**Student’s Name: Nikhil**

**Roll Number: B20219**

**Mobile No: 8949463760**

**Branch:**

**CSE**

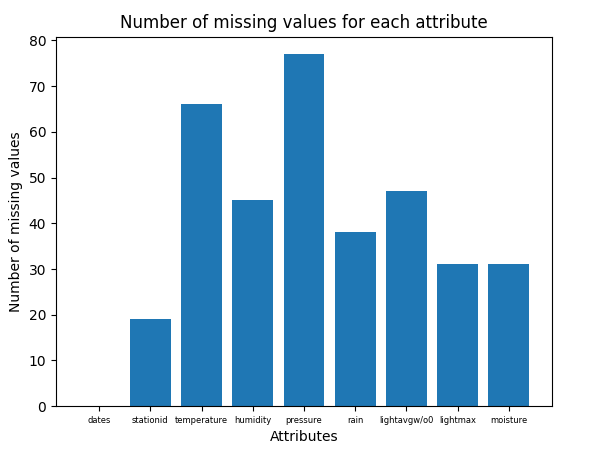


Figure 1 Number of missing values vs. attributes

**Inferences:**

1. The attribute ‘pressure’ has the highest number of missing values while the attribute ‘dates’ has the lowest number of missing values.
2. We see that temperature and pressure have high number of missing values (66 and 77 respectively). Humidity. Rain, ghtavgw/o0, ghtmax and moisture have moderate number of missing values (45,38,47,31 and 31 respectively) while dates and stationid have low number of missing values (0 and 19 respectively)

# a.

**Inferences:**

1. We usually delete the tuples with missing value of the target attribute because our main concern in data is of this attribute. So if any row has missing value in this attribute, there is no point in doing calculations and taking that row into consideration.
2. A total of 19 tuples had to be deleted from the data-set.
3. 2% of the total number of tuples have been deleted

**b.**

**Inferences:**

1. 35 tuples have been deleted in this step.
2. 3.78% of the total number of tuples have been deleted.
3. Comment on the data loss
4. Since these rows contained very low amount of data, they are almost useless, as taking them into consideration will change the net properties of non missing attributes in these and also require extra time for computation. So we delete these kind of rows

Table 1 Number of missing values per attribute after removing missing values

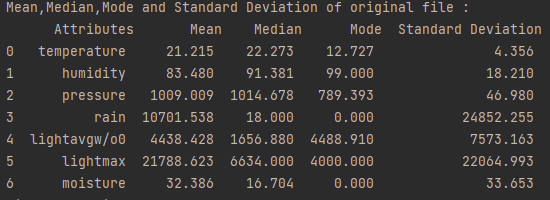
|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Attribute** | **Number of missing values** | **Percentage of missing values** |
| **1** | dates | 0 | 0% |
| **2** | stationid | 0 | 0% |
| **3** | temperature (in °C) | 34 | 3.6% |
| **4** | humidity (in g.m−3) | 13 | 1.37% |
| **5** | pressure (in mb) | 41 | 4.334% |
| **6** | rain (in ml) | 6 | 0.634% |
| **7** | lightavgw/o0 (in lux) | 15 | 1.585% |
| **8** | lightmax (in lux) | 1 | 0.1% |
| **9** | moisture (in %) | 6 | 0.634% |

**Inferences:**

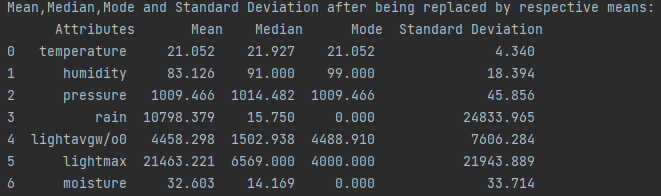
1. ‘pressure’ has maximum number of missing values. ‘dates’ and ‘stationid’ have minimum number of missing values.
2. The file has 116 missing values.

