資科系 111016011 陳奕

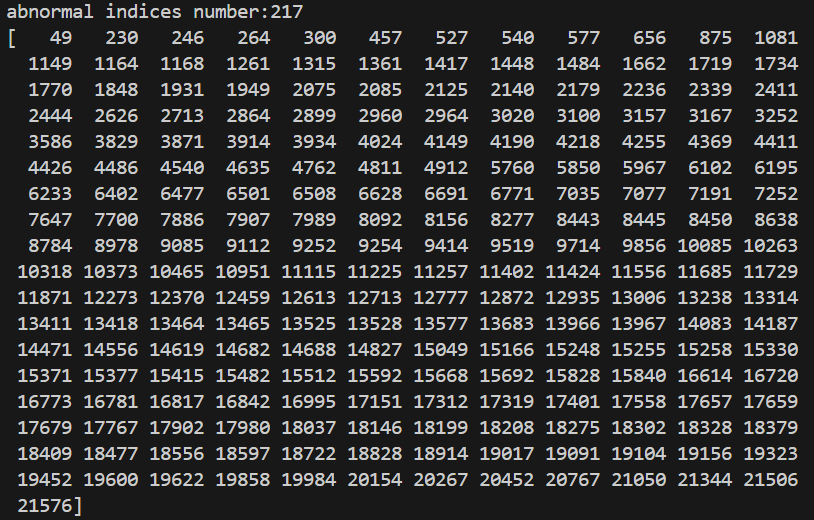
機器學習HW1\_Regression Analysis

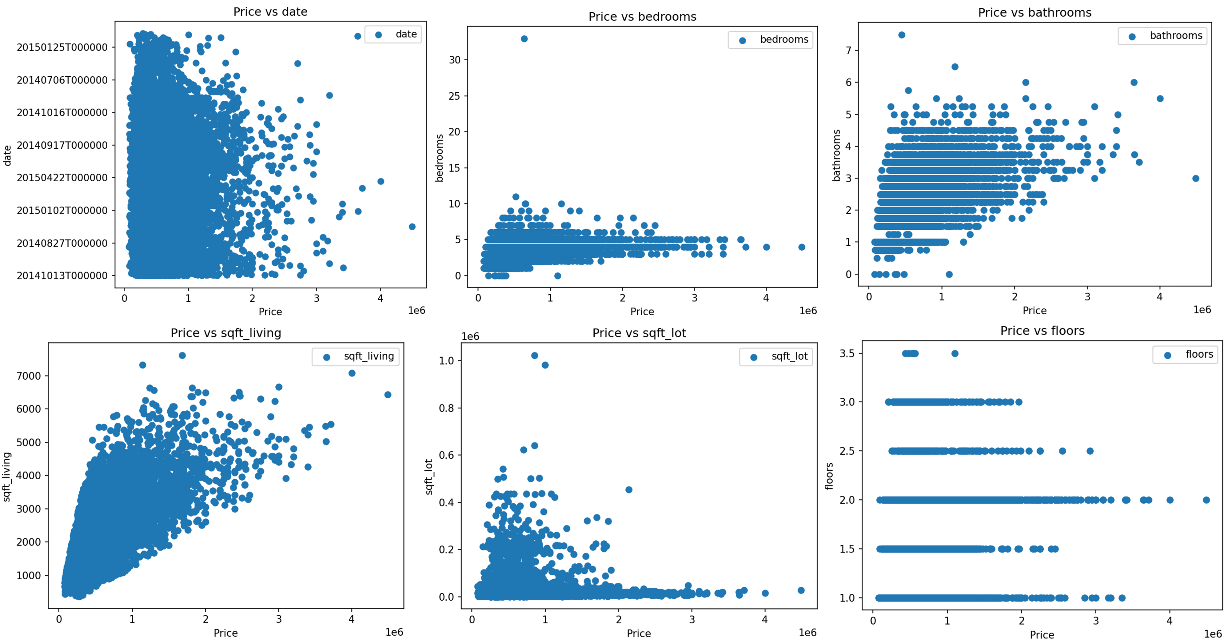
程式碼語言: Python

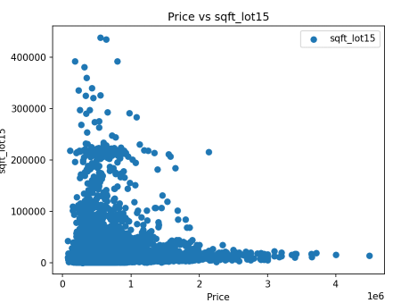
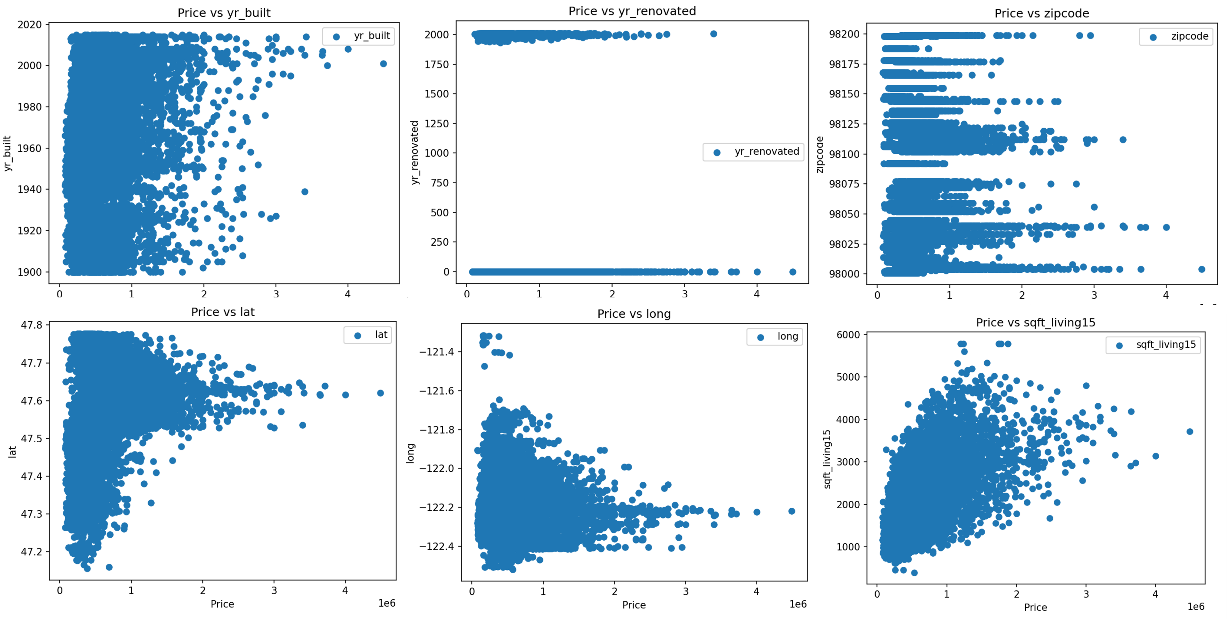
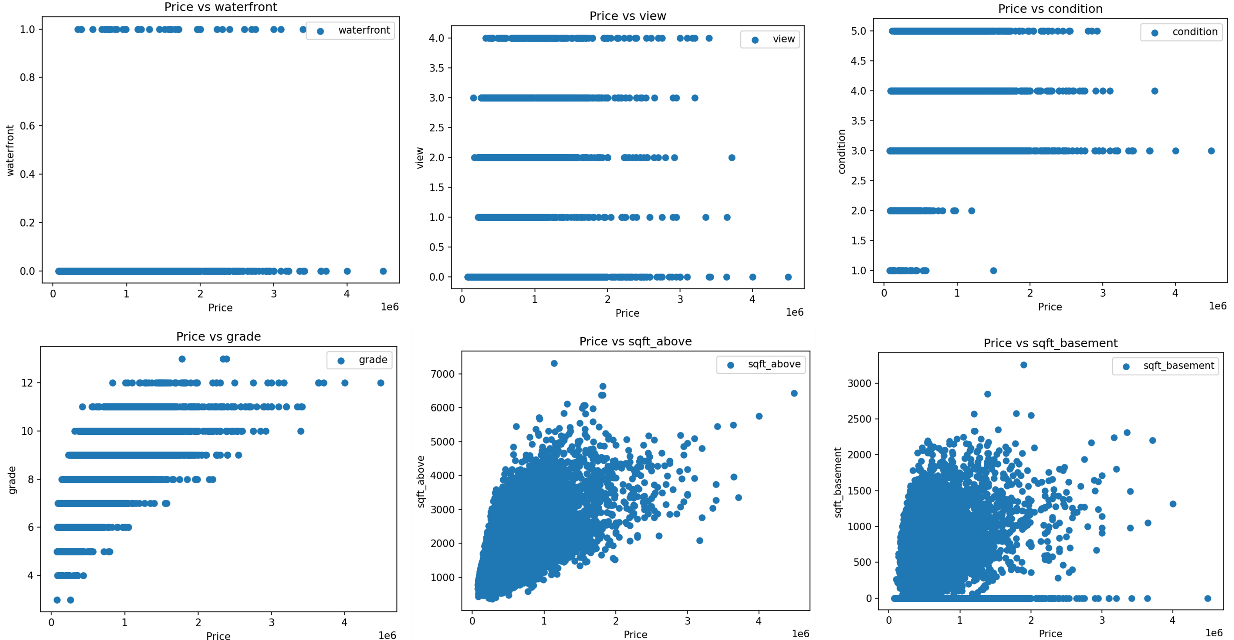
內容:

Part I:

1. 資料清理與視覺化圖表
2. 首先先確定資料是否有缺失(function 1)，每個特徵都有21613筆，沒有資料遺失
3. 接下來透過function 2來檢測離群值，由於特徵值較多，所以這裡使用孤立森林法來檢測，結果如下，共有217個離群值，透過function 3將其全部刪除並儲存成新的csv檔，新的csv檔即為清理過的資料



1. 最後使用function 4繪製視覺化圖表，由於id不怎麼重要所以將其刪除，在function4中，由於輸出date時，y軸會因為刻度太小而糊再一起，所以特別將date與其他特徵分開來。



