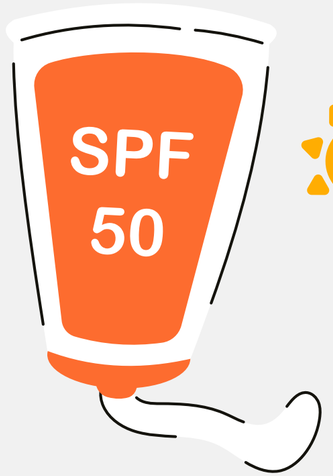
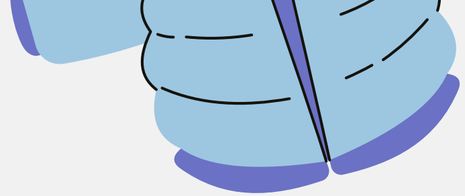
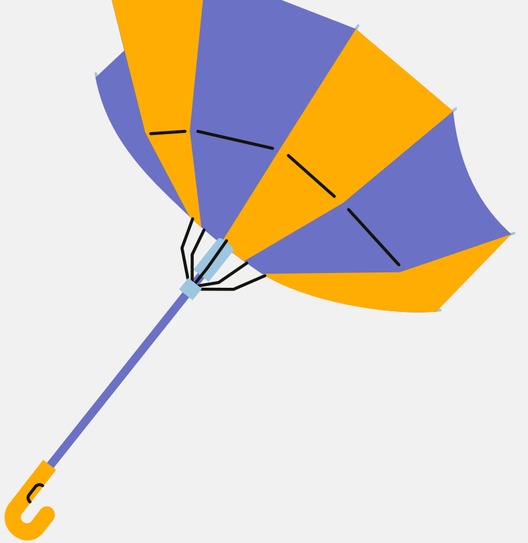






WEATHER FORECAST

B1228006 鄭暐薰
B1228022 梁釗豪
B1228029 蘇琪文

競賽任務-預測

Rain in Australia

分析

降雨預測模型
時間序列分析與趨勢預測
特徵工程與模型解釋
地區氣候比較與分群

應用

農業與能源應用
異常偵測

降雨預測模型

根據當天的氣象條件預測隔天是否會下雨。

時間序列分析與趨勢預測

探索季節性與週期性模式預測未來天氣

特徵工程與模型解釋

分析哪些變數(如濕度、氣壓)對降雨影響最大

地區氣候比較與分群

根據 Location 欄位比較不同城市的氣候特性

應用

農業與能源應用

預測農業灌溉時機，太陽能與風能產量

異常偵測

偵測異常氣象事件,eg:異常高溫或低溫、突然暴雨或乾旱

資料集

Rain in Australia

https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Rain in Australia

資料長相

表格(CSV)格式
每筆資料代表澳洲某地某日的氣象觀測值
包含超過20 個欄位
涵蓋溫度、濕度、風速、氣壓、雲量、日照等

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0	23.2	No	No

EDA(資料分析)

資料集的基本資訊和結構

資料筆數：145460 筆
欄位數量：23 欄
dtypes:
float64(16), object(7)
memory usage:
25.5+ MB

欄位名稱	資料型態	說明（補充）
Date	object	日期（需轉換爲 datetime）
Location	object	城市或地區名稱
MinTemp	float64	當日最低氣溫（攝氏）
MaxTemp	float64	當日最高氣溫（攝氏）
Rainfall	float64	降雨量（毫米）
Evaporation	float64	蒸發量（毫米）
Sunshine	float64	日照時數（小時）
WindGustDir	object	最大陣風方向（類別）
WindGustSpeed	float64	最大陣風速度（km/h）
WindDir9am	object	上午 9 點風向（類別）
WindDir3pm	object	下午 3 點風向（類別）
WindSpeed9am	float64	上午 9 點風速（km/h）
WindSpeed3pm	float64	下午 3 點風速（km/h）
Humidity9am	float64	上午 9 點濕度（%）
Humidity3pm	float64	下午 3 點濕度（%）
Pressure9am	float64	上午 9 點氣壓（hPa）
Pressure3pm	float64	下午 3 點氣壓（hPa）
Cloud9am	float64	上午 9 點雲量（0–8）
Cloud3pm	float64	下午 3 點雲量（0–8）
Temp9am	float64	上午 9 點氣溫（攝氏）
Temp3pm	float64	下午 3 點氣溫（攝氏）
RainToday	object	當天是否下雨（Yes/No）
RainTomorrow	object	隔天是否下雨（Yes/No）

