**Student’s Name:** Sachin Mahawar

**Roll Number:** b20129

**Mobile No:** 9166843951

**Branch:** CSE

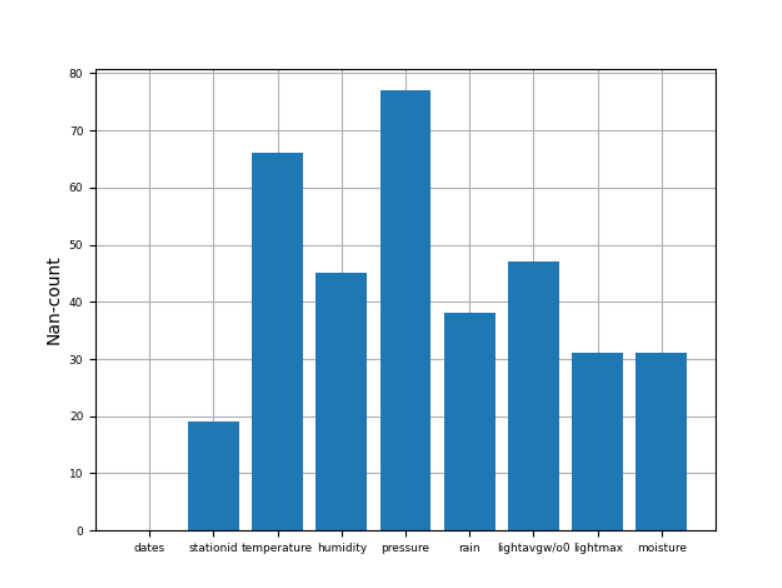


Figure Number of missing values vs. attributes

**Inferences:**

1. Attribute ‘pressure’ has the maximum missing values and attribute ‘dates’ has minimum.
2. The frequency of missing values is very less compared to rows in dataframe.

# a.

**Inferences:**

1. If ‘station id’ attribute is missing then there is no meaning of all its data as we don’t know which station’s data it is.
2. Total number of rows deleted is 19.
3. Percentage of rows deleted: 2.01%.

**b.**

**Inferences:**

1. Total number of tuples deleted is 39 which have missing values more than 2.
2. Percentage of rows deleted: 4.12%.
3. Data loss is very less so we can use this data.
4. Since rows which are deleted have 3 or more missing values which makes that tuple less informative and it just cause problem for whole dataset so removing such row is better.

Table 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. Attribute ‘pressure’ has maximum missing values which is 41 and attribute ‘dates’ and ‘Stationid’ has minimum missing values which is 0.
2. “Dates” and “Stationid” have 0% missing data, ”temperature” has 3.59% missing data, “humidity” has 1.37% missing data, “pressure” has 4.33% missing value, “rain” has 0.63% missing data, “lightavg” has 1.58% missing data, “lightmax” has 0.1% missing data and “moisture” has 0.63% missing data.
3. Total number of missing values: 116.

# a. i.

Table 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 | NA | NA | NAN | NA | NAN | NA | NA | NA |
| **2** | stationid | NA | NA | NA | NA | NA | NA | NA | NA |
| **3** | temperature (in °C) | 21.21 | 12.72 | 22.27 | 4.35 | 21.07 | 21.07 | 21.80 | 4.24 |
| **4** | humidity (in g.m−3) | 83.47 | 99.00 | 91.38 | 18.20 | 83.26 | 99.00 | 90.11 | 17.95 |
| **5** | pressure (in mb) | 1009.00 | 789.39 | 1014.67 | 46.95 | 1009.22 | 1009.22 | 1014.07 | 45.19 |
| **6** | rain (in ml) | 10701.53 | 0.00 | 18.00 | 24839.10 | 10942.72 | 0.00 | 24.75 | 24561.24 |
| **7** | lightavgw/o0 (in lux) | 4438.42 | 4488.91 | 1656.88 | 7569.15 | 4430.92 | 4488.91 | 1911.23 | 7396.66 |
| **8** | lightmax (in lux) | 21788.62 | 4000.00 | 6634.00 | 22053.31 | 21650.16 | 4000.00 | 7544.00 | 21666.72 |
| **9** | moisture (in %) | 32.38 | 0.00 | 16.70 | 33.63 | 32.67 | 0.00 | 17.72 | 33.39 |

**Inferences:**

1. Maximum change: (1) Mean: rain (2) Median: lightavgw/0 (3) Mode: pressure (4) S.D.: lightmax

Minimum change: (1) Mean: temperature (2)Median: temperature (3)Mode: humidity (4) S.D.: temperature

1. There is very less change in data for attributes having less values and large change in data for attributes having large values.
2. Since for most of the attributes, change in values is very small, so data is still reliable.