1. 敘述性統計分析
2. 使用function 5 ，可以得出下列表格，而表格代表的含意由上到下代表：

count（計數）：非缺失值的數量。

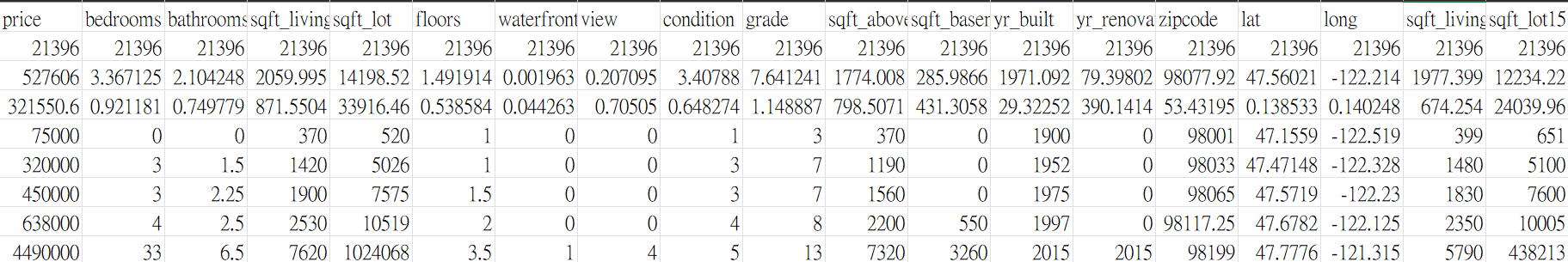
mean（平均值）：數值欄位的平均值。

std（標準差）：數值欄位的標準差。

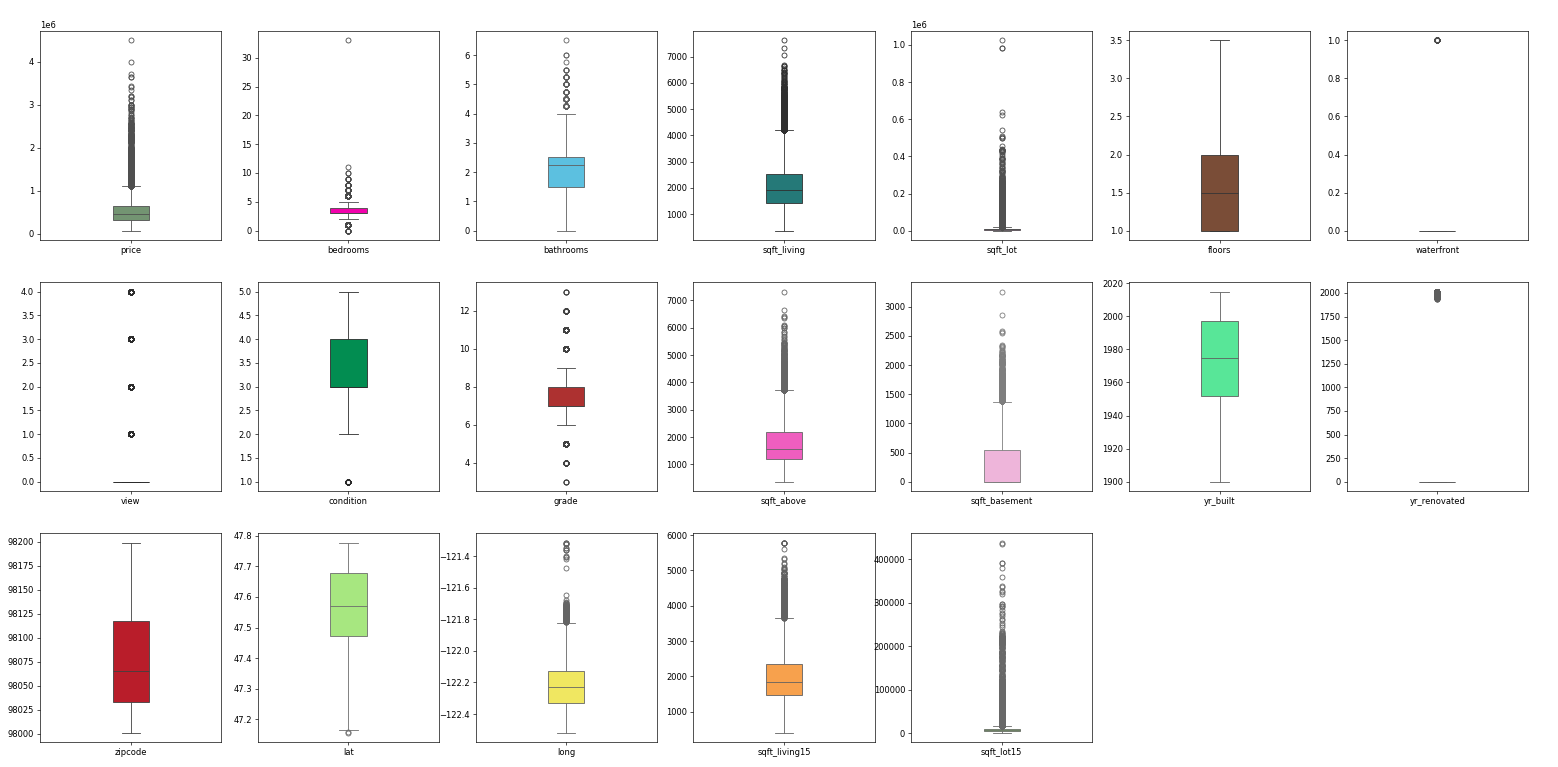
min（最小值）：數值欄位的最小值。

