中**華大學資訊工程學系**

**專題製作報告**

**大數據分析-空氣品質與癌症之間的關聯**

指導教授：歐陽雯

組員：陳楷勳 B10902202

中 華 民 國 114年 6 月 21 日

**目 錄**

**第一章 前言**

**1-1. 研究動機 ………………....................03**

**1-2. 研究目標 ………………....................03**

**第二章 系統流程與動作流程**

**2-1. 系統流程 ………………....................04**

**2-2. 流程說明 ………………....................05**

**第三章 成果**

**3-1. 成果 ………………........................13**

**第四章** **問題與討論**

**4-1. 問題 ………………….........................22**

**第五章 結論**

**5-1. 結語 ………………….........................23**

**5-2. 參考資料 ………………....................23**

**第一章 前言**

**1.1 研究動機**

空氣品質之評估，需要根據多項空氣污染物濃度來計算的，主要包括PM2.5、PM10、臭氧(O3)、二氧化硫(SO2)、二氧化氮(NO2)和一氧化碳(CO)，最終得到AQI，以此作為衡量空氣品質重要指標，AQI衡量方式並非為一線性公式，每個污染物的濃度可以用以下公式來計算如圖 1

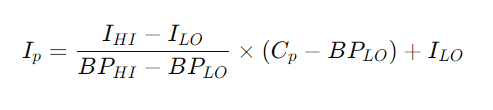


圖 1 子指標計算公式

其中Ip​ 是污染物 p 的子指標，C p​ 是污染物 p 的濃度，BP HI​ 和 BP𝐿𝑂是污染物濃度對應的高低標準值，IHI​ 和 ILO ​ 是對應的高低AQI標準值，再來各種濃度算完取最大值如圖 2。



