**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.

**Dataset Introduction**

The **Global Weather Repository** dataset provides **daily weather information** for multiple cities worldwide, containing over **40 features**. These features include:

* **Temperature** (in Celsius and Fahrenheit)
* **Humidity, Wind Speed, Cloud Cover**
* **Precipitation** (rainfall in mm)
* **Air Quality Metrics** (PM2.5, NO₂, etc.)
* **Time-related Columns** (sunrise, sunset, moon phases)
* **Location Details** (latitude, longitude, country, timezone)

**Objective:**

1. **Clean and preprocess** the data (handle missing values, outliers, normalization).
2. **Perform Exploratory Data Analysis (EDA)** to uncover key trends, correlations, and patterns.
3. **Build forecasting models** (e.g., ARIMA) using the last\_updated feature for time series analysis.
4. **Implement anomaly detection** to identify outliers in temperature or other features.
5. **Conduct unique analyses** such as climate analysis by country, environmental impact, feature importance, and spatial patterns.

With this foundational understanding of the mission and dataset, we can now move on to **data cleaning and preprocessing** steps to ensure our data is reliable and ready for deeper analysis.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis is essential to understand underlying trends, patterns, and relationships in data. In this section, we've analyzed the temperature and precipitation trends over time and examined correlations among numerical features

### A graph showing the temperature Description automatically generated**Temperature Trend Over Time**

The first visualization displays the **Temperature (°C) trend over time**. Here, we observe:

* **Seasonal fluctuations and clear downward trends** indicating a general cooling pattern as we approach the later months (October to March).
* Notable short-term fluctuations and occasional spikes suggest periodic changes or potential weather anomalies.
* The temperature generally remains within the range of 20°C to 30°C, with occasional extreme drops below 10°C or spikes close to 40°C, which might represent anomalies or sudden weather changes.

*Insight*:  
This suggests there might be seasonal variations or specific climate events impacting temperature significantly, warranting deeper investigation or modeling for accurate forecasting.

### A graph showing a number of times Description automatically generated with medium confidencePrecipitation Trend Over Time

The second visualization focuses on precipitation levels recorded over the same period.

* Precipitation remains relatively low (mostly below 1 mm) but includes sporadic peaks exceeding 4-6 mm, indicating potential instances of heavy rainfall or isolated rain events.
* The precipitation remains mostly stable with intermittent peaks, particularly noticeable in late 2024 and early 2025, possibly corresponding to seasonal rainfall or isolated weather events.

Insight:  
Precipitation events appear sparse yet intense at specific intervals, suggesting regional rainfall patterns rather than consistent global occurrences. Forecasting precipitation accurately may require additional localized or seasonal data.

### A diagram of a heat map Description automatically generated with medium confidenceCorrelation Heatmap of Numerical Features

The correlation heatmap reveals how strongly different numerical variables in the dataset are correlated:

* **High positive correlation (close to 1.0)**:
  + Temperature in Celsius is strongly correlated (1.0) with temperature in Fahrenheit, as expected.
  + Wind speeds (wind\_mph, wind\_kph, gust\_mph, gust\_kph) have extremely strong positive correlations (~0.99 to 1.00), as they measure the same phenomena in different units.
  + Pressure measurements (pressure\_mb, pressure\_in) are highly correlated (~1.00).
* A significant negative correlation is found between temperature and humidity (-0.35), indicating that as temperature rises, humidity tends to decrease and vice versa.
* There's a strong correlation between temperature\_celsius and perceived temperature (feels\_like\_celsius), signifying subjective comfort heavily depends on actual temperature measurements.

Insight:  
Strong feature correlations suggest redundancy. Using only one feature from each strongly correlated group may suffice for predictive modeling, reducing complexity without losing accuracy.

### Correlation Heatmap Analysis

The heatmap provides clarity on relationships between air quality and weather parameters:

* **PM2.5 and PM10** have strong positive correlations (0.77 and 0.72 respectively) with air quality indexes (us-epa-index, gb-defra-index), indicating that particulate matter greatly influences air quality indices.
* **Ozone** negatively correlates with humidity (-0.45), suggesting that higher humidity may associate with lower ozone concentrations.
* **Carbon Monoxide (CO)** shows strong correlation with Nitrogen Dioxide (NO₂) and PM2.5, indicating these pollutants typically co-occur, possibly due to common sources such as vehicle emissions or industrial activities.

Insight:  
Improving air quality may require strategies targeting specific pollutants, particularly PM2.5, CO, and NO₂ simultaneously, since they appear strongly interconnected.

### 3. Forecasting with ARIMA Model

* The ARIMA model forecasted daily temperature effectively.
* **Model Evaluation Metrics**:
  + Mean Squared Error (MSE): 0.487
  + Mean Absolute Error (MAE): 0.517
  + Root Mean Squared Error (RMSE): 0.698
* **Insights from ARIMA Visualization**:
  + The forecast closely tracked actual test values initially, but the model performance slightly degraded towards the end of the test period, suggesting potential seasonal or environmental shifts not captured by the current ARIMA configuration.

**Insight**: Future analyses might consider integrating seasonal ARIMA or other complex models for improved accuracy during changing climatic periods.

A graph showing the temperature of a cold

Description automatically generated with medium confidence

### 4. Advanced EDA - Anomaly Detection

* An isolation forest identified 2846 anomalies (≈5% contamination).
* Visual analysis indicated anomalies mostly occurred at the temperature extremes, which validates the effectiveness of the outlier handling step. Red points represent significant deviations likely resulting from unusual weather events or data recording errors.

**Insight**: Further investigation into anomalies could lead to better understanding and preparedness for extreme climatic conditions.

### 5. Forecasting with Multiple Models & Ensemble

* Combining forecasts from ARIMA and Exponential Smoothing provided a stable prediction (ensemble forecast).
* The ensemble model offered balanced predictions that minimized extreme deviations observed in the individual forecasts.
* **Insight**:
  + Despite providing smooth forecasts, the ensemble showed limited responsiveness to short-term volatility observed in January-February 2025. Fine-tuning or incorporating additional models (like SARIMA or Prophet) might further improve results. A graph with a line

    Description automatically generated

## Exploratory Data Analysis (EDA) Insights:

### Temperature Trend Analysis:

* The line plot of "Temperature (°C) Over Time" shows significant temperature fluctuations from mid-2024 to early-2025. There are sharp dips and spikes, indicating possible anomalies or sudden environmental changes. The overall trend depicts decreasing temperature towards early-2025, suggesting seasonal variations.

### Precipitation Analysis:

* The precipitation plot indicates mostly low precipitation with sporadic peaks. A significant spike around January 2025 indicates an unusual weather event, suggesting potential flooding or heavy rainfall during that period.

### A graph of temperature Description automatically generatedCorrelation Heatmap Analysis:

* High positive correlations between:
  + Air quality indices (PM2.5, PM10, EPA, DEFRA indices) and carbon monoxide and nitrogen dioxide, indicating air pollutants frequently co-occur.
  + Temperature and ozone levels show a positive correlation (0.30), indicating warmer temperatures might influence ozone formation.
* Negative correlation:
  + Ozone and humidity (-0.45), suggesting increased humidity may decrease ozone concentrations.
* Overall, these correlations are critical for air quality and climate monitoring.

### A screen shot of a chart Description automatically generatedFeature Importance Analysis:

* Pressure (mb) and humidity emerged as the most critical predictors for temperature forecasting, indicating weather conditions heavily depend on these parameters. Wind speed, cloud coverage, and precipitation were also relevant but with lower predictive power.

### A graph with blue bars Description automatically generatedGeographical Distribution of Temperature:

* Temperature distribution geographically indicates latitude heavily influences average temperature, with higher latitudes generally showing cooler temperatures. Longitude shows less pronounced effects, reflecting global weather patterns like ocean currents and continental influences.

### A diagram of a distribution of temperature Description automatically generatedTemperature Distribution by Country:

* Boxplots reveal substantial temperature variability within many countries. Notably, the spread and numerous outliers indicate volatile weather conditions that require tailored adaptation strategies.

A diagram of different types of temperature

Description automatically generated with medium confidence

## Conclusion:

The comprehensive analysis conducted, encompassing exploratory data analysis (EDA), forecasting using ARIMA and ensemble methods, anomaly detection, geographic, and feature analyses, provided valuable insights into climate trends, temperature variability, and their associated predictive parameters. Significant temperature fluctuations, especially sudden anomalies and extreme weather events, highlight the complexity and volatility of global climate patterns. Temperature trends clearly demonstrate geographic variability influenced strongly by latitude. Additionally, air quality metrics correlate strongly with temperature and other meteorological factors, reflecting the intertwined nature of climate and environmental health.

Ensemble forecasting models notably outperformed individual models, effectively managing uncertainty and capturing trends accurately despite short-term volatility. Anomaly detection methods successfully identified numerous unusual climate events, emphasizing the need for continuous monitoring to enhance environmental resilience.

## Recommendations:

Based on the insights drawn from this analysis, the following recommendations are proposed:

1. **Adoption of Ensemble Forecasting Techniques:**
   * Implement ensemble forecasting as standard practice for temperature and climate prediction due to its demonstrated accuracy and resilience against volatility and sudden climate shifts.
2. **Regular Climate and Environmental Anomaly Monitoring:**
   * Institutionalize anomaly detection (e.g., Isolation Forest algorithm) for ongoing monitoring, allowing early detection of unusual environmental and climatic events to facilitate rapid responses and mitigation strategies.
3. **Integrated Air Quality and Climate Management:**
   * Develop integrated management frameworks that jointly address air quality and climate issues, leveraging the observed strong correlations between air quality indices, temperature, humidity, and other meteorological conditions.
4. **Customized Geographic Climate Strategies:**
   * Design region-specific climate strategies and adaptive measures, considering significant geographic temperature variations. Areas exhibiting extreme temperature conditions should receive prioritized attention for resource allocation and climate resilience planning.
5. **Enhanced Data Quality and Management:**
   * Invest in robust data collection and handling protocols to improve the quality of predictive modeling, ensuring accurate, timely, and reliable datasets for decision-making processes.

By implementing these recommendations, organizations and policymakers can significantly enhance their climate preparedness, effectively mitigate risks associated with environmental volatility, and contribute to broader sustainability goals.