PM Accelerator Tech Assessment Report

Author: Abhishek Maher

PM Accelerator’s mission:

By making industry-leading tools and education available to individuals from all backgrounds, **we level the playing field for future PM leaders.** This is the PM Accelerator motto, as we grant aspiring and experienced PMs what they need most – Access. We introduce you to industry leaders, **surround you with the right PM ecosystem**, and discover the new world of AI product management skills.

Introduction

In this project, I have conducted a comprehensive weather and environmental data analysis pipeline—from data cleaning to advanced forecasting. The goal was to extract meaningful insights from historical weather and pollution data, understand environmental trends, and build accurate models for forecasting.

The project begins with data cleaning, pre-processing, and transformation, including handling missing values, outliers, and normalizing the data for modelling. I then performed both basic and advanced exploratory data analysis (EDA) to uncover hidden patterns, seasonal trends, and correlations.

Using rich visualizations, I analysed a wide range of weather parameters and air pollutants, exploring their individual behaviour as well as their environmental impact. I also performed geographical analysis to understand how weather conditions and pollution levels vary across different regions and countries.

To ensure robustness, I applied anomaly detection techniques to identify and interpret unusual data points. For the forecasting component, I built and evaluated multiple time series forecasting models, utilizing the lastupdated feature as the temporal reference. These models were assessed using various evaluation metrics to compare performance and improve prediction accuracy.

Data Cleaning & Pre-processing:

To ensure the dataset was well-prepared for analysis and modelling, I performed the following data cleaning and pre-processing steps:

1**. Redundant Feature Removal**

The dataset contained multiple weather-related variables in different units (e.g., metric vs. imperial):

- temperature_celsius vs. temperature_fahrenheit
- wind_kph vs. wind_mph
- gust_kph vs. gust_mph
- precip_mm vs. precip_in
- visibility_km vs. visibility_miles
- pressure_mb vs. pressure_in
- feels_like_celsius vs. feels_like_fahrenheit

Highly correlated pairs were identified and one column from each pair was retained based on standardization.

2. Handling Missing Values & Duplicates

No missing values or duplicate records were found, eliminating the need for imputation or deduplication.

3. Standardizing Country Names

Country names were standardized to English to ensure consistency.

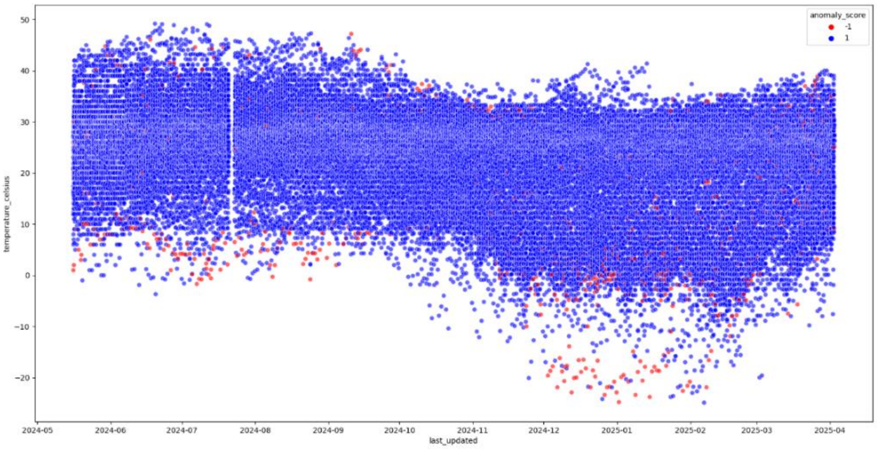
4. Timestamp Conversion

The last_updated column (originally a string) was converted into a datetime object for temporal analysis.

5. Anomaly Detection & Outlier Removal

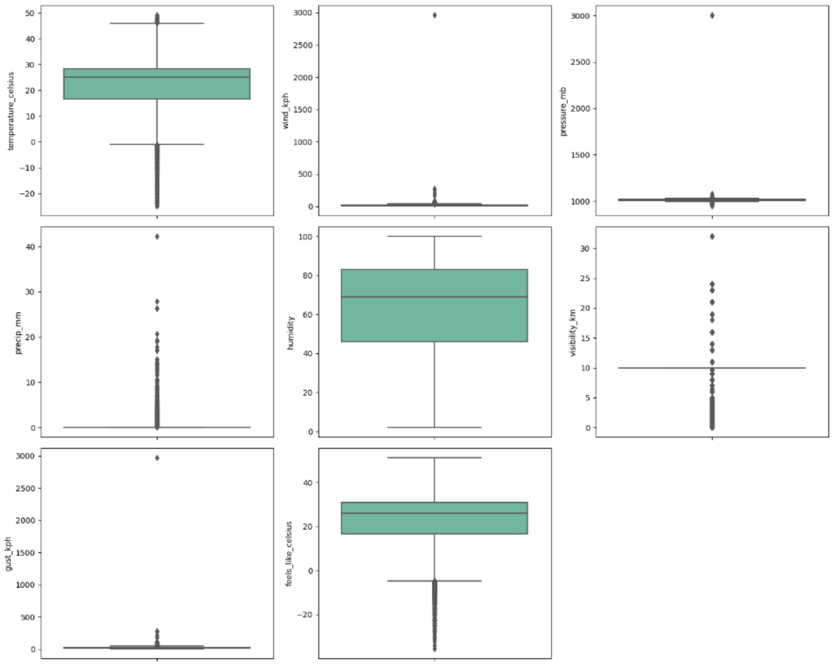
To improve data quality and reduce the impact of extreme values on statistical analysis and models:

1. Applied Isolation Forest, an unsupervised learning algorithm for detecting anomalies in multivariate data.



Using this algorithm, I found 1248 outliers in the data.

2. Used IQR-based filtering (Interquartile Range) for univariate outlier detection on numerical features



Outliers from both methods were removed to improve model robustness

Exploratory Data Analysis (EDA)

Dataset Overview

1. There are 23 numerical columns in data. They are ['latitude', 'longitude', 'last_updated_epoch', 'temperature_celsius', 'wind_kph', 'wind_degree', 'pressure_mb', 'precip_mm', 'humidity', 'cloud', 'feels_like_celsius', 'visibility_km', 'uv_index', 'gust_kph', 'air_quality_Carbon_Monoxide', 'air_quality_Ozone', 'air_quality_Nitrogen_dioxide', 'air_quality_Sulphur_dioxide', 'air_quality_PM2.5', 'air_quality_PM10', 'air_quality_us-epa-index', 'air_quality_gb-defra-index', 'moon_illumination']

2. There are 11 categorical columns in data. They are ['country', 'location_name', 'timezone', 'last_updated', 'condition_text', 'wind_direction', 'sunrise', 'sunset', 'moonrise', 'moonset', 'moon_phase']

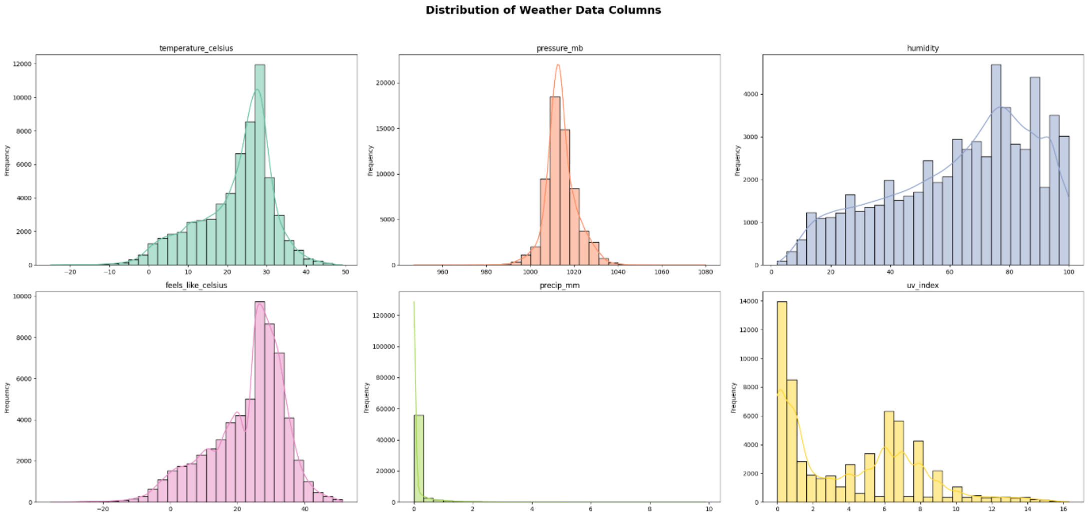
3. Geographical features:
   a. There are **383** unique latitudes and **389** unique longitudes in our data. Number of unique pairs of these are **424**.
   b. We have 191 different countries and 248 locations in our data.
   c. 292 of the latitudes are from Northern hemisphere while only 91 are from southern part.

4. There are 16 unique wind directions in the data. East (6027) and East North-East (5230) being the most frequent. This could be due to most of our location being in northern hemisphere where eastern and north eastern winds are more prominent.

5. The time range of our data is from 16 May 2024 to 2 April 2025.

Numerical features Analysis

To understand the underlying distributions and relationships within the dataset, I began by analyzing the numerical features.



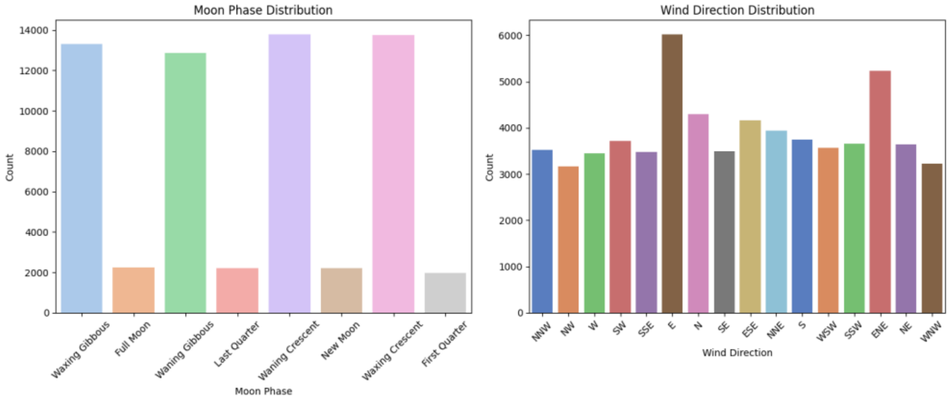
Insights:

a. **Temperature vs Feels Like:**
Both features have very similar distributions, which confirms that the feels-like temperature is closely aligned with the actual temperature across the dataset. This redundancy led to the removal of feels like feature.

b. **Pressure:**
The distribution of pressure readings is approximately normal, suggesting stable atmospheric conditions throughout the dataset.

c. **Precipitation:**
Precipitation values are heavily right-skewed, indicating that most days had little to no precipitation, while a few days experienced very high rainfall or snowfall.

Categorical Feature Analysis



Insights:

a. Waning and waxing periods are much more in the data. This is due to these periods being transitional phases that last several days. Meanwhile Full moon, new moon, first quarter, and last quarter are momentary phases — they occur at a single point in time.

b. East and East North-eastern winds are the most frequently appearing directions in the data. This is due to most of the location being in northern hemisphere where these directions are most common

Geographical patterns

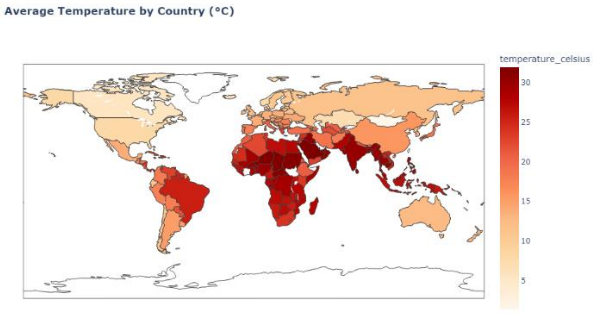
To explore how climate metrics, vary across different regions, I grouped the dataset by country and city, calculating the mean values of all relevant weather parameters. This allowed for regional comparisons and deeper geographical insight.

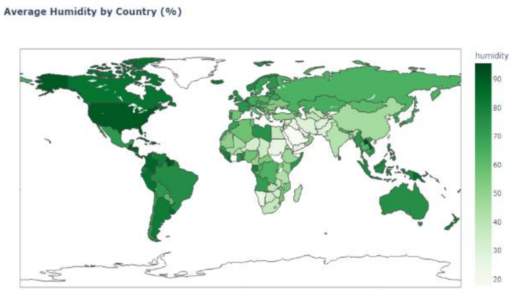
Key Country-wise Insights:

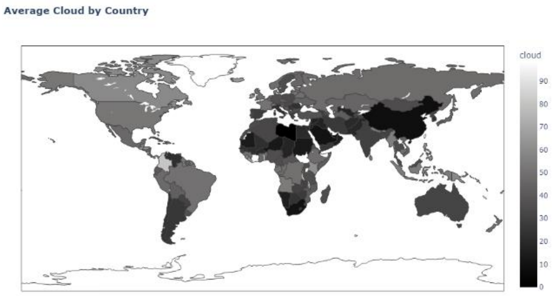
Parameter	Highest(Location)	Value	Lowest(Country)	Value
Temperature	Riyadh, Saudi Arabia	**45.0°C**	Ulaanbaatar, Mongolia	1.31 C
Precipitation	Vientiane, Laos	1.97mm	Melbourne, Australia	0.0
Humidity	Belmopan, Belize	95.52%	Riyadh, Saudi Arabia	7.0%

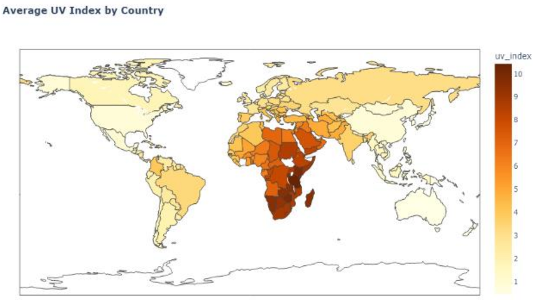
To better understand the spatial variation in climate conditions, I visualized global distributions using choropleth maps for the following features:

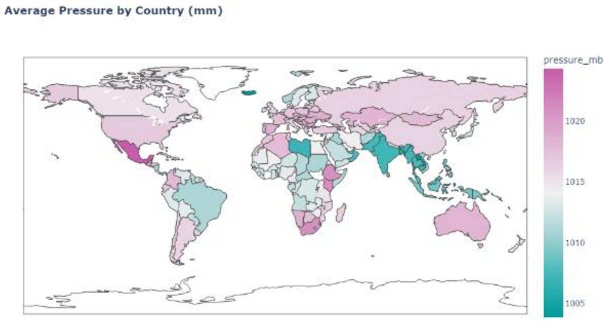
Temperature**,** Pressure, Cloud Cover, UV Index, Humidity









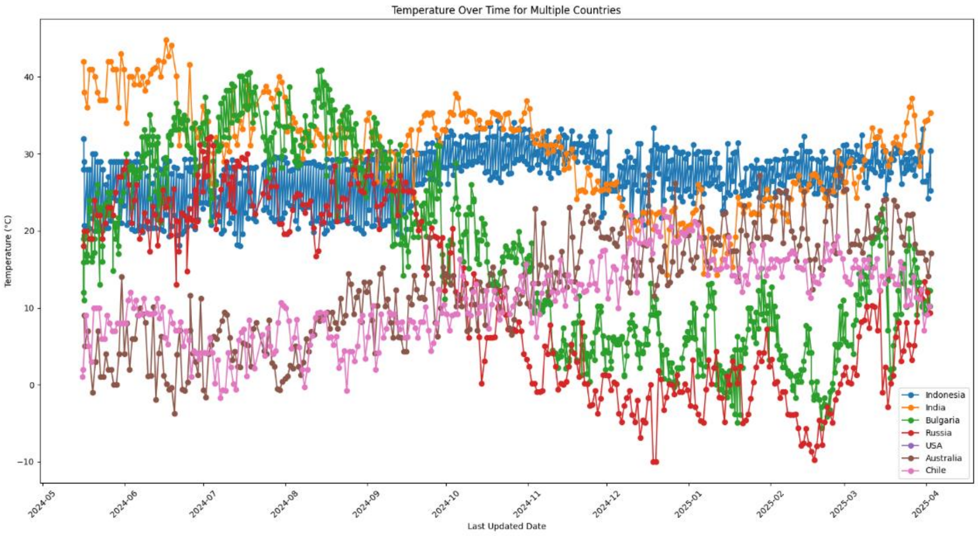


Observed Patterns:

1. **Latitude Effects on Temperature:**

    Tropical regions consistently show higher average temperatures throughout the year compared to higher latitude countries.

    To further validate this pattern, I visualized the year-long temperature variation for a selection of countries from different regions and latitudinal zones.



a. Countries like **Bulgaria** and **Russia** exhibit strong seasonal variation, with warm summers and cold winters. These countries show a distinct peak in mid-year (June–August) and **a** low during early year and end (January–February, December)**.**

b. In contrast, countries such as **Australia** and **Chile** display the opposite pattern, with cooler temperatures mid-year and warmer temperatures at the start and end of the year. This reflects the reversed seasons due to their position below the equator.

c. Countries like **India** experience relatively stable temperatures year-round. Though there are fluctuations due to seasonal monsoons, the temperature range remains narrow compared to high-latitude countries**.**

These patterns reinforce the impact of latitude and hemisphere on regional climate behaviour.

2. **Temperature vs. Pressure:**

a. There is a slight inverse correlation between temperature and pressure.

b. Hotter regions (e.g., South Asia, Sub-Saharan Africa) tend to have lower atmospheric pressure, contributing to monsoon-like systems.

3. **Humidity Distribution:**
    a) Arid zones like the Middle East, Sub-Saharan Africa, and Namibia exhibit very low humidity**.**
    b) Rainforest regions such as the Amazon, Congo Basin, and Southeast Asia show very high humidity levels due to consistent precipitation and dense vegetation.

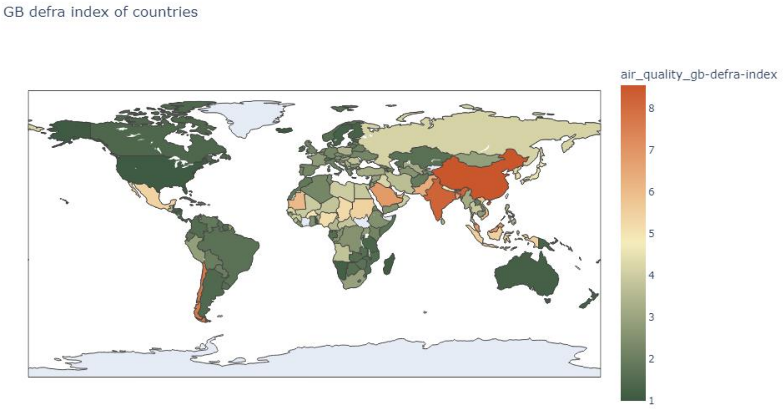
4. **UV Index & Cloud Cover:**
    a) Regions with low cloud cover and high temperature—particularly parts of Africa and the Middle East—correlate with high UV index values, highlighting higher exposure to solar radiation.

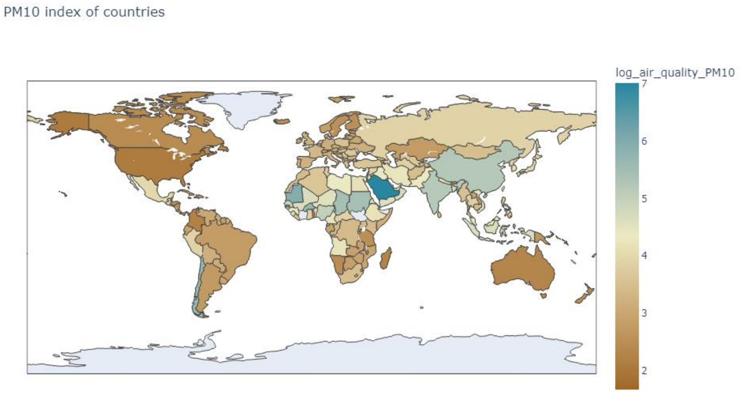
Environmental Analysis

To understand the **global distribution of air pollutants** and their **relationship with weather parameters**, I conducted a detailed analysis using country-wise aggregated data and correlation techniques.

Pollutant	Highest (Country)	Value	Lowest(Country)	Value
Carbon Monoxide	Malaysia	3482.25	USA	176.90
Ozone	Bahrain	132.67	Colombia	3.50
Nitrogen Dioxide	China	86.36	Kiribati	0.0
Sulphur Dioxide	China	117.15	Kiribati	-31.03
PM2.5	Chile	241.51	Solomon Islands	2.93
PM10	Saudi Arabia	114.87	Solomon Islands	4.29
US-EPA-INDEX	China	4.17	Colombia	1.0
GB-DEFRA-INDEX	China	8.57	Kosovo	1.0

To visualize overall pollution patterns, I plotted **choropleth maps** for composite indices such as: PM10, GB DEFRA Index

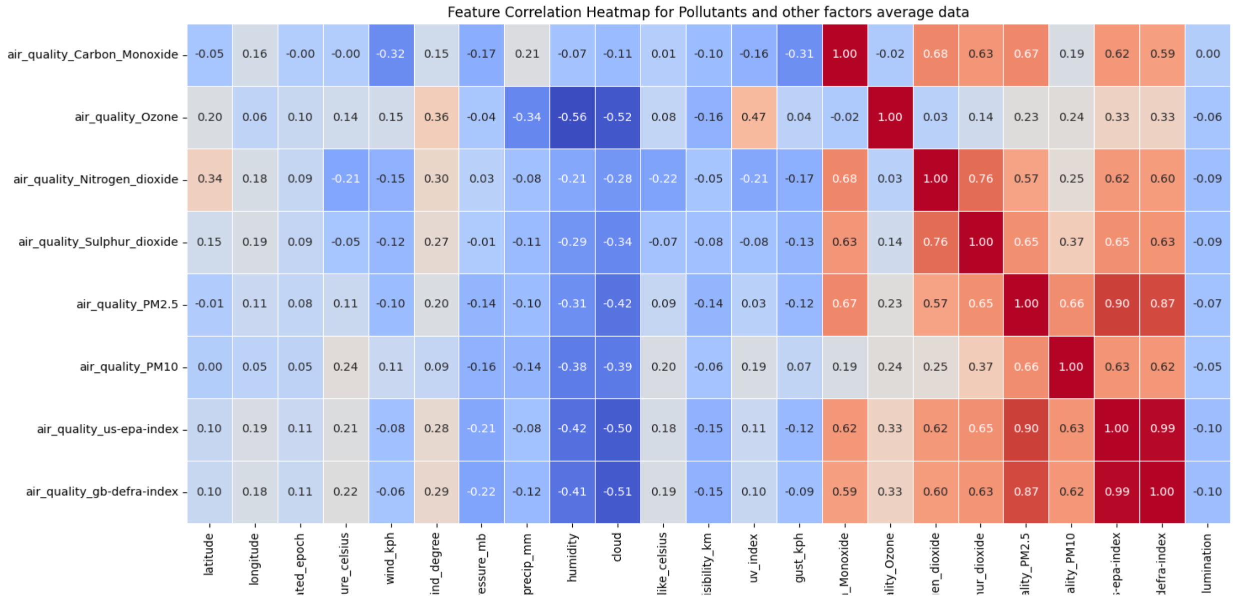




Observations:

a. Developing countries in low-latitude regions (e.g., parts of Africa, South Asia, and the Middle East) tend to exhibit higher levels of pollution across most indicators.
b. These regions are often associated with rapid urbanization, industrial activity, and fewer environmental regulations.

Then to explore how pollutants interact with atmospheric conditions, I generated a heatmap showing correlations between pollutants and weather features.



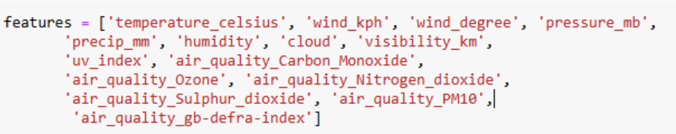
Key Insights:

- **Ozone & Humidity:**
    - Negative correlation: High humidity suppresses ozone formation as water vapour can block UV radiation—necessary for ozone production—and promote cloud formation.
- **Humidity & Cloud Cover vs. Pollutants:**
    - Most pollutants (CO, NO_2, SO_2, PM) are negatively correlated with humidity and cloud cover. Humid and cloudy conditions often aid in dispersing or diluting air pollutants.
- **Primary Pollutants Interrelation:**
    - CO, NO_2, and SO_2 show positive correlations with each other as they are all primary pollutants released directly from fossil fuel combustion and industrial processes.
- **Pollution Indices:**
    - US EPA Index, GB DEFRA Index, PM2.5, and PM10 show strong positive correlation with other pollutant levels, as these indices are composite scores derived from various pollutant measurements.

Country Clusters

To identify patterns and similarities between countries based on their weather and environmental conditions, I applied K-Means clustering to the aggregated country-level dataset.

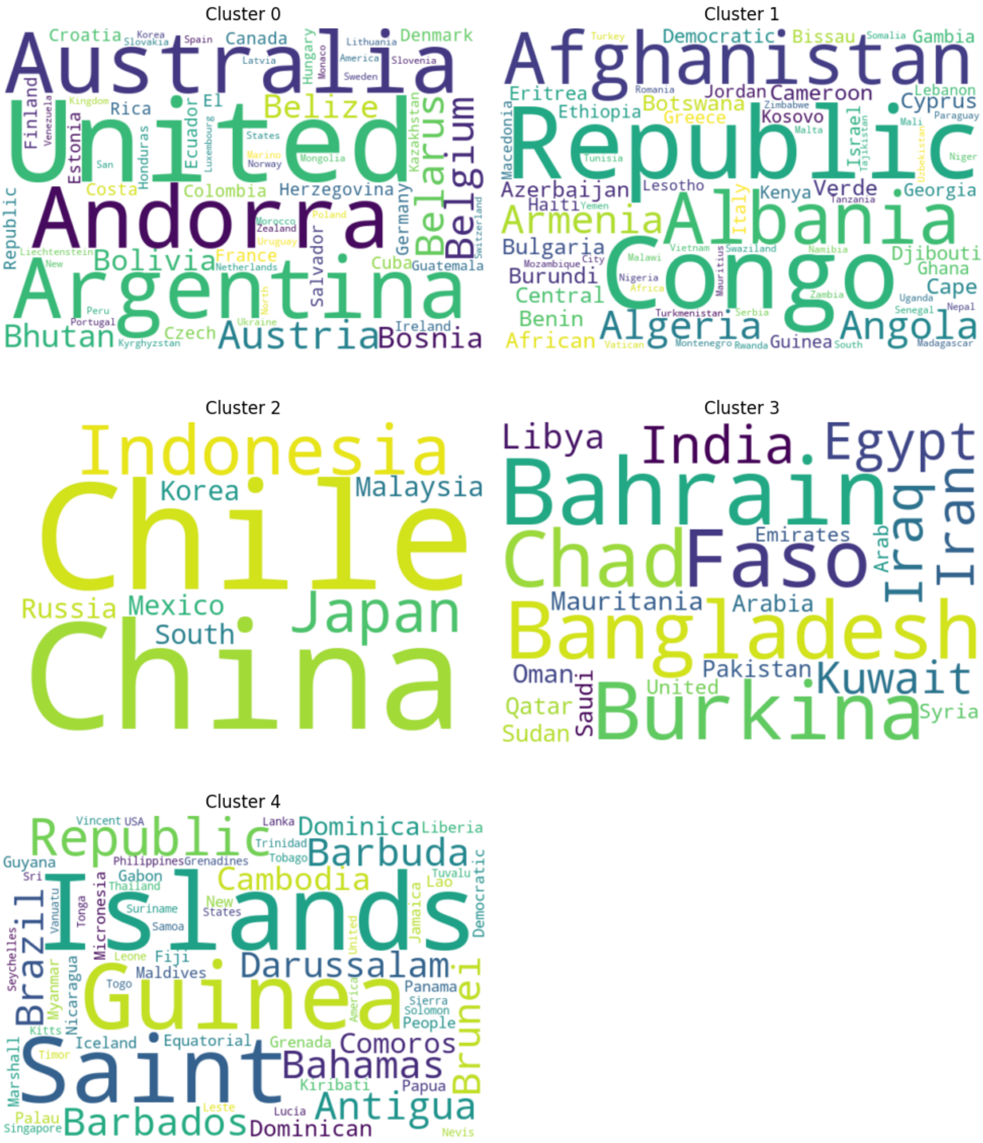
For clustering, I used a diverse set of features including:



```
features = ['temperature_celsius', 'wind_kph', 'wind_degree', 'pressure_mb',
        'precip_mm', 'humidity', 'cloud', 'visibility_km',
        'uv_index', 'air_quality_Carbon_Monoxide',
        'air_quality_Ozone', 'air_quality_Nitrogen_dioxide',
        'air_quality_Sulphur_dioxide', 'air_quality_PM10',
         'air_quality_gb-defra-index']
```

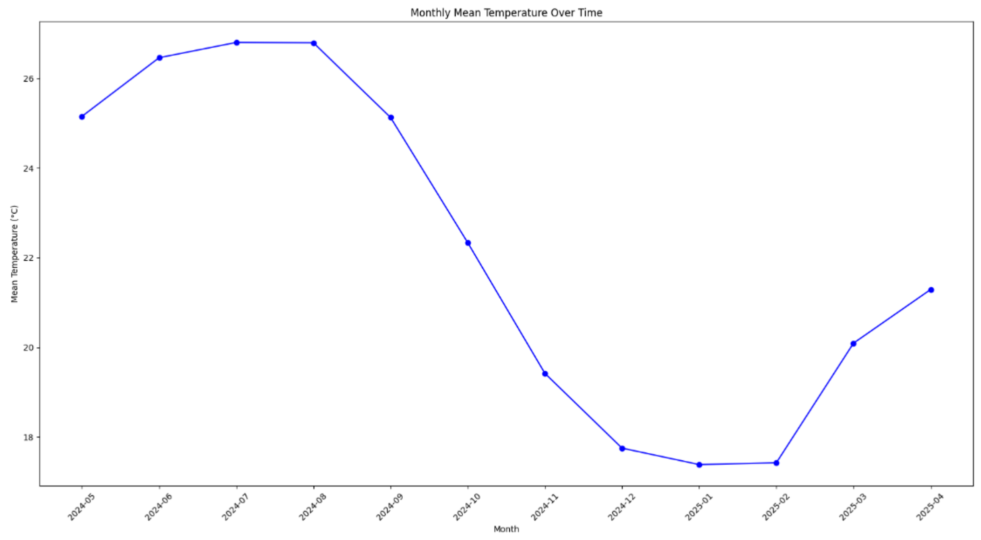
To ensure that all variables contribute equally to the clustering process and are not dominated by features with larger scales, I used StandardScaler for normalization. This pre-processing step standardized each feature to have zero mean and unit variance, which is crucial for distance-based algorithms like K-Means**.**

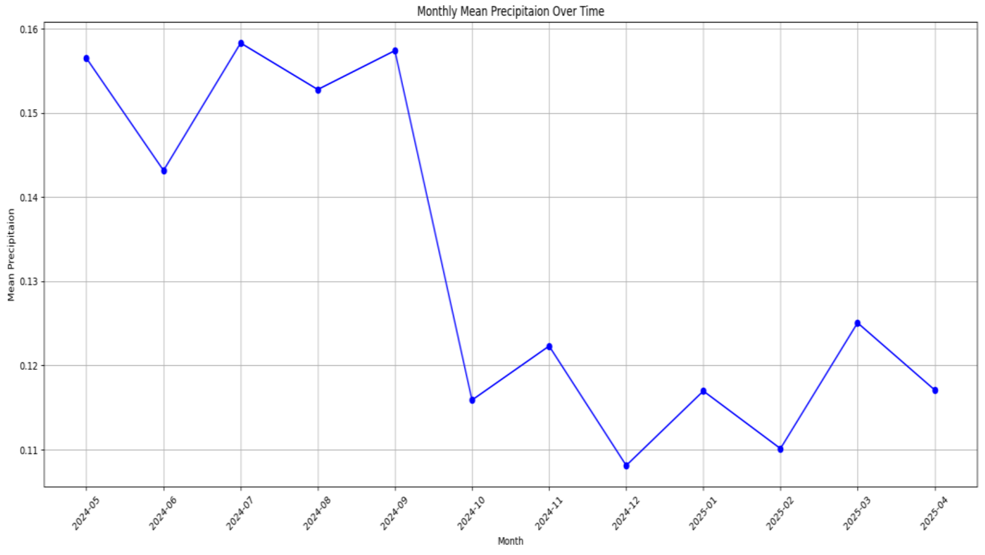
After clustering, the following clusters were created:



Precipitation & Temperature Analysis

To understand how temperature and precipitation behave across time, I aggregated the data by month and visualized the trends using line charts.





Key Insights:

- Both temperature and precipitation exhibit a clear seasonal pattern.
- They peak during the middle of the year (typically around June–August), corresponding to summer months in the northern hemisphere**.**
- The lowest values are observed at the start and end of the year (January, December), marking winter months.
- This alignment suggests that higher temperatures often coincide with increased precipitation, which is typical of monsoonal or tropical climates in many regions.

Weekly Temperature Forecasting

To understand and predict future temperature trends, I focused on weekly-level forecasting using historical temperature data.

Data Pre-processing

- The dataset was resampled to compute weekly average temperatures.
- A stationarity check was performed using the Augmented Dickey-Fuller (ADF) Test.
    - **Result:** p-value < 0.05
    - **Interpretation:** The series is stationary and suitable for ARIMA/SARIMA modelling.

Models used and parameters found after grid search:

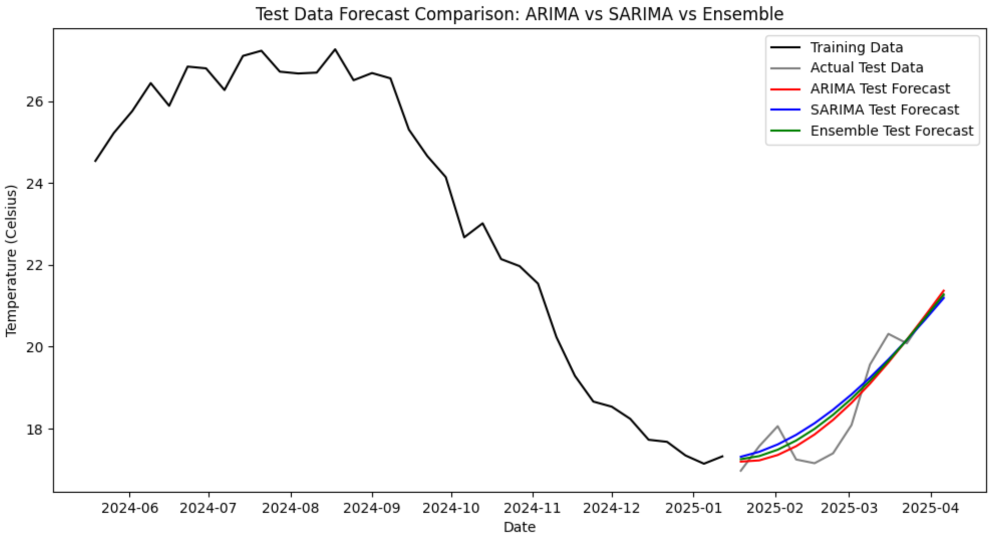
a. **ARIMA**:
A classical time series model assuming stationarity.
Optimal Parameters (p, d, q)**:** (3, 0, 2)

b. **SARIMA**:
An extension of ARIMA that supports seasonal patterns.
Optimal Parameters (p, d, q)(P, D, Q, s): (2, 0, 2)(1, 0, 1, 52)

c. **Ensemble Model**:
A simple average of the forecasts from ARIMA and SARIMA models to enhance stability.

Model Evaluation

The models were trained and tested on the resampled data. Performance was measured using several standard metrics.



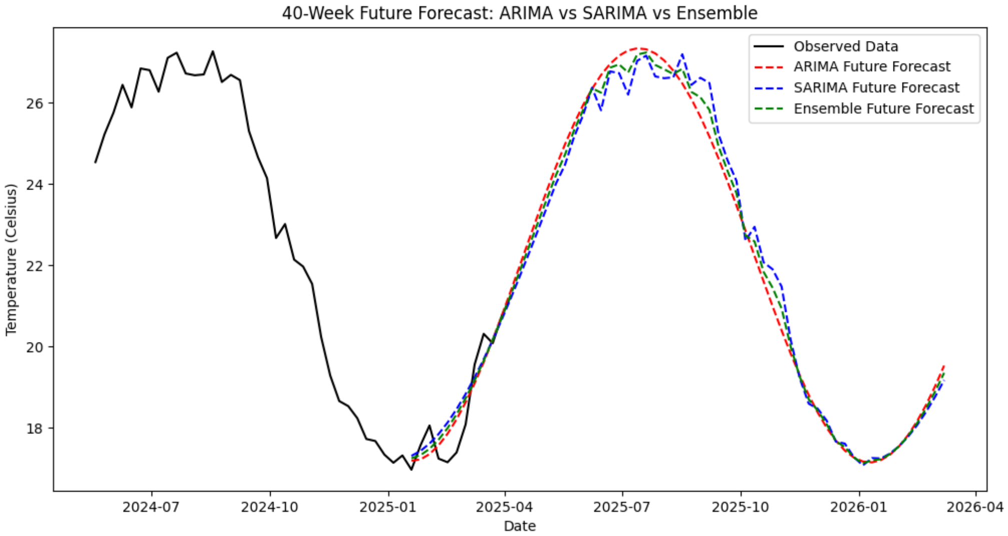
MODELS	MAE	MSE	RMSE	R2	AIC
ARIMA	0.421	0.246	0.496	0.891	60.50
SARIMA	0.456	0.321	0.567	0.858	75.80
ENSEMBLE	0.434	0.274	0.524	0.879	

Insights:

- ARIMA performed best overall based on all evaluation metrics.
- SARIMA underperformed slightly, likely due to seasonality patterns not being strong enough at the weekly level.
- The Ensemble model offered a good middle ground in terms of accuracy and stability.

Future Forecast

I proceeded to forecast the temperature for next 40 weeks using all the models and got the following results:



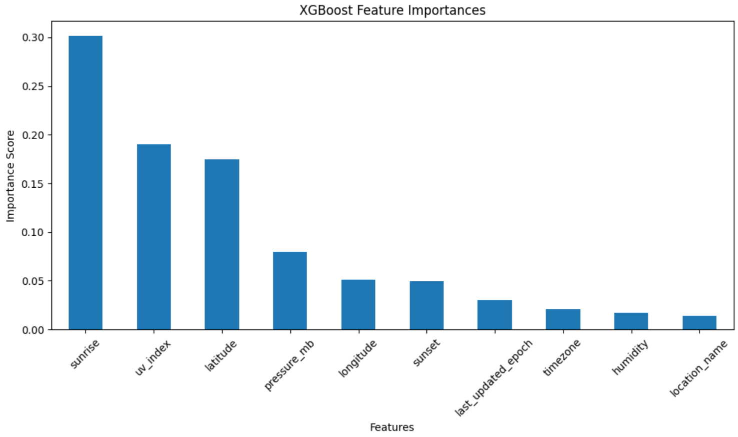
The forecast shows a clear seasonal fluctuation, with temperatures expected to rise and fall in a cyclical pattern, matching historical behaviour.

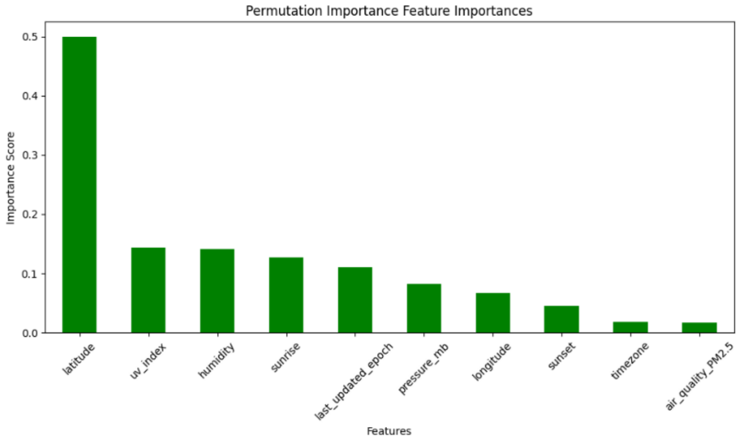
Feature Importance

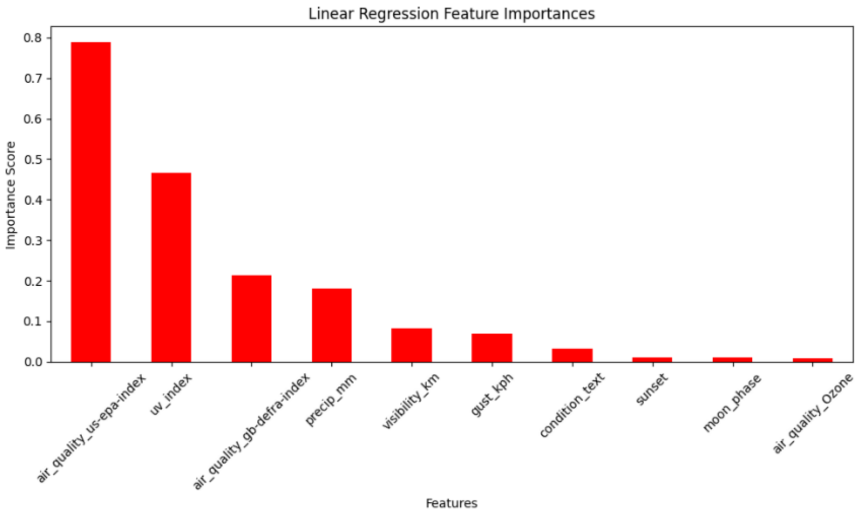
To understand which variables influence temperature the most, multiple feature importance techniques were applied:

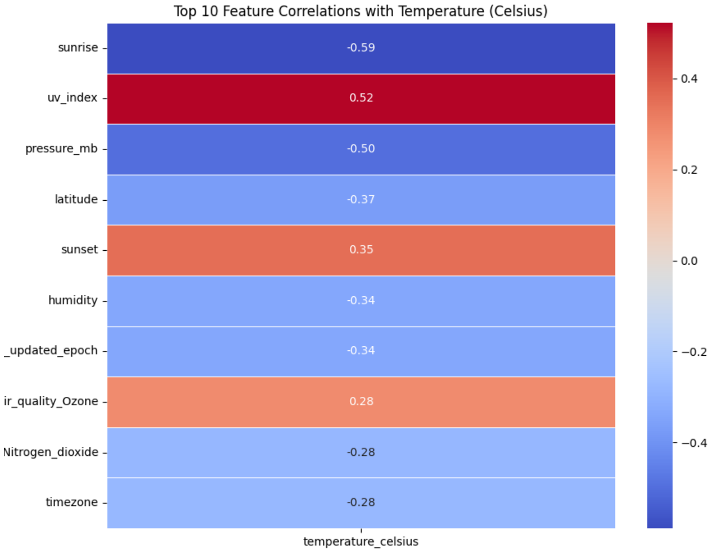
Methods Used:

- **XGBoost Feature Importance:** Captured the importance of each feature based on tree-based splits.
- **Permutation Importance (XGBoost):** Provided insight into the drop in model performance when individual features are shuffled.
- **Linear Regression Coefficients:** Helped understand the linear impact of each variable.
- **Correlation Analysis:** Showed how strongly each variable is linearly associated with temperature.









Results:

- **XGBoost**: Top features included sunrise, uv_index, latitude, and pressure_mb, indicating a strong influence of solar exposure and geographic location.
- **Permutation Importance**: Highlighted latitude, humidity, uv_index, and sunrise as the most impactful.
- **Linear Regression**: The most important features were air_quality_us-epa-index, uv_index and air_quality_gb-defra_index.
- **Correlation Analysis**: sunrise, uv_index, pressure_mb, and latitude had the highest absolute correlations with temperature.

These analyses collectively highlighted that solar-related features (sunrise, sunset, uv_index), geographic variables (latitude, longitude), and atmospheric measures (humidity, pressure_mb) are key predictors of temperature.