Contents

[**DATA PREPARATION PHASE:** 2](#_Toc60495266)

[**1.** **Introduction** 2](#_Toc60495267)

[**1.1.** **Objective** 2](#_Toc60495268)

[**1.2.** **Datasets** 2](#_Toc60495269)

[**1.2.1.** **Historical Weather Data** 2](#_Toc60495270)

[**1.2.2.** **Historical Wildfires Data** 2](#_Toc60495271)

[**1.2.3.** **Historical Weather Forecast Data** 3](#_Toc60495272)

[**1.2.4.** **Historical Vegetation Data** 3](#_Toc60495273)

[**1.2.5.** **Land Class Data** 3](#_Toc60495274)

[**1.3.** **Approach Introduction** 3](#_Toc60495275)

[**2.** **Cleaning and Preprocessing** 3](#_Toc60495276)

[**2.1.** **Normalizing Data** 3](#_Toc60495277)

[**2.2.** **Handling Missing (NaN) Values** 4](#_Toc60495278)

[**2.3.** **Handling Duplicate Data** 4](#_Toc60495279)

[**2.4.** **Transforming Data** 4](#_Toc60495280)

[**DATA UNDERSTANDING PHASE:** 4](#_Toc60495281)

[**3.** **Exploratory Data Analysis (EDA)** 4](#_Toc60495282)

[**3.1.** **Trends & Seasonality** 5](#_Toc60495283)

[B) Historical Weather – Trends & Seasonality **by Region with Parameters** – Year by Year: 6](#_Toc60495284)

[5. Modeling 9](#_Toc60495285)

[5.1. Regression Models 9](#_Toc60495286)

[5.1.1. Applying algorithm 9](#_Toc60495287)

[5.1.2. Solution to the problem 9](#_Toc60495288)

[5.1.3. Performance of Different Models 9](#_Toc60495289)

[5.2. Classification Models 10](#_Toc60495290)

[6. Conclusions 10](#_Toc60495291)

[7. Future Directions 10](#_Toc60495292)

To Dos’

>> Examine wildfires data further!

>> Understand how Relative Humidity, Soil Water Content, and Wind Speed are related to the other three parameters Precipitation, Solar Radiation and Temperature.

>> How to connect each region’s weather with the fires, and the vegetation?

# **DATA PREPARATION PHASE:**

1. **Introduction**
   1. **Objective**
   2. **Datasets**
      1. **Historical Weather Data**

Data Elements of Weather Data:

* All variables are aggregated to daily values from YYYY-mm-ddT01:00:00Z to YYYY-mm-(dd+1)T00:00:00Z
* **`Precipitation`**
  + is derived from total precipitation.
  + Hourly raw data is converted from m/hour to **mm/hour**
* **`Relative humidity`**
  + is derived from the temperature and dewpoint
* **`Soil Water Content`** 
  + is given for 0 - 7 cm below the surface
* **`Solar Radiation`** or Surface Solar Radiation Downwards.
  + Units are converted from J/h to **MJ/h**
* **`Temperature`**
* **`Wind Speed`** 
  + is calculated for every hour from the Easterly and Northerly 10 meter wind components
    1. **Historical Wildfires Data**

Data Elements of Wildfires Data:

* **`Region`**:
  + the 7 regions - 'NSW', 'NT', 'QL', 'SA', 'TA', 'VI', 'WA'
* **`Date`:**
  + in UTC and provide the data for 24 hours ahead
* **`Estimated\_fire\_area`:** **\*\* This data element can be used to train the model with however, other elements will also be considered.**
  + daily sum of estimated fire area for presumed vegetation fires
  + with a confidence > 75% for a each region
  + **in km^2**
* **`Mean\_estimated\_fire\_brightness`:**
  + daily mean (by flagged fire pixels(=count)) of estimated fire brightness for presumed vegetation fires
  + with a confidence level > 75%
  + in **Kelvin**
* **`Mean\_estimated\_fire\_radiative\_power`:**
  + daily mean of estimated radiative power for presumed vegetation fires
  + with a confidence level > 75% for a given region
  + in **megawatts**
* **`Mean\_confidence`:**
  + daily mean of confidence for presumed vegetation fires
  + with a confidence level > 75%
* **`Std\_confidence`:**
  + standard deviation of estimated fire radiative power
  + in **megawatts**
* **`Var\_confidence`:** 
  + Variance of estimated fire radiative power
  + in megawatts
* **`Count`:** 
  + daily numbers of pixels for presumed vegetation fires
  + with a confidence level of larger than 75% for a given region
* **`Replaced`:**
  + Indicates with an Y whether the data has been replaced with standard quality data when they are available (usually with a 2-3 month lag).
  + **Replaced data has a slightly higher quality** in terms of locations
    1. **Historical Weather Forecast Data**
    2. **Historical Vegetation Data**
    3. **Land Class Data**
  1. **Approach Introduction**

