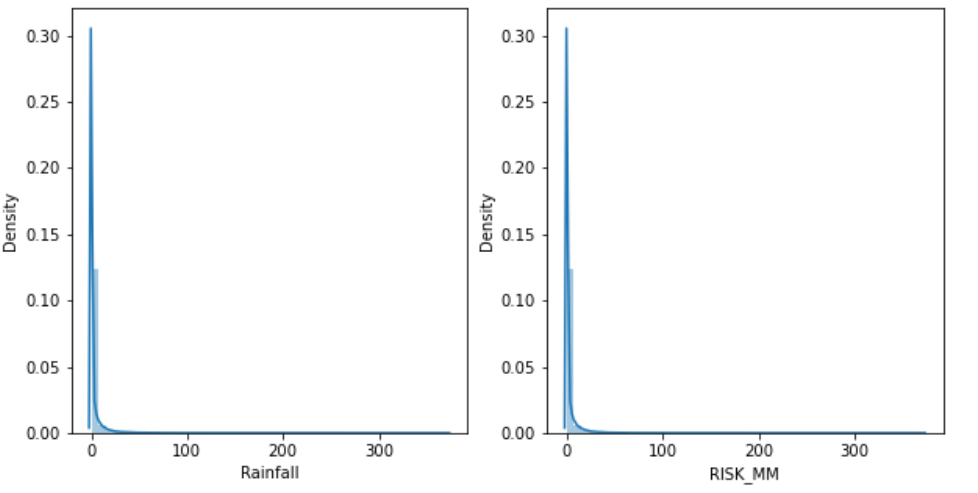
1. **Meeting 1: - (29th May, 2021)**

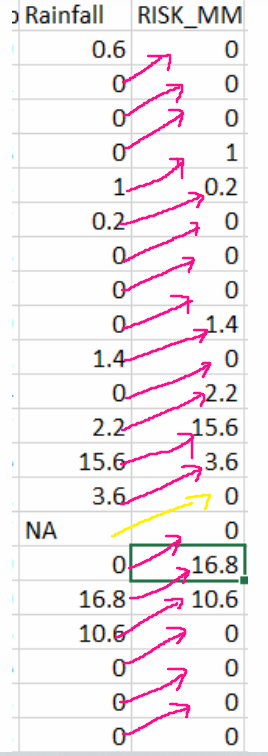
We found out in this meeting that both Rainfall and RISK\_MM features in our data are basically same. For that we can prove them by using Excel and we also saw the distplot for the same and we are getting the same graph as well which gives us enough evidence that both these features are same. So, we have decided that we will be dropping RISK\_MM feature since the null values which correspond to Rainfall feature is filled 0 in the RISK\_MM feature which might not be accurate.

Distplot graph for both Rainfall and RISK\_MM:



Here in this graph, we can see that both these features have the same density plotting with the same scale.

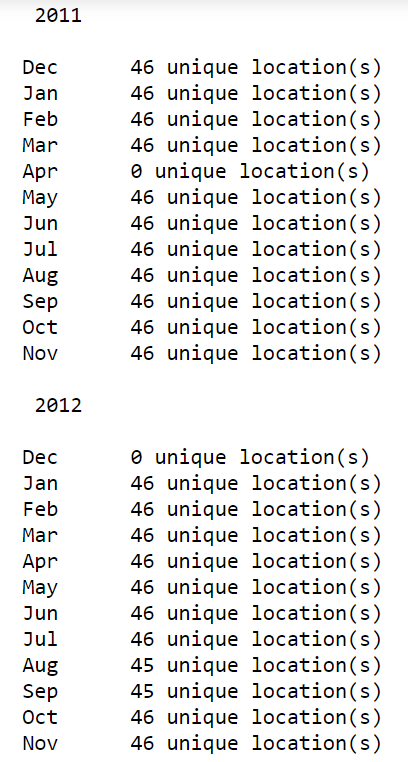
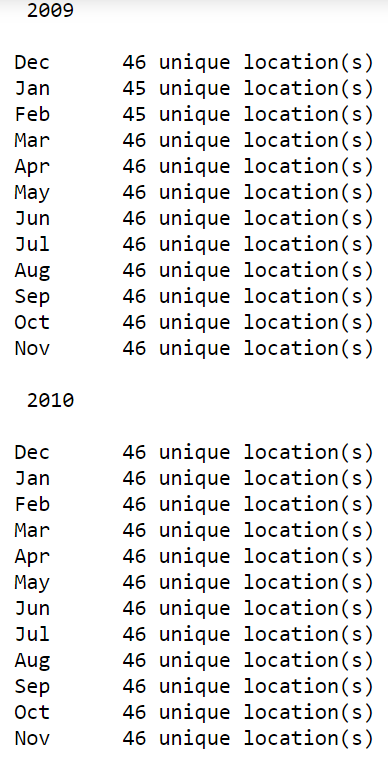
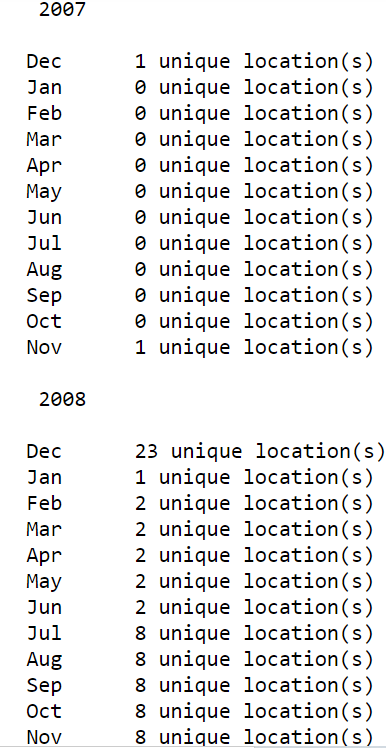
EXCEL proof for both Rainfall and RISK\_MM features:-