EDA(資料分析)

資料內容

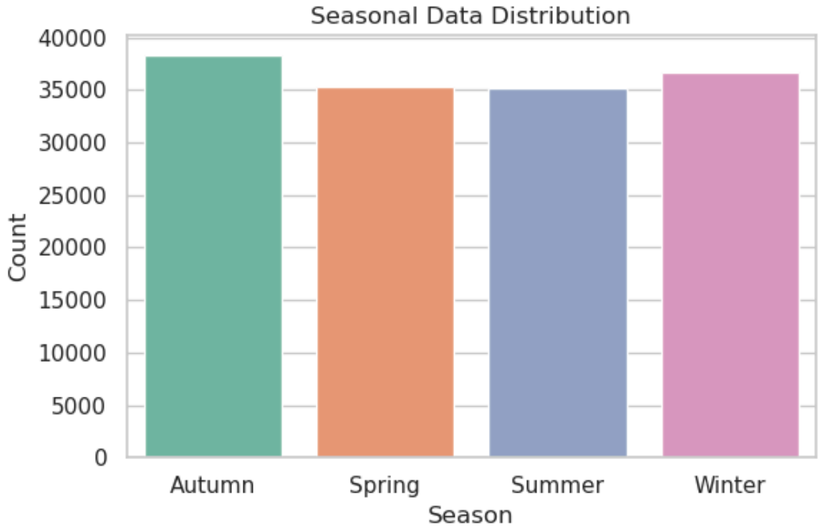
Date:
最早日期：2007-11-01
最晚日期：2017-06-25
locations:(49 個地點)
'Adelaide', 'Albany',
'Albury', ... , 'Witchcliffe',
'Wollongong', 'Woomera'

	mean	std	min	max
MinTemp	12.194034	6.398495	-8.5	33.9
MaxTemp	23.221348	7.119049	-4.8	48.1
Rainfall	2.360918	8.47806	0	371
Evaporation	5.468232	4.193704	0	145
Sunshine	7.611178	3.785483	0	14.5
WindGustSpeed	40.03523	13.607062	6	135
WindSpeed9am	14.043426	8.915375	0	130
WindSpeed3pm	18.662657	8.8098	0	87
Humidity9am	68.880831	19.029164	0	100
Humidity3pm	51.539116	20.795902	0	100
Pressure9am	1017.64994	7.10653	980.5	1041
Pressure3pm	1015.255889	7.037414	977.1	1039.6
Cloud9am	4.447461	2.887159	0	9
Cloud3pm	4.50993	2.720357	0	9
Temp9am	16.990631	6.488753	-7.2	40.2
Temp3pm	21.68339	6.93665	-5.4	46.7

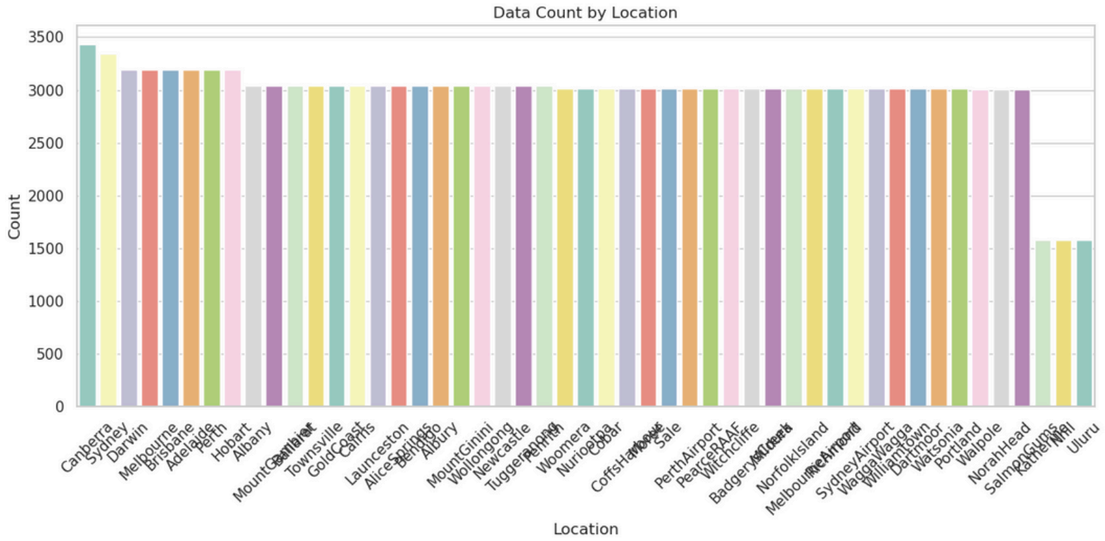
EDA(資料分析)

