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 in 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. Therefore, I have decided to drop total-bedrooms from the dataset.

The ocean_proximity column contains only string values. 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 somewhat close results. The f-statistics are about 4586 and 4341, respectively.

The Decision Tree Regression is not suitable for this dataset, the f-statistic is 2776.

The Random Forest Regression, on the other hand, produces the best result. The f-statistic is 7650.

The SVR has f-statistic 1373.

## Model Fine Tuning

I have used GridSearchCV and RandomizedSearchCV to fine-tune the Random Forest Regression model. The f-statistics have reached 7723, slightly better than 7650 as before.