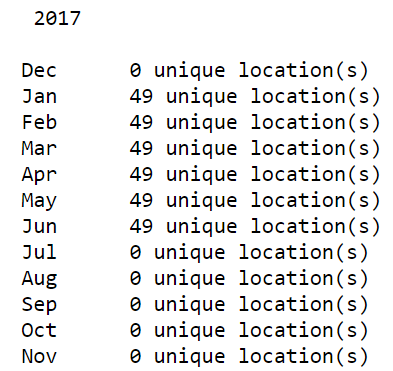
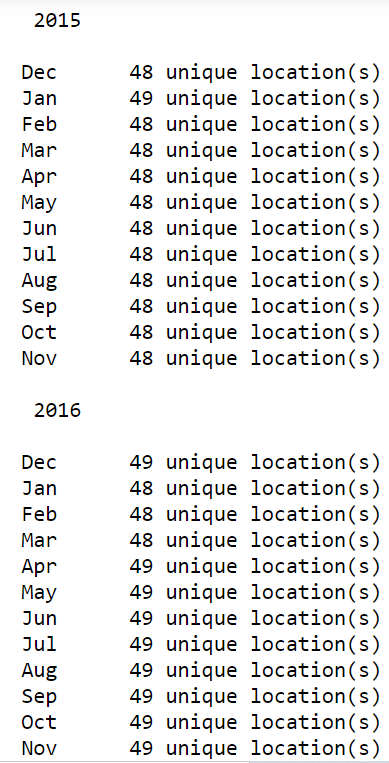
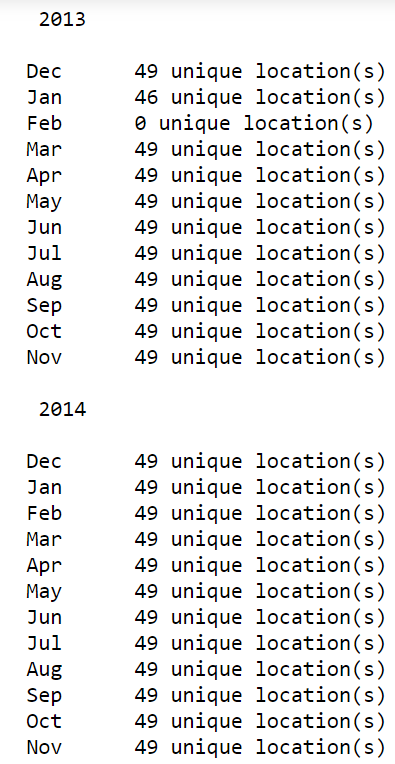


This image basically sums up what we want to try to convey about the data.

1. **Meeting 2: - (30th May, 2021)**

* Here in this meeting, we decided to explore the features **“MinTemp”** and **“MaxTemp”** and study more about this data. We explored the features with respect to both date and locations.
* The images below show how many unique locations has been given year wise and month wise as well.

