Rainfall prediction-Weather Forecasting

Problem statement:

In the present scenario where the dynamicity is at its utmost level, global climate change is something that can not remain untouched by it. These sudden shifts in the climate often increase the degree of uncertainty in rainfall/precipitation which tends to impact other components of the terrestrial water cycle. Some of the ways that uncertainty in rainfall and precipitation can affect us include:

1. Agricultural production: Uncertainty in rainfall and precipitation can affect agricultural production, as crops and livestock rely on consistent access to water to thrive. In regions where rainfall is unpredictable or insufficient, farmers may struggle to produce enough food to meet the needs of their communities.

2. Water resources: Uncertainty in rainfall and precipitation can also affect the availability of fresh water for drinking, irrigation, and other purposes. In regions with unreliable rainfall, communities may struggle to access enough, fresh water.

3. Flooding and drought: Changes in rainfall and precipitation patterns can also lead to extreme weather events such as floods and droughts. These events can have serious consequences

including damage to infrastructure, destruction of crops, and loss of life.

4. Health impacts: Changes in rainfall and precipitation can also affect public health. For example, increased rainfall can lead to the proliferation of waterborne diseases, while drought can cause food and water shortages that lead to malnutrition and other health problems.

Overall, rainfall and precipitation uncertainty can significantly impact individuals, communities, and the environment. It is important to monitor and manage these risks.

This is where the need for precise rainfall predictions and weather forecasting unfolds among us, as it helps us monitor climate changes by providing accurate and up-to-date information about current and future weather conditions. This information can help us understand how the climate is changing and how it may continue to change in the future.

For example, by predicting the amount of rainfall that is expected in each region, we can better understand the potential for flooding or drought. This information can help us prepare for and mitigate the impacts of these events. Similarly, by forecasting the temperature, humidity, windspeeds, pressure, and many other environmental factors in each region, we can better understand how these factors may be impacting the local ecosystem and the people who live there.

DATA ANALYSIS:

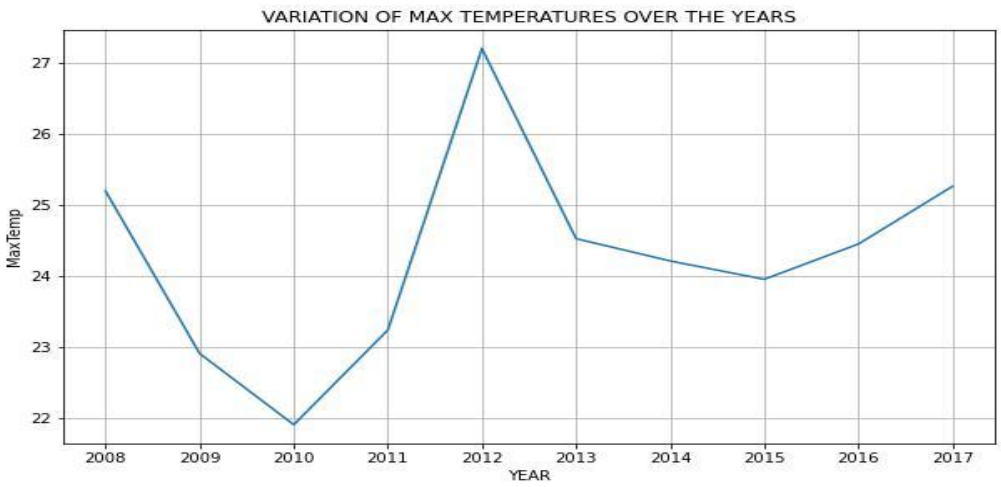
Data plays a crucial role in analysis because it is the raw material that is used to draw conclusions and make decisions. To perform any kind of analysis, you need to have data to work with. This data can come from a variety of sources, such as experiments, surveys, observations, or simulations.

The data we are using for our analysis and predictions consists of 10 years of daily weather observations from different locations in Australia.

TEMPERATURE

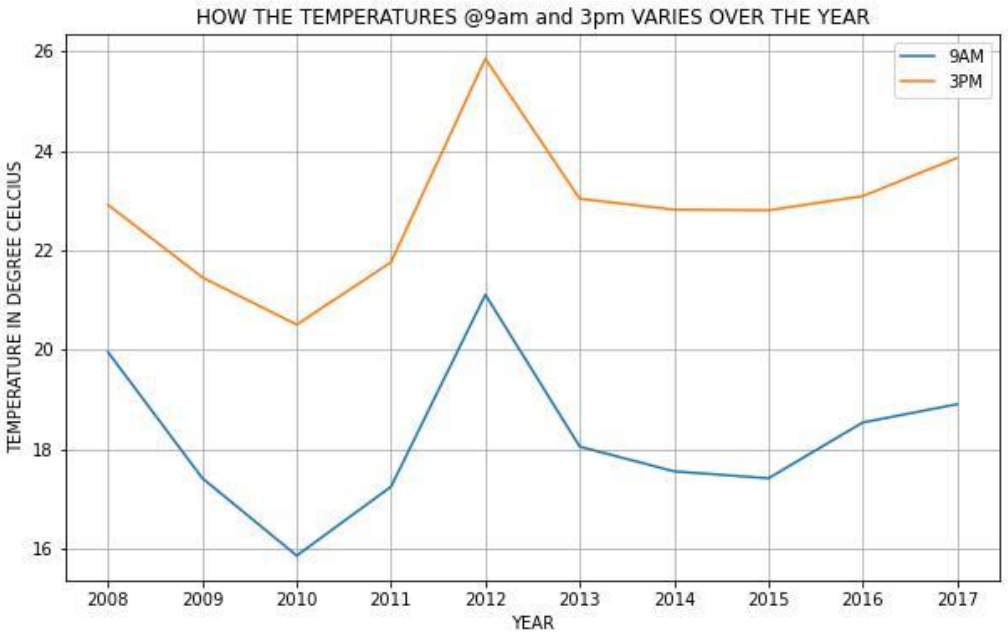
Temperature is a crucial factor that can have significant impacts on the ecosystem and environment. Higher temperatures can lead to a range of impacts, including Changes in the distribution and behavior of plants and animals, alterations in the water cycle, availability of resources such as food and water for plants and animals, and some components of the ecosystem,

So, it is very important to analyze the temperatures over the years and across the months at different timings to get a clear idea about the variations.

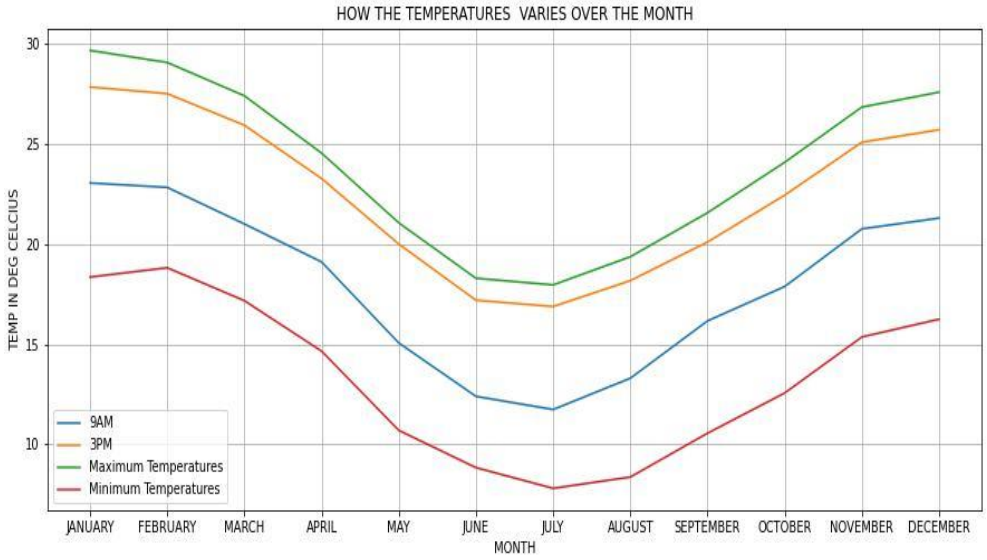


A steep fall in mean maximum temperature from 25.2 degrees Celsius to 22.9 degrees can be seen from 2008 to 2009 which is followed by a gradual decline to over 22 degrees in 2010. Maximum temperatures were recorded in 2012 and claiming it as one of the hottest years in Australian history.

