Problem Statement

Given a set of 1990 data on California house characteristics (age, number of bedrooms, etc.) and location (coordinates, distance from major cities, etc.) versus median price, can we create a model to estimate the value of a California house in an area not listed in the data, assuming we are provided with its own characteristics, location, and other info? What are the most significant characteristics to include when building a house in California such that its estimated price is in the 70th percentile of California house prices?

1) Context

A California real estate investor is interested in flipping properties that the market highly undervalues. He would like a model that details what characteristics of a house will especially entice buyers to pay premiums for. Ideally he would use this model to search for hidden gems - houses that contain high-value characteristics but are for sale at prices far below expected. To maximize the cost-effectiveness of his efforts, he would like to primarily consider houses that he could first buy at below the 50th percentile of California prices and then sell at at least the 70th percentile of prices.

2) Criteria for success

The investor is able to sell at least 5 houses within the next year priced at at least the 70th percentile of California house prices, after buying them at below the 50th percentile of prices.

3) Scope of solution space

The solution applies to any house within California. When searching for houses to flip, the investor will foremost consider the characteristics that are most correlated to price, in order to maximize the price he can set the house at.

4) Constraints within solution space

The dataset does not contain numerous fields that would likely have an impact on house price, like proximity to other popular natural features (national parks, rivers, mountains, etc.), proximity to public transportation, number of large trees on the property, etc.

5) Stakeholders to provide key insight

Whoever was able to provide the data source. The investor's boss, if he has one.

6) Key data sources

Dataset on California houses containing the following fields:

- Median house value: Median house value for households within a block
- Median income: Median income for households within a block of houses
- Median Age: Median age of a house within a block; a lower number is a newer building
- Total Rooms: Total number of rooms within a block
- Total Bedrooms: Total number of bedrooms within a block
- Population: Total number of people residing within a block
- Households: Total number of households, a group of people residing within a home unit, for a block
- Latitude: A measure of how far north a house is; a higher value is farther north
- Longitude: A measure of how far west a house is; a higher value is farther west
- Distance to coast: Distance to the nearest coast point (in meters)
- Distance to Los Angeles: Distance to the centre of Los Angeles (in meters)
- Distance to San Diego: Distance to the centre of San Diego (in meters)
- Distance to San Jose: Distance to the centre of San Jose (in meters)
- Distance to San Francisco: Distance to the centre of San Francisco (in meters)

Source:
https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features