Sub-Project 1
Data Visualisation and Statistics from Data

Anmol Bishnoi
IIT Mandi, b19069@students.iitmandi.ac.in
7042845211

QUESTION 1

Mean, median, mode, minimum, maximum and standard deviation for all the attributes (excluding dates and stationid).

Attributes	Mean	Median	Mode	Minimum	Maximum	Standard Deviation
Temperature	21.214888	22.272730	[12.727269999999999]	7.672900	31.375	4.353513
Humidity	83.479932	91.380950	[99.0]	31.000000	99.720	18.200427
Pressure	1009.008774	1014.677832	[789.3926923077]	452.097887	1079.162	46.955613
Rain	10701.538370	18.000000	[0.0]	0.000000	82037.250	24839.102466
Average Light	4438.428453	1656.880000	[4488.9103]	0.000000	54612.000	7569.154781
Maximum Light	21788.623280	6634.000000	[4000]	2259.000000	54612.000	22053.315399
Moisture	32.386053	16.704200	[0.0]	0.000000	100.000	33.635434

Output

I. Temperature (in degrees Celsius)

Mean	21.215
Median	22.273
Mode	12.727
Minimum	07.673
Maximum	31.375
Standard Deviation	04.354

The maximum and minimum values show a massive difference, physically impossible over a single month. The mode is closer to the minimum than the maximum and hence suggests the data being positively skewed.

II. Humidity (relative)

Mean	83.480
Median	91.381
Mode	99.000
Minimum	31.000
Maximum	99.720
Standard Deviation	18.200

The values are heavily skewed towards the maximum value and thus the data is negatively skewed. The mode is at 99 which makes us think that the data was collected when it was relatively humid in the region, probably in the monsoon season. The relatively higher mean and the small standard deviation also confirm our thoughts.

III. Pressure (atmospheric in millibars)

Mean	1009.009
Median	1014.678
Mode	789.393
Minimum	452.098
Maximum	1079.162
Standard Deviation	46.956

The mean at almost 1 Bar and the low standard deviation suggest most of the sensors are located at relatively lower heights. The sensor giving the Minimum value at almost 450mB should be a cause of concern though as this pressure is usually found at elevations of around 6000m whereas the hills surrounding our campus have a height of around 2000m.

IV. Rain (in millilitres)

Mean	10701.538
Median	18
Mode	0.0
Minimum	0.0
Maximum	82037.250
Standard Deviation	24839.102

Although the massive range and standard deviation look alarming, it is actually perfectly natural and explainable. There are days when it doesn't rain at all while there are days when it rains heavily. The median suggests that there were more days that saw little to no rain than the number of days when it rained a lot. While the mean suggests that on the days it actually rained, it did so very heavily.

V. Average Light (average in a day in lux)

Mean	4438.428
Median	1656.880
Mode	4488.910
Minimum	2259.000
Maximum	54612.000

Standard Deviation	7569.155

The range of values suggests that on some days, the skies would be relatively clearer while other days it would be relatively overcast throughout the day which perfectly extends on our theory that it would rain on some days heavily while it wouldn't rain at all on other days.

VI. Maximum Light (max in a day in lux)

Mean	21788.623
Median	6634.000
Mode	4000.000
Minimum	2259.000
Maximum	54612.000
Standard Deviation	22053.315

The minimum value of maximum lux in a day having the low value of 2259 suggests that on days when it would be cloudy and rainy, the clouds would remain overhead throughout the day and not let much sunlight fall on the sensors.

VII. Moisture (relative)

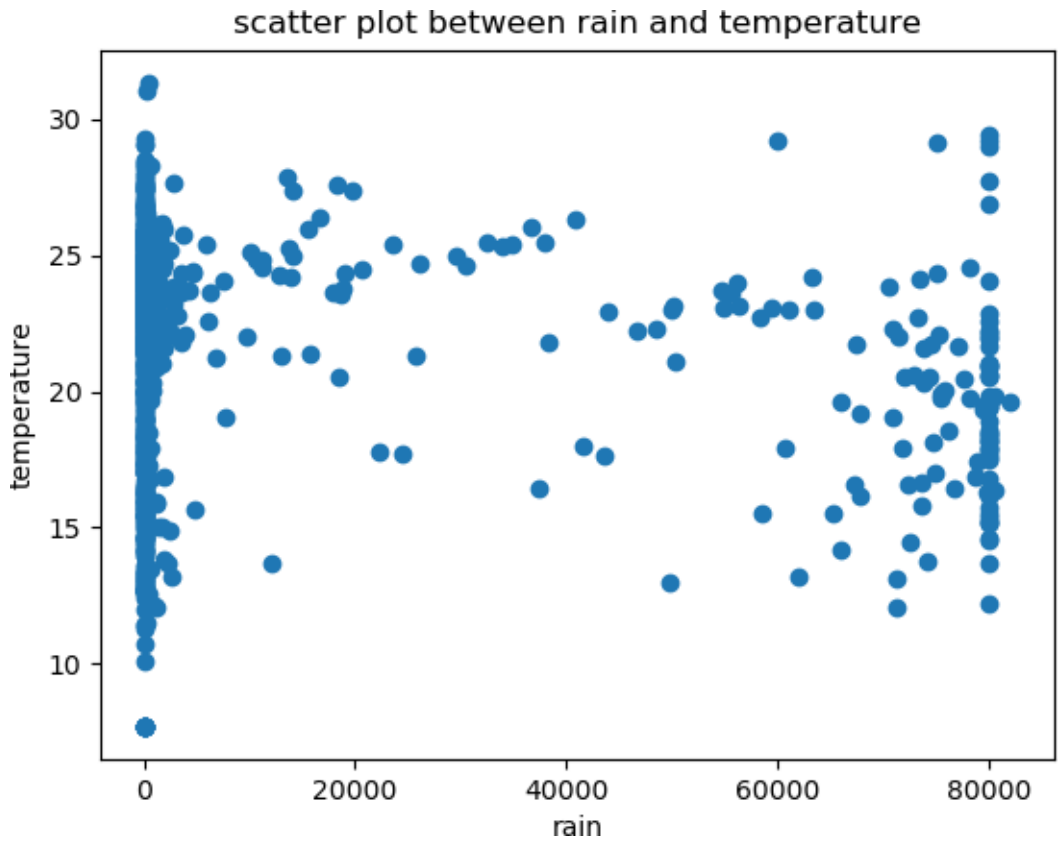
Mean	32.386
Median	16.704
Mode	0.000
Minimum	0.000
Maximum	100.000
Standard Deviation	33.635

The relatively lower average value of soil moisture even in the season when average rains were over 10 litres in a day suggests that the soil where such sensors are kept is bad at retaining moisture content and can get dry fairly quickly.

QUESTION 2A

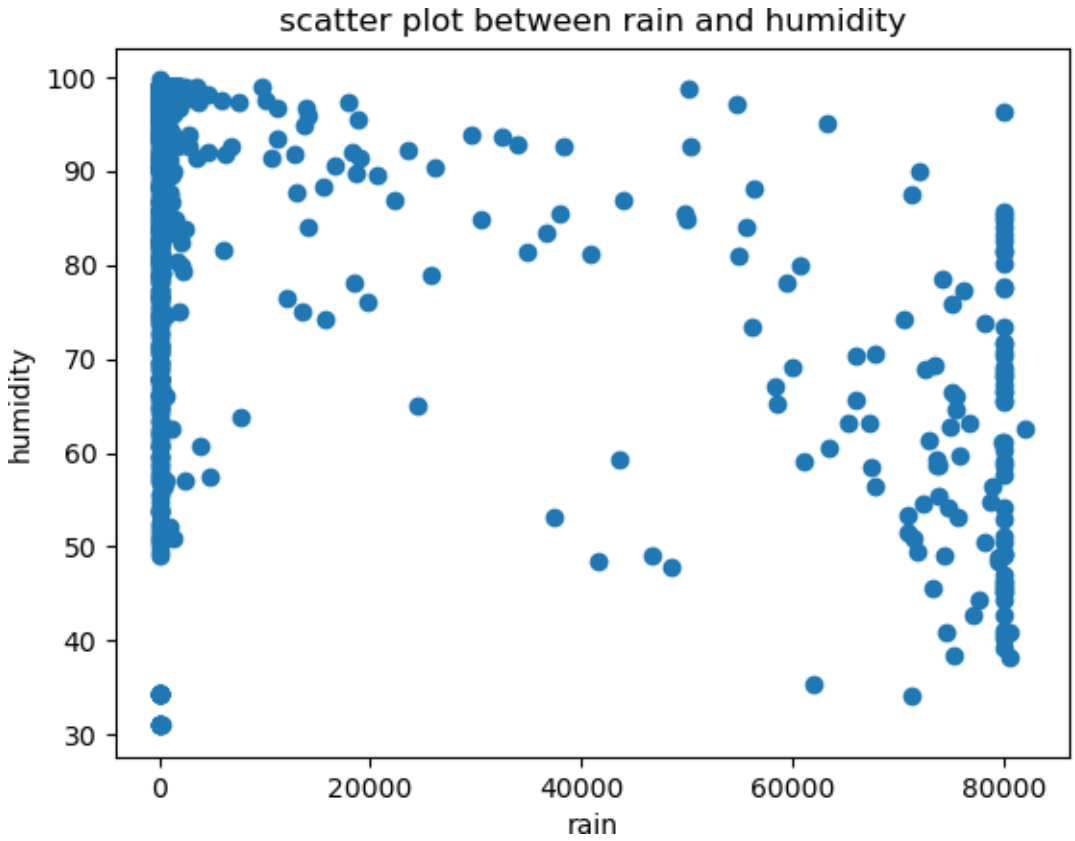
Obtain the scatter plot between 'rain' and each of the other attributes, excluding 'dates' and 'stationid'. Consider 'rain' in x-axis and other attributes in y-axis.