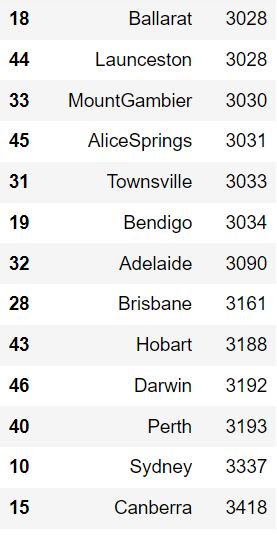
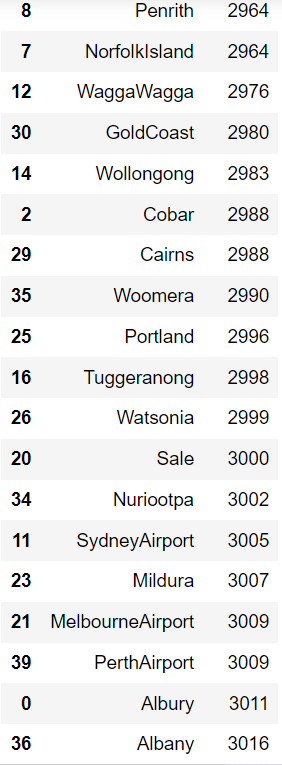
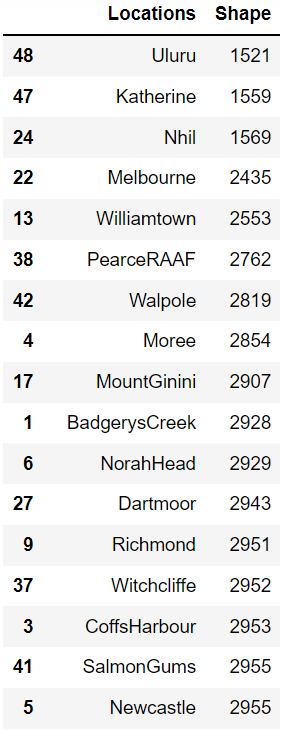




From the above graph, we can say that our actual data starts **from November 2007** and ends **with June 2017** with uneven data recorded for each month. Here the data of locations are not uniform.

* Here, we are looking at the no. of records for each location sorted in ascending order.

We can see that **Uluru has the least amount of data** and **Canberra has the highest amount of data** in the whole dataset.



* We also plotted the histogram of “MinTemp” and “MaxTemp” at each month and observed that the data follows an almost similar pattern (somewhat normal) throughout each month.
* On plotting the histogram of the same features with respect to each location, we observed that the data does not follow similar pattern for all the location.

1. **Meeting 3 (2nd June, 2021)**

Here we have decided to explain each variable.

ASHWIN: - Humidity9am and Humidity3pm

SIMRAN: - Pressure9am and Pressure3pm

ANSHIKA: - Cloud9am and Cloud3pm

ASHUTOSH: - Temp9am and Temp3pm

1. ***Humidity9am and Humidity3pm***

First, we need to understand about this feature. In this data, basically it is RELATIVE HUMIDITY.

* **WHAT IS RELATIVE HUMIDITY?**

**Relative humidity** tells us how much water vapor is in the air, compared to how much it could hold at that temperature. It is shown as a percent. For example, a relative humidity of 50 percent means the air is holding one half of the water vapor it can hold.

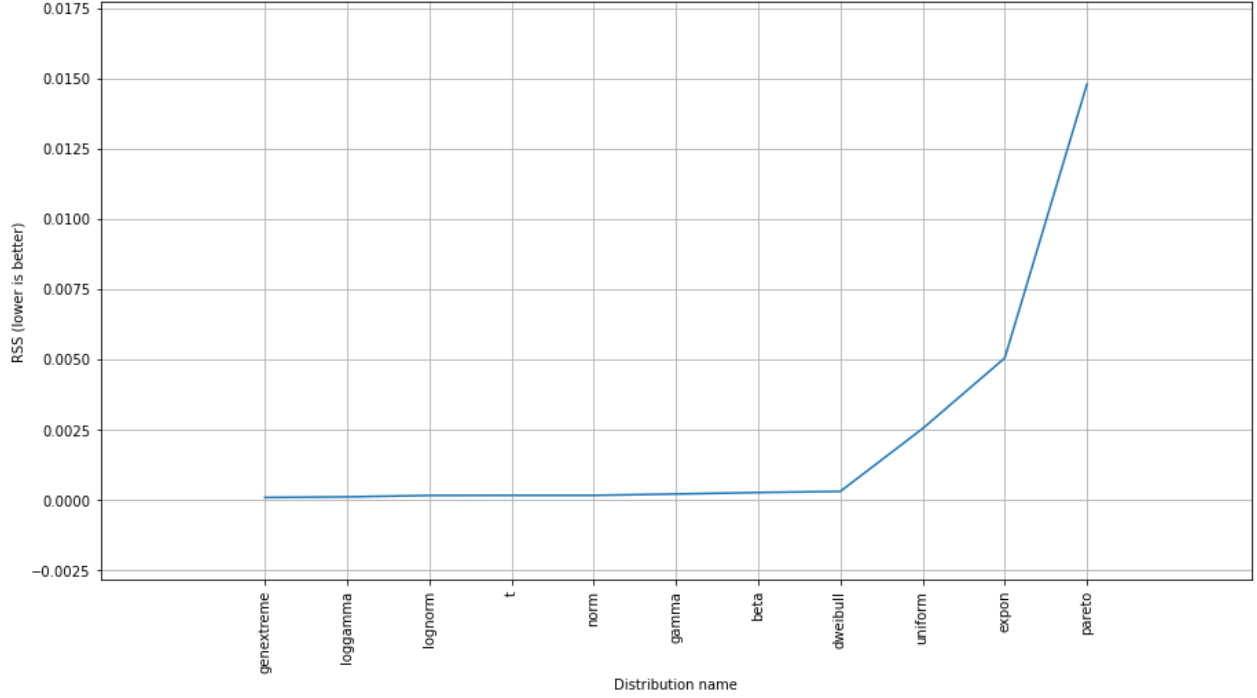
* **HOW IS IT DIFFERNENT FROM HUMIDITY?**

