Analyzing Key Factors Influencing Housing Prices in California

INF2190 Project Paper

Professor: Periklis Andritsos

Group 28

Jiayi Guo (1011265630)

Lanyue Hu (1011745462)

Hao Ni (1008283457)

Hanren Tu (1011265523)

Code Base: https://colab.research.google.com/drive/17KJeg3m1OryqS4-3xPcrEnv9cH5xu3QO?usp=sharing

Introduction

The housing market in California is shaped by numerous socio-economic and geographic factors. Understanding these determinants is crucial for homebuyers, real estate investors, policymakers, and urban planners. This project investigates the major factors influencing house prices in California using the "California Housing Prices" dataset from Kaggle, which includes details on geographic location, median income, age of housing, number of rooms, and proximity to the ocean.

Key research questions include: What are the most influential factors affecting housing prices in California? How does proximity to the ocean affect median housing prices? What is the relationship between socio-economic variables (e.g., median income) and housing prices? Are there distinct regional trends or clusters in California's housing market?

To achieve these objectives, we will employ analytical methodologies such as correlation analysis, multiple linear regression, decision tree modeling, and cluster analysis. These techniques will help quantify the influences, predict housing prices, and identify clusters of neighborhoods with similar pricing characteristics. The outcomes will provide actionable insights for investment strategies, policy decisions, and a deeper understanding of California's housing market dynamics.

Problem definition

California's real estate market is highly heterogeneous and dynamic, driven by various socio-economic and geographic factors. Understanding the reasons behind price differences across regions is challenging for homebuyers, real estate investors, and policymakers.

Key questions include:

- What factors have the greatest impact on California home prices?
- How does proximity to the ocean or urban centers affect median housing prices?
- Do socio-economic indicators such as median income and population density impact housing prices?

Example: Consider two neighborhoods, one near the coastline and one inland. While both may have similar housing characteristics, the coastal neighborhood often has higher prices. Is it merely proximity to the ocean, or do other variables like household income or population density play a role?

Motivation

Understanding and predicting home prices is critical for:

- Homebuyers seek informed choices when purchasing a home.
- Real estate investors need reliable forecasts to identify profitable markets.
- Policymakers and urban planners aiming to promote affordable housing and sustainable urban development.

Example: A family seeking to purchase a house in high-demand San Francisco might overlook inland neighborhoods with similar amenities but lower costs. Similarly, an investor might prioritize properties near urban centers without realizing that densely populated areas or proximity to amenities can yield higher returns even in less noticeable locations.

Description of the data

The "California Housing Prices" dataset from Kaggle provides information about housing prices in various California areas. It includes attributes like the number of rooms, population density, and house age, revealing how location, median income, and proximity to the ocean affect prices. The dataset consists of 20,640 rows and 10 variables:

- Longitude: Indicates how far west a house is.
- Latitude: Indicates how far north a house is.

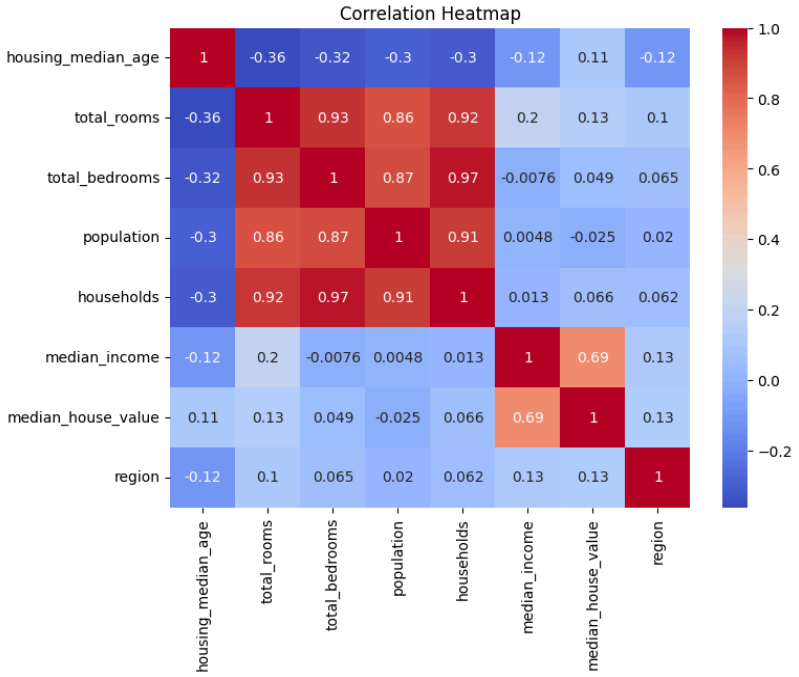
- Housing Median Age: The median age of houses in the neighborhood.
- Total Rooms: Total number of rooms in all blocks.
- Total Bedrooms: Total bedrooms within a block.
- Population: Number of residents in a block.
- Household: Total number of households in a block.
- Median Income: Median income for households within a block (in tens of thousands of USD).
- Median House Value: Median house value for households within a block (in USD).
- Ocean Proximity: Location of the house relative to the ocean.

