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

Student's Name: Vivek jaiswal **Mobile No: 9887552039**

Roll Number: B20172 **Branch:DSE**

1

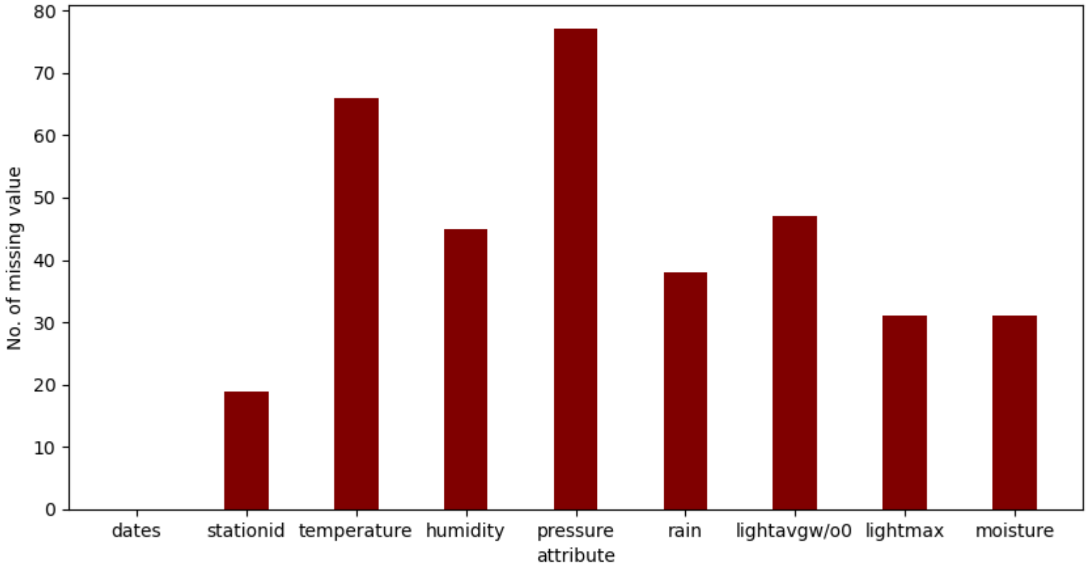

Figure 1 Number of missing values vs. attributes

Inferences:

1. Pressure has max missing and date has min missing .

2. The attribute 'dates' have no missing value while attributes 'lightmax' and 'moisture' have equal number of missing values around 30, 'stationid' has number of missing values around 20, 'temperature' has number of missing values around 65, 'humidity' has number of missing values around 45, 'pressure' has number of missing value 75, 'rain' has number of missing value 40 and 'lightavgw/o0' has number of missing value 50 .

2 a.

Inferences:

1. 19 tuples were deleted at this step.

2. 2 percent of the total number of tuples are deleted.

b.

Inferences:

1. 35 tuples were deleted this step.

2. 3.76 % of the total number of tuples are deleted.

3. The data lost wasn’t of high weightage as it contained no specific information because they had more than or equal to 1/3 missing values.

4. As these tuples are very less as compared to the total tuples in dataframe,so they have so many missing values dropping them is better optioon rather than predictig their values as it might be less accurate and more time consuming on the other hand dropping them entirely can increase accuracy of our model.

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. Pressure has max and stationid and dates have min missing value.
2. Pressure has max percent data missing 4.6%, dates and stationid have 0% missing values, lightmax has 0.12%, rain and moisture has 0.67%, temperature has 3.8%, lightavgw/o0 has 1.68% and humidity has 1.46% missing values .
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)	21.215	12.727	22.273	4.356	21.052	12.727	21.927	4.340
4	humidity (in g.m^{-3})	83.480	99.000	91.381	18.21	83.126	99	91	18.394
5	pressure (in mb)	1009.009	789.393	1014.678	46.98	1009.466	1009.466	1014.482	45.856
6	rain (in ml)	10701.538	0	18	24852.255	10798.379	0	15.75	24833.965
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.88	7573.163	4458.298	4488.910	1502.938	7606.284
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21463.221	4000	6569	21943.889
9	moisture (in %)	32.386	0	16.704	33.653	32.603	0	14.169	33.714

Inferences:

1. In the mean ‘lightmax’ has max change and ‘temperature’ has min change. In mode ‘pressure’ has max change and all other attributes have min change. In median ‘lightavgw/o0’ has maximum change and

‘pressure’ has minimum change and in standard deviation ‘lightmax’ has maximum change and ‘temperature’ has minimum change.

