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

Student's Name: Ishaan Gupta **Mobile No: 9179242114**

Roll Number: B20292 **Branch: Mechanical Engineering**

1

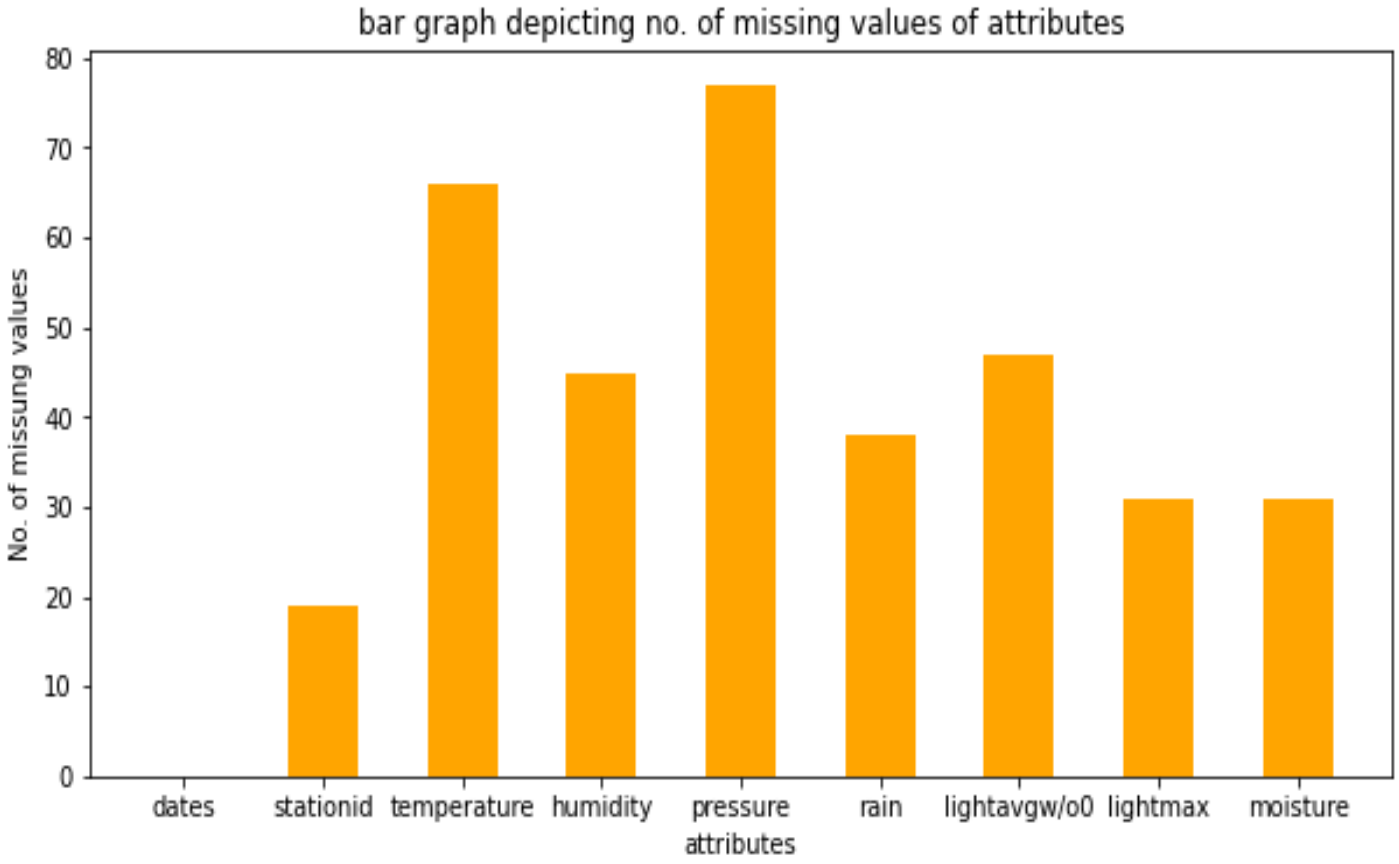

Figure 1 Number of missing values vs. attributes

Inferences:

1. Attribute 'pressure' is having maximum and attribute 'dates' is having minimum missing values.
2. Attribute 'pressure' is having highest frequency and attribute 'dates' is having lowest frequency.