Humidity is the water content of the mixture of water vapor and other elements found in the air while relative humidity is the percentage of water vapor in the air at a given temperature.

Secondly, we need to answer sub questions which well explains the humidity data.

* **WHAT IS THE DISTRIBUTION OF THE DATA?**

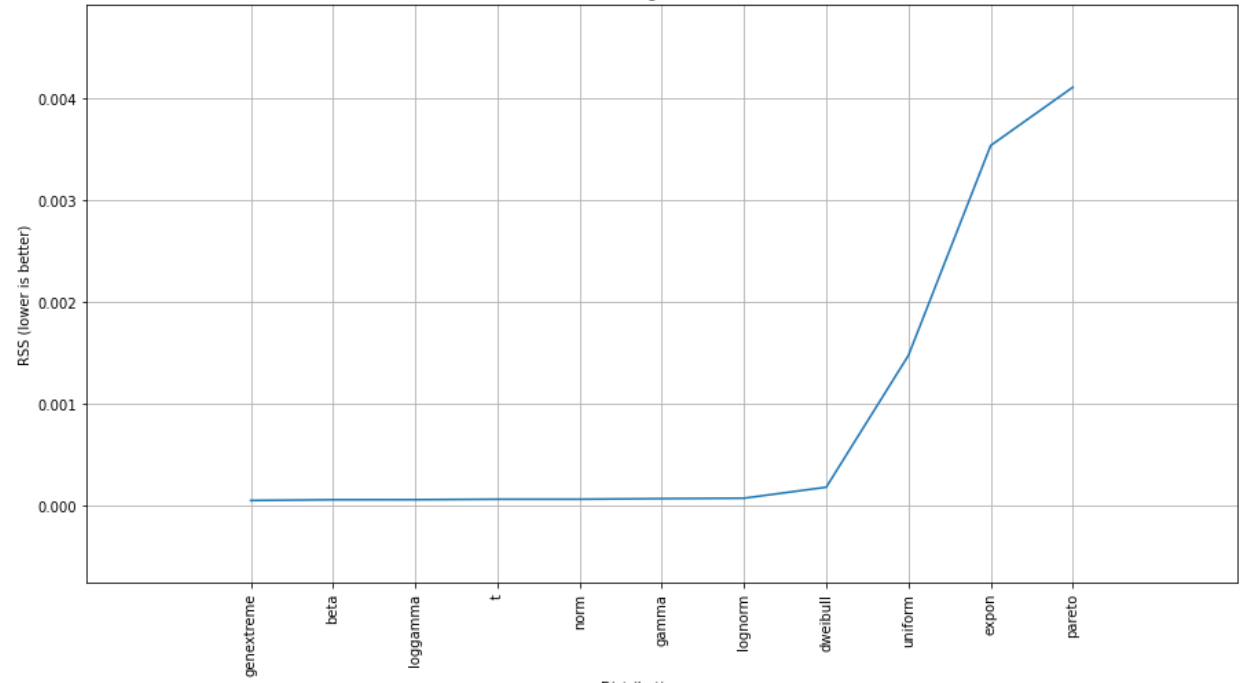
For the feature **Humidity9am**, we can see the graph below in order to understand what all distributions are quite available for us and which one is the best fit.



Here in this graph, we can see that on the x-axis we have the names of the distributions and on y-axis we have the RSS (Residual Sum of Squares).

Here from the graph, it is quite obvious that we should be choosing genextreme distribution since it has the lowest RSS. But that is something which we have to look on. If we also see RSS of normal distribution, we can say that there is not much difference in RSS between genextreme and normal distribution. So, we can say that Humidity9am feature follows normal distribution.

