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

Student's Name: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: CSE**

1

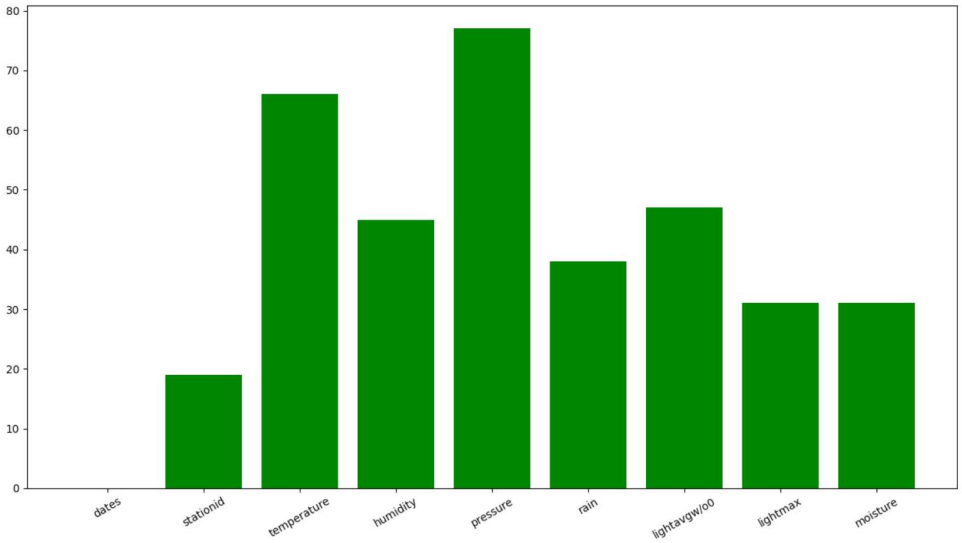

Figure 1 Number of missing values vs. attributes

Inferences:

1. Atmospheric pressure and date has maximum and minimum data values missing respectively.
2. 68-78 values of pressure and temperature data, while some around 40 values are missing of humidity, rain, lightavgw/o0 and lightmax and around 20 values are missing of stationid and no values is missing of date.
3. One have to see the data collection, transfer technique to measure atmospheric pressure and temperature, since most data is missing here, so these techniques should be improved.

2 a.

Inferences:

1. We decided to erase the tuple when stationid is missing in light of the fact that without it we will not have the option to know from which area the data are. So it's silly to know data when we don't have the idea about the area of data.
2. Number of tuples we have to delete are 19.

3. A total about of 2 percentage of data is missing.

b.

Inferences:

1. There are total 35 data deleted.
2. Hence total about 4 percentage data is missing.
3. 4 percent is a big number in terms of loss.
4. Since If this were not to done then one cannot analyse the data and ultimately we can not get any inferences from the data as it reduces the statistical power of the data.
5. We can see that a large amount of data is lost in cleaning process, hence one can never say that his/her analysis is the best 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. Atmospheric pressure has max while dates had minimum missing values.
2. Dates and stationid has no percentage of data missing, similarly rain, lightmax and moisture has very less data missing about 0.03 percentage data missing, humidity and lightavgw/o0 has there about 1.23 percentage data missing and temperature and pressure has their 4-4.23 percentage data missing.
3. There are total 116 data values missing now.

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	temperature (in °C)	21.05	21.05	21.79	4.26	21.21	12.727	22.27	4.355
2	humidity (in $g.m^{-3}$)	83.184	99.0	90.17	18.08	83.48	99.0	91.38	18.21
3	pressure (in mb)	1009.34	1009.35	1014.09	45.126	1009	789.4	1014.6	47
4	rain (in ml)	11080.66	0	22.500	24878.305	10701.53	0	18	24852.25
5	lightavgw/o0 (in lux)	4448.52	4488.9103	1838.02	7464.66	4438.42	4488.9103	1656.9	7573.16
6	lightmax (in lux)	21587.28	4000	7186.000	21681.8	21788.6	4000	6634	22064
7	moisture (in %)	32.5	0	17.70	33.285	32.38	0	16.7	33.65

Inferences:

1. All the data are quite similar only the remarkable changes in mode data of temperature and atmospheric pressure can be seen.