圖 2 各種濃度算完取最大值

最終可以得到一個對應的分數範圍，如圖 3。

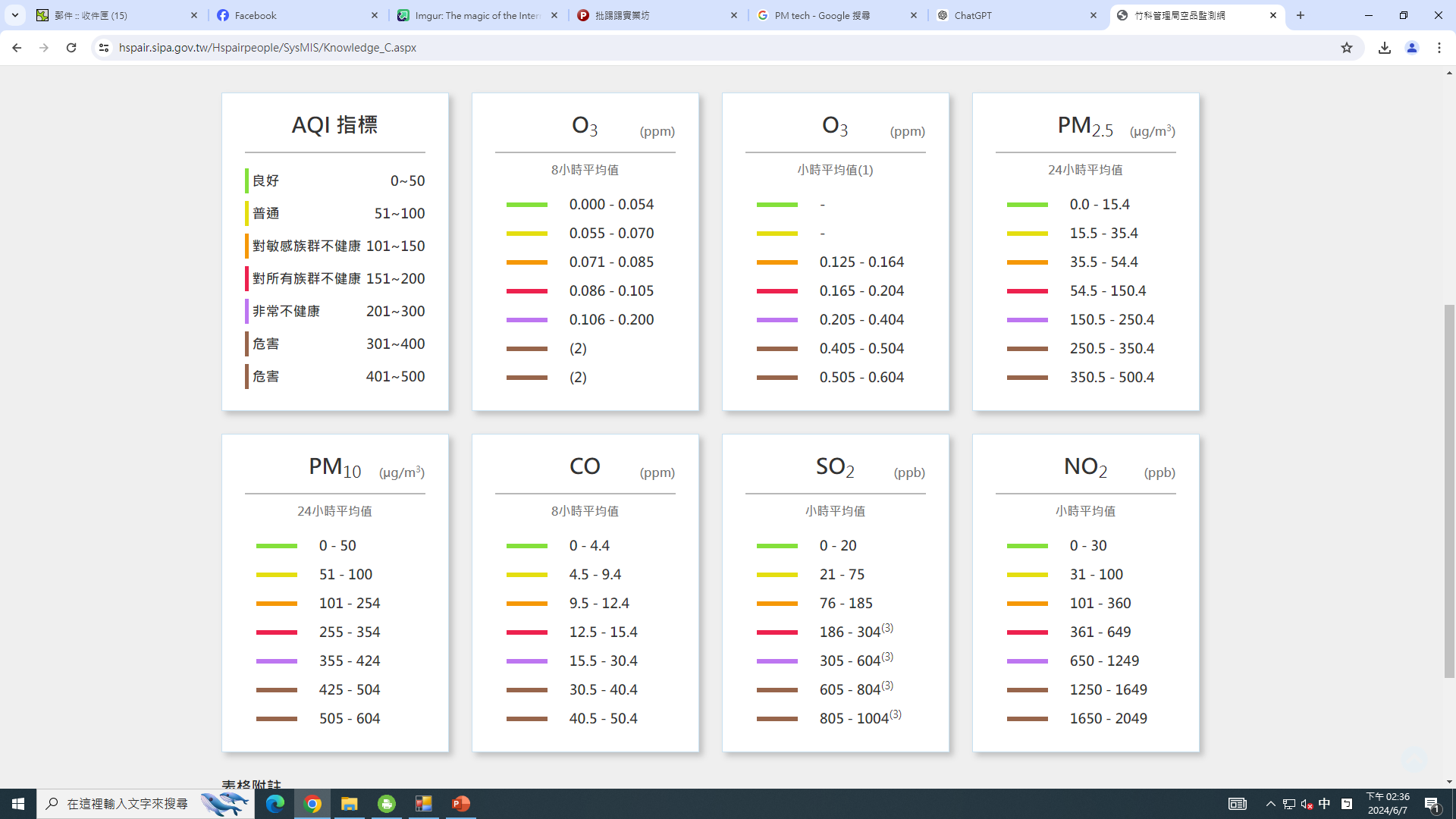


圖 3 各種汙染物對應分數

因此本研究希望直接用原始數據進行基礎統計以及建立預測模型，並將原始數據與肺癌人口做比較，嘗試尋求找到空氣品質數據與罹癌人口是否存在一定關聯性。

**1.2 研究目標**

本研究會將數據整理後，預計完成以下：

1. 做一系列可視化之圖表解釋可能之關聯。
2. 建立AQI預測模型。
3. 運用地圖資訊及罹癌人數繪製分布圖。

**第二章 系統流程**

**2-1. 系統流程**

放入MongoDB

用政府API抓資料

Kafka-producer

Kafka-consumer

提出資料

視窗應用程式

由MongoDB取資料

資料清理及統計

分train/test

訓練模型

模型評估

**2-2 流程說明\_爬蟲**

本專題所使用的資料為政府資料開放平臺中的資訊。

網址為https://data.moenv.gov.tw/api/v2/aqx\_p\_432?api\_key=ca6e7784-9cec-4631-b038-ce4ab6e075d1&limit=1000&sort=ImportDate%20desc&format=CSV，該資料每日抓取一份csv檔案，一份CSV大約含有85筆資料，日期區間為2023.9.12至2024.1.6

**2-2 流程說明\_kafka**

Kafka是一種開源的分散式串流處理平台和訊息佇列系統，最初由LinkedIn開發，並於2011年貢獻給Apache軟體基金會。 Kafka被設計用於處理即時資料流，具有高吞吐量、持久性、可擴展性和容錯性。

再將資料放入kafka前，我們需要先架設kafka，在這之前我們必須要架設ZooKeeper，ZooKeeper是一個開源的分散式協調服務，它為分散式應用程式提供了高度可靠的協作服務和資料管理。接下來就能架設kafka了，我們可以透過producer去輸入資料，當資料輸入後可以透過consumer去及時查看輸入的資料，此時我們只要 去訂閱這個consumer，就能獲得獲得經過kafka整理後所獲得的資料並將儲存下來，如果資料龐大且快速的話，有可能因為來不及儲存而讓資料遺失，這時kafka就能起到龐大的作用。

步驟:1.下載並解壓縮下載的kafka檔案

2.透過你解壓縮的資料夾去打開cmd

3.輸入.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties(架設zookeepr)

4.一樣的資料夾另外的cmd輸入.\bin\windows\kafka-server-start.bat .\config\server.properties(架設kafka)

5.創建topic kafka-topics.bat --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic testtopic

一張含有 文字, 螢幕擷取畫面, 字型, 黑與白 的圖片

自動產生的描述

Zookeeper架設成功

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

Kafka架設成功



本專題設立了”kafka\_topic”為” air-data”讓cosumer能夠去取得最新的資料

**2-2 流程說明\_記錄成CSV檔及資料預處理**

運用”pymongo”提取資料後存成csv，以做後續運用。有時候會因為種種原因而資料缺失，以及數字有離群值，因此針對比較怪異的數值及缺失就整列刪除。

**第三章 成果**

**3-1統計資料分析**

將AQI依照大小分出最大前五名縣市為左營、林園、仁武、琉球、大寮，最小前五名縣市為關山、台東、花蓮、冬山、宜蘭，挑選其中幾個資料作關聯係數矩陣。

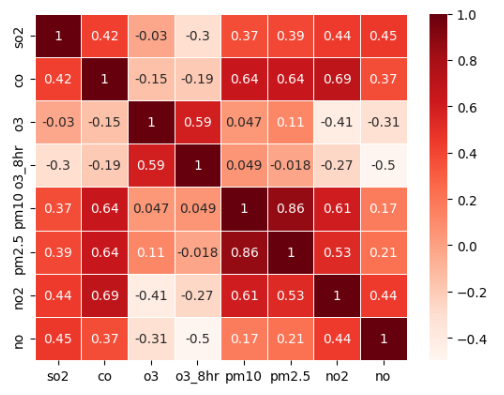
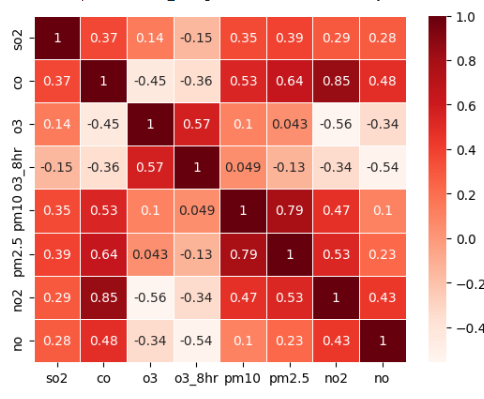


