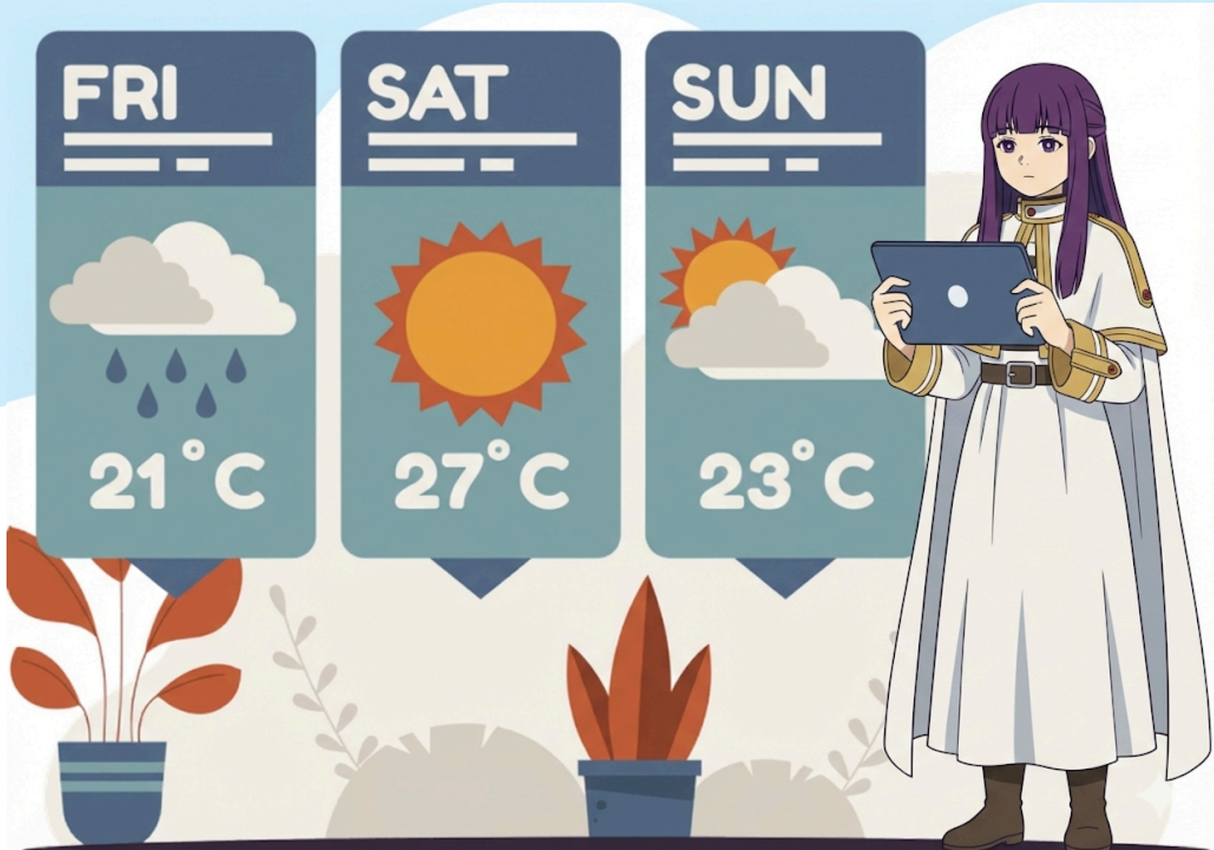
Based on Philippine Cities Weather Data (2020-2023)

WEATHER FORECAST & TEMPERATURE PREDICTION IN THE PHILIPPINES (2026)

Naufal Dzakia Raiffaza — Data Scientist



WHO AM I??

My Name is Naufal Dzakia Raiffaza, And I am a Data Analyst and Scientist Researcher driven by a passion for Data Science and Python.

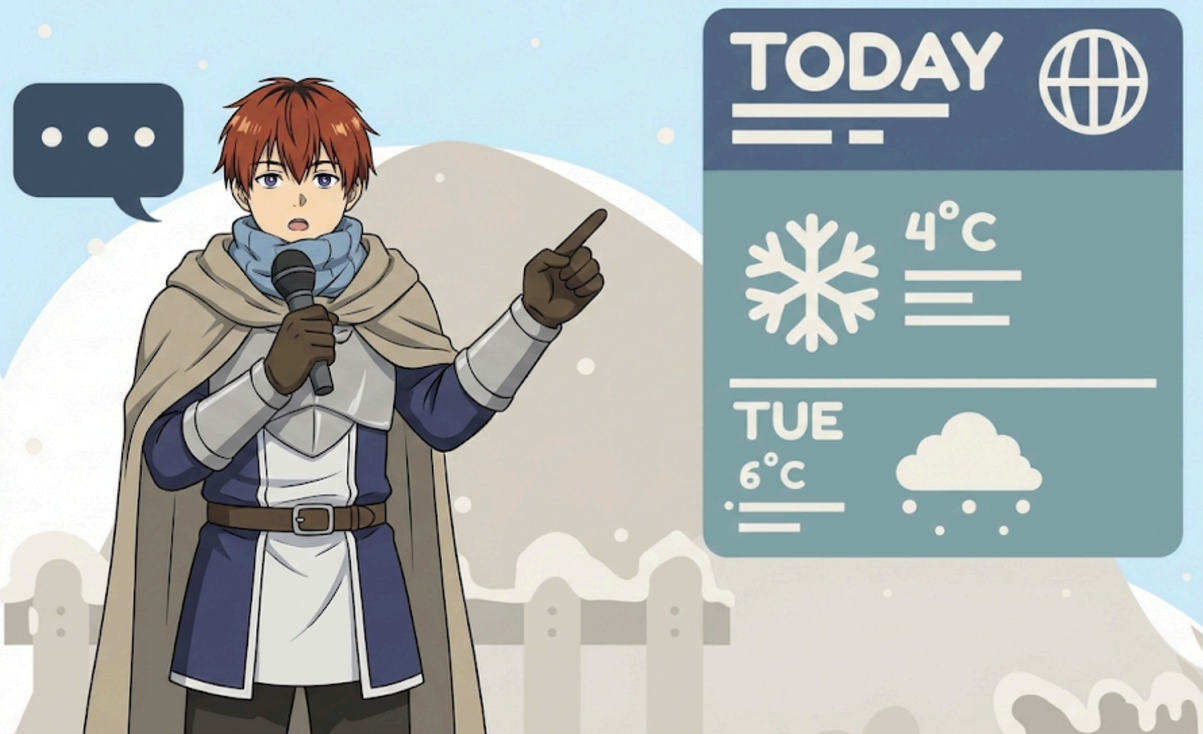
With a background in international collaboration and published work in software ecosystems,
I specialize in turning raw data into strategic assets.

WHAT WE GONNA DO??

This project analyzes daily weather data from the Philippines to predict maximum air temperature using machine learning techniques.

The dataset covers the period from 2020 to 2023 and includes key meteorological variables such as temperature, wind, precipitation, solar radiation, and evapotranspiration.

Based on historical patterns, the trained models generate forward-looking temperature forecasts for 2026 across three regions: North, Central, and South.



PROBLEM STATEMENT & MOTIVATION

The Philippines experiences strong seasonal and regional variations in temperature, which can impact agriculture, energy demand, and public health.

Traditional descriptive analysis is often insufficient to anticipate future temperature patterns, especially under changing climate conditions.

This project addresses the need for a predictive, data-driven approach to temperature forecasting by leveraging machine learning models trained on historical weather data.

DATASET OVERVIEW & SCOPE

The dataset consists of daily historical weather observations collected across multiple regions in the Philippines.
It spans from January 2020 to December 2023 and serves as the foundation for training and evaluating machine learning models.
Each observation represents a single day and contains a combination of temperature, atmospheric, and solar-related variables.

- Time range: 2020 – 2023 (daily frequency)
- Geographic scope: Philippines (North, Central, South)
- Target variable: Daily maximum temperature
- Number of observations: ~206,000+

FEATURE OVERVIEW (GROUPED)

city_name	datetime	weather_code	temperature_2m_max	temperature_2m_min	temperature_2m_mean	apparent_temperature_max	apparent_temperature_min	apparent_temperature_mean	sunrise	sunset
Alaminos	2020-01-01	1.0	32.4	23.1	26.9	35.6	25.1	30.1	2020-01-01T06:27	2020-01-01T17:38
Alaminos	2020-01-02	1.0	32.7	25.4	27.8	35.6	27.1	30.5	2020-01-02T06:28	2020-01-02T17:39
Alaminos	2020-01-03	3.0	31.3	23.2	26.5	33.3	25.2	28.7	2020-01-03T06:28	2020-01-03T17:40
Alaminos	2020-01-04	0.0	30.2	21.6	25.5	32.4	22.8	27.7	2020-01-04T06:28	2020-01-04T17:40
Alaminos	2020-01-05	1.0	31.9	23.8	26.7	33.1	25.3	28.6	2020-01-05T06:29	2020-01-05T17:41

daylight_duration	sunshine_duration	precipitation_sum	rain_sum	snowfall_sum	precipitation_hours	wind_speed_10m_max	wind_gusts_10m_max	wind_direction_10m_dominant	shortwave_radiation_sum
40263.84	36299.32	0.0	0.0	0.0	0.0	18.0	33.5	82.0	18.68
40277.32	36381.88	0.0	0.0	0.0	0.0	18.0	35.6	91.0	18.57
40292.00	36385.75	0.0	0.0	0.0	0.0	17.6	33.8	103.0	18.02
40307.85	36506.23	0.0	0.0	0.0	0.0	15.5	30.2	73.0	19.51
40324.83	36564.63	0.0	0.0	0.0	0.0	19.8	38.9	96.0	19.56

- Inside the dataset

FEATURE OVERVIEW (GROUPED)

The dataset contains multiple meteorological variables capturing atmospheric, thermal, and solar conditions.
These features are grouped based on their physical meaning to support interpretable and robust machine learning modeling.

Raw Meteorological Features

- Air temperature (max, min, mean, apparent)
- Wind speed, gusts, and direction
- Precipitation and snowfall (Snowfall is consistently zero for the Southeast Asia region)
- Solar radiation and evapotranspiration
- Daylight and sunshine duration

