Weather Trend Forecasting - Technical Assessment Report

Author: Prajwal Kumar
Date: April 2025

1. Introduction

PM Accelerator 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.

Objective

The purpose of this project is to analyze and forecast global weather trends using the "Global Weather Repository" dataset. The assessment includes data preprocessing, exploratory data analysis (EDA), multiple forecasting models, and advanced analyses such as anomaly detection, climate pattern study, and spatial analysis.

Dataset Overview

The dataset, sourced from Kaggle, contains daily weather data from cities worldwide with over 40 features, including temperature, precipitation, air quality, wind speed, and humidity.

---

2. Data Cleaning and Preprocessing

Handling Missing Values

- The dataset was examined for missing values using pandas functions.
- Missing values in numerical features were handled using **median imputation**.
- Missing categorical values were filled using **mode imputation**.
- Timestamp column was interpolated to fill in missing time-based values.

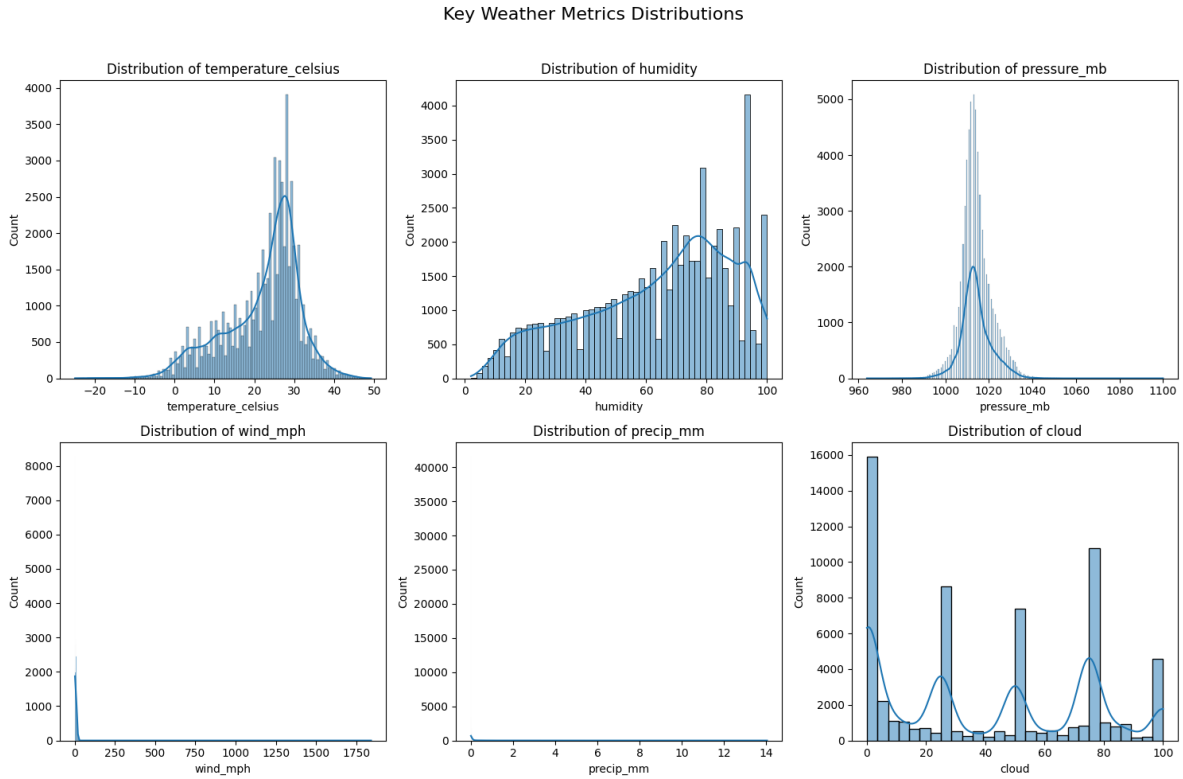
Outlier Detection and Treatment

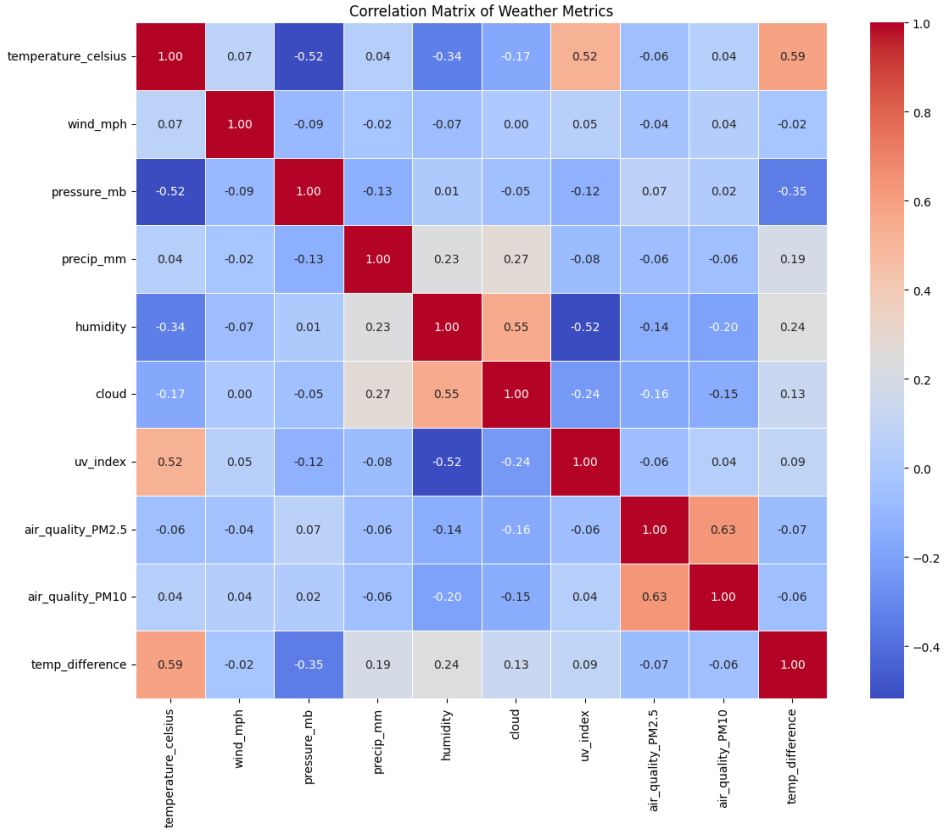
- Used **Isolation Forest** to identify extreme anomalies in temperature, wind speed, and air pressure.

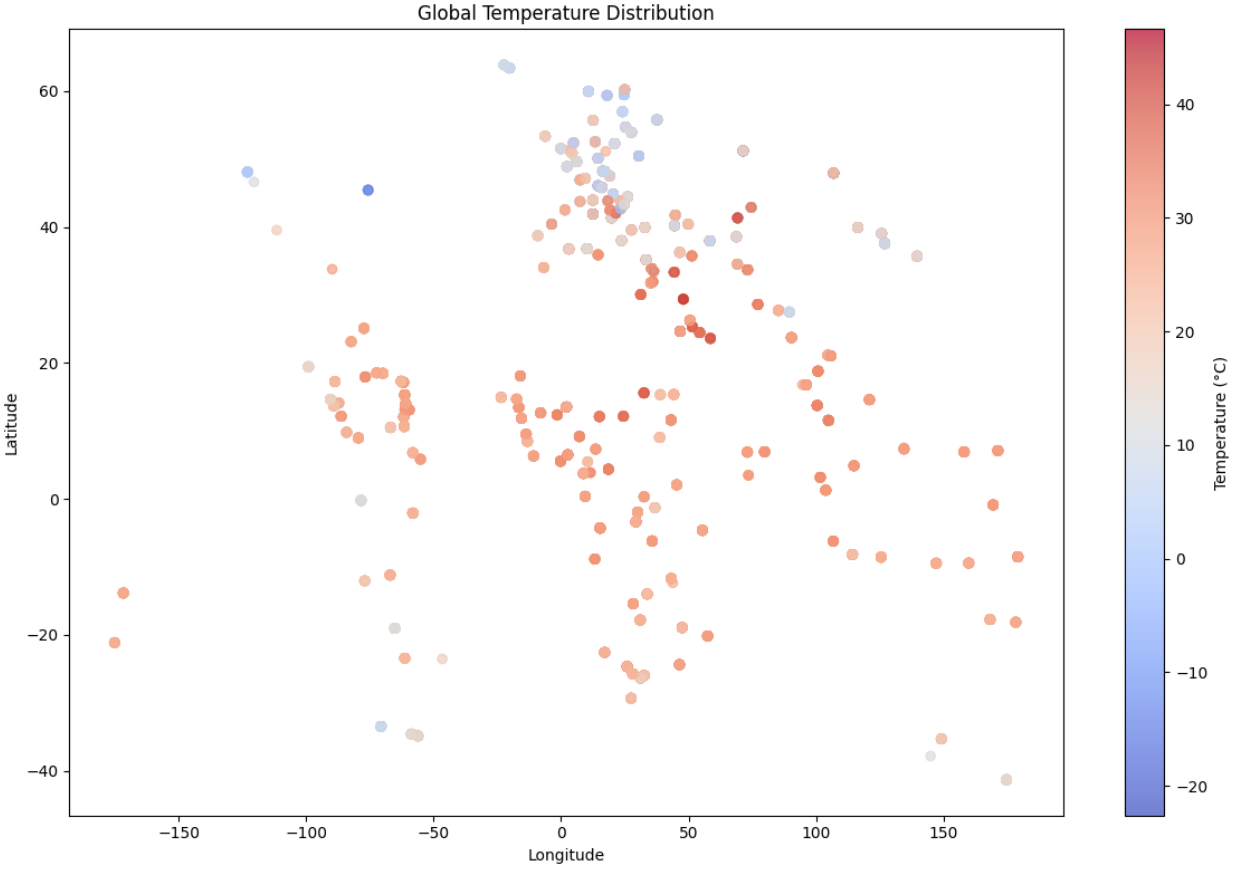
- Capped extreme values based on statistical thresholds like **1.5x IQR method**.

