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

Student's Name: Yash Sharma **Mobile No: 8802131138**

Roll Number: B20241 **Branch: CSE**

1

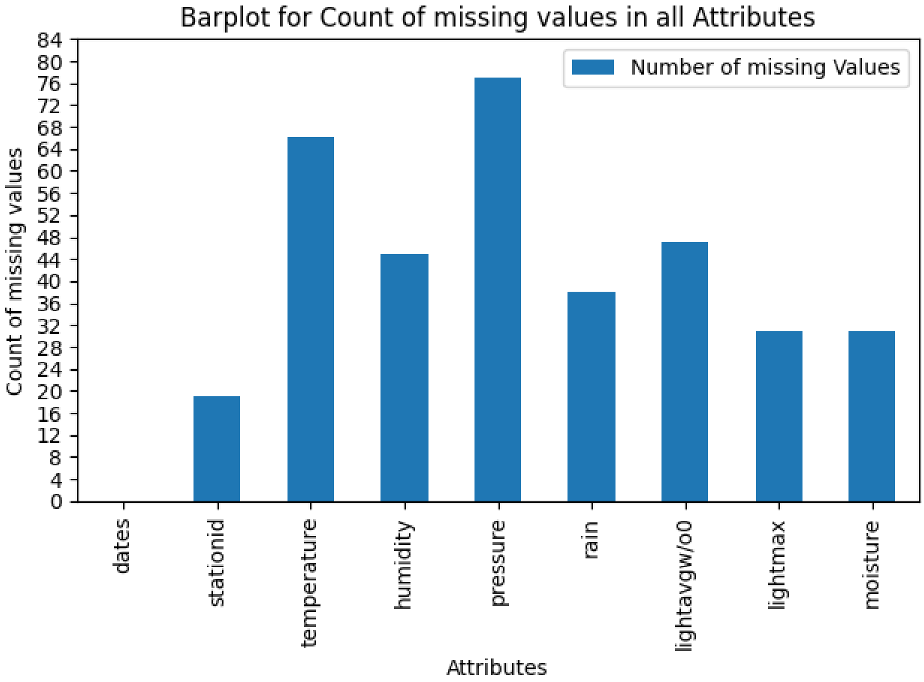

Figure 1 Number of missing values vs. attributes

Inferences:

1. Attribute "dates" has minimum or no missing value while attribute "pressure" has maximum missing values.
2. The bar "heights" depicts the frequency of missing data for each attribute. On average except attribute "dates", 39 values are missing from the attribute column.
3. We may observe that the target column "stationid" also have many missing values.
4. We see there are a lot of attributes with a good amount of missing values. Thus, if any of the row has majority of the attributes or values missing than it may not result in a good ML model. Thus, before imputation of missing values we will be required to drop certain rows which have majority (above some threshold) of values missing.

2 a.

Inferences:

1. We choose to delete the tuple if the target attribute is missing because during fitting our model (training set) for making predictions on test set we must not have any missing value in target attribute. We need all target values in order to train our model. It's a must condition while training a ML model.
2. A total 19 tuples (or vectors) were deleted during this procedure.
3. Approximately 2.01 % of the total number of tuples initially were deleted in this step.

b.

Inferences:

1. A total 35 tuples (or vectors) were deleted during this procedure.
2. Approximately 3.78 % of the total number of tuples initially were deleted in this step.
3. We see that around 4 % of the data still had many tuples with many attributes (at least 3 values) missing.
4. The predictions can not be made by our model if the input data is missing. Moreover, if made it will result in making inaccurate and unrealistic predictions. In the case of multivariate analysis, if there is a larger number of missing values, then it can be better to drop those cases (rather than do imputation) and replace them.
5. About 4 % of Data have been deleted in this step or is lost but that was necessary step as the vectors or rows had at least 3 attributes missing so it will make our predicted model make unintended predictions or if they weren't removed instead were imputed than our resultant model would most like make unrealistic prediction which may not be useful at all in real scenarios.
6. A model trained with the removal of all missing values creates a robust model. The same thing is done 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. Attribute “dates” and “stationid” has minimum or no missing values while attribute “pressure” has maximum missing values.
2. The percentage of data missing for each attribute is as follow:

S. No	**Attribute**	**Number of missing values**
1	dates	0
2	stationid	0
3	temperature (in °C)	3.816
4	humidity (in g.m^{-3})	1.459
5	pressure (in mb)	4.601
6	rain (in ml)	0.673
7	lightavgw/o0 (in lux)	1.683
8	lightmax (in lux)	0.112
9	moisture (in %)	0.673

3. A total of 116 values from all the attributes are missing in the data file.

4 a. i.

