DATA-614 Project Report - Australia Weather Prediction

Group 4: Deshant Sachdeva, Ketan Bassi, Nanditha Sriram, Saurabh Anand

Introduction

Climate change can affect the intensity and frequency of precipitation. Warmer oceans increase the amount of water that evaporates into the air. This in turn can produce more intense precipitation. The potential impacts of heavy precipitation include crop damage, soil erosion, and an increase in flood risk due to heavy rains, which in turn can lead to injuries, drownings, and other flooding-related effects on health. The prediction helps people to take preventive measures and can help reduce possible financial loss. Accuracy of rainfall statements also has importance for countries whose economy relies heavily on agriculture. The stakeholders for our analysis can extend to a farmer who wants to know which is the best month to start planting and also for the government policy makers who need to prepare any policy for preventing floods during the rainy season & drought during the dry season in Australia.

Objective

Australia is experiencing higher temperatures, more extreme droughts, fire seasons, floods and more extreme weather due to climate change. Higher temperatures create a range of extreme weather and climate events: longer droughts in some areas of the continent, and in others, heavier rain storms due to greater evaporation. Marine heatwaves are on the rise devastating Australia's kelp forests, seagrass meadows, coral reefs and all the underwater creatures that depend on them. These impacts also affect humans—the creatures and habitats are sources of food and income. Coral bleaching has increased in frequency and severity on the Great Barrier Reef. It is now occurring so frequently that large areas are unlikely to ever recover. It can adversely affect the businesses and increase financial losses.

Our aim is to conduct a series of studies by using machine learning techniques to create models that can predict the amount of rainfall and the maximum temperatures based on the weather data available for the next one year in order to understand the impact of climate change in major cities in Australia.

Guiding Questions

1. Which cities have the highest average rainfall and average maximum temperature across Australia?
2. To predict if it will rain or not in the upcoming weeks.
3. To predict the maximum temperature for future dates.

Data collection

This dataset contains about 10 years of daily weather observations from numerous Australian weather stations. The data was sourced from Kaggle in csv format and is open for public use.

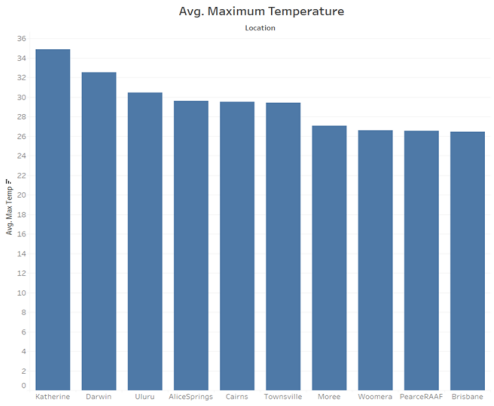
Attribute Information:

Date: The date of observation **| Location:** The common name of the location of the weather station
MinTemp: The minimum temperature in degrees celsius **| MaxTemp:** The maximum temperature in degrees celsius **| Rainfall:** The amount of rainfall recorded for the day in mm
Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am

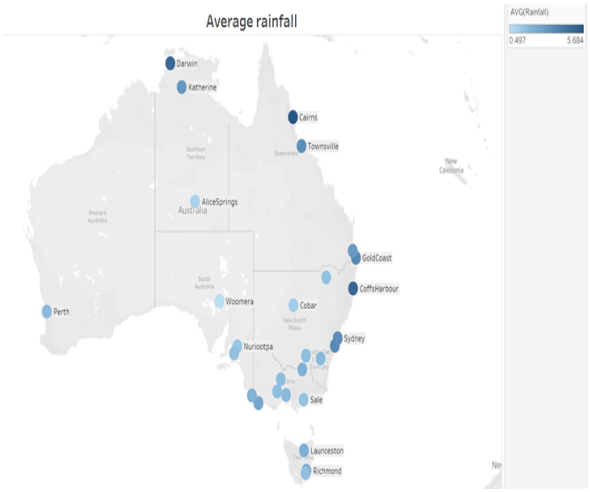
Sunshine: The number of hours of bright sunshine in the day **| WindGustDir:** The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am: Direction of the wind at 9am **| WindDir3pm:** Direction of the wind at 3pm
WindSpeed9am: Speed of the wind at 9am **| WindSpeed3pm:** Speed of the wind at 3pm
Humidity9am: Humidity at 9am **| Humidity3pm:** Humidity at 3pm
Pressure9am: Atmospheric Pressure at 9am **| Pressure3pm:** Atmospheric Pressure at 3pm
Cloud9am: Cloud cover at 9am **| Cloud3pm:** Cloud cover at 3pm
Temp9am: Temperature at 9am **| Temp3pm:** Temperature at 3pm
RainToday: Whether it rained today or not **|RainTomorrow:** Target variable

