國立成功大學
113 學年度第1學期
期末報告

課程名稱:資料科學導論
授課老師:李政德
報告主題:酒店價格趨勢與機票價格趨勢的關聯性--
以桃園-東京國際機場為例

組名:資科一
成員:資訊**115 F34115043** 賴鵬豐
統計**114 H24101337** 陳昱瑄

Table of Contents

0. Brief introduction of the problem

在現代人忙碌的生活中，總會想在空閒時段去其他國家旅遊，在旅遊的時候價格會是考慮的一個重點。其中，有許多因素會造成酒店價格的浮動，像是機票優惠、人力成本、當地物價等等所影響。我們認為機票的價格會是影響酒店價格的一個重要因素。我們會藉由分析酒店價格與機票價格關聯，判斷兩者之間的關聯性，如果有關聯的話，消費者就可以藉由目前的機票價格，推斷目前的酒店價格是不是比較合理來決定是否選擇該酒店。

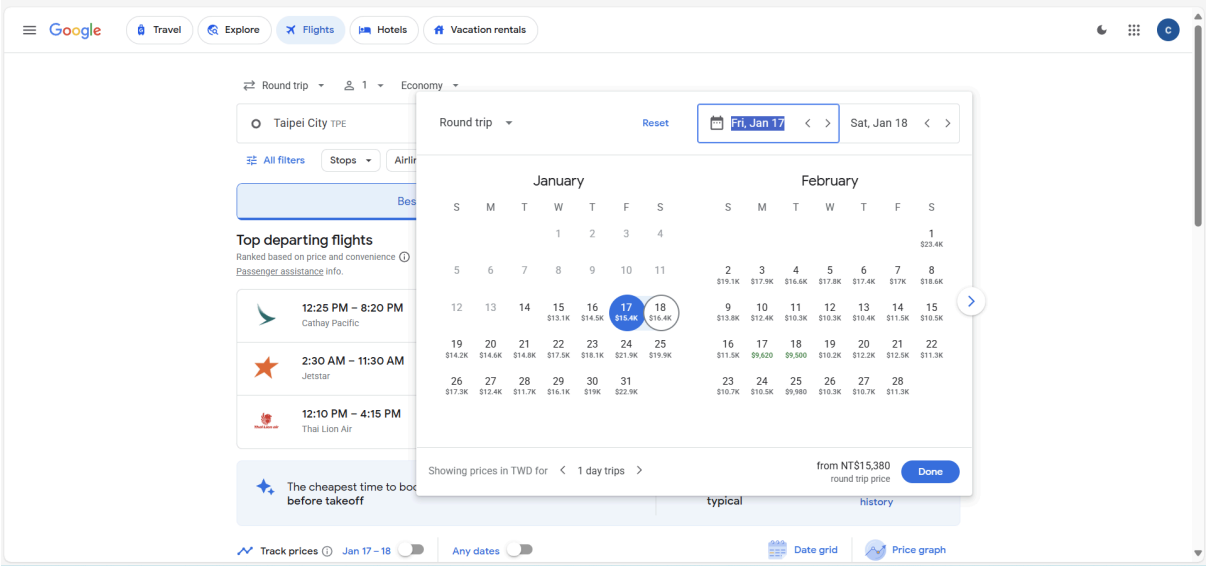
1. Data Integration and preprocessing

Google travel中有提供實時酒店以及機票價格資訊，在爬取flight頁面與hotel頁面後，使用beautifulsoup來獲取酒店資料及價格，不過beautifulsoup僅能爬取靜態網站，但Google travel是動態網站，所以在這裡使用了selenium進行搭配進行爬蟲。

