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

Student's Name: Gaurav Kumar **Mobile No: 8529143452**

Roll Number: B20197 **Branch:**EE

1

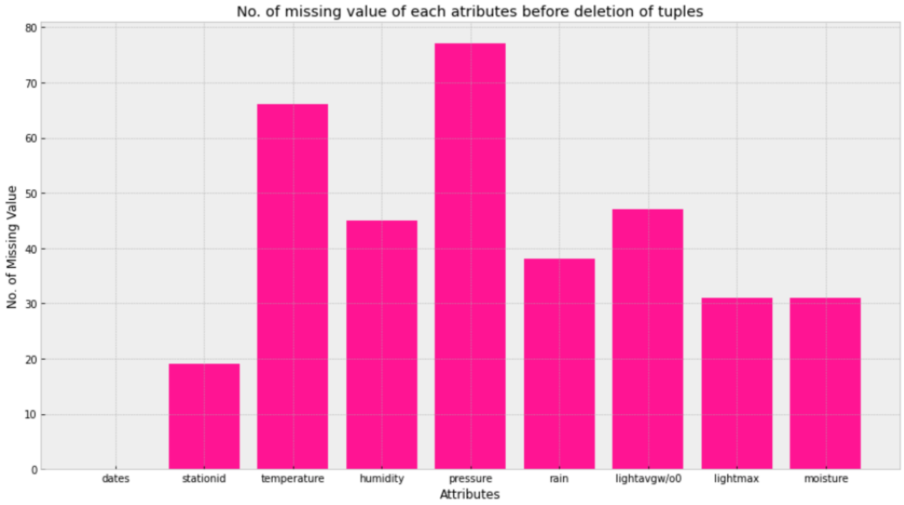

Figure 1 Number of missing values vs. attributes

Inferences:

1. Pressure have maximum and dates have minimum
2. 'temperature' and 'pressure' have distinctively high number of missing values while 'stationid' has lowest number of missing values, the other attributes have more or less comparable missing values around 35.

2 a.

Inferences:

1. 'stationid' is the target variable here. All the other attributes have values that are defined only in the context of this because they are readings of the sensors whose id is given by the target variable itself. To elaborate, say an ML model is being trained to recognize station by given readings or to predict attribute values for a station. In this context rows having no 'stationid' are useless because they cannot be used.
2. Number of rows deleted = 19.

3. 19 x 100/945 = 2.010 % rows have been deleted

b.

Inferences:

1. 35 rows have been deleted after this step.
2. 35 x 100/945 = 3.703 % rows deleted.
3. Compared to original data set, only 3.703 + 2.010 = 5.713 % data has been lost, which is not quite significant considering that those rows had less than 2/3 of the attribute values present in them according to the program, hence data loss can be safely ignored.
4. Given the condition that rows having more than or equal to 1/3 of attribute values missing must be dropped (>= 3 here), this is a necessary step, such rows have little informative value and most likely cannot be representative of the properties of the target variable / variable in context ; say 'stationid'

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)	58
4	humidity (in $g.m^{-3}$)	38
5	pressure (in mb)	65
6	rain (in ml)	30
7	lightavgw/o0 (in lux)	39
8	lightmax (in lux)	24
9	moisture (in %)	27

Inferences:

1. 'pressure' has maximum missing values. 'Dates' and 'stationid' both have no missing values (minimum).

2. ‘stationid’ and ‘dates’ have no missing values. There are 891 rows left after previous steps so, temperature, humidity, pressure, rain, lightavgw/o0, lightmax and moisture have 6.509,9.935,5.80,2.680,3.485,2.144,2.412% values missing
3. Total missing values = 281

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)	21.214	12.727	22.272	4.356	21.016	21.016	21.73	4.264
4	humidity (in $g.m^{-3}$)	83.480	99.0	91.380	18.210	83.048	99.0	90.021	17.960
5	pressure (in mb)	1009.00	789.392	1014.677	46.980	1009.279	1009.279	1014.270	45.418
6	rain (in ml)	10701.538	0.0	18.0	24852.255	11299.279	0.0	22.5	24978.192
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.88	7375.16	4480.880	0.0	1829.648	24978.192
8	lightmax (in lux)	21788.623	6634.0	6634.0	4000	21694.087	4000.0	6921.0	21797.706
9	moisture (in %)	32.386	0.0	16.704	33.653	32.781	0.0	17.085	33.531

Inferences:

1. Maximum change in mean, mode, median and SD are ‘rain’,’lightmax’ ,’lightmax’ ,’lightmax’ respectively. Minimum change in mean, median, mode and SD are in ‘temperature’ ,’humidity’ ,’pressure’ ,’moisture’ respectively.
2. yes, attributes having maximum no. of missing values have generally maximum relative change in mean, mode, median, SD. And attributes having minimum no. of missing values have generally minimum relative change in mean, mode, median, SD.

ii.