Manmade global warming was likely a significant contributing factor in Australia's "Angry Summer" of 2012-2013, according to a study. It was the country's hottest on record and featured devastating wildfires as well as widespread flooding.



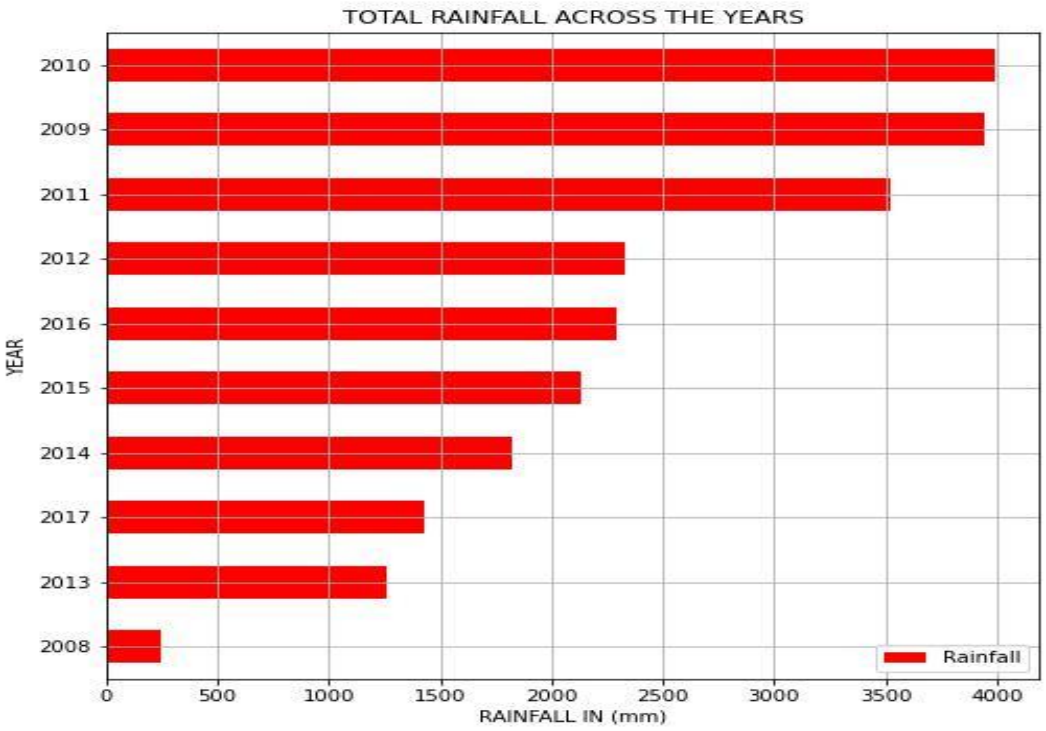
Variation over two different timings 9 am and 3 pm can be seen through the plot with higher temperatures being recorded @ 3 pm and it is obvious as the sun is highest in the sky at that period, what is interesting to note is that the pattern followed by both the curves (9 am and 3 pm) over the years is almost same.



Variations in the temperatures over the months can be seen from the above line plots. With maximum temperatures reaching around 30 degrees Celsius in summer. Minimum temperatures reach 5 degrees in the winter season (June, July, and August).

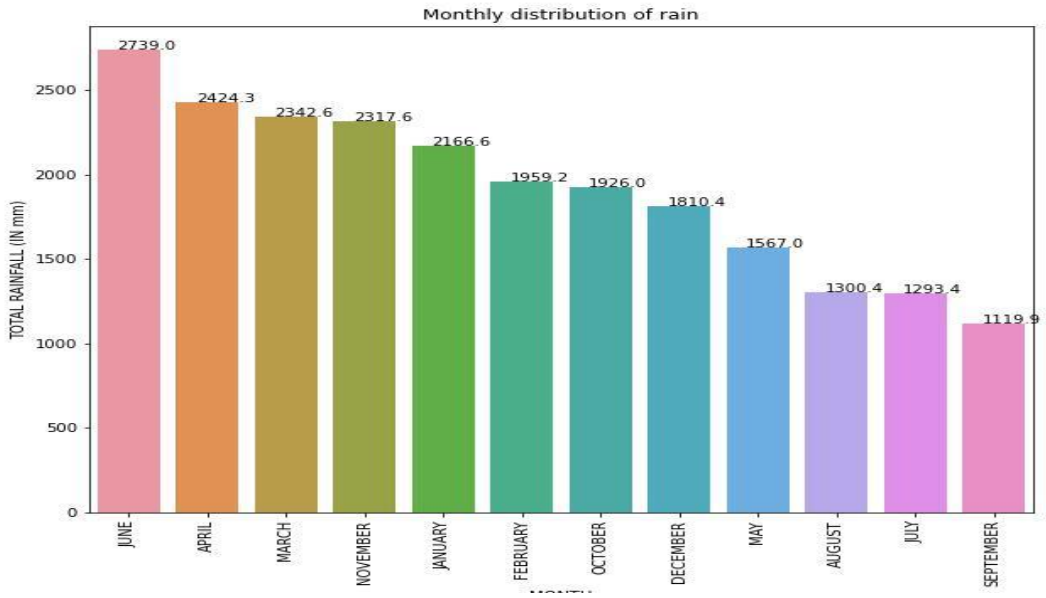
RAINFALL ACROSS THE YEARS/MONTHS BASED ON THE DATA

A good ecosystem can support a diverse range of plant and animal life, and where the various species within the ecosystem can coexist and thrive. This often requires a balance of different environmental factors, including temperature, humidity, and nutrient levels, as well as the ***amount of rainfall***



Maximum rainfall (in mm) has been recorded in the year 2010 with the reading reaching 4000mm throughout the year.

The year 2008 shows drastically low values of only around 300mm which is considerably low.

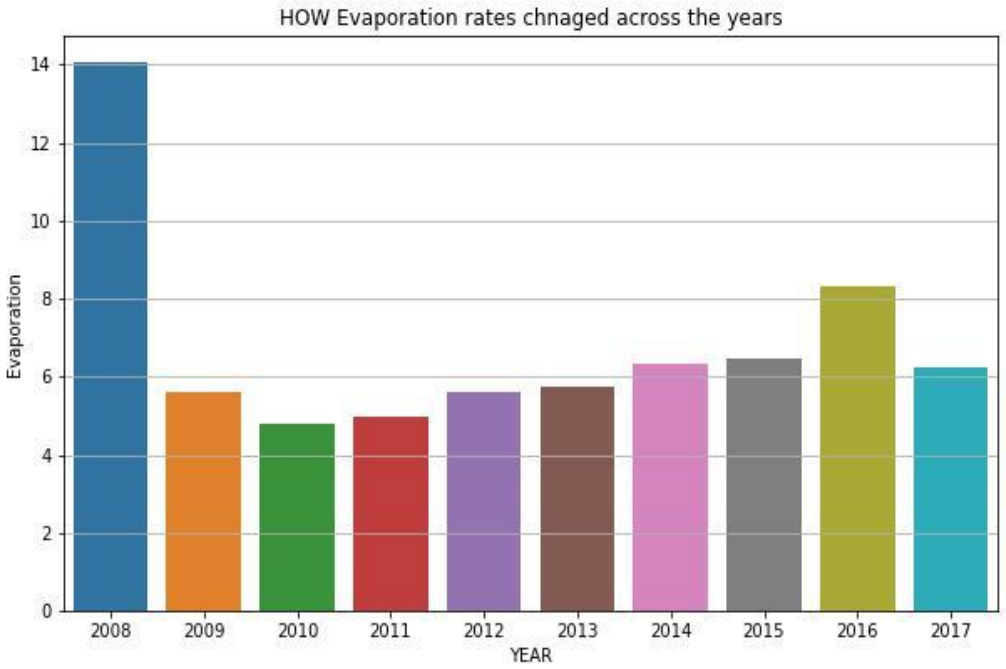


June, April, and March are ranked 1,2, and 3 respectively across the years in terms of the highest rainfall. Summer months including January, December, and February also record a significant amount of rainfall. September remains the driest month recording only 1119.9 mm of rains

