

大數據分析概觀與實務研討

影響空氣品質的因素之分析

四工管三乙 B10601104 許喬雅

四電機二甲 B10707032 楊政憲

四資工二甲 B10715029 陳彥瑋

四資工一 B10815034 吳承翰

目錄

一、	動機.....	3
二、	資料蒐集.....	4
三、	分析方法.....	7
(一)	數據分析.....	7
(二)	機器學習.....	7
四、	預期結果.....	7
五、	分析過程.....	8
(一)	資料採集.....	8
(二)	資料整理.....	10
(三)	數據分析.....	12
(四)	機器學習.....	23
六、	分析結果.....	26
七、	分析結論.....	29
八、	未來展望.....	30
九、	我們的看法.....	30
十、	參考資料.....	31
十一、	附錄.....	32
(一)	網頁爬蟲程式碼.....	32
(二)	資料整理程式碼.....	35
(三)	機器學習程式碼.....	38

一、動機

近年來環保意識提升，不只有民眾自發性的減塑、節電等活動，政府單位也提出了相應的政策與口號，例如：台中市政府核定台中火力發電廠全廠全年生煤使用量上限為 1104 萬噸，另要求中火在空污季減少 5 部機組用量，希望能因此改善台中的空氣品質^[1]。

另外，2018 年中華民國全國性公民投票(1124 公投)中有幾項與環境有關的題目，其中一項提案為：「您是否同意以平均每年至少降低 1%之方式逐年降低火力發電廠發電量？」提案的理由大致是因為火力發電造成了 PM2.5 的上升，進而影響了空氣的品質。其結果為：同意票 73.8%、不同意票 19.6%、無效票 6.6%^[2]，因此此提案通過並送交相關單位審查。

從上述兩例當中，可以發現空氣品質好像深受火力發電的影響，但是我們認為影響空氣品質的因素眾多，不可能藉由停止火力發電而能全面改善空氣品質。影響空氣品質的主因真的是火力發電嗎？如果不是的話，那又會是什麼？因此，我們想要找出真正影響空氣品質的元兇。

^[1] 資料來源: <https://news.ltn.com.tw/news/politics/breakingnews/2840027>

^[2] 公投結果來源為以下網址之「107 年全國性公民投票案第 7-16 案結果.pdf」：
<https://web.cec.gov.tw/central/cms/RefResults7to16>，此處之數據為同意票、不同意票、無效票人數分別占總實際投票人數之比例，與資料來源之同意票、不同意票分別占有效票數之比例的計算方式不同，並非數據之誤植。

二、資料蒐集

我們想探討影響空氣品質的關鍵變數，經過資料調查後，我們整理出幾項可能與空氣品質相關的變因並設定為自變量，包含：空氣、火力發電發電量、交通量、工業區數目與面積。另外，因變量我們設定為空氣品質指數(AQI)，以下將簡介各變量之介紹、選擇原因及資料來源。

(一) 天氣

1. 簡介：蒐集全台 22 縣市之歷年報表，並統計期每月資料，由於氣象站的監測值眾多，經評估後我們選擇五項可能與空氣品質相關的監測值來做分析，包含：氣溫、降水量、相對溼度、氣壓、風速。
2. 選擇原因：猜測污染源濃度可能與空氣對流或境外擴散(季風)有關，期望找到各地之天氣評估數值與空氣品質間的相關性。
3. 資料來源：交通部中央氣象局 觀測資訊查詢系統 (網址：<https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp>)

(二) 火力發電發電量

1. 簡介：台灣電力公司火力發電各機組歷年發電量，統計其每月總量，因為只有 10 個縣市設有台電的火力發電廠，因此我們統計此 10 個縣市的資料，其他縣市不做分析。
2. 選擇原因：空氣品質似乎與火力發電的比例息息相關，期望找到火力發電各機組發電量與其所在位置之空氣品質的相關性。
3. 資料來源：臺灣電力公司 過去電力供需資訊 (網址：<https://www.taipower.com.tw/tc/page.aspx?mid=210>)

(三) 交通量

1. 簡介：各縣市公路交通量統計表，以縣市 9 年之每日平均車流量評估(臺北市、金門縣、連江縣資料缺失)。
2. 選擇原因：猜測空氣品質可能與汽機車廢氣有關係，期望找到各縣市不同的交通量對與空氣品質，有沒有直接的影響性。
3. 資料來源：交通部公路總局 公路交通量調查統計 (網址：https://www.thb.gov.tw/sites/ch/modules/download/download_list?node=bc520be-3e03-4e28-b4cb-7e338ed6d9bd&c=83baff80-2d7f-4a66-9285-d989f48effb4)

(四) 工業區數目與面積

1. 簡介：以各縣市當前工業區數目、面積及佔各縣市面積比例評估
2. 選擇原因：猜測空氣品質可能與工業區排放之廢氣有關聯，期望找到各縣市不同的工業區面積與比例對空氣品質的影響。
3. 資料來源：台灣工業用地供給與服務資訊網 工業區統計 (網址：<https://idbpark.moeaidb.gov.tw/Environ/statistical>)

(五) 空氣品質指數

1. 簡介：空氣品質指數(Air Quality Index, AQI)是政府從 2016 年 12 月開始實施的空氣品質監測標準，測量的污染項目有六種，分別是臭氧(O₃)、細懸浮微粒(PM_{2.5})、懸浮微粒(PM₁₀)、一氧化碳(CO)、二氧化硫(SO₂)及二氧化氮(NO₂)，其能更全面地監測多項污染物。
2. 計算方式：將不同污染物對人體健康的影響程度分別換算出副指標值，再以當日各副指標之最大值為該測站當日之 AQI，以確保不會因為只關注單一指標而忽視了其他污染物造成的影響。

3. 舉例：如果今天的 $PM_{2.5}$ 小於 15.4，落在「良好」的綠色區間，但是 O_3 的小時平均值是 0.166，落在「對所有族群不健康」的紅色區間，那 AQI 會取最高值，顯示為「對所有族群不健康」的紅色區間。下圖為**污染物濃度及污染副指標對照表**（圖一），可以以此為對照。

空氣品質指標 (AQI)							
AQI 指標	O_3 (ppm) 8小時平均值	O_3 (ppm) 小時平均值 ⁽¹⁾	$PM_{2.5}$ ($\mu g/m^3$) 24小時平均值	PM_{10} ($\mu g/m^3$) 24小時平均值	CO (ppm) 8小時平均值	SO ₂ (ppb) 小時平均值	NO ₂ (ppb) 小時平均值
良好 0~50	0.000 - 0.054	-	0.0 - 15.4	0 - 54	0 - 4.4	0 - 35	0 - 53
普通 51~100	0.055 - 0.070	-	15.5 - 35.4	55 - 125	4.5 - 9.4	36 - 75	54 - 100
對敏感族群 不健康 101~150	0.071 - 0.085	0.125 - 0.164	35.5 - 54.4	126 - 254	9.5 - 12.4	76 - 185	101 - 360
對所有族群 不健康 151~200	0.086 - 0.105	0.165 - 0.204	54.5 - 150.4	255 - 354	12.5 - 15.4	186 - 304 ⁽³⁾	361 - 649
非常不健康 201~300	0.106 - 0.200	0.205 - 0.404	150.5 - 250.4	355 - 424	15.5 - 30.4	305 - 604 ⁽³⁾	650 - 1249
危害 301~400	⁽²⁾	0.405 - 0.504	250.5 - 350.4	425 - 504	30.5 - 40.4	605 - 804 ⁽³⁾	1250 - 1649
危害 401~500	⁽²⁾	0.505 - 0.604	350.5 - 500.4	505 - 604	40.5 - 50.4	805 - 1004 ⁽³⁾	1650 - 2049

- 一般以臭氧 (O_3) 8小時值計算各地區之空氣品質指標 (AQI)。但部分地區以臭氧 (O_3) 小時值計算空氣品質指標 (AQI) 是更具有預警性，在此情況下，臭氧 (O_3) 8小時與臭氧 (O_3) 1小時之空氣品質指標 (AQI) 則皆計算之，取兩者之最大值作為空氣品質指標 (AQI)。
- 空氣品質指標 (AQI) 301以上之指標值，是以臭氧 (O_3) 小時值計算之，不以臭氧 (O_3) 8小時值計算之。
- 空氣品質指標 (AQI) 200以上之指標值，是以二氧化硫 (SO_2) 24小時值計算之，不以二氧化硫 (SO_2) 小時值計算之。

污染物濃度及污染副指標對照表（圖一）

4. 資料來源：行政院環境保護署 環保統計查詢網（網址：<https://stat.epa.gov.tw/>）

三、分析方法

- (一) 數據分析：蒐集各變因的資料，並做整理把無效資料刪除，使用 excel 得到各自變量與 AQI 的相關矩陣，並從相關係數判斷各項自變量與 AQI 的相關性。另外，也可以使用自變量對比 AQI 的數據繪製直方圖或折線圖，觀察其中的關係與趨勢。
- (二) 機器學習：使用邏輯回歸 (Logistic Regression) 預測變因在特定情況下時，會有怎樣的結果，並且使用準確度 (Accuracy) 與召回率 (Recall) 評估機器學習的成效。

