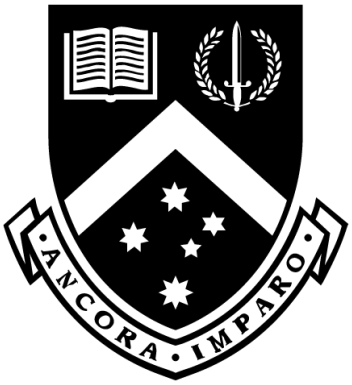


MONASH UNIVERSITY

ADS1002 - DATA CHALLENGES 2

GROUP 4

THE EFFECT OF CLIMATE DRIVERS ON AUSTRALIAN WEATHER

CONTRIBUTORS
AYDA ARYA (AA)
JACK LE (JL)
KHA DOAN (KD)
STEPHANIE WAINAINA (SW)

SUPERVISOR
DR SIMON BOWLY

OCTOBER 22, 2021

TABLE OF CONTENTS

CONTRIBUTIONS

SECTION	CODING	PRESENTATION	REPORT
Project Introduction			
Description		JL	JL
Aim		KD	KD
Background Information			
The Meaning Behind ENSO, IOD & SAM		JL	JL
What Are The Climate Drivers in the Dataset?		JL	JL
The Dataset		KD, AA	KD
Formatting/ Merging Data			
Importing Libraries and Datasets	AA, JL	AA	AA
Merging the Data	AA, JL	AA	AA
Manipulation/Formatting	AA	AA	AA
Exploratory Data Analysis			
Descriptive Statistics	JL, AA	JL, AA	JL, AA
Time Series	KD	KD	KD
Outliers	SW	SW	SW
Correlations	SW	SW	SW
Modelling The Data			
Initial Modelling and Accuracy Comparison	JL	JL	JL
Further Investigation	AA	AA	
Conclusion	SW	SW, AA	SW

CLIMATE DRIVERS REPORT

Group 4

Project Introduction

Description

This project looked at the effects of drivers on the Australian climate. We first had to format and merge the datasets into one master dataset. We then performed exploratory data through descriptive statistics, timeseries, outliers and correlations to identify trends before modelling the data to compare accuracy/correlation for Melbourne and Cairns. Since all data was sourced from the Bureau of Meteorology all data points in regards to temperature will be in degrees celsius (°C).

Aim

The aim of the project is to investigate the effects of three drivers - El Nino-Southern Oscillation (ENSO), the Indian Ocean Dipole (IOD) and Southern Annular Mode (SAM) – on the climate of Australia. This could be done by measuring the anomalies of these drivers each month and modelling the data for Melbourne and other Australian cities.

1. Background Information

1.1 The Meaning Behind ENSO, IOD & SAM

El Niño–Southern Oscillation (ENSO)

ENSO is a recurring climate pattern involving changes in the temperature of waters in the central and eastern tropical Pacific Ocean. There are three states:

El Niño: A warming of the ocean surface
La Niña: A cooling of the ocean surface
Neutral: Neither El Niño or La Niña.

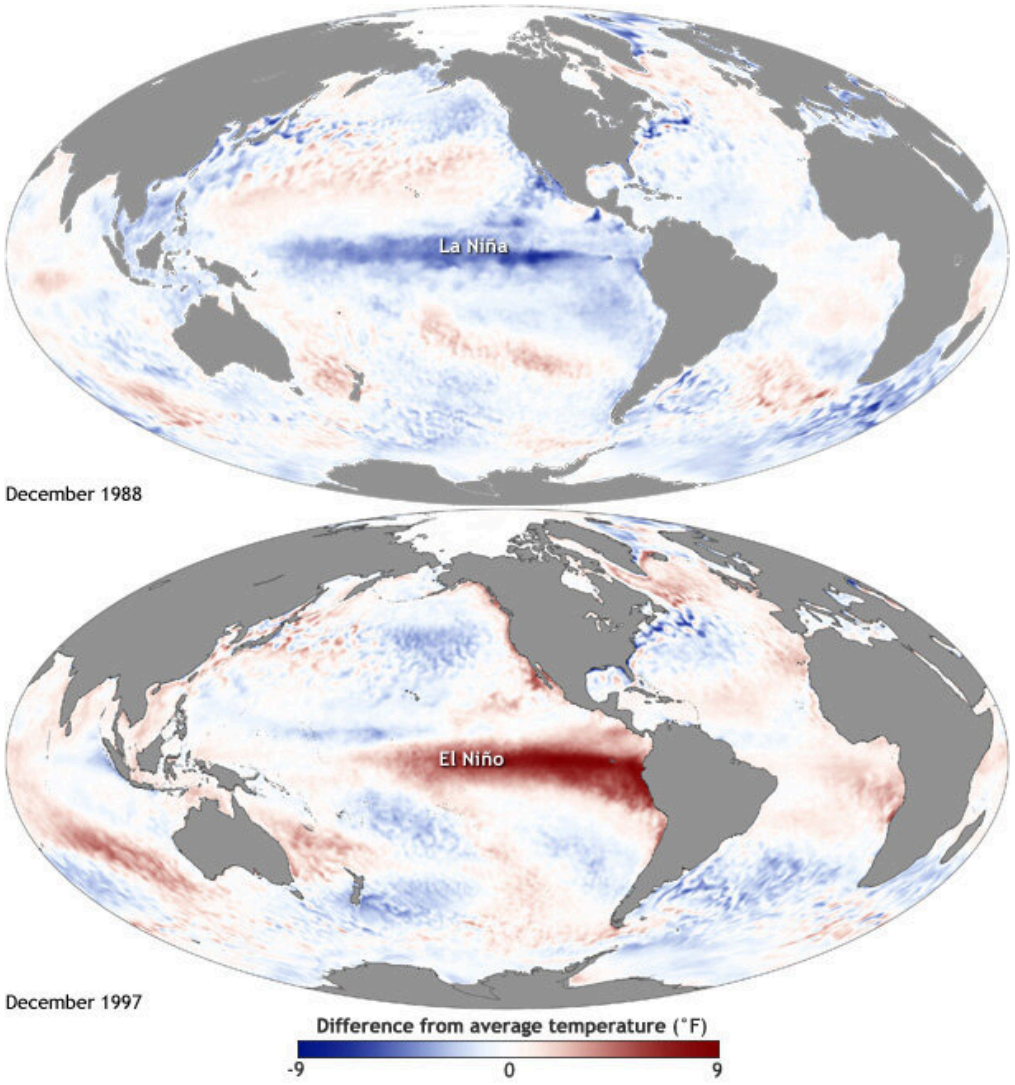
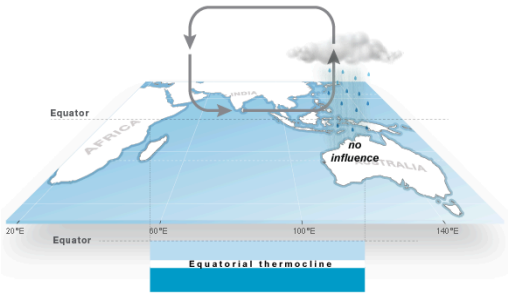


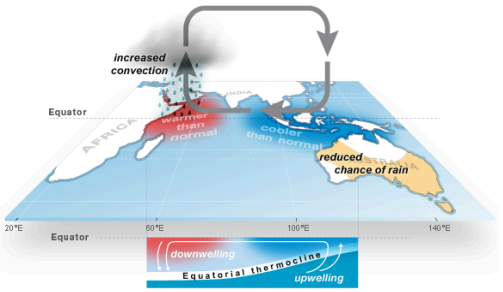
Figure 1.1.1
Source Image: http://www.climate.gov/media/6821

Indian Ocean Dipole (IOD)

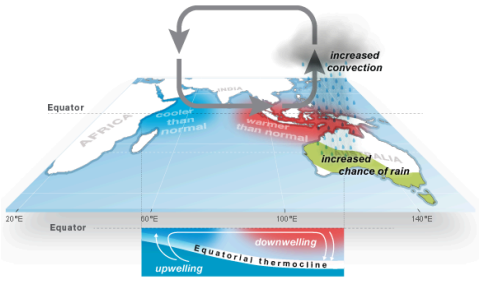
IOD are the sustained changes in the difference between sea surface temperatures of the tropical western and eastern Indian Ocean. The IOD is one of the key drivers of Australia's climate and can have a significant impact on agriculture. This is due to events generally coinciding with the winter crop growing season. The IOD has three phases: neutral, positive and negative.



