Weather and Wildfire Descriptive Analysis

Southern Australia (2009 – 2017)

MIS 633, Winter 2020

GROUP 1: "Wildfire"

Russell Destremps, Saravanan Jayakumar, Padma Priya Jayaraj, Saniya Khan, Soumya Sinha

Executive Summary: Trends in both general weather conditions and incidents of wildfire were analyzed in southern Australia over a time period spanning 2009 to 2017. Through the collection of various data sets and appropriate cleaning and merging, a comprehensive data set of observations was constructed over this time period. Based on extensive exploratory analysis and regression, trends in weather, such as heightened fire risks due do seasonality of temperature and rainfall factors was clearly demonstrated. All data preprocessing and analysis was focused within the SAS Studio University Edition software.

Outline

Team Name: Wildfire

PROBLEM STATEMENT

Topic. Wildfire and Weather Analysis in Southern Australia.

Problem Statement. This study aims to address several research questions focused on identifying trends in weather and incidents of wildfire in Southern Australia over the time period of 2009 to 2017. Further, quantitative examination of the relationship between these two observations – weather and incidents of wildfire – will follow to determine any correlation between various weather factors and fire occurrence.

Relevant Research Questions.

- What are the overall trends of incidents of wildfire in Southern Australia – frequency, time of year, type of incident etc?
- What is the overall trend of basic weather data over the period of analysis in Southern Australia?
- What weather factors are significant indicators of incidents of wildfire in Southern Australia?

Background. The research team has identified this as a topic of interest based on the recent severity of wildfire that effected much of Southern Australia. Further, there has been increased media coverage on the effects of climate change, to include proposing links between climate change and occurrences of natural disaster that may be linked to weather. In particular, "numerous reports, ranging from popular media through to peer-reviewed scientific literature, have led to a common perception that fires have increased or worsened in recent years around the world."[i] [iiiii] Therefore, the research focus will be geographically consistent with the recent events in Southern Australia and examine any linkage with weather data in that area over the same period. The research will not focus on creating any prediction models, but instead on descriptive analytics on the data sets that follow.

Team Name: Wildfire

METHODOLOGY

I. Collect Data

a. Wildfire Data - South Australian Government Data Directory: South Australian Country Fire Service Brigade Incidents.[iv]

i. **Data Source***.* The problem statement is primarily centered on fire data and therefore initial data research focused in this area. The South Australian Government Data Directory contains open datasets for all government organizations like Department of Health, Environment & water, fire service etc. These datasets are maintained by Government agencies which regularly update the publicly available database. In particular, The South Australian County Fire Service Brigade Incidents database contains location-based record keeping of categorically indexed emergency responses. The locations are identified following the organizational structure of the government fire department Regions and Brigades, which then have names corresponding to the specific towns/areas they serve.

ii. **Data Collection***.* The datasets from SA Country Fire Service (CFS) department are readily available for direct CSV file download and provide information on incidents attended by between (2009 – 2017), including the incident date, the type of incident and the primary attending CFS brigade. The time period chosen maximizes availability of complete data and is assessed as satisfactory to produce sufficient cases of fire incidents, enabling an improved analysis when comparing with weather data.

b. Geo Data - World Geocoding Service.[v]

i. **Data Source.** Next, the city/town location variable from the South Australian CFS Brigade Incidents dataset was utilized to research latitude and longitude location. This unit of location was required to readily research next steps of compiling weather data. The collection method for matching location and latitude/longitude readings follows, however, the ArcGIS database from the World Geocoding Service was utilized for this purpose.

ii. **Data Collection.** The Fire Incident dataset contained over 400 unique locations. Therefore, a programming script was written in order to streamline this step utilizing the Python programming language. In particular, the Geopy package in Python allows abstraction via an application programming interface (API) access of geo data including coordinates, addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

c. Weather Data - Australia Bureau of Meteorology.[vi]

i. **Data Source***.* Finally, with locations for all fire incidents identified in latitude/longitude format, a third dataset could now be collected for weather data over the same period. The Australian government repository for weather data is expansive in both location, weather variables and date ranges back to the late 1800's. The robustness and general reliability of this data source were the leading factors for its selection. The data consists of 28 weather related variables, including mostly continuous factors and several identification variables. When collected across all fire incident locations over the period of interest, the number of daily observations results in over 900k observations.

ii. **Date Collection**. Due to the expansiveness of the data set, a more technical approach was taken in order to extract the data from the Australian Bureau of Meteorology. Therefore, an API-based program was created utilizing Python. The API enabled the data to be extracted by location for the date range focus of the study. In order to accomplish this, the location names originally found in the wildfire were converted to latitude/longitude to interface with the weather database.

II. **Clean and Collate Data**. This step starts when all data to support the Problem Statement and Research Questions has been collected and will end when data sets are cleaned and sorted for ready use in subsequent phases and are combined as required to facilitate analysis of data that has been collated by both location and date.

a. **Clean.** *Format, remove unwanted data, etc*.

i. **Wildfire Data.** Two sets of data were extracted from the source:

- Dataset 1: Fire incident recorded by County Fire Service (CFS) between 2014 – 2017 with incident date, the type of incident, primary attending CFS brigade and Region categorized based on Brigade Group.
- Dataset 2: Fire incident recorded by CFS between 2009 – 2013 with incident date, the type of incident and primary attending CFS brigade (Without region Information).

In order to find the Region information for Dataset 2, the VLOOKUP function is used to extract the Region values corresponding to Brigade from Dataset 1. Both the Datasets are imported in SAS to further format, clean and remove unwanted data.

```
%web_drop_table(WORK.Firedata_20092013);


FILENAME REFFILE '/folders/myfolders/sasuser.v94/Project/Firedata_20092013.xlsx';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLSX
    OUT=WORK.Firedata_20092013;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.Firedata_20092013; RUN;


%web_open_table(WORK.Firedata_20092013);
```

```
%web_drop_table(WORK.Firedata_20142017);


FILENAME REFFILE '/folders/myfolders/sasuser.v94/Project/Firedata_20142017.xlsx';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLSX
    OUT=WORK.Firedata_20142017;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.Firedata_20142017; RUN;


%web_open_table(WORK.Firedata_20142017);
```

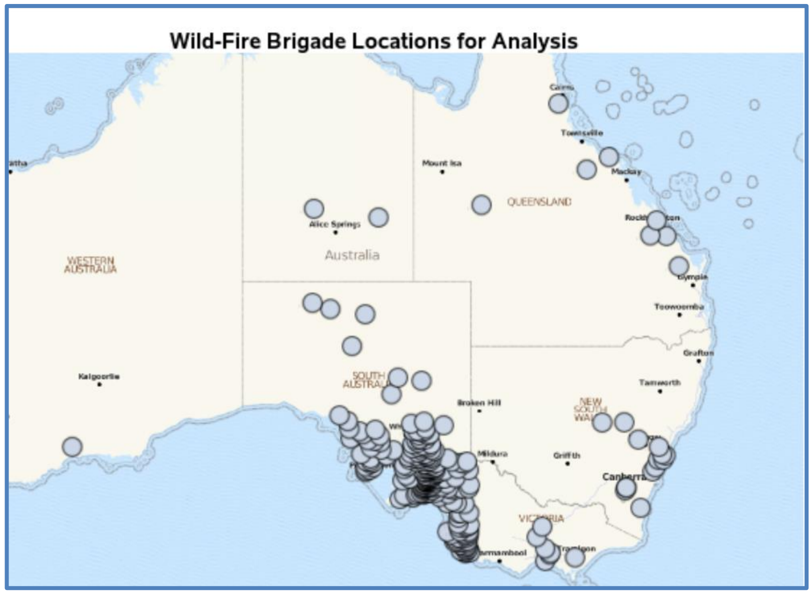
Dataset 1 & 2 are concatenated into single dataset called “Fire_incident” in FIRE library (year between 2009 – 2017) for further Analysis.

```
data WORK.combined; set WORK.firedata_20092013 work.firedata_20142017;
run;
```

Now the “Fire_incident” dataset has 53,358 observations which includes all the incidents attended by CFS Brigade such as vehicle accident, animal rescue etc. However, the focus of interest is to examine the relationship between weather dataset and wildfire specific incidents. So, the unrelated incidents were removed which are not associated with weather observations, reducing the observations to 2734. These final observations included unique Brigade locations were used to extract only the required latitude and longitude information.

```
data Fire.Fire_incident; set Fire.Fire_incident;
    Where TypeOfIncident in ("TREE_FIRE","SCRUB_AND_GRASS_FIRE","FOREST_FIRE");
run;
```

ii. **Geodata**. The geolocations were found using a simple package tool (geopy) to extract latitude and longitude using the Brigade /Region combinations for the Southern Australian locations. The APIs used have the information without any junk values. For some geocodes, the API program timed out or did not produce accurate results. For these instances, latitude & longitude for those regions/locations were collected manually.[vii]



The following map represents the locations that were identified as a result of the brigade lat/lon extraction. Approximately ten of the city locations were found to be outside of the intended study region of Southern Australia. The reason for this faulty result is most likely due to city names that may have multiple locations within Australia. In these cases, the city/town with multiple matching locations within the country of Australia was improperly selected. In order to prevent a mismatch in fire and weather data when these data sets are later merged, these locations will be manually removed from the analysis. This will slightly reduce the overall cities and therefore fire incidents in the data set, however this tradeoff is acceptable to improve quality of results.

iii. **Weather Data.** The API extraction of the weather dataset described earlier resulted in data with superfluous information in the "header" and "index" parts. The data also had the data definition dictionary prepended along with actual weather data corresponding to a specific latitude/longitude.