2 a.

Inferences:

1. We chose to delete the tuple if the target attribute is missing because data will be of no use is station id is not known.
2. The number of tuples deleted after this step-19.
3. Percentage of the total number of tuples deleted is 2.02%.

b.

Inferences:

1. The number of tuples deleted after this step-35.
2. Percentage of the total number of tuples deleted is 3.78%.
3. We lost about 3.78% of data after this step.
4. We did this step to clear the tuples in which more amount of data is missing to get almost clean data.

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. Attribute 'temperature' is having maximum and attributes 'dates','sationid' are having minimum missing values.
2. Maximum percentage of data missing is from 'temperature' and minimum percentage of data missing are from 'dates' and 'stanoid'.
3. The total number of missing attributes in the file- 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								
2	stationid								
3	temperature (in °C)	21.079	12.727	22.111	4.399	21.078	21.078	21.8	4.243
4	humidity (in $g.m^{-3}$)	83.262	99	91.367	18.412	83.262	99	90.119	17.967
5	pressure (in mb)	1009.225	789.393	1014.932	47.180	1009.225	1009.225	1014.070	45.214
6	rain (in ml)	10942.726	0	15.75	25084.313	10942.726	0	24.75	24574.252
7	lightavgw/o0 (in lux)	4430.928	4488.910	1461.774	7591.994	4430.927	4488.910	1911.233	7400.586
8	lightmax (in lux)	21650.163	4000	6569	22043.154	21650.163	4000	7544	21678.196
9	moisture (in %)	32.671	0	13.910	33.978	32.671	0	17.723	33.415

Inferences:

1. Mean-max=lightmax , min=temperature
2. Mode-max=lightavgw/o0, min=pressure
3. Median-max= pressure , min= -
4. Standard deviation- max=lightmax, min=temperature

ii.