Indian Ocean Dipole (IOD): **Neutral phase**



Indian Ocean Dipole (IOD): **Positive phase**



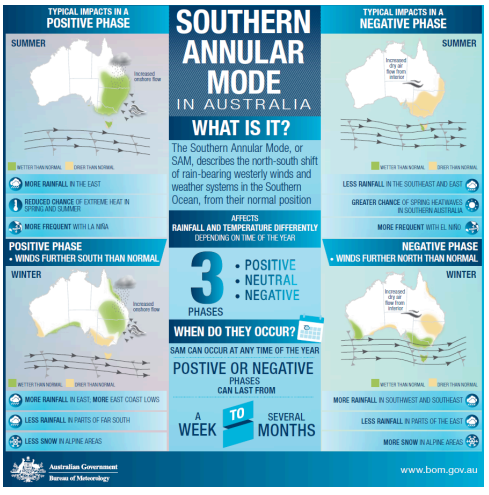
Indian Ocean Dipole (IOD): **Negative phase**

Figure 1.1.2

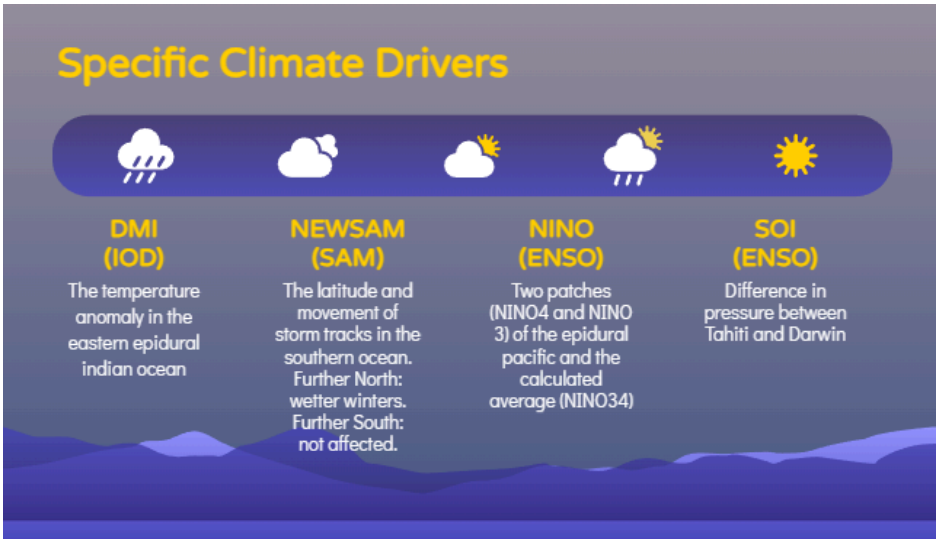
Southern Annular Mode (SAM)

SAM is a climate driver that can influence rainfall and temperature in Australia. The SAM refers to the (non-seasonal) north-south movement of the strong westerly winds that blow almost continuously in the mid- to high-latitudes of the southern hemisphere.

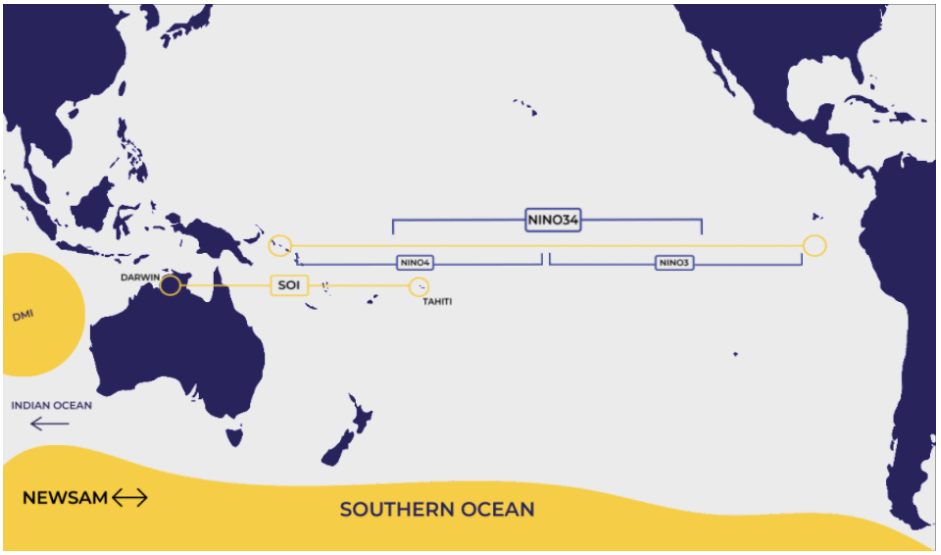
Figure 1.1.3



1.2 What Are The Climate Drivers in the Dataset?



1.3 Visual Diagram of the Drivers Around the World



1.4 The Dataset

The raw data we were provided gave us monthly measures of rainfall, maximum temperature and minimum temperature at two different stations in Melbourne, the Olympic Park and the Regional Office. These came in wide format csv files with rows representing years and, columns representing months, with records from 1856 to 2020. For the Melbourne data, we used an edited version of the raw data which melted the wide dataframes down. This in turn made it easier to incorporate all the data into one table as we could now include a month column.

To compare to the accuracy of the Melbourne model, we used a dataset from Cairns found on the Bureau of Meteorology website which had about a dozen missing values.

2. Formatting/ Merging Data

2. 1 Importing Libraries and Datasets

The first step to this project was importing the libraries and datasets. Since there were numerous datasets to import, and some were in txt while others were in csv, we used loops to import them. This greatly improved the neatness and readability of the code.

However we noticed that there was a problem with the code such that, if there were already headings on the data tables then these would transfer over to the melted version of the tables. The tables would look like this:

Figure 2.1.1

	Year	1	2	3	4	5	6	7	8	9	10	11	12
0	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1	1957	-0.87	-2.27	0.07	-1.97	-2.50	-0.87	1.00	-1.73	0.72	-3.12	-5.42	-2.63
2	1958	0.04	-2.84	-2.52	-0.30	-2.72	-0.97	-1.65	0.97	1.90	-0.16	0.30	0.82
3	1959	1.59	-0.19	-0.54	-1.25	-0.32	-1.12	0.18	-1.64	2.46	-0.46	3.28	-0.38
4	1960	0.90	0.85	3.40	-0.35	-0.60	-0.17	0.31	0.43	1.41	-1.17	1.18	1.36
...	...	...	...	...	...	...	...	...	...	...	...	...	...
60	2016	3.13	1.35	4.36	1.06	-0.92	3.66	0.81	-1.28	2.46	-0.89	-3.12	-1.52
61	2017	-1.12	-1.09	-1.56	1.65	1.82	1.98	0.16	0.31	0.42	-0.64	3.18	1.44
62	2018	2.72	1.02	-0.03	-1.66	0.01	-1.57	0.55	-0.78	1.83	2.76	2.84	1.44
63	2019	2.79	-1.87	1.47	0.86	-0.11	2.21	-2.20	-2.04	0.81	-1.97	-4.42	-1.78
64	2020	0.57	-0.36	2.05	-1.72	1.03	1.18	-0.97	-2.20	-0.25	1.79	1.14	2.28