Exploratory Data Analysis

On our initial data analysis we checked the maximum temperature of different locations across Australia. Overall the average maximum temperature of top 10 locations(fig 1.) remained to be between 25◦C-35◦C. Katherine had the highest maximum temperature and Cairns had the maximum rainfall.



1.



2.

From the data and the plot(2.) we can see most rainfall is received over the east part of Australia. The overall summary for the various columns can be found in figure 3.
The results show the maximum average temperatures reaching 23◦C while the average minimum temperatures to be 13◦C. Overall the average rainfall received over Australia is 2.3mm with the maximum at 371mm.

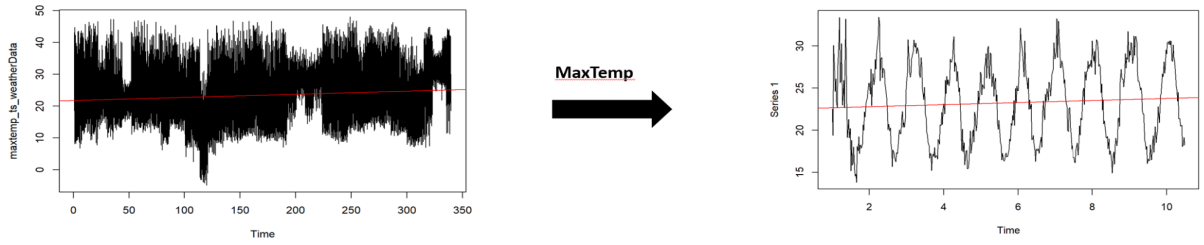
```
     Date              Location           MinTemp          MaxTemp
 Length:145460      Length:145460      Min.   :-8.50   Min.   :-4.80
 Class :character   Class :character   1st Qu.: 7.60   1st Qu.:17.90
 Mode  :character   Mode  :character   Median :12.00   Median :22.60
                                       Mean   :12.19   Mean   :23.22
                                       3rd Qu.:16.90   3rd Qu.:28.20
                                       Max.   :33.90   Max.   :48.10
                                       NA's   :1485    NA's   :1261
    Rainfall        Evaporation        Sunshine      WindGustDir
 Min.   :  0.000   Min.   :  0.00   Min.   : 0.00   Length:145460
 1st Qu.:  0.000   1st Qu.:  2.60   1st Qu.: 4.80   Class :character
 Median :  0.000   Median :  4.80   Median : 8.40   Mode  :character
 Mean   :  2.361   Mean   :  5.47   Mean   : 7.61
 3rd Qu.:  0.800   3rd Qu.:  7.40   3rd Qu.:10.60
 Max.   :371.000   Max.   :145.00   Max.   :14.50
 NA's   :3261      NA's   :62790    NA's   :69835
```

3.

Data Wrangling

For the analysis a lot of data wrangling of the data was required in order to make it compatible for time series analysis. The data has over 10 years of records, so it was decided to first convert the data to a time series with 365.25 frequency. Further each of the columns were interpolated to counter any kind of NA or missing values that it might have.