1. **Cleaning and Preprocessing**

***Related Juypter Notebooks*:**

*2.1 Historical\_Wildfires – Cleaning and Preprocessing*

*2.2 Historical\_Weather – Cleaning and Preprocessing*

*2.3 Historical\_Weather Forecasts – Cleaning and Preprocessing*

***Ouput Datasets “C&P\_Datasets”***

*C&P\_Wildfires.csv*

*C&P\_Weather.csv*

*C&P\_Forecasts.csv*

* 1. **Normalizing Data**

1 – Date data type in all three files (Historical Wildfires, Historical Weather, and Historical Wildfires Forecasts) set to YYYY-MM-DD

* 1. **Handling Missing (NaN) Values**

1 – Historical **Wildfires** Data:

2207 Missing (NaN) values for two columns, Std\_confidence and var\_confidence, replaced with the mean value of each respectively

* 1. **Handling Duplicate Data**
  2. **Transforming Data**
  3. **Asfd**

# **DATA UNDERSTANDING PHASE:**

1. **Exploratory Data Analysis (EDA)**

Basic tools of exploratory data analysis are plots, graphs and summary statistics. Below are the possible ways to perform exploratory data analysis:

1 – Plotting distributions of all variables.

2 – Plotting time series of data.

3 – Transforming variables.

4 – Looking at pairwise relationships between variables.

5 – Generating summary statistics for all variables of interest.

6 – Computing mean, minimum, maximum, upper/lower qualifiers.

7 – Identifying outliers.

Exploratory Data Analysis is a **major and significant** portion of the Data Understanding phase to ensure understanding of the datasets and how the variables relate to one another, where correlations may exist, examine patterns that may become obvious through examination of the data, to make comparisons between distributions and summarize the overall events of the data as observed through the exploration of the data. Most importantly the findings of this analysis can and will likely be used to inform (and later improve) the development of the algorithm.

**Types of Exploratory Data Analysis performed in this project:**

* Time Series **Line Plots**:
  1. Time series line plots were used to plot trends / seasonality of weather parameters, month by month, year over year comparisons for each parameter: **precipitation, relative humidity, soil water content, solar radiation, temperature and windspeed.** (Jupyter notebook 3.1 A)
  2. Time series line plots were also used to plot trends/seasonality of weather parameters, month by month, year over comparison **for each region** by each parameter. (Jupyter notebook 3.1 B)
  3. Time series were further used in looking at the wildfires data, in a very similar manner to the weather data, by each data element of fires: mean estimated fire area, mean estimated fire brightness, and mean fire radiative power.
* **Bar Plots**:
  1. **Trends & Seasonality**

***Related Juypter Notebooks*:**

*3.1 A) EDA – Historical Weather – Trends & Seasonality by Weather Parameter – Year by Year*

*3.1 B) EDA – Historical Weather – Trends & Seasonality by Region*

*3.1 C – 1) EDA – Historical Wildfires – Mean Estimated Fire Area – Trends & Seasonality Year by Year and Regions*

*3.1 C – 2) EDA – Historical Wildfires – Mean Estimated Fire Brightness – Trends & Seasonality Year by Year and Regions*

*3.1 C – 3) EDA – Historical Wildfires – Mean Estimated Fire Radiative Power – Trends & Seasonality Year by Year and Regions*

*3.2.1) EDA – Historical Weather – Trends & Seasonality – NSW – 2006 through 2020*

Wildfires occur under a set of specific circumstances and conditions, in a particular period of time during the year. Understanding the seasonality of weather year over year, as well as the trends for different parameters year over year, and then putting both weather seasonality and trends in parameter in combination with regions, can give us an insight into the features we should be associating to better understand how we could predict wilfires.

