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

Student's Name: Priyanka Kumari **Mobile No:** 8328354314

Roll Number: B20307 **Branch:** Electrical Engineering

1

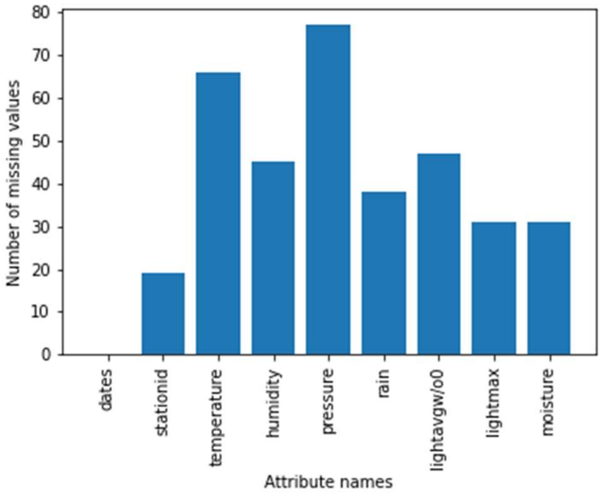

Figure 1 Number of missing values vs. attributes

Inferences:

1. From the above graph we can infer that attribute pressure have maximum number of missing values and attribute dates have minimum number of missing values.
2. From the above bar graph, we can infer that attribute dates have no missing values and attribute lightmax and moisture have same number of missing values also attribute temperature and pressure have maximum number of missing values.

2 a.

Inferences:

1. If the targeted attribute has missing values, it can provide wrong and misleading statistics about the data so to get more accurate and good analysis of the data, we tend to delete the attributes with missing values.
2. Number of tuples deleted equals 19

3. Percentage of total number of tuples deleted is 2.01

b.

Inferences:

1. Number of tuples deleted equals 35
2. 3.65% of the total number of tuples is deleted.
3. Data lost in b part is more than data lost in part a
4. This step is required for cleansing and organizing the data. Since many of the tuple had a lot of missing values thus, we can ignore those tuples without losing the quality of 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	14
3	temperature (in °C)	25
4	humidity (in g.m^{-3})	5
5	pressure (in mb)	39
6	rain (in ml)	4
7	lightavgw/o0 (in lux)	11
8	lightmax (in lux)	0
9	moisture (in %)	4

Inferences:

1. For the above table we can see that attribute pressure have the max value equals to 39 and attribute dates and lightmax have minimum value equal to 0
2. The percentage of missing values for attribute of dates: 0.00%, stationed:1.47%, temp: 2.6%, humidity: 0.52%, pressure: 4.12%, rain: 0.42%, lightavg:1.16%, lightmax:0.00%, moisture: 0.42%

3. Total number of missing attributes in the file is 102, the attribute date has the 0-missing value.

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	22.273	4.356	21.196	12.727	22.169	4.33
4	humidity (in g.m^{-3})	83.480	99.00	91.381	18.210	83.538	99.00	91.381	18.207
5	pressure (in mb)	1009.009	789.393	1014.67	46.980	1009.26	789.39	1014.678	45.99
6	rain (in ml)	1070.009	0.00	18.00	24852.25	10651.63	0.00	22.500	24779.512
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.88	7573.16	4486.34	4488.91	1623.49	7573.79
8	lightmax (in lux)	21788.623	4000.00	6634.00	22065	21517.191	4000.00	6569.00	21935.166
9	moisture (in %)	32.386	0.00	16.704	33.659	32.327	0.00	16.307	33.603

Inferences:

1. max change in attribute for mean value is rain and min is in temperature,
   max change in attribute for median value is lightmax (in lux) and min is in temperature
   max change in attribute for mode value is pressure and min or no change is observed in rest of the attributes.
   Max change in attribute for standard deviation is in lightmax (in lux) and min in moisture (in %).
   Overall max change is observed in attribute lightmax (in lux) and min in attribute moisture.
2. There is no such relationship between the maximum or minimum change in values of mean, median, mode, and standard deviation with the number of missing values. Also, this time, the attributes having average number of missing values as compared to all show the maximum change

3. Due to very less change in the values of mean, median , mode and standard deviation , the reliability of data has not been affected and it's quality has been maintained.

ii.