四、預期結果

- (一) 火力發電與 AQI 無關聯：氣候也是影響空氣品質的因素之一，我們認為有許多污染物是隨著東北季風而來，把源自於中國的黃土與污染粒子帶來台灣，因此需考慮境外的污染物排放量與氣候擴散因素。
- (二) 污染源大多來自境內，而電力帶來的污染只佔其中的一小部分：我國污染源可分為境外與境內來源，其中境外來源約佔 34 ~ 40%，而境內污染源約佔 60 ~ 66%。在境內污染源中，移動源影響約為 30 ~ 37%、工業源約為 27 ~ 31%、其他約為 32 ~ 43%，而電力業僅佔境內污染源 10%! 因此我們認為製電過程中對空氣造成的污染應該只佔一小部分而已。
- (三) 空氣污染有季節性：我國空污應該有季節性，且多集中在 10~3 月，因為東北季風的季節大致為秋季到春季之間，因此我們推論空氣污染應在 10~3 月較為嚴重。

五、分析過程

（一） 資料採集

在取得資料的部分我們使用了兩種方式，分別是直接下載以及網頁爬蟲，以下是細節的部分。

1. 直接下載

除了天氣資料之外，其餘的資料都只需要進行少數的點擊就足夠滿足我們的需求，因此這部分的資料非常容易取得。

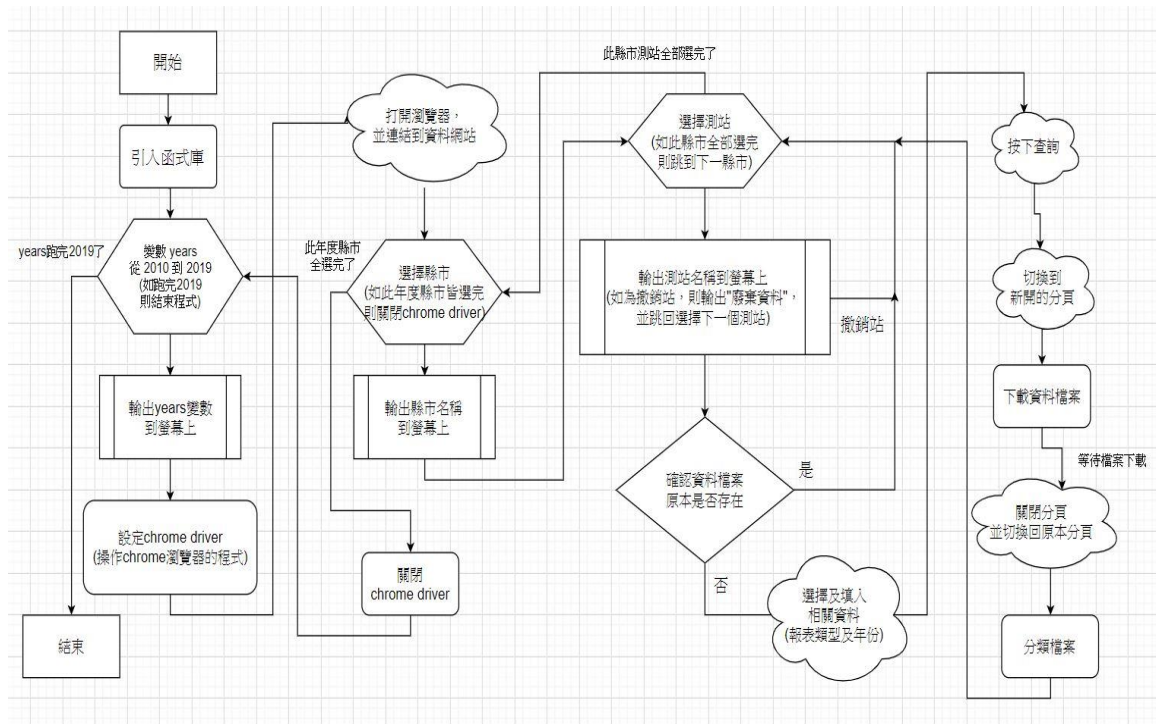
2. 網頁爬蟲

對於天氣資料而言，因為我們取得資料的網頁一次只能取得特定年份且特定天氣測站的資料，如果以直接下載的方式會變得非常的繁瑣，因此我們透過 **Python** 語言使用網頁爬蟲的技術自動地進行資料的下載，以下是詳細的說明。

在（十一）附錄、（一）網頁爬蟲程式碼中，我們使用了 **os**, **time**, **selenium** 函式庫。**selenium** 是用來自動化操作網頁的一個函式庫，我們主要是用來做資料檔案的自動下載。而 **time** 函式庫是用來取得系統時間，我們是用在等待檔案下載的部分。**os** 函式庫是一個用來控制系統指令的函式庫，我們是做檔案下載後的自動分類部分。

以下是程式的執行流程^[1]，詳細的程式碼在附錄(十一)、（一）網頁爬蟲程式碼中。

^[1] 完整大圖: <https://franktaiwan123.github.io/getWeatherFlowChart/Flow.html>



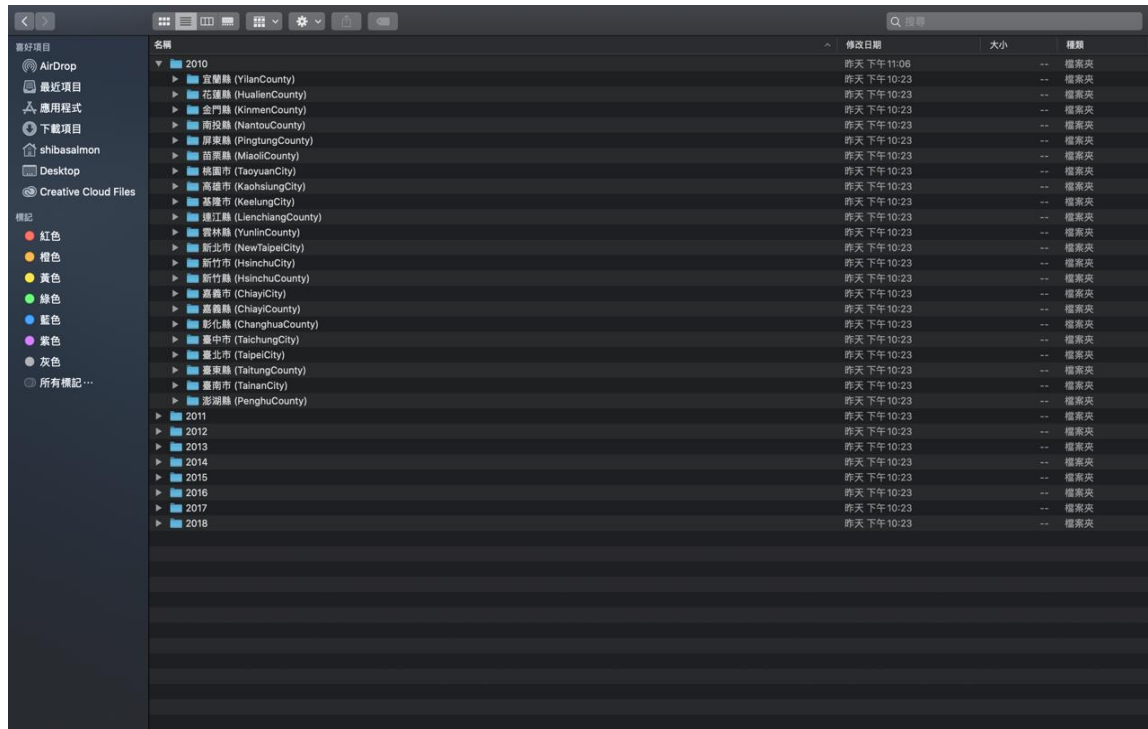
流程圖中的方框是一般的程式處理，六角形框是迴圈的部分，左
 右有雙線的方框是程式的輸出，雲形狀框是 selenium 操作瀏覽器的部
 分，菱形框是條件處理。

(二) 資料整理

1. 資料整理選擇：我們用於機器學習的資料，必須將其做整理。於是我們將「天氣資訊」利用 Python 整理到跟 AQI 資訊的格式一模一樣，如下圖

時間	縣市	AQI大於100之比率(%)	AQI大於100之日數
2010年 1月	臺北市	0	0
	新北市	0	0
	桃園市	0	0
	基隆市	0	0
	新竹縣	0	0
	新竹市	0	0
	苗栗縣	0	0
	臺中市	1.29	2
	彰化縣	1.61	1
	南投縣	3.23	2
	雲林縣	4.84	3
	嘉義縣	8.06	5
	嘉義市	9.68	3
	臺南市	8.87	11
	高雄市	8.15	19
	屏東縣	8.79	8
	宜蘭縣	0	0
	花蓮縣	0	0
	臺東縣	0	0
	金門縣	0	0
	連江縣	0	0
	澎湖縣	0	0
2010年 2月	臺北市	0.71	1
	新北市	0.79	2
	桃園市	0	0
	基隆市	0	0
	新竹縣	0	0
	新竹市	0	0
	苗栗縣	0	0
	臺中市	0	0
	彰化縣	0	0
	南投縣	0	0
	雲林縣	0	0
	嘉義縣	0	0
	嘉義市	0	0
	臺南市	0	0
	高雄市	0	0
	屏東縣	0	0
	宜蘭縣	0	0
	花蓮縣	0	0
	臺東縣	0	0