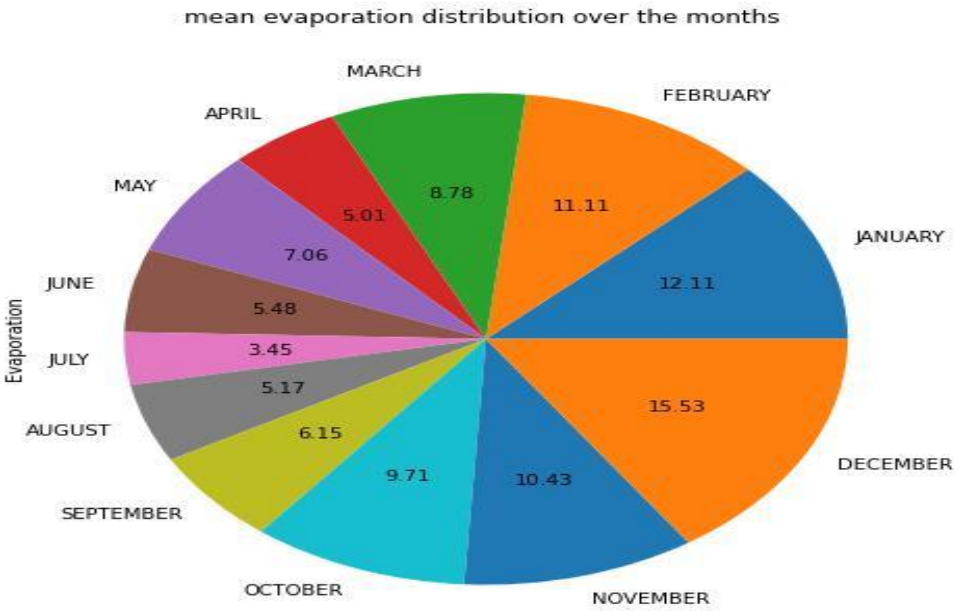
EVAPORATION

Evaporation is the process by which water is converted from its liquid form to a gas or vapor. It plays a significant role in the water cycle, which is the continuous movement of water between the Earth's surface and the atmosphere.

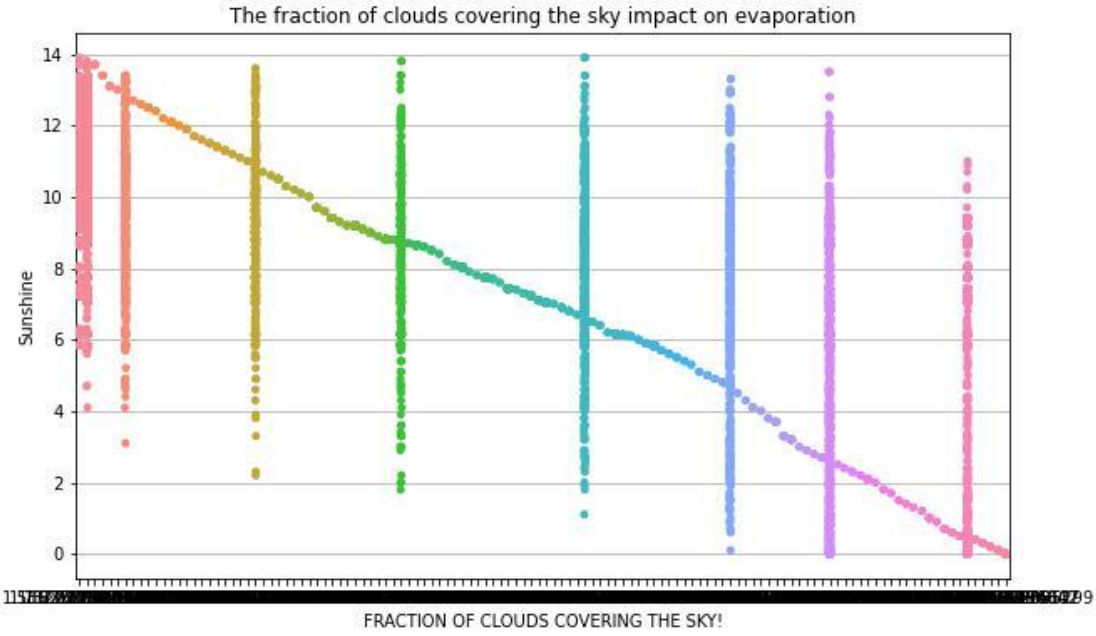
The amount of evaporation that occurs in a region depends on a variety of factors, including temperature, humidity, and wind. Warmer temperatures and lower humidity levels tend to promote more evaporation, while cooler temperatures and higher humidity levels tend to inhibit evaporation.



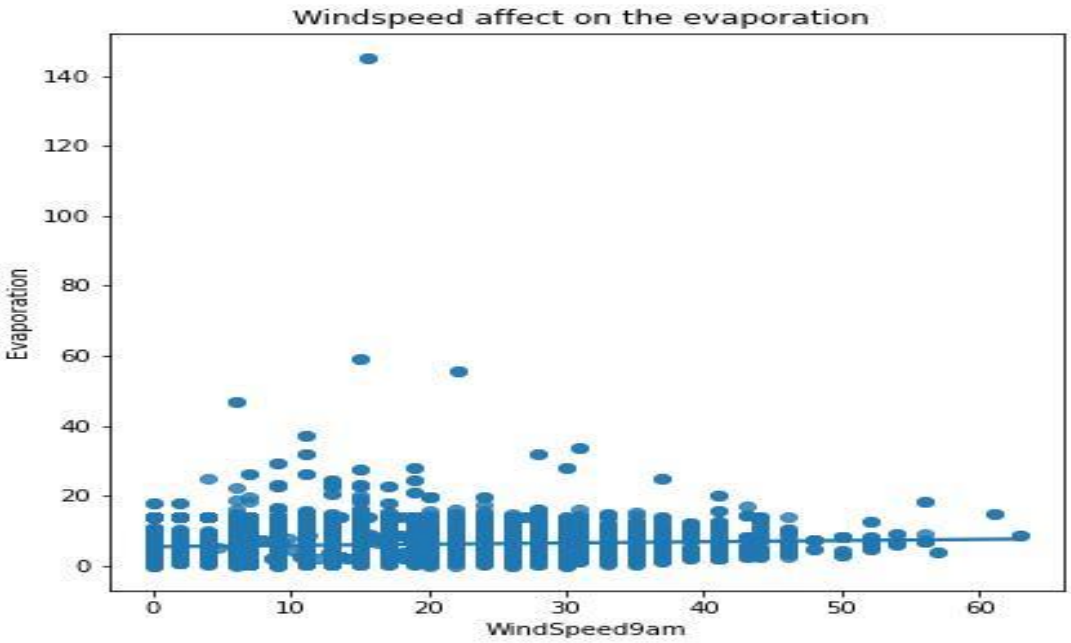
Evaporation rates over the years can be seen increasing from 2010 onwards by a significant amount.



