Solar Performance Analysis – Methodology Documentation

1. Introduction

The Solar Performance Analysis project aims to evaluate the energy generation efficiency of multiple solar sites, identify anomalies, and derive actionable recommendations for performance optimization. This project uses a data-driven approach by integrating time-series data analytics, database solutions, and visualization frameworks.

2. Project Goals

- Design a robust data pipeline to process, store, and retrieve large-scale time-series solar data efficiently.
- Develop an analysis framework to assess performance metrics, compare site data, and detect operational anomalies.
- Create interactive dashboards and reports for decision-making and performance monitoring.

3. Data Cleaning and Preprocessing

Before performing the analysis, the raw data underwent a comprehensive cleaning and preprocessing workflow to ensure consistency and accuracy. The steps followed were:

1. Handling Missing or Empty Columns:
    - The datasets for Site 2 and Site 3 contained several columns with entirely missing values. These columns were removed completely as they provided no relevant information.
    - The dataset for Site 1 did not have such empty columns.
2. Row Alignment:
    - The Site 3 dataset included an extra timestamp row that did not align with the other two datasets. This additional row was removed to maintain consistency across all datasets.
3. Column Formatting:
    - The column headers were standardized for clarity and ease of use, such as renaming fields for better understanding and ensuring uniform naming conventions across datasets.
    - Unnecessary fields and redundant information were removed.
    - Data type conversions were applied, such as transforming the timestamp column from the object data type to datetime for efficient time-series operations.
4. Handling Missing Values and Outliers:

- Missing values were handled using linear interpolation to maintain the continuity of time-series data.
- Duplicate values were checked and removed if identified.
- Potential outliers were detected and analyzed to minimize their impact on the final analysis.

5. Repetition for Consistency:

   - The above steps were systematically applied to the datasets of all three sites to ensure consistency in cleaning and formatting.

4. Data Transformation and Storage

After cleaning, the data was transformed and prepared for efficient querying and analysis by uploading it to a relational database.

1. Database Creation:

   - A PostgreSQL database named solar_analysis was created to store the cleaned data. Within this database, three separate tables were set up, one for each solar site.The primary key for each tables is their 'timestamp' column which can be used to distinguish all records and query effectively.
   - A new table 'site_metadata' is also created to efficiently query the data by using the concept of foreign keys and primary keys.

2. Data Upload Process:

   - Using Python libraries like SQLAlchemy, an engine was created to facilitate uploading the cleaned data to PostgreSQL.
   - Psycopg2 was used to establish a connection to the database and execute queries.

3. Database Efficiency:

   - The cleaned datasets were uploaded into PostgreSQL in a structured and easy-to-understand format. This structured database allowed for efficient querying and analysis of large time-series data.
   - This database architecture provided a robust foundation for conducting further performance analysis across all three solar sites.

5. Data Structure

1. Files:
   - cleaned_solar_data_Site_1.csv
   - cleaned_solar_data_Site_2.csv
   - cleaned_solar_data_Site_3.csv
2. Key Columns: (X can be any value from 1,2 and 3)
   - Timestamp (indregTCX_timestamp). It is also used as primary key.
   - Energy Generated (indregTCX_Energy_kWh_sum)
   - Power Factor (indregTCX_Power_Factor_avg_avg)
   - Voltage (indregTCX_Voltage_LL_V_avg

6. Analysis Framework

This part talks about the analysis done on data to gain useful insights:

1. Summary Statistics

A detailed statistical analysis was performed on the cleaned datasets to understand the key characteristics and variability in the data. The following metrics were calculated for all performance parameters across the three solar sites:

- Mean: To identify the average value of parameters such as energy generation, power factor, and voltage.
- Standard Deviation: To assess the variability or consistency of performance metrics over time.

For example:

- The mean power factor was calculated to understand the efficiency of energy conversion.
- The standard deviation of voltage levels helped assess voltage stability and consistency across sites.

2. Performance Metrics Calculation

- Energy Yield:
  - Calculated total energy generation (kWh) per site and for all sites combined.
  - Aggregated daily, monthly, and overall trends.

- Efficiency:
  - Derived efficiency by normalizing energy generation based on installed capacity and environmental conditions.

3. Site Comparison Methodology

- Key Metrics:
  - Total energy yield.
  - Average power factor.
  - Voltage stability (standard deviation of voltage values).
- Comparison Logic:
  - Aggregated site-wise data for side-by-side comparison using bar plots.
  - Highlighted top-performing and underperforming sites.

4. Anomaly Detection
    - Energy Anomalies:
        - Threshold-based: Energy generation values below the user-defined threshold flagged as anomalies.
    - Power Factor Anomalies:
        - Applied statistical bounds:
            - Lower: μ−3σ\mu - 3\sigmaμ−3σ
            - Upper: μ+3σ\mu + 3\sigmaμ+3σ
            - Flagged outliers visually in scatter plots.
5. Power Quality Analysis
    - Voltage Stability:
        - Computed standard deviation to assess load balance and voltage fluctuations.
    - Power Factor Impact:
        - Assessed the correlation between power factor anomalies and energy yield.

7. Visualization and Dashboard

An interactive dashboard was developed using Streamlit, pandas and plotly to provide insights into solar site performance. The key features and visualizations are as follows:

1. Performance Overview:
    - Energy Trends: A line chart visualized daily energy generation, showing fluctuations over time. Bar charts displayed aggregated energy for each site, allowing for easy comparison.
2. Anomaly Visualization:
    - Identifying Anomalies: Queries were used to filter out specific periods of irregular performance (e.g., dips in energy or low power factor). Scatter plots highlighted these anomalies in red, showing points of concern.
3. Site Comparisons:
    - Energy Yield & Efficiency: A grouped bar chart compared total energy generation across all sites. Queries filtered data to focus on specific sites, highlighting the differences in energy output and inefficiencies.
4. Interactive Filters:
    - Users could filter data by:
        - Date Range: Focused on specific time periods to visualize trends and anomalies.
        - Site Selection: Allowed users to choose one or multiple sites for comparison.

- Anomaly Sensitivity: Enabled dynamic adjustment of anomaly detection thresholds.

8. Results and Insights

This part talks about the results and insights derived from analysis of data

1. Key Findings:
   - Energy generation trends were visualized daily and over time for each site.
   - Detected significant anomalies in energy generation and power factor data.
   - Compared total energy yield across sites to highlight performance discrepancies.
2. User Benefits:
   - Easy-to-understand visualizations to monitor solar site performance.
   - Proactive detection of anomalies to aid in maintenance and optimization.
   - Streamlined comparative analysis across sites.

9. Future Enhancements

1. Real-time Monitoring**:** Integration of real-time data for continuous tracking of energy generation and system performance. Reproducible data import process is not used due to time constraints.
2. Predictive Maintenance**:** Introducing predictive analytics to detect potential issues early, minimizing downtime and improving reliability.
3. Custom Efficiency Metrics**:** Enabling users to define and track tailored efficiency metrics based on site-specific needs.

10. Conclusion

The project successfully integrated engineering and analytics to develop a comprehensive solar performance analysis framework. By identifying inefficiencies, detecting anomalies, and comparing site performance, the analysis provides actionable recommendations for optimizing energy yield and improving solar site operations.