greatlearning

CAPSTONE PROJECT

RAIN PREDICTION IN AUSTRALIA -GROUP 5

- Interim Report

Mentored by:
Vikash Chandra

Submitted by:
T Balaji
K B Charles
D Harun Raj
S Melvin
B Nitin Somasundar
Pranav Hari

Table of Contents

Table of Figure

Industry Review:

Background Research:

In recent years, Australia has experienced numerous floods, storms and bushfires that have had a devastating effect on its properties, businesses and lives. As the climate continues to change, the frequency and severity of these extreme weather conditions is likely to increase as well. Many organizations are already finding it increasingly difficult to respond proactively to these unpredictable disasters. Major natural catastrophes are a heavy burden even to successful economies like Australia. Data from previous years suggests that 97% of disaster funding is spent on post-disaster relief and recovery, with only 3% invested in mitigating a disaster before it happens.

Despite regular loss events, the consequences of flooding are still often underestimated both in Australia and around the world. One reason for this is that almost no one expects the unexpected, even if it happens often, as is the case with this phenomenon. Floods, in particular the flash floods are common in Australia, can happen anywhere, also in extremely dry regions. This element of surprise makes these events even more dangerous than they would otherwise already be. On the other hand, lack of rainfall has severely affected the agricultural industry, which is one of the major contributors to Australia's economy. Dry conditions and lack of rainfall are the major cause for the infamous bushfires. Australian livestock which is sought after across the world for its quality meat is also greatly affected by lack of rainfall. The drought and low rainfall have been a major issue not only in Australia but across the globe.

Need for the study:

Hence, predicting rainfall in the forthcoming years is vital, as it not only allows world governments to formulate policies accordingly in many fields but also helps in taking precautionary measures. It is also necessary for organizations and even individuals to know beforehand to plan accordingly. It is important to exactly determine rainfall for the effective use of water resources, crop productivity and preplanning of water structures.

Current practices:

- Presently, weather is predicted using various techniques and instruments. A huge amount of data has to be processed and patterns are found and then compared to historical data to predict phenomena such as rainfall and calamities.
- There are various methods used in predicting rainfall, each used by various organizations in predicting weather and climate in the short term and long term respectively.
- The current method used by the Australian Bureau of Meteorology is called the Earth system models (ESM) which have been in development for around 10 years. The Australian version of the ESM is called ACCESS(Australian Community Climate and Earth System Simulator). This method is used for the daily weather predictions and the climatic conditions for a foreseeable future.

Literature Survey - Publications, Application, past and undergoing research:

Rainfall in Queensland Part 1: A literature survey of key rainfall drivers in Queensland, Australia: Rainfall variability and change Prepared by Nicholas Klingaman February 2012

Application:

Predicting rainfall has always been a very difficult science in itself to master. For thousands of years, humans have been trying to predict the rainfall for different purposes. In a country like Australia, where agriculture as an industry is a mass employer and contributes a substantial amount to the country's GDP inaccurate predictions of rains can have huge ramifications economically and socially. Occasionally a dust storm will blanket a region and there are reports of the occasional tornado. Tropical cyclones, heat waves, bushfires and frosts in the country are also associated with the Southern Oscillation. Furthermore, the country also suffers from bushfires which cause a massive loss to life and property. These bushfires are linked to a scarcity of rains. Climate change in Australia is a highly contentious political issue. Temperatures in the country rose by approximately 0.7 °C between 1910 and 2004, following an increasing trend of global warming. Accurate predictions will enable people and emergency services to be better prepared for future disasters.

Past and undergoing research:

Queensland's climate experiences considerable natural inter-annual and decadal variability in its rainfall. To determine how Queensland's rainfall is going to change in the coming decades as the planet warms, it is critical to establish which global and regional climate phenomena are driving this variability. Understanding the potential impacts of climate change is essential to inform strategies and actions to avoid or manage dangerous levels of change. The impacts of these changes are important especially for those sectors that are vulnerable to changes in rainfall, such as water-resource management and agriculture. In 2007, the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC, 2007) confirmed that there is currently substantial uncertainty in rainfall projections for the Australian region for the coming century. This reinforces the urgent need to reduce that uncertainty.

Improving the current understanding of key processes and phenomena that influence Queensland rainfall on timescales from days to decades will help to address that uncertainty, as there may be greater confidence in the impacts of climate change on these phenomena than for rainfall. A reduced uncertainty in rainfall changes would support effective planning to address potential changes in the hydrological cycle at the regional and local level. Part of the uncertainty in future rainfall changes arises from the various competing and interacting

influences on Queensland rainfall, which are associated with synoptic and climate drivers across various time scales, such as tropical cyclones, the Madden-Julian Oscillation (MJO), the El-Niño Southern Oscillation (ENSO) and the Interdecadal Pacific Oscillation (IPO). The MJO controls the sub-seasonal variations in the summer monsoon rainfall with a period of 30-60 days.

It primarily affects northern Queensland and modulates the monsoon and trade-wind circulations. Active periods of the MJO also increase the probability of tropical cyclone formation in the Coral Sea.

Data Dictionary and Pre-processing Data Analysis:

Data Attribute Details

RangeIndex: 145460 entries, 0 to 145459

Data columns (total 23 columns):

Sr.No.	Variables Names	Categorization of Variable	Null values Check
1	Date	Categorical	1,45,460 non-null object
2	Location	Categorical	1,45,460 non-null object
3	MinTemp	Numerical /Discrete	1,43,975 non-null float64
4	MaxTemp	Numerical /Discrete	1,44,199 non-null float64
5	Rainfall	Numerical /Discrete	1,42,199 non-null float64
6	Evaporation	Numerical /Discrete	82,670 non-null float64
7	Sunshine	Numerical /Discrete	75,625 non-null float64
8	WindGustDir	Categorical	1,35,134 non-null object
9	WindGustSpeed	Numerical /Discrete	1,35,197 non-null float64
10	WindDir9am	Categorical	1,35,197 non-null object
11	WindDir3pm	Categorical	1,41,232 non-null object
12	WindSpeed9am	Numerical /Discrete	1,43,693 non-null float64
13	WindSpeed3pm	Numerical /Discrete	1,42,398 non-null float64
14	Humidity9am	Numerical /Discrete	1,42,806 non-null float64
15	Humidity3pm	Numerical /Discrete	1,40,953 non-null float64

16	Pressure9am	Numerical /Discrete	1,30,395 non-null float64
17	Pressure3pm	Numerical /Discrete	1,30,432 non-null float64
18	Cloud9am	Numerical /Discrete	89,572 non-null float64
19	Cloud3pm	Numerical /Discrete	86,102 non-null float64
20	Temp9am	Numerical /Discrete	1,43,693 non-null float64
21	Temp3pm	Numerical /Discrete	1,41,851 non-null float64
22	RainToday	Categorical	1,42,199 non-null object
23	RainTomorrow	Categorical	1,42,193 non-null object

Dtypes: float64(16), object (7)

Project Justification:

- This is a data set for predicting whether there will be rains in a town in Australia on a given day.
- This is a classification problem. The dependent variable is Rain Tomorrow.
- We can use Classification model algorithms like Logistic Regression, K-NN, Decision Tree, Random Forest, etc.,
- We can use bagging and boosting techniques for increasing the accuracy and performance of the model.

Data Exploration (EDA):

Relationship between the variables:

There are a total of 23 features among which RainTomorrow is the Target variable. Let us see the various visual analytics as follows:

Univariate Analysis:

Rain Tomorrow:

This feature describes the number of days where it rained in a given location during the next day.

