

旅遊小幫手

Python資料分析與機器學習應用_期末報告

組員：王柏詒、簡詩汶、賴欣妤、黃韻文、黃敏瑄

I. 旅遊小幫手 – 動機與背景

a. 最終成果呈現

各景點人流預測 | Airbnb 未來房價預測 | 景點分群分析

b. 為什麼是「旅遊」、「觀光」？

疫情過後，國內旅遊業開始逐漸復甦，同時有吸引許多國外旅客至台灣遊玩，雖然這將帶給整體經濟和觀光業正面的影響，卻同時導致許多觀光景點人潮過多和特定區域房價變高的負面情況。因此我們希望能夠利用預測各景點的人流、Airbnb 為未來房價和景點分群分析來提升國內景點旅遊品質以及增加個地區觀光收入，更重要的是期望能夠創造與提升台灣觀光品牌的國際形象。

c. 網路上已有許多相關觀光資訊，我們可以基於現有資訊創造出何種價值？

經過我們的研究發現，現有的數據存在兩個很大的問題：

1. 觀光統計資訊量過大

觀光局雖然提供很大量的相關資訊，年度研究報告高達 480 頁，但因缺乏視覺化統計圖表，資訊太多且龐雜導致民眾無法快速的了解目前的情況以及想要得到的訊息。另外，目前的趨勢分析僅依據季度與區域分析單向旅遊因子，無針對個別景點進行進一步的深入分析，且整體建議僅針對大方向，沒有提供個別地區 / 細緻化的發展建議，造成目前的數據無法提供一般民眾有用旅遊資訊與建議。

2. 少有以數據為主的客觀建議

目前網路上的研究多以趨勢、案例分析、消費者知覺態度、對景點的喜好、旅遊型態分析為主，較少有以實證資料為主的研究建議。而各方報導、論壇、KOL 影音與部落客文章分享的旅遊建議與評價多以「主觀感受」為內容，資訊略顯偏頗。

綜觀以上兩點現階段網路上資訊的缺失，可以明顯的觀察到，對於旅客而言，雖能夠取得許多相關的資料，但大多缺乏主規劃旅遊行程的客觀資料依據。

II. 專案內容介紹

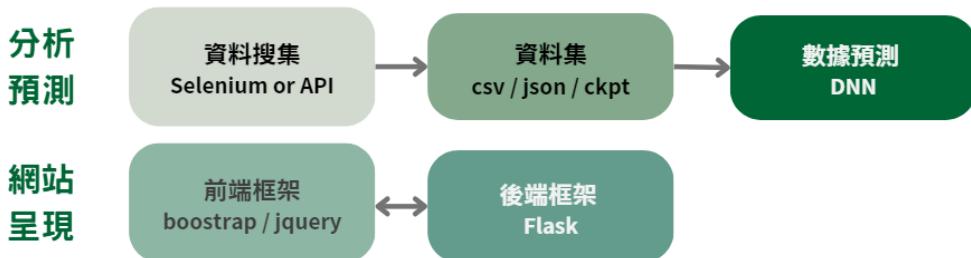
● 期中期末比較

專案 主題	期中規劃	期末調整	完成度
預測	<ul style="list-style-type: none">• 使用政府開放資料• 爬取天氣、平假日、房價• 預測未來每日人流量	<ul style="list-style-type: none">• 景點每月人流量預測• 旅館每日房價預測	90%

分析	<ul style="list-style-type: none"> 觀光局每月景點人流量 爬取天氣、平假日 景點分群、行銷建議 	無	100%
呈現	自架網站	無	100%

- 專案介紹

最後產出一個自架網站，主要的功能為房價查詢和景點分群景點分群分析。

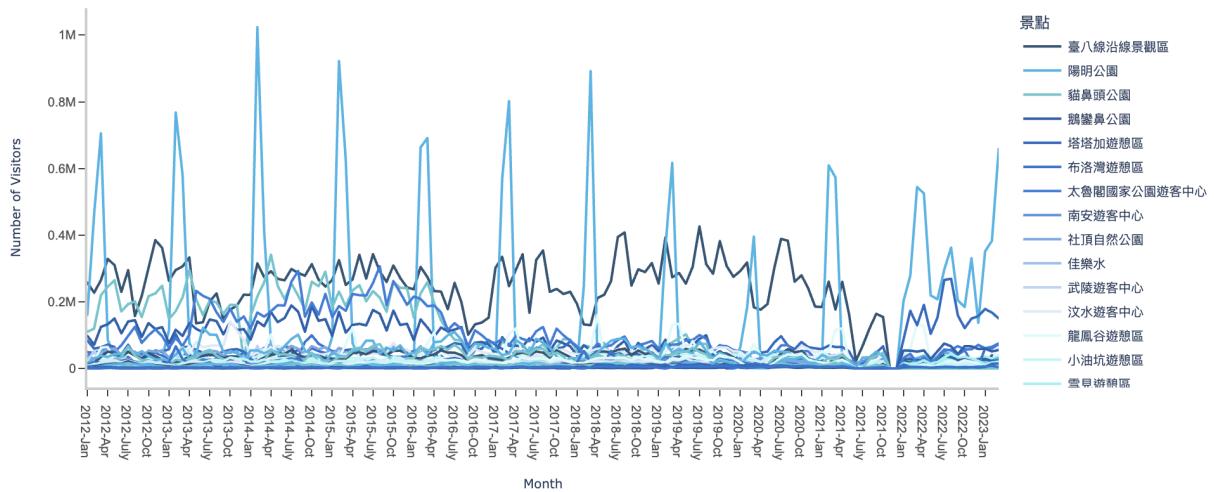


III. 探索式資料分析

我們使用交通部觀光局行政資訊網的國內主要觀光遊憩據點遊客人數月別統計資料，共 351 個景點(筆資料)，時間由 2012 年 1 月至 2023 年 3 月。在整理資料的過程中遇到的困難包含遊憩區更名、疫情、人流量紀錄為 0、增加遊憩區等。因此在建模前我們先對原始資料進行資料分析，以觀察其趨勢與思考可能的分群類型。在資料清理後使用觀光局的八個分類與各月人流量做圖，可以得到以下結果：

1. 國家公園

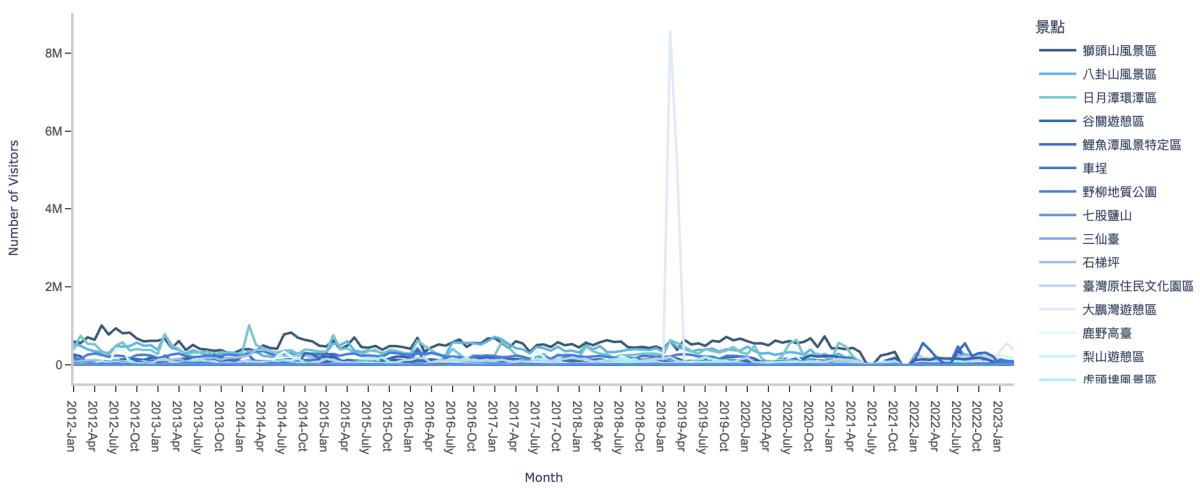
國家公園



由上圖可以看出陽明公園每年2至3月皆有人流量高峰，應為陽明山每年2至3月的櫻花季所帶來的人流。櫻花季僅在疫情在2020年初剛爆發時影響陽明山的人流，而在疫情後人流量有逐漸回升的趨勢。

2. 國家級風景特區

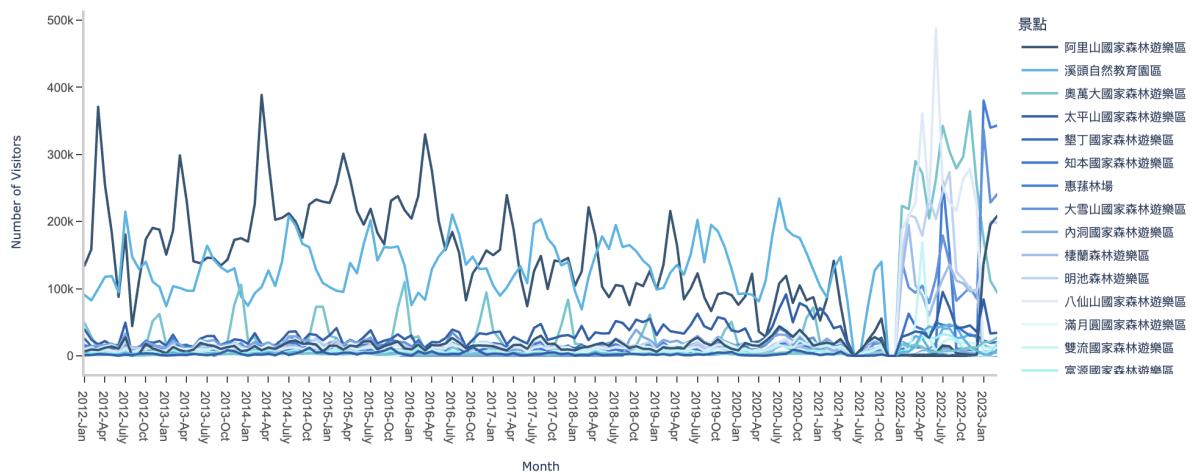
國家級風景特定區



國家級風景特定區中唯一的極端值為 2019 年大鵬灣遊憩區，其餘景區人流量皆屬穩定。2019 年台灣燈會在屏東大鵬灣，根據鵬管處歷年統計，燈會期間一天的人流量即超過 3 年的總和；由此可見燈會所帶來的人流量可引發地區觀光與關注。

