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

Student's Name: SARTHAK SAPTAMI KUMAR JHA **Mobile No: 8825319259**

Roll Number: B20317 **Branch: CSE**

1

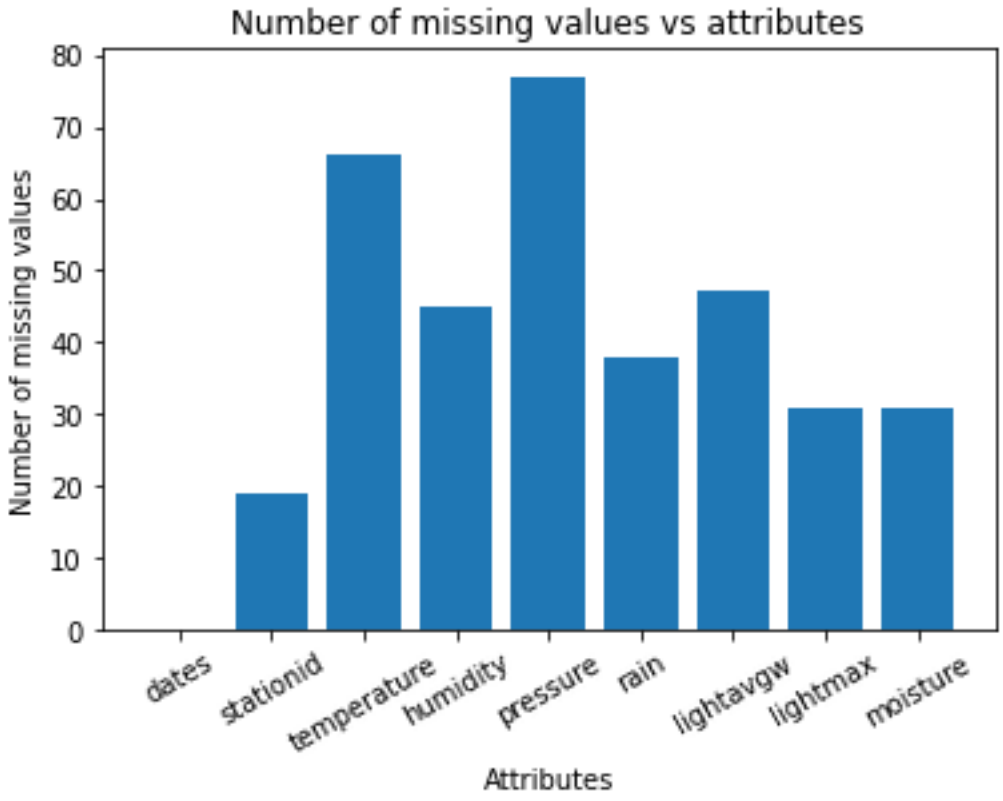

Figure 1 Number of missing values vs. attributes

Inferences:

1. The maximum number of missing values are in pressure and minimum in stationid.
2. Each attribute has at least 30 missing readings, and there are no missing dates in the data.
3. Inference 3(You may add or delete the number of inferences)

2 a.

Inferences:

1. We must delete the tuples where stationid is missing, because it signifies the location from where the data is taken, and so without it, the data is of no use.
2. A total of 19 tuples were deleted in this step.
3. 2% of data was deleted in this step.

b.

Inferences:

1. Another 30 tuples are deleted in this step.
2. 3.23% of data was deleted in this step.
3. The data loss is quite significant in this step.
4. This step is needed to erase tuples which have too few data to do any meaningful analysis.

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)	37
4	humidity (in $g.m^{-3}$)	16
5	pressure (in mb)	45
6	rain (in ml)	7
7	lightavgw/o0 (in lux)	17
8	lightmax (in lux)	2
9	moisture (in %)	7

Inferences:

1. Pressure has the maximum number of missing values and lightmax has the minimum number of missing values, while data and stationid have no missing values.
2. Temperature, humidity, pressure and lightavg contribute the highest percentage of missing values.

3. A total of 131 values are missing in the data.

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		19-07-2018				19-07-2018		
2	stationid		t9				t9		
3	temperature (in °C)	21.21	12.73	22.27	4.35	21.05	21.05	21.92	4.33
4	humidity (in $g.m^{-3}$)	83.48	99.00	91.38	18.21	83.14	99.00	90.86	18.34
5	pressure (in mb)	1009.01	789.39	1014.68	46.98	1009.47	1009.47	1014.43	45.72
6	rain (in ml)	10701.54	0.00	18.00	24852.25	10860.54	0.0	16.87	24878.70
7	lightavgw/o0 (in lux)	4438.43	4488.91	1656.88	7573.16	4451.45	4488.91	1516.01	7588.04
8	lightmax (in lux)	21788.62	4000.00	6634.00	22064.99	21498.31	4000.00	6569.00	21954.04
9	moisture (in %)	32.39	0.0	16.70	33.65	32.58	0.0	14.25	33.73

