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

Student's Name: Vishal Sharma

Mobile No: 9540140310

Roll Number: B20239

Branch: Data Science and Engineering

1

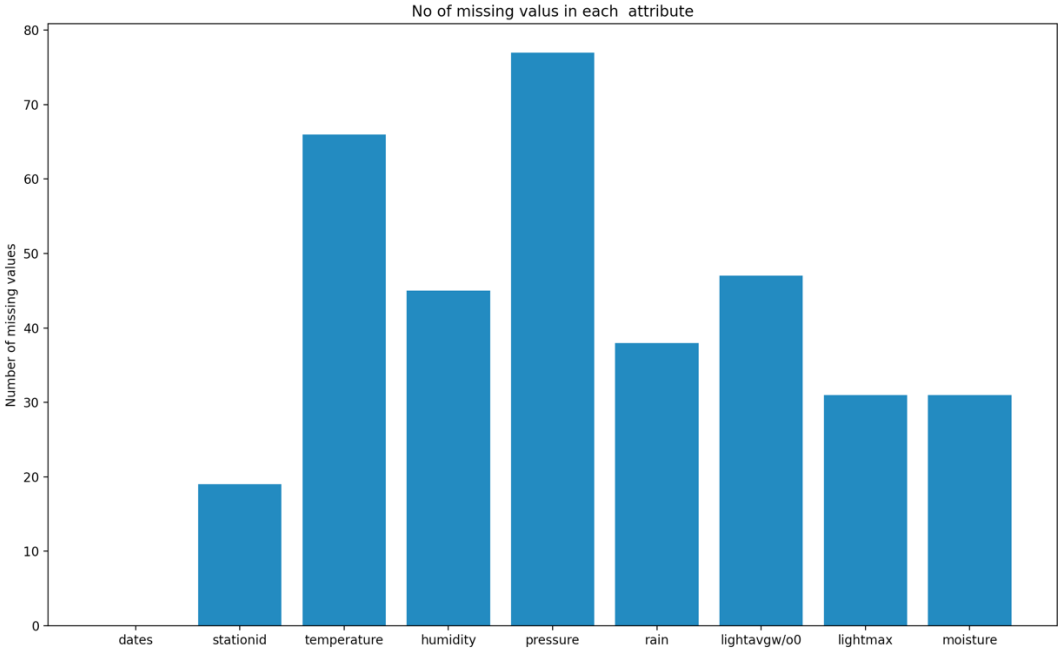

Figure 1 Number of missing values vs. attributes

Inferences:

1. From the above plot, Pressure has maximum number of missing values, while dates have minimum number of missing values.
2. Dates have no missing values, thus have a zero frequency. Attributes 'lightmax' and 'mositure' have same frequency, which is around 30. Attribute 'rain' has a frequency of about 40, while 'stationid' has a frequency of only 20. 'Humidity' have a frequency of around 45, 'temperature' at around 65, while 'pressure' has maximum frequency, around 76.

2 a.

Inferences:

1. Among those attributes which have missing values, 'stationid' has the minimum number of missing values. Also, it is not a numeric data.
2. Around 19 tuples were deleted.
3. The percentage of tuples deleted is roughly 2%.

b.

Inferences:

1. 35 tuples were deleted after this step.
2. The percentage of tuples deleted is roughly around 3.77%.
3. The data lost in this step couldn't provide us with much information. It is because more than 1/3rd of the attributes in a tuple didn't have any value.
4. Dropping these tuples were justified because many attributes didn't have any value in them. And even if we have guessed the value and filled the cells, there might be a possibility that the data may deviate quite a lot from the original one