Cleaning steps involved extracting weather info for each latitude/longitude combination for a specific time period and then concatenating weather data per location to a single file. Furthermore, the response was in the form of space delimited data, which was converted to CSV format for ease of use.

b. Collate. With the data from each set cleaned and sorted, data from the various sources are combined into a single data frame in order to enable subsequent exploration and analysis. That is, the focus of the problem statement is to examine correlation between weather data and incidents of fire, therefore weather by-location and fire incidents by-location data must be combined in order to begin analysis.

Weather data has some unnecessary variables when extracted using API. Those variables were Dropped and there are some variables renamed as per our ease of use.

```
/* Dropping variables */
data Fire.Weather; set Fire.Weather;
    drop A Date__yyyymmdd_ Day Smx Smn Srn Sev Ssl Svp ;
run;

/*Renaming the variables name */
data Fire.Weather; set Fire.Weather;
    rename VAR4=Temp_Max VAR6=Temp_Min VAR12=Radn VAR16=RH_MaxT VAR17=RH_MinT Date2_ddmmyyyy_=Date;
run;
```

```
/* Changing the Date format as per weather data*/
data Fire.Fire_incident; set Fire.Fire_incident;
    format Date ddmmyy10.;
    informat Date ddmmyy10.;
run;
```

In Data Fire_incident, Date was in the format of "MMDDYYYY" and it is changed to date format as Weather data to perform merging operation in next step.

WEATHER

Date
Brigade
Latitude
Longitude
Temp_Max
Temp_Min
Rain_mm
EVAP_mm
Radn
VP_Hpa
RH_Max
RH_Min

Merging Condition **LEFT JOIN**

FIRE_INCIDENT

Date
Region
Brigade
Type of Incident

Both datasets are merged with LEFT JOIN with the condition "Brigade" and "date" variable. Now the resultant combined dataset will have all the weather data sorted in date wise (from 2009 to 2017) for all the brigade Location with certain locations have fire incidents recorded. To develop a prediction model, a new Binary variable "Fire" is created with value 1 if there is a fire incident and value 0 if there is no fire incident.

```
PROC SQL;
CREATE TABLE Fire.combined
AS
SELECT WEATHER.Date, WEATHER.Brigade, WEATHER.Latitude, WEATHER.Longitude, WEATHER.Temp_Max, WEATHER.Temp_Min,
WEATHER.Rain__mm_, WEATHER.Evap_mm_, WEATHER.Radn, WEATHER.VP__hPA_, WEATHER.RH_MaxT, WEATHER.RH_MinT, FIRE_in
FROM FIRE.WEATHER WEATHER
LEFT JOIN FIRE.Fire_incident Fire_incident
ON
    (
      ( WEATHER.Brigade = FIRE_incident.Brigade ) AND
      ( WEATHER.Date = FIRE_incident.Date )
    ) ;
QUIT;

/* Creating a new Binary variable "Fire" for every observations with and without Fire incidents */
data Fire.combined; set Fire.combined;
    format Fire 8.;
    if TypeOfIncident = " " then Fire = 0;
    else Fire = 1;
run;
```

III. **Data Exploration.** This step will start with complete and clean data sets and end with a refined focus on the Problem Statement and initial findings that will be further examined in detail during the Data Analysis step.

a. Wildfire Data.

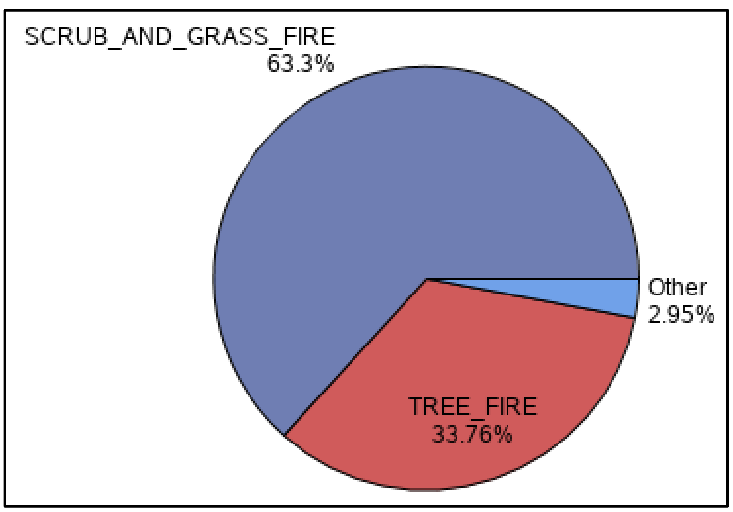
The fire dataset has 53360 rows and 5 columns as SubID, Date, Brigade, Region and TypeOfIncident. This dataset basically helps us understand which region experienced the maximum fire incidents over a period of time. Also, we can look into the specific brigades under those regions which are having the maximum number of fir incidents.

The data for the entire country of Australia is rather large, with Region 1 and Region 2 almost contributing to 80% for the country as a whole. Therefore, to gain the most insights, while reducing the amount of corresponding weather data required to pair with these locations, Region 1 and 2 only were selected. This results in 9 years of data from 2009(May till December) to 2017(Jan till June) for fire and non-fire incidents.

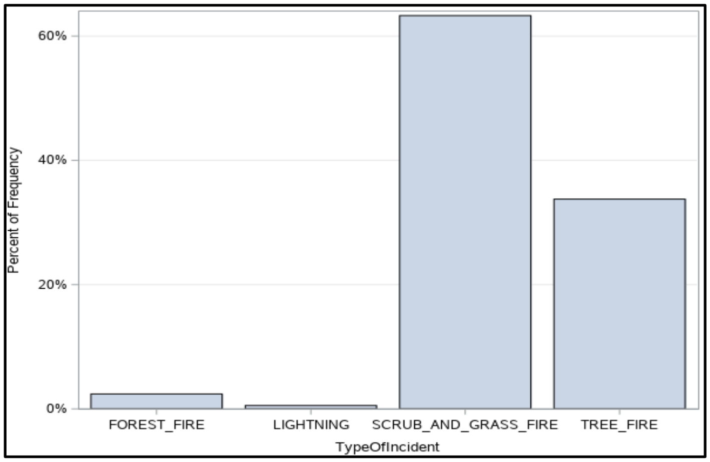
Fire incidents is originally brought in as a categorical variable with many levels, such as regular house fires or building fires and medical emergencies. In order to focus on values that pertained specifically to wildfires, the three major types values were filtered, and associated observations retained:

1. Scrub and Grass fire
2. Tree fire
3. Forest fire

Below are a few graphs which helps us interpret the data better for fire incidents.



Pie Chart showing the distribution of fire incidents in Southern Australia. **Scrub and grass fire contribute the most i.e. 63% followed by tree fire which contributes to 33%.** Forest fire contributes the least i.e. 2.41%.



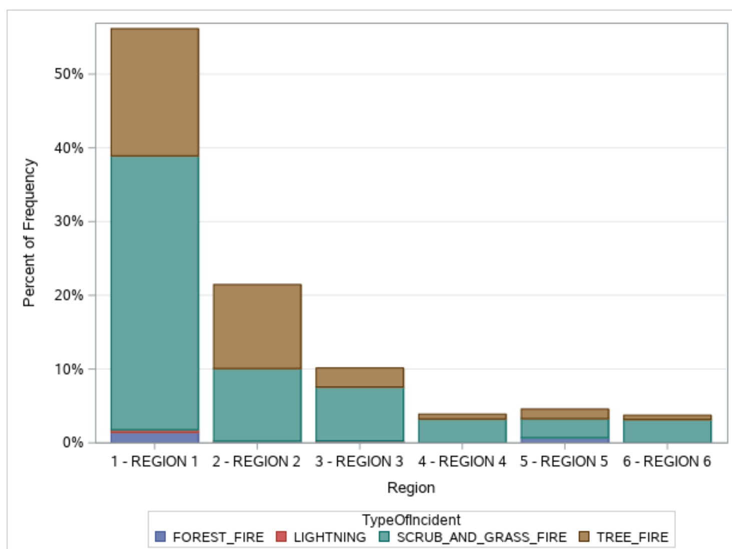
In order to see the contribution of other category in causing fire we plot a bar graph. It is clearly observed that forest fire does not contribute much in causing fire.