**ii.**

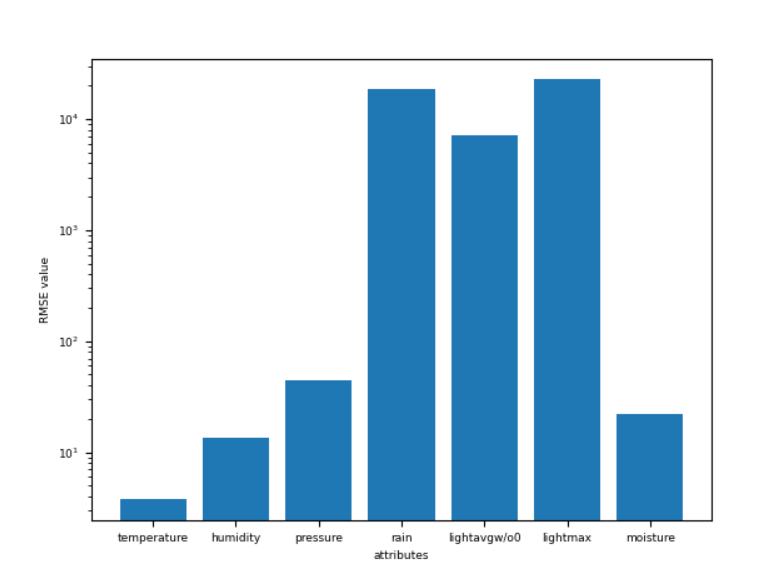


Figure RMSE vs. attributes

**Inferences:**

1. Attribute ‘lightmax’ has maximum RMSE value which is 22711.61 and attribute ‘temperature’ has minimum RMSE value which is 3.7.
2. There is no specific relation between missing values and maximum RMSE and also for minimum missing values and minimum RMSE values.
3. Since RMSE values are quite high for almost all attributes, data is not reliable.

# b. i.

Table 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 | NA | NA | NA | NA | NA | NA | NA | NA |
| **2** | stationid | NA | NA | NA | NA | NA | NA | NA | NA |
| **3** | temperature (in °C) | 21.196 | 12.727 | 22.169 | 4.327 | 21.214 | 12.727 | 22.272 | 4.353 |
| **4** | humidity (in g.m−3) | 83.538 | 99.00 | 91.380 | 18.197 | 83.479 | 99.00 | 91.380 | 18.20 |
| **5** | pressure (in mb) | 1009.264 | 789.392 | 1014.677 | 45.974 | 1009.008 | 789.392 | 1014.677 | 46.955 |
| **6** | rain (in ml) | 10651.638 | 0 | 22.500 | 24766.397 | 10701.538 | 0 | 18.00 | 24839.102 |
| **7** | lightavgw/o0 (in lux) | 4486.340 | 4488.910 | 1623.494 | 7569.787 | 4438.428 | 4488.910 | 1656.880 | 7569.154 |
| **8** | lightmax (in lux) | 21517.191 | 4000.00 | 6569.00 | 21923.55 | 21788.623 | 4000.00 | 6634.00 | 22053.315 |
| **9** | moisture (in %) | 32.327 | 0 | 16.306 | 33.584 | 32.386 | 0 | 16.704 | 33.635 |

**Inferences:**

1. Attribute “lightmax” has maximum change in values and attributes “dates” and “stationed” have minimum change.
2. There is very less change in data for attributes having less values and large change in data for attributes having large values.
3. Since for most of the attributes, change in values is very small, so data is still reliable.

