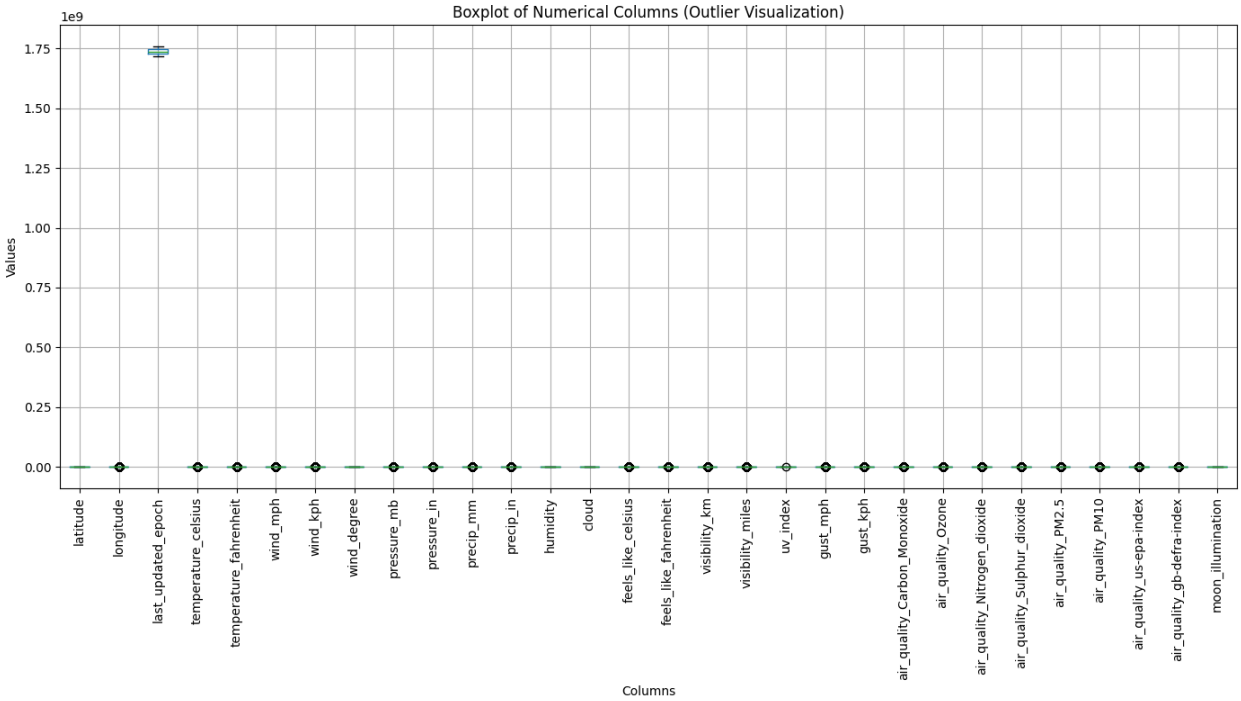
Weather Trend Forecasting

This report summarizes the exploratory data analysis (EDA) performed on the weather dataset, highlighting variable distributions, correlations between key features, and initial data quality assessment. This foundational analysis is crucial for building and evaluating any subsequent weather forecasting model.

1. Data Quality and Distribution Analysis

Outlier Visualization (Boxplot)

The boxplot reveals that, for most numerical weather variables (temperature, wind, pressure, humidity, air quality metrics), the data is concentrated near zero (or their typical ranges), but there are significant **outliers** present in several columns.

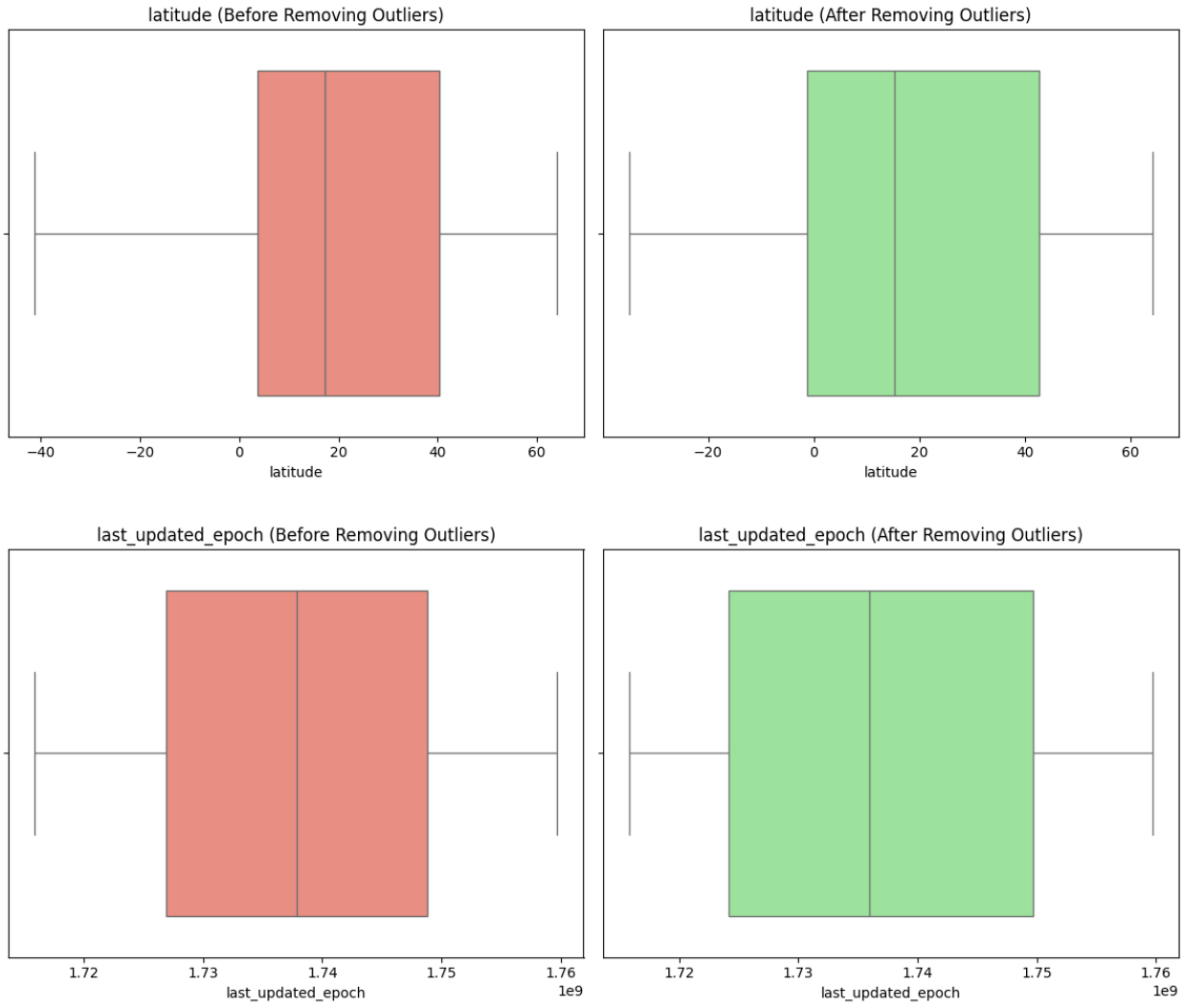


- **Extreme Outliers:** Variables like latitude, longitude, and last_updated_epoch have extreme values, suggesting potential data collection errors or inappropriate scaling in this visualization. The values for all other features are tightly clustered near the

zero axis of the y-scale, indicating that they have far smaller absolute magnitudes than these time/location features.

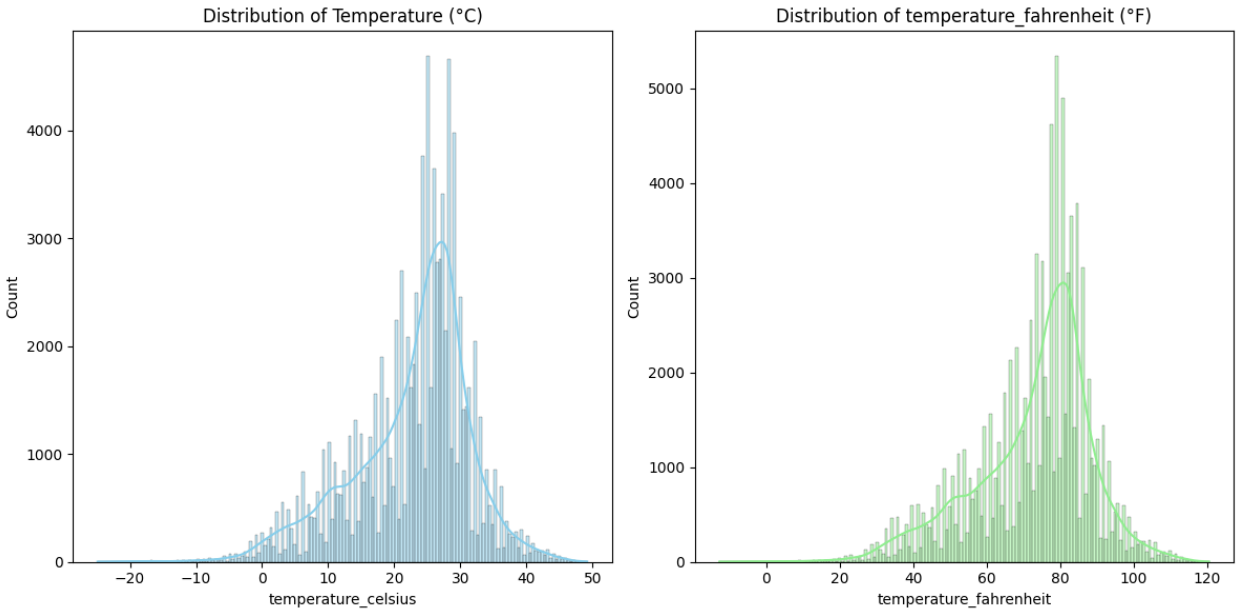
- **Actionable Insight:** A robust data cleaning and preprocessing step is mandatory to handle these outliers before training any predictive model.

  Some Example:



Distribution of Temperature

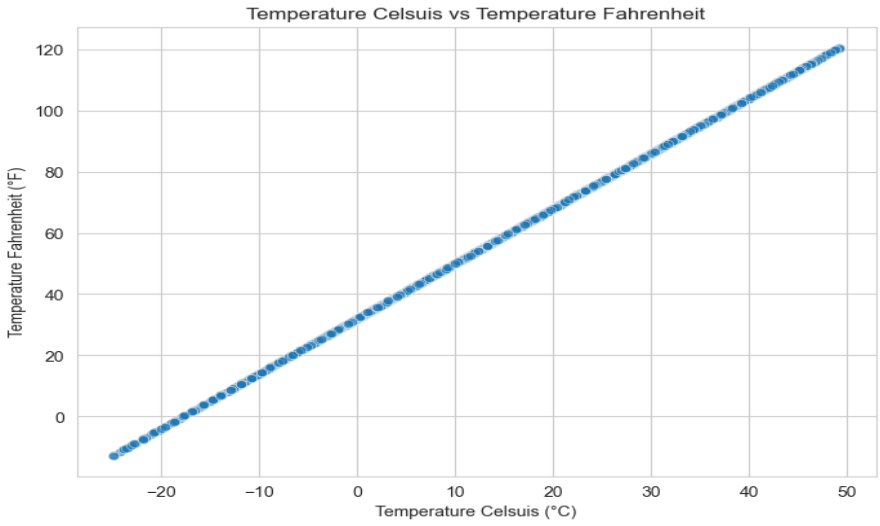
The histograms for temperature in Celsius and Fahrenheit show an expected **normal-like distribution** (bell-curve shape).



- The data is **centered around the mid-20s** (~25∘C or ~77∘F), indicating that the collected data points likely represent a region with a warm-to-hot climate.
- This also confirms the **perfect linear relationship** between the two units of temperature, which is further visualized in a scatter plot (see below).

2. Key Relationship Visualizations

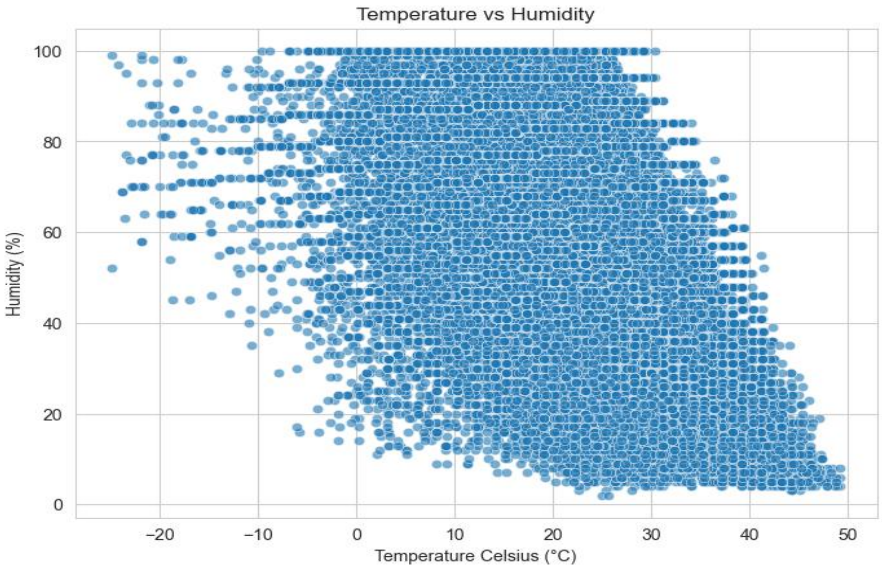
Temperature Correlation (C vs. F)



The scatter plot of Temperature Celsius vs. Temperature Fahrenheit displays a **perfectly linear positive correlation**.

- **Insight:** This confirms that temperature_celsius and temperature_fahrenheit are redundant features (collinear), and only one should be used for model building to prevent multicollinearity.

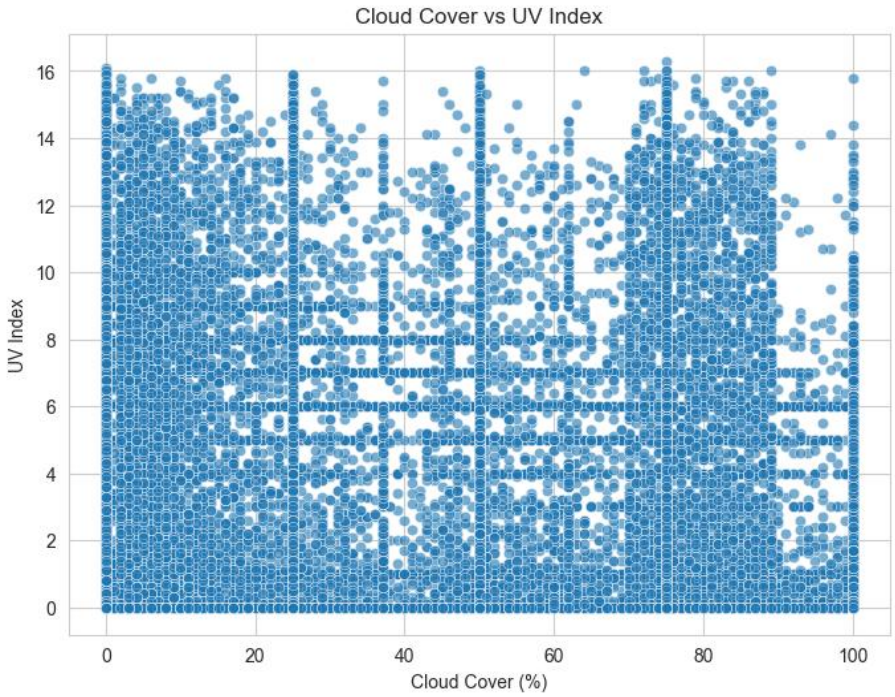
Temperature vs. Humidity



This scatter plot shows a clear **negative relationship** (inverse correlation) between temperature and humidity.

- **Observation:** As temperature increases (moving right on the x-axis), the maximum and average humidity tends to decrease, forming an inverse cone shape.
- **Interpretation:** The **highest humidity** (100%) occurs at or below **30∘C**, suggesting a clear upper bound on the amount of moisture the air can hold at high temperatures in this dataset, which is consistent with atmospheric physics (higher temperatures allow for lower relative humidity before reaching saturation).

Cloud Cover vs. UV Index



This plot shows a scattered but notable relationship:

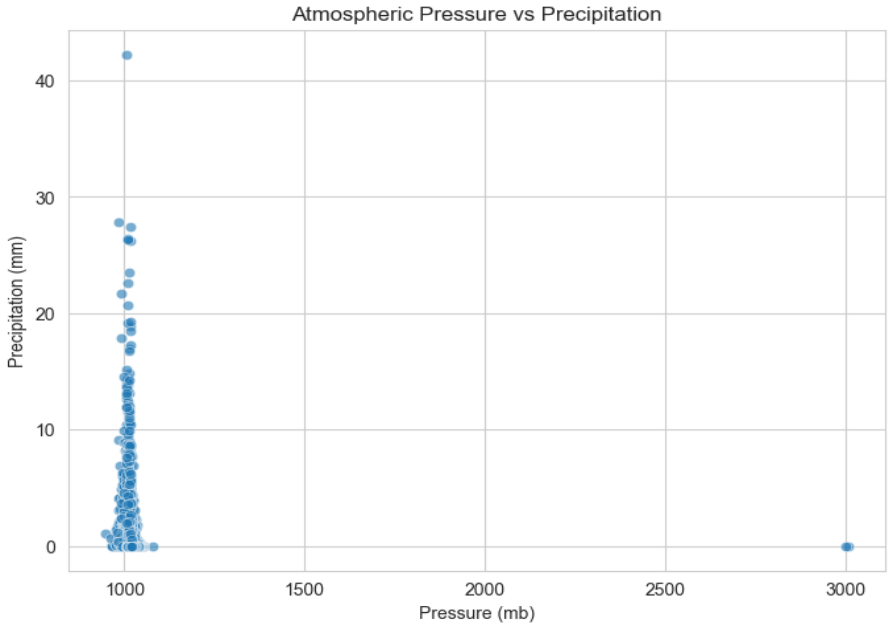
- **Observation:** High UV Index values (above 10) are possible across **all cloud cover levels**, including 0% (clear sky) and 100% (overcast).
- **Interpretation:** While cloud cover is a factor, it is **not the sole determinant** of the UV Index. High UV values at high cloud cover (or vice-versa) may be due to thin, high-altitude clouds that do not block UV rays, or other factors like sun angle and ozone concentration.

Temperature vs. Precipitation

The relationship between temperature and precipitation is highly clustered:

- **Observation:** The vast majority of precipitation events (non-zero precip_mm) occur at **temperatures between 10∘C and 35∘C**.
- **Interpretation:** This is common, as extremely high or low temperatures often suppress the conditions required for rain (convective lifting, water vapor content, etc.). The maximum precipitation value (over 40 mm) occurs at a temperate range (~25∘C).

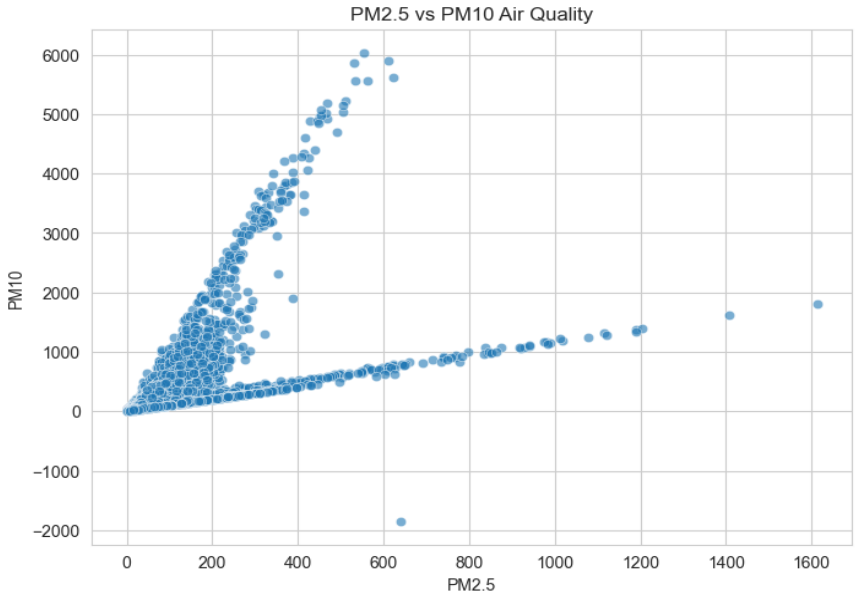
Atmospheric Pressure vs. Precipitation



This scatter plot demonstrates that precipitation is overwhelmingly associated with a **narrow range of atmospheric pressure**.

- **Observation:** Significant precipitation primarily occurs when the pressure is **around 1000 mb** (low-pressure systems).
- **Outlier/Anomalies:** The isolated data point at Pressure≈3000 mb with Precipitation≈0 mm is an **extreme anomaly** and should be investigated/removed as part of the data cleaning process.

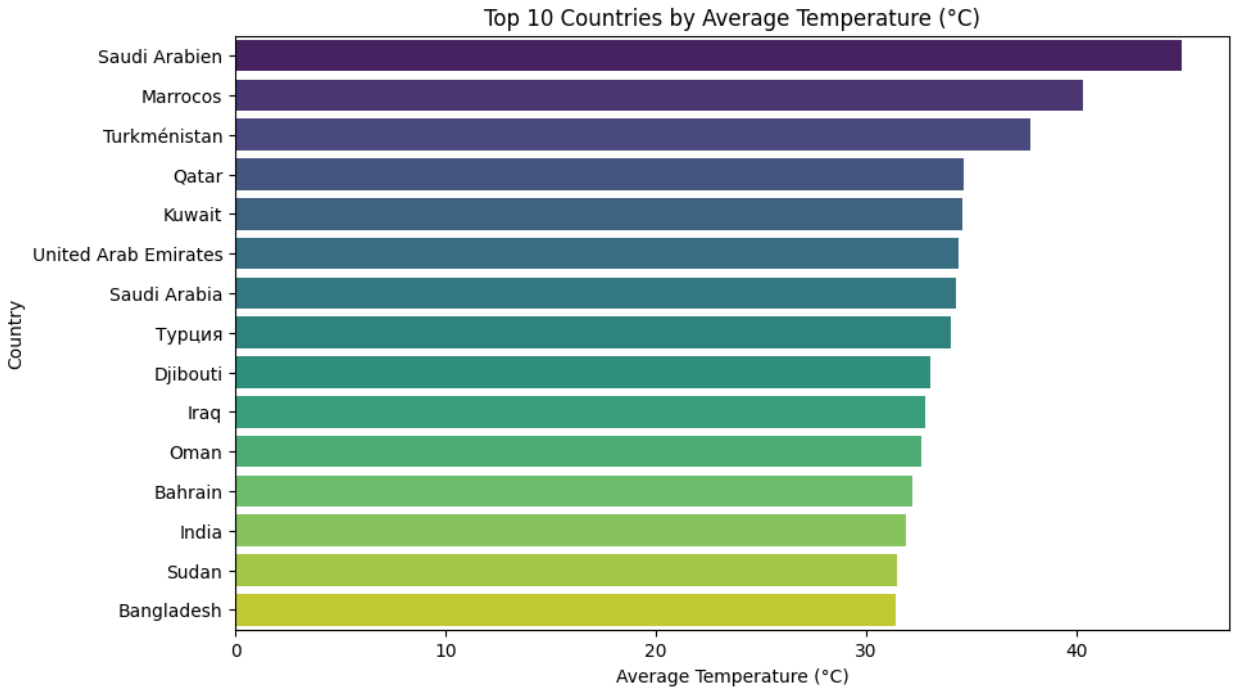
PM2.5 vs. PM10 Air Quality



There is a complex, non-linear relationship between the concentrations of particulate matter (PM).

- **Observation:** In general, higher PM2.5 (fine particles) correlates with higher PM10 (coarse particles). However, there is a distinct split where **very high PM10 levels** exist for relatively **moderate PM2.5 levels** (forming a distinct upward curve on the left of the graph), suggesting a significant source of coarser particles.
- **Interpretation:** The distribution shows that PM10 levels can be highly variable and are not simply proportional to PM2.5 levels, indicating different sources or atmospheric transport mechanisms for fine versus coarse particles.

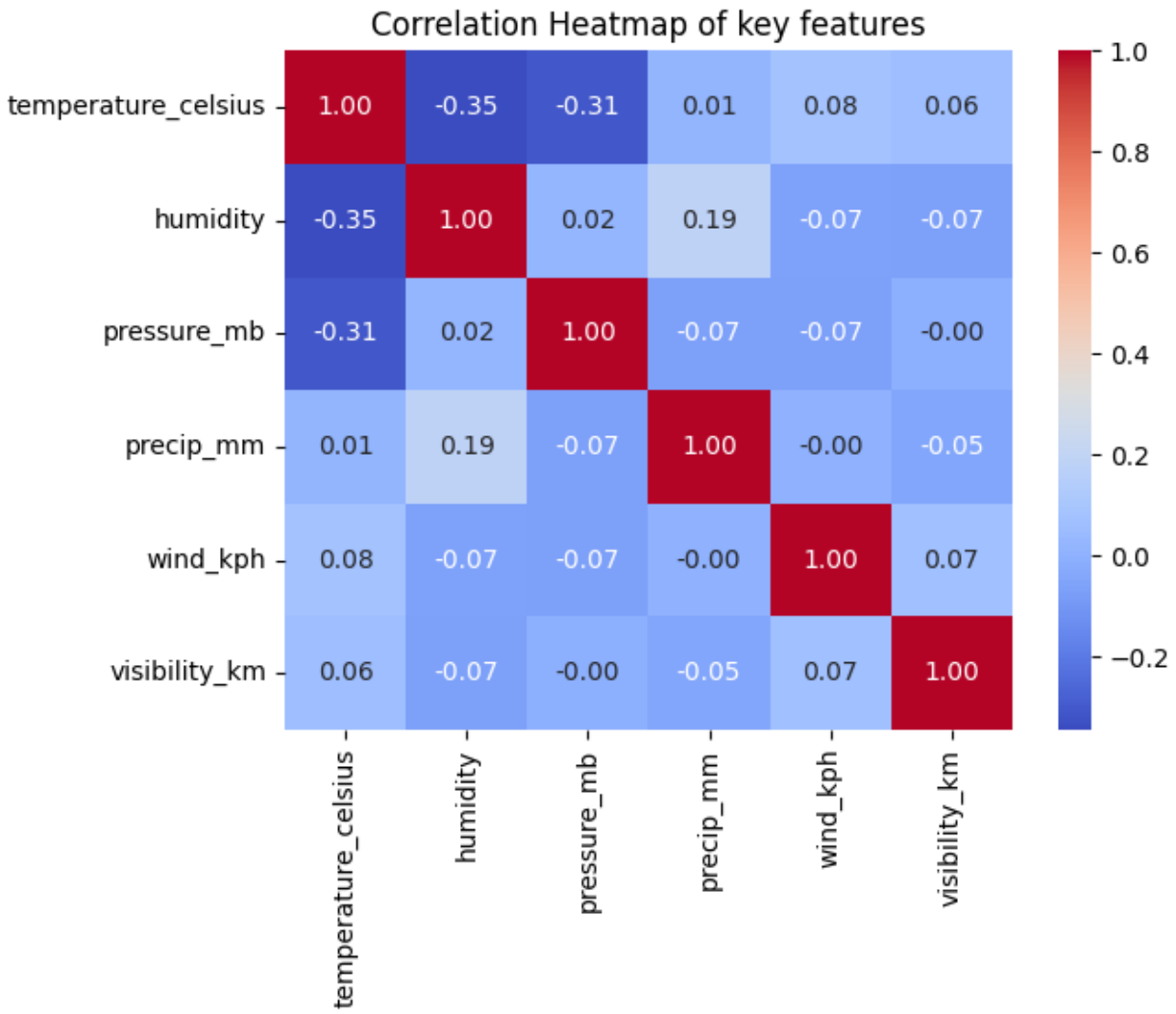
Top 15 Countries by Average Temperature



This bar chart provides an important **climatological context** for the dataset.

- **Observation:** The countries with the highest average temperatures include **Saudi Arabia, Morocco, Turkmenistan, Qatar, and Kuwait**.
- **Interpretation:** This strongly reinforces the earlier finding from the temperature distribution that the source data is heavily skewed towards regions with a **hot climate**. This must be accounted for when building a generalizable forecasting model.

. 3. Correlation Heatmap Analysis



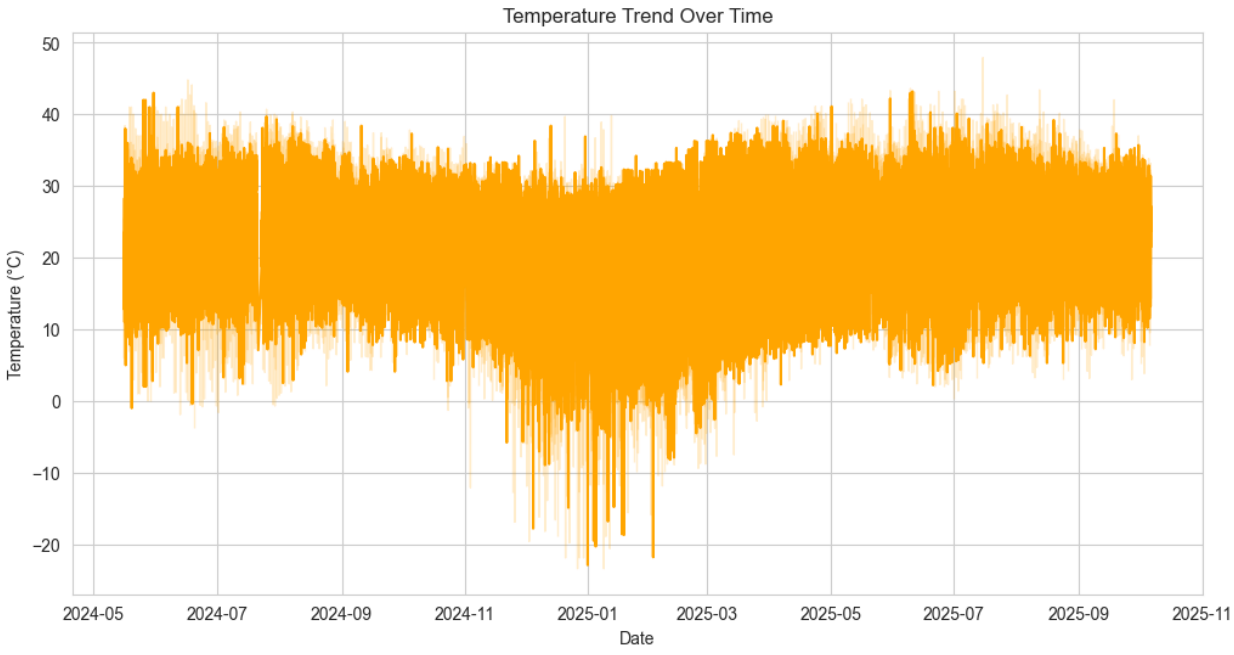
The correlation heatmap quantifies the linear relationships between the key weather features selected for analysis.

Relationship	Correlation Value (r)	Interpretation
Temperature vs. Humidity	**-0.35**	**Moderate Negative:** As temperature rises, humidity tends to decrease (consistent with scatter plot).
Temperature vs. Pressure	**-0.31**	**Weak Negative:** Slightly lower temperatures are associated with slightly higher pressures (or vice-versa), though the effect is weak.

Relationship	Correlation Value (r)	Interpretation
Humidity vs. Precipitation	**0.19**	**Weak Positive:** Higher humidity shows a weak association with higher precipitation, as expected, but the relationship is not strong.
Humidity vs. Wind	**-0.07**	**Negligible:** Humidity is largely independent of wind speed.
All other pairs	~0.00 to ±0.08	**Very Weak to Negligible:** The remaining pairs have very little linear correlation, suggesting they act largely independently of each other.

Exploratory Data Analysis (EDA) and Data Trends

A. Temperature Trend Over Time



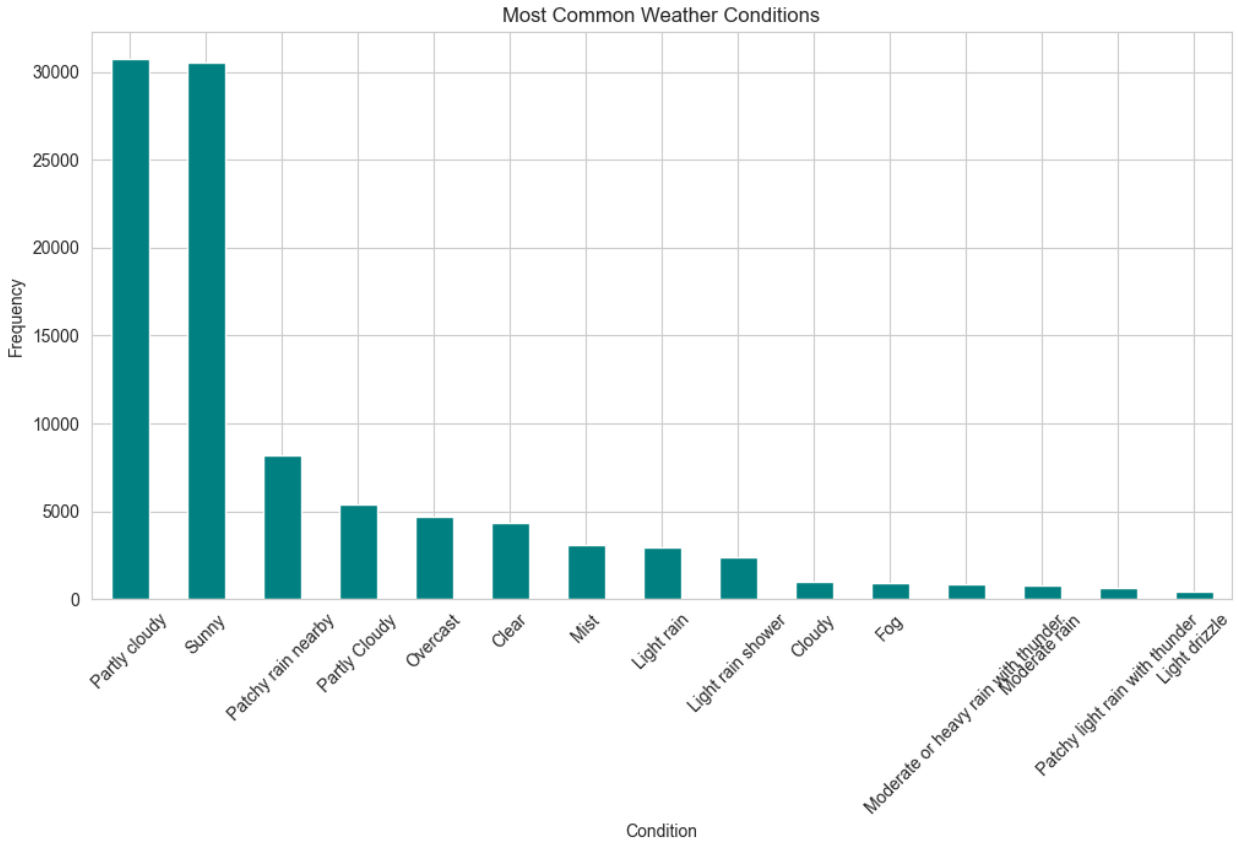
B.

The time series plot reveals the temperature pattern from mid-2024 to late-2025.

- **Strong Seasonality:** The data clearly exhibits a strong annual seasonal trend. Temperatures reach their **highest peak** (approaching 50∘C) during the summer months (around June-August of both 2024 and 2025).

- **Cold Period:** The **lowest temperatures** (dropping to −20∘C and below) occur during the winter months (November 2024 to March 2025).
- **Data Consistency:** The overall pattern of peaks and troughs appears consistent across the observed periods, confirming the data captures regular weather cycles.

C. Geographic and Climatic Context



- **Latitude vs. Temperature:** The scatter plot shows a wide range of latitudes are included in the dataset (from ∼−40 to ∼65). As expected, the **highest average temperatures** are clustered around the **Equator (Latitude 0∘)** and mid-latitudes (up to ∼30∘), while the **lowest temperatures** are primarily found in the high-latitude regions (above 40∘).
- **Most Common Conditions:** The most frequently recorded weather conditions are **Partly cloudy** and **Sunny**, followed by **Patchy rain nearby** and **Partly Cloudy** (categorized differently), indicating a climate that is frequently clear or with partial cloud cover.

Rank	Condition	Frequency
1	**Partly cloudy**	~30,500
2	**Sunny**	~30,000
3	**Patchy rain nearby**	~8,000

D. Feature Relationships

- **Feels Like vs. Actual Temperature (Heat Index):** This plot illustrates the "Feels Like" temperature (or Heat Index/Wind Chill) as a function of Actual Temperature, colored by Humidity.
    - **Cold Weather (Wind Chill):** At temperatures below ~20◦C, lower humidity (blue dots) generally makes the air feel **warmer** than the actual temperature, while higher humidity (red dots) makes it feel **colder** due to wind chill in this regime.
    - **Hot Weather (Heat Index):** At temperatures above ~20◦C, the trend dramatically changes. Higher humidity (red dots) causes the air to feel **much hotter** (Heat Index) than the actual temperature.
- **Wind Speed vs. Visibility:** The plot shows that high visibility is possible at low wind speeds. Critically, there is an **extreme outlier** at a wind speed of ~3000 km/h with a visibility of 10 km. This value is physically impossible for surface wind speed and must be **removed** from the dataset.

Feature Selection and Model Evaluation

A. Feature Importance

The **XGBoost** model's feature importance score highlights the most critical factors for predicting temperature:

1. **Feels Like Celsius (feels_like_celsius):** This is overwhelmingly the **most important feature** (Importance Score ~0.9), suggesting that the relationship between ambient temperature and thermal comfort indices is nearly constant across the dataset, making the 'feels like' metric a strong predictor of the actual temperature.

2. **Humidity:** The second most important feature, though its score is drastically lower (around 0.08).

The other features, including wind, pressure, and time components (hour, month), had negligible impact on the prediction.

B. Model Performance Comparison (R2 Score)

A comparison across several regression models was performed using the R2 score (Coefficient of Determination), where a score of 1.0 indicates a perfect fit.

Model Category	Top Performing Models	R2 **Score**
Ensemble/Boosting	**XGBoost, Gradient Boosting, AdaBoost, Random Forest**	1.00 (Near Perfect)
Simple Regression	Linear Regression, Decision Tree	1.00 (Near Perfect)
Weak Models	SVR, KNN	~0.3−0.4 (Poor)