Table 2 Mean, mode, median and standard deviation before (original) and after (filling mean) replacing missing values by mean

S. No	**Attribute**	**Before**				**After**			
		Mean	**Mode**	**Median**	**S.D.**	**Mean**	**Mode**	**Median**	**S.D.**
1	dates	Primary Key							
2	stationid	Primary Key							
3	temperature (in °C)	21.098	12.727	22.14	4.396	21.052	21.052	21.927	4.337
4	humidity (in g.m^{-3})	83.123	99.0	91.18	18.403	83.126	99.0	91.0	18.384
5	pressure (in mb)	1010.052	789.393	1014.925	46.049	1009.466	1009.466	1014.482	45.83
6	rain (in ml)	10727.43	0.0	15.75	24834.988	10798.379	0.0	15.75	24820.025
7	lightavgw/o0 (in lux)	4442.747	4488.910	1464.627	7606.675	4458.298	44889103	1502.938	7602.014
8	lightmax (in lux)	21473.799	4000	6569.0	21933.842	21463.221	4000.0	6569.0	21931.572
9	moisture (in %)	32.572	0.0	13.894	33.832	32.603	0.0	14.17	33.695

Inferences:

1. For mean, attribute “rain” has maximum change while attribute “humidity” has minimum change (magnitude wise).
2. For mode, attribute “temperature” has maximum change while rest all attributes have minimum change or no change (are equal magnitude wise).
3. For median, attribute “lightavgw/o0” has maximum change while attribute ““rain” and “lightmax”” have minimum change (magnitude wise).
4. For standard deviation, attribute “rain” has maximum change while attribute “humidity” has minimum change (magnitude wise).
5. As we may observe that there is no consistent relation visible between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values. But it may be observed that attributes that had a considerable amount of values missing from their dataset before are seen to have their calculated parameters closer to their original values.
6. As we may see the observed mean, mode, median and standard deviation and that from the original data frame are very closer to each other even some of them are exactly similar. Thus, the data thereby obtained is much similar to original data (though differs slightly) and hence we may state that the data obtained is quite reliable for further investigations.

ii.

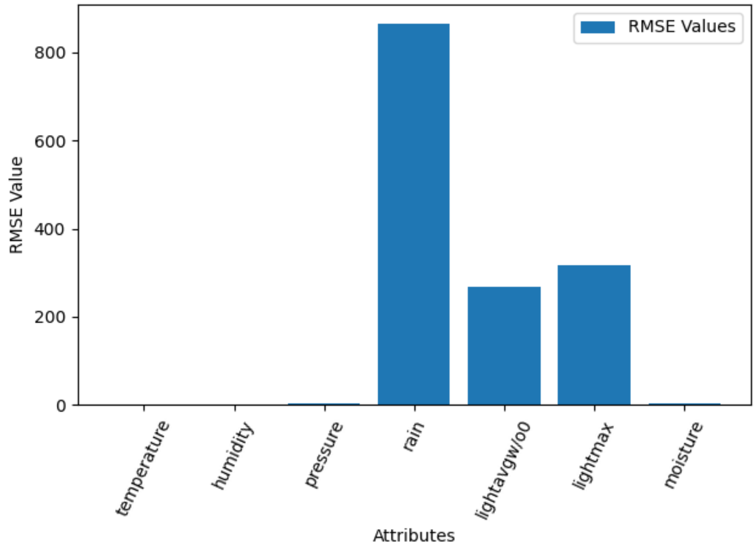

Figure 2 RMSE vs. attributes

Inferences:

1. Attribute “rain” has maximum RMSE while attribute “temperature” has minimum RMSE.
2. As we may observe that there is no consistent relation visible between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values. But it may be observed that attributes that had a considerable amount of values missing from their dataset before are seen to have their calculated parameters closer to their original values.
3. The RMSE value for nearly all attributes except { rain, lighmax, lightavg } are very low in magnitude. Thus, there is large probability that our model may be capable of prediction accurate prediction. Hence, the data is quite reliable for further investigation.

b. i.

