Advance Python Programming

An End-to-End Machine Learning Project: I

Dionisios N. Sotiropoulos

November 9, 2021

University of Piraeus
Department of Informatics

Table of contents

Working with Real Data

Working with Real Data I

In this teaching unit we will use the California Housing Prices dataset from the StatLib repository.

- This dataset is based on data from the 1990 California census.
- It is not exactly recent (a nice house in the Bay Area was still affordable at the time), but it has many qualities for learning.

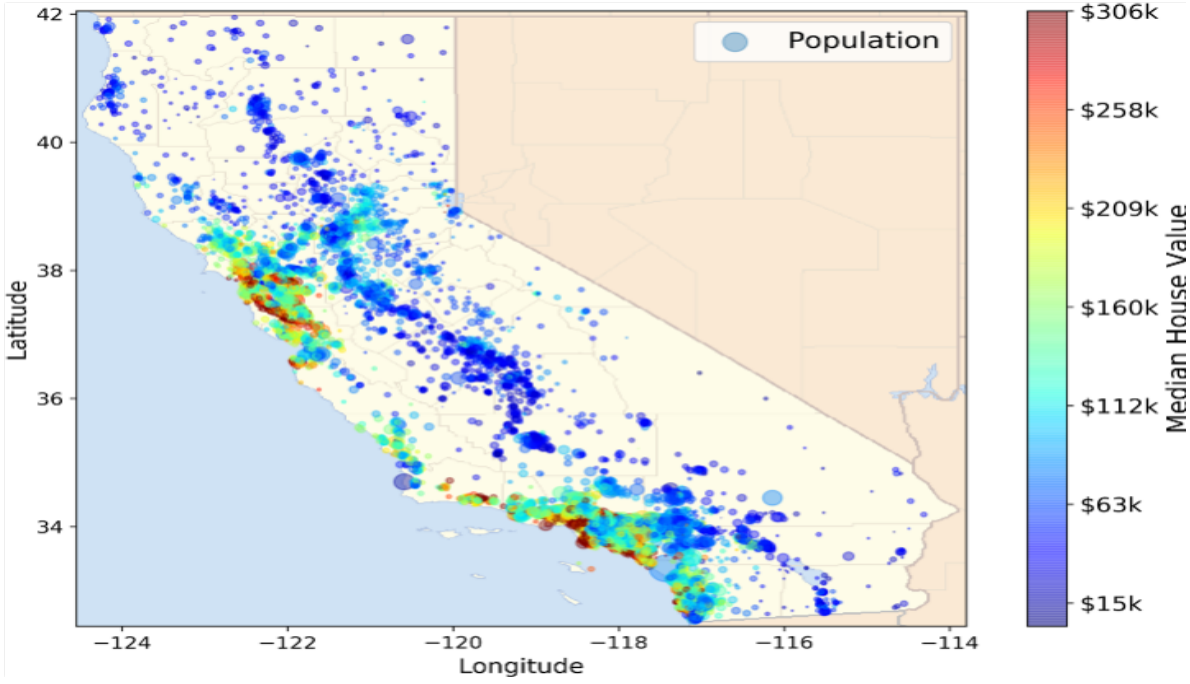

Figure 1: California housing prices

Working with Real Data II

Our first task is to use California census data to build a model of housing prices in this state.

- This data includes metrics such as the **population**, **median income**, and **median housing price** for each block group in California.
- Block groups are the smallest geographical unit for which the US Census Bureau publishes sample data.
- A **block group** or **district** typically has a population of 600 to 3,000 people.

Modelling Task:

Our model should learn from this data and be able to predict the median housing price in any district, given all the other metrics.

Modelling Objective

Modelling Objective I

The first question is what exactly the **modelling objective** is.

- Building a model is probably not the end goal.

Knowing the objective is important because it will **determine**:

- how you **frame** the problem.
- which **algorithms** to select.
- which **performance measure** to utilize in order to **evaluate** the selected model.
- how much **effort** is required in order to **tweak** the model.

Modelling Objective II

Our model's output (a prediction of a district's median housing price) will be fed to another Machine Learning system (see Fig. 2), along with many other signals.