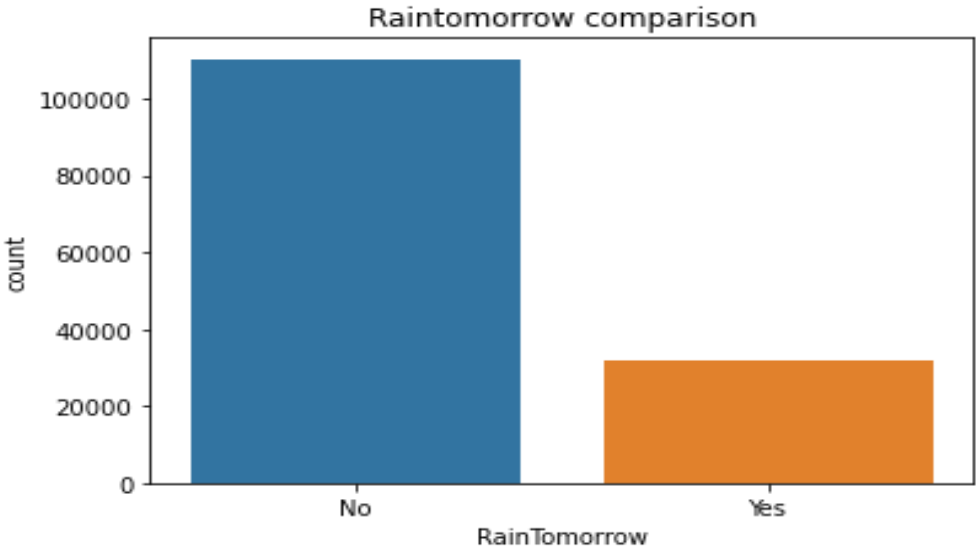

Figure 1: Distribution of Rain Tomorrow

- It is found that there is a 22.5% of raining and 77.5% chance of raining the following day.
- It is seen that there were more days where it has not rained than it has rained

Rain Today:

This feature describes the number of days where it rained in a given location during the present day.

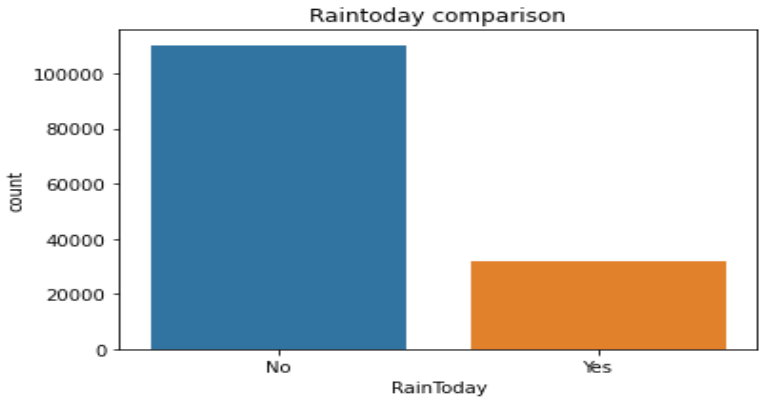

Figure 2: Distribution of Rain Today

- It is found that there is a 22.5% of raining and 77.5% chance of raining the on the present day.
- It is seen that there were more days where it has not rained than it has rained.

Minimum Temperature:

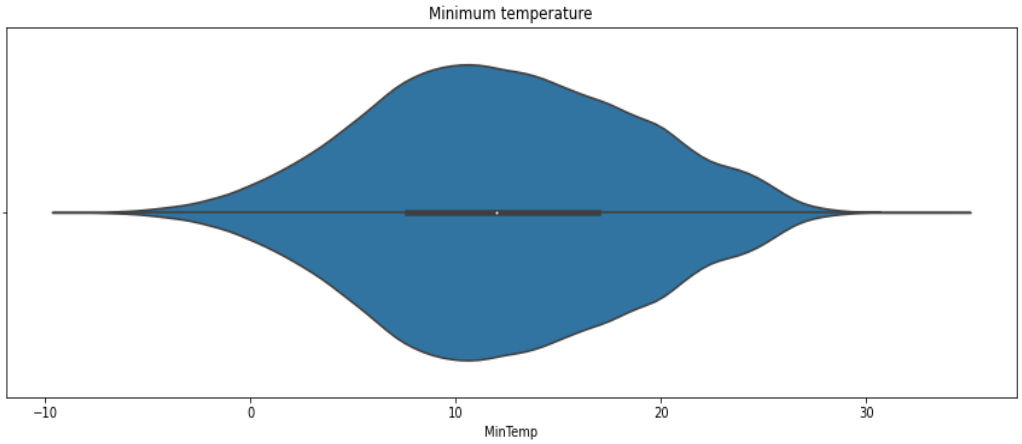

Figure 3: Minimum Temperature

- Minimum temperature seems normally distributed and outliers present on the variable and liesbetween 8 to 18 mean and 50th percentile are equal therefore the minimum temperature variable is normally distributed.
- In a minimum temperate variable, the maximum point is 33.8.

Maximum Temperature:

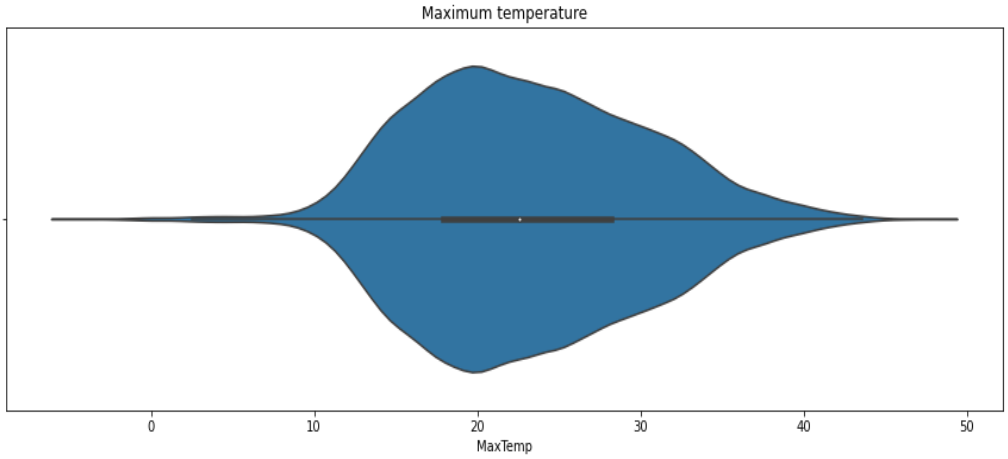

Figure 4 : Maximum Temperature

- Maximum temperature seems normally distributed and outliers present on the variable and lies between 18 to 28 mean and 50th percentile are equal so maximum temperature is also normally distributed.
- In the maximum temperature variable the maximum point is 48 in a whole data set.

Rainfall Amount:

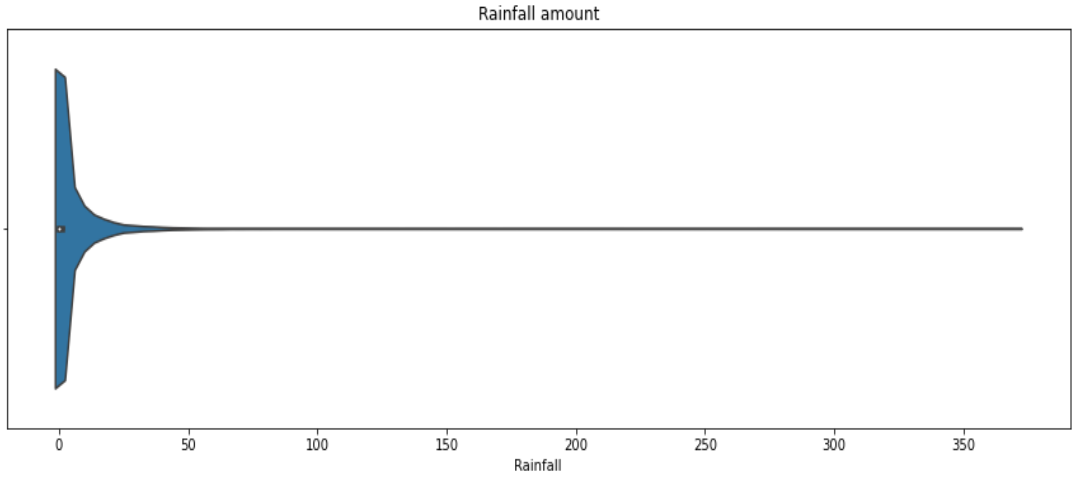

Figure 5: Rainfall Amount

- It is found that the data is right skewed in its distribution.
- Total rainfall falls between the range from 0 to 1 percent from the overall values.

Bivariate Analysis:

Distribution of Rainfall:

This feature describes the distribution of the rainfall which is compared to the categorical variable Rain Tomorrow.