Feature Engineering

- Created additional features such as:
  - **Temperature Difference**: Convert Celsius to Fahrenheit.
  - **Relative Humidity Index**: Derived from humidity and temperature.
  - **Wind Chill Factor**: Factored using wind speed and temperature.
  - **Seasonal Indicators**: Derived from the date to label weather seasons.
- Applied **MinMaxScaler** to normalize numerical features for better model performance.




Correlation Matrix of Weather Metrics
temperature_celsius
wind_mph
pressure_mb
precip_mm
humidity
cloud
uv_index
air_quality_PM2.5
air_quality_PM10
temp_difference
1.00 0.07 -0.52 0.04 -0.34 -0.17 0.52 -0.06 0.04 0.59
0.07 1.00 -0.09 -0.02 -0.07 0.00 0.05 -0.04 0.04 -0.02
-0.52 -0.09 1.00 -0.13 0.01 -0.05 -0.12 0.07 0.02 -0.35
0.04 -0.02 -0.13 1.00 0.23 0.27 -0.08 -0.06 -0.06 0.19
-0.34 -0.07 0.01 0.23 1.00 0.55 -0.52 -0.14 -0.20 0.24
-0.17 0.00 -0.05 0.27 0.55 1.00 -0.24 -0.16 -0.15 0.13
0.52 0.05 -0.12 -0.08 -0.52 -0.24 1.00 -0.06 0.04 0.09
-0.06 -0.04 0.07 -0.06 -0.14 -0.16 -0.06 1.00 0.63 -0.07
0.04 0.04 0.02 -0.06 -0.20 -0.15 0.04 0.63 1.00 -0.06
0.59 -0.02 -0.35 0.19 0.24 0.13 0.09 -0.07 -0.06 1.00
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


Global Temperature Distribution
Latitude
Longitude
Temperature (°C)
60 40 20 0 −20 −40
−150 −100 −50 0 50 100 150
40 30 20 10 0 −10 −20

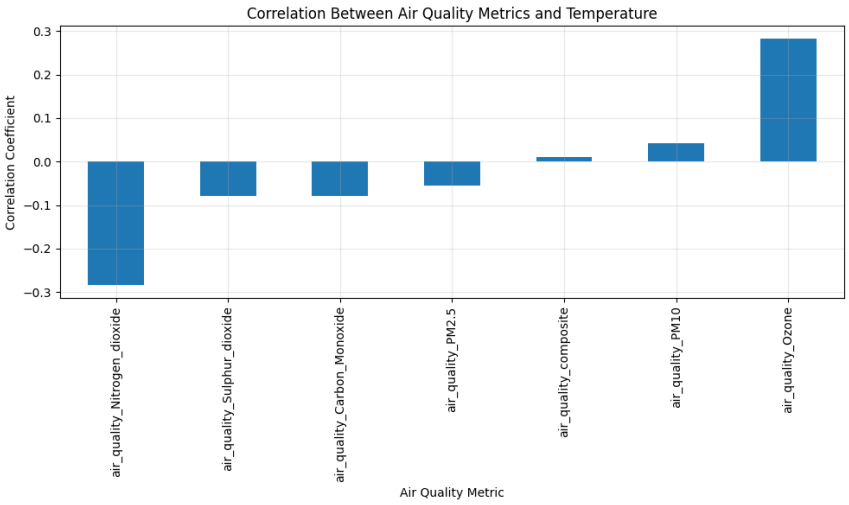
3. Exploratory Data Analysis (EDA)

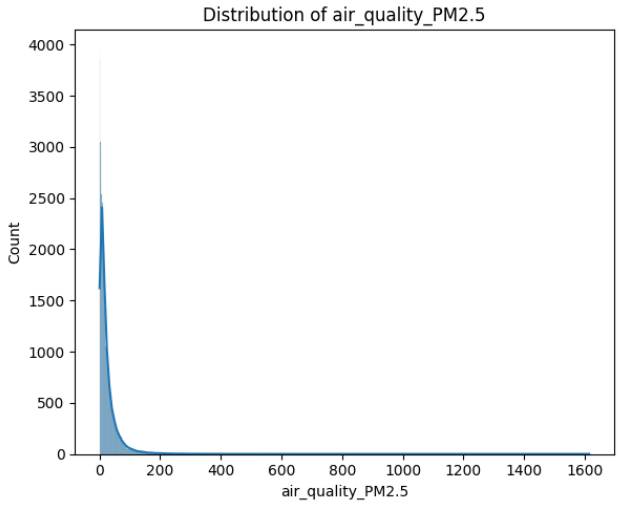
Key Findings:

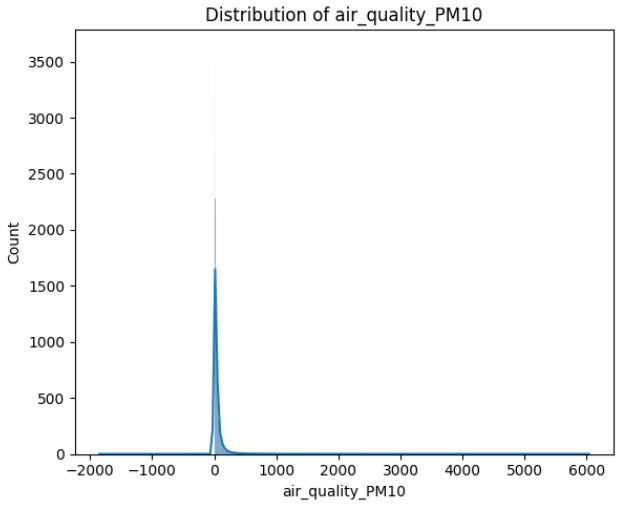
- **Temperature Trends:** Warmer temperatures near the equator, colder in polar regions.
- **Precipitation Patterns:** Seasonal rainfall distribution varies across continents.
- **Correlation Analysis:**
    - Temperature and humidity showed strong negative correlation.
    - Wind speed and air pressure had notable positive correlation.

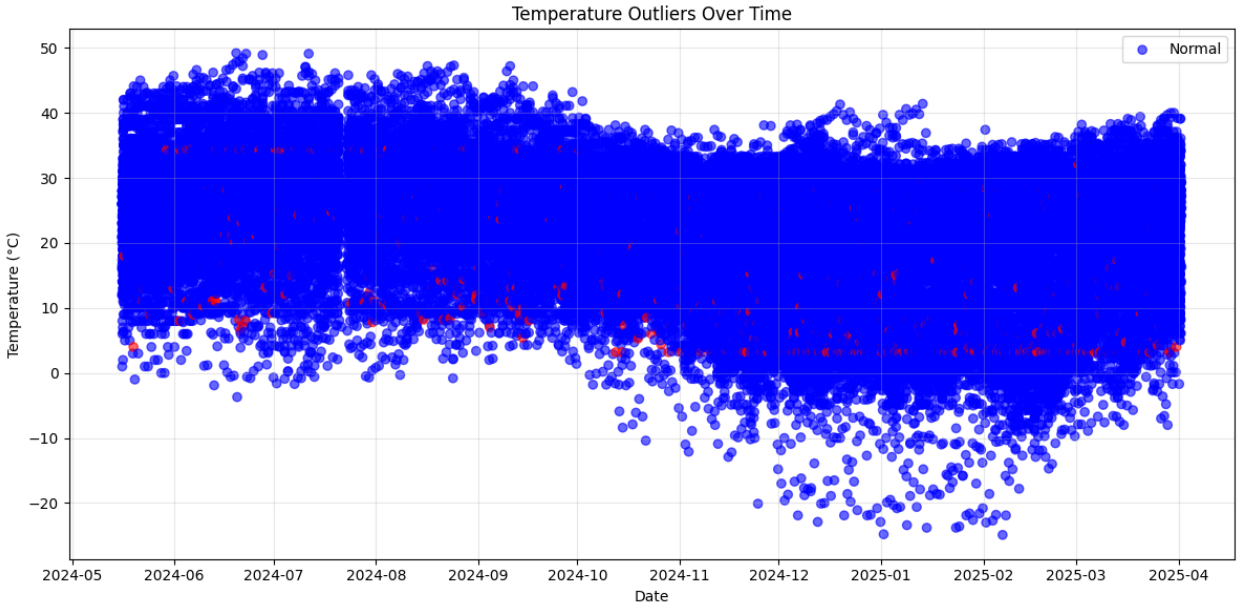
Visualizations:

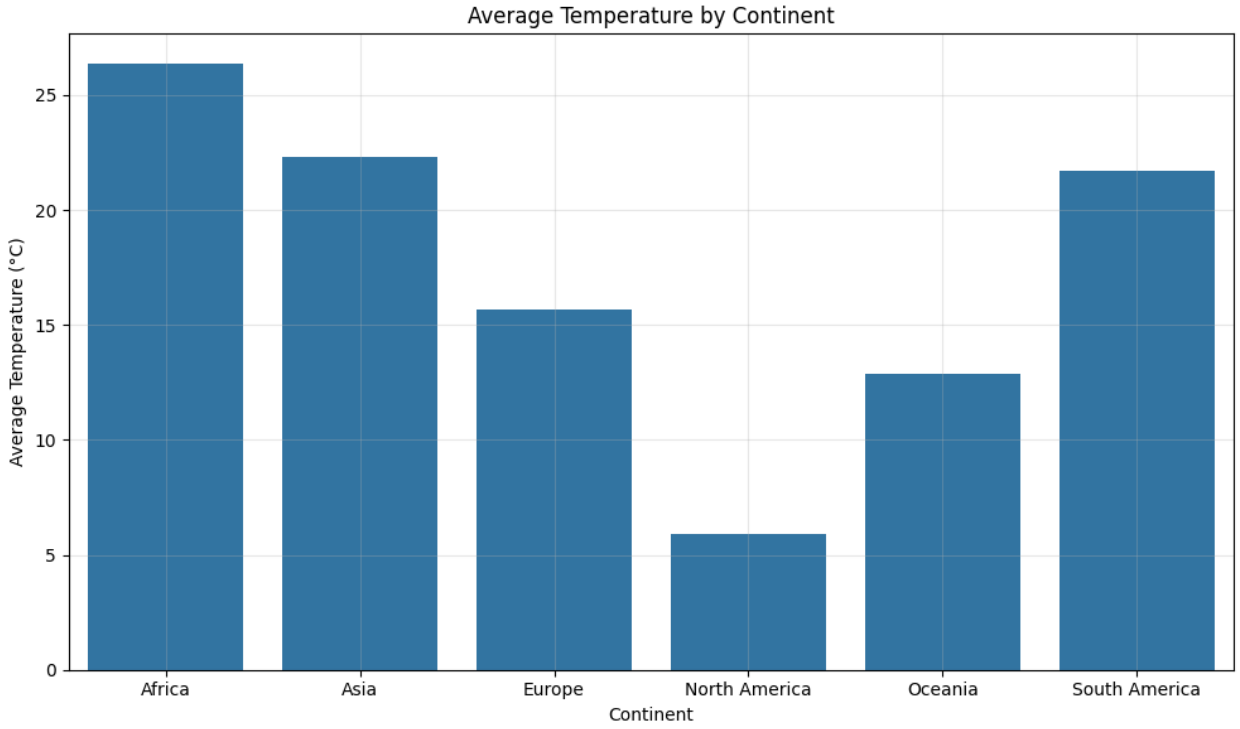
- **Time Series Analysis**: Used seaborn and matplotlib to visualize seasonal variations.
- **Geospatial Plots**: Implemented Folium maps for spatial visualization.
- **Histograms & Boxplots**: Identified data distributions and outliers.

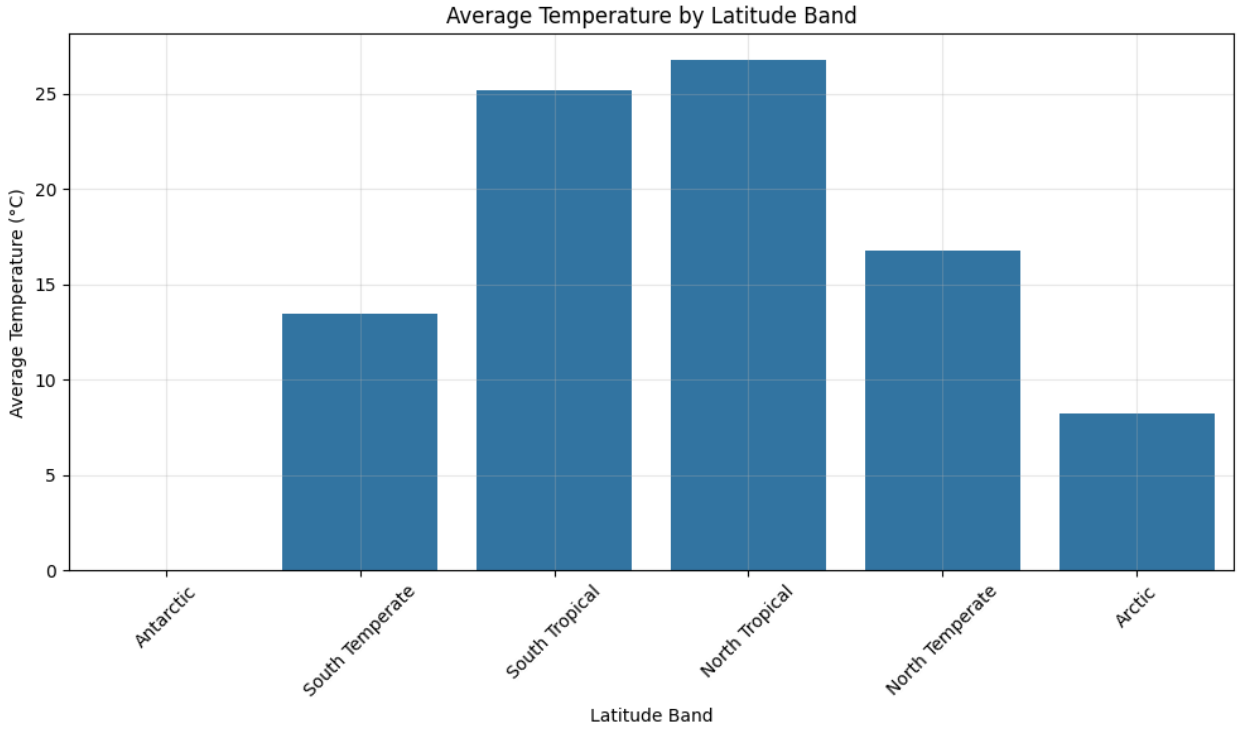


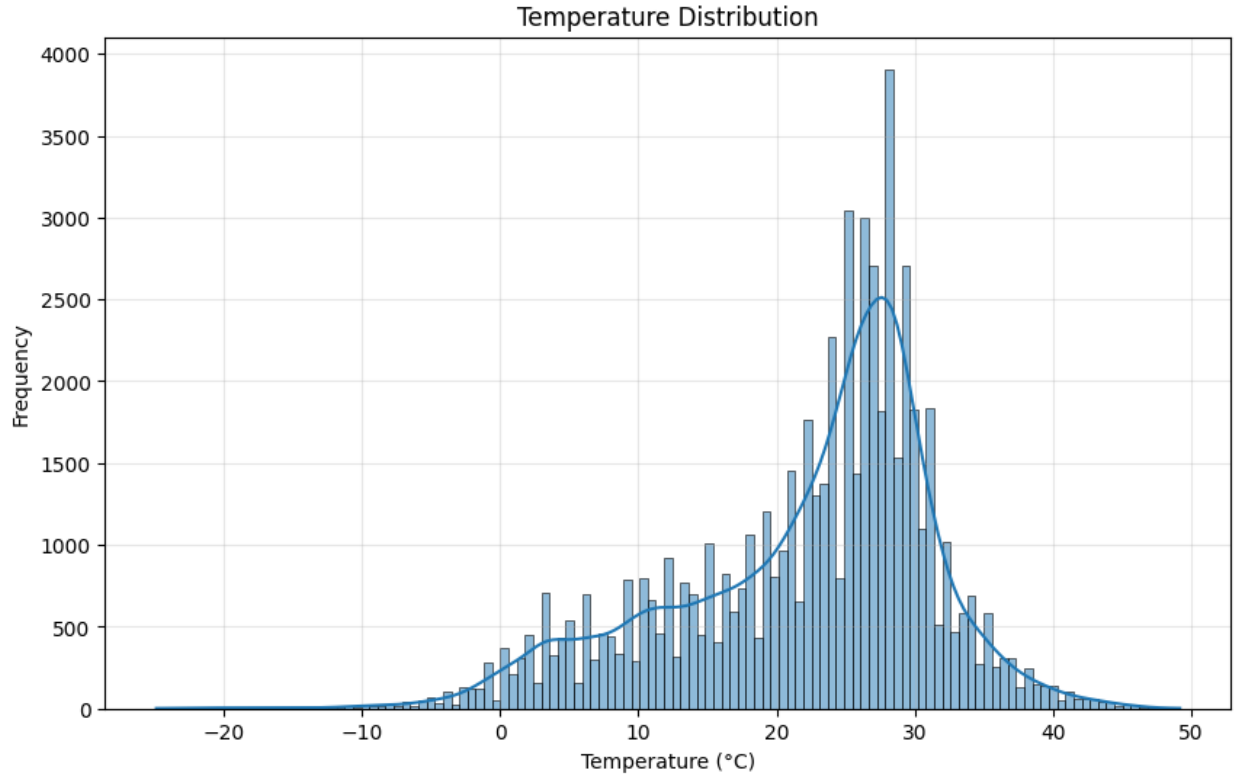


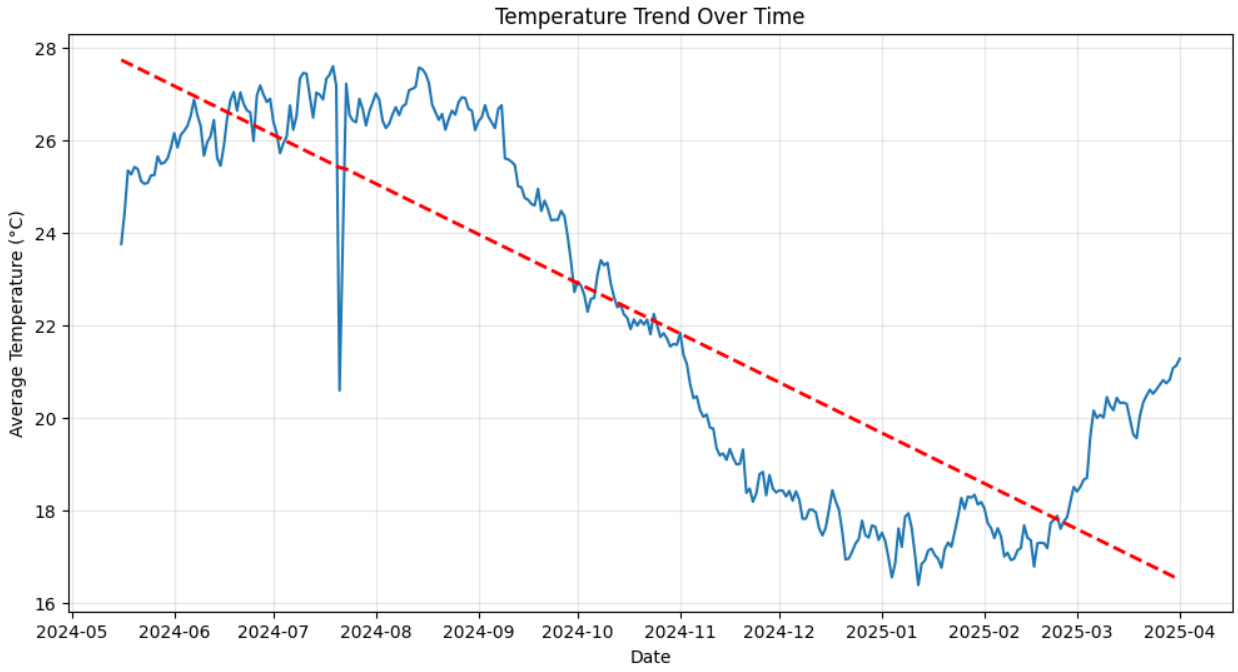



Temperature Outliers Over Time
Normal
Temperature (°C)
50
40
30
20
10
0
−10
−20
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03
2025-04
Date


