# Case Study 2: Linear regression exercise

### Initial data inspection

The below information is provided to help you work with the data set.

California Housing dataset characteristics:

Number of Instances: 20640

Number of Attributes: 8 numeric, predictive attributes and the target

Attribute Information:

* MedInc median income in block group
* HouseAge median house age in block group
* AveRooms average number of rooms per household
* AveBedrms average number of bedrooms per household
* Population block group population
* AveOccup average number of household members
* Latitude block group latitude
* Longitude block group longitude

This dataset was obtained from the StatLib repository. <https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html>

The target variable is the median house value for California districts, expressed in hundreds of thousands of dollars ($100,000).

This dataset was derived from the 1990 U.S. census, using one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

A household is a group of people residing within a home. Since the average number of rooms and bedrooms in this dataset are provided per household, these columns may take surprisingly large values for block groups with few households and many empty houses, such as vacation resorts.

The goal is to build a Linear Regression Model on the California Housing Dataset. We will be trying to predict the MedianPrice per house in each block. Denoted as the target column. This price is in $100,000s

We have two datasets available:

* housing\_data.csv is the original dataset
* housing\_data\_cleaned\_STD.csv is the same dataset but having had 114 rows removed with missing AveRooms. It has also had all points above or below 3 standard deviations removed.

In your lab environment, load in the dataset with an appropriate method and create a new notebook to work with the data, completing the following tasks.

### Data Familiarisation

### Using your statistical analysis and data visualisation (EDA) techniques, familiarise yourself with the dataset. Seek to understand:

### the types of data.

### distributions of features.

### potential correlating features.

### obvious patterns.

### errors, nulls, outliers.

### Assumptions

1. Select the feature you would like to use for linear regression.
2. Would you use all the rows in the data set or a sample of your data for linear regression? If using a sample, which sampling method would you use?

**Preparation of the data**

### To fit a Linear Regression model, we will need to remove any rows with missing values. We would typically investigate these further and treat them, but for convenience just drop the rows that have empty values in either chosen feature or target columns.

### Linear regression

1. After preparing the dataset, we can fit our regression model. Import the Linear Regression model from scikit-learn’s linear\_model module.
2. Create a Linear Regression model, using the fit method to train the model on x and y.
3. Generate predicted values for some sample or example data.
4. Create a scatter plot with x,y and your predictions showing.

**Stretch exercise**

1. Calculate the pearson correlation coefficient and p value of your chosen linear regression method.
2. Calculate the R2 score and residuals of your linear regression model.
3. Assess if the chosen feature is useful for predicting the target and what might be missing.