分佈情況

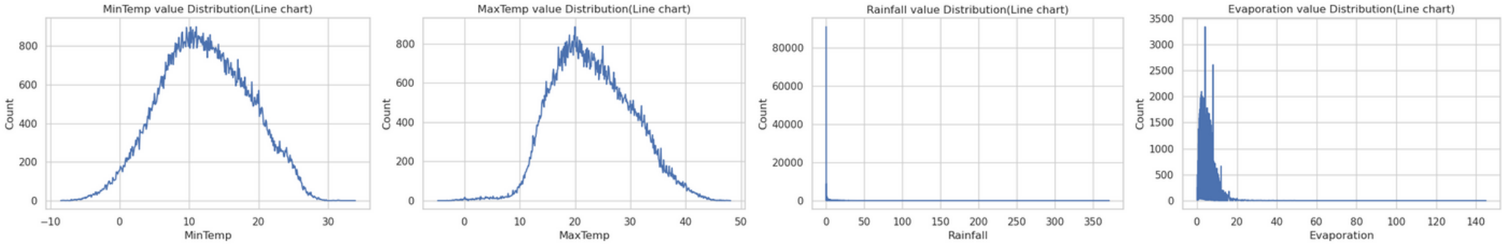
各季節資料筆數分佈

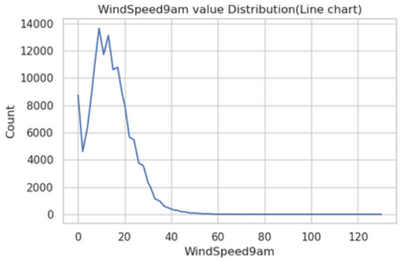


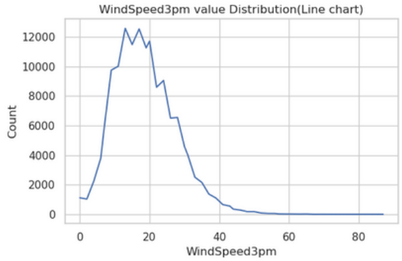
各地區資料筆數分佈

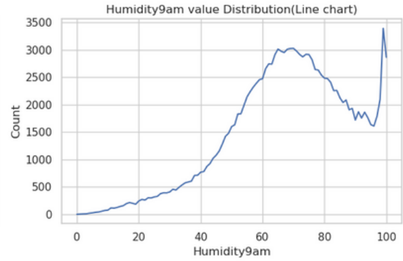


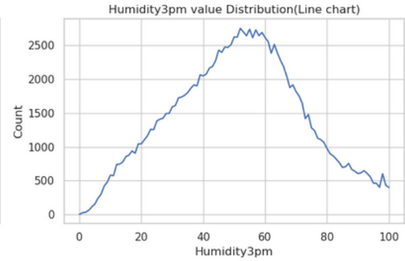
