ChatGPT

空氣品質監測與預測 – 期末專案規劃

1. 資料科學研究目標

本專案旨在運用資料科學方法對臺灣的空氣品質進行監測與預測分析，提供環境管理與決策參考。主要研究目標包括：

1.1 空氣品質季節性預測

分析空氣污染物隨季節變化的趨勢，建立季節性預測模型。透過時間序列分解和預測技術，瞭解全年不同季節的空氣品質變化規律，例如預測冬季與夏季的污染水準差異，提供**季節性趨勢**的洞察。此目標有助於提前預判高污染季節來臨時的空品狀況，協助相關單位採取主動的管制措施。

1.2 污染源分類

利用**無監督學習**將空氣品質資料進行分群，從中歸納出不同**污染源型態**。透過分析各測站污染物組成與特徵（例如某些地區可能以交通汙染為主、某些地區工業排放較多），將相似污染特徵歸類，達到**污染源分類**的目的。此目標將有助於辨識各地區主要的污染來源類別，提供環境治理時區域差異化策略的依據。

1.3 異常檢測

建立空氣品質**異常事件**的自動偵測機制，及早發現空汙異常尖峰或資料異常。透過異常檢測模型偵測出明顯偏離正常模式的事件（例如突發性污染尖峰、儀器錯誤數據），並對這些異常情境進行研判與說明。此目標可用於環境預警系統，及時識別嚴重污染事件或資料品質問題，進而分析其成因與影響。

2. 資料來源

專案所使用的資料主要來自臺灣政府的**環境資料開放平臺**之空氣品質開放資料集。例如環境部監測資訊司提供的「空氣品質指標(AQI)歷史資料」(資料集代碼AQX_P_488) [1]。該資料集每小時提供全臺各環保署空氣品質測站的即時監測數據，包括空氣品質指標AQI以及詳細的污染物濃度與氣象資訊 [2]。主要欄位涵蓋測站名稱、縣市、AQI指數、主要污染物種類、各項污染物濃度（如SO_2、CO、O_3、PM_{10}、$PM_{2.5}$、NO_2等）以及氣象參數如風速、風向，另外還有資料時間戳記與測站座標等 [2]。

本專案將下載**全年逐時資料**進行離線分析，同時也使用**即時資料**進行即時預測展示。資料取得方式可透過開放平臺提供的API或批次下載歷史資料（如CSV檔）。在使用資料前，將確認資料授權方式符合政府資料開放平臺規範並遵循相關使用條款。

由於資料來自全臺各測站，資料量龐大且具有**時間序列**特性。我們預計選取近一年的完整逐時資料作為分析範圍，涵蓋各個測站在不同時間的污染物讀值。必要時也會輔以相關的輔助資料來源（例如地理資訊或天氣資料）進行交叉分析，以豐富預測模型的輸入特徵。

3. 資料處理流程

為確保資料品質並提取有效資訊，專案將執行以下資料前處理步驟：

3.1 資料清洗

對取得的原始空氣品質資料進行清洗，處理**遺漏值**、**異常值**以及單位或格式不一致的情形。可能的清洗步驟包括： - **遺漏值補植**：針對少量缺失資料，以線性插值或以相鄰時刻/相鄰測站數據估計填補，必要時也可標記缺失以避免誤用。 - **異常值過濾**：偵測明顯超出合理範圍的數值（例如儀器故障導致的極端值），可依環保署污染物合理區間設定閾值篩選，或結合異常檢測模型（如Isolation Forest）初步標記可能的異常資料點，後續分析可排除或特別關注這些點。 - **單位與格式統一**：確認所有測站的各項污染物單位一致（開放資料已定義各欄位單位），如有需要將單位換算統一。另外將時間欄位轉為適當的時間戳記格式，確保不同資料集（即時與歷史資料）時間格式一致，方便後續整合。