2. The attribute ‘lightmax’ having minimum missing values has maximum change in mean and standard deviation. The attribute ‘temperature’ having maximum missing values has minimum change in mean and standard deviation.

ii.

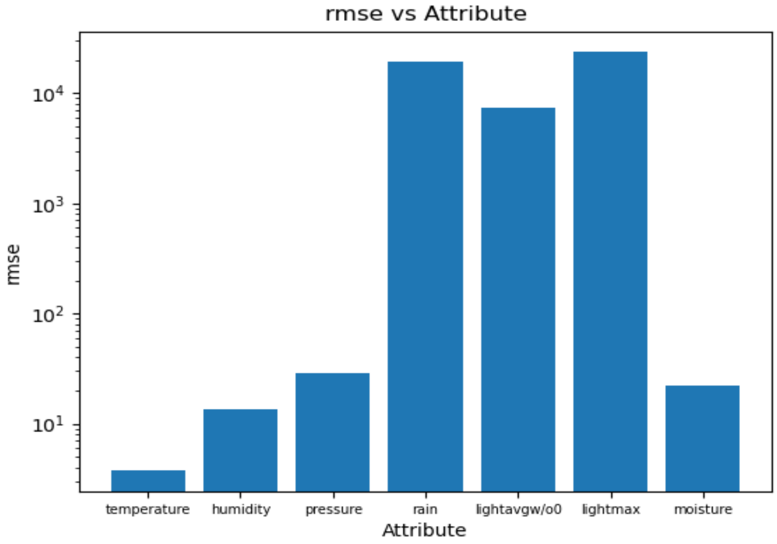

Figure 2 RMSE vs. attributes

Inferences:

1. The attributes ‘rain’ and ‘temperature’ have maximum and minimum rmse .

2. The attribute ‘temperature’ had maximum missing values and minimum change in mean and standard deviation and here it has minimum RMSE. The data is not reliable for further investigation as the values of RMSE for some a6ributes are quite high while ideally it shouldn’t be that high.

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)	21.215	12.727	22.273	4.356	21.115	12.727	22.140	4.399
4	humidity (in g.m^{-3})	83.480	99.000	91.381	18.21	83.166	99	91.180	18.408
5	pressure (in mb)	1009.009	789.393	1014.678	46.98	1009.968	789.393	1014.925	45.999
6	rain (in ml)	10701.538	0	18	24852.255	10727.959	0	15.750	24848.715
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.88	7573.163	4496.754	4488.91	1500.5	7649.458
8	lightmax (in lux)	21788.623	4000	6634	22064.993	21473.799	4000	6569	21946.161
9	moisture (in %)	32.386	0	16.704	33.653	32.529	0	13.894	33.791

Inferences:

1. In the mean ‘lightmax’ has maximum change and ‘temperature’ has minimum change. In mode there’s no change in before and after. In median ‘lightavgw/o0’ has maximum change and ‘temperature’ has minimum change and in standard deviation ‘lightmax’ has maximum change and ‘temperature’ has minimum change.

2. The a6ribute ‘lightmax’ having minimum missing values has maximum change in mean and standard deviation. The a6ribute ‘temperature’ having maximum missing values has minimum change in mean and standard deviation.

3. yes as the difference of mean, median, mode and standard deviation before and afterer is not abruptly high but still considerable high, the data cannot be considered reliable for further analysis.

4. From the observed changes in mean, mode, median and standard deviation, the difference in modes have become zero, the difference in standard deviation has significantly increased and for other a6ributes there isn’t much change after replacing missing values by linear interpolation as compared to when we replaced missing values by its mean.

ii.

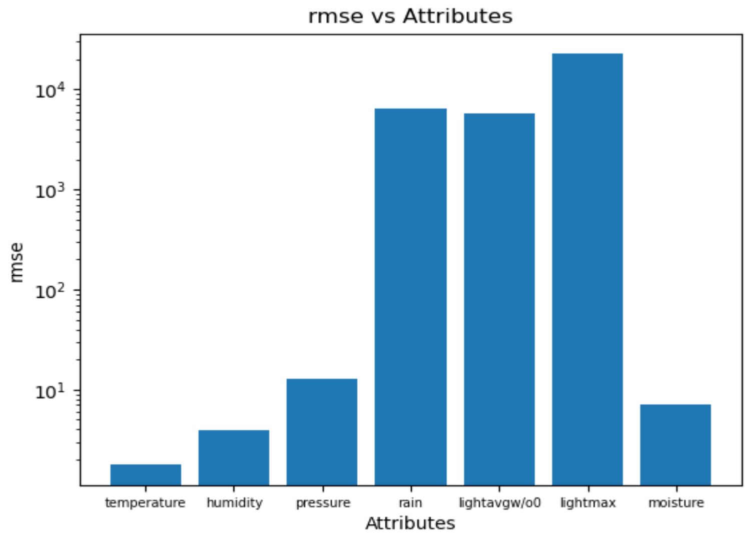

Figure 3 RMSE vs. attributes

Inferences:

1. The attributes lightavgw/o0 and lightmax have maximum and minimum rmse .
2. The attribute ‘lightmax’ had minimum missing values and maximum change in mean and standard deviation and here it has minimum RMSE.

