Project Report

B5W0: Solar Data Discovery

Gashaye Adugna

1. Introduction

The objective of this project is to analyze solar datasets from various countries, assessing solar potential and enabling cross-country comparisons. The project is structured into three main tasks: setting up the development environment using Git, conducting data profiling and exploratory data analysis (EDA), and synthesizing insights through comparative analysis of cleaned datasets. Initially, the team focused on familiarizing themselves with version control, creating a GitHub repository, establishing a local development environment, and configuring automated workflows for continuous integration. Essential files were added, and the setup process was documented in a README for future reference.

The second phase involved profiling, cleaning, and exploring the solar datasets to prepare them for analysis. Each country's data was handled in dedicated branches, where summary statistics, outlier detection, and time series visualizations were performed. The third task synthesized cleaned datasets from Benin, Sierra Leone, and Togo for comparative analysis, utilizing boxplots and statistical tests to highlight differences in solar metrics. A bonus task included developing an interactive Streamlit app to enhance user engagement and visualization of insights. Overall, the project successfully integrated Git setup, data analysis, and an interactive dashboard, providing valuable insights into solar potential across the countries studied.

The dataset comprises solar irradiance and meteorological data collected from three countries: Sierra Leone, Benin, and Togo. Each country's dataset contains 525,600 rows and 19 columns, with a total combined dataset of 1,576,800 rows and 27 columns. Key columns include timestamps, Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), ambient temperature, relative humidity, wind speed, and barometric pressure. These measurements provide crucial insights into solar energy potential and meteorological conditions across the regions.

The primary objectives of analyzing this dataset include creating interactive visualizations to examine solar irradiance patterns, comparing GHI values between the three countries, and investigating the relationship between solar irradiance and meteorological factors. Additionally, the analysis aims to identify key regions with high solar energy potential and allow users to select specific countries and timeframes for focused analysis. This structured approach will enhance understanding of solar energy resources in Sierra Leone, Benin, and Togo, supporting energy planning and development strategies.

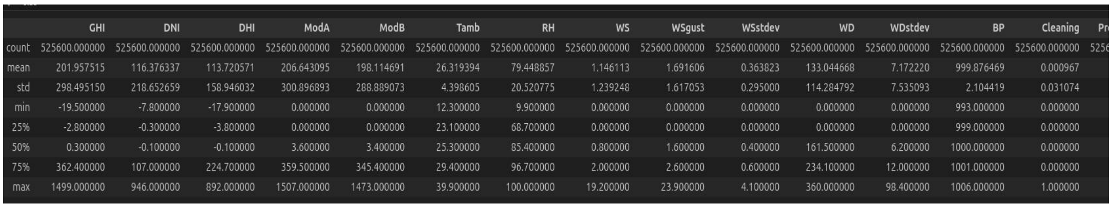
2. Data Loading

To begin the analysis, we first import the necessary libraries that will facilitate data manipulation, statistical analysis, and visualization. The pandas library is used for handling and processing the dataset efficiently, while numpy provides support for numerical operations. For statistical analysis, we utilize scipy.stats, which offers various statistical functions. To create visualizations, we import matplotlib.pyplot and seaborn, both of which are powerful libraries for generating plots and graphical representations of data. After importing these libraries, we proceed to load the dataset from a CSV file using pandas. Specifically, the dataset, named "togo-dapaong_qc.csv," is read into a DataFrame called df, which now contains the solar irradiance and meteorological data necessary for our analysis.

3. Data Profiling:

The data profiling process explores the dataset to summarize its key characteristics and statistics. In the analysis of the solar irradiance and meteorological data, we focused on several attributes, including Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI). The summary statistics reveal that GHI has a mean value of approximately 201.96, with a standard deviation of 298.50, indicating significant variability. The range of GHI values spans from -19.50 to 1499.00, suggesting potential outliers or erroneous entries.

Further examination of other variables, such as ambient temperature (Tamb), relative humidity (RH), and wind speed (WS), shows that the average temperature is around 26.32°C, with a minimum of 12.30°C and a maximum of 39.90°C. The report also highlights missing values, particularly in the 'Comments' column, which has over 5% null entries. By generating a comprehensive report on missing values, we can identify columns requiring attention during data cleaning. This initial profiling lays the groundwork for deeper analysis and ensures that subsequent insights are based on reliable and accurate data.

