Data Quality Management – Imputation Component

1.1 What is Data imputation?

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

1.2 Why Data Imputation?

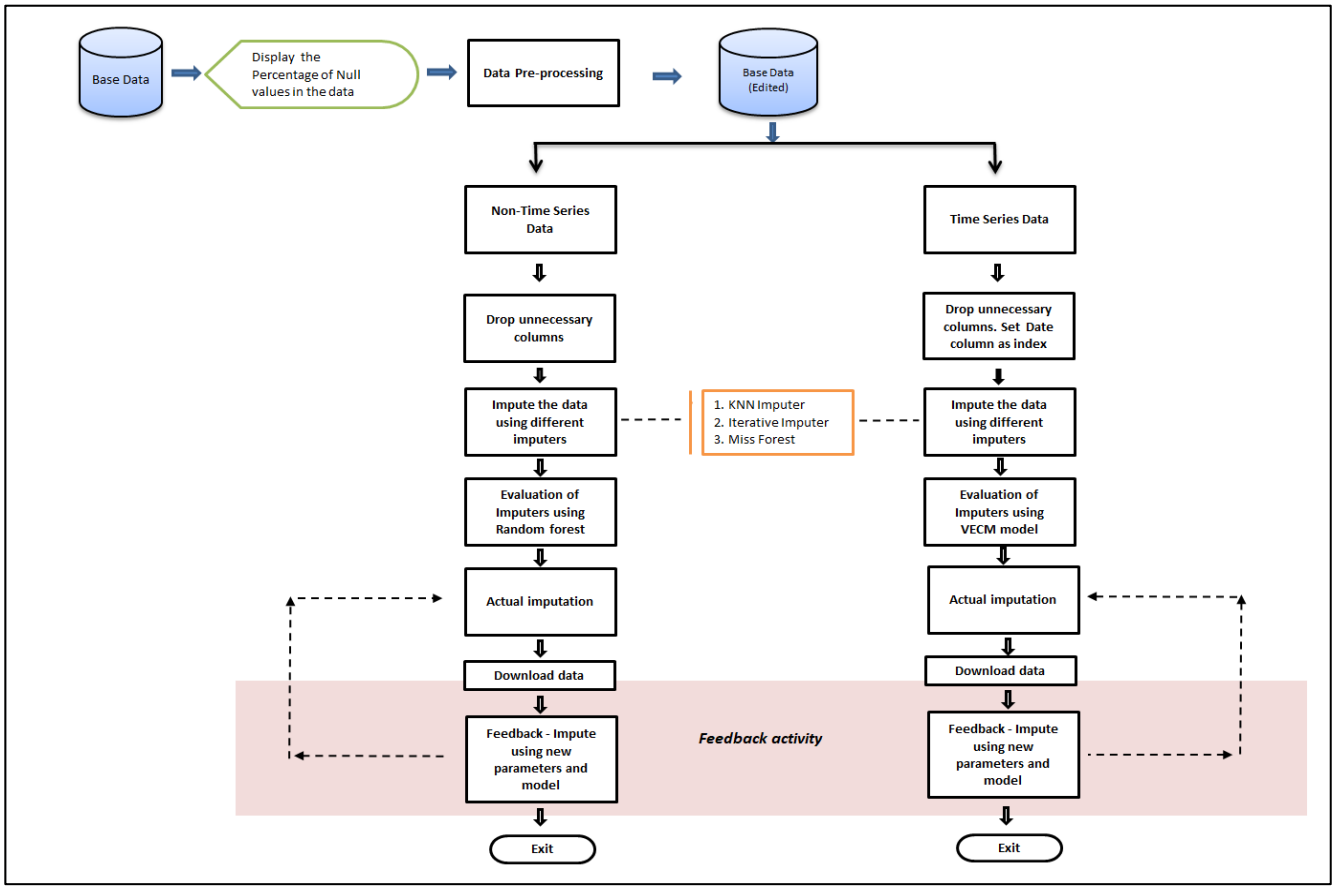
1. *Incompatible with most of the Python libraries used in Machine Learning:* While using the libraries for ML, they don't have a provision to automatically handle these missing data and can lead to errors.
2. *Distortion in Dataset*: A huge amount of missing data can cause distortions in the variable distribution i.e. it can increase or decrease the value of a particular category in the dataset.
3. *Affects the Final Model*: the missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.