EDA(資料分析) 分佈情況

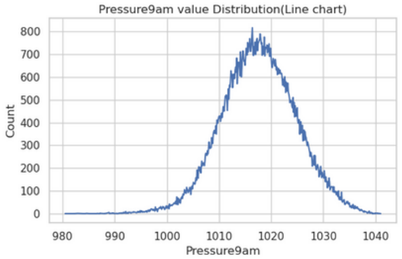


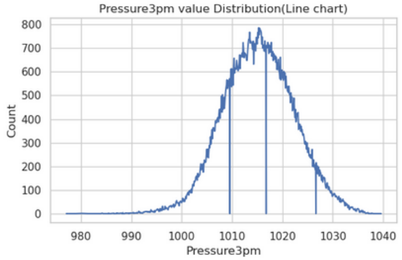


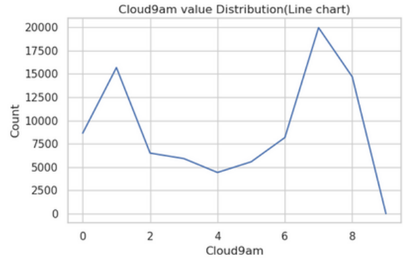


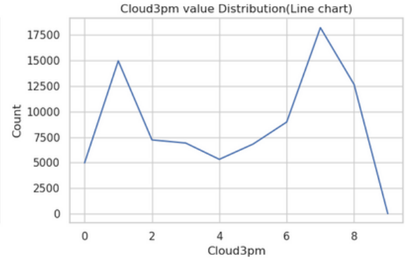


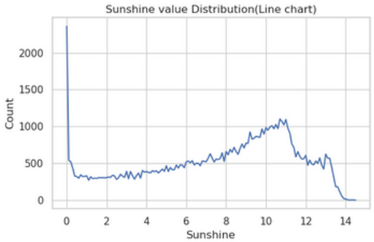