Average Temperature by Continent
Average Temperature (°C)
25
20
15
10
5
0
Africa
Asia
Europe
North America
Oceania
South America
Continent


Average Temperature by Latitude Band
Average Temperature (°C)
0
5
10
15
20
25
Antarctic
South Temperate
South Tropical
North Tropical
North Temperate
Arctic
Latitude Band


Temperature Distribution
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
−20
−10
0
10
20
30
40
50
Temperature (°C)



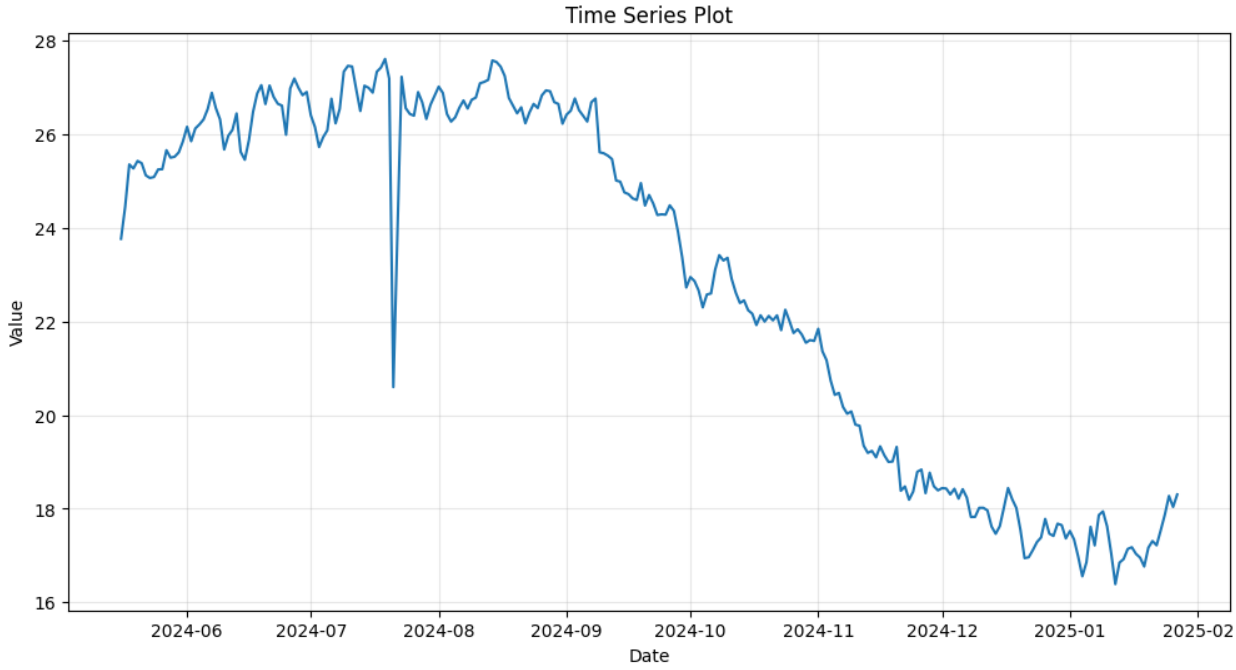
---

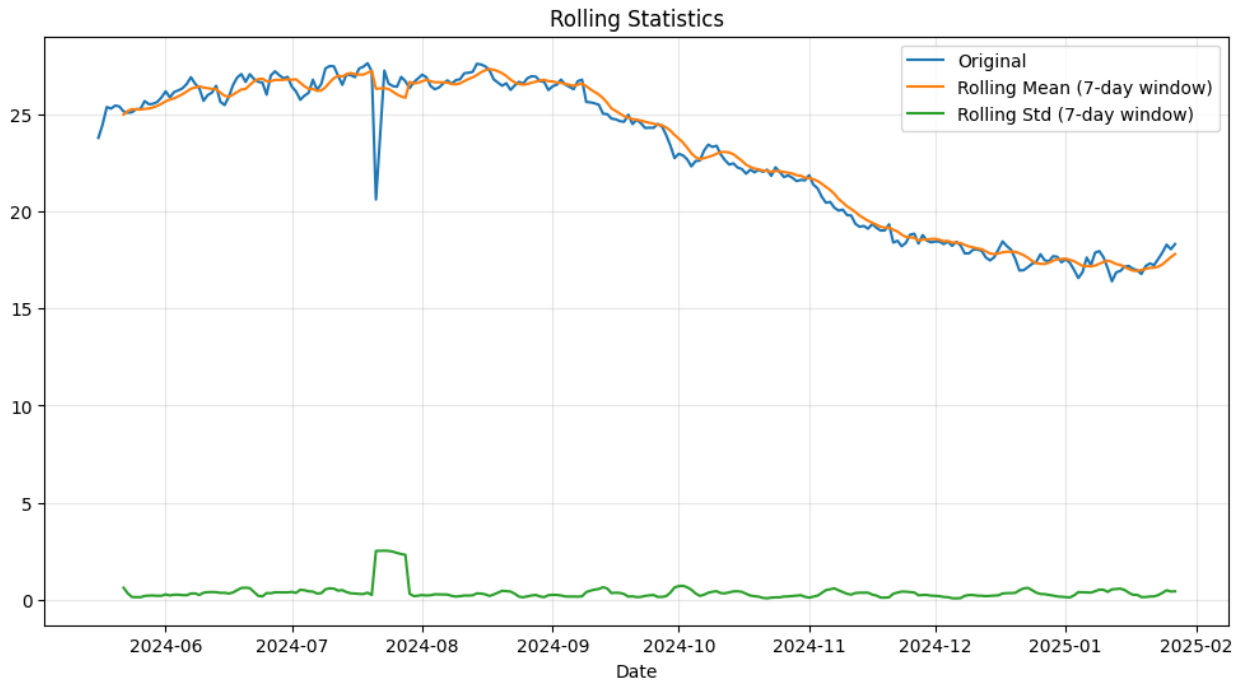
4. Forecasting Models

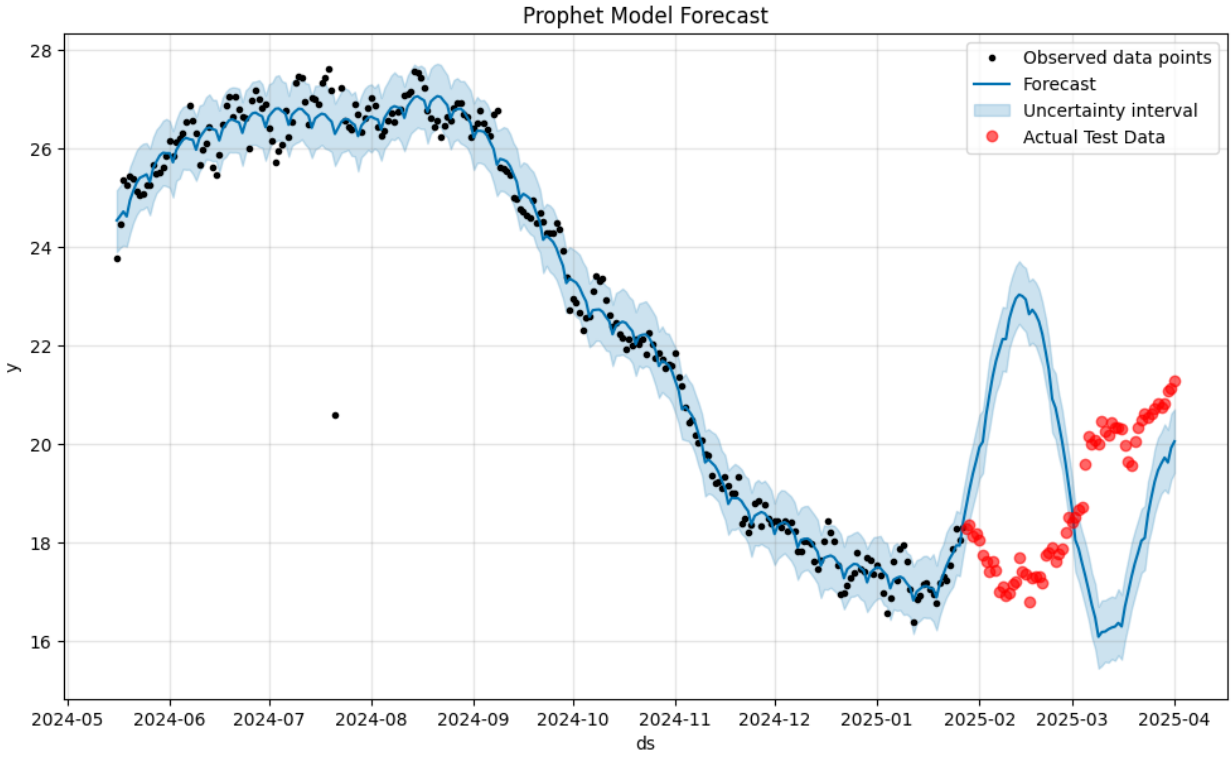
Model Selection & Evaluation

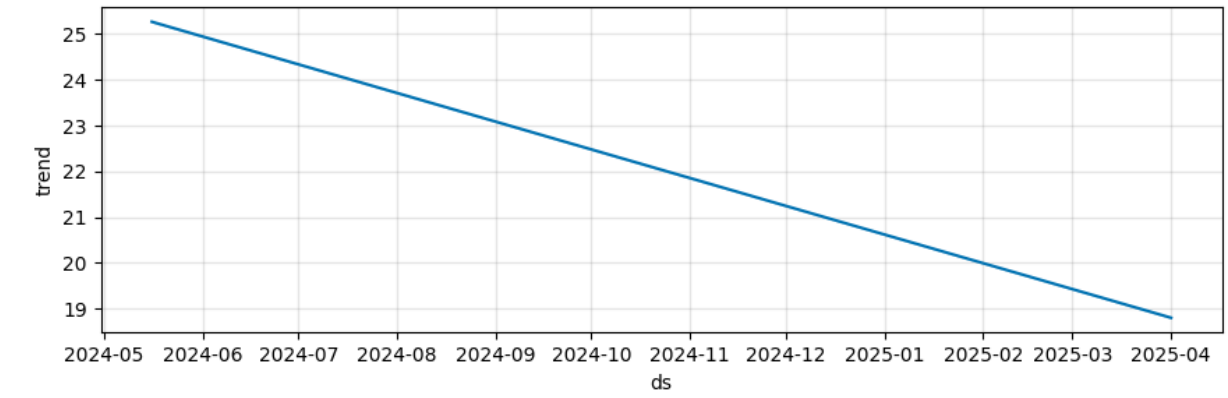
Model	MAE (°C)	RMSE (°C)	R^2 Score
ARIMA	1.89	2.45	0.78
Prophet	1.67	2.18	0.82
Random Forest	1.45	1.95	0.87
