IC 272: DATA SCIENCE - III

LAB ASSIGNMENT - I

Data cleaning – handling missing values and outlier analyses

Student's Name: Amit Maindola **Mobile No: +91 7470985613**

Roll Number: B20079 **Branch: Computer Science & Engineering**

1

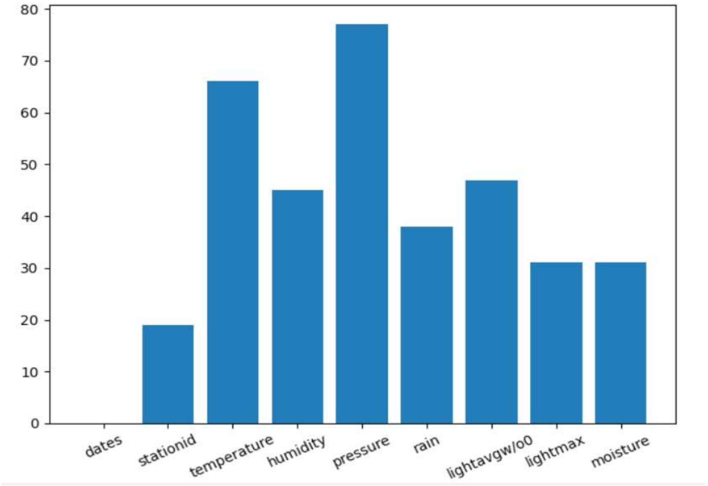

Figure 1 Number of missing values vs. attributes

Inferences:

1. 'pressure' and 'dates' attributes have maximum and minimum missing values respectively.
2. Number of missing values in following attributes are:
   dates: 0, **stationid:** 19, **temperature:** 66, **humidity:** 45, **pressure:** 77, **rain:** 38, **lightavgw/o0:** 47, **lightmax:** 31, **moisture:** 31.

2 a.

Inferences:

1. We choosed to delete the touples whose target attribute values are missing to remove the nuisancr in our data.
2. 19 tuples are deleted in this step.
3. 2.01 percentage of the total number of tuples is deleted.

b.

Inferences:

1. 35 tuples are deleted after this step.
2. 3.77 percentage of the total number of tuples is deleted.
3. Some data is lost in this step for other attributes having a non null value.
4. This step was needed for cleaning the data, as the tuples having 3 or more null values is deleted in this 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. ‘pressure’ has maximum and ‘dates’ and ‘stationid’ have minimum number of missing values.
2. Percantage of missing values in following attributes are:
   dates: 0, **stationid: 0**, **temperature: 3.81**, **humidity:** 1.45, **pressure:** 4.6, **rain:** 0.67, **lightavgw/o0:** 1.67, **lightmax: 0.11**, **moisture: 0.67**.
3. Total 116 values in table are missing.

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	21.079	21.800	4.243	21.215	12.727	22.273	4.356
4	humidity (in g.m^{-3})	83.262	99.000	90.119	17.968	83.480	99.000	91.381	18.210
5	pressure (in mb)	1009.225	1009.225	1014.071	45.215	1009.009	789.393	1014.678	46.980
6	rain (in ml)	10942.726	0.000	24.750	24574.253	10701.538	0.000	18.000	24852.255
7	lightavgw/o0 (in lux)	4430.928	4488.910	1911.234	7400.586	4438.428	4488.910	1656.880	7573.163
8	lightmax (in lux)	21650.163	4000.000	7544.000	21678.196	21788.623	4000.000	6634.000	22064.993
9	moisture (in %)	32.672	0.000	17.723	33.416	32.386	0.000	16.704	33.653

Inferences:

1. **Mean:** maximum change in 'rain' & minimum change in 'temperature'.
   Median: maximum change in 'pressure' & minimum change in 'temperature'.
   Mode: maximum change in 'lightmax' & minimum change in 'humidity', 'rain', 'lightmaxavgw/o0, 'lightmax', and 'moisture'.
   Standard deviation: maximum change in 'lightmax' & minimum change in 'temperature'.
2. There are minimum missing values in 'lightmax' and max. change in median and standard deviation, max missing value in 'pressure' and max. change in mode, there are second most maximum number of missing values in 'temperature' and minimum change in mean, median and standard deviation.
3. As we can notice for most of the attributes the change is minimum, so this data is reliable for further investigation.

ii.

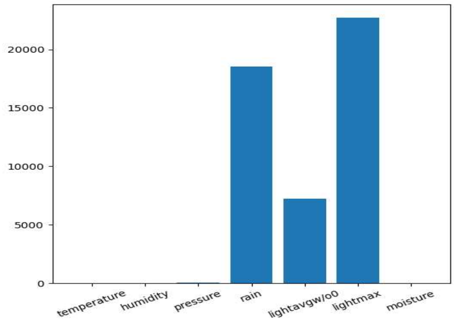

Figure 2 RMSE vs. attributes

Inferences:

1. ‘lightmax’ and ‘temperature’ attributes have maximum and minimum RMSE respectively
2. ‘lightmax’ has highest number of missing values and have maximum RMSE, while ‘temperature’ has lowest RMSE and second highest number of missing values
3. RMSE for three attributes is much higher than expected, and other have very low value. Due to high RMSE in those three attributes data is not reliable.