I. Temperature (in degrees Celsius)



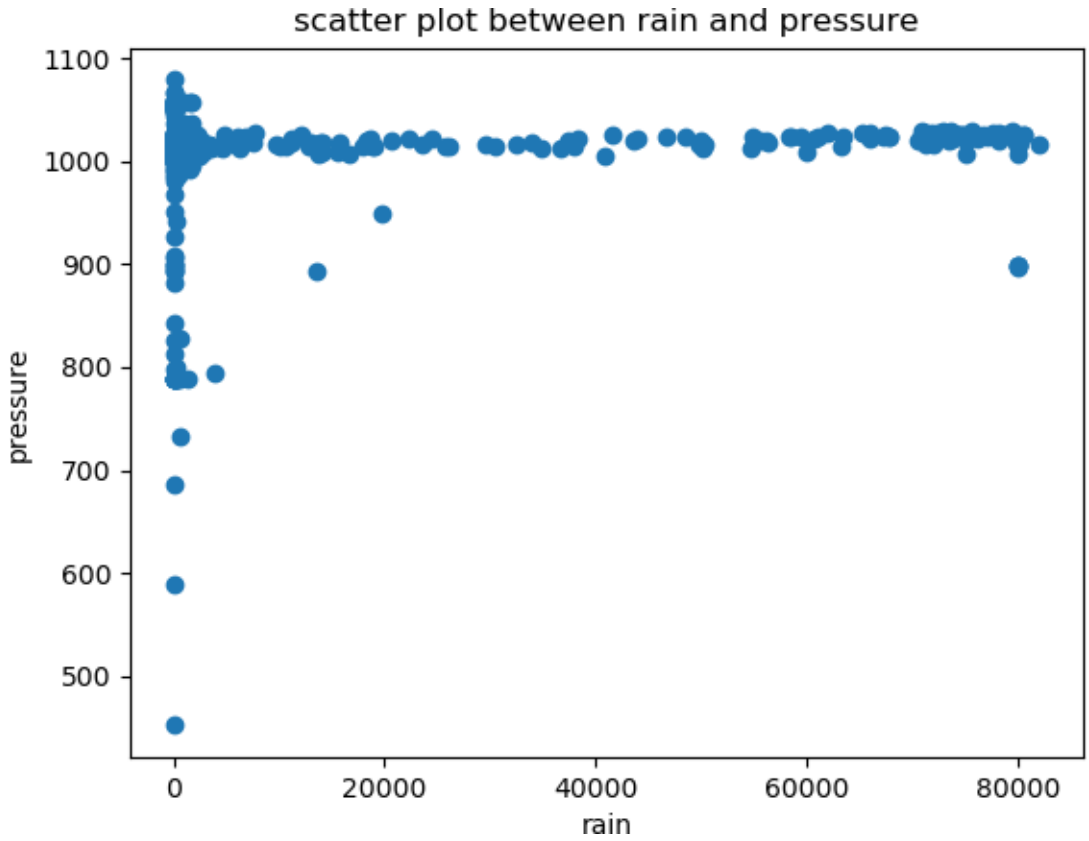
Although, the temperature recorded on the days when it did not rain, is spread over a very huge range, the cluster of points can be clearly observed to have moved towards lower temperatures on the days it rained which perfectly matches real life observations

II. Humidity (relative)



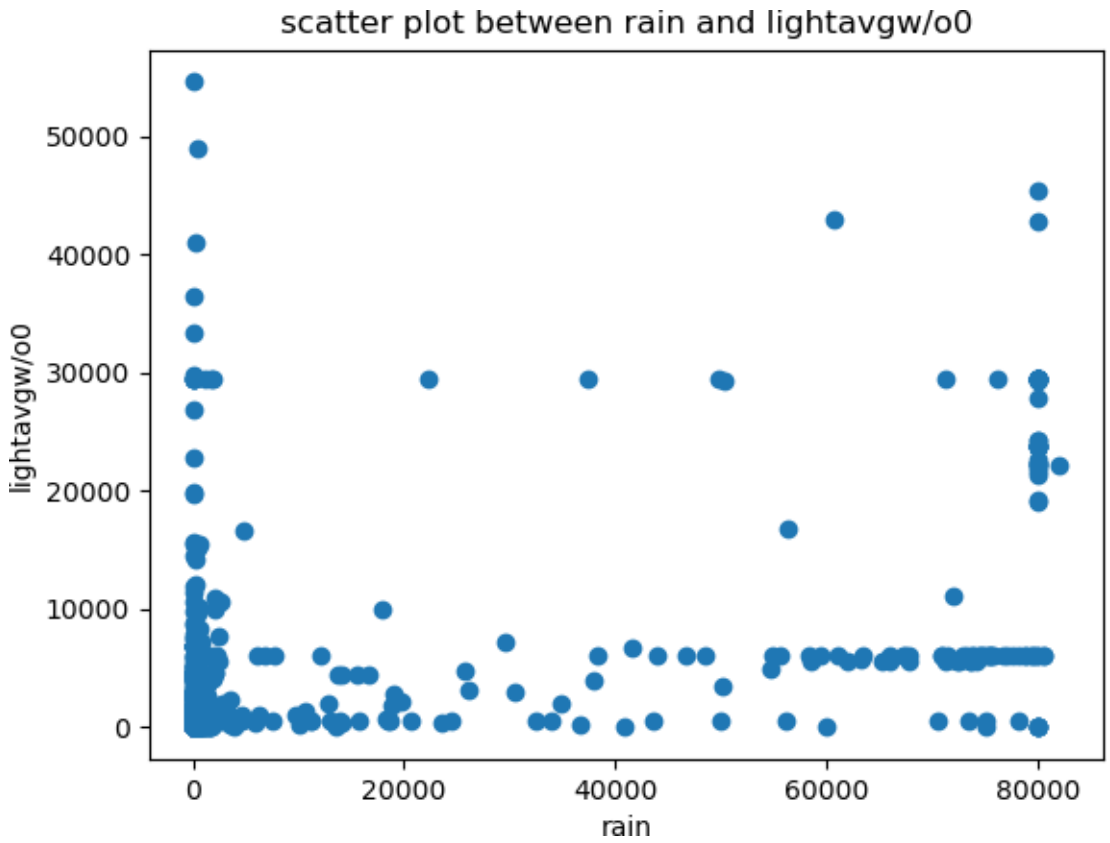
The plot can be clearly read as the relative atmospheric humidity going lower with the increase in rainfall. This makes sense as the rain will remove water vapour from air through condensation and deposit on the surface.

III. Pressure (atmospheric in millibars)



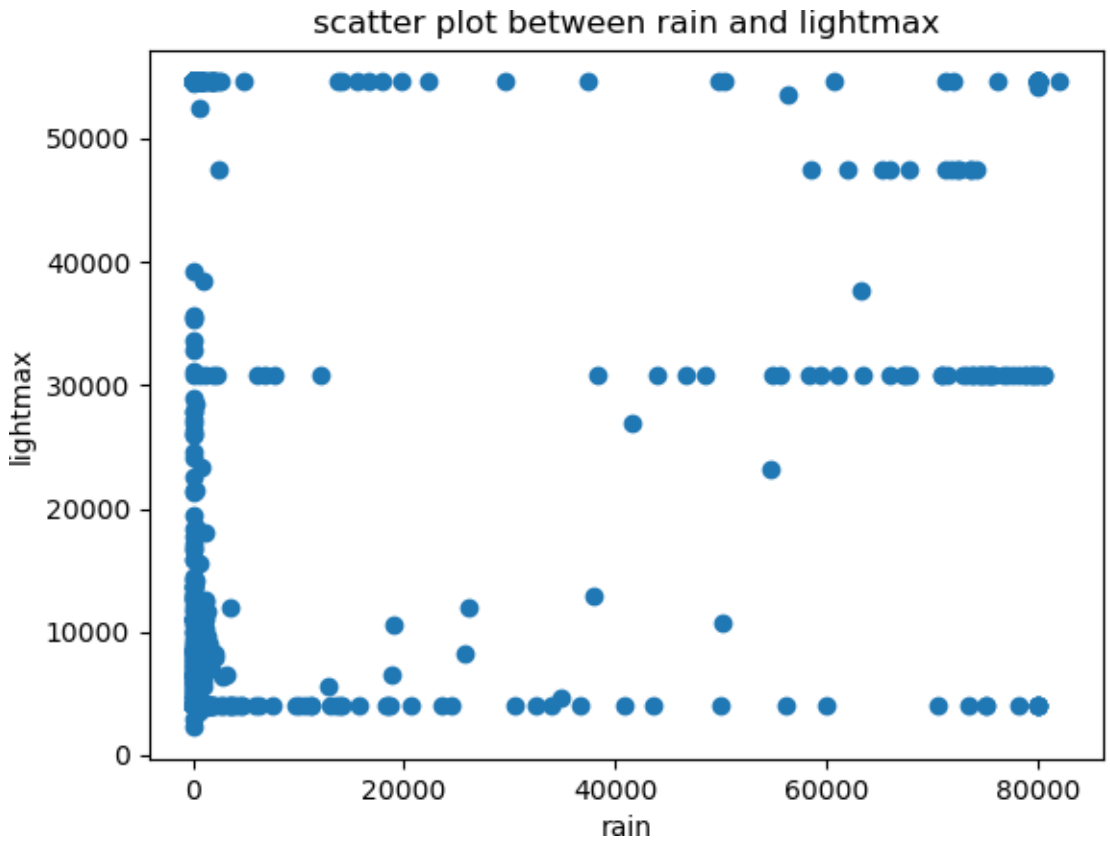
The amount of rainfall has no major effects on the atmospheric pressure as the pressure solely depends on the elevation from sea level. Rain does seem to reduce the vague and incorrect values though, which is an interesting observation and has something to do with how the barometers work and this should be investigated further.