Inferences:

1. **Mean –** Maximum change is in rain at -1.48% and minimum change is in pressure at -0.04%
   Median – Maximum change is in moisture at 14.67% and minimum change is in pressure at 0.02%
   Mode – Maximum change is in temperature at -65.35% and minimum change is in various attributes at 0%
   Standard deviation – Maximum change is in pressure at 2.66% and minimum change is in rain at -0.10%
2. The least amount of change is visible in pressure, which had the maximum number of missing values.
3. The change in the statistics of data is quite high in some attributes like temperature and moisture and so can't be considered as reliable for further investigation.

ii.

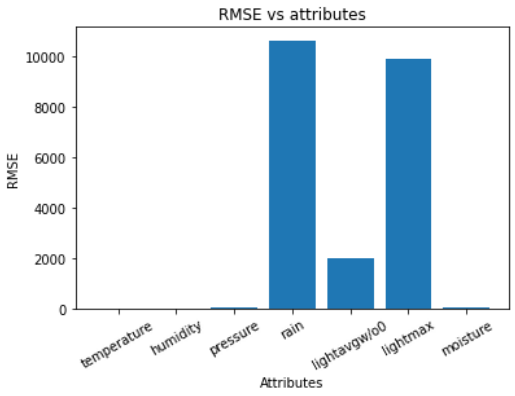

Figure 2 RMSE vs. attributes

Inferences:

1. Rain has maximum RMSE and temperature has minimum RMSE.
2. Since rain has maximum change in its mean, so it has the highest values of RMSE, the other attributes don't have an effect on RMSE.
3. Since the RMSE values are very large for at least 3 attributes, so we can't claim the data to be reliable.
4. Since Rain has maximum variation, replacing missing values with mean, leads to very high RMSE values, same is true for lightavg and lightmax, and we can say that standard deviation has the highest effect on RMSE values.

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		19-07-2018				19-07-2018		
2	stationid		t9				t9		
3	temperature (in °C)	21.21	12.73	22.27	4.35	21.11	12.72	22.15	4.39
4	humidity (in $g.m^{-3}$)	83.48	99.00	91.38	18.21	83.15	99.00	91.06	18.37
5	pressure (in mb)	1009.01	789.39	1014.68	46.98	1009.94	789.39	1014.93	45.91
6	rain (in ml)	10701.54	0.00	18.00	24852.25	10777.98	0.0	15.75	24896.12
7	lightavgw/o0 (in lux)	4438.43	4488.91	1656.88	7573.16	4492.28	4488.91	1501.71	7631.52

8	lightmax (in lux)	21788.62	4000.00	6634.00	22064.99	21497.18	4000.00	6569.00	21959.03
9	moisture (in %)	32.39	0.0	16.70	33.65	32.49	0.0	13.91	33.81

Inferences:

1. **Mean –** Maximum change is in lightmax at -1.33% and minimum change is in pressure at 0.09%
   Median – Maximum change is in moisture at -16.72% and minimum change is in pressure at 0.02%
   Mode – Maximum change is in temperature at -0.07% and minimum change is in various attributes at 0%.
   Standard deviation – Maximum change is in pressure at -2.26% and minimum change is in rain at -0.17%
2. The least amount of change is visible in pressure, which had the maximum number of missing values.
3. Yes, the data is reliable for further investigation as the difference in data statistics from the original data is very less even in cases of maximum change, and also in various attributes, the change is very less, and so the data can be considered reliable for further investigation.
4. From the above changes in data from filling of values using mean and interpolation, we can say that, filling missing values using mean degrades the statistical performance of the data, and we can clearly infer that, interpolation method is better for filling of missing values.

ii.

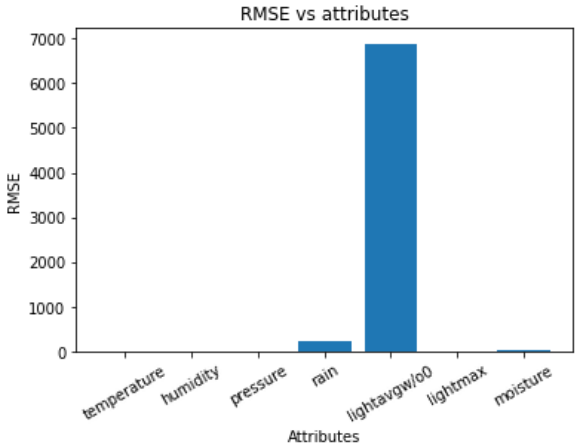

Figure 3 RMSE vs. attributes

Inferences:

1. lightavgw/o0 has maximum RMSE, and lightmax has minimum RMSE.

