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

Student's Name: Mayank Bansal **Mobile No: +919636993445**

Roll Number: B20156 **Branch:CSE**

1

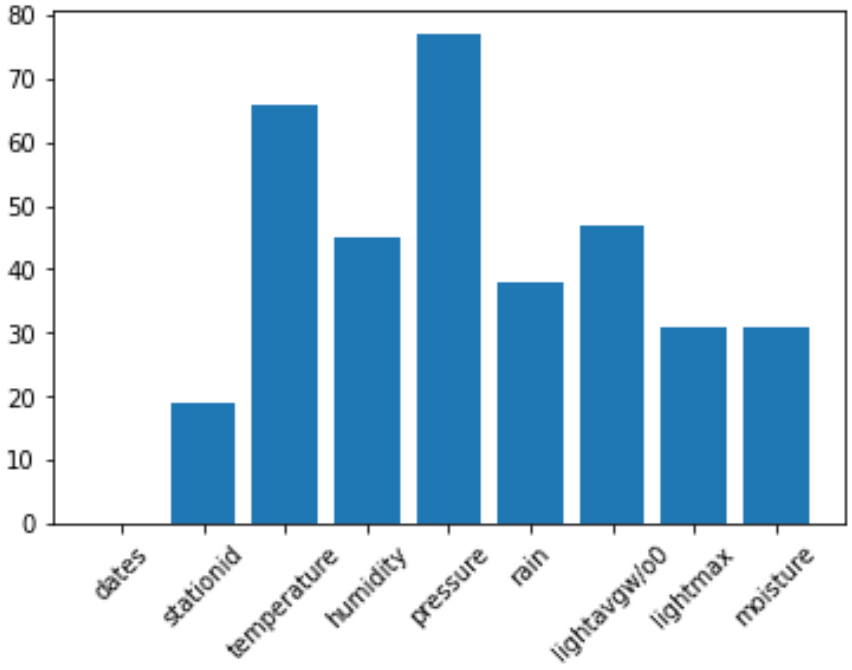

Figure 1 Number of missing values vs. attributes

Inferences:

1. Dates have no missing value whereas Pressure and temperature have a high number of missing values.
2. Pressure and temperature have more missing which indicates the data collection system might have a problem.IN case of other values the missing values are less which might be the case of temporary data disruptions.

2 a.

Inferences:

1. Because stationed and Dates together give us the index for data which provides a uniqueness to the tuple. Also stationid can never be assumed or calculated.
2. 19 tuples are deleted in this step.
3. 19 tuples refer to a total of 199.16 percent of data.

b.

Inferences:

1. Total of 30 tuples are deleted.
2. 3.17% of the total number of tuples are deleted.
3. Yes, we lose data but we can't afford to get wrong information
4. Data which has 3 or more attributes missing can harm the useful data we try to obtain. Which can further harm our project based on data we use.

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(37) have maximum missing values whereas stationid(0) , date(0) and lightmax(2) have minimum number of missing values
2. Temperature=>3.9% Humidity=>1.6% Pressure=>4.7% rain=>.7% moisture=>.7% lightavgw=>1.7% lightmax=>.2%
3. Total 131 number of attributes are missing.
4. Inference 4(You may add or delete the number of inferences)

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	temperature (in °C)	21.05	21.05	21.922	4.328	21.21	12.727	22.27	4.355
2	humidity (in $g.m^{-3}$)	83.141	99.0	90.85	18.3485	83.48	99.0	91.38	18.21
3	pressure (in mb)	1009.5	1009.47	1014.4	45.727	1009	789.4	1014.6	47
4	rain (in ml)	10860.5	0	16.875	24878.7	10701.53	0	18	24852.25
5	lightavgw/o0 (in lux)	4451.45	4488.9103	1516	7588.04	4438.42	4488.9103	1656.9	7573.16
6	lightmax (in lux)	21498.3	4000	6569	21954	21788.6	4000	6634	22064
7	moisture (in %)	32.58	0	14.25	33.73	32.38	0	16.7	33.65

Inferences:

1. Attributes like lightmax and rain have a severe change in mean as compared to others. Whereas attributes like temperature, humidity , pressure and moisture have a verry less difference in their means.
2. SD and median are approximately same for all the attributes
3. Mode is exactly same for all attributes except pressure which also has a huge difference.
4. Mode for pressure has very fast difference because it has the most number of missing values.
5. Data is reliable for getting the idea of center and the spread of data but not reliable for frequency.

ii.