Region wise fire distribution:

As stated above, observations for Region 1 and Region 2 are very high. From the stacked bar chart below, we can observe the same.

Another important thing to note is that **scrub and grass fire is the most common type of fire incident causing fire in Region 1** of Southern Australia which includes a few brigade groups like:

- Brigades in the East Torrens Group
- Brigades in the Heysen group
- Brigades in the Kangaroo Island Group
- Kyeema Group
- Mawson Group
- Brigades in Mt Lofty CFS Group
- Mundoo Group
- Onkaparinga Group
- Southern Fleurieu Group
- Strathalbyn Group
- Sturt Group
- Victor Harbor Group

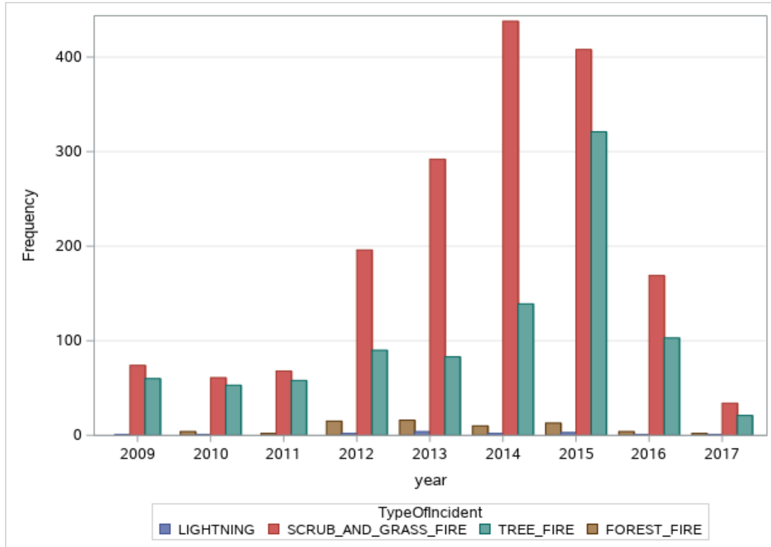


Region 2 has a greater number of tree fire incidents and includes few brigade groups like:

- Northern Barossa Group
- Barossa Group
- Gilbert Group
- Gumeracha Group
- Horrocks Group
- Light Group
- Northern Yorke Peninsula Group
- Para Group
- Southern Yorke Group
- Wakefield Plains Group
- Yorke Valley Group

Frequency of fire incidents over time:

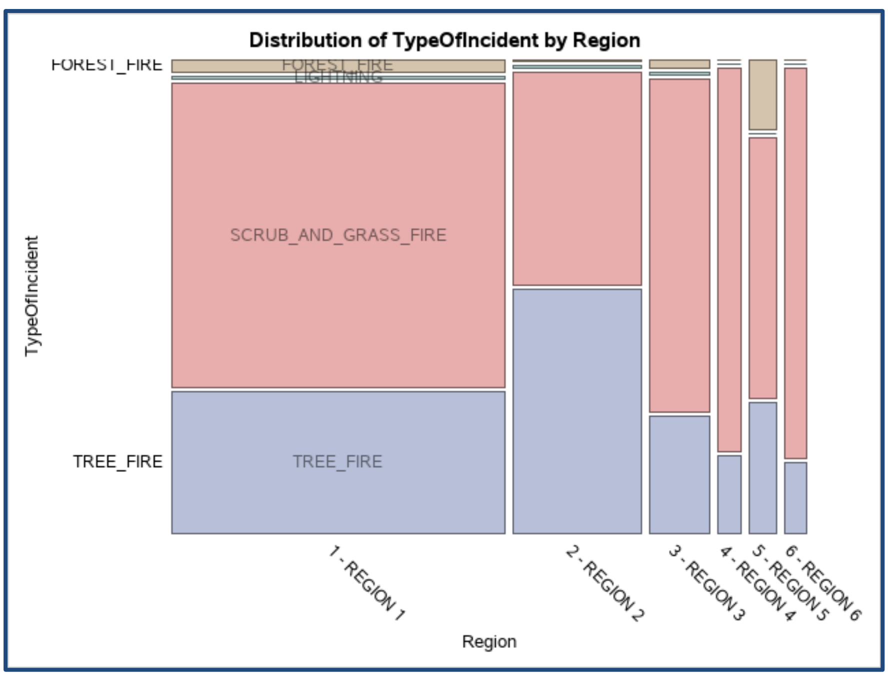
Below are the important observations from the time series plot for fire incidents over time:



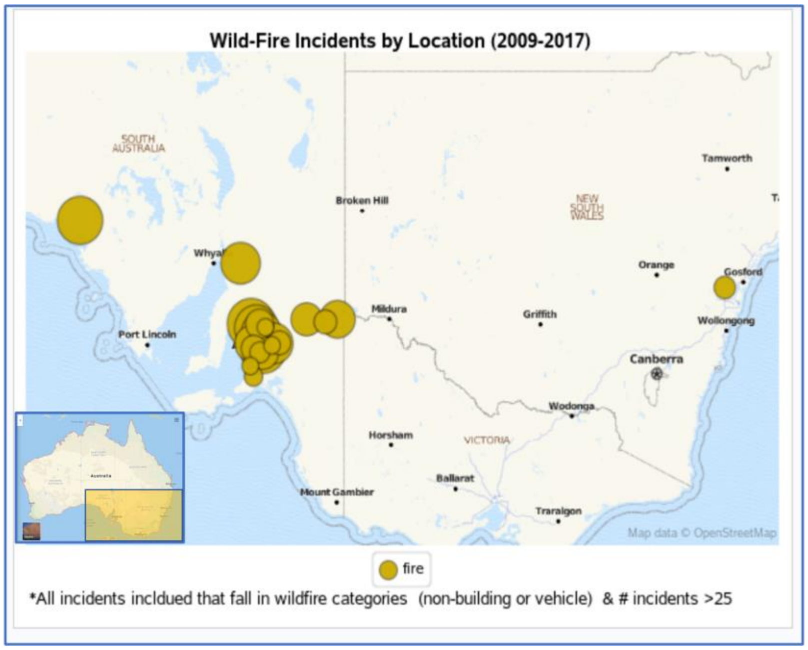
1. Overall, 2014 had the greatest number of fire incidents over time followed by 2015.
2. Fire incidents increased from 2012 onwards and was recorded as the highest in 2014, followed by significant drop in 2016(almost 50%).
3. Scrub and grass fire were the most common cause for fire incident across the 9 years and tree fire significantly became an important factor in the year 2015.
4. Very few fire incidents are recorded for 2009 and 2017. This is partly because we do not have the complete data for all the months for these 2 years.

Mosaic Plot:

Mosaic Plot is used to depict relationship between two categorical variables (in this case, Type of Fire by Region). Area of the box denotes the number of observations. For Region 1 and Region 2 the rectangular box is comparatively bigger that Region 3, 4 and 5.



Next, high frequency of wildfire incidents were identified through overlaying frequencies with geographic data included in the set and filtering for frequency. The following bubble map provides insights into where these high frequency areas are geographically focused.



Going forward with future phases of analysis, all four wildfire categories will be identified as only as a binary fire/non-fire event for that date/location. This binary classification will be used as the dependent variable, with the next section, a detailed exploratory analysis on weather, comprising the independent variables to be used for analysis in determining conditions that may contribute to incidents of wildfire.

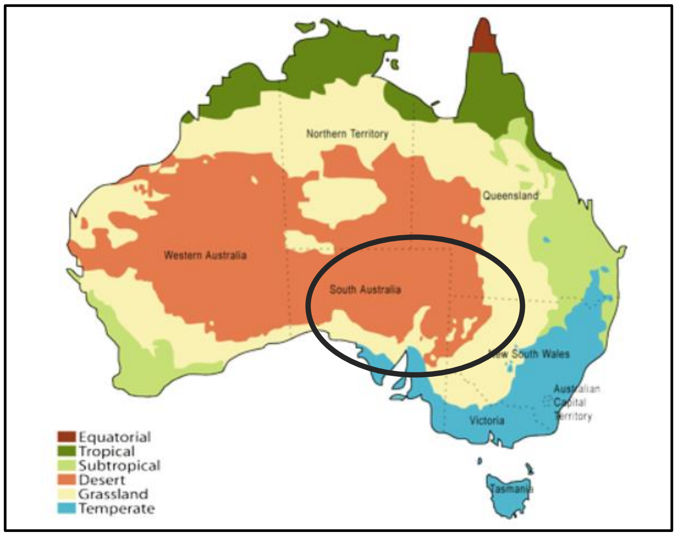
b. Weather Data.

