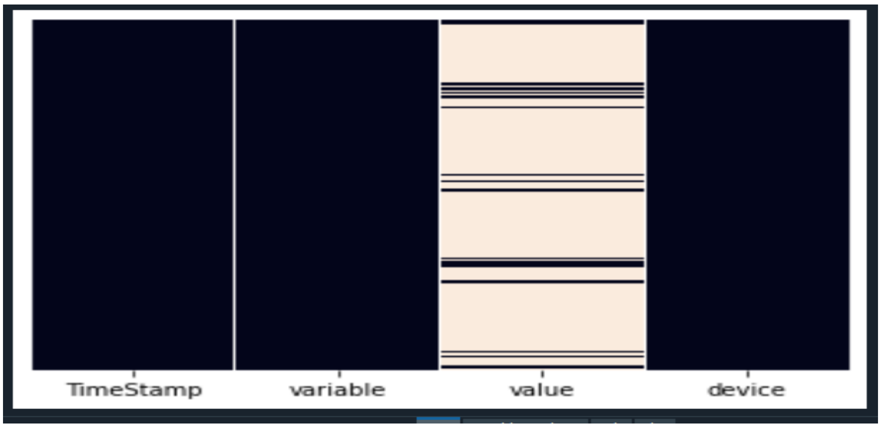
Summary Report

Name – Saurabh L. Bonde

Email – saurabh.090395@gmail.com

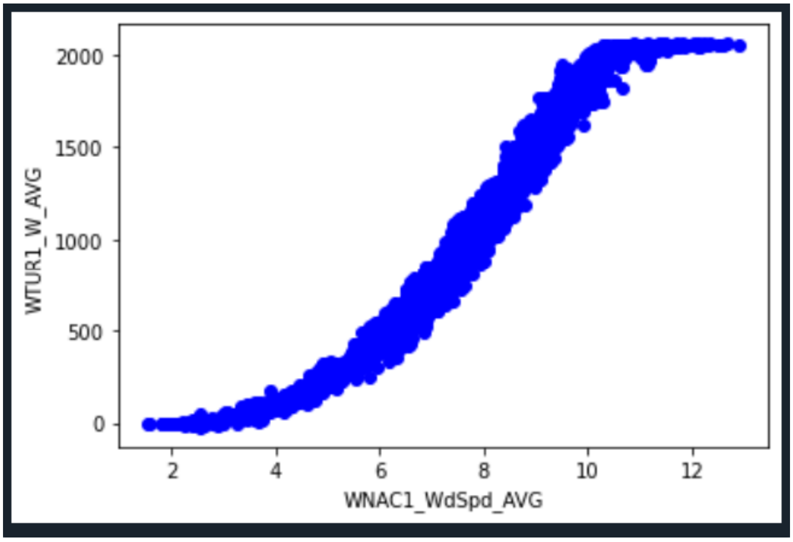
Phone - 7798748934

- Data processing Method :
    1. Importing a Library.
    2. Created a list called “filename” containing names of all the files to be read.
    3. Using For loop read the files & concatenate them into the dataframe “data”.
    4. Applying basic operation on dataset to understand the data.
    5. Using seaborn visualise the raw data which help to identify NAN value easily.



6. “variable” is the column with categorical data, so performs basic operations like unique(), describe() etc. to get the sense of number of variable & their count.
7. Change the data type of the TimeStamp Column which is required to aggregate the data by 10 minute timeframe.
8. Sorting dataset for proper timeseries aggregation of the data.
9. Defined new dataframe to contain processed data “agg_10m”
10. Using Groupby function grouped a data with grouper key = “TimeStamp” & frequency = 10, aggregate function used to calculate average, minimum, maximum & Standard deviation except last value which calculated directly from dataframe function.
11. After 10 minute aggregation of the dataset the TimeStamp is converted from index to column as it has converted to index, cause we have used it as a key in grouper.
12. Adding new column to the dataset as per requirement using tz_localised & dt.tz_comvert function
13. Taking care of NaN values

14. Created variable var1 & var2 to find the scattered plot x & y i.e. WNAC1_SdSpd_AVG & WTUR1_W_AVG , visualised a scattered plot.



- Issues need to be careful about :

    1. NaN value need to be taken care at the end, otherwise It could have been caused calculation error in average, minimum & standard deviation.
    2. Need to define data-frame variable otherwise, it could give error while accessing pandas function over the data
    3. Data could not be processed without sorting it.
    4. While using groupby function it is always better to perform each operation separately, which help while adding operated data column to the data-frame
    5. Most important issue is dataset is huge and machine hang while working on jupyter notebook so preferred spyder still hangs but comparatively less.

- Aggregation doesn't make sense:

After calculating mean, minimum & maximum there is no need to calculated standard deviation, or vice versa if we calculated mean & standard deviation, make process redundant for same evaluation.

- Data Quality measure:

    1. Incompleteness – there are lots of NaN values in value column of the data
    2. Inaccuracy – Not Occurred
    3. Inconsistency – Data was not consistent as per Time-Stamp
    4. Invalid – Not occurred
    5. Non Standard – data is in non-standard format.