	Year	Month	newsam
0	YEAR	1	JAN
65	YEAR	2	FEB
130	YEAR	3	MAR
195	YEAR	4	APR
260	YEAR	5	MAY
325	YEAR	6	JUN
390	YEAR	7	JUL
455	YEAR	8	AUG
520	YEAR	9	SEP
585	YEAR	10	OCT
650	YEAR	11	NOV
715	YEAR	12	DEC

2.2 Merging the Data

The next step was merging the datasets on the dates to make one master dataset in order to analyse the data and later begin modelling. We melted the datasets and made the year the index to simplify the process. After merging, our dataset looked like this:

Figure 2.2.1

	Year	Month	monthly_rain	max_temp	min_temp	dmi	newsam	nino	soi
0	1957	1	6.8	25.8	11.8	-0.371	-0.87	-0.56	0.6
1	1958	1	13.1	23.6	13.0	-0.331	0.04	1.54	-1.9
2	1959	1	13.9	29.5	15.8	0.038	1.59	0.43	-0.9
3	1960	1	45.8	28.8	16.1	-0.171	0.90	0.02	0.1
4	1961	1	33.6	29.4	16.6	-0.079	0.91	-0.15	-0.3
...	...	...	...	...	...	...	...	...	...
762	2016	12	39.6	25.5	14.6	-0.241	-1.52	-0.51	0.3
763	2017	12	128.4	25.0	15.2	0.179	1.44	-0.85	-0.1
764	2018	12	104.8	25.7	16.0	0.379	1.44	0.97	1.0
765	2019	12	6.2	24.6	14.0	0.312	-1.78	0.51	-0.6
766	2020	12	42.0	22.9	13.1	0.100	2.28	-0.98	1.8

767 rows × 9 columns

2.3 Manipulation/Formatting

The data types within the DataFrame then required some manipulation to make a datetime column, and then to convert the rest of the data types into integers so they could be used in the modelling stage (prior to this they were inconsistent as shown below).

Figure 2.3.1

```
Year             int64
Month           object
monthly_rain    object
max_temp        object
min_temp        object
dmi            float64
newsam          object
nino           float64
soi             object
dtype: object
```

However, the Year and Month values had to be strings so that we could construct our datetime column. After that was constructed the object data types were converted to numeric values and our final table was this:

Figure 2.3.2

	Date	Year	Month	monthly_rain	max_temp	min_temp	dmi	newsam	nino	soi
0	1957-01-01	1957	1	6.8	25.8	11.8	-0.371	-0.87	-0.56	0.6
1	1957-02-01	1957	2	30	25.9	13.7	-0.484	-2.27	-0.24	-0.1
2	1957-03-01	1957	3	37.4	23.4	11.8	-0.097	0.07	0.08	0.2
3	1957-04-01	1957	4	29.4	20.7	10.1	-0.225	-1.97	0.42	0.2
4	1957-05-01	1957	5	44.2	16.8	8	-0.247	-2.50	0.45	-0.7
...	...	...	...	...	...	...	...	...	...	...
762	2020-08-01	2020	8	61.6	15.1	7.6	-0.070	-2.20	-0.42	1.1
763	2020-09-01	2020	9	33.6	18.1	9.8	-0.084	-0.25	-0.66	0.9
764	2020-10-01	2020	10	74.2	19.1	10.9	0.233	1.79	-1.19	0.5
765	2020-11-01	2020	11	48.8	24.5	14	0.143	1.14	-1.01	0.7
766	2020-12-01	2020	12	42	22.9	13.1	0.100	2.28	-0.98	1.8

767 rows × 10 columns

Using datetime module to convert “Date” column into datetime format and set as the index.

Figure 2.3.3

Date	Year	Month	monthly_rain	max_temp	min_temp	dmi	newsam	nino	soi
1960-01-01	1960	1	45.8	28.8	16.1	-0.171	0.90	0.02	0.1
1960-01-02	1960	2	61.2	24.9	13.7	0.036	0.85	-0.24	0.1
1960-01-03	1960	3	10.8	25.1	13.7	-0.333	3.40	-0.08	1.0
1960-01-04	1960	4	195.0	19.7	10.6	-0.670	-0.35	0.01	0.8
1960-01-05	1960	5	95.2	14.8	8.0	-0.659	-0.60	0.05	0.5
...	...	...	...	...	...	...	...	...	...
2020-01-08	2020	8	61.6	15.1	7.6	-0.070	-2.20	-0.42	1.1
2020-01-09	2020	9	33.6	18.1	9.8	-0.084	-0.25	-0.66	0.9
2020-01-10	2020	10	74.2	19.1	10.9	0.233	1.79	-1.19	0.5
2020-01-11	2020	11	48.8	24.5	14.0	0.143	1.14	-1.01	0.7
2020-01-12	2020	12	42.0	22.9	13.1	0.100	2.28	-0.98	1.8

3. Exploratory Data Analysis

3.1 Descriptive Statistics

Figure 3.1.1

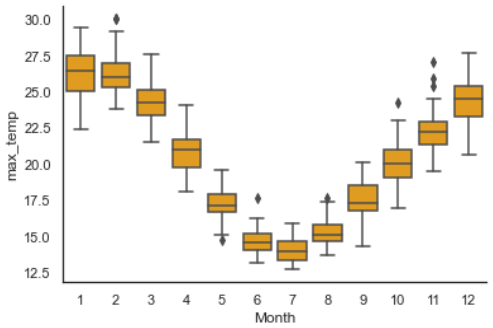
	monthly_rain	max_temp	min_temp	dmi	newsam	nino	soi
count	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000
mean	51.650326	20.282529	11.224250	-0.013824	0.039648	-0.006102	0.106910
std	32.076314	4.502626	3.232971	0.352919	1.787256	0.825724	0.941537
min	0.500000	12.700000	4.600000	-1.197000	-7.650000	-2.180000	-3.600000
25%	29.200000	16.300000	8.300000	-0.239500	-1.135000	-0.530000	-0.500000
50%	46.400000	20.400000	11.000000	-0.026000	0.090000	-0.060000	0.100000
75%	66.800000	24.250000	13.900000	0.208500	1.330000	0.480000	0.700000
max	238.200000	30.100000	18.600000	1.402000	4.920000	2.570000	2.900000

First laying down the foundations with the .describe() function in order to get general values for all the columns.

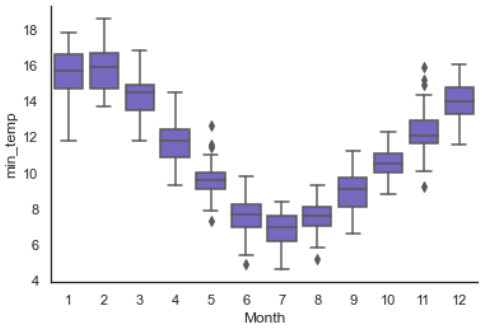
Furthermore, NEWSAM, NINO and DMI have more outliers on the positive end of the graph. Contrastingly, SOI and NEWSAM have more outliers on the negative end,

Figure 3.1.2

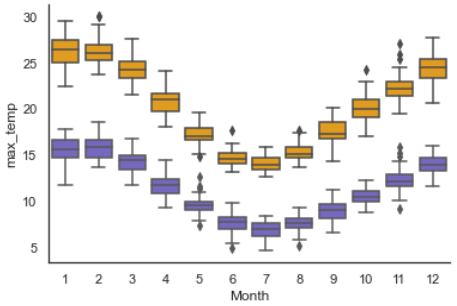
MELBOURNE'S MAX TEMPERATURE



MELBOURNE'S MIN TEMPERATURE



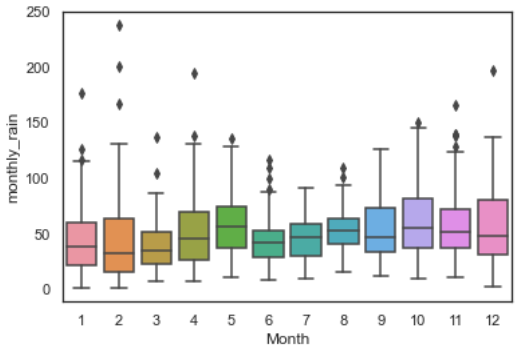
MERGED MELBOURNE'S MAXIMUM AND MINIMUM TEMPERATURE



Melbourne's maximum and minimum temperatures have been overall pretty normal. It should be noted that these temperatures change over the year based on the season. This explains the dip observed in the maximum and minimum temperatures in the months of 6-8 with a slow steady increase from 9-12.