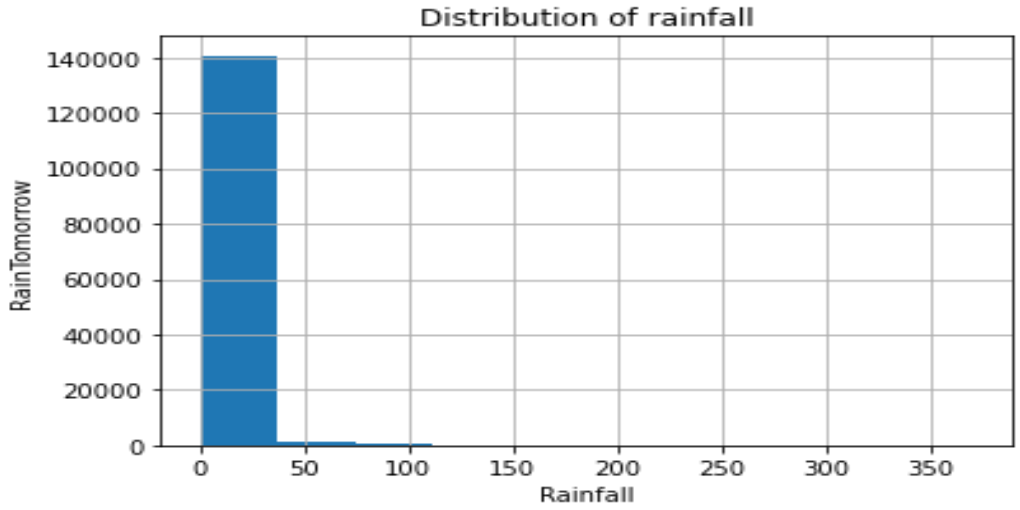

Figure 6: Rainfall vs Rain Tomorrow

- The rainfall is positively distributed on the Rain Tomorrow feature.

Distribution of Rain Tomorrow on Evaporation:

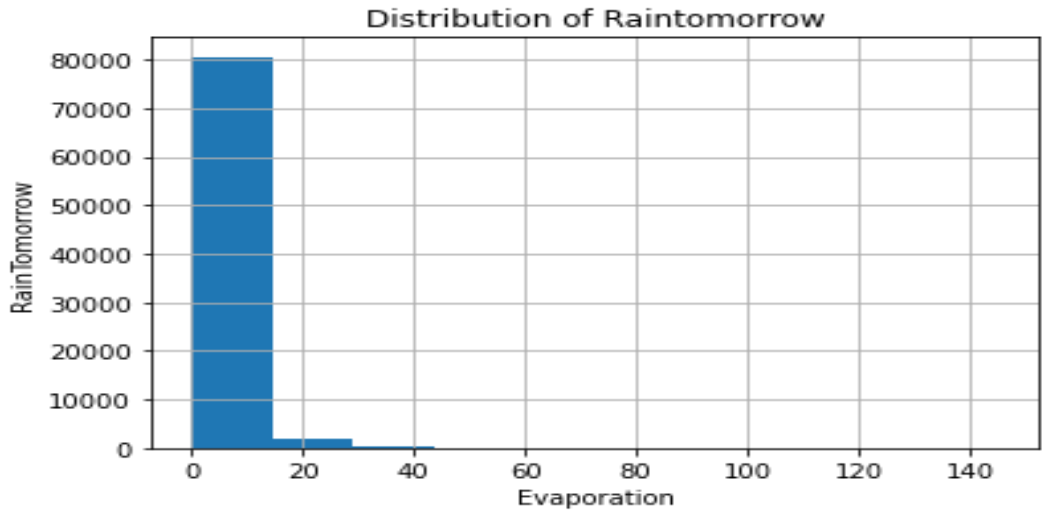

Figure 7: Rain Tomorrow vs Evaporation

This feature describes the distribution of the evaporation which is compared to the categorical variable Rain Tomorrow.

Rain Tomorrow vs Wind Speed 9am :

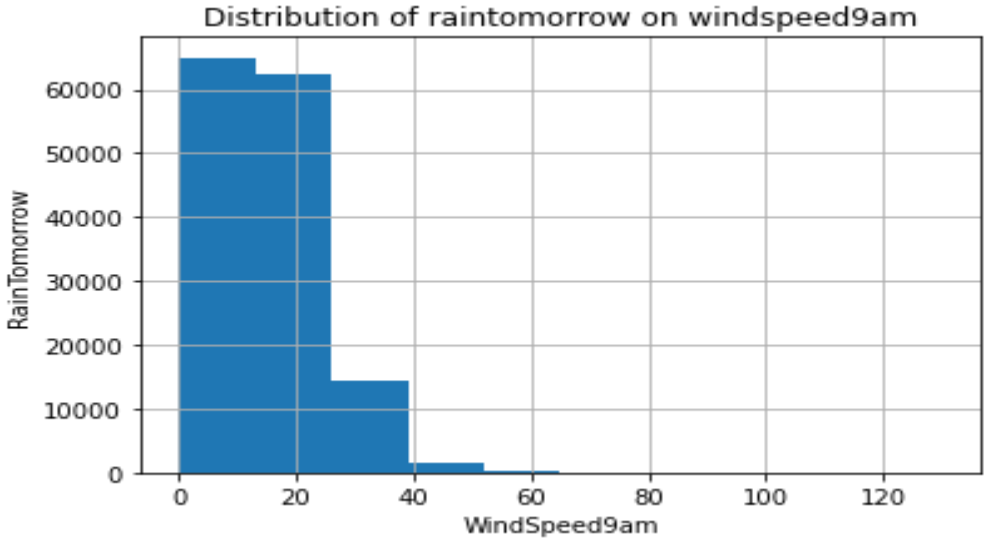

Figure 8: Rain Tomorrow vs Wind Speed 9 am

- It is seen that the wind speeds at 9 am were lower than 50 units.
- Therefore, overall rain tomorrow has positively skewed on the evaporation.

Rain Tomorrow vs Wind Speed3pm:

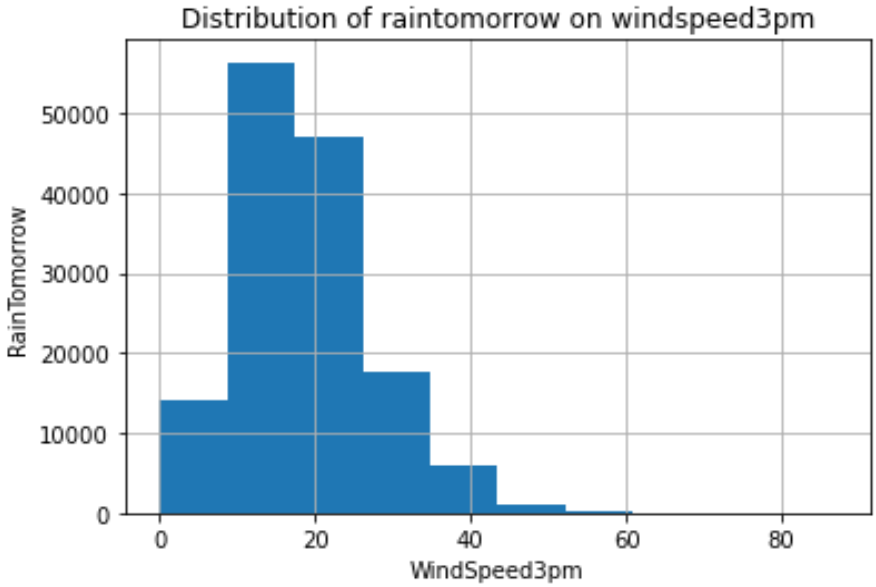

Figure 9: Rain Tomorrow vs Wind Speed 3pm

- It is seen that the wind speeds at 3 pm were lower than 50 units.
- Therefore, overall rain tomorrow has positively skewed on the evaporation.
- The overall rain tomorrow falls between 15 to 35 km speed.