- This downstream system will determine whether it is worth investing in a given area or not.

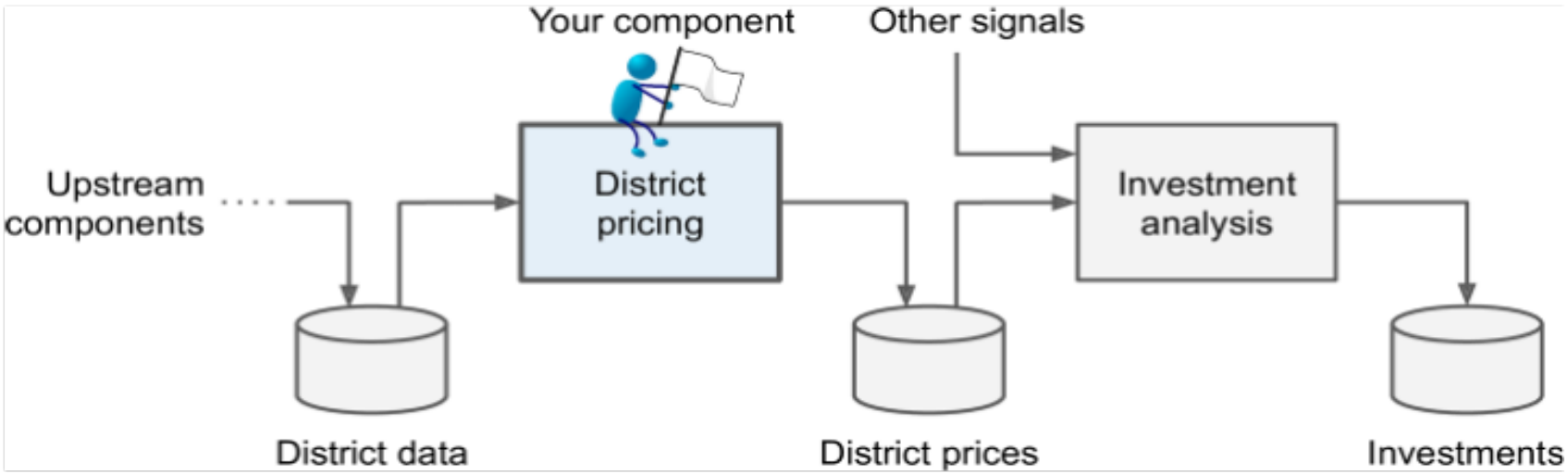

Figure 2: A Machine Learning pipeline for real estate investments.

Pipelines

Pipelines:

A sequence of data processing components is called a data pipeline.

Pipelines are very common in Machine Learning systems, since there is a lot of data to manipulate and many data transformations to apply.

- Components typically run asynchronously.
- Each component pulls in a large amount of data, processes it, and spits out the result in another data store.
- The next component in the pipeline pulls this data and spits out its own output.
- Each component is fairly self-contained: the interface between components is simply the data store.
- If a component breaks down, the downstream components can often continue to run normally (at least for a while) by just using the last output from the broken component.

Performance Measures

Machine Learning models will be trained according to the RMSE and will be evaluated according to both RMSE and MAE.

Root Mean Squared Error:

$$RMSE(\boldsymbol{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}^{(i)}) - y^{(i)})^2} \tag{1}$$

Mean Absolute Error:

$$MAE(\boldsymbol{X}, h) = \frac{1}{m}\sum_{i=1}^{m}|h(\boldsymbol{x}^{(i)}) - y^{(i)}| \tag{2}$$

Housing Dataset

Housing Dataset

The housing dataset is a collection of **20640** records containing the following attributes:

1. **longitude**: numeric
2. **latitude**: numeric
3. **housing median age**: numeric
4. **total rooms**: numeric
5. **total bedrooms**: numeric (contains missing values)
6. **population**: numeric
7. **households**: numeric
8. **median income**: numeric
9. **median house value**: numeric
10. **ocean proximity**: numeric

Dataset Investigation I

The **median income** attribute does not look like it is expressed in US dollars (USD).