2. There is slight changes In the statistical data.
3. For other attributes mode data is exactly same but for atmospheric pressure data there is remarkable difference we can see.
4. Data analysis by changing the missing values with the mean value can be done as it gives approximate inference.

ii

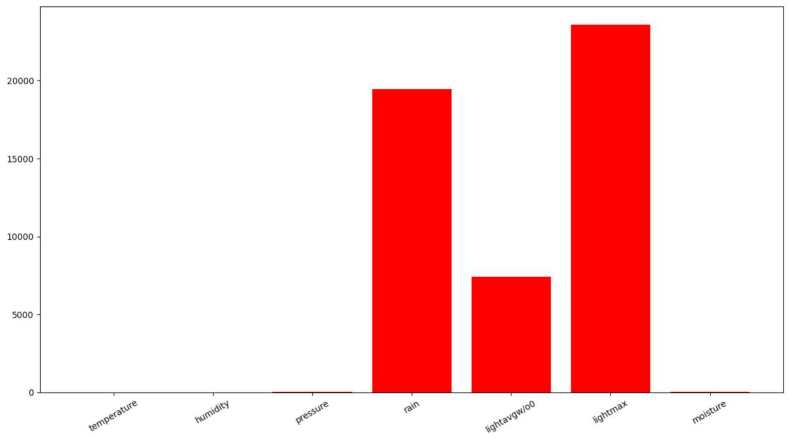

Figure 2 RMSE vs. attributes

Inferences:

1. RMSE value of lightmax is very highest among all and moisture, pressure, humidity, and temperature has very low RMSE values.
2. Since one can see the large RMSE values for lightmax, rain and lightavgw/o0, hence we can say that these data when filled with mean has a drastic change with the original one.
3. For lightmax, rain and lightavgw/o0 like data one should not use method of mean to replace the missing data, otherwise it will give a changed references.

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	temperature (in °C)	21.17	12.727	21.15	4.39	21.214	12.727	22.27	4.35
2	humidity (in $g.m^{-3}$)	83.41	99	91.367	18.37	83.48	99	91.38	18.21
3	pressure (in mb)	1009.73	789.39	1014.68	45.915	1009	789.39	1014.67	46.98
4	rain (in ml)	10836.05	0	19.689	24896.12	10701.5	0	18	24852.25
5	lightavgw/o0 (in lux)	4521.71	4488.91	1579.86	7631.52	4438.42	4488.91	1656.8	7573.162
6	lightmax (in lux)	21529.65	4000	6569	21959	21788.62	4000	6634	22064
7	moisture (in %)	32.167	0	16.220	33.8	32.38	0	16.7	33.65

Inferences:

1. All the attribute has about same mean value to original one, mode is exactly same, and mean and standard deviation is also same. Hence overall all the statistical data is about same.

2. Atmospheric pressure and temperature were having the largest missing values among the attribute but here one can see that all the statistical data is same to the real one, hence the method of interpolation works fine to analyze data.

3. One can reliable on the method of interpolation to analyze a data. Since the data are quite same.

4. It is reliable to get inference or fill your data by mean in case of mean, median and standard deviation but for all the data if you want to get more approximate to original one then you should must choose interpolation method to analyze.

ii.

