TAKE-HOME ASSESSMENT

Objective:

To preprocess the data and create a visualization using the preprocessed dataset.

Dataset:

You must download and extract the zip file. The zip file contains CSV files that have 5 days of PR and GHI data, organized into directories based on the start month and parameter (PR or GHI).

PR (Performance Ratio) – This parameter is used to track the daily performance of the PV plant. A high value indicates that the plant is performing well and there are no issues.

GHI (Global Horizontal Irradiance) – This parameter tracks the total irradiation for a particular day. A high value indicates a sunny day.

Below is an example of the folder structure:

PR/

2023-01/

2023-01-01_PR.csv

2023-01-06_PR.csv

GHI/

2023-01/

2023-01-01_GHI.csv

2023-01-06_GHI.csv

Link:

https://drive.google.com/file/d/1KdpHt7GVtWUAH9vvJMNgBSfQ8ochrsx1/view?usp=sharing

Data preprocessing:

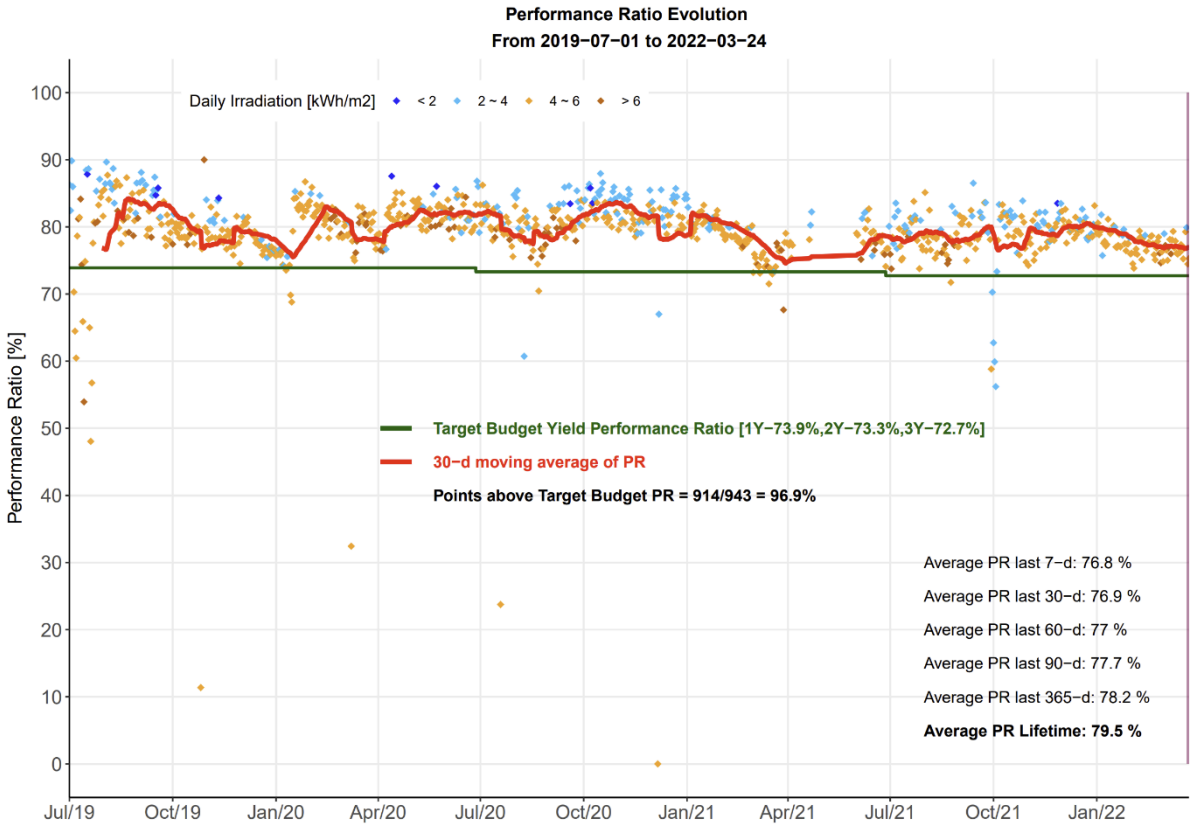
Generate a **single CSV** file containing all the data from both the PR and GHI folders. The new CSV file should contain 3 columns: Date, GHI, PR.

Important Notes:

- The data has to be collated into a single file. The file should contain 982 rows.
- Create a single function to preprocess the data. Make sure you organize your code and it is readable.

Data Visualization:

Once the data has been processed, the below graph must be generated using the preprocessed data:



Important Notes:

- Create a single function to generate the graph. Make sure you organize your code and it is readable.

- The **red line** on the graph represents the 30-d moving average of the PR (Performance Evolution) whereas the scatter points depict the PR value of that day, shaded (colored) with GHI based on GHI value.

- The **red line** on the graph represents the 30-d moving average of the PR (Performance Evolution) whereas the scatter points depict the PR value of that day, shaded (colored) with GHI based on GHI value.

- The **dark green line** represents the budget line. The value begins from 73.9 and should reduce by 0.8% every year (**Do not hardcode the values**). As you can see, the values are:

    - 73.9 for the first year (July 2019 to June 2020)
    - 73.3 for the second year (July 2020 to June 2021)

    - 72.7 for the third year (July 2021 to present)

  This should happen dynamically via code.

- The points for the scatter plot are **color-coded** (as per the legend above). That is: if the GHI [Daily Irradiation] is:
    - Less than 2: Navy blue
    - 2-4: Light blue
    - 4-6: Orange
    - >6: Brown

- The points above **Target Budget PR** represent the number of PR points above the Budget PR for that particular year.

- The **bottom right section** of the graph simply shows the average PR for the last 7 days, the last 30 days, the last 60 days, and so on.

- Please note that the values and the trends will not match the graph exactly since we have changed the data slightly.

Bonus Points:

Enhance the script to accept start and end date arguments for generating a PR graph based on the specified date range. For example, you could run the script with --start_date 2024-01-01 --end_date 2024-06-30 to visualize PR data between January 1, 2024, and June 30, 2024.

Files that you have to submit:

- A single CSV file containing all data points

- The completed output graph generated

- The code you used to generate the CSV and graph (preferably in Python)