IV. Average Light (average in a day in lux)



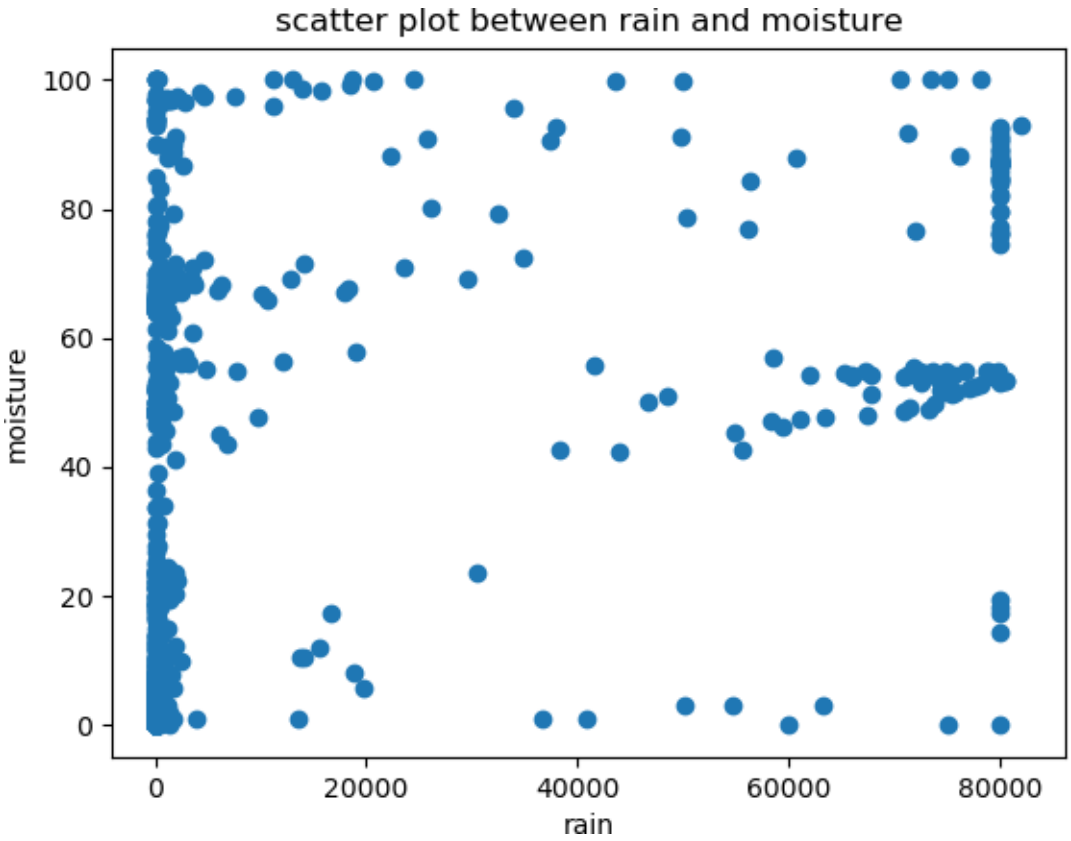
The average light forms very straight lines and shows a very prominent value of around 8000 lux at the higher rainfall regions.

V. Maximum Light (max in a day in lux)



Again, the maximum light in a day forms very straight characteristic lines in the plot. This makes us wonder on how exactly do these sensors work and how they might be oriented that they give such exact straight lines except a few values here and there which shouldn't have been the case in the randomness of nature.

VI. Moisture (relative)

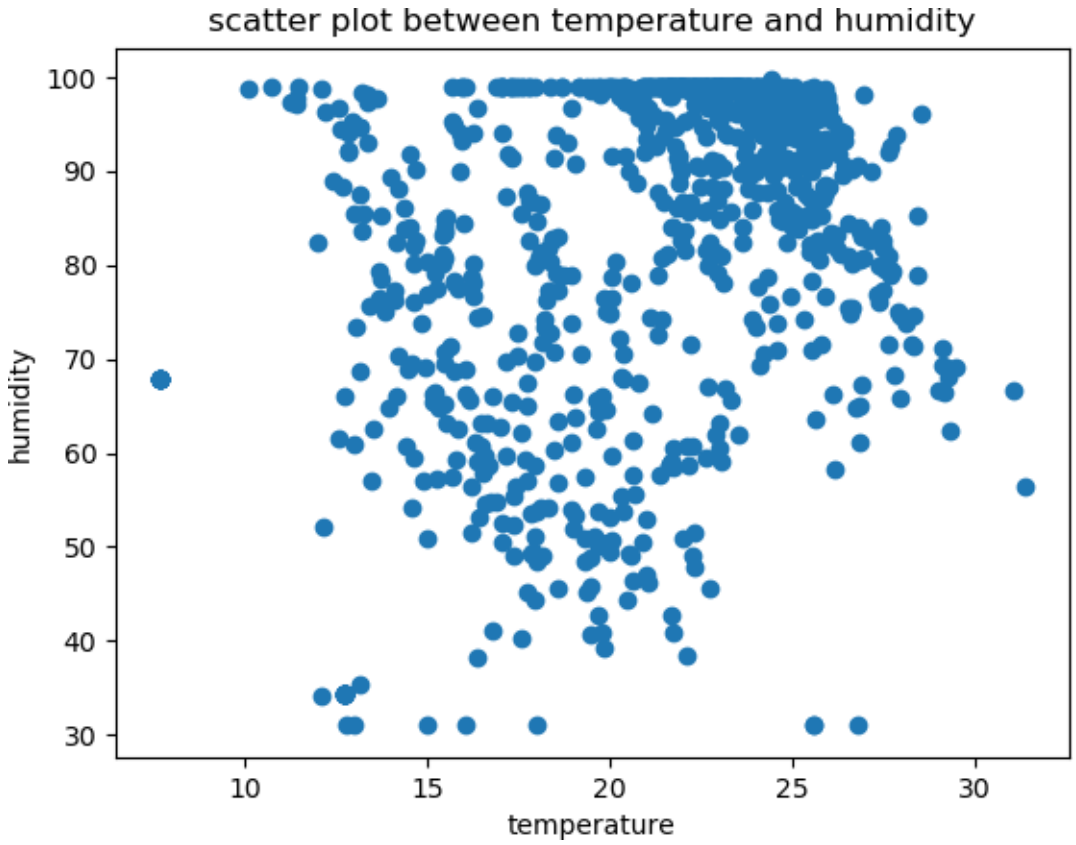


The soil moisture content can be clearly seen to increase with the increase in rain which is exactly what anyone would have guessed. What is interesting is on days when it did not rain, the soil moisture forms almost a well distributed pattern before forming a cluster at the lower values. This probably indicates that the after a day of rain followed by multiple days of little to no rain, the soil slowly drains to lower values of moisture content.

QUESTION 2B

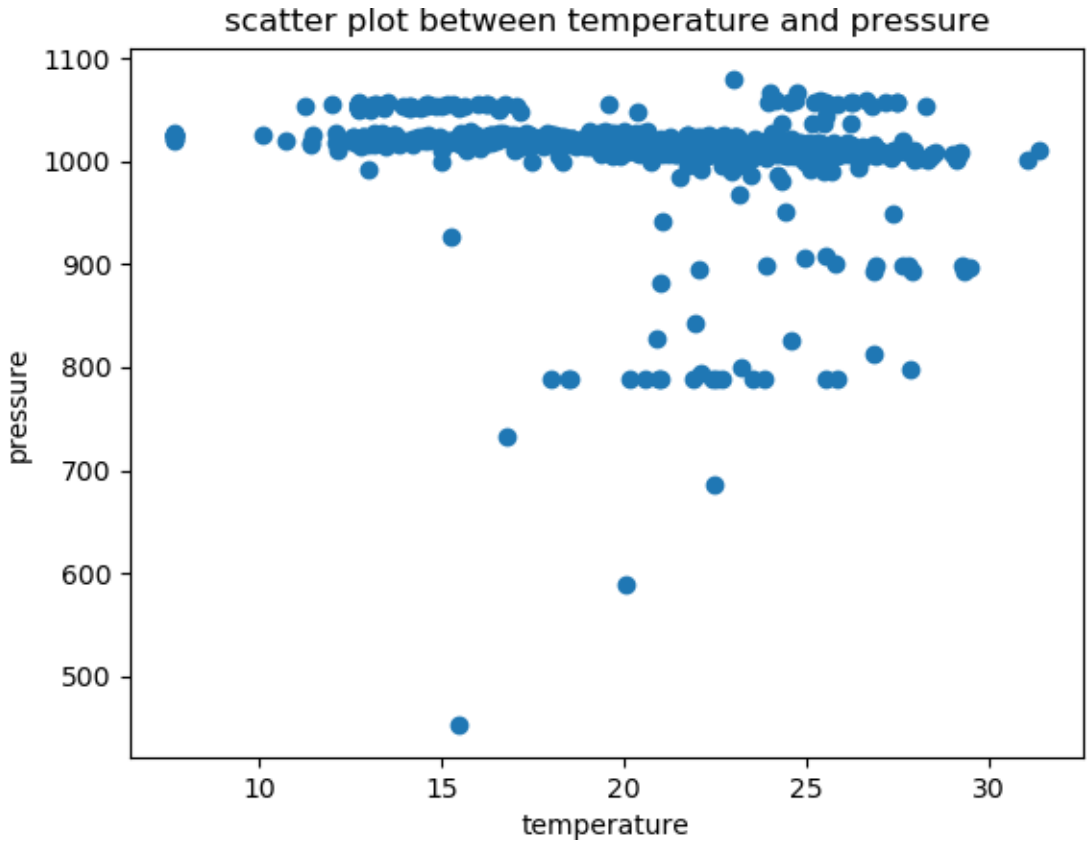
Obtain the scatter plot between 'temperature' and each of the other attributes, excluding 'dates' and 'stationid'. Consider 'temperature' in x-axis and other attributes in y-axis.

