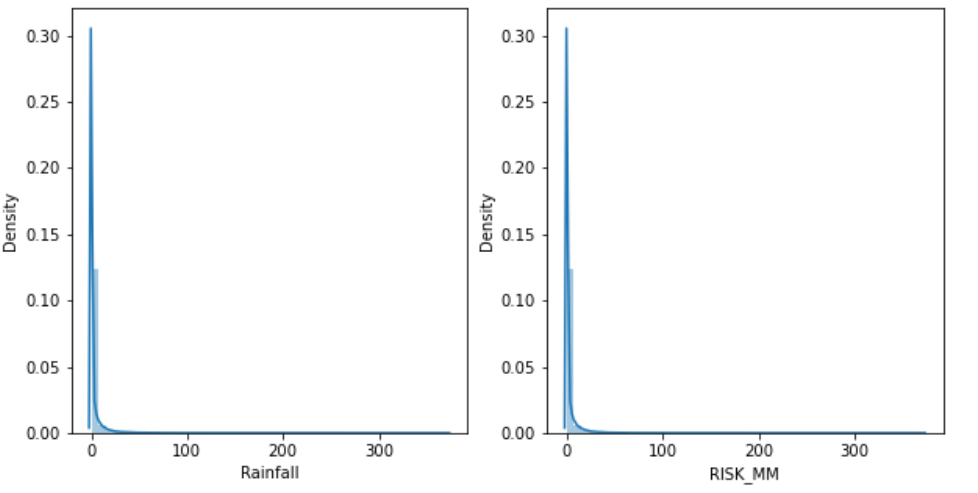
1. **Meeting 1: - (29thMay, 2021)**

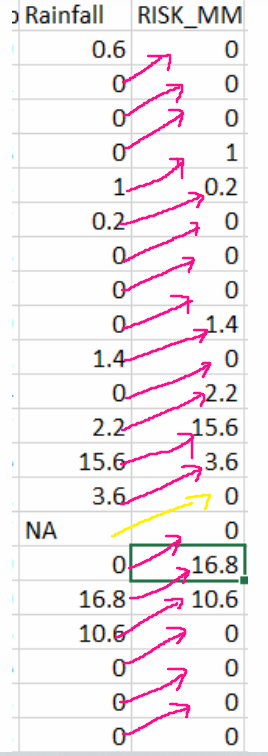
We found out in this meeting that both Rainfall and RISK\_MM features in our data are basically same. For that we can prove them by using Excel and we also saw the distplot for the same and we are getting the same graph as well which gives us enough evidence that both these features are same. So, we 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.

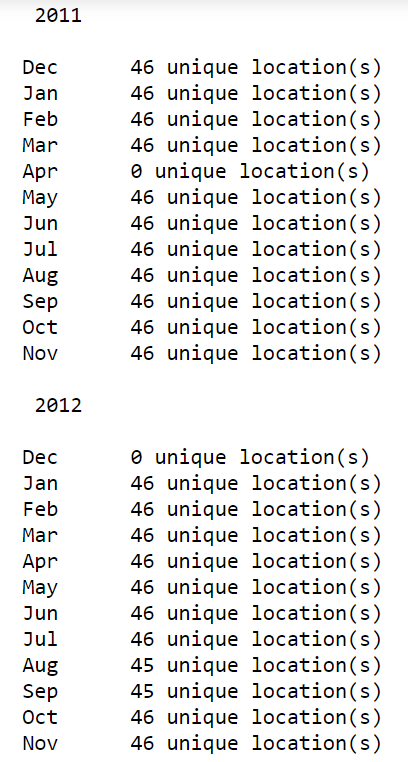
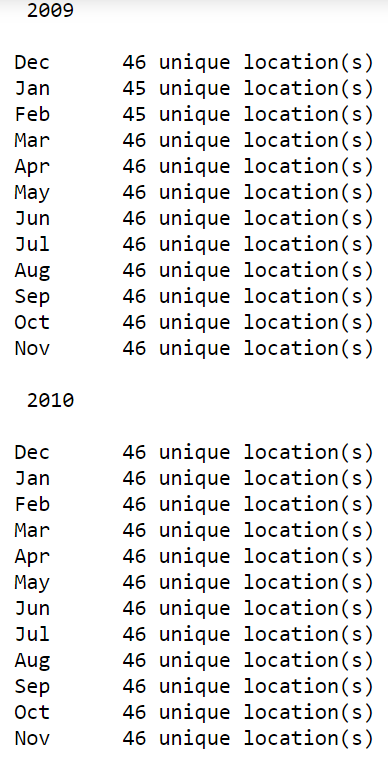
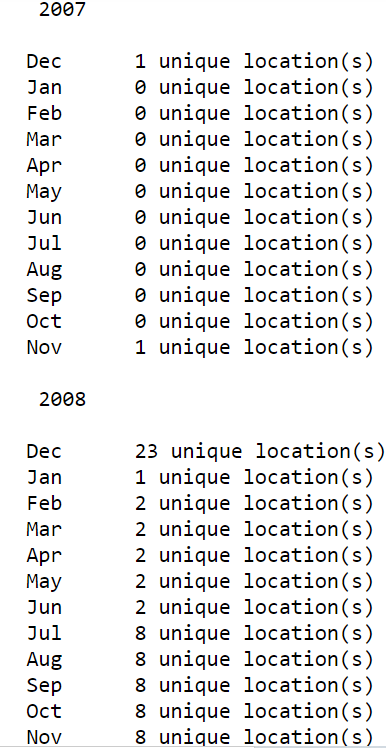
EXCEL proof for both Rainfall and RISK\_MM features:-

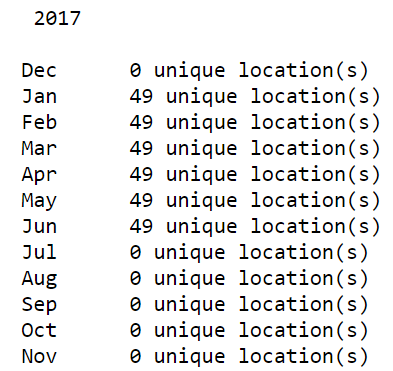
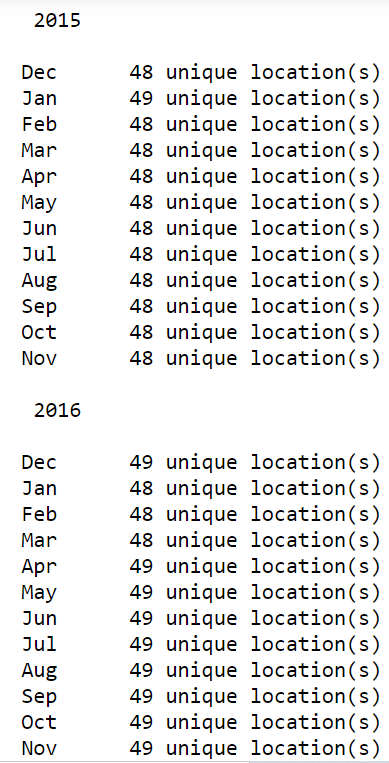
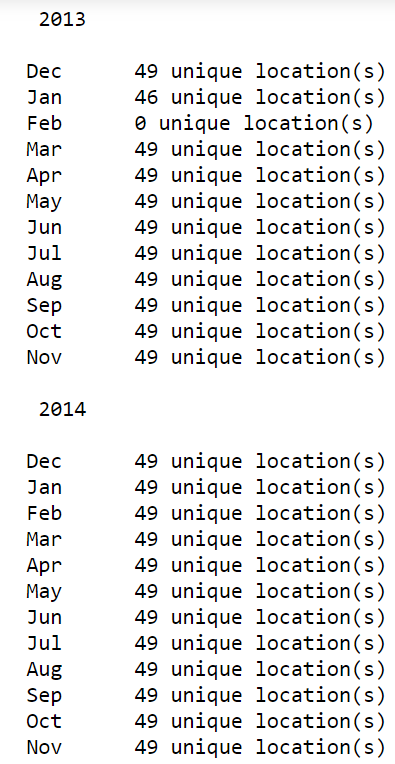


This image basically sums up what we want to try to convey about the data.