However, it must be remembered that **correlation does not mean causation**. Therefore, it will require further evaluation and feature engineering to get closer to causation and successful prediction.

* + 1. **Historical Weather Data -Trends & Seasonality**

1. Historical Weather – Trends & Seasonality **by Weather Parameters** – Year by Year:

Min & Max Values of All 6 Weather Parameters, month by month, year by year comparison.

**Background:**

Australian Summer months are **December through January.**

*March – May is Autumn; June – August is Winter; and September – November is Spring*

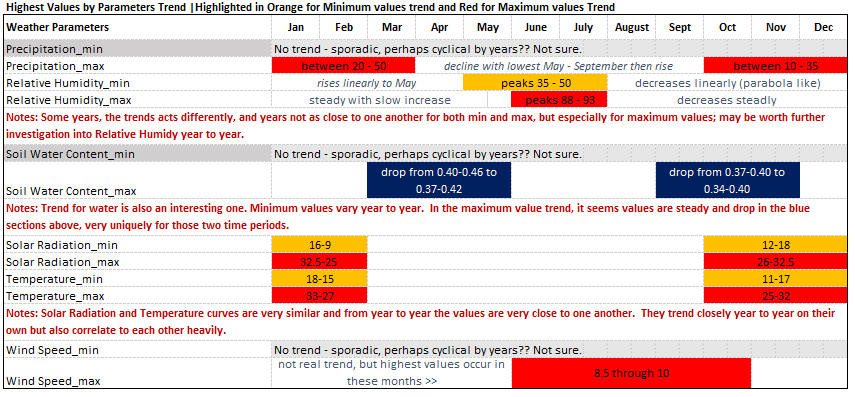
**Objective of Analysis:**

To observe and gain insight on the overall trend, for all 6 parameters, Precipitation, Relative Humidity, Soil Water Content, Solar Radiation, Temperature and Wind Speed, over the period of a year month to month, with year-by-year comparison, by looking at both the minimum and maximum values.

**Summary of Examination:**

Certain parameters have a seasonality, specifically **Precipitation, Solar Radiation and Temperature**. These three parameters are the highest in February through March and October through December, which represent the Summer months in Australia.

However, it’s important to further understand why **Relative Humidity, Soil Water Content and Windspeed** have slightly different distributions.



Note: This figure can also be found in the Tables excel spreadsheet [here](https://github.com/ElenaE873/WildfiresAustralia_CallforCode2020/blob/main/Trends%20%26%20Seasonality/Trends%20%26%20Seasonality%20Tables.xlsx).

1. Historical Weather – Trends & Seasonality **by Region with Parameters** – Year by Year:

Min & Max Values of All 6 Weather Parameters, looking by Region, by month, year by year comparisons

