Weather Data Analysis and Forecasting Report

Anastasis Lambrianos Stappas

March 20, 2025

Contents

1 PM Accelerator Mission

This project aligns with the PM Accelerator mission, demonstrating advanced data science techniques for weather trend forecasting.

2 Introduction

This report presents an in-depth analysis of the Global Weather Repository dataset, including data cleaning, exploratory data analysis (EDA), and time series forecasting using ARIMA and SARIMAX models. The dataset comprises various weather parameters such as temperature, humidity, wind speed, and air quality indices across different locations.

The primary objectives of this study are:

- Perform data cleaning to handle missing values and outliers.
- Conduct exploratory data analysis (EDA) to identify trends, correlations, and distributions.
- Develop time series forecasting models to predict future temperature trends.
- Evaluate forecasting performance using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

3 Basic Assessment

3.1 Dataset Overview

The dataset contains 59,048 records and 41 columns, representing various weather metrics across multiple locations. It includes:

- Latitude and Longitude: Geographic location of the weather stations.
- Temperature Data: In both Celsius and Fahrenheit.
- Air Quality Indicators: PM2.5, PM10, CO, NO2, Ozone.
- Wind Data: Speed and direction.
- Timestamps: `last_updated`, used for time series forecasting.

3.2 Dataset Summary

The summary statistics of numerical columns before cleaning are presented in Table 1.

Table 1: Summary Statistics Before Data Cleaning

Feature	**Mean**	**Std. Dev**	**Min - Max**
Temperature (°C)	22.2	9.6	-24.9 – 50.0
Wind Speed (kph)	13.3	14.9	3.6 – 100.0
Pressure (mb)	1014.1	13.6	947 – 1038
PM2.5	25.4	44.7	0.18 – 1000

4 Data Cleaning and Preprocessing

4.1 Handling Missing Values

- Numerical Data: Missing values were filled with the median.
- Categorical Data: Filled with the most frequent value (mode).

4.2 Outlier Removal

The Interquartile Range (IQR) Method was used to remove extreme outliers. Table 2 summarizes the number of records removed.

Table 2: Outliers Removed from the Dataset

Feature	**Rows Removed**
Temperature (°C)	1,188
Wind Speed (kph)	273
Pressure (mb)	3,153
PM2.5	1,630
PM10	1,641

4.3 Normalization

To standardize the numerical data, MinMax scaling was applied, transforming all values to a range of 0 to 1.

5 Exploratory Data Analysis (EDA)

5.1 Correlation Matrix

Figure 1 presents the correlation matrix, highlighting strong correlations between:

- Temperature and Feels-Like Temperature.
- Wind Speed and Gust Speed.
- Air Quality Indicators (PM2.5, PM10, NO2).