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.



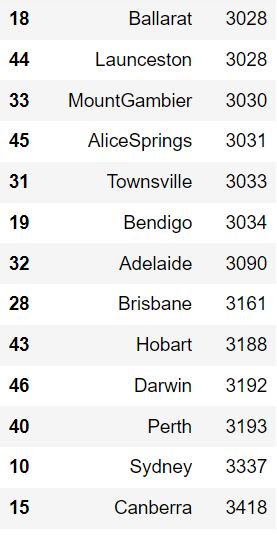
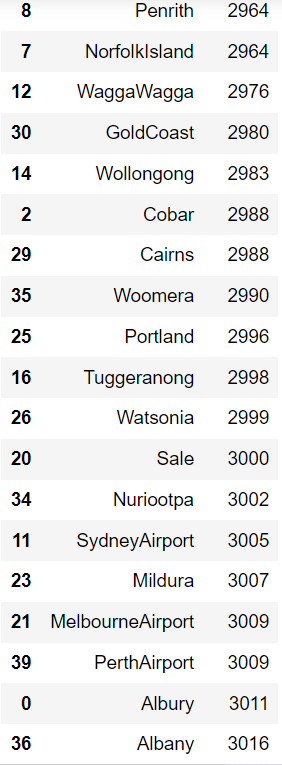
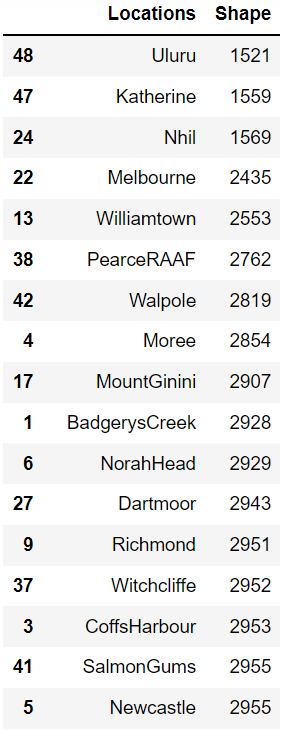


From the above graph, we can say that our actual data starts **from November 2007** and ends **with June 2017** with uneven data recorded for each month. Here the data of locations are not uniform.

Doubt here – The data for only 1 location is available in 2007 which makes 61 rows of the data and similarly there less data available for 2008 out of 49 total locations, which makes 2246 rows of the data. In total, these are 2307 rows which makes about 1.6% of the rows. Now since the entire data is not available, and we are not supposed to increase the number of rows, so can we delete the rows?

* 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.
* On plotting “MinTemp” and “MaxTemp” with respect to year for each location, we noticed that in most of the cases the both the temperature starts decreasing, reaches its lowest value at 2012 and then increases.

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.

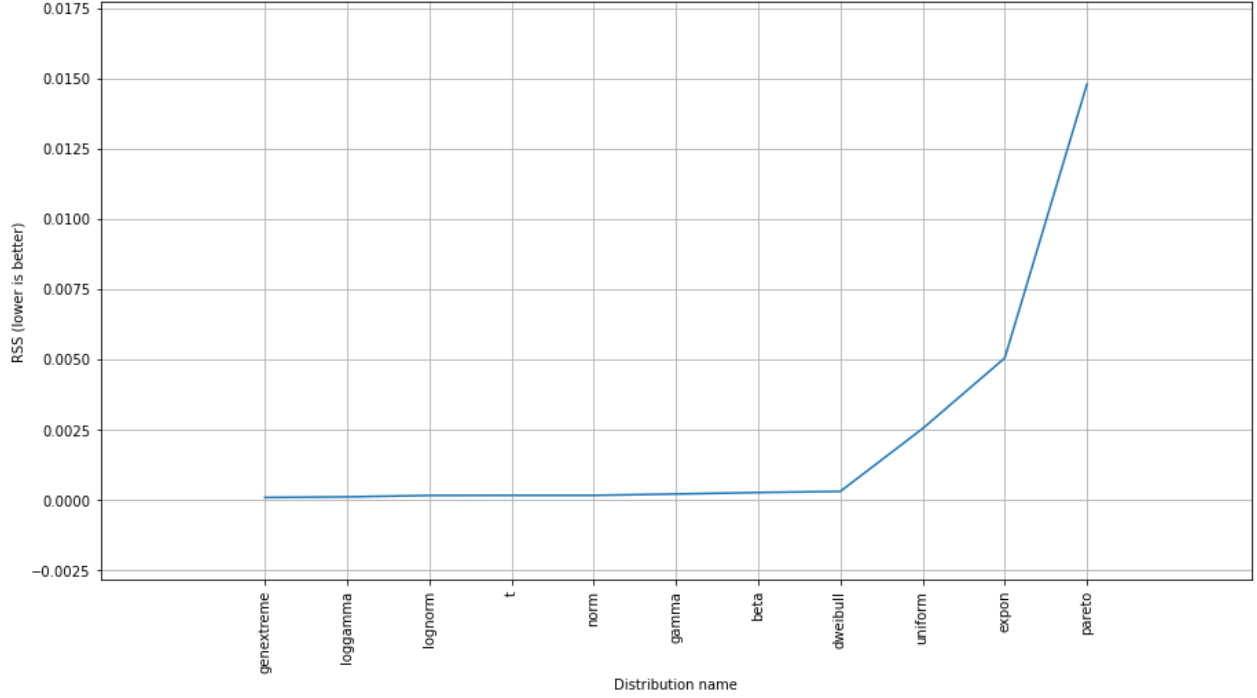
* **HOW IS IT DIFFERNENT FROM HUMIDITY?**