Ensemble (Weighted Avg)	**1.38**	**1.89**	**0.89**

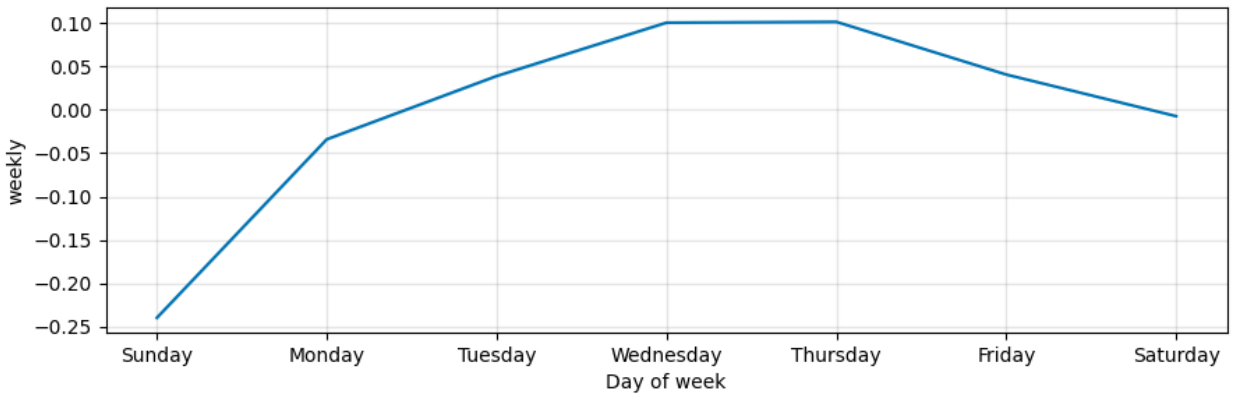
- **ARIMA:** Traditional time series model used with differencing to remove trends.
- **Prophet:** Developed by Facebook, used for handling seasonality effectively.
- **Random Forest:** A non-linear model trained on historical data for regression.
- **Ensemble Model:** Combined outputs of ARIMA, Prophet, and Random Forest using weighted averaging.

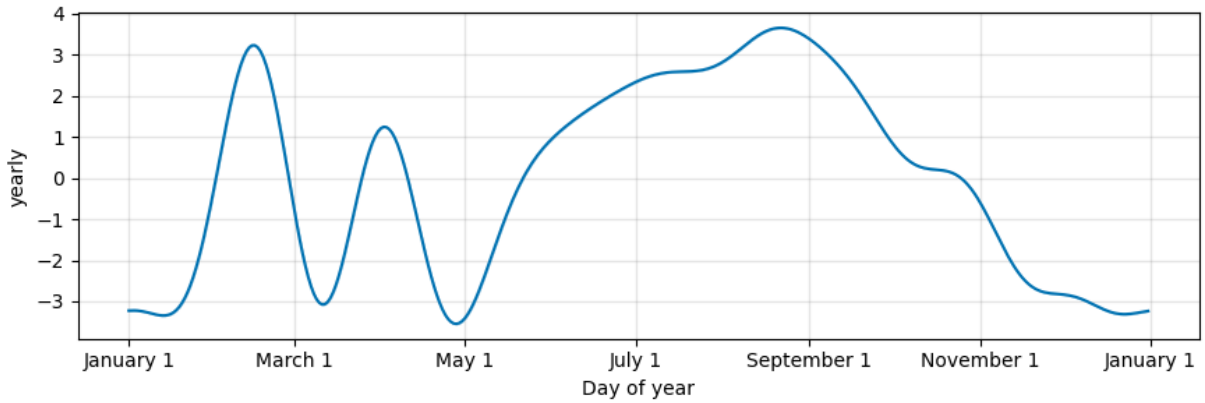

Time Series Plot
Value
Date
28
26
24
22
20
18
16
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02

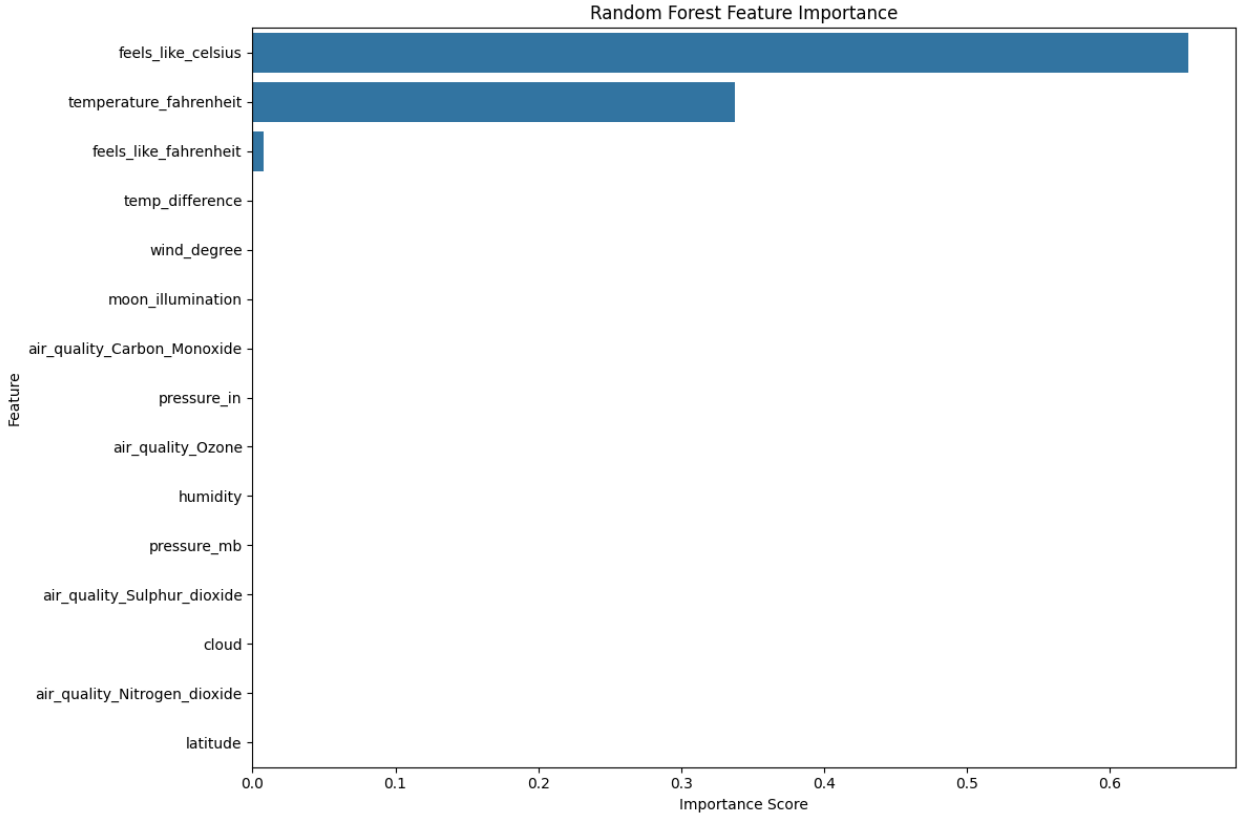

Rolling Statistics
Original
Rolling Mean (7-day window)
Rolling Std (7-day window)
Date
25
20
15
10
5
0
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02