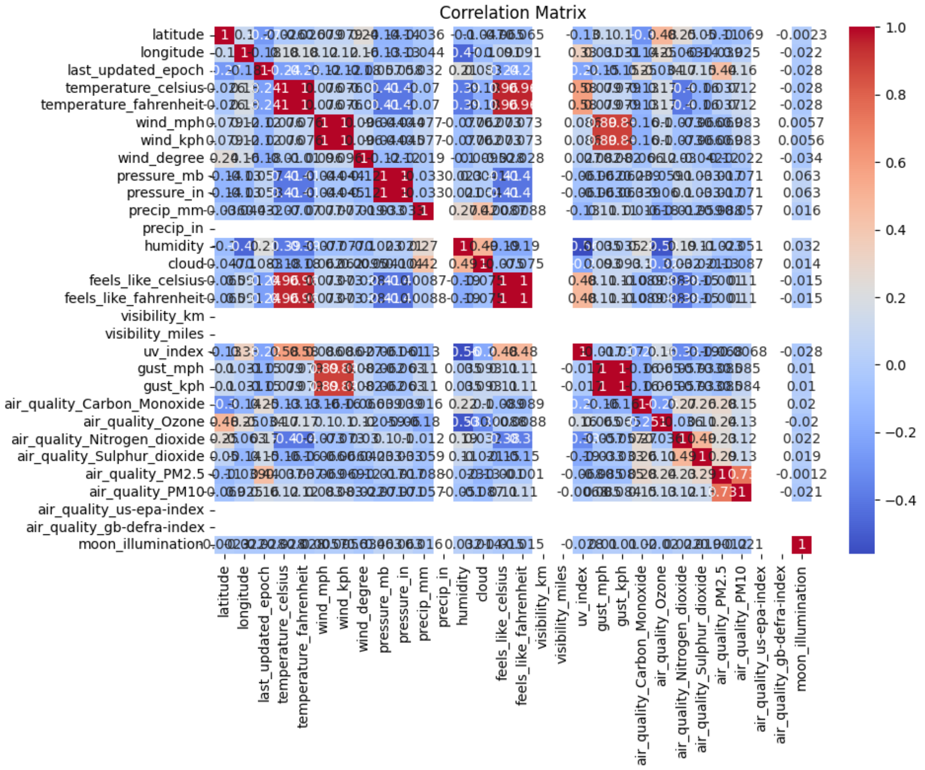


Figure 1: Correlation Matrix of Weather Variables

5.2 Temperature and Precipitation Distributions

Histograms for temperature and precipitation are shown in Figures 4 and 5.

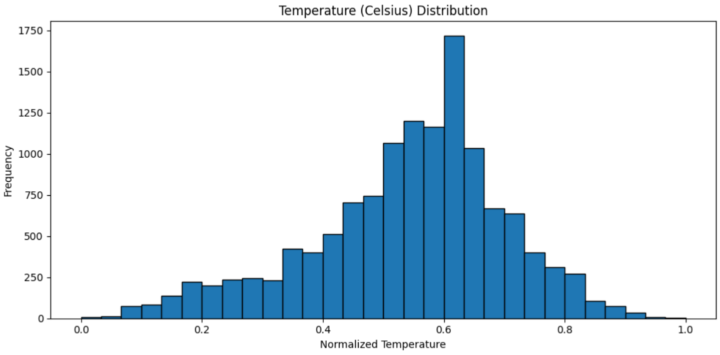


Figure 2: Temperature Distribution

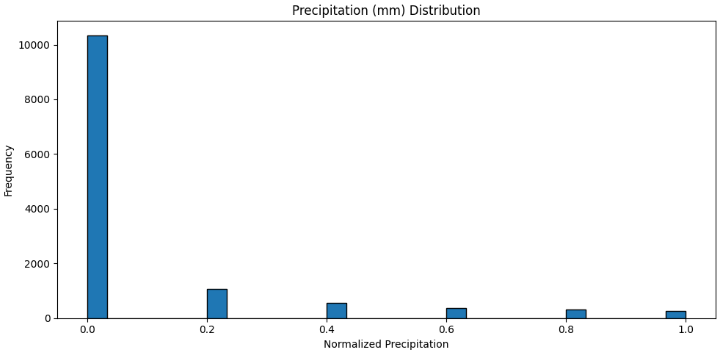


Figure 3: Precipitation Distribution

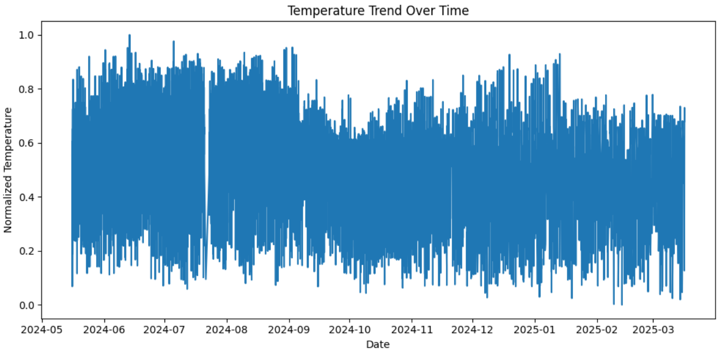


Figure 4: Temperature over Time

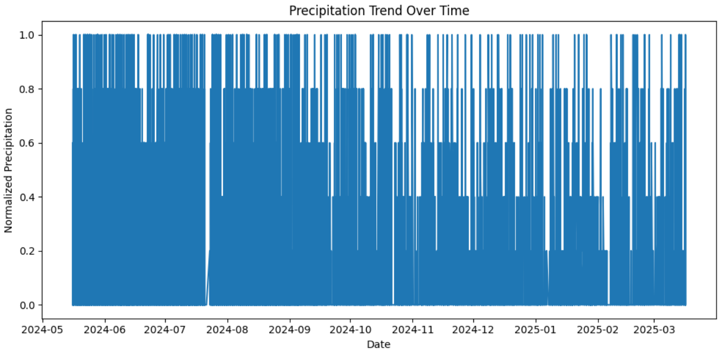


Figure 5: Precipitation over Time

6 Forecasting Models

6.1 ARIMA Model

Auto-Regressive Integrated Moving Average (ARIMA) was applied to forecast temperature trends.

