WEEK 0 INTERIM REPORT : SOLAR DATA DISCOVERY CHALLENGE

10 Academy - Kifiya AI Mastery Training

Date: November 6, 2025
Trainee: Hermona Addisu
GitHub Repository: https://github.com/HermonaDev/solar-challenge-week0

1.0 INTRODUCTION

This interim report outlines my planned approach for the Week 0 Solar Data Discovery Challenge. The project involves analyzing solar farm data from Benin, Sierra Leone, and Togo to support MoonLight Energy Solutions in identifying high-potential regions for solar investment. This report covers my completed setup work and the methodology I plan to implement for the remaining tasks.

2.0 TASK 1: GIT & ENVIRONMENT SETUP

2.1 Summary of Completed Setup

I have successfully established the foundational infrastructure for this project:

Repository Structure:

- Created a professional GitHub repository with organized project structure
- Implemented a `.github/workflows` directory for CI/CD automation
- Set up separate directories for notebooks, scripts, data, and documentation

Development Environment:

- Configured a Python virtual environment to ensure consistent dependency management
- Installed essential libraries including pandas, numpy, matplotlib, seaborn, and scikit-learn
- Created a `requirements.txt` file for reproducibility

Version Control Strategy:

- Implemented a feature branching workflow for organized development
- Established clear commit message conventions for tracking progress
- Set up GitHub Actions for continuous integration testing

Current Status: The development environment is fully operational and ready for the analysis phase.

3.0 TASK 2: DATA PROFILING, CLEANING & EDA APPROACH

3.1 Planned Methodology

For the exploratory data analysis phase, I plan to follow this systematic approach:

Data Profiling:

- Load datasets for Benin, Sierra Leone, and Togo
- Examine data structure, column types, and basic statistics
- Identify missing values, duplicates, and data quality issues
- Document the shape and characteristics of each dataset

Data Cleaning Strategy:

- Address physically impossible values (e.g., negative solar irradiance readings)
- Handle missing data using appropriate imputation/ removal techniques where necessary
- Detect and treat outliers using Z-score methodology (threshold: $|Z| > 3$)
- Standardize data formats and ensure consistency across datasets

Exploratory Data Analysis Plan:

- **Summary Statistics:** Calculate mean, median, standard deviation, and ranges for key solar metrics (GHI, DNI, DHI)
- **Time Series Analysis:** Examine temporal patterns at daily and monthly scales to identify seasonal trends
- **Correlation Analysis:** Generate correlation matrices and heatmaps to understand relationships between environmental variables
- **Distribution Analysis:** Create histograms and density plots to understand the distribution of solar radiation values
- **Visual Analytics:** Develop boxplots, scatter plots, and line graphs to reveal patterns and anomalies

Key Metrics to Analyze:

- Global Horizontal Irradiance (GHI)
- Direct Normal Irradiance (DNI)
- Diffuse Horizontal Irradiance (DHI)
- Ambient Temperature
- Module Temperature
- Wind Speed and Direction
- Relative Humidity

3.2 Expected Deliverables

- Clean, processed datasets for all three countries
- Comprehensive statistical summary tables
- Visualization suite documenting key patterns and relationships
- Written insights documenting data quality issues and preliminary findings

3.3 Risk Assessment & Mitigation Strategy

Technical & Operational Challenges Anticipated:

Computational Resources:

- *Risk:* Large dataset size (1.5M+ records total) may strain local processing
- *Mitigation:* Implement batch processing and sampling for initial exploration; use efficient data types (category, float32)

Software Dependencies:

- *Risk:* Version conflicts or package compatibility issues during analysis
- *Mitigation:* Pinned dependency versions in requirements.txt

Infrastructure Reliability:

- *Risk:* CI/CD pipeline failures or GitHub Actions downtime
- *Mitigation:* Local testing scripts as backup; regular pipeline validation

Development Velocity:

- *Risk:* Complex visualizations or statistical methods requiring iterative refinement
- *Mitigation:* Modular code design with rapid prototyping approach; version-controlled experimentation

Deployment Challenges:

- *Risk:* Streamlit Cloud compatibility or deployment issues
- *Mitigation:* Early staging deployment; comprehensive logging and error handling

Time Management:

- *Risk:* Unforeseen complexity in cross-country statistical analysis
- *Mitigation:* Buffer time allocation; MVP-first approach for core deliverables

4.0 CURRENT PROGRESS & REMAINING WORK

4.1 Progress to Date (Nov 7-8)

Already Completed:

✅ Established professional development environment and CI/CD pipeline

✅ Developed reusable data cleaning classes and validation framework

✅ Conducted initial exploratory analysis for all three countries

✅ Implemented core statistical comparison methodology

In Progress:

🔄 Comprehensive data quality assessment and outlier analysis

🔄 Advanced time series pattern identification

🔄 Cross-country statistical significance testing

🔄 Streamlit dashboard development

4.2 Remaining Work Plan (Nov 9-12)

Finalize Analysis & Validation (Nov 9)

- Complete comprehensive EDA with statistical rigor

- Validate findings through multiple analytical approaches

- Ensure data quality and methodological soundness

Enhance Deliverables (Nov 10)

- Polish visualizations and business insights

- Develop strategic investment recommendations

- Prepare production-ready dashboard deployment

Documentation & Submission (Nov 11)

- Compile comprehensive final report

5.0 CONCLUSION

The project foundation has been successfully established with a professional development environment and version control system in place. I have outlined a clear methodology for the data analysis work ahead, focusing on thorough data cleaning, comprehensive exploratory analysis, and meaningful cross-country comparisons. The planned approach will provide MoonLight Energy Solutions with actionable insights to guide their solar investment strategy. I am confident this systematic methodology will yield high-quality analytical results by the final submission deadline.