Indian Institute of Technology Mandi

IC 272: DATA SCIENCE - III

LAB ASSIGNMENT - II

Data cleaning – handling missing values and outlier analyses

Student's Name: Arpit Singh **Mobile No: 6265104315**

Roll Number: B20084 **Branch:CSE**

1

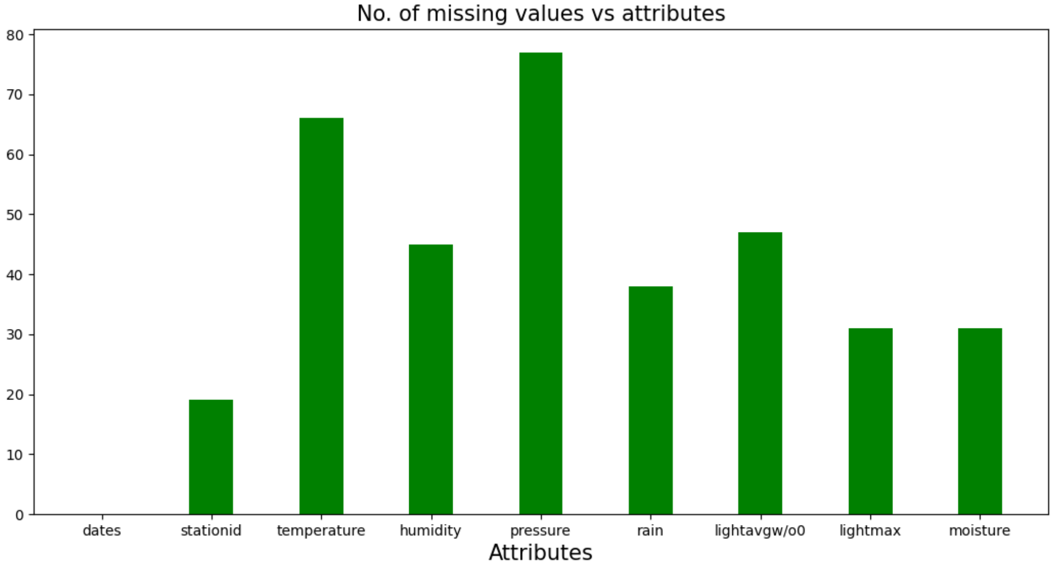

Figure 1 Number of missing values vs. attributes

Inferences:

1. Pressure and dates have maximum and minimum no. of missing values respectively.
2. Dates have 0 no. of missing values. stationid have 2nd lowest no. of missing values 19. lightmax and moisture have equal no. of missing values i.e, 31. pressure and temperature have comparatively higher no. of missing values than other attributes 66, 77 respectively. no. of missing values of attributes, humidity, rain, lightavgw/o0 lies in the range 35-50.

2 a.

Inferences:

1. Target attribute “stationid” which is the location of sensors. If the location is missing in data we can’t determine it by any means so we will have to drop the tuple for the missing values of the targeted attribute.
2. Number of tuples deleted after this step is 19.
3. 2.01 percent of the total number of tuples is deleted.

b.

Inferences:

1. Number of tuples deleted after this step is 35.
2. 3.78 percent of the total number of tuples is deleted in this step.
3. 3.78 percent of tuples were removed because it’s data was missing which is a small percentage of data and missing values can result in errors in our interpretation of the data.
4. Deletion of tuples with missing values was required as it can give errors so it was a necessary step.

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. Dates, stationid have minimum missing values and pressure have maximum missing values.
2. For dates and stationid 0 percent data is missing and for temperature, humidity, pressure, rain, lightavgw/o0, lightmax, moisture 3.81, 1.46 ,4.6, 0.673, 1.68 , 0.112, 0.673 percent data is missing respectively.
3. The total number of missing attributes in the file is 7.

