

CALIFORNIA HOUSING ANALYSIS

Using Data to Identify the Key Drivers of Property Value

Giovanna Calabrese Villanueva

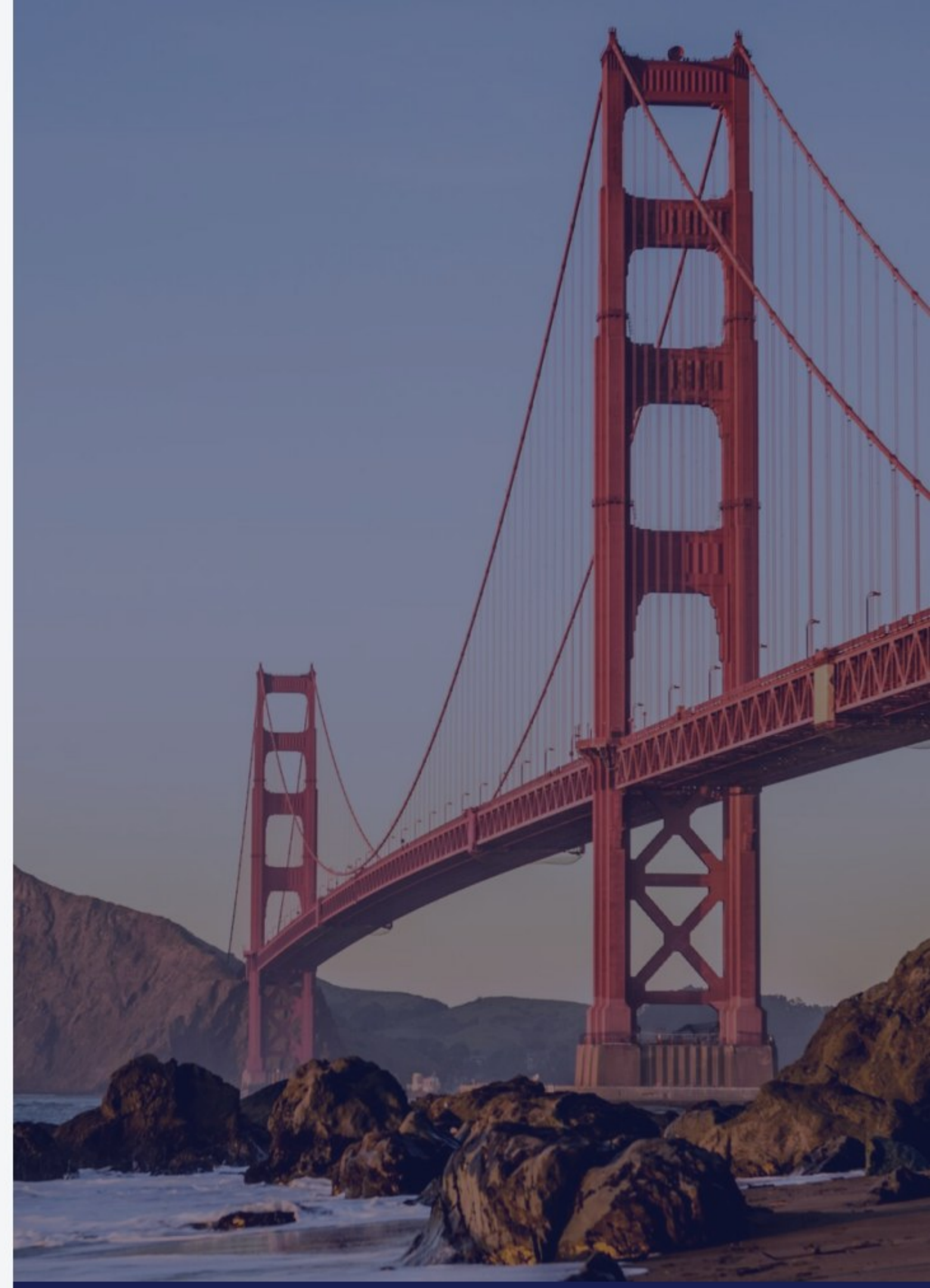


1. Business Problem

A real estate investment agency is considering investment opportunities in California. The agency wants to understand which economic, housing, demographic, and geographic characteristics are associated with higher median housing values.