圖 4 左邊為左營右邊為林園

其中越深色表示越有關聯如圖 4，大部分情況關聯係數小於0.6，其中比較特殊的為PM2.5與PM10，關聯係數大約為0.8，推測可能是PM2.5和PM10都是空氣中的顆粒物，可能來自於類似的污染源，因此才會有同時升高減少的現象發生，另外是NO2與CO也是有比較高的相關性，由於這兩者是常見的汽機車廢氣，可能也是因為同為一個汙染源才導致相關係數較高。

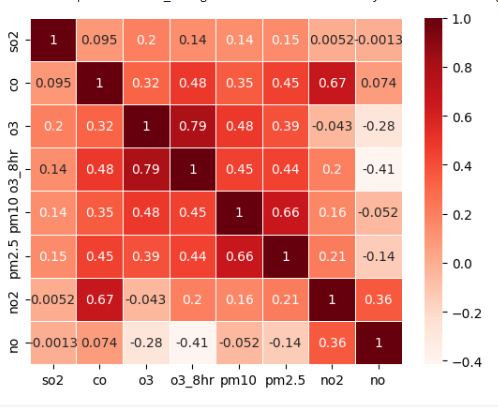
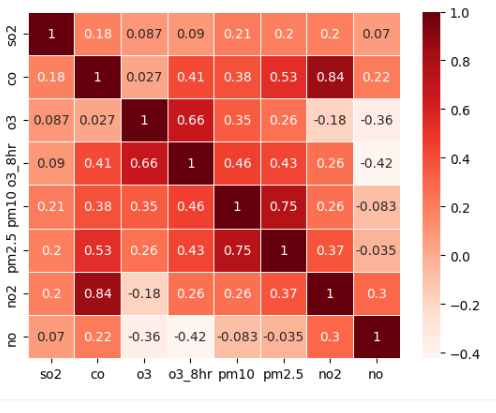


圖 5 左邊為花蓮右邊為台東

圖 5為AQI較小之區域，關聯係數與圖 4雷同。

以上數據是考慮單一地區全部時間所繪製的關聯係數矩陣，若是以日為單位，針對每一日AQI與各種汙染物(SO2 CO O3 PM10 PM2.5)算相關係數，再以日期做折線圖，以AQI依照大小分出最大前五名縣市為左營、林園、仁武、琉球、大寮，最小前五名縣市為關山、台東、花蓮、冬山、宜蘭作圖，可以得到如下

