數資系 110813037 余珮綺

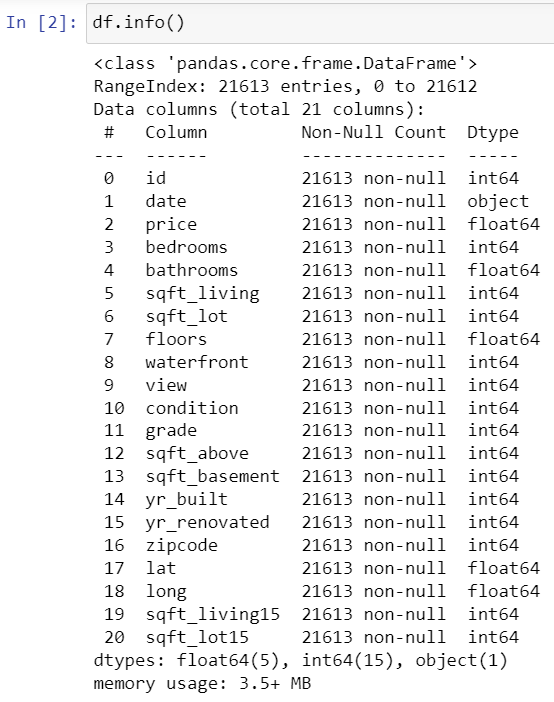
機器學習HW1\_Regression Analysis

程式碼語言: Python

內容:

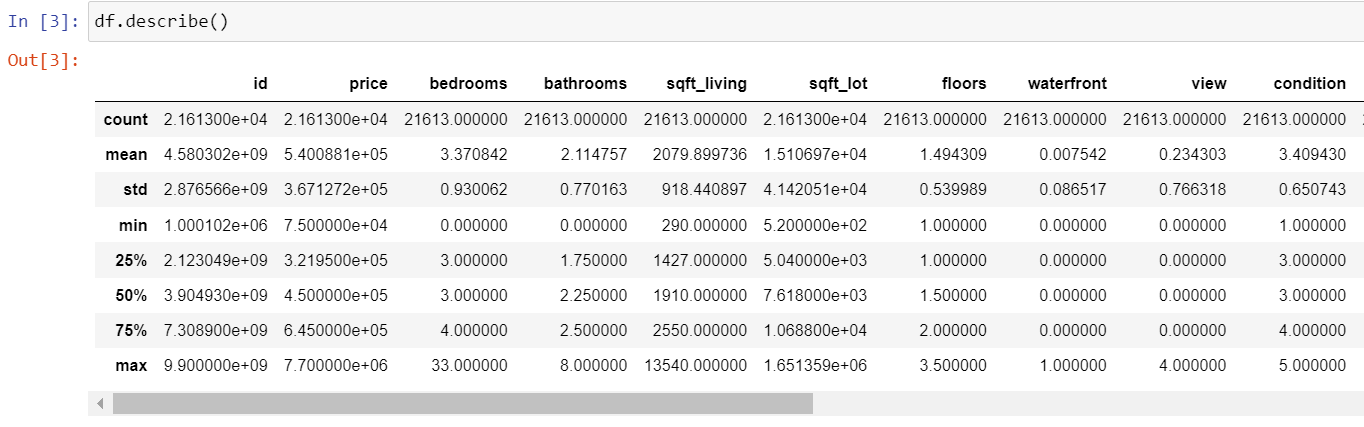
Part I:

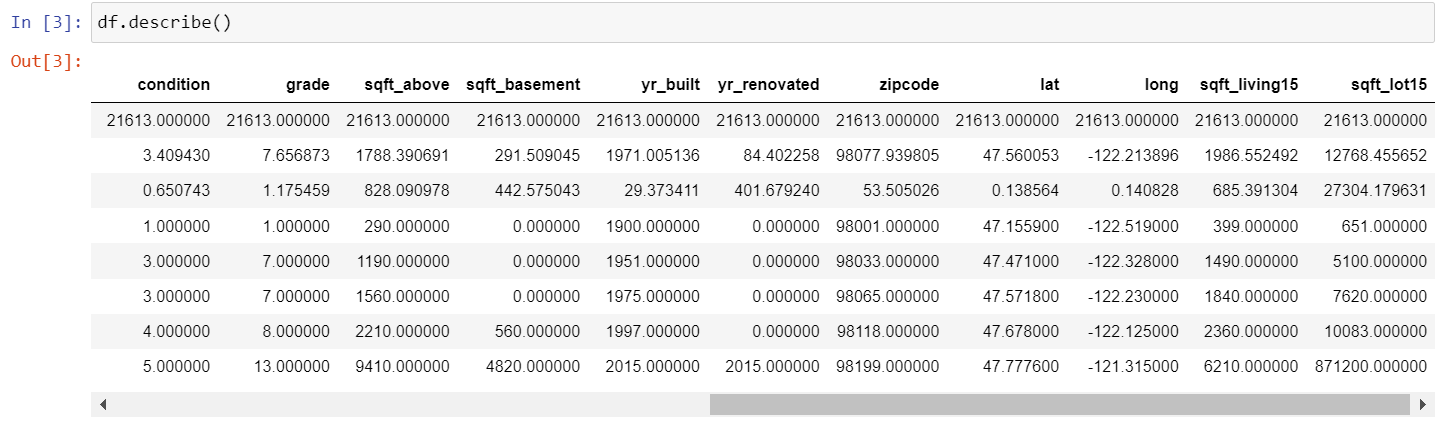
1. 資料清理與視覺化圖表



首先查看本house data的基本資訊，可看見除了日期外沒有類別型資料，也無缺失值，不須額外進行判斷是否要刪除或填補缺失值的動作。

1. 敘述性統計分析





從敘述統計可看見，有些數值較低者透過回去翻閱欄位說明檔案，可知像bedrooms和bathrooms為房間數量、waterfront的min為max為1則可知曉此欄為代表的是有無、condition分布為1至5的整數評分，諸如此類具有不同意義的數值。

根據「機器學習歷程 Chap.1 如何判斷變數的相關性(Pearson, Spearman or kendall tau?)」

(<https://medium.com/機器學習知識歷程/機器學習歷程-chap-1-eda-如何判斷變數的相關性-pearson-spearman-or-kendall-tau-c660e6465601>)上提及的測量尺度：名目尺度、順序尺度、區間尺度和比例尺度，再加上上一段所述具有不同意義的數值，可知在接下來進行相關係數的分析時，要考量到是適合使用Pearson還是Spearman。