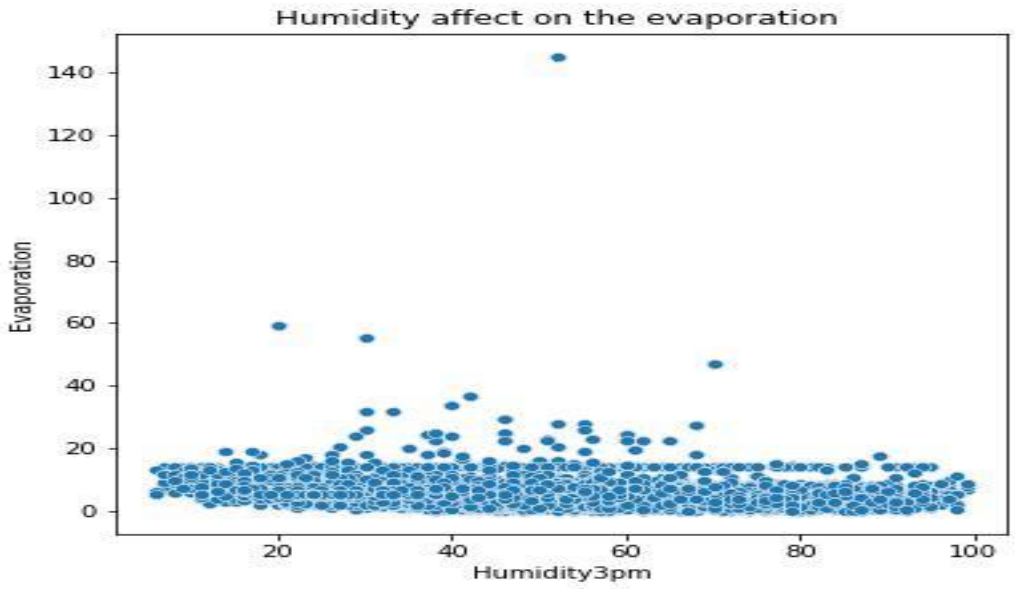
Mean Evaporation rates over different months can be seen in the above pie plot. The evaporation rate is at its peak in summer reaching values of 15.53 mm. Rates decrease to 3.45mm in July which is the coldest month of the season. It has a direct correlation with the amount of direct sunshine available over the water bodies causing the evaporation.



As the fraction of clouds covering the sky increases a decent drop in Direct sunshine can be seen from the above plot. Low direct sunshine further leads to poor evaporation rates.



When wind speeds are high, they can help to mix the air and bring in drier air from other areas. This can help to lower the humidity levels in the air, which in turn can increase the rate of evaporation. In addition, high wind speeds can help to cool the surface of the water, which can also increase the rate of evaporation.

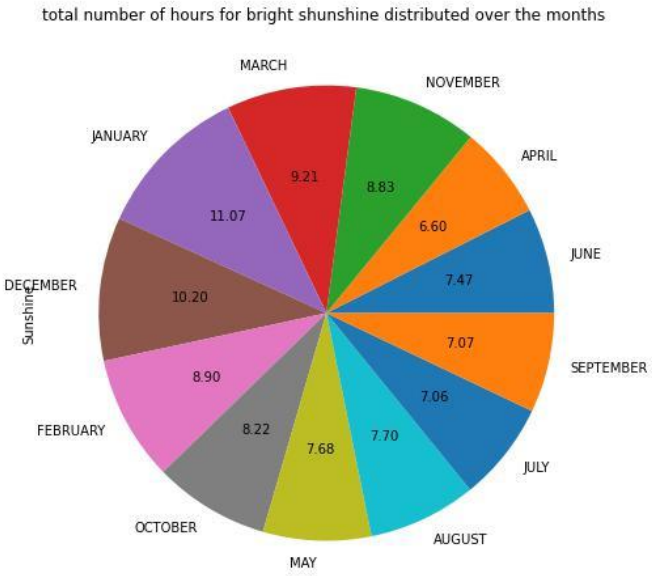


In general, high humidity levels tend to inhibit the rate of evaporation, while low humidity levels tend to promote the rate of evaporation. This is because the air can only hold a certain amount of water

vapor before it becomes saturated, and the amount of water vapor that the air can hold increases as the temperature increases. It can be further seen decreasing from the plot as the density keeps on decreasing as the humidity level increases.

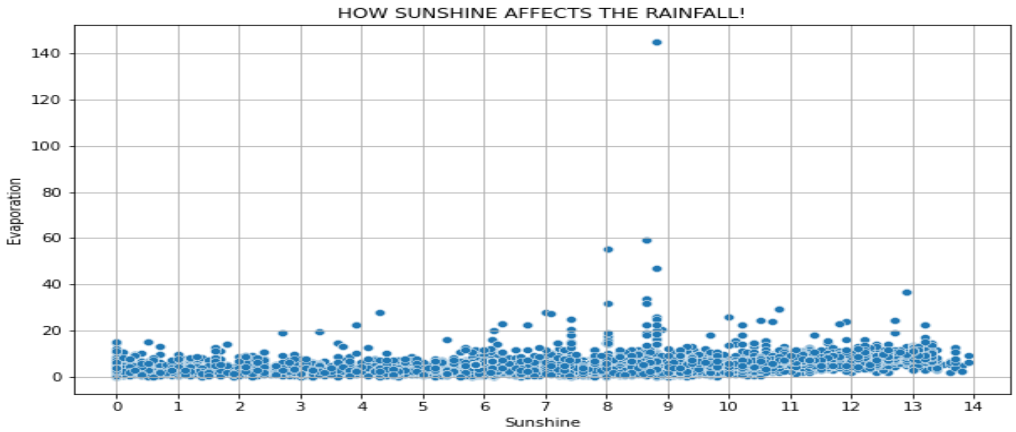
SUNSHINE

Sunshine plays a key role in the water cycle and can have a significant impact on the rainfall and weather patterns in different areas.



The total number of hours of direct sunshine reaching the ground even reaches 11.07 during the day in the month of January which is one of the hottest months in Australia.

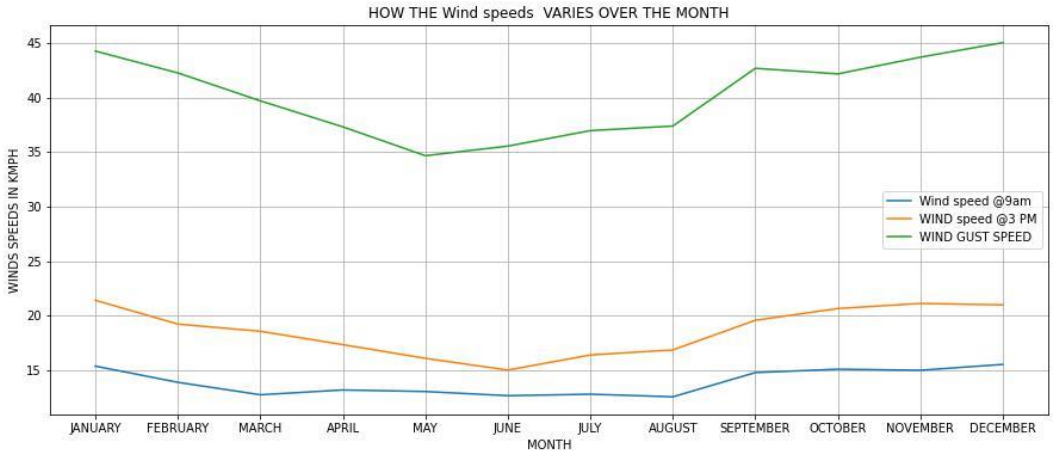
The sunshine can drop as low as 6 hours a day during the winter season.



With the increased number of direct sunshine over the terrestrial bodies, the rate of evaporation is seen increasing as the density can be seen increasing in the plot.

WINDSPEED

As the factors like wind speed, humidity, and pressure have no major changes over the years we are more interested in observing their patterns over the months (seasonal).

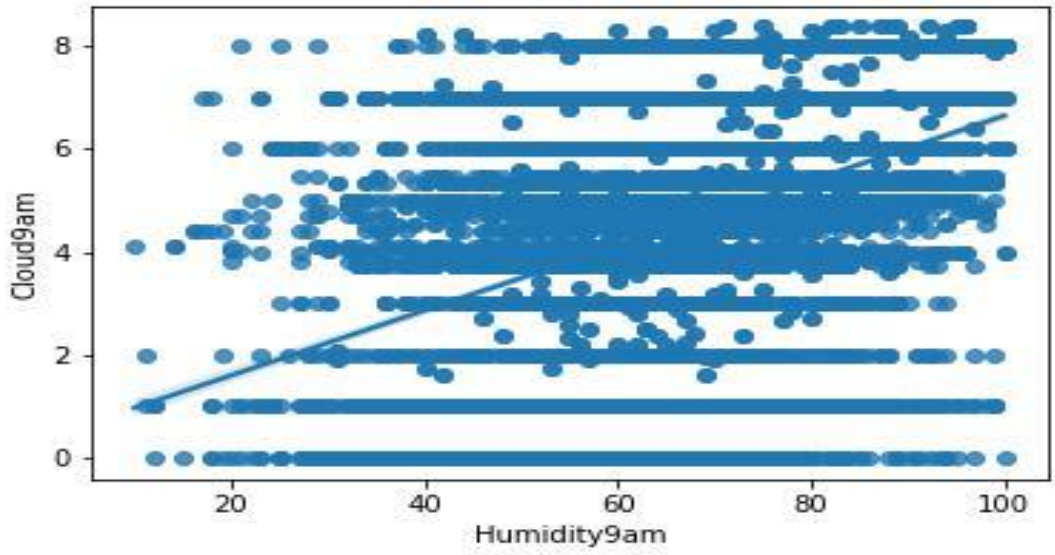


Wind gust speed can be seen reaching 45 kmph in the months of December and January and can be used to produce electricity using wind turbines.