Figure 3.1.3

MELBOURNE'S MONTHLY RAINFALL

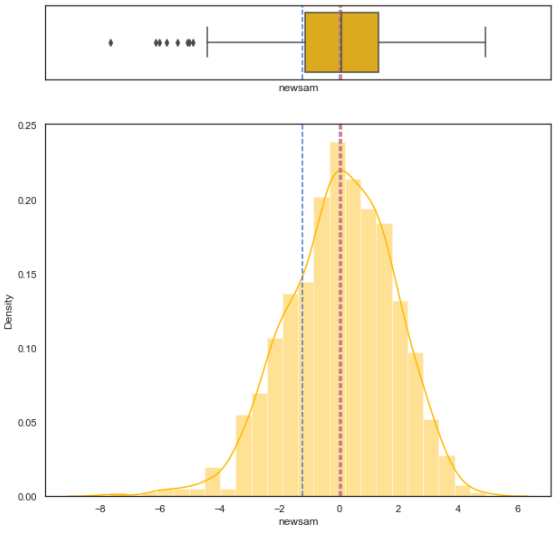


As expected, Melbourne's monthly rainfall is normal, with higher rainfall in the seasons of winter (Months 6-8) and Autumn (Months 3-5).

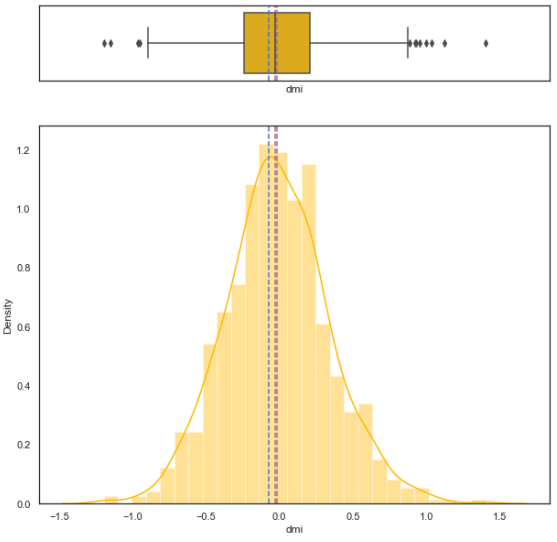
We experimented with sns.distplot to understand and visualize what the function would look like. This resulted in trying to subplot distribution plots, it was extremely difficult:

Figure 3.1.4

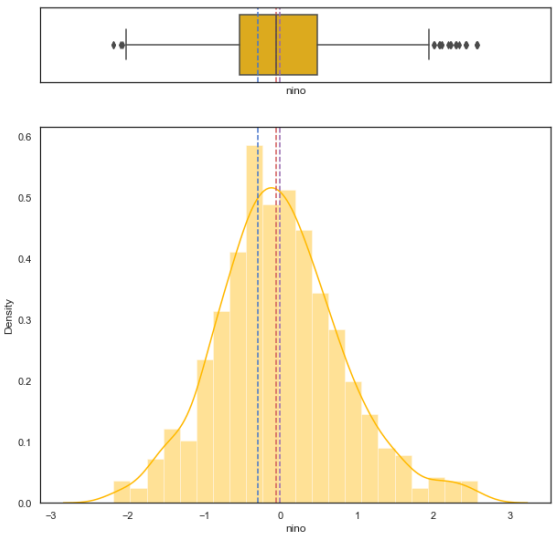
NEWSAM



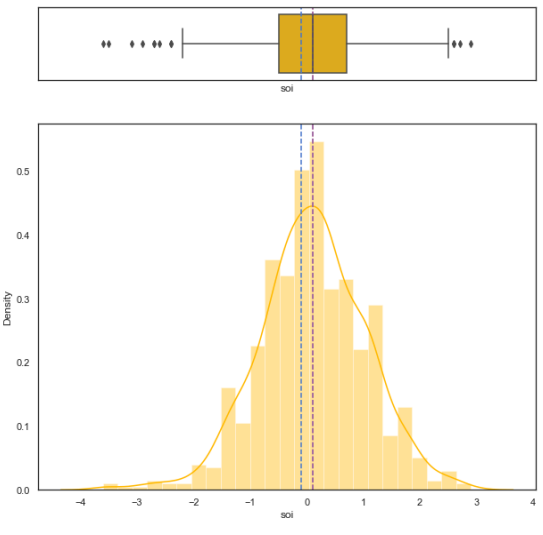
DMI



NINO



SOI



Here we can see a boxplot and normal distribution graph combined into one subplot, this allows us to further visualise not only the distribution of these drivers, but it also allows us to see anomalies prominent throughout the dataset. We can see clearly from Figure [...] that the distributions of the driver data appear to be normal, with some outliers on either side of the graphs.

3.2 Time Series

We used the resampled data for the mean of each driver in each decade, setting the start as 1960-01-01 (first date of first full decades) and the end as 2020-01-21 (latest date in data).The years 1957-1959 were removed from the dataset.

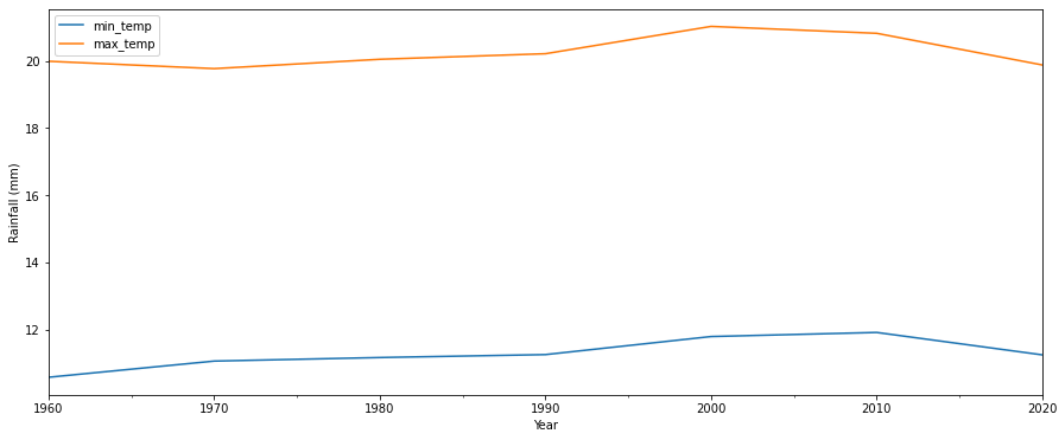
Figure 3.2.1

Date	5-year span	Unnamed: 0	monthly_rain	max_temp	min_temp	dmi	newsam	nino	soi
1960-12-31	1960	41.5	70.991667	19.791667	10.016667	-0.496417	0.629167	-0.090833	0.500000
1965-12-31	1965	77.5	53.110000	20.168333	10.540000	-0.032417	-0.325000	-0.021167	0.156667
1970-12-31	1970	137.5	49.631667	19.730000	10.746667	-0.047383	-0.527667	0.006833	0.155000
1975-12-31	1975	197.5	61.313333	19.908333	11.145000	-0.015550	-0.411000	-0.559333	0.760000
1980-12-31	1980	257.5	52.721667	19.826667	11.086667	-0.100633	-0.198167	0.113000	-0.061667
1985-12-31	1985	317.5	47.990000	20.033333	11.066667	-0.087317	0.083667	-0.051833	-0.193333
1990-12-31	1990	377.5	54.653333	20.056667	11.336667	-0.093850	-0.230000	-0.064333	0.078333
1995-12-31	1995	437.5	61.153333	19.948333	11.051667	0.030183	0.142167	0.402667	-0.578333
2000-12-31	2000	497.5	49.363333	20.593333	11.513333	0.038983	0.612667	-0.214833	0.323333
2005-12-31	1005	557.5	45.116667	20.815000	11.705000	-0.034300	0.148000	0.179333	-0.120000
2010-12-31	2010	617.5	43.026667	21.228333	11.885000	0.146200	0.489833	-0.251333	0.498333
2015-12-31	2015	677.5	49.096667	20.888333	11.966667	0.144450	0.676500	0.124500	0.211667
2020-12-31	2020	731.5	45.575000	20.810417	11.860417	0.272042	0.340417	0.192917	-0.020833