WHY IS IT IMPORTANT?

- California contains diverse local housing markets
- Investment decisions based only on statewide averages may overlook important differences between areas.



Where should a real estate agency focus when evaluating investment opportunities in California?

- What drives value?
- How important is location?
- Can we predict market value?

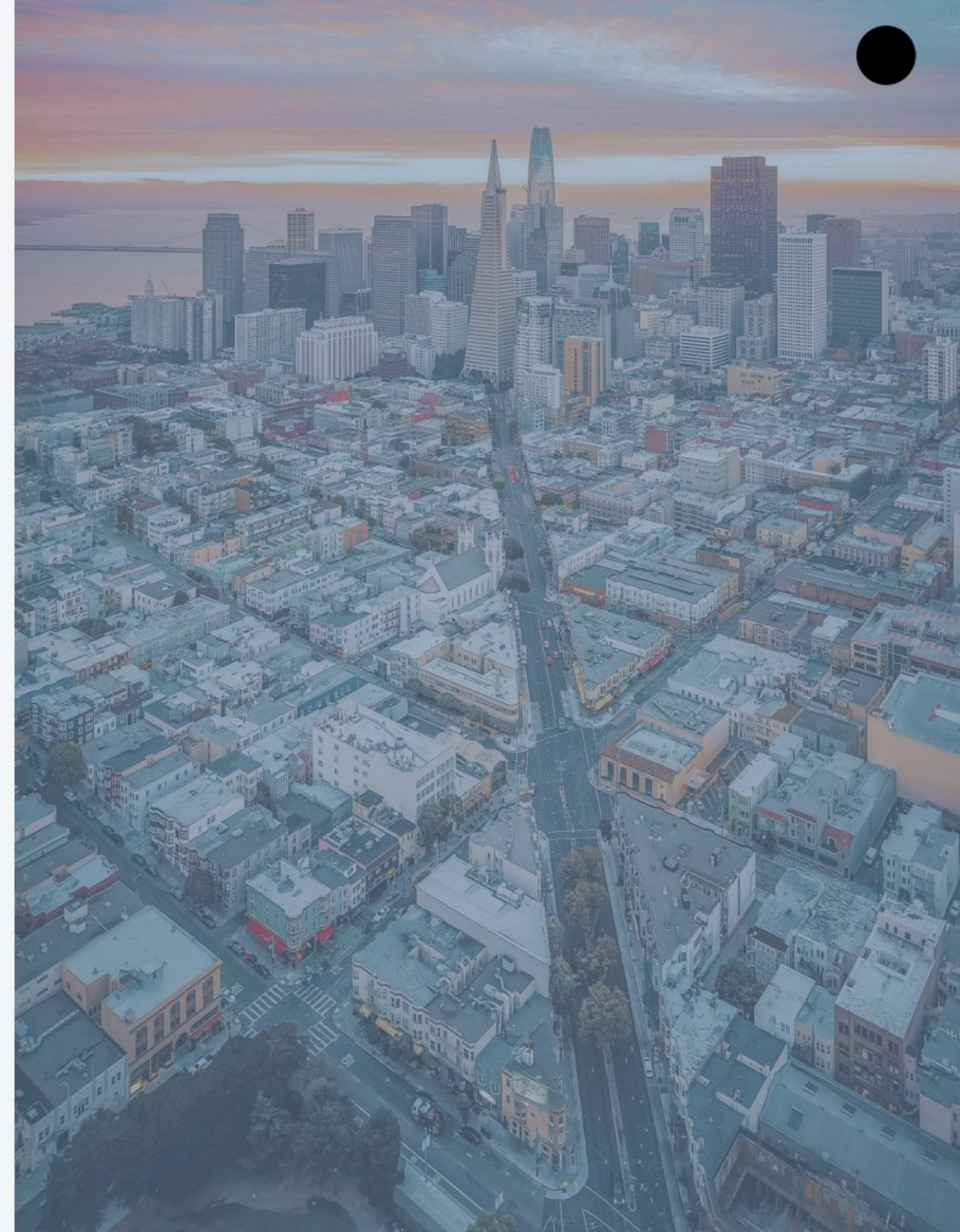
Who benefits from the results?