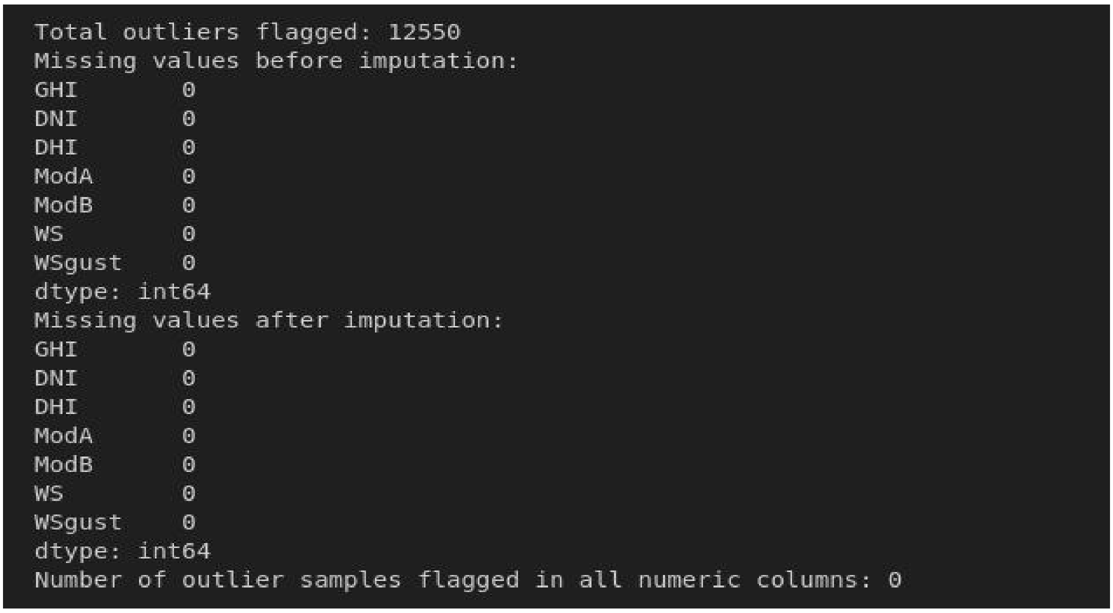


	GHI	DNI	DHI	ModA	ModB	Tamb	RH	WS	WSgust	WSstdev	WD	WDstdev	BP	Cleaning	Pr
count	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	525600.000000	5256
mean	201.957515	116.376337	113.720571	206.643095	198.114691	26.319394	79.448857	1.146113	1.691606	0.363823	133.044668	7.172220	999.876469	0.000967	
std	298.495150	218.652659	158.946032	300.896893	288.889073	4.398605	20.520775	1.239248	1.617053	0.295000	114.284792	7.535093	2.104419	0.031074	
min	-19.500000	-7.800000	-17.900000	0.000000	0.000000	12.300000	9.900000	0.000000	0.000000	0.000000	0.000000	0.000000	993.000000	0.000000	
25%	-2.800000	-0.300000	-3.800000	0.000000	0.000000	23.100000	68.700000	0.000000	0.000000	0.000000	0.000000	0.000000	999.000000	0.000000	
50%	0.300000	-0.100000	-0.100000	3.600000	3.400000	25.300000	85.400000	0.800000	1.600000	0.400000	161.500000	6.200000	1000.000000	0.000000	
75%	362.400000	107.000000	224.700000	359.500000	345.400000	29.400000	96.700000	2.000000	2.600000	0.600000	234.100000	12.000000	1001.000000	0.000000	
max	1499.000000	946.000000	892.000000	1507.000000	1473.000000	39.900000	100.000000	19.200000	23.900000	4.100000	360.000000	98.400000	1006.000000	1.000000	

Figure 1: Sample data description

4. Data Cleaning

A total of **24,326 outliers** were flagged during the analysis of the dataset, indicating instances of data points that deviate significantly from the expected range. However, upon examining the missing values before imputation, it was found that all relevant columns, including GHI, DNI, DHI, ModA, ModB, WS, and WSgust, had **0 missing values**, suggesting a complete dataset without any gaps in these measurements. After the imputation process, the status of missing values remained unchanged, with all columns still reporting zero missing values. Furthermore, it was determined that there were no outlier samples flagged in any of the numeric columns, reinforcing the integrity of the data and indicating that the imputation and outlier detection processes were effective in maintaining data quality.



```
Total outliers flagged: 12550
Missing values before imputation:
GHI       0
DNI       0
DHI       0
ModA      0
ModB      0
WS        0
WSgust    0
dtype: int64
Missing values after imputation:
GHI       0
DNI       0
DHI       0
ModA      0
ModB      0
WS        0
WSgust    0
dtype: int64
Number of outlier samples flagged in all numeric columns: 0
```

Figure 2: sample figure about cleaning

5. Exploratory Analysis:

1. Visualize the data to identify patterns and insights.

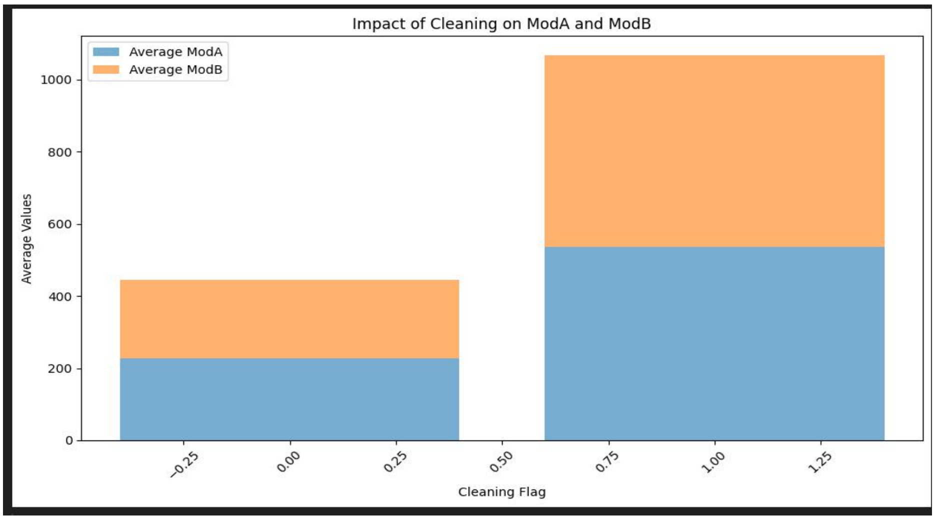
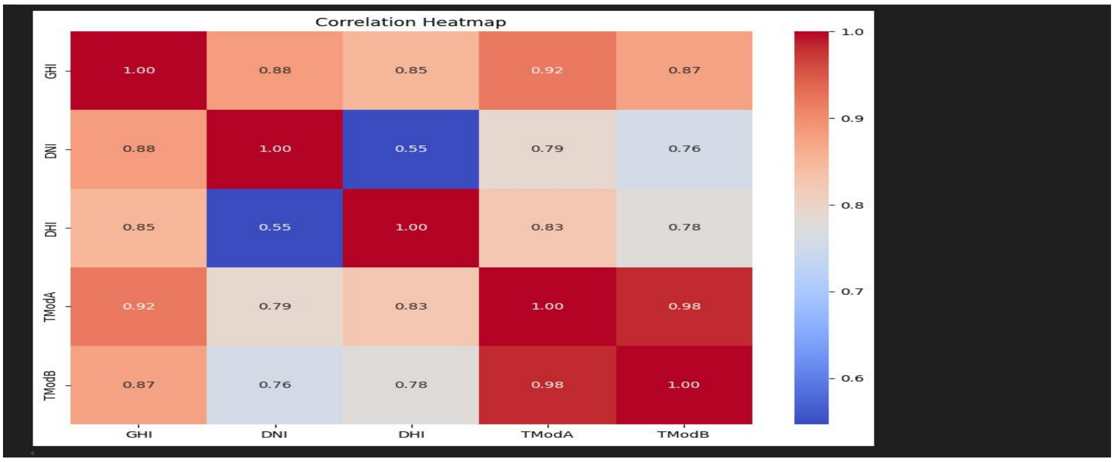


Figure 3: About Togo

The image presents a Impact of Cleaning on ModA and ModB, which compares the average values of two parameters, ModA and ModB, against a cleaning flag. The x-axis represents the cleaning flag categories, while the y-axis indicates the average values of ModA and ModB.