For the timeseries, we first created a date index with the date as the datetime format rather than object format. Due to the large number of datapoints, we decided to group each date record into 5-year means starting in 1960 and ending in 2020. This gave us the average for each driver, as well as rainfall and minimum and maximum temperatures.

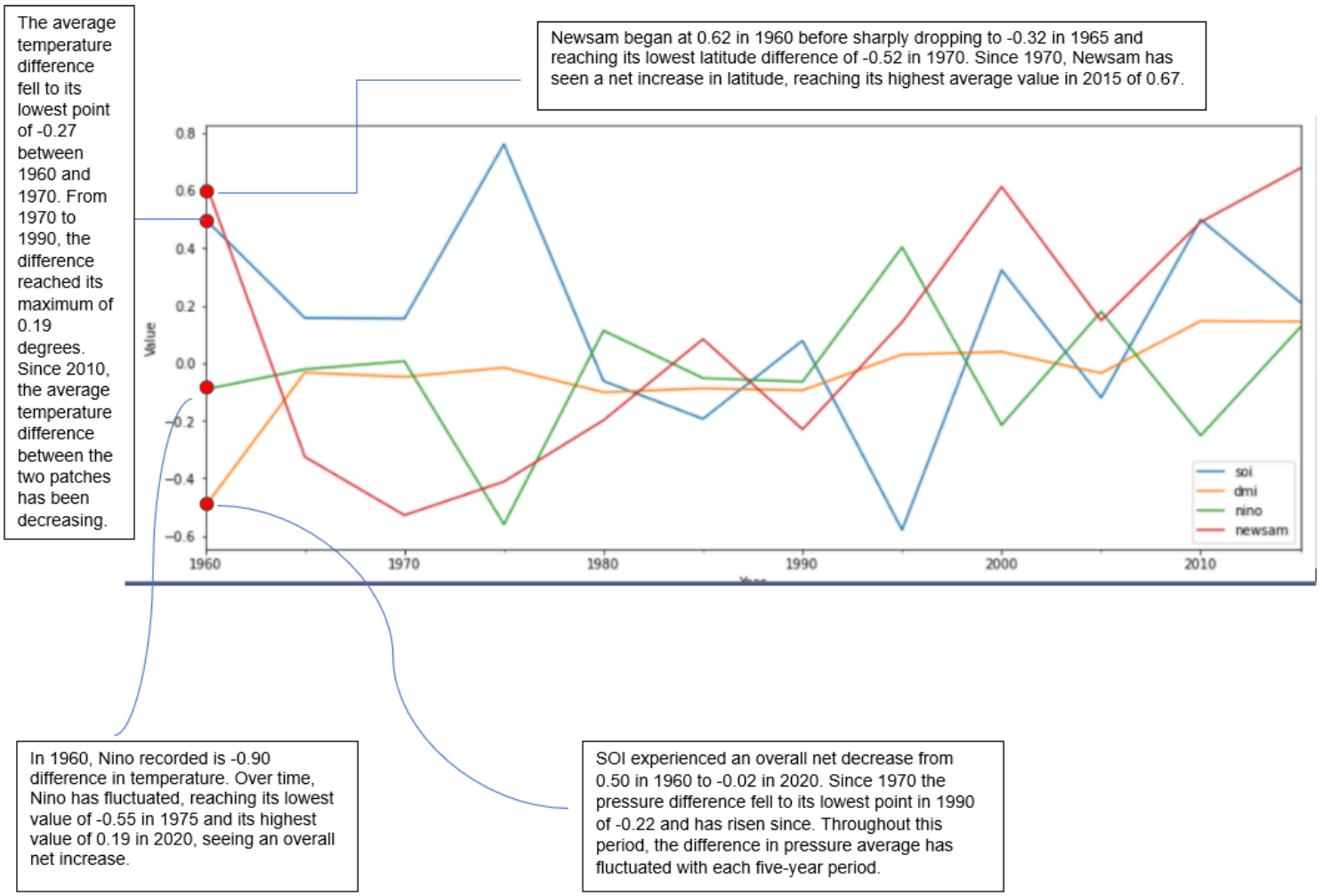
From 1970 to 1995, the average monthly rainfall fluctuated, before dropping to its lowest point of 43.02mm in 2010. It has been rising since to a value of 49mm in 2015.

Figure 3.2.2



The minimum and maximum temperatures have gradually risen until the 2000s decade. Temperatures then plateaued during 2000 and 2010 and began to decrease between 2010 and 2020.

Figure 3.2.3

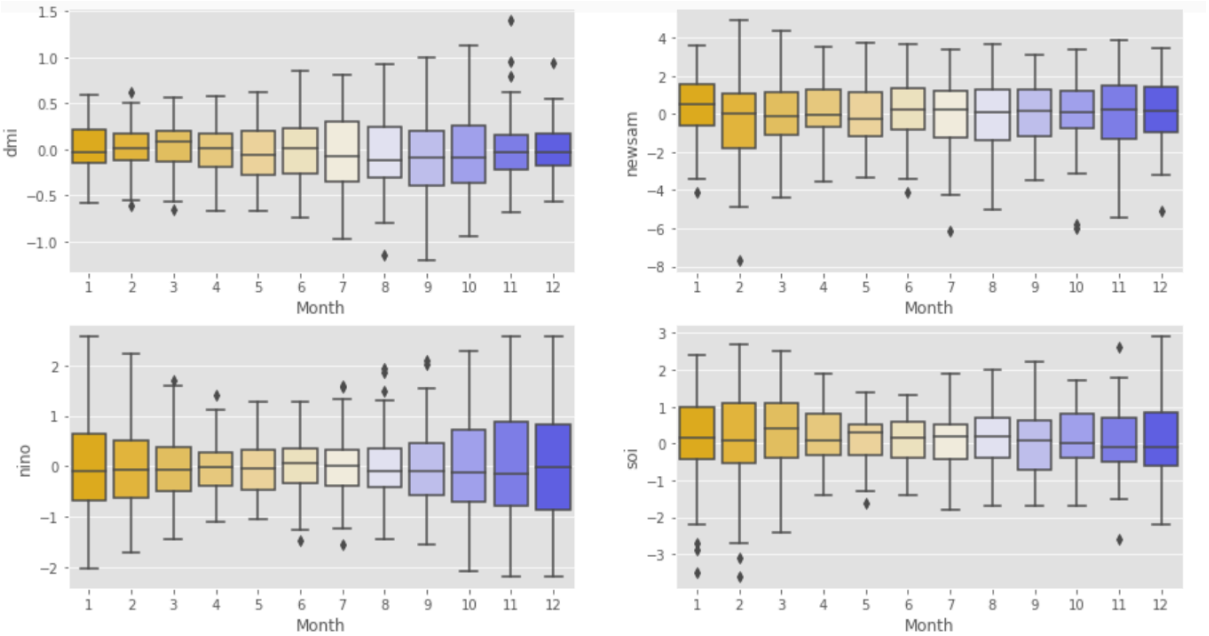


3.3 Outliers

As part of exploratory data analysis, we checked for outliers. This was done by using the 'seaborn' library and the 'boxplot' function. In addition to that, the highest and lowest points of the variables with outliers were outlined.

```
Highest allowed for nino: 2.4710717966520956
Lowest allowed for nino: -2.4832751864826044
Highest allowed for dmi: 1.0449339242309779
Lowest allowed for dmi: -1.0725819033704824
Highest allowed for monthly rain: 147.87926665291727
Lowest allowed for monthly rain: -44.57861476243485
Highest allowed for newsam: 5.401416476185314
Lowest allowed for newsam: -5.322120517906304
```

Figure 3.3.1



As seen, there are 5 variables with outliers, namely monthly rain,NINO,Newsam,DMI and SOI.

