Housing Price Prediction

Shrut Dalwadi

Introduction

The California Housing dataset provides insights into how location, income, and household characteristics affect housing prices. This dataset is intriguing to me since it contains numerical as well as categorical features.

Research Questions Overview

1. How does the number of total bedrooms impact median house value across regions?
2. How does the ratio of bedrooms per room affect median house value?
3. What is the impact of ocean proximity on median house value?
4. How does the number of total rooms affect households in predicting median house value?
5. How does median income compare to housing median age in impacting median house value?

Data Exploration and Cleaning

Feature	Description	Data Type
longitude	The longitude of the region.	float64
latitude	The latitude of the region.	float64
housing_median_age	The median age of the houses in the region.	float64
total_rooms	The total number of rooms in all houses in the region.	float64
total_bedrooms	The total number of bedrooms in all houses in the region.	float64
population	The total population of the region.	float64
households	The total number of households in the region.	float64
median_income	The median income of households in the location (in tens of thousands of dollars).	float64
median_house_value	The median house value for households in the location (target variable).	float64
ocean_proximity	The proximity of the location to the ocean (categorical feature).	object

- **Handling Missing Values:** Only the total_bedrooms feature had missing values (207 missing entries). Rows with missing values were removed.

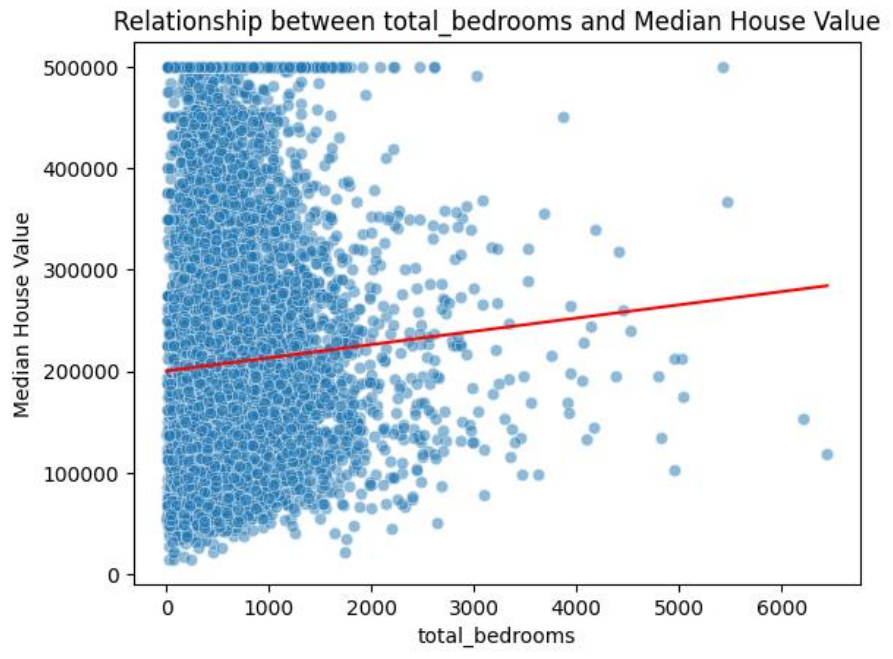
- **Feature Creation:** Two new features were created to improve predictive power:
  - bedrooms_per_room: Created by dividing total_bedrooms by total_rooms, this feature showed a stronger correlation with median_house_value than total bedrooms or rooms individually.

  - rooms_per_household: Derived by dividing total_rooms by households, offering a more detailed measure of housing density than total_rooms alone.

- **Handling Categorical Data:** ocean_proximity was one-hot encoded to convert this categorical feature into numerical format.

Research Question 1

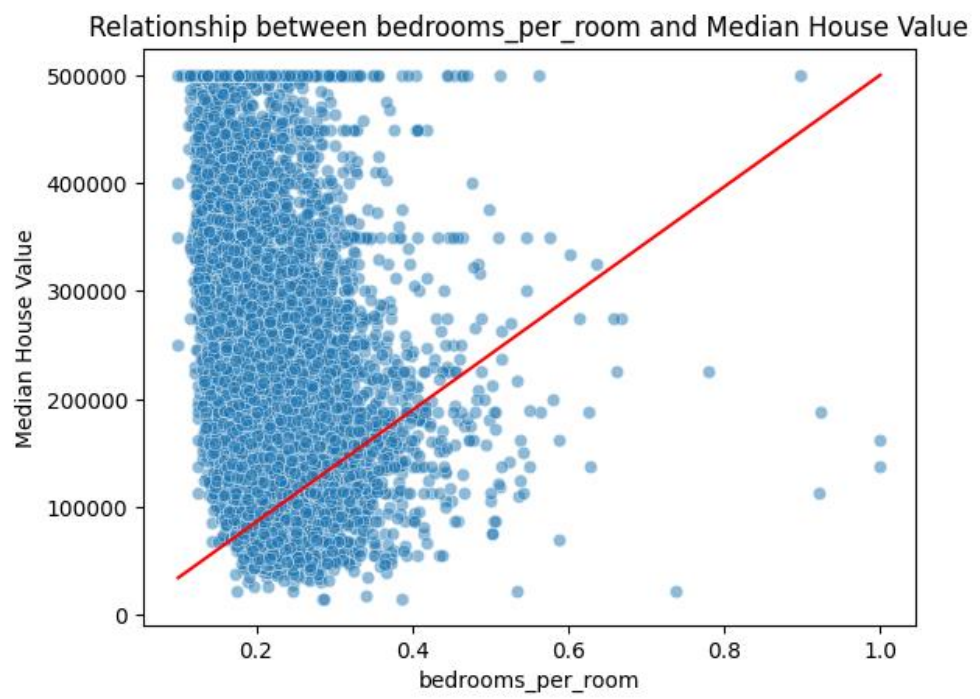
Q. How does the number of total bedrooms impact median house value across regions?



Total bedrooms show a positive but weak effect on median house value, with a higher p-value (0.056), suggesting it isn't a strong predictor on its own and has a limited impact compared to other factors.

Research Question 2

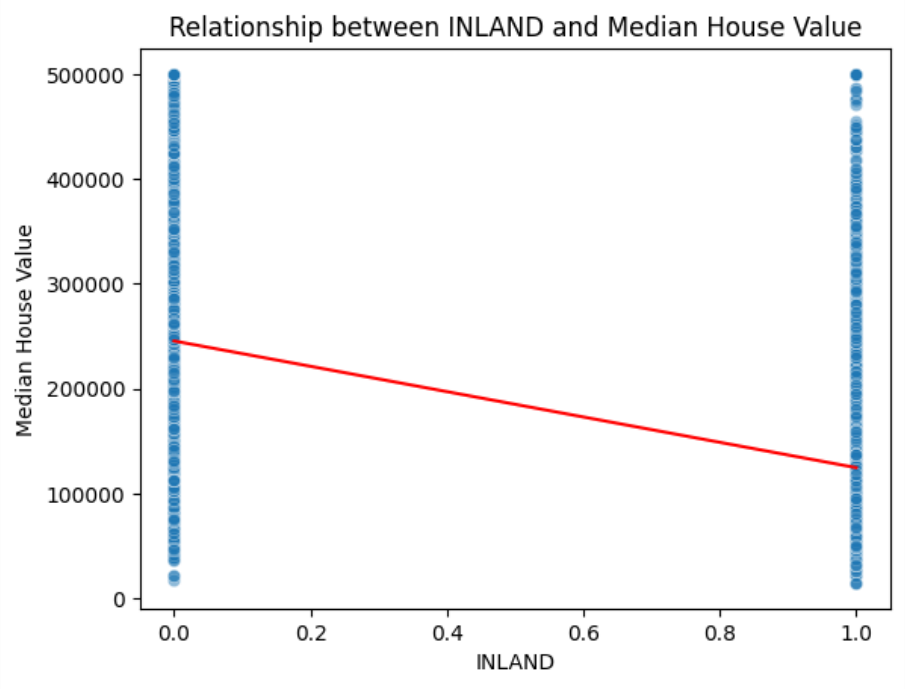
Q. How does the ratio of bedrooms per room affect median house value?



The bedrooms per room ratio has a strong positive effect on median house value, with a high coefficient (278,700) and strong statistical significance. This suggests that homes with more bedrooms relative to rooms tend to be larger or more luxurious, adding to their value.

Research Question 3

Q. What is the impact of ocean proximity on median house value?

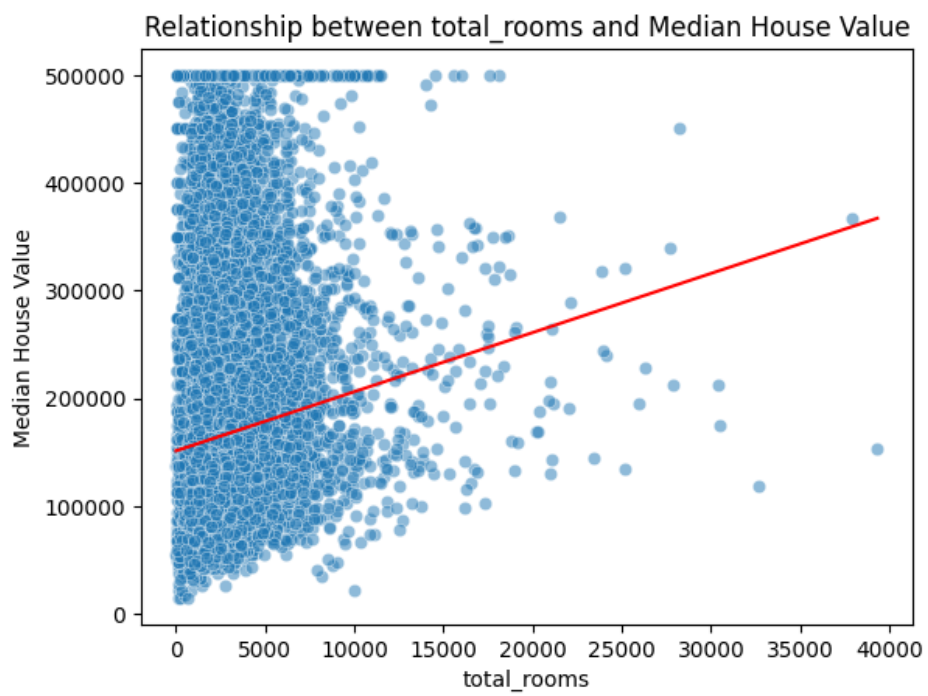


Proximity to the ocean is highly influential, with inland areas having a significant negative impact on house prices. Homes closer to the ocean generally hold higher values, emphasizing the premium associated with ocean proximity.