經過清洗後，將得到一份較為完整且可靠的時序資料，為特徵提取和模型分析打下基礎。

3.2 特徵工程

為提升模型的預測能力與分析的洞察力，我們將對清洗後的資料進行特徵工程，包括生成新的衍生變數和統計量： - **時間相關特徵**：從時間戳記提取出相關特徵，例如小時（判斷日夜或通勤時段）、星期幾、月份等，捕捉日週週期性模式。也可加入節假日標記（例如農曆新年、跨年夜等可能影響空污的節日），以供模型考量特殊事件對污染的影響。 - **移動平均與滯後特徵**：基於原始濃度值計算移動平均值，平滑短期波動。例如計算$PM_{2.5}$的24小時移動平均以捕捉日尺度趨勢（資料中已提供部分污染物的移動平均，如8小時臭氧、$PM_{2.5}$日均值等，可再次確認計算公式）。同時產生**滯後特徵**，例如將前1小時、前3小時、前24小時的污染物濃度作為預測當前時刻的特徵，用以引入時間序列的自相關信息，這對於像XGBoost等機器學習模型預測時間序列非常重要。 - **風速風向特徵**：將風向與風速資訊轉換為更適合模型的特徵。例如將風速 (WindSpeed) 與風向 (WindDirec) 轉換為**平面向量**的兩個分量：$u = \text{WindSpeed} \times \cos(\text{WindDirec})$，$v = \text{WindSpeed} \times \sin(\text{WindDirec})$，代表東西向和南北向的風速分量。這樣可更直觀地反映風吹拂帶來的污染物傳輸效應。此外，也可依風向將數據分類（例如將風向離散為八個方位），探討不同季風方向下的污染水平差異。 - **跨測項特徵**：計算不同污染物之間的比值或組合，以捕捉污染來源線索。例如$PM_{2.5}/PM_{10}$的比值可反映細顆粒物在總懸浮微粒中所佔比例，較高可能代表燃燒來源；再如NO_2/NO比值可反映交通排放特性等。這些衍生特徵將有助於後續的污染源分群分析。 - **統計特徵**：對於需要以測站為單位分析的任務（如污染源分類），可先為每個測站計算統計特徵值，例如該站在分析期間內各污染物的平均值、峰值、中位數等，或週末與平日平均值差異等，作為該測站的特徵向量，用以進行測站層級的比較。

透過上述特徵工程，我們將獲得豐富的特徵資料集，以供不同模型使用，提高模型對複雜模式的捕捉能力。

3.3 資料整合

考量本專案需分析**跨測站**與**長時間序列**的資料，我們將進行資料整合與重組： - **測站資料彙總**：將各測站的資料按照時間對齊，形成一個包含所有測站的統一時間索引。如有需要進行全域性的分析（例如以空間為維度的聚類），可建立「時間 × 測站 × 污染物」的三維資料表。針對污染源分群，我們可能對每個測站計算特徵後形成「測站 × 特徵」的資料表格，用於後續的分群模型。 - **資料集結合**：如果有來自不同資料集的資訊需要結合（例如環境氣象資料，如溫度、降雨等），則按照時間和地點將其與空氣品質資料合併。在本專案中，基本的風速風向已包含於AQI資料集中，若有需要更完整的氣象要素（如氣溫、降雨），可從中央氣象局等開放資料取得並根據測站位置匹配。然而在目前計畫中，以空氣品質測站自帶的參數為主。 - **地理資訊整合**：為了製作污染地圖與進行區域分析，我們需要測站的經緯度座標（資料集中已提供Longitude/Latitude）。將測站地理資訊與污染數據關聯，方便後續在Leaflet地圖上標示測站位置並呈現污染程度。另外，可整合各測站所屬行政區（縣市）資訊，便於依城市或區域彙總統計。例如，可計算各縣市的平均AQI或污染物濃度以供比較。 - **資料庫與更新**：為支持Shiny應用的即時瀏覽與互動，我們考慮將清理和特徵處理後的資料存入資料框或輕量資料庫中。歷史逐時資料可保存為R的資料檔或CSV供Shiny讀取，即時資料則可透過API定期抓取最新值並動態更新。在整合過程中確保歷史資料與新資料的格式一致，以便一併餵入模型或視覺化模組使用。

完成整合後，我們將擁有完善的分析資料集，可滿足不同分析與建模步驟的輸入需求。

4. 資料分析與建模方法

本專案將同時採用傳統統計時間序列方法與機器學習方法，以全面分析和預測空氣品質。具體方法包括時間序列分解分析、預測模型建置、分群分析辨識污染源，以及異常檢測，詳述如下：

4.1 時間序列分析 (季節性分解與趨勢分析)

首先對各測站的空氣品質時間序列資料進行探索性分析，以了解長期趨勢及週期性。採用**STL (Seasonal and Trend decomposition using Loess)季節性分解**技術將時間序列拆解為**長期趨勢**、**季節性周期**和**殘差**成分。透過STL分解，我們可以觀察例如每日週期（可能的上下班尖峰模式）、每週模式（週末/平日差異）以及年度季節性（夏季與冬季空污差異）等。在空氣品質資料中，預期可看到**顯著的季節效應**，例如冬季$PM_{2.5}$濃度明顯高於夏季，且每年約呈現周期性變化 [3] 。季節性分解圖將以視覺化方式展示這些規律，有助於理解污染變化的自然模式。

接下來基於時間序列特性建立**統計預測模型**。我們將嘗試**ARIMA模型**（自我迴歸整合移動平均）捕捉資料的自相關和趨勢。對於具有季節性的污染物（例如AQI或$PM_{2.5}$呈現年度周期），會使用**季節性ARIMA (SARIMA)**，將年周期納入模型階數中。ARIMA模型適用於平穩化後的時間序列，因此在建模前會對序列進行差分以去除單位根、並以ACF/PACF圖來選擇適當的AR和MA項。

除了傳統ARIMA，本專案亦計畫採用**Prophet模型**進行時間序列預測。Prophet是Facebook開發的開源預測工具，擅長處理包含多重季節性與假日效應的資料 [4] 。我們可利用Prophet自帶的週期性成分（年季節性、週期、每日週期）來擬合空氣品質指數的變化，並可將重要節日（如春節期間可能有燃放鞭炮導致短期$PM_{2.5}$升高）作為假日效應加入模型。相比ARIMA，Prophet對缺失值和異常值較具魯棒性，且參數易於調整，非常適合快速建立基線預測。