- The corresponding data has been **scaled** and **capped** at 15 (actually, 15.0001) for higher median incomes, and at 0.5 (actually, 0.4999) for lower median incomes.
- Thus, the numbers represent roughly tens of thousands of dollars (e.g., 3 actually means about $30,000).

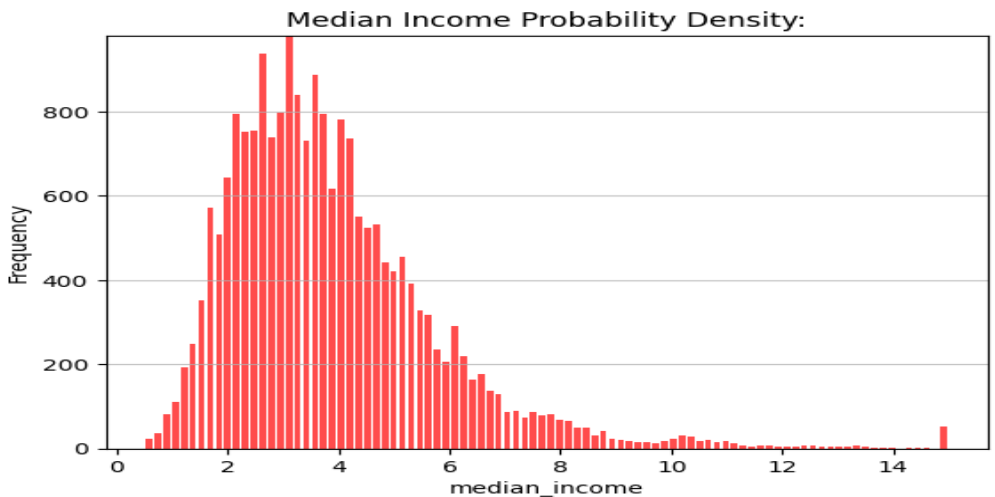

Figure 3: Probability Density Distribution for the Median Income

Dataset Investigation II

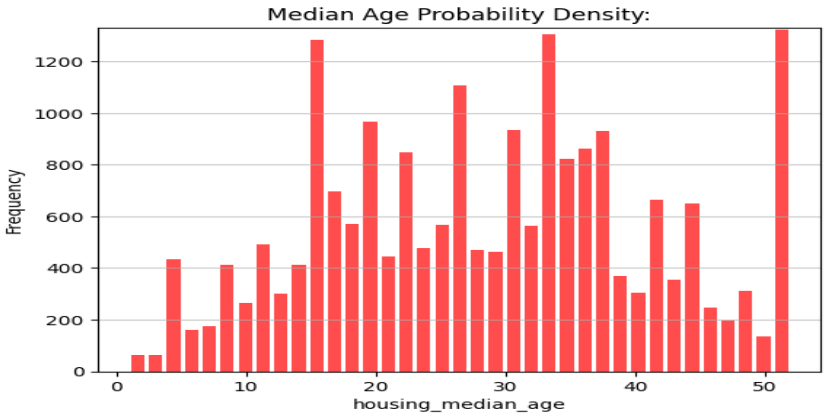

Figure 4: Probability Density Distribution for the Median Age

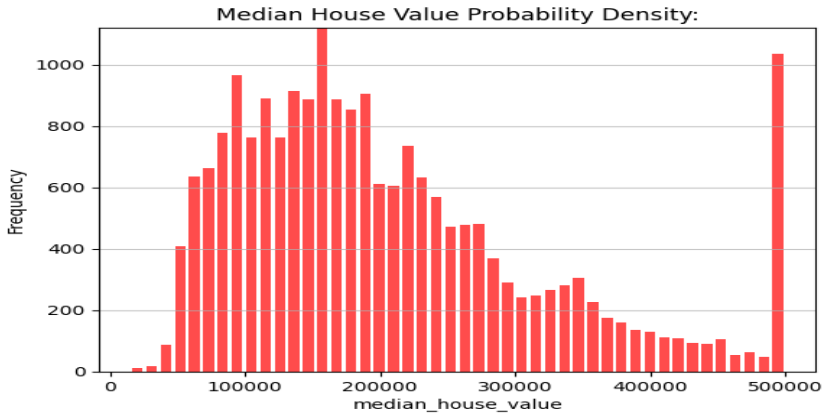

Figure 5: Probability Density Distribution for the Median House Value.

