**Dataset:**

Context

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

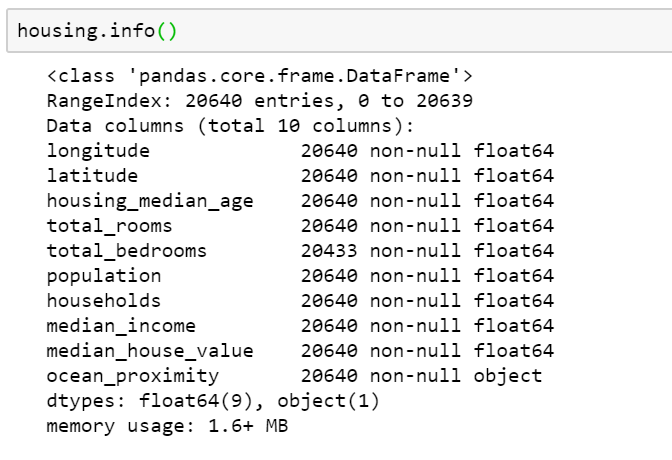
Content

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. Be warned the data aren't cleaned so there are some preprocessing steps required! The columns are as follows, their names are self explanitory:

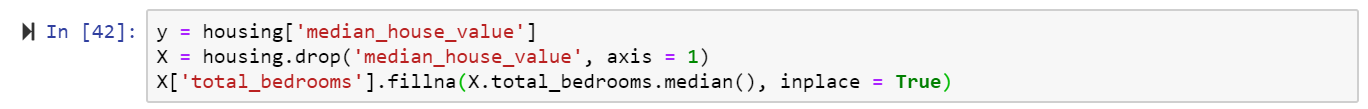
1. **longitude**: A measure of how far west a house is; a higher value is farther west
2. **latitude**: A measure of how far north a house is; a higher value is farther north
3. **housingMedianAge**: Median age of a house within a block; a lower number is a newer building
4. **totalRooms**: Total number of rooms within a block
5. **totalBedrooms**: Total number of bedrooms within a block
6. **population**: Total number of people residing within a block
7. **households**: Total number of households, a group of people residing within a home unit, for a block
8. **medianIncome**: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. **medianHouseValue**: Median house value for households within a block (measured in US Dollars)
10. **oceanProximity**: Location of the house w.r.t ocean/sea

**Implementation**

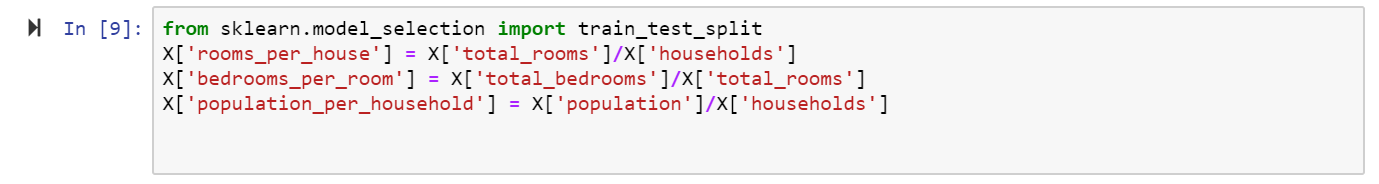
First of all in my dataset there is 171 data that is missing in total\_bedrooms features. So first we have to fix that problem because we can not split as well as we can not predict data if our dataset has some missing values.



So here we must use total\_bedrooms median to fill those missing data. And for that I’m using fillna method of pandas to fill total\_bedrooms.



After that I add some more features to make little bit more perfect dataset. Features like rooms\_per\_house, bedrooms\_per\_house, and population\_per\_house. And we can calculate all these features by below equations.



After that I use Linear Regression method to calculate mean squared error and it’s 84222 and that’s huge, we can do better than that one way is we can add more features, or we can handle **ocean\_Proximity feature. After handling this feature, I again use** Linear Regression method and this time the error is 61083. So here we can tell that we improved our error by approximately 20000 and that is huge. I also tried different method like lasso, ridge, elasticnet, random forest regression.

**And the result is**

|  |  |
| --- | --- |
| **Method** | **mean\_squared\_error** |
| **Lasso Regression** | **61509** |
| **Ridge Regression** | **68229** |
| **Elasticnet Regression** | **69873** |
| **Random forest Regression** | **50109** |
| **linear Regression** | **61083** |

**Observation:**

If we want to predict house values and make this dataset highly accurate then we have to add some features like

1. The nearest city.
2. Distance from the region's coordinates to the next-door city.
3. The nearest 'big city' (a big city is categorized as >300,000 residents in the given year, 1990).
4. The distance to the nearest 'big city'.
5. The mall