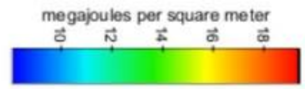
Summary Statistics for the variables in Weather data shows following results

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Day	877362	183.05600	105.39893	160606374	1.00000	366.00000	Day
TMax	877362	22.46831	7.29482	19712840	2.00000	48.00000	T.Max(oC)
Tmin	877362	10.49974	5.22180	9212071	-7.50000	34.00000	T.Min(oC)
Rain__mm_	877362	1.52424	5.01871	1337313	0	527.40000	Rain (mm)
Evap_mm_	877362	4.83292	3.15166	4240216	0.20000	33.80000	Evap(mm)
Radiation	877362	16.93588	7.77303	14858898	2.00000	35.00000	Radn(MJ/m2)
VP__hPA_	877362	12.05151	3.50840	10573537	1.00000	38.00000	VP (hPA)
RHmaxT	877362	46.87310	17.10751	41124675	1.30000	100.00000	RHmaxT(%)
RHminT	877362	88.56385	16.18795	77702555	4.10000	100.00000	RHminT(%)
Date	877362	19541	948.58699	1.71441E10	17898	21183	Date2(mmddyyyy)
Latitude	877362	-34.33844	2.97077	-30127242	-38.37035	-17.35250	Latitude
Longitude	877362	139.60960	3.70982	122488158	120.07625	151.90434	Longitude

The mean values for Tmax and Tmin variables suggest the average values are roughly in the range [10,22] (in degree Celsius) indicating the temperatures are moderate which suggests that high temperature is not the only reason that can be attributed to high incidence of forest fires. However, average precipitation (Rain) is approximately 1.5 mm, suggesting the region is extremely dry[1] which can be seen from the descriptive map below (South Australian region classified as "Desert"):



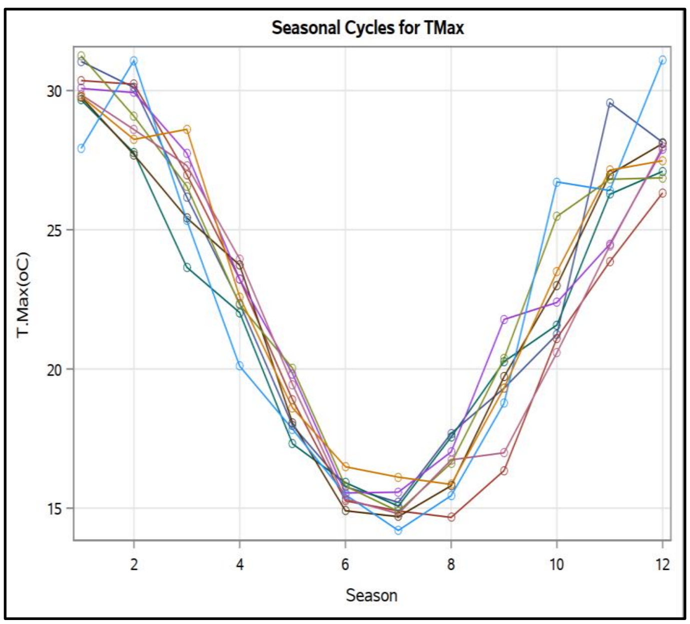


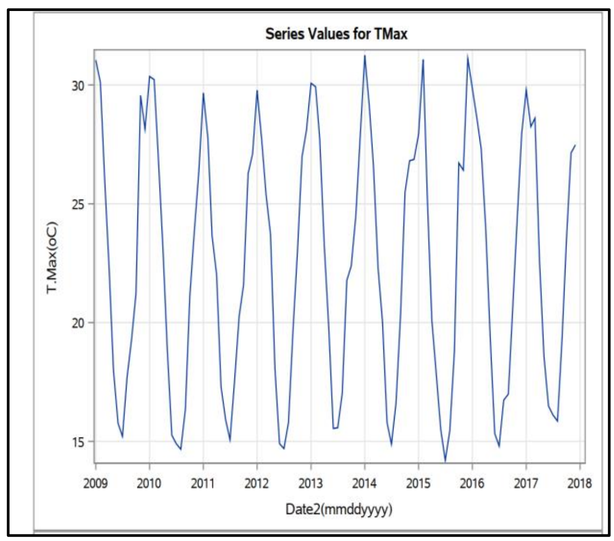
As per the Mesonet[2] Solar Radiation description, the values for variable Radn (MJ/m2) follow the scale given in the figure above. The summary statistics for “Radiation” variable is 16.9 MJ/m2 which indicates that solar radiation is at the higher end for the given region.

Time Series Analysis

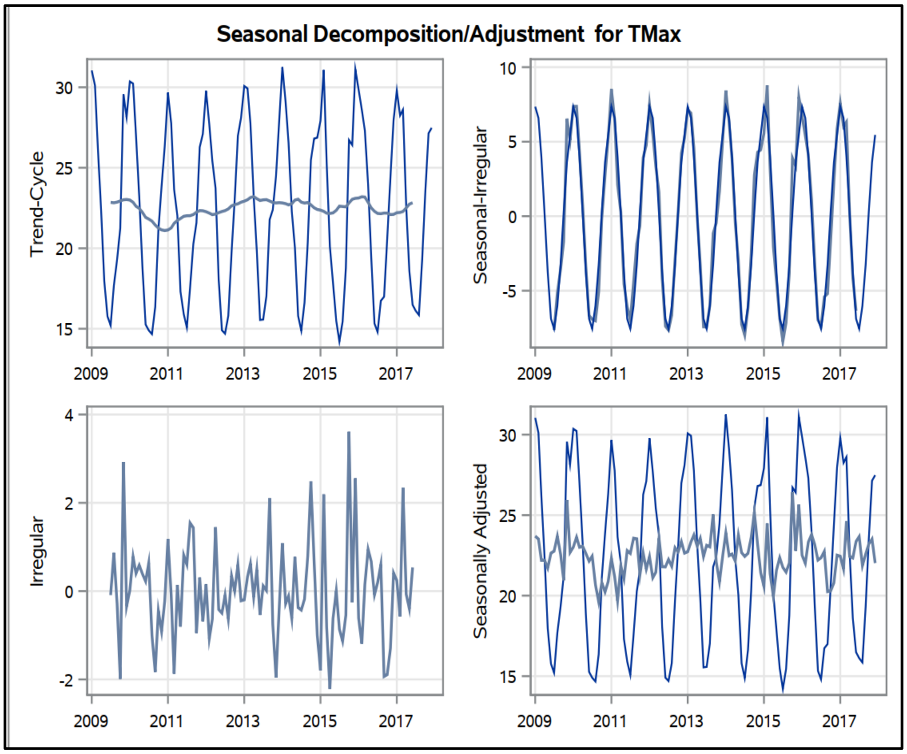
Tmax (Maximum Temperature)

To understand the trends and seasonal cycles for maximum temperature, time series analysis was done which involved data preparation task (to set the Date as the “index” variable for analysis) and data exploration was done for monthly intervals, taking the mean monthly temperature values .





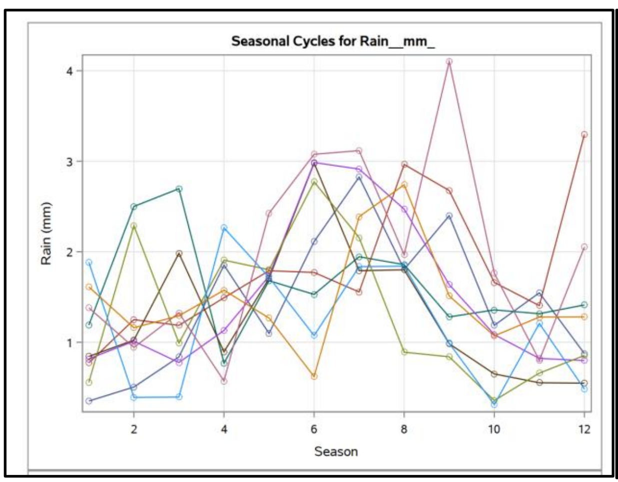
From the above graph, it seems months 6-8 (June-August) are coldest whereas December to February are warmest for the regions. And the fluctuations are higher in year 2014,2015 and 2016.

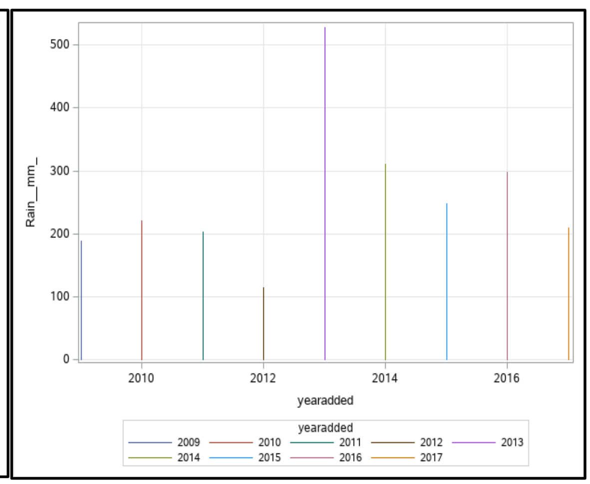


On doing seasonal decomposition analysis, it can be seen that Tmax shows slightly upward movement in trend statistics which can probably be due to global warming. The maximum temperatures are cyclical over the period 2009-2017, with slight deviations.