3.The data is not reliable for further investigation as the values of RMSE for some a6ributes is still quite high while ideally it shouldn’t be that high.

4.The replaced values are more closer to the original ones when we replaced by interpolation as compared to when we replaced by mean.

.

5 a.

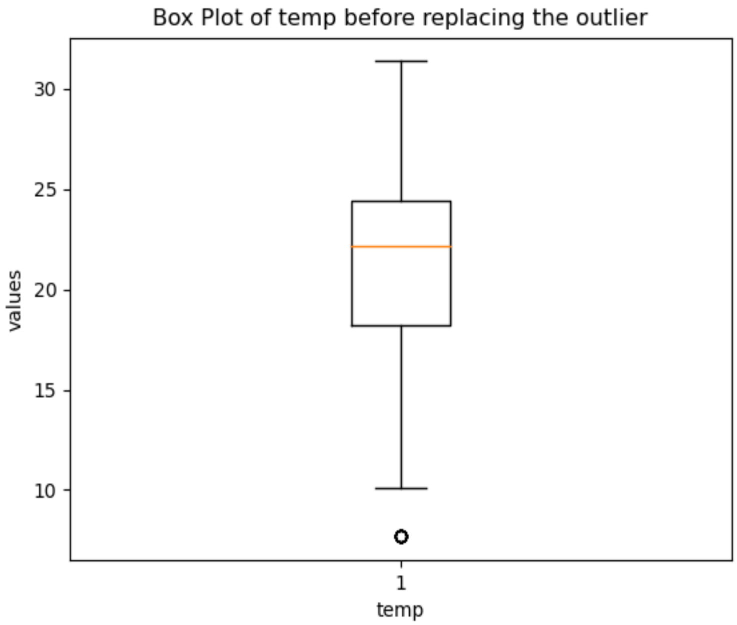

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers and their row numbers are 509, 510, 511, 512, 513, 514, 515, 516, 517, 518.

2. The Inter quartile range is around 6.

3. The spread is around 32 .

4.Data is positively skewed

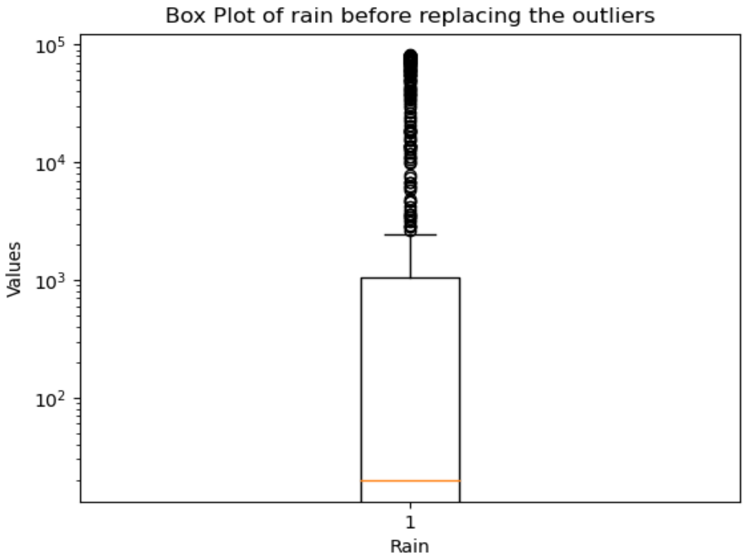

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. The number of outliers are 175 and their row numbers are 135, 199, 200, 201, 206, 322, 323, 324, 630, 631, 632, 636, 637, 638, 693, 694, 696, 697, 699, 702, 704, 705, 711, 742, 743, 744, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 785, 788, 789, 790, 791, 792, 793, 794, 795, 796, 798, 799, 800, 801, 802, 803, 825, 826, 827, 828, 829, 831, 835, 836, 840, 841, 842, 843, 846, 847, 851, 853, 854, 855, 856, 857, 858, 859, 862, 863, 864, 865, 866, 867, 868, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 923, 924, 925, 926, 927, 928, 929, 930, 931, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944.