不難看出 AQI 的資訊是以「年份-月份-縣市」按順序排列，我們必須將現有的天氣資訊合併處理成一個檔案，並按照上述排列，如下圖



2. 整理結果：最後可得整理結果如下

Object	Mouth	Temperature	Precp	RH	StnPres	WS
2010-臺北市-1	1	15.18	127.44	81.91	992.55	2.16
2010-新北市-1	1	15.32	191.78	85.24	1001.76	2.05
2010-桃園市-1	1	15.15	120.56	76.33	999.78	2.07
2010-基隆市-1	1	16.25	197.45	80	1013.65	5.55
2010-新竹縣-1	1	14.04	100.99	79.67	977.69	1.38
2010-新竹市-1	1	15.6	71.5	81	1019.8	1
2010-苗栗縣-1	1	14.5	72	79.75	988.22	1.38
2010-臺中市-1	1	13.47	40.76	76.5	1013.6	2.73
2010-彰化縣-1	1	16.8	23.17	0	0	2.25
2010-南投縣-1	1	8.84	50.8	70.5	773.95	2.04
2010-雲林縣-1	1	15.77	24.33	0	0	1.3
2010-嘉義縣-1	1	12.37	29.52	82	875.8	0.67
2010-嘉義市-1	1	17.2	19.5	75	1016.6	2
2010-臺南市-1	1	17.73	13.45	70.5	1012.68	2.07
2010-高雄市-1	1	18.98	18.75	72	1018.13	0.9
2010-屏東縣-1	1	18.86	30.19	68	998.38	2.04
2010-宜蘭縣-1	1	15.29	191.28	80	980.92	1.95
2010-花蓮縣-1	1	15.29	95.93	75	967.32	1.22
2010-臺東縣-1	1	16.57	69.17	75.5	963.34	2.18
2010-金門縣-1	1	13.3	25.6	75	1019.8	3.2
2010-連江縣-1	1	10.5	41.3	88	1013	3.5
2010-澎湖縣-1	1	18.15	6.4	75.5	1017.05	6.95
2010-臺北市-2	2	16.08	186.15	87	989.01	1.79
2010-新北市-2	2	16.16	275.06	90.32	998.06	1.68
2010-桃園市-2	2	15.78	236.11	85.33	996.3	1.62
2010-基隆市-2	2	16.7	244.2	86.5	1009.55	5
2010-新竹縣-2	2	15.18	269.25	85.44	974.56	1.34
2010-新竹市-2	2	16.5	208	86	1016.4	1.2
2010-苗栗縣-2	2	15.78	236.73	84.25	985.03	1.48
2010-臺中市-2	2	15.1	139.22	81	1010.4	3.17
2010-彰化縣-2	2	18.4	74.33	0	0	2.25
2010-南投縣-2	2	10.09	156.13	78.5	772.55	2.62
2010-雲林縣-2	2	17.63	72.33	0	0	1.7
2010-嘉義縣-2	2	14.47	102.8	79	874	0.93
2010-嘉義市-2	2	19.1	53.8	77	1013.4	2.3
2010-臺南市-2	2	19.78	39.95	74.5	1009.52	2.2
2010-高雄市-2	2	20.78	60.75	74	1015.1	0.93
2010-屏東縣-2	2	20.61	19.35	76	995.44	1.49
2010-宜蘭縣-2	2	17.02	167.11	82	976.94	1.88
2010-花蓮縣-2	2	17.17	134.75	77	963.55	1.28
2010-臺東縣-2	2	18.76	59.32	79.75	959.84	2

（三）數據分析

1. 天氣

在分析的過程中，我們首先將天氣資料以及 AQI 資料做成表格，以方便進行資料的排序及篩選。

再來，我們將做分析用的表格（以下簡稱為分析表格）的資料編排格式訂為先將單一年月份的 22 縣市排完再排下一月份 22 縣市的資料欄位格式^[1]，如下圖所示。

時間	縣市	A
2010年 1月	臺北市	
2010年 1月	新北市	
2010年 1月	桃園市	
2010年 1月	基隆市	
2010年 1月	新竹縣	
2010年 1月	新竹市	
2010年 1月	苗栗縣	
2010年 1月	臺中市	
2010年 1月	彰化縣	
2010年 1月	南投縣	
2010年 1月	雲林縣	
2010年 1月	嘉義縣	
2010年 1月	嘉義市	
2010年 1月	臺南市	
2010年 1月	高雄市	
2010年 1月	屏東縣	
2010年 1月	宜蘭縣	
2010年 1月	花蓮縣	
2010年 1月	臺東縣	
2010年 1月	金門縣	
2010年 1月	連江縣	
2010年 1月	澎湖縣	
2010年 2月	臺北市	
2010年 2月	新北市	
2010年 2月	桃園市	
2010年 2月	基隆市	

在決定完分析表格格式之後，我們將資料整理進此格式，下圖是部分的資料，圖中資料欄位的 AQI 大於 100 之比例(%)是指當月此縣市

^[1] 舉例來說，就是先排西元 2010 年 1 月的所有縣市之後再排西元 2010 年 2 月的所有縣市，依序下去

所有測站測出 AQI 大於 100 (不良)的總日數加總後，再除以當月此縣市所有測站的總運作日數並取百分比表示^[2]。

其他欄位如 Temperature 是溫度、Precp 是降水量、RH 是濕度、StnPres 是測站氣壓、WS 是風速。

時間	縣市	AQI大於100之比率(%)	Temperature	Precp	RH	StnPres	WS
2010年 1月	臺北市	0	15.18	127.44	81.91	992.55	2.16
2010年 1月	新北市	0	15.32	191.78	85.24	1001.76	2.05
2010年 1月	桃園市	0	15.15	120.56	76.33	999.78	2.07
2010年 1月	基隆市	0	16.25	197.45	80	1013.65	5.55
2010年 1月	新竹縣	0	14.04	100.99	79.67	977.69	1.38
2010年 1月	新竹市	0	15.6	71.5	81	1019.8	1
2010年 1月	苗栗縣	0	14.5	72	79.75	988.22	1.38
2010年 1月	臺中市	1.29	13.47	40.76	76.5	1013.6	2.73
2010年 1月	彰化縣	1.61	16.8	23.17	0	0	2.25
2010年 1月	南投縣	3.23	8.84	50.8	70.5	773.95	2.04
2010年 1月	雲林縣	4.84	15.77	24.33	0	0	1.3
2010年 1月	嘉義縣	8.06	12.37	29.52	82	875.8	0.67
2010年 1月	嘉義市	9.68	17.2	19.5	75	1016.6	2
2010年 1月	臺南市	8.87	17.73	13.45	70.5	1012.68	2.07
2010年 1月	高雄市	8.15	18.98	18.75	72	1018.13	0.9
2010年 1月	屏東縣	8.79	18.86	30.19	68	998.38	2.04
2010年 1月	宜蘭縣	0	15.29	191.28	80	980.92	1.95
2010年 1月	花蓮縣	0	15.29	95.93	75	967.32	1.22
2010年 1月	臺東縣	0	16.57	69.17	75.5	963.34	2.18
2010年 1月	金門縣	0	13.3	25.6	75	1019.8	3.2
2010年 1月	連江縣	0	10.5	41.3	88	1013	3.5
2010年 1月	澎湖縣	0	18.15	6.4	75.5	1017.05	6.95
2010年 2月	臺北市	0.71	16.08	186.15	87	989.01	1.79
2010年 2月	新北市	0.79	16.16	275.06	90.32	998.06	1.68
2010年 2月	桃園市	0	15.78	236.11	85.33	996.3	1.62
2010年 2月	基隆市	0	16.7	244.2	86.5	1009.55	5

在填完資料之後，我們使用相關矩陣來分析 AQI 大於 100 之比例與其他天氣指標的關聯性，下圖是運算出來的結果。

	AQI大於100之比率(%)	Temperature	Precp	RH	StnPres	WS
AQI大於100之比率(%)	1					
Temperature	-0.081446208	1				
Precp	-0.206462914	0.281359408	1			
RH	-0.02081299	0.045355338	0.21029872	1		
StnPres	0.047261075	0.095109633	0.01273313	0.790082	1	
WS	0.02919699	-0.094565244	-0.15140076	-0.02626	0.121231	1

^[2] 舉例來說，西元 2016 年 5 月台北市 5 個測站中的中山、古亭、萬華在當月分別有 1 天、2 天、2 天的 AQI 超過 100，另外的 2 個測站在當月的 AQI 皆沒有任何一日超過 100，且所有測站在當月的 31 天內皆有運作，所以 2016 年 5 月台北市 AQI 大於 100 之比例為 $(1 + 2 + 2) / (31 * 5) \approx 0.0323 = 3.23\%$ ，該欄位之數值為 3.23。

