一、資料分類

在讀入 input.csv 後，首先要先將資料進行分類。我一共將資料分成三大類，分別是要預測的 testing_data(2021.10.15~2021.11.11 的資料)、用來驗證 model 好壞的 validation_data(2021.9.13~2021.10.14 的資料)以及訓練模型用的 training_data(2021.9.12 以前的資料)。

在 training_data 的部分我又將其分為 60days, 120days 及 168days，分別代表用來訓練的資料是 2021.9.12 以前共 60,120,168 天的數值。會這樣分類是因為長期與短期資料訓練出來的結果可能不盡相同，如何進行選擇也會在後面提到。

二、Regression

我使用的是 OLS 的方式實作 Regression。參考老師的投影片可以知道我們需要在得到 X 後預測 Y （如下圖）。

$$\text{predict with: } \hat{y}^i = \sum_j^n w_j \phi_j(\mathbf{x}^i)$$

為此，我們需要得到 W 的數值。同樣參考老師投影片，我們可以得到下列算式，透過矩陣的轉換及運算獲得 W 的數值。

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

在實作 code 的部分，我先將要使用的 X,Y 以 Array 而非 List 的方式儲存，方便之後做矩陣的運作。

接著先將 X 的部分進行處理，得到一個 n*2 的矩陣，然後同上方算式的運算方式，透過 numpy 函式庫裡的 function，進行矩陣乘法、逆矩陣、轉置矩陣。(詳細實作可見下圖)

```
def Regression(train):
# TODO 4: Implement regression
    x = np.array(PreprocessData(train)[0])
    y = np.array(PreprocessData(train)[1])


    x_vec = np.concatenate( (np.ones( (x.shape[0], 1) ), x[:,np.newaxis]), axis=1)
    y_vec = y[:,np.newaxis]


    w_vec = np.matmul( np.matmul( np.linalg.inv( np.matmul(x_vec.T,x_vec) ),x_vec.T ),y_vec )


    return w_vec
```

三、選擇模型

前述提使用三種不同時間長短的資料進行訓練，在完成 Regression 後，會使用 validation_data 來判斷預測模型的好壞。透過 CountLoss()，可以得到 validation_data 的預測結果與實際值的差別。我這次使用 MSE 與 MAPE 的方式來計算 loss，然後比較三種模型的 loss 大小，數值最小的將會成為最後預測用的模型。(實作如下圖)

```
def CountLoss(model):
# TODO 5: Count loss of training and validation data
    #MSE
    #validation_data.dtype = 'float64'
    data = []
    estimate = []
    theory = []

    data = PreprocessData(validation_data)
    theory = data[1]

    beta = Regression(model)
    estimate = beta[0]+beta[1]*data[0]
    loss = (1/(2*estimate.size)) * np.sum((estimate-theory)**2)

    MAPE = 0
    for i in range(0,validation_data.shape[0]):
        MAPE += (1/estimate.size) * abs(((estimate[i]-theory[i])/theory[i]))
```

另外，根據這次訓練結果，60days 的模型訓練可能是較佳的模型，以下附上三種模型的 loss 值

day60	float64	1	0.01783123146694892
day120	float64	1	0.01809965954408564
day168	float64	1	0.022207361729849558

(MAPE 值)

day60	float64	1	134.0782813706305
day120	float64	1	146.28348994235742
day168	float64	1	212.41708921015032

(MSE 值)

以上就是這次實作的主要部分，在選定模型後只要將 testing_data 放入模型中就可以得到預測值，並將其寫入 output_CSV 中即完成本次作業。