# 期中報告

M132040019 廖廣筑

M122040017 吳俞憲

## I. Abstract

這篇報告主要是在探討各種天氣相關的變數中，哪些變數最影響到健康風險指標的數值。透過 EDA 去分析包含美國多個城市的綜合氣象數據，去找出與健康風險指標變動密切相關的特定天氣屬性，例如氣溫、濕度、風速、太陽輻射及降水量……等。然後進行建模，並將各個模型進行比較，最後解釋模型。

## II. Introduction and related work

隨著近年來各個國家發展快速，空氣汙染與極端氣候的現象越來越嚴重，這個現象也悄悄的影響著大家的健康，因此有必要去了解各個天氣變數與健康風險指數的相關性。這份報告探討了美國城市 2024 年 9 月的詳細天氣資料與健康風險指標。

## III. Dataset and methods

此資料集含有 1000 筆資料與 46 項變數，提供了美國主要城市的城市空氣品質及其潛在健康影響相關的各項數據。

變數為 10 大類：

目標變數(健康風險指標)、天氣變數(天氣風險、天氣詳細描述、天氣圖示表示、特定天氣狀況代碼、天氣條件嚴重程度分數、總體天氣狀況）、降水(降水類型、總降水量、降水機率、降水覆蓋範圍）、風(當天最大風速、當天平均風速、風向）、溫度(當天最高溫度、當天最低溫度、當天平均溫度、體感最高溫、體感最低溫、體感平均溫、露點溫度、當天溫差、熱指數）、太陽(太陽輻射、接收的太陽能量、UV 指數等級、日出時間、日落時間）、時間(月份、季節、星期幾、是否為周末）、地點(城市、提供資料的氣象站）、雪(降雪量、雪深度）、其他(月相、資料來源、相對溼度、雲量覆蓋率、氣壓）

我們使用挑選過的變數，用 EDA 的方式去分析資料，最後用 pycaret 去篩選模型，最後用 SHAP 的方式去解釋模型。

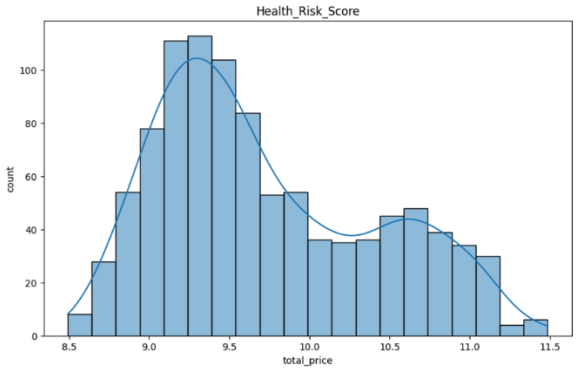
# IV. Experiments and results

## a. 處理缺失值

1. 特定天氣狀況代碼：我們查不到此代碼表示的涵義，故後續部會將這個變數加入討論。
2. 提供資料的氣象站：缺失值太高，而且有城市這個變數，所以也不討論此變數。
3. 降水類型：資料中只有 NA 跟 rain，所以資料在填寫的時候只有填寫下雨，所以我們將 NA 值改完 0 代表沒有降水。
4. 雪深度：他不是 nan 就是 0，所以也不考慮此變數