# a. i.

Table 2 Mean, mode, median and standard deviation before and after replacing missing values by mean



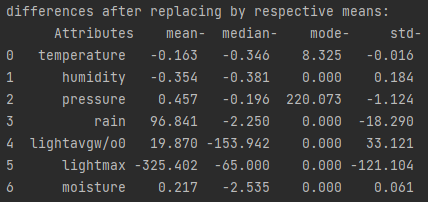
**Before**



**After**

**Inferences:**

Differences in the properties after replacing the missing values by respective means:



1. Attributes having the maximum and the minimum change :

|  |  |  |
| --- | --- | --- |
| **Property** | **Attribute having max change** | **Attribute having min change** |
| Mean | lightmax | temperature |
| Median | lightavgw/o0 | pressure |
| Mode | pressure | All except temperature |
| Standard Deviation | lightmax | temperature |

1. Pressure had most number of missing values and it showed minimum change in median and maximum change in mode. Lightmax had least number of missing values and it showed maximum change in mean and standard deviation and minimum change in mode. It doesn’t seem a good relation.
2. The change observed in almost all attribute is significantly high for at least one property so the data is not fully reliable for further experimental analysis.

**ii.**

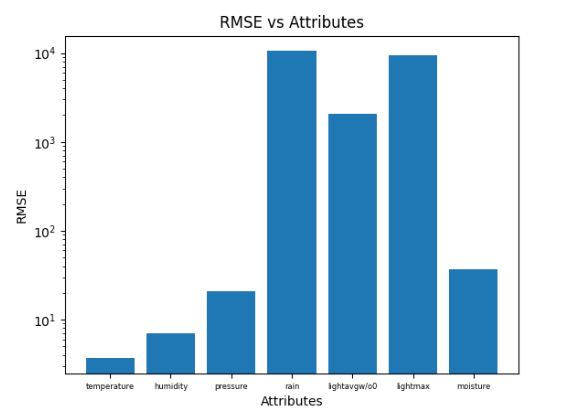


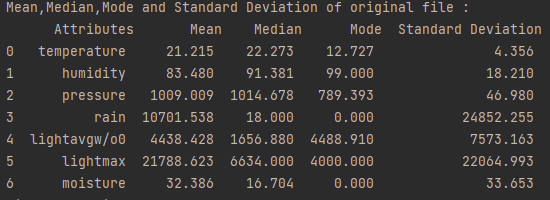
Figure 2 RMSE vs. attributes

**Inferences:**

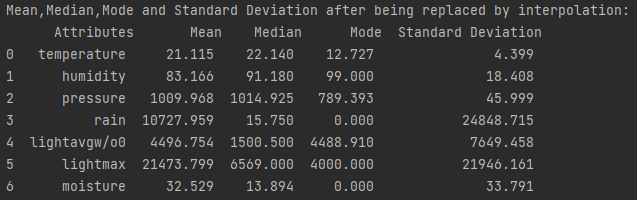
1. The attributes rain and temperature have maximum and minimum RMSE values respectively.
2. Rain had neither any maximum nor minimum change in any property but here it shows maximum RMSE value. Some attributes like pressure had highest change in some properties but they show very less change in RMSE. Pressure had most missing value but very less change in RMSE. So there is no appreciable relation.
3. Since many attributes show high RMSE values, the data is not so reliable for further analysis.

# b. i.

Table 3 Mean, mode, median and standard deviation before and after replacing missing values by linear interpolation technique



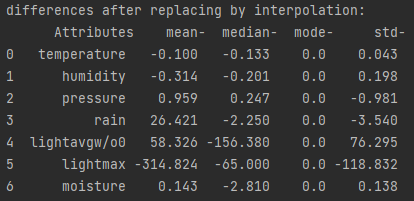
**Before**



**After**

**Inferences:**

Differences in the properties after replacing the missing values by interpolation:



1. Attributes having the maximum and the minimum change :

|  |  |  |
| --- | --- | --- |
| **Property** | **Attribute having max change** | **Attribute having min change** |
| Mean | lightmax | temperature |
| Median | lightavgw/o0 | temperature |
| Mode | All show no change | All show no change |
| Standard Deviation | lightmax | temperature |