- Market analysts and investment decision-makers.
- Supports decisions on which California markets should receive further investigation before investment.

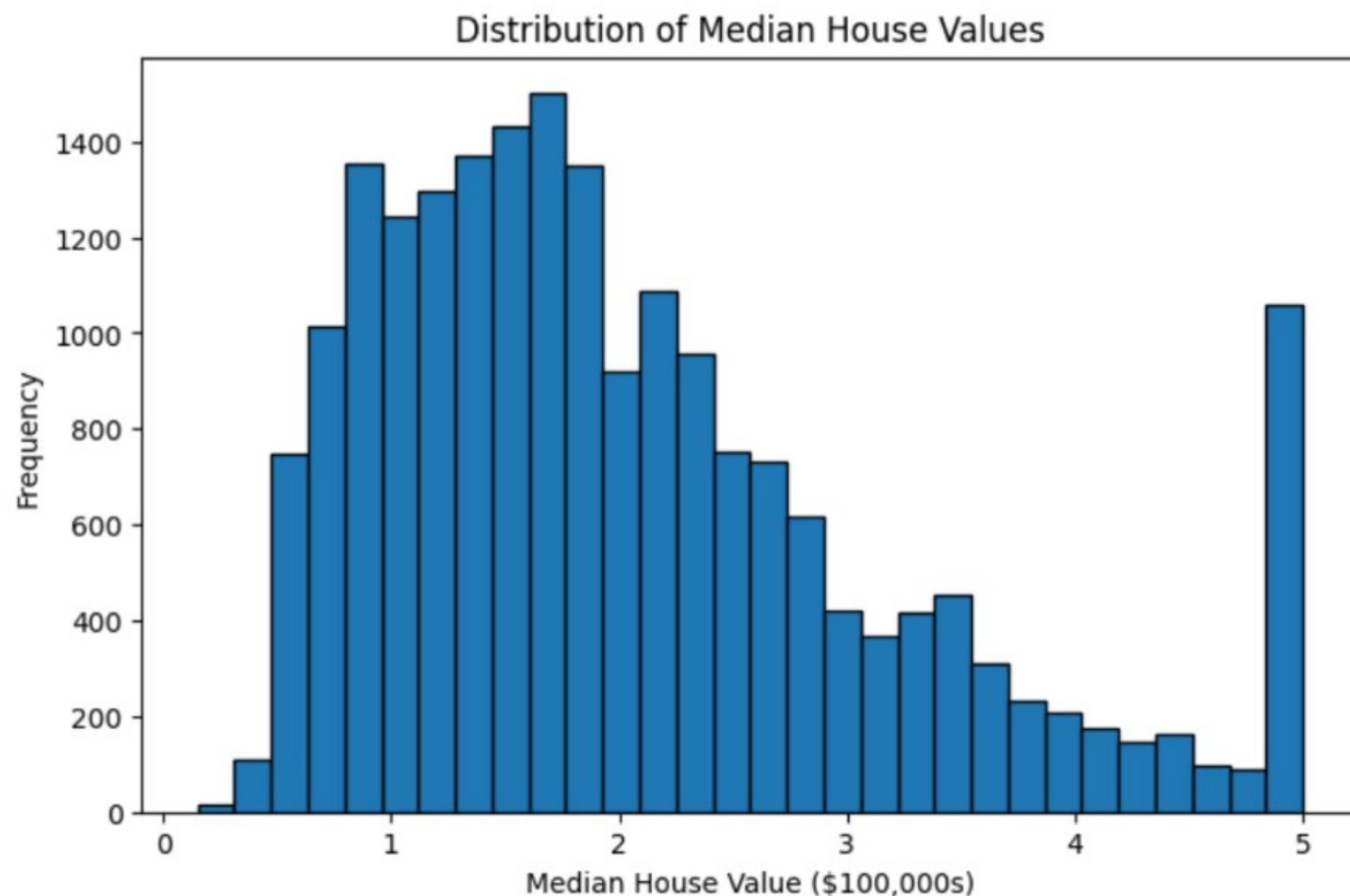


2. Dataset Overview

Item	Description
Dataset name	California Housing Dataset
Number of records	20,640.
Number of input variables	8
Target variable	MedHouseVal
Numerical variables	MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude, MedHouseVal
Categorical variables	None



