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

Student's Name: Abhay Gupta **Mobile No:** 9511334630

Roll Number: B20075 **Branch:**CSE

1

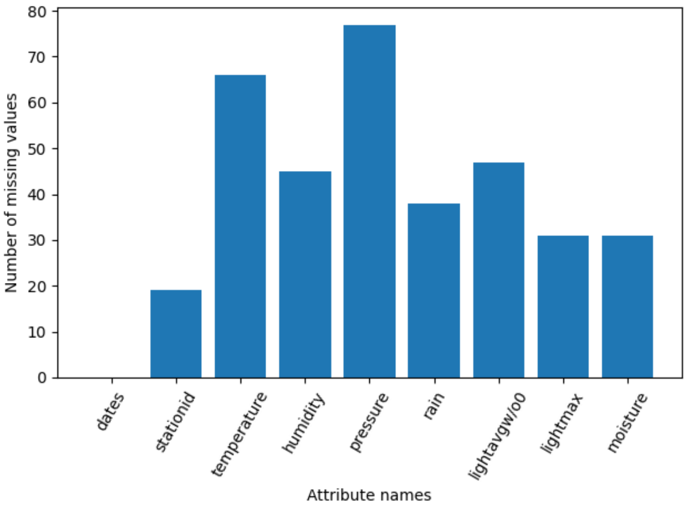

Figure 1 Number of missing values vs. attributes

Inferences:

1. Attribute 'Pressure' has maximum and 'dates' has minimum missing values.
2. Attribute dates has 0 missing value, stationid has 19 missing values, temperature has 66, humidity has 45, pressure has 77, rain has 38, lightavgw/o0 has 47, light max has31 and moisture has 31 missing values.

2 a.

Inferences:

1. We choose to delete the tuple if the target attribute is missing because without the target attribute, rest of the values of that tuple are useless. As we don't know to which class that data belongs to.
2. Total number of tuples deleted after this step are 19
3. Percentage of total number of tuples deleted = (19/945) * 100 = 2.01%

b.

Inferences:

1. Total number of tuples deleted after this step are 30
2. Percentage of total number of tuples deleted = (30/926) * 100 = 3.23%
3. There is some data loss. But effectively we reach more closer to the original distribution.
4. This step was needed to prevent the non-uniformity of data. As, for some tuples, some values were known and some were not, which could lead to the wrong result.

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 maximum and dates has minimum missing vakues.
2. Percentage of data missing for:

dates	0.000000 %
stationid	0.000000%
temperature	4.129464%
humidity	1.785714%
pressure	5.022321%
rain	0.781250%

lightavgw/o0 1.897321%
lightmax 0.223214%
moisture 0.781250%

3. The total number of missing attributes are 131.

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		
0.2	stationid		T9				T9		
3	temperature (in °C)	21.214	12.7273	22.272	4.355	21.050	21.0508	21.922	4.328
4	humidity (in $g.m^{-3}$)	83.479	99	91.380	18.210	83.141	99	90.859	18.348
5	pressure (in mb)	1009.008	789.393	1014.677	46.980	1009.470	1009.47	1014.433	45.727
6	rain (in ml)	10701.538	0	18	24852.255	10860.547	0	16.875	24878.702
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.880	7573.162	4451.454	4488.91	1516.011	7588.040
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21498.312	4000	6569	21954.040
9	moisture (in %)	32.386	0	16.704	33.653	32.583	0	14.252	33.734

Inferences:

1. Mean: Maximum change is in lightmax and minimum change is in temperature.
   Median: Maximum change is in lightavgw/o0 and minimum change is in temperature
   Mode: Maximum change is in pressure and minimum change is in date, stationed, humidity, rain, lightmax, moisture.
   Standard deviation: Maximum change is in lightmax and minimum change is in temperature

2. We can observe that mode & mean of most of the attributes are approximately same. However same cannot be said about other attributes because of large differences in values. Attributes having largest number of missing values has largest change in mode.
3. As the change in mean, median, mode, standard deviation is small, so the data is reliable for further analysis.

ii.