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)	21.196	12.727	22.169	4.329	21.215	12.727	22.273	4.356
4	humidity (in g.m^{-3})	83.538	99.000	91.381	18.207	83.480	99.000	91.381	18.210
5	pressure (in mb)	1009.265	789.393	1014.678	45.999	1009.009	789.393	1014.678	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)	4486.341	4488.910	1623.494	7573.795	4438.428	4488.910	1656.880	7573.163
8	lightmax (in lux)	2151	4000.0	6569.00	219	21788	4000.	6634.00	2206

		7.191	00	0	35.166	.623	000	0	4.993
9	moisture (in %)	32.327	0.000	16.307	33.603	32.386	0.000	16.704	33.653

Inferences:

1. **Mean:** maximum change in 'lightmax' & minimum change in 'temperature'.
   Mode: all chages are 0.
   Median: maximum change in 'lightmax' & minimum change in 'humidity' and 'pressure'.
   Standard deviation: maximum change in 'lightmax' & minimum change in 'humidity'.
2. There are minimum missing values in 'lightmax' and max. change in median and standard deviation, max missing value in 'pressure' and max. change in mode, there are second most maximum number of missing values in 'temperature' and minimum change in mean, median and standard deviation.
3. As we can notice for most of the attributes the changes are low, so this data is reliable for further investigation.

ii.

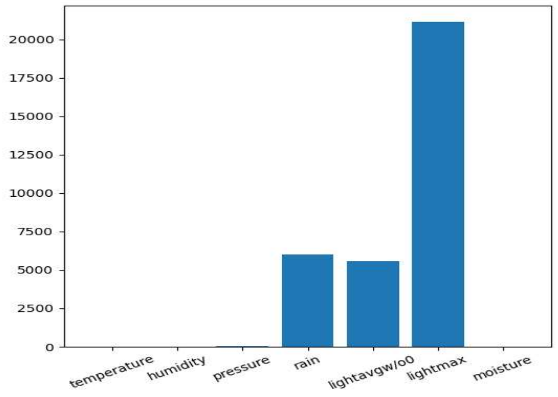

Figure 3 RMSE vs. attributes

Inferences:

1. 'lightmax' and 'temperature' attributes have maximum and minimum RMSE respectively
2. 'lightmax' has highest number of missing values and have maximum RMSE, while 'temperature' has lowest RMSE and second highest number of missing values

3. RMSE for three attributes is much higher than expected, and other have very low value. Due to high RMSE in those three attributes data is not reliable.
4. In this case replacing the missing values by interpolation is more effective than replacing by mean.

5 a.

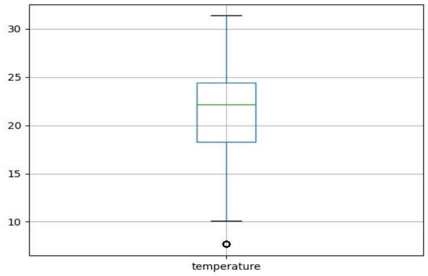

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. Number of ouliers are 10 with all value as 7.67 and row number 509 to 518.
2. Inter quartile range is 6.1.
3. Data is spreaded in the range of 7.67 to 31.375.
4. Data is negetively skewed.

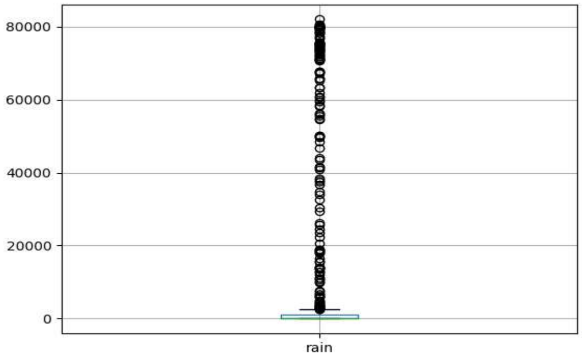

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. Number of ouliers are 185.
2. Inter quartile range is 987.75.
3. Data is spreaded in the range of 22.5 to 82037.25.
4. Data is positively skewed.

b.

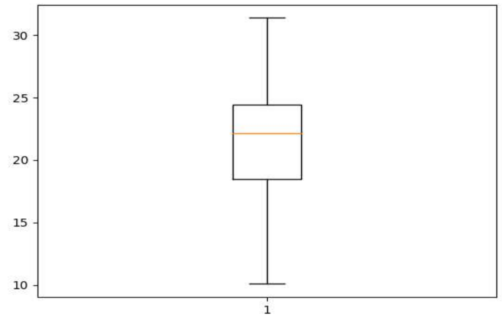

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. Number of ouliers now are reduced to zero which was 10 previously.
2. Inter quartile range now is reduced to 5.9344 from 6.101.
3. Data is spreaded from 10.085 to 31.375.
4. Data is negetively skewed same as before.

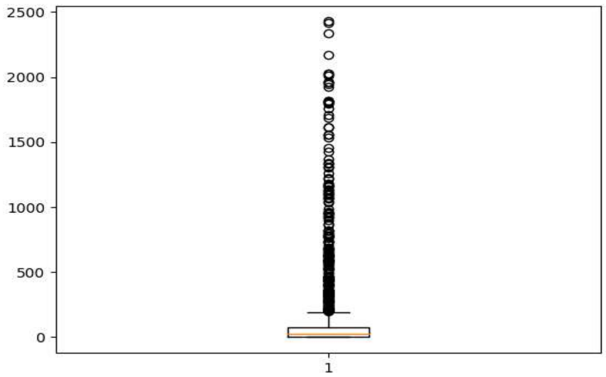

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. Now number of outliers is 233 which was 185 previously.
2. Inter quartile range is reduced to 76.5 from 987.75.
3. Data is spreaded from 0 to 2427.750.
4. Data is positively skewed as before.