**Data Mining HW2 Time Series Regression**

**0853412 吳宛儒**

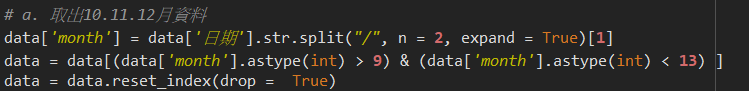
**讀取資料：**



**前處理：**

**1. 資料前處理**

a. 取出10.11.12月資料



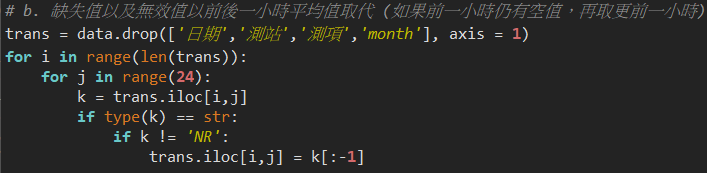
 b. 缺失值以及無效值以前後一小時平均值取代 (如果前一小時仍有空值，再取更前一小時)

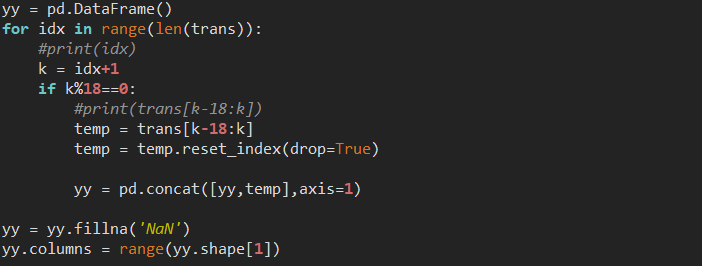
首先將奇異值的特殊符號去掉，並將data變為row數為18(features)的dataframe；