**ii.**

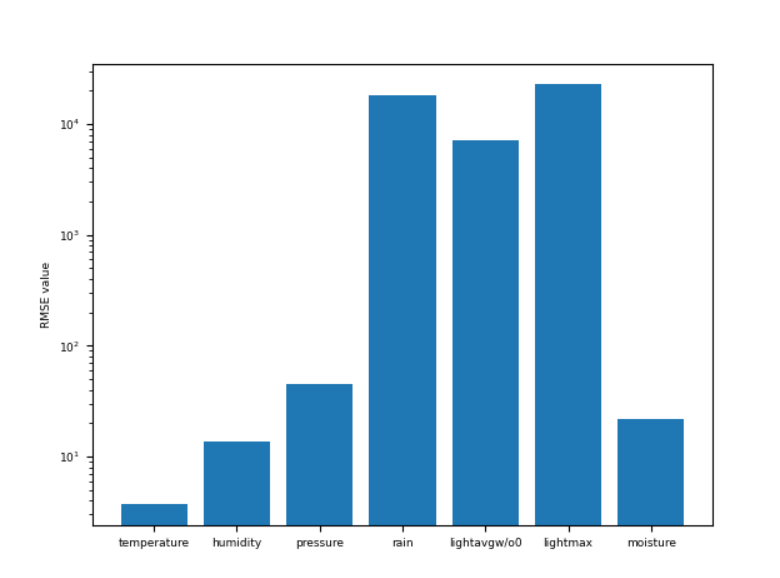


Figure RMSE vs. attributes

**Inferences:**

1. Attribute ‘lightmax’ has maximum RMSE value which is 22736.704 and attribute ‘temperature’ has minimum RMSE value which is 3.695
2. There is no specific relation between missing values and maximum RMSE and also for minimum missing values and minimum RMSE values.
3. Since RMSE values are quite high for almost all attributes, data is not reliable.

**5 a.**

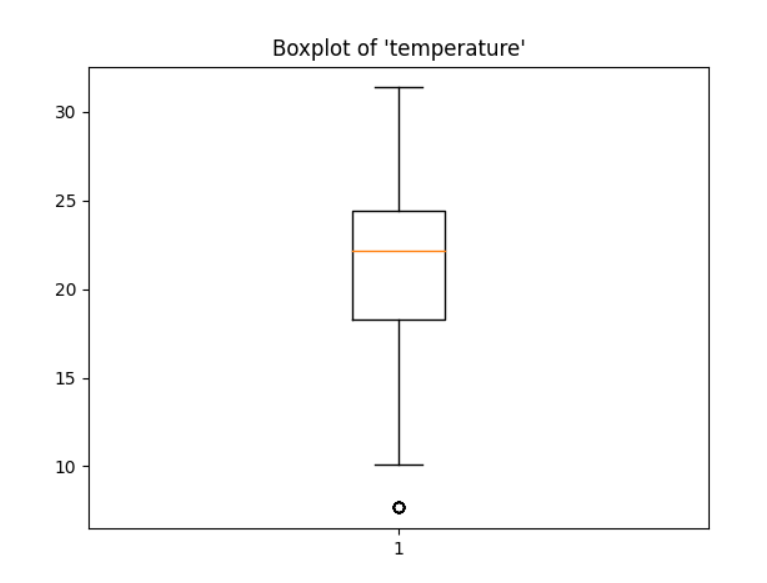


Figure Boxplot for attribute temperature (in °C)

**Inferences:**

1. Value of outliers is 7.6729 and it’s count is 10.
2. Inter quartile range is approximately 6.
3. Due to very few outliers present in this data set, variance/spread is very low.
4. Since the median line lies above the middle point so it is negatively skewed data.

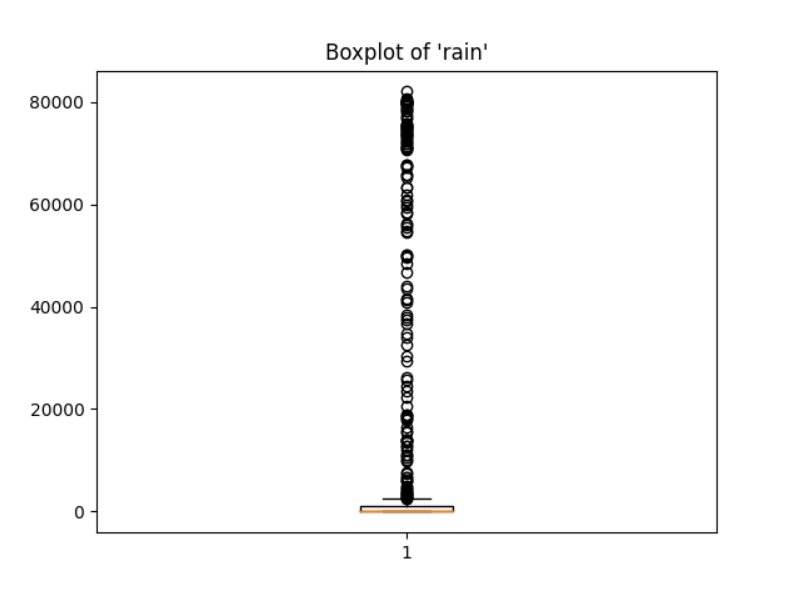


Figure Boxplot for attribute rain (in ml)