Sunshine vs Evaporation:

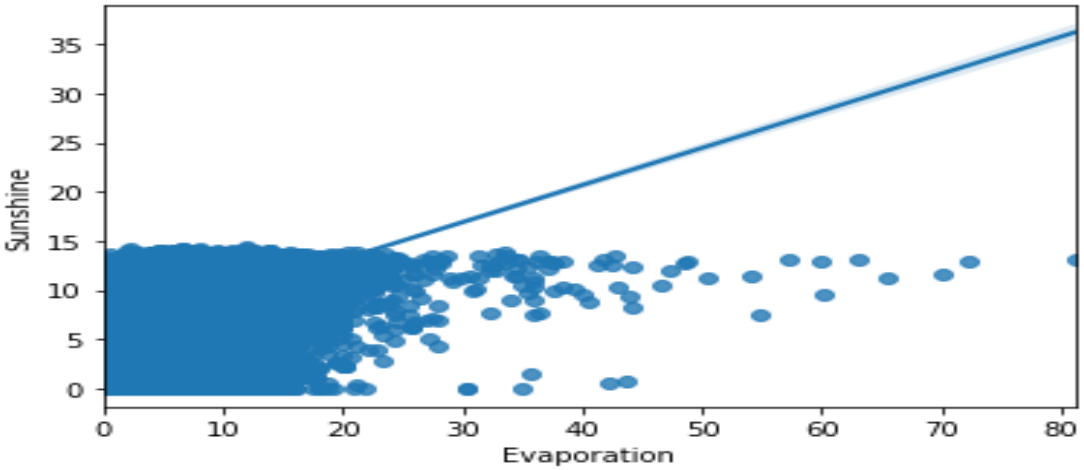

Figure 10: Sunshine vs Evaporation

- It is clearly shown that the evaporation and sunshine have a positive relationship
- When sunshine increases the evaporation also increases.
- It shows a positive trend.
- Major sunshine and evaporation points falls on 0 to 15, The sunshine and evaporation between 0 to 35.

MinTemp vs Evaporation:

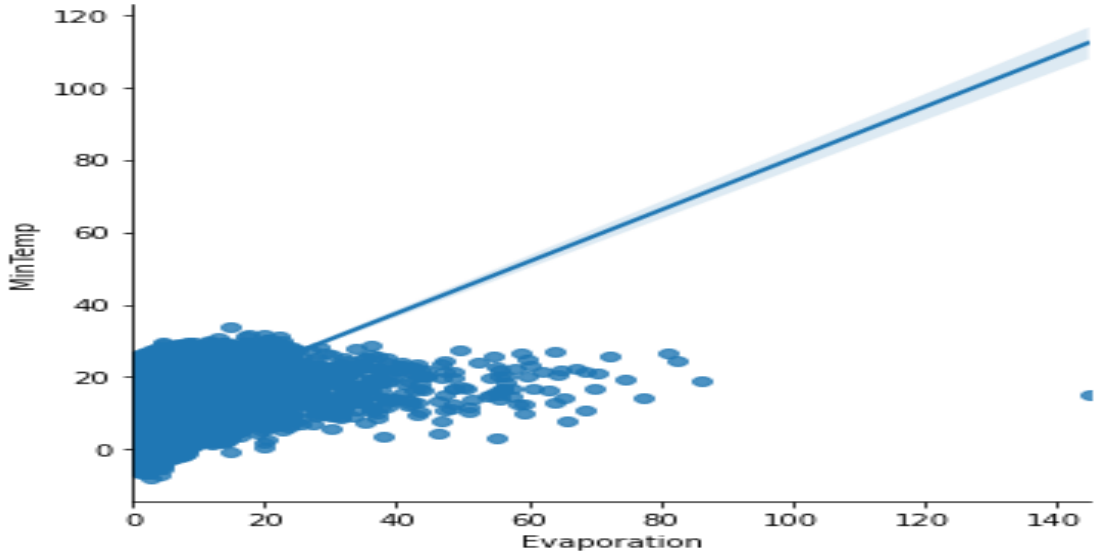

Figure 11: MinTemp vs Evaporation

- It is clearly shown that the evaporation and Mintemp have a positive relationship.
- When Mintemp increases the evaporation also increases.
- It shows a positive trend.
- Major Mintemp and evaporation points falls between 0 to 25, The evaporation and Mintemp between 0 to 55.

Rainfall vs Year:

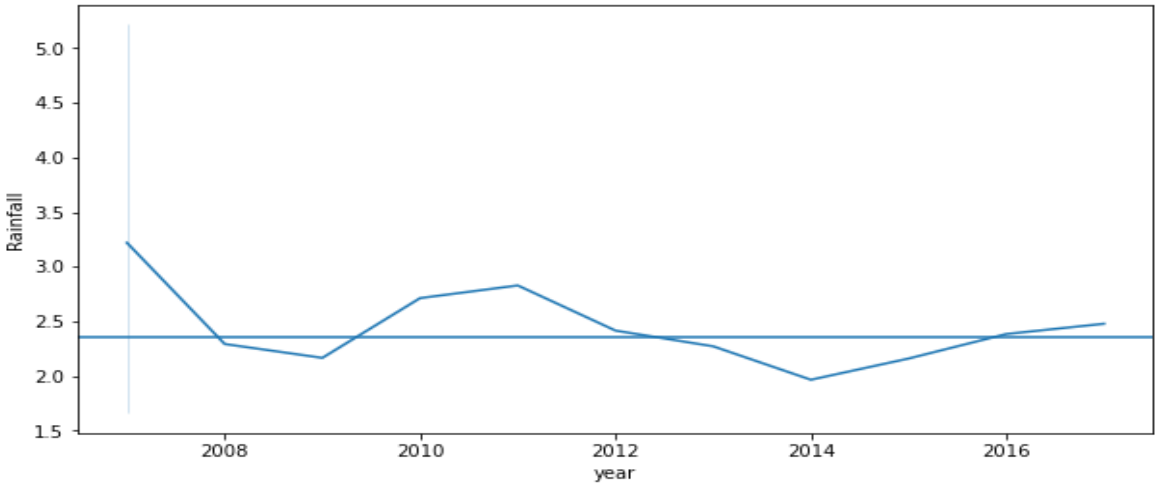

Figure 12: Rainfall Vs Year

- It clearly shows that the rainfall and year have up and down relationship.
- In the year 2007 and 2011 has the highest amount of rainfall.
- During 2014 and 2007 has the lowest amount of rainfall during these years.
- While average amount of rainfalls attained in the years 2008 ,2009,2012 and 2016 had an average amount of rainfall.

Rainfall vs Month:

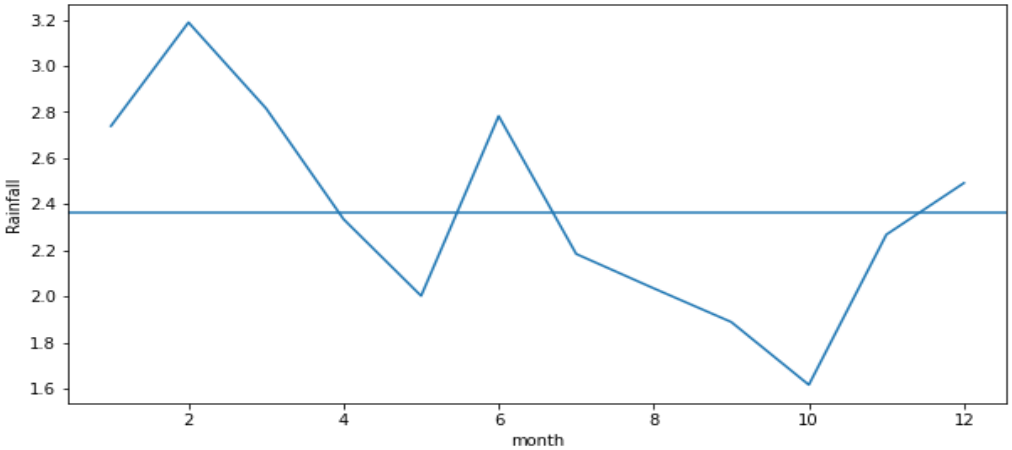

Figure 13: Rainfall vs Month

- It clearly shows that the rainfall and month have up and down relationship.
- In the month of 2 and 6 has a highest number of rainfalls.
- During 10th and 5th month it has the lowest number of rainfalls during this month.
- While average number of rainfalls attain on 4,5,7 and 11th months had an average number of rainfalls.

