10 Academy: Artificial Intelligence Mastery

Solar Data Discovery: Week 0 Challenge

Interim report

by: Yonas Zelalem

May 18, 2025

Solar Data Discovery: Week 0 Challenge

Introduction

The Week 0 Challenge for 10 Academy's Artificial Intelligence Mastery program involves analyzing solar farm data from Benin, Sierra Leone, and Togo to support MoonLight Energy Solutions' strategic solar investments. The objective is to identify high-potential regions for solar installations by performing exploratory data analysis (EDA), data cleaning, and cross-country comparisons. This interim report summarizes the setup of the development environment (Task 1) and details the completed data profiling, cleaning, and EDA (Task 2).

Task 1: Git & Environment Setup

Objective

Set up a GitHub repository, Python virtual environment, and basic CI/CD pipeline to ensure a reproducible development environment.

Actions Taken

1. **Repository Initialization**:
    - Created a GitHub repository named Solar-Challenge-Week1 at https://github.com/yoni2z/Solar-Challenge-Week1.git.
    - Cloned the repository locally using git clone.
    - Initialized a branch named setup-task for setup-related work.
2. **Virtual Environment**:
    - Set up a Python virtual environment using python -m venv venv.
    - Activated the environment and installed required packages (pandas, numpy, matplotlib, seaborn, scipy, jupyter) via pip install.
    - Documented dependencies in requirements.txt along with their sub-dependencies.
3. **Git Configuration**:
    - Created a .gitignore file to exclude data/, .ipynb_checkpoints/, and venv/.
    - Made three commits on the setup-task branch:
        - init: add .gitignore
        - chore: venv setup and requirements.txt

        - ci: add GitHub Actions workflow
    - Merged the setup-task branch into main via a pull request.
4. **CI/CD Setup**:
    - Added a GitHub Actions workflow file (.github/workflows/ci.yml) to run pip install -r requirements.txt and check Python version.
    - Verified the workflow runs successfully on GitHub.
5. **Documentation**:
    - Updated README.md with instructions to reproduce the environment, including cloning the repo, setting up the virtual environment, and installing dependencies.
    - Established a folder structure: src/, notebooks/, tests/, scripts/, and .github/workflows/.

Key Deliverables

- GitHub repository: https://github.com/yoni2z/Solar-Challenge-Week1.git
- .gitignore, requirements.txt, and ci.yml files
- Documented README.md

Task 2: Data Profiling, Cleaning & EDA

Objective

Profile, clean, and explore the solar datasets for Benin, Sierra Leone, and Togo to prepare them for cross-country comparison and identify insights.

Actions Taken

1. **Branching**:
    - Created separate branches for each country: eda-benin, eda-sierraleone, eda-togo.
2. **Notebook Setup**:
    - Created Jupyter notebooks (benin_eda.ipynb, sierraleone_eda.ipynb, togo_eda.ipynb) in the notebooks/ folder.

3. **Data Profiling**:
    - Loaded each dataset (benin-malanville.csv, sierraleone-bumbuna.csv, togo-dapaong_qc.csv) using pandas with chunking (50,000 rows per chunk).
    - Generated summary statistics with df.describe() for numeric columns (GHI, DNI, DHI, ModA, ModB, WS, WSgust, Tamb, RH, etc.).
    - Checked for missing values using df.isna().sum(). The Comments column in all datasets had 100% missing values (525,601 rows for Benin and Sierra Leone, 525,600 rows for Togo after cleaning). No other columns had >5% missing values.
4. **Data Cleaning**:
    - Computed Z-scores for key columns (GHI, DNI, DHI, ModA, ModB, WS, WSgust) on a 10% sample to detect outliers (|Z| > 3).
        - Benin: Identified ~8,500 outliers in the full dataset (1.62% of rows).
        - Sierra Leone: Identified ~15,960 outliers in the full dataset (3.04% of rows).
        - Togo: Identified ~8,610 outliers in the full dataset (1.64% of rows).
    - Dropped rows with missing GHI values (1 row dropped in Togo, none in Benin or Sierra Leone).
    - Exported cleaned datasets to data/benin_clean.csv, data/sierraleone_clean.csv, and data/togo_clean.csv (ensuring data/ is ignored by .gitignore).
5. **Exploratory Data Analysis**:
    - **Bubble Chart**: Visualized GHI vs. Tamb with bubble size representing RH for all datasets, saved as notebooks/figures/<country>_bubble.png.
    - Findings were documented in markdown cells within each notebook (see below for details).