There are some missing values in the "Total Bedrooms" column. The dataset provides valuable information for predicting house prices and analyzing trends in the real estate market.

Data Analysis Task

Data Preprocessing and Correlation Analysis

Handling Missing Values: Missing values in the 'total bedrooms' column were imputed using the median to preserve central tendency and robustness against outliers.
Label Encoding: Transformed the categorical 'ocean proximity' feature into numerical values to simplify training while maintaining computational efficiency.
Exploring Correlations: Computed a correlation matrix and visualized it using a heatmap to identify relationships between features. 'Median income' exhibited the highest positive correlation with 'median house value' (correlation coefficient = 0.688), confirming its role as a key predictor.

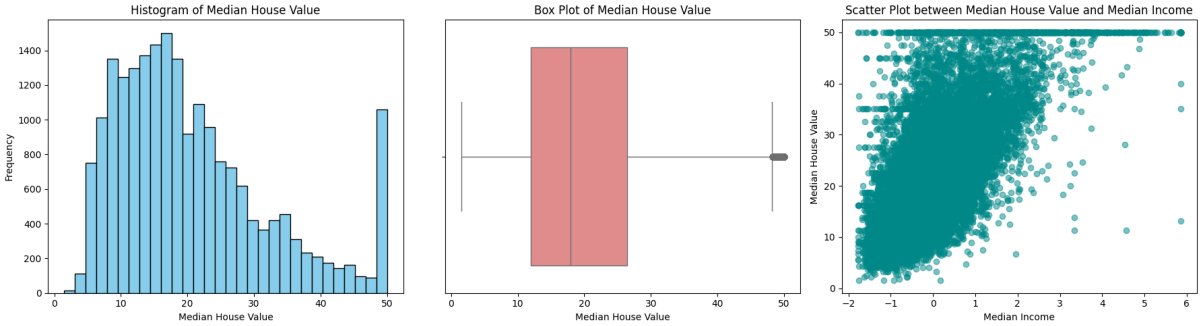


Addressing Skewness and Kurtosis: Several features showed significant skewness and kurtosis. Experimented with square root, log, and Box-Cox transformations, but these methods were limited by the capped values in 'median income' and 'median house value'.
Feature Importance: Identified 'median income' as the most significant predictor, aligning with the correlation analysis. Other moderately influential features included 'total rooms' and 'households'.

Visualization:

- A histogram revealed the positive skew of 'median house value', with a concentration near the upper limit of $500k, confirming the effect of capping.
- A box plot highlighted outliers and asymmetry in the housing price distribution.

- A scatter plot between 'median income' and 'median house value' demonstrated their strong positive correlation, with noticeable clustering near the upper capped values.



