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

Student's Name: Madhur Jajoo **Mobile No: 7597389137**

Roll Number: B20211 **Branch:** DSE

1

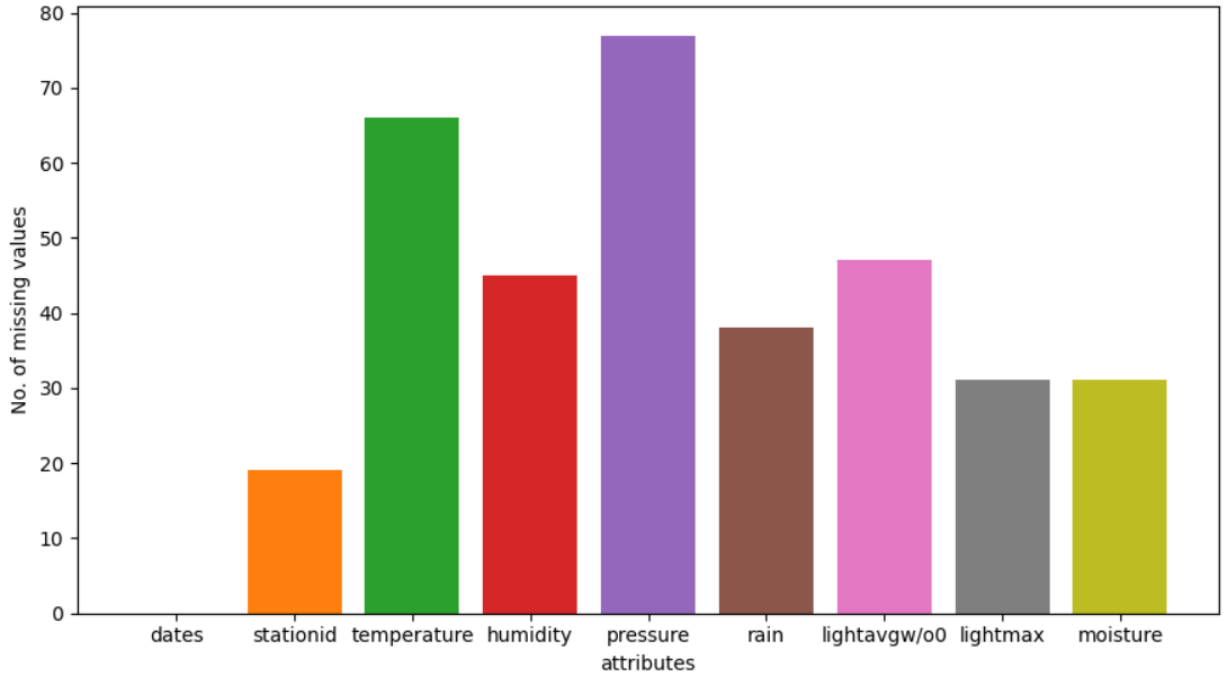

Figure 1 Number of missing values vs. attributes

Inferences:

1. Pressure and dates have maximum and minimum values respectively.
2. Pressure, humidity, lightavgw/o0 and temperature are missing too much values while rain lightmax moisture and stationid are missing some less values and dates is missing no values.

2 a.

Inferences:

1. we deleted the tuples which were not having the station id because without it we cannot predict **where** landslide is going happen.
2. 19 tuples deleted after this step.
3. 2.01 percent of the total number of tuples were deleted.

b.

Inferences:

1. 35 tuples deleted after this step.
2. 3.77 percent of the total number of tuples were deleted.
3. We lost 5.7% of data in total.
4. This step is needed because when we don't have much data we can't predict future events or it will make our future predictions wrong.

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 and lightmax have maximum and minimum missing values as dates are not missing at all and we deleted the tuples which were not having the value for stationed.

