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

**Student's Name: Garvit Verma** **Mobile No: 8272840777**

**Roll Number: B20098** **Branch: CSE**

**1**

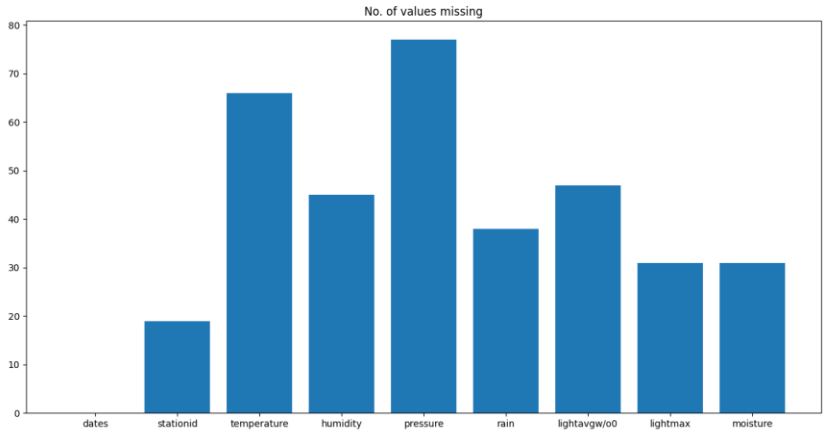


Figure 1 Number of missing values vs. attributes

**Inferences:**

1. Attribute 'pressure' has maximum no. of missing values and 'dates' has 0 missing values.
2. Frequency of missing values as per attributes are as follows:
   **dates:** 0, **stationed:** 19, **temperature:** 66, **humidity:** 45, **pressure:** 77, **rain:** 38, **lightavgw/o0:** 47, **lightmax:** 31**, moisture:** 31.

**2 a.**

**Inferences:**

1. We deleted the tuple whose target attribute values are missing so as to remove any nuisance in our data because there are no values in that particular cell.
2. Number of tuples deleted are 19.
3. 2.01 percent of the total no. of tuples are deleted.

**b.**

**Inferences:**

1. Total no. of tuples deleted are 35.
2. 3.77 percent of the total no. of tuples are deleted.
3. Some data has been lost for other attributes as well whose values were not nan.
4. The need for this step is because there are equal to or than 3 attributes whose values were missing have to be deleted so as to analyze the data clearly.

**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. Attributes ‘pressure’ has maximum and ‘dates’, ‘stationid’ has minimum no. of missing values.
2. Percentage of missing values as per the attributes are as follows:
   **dates: 0, stationed: 0, temperature: 3.81, humidity: 1.45, pressure: 4.6, rain: 0.67, lightavgw/o0: 1.67, lightmax: 0.11, moisture: 0.67.**
3. Total no. of missing values in 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 | After | | | | Before | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Mode | Median | S.D. | Mean | Mode | Median | S.D. |
| 1 | dates | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | stationid | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | Temperature (in °C) | 21.078 | 21.078 | 21.8 | 4.243 | 21.215 | 12.727 | 22.273 | 4.356 |
| 4 | Humidity(in $g.m^{-3}$) | 83.262 | 99 | 90.119 | 17.968 | 83.479 | 99 | 91.381 | 18.210 |
| 5 | pressure(in mb) | 1009.225 | 1009.225 | 1014.071 | 45.215 | 1009.008 | 789.392 | 1014.678 | 46.980 |
| 6 | rain (in ml) | 10942.726 | 0 | 24.750 | 24574.252 | 10701.538 | 0 | 18 | 24852.255 |
| 7 | lightavgw/o0 (in lux) | 4430.928 | 4488.910 | 1911.234 | 7400.586 | 4438.428 | 4488.910 | 1656.88 | 7573.163 |
| 8 | lightmax(in lux) | 21650.163 | 4000 | 7544 | 21678.196 | 21788.623 | 4000 | 6634 | 22064.993 |
| 9 | moisture(in %) | 32.672 | 0 | 17.723 | 33.416 | 32.386 | 0 | 16.704 | 33.653 |

**Inferences:**

1. **Maximum** change in mean is of attribute ‘rain’, in mode is in ‘pressure’, in median is in ‘lightmax’, in stp.dev is in ‘lightmax’. **Minimum** change in mean is in attribute ‘temperature’, in mode is in ‘humidity’, ’rain’, ‘lightmaxavgw/o0’, ‘lightmax’, ‘moisture’ i.e., 0, in median is in ‘temperature’, in stp.dev is in ‘temperature’.
2. There are minimum missing values in ‘lightmax’ and max. change in median and stp.dev, max missing values in ‘pressure’ and max change in mode, there are second most maximum no. of missing values in ‘temperature’ and minimum change in mean, median and stp.dev.
3. As we can see for many of the attributes, the change is minimum, so the data is reliable for further investigations.