Engineered Features

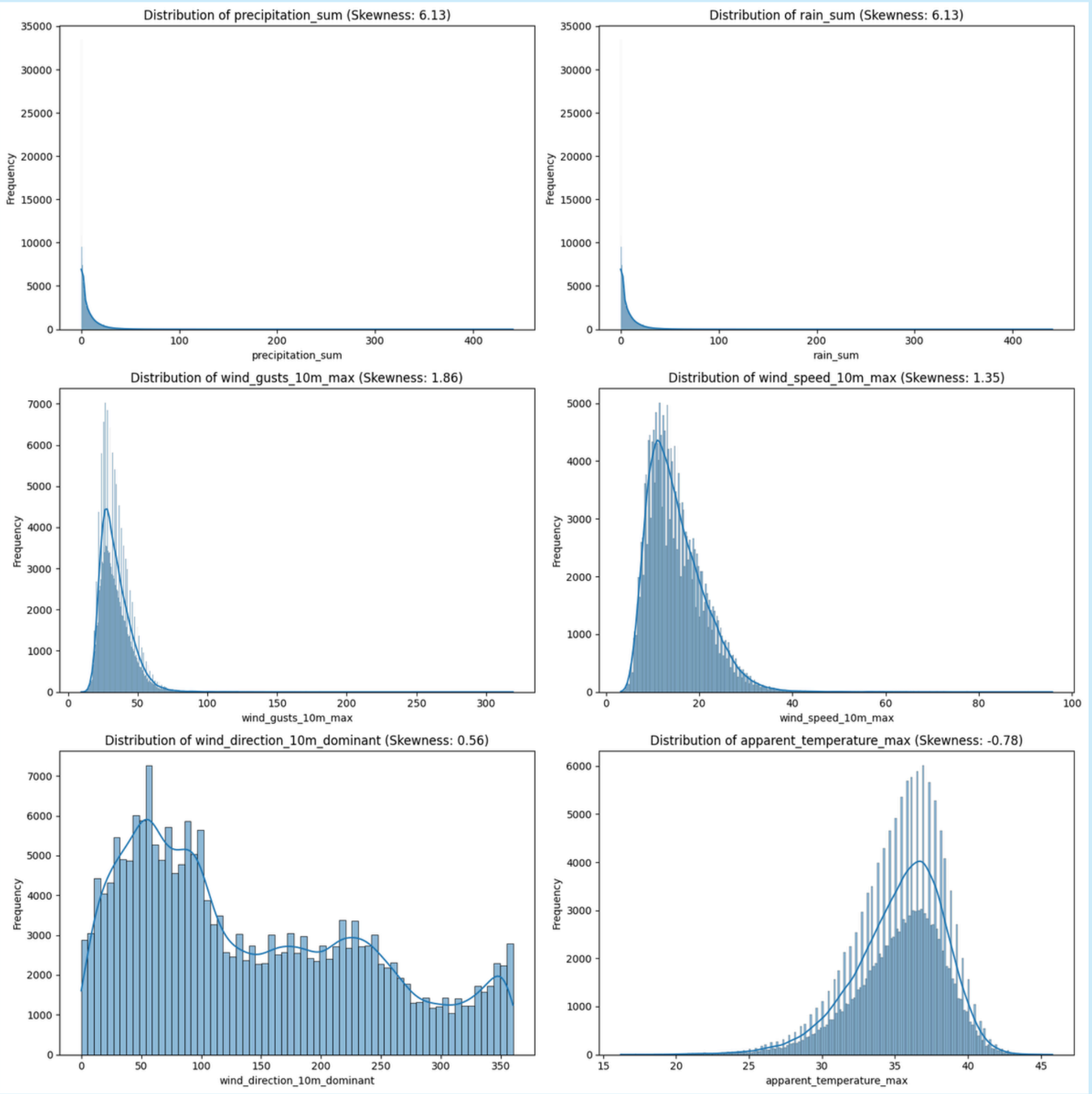
- Regional grouping derived from city-level data (North, Central, South)
- Time-based features extracted from datetime
- Filtered wind-related anomalies

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the distribution, relationships, and potential anomalies in the weather data.
Visual analysis was used to identify seasonal patterns, correlations between variables, and data quality issues prior to modeling.

The condition of our data now is:

1. Safe from missing values
2. Safe from duplicated data



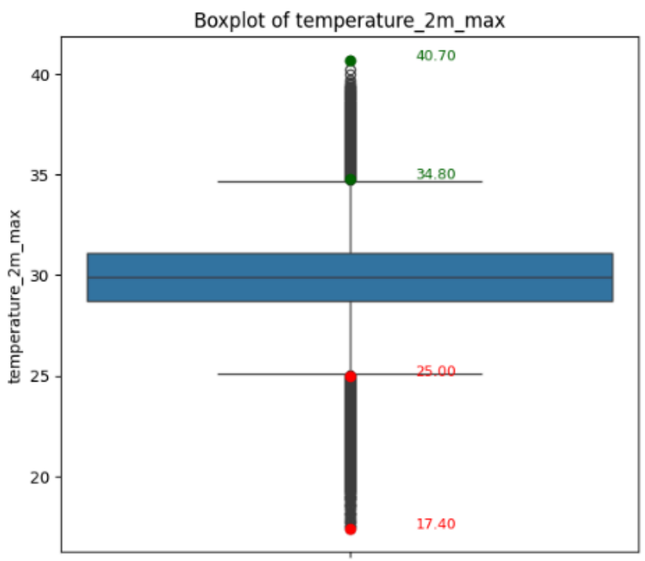
SKEWNESS

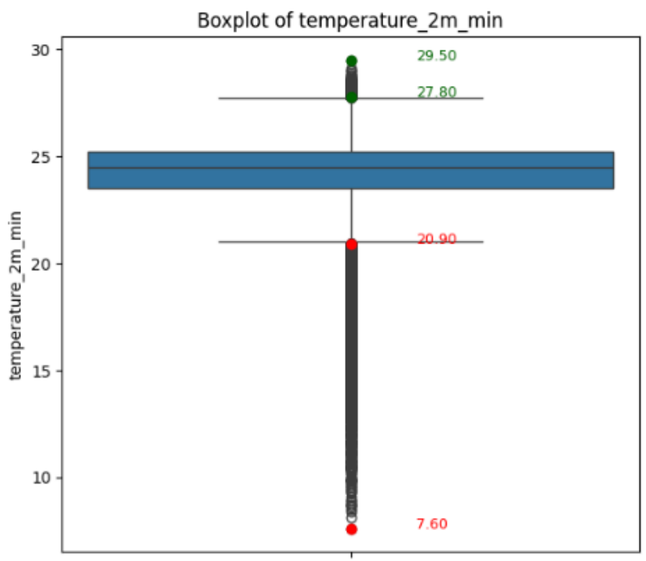
Feature	Meaning
precipitation_sum	Extremely right-skewed
rain_sum	Zero-inflated, heavy tail
wind_gusts_10m_max	Moderately skewed
wind_speed_10m_max	Right-skewed
wind_direction_10m_dominant	Near-uniform / circular
apparent_temperature_max	Slight left-skewed

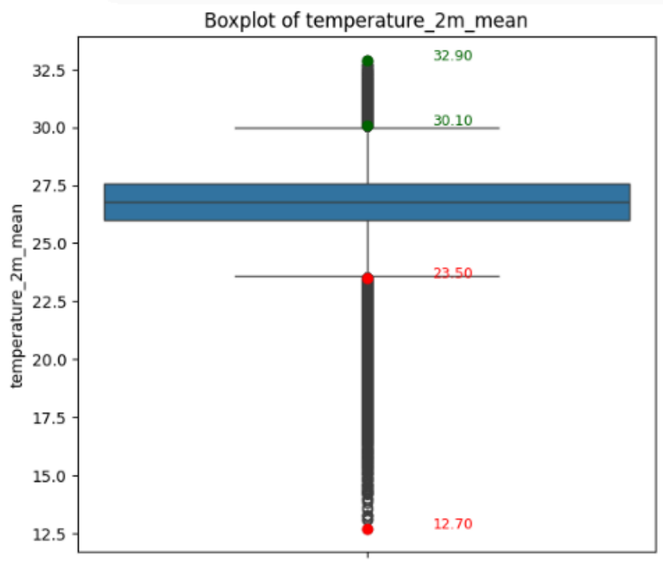
Exploratory data analysis shows that all numerical features exhibit skewed distributions, with no variables following a normal distribution.
Therefore, boxplot-based (IQR) analysis was consistently used for outlier identification, and median-based statistics were preferred for robust analysis.

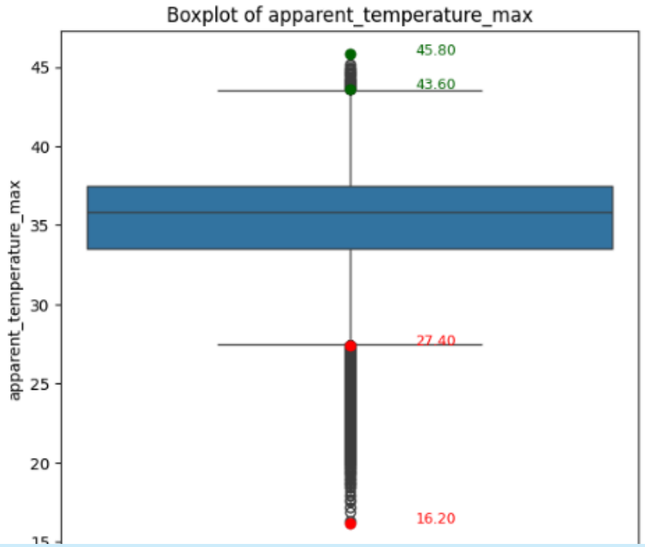
Points:

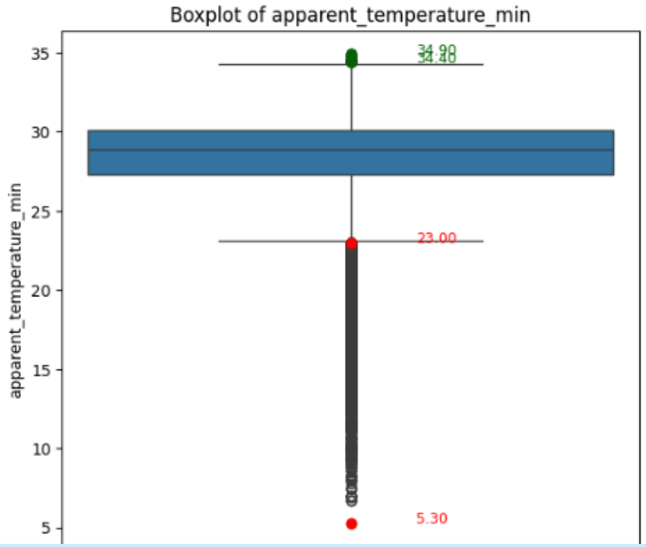
- **No normality assumption applied**
- **Boxplot used instead of z-score**
- **Median preferred over mean**
- **Outlier handling based on IQR thresholds**

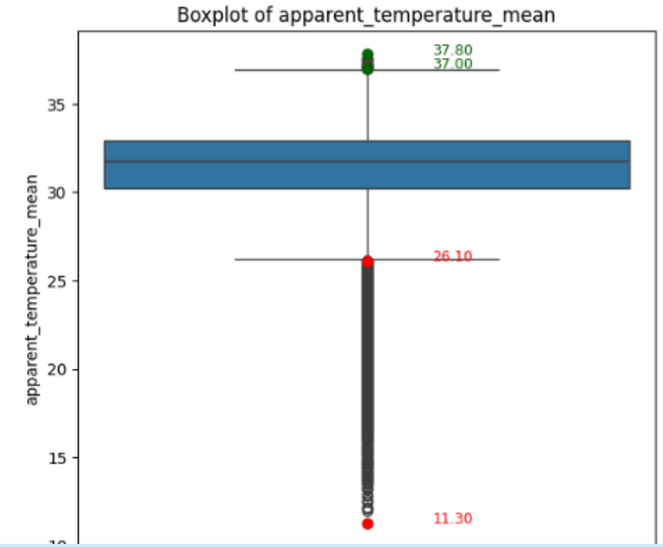












OUTLIERS

Column	Method	Q1	Q3	IQR	Lower Bound	Upper Bound	Num Lower Outliers	Min Lower Outlier	Max Lower Outlier	Num Upper Outliers	Min Upper Outlier	Max Upper Outlier
temperature_2m_max	IQR	28.70	31.10	2.40	25.10	34.70	3947	17.40	25.00	4527	34.80	40.70
temperature_2m_min	IQR	23.50	25.20	1.70	20.95	27.75	8557	7.60	20.90	394	27.80	29.50
temperature_2m_mean	IQR	26.00	27.60	1.60	23.60	30.00	8327	12.70	23.50	2476	30.10	32.90
apparent_temperature_max	IQR	33.50	37.50	4.00	27.50	43.50	3607	16.20	27.40	67	43.60	45.80
apparent_temperature_min	IQR	27.30	30.10	2.80	23.10	34.30	8254	5.30	23.00	17	34.40	34.90
apparent_temperature_mean	IQR	30.20	32.90	2.70	26.15	36.95	8774	11.30	26.10	29	37.00	37.80

DIFFERENT BETWEEN TEMPERATURE 2M VS APPARENT TEMPERATURE

Feature	Temperature 2m	Apparent Temperature
Measurement	Thermometer reading	Calculated formula
Main Factor	Heat of the air	Heat + Humidity + Wind
Best For	Science & Records	Safety & Comfort
Example	"It is 32°C."	"It feels like 41°C because it's humid."

OUTLIERS, TEMPERATURE

TEMPERATURE FEATURES

Let's check from temperatures features, do re-check based on their city!