The **housing median age** and the **median house value** were also capped. The latter may be a serious problem since it is out **target attribute** (our **labels**).

Data Investigation III

Data Preprocessing:

- Working with preprocessed attributes constitutes a common practise in Machine Learning and is not necessarily a problem. It is important, however, to acquire a deeper understanding on the exact mechanism that generated the data.
- Capped values may induce a serious problem when such attributes provide the target labels for the utilized Machine Learning algorithms (e.g. Median House Value).
  - Machine Learning algorithms may learn that prices never go beyond the limit of $500,000.
  - Collect proper labels for the districts whose labels were capped.
  - Remove those districts from the training set (and also from the test set, since our system should not be evaluated poorly if it predicts values beyond $500,000).

Data Investigation IV

Data Preprocessing:

- These attributes have very different scales. Therefore, **feature scaling** should be explored.
- Many histograms are **heavy-tailed**:
  - the corresponding attributes' values extend much farther to the right of the median than to the left.
  - Machine Learning algorithms may face a harder problem in detecting patterns for such cases.
  - We will employ **data transformation** techniques so that the utilized features follow a more **bell-shaped** distribution.

Training & Testing Subsets

Generating Training and Testing Datasets I

Generating training & testing subsets of data may be conducted by employing **pure random sampling** methods on the original dataset.

- This is, in general, a well-accepted approach when the volume of the available data is sufficiently large, especially, relative to the number of attributes.
- On the opposite case, we run the risk of introducing a significant **sample bias**.
- When considering the problem of predicting the median house price, the median income constitutes an extremely important feature.
- Therefore, it is of critical importance to ensure that the training and testing subsets of data are representative of the various categories of incomes in the whole dataset.

Generating Training & Testing Datasets II

Stratified Sampling:
Performing a thorough investigation on the median income histogram, reveals that most of the median income values are contained within the $[1.5, 6.0]$ interval (i.e. $[\$15,000, \$60,000]$). Some of the median incomes, however, go far beyond 6.0

- It is important to preserve a sufficient number of instances from each stratum, or else the estimate of a stratum's importance may be biased.
- This means that you should not have too many strata, and each stratum should be large enough.

