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

**Student's Name: Vikas Dangi** **Mobile No: 9406661661**

**Roll Number: B20238** **Branch: Electrical Engineering**

**1**

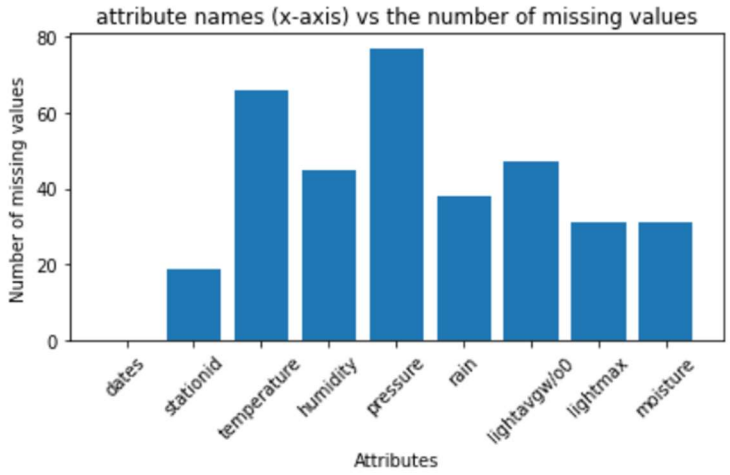


Figure 1 Number of missing values vs. attributes

**Inferences:**

1. Pressure and Dates have maximum and minimum missing values respectively
2. From the bar chart the frequency of missing values of stationed and date is much less as these attributes are not received from any sensors in real life

**2 a.**

**Inferences:**

1. We choose to delete the tuple if the target attribute is missing because even if we get some insight in the data, we won't know the about who holds it so it is of no use.
2. 19 tuples deleted after this step.
3. 2.01 percentage of the total number of tuples is deleted.

**b.**

**Inferences:**

1. 35 number of tuples got deleted after this step.
2. 3.7 percentage of the total number of tuples is deleted?
3. The data loss is not much it is worth the correct information we would get after cleaning it
4. The tuple where we get 3 more than 3 data values missing is can give bad results because filling the values is just a compromise, we make not so it is highly probable that it will interfere with the correct results.

**3**

| 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 |

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

**Inferences:**

1. Pressure and temperature have maximum values missing and dates and stationed, dates and lightmax have minimum missing values.
2. Here is the percentage change in the number of values:

| Attributes | Percentage of missing values |
|---|---|
| dates | 0 |
| stationid | 0 |
| temperature | 3.8159 |
| humidity | 1.4590 |
| pressure | 4.6015 |
| rain | 0.6734 |
| lightavgw/o0 | 1.6835 |
| lightmax | 0.1122 |
| moisture | 0.6734 |

3. 116 is the total number of missing attributes in the file.

**4 a. i.**

Table 2 Mean, mode, median and standard deviation of the cleaned (By filling NA with mean) and original file

| S. No | Attribute | (Calculated using filling mean) | | | | (Original file: Real Data) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Mode | Median | S.D. | Mean | Mode | Median | S.D. |
| 1 | dates | NA | NA | NA | NA | NA | NA | NA | NA |
| 2 | stationid | NA | NA | NA | NA | NA | NA | NA | NA |
| 3 | temperature (in °C) | 21.052 | 21.052 | 21.927 | 4.340 | 21.05 | 12.72 | 22.27 | 4.36 |
| 4 | humidity (in $g.m^{-3}$) | 83.126 | 99.000 | 91.000 | 18.394 | 83.12 | 99.00 | 91.38 | 18.21 |
| 5 | pressure (in mb) | 1009.466 | 1009.466 | 1014.482 | 45.856 | 1009.46 | 789.39 | 1014.67 | 46.98 |
| 6 | rain (in ml) | 10798.379 | 0.000 | 15.750 | 24833.965 | 10798.38 | 0.00 | 18.00 | 24852.25 |
| 7 | lightavgw/o 0 (in lux) | 4458.298 | 4488.910 | 1502.938 | 7606.284 | 4458.29 | 4488.91 | 1656.88 | 7573.16 |
| 8 | lightmax (in lux) | 21463.221 | 4000 | 6569 | 21943.889 | 21463.22 | 4000.00 | 6634.00 | 22064.99 |
| 9 | moisture (in %) | 32.603 | 0.000 | 14.170 | 33.714 | 32.60 | 0.00 | 16.70 | 33.65 |

**Inferences:**

1. Mean: Maximum change - Rain attribute; Minimum Change- Pressure Attribute
   Mode: Maximum change- Temperature; Minimum Change- All other than pressure and temperature have not changed.
   Standard Deviation: Maximum Change-Pressure; Minimum Change- Humidity
   Median: Maximum change - lightavgw; Minimum Change- No change in lightmax and rain
   Standard Deviation: Maximum Change-Pressure; Minimum Change- Humidity
2. It was discovered that the two properties with the most missing values were pressure and temperature. The highest change in standard deviation occurred in pressure, while the maximum change in mode value occurred in temperature. Lightmax only has one missing value, thus has minimum change. The quantity of missing values in an attribute correlates with the amount of change in the values of central tendencies.
3. We can declare that yes, this data can be used for further research because the percentage change values are minor and the percentage of missing values does not exceed a lot.

