IC 272: DATA SCIENCE - III
LAB ASSIGNMENT - II
Data cleaning – handling missing values and outlier analyses

Student's Name: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:EE**

1

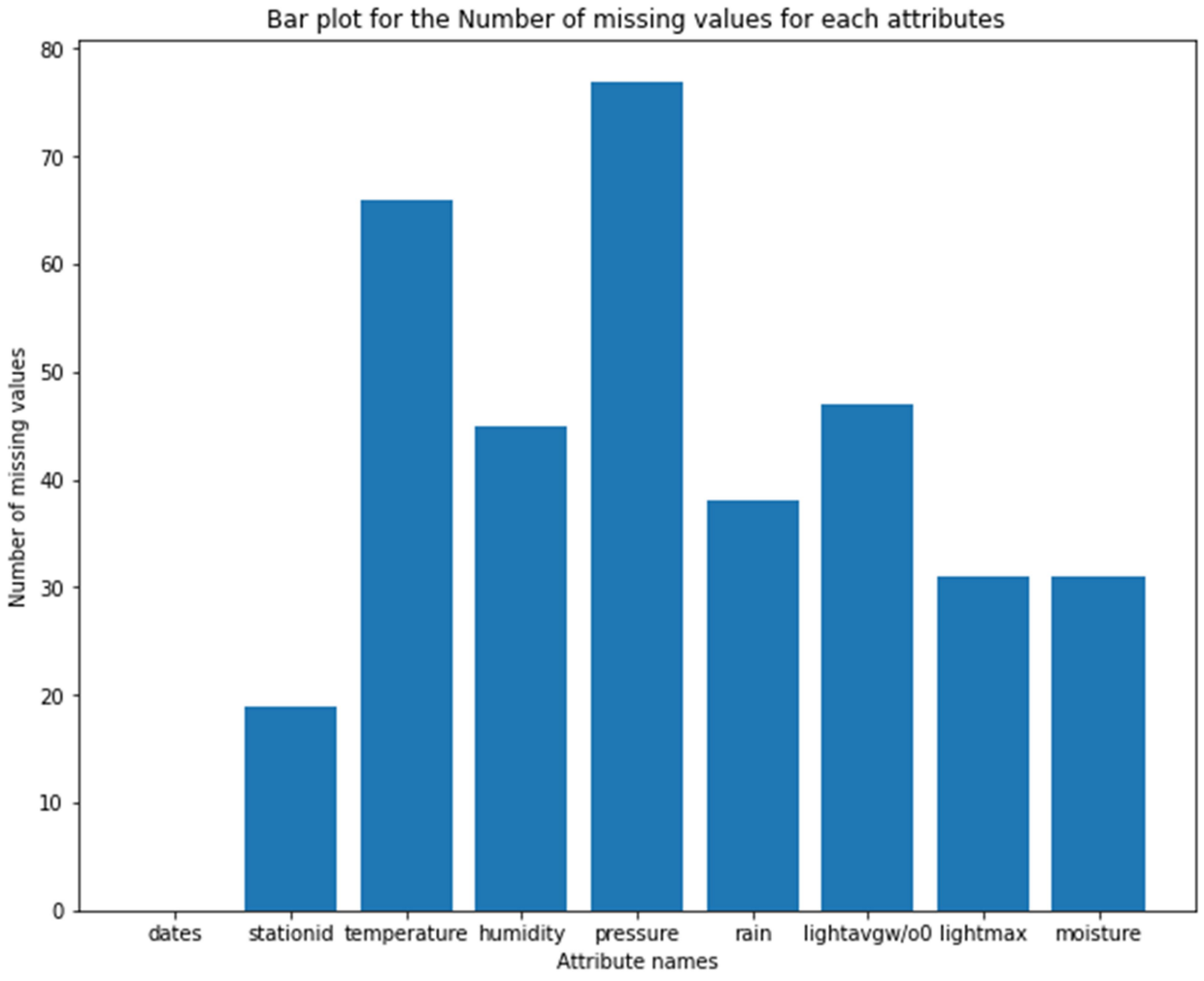

Figure 1 Number of missing values vs. attributes

Inferences:

