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

Student's Name: Kuldeep Jain Dugar **Mobile No: 8986388665**

Roll Number: B20112 **Branch:**CSE

1

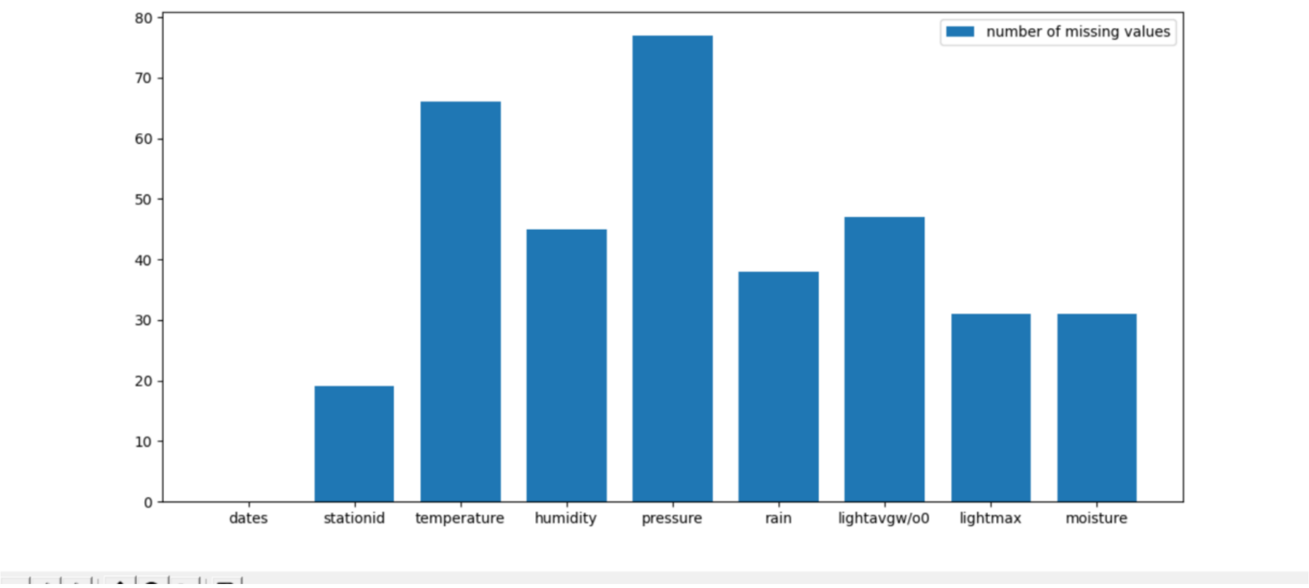

Figure 1 Number of missing values vs. attributes

Inferences:

1. Max – Pressure Min- Stationid
2. stationid 19
3. temperature 66
4. humidity 45
5. pressure 77
6. rain 38
7. lightavgw/o0 47
8. lightmax 31
9. moisture 31

2 a.

Inferences:

1. It is a categorical attribute and we cant replace it with mean/median/mode
2. State the number of tuples deleted after this step.- 19
3. percentage of the total number of tuples is deleted - 2%

b.

Inferences:

1. State the number of tuples deleted after this step.- 19
2. What percentage of the total number of tuples is deleted?2.05%
   Tuples having more than 1/3 values as NaN were elimiated

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)	48
4	humidity (in $g.m^{-3}$)	27
5	pressure (in mb)	59
6	rain (in ml)	19
7	lightavgw/o0 (in lux)	29
8	lightmax (in lux)	13
9	moisture (in %)	16

Inferences:

maximum – Temperature and minimum – Moisture/ stationid/dates

1. For each attribute, comment on the percentage of data missing.-

2. temperature 5.2%
3. humidity 2.9%
4. pressure 6.4%
5. rain 2%
6. lightavgw/o0 3.1%
7. lightmax 1.4%
8. moisture 1.7%

9. State the total number of missing attributes in the file.- 211

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	21.05	21.05	21.89	4.29
4	humidity (in $g.m^{-3}$)	83.47	99	91.38	18.21	83.18	99	90.5	18.21
5	pressure (in mb)	1009	789.39	1014.67	46.98	1009	1014	1009	45.47
6	rain (in ml)	10701.5	0	18	24852	11080	0	18.0	24978
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.88	7573.162	4448.52	4488.91	1695.43	7521.80
8	lightmax (in lux)	21788.62	4000	6634.0	22064.993	21587.28	4000.0	6607.0	21847.76
9	moisture (in %)	32	0.0	16.70	33.65	32.49	0	15.7	33.533.65

Inferences:

Mean	Mode	Median	S.D.
Max rain	**pressure**	lightmax (in lux)	lightmax (in lux)
Min- pressure	Rain moisture (many)	rain	humidity

1. The attributes with most missing values must have more max difference
2. The difference is less so it can be trusted

ii.