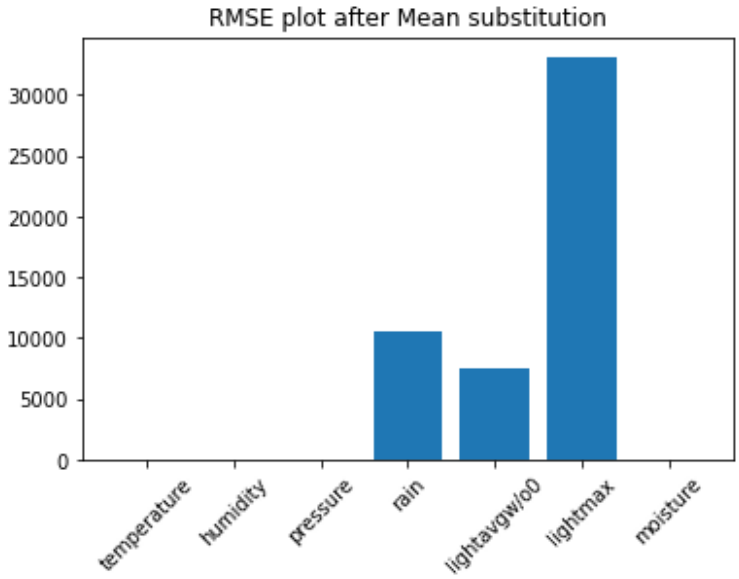

Figure 2 RMSE vs. attributes

Inferences:

1. RMSE value of lightmax is very high. Attributes like moisture, pressure, humidity, and temperature are very low as compared to others.
2. Lightmax and rain have very high RMSE value which can be related to their differences in their respective means.
3. It is unreliable for selected attributes not for the others.

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	temperature (in °C)	21.11	12.727	21.15	4.39	21.214	12.727	22.27	4.35
2	humidity (in $g.m^{-3}$)	83.15	99	91	18.37	83.48	99	91.38	18.21
3	pressure (in mb)	1009.94	789.39	1014.93	45.915	1009	789.39	1014.67	46.98
4	rain (in ml)	10777	0	15.75	24896.12	10701.5	0	18	24852.25
5	lightavgw/o0	4492.28	4488.91	1501.7	7631.52	4438.42	4488.91	1656.8	7573.162

	(in lux)								
6	lightmax (in lux)	21497.2	4000	6569	21959	21788.62	4000	6634	22064
7	moisture (in %)	32.5	0	13.9	33.8	32.38	0	16.7	33.65

Inferences:

1. Lightmax has very huge difference in their means as compared to others. Modes are exactly same for all the others. Median for lightavgw has high difference compared to the others.
2. Pressure had very high missing values. But it can't be seen in the data of the interpolation.
3. It is reliable for data like Mode, Mean and SD but not for data like median.
4. In case of interpolation we get a better idea of frequencys.

ii.

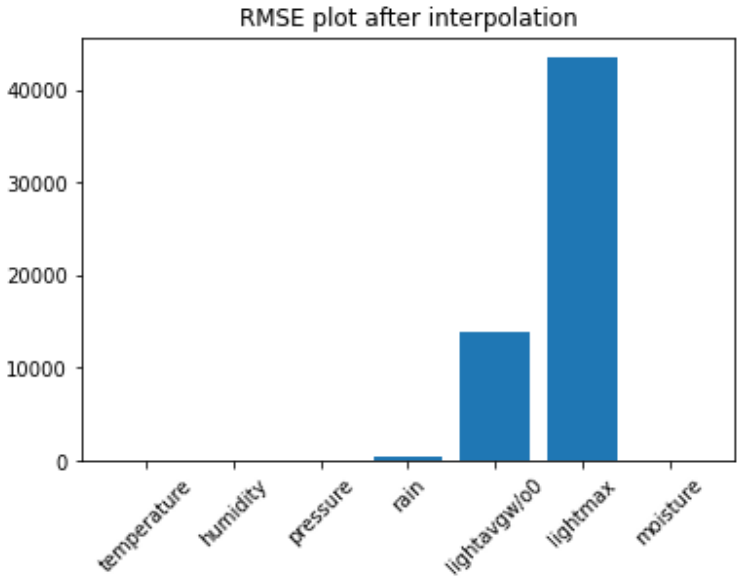

Figure 3 RMSE vs. attributes

Inferences:

1. Lightmax and lightavgw has very high RMSE value whereas others have very low RMSE values
2. Lightmax has very high error because it also has high number of missing values. and a huge difference in their means.

3. It can be reliable upto a certain extent. Moisture, pressure, humidity and temperature give us a better assumption.

5 a.

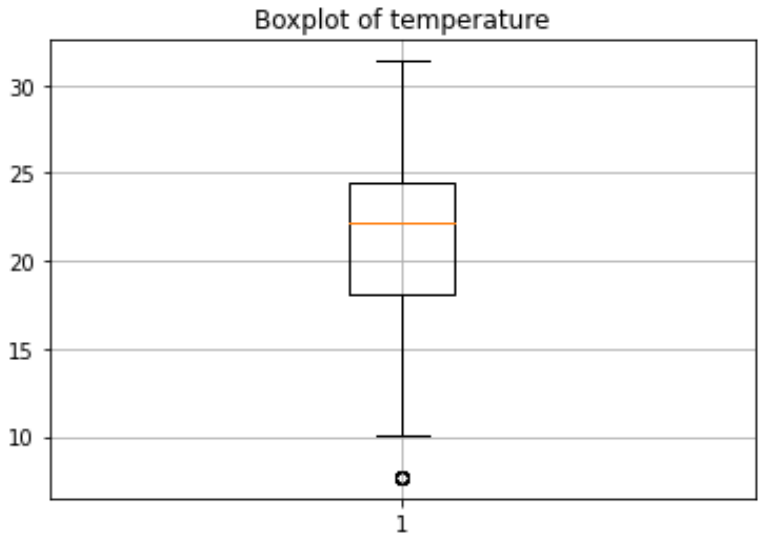

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers.
2. Outliers [7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729]
3. Inter Quartile Range= 6.371
4. We can see that only 10 values are outliers. The data is very slightly spread
5. The Data is negatively skewed.

