Table 1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Median	Max
Median_House_Value	192,477.90	97,711.51	14,999	173,800	500,000
Median_Income	3.68	1.57	0.50	3.45	15.00
Median_Age	28.39	12.51	1	28	52
Tot_Rooms	2,619.76	2,181.35	2	2,111	39,320
Tot_Bedrooms	539.65	422.29	2	436	6,445
Population	1,440.81	1,143.65	3	1,179	35,682
Latitude	35.65	2.15	32.54	34.27	41.95
Longitude	−119.56	2.01	−124.35	−118.50	−114.31
Distance_to_coast	41,920.87	49,832.47	120.68	21,303.29	333,804.70
Distance_to_central_city	1,408,546.00	253,934.40	1,230,026.00	1,316,133.00	3,249,804.00

Table 2: Correlation Matrix

	Median_House_Value	Median_Income	Median_Age	Tot_Rooms	Tot_Bedrooms
Median_House_Value	1	0.64	0.07	0.14	0.08
Median_Income	0.64	1	-0.19	0.22	0.02
Median_Age	0.07	-0.19	1	-0.37	-0.33
Tot_Rooms	0.14	0.22	-0.37	1	0.93
Tot_Bedrooms	0.08	0.02	-0.33	0.93	1
Population	0.01	0.04	-0.30	0.86	0.88
Latitude	-0.15	-0.08	0.01	-0.03	-0.07
Longitude	-0.05	-0.01	-0.10	0.04	0.07
Distance_to_coast	-0.48	-0.23	-0.22	0	-0.02
Distance_to_central_city	-0.31	-0.17	-0.17	0	-0.02

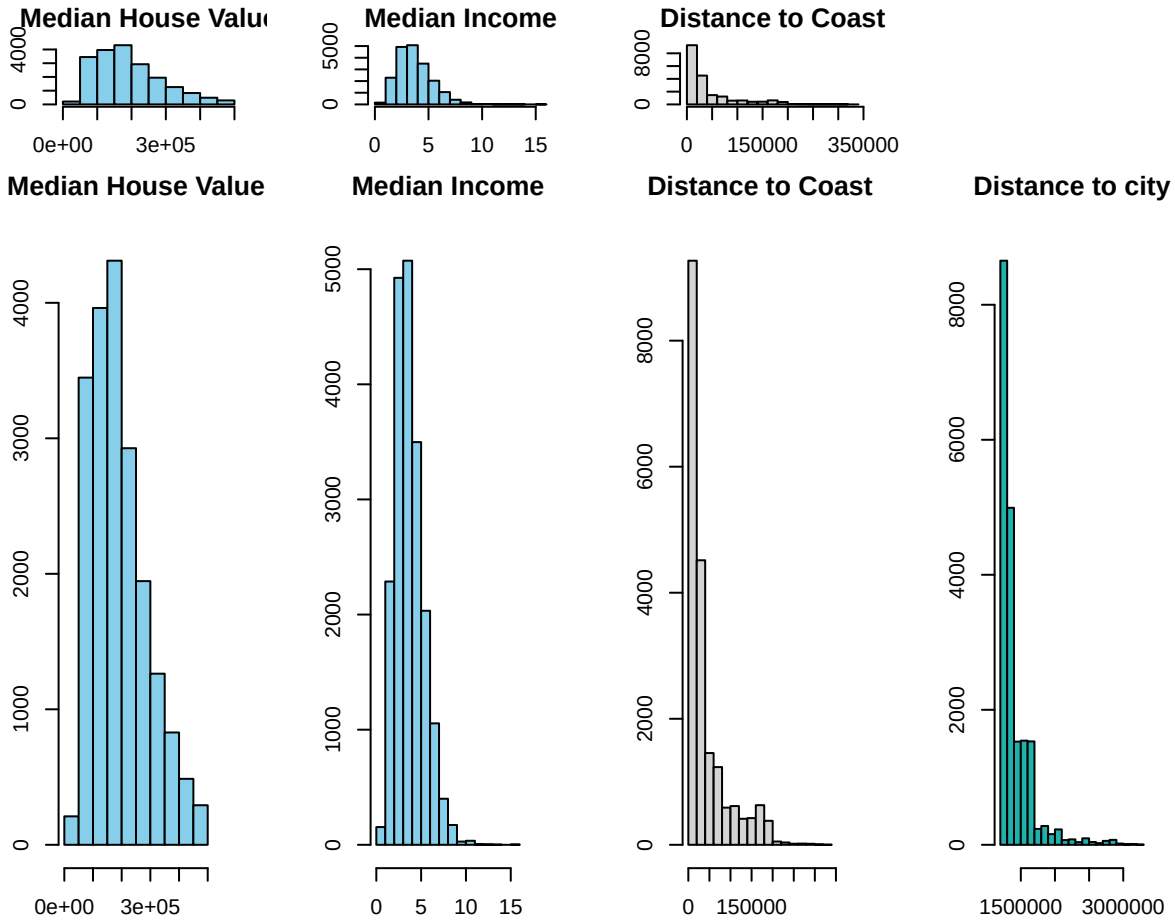
The correlation matrix shows a strong positive correlation (0.64) between median house and median income,