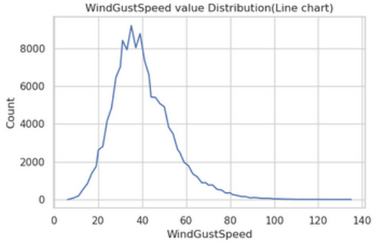


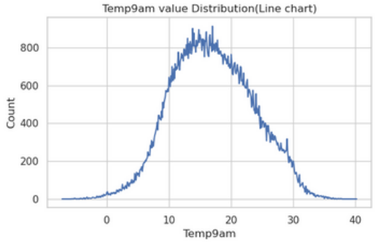


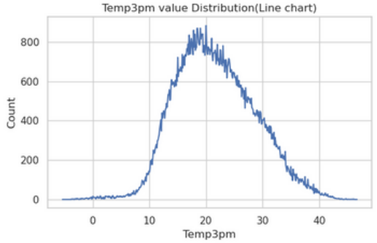






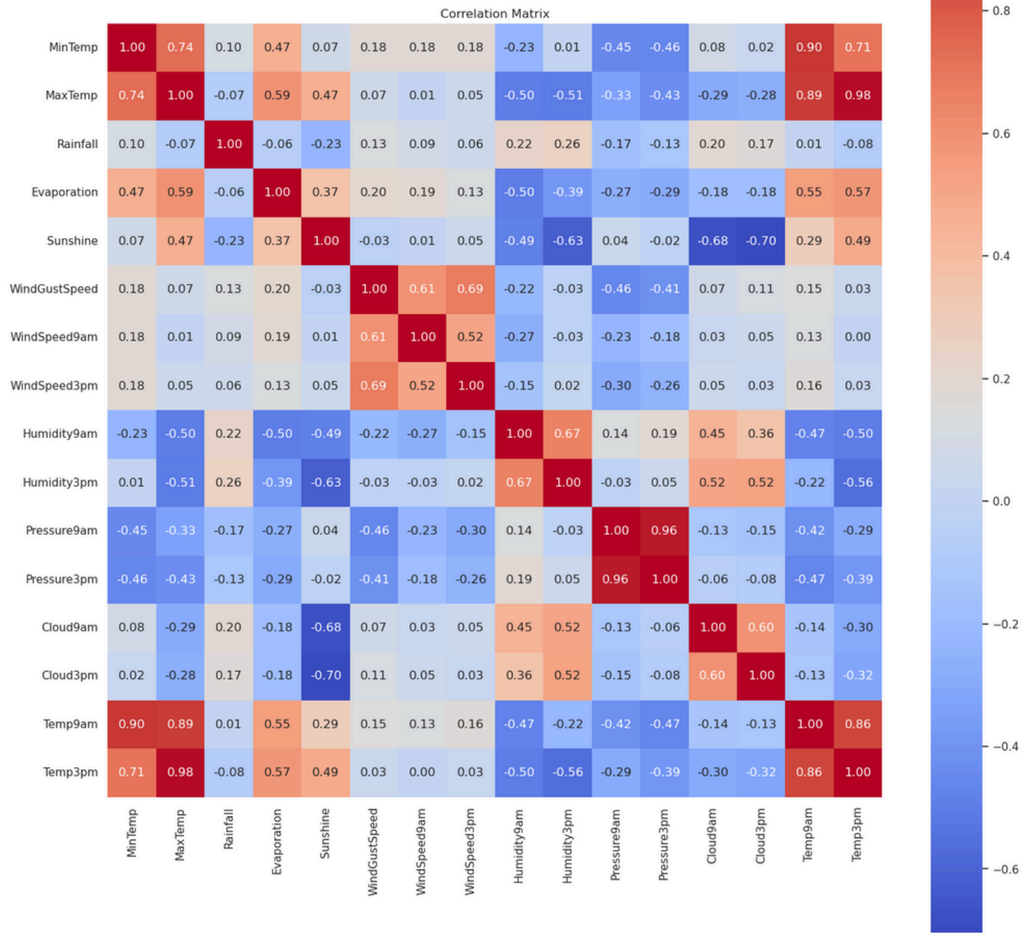




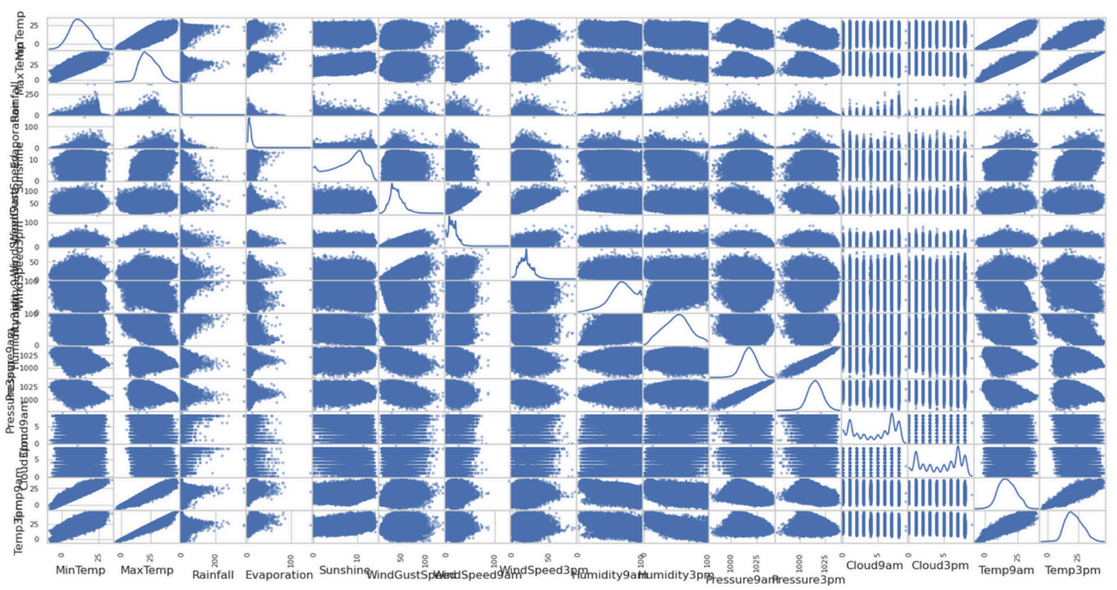


EDA(資料分析)