The results that were obtained for rainfall remained not so easy to visualize. Hence data was aggregating from the daily data to weekly records. Since the aggregation was done the data had to be converted to time series again. Once the data time series analysis was done results were plotted.. Below are the plots(fig 4. & 5.) of the resultant data,

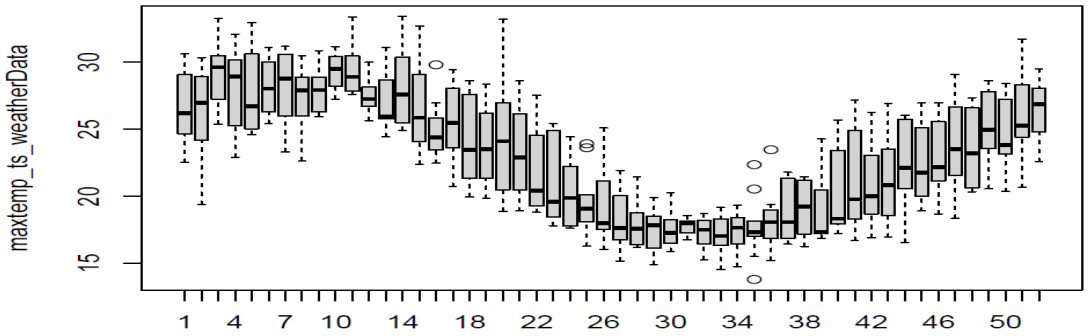


4.

From the results obtained the graph seems to be stationary for Rainfall(fig 5.) however it was required to perform further tests like Augmented Dickey-Fuller test in order to make sure

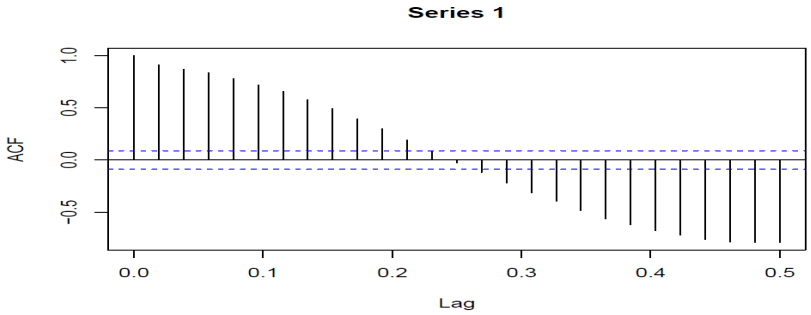
5.

whether the data is stationary or not. Initially the box plot was used inorder to check any outliers for the data. Since the dataset was huge and the outlier numbers came were pretty low hence they were insignificant enough to ignore.



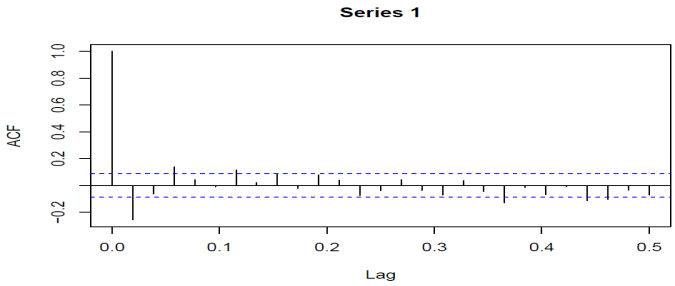
6.

Next ACF and PACF plots were used in order to check whether the elements of a time series are positively correlated, negatively correlated, or independent of each other.



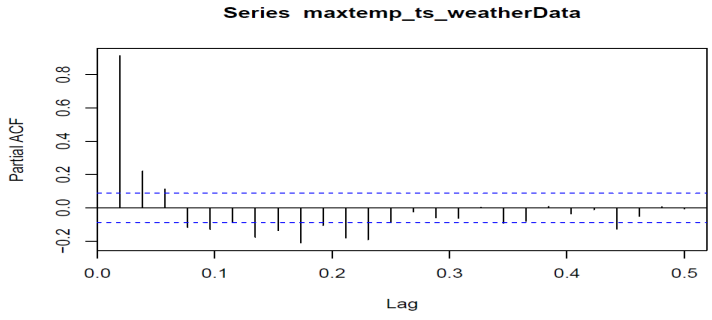
7.

From the above graph(fig 7.) of autocorrelation it can be inferred that the variables are statistically correlated with each other. It can be seen that there is correlation among all lags. This is expected, because there is a trend in our dataset. If lags in acf are taken into account, then the autocorrelation between most lags goes away.



