This project uses the California Housing Prices dataset at <https://www.kaggle.com/camnugent/california-housing-prices>. The data contains information from the 1990 California census. It has an easily understandable list of variables, and has 20640 rows of data, which is manageable size.

I have borrowed heavily from Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'.

## Data Cleaning and Wrangling

There are 207 non-available values in the total-bedrooms column, and the correlation of total-bedrooms to medium\_housing\_value is very low, at 0.049686, I have decided to drop total-bedrooms from the dataset.

The ocean\_proximity column contains only string values [‘NEAR BAY’, ‘<1H OCEAN’, ‘INLAND’, ‘NEAR OCEAN’, ‘ISLAND’]. To make it easier to plot, I have replaced the string values by integers[‘ISLAND’=0, ‘NEAR BAY’=1, ‘NEAR OCEAN’=2, ‘1H OCEAN’=3, ‘INLAND’=4].

Since the correlations of medium\_housing\_value to longitude(-0.0459), population(-0.0246), and households(0.065843) are also low, I have dropped these 3 columns as well.

## Different Models

Using Statsmodel to fit Linear Regression is very easy and straightforward. Sklearn, on the other hand, produces surprisingly close results. The f-statistics are about 5673 for both.

The Decision Tree Regression is not suitable for this dataset.

The Random Forest Regression produces the best result. The f-statistic is 100054.

The SVR has f-statistic as 2095.

## Model Fine Tuning

I have used GridSearchCV and RandomizedSearchCV to fine-tune the Random Forest Regression model. The f-statistics have reached 144319.