

Abstract geometric lines in the top left corner, consisting of several overlapping, irregular polygons and lines in a light beige color.

空氣品質與健康狀況

M132040019 廖廣筑

M122040017 吳俞憲

資料集介紹

Urban Air Quality and Health Impact

此資料集含有1000筆資料，並且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

| data | | | | | | | | | | | | | | | | |
|------|------------|---------------|------------|-----------|-----------|--------------|--------------|-----------|-----------|-----------|-----|--------------|------------|------------|----------------|----------------|
| | datetime | datetimeEpoch | tempmax | tempmin | temp | feelslikemax | feelslikemin | feelslike | dew | humidity | ... | City | Temp_Range | Heat_Index | Severity_Score | Condition_Code |
| 0 | 2024-09-07 | 1.725692e+09 | 106.100000 | 91.000000 | 98.500000 | 104.000000 | 88.100000 | 95.900000 | 51.500000 | 21.000000 | ... | Phoenix | 15.100000 | 95.918703 | 4.430000 | NaN |
| 1 | 2024-09-08 | 1.725779e+09 | 103.900000 | 87.000000 | 95.400000 | 100.500000 | 84.700000 | 92.300000 | 48.700000 | 21.500000 | ... | Phoenix | 16.900000 | 92.281316 | 3.880000 | 0.0 |
| 2 | 2024-09-09 | 1.725865e+09 | 105.000000 | 83.900000 | 94.700000 | 99.900000 | 81.600000 | 90.600000 | 41.700000 | 16.900000 | ... | Phoenix | 21.100000 | 90.599165 | 3.630000 | 0.0 |
| 3 | 2024-09-10 | 1.725952e+09 | 106.100000 | 81.200000 | 93.900000 | 100.600000 | 79.500000 | 89.800000 | 39.100000 | 15.700000 | ... | Phoenix | 24.900000 | 89.638811 | 2.851200 | 0.0 |
| 4 | 2024-09-11 | 1.726038e+09 | 106.100000 | 82.100000 | 94.000000 | 101.000000 | 80.000000 | 90.000000 | 40.100000 | 15.900000 | ... | Phoenix | 24.000000 | 89.760414 | 3.390800 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 995 | 2024-09-18 | 1.726633e+09 | 76.060546 | 64.359387 | 69.002142 | 77.673823 | 63.510920 | 67.003338 | 59.936776 | 73.189130 | ... | Philadelphia | 12.886665 | 71.837558 | 1.957318 | NaN |
| 996 | 2024-09-17 | 1.726550e+09 | 68.409198 | 65.939319 | 66.567410 | 68.956722 | 64.805635 | 65.992526 | 59.010257 | 74.137401 | ... | San Diego | 2.613629 | 72.463491 | 2.537413 | 0.0 |
| 997 | 2024-09-12 | 1.726122e+09 | 69.756690 | 65.286919 | 65.919492 | 68.158536 | 63.662942 | 67.313322 | 62.024442 | 84.650482 | ... | San Diego | 4.598936 | 67.560060 | 3.595470 | NaN |
| 998 | 2024-09-14 | 1.726284e+09 | 77.106797 | 61.481724 | 68.106569 | 76.426959 | 60.901526 | 68.094309 | 63.169608 | 86.860261 | ... | Los Angeles | 15.477717 | 67.930437 | 3.498942 | 0.0 |
| 999 | 2024-09-18 | 1.726618e+09 | 90.923080 | 79.296868 | 81.636991 | 94.180423 | 78.071851 | 84.987113 | 73.393045 | 74.734715 | ... | Houston | 11.017871 | 86.802712 | 3.040020 | 0.0 |

資料集變數介紹

天氣：

1. Condition_Code(特定天氣狀況的代碼)

降水：

1. Precip_Type(降水類型)
2. Precipitation(當天的總降水量)
3. Precip_Prob(降水的機率)
4. Precip_Cover(降水的覆蓋範圍)

風：

1. Wind_Gust(當天的最大陣風速度)
2. Wind_Speed(當天的平均風速)
3. Wind_Direction(風向)

溫度：

1. Temp_Max(當天的最高溫度)
2. Temp_Min(當天的最低溫度)
3. Temp_Avg(當天的平均溫度)
4. Temp_Range(當天的溫差)

太陽：

1. Solar_Radiation(太陽輻射)
2. Solar_Energy(接收到的太陽能量)

資料集變數介紹

時間：

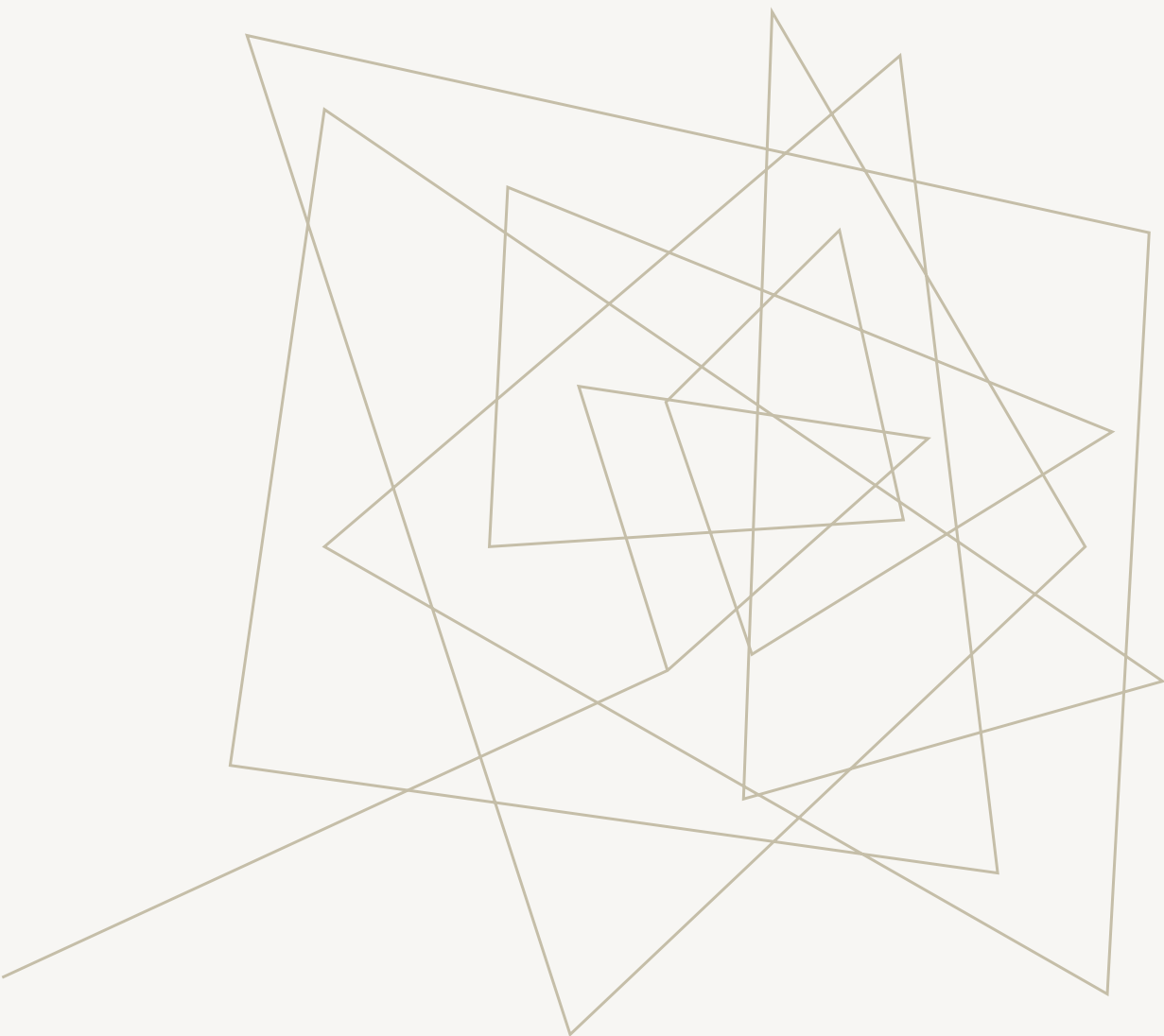
1. Day_of_Week(星期幾)
2. Is_Weekend(是否為週末)
3. Month(月份)
4. Season(季節)
5. datetime(確切日期)

地點：

1. Stations(提供資料的氣象站)
2. City(城市名稱)

其他：

1. Dew_Point(露點溫度)
2. Humidity(相對濕度)
3. Pressure(氣壓)
4. Cloud_Cover(雲量覆蓋率)
5. Heat_Index(熱指數)
6. Snow_Depth(雪深)
7. Source(資料來源)