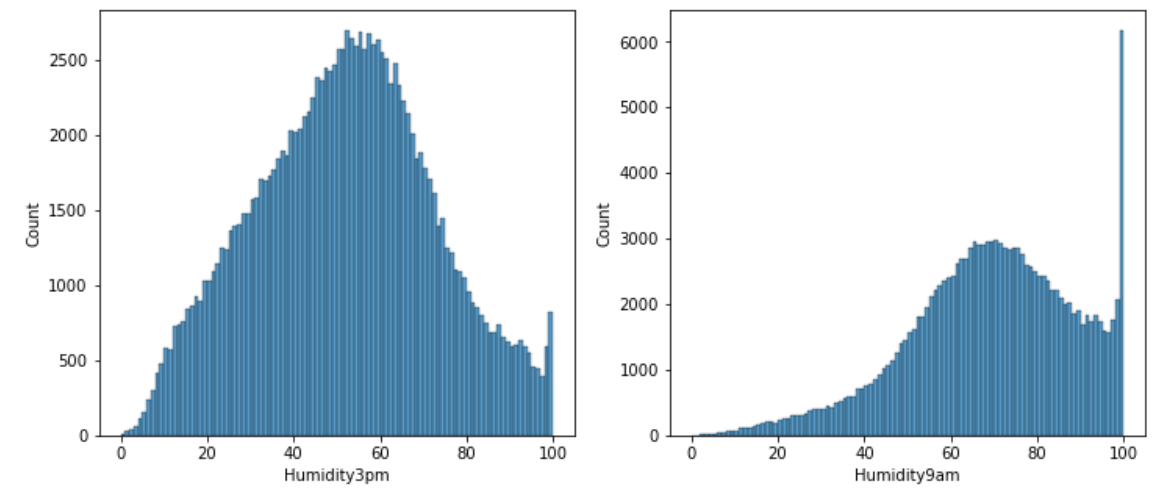
For the feature **Humidity3pm**, we can see the graph below in order to understand what all distributions are quite available for us and which one is the best fit.



Here in this graph, we can see that on the x-axis we have the names of the distributions and on y-axis we have the RSS (Residual Sum of Squares).

Here from the graph, it is quite obvious that we should be choosing genextreme distribution since it has the lowest RSS. But that is something which we have to look on. If we also see RSS of normal distribution, we can say that there is not much difference in RSS between genextreme and normal distribution. So, we can say that Humidity9am feature follows normal distribution.

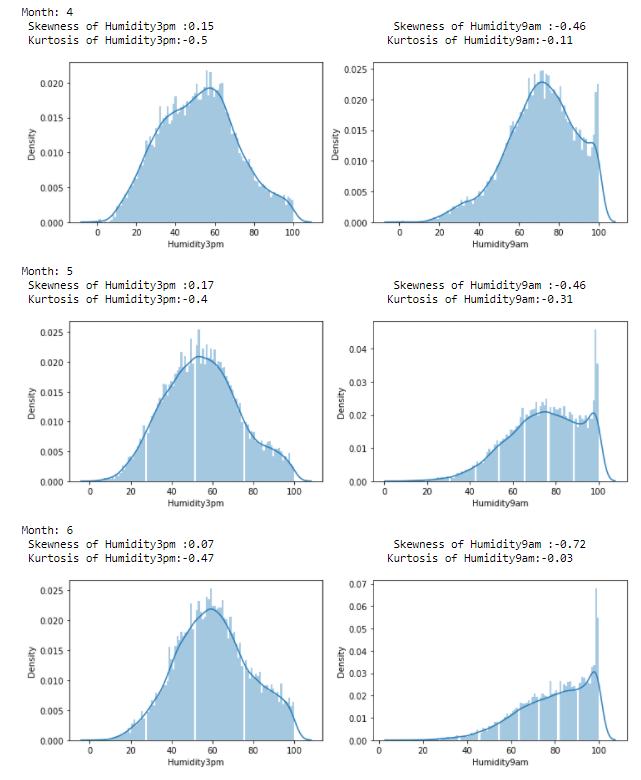
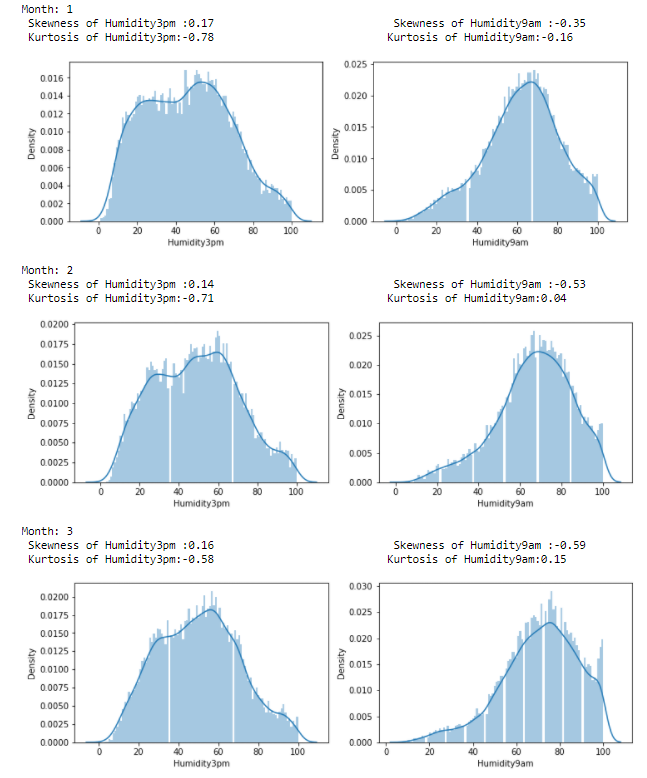
* **DISTRIBUTION PLOT OF THE HUMIDITY FEATURE?**