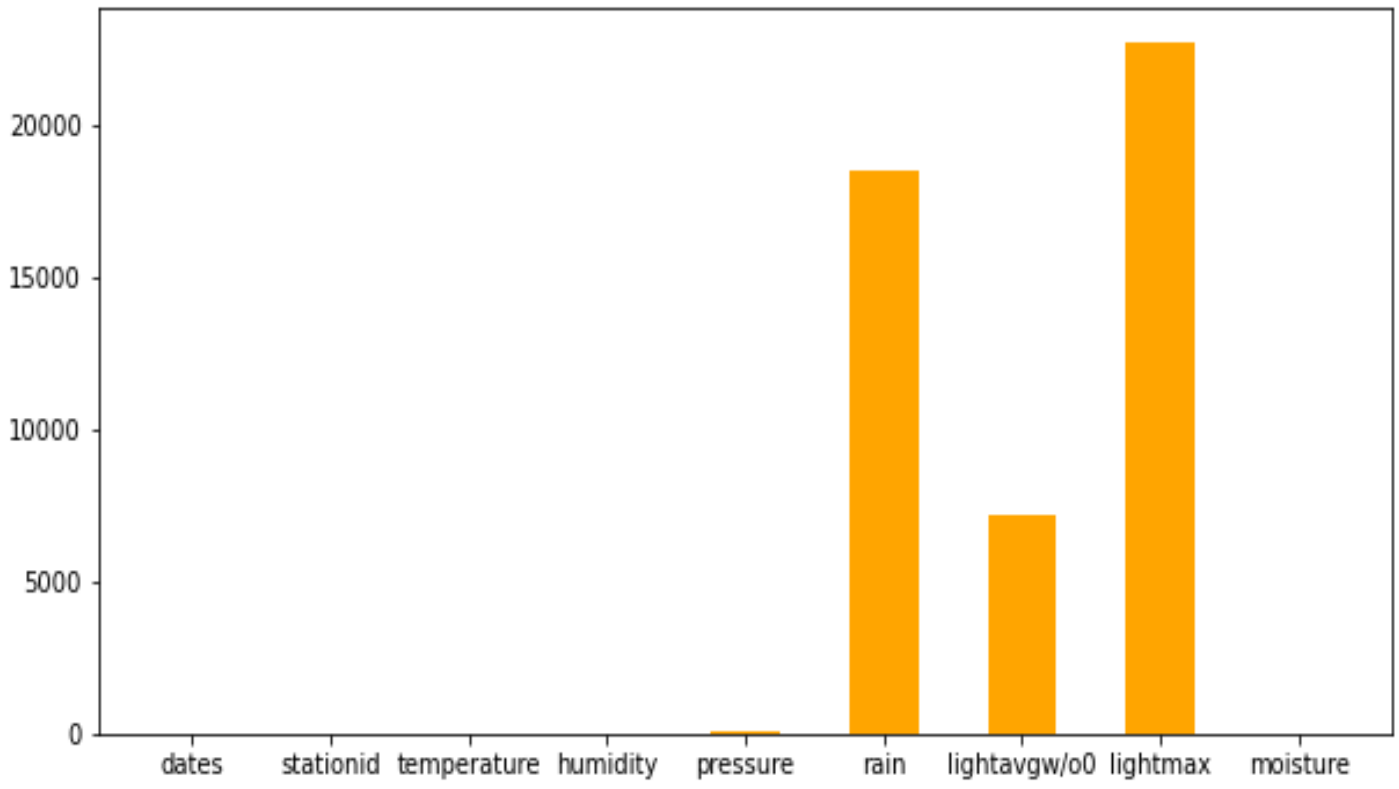

Figure 2 RMSE vs. attributes

Inferences:

1. Max RMSE=lightmax.

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								
2	stationid								
3	temperature (in °C)	21.079	12.727	22.111	4.399	21.196	12.727	22.169	4.329
4	humidity (in $g.m^{-3}$)	83.262	99	91.367	18.412	83.538	99	91.380	18.206
5	pressure (in mb)	1009.225	789.393	1014.932	47.180	1009.264	789.392	1014.677	45.998

6	rain (in ml)	10942.726	0	15.75	25084.313	10651.638	0	22.5	24779.512
7	lightavgw/o0 (in lux)	4430.928	4488.910	1461.774	7591.994	4486.340	4488.910	1623.494	7573.795
8	lightmax (in lux)	21650.163	4000	6569	22043.154	21517.191	4000	6569	21935.165
9	moisture (in %)	32.671	0	13.910	33.978	32.327	0	16.306	33.602

Inferences:

1. Mean-max=lightmax , min=temperature
2. Mode-max=lightavgw/o0, min=temperature
3. Median-max= - , min= -
4. Standard deviation- max=lightmax, min=temperature

ii.