Table 3: Correlation Matrix

	Population	Latitude	Longitude	Distance_to_coast	Distance_to_central_city
Median_House_Value	0.01	-0.15	-0.05	-0.48	-0.31
Median_Income	0.04	-0.08	-0.01	-0.23	-0.17
Median_Age	-0.30	0.01	-0.10	-0.22	-0.17
Tot_Rooms	0.86	-0.03	0.04	0	0
Tot_Bedrooms	0.88	-0.07	0.07	-0.02	-0.02
Population	1	-0.11	0.10	-0.05	-0.09
Latitude	-0.11	1	-0.92	0.31	0.45
Longitude	0.10	-0.92	1	0	-0.24
Distance_to_coast	-0.05	0.31	0	1	0.30
Distance_to_central_city	-0.09	0.45	-0.24	0.30	1

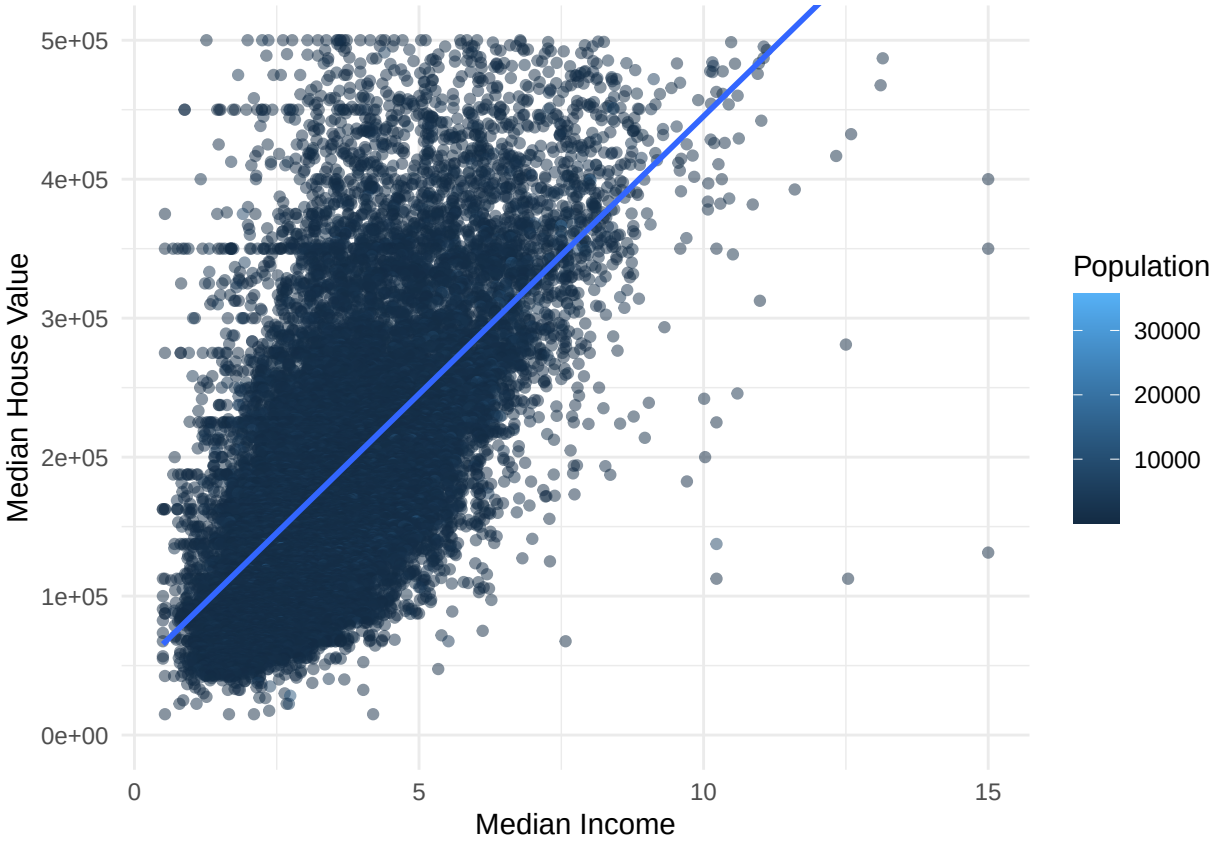
indicating that areas with higher incomes tend to have more expensive houses. There are also significant negative correlations with distance to the coast (-0.48) and central cities (-0.31), suggesting that proximity to these areas generally increases home values. Other variables, such as total number of rooms and total number of bedrooms, show weaker correlations, suggesting that these factors have smaller impact on housing value in California.

