Exploratory Data Analysis (EDA) Findings

Dataset Summary

- Duration: June 2023 to May 2025
- Scope: Rajasthan-level solar generation (CEA) + NASA POWER climate data
- Records: 731 daily entries

Column Renaming & Structure

Cleaned and normalized columns:

- Solar generation: `solar_mwh`
- Temperature: `temp_avg_c`, `temp_max_c`, `temp_min_c`
- Humidity: `humidity_pct`
- Wind: `wind_speed_ms`
- Precipitation: `precip_mm`
- Solar irradiance: `solar_rad_allsky_mj_m2`, `solar_rad_clrsky_mj_m2`
- Cloudiness proxy: `cloudiness_index = clrsky - allsky`

---

1. 📊Correlation Heatmap Insights

**Target: **``

Predictor	Correlation	Direction	Strength
All-sky Irradiance	+0.86	↗	Very Strong
Clear-sky Irradiance	+0.81	↗	Strong
Cloudiness Index (derived)	−0.72	↘	Strong Inverse
Max Temperature	+0.17	↗	Weak
Humidity	−0.13	↘	Weak
Rainfall	−0.08	↘	Negligible
Wind Speed	−0.03	↔	None

Conclusion:

> Solar radiation (especially All-sky) is the dominant driver. Cloudiness has strong negative effect. Other variables provide minor seasonal context.

---

2. 📈Temporal Trend Analysis

- `solar_mwh` and `solar_rad_allsky_mj_m2` show strong seasonal alignment.
- Peak: April to June
- Dips: July (monsoon) & December–January (winter haze)
- `precip_mm` spikes during monsoon (June–September) correlate with solar troughs.
- `cloudiness_index` surges during same low solar periods, confirming its suppressive role.
- Temperatures peak May–June but are not strongly aligned with solar output drops.

---

3. 🧦Cloudiness as Key Inhibitor

- Difference between clear-sky and all-sky irradiance reveals true cloud burden.
- Cloudiness index is highly anti-correlated with solar output (−0.72).

---

4. 📊Correlation Matrix Summary

```
solar_mwh           1.00
solar_rad_allsky    0.86
solar_rad_clrsky    0.81
cloudiness_index   -0.72
temp_max_c          0.17
humidity_pct       -0.13
precip_mm          -0.08
wind_speed_ms      -0.03
```

---

5. 🧰Pairplot Insights

- **Strong Linear Cluster:**
- `solar_mwh` vs. `solar_rad_allsky_mj_m2`
- `solar_mwh` vs. `solar_rad_clrsky_mj_m2`
- **Clear Negative Slope:**
- `solar_mwh` vs. `cloudiness_index`

- **Wide Scatter / Weak Correlation:**

- `solar_mwh` vs. `humidity_pct`
- `solar_mwh` vs. `precip_mm`
- `solar_mwh` vs. `wind_speed_ms`

- **Outliers & Distribution Checks:**

- All variables generally show Gaussian distributions except for precipitation, which is heavily right-skewed.

Recommendations for Modeling

- **Primary predictors:** `solar_rad_allsky_mj_m2`, `cloudiness_index`
- **Optional add-ons:** `temp_max_c`, `humidity_pct`
- **Exclude:** `wind_speed_ms`, `precip_mm` (low signal)

Would you like me to proceed with feature engineering (lags, rolling means) or start building a baseline predictive model next?