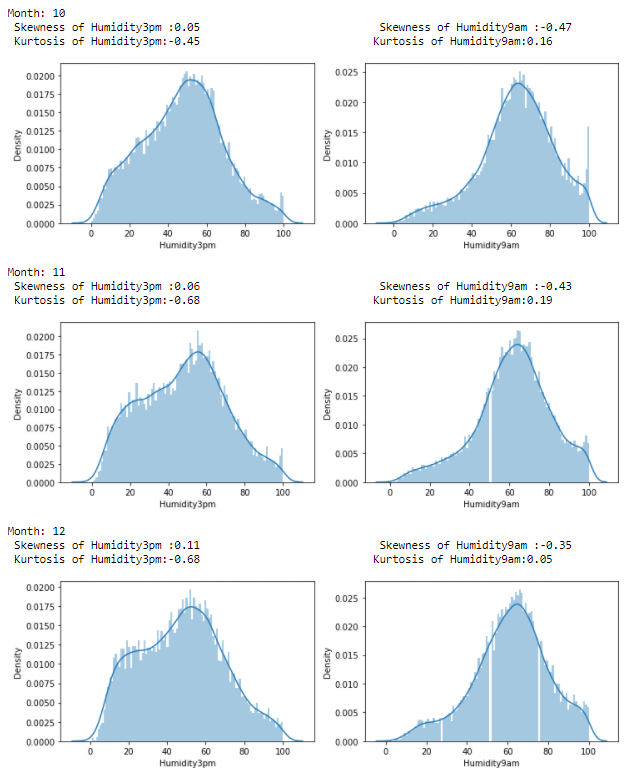
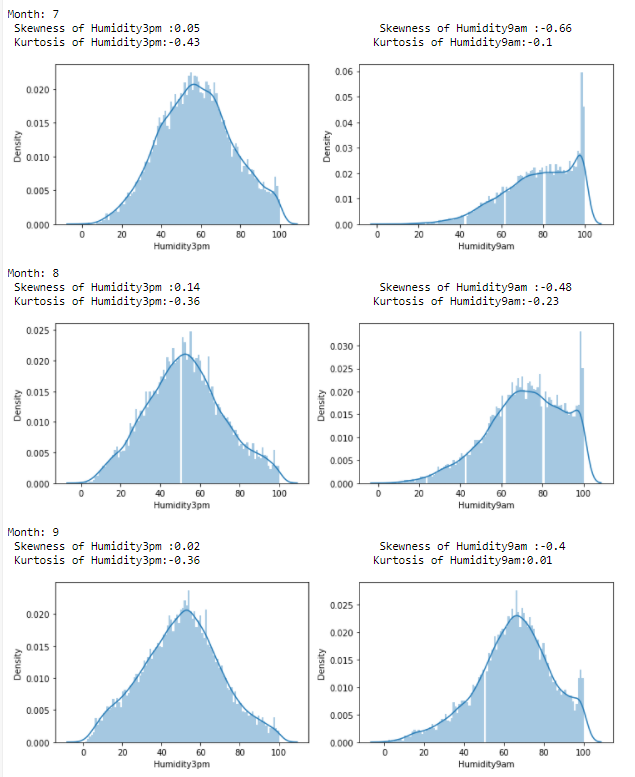


On the left, we can see the histogram of Humidity3pm feature and on the right we can see the histogram of Humidity9am feature.

* **DISTRIBUTION OF BOTH HUMIDITY FEATUURES(MONTHWISE)?**

In the below graphs, we have created distributions of our features with respect to every month.

