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

Student's Name: **Mobile No:**

Roll Number: **Branch:**

1

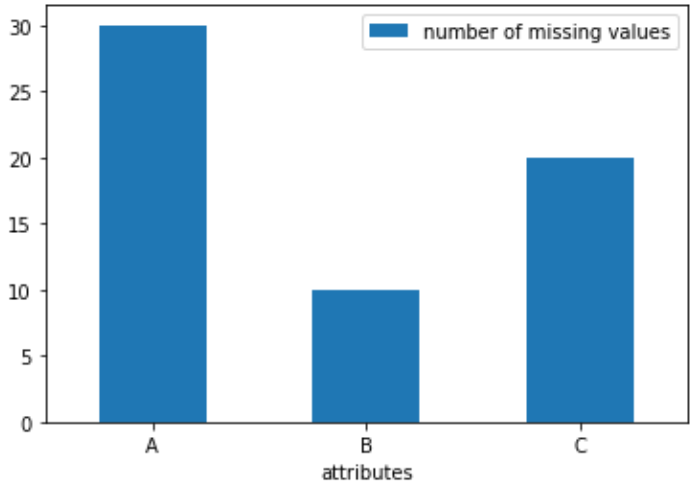

Figure 1 Number of missing values vs. attributes

Inferences:

1. Which attributes have maximum and minimum missing values respectively?
2. From the bar chart comment on the frequency of missing values for each attribute.
3. Inference 3(You may add or delete the number of inferences)
   Note: The bar chart above is for illustration purposes. Replace it with the bar chart obtained by you. Rename x-axis legend and y-axis legends with attribute names and number of missing values respectively.

2 a.

Inferences:

1. Ponder upon; why do we choose to delete the tuple if the target attribute is missing.
2. State the number of tuples deleted after this step.
3. What percentage of the total number of tuples is deleted?
4. Inference 4(You may add or delete the number of inferences)

b.

Inferences:

1. State the number of tuples deleted after this step.
2. What percentage of the total number of tuples is deleted?
3. Comment on the data loss
4. Justify the need for this step
5. Inference 5(You may add or delete the number of inferences)

3

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

S. No	Attribute	Number of missing values
1	dates	
2	stationid	
3	temperature (in °C)	
4	humidity (in $g.m^{-3}$)	
5	pressure (in mb)	
6	rain (in ml)	
7	lightavgw/o0 (in lux)	
8	lightmax (in lux)	
9	moisture (in %)	

Inferences:

1. Which attributes have maximum and minimum missing values respectively?
2. For each attribute, comment on the percentage of data missing.
3. State the total number of missing attributes in the file.
4. Inference 4(You may add or delete the number of inferences)

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

Inferences:

1. Which attributes have the maximum and the minimum change in the mean, mode, median and standard deviation respectively?
2. Is there any relation between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values? (In reference to Q1. and Q3.)
3. From the change observed in mean, mode, median and standard deviation ponder is the datareliable for further investigation or experimental analyses.
4. Inference 4(You may add or delete the number of inferences)

ii.

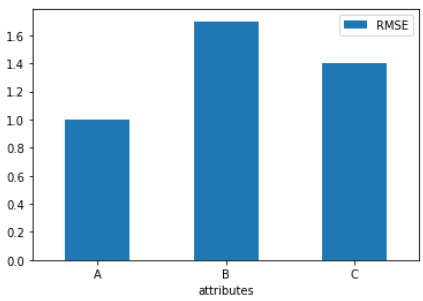

Figure 2 RMSE vs. attributes

Inferences:

1. Which attributes have maximum and minimum RMSE respectively?
2. Is there any relation between RMSE, maximum and minimum change in mean, mode, medianand standard deviation and maximum and minimum missing values? (In reference to Q1, Q3 and Q4. a. i.)
3. From RMSE ponder is the data reliable for further investigation or experimental analyses.
4. Inference 4(You may add or delete the number of inferences)
   Note: The bar chart above is for illustration purposes. Replace it with the bar chart obtained byyou. Rename x-axis legend and y-axis legends with attribute names and RMSE respectively

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

Inferences:

1. Which attributes have the maximum and the minimum change in the mean, mode, median and standard deviation respectively?
2. Is there any relation between maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values? (In reference to Q1. and Q3.)
3. From the change observed in mean, mode, median and standard deviation ponder is the data reliable for further investigation or experimental analyses.
4. From the observed changes in mean, mode, median and standard deviation compare and contrast replacing missing values by mean and linear interpolation technique.
5. Inference 5(You may add or delete the number of inferences)

ii.