Another and the most important reason is "We want to restore the complete dataset". This is mostly in the case when we do not want to lose any (more of) data from our dataset as all of it is important, & secondly, dataset size is not very big, and removing some part of it can have a significant impact on the final model.

1.3 Imputation methods and Advance imputers

1. Imputation of continuous variables using mean and median and categorical features using mode. These steps can be implemented using the Simple Imputer library.
2. Advance imputers like
    - K-nearest Neighbour Imputer
    - Iterative Imputer
    - Miss Forest Imputer, are used in this component. Detailed explanation about the imputers and its working are given a separate document.

1.4 Imputation component flow chart



1.5 Requirements to run the component

Installations:

pip install tk

pip install MissForest

pip install pmdarima

Command to run:

Imputation_Model.py

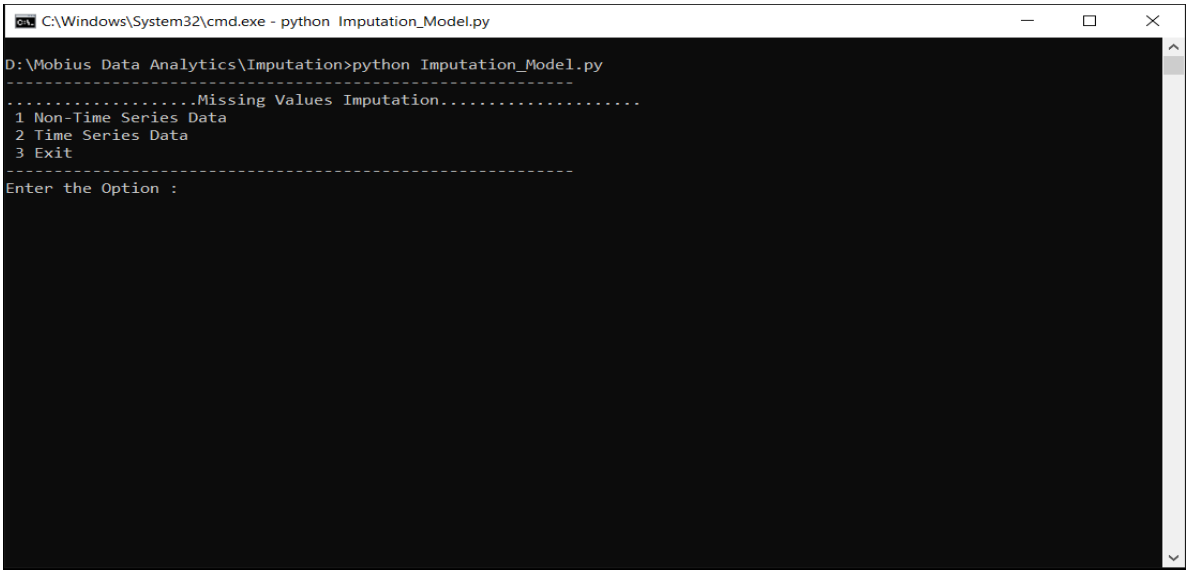
Datasets:

1. Non – Time Series data: Petromin_Non_Time_Series_data
2. Univariate Time Series data: Univariate Time series data
3. Multivariate Time Series data: Multivariate Time series data