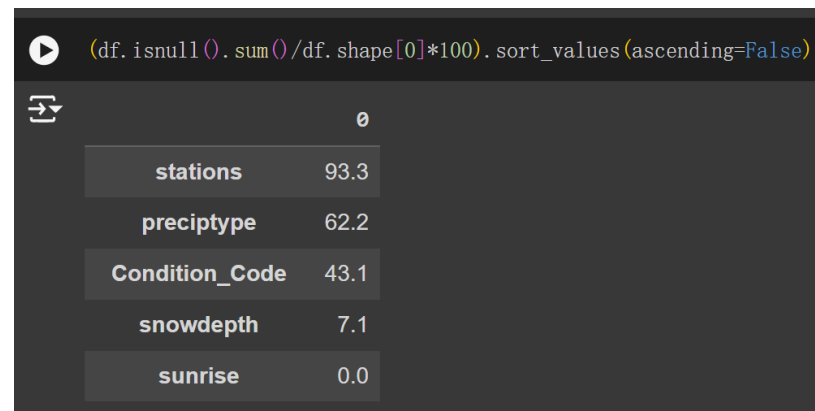


DATA PREPARATION

處理缺失值

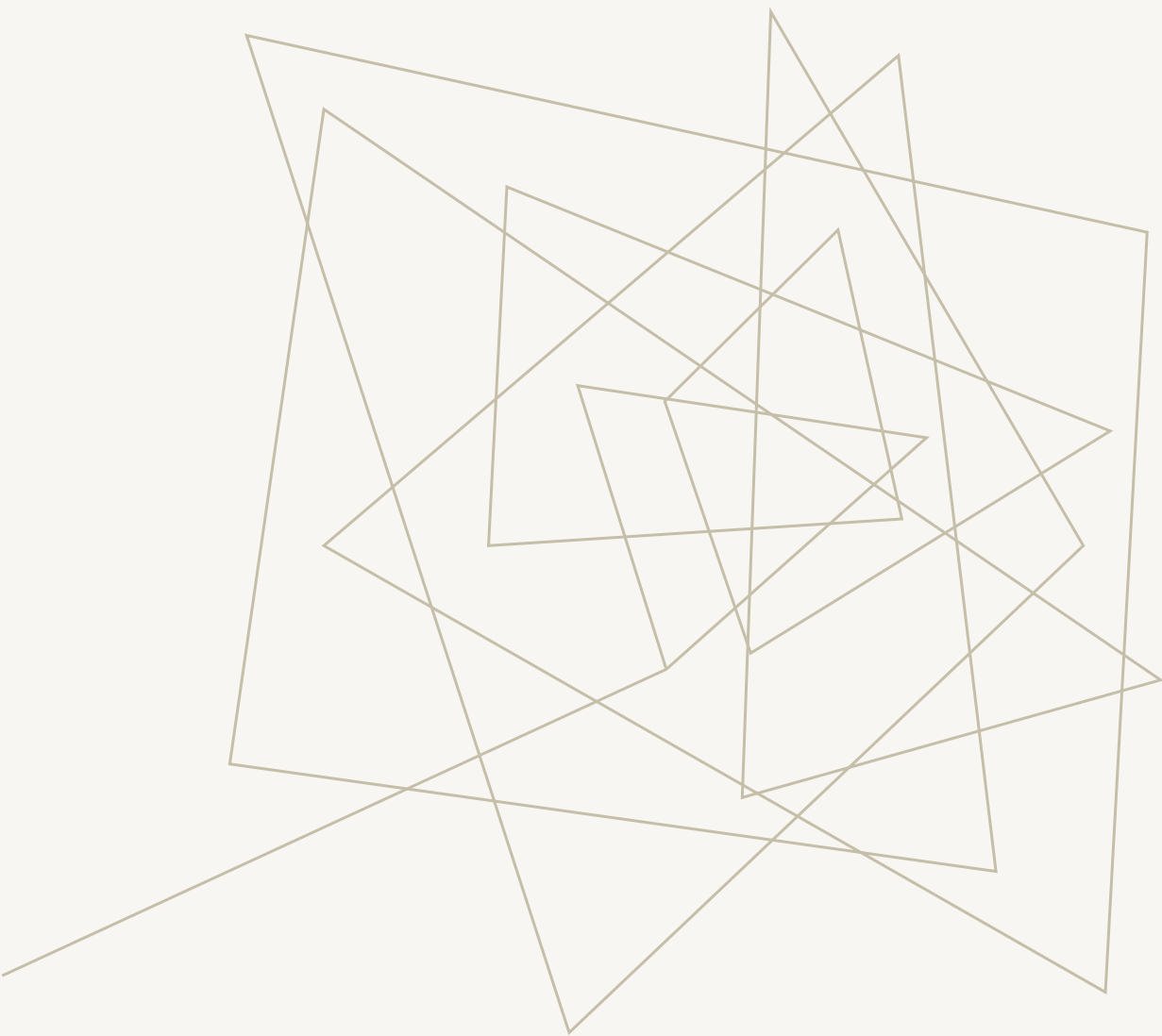
有缺失值的變數：

1. **Condition_Code**：因為不確定是什麼，而且有其他天氣相關的變數，所以不考慮此變數
2. **Stations**：缺失值太高直接不考慮此變數
3. **Preciptype**：只要是否紀錄下雨，所以處理得方式就是把缺失值填補為0代表沒下雨
4. **Snowdepth**：因為他不是nan就是0，所以也不考慮此變數



The screenshot shows a Jupyter Notebook interface. At the top, a code cell contains the command: `(df.isnull().sum()/df.shape[0]*100).sort_values(ascending=False)`. Below the code cell, the output is displayed as a table. The table has two columns: the variable name and its percentage of missing values. The variables are listed in descending order of missing value percentage.

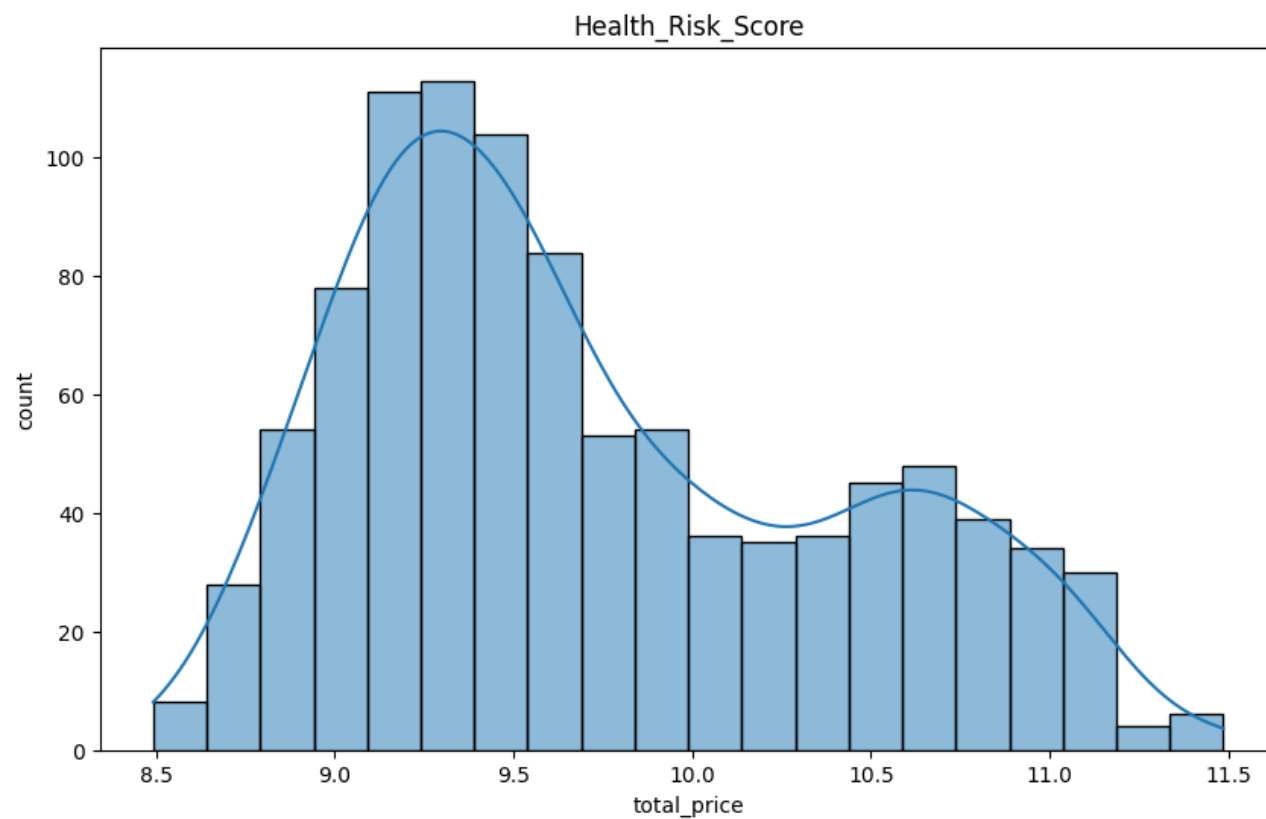
| | 0 |
|----------------|------|
| stations | 93.3 |
| preciptype | 62.2 |
| Condition_Code | 43.1 |
| snowdepth | 7.1 |
| sunrise | 0.0 |