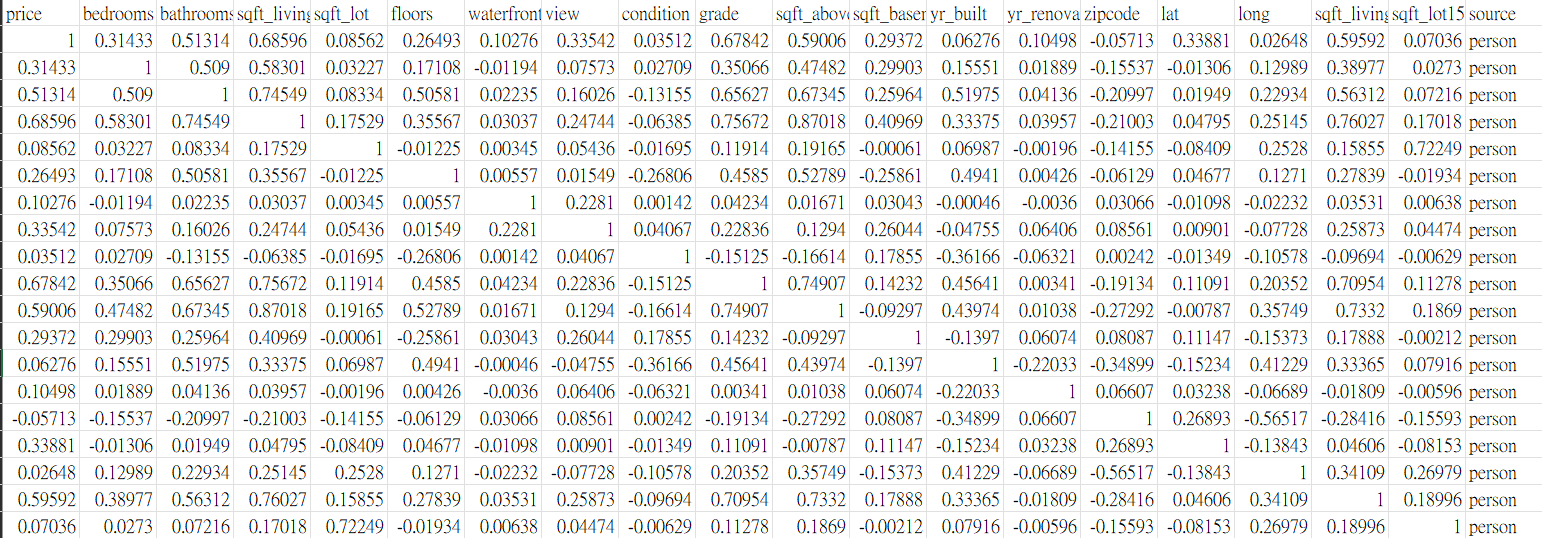
25%、50%、75%（四分位數）：四分之一、中位數和四分之三的百分位數。

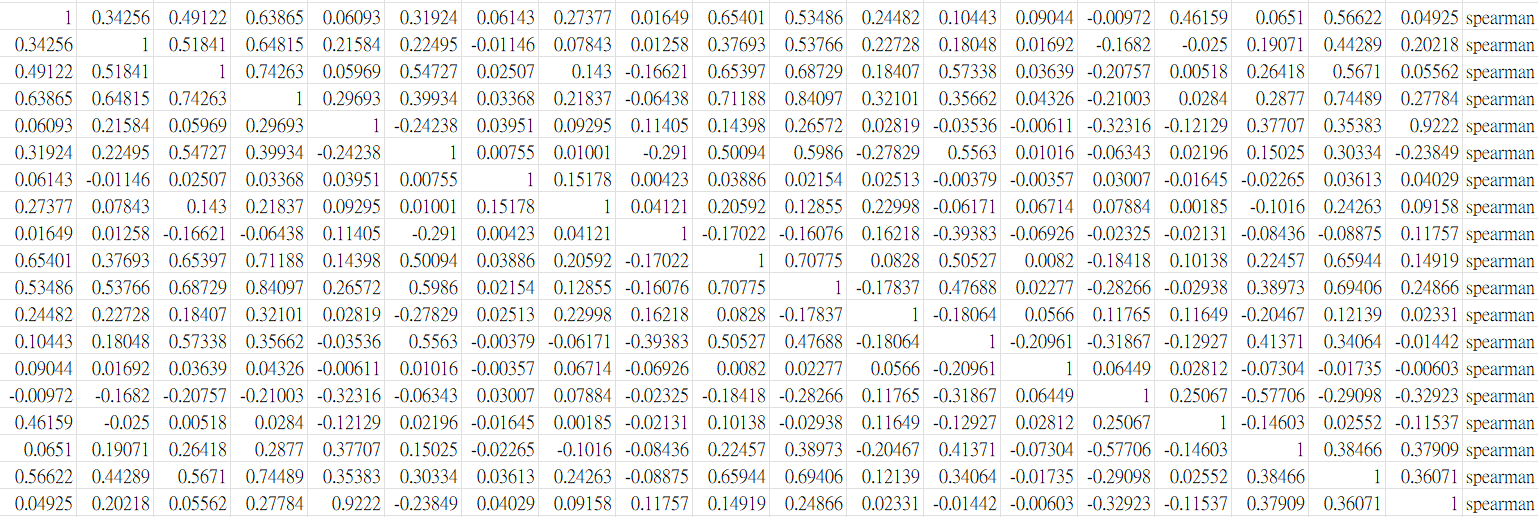
max（最大值）：數值欄位的最大值。

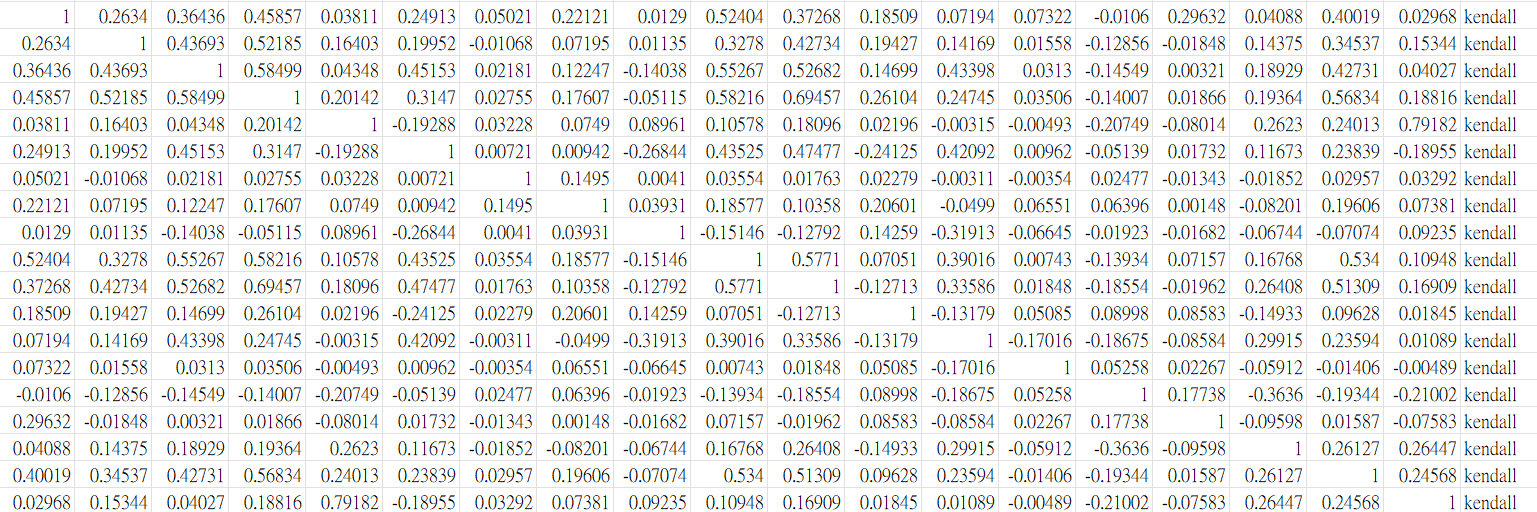