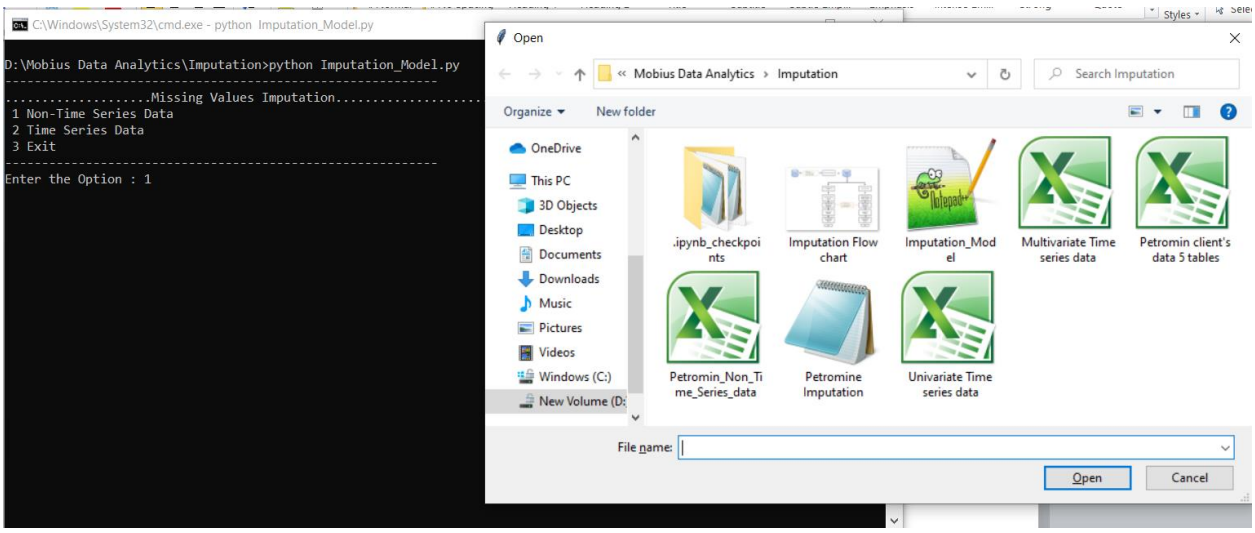
1.6 Steps:

Data imputation for Non – Time series Data:

1. Select Non – Time series data type by giving the input as 1



2. Get the data from the respective folder



3. Component displays the dimensions of the data, percentage of null values in the dataset, the data types and a statistical summary of the original data

```
C:\Windows\System32\cmd.exe

 Total Rows: 52561
 Total Columns: 13

--------Percentage of Missing values---------

   CustomerID : 0.0 %   ( int64 )
   Activity : 6.4 %   ( object )
   CustomerSegment : 0.09 %   ( object )
   DailyRun : 4.6 %   ( float64 )
   ExpectedOilType : 69.39 %   ( object )
   ExpectedService : 4.6 %   ( object )
   ExpectedServiceDate : 45.63 %   ( datetime64[ns] )
   Loyalty : 6.4 %   ( object )
   MultiBrand : 0.0 %   ( int64 )
   OwnershipSegment : 4.6 %   ( object )
   Satisfaction : 0.0 %   ( object )
   ServiceBehavior : 4.6 %   ( object )
   VINProfile : 4.6 %   ( object )

-------------Data Summary-------------


                    unique                          top   freq          mean   min       max
CustomerID             NaN                          NaN    NaN  73039.842278   6.0  157674.0
Activity                 5                     Inactive  20775           NaN   NaN       NaN
CustomerSegment         16   OCCASIONAL AND LOW SPENDER  24638           NaN   NaN       NaN
DailyRun               NaN                          NaN    NaN     76.508264  4.04     201.0
ExpectedOilType          2                      Mineral   8874           NaN   NaN       NaN
ExpectedService         62      Not expected to service  21565           NaN   NaN       NaN
ExpectedServiceDate    771          2022-05-11 00:00:00    301           NaN   NaN       NaN
Loyalty                  8                       Low_WB  13344           NaN   NaN       NaN
MultiBrand             NaN                          NaN    NaN      0.018474   0.0       1.0
OwnershipSegment         6            Losing engagement  26792           NaN   NaN       NaN
Satisfaction             4                      Unknown  44787           NaN   NaN       NaN
ServiceBehavior          7  Not recent and no engagement  12169           NaN   NaN       NaN
VINProfile               3                  Current car  43978           NaN   NaN       NaN
```