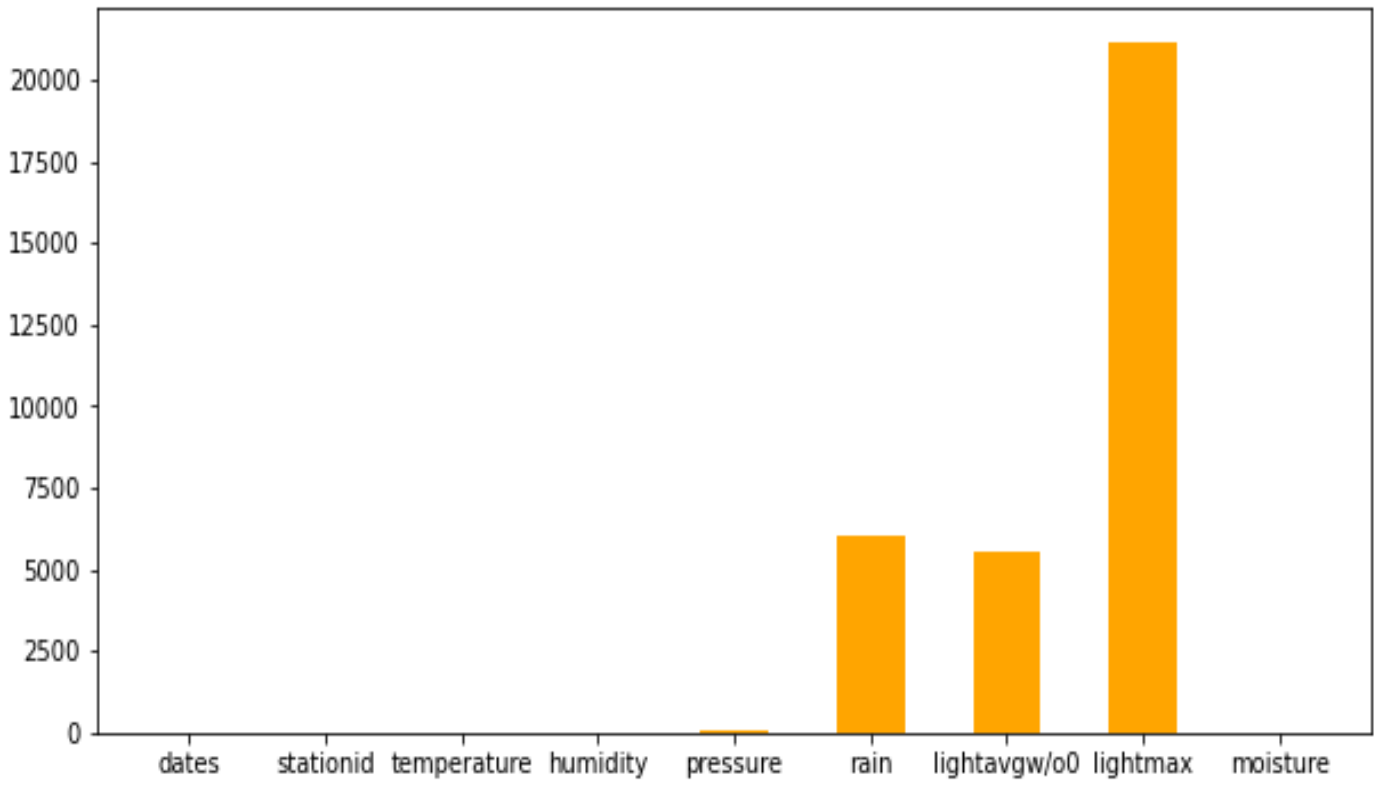

Figure 3 RMSE vs. attributes

Inferences:

1. Max RMSE=lightmax.

5 a.