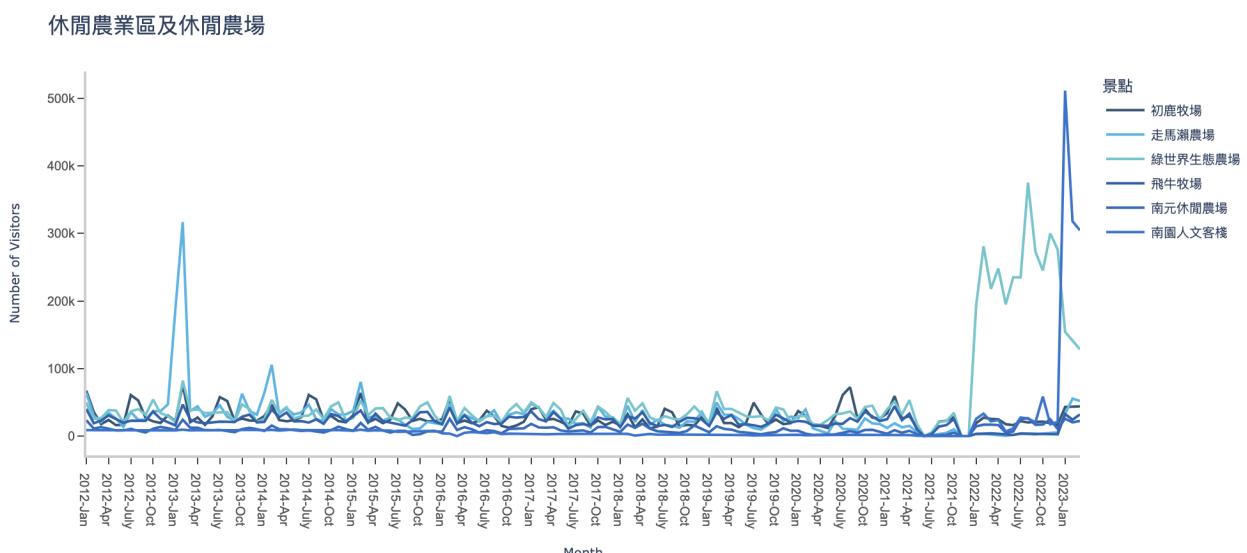
3. 森林遊樂區

森林遊樂區



從以上的圖表可以看出阿里山 2012-2019 每年 3 月人流量皆有高峰，因阿里山每年 3 月至 4 月初是櫻花季。而在疫情間可以看出 2021 年 7 月時人流量有成長的趨勢，此時為指揮中心公布「微解封」的時期。大幅成長發生在 2022 年 1 月後，由於國人無法出國旅遊，使得大量民眾走入山林。

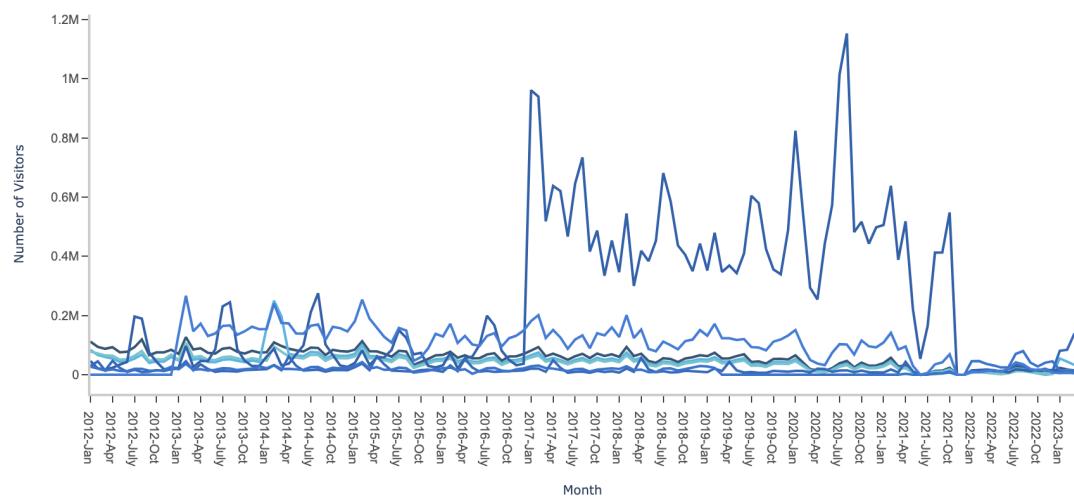
4. 休閒農業區及休閒農場



2013年1月走馬瀨農場出現人流高峰，了解過後發現為冬季熱氣球嘉年華，創造了逾30萬的人流。在後疫情時代，綠世界生態農場在2022年有顯著成長；而南園人文客棧則是在2022年底有50萬的人流，推測是全新的旅遊提案所帶來的人流成長。

5. 觀光地區

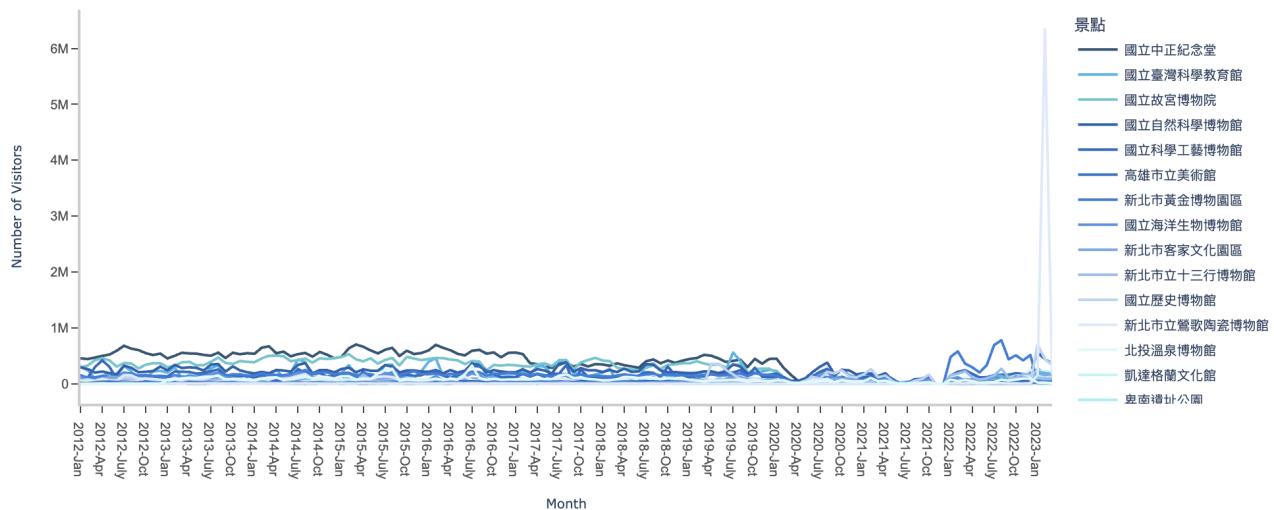
觀光地區



圖中麗寶樂園的人流在2017-2020的月平均都大於40萬，而在2017前可以發現每年7、8月時皆有人流高峰，判斷是暑假與季節性開園的馬拉灣使得人流增長。而在2017後持續性的高平均人流，推測是新增的園區與設施所帶來的人潮。

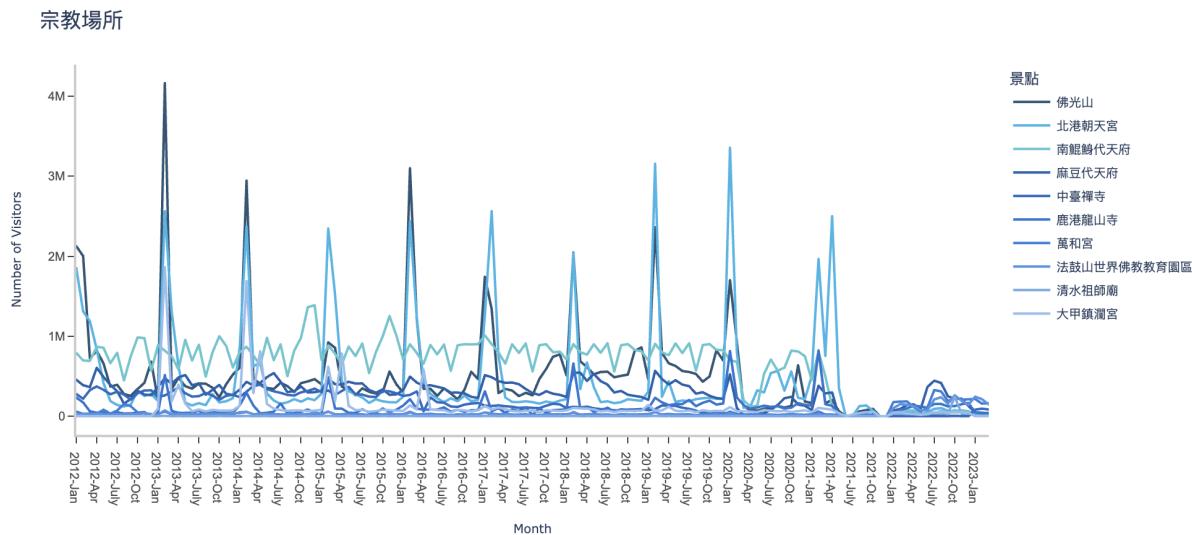
6. 博物館

博物館



各博物館在疫情爆發時人流量皆銳減，而在疫情後僅有新北市黃金博物園區有明顯人流成長。其中高峰為2022年8月，黃金博物館舉辦礦山藝術季，帶來自疫情爆發後前所未有的人潮。

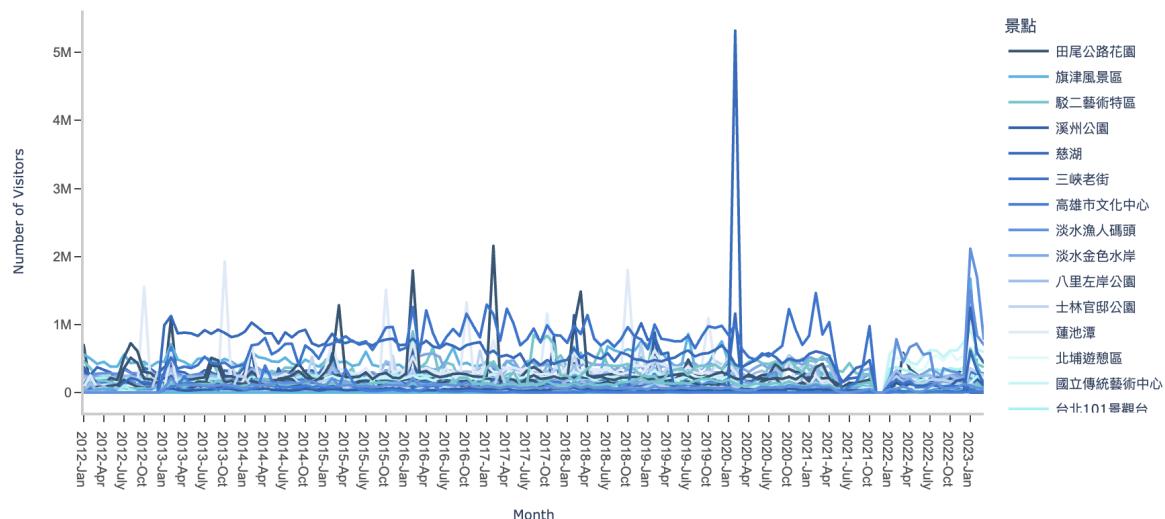
7. 宗教場所



宗教場所在每年2月皆有高峰，其中又以佛光山與北港朝天宮人潮最多。台灣的民俗之一為在過年時至廟裡拜拜，因次宗教場所在過年時有規律性的高峰。而在疫情後，僅有2022年7月的麻豆代天府有些微人潮成長。研究過後發現是南美館的《亞洲的地獄與幽魂》展覽，一併帶動了麻豆代天府的人潮。我們覺得宗教場所的人流是受疫情影響最嚴重的，且在後疫情時代人流也並無回歸的趨勢。

