Homework 1 統計110張孟涵 h24064064

1. 從Kaggle中找一個資料集, 描述此資料相關 資訊, 包含:

• 問題定義

• 潛在問題

• 分析與預測難度

• 價值

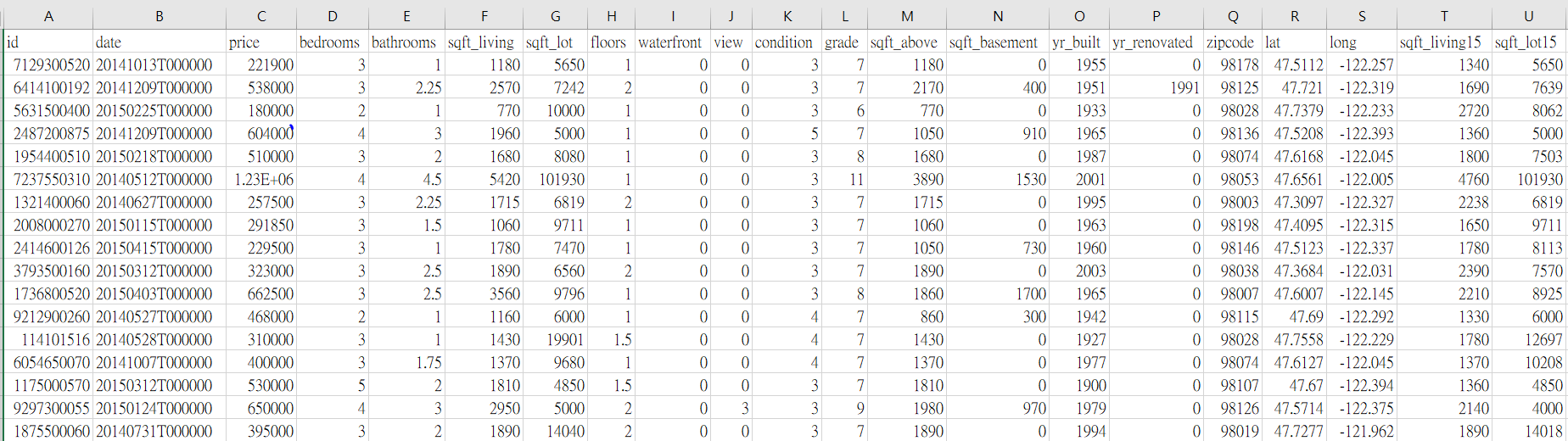
我從kaggle中找到一個資料集:House Sales in King County, USA

資料集簡介:

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

(<https://www.kaggle.com/harlfoxem/housesalesprediction>)

以下是資料大概的內容:



• 問題定義:

預測未來在 King County的房子的房價。

• 潛在問題:

1. 雖然房價通常會受到房子的大小、位置 、屋齡……等等在資料集有的feature影響，但是也可能會受到政治因素或是天災影響，例如:政府突然實施奢侈稅、該地區突然發生地震，這些都會使真實的房價和我們用這個資料集所建的模型有落差。
2. 這個資料集只有包含2014年5月及2015年5 月的房價資料，資料已經距離現在將近五年，且只有5月的資料，要考量他是否適用在預測未來的房價。

• 分析與預測難度:

1. 有些features 含有缺失值，例如: yr\_renovated (翻修年份)，但有些房子可能沒有翻修過，資料集是把它的值設零，但我覺得這不是個好方法，可能需要些缺失值處理。
2. 各個Feature的range相差太大，例如:bedroom的數量介在0~10，而sqft\_lot (停車場坪數) 介在690~1651359，所以可能需要將資料Normalize。
3. 若我們想把很多feature放到我們的模型，可能需要降維以提升效能。

• 價值

1. 建立能準確預測房價的模型，可提供房地產投資人作投資的決策。
2. 可讓想買房的人了解行情，不會當冤大頭。
3. 讓屋主知道行情，定出合理的價錢，增加競爭力。
4. 描述一個在你學習領域的資料或是日常生活 中可取得的資料, 說明可以用來產生何種價值 (預測, 分析, 別人不知道的事…)

整合政府及民營停車場提供的停車場即時剩餘車位數，可幫助駕駛尋找在目前位置最近的停車位，不會有到達停車場卻沒有停車位的窘境。