變數探索

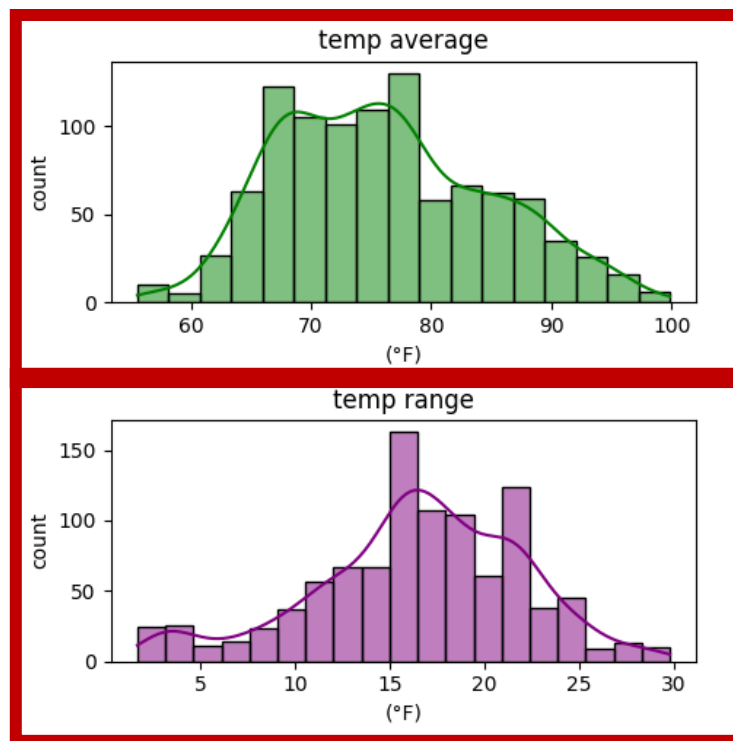
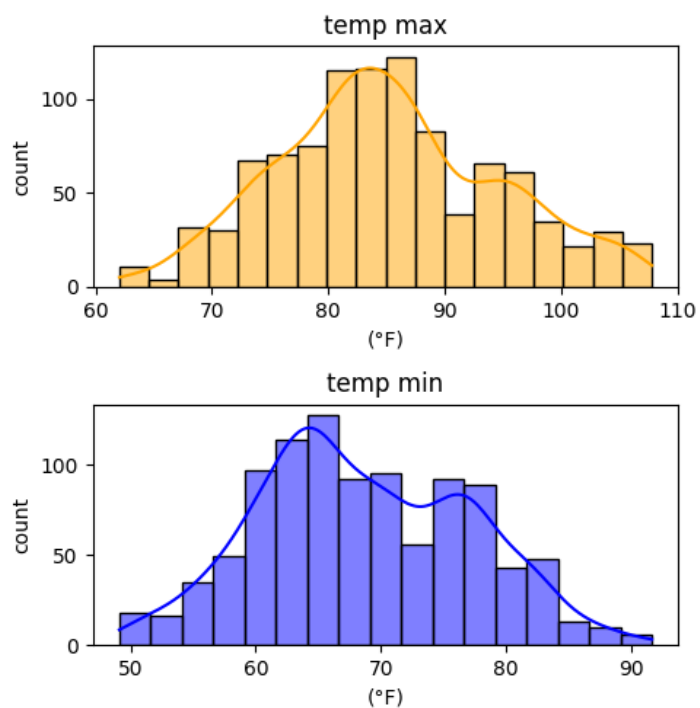
數值變數：HEALTH_RISK_SCORE

健康風險指標：



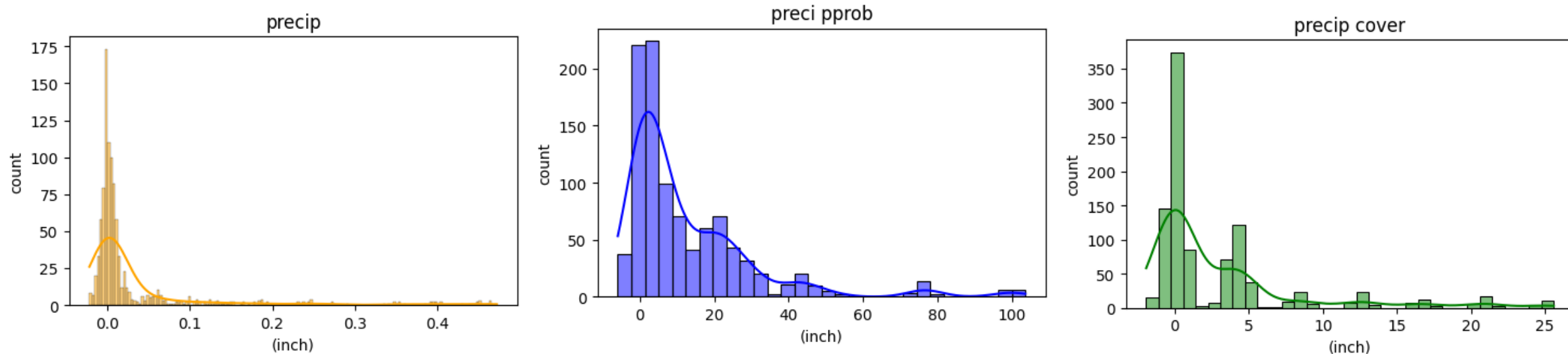
健康風險指標數值界在8.5~11.5之間，
數據大部分介在9~10之間

數值變數：溫度



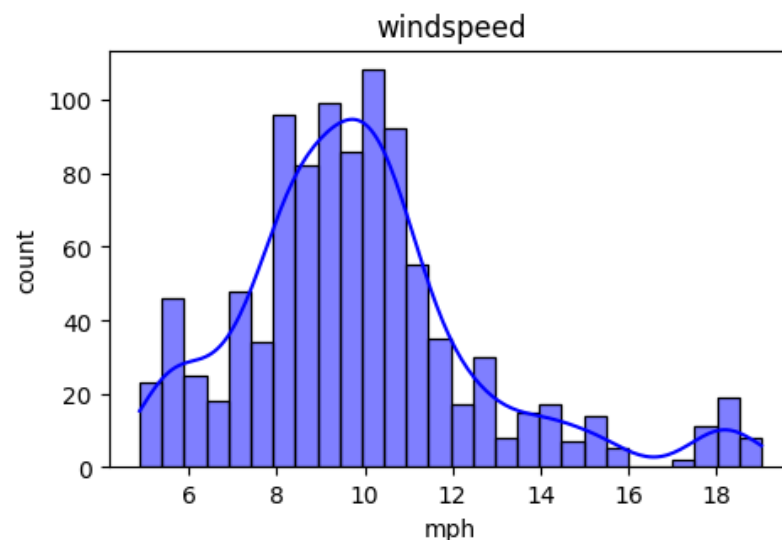
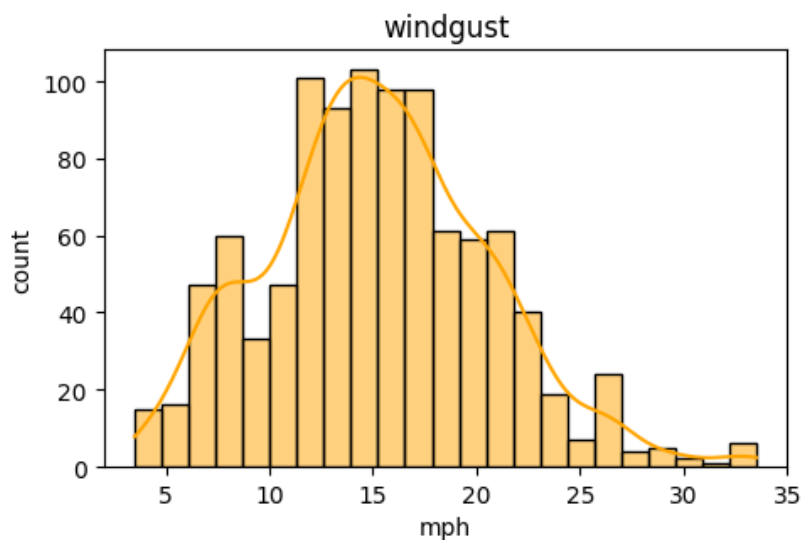
這四個溫度的分布都不極端，且平均溫度跟溫差較可以代表當日溫度狀況