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?**

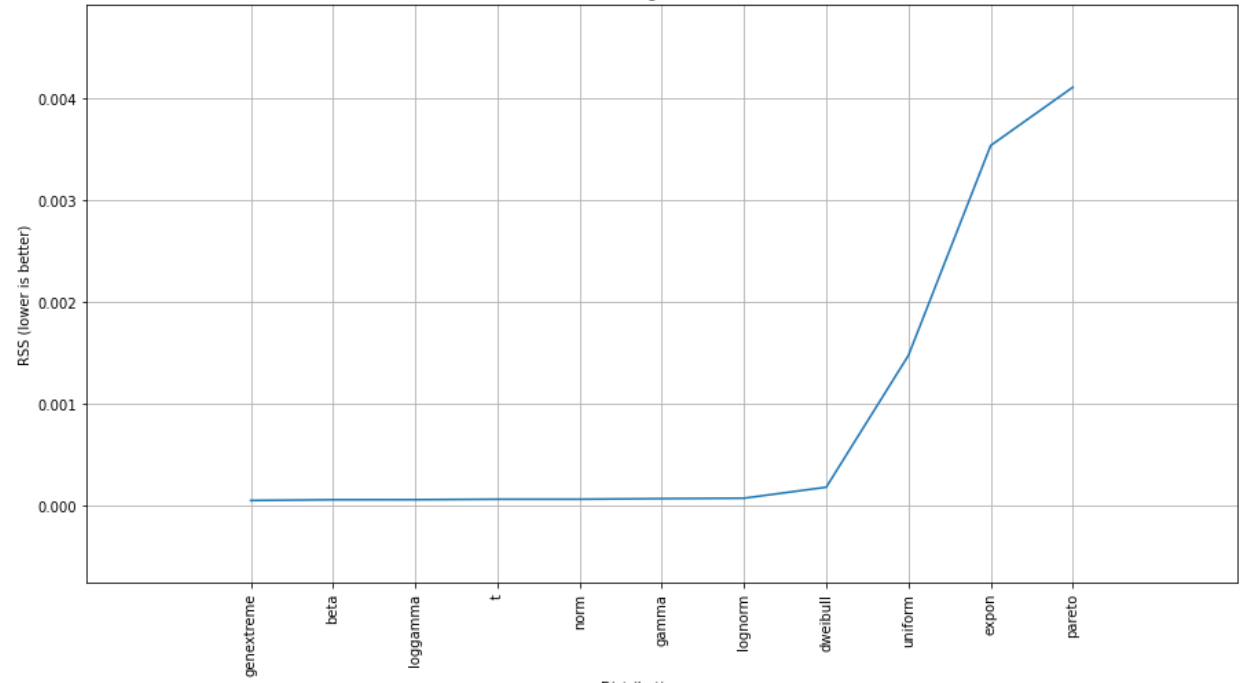
For the feature **Humidity9am**, we can see the graph below in order to understand what all distributions are quite available for us and which one is the best fit.



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.

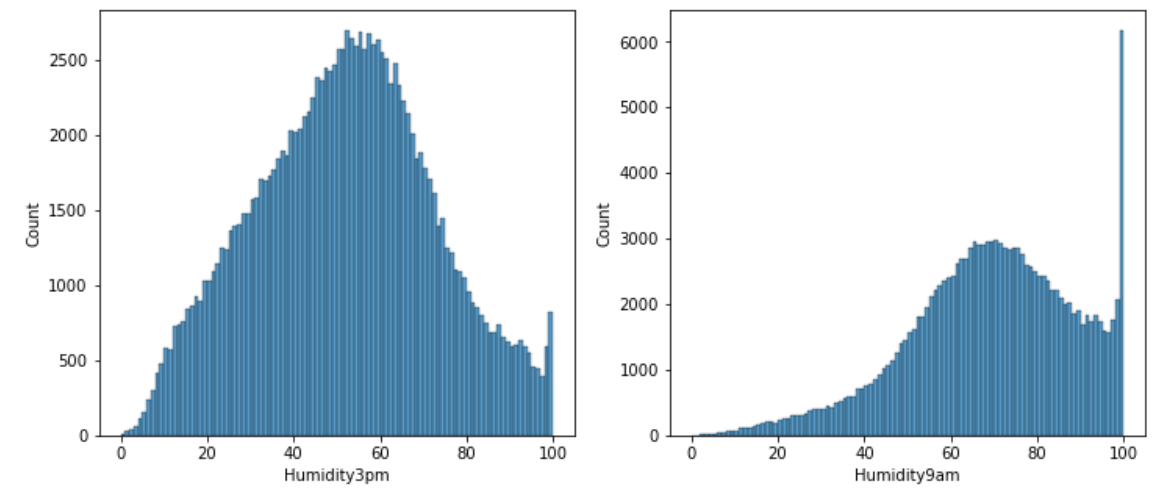
For the feature **Humidity3pm**, we can see the graph below in order to understand what all distributions are quite available for us and which one is the best fit.



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.

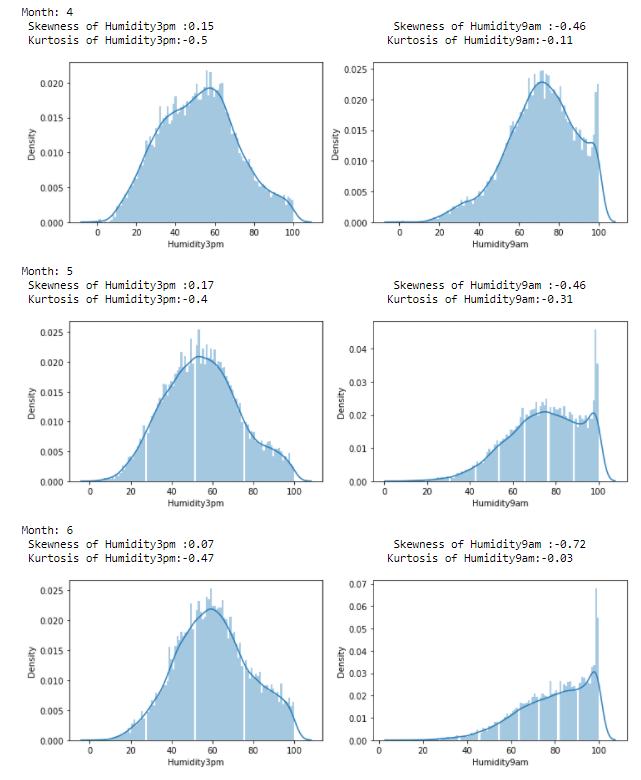
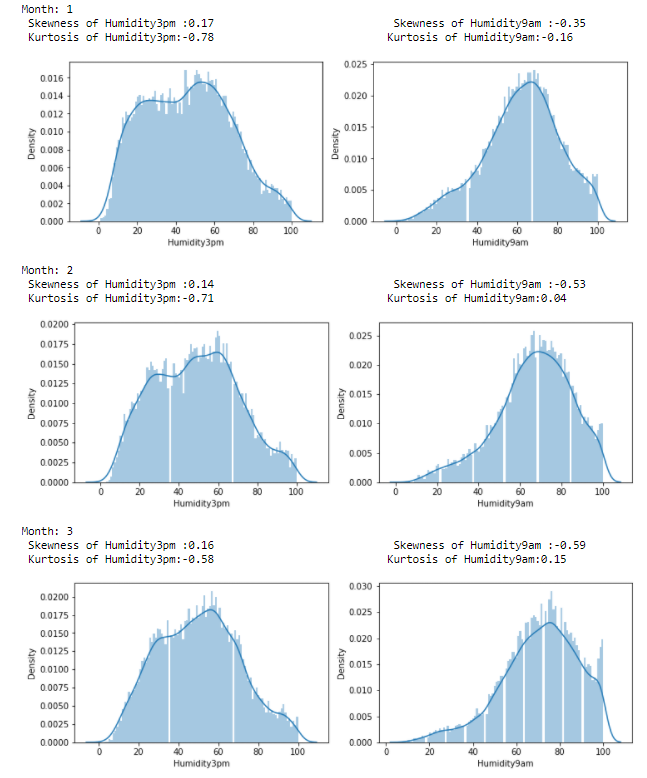
* **DISTRIBUTION PLOT OF THE HUMIDITY FEATURE?**