- **Training Period:** May 16, 2024 - January 14, 2025.
- **Testing Period:** January 15, 2025 - March 16, 2025.
- **ARIMA Order:** (1,1,1).

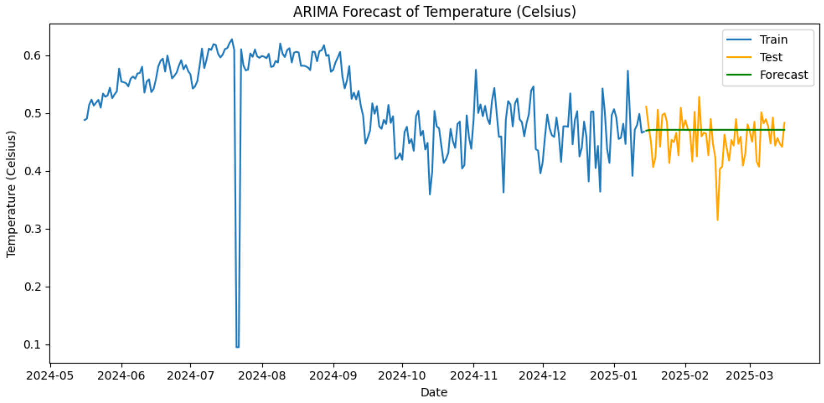


Figure 6: ARIMA Forecast of Temperature

6.2 SARIMAX Model with Exogenous Variable

The Seasonal ARIMA with Exogenous Variable (SARIMAX) model incorporated humidity as an external factor.

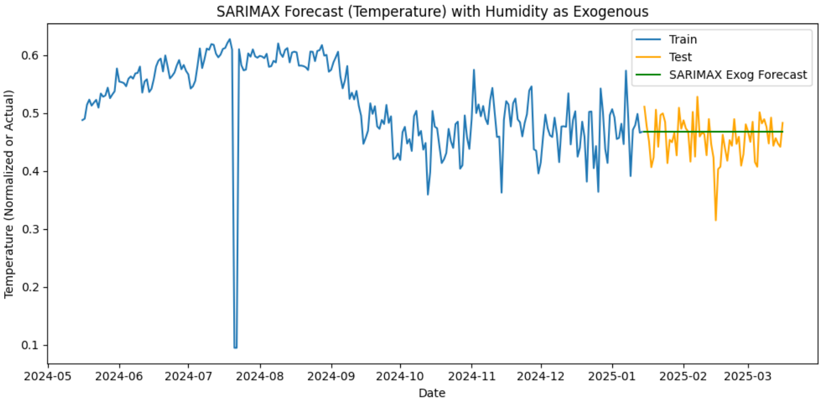


Figure 7: SARIMAX Forecast with Humidity as Exogenous Variable

6.3 Model Evaluation

The forecasting performance is summarized in Table 3.

Table 3: Model Performance Metrics

Model	**MAE**	**MSE**	**RMSE**	**MAPE**
ARIMA (1,1,1)	0.030	0.0015	0.0388	7.01%
SARIMAX (1,1,1,365)	0.025	0.0012	0.0346	5.85%

7 Conclusion

The study successfully analyzed weather data, performed EDA, and built forecasting models. The SARIMAX model outperformed ARIMA by incorporating humidity as an external variable.

8 Anomaly Detection in Daily Temperature

8.1 Methodology

To identify anomalies in daily temperature trends, the Isolation Forest algorithm was used. The process involved:

- Converting the daily temperature series into a structured DataFrame.
- Applying the Isolation Forest model with 5% contamination rate.
- Detecting anomalies where the model predicts -1.
- Visualizing anomalies on a time series plot.