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)	21.215	12.727	22.272	4.355	21.078	21.078	21.8	4.243
4	humidity (in $g.m^{-3}$)	83.479	99.0	91.380	18.210	83.261	99.0	90.119	18.134
5	pressure (in mb)	1009.008	789.392	1014.677	46.980	1009.339	1009.339	1014.070	46.032
6	rain (in ml)	10701.538	0.0	18	24852.255	10663.726	0.0	18	24663.252
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.88	7573.162	4440.927	4488.910	1579.859	7554.680
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21528.763	4000	6569	21983.526
9	moisture (in %)	32.386	0.0	16.704	33.653	32.671	0.0	13.982	33.859

Inferences:

1. Attributes having maximum and the minimum change in the mean are moisture and pressure. In mode temperature have maximum change and humidity, rain, lightavgw/o0, lightmax, moisture have 0 change. In the median rain and pressure have maximum and the minimum change. In standard deviation pressure and moisture respectively.
2. The one having maximum change is mean, mode is also having minimum change in standard deviation.

3. The mean mode median are nearly same in some attributes and are very far in some others which makes the data unreliable.

ii.

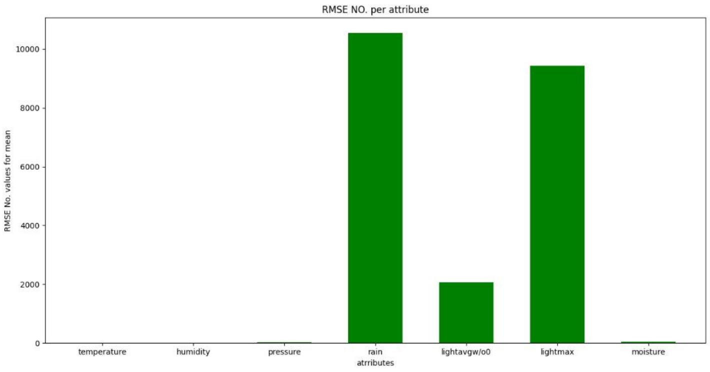

Figure 2 RMSE vs. attributes

Inferences:

1. Attribute Temperature has minimum and rain has maximum RMSE respectively.
2. We can see that rmse no. is the sqrt of the submission of difference between the original and filled values of the data and we can clearly see that for rain mode of data before and after varies with a large difference and for attributes with small values of RMSE there are negligible changes in median and mean of data. Except for temp for which median varies enough and we can see that the rmse value of that is also more than other small value attribute. And after normalisation the values with more no. of missing values have higher rmse values and for less we have lower rmse values.
3. For attributes with small values of rmse no. it's eligible for further investigation but for values bigger than others it’s not that much appreciable to do analysis.

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)	21.215	12.727	22.272	4.355	21.074	12.074	22.959	4.331

4	humidity (in g.m^{-3})	83.479	99.0	91.380	18.210	83.204	99.0	90.980	18.314
5	pressure (in mb)	1009.008	789.392	1014.677	46.980	1009.339	1009.339	1014.513	46.032
6	rain (in ml)	10701.538	0.0	18	24852.255	10663.616	0.0	18	24663.876
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.88	7573.162	4440.397	4488.910	1579.859	7554.680
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21528.781	4000.0	6569.0	21983.565
9	moisture (in %)	32.386	0.0	16.704	33.653	32.665	0.0	13.982	33.859

Inferences:

1. Lightmax and temperature have the maximum and the minimum change in the mean respectively, any attribute doesn’t have change in the mode, lightavgw/o0 and pressure temperature have the maximum and the minimum change in the median respectively, lightmax and temperature temperature have the maximum and the minimum change in the standard deviation respectively
2. mean, mode has maximum change, also having minimum change in standard deviation.
3. 3. Yes this data is reliable as it does not have many changes.
4. Replacing missing values with mean makes huge changes in mode and median which makes the data more irreliable for analysis, while in interpolation it doesn’t happen

ii.