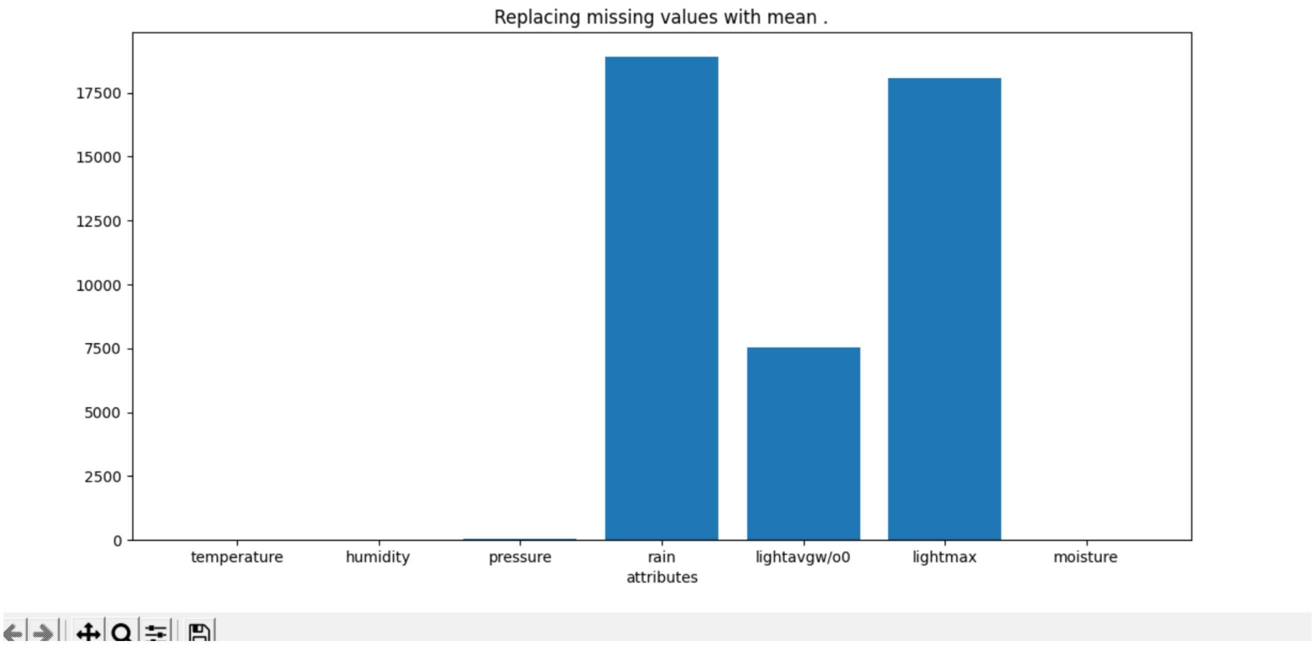

Figure 2 RMSE vs. attributes

Inferences:

1. Rain has the maximum RMSE.

2. The attributes with most missing values must have more max difference
3. Some attributes have very less error which implies data is reliable in the other hand some have high RMSE values

b. i.

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)	21.214	12.72	22.2	4.35	21.15	12.72	22.17	4.36
4	humidity (in g.m^{-3})	83.47	99	91.38	18.21	83.19	99	91.14	18.35
5	pressure (in mb)	1009.0	789.39	1014	46.98	1009.70	789.34	1014.29	45.77
6	rain (in ml)	10701.53	0	18	24852.25	10951.43	0	18	25125.15
7	lightavgw/o0 (in lux)	4438.428	4488.91	1656.88	7573.162	4497.14	4488.91	1579.85	7604.61
8	lightmax (in lux)	21788.623	4000.	6634.0	22064.99	21577.28	4000.	6569.0	21971.76
9	moisture (in %)	32.38	0	16.704	33.65	32.49	0	14.25	33.65

Inferences:

1. Which attributes have the maximum and the minimum change in the mean, mode, median and standard deviation respectively?