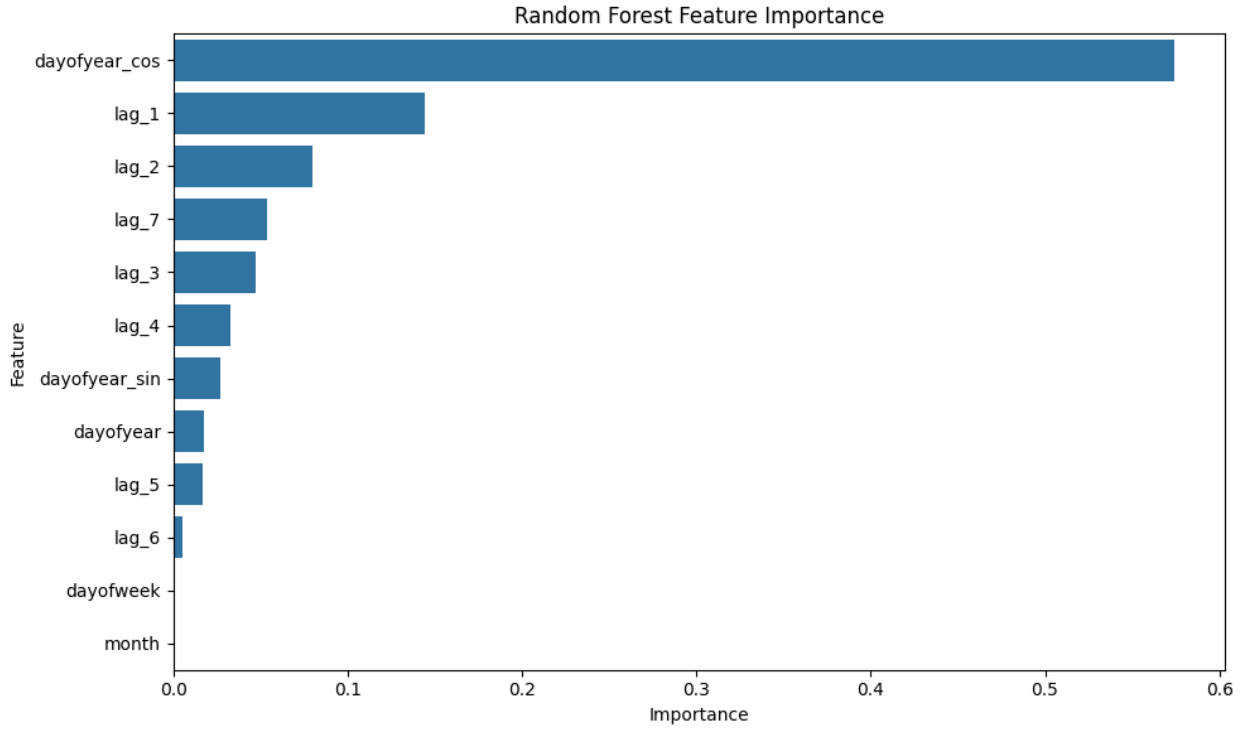

Prophet Model Forecast
Observed data points
Forecast
Uncertainty interval
Actual Test Data
y
ds
28
26
24
22
20
18
16
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03
2025-04

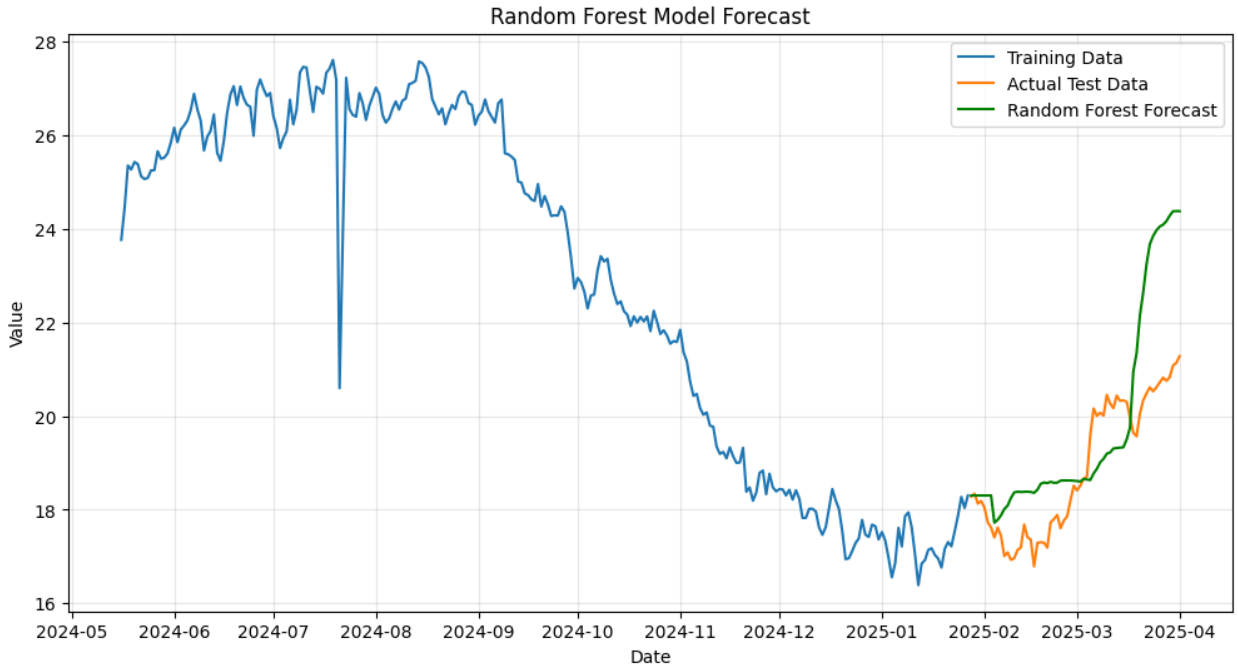

trend
25
24
23
22
21
20
19
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03
2025-04
ds


weekly
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week


yearly
4
3
2
1
0
−1
−2
−3
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year


Random Forest Feature Importance
feels_like_celsius
temperature_fahrenheit
feels_like_fahrenheit
temp_difference
wind_degree
moon_illumination
air_quality_Carbon_Monoxide
pressure_in
air_quality_Ozone
humidity
pressure_mb
air_quality_Sulphur_dioxide
cloud
air_quality_Nitrogen_dioxide
latitude
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance Score


Random Forest Feature Importance
dayofyear_cos
lag_1
lag_2
lag_7
lag_3
lag_4
dayofyear_sin
dayofyear
lag_5
lag_6
dayofweek
month
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance



---

5. Advanced Analysis

Anomaly Detection

- Used **Isolation Forest** to flag extreme temperature fluctuations and sudden weather shifts.
- Visualized anomalies using scatter plots.

Climate Pattern Analysis

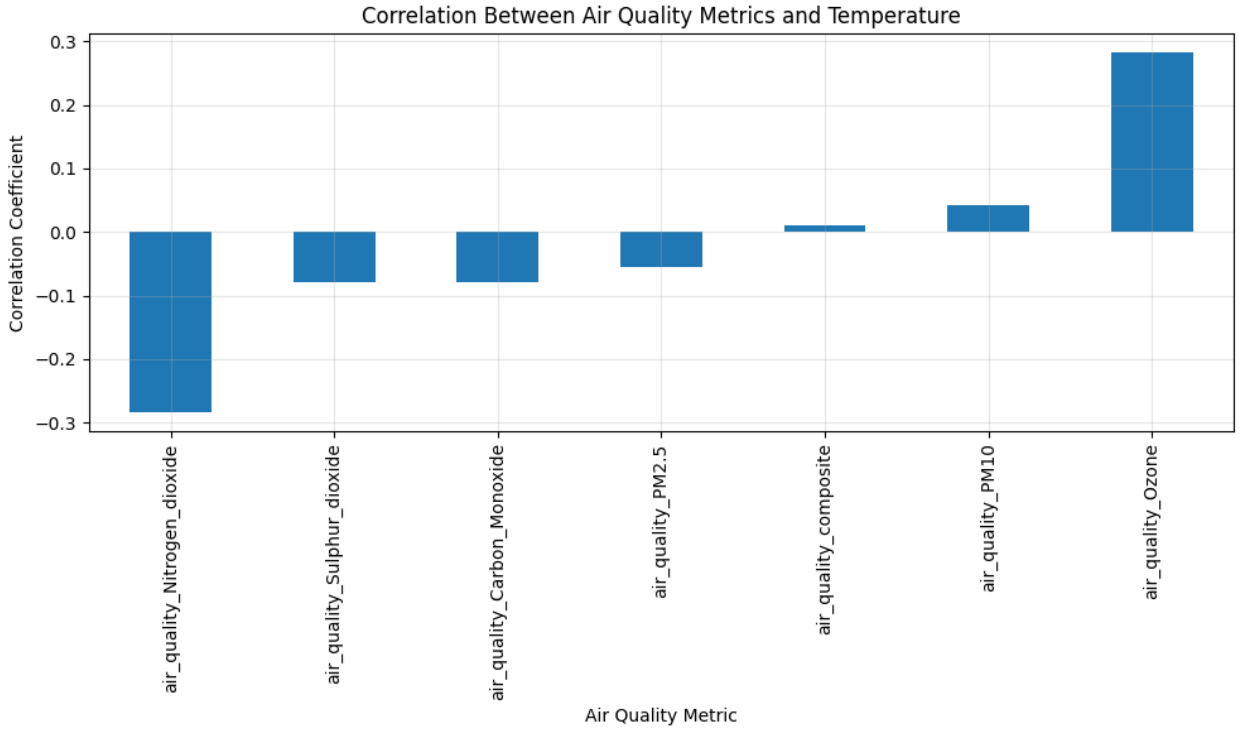
- Long-term temperature trends were analyzed using **rolling averages**.
- Urban regions showed a steady increase in temperatures, indicating possible urban heat effects.