1. Attribute pressure has maximum number of missing values , and attribute stationid has minimum number of missing values.
2. For dates there is no missing value , for stationid the number of missing values are very so this sensor works fine , for temperature its missing value equals 66 , for humidity the number of missing values is close to 45 , and for pressure It is around 79 which is very high (means sensor is not working properly) for the rain and light avgw/o0 sensor works moderately , and for light max and moisture it works fine.
3. From the barchart it can be observed that pressure sensor don't works well , and light max sensor works very well.

2 a.

Inferences:

1. Since the target attribute is stationid and if there are missing values in it then we are not able to detect which sensor is faulty as there are 10 location where these sensors are situated in mandi town .
2. number of tuples deleted are 19.
3. 2.01% of tuples are deleted from total tuples available.

b.

Inferences:

1. 35 tuples are deleted after this step.
2. 3.703% of tuples are deleted from total tuples available.
3. Since there is very few percent of tuples are deleted so the data loss is very low.
4. Since there are 9 total attributes available and we drop that tuple which contains missing value more or equal to one third of number of attributes available to ensure that there should be no data loss happens.

3

Table 1 Number of missing values per attribute after removing missing values

S. No	Attribute	Number of missing values
1	dates	0
2	stationid	0

3	temperature (in °C)	34
4	humidity (in $g.m^{-3}$)	13
5	pressure (in mb)	41
6	rain (in ml)	6
7	lightavgw/o0 (in lux)	15
8	lightmax (in lux)	1
9	moisture (in %)	6

Inferences:

1. The attribute pressure has maximum number of missing values available , and the attribute dates and stationid has minimum number of missing values available.
2. For dates there is 0% of data missing ,for stationid there is 0% of data missing , for temperature there is 29.3% of missing values , for humidity theres is 11.2% of missing values , for pressure there is 35.34% of missing values , for rain there is 5.71% of missing values ,for lightavgw/o0 there is 12.93% of missing values , for light max there is 0.82% of missing values , for moisture there is 5.17% of missing values available.
3. Total number of missing values in the file is 116.

4 a. i.

Table 2 Mean, mode, median and standard deviation before and after replacing missing values by mean

S. No	**Attribute**	**Before**				**After**			
		Mean	**Mode**	**Median**	**S.D.**	**Mean**	**Mode**	**Median**	**S.D.**
1	dates	Na	na	Na	NA	NA	NA	NA	NA
2	stationid	Na	na	Na	NA	NA	NA	NA	NA
3	temperature (in °C)	21.073	21.000	21.800	4.243	21.214	12.727	22.272	4.355
4	humidity (in $g.m^{-3}$)	83.249	99.000	90.119	17.967	83.479	99.000	91.380	18.210
5	pressure (in mb)	1009.206	1009.000	1014.070	45.214	1009.008	789.392	1014.677	46.980

6	rain (in ml)	10942.697	0.000	24.750	24574.252	10701.538	0.000	18.000	24852.255
7	lightavgw/o0 (in lux)	4430.881	4488.910	1911.233	7400.586	4438.428	4488.910	1656.880	7573.162
8	lightmax (in lux)	21650.157	4000.000	7544.000	21678.196	21788.623	4000.000	6634.000	22064.993
9	moisture (in %)	32.649925	0.000	17.723	33.416	32.386	0.000	16.704	33.653

Inferences:

1. for the mean rain has maximum change , temperature has minimum change, for the median lightavgw/o0 has maximum change and pressure has minimum change , for the mode pressure has maximum change and moisture has minimum change, for the std the light max have maximum change and moisture has minimum change .

2. Since the data after changes is mostly similar to the orginal data so this data is reliable for further uses.

ii.

