心得與感想

這個研究讓我證明自己在面對無法應用常用的方法解決問題時，有能力自行構想出新的解決方案，並持之以恆付諸實現。在這個與課業毫無關聯的領域，持續發現新問題所帶來的驚喜，以及解決問題所獲得的成就感，是自己在課業壓力下還能擠出時間完成研究的主要原因。

從去年四月決定參加旺宏科學獎開始，五月趕工交件又落榜後，能支撐著自己在繁重的科學班課業壓力下繼續持之以恆的主要原因，就是在這個與課業毫無關聯的領域，持續發現新問題所帶來的驚喜，以及解決問題所獲得的成就感。

雖然最終結果不盡理想，但是在這一個築夢、努力的過程中，學習、看見平常求學路上看不見的風景與知識。在這次專題研究，讓我學習處理機器學習問題的方法與技巧，讓我能在之後的學習中更加有競爭力。

# 第二十屆旺宏科學獎

# 創意說明書

參賽編號： SA20-030
研究題目：線上訂房平台調價策略之研究與預測
姓名：鄭恆安
關鍵字：爬蟲、大數據分析、機器學習

# 壹、研究動機

小時候一直都期待暑假的到來，與家人一起出遊。記得有一年的四、五月份，知道父親開始規劃暑假旅遊，便自告奮勇的幫忙上網找尋飯店，看著不同飯店每天不規律跳動的房價，根本不知道該如何選擇。即使等到父親送出訂單後，也想知道我們是用優惠的價格訂到飯店，還是當了冤大頭？自此之後，每當出遊入住飯店時，會想知道這些飯店有沒有標準的調價規則，但總是沒有辦法靠自己看出一些端倪。在嘗試幾次之後，我領悟出一個道理，雖然我知道商品在快到期前或是乏人問津時，往往有破盤價的出現。反之，如果需求孔急，當庫存不足只剩下最後商品時，物以稀爲貴，自然造成價格上揚，但是因爲我並不知道飯店房間的庫存，也只知道自己的需求，所以是不可能靠自己觀察來了解飯店的調價規則的。

高中加入網管社，跟學長學習維護校園網路與學校網頁，常聽學長提到爬蟲程式、機器學習等酷炫技術。偶然機會看到電視上Trivago的廣告詞，它提到同一個房間在不同的網站上的價格可能不一樣，突然小時候對飯店調價規則的疑問又浮出腦海，除了訂房的時間外，還有哪些原因會導致房價改變呢？

上網找了幾篇論文後發現，國內的研究主要是餐飲與旅館科系學生以訪談飯店業者的方式爲主，但看不出飯店實際的價格變化，更不知道飯店之間調價策略是否有關聯。爲解決心中的疑惑，因此決定趁這個機會學習爬蟲程式，直接從線上訂房平台(Online Travel Agent, OTA)所提供的大量數據來研究各飯店的調價策略，希望也能用迴歸分析或是機器學習等工具找出不同飯店在不同狀況下的調價策略。

研究結果可以提供飯店業者動態房價調整建議，讓業者根據競爭對手的調價策略動態調價，提升飯店的訂房量與營收，或許還可以作爲消費者訂房前的參考依據。

# 貳、研究目的

本研究的目的有二點：

一、設計程式自動爬取線上訂房平台的房價，找出影響飯店調價策略之變因。

二、利用統計迴歸與機器學習建構房價調價模型，根據歷史資料進行單日房價預測，比較各方法預測結果的準確度與差異。

# 參、研究方法

本研究預定借助統計迴歸與機器學習技術，利用線上訂房網站的資料來建構調價模型，在此先介紹相關背景知識與預定採用的方法。在背景知識部分，除了簡單介紹國內現有以訪談飯店業者的參考文獻之外，我也嘗試從線上訂房產業服務供應商的角度切入，從他們開發的飯店管理(Hotel Manager)工具，深入了解飯店業者的需求，並說明本研究可能的實際應用。另簡單介紹統計迴歸與機器學習相關方法，說明各方法的優缺點與應用時機。在預定採用的方法與步驟部分，將依序介紹本研究的「蒐集房價資料流程」、「資料視覺化方法」，以及「建構統計迴歸與機器學習模型的方法」，詳細說明本研究採取之方法與步驟。

## 一、背景知識

在開始進行研究之前，我先研讀網路上線上訂房平台的文章 [1,2]，從中了解到目前觀光旅遊等消費型態已經大幅改變，飯店業者爲了增加房間曝光率以吸引來自各國的自助旅行者，紛紛將房間上架至各國際大型線上訂房平台進行販售，希望透過訂房平台版面曝光，接觸更多的客源 [3]。也從經驗與數據兩個面向來探討飯店定價策略的文章中 [4]，看到兩家飯店的經營者，各自從不同的角度切入，做出完全相反定價策略的不同觀點。在閱讀了這些文章之後，我進一步從參考資料 [5]與 [6]中，深刻了解線上旅遊產業鏈以及主要的供應者，我才發現原來線上旅遊產業很大，競爭激烈，而我熟知的訂房網站不過是線上旅遊產業鏈的一小部分。深入瞭解後發現，爲了讓飯店業者在全球人氣訂房平台上同時銷售他們的房間，控管各個訂房平台房間上架的價格與數量，還有各種提供管理訂房平台通路工具的軟體服務業者，分工細密。在許多文章中不約而同的都提到成立於2006年的SiteMinder公司，研究此公司我發現，它除了開發通路管理(Channel Manager)、飯店訂房引擎(Hotel Booking Engine)、飯店網頁設計、飯店付款處理、飯店商業洞察力(Hotel Business Insight)、飯店多元搜索(Hotel Metasearch)等各式各

樣幫助飯店業者提升營運效率的各種軟體工具，還提供許多免費的數據統計報告供人下載。由於SiteMinder是這個產業的龍頭，與目前國際知名的飯店以及線上訂房平台都有合作關係，因此擁有全球飯店售價與銷量的完整第一手資料。圖1是從SiteMinder網站 [7]中所找到的全球與台灣飯店訂房量與2019年同期比率走勢圖，根據這個統計資料，因爲COVID-19的緣故，在2020年的十至十二月間，全球飯店的預訂量只有2019年同期預訂量的41%，台灣飯店的預訂量也下降爲2019年同期預訂量的75%。根據這些數據可看出，嚴峻的市場造成飯店間競爭的白熱化，如何在銷售率與利潤之間做出取捨，隨時根據市場環境與競爭對手策略調整房價，對飯店業者來說會是決定生死存亡的關鍵。從SiteMinder的例子中，我發現提供管理工具給飯店的軟體業者，因爲手邊擁有大量的訂房數據，反而比實際經營的飯店業者更能精確掌握旅遊住房市場的調價策略與趨勢，因此更確定我直接從線上訂房平台取得數據進行調價策略研究，是一個符合未來趨勢的方向。

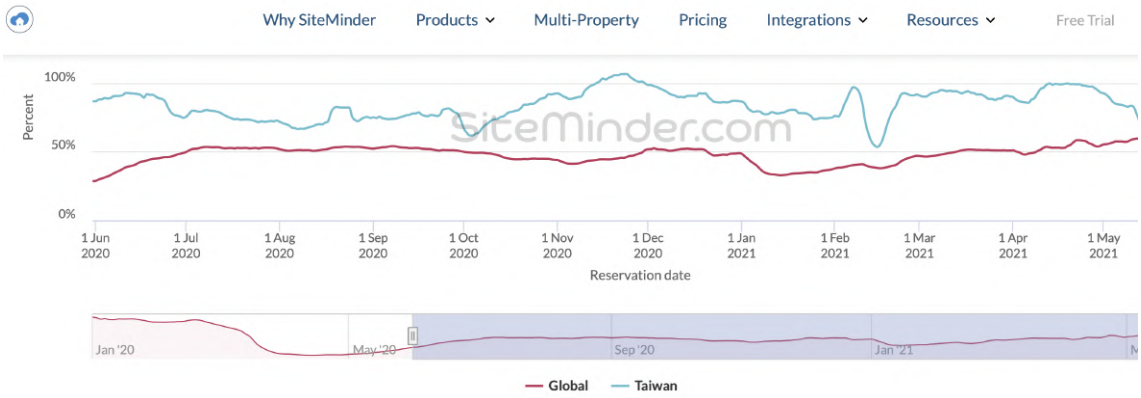



圖 1: 全球與台灣飯店訂房量與2019年同期比率走勢圖

在爬取線上訂房平台房價之前，需要決定要找幾個訂房平台？以及需要蒐集那些資料？在選擇訂房平台方面，我從一篇研究法國價格平價協議法案(Price Parity Agreements)對飯店與民宿房價前後影響的論文中讀到 [8]，在做房價分析時只要任意找一個訂房平台來蒐集資料就已具備足夠的代表性，這是基於作者蒐集三年巴黎地區不同線上訂房平台的房價發現，不同平台上的房價最多僅有4%的差異。本研究假設這個結論也可適用於台灣，所以我選擇最熟悉的Hotels.com進行資料蒐集。在資料項目的選擇上，國內外的研究都差不多，大多會考慮飯店的星級數、地點、客群等因素。因此，我選擇台北市飯店的一級戰區–交通要衝(台北車站)與商業辦公區域(101大樓)周遭一公里內的3、4、5星級飯店爲研究對象，除了有地點的差異，亦有星級數不同變因。蒐集資料的前置作業流程可參考圖2。

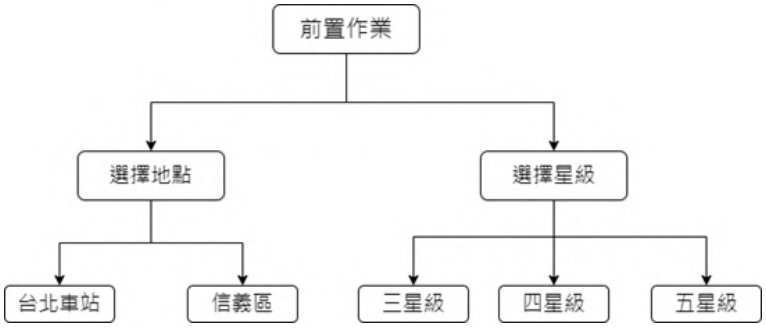



圖 2: 蒐集資料的前置作業流程

## 二、模型介紹

因爲機器學習中常會使用線性的方法作爲最基礎的預測方法，因此我先採用Linear Regression、Ridge Regression。再來，因爲tree base的演算法在機器學習中是很常被用到的非線性模型，其中Random Forest是其中較不容易overfitting與效果最好的決策樹模型。XGBoost則是近年來機器學習比賽的常勝軍。而LSTM是近幾年人工智慧不斷發展的重要模型之一，其彈性也很大，而且LSTM是設計給時間序列問題的模型，因爲房價調整問題有時間序列的特性，本研究未來也會嘗試LSTM模型，但目前尚未使用LSTM進行訓練。