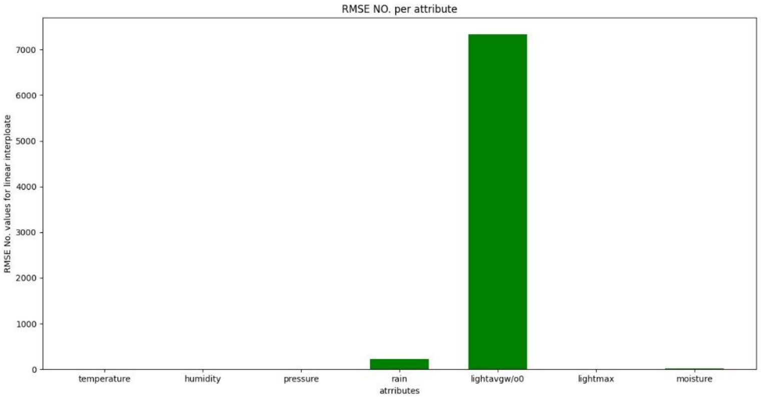

Figure 3 RMSE vs. attributes

Inferences:

1. Attribute lightavgw/o0 has maximum and temperature has minimum RMSE respectively.
2. We can see that rmse no. is the sqrt of the submission of difference between the original and filled values of the data and we can clearly see that for lightmax the median of data before and after varies with a little difference and for attributes with small values of RMSE there are negligible changes in median and mean of data.
3. For attributes with small values of rmse no. it is eligible for further investigation and the data we obtained here are better than the mean method used in the last part.
4. From the calculated RMSE of mean and linear inter-plotation method we can see that in case of linear inter-plotation method the values obtained are better than the values obtained from mean method technique but for lightavgw/o0 the values of rmse increases. After normalization we can see that the values with more no. of missing values the rmse ia higher and for less rmse is lower.

5 a.

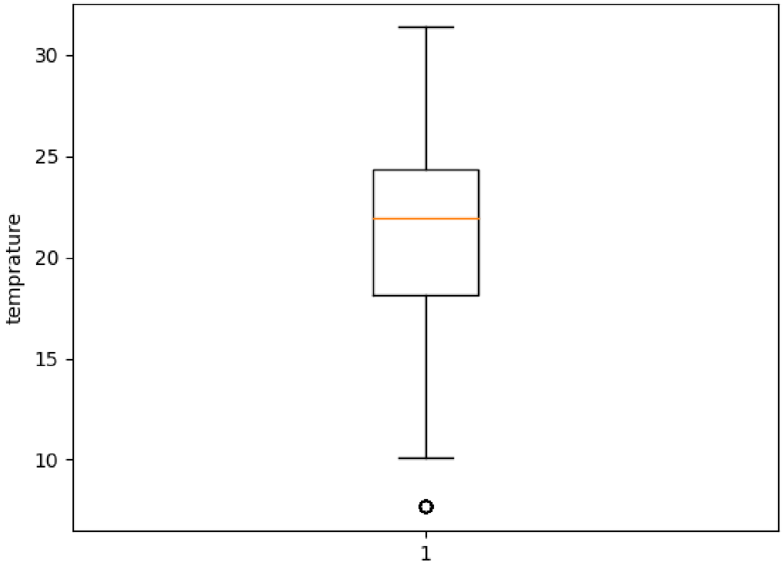

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. Number of outliers is 1.
2. The Inter quartile range is 6.1830475.
3. values of temperature vary from 5 to 33.
4. The data is left skewed.