Table 3 Mean, mode, median and standard deviation before (original) and after (filling interpolated values) replacing missing values by linear interpolation technique

S. No	Attribute	Before				After			
		Mean	Mode	Median	S.D.	Mean	Mode	Median	S.D.
1	dates	Primary Key							
2	stationid	Primary Key							
3	temperature (in °C)	21.098	12.727	22.14	4.396	21.115	12.727	22.140	4.396
4	humidity (in g.m^{-3})	83.123	99.0	91.18	18.403	83.166	99.0	91.18	18.398
5	pressure (in mb)	1010.052	789.393	1014.925	46.049	1009.968	789.393	1014.925	45.973
6	rain (in ml)	10727.43	0.0	15.75	24834.988	10727.959	0.0	15.75	24834.767
7	lightavgw/o0 (in lux)	4442.747	4488.910	1464.627	7606.675	4496.754	4488.910	1500.5	7645.164
8	lightmax (in lux)	21473.799	4000	6569.0	21933.842	21473.799	4000.0	6569.0	21933.842
9	moisture (in %)	32.572	0.0	13.894	33.832	32.529	0.0	13.894	33.772

Inferences:

1. For mean, attribute “lightavgw/o0” has maximum change while attribute “lightmax” has minimum change (magnitude wise).
2. For mode, surprisingly all attributes have minimum change or no change (are equal magnitude wise).
3. For median, attribute “lightavgw/o0” has maximum change while rest all the attributes have minimum change or no change (are equal magnitude wise).

4. For standard deviation, attribute “lightavgw/o0” has maximum change while attribute “temperature” has minimum change or no change (are equal magnitude wise).
5. As we may observe that there is no consistent relation visible between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values. But it may be observed that attributes that had a considerable amount of values missing from their dataset before are seen to have their calculated parameters closer to their original values i.e. temperature etc.
6. As we may see the observed mean, mode, median and standard deviation and that from the original data frame are very closer to each other even some of them are exactly similar. Thus, the data thereby obtained is much similar to original data (though differs slightly) and hence we may state that the data obtained is quite reliable for further investigations.
7. We may clearly observe that imputation done via filling by interpolation method is much better and reliable than by filling values with mean. The observed maximum and the minimum change in the mean, mode, median and standard deviation, is 0.0 or exactly same for the imputation by interpolation method while for the same attribute imputed via mean values differs a lot in magnitude. Hence, Interpolation method is more reliable and precise.

ii.

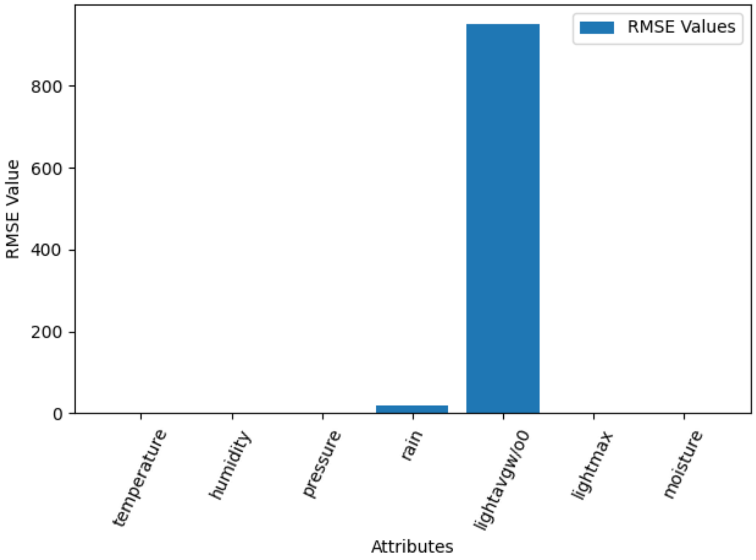

Figure 3 RMSE vs. attributes

Inferences:

1. Attribute “lightavg” has maximum RMSE while attribute “lightmax” has minimum RMSE.
2. It may be seen that there is no consistent relation visible between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values. But it may be

observed that attributes that had a considerable amount of values missing from their dataset before are seen to have their calculated parameters closer to their original values.