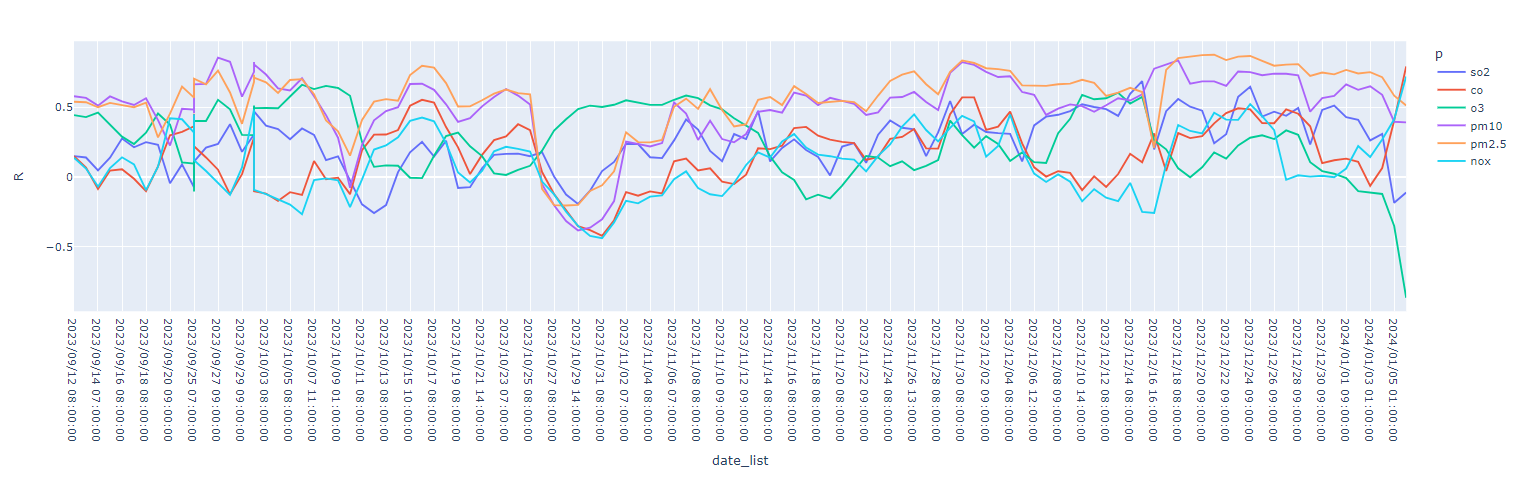


圖 6 左營關聯係數隨時間變化圖

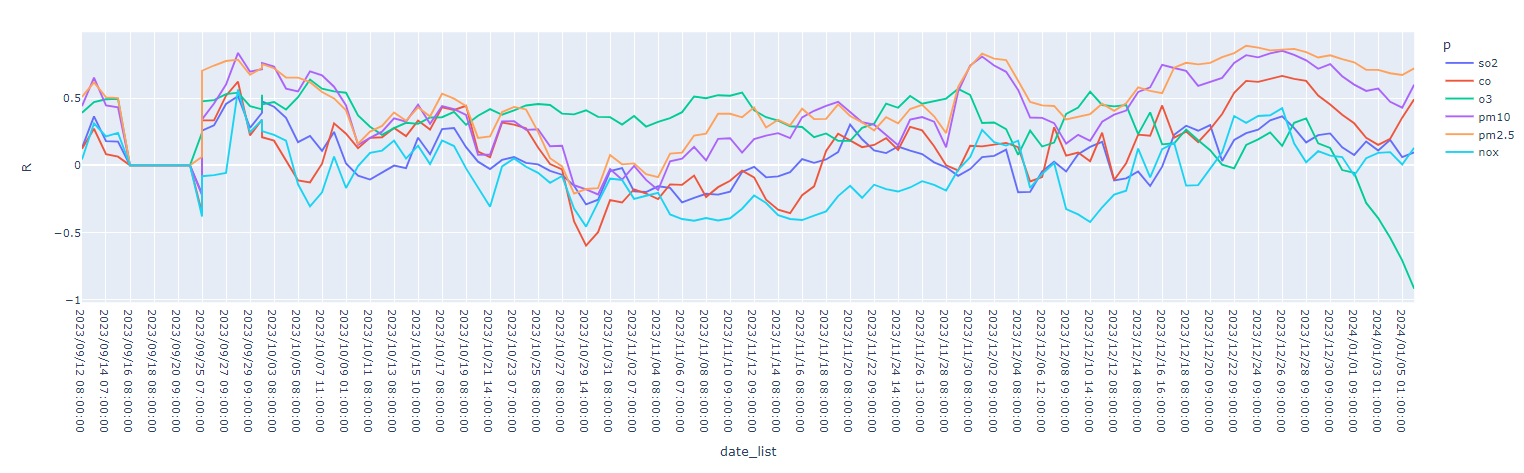


圖 7 林園關聯係數隨時間變化圖



圖 8 仁武關聯係數隨時間變化圖

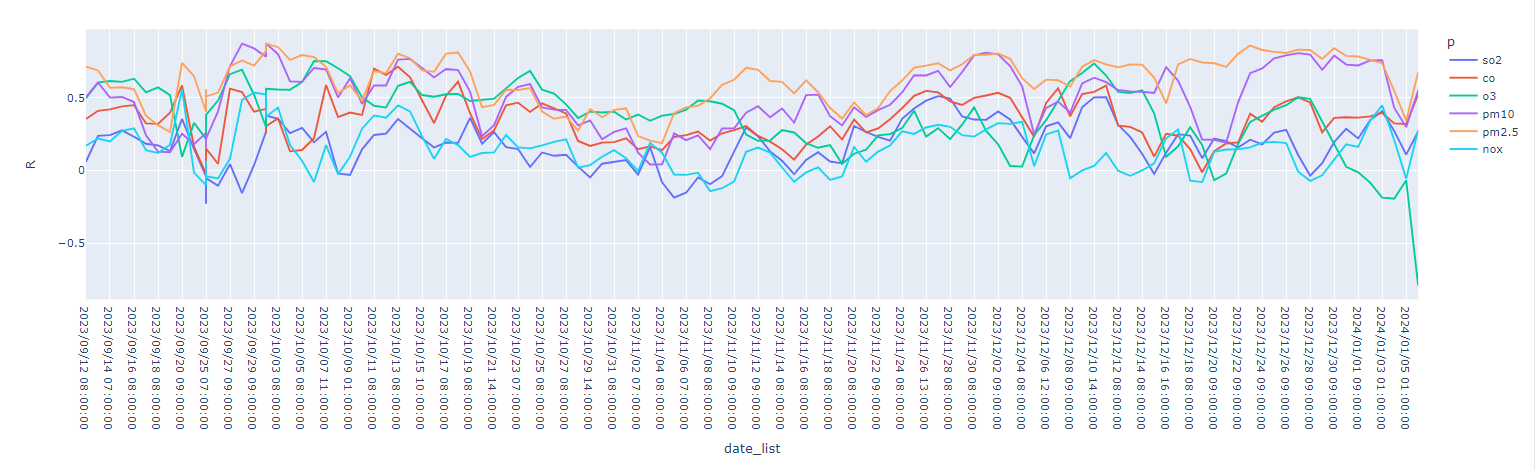


圖 9 琉球關聯係數隨時間變化圖



圖 10 大寮關聯係數隨時間變化圖

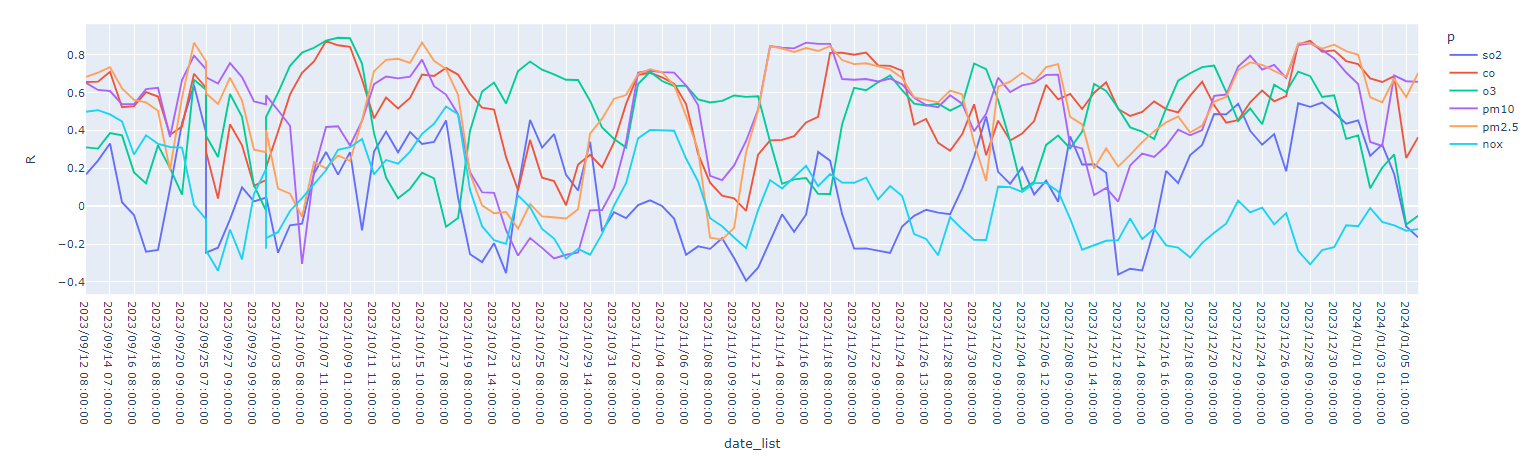


圖 11 關山關聯係數隨時間變化圖

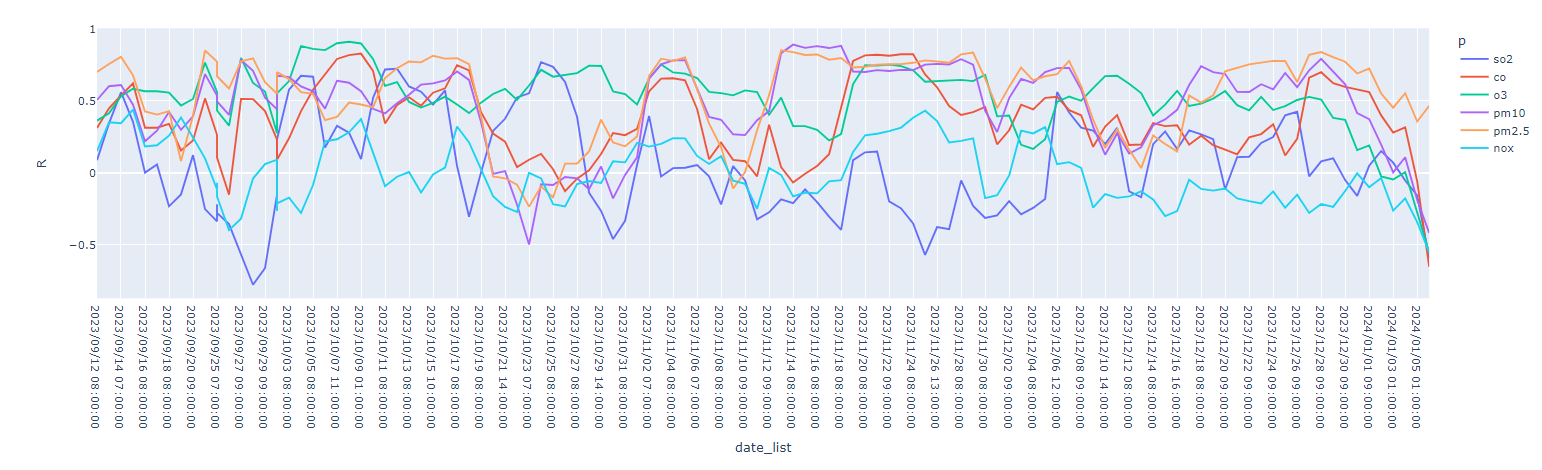


圖 12 台東關聯係數隨時間變化圖



圖 13 花蓮關聯係數隨時間變化圖