1. 使用function 6，畫出除了id外的所有四分圖，由於date的資料並非數值所以畫不出來，所以此處只畫出數值資料。

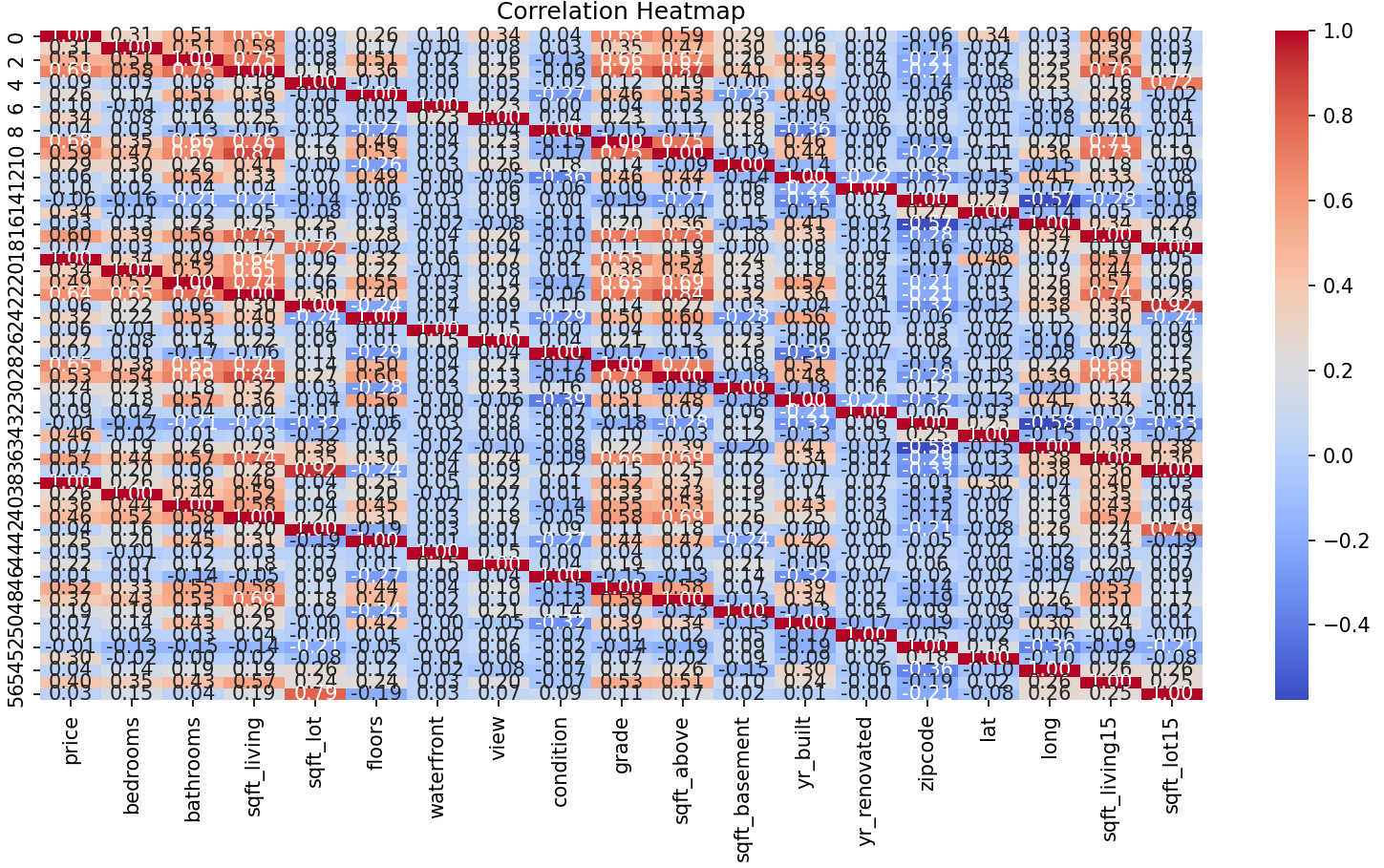


1. 特徵相關性分析
2. 首先透過function 7 將課本上的三種取相關係數(Person、Spearman、Kendall)的方式存在csv檔裡，由於出現的資料會全部混在一起，所以在此表格中，新增一欄來區分此係數是用何種方式取得。

