資四甲110916004 蔡奇傑

機器學習HW1\_Regression Analysis

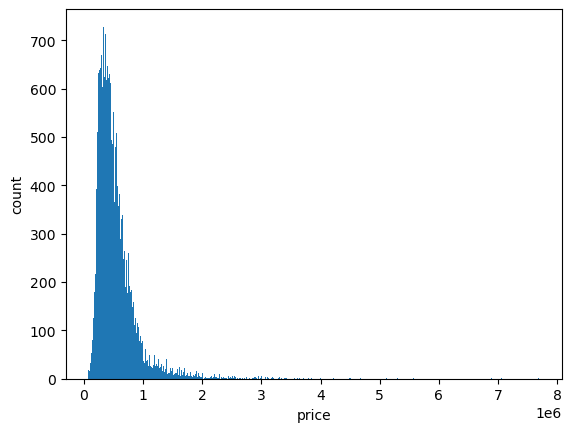
**Part 1:**

1. **資料清理與視覺化圖表**
   1. 清理遺漏值:

****處理數據時常遇到數據不完整或遺漏的情況，且遺漏值容易造成分析結果的偏誤並難以直接交由模型來擬合，因此我利用pandas載入後透過isnull().sum來了解此數據是否有出現數值遺漏的情形，結果如下:

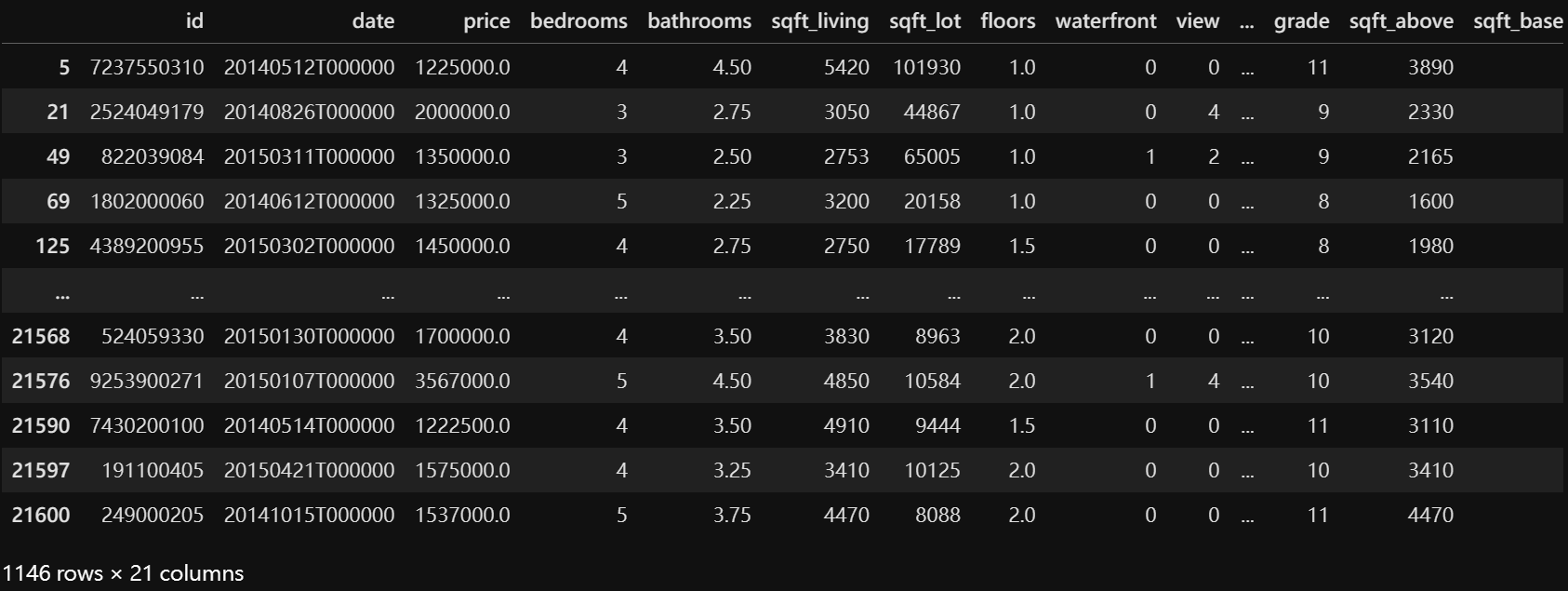
執行完畢後未發現任何遺漏值，因此無需作數據處理

* 1. 視覺化圖表:
     1. 利用長方圖顯示各價格區間的房子個數：

