

Insight yang Bisa Didapat dari Dataset Cuaca Anda (Sebelum Modeling)

1. Pola Musiman Curah Hujan

- Apakah ada bulan-bulan tertentu dengan curah hujan tinggi (musim hujan)?
 - Contoh: Puncak hujan di bulan Desember–Februari?
 - Bisa divisualisasikan dengan:
-

2. Hubungan Suhu & Curah Hujan

- Apakah curah hujan cenderung lebih tinggi saat suhu lebih rendah?
 - Misalnya: RR tinggi saat T_x atau T_{avg} rendah → indikasi cuaca mendung/berawan
 - Visualisasi: scatter plot antara T_{avg} dan RR
-

3. Pengaruh Kelembaban terhadap Hujan

- Apakah kelembaban tinggi (RH_{avg}) berkorelasi positif dengan RR?
 - Bisa dicek melalui korelasi atau scatter plot.
 - Jika kuat, kelembaban bisa jadi fitur penting dalam model prediksi.
-

4. Pengaruh Angin terhadap Hujan

- Apakah arah atau kecepatan angin berkaitan dengan hujan?
 - Arah angin tertentu bisa menjadi indikasi datangnya hujan dari wilayah tertentu.
 - Kecepatan angin tinggi bisa dikaitkan dengan badai atau hujan lebat.
-

5. Hubungan Penyinaran Matahari dengan Hujan

- Apakah ss (lama penyinaran matahari) rendah saat RR tinggi?
 - Logis secara fisik: hujan → langit mendung → matahari terhalang
-

6. Distribusi Data

- RR sering punya distribusi **skewed** (condong ke 0, banyak hari tidak hujan).
- Mungkin perlu di-log-transform sebelum modeling agar distribusi lebih normal.

7. 📌 Outlier / Nilai Tidak Wajar

- Nilai-nilai suhu atau curah hujan yang sangat ekstrem perlu ditandai.
- Bisa berdampak besar pada model regresi atau prediksi, apalagi jika outlier adalah error pencatatan.

8. 📈 Trend Jangka Panjang

- Apakah ada tren peningkatan/pengurangan suhu atau curah hujan dari 2023 ke 2024?
- Bisa berkaitan dengan perubahan iklim lokal.

9. 🛠️ Kualitas Data

- Seberapa banyak nilai hilang atau tidak terukur (8888)?
- Apakah perlu diimputasi? Di-drop?

🏠 Kesimpulan Awal:

Insight ini membantu Anda menjawab:

- Fitur apa yang **paling relevan** untuk prediksi?
- Apakah perlu **transformasi khusus** untuk fitur tertentu?
- Apa strategi **splitting** dan **validasi waktu** yang cocok (karena ini data time series)?