爬取方法-機票：

價格統一爬取資訊頁面的宣傳價格，下圖顯示每天去程的機票價格

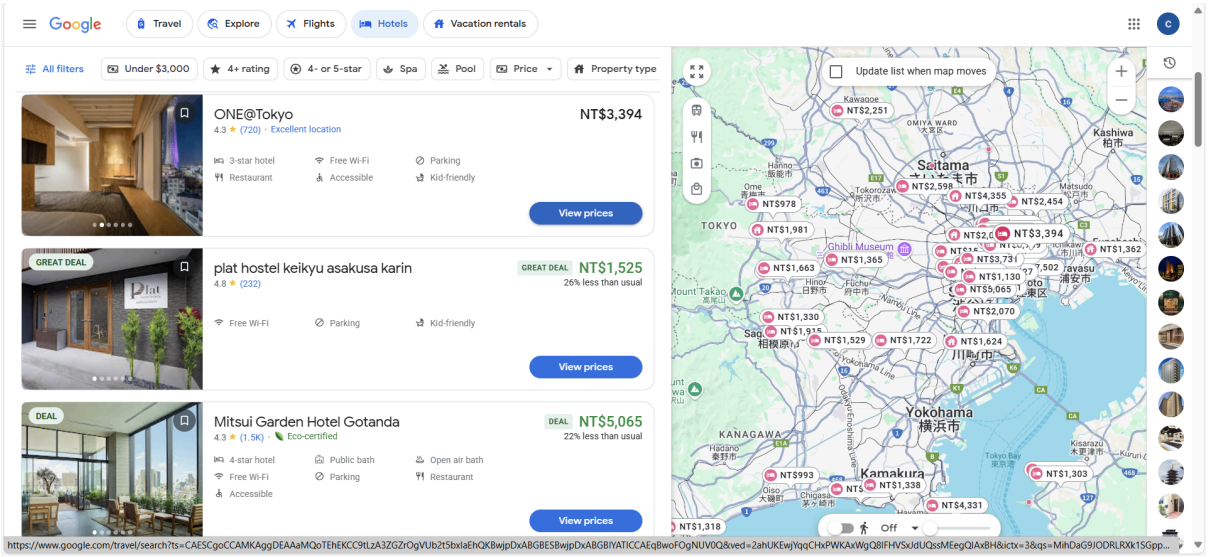
在這裡設定的條件有：直達、最經濟的機票價格(不考慮飛航公司，只比價格)、爬取未來一整年的機票價格。



爬取方法-酒店:

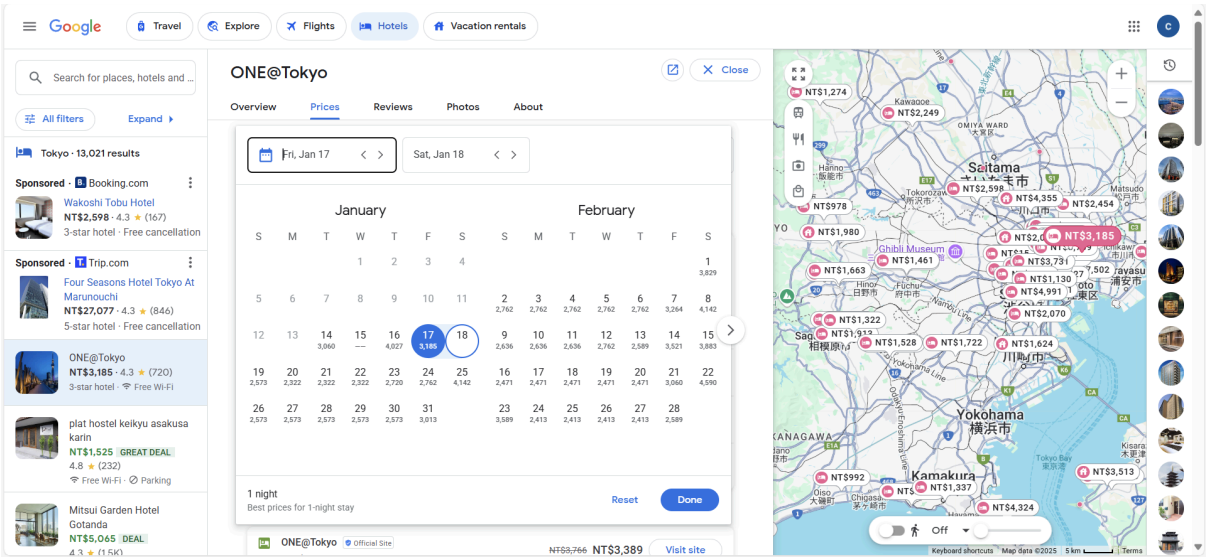
按照星級分別爬取（3星、4星、5星）

先爬取酒店的基本資訊和網址，之後再使用這份資料集逐一訪問各酒店的頁面，並爬取酒店的價格，這樣的爬取方法除了比較節省時間，也是爲了避免網站頻繁刷新導致酒店顯示的順序會打亂，從而影響後續資料集整理的困難。