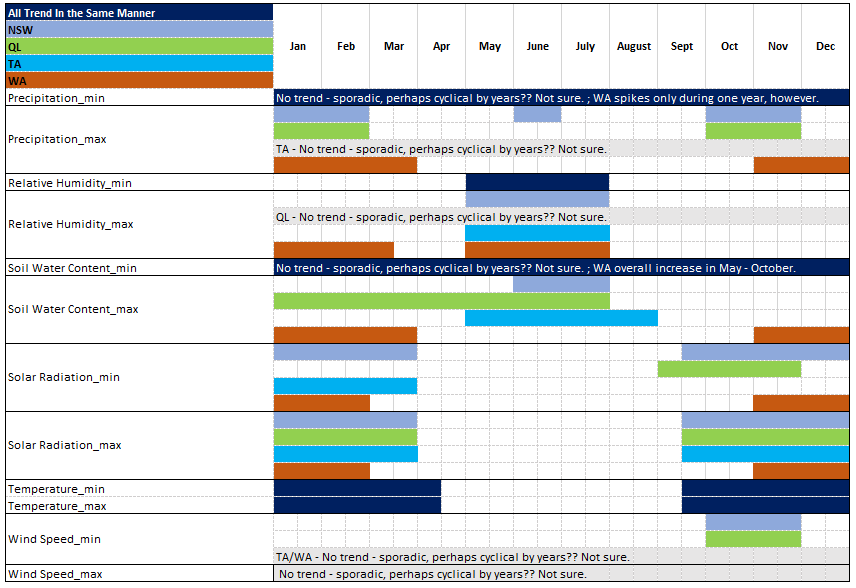
**Objective of Analysis:**

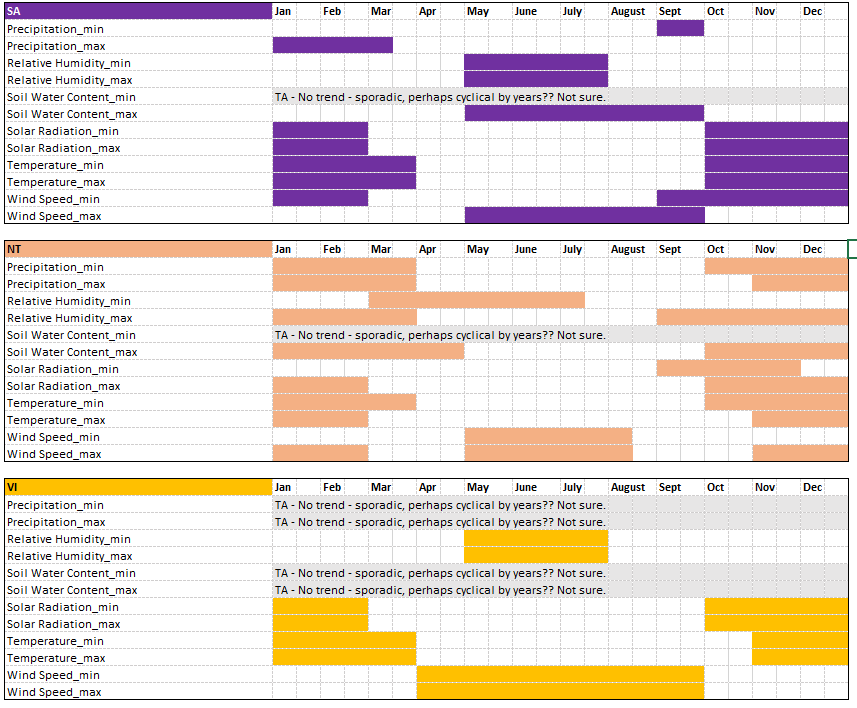
To examine how the trend or seasonality over the 6 parameters, by the 7 separate regions. This meaning that each of the 6 trend examinations are repeated area by area, to determine if weather differs and to what extent, area by area. \*\* With this analysis, the focus will only be on Max values.

Regions: 'NSW', 'NT', 'QL', 'SA', 'TA', 'VI', 'WA'

**Summary of Examination:**

Information on regions is slightly inconclusive, need to examine further. Not sure exactly how to use this regional data with the fires, and vegetation, yet.





1. Historical Wildfires - Trends & Seasonality Year by Year and Regions:

**Objective of Analysis:**

To observe and gain insight and an understanding of wildfires per region. Understanding the historical files year to year, as well as region to region, will aid in associating this data with weather and vegetation index.

**Summary of Examination:**

Findings are slightly inconclusive by looking at this type of time series, even when compared to weather parameter trends. It all just blends together. Need to examine at greater granularity with further questioning.

* 1. Asf

1. Asfd
   1. Calculation of Target Variables
   2. Relationship between Variables
   3. Asdfasdf
2. Modeling
   1. Regression Models
      1. Applying algorithm
      2. Solution to the problem
      3. Performance of Different Models
      4. AsdASD
   2. Classification Models
3. Conclusions
4. Future Directions