透過上述時間序列分析，我們將獲得對空氣品質數據本身結構的深刻理解，並為後續的預測建模提供基線和參考。如果某些污染物在特定季節**趨勢明顯**或波動增加，我們會特別關注（例如冬季極端高值頻率較高 [3] ），並在預測時適當調整模型以因應這些變化。

4.2 預測模型 ($PM_{2.5}$ / AQI 預測)

在建立了時間序列基線模型後，我們將引入機器學習模型進行空氣品質的預測，重點針對細懸浮微粒$PM_{2.5}$濃度或AQI指數做短期預測。主要採用的模型是**XGBoost**（Extreme Gradient Boosting），一種基於梯度提升的樹狀模型。選擇XGBoost的原因在於其在迴歸預測任務中表現優異，能自動捕捉特徵與目標之間的非線性關係，並處理高維特徵及特徵交互。過去研究也表明XGBoost在空氣污染預測上可取得不錯的準確率，例如某研究中XGBoost模型預測$PM_{2.5}$濃度的測試集R^2超過0.84 [5] 。

我們將以前述整理的特徵資料作為XGBoost的輸入，包括時間滯後的污染物濃度、氣象條件（風速風向等）、日夜與季節等指標。模型輸出可以是**連續數值預測**（如下一小時的$PM_{2.5}$濃度或AQI值），也可以延伸做**多步預測**（如預測未來24小時的濃度曲線）。在訓練過程中，我們會將資料集區分為訓練集與測試集，以歷史資料訓練模型，並用保留的近期資料檢驗模型效能，避免過度擬合。必要時也會採用**交叉驗證**評估模型穩定性。

除了XGBoost，本專案也可嘗試其他機器學習或深度學習模型進行預測，例如隨機森林、長短期記憶神經網路（LSTM）等，並與XGBoost進行比較。不過考量專案時間與重點，XGBoost作為重點模型應足以提供精準度和可解釋性的均衡。

在模型完成訓練後，我們將評估其預測表現，詳情於後述**分析指標**章節說明。若發現預測誤差較大的時段（例如冬季高峰時段），我們將分析其原因，可能是模型未能充分捕捉極端狀況的特徵，此時可考慮針對該情境增強特徵或調整模型參數（例如增加樹的深度以擷取非線性模式）。總體而言，此預測模型將為空氣品質提供短期預報，可在Shiny介面中與實際值比較，供使用者掌握未來趨勢。

4.3 污染源分類 (分群分析)

為了探討**各地區污染源的異同**，本專案將對空氣品質資料進行**分群分析**。我們計畫採用**K-Means**與**階層式分群(Hierarchical Clustering)**等無監督學習方法，將性質相近的資料點歸為一群。分群的對象有兩種考量：

- **以測站為單位進行分群**：將每個測站在分析期間內的污染特徵向量進行聚類。特徵向量可由該測站的主要污染物平均濃度組成（例如$PM_{2.5}$、SO_2、NO_2等長期均值或某種典型日型），或加入該站所在位置類型（都會區/工業區/郊區）。透過對測站分群，我們期望將測站劃分為若干類型，例如「交通汙染型測站」（可能特徵是CO、NO_2偏高）、「工業汙染型測站」（SO_2、NOx偏高）、「背景測站」（各污染物濃度普遍較低但O_3較高）等。這將對應到不同的主要污染來源。**階層式分群**可協助我們了解分群的層次結構，例如先大致分成都市與非都市，再細分不同類型。已有研究證明，以微粒組成特徵對監測站分群能夠得到在空污研究上具有意義的群組劃分 [6] 。

- **以時間/事件為單位進行分群**：另一種方式是將每個時間點（或每天、一段期間）的多項污染物組成視為一個資料點，對這些點聚類。如此可將**污染事件**按成因或特性分組，例如一群事件是「高臭氧高溫日」，另一群可能是「春季沙塵暴事件」（高PM_{10}）、或「秋冬交通污染尖峰」等。此分析需要將同一時間不同污染物的值組合在一起進行聚類，可能需要先篩選特定測站或區域作為代表。此方式有助於了解同類型的污染事件在時間上的分布，從而推斷其來源（如沙塵事件多發於春季特定天氣）。

在實施分群時，我們會先利用**K-Means**嘗試將資料分成數個集群，透過肘部法(Elbow method)或輪廓係數(Silhouette Coefficient)來選擇適當的分群數k。K-Means可快速收斂並給出初步分組結果。接著可應用**階層式分群**對K-Means結果做驗證與細化，例如使用凝聚式階層分群對相似群組進一步合併或拆分，並繪製樹狀圖觀察群之間的距離。分群結果將連結回實際地理位置和可能的污染源：我們會分析每個群組的特徵污染物組成，以及群組中測站的地理分布或時間分布，進一步推斷其代表的污染來源類型。理想情況下，每個群組都能對應到**明確且可解釋**的污染源型態，例如群組A代表「都會區交通型污染」、群組B代表「工業區排放型污染」等，驗證這些歸類是否符合我們對各區域的認知。