Exploratory data analysis

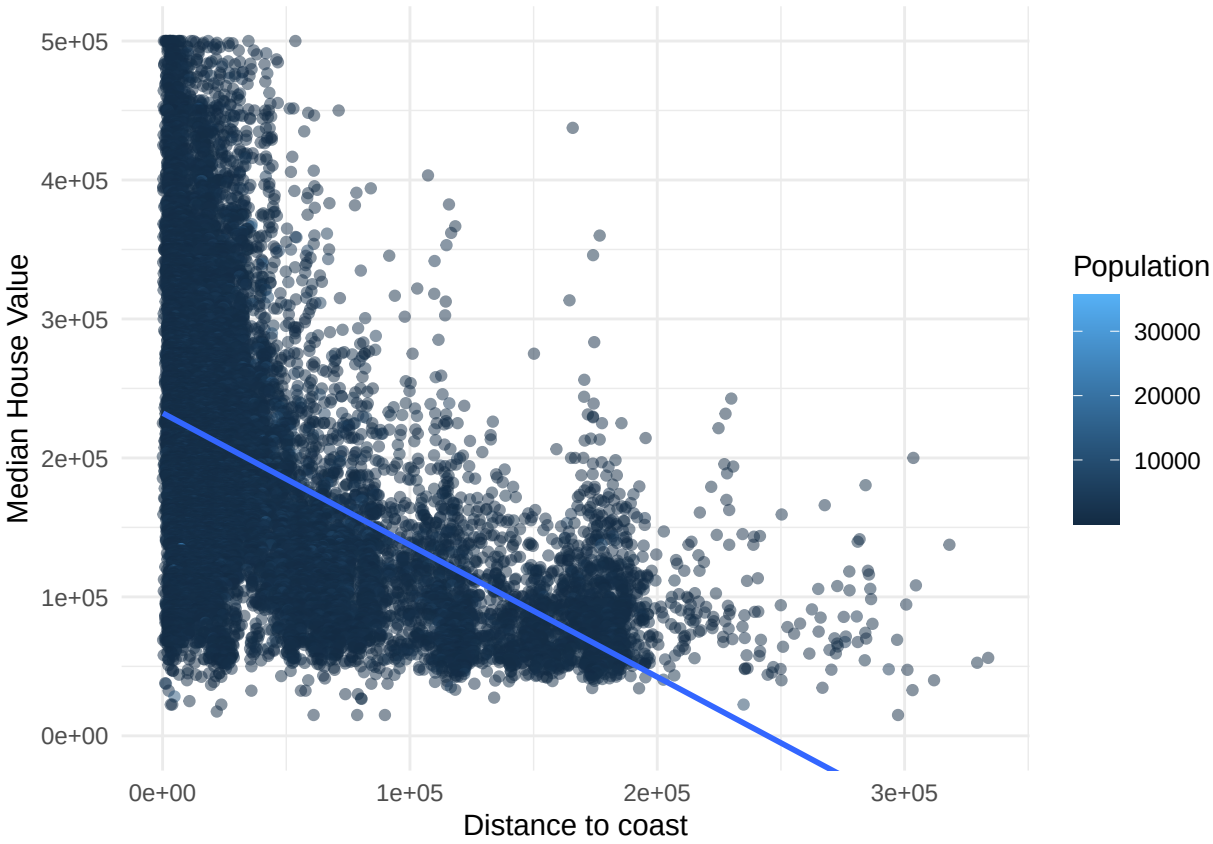


Median house value and median income are left-skewed, indicating that most of the blocks have lower values

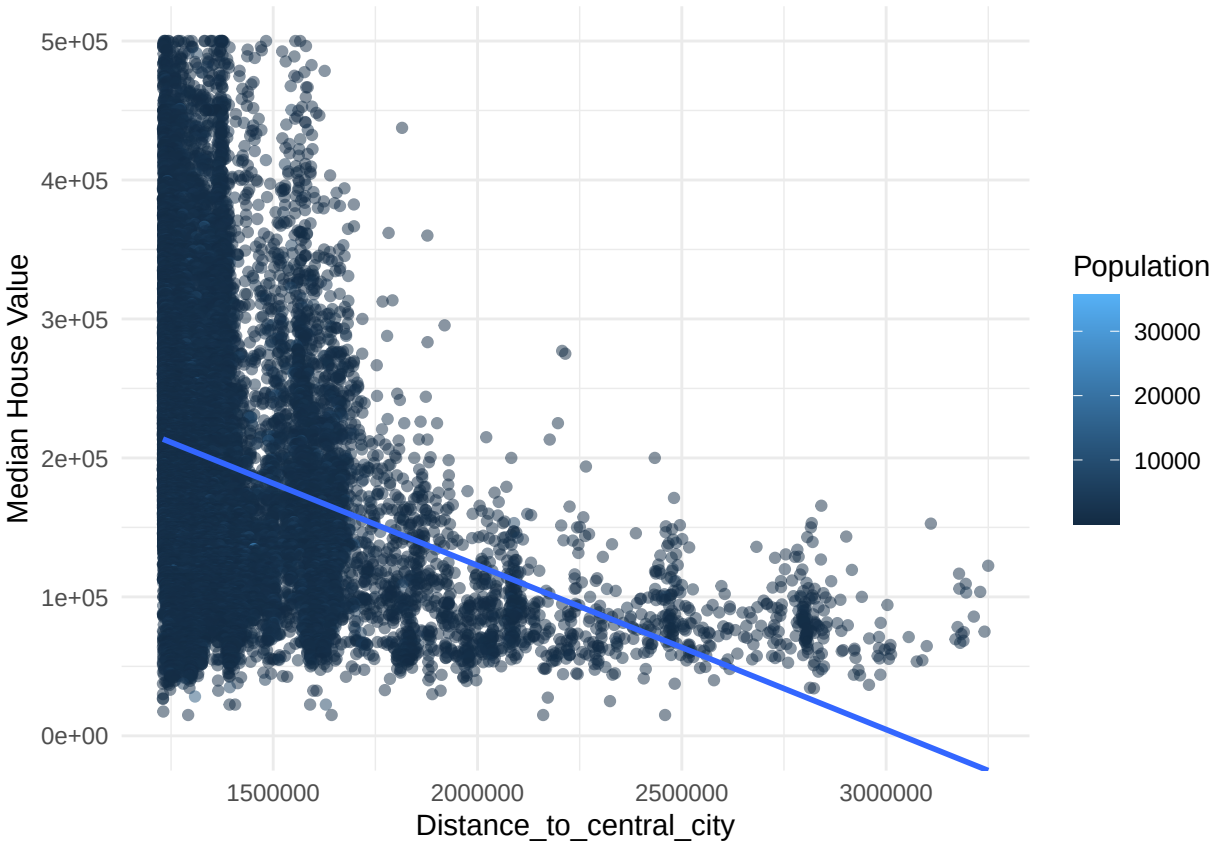
and incomes, while distance to coast and distance to city show that a large number of blocks are situated close to the coast and central city, with very few blocks located farther away.