**ii.**

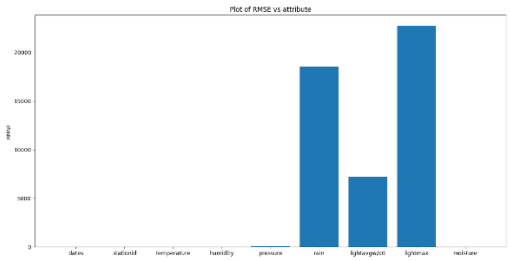


Figure 2 RMSE vs. attributes

**Inferences:**

1. Maximum RMSE is of attribute 'lightmax' and minimum is of attribute 'temperature'.
2. Maximum rmse is of 'lightmax' which has minimum number of missing values and maximum change in median and stp.dev. Minimum rmse is of 'temperature' which also has minimum change in mean, median and stp.dev.
3. RMSE values for three attributes are much higher than expected, and others are small which can be adjusted So, if we neglect those attributes, data can be used for further investigations.

**b. i.**

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

| S. No | Attribute | After | | | | Before | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Mode | Median | S.D. | Mean | Mode | Median | S.D. |
| 1 | dates | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | stationid | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | temperature (in °C) | 21.196 | 12.727 | 22.169 | 4.329 | 21.215 | 12.727 | 22.273 | 4.356 |
| 4 | humidity (in $g.m^{-3}$) | 83.538 | 99 | 91.380 | 18.206 | 83.479 | 99 | 91.381 | 18.210 |
| 5 | pressure (in mb) | 1009.265 | 789.392 | 1014.677 | 45.998 | 1009.008 | 789.392 | 1014.678 | 46.980 |
| 6 | rain (in ml) | 10651.638 | 0 | 22.500 | 24779.512 | 10701.538 | 0 | 18 | 24852.255 |
| 7 | lightavgw/o0 (in lux) | 4486.340 | 4488.910 | 1623.494 | 7573.795 | 4438.428 | 4488.910 | 1656.88 | 7573.163 |
| 8 | lightmax (in lux) | 21517.191 | 4000 | 6569 | 21935.165 | 21788.623 | 4000 | 6634 | 22064.993 |
| 9 | moisture (in %) | 32.327 | 0 | 16.306 | 33.602 | 32.386 | 0 | 16.704 | 33.653 |

**Inferences:**

1. Maximum change in mean is in attribute 'lightmax', in mode is in all changes are 0, in median 'lightmax', in stp.dev 'lightmax'. Minimum change in mean is in attribute 'temperature', in mode all changes are 0, in median 'humidity' and 'pressure', in stp.dev 'humidity'.
2. 'lightmax' has minimum no. of missing values and maximum change in mean, median, stp.dev. 'pressure' has maximum no. of missing values and minimum change in median.
3. As we can see for many of the attributes, the change is minimum, so the data is reliable for further investigations.
4. As we can see that the change after replacing missing values by interpolation method in mean, median, mode and stp.dev is less as compared to that when replaced by mean. So, here the interpolation method is useful for further investigations.

**ii.**

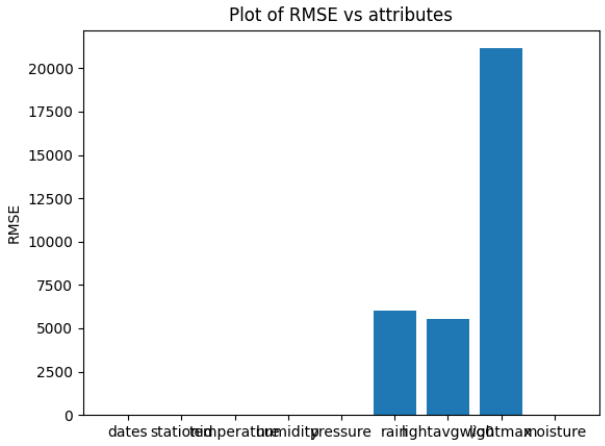


Figure 3 RMSE vs. attributes

**Inferences:**

1. Maximum RMSE is of attribute 'lightmax' and minimum is of 'temperature'.
2. 'lightmax' has max rmse and also has minimum no. of missing values, while 'temperature has minimum rmse and second most no. of missing values.
3. RMSE values for three attributes are much higher than expected, and others are small which can be adjusted So, if we neglect those attributes, data can be used for further investigations.
4. Replacing by interpolation method is useful here as the values of rmse has been reduced here as compared to that when replacing by mean.