相關係數矩陣



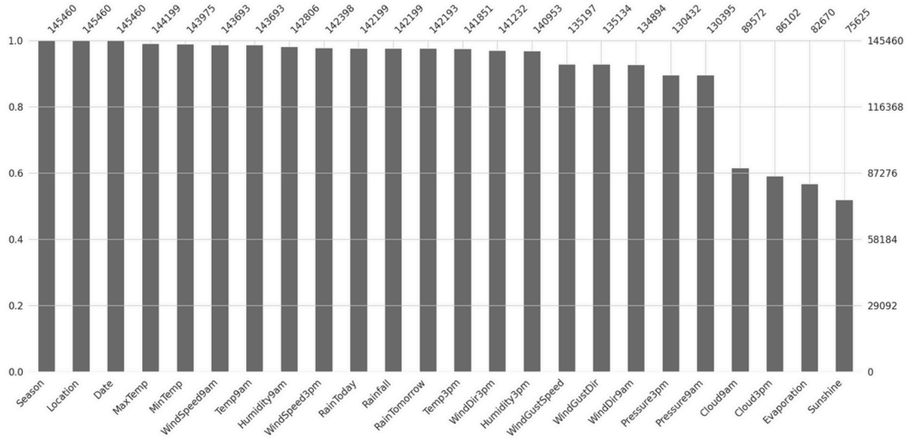
散點矩陣

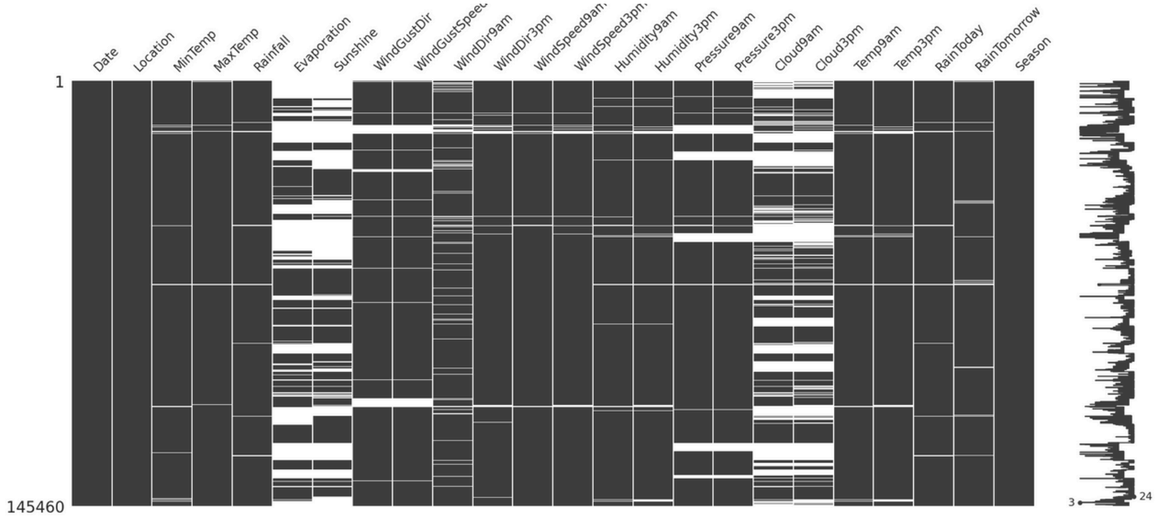


EDA(資料分析)

缺失值

Date	0
Location	0
MinTemp	1485
MaxTemp	1261
Rainfall	3261
Evaporation	62790
Sunshine	69835
WindGustDir	10326
WindGustSpeed	10263
WindDir9am	10566
WindDir3pm	4228
WindSpeed9am	1767
WindSpeed3pm	3062
Humidity9am	2654
Humidity3pm	4507
Pressure9am	15065
Pressure3pm	15028
Cloud9am	55888
Cloud3pm	59358
Temp9am	1767
Temp3pm	3609
RainToday	3261
RainTomorrow	3267
Season	0





EDA(資料分析)

缺失值補法

類型	欄位範例	補法建議
數值型	MinTemp, MaxTemp, Rainfall, Humidity9am, Pressure3pm 等	使用 中位數 或 地區分組平均值 補值
高缺失	Evaporation, Sunshine, Cloud9am, Cloud3pm	若分析重要 → 分地區補平均；否則 → 考慮刪除欄位
類別型	WindGustDir, WindDir9am, WindDir3pm,RainToday	使用 眾數 或 "Unknown" 補值
預測目標	RainTomorrow	若缺失 → 刪除整筆資料（避免影響模型）

參考來源

https://www.kaggle.com/code/shraddhaa/starter-rain-in-australia-e771a48c-1

https://www.kaggle.com/code/siddheshera/rain-in-australia-with-eda-h2o-88-4-auc

資料概述

特徵目標：22欄

- 區分數值與類別的特徵

目標欄位：RainTomorrow(編碼)