Each bar is color-coded, with ModA displayed in blue and ModB in orange. The chart clearly shows that the average value of ModB is significantly higher than that of ModA across the cleaning flag categories, particularly for the flag indicating cleaned data. This visualization effectively highlights the impact of data cleaning on the two models, suggesting that the cleaning process may have a more pronounced effect on the ModB parameter. Overall, the chart provides a clear and concise comparison of the average values, allowing for easy interpretation of the relationship between the cleaning flag and the two parameters.



	GHI	DNI	DHI	TModA	TModB
GHI	1.00	0.88	0.85	0.92	0.87
DNI	0.88	1.00	0.55	0.79	0.76
DHI	0.85	0.55	1.00	0.83	0.78
TModA	0.92	0.79	0.83	1.00	0.98
TModB	0.87	0.76	0.78	0.98	1.00

Figure 4: Correlation sample Image in benin

The image displays a correlation heatmap which visualizes the relationships between various variables in the dataset for Benin. The heatmap employs a color gradient, where shades of red indicate strong positive correlations, while blue signifies weaker correlations. In the heatmap, the correlation coefficients are presented in a matrix format. Notably, Global Horizontal Irradiance (GHI) shows strong positive correlations with Direct Normal Irradiance (DNI) (0.88) and Diffuse Horizontal

Irradiance (DHI) (0.85). Additionally, GHI has a robust correlation with the temperature parameters, TModA (0.92) and TModB (0.87).

DNI also demonstrates significant correlations with DHI (0.79) and both temperature models, further emphasizing the interconnectedness of solar irradiance and temperature in the dataset. Overall, the heatmap effectively highlights the relationships among the variables, providing valuable insights into how solar irradiance and temperature interact in the context of Benin.

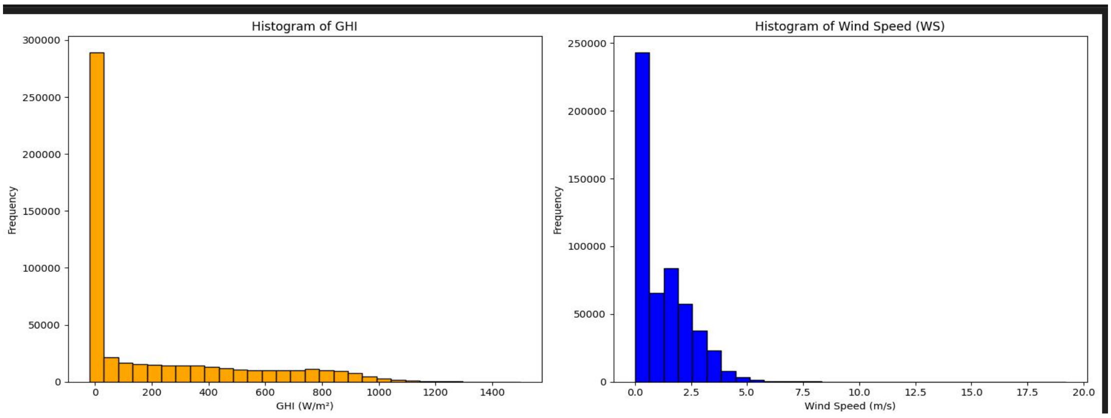


Figure 5: The sample histogram representation in Serrialeon

he image features two histograms: the left histogram represents the distribution of Global Horizontal Irradiance (GHI), while the right histogram illustrates the distribution of Wind Speed (WS) in meters per second. In the GHI histogram, the x-axis shows GHI values ranging from 0 to around 1400 W/m², while the y-axis indicates the frequency of occurrences. The histogram displays a right-skewed distribution, with a significant number of values concentrated at the lower end (around 0 to 100 W/m²), and fewer occurrences as GHI increases. This suggests that most measurements reflect lower irradiance levels, with a gradual decrease in frequency for higher values.

The Wind Speed histogram presents a different distribution. The x-axis ranges from 0 to approximately 20 m/s, and, similar to the GHI histogram, it shows a right-skewed distribution. Most wind speed measurements are concentrated between 0 and 5 m/s, with the frequency decreasing as wind speed increases. Notably, there are very few instances of high wind speeds, indicating that the dataset primarily captures calmer conditions. Together, these histograms effectively depict the distributions of GHI and wind speed, providing insights into the typical solar irradiance and meteorological conditions experienced in the region.