Rainfall (*Precipitation)*:

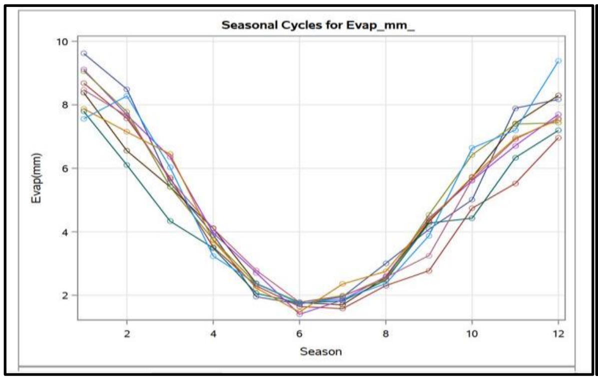
It seems that the distribution of rainfall is spread out throughout the year having some peaks in June-August periods. It is also observed that year 2013 had very heavy rainfall.

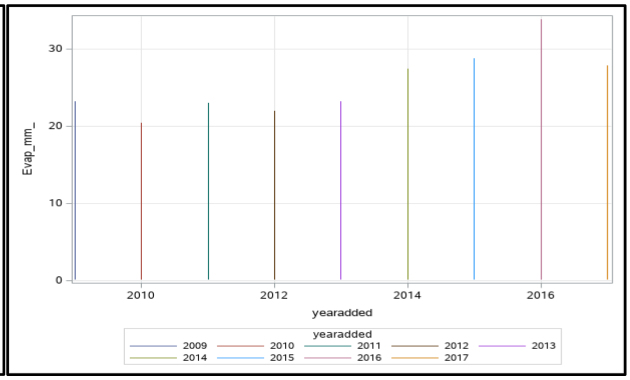




Evaporation:

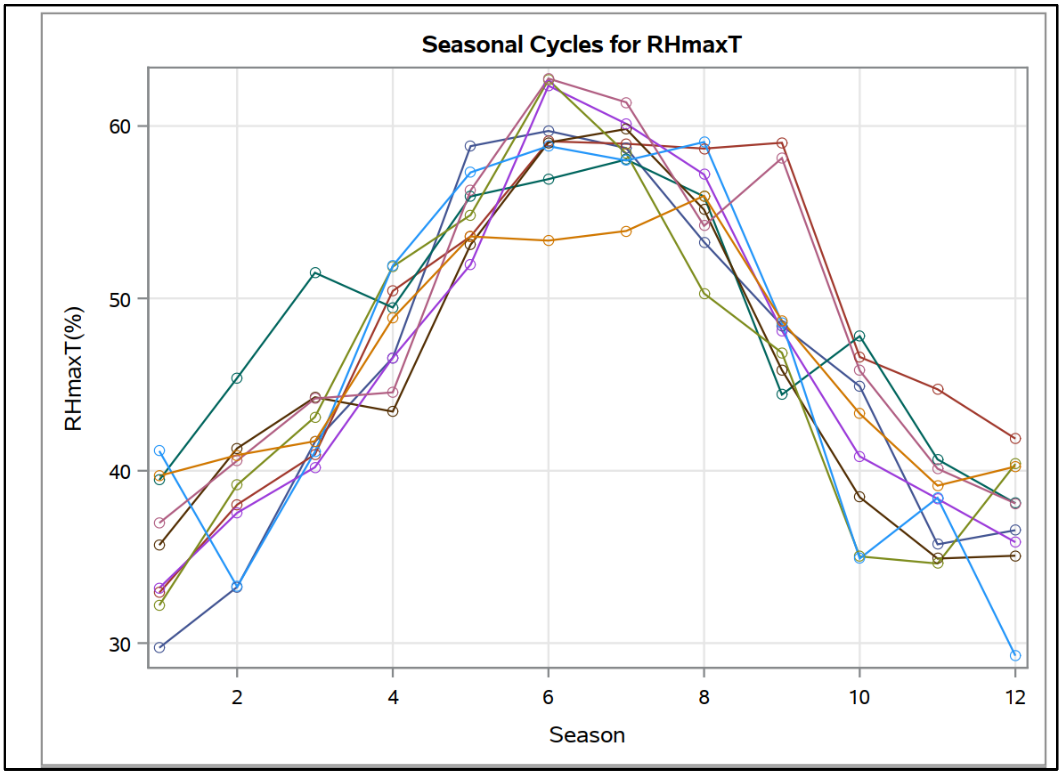
Evaporation is more in 2016. The rate of evaporation increases with an increase in temperature. So, when temperature is high, there is quick evaporation when compared to freezing temperatures.





Relative Humidity (RhMax)

The relative humidity variable was analyzed to find its seasonal cycles. We believe that a dry, less humid climate could trigger a bushfire. On doing time series exploration, it was found that the humidity is maximum during the months of June-August and minimum during the month of December-February.



Interpretation

As per the time series exploration of weather data, it can be findings can be summarized as:

1. The region has hot and dry summer months of December- February based on maximum temperatures and relative humidity values.
2. Conversely, the winter months (June-August) are cool and wet based on the Tmax, Rhmax and precipitation variables.

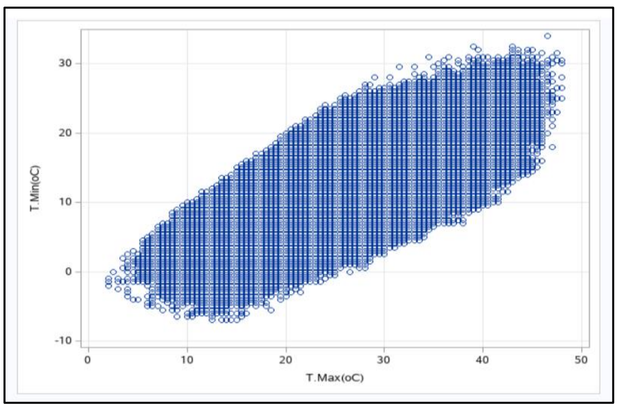
IV. <u>Analysis.</u> This step continues with refined focus on the Problem Statement based on outcomes from Data Exploration and concludes following a robust series of analytics methods have been applied to reach conclusions in support of the Problem Statement. This step may require additional research and data exploration as may be required to support findings.

a. Correlation insights and Interpretation

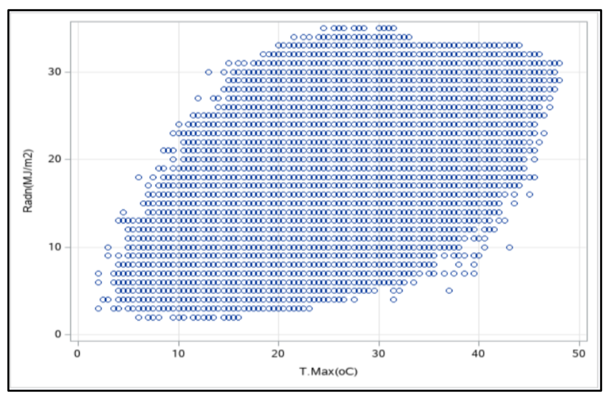
b. Logistic Regression.

<u>Correlation Insights and Interpretations:</u>

	Temp_Max	Temp_Min	Rain__mm_	Evap_mm_	Radn	VP__hPA_	RH_MaxT	RH_MinT	Fire
Temp_Max	1	0.69763	-0.18563	0.8485	0.68594	0.33965	-0.78027	-0.55008	0.04806
Temp_Min	0.69763	1	0.03929	0.59075	0.33092	0.63596	-0.25543	-0.5525	0.03562
Rain__mm_	-0.18563	0.03929	1	-0.17206	-0.19629	0.15718	0.33469	0.13188	-0.01244
Evap_mm_	0.8485	0.59075	-0.17206	1	0.80492	0.16161	-0.73269	-0.58162	0.04544
Radn	0.68594	0.33092	-0.19629	0.80492	1	0.08387	-0.68054	-0.33611	0.03748
VP__hPA_	0.33965	0.63596	0.15718	0.16161	0.08387	1	0.25594	0.23613	-0.00214
RH_MaxT	-0.78027	-0.25543	0.33469	-0.73269	-0.68054	0.25594	1	0.63348	-0.04549
RH_MinT	-0.55008	-0.5525	0.13188	-0.58162	-0.33611	0.23613	0.63348	1	-0.04803
Fire	0.04806	0.03562	-0.01244	0.04544	0.03748	-0.00214	-0.04549	-0.04803	1



Temperature Max and Min are highly correlated among each other (0.697). It can be interpreted in a way that there is almost a constant difference in Max and Min temperature of any day that leads to have high positive correlation. There is no drastic drop in a temperature from Max to Min for any particular day.



Similarly, Temperature affects the Evaporation and radiation. It can be explained by the series of subsequent process.