以下簡單介紹本研究使用的統計迴歸與機器學習方法:

- **1. Linear Regression**： 線性迴歸，是最基本的統計迴歸模型，利用最小平方法找出對一個或多個變數最佳的線性迴歸方程式。
- **2. Ridge Regression**： Ridge Regression是Linear Regression的改良統計迴歸方法 。Lasso可以同時進行變數篩選與複雜度調整(正規化) 。它是在最小平方和中對模型加上正規化(regularizer)或是稱作懲罰項 (penalty term)，以減少模型的複雜性和防止過度擬合的簡單線性迴歸模型。
- **3. Random Forest :** 隨機森林(Random Forest)是裝袋演算法(Bagging)加上決策樹 (Decision Tree)所組成的演算法。Bagging會從訓練資料中隨機抽取(取出後放回，n<N)樣本訓練多個分類器，再用多數決(Majority vote)得到最終結果。每個決策樹都是由隨機分配的訓練樣本創建的，再透過取平均的方式加強準確率與減少overfitting的問題。

- **4. XGBoost :** XGBoost是一種加強版本的gradient boosting機器學習演算 法，與Random Forest同爲多個迴歸樹組合而成的。但與Random Forest不同的是，XGBoost是有序的產生迴歸樹，根據前面的樹，來調整下一棵樹。然而，因爲它預先將數據進行排序，使得它的計算量大幅減少、運行速度也較其他boosting方法更快、更有效率。

- **5. Long Short-Term Memory (LSTM) :** 一般的機器學習模型不適合處理時間序列的問題，RNN模型透過隱藏層將上一個輸入存起來，並在下一次輸入時同時考慮上一次的值，可以描述時間序列的關聯性。LSTM是進階的RNN模型，透過增加forget gate，決定是否忘記上次的結果，可以同時考慮很長時間的值並進行計算。

## 三、預定採用的方法與步驟

本研究使用的設備是筆記型電腦，軟體工具包括Jupyter Lab、 Python，Tensorflow等。相關知識除了教授、老師與學長的指導，大部份是從網路上自己學習摸索而來。以下介紹基本假設與名詞定義：

- 基本假設：在訂房平台上查詢的房價是飯店業者在前一日上架就不調整的實際售價。

- 短天期與長天期預測： 在本研究中，短天期預測的定義爲隔日的房價預測，而長天期預測就是針對兩天以上的預測。

- 歷史資料天數**n**： 是本研究房價預測模型所需要的輸入日數長度。以參考資料 [2]爲例，它的工具利用過去90天的數據進行評估、預測($n = 90$)。

- 歸一化：將訓練資料根據資料的最大最小值，將所有數值縮放到0和1之間。

- 變因：影響房價變動的原因。

- 搜尋項目：針對變因所制定的蒐集資料所需參數。

- 特徵：針對變因所細分過的調價模型的欄位。

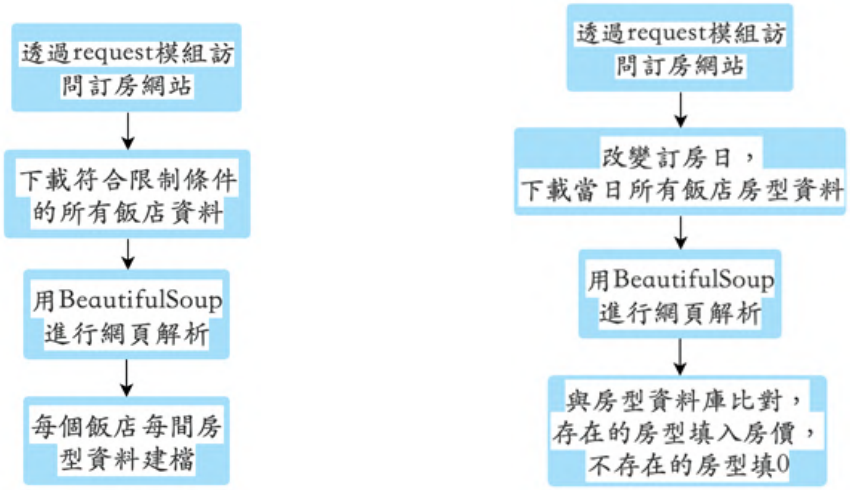
- 訓練集/測試集：用來建立訓練/測試資料的房價資料。

- 間隔天數：訂房日與入住日的天數差。

本研究的預定採用的方法可分成「蒐集房價資料流程」、「資料視覺化」，以及「建構統計迴歸與機器學習模型」三大步驟。蒐集房價資料流程可分成建立房型資料庫與爬取房價兩部分，流程如圖3(a)與圖 3(b)。前者是根據研究參考文獻所獲得的初步結論，先建立房價資料庫所需要的搜尋項目，後者則是利用爬蟲程式每天模擬消費者上網訂房行爲，並將飯店、房型、訂房日、入住日、入住日房價等搜尋項目的數值，逐一紀錄在房價資料庫中。蒐集房價資料流程的重點是設計能自動爬取線上訂房平台房價的程式。我的做法是利用Python模組訪問訂房平台，先下載符合限制條件(台北車站或101大樓半徑1公里內)的所有飯店資料，進行網頁解析，再把各飯店資料依房型建檔。需注意的是，每間飯店會有多種房型，相同房型還會因爲服務條件不同而有不同的房價，因此都需要分別紀錄建檔。在建立房型資料庫的過程中，爲了避免因爲某些房型已被預訂而查無資料或是飯店房型尚未上架，所以本研究選擇以初次下載資料日(2020年7月1日)起算4個月後的日期(2020年11月1日)所搜尋到每間飯店的房型爲基準(註：假設此時上架的房型爲每間飯店所有種類的房型)，逐一爲每一間搜尋到的飯店與所有房型，逐一於資料庫建檔。接下來再讓爬蟲程式每天訪問訂房平台爬取房價，在此，本研究把爬取房價當天日期當作搜尋日，再以當天與未來每一天輪流作爲入住日，逐日蒐集每間飯店每間房型的房價，將所爬取的資料與房型資料庫進行比對，並填入對應房型的房價；針對不存在的房型(已無空房或暫時下架)，則將房價設爲0。接著，再重複以上步驟，逐日爬取房價(註：本研究自2020年7月1日開始每天爬取以2020/7/1至2021/7/1爲入住日每天的房價持續至今)。

在資料視覺化部分，主要負責將蒐集的房價資料進行分類、繪圖，再從圖中進一步篩選影響房價變動的關鍵變因，並設爲調價模型的特徵，詳細的分析結果紀錄於「研究過程」中。透過資料的視覺化分析結果，我發現房價與許多變因有關，如:飯店位置、飯店星級、淡旺季、住房率、疫情、政策等等，這也導致房價不規律的變動，不容易用判斷式的敘述或特定函式寫出程式來預測。爲解決這個問題，我決定採用統計迴歸與機器學習演算法來建立模型，以預測飯店房價。

在建構統計迴歸與機器學習模型部分，我的設計概念是用矩陣**X** 與 **Y**分別代表模型的輸入與輸出，其中輸入是一個$M \times (4n+2)$或$(4n+3)$的矩陣(後者多增加




(a) 建立房型資料庫　　(b) 爬取房價

圖 3: 蒐集房價資料流程

了一個房型資訊納入特徵)，$M$是總共的訓練資料筆數，每筆訓練資料包含n天的歷史房價。矩陣$\mathbf{X}$ 的公式爲

$$\mathbf{X} = \begin{bmatrix} \vec{\mathbf{x}}_{1,1} & \dots & \vec{\mathbf{x}}_{1,j} & \dots & \vec{\mathbf{x}}_{1,n} & \vec{\mathbf{w}}_1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \vec{\mathbf{x}}_{M,1} & \dots & \vec{\mathbf{x}}_{M,j} & \dots & \vec{\mathbf{x}}_{M,n} & \vec{\mathbf{w}}_M \end{bmatrix}, \tag{1}$$

$\vec{\mathbf{x}}_{i,j}$是第$i$筆訓練資料的第$j$天的訂房資訊 $(1 \leq i \leq M, 1 \leq j \leq n)$，它是 $1 \times 4$的行向量(row vector)，存放房價、間隔天數 、入住日星期別與入住日節日別等四個元素。$\vec{\mathbf{w}}_i$是 第$i$筆訓練資料的訂房資訊，它是$1 \times 2$(沒有考慮房型)或$1 \times 3$(有考慮房型)的行向量，存放包括訂房日星期別、訂房日節日別，以及房型等資訊。而輸出矩陣$\mathbf{Y}$是$M \times 1$的列向量，表示爲：

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}, \tag{2}$$

$y_i$ 是第$i$筆訓練資料的預期輸出結果，也就是我想要估計的房價。

建構統計迴歸與機器學習模型分成「切割訓練集與測試集」、「資料歸一化」、「選定歷史資料天數」，以及「實際預測與驗證」等四個步驟。首先，我將蒐集房價資料切割成訓練集與測試集，以建構矩陣$\mathbf{X}$與$\mathbf{Y}$，其中，訓練集的目的是用來訓練模型，測試集的目的是驗證訓練模型的效果，防止所訓練的模型過

度擬合訓練資料，在大部分的研究中會保留60%~80%的資料當作訓練集，爲了保留時間的連續性，所以我決定將手邊一整年的資料切成兩段，取前九個月作爲訓練集，後三個月作爲測試集。資料歸一化的目的是爲了把不同類型的數據統一到一個參考系下，這樣比較起來才有意義。目前參考文獻的做法是採用「特徵歸一化」，也就是先把所有房型的原始數據組合成訓練資料，再將所有訓練資料的數據進行最大最小標準化，把最大值對應到1，最小值對應到0。但因爲各種房型的價差很大，實際應用後發現這種作法的平均絕對值百分誤差(mean absolute percentage error, MAPE)會高達36%以上。爲了解決這個問題，我改採用「房型歸一化」，先將每個房型的資料各自做最大最小標準化，再將所有房型各自標準化後的數據組合成訓練資料。使用「房型歸一化」之後，就可成功將MAPE降到5%以下。在完成以上兩步驟之後，接下來需要選擇適當的歷史資料長度來進行單日房價預測，以得到預測結果較精準的模型。根據資料視覺化分析的結果，我發現房價常以七天爲單位變動，因此以7爲基底，選定不同倍率的歷史資料長度n，找出測試集的MAPE最小的n值，作爲模型的最佳歷史資料天數。最後，我選擇入住日爲 2021/6/1的資料，把從2020/7/1開始每天訂房所爬取的房價當作實際值以及房價預測模型所需的歷史資料，再用我提出的房價預測模型進行隔日房價預測，再找出預測值與實際值的平均絕對值百分誤差MAPE，以驗證房價預測模型的準確性。

