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

**Student's Name: Akshar Singh** **Mobile No: 7428357700**

**Roll Number: B20147** **Branch: Computer Science**

**1**

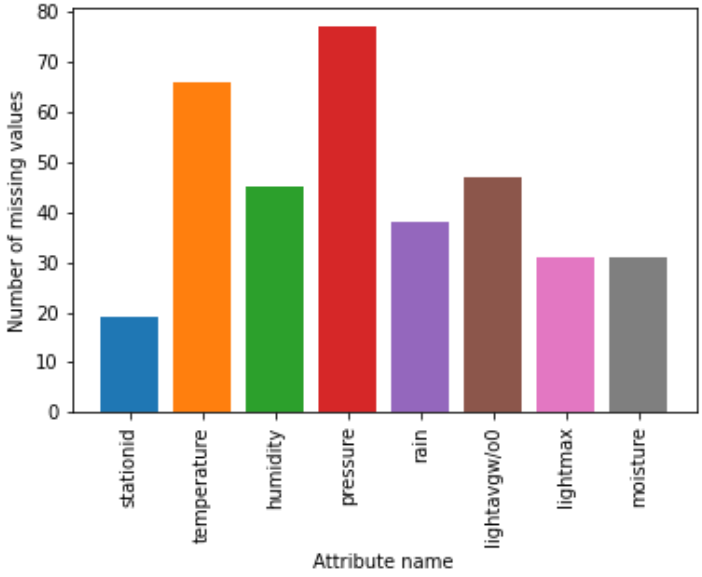


Figure 1 Number of missing values vs. attributes

**Inferences:**

1. Pressure has the highest frequency and station id has the lowest frequency of missing values
2. Frequency for Stationid - 19
3. Frequency for Temperature - 66
4. Frequency for Humidity - 45
5. Frequency for Pressure - 77
6. Frequency for Rain - 38
7. Frequency for Lightavgw/o0 - 47
8. Frequency for Lightmax - 31
9. Frequency for Moisture - 31

**2 a.**

**Inferences:**

1. We choose to delete it because if stationed is missing then we don't know the location of which the data has been collected hence the data becomes useless
2. 19 tuples have been deleted.
3. 2.010 percent of the total number of tuples is deleted.

**b.**

**Inferences:**

1. 35 tuples have been deleted.
2. 3.780 percent of the total number of tuples is deleted.
3. The data which is lost in this step would be of no use because most the other data relating to this is missing.
4. This step was needed in order to preserve the relation between one attribute and the other attribute.

**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) | 57 |
| **4** | humidity (in $g.m^{-3}$) | 37 |
| **5** | pressure (in mb) | 66 |
| **6** | rain (in ml) | 29 |
| **7** | lightavgw/o0 (in lux) | 39 |
| **8** | lightmax (in lux) | 23 |
| **9** | moisture (in %) | 25 |

**Inferences:**

1. Pressure has the maximum missing value and stationed, dates have the minimum missing values.
2. Percentage of data missing for each attribute:-.
   a) Temperature – 6.397
   b) Humidity – 4.152
   c) Pressure – 7.407
   d) Rain – 3.255
   e) lighavgw/o0 – 4.377
   f) lightmax – 2.581
   g) moisture – 2.806
3. 276 total number of missing attributes in the file.

**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) | 21.214 | 12.727 | 22.273 | 4.354 | 21.088 | 21.088 | 21.8 | 4.198 |
| 4 | humidity (in $g.m^{-3}$) | 83.480 | 99 | 91.381 | 18.200 | 83.072 | 99 | 90.177 | 18.236 |
| 5 | pressure (in mb) | 1009.009 | 789.393 | 1014.679 | 46.956 | 1009.117 | 1009.117 | 1014.179 | 45.947 |
| 6 | rain (in ml) | 10701.538 | 0 | 18 | 24839.103 | 11292.896 | 0 | 22.5 | 24979.244 |
| 7 | lightavgw/o0 (in lux) | 4438.428 | 4488.910 | 1656.88 | 7569.155 | 4440.388 | 4488.910 | 1718.13 | 7510.060 |
| 8 | lightmax (in lux) | 21788.623 | 4000 | 6634.0 | 22053.315 | 21497.182 | 4000 | 6801.0 | 21701.667 |
| 9 | moisture (in %) | 32.386 | 0 | 16.704 | 33.635 | 32.765 | 0 | 17.5472 | 33.565 |

