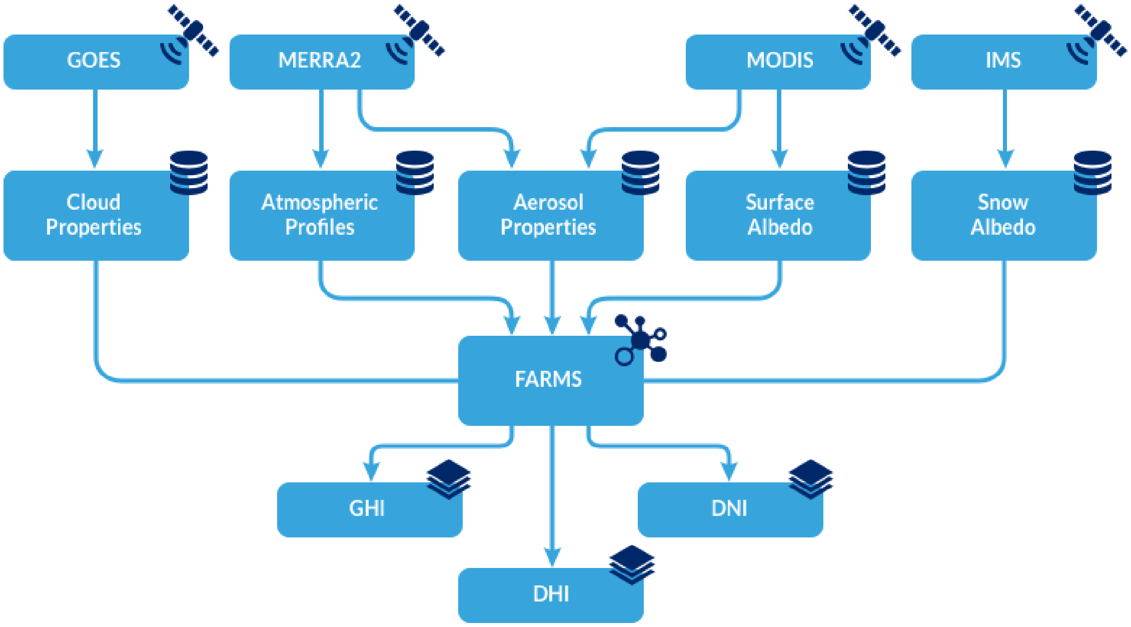
Exploratory Data Analysis (EDA) of the National Solar Radiation Database (NSRDB) Data

1. Introduction

Exploratory Data Analysis (EDA) is a technique used to analyze datasets by summarizing their main characteristics, often with visual methods. EDA serves to understand the data, discern relationships between variables, identify outliers, missing values, and other data quality issues. It's a critical step in the data analysis process, guiding the modeling approach and highlighting key variables. In this project, we utilize the National Solar Radiation Database (NSRDB), celebrated for its detailed temporal and spatial resolution, and comprising crucial measurements such as global horizontal (GHI), direct normal (DNI), and diffuse horizontal irradiance (DHI), along with other meteorological data. Compiled by the National Renewable Energy Laboratory's (NREL) Physical Solar Model (PSM), and integrating data from NOAA's GOES satellites, NIC's IMS, and NASA's MODIS and MERRA-2, the NSRDB employs the PSM's computations of GHI which take into account atmospheric conditions and cloud data via the FARMS algorithm. The aim is to harness this dataset to assess solar energy potential across the United States and to identify prime regions for solar power generation. Additionally, we intend to develop a predictive model using the NSRDB data to forecast solar radiation, facilitating the optimization of solar energy production.



