Exercise 7

Note: This exercise should be done by forming a group of 2 or 3 (not exceeding 3). All groups should register themselves using the form: https://tinyurl.com/ds203-e7-group (submissions will NOT be assessed if this form is not filled)

Due date : October 10, 2023, 23:55 Hrs.

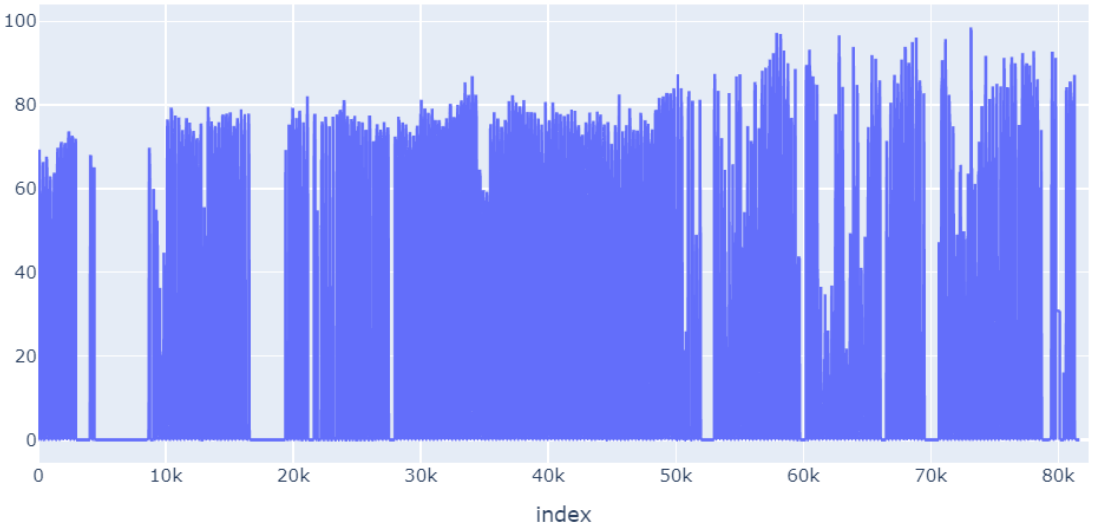
Marks: 20; Weightage: 15%

Data to be used: e7-htr-current.csv

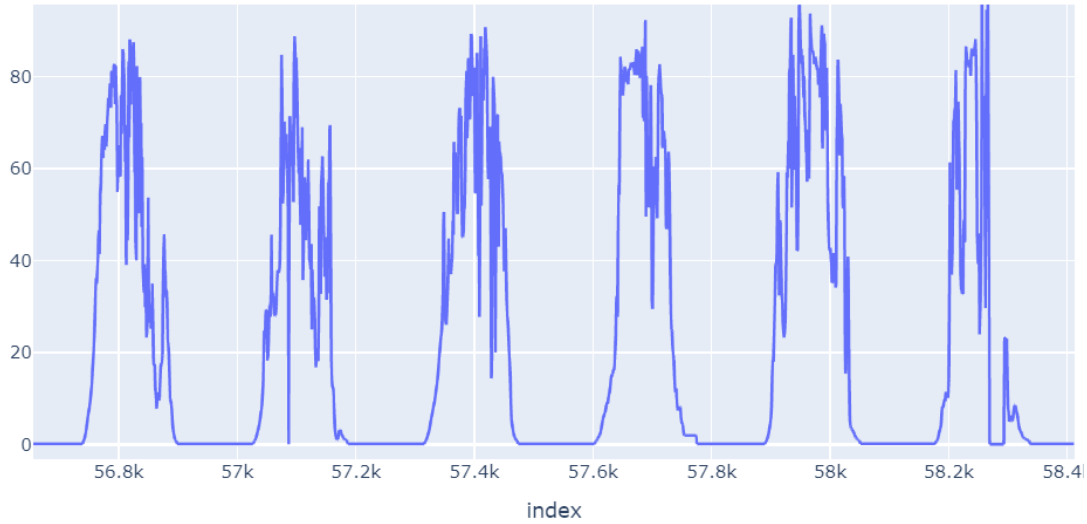
Data Description

- Transformer current data, sampled every 5 minutes
- Data source: a solar power generation site

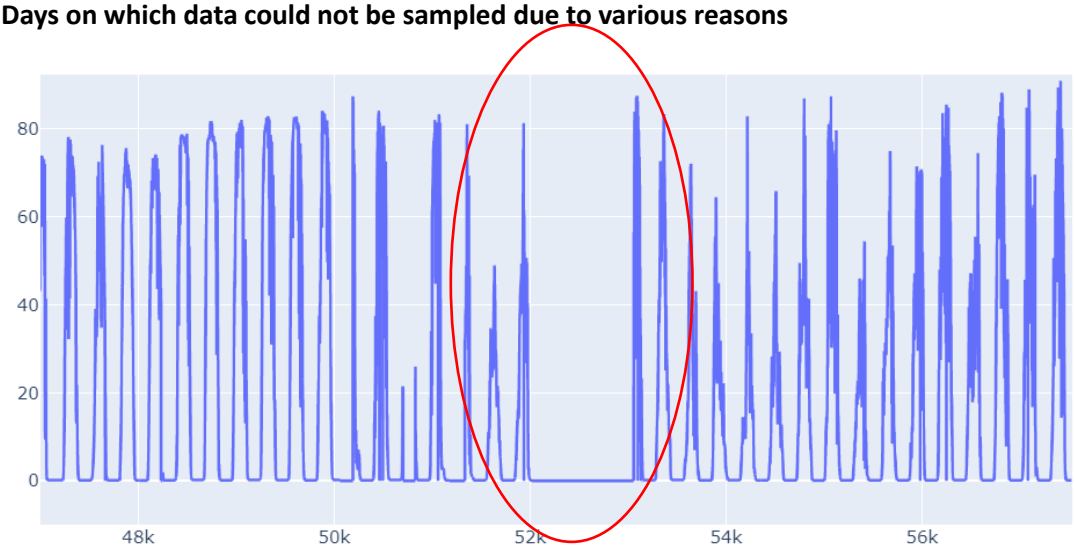
Line plot of all data points (the complete data – about 280 days).