4.4 異常檢測

針對空氣品質資料中的**異常狀況**（如極端高值、異常波動），本專案將引入**異常偵測模型**予以自動化識別。我們將主要考慮兩種無監督的異常檢測演算法： - **Isolation Forest (孤立森林)**：Isolation Forest通過隨機選擇特徵與切分值構建決策樹來分離觀測值，異常點因較易被隔離而具有較短的平均路徑長度。我們會以歷史資料訓練一個Isolation Forest模型，使其學習正常空氣品質讀數的分佈，然後對每筆資料計算一個異常分數。如果某時刻某測站的多項污染物讀值組合在模型下得到高異常分數，則判定其為潛在異常事件。由於Isolation Forest能同時考慮多變數，我們可以利用它來偵測**複合型異常**（例如某時段$PM_{2.5}$和SO_2都異常升高）。文獻顯示Isolation Forest在偵測空汙尖峰事件方面相當有效，可準確找出污染劇增的時刻 [5] 。 - **Local Outlier Factor (LOF，局部離群因子)**：LOF透過比較一個點與其鄰近點的密度差異來判定異常程度。如果一筆資料的局部密度明顯低於周圍鄰居（即該點相對孤立），則LOF值偏高表示異常。我們可針對各測站各污染物建立LOF模型來發現時間序列中的離群點。此方法在處理單變數時間序列的異常（例如某站$PM_{2.5}$單一濃度激增）時直觀有效。LOF也適合偵測**局部異常**，例如大部分測站在某時刻污染值都高（區域性事件），只有個別站特別低或高，這種情況亦可被LOF識別出來。

實際應用中，我們可能結合兩種方法的優點：Isolation Forest用於全域異常篩查，LOF用於局部微調和確認。異常偵測的結果將提供一份可疑異常事件清單，包括事件發生的時間、地點、涉及的污染物及其數值。我們將對這些偵測出的異常進行分析與標記，例如確認它們是否為環境上的**真實異常**（如沙塵暴入侵、境外移入煙霧或重大工安事故排放），抑或只是儀器校正/資料遺漏導致的假異常。藉由與氣象資料或新聞事件比對，我們可以為主要的異常事件提供說明。例如，若Isolation Forest偵測到某年夏季某天多個測站PM_{10}濃度極高，且範圍涵蓋全台，可能對應一次境外沙塵事件；再如春節期間特定城市$PM_{2.5}$深夜出現異常峰值，可能與煙火燃放有關。透過這種方式，異常檢測模組不僅能保障資料品質（偵測出異常資料點避免對模型造成干擾），更能從中發現**突發環境事件**並納入研討。

5. 呈現方法 (Shiny互動應用介面)

為了直觀展示分析結果並提供互動式操作，本專案將使用R語言的**Shiny框架**開發一套網頁應用介面。此Shiny應用將匯集資料瀏覽、圖表展示、地圖視覺化和模型結果於一處，方便使用者探索。我們計畫的介面包含以下主要模組：

5.1 資料瀏覽與篩選模組

提供一個資料查詢介面，使使用者能夠瀏覽原始數據和清理後的資料集。介面上將有篩選選項，允許使用者按**縣市、測站、污染物類型、時間範圍**等條件篩選資料。例如，使用者可以選擇「臺北市 – 古亭站 – $PM_{2.5}$ – 2024年1月至3月」，介面將即時顯示符合條件的資料表格或摘要統計。表格中將列出時間、各項污染物濃度、AQI等，並高亮超標或異常的值（如AQI>100的格子以顏色標示）。此外，可提供基本的資料下載功能，讓使用者將篩選後的資料匯出CSV以供進一步分析。此模組確保了**資料透明度**，讓使用者先熟悉數據概況。

5.2 變化趨勢圖表模組

在這部分，我們將展示空氣品質指標和污染物濃度隨時間的**變化趨勢圖**。使用者可在介面上選擇欲查看的**縣市或測站**以及**污染物項目**，系統即生成對應的時間序列圖。圖表形式包括： - **時序折線圖**：顯示所選測站在指定時段的某污染物濃度走勢。例如展示某測站$PM_{2.5}$在過去一年的逐日平均變化折線，方便觀察高低起伏和季節趨勢。 - **季節性分解圖**：對於長時間序列資料，可按使用者選定的測站，生成該測站AQI或主要污染物的STL分解圖。圖中將包含原始序列、趨勢、季節性、殘差四個子圖，讓使用者明確看到季節循環模式與長期變化。透過此圖，使用者可以驗證例如冬季是否明顯高於夏季、每日下午是否有臭氧高峰等直觀現象。 - **比較圖**：可以允許選取多個測站或多種污染物進行重疊比較。例如在同一圖上繪製臺北與高雄的AQI走勢對比，或將$PM_{2.5}$與PM_{10}的走勢疊加比較其相關性。這有助於發現區域同步性或污染物間的關聯。 此模組的圖表均會加入互動功能，如滑鼠懸停顯示精確數值、拉伸選取子區間放大檢視等。使用者也可以切換圖表類型（折線、長條、盒鬚圖等）以不同角度檢視資料分佈。例如，可切換為月度盒鬚圖察看每月$PM_{2.5}$分佈及中位數走勢。