1.1. Data

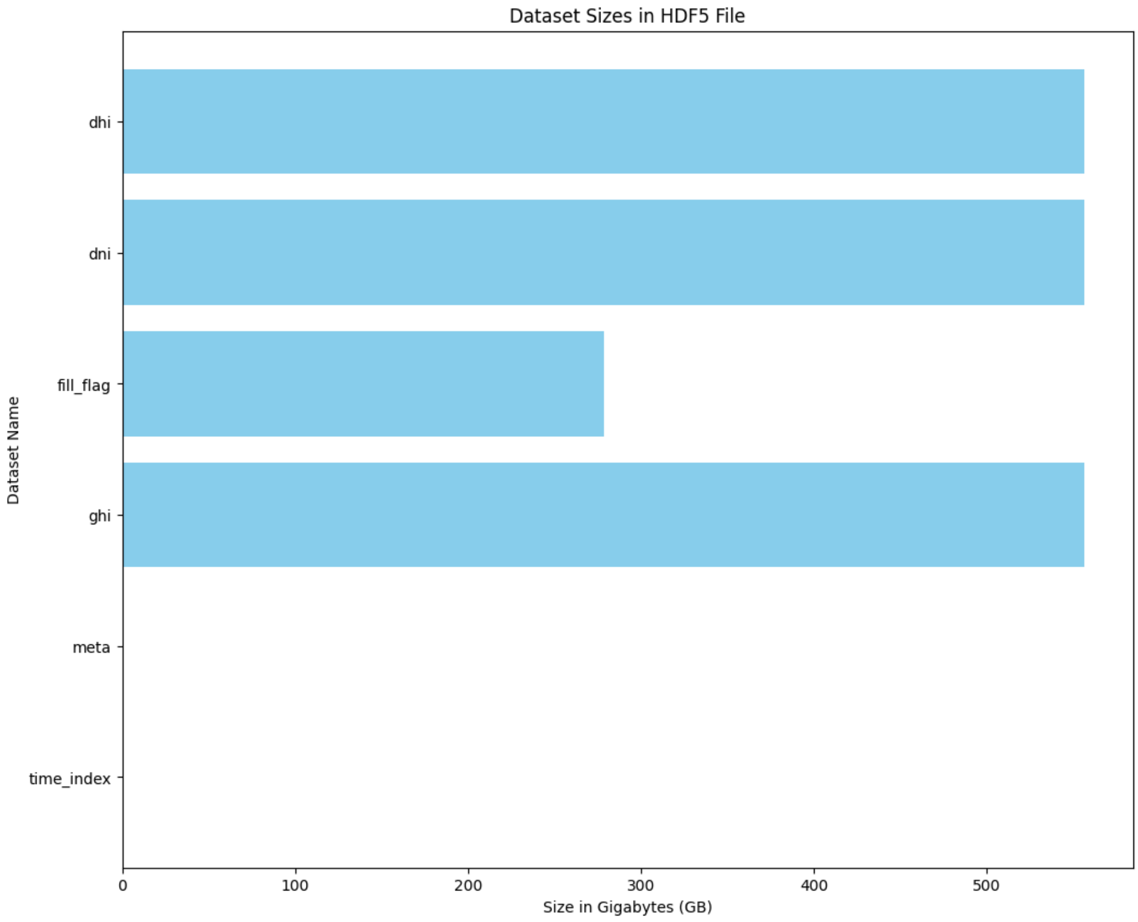
The dataset is available in various intervals and resolutions, with the project data for 2022 provided in 5-minute intervals at a 2km resolution. Available for download in CSV format from the NREL website, for this project, we opt for the HDF format, obtained from data.openei.org which offers unfettered download access. It's noteworthy that each file is roughly 2GB in size and is supplied in the h5 format, a binary Hierarchical Data Format designed for the storage of extensive scientific data in multidimensional arrays, becoming a standard for large datasets.

1.2 Data Description

The selected dataset for this project includes the 2022 CONUS (Continental United States) data, accessible in 5-minute intervals at a 2km resolution (`nsrdb_conus_irradiance_2022.h5`). The file comprises various datasets:

- `ghi` : Global Horizontal Irradiance (W/m^2), 556.608 GB.
- `dni` : Direct Normal Irradiance (W/m^2), 556.608 GB.
- `dhi` : Diffuse Horizontal Irradiance (W/m^2), 556.608 GB.
- `fill_flag` : Indicates the presence of missing data, 278.304 GB.
- `meta` : Metadata including location, elevation, and timezone information, 0.323 GB.
- `time_index` : Timestamps of the data entries, 0.002 GB.

The data is organized hierarchically, with datasets stored at the root level. The total size of the h5 files is 2TB, and the size of each dataset is illustrated below.



Below are the more information about the dataset and its structure:

1. `dhi` **(Diffuse Horizontal Irradiance)**: This dataset records the amount of solar radiation received per unit area by a surface horizontal to the ground, diffused or scattered by the atmosphere. It's essential for evaluating the solar energy available in shaded areas or on cloudy days. The shape `(105120, 2842719)` indicates the dataset contains 105,120 time intervals (entries over time) and data for 2,842,719 geographic locations, with data type unsigned 16-bit integer (`<u2`).