The plot shows a positive correlation between the two variables, as indicated by the upward trend and the fitted line. Most of the data is clustered around lower median incomes and median house values, with fewer districts having high values of both. There are also few outliers with high median incomes not following the overall trend, indicating variations in the data that might need further investigation.



A negative correlation is suggested by the downward trend of the fitted line, indicating that as the distance to the coast increases, the median house value tends to decrease. Most data points are concentrated at lower distances to the coast, which also correspond to higher median house values, while areas with a smaller population appear to be more spread out.

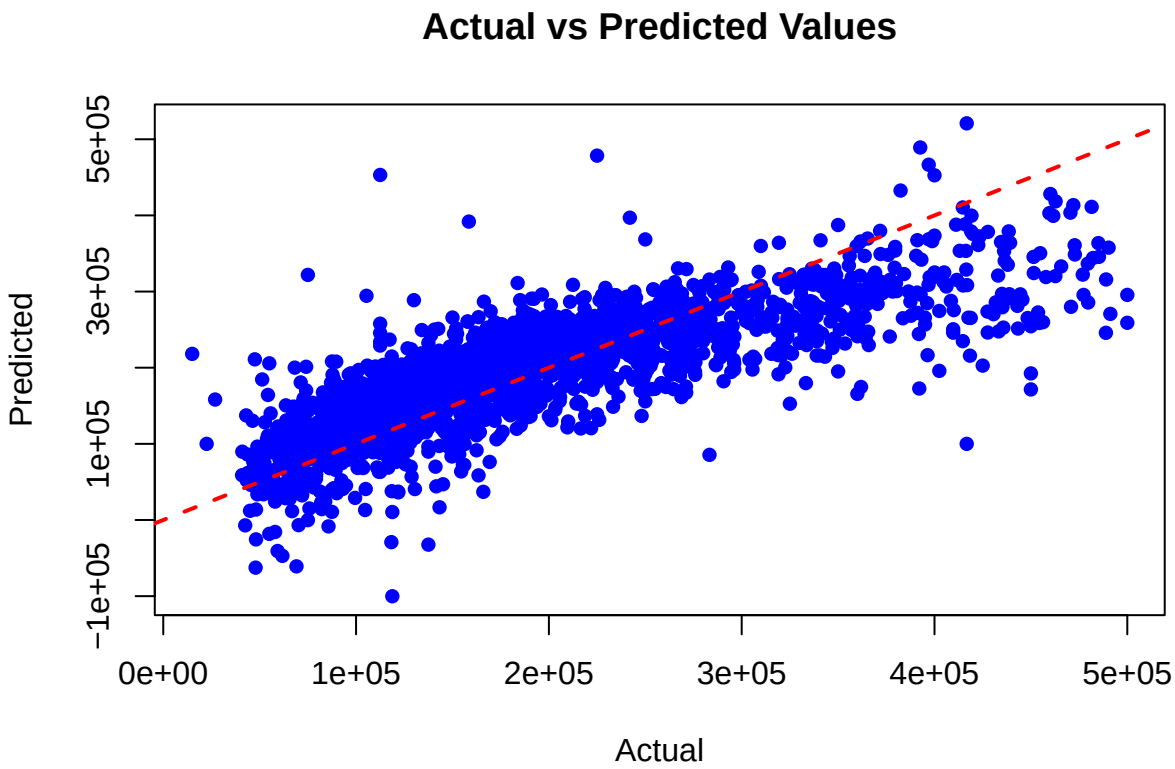


The scatter plot suggests a negative correlation between the distance to the central city and the median house value. As the distance increases, the median house value tends to decrease. The data points concentrated at the lower end of the distance to the central city implies that many of the districts are located relatively close to central urban areas, where housing prices are higher.

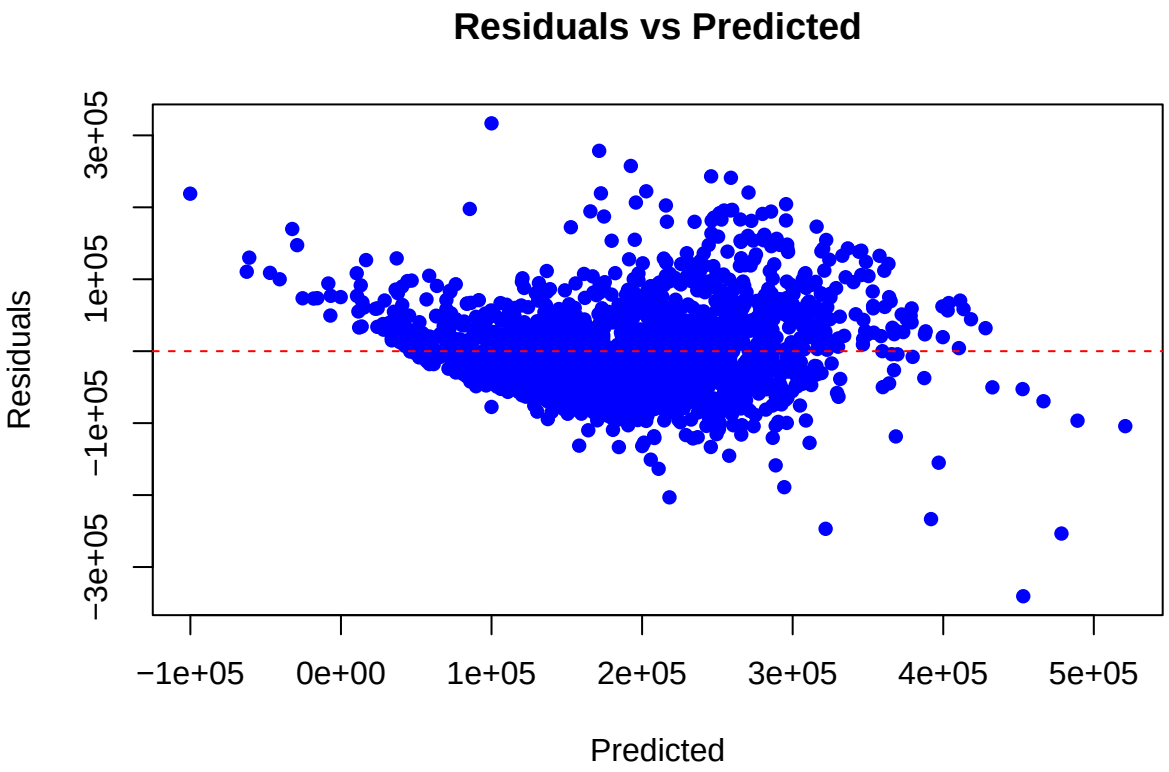
Table 4: Regression analysis of treatment on prejudice