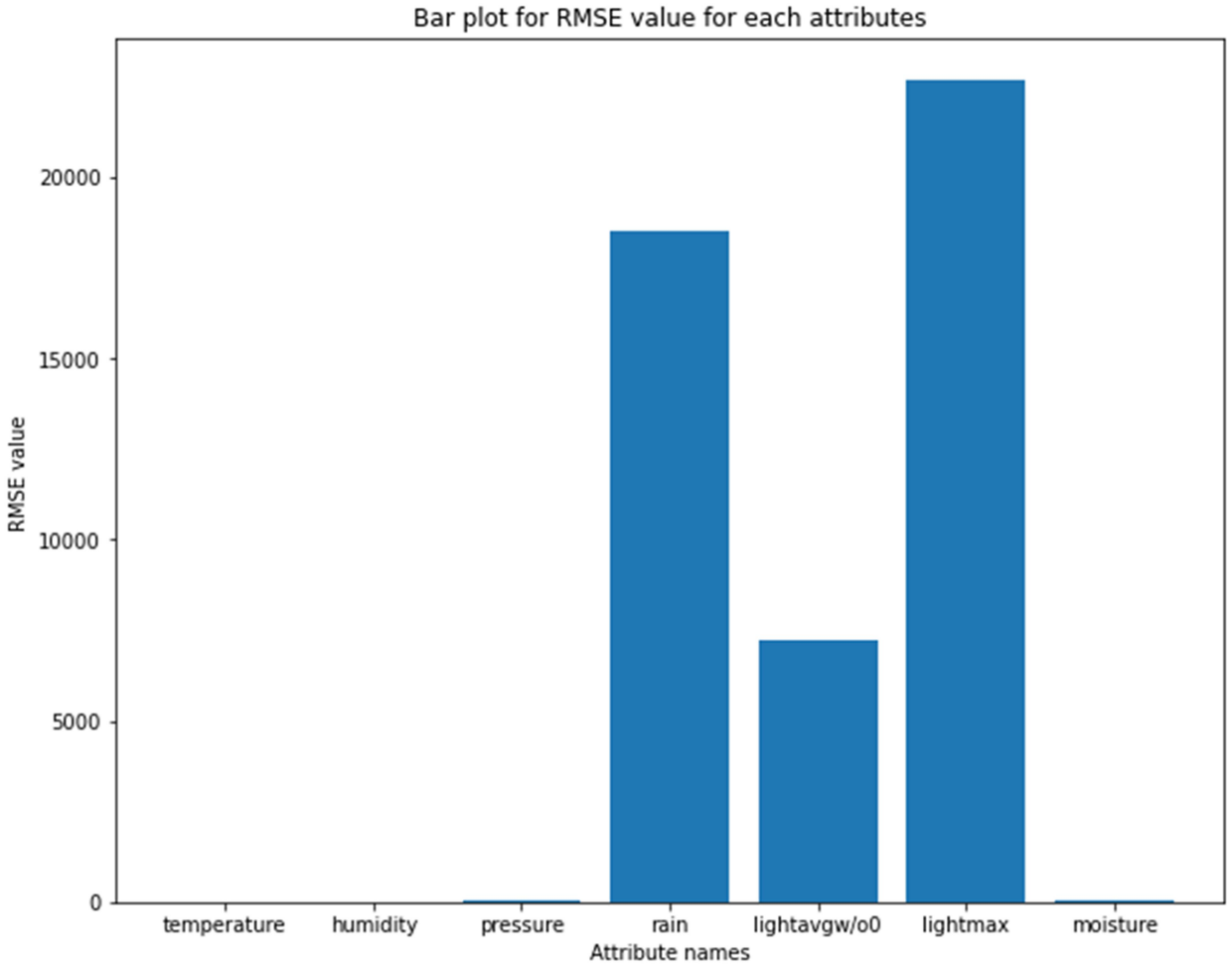

Figure 2 RMSE vs. attributes

Inferences:

1. The maximum value of RMSE is for lightmax and minimum for the temperature and humidity.
2. For an attribute if mean is maximum then it have more RMSE, and for the attribute if median is maximum then that attribute have highest RMSE value , and for an attribute if mode is maximum then RMSE value is minimum for that attribute, and if for a attribute if it shows maximum change for the std then its RMSE value should be low.
3. No, data is not suitable for further analysis as it can been seen from RMSE vs attributes the RMSE value for some attribute is very high as compared to others .

b. i.