# 肆、研究過程

接下來介紹視覺化分析、建構模型與實際預測的研究過程與初步成果。其中，「視覺化分析」會評估不同入住日、不同訂房日、同星級飯店、不同星級飯店、同區飯店、不同區飯店對房價調整的關係與影響，再從中選擇建構模型所需的變數。「建構模型」會介紹我在切割訓練集與測試集、資料歸一化方法、選定合適的歷史資料天數n，以及評估模型好壞的做法。「實際預測」會介紹不同方法建構出的模型的調價預測結果。

## 一、視覺化分析

在搜集資料的過程，我根據參考文獻選擇紀錄不同入住日、不同訂房日、同星級飯店、不同星級飯店、同區飯店、 不同區飯店等變因，再根據蒐集的 資料畫

出房價間統計關聯圖，分析不同變因對房價走勢的影響與關係，從中篩選與房價變動呈中、高度相關的變因，作爲建構模型所需的變數(特徵)。

圖4是自2020-07-01起，君悅酒店不同房型每日入住價格變化圖。由圖中可發現，不同房型價格不同，但價格調整有一致的趨勢，經過比對數據後發現，房價會在周五、周六兩天調漲，且調價幅度類似，但在特殊假期(元旦、端午連假等)則會根據房型而有不同程度的大幅調漲。此外，還可從圖中看出因爲疫情的不確定性，導致房價有許多不規律的波動，這也顯現飯店業者對調價系統的需求。

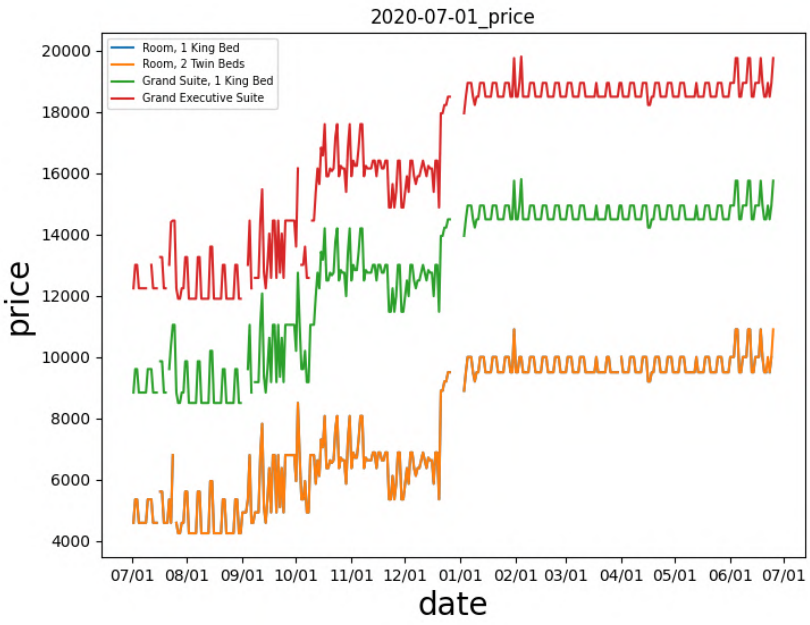



圖 4: 君悅酒店不同房型每日入住價格變化圖

圖5是君悅酒店，相同房型(King Bed)的每日入住房價走勢圖。圖上四條線分別代表在2020/7/1、2020/8/1、2020/9/1、2020/10/1等四個不同日期進行訂房所得到的入住日(註：從訂房日隔日到2021/7/1)房價。由圖中可發現，針對同一房型，選擇在2020年8月份入住，在2020/7/1日訂房 (藍線)會比 2020/8/1日訂房(橘線)的價格便宜約1500元；但如果是在2021年1月份之後入住，2020/7/1日訂房 (藍線)卻要比2020/10/1日訂房(紅線)的價格多付約3500元，跟大眾認知的早鳥優惠不同。以同一訂房日來看，除了10至12月的區間可看到價格下降之外，其餘入住日的房價都有逐步墊高的相近趨勢，乍看之下會讓人得到訂房日與入住日間隔越長價格越高，與早鳥優惠概念正好相反的結果，但這應該是因爲COVID-19導致在前半年旅

遊業蕭條所造成飯店房價跳降。爲了確認，我另外從交通部觀光局網站所提供的觀光業務統計資訊 [10]中，找到2019年與2020年台北地區觀光旅館的住用率，前者皆高於75%，後者卻急遽下滑至24%，證實我的推論。

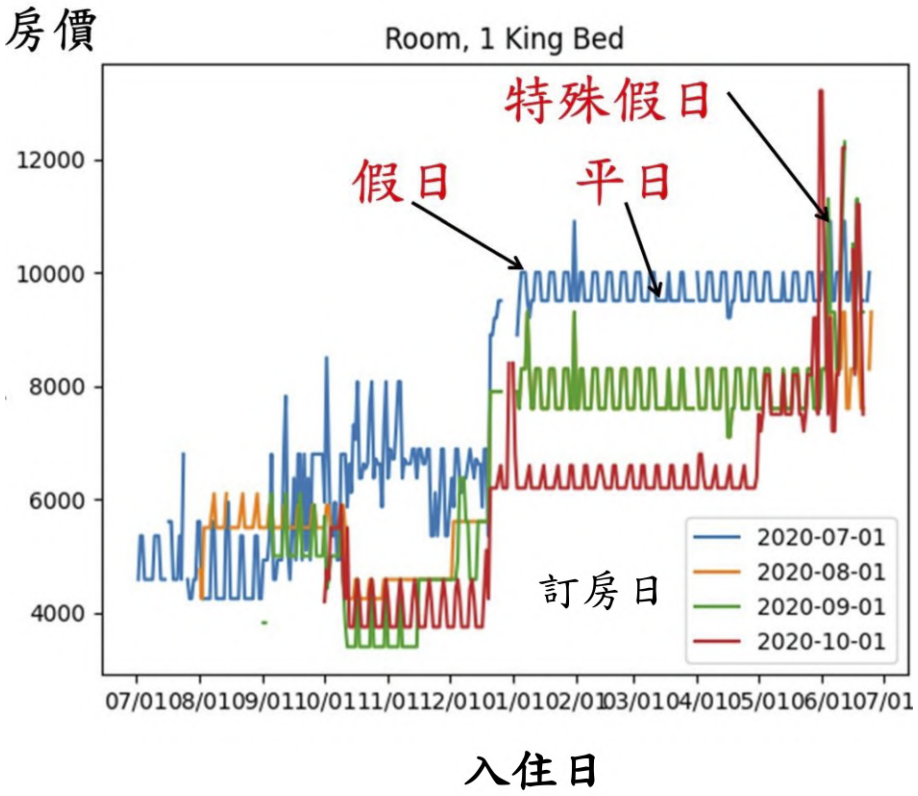



圖 5: 君悅酒店不同訂房日每日入住房價走勢圖

針對不同入住日所觀察到的房價變化，雖然可以看到不同訂房日的房價隨變動趨勢，但是這個價格的變動很可能是因爲淡、旺季的供給需求這些從數據上看不出來的因素所造成的，並不容易有一個客觀的比較基準。因此，本研究以君悅酒店(Grand Hyatt Taipei)爲例，進一步從數據中，比較相同入住日的房價如何隨訂房日的不同而改變，結果如圖6所示。圖6是2020/09/01~2021/01/01等三個不同入住日隨「訂房日」改變所獲得的房價，圖中的每一條線代表「某一個入住日，在不同訂房日的房價」，圖上的斷點代表數據庫中沒有資料(房間已經完售)。可以從圖上看出，不同入住日的房價都可能會在特定日期(如7/29)同時調價，原因可能是飯店或訂房平台針對特定假期的集體促銷活動。

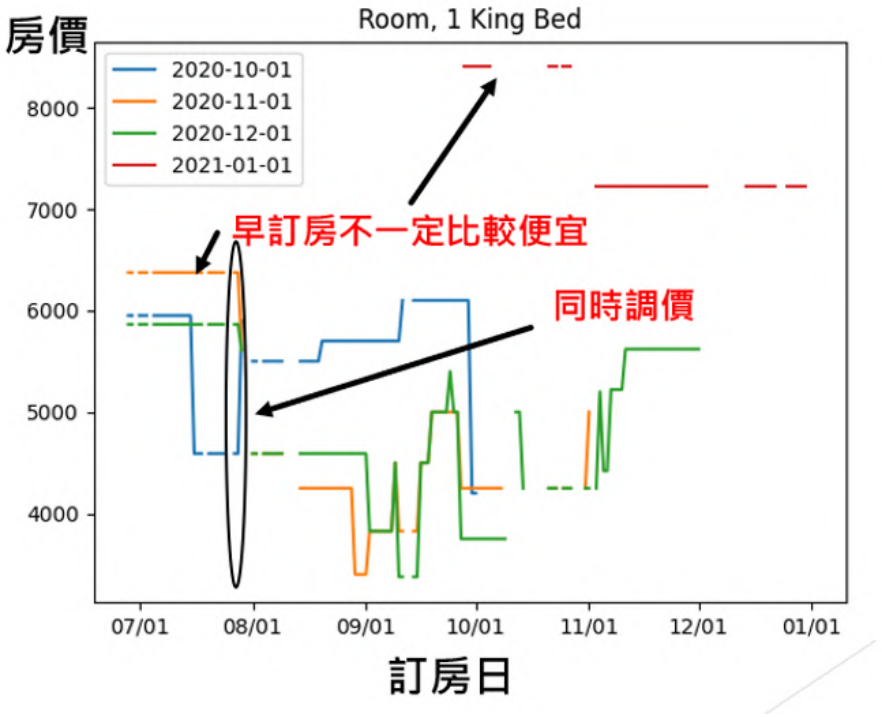



圖 6: 君悅酒店訂房日與房價的關係

不同訂房日與入住日的間隔天數(幾天前訂房)與房價變動的關係，結果如圖7。圖7是2020/10/01~2021/01/01等四個不同入住日的房價，隨著「訂房日與入住日的間隔(interval)」改變的結果，圖中的每一條線代表「某一個入住日，在入住日前90天內，每一天的訂房價格」，圖上的斷點代表數據庫中沒有資料(房間已經完售)。從圖中可看出，房價的調整並不是嚴格遞增函數(早鳥訂房優惠並不存在)，且不同入住日的訂房價格波動趨勢大不相同，顯示房價與入住日相關。以入住日是2020/10/01日爲例，大約要在入住日前60~80日訂房才可獲得最低價；如果是2020/11/01日，最低價會落在入住前60日附近；2020/12/01日，最低價會落在入住前85日左右;如果是2021/01/01日入住，要越早訂房越好，因爲可能訂不到房間，但是可能因爲疫情緣故造成需求下降，所以才會越晚訂房價錢越低。從圖上還可看出，每一條線在特定Interval會持平，在接近入住日則有上揚的趨勢。整體來說，入住日房價的調整趨勢主要與入住日當日特性(「星期幾」、「節日別」)等因素相關。我另外用圖8來同時觀察入住/訂房日對君悅酒店房價的影響。