**Inferences:**

1. There are total 185 outliers present in this data set ranging from 82037.25 to 2470.5.
2. Inter quartile range is around 100.
3. Since the number of outliers are 185, which is very high so it has quite high spread.
4. Since the median line lies below the middle point of iqr, it is positively skewed dataset.

**b.**

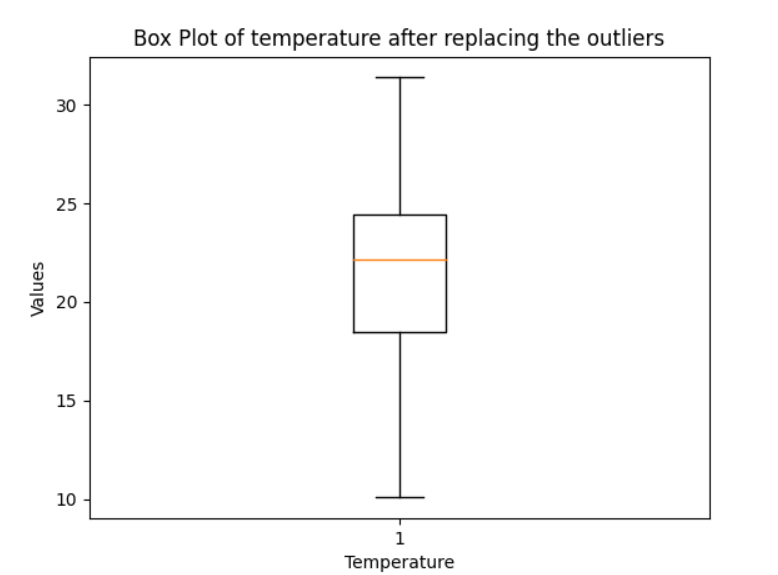


Figure Boxplot for attribute temperature (in °C) after replacing median with outliers

**Inferences:**

1. No outliers are present in this boxplot.
2. IQR is same as before which his around 6.
3. Variance is also almost same as before.
4. It is still negatively skewed same as before.

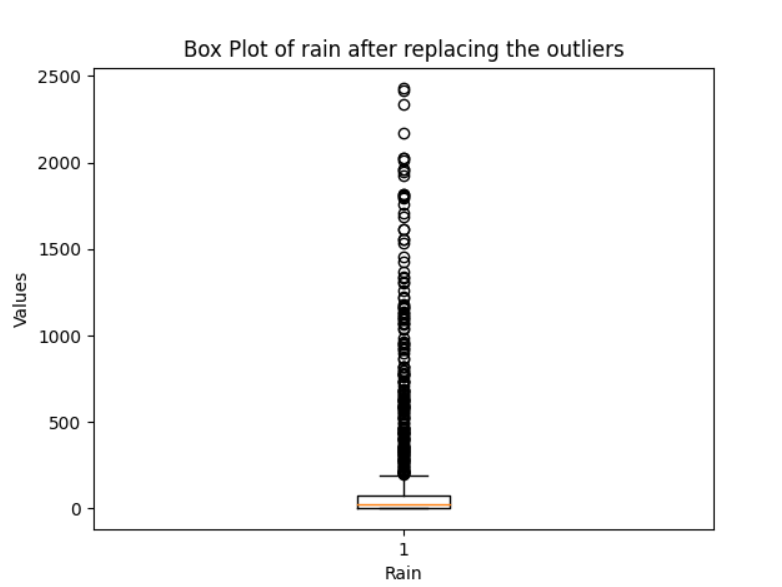


Figure Boxplot for attribute rain (in ml) after replacing median with outliers

**Inferences:**

1. There are total 193 outliers present in this data set which has increased from previous case ranging from 200.25 to 2427.75.
2. Inter quartile range is around 70 which get decreased from previous case.
3. Since the number of outliers has increased, variance is still high.
4. Skewness of data set does not change and is same as before.