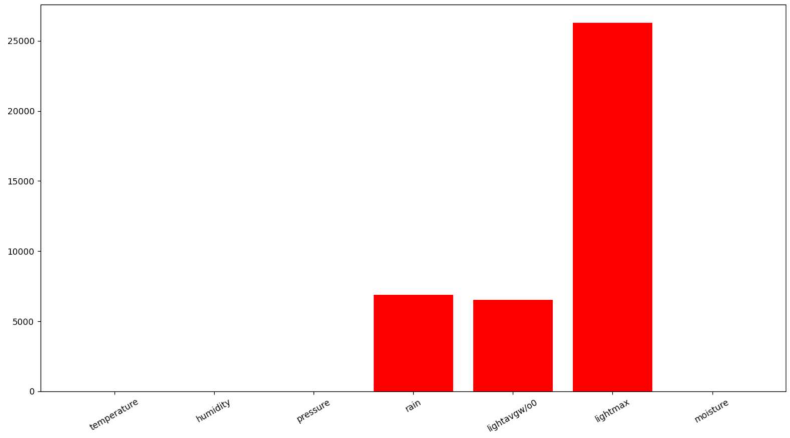

Figure 3 RMSE vs. attributes

Inferences:

1. lightmax has highest RMSE value while attributes like temperature, humidity, pressure and moisture has very low RMSE value.
2. Pressure and temperature data has very high missing values while lightavgw/o0, rain and lightmax has less missed data but RMSE values of pressure and temperature is low and lightavgw/o0, rain and lightmax has high RMSE values. Hence one can say that the things gone opposite.
3. From RMSE data we get that interpolation should not be done with rain, lightavgw/o0 and lightmax attributes.
4. Seeing both the graphs we can easily say that interpolation is a better technique to find missing data values.
5. One should use interpolation instead of mean for missing data.

5 a.

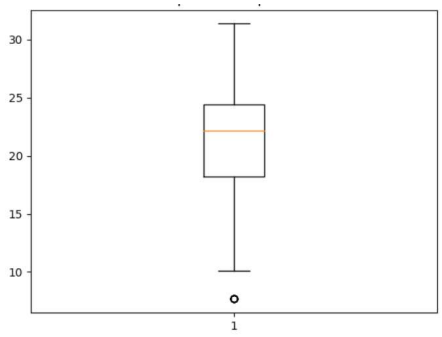

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. There are 10 outliers whose indexes with values are :
   {509: 7.6729, 510: 7.6729, 511: 7.6729, 512: 7.6729, 513: 7.6729, 514: 7.6729, 515: 7.6729, 516: 7.6729, 517: 7.6729, 518: 7.6729}
2. Since the first quartile is at 16 and third is at 24, hence interquartile range is of 8 unit.
3. Since we can see that only 10 values are outliers and the range of data and values of interquartile range is similar, hence the data is very slightly spread.
4. Since the median line is slightly above from to mid of the box, hence very slight data is more on below side so the data is slightly negatively skewed.