**ii.**

**Note:** The graph has logarithm scale on the Y-axis as the range was quite high

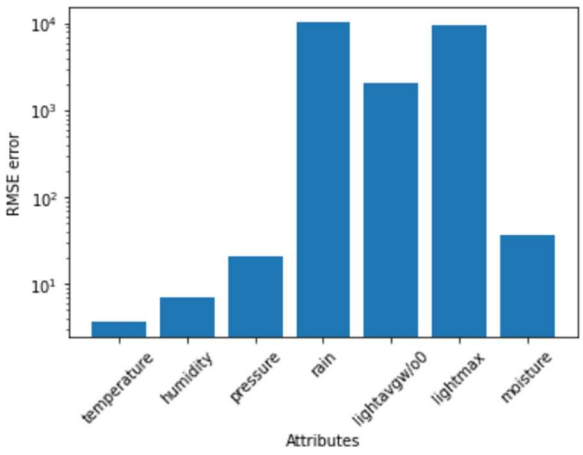


Figure 2 RMSE vs. attributes

**Inferences:**

1. Rain attributes has maximum and temperature has minimum RMSE respectively.
2. Although there is no as such direct relation between RMSE and the change in other parameters. It can be observed a bit that the RMSE values of attributes that experienced the most or least change in other central tendencies are comparatively small.
3. The data isn't particularly relevant for further research because some attributes have extremely high RMSE values.

**b. i.**

Table 3 Mean, mode, median and standard deviation b/w cleaned by linear interpolation technique and original file

| S. No | Attribute | (Calculated using Linear Interpolation) | | | | (Original file: Real Data) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Mode | Median | S.D. | Mean | Mode | Median | S.D. |
| 1 | dates | NA | NA | NA | NA | NA | NA | NA | NA |
| 2 | stationid | NA | NA | NA | NA | NA | NA | NA | NA |
| 3 | temperature (in °C) | 21.115 | 12.727 | 22.140 | 4.399 | 21.05 | 12.72 | 22.27 | 4.36 |
| 4 | humidity (in $g.m^{-3}$) | 83.166 | 99 | 91.180 | 18.408 | 83.12 | 99.00 | 91.38 | 18.21 |
| 5 | pressure (in mb) | 1009.968 | 789.393 | 1014.925 | 45.999 | 1009.46 | 789.39 | 1014.67 | 46.98 |
| 6 | rain (in ml) | 10727.959 | 0 | 15.75 | 24848.715 | 10798.38 | 0.00 | 18.00 | 24852.25 |
| 7 | lightavgw/o 0 (in lux) | 4496.754 | 4488.910 | 1500.5 | 7649.458 | 4458.29 | 4488.91 | 1656.88 | 7573.16 |
| 8 | lightmax (in lux) | 21473.799 | 4000 | 6569 | 21946.160 | 21463.22 | 4000.00 | 6634.00 | 22064.99 |
| 9 | moisture (in %) | 32.529 | 0.000 | 13.894 | 33.791 | 32.60 | 0.00 | 16.70 | 33.65 |

**Inferences:**

1. Mean: Maximum change- Rain attribute Minimum Change- Pressure Attribute
   Mode: Maximum change- None Minimum Change- No change in any attribute
   Standard Deviation: Maximum Change-Pressure Minimum Change- Humidity and Temperature
   Median: Maximum change - lightavgw Minimum Change- No change in lightmax, rain and pressure
2. The change in values of central tendencies after replacement is less for attributes with fewer missing data. Although there is no such sharp line of distinguishment.
3. We can declare that yes, this data can be used for further research because the percentage change values are very less and the proportion of missing values does not exceed much.
4. As we know that if we replace the missing values by mean the mean of the data doesn't change much compared to the original data. But other parameters seem to vary to a fair extend in that. In the interpolation very less difference is seen in the mode as compared to the previous one. We can say that each of them is better than the other in some aspects. But interpolation seems to have much more balanced change if we see the other attributes other than mean.