數值變數：降雨



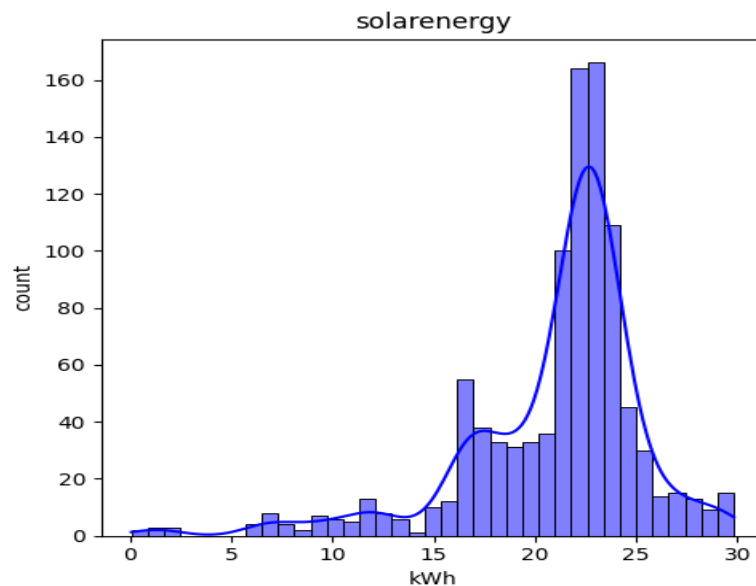
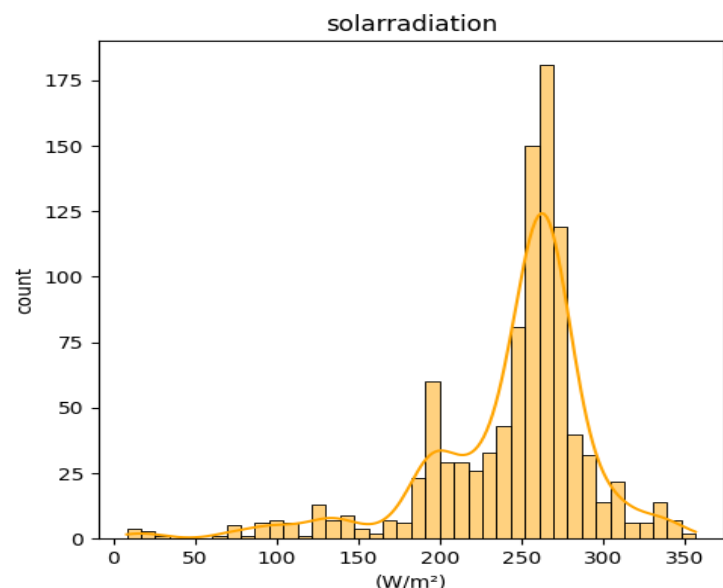
這三個降雨相關的圖有嚴重的右偏，而且還有負數的出現，可能是資料填的時候就有錯誤，所以我們不會考慮這三個變數

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

數值變數：太陽變數



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

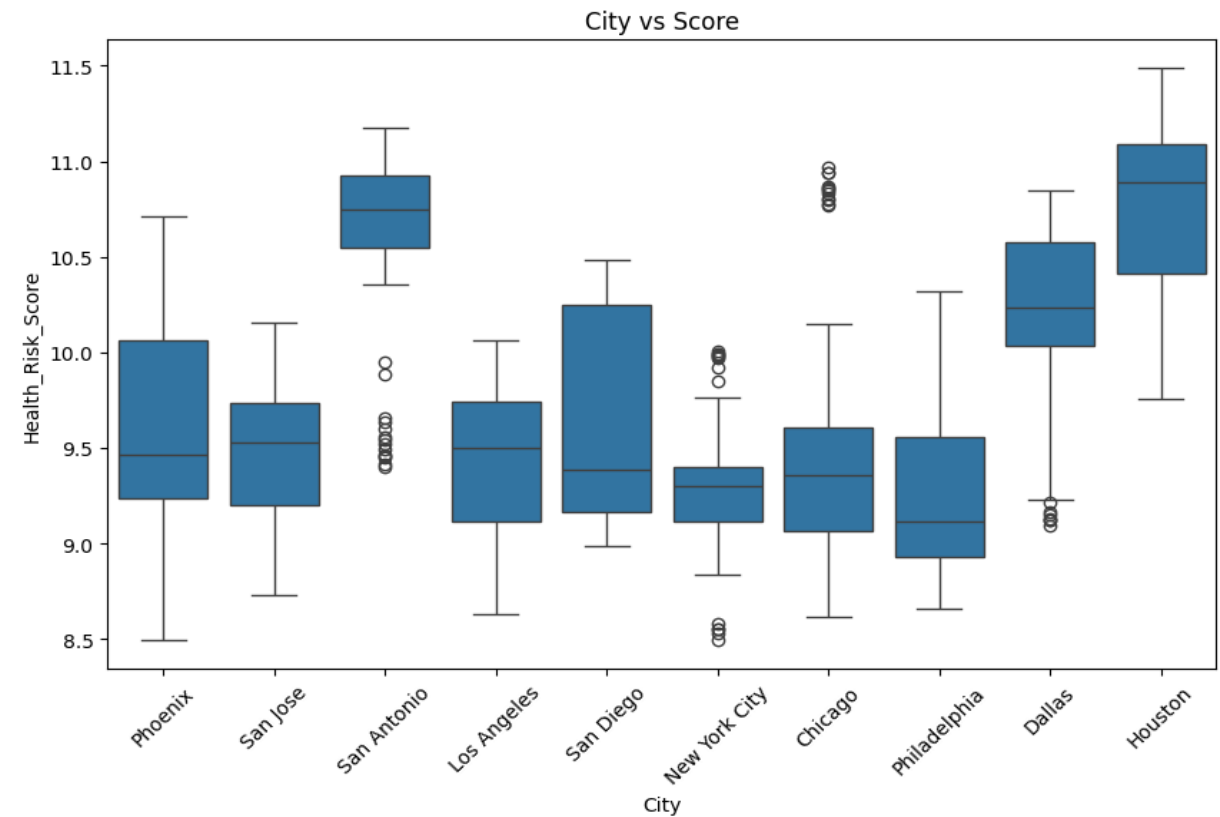
```
0.9912955859677877
```

這兩個變數的分配一樣是因為：

1. **Solar Radiation**(太陽輻射)：每平方公尺在單位時間內的能量輸入
2. **Solar Energy**(接收到的太陽能量)：每小時的能量輸入

他們的相關性很高，因此只考慮**Solar Radiation**(太陽輻射)

類別變數：地區與健康風險指標

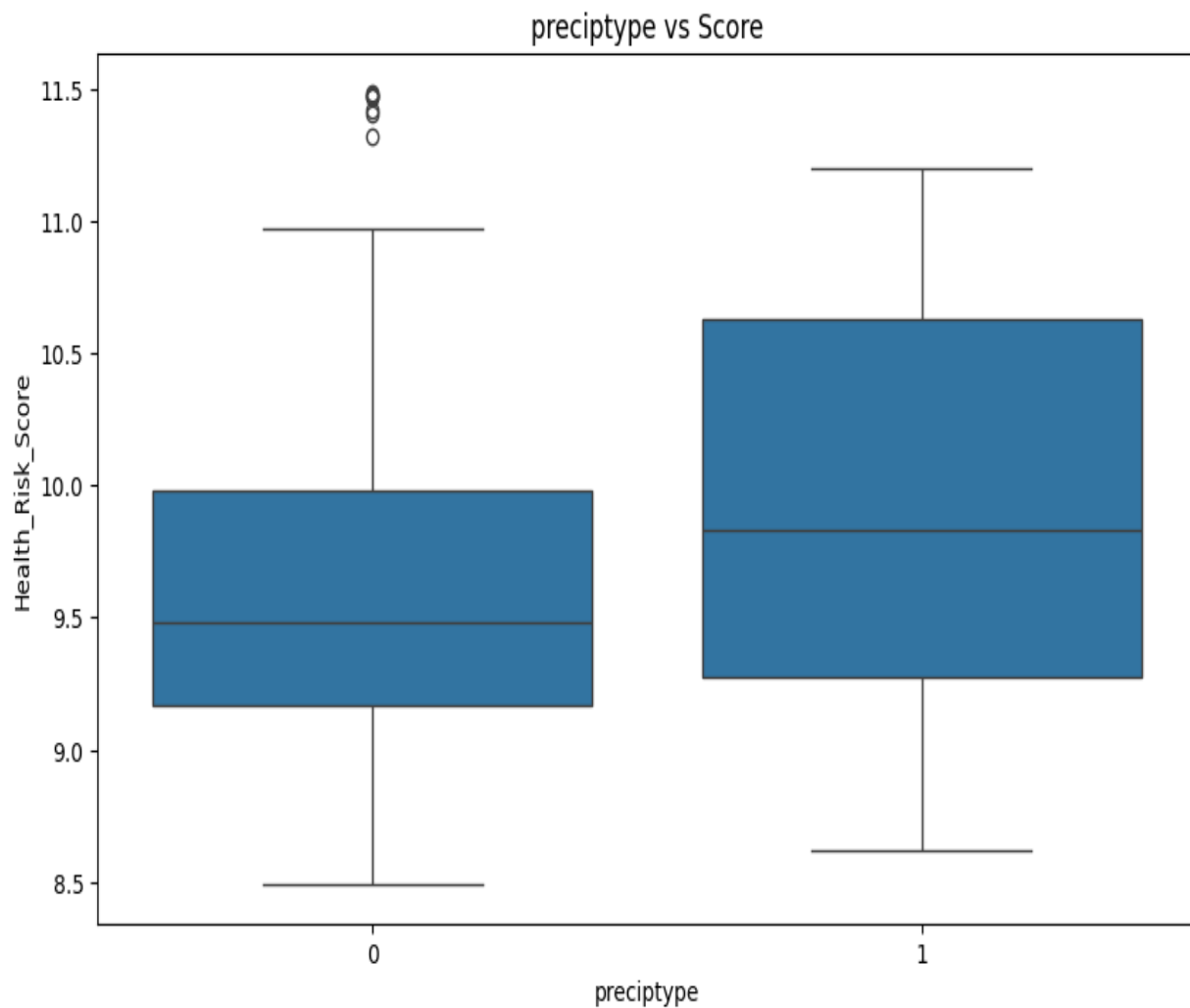


健康風險指標較高的地區

1. San Antonio
2. Dallas
3. Houston

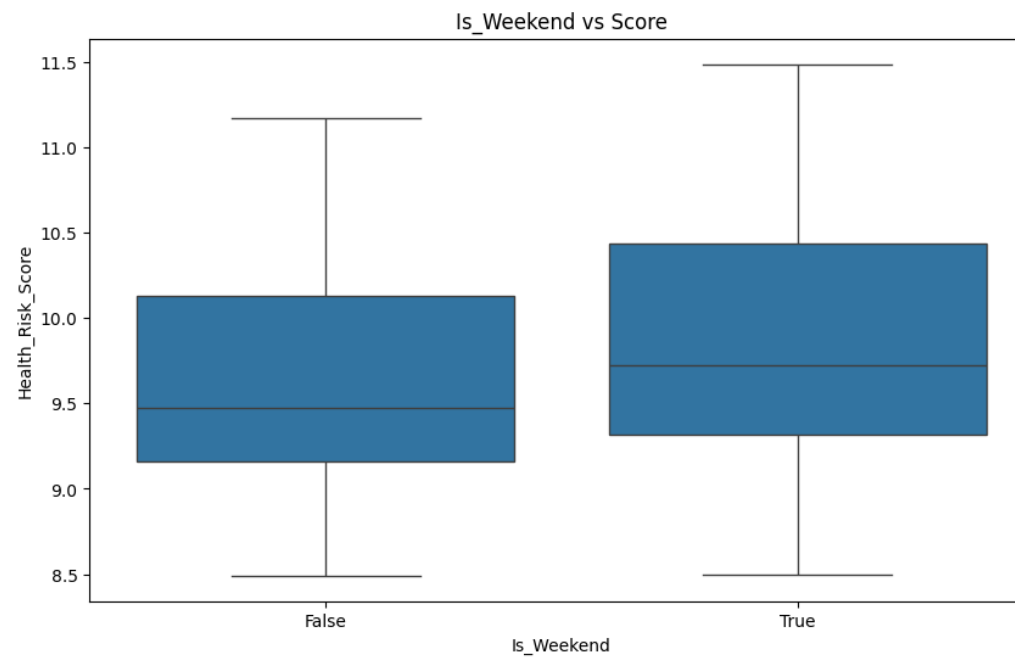
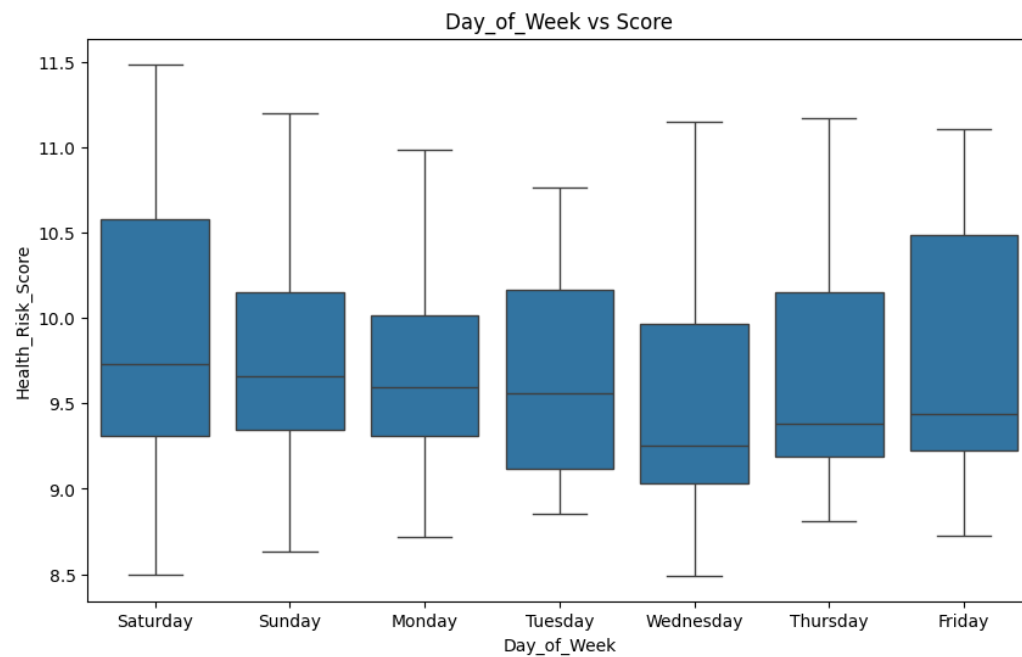