- 0=不下雨
- 1=下雨
- 2=缺值或未知

資料前處理

特徵工程

日期拆解

- 將 Date 欄位拆成：
  - Year、Month、Day、DayOfWeek

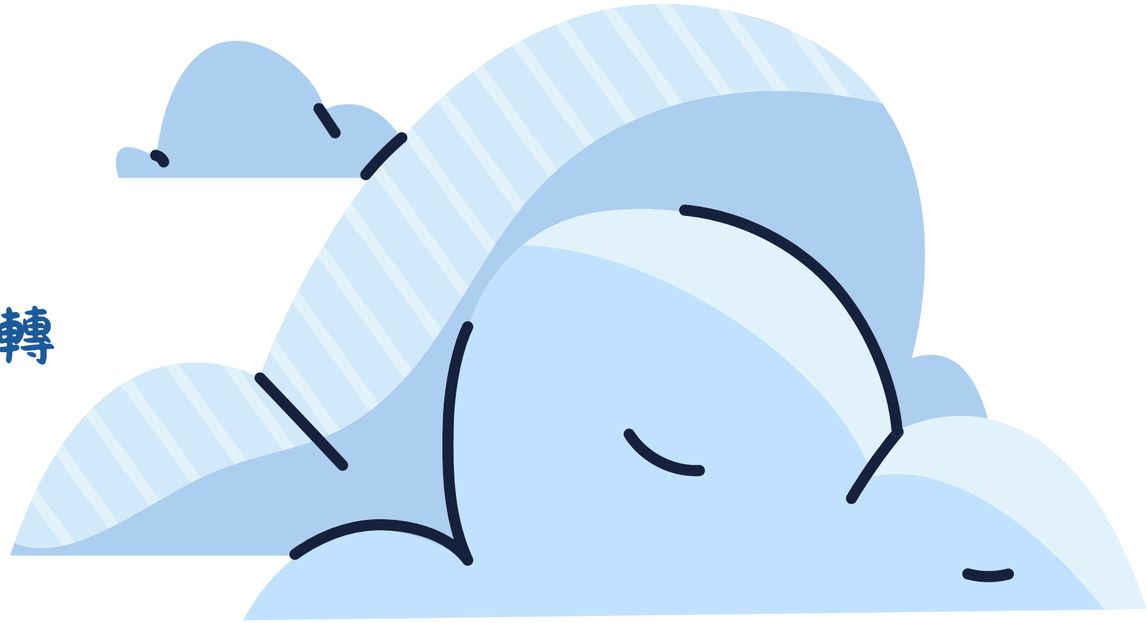
- 有助於模型捕捉季節性變化 (如雨季、冬季等)

資料前處理

特徵工程

類別變數 One-Hot 編碼

- 使用 pd.get_dummies() 轉換所有類別欄位
- 避免模型誤將類別當作數值距離計算



資料前處理

滯後與滾動特徵

滯後特徵（Lag Features）

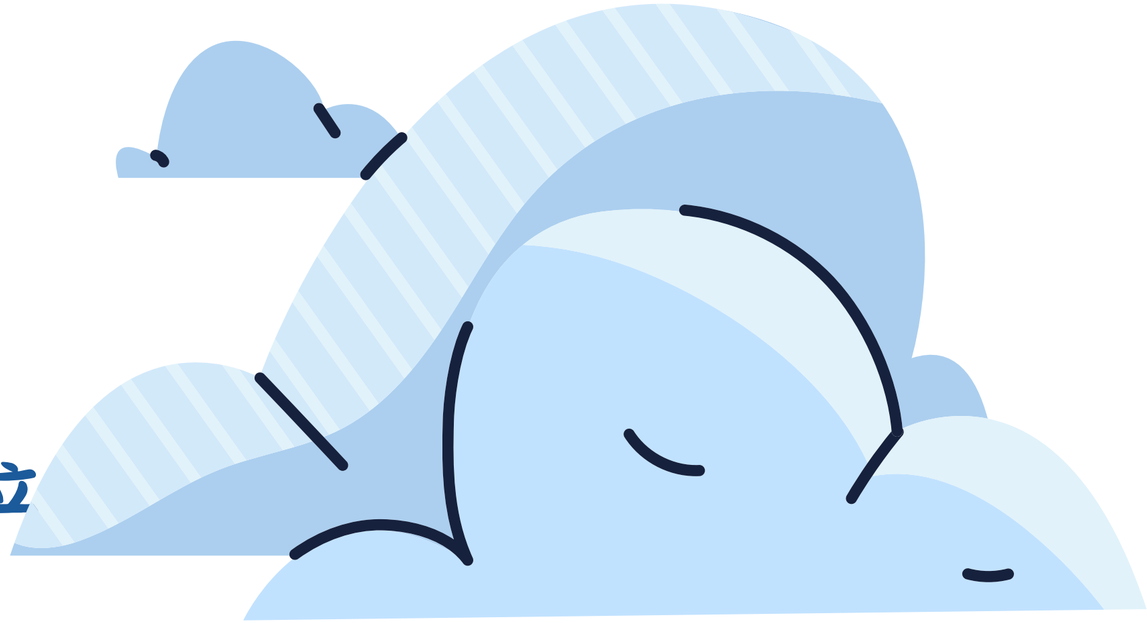
- 為每個Location建立：
  - Rainfall_lag1、MinTemp_lag1、MaxTemp_lag1
- 幫助模型捕捉「昨日天氣」對「明日是否下雨」的影響。