Decision Tree Regressor

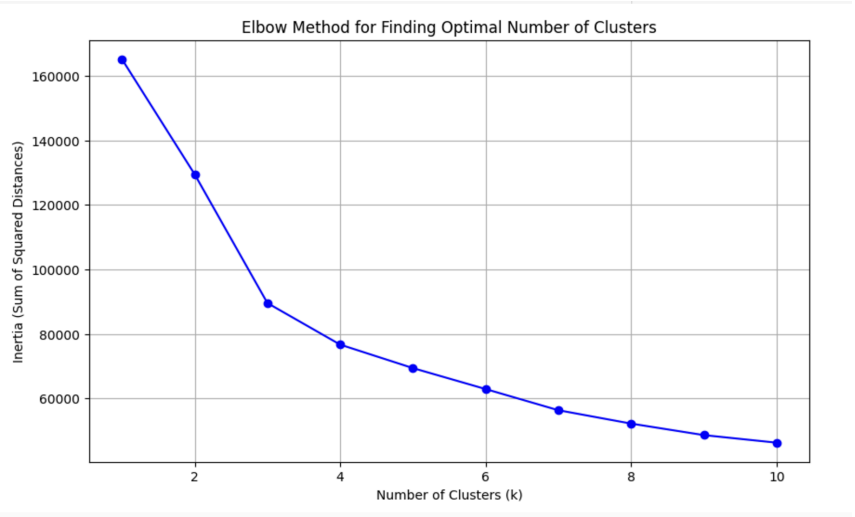
To predict median housing prices, we applied a Decision Tree Regressor, emphasizing its interpretability and ability to capture non-linear relationships. The data preprocessing focused on addressing skewed features, such as 'total rooms' and 'population', using log and Box-Cox transformations to improve normality. 'latitude' and 'longitude', which exhibited non-linear trends, were converted into regional clusters using KMeans, simplifying spatial data representation.

Exploratory analysis identified 'median income' as the most significant feature influencing housing prices, with weaker contributions from variables like 'total rooms' and 'housing median age'. These insights guided feature selection for the model. By incorporating preprocessed and engineered features, the initial Decision Tree Regressor achieved an R-squared score of 0.49, demonstrating moderate predictive accuracy.

```
R-squared score: 0.4946398348252846
```

Clustering Analysis

To uncover patterns in the California housing dataset, we employed Hierarchical clustering and K-Means clustering, focusing on geographic, demographic, and economic variables. Hierarchical clustering, using Ward's linkage method, revealed nested relationships and cluster formations, visualized through dendrograms, making it ideal for exploratory analysis. K-Means clustering segmented the data into distinct clusters, with the elbow method and silhouette scores confirming that five clusters were optimal. Preprocessing steps, including imputing missing values and standardizing features, ensured balanced contributions from all variables. Together, these methods offered meaningful insights and precise segmentation of the housing data.



Multiple Linear Regression

The analysis was conducted using Multiple Linear Regression (MLR) to explore the relationship between various housing-related features and the median house values in California. MLR was chosen for its ability to quantitatively assess the impact of multiple independent variables on a single dependent variable while accounting for the effects of other predictors. This method also allows for a clear interpretation of the relative importance of each predictor through its regression coefficients.

The data was first cleaned to address any missing values, irrelevant columns, and inconsistencies. The dataset was then split into training (80%) and testing (20%) sets to enable performance evaluation on unseen data. The training set was used to fit a LinearRegression model from scikit-learn, which was subsequently evaluated using standard metrics such as Mean Squared Error (MSE) and R-squared (R^2).

Experimental Results

Data Preprocessing and Correlation Analysis

Preprocessing and correlation analysis ensured that the dataset was prepared for analysis. Correlation heatmaps confirmed median_income as the strongest predictor of median_house_value. Histograms and scatter plots revealed data capping and outliers, influencing clustering and predictive performance.