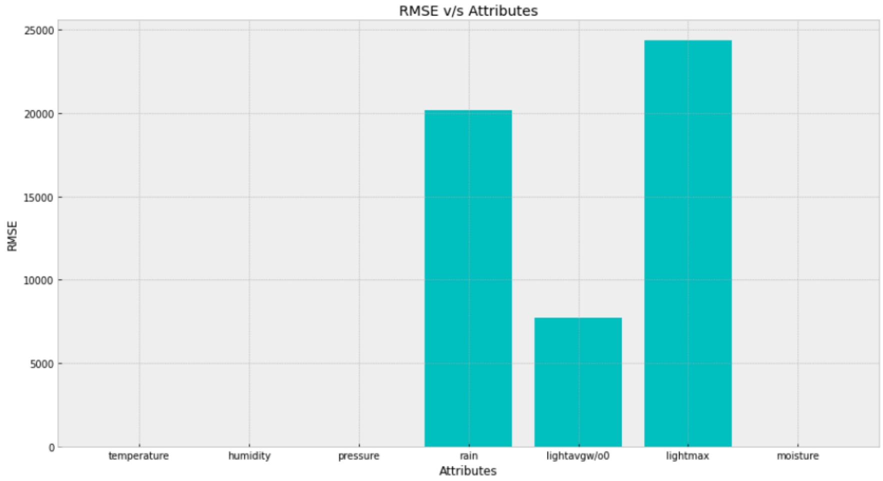

Figure 2 RMSE vs. attributes

Inferences:

1. ‘lightmax’ have maximum and ‘temperature’ have minimum RMSE value.
2. Attribute whose mean, mode, median, SD change maximum and have least no. of missing values have high RMSE and vice-versa.
3. For attributes , ’temperature’ ,’humidity’, ’pressure’ , ‘moisture’ data is reliable for further investigation.

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								
2	stationid								
3	temperature (in °C)	21.214	12.727	22.272	4.356	21.144	12.727	22.169	4.354
4	humidity (in $g.m^{-3}$)	83.480	99.0	91.380	18.210	83.255	99.0	91.0	18.180
5	pressure (in mb)	1009.00	789.392	1014.677	46.980	1009.651	789.392	1014.881	45.728
6	rain (in ml)	10701.538	0.0	18.0	24852.255	11075.10	0.0	20.25	25175.756
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.88	7375.16	4561.720	4488.910	1623.494	7697.950

8	lightmax (in lux)	21788.623	6634.0	6634.0	4000	21709.493	4000.0	6569.0	22020.154
9	moisture (in %)	32.386	0.0	16.704	33.653	32.461	0.0	15.138	33.722

Inferences:

1. Maximum change in mean, mode, median and SD are ‘rain’,’lightmax’ ,’lightmax’ ,’lightmax’ respectively. Minimum change in mean, median, mode and SD are in ‘temperature’ ,’humidity’ ,’pressure’ ,’moisture’ respectively.
2. yes, attributes having maximum no. of missing values have generally maximum relative change in mean, mode, median, SD. And attributes having minimum no. of missing values have generally minimum relative change in mean, mode, median, SD.

ii.