2. Temperature is missing 3.82% data; Humidity is missing 1.46% data; Pressure is missing 4.6%data; rain is missing 0.67% data; lightavgw/o0 is missing 1.68% data; lightmax is missing 0.011% data; moisture is missing 0.67% data.
3. 116 values are still 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	Median	Mode	S.D.	Mean	Median	Mode	S.D.
1	dates								
2	stationid								
3	temperature (in °C)	21.052	21.927	21.05244	4.34	21.215	22.273	12.727	4.356
4	humidity (in $g.m^{-3}$)	83.126	91.0	99.0	18.394	83.48	91.380	99.0	18.210
5	pressure (in mb)	1009.465	1014.482	1009.465	45.856	1009.008	1014.677	789.392	46.980
6	rain (in ml)	10798.37	15.75	0.0	24833.96	10701.54	18.0	0.0	24852.25
7	lightavgw/o0 (in lux)	4458.297	1502.939	4488.910	7606.284	4438.428	1656.88	4488.9103	7573.163
8	lightmax (in lux)	21463.221	6569.0	4000.0	21943.889	21788.623	6634	4000	22064.993
9	moisture (in %)	32.603	14.17	0.0	33.714	32.386	16.702	0.0	33.653

Inferences:

1. ligtmax and humidity have maximum and minimum change in mean respectively; ligtmax and pressure have maximum and minimum change in median respectively; pressure and humidity, lightmax, moisture, lightavgw/o0, rain have maximum and minimum change in mode respectively. ligtmax have maximum and minimum change in standard deviation respectively.
2. The one having maximum change in mean, mode is also having maximum 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.

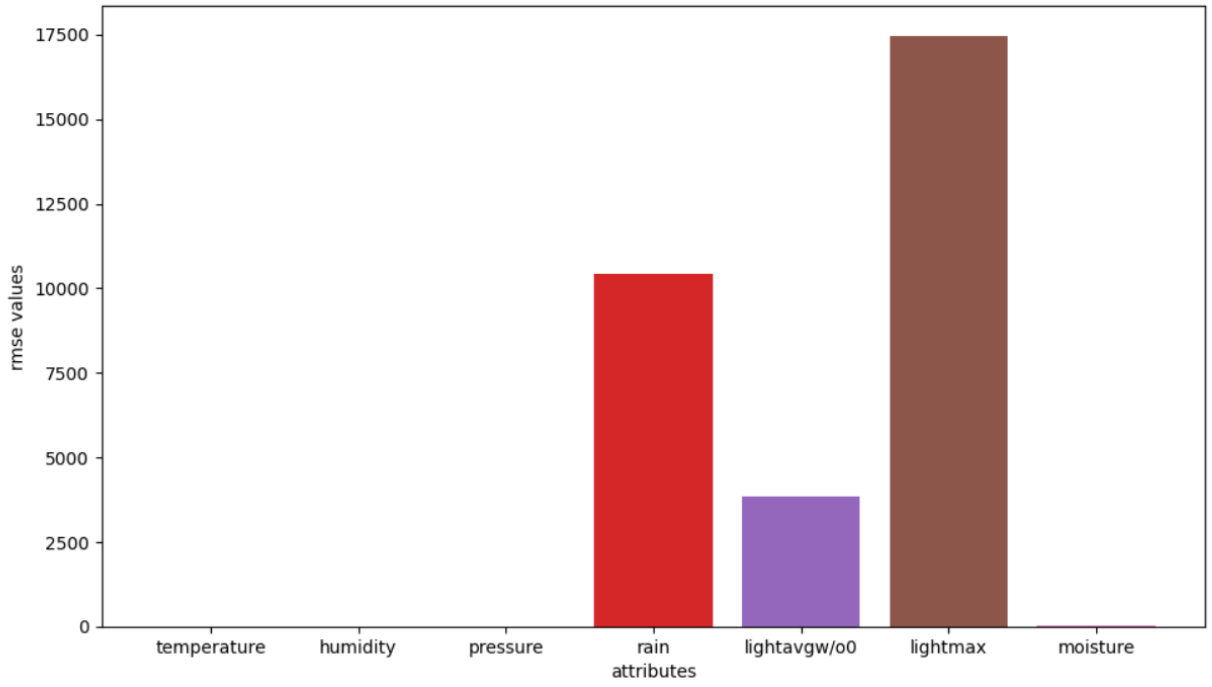

Figure 2 RMSE vs. attributes

Inferences:

1. Lightmax and pressure have maximum and minimum value of rmse respectively.
2. Lightmax the attribute having maximum change in mean and mode is the attribute having maximum value of RMSE
3. The data is not reliable as RMSE is very high for some pf the attributes.