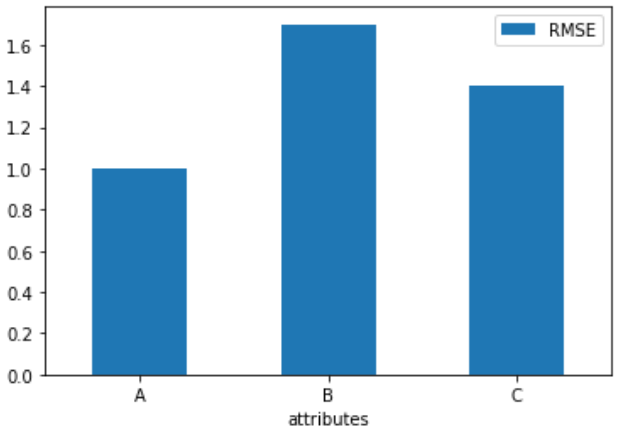

Figure 3 RMSE vs. attributes

Inferences:

1. Which attributes have maximum and minimum RMSE respectively?
2. Is there any relation between RMSE, maximum and minimum change in mean, mode, median and standard deviation and maximum and minimum missing values? (In reference to Q1, Q3 and Q4 b. i.)
3. From RMSE ponder is the data reliable for further investigation or experimental analyses.
4. From the calculated RMSE compare and contrast replacing missing values by mean and linear interpolation technique.
5. Inference 5(You may add or delete the number of inferences)
   Note: The bar chart above is for illustration purposes. Replace it with the bar chart obtained by you. Rename x-axis legend and y-axis legends with attribute names and RMSE respectively.

5 a.

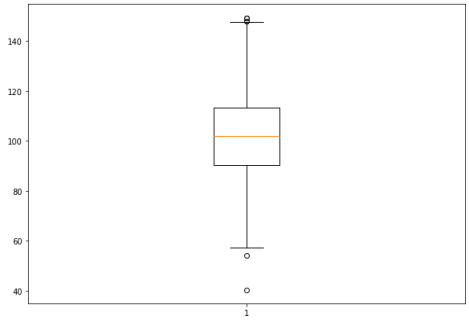

Figure 4 Boxplot for attribute temperature (in °C)

Inferences:

1. List the number of outliers and their row numbers.
2. Infer the Inter quartile range.
3. Infer the spread/variance.
4. Infer the skewness of the data.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units.

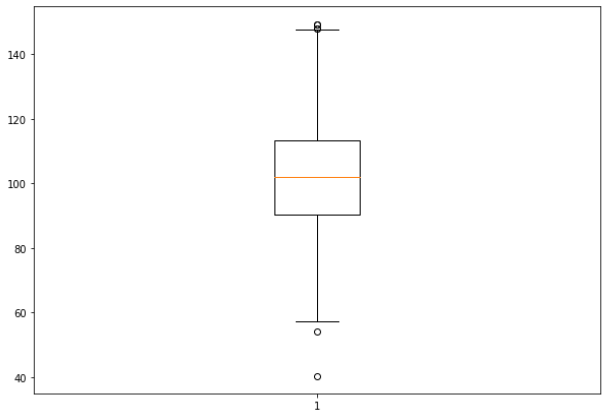

Figure 5 Boxplot for attribute rain (in ml)

Inferences:

1. List the number of outliers and their row numbers.
2. Infer the Inter quartile range.
3. Infer the spread/variance.
4. Infer the skewness of the data.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units.

b.

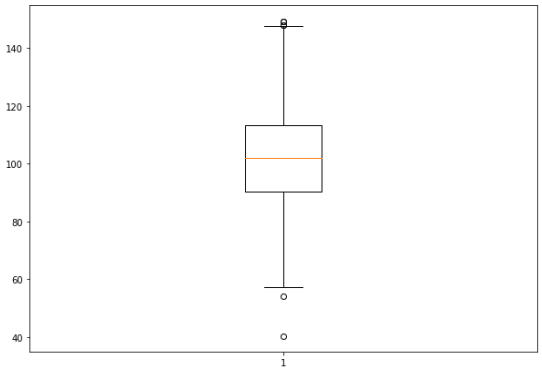

Figure 6 Boxplot for attribute temperature (in °C) after replacing median with outliers

Inferences:

1. List the number of outliers, their row number and compare with Q5. a.
2. Infer the Inter quartile range compare with Q5. a.
3. Infer the spread/variance compare with Q5. a.
4. Infer the skewness of the data compare with Q5. a.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units

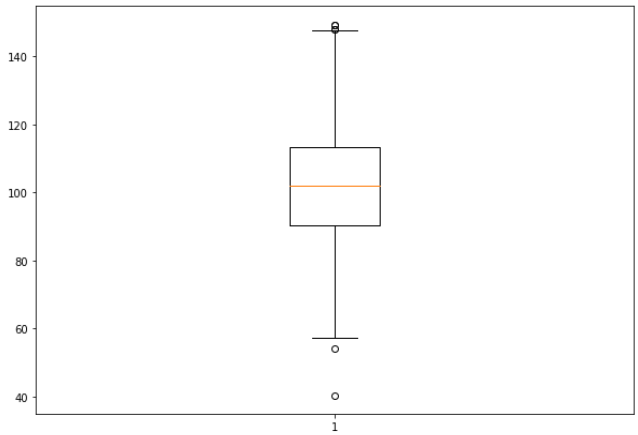

Figure 7 Boxplot for attribute rain (in ml) after replacing median with outliers

Inferences:

1. List the number of outliers, their row number and compare with Q5. a.
2. Infer the Inter quartile range compare with Q5. a.
3. Infer the spread/variance compare with Q5. a.
4. Infer the skewness of the data compare with Q5. a.
5. Inference 5 (You may add or delete the number of inferences)
   Note: The boxplot above is for illustration purposes. Replace it with the boxplot obtained by you. Rename legends with appropriate attribute names with units

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.**