Contents

[**DATA PREPARATION PHASE:** 2](#_Toc60500340)

[**1.** **Introduction** 2](#_Toc60500341)

[**1.1.** **Objective** 2](#_Toc60500342)

[**1.2.** **Datasets** 2](#_Toc60500343)

[**1.2.1.** **Historical Weather Data** 2](#_Toc60500344)

[**1.2.2.** **Historical Wildfires Data** 2](#_Toc60500345)

[**1.2.3.** **Historical Weather Forecast Data** 3](#_Toc60500346)

[**1.2.4.** **Historical Vegetation Data** 3](#_Toc60500347)

[**1.2.5.** **Land Class Data** 3](#_Toc60500348)

[**1.3.** **Approach Introduction** 3](#_Toc60500349)

[**2.** **Cleaning and Preprocessing** 3](#_Toc60500350)

[**2.1.** **Normalizing Data** 4](#_Toc60500351)

[**2.2.** **Handling Missing (NaN) Values** 4](#_Toc60500352)

[**2.3.** **Handling Duplicate Data** 4](#_Toc60500353)

[**2.4.** **Transforming Data** 4](#_Toc60500354)

[**DATA UNDERSTANDING PHASE:** 4](#_Toc60500355)

[**3.** **Exploratory Data Analysis (EDA)** 4](#_Toc60500356)

[**PHASE 1 – Explore and gain insight into trends and seasonality of weather and wildfires data by looking at data year by year, and within regions year by year and by specific parameter comparison** 5](#_Toc60500357)

[3.1 Historical Weather – Trends & Seasonality 5](#_Toc60500358)

[3.2 Historical Wildfires – Trends & Seasonality 10](#_Toc60500359)

[5. Modeling 11](#_Toc60500360)

[5.1. Regression Models 11](#_Toc60500361)

[5.1.1. Applying algorithm 11](#_Toc60500362)

[5.1.2. Solution to the problem 11](#_Toc60500363)

[5.1.3. Performance of Different Models 11](#_Toc60500364)

[5.2. Classification Models 11](#_Toc60500365)

[6. Conclusions 11](#_Toc60500366)

[7. Future Directions 11](#_Toc60500367)

# **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**:

## **PHASE 1 – Explore and gain insight into trends and seasonality of weather and wildfires data by looking at data year by year, and within regions year by year and by specific parameter comparison**

***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*

## 3.1 Historical Weather – Trends & Seasonality

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 wildfires.

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 – 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 February.**

*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.

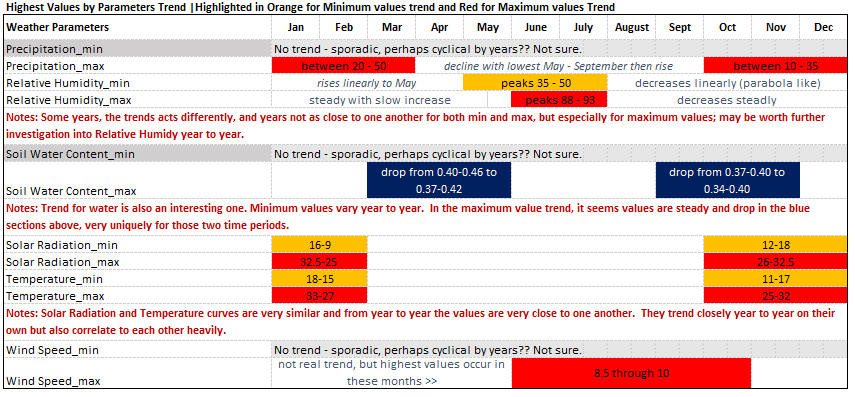
**Summary of Findings:**

**Parameters with Highest values during the Summer Months:**

* **Precipitation\_max, SolarRadiation\_max and Temperature\_max** have the highest values beginning in October and ending in Early March.
* **Soil Water Content** appears to be at its highest form from December through end of February; although slightly higher in June-August, it’s slightly lower than in December through February.

**Parameters with Lowest values during the Summer Months:**

* **RelativeHumidity\_max and WindSpeed\_max** are two of the parameters that appear to be in opposite trend to the four above, with some of the lowest maximum values during these Australian summer months. Additionally, WindSpeed\_min doesn’t have a particular trend. Further investigation may be needed here.