Generating Training & Testing Datasets III

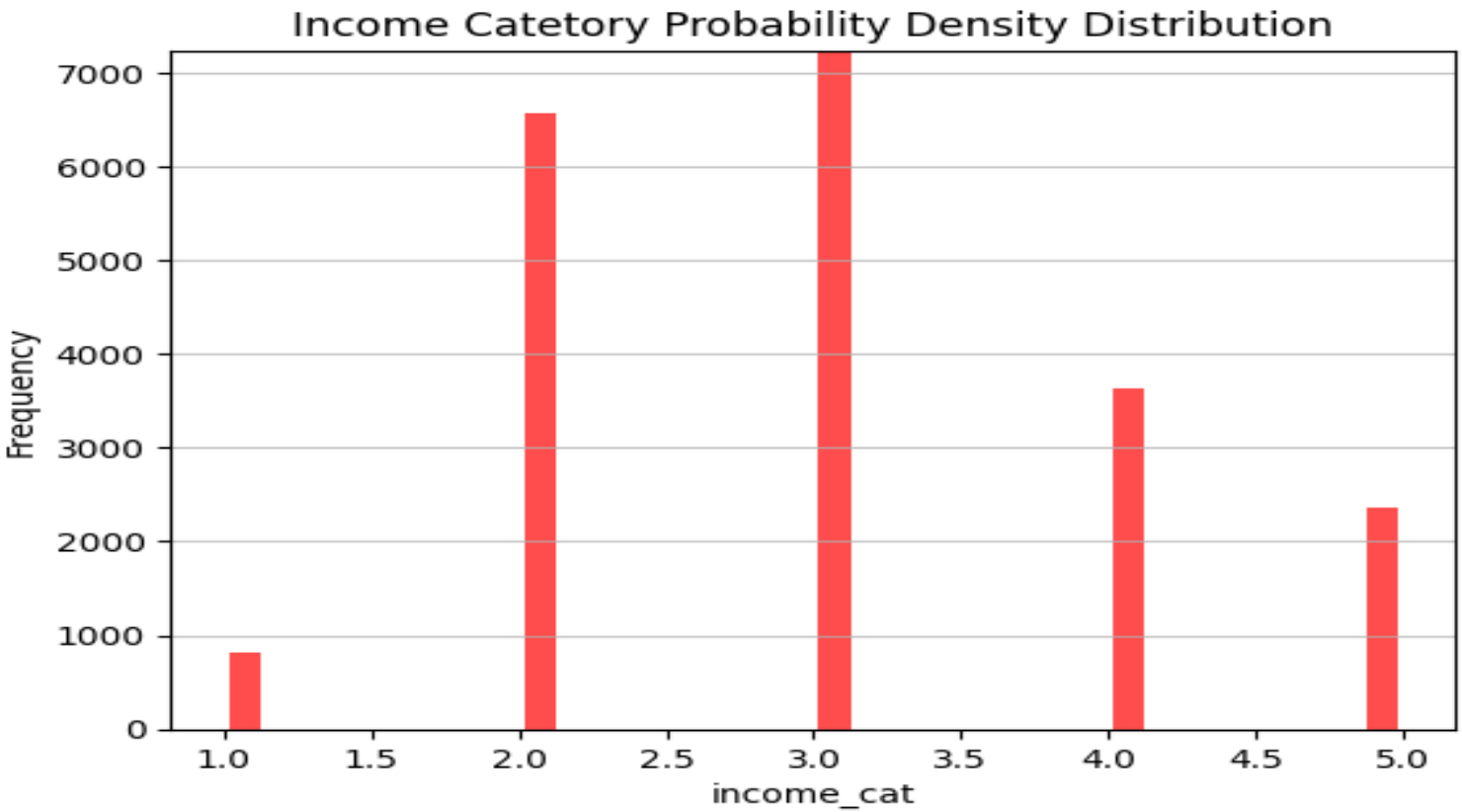

Figure 6: Stratified Median Income

Generating Training & Testing Datasets IV

Income Class	Overall	Stratified	Random
1	0.039826	0.039729	0.040213
2	0.318847	0.318798	0.324370
3	0.350581	0.350533	0.358527
4	0.176308	0.176357	0.167393
5	0.114438	0.114583	0.109496

Table 1: Stratified vs Random Sampling

Frequency of each income category within the original dataset, the stratified test dataset and the randomly sampled test dataset.

