**Rain in Australia**

**Data Analytics Technical Report**

General Assembly DAIFXR01 Capstone Project

- Dylan Li

Table of Contents

[1. About Dataset 1](#_Toc140257811)

[1) 2008-2017 Dataset 1](#_Toc140257812)

[2) 2022-2023 Dataset 1](#_Toc140257813)

[2. Data Cleaning 1](#_Toc140257814)

[1) Data Cleaning Tools – Excel & Python 1](#_Toc140257815)

[2) Data Cleaning Brief 1](#_Toc140257816)

[3. Non-Technical Presentation 2](#_Toc140257817)

[1) Presentation Tool – Tableau 2](#_Toc140257818)

[2) Presentation Brief 2](#_Toc140257819)

[4. Technical Analysis – Modelling 4](#_Toc140257820)

[1) Exploration 4](#_Toc140257821)

[2) KNN Modelling 4](#_Toc140257822)

[5. Summary 5](#_Toc140257823)

# About Dataset

## 2008-2017 Dataset

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

This dataset contains about 10 years of daily weather observations from many locations across Australia. The daily observations were drawn from numerous weather stations sourced from Bureau of Meteorology website.

## 2022-2023 Dataset

Bureau of Meteorology -<http://www.bom.gov.au/climate/dwo/index.shtml>

To make comparison with current year weather, monthly observations were downloaded from BOM for 5 major cities in Australia.

Copyright Commonwealth of Australia, Bureau of Meteorology.

# Data Cleaning

## Data Cleaning Tools – Excel & Python

Most data cleaning works were completed in Excel. Cleaning in Python was mainly for the purpose of modelling, which includes change of data type, removal of null value etc. This will be discussed in details in modelling section.

## Data Cleaning Brief

* + 1. Missing and “NA” were checked for all columns of 2008-2017 data set. More than 5000 missing values were found in columns Evaporation, Sunshine, Cloud 9am and Cloud 3pm. These 4 columns were checked separately to see if the missing values were concentrated on the same locations. In the end 20 out of 50 locations were left for analysis as they all had less than 1000 missing values in every column.
    2. The monthly median values for each feature and each location were calculated using array formula, then populated into the cells missing value.
    3. Value “Calm” in Wind Speed columns were change to 0 to match the numeric format.
    4. Missing values in Wind Directions columns were left as blank as they cannot be populated using median values.
    5. Column Rainfall Category was added based on the rainfall amount on Column Rainfall.
    6. Data Dictionary was added referencing to BOM websites:

<http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>

<http://www.bom.gov.au/climate/data-services/content/faqs-elements.html>

* + 1. Similar cleaning process was carried out for 2022-2023 dataset.

# Non-Technical Presentation

## Presentation Tool – Tableau

Multiple sheets and dashboard were built in tableau to form into a 7 pages story discussing climate in Australia.

## Presentation Brief

* + 1. The definition of a rainy day is any day with rainfall total of at least 0.2 mm. 0.2 mm is quite small amount of rain and unlikely to have much impact on many activities. BOM has set a few thresholds as indicators of rains. This analysis follows the same definition.

We can see that in 10 years period, most days are counted as no rain across different locations. There’s only 1304 day counts in total of heavy rains and storm.

Putting the 10 years total rainfall on a map we can see the geographical pattern. Darwin and Cairns on the north have the largest amount of rainfall, followed by Brisbane and Sydney on the eastern coast, then Melbourne, Perth on the south. Alice Springs in the middle had only 26 hundred rainfall during 10 years.

* + 1. Comparing this to the average annual rainfall map provided by BOM we see the same trend. Northern spot has above 2000mm annual rainfall. Cairns has more than 3000mm. Alice Springs has around 200 to 300 mm annual rainfall, which matches our 10 years number.

The main reason for this pattern is the climate zone across the Australia continent. The total rainfall is clearly differentiated by the climate zone boundaries. Darwin and Cairns are in the tropical zone, Brisbane in subtropical, Sydney and Perth in between subtropical and temperate, lastly Melbourne and Portland in temperate zone.

* + 1. Put these locations in a scattered plot and trying to find clusters for them, it’s also quite accurate. There are only 3 features used including average temperature, rainfall amount and humidity.

Wagga Wagga and Melbourne on this chart are almost overlapping, which means they must have a large difference in humidity to be put in different clusters. This is also true in fact because Wagga Wagga is in a semi-arid climate which means it’s close to the dessert.

* + 1. Now we look back to the total rainfall over 10 years and the day counts of different rainfall types.

Alice Springs has the lowest rainfall and the most days without rain. Cairns has the large amount of rainfall. Portland, interestingly has the lowest no rain days, however its rain amount is only in the middle range.

We can conclude that the amount of heavy rains and moderate rains contribute most to rainfall amount. They are mostly in the same trend.

* + 1. There’s also characteristic in the seasonality of rain. We’re more likely to see seasonality where rainfall amount is large, especially in Darwin and Cairns.
    2. See the seasonality in a monthly average point of view, the rainfall in tropical zone follows same trend, when it moves to the south, the largest rain happens with the lowest temperature.

Compare different locations on dashboard with all features. Some key points are: Clouds coverage are somehow towards two extreme ends; Winds are mostly ocean wind;

* + 1. Last but not least we compare the old data to current year. The current year average temperature is lower than 10-year average; Rainfall is higher; Humidity and pressure are both higher.

The reason is the alteration of El Nino and La Nina, which is the warming and cooling of tropical Pacific.

Between 2008 and 2017, there were a few El Nino and La Nina. They eliminated each other’s impact in the average line. We experienced 3-year consecutive La Nina until 2023, which explains the lower temperature and more rainfall. However, BOM has announced that 2022-2023 La Nina has ended. There’ a 70% chance of El Nino forming in late 2023.

# Technical Analysis – Modelling

## Exploration

* + 1. Pandas, matplot and seaborn library are imported.
    2. There are 60926 rows and 22 columns in the 2008-2017 dataset.
    3. Date column is changed from data type “object” to “datetime”.
    4. No outliers are found in the dataset.
    5. Rainfall values are right skewed. Most rainfall are concentrated to lower than 5 mm.
    6. There are some low to midium correlations between features:
       1. Temp3pm – Evaporations: 0.57
       2. Temp3pm – sunshine: 0.37
       3. Humidity3pm – MaxTemp: -0.45
       4. Humidity9am – Sunshine: -0.40

But correlations between Rainfall and other columns are at most 25%

* + 1. Average Temperature, WindSpeed, Humidity and Month columns are created for the purpose of modelling.

## KNN Modelling

* + 1. Sklearn library is imported.
    2. Categorical columns such as Location and Rain Category are converted to dummies.
    3. Null values are checked for selected columns. There are 696 missing values for Humidity.
    4. Humidity missing values are removed.
    5. 2008-2017 data is assigned as training data, while 2022-2023 is test data.
    6. The model accuracy is 0.6405 with k = 10.
    7. The null accuracy is 0.6153 which is lower than model accuracy.
    8. Use k plot to find the best fit k value.
    9. k = 5 was selected using elbow method.
    10. The final model accuracy is 0.6181, which is still higher than null accuracy.

# Summary

This analysis shows that there’s distinguishable climate patterns in Australia based on the locations.

Coastal areas have higher rainfall comparing to the central areas. Among coastal areas, rainfall amount is the highest in the tropical areas and it decreases as the location moves towards south east.