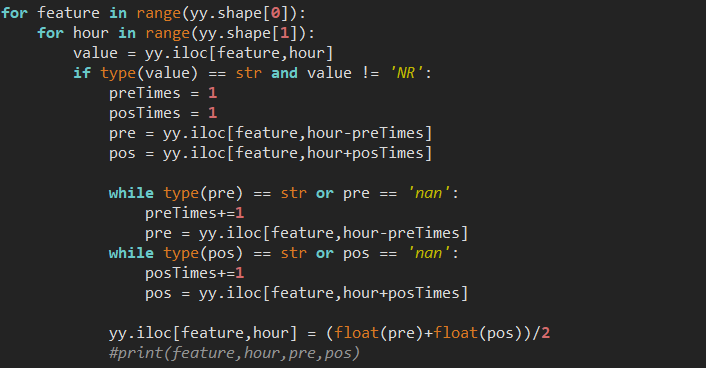
再透過判斷是否為string且不是為’NR’(沒有降雨的意思，不代表奇異值)；

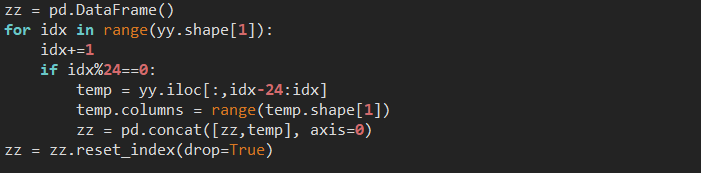
如果是奇異值，便將前後非奇異值的加減取平均作為替代值；

最後再將資料維度復原。





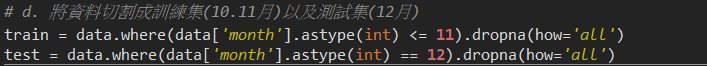




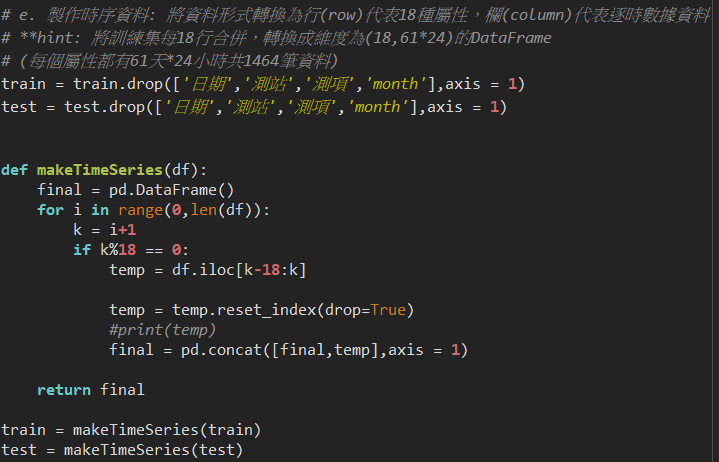
 c. NR表示無降雨，以0取代



 d. 將資料切割成訓練集(10.11月)以及測試集(12月)

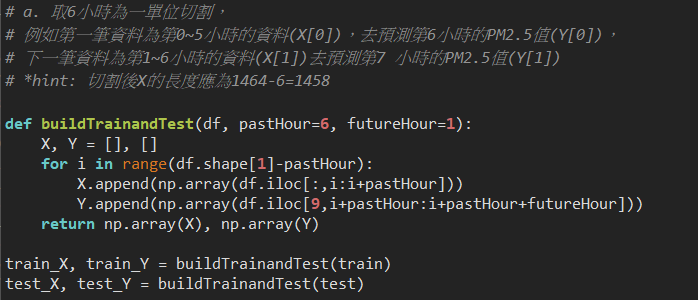


 e. 製作時序資料: 將資料形式轉換為行(row)代表18種屬性，欄(column)代表逐時數據資料



**2. 時間序列**

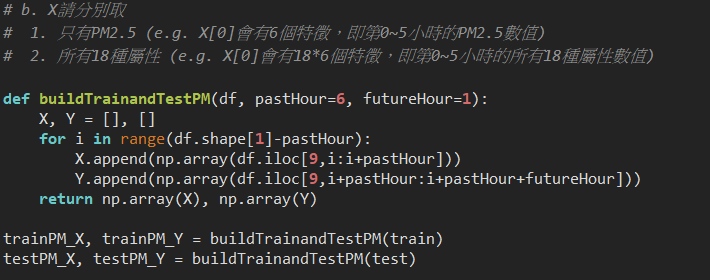
 a. 取6小時為一單位切割，例如第一筆資料為第0~5小時的資料(X[0])，去預測第6小時的PM2.5值(Y[0])，下一筆資料為第1~6小時的資料(X[1])去預測第7 小時的PM2.5值(Y[1])  \*hint: 切割後X的長度應為1464-6=1458



 b. X請分別取

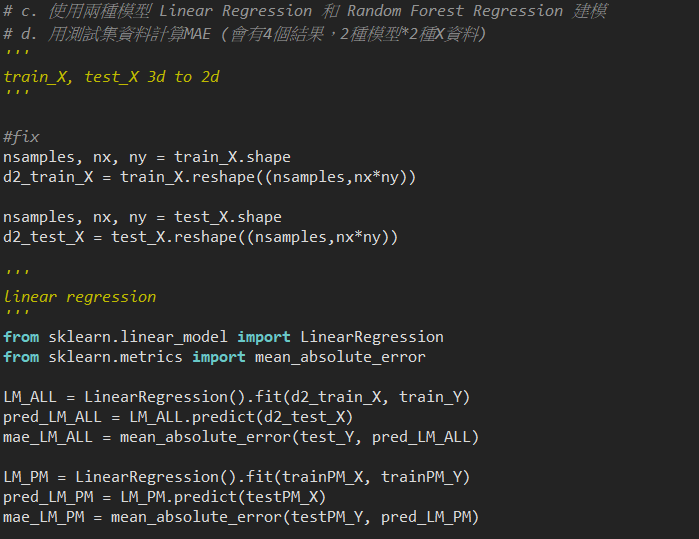
     1. 只有PM2.5 (e.g. X[0]會有6個特徵，即第0~5小時的PM2.5數值)

