In this dataset, we notice that the entire nature of the columns originally presented are numeric (float).

The columns being very heterogeneous by their nature, we are confronted from the start by the lack of clarity imposed by the different units of measurement.

It can be assumed (from the data displayed) that the columns relate to the following units:

`Median\_House\_Value` = Dollars

`Median\_Income` = Dollars

- Most of whose values have a digit before the decimal point (composed of tens).

Outlier values (notably distorting the average.

`Median\_Age` = Age

Disparities in low outliers: 1 year

`Tot\_Rooms` = Integer Number

`Tot\_Bedrooms` = Integer Number

`Population` = Number of people, inhabitants

`Households` = Number of households. Low outlier value wrong because it is negative

`Latitude, Longitude` = Geographic coordinates

`Distance\_to\_coast` = Distance units, probably Miles

`Distance\_to\_LA`, `Distance\_to\_SanDiego =` Distance, but expressed as an exponential (different from the other two columns listing distances):

`Distance\_to\_SanJose', 'Distance\_to\_SanFrancisco` = Distance in Miles

`Max\_Age` = Maximum age depending on the Median\_Age column

Concerning the Average; the values are significantly brought upwards, by a considerable number of high-value households (500,000, for 800,000 counts), which causes the average to tend to a much greater amount than the median. Median Income is very thinly dispersed and follows a logic that makes sense, despite a slight over-representation of very high value Incomes.

For greater relevance, we chose to use the interquartile range (IQR) method to preprocess our dataset. More than 2000 lines have been deleted.

The “Closest\_City” column tells us that the values are very unevenly distributed geographically, and that there are sectors essentially close to the cities bordering the west coast (by the sea). Almost half of the sectors have the closest city: Los Angeles (9829 units), followed by San Francisco (5057), San Jose (3766) and San Diego (2002).

*Description of each column with a sentence:*

Median\_House\_Value = Average housing value by sector

Median\_Income = Average household income by sector

Median\_Age = Average age of sectors

Tot\_Rooms = Sum of household rooms by sector

Tot\_Bedrooms = Sum of household bedrooms by sector

Population = Number of inhabitants per sector

Households = Number of households per sector

Latitude = Geographic coordinates Y

Longitude = Geographic coordinates

Distance\_to\_coast = Distance expressed in miles from the sector to the nearest sea point.

Closest\_City = Distance from sector to nearest major city

Correlation matrices:

We can focus on material correlations which suggest that the structural elements of households are all very correlated. Depending on the number of households per sector, the number of rooms and bedrooms will be higher, as will the number of rooms and bedrooms.

Logically, the number of inhabitants (population) is increasing, the more rooms and households there are.

In addition to these mundanely understandable correlations, we note that as the average income of households in a sector increases, the median value of households increases (0.689597).

Additionally, Median\_House\_Value and Distance\_to\_Coast are significantly closely correlated (-0.466959), which means, literally, that households with high monetary value will tend to be close to the sea.

Los Angeles (which is the closest major city for almost half of the sectors counted) also brings together the most wealthy sectors. Even if the other 3 cities concentrate sectors that are very slightly higher in terms of income and common values, compared to the average for the entire Californian state.

On the other hand, we can affirm that there are large disparities in terms of size (population and households covered) between the different sectors. Indeed, an average apparently wealthy household does not necessarily mean that it has a lot of rooms and accommodation in this sector. Hence the assumption that a good number of relatively wealthy sectors are small.

Linear Regression:

The MSE is at 0.211725 meaning that on average, the squared prediction errors of your model are of the order of 0.211725. In this case, an MSE of 0.211725 suggests that your linear regression model is doing a reasonably good job of fitting the data, but there is still some unexplained variability.

Our MAE of 0.239011 means that, on average, your model's prediction errors have an absolute magnitude of approximately 0.239011. Knowing that our MAE is less sensitive to outliers than MSE, it suggests that your model's predictions are, on average, 0.239011 units away from the response variable (average household value) from the actual values.

In summary, our linear regression model is able to predict average household value as a function of distance to the sea and average income by sector reasonably well. However, there remains some residual error, as indicated by the MSE and MAE.

See further analysis along Part 3.