I. Humidity (relative)



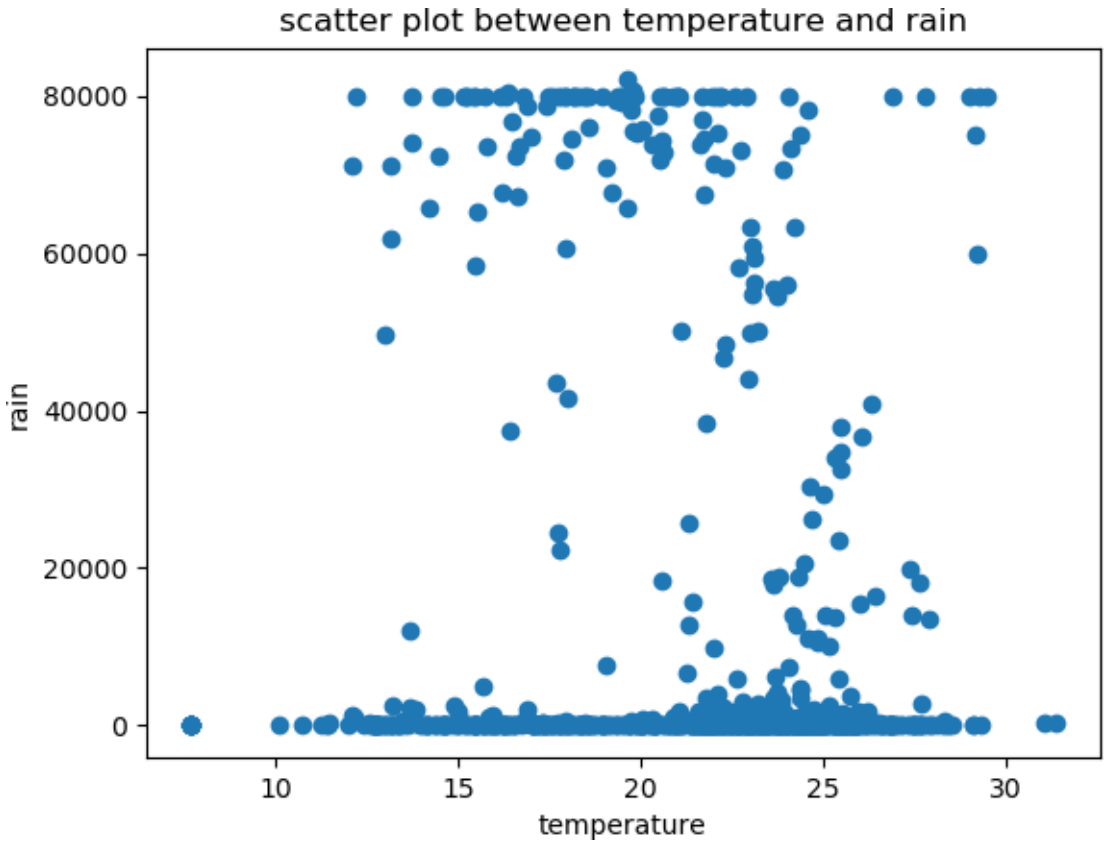
From the plot it can be clearly observed that the atmospheric humidity is spread uniformly at lower temperatures while it forms a distinct cluster with the humidity being high at higher temperatures. Overall, we could say that an uphill pattern is formed which shows the attributes are positively correlated. This extends from the fact that on hotter days, there is more evaporation and hence higher humidity levels in the atmosphere.

II. Pressure (atmospheric in millibars)



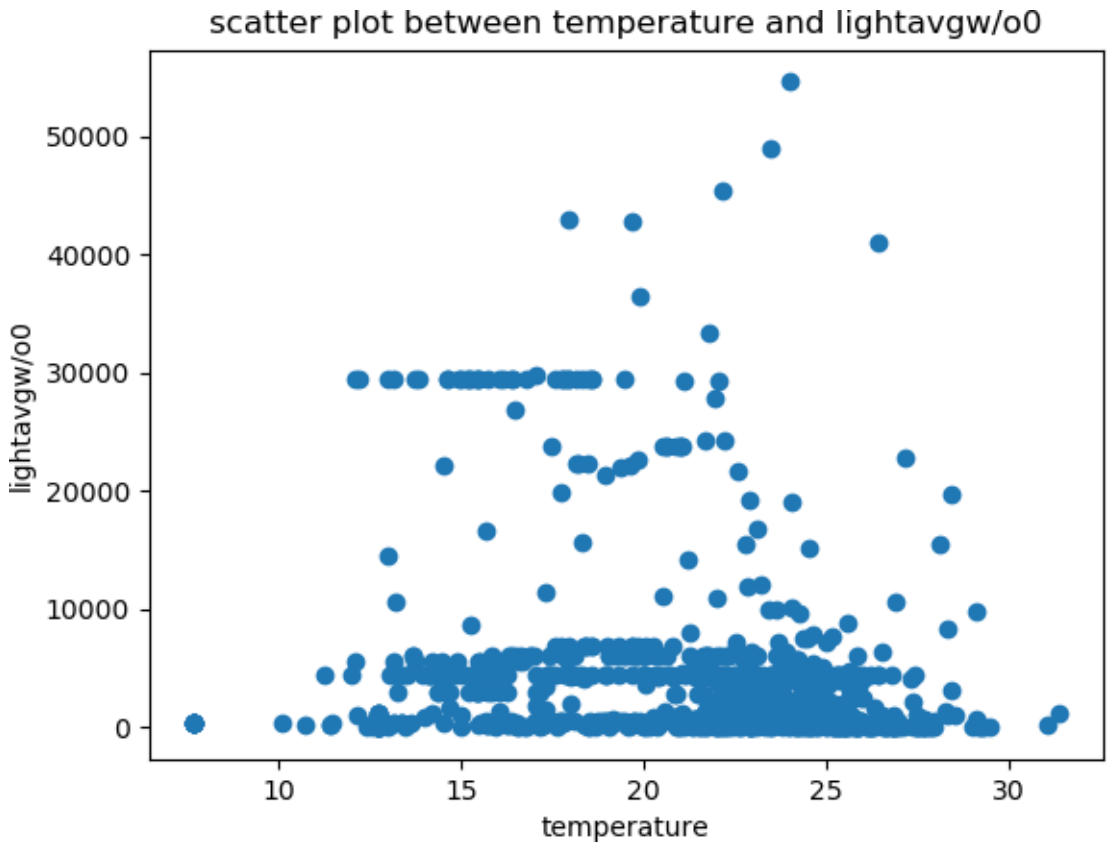
The difference in temperature does not really cause any major change in the atmospheric pressure band which is also true as the pressure mainly depends on the elevation from sea level. The randomness in the pressure is reduced at lower temperatures but this could be due to how the actual barometers function.

III. Rain (in millilitres)



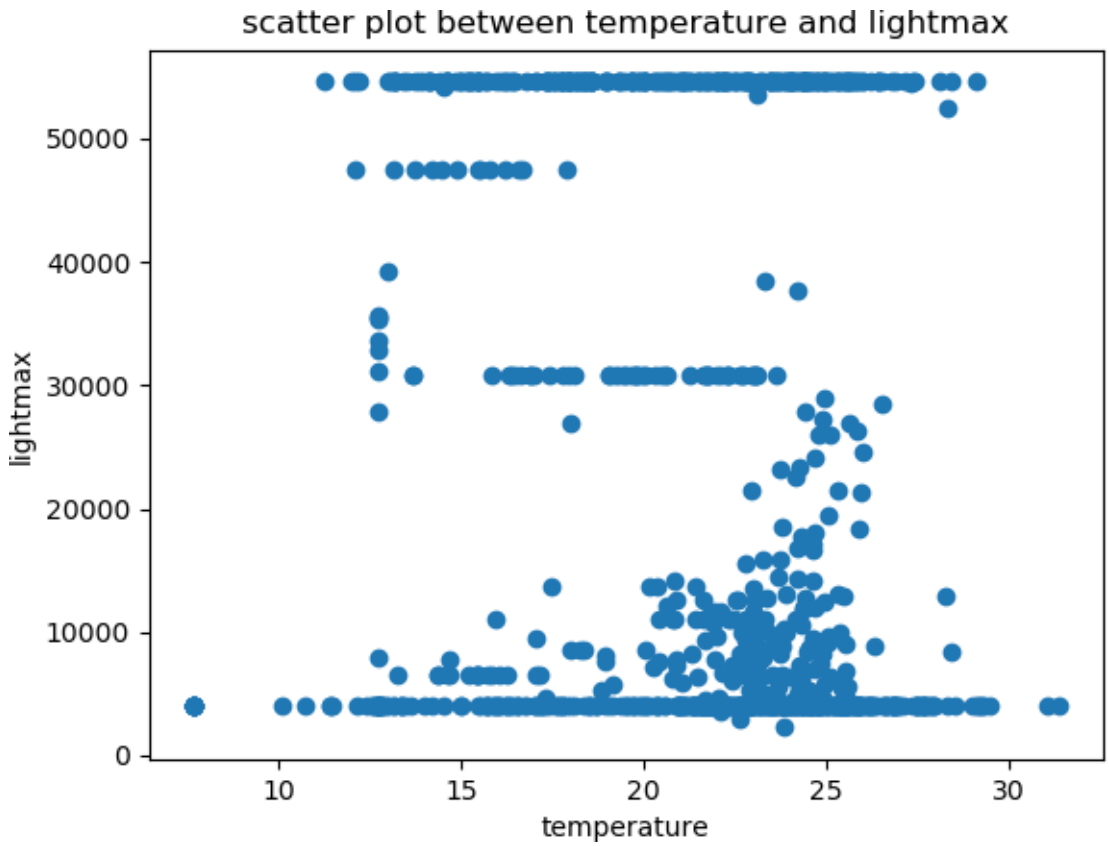
The mass of the graph can be seen to have a negative correlation as the datapoints of higher rain recorded lower temperatures and vice versa. This is also observed as rain does indeed bring down the average temperature of the region by a few degrees.

IV. Average Light (average in a day in lux)



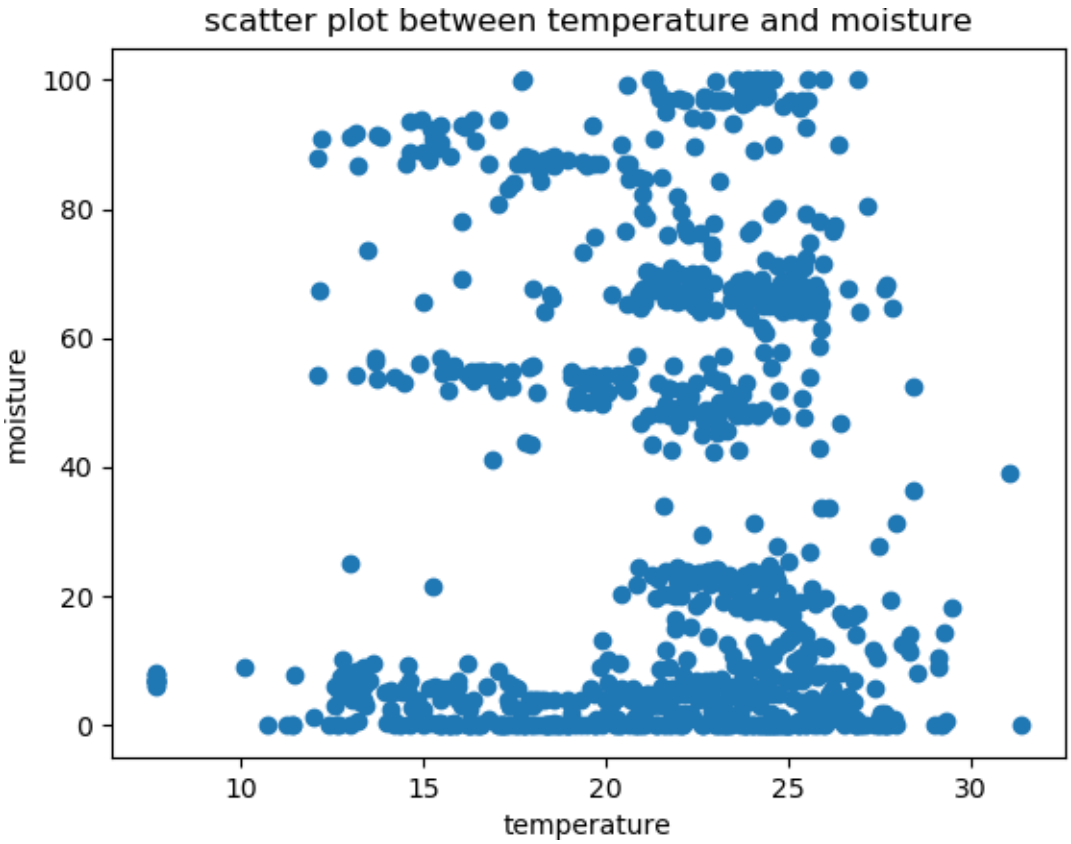
The scatter plot shows clear straight lines. It seems to show more lower values at higher temperatures but the bulk of the points stay in the same region more or less

V. Maximum Light (max in a day in lux)



This scatter plot too shows very distinct straight lines with no direct or indirect correlations to be observed. It does however show more randomness at higher temperatures but this is more likely because of how the actual instrument functions.