Data Visualization

Data Visualization

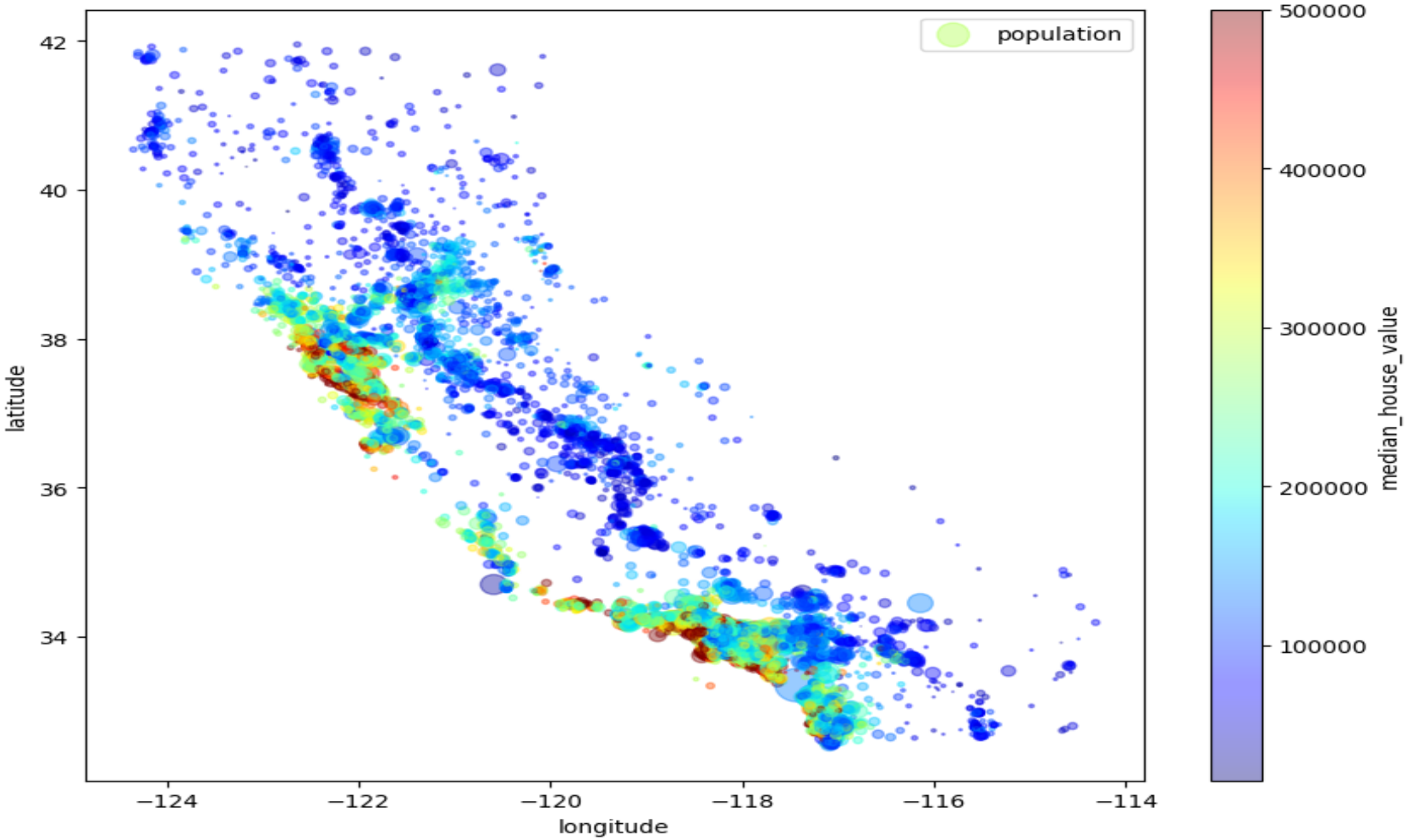

Figure 7: What can we understand from the above Figure?

Looking for Correlations

Looking for Correlations I

- The previous image tells us that the housing prices are very much related to the location (e.g., close to the ocean) and to the population density.
- Since the dataset is not too large, we can easily compute the standard correlation coefficient (also called Pearson's r) between every pair of attributes given by the following formula:

Pearson Correlation:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{3}$$

Looking for Correlations II

Correlation Coefficients:

- Correlation coefficients range within the $[-1, +1]$ interval.
- Coefficients close to $+1$ indicate the existence of a strong positive correlation. For example, the median house value tends to go up when the median income goes up.
- Coefficients close to -1 indicate the existence of a strong negative correlation. For example, there exists a small negative correlation between the latitude and the median house value (i.e., prices have a slight tendency to go down when you go north).
- Coefficients close to 0 indicate that there is no linear correlation.