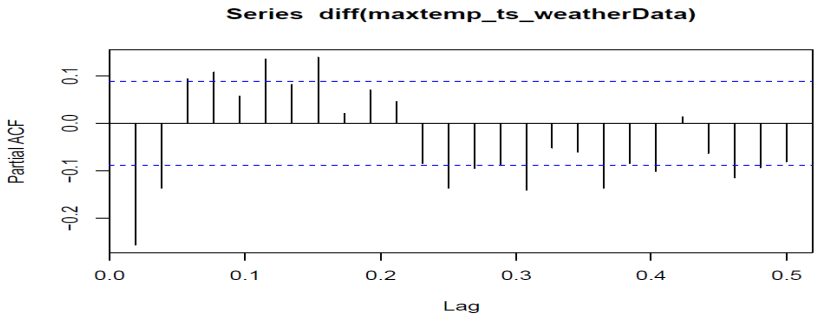
8.

From the above graph of autocorrelation it can be said that now variables are statistically not much correlated with each other. The partial autocorrelation (pacf) removes the correlations that are due to an indirect effect, and just focuses on direct effects.



9.

From the above pacg graph it can be seen that the lags are not much statistically significant except the first lag and hence it can be said that there are not many direct correlations in our data.



10.

From the above plot of pacf the partial autocorrelations for lags 1 and 2 are statistically significant. The subsequent lags are nearly significant and are negatively correlated suggesting an inverse effect from one observation with respect to the next observation.

Next, Augmented Dickey-Fuller test is performed in order to determine whether the data is stationary or not, giving us more insights about any trend in data.
Starting with MaxTemp, below results are obtained from the ADT test. As it can be seen that the result obtained(Fig. 11) for maximum temperature data is not stationary. Hence taking the diff of the data in order to check whether the data was further stationary or not. From the results it can be said the data became stationary and further regression analysis for time series can be performed.

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -0.995   0.323
[2,]   1 -0.737   0.415
[3,]   2 -0.900   0.357
[4,]   3 -0.938   0.344
[5,]   4 -0.959   0.336
[6,]   5 -0.989   0.325
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -4.65  0.0100
[2,]   1 -3.61  0.0100
[3,]   2 -3.14  0.0243
[4,]   3 -3.53  0.0100
[5,]   4 -4.00  0.0100
[6,]   5 -4.31  0.0100
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -4.64  0.0100
[2,]   1 -3.59  0.0333
[3,]   2 -3.15  0.0970
[4,]   3 -3.54  0.0385
[5,]   4 -4.00  0.0100
[6,]   5 -4.31  0.0100
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

11.

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -8.86    0.010
[2,]   1 -5.95    0.010
[3,]   2 -4.07    0.010
[4,]   3 -3.41    0.010
[5,]   4 -2.81    0.010
[6,]   5 -2.42    0.017
Type 2: with drift no trend
     lag     ADF p.value
[1,]   0 -17.95    0.01
[2,]   1 -13.79    0.01
[3,]   2 -10.32    0.01
[4,]   3  -9.11    0.01
[5,]   4  -7.75    0.01
[6,]   5  -7.07    0.01
Type 3: with drift and trend
     lag     ADF p.value
[1,]   0 -17.94    0.01
[2,]   1 -13.79    0.01
[3,]   2 -10.32    0.01
[4,]   3  -9.12    0.01
[5,]   4  -7.76    0.01
[6,]   5  -7.08    0.01
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

12.

For rainfall data the ADF test(Fig 11.) provides us results that the data is stationary,