2. The Inter quartile range is 10^3.

3. The spread is close to 10^5.

4. Data is positive skewed.

b.

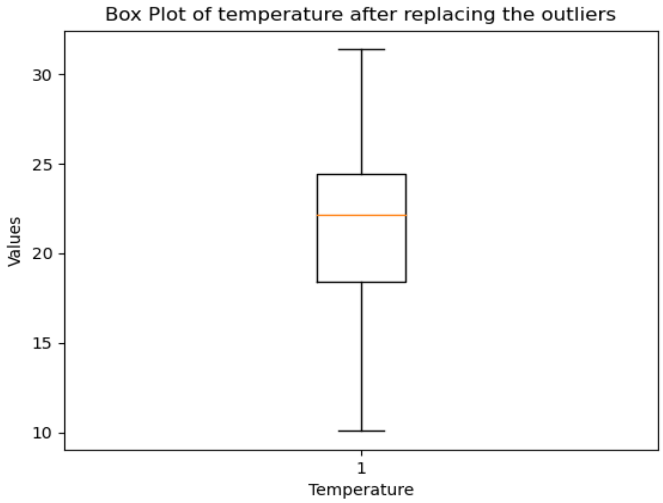

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers.
2. The Inter quartile range is around 6 which is almost same as before replacing the outliers.
3. The spread is around 25 which is less than the spread before replacing the outliers.
4. The data is positively skewed which as same as the skewness before replacing the outliers.

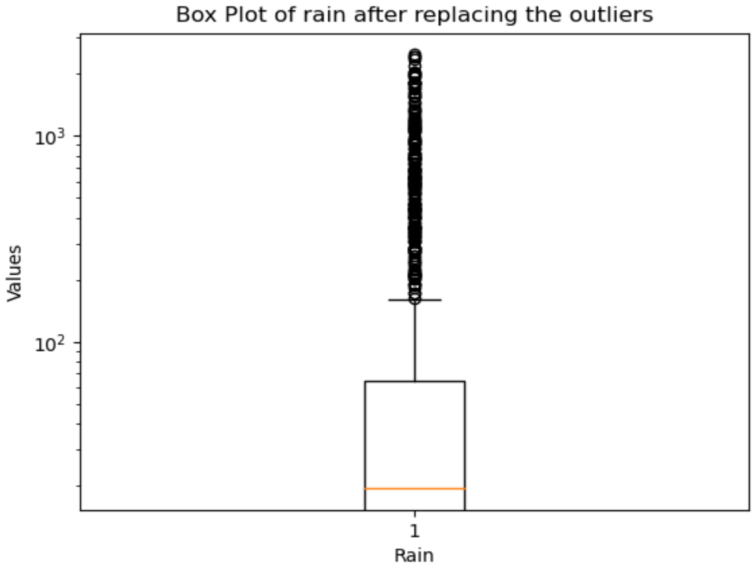

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. The number of outliers are 182 which are more than before replacing the outliers and their row numbers are 1, 2, 3, 4, 5, 11, 12, 13, 15, 16, 17, 20, 21, 23, 24, 25, 26, 27, 30, 31, 36, 38, 39, 40, 41, 43, 44, 48, 51, 53, 56, 60, 62, 70, 71, 72, 73, 90, 141, 142, 144, 145, 149, 154, 198, 202, 203, 204, 205, 207, 208, 209, 213, 218, 219, 227, 229, 230, 231, 232, 235, 237, 238, 239, 246, 248, 250, 265, 321, 325, 328, 377, 381, 382, 384, 385, 388, 389, 393, 394, 395, 397, 399, 400, 401, 409, 411, 412, 413, 419, 426, 428, 432, 442, 448, 452, 455, 464, 467, 470, 484, 489, 496, 507, 522, 523, 525, 526, 527, 528, 529, 533, 534, 535, 536, 550, 561, 633, 634, 641, 669, 670, 671, 672, 673, 676, 680, 681, 685, 689, 691, 698, 700, 701, 707, 718, 719, 720, 721, 722, 724, 727, 728, 729, 730, 732, 734, 735, 736, 739, 740, 745, 746, 747, 786, 787, 797, 812, 814, 818, 819, 820, 821, 822, 823, 824, 830, 832, 833, 834, 838, 839, 844, 845, 849, 850, 852, 881, 882, 921, 922, 932

2. The Inter quartile range is 50 which is less than as compared to before replacing the outliers.

3. The spread is close to 10^5 which is almost same as before replacing the outliers.

4. The data is more positive skewed as compared to the data before replacing the outliers.