California Housing Data Analysis

Lucy Zhao & Keon Dibley

2023-10-22

Dataset: California Housing Prices Data Source: Kaggle

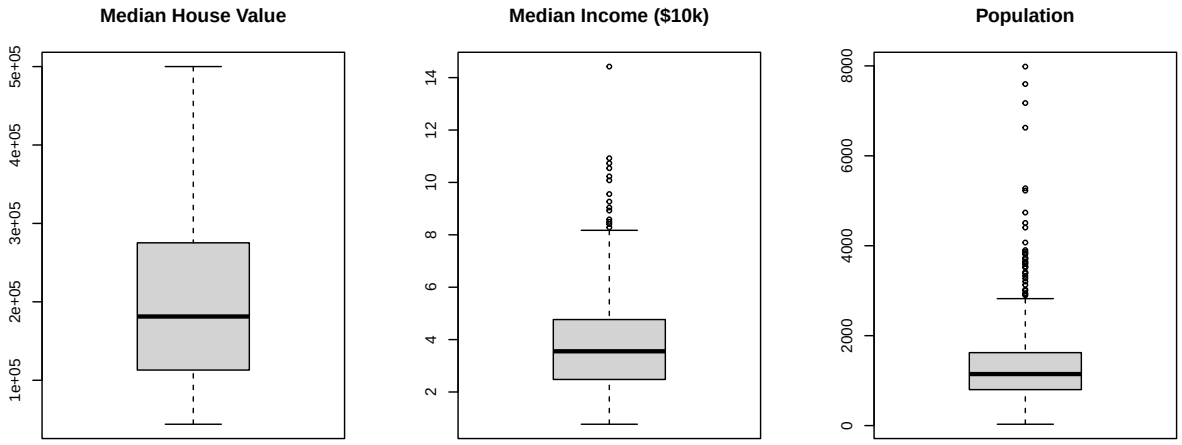
The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data.

An observational unit in this dataset is a block of houses

Variables:

- Median House Value: Median house value for households within a block (measured in US Dollars) [$]
- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
- Median Age: Median age of a house within a block; a lower number is a newer building [years]
- 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 [m]
- Distance to Los Angeles: Distance to the centre of Los Angeles [m]
- Distance to San Diego: Distance to the centre of San Diego [m]
- Distance to San Jose: Distance to the centre of San Jose [m]
- Distance to San Francisco: Distance to the centre of San Francisco [m]
- North South CA: Decides whether a block is in Northern or Southern California, based on the northern border of San Luis Obispo County
- Near Center of Top 4 Big City: Decides how far a block is from the center of its closest big city (SF, LA, SD, SJ): close, somewhat close, somewhat far, or far.

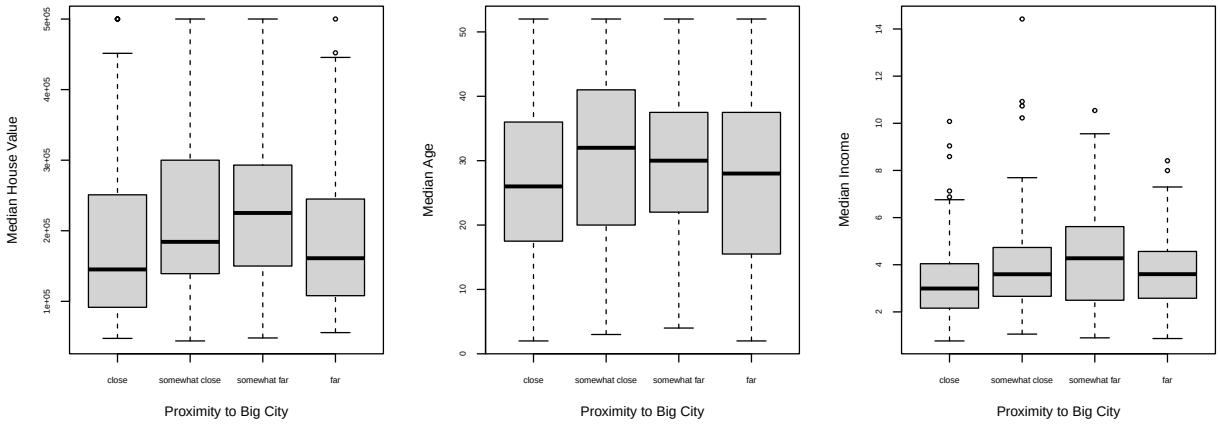
Individual Distributions:



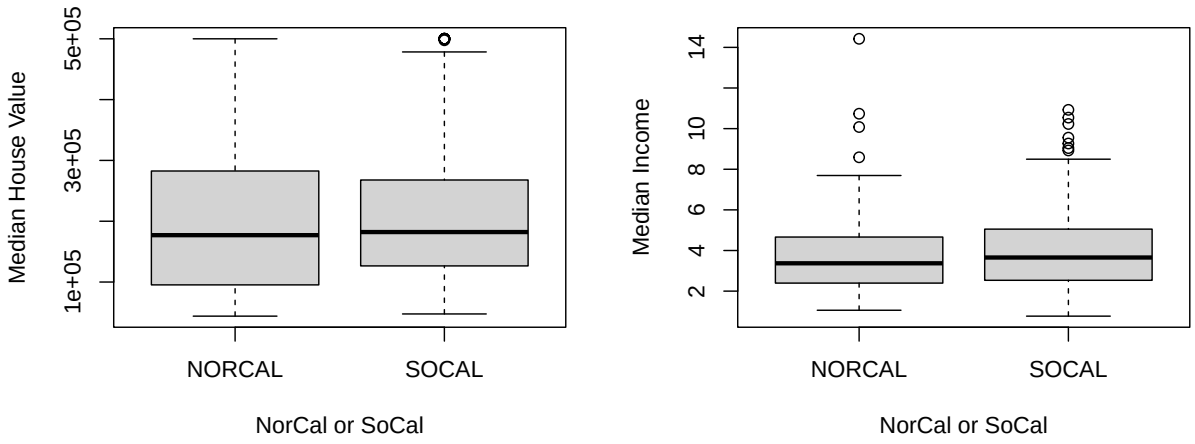
(Figure 1:Boxplots of the Median House Value, Median Income, and Population variables)

Binary Distributions:

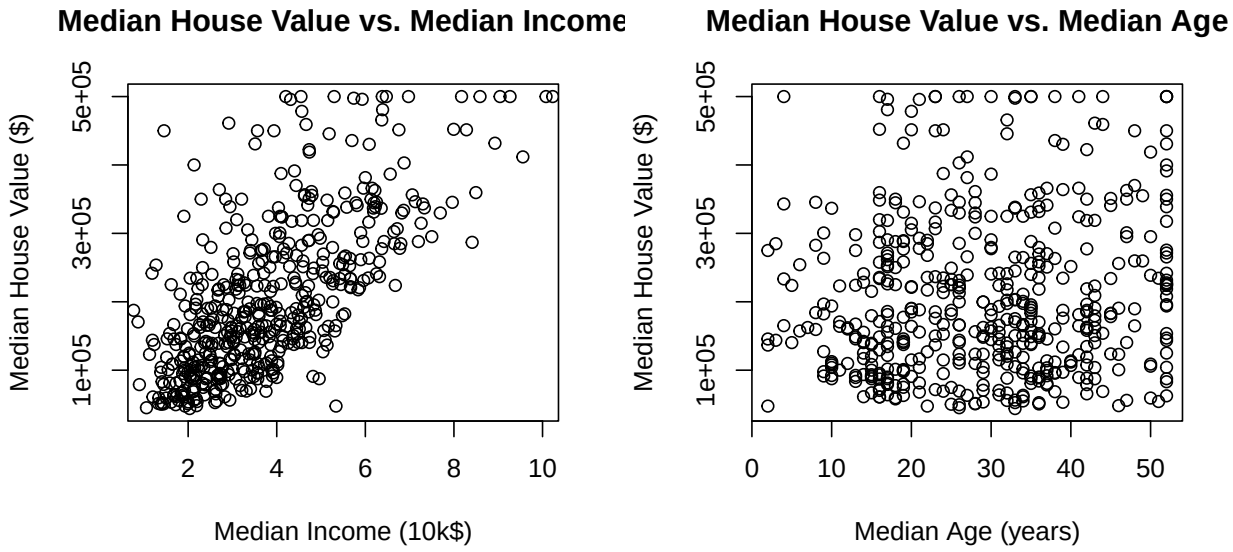
Below is the relationship between a block's proximity to a top 4 big city in California, and its Median House Value, Median House Age, and Median Income:



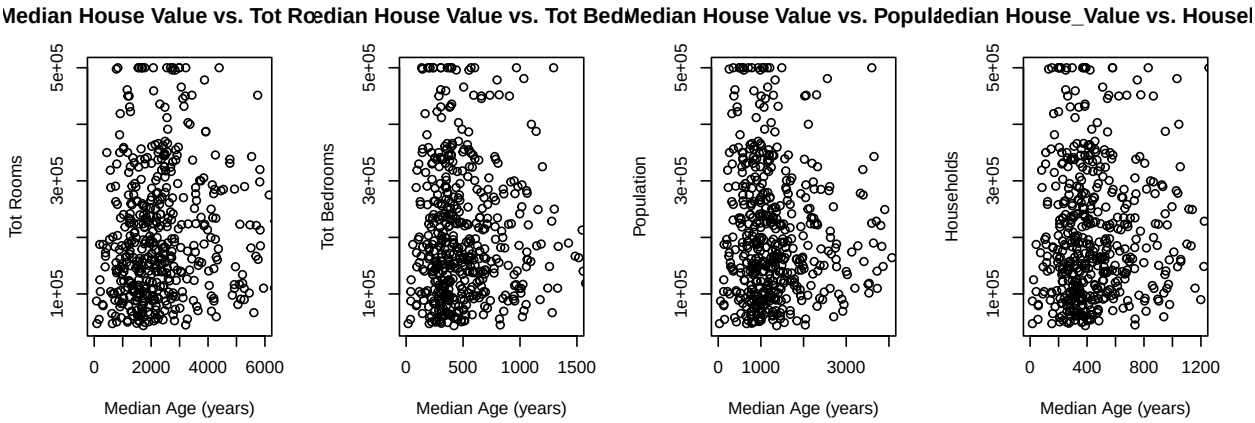
(Figure 2: A block's proximity to a big city vs. its Median House Value, Median Age, and Median Income)
Houses that are very close to city centers tend to be newer according to this graph.
Interestingly, people who are 'close' to city centers are associated with lower median house value and income than those who live 'somewhat close' or 'somewhat far'. This may be attributed to urban sprawl and the diffusion of people with money from urban centers into the suburbs.



(Figure 3 : Comparing Median House Value and Median Income in Northern and Southern California)
There is no obvious difference between Northern California and Southern California in terms of median house value and median income, they have similar statistical properties.



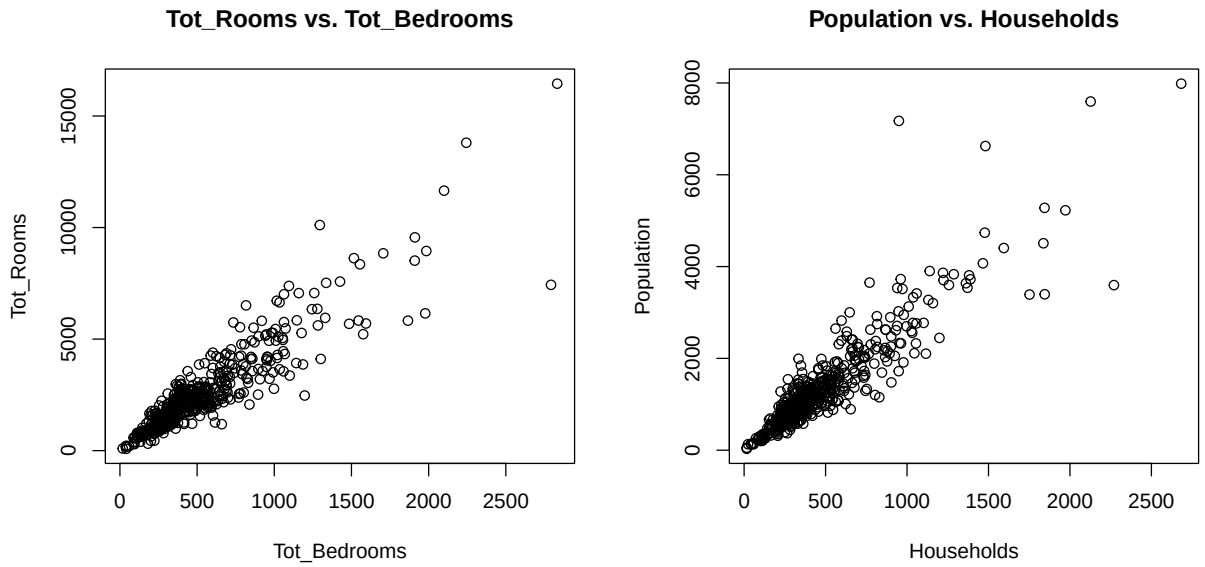
(Figure 4 : How do Median Income and Median Age impact Median House Value)
It seems that there's a linear relationship between Median House Value and Median Income. However, there seems to be no relationship between Median House Value and Median Age.



(Figure 5 :Relationship Between Median Age and Four Other Variables)

Intuitively, these four scatter plots share the same pattern, which suggests a relationship between the four response variables. Therefore, Total Rooms, Total Bedrooms, Population, and Households should be considered together.

Below is the plot to show this more clearly.



(Figure 6 : Correlation between Total Rooms and Total Bedrooms, as well as between Populations and Households)
From the graph's slope, we can see that
Tot_Rooms = 4 * Tot_Bedrooms
Population = 2 * Households