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

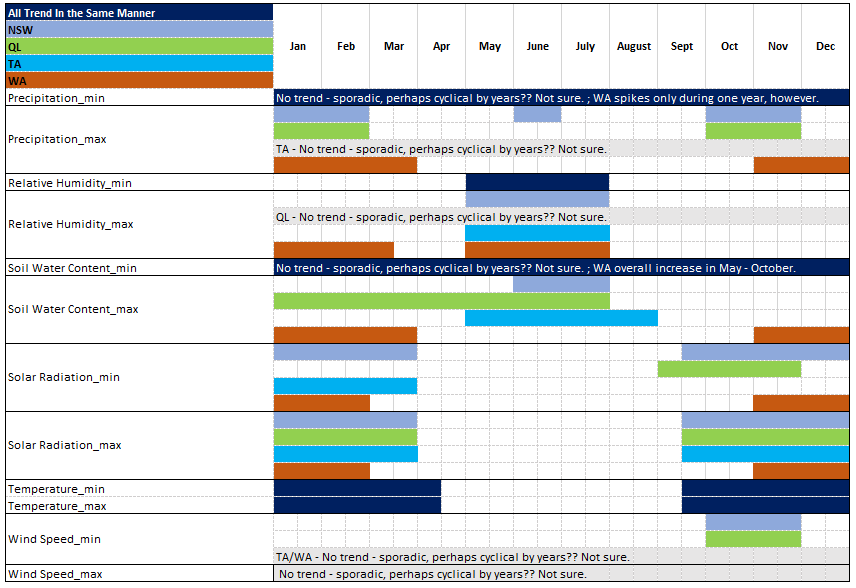
**Objective of Analysis:**

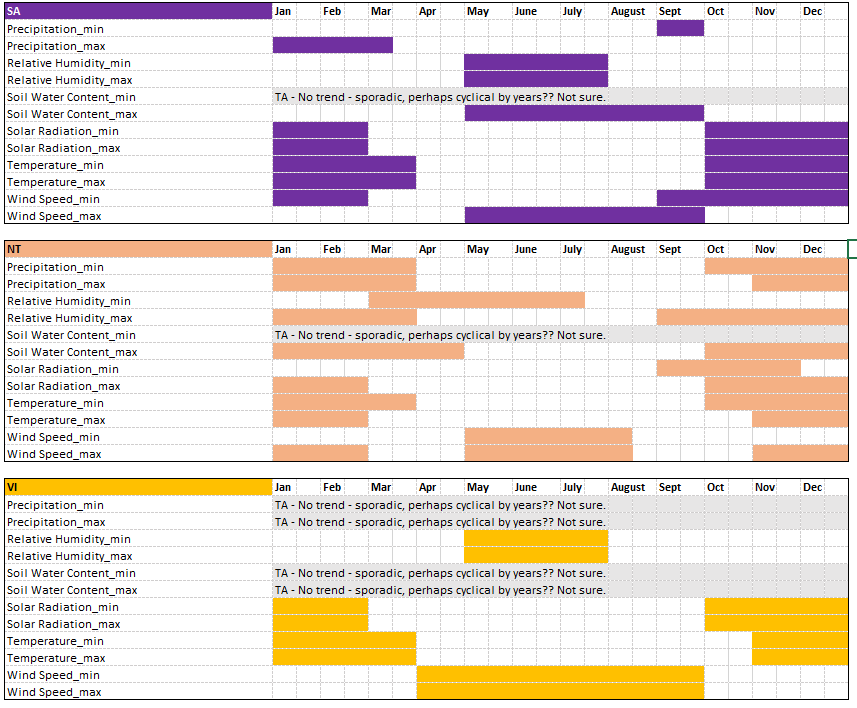
To examine how the trend or seasonality over the 6 parameters, by the 7 separate regions. This meaning that each of the 6 trend examinations are repeated area by area, to determine if weather differs and to what extent, area by area. \*\* With this analysis, the focus will only be on Max values.

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, yet.