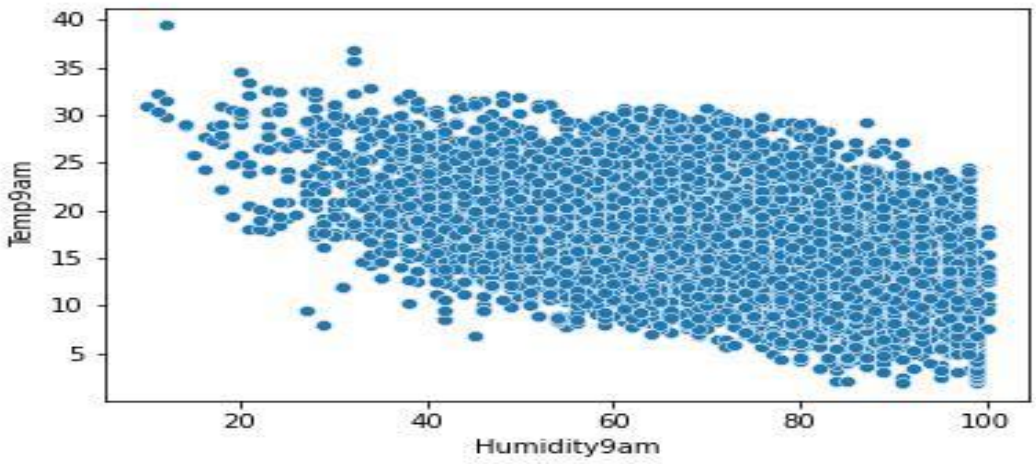
Winds @ 3 pm record larger speeds when compared to winds @ 9 am, and both follow the same patterns over the months.

HUMIDITY:

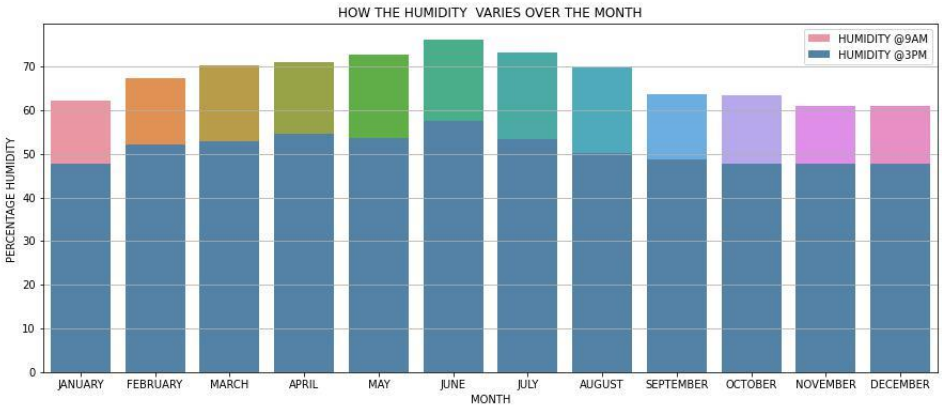
Humidity refers to the amount of water vapor present in the air, and it can have a significant impact on the rainfall and weather patterns in each area.



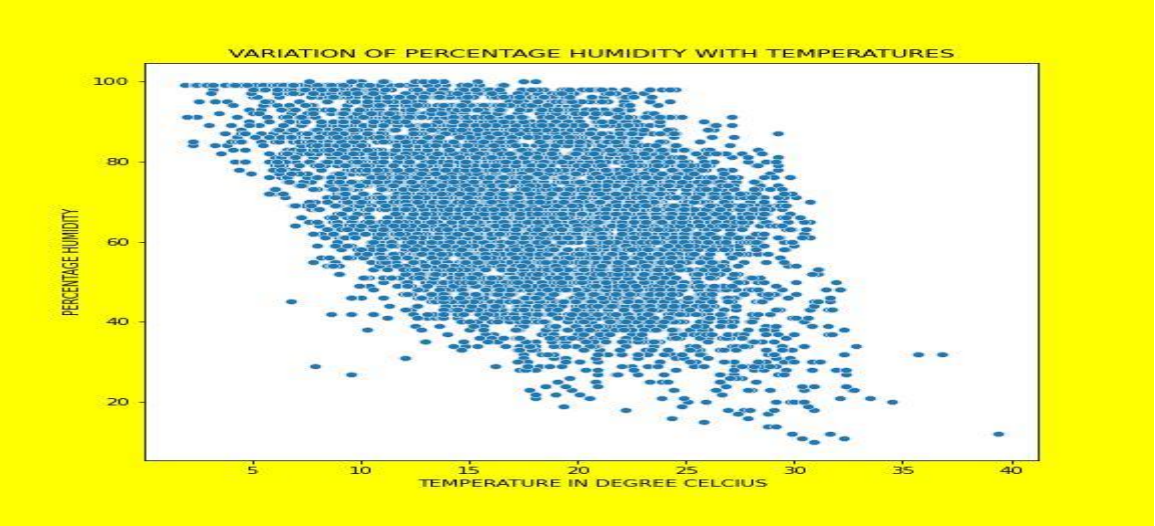
High humidity levels can lead to the formation of clouds, which can eventually produce precipitation in the form of rain or snow. This is because the air becomes saturated with water vapor and is unable to hold anymore, leading the excess water vapor to condense into clouds.



Increased percentage humidity over the region tends to bring down the temperature as can be seen from the above plot. The temperature even dropped below 5 degrees in some cases.

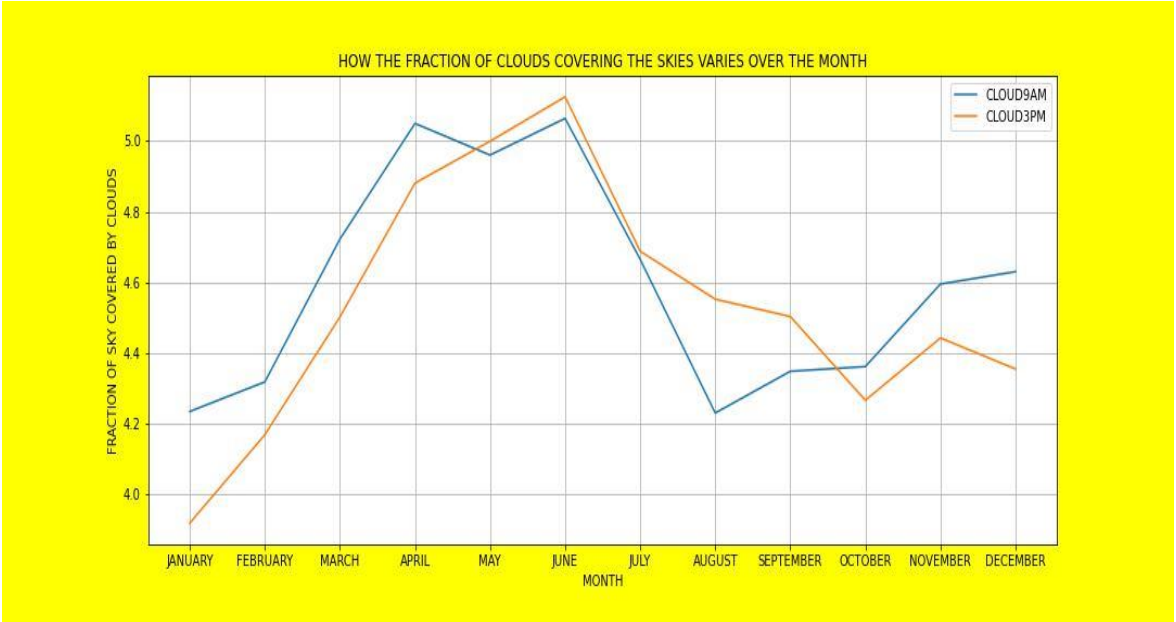


Humidity is always greater @ 9 am when compared to humidity @ 3 pm. As it is also very clear from the temperature-humidity plot that as the temperature increases the humidity tends to decrease. This is one of the reasons for higher humidity levels in the morning as compared to the afternoon. June and July record the maximum humidity as the conditions are most suitable for high humidity (low temperatures).