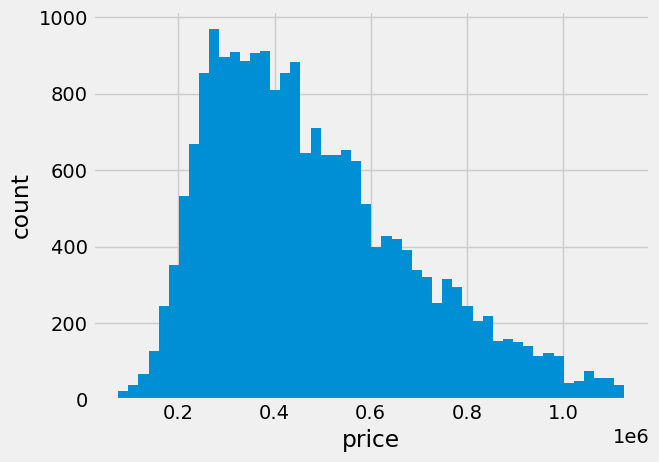


* + 1. 清除離群值：

利用IQR測試後發現1146個離群值

****

將其刪除後再列出其價格的長方圖:



由此可發現數遽變的更為集中且部分數值間的相關度也有些微的提高，皆有利於後續的模型建立。

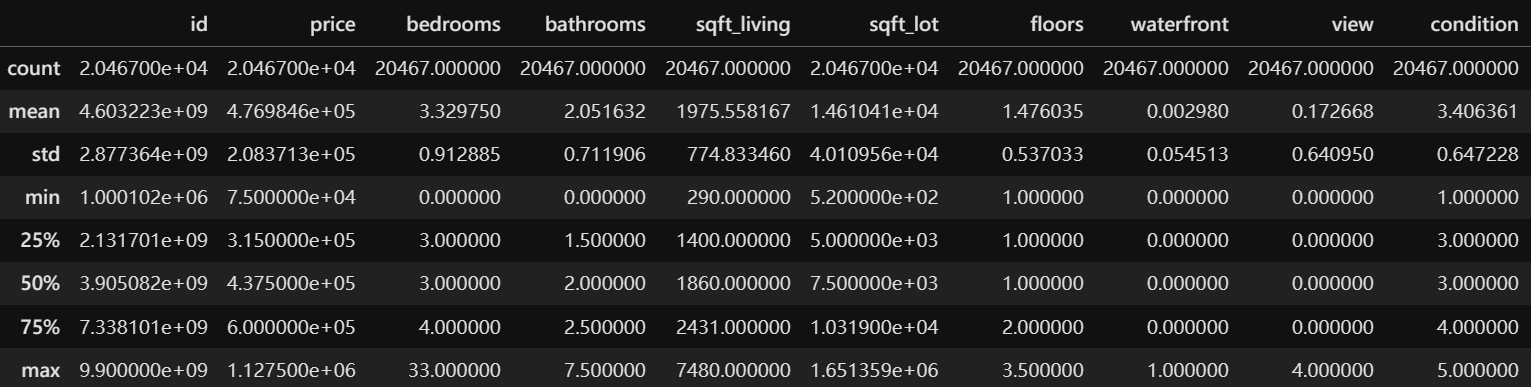
**2. 敘述性統計分析**

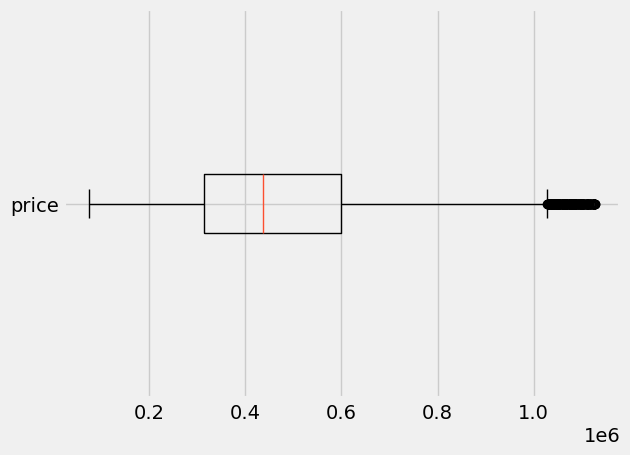
**** 透過敘述性統計分析，可以來了解樣本的屬性，樣本之敘述統計量包含平均數、中位數、標準差、變異數、最大值/最小值等等，以下為針對price特徵的計算結果:

房價(price)的總平均數為476984，中位數為437500，平均數較中位數高，表示數據分佈右偏(正偏)，即有正態分佈的長尾向右延伸，在房價的情境中，這代表有一部分高價房屋使得平均數被拉高，也就是因為少數的昂貴房屋使得整體房價的平均數偏高。

另外，中位數因為是處於中間位置的值且不受極端值的影響，所以它可以更好地代表數據的集中趨勢。而在這種情況下，中位數較低，顯示了大多數房價集中在較低的區域。

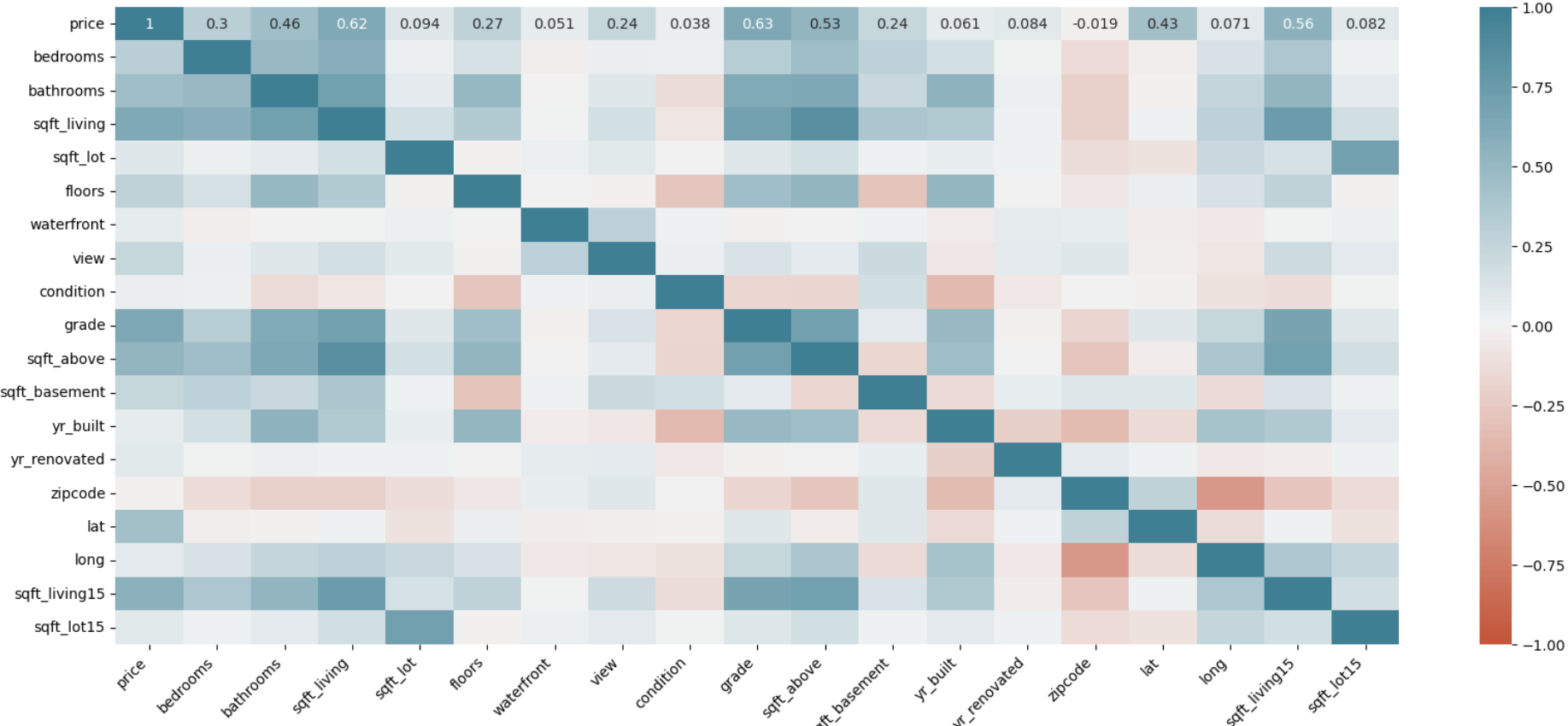
以下我也透過df.describe()來取得其他特徵的相關統計:

 ****

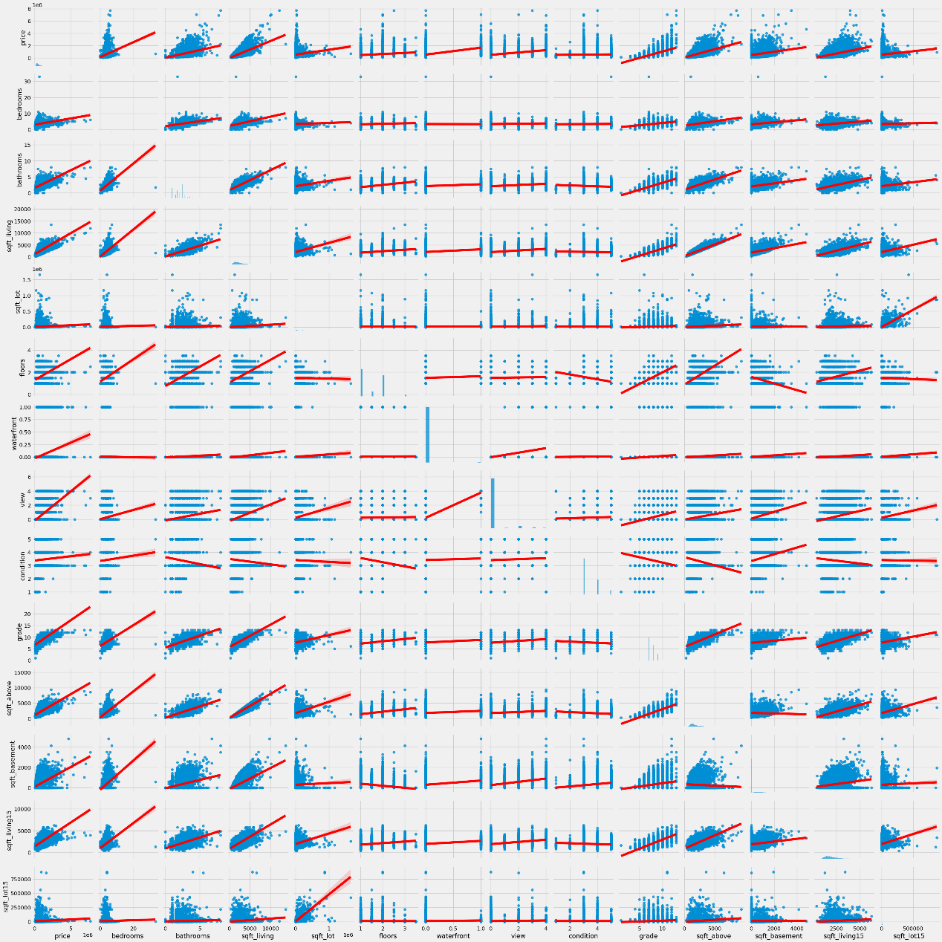
****透過price的四分位箱型圖發現price的箱子落在約0.3到0.6之間，表示房價的中間50%數據相對集中在這個範圍內。大多數房價分佈在這個區間內，不是偏向於極高或極低的價格，而圖的右側仍有一些的離群值。

**3.** **特徵相關性分析**

* 1. 相關係數熱度圖

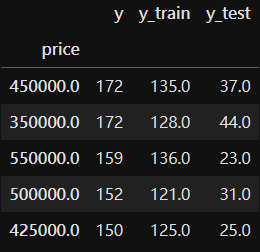
透過相關係數熱度圖可以觀察到，和price正相關係數最高的是grade(0.63)，其次是sqft\_living (0.62)(房子的面積)，也就是這兩者呈高度的正相關，但因為grade的數值並不連續(為離散、正整數的數值組成)，因此，我將使用相關係數次高的sqft\_living用來建立下述的簡單線性回歸線。

* 1. 散點圖矩陣分析:

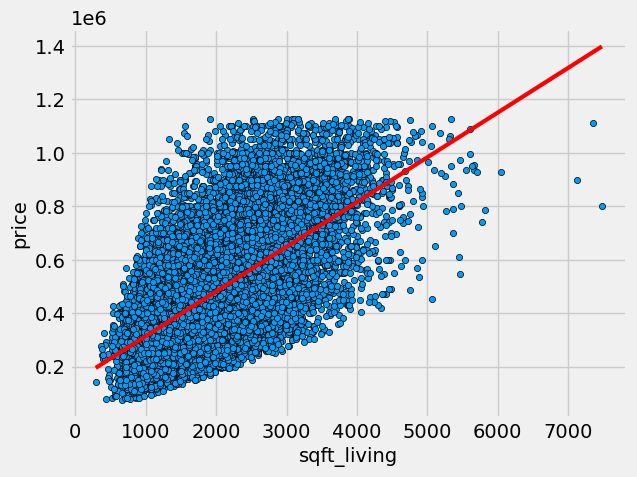
****在這裡使用sns.pairplot()函式來呈現特徵與特徵間兩兩的相互關係，我去除了id、date等有關編號的特徵，保留了可能對price有影響的13個特徵來比對，其中對角線上為該特徵的分布圖，然後紅線代表特徵和特徵間的回歸線。