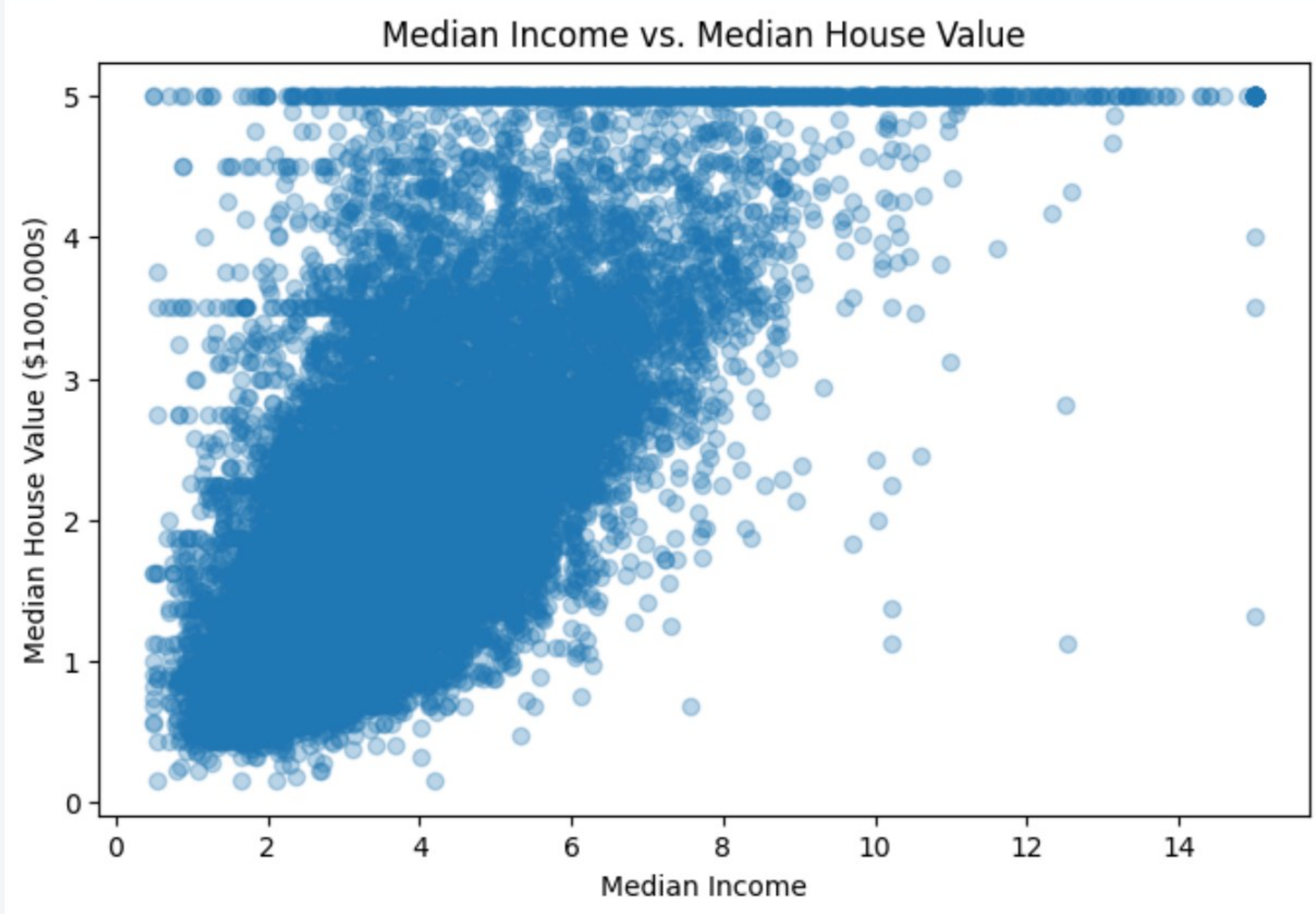
3. California Is Not One Housing Market



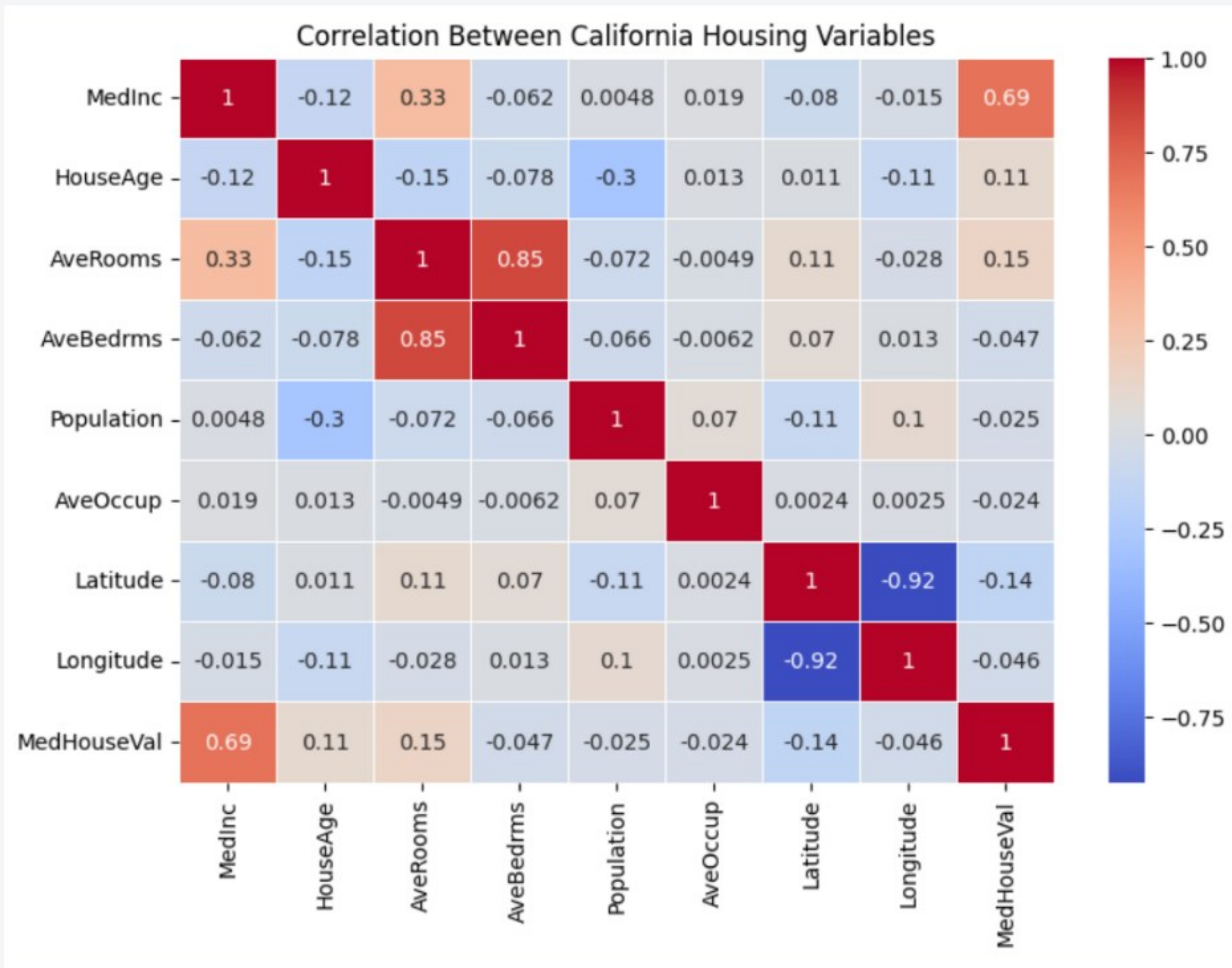
- Housing values vary substantially across California.
- Investment opportunities should be evaluated at the local market level, not the state level.
- Spike at the upper end (\$500,000) - an important limitation if investing in luxury markets.