4. List of variables that can be dropped before imputation:
    - If the variables in the given dataset contains > 40% of missing values, that particular variable can be dropped before imputing.
    - Any date or date-time variables
    - Unique Identification variables like Customer ID, Account ID, Name, Employee No. etc.

```
C:\Windows\System32\cmd.exe
---------Drop unnecessary features----------
                      0
0            CustomerID
1              Activity
2        CustomerSegment
3              DailyRun
4        ExpectedOilType
5        ExpectedService
6    ExpectedServiceDate
7                Loyalty
8             MultiBrand
9       OwnershipSegment
10          Satisfaction
11       ServiceBehavior
12            VINProfile

Enter the number of features to be dropped:3
Enter the Feature index :0
Enter the Feature index :4
Enter the Feature index :6
```

5. Select the type of approach to evaluate the imputers
    - Choose Classification if the dataset contains only Categorical or mixed (Categorical and continuous) variables
    - Choose Regression if the dataset contains only Continuous variables.
6. After choosing the approach, select the target variable.
    - To run any ML model X and y separation is required.
    - X contains all independent variables and y contains the dependent variable.

- The ML model predicts the y using the X independent variables by splitting into train and test set. And then model evaluation score is displayed.

7. Evaluation scores of different imputers are displayed. The user can select the imputer with high ROC_AUC score.

```
C:\Windows\System32\cmd.exe

 Select Classification if the dataset has Mixed features (Both Numerical and Categorical)
 Select Regression if dataset has only Numerical features

 1 Classification
 2 Regression

Enter the option:1
                  0
0          Activity
1   CustomerSegment
2          DailyRun
3   ExpectedService
4           Loyalty
5        MultiBrand
6  OwnershipSegment
7      Satisfaction
8   ServiceBehavior
9        VINProfile
Enter the index of the target variable :5

-------- Evaluation scores for different Imputers---------

Roc_auc_score for KNN_Imputer:(k=5) 0.875
Roc_auc_score for KNN_Imputer:(k=7) 0.877
Roc_auc_score for KNN_Imputer:(k=9) 0.877
Roc_auc_score for Iterative_Imputer 0.89
Iteration: 0
Iteration: 1
Roc_auc_score for Miss_Forest_Imputer 0.887
```

8. Proceed to impute the missing values using the Advance imputer based on their performance.
   - After selecting the type of imputer, component asks for additional parameter. The user can choose the parameters from the given options
   - The component imputed the missing values and displays the count of null values and summary of data after imputation.