在分析完整體的資料後，我們也分別對單一縣市做同樣的分析，

以下是其中一縣市的結果。

	<i>AQI大於100之比率(%)</i>	<i>Temperature</i>	<i>Precp</i>	<i>RH</i>	<i>StnPres</i>	<i>WS</i>
AQI大於100之比率(%)	1					
Temperature	-0.227668173	1				
Precp	-0.276304826	0.261483008	1			
RH	-0.253435652	0.424618275	0.511556607	1		
StnPres	0.189847728	-0.920973821	-0.38013713	-0.64862	1	
WS	0.610797884	-0.26085483	-0.33522255	-0.35813	0.311367	1

2. 交通量

對於交通量來說，因為原始資料形式的緣故，我們將分析表格的資料順序訂為以各縣市來區分，如下圖所示，此圖已將缺失交通量的縣市排除。

縣市	
新北市	
桃園市	
基隆市	
新竹縣	
新竹市	
苗栗縣	
臺中市	
彰化縣	
南投縣	
雲林縣	
嘉義縣	
嘉義市	
臺南市	
高雄市	
屏東縣	
宜蘭縣	
花蓮縣	
臺東縣	
澎湖縣	

決定了分析表格排序方式之後，我們將資料的 AQI 欄位又另外分為 2010~2012、2013~2015、2016~2018 這三個時間區間的平均每月 AQI 大於 100 比例(%)，也就是將「每月 AQI 大於 100 比例(%)」這一個 AQI 的原始項目再依據時間的範圍做一個平均^[1]，原因是因為我們要避免掉直接將此 AQI 項目以 9 年的時間做平均可能造成出來的數值受單一年度的極端值影響的風險。

^[1] 舉例來說，2013~2015 台北市的平均每月 AQI 大於 100 比例(%)計算的方式就是將台北市在 2013~2015 期間的「每月 AQI 大於 100 比例(%)」這一個項目做算術平均

決定完了資料的格式之後，我們將資料整理之後填入分析表格，
 下圖是填完後的結果，圖中下方的部分是缺乏平均每日車流量的縣市。

縣市	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	每日平均車流量
新北市	1.08	0.61	5.62	2879092
桃園市	0.36	0.2	6.58	1806048
基隆市	0.09	0	3.39	339965
新竹縣	0.36	0.41	5.75	517476
新竹市	0.27	0.19	6.37	73081
苗栗縣	0.32	0.18	7.56	620031
臺中市	1.07	0.76	11.26	1926420
彰化縣	0.78	0.37	11.82	1217233
南投縣	1.46	0.92	19.3	774656
雲林縣	1.37	1.51	20.78	558184
嘉義縣	2.32	2.38	17.53	588689.5
嘉義市	1.83	1.19	20.4	114420
臺南市	1.6	1.69	17.61	1572536
高雄市	3.77	2.82	23.22	1483805
屏東縣	3.73	3.2	18.41	1036275
宜蘭縣	0.13	0	1.1	663735
花蓮縣	0.18	0	0.91	382737
臺東縣	0.27	0.09	0.27	166757
澎湖縣	0	0	4.65	47322
無法處理				
金門縣	0.73	2.71	18.31	
連江縣	0	1.08	15.21	
臺北市	1.07	0.54	3.52	

填完資料後，我們一樣使用相關矩陣來分析他們的關聯性，下圖是運算後的結果。

	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	每日平均車流量
2010-2012每月平均AQI大於100比例(%)	1			
2013-2015每月平均AQI大於100比例(%)	0.972688908	1		
2016-2018每月平均AQI大於100比例(%)	0.829980078	0.830645205	1	
每日平均車流量	0.27608534	0.222199957	0.156711044	1

3. 火力發電量

我們定義 AQI 與火力發電量分析表格的順序和 AQI 與天氣的順序是相同的，在此不再贅述。

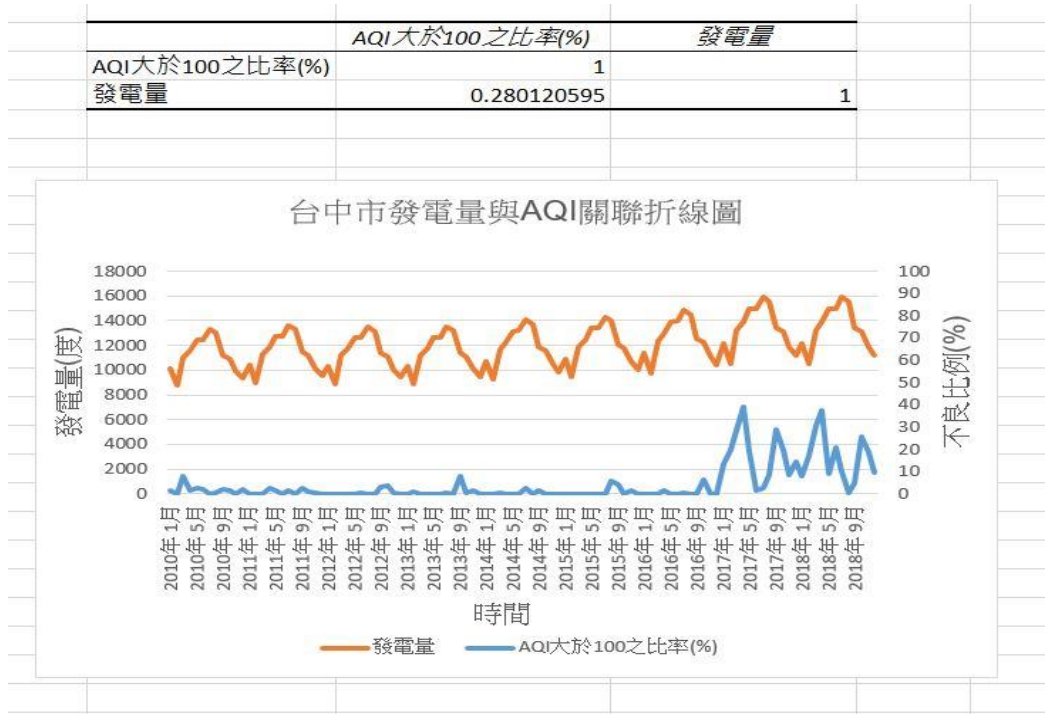
定義完之後，我們將資料填入，下圖是填入後的結果。

時間	縣市	AQI大於100之比率(%)	發電量
2010年 1月	新北市	0.0	3944.3
2010年 1月	桃園市	0.0	7944.0
2010年 1月	苗栗縣	0.0	3576.7
2010年 1月	臺中市	1.3	10166.7
2010年 1月	雲林縣	4.8	4100.4
2010年 1月	高雄市	8.2	12836.5
2010年 1月	花蓮縣	0.0	1532.1
2010年 1月	金門縣	0.0	96.6
2010年 1月	連江縣	0.0	41.4
2010年 1月	澎湖縣	0.0	137.9
2010年 2月	新北市	0.8	3464.4
2010年 2月	桃園市	0.0	6540.3
2010年 2月	苗栗縣	0.0	4225.8
2010年 2月	臺中市	0.0	8754.4
2010年 2月	雲林縣	0.0	3387.8
2010年 2月	高雄市	0.0	11348.0
2010年 2月	花蓮縣	0.0	707.7
2010年 2月	金門縣	0.0	88.7
2010年 2月	連江縣	0.0	38.0
2010年 2月	澎湖縣	0.0	126.8
2010年 3月	新北市	8.2	2982.0
2010年 3月	桃園市	8.1	9574.5
2010年 3月	苗栗縣	6.5	4711.5

填入資料後，我們一樣使用相關矩陣進行分析，以下是相關矩陣的內容。

	AQI大於100之比率(%)	發電量
AQI大於100之比率(%)	1	
發電量	0.115178681	1

做完整體的分析之後，我們也使用同樣的方法對單一縣市做分析，並且也將火力發電量及 AQI 大於 100 比例(不良比例)依據時間軸製作出折線圖，以下是其中一縣市的相關矩陣及折線圖。



4. 工業區數目與面積

我們定義此項目分析表格的排序方式和 AQI 與交通量的表格順序是相同的，皆是以縣市做分類。

和交通量一樣，在工業區這一個比較項目中我們也將 AQI 分成 3 個時間區段，而我們也將工業區在各縣市的數量、面積、占縣市面積比例列入分析表格的欄位中。

下圖是填入資料後的結果。

A	B	C	D	E	F	G
縣市	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	工業區數目	面積(平米)	佔各縣市面積比例(%)
臺北市	1.07	0.54	3.52	5	3057900	1.13%
新北市	1.08	0.61	5.62	51	23956002	1.17%
桃園市	0.36	0.2	6.58	32	45661947	3.74%
基隆市	0.09	0	3.39	2	1001600	0.75%
新竹縣	0.36	0.41	5.75	5	5930000	0.42%
新竹市	0.27	0.19	6.37	1	6250000	6.00%
苗栗縣	0.32	0.18	7.56	10	9001342	0.49%
臺中市	1.07	0.76	11.26	16	22477591	1.01%
彰化縣	0.78	0.37	11.82	10	78143693	7.27%
南投縣	1.46	0.92	19.3	4	7096400	0.1728116547%
雲林縣	1.37	1.51	20.78	8	178469957	13.8259567507%
嘉義縣	2.32	2.38	17.53	8	7249200	0.3808079556%
嘉義市	1.83	1.19	20.4	0	0	0.00%
臺南市	1.6	1.69	17.61	14	29388943	1.3409486656%
高雄市	2.77	2.87	22.22	18	40637086	1.2770374128%