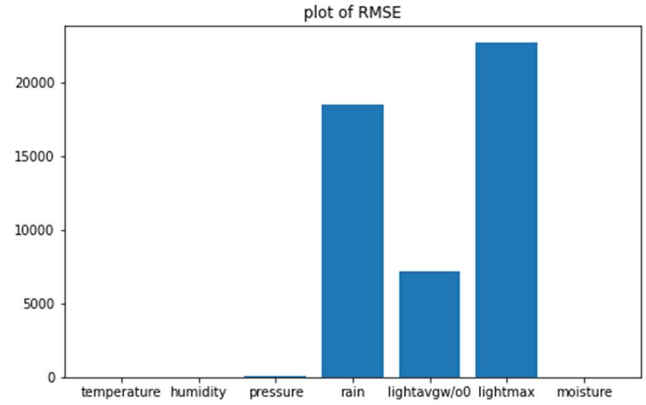

Figure 2 RMSE vs. attributes

Inferences:

1. Max value: 'Lightmax' and Min value: 'temperature'
2. There seems to be high values of RMSE for attributes having maximum change in mean, median , mode and standard deviation whereas RMSE values are low for attributes with less change in magnitudes of mean, median, mode and standard deviation
3. Yes, the data becomes a bit reliable but still the high RMSE values of some of the attributes decreases the quality of our dataset.

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	22.273	4.356	21.196	12.727	22.169	4.33
4	humidity (in $g.m^{-3}$)	83.480	99.00	91.381	18.210	83.538	99.00	91.381	45.99
5	pressure (in mb)	1009.009	789.393	1014.678	46.980	1009.3	789.39	1014.67	24779.512
6	rain (in ml)	10701.538	0.00	18.00	24852.255	10652	0.00	22.500	7573.795
7	lightavgw/o0 (in lux)	4438.428	4488.910	1656.880	7573.163	4486.3	4488.9	1623.5	21935.166
8	lightmax (in lux)	21788.623	4000.000	6634.00	22064.993	21517	4000.0	6569.0	33.603
9	moisture (in %)	32.386	0.00	16.704	33.653	32.327	0.00	16.307	33.603

Inferences:

1. max change in attribute for mean value is lightmax (in lux) and min is in temperature,
   max change in attribute for median value is lightavgw/o0 (in lux) and min is in pressure
   for mode no change is observed,
   Max change in attribute for standard deviation is in lightmax (in lux) and min in humidity.
   Overall max change is observed in attribute lightmax (in lux) and min in attribute humidity.
2. There is no such relationship between the maximum or minimum change in values of mean, median, mode, and standard deviation with the number of missing values. Also, this time, the attributes having average number of missing values as compared to all show the maximum change.
3. Due to very less change in the values of mean, median, mode and standard deviation, the reliability of data has not been affected and its quality has been maintained.

4. I n case of replacing the missing values by mean a change of high magnitude is observed in the values of mean, median, mode and standard deviation whereas in case of replacing the missing values by linear interpolation technique this change observed is of low magnitude and hence the quality of the data is better than the former.

ii.

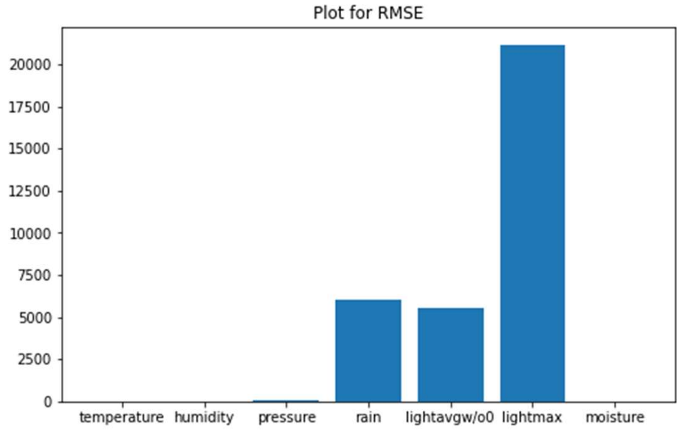

Figure 3 RMSE vs. attributes

Inferences:

1. From the above bar graph, we can infer that Lightmax' have maximum and 'temperature' has minimum RMSE.
2. There seems to be high values of RMSE for attributes having maximum change in mean, median, mode and standard deviation whereas RMSE values are low for attributes with less change in magnitudes of mean, median, mode and standard deviation
3. Yes, the data is reliable for further investigation and experimental analyses since the change observed is less and data is consistent.
4. After replacing values by interpolation technique, the RMSE values comes out to be low whereas in case of replacing the missing values by mean the RMSE values are significantly higher. This makes the data obtained using linear interpolation technique more reliable.

5 a.