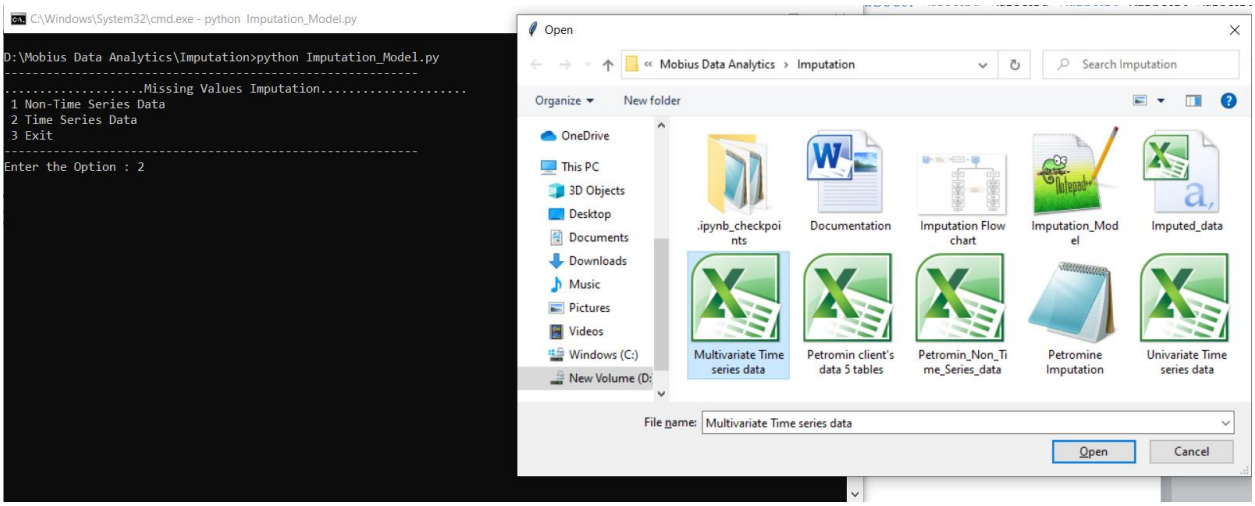
```
C:\Windows\System32\cmd.exe
Proceed to Impute the data? YES/NO : yes
List of Imputers:
 1 KNNImputer
 2 IterativeImputer
 3 MissForest
Enter the index of the Imputer:2
Enter the imputation order (descending or ascending or random):random

------------Count of null values after Imputation----------
 Activity            0
CustomerSegment     0
DailyRun            0
ExpectedService     0
Loyalty             0
MultiBrand          0
OwnershipSegment    0
Satisfaction        0
ServiceBehavior     0
VINProfile          0
dtype: int64

--------------Imputed Data Summary--------------

                      count                          top   freq          mean       max
CustomerID          52561.0                          NaN    NaN  73039.842278  157674.0
ExpectedOilType       16091                      Mineral   8874           NaN       NaN
ExpectedServiceDate   28577           2022-05-11 00:00:00    301           NaN       NaN
Activity              52561                     Inactive  22690           NaN       NaN
CustomerSegment       52561    OCCASIONAL AND LOW SPENDER  24638           NaN       NaN
DailyRun            52561.0                          NaN    NaN     76.058076     201.0
ExpectedService       52561       Not expected to service  21565           NaN       NaN
Loyalty               52561                         LowWB  13775           NaN       NaN
MultiBrand          52561.0                          NaN    NaN      0.018474       1.0
OwnershipSegment      52561             Losing engagement  27107           NaN       NaN
Satisfaction          52561                       Unknown  44787           NaN       NaN
ServiceBehavior       52561  Not recent and no engagement  12171           NaN       NaN
VINProfile            52561                   Current car  46358           NaN       NaN
```

9. Feedback activity:
    - After the first level imputation the imputed dataset is downloaded to the current working directory of the user.
    - *The user can compare the data summary before and after imputation and understand if the same behavioural pattern is been followed in the variables after imputation.*
    - If the user is not satisfied with the output, the user can input the option as Re-Run and impute the original dataset with different imputers and parameters.
    - After the second level imputation, the user can download the imputed dataset by giving input as 'yes' to download the data.