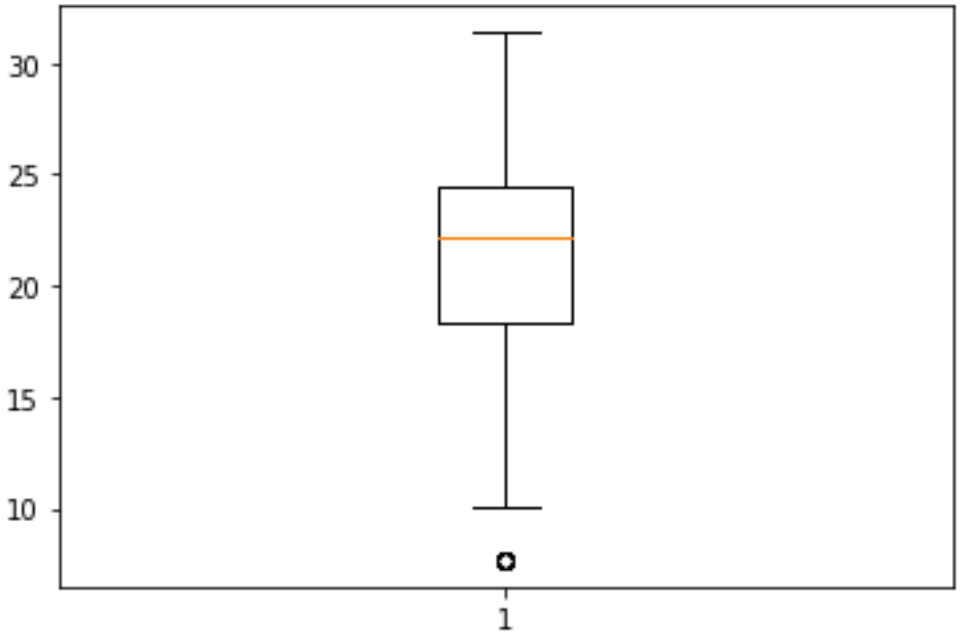

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. 1 outlier is present in 0-5 range.
2. It is not much spreaded.
3. It is positively skewed.

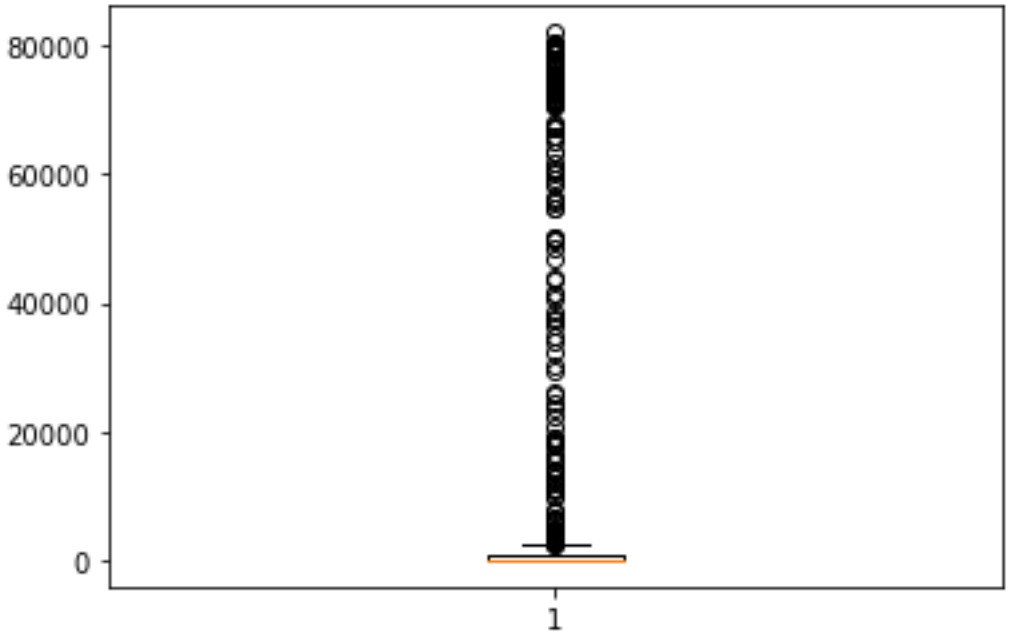

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. There are many outliers in range 0-80000.
2. It is not spreaded as box is close to 0.
3. It is negatively skewed.

b.