8. 其他

其他



由上圖可以看到蓮池潭每年10月皆有人群高峰，為每年舉辦的左營萬年季的地點。而在2014年由於南高雄發生氣爆事件，因此停辦一年，當年也就沒有人流高峰。左營萬年季有多年歷史，2021年因疫情停辦，雖然2022年復辦，但是人流並沒有回復到以往的峰值，其人流成長也不像以往有顯著的趨勢。

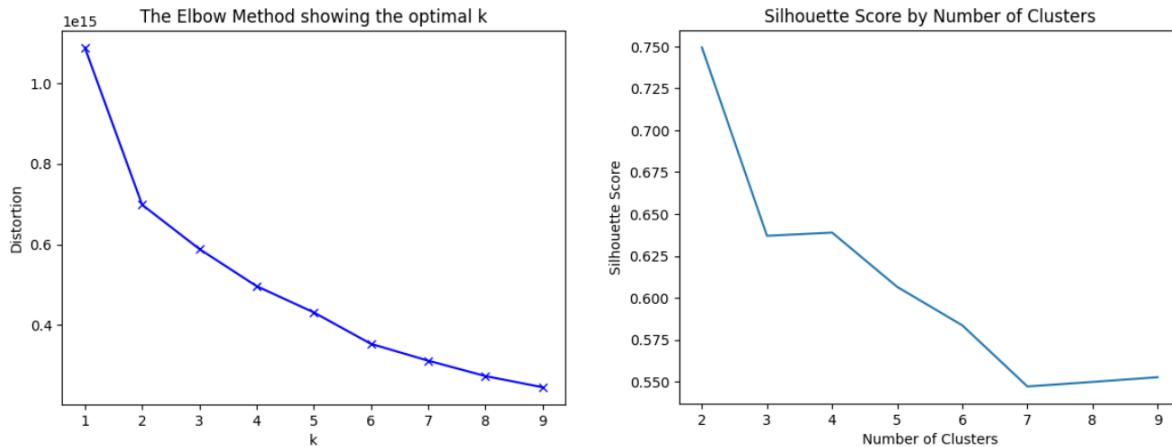
IV. 觀光景點分群模型 - 建模方式和調整方式

我們在景點類型的劃分上並未將「其他」更細分，原因乃這份資料是從交通部觀光局的網站中下載的開放式資料(open data)，考量公部門有其正式的分類考量，如國家公園、國家級風景特定區的層次、使用法規、規劃可能不同於博物館的管理方式，故我們無多做類型的劃分，以免混淆。

另外，有部分景點(暫且稱為「新興景點」，以下將會再詳述)是在 2020 年後才陸續有人流資料，我們推測是因為這些景點是近年來才開始發展而受重視的。在我們寫信予交通部觀光局尋求單日人流資料時，其於信中亦表示「人流資料是由觀光景點的主管機關所填報，計算方式可能由管理人員手動估算、碼錶自動計算等」，故於本次報告中，會有部分景點的資料在時間序列的圖上會呈現斷裂、缺失的情況，特此說明。

清理完資料後，我們接著想用非監督式學習的 K-means Clustering 方法找出各景點的分群類型，以了解是否有特定人流趨勢。使用不同的分群方法是為了瞭解不同分群方法之間是否形成不同的分群結果，但在本報告中，兩分群方法並無明顯分群結果，幾乎一樣。

我們首先要找最適合分群的 K 值，也就是 optimal K。第一個方法採用 Elbow Method，發現在 K=2 時有一個明顯的轉折，我們同樣用 Silhouette Score 尋找最適 K，同樣發現 K=2 時分數最高，並出現第二高峰在 K=4 之處。



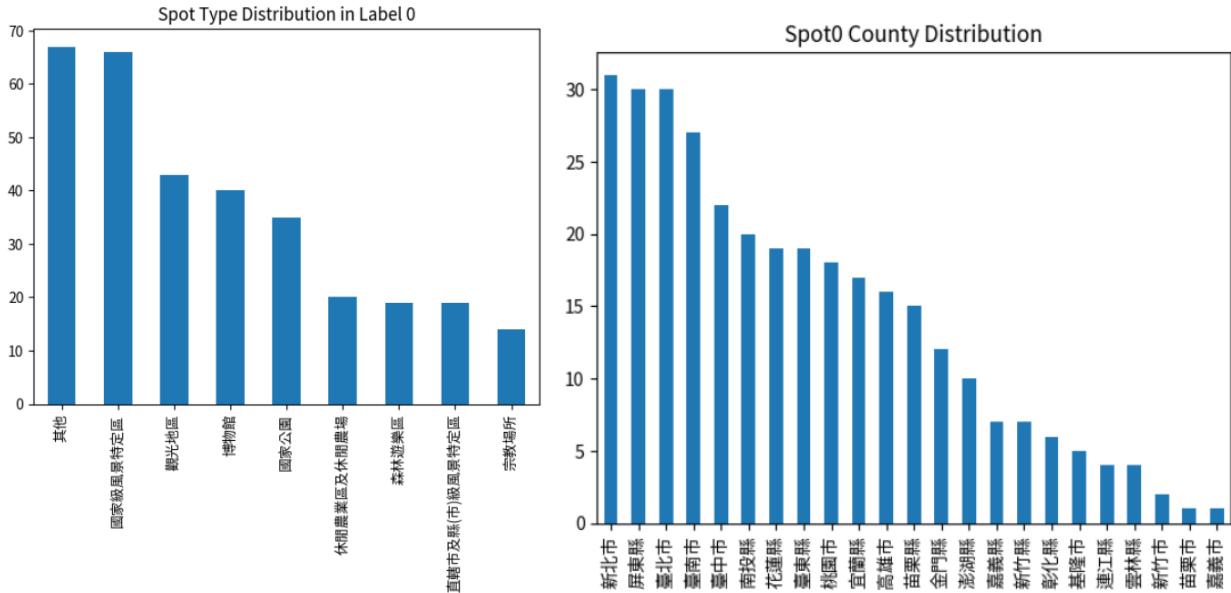
V. 分群模型 — 模型的結果/成果

在兩個方法顯示最適合分成兩群(K=2)的情況下，我們決定先把觀光景點的人流用 K-means 方法資料分成兩群看。

總體而言，在兩張圖中，我們都可以觀察到兩波下降趨勢，分別在 2020 年 1 至 3 月與 2021 年的 5 至 7 月，皆是受新冠肺炎嚴峻時期影響。當時曾發布三級警戒除了強制性造成人流減少，民眾也會受社會心理因素盡量避免外出，減少染疫的可能，故這兩波明顯的驟降反映在整體大環境的不可抗力。

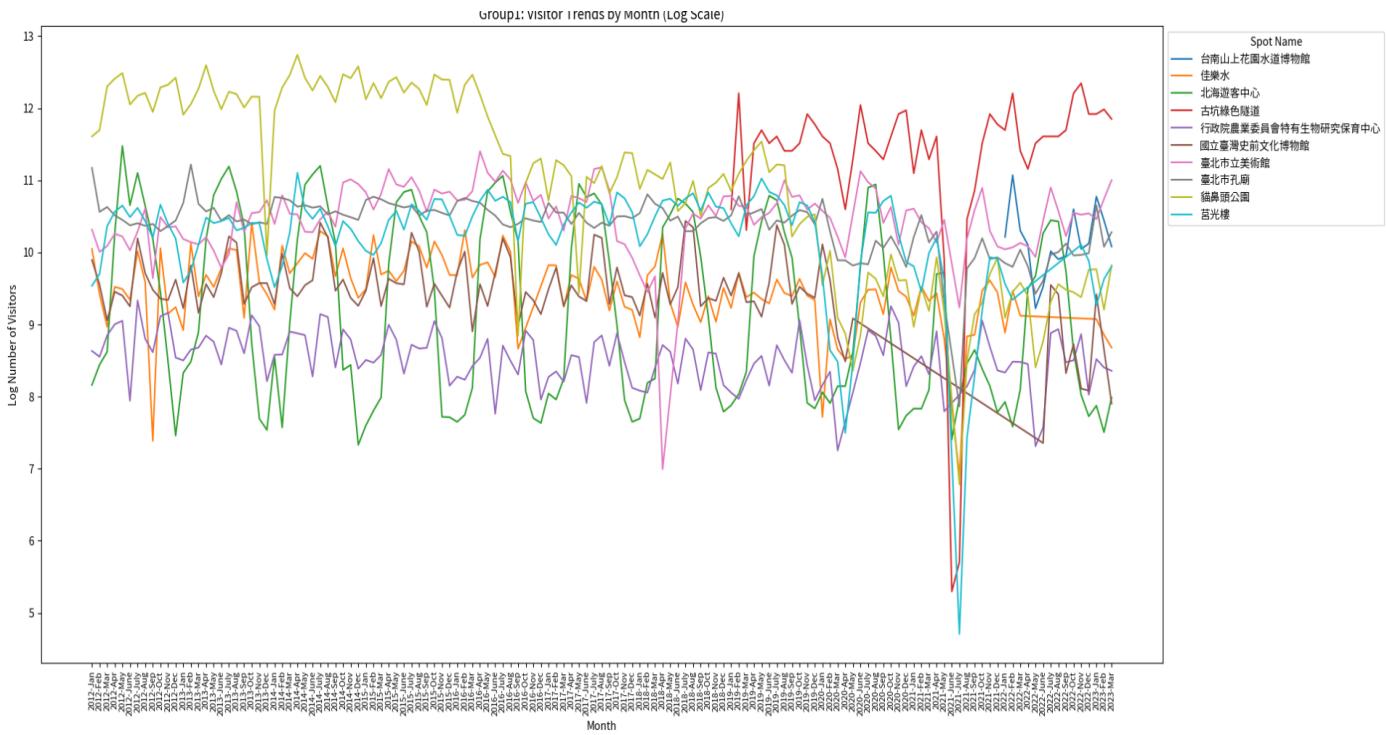
接著分群詳述：

第一群共 323 個景點被包覆，且景點類型多為「其他」，其次依序為：國家級風景特定區、觀光地區、博物館、國家公園、休閒農業區與休閒農場、森林遊樂區、直轄市及縣(市)級風景特頂區宗教場所。而新北市佔有分部最多的景點，第二至五名為屏東縣、臺北市、臺南市與臺中市，多為直轄市。