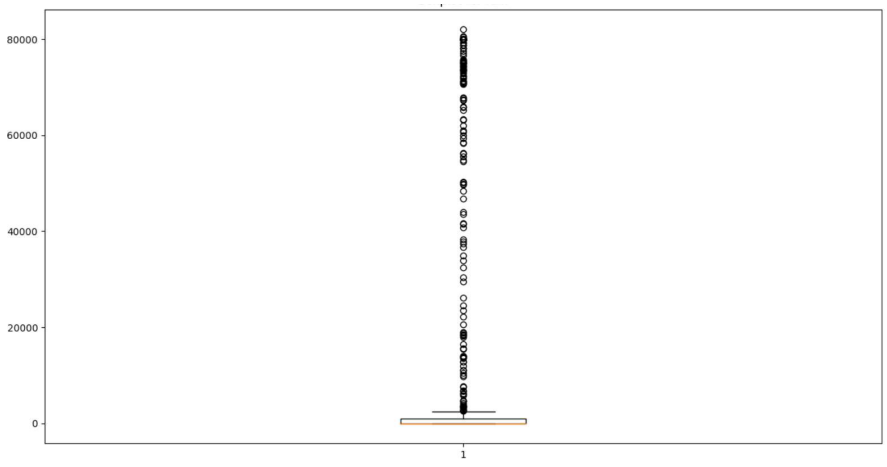

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. There are total 183 outliers whose index and values are :
   {135: 13583.25, 136: 6791.625, 199: 15459.75, 200: 14001.75, 201: 16571.25, 206: 13666.5, 322: 59982.75, 323: 80000.0, 324: 75048.75, 367: 80000.0, 368: 80000.0, 369: 80000.0, 370: 80000.0, 630: 3930.5, 631: 36636.75, 632: 40789.0, 636: 63256.5, 637: 54616.5, 638: 50172.75, 693: 37928.25, 694: 26178.75, 696: 3138.75, 697: 3449.25, 699: 18884.25, 702: 9765.0, 704: 18976.5, 705: 30393.0, 711: 2814.75, 742: 80000.0, 743: 82037.25, 744: 56319.75, 748: 71968.5, 749: 80000.0, 750: 80000.0, 751: 50242.5, 752: 80000.0, 753: 80000.0, 754: 80000.0, 755: 80000.0, 756: 80000.0, 757: 80000.0, 758: 80000.0, 759: 80000.0, 760: 80000.0, 761: 80000.0, 762: 80000.0, 763: 80000.0, 764: 80000.0, 765: 80000.0, 766: 80000.0, 767: 80000.0, 768: 80000.0, 769: 80000.0, 770: 80000.0, 771: 80000.0, 772: 80000.0, 773: 80000.0, 774: 80000.0, 775: 80000.0, 776: 80000.0, 777: 80000.0, 778: 80000.0, 779: 80000.0, 780: 80000.0, 781: 80000.0, 782: 80000.0, 783: 60675.75, 784: 41463.0, 785: 22250.25, 788: 2637.0, 789: 80000.0, 790: 80000.0, 791: 80000.0, 792: 80000.0, 793: 80000.0, 794: 80000.0, 795: 80000.0, 796: 37392.75, 798: 49725.0, 799: 80000.0, 800: 80000.0, 801: 71154.0, 802: 80000.0, 803: 80000.0, 825: 12854.25, 826: 34879.5, 827: 4610.25, 828: 6210.0, 829: 10557.0, 831: 3451.5, 835: 3312.0, 836: 18285.75, 840: 3613.5, 841: 2893.5, 842: 23474.25, 843: 14042.25, 846: 3647.25, 847: 5877.0, 851: 10062.0, 853: 17997.75, 854: 29517.75, 855: 32514.75, 856: 13943.25, 857: 4212.0, 858: 4691.25, 859: 7519.5, 862: 11112.75, 863: 2821.5, 864: 33941.25, 865: 43643.25, 866: 20664.0, 867: 11144.25, 868: 18587.25, 869: 18373.5, 870: 15646.5, 871: 12915.0, 872: 49916.25, 873: 24522.75, 874: 75105.0, 875: 73417.5, 876: 70580.25, 877: 78126.75, 878: 56097.0, 879: 6061.5, 883: 38355.75, 884: 55509.75, 885: 43974.0, 886: 6747.75, 887: 54843.75, 888: 59377.5, 889: 58320.0, 890: 60963.75, 891: 63342.0, 892: 67378.5, 893: 70929.0, 894: 73158.75, 895: 71367.75, 896: 73838.25, 897: 46732.5, 898: 48429.0, 899: 67830.75, 900: 75447.0, 901: 74646.0, 902: 75402.0, 903: 75723.75, 904: 74254.5, 905: 75201.75, 906: 77044.5, 907: 74472.75, 908: 77503.5, 909: 78180.75, 910: 79915.5, 911: 80583.75, 912: 80482.5, 913: 79337.25, 914: 79317.0, 915: 70823.25, 916: 75638.25, 917: 73752.75, 918: 65893.5, 919: 72774.0, 920: 7773.75, 923: 12037.5, 924: 79839.0, 925: 78633.0, 926: 78779.25, 927: 76662.0, 928: 67252.5, 929: 74913.75, 930: 4869.0, 931: 41618.25, 933: 58443.75, 934: 74173.5, 935: 72445.5, 936: 65873.25, 937: 67675.5, 938: 61989.75, 939: 71237.25, 940: 73577.25, 941: 65301.75, 942: 73534.5, 943: 72283.5, 944: 71799.75}
2. The interquartile range is in order of 1000.
3. Seeing the outliers, interquartile range infront of range value of the data one can easily say that data is largely spread.
4. We can see that so many data is on upper side hence it positively skewed.
5. These so many outliers having high values means that the most of the day are not rainy days but when the days it rains, it does heavily.