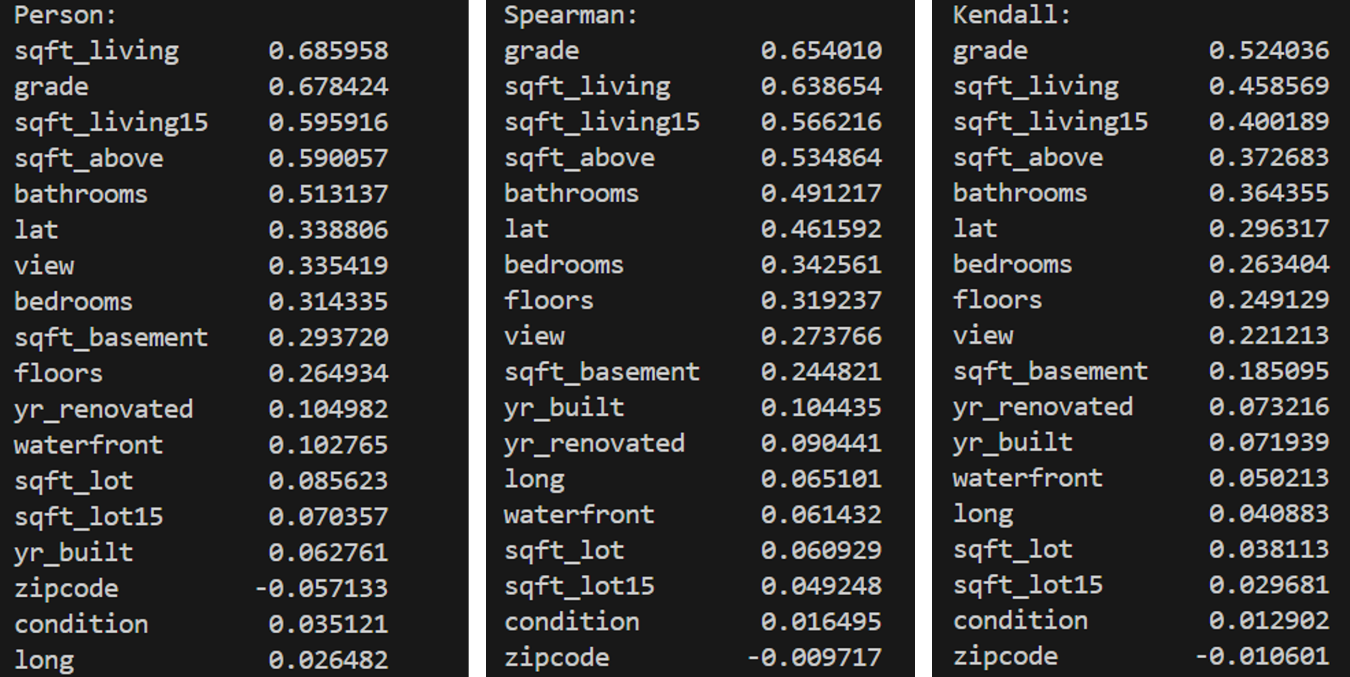




1. 再來用function 8 畫出熱度圖



1. 對price 跟各特徵值用上述三種方法的所得出的相關係數如下(使用function 9, 已經對所有值的絕對值進行排序)



可以發現，sqft\_living、grade、sqft\_living15、sqft\_above、bathrooms不論使用哪種計算方式都是與price 最相關的前五項。

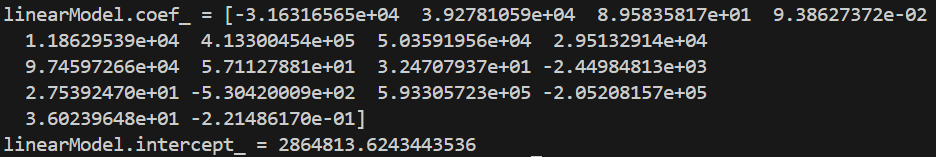
Part II:

1. 資料分割與建置迴歸模型 (線性 or 非線性)
2. 資料分割

使用function 2-1 將資料分成8：2( 80%為訓練集,20%為測試集,同時設定random\_state,以固定每次分割情形)

1. 建置迴歸模型(線性)

首先，由於各特徵的尺度都不同。由於自變數有很多個，所以使用多元線性迴歸，以下是結果



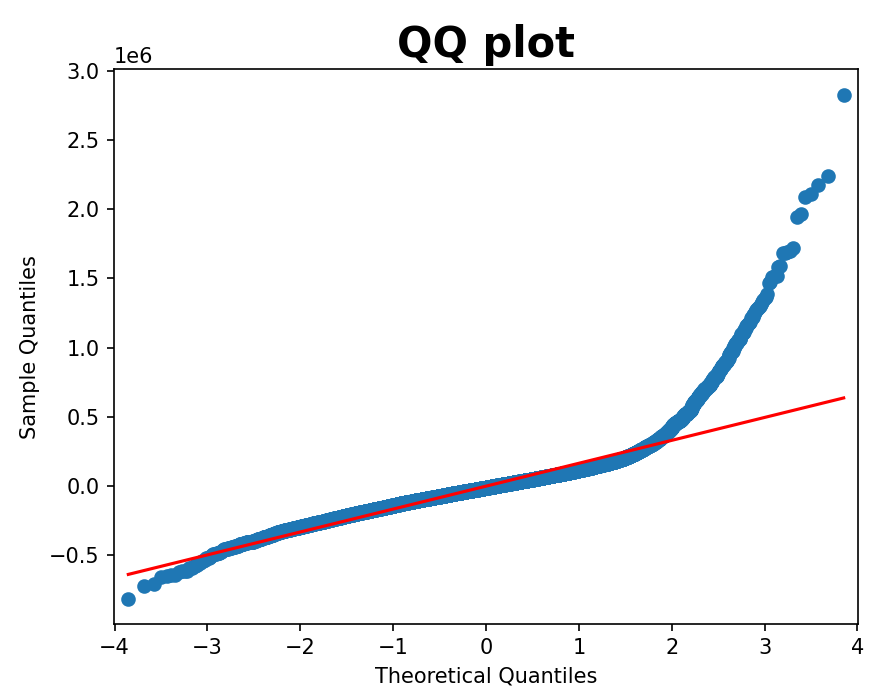
1. 模型效能評估
2. MAR 與 R^2

透過function 2-3 可以算出，此模型的MAE與R^2如下，可以發現R^2 在測試集比在訓練集還略低一點，但兩者都還是有0.6以上的成績。



1. QQ圖

透過function 2-4可以得到此模型的QQ圖，如下



1. 預測結果分析

模型的R^2為0.685，測試集為0.706，分數都接近0.7，而且分數並沒有相差太多，代表模型表現不錯且可以推論並沒有發生過擬合的情況，根據QQ圖的結果，在大部分的區間藍點與紅線都吻合，只在後段有些偏離。