VI. Moisture (relative)



The soil moisture content is more or less regularly distributed throughout the graph and no direct linear relationship can be formed between moisture content and temperature.

QUESTION 3A

Find the value of correlation coefficient of 'rain' with all other attributes (excluding dates and stationid).

```
Correlation Coefficents with Rain:
temperature    -0.108893
humidity       -0.434917
pressure        0.070785
rain            1.000000
lightavgw/o0    0.527490
lightmax        0.312843
moisture        0.426928
```

Output

- Rain and Temperature are negatively weakly correlated as we know rainfall usually brings down temperatures.
- Humidity on the other hand is negatively correlated with rain moderately. Rainfall does lower atmospheric humidity by causing more condensation.
- Rain and pressure have no real correlation
- And, although both measures of light over a day give moderate positive correlation with rain, we can't really justify it.
- Rain and soil moisture content do have a moderate positive correlation and this is obvious as the rainwater is the one that irrigates the soil in such regions.

QUESTION 3B

Find the value of correlation coefficient of 'temperature' with all other attributes (excluding dates and stationid).

```
Correlation Coefficents with Temperature:
temperature     1.000000
humidity        0.401570
pressure       -0.181389
rain           -0.108893
lightavgw/o0   -0.181400
lightmax       -0.145884
moisture        0.080660
```

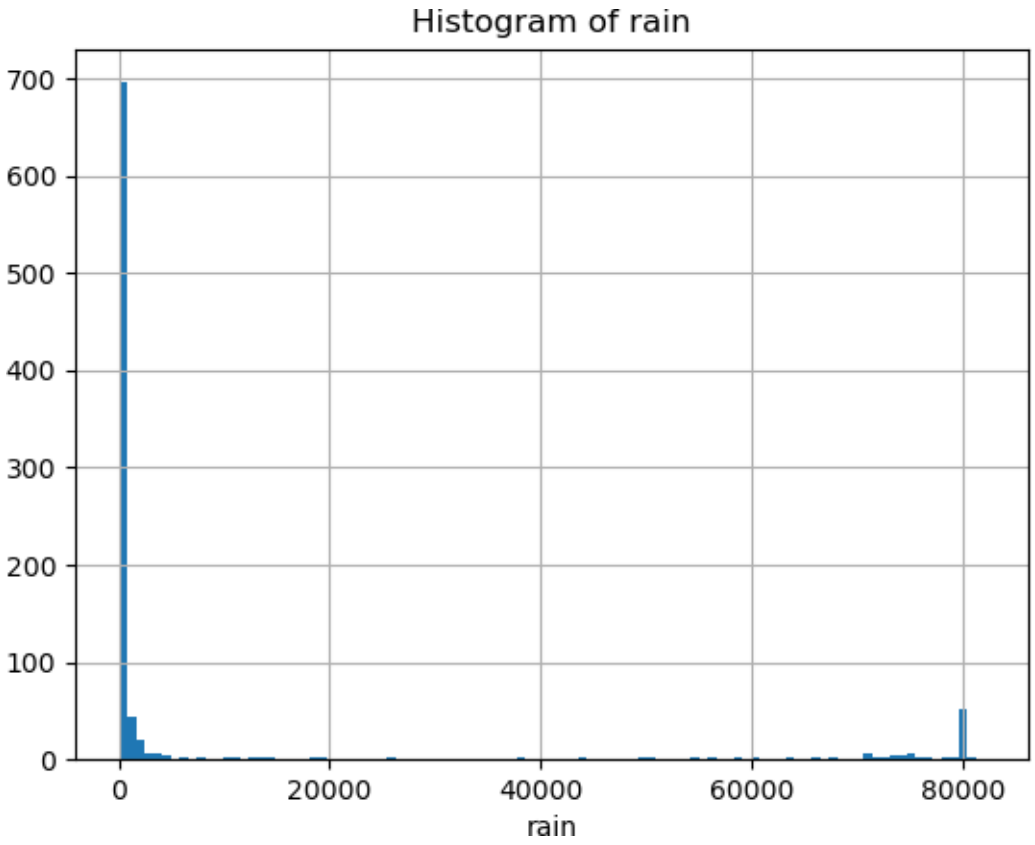
Output

- Temperature and Humidity have a moderate positive correlation. This can be explained as with the increase in temperature, the levels of evaporation increases and with greater evaporation, the moisture content of the atmosphere increases.

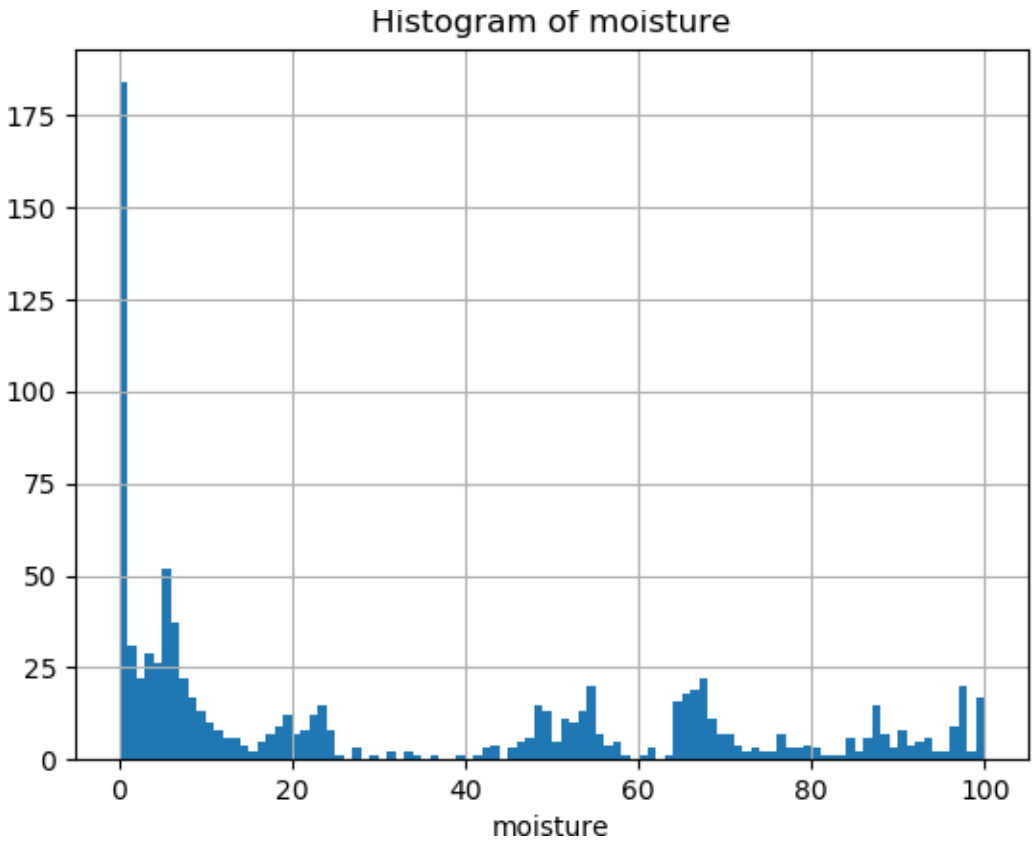
- Pressure, Rain and the Luminous Intensity all show weak negative correlation with Temperature. This might be due to several indirect factors affecting the environment of the region where the monitoring stations are set up.
- Moisture and Temperature show negligible positive correlation and this can be assumed to be because of the randomness in the data.