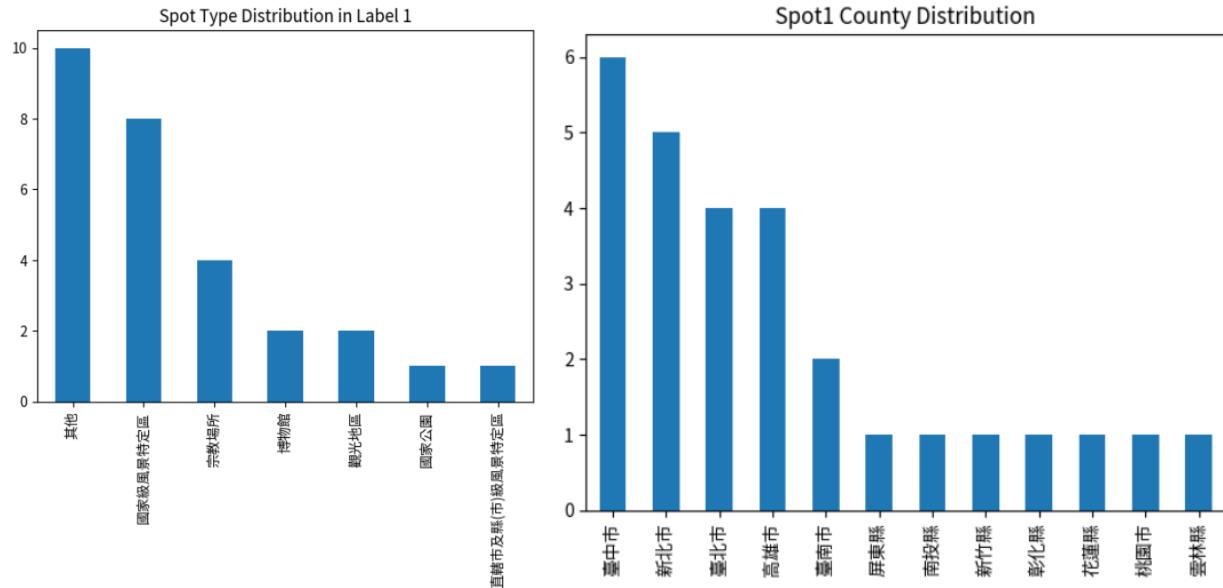


由於第一個分群有過多資料無法一一檢視，故我們取隨機十筆資料來觀察景點的趨勢。本次被挑中的景點有台南山上花園水道博物館、佳樂水、北海遊客中心、古坑綠色隧道、行政院農委會特有生物研究保育中心、國立台灣史前文化博物館、台北市立美術館、台北市孔廟、貓鼻頭公園與莒光樓。

當我們第一眼看見這張人流的時間序列圖時，可以直觀地感受到各景點有季節性(或夏或冬)的規律地波動變化，且時間範圍常跨度 3-6 個月。以位於澎湖的「北海遊客中心」為例，其每年的人流約在 3 月開始成長，至 7、8 月時會達到高峰，冬天人流明顯較少，其餘景點亦有其規律波動。



第二群共 28 個景點被囊括，景點類型亦多為「其他」，其次為國家級風景特定區、宗教場所博物館、觀光地區、國家公園與直轄市及縣市級風景特定區。臺中市擁有最多的第二群景點，其次五名為新北市、臺北市、高雄市與臺南市，全位在直轄市。



第二群的資料點雖少於第一群，惟考量 28 個資料點的視覺化過於繁雜與抽樣統一性，我們一樣取隨機十筆資料點觀察景點的人流趨勢。本次被抽取的景點有獅頭山風景區、麻豆代天府、東豐自行車綠廊及后豐鐵馬道、林口三井 Outlet、國立故宮博物院、國立自然科學博物館、臺中公園、草悟道、北港朝天宮與八里左岸公園。

與第一群不同的是，我們可以看出景點有著時間範圍較小(單月)的規律高峰特徵，並有較明顯的波峰波谷，白話一點稱「單月衝高」，且人流密度高之越集中每年初。仔細觀察特徵最明顯的「北港朝天宮」，1 至 4 月有著先上升後趨緩，並於 2 月達到最高峰的樣態。另外我們也可

以發現，第二群中的規律高峰多發生在 1 至 4 月，從景點名稱與其類型可管窺，第二群中的景點多適合於過年期間、家庭走春出遊拜拜的類型。

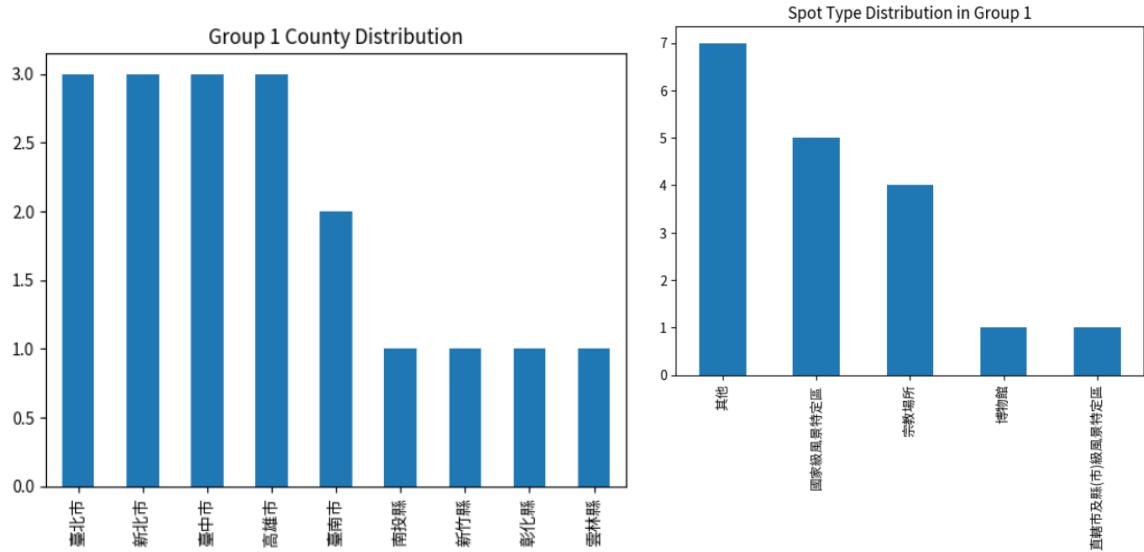


綜上總結，分兩群最大的不同在於高峰的時間點長短與人流高峰、低谷的截距大小：第一群通常以一個季節為單位，常見「夏季」暑假期間、人流截距較小；第二群以單月突然衝高為特徵，且發生期間多為「過年」冬春之際，人流截距大。

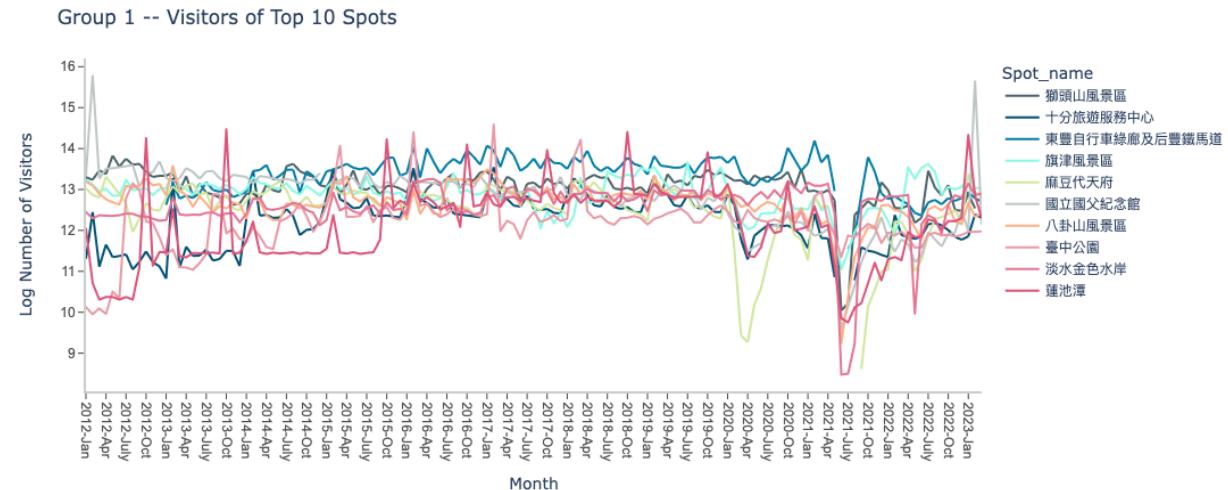
在概覽過最適分群 $K=2$ 的呈現結果後，我們肯定些許群內的不同，但想要進一步觀察更仔細的不同類型的觀光景點特徵，並提出我們的反思與建議。所以我們將資料分成分群分數次高的 $K=4$ 群，藉此以更小的微度瞭解臺灣人民對觀光的行為、型態與喜好。

1. 第一群：事件規律型

第一群共 18 個景點被分配進來，主要分部在直轄市，以臺北市居冠之位；景點類型以「其他」為主，其次依序為國家級風景特定區、宗教場所、博物館、直轄市及縣（市）級風景特定區。



在這一群中，我們發現這些景點已經有找到屬於自己的既有地位，可能會有固定的節日、活動吸引人潮，並且穩定規律地發展，在某些月份中人流會固定達到高峰。如十分旅遊服務中心會於每年的元宵節固定達到整年人流量最高，乃「放天燈」效應所致。這些景點已有長久培養的名聲與人潮，是典型的觀光景點。其中我們認為可更關注「宗教場所」，如前述的「北港朝天宮」、「麻豆代天府」都有其規律地繞境、朝拜活動，會吸引信眾前往。