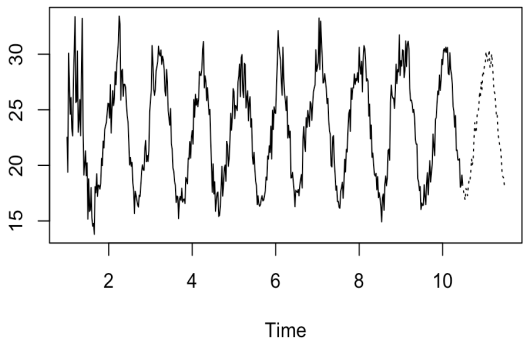
```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -8.86    0.010
[2,]   1 -5.95    0.010
[3,]   2 -4.07    0.010
[4,]   3 -3.41    0.010
[5,]   4 -2.81    0.010
[6,]   5 -2.42    0.017
Type 2: with drift no trend
     lag     ADF p.value
[1,]   0 -17.95    0.01
[2,]   1 -13.79    0.01
[3,]   2 -10.32    0.01
[4,]   3  -9.11    0.01
[5,]   4  -7.75    0.01
[6,]   5  -7.07    0.01
Type 3: with drift and trend
     lag     ADF p.value
[1,]   0 -17.94    0.01
[2,]   1 -13.79    0.01
[3,]   2 -10.32    0.01
[4,]   3  -9.12    0.01
[5,]   4  -7.76    0.01
[6,]   5  -7.08    0.01
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

13.

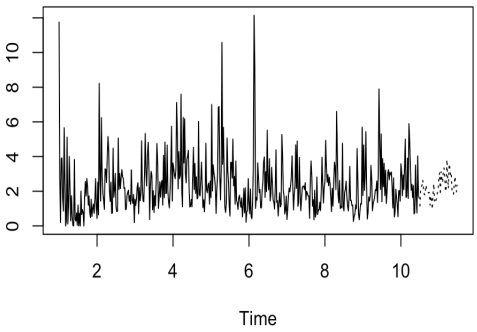
From the above results obtained for rainfall it can be said that there is no drift or trend in the final data that was obtained after doing necessary data wrangling and hence resulting data is stationary. Further the time series analysis and predictions on rainfall data using ARIMA regression models will be done to forecast a 52 week rainfall.

Analysis

After the data wrangling, the variables MaxTemp and Rainfall are used to predict values for the next 52 weeks. The arima() function is used with the order of the AR as 0, the number of times to difference the data as 1, and the order of MA as 1 with seasonality as (0,1,1).



Prediction for MaxTemp



Prediction for Rainfall

Since the data has been aggregated weekly, the number of data points has been drastically reduced and currently there are 494 records. Each attribute has also been converted to a time series class. These values can be used to build the time series regression models. The data has been split to test and train set with 350 records in the test set and 144 records in the test set. While the data has been converted to time series class, they are in individual vectors. Hence, all the vectors are combined using cbind() to build the train set.

The auto.arima() is used to build the regression model with the first 350m records from Rainfall as the predictor and the train set as the regressors. Thus the full model is obtained. The summary and coefficients of the model are given below:

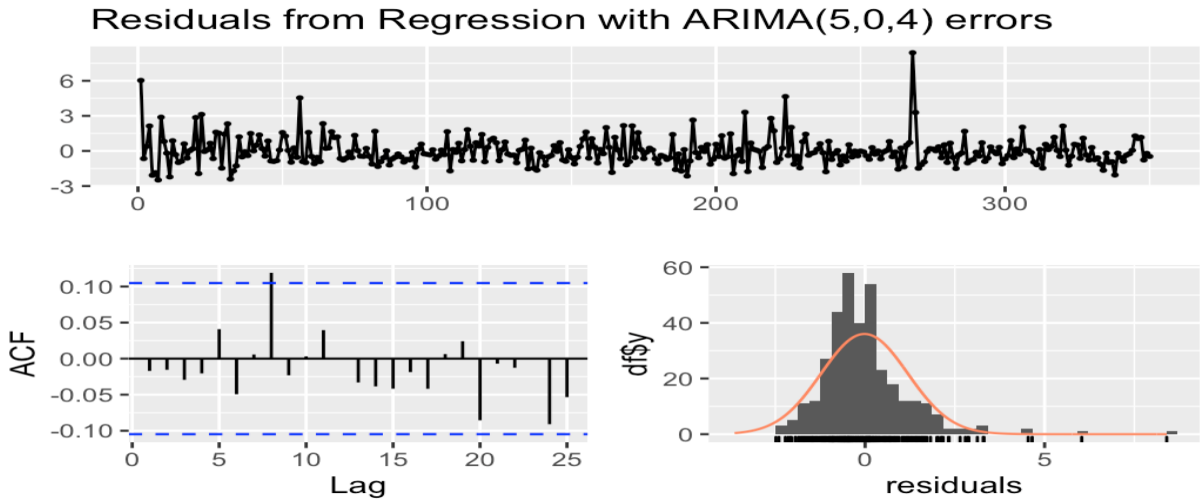
```
Series: rainfall_ts_weatherData[1:350]
Regression with ARIMA(1,0,2) errors