A strong relationship can be observed between temperature and percentage humidity. With the increasing temperatures, a considerable drop can be observed in the percentage humidity. As we know high temperature removes water from the air making it less humid and drier.

Clouds:



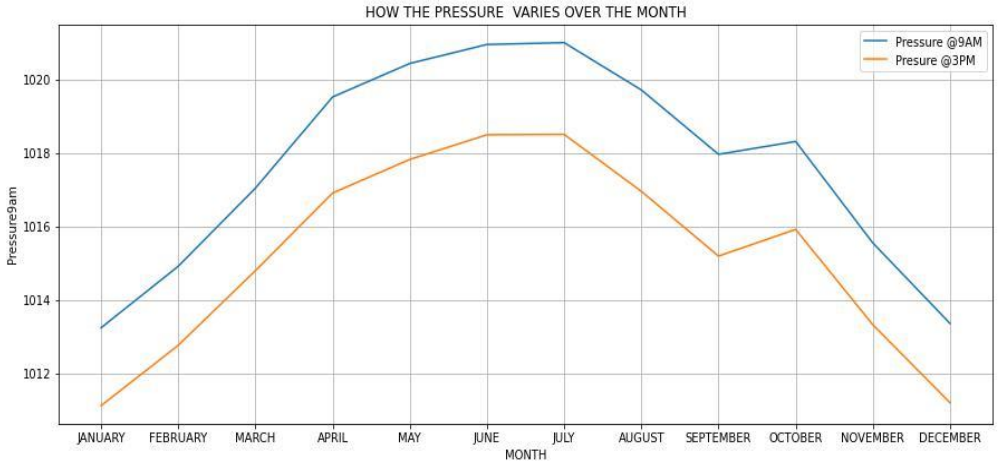
IN THE WINTERS/AUTUMN SEASONS CLOUDS COVER A LARGE FRACTION OF THE SKY REACHING A VALUE AROUND 6

FROM JANUARY TO APRIL THERE IS AN INCREASE IN CLOUDS FRACTION FOR BOTH THE TIMINGS AND CLOUDS @9AM HAVE A GREATER FRACTION AS COMPARED TO CLOUDS @3PM

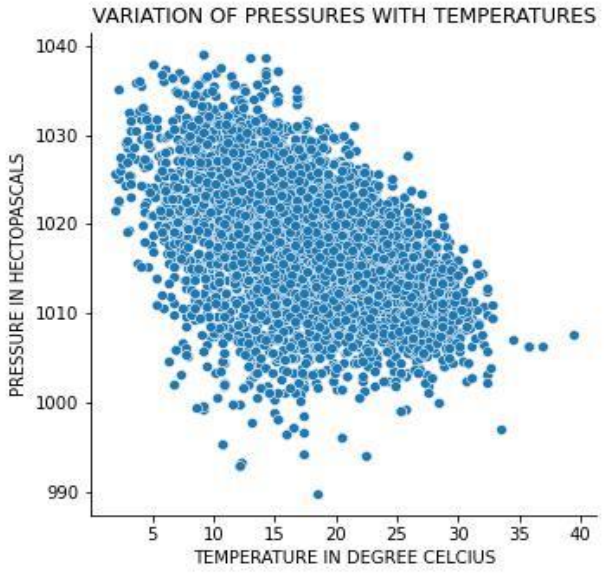
POST-APRIL CLOUDS @3PM TEND TO COVER A LARGER FRACTION TILL OCTOBER WITH BOTH ATTAINING THE PEAK IN JUNE!!

DECREASE IN THE CLOUD FRACTIONS POST WINTERS AND AS THE SUMMERS BEGIN AGAIN CLOUDS @9AM TAKE THE LEAD

PRESSURE :

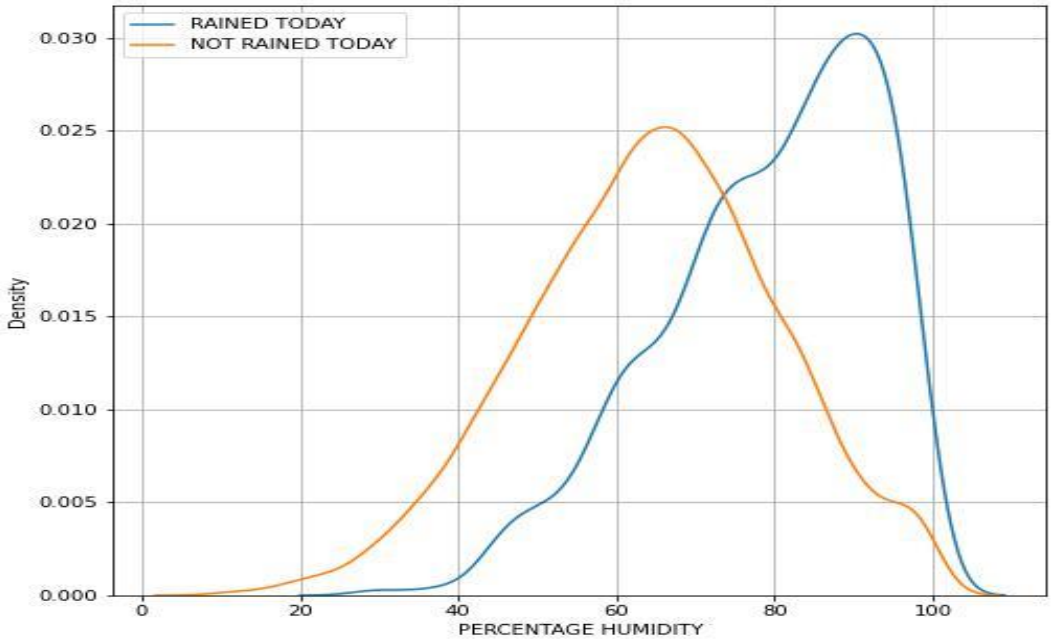


The surge in the pressure can be seen during winter with the pressures reaching 1021 hectopascals around 9 am. Pressure is relatively low in the afternoon though it follows the same pattern as followed by pressure in the morning over the years.

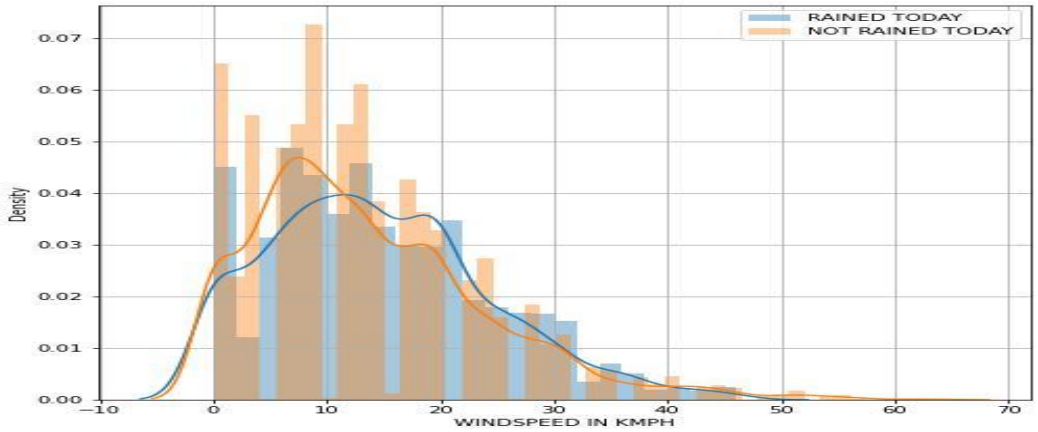


It can be inferred from the above plot that with the increasing temperature significant drop in the pressure can be observed.