     2. 所有18種屬性 (e.g. X[0]會有18\*6個特徵，即第0~5小時的所有18種屬性數值)

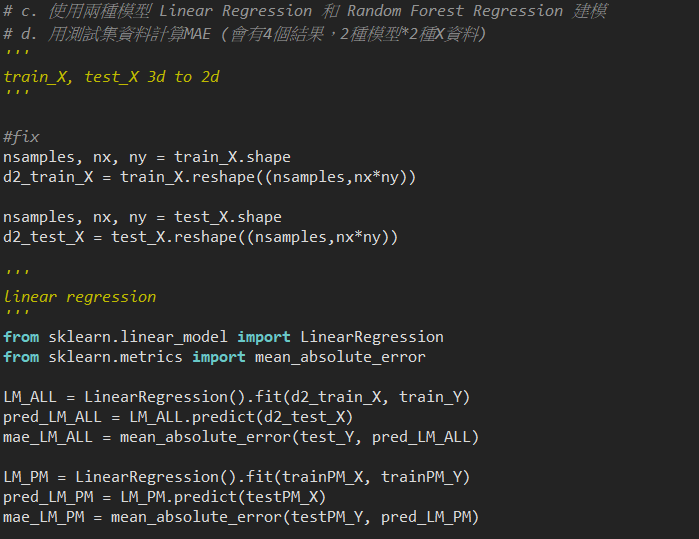


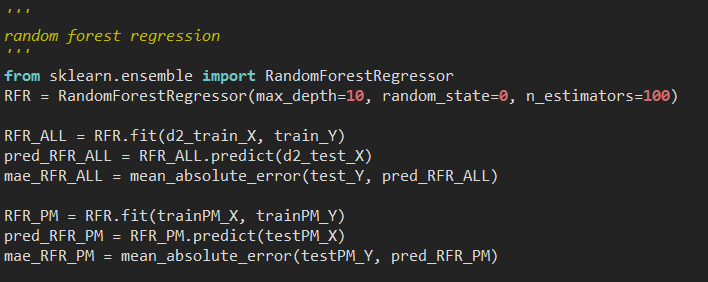
 c. 使用兩種模型 Linear Regression 和 Random Forest Regression 建模

 d. 用測試集資料計算MAE (會有4個結果，2種模型\*2種X資料)

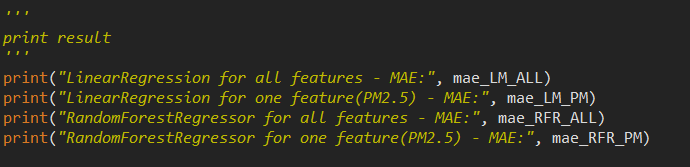


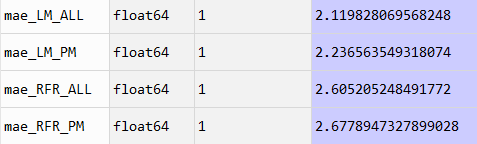
因為all features的train data是三維，因此在放進model之前應該做reshape處理成二維。





**印出結果：**





**LinearRegression for all features - MAE:** 2.119828069568248

**LinearRegression for one feature(PM2.5) - MAE:** 2.236563549318074

**RandomForestRegressor for all features - MAE:** 2.605205248491772

**RandomForestRegressor for one feature(PM2.5) - MAE:** 2.6778947327899028

**討論：**

從結果可以看出來Linear Regression with all features跑出來的error是最小的，這可能與Random Forest的max\_depth有關(不夠深、參數調整)。而將兩種model分開來看的話，可以看到將所有features都考量進去普遍會比只考量一個feature(PM2.5)表現還要好，也符合個人的預期：較多特徵能比單一特徵預測得更好。