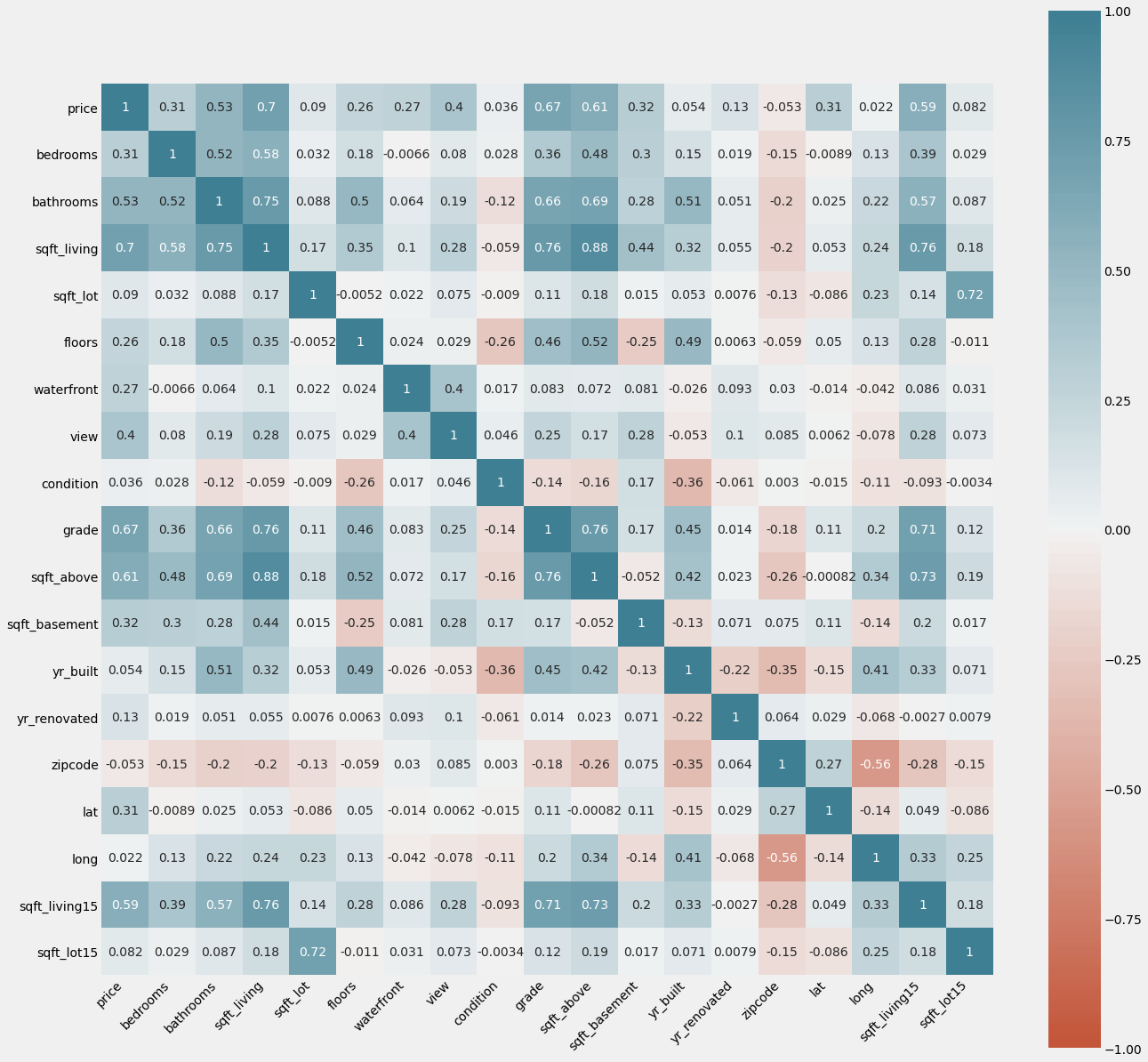
以下初步整理：

名目尺度：waterfront ->可使用Spearman

區間尺度：condition, grade, yr\_built, yr\_renovated -> 可使用Pearson

1. 特徵相關性分析

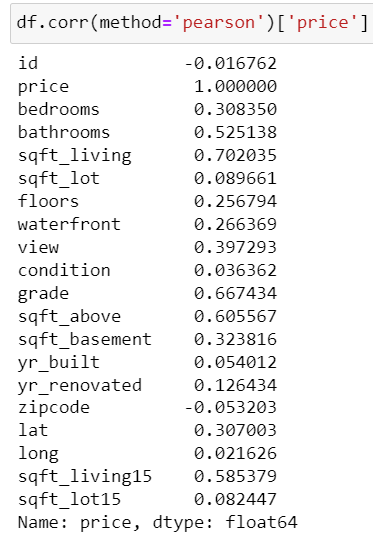
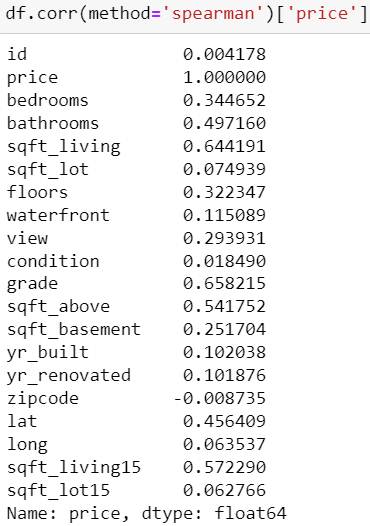
由於沒有需要特別處理的資料，因此直接使用熱圖查看各個欄位與price的相關係數