Table 3 Mean, mode, median and standard deviation before and after replacing missing values by linear interpolation technique

S. No	Attribute	Before				After			
		Mean	Mode	Median	S.D.	Mean	Mode	Median	S.D.
1	dates	NA	NA	NA	NA	NA	NA	NA	NA
2	stationid	NA	NA	NA	NA	NA	NA	NA	NA
3	temperature (in °C)	21.196	21.000	22.169	4.329	21.2148	12.727	22.272	4.3558
4	humidity (in $g.m^{-3}$)	83.538	99.000	91.380	18.206	83.479	99.000	91.380	18.210
5	pressure (in mb)	1009.264	1009.000	1014.677	45.998	1009.0087	789.392	1014.677	46.980
6	rain (in ml)	10651.638	0.000	22.500	24779.512	10701.538	0.000	18.000	24852.255
7	lightavgw/o0 (in lux)	4486.340	4488.910	1623.494	7573.795	4438.4284	4488.910	1656.880	7573.1628
8	lightmax (in lux)	21517.191	4000.000	6569.000	21935.165	21788.623	4000.000	6634.000	22064.993
9	moisture (in %)	32.32	0.000	16.306	33.602	32.386	0.000	16.704	33.6532

Inferences:

1. for the mean rain has maximum change , temperature has minimum change, for the median lightavgw/o0 has maximum change and pressure has minimum change , for the mode pressure has maximum change and moisture has minimum change, for the std the light max have maximum change and moisture has minimum change
2. Since the data after changes is mostly similar to the orginal data so this data is reliable for further uses.

ii.