In the plot, 1 signifies the ocean_proximity of the district as INLAND and 0 means other class of ocean_proximity.

Research Question 4

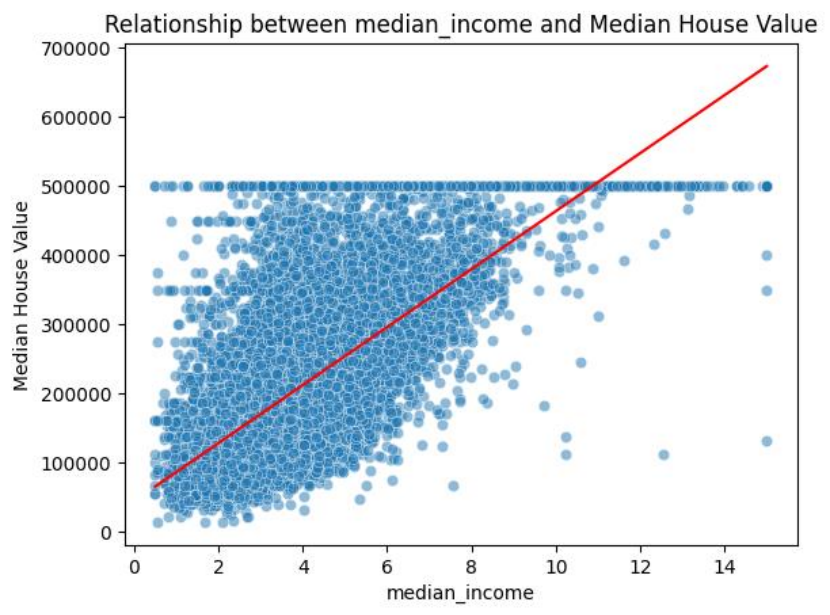
Q. How does the number of total rooms affect households in predicting median house value?

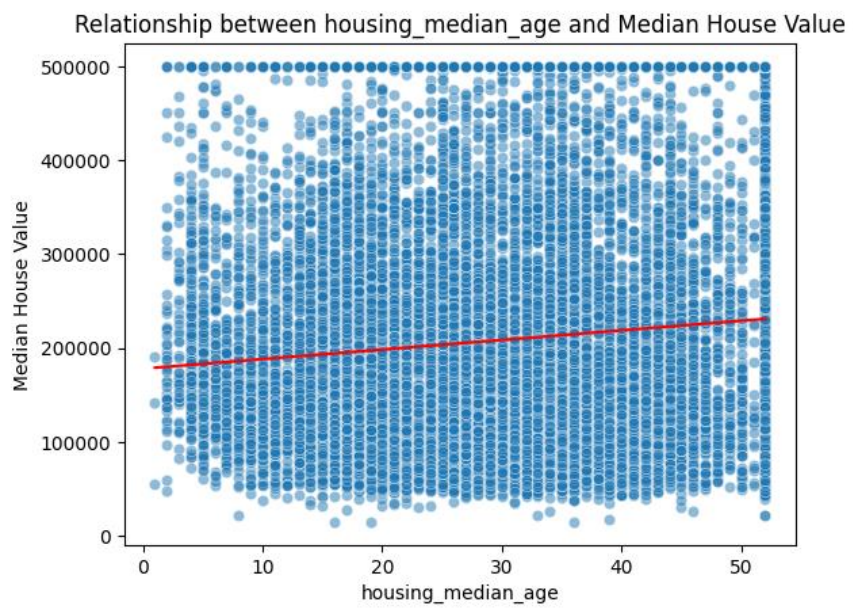


The total number of rooms has a minimal effect on house value (p-value of 0.115) which indicates that on its own, room count isn't a strong predictor of house value.

Research Question 5

Q. How does median income compare to housing median age in impacting median house value?





Research Question 5

Median income has a much greater impact on house values than housing age, with a large positive coefficient (41,510) and strong significance. While housing age also has a positive effect, its influence is smaller, suggesting that income levels are a more crucial factor in determining housing value than the age of the homes.

Conclusion

- The linear regression model shows that median house values are mainly driven by income levels, proximity to the ocean, and household characteristics. Higher incomes and closer ocean access significantly raise house values, while inland areas generally see lower prices.

- A higher bedrooms-per-room ratio also adds value, likely indicating larger or more luxurious homes. Meanwhile, total rooms and bedrooms have a minimal effect on their own, and housing age has a positive but smaller impact compared to income.

- However, there are limitations to consider, the residuals suggest some non-linear relationships that the linear model may not fully capture and there are potential outliers that might skew results.