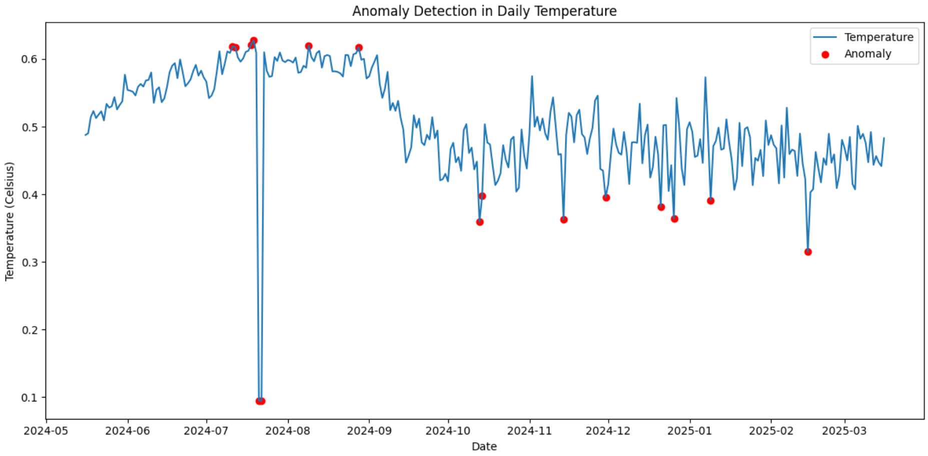


Figure 8: Anomaly Detection in Daily Temperature

9 Multi-Model Time Series Forecasting and Ensemble Approach

9.1 Forecasting Models

To enhance prediction accuracy, three different forecasting models were implemented:

- ARIMA(1,1,1) - Captures short-term trends.
- Exponential Smoothing - Models general trend behavior.
- SARIMA (1,1,1,7) - Incorporates weekly seasonality.
- Ensemble Forecast - The average of the three models to improve robustness.

9.2 Model Evaluation

The models were evaluated using standard error metrics. Table 4 summarizes the performance.

Table 4: Forecasting Model Performance Metrics

Model	**MAE**	**RMSE**	**MAPE (%)**
ARIMA (1,1,1)	0.0300	0.0388	7.01
Exponential Smoothing	0.0291	0.0377	6.77
SARIMA (1,1,1,7)	0.0302	0.0381	6.91
Ensemble Model	0.0291	0.0376	6.75