**ii.**

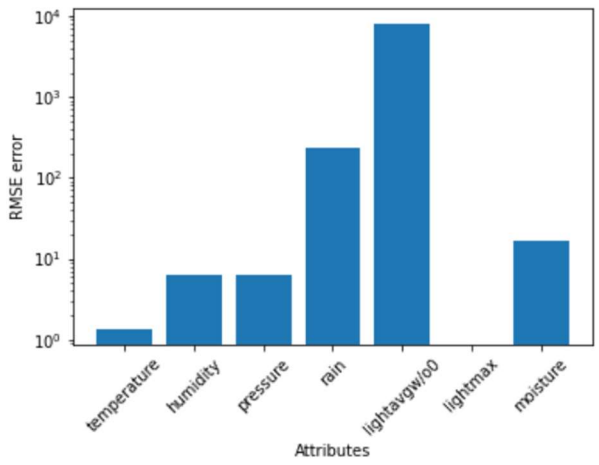


Figure 3 RMSE vs. attributes

**Inferences:**

1. The RMSE of lightavgw is the highest, and the RMSE of lightmax is the lowest.
2. No such direct full proved relation can be made for RMSE although and others, maximum minimum change in mean, mode, median and standard deviation and maximum and minimum missing values had some relations previously.
3. When values were substituted using interpolation instead of the mean, the RMSE values were lower in general. Lightmax experienced the most dramatic change, with the RMSE value for the interpolation technique becoming extremely small. It conveys to us that interpolation is somehow good in this sense.
4. The data isn't particularly relevant for further research because some attributes have extremely high RMSE values.

**5 a.**

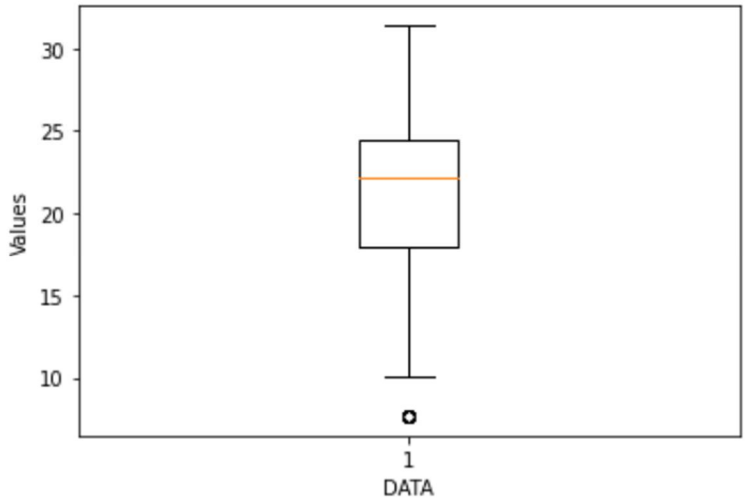


Figure 4 Boxplot for attribute temperature (in °C)

**Inferences:**

1. The following are the row numbers for the outliers: 462, 463, 464, 465, 466, 467, 468, 469, 470, 471.
2. The Inter quartile range is 6.40
3. Range = Highest value – Smallest Value = 23.7 and the variance is 19.3
4. As the middle orange line is not in the middle and symmetric it is skewed and as the bottom part is wider it is a left skewed.

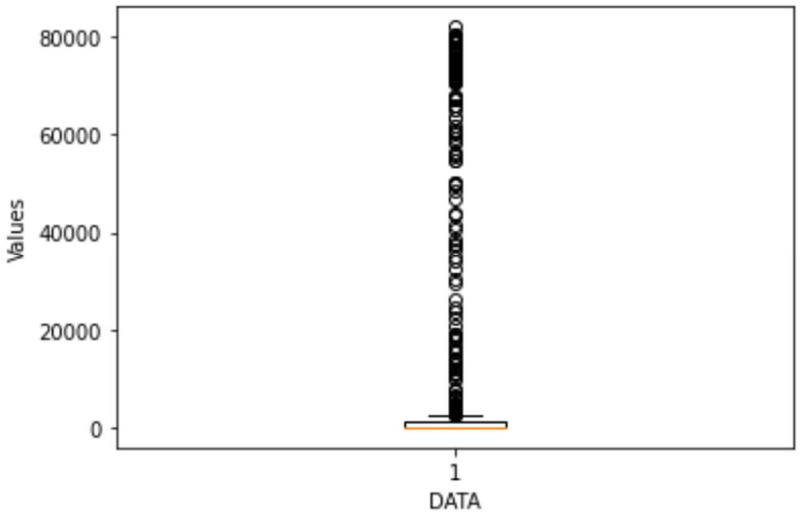


Figure 5 Boxplot for attribute rain (in ml)