Looking for Correlations III

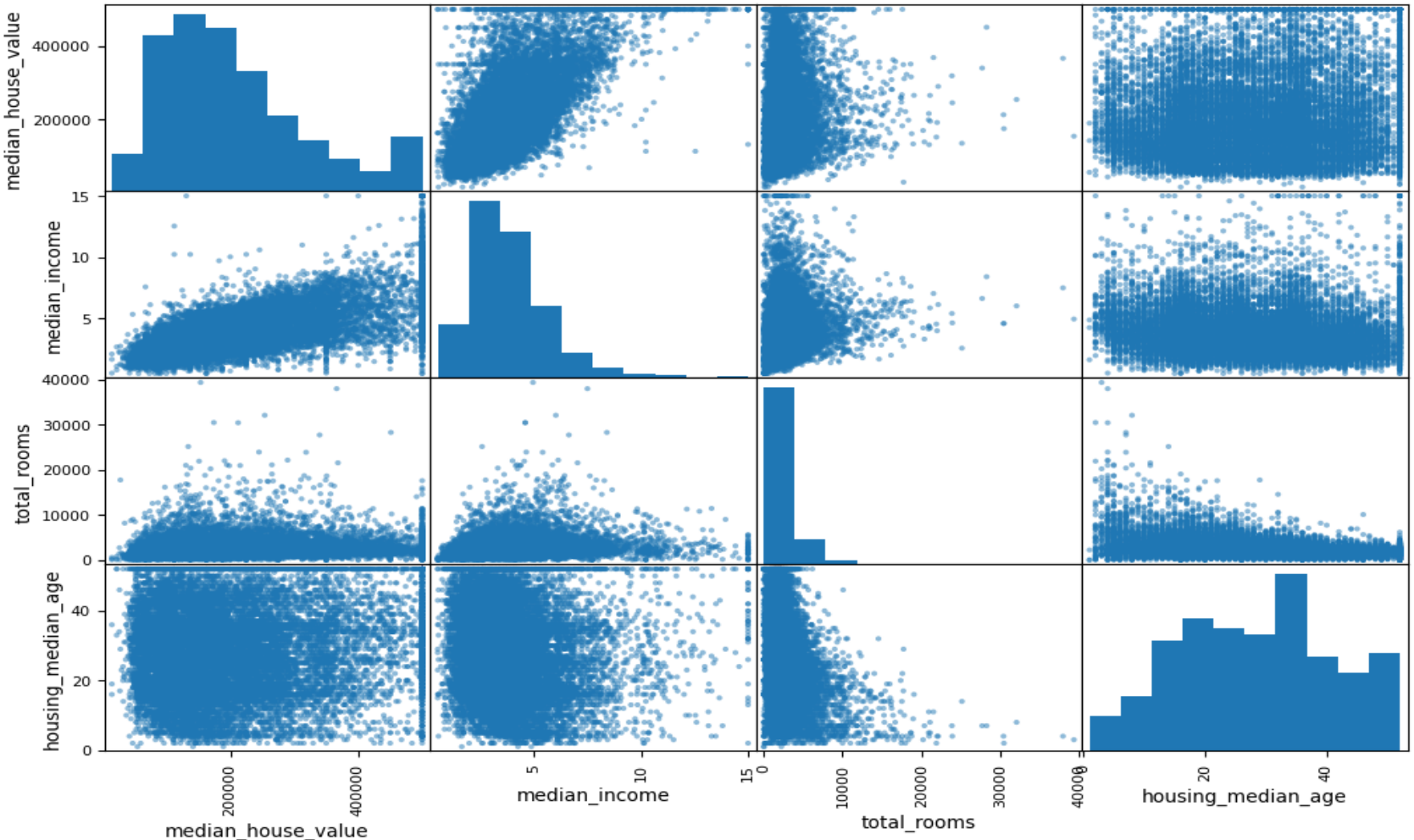

Figure 8: This scatter matrix plots every numerical attribute against every other numerical attribute, plus a histogram of each numerical attribute.