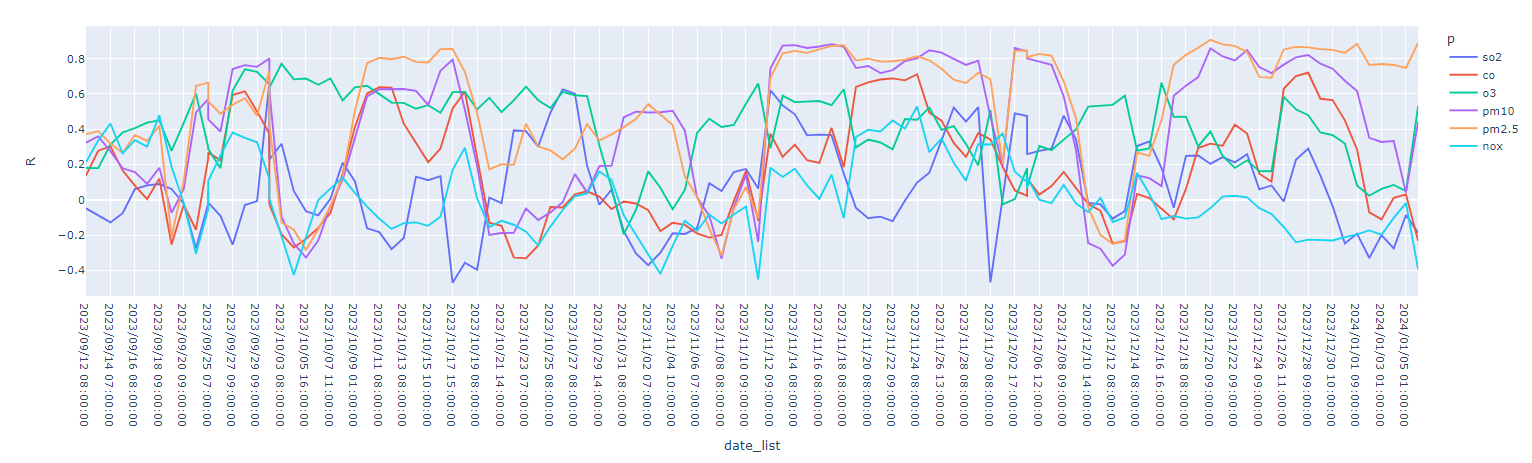


圖 14 冬山關聯係數隨時間變化圖

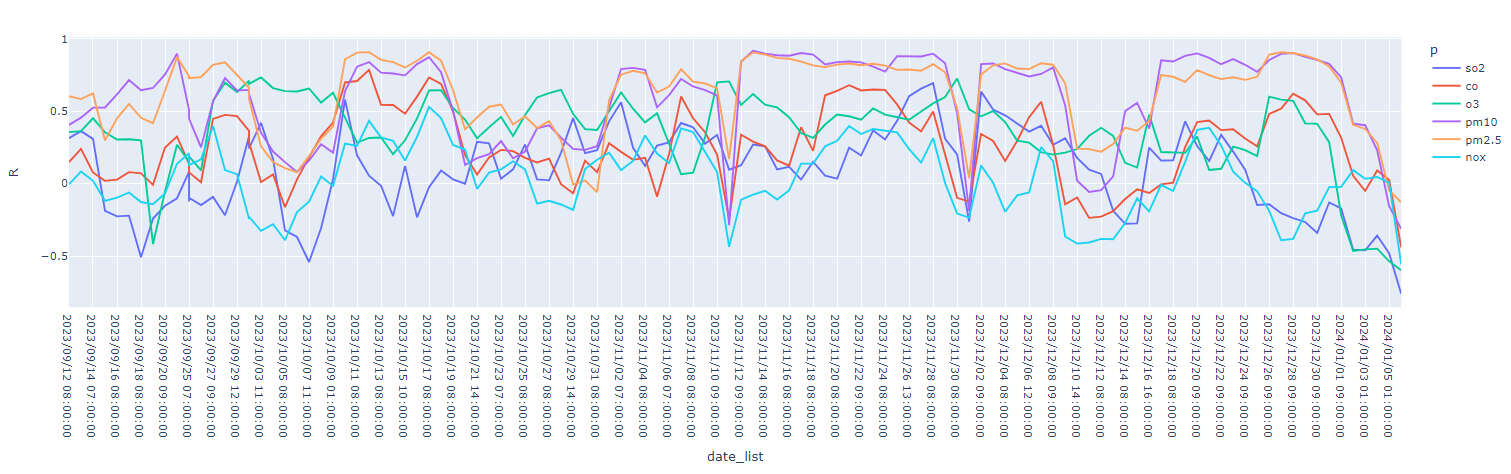


圖 15 宜蘭關聯係數隨時間變化圖

由於AQI本身是將各種汙染物指標取其最大值，因此總會有某些時間某種汙染物會與AQI最有關聯，例如以左營與林園為例子，在2023.12.4~2024.1.6這段期間，PM2.5與AQI會最有關聯，而PM2.5與PM10也有強相關，所以一樣可以在這個圖中發現PM10也是處於相對高值。汙染嚴重區域與污染不嚴重之區域有一個明顯差異在於，汙染嚴重的區域其關聯係數曲線大部分情況是偏上邊，但污染不嚴重之區域其曲線上下變化較為劇烈，推測原因是汙染物較低，一點波動就會讓相關係數變化劇烈，但因為不是常態高或低值，所以重要汙染物一直變換。