2. `dni` **(Direct Normal Irradiance)**: Represents solar radiation received per unit area by a surface that is always held perpendicular (normal) to the rays coming from the direction of the sun. This measure is crucial for concentrating solar power (CSP) and concentrating photovoltaic (CPV) technologies. Like `dhi`, it has a shape of `(105120, 2842719)`, indicating the same dimensionality and data type.

3. `ghi` **(Global Horizontal Irradiance)**: This key refers to the total amount of shortwave radiation received from above by a surface horizontal to the ground. This measure includes both the direct sunlight and the diffuse sunlight scattered by the atmosphere. It is a critical parameter for photovoltaic systems and for estimating

solar energy potential. Its data structure is similar to `dhi` and `dni`, with a shape of `(105120, 2842719)` and data type `<u2`.

4. **`fill_flag`**: This dataset indicates the quality or status of the other data points (`dhi`, `dni`, `ghi`). It might flag data points that are estimated rather than measured, or indicate the presence of errors or anomalies. Its shape `(105120, 2842719)` matches the other datasets, but it uses an unsigned 8-bit integer type (`|u1`), suitable for flag values that typically don't require large numerical ranges.

5. The `meta` dataset in the HDF5 file contains essential metadata for each geographic location included in the National Solar Radiation Database (NSRDB). This metadata provides critical context for the irradiance data (`dhi`, `dni`, `ghi`) and other measurements, allowing for precise geographical and environmental characterization. Here's a breakdown of the metadata fields:

   A. **`latitude` (`<f4`)**: Represents the latitude of the location in degrees, using a 32-bit floating-point format. Latitude values indicate how far north or south a place is relative to the equator.

   B. **`longitude` (`<f4`)**: Represents the longitude of the location in degrees, also in a 32-bit floating-point format. Longitude values indicate how far east or west a place is from the prime meridian.

   C. **`elevation` (`<i2`)**: The elevation of the location above sea level, measured in meters, using a 16-bit integer format. Elevation can significantly affect solar irradiance levels, with higher altitudes generally receiving more solar energy.

   D. **`timezone` (`<f4`)**: The time zone of the location, represented as hours offset from Coordinated Universal Time (UTC), in a 32-bit floating-point format. Time zone information is crucial for aligning time series data with local solar time.

   E. **`country` (`S36`)**: The country where the location is situated, encoded as a fixed-length string of 36 characters. This information is essential for regional analysis and comparison.

   F. **`state` (`S30`)**: The state, province, or equivalent administrative region of the location, encoded as a fixed-length string of 30 characters. State-level information can be useful for more localized solar energy studies.

   G. **`county` (`S38`)**: The county or equivalent administrative subdivision of the location, encoded as a fixed-length string of 38 characters. County information allows for even more granular analysis of solar irradiance data.

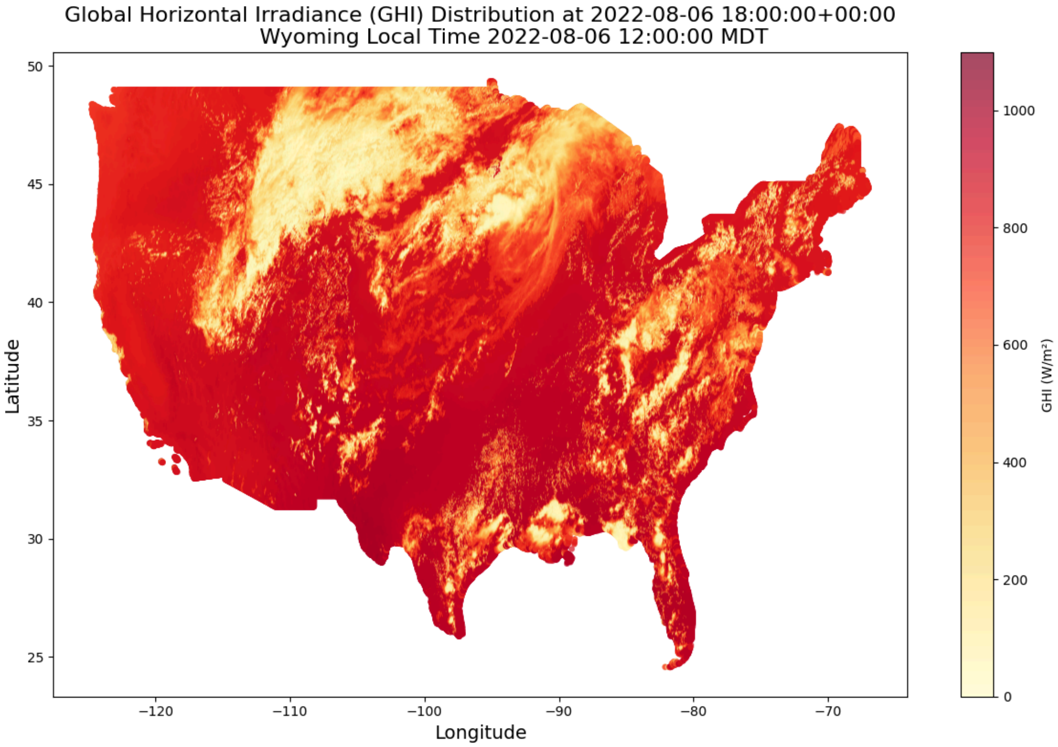
   H. **`gid_full` (`<i4`)**: A unique identifier for the geographic location, using a 32-bit integer format. This identifier can be used to cross-reference locations within the dataset or with external datasets.