根據手邊的資料，五星級的King、四星級的Twin、三星級的Double分別是各星級飯店中最多家飯店上架的房型。五星級飯店King的房價波動的週期性變

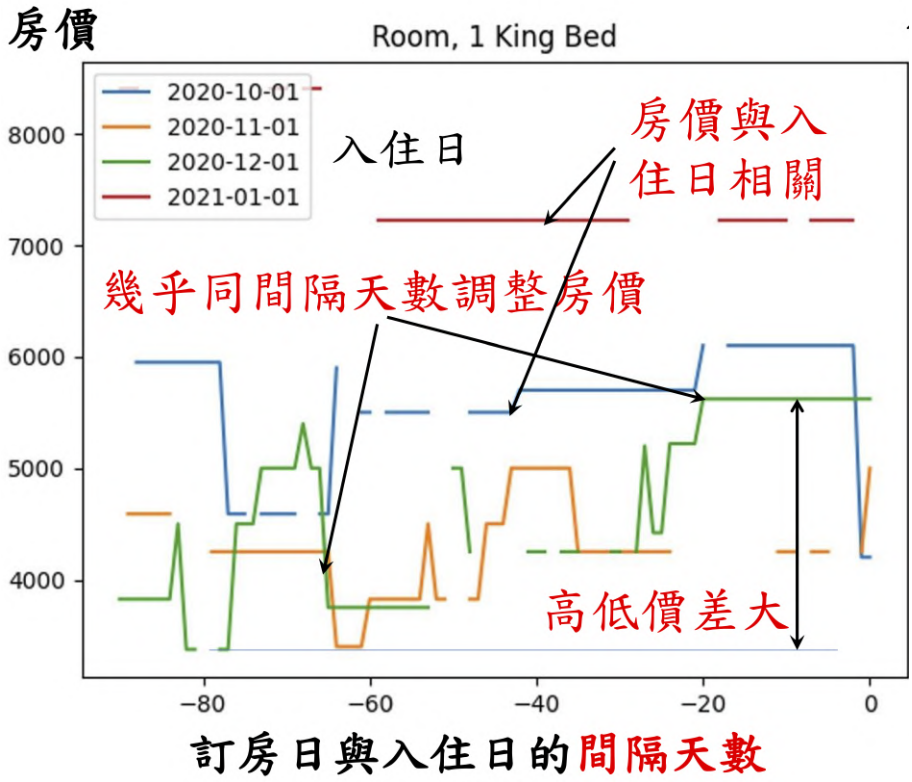



圖 7: 君悅酒店入住日天數間隔與房價的關係

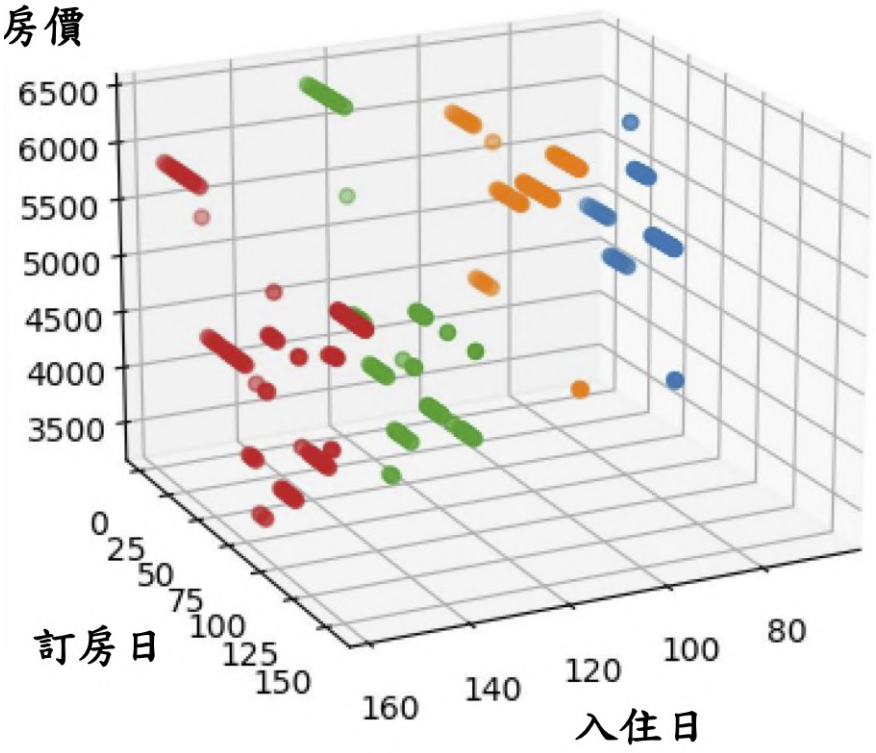



圖 8: 君悅酒店入住/訂房日對房價的影響

化非常明顯，而且房價隨入住日起伏並不大，整體呈中、高度相關。四星級飯店Twin的房價波動的週期性變化比較不明顯，部分房價關聯性極高(Home Hotel, La Meridien Taipei, Sheraton Grand Taipei Hotel, Palais de Chine Hotel)，但與地區較沒有太大關係。三星級的Double房型則有非 常規律的調價趨勢，彼此的房價關聯性也最高，推估應是每間飯店價錢非常相近，爲了吸引顧客，必須隨時依照市場、競爭對手調價。而從以上觀察發現，不同飯店但相同星級間調價似乎有關聯，但與所在地區關聯較小，這可能是因爲台北市交通便利，所以地區的影響性較不明顯。總結來說，從以上幾點分析，我發現影響房價制定與調整變因有下列幾點，飯店星級、房型、入住日是星期幾、入住日的節日別、訂房日、訂房日與入住日天數間隔等幾點。我也從中發現地區(客群)對於台北車站及101大樓周圍飯店房價調整沒有太大的關連以及四、五星級間飯店房價關聯程度較不明顯。這個現象也呈現出四、五星級的飯店在定價時可能沒有考慮競爭對手的定價與應對。反而是三星級可能因爲彼此同質性過高與競爭對手太多，所以必須針對其他飯店的定價予以調整，保持競爭力。由於篇幅限制，五星級的King、四星級的Twin、三星級的Double的價格走勢圖與關聯圖就不一一列出，留在附錄一供參考。

依據視覺化分析的結果可看出，入住日(星期幾、是否爲特殊節日)與訂房日是兩項影響房價的重要因素，但根據資料分析的結果，看不出房價調整的明顯規律性。此外，影響房價因素還有飯店星級、房型等其他因素，使結果更加複雜，所以接下來將借助統計迴歸或機器學習來解決此問題。

## 二、建構模型與實際預測驗證

以下建構模型與實際預測驗證皆以君悅飯店爲例。根據研究方法的規劃，我在切割訓練集與測試集之後，完成資料歸一化，找出了歷史資料天數n、比較有、無將房型納入特徵的兩種模型，並以所建構的模型用手邊的資料進行實際預測與驗證，相關結果整理如下。在研究過程中，我嘗試了Linear Regression、Ridge Regression 、Random Forest、XGBoost等不同方法，發現效果最好的線性迴歸與機器學習演算法分別是Ridge Regression與XGBoost，以下就用這兩個模型爲例，說明採用不同歸一化的方法、決定歷史資料天數n，以及是否考慮房型的研究過程，用平均絕對值百分誤差(MAPE)以及預測結果來比較這兩種方法的結果，其餘的結果則留在附錄二供參考，不再詳述。圖9(a)與圖9(b)是Ridge Regression採用特徵歸一化的結果。圖9(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線

與橘線分別代表訓練集與測試集的MAPE。在選擇歷史資料天數n的時候，我是取在測試集MAPE最小的n值做爲最佳歷史資料天數n，在本例中，可以從圖9(a)看出$n = 175$時，有最小的MAPE。圖9(b)是用最佳歷史資料天數$n = 175$訓練的模型，進行預測並與實際資料驗證的測試結果，其中，藍線與橘線分別代表模型預測結果以及實際資料。從圖9(b)可看出，預測結果並不理想。


(a) MAPE　　(b) 測試結果

圖 9: Ridge Regression採用特徵歸一化的結果

圖10(a)與圖10(b)是機器學習方法XGBoost採用特徵歸一化的結果。圖10(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線與橘線分別代表訓練集與測試集的MAPE，從圖中可看出，在測試集，$n = 175$有最小的MAPE。圖10(b)是從測試集中，任意取出一組相同入住日、不同訂房日所取得的房價變化資料，以$n = 175$訓練模型所獲得的預測結果，並把預測值與實際值比對驗證的結果，其中，藍線與橘線分別代表模型預測結果以及實際房價走勢，從圖10(b)可看出，預測的結果的MAPE很高，也與實際房價走勢差很多。

就預測結果來看，不管採用哪一種線性迴歸與機器學習方法，或不同的歷史資料天n值，採用特徵歸一化的預測結果都不理想，MAPE很大，與實際房價走勢結果也差很多。經過比對數據與深入思考後，我發現這是因爲特徵歸一化所導致的。首先，大部分的應用所考慮的訓練資料是同一種資料(如，同一檔股票)，而我的訓練資料則爲多種(房型)的房價資料彙整而成的。其次，特徵歸一化方法會針對每一個特徵進行歸一化，所以在實際資料驗證時，因爲輸入房型的資料不同、多寡，導致歸一化後的值也就不同，進而影響預測結果。爲了解決這個問題，我另外提出「先將每個房型的房價做歸一化後，再直接轉爲輸入資料」的房