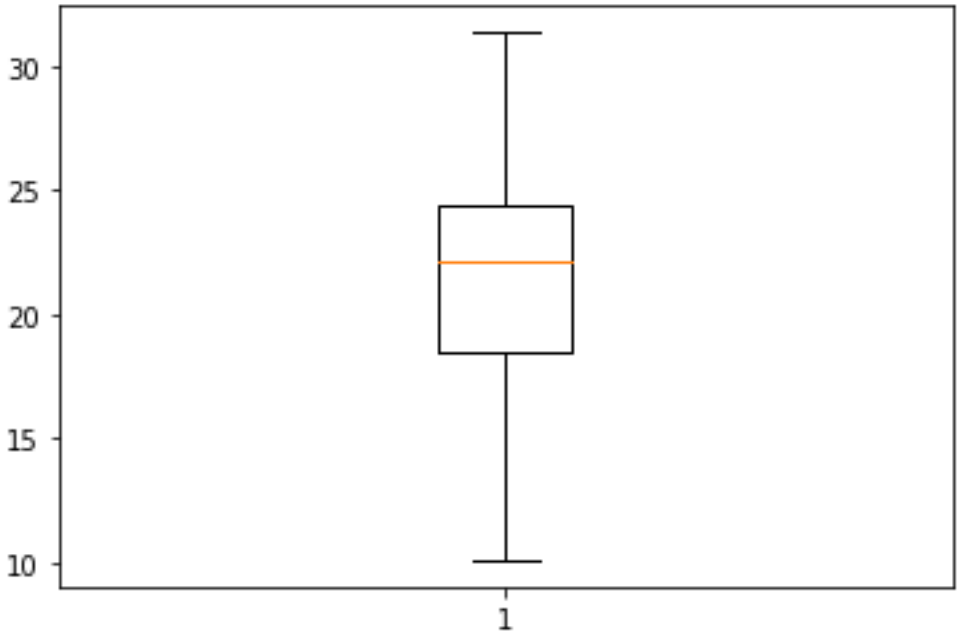

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. No outlier is present.
2. It is not much spreaded.
3. It is weak positively skewed.

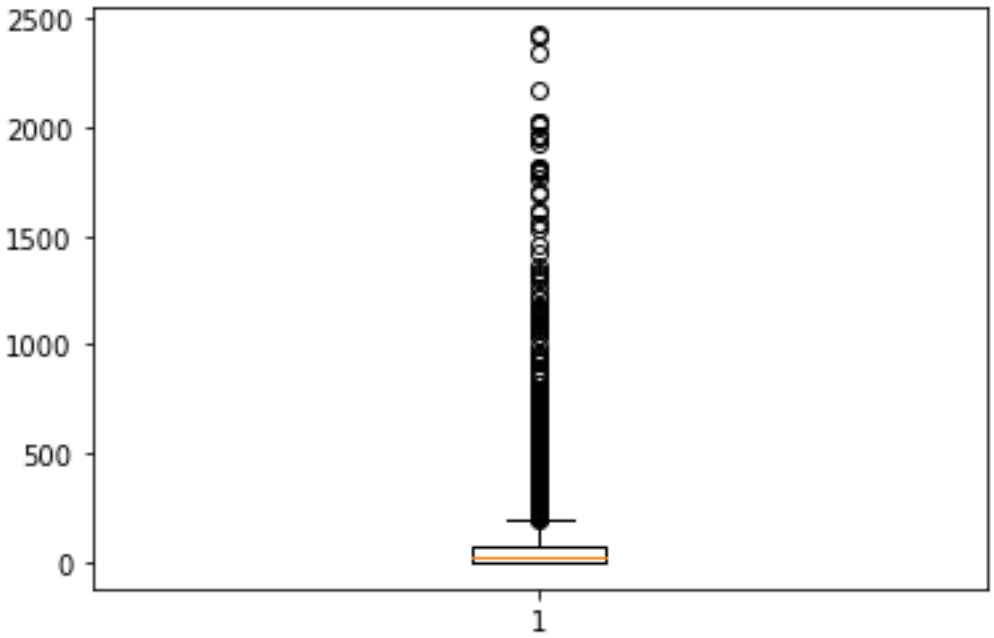

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. There are many outliers in range 0-2500.
2. It is not spreaded as box is close to 0.
3. It is negatively skewed.