6. `time_index` : Provides the timestamps for the data entries, likely in a string or datetime format, indicated by the type `|S25` (a fixed-length string of 25 characters). Its shape `(105120,)` shows there are 105,120 time points across the dataset, correlating with the temporal resolution of the other data keys.

Data Structure

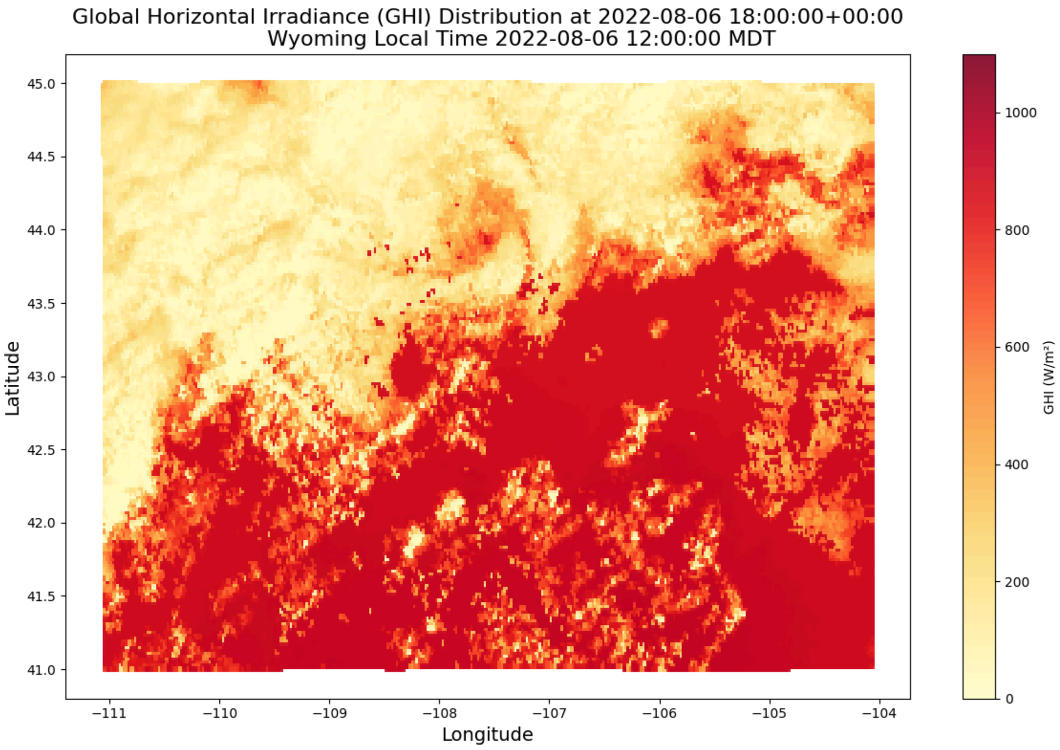
2. Visualizing the Data

The NSRDB dataset is vast, with millions of data points across the continental United States. To make sense of this data, we'll start by visualizing a small subset of the data to understand the structure and characteristics of the dataset. The below plot shows the solar radiation for entire United States for 6th Agust 2022 at 12:00 PM (UTC).