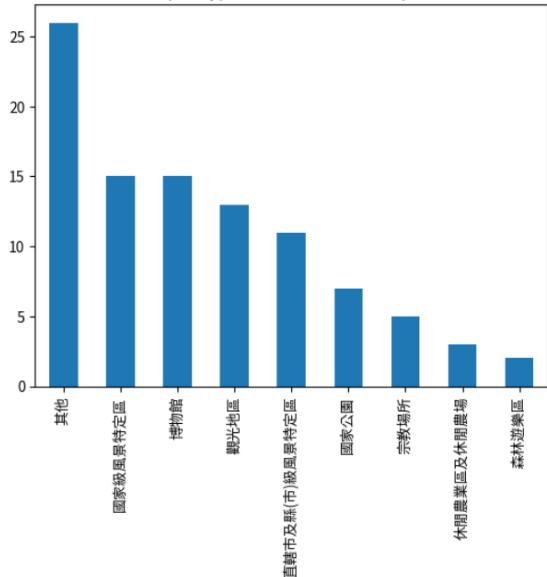


其中，值得注意的是 2023 年國父紀念館的「燈會」人潮效應十分顯著。

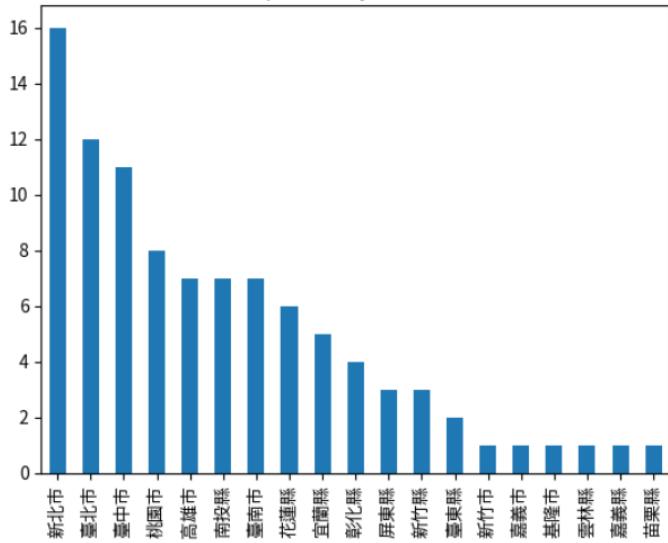
2. 第二群：季節穩定型

第二群共 97 個景點被囊括，亦多分布於直轄市，以新北市擁有最多屬於季節穩定型的景點。而此群的景點類型多為「其他」，其次依序為國家級風景特定區、博物館、觀光地區、直轄市及縣(市)級風景特定區、國家公園、宗教場所、休閒農業區及休閒農場與森林遊樂區。

Spot Type Distribution in Group 2



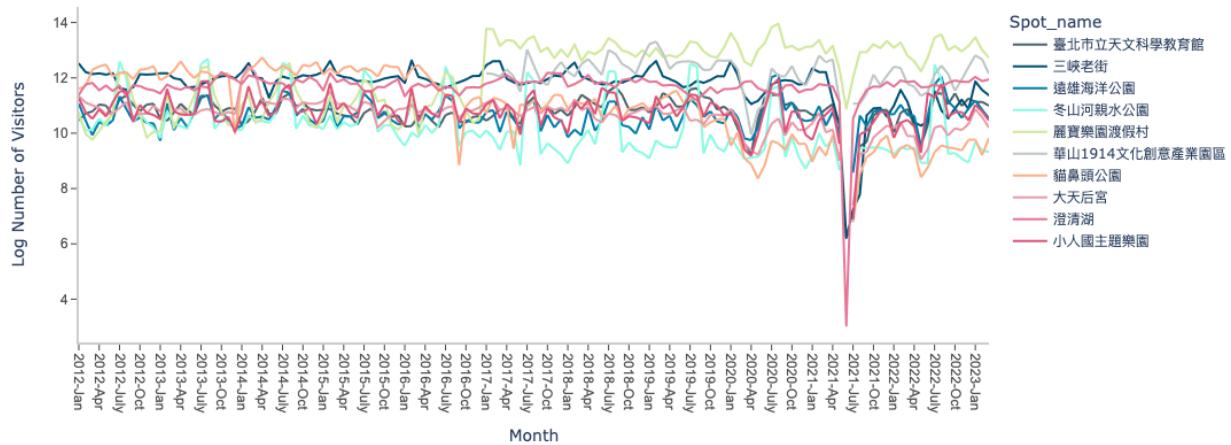
Group 2 County Distribution



在這一群中我們發現其人流截距相比前一群較小，並無特別的事件誘發人潮，而是以「季節」為界，穩定的波動。且我們進一步發現，景點人潮會分別在暑假、寒假達到高峰；有連續假期的月份(四月清明連假、雙十國慶連假)達到次高峰。我們推測這類型的景點以親子闔家出遊為主，甚至會配合小孩假期，規劃安排家庭觀光行程。典型的景點如遠雄海洋公園、小人國。

更進一步我們發現「麗寶樂園度假村」在 2017 年 1 月突然竄升，瞭解過後才知道，2017 年是臺中市第一年在麗寶樂園舉辦跨年晚會，並揭牌新蓋好的全台最大的摩天輪，成功吸引人潮。

Group 2 -- Visitors of Top 10 Spots



[三立新聞]

搭接駁車免塞！麗寶樂園跨年晚會交通資訊必看| 生活

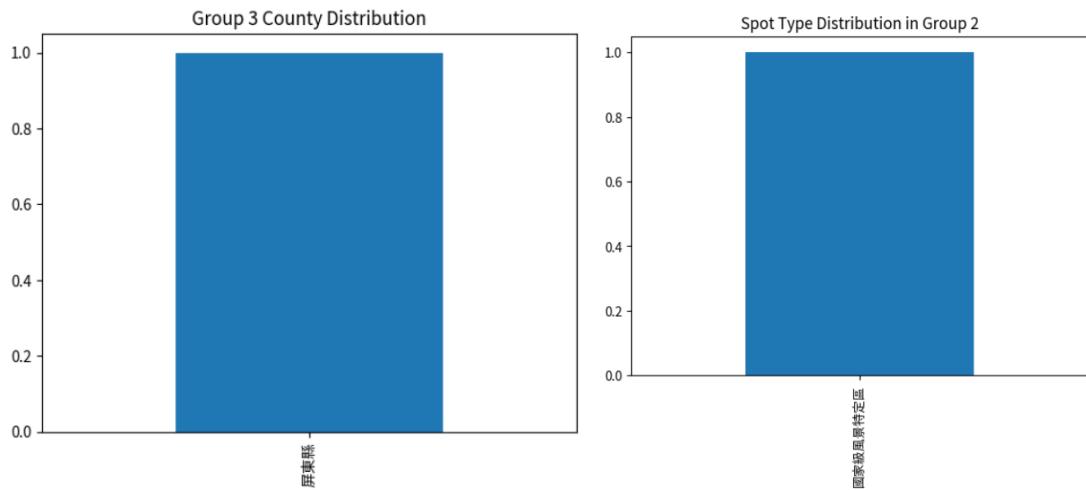
台中市政府舉辦「2017花Young台中」跨年晚會將以雙主場登場，麗寶樂園第二停車場部分，大甲警分局為避免塞車困擾，實施交通疏導，呼籲民眾進入會場搭乘大眾交通工具。



2016年12月31日

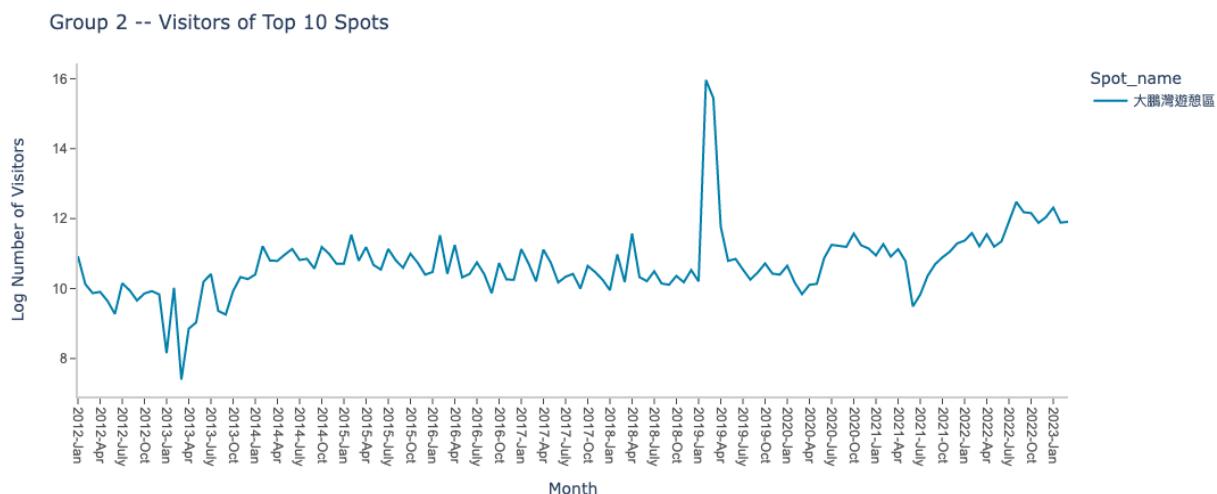
3. 第三群：大鵬灣特殊景點個案(單事件觸發類型)

大鵬灣位於屏東縣，並為國家級風景特定區。



被分出來的第三群大鵬灣是一個很特別的案例，在我們探索資料、分多群(K)觀察時，不管怎麼分，大鵬灣永遠特立獨行地被分出來。可以觀察到相對其他景點群，大鵬灣有其特色。在2019年之前，觀光人流量一直維持在平均每個月5-6萬人次，僅有小幅度的人流波動，而相對高峰出現在每年冬末春初。但在2019年1-3月突然有唯一的高峰出現，2019年4月後便持續維持高峰前的穩定5-6萬人流，直到近兩年新冠肺炎疫情後才有慢慢成長的趨勢。

我們細就2019年1-3月的高峰瞭解其背後的原因，發現2019年的臺灣燈會，就是在屏東縣舉辦，而大鵬灣是當年的主燈區所在之處，故有一瞬間湧入人潮的高峰態樣，顯見燈會的影響力在臺灣十分巨大，是吸引人潮的重要因子。



LTN 自由時報

台灣燈會》人潮擠爆！大鵬灣今突破81萬人 首度延後關燈

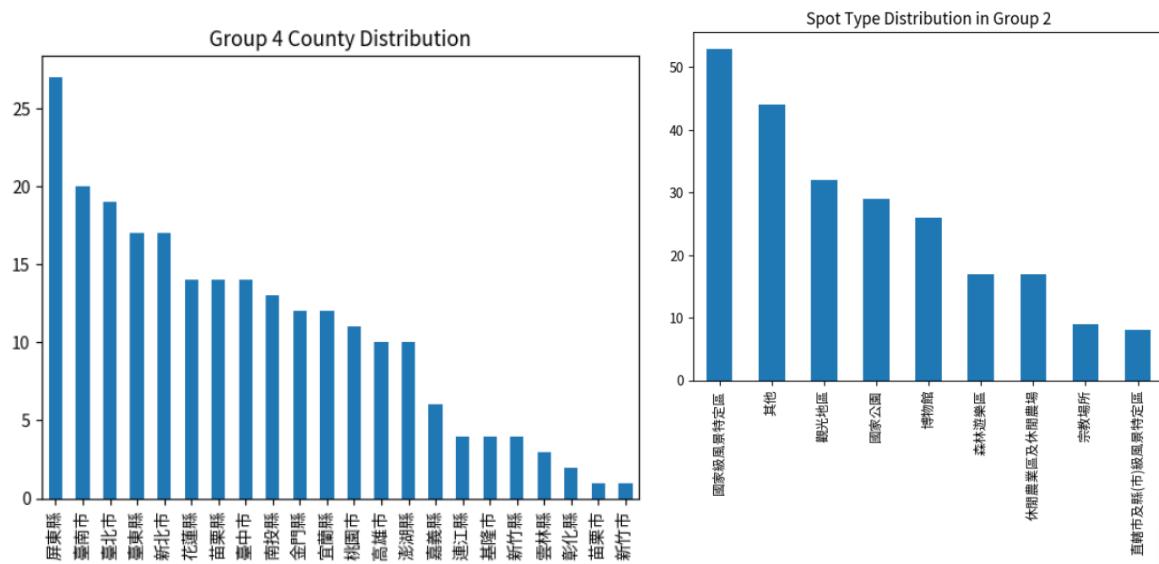
台灣燈會迎來第一個周末，雖然是補班日，入園人數卻創下單日新高，一天之內就湧入81萬人次。（記者陳彥廷翻攝）。2019/02/23 22:50. [記者陳彥廷 / 屏東報導] 2019台灣…