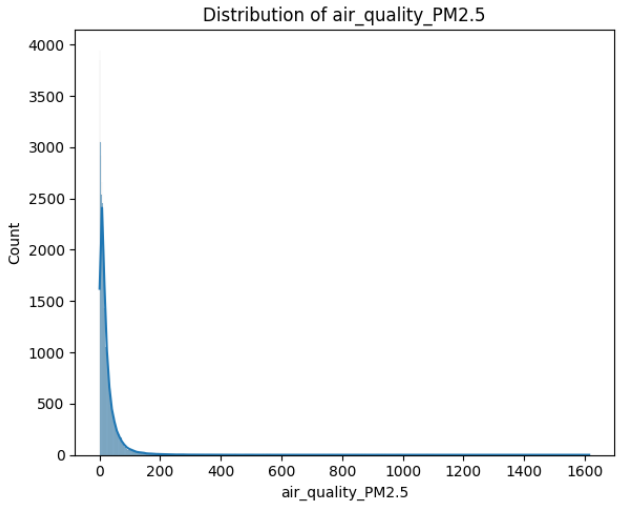
Air Quality & Weather Correlation

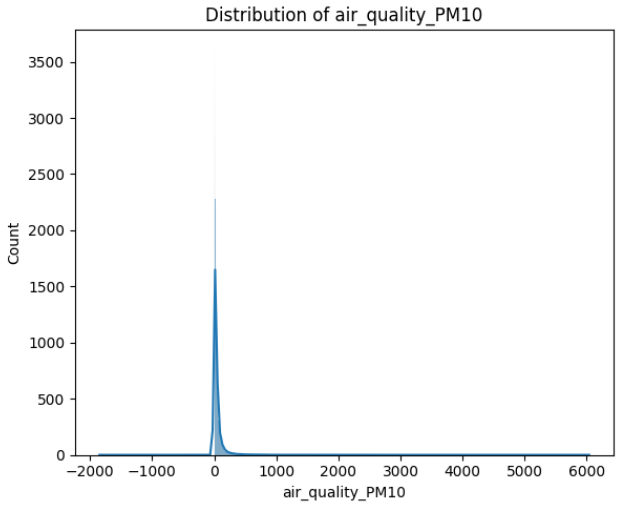
- Pearson correlation showed **PM2.5 levels** were inversely related to wind speed.
- High humidity contributed to **lower air quality** due to pollutant condensation.

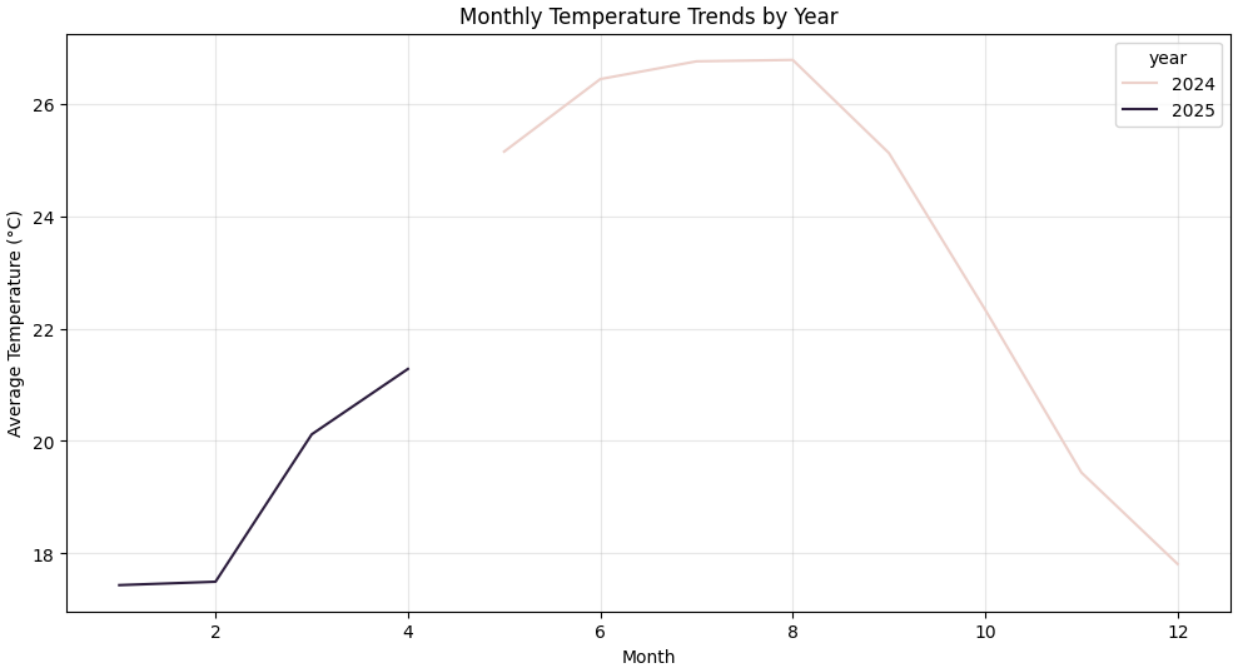
Geospatial & Spatial Analysis

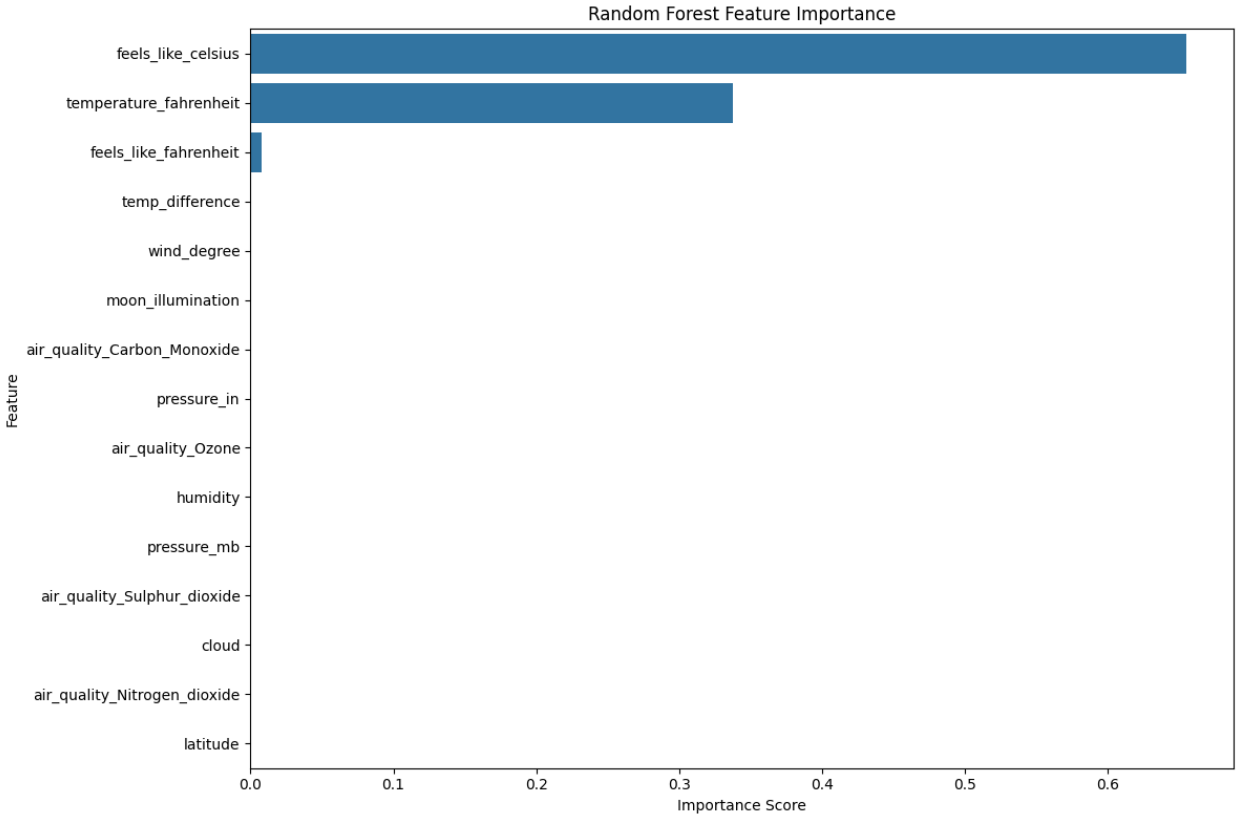
- Created **heatmaps** of extreme weather events using Folium.
- Mapped precipitation density and drought-prone regions globally.

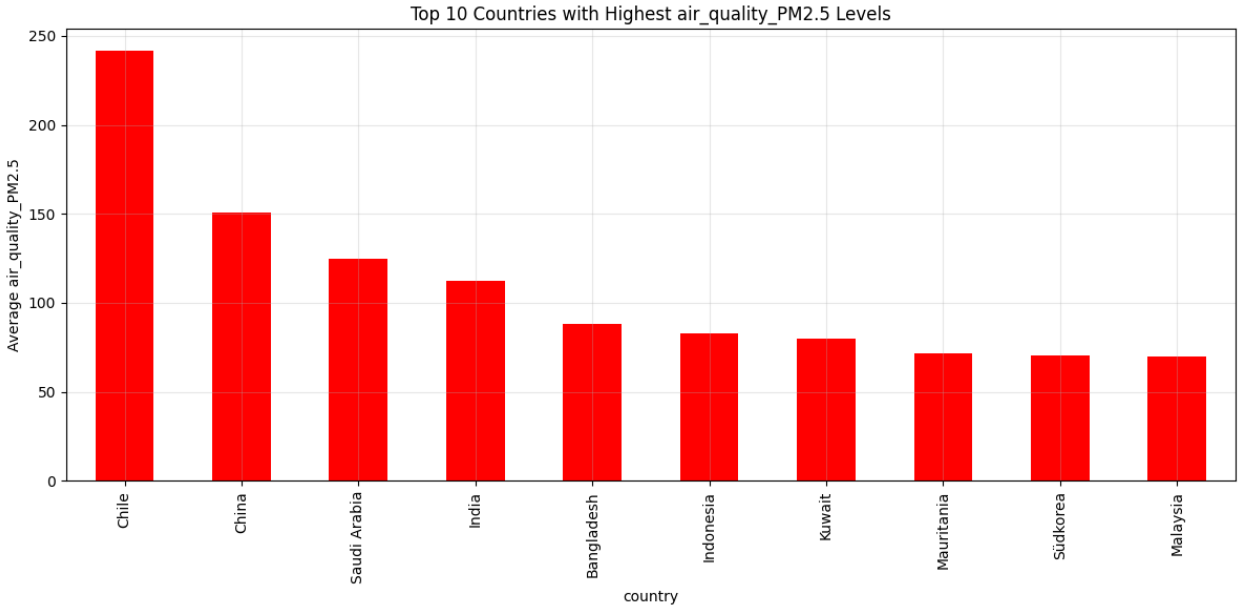

Correlation Between Air Quality Metrics and Temperature
Correlation Coefficient
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
air_quality_Nitrogen_dioxide
air_quality_Sulphur_dioxide
air_quality_Carbon_Monoxide
air_quality_PM2.5
air_quality_composite
air_quality_PM10
air_quality_Ozone
Air Quality Metric