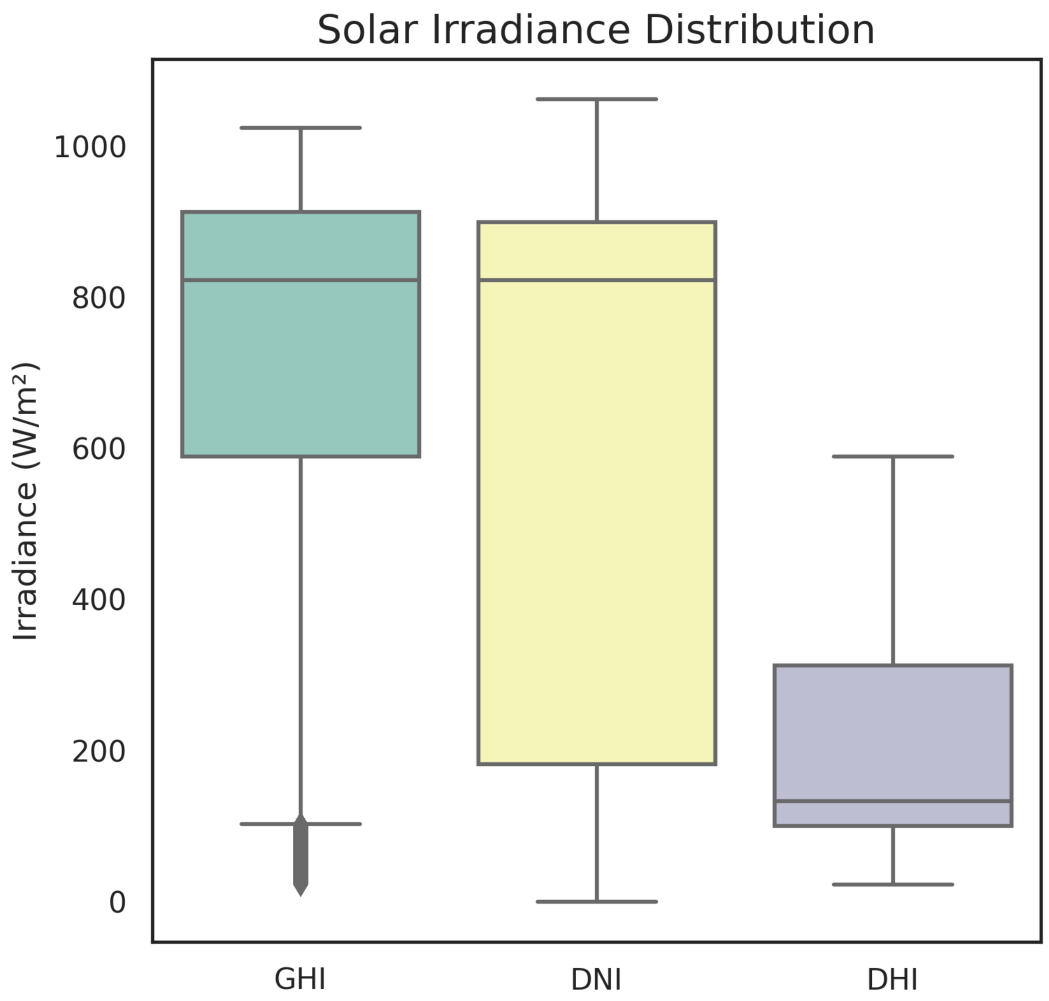

Global Horizontal Irradiance (GHI) Distribution at 2022-08-06 18:00:00+00:00
Wyoming Local Time 2022-08-06 12:00:00 MDT
Latitude
50
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
GHI (W/m²)
1000
800
600
400
200
0

For obtaining the plot mentioned above, we have joined the `ghi` dataset with the `meta` dataset to acquire the latitude and longitude of each location, and filtered the data based on the time_index and country, excluding Hawaii and Alaska. For memory and time efficiency, we have first filtered data in the metadata and then joined the `ghi` dataset with the filtered metadata.

Also if we want to visualize the solar radiation for Wyoming, for the same time, we will get the below plot.