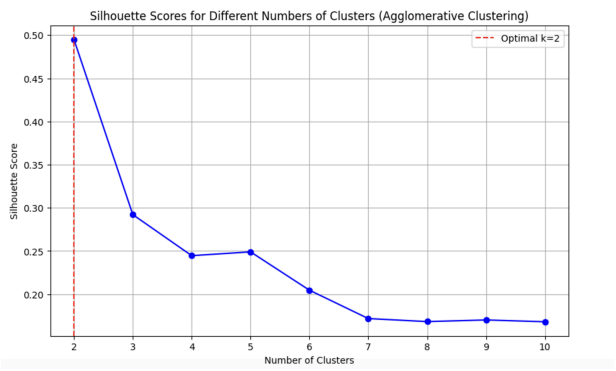
Decision Tree Regressor

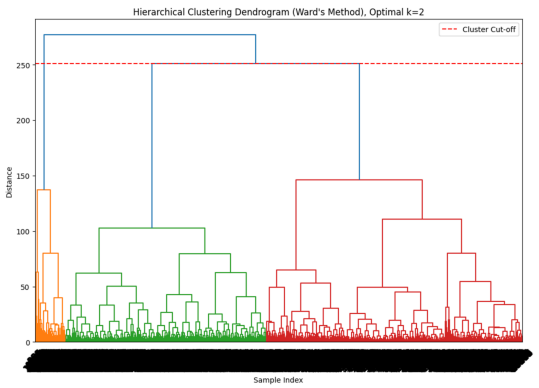
The Decision Tree Regressor was optimized using GridSearchCV to improve its performance. Key hyperparameters, including max_depth, min_samples_split, and min_samples_leaf, were tuned to mitigate overfitting and enhance generalizability. The optimized model achieved an R-squared score of 0.67, reflecting a significant improvement. 'median income' emerged as the dominant predictor, consistent with its strong correlation with housing prices. Regional labels and 'ocean proximity' also contribute meaningfully. While the model fell slightly short of the target score of 0.70, it provided useful insights into feature importance.

```
The best params: {'max_depth': 20, 'min_samples_leaf': 22, 'min_samples_split': 50}
Optimized R-squared score: 0.6749994004565936
```

Clustering Analysis

The clustering analysis revealed key patterns in California's housing market by testing clusters from 2 to 10 using hierarchical and K-Means clustering. Both methods identified five clusters as optimal, validated by the elbow method, silhouette scores, and dendrograms. Clusters reflected socioeconomic stratification, with high-income coastal neighborhoods forming distinct groups, while latitude and longitude highlighted urban and coastal trends. The results segmented the data into meaningful groups, uncovering drivers of housing trends and providing actionable insights for investment, policy, and planning.

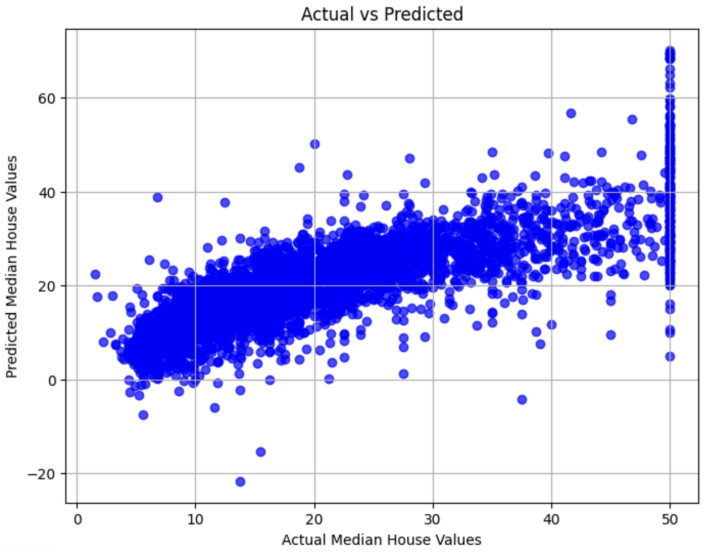




Multiple Linear Regression
The Multiple Linear Regression model produced an MSE of 50.31 and an R-squared of 0.62, showing that the model explains 62% of the variance in median house values. This indicates moderate accuracy but highlights room for improvement.

	Feature	Coefficient
6	households	27.644651
7	median_income	7.616057
2	housing_median_age	1.557043
8	ocean_proximity	0.571795
4	total_bedrooms	0.011493
1	latitude	-3.750961
0	longitude	-3.957807
3	total_rooms	-27.924742
5	population	-140.970317

The regression coefficients revealed important predictors of housing prices. Among the positive predictors, households (27.64) and median income (7.62) emerged as the most significant, with median income being a strong driver of higher house values. Other positive predictors included housing median age (1.56) and ocean proximity (0.57), suggesting that older housing stock and proximity to the ocean have modest positive impacts. The effect of total bedrooms (0.01) was negligible, indicating it has little influence on housing prices.