資料前處理

滯後與滾動特徵

滾動平均特徵（Rolling Mean）

- 為主要氣象變數（MinTemp、MaxTemp、Rainfall、Humidity）建立 3 日滾動平均
- 平滑極端值，提升模型穩定性。



資料前處理

異常值處理

- 對數值欄位進行 四分位距截斷法（IQR Capping）

- 將超出範圍的值限制在上下限內，避免模型被極端氣候誤導。

```python
# 定義異常值處理函數（Capping）
def cap_outliers(df, column, factor=3):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_fence = Q1 - factor * IQR
    upper_fence = Q3 + factor * IQR
    # 將超出上下界的值截到上下界
    df[column] = df[column].clip(lower_fence, upper_fence)
    print(f"{column} 已處理：下界={lower_fence}，上界={upper_fence}")
```

資料前處理

資料不平衡+數值標準化

RainTomorrow結果不平衡

- 使用 SMOTE 技術平衡類別，避免模型偏向「不下雨」類別

保證各特徵在模型中權重一致

- 使用 StandardScaler 將數值特徵轉為平均值 0、標準差 1

資料前處理

資料切分

資料集	筆數	備註
訓練集（平衡後）	約 151,920 筆	已 SMOTE 過採樣
測試集	約 36,000 筆	保持原始比例
特徵數量	約 60+	含滯後、滾動與 One-hot 特徵

Data Mining方法

1. 關聯性分析：分析每個分類之間的關聯性

使用：

a. Pearson相關係數：

$$r = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

觀察兩個連續變數之間的線性關係，對極端值敏感→要前處理

數值在-1~1之間，適合快速找出「線性相依」的候選特徵

只能捕捉線性關係，且受資料分配的影響

但計算簡單方便，可以作為初步關聯篩選工具

Data Mining方法

1. 關聯性分析：分析每個分類之間的關聯性

使用：

b. Spearman相關係數：

兩個變數之間的排序關係，會先把資料轉成rank再去計算相關性，對非線性且單調的資料來說可以穩定分析

當資料有離群值或分布不常態時適合使用

c. 熱點圖：觀察相關程度

2. 尋找關聯規則
(例如：高濕度+低溫→明天下雨)

使用：(可擇一)

a. Apriori：操作簡單，可以自己控制support和confidence，組成自己想要的組合數再去篩選，但效率在變數很多時會很差

b. FP-Growth：效率比Apriori好，會建立一顆tree不用一直重複掃描資料集但結構上比較複雜，需要較多記憶體

去尋找氣象條件與是否降雨之間的潛在規則

3. 預測模型

使用：

a. 邏輯迴歸

b. 決策樹

c. SVM

d. 隨機森林

擇一或二評估準確率、混淆矩陣、F1-score等，或是進行模型比較，分析優缺點，去確認預測值(規則)跟實際值之間的準確度

預期成果

1. 找出主要氣象變數之間的關聯性
2. 建立一個有效的天氣預測模型
3. 製作一個簡單互動式詢問天氣的介面

参考来源

https://www.kaggle.com/code/chandrimad31/rainfall-prediction-7-popular-models

https://www.kaggle.com/code/prashant111/extensive-analysis-eda-fe-modelling#13.-Predict-results-

https://stackoverflow.com/questions/19966018/filling-missing-values-by-mean-in-each-group

https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial

https://www.cnblogs.com/zhoumengyao-2103840159/p/16989065.html