**California Housing Data Set Description**

Many of the Machine Learning Crash Course Programming Exercises use the California housing data set, which contains data drawn from the 1990 U.S. Census. The following table provides descriptions, data ranges, and data types for each feature in the data set.

| **Column title** | **Description** | **Range\*** | **Datatype** |
| --- | --- | --- | --- |
| longitude | A measure of how far west a house is; a more negative value is farther west | * Longitude values range from -180 to +180 * Data set min: -124.3 * Data set max: -114.3 | float64 |
| latitude | A measure of how far north a house is; a higher value is farther north | * Latitude values range from -90 to +90 * Data set min: 32.5 * Data set max: 42.5 | float64 |
| housingMedianAge | Median age of a house within a block; a lower number is a newer building | * Data set min: 1.0 * Data set max: 52.0 | float64 |
| totalRooms | Total number of rooms within a block | * Data set min: 2.0 * Data set max: 37937.0 | float64 |
| totalBedrooms | Total number of bedrooms within a block | * Data set min: 1.0 * Data set max: 6445.0 | float64 |
| population | Total number of people residing within a block | * Data set min: 3.0 * Data set max: 35682.0 | float64 |
| households | Total number of households, a group of people residing within a home unit, for a block | * Data set min: 1.0 * Data set max: 6082.0 | float64 |
| medianIncome | Median income for households within a block of houses (measured in tens of thousands of US Dollars) | * Data set min: 0.5 * Data set max: 15.0 | float64 |
| medianHouseValue | Median house value for households within a block (measured in US Dollars) | * Data set min: 14999.0 * Data set max: 500001.0 | float64 |

\* Min and max values in the table below were obtained from the Exercise notebooks using [pandas.DataFrame.describe()](https://web.archive.org/web/20240602065639/https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html) on the California Housing data set

## **Reference**

Pace, R. Kelley, and Ronald Barry, "Sparse Spatial Autoregressions," Statistics and Probability Letters, Volume 33, Number 3, May 5 1997, p. 291-297.

The following is the data methodology described in the paper:

We collected information on the variables using all the block groups in California from the 1990 Census. In this sample a block group on average includes 1425.5 individuals living in a geographically compact area. Naturally, the geographical area included varies inversely with the population density. We computed distances among the centroids of each block group as measured in latitude and longitude. We excluded all the block groups reporting zero entries for the independent and dependent variables. The final data contained 20,640 observations on 9 characteristics.