2. There is very little relation between change in statistics of the data and its RMSE values.
3. Since the RMSE values are very low for all attributes except lightavg, we can say that the data is reliable for further investigation.
4. The use of interpolation for replacing missing values leads to much lower values of RMSE for all attributes, when compared to mean method.

5 a.

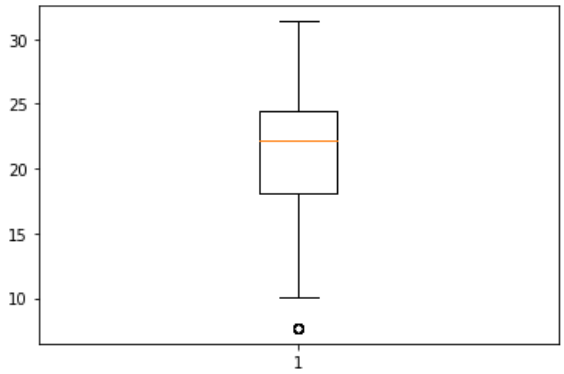

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers of values 7.6729 each. They are located at row 511 to 520.
2. The inter quartile range is 6.37.
3. The variance is 19.27.
4. The data is skewed towards higher temperatures.
5. There are very few outliers, all having the same values, and so are represented by a single circle in the boxplot.

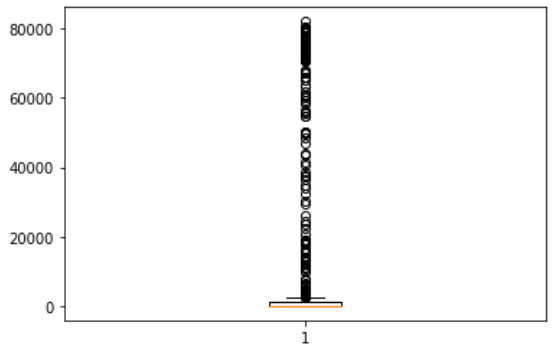

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. There are 177 outliers with a large number of them having indices above 650.
2. The inter quartile range is 1048.5.
3. The variance is 619817205.84.
4. The data is heavily skewed towards smaller values
5. A large number of outliers are present above 10000ml and need to be removed to get a better understanding of the data.

b.

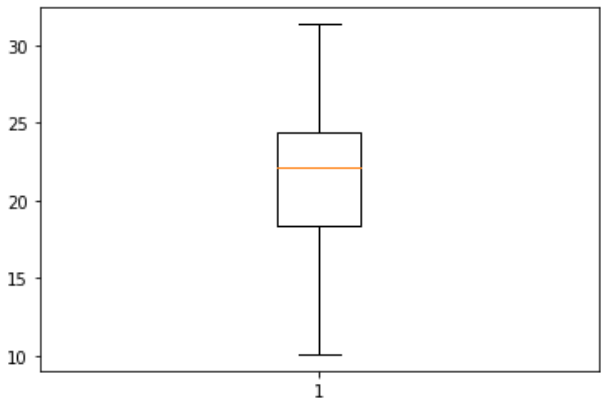

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are 0 outliers.
2. The inter quartile range is 6.08
3. The variance is 17.24
4. The data is lightly skewed.
5. In the temperature data, we have successfully removed all outliers in one pass.

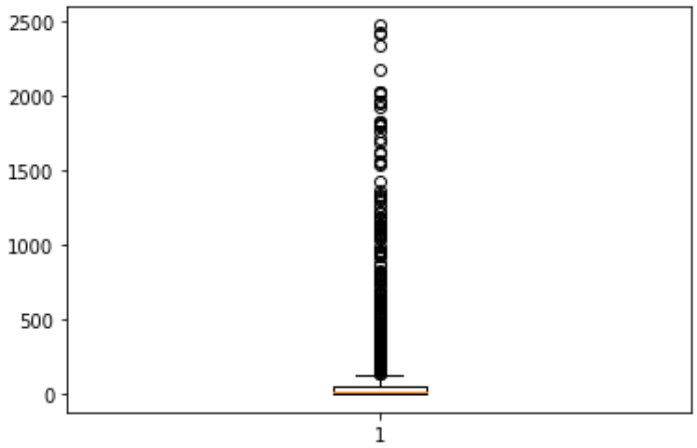

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. There are 182 outliers after removing the outliers.
2. The inter quartile range is 51.75.
3. The variance is 156322.01.
4. The data is heavily skewed towards lower values.
5. There are still a large number of outliers even after removing the outliers in one pass, because the data is heavily spread, so according to the new quartiles, there are still many outliers. Though, a large number of outliers have been removed as previously a large number of outliers were above 10000, but now all outliers are limited to below 2500.