3. The RMSE value for nearly all attributes except {“lightavg”} is very low in magnitude. Thus, there is large probability that our model may be capable of prediction accurate prediction. Hence, the data is quite reliable for further investigation.
4. It is visible from the RMSE data for both replacing missing values by mean and linear interpolation that linear interpolation method has drastically reduce RMSE compared to that of missing values by mean. Even the attribute “lightmax” which had RMSE in method 1 has 0 RMSE while interpolating. Thus, this concludes that linear interpolation method is much better that filling mean values with respect to this problem dataset.
5. The linear method ignores the index and treats missing values as equally spaced and finds the best point to fit the missing value after previous points. This, drastically reduced RMSE as visible from the data.

5 a.

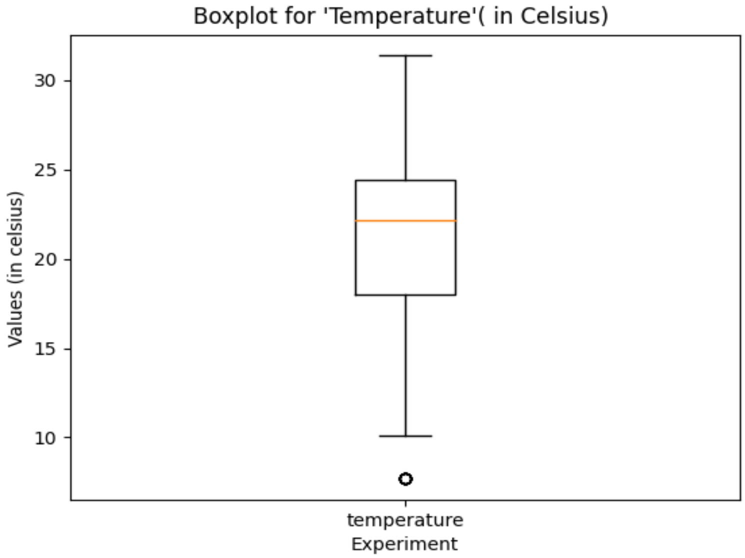

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. It can be observed that 6 outliers (same value) are present below the bottom whisker with values lower than QUARTILE 1 -1.5 * Interquartile range or 10.10 °C.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=24.412-18.005=6.407 °C.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=22.139 °C).

4. It can be observed that the longer part of the box is below the median, thus the data is left-skewed (Negative skewness).

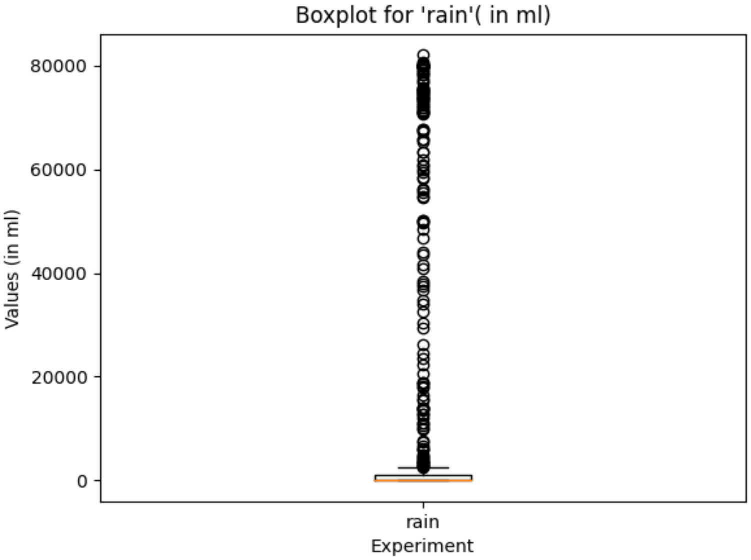

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. It can be observed that a large number of outliers are present above the top whisker with values greater than QUARTILE 3 +1.5 * Interquartile range or 2471 ml.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=1040-(0) =1040 ml.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=15 ml).
4. It can be observed that the longer part of the box is above the median, thus the data is right-skewed (Positive skewness).

b.