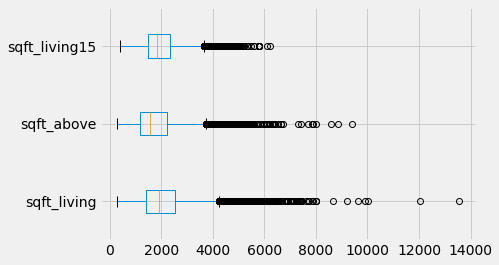
透過熱圖看見與price具較高度相關的幾個欄位類別為：

**bathrooms: 0.53、sqft\_living: 0.7、grade: 0.67、sqft\_above: 0.61、sqft\_living15: 0.59**

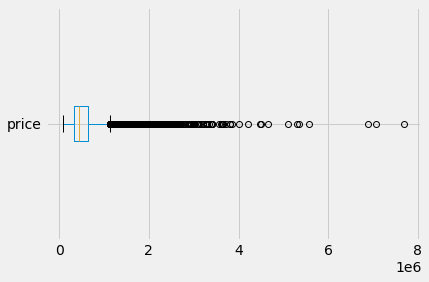
然而上述提及類別型資料與連續型資料，因此想再進一步確認接下來選取的特徵欄位：

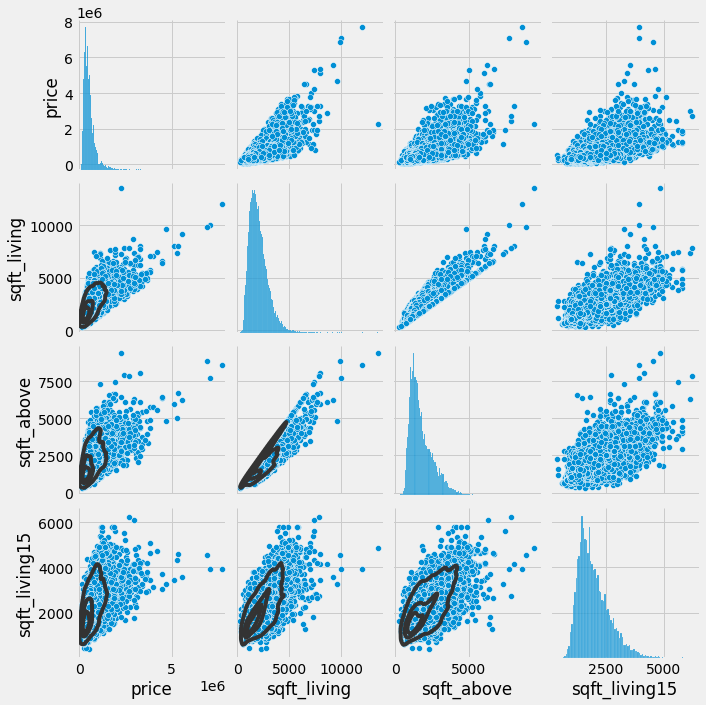
用紅框選取起來的可再次確認除了grade外其他的欄位對price的高低相關性很小



接下來，透過盒鬍圖查看sqft\_living, sqft\_above, sqft\_living的數值分布情形，發現不少離群值，因此想在進一步確認price是否也有一樣的情況，若無則需要針對這些離群值做額外的資料處理。



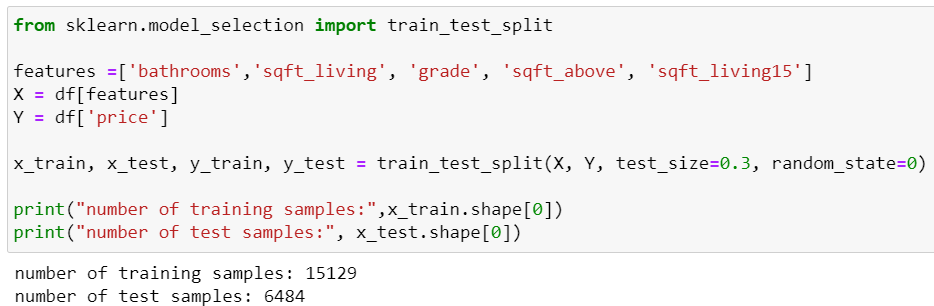
然而可發現，透過盒鬚圖知曉price也有不少離群值，因此可判斷不需做額外資料處理。



再用pairplot確認sqft\_living, sqft\_above, sqft\_living與price確實具一定相關性。

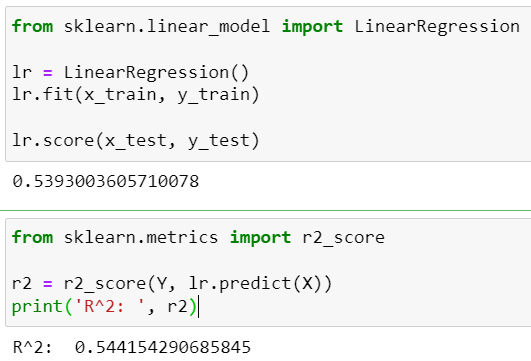
Part II:

1. 資料分割與建置迴歸模型 (線性 or 非線性)