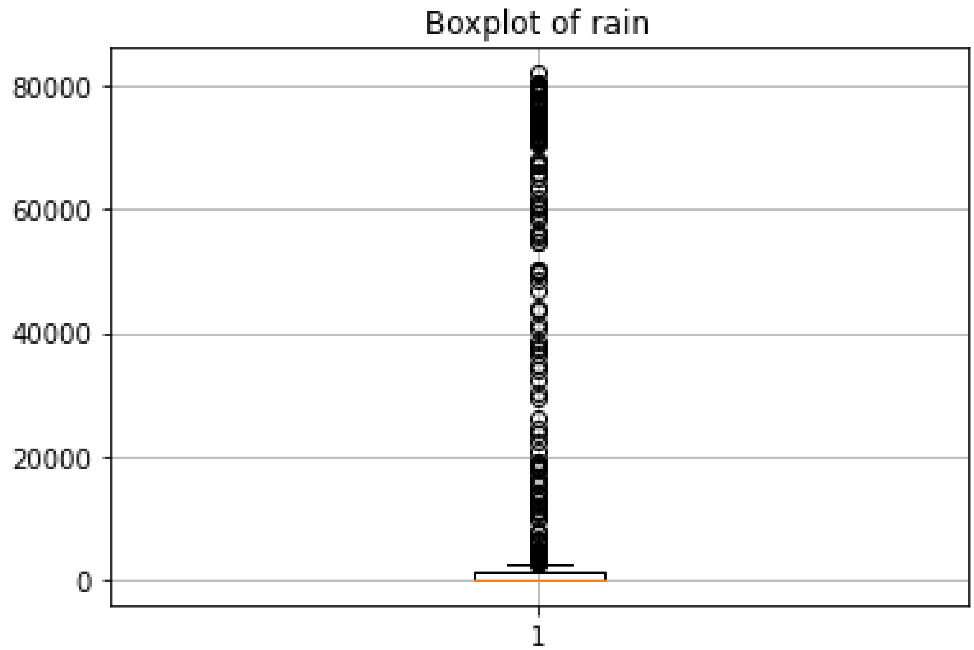

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. Number of outliers in rain 177
2. Outliers=[13583.25, 15459.75, 14001.75, 16571.25, 13666.5, 59982.75, 80000.0, 75048.75, 80000.0, 3930.5, 36636.75, 40789.0, 63256.5, 54616.5, 50172.75, 37928.25, 26178.75, 3138.75, 3449.25, 18884.25, 9765.0, 18976.5, 30393.0, 2814.75, 80000.0, 82037.25, 56319.75, 71968.5, 80000.0, 80000.0, 50242.5, 80000.0, 60675.75, 22250.25, 2637.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 37392.75, 49725.0, 80000.0, 80000.0, 71154.0, 80000.0, 80000.0, 12854.25, 34879.5, 4610.25, 6210.0, 10557.0, 3451.5, 3312.0, 18285.75, 3613.5, 2893.5, 23474.25, 14042.25, 3647.25, 5877.0, 10062.0, 17997.75, 29517.75, 32514.75, 13943.25, 4212.0, 4691.25, 7519.5, 11112.75, 2821.5, 33941.25, 43643.25, 20664.0, 11144.25, 18587.25, 18373.5, 15646.5, 12915.0, 49916.25, 24522.75, 75105.0, 73417.5, 70580.25, 78126.75, 56097.0, 6061.5, 38355.75, 55509.75, 43974.0, 6747.75, 54843.75, 59377.5, 58320.0, 60963.75, 63342.0, 67378.5, 70929.0, 73158.75, 71367.75, 73838.25,

46732.5, 48429.0, 67830.75, 75447.0, 74646.0, 75402.0, 75723.75, 74254.5, 75201.75, 77044.5, 74472.75, 77503.5, 78180.75, 79915.5, 80583.75, 80482.5, 79337.25, 79317.0, 70823.25, 75638.25, 73752.75, 65893.5, 72774.0, 7773.75, 12037.5, 79839.0, 78633.0, 78779.25, 76662.0, 67252.5, 74913.75, 4869.0, 41618.25, 58443.75, 74173.5, 72445.5, 65873.25, 67675.5, 61989.75, 71237.25, 73577.25, 65301.75, 73534.5, 72283.5, 71799.75]

3. Inter Quartile Range= 1048.5
4. Seeing the outliers above the Inter Quartile range of the data one can easily say that data is largely spread.
5. Many of outliers are above the IQR, Hence it is positively skewed.

b.

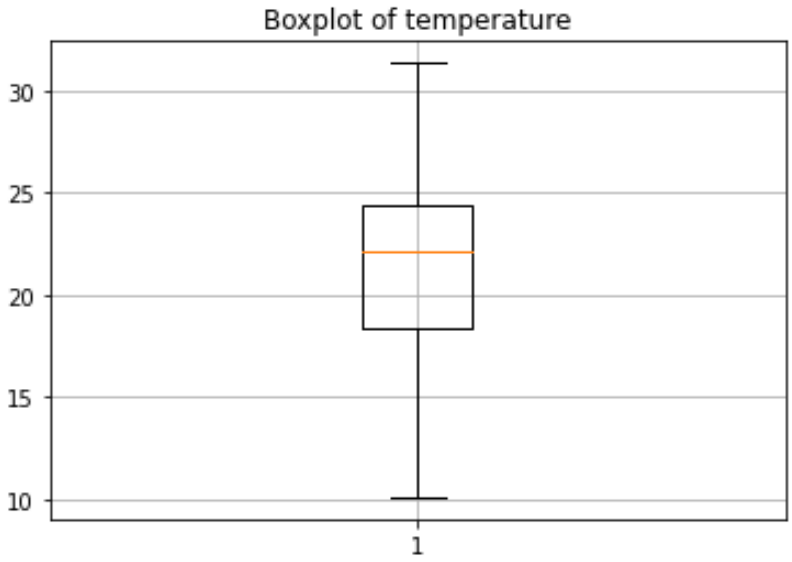

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers in this graph as compared 10 in previous case.
2. Inter Quartile Range= 6.080
3. IQR is same here with less range hence spread is less.
4. The Boxplot is negatively skewed.