5.3 預測結果展示模組

本模組重點呈現**模型預測**的結果，包括未來趨勢預報和模型性能比較。具體設計如下： - **未來趨勢預報圖**：使用者選擇某測站與污染物後，系統將利用先前訓練的預測模型（如XGBoost或ARIMA）計算未來若干時段的預測值，並與歷史實測值拼接繪圖。例如，可以在圖上顯示「截至昨日的實際$PM_{2.5}$濃度」以及「未來三天的每小時$PM_{2.5}$預測走勢」，預測部分以虛線標示，並給出模型預測的**信心區間**（例如95%預測區間帶）。這讓使用者對短期未來的空氣品質有直觀瞭解。如果預測模型有多種，我們也可同時顯示不同模型的預測曲線以供比較（例如Prophet vs. XGBoost）。 - **預測精度比較**：在此亦提供模型表現的數值指標。可製作一個表格或圖表列出不同模型在測試集上的誤差評估，例如ARIMA、Prophet、XGBoost各自的RMSE、MAE值。使用**條形圖**表示也直觀，如條形長度代表誤差大小，誤差越小的模型條形越短，幫助使用者識別最佳模型。此外，若我們有即時更新的預測，還可以實時比較「今日預測的AQI與今日實際AQI」的差異，以檢視模型在現實中的表現。 - **數值預報查詢**：除了圖形，對於預測的數值我們也提供查詢介面。使用者可選擇日期時間，看到模型對該時段各污染物的預測值，方便取得精確預報數字。例如明天早上8點某測站$PM_{2.5}$預測為35 μg/m³，AQI預測為80等。

預測結果展示模組的價值在於**將複雜模型輸出以人性化方式呈現**。使用者不用深入模型內部即可理解未來空氣品質走向，同時對模型可靠度有明確認識（透過誤差指標和歷史比較)。這部分功能對環境決策者相當重要，例如當模型預示未來幾天空污將惡化時，可提前發布警報或啟動應變措施。

5.4 污染源地圖展示模組

空氣品質具有明顯的空間分佈特性，因此我們將利用地圖視覺化來呈現各地區的污染狀況和可能的污染源分佈。使用**Leaflet地圖**在Shiny中進行開發，提供互動式的**空品地圖**： - **測站即時狀況地圖**：地圖上標示全臺各空品測站的位置（依經緯度）。每個測站以圓點或圖示表示，並以顏色或大小代表某污染指標。例如，以AQI顏色標準渲染測站顏色：綠色表示良好、紅色表示空污嚴重。當使用者改選不同污染物時，也可切換顏色代表的指

標（如選$PM_{2.5}$時，顏色深淺按$PM_{2.5}$濃度範圍顯示）。使用者可點擊任意測站，彈出氣泡窗口顯示該站的名稱及最新各項污染物數值。 - **熱區/熱點地圖**：基於測站數值生成**污染熱點圖**。例如利用空氣品質資料做空間內插，在地圖上以顏色漸層呈現某污染物濃度的地理分布。Leaflet可結合外部套件（如heatmap layer）將測站濃度映射為連續的色斑圖，讓使用者一目了然哪裡是高污染熱區。例如在東北季風影響下，可能呈現西部濱海地區污染累積、而東部清淨的圖像。此功能有助於理解污染物的**地理擴散**態勢。 - **污染源群組地圖**：結合前述**污染源分類**結果，在地圖上以不同顏色或形狀標記不同群組的測站。例如將聚類分析分得的三類測站，用紅三角表示第一類（如交通型）、藍方形表示第二類（工業型）、綠圓形表示第三類（背景型）等。這樣可以直觀呈現不同類型污染源在空間上的分布格局，驗證某類型站點是否集中於特定區域。例如若工業型測站集中在中南部沿海工業區，地圖將清楚反映此現象。 - **互動過濾**：地圖介面亦提供篩選控制，使用者可以勾選欲顯示的污染物或群組類別，或者拖動時間軸來查看不同時間的空污空間分布（如播放一日當中AQI的地理變化動畫）。這種互動過濾讓使用者探尋特定情境，例如僅看某日時段的O_3分布，或只看屬於交通污染群組的測站分布。

透過Leaflet地圖模組，抽象的數據將轉化為易於理解的空間資訊。**地圖化呈現**能強調區域差異，幫助使用者連結地理位置與污染特性，例如發現**污染熱點**與工業區重疊，或山區背景站空氣較佳等，從而為研擬區域環境治理策略提供直觀依據。