	Median_House_Value			
	Model 1	**Model 2**	**Model 3**	**Model 4**
Median_Income	39987.010***	34866.500***	34094.770***	34151.840***
	(339.918)	(312.226)	(310.132)	(309.762)
Distance_to_coast		−0.694***	−0.631***	−0.633***
		(0.010)	(0.010)	(0.010)
Distance_to_central_city			−0.044***	−0.045***
			(0.002)	(0.002)
Population				−3.180***
				(0.413)
Constant	45457.000***	93381.440***	155953.000***	161893.000***
	(1358.985)	(1391.200)	(3088.163)	(3178.534)
N	19675	19675	19675	19675
R-squared	0.413	0.531	0.543	0.545
Adj. R-squared	0.413	0.531	0.543	0.545

***p < .01; **p < .05; *p < .1



From the scatter plot, we can observe that the majority of predictions appear closely aligned with the actual values, as indicated by the concentration of points around the dashed red line, which represents perfect prediction accuracy. However, the spread of the residuals increases with higher value houses, suggesting the model is less accurate for higher-priced homes. The summary of the model shows significant predictors, including median income and location features, explaining approximately 59.45% of the variance in median house values.



From the plots, we can observe that the residuals are scattered around the horizontal line at zero, which represents little error between the predicted and actual values. While the residuals appear randomly dispersed, there's a visible pattern where variance increases with higher predicted values. This pattern suggests that the model may not be equally accurate across all levels of house values and less accurate at the higher end of the prediction range.

Results discussion

The evidence provided in the regression analysis strongly supports rejecting the null hypothesis H1 in favor of the alternative hypothesis H1. This is indicated by the significance levels ($p < 0.01$) of the coefficients for these variables across all models.

The evidence is consistent across different model specifications, with each additional variable included in the successive models maintaining its significance and the expected sign. Median income consistently shows a positive correlation with median house values across all models,and negative correlation with distance to coast, distance to central city, and population.

Conclusion

In this research, the California Housing Prices dataset was meticulously analyzed to determine the influence of various factors on housing prices within the state. The linear regression model developed for this purpose utilized data from around 20,000 blocks of housing from 1997, encompassing a broad spectrum of variables such as median income, population density, housing age, proximity to the coast and city centers, and more. The analysis was guided by two hypotheses:

H1, proposing a positive correlation between median house values and specific housing conditions and socioeconomic and demographic factors; and H2, suggesting a negative correlation with other variables.

The results from the regression models clearly aligned with the first hypothesis H1, revealing a robust positive correlation between median income and median house values. In contrast, a negative correlation was observed with distance to the coast and central city, as well as population density, which supports the assumption

that while affluence and desirable locations elevate house prices, higher population densities tend to correlate with lower house values. These findings were consistent across multiple model specifications, reinforcing the reliability of the results. The adjusted R-squared values indicated that a significant proportion of the variability in housing prices was accounted for by the predictors included in the model.

Conclusively, this study has successfully illuminated the relationships between key factors and median house values in California. The evidence gathered underpins the theoretical framework and confirms the initial hypotheses, offering valuable insights into the complexities of the real estate market in California. The consistency of the results across different model specifications not only strengthens the validity of the conclusions drawn but also emphasizes the importance of considering a variety of factors when predicting housing prices. Future research could build upon these findings by incorporating additional data, applying more advanced modeling techniques, or exploring the impact of temporal changes, to enhance our understanding of housing market dynamics further.