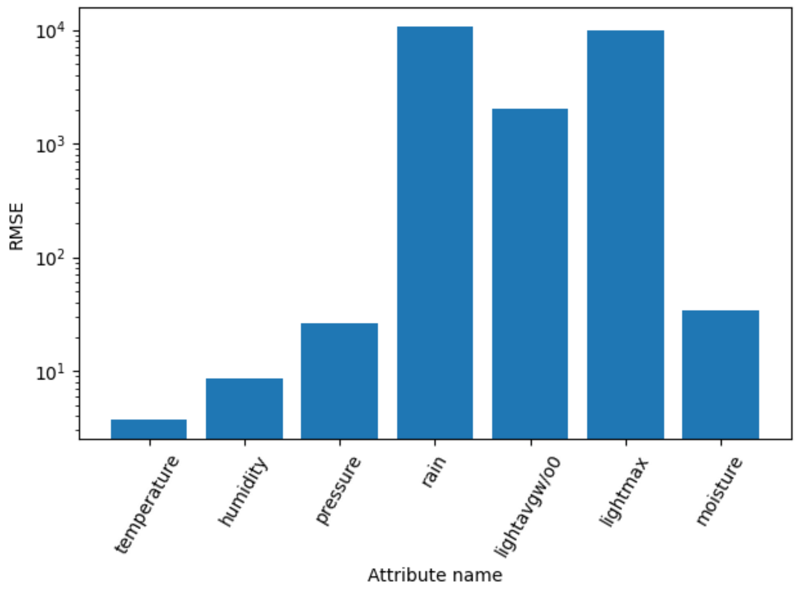

Figure 2 RMSE vs. attributes

Inferences:

1. Rain has maximum and temperature has minimum RMSE value.
2. There is a relation between the RMSE value and the no. of missing values as the attributes having more missing values have higher values of RMSE. Also, the attribute with higher number of missing values have significant value of RMSE.
3. The data is reliable for further investigation except the rain and lightmax 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		19-07-2018				19-07-2018		
2	stationid		T9				T9		
3	temperature (in °C)	21.214	12.7273	22.272	4.355	21.116	12.727	22.157	4.390
4	humidity (in g.m^{-3})	83.479	99	91.380	18.210	83.156	99	91.060	18.372
5	pressure (in mb)	1009.008	789.393	1014.677	46.980	1009.942	789.393	1014.936	45.915
6	rain (in ml)	10701.538	0	18	24852.255	10777.983	0	15.750	24896.128
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.880	7573.162	4492.283	4488.91	1501.719	7631.524
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21497.189	4000	6569	21959.033
9	moisture (in %)	32.386	0	16.704	33.653	32.498	0	13.910	33.812

Inferences:

1. Mean: Maximum change is in lightmax and minimum change is in temperature.
   Median: Maximum change is in lightavgw/o0 and minimum change is in temperature
   Mode: There is no change in mode.
   Standard deviation: Maximum change is in lightmax and minimum change is in temperature
2. We can observe that mode & mean of most of the attributes are approximately same. However same cannot be said about other attributes because of large differences in values.
3. As, the changes are small so the data is reliable for further analysis, except for few attributes.
4. The change in mean, median, mode is less in the case of replacing by interpolation as compared to that of mean.

ii.

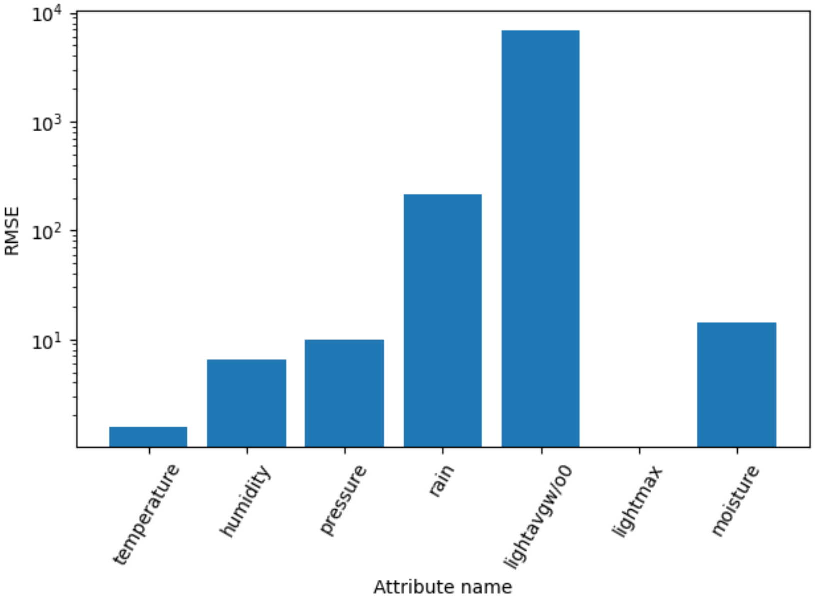

Figure 3 RMSE vs. attributes

Inferences:

1. Lightavgw/o0 has maximum and lightmax has minimum RMSE value.
2. There is a relation between the RMSE value and the no. of missing values as the attributes having more missing values have higher values of RMSE. Also, the attribute with higher number of missing values have significant value of RMSE.
3. The data is reliable for further investigation except lightavgw/o0.
4. RMSE is less in case of replacing missing values by linear interpolation as compared to that of replacing by mean. So, here interpolation is better than mean.

5 a.