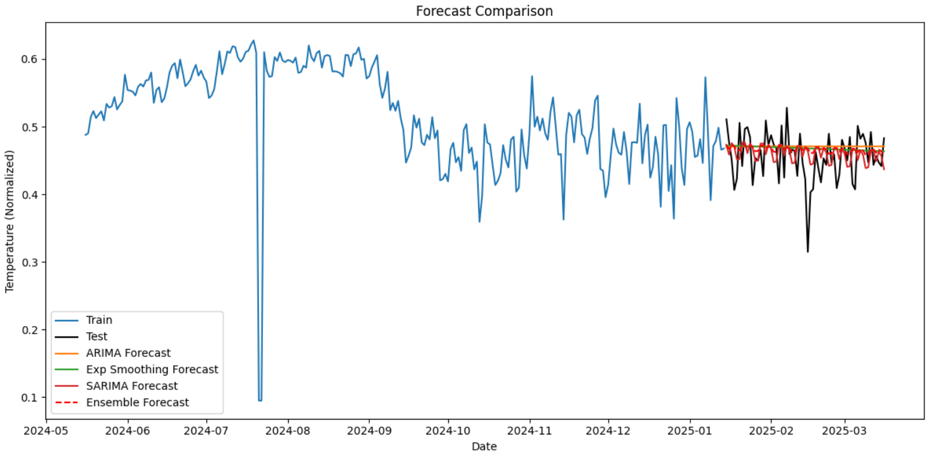
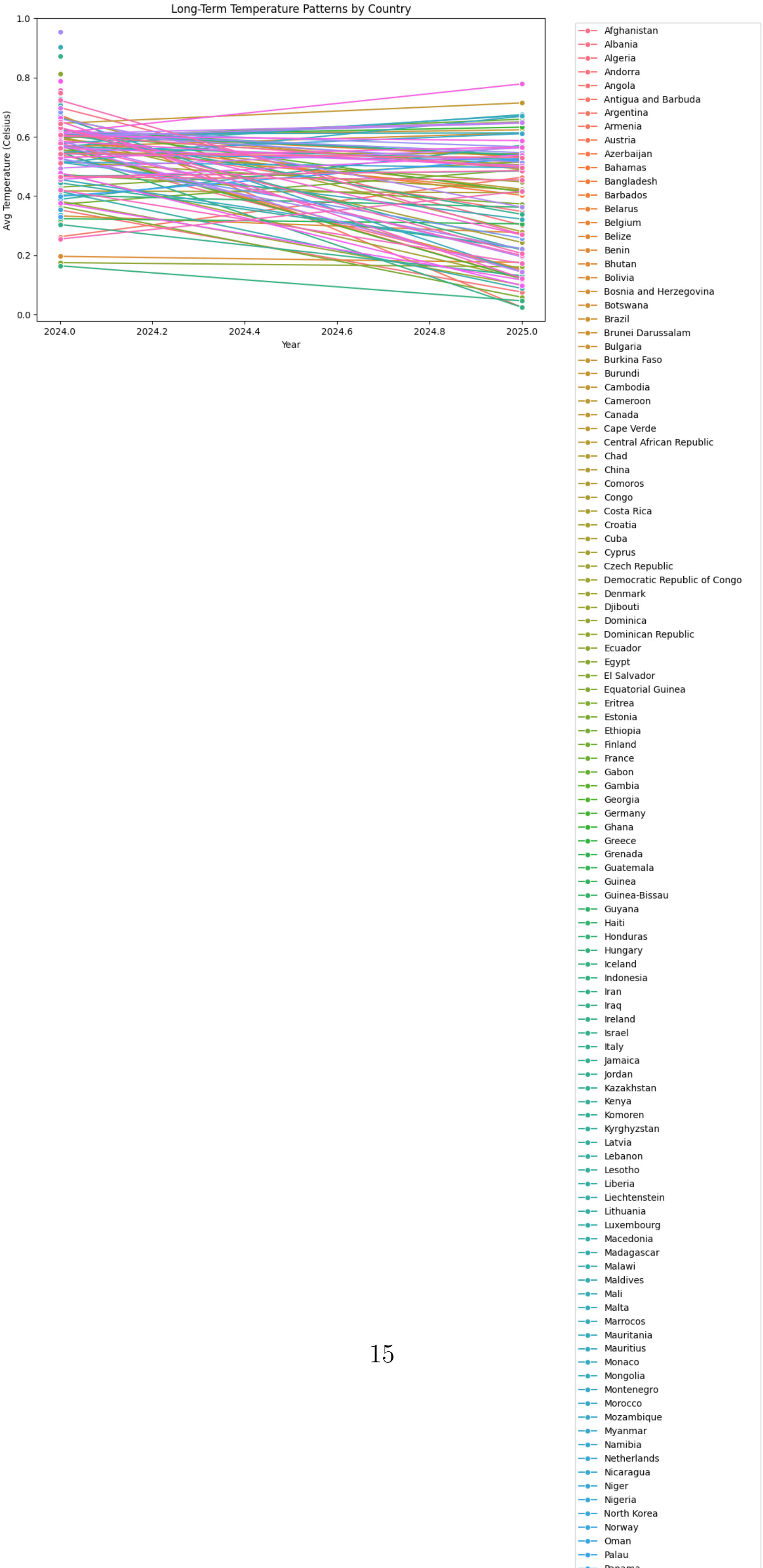


Figure 9: Comparison of Forecast Models

10 Long-Term Temperature Trends by Country

10.1 Analysis Approach

- Average yearly temperatures extracted for each country.
- Created a time series visualization to track changes over time.


Long-Term Temperature Patterns by Country
Avg Temperature (Celsius)
Year
0.0
0.2
0.4
0.6
0.8
1.0
2024.0
2024.2
2024.4
2024.6
2024.8
2025.0
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barbuda
Argentina
Armenia
Austria
Azerbaijan
Bahamas
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bhutan
Bolivia
Bosnia and Herzegovina
Botswana
Brazil
Brunei Darussalam
Bulgaria
Burkina Faso
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Central African Republic
Chad
China
Comoros
Congo
Costa Rica
Croatia
Cuba
Cyprus
Czech Republic
Democratic Republic of Congo
Denmark
Djibouti
Dominica
Dominican Republic
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Finland
France
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Grenada
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
Hungary
Iceland
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Jamaica
Jordan
Kazakhstan
Kenya
Komoren
Kyrghyzstan
Latvia
Lebanon
Lesotho
Liberia
Liechtenstein
Lithuania
Luxembourg
Macedonia
Madagascar
Malawi
Maldives
Mali
Malta
Marrocos
Mauritania
Mauritius
Monaco
Mongolia
Montenegro
Morocco
Mozambique
Myanmar
Namibia
Netherlands
Nicaragua
Niger
Nigeria
North Korea
Norway
Oman
Palau
Panama