b.

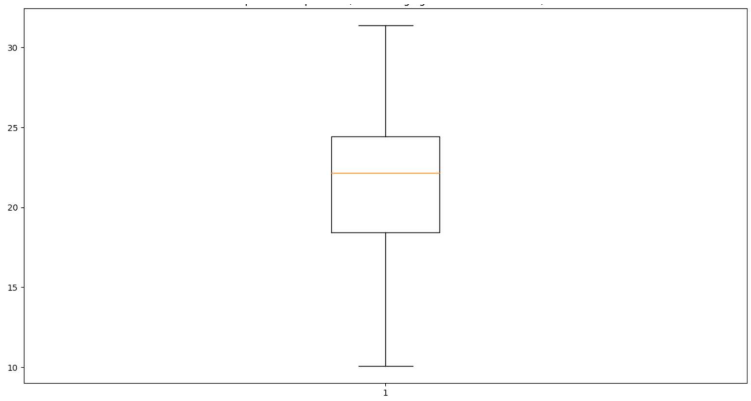

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. There are no outliers left now, initially it were 10 .
2. Now the first quartile range is at 18 and third quartile is at 24, hence inter-quartile range becomes 6 while initially it was at 8, with same third quartile but now there is slightly increase in first quartile.
3. Now the inter-quartile range becomes 6 but the range is same hence the spread is more now.
4. The median line is above to mid of the box, hence this is negatively skewed data.