Mean	**Mode**	**Median**	**S.D.**
Max rain	**pressure**	lightmax (in lux)	lightmax (in lux)
Min- humidity	Rain moisture (many)	rain	temperature

2. The attributes with most missing values must have more max difference
3. From the change observed in mean, mode, median and standard deviation ponder is the data reliable for further investigation or experimental analyses- There is very less difference in most of the attributes.
4. Interpolation is a better method to replace the value

ii.

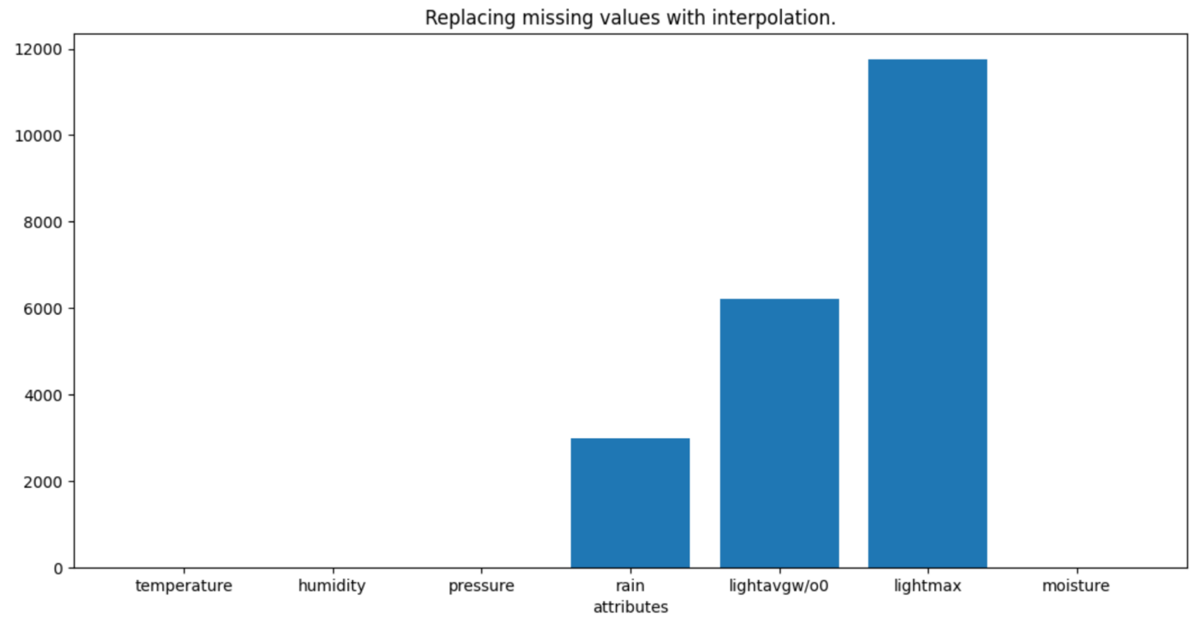

Figure 3 RMSE vs. attributes

Inferences:

1. Max=lightmax min- temperature
2. The attributes with most missing values must have more max difference
3. Some attributes have very less error which implies data is reliable in the other hand some have high RMSE values
4. From the calculated RMSE compare and contrast replacing missing values by mean and linear interpolation technique.-

	Mean	Interpolation
temperature (in °C)	3.91	1.66
humidity (in g.m^{-3})	12.86	4.69
pressure (in mb)	31.79	14.39
rain (in ml)	18921.42	3004.14
lightavgw/o0 (in lux)	7514.49	6209
lightmax (in lux)	18070	11762
moisture (in %)	24.75	9.59

5 a.