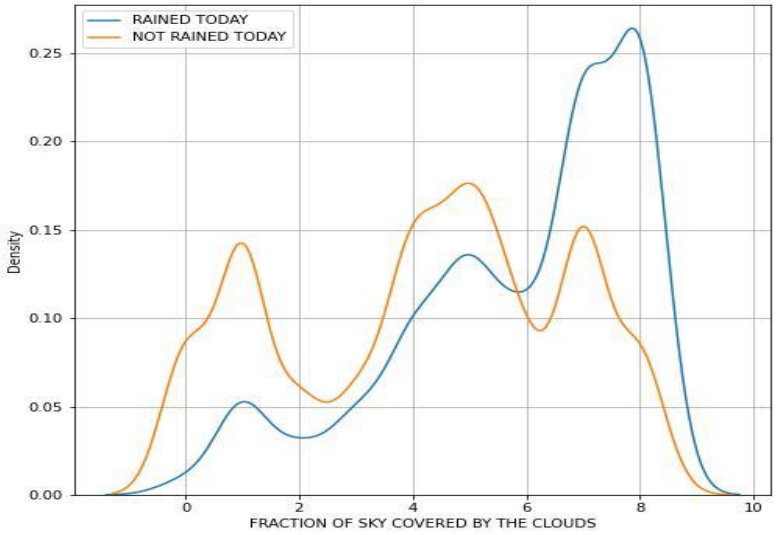
Analyzing the patterns when it rained and when it did not



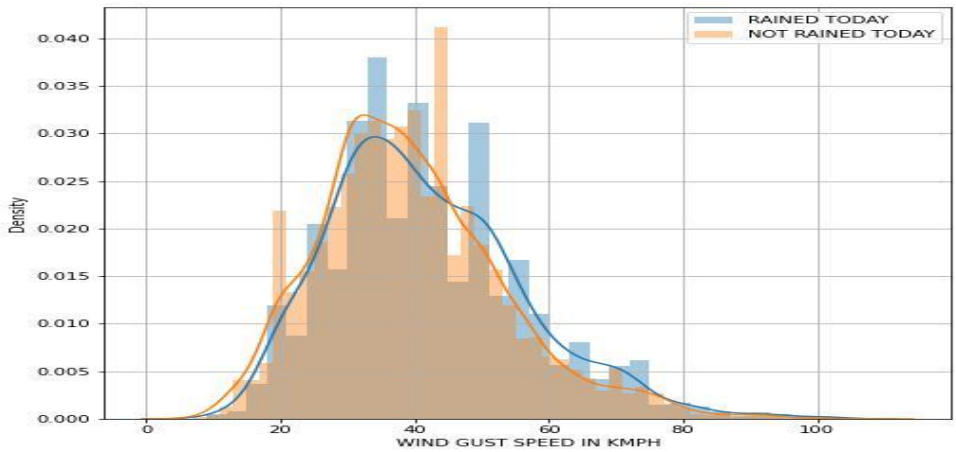
High values for percentage humidity were observed when it rained as compared to when it did not. Increased humidity can be a clear sign of the occurrence of rain or precipitation.



The mean wind speeds can be witnessed increasing when it has rained or is about to rain as compared to when it did not rain thus increased wind speeds can be a clear sign of rain or precipitation in the coming period or interval.



A high skewness to the right of the plot that describes the density distribution of clouds covering the skies can be observed indicating a direct relationship between clouds and rain or precipitation. More clouds are often linked with a good amount of rainfall or precipitation.



A plot describing the distribution of wind gust speed

EDA CONCLUDING REMARKS:

- Temperature, humidity, precipitation, air pressure, wind speed, and wind direction are key observations of the atmosphere that helps in predicting the weather conditions in any region.
- 2012 was observed to be the hottest year in Australia with the maximum temperatures attaining the values of 26 degrees Celsius.
- 2010 recorded the highest rainfall among all the analyzed years by measuring 4000 mm of rainfall across Australia
- When the relative humidity is high, it means that the air is holding a lot of water vapor, which can lead to rain or other forms of precipitation. When the relative humidity is low, it means that the air is relatively dry and will not lead to precipitation.
- Wind speed plays an important role in predicting rainfall or precipitation. Strong winds can cause air to rise and cool, which can cause condensation and the formation of rain clouds. This is especially true when the wind is coming from a cooler area. Strong winds can also cause existing rain clouds to move faster and cover a larger area, resulting in more precipitation. On the other hand, weak winds can cause the precipitation to slow down, reducing the amount of rainfall.
- Air pressure affects rainfall and precipitation in a variety of ways. Low air pressure causes air to rise, cool, and condense into rain clouds. Therefore, we have observed areas of low pressure, and rain often.
- Clouds can block sunlight from reaching the Earth's surface, cooling the air, and making it more likely for precipitation to occur. Clouds can also trap heat from the Sun, warming the air and making it more likely for moisture to condense into rain droplets

- Temperature has a significant impact on weather conditions. Temperature affects the air's ability to absorb and release moisture, which can lead to precipitation in each area. Temperature also affects air pressure, which can cause air to rise or fall and affect the formation of clouds and rain.

ABOUT THE DATASET WE HAVE USED:

The Dataset contains about 10 years of daily weather observations of different locations in Australia.

- Link- https://github.com/dsrscientist/dataset3

DATA DESCRIPTION

Number of columns: **23**

1. Date - The date of observation
2. Location -The common name of the location of the weather station
3. MinTemp -The minimum temperature in degrees Celsius
4. MaxTemp -The maximum temperature in degrees Celsius
5. Rainfall -The amount of rainfall recorded for the day in mm
6. Evaporation -The so-called Class A pan evaporation (mm) in the 24 hours to 9 am
7. Sunshine -The number of hours of bright sunshine in the day.
8. WindGustDir- The direction of the strongest wind gust in the 24 hours to midnight
9. WindGustSpeed -The speed (km/h) of the strongest wind gust in the 24 hours to midnight
10. WindDir9am -Direction of the wind at 9am

11. WindDir3pm -Direction of the wind at 3 pm
12. WindSpeed9am -Wind speed (km/hr) averaged over 10 minutes before 9 am
13. WindSpeed3pm -Wind speed (km/hr) averaged over 10 minutes before 3 pm
14. Humidity9am -Humidity (percent) at 9 am
15. Humidity3pm -Humidity (percent) at 3 pm
16. Pressure9am -Atmospheric pressure (hpa) reduced to mean sea level at 9 am
17. Pressure3pm -Atmospheric pressure (hpa) reduced to mean sea level at 3 pm
18. Cloud 9 am - Fraction of sky obscured by cloud at 9 am.
19. Cloud3pm -Fraction of sky obscured by cloud
20. Temp9am-Temperature (degrees C) at 9 am
21. Temp3pm -Temperature (degrees C) at 3 pm
22. RainToday -Boolean: 1 if precipitation (mm) in the 24 hours to 9 am exceeds 1mm, otherwise 0
23. RainTomorrow -The amount of next day rain in mm. Used to create response variable. A kind of measure of the "risk".

Pre-Processing Pipeline:

Pre-processing of data is necessary before building a model because it helps to ensure that the data is in the right format and is of the highest quality. Pre-processing includes tasks such as cleaning the data, removing any inconsistencies (handling the nulls), and transforming the data into a format that the model can understand. Pre-processing can also help to reduce noise and improve the accuracy and performance of the model. Additionally, pre-processing can help to identify and remove any outliers or incorrect data points that could affect the model's accuracy.

Handling the null values

- A significant number of null values can be seen in the raw dataset and must be treated using techniques like simple imputation, iterative imputers, and KNN imputation techniques.