1. Temperature shows least change in all properties and it has 2nd highest number of missing values, but pressure, which has most number of missing values shows a significant amount of change in all the attributes. This shows that there is no appreciable relation.
2. The change observed in properties is low as compared to mean for most of the attributes, and that’s why it can be taken as a little reliable for further investigation.
3. Almost all attributes except lightavgw/o0 show less change in the properties than the previous case.
4. Inference 5(You may add or delete the number of inferences)

**ii.**

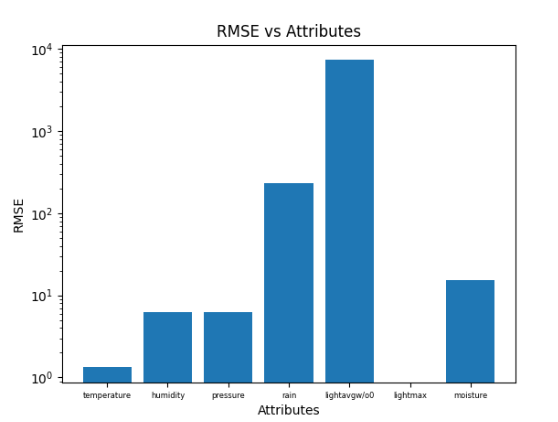


Figure 3 RMSE vs. attributes

**Inferences:**

1. The attributes lightavgw/o0 and lightmax have maximum and minimum values of RMSE respectively.
2. The RMSE and missing values do not show any following trend. But lightavgw/o0 has highest RMSE and also shows high difference in properties, same is with rain, it has 2nd highest RMSE and also shows high difference in properties. Temperature has low RMSE and shows very less differences in properties. Hence there is an appreciable relation between RMSE and difference in properties.
3. Since RMSE is significantly high for most attributes, this is not reliable for further experimental analysis.
4. For most attributes( except humidity and lightavgw/o0), RMSE was higher when replaced by mean than when replaced by interpolation.

# a.

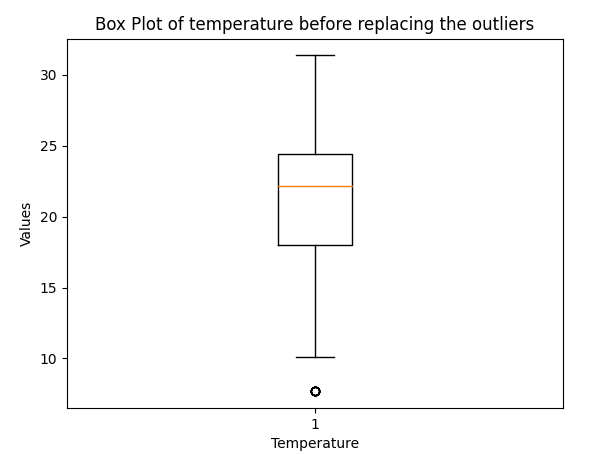


Figure 4 Boxplot for attribute temperature (in °C)

**Inferences:**

1. There is 1 outlier with value nearly 7.7.
2. The Inter quartile range is 24.5-18 = 6.
3. The data has maximum value 31.33 and minimum value 7.7. The first quartile lies at temperature 18, the second quartile(median) lies at temperature 22 and the third quartile lies at temperature 24.5. The mode of the data lies between the 2nd and 3rd quartile.
4. Since median is smaller than the mode, the data is left(negative) skewed.

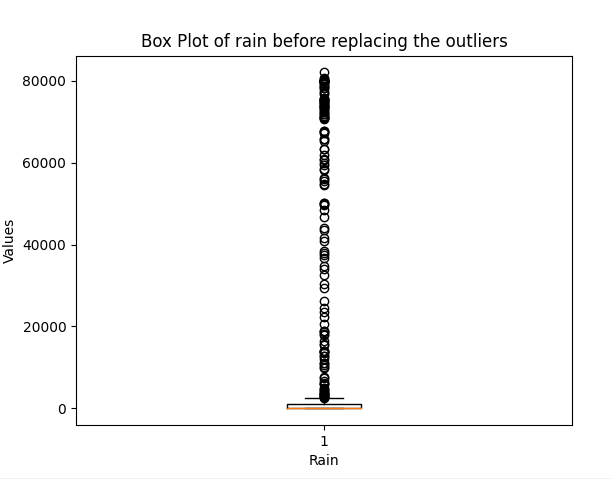


Figure 5 Boxplot for attribute rain (in ml)