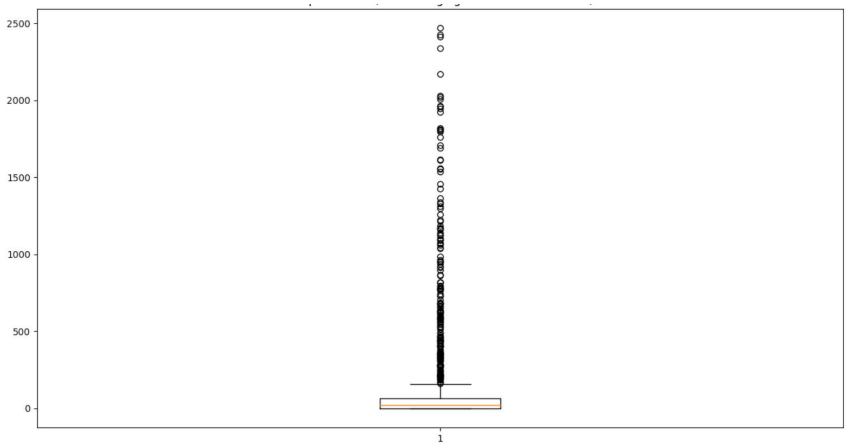

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. Total outliers are now : 191, the index with the values are :
   {1: 1761.75, 2: 652.5, 3: 963.0, 4: 254.25, 5: 339.75, 11: 607.5, 12: 560.25, 13: 513.0, 15: 474.75, 16: 817.875, 17: 1161.0, 20: 240.75, 21: 398.25, 23: 816.75, 24: 776.25, 25: 681.75, 26: 441.0, 27: 274.5, 30: 1341.0, 31: 1804.5, 36: 2171.25, 37: 1456.875, 38: 742.5, 39: 443.25, 40: 774.0, 41: 1167.75, 42: 898.875, 43: 630.0, 44: 594.0, 48: 546.75, 49: 576.0, 50: 605.25, 51: 634.5, 53: 1091.25, 56: 162.0, 62: 366.75, 63: 183.375, 70: 589.5, 71: 207.0, 72: 281.25, 73: 1215.0, 90: 315.0, 141: 1260.0, 142: 324.0, 144: 360.0, 145: 679.5, 198: 1710.0, 202: 1183.5, 203: 1962.0, 204: 1071.0, 205: 438.75, 207: 864.0, 208: 816.75, 209: 796.5, 213: 191.25, 218: 202.5, 219: 1611.0, 227: 353.25, 229: 533.25, 230: 213.75, 231: 434.25, 232: 191.25, 235: 202.5, 237: 594.0, 238: 409.5, 246: 333.0, 248: 468.0, 250: 222.75, 265: 263.25, 321: 459.0, 328: 272.25, 377: 621.0, 381: 587.25, 382: 468.0, 384: 778.5, 385: 987.75, 388: 623.25, 389: 330.75, 393: 1075.5, 394: 308.25, 395: 337.5, 397: 1617.75, 400: 402.75, 401: 2414.25, 409: 1044.0, 411: 211.5, 412: 285.75, 413: 400.5, 419: 1426.5, 426: 209.25, 428: 551.25, 432: 344.25, 442: 1140.75, 448: 357.75, 452: 308.25, 455: 774.0, 456: 703.125, 457: 632.25, 458: 561.375, 459: 490.5, 460: 419.625, 461: 348.75, 463: 277.875, 464: 207.0, 467: 1172.25, 470: 427.5, 471: 213.75, 480: 187.5, 481: 273.375, 482: 359.25, 483: 445.125, 484: 531.0, 489: 1311.75, 496: 247.5, 507: 454.5, 522: 283.5, 523: 1062.0, 525: 1554.75, 526: 569.25, 527: 357.75, 528: 1795.5, 529: 382.5, 533: 353.25, 534: 918.0, 535: 677.25, 536: 1689.75, 561: 213.75, 633: 637.5, 634: 2470.5, 641: 580.5, 669: 951.75, 670: 281.25, 671: 684.0, 672: 463.5, 673: 420.75, 676: 1329.75, 680: 173.25, 681: 211.5, 685: 173.25, 689: 1300.5, 691: 326.25, 698: 621.0, 700: 1818.0, 701: 783.0, 707: 949.5, 718: 438.75, 719: 1559.25, 720: 1039.5, 721: 405.0, 722: 582.75, 724: 234.0, 727: 666.0, 728: 625.5, 729: 1365.75, 730: 1129.5, 732: 524.25, 734: 492.75, 735: 920.25, 736: 218.25, 739: 2022.75, 740: 2009.25, 745: 438.75, 746: 285.75, 747: 225.0, 786: 1809.0, 787: 1226.25, 797: 1964.25, 812: 321.75, 814: 688.5, 818: 765.0, 819: 1125.0, 820: 868.5, 821: 1107.0, 822: 405.0, 823: 731.25, 830: 794.25, 832: 1536.75, 833: 954.0, 834: 731.25, 838: 1926.0, 839: 1818.0, 844: 243.0, 845: 373.5, 849: 308.25, 850: 936.0, 852: 2029.5, 881: 661.5, 882: 1946.25, 921: 1095.75, 922: 2340.0, 932: 2427.75}
2. The inter-quartile range now is of the order of 100 to 10.
3. Since initially range was of order about 80000 which now becomes 2500, hence there is huge change in spead of data.
4. Since the data is less spread now but wrt previous one but still we can see a large number of outliers with larger values, hence range of data is still high wrt inter-quartile range, hence it is still largely spread.