Coefficients:
          ar1     ma1     ma2  Series 1  Series 2  Series 3  Series 4  Series 5
Series 6  Series 7  Series 8  Series 9  Series 10  Series 11  Series 12
      -0.8281  1.0570  0.1522   -0.0108    0.0859   -0.2508    0.2472    0.2235
0.1094   -0.1800    0.0092    0.2377    -0.1639     0.1388     0.1495
s.e.   0.1963  0.2081  0.0949    0.1476    0.3776    0.1272    0.1687    0.0336
0.0590    0.0585    0.0395    0.0487     0.2061     0.2058     0.1555
      Series 13  Series 14  Series 15
         0.2517    -0.3673     0.4913
s.e.     0.1947     0.1976     0.4263

sigma^2 = 1.379:  log likelihood = -543.62
AIC=1125.25   AICc=1127.55   BIC=1198.55

Training set error measures:
                      ME     RMSE       MAE MPE MAPE      MASE         ACF1
Training set -0.00301591 1.143542 0.8007887 NaN  Inf 0.5412771 -0.01453192
```

Summary of full model

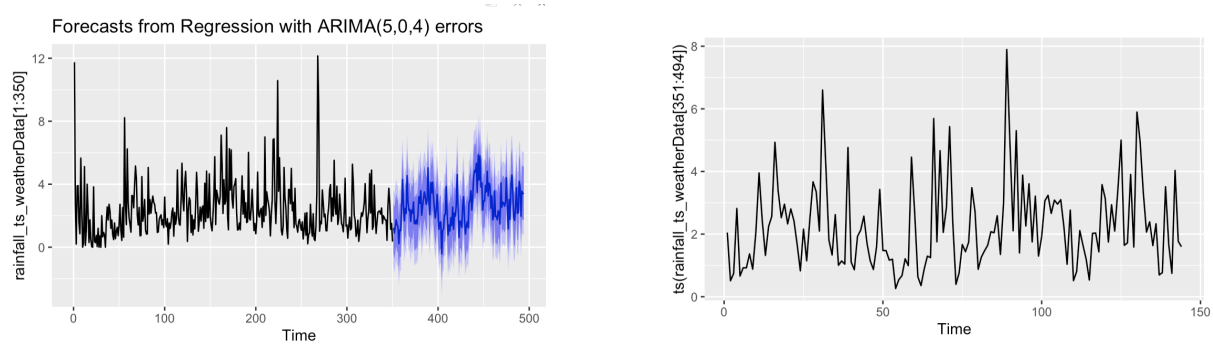
```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1       -0.8281010  0.1963172 -4.2182 2.463e-05 ***
ma1        1.0569846  0.2080610  5.0802 3.771e-07 ***
ma2        0.1522360  0.0948644  1.6048  0.108543
Series 1  -0.0108074  0.1476003 -0.0732  0.941631
Series 2   0.0858911  0.3775565  0.2275  0.820041
Series 3  -0.2507953  0.1271636 -1.9722  0.048584 *
Series 4   0.2471898  0.1686783  1.4655  0.142798
Series 5   0.2234962  0.0335858  6.6545 2.843e-11 ***
Series 6   0.1093976  0.0590279  1.8533  0.063836 .
Series 7  -0.1799914  0.0584602 -3.0789  0.002078 **
Series 8   0.0091778  0.0394537  0.2326  0.816054
Series 9   0.2376978  0.0486509  4.8858 1.030e-06 ***
Series 10 -0.1639480  0.2061385 -0.7953  0.426422
Series 11  0.1388150  0.2058186  0.6745  0.500023
Series 12  0.1494642  0.1555090  0.9611  0.336487
Series 13  0.2516925  0.1947136  1.2926  0.196139
Series 14 -0.3672676  0.1976034 -1.8586  0.063082 .
Series 15  0.4913430  0.4263177  1.1525  0.249104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coefficients of full model

Upon observing the coefficients, many of the variables are insignificant. The variables that have a p-value less than alpha, 0.05, are considered to build a new model. These variables include Evaporation, WindGustSpeed, WindSpeed3pm and Humidity3pm. They are once again combined to form another train set using cbind(). We continue to check the error using the Ljung-Box test. It is observed that the p-value is less than 0.05. This implies that the model does not pass the test but the plots are observed.