類別變數：是否降雨與健康風險指標

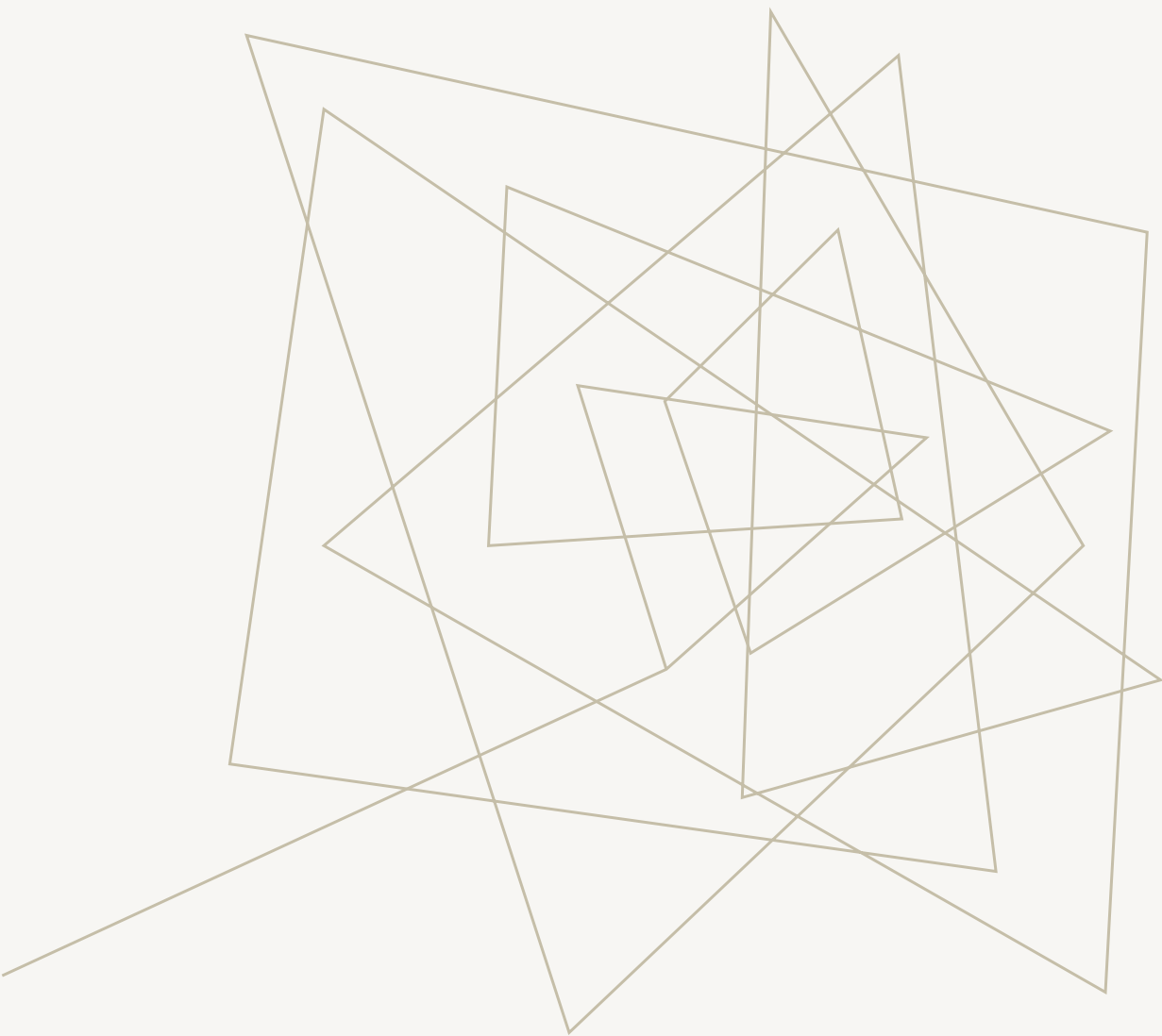


Precip type (0：沒降雨，1：有降雨)
兩者差異不大，
但是有降雨的部分健康風險指標稍微
高一點

類別變數：星期幾與健康風險指標



圖看起來都差不多，這或許沒有特別的影響



特徵工程

特徵工程

- 新變數風速比率: max/average

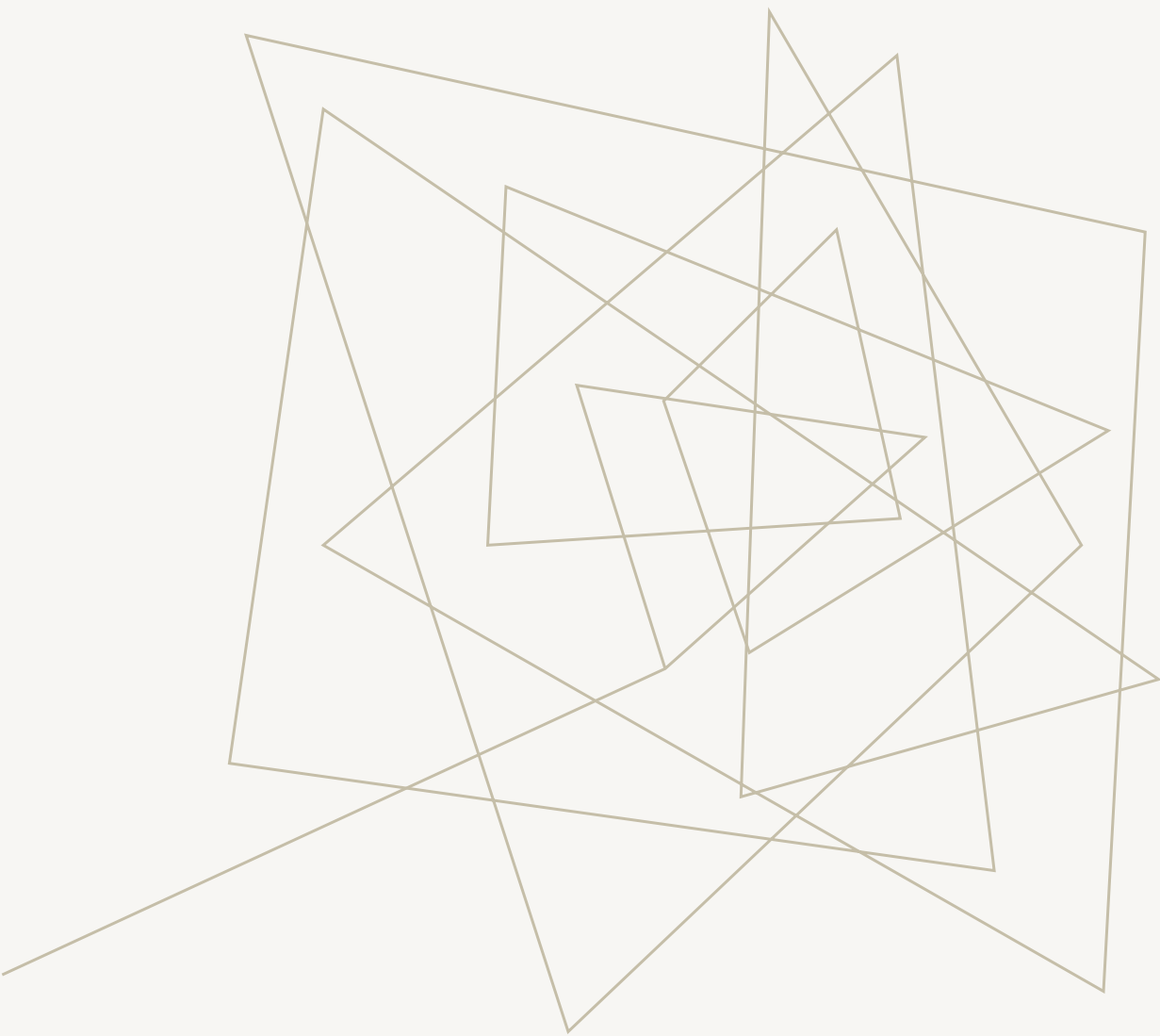
```
[ ] df['windratio'] = df['windgust'] / df['windspeed']
```

- 對城市做Target encoding

```
[ ] mean_target = df.groupby('City')['Health_Risk_Score'].mean()  
    df['city_target_encoded'] = df['City'].map(mean_target)
```

- 是否為周末改為0,1

```
[ ] df['Is_Weekend'] = df['Is_Weekend'].apply(lambda x: 1 if str(x).upper() == "TRUE" else 0)
```



相關係數熱力圖

相關係數熱力圖

先刪除未知變數與同類型變數，剩餘變數：

Dew_Point(露點溫度)、Humidity(相對濕度)、Pressure(氣壓)、

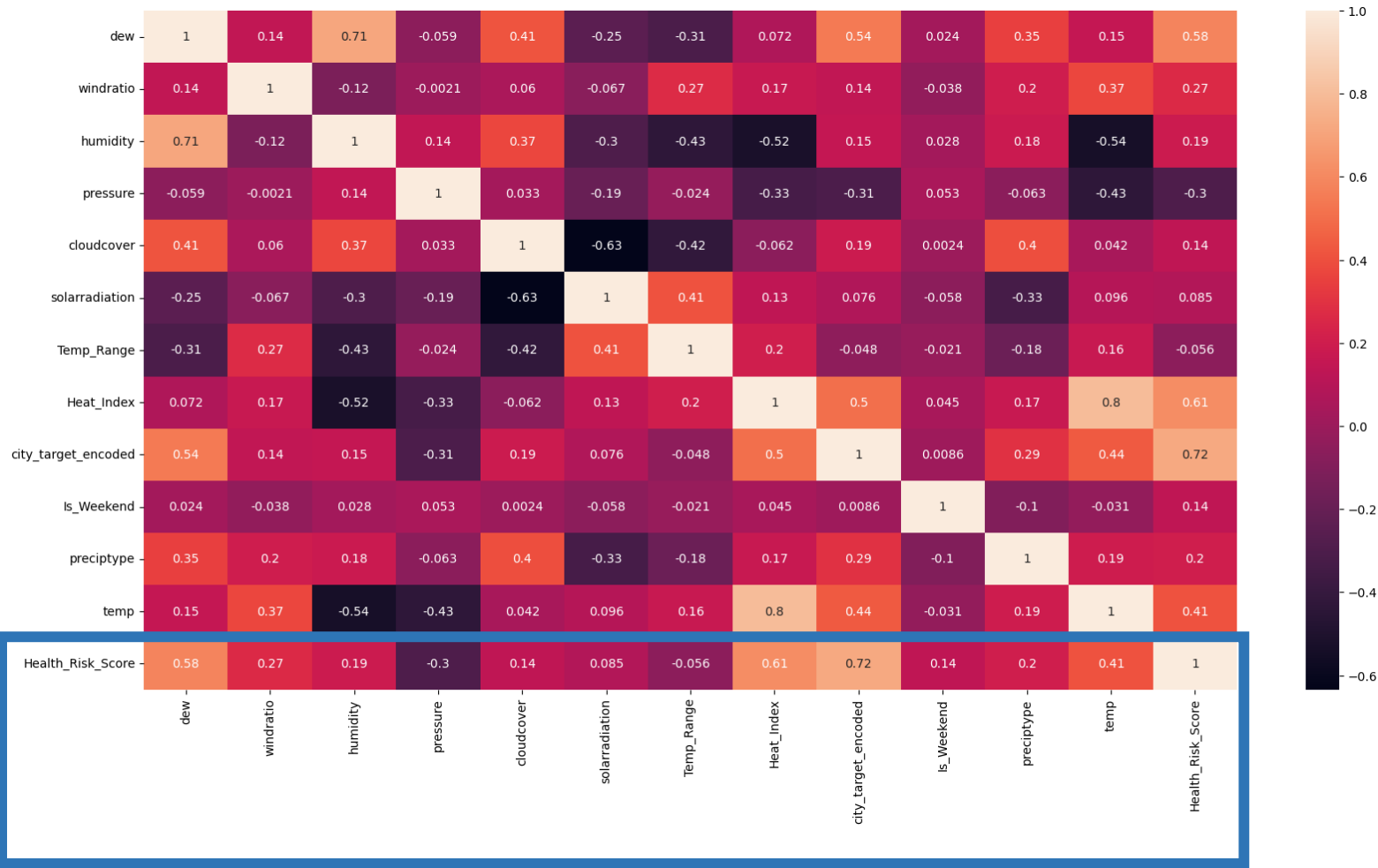
Cloud_Cover(雲量覆蓋率)、Solar_Radiation(太陽輻射)、