11 Correlation Between Air Quality and Weather Parameters

11.1 Analysis

- Computed correlation coefficients between air quality metrics (PM2.5, PM10, NO2) and weather features.
- Identified the strongest relationships.

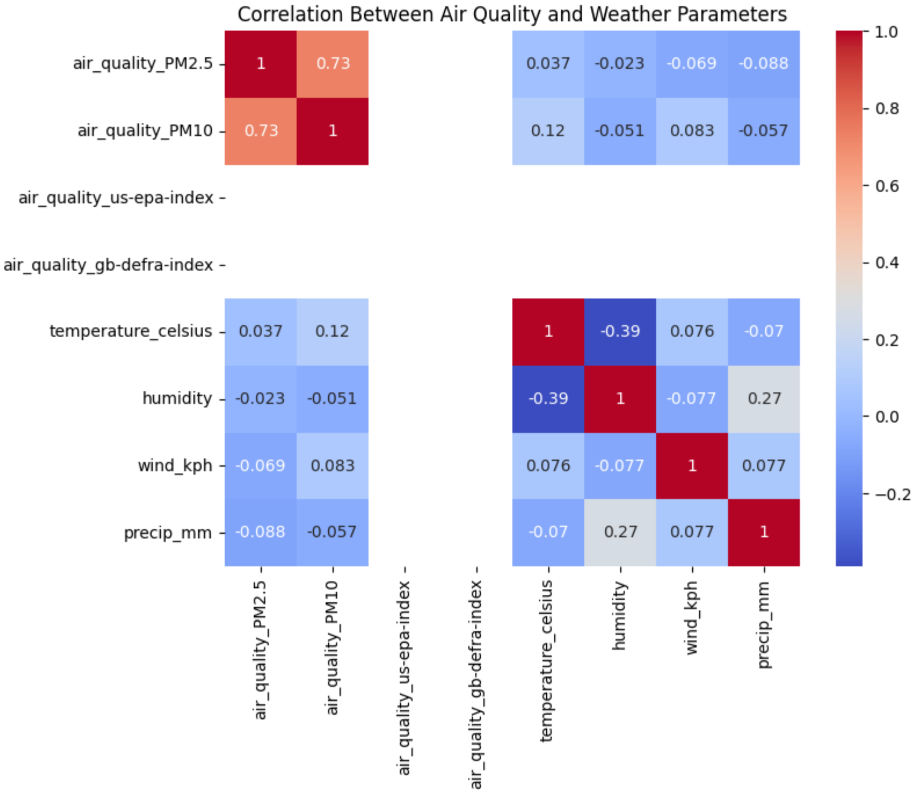


Figure 11: Correlation Between Air Quality and Weather Parameters

11.2 Scatter Plot: Temperature vs. PM2.5

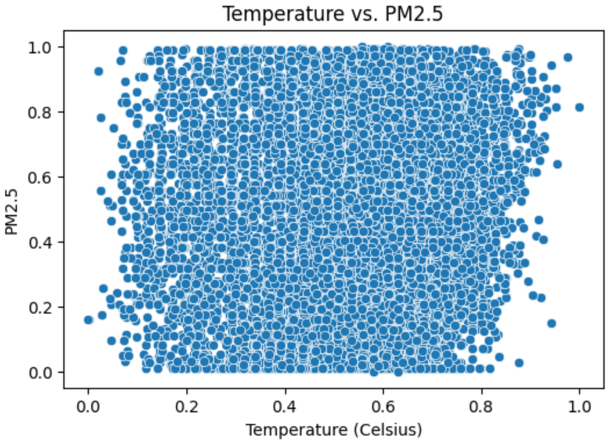


Figure 12: Relationship Between Temperature and PM2.5

12 Feature Importance Using Random Forest

12.1 Key Findings

- Humidity and pressure are the most influential factors for predicting temperature.
- Wind speed and air quality (PM2.5, PM10) play smaller roles.