Part III:

1. 非線性迴歸模型 or 特徵變數變換
2. 非線性迴歸模型

這裡使用function 3-1進行多項式迴歸，以下為degree為2與3的結果

Degree = 2

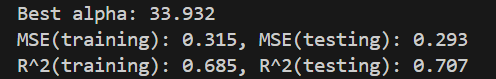


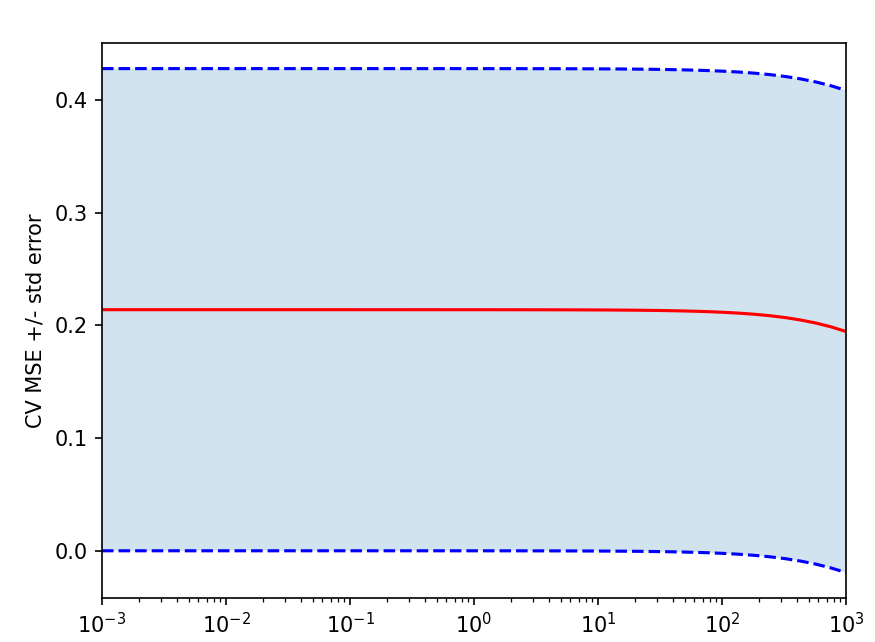
Degree = 3



在二次迴歸後，決定係數皆上升，但三次迴歸後，只有training 的上升，testing 的反而下降，可以推測發生過擬合的狀況。

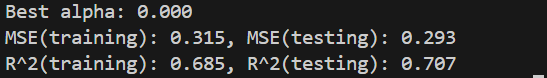
1. 迴歸模型之正規化
2. 脊迴歸

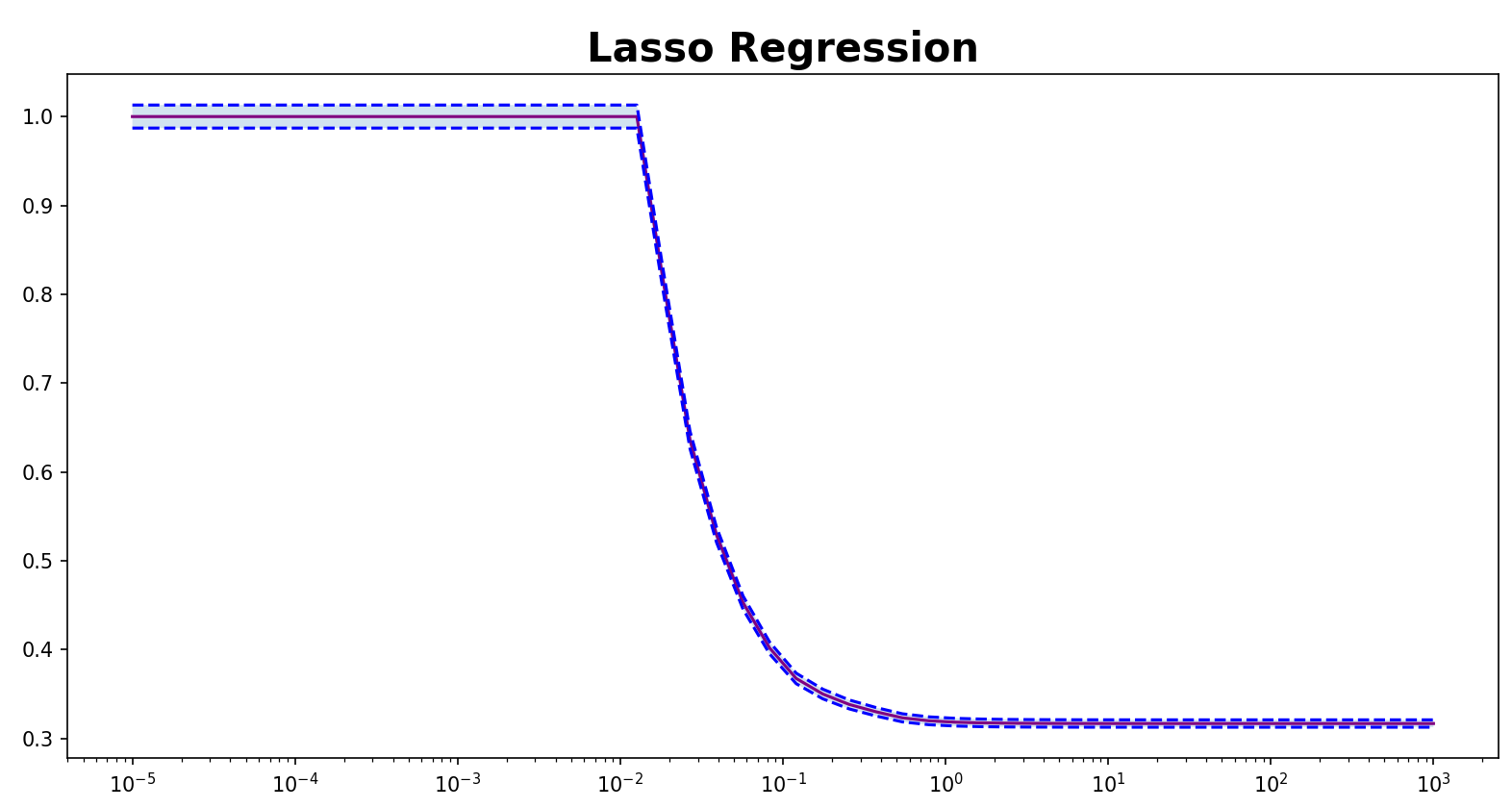