Rainfall vs Location:

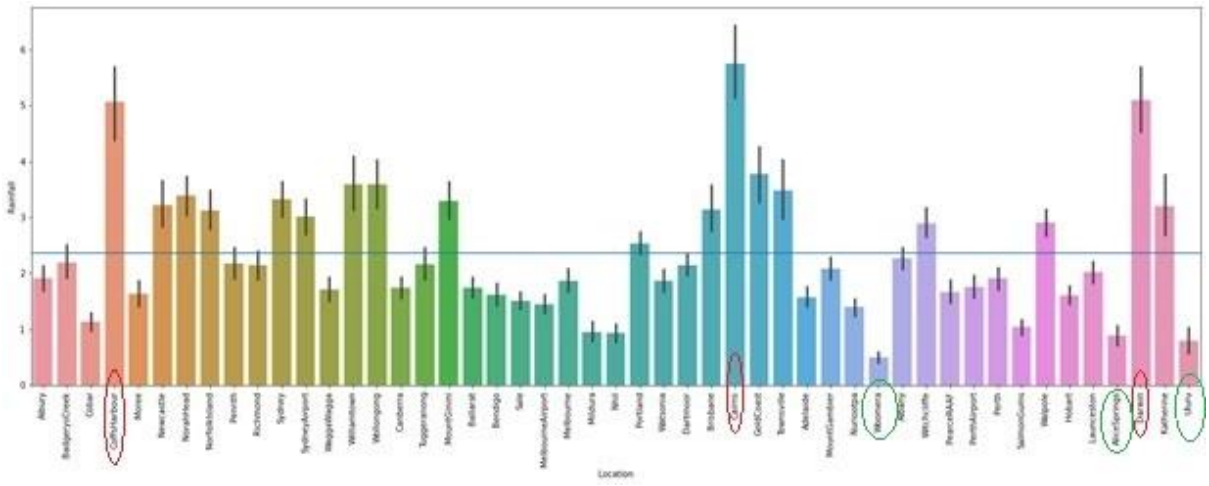

Figure 14: Rainfall vs Location

- The highest rainfall for the top three cities is Caims followed by Darwin and Caffsharbour, which are circled in red.
- The lowest rainfall for the top three cities is Woomera followed by Ulluru and Alicesprings. There are 18 cities have more than average rainfall in the Australia country, which are circled in green.

Wind Gust Direction vs Cloud 9 am:

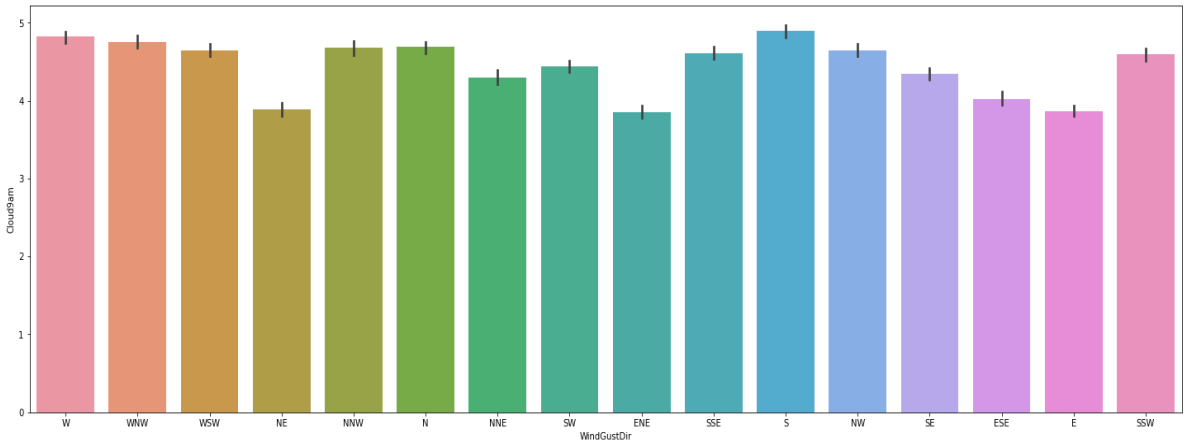

Figure 15: WindGusDir vs Cloud9am

- The South (S) direction has a highest cloud forming during morning timeline followed by west direction and west-northwest (WNW).
- The East-northeast (ENE) direction and North east (NE) direction have a lowest cloud forming during morning timeline.

Wind Gust Direction vs Cloud 3 pm:

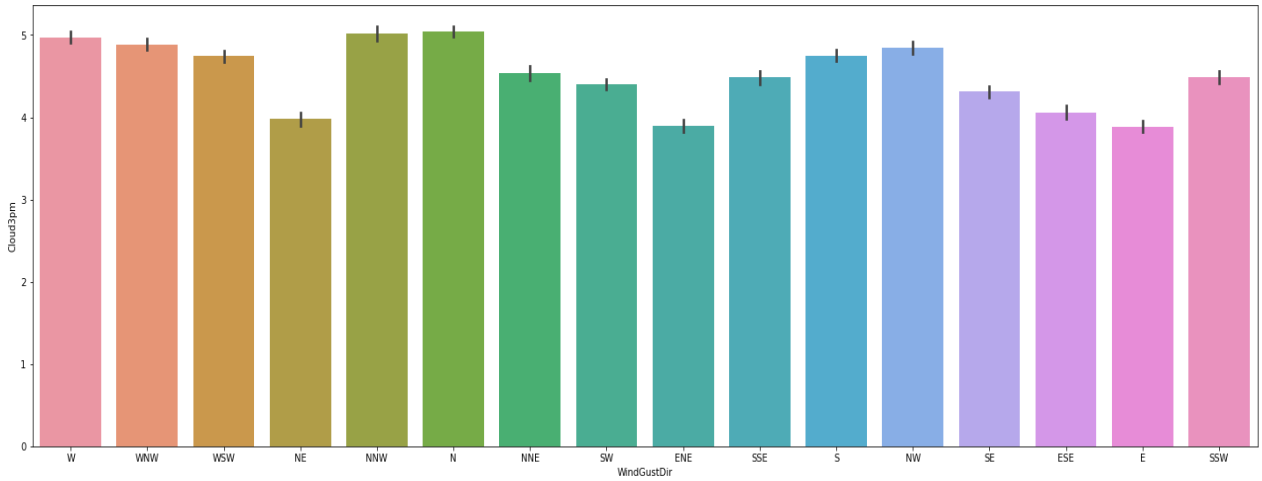

Figure 16: WindGusDir vs Cloud3pm

- The North (N) direction and north-northwest (NNW) direction have the highest cloud forming during evening timeline followed by west direction and north west direction.
- The East-northeast (ENE) direction and North east (NE) direction have the lowest cloud forming during evening timeline as same as morning timeline.

Wind Direction 3pm vs Cloud 3pm:

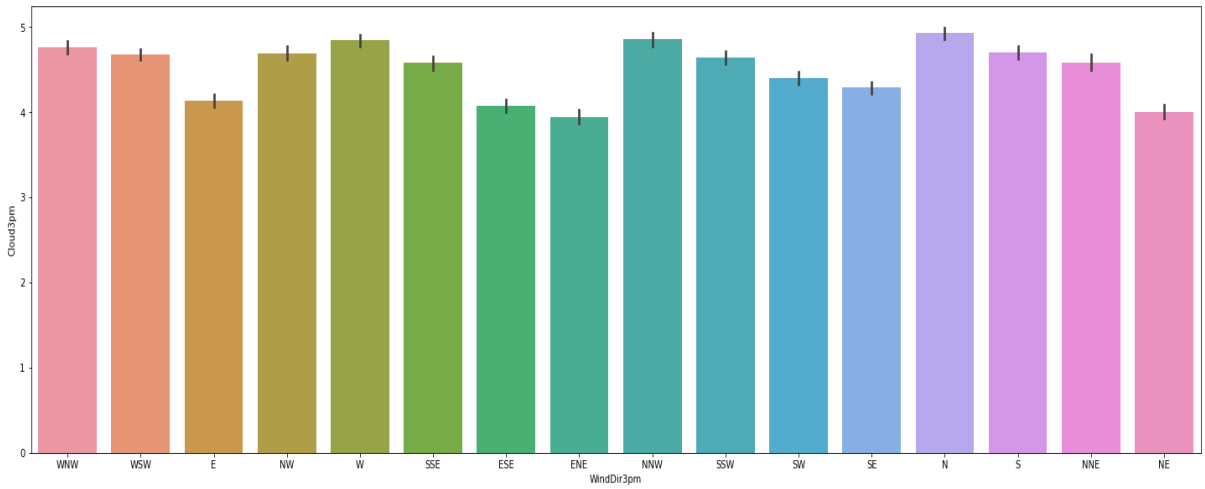

Figure 17: WindDir3pm vs Cloud3pm

- The North (N) direction and north-northwest (NNW) direction have a highest cloud forming during evening timeline followed by west direction and north west (NW) direction.
- The East-northeast (ENE) direction and North east (NE) direction have a lowest cloud forming during evening timeline as same as morning timeline.