The data is highly accurate and valid . The outliers perfectly match the geography and climate profiles of the Philippines.

Here is the breakdown:

1. Highlands (The Cold Outliers).

- **Baguio** (Min 7.6°C): Correct. Baguio is at ~1,500 meters elevation. While 7.6°C is very cold, it is within historical records (the record low is ~6.3°C). The absence of upper outliers here confirms it stays cool.
- **Malaybalay**: Located in the mountains of Mindanao, explaining the cool lower outliers**[1]**.

2. The "Heat Bowls" (The Hot Outliers)

Tuguegarao & Cabanatuan (Max ~39°C - 40°C): Correct. These cities are in inland valleys (**Cagayan Valley** and **Central Luzon plains**).

They are historically the hottest parts of the country because mountains block the breeze, trapping heat**[2]**.

3. Metro Manila (Urban Heat)

Quezon City/Manila (Max ~38°C): Correct. The high density of concrete creates an "Urban Heat Island" effect, making it hotter than surrounding rural areas.

Verdict: The dataset is clean. There are no impossible values (e.g., 100°C), and the extremes follow the physical geography of the country.

Temperature Outlier Summary by City (Colored)

city_name	Count_Lower	Min_Lower	Max_Lower	Count_Upper	Min_Upper	Max_Upper	Total
Baguio	4352	7.60	25.00	0	-	-	4352
Malaybalay	3358	14.80	25.00	0	-	-	3358
Marawi	2813	15.50	25.00	0	-	-	2813
Tagaytay City	1973	17.20	25.00	0	-	-	1973
Canlaon	1052	17.30	25.00	0	-	-	1052
Lipa City	910	17.50	25.00	2	35.60	35.70	912
Tabuk	826	14.50	25.00	32	30.10	37.00	858
Tuguegarao	431	16.50	25.00	224	27.80	38.80	655
Mabalacat City	304	18.00	25.00	318	30.10	38.80	622
Dagupan	460	15.80	25.00	157	30.10	39.10	617
Ilagan	474	16.20	25.00	136	28.00	38.30	610
Cauayan	418	15.40	25.00	158	27.80	38.60	576
San Carlos	454	17.90	25.00	3	34.80	35.30	457
Urdaneta	69	17.90	24.60	329	28.00	39.40	398
Tayabas	372	19.40	25.00	0	-	-	372
Palayan City	127	17.10	24.80	231	27.80	39.40	358
Gapan	76	18.10	24.80	265	28.20	40.20	341
Cabanatuan City	87	18.30	25.00	253	27.80	40.00	340
Tarlac City	43	19.10	23.50	271	28.10	39.70	314
Antipolo	108	17.30	24.70	181	30.10	39.10	289
Muñoz	18	19.90	24.90	266	27.80	40.70	284
Laoag	92	16.90	24.90	192	27.80	38.30	284
Angeles City	128	17.80	24.80	154	30.10	38.40	282
City of Marikina	70	17.80	25.00	197	30.10	38.60	267
Iligan City	261	18.70	24.90	0	-	-	261
Malabon	30	18.60	24.70	216	27.90	38.40	246
Caloocan City	30	18.60	24.70	215	27.90	38.40	245
Quezon City	43	18.40	24.50	190	27.80	38.20	233

sample of temperature outliers table

OUTLIERS, TEMPERATURE

APPARENTS TEMPERATURE FEATURES

1. Highlands (The Cold Outliers)

- Observation: Baguio (Min 7.6°C) and Malaybalay.
- Context: Baguio sits at ~1,500 meters elevation. A minimum of 7.6°C is consistent with historical records (record low is ~6.3°C). The absence of upper outliers here confirms the region's distinct highland climate [3].

2. The "Heat Bowls" (The Hot Outliers)

- Observation: Tuguegarao & Cabanatuan (Max ~39°C - 40°C).
- Context: These cities are located in inland valleys (Cagayan Valley and Central Luzon plains). Major mountain ranges block cooling sea breezes, effectively trapping heat and creating the hottest microclimates in the country [4].

3. Metro Manila (Urban Heat)

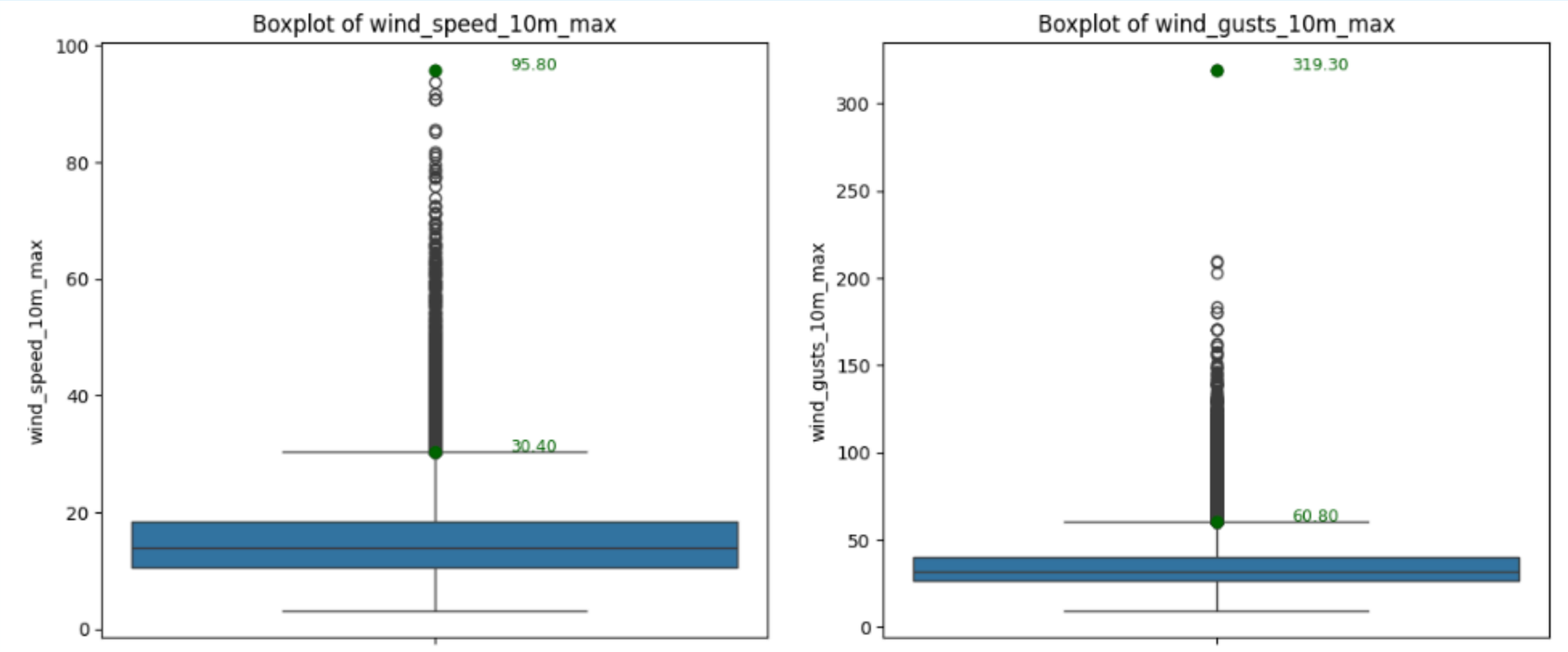
- Observation: Quezon City & Manila (Max ~38°C).
- Context: The high density of concrete and asphalt creates an "Urban Heat Island" effect, causing temperatures to remain significantly higher than in surrounding rural areas [5].

Verdict The dataset is clean. There are no impossible values (e.g., >55°C), and the extreme variances strictly follow the physical geography of the country.

city_name	Count_Lower	Min_Lower	Max_Lower	Count_Upper	Min_Upper	Max_Upper	Total
Baguio	4247	5.30	27.40	0	-	-	4247
Malaybalay	2825	15.00	27.40	0	-	-	2825
Tagaytay City	1938	14.80	27.40	0	-	-	1938
Marawi	1456	16.00	27.40	0	-	-	1456
Lipa City	955	17.60	27.40	0	-	-	955
Tabuk	682	14.70	27.40	0	-	-	682
Canlaon	557	18.30	27.40	0	-	-	557
Tayabas	498	18.70	27.40	0	-	-	498
Dagupan	427	16.20	27.40	13	37.10	45.10	440
Mabalacat City	440	17.60	27.40	0	-	-	440
Tuguegarao	387	17.50	27.40	3	44.00	44.50	390
Ilagan	367	17.30	27.40	3	37.30	43.90	370
Cauayan	357	16.30	27.40	2	43.80	44.80	359
Lucena	248	20.50	27.40	0	-	-	248
Laoag	243	16.90	27.40	2	43.60	44.50	245
Batac City	234	16.70	27.30	1	43.60	43.60	235
San Juan	218	19.70	27.40	0	-	-	218
Antipolo	198	17.90	27.40	2	37.20	43.70	200
Angeles City	199	17.60	27.40	0	-	-	199
Palayan City	183	17.00	27.30	4	37.10	44.50	187
Vigan	164	18.30	27.40	2	34.50	37.00	166
Muñoz	157	18.60	27.10	8	37.20	44.70	165
San Carlos	154	18.60	27.40	0	-	-	154
Maasin	152	20.50	27.40	0	-	-	152
Samal	142	19.30	27.30	0	-	-	142
Gapan	119	18.30	27.40	10	37.00	44.90	129
Cabanatuan City	114	18.70	27.30	12	37.20	45.80	126
Tarlac City	115	19.40	26.40	6	43.80	45.20	121
Urdaneta	106	17.90	26.10	8	37.00	44.70	114
Dasmariñas	113	20.70	27.20	0	-	-	113

sample of apparent temperature outliers table

OUTLIERS, WIND



OUTLIERS, WIND

Many outliers that actually pretty sus! The world extreme three-second average wind gust record of 113.3 m s-1, measured on Barrow Island, Australia![6].

And in this case! many of them actually 100 above? woaa??? and there is 300 m/s like seriously???

CHECK ANOMALY!
To assess extreme wind anomalies, the top five highest wind speed and wind gust records were examined against real-world events.

While major storms occurred on the corresponding dates, the reported magnitudes were not physically plausible as surface observations. Based on this assessment, wind speed values exceeding 32 m/s and wind gust values exceeding 70 m/s were flagged as invalid.

These records were retained in the dataset for transparency but excluded from the modeling stage due to their high frequency and lack of verifiable references

Wind Extreme Crosscheck by City (Physics-Based Thresholds)

	city_name	Variable	Rule	Threshold	Count	Min_Value	Max_Value	Total_Suspicious
0	Baybay	wind_gusts_10m_max	Wind Gust > 100 m/s	100	36	100.40	319.30	37
1	Tabaco	wind_gusts_10m_max	Wind Gust > 100 m/s	100	36	100.40	319.30	37
2	Baybay	wind_speed_10m_max	Wind Speed > 60 m/s	60	1	63.80	63.80	37
3	Tabaco	wind_speed_10m_max	Wind Speed > 60 m/s	60	1	63.80	63.80	37
4	Olongapo	wind_gusts_10m_max	Wind Gust > 100 m/s	100	9	100.10	157.00	9
5	Tagaytay City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	7	100.80	130.30	7
6	Santa Rosa	wind_gusts_10m_max	Wind Gust > 100 m/s	100	5	103.00	183.20	7
7	Navotas	wind_speed_10m_max	Wind Speed > 60 m/s	60	4	67.40	85.60	7
8	Navotas	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	107.60	124.60	7
9	Santa Rosa	wind_speed_10m_max	Wind Speed > 60 m/s	60	2	60.90	71.10	7
10	Lucena	wind_gusts_10m_max	Wind Gust > 100 m/s	100	5	100.40	140.00	6
11	Cabuyao	wind_gusts_10m_max	Wind Gust > 100 m/s	100	4	100.10	120.20	6
12	Iriga City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	4	100.80	209.90	6
13	Cavite City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	103.00	114.80	6
14	Cavite City	wind_speed_10m_max	Wind Speed > 60 m/s	60	3	68.60	81.40	6
15	Makati City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	118.80	139.70	6
16	Makati City	wind_speed_10m_max	Wind Speed > 60 m/s	60	3	63.00	77.40	6
17	Paranaque City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	118.80	139.70	6
18	Paranaque City	wind_speed_10m_max	Wind Speed > 60 m/s	60	3	63.00	77.40	6
19	Taguig	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	118.80	139.70	6
20	Taguig	wind_speed_10m_max	Wind Speed > 60 m/s	60	3	63.00	77.40	6
21	Cabuyao	wind_speed_10m_max	Wind Speed > 60 m/s	60	2	66.00	67.00	6
22	Iriga City	wind_speed_10m_max	Wind Speed > 60 m/s	60	2	61.20	93.70	6
23	Lucena	wind_speed_10m_max	Wind Speed > 60 m/s	60	1	60.90	60.90	6
24	Ligao	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	110.20	208.80	5
25	Mandaluyong City	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	120.60	149.00	5
26	Manila	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	120.60	149.00	5
27	Meycauayan	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	116.30	142.20	5
28	Pasig	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	115.20	133.90	5
29	Valenzuela	wind_gusts_10m_max	Wind Gust > 100 m/s	100	3	120.60	149.00	5

sample of wind outliers table

OUTLIERS, WIND

Wind speed

According to the Beaufort wind scale adopted by the World Meteorological Organization, surface wind speeds up to approximately 10.7 m/s (Beaufort force 5) are considered normal, non-storm conditions. Wind speeds exceeding this range indicate increasingly severe weather[7].