透過function 3-3 可以得出下列結果



1. 套索迴歸

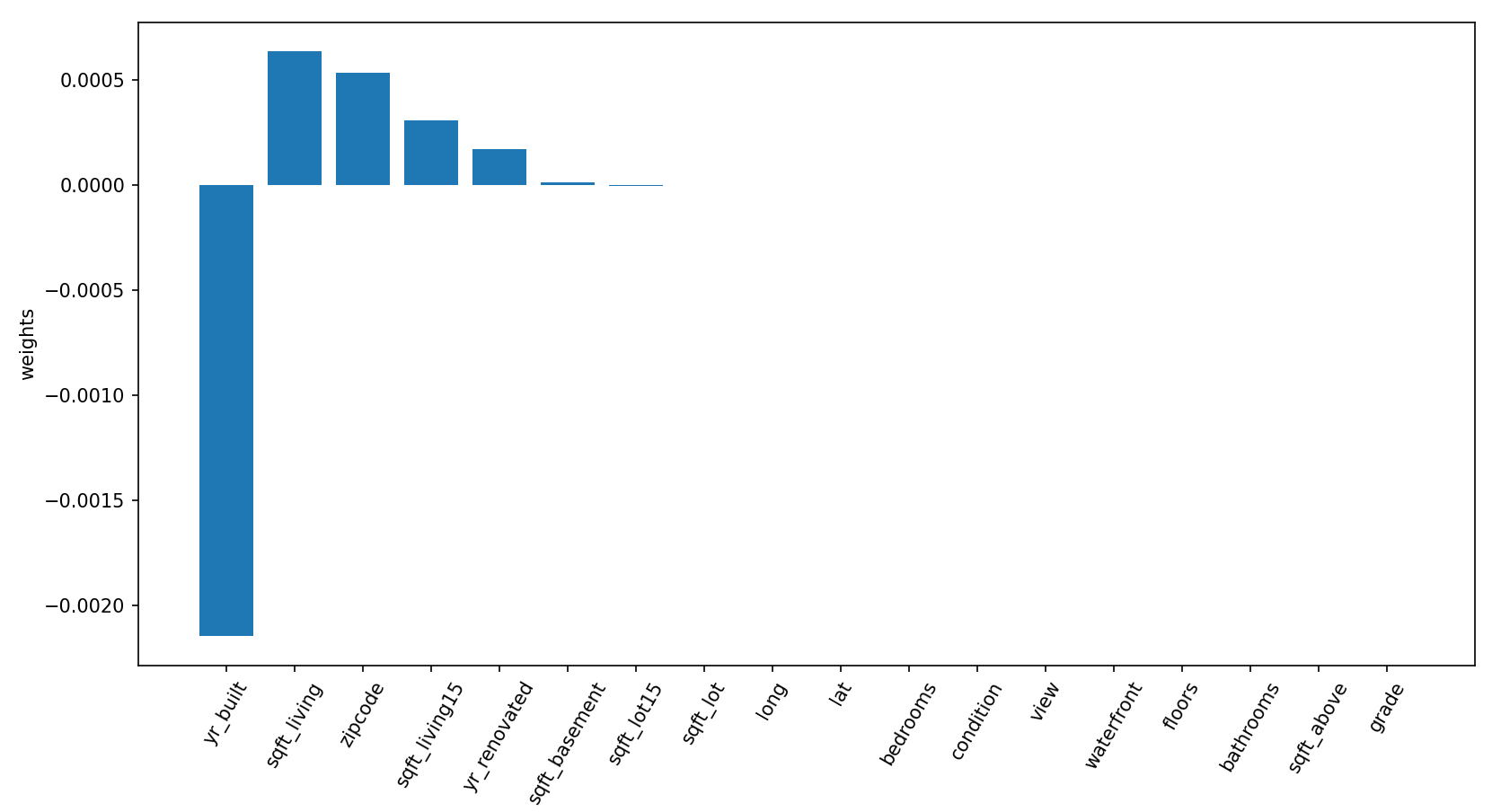
利用function 3-4，可得出下列結果





(3) 彈性網  
 使用function 3-5，可得下列結果





9. 相關模型結果比較分析

三個模型的結果如下，

|  |  |  |
| --- | --- | --- |
| 模型 | R^2 training | R^2 testing |
| 脊迴歸 | 0.685 | 0.707 |
| 套索迴歸 | 0.685 | 0.707 |
| 彈性網 | 0.509 | 0.540 |

脊迴歸跟套索迴歸的結果一樣，而且訓練集跟測試集的結果都沒有相差太多，代表沒有過度擬合的情形發生。