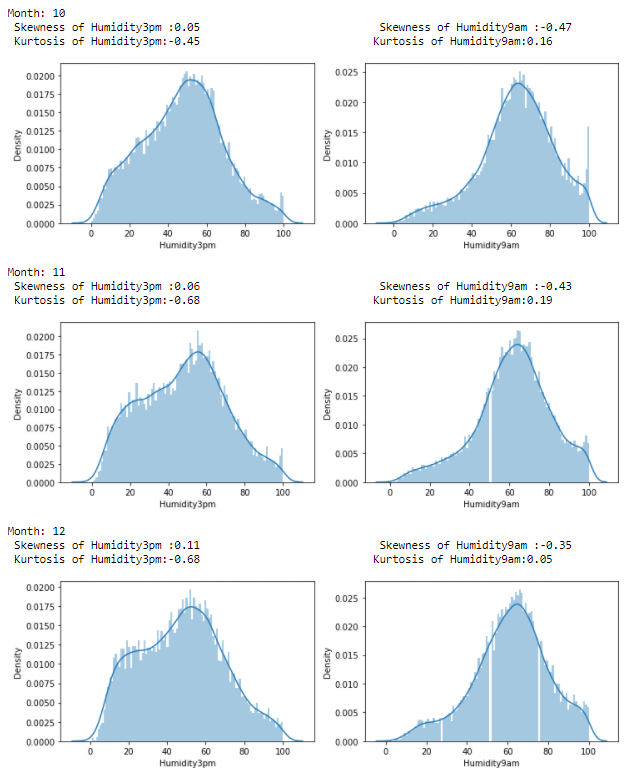
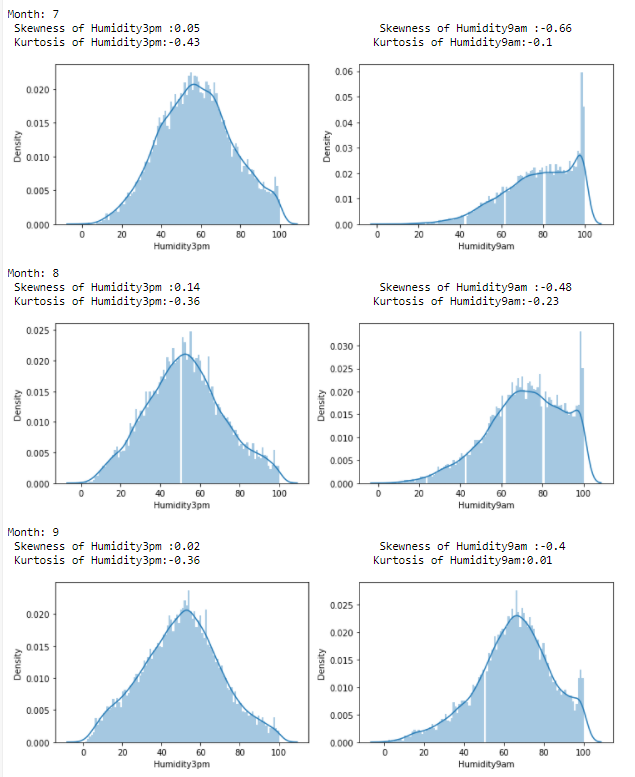


On the left, we can see the histogram of Humidity3pm feature and on the right we can see the histogram of Humidity9am feature.

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

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

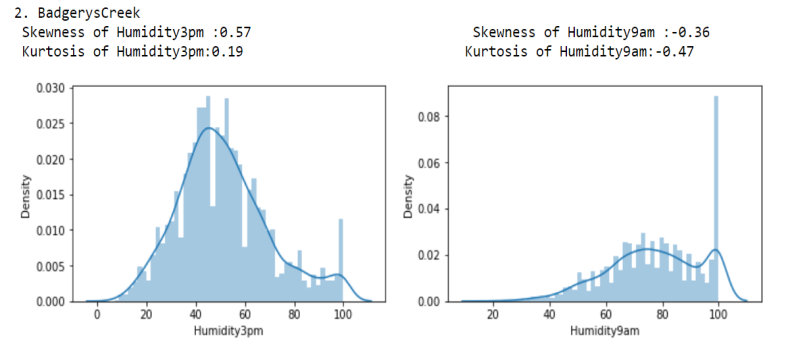
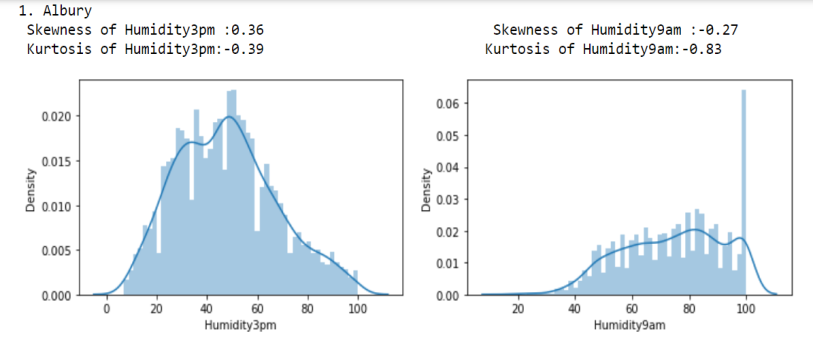




* **DISTRIBUTION OF BOTH HUMIDITY FEATURES(LOCATION WISE)?**

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

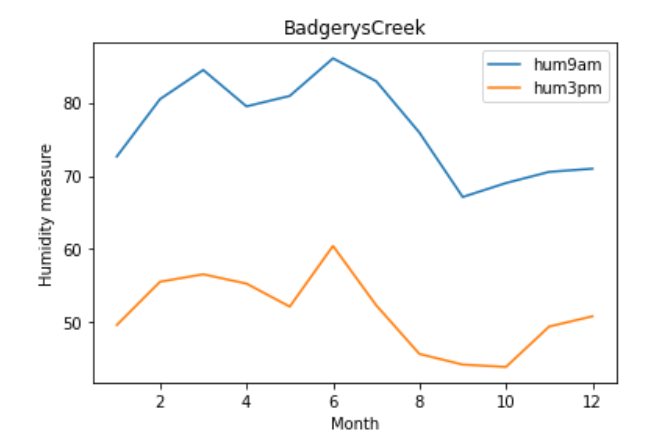
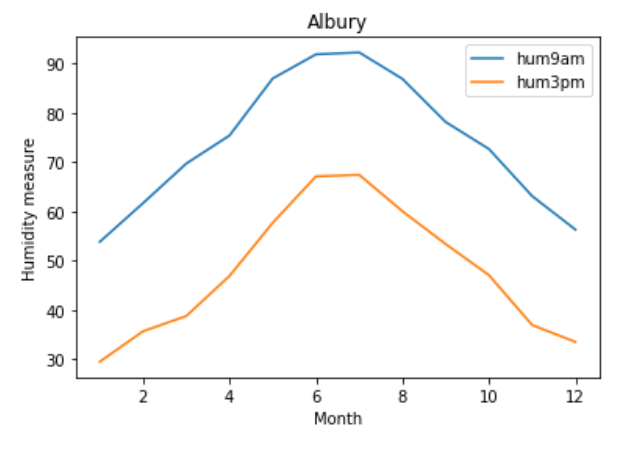
But due to a lot of unique features (49 unique locations), we are just going to show a sample of 2-3 locations.