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)	21.114	12.727	22.14	4.39	21.214	12.727	22.273	4.355
4	humidity (in $g.m^{-3}$)	83.165	99.0	91.179	18.408	83.48	99.0	91.380	18.210
5	pressure (in mb)	1009.968	789.329	1014.925	45.999	1009.008	789.392	1014.677	46.980
6	rain (in ml)	10727.96	0.0	15.75	24848.71	10701.54	0.0	18.0	24852.25
7	lightavgw/o0 (in lux)	4496.753	4488.91	1500.5	7649.45	4438.428	4488.9103	1656.88	7573.163
8	lightmax (in lux)	21473.799	4000	6569	21946.16	21788.623	4000	6634	22064.993
9	moisture (in %)	32.528	0.0	13.894	33.791	32.386	0.0	16.702	33.653

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. The one having maximum change in mean, mode is also having maximum change in standard deviation .
3. Yes this data is reliable as it does not have much changes.
4. Replacing missing values with mean makes huge changes in mode and median which makes the data more inreliable for analysis, while in interpolation it doesn't happen

ii.

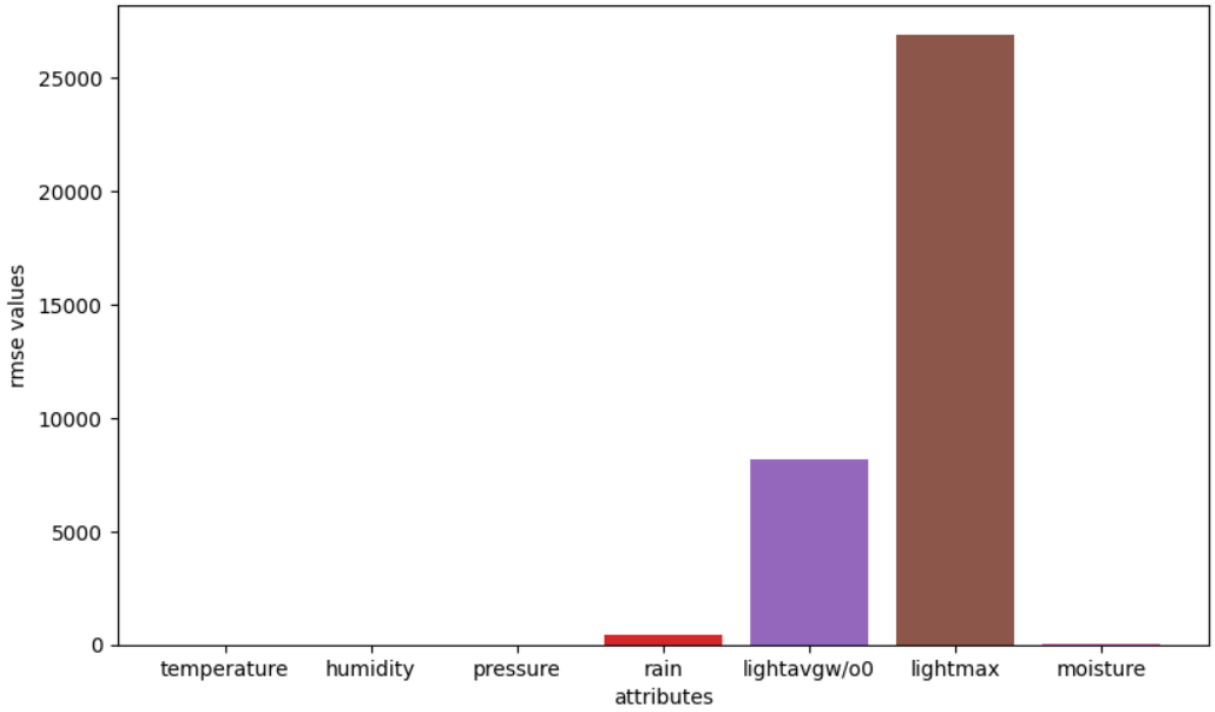

Figure 3 RMSE vs. attributes

Inferences:

1. Lightmax and temperature have maximum and minimum value of rmse respectively.
2. Lightmax the attribute having maximum change in mean and mode is the attribute having maximum value of RMSE.
3. The data is reliable as only lightwave is having high value of rmse.
4. From the calculated RMSE compare and contrast replacing missing values by mean and linear interpolation technique.

Had no time for the rest of the work:c

5 a.

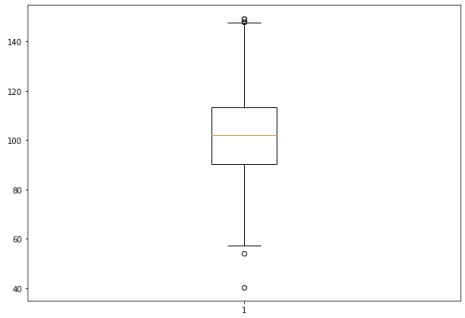

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. List the number of outliers and their row numbers.
2. Infer the Inter quartile range.
3. Infer the 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.

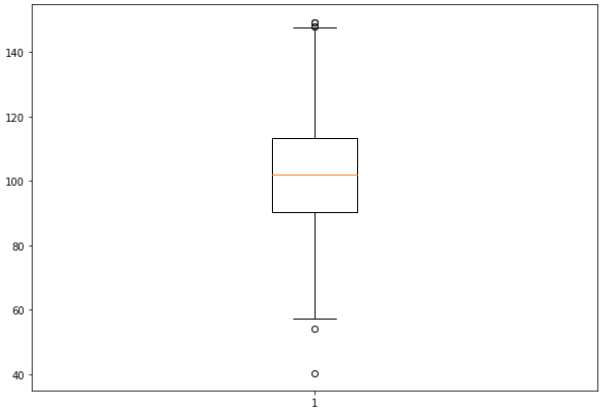

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. List the number of outliers and their row numbers.
2. Infer the Inter quartile range.
3. Infer the 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.

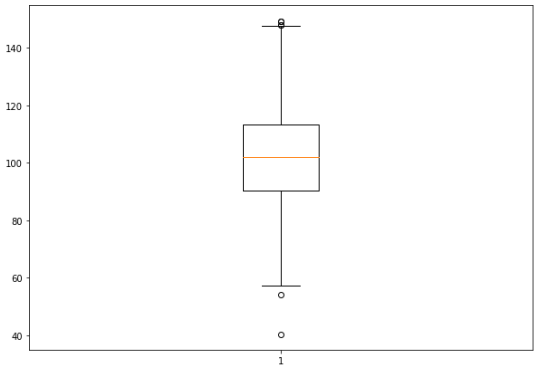

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

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

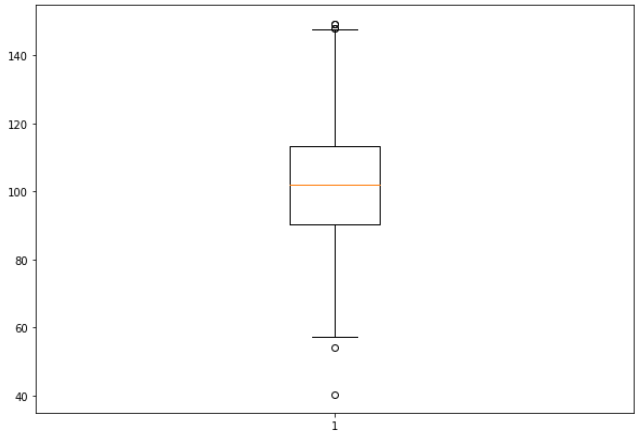

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

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

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and title.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places**
- **The quantities which have units should be written with units.**