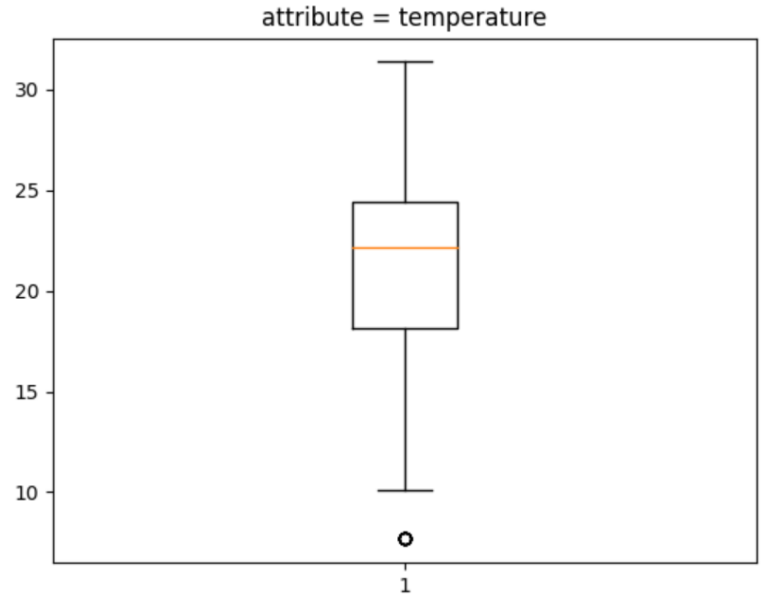

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. No. of quartile = 10
2. Rows -[509, 510, 511, 512, 513, 514, 515, 516, 517, 518]
   Outliers -[7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729, 7.6729]
3. IQR - 6.24
4. Infer the spread/variance.- Less outliers so less spreaded
5. Infer the skewness of the data.- Positive

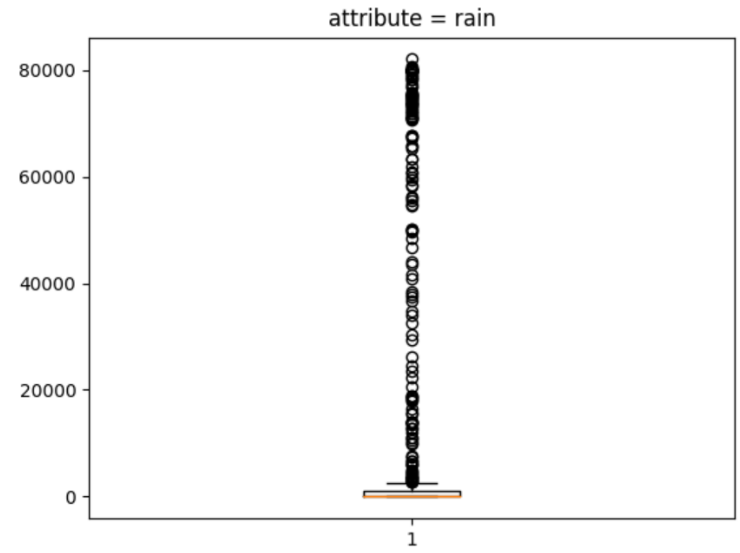

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. List the number of outliers and their row numbers.- 181
   Row Numbers
   [135, 136, 199, 200, 201, 206, 322, 323, 324, 367, 368, 369, 370, 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, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 785, 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, 869, 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]
   outliers of rain
   [13583.25, 6791.625, 15459.75, 14001.75, 16571.25, 13666.5, 59982.75, 80000.0, 75048.75, 80000.0, 80000.0, 80000.0, 80000.0, 3930.5, 36636.75, 40789.0, 63256.5, 54616.5, 50172.75,

