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

**Student's Name: Aryan Tiwari** **Mobile No: 8982562898**

**Roll Number: B20187** **Branch:** Electrical Engineering

**1**

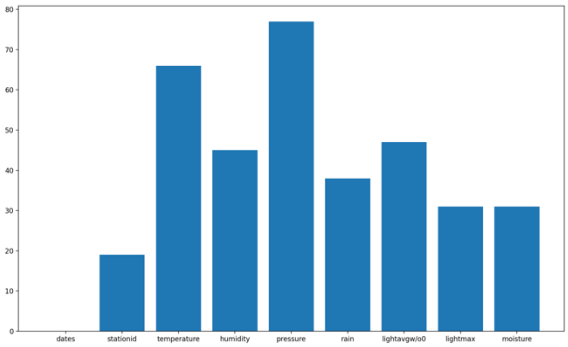


Figure 1 Number of missing values vs. attributes

**Inferences:**

1. Maximum missing values: pressure

   Minimum missing values: dates

2.

```
b 2/Lab Assignment 2-20210903/q1.py"
dates             0
stationid        19
temperature      66
humidity         45
pressure         77
rain             38
lightavgw/o0     47
lightmax         31
moisture         31
```

**2 a.**

**Inferences:**

1. Target attribute is station-id which is crucial to know the location of landslide prone area, unavailability of location prevents us to analyze landslide prone areas.
2. 19 tuples are deleted in this step
3. Percentage of the total number of tuples deleted is 2.01%

**b.**

**Inferences:**

1. The number of tuples deleted after this step are 35.
2. Percentage of the total number of tuples deleted is 3.70%
3. The data loss is less than 5% and had more than 33% values missing.
4. Evaluating and analyzing missing data is a waste of time and rows with more than 33% missing values don't give much information.

**3**

Table 1 Number of missing values per attribute after removing missing values

```
Number of missing values in each attribute after step 2
 dates            0
stationid         0
temperature      34
humidity         13
pressure         41
rain              6
lightavgw/o0     15
lightmax          1
moisture          6
```

**Inferences:**

1. Maximum missing values: pressure

   Minimum missing values: date, station-id

2.

| S. No | Attribute | % of missing values |
|---|---|---|
| **1** | dates | 0 |
| **2** | stationid | 0 |
| **3** | temperature (in °C) | 3.59 |
| **4** | humidity (in $g.m^{-3}$) | 1.37 |
| **5** | pressure (in mb) | 4.33 |
| **6** | rain (in ml) | 0.63 |
| **7** | lightavgw/o0 (in lux) | 1.58 |
| **8** | lightmax (in lux) | 0.1 |
| **9** | moisture (in %) | 0.63 |

3. The total number of missing attributes in the file are 116.

**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) | 21.21 | 12.72 | 22.27 | 4.35 | 21.07 | 21.05 | 21.92 | 4.34 |
| **4** | humidity (in $g.m^{-3}$) | 83.48 | 99 | 91 | 18.21 | 83.12 | 99 | 91 | 18.39 |
| **5** | pressure (in mb) | 1009.01 | 789.39 | 1014.67 | 46.98 | 1009.46 | 1009.46 | 1014.48 | 45.85 |
| **6** | rain (in ml) | 10701.54 | 0 | 18 | 24852.2 | 10798.38 | 0 | 15.75 | 24833.96 |
| **7** | lightavgw/o0 (in lux) | 4438.42 | 4488.91 | 1656.88 | 7573.16 | 4458.29 | 4488.91 | 1502.94 | 7606.28 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **8** | lightmax (in lux) | 21788.62 | 4000 | 6634.0 | 22064.9 | 21463.221 | 4000 | 6569 | 21943.88 |
| **9** | moisture (in %) | 32.38 | 0 | 16.704 | 33.65 | 32.6 | 0 | 14.17 | 33.71 |

**Inferences:**

1. Mean: Max(Lightmax); Min(Temp)
   Median : Max(lightavgw/o0) ; Min(pressure)
   Mode: Max(pressure); Min()
   Std-dev: Max(lightmax); Min(temp)
2. The deviation is quite small and hence is reliable.

**ii.**

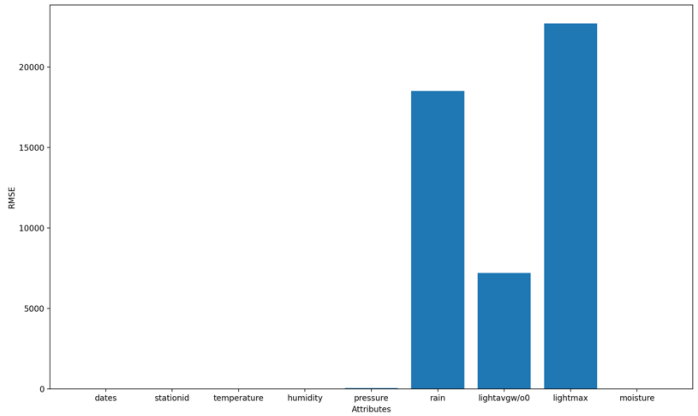


Figure 2 RMSE vs. attributes

**Inferences:**

1. Max RMSE: ‘rain’; Min: ‘temperature’
2.

```
RMSE of temperature is 3.754920706652186
RMSE of humidity is 13.593367701425382
RMSE of pressure is 44.67907886243912
RMSE of rain is 18514.713587699505
RMSE of lightavgw/o0 is 7204.406021900535
RMSE of lightmax is 22711.616895729538
RMSE of moisture is 22.07774631718766
```