WindGustDirection vs Rain Tomorrow:

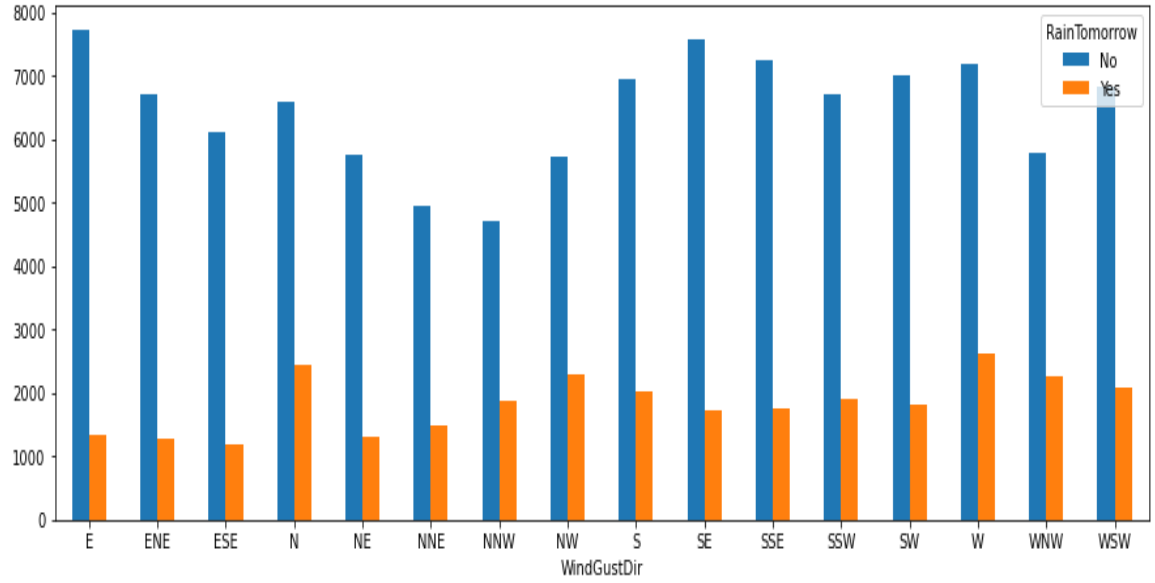

Figure 18: Wind Gust Direction vs Rain Tomorrow

- During East (E) direction and Southeast (SE) direction has a highest Number of No value counts present Which means These two directions has no rain tomorrow.
- We can expect rain tomorrow in west direction followed by North and North west (NW) direction.

Wind Gust Direction vs Rain Today:

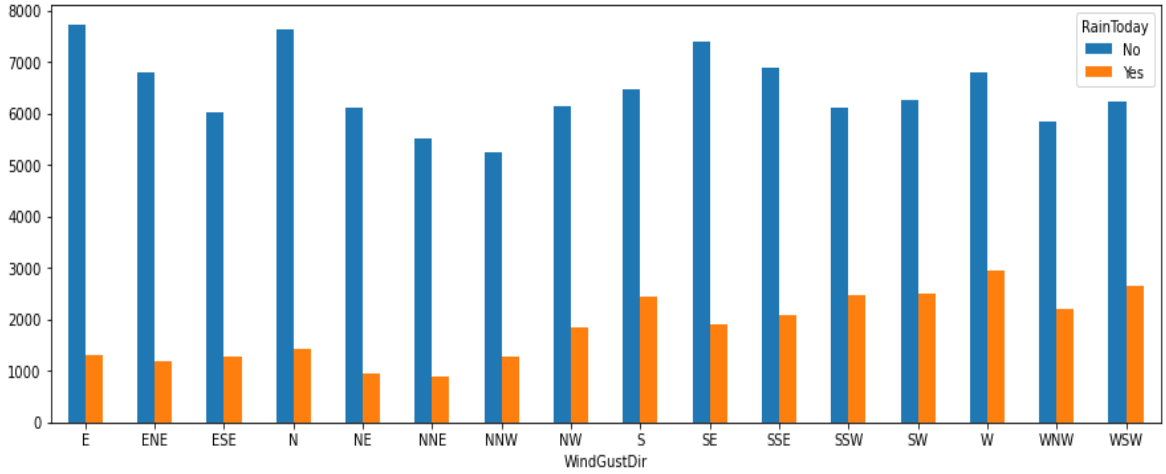

Figure 19: Wind Gust Direction vs Rain Today

- We can expect No rain on today during East (E) direction and North (N) direction followed by South east (SE) direction.
- As Opposite to the No rain today we can expect high RainToday on West direction followed by West south-west (WSW) direction and South direction.

Multivariate Analysis:

Month vs Mintemp and Month vs Maxtemp(RainTomorrow):

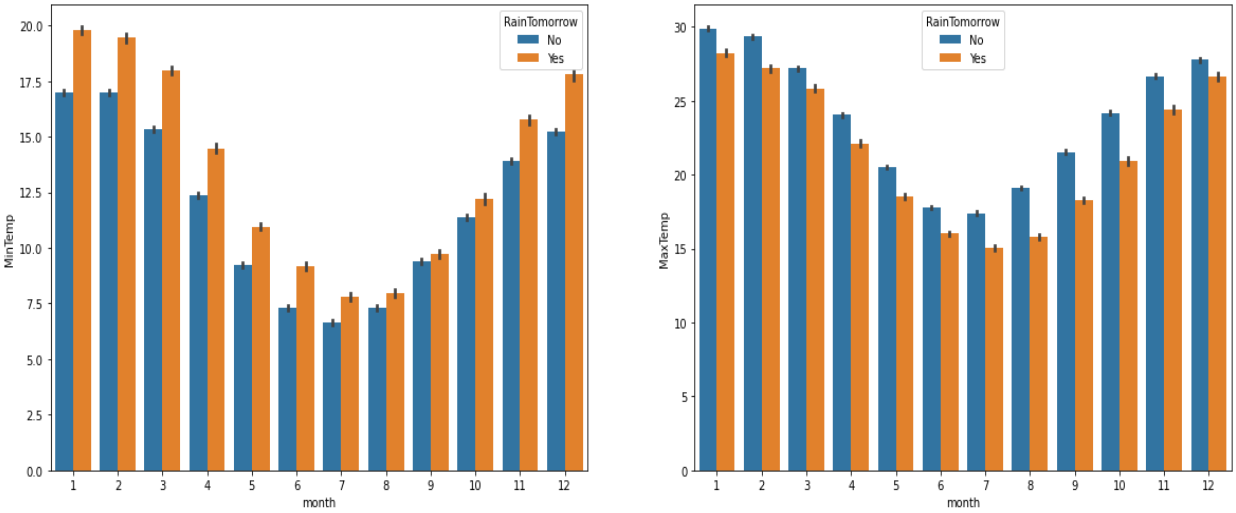

Figure 20: Month vs Mintemp and Month vs Maxtemp (RainTomorrow)

- In the month of January, February and March, December have a high chance of Raintomorrow during minimum temperature and low chances of No rain tomorrow during the month of July, June and august.
- In the month of January, February and March, December have a high chance of No Rain Tomorrow during maximum temperature and low chances of rain tomorrow during the month of July, June and august.