1. Temperature is directly proportional to Radiation, i.e. radiation increases with increase in temperature and that is well inferred by high positive correlation (0.686)
2. As the radiation increases, Thermal heat energy increases which causes the evaporation to happen from earth surface and that is well inferred by high positive correlation (0.8485).

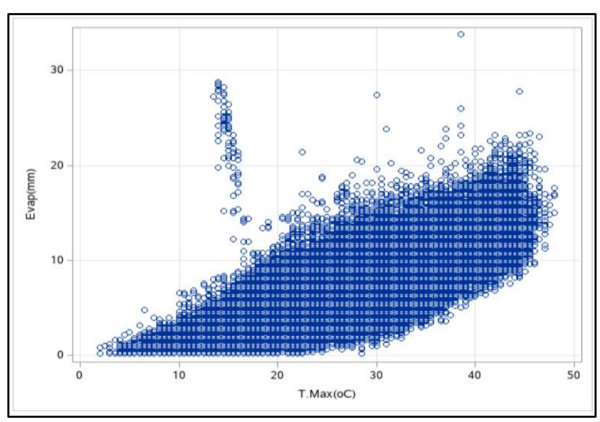
Temperature

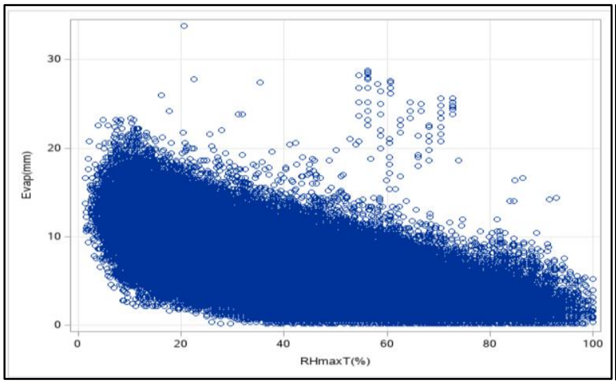
↓

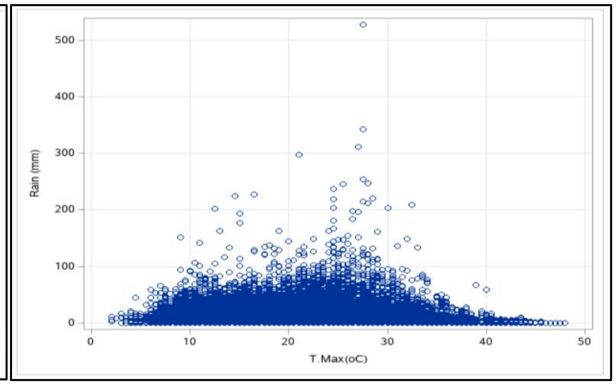
Radiation

↓

Evaporation







As the temperature increases, water content in Air (RH) also evaporates and thus relative Humidity increases. That's the reason RH is negativey correlated (-0.73 & -0.78) with Evaportaion and Temperature. It can also be interpeted that with high humididty, chance of getting rain also increases which is directly proportional to RH and inversely proportional to Evaporation and Temperature.

All the above variable with high correlation will lead to Multicollinearity which affcets the final model. So, It need to be taken care by Feature slection methods in later steps.

Correlation explanation with dependant variable – Fire:

	Fire
Temp_Max	0.04806
Temp_Min	0.03562
Rain__mm_	-0.01244
Evap_mm_	0.04544
Radn	0.03748
VP__hPA_	-0.00214
RH_MaxT	-0.04549
RH_MinT	-0.04803
Fire	1

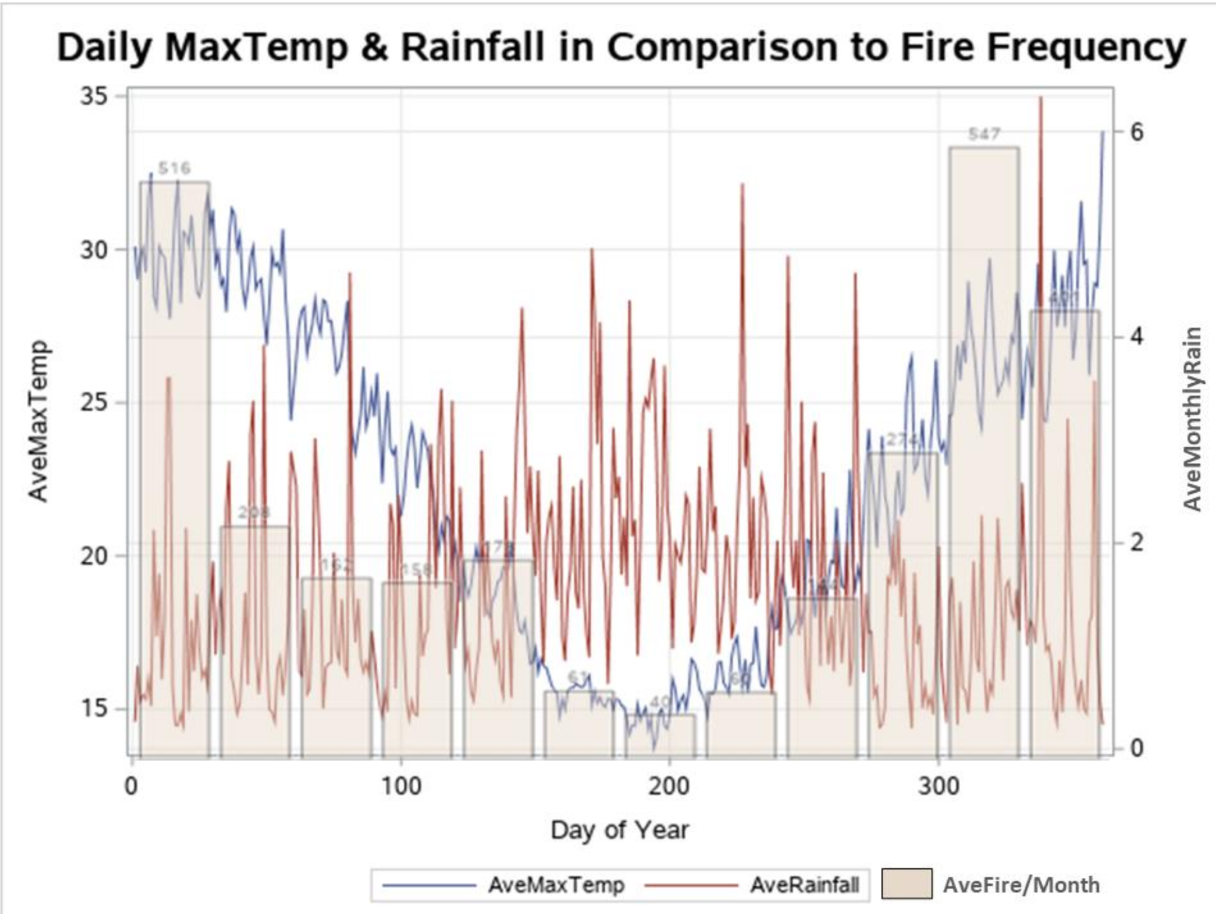
From the table, it is very clear Max Temperature is having positive correlation with Fire incident (0.04806). The reason for correlation very close to zero is because ratio of 1's to 0's in our dataset is very low. 2789 / 8777389 = 0.003178. Due to imbalance of 1's and 0's in dataset correlation is not having higher magnitude but it sows the direction of relation between dependent and independent variables.

Similarly, the rain is negatively correlated with Fire incident (-0.01244) and it is very evident that when there is a rain on any particular day, there is no fire incident. Again, due to imbalance of data explained previously magnitude of correlation values is very low.

It makes sense that Evaporation and Radiation also having positive correlation with Fire incident because Temperature is directly proportional to Evaporation and radiation.

And also, relative humidity is negatively correlated with Fire incident which is because RH is directly proportional to Rain and with high humididty, chance of getting rain also increases.

A visualization of these relationships is demonstated when overlaying an average daily temperature and rainfall series plot. It is clear that the seasonality of rainfall and temperature are inverse. As demonstrated through correlation, it is expected then that this relationship supports increase rates of wildifre incidents in the months correpsonding to high temperatures and lowe rainfall. This is clearly demonstated when evaluating the below graph which overlays the monthy freuency of fires across the data set.



Logistic Regression:

Logistic regression model performance decreases when we include all variables or few variables. Algorithm won’t converge, it means that the parameters being estimated in the model don't change between iterations. Therefore, we must take only the relevant variables in the model to get better accuracy.

For feature selection, with subject matter expertise we can select the important variable to use in model. With above correlation explanation, some variables are explained by series of subsequent processes. For e.g.: Temperature, Radiation and Evaporation are subsequent processes which cannot be removed as redundant variables. Same for Rain and RH variable.