**Inferences:**

1. Attributes have the maximum and the minimum change in the mean, mode, median and standard deviation respectively:-

Min :- Lightmax
Max - pressure

2. The maximum change occurred with the attribute having maximum number of missing values and minimum change occurred with attribute having minimum number of missing values
3. The data seems reliable as there is very less change in the central values.
4. Inference 4(You may add or delete the number of inferences)

**ii.**

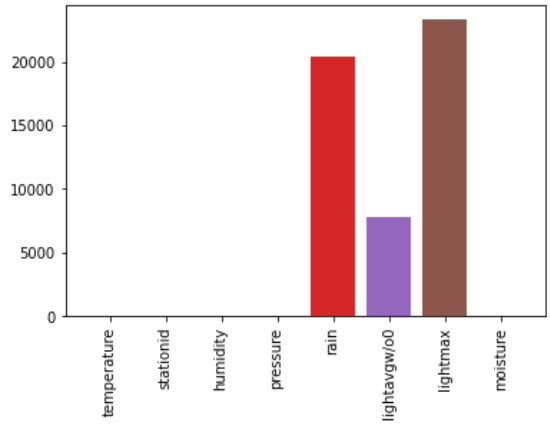


Figure 2 RMSE vs. attributes

**Inferences:**

1. Attributes have maximum and minimum RMSE respectively:-
   Max – Lightmax
   Min - Pressure
2. The maximum change occurred with the attribute having maximum number of missing values and minimum change occurred with attribute having minimum number of missing values which was same as above
3. The data seems reliable since its very low for most of the values and its high value for some attributes is due to the high data value of that data attribute.

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) | 21.214 | 12.727 | 22.273 | 4.354 | 21.192 | 12.727 | 22.134 | 4.283 |
| **4** | humidity (in $g.m^{-3}$) | 83.480 | 99 | 91.381 | 18.200 | 83.247 | 99 | 99.349 | 18.439 |
| **5** | pressure (in mb) | 1009.009 | 789.393 | 1014.679 | 46.956 | 1009.510 | 789.392 | 1014.832 | 46.256 |
| **6** | rain (in ml) | 10701.538 | 0 | 18 | 24839.103 | 11080.413 | 0 | 20.25 | 25174.140 |
| **7** | lightavgw/o0 (in lux) | 4438.428 | 4488.910 | 1656.88 | 7569.155 | 4503.686 | 4488.910 | 1464.629 | 7679.810 |
| **8** | lightmax (in lux) | 21788.623 | 4000 | 6634.0 | 22053.315 | 21417.007 | 4000 | 6569 | 21918.632 |
| **9** | moisture (in %) | 32.386 | 0 | 16.704 | 33.635 | 32.493 | 0 | 15.157 | 33.743 |

**Inferences:**

1. Attributes have the maximum and the minimum change in the mean, mode, median and standard deviation respectively :-
   Min – Pressure, Max- Lightmax
2. There is no relation between missing values and the central value of the data
3. The data seems reliable since its very low for most of the values and its high value for some attributes is due to the high data value of that data attribute.
4. By interpolation method, the error was minimum as compared to that by the mean method.

**ii.**

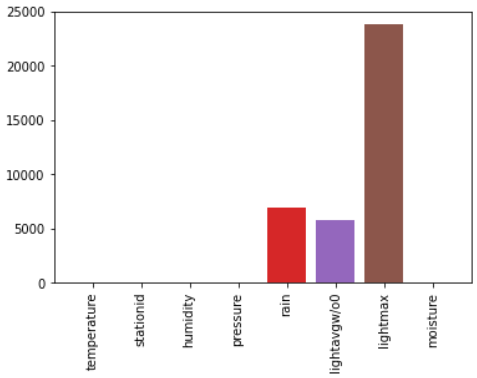


Figure 3 RMSE vs. attributes

**Inferences:**

1. Attributes have maximum and minimum RMSE respectively :-
   Min : temperature
   Max : Lightmax
2. The maximum change occurred with the attribute having maximum number of missing values and minimum change occurred with attribute having minimum number of missing values
3. The data seems reliable since its very low for most of the values and its high value for some attributes is due to the high data value of that data attribute.
4. By interpolation method, the RSME was less as compared to that by the mean method.