填入資料後，與之前相同，我們也使用相關矩陣進行關聯性的分析，以下是相關矩陣內容。

	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	工業區數目	面積(平米)	佔各縣市面積比例(%)
2010-2012每月平均AQI大於100比例(%)	1					
2013-2015每月平均AQI大於100比例(%)	0.836153427	1				
2016-2018每月平均AQI大於100比例(%)	0.721574433	0.825855343	1			
工業區數目	0.174874264	0.008827665	-0.032374474	1		
面積(平米)	0.158300072	0.123836884	0.337401856	0.233449224	1	
佔各縣市面積比例(%)	-0.001350565	-0.02510639	0.199030262	0.104539354	0.91308696	1

5. 地理區域

對於地理區域的資料，我們是將 AQI 的資料整理成在相同的月份
各縣市的 AQI 大於 100 比例，每一行是 2010~2018 的每個月份，每一
列是各個縣市的 AQI 大於 100 比例。

以下是部分的資料。

時間	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
時間(AQI不良比例%)	臺北市	新北市	桃園市	基隆市	新竹縣	新竹市	苗栗縣	臺中市	彰化縣	南投縣	雲林縣	嘉義縣	嘉義市	臺南市	高雄市	屏東縣	宜蘭縣	花蓮縣
2010年1月		0	0	0	0	0	0	0	1.29	1.61	3.23	4.84	8.06	9.68	8.87	8.15	8.79	0
2010年2月		0.71	0.79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2010年3月		9.03	8.24	8.06	3.23	8.06	6.45	6.45	7.74	8.06	6.45	9.68	11.29	9.68	10.48	11.69	8.6	4.84
2010年4月		3.33	3.7	0.83	0	0	0	1.67	1.33	1.67	1.67	1.67	1.67	0	0.83	2.08	4.44	0
2010年5月		2.58	1.08	0.81	0	0	0	3.23	2.58	0	3.23	0	0	0	0.81	3.23	3.23	0
2010年6月		0.67	0.37	0	0	0	0	0	2	1.67	0	0	0	0	0	2.92	2.22	0
2010年7月		0	0.36	0	0	0	0	0	0	0	0	0	0	0	0	2.43	0	0
2010年8月		0	1.08	0	0	0	0	0	0.68	0	1.61	0	3.23	0	0	2.42	2.15	0
2010年9月		0.67	1.48	0	0	0	0	0	2	0	0	3.33	0	0	0	2.5	1.11	0
2010年10月		0	0	0	0	0	0	0	1.29	0	1.61	3.23	0	0	3.23	10.89	9.68	0
2010年11月		0	0	0	0	0	0	0	0	0	0	1.67	10	0	1.67	5.83	5.56	0
2010年12月		0	0.36	0	0	0	0	0	1.94	1.61	3.23	1.61	11.29	16.13	8.87	7.26	4.3	0
2011年1月		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.23	0
2011年2月		0	0	0	0	0	0	0	0	0	3.57	1.79	1.79	10.71	8.04	6.7	5.95	0
2011年3月		0	1.79	0	0	0	0	0	0	0	3.23	0	1.61	3.23	0	6.85	8.6	0
2011年4月		0	1.85	0	0	0	3.33	0	2.67	3.33	6.67	6.67	6.67	13.33	2.5	6.25	3.33	0
2011年5月		3.23	3.94	0	0	0	0	0	1.29	0	0	3.23	0	0	0	1.61	2.15	0
2011年6月		0	0.74	0	0	0	0	0	0	0	0	0	0	0	0	0	1.11	0
2011年7月		0	0	0	0	1.61	0	0	1.29	0	0	0	0	0	0	0.81	1.08	0
2011年8月		2.58	1.43	0	0	0	0	0	0	0	0	0	0	0	0	0.4	2.17	0
2011年9月		4.67	3.33	1.67	0	0	0	0	2.5	3.33	3.33	0	6.67	0	0.83	7.92	8.89	0
2011年10月		0	0	0	0	0	0	0	0.81	0	1.61	0	1.61	0	1.61	6.85	11.83	0
2011年11月		0	0	0	0	0	0	0	0.68	1.67	1.67	0	1.67	0	1.67	6.25	2.22	0
2011年12月		0	0	0	0	0	0	0	0	0	0	1.61	0	0	0	0.4	1.08	0
2012年1月		0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.62	0	0

整理了資料之後，我們仍然是使用相關矩陣進行關聯性分析，以下是相關矩陣的內容。

	臺北市	新北市	桃園市	基隆市	新竹縣	新竹市	苗栗縣	臺中市	彰化縣	南投縣	雲林縣	嘉義縣	嘉義市	臺南市	高雄市	屏東縣	宜蘭縣	花蓮縣	臺東縣	金門縣	連江縣	澎湖縣
臺北市	1																					
新北市	0.926354	1																				
桃園市	0.811313	0.884271	1																			
基隆市	0.806292	0.900943	0.813932	1																		
新竹縣	0.764199	0.830905	0.963383	0.787842	1																	
新竹市	0.610952	0.698307	0.89563	0.618938	0.898627	1																
苗栗縣	0.727932	0.802449	0.929114	0.749736	0.92081	0.865214	1															
臺中市	0.779899	0.883468	0.933795	0.822477	0.910637	0.821922	0.92625	1														
彰化縣	0.695392	0.8122	0.939656	0.749488	0.925843	0.894187	0.893374	0.928741	1													
南投縣	0.620881	0.759063	0.830972	0.710348	0.828081	0.807863	0.777247	0.875171	0.898806	1												
雲林縣	0.674977	0.806467	0.884301	0.750962	0.865122	0.843975	0.841501	0.924999	0.944417	0.968546	1											
嘉義縣	0.620032	0.758502	0.878414	0.709581	0.865207	0.850832	0.840997	0.896103	0.944394	0.932897	0.965404	1										
嘉義市	0.567146	0.712716	0.793207	0.655974	0.784237	0.775814	0.736303	0.841667	0.878974	0.974195	0.959482	0.940047	1									
臺南市	0.534369	0.676567	0.770598	0.617236	0.757927	0.761987	0.722163	0.817852	0.874031	0.961936	0.951087	0.941826	0.983616	1								
高雄市	0.444383	0.582663	0.707859	0.520628	0.702745	0.749241	0.691384	0.753392	0.823112	0.937456	0.906401	0.892457	0.941155	0.96162	1							
屏東縣	0.433948	0.56063	0.686591	0.489903	0.672735	0.716121	0.657064	0.727243	0.766841	0.911921	0.864701	0.854578	0.908956	0.932785	0.97539	1						
宜蘭縣	0.790027	0.799346	0.761459	0.806092	0.758666	0.581262	0.70372	0.751871	0.677633	0.646229	0.683483	0.663427	0.620674	0.606643	0.5198	0.512899	1					
花蓮縣	0.778084	0.784558	0.655664	0.753068	0.592917	0.431965	0.586619	0.641771	0.52841	0.477207	0.522366	0.504324	0.439643	0.412378	0.321486	0.300863	0.834395	1				
金門縣	1.138919	0.07463	0.04895	0.058014	0.006685	0.024301	0.035055	0.043526	0.000849	0.047853	0.029553	0.062941	0.008109	0.011746	0.024924	0.033689	0.162595	0.216118	1			
澳門	0.602271	0.73418	0.817391	0.711711	0.812699	0.761271	0.746543	0.813304	0.866663	0.873125	0.900976	0.875759	0.886329	0.836506	0.78515	0.635019	0.500215	0.060575		1		
連江縣	0.680564	0.806377	0.811683	0.827068	0.783777	0.695242	0.748965	0.821645	0.843457	0.840946	0.873446	0.840963	0.8243	0.808308	0.777502	0.733469	0.717432	0.616128	0.10719	0.916742	1	
澎湖縣	0.487078	0.611104	0.623463	0.617741	0.563677	0.41699	0.5953	0.653765	0.654877	0.580077	0.631772	0.641679	0.60045	0.63183	0.571163	0.552577	0.665966	0.562455	0.061077	0.643541	0.754436	1
									>=0.7	高度相關												
									>=0.85	特別相關												

6. 月份

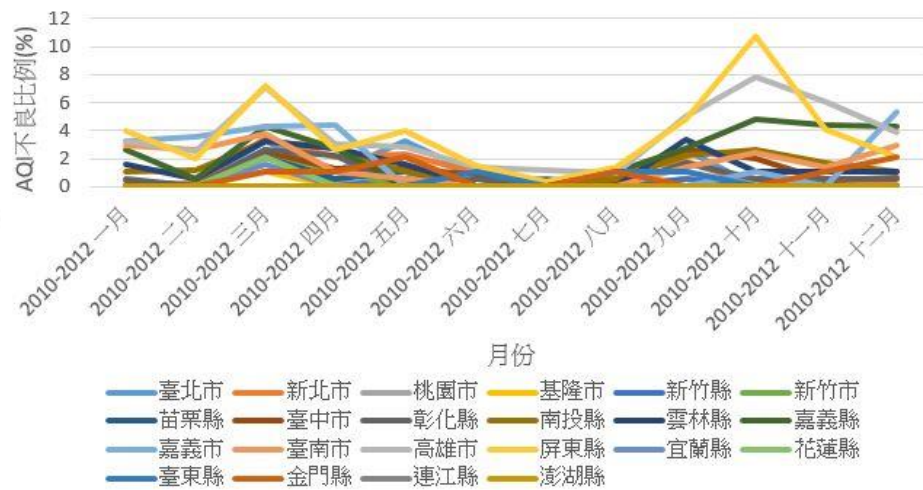
我們整理 AQI 與月份分析表格的方式也是用 AQI 與地理區域的方式進行，不一樣的是，我們每一列代表的是 2010~2012、2013~2015、2016~2018 年這三個時間區段它們在每個月算出它們這區段 3 年同一個月份的平均 AQI 大於 100 比例^[1]，如下圖所示。

