

Data Scientist I Assignment

This assignment involves the detection of photometric changes in satellite observations. You are provided with datasets representing three eras, each containing multiple CSV files. Each file corresponds to a specific satellite identified by a unique five-digit NORAD ID and contains photometric data collected over an 8-day duration.

Data Description

Each CSV file contains the following columns:

Column	Description
Norad_id	Unique identifier of a satellite
Timestamp	Observation time in UTC
Equatorial_phase	Solar Equatorial Phase Angle (deg) (or SEPA)
Magnitude	Observed magnitude of the satellite
Magnitude_unc	Uncertainty in calculation of the magnitude
Sensor	Unique identifier of the sensor
Zeroptd	Zero point used to calibrate the sensor

Objective

Compare the reference measurements taken over the previous seven days to the current measurements collected over a six-hour window and identify anomalies or deviations. The results must be summarized as a JSON file per era, listing the detected anomalies, the corresponding range of Equatorial Phase Angles, and the range of timestamps of the occurrence. The table below shows the start and end times of the reference and current periods for each era.

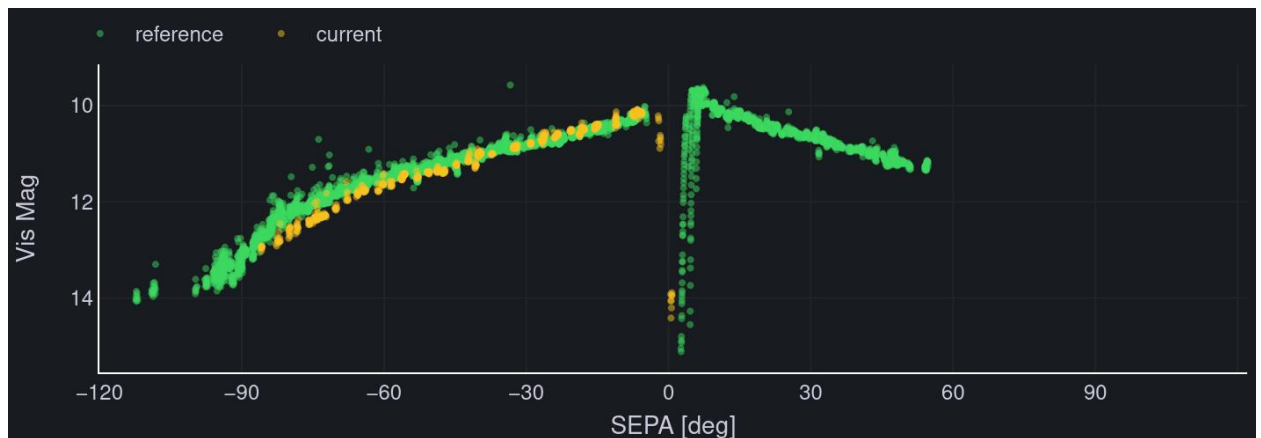
Note that reference start to end is 7 days and the current period from start to end is 6 hours.

Era	Reference Start	Reference End	Current Start	Current End
1	2025-10-31 03:20:19	2025-11-07 03:20:19	2025-11-07 03:20:19	2025-11-09 03:20:19
2	2025-10-28 20:21:50	2025-11-04 20:21:50	2025-11-04 20:21:50	2025-11-06 20:21:50
3	2025-10-08 11:15:42	2025-10-15 11:15:42	2025-10-15 11:15:42	2025-10-17 11:15:42

Example output format:

```
{  
  "norad_id": 12345,  
  "equatorial_phase": [-82.5, -64.4],  
  "timestamp": ["2025-10-14 11:57:20", "2025-10-14 13:10:16"]  
}
```

The figure below illustrates the comparison between reference and current measurements for object with norad_id 12345. It can be seen from the current measurements have deviated from the reference measurements on the left arm of the data between equatorial phase angles of [-82.5, -68.8]. Note that the sudden increase in the visual magnitude around zero degrees is because of an effect known as “glinting”. These effects need to be excluded from the output data.



Write a report not exceeding four pages with a focus on practical analysis, thorough methodology selection, and robust explanation.

Task Breakdown

1. **Provide a python class** that has a method called `run` which takes the input in the form of a json file and provides the output in the form of a json file containing changes detected. Please add comments wherever necessary to improve the readability of code. The code must have the following functions:
 - a. **Outlier Handling**
 - b. **Account for Glinting Effects Near Zero Equatorial Phase Angle:** Identify and adjust for "glinting" in measurements around zero SEPA. Explain how glinting confounds change detection and describe your strategy to mitigate its impact.

- c. **Assign Degree of Deviations to Anomalies:** For each detected anomaly, quantify the degree of deviation (e.g. probability) from the reference norm. Present results in a table or graph and annotate your scoring scheme.
 - d. **Detect if Objects Have Deviated from Normal Element Set:** Define what constitutes the "normal element set" (e.g., typical pattern or grouping). Analyze if any objects/elements in the "current" dataset have shifted group memberships or display new/unusual behaviors compared to the reference.
- 2. **Explain Method Choice, Advantages, and Disadvantages:** For each step (outlier handling, anomaly detection, glinting adjustment, deviation scoring, set deviation), explain why the approach was selected. List any other methods that could be used and describe why this method was chosen over the others.
- 3. Include figures to show that anomalies are being detected correctly.
- 4. **List All Assumptions Made:** Explicitly state every assumption in your workflow (e.g., measurement reliability, distributional assumptions, independence of errors, nature of glinting, etc.). Discuss how these assumptions might affect the validity and robustness of your findings.

Note: The norad_id in the data has been altered to protect the names of the actual satellites.