Week 0 – Solar Challenge Report

Name: Chalie Lijalem

Submission Date: 5/20/2025

Introduction

The Week 0 phase of the Solar Challenge focused on setting up the foundational infrastructure for analyzing solar power datasets across multiple African regions. This week's objective was to explore, clean, and visualize data from solar installations in countries such as Benin, Sierraleone, and Togo. The goal was to establish a reproducible, scalable, and maintainable pipeline for environmental data analysis.

Project Structure and Design Philosophy

From the start, a clean and modular project structure was prioritized. The repository was structured to separate raw datasets, Jupyter notebooks, reusable utility functions, and automation scripts. The core logic for data loading, preprocessing, and visualization was abstracted into a single src/utils.py file, making the codebase easy to reuse and extend.

Each country's analysis was organized in its dedicated notebook under the notebooks/ directory. This approach enabled consistent workflows across multiple datasets and promoted collaboration in a version-controlled setting.

Data Overview and Preprocessing

We began by loading solar datasets such as benin-malanville.csv and inspecting them for issues such as missing values, outliers, and inconsistent formatting. A dedicated utility function, generate_missing_value_report, was implemented to quickly generate missing data summaries, helping us identify the columns that needed attention.

Outliers were handled using z-score–based detection through the detect_outliers_zscore function. Missing values were addressed via impute_median, which ensured that the imputation retained the underlying data distributions.

In the spirit of data integrity, cleaned datasets were saved under the data/ directory to preserve the raw data separately for reproducibility and auditing purposes.

Exploratory Data Analysis (EDA)

To better understand sensor trends and relationships, a series of custom visualization utilities were developed:

- **plot_time_series**: Allowed us to examine temperature, humidity, and solar output across time.
- **plot_correlation_heatmap**: Gave insight into how sensor values were interrelated.
- **plot_pairwise**: Helped spot linear and non-linear trends among variables.

- **plot_bubble**: Made it easy to compare multivariate relationships like RH vs Tamb with bubble size indicating humidity levels.

Each notebook generated meaningful insights per region for instance, the Benin dataset revealed temperature dips aligned with drops in solar output, while Sierra Leone showed a different pattern.

Code Reusability and Best Practices

To avoid repetitive code and maintain clarity, all preprocessing and plotting functions were moved into the src/utils.py module. This aligns with Pythonic principles such as DRY (Don't Repeat Yourself) and separation of concerns. These reusable methods will streamline further development in later weeks of the challenge.

The repository also includes a requirements.txt file and uses venv for managing the environment locally. This ensures that the project can be reproduced easily across different machines and contributors.

Additionally, a GitHub Actions workflow was set up to enable basic continuous integration, ensuring code quality and environment compatibility moving forward.

Reflection and Next Steps

Week 0 laid a strong foundation by establishing a structured codebase and performing initial EDA on solar data. This week wasn't about building machine learning models it was about deeply understanding the data and setting up the environment needed to scale that understanding.

Next steps include:

- Adding more advanced feature engineering (e.g., rolling averages, time-window stats)
- Introducing object-oriented wrappers for datasets
- Writing tests for all utility functions
- Building dashboards or report generators to visualize EDA findings interactively

Conclusion

Overall, the work done in Week 0 has prepared the project for the weeks ahead. With a clean structure, reusable code, and meaningful insights from initial data exploration, the Solar Challenge team is now well-positioned to tackle more complex analytics, modeling, and deployment tasks in the upcoming phases.