2010-2012	一月
2010-2012	二月
2010-2012	三月
2010-2012	四月
2010-2012	五月
2010-2012	六月
2010-2012	七月
2010-2012	八月
2010-2012	九月
2010-2012	十月
2010-2012	十一月
2010-2012	十二月
2013-2015	一月
2013-2015	二月
2013-2015	三月
2013-2015	四月
2013-2015	五月
2013-2015	六月
2013-2015	七月
2013-2015	八月
2013-2015	九月
2013-2015	十月
2013-2015	十一月
2013-2015	十二月
2016-2018	一月
2016-2018	二月
2016-2018	三月
2016-2018	四月
-----	-----

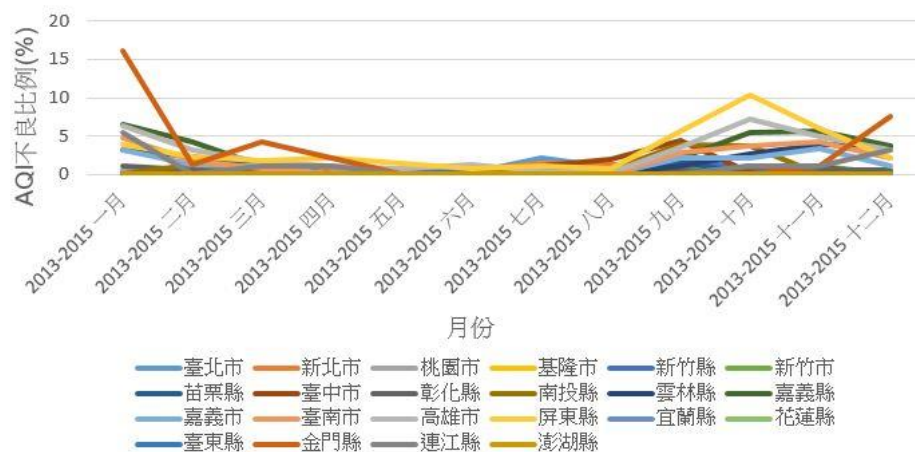
在月份的部分，我們分析的方式並不是使用相關矩陣，而是直接對三個時間區段畫出每個縣市在 1~12 月的平均 AQI 大於 100 比例折線圖，以下是折線圖的內容。

^[1] 舉例來說，2010~2012 這三年 1 月的「AQI 大於 100 比例」做算術平均就會變成其中一列，列名為「2010-2012 一月」

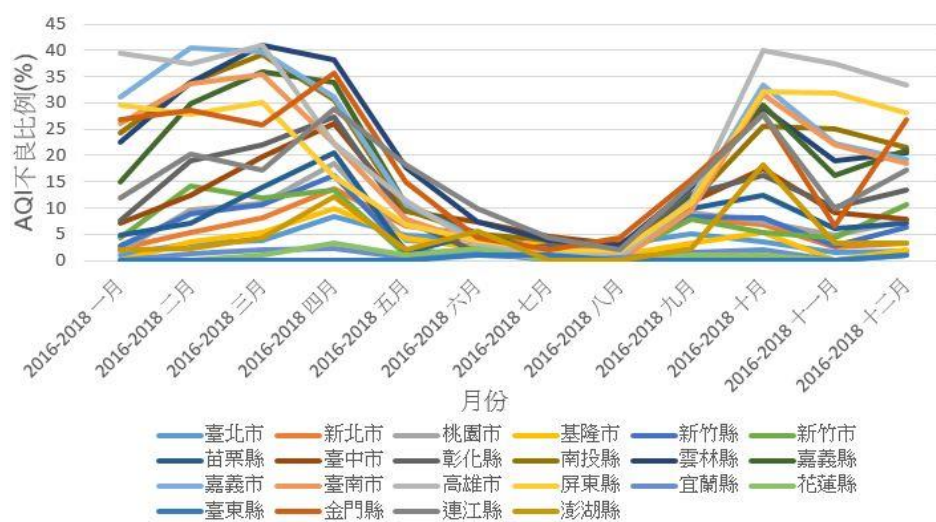
2010-2012各縣市每月平均AQI折線圖



2013-2015各縣市每月平均AQI折線圖



2016-2018各縣市每月平均AQI折線圖



（四）機器學習

1. 資料選擇：

基於資料處理的「方便性」及資料的「完整性」，我們最終只選擇了天氣資料及 AQI 作為機器學習的學習資料（Training Set）。

2. 學習方法：

我們試過利用線性迴歸（Linear Regression），作為資料分析的方法，但基於 AQI 數據呈現「許多資料為 0」的狀態，我們發現並不適合用

線性迴歸。如下圖所示

時間	縣市	AQI大於100之比率(%)	AQI大於100之日數
2010年1月	臺北市	0	0
	新北市	0	0
	桃園市	0	0
	基隆市	0	0
	新竹縣	0	0
	新竹市	0	0
	苗栗縣	0	0
	臺中市	1.29	2
	彰化縣	1.61	1
	南投縣	3.23	2
	雲林縣	4.84	3
	嘉義縣	8.06	5
	嘉義市	9.68	3
	臺南市	8.87	11
	高雄市	8.15	19
	屏東縣	8.79	8
	宜蘭縣	0	0
	花蓮縣	0	0
	臺東縣	0	0
	金門縣	0	0
	連江縣	0	0
	澎湖縣	0	0
	臺北市	0.71	1
	新北市	0.79	2
	桃園市	0	0
	基隆市	0	0

故我們決定以 AQI 大於 100 的天數比例是否大於 0%為分界點，使用了邏輯迴歸（Logistic Regression）來進行「二元分類」。

3. 邏輯迴歸：

- A. 利用 Sigmoid 函數「對參數的變化很敏感」的特性可將資料進行「二元分類」。
- B. 使用的假設函數（Hypothesis），用來表示的數學模型，如下圖

$$h(\theta) = \frac{1}{1 + e^{-\theta^T x}}$$

- C. 使用成本函數（Cost Function），用來表示每個參數的誤差量，判斷學習是否收斂，如下圖

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

- D. 利用梯度下降法（Gradient Descent）進行訓練。

4. 使用混淆矩陣（Confusion Matrix）的「召回率（Recall）」進行分析：

	實際 YES	實際 NO
預測 YES	TP (True Positive)	FP (False Positive) Type I Error
預測 NO	FN (False Negative) Type II Error	TN (True Negative)

- A. TP (True Positive)：預測是真，實際也是真（預測對）
- B. TN (True Negative)：預測是假，實際也是假（預測對）
- C. FP (False Positive)：預測是真，實際是假（預測錯）
- D. FN (False Negative)：預測是假，實際是真（預測錯）

Recall： $TP / (TP+FN)$ 。

意義：（預測是真，實際也是真） / （所有真實發生真的情況），代表實際是真的情況，能夠被預測多少比例，因為我們很不希望壞空氣的天數沒有被偵測到，所以要比較看重「實際是壞天氣，能夠被預測出的比例」。

5. 結果：

- A. 準確度 (Accuracy)： 65.99%
- B. 召回率 (Recall)： 33.04%
- C. 以「二元分類」來說，此結果差強人意，表示出天氣資料與 AQI 的值並沒什麼太大關係。

六、分析結果

(一) 天氣

1. 就總體而言，AQI 與天氣幾乎沒有關連（0.04 左右），只有與降水量可能比較有相關(-0.2 左右)，如下圖。

	AQI大於100之比率(%)	Temperature	Precp	RH	StnPres	WS
AQI大於100之比率(%)	1					
Temperature	-0.081446208	1				
Precp	-0.206462914	0.281359408	1			
RH	-0.02081299	0.045355338	0.21029872	1		
StnPres	0.047261075	0.095109633	0.01273313	0.790082	1	
WS	0.02919699	-0.094565244	-0.15140076	-0.02626	0.121231	1

2. 就區域來說，北部(新北)、南部(高雄)、東部(花蓮)部分 AQI 與天氣較無關聯；中部(台中)部分 AQI 與測站氣壓有中度相關(-0.58)；離島(金門)部分 AQI 與風速有中度相關(0.61)

(二) 火力發電量

1. 就總體而言，AQI 與火力發電量幾乎沒有關聯（0.11 左右）

	AQI大於100之比率(%)	發電量
AQI大於100之比率(%)	1	
發電量	0.115178681	1

2. 從單一縣市來看，雖然大多城市的發電量與 AQI 皆沒有太大的關連性，但是有些縣市會比其他縣市有更高的相關係數，像是台中的相關係數有達到 0.28，但是新北的相關係數卻只有 0.01