QUESTION 4

Plot the histogram for the attributes 'rain' and 'moisture'



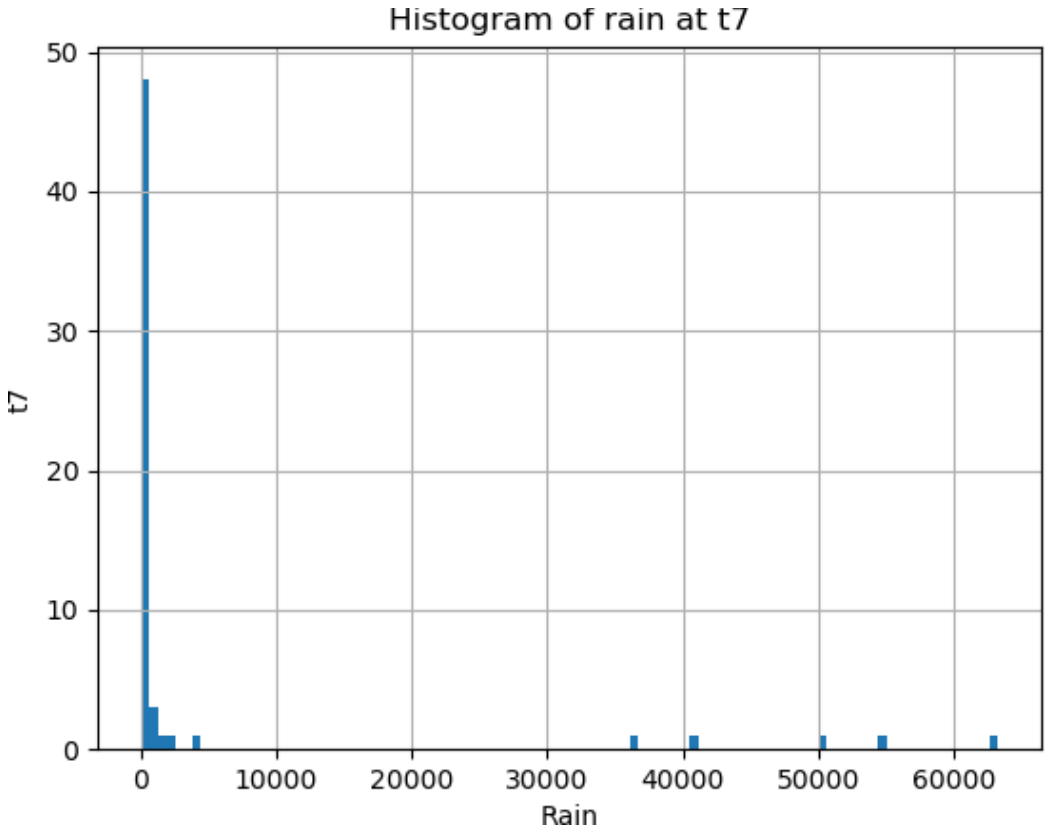
We can observe that the data is mostly dominated by days of no rain. But there are also a few days at the upper end of the spectrum with very heavy rainfall with a few days here and there with moderate rainfall.

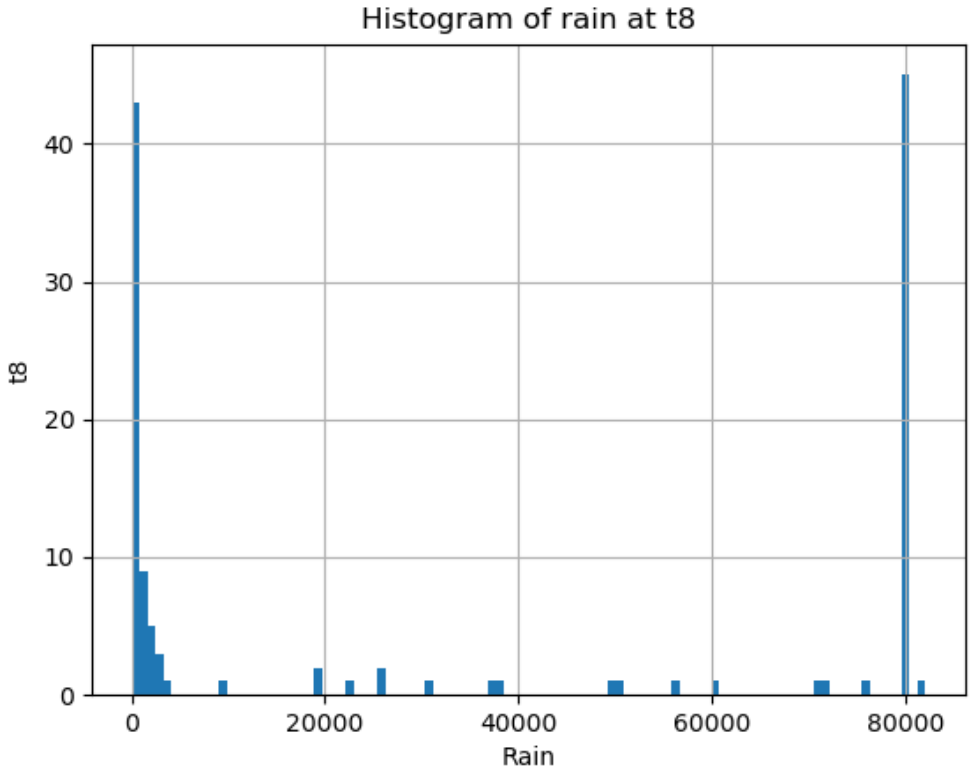


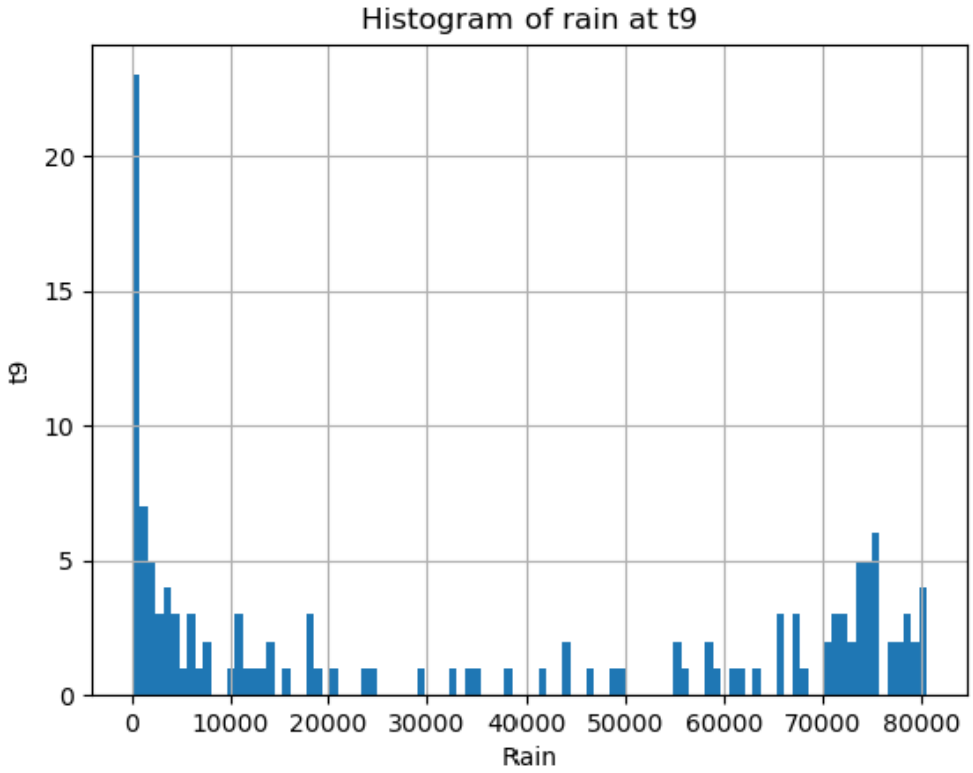
We can observe that the data is positively skewed with the mode being at 0. The soil moisture content is more fairly distributed over the entire region when compared to rainfall though. This is possibly due to the fact that soil loses its moisture to seeping over time.

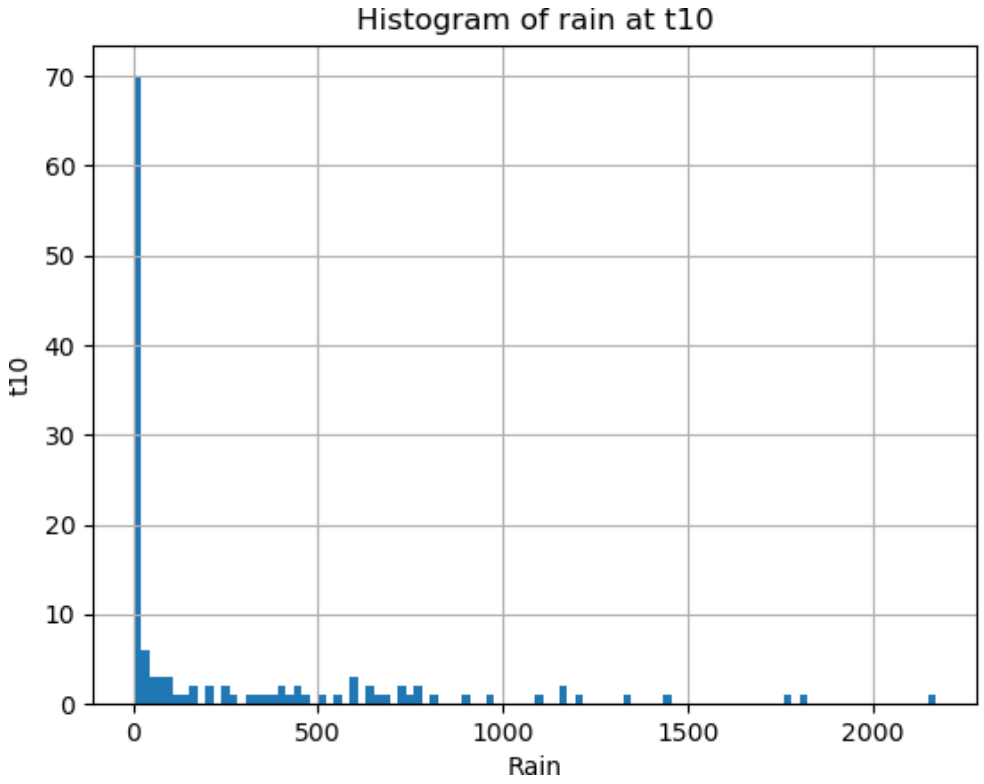
QUESTION 5

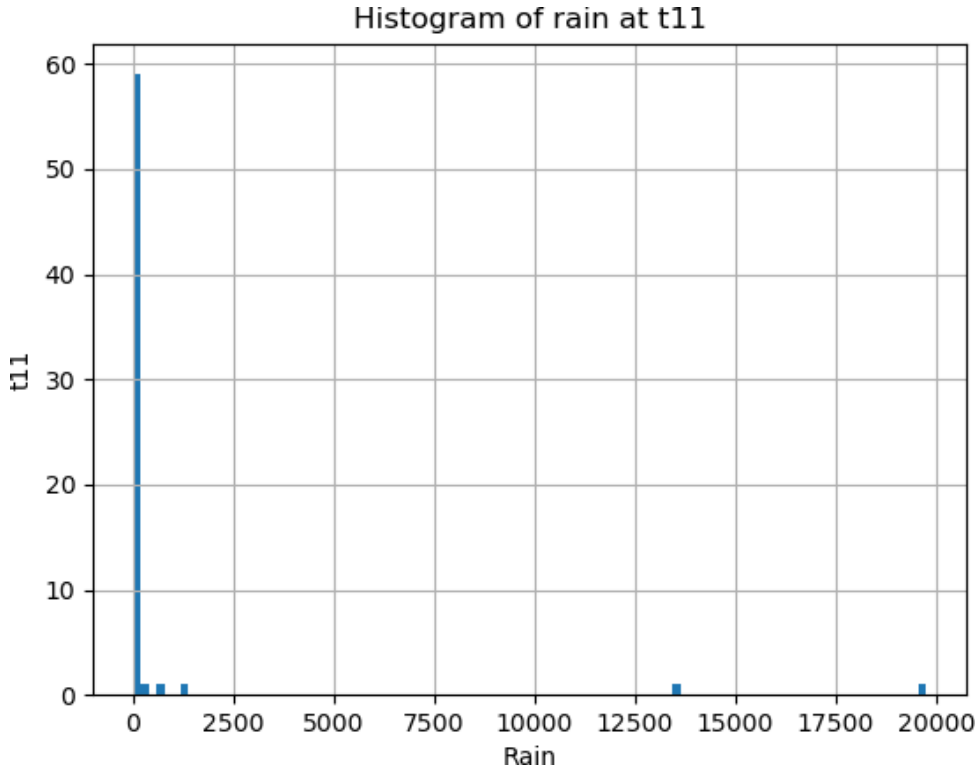
Plot the histogram of attribute 'rain' for each of the 10 stations

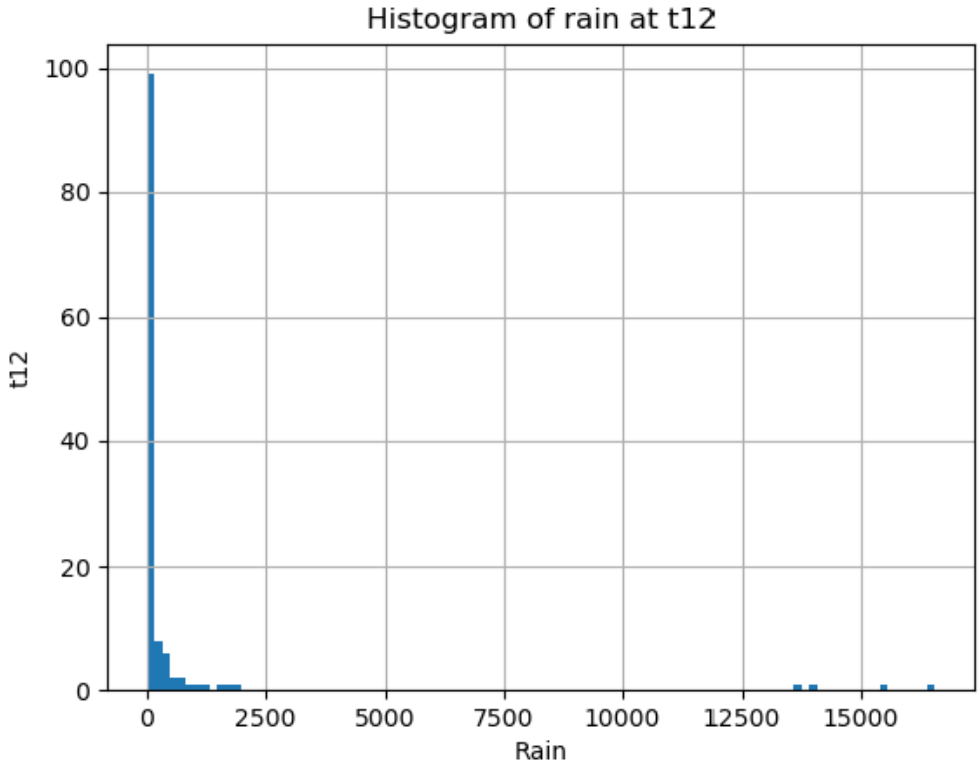


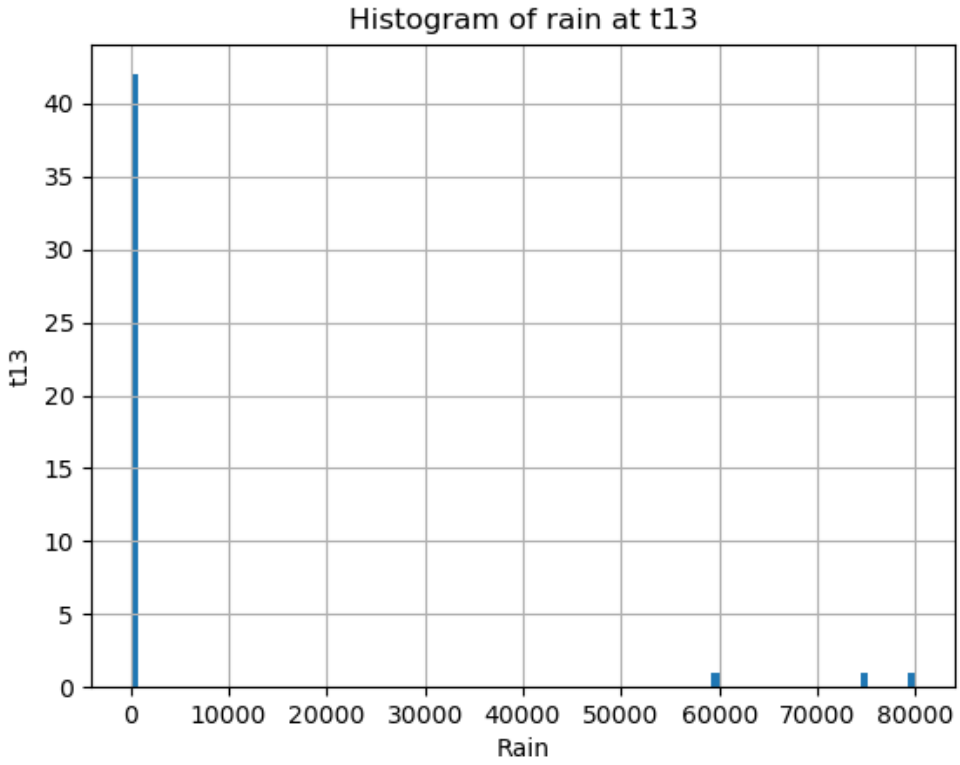

Histogram of rain at t8
t8
Rain
0
10
20
30
40
0
20000
40000
60000
80000

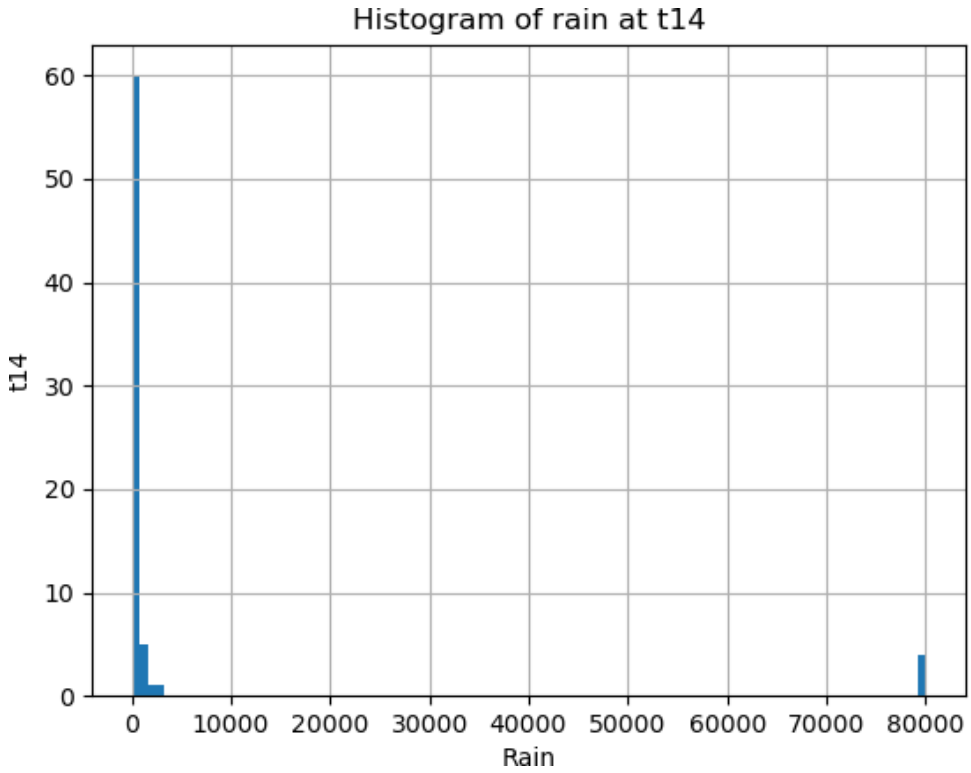

Histogram of rain at t9
t9
Rain
0
5
10
15
20
0
10000
20000
30000
40000
50000
60000
70000
80000

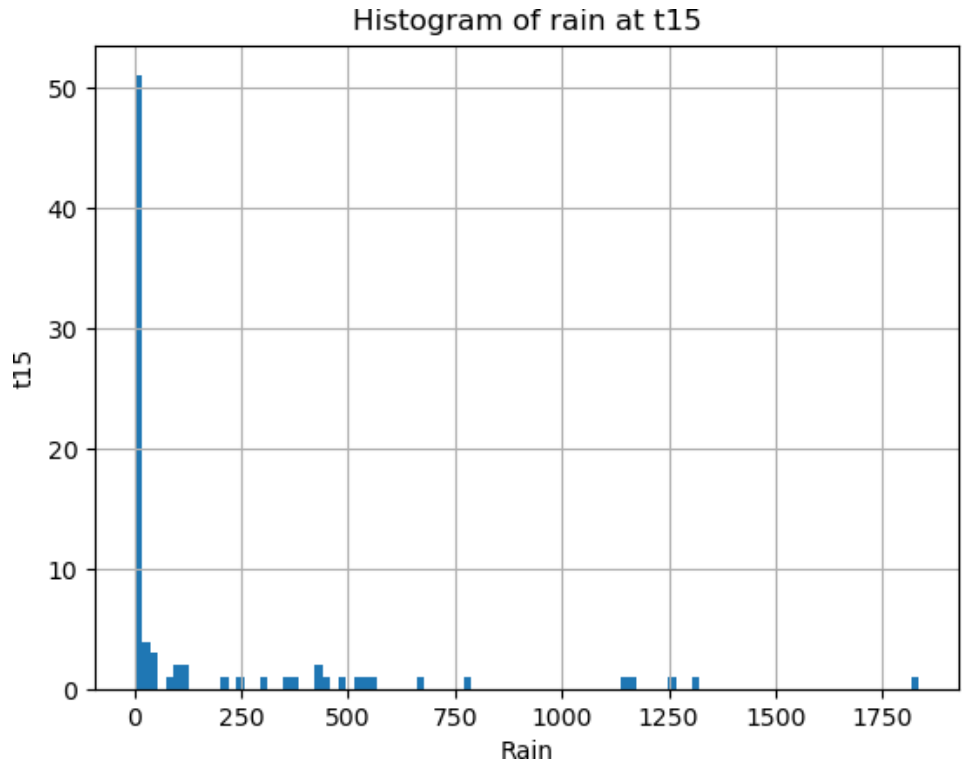

Histogram of rain at t10
t10
Rain
0
10
20
30
40
50
60
70
0
500
1000
1500
2000


Histogram of rain at t11
t11
Rain
0
10
20
30
40
50
60
0
2500
5000
7500
10000
12500
15000
17500
20000


Histogram of rain at t12
t12
Rain
0
20
40
60
80
100
0
2500
5000
7500
10000
12500
15000

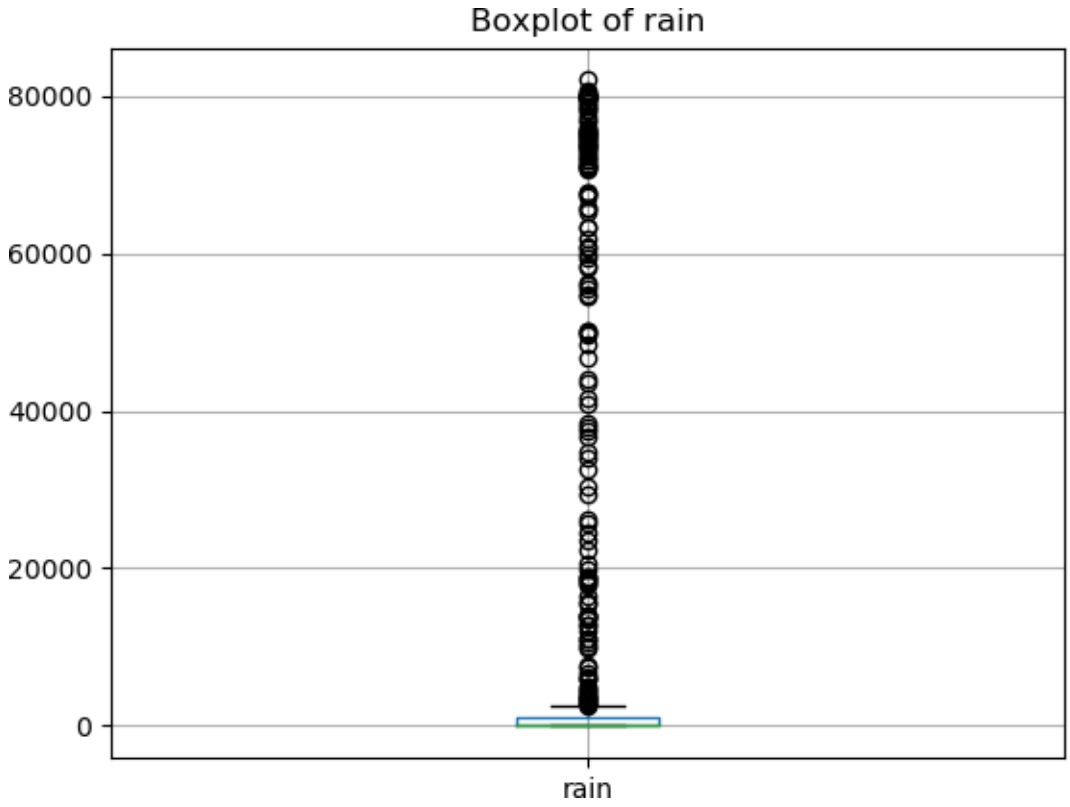

Histogram of rain at t13
t13
Rain
0
5
10
15
20
25
30
35
40
0
10000
20000
30000
40000
50000
60000
70000
80000


Histogram of rain at t14
t14
Rain
0
10
20
30
40
50
60
0
10000
20000
30000
40000
50000
60000
70000
80000

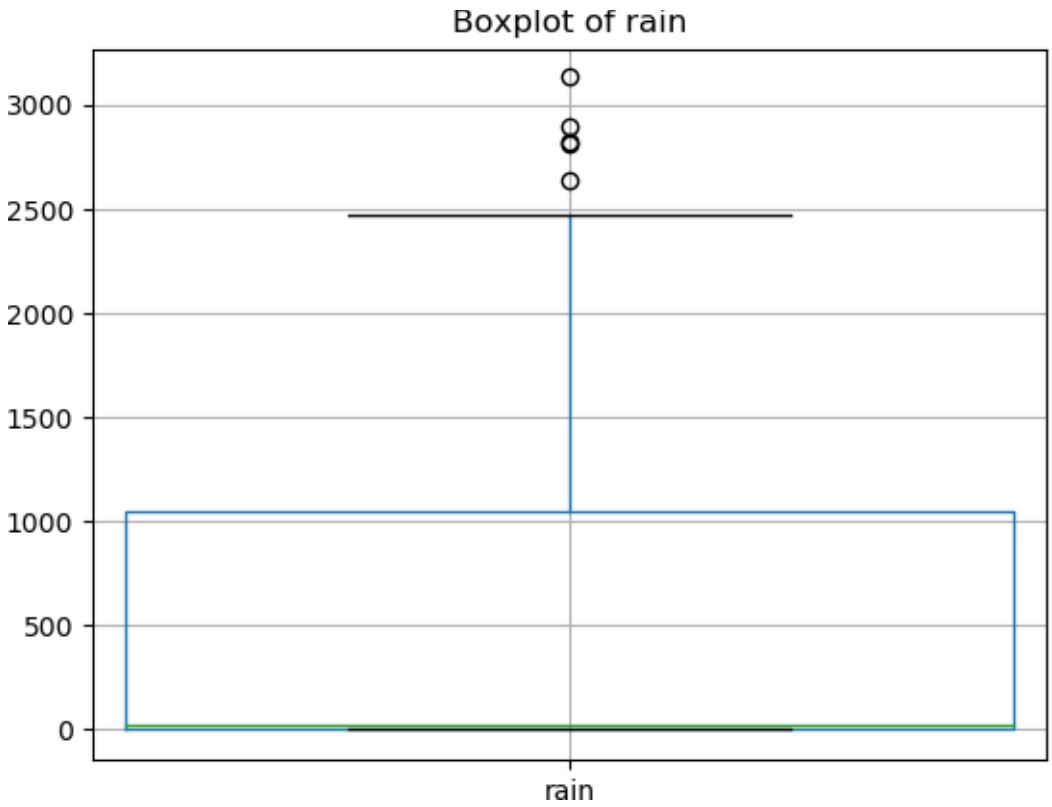

Histogram of rain at t15
t15
Rain
0
10
20
30
40
50
0
250
500
750
1000
1250
1500
1750

QUESTION 6

Obtain the boxplot for the attributes 'rain' and 'moisture'

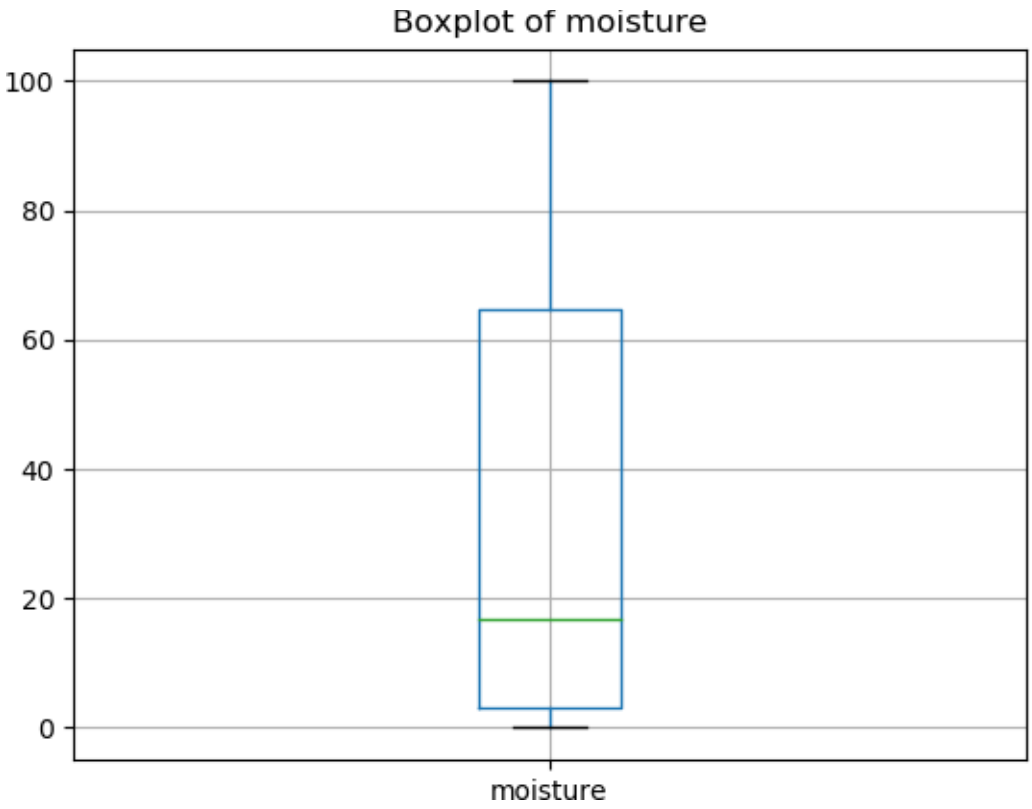


This is the boxplot for rain. It is heavily positively skewed as can be confirmed from the histogram presented in question 4. Most of the graph can be observed as being full of outliers. This is due to the fact that the data collection period had most days of little to no rain while there were a few with very heavy rainfall. To have a better look at the main box, we'll zoom into the figure.



Boxplot with outliers removed

Here we can notice that the median, and first quartile almost coincide at the minimum value, zero, because of the large number of observations with value 0.



This is the boxplot of soil moisture content with its comparatively smaller positive skew. The minimum and maximum values are 0% and 100% respectively.