5.5 分群與異常結果展示模組

最後，Shiny應用將包含對**分群分析結果**與**異常檢測結果**的視覺化展示，並提供互動式的探索工具： - **分群分析視覺化**：對於污染源分群結果，我們將提供數種圖表來解讀群組特性。例如使用**雷達圖**或**平行坐標圖**顯示各群組的特徵污染物平均值，讓使用者比較群組間的異同；使用**降維散佈圖**（如PCA或t-SNE降維後的2D散點圖）將測站投影到二維空間，以不同顏色表示群組，直觀展示群組的分離度和緊密度。另一個圖表是**樹狀圖**（dendrogram），特別對階層式分群有用，使用者可藉此調整截斷距離觀察不同數目的群組分類。介面上允許點擊選擇某個群組，並高亮地圖上屬於該群組的測站，以及在資料瀏覽模組中篩選出該群組的測站資料，以進一步檢視。 - **異常事件視覺化**：將異常偵測模型找出的異常點在介面上清晰標示。一種方式是在**時間序列圖**上將異常時刻做標記（例如以紅色圓點標出Isolation Forest判斷的離群點），使用者可以滑鼠點選該點以顯示詳情（時間、測站、各污染物值及異常分數）。此外，可提供**異常事件列表**，列出重大異常事件（依異常分數或影響範圍排序），每項事件描述其時間範圍、涉及區域和可能成因。對於範圍影響較廣的事件，也可在地圖上提供播放功能，展示事件發生期間污染擴散的情形。例如一場沙塵事件期間，各測站PM_{10}濃度同步升高的地理推進。透過視覺化的輔助，使用者能更好地理解異常事件的特性，例如其空間影響範圍是局部還是全區域性的，以及異常期間不同污染物的關聯變化。

分群與異常結果的互動展示，能將**抽象的分析成果轉化為具體的畫面**。使用者不僅可以看模型輸出的群組或異常標籤，還能直接觀察這些結果在時空上的意義。例如，透過點選異常事件列表中的「2024/03/05 沙塵暴事件」，地圖和時間序列同步顯示該事件的影響，有助於驗證模型結果並增進對空污機制的理解。

6. 分析指標

為評估模型性能以及分析結果的有效性，本專案將採用多種量化指標：

6.1 預測模型準確率指標

對於預測模型（ARIMA、Prophet、XGBoost等）的表現評估，我們將使用**迴歸誤差指標**來衡量預測值與實際值之間的差距： - **RMSE (Root Mean Square Error，均方根誤差)**：計算預測誤差的均方根值，公式為$ \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{y}i - y_i)^2} $。RMSE對較大誤差給予更高的懲罰，適合評估模型對峰值的預測能力。空氣品質預測中，RMSE可反映預測的AQI或濃度值與實測的偏差程度，我們期望RMSE越小越好。 - **MAE (Mean Absolute Error，平均絕對誤差)**：計算誤差絕對值的平均，公式為$\frac{1}{n}\sum_i - y_i|$。相比RMSE，MAE對每個誤差給予相同比重，能反映模型預測的一般準確度而不被極端值過度影響。在空污預測中，MAE小表示模型整體預測值與實際值接近。 - }^n |\hat{y**R^2 決定係數**：在某些模型評估中我們也會引

用R^2來判斷模型解釋變異的能力（但對非線性模型而言需注意計算方式）。如前述XGBoost模型在相關研究中的$R^2>0.84$ [5]，我們也將計算本專案模型的R^2以比較。

我們會將上述指標以表格或圖表形式呈現，例如「表：各模型$PM_{2.5}$預測誤差比較」，包含測試集上的RMSE、MAE值，以便清晰比較哪種模型表現最佳。同時，也會針對不同期間計算指標，如分季節評估模型（檢視模型是否在冬季預測誤差明顯大於夏季）。倘若發現某模型在某條件下誤差偏高，將在報告中討論其原因（例如冬季污染較複雜難以預測所致）。整體而言，預測模型的目標是讓主要污染物濃度預測的RMSE控制在一個可接受的範圍內（例如$PM_{2.5}$濃度誤差在幾微克以內），以確保預報具有實用價值。

6.2 分群績效指標

對於污染源分群結果，我們將利用**聚類品質**指標來評估分群的有效性： - **輪廓係數 (Silhouette Coefficient)**：此指標綜合衡量群內緊密度與群間分離度。對於每個資料點計算輪廓值s，定義為$s = (b - a) / \max(a, b)$，其中a是該點與同群其他點的平均距離，b是該點與最近鄰群所有點的平均距離。s取值範圍在-1到1之間，越接近1表示該點與自身群組距離遠小於與其他群組，分群結果越好。我們會計算整體的平均輪廓係數作為模型性能指標。如平均輪廓係數接近或大於0.5，表示有明顯分群結構；若低於0或接近0則表示分群可能不明顯或有重疊。 - **群內統計與可解釋度**：除了數量指標，分群結果還需要在**語義上有意義**。我們會檢視每個群組的特徵均值、特徵分佈，判斷群組是否對應明確的污染源型態。如果某些群組缺乏明顯特徵，或不同群組特徵差異很小，則可能需要調整分群數或方法。我們也計劃列出每個群組中的測站清單，並檢查其空間分布是否具有聚集性，以此作為對分群合理性的輔助評估（例如若一群包含的測站遍布各種不同環境類型，則該群可能過於雜湊）。 - **對比分析**：將聚類結果與先驗知識對比也是重要手段。例如，我們知道某工業區測站群應該具有高SO_2特徵，若模型分群結果確實反映此點，則表示模型有效捕捉到了來源差異。我們可能針對已知幾個特殊污染源（如石化區、交通壅塞區）的測站預先標記類型，然後比較機器分群的結果與這些標記的吻合程度，作為效度檢驗。

