* 主題：環境與人口密度對登革熱確診數影響
* 動機：

世界衛生組織表示，2023年全球報告了超500萬登革熱感染病例，包括5000例死亡病例。這一出乎意料的病例激增，對公共衛生構成了潛在的嚴重威脅。尤其是像台灣這樣擁有溫暖濕潤氣候的地區，登革熱的季節性爆發尤為顯著。

聯合新聞網中有提到，近日每天都可以看到登革熱相關的新聞，也出現越來越多登革熱的病例，光是新北在2024/09/30日就再增6例本土登革熱病例，累計達49例。( https://udn.com/news/story/123735/8259545 )

* 資料：

Summary：

Y（應變數）：登革熱確診數

X（自變數）：空氣品質、人口密度及其他環境因素

1. 登革熱確診數 ( 應變數 )

資料來源：疾病管制署資料開放平台 – 登革熱1998年起每日確定病例統計

( <https://data.cdc.gov.tw/dataset/dengue-daily-determined-cases-1998>)

資料時間：1998年至今登革熱每日確定病例統計

資料的時間粒度：日

抓取條件：發病日介於2013年 ~ 2023年且不為境外移入

資料解釋：登革熱2013-2023年每日確定病例分組統計為整理後的資料

Row data抓取後總比數：86702筆

一張含有 螢幕擷取畫面, 文字, 多媒體軟體 的圖片

自動產生的描述

|  |  |  |
| --- | --- | --- |
| item | type | note |
| 發病日 | datetime | YYYY/MM/DD |
| 個案研判日 | datetime | YYYY/MM/DD |
| 通報日 | datetime | YYYY/MM/DD |
| 性別 | string | F/M |
| 年齡層 | string | 最小0歲, 五歲一個單位, 最大70+ |
| 居住縣市 | string |  |
| 居住鄉鎮 | string |  |
| 居住村里 | string | 很少有資料應刪除 |
| 最小統計區 | string | 無關刪除 |
| 最小統計區中心點X | float | 無關刪除 |
| 最小統計區中心點Y | float | 很少有資料應刪除 |
| 一級統計區 | string | 很少有資料應刪除 |
| 二級統計區 | string | 很少有資料應刪除 |
| 是否境外移入 | string | 是/否  應只留下是否境外移入=否 |
| 確定病例數 | int |  |
| 居住村里代碼 | string | 無關刪除 |
| 血清型 | string | 無關刪除 |
| 內政部居住縣市代碼 | string | 無關刪除 |
| 內政部居住鄉鎮代碼 | string | 無關刪除 |
| 內政部感染縣市代碼 | string | 無關刪除 |
| 內政部感染鄉鎮代碼 | string | 無關刪除 |

後續須group by 發病日/居住縣市/居住鄉鎮

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

自動產生的描述

大概看一下資料分布，整理後資料筆數為13518筆，後續再決定要不要刪除離群值、對數變換等等的。

1. 空氣品質 ( 自變數 )

資料來源：空氣品質監測網( https://airtw.moenv.gov.tw/CHT/Query/His\_Data.aspx )

Raw data：

一張含有 文字, 黑與白, 黑色, 白色 的圖片

自動產生的描述

資料處理：將資料轉製成以下格式

一張含有 文字, 黑與白, 黑色, 樣式 的圖片

自動產生的描述

並留下每日最大值、最小值、平均值作為變數

抓取條件：2013年 ~ 2023年空氣品質資料 ( 需合併 )

資料的時間粒度：日

資料解釋：每個測站都有一個data，這裡以”空氣品質資料\_二林\_2023”為例，後續再做合併。

整理後：

|  |  |  |  |
| --- | --- | --- | --- |
| item | 定義 | 單位 | type |
| 測站 | - | - | string |
| 日期 | YYYY-mm-dd HH:MM:SS | - | datetime |
| AMB\_TEMP | 溫度 |  | float |
| CH4 | 甲烷 | ppm | float |
| CO | 一氧化碳 | ppm | float |
| NMHC | 非甲烷碳氫化合物 | ppm | float |
| NO | 一氧化氮 | ppb | float |
| NO2 | 二氧化氮 | ppb | float |
| NOx | 氮氧化物 | ppb | float |
| O3 | 臭氧 | ppb | float |
| PM10 | 懸浮微粒 |  | float |
| PM2.5 | 細懸浮微粒 |  | float |
| RAINFALL | 雨量 | mm | float |
| RH | 相對濕度 | % | float |
| SO2 | 二氧化硫 | ppb | float |
| THC | 總碳氫化物 | ppm | float |
| WD\_HR | 風向小時值  (以整個小時向量平均) | m/sec | float |
| WIND\_DIREC | 風向  (以每小時最後10分鐘向量平均) | degress | float |
| WIND\_SPEED | 風速  (以每小時最後10分鐘向量平均) | m/sec | float |
| WS\_HR | 風速小時值  (以整個小時向量平均) | degress | float |

1. **人口密度**( 自變數 )

資料來源：內政部戶政司全球資料網 ( <https://www.ris.gov.tw/app/portal/346> )

抓取條件：2013年 ~ 2023年戶數、人口數、性別比例及人口密度統計表( 需合併 )

資料解釋：每個月都有一個data，這裡以” 縣市人口性比例及人口密度-112年1月”為例，後續再做合併。

資料的時間粒度：月 ( 雖然是月的資料，但人口密度應不大，可以直接把每月的人口密度資料分配到每日資料吧 )

**並從中抓取資料：**

|  |  |  |
| --- | --- | --- |
| item | type | 單位 |
| 月份 | string |  |
| 區域別 | string |  |
| 人口密度 | float | 每平方公里人口數 |

* **資料處理：**

上述資料依據日期、地區合併，再做相對應的分析處理，目前都為raw data。

* **考慮問題：**

1. 要不要做刪除離群值或是對數轉換等等
2. 就結果準確度再看看要不要分區進行建模
3. 人口密度的資料時間粒度為月，直接拼組再前面的每日資料中是否會不準確
4. 或許還有別的自變數可以加入