3. The large values of rmse for rain and lightmax shows that the data varies a lot from original and hence is unfit for analysis.

**b. i.**

| S. No | Attribute | Before | | | | After | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Mode | Median | S.D. | Mean | Mode | Median | S.D. |
| 1 | dates | | | | | | | | |
| 2 | stationid | | | | | | | | |
| 3 | temperature (in °C) | 21.21 | 12.72 | 22.27 | 4.35 | 21.115 | 12.727 | 22.14 | 4.399 |
| 4 | humidity (in $g.m^{-3}$) | 83.48 | 99 | 91 | 18.21 | 83.166 | 99 | 91.18 | 18.408 |
| 5 | pressure (in mb) | 1009.01 | 789.39 | 1014.67 | 46.98 | 1009.968 | 789.393 | 1014.925 | 45.99 |
| 6 | rain (in ml) | 10701.54 | 0 | 18 | 24852.2 | 10727.959 | 0 | 15.75 | 24848.71 |
| 7 | lightavgw/o0 (in lux) | 4438.42 | 4488.91 | 1656.88 | 7573.16 | 4496.754 | 4488.91 | 1500.5 | 7649.458 |
| 8 | lightmax (in lux) | 21788.62 | 4000 | 6634.0 | 22064.9 | 21473.799 | 4000 | 6569 | 21946.16 |
| 9 | moisture (in %) | 32.38 | 0 | 16.704 | 33.65 | 32.529 | 0 | 13.894 | 33.791 |

Table 3 Mean, mode, median and standard deviation before and after replacing missing values by linear interpolation technique

**Inferences:**

1. Mean: Max(Lightmax); Min(Temp)
   Median : Max(lightavgw/o0) ; Min(temp)
   Mode: No-change
   Std-dev: Max(lightmax); Min(temperature)
2. The variation of stats is less after replacement.
3. Linear interpolation is better as it has lesser deviation than mean replacement.

**ii.**

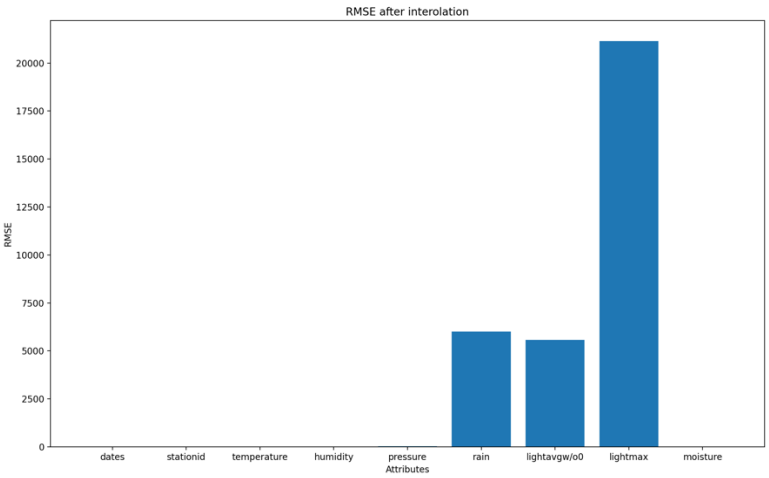


Figure 3 RMSE vs. attributes

**Inferences:**

1. Max RMSE: 'lightavgw/o0 '; Min: 'lightmax'
2.

```
b 2/Lab Assignment 2-20210903/q4_b.py"
RMSE of temperature is 1.8556170639412197
RMSE of humidity is 3.846215375271329
RMSE of pressure is 24.63997651139289
RMSE of rain is 6014.847129201579
RMSE of lightavgw/o0 is 5567.3136390853
RMSE of lightmax is 21150.851782705126
RMSE of moisture is 7.031077545743664
```

3. The high RMSE of lightmax shows unreliability of data.
4. The RMSE attributes have reduced with interpolation

**5 a.**

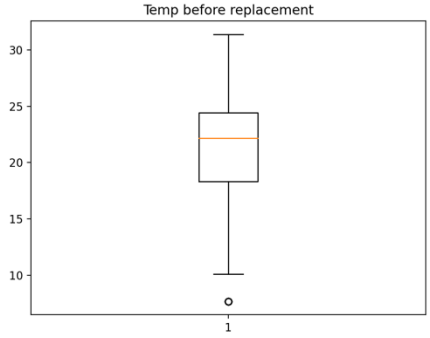


Figure 4 Boxplot for attribute temperature (in °C)

