2.Introduction

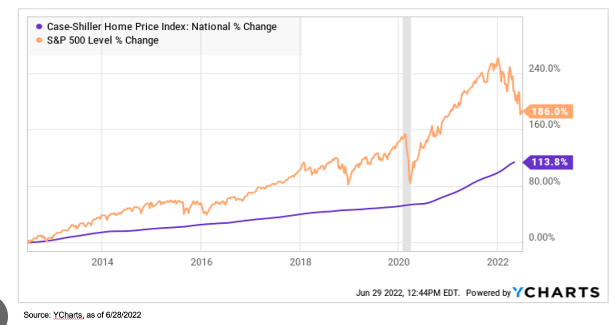
(1)動機:

根據近年來不動產業者對於房價走勢以及股市漲跌之間的研究發現，兩者的相關係數從十年前的62%上升到近幾年的98%創下歷史新高，對於以往股市不太影響房價的觀念可能需要重新檢視。一直以來房價的波動被視為股票的落後指標，舉例來說前陣子的台股指數飆破萬八時，隨後許多蛋白區的房價都飆升至鄰近原蛋黃區房價的價格，但此種現象在目前股市緊縮的情況下也造成房價開始持平的狀態。因此，此project的研究動機主要是依據美國10州歷史的房價交易紀錄，來預測未來房價價格，下圖為近20年房價指數以及台股加權指數比較。



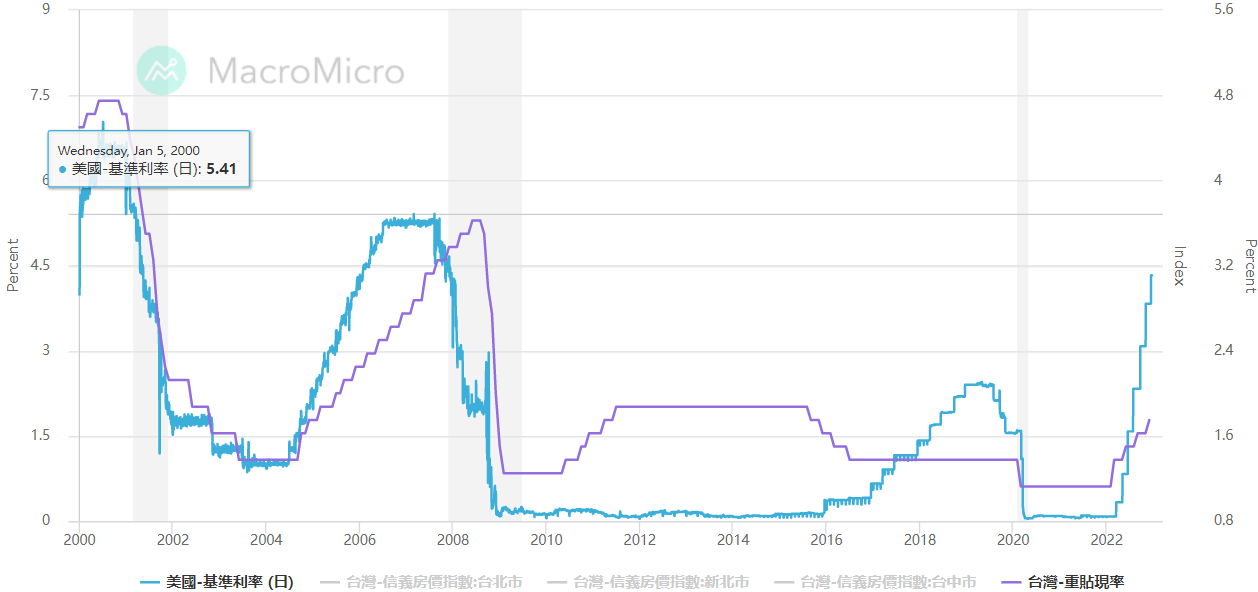
(2)背景:

由於台灣法律較嚴格保護個人的房地場交易細目，資料取得不易，所以資料的蒐集以及處理上只能藉由美國的資料來訓練模型，但同時我們也確認美國的房價以及股市走向跟台灣一樣有很大的相關性，藉此利用他國的預測結果來踹測癌灣的房價走勢，如下圖所示為美國S&P 指數與美國房價指數的比較。端看2014~2021的股市以及房價走向，可以說相似度與上塗有非常大的相似之處。



(3)目標

此project希望能藉由美國的歷史房價交易紀錄，以及多種不同的機器學習方法訓練出模型(linear refression, XGBoost, Random Foreast……)，預測未來房價的結果，藉此來當作台灣未來法律同意釋出歷史房價資料後，能夠利用類似的方法處理台灣房價預測的模型，如下圖可以清楚觀察到台灣央行的貨幣政策(rediscount rate)幾乎與美國有相同的走向，這讓一般投資房地產要向銀行貸款時，台灣與美國會有相同的投資策略思維，因此我們判斷，對於機器學習的模型，利用相同的訓練模式在未來也可以應用在台灣。



4.Method

Flowchart:

**Raw Data of historical houseprice in US**

**Feature Engineering**

1. **filter data from 2010/01/01 ~ 2022/06/30**
2. **calculate the correlation matrix between each colums, drop the colums whose correlation with price is less than 0.1.**
3. **deal with the nan value in column “bed” & “bath”, using method of statistic of similar type of house.**
4. **survey the houseprice distribution of each states in data, and scored it from 1 ~ 4, 1 for the highest and 4 for the lowest price in average。**
5. **label the time interval(2 months) for each data from 2010/01/01(label = 0) to 2022/12/3(label = 78)**

**Split data(after feature engineering)**

1. **train data(2010/01/01 ~ 2022/06/30)**
2. **test data(2022/06/30 ~ 2022/12/31)**
3. **Kfold = 10 (train : validation = 9 : 1)**

**models**

**………….**

**XG Boosting**

**Linear regression**

**Random Foreast**

**Comparison of Prediction result**

1. **Do experiment of the different data(drop colums)to test the training result**
2. **Compare the result of each model, and discuss the difference between them.**
3. **Choose the best model to predict the test data.**