價格爬取:

繼上文含網址的資料集，逐一訪問酒店頁面並爬取最經濟房型價格和其他基本資訊如:日期、評分等。



2.Insights discovered from the data

將資料整合後，我們需要先分析機票與酒店的關聯性，這裡我們主要考慮機票價格趨勢會不會和酒店價格趨勢同步（正相關或負相關），比如機票價格上升同時酒店價格上升或機票價格下降酒店同時價格下降等等的情況。
考慮到不同間業者的優惠策略不同，這裡會使用多筆酒店價格作爲參考、每個星級有36間，以降低單一業者價格偏差對整體分析的影響，並確保數據的代表性和可靠性。這樣可以更準確反映該星級酒店在市場上的平均價格。

3. Methodology details

我們先對酒店價格進行資料預處理（不同星級獨自處理），這裡使用的方法是先繪製觀察所有酒店價格趨勢折線圖，剔除異常走勢的酒店。因為每個酒店的價格範圍都不一樣，需要將每個酒店的價格範圍進行標準化後再和每個酒店平均起來以分析趨勢走向。接著會得到三筆平均資料集（3，4，5星級酒店），開始和機票價格做分析對比，這裡使用的方法是Correlation（相關性分析）、Granger因果檢驗（評估一個變量的歷史值是否能用來預測另一個變量），以及繪製可視化圖來進行直觀比較。對比後發現整體分數和表現都不理想，機票價格和酒店價格并沒有出現肉眼可見的規律。

相關性分析顯示，整體酒店價格與機票價格的相關係數為-0.0158，顯示兩者之間幾乎沒有線性相關性。可能是因為機票和酒店的價格受到不同市場因素影響，例如機票價格更依賴航班供需和燃油價格，而酒店價格則受到地區旅遊熱度、星級等多重因素影響。因果檢驗顯示，不論是F-test或是Chi-square test中，在10輪訓練中只有兩輪的p值小於0.05，其餘8輪中均大於0.05，代表酒店價格和機票價格之間沒有顯著的因果關係，酒店價格和機票價格的波動來源獨立，且彼此之間的影響較小。使用相關性分析和因果檢驗的限制在於這些方法僅能捕捉線性相關性和直接因果關係，可能無法判斷非線性關係或複雜的交互效應。

進一步驗證：
將各星級的平均價格趨勢來做Correlation和Granger分析來觀察各級酒店有沒有相關性，以此先確認各酒店平均價格趨勢會不會混亂，從而影響前面的分析，並考慮是否需要更換分析方式。

不同星級飯店和機票價格間的比較:

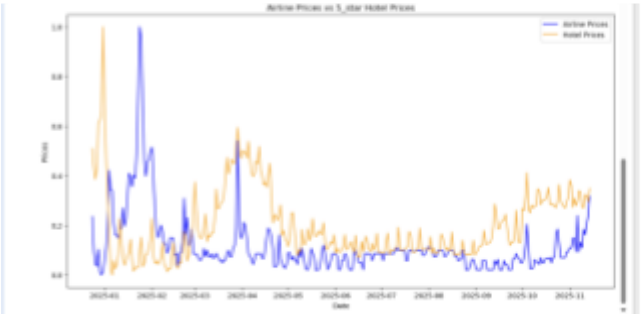
機票vs三星酒店

藍線:機票價格
黃線:三星酒店價格

機票vs四星酒店

藍線:機票價格
黃線:四星酒店價格

機票vs五星酒店

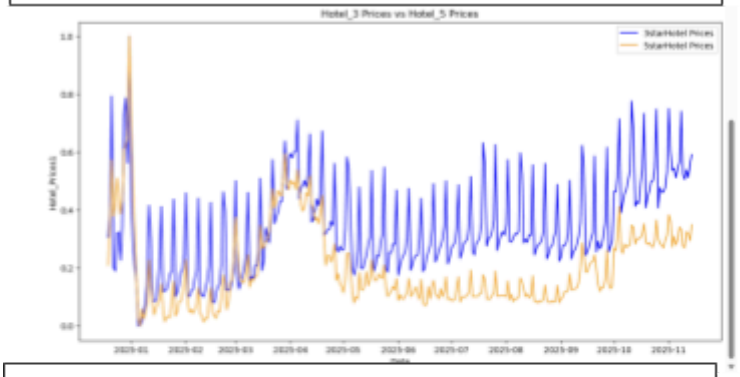


藍線:機票價格
黃線:五星酒店價格

	Granger因果檢驗:lags 1 p-value	Granger因果檢驗:lags 2 p-value	Granger因果檢驗:lags 3 p-value	Correlation
機票/3星級	0.2063	0.1070	0.1808	-0.1235
機票/4星級	0.3421	0.5241	0.7343	-0.1295
機票/5星級	0.5606	0.8887	0.6014	-0.0986

不同星級飯店間的比較:

三星酒店vs五星酒店



藍線:三星酒店
黃線:五星酒店

四星酒店vs五星酒店

藍線:四星酒店
黃線:五星酒店

三星酒店vs四星酒店

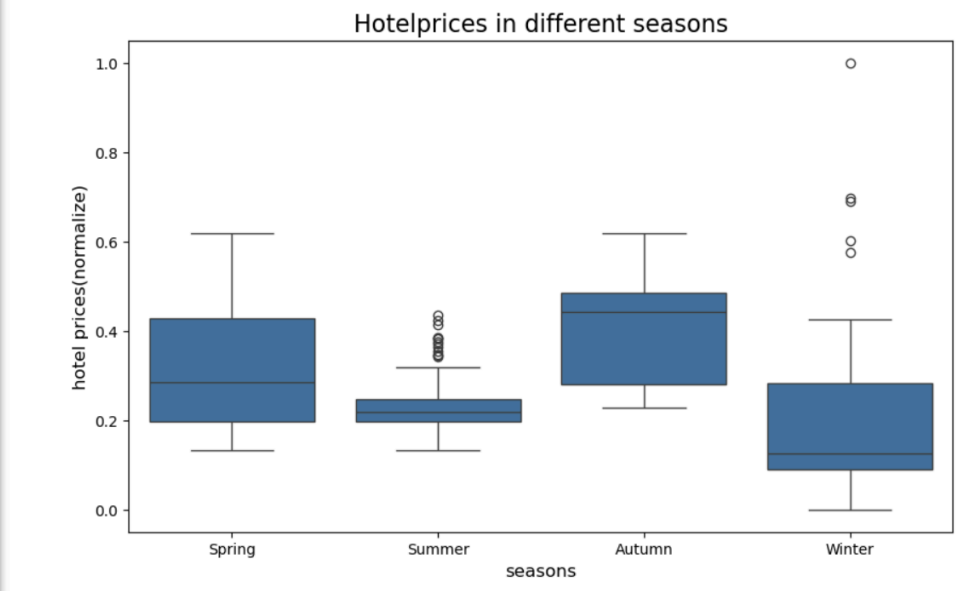
藍線:三星酒店
黃線:四星酒店

	Granger因果檢驗:lags 1 p-value	Granger因果檢驗:lags 2 p-value	Granger因果檢驗:lags 3 p-value	Correlation
3星級/5星級	0.0002	0.0003	0.0008	0.7214
4星級/5星級	0.8467	0.0078	0.0043	0.7162
3星級/4星級	0.0000	0.0000	0.0000	0.8902

如上表所示，各星級的平均價格趨勢呈近同步（正相關）的狀況，由此可見這樣的價格趨勢可以認爲是合理的。機票價格趨勢多少會影響酒店價格趨勢，但這其中的規律和相關特徵可能更加複雜（天氣因素、業者策略、股市等），這是我們能在開始這個課題前沒有考慮到的問題。