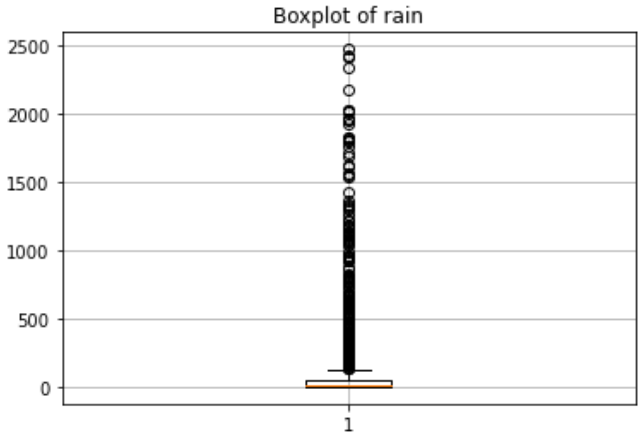

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

s**Inferences:**

1. There are total of 182 outliers in Boxplot
2. Outliers=[1761.75, 652.5, 963.0, 254.25, 339.75, 607.5, 560.25, 513.0, 474.75, 817.875, 1161.0, 240.75, 398.25, 816.75, 776.25, 681.75, 441.0, 274.5, 1341.0, 1804.5, 2171.25, 742.5, 443.25, 774.0, 1167.75, 630.0, 594.0, 546.75, 634.5, 1091.25, 162.0, 157.5, 366.75, 589.5, 207.0, 281.25, 1215.0, 315.0, 1260.0, 324.0, 360.0, 679.5, 130.5, 159.75, 1710.0, 1183.5, 1962.0, 1071.0, 438.75, 864.0, 816.75, 796.5, 191.25, 202.5, 1611.0, 353.25, 533.25, 213.75, 434.25, 191.25, 202.5, 594.0, 409.5, 139.5, 333.0, 468.0, 222.75, 263.25, 459.0, 158.0, 272.25, 621.0, 587.25, 468.0, 778.5, 987.75, 623.25, 330.75, 1075.5, 308.25, 337.5, 1617.75, 144.0, 402.75, 2414.25, 1044.0, 211.5, 285.75, 400.5, 1426.5, 209.25, 551.25, 344.25, 1140.75, 357.75, 308.25, 774.0, 207.0, 1172.25, 427.5, 531.0, 1311.75, 247.5, 454.5, 283.5, 1062.0, 1554.75, 569.25, 357.75, 1795.5, 382.5, 353.25, 918.0, 677.25, 1689.75, 141.75, 213.75, 637.5, 2470.5, 580.5, 951.75, 281.25, 684.0, 463.5, 420.75, 1329.75, 173.25, 211.5, 173.25, 1300.5, 326.25, 621.0, 1818.0, 783.0, 949.5, 438.75, 1559.25, 1039.5, 405.0, 582.75, 234.0, 666.0, 625.5, 1365.75, 1129.5, 524.25, 492.75, 920.25, 218.25, 2022.75, 2009.25, 438.75, 285.75, 225.0, 1809.0, 1226.25, 1964.25, 321.75, 688.5, 765.0, 1125.0, 868.5, 1107.0, 405.0, 731.25, 157.5, 794.25, 1536.75, 954.0, 731.25, 1926.0, 1818.0, 243.0, 373.5, 308.25, 936.0, 2029.5, 661.5, 1946.25, 1095.75, 2340.0, 2427.75]
3. Inter Quartile Range= 51.75
4. Since initially range was of order about 80000 which now becomes 2500, hence spread of data of is less.
5. Data is positively skewed. But still has a large spread.
6. New Dataset can new outliers because earlier dataset was more concentrated around the median and exchanging outliers with median can make other points to be outliers.