Temp_Range(溫差)、Temp_Avg(平均溫度)、

Heat_Index(熱指數)、City(城市)、Is_Weekend(是否周末)、

Precip_Type(降水類型)

相關係數熱力圖



Solar_Radiation跟
Temp_Range與
健康風險指標相關性
太低，所以後續跑模
型時將他們去除



MODEL BUILDING

模型比較

我們將資料分為90%訓練集與10%測試集，並且有做標準化，之後篩選模型：

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

可以發現 **rf** 跟 **svm** 在各項目中都是相對較好的

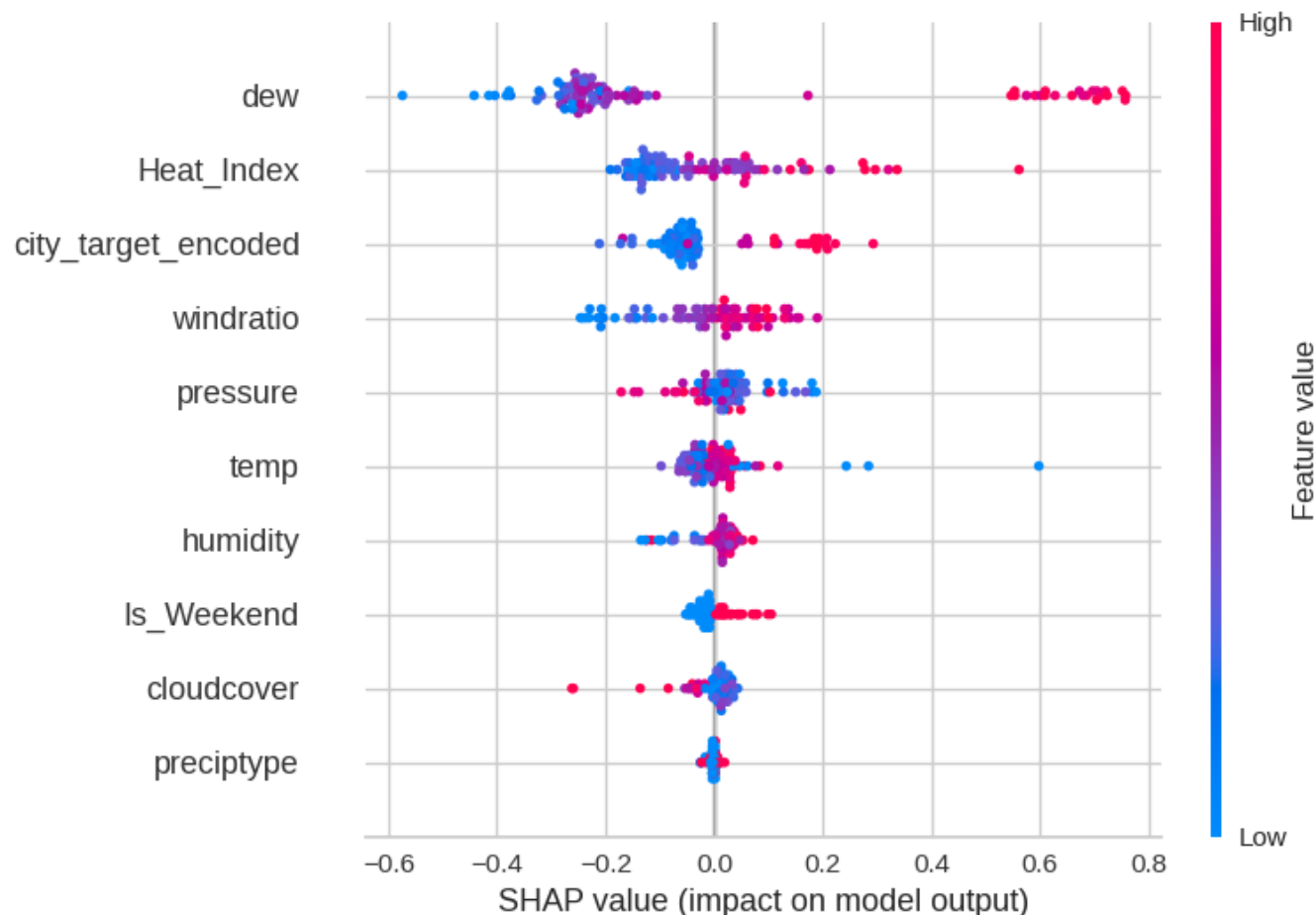


MODEL INTERPRETATION

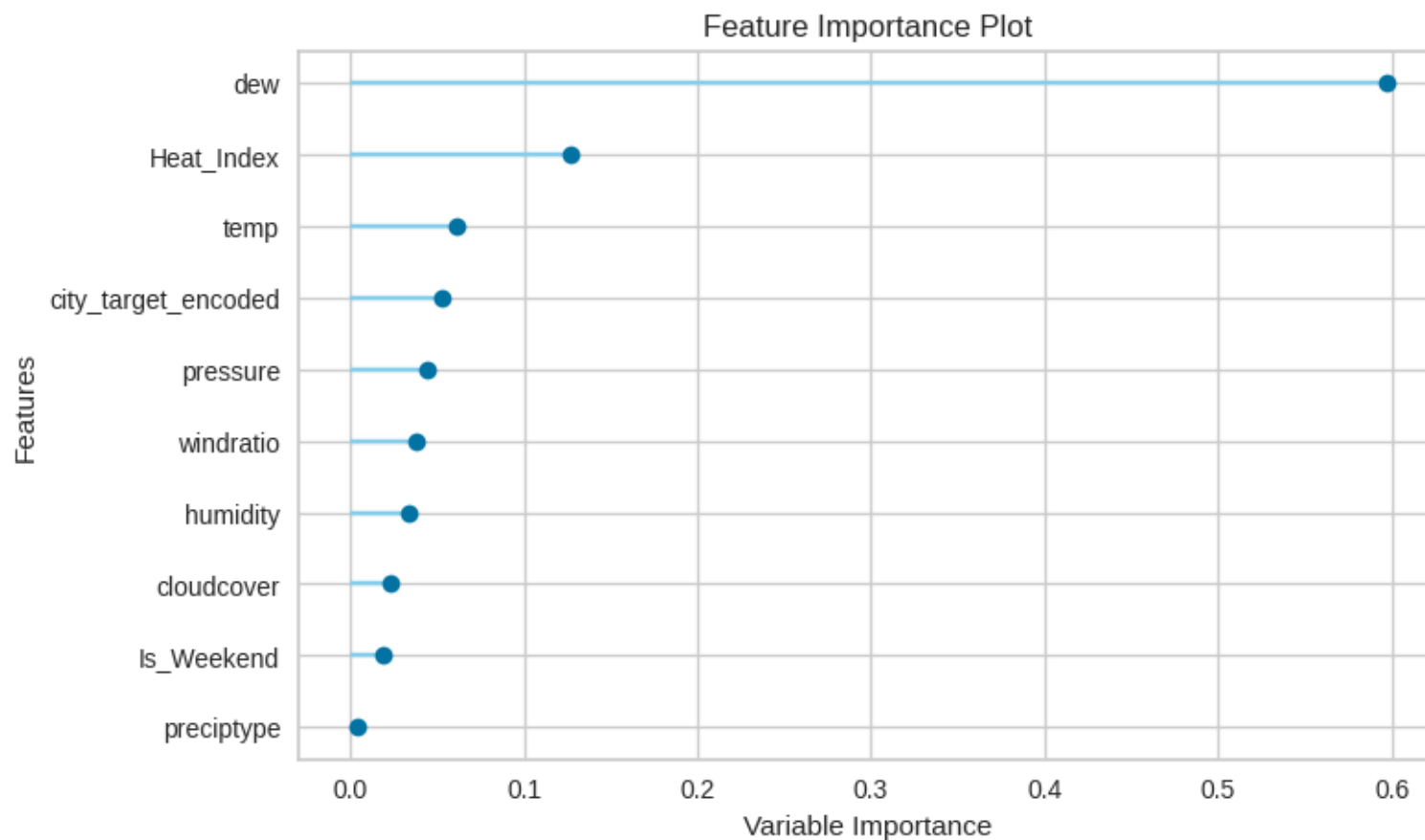
模型解釋：RANDOM FOREST REGRESSOR

我們用SHAP的方式去解釋RF模型

1. Dew_Point(露點溫度)跟 Heat_Index(熱指數)的SHAP值分佈較廣，所以對模型預測有較大的影響
2. Pressure(氣壓)紅色點幾乎都分佈在左邊，所以他的數值增加，有潛在可能健康風險指標會降低



模型解釋：RANDOM FOREST REGRESSOR