總之，聚類績效指標不僅包括數值評估（如輪廓係數），更包含對結果合理性的討論。本專案將綜合這些面向，確保分群分析所得的分類對污染源區分確有參考價值，而非隨機分組。

6.3 異常偵測效果評估

異常檢測因其無監督性質，評估較具挑戰。我們將從以下角度描述異常偵測的效果： - **偵測率與誤報率**：如果手頭有已知的歷史異常事件（例如環保署公告的重大污染事件或明顯超標事件），可作為**真實異常**對照，評估模型是否成功偵測（偵測率）以及有無將正常事件錯判為異常（誤報）。例如某次沙塵暴事件期間全臺PM_{10}驟升，我們檢視Isolation Forest是否對該期間多數點給出高異常分數；又或者平常空品良好的日子模型不應報告大規模異常。我們將以案例方式列舉：若模型對某已知事件給出異常警示且無明顯漏報，即屬有效。 - **空間與時間覆蓋**：觀察模型偵測出的異常在空間和時間上的分布。我們期望異常事件多半是**短時且局部**的，或是**廣域但特定時間**的。若偵測結果顯示幾乎所有測站長期間都被標為異常，則模型參數可能太過敏感，需要調整。反之，若僅極少數點出現異常，可能錯失一些應關注的情況。我們會調節Isolation Forest的異常分數門檻或LOF鄰居數等參數，權衡偵測的靈敏度與專一性，並透過視覺化（第5.5節所述）來確認異常點的恰當性。 - **案例分析**：挑選若干典型的異常事件進行深入分析，以質性方式評估偵測效果。例如： - **區域性長程污染事件**：如模型偵測到某幾日全臺多個測站同步高$PM_{2.5}$，我們查證這可能是境外霧霾輸入事件。若與氣象後軌跡分析相符，證明模型有效抓住了此類異常 [7]。 - **局部性突發事件**：如某工業區測站在非典型時段SO_2飆高且僅限該區，模型將其標為異常。我們可查閱當地新聞或排放記錄，若確有事故排放發生則驗證了模型的實用性。 - **資料異常**：模型可能標記某測站長時間數值為極端低值或零值為異常，這通常暗示儀器故障或維護停機。我們將此類結果與環保署站點維護記錄比對，若符合則模型成功偵測資料品質問題。 - **F1分數**：在某些可以取得標註的情況下（例如我們人工標記了一些異常事件作測試），可計算異常檢測的Precision、Recall和F1分數。先前提到的研究顯示Isolation Forest檢測污染尖峰可達到F1>0.8 [5]，我們的目標是在重要異常事件的檢出上取得同級水準的精度。

最後，我們會綜合上述量化與質化結果，給出異常檢測模組效能的結論。例如，「模型成功偵測了全年發生的5起已知重大異常事件，未發生明顯誤報，顯示出良好的異常辨識能力」。若有不足（如對逐漸惡化的慢性事件不敏感），也將說明原因並提出改進建議（例如引入趨勢異常檢測方法等）。

7. 預期成果與研討方向

綜合上述分析方法，我們預期本專案將產出豐富的發現與成果，並引發進一步探討的議題，包括但不限於：

- **不同季節的污染變化趨勢**：透過時間序列分析，我們將量化春夏秋冬空氣品質的差異。預期結果是冬季因大氣擴散條件較差及境外污染輸入，$PM_{2.5}$等指標明顯升高，而夏季在對流旺盛及降雨沖刷下細懸浮微粒濃度相對較低 [3] ；臭氧則可能在陽光強烈的夏季出現較高峰值。這些季節性趨勢將在報告中以圖表清晰呈現，並可解釋其背後原因（如冬季頻繁出現極端高值與大氣穩定度相關 [3] ）。此成果可供研擬**季節性控污對策**參考，如在秋冬提前啟動減排措施。