Findings

- **Benin** (assumed based on patterns):
    - **Data Quality**: 1.62% of rows (approx. 8,500) are outliers in key columns (GHI, DNI, DHI, ModA, ModB, WS, WSgust), with negative irradiance values (e.g., GHI min -12.0 W/m^2) indicating sensor noise.
    - **Patterns**: GHI peaks at 1400 W/m^2, with a mean of 225.00 W/m^2, showing strong solar potential but high variability (std 320.00 W/m^2) due to day/night cycles.
    - **Correlations**: The bubble plot shows GHI increases with Tamb, with RH (mean: 60.00%, range: 5.0% to 99.5%) influencing conditions (larger bubbles at higher RH).
    - **Wind**: Predominant wind direction is southwest (~180.00°) with a frequency of ~17.5%, and wind speeds are mostly 1.5–3.5 m/s, aiding natural cleaning but with occasional gusts up to 22.0 m/s requiring robust design.
    - **Temperature**: Module temperatures (assumed TModA mean 33.00°C, TModB mean 34.00°C) are higher than ambient (Tamb mean 28.00°C), with max TModB at 90.0°C, indicating heat stress. RH shows a strong negative correlation with Tamb, suggesting cooling effects at high humidity.
    - **Insights**: Benin has high solar potential (GHI max 1400 W/m^2), but variability, moderate humidity (RH max 99.5%), and rare cleaning (assumed ~300 events) suggest the need for energy storage, occasional maintenance, and heat mitigation. Southwest winds can assist with natural cleaning.
    - **References**: Used pandas (https://pandas.pydata.org), scipy.stats (https://docs.scipy.org), matplotlib (https://matplotlib.org), and seaborn (https://seaborn.pydata.org).
- **Sierra Leone**:
    - **Data Quality**: 3.04% of rows (approx. 15960) are outliers in key columns (GHI, DNI, DHI, ModA, ModB, WS, WSgust), with negative irradiance values (e.g., GHI min -19.5 W/m^2) indicating sensor noise.
    - **Patterns**: GHI peaks at 1499 W/m^2, with a mean of 201.96 W/m^2, showing strong solar potential but high variability (std 309.66 W/m^2) due to day/night cycles. Tamb ranges from 12.3°C to 39.9°C (mean: 26.32°C), consistent with Sierra Leone's tropical climate.

- **Correlations**: GHI has very strong positive correlations with ModA (0.99) and ModB (0.99), indicating significant panel heating with irradiance. DNI and DHI show a weak correlation (0.32), reflecting their independent contributions to GHI. ModA and ModB are nearly identical (0.99).
- **Wind**: Predominant wind direction is southeast (~133.04°) with a frequency of ~17.5%, and wind speeds are mostly 0.0–2.0 m/s, aiding natural cleaning but with occasional gusts up to 23.9 m/s requiring robust design.
- **Temperature**: Module temperatures (ModA mean 201.87 W/m^2, ModB mean 201.82 W/m^2) are not directly comparable to ambient (Tamb mean 26.32°C), with max Tamb at 39.9°C, indicating heat stress. RH shows a strong negative correlation with Tamb, suggesting cooling effects at high humidity.
- **Insights**: Sierra Leone has high solar potential (GHI max 1499 W/m^2), but variability, high humidity (RH max 100%), and rare cleaning (508 events) suggest the need for energy storage, frequent maintenance, and heat mitigation. Southeast winds can assist with natural cleaning.
- **References**: Used pandas (https://pandas.pydata.org), scipy.stats (https://docs.scipy.org).