(三) 交通量

1. 總的來說，AQI 與交通流量關聯不大，數值都落在 0.27 左右

	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	每日平均車流量
2010-2012每月平均AQI大於100比例(%)	1			
2013-2015每月平均AQI大於100比例(%)	0.972688908	1		
2016-2018每月平均AQI大於100比例(%)	0.879980078	0.830645205	1	
每日平均車流量	0.27608534	0.222199957	0.156711044	1

(四) 工業區

1. 從數據來看，工業區的面積與 AQI 也沒有什麼關聯(0.12~0.33)

	2010-2012每月平均AQI大於100比例(%)	2013-2015每月平均AQI大於100比例(%)	2016-2018每月平均AQI大於100比例(%)	工業區數目	面積(平米)	佔各縣市面積比例(%)
2010-2012每月平均AQI大於100比例(%)	1					
2013-2015每月平均AQI大於100比例(%)	0.836153427	1				
2016-2018每月平均AQI大於100比例(%)	0.721574433	0.825855343	1			
工業區數目	0.174874264	0.008827665	-0.032374474	1		
面積(平米)	0.158300072	0.123836884	0.337401856	0.233449224	1	
佔各縣市面積比例(%)	-0.001350565	-0.02510639	0.199030262	0.104539354	0.91308696	1

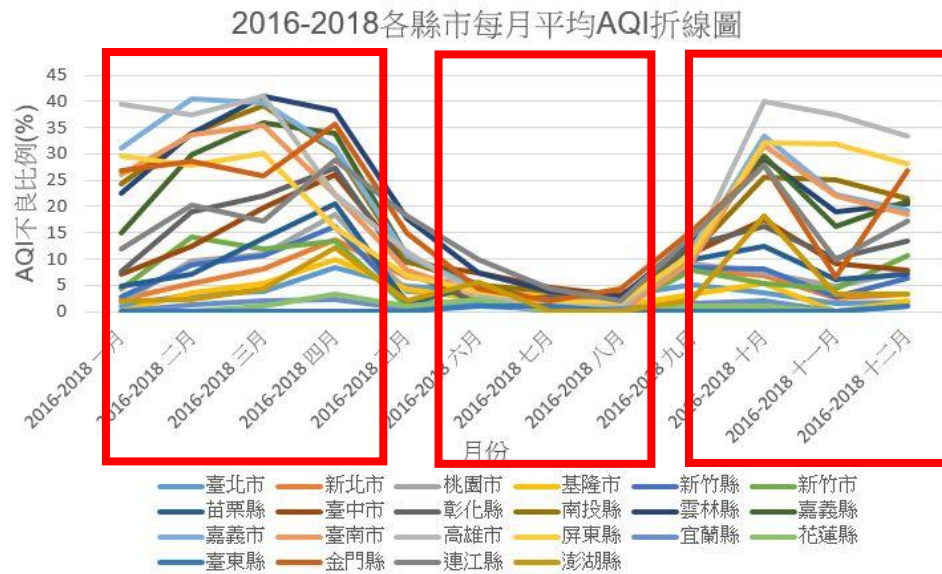
(五) 地理區域

1. 從相關矩陣中可以看到，鄰近的區域在相同的時間點對於 AQI 的值會有非常高度的相關，但是東半部卻沒有這樣的現象

	臺北市	新北市	桃園市	基隆市	新竹縣	新竹市	苗栗縣	臺中市	彰化縣	南投縣	雲林縣	嘉義縣	嘉義市	臺南市	高雄市	屏東縣	宜蘭縣	花蓮縣	臺東縣	金門縣	連江縣	澎湖縣
臺北市	1																					
新北市	0.926354	1																				
桃園市	0.811313	0.884271	1																			
基隆市	0.806292	0.900943	0.813932	1																		
新竹縣	0.764199	0.830905	0.963383	0.787842	1																	
新竹市	0.610952	0.698307	0.89563	0.618938	0.898627	1																
苗栗縣	0.727932	0.802449	0.929114	0.749736	0.92081	0.865214	1															
臺中市	0.778899	0.883468	0.933795	0.822477	0.910637	0.821922	0.92625	1														
彰化縣	0.695392	0.8122	0.939656	0.749488	0.925843	0.894187	0.893374	0.928741	1													
南投縣	0.620881	0.759063	0.830972	0.710348	0.820801	0.807863	0.777247	0.875171	0.898806	1												
雲林縣	0.674977	0.806467	0.884301	0.750962	0.865121	0.843975	0.841501	0.924999	0.944417	0.968546	1											
嘉義縣	0.620032	0.758502	0.878414	0.709581	0.865207	0.850832	0.840997	0.896103	0.944394	0.932897	0.965404	1										
嘉義市	0.567146	0.712716	0.792307	0.655974	0.784237	0.775814	0.736303	0.841667	0.878974	0.974195	0.959482	0.940047	1									
臺南市	0.534369	0.676567	0.770598	0.617236	0.757927	0.761987	0.722163	0.817852	0.874031	0.961936	0.951087	0.941826	0.983616	1								
高雄市	0.444383	0.582663	0.707859	0.520628	0.702745	0.749241	0.691384	0.753392	0.823112	0.937456	0.906401	0.892457	0.941155	0.96162	1							
屏東縣	0.433948	0.56063	0.686591	0.489303	0.672735	0.716121	0.657064	0.727243	0.766841	0.911921	0.864701	0.854578	0.908956	0.932785	0.97539	1						
宜蘭縣	0.790027	0.799346	0.761459	0.806092	0.758666	0.581262	0.70372	0.751871	0.677633	0.646229	0.683483	0.663427	0.620674	0.606433	0.5198	0.51283	1					
花蓮縣	0.778084	0.784558	0.655664	0.753068	0.592917	0.431965	0.586619	0.641771	0.52841	0.477207	0.522366	0.504324	0.439643	0.412378	0.321486	0.300863	0.834395	1				
臺東縣	0.138919	0.07463	0.048595	0.058014	0.006685	0.024301	0.035055	0.043526	0.008489	0.047853	0.029553	0.062941	0.008109	0.011746	0.024924	0.033688	0.162595	0.216118	1			
金門縣	0.602271	0.73418	0.817391	0.711711	0.812699	0.761271	0.746543	0.813304	0.866663	0.873125	0.900976	0.875759	0.88062	0.863269	0.836506	0.78515	0.653019	0.300223	0.006875	1		
連江縣	0.680564	0.806377	0.816813	0.827068	0.783777	0.695242	0.748965	0.821645	0.843457	0.840946	0.873446	0.840963	0.8243	0.808308	0.777502	0.733469	0.717432	0.616128	0.10719	0.916742	1	
澎湖縣	0.487078	0.611104	0.623463	0.617741	0.563677	0.41699	0.5953	0.653765	0.654877	0.580077	0.631772	0.641679	0.60445	0.63183	0.571163	0.552577	0.665966	0.562455	0.061077	0.643541	0.754436	1

(六) 月份

1. 從圖中可以看到 AQI 與月份有顯著的關係，在大約 10 月到隔年 4 月會有較高的 AQI 不良比例，而在 6~8 月時比例較低



七、分析結論

（一） 各縣市之 AQI 受不同因素影響

1. 台中市-氣壓和火力發電量
2. 苗栗縣-火力發電量
3. 嘉義市-相對溼度
4. 金門縣和連江縣-風速
5. 其他縣市-其他待探討因素

（二） 天氣

1. 依據各地區氣候地形條件的不同，對於 AQI 會有不同的影響。在某些特定氣候地形情況下 AQI 會與天氣的一些參數有顯著的關聯，但在其他的情況下可能就不會有太大的關聯性。例如：金門與風速有相關性，可能是因為離島地形平坦易受中國大陸污染物的擴散影響
2. 在特定的情況下，AQI 確實會與天氣有關聯，與預期結果相符，但在其他情況下則不一定

（三） 火力發電

AQI 與火力發電量確實沒有什麼關聯，與預期結果相符

（四） 月份

與我們的預期相符，AQI 有季節性

八、未來展望

我們希望我們做得這份研究能為台灣空氣品質的改善盡一份心力，讓政府、民間的團體或媒體注意到我們分析出來的結果，讓以後在討論這個話題時能把重點放在更準確的方向，而不是不斷的處理些並不是那麼重要的問題，期望有更準確的研究團隊對我們的題目做更深入的研究並發表，讓政府能對不同縣市不同地區的空气品質問題的根本原因去改進。

九、我們的看法

依據分析的結果，其實我們沒有找到一個影響 AQI 的決定性因素，我們推測是因為影響 AQI 的因素實在是太多了，導致最多只能找到中度相關的因素，但還是有比較明顯的因素那就是月份，跟月份最相關的推測是季風的因素，再將資料與季風比對發現冬季吹東北季風時 AQI 特別高空氣品質特別差，東北季風帶來的污染源可追溯到中國內陸，所以我們覺得台灣最大的空氣問題還是從大陸隨著季風帶來的污染物。