4. Income as Major Value Indicator

- A positive relationship exists between median income and median house value.
- Income can help real estate companies identify markets that are more likely to support higher property values.



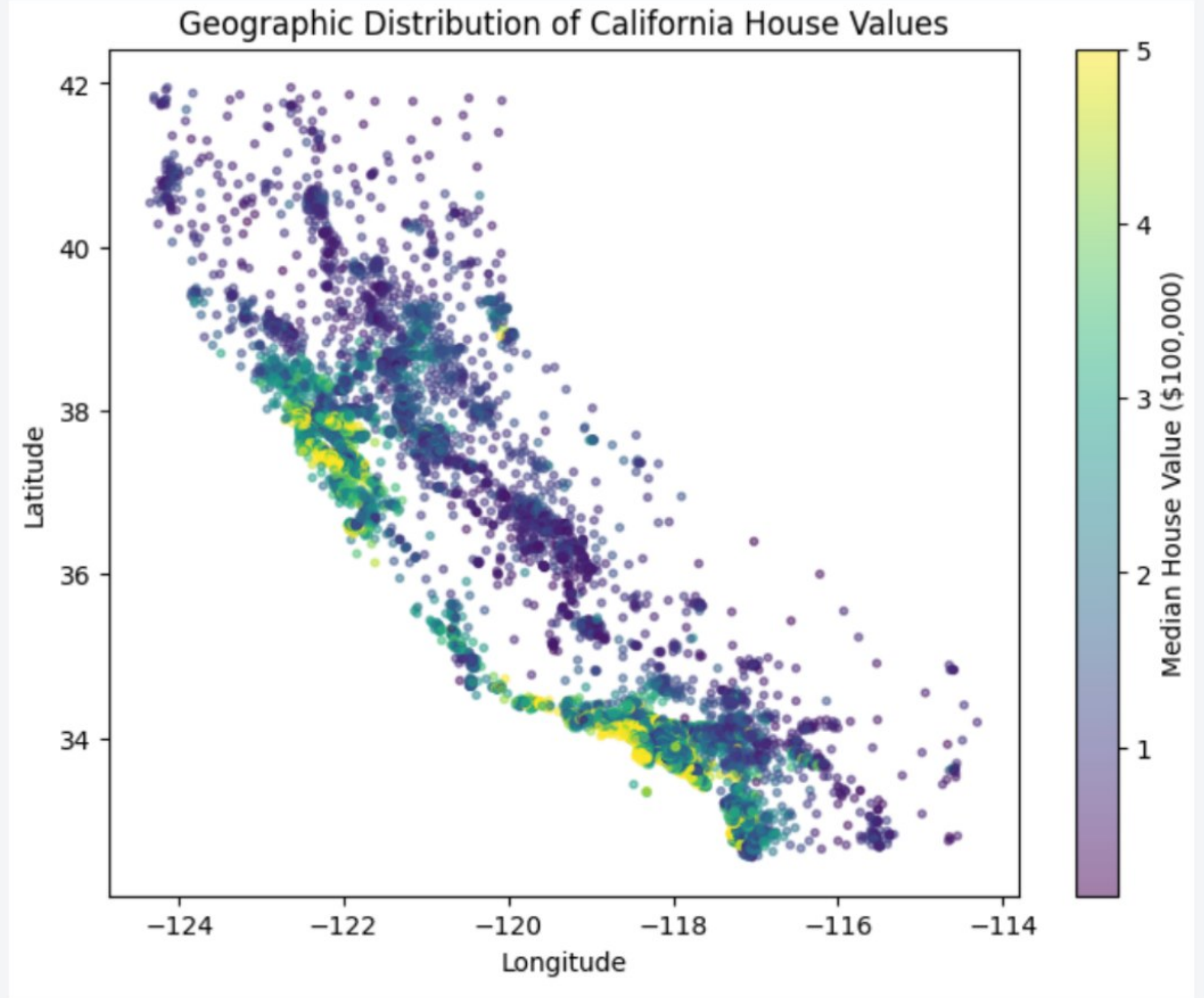
5. What Factors Are Most Related to House Values?