The outlier values could be removed by using the drop() to drop the outlier values found or by replacing the outlier values using IQR.

It is worth noting that the dataset only has around 700 entries, therefore removing the outliers could prove to be detrimental when modelling.

3.4 Correlations

We did the correlation analysis of two datasets, melb_weather and melb_anom. The correlation analysis was done using df.corr(), vmin and vmax to set the range of values (-1,1) and using sort_values() to plot a heatmap to check for any correlation. However, the heatmap did not give any concrete results thus prompting further analysis to be done. This is evident in the pictures below.

Figure 3.4.1

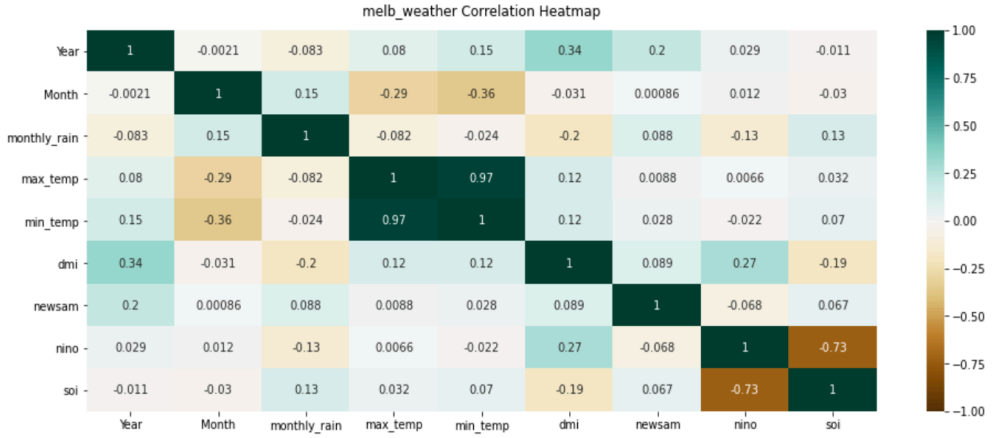
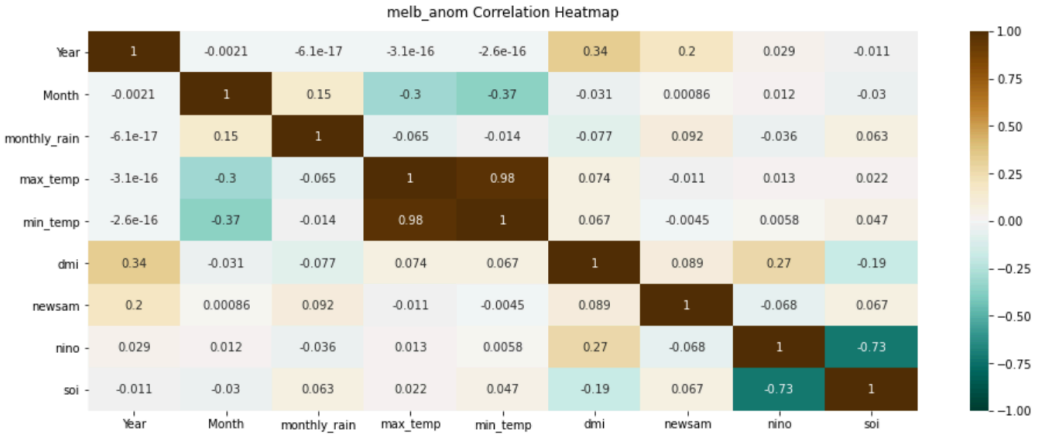


Figure 3.4.2



As seen above, there is no general correlation between the drivers and, Month,monthly rain, max and min temperatures.

Further analysis was done using the Pearson and the Spearman Rank correlation coefficient. This entailed exploring the correlation of specific variables in order to gain a better understanding. We used the Pearson and Spearman Rank correlation because the Pearson correlation coefficient measures the linear relationship between the two datasets. The Spearman Rank does not assume that data is from a specific distribution, so it is a non-parametric correlation measure that can highlight non-linear association.

The Pearson correlation coefficient was used to explore the relationship between the variables that showed any correlation. This is because it helps capture the strength of the linear relation between two variables hence, making it easier to draw conclusions.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but they are reasonable for datasets larger than 500 or so of which this dataset is.

3.5 Correlation analysis for melb_weather

This was done by creating a new dataframe ,df2, with the variables that showed significant correlation, namely:

- NINO
- SOI
- 'Min_temp'
- 'Max_temp'
- 'Month'

Figure 3.5.1

:

	nino	soi	min_temp	max_temp	Month
nino	1.000000	-0.695979	-0.030162	-0.009293	-0.000261
soi	-0.695979	1.000000	0.059108	0.029652	-0.060415
min_temp	-0.030162	0.059108	1.000000	0.968473	-0.339439
max_temp	-0.009293	0.029652	0.968473	1.000000	-0.287935
Month	-0.000261	-0.060415	-0.339439	-0.287935	1.000000

Exploring the correlation between Max and Min temp

```
Correlation coefficient:  0.9677013662932092
P-value:  0.0
```

The function produced correlation coefficient 0.9677013662932092 that is close to 1, thus confirming a strong correlation. The returned p-value is < 0.001, and so it confirms strong certainty in the result.

Exploring the correlation between Month and min_temp

```
Correlation coefficient:  -0.35872320413542874
P-value:  1.0446552568610767e-24
```

The function produced correlation coefficient -0.35872320413542874, that is close to 0, and it confirms a general negative correlation. The returned p-value is < 0.001, and so it confirms strong certainty in the result.

Exploring the correlation between nino and soi

```
Correlation coefficient:  -0.7260369503275145
P-value:  1.539387961685675e-126
```

The function produced correlation coefficient -0.7260369503275145, that is close to -1, and it confirms a strong negative correlation. The returned p-value is < 0.001, and so it confirms strong certainty in the result.

3.6 Correlation analysis for melb_anom

This was done by creating a new dataframe ,df1 with the variables that showed any correlation. The overall coefficient for df1 was done using the spearman rank coefficient correlation.

Figure 3.6.1

	nino	soi	min_temp	max_temp	Month
nino	1.000000	-0.695979	-0.006465	-0.006888	-0.000261
soi	-0.695979	1.000000	0.040081	0.022760	-0.060415
min_temp	-0.006465	0.040081	1.000000	0.975982	-0.349920
max_temp	-0.006888	0.022760	0.975982	1.000000	-0.295421
Month	-0.000261	-0.060415	-0.349920	-0.295421	1.000000

It appears that the spearman correlation coefficient of df1 is the same as that of df2 above. This is odd as in the heatmaps, melb_anom is more correlated than melb_weather.

Exploring the correlation between Max and Min temp

```
Correlation coefficient:  0.9759410185659803
P-value:  0.0
```

The function produced correlation coefficient 0.9759410185659803, that is close to 1, and it confirms a strong correlation. The returned p-value is < 0.001, and so it confirms strong certainty in the result.

