# **案例:The Boston Housing Dataset**

1.<https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>

2.<https://www.kaggle.com/code/linjoe1219/sklearn-boston-house-prices-regressiontree>(中文版)

3.<https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7>

# 波士頓房價資料集特徵說明

* 'CRIM': 城鎮人均犯罪率
* 'ZN': 占地面積超過2.5平方英尺的住宅用地比例
* 'INDUS': 城鎮非零售業務地區的比例
* 'CHAS': 查爾斯河（Charles River）變量（若土地位於河流邊界，則為 1；否則為 0）
* 'NOX': 環保指數
* 'RM': 每棟住宅的房間數
* 'AGE': 1940 年以前建成的自住單位的比例
* 'DIS': 距離 5 個波士頓的就業中心的加權距離。
* 'RAD'距離高速公路的便利指數
* 'TAX'每一萬美元的不動產稅率
* 'PTRATIO'城鎮中的教師學生比例
* 'B'城鎮中的黑人比例
* 'LSTAT'地區中有多少房東屬於低收入人群

# 步驟說明

1. 使用sklearn的boston資料集，發現字典格式的資料中有'data'、'target'、'target\_names'、'feature\_names'可加以利用。
2. 用pandas整理成表格
3. 用df.corr()畫出特徵係數矩陣
4. 用matplotlib.pyplot及seaborn劃出feature\_names的熱度圖，以便觀察特徵之間的關係。
5. 用train\_test\_split分出測試資料以及目標資料(輸入輸出資料)
6. 將測試資料fit進DecisionTreeRegressor
7. 使用export\_graphviz以及graphviz畫出分類樹，可解釋性強
8. 防止過擬合: 先剪枝或後剪枝(設定max\_depth)
9. 決策樹優點: 說服力強，缺點: 調整參數需要經驗法則。