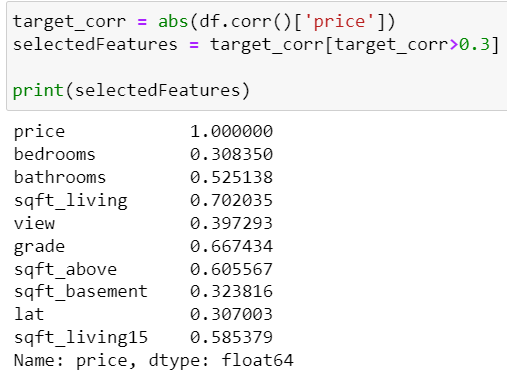
首先，將以上挑選出的幾個欄位作為特徵，並將資料進行7:3的分割

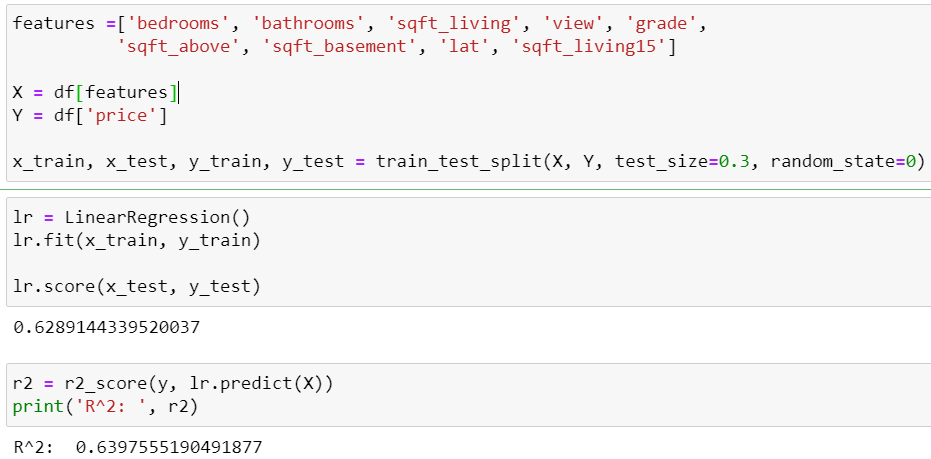
1. 模型效能評估



然而可見得建立簡單線性回歸模型的預測結果並不太好，score僅有0.54左右，R^2 score也在0.54左右

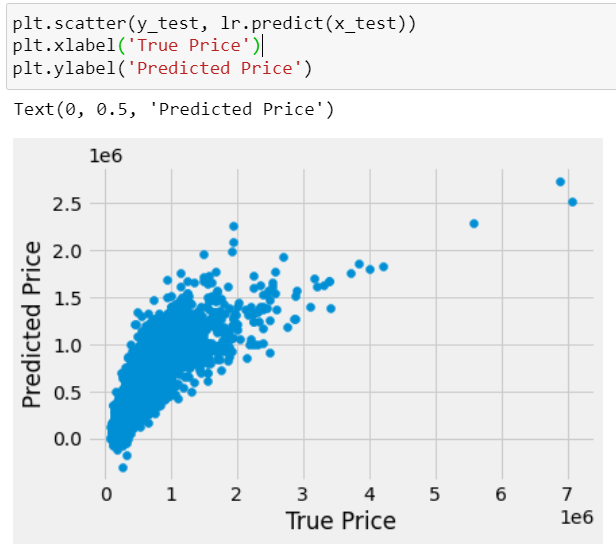
因此決定再重新篩選特徵欄位，將相關係數大於0.3的欄位直接納入分析





並可見得score獲得0.63左右，R^2則在0.64左右

1. 預測結果分析

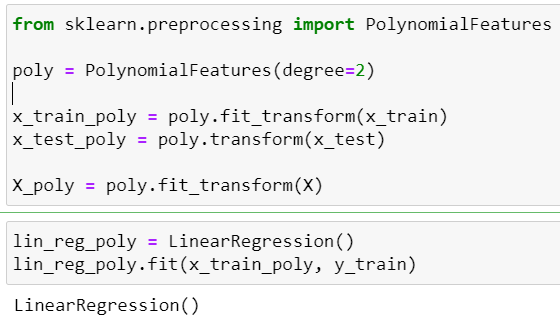


由本張圖可看見預測的價格仍與真實價格具有落差，或許該房價資料並不那麼適合使用簡單的線性回歸進行分析。

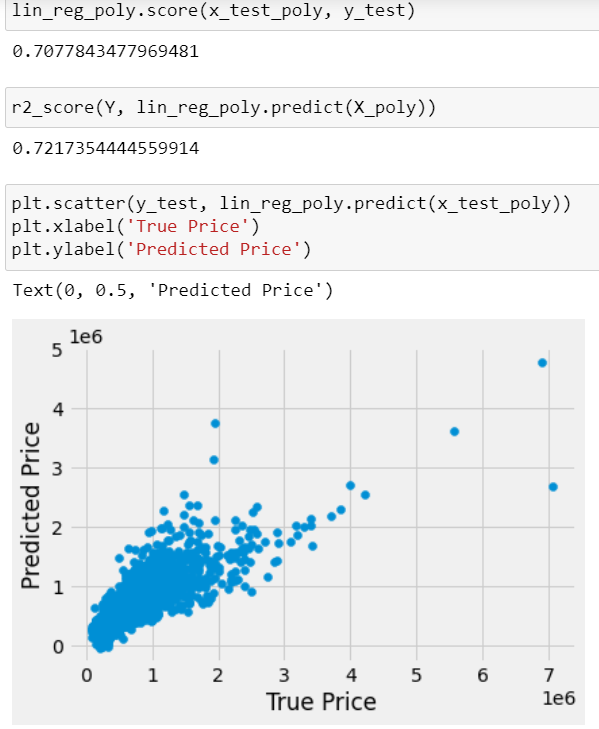
Part III:

1. 非線性迴歸模型 or 特徵變數變換

接續上個段落的分析結果，決定以polynomial作為優化該模型的方法。



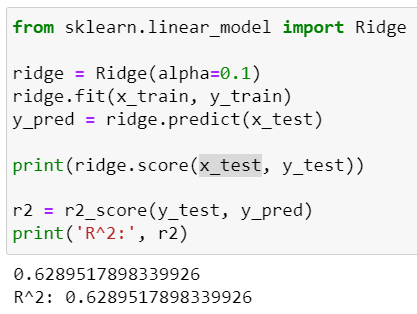