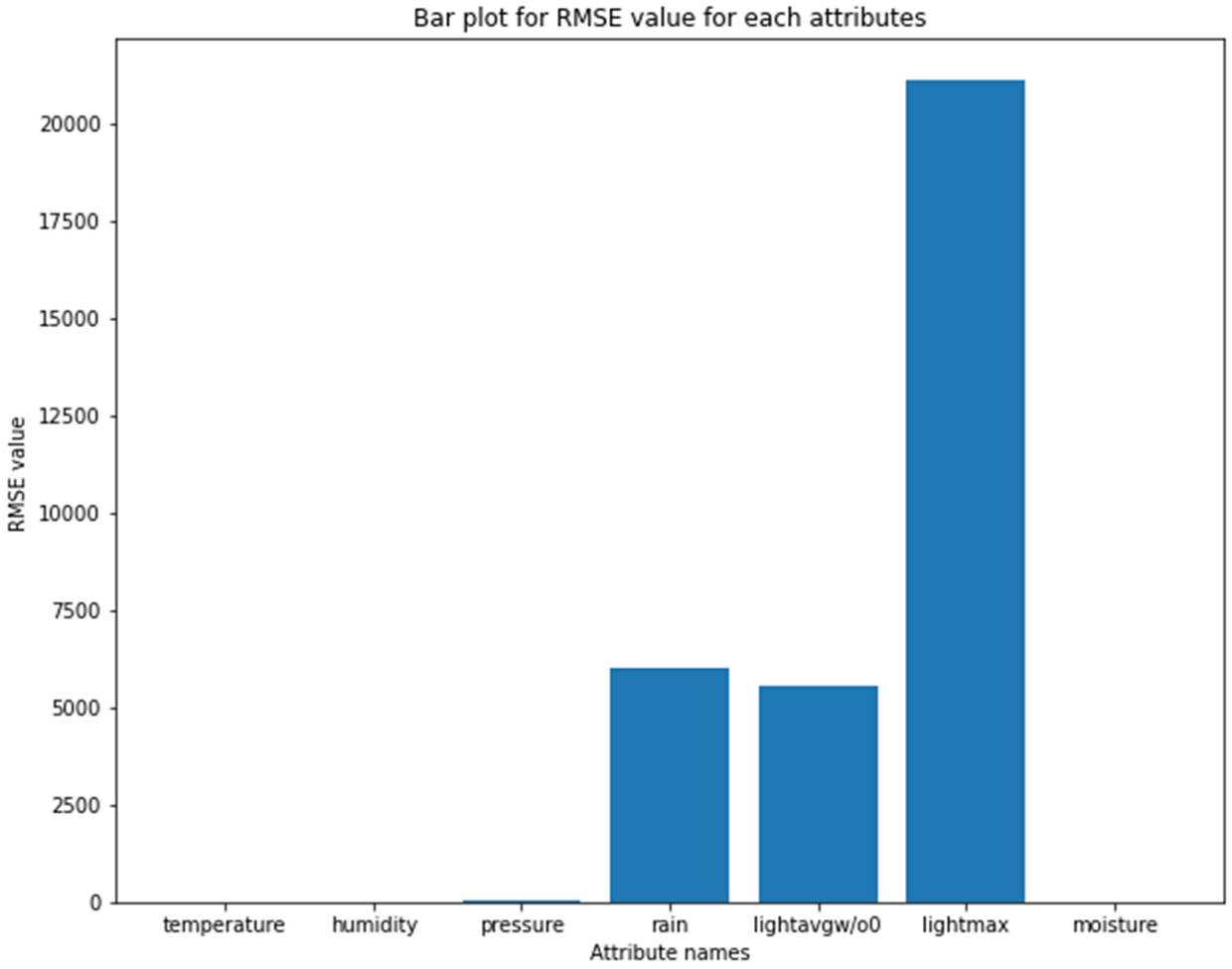

Figure 3 RMSE vs. attributes

Inferences:

1. Light max has maximum while rain has minimu?
2. For an attribute if mean is maximum then it have more RMSE, and for the attribute if median is maximum then that attribute have highest RMSE value , and for an attribute if mode is maximum then RMSE value is minimum for that attribute, and if for a attribute if it shows maximum change for the std then its RMSE value should be lowFrom RMSE ponder is the data reliable for further investigation or experimental analyses.
3. No, data is not suitable for further analysis as it can been seen from RMSE vs attributes the RMSE value for some attribute is very high as compared to others .

5 a.