**Part 2:**

1. **資料分割與建置迴歸模型 (線性 or 非線性)**
   1. 資料分割:

****首先將dataset中以80:20(訓練集:測試集)的比例切割，接者使用表格查看數據集(y)、訓練(y\_train)與測試集(y\_test)的數量。

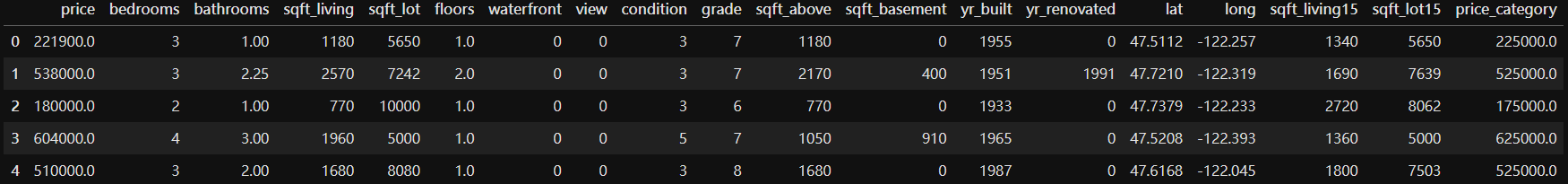
* 1. 建立簡單線性迴歸模型:

在這裡先建立只有一個特徵預測price特徵的簡單線性回歸，在特徵相關性分析中，我選用sqft\_living來當作自變量，結果如下:



此迴歸係數中，W1為斜率，W0為截距，在這邊我得到斜率為167.36，即當sqft\_living增加一個單位後，price會上漲167.36個單位。

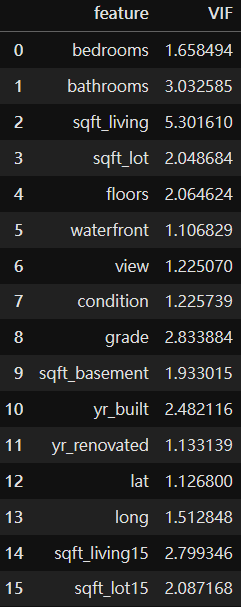
* 1. 建立多元線性迴歸模型:
     1. 建立多元迴歸模型

在建立多元線性回歸前，我將和price無關的特徵'id', 'date', 'zipcode'先從dataset中去除，再以scikit-learn庫中LinearRegression()建立多元線性迴歸。

* + 1. 查看VIF

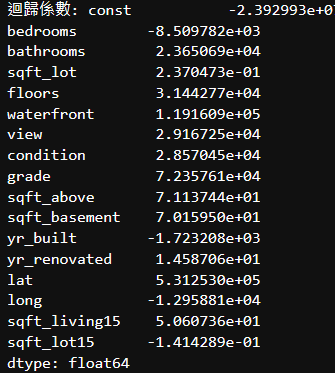
增加多個自變數後，為了避免「多元共線性」問題，讓某些自變數提高解釋力，進而影響模型造成偏差，我會用變異膨脹因子(VIF)來查看是否共線性較嚴重，課本提到statsmodels套件計算VIF的官方API建議大於5及有共線性存在，但大多數人認為其標準可以放寬到10，以下為執行結果:

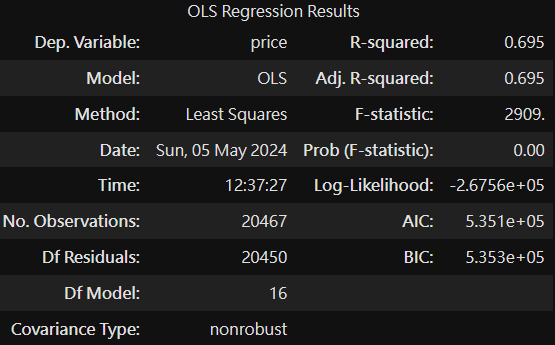
從上面可以觀察到sqft\_living、sqft\_above、sqft\_basement這三個特徵的VIF非常大，存在嚴重的共線性，因此我選擇將三個特徵依序從dataset中移除，在執行一次VIF後結果如下:



最後可以發現刪除sqft\_living後整體VIF表現較優，因此我決定刪除sqft\_living

* + 1. 修正多元迴歸模型

修正完後的迴歸模型的每個特徵迴歸係數如下，每個係數都不為零，表示這些自變數都有解釋力，並且使用statsmodels套件進行迴歸分析，以下為檢定結果:

由上表可以知道R-square為0.695，依照課本所述在合理範圍。另外"Prob (F-statistic)" 的值為0.00，這意味著在零假設下觀察到給定的F統計量（2909）或更極端情況的概率非常低。

1. **模型效能評估**

在這裡我利用了MSE(均方誤差)及決定係數(coefficient of determination)R^2來評估模型效能

* 1. 簡單線性回歸模型



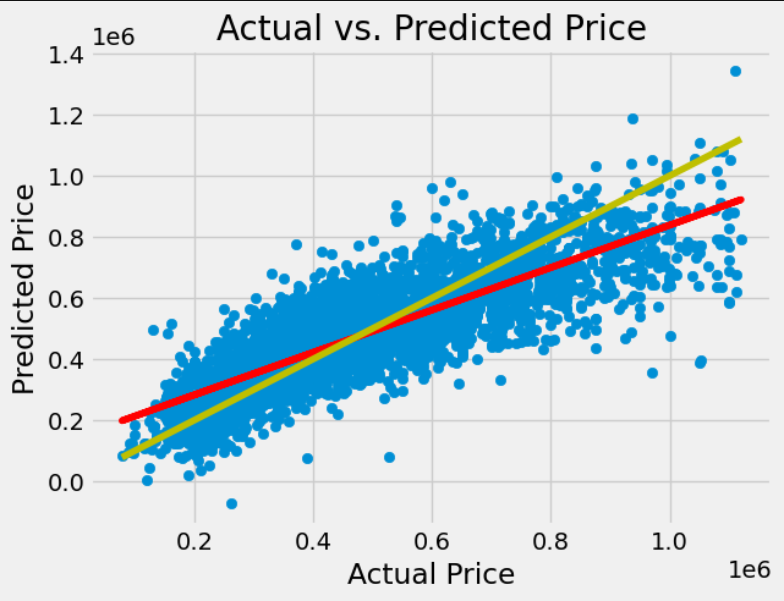
* 1. 多元線性回歸模型

在多元線性回歸中，因為MSE衡量了模型預測值與實際值之間的平均誤差的平方且可能數值較大，所以MSE較大。但和簡單線性迴歸模型相比，MSE明顯減少了許多，在R^2中也驗證了課本所提到的增加自變數，能夠大幅提升R^2值，另外模型在訓練集和測試集上的性能相對接近，沒有出現明顯的過擬合或欠擬合現象。

1. **預測結果分析**

以下針對多元線性迴歸模型做分析:

* 1. 實際值與預測值的散點圖

下圖中以x軸當作price實際值，y軸當作price預測值，來觀察兩者之間的關係。紅色線表示通過線性回歸模型擬合的回歸線，代表實際價格與模型預測價格之間的線性關係。

coefs[0](斜率): 0.695

由圖及斜率可發現預測值和實際值存在一定的誤差

* 1. 殘差分析
     1. 常態性檢定

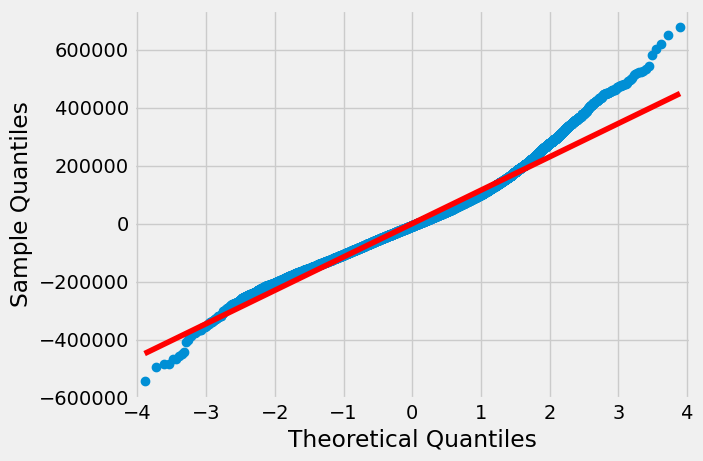
因為樣本數較多，我採用Kolmogorov-Smirnov常態性檢定來檢查殘差值是否呈現常態分佈:



根據輸出結果，統計量為0.532，p值為0.000 ，p小於顯著性水平，說明殘差不符合正態分佈。這意味著整體的線性回歸模型的誤差項在不同觀測值上可能不滿足獨立且具有相同方差的假設。

* + 1. QQ圖

使用 qqplot 函數來繪製殘差的 QQ 圖，用來檢查一個樣本是否符合常態分佈，其中藍色的點表示觀測值的分位數，紅色的直線表示理論分佈（在這裡是常態分佈）。

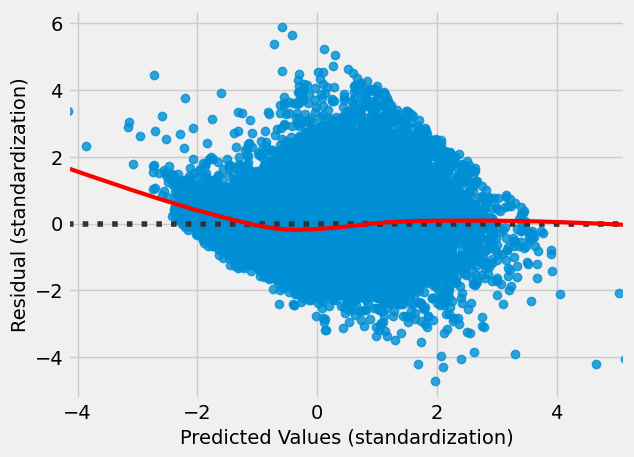
由圖可以發現，theoretical Quantiles中約 -3至1.5的範圍間其藍色的點和紅色線非常接近，代表這段區間這些觀測值的數值與常態分佈的期望值和變異數十分相符。然而超出-3至1.5的範圍時，可以看到明顯的偏差。這表示在這些區域內，觀測值的分佈與常態分佈有所偏離，可能存在一些異常值或系統性的偏差。這代表模型可能在處理極端值或在這些區域內的預測能力較差。

* + 1. 檢查獨立性

使用Durbin-Watson統計量DW來檢驗自相關性，檢視是否殘差間相互獨立，以避免在估計迴歸係數時降低檢定立，結果如下: 

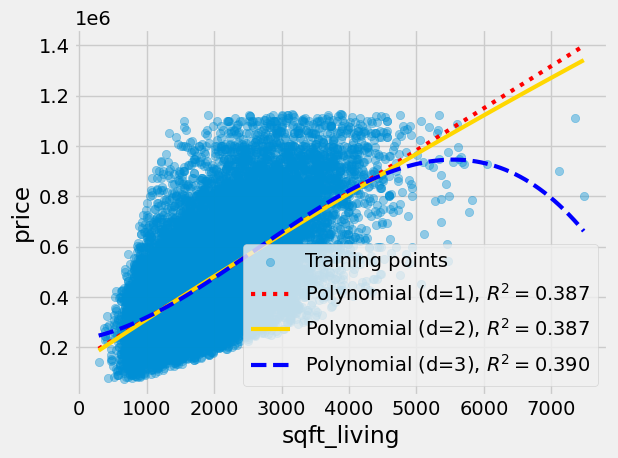
在課本中提到dw的合法範圍須在0-4之間，而此迴歸模型的dw為1.973，符合課本提到的接近2表示具有獨立性。

* + 1. 繪製殘差圖

這裡使用殘差圖，可以觀察模型的殘差與預測值之間的關係，紅色線為估計值，是利用局部加權散點平滑法對殘差繪製而成，即標準化後的預測值和殘差進行擬合。在結果中顯示X軸在約-1.5之後，紅色線十分接近0的水平線，代表了殘差變異數符合同質性，預測值為較精準，而在-1.5左側紅色線偏離了水平線，可能導致模型在低預測值範圍內較為不準。

**Part 3:**

1. **非線性迴歸模型 or 特徵變數變換**
2. 簡單非線性回歸模型:

以下對sqft\_living(自變數)和price(應變數)做degree為1到3的迴歸模型製作結果:

圖表右下角的R^2三個數值分別為degree 1到3迴歸模型的 Adj. R，在本次實做中，這三個並無明顯的變化，而在deg為3的時候可看出右側受到一點離群值影響，讓迴歸線轉彎的幅度比較大。