將維度設為2，並將X套入polynomial transform後建立模型。

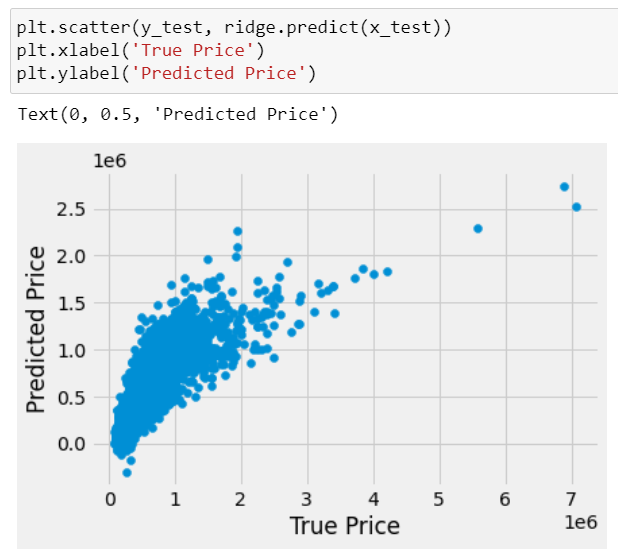


可發現此時不管是score還是R^2 score都比簡單線性回歸的模型表現更佳，至少都在0.7以上的準確率。

1. 迴歸模型之正規化

此外也使用正規化的回歸模型Ridge Regression進行分析。





可觀察到表現上遜於polynomial，評分數在0.63左右，和簡單線性回歸的表現差不多。

1. 相關模型結果比較分析

總地來說，本房價分析總共進行了四次的分析結果：

1. Simple linear regression, 使用5個特徵欄位，評分約為0.54
2. Simple linear regression, 使用相關係數大於0.3之所有欄位，評分約為0.64
3. Polynomial regression, 使用相關係數大於0.3之所有欄位，評分約為0.71
4. Ridge regression, 使用相關係數大於0.3之所有欄位，評分約為0.64

由以上結果可得對於本資料，利用Polynomial regression建立模型並分析更為合適。