由左圖可知：

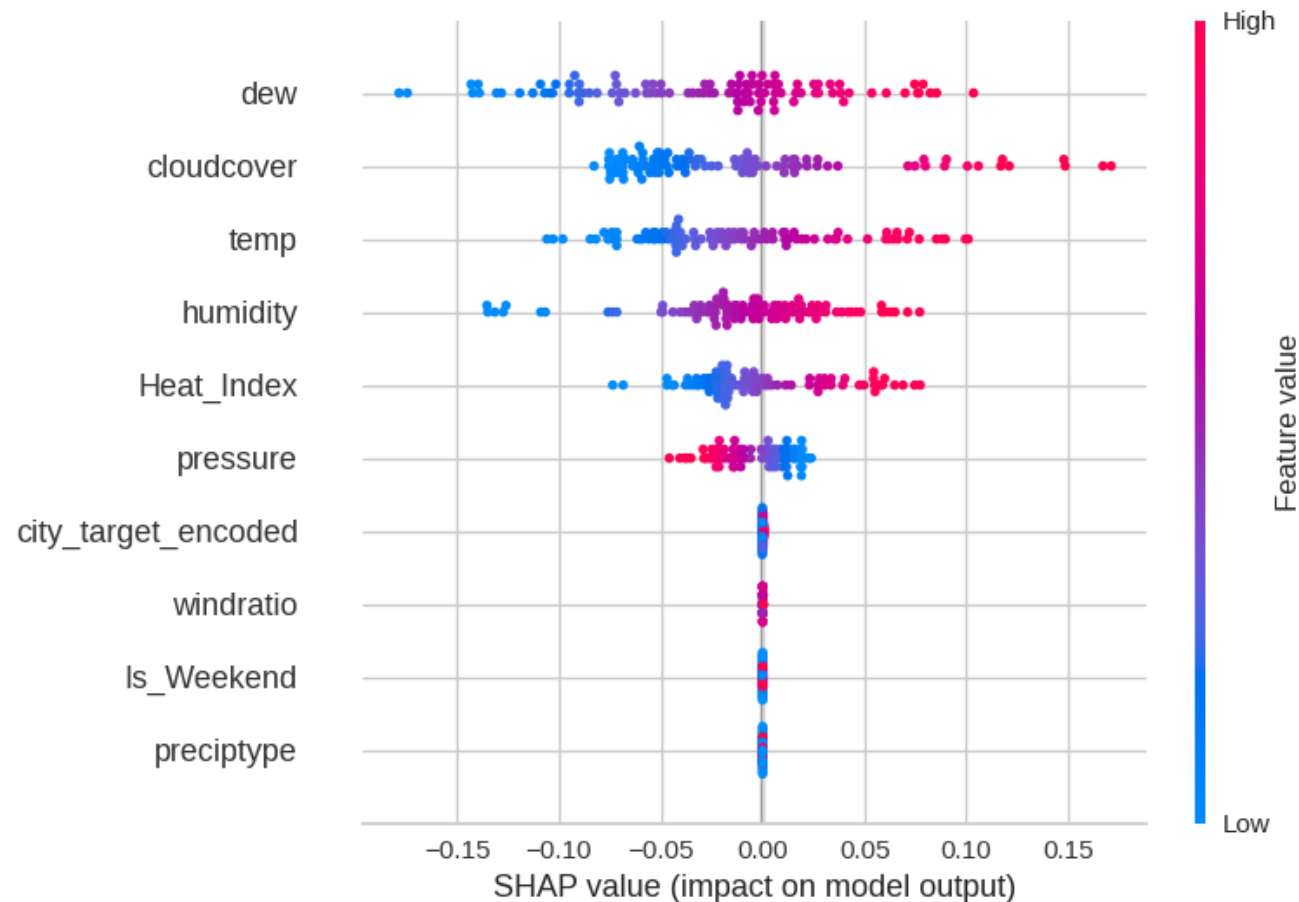
Dew_Point(露點溫度) 的確對預測健康風險指標是一個很重要的變數

而Heat_Index(熱指數)是其次重要

模型解釋：SUPPORT VECTOR REGRESSION

我們用SHAP的方式去解釋SVR模型，

1. Dew_Point(露點溫度)跟 Cloud_Cover(雲量覆蓋率) SHAP值分佈較廣，所以對模型預測有較大的影響
2. Pressure(氣壓)紅色點幾乎都分佈在左邊，所以他的數值增加，有潛在可能健康風險指標會降低



結論

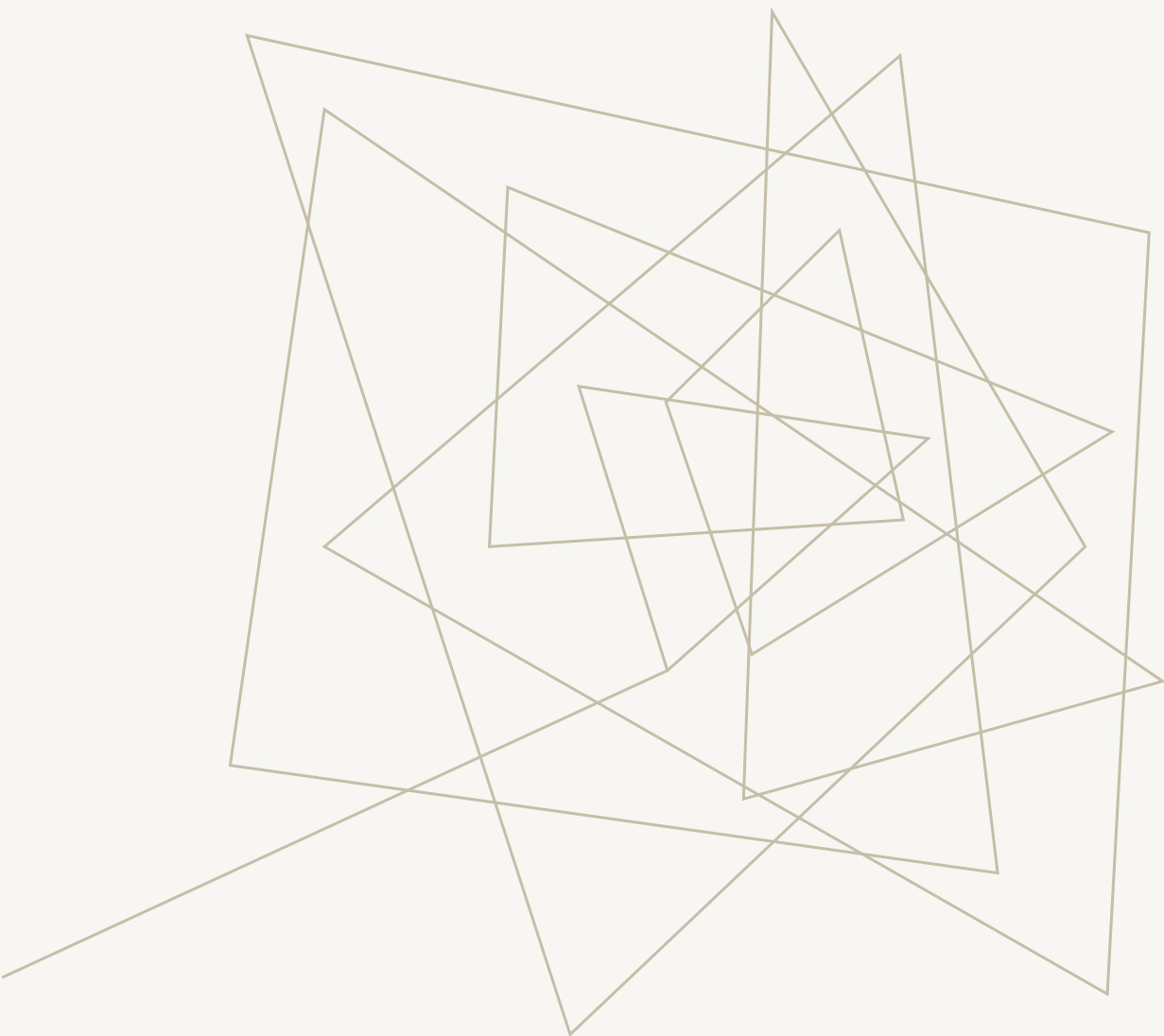
1. SVM :

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

2. RF :

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

- 在這兩個模型中，Dew_Point(露點溫度) 都是影響預測結果的重要特徵
- 在這兩個模型中，Random Forest Regressor的指標優於Support Vector Regression，所以我們會選擇Random Forest Regressor



THANKS