Variable Clustering:

One another way to eliminate redundant variable is through variable clustering.

```
ods noproctitle;

proc varclus data=FIRE.WEATHER hierarchy plots outtree=Fire.Varclus_tree
        outstat=Fire.Varclus_stats;
    var Temp_Max Rain__mm_ Evap_mm_ Radn VP__hPA_ RH_MaxT RH_MinT Temp_Min;
run;
```

2 Clusters		R-squared with			
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	Temp_Max	0.8190	0.3288	0.2697	T.Max(oC)
	Rain__mm_	0.1024	0.0118	0.9084	Rain (mm)
	Evap_mm_	0.8559	0.1730	0.1742	Evap(mm)
	Radn	0.6804	0.0525	0.3373	Radn(MJ/m2)
	RH_MaxT	0.8161	0.0000	0.1839	RHmaxT(%)
	RH_MinT	0.4869	0.0305	0.5292	RHminT(%)
Cluster 2	VP__hPA_	0.8182	0.0005	0.1819	VP (hPA)
	Temp_Min	0.8182	0.2981	0.2590	T.Min(oC)

Based on the results of variable clustering, two clusters are formed with 6 and 2 variables. A variable selected from each cluster should have a high correlation with its own cluster and a low correlation with the other clusters. The 1-R^2 ratio can be used to select these types of variables. To interpret and select variables from variable clustering, 1-R^2 ratio needs to be very low.

Most Significant variable from each clusters. Most proportion of variation in cluster are explained by these variables.

Removed

It is observed that Temp_Min have high 1-R^2 ratio in cluster 2. So, it can be removed from the model. Similarly, in cluster 1, Evap_mm is the most significant variable. Other variables in cluster 1 cannot be removed from the model because with subject matter research these are all subsequent processes which cannot be removed as redundant variables.

With the final selected variables, Logistic regression model is run for dependent Binary variable "Fire" and other independent variables selected through above feature selection method. There are three methods in which Logistic regression can be done.

1. **Entry method** (runs with all independent variables at the same time)
2. **Backward elimination method** (All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation)
3. **Stepwise Selection method** (involves analysis at each step to determine the contribution of the predictor variable entered previously in the equation)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.0553	0.1373	872.4129	<.0001
Rain__mm_	1	-0.0324	0.00952	11.6114	0.0007
Evap_mm_	1	-0.0591	0.0123	23.0627	<.0001
Radn	1	0.0432	0.00441	95.9578	<.0001
VP__hPA_	1	0.0787	0.00649	147.0564	<.0001
RH_MaxT	1	-0.0281	0.00247	130.0354	<.0001
RH_MinT	1	-0.0226	0.00155	214.4183	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Rain__mm_	0.968	0.950	0.986
Evap_mm_	0.943	0.920	0.966
Radn	1.044	1.035	1.053
VP__hPA_	1.082	1.068	1.096
RH_MaxT	0.972	0.968	0.977
RH_MinT	0.978	0.975	0.981

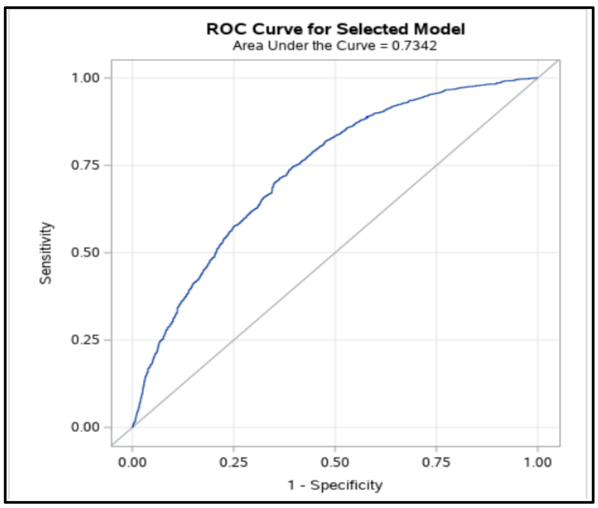
Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.000	2749	0	792E3	0	0.3	100.0	0.0	0.3	.
0.020	64	79E4	2302	2685	99.4	2.3	99.7	2.7	99.7
0.040	0	792E3	0	2749	99.7	0.0	100.0	.	99.7

From the above results, through backward elimination, it is found that algorithm is converging to probability level of 0.04, which is not good and also from classification table, it can be seen that either the model is either predicting all the values as 1 or all the values as 0. It is the same for other logistic method as well.

It is mainly because of imbalance of data because ratio of 1's to 0's in our dataset is very low. 2789 / 8777389 = 0.003178.

To rectify above shortcomings of logistic model, random sample of 3000 observation is selected from the final dataset which has fire incidents = 0. The above method is performed with **PROC surveyselect**. Random sample is stratified using each month of data, which means random sample has data from each month to avoid the bias in final model.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.1153	0.5082	37.5840	<.0001
Temp_Max	1	-0.0606	0.0143	17.8677	<.0001
Rain__mm_	1	-0.0500	0.0120	17.3737	<.0001
Radn	1	0.0314	0.00504	38.8286	<.0001
VP__hPA_	1	0.1418	0.0198	51.3456	<.0001
RH_MaxT	1	-0.0477	0.00668	51.1101	<.0001
RH_MinT	1	-0.0244	0.00234	108.5076	<.0001



The above graph is called a **Receiver Operating Characteristic curve** (or ROC curve.) It is a plot of the true positive rate against the false positive rate for the different possible Threshold points from Classification table. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The area under the curve is a measure of accuracy which is 73.4%.

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Correct	Sensitivity	Specificity	Pos Pred	Neg Pred
0.500	1783	2043	957	966	66.6	64.9	68.1	65.1	67.9
0.520	1688	2123	877	1061	66.3	61.4	70.8	65.8	66.7
0.540	1595	2212	788	1154	66.2	58.0	73.7	66.9	65.7
0.560	1505	2295	705	1244	66.1	54.7	76.5	68.1	64.8
0.580	1374	2378	622	1375	65.3	50.0	79.3	68.8	63.4
0.600	1277	2439	561	1472	64.6	46.5	81.3	69.5	62.4
0.620	1178	2500	500	1571	64.0	42.9	83.3	70.2	61.4
0.640	1081	2575	425	1668	63.6	39.3	85.8	71.8	60.7
0.660	984	2635	365	1765	63.0	35.8	87.8	72.9	59.9
0.680	865	2685	315	1884	61.7	31.5	89.5	73.3	58.8
0.700	766	2743	257	1983	61.0	27.9	91.4	74.9	58.0
0.720	677	2791	209	2072	60.3	24.6	93.0	76.4	57.4
0.740	557	2835	165	2192	59.0	20.3	94.5	77.1	56.4
0.760	436	2884	116	2313	57.7	15.9	96.1	79.0	55.5
0.780	280	2922	78	2469	55.7	10.2	97.4	78.2	54.2
0.800	135	2954	46	2614	53.7	4.9	98.5	74.6	53.1
0.820	36	2983	17	2713	52.5	1.3	99.4	67.9	52.4
0.840	7	2997	3	2742	52.3	0.3	99.9	70.0	52.2
0.860	0	3000	0	2749	52.2	0.0	100.0	.	52.2

1. The values of Event and Non-event for correct and Incorrect at cut-off point
2. The cutoff based should be based on our objective, level of impact and the tradeoff between sensitivity, specificity and false positivity values.
3. We should select cutoff value such that we can improve sensitivity of the model by restricting the false positive rate to the lowest minimum value.
4. The tabular view will allow you to analyze effect of minute change in probability cutoff value and select value up to two decimal places (for e.g. 0.70). Notice that, as we try to increase Nonevent identified correctly, Event identified incorrectly also increases accompanied by a decrease in Event identified incorrectly.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Temp_Max	0.941	0.915	0.968
Rain__mm_	0.951	0.929	0.974
Radn	1.032	1.022	1.042
VP__hPA_	1.152	1.108	1.198
RH_MaxT	0.953	0.941	0.966
RH_MinT	0.976	0.971	0.980

Interpretation of ODD Ratios:

- Estimate of Radn (1.032) can be interpreted as for **every unit increase in Radiation, odds of Fire incident happening is 3.2% increased.**
- Estimate of VP__hPA (1.152) can be interpreted as for **every unit increase in vapor pressure, odds of Fire incident happening is 15.2% increased.**
- Estimate of Rain__mm (0.951) can be interpreted as for **every mm increase in Rain, odds of Fire incident happening is 4.49% decreased.**
- Estimate of RhMax (0.953) can be interpreted as for **every unit increase in Relative Humidity, odds of Fire incident happening is 4.47% decreased.**

V. Summary of Findings.

a. Summary.