Global Horizontal Irradiance (GHI) Distribution at 2022-08-06 18:00:00+00:00
Wyoming Local Time 2022-08-06 12:00:00 MDT
Latitude
45.0
44.5
44.0
43.5
43.0
42.5
42.0
41.5
41.0
Longitude
−111
−110
−109
−108
−107
−106
−105
−104
GHI (W/m²)
1000
800
600
400
200
0

Because of the vastness of the dataset, we will be working with a subset of the data to perform our analysis. Solar irradiance distribution for above mentioned time and date is shown in the plot above in box plot below for `ghi, dni, dhi` for the entire United States.

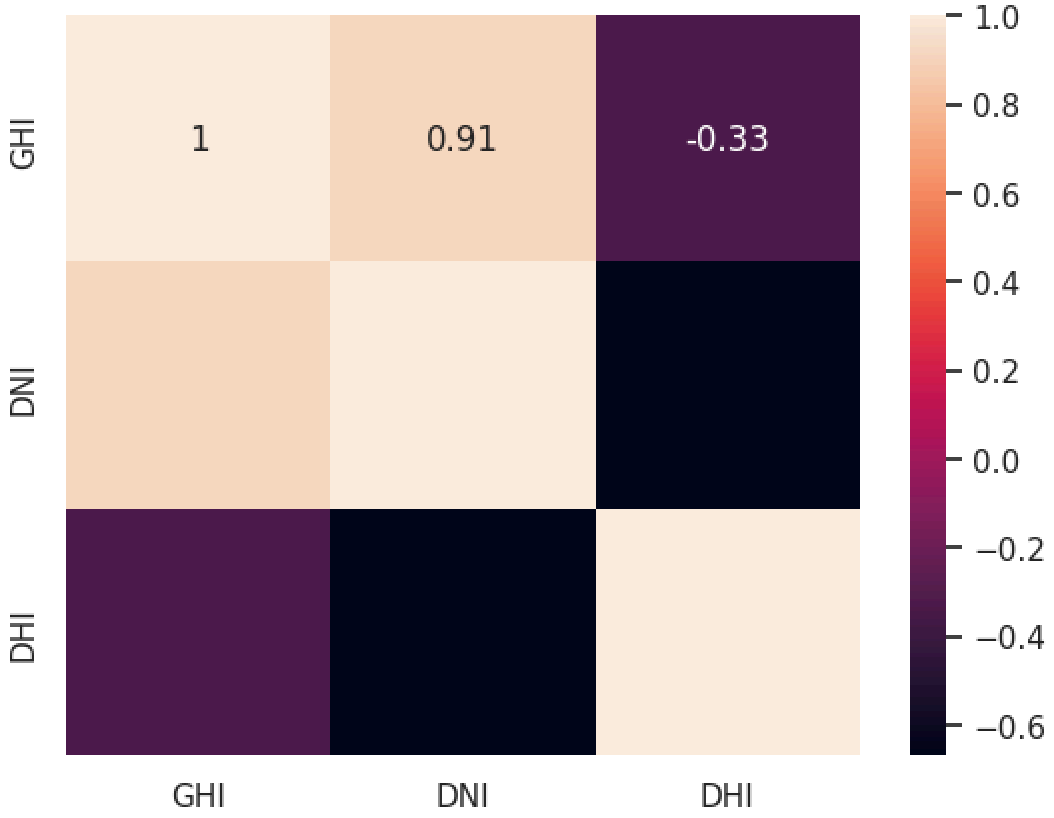


2.1 Correlation Between Solar Irradiance Measures

A Correlation Heatmap is a graphical representation of a correlation matrix, which shows the correlation coefficients between a set of variables. In your case, with `ghi, dni, dhi`, a Correlation Heatmap would illustrate how these three different measures of solar irradiance are related to each other.

Each cell in the heatmap would contain a value between -1 and 1, representing the Pearson correlation coefficient for the pair of variables it represents. Here's what the values mean:

A value close to 1 implies a strong positive correlation: as one variable increases, the other variable tends to also increase. A value close to -1 implies a strong negative correlation: as one variable increases, the other variable tends to decrease. A value around 0 implies no correlation: the variables do not have a strong relationship in how they change together.