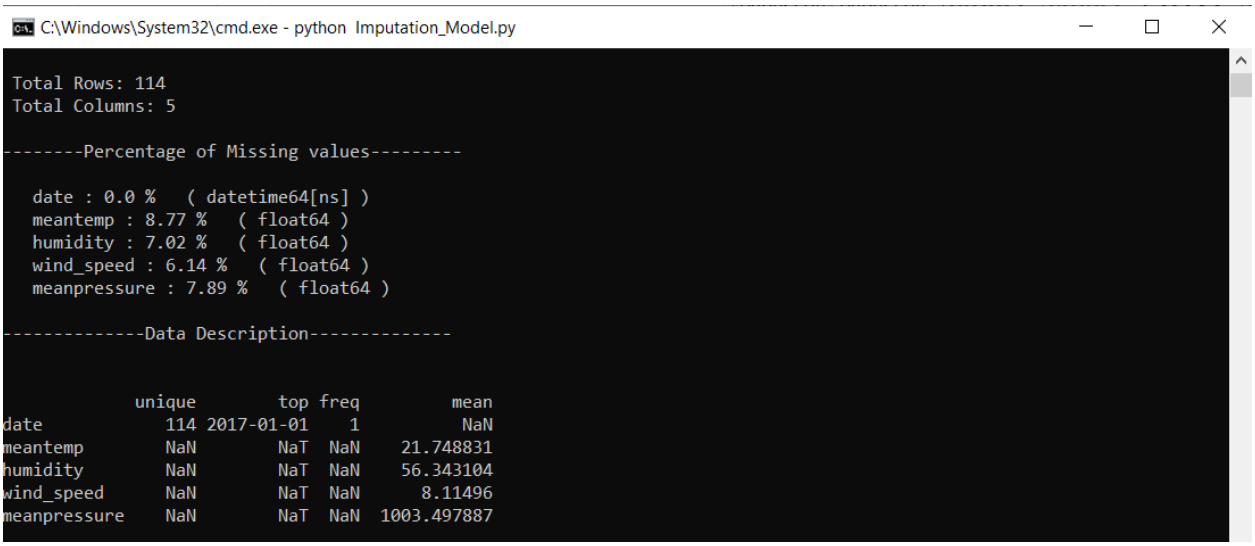
```
-------------Do you want to Exit or run the model with different parameters?-------------

Enter the Option(1. EXIT/n 2.RE-RUN): 2

Inpute the dataset using different Imputers / Parameters


Proceed to Impute the data? YES/NO : yes
List of Imputers:
 1 KNNImputer
 2 IterativeImputer
 3 MissForest
Enter the index of the Imputer:1
Enter the value for n_neighbors (5 or 7 or 9):5

------------Count of null values after Imputation----------
 DailyRun            0
MultiBrand          0
Activity            0
CustomerSegment     0
ExpectedService     0
Loyalty             0
OwnershipSegment    0
Satisfaction        0
ServiceBehavior     0
VINProfile          0
dtype: int64

-------------Imputed Data Summary-------------

                      count                          top   freq          mean       max
CustomerID          52561.0                          NaN    NaN  73039.842278  157674.0
ExpectedOilType       16091                      Mineral   8874           NaN       NaN
ExpectedServiceDate   28577           2022-05-11 00:00:00    301           NaN       NaN
DailyRun            52561.0                          NaN    NaN     74.968924     201.0
MultiBrand          52561.0                          NaN    NaN      0.018474       1.0
Activity              52561                     Inactive  24137           NaN       NaN
CustomerSegment       52561    OCCASIONAL AND LOW SPENDER  24687           NaN       NaN
ExpectedService       52561       Not expected to service  23984           NaN       NaN
Loyalty               52561                         LowWB  16706           NaN       NaN
OwnershipSegment      52561             Losing engagement  29209           NaN       NaN
Satisfaction          52561                       Unknown  44787           NaN       NaN
ServiceBehavior       52561  Not recent and no engagement  14586           NaN       NaN
VINProfile            52561                   Current car  46395           NaN       NaN

 Do you want to download the data? YES/NO : yes
Downloaded successfully!
```

Data imputation for Time series Data:

1. Select Non – Time series data type by giving the input as 2
2. Get the data from the respective folder



3. Component displays the dimensions of the data, percentage of null values in the dataset, the data types and a statistical summary of the original data



4. List of variables that can be dropped before imputation:
    - If the variables in the given dataset contains > 40% of missing values, that particular variable can be dropped before imputing.
    - Unique Identification variables like Customer ID, Account ID, Name, Employee No. etc.
5. For time series data, date or data-time variable has to be set as index. The user can input the Date variable to convert it to index.

```
C:\Windows\System32\cmd.exe - python  Imputation_Model.py

             unique        top freq         mean
date            114 2017-01-01    1          NaN
meantemp        NaN        NaT  NaN    21.748831
humidity        NaN        NaT  NaN    56.343104
wind_speed      NaN        NaT  NaN      8.11496
meanpressure    NaN        NaT  NaN  1003.497887

---------Drop unnecessary features----------
              0
0          date
1      meantemp
2      humidity
3    wind_speed
4  meanpressure

Enter the number of features to be dropped:0
              0
0          date
1      meantemp
2      humidity
3    wind_speed
4  meanpressure

 Select the index of the Date column to set as index

Enter the index:0
date
```