**5 a.**

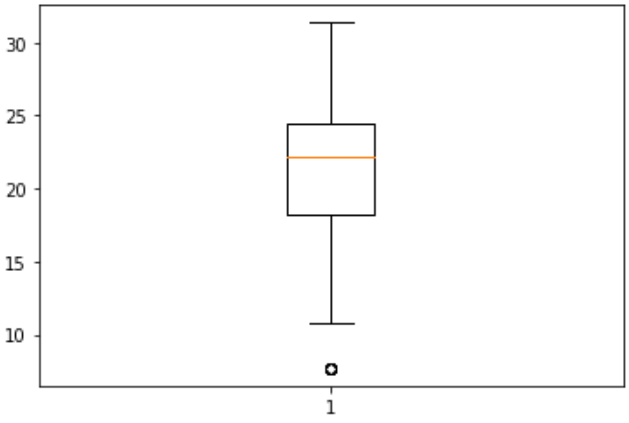


Figure 4 Boxplot for attribute temperature (in °C)

**Inferences:**

1. There are a few outliers below the lower whisker
2. IQR = 6.102
3. Variance = 18.727
4. The data seems to be negatively skewed

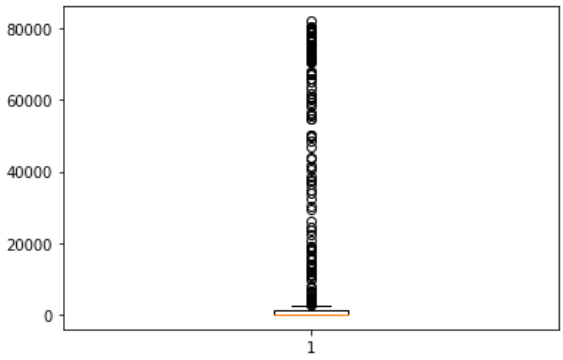


Figure 5 Boxplot for attribute rain (in ml)

**Inferences:**

1. There are many outliers above upper whisker.
2. IQR = 987.75
3. Variance = 634435344 approx.
4. Data seems to be positively skewed.

**b.**

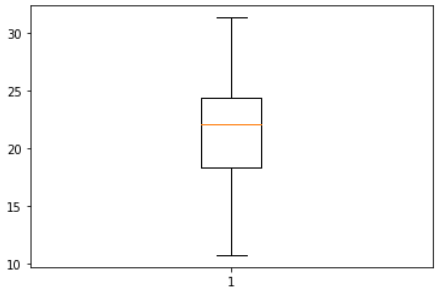


Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

**Inferences:**

1. There are no outliers
2. IQR = 5.993 less than above
3. Variance = 16.917 less than above
4. Data seems to be somewhat negatively skewed but less skewed than above

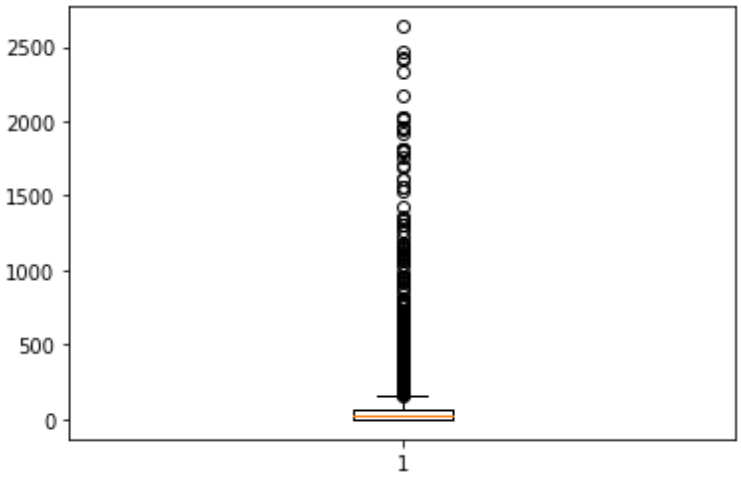


Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

**Inferences:**

1. There are many outliers above the upper whisker.
2. IQR = 64.125 very much less than above
3. Variance = 164025 approx.

4. Data seems to be positively skewed but less skewed than above.