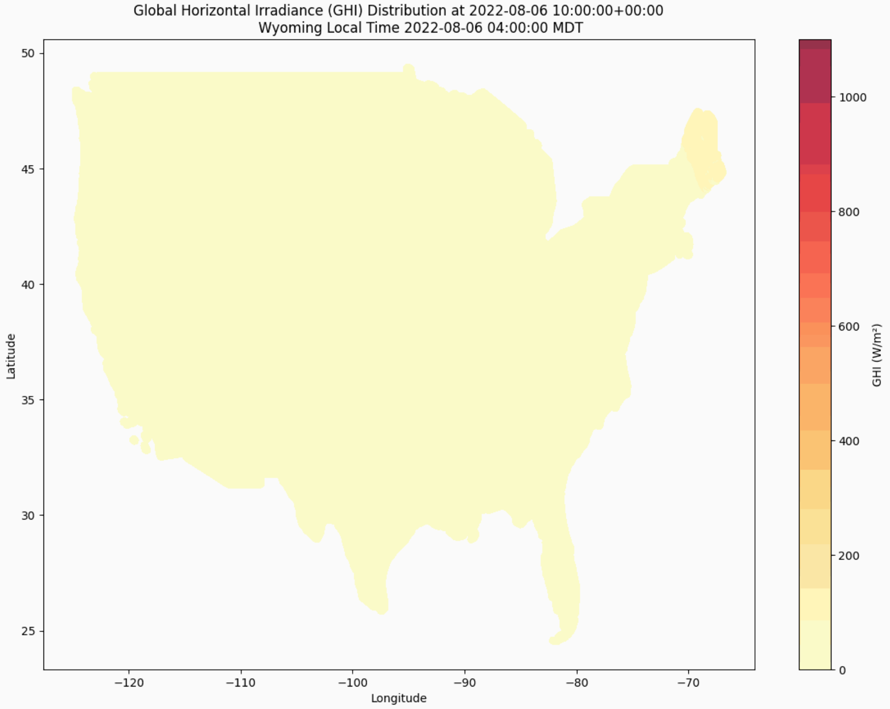
2.2 Frequency of Values in Each Dataset

The distribution of `ghi`, `dni`, and `dhi` is represented through the density of data points at various irradiance levels, blending elements of both box plots and kernel density plots within a Violin Plot.

The broader sections of the violin signify areas where data is more densely concentrated, indicating a higher frequency of irradiance values. Conversely, the narrower sections depict areas of lower data concentration, implying that fewer data points fall within those specific ranges of irradiance. Within the violin, an embedded box plot delineates the interquartile range, with a distinct line representing the median value.

2.3 Time Series Analysis

Upon visualizing the data across the United States over a 24-hour period, we obtain the plot depicted below. This visualization specifically leverages the `ghi` (Global Horizontal Irradiance) dataset. It provides a granular view of how solar irradiance fluctuates throughout a single day, reflecting the natural cycle of sunrise, peak solar noon, and sunset. By examining this time series plot, we can identify patterns of solar energy availability, including the times of day when irradiance begins to increase, reaches its maximum, and subsequently decreases. This analysis is crucial for understanding solar energy potential and optimizing solar panel operation to maximize energy capture during peak sunlight hours. Furthermore, it can shed light on the impact of geographical location and atmospheric conditions on solar irradiance levels across different regions of the United States.


Global Horizontal Irradiance (GHI) Distribution at 2022-08-06 10:00:00+00:00
Wyoming Local Time 2022-08-06 04:00:00 MDT
Latitude
50
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
GHI (W/m²)
0
200
400
600
800
1000

3. Conclusion

The National Solar Radiation Database (NSRDB) offers a valuable compilation of solar radiation measurements, encompassing global horizontal (GHI), direct normal (DNI), and diffuse horizontal irradiance (DHI). This extensive dataset is instrumental for gauging solar energy potential throughout the United States and pinpointing regions optimal for solar power infrastructure. Through exploratory data analysis (EDA) of the NSRDB, valuable insights can be extracted about the variance and distribution of solar irradiance, the interrelationships between different types of solar radiation, and the temporal dynamics of solar energy availability. Such analysis aids in identifying prime locations for solar power installations and in fine-tuning the operation of solar panels to maximize energy harvest. Furthermore, this project endeavors to develop a predictive model that utilizes GOES-16 satellite data to estimate solar radiation levels. This predictive tool has the potential to enhance solar radiation forecasting and thereby optimize solar energy production. Nonetheless, challenges arise in managing the voluminous dataset and executing the analysis with efficiency and efficacy. Visualizing data for specific locations and times further complicates matters, given the dataset's breadth and the computational demands of time and memory resources.