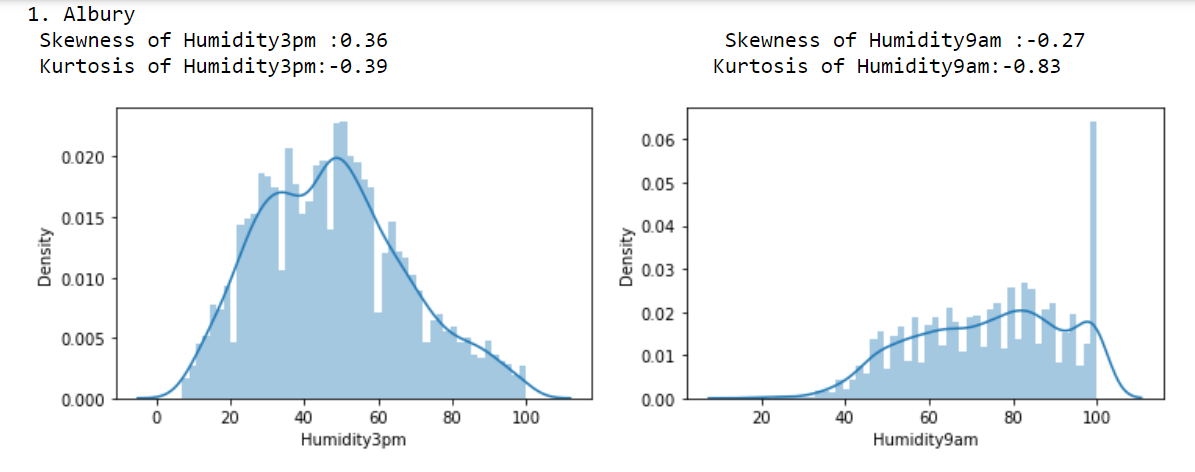
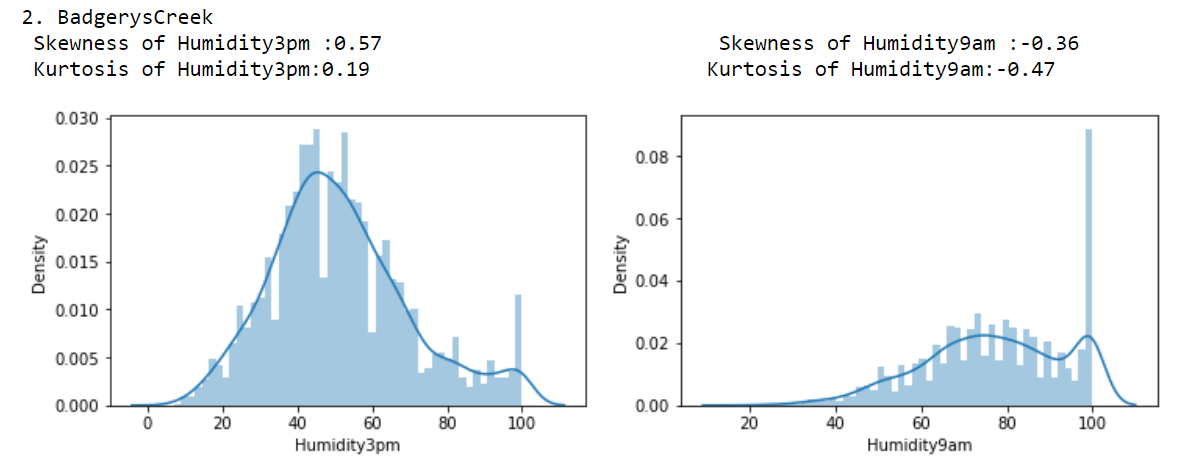




* **DISTRIBUTION OF BOTH HUMIDITY FEATURES (LOCATION WISE)?**

We have also created distributions with respect to different locations as well.