Month vs Mintemp and Month vs Maxtemp Inferences on Today Rain:

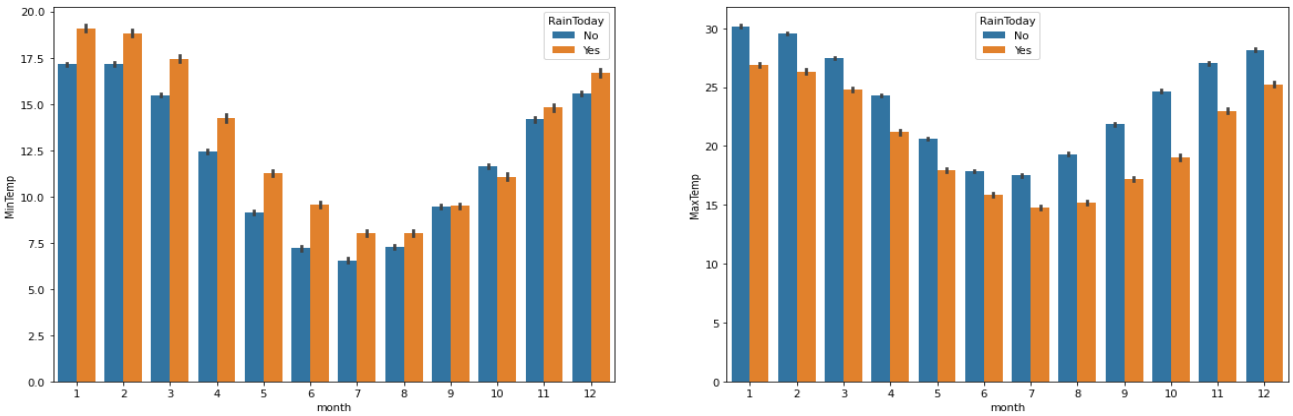

Figure 21: Month vs Mintemp and Month vs Maxtemp (RainToday)

- In the month of January, February and March, December has a high chance of Raintoday during minimum temperature and low chances of No rain tomorrow during the month of July, June and August.
- In the month of January, February and March, December have a high chance of No Rain Today during maximum temperature and lowest chances of rain tomorrow during the month of July,June and August

Year vs Cloud3pm (hue=RainToday) and Year vs Cloud3pm (hue=RainTomorrow):

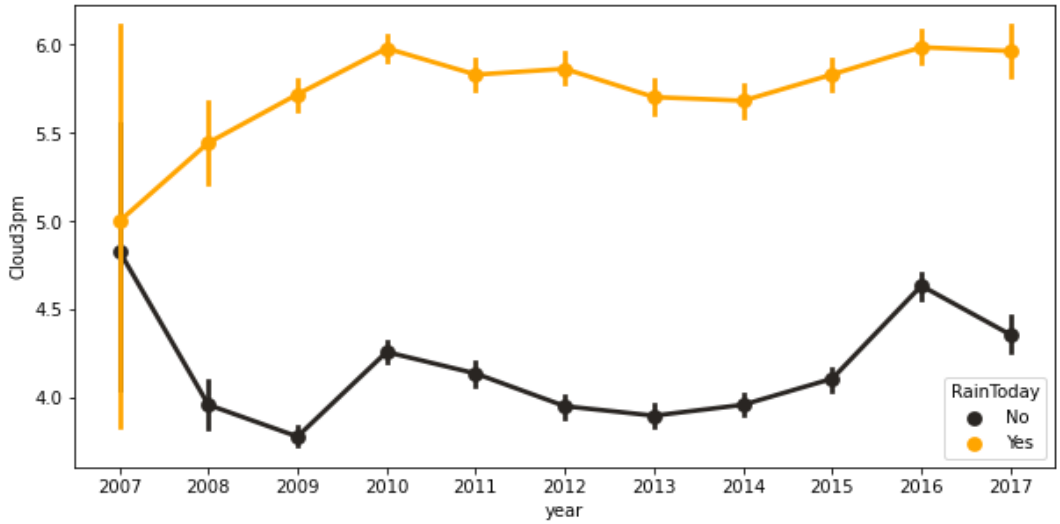

Figure 22: Year vs Cloud 3pm (hue=RainToday)

- There is a gradual rise in year 2007 to 2017 contributing rains with cloud3pm values over 5.0 has a higher chance of raining today.
- There is a slow drop in year from 2007 to 2017 with cloud3pm values less than 5.0 has no chance of raining today.

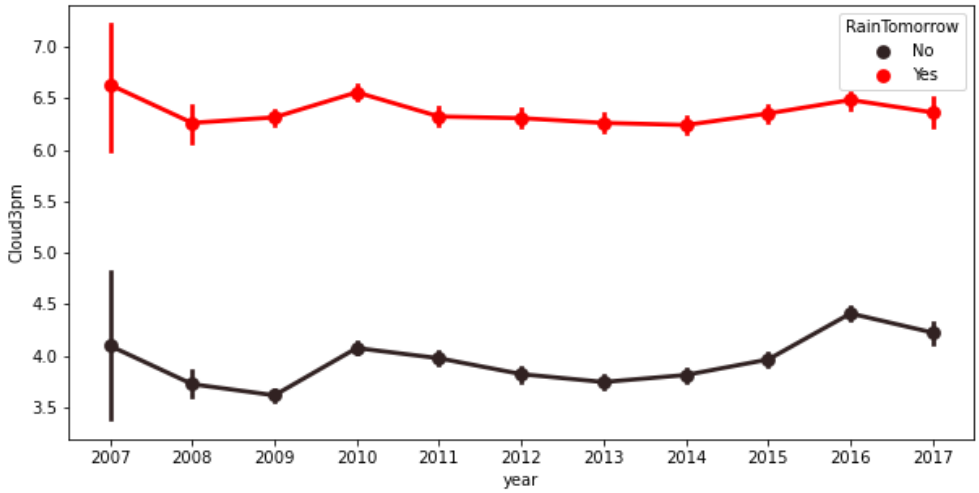

Figure 23: Year vs Cloud 3pm (hue=RainTomorrow)

- There is a slight decrease with steady movement in year 2007 to 2017 which has higher chance of rain with Cloud3pm values between 6.0 to 7.0.
- There is a slight decrease with steady movement in year 2007 to 2017 which has No chance of rain with Cloud3pm values between 3.5 to 4.5.

Sunshine vs Rainfall:

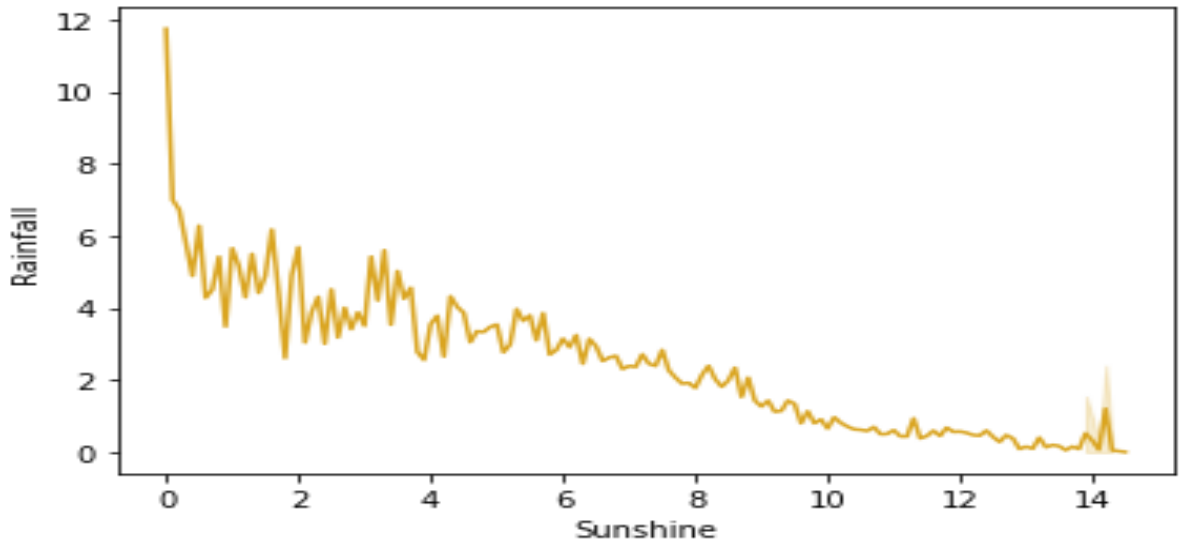

Figure 24: Sunshine vs Rainfall

- There is a steady drop from 0 to 14 in the X-axis (Sunshine).and gradual rise from 0 to 12 in Y-axis (Rainfall).
- When Sunshine decreases rainfall gradually increases this clearly shows that rainfall and sunshine are inversely proportional to each other.

Feature engineering:

There are some irrelevant Features which are dropped. For example,'WindGustDir','WindDir9am','WindDir3pm' has the unique and same value for 3 variables for all the record. Therefore they has been dropped. 'Wind Gust Dir' has components Like 'W', 'WNW', 'WSW', 'NE', 'NNW', 'N', 'NNE', 'SW', 'ENE', 'SSE', 'S', 'NW', 'SE', 'ESE', 'E', 'SSW'. These components also there in 'WindDir9am' and 'WindDir3pm'. "Location" feature has unique of 49 different names which is not used for the next model building hence this feature is dropped.