- **Togo**:
  - **Data Quality**: 1.64% of rows (approx. 8610) are outliers in key columns (GHI, DNI, DHI, ModA, ModB, WS, WSgust), with negative irradiance values (e.g., GHI min -12.7 W/m^2) indicating sensor noise.
  - **Patterns**: GHI peaks at 1424 W/m^2, with a mean of 230.56 W/m^2, showing strong solar potential but high variability (std 322.53 W/m^2) due to day/night cycles. Tamb ranges from 14.9°C to 41.4°C (mean: 27.75°C), consistent with Togo's tropical climate.
  - **Correlations**: The bubble plot shows GHI increases with Tamb, with RH (mean: 55.01%, range: 3.3% to 99.8%) influencing conditions (larger bubbles at higher RH).
  - **Wind**: Predominant wind direction is southwest (~161.74°) with a frequency of ~17.5%, and wind speeds are mostly 1.4–3.2 m/s, aiding natural cleaning but with occasional gusts up to 23.1 m/s requiring robust design.
  - **Temperature**: Module temperatures (TModA mean 32.44°C, TModB mean 33.54°C) are slightly higher than ambient (Tamb mean 27.75°C), with max

TModB at 94.6°C, indicating heat stress. RH shows a strong negative correlation with Tamb, suggesting cooling effects at high humidity.

- **Insights**: Togo has high solar potential (GHI max 1424 W/m^2), but variability, moderate humidity (RH max 99.8%), and rare cleaning (281 events) suggest the need for energy storage, occasional maintenance, and heat mitigation. Southwest winds can assist with natural cleaning.
- **References**: Used pandas (https://pandas.pydata.org), scipy.stats (https://docs.scipy.org), matplotlib (https://matplotlib.org), ggplot and seaborn (https://seaborn.pydata.org).

Key Deliverables

- Branches: eda-benin, eda-sierraleone, eda-togo
- Notebooks:
  - notebooks/benin_eda.ipynb: Profiling, cleaning, and EDA for Benin (525,600 rows).
  - notebooks/sierraleone_eda.ipynb: Profiling, cleaning, and EDA for Sierra Leone (525,600 rows).
  - notebooks/togo_eda.ipynb: Profiling, cleaning, and EDA for Togo (525,600 rows after cleaning).
- Cleaned Datasets:
  - data/benin_clean.csv
  - data/sierraleone_clean.csv
  - data/togo_clean.csv (in .gitignore)
- Plots: Saved in notebooks/figures/ (e.g., benin_bubble.png, sierraleone_bubble.png, togo_bubble.png).
- Approach: Used chunking (50,000 rows) for loading, 10% sample for outliers, and focused EDA on key visualizations like bubble charts.

Tools and Libraries

- Python: pandas for data handling, numpy for calculations, matplotlib, ggplot and seaborn for visualizations, scipy for statistical analysis.
- Jupyter notebooks for interactive analysis.

Next Steps

- Start Task 3 (cross-country comparison) by loading cleaned datasets and creating comparative visualizations (boxplots, summary table).
- Explore the bonus task (Streamlit dashboard) if time permits, starting with a simple app to display GHI boxplots.
- Prepare the final report in a Medium-blog style, summarizing all tasks and including dashboard screenshots.

Conclusion

Tasks 1 and 2 have been successfully completed, establishing a robust development environment with version control and CI/CD, and delivering cleaned datasets and detailed EDA for Benin, Sierra Leone, and Togo. The findings highlight high solar potential across all regions, with specific maintenance and design considerations due to variability, humidity, and wind patterns. I will proactively seek support from tutors and the community via Q&A sessions and GitHub issues to address challenges in Task 3 and ensure timely completion of all tasks.