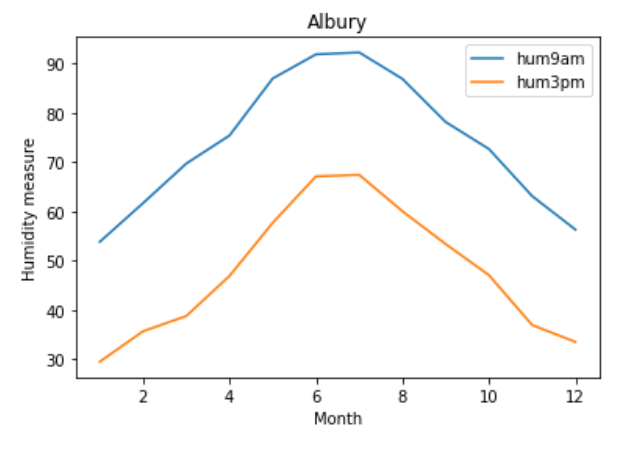
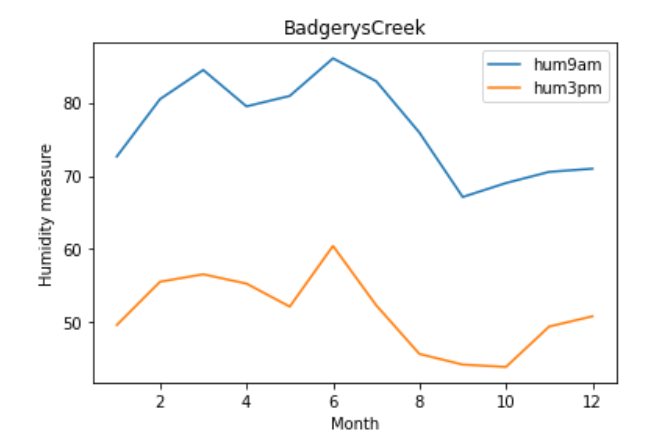
But due to a lot of unique features (49 unique locations), we are just going to show a sample of 2-3 locations.

So here we have shown the distributions of locations Albury and BadgerysCreek. Similarly, we have created distributions for different locations.

* **FURTHER ANALYSIS**

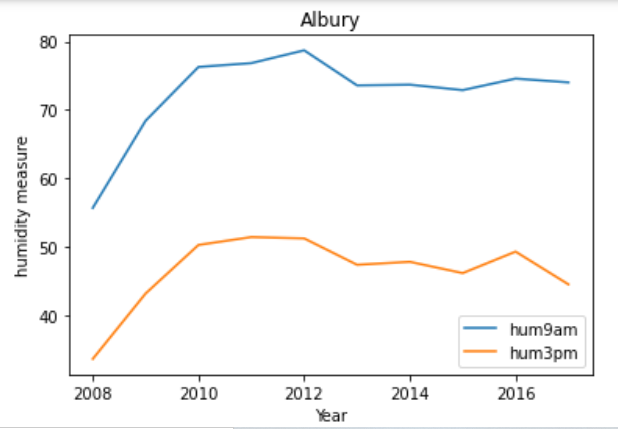
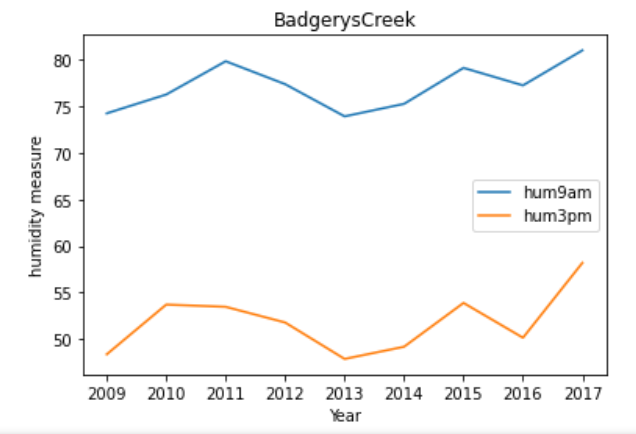
Here, we have created few more analysis for our data. With respect to every location, we took mean of both humidity9am and humidity3pm (both with respect to mean) and with that we have plotted for every location.

Here we can see on the graph that on the x axis we have the month wise label and on y axis we have the mean humidity measure. From the above graph we can clearly see that both these features are behaving identical which is what we need in these features.

It is quite obvious that since there are 49 unique locations, it is quite difficult and lengthy in order to show every location. So, we took two different locations and portrayed here.

The same procedure has been done year wise as well.

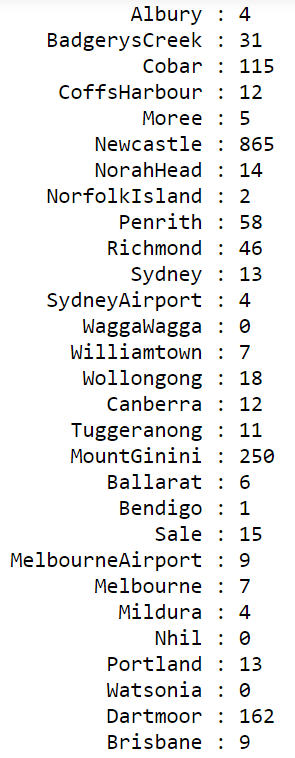
 

Here we can see on the graph that on the x axis we have the year wise label and on y axis we have the mean humidity measure. From the above graph we can clearly see that both these features are behaving identical which is what we need in these features.

* **WHAT ABOUT THE MISSING VALUES?**

Here we are checking the number of missing values for the respective locations.

**For Humidity3pm**

**For Humidity9am**