(a) MAPE (b) 測試結果

圖 10: XGBoost採用特徵歸一化的結果

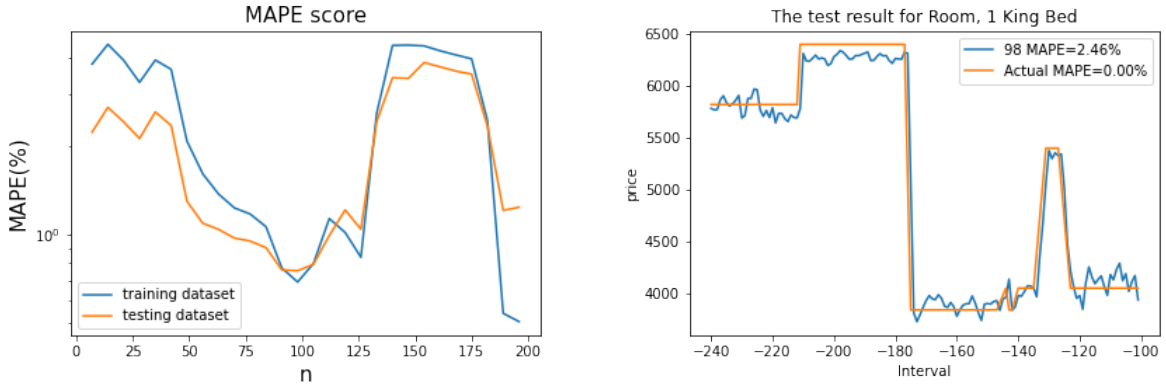
型歸一化方法。

從視覺化分析中，我發現同一家飯店的調價模式很相近，如果採用房型歸一化將每種房型的房價資料先做最大最小歸一化，這樣所有房型的數據皆是介於0、1間，或許可消除因爲房型所導致房型間的價差，所以在研究過程中，我同時建構「不考慮房型特徵」與「考慮房型特徵」這兩種模型，並將Ridge Regression與XGBoost的預測結果分別整理在圖11、12與圖13、14。圖11(a)與圖11(b)是Ridge Regression採用房型歸一化、不考慮房型特徵的結果。圖11(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線與橘線分別代表訓練集與測試集的MAPE，從圖中可看出，在測試集，同樣在$n = 175$有最小的MAPE。圖11(b)是以$n = 175$訓練的模型，進行預測並與實際資料驗證的測試結果，其中，藍線與橘線分別代表模型預測結果以及與實際資料的比對。從圖中可看出，這次預測的結果很好，可以將MAPE降到5%，預測的結果也與房價變化趨勢一致。圖12(a)與圖12(b)是Ridge Regression採用房型歸一化並考慮房型特徵的結果。圖12(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線與橘線分別代表訓練集與測試集的MAPE，從圖中可看出，在測試集，在$n = 98$有最小的MAPE。圖12(b)是以$n = 98$訓練的模型，進行預測並與實際資料驗證的測試結果，其中，藍線與橘線分別代表模型預測結果以及實際資料。比較圖11(b)與圖12(b)可看出，將房型列入特徵之後，模型在測試集的MAPE更小、預測的結果更好，

圖13(a)與圖13(b)是XGBoost採用房型歸一化、不考慮房型的結果。圖13(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線與橘線分別代表訓練集與測


(a) MAPE (b) 測試結果

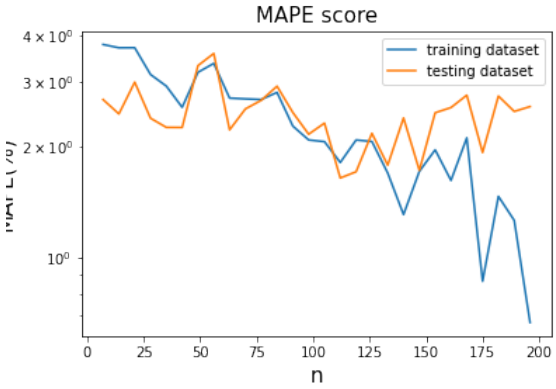
圖 11: Ridge Regression採用房型歸一化，不考慮房型特徵的結果



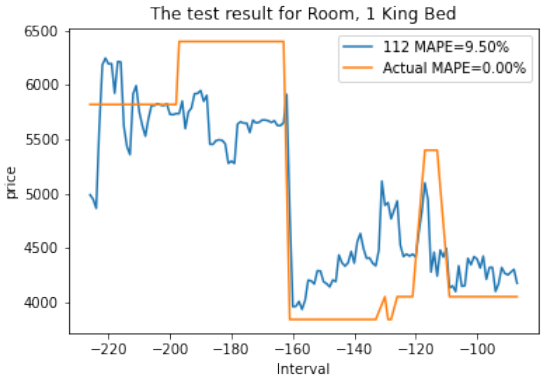

(a) MAPE (b) 測試結果

圖 12: Ridge Regression採用房型歸一化，考慮房型特徵的結果

試集的MAPE。從圖中可看出，在測試集，在$n = 112$有最小的MAPE。圖13(b)是以$n = 112$訓練的模型，進行預測並與實際資料驗證的測試結果，其中，藍線與橘線分別代表模型預測結果以及實際資料。從圖中可看出，這次預測的結果比特徵歸一化(圖10(b))的結果好很多。圖14(a)與圖14(b)是XGBoost採用房型歸一化、考慮房型的結果。圖14(a)是MAPE隨著歷史資料天數n的變化關係，其中，藍線與橘線分別代表訓練集與測試集的MAPE。從圖中可看出，在測試集，在$n = 98$有最小的MAPE。圖14(b)是以$n = 98$訓練的模型，進行預測並與實際資料驗證的測試結果，其中，藍線與橘線分別代表模型預測結果以及實際資料。而我們可以發現有考慮房型特徵(圖14(b))比沒有考慮房型特徵(圖13(b))的MAPE少將近一半。

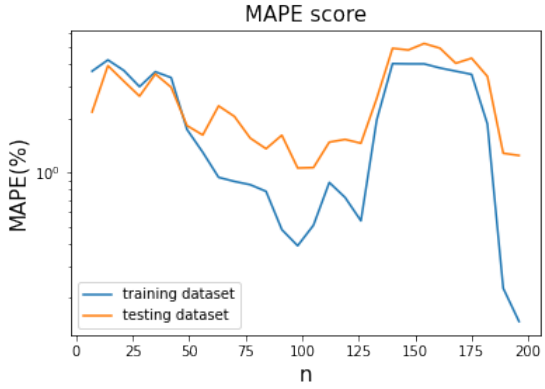



(a) MAPE

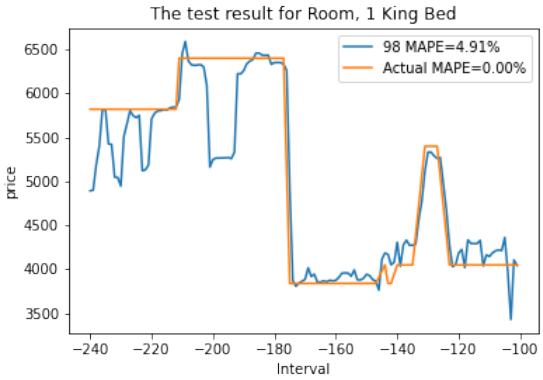



(b) 測試結果

圖 13: XGBoost採用房型歸一化，不考慮房型特徵的結果




(a) MAPE




(b) 測試結果

圖 14: XGBoost採用房型歸一化，考慮房型特徵的結果

總結來說，房型歸一化能解決特徵歸一化在預測結果與測試集上的MAPE偏差很大的問題，預測結果也比特徵歸一化方法好很多；將房型特徵納入模型的預測效果比較好；統計迴歸的Ridge Regression比機器學習的XGBoost在測試案例的預測結果有更好的表現。從結果還可看出，同一家飯店不同房型的調價策略仍有差異。有趣的是，$n$ =175(25週)、112(16週)、98(14週)是較常出現的最佳的歷史資料天數，詳細原因有待之後再研究。本研究所開發的程式碼與房價資料庫放在附錄三與附錄四供參考。

## 三、不同訓練方法預測結果比較

圖15是房型歸一化、不考慮房型(no id)與考慮房型(id)兩種方法在測試集的MAPE，圖16(a)與圖16(b)則分別是房型歸一化、不考慮房型與考慮房型兩種方

法在2021/06/01爲入住日的預測值與實際值的結果。經過比較發現，有考慮房型的MAPE皆比沒有考慮房型的小、統計迴歸方法的MAPE也都比機器學習方法小，但Linear與Ridge Regression的MAPE結果相差不大。機器學習方法則是XGBoost的表現較好。初步推論可能是因爲每次調價過程都是線性，或者是因爲我對機器學習方法訓練技巧還不夠熟練，在訓練機器學習模型時皆使用預設參數，導致機器學習模型結果比統計回歸方法差。雖然在圖16(a)與圖16(b)中會看到Linear Regression的MAPE比ridge Regression小或是Random Forest的MAPE比XGBoost小，但是這僅是一天的預測結果。就比較模型好壞，主要還是應該參考在測試集上MAPE的結果。

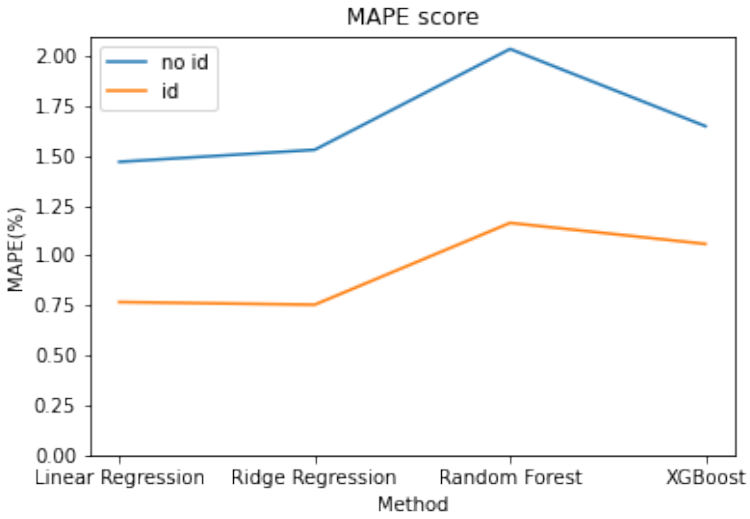
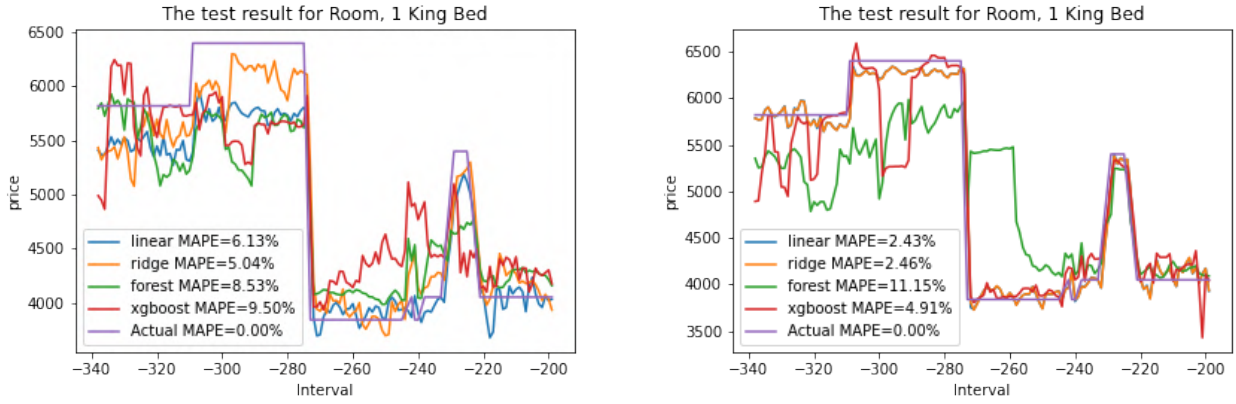



圖 15: 不同模型的在測試集上的預測結果比較




(a) 不考慮房型特徵的結果

(b) 考慮房型特徵的結果

圖 16: 不同模型在2021/06/01爲入住日的預測結果比較

# 伍、結論與應用

- 本研究的創見性是捨棄傳統以「經營者經驗」爲主的定性研究，採用隨手可得的訂房網站，基於「數據」，以定量方式研究飯店業者的調價策略。
- 本研究的貢獻是從訂房網站的房價數據下手，研究飯店的即時調價策略；自行提出「房型歸一化」的資料預處理方法，應用於統計迴歸與機器學習方法建構調價策略模型；最後再以君悅酒店的54萬筆資料爲例，實際建立Ridge Regression與XGBoost建構房價預測模型，並以一整年的房價資料進行驗證，已達到預測值與實際值的估計誤差MAPE僅0.75%。
- 本研究的發現如下：從數據分析結果發現，飯店的調價策略與入住日與訂房日(星期幾、是否爲特殊節日)、間隔天數、房型等變因關係密切。就資料分析結果可發現，三星級飯店間競爭激烈，定價與調價會彼此影響；四、五星級的飯店因爲間數較少，競爭較小，所以彼此的定價與調價關聯性低；即使是同一間飯店，不同房型的調價策略相近但不完全相同。根據實際測試結果發現，利用統計回歸或機器學習方法，來建立飯店調價模型進行房價預測是具體可行的；資料預處理的方法對於預測結果影響很大，考慮房型特徵，使用房型歸一化預處理可有效降低估計誤差。
- 本研究的實際應用爲提供飯店業者調價的參考模型，讓業者可透過分析自家的歷史資料找出自己可能忽略的調價變因，或針對同區域、同星級的飯店，深入分析對手的調價變因或利用隔日房價預測結果，進行更精確的動態價格設定。

在長達13個月的研究期間，適逢COVID-19疫情，造成住房率大幅衰減。我原本會擔心COVID-19疫情會讓本研究的結論有所偏頗，但從實際數據看出房價隨著疫情好壞、政府紓困/振興政策，防疫旅館的設置，以及國內、國際旅遊人潮的影響而變動著，這些變動遠遠超過飯店業者的經驗與想像，也更加確立本研究的重要性。我的這份研究，正好透過數據紀錄了飯店業者在這段時間，因應所有突發狀況當下所做出的即時調價策略。爲了驗證本研究的價值，我特別以君悅酒店入住日爲2021/6/1近一年的房價走勢爲例，利用模型進行逐日房價預測進行實際驗證，MAPE誤差僅2.5%，足以佐證本研究在實際應用上的可行性。

# 陸、參考資料


[1] “OTA學：時代趨勢下的旅宿產業思維,” (https://solomo.xinmedia.com/globaltourismvision/141754)

[2] “Expedia幫旅宿業帶進60億營收的工具大揭密,” (https://www.bnext.com.tw/article/55536/explore-

[3] “跟風串接OTA，對旅宿業者是災難一場還是黑暗中的一道曙光？” (https://market.ltn.com.tw/article/4212)

[4] “飯店訂價策略，重數據？重經驗,” (https://www.hbrtaiwan.com/article_content_AR0009975.html)

[5] “解析線上旅遊產業鏈,” (https://cornerstonevc.tw/brandon-023-value-network-of-online-travel-industry/)

[6] “線上旅遊產業的主要 Players,” (https://cornerstonevc.tw/brandon-022-major-players-in-online-travel-industry/)

[7] “SiteMinder World Hotel Index,” (https://www.siteminder.com/world-hotel-index/data/)

[8] T. Larrieu, “Pricing strategies in online market places and Price Parity Agreements: Evidence from the hotel industry,” Working Paper, July 2019.

[9] “Random Forest Regression,” (https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why)

[10] “交通部觀光局，觀光業務統計,” (https://bit.ly/34t6cbA)

[11] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin, “Learning from data,” Vol. 4. New York, NY, USA:: AMLBook, 2012.

# 附錄一：視覺化分析補充資料

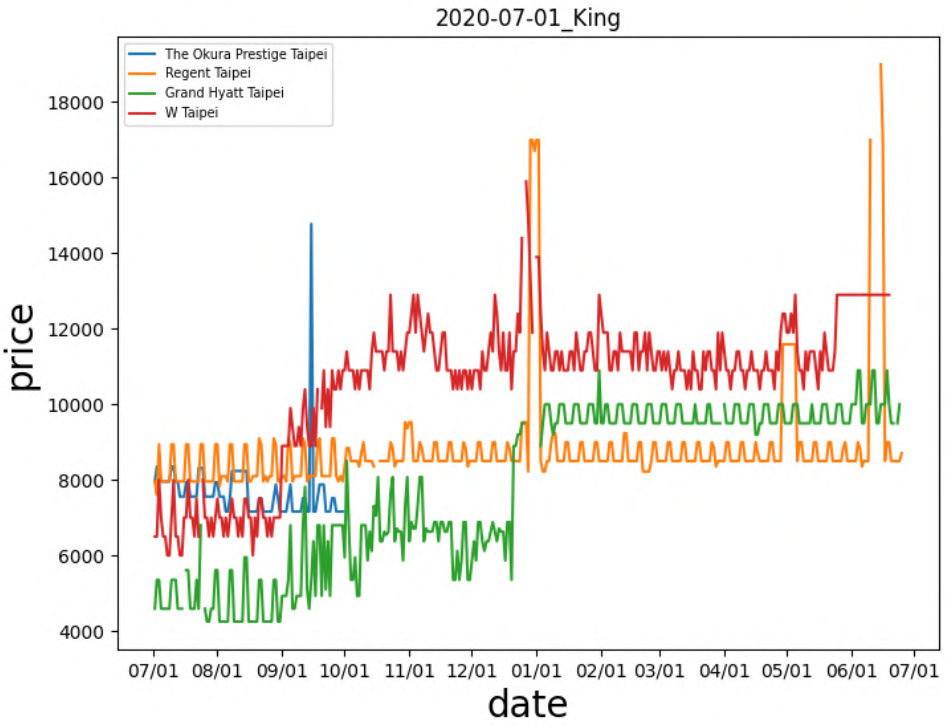



圖 17: 五星級飯店King(單人房)價格走勢圖(前二個在北車、後兩個在101)

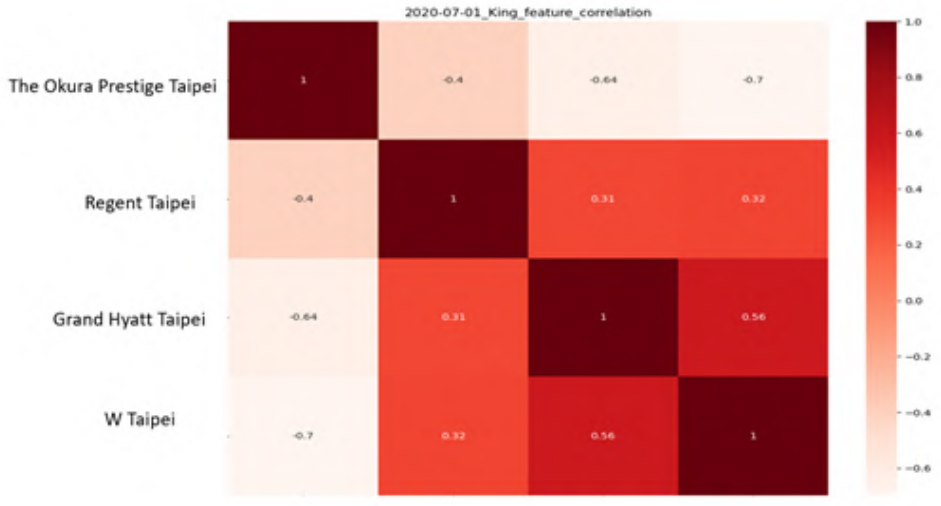



圖 18: 五星級飯店King(單人房)價格關聯(前二個在北車、後兩個在101)

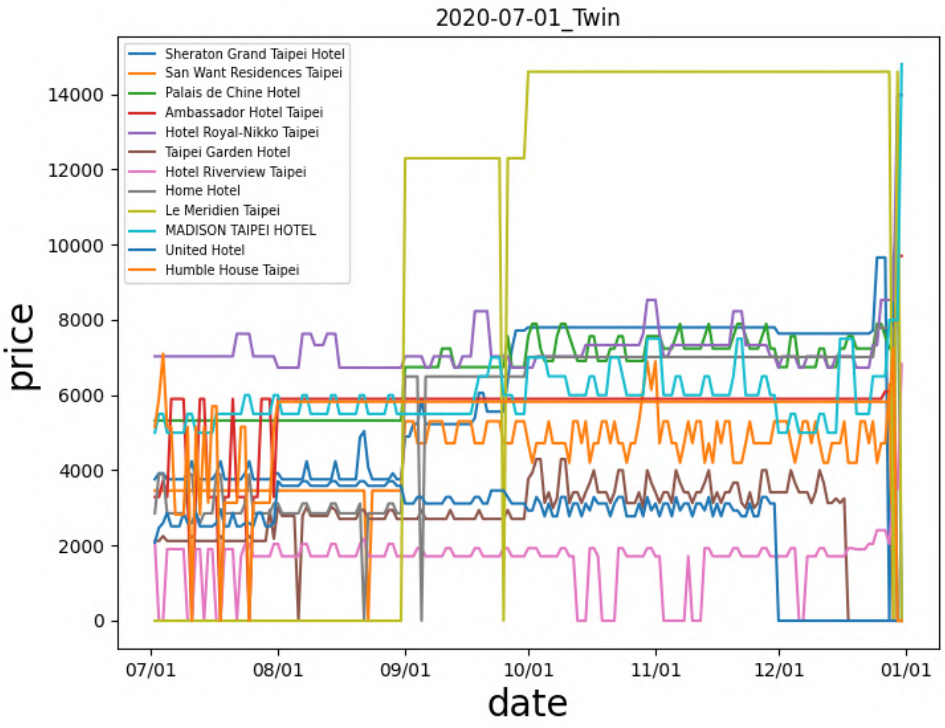



圖 19: 四星級飯店Twin(單床雙人房) 價格走勢圖(前七個在北車、後五個在101)

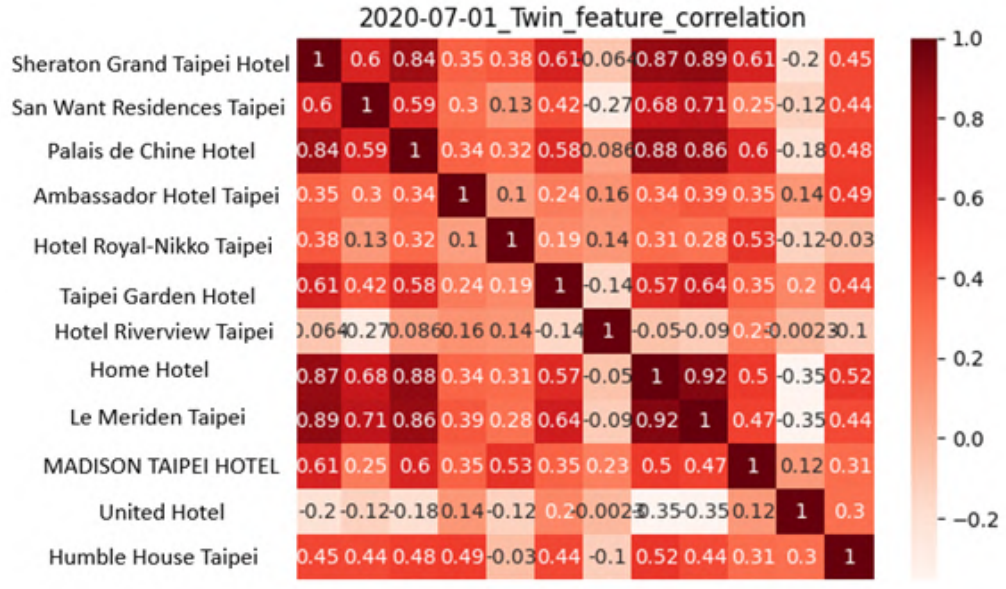



圖 20: 四星級飯店Twin(單床雙人房) 價格關聯圖(前七個在北車、後五個在101)

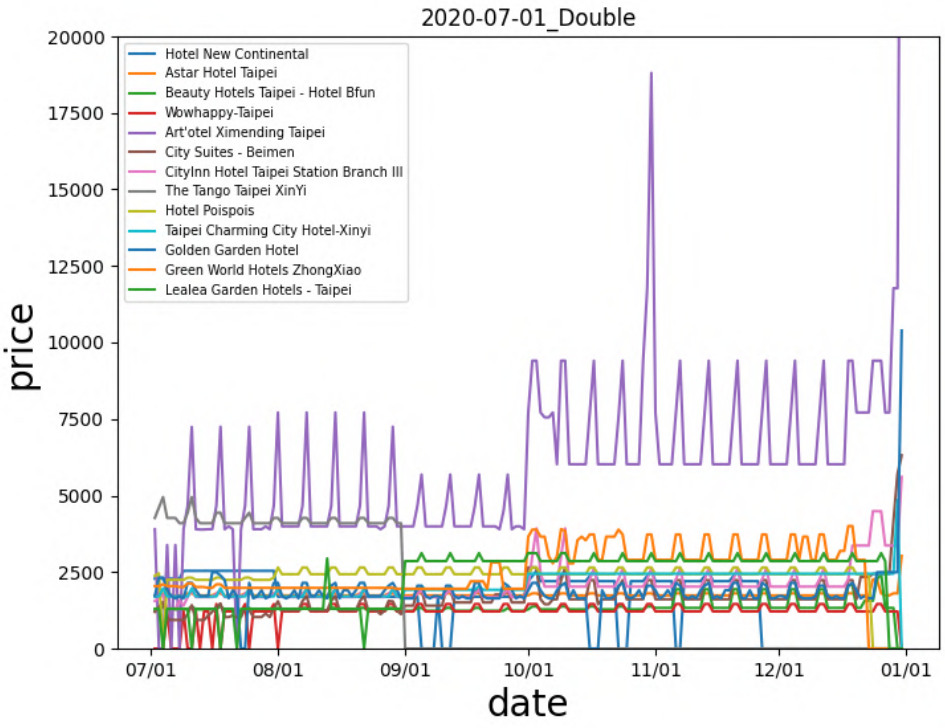



圖 21: 三星級飯店Double(雙床雙人房)價格走勢圖(前七個在北車、後六個在101)

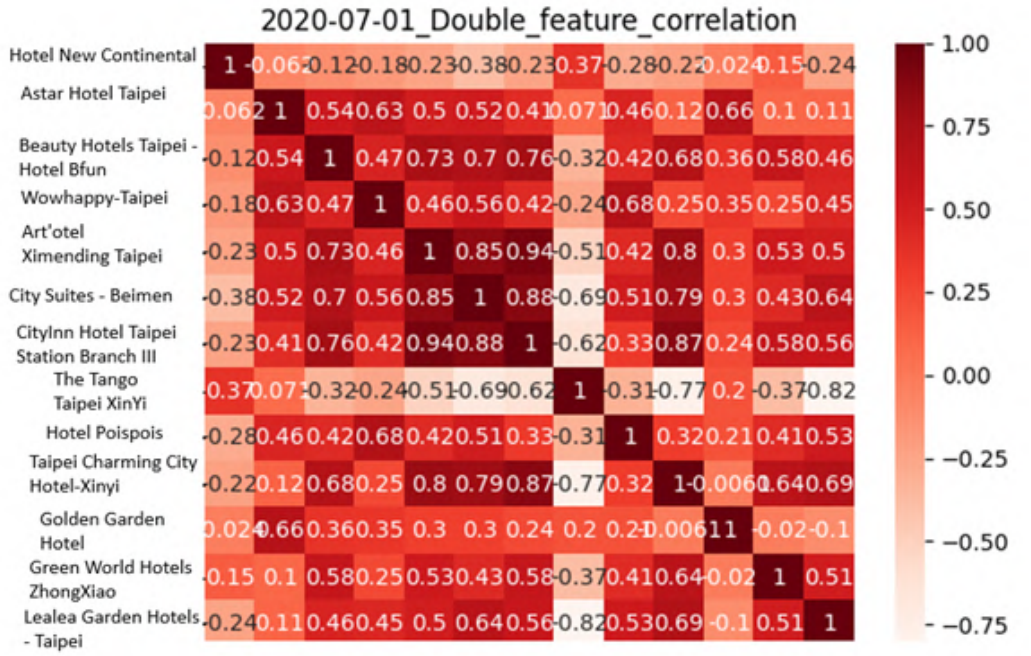
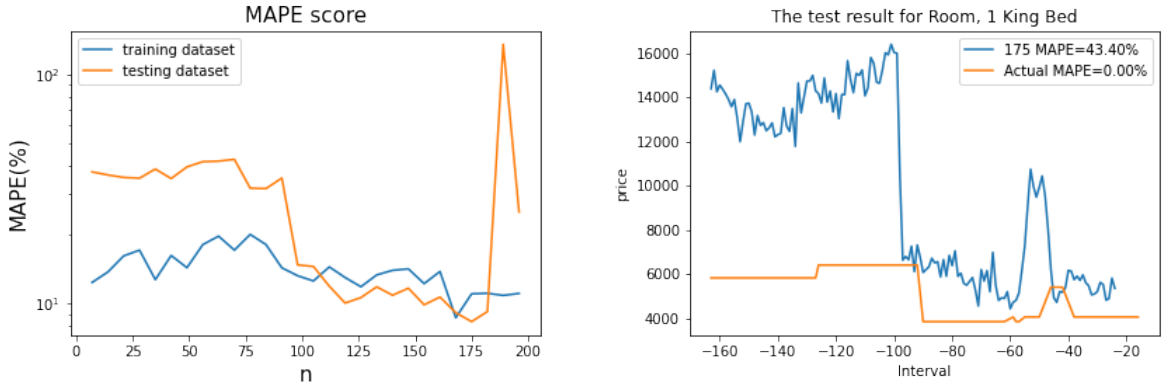


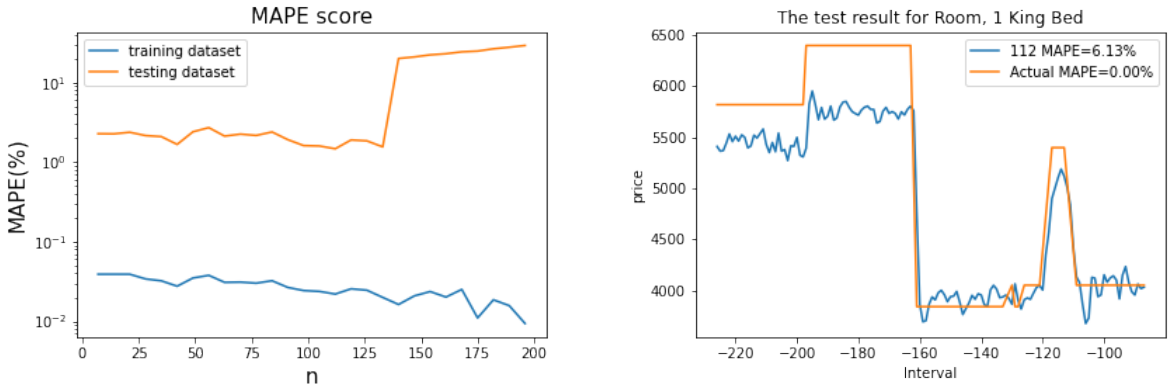

圖 22: 三星級飯店Double(雙床雙人房)價格關聯圖(前七個在北車、後六個在101)