Correlation of the Dataset:

The correlation of the full dataset is as below:

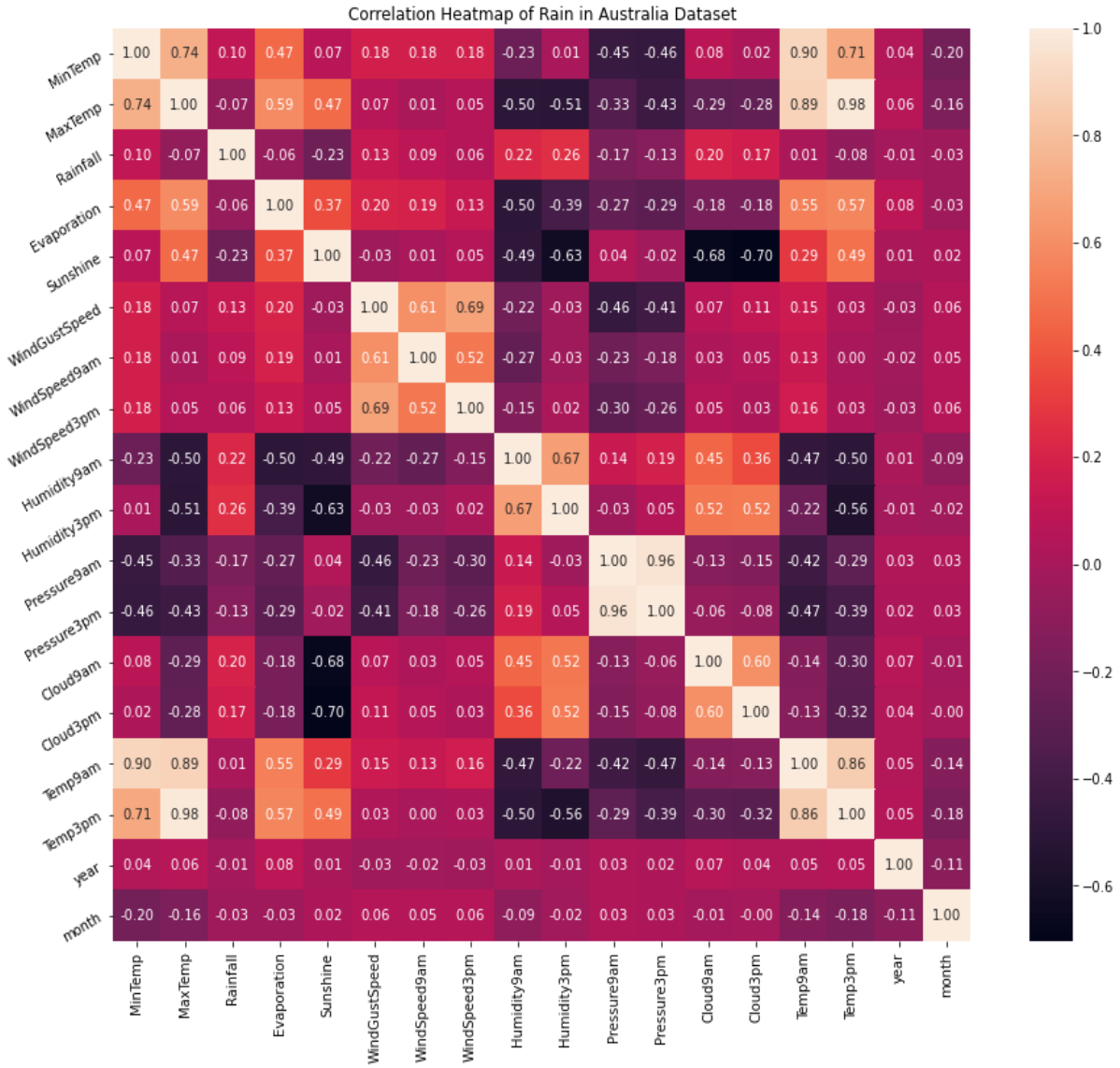

Figure 25: Correlation matrix heatmap.

- The above correlation plot tells that maxtemp and mintemp have a high correlated compared to other features.
- Temp 3pm and temp 9am features are also highly correlated with each other.
- Pressure 9am and pressure 3pm also highly correlated with each other So They can't contribute to model accuracy instead they are duplicate values.
- Correlated with each other variables would have been remove because they can't be involved in Contribution to the model.
- The final conclusion mintemp, maxtemp, temp9am and temp3pm are highly correlated with each variable.

Checking of outliers:

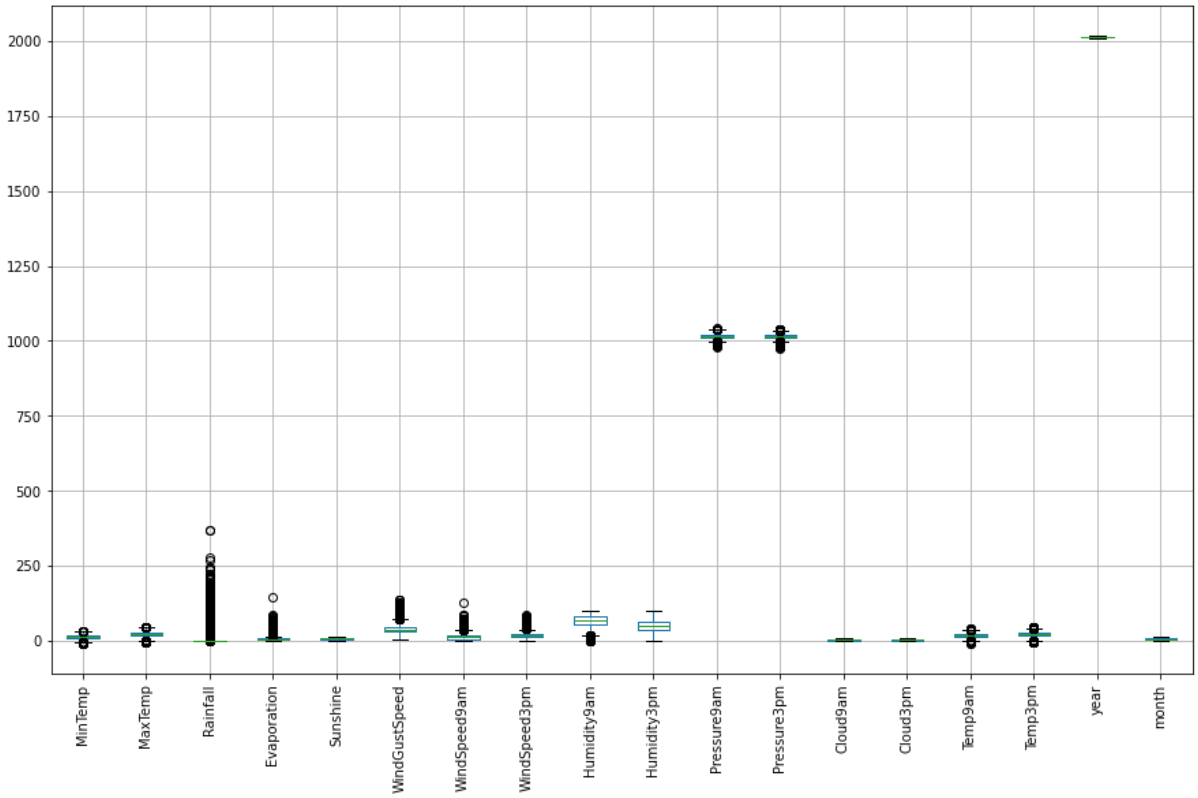

Figure 26: Checking outliers

- The rainfall feature has a high number of outliers detect followed by Windgustspeed and windspeed9am
- There are no outliers detect on the cloud 9 am,cloud 3pm, humidity, month and year.

Future Tasks:

1. After Data analyzing, Cleaning, Missing values and Outliers treatment we proceed to start building the machine learning model i.e. start with base model called as Logistic Regression.
2. After Built on Logistic Regression model we would like to try several classification models
3. Such as Decision Tree, Random Forest, Naïve bayes.
4. We would like to Use hyperparameter tuning to the model to increase the model accuracy.
5. After the above particular tasks done we use ensemble technique to the models to select which model is giving maximum accuracy to the target variable.