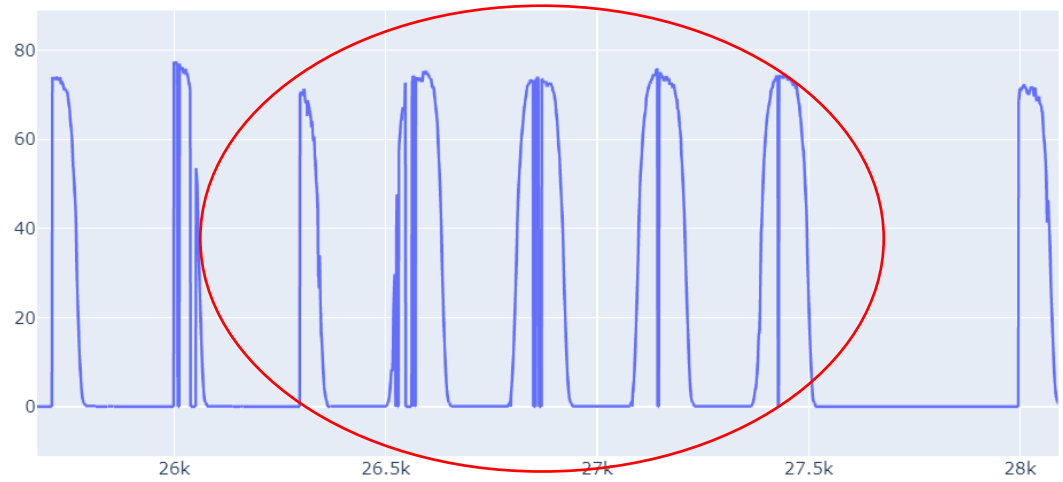
Days on which the data was choppy (bad) due to various reasons



Days on which data could not be sampled due to various reasons



Relatively good days! There may be slightly noisy readings, but seemingly easy to fix.



The challenge:

- Can the data be “de-noised”? That is, can the data set’s quality be improved with the help of the data itself?

The task: Data Quality Improvement

1. Identify the ‘good’, ‘bad’ and ’missing’ days
2. Make the ‘good days’ ‘really good’, by fixing the observed data problems
3. Use data from about 80% of the ‘good’ days to create an ML model.
4. Validate the model on the remaining 20% of the “good” days (How will you validate? What metrics will you use for validation?)
5. Use the ML model thus created to “fix” the data from the “bad” and “missing” days

6. Create appropriate metrics, and a plot of the "improved data set", and compare it – using the metrics, and visually - with the original. Highlight your achievements, and justify the methods and steps that you employed to achieve the results.
7. Create an exhaustive presentation (slide deck) that includes:
    a) A description of the problem (**1-2 slides**)
    b) An executive summary of your solutions and achievements (**1-2 slides**)
    c) Slides outlining every major step, every major decision taken, every tool / automation used, every major outcome /observation – with crisp justification / explanation (**as many slides as required – see the evaluation criteria mentioned in the Table below**)
    d) The reasons why you think your approach is the best one (**one slide**)
    e) One alternate approach that could be equally good, in your opinion (**one slide**).
    f) The major challenges faced and how they were overcome (**1-2 slides**)
    g) Can such an "improved data set" be used for creating ML models? Justify your answer. (**1-2 slides**)
8. The presentation should be adequately illustrated by using appropriate figures / graphs / tables / analysis. Please DO NOT state the obvious!
9. Submit the following:
    a) PDF of your Jupyter Notebook(s) (after removing all the debug dumps)
    b) PDF of your presentation
    c) Group details using the form: **https://tinyurl.com/ds203-e7-group** (your submission will NOT be assessed if group details are not submitted)

Do not forget to include the roll numbers of all group members in the file names!

Assessment criteria

Criteria	**Marks**
Data analysis and diagnosis	5
The problem-solving steps, and the outcome(s)	5
The level of automation (programming) used to achieve the results	5
Quality of the presentation (adherence to points '7' through '9')	5

Suggestions and hints:

- Understand and use the Python library **plotly** for interactive data exploration.
- The data covers about 280 days of operation of the transformer. Treat the data for one day as one set of observations – so you have about 200 sets of observations (after accounting for completely missing data) that can be used to predict the 'current' through the transformer at any given 'active' time of the day.
- What independent features will be required to predict the 'current' through the transformer at any given 'active' time of the day? Feature Engineering required?
- How can you automate the process of identifying 'good', 'bad' and 'missing data' days?
- How can you use **descriptive statistics** to assist in some of the analysis steps?

***