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 ‘dates’ and ‘stationid’ have minimum number of missing values (0 missing values), while ‘pressure’ has maximum number of missing values (41 missing values).
2. Attribute ‘pressure’ has around 4.6% of missing values, ‘temperature’ has around 3.8%, ‘humidity’ has around 1.4%, ‘rain’ and ‘moisture’ have around 0.6%, ‘lightavw/o0’ has around 1.8%, ‘lightmax’ has around 0.1% of missing values respectively.
3. The total number of missing values is 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	Median	Mode	S.D.	Mean	Median	Mode	S.D.
1	temperature (in °C)	21.052	21.927	21.052	4.339	21.097	22.139	12.727	4.398
2	humidity (in $g.m^{-3}$)	83.125	91.0	99.0	18.393	83.122	91.179	99.0	18.413
3	pressure (in mb)	1009.465	1014.482	1009.465	45.855	1010.051	1014.925	789.392	46.074
4	rain (in ml)	10798.37	15.75	0	24833.9	10727.43	15.75	0	24848.9
5	lightavgw/o0 (in lux)	4458.29	1502.93	4488.91	7606.28	4442.74	1464.62	4488.91	7610.9
6	lightmax (in lux)	21463.79	6569	4000	21943.8	21473.79	6569	4000	21946.1
7	moisture (in %)	32.6	14.16	0	33.71	32.57	13.89	0	33.85

Inferences:

1. For mean, attribute ‘temperature’ has minimum change in mean, while attribute ‘rain’ have maximum change. For median, attributes ‘lightmax’ and ‘rain’ have minimum change, while attribute ‘lightavgw/o0’ has a maximum change. For mode, attributes ‘moisture’, ‘lightmax’, ‘lightavgw/o0’, ‘rain’ and ‘humidity’ has minimum change. While attribute ‘pressure’ has a maximum change. For Standard Deviation, attribute ‘humidity’ has the minimum change, while attribute ‘rain’ has maximum change.
2. The attribute ‘temperature have maximum number of missing values, has a minimum change in its mean.
3. As the difference is huge in 3 attributes, the data is not so reliable.

ii.