- **各地區污染源差異**：藉由分群分析和地圖呈現，我們將揭示不同城市和區域間空氣污染成因的差異。可能的發現包括：都會區（如臺北都會區）因交通流量大，NO_2、CO等交通污染物水準普遍較高；重工業密集的地區（如雲嘉南沿海、高屏地區）硫氧化物(SO_2)及臭氧前驅物NOx濃度偏高，顯示工業排放影響明顯；相對地，東部離島等測站群組則維持較低的$PM_{2.5}$與NOx，空品較佳。這些**區域特性**將在群組特徵中體現，並可從地圖上看到城市中心污染濃度高於郊區、沿海工業帶高於山區背景值的空間梯度 [8] 。我們會討論這些差異的成因，如地形與盛行風向導致污染物易蓄積於特定盆地，或不同產業活動帶來的排放強度差別。同時，此部分成果也啟發**在地化**的空污治理建議，例如針對交通型污染源嚴重的都會區強化公共運輸和車流管理，針對工業源集中的區域落實排放管制。

- **高污染異常情境說明**：透過異常檢測，我們將鎖定數個全年中最顯著的污染異常事件並深入剖析。預期能夠說明的情境包括：**區域性沙塵暴或境外污染輸送**（例如每年春季可能出現的沙塵事件，導致全臺PM_{10}短時間飆升，我們將利用氣團後推等輔助資料佐證其來源）；**重大節日活動**（如春節期間因煙火或燒香造成局部$PM_{2.5}$升高，時間集中在除夕夜至清晨，我們將證明模型成功捉到這類短暫異常並量化其影響程度）；**工廠突發排放事件**（若資料期間內發生過某工安事故或非法排放導致某站污染激增，我們將分析該事件的監測數據特徵及模型偵測表現）。透過對這些案例的說明，我們不僅驗證模型的實用性，也對**極端污染事件**的形成機制和影響範圍有更深入的了解。這將引出討論，例如：沙塵暴對臺灣PM_{10}年均值的貢獻程度，未來是否需將境外傳輸納入空品標示考量；節慶活動的短期影響是否需要管制或提醒民眾防範；以及如何加強監控工業區以避免未經通報的排放。

- **模型限制與未來研究方向**：在成果討論部分，我們也將誠實面對本專案的限制，並提出後續研究建議。例如，本專案僅運用了基本的氣象參數（風速風向），未來可引入更多氣象因子（如大氣邊界層高度、降雨量）以及交通流量、排放清單等數據，以進一步提高預測精度和源解析能力。另外，時間序列模型方面可嘗試更先進的深度學習模型（如LSTM或Transformer）處理長序列依賴；分群分析可結合**來源解析**技術（如正矩陣因子分解 PMF）以更定量地對應污染源；異常檢測則可拓展至**多變量控制圖**或深度學習自編碼器的方法，以提升對複雜異常模式的辨識。這些延伸方向將在報告末提出，以供未來的研究或專案參考。

最後，本專案預期透過完整的資料科學流程，產出對臺灣空氣品質有價值的見解。例如，從結果中我們可以清楚指出「冬季污染較夏季高出X倍，主要由境外輸入和不利氣象條件導致」、「高雄某工業區的SO_2年均值為臺北市區的Y倍，顯示區域排放差異」、「全年共偵測到Z起重大異常事件，其中沙塵事件對PM_{10}月均值影響可達多少」等具體結論。這些發現將有助於環境單位制定科學根據的空污改善策略，也為後續學術研究提供了可進一步鑽研的課題。通過本期末專案的實踐，我們將深化對空氣品質數據的理解，並體現資料科學在環境領域中的強大應用價值。

參考資料：

- 臺灣環境部環境資料開放平臺: 空氣品質指標(AQI)逐時資料 [1] [2]
- Wang, Z. 等 (2023). 以上海為例的PM2.5時空預測研究：發現空氣污染呈現顯著季節周期，冬季極端高值頻率明顯高於其他季節 [3] ；市中心濃度高於郊區，向外遞減 [8] 。
- 某工業城市空氣品質AI預測系統: 結合XGBoost預測與Isolation Forest異常偵測，XGBoost預測 R^2>0.84，Isolation Forest檢測污染尖峰F1-score>0.80 [5] 。
- Austin, E. 等 (2013). 利用分群分析根據$PM_{2.5}$組成差異對監測站進行分類，所得群組反映不同排放源特性 [6] 。

---

[1] [2] 空氣品質指標(AQI)(歷史資料) | 環境部環境資料開放平臺
https://data.moenv.gov.tw/dataset/detail/aqx_p_488

[3] [8] A spatiotemporal XGBoost model for PM2.5 concentration prediction and its application in Shanghai - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10696222/

[4] GitHub - grtvishnu/Air-Pollution-Prediction-and-Forecasting: :octocat: Detection and Prediction of Air quality Index :octocat:
https://github.com/grtvishnu/Air-Pollution-Prediction-and-Forecasting

[5] AI-Enhanced air quality assessment and prediction in industrial cities: A case study of Kryvyi Rih, Ukraine
https://www.ecoeet.com/AI-Enhanced-Air-Quality-Assessment-and-Prediction-in-Industrial-Cities-A-Case-Study, 203725,0,2.html

[6] A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC3878877/

[7] PM2.5 anomaly detection for exceptional event demonstrations: A Texas case study - PubMed
https://pubmed.ncbi.nlm.nih.gov/39231245/