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

* **FURTHER ANALYSIS**

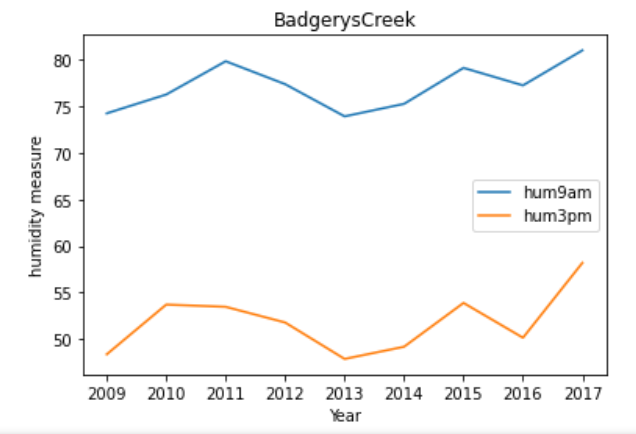
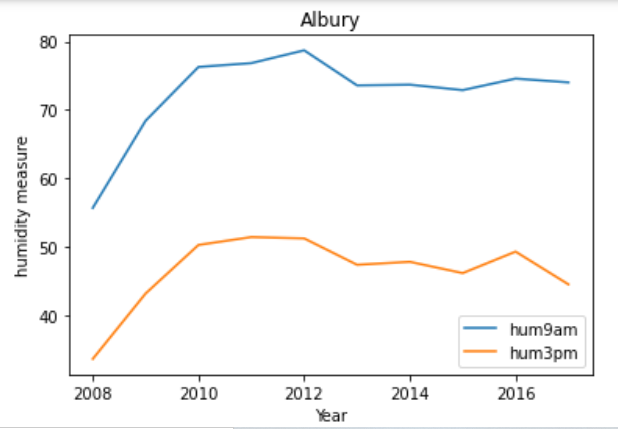
Here, we have created few more analysis for our data. With respect to every location, we took mean of both humidity9am and humidity3pm (both with respect to mean) and with that we have plotted for every location.



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.

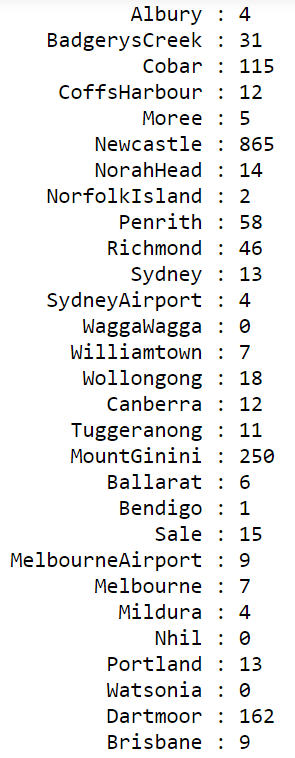


Here we can see on the graph that on the x axis we have the year wise label and on y axis we have the mean humidity measure. From the above graph we can clearly see that both these features are behaving identical which is what we need in these features.

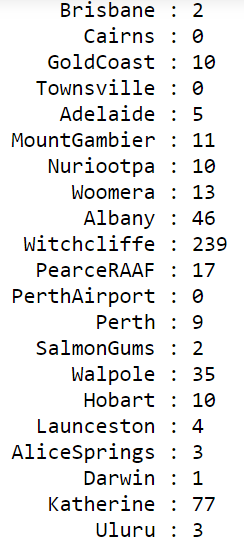
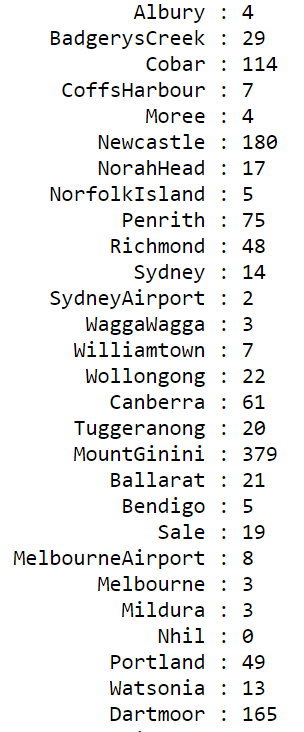
* **WHAT ABOUT THE MISSING VALUES?**

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

**For Humidity3pm**



**For Humidity9am**



1. ***Pressure9am and Pressure3pm***

These variables measure the atmospheric pressure at 9am and 3pm respectively observed at the mean sea level. The values are given in hectopascals ( or millibar ). Hectopascals is the SI unit for measuring atmospheric pressure.

The average pressure at mean sea-level in the International Standard Atmosphere is 1013.25 hPa ~ 1 atm ~ 29.92 inches of mercury.

As air rises in low-pressure systems, it cools and often condenses into clouds and precipitation, resulting in storms. In high-pressure systems, the air sinks toward the Earth and warms upward, leading to dry and fair weather.

**Predicting weather with a Barometer**

* High Pressure

A barometric reading over 30.20 inHg is generally considered high, and high pressure is associated with clear skies and calm weather.

If the reading is over 30.20 inHg (102268.9 Pa or 1022.689 mb):

* Rising or steady pressure means continued fair weather.
* Slowly falling pressure means fair weather.
* Rapidly falling pressure means cloudy and warmer conditions.
* Normal Pressure

A barometric reading in the range of 29.80 and 30.20 inHg can be considered normal, and normal pressure is associated with steady weather.

If the reading falls between 29.80 and 30.20 inHg (100914.4–102268.9 Pa or 1022.689–1009.144 mb):

* Rising or steady pressure means present conditions will continue.
* Slowly falling pressure means little change in the weather.
* Rapidly falling pressure means that rain is likely, or snow if it is cold enough.
* Low Pressure

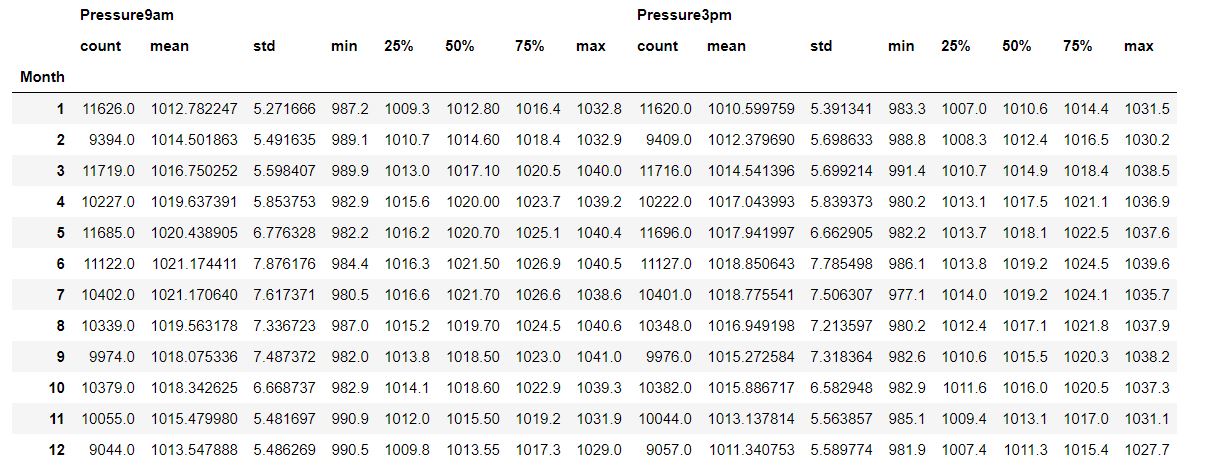
A barometric reading below 29.80 inHg is generally considered low, and low pressure is associated with warm air and rainstorms.