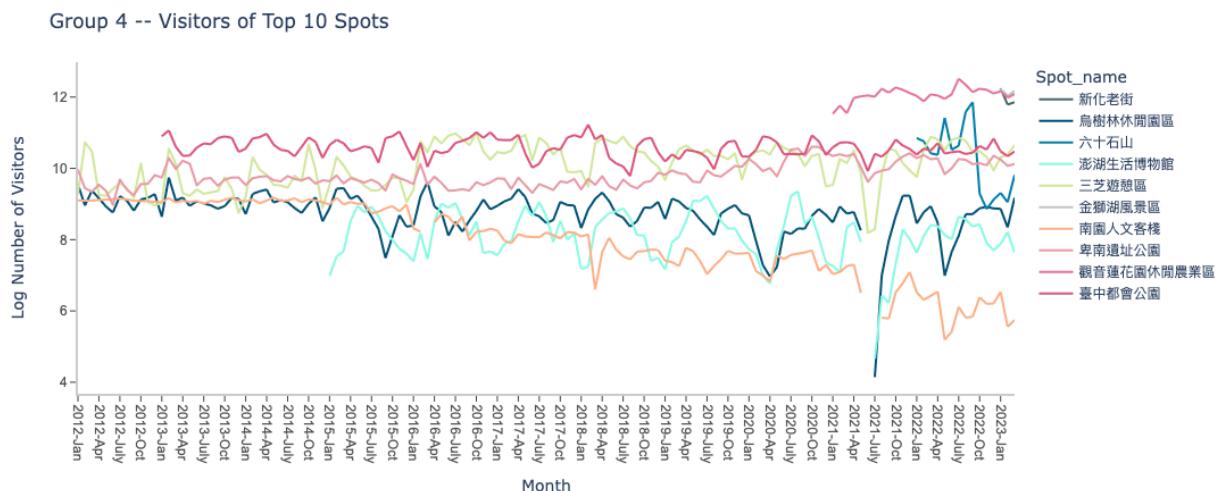
2019年2月23日

4. 第四群：潛力發展景點

第四群共有 235 筆資料景點被納入，其中以屏東縣擁有最多的觀光景點，其次依序為臺南市、臺北市、臺東縣與新北市。觀光景點類型則以「國家級風景特定區」最多，接著為、其他類型、觀光地區、國家公園、休閒農業區及休閒農場、宗角場所與直轄市及縣(市)級風景特定區。



我們發現許多近三年才開始統計人流資料的景點都被歸類進了第四群；景點之間並無特別顯著的月份人流高峰或相似特徵，相對平穩，無大起，無大落。從景點名稱看來，也無直觀地既定景點印象，推測第四群中的景點尚未找到屬於自己的定位與既有的活動、行銷、吸引人的手段。但隨著人流相對平穩，或許有機會、有潛力，於疫情之後發展成新興的觀光景點。



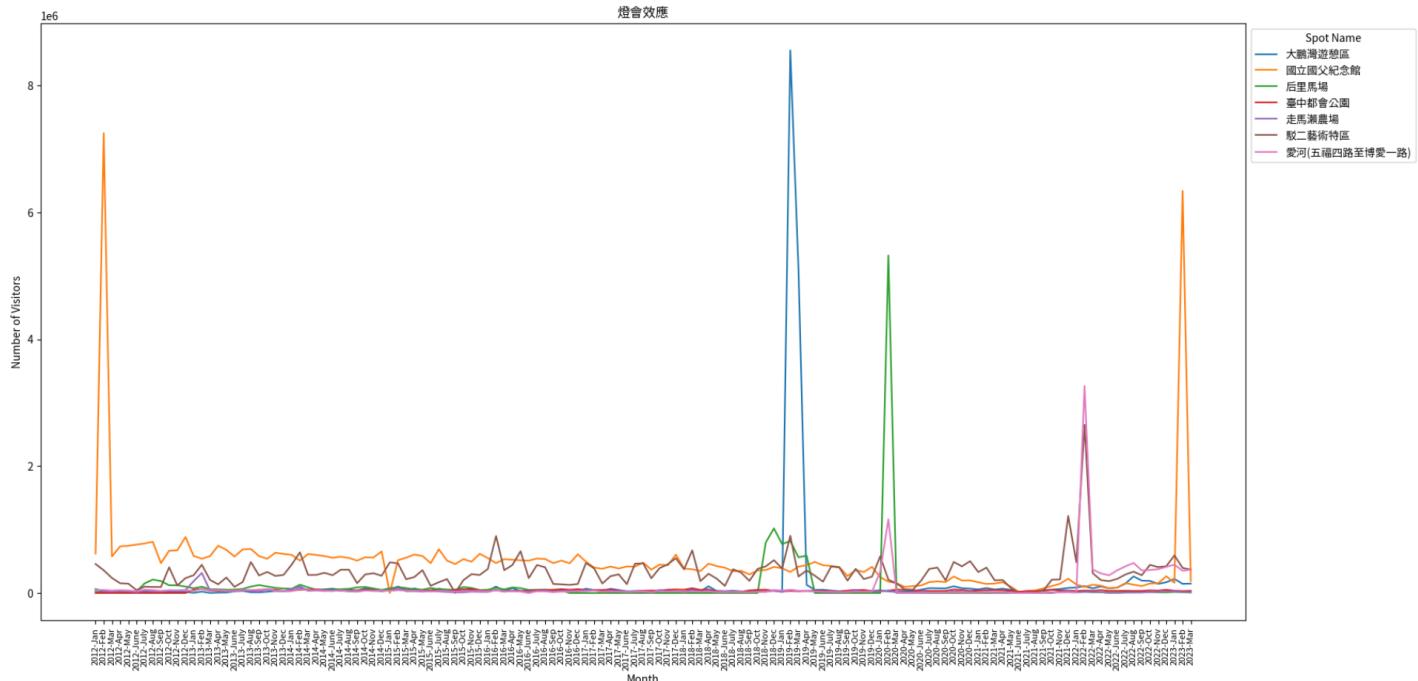
上述四群不同類型的觀光景點的分析外，我們也發現每到「春節期間」，必有顯著的人流成長，比起其他連續假期變化更加劇烈。推測因春節期間除了傳統的文化習俗如拜年、走春外，「長假期」也是一個影響觀光出遊的因素。

VI. Insight 總歸納與反思

a. 「燈會」作為單事件人流高峰觸發點

本次分群人流趨勢圖最讓我們驚訝的是「燈會」在臺灣觀光景點的影響力十分巨大。因為小組成員本身不是會去看燈會的人，只是在今年特別感受到臺北市的人潮眾多；沒想到攤開曾經舉辦過燈會縣市的主燈區原始資料，發現燈會確實是人潮的流量密碼。2019

年屏東縣燈會還有「賽車」的加成作用，因此人流量居最高；其次為 2020 年臺中燈會、2022 高雄燈會與今年 2023 的臺北燈會。不過我們認為今年的臺北燈會人流有低估的可能，因為今年燈會的範圍還包括忠孝東路到忠孝敦化站、松菸文創園區、臺北 101，極有可能被稀釋掉。但撇除此疑點，燈會對一個縣市、一個觀光景點的人潮、錢潮影響確實大。



b. 針對景點四群的反思建議

分群類型	特徵	反思建議
第一群： 事件規律 型	<ul style="list-style-type: none"> 各景點有季節性（或夏或冬）的規律地波動變化，且時間範圍常跨度 3-6 個月。 以位於澎湖的「北海遊客中心」為例，其每年的人流約在 3 月開始成長，至 7、8 月時會達到高峰，冬天人流明顯較少，其餘景點亦有其規律波動。 	<ul style="list-style-type: none"> 單事件／單月高峰的景點類型，可注意大量人流短期湧入的因應措施，如周邊交通配置、流動廁所、糧食量能等。
第二群： 季節穩定 型	<ul style="list-style-type: none"> 景點有著時間範圍較小（單月）的規律高峰特徵，並有較明顯的波峰波谷，白話一點稱「單月衝高」，且人流密度高之越多人集中在每年初。 規律高峰多發生在 1 至 4 月，從景點名稱與其類型可管窺，景點多適合於過年期間、家庭走春出遊拜拜的類型。 	<ul style="list-style-type: none"> 季節穩定型有明顯的假期淡旺季，多數景點周邊的商圈住宿也會因此而調整。 我們認為這類型的景點可以適時加入新元素吸引不同的旅客，打破既有的框架旅遊模式，發展不僅適合親子出遊，也適合獨旅、深度旅遊的路線。

第三群： 特殊景點 個案	<ul style="list-style-type: none"> 在我們探索資料、分多群觀察時，不管怎麼分，大鵬灣永遠特立獨行地被分出來 其背後的原因是：2019 年的臺灣燈會，就是在屏東縣舉辦，而大鵬灣是當年的主燈區所在之處，故有一瞬間湧入人潮的高峰態樣，顯見燈會的影響力在臺灣十分巨大，是吸引人潮的重要因子。 	<p>「燈會」是一個很有影響力的觀光活動，各縣市、地區若有機會，可透過舉辦燈會進行縣市的總體行銷。</p>
第四群： 潛力發展 景點	<ul style="list-style-type: none"> 景點之間並無特別顯著的月份人流高峰或相似特徵，相對平穩，無大起，無大落。 從景點名稱看來，也無直觀地既定景點印象，推測第四群中的景點尚未找到屬於自己的定位與既有的活動、行銷、吸引人的手段。 隨著人流相對平穩，或許有機會、有潛力，於疫情之後發展成新興的觀光景點。 	<p>可針對具有潛力的景點提供更多行銷、活動策略方面的幫助，使之能夠更好地發展成成熟的新景點，找到定位與路線後，深耕。</p>

VII. 預測模型(簡)

a. 景點人流量預測模型

資料來源：交通部觀光局

資料處理 (只列特別的部分): **label / one-hot encoding & pandas.melt**

label / one-hot encoding			pandas.melt		
Spot_code	County_encode	Spot_type_encode	population	Year	month
0	13	3	12187.0	2012	1
1	13	3	2597.0	2012	1
2	13	3	161000.0	2012	1
3	13	3	17047.0	2012	1
4	13	3	21201.0	2012	1