# 附錄二：建構模型補充資料



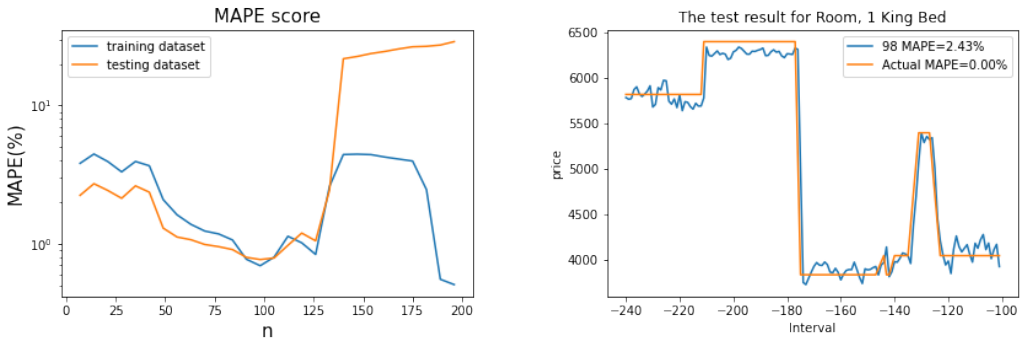

(a) MAPE (b) 測試結果

圖 23: Linear Regression採用特徵歸一化的結果



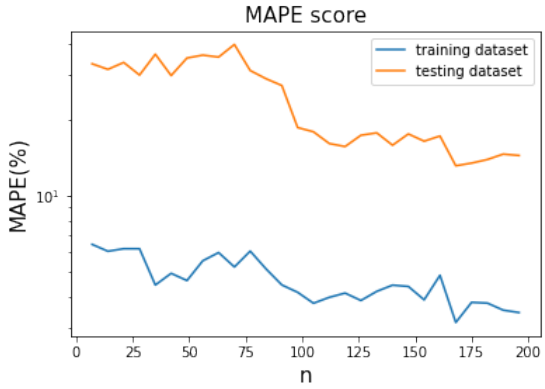

(a) MAPE (b) 測試結果

圖 24: Linear Regression採用房型歸一化，不考慮房型特徵的結果

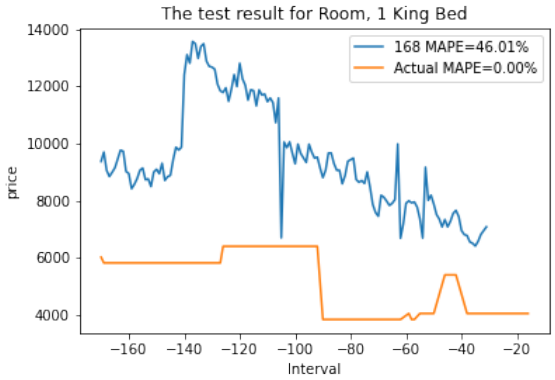



(a) MAPE (b) 測試結果

圖 25: Linear Regression採用房型歸一化，考慮房型特徵的結果

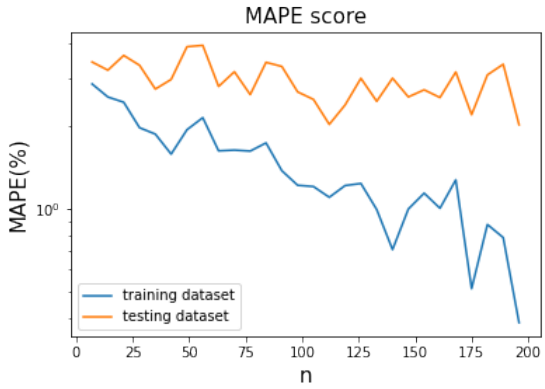



(a) MAPE

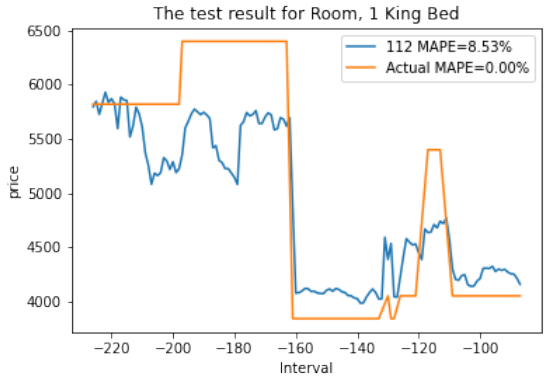



(b) 測試結果

圖 26: Random Forest採用特徵歸一化的結果

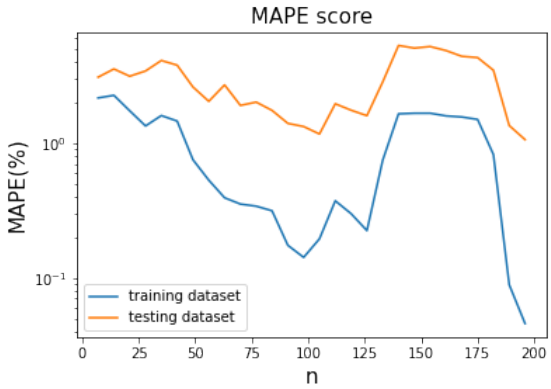



(a) MAPE

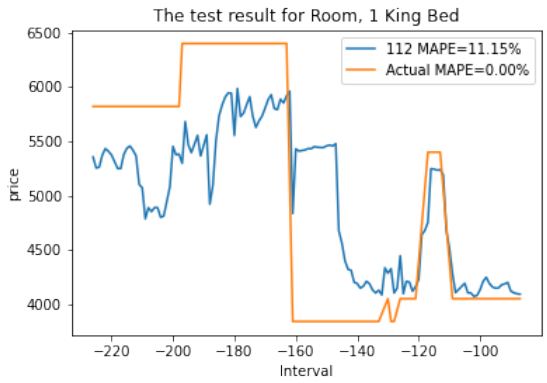



(b) 測試結果

圖 27: Random Forest採用房型歸一化，不考慮房型特徵的結果




(a) MAPE




(b) 測試結果

圖 28: Random Forest採用房型歸一化，考慮房型特徵的結果

# 附錄三：程式碼

程式碼請參考我的HackMD筆記(https://hackmd.io/NR_tEFjXTc-xWkqIdJ_orQ?view)，其中，

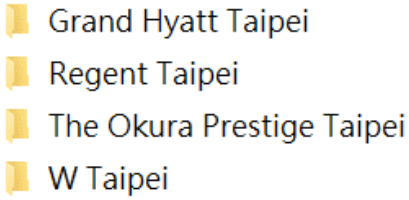
- create_hotel_roomtype_database.py : 建立房型資料庫
- crawhotels.py : 自動化蒐集房價資料
- data_preprocessing.py : 資料預處理
- create_training_data.py : 建立訓練資料
- create_testing_data.py : 建立測試資料
- linear_training.py : 訓練Linear Regression 模型

# 附錄四：房價資料庫

以下爲我建立的房價資料庫部分內容


(a) 三星級


(b) 四星級




(c) 五星級

圖 29: 飯店名稱

```
Grand Hyatt Taipei: [] 15 items
  0: "Room, 1 King Bed"
  1: "Room, 2 Twin Beds"
  2: "Deluxe Room, 1 King Bed"
  3: "Deluxe Room, 2 Twin Beds"
  4: "Room, 2 Twin Beds, View"
  5: "Room, 1 King Bed, View"
  6: "Premium Room, 1 King Bed"
  7: "Deluxe Room, 1 King Bed (Club Access)"
  8: "Deluxe Room, 2 Twin Beds (Club Access)"
  9: "Club Room, 2 Twin Beds, View"
  10: "Club Room, 1 King Bed, View"
  11: "Premium Room, 1 King Bed, View (Club Access)"
  12: "Grand Suite, 1 King Bed"
  13: "Grand Executive Suite"
  14: "Grand Executive View Suite"
```

圖 30: 房型列表

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | searchtime | 2020/6/28 | 2020/6/29 | 2020/6/30 | 2020/7/1 | 2020/7/2 | 2020/7/3 | 2020/7/4 | 2020/7/5 | 2020/7/6 | 2020/7/7 | 2020/7/8 | 2020/7/9 | 2020/7/10 | 2020/7/11 | 2020/7/12 | 2020/7/13 | 2020/7/14 | 2020/7/15 | 2020/7/16 | 2020/7/17 | 2020/7/18 | 2020/7/19 | 2020/7/20 | 2020/7/21 |
| 2 | 2020/6/28 | 4930 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 3 | 2020/6/29 | | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 4 | 2020/6/30 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 2020/7/1 | | | | 4590 | 4590 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | | 5610 | 5610 | 4590 | 4590 | 4590 |
| 6 | 2020/7/2 | | | | | 4590 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 7 | 2020/7/3 | | | | | | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 8 | 2020/7/4 | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 2020/7/5 | | | | | | | | 4590 | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 10 | 2020/7/6 | | | | | | | | | 4590 | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 11 | 2020/7/7 | | | | | | | | | | 4590 | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 12 | 2020/7/8 | | | | | | | | | | | 4590 | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 13 | 2020/7/9 | | | | | | | | | | | | 4590 | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 14 | 2020/7/10 | | | | | | | | | | | | | 5355 | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 15 | 2020/7/11 | | | | | | | | | | | | | | 5355 | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 16 | 2020/7/12 | | | | | | | | | | | | | | | 5355 | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 17 | 2020/7/13 | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 18 | 2020/7/14 | | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 5610 | 5610 | 4590 | 4590 | 4590 |
| 19 | 2020/7/15 | | | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 4590 | 4590 | 4590 | 4590 |
| 20 | 2020/7/16 | | | | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 4590 | 4590 | 4590 |
| 21 | 2020/7/17 | | | | | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 4590 | 4590 |
| 22 | 2020/7/18 | | | | | | | | | | | | | | | | | | | | | 4590 | 4590 | 4590 | 4590 |
| 23 | 2020/7/19 | | | | | | | | | | | | | | | | | | | | | | | | |
| 24 | 2020/7/20 | | | | | | | | | | | | | | | | | | | | | | | 4590 | 4590 |
| 25 | 2020/7/21 | | | | | | | | | | | | | | | | | | | | | | | | 4590 |

圖 31: 房價表格