- As we can observe from the analysis and applying some basic knowledge about the features, features like temperature, and wind gust speed, showed a pattern that was changing over both months and years.
- Following it, we will be imputing “Temp9am”, “Temp3pm”, “MaxTemp”, “MinTemp”, and “WindGustSpeed”, using the grouped mean of year and month. This will be the most suitable way of imputing these as they show similar kinds of patterns over different months and across the years.

  For ex-imputing MaxTemp

  dft['MaxTemp']=dft.groupby(['YEAR','MONTH'],sort=False)['MaxTemp'].apply(lambda:x.fillna(x.mean()))

- As the amount of rainfall (Rainfall) showcased a pattern over the years it will be much better if we consider only the year-wise mean to impute the missing values in the amount of rainfall.

- “Sunshine”,” Pressure9am”, and “Pressure3pm” exhibits a meaningful pattern over different months, imputing the nulls with the month-wise mean can do justice to it.

- Evaporation is a continuous feature and is measured using class-A pan evaporation taking place in mm.
  A strong relationship is found between sunshine and evaporation, Using the KNN Imputation technique to fill the nulls present in Evaporation will be a good idea.
  For example:
  fkk=pd.DataFrame(knn.fit_transform(dft[['Sunshine','Evaporation']]))
  dft['Evaporation']=fkk[1]

- “WindGustDir”,” WindDir9pm”,” WindDir3pm”,”RainToday”, “RainTomorrow” is witnessed as categorical columns and it will be wise to impute them using a simple imputation technique with strategy as “most frequent”.
- “WindSpeed9am”, and” WindSpeed3pm” exhibited a high correlation with the WindGustSpeed convincing us to fill the nulls in both features using the Iterative Imputation technique.

  For example:

  lr=LinearRegression()

  ii=IterativeImputer(lr)

  new=pd.DataFrame(ii.fit_transform(dft[['WindGustSpeed','WindSpeed9am']]),columns=['WindGustSpeed','WindSpeed9am'])

  dft['WindSpeed9am']=new['WindSpeed9am']

- “Humidity9am”,” Humidity3pm”,” Cloud9am”, and” Cloud3pm” were seen experiencing high correlation with the feature Sunshine so it will be better if we use iterative imputation to impute the nulls present in these 4 columns.

Encoding the categorical columns

It is necessary to encode categorical columns because most machine-learning models only accept numerical variables. Since categorical variables are usually represented as strings or categories, they need to be converted into numerical values for the model to understand and extract valuable information from them.

- Date being an ordinal entity must be ordinally encoded.

- WindGustDir, WindDir3pm, and WindDir9am have many categories that must be encoded using a binary encoding technique

- Mapping RainToday, RainTomorrow with 1 if it rained and 0 if it did not.

Data Transformation

From the analysis, some of the features have high skewness deviating from a Gaussian distribution. An alarming number of outliers were also detected in the distribution plots of some features.

So, we have tried scaling the data using the PowerTransformer technique.

It is a technique for transforming numerical input or output variables to have a Gaussian or more Gaussian-like probability distribution. An important feature of this scaling method is that it can also be used to reduce the effect of outliers on the model by transforming the data into a more uniform distribution.

Outlier handling in the dataset

It is necessary to handle outliers in the dataset because they can have a significant impact on the accuracy and performance of the model. Outliers can skew the results of the model, lead to inaccurate predictions, and reduce the accuracy of the model. Outliers can also cause instability and bias in the model, and can even lead to overfitting. Therefore, it is important to identify and remove outliers from the dataset before building a model to ensure the highest accuracy and performance.

We will be using ***Datasist.structdata*** a Python library for outlier handling. It provides a range of powerful functions and methods for quickly manipulating, transforming, and visualizing data. It also provides powerful tools for analysis, such as statistical tests and machine learning algorithms.

Out ruling multicollinearity

Checking for multicollinearity among features is an important step in machine learning because it can help to identify features that are highly correlated with each other. These features may not be providing any additional information to the model and can even cause the model to be unstable or inaccurate.

We will be using ***variance_inflaton_factor*** technique to check the z-scores among the features.

- Date, Year
- MaxTemp, Temp9am, Temp3pm
- Pressure9am, Pressure3pm

These features were seen having z-scores greater than 10 hinting toward the high multi-collinearity among these features and must be treated before proceeding to the modeling as they can hinder the stability and lead to the model overfitting.

As a result, we will be dropping

- Date
- MaxTemp
- Temp9am
- Pressure3pm

Helped in controlling the z-scores and all the scores were less than 10 confirming reduced multi-collinearity.

Selecting best features

Selecting the best features for a model is important because it increases the accuracy and performance of the model. By carefully selecting the most relevant and appropriate features, the model will be able to make better decisions and produce more accurate results. Moreover, selecting the best features also helps reduce the computational overhead and improves the efficiency of the system.

We will be selecting the best features using **SelectKBest** method. The top 28 features with corresponding ANOVA scores are sorted. We have used score_func as f_classif to iterate over the features using the ANOVA method.

Building Machine Learning Models

We have the problem in two parts

1) Predicting the amount of rainfall
2) Predicting if it will rain Tomorrow

Part 1. Predicting the rainfall is a regression problem and Part 2. Predicting if it will tomorrow is a classification problem.

Part1) Predicting the amount of rainfall

After pre-processing the data, we are now ready to split the data into train and test.

Now the training dataset will be used to train our models.

Since it is a regression problem, we will start with LinearRegression model.

Model 1) **LinearRegression**

- Post-training the model with the train dataset now we will start predicting the values for test dataset.
- Evaluation of the accuracy scores using r2_score
- Determining the random state at which the difference between the testing and the training r2_score is the least.
- When the random state is fixed, we will check for cross-Val scores for different cross-validation values.
- Selection of the best cross-validation value will depend on the least difference between testing values r2_score and cross-Val score at that corresponding cross-validation value.
- Once the cross-validation value is fixed we can use the same value of it and random state over different models to compare the performance among them.

Evaluation of

- Mean_Squared_error
- Mean_absolute_error
- Root_mean_Sqaured_error
- Difference between root_mean_squared_error and mean_absolute_error.

1. High r2_scores for both training and testing as well as the least difference between the two
2. Lowest values for Root Mean Squared Error and Difference between Root Mean Squared Error and Mean Absolute Error

These will be the deciding factor for the best model for this regression problem.

Evaluating the same scores and errors for

Model 2) **DecisionTreeRegressor**

Model 3) **AdaBoostRegressor**

Model 4) **RandomForestRegressor**

Model 5) **SVR**

Model 6) **KNN**

Good and repeated hyper-parameter tuning of parameters will help us bring the best conclusions among all models.

Conclusion

Among all the LinearRegression had high and close values for training and testing r2_scores along with the least difference between root mean squared error and mean absolute error depicting closer values to the fitting line

Part 2) Predicting if it will be Tomorrow or not

- The same steps as we followed for the regression problem post-pre-processing
- Instead of LinearRegression we will be using LogisticRegression
- Model 1) **LogisticRegression**
- Instead of calculating r2_scores we will be calculating accuracy_scores.
- After selecting a suitable random state and cross-validation value we will be proceeding with using them in each model to compare the performance.

Evaluation of

- Confusion matrix
- Classification report

1. High accuracy_score and the lowest difference between training and testing accuracy_score
2. Low false positives and false negatives depending upon the type of problem
3. Closer values for Area Under the Curve when plotted for both training and testing datasets.

These will be the deciding factor for choosing the best classification model for this problem.

Evaluating the same as discussed above for

Model 2) **DecisionTreeClassifier**

Model 3) **RandomForestClassifier**

Model 4) **BaggingClassifier**

Model 5) **KNN**

Proper Hyperparameter tuning of the parameters involved in each model will help us fine-tune the model and increase its accuracy.

Conclusion

1. Since for LogisticRegression the difference between the cross-Val scores and the accuracy score is considerably lower involving the difference between the training and testing score is the lowest.
2. Individual Training and testing scores are among the highest giving it an advantage over other models.
3. Same AUC scores are being recorded for both the training and testing datasets for LogisticRegression Model.

These factors make us choose LogisticRegression as our best model for this problem.