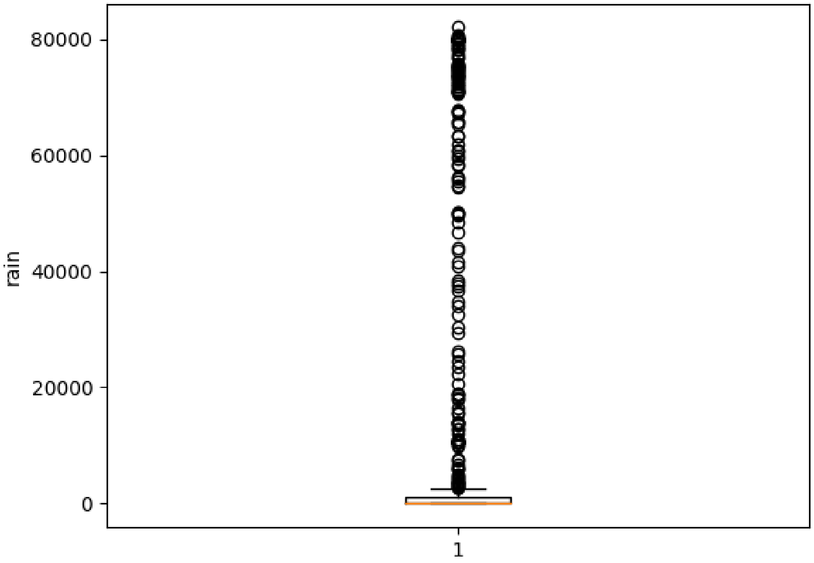

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. Rain has multiple no. of outliers.
2. The Inter quartile range is 1104.1875.
3. The value of rain varies from 0.0 to 82000.
4. The data is right skewed.

b.

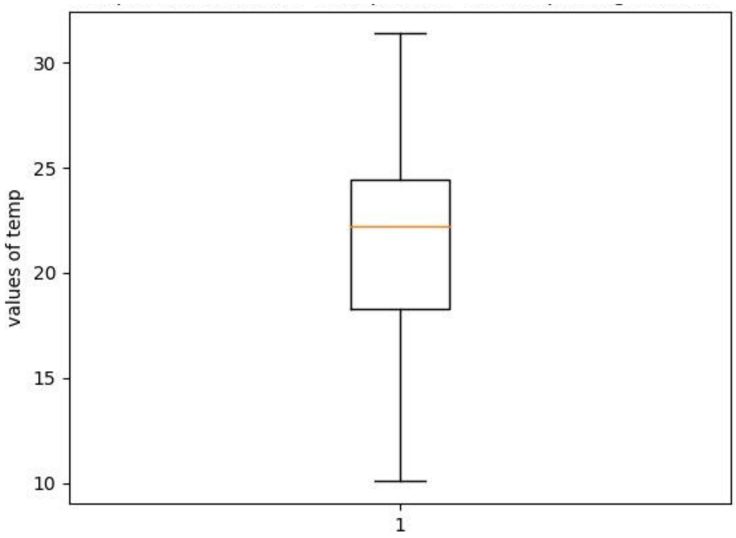

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. No outliers.
2. The Inter quartile range is 6.1830475.
3. values of temperature vary from 10 to 33.
4. The data is left skewed.

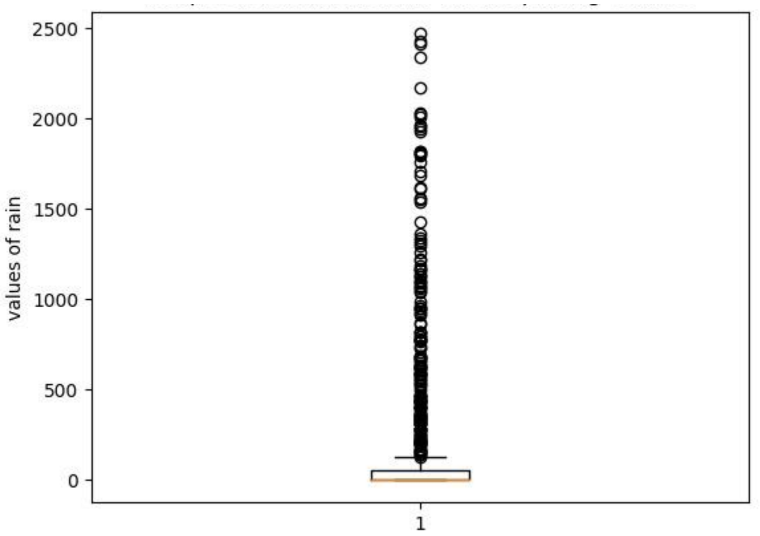

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. Rain has multiple no. of outliers.
2. The Inter quartile range is 1104.1875.
3. The value of rain varies from 0.0 to 2400.
4. The data is right skewed.