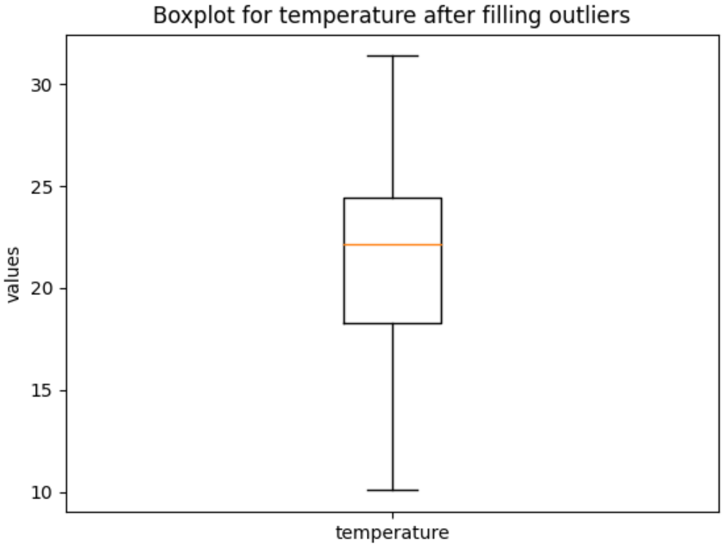

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. We see after there are no outliers visible above after replacing the previous outliers with median.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=24.412-18.291=6.121 °C.
   We may see that previously IQR was 6.407 °C which has now changed to 6.121 °C. Thus, the IQR has decreased slightly.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=22.139 °C). We see the median for the dataset and it's variation about it remains unchanged.
4. It can be observed that the longer part of the box is below the median, thus the data is left-skewed (Negative skewness). The skewness also remains unchanged after replacing outliers with median.

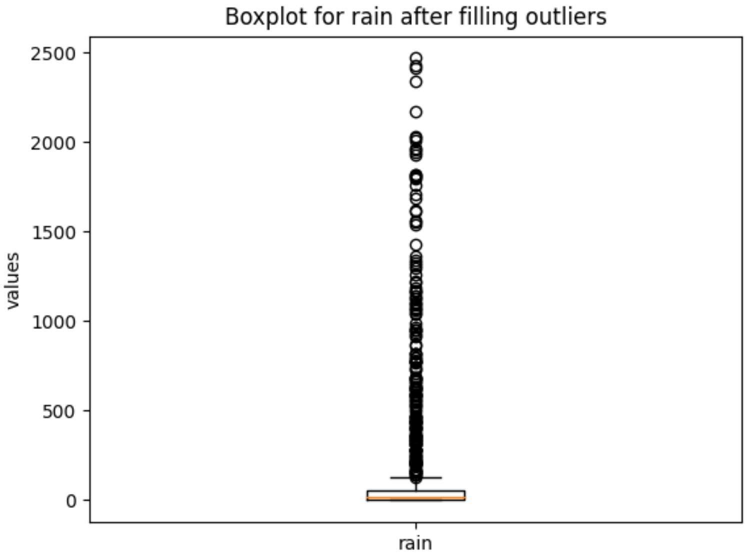

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. It can be observed that a large number of outliers are present above the top whisker with values greater than QUARTILE 3 +1.5 * Interquartile range or 128.3 ml. As compared to previous boxplot for the same attribute before we replaced the outliers, we are getting no noticeable change or the process seems to have failed in this scenario as no outliers seems to be removed. The boxplot seems to have remained unchanged.
2. The Inter quartile range may be calculated as, IQR=QUARTILE 3-QUARTILE 1=51.7-0 =51.7 ml. It is observed that the interquartile range (IQR) has reduced drastically from 1040 ml to 51.7 ml after we replaced the outliers with median of the dataset.
3. In the given boxplot, the QUARTILE 1 and QUARTILE 3 measures the variability of given attribute about the median (QUARTILE 2=15.8 ml). We see the median and its variation about it have changed slightly but no major change has occurred.
4. It can be observed that the longer part of the box is above the median, thus the data is right-skewed (Positive skewness). The skewness also remains unchanged after replacing outliers with median.
5. The reason for the strange behavior of this data and no change in frequency of outliers even after replacing them with median is that we may have a data that is not normally distributed. We have that many outliers that may infer that they aren't outliers, we have a non-normal distribution.
6. Even if we continue to replace the outliers than IQR changes respectively which will keep giving us new outliers.