圖 15有一個比較特殊的現象，其關聯係數在2024.1.5每種汙染物都下降，若攤開原始數據可以發現此時各種汙染物變化很小，沒有汙染物作為先行指標，所以顯得AQI值與任何一種汙染物都無關。

**3-2 預測模型**

本研究將使用LSTM作為預測之模型其結構如圖 16，輸入的數據由圖 17做舉例，其中AQI為要預測的對象，輸入的數據會用0~11筆數據，共12筆包含AQI、SO2、CO、O3、O3(8hr)、PM10、PM2.5、NO2、NO這些欄位共108個數據，去預測第12筆也就是33，再輸入會用1~12筆數據來預測第13筆也就是34，使用資料數為1841筆，參數調整epochs設100，batch\_size設30。

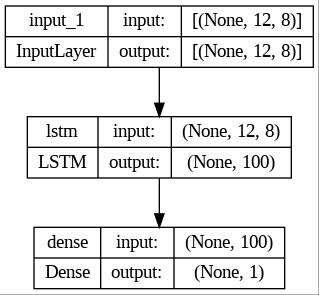


圖 16

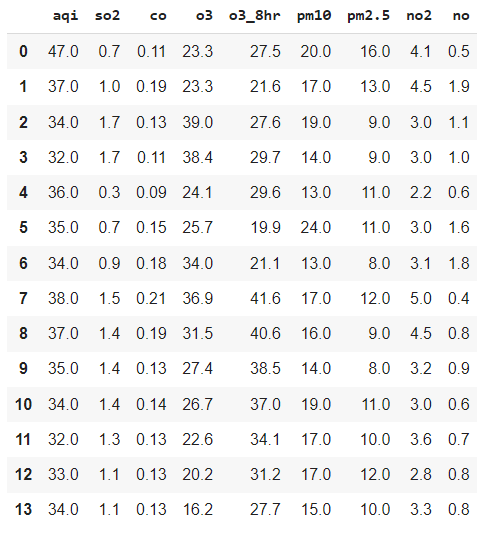


圖 17

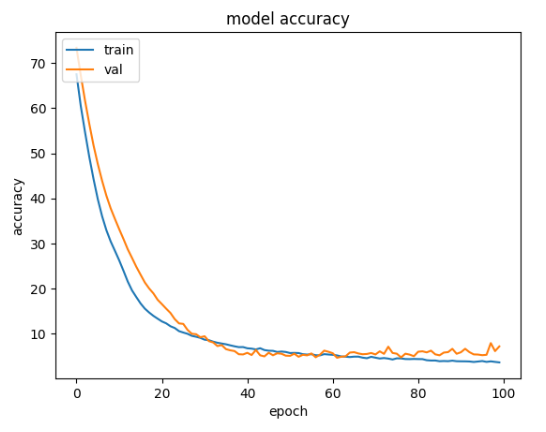


圖 18

程式碼 : 透過讀取MongoDB獲取資料獲得停車場資訊

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述**3-1 專案\_指定資訊**

透過下拉式選單能夠查詢在指定的時間點該停車場有車位的機率及後30分鐘的機率的折線圖，下拉式選單由左至右分別是停車場名稱,時間(小時),時間(分鐘),星期

一張含有 文字, 地圖, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 地圖, 地圖集, 圖表 的圖片