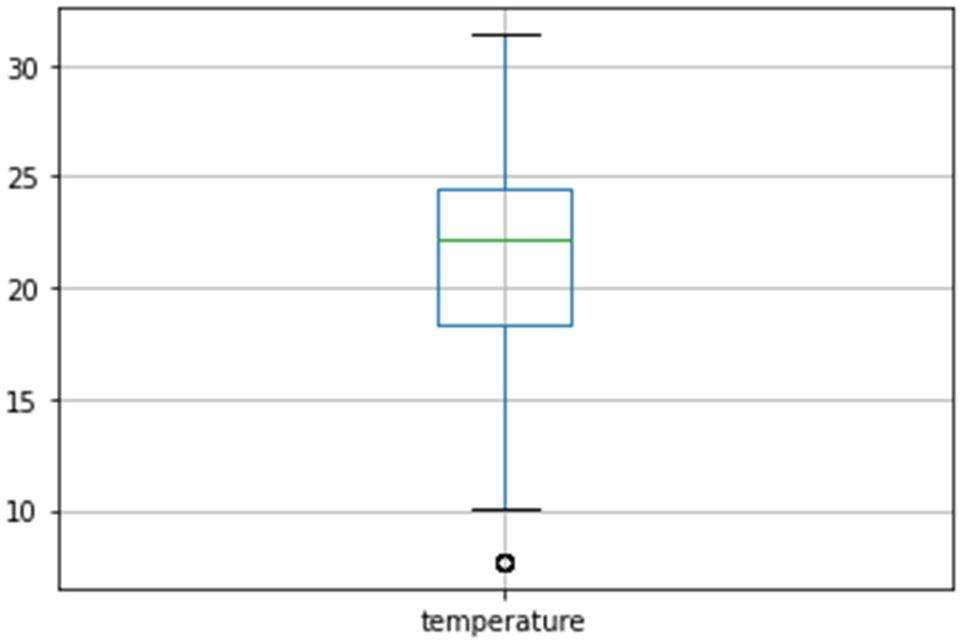

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. number of ouliers are less than 10.
2. Inter quartile range is 10 .C
3. The data is spread from 18 to 24.C
4. It is positively skewed

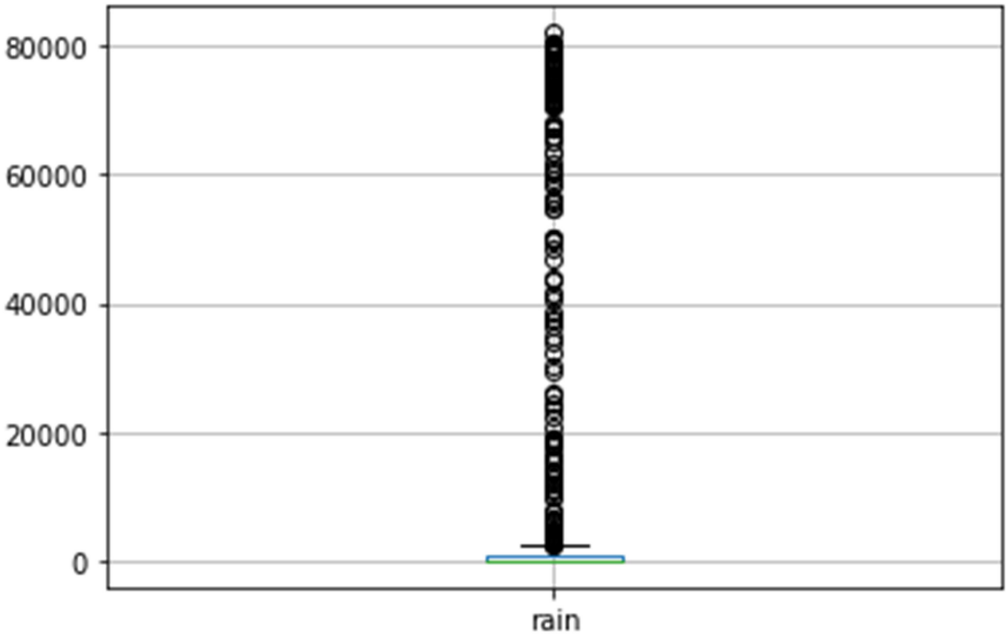

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. From 1 the outliers began and goes till 80000.ml
2. Q1 is 0.(that means the 25 percent of rainfall lies close to 10 to 100ml
3. The spread is very less.
4. It is symmetric

b.