模型架構: DNN模型

```

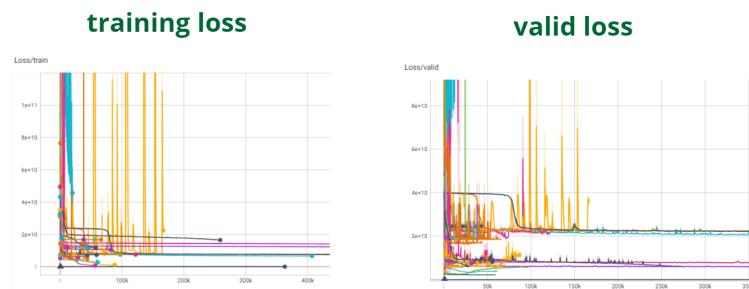
class My_spot_Model(nn.Module):
    def __init__(self, input_dim):
        super(My_spot_Model, self).__init__()
        self.layers = nn.Sequential(
            nn.Linear(input_dim, 128),
            nn.LeakyReLU(),
            nn.Linear(128, 48),
            nn.LeakyReLU(),
            nn.Linear(48, 16),
            nn.LeakyReLU(),
            nn.Linear(16, 24),
            nn.LeakyReLU(),
            nn.Linear(24, 8),
            nn.LeakyReLU(),
            nn.Linear(8, 1),
            nn.ReLU(),
        )

```

可調參數:

- `valid_ratio: 0.2`
- `n_epochs: 3000`
- `batch_size: 256`
- `learning_rate: 1e-3`
- `early_stop: 300`

各種模型訓練過程:



模型評估:

```

from sklearn.metrics import mean_squared_error
mean_squared_error(test_data[:, -1], preds, squared=False)

```

61861.75

不是特別好的模型，可能原因為缺失值過多，加上疫情期間多處人流量劇烈受影響，導致模型學到特別的規則，而疫情後又爆發一段旅遊潮，因此此時預測下個月的人流量確實有瑕疵。期待未來待疫情完全結束，可以將此時的資料標為特殊值，增進模型的能力。

b. Airbnb 預測模型

資料來源: 每日 Airbnb 房價爬蟲 | 每日天氣預報爬蟲 | 捷運站點資料蒐集

```

def scrap_this_page(dataframe, url, checkin, checkout, adults, children, infants, pets):
    # Create selector
    html = requests.get(url).content
    sel = Selector(text=html)

    hotels = sel.css('div.c4mnd7w')

    # Select the first announcement from the previous list of 20
    for i in range(len(hotels)):
        hotel = hotels[i]
        # Get main information
        title = hotel.css('div[data-testid="listing-card-title"] ::text').extract_first()
        price = hotel.css('span.tyxjp1 ::text').extract_first()
        if price == None:
            ori_price = hotel.css('span.lks8gb ::text').extract_first()
            dis_price = hotel.css('span.ly74zjx ::text').extract_first()
        else:
            ori_price = price
            dis_price = price
        rating = hotel.css('span.cldllyb ::text').extract_first()
        url = hotel.css('a.bn2bl2p ::attr(href)').extract_first()

        # Add data to the dataframe
        dataframe.loc[i] = [title, ori_price, dis_price, rating, checkin, checkout, adults, children, infants, pets, timestamp, main_url]
    return dataframe

def to_next_page(sel, page_1):
    next_page = sel.css('a.black0h ::attr(href)').extract()[page_1]
    return f'{main_url}{next_page}'

```

爬此頁

加入 dataframe

進入下一页

```

def scrap_weather():
    import requests
    import pandas as pd
    from datetime import datetime

    df_pre = pd.DataFrame(columns=['date','region',
                                    'rain_0','rain_1','rain_2','rain_3','rain_4','rain_5','rain_6',
                                    'temp_0','temp_1','temp_2','temp_3','temp_4','temp_5','temp_6',
                                    'humi_0','humi_1','humi_2','humi_3','humi_4','humi_5','humi_6'])

    today = datetime.today().date()
    url = 'https://opendata.cwb.gov.tw/api/v1/rest/datastore/F-D0047-063?Authorization=CWB-8CB6'
    date = requests.get(url)
    data_json = date.json()
    location = data_json['records'][0]['locations'][0]['location']

    for i in range(len(location)):
        rain = location[i]['weatherElement'][0]
        temperature = location[i]['weatherElement'][1]
        humidity = location[i]['weatherElement'][2]

```

爬取當天-六天後的
雨量、氣溫、濕度

資料從氣象局
開放的 api 撷取

資料處理 (只列特別的部分): Replace string & Normalization

Replace string with regex

```

df = pd.read_csv('airbnb_taipei_spec_date_clean_copy.csv', encoding='unicode_escape')
df = df.replace({"Ã": "", "ƒ": "", "Ã": "", ":"}, {"TWD": "", ":"}, {"": "\xa0": "", "\$": ""}, regex=True)

```

Normalization

```

from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
train.iloc[:, :-1] = scaler.fit_transform(train.iloc[:, :-1])
test.iloc[:, :-1] = scaler.transform(test.iloc[:, :-1])

```

模型架構: DNN模型

```

class My_airbnb_Model(nn.Module):
    def __init__(self, input_dim):
        super(My_airbnb_Model, self).__init__()
        self.layers = nn.Sequential(
            nn.Linear(input_dim, 128),
            nn.LeakyReLU(),
            nn.Linear(128, 48),
            nn.LeakyReLU(),
            nn.Linear(48, 16),
            nn.LeakyReLU(),
            nn.Linear(16, 24),
            nn.LeakyReLU(),
            nn.Linear(24, 8),
            nn.LeakyReLU(),
            nn.Linear(8, 1)
        )

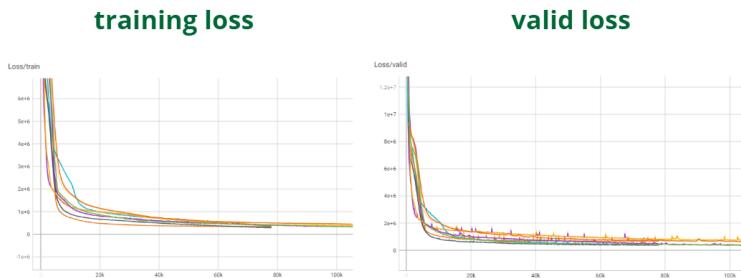
```

可調參數:

- valid_ratio: 0.2
- n_epochs: 3000
- batch_size: 256
- learning_rate: 1e-3

- early_stop: 300

各種模型訓練過程:



模型評估:

```
from sklearn.metrics import mean_squared_error
mean_squared_error(test_data[:, -1], preds, squared=False)
```

549.04

由於最多可以選到16位大人，金額可以高達10萬，因此我們認為549的MSE(squared=False)算是還不錯的模型。

VIII. 遇到的困難和解決方法

我們原先在期中報告中以「分析和預測觀光景點的人流」為主題，以下將詳細說明我們最終在期末報告中稍作改變的原因、遇到的困難和解決辦法。

遇到的困難	解決辦法與改變
觀光景點資料沒有日資料，難以做我們原先的分析	觀光景點人流改成使用月資料，增加 Airbnb 房價相關資料分析與預測。

IX. 未來可發展: 日資料的單日預測(可以呼應原來做不出來的期中提案主題)

回應旅遊型態分析主題的命題初衷，起初覺察的痛點是每當連假、例假日時節觀光旅遊景點的人潮總是大量湧現，然而，總是需要到旅行當天查看 google 忙碌程度、或是抵達該景點時才能掌握當地人流量多寡，旅遊景點人潮無法事先預期的特性導致出遊時難以做出最適合自己的判斷，與此同時也因大量人潮的湧入降低了景點的旅遊品質，然而，目前政府機構尚無統整性的每日人流資料可作分析，不過，近年來政府開始著力於電信信令人口統計資料的應用，這份數據源自各大電信公司所蒐集到的手機訊號數據，由於當前的台灣社會幾乎人手一台手機且習慣隨身攜帶，因此用此數據作為人流量統計相當準確，近期政府單位曾開放短期的每日人流數據作為競賽使用，因此，若未來有機會可以進一步申請這份資料用作分析即能達成最初目標，還能整合房價預測的功能，讓遊客能夠更有依據的挑選出最適合自己的出遊景點與下榻旅館。

不僅如此，單日人流資料也能與景點分析相互作用，觀察十年來更細緻的人流量波動，找出更深入的影響因素，搭配實時的預測人流量，為各地的景點擬定更精準的行銷策略。

X. 心得感想

回顧整個學期的學習，我們深刻體會到了機器學習在現代科技中的重要性和應用廣泛性。透過這門課，獲得了許多寶貴的知識和技能，並且能夠運用所學來解決現實世界的問題。在課程中，我們學習了多種機器學習模型，包括監督式學習、非監督式學習和強化學習等。而通過這次的報告，我們更深入地了解了這些模型的運作原理、優缺點以及在不同場景中的應用。例如，我們學習了線性回歸、邏輯回歸和支持向量機等監督式學習模型，這些模型在預測、分類和回歸等任務中具有廣泛的應用。此外，我們還學習了聚類算法如 K-means 和層次聚類，這次報告當中也實際應用了當中的幾個模型，讓我們更加深入地了解到機器學習的全貌，並且能夠更自信地應對實際問題。這次實作的經驗不僅加深了我們對模型的理解，還讓我們意識到了機器學習在解決現實問題中的潛力。通過實際的案例和數據分析，我們意識到機器學習可以應用於各種領域，對於提高效率和創造價值有著巨大的影響力。總的來說，這學期的機器學習課程是一次非常寶貴的學習經歷。不僅學到了豐富的知識，還獲得了實際的經驗，讓我們更加深入地了解了機器學習的模型和應用。這將對於不管是未來的學習和或是職業發展都有著長遠的影響。相信機器學習將在未來繼續發展和創新，並為我們帶來更多驚喜和機遇。

XI. Appendix — website

- Airbnb 房價預測

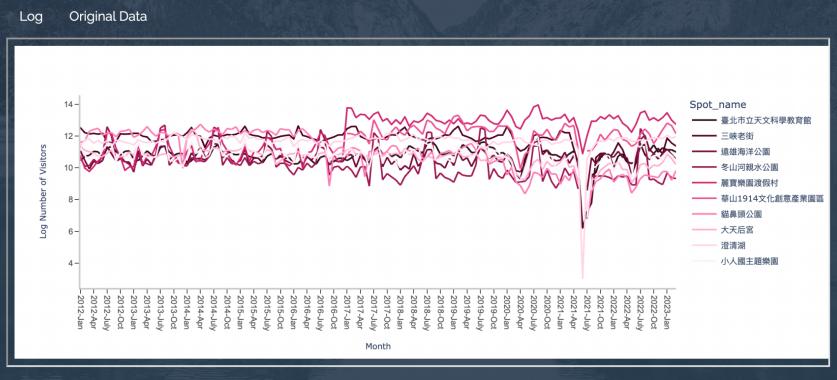


- 人流分群分析



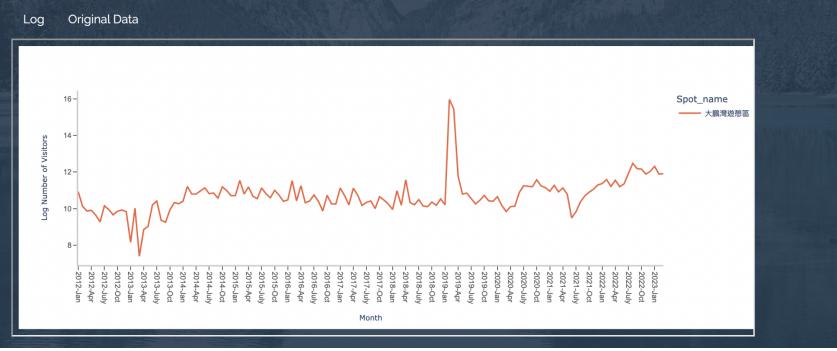
Group 2 – 季節穩定型

- 景點有著明顯較小（單月）的規律高峰特徵，並有較明顯的波峰波谷，白話一點稱「單月衝高」，且人流密度高之越多人集中在每年初。
- 規律高峰多發生在 3 至 4 月，從景點名稱與其類型可管窺，景點多適合於過年期間、家庭走春出遊拜拜的類型。