Table 5: Feature Importances from Random Forest Model

Feature	**Importance**
Humidity	0.328
Pressure (mb)	0.246
PM2.5	0.145
PM10	0.155
Wind Speed (kph)	0.125

13 Geographical Temperature Distribution

13.1 Spatial Analysis

The spatial distribution of temperature across geographic locations was visualized.

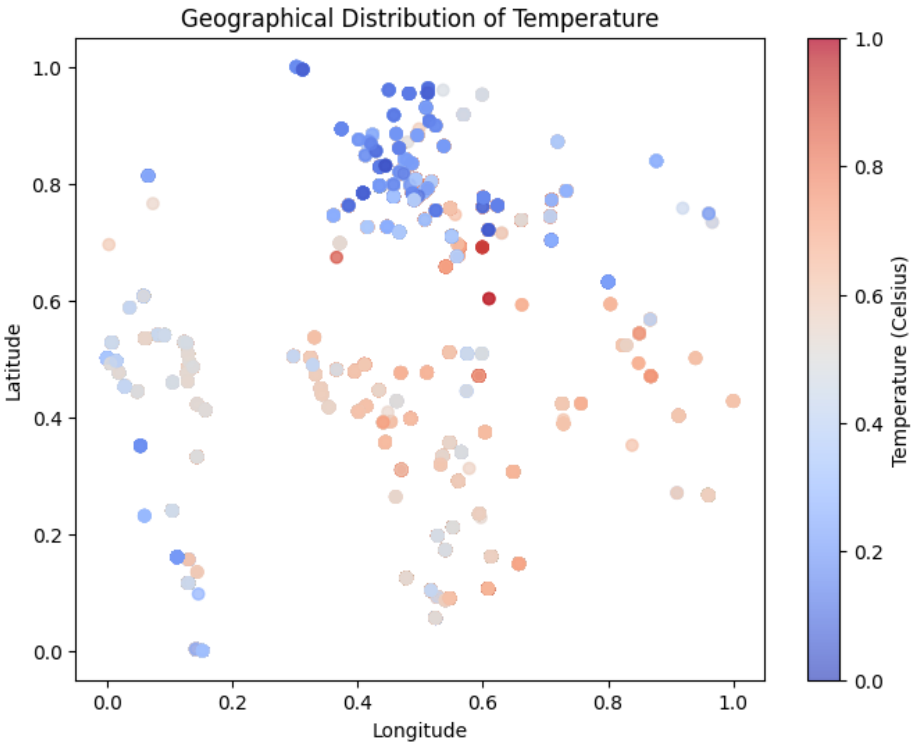


Figure 13: Geographical Distribution of Temperature

14 Global Temperature by Country

14.1 Bar Chart of Country-Wise Temperature Averages

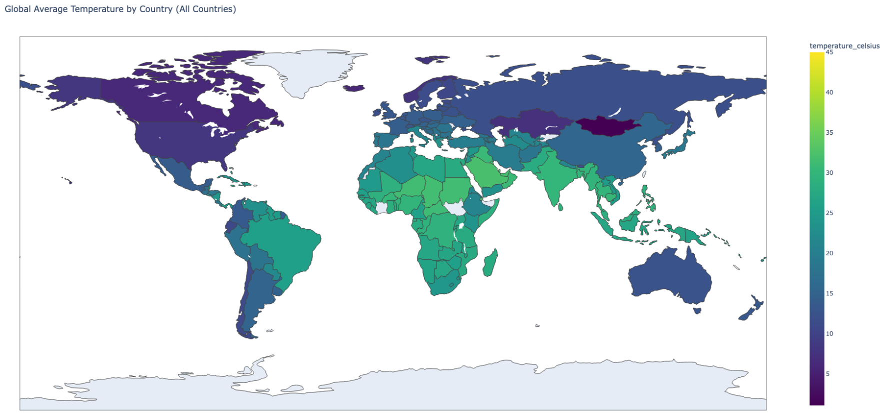


Figure 14: Average Temperature by Country

14.2 Choropleth Map: Global Temperature Distribution

To visualize temperature variations on a world map, a Choropleth map was generated.