The plots obtained meet all the requirements. Hence we proceed to predict the data using the test set. The test set is formed using cbind() with the last 144 records of the significant variables. The ex-poste prediction method is used to forecast. The autoplot() function is used to plot both the predicted values and the actual values.



Predicted values Actual values

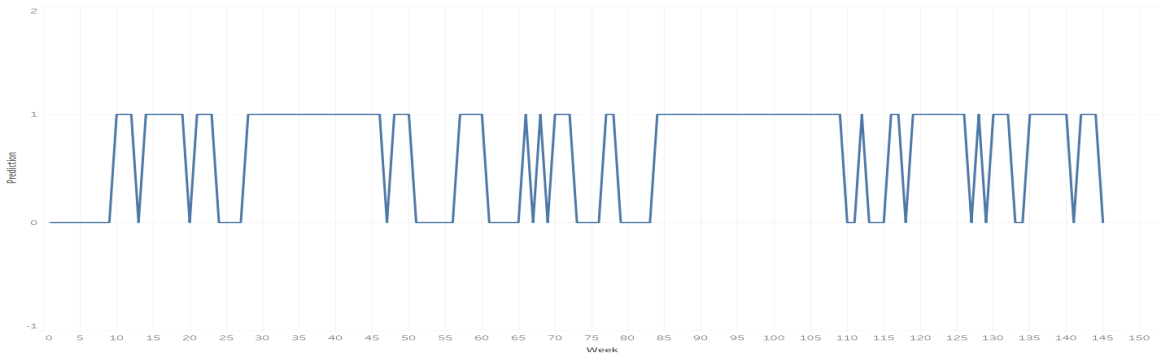
Once the model is built and the values are predicted, the accuracy and RMSE is found using the accuracy() function.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.01810153	1.204343	0.8408880	NaN	Inf	0.5683815	-0.01705439
Test set	-0.35764014	1.117950	0.8708846	-37.32126	55.88482	0.5886571	NA

From the results the RMSE values for the test set are low which concludes that the model performed well. The MPE of -37% gives us an overall accuracy of 63%. However by using further correlations and deep neural networks we can reduce the overall errors.

From the model, the amount of predicted rainfall for the subsequent weeks is obtained. This data can be extrapolated to say if there is a chance of rainfall in a particular week or not. A threshold value of 2 mm is taken. Amount of rainfall obtained is compared to this threshold. If the value is greater than 2mm, it is concluded that there will be rain that week. If not, it is concluded that it will not rain.

The data obtained from prediction is used to build a data frame with the week number and a boolean column that contains values pertaining to having rainfall that week. This data is exported to tableau to obtain the graph below.



Results and Discussion

Starting from a basic weather dataset, the analysis went through an interesting data story involving exploratory analysis and model building. From the predictions on rainfall, the amount of rainfall for each of the subsequent weeks was obtained. This amount of rainfall predicted across the coming weeks can be used to check if there will be floods or not which can be helpful for many businesses that rely on weather for example Coca Cola sales tend to be higher during summers etc. Using the amount of rainfall, the chance of rainfall for the future weeks was also forecasted by setting a threshold of 2mm. This will help government policymakers in Australia to make decisions incase of drought or torrential rainfall.

Arima modeling was used for the analysis, however exploring different models like neural networks and Tensorflow would definitely be a viable option for the future scope of this project. We would recommend using more data for training the model since forecasting is based only on the historical trend, the more accurate prediction must be combined using meteorological data and some expertise from climate experts.

References

1. US EPA, O. (2016). Climate Change Indicators: Heavy Precipitation. [online] US EPA. Available at: https://www.epa.gov/climate-indicators/climate-change-indicators-heavy-precipitation#:~:text=Climate%20change%20can%20affect%20the.

2. kaggle.com. (n.d.). Rain in Australia. [online] Available at: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.