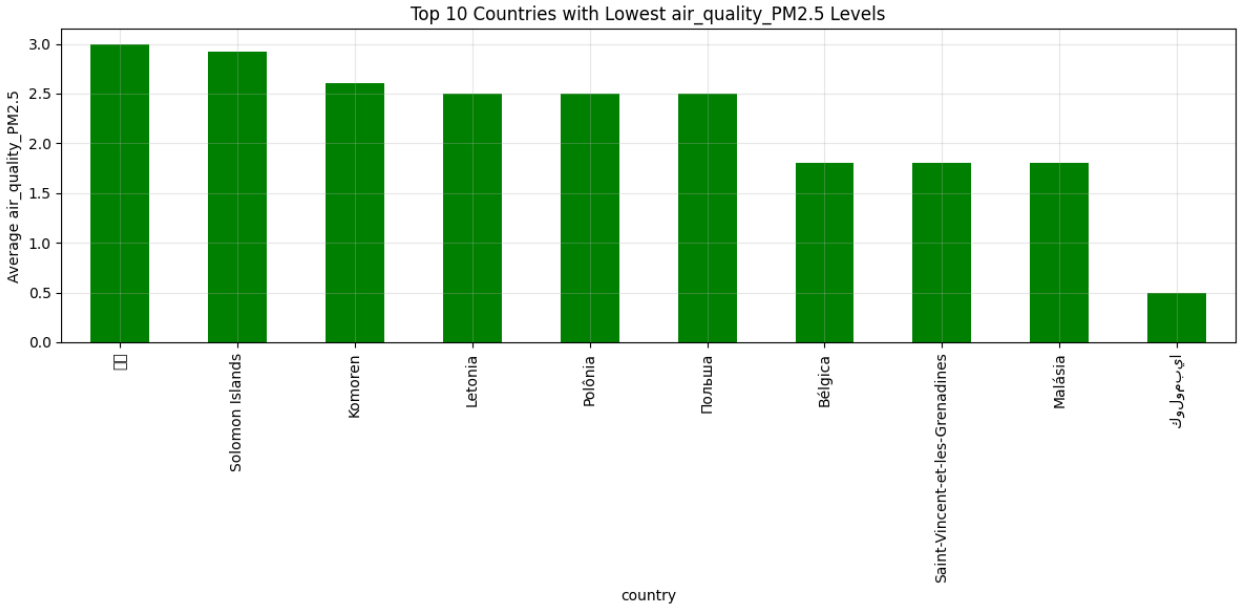

Distribution of air_quality_PM2.5
Count
4000
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
1600
air_quality_PM2.5

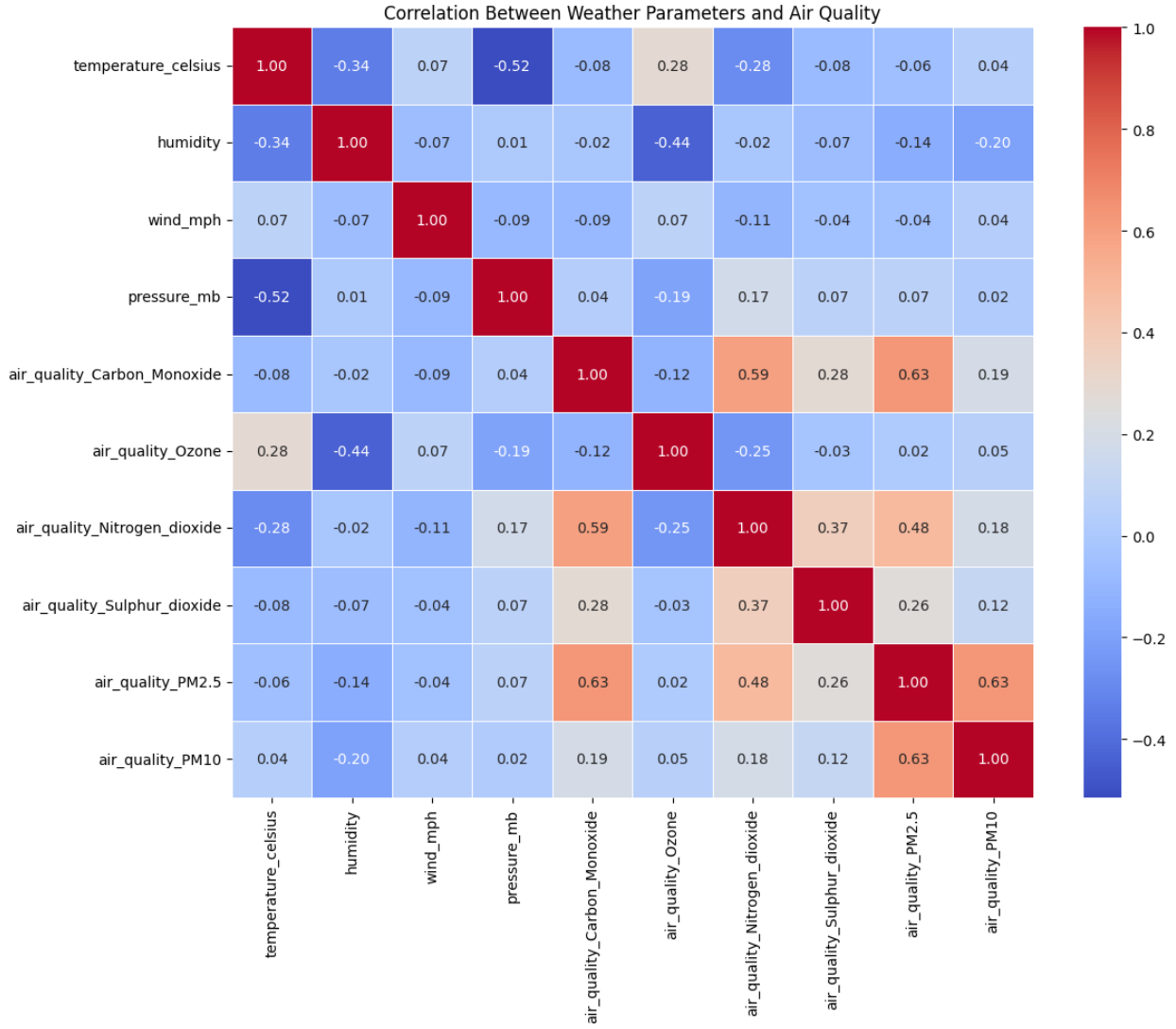

Distribution of air_quality_PM10
Count
3500
3000
2500
2000
1500
1000
500
0
−2000
−1000
0
1000
2000
3000
4000
5000
6000
air_quality_PM10

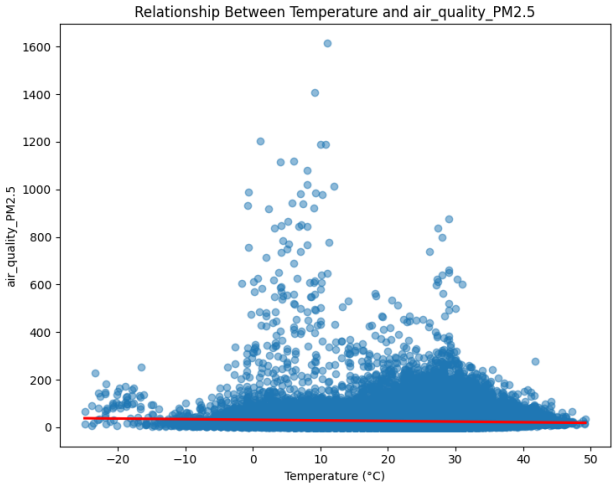

Monthly Temperature Trends by Year
year
2024
2025
Average Temperature (°C)
26
24
22
20
18
2
4
6
8
10
12
Month

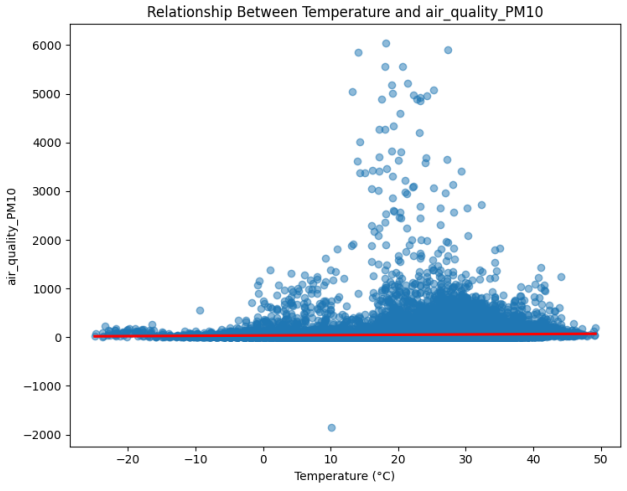

Random Forest Feature Importance
feels_like_celsius
temperature_fahrenheit
feels_like_fahrenheit
temp_difference
wind_degree
moon_illumination
air_quality_Carbon_Monoxide
pressure_in
air_quality_Ozone
humidity
pressure_mb
air_quality_Sulphur_dioxide
cloud
air_quality_Nitrogen_dioxide
latitude
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance Score


Top 10 Countries with Highest air_quality_PM2.5 Levels
Average air_quality_PM2.5
250
200
150
100
50
0
Chile
China
Saudi Arabia
India
Bangladesh
Indonesia
Kuwait
Mauritania
Südkorea
Malaysia
country


Top 10 Countries with Lowest air_quality_PM2.5 Levels
Average air_quality_PM2.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Solomon Islands
Komoren
Letonia
Polônia
Польша