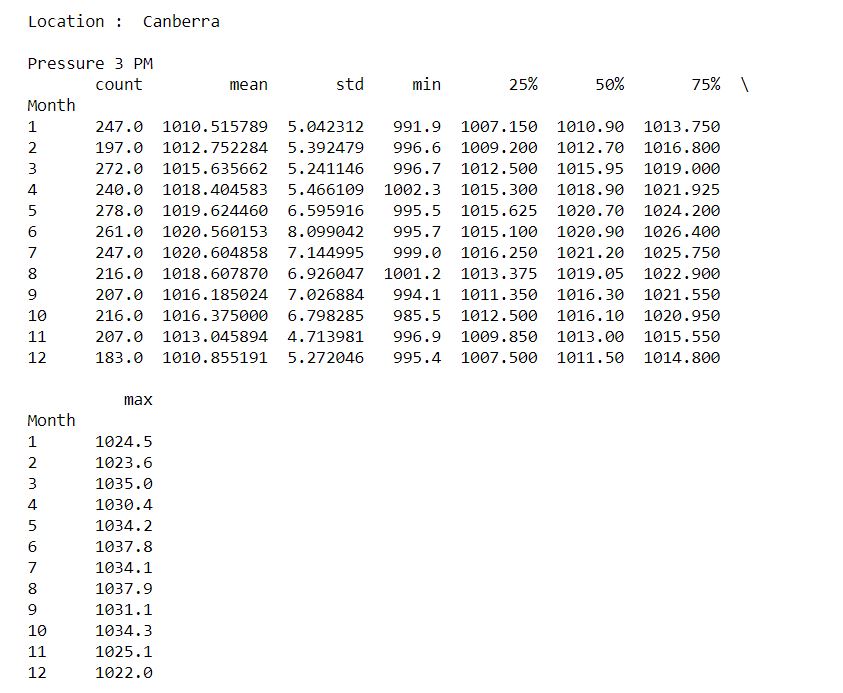
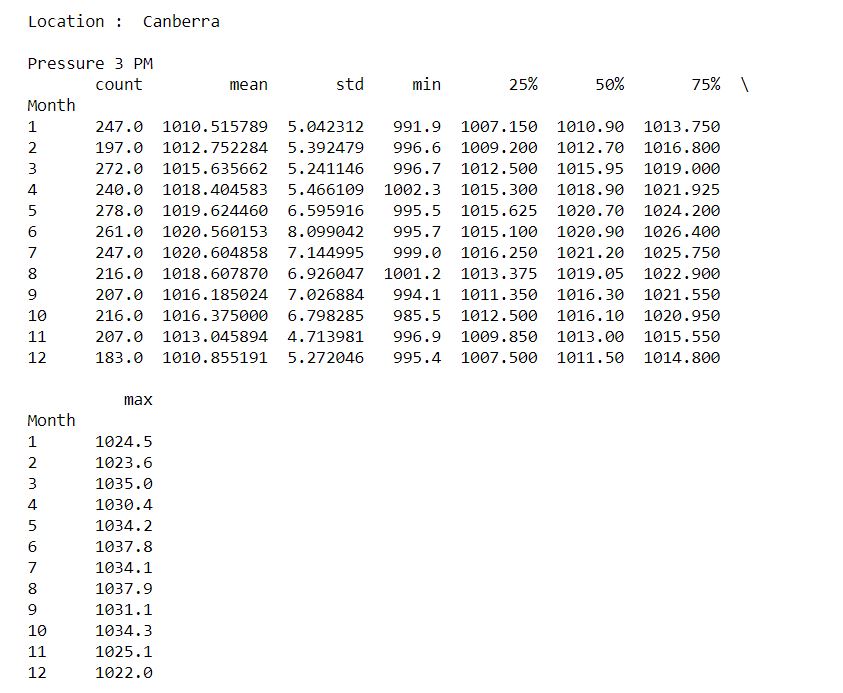
If the reading is under 29.80 inHg (100914.4 Pa or 1009.144 mb):

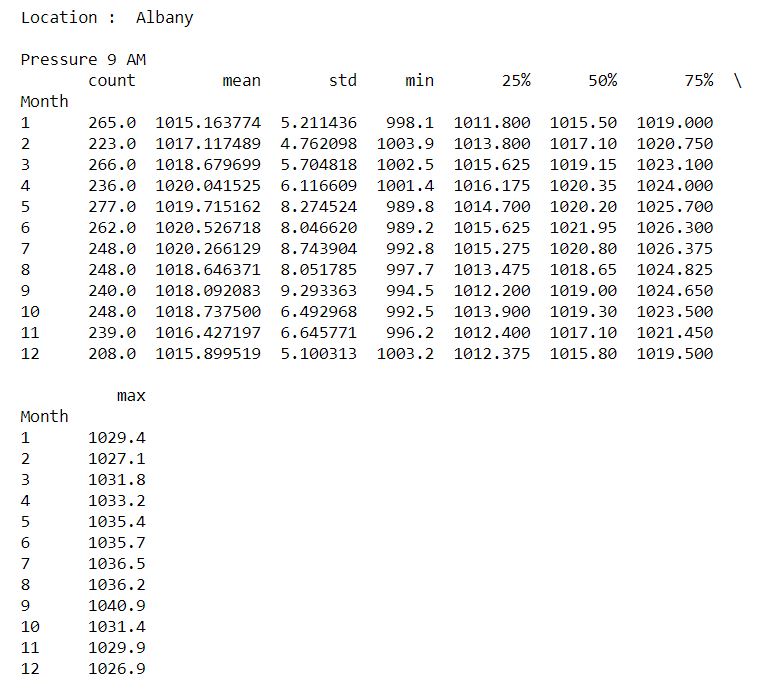
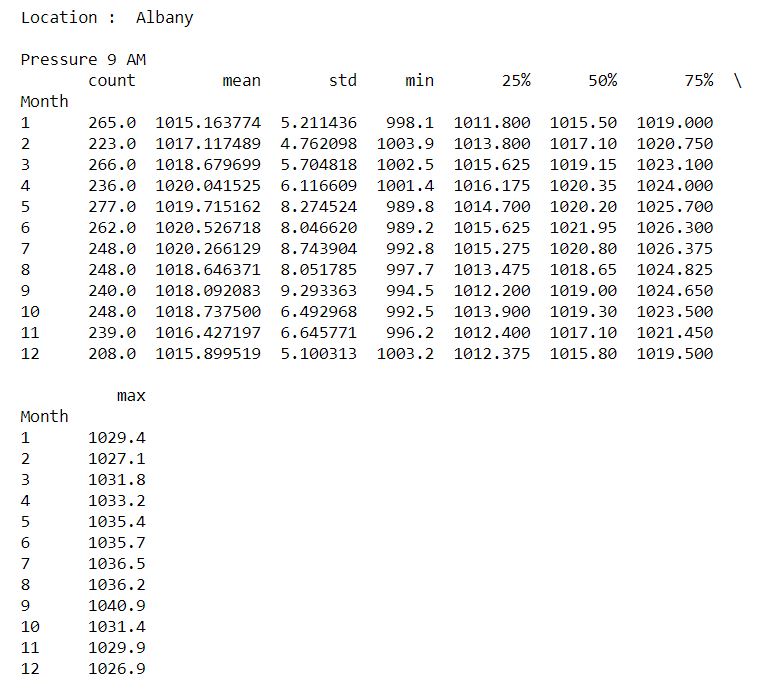
* Rising or steady pressure indicates clearing and cooler weather.
* Slowly falling pressure indicates rain.
* Rapidly falling pressure indicates a storm is coming.