**5 a.**

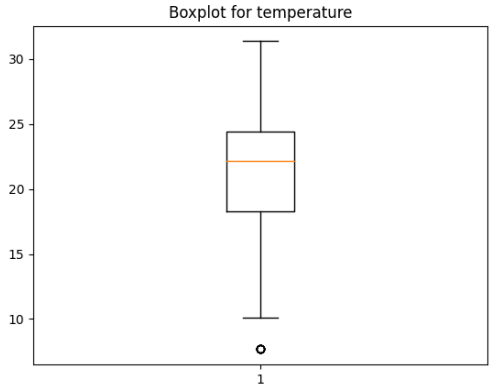


Figure 4 Boxplot for attribute temperature (in °C)

**Inferences:**

1. Number of outliers are 10 with all values as 7.6729, with row no. 509-518.
2. Inter quartile range is 6.10198.
3. Data lies in the range 31.375-7.6729.
4. Data is Negatively skewed.

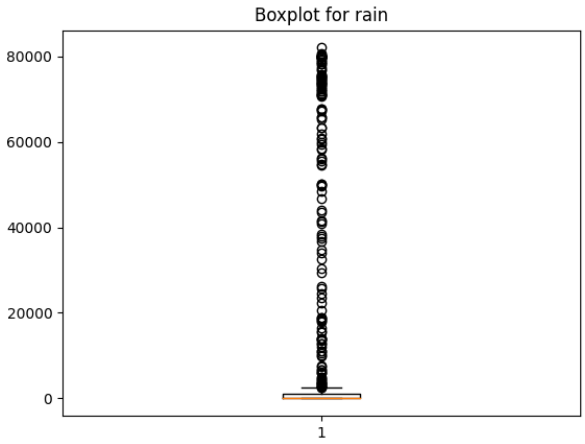


Figure 5 Boxplot for attribute rain (in ml)

**Inferences:**

1. Number of outliers are 185, where outliers above upper bound only exists.
2. Inter quartile range is 987.75.
3. Data lies in the range 82037.25-0.
4. Data is Positively skewed.

**b.**

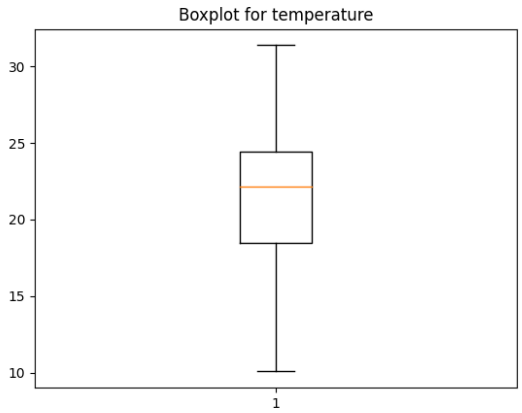


Figure 6 Boxplot for attribute temperature (in °C) after replacing outliers with median

**Inferences:**

1. There are no outliers left after replacing outliers with median.
2. Interquartile range is 5.9344, which is smaller as compared to Q5. a.
3. Data lies in the range 31.375-10.08511, where the minimum value has been changed as compared to Q5. a.
4. Data is Negatively skewed.

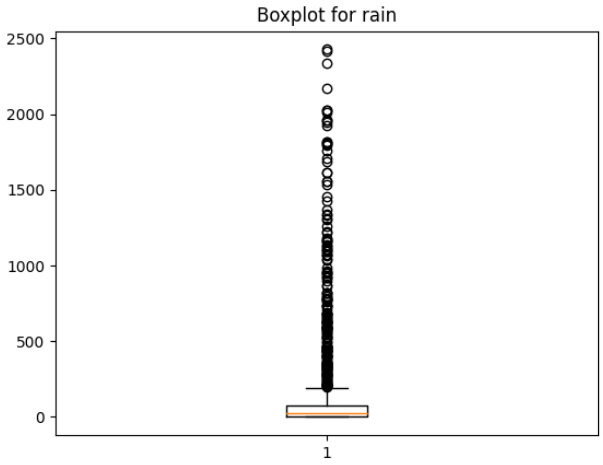


Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

**Inferences:**

1. Total no. of outliers is 193.
2. Interquartile range is 76.5, which is very much smaller than Q5. a.
3. Data lies in the range 2427.75-0, where the maximum value have been changed as compared to Q5. a.
4. Data is Positively skewed.