A scatter plot of actual vs. predicted values showed alignment but also revealed some outliers, suggesting room for refinement. Residuals were randomly distributed around zero, confirming the model's assumptions. Overall, the model provides useful insights into housing price drivers. However, improving accuracy may require non-linear models, interaction terms, or additional features. The findings offer a strong baseline for further analysis.

Discussion

This analysis employed three models—Decision Tree Regressor, Clustering Analysis, and Multiple Linear Regression—to explore factors influencing California's housing market.

- The Decision Tree Regressor captured non-linear relationships, with median_income emerging as the most critical predictor. After hyperparameter optimization, the model

explained 68% of the variance in housing prices, significantly improving from the baseline. However, Decision Trees remain prone to overfitting, and the capped values for median_income and median_house_value constrained predictions at the upper price range. These limitations stem from dataset characteristics rather than model deficiencies. Future improvements could address capped values through data preprocessing or scaling. Additionally, ensemble methods like Random Forests or Gradient Boosting could enhance robustness and better generalize pricing trends.

- Clustering Analysis: Clustering segmented neighborhoods into meaningful groups, with K-Means identifying five clusters and hierarchical clustering providing additional context. The results highlighted socio-economic and geographic patterns, such as high-income coastal clusters, offering insights for investment and policy. K-Means assumes spherical clusters, and hierarchical clustering struggles with large datasets. Future work could use advanced methods like time-series data to capture complex trends.
- Multiple Linear Regression: MLR quantified feature impacts, with median income as the strongest predictor. The model explained 62% of price variance, highlighting key trends in housing markets. MLR's linearity limits complex interactions, and capped values reduce accuracy. Future improvements could include polynomial regression or expanded datasets for better predictions.

Challenges, such as capped values in median income and median house value, constrained model performance. Addressing these limitations through advanced models like Random Forest or Gradient Boosting could further enhance accuracy. Additionally, exploring datasets without capped values would provide more robust predictions. The combination of models and visualizations in this study offers a comprehensive understanding of housing price dynamics in California.

Conclusion

This project examines the key factors that influence home prices in California using techniques and models such as Decision Tree Regression, Cluster Analysis, and Multiple Linear Regression that we have learned during the semester. Our findings indicate that of all the factors, median income is the most important of all the variables that predict future home prices. Income pushes up house prices when residential neighborhoods have higher socioeconomic status. The type of waterfront housing and its location remain quite important factors for overall valuation. And cluster analysis provided meaningful insights into regional and socioeconomic segmentation. The regression models quantify the relationship between various factors and house prices.

Based on the results of our analysis, we offer the following recommendations for homebuyers, real estate investors and policymakers. Homebuyers should prioritize middle- and upper-income areas for property appreciation potential, while exploring areas farther away from the coast but with better value for money, and choosing homes that are newer or in established neighborhoods. Real estate investors can focus on high-income neighborhoods and potential markets at the urban-suburban interface, combining the results of a cluster analysis to prioritize investments in high-end areas in and around urban centers. For policymakers, it is recommended that the construction of affordable housing be accelerated, especially in areas with a concentration of low- and middle-income households, and that investments be made in transportation and infrastructure in non-coastal areas to alleviate price pressures in high-demand areas.

AI Appendix

We used generative AI (ChatGPT) to polish the final version of this report. It helped enhance sentence structure, coherence, and flow. The content, analysis, and ideas presented in this proposal are entirely based on our discussions and independent thought. AI was used to polish wording and improve expression, as well as to check grammar and sentence construction.

Prompt for Refinement: "Please read the paragraph I provided and polish the text while ensuring that the meaning of the original text is not changed. Check the sentence structure and ensure coherence and fluency: XXXXXX"

References

- Nugent, Cameron. *California Housing Prices*. 2018, Kaggle, *https://www.kaggle.com/datasets/camnugent/california-housing-prices*.
- OpenAI. (2024). *ChatGPT (November 2024 version)*. Retrieved from *https://chat.openai.com/*