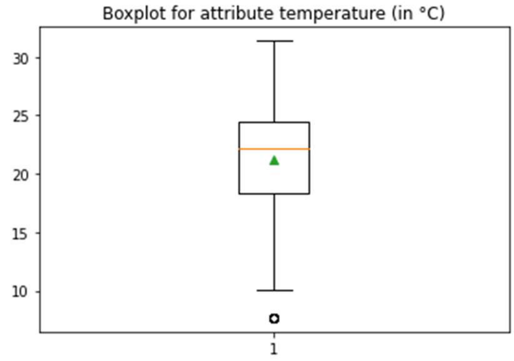

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. Outliers present are of value 7.6729°C and their row numbers ranges from 509°C to 518°C..
2. Inter quartile range is from 18.314°C to 24.416°C with a magnitude of 6.10198° (IQR=Q3-Q1).
3. From the above boxplot the data is negatively skewed.
4. Size of the box is good hence more variability of attribute.

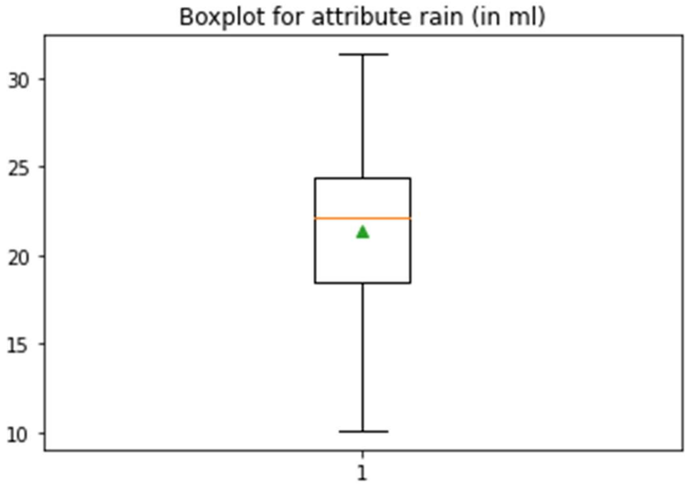

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. There are very large number of outliers in the dataset and all have values greater than 987.75ml.

2. The Inter quartile range is from 0.0ml to 987.75ml and with magnitude of 987.75ml.

3. From the above boxplot the data is negatively skewed.

4. This attribute has very low spread looking at the size of the box but the high number of outliers makes the study of this data difficult.

b.

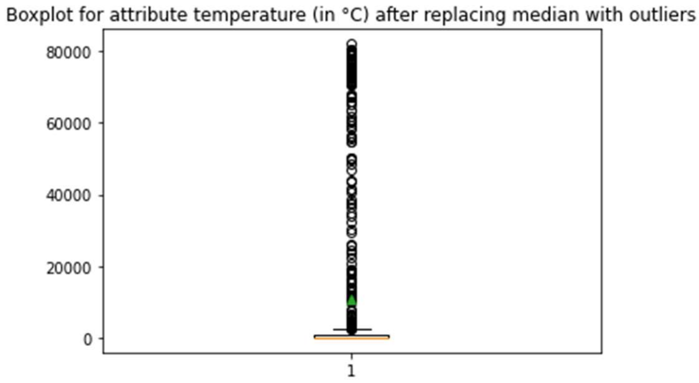

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers in this data since all outliers are replaced with median values.

2. The inter quartile range is from 18.48227°C to 24.41667°C with a magnitude of 5.934°C. It is quite similar to the previous values.

3. The variability of data is high as same with the previous. 4. Data is moderately negatively skewed

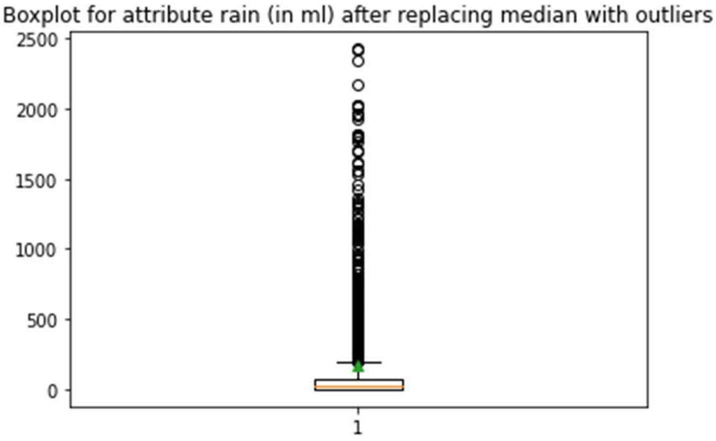

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. It has also very large numbers of outliers as the same with q5. a.

2. Interquartile range is from 0ml to 76.5ml whereas in previous case it was from 0 to 987.75ml.

3. This attribute has very low spread looking at the size of the box but the high number of outliers makes the study of this data difficult. In this case magnitude of outliers has decreased