**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 pretty self explanitory:

longitude

latitude

housing\_median\_age

total\_rooms

total\_bedrooms

population

households

median\_income

median\_house\_value

ocean\_proximity

**Acknowledgements**

This data was initially featured in the following paper:  
Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297.

and I encountered it in 'Hands-On Machine learning with Scikit-Learn and TensorFlow' by Aurélien Géron.  
Aurélien Géron wrote:  
This dataset is a modified version of the California Housing dataset available from:  
[Luís Torgo's page](http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html) (University of Porto)