十、參考資料

- (一) 盧市府防空污 擬卡台中火力電廠生煤許可證、用量
(<https://news.ltn.com.tw/news/politics/breakingnews/2840027>)
- (二) 107 年公民投票第 7 至 16 案投票結果
(<https://web.cec.gov.tw/central/cms/RefResults7to16>)
- (三) 交通部中央氣象局 觀測資訊查詢系統
(<https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp>)
- (四) 臺灣電力公司 過去電力供需資訊
(<https://www.taipower.com.tw/tc/page.aspx?mid=210>)
- (五) 交通部公路總局 公路交通量調查統計
(https://www.thb.gov.tw/sites/ch/modules/download/download_list?node=bcc520be-3e03-4e28-b4cb-7e338ed6d9bd&c=83baff80-2d7f-4a66-9285-d989f48effb4)
- (六) 台灣工業用地供給與服務資訊網 工業區統計
(<https://idbpark.moeaidb.gov.tw/Environ/statistical>)
- (七) 行政院環境保護署 環保統計查詢網
(<https://stat.epa.gov.tw/>)
- (八) Vol.5_空氣品質指標 AQI 是什麼？跟常聽見的 PM2.5 又有什麼關係？
(https://www.xpure-tw.com/blog/vol5_aqipm25)
- (九) Machine Learning 學習日記 — Coursera 篇 (Week 3.2):Cost Function, Simplified Cost Function and Gradient Descent, Advanced Optimization
(https://medium.com/@ken90242/machine-learning_學習日記-coursera_篇-week-3-2-cost-2e7aa00e8c4a)
- (十) 如何辨別機器學習模型的好壞？秒懂 Confusion Matrix
(<https://www.ycc.idv.tw/confusion-matrix.html>)

十一、 附錄

(一) 網頁爬蟲程式碼

```
import os

import time

from selenium import webdriver

from selenium.webdriver.support.ui import Select

yearlist = [2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]
#yearlist = [2016, 2017, 2018, 2019]

for Years in yearlist:

    print(Years)

    options = webdriver.ChromeOptions()

    prefs = {'profile.default_content_settings.popups': 0,
            'download.default_directory':
'D:\\Main\\NTUST\\bigData_finalTeamProj\\weather\\'+str(Years)}

    options.add_experimental_option('prefs', prefs)

    driver = webdriver.Chrome(
        executable_path="D:\\Main\\NTUST\\bigData_finalTeamProj\\chromedriver79.exe",
        chrome_options=options)

    driver.get("https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp")

    assert "觀測資料查詢系統" in driver.title

    driver.implicitly_wait(30)

    selectCountry = Select(driver.find_element_by_id('stationCounty'))

    for city in range(len(selectCountry.options)):

        currCity = selectCountry.options[city].text

        print(currCity)

        selectCountry.select_by_index(city)

        driver.implicitly_wait(30)

    selectStation = Select(driver.find_element_by_id('station'))

    for station in range(len(selectStation.options)):

        currStation = selectStation.options[station].text

        print("\t"+currStation)
```



```

if "撤銷站" in currStation:
    print("廢棄資料")
    continue

fileExist = False

dir_link = 'D:\\Main\\NTUST\\bigData_finalTeamProj\\weather\\'+str(Years)
newName = currCity.split(" ")[0]+"_"+currStation.split(" ")[0]+"_"+str(Years)+".csv"
if "?" in newName:
    newName = newName.replace("?", "[E]")
for file in os.listdir(dir_link+"\\ "+currCity+"\\"):
    if newName in file:
        print(currCity.split(" ")[0]+"_"+currStation.split(" ")[0]+"_"+str(Years)+".csv"+"
exist!")

        fileExist = True
        break

if fileExist:
    continue

selectStation.select_by_index(station)
driver.implicitly_wait(30)

selectClass = Select(driver.find_element_by_id('dataclass'))
selectClass.select_by_value("inquire")
driver.implicitly_wait(30)

selectType = Select(driver.find_element_by_id('datatype'))
selectType.select_by_value("year")
driver.implicitly_wait(30)

year = driver.find_element_by_id("datepicker")
year.clear()
year.send_keys(str(Years))

driver.find_element_by_id("doquery").click()
driver.implicitly_wait(30)
time.sleep(1)

```

```
windows=driver.window_handles  
driver.switch_to.window(windows[-1])
```

```
driver.find_element_by_css_selector('a#downloadCSV>input').click()  
time.sleep(2)  
driver.close()  
driver.switch_to_window(windows[0])
```

```
dir_link = 'D:\\Main\\NTUST\\bigData_finalTeamProj\\weather\\'+str(Years)  
dir_lists = os.listdir(dir_link)  
dir_lists.sort(key=lambda fn: os.path.getmtime(dir_link + '\\'+ fn))  
os.chdir(dir_link)  
newName = currCity.split(" ")[0]+"_"+currStation.split(" ")[0]+"_"+str(Years)+".csv"  
if "?" in newName:  
    newName = newName.replace("?", "[E]")  
os.rename(dir_lists[-1], newName)  
os.replace(dir_link+"\\"+newName, dir_link+"\\"+currCity+"\\"+newName)
```

```
print()
```

```
driver.quit()  
print()
```

(二) 資料整理程式碼

```
import pandas as pd
import os
```

```
def step1():
    for year in os.listdir():
        if os.path.isdir(year):
            os.chdir(year)

            for city in os.listdir():
                if os.path.isdir(city):
                    os.chdir(city)
                    all_data = pd.DataFrame()
                    for csvFile in os.listdir():
                        if '.csv' in csvFile and 'all' not in csvFile:
                            df = pd.read_csv(csvFile)
                            all_data = pd.concat([all_data, df])
                            print(os.getcwd())
                    all_data.to_csv('all.csv')
                    os.chdir('../')

            os.chdir('../')
```

```
def step2():
    for year in os.listdir():
        if os.path.isdir(year):
            os.chdir(year)

            for city in os.listdir():
                if os.path.isdir(city):
                    os.chdir(city)
                    try:
                        os.remove('result.csv')
                    except:
                        print()
                    os.chdir('../')
```

```

os.chdir('../')

def step3():
    import pandas as pd
    import os
    pd.set_option('mode.chained_assignment', None)
    for year in os.listdir():
        if os.path.isdir(year):
            os.chdir(year)

            for city in os.listdir():
                if os.path.isdir(city):
                    os.chdir(city)

                    items = ['Temperature', 'Precp', 'RH', 'StnPres', 'WS']
                    df = pd.read_csv('all.csv')
                    new_df = pd.DataFrame()
                    new_df['Mouth'] = range(1, 13)
                    for item in items:
                        new_df[item] = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

                        length = df.shape[0]
                        l = [0,0,0,0,0,0,0,0,0,0,0,0]
                        for i in range(length):
                            try:
                                float(df[item][i])
                                mouth = df['ObsTime'][i]
                                l[mouth-1] += 1
                                new_df[item][mouth-1] += float(df[item][i])
                            except ValueError:
                                a=0

                        for i in range(12):
                            if l[i] != 0:
                                new_df[item][i] = float(new_df[item][i]) / l[i]
                                new_df[item][i] = round(new_df[item][i], 2)
                    new_df.to_csv('result.csv')
                    os.chdir('../')

os.chdir('../')

```

step1()

step2()

step3()

(三) 機器學習程式碼

```
import numpy as np

import pandas as pd

from sklearn.linear_model import LogisticRegression

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import confusion_matrix


LR = LogisticRegression()


df1 = pd.read_csv('天氣總表.csv')

yi = pd.read_csv('for_MC.csv')['AQI 大於 100 之比率(%)'][:2376]


xi = df1[['Temperature', 'Precp', 'RH', 'StnPres', 'WS']]

xi.insert(0, 'Constant', 1)


# 用來增加參數

xi['T^2'] = df1['Temperature'] * df1['Temperature']

xi['P^2'] = df1['Precp'] * df1['Precp']

xi['R^2'] = df1['RH'] * df1['RH']

xi['S^2'] = df1['StnPres'] * df1['StnPres']

xi['W^2'] = df1['WS'] * df1['WS']

xi['TP'] = df1['Temperature'] * df1['Precp']

xi['TR'] = df1['Temperature'] * df1['RH']

xi['TS'] = df1['Temperature'] * df1['StnPres']

xi['TW'] = df1['Temperature'] * df1['WS']

xi['PR'] = df1['Precp'] * df1['RH']

xi['PS'] = df1['Precp'] * df1['StnPres']

xi['PW'] = df1['Precp'] * df1['WS']

xi['RS'] = df1['RH'] * df1['StnPres']

xi['RW'] = df1['RH'] * df1['WS']

xi['SW'] = df1['StnPres'] * df1['WS']

# =====
```

```
x = xi.to_numpy()
y = yi.to_numpy(dtype=int)

sc = StandardScaler()
sc.fit(x)
x = sc.fit(x).transform(x)

LR.fit(x, y)

# =====

xt = x
yt = LR.predict(xt)
accuracy = LR.score(x, y)
tn, fp, fn, tp = confusion_matrix(y, yt).ravel()
recall = tp / (tp + fn)

print(str(accuracy) + ' ' + str(recall))

df = pd.DataFrame(y)
df.to_csv('result.csv')
```