37928.25, 26178.75, 3138.75, 3449.25, 18884.25, 9765.0, 18976.5, 30393.0, 2814.75, 80000.0, 82037.25, 56319.75, 71968.5, 80000.0, 80000.0, 50242.5, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 60675.75, 22250.25, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 80000.0, 37392.75, 49725.0, 80000.0, 80000.0, 71154.0, 80000.0, 80000.0, 12854.25, 34879.5, 4610.25, 6210.0, 10557.0, 3451.5, 3312.0, 18285.75, 3613.5, 2893.5, 23474.25, 14042.25, 3647.25, 5877.0, 10062.0, 17997.75, 29517.75, 32514.75, 13943.25, 4212.0, 4691.25, 7519.5, 11112.75, 2821.5, 33941.25, 43643.25, 20664.0, 11144.25, 18587.25, 18373.5, 15646.5, 12915.0, 49916.25, 24522.75, 75105.0, 73417.5, 70580.25, 78126.75, 56097.0, 6061.5, 38355.75, 55509.75, 43974.0, 6747.75, 54843.75, 59377.5, 58320.0, 60963.75, 63342.0, 67378.5, 70929.0, 73158.75, 71367.75, 73838.25, 46732.5, 48429.0, 67830.75, 75447.0, 74646.0, 75402.0, 75723.75, 74254.5, 75201.75, 77044.5, 74472.75, 77503.5, 78180.75, 79915.5, 80583.75, 80482.5, 79337.25, 79317.0, 70823.25, 75638.25, 73752.75, 65893.5, 72774.0, 7773.75, 12037.5, 79839.0, 78633.0, 78779.25, 76662.0, 67252.5, 74913.75, 4869.0, 41618.25, 58443.75, 74173.5, 72445.5, 65873.25, 67675.5, 61989.75, 71237.25, 73577.25, 65301.75, 73534.5, 72283.5, 71799.75]

2. Infer the Inter quartile range.- 1072.125
3. Infer the spread/variance.- more outliers more spreaded
4. Infer the skewness of the data.- Negatively

b.

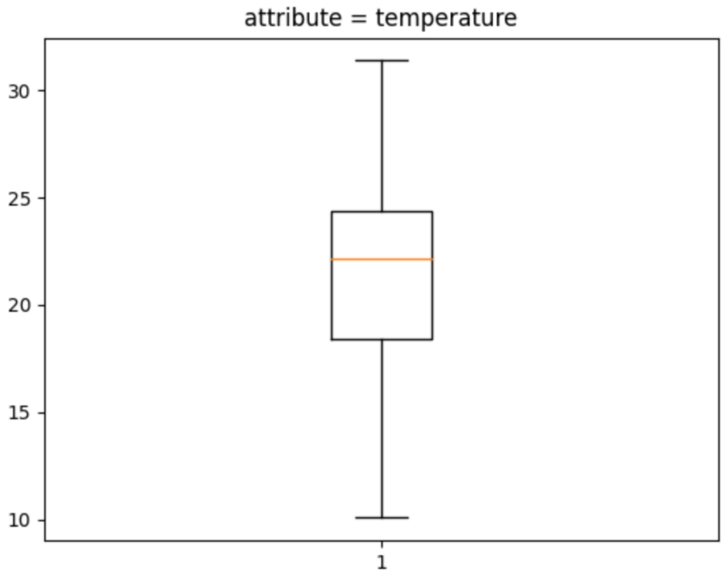

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. List the number of outliers =0 , Previously it was 10
2. Infer the Inter quartile range = 6.01
3. Infer the spread/variance- less spreaded
4. Infer the skewness of the data compare with Q5. a. Positive vs Positive