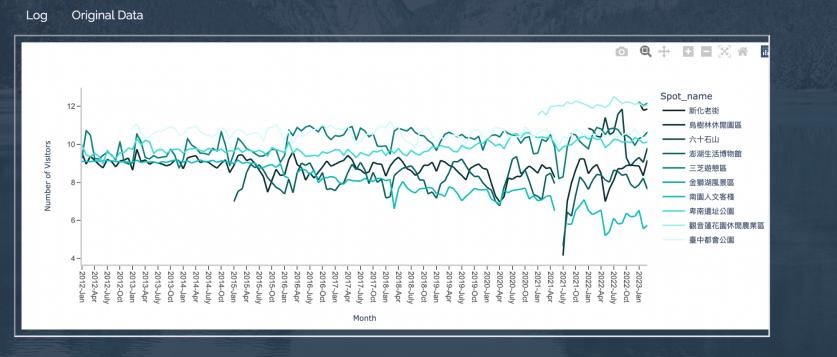
Group 3 – 大鵬灣特殊景點個案（單事件觸發類型）

- 在我們探索資料、多分群觀察時，不管怎麼分，大鵬灣永遠特立獨行地被分出來
- 其背後的原因是：2019 年的臺灣燈會，就是在屏東縣舉辦，而大鵬灣是當年的主燈區所在之處，故有一瞬間湧入人潮的高峰態樣，顯見燈會的影響力在臺灣十分巨大，是吸引人潮的重要因子。



Group 4 – 潛力發展景點

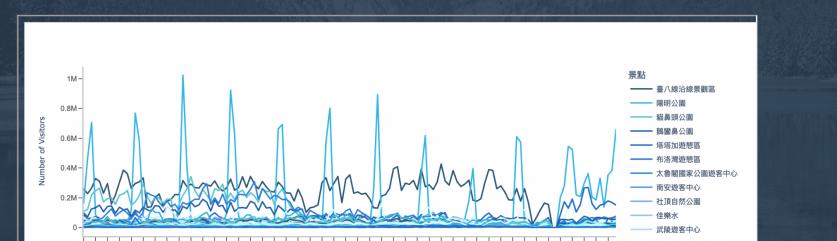
- 景點之間並無特別顯著的月份人流高峰或相似特徵，相對平穩，無大起，無大落。
- 從景點名稱無求，也無直觀地既定景點印象，推測第四群中的景點尚未找到屬於自己的定位與既有的活動、行銷、吸引人的手段。
- 隨著人流相對平穩，或許有機會、有潛力，於疫情之後發展成新興的觀光景點。



● 人流資料分析

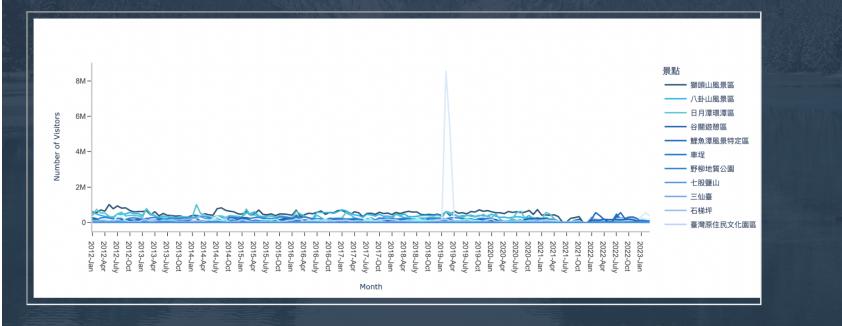
國家公園

由圖中可以看出陽明公園每年 2 至 3 月皆有人流量高峰，應為陽明山每年 2 至 3 月的櫻花季所帶來的人流。櫻花季僅在疫情在 2020 年初剛爆發時影響陽明山的人流，而在疫情後人流量有逐漸回升的趨勢。



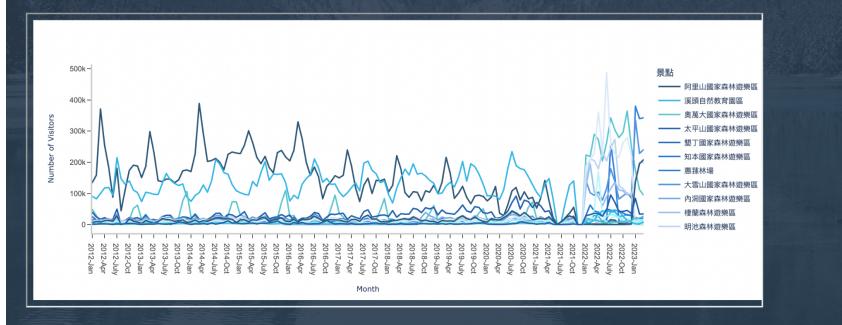
國家級風景特定區

國家級風景特定區中唯一極端值為 2019 年大鵬灣遊憩區，其餘景區人流量皆屬穩定。2019 年台灣燈會在屏東大鵬灣，根據謄管處歷年統計，燈會期間一天的人流量即超過 3 年的總和；由此可見燈會所帶來的人流量可引發地區觀光與關注。



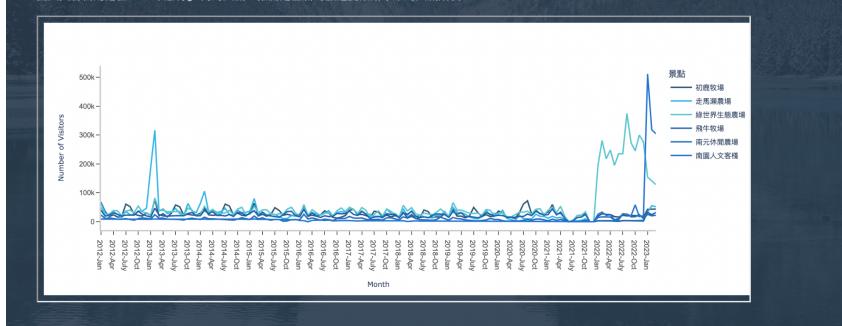
森林遊樂區

從圖表中可以看出阿里山 2012~2019 每年 3 月人流皆有高峰，因阿里山每年 3 月至 4 月初是櫻花季。而在疫情間可以看出 2021 年 7 月時人流有成長的趨勢，此時為指揮中心公布「微解封」的時期。大幅成長發生在 2022 年 1 月後，由於國人無法出國旅遊，使得大量民眾走入山林。



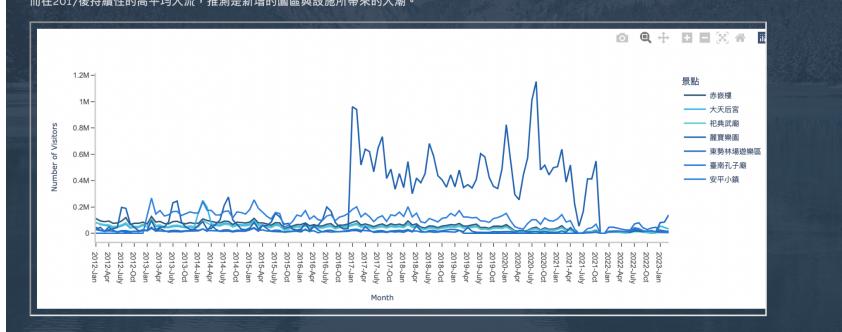
休閒農業區與休閒農場

2013 年 1 月走馬瀨農場出現人流高峰，了解過後發現為冬季熱氣球嘉年華，創造了逾 30 萬的人流。在後疫情時代，綠世界生態農場在 2022 年有顯著成長；而南園人文客棧則是在 2022 年底有 50 萬的人流，推測是全新的旅遊提案所帶來的人流成長。



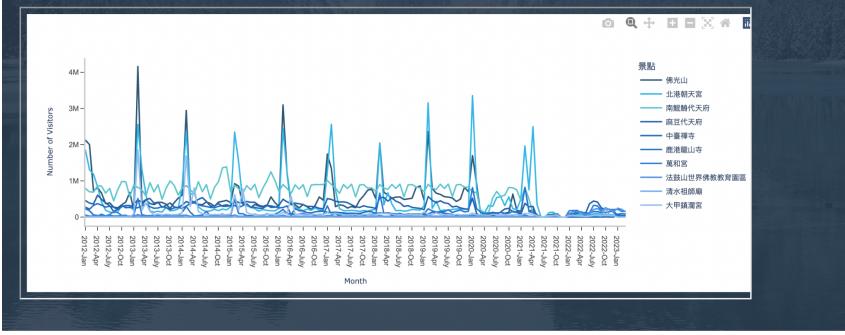
觀光地區

圖中確實樂園的人流在 2017~2020 的月平均都大於 40 萬，而在 2017 年前可以發現每年 7、8 月時皆有人流高峰，判斷是暑假與季節性開園的馬拉松使得人流增長。而在 2017 後持續性的高平均人流，推測是新增的園區與設施所帶來的人潮。



宗教場所

宗教場所在每年2月皆有高峰，其中又以佛光山與北港朝天宮人潮最多。台灣的民俗之一為在過年時至廟裡拜拜，因次宗教場所在過年時有規律性的高峰，而在疫情後，僅有2022年7月的麻豆代天府有些微人潮成長。研究過後發現是南美館的《亞洲的地獄與幽魂》展覽，一併帶動了麻豆代天府的人潮。我們覺得宗教場所的人流是受疫情影響最嚴重的，且在後疫情時代人流也並無回歸的趨勢。



其他

由圖中可以看到蓮池潭每年10月皆有人群高峰，為每年舉辦的左營萬年季的地點。而在2014年由於南高雄發生氣爆事件，因此停辦一年，當年也就沒有人流高峰。左營萬年季有多年歷史，2021年因疫情停辦。雖然2022年復辦，但是人流並沒有回復到以往的峰值，其人潮成長也不像以往有顯著的趨勢。