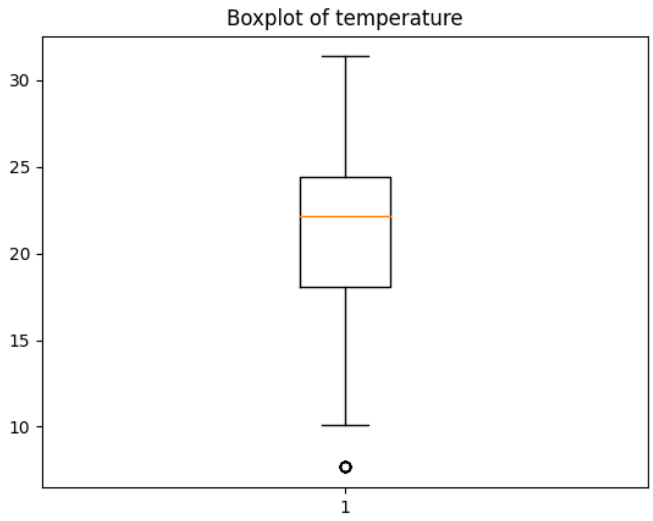

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers.
2. IQR = 6.37.
3. Variance = 19.27
4. Data is negatively skewed

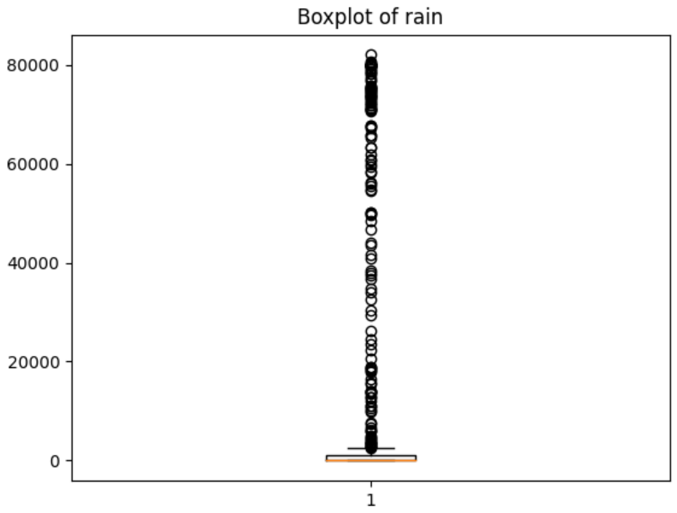

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. The number of outliers are 177
2. IQR = 1048.5
3. Variance = 619817205.84
4. Data is positively skewed

b.

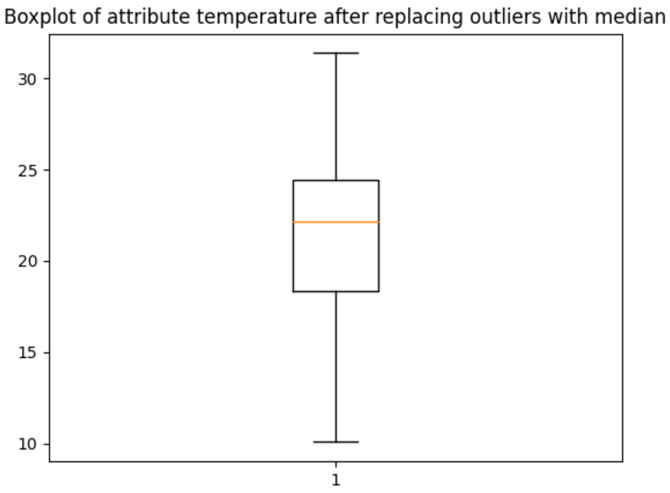

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers. But in Q5(a) there was some outliers present
2. IQR = 6.080
3. Variance = 17.245. This is less than the variance of Q5(a)
4. Data is negatively skewed similar to that of Q5(a)

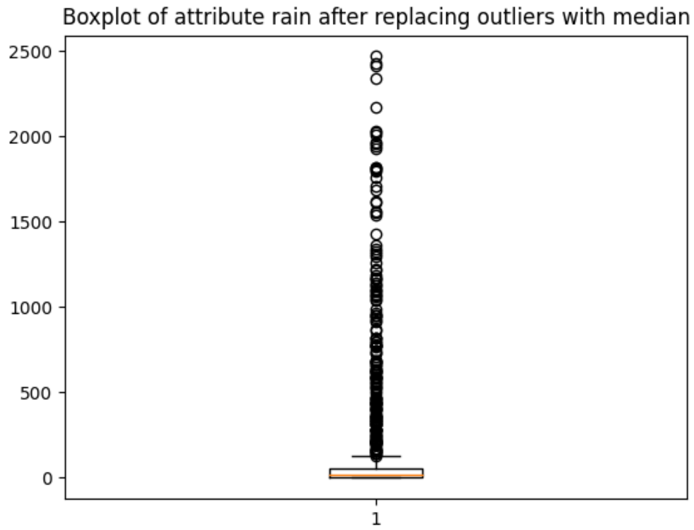

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. The number of outliers are 182. They are more than that of Q5(a)
2. IQR = 51.75 . This is much smaller than IQR of Q5(a)
3. Variance = 156322.013. This is much smaller than that of Q5(a)
4. Data is positively skewed similar to that of Q5(a)