**Inferences:**

1. There are a large number of outliers lying in the range 2500-80000.
2. The Inter quartile range is 1050-0=1050.
3. The data has maximum value 80000 and minimum value 0. The first quartile lies at 0, the second quartile(median) lies at 15.7 and the third quartile lies at 1050. The mode of the data lies between the 1st and 2nd quartile.
4. Since median is greater than the mode, the data is right(positive) skewed.

**b.**

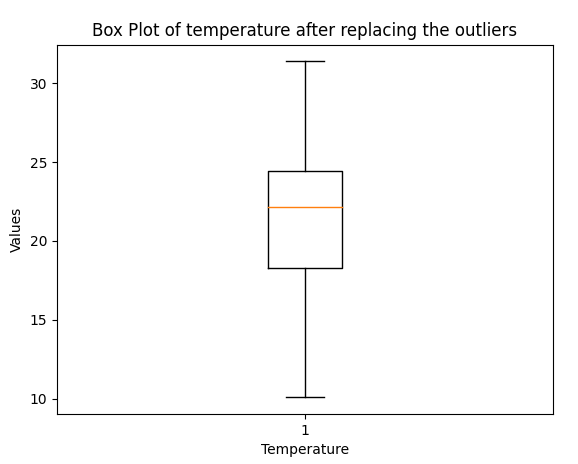


Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

**Inferences:**

1. There are no outliers, unlike the data before replacing outliers.
2. The inter quartile range is 24.4-18.2 = 6.2, which is almost same as previously obtained value.
3. The data has maximum value 31.33, which is same as the previous data and minimum value 10, which is almost 2.3 greater than the previous data. The first quartile lies at temperature 18.2, the second quartile(median) lies at temperature 22 and the third quartile lies at temperature 24.5, which are all almost same as the previous data. The mode of the data lies between the 2nd and 3rd quartile, same as the previous data.
4. Since median is smaller than the mode, the data is left(negative) skewed, as in the previous case.

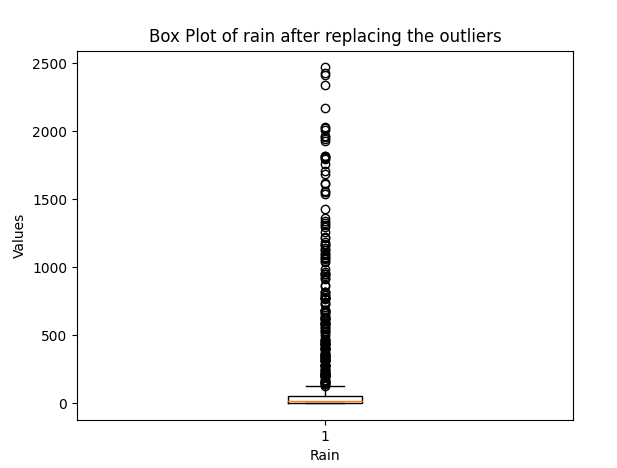


Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

**Inferences:**

1. There are a large number of outliers lying in the range 130-2500, which is way less than the previous case( almost 30 times smaller range)
2. The Inter quartile range is 52-0 = 52, which is way less than the previous case( almost 20 times less).
3. The data has maximum value 2450, which is way less than the previous case( almost 32 times less) and minimum value 0, which is same as the previous case. The first quartile lies at 0( same as previous case), the second quartile(median) lies at 15.7( same as previous case) and the third quartile lies at 52 (at almost 20 times smaller value than the previous case) . The mode of the data lies between the 1st and 2nd quartile, same as in previous case.
4. Since median is greater than the mode, the data is right(positive) skewed, same as the previous case.