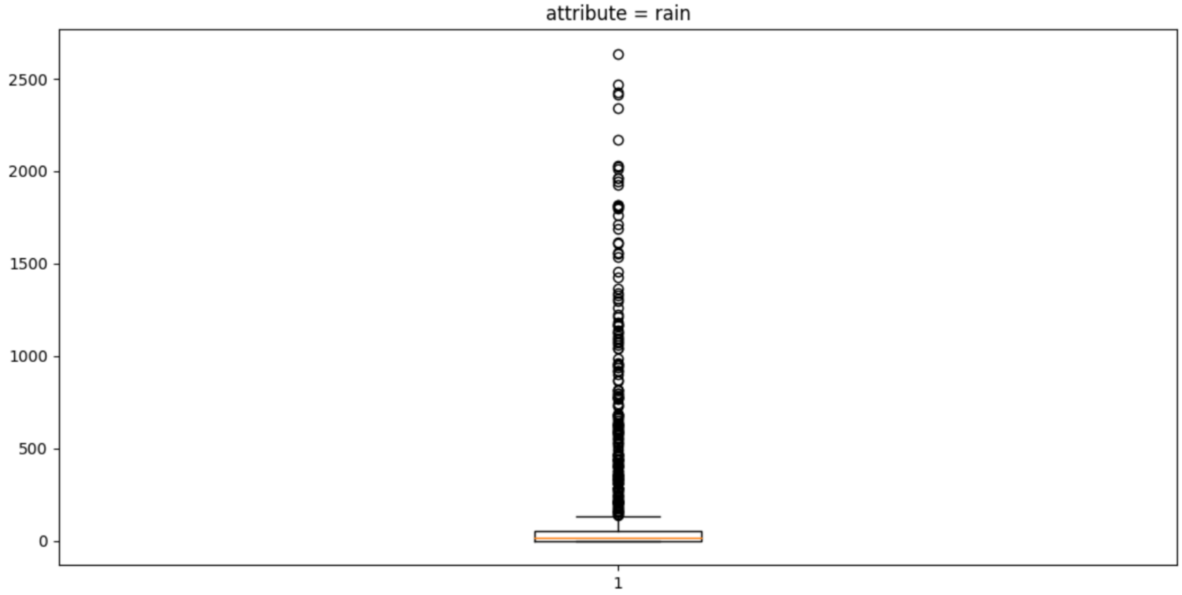

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. List the number of outliers = 187, their row number and compare with Q5.= previously it was 181
   [1, 2, 3, 4, 5, 11, 12, 13, 15, 16, 17, 20, 21, 23, 24, 25, 26, 27, 30, 31, 36, 37, 38, 39, 40, 41, 42, 43, 44, 48, 49, 50, 51, 53, 56, 60, 62, 63, 70, 71, 72, 73, 90, 141, 142, 144, 145, 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, 788, 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]
   outliers of rain
   [1761.75, 652.5, 963.0, 254.25, 339.75, 607.5, 560.25, 513.0, 474.75, 817.875, 1161.0, 240.75, 398.25, 816.75, 776.25, 681.75, 441.0, 274.5, 1341.0, 1804.5, 2171.25, 1456.875, 742.5, 443.25, 774.0, 1167.75, 898.875, 630.0, 594.0, 546.75, 576.0, 605.25, 634.5, 1091.25, 162.0, 157.5, 366.75, 183.375, 589.5, 207.0, 281.25, 1215.0, 315.0, 1260.0, 324.0, 360.0, 679.5, 159.75, 1710.0, 1183.5, 1962.0, 1071.0, 438.75, 864.0, 816.75, 796.5, 191.25, 202.5, 1611.0, 353.25, 533.25, 213.75, 434.25, 191.25, 202.5, 594.0, 409.5, 139.5, 333.0, 468.0, 222.75, 263.25, 459.0, 158.0, 272.25, 621.0, 587.25,

468.0, 778.5, 987.75, 623.25, 330.75, 1075.5, 308.25, 337.5, 1617.75, 144.0, 402.75, 2414.25, 1044.0, 211.5, 285.75, 400.5, 1426.5, 209.25, 551.25, 344.25, 1140.75, 357.75, 308.25, 774.0, 207.0, 1172.25, 427.5, 531.0, 1311.75, 247.5, 454.5, 283.5, 1062.0, 1554.75, 569.25, 357.75, 1795.5, 382.5, 353.25, 918.0, 677.25, 1689.75, 141.75, 213.75, 637.5, 2470.5, 580.5, 951.75, 281.25, 684.0, 463.5, 420.75, 1329.75, 173.25, 211.5, 173.25, 1300.5, 326.25, 621.0, 1818.0, 783.0, 949.5, 438.75, 1559.25, 1039.5, 405.0, 582.75, 234.0, 666.0, 625.5, 1365.75, 1129.5, 524.25, 492.75, 920.25, 218.25, 2022.75, 2009.25, 438.75, 285.75, 225.0, 1809.0, 1226.25, 2637.0, 1964.25, 321.75, 688.5, 765.0, 1125.0, 868.5, 1107.0, 405.0, 731.25, 157.5, 794.25, 1536.75, 954.0, 731.25, 1926.0, 1818.0, 243.0, 373.5, 308.25, 936.0, 2029.5, 661.5, 1946.25, 1095.75, 2340.0, 2427.75]

2. Infer the Inter quartile range- 54.0
3. Infer the spread/variance – have many outliers still less spreaded than earlier.
4. Infer the skewness – negative .

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and title.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places**
- **The quantities which have units should be written with units.**