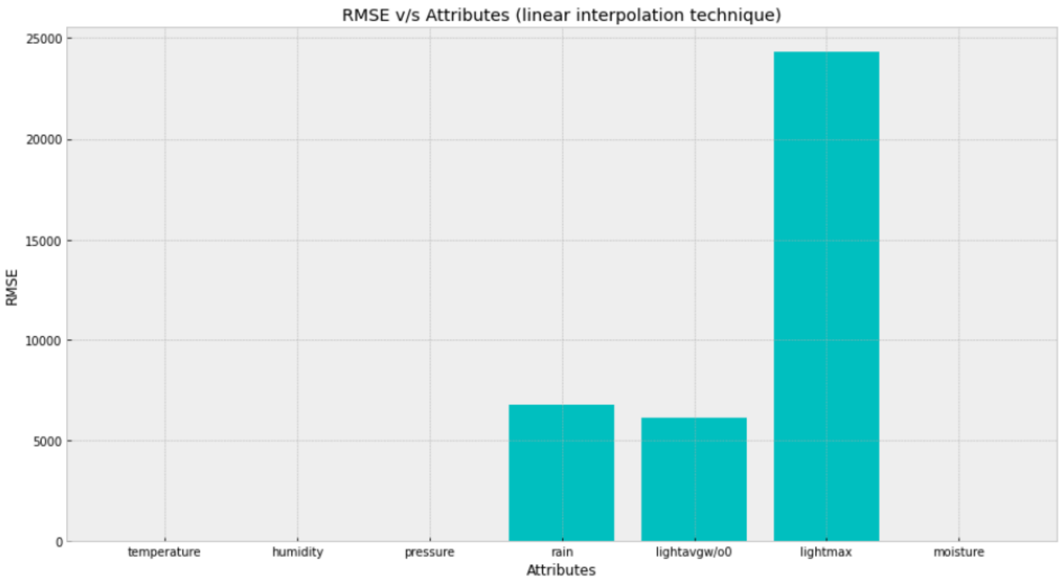

Figure 3 RMSE vs. attributes

Inferences:

1. ‘lightmax’ have maximum and ‘temperature’ have minimum RMSE value.
2. Attribute whose mean, mode, median, SD change maximum and have least no. of missing values have high RMSE and vice-versa.

3. For attributes , ’temperature’ ,’humidity’, ’pressure’ , ‘moisture’ data is reliable for further investigation.

5 a.

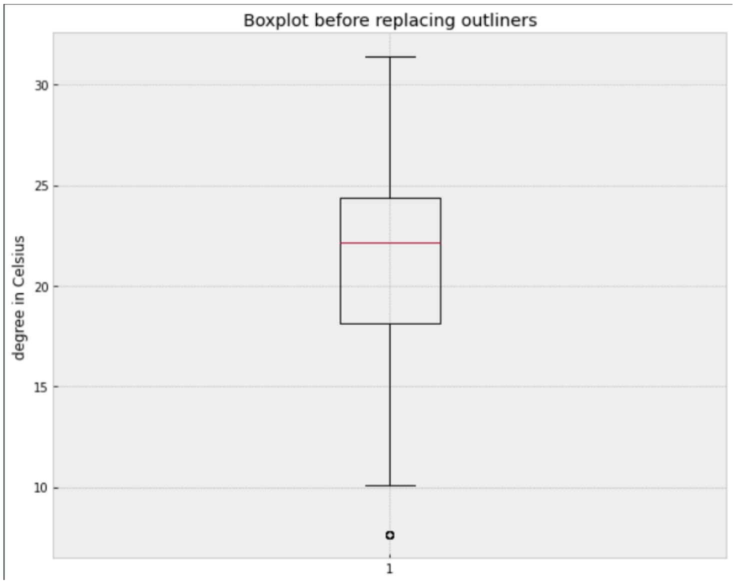

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. No. of outliners = 10
2. IQR = 6.278
3. IQR denotes spread/variance.
4. Infer the skewness of the data.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units.

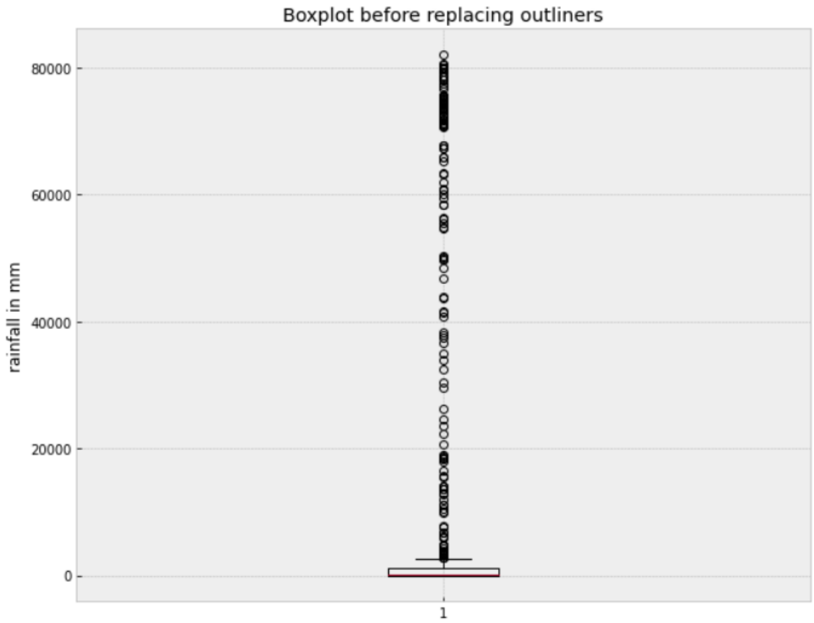

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. No. of outliners = 180
2. IQR = 1093.5
3. IQR denotes spread/variance.
4. Infer the skewness of the data.
5. Inference 5 (You may add or delete the number of inferences)

   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units.

b.