自動產生的描述

**一張含有 文字, 地圖, 地圖集, 圖表 的圖片

自動產生的描述**

**一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述**

程式碼 : 透過下拉式選單的選項去查詢資料庫裡指定資料

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

程式碼 : 繪製黑色的折線圖

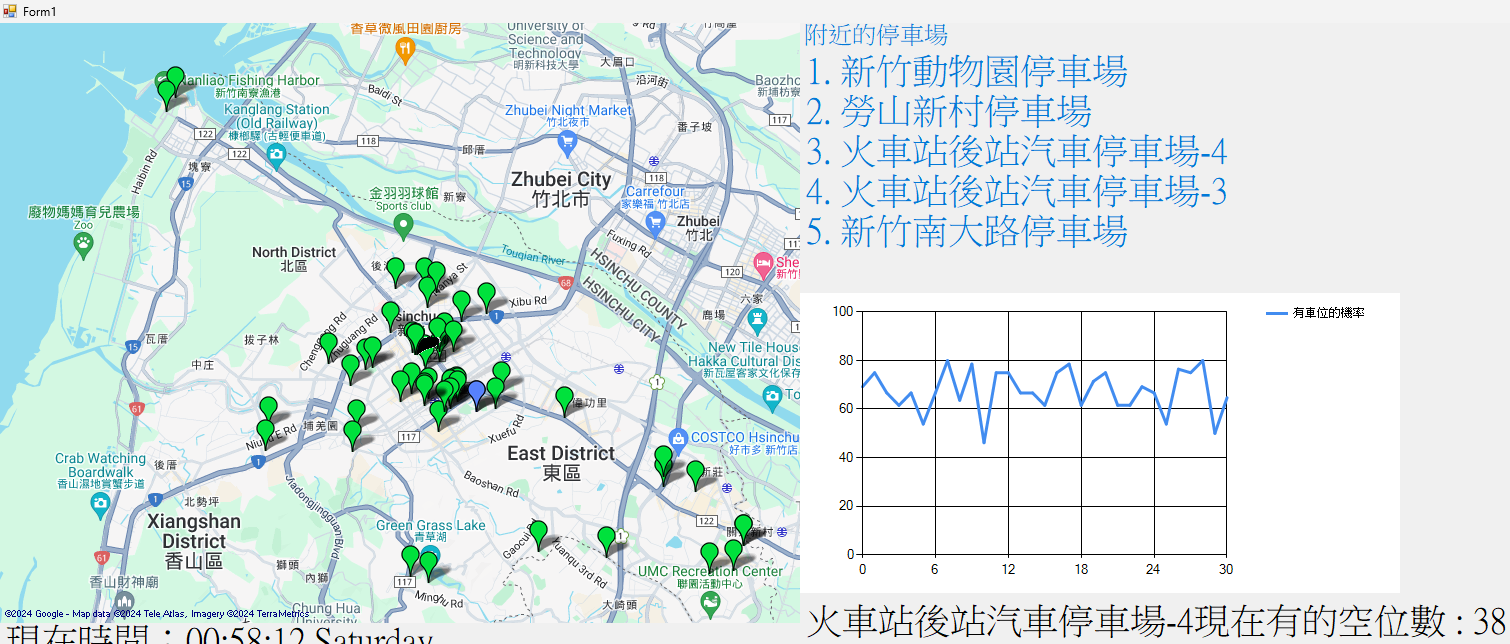
一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

**3-1 專案\_PIN**

**一張含有 文字, 地圖, 地圖集, 字型 的圖片

自動產生的描述**

在地圖上點擊兩下會產生藍色標點，此時旁邊的附近的停車場會顯示離該藍點前五近的停車場。

此時點擊其中一個停車場後(這裡點擊了火車後站停車場-4)，就會產生現在時間到後30分鐘的有剩餘空位的機率的折線圖，以及折現圖下方會透過即時爬蟲所抓下來的該停車場現在的空位數。

程式碼 : 標記藍點

**一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述**

程式碼 : 產生前五近的停車場

****

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

程式碼 : 繪製藍色的折線圖

**一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述**

程式碼 : 取得時間當下該停車場的空位數

**一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述**

**第四章 問題與討論**

**4-1. 問題**

本來想要使用機器學習去預測數據，也進行了許多嘗試，所使用的是LSTM，在這不斷嘗試的期間，調整了許多不同種的數值(learning rate,epochs,batch\_size,units)，

也改過許多次fetcher內容，但成果皆沒有起色，於是只能把此問題已人心莫測畫下結尾。

**第五章 結論**

**5-1. 結語**

這次的專題還有許多進步的空間，例如程式碼太繁瑣，其中能夠省下許多步驟，也能結合其他資料進行更多元的功能，例如:停車場附近的餐廳OR景點。我相信這個視窗能夠在某些時候給予幫助吧。

**5-2. 參考資料**

[<https://data.gov.tw/dataset/129136>](https://data.gov.tw/dataset/129136)