# Description of Problem

We want to be able to predict the average price of a neighbourhood in California based on the characteristics of the given census block. We will be using the California Housing dataset which is obtained through the StatLib repository (but this dataset was originally derived from the 1990 U.S census).

**Dataset link:** https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\_california\_housing.html#sklearn.datasets.fetch\_california\_housing

# How are we going to solve the problem?

We will be implementing two different algorithms (in python) a ***Regression Decision Tree*** and a ***KNN Model*** to help predict the housing price of a given census block. Both algorithms will help see what attributes affect an average housing price more or less than another. We will be using ***K-Fold cross validation*** to train both these given algorithms so we make sure we are using these algorithms to their fullest.

# Comparison

By using two different algorithms we will be able to compare the results from one another which will help us to see if we can find any correlation between particular attributes. Thus affecting an average housing price in a census block and also being able to predict their average price.

We will also be able to see which algorithm gives us more accurate results than another by a couple of different ways. Firstly a confusion matrix will be used as a visually simple way of seeing which algorithms are classifying data incorrectly and correctly. We will also use an ROC curve to see how precise the algorithm is (and the AUC). We will then be able to see how well the two algorithms hold up against each other and if one is ‘better’ than the other (Better in terms of efficiency like time, processing power etc).