Understanding the Pressure data :

We observe the patterns in the pressure features over the months and seasons and also for different locations :

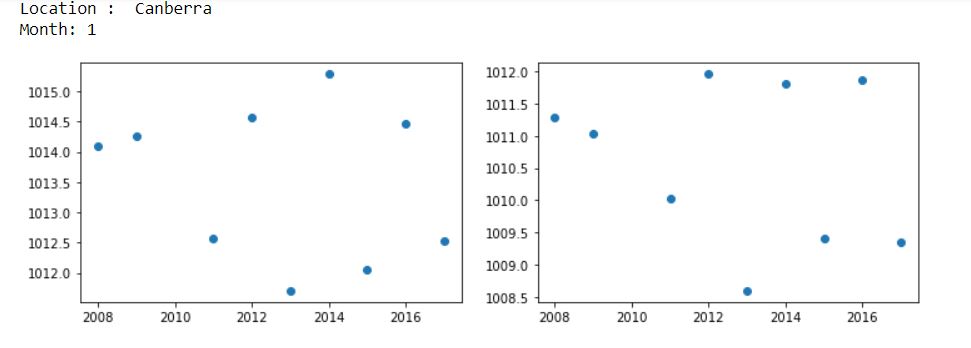






These tables show the monthwise values for pressure features for two different locations. Similarly we can observe these for other locations as well.

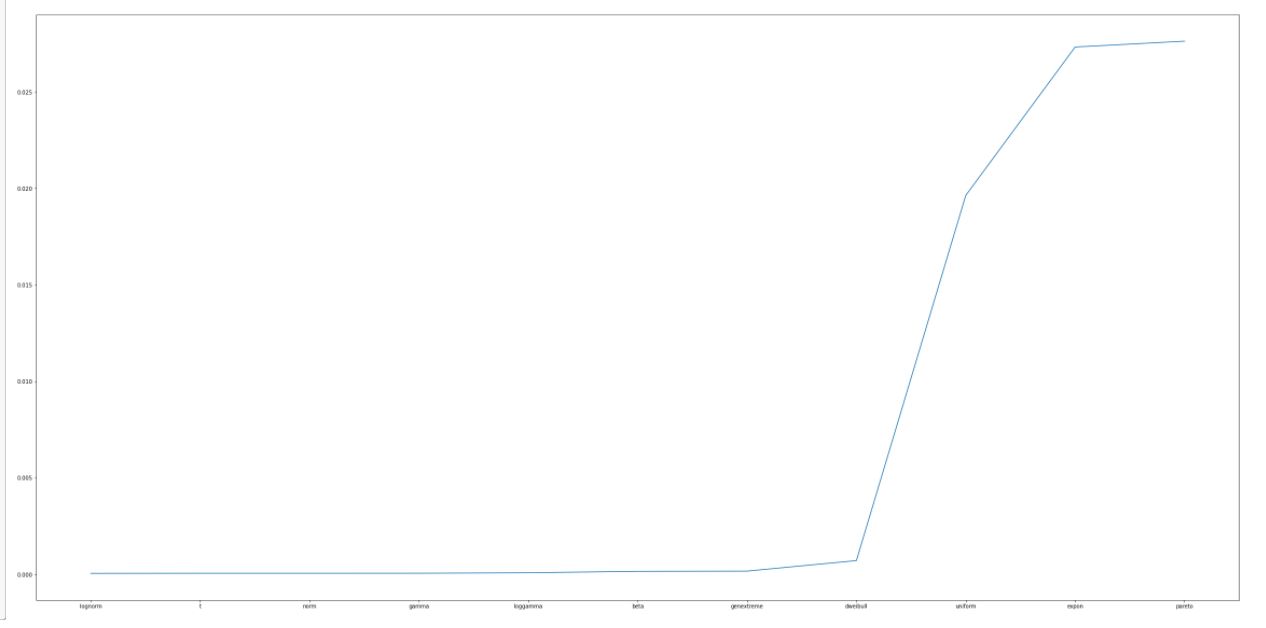
The values for Pressure in a location for different months is more or less similar. In fact, over the years, there hasn’t been much change in the Pressure levels for a given location in a month. The following figure for Canberra shows the same. We can see this for other locations and months as well. The range of variation of Pressure9am in January in Canberra are similar and in the normal pressure category – (1009-1022 hPa).



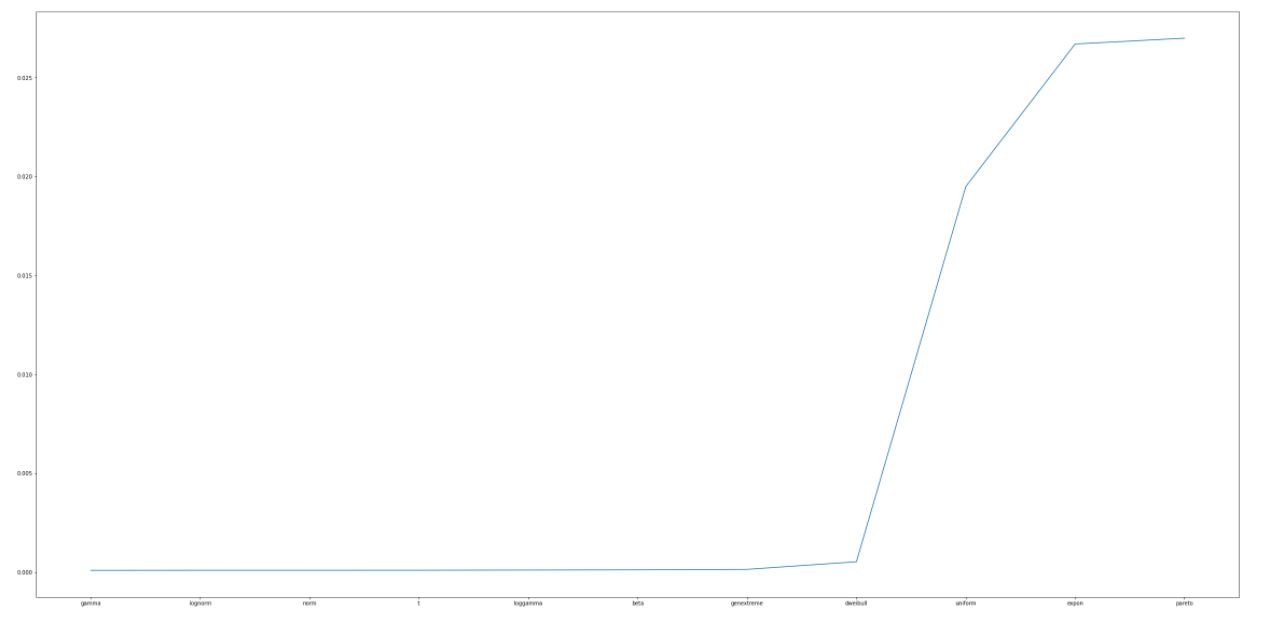
* **WHAT IS THE DISTRIBUTION OF THE DATA?**

From the following figures and graphs we can see that the Pressure9am and Pressure3pm columns follow Normal distribution as they have the lowest value for RSS.