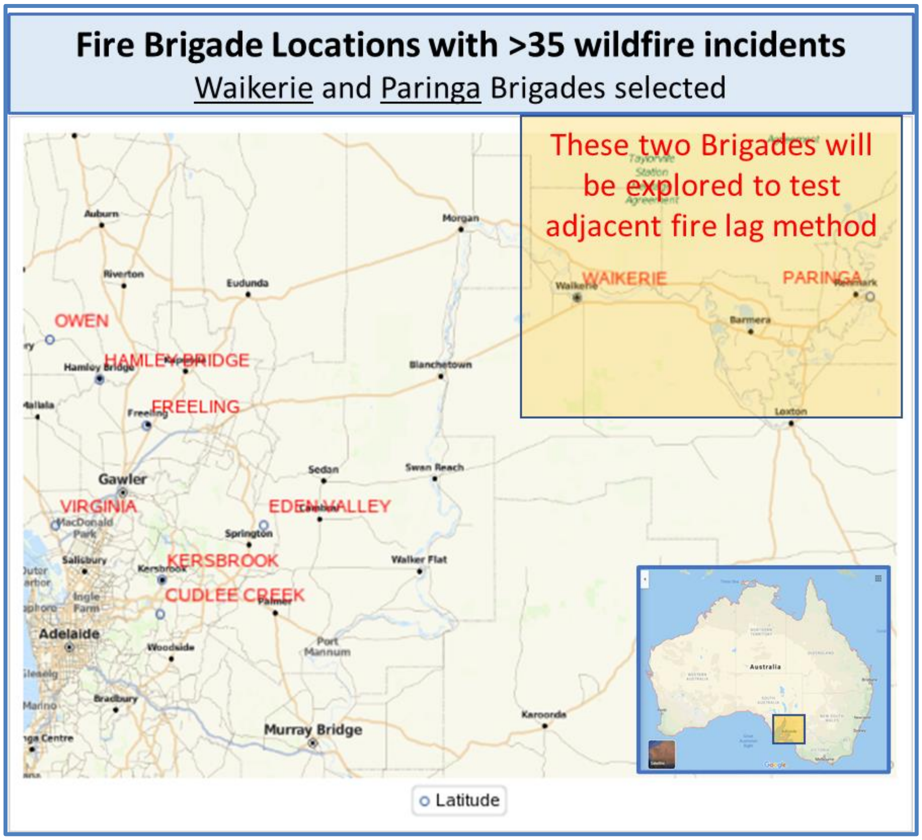
Temp_Max is the most important variable with the highest p-value and correlation with Fire. It is a positive indicator of Fire incident. **With a unit increase of radiation and vapor pressure** will increase the odds of fire incident happening by 3.2% and 15% thus making these variables also as an important variable and it is well supported by positive correlation results. Similarly, **Rain and Relative Humidity** decrease the odds of fire incident happening by 4.5%.

The above interpretation and result are valid for South Australia Locations and it is limited to set of variables used in analysis and there might be other factors which might influence the Fire incident happening

b. Next Steps.

- The next steps of this analysis may include expanding the time horizon to 20 or even 50 years if reliable data were available. This would enable more robust conclusion on weather and overall fire trends.
- Having a dependent binary variable is one reason that performing any type of time series analysis and prediction is challenging. In order to overcome this, a continuous variable may be collected such as size, severity, duration, etc. of fire events. However, to improve the logistic regression model without pursuing a time series prediction analysis, several new variables can be constructed from the existing dataset. These include adding variables that account for adjacent location's previous (e.g., day minus 1:3) "fire" values. In a similar manner, extra variables can be included that include the previous days' (e.g., d-1:3) weather values for each value. This would incorporate previous days' weather patterns and determine their significance in addition to the actual day's weather variables in a logistic regression model.
- First, cities with high frequency fire events are identified. Then these cities are plotted on a map, selecting tightly grouped city pairs to utilize for this hypothesis testing. The Brigade locations of Waikerie and Paringa were selected. An additional variable is then created, which counts fire incidents in the preceding three days in the city itself, as well as the adjacent city. The theory behind creating this variable is that proximity may result in fires spreading or conditions being similar enough to affect the likelihood of another fire event. To limit the introduction of new variables in this hypothesis test, the additional lag weather variable was not created at this time. The below map was created to aid in the city pair selection and in order to better visualize the concept.

Team Name: Wildfire



This concept of creating additional lag variables from existing data will be briefly explored by creating a second model to test against the original full variable set at these two locations. The null hypothesis will be that the model is not improved by adding this lag variable for adjacent locations fire incidents for d – 1:3. As shown below, the null hypothesis is rejected and therefore the model as a whole may be improved by expanding this adjacent city lag variable.

Model Comparison Comments

The P value shows the significance of adding the lag weather variable, which, in this case was heavily weighted on the model.

The odds ration further shows the significance adding this lag variable, considering increasing the odds of a fire event happening if this value is observed.

The model as a whole performed slightly better than the original weather variables only.

Logistic Regression Model & Performance
Including Adjacent Fire Lag Variable

Response Profile

Ordered Value	Fire	Total Frequency
1	0	6492
2	1	81

Probability modeled is Fire='1'.

Model Fit

Criterion	Intercept Only
AIC	875.194
SC	881.985
-2 Log L	873.194

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.0051	1.1385	12.3750	0.0004
Temp_Max	1	0.0415	0.0343	1.4648	0.2262
Rain__mm_	1	-0.0933	0.1214	0.5906	0.4422
Evap_mm_	1	-0.0751	0.0699	1.1547	0.2826
RH_MaxT	1	-0.0338	0.0153	4.8629	0.0274
Prec_and_Adj_Fire	1	0.9600	0.2432	15.5824	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Temp_Max	1.042	0.975	1.115
Rain__mm_	0.911	0.718	1.156
Evap_mm_	0.928	0.809	1.064
RH_MaxT	0.967	0.938	0.996
Prec_and_Adj_Fire	2.612	1.622	4.207

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.8	Somers' D	0.356
Percent Discordant	32.2	Gamma	0.356
Percent Tied	0.0	Tau-a	0.009
Pairs	525852	c	0.678

Logistic Regression Model & Performance
Original Variables Only

Response Profile

Ordered Value	Fire	Total Frequency
1	0	6492
2	1	81

Probability modeled is Fire='1'.

Model Fit

Criterion	Intercept Only
AIC	875.194
SC	881.985
-2 Log L	873.194

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.2367	4	<.0001
Score	30.0138	4	<.0001
Wald	27.3593	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.7939	1.1334	11.2041	0.0008
Temp_Max	1	0.0372	0.0340	1.1957	0.2742
Rain__mm_	1	-0.0910	0.1192	0.5831	0.4451
Evap_mm_	1	-0.0619	0.0702	0.7776	0.3779
RH_MaxT	1	-0.0358	0.0154	5.3906	0.0202

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Temp_Max	1.038	0.971	1.109
Rain__mm_	0.913	0.723	1.153
Evap_mm_	0.940	0.819	1.079
RH_MaxT	0.965	0.936	0.994

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.9	Somers' D	0.338
Percent Discordant	33.1	Gamma	0.338
Percent Tied	0.0	Tau-a	0.008
Pairs	525852	c	0.669

c. Limitations. Overall limitations in methods and findings, include the inability to construct a proper time series analysis model. This resulted from the group's collective knowledge gap on this advanced analytical technique. However, it can reasonably be expected that trends and cumulative effects of weather factor over time may be more significant than analyzing single day observations and drawings conclusions therein. Further, the data set on fire incidents selected only contained a binary incident of fire for the day the fire occurred. There was not an ability to analyze severity or duration of the fire event, which further limits any conclusions that can be drawn from the descriptive data analytics performed within this research.

---

[i] ENVIRONMENTAL SCIENCE. Reform forest fire management. *North MP, Stephens SL, Collins BM, Agee JK, Aplet G, Franklin JF, Fulé PZ Science. 2015 Sep 18; 349(6254):1280-1.*

[ii] Moreira N. 2006. Study links increase in wildfires to global warming. Boston Globe, 7 July 2006. Retrieved from: http://archive.boston.com/news/nation/articles/2006/07/07/study_links_increase_in_wildfires_to_global_warming/.

[iii] Northoff E. 2003. Fires are increasingly damaging the world's forests. Rome, Italy: FAO. Retrieved from: See http://www.fao.org/english/newsroom/news/2003/21962-en.html.

[iv] The State of Queensland. Data.SA: South Australian Government Data Directory. Retrieved from: https://data.sa.gov.au/data/dataset/south-australian-country-fire-service-brigade-incidents

[v] World Geocoding Service. Retrieved from: https://developers.arcgis.com/rest/geocode/api-reference/overview-world-geocoding-service.htm.

[vi] The Government of South Australia. SILO – Australian Climate Data from 1889 to Yesterday. Retrieved from: https://www.longpaddock.qld.gov.au/silo

[vii] Latitude and Longitude Finder. Retrieved from: https://www.latlong.net/

[viii] Average Precipitation for deserts (https://earthobservatory.nasa.gov/experiments/biome/biodesert.php)

[ix] Mesonet Scale (https://www.mesonet.org/images/site/DescriptionSolar.pdf)

[x] Condensation (http://www.atmo.arizona.edu/students/courselinks/fall12/atmo336/lectures/sec1/evap_cond.html)