## 3.2 Historical Wildfires – Trends & Seasonality

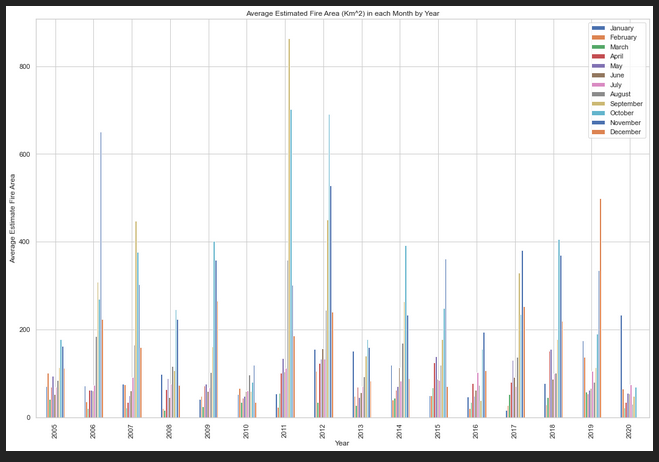
1. Historical Wildfires - Trends & Seasonality Year by Year and Regions:

**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:**

1. Based on the Average Estimated Fire Area in each Month by Year you can see the highest fire areas are during the latter months of the year. This is a very unique trend, that is obviously seen year to year. Perhaps it will be important to look at each year data to see when fire areas, brightness



1. When examining the estimated fire area per region on a yearly basis, we can tell which regions are most affected (in order of highest area to lowest area, consistently year over year, and some observations worth of examining further):

1 – **WA**(cream) **& NT**(orange) [North West] – These two areas are comparable to one another. However, I’d say that NT is generally a more affected area than WA in **most years.**

**Review further:**

* Are weather conditions for these areas similar year to year?
* 2011 and 2012 they seem almost inversed. Review the data for these two year that might inform what happened here.

2 – **QL**(blue) – This area is next in line as far as the next largest estimated fire areas year over year and the areas year over year seems to be consistent, approximately!

* Geographically, this makes sense that it is next on the list, because it is located right next to WA/NT and further north from SA

3 – **NSW** (teal)  **and VI** (yellow)– These are the next two in line with NSW consistently having larger estimated fire areas than VI, but in some years remarkably close to one another.

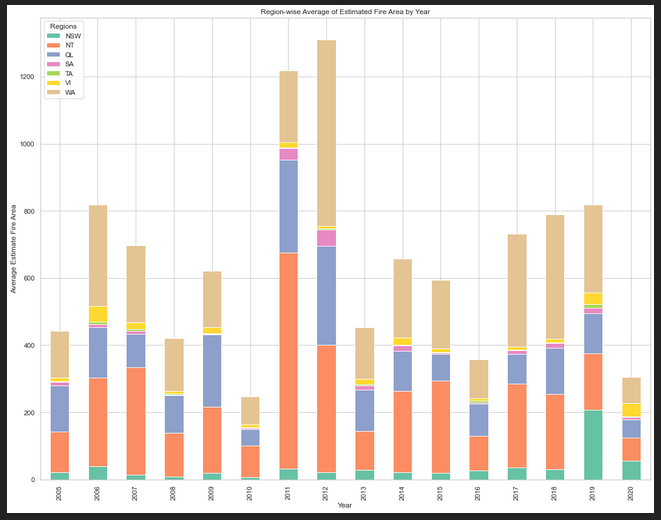
* Geographical placement of these two regions further explains why they land very much in the middle of this list, because it is right below QL but rather more east.
* **Review further:** The vegetation of these two regions probably contribute to them having more fires than maybe you’d expect since they are so far south, closer to TA. Worthy of investigation.

**Review further:**

* **NSW:** one thing worthy of note is that in the year 2019, the area for this region spiked more than before, but seems to have normalized in 2020. It would be wise to explore this region in the year 2019 to see which characteristics might have contributed possibly to the spike.
  + Also, might be good to observe when it started and how long the fire lasted based on the distribution of the fire area, brightness and radiative power of that year.

4 –**SA** (magenta)**and TA** (lime green) – These regions are least affected, that’s for sure. However, **SA** having more fire area than TA consistently year over year.

* What’s really interesting about SA is that it’s right below WA/NT (which are the top highest with fire areas), however, SA being so south, it seems to be protected.
* Geographically SA is the further south and that seems to have even more protection from fires.





* 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