1. 多元非線性回歸模型:

|  |  |
| --- | --- |
| degree = 1 | degree = 2 |
| 迴歸模型效能評估: | 迴歸模型效能評估: |
|  |  |
| degree = 3 |
| 迴歸模型效能評估: |
|  |

在degree為2的非線性多元迴歸模型中，R^2和MSE相較線性的都有上升，另外，我使用測試數據X\_test[' sqft\_above'] 作為x軸，實際目標變量price作為散點圖上的點，並使用多項式多元迴歸模型的預測結果y\_pred\_poly作為迴歸曲線。

圖中的藍色點代表實際的價格值，而紅色的曲線則代表使用多項式多元迴歸模型對價格進行的預測。

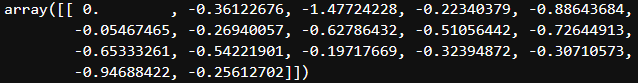
觀察兩者圖發現，deg=2的紅線涵蓋藍色的點相較deg=1,deg=3的多，代表deg=2的模型預測來的比deg=1理想，並且deg=3的預測值常出現與實際值無關的值，顯而易見的deg=3並無法有效地預測價格

1. **迴歸模型之正規化**

下面針對多元線性迴歸模型進行標準化的實作:

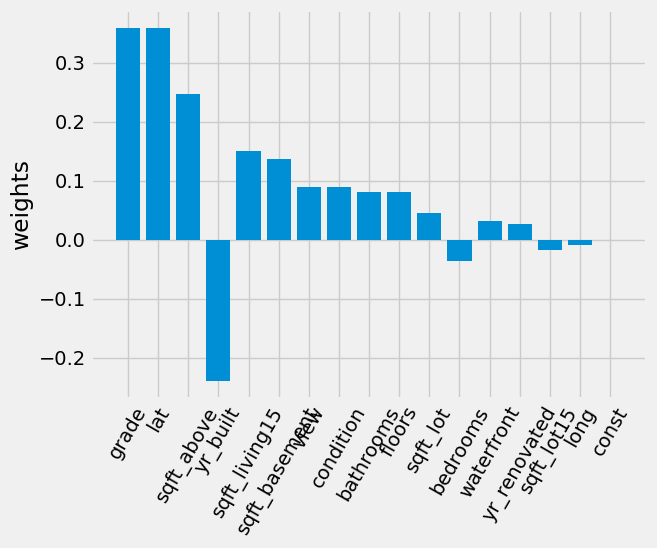
1. 資料正規化

這裡將使用fit\_transform方法對特徵矩陣(X)和目標變量(y)進行標準化處理，將特徵縮放到平均值為0，標準差為1的標準正態分佈。確保不同特徵之間的數值範圍相似，檢視以下輸出的正規化後特徵矩陣(X\_std)，來確保特徵值已經被正確縮放:



1. 接者使用Ridge回歸模型和交叉驗證(Cross-Validation)來選擇最佳的正則化參數alpha值:



1. 結果:

此長條圖提供了Ridge回歸模型中每個特徵變數的迴歸係數的大小和相對重要性，迴歸係數為正，表示該特徵與預測目標之間存在正向關係，即該特徵的增加將導致預測目標的增加。

1. 效能評估:



1. **相關模型結果比較分析**

以下我使用表格來呈現上述非線性回歸模型的MSE即R^2數值:

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | | MSE | R^2 |
| 簡單迴歸模型 | deg = 1 | 26601340694.466 | 0.387 |
| deg = 2 | 26597441116.492 | 0.387 |
| deg = 3 | 26466693655.679 | 0.390 |
| 多元迴歸模型 | deg = 1 | 13409899428.377 | 0.692 |
| deg = 2 | 11509532227.251 | 0.736 |
| deg = 3 | 47498847238.093 | -0.141 |
| Ridge回歸模型(多元線性迴歸模型之正規化) | | 13251671402.462 | 0.695 |

上述6個模型進行比較後，對於簡單迴歸而言degree數達到2或3可以得到更好的預測表現，然而多元迴歸模型表現相比簡單迴歸更能大幅提升效能，這代表以單一個特徵當影響因素不足以讓模型順利預測price值，然而在多元迴歸中，我嘗試使用deg = 3建置多元迴歸發現，MSE和R^2指標效能會大幅下降，因此也不能一昧地提高degree數，需要挑選剛好的degree才能符和預期。