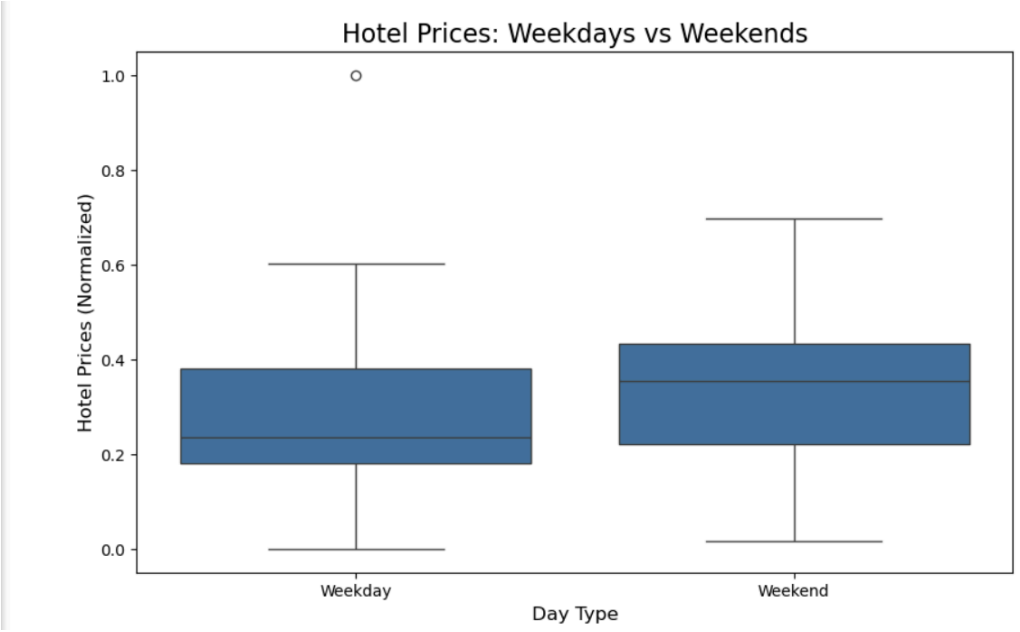
因此我們需要找出到底是什麼主要因素影響酒店價格趨勢，透過對應日期來新增特徵像是是否是周末、屬於什麼季節等等。

透過整合三筆資料集（各星級平均價格趨勢）來做分析：
分析季節和價格的關係:



季節可分辨出價格大致的波動。如圖所示春和秋天的平均趨勢比較高，而夏和冬則相反，可推斷出春和秋為當地的旅游旺季，夏和冬則爲旅游淡季。異常值代表酒店價格大幅漲調，也可能代表假期或旅游旺季。

分析平日/假日和價格的關係:



平日平均價格比假日平均價格來得低，可解釋爲什麼折線圖每個月會有約四次的峰值。

4.. Conclusions

我們以酒店價格和機票價格之間的關聯性作為研究主題，進而為旅游規劃提供參考。然而分析結果顯示酒店價格與機票價格的相關性較低，兩者之間並未呈現明顯的線性關係。酒店價格則更容易平/假日以及季節的影響，雖然本研究未能驗證酒店價格與機票價格的強相關性，但仍為未來的研究提供了思路。未來可以考慮以下方向進行深入探索：

- 細分市場分析：針對特定地區或特定時段進行更具針對性的分析，以發現潛在的局部相關性。
- 更多數據維度：加入其他影響旅遊旺季和淡季的因素（如天氣、節日、當地活動等）進行多變量分析，以便更全面地解釋價格波動。
- 跨地區研究：擴大研究範圍，分析不同地區和國家的情況，對比酒店與機票價格的關係是否存在區域性差異。

遇到的問題：

- 撰寫爬蟲耗費太多的時間
    - 動態網站在每個動作操作完后，html都有可能變動，需要合理安排操作方式
    - 動態網站逐一抓取需要穩定的網絡，每次試錯都會耗費很多時間
    - 正式抓取耗費的時間會隨著抓取目標的數量上升

- 確定主題不全面
    - 一開始沒考慮到後續建構模型部分
    - 抓取的情況不能直接斷定解讀，比如說價格上調下降都沒辦法百分百確定一個方向，背後需要考慮到太多層面（業者策略、人為因素等）

5. The contribution of each team member

賴鵬豐: 爬蟲、資料分析、撰寫書面報告
陳昱瑄: 資料預處理及分析、撰寫書面報告

報告影片連結:
https://www.youtube.com/watch?v=FA9FYOnWbPc