**Inferences:**

1. There are 10 outliers.

The outliers for temperature are as follows

| | dates | stationid | temperature | humidity | pressure | rain | lightavgw/o0 | lightmax | moisture |
|---|---|---|---|---|---|---|---|---|---|
| 509 | 16-11-2018 | t15 | 7.6729 | 67.8972 | 1022.555421 | 0.0 | 417.0392 | 4000.0 | 8.1028 |
| 510 | 17-11-2018 | t15 | 7.6729 | 67.8972 | 1020.869643 | 0.0 | 417.0392 | 4000.0 | 7.9286 |
| 511 | 18-11-2018 | t15 | 7.6729 | 67.8972 | 1022.956262 | 0.0 | 417.0392 | 4000.0 | 7.9159 |
| 512 | 19-11-2018 | t15 | 7.6729 | 67.8972 | 1026.069680 | 0.0 | 417.0392 | 4000.0 | 7.0560 |
| 513 | 20-11-2018 | t15 | 7.6729 | 67.8972 | 1026.240417 | 0.0 | 417.0392 | 4000.0 | 7.0000 |
| 514 | 21-11-2018 | t15 | 7.6729 | 67.8972 | 1025.567290 | 0.0 | 417.0392 | 4000.0 | 7.0000 |
| 515 | 22-11-2018 | t15 | 7.6729 | 67.8972 | 1026.253462 | 0.0 | 417.0392 | 4000.0 | 6.9327 |
| 516 | 23-11-2018 | t15 | 7.6729 | 67.8972 | 1028.316140 | 0.0 | 417.0392 | 4000.0 | 7.0000 |
| 517 | 24-11-2018 | t15 | 7.6729 | 67.8972 | 1025.770238 | 0.0 | 417.0392 | 4000.0 | 6.2143 |
| 518 | 25-11-2018 | t15 | 7.6729 | 67.8972 | 1029.755156 | 0.0 | 417.0392 | 4000.0 | 6.0000, |

2. IQR: 6.1
3. Left skewed.

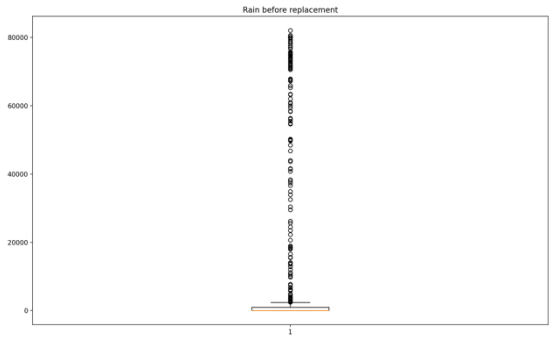


Figure 5 Boxplot for attribute rain (in ml)

**Inferences:**

1. Number of outliers is 184
2. IQR: 987.75
3. Right skewed

**b.**

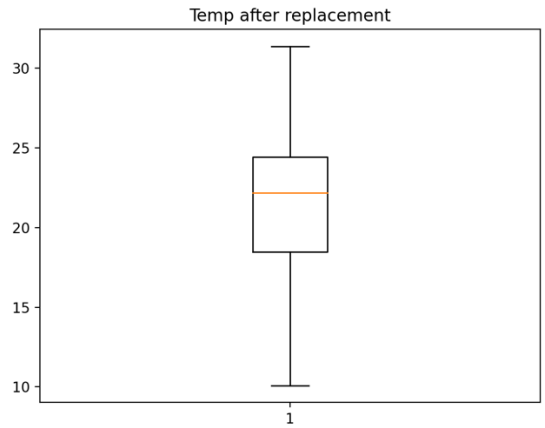


Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

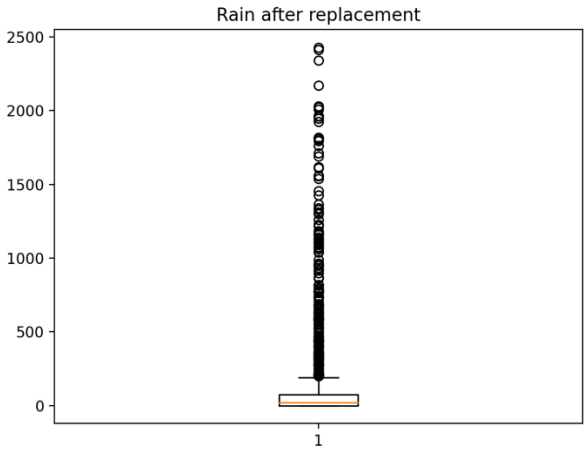


Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers