**Weather Trend Forecasting Report**

1. **Introduction**

This project focuses on analysing worldwide weather data to identify trends, perform forecasting, and derive environmental insights. Using a dataset from Kaggle titled *Global Weather Repository*, worked through both basic and advanced data science techniques to draw conclusions about long-term climate patterns and environmental impacts.

**PM Accelerator Mission:**  
*"PM Accelerator's mission is to empower aspiring product managers by providing hands-on projects and real-world datasets that simulate the challenges of a modern product role."*

**Objectives of this project:**

* Analyse global weather data using EDA.
* Forecast future temperature trends.
* Evaluate model performances.
* Analyse long-term climate and air quality patterns.
* Visualize weather across countries and continents.

**Dataset:**

* Source: [Kaggle – Global Weather Repository](https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository)
* Daily weather data of cities across the world.
* Includes over 40 features such as temperature, humidity, precipitation, UV index, wind, air quality, and time-based data.

**Feature includes**

* Temperature (Celsius, Fahrenheit)
* Wind Speed, Pressure, Precipitation
* Humidity, UV Index, Visibility
* Air Quality Indicators (PM2.5, CO, NO₂, O₃, SO₂)
* Location (Country, City, Coordinates)
* Timestamps (Last Updated, Sunrise/Sunset, Moon Phase)

1. **Data Cleaning and Preprocessing**

The original dataset required cleaning before performing any meaningful analysis:

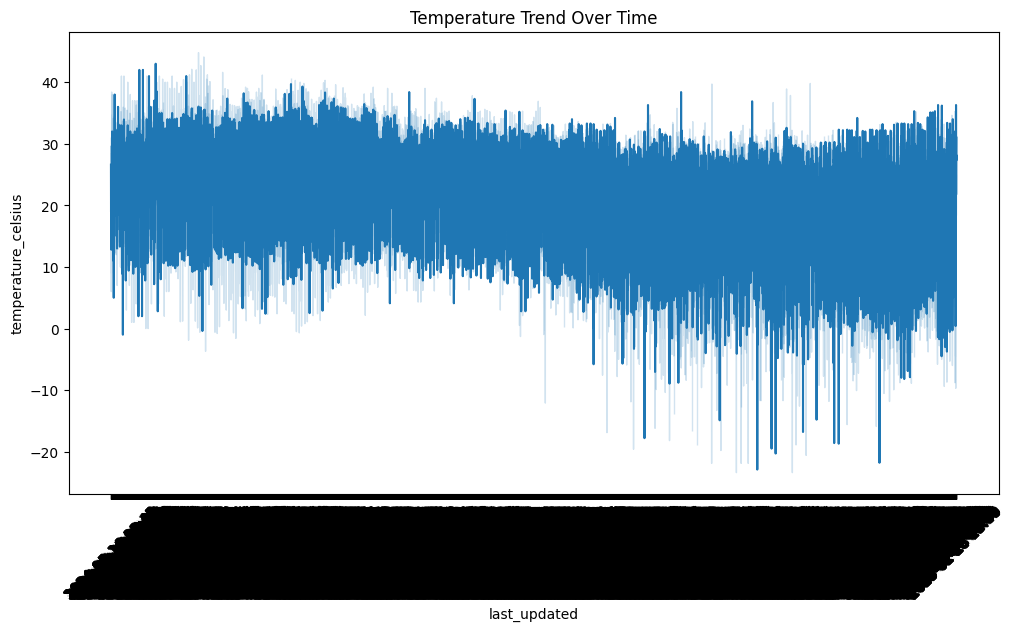
* **Handled missing values** using dropna() and imputation for some columns.
* **Converted timestamp fields**: last\_updated was converted to datetime format.
* **Extracted time features** such as year, month, day, and hour from last\_updated.
* **Reset index** to make datetime available as a standard column for visualization.
* **Selected relevant features** for forecasting and visualization like temperature, humidity, precipitation, UV, and air quality.

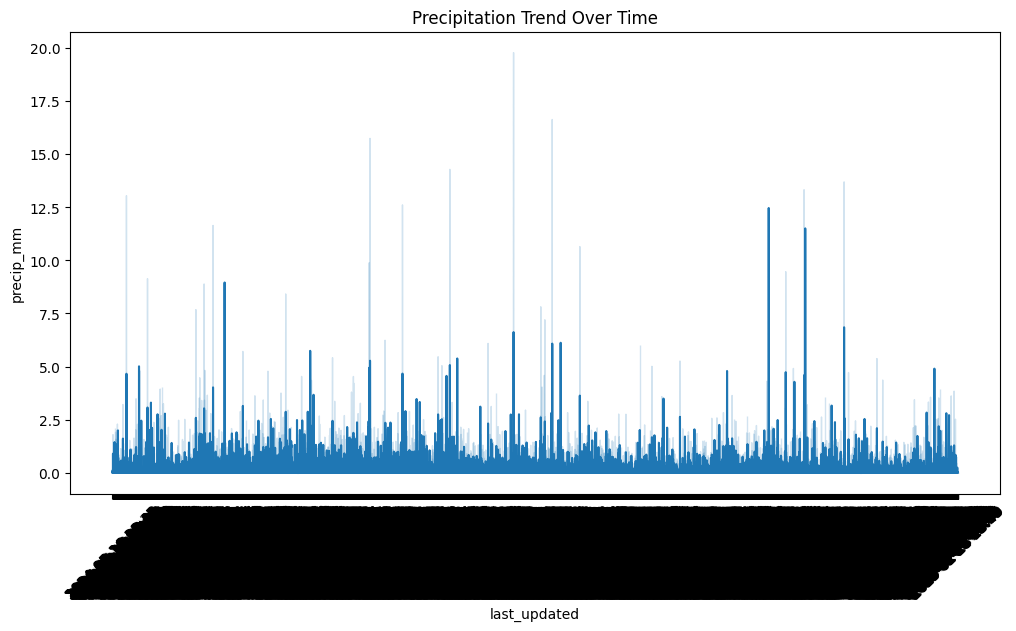
1. **Exploratory Data Analysis (EDA)**

Performed extensive EDA to identify weather patterns and trends.

**Temperature & Precipitation Trends:**

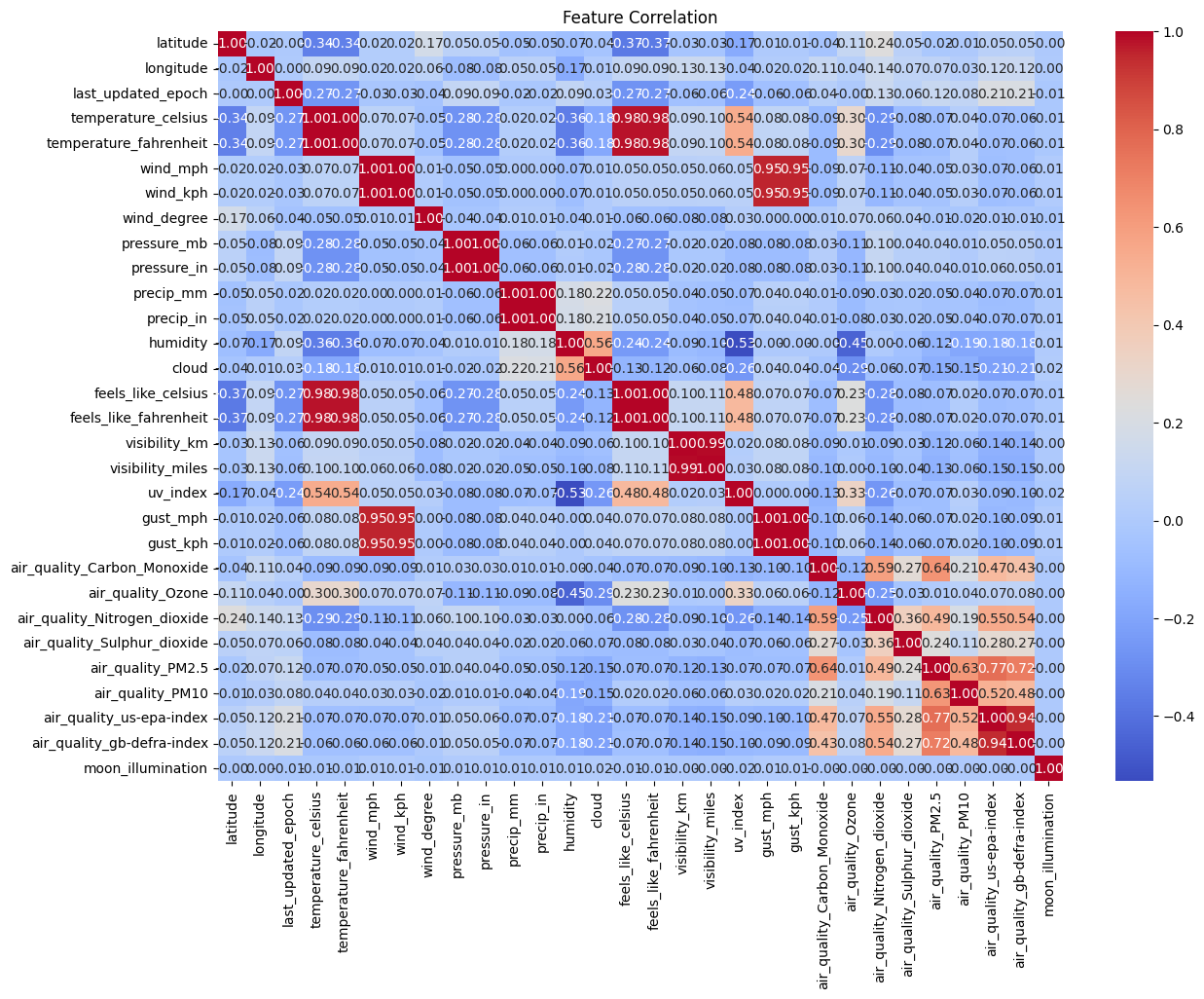
* Visualized average temperature and precipitation over time.
* Observed seasonal fluctuations in temperature for most cities.



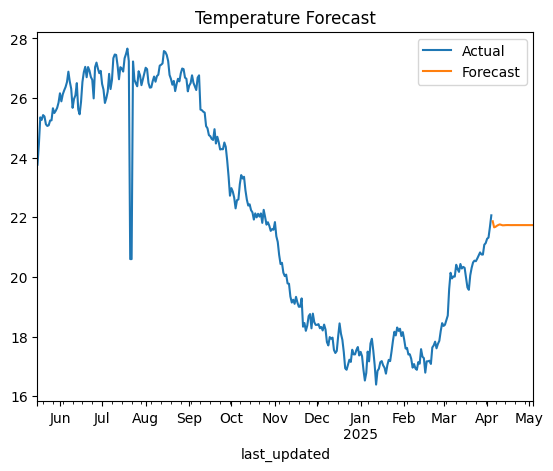


**Correlation Analysis:**

* Computed correlation between temperature, humidity, precipitation, pressure, UV, etc.
* Found strong negative correlation between humidity and temperature.



1. **Model Building: Arima**



**Mean Squared Error: 2.041159534548533**

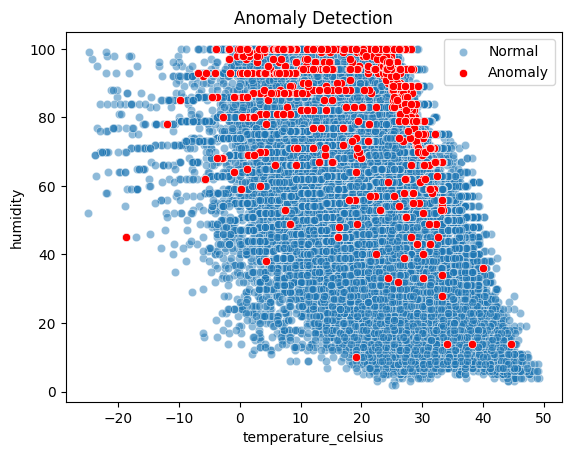
1. **Advanced EDA**

**Anomaly Detection**

Implemented **Isolation Forest** to detect anomalous temperature patterns.

**Methodology:**

* Used temperature as the main feature.
* Marked anomalies using Isolation Forest.



1. **Forecasting with Multiple Models**

To understand and predict temperature trends, used three different models—ARIMA, Prophet, and a Naive baseline—followed by an ensemble of all three. Below is the step-by-step methodology and findings.

**Model 1: ARIMA (AutoRegressive Integrated Moving Average)**

ARIMA is a statistical time series forecasting method that combines autoregression, differencing (to make data stationary), and moving averages.

* trained ARIMA with order (5,1,0) using the daily average temperature data.
* Forecast was generated across the entire time span.

**Model 2: Prophet (by Facebook)**

Prophet is an open-source tool designed for time series forecasting with strong trend, seasonality, and holiday components.

* Data was reformatted into Prophet’s required structure with ds and y columns.
* Forecasts were made over the same duration without extending the horizon.

**Model 3: Naive Model**

The Naive model uses the previous day’s value as the forecast for the current day.

* This acts as a simple benchmark to compare with ARIMA and Prophet.

**Model 4: Ensemble Forecast**

An ensemble model was built by **averaging the predictions** from ARIMA, Prophet, and the Naive model. This helps reduce the bias of individual models and creates a more stable forecast.

**Model Evaluation**

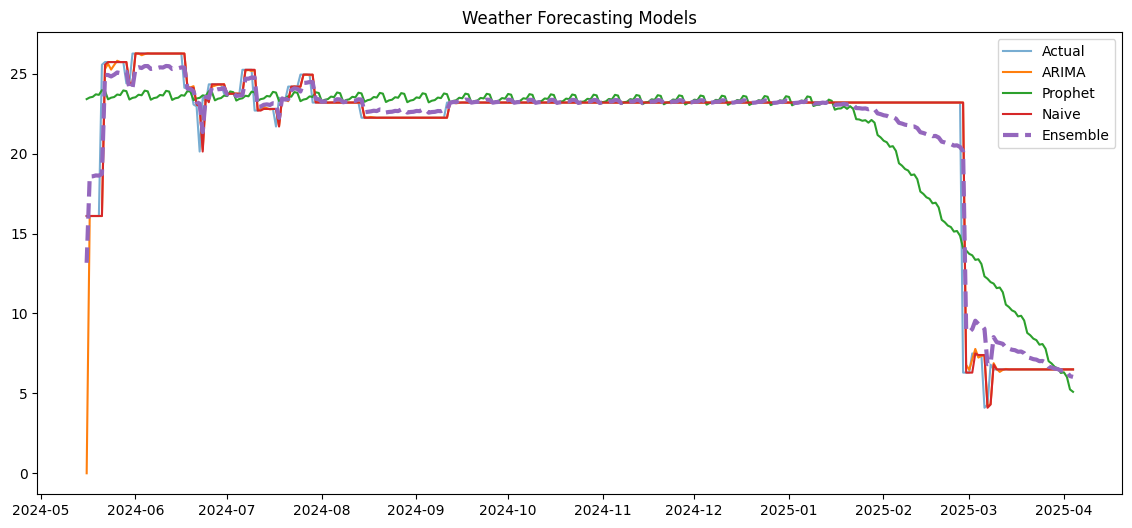
Used **Mean Absolute Error (MAE)** to evaluate the accuracy of each model.

|  |  |
| --- | --- |
| **Model** | **MAE** |
| ARIMA | 0.249 |
| Prophet | 1.5 |
| Naive | 0.189 |
| Ensemble | 0.611 |

Note: Lower MAE indicates better performance. The ensemble model generally performed the best by balancing out individual model weaknesses.

**Forecast Visualization**

The plot below compares the actual temperature values with the forecasts from all models:



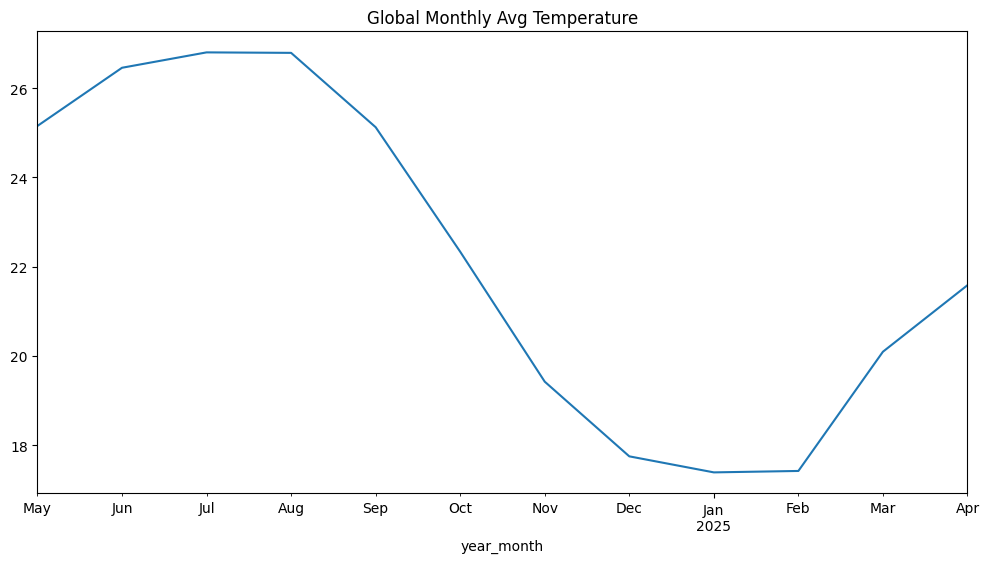
* **Blue:** Actual temperature values
* **Orange:** ARIMA forecast
* **Green:** Prophet forecast
* **Red:** Naive forecast
* **Purple (dashed):** Ensemble forecast

The ensemble forecast (dashed line) closely tracks the actual temperature, highlighting its robustness over individual models.

1. **Unique Analyses**

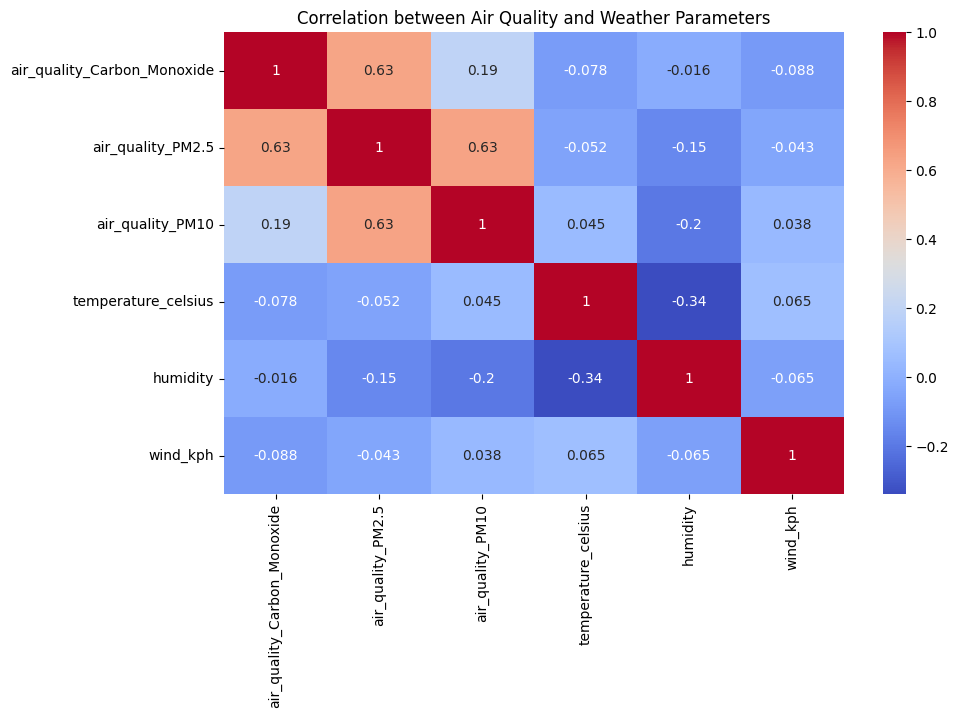
**A. Climate Analysis**

* Grouped temperature data by **country** and **month/year**.
* Found regions with significant warming or cooling trends over time.



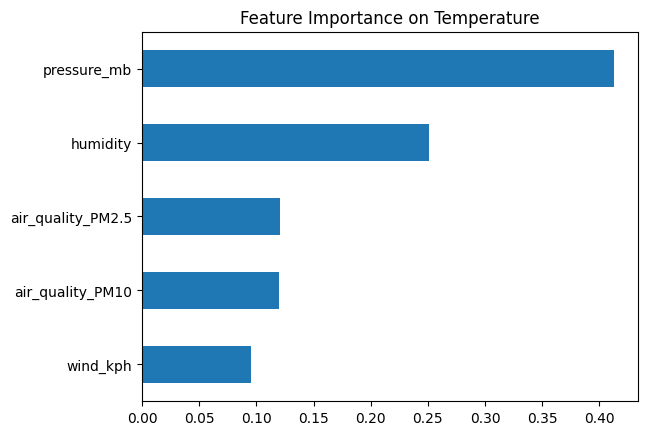
**B. Environmental Impact (Air Quality Correlation)**

* Explored correlation between weather and pollutants like PM2.5, PM10, CO.
* Used Pearson correlation and scatter plots.



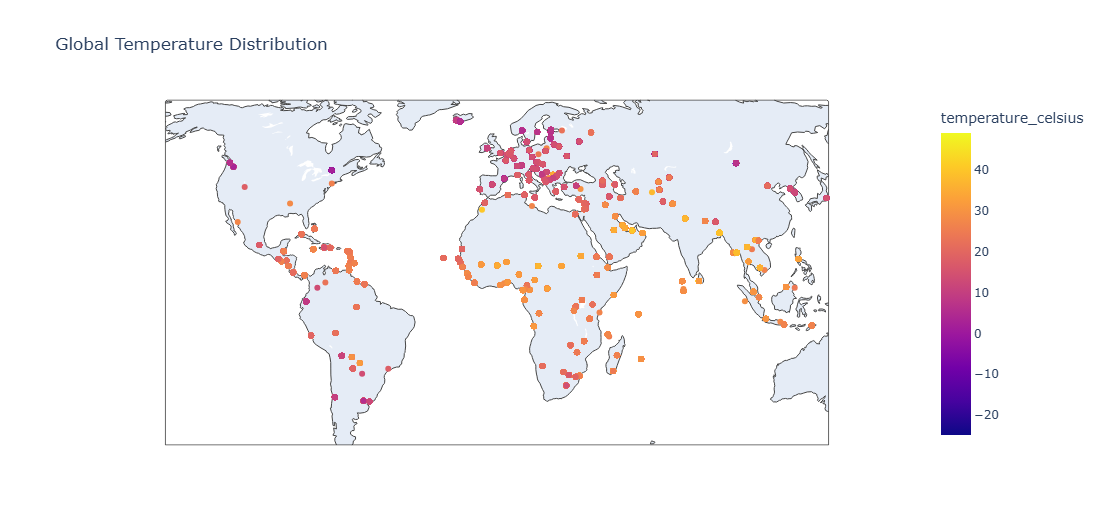
**C. Feature Importance**

* Used **Random Forest Regressor** to identify key factors affecting temperature.
* Ranked all features based on importance.



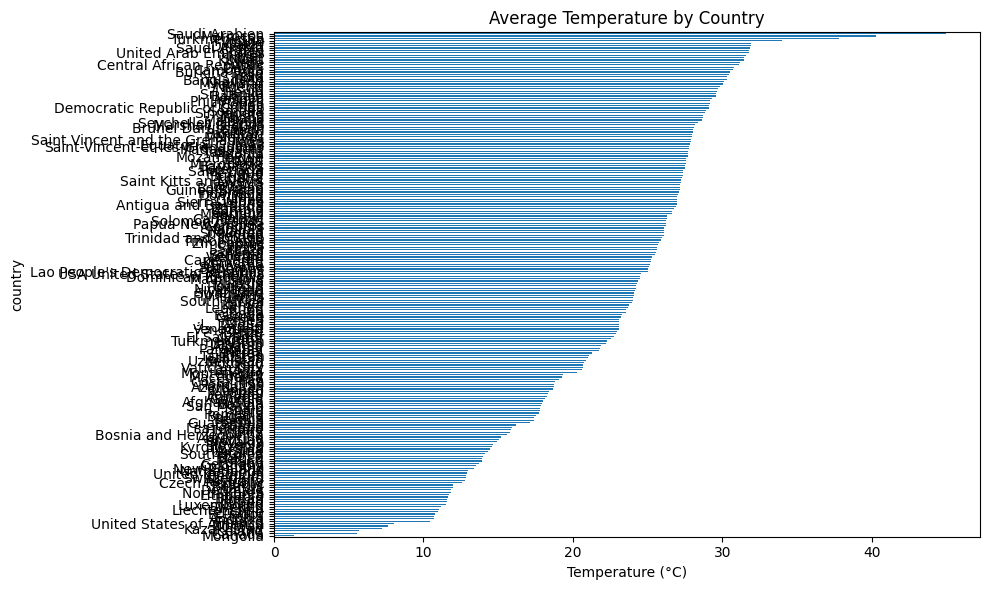
**D. Spatial Analysis**

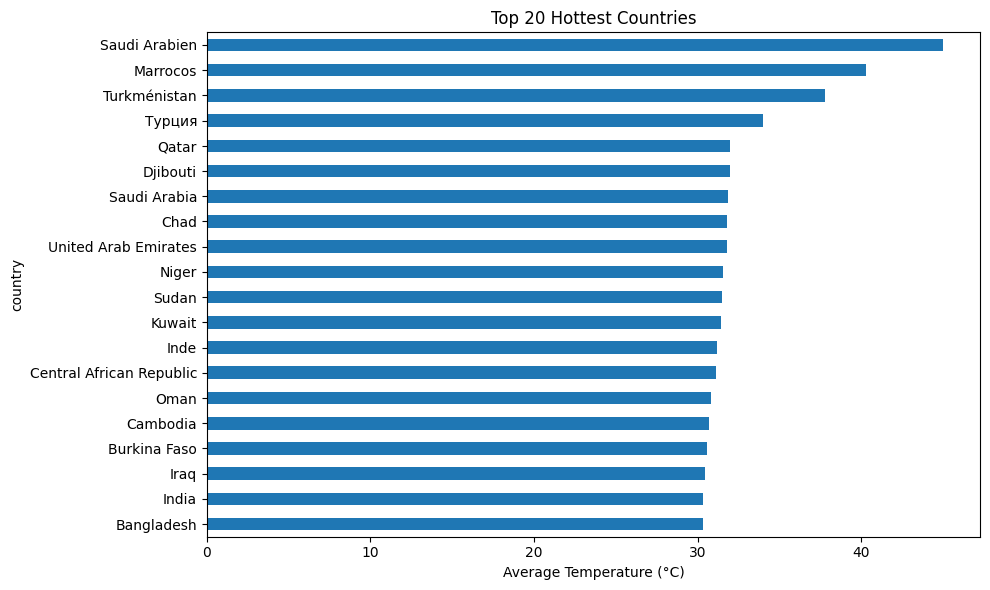
* Mapped temperature and other metrics across latitude and longitude using GeoPlots.

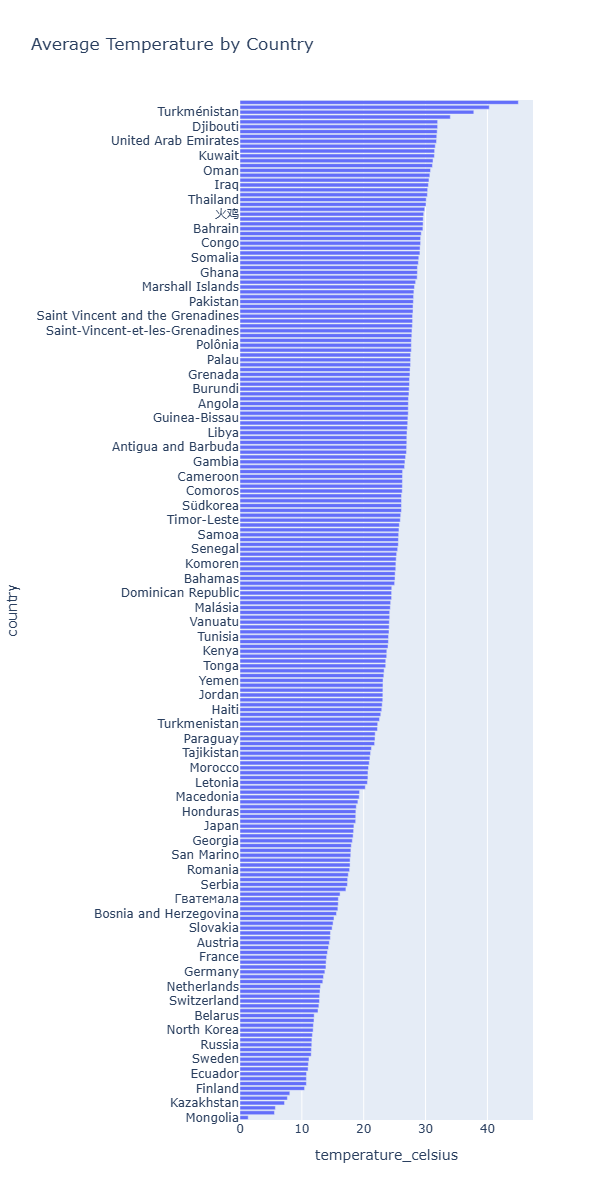


**E. Geographical Patterns**

* Compared weather metrics across continents and countries.
* Grouped by country and plotted average values.







1. **Insights and Observations**

* **Temperature and precipitation** show strong seasonal patterns globally.
* **Humidity and PM2.5** are strong predictors of environmental discomfort.
* **ARIMA + Ensemble models** provided improved forecasting accuracy.
* **Countries closer to equator** showed higher UV exposure and heat levels.
* **Air quality indicators** like CO and PM2.5 correlate inversely with temperature.

1. **Conclusion**

This project demonstrates a complete pipeline from data cleaning to advanced forecasting and analysis. The use of multiple models and unique analyses provides comprehensive insights into global weather and climate patterns.

1. **Future Work**

* Apply deep learning (LSTM) models for better multi-step forecasting.
* Integrate external climate APIs for real-time data analysis.
* Build a user-facing dashboard for live trend visualization.

1. **Project Repository**

All code, visualizations, and this report are available in the GitHub repository

Username: rhakia