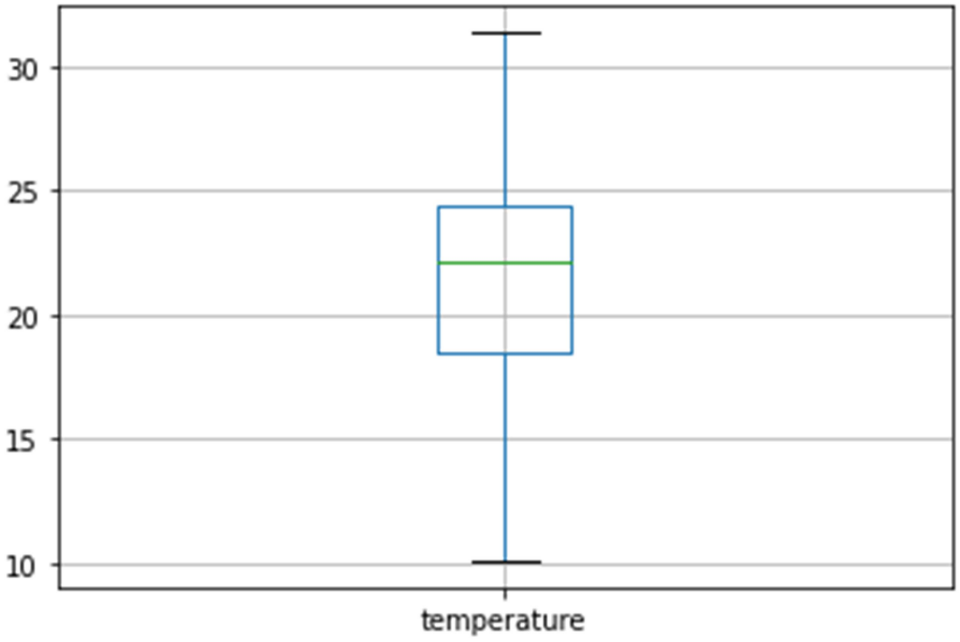

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There is no outlier present for the attribute temperature.
2. Q1 is 10.(that means 25 percent of the temperature values lies equal or less than 10 C
3. The data is spread from 19 to 24 C
4. It is symmetric

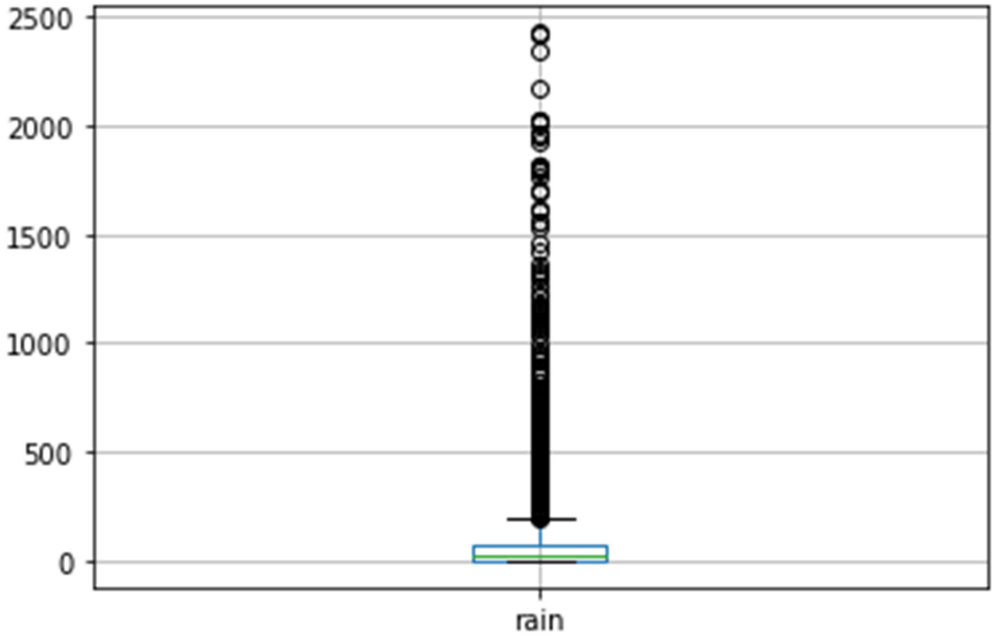

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. Outliers began from 0 and end till 2500ml
2. Q1 is 0. (that means 25 % of the rain lies close to 0ml
3. The data is spreaded from 0 to 80ml