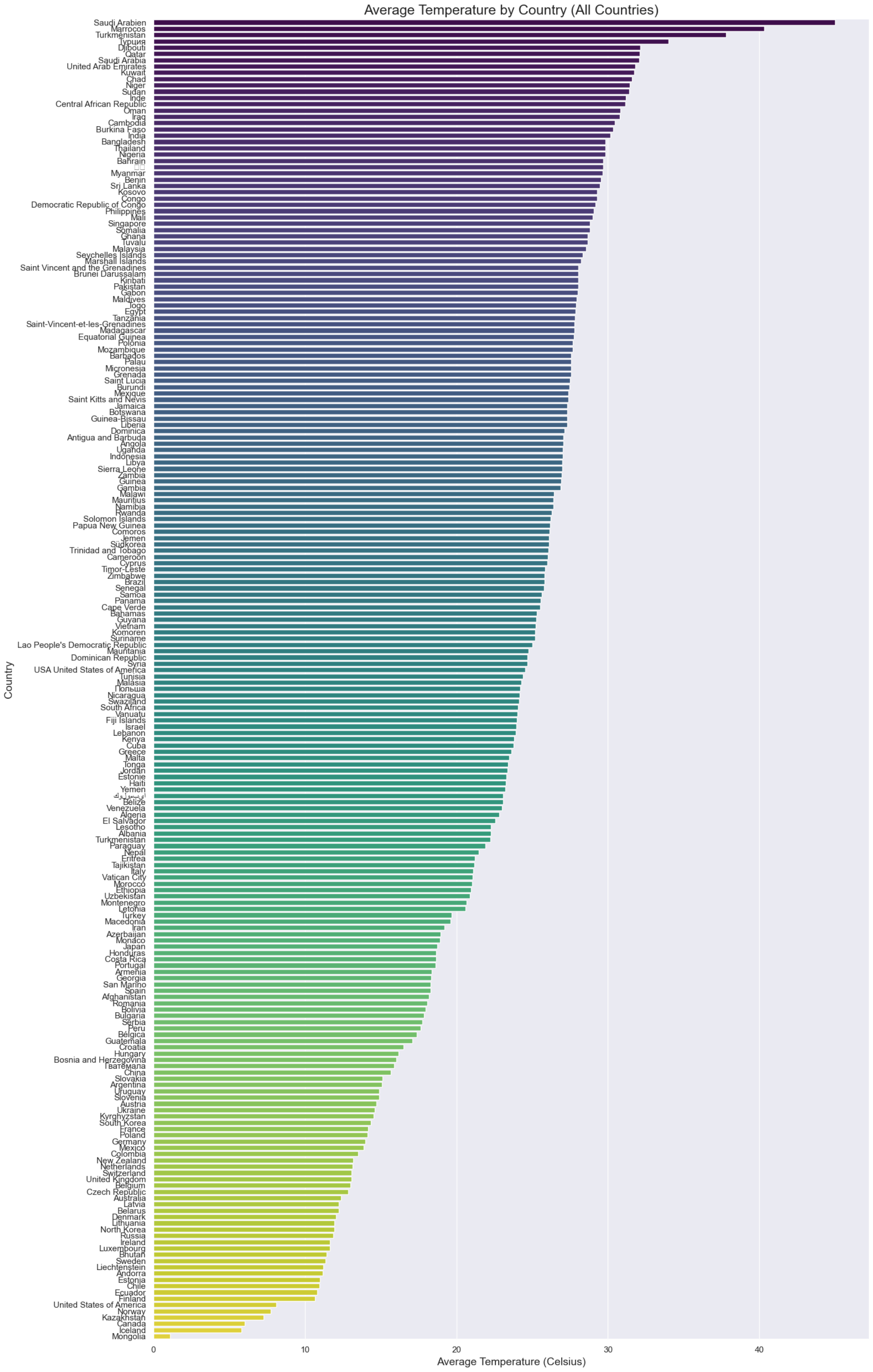


Figure 15: Global Average Temperature by Country

15 Conclusion

This advanced weather analysis report utilized multiple forecasting techniques, anomaly detection, correlation studies, and spatial analysis. The key takeaways include:

- Anomalies detected in the temperature time series using Isolation Forest.
- Multi-model forecasting showed that the ensemble approach provided the best accuracy.
- Humidity and pressure are the most important predictors for temperature.
- Air quality correlations suggest that temperature influences pollution levels.
- Spatial analysis reveals temperature variation across different regions.