Pressure 9am :



Pressure 3pm :

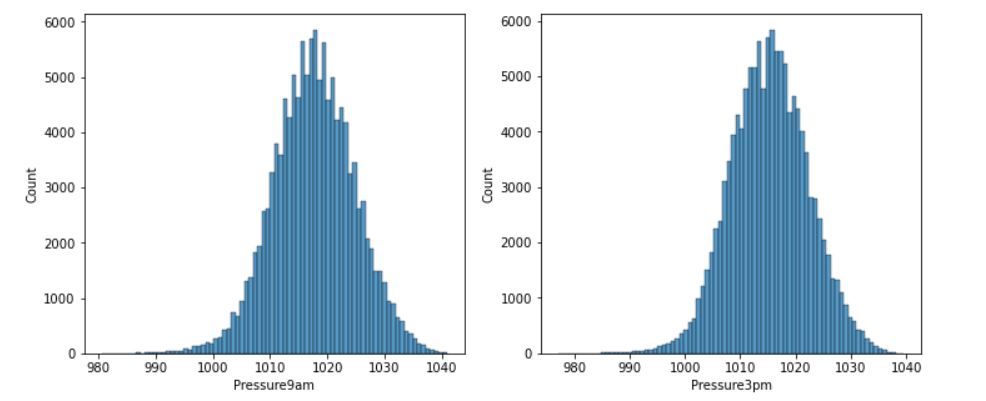


Here in this graph, we can see that on the x-axis we have the names of the distributions and on y-axis we have the RSS (Residual Sum of Squares).

Here from the graph of Pressure9am, we see that lognorm distribution has the lowest RSS value and for Pressure 3pm, gamma distribution has the lowest value for RSS. Also, there isn’t much difference between their and Normal distribution RSS, so we can say that they follow Normal distribution.

Hence, we can conclude that the Pressure variables follow Normal distribution.

* **DISTRIBUTION PLOT OF THE PRESSURE FEATURE?**



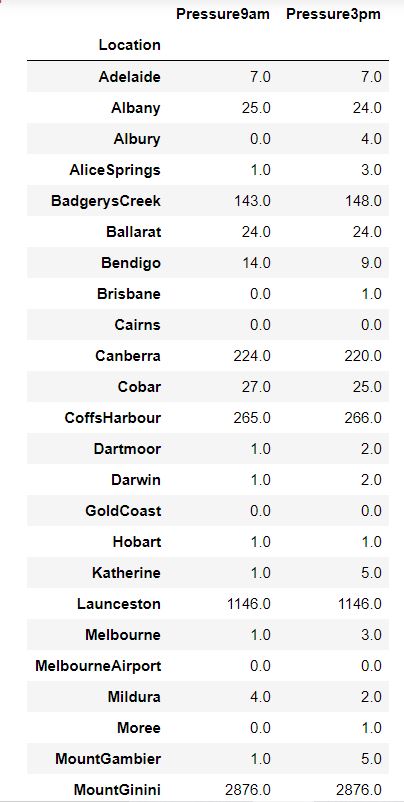
We observe from the above graphs, that the Pressure features follow Normal distribution.

Also, when we observe these variables for different months, they are identical to quite an extent.

Similarly, for different locations also, the features behave in a similar manner.

* **WHAT ABOUT THE MISSING VALUES?**

Following table shows the null values in the pressure features for different locations:

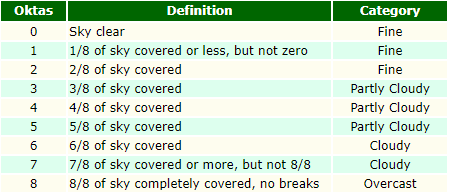
****

1. ***Cloud 9am and Cloud3pm***

Fraction of sky obscured by clouds at 9 am and 3pm respectively. This is measured in 'oktas' which are a unit of eights.It

records how many eights of the sky are obscured by cloud.

A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

****

* By analysing the missing values of these two features,it was found that there are **12** cities for which

we have no record for Cloud9am and Cloud3pm for any year. These cities are:



Possible way to impute the values for these cities:

We can refer to a map to find neighbouring locations around these 12 locations.If any of them is present in our

dataset and has sufficient information on our concerned features,then we can use mean/median of that

location to fill in our missing values month wise.

* As we know,Cloud9am and Cloud3pm are features with integral values in the interval [0,8],thus it is very

clear that any value below 0 or above 8 is considered an erroneous data point.

On inspecting our data,we found 2 such erroneous data points in Cloud9am in the following locations:

Sydney on 23rd Sept 2009 and Canberra on 27th May 2012

Similiarly,we found one erroneous data point in the Cloud3pm in Woomera on 2nd Nov 2012.

In both the cases,data points were 9.

How to deal with these data points?

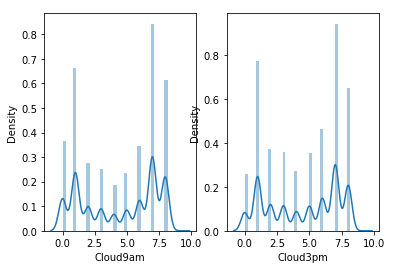
Since they appear in just 3 observations,we assume that it might have been entered by mistake instead of an 8

for a sky completely outcasted by clouds.

Hence,we intend to replace these 9s by 8s .

* **Distribution of the features**

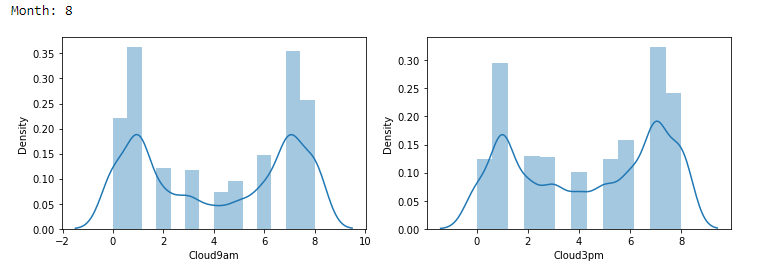
The following plot shows the distribution of these features based on overall dataset



Distribution of the features month wise:

The plot displayed here is for the month of August.We can see that the distribution of both of the features is **bimodal** and it is clear from the plot that majority of the values recorded are towards the extreme values.

This is a common insight for almost every month.

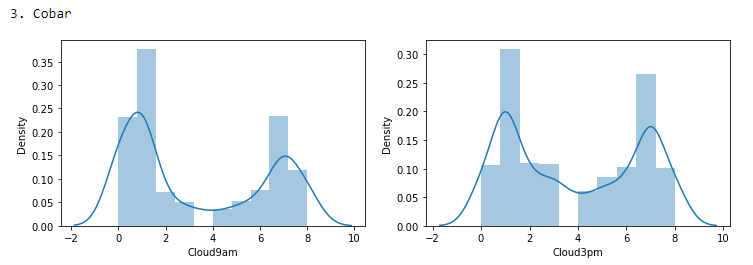


Distribution of the features location wise:

The plot displayed here is for the loaction of Cobar.We can see that even location wise, distribution of both of

the features is **bimodal** and most of the values recorded are towards the extreme values.And here,bimodality

of distribution is observed for most of the locations.



1. ***Temp9am and Temp3am***

* These two variables behaved very similar to the variables “MinTemp” and “MaxTemp” which is pretty obvious.
* On plotting these variables for each month we obained an almost normal curve for each month but different distribution when plotted against locations.
* Average value of Temp3am always came out to be greater than Temp9am.