And for the hurricane is 32 m/s above!

	city_name	datetime	wind_speed_10m_max
105497	Ligao	2020-11-01	95.8
89426	Iriga City	2020-11-01	93.7
92759	Kabankalan	2021-12-17	91.8
64999	Dapitan	2021-12-16	90.7
101524	Lapu-Lapu City	2021-12-16	90.7
53311	Cebu City	2021-12-16	90.7
126361	Mandaue City	2021-12-16	90.7
136008	Navotas	2020-05-15	85.6
195028	Toledo City	2021-12-16	85.2
63538	Danao	2021-12-16	81.8

✅ **1) Ligao — 2020-11-01 — 95.8 m/s**
Region & Date: Early November 2020

Storm context:
Typhoon Goni (local name: Rolly) was active around Nov 1, 2020 and is recorded as one of the strongest typhoons of the year. Goni brought extremely strong winds and widespread damage to the Philippines.

- Reality check:
Super Typhoon Goni’s maximum sustained winds at landataall were extraordinarily high, but official surface station measurements rarely exceed ~70–80 m/s in open coastal exposure.

A value like ~95.8 m/s sustained at city surface level is not reported in official observations.

- Conclusion:
Storm exists, but value is likely overestimated.
Dataset number far exceeds typical documented winds.

OUTLIERS, WIND

✅ **2) Iriga City — 2020-11-01 — 93.7 m/s**
Context: Same date & region as Ligao.
Goni was tracking through the Philippines on Nov 1, 2020.

- Reality check.
Iriga City is inland relative to typhoon core track; even at strongest landataall, surface winds decay inland.
A 93.7 m/s wind speed at surface for an inland city is not supported by observational weather records.

- Conclusion:
Storm conditions were real, but the magnitude recorded is not realistic for that location's surface measurements.

✅ **3) Kabankalan — 2021-12-17 — 91.8 m/s**
Storm context: Mid-December 2021 was dominated by ***Super Typhoon Rai*** (local name: Odette) one of the strongest storms of 2021.

Rai's effects were widely reported across Visayas and parts of Mindanao on Dec 16–17.

- Reality check:
Rai brought very strong winds, but even Category 5 typhoons usually have maximum sustained surface winds ≤ ~60–70 m/s at land stations.
~91.8 m/s far exceeds verified surface documentation.

- Conclusion:
Storm present, but value is much higher than observed reality — likely data error.

	city_name	datetime	wind_speed_10m_max
105497	Ligao	2020-11-01	95.8
89426	Iriga City	2020-11-01	93.7
92759	Kabankalan	2021-12-17	91.8
64999	Dapitan	2021-12-16	90.7
101524	Lapu-Lapu City	2021-12-16	90.7
53311	Cebu City	2021-12-16	90.7
126361	Mandaue City	2021-12-16	90.7
136008	Navotas	2020-05-15	85.6
195028	Toledo City	2021-12-16	85.2
63538	Danao	2021-12-16	81.8

OUTLIERS, WIND

✅ 4) **Cebu City — 2021-12-16 — 90.7 m/s**
Storm context: Also during ***Super Typhoon Rai / Odette landataall*** period.
Cebu City experienced severe weather during the storm's passage.

- Reality check:
Damaging winds and infrastructure destruction were reported, but no official station data supports ~90.7 m/s at city surface height.

- Conclusion:
Storm conditions real, but hash magnitude is not supported by observational logs.

✅ **5) Navotas — 2020-05-15 — 85.6 m/s**
Storm context: Around May 14–15, 2020, ***Typhoon Vongfong*** (local name: Ambo) made landataall and affected eastern Philippines.

It was an active storm moving across Visayas/Luzon region.

- Reality check:
Typhoon Vongfong had significant winds (sustained up to ~50–60 m/s at worst coastal exposure prior to landataall), but an 85.6 m/s wind recorded at Navotas — especially far from landataall center — is not corroborated by verified historical measurements.

- Conclusion:
Storm was real, but dataset value is too high to be a plausible surface measurement.

	city_name	datetime	wind_speed_10m_max
105497	Ligao	2020-11-01	95.8
89426	Iriga City	2020-11-01	93.7
92759	Kabankalan	2021-12-17	91.8
64999	Dapitan	2021-12-16	90.7
101524	Lapu-Lapu City	2021-12-16	90.7
53311	Cebu City	2021-12-16	90.7
126361	Mandaue City	2021-12-16	90.7
136008	Navotas	2020-05-15	85.6
195028	Toledo City	2021-12-16	85.2
63538	Danao	2021-12-16	81.8

OUTLIERS, WIND

CONCLUSION!

- A large number of wind speed observations were identified as anomalous. Although storm events occurred on the corresponding dates, there is no reliable reference confirming surface wind speeds of such magnitude. As these values exceed physically plausible limits for storm conditions, they were excluded from the modeling stage[8],[9],[10].

City	Date	Storm Present?	Realistic Wind?
Ligao	2020-11-01	✔ Goni	✘ Value too high
Iriga City	2020-11-01	✔ Goni	✘ Too high
Kabankalan	2021-12-17	✔ Rai	✘ Too high
Cebu City	2021-12-16	✔ Rai	✘ Too high
Navotas	2020-05-15	✔ Vongfong	✘ Too high

OUTLIERS, WIND

For wind guts
use more than 70 m/s!

- Why 70 m/s IS the right cutoff
> 70 m/s gust at 10 m height:
is extraordinarily rare, exceeds almost all documented surface observations would imply near world-record conditions

1. **Baybay & Tabaco — 1 November 2020 — 319.3 m/s**

Storm context:
Around 1 November 2020, the Philippines was affected by ***Super Typhoon Goni*** (local name: Rolly), one of the strongest tropical cyclones of the 2020 season. Goni brought violent winds and heavy rainfall to parts of the Visayas and southern Luzon. Reports from international disaster monitoring agencies confirm the severity of the event.

Although Typhoon Goni had an exceptionally strong wind field, maximum wind gusts recorded at surface meteorological stations during the event were on the order of hundreds of kilometers per hour, not exceeding known physical limits for surface observations.

Reality check:
A reported wind gust of 319.3 m/s (approximately 1,150 km/h) is far beyond any verified or physically plausible surface wind measurement, even under the most extreme typhoon conditions.
For context, historical records show that tropical cyclone surface **wind gusts rarely exceed approximately 75–85 m/s** at exposed coastal stations. No known verified surface observation approaches a value of 319 m/s.

Conclusion:
The storm event (Typhoon Goni) was real.
However, the reported wind gust magnitude is unrealistic and not supported by any verified meteorological observation.

	city_name	datetime	wind_gusts_10m_max
17837	Baybay	2020-11-01	319.3
175625	Tabaco	2020-11-01	319.3
89426	Iriga City	2020-11-01	209.9
105497	Ligao	2020-11-01	208.8
46006	Carcar	2021-12-16	203.4
168320	Santa Rosa	2020-11-01	183.2
17847	Baybay	2020-11-11	180.4
175635	Tabaco	2020-11-11	180.4
126361	Mandaue City	2021-12-16	170.6
101524	Lapu-Lapu City	2021-12-16	170.6

OUTLIERS, WIND

2. **Iriga City — 1 November 2020 — 209.9 m/s**

Storm context:
This record corresponds to the same date as the Baybay and Tabaco observations and coincides with the landfall and peak intensity period of ***Typhoon Goni*** in the Philippines.

Reality check:
Typhoon Goni did produce very strong winds. However, sustained surface winds or gusts exceeding approximately **90–100 m/**s have not been documented at inland or near-coastal observation stations in the Visayas region.
A gust value of **209.9 m/s** (approximately 755 km/h) is physically impossible for surface wind measurements in tropical cyclones.

Conclusion:
The storm event was real.
The recorded gust magnitude is unrealistic and should not be interpreted as a true meteorological observation.

	city_name	datetime	wind_gusts_10m_max
17837	Baybay	2020-11-01	319.3
175625	Tabaco	2020-11-01	319.3
89426	Iriga City	2020-11-01	209.9
105497	Ligao	2020-11-01	208.8
46006	Carcar	2021-12-16	203.4
168320	Santa Rosa	2020-11-01	183.2
17847	Baybay	2020-11-11	180.4
175635	Tabaco	2020-11-11	180.4
126361	Mandaue City	2021-12-16	170.6
101524	Lapu-Lapu City	2021-12-16	170.6

3. **Ligao — 1 November 2020 — 208.8 m/s**

Storm context:
This observation is dated 1 November 2020, aligning with the maximum impact period of ***Typhoon Goni*** across southern Luzon and nearby regions.

Reality check:
While typhoon conditions were present, surface wind gusts in city-level observations exceeding **200 m/s** do not align with any verified meteorological records or historical data.
Such values are far beyond known physical limits for near-surface wind measurements during tropical cyclones.

Conclusion:
The storm event was real.
The reported wind gust magnitude is unrealistic and should be treated as anomalous.

OUTLIERS, WIND

5. **Carcar 16 December 2021 — 203.4 m/s**

Storm context:

On 16 December 2021, ***Super Typhoon Rai*** (local name: Odette) was active and made landfall in multiple areas of the central Philippines, including locations near Carcar, Cebu. Official reports indicate that Rai had maximum sustained winds of approximately 195 km/h (about 54 m/s), with gusts reaching up to approximately 270 km/h (about 75 m/s) near the storm core.

Reality check:

A reported gust of **203.4 m/s** (approximately 732 km/h) at Carcar is not supported by any verified typhoon measurement records.

Even in the strongest tropical cyclones, observed surface gusts rarely exceed **80–90 m/s** at exposed coastal stations, and are significantly lower in inland or urban environments.

Conclusion:

The storm event was real.

The reported gust magnitude is physically implausible and inconsistent with documented observations.

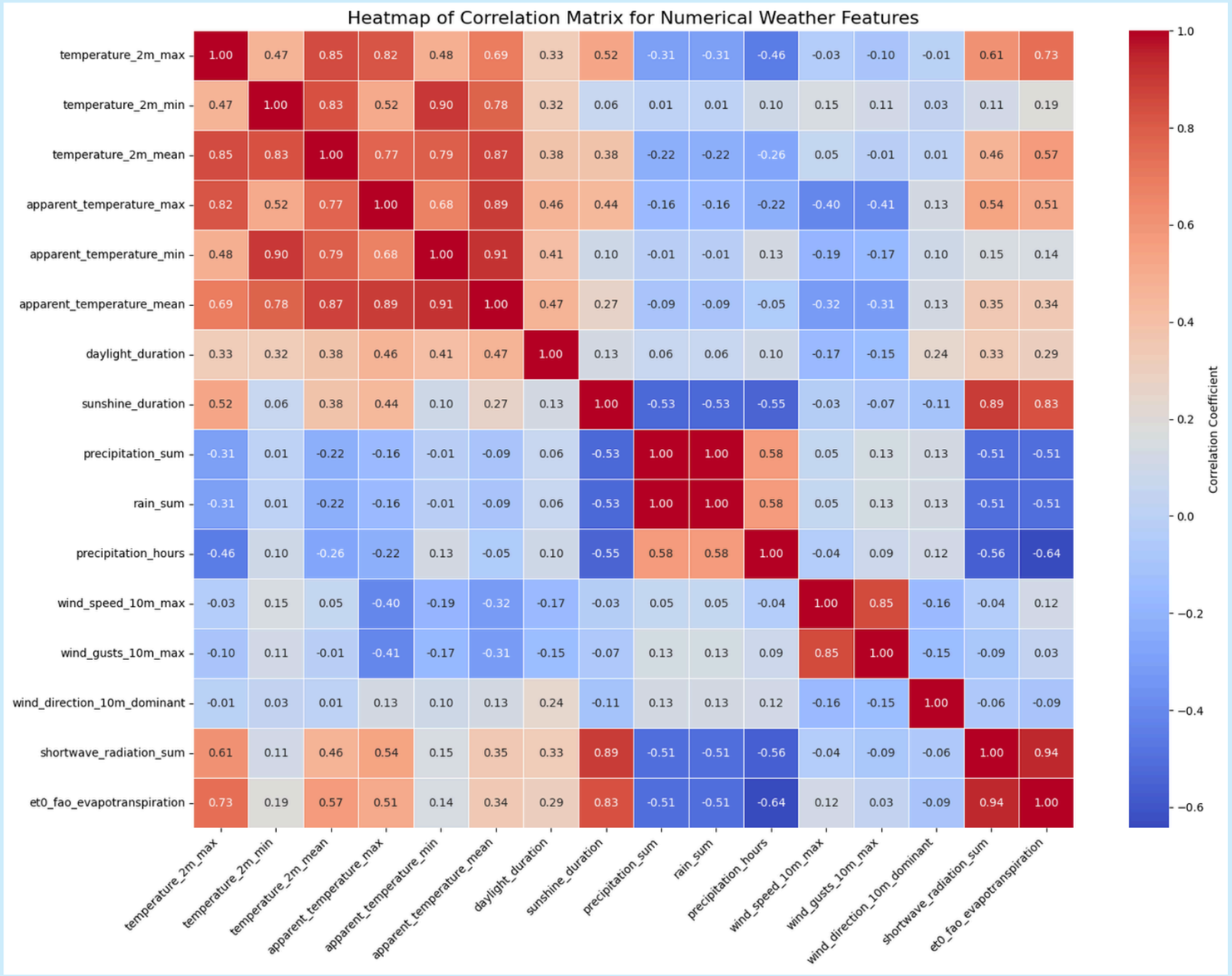
Final Conclusion

The storm events associated with the listed dates, including Typhoon Goni in November 2020 and Super Typhoon Rai in December 2021, were real and caused significant impacts across the Philippines.

However, the recorded wind gust magnitudes exceeding 200 m/s far exceed any known or verified surface wind observations. These values are therefore considered physically implausible and should be treated as data quality anomalies rather than real meteorological extremes[11],[12],[13].

	city_name	datetime	wind_gusts_10m_max
17837	Baybay	2020-11-01	319.3
175625	Tabaco	2020-11-01	319.3
89426	Iriga City	2020-11-01	209.9
105497	Ligao	2020-11-01	208.8
46006	Carcar	2021-12-16	203.4
168320	Santa Rosa	2020-11-01	183.2
17847	Baybay	2020-11-11	180.4
175635	Tabaco	2020-11-11	180.4
126361	Mandaue City	2021-12-16	170.6
101524	Lapu-Lapu City	2021-12-16	170.6

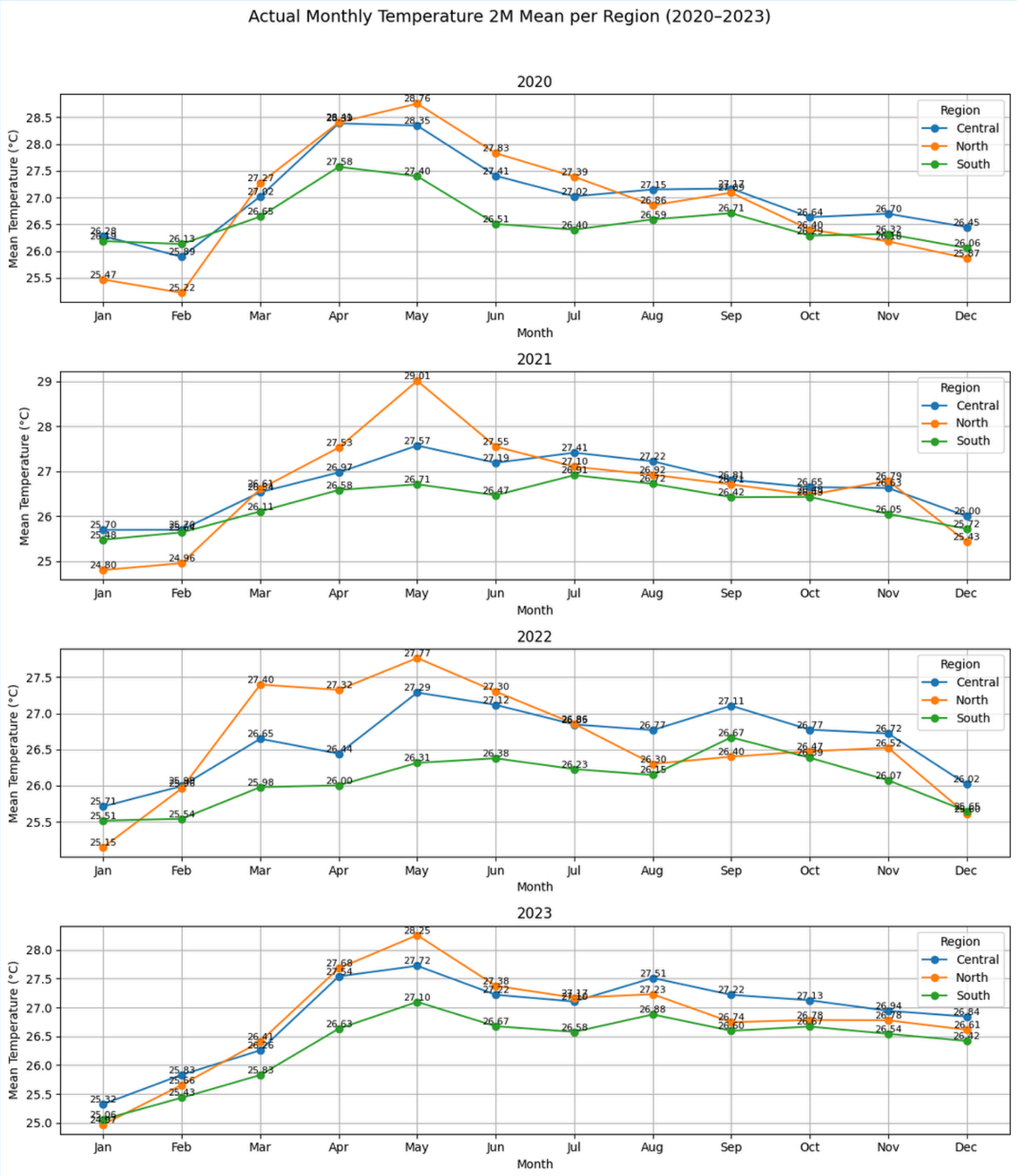
HEATMAP



As shown in the correlation analysis, the variables precipitation_sum and rain_sum exhibit a perfect correlation (1.00), indicating that they contain identical information in this dataset.

Although the two variables are conceptually different, their values are the same due to the absence of non-rain precipitation events.

To avoid redundancy and potential multicollinearity issues, one of these variables was removed prior to modeling. This step is particularly important for models sensitive to correlated features, while tree-based models are generally less affected by such redundancy but still better to drop one of them because on our data POV they actually same.



MEAN TEMPERATURE

As shown in the monthly temperature visualization (2020–2023), all regions exhibit a clear seasonal pattern, with temperatures increasing from the start of the year, peaking around April–May, and then declining toward year-end.

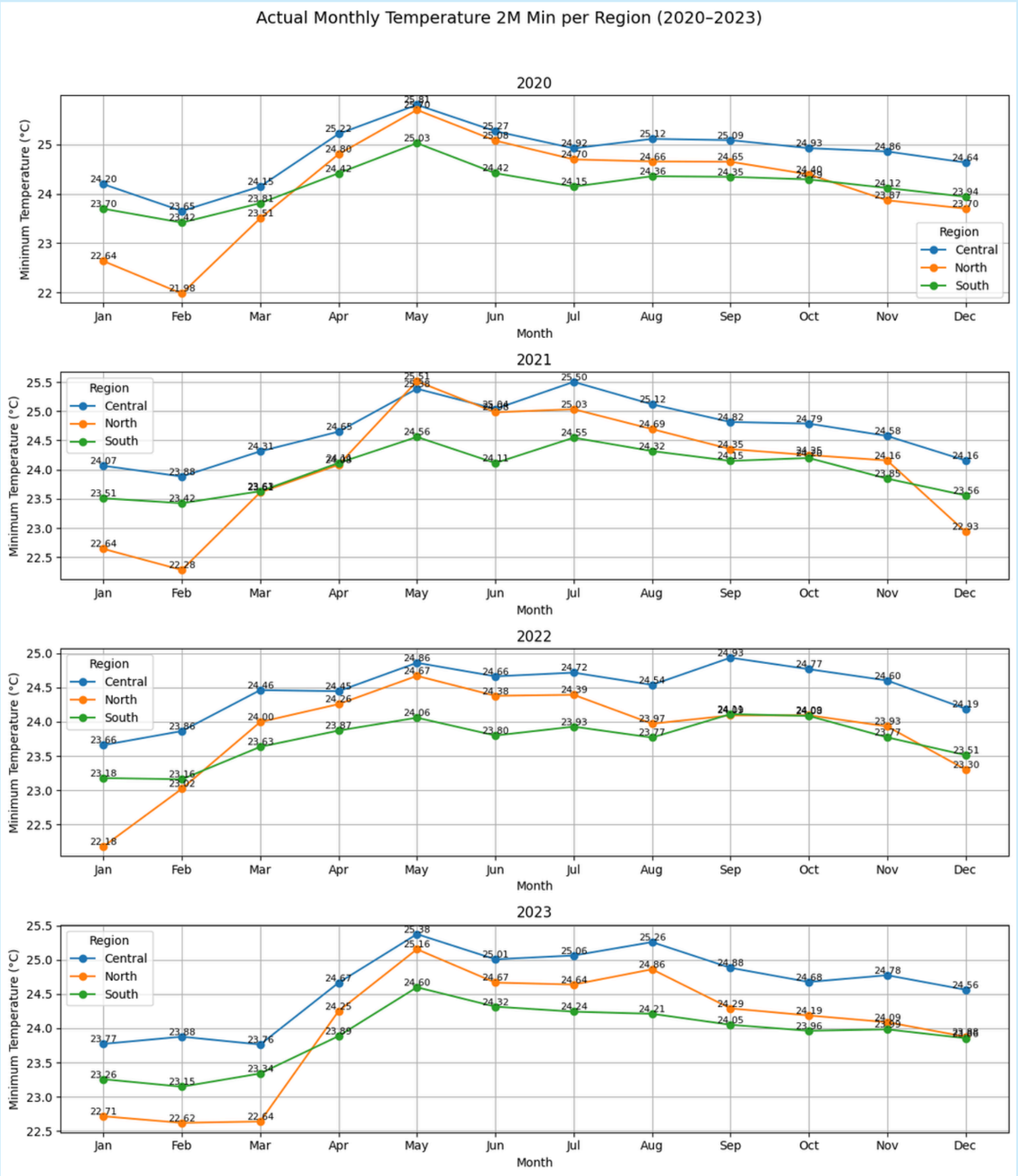
The North region consistently records higher mean temperatures, particularly during peak months, indicating stronger seasonal heating effects.

In contrast, the South region shows the most stable temperature profile, with smaller monthly fluctuations, while the Central region lies between the two.

Although some interannual variation is observed (e.g., higher peaks in 2021 and 2023), the overall temporal trend remains consistent, suggesting that temperature dynamics are primarily driven by recurring seasonal patterns rather than random variability.

This consistency supports the reliability of the dataset for time-based modeling, and the clear regional differences justify the inclusion of region-specific features to capture spatial variability that would otherwise be obscured by aggregation.

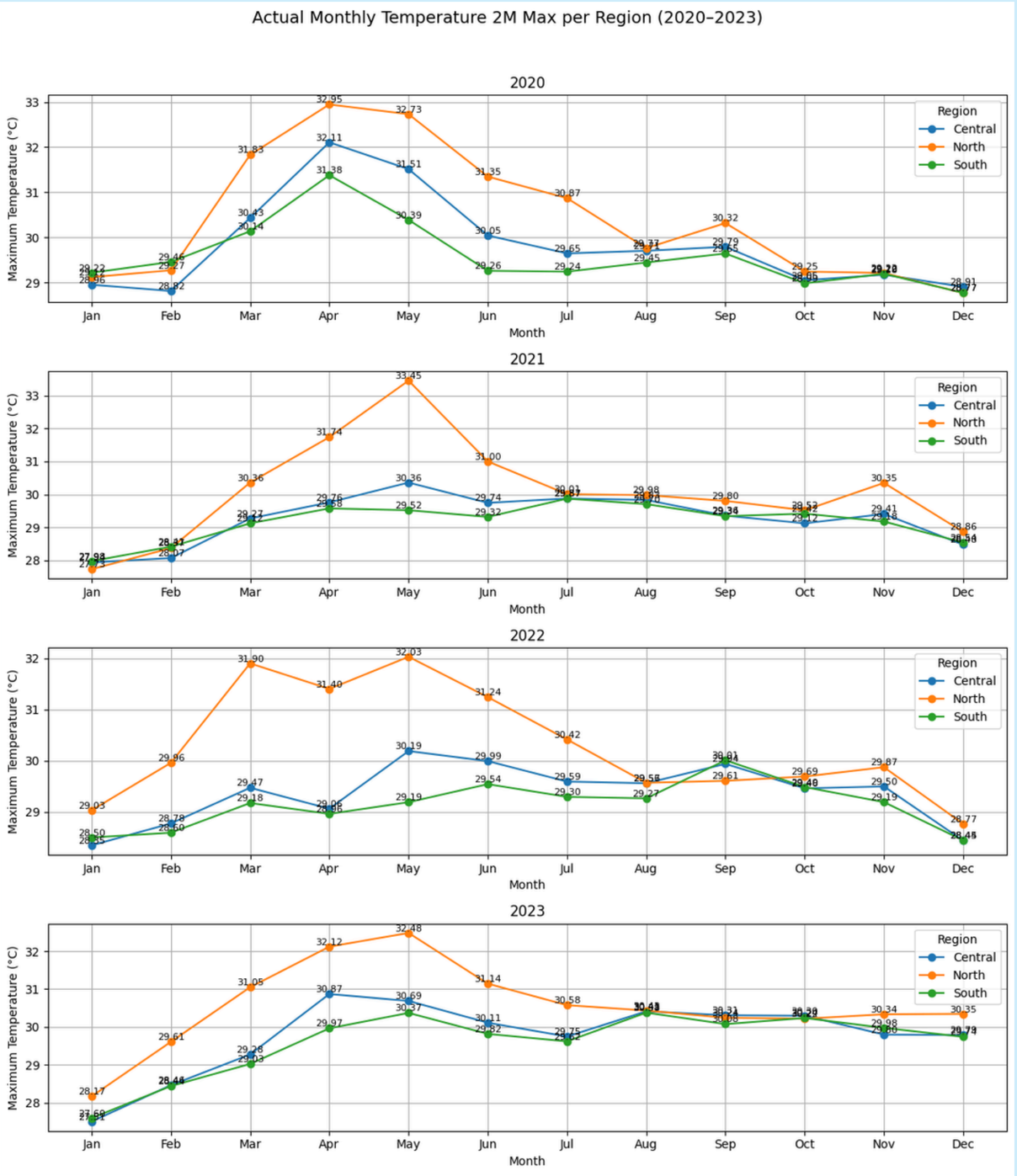
MIN TEMPERATURE



As shown in the monthly minimum temperature visualization (2020–2023), all regions exhibit a clear seasonal pattern, with minimum temperatures rising toward April–May and declining afterward.

The Central region consistently records the highest minimum temperatures, indicating warmer nighttime conditions, while the North region shows the lowest values, especially in January–February. In contrast, the South region displays the most stable pattern with limited month-to-month variability.

Overall, the temperature trends remain highly consistent across years, suggesting that minimum temperature variations are mainly driven by seasonal and regional factors. This consistency supports the use of temporal and region-based features in modeling and indicates that minimum temperature is a relatively predictable variable.



MAX TEMPERATURE

As shown in the monthly maximum temperature visualization (2020–2023), all regions display a pronounced seasonal pattern, with maximum temperatures rising early in the year, peaking around April–May, and gradually decreasing afterward.

The North region consistently records the highest maximum temperatures, particularly during peak months, indicating stronger daytime heating compared to the Central and South regions. The South region remains relatively stable, with lower peaks and smaller monthly variability, while the Central region shows moderate fluctuations.

Despite some interannual differences (e.g., higher peaks in 2021 and 2023), the overall temporal trend is stable across years, suggesting that maximum temperature variability is primarily driven by recurring seasonal and regional climatic factors.

This clear seasonality and regional differentiation support the use of time-based and region-specific features in modeling, and indicate that maximum temperature patterns are systematic rather than random.

PREPARING BEFORE CREATE THE ML MODEL!

Before we go through modeling, let's combine the city columns because we got a lot of city.

so it's better to grouping it to based on North, Central and South region.

```
north = [
    "Mabalacat City","Lipa City","Lucena","Bacoor","Angeles City","Antipolo",
    "Alaminos","Balanga","Baguio","Batac City","Batangas City","Biñan",
    "Cabanatuan City","Caloocan City","Candon","Cauayan","Cavite City",
    "Calamba","Calapan","Dagupan","Gapan","General Trias","Ilagan","Imus",
    "Iriga City","Isabela","Laoag","City of Marikina","Malolos","Makati City",
    "Malabon","Meycauayan","Muñoz","Naga","Navotas","Olongapo","Palayan City",
    "Paranaque City","Pasig","Puerto Princesa City","Quezon City","San Fernando",
    "San Jose","San Jose del Monte","San Juan","San Pablo","San Pedro",
    "Santa Rosa","Santo Tomas","Tabaco","Tabuk","Tagaytay City","Taguig",
    "Tanauan","Tarlac City","Tayabas","Tuguegarao","Urdaneta","Valenzuela",
    "Vigan","Las Piñas","Manila"
]

central = [
    "Bacolod","Bago City","Bais","Bayawan","Baybay","Borongan","Bogo",
    "Calbayog City","Canlaon","Carcar","Catbalogan","Cebu City","Cadiz",
    "Danao","Dumaguete","Escalante","Himamaylan","Iloilo City","Kabankalan",
    "La Carlota","Lapu-Lapu City","Mandaue City","Ormoc","Roxas","Sagay",
    "San Carlos","Silay City","Sipalay","Tacloban City","Talisay","Tanjay",
    "Toledo City","Victorias","City of Masbate","City of Passi",
    "City of Sorsogon"
]

south = [
    "Butuan","Cabadbaran","Bayugan","Bislig","Cagayan de Oro","Cotabato",
    "Dapitan","Davao","Digos","Dipolog","El Salvador","General Santos",
    "Gingoog City","Iligan City","Kidapawan","Koronadal","Lamitan","Malaybalay",
    "Marawi","Mati","Oroquieta","Ozamiz City","Pagadian","Panabo","Samal",
    "Surigao City","Tacurong","Tandag","Tangub","Valencia","Zamboanga City"
]
```

PREPARING BEFORE CREATE THE ML MODEL!

✅ **Drop Decisions**
We gonna drop useless column, and they are:

1. Weather Code
Weather code features were removed because they represent categorical daily weather conditions. Since the data were aggregated to a monthly level, these daily categorical indicators no longer retain meaningful information and could introduce noise into the model.

2. Rain Sum
The rain_sum variable was dropped due to its near-perfect correlation (≈ 1.00) with precipitation_sum. Keeping both features would introduce redundant information and increase the risk of multicollinearity without improving predictive performance.

3. City Name
The city_name feature was excluded because it has high cardinality and primarily serves as an identifier rather than a predictive feature. Spatial information is already captured through the region variable, which provides a more generalized and climate-relevant representation for modeling.

MACHINE LEARNING MODEL

A. Time-Based Train–Test Split

- Train = 2020 - 2022
- Test = 2023

Why not use an 80/20 split?
An 80/20 random split is not appropriate for time-dependent data because it would break the temporal order and potentially introduce data leakage.

Why use 2023 as the test set?
The year 2023 is used as a proxy for the future to evaluate how well the model generalizes to unseen time periods.

MACHINE LEARNING MODEL

A. Time-Based Train–Test Split

- Train = 2020 - 2022
- Test = 2023

Why not use an 80/20 split?
An 80/20 random split is not appropriate for time-dependent data because it would break the temporal order and potentially introduce data leakage.

Why use 2023 as the test set?
The year 2023 is used as a proxy for the future to evaluate how well the model generalizes to unseen time periods.
The dataset is split using a datetime boundary to preserve temporal order.
The year 2023 is used as a temporal hold-out set to evaluate model generalization.
Categorical encoding is applied after splitting to prevent data leakage.

MACHINE LEARNING MODEL

B. FILTER WIND SPEED AND GUTS THAT ACTUALLY SUS

- Max Wind Speed = 32
- Max Wind Guts = 70

C. ENCODING OHE (ONE HOT ENCODING)

- Only for Region

D.Cyclical Encoding (Sine–Cosine Encoding)

- datetime
- convert it to month (month_sin and month_cos)

MACHINE LEARNING MODEL

What are month_sin and month_cos?
month_sin and month_cos are cyclical representations of the month.
Months are cyclical, not linear. After December (month 12), the cycle returns to January (month 1). January and December are close in time, even though their numerical values (1 and 12) are far apart.

To represent this cyclical nature correctly, each month is mapped onto a circle using sine and cosine functions. This allows the model to understand that the end and the beginning of the year are connected.

How to read the values?
For example:

- Month 1 (January):

month_sin ≈ 0.50
month_cos ≈ 0.87
This represents the early part of the year and places January close to December on the circle.

- Month 3 (March):

month_sin ≈ 1.00
month_cos ≈ 0.00
This corresponds to the top of the cycle.

- Month 6 (June):

month_sin ≈ 0.00
month_cos ≈ -1.00
This represents the middle of the year.

month	month_sin	month_cos
1	5.000000e-01	8.660254e-01
2	8.660254e-01	5.000000e-01
3	1.000000e+00	6.123234e-17
4	8.660254e-01	-5.000000e-01
5	5.000000e-01	-8.660254e-01
6	1.224647e-16	-1.000000e+00
7	-5.000000e-01	-8.660254e-01
8	-8.660254e-01	-5.000000e-01
9	-1.000000e+00	-1.836970e-16
10	-8.660254e-01	5.000000e-01
11	-5.000000e-01	8.660254e-01
12	-2.449294e-16	1.000000e+00

MACHINE LEARNING MODEL

WHY IS THIS USEFUL FOR MACHINE LEARNING?

Using the month number directly (1 to 12) creates a false distance, where December (12) appears far from January.

By using month_sin and month_cos, December and January have similar numerical representations. This preserves seasonal continuity and allows machine learning models to learn smooth seasonal patterns without artificial breaks.

This approach is particularly useful when predicting future periods, such as the year 2026, because the model learns seasonal behavior rather than memorizing specific years.

month	month_sin	month_cos
1	5.000000e-01	8.660254e-01
2	8.660254e-01	5.000000e-01
3	1.000000e+00	6.123234e-17
4	8.660254e-01	-5.000000e-01
5	5.000000e-01	-8.660254e-01
6	1.224647e-16	-1.000000e+00
7	-5.000000e-01	-8.660254e-01
8	-8.660254e-01	-5.000000e-01
9	-1.000000e+00	-1.836970e-16
10	-8.660254e-01	5.000000e-01
11	-5.000000e-01	8.660254e-01
12	-2.449294e-16	1.000000e+00

MACHINE LEARNING MODEL

WHY IS THIS USEFUL FOR MACHINE LEARNING?

Using the month number directly (1 to 12) creates a false distance, where December (12) appears far from January.

By using month_sin and month_cos, December and January have similar numerical representations. This preserves seasonal continuity and allows machine learning models to learn smooth seasonal patterns without artificial breaks.

This approach is particularly useful when predicting future periods, such as the year 2026, because the model learns seasonal behavior rather than memorizing specific years.

month	month_sin	month_cos
1	5.000000e-01	8.660254e-01
2	8.660254e-01	5.000000e-01
3	1.000000e+00	6.123234e-17
4	8.660254e-01	-5.000000e-01
5	5.000000e-01	-8.660254e-01
6	1.224647e-16	-1.000000e+00
7	-5.000000e-01	-8.660254e-01
8	-8.660254e-01	-5.000000e-01
9	-1.000000e+00	-1.836970e-16
10	-8.660254e-01	5.000000e-01
11	-5.000000e-01	8.660254e-01
12	-2.449294e-16	1.000000e+00

MODEL

We Chose Three Tree-Based Models

- Decision Tree
- Random Forest
- XGBoost

WHY WE CHOOSE THEM?

- **Decision Tree (Baseline Model)**

The Decision Tree model was selected as the baseline model in this study. It provides a simple and interpretable structure that helps illustrate how input features are split to generate predictions. Despite its simplicity, a single decision tree is capable of capturing non-linear relationships within the data.
The primary purpose of using a Decision Tree is to establish a reference point for model performance. By evaluating this baseline, it becomes possible to quantify the performance improvements achieved by more advanced ensemble-based methods.

- **Random Forest (Stable Ensemble Model)**

Random Forest was chosen to address the limitations of a single Decision Tree, particularly its tendency to overfit the training data. This model constructs multiple decision trees using bootstrap sampling and aggregates their predictions through averaging.
By combining many trees, Random Forest reduces variance and produces more stable and reliable predictions. This makes it well-suited for noisy, real-world weather data, especially when working with a limited historical time span such as the 2020–2023 period.
As a result, Random Forest serves as a robust intermediate model between simple baseline approaches and more complex boosting techniques.

- **XGBoost (High-Performance Model)**

XGBoost was selected as the high-performance model in this analysis. Unlike bagging-based methods, XGBoost builds trees sequentially, where each new tree is trained to correct the errors made by previous ones.
This boosting strategy allows XGBoost to capture complex feature interactions and subtle non-linear patterns in the data. In many predictive modeling tasks, XGBoost consistently delivers superior accuracy compared to other tree-based methods.
In this study, XGBoost represents the upper performance bound of tree-based machine learning models and serves as the benchmark for optimal predictive performance.

WHAT WE DO?

- **Target:**
  - "temperature_2m_min",
  - "temperature_2m_mean",
  - "temperature_2m_max"
- **Drop:**
  - "apparent_temperature_max",
  - "apparent_temperature_min",
  - "apparent_temperature_mean"

The dropped columns were excluded because they:

- Are explicit prediction targets
- Act as direct or indirect proxies for the targets
- Introduce data leakage and invalidate model evaluation

Removing these variables ensures that the model learns genuine relationships between atmospheric drivers and daily temperature characteristics, resulting in reliable and interpretable performance estimates.

RESULTS

Non-Tuning

- The best model based on non-hyperparameter tuning is XGBoost with Score of
- R2 = 0.734893 that actually better rather than others model, Show the model is learning and not memorizing. with MAE = 0.56 Celcius and RMSE = 0.77 Celcius

Hyperparameter Tuning

- Based on the evaluation results, XGBoost without extensive hyperparameter tuning achieved the best overall performance, with the highest R^2 and the lowest error values.
- Random Forest also showed strong and stable performance using a baseline configuration.
- Hyperparameter tuning significantly improved Decision Tree performance, while tuning XGBoost resulted in reduced generalization, indicating that its baseline configuration was already well-optimized.

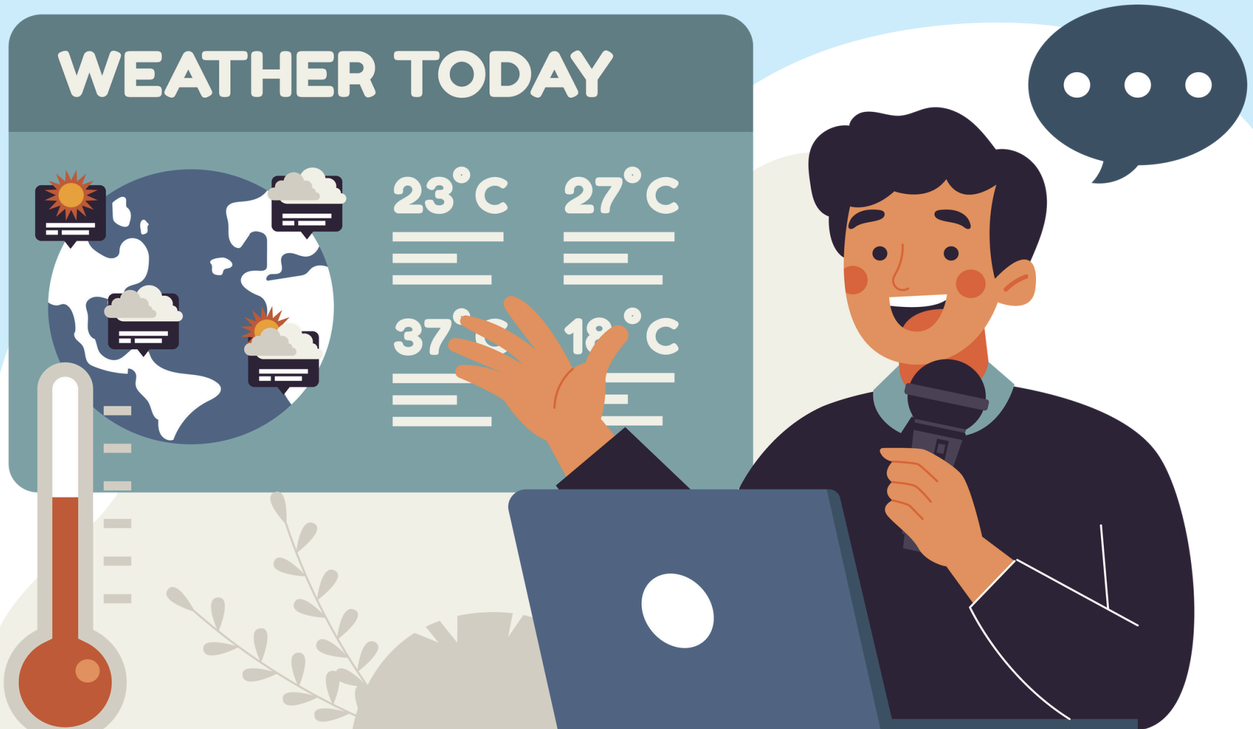
	R2	MAE	RMSE
Decision Tree	0.418474	0.987094	1.353206
Random Forest	0.719331	0.689988	0.941017
XGBoost	0.734893	0.669261	0.912266

Non - Tuning Results

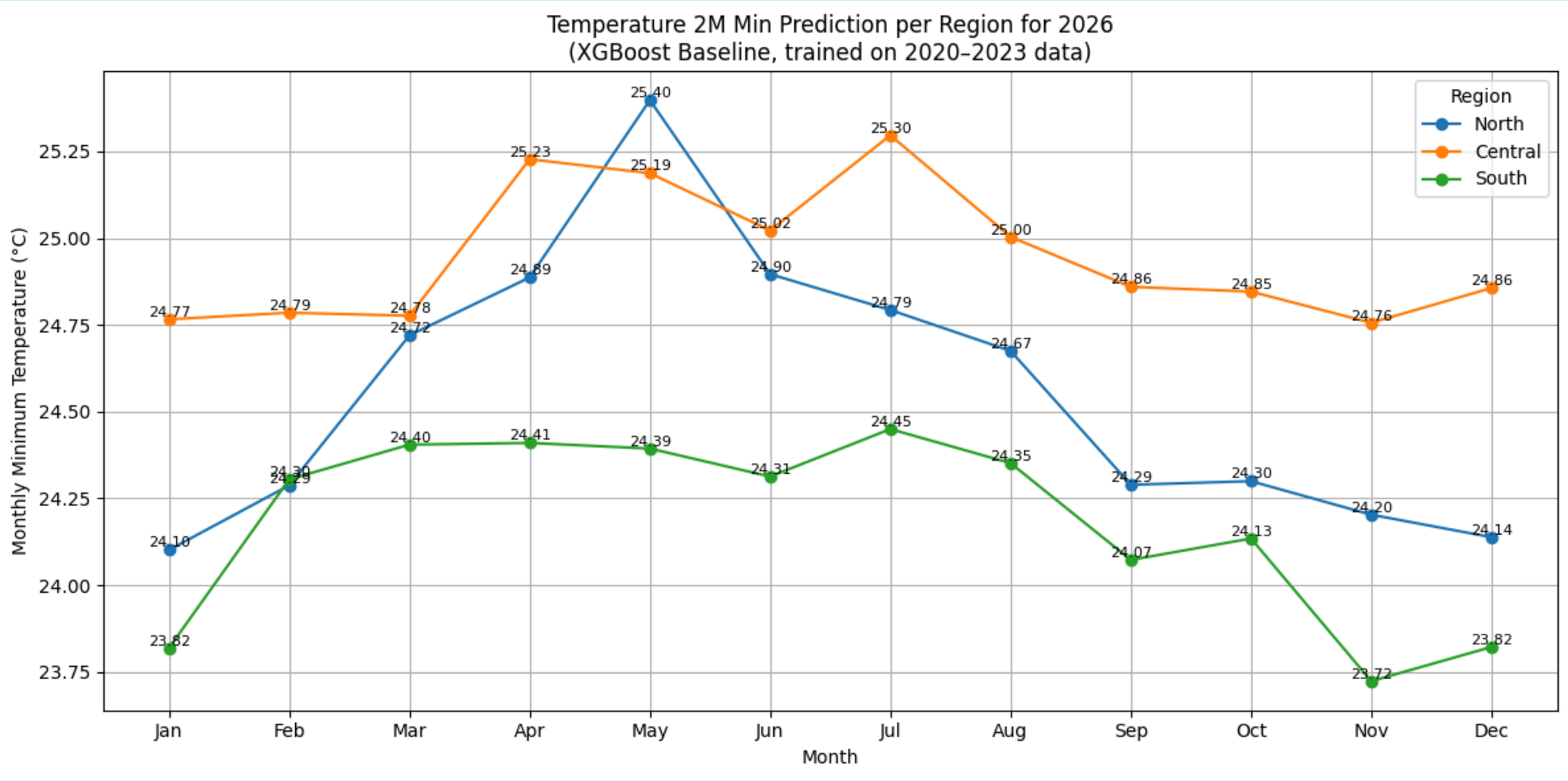
	R2	MAE	RMSE
Decision Tree (Tuned)	0.552606	0.867859	1.199438
Random Forest (Tuned)	0.718718	0.690784	0.942024
XGBoost (Tuned)	0.688646	0.719445	0.990002

Tuning Results

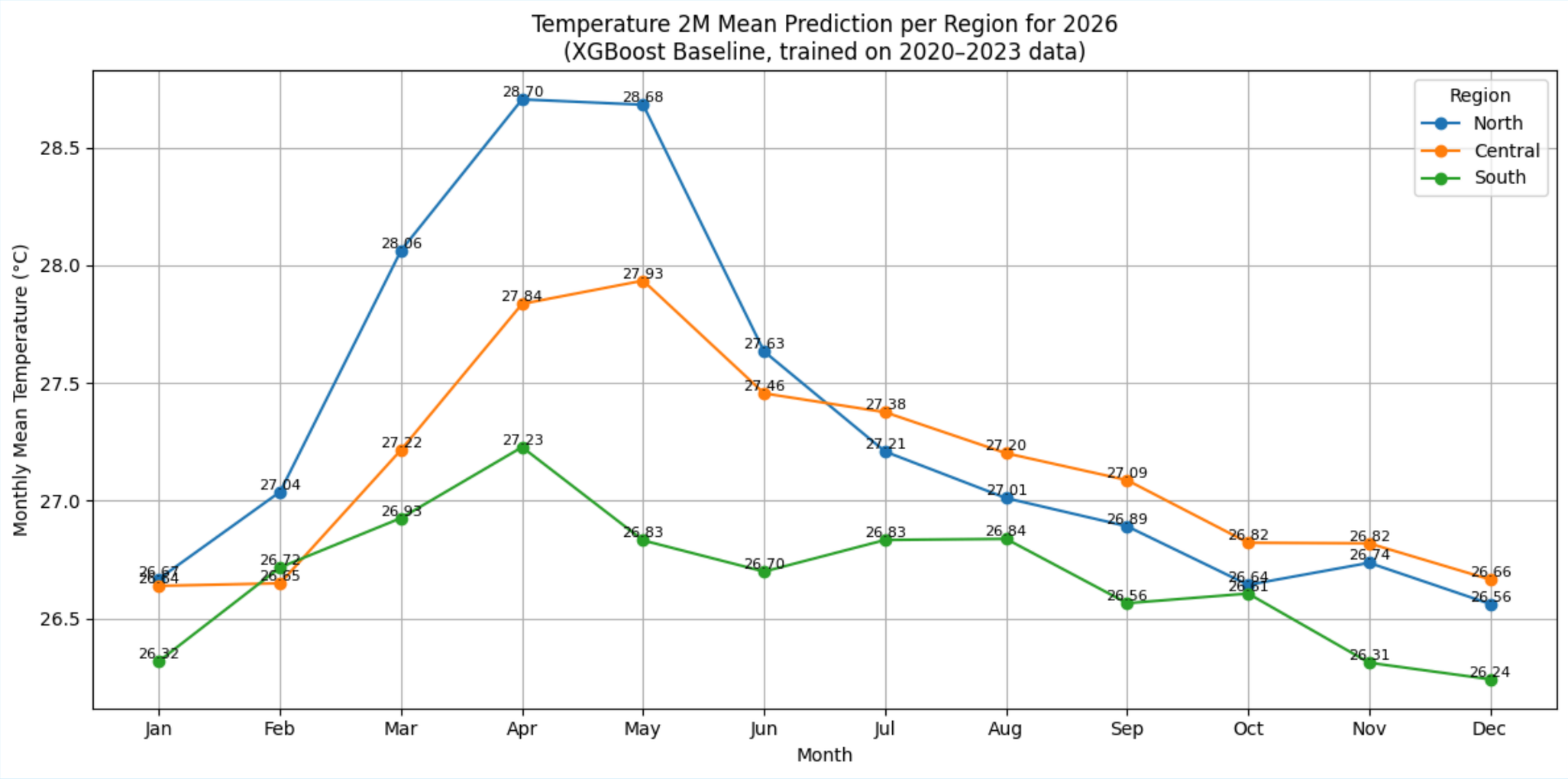
LINESCHART PREDICTION USING XGBOOST!



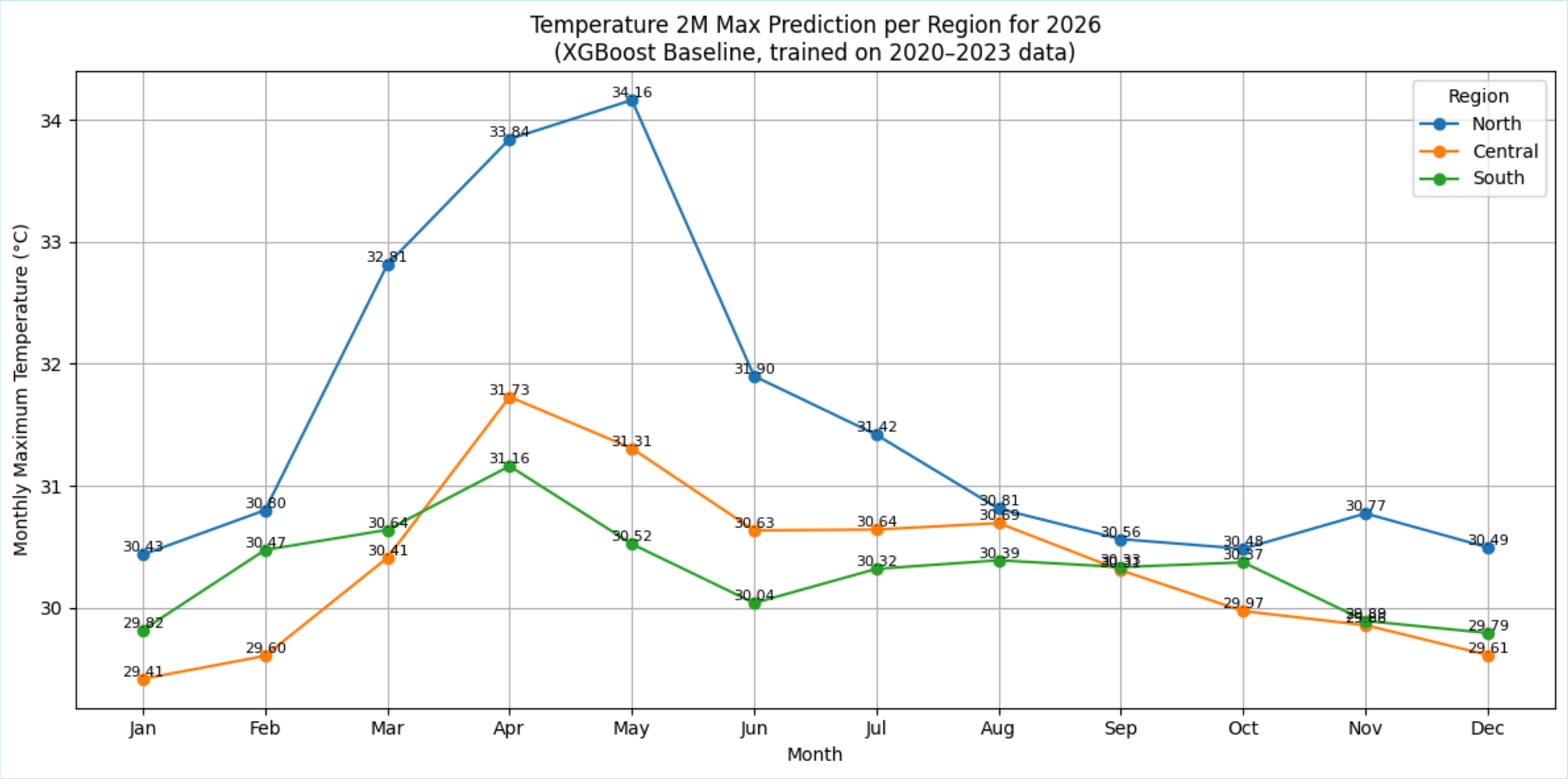
LINESCHART PREDICTION USING XGBOOST!



LINESCHART PREDICTION USING XGBOOST!

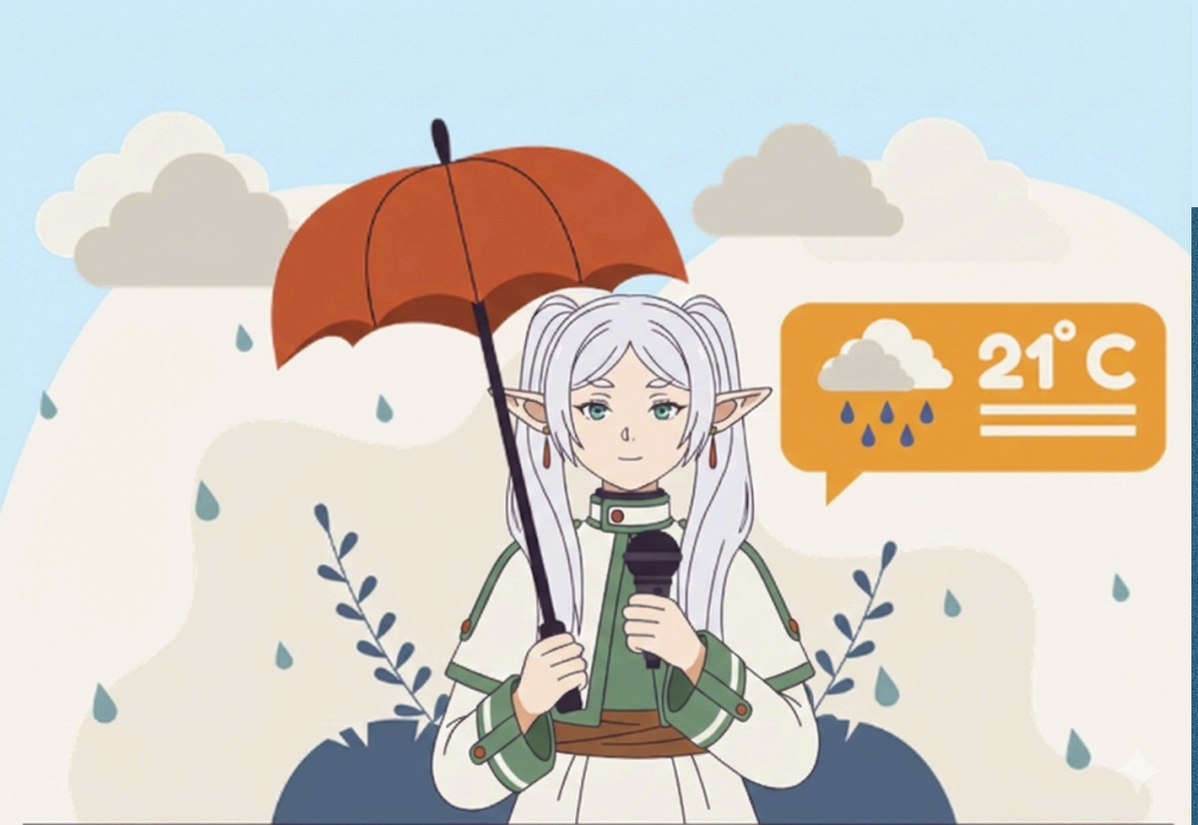


LINESCHART PREDICTION USING XGBOOST!



CONCLUSION OF PREDICTION ON 2026

- 🔥 Maximum Temp (The Heat is Real): The North region is predicted to practically "boil" in May 2026 with a peak of 34.16°C.
- 🌡️ Mean Temp (The Baseline): The seasonal patterns remain consistent. The North region's average remains high in April (28.70°C) and May (28.68°C). This consistency is a sign that the model is actually "learning" the seasonal rhythm, not just guessing wildly.
- ❄️ Minimum Temp (The Cool Down): Interestingly, in July, the Central region is predicted to see a slight rise in minimum temperature (25.30°C) , unlike the South which stays cooler at 24.45°C.



TERIMA KASIH

ありがとう