Looking for Correlations IV

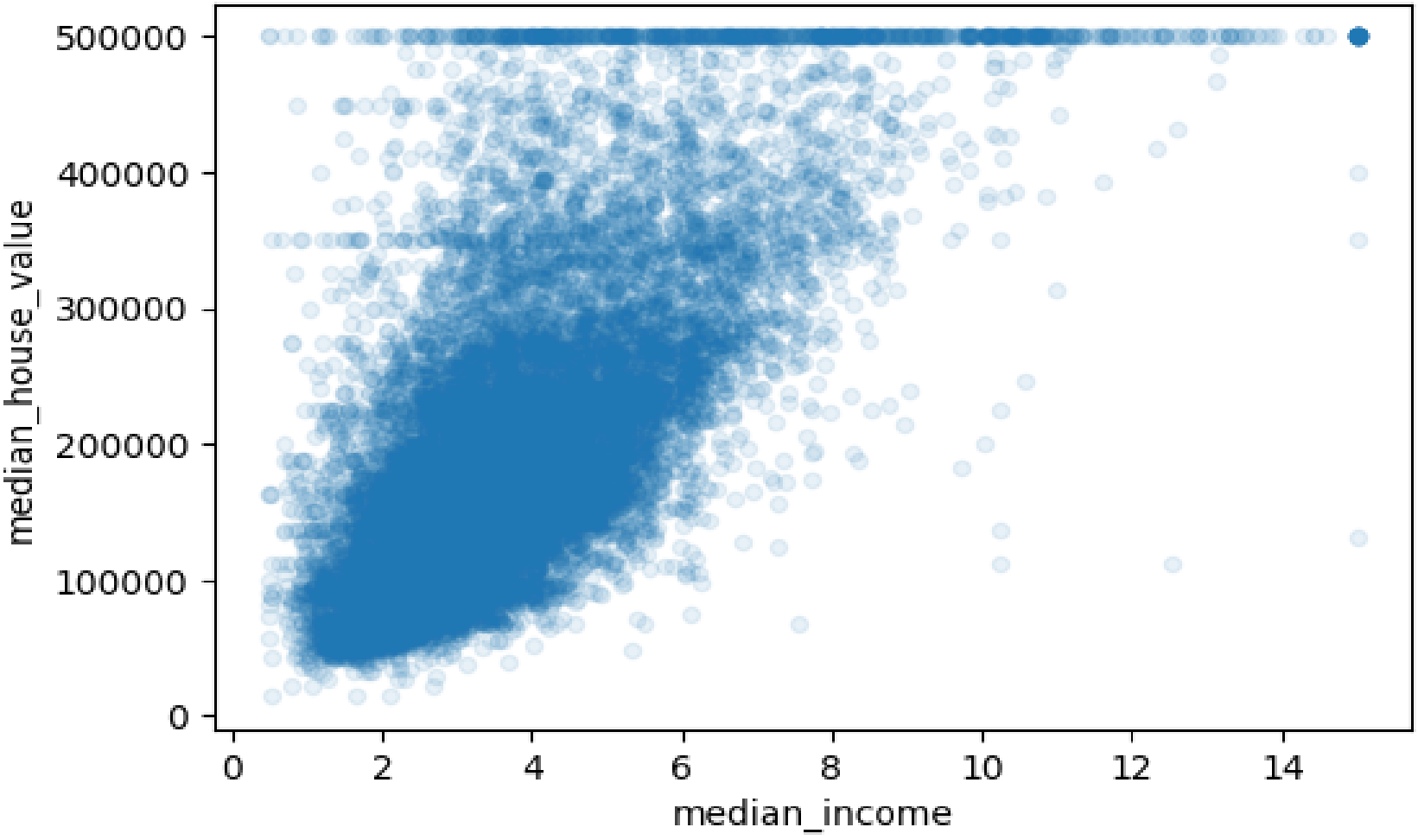

Figure 9: Median income versus median house value.

Looking for Correlations V

The previous plot allows us to arrive to the following conclusions:

- The correlation is indeed very strong; you can clearly see the upward trend, and the points are not too dispersed.
- The price cap that we noticed earlier is clearly visible as a horizontal line at $\$500,000$.
- This plot also reveals other less obvious straight lines:
  - A horizontal line around $\$450,000$, another around $\$350,000$ perhaps one around $\$280,000$, and a few more below that.
  - One may want to try removing the corresponding districts to prevent machine learning algorithms from learning to reproduce these data quirks.

Feature Engineering

Feature Engineering I

One could consider to form various combinations of attributes before preparing the data for any machine learning algorithm.

Feature Combinations:

- For example, the total number of rooms in a district is not very useful if you don't know how many households there are. What one would really want is the number of rooms per household.
- Similarly, the total number of bedrooms by itself is not very useful. It would be wiser to compare it to the number of rooms.
- Finally, the population per household also seems like an interesting attribute combination to look at.

Feature Engineering II

- The new **bedrooms per room** attribute is much more correlated with the median house value than the total number of rooms or bedrooms.
- Apparently houses with a lower **bedroom / room ratio** tend to be more expensive.
- The number of **rooms per household** is also more informative than the total number of rooms in a district obviously the larger the houses, the more expensive they are.