**Inferences:**

1. There are 182 outliers in the Rain and there row numbers are 122, 183, 184, 185, 190, 300, 301, 302, 583, 584, 585, 589, 590, 591, 646, 647, 649, 650, 652, 655, 657, 658, 664, 691, 692, 693, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 735, 736, 737, 738, 739, 740, 741, 742, 743, 745, 746, 747, 748, 749, 750, 772, 773, 774, 775, 776, 778, 782, 783, 787, 788, 789, 790, 793, 794, 798, 800, 801, 802, 803, 804, 805, 806, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 869, 870, 871, 872, 873, 874, 875, 876, 877, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890
2. IQR = q3 – q1 = 1041.75– 0 = 1041.75
3. Range = Highest value – Smallest Value =82037.25
   Variance = 617458628.2
4. The data is highly right skewed/positively skewed as the orange lies very low without showing any symmetry.

**b.**

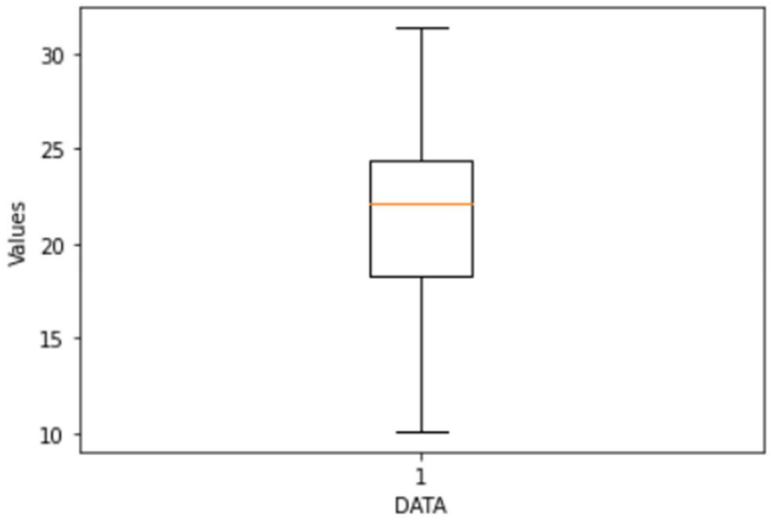


Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

**Inferences:**

1. The following are the row numbers for the outliers: 509, 510, 511, 512, 513, 514, 515, 516, 517, and 518. The number of outliers has remained constant only the row numbers have altered. The outliers have relocated to the bottom half of the distribution (500-900 range).
2. IQR = q3 – q1 = 6.1, It has not changed much as the outliers were replaced with the median
3. Variance = 17.3, variance and range have decreased after median replaced the outliers.
4. The data is left skewed or negatively skewed as the middle orange line lies below making it asymmetric.

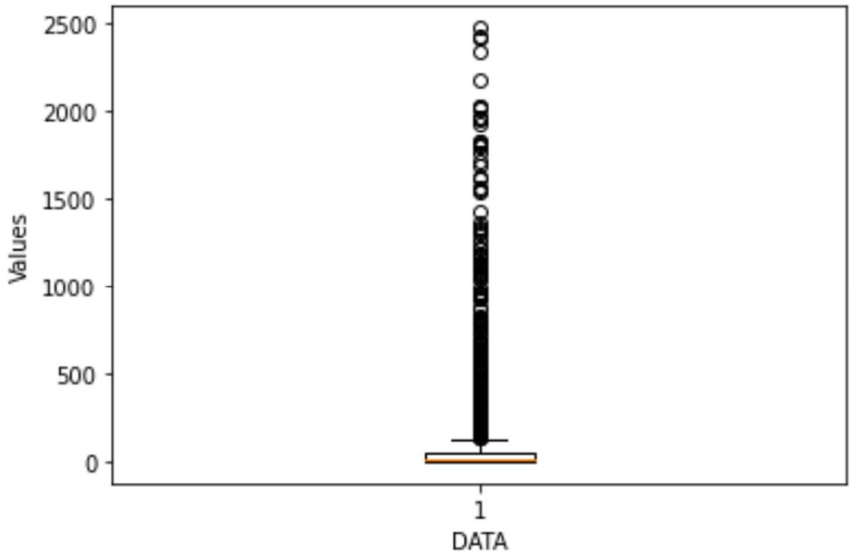


Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

**Inferences:**

1. There are 181 outliers.
2. IQR = q3 – q1 = 51.75 IQR and q3 are reduced by a huge extend after outlier values have been replaced with median values.
3. Range = 2470.5 - 0.0= 2470.5 and Variance = 157082.85. Both spread and variance have got reduced a lot.
4. The data is still right skewed or positively skewed as the orange line lies a lot below making it unsymmetric.

Thank you