- Median income has the strongest positive correlation with house value ($r = 0.69$)
- Average Rooms has a weak positive correlation with house value ($r = 0.15$).
- House Age also shows a weak positive relationship ($r = 0.11$).
- Most other variables have little direct linear correlation with house value.
- Income appears to be the most useful individual predictor of housing value. However, housing prices cannot be explained by a single factor alone.

6. Does Location Change the Picture?

- Geographic concentrations of higher and lower housing values.
- Similar physical housing characteristics → not necessarily similar values across different locations.
- After income, location should be another consideration for investment markets.



7. Preparing the Data for the Model



- No missing values were identified
- No duplicates identified
- All 9 variables were stored as numerical data.
- No unnecessary columns.

- Median House Value was left in its original \$100,000 units for modeling.
- No scaling was required because basic Linear Regression can work with predictors measured on different scales.



- The data was divided into 80% training data and 20% testing data, providing a more realistic estimate of how the model may perform on new data.

8. Turning Market Insights into a Prediction

8 Market Indicators



the outcome is
continuous



Linear Regression

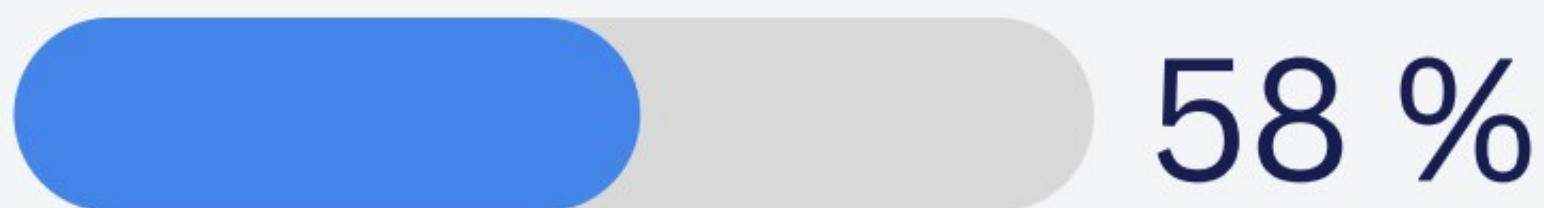


Predicted Median House Value

The model combines all eight economic, housing, demographic, and geographic indicators to predict median house value.



9. How Well Does the Model Predict House Values?



The model explains approximately 58% of the variation in median house values.

RMSE	R ²
≈ 0.746	≈ 0.576
≈ \$74,600	≈ 58%
Typical Prediction Error	Variation Explained

The model provides meaningful predictive information, but housing prices are influenced by additional factors not included in the dataset.

10. From Data to Business Decisions

KEY FINDINGS

1. **Income matters**

Higher median income is strongly associated with higher median house values.

2. **Location matters**

Housing values show clear geographic differences across California.

3. **Prediction is possible, but imperfect**

Linear Regression explains approximately 58% of house-value variation.

RECOMMENDATIONS

1. **Prioritize income when evaluating markets**

Use neighborhood income characteristics as an important indicator when comparing potential housing markets.

2. **Combine economic + geographic data**

Avoid evaluating properties using income alone; include location when assessing market opportunities.

3. **Use the model as decision support, not as the only pricing tool**

Linear Regression can provide an initial estimate, but additional market information should be considered.

4. **Improve future models with additional data**

Consider school quality, neighborhood amenities, property condition, crime/safety indicators, employment access, and other local factors.

Thank you