## b. Exploratory Data Analysis

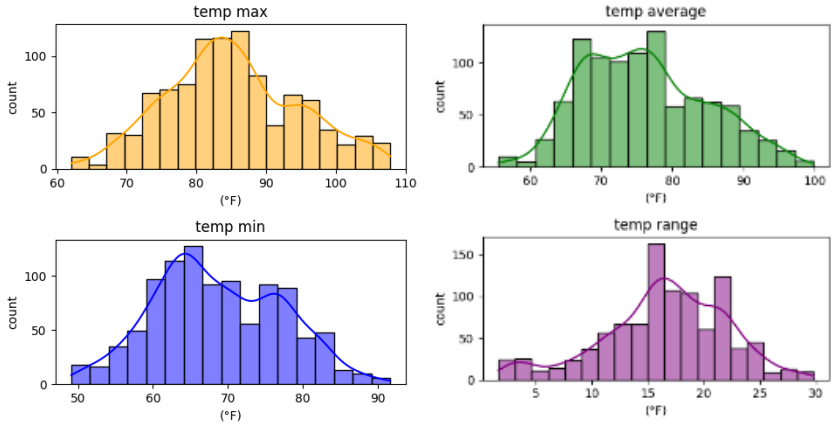
1. 目標變數：健康風險指標




健康風險指標：代表基於天氣和空氣品質狀況的潛在健康風險的分數。

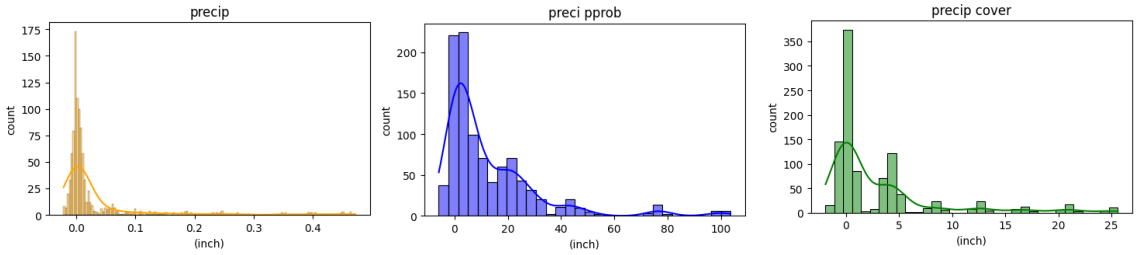
健康風險指標數值界在 8.5~11.5 之間，而數據大部分都落在 9~10 之間。所以如果今天的健康風險大於 10，代表今天的健康風險是偏高的，要特別注意。

2. 數值變數：溫度



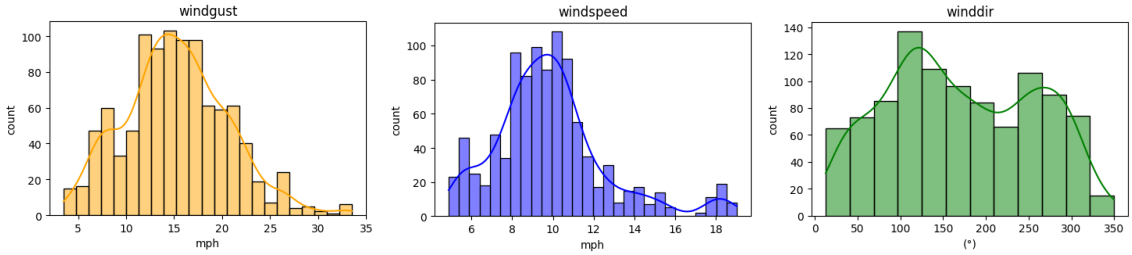

這四個溫度的分布都不極端，但平均溫度跟溫差較可以代表當日溫度狀況，而且當日的溫差也會影響到健康。所以我們會把平均溫度跟溫差加入後續討論。而體感溫度會受到相對濕度與風速與氣溫與個別身體差異的影響，所以不納入後續討論。

3. 數值變數：降雨



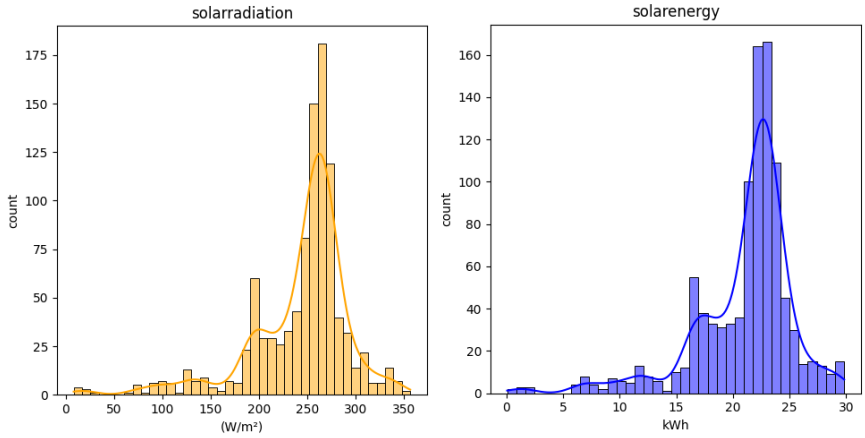

這三個降雨相關的圖有嚴重的右偏，而且還有負數的出現，可能是資料填的時候就有錯誤，所以我們不會考慮這三個變數。

4. 數值變數：風速




最大風速與平均風速的分布較接近，我們想要利用最大風速跟平均風速來製造一個新變數，可以了解風速有沒有很大的波動，因為極端氣候也有可能影響到健康風險。當初查看資料時不知道該如何對風向(0~360 度)這個變數進行處理，所以後續就沒有把風向加入討論。

5. 數值變數：太陽變數

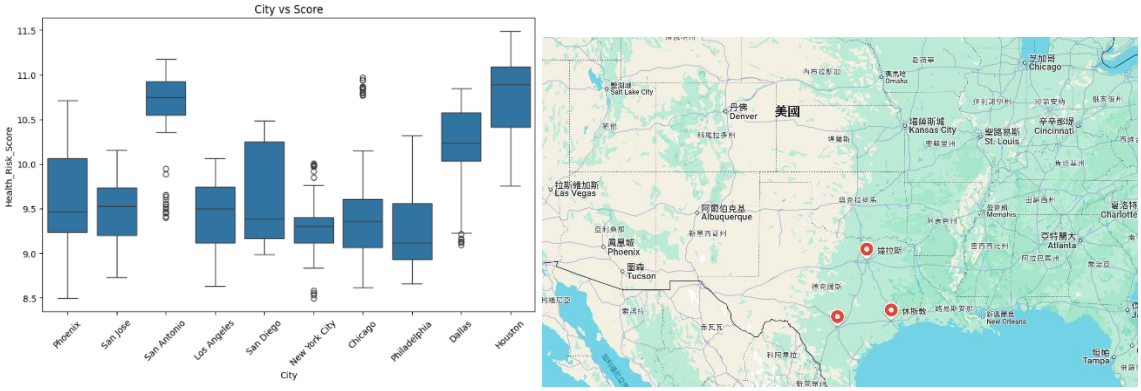



```
correlation = df['solarradiation'].corr(df['solarenergy'])
print(correlation)
```

```
0.9912955859677877
```

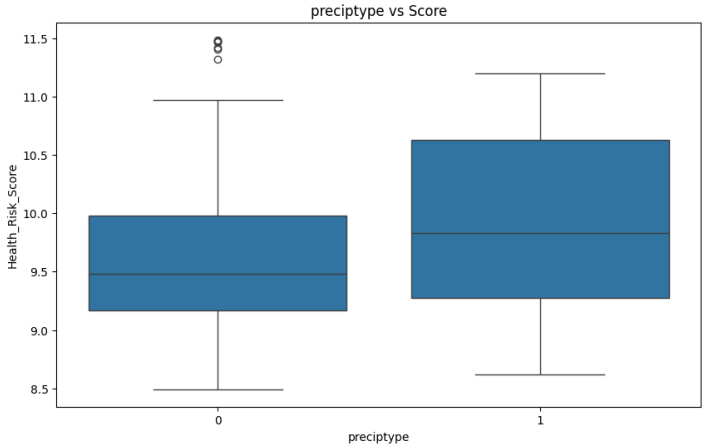
這兩個變數分配看起來很相似，是因為太陽輻射代表每平方公尺在單位時間內的能量輸入，而太陽能量代表每小時的能量輸入。而且他們倆個變數的相關性很高。所以後續只會把太陽輻射加入討論。而 UV 指數應該要是整數數值，但資料中的 UV 指數有小數，應該是資料錯誤，因此不加入討論。

6. 類別變數：地區與健康風險指標



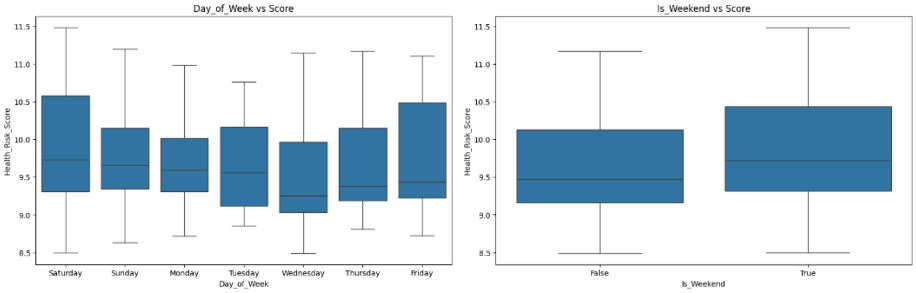

分數偏高的三個地區分別是：San Antonio、Dallas、Houston，而這三區都位在美國德克薩斯州，因此可以推測出這個區域的健康風險都是偏高，可以這對此區域做更多的探討。

7. 類別變數：是否降雨與健康風險指標




0 代表沒降雨、1 代表有降雨，從這張圖可以看出兩者差異不大，但是有降雨的部分健康風險指標稍微高一點。

8. 類別變數：星期幾與健康風險指標




健康風險指數看起來都差不多，星期幾應該是對目標變數沒有特別影響。

9. 時間的變數不會加入後續討論，因為這份資料只包含 2024 年 9 月秋季的資料。

c. 特徵工程

1. 新變數風速比率: max/average

```
df['windratio'] = df['windgust'] / df['windspeed']
```

2. 對城市做 Target encoding

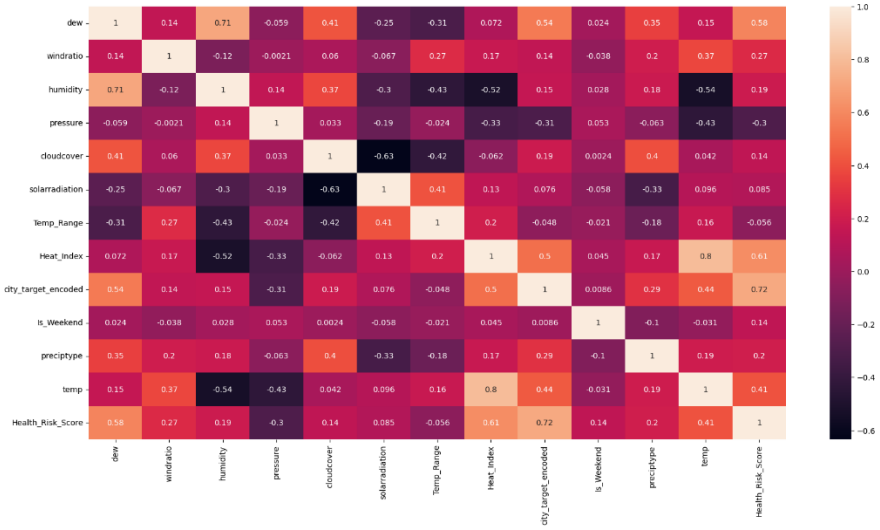
```
df['windratio'] = df['windgust'] / df['windspeed']
```

3. 是否為周末改為 0,1

```
df['windratio'] = df['windgust'] / df['windspeed']
```

d. 相關係數熱力圖:

我們去掉未知變數、部分時間變數與上述討論出來的不保留的變數




可以發現溫差跟太陽輻射這兩個與健康風險指標的相關性很低。

# v. Model analysis and comparison

## a. 模型前資料處理：

1. 資料拆分：我們將數據拆分為 90%訓練集、10%測試集。
2. 變數比較：從相關性熱力圖我們可以發現有兩個變數與健康風險指標的相關性較低，所以我們想比較看看熱力圖中的全部變數模型與刪除二個相關性較低的變數模型。
3. 利用 pycaret 建立模型，從訓練集再拆分出 10%驗證集，且資料有經過標準化。

## b. 全變數模型：

1.

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
|---|---|---|---|---|---|---|---|---|
| **svm** | Support Vector Regression | 0.0947 | 0.0159 | 0.1259 | 0.9654 | 0.0116 | 0.0097 | 0.1080 |
| **rf** | Random Forest Regressor | 0.0941 | 0.0168 | 0.1285 | 0.9636 | 0.0120 | 0.0097 | 1.1760 |
| **knn** | K Neighbors Regressor | 0.0939 | 0.0183 | 0.1348 | 0.9605 | 0.0125 | 0.0097 | 0.0940 |
| **xgboost** | Extreme Gradient Boosting | 0.0993 | 0.0184 | 0.1348 | 0.9595 | 0.0126 | 0.0102 | 0.2420 |
| **dt** | Decision Tree Regressor | 0.1300 | 0.0474 | 0.2155 | 0.8977 | 0.0199 | 0.0134 | 0.0800 |
| **lr** | Linear Regression | 0.2059 | 0.0729 | 0.2691 | 0.8422 | 0.0248 | 0.0211 | 1.1100 |
| **ridge** | Ridge Regression | 0.2061 | 0.0729 | 0.2691 | 0.8422 | 0.0248 | 0.0211 | 0.0720 |
| **lasso** | Lasso Regression | 0.5781 | 0.4710 | 0.6854 | -0.0161 | 0.0629 | 0.0588 | 0.0780 |

可以發現 SVM 模型評分下幾乎都是最好的，而第二名是 Random Forest，且這兩者的差異不是很大。

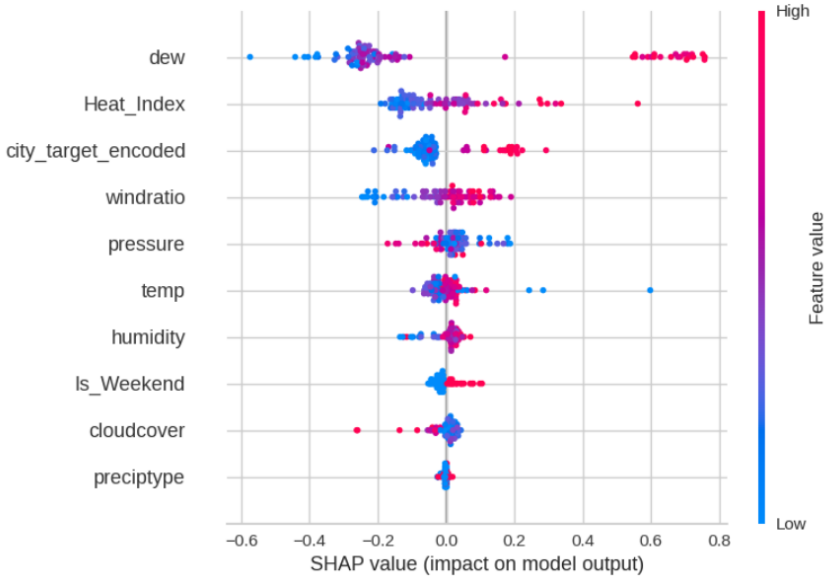
c. 去除相關性較低模型：

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
|---|---|---|---|---|---|---|---|---|
| **rf** | Random Forest Regressor | 0.1010 | 0.0189 | 0.1367 | 0.9588 | 0.0128 | 0.0104 | 0.6580 |
| **svm** | Support Vector Regression | 0.1045 | 0.0194 | 0.1391 | 0.9578 | 0.0128 | 0.0107 | 0.1000 |
| **xgboost** | Extreme Gradient Boosting | 0.1091 | 0.0220 | 0.1478 | 0.9520 | 0.0138 | 0.0113 | 0.1940 |
| **knn** | K Neighbors Regressor | 0.1045 | 0.0241 | 0.1536 | 0.9488 | 0.0144 | 0.0108 | 0.0860 |
| **dt** | Decision Tree Regressor | 0.1335 | 0.0476 | 0.2156 | 0.8972 | 0.0201 | 0.0138 | 0.0360 |
| **lr** | Linear Regression | 0.2253 | 0.0819 | 0.2858 | 0.8223 | 0.0265 | 0.0232 | 0.0580 |
| **ridge** | Ridge Regression | 0.2254 | 0.0819 | 0.2858 | 0.8223 | 0.0265 | 0.0232 | 0.0660 |
| **lasso** | Lasso Regression | 0.5781 | 0.4710 | 0.6854 | -0.0161 | 0.0629 | 0.0588 | 0.0640 |

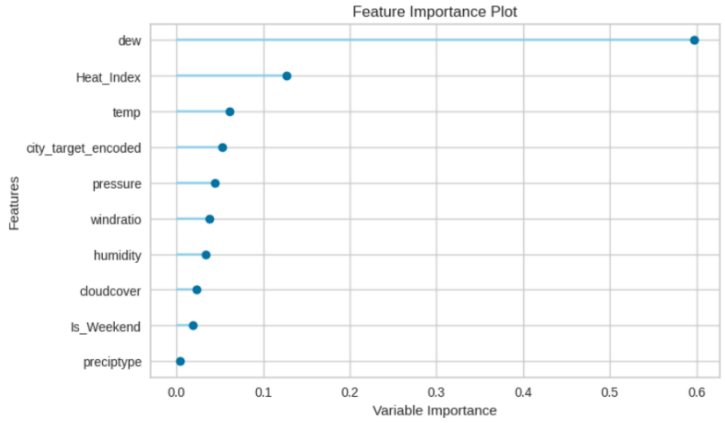
結果變成 Random Forest 是最好的，其次是 SVM 一樣都是這兩個模型是最好的。我們比較了全變數與去除相關性較低的變數，其實刪除變數之後模型評分也不會下降太多，所以我們決定最後選則刪除兩個相關性較低的變數。

## d. 模型解釋性：

1. Random Forest：

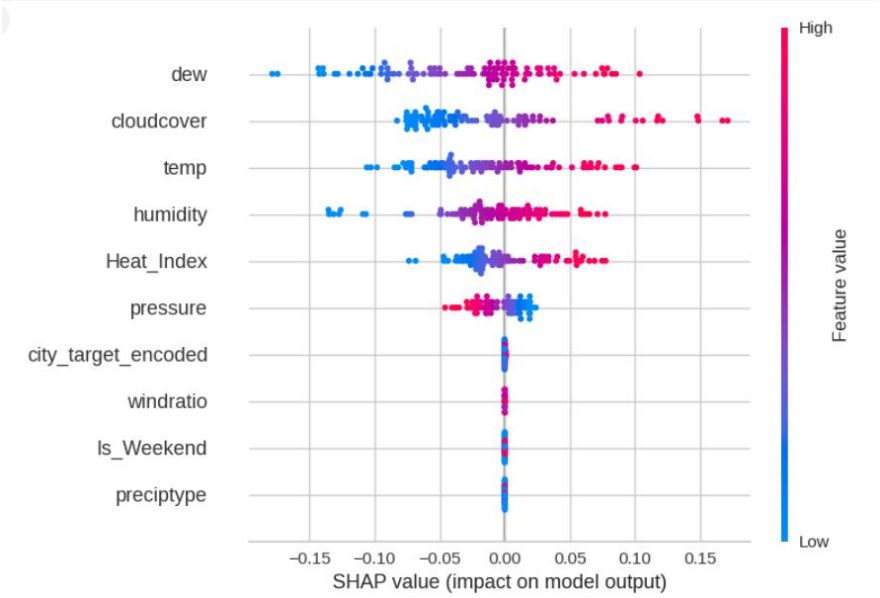



可以看到在 Shap 圖之下，大部分數據都是變數的值上升，風險也是上升的 只有 pressure 有潛在可能如果值下降，風險是上升的。在 dew 有很明確的分布，代表這可能是影響風險很重要的因素




從特徵重要性中也可以看到露點溫度確實是一個對健康風險很重要的變數，其次是熱指數。

2. SVM：




與 Random Forest 一樣，dew 看起來都是模型覺得很重要的變數，不同的地方在於 cloudcover、temp 與 humidity 這幾個原本在 rf 中沒有明顯影響的變數在此都有明顯的趨勢。

# VI. 結論

1. 我們發現兩個模型中，露點溫度都是其最重要的變數，所以我們可以針對這個變數去理解其跟健康風險的關係。
2. 前面也有看到在某個州的健康風險分數可能較高，未來也可以加入更多其它變數，例如:產業、地形等等，說不定我們就可以更清楚的知道造成這個狀況的原因。
3. 最終模型比較:

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
|---|---|---|---|---|---|---|---|
| Support Vector Regression | | 0.0886 | 0.0122 | 0.1106 | 0.9720 | 0.0104 | 0.0092 |

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
|---|---|---|---|---|---|---|---|
| Random Forest Regressor | | 0.0737 | 0.0097 | 0.0984 | 0.9779 | 0.0094 | 0.0077 |

在測試集中，可以發現 Random Forest 整體的模型評分都較好，所以我們會選擇此模型來預測健康風險分數。

# VII. 參考資料:


資料來源: https://www.kaggle.com/datasets/abdullah0a/urban-air-quality-and-health-impact-dataset/data

風險分數: https://www.alberta.ca/about-the-air-quality-health-index

熱指數: https://www.weather.gov/ama/heatindex

體感溫度：https://pansci.asia/archives/93280

紫外線指數：https://www.cwa.gov.tw/Data/knowledge/announce/service13.pdf

Chat gpt

- 分工情形:

吳俞憲(50%):整體溝思、程式處理、書面報告最後部分

廖廣筑(50%):提供意見、製作 ppt 與報告、書面報告前大半部分