Exploring the correlation between Month and min_temp

```
Correlation coefficient:  -0.365102580816332
P-value:  1.3451806606325532e-25
```

As expected, the function produced correlation coefficient -0.365102580816332, that is close to 0, and it confirms that it has a general negative correlation. The returned p-value is < 0.001, and so it confirms strong certainty in the result.

Exploring the correlation between nino and soi

```
Correlation coefficient:  -0.7260369503275145
P-value:  1.539387961685675e-126
```

As expected, the function produced correlation coefficient -0.7260369503275145, that is close to -1, and it confirms a strong negative correlation. The p-value is < 0.001, and so it confirms strong certainty in the result.This is further seen in the regression plot below.

3.7 Findings

In melb_weather the only data that is correlated is nino and soi have a strong negative correlation, max_temp and min_temp have a strong positive correlation and, Month and min_temp have a negative correlation. While, in melb_anom the data is more correlated than in melb_weather. It is more correlated by 0.1. This may be due to the included anomalies hence making the dataset larger. Thus,it has been confirmed that there is certainty in the results despite the outliers being included. However, no conclusion can be drawn as to why they are correlated.

4. Modelling The Data

4.1 Initial Modelling and Accuracy Comparison

Figure 4.1.1

Using variables Month, Year, DMI, Newsam, NINO, SOI to predict anomalies.
Melbourne's testing and training scores are clearly not optimal. Although scores are mostly positive, it is extremely low.

Here we used nearly all the variables that were in the dataset, we initially approached this analysis without much thought of isolating and finding correlation between pairs of variables. Our initial thoughts were to attempt to find correlations between all variables to create a model that encapsulates all parts of the dataset.

So, in an effort to find the root of these low scores, we decided to use data from a different city in Australia more closely situated to the driver locations. This would give us an idea of if the low accuracy score was because of a low impact of drivers on Melbourne weather or not.

4.2 Further Investigation

We wanted to figure out if using different variables for the input of the model would improve the accuracy scores. We noticed that the correlation between NINO and SOI was quite high, so it's possible that they are not both needed in order to get similar results.
Below are tables representing the results we got, with the input variables as the titles, the output variables as rows and testing and training scores as columns.

Figure 4.2.1

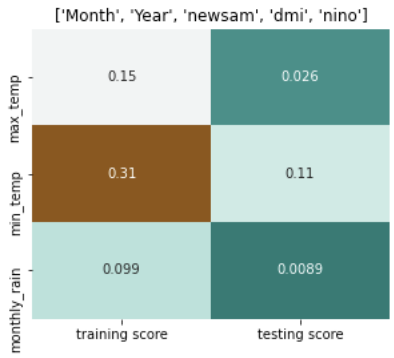


Figure 4.2.2

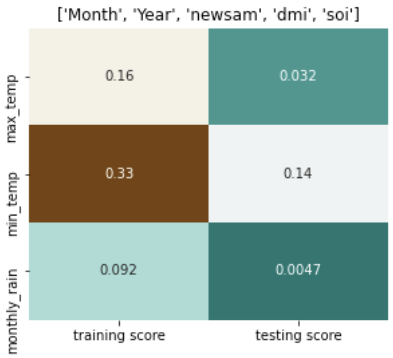


Figure 4.2.3

Figure 4.2.4

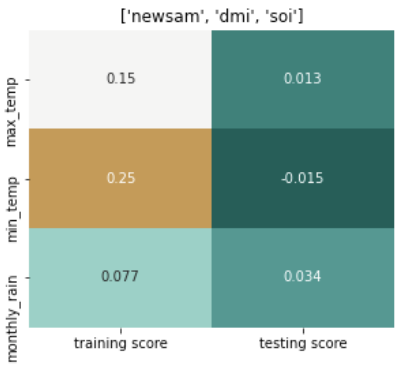


Figure 4.2.5

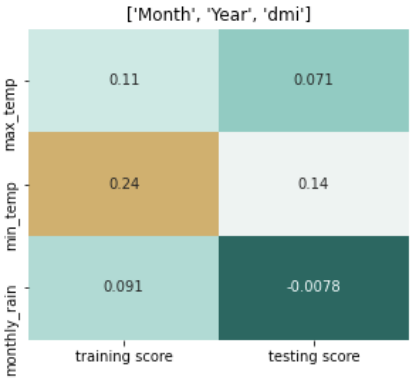
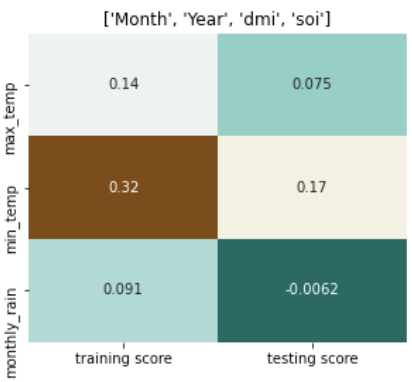


Figure 4.2.6



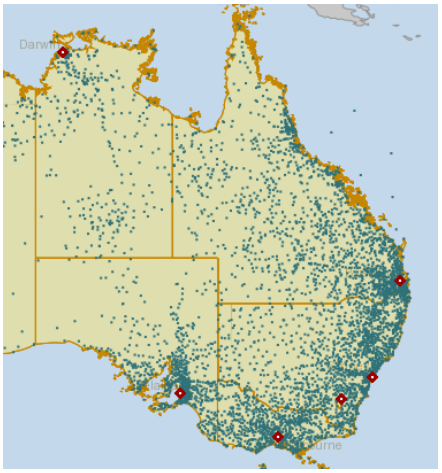
4.3 Findings

We observed that if you were to use a model to just predict monthly rainfall, then using the model from [...] would suffice. Similarly, if we were just wanting to model the minimum and maximum temperatures, we could use the models from [...] and [...] respectively.

Additionally, if you wanted to only find values for weather in Melbourne using just one driver, it would be best to use the model from [...]. For two drivers, we would recommend the [...] and for three drivers [...]. For the comparison below however, we decided that it would be better to assume that we had all the values for the drivers.

4.4 Using a Different State - Cairns

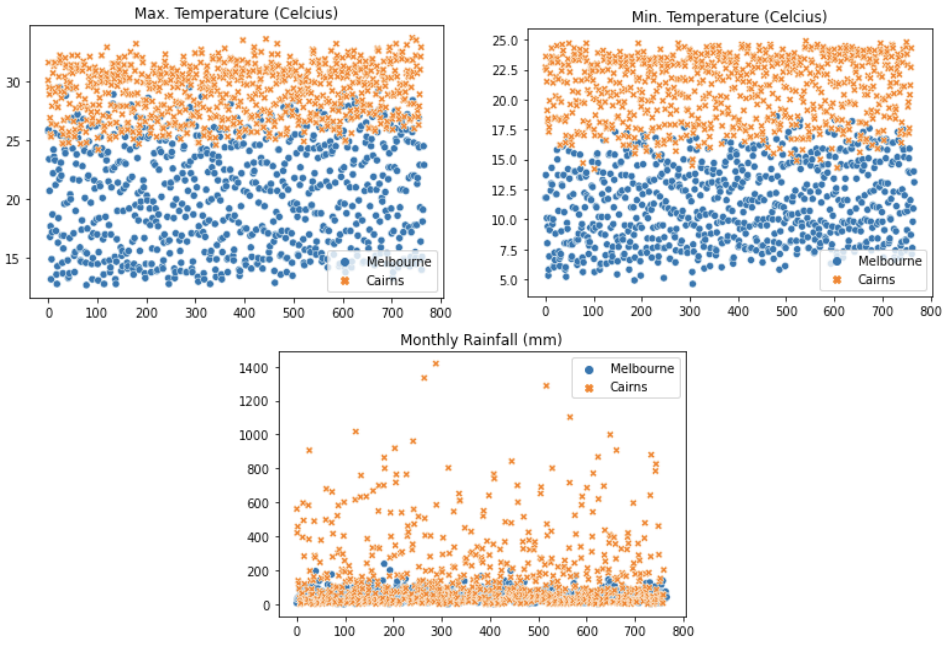
Figure 4.4.1



As seen from Fig. [...] the weather stations are situated across the beach line of Australia, however, Darwin's weather station had several null values, and so did Brisbanes. We wanted the weather stations to be as close to the drivers as possible, in order to get the most accurate output. We mainly wanted to focus on the ENSO region, which is closest to Darwin or Cairns, as the other weather stations aren't closer to any other drivers.