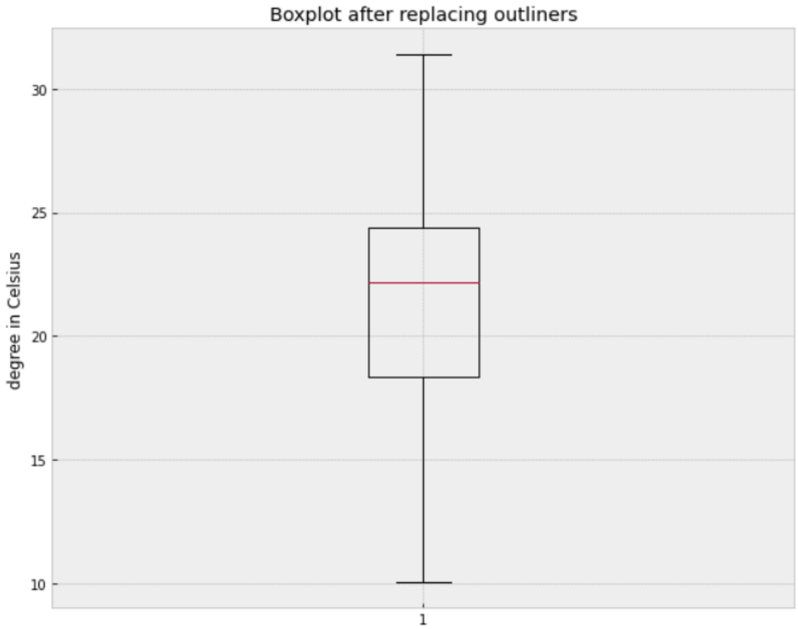

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. List the number of outliers, their row number and compare with Q5. a.
2. Infer the Inter quartile range compare with Q5. a.
3. Infer the spread/variance compare with Q5. a.
4. Infer the skewness of the data compare with Q5. a.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units

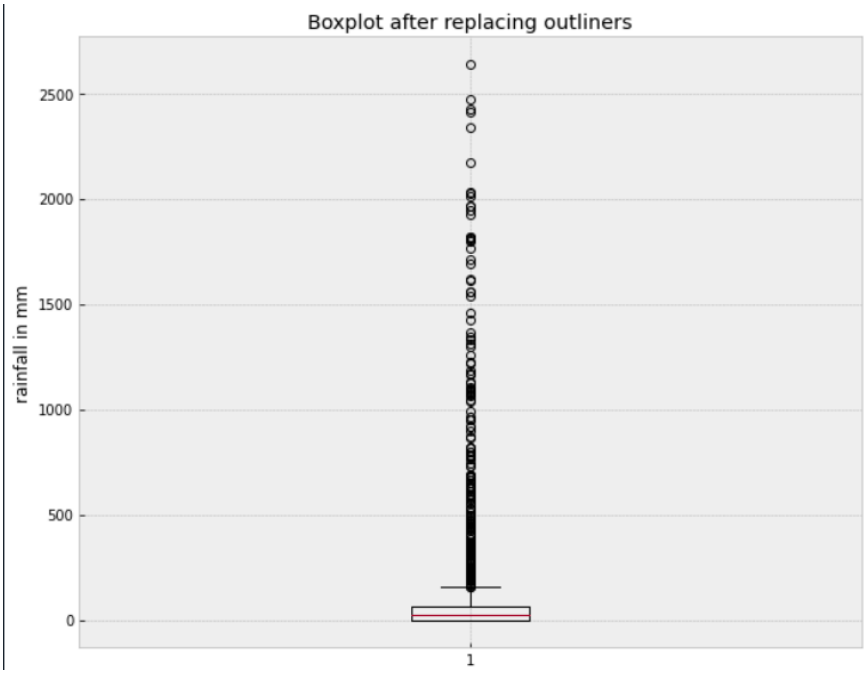

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. List the number of outliers, their row number and compare with Q5. a.
2. Infer the Inter quartile range compare with Q5. a.
3. Infer the spread/variance compare with Q5. a.
4. Infer the skewness of the data compare with Q5. a.
5. Inference 5 (You may add or delete the number of inferences)

   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units