6. User can select the type of time series data.
    - Choose Univariate if only one time dependent variable has to be predicted
    - Choose Multivariate if more than one time dependent variables has to be predicted

```
C:\Windows\System32\cmd.exe - python  Imputation_Model.py

 Is the imported Time series data Univariate or Multivariate?

 1 Univariate
 2 Multi-variate
 3 Exit
Enter the Option :
```

7. Evaluation scores of different imputers are displayed.
    - Auto Arima is used to evaluate Univariate data
    - Vector Error Correction Modelling is used to evaluate Multivariate time series data
    - The evaluation scores of different Imputers are displayed. User can choose the imputer with least RMSE score.
    -
8. Proceed to impute the missing values using the Advance imputer based on their performance.
    - After selecting the type of imputer, component asks for additional parameter. The user can choose the parameters from the given options
    - The component imputed the missing values and displays the count of null values and summary of data after imputation.

```
Select C:\Windows\System32\cmd.exe
-------- Evaluation scores for different Imputers---------

RMSE for the KNN_Imputer:(k=5) [4.428, 5.959, 2.237, 3.37]
RMSE for the KNN_Imputer:(k=7) [4.427, 5.851, 2.101, 3.443]
RMSE for the KNN_Imputer:(k=9) [4.477, 5.655, 2.115, 3.516]
RMSE for the Iterative_Imputer [4.111, 5.644, 2.255, 3.846]
Iteration: 0
Iteration: 1
Iteration: 2
Iteration: 3
Iteration: 4
RMSE for the Miss_Forest_Imputer [4.807, 5.753, 2.184, 3.935]

Proceed to Impute the data? YES/NO : yes
List of Imputers:
 1 KNNImputer
 2 IterativeImputer
 3 MissForest
Enter the index of the Imputer:1
Enter the value for n_neighbors (5 or 7 or 9):5

------------Count of null values after Imputation----------
 meantemp        0
humidity        0
wind_speed      0
meanpressure    0
dtype: int64

--------------Imputed Data Summary--------------

               count         mean      min          max
meantemp      114.0    21.772506  11.0000    34.500000
humidity      114.0    56.737119  17.7500    95.833333
wind_speed    114.0     8.134526   1.3875    19.314286
meanpressure  114.0  1004.030870  59.0000  1022.809524


**************The missing values in the given dataset is imputed and downloaded! Check the data!! **************
```

9. Feedback activity: The feedback mechanism is similar to Non-time series data. Refer to point 9 in Page no. 6 to understand the feedback activity.

```
Select C:\Windows\System32\cmd.exe

-------------Do you want to Exit or run the model with different parameters?-------------

Enter the Option(1. EXIT/n 2.RE-RUN): 2

Inpute the dataset using different Imputers / Parameters


Proceed to Impute the data? YES/NO : yes
List of Imputers:
 1 KNNImputer
 2 IterativeImputer
 3 MissForest
Enter the index of the Imputer:2
Enter the imputation order (descending or ascending or random):random

------------Count of null values after Imputation----------
 meantemp        0
humidity        0
wind_speed      0
meanpressure    0
dtype: int64

--------------Imputed Data Summary--------------

               count         mean      min          max
meantemp      114.0    21.707512  11.0000    34.500000
humidity      114.0    56.416158  17.7500    95.833333
wind_speed    114.0     8.122280   1.3875    19.314286
meanpressure  114.0  1003.531545  59.0000  1022.809524

D:\Mobius Data Analytics\Imputation>
```