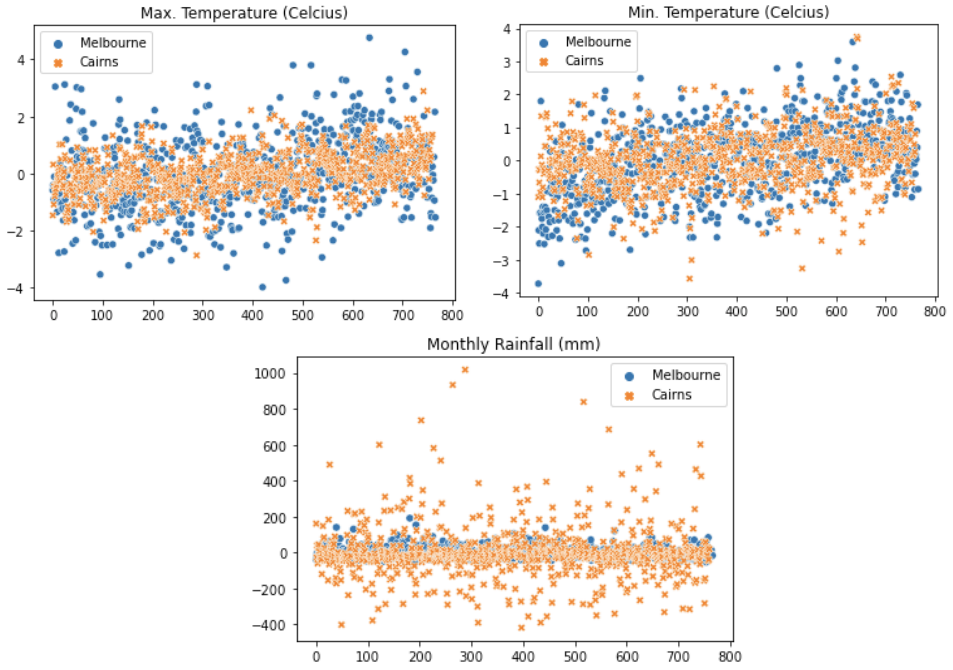
We started by importing our data from Cairns using the same source that we did for Melbourne weather, the station being Cairns AERO. It was important to first have a look at how Melbourne and Cairns weather differed, so we knew what to expect from our modelling. Below are the values for maximum temperature, minimum temperature and monthly rainfall for the two cities, with the index of the data points laid out at the bottom.

Figure 4.4.2



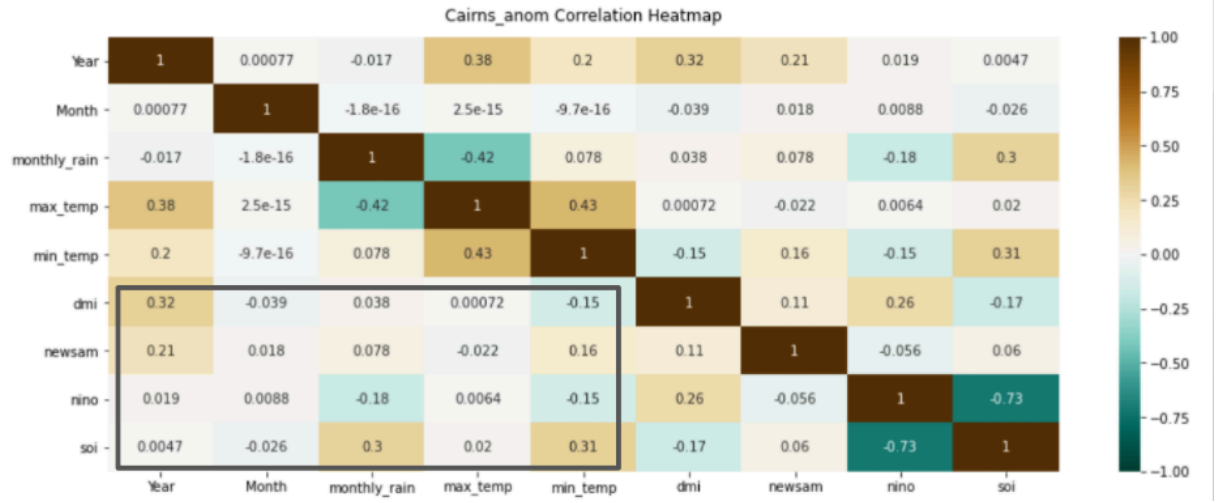
As we can see from Figure 4.4.2, minimum and maximum temperatures between Cairns and Melbourne differ drastically and there are jumps in values for Cairns data also in monthly rainfall. We should also take a look at the difference between values in the anomalies datasets as those were the ones we would use in analysis.

Figure 4.4.3



From the figures above we can see that by getting the anomalies of the values, the effect of getting the anomalies is that values of Melbourne and Cairns balance out, with slight differences.

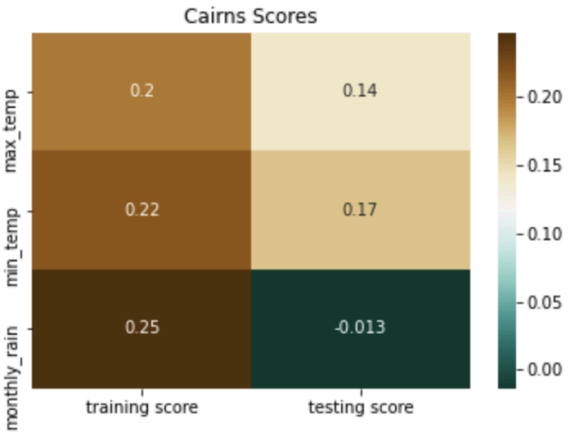
Figure 4.4.4



With a new correlation table for Cairns' set of data, the correlation scores were still pretty low, and very surprising. There's still not a lot of connection between the indices and the output variables.

Figure 4.4.5

Figure 4.4.6



Between the two heatmaps, you'll notice a general improvement to the scores, but because this improvement is insubstantial, we can't say with confidence that using drivers that are perhaps more influential to Victorian weather will send the scores for Melbourne up to the higher end.

Conclusion

This investigation shows that the location of drivers in relation to a city does not necessarily imply a direct impact of these indices on our variables. We have further shown this by comparing the accuracy scores of anomaly predictions between Melbourne and Cairns, which is much closer to where the driver measurements are made than Melbourne, and not seeing much change in the result.

The results help give a general idea of the trends in rainfall and changes in temperature. However, it is worth noting that weather prediction models comparatively require a lot of complexity to give accurate results. Thus making it quite difficult to make any definitive conclusions with only a handful of variables. This in turn leads to creating a very simple linear model.

We recommend that elements of mathematical modelling for variables such as air velocity,pressure, density and humidity should be included as it would greatly increase the accuracy of the model built. Additionally, if provided with ample time and skill, a project like this would be able to see much more accurate results by adding layers of complexity or using more complicated model templates.

Nevertheless, if it were not for the time constraint we would have also investigated the effects of splitting the data up into seasons, and also the effect of removing outliers from the Cairns dataset in particular.