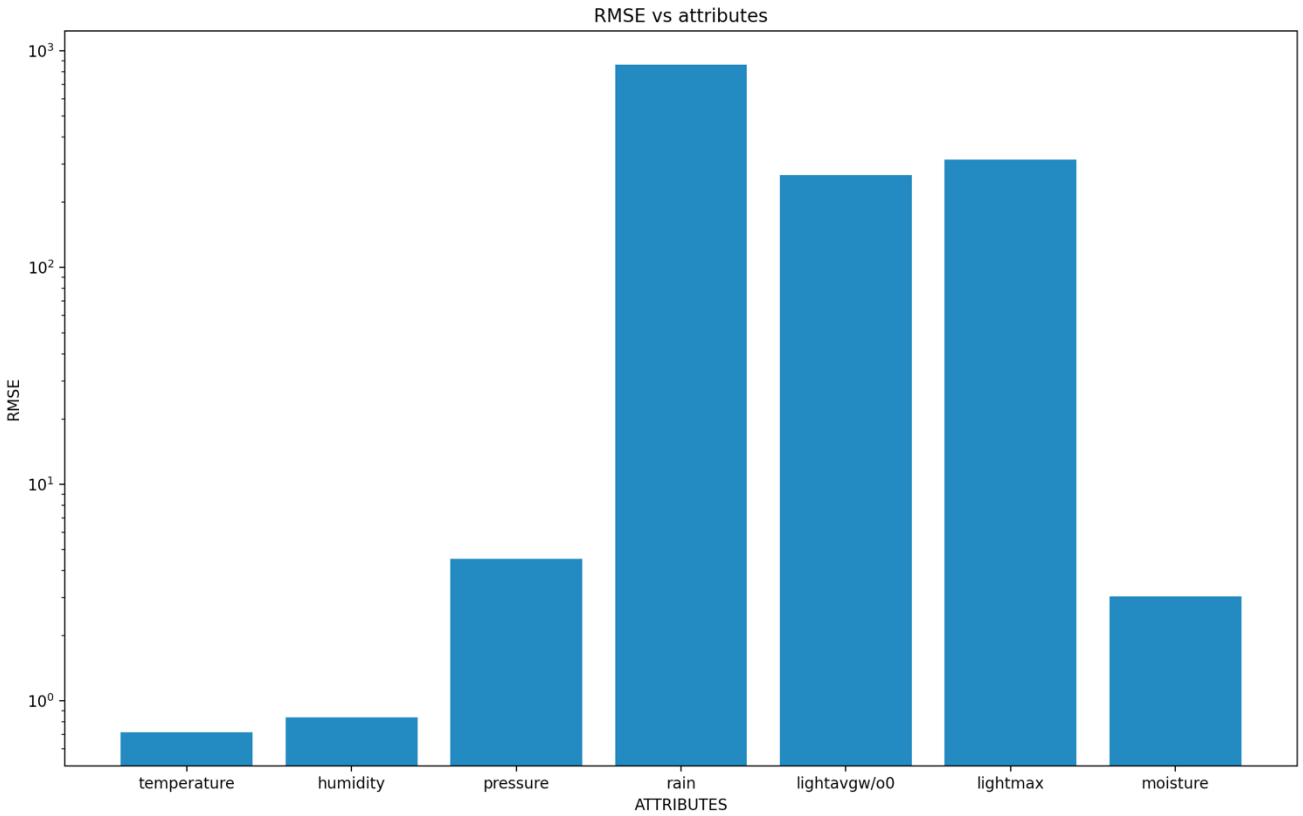

Figure 2 RMSE vs. attributes

Inferences:

1. Attributes 'temperature' and 'rain' have minimum and maximum RMSE values respectively.
2. The attribute 'temperature' has maximum number of missing values, but the change in mean and its RMSE is minimum.
3. The data is not reliable as many attributes have a high value of RMSE, ideally which should be minimum.

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	Median	Mode	S.D.	Mean	Median	Mode	S.D.
1	temperature (in °C)	21.05	21.92	21.05	4.33	21.09	22.13	12.72	4.39
2	humidity (in $g.m^{-3}$)	83.12	91	99	18.39	83.12	91.17	99	18.41
3	pressure (in mb)	1009.46	1014.48	1009.46	45.85	1010.05	1014.92	789.39	46.07
4	rain (in ml)	10798.3	15.75	0	24833.9	10727.4	15.75	0	24848.9
5	lightavgw/o0 (in lux)	4458.29	1502.93	4488.91	7606.28	4442.74	1464.62	4488.9	7606.28
6	lightmax (in lux)	21463.2	6569	4000	21943.8	21473.7	6569	4000	21943.8
7	moisture (in %)	32.6	14.16	0	33.71	32.57	13.89	0	33.85

Inferences:

1. For mean, attribute ‘humidity’ has a minimum change, while attribute ‘rain’ has a maximum change. For median, attributes ‘rain’ and ‘lightmax’ have a minimum change, while attribute ‘lightavgw/o0’ has a maximum change. For mode, attributes ‘humidity’, ‘rain’, ‘lightmax’ and ‘moisture’ have a minimum change, while attribute ‘pressure’ has a maximum change. For standard deviation, attributes ‘lightavgw/o0’ and ‘lightmax’ have a minimum change, while attribute ‘rain’ has a maximum change.
2. The attribute ‘lightmax’ have minimum number of missing values and also have minimum change in median, mode, and standard deviation.
3. Thought the change in many attributes is quite less, but the change in some attributes is still high, thus the new data is not so much reliable.
4. From the changes made, change in mode and standard deviation is quite low, though some have a large difference.

ii.

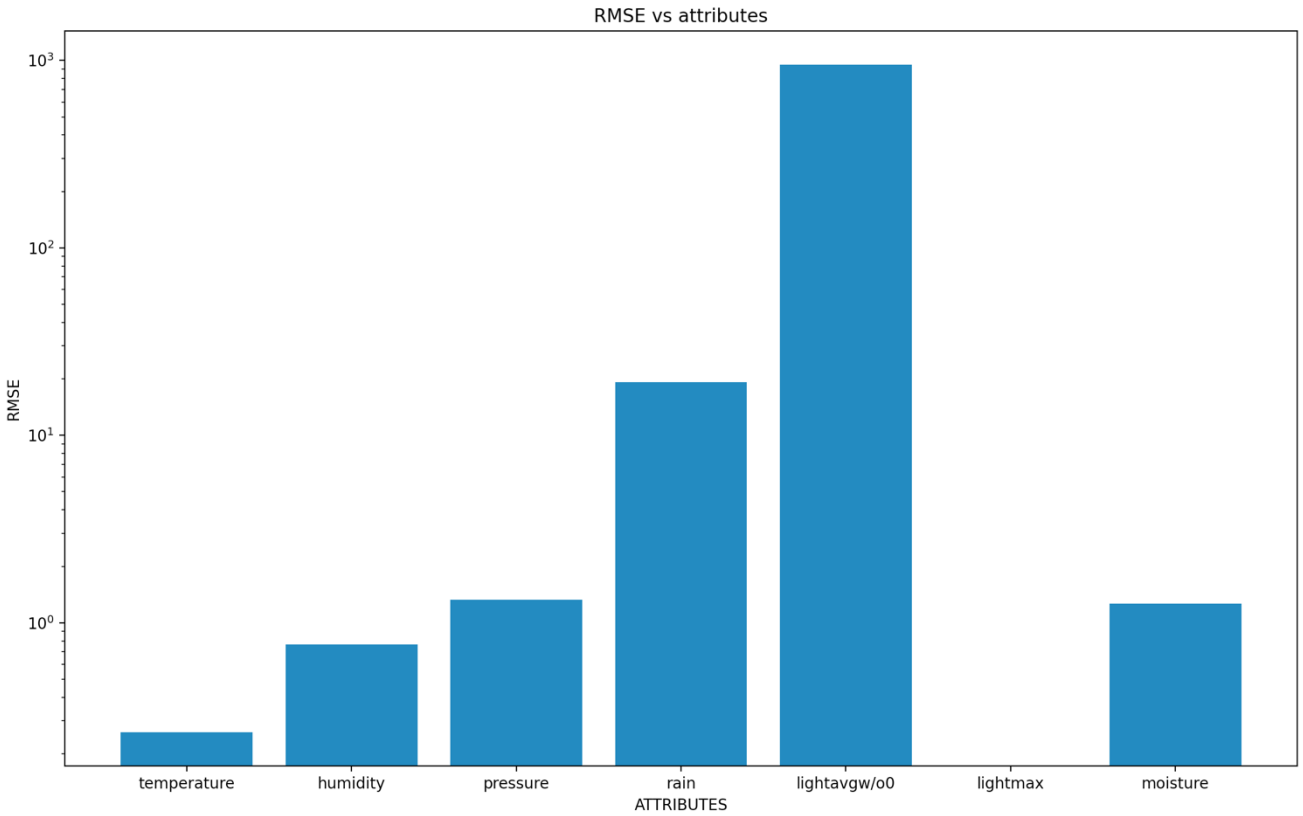

Figure 3 RMSE vs. attributes

Inferences:

1. The attributes 'lightavgw/o0' and 'lightmax' have maximum and minimum RMSE values respectively.
2. The attribute 'lightmax' has minimum number of missing values and has minimum change in median, mode and standard deviation, and its RMSE value is also zero.
3. The data is not so reliable as some attributes still have a high RMSE value.
4. The replaced value through interpolation is much closer than replacing by mean.

5 a.

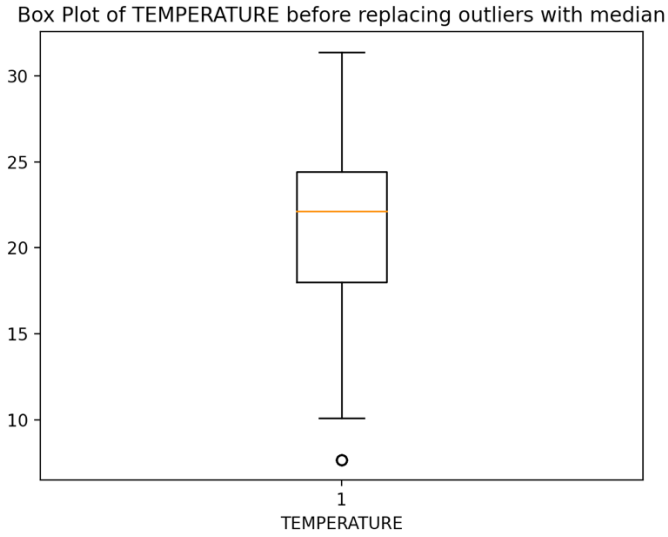

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers, and their row numbers are 462, 463, 464, 465, 466, 467, 468, 469, 470, 471
2. The inter quartile range (IQR) is around 6.
3. The spread is around 32
4. Data is positively skewed.

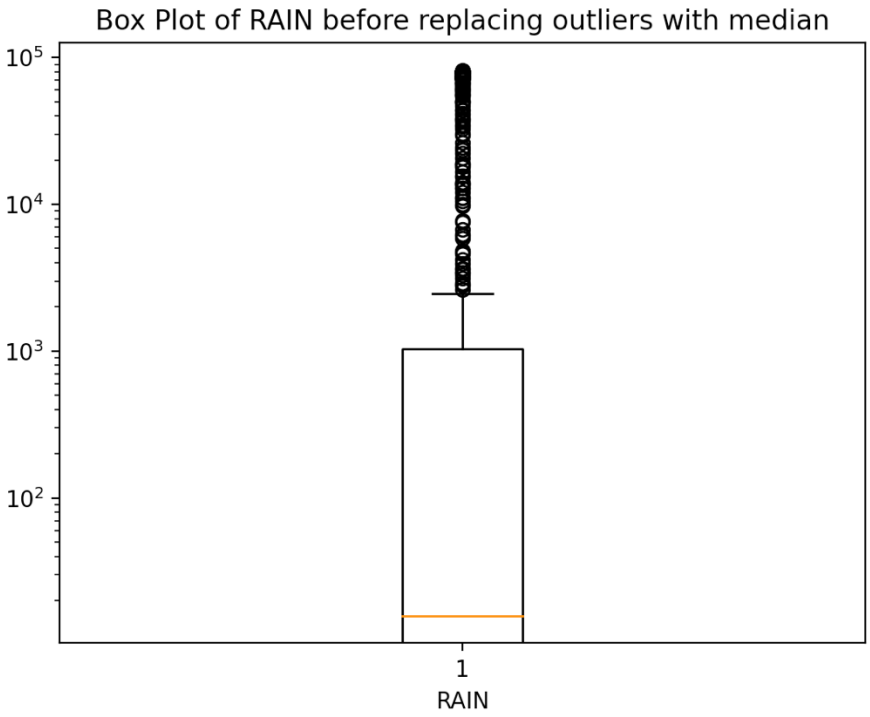

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. There are 175 outliers and their 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. The inter quartile range is around 10^3.
3. The spread is around 10^5.
4. The data is negatively skewed.

b.

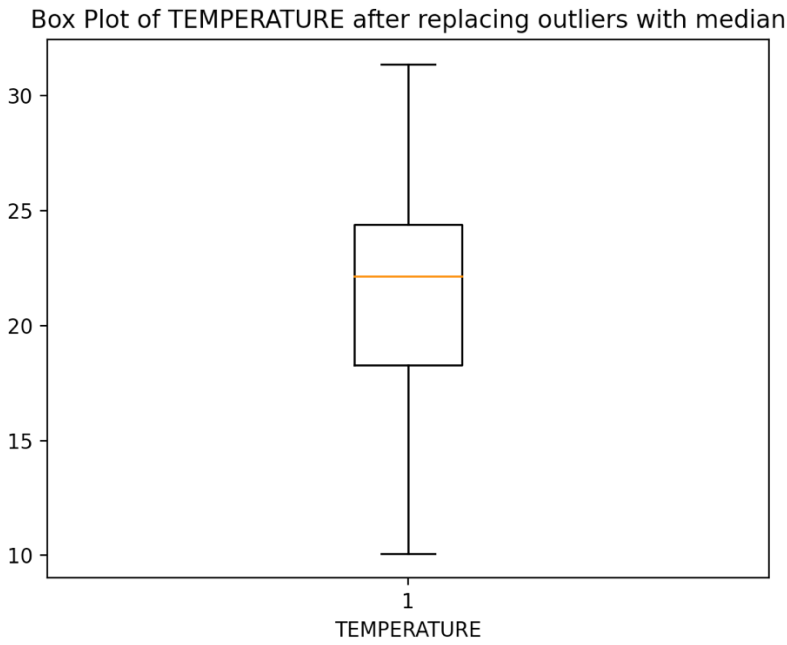

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers.
2. The inter quartile range is around 6 which is almost same as the previous value.
3. The spread is around 25 which is less than the previous value.
4. It is positively skewed like before.

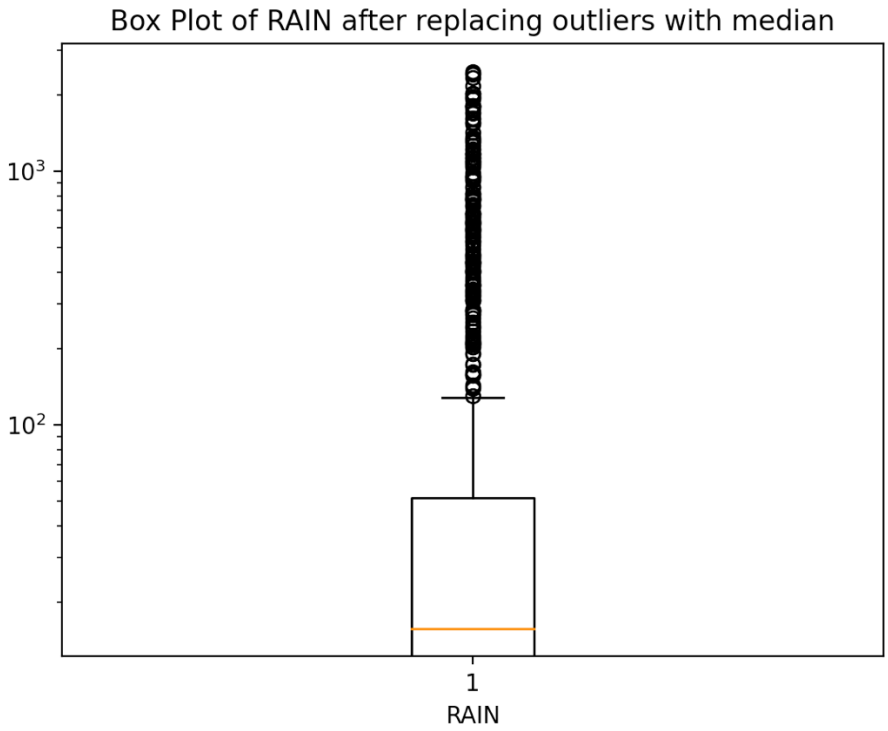

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. There are 182 outliers and their row numbers are 1, 2, 3, 4, 5, 10, 11, 12, 13, 14, 15, 18, 19, 21, 22, 23, 24, 25, 28, 29, 33, 34, 35, 36, 37, 38, 39, 43, 44, 46, 49, 53, 54, 60, 61, 62, 63, 79, 127, 128, 130, 131, 135, 140, 182, 186, 187, 188, 189, 191, 192, 193, 197, 201, 202, 210, 212, 213, 214, 215, 218, 220, 221, 222, 228, 230, 232, 245, 299, 303, 306, 350, 354, 355, 356, 357, 360, 361, 365, 366, 367, 369, 371, 372, 373, 380, 382, 383, 384, 387, 394, 396, 400, 410, 416, 420, 423, 424, 426, 429, 437, 442, 449, 460, 475, 476, 478, 479, 480, 481, 482, 486, 487, 488, 489, 503, 514, 586, 587, 594, 622, 623, 624, 625, 626, 629, 633, 634, 638, 642, 644, 651, 653, 654, 660, 671, 672, 673, 674, 675, 677, 680, 681, 682, 683, 684, 685, 686, 687, 689, 690, 694, 695, 696, 733, 734, 744, 759, 761, 765, 766, 767, 768, 769, 770, 771, 777, 779, 780, 781, 785, 786, 791, 792, 796, 797, 799, 827, 828, 867, 868, 878.
2. The inter quartile range is around 52 which is very less than the previous value.
3. The spread is close to 10^5 which is almost same as the previous value.
4. The data is negatively skewed.