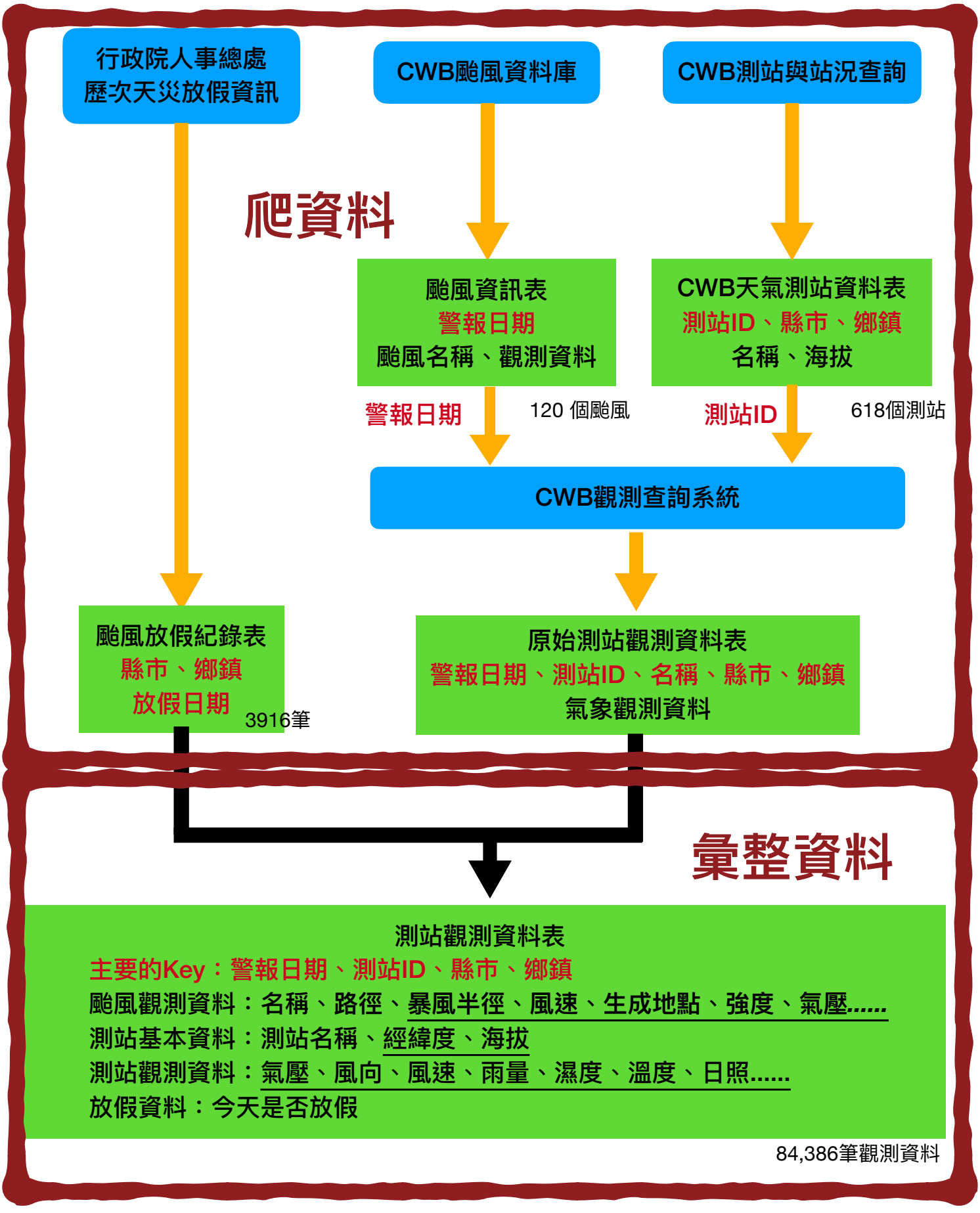
資料取得與彙整


行政院人事總處
歷次天災放假資訊
CWB颱風資料庫
CWB測站與站況查詢
爬資料
颱風資訊表
警報日期
颱風名稱、觀測資料
CWB天氣測站資料表
測站ID、縣市、鄉鎮
名稱、海拔
警報日期
120 個颱風
測站ID
618個測站
CWB觀測查詢系統
颱風放假紀錄表
縣市、鄉鎮
放假日期
3916筆
原始測站觀測資料表
警報日期、測站ID、名稱、縣市、鄉鎮
氣象觀測資料
彙整資料
測站觀測資料表
主要的Key：警報日期、測站ID、縣市、鄉鎮
颱風觀測資料：名稱、路徑、暴風半徑、風速、生成地點、強度、氣壓.....
測站基本資料：測站名稱、經緯度、海拔
測站觀測資料：氣壓、風向、風速、雨量、濕度、溫度、日照.....
放假資料：今天是否放假
84,386筆觀測資料
資料來源
Dataframe
Python
requests, bs4
Python Pandas

資料前處理 - 分組

測站觀測資料表

主要的Key：警報日期、測站ID、縣市、鄉鎮

颱風觀測資料：名稱、路徑、暴風半徑、風速、生成地點、強度、氣壓.....

測站基本資料：測站名稱、經緯度、海拔

測站觀測資料：氣壓、風向、風速、雨量、濕度、溫度、日照.....

放假資料：今天是否放假

84,386 筆觀測資料

風向&風速轉換成向量

新增欄位：明天是否放假

依鄉鎮分組

警報日期、測站ID、縣市、鄉鎮
颱風觀測資料、是否放假
(數值型、類別型混合)

丟棄重複的row

測站觀測資料
(都是數值型)

依照縣市/鄉鎮分組後取平均

資料整合整合

資料集 - 鄉鎮版

主要的Key：警報日期、測站ID、縣市

颱風觀測資料：名稱、路徑、暴風半徑、風速、生成地點、強度、氣壓.....

測站基本資料：測站名稱、經緯度、海拔

測站觀測資料：氣壓、風向量、雨量、濕度、溫度、日照.....

放假資料：今天是否放假、明天是否放假

25,909 筆資料
44 欄位

DataFrame　Processed DataFrame　R

資料前處理、特徵工程與建模評估 - 1

資料集 - 以 鄉鎮 版為例

25,909 筆資料
44 欄位

非訓練特徵：~~警報日期、測站ID、各種名稱、所屬行政區域...~~

類別型特徵：颱風路徑分類、颱風強度分類

數值型特徵：

1. 地點基本資料：經緯度、海拔
2. 颱風相關特徵：暴風半徑、颱風風速、生成地點、強度、氣壓.....
3. 測站觀測資料：氣壓、風向量、雨量、濕度、溫度、日照.....
4. 放假情形：今天是否放假、明天是否放假

NA前處理

1. 刪除NA過多的Row
2. 確認不合理的數值並強迫令為NA等待補值

(11050, 31)

特徵前處理

3. 丟棄非訓練特徵
4. 將類別型特徵轉為dummy

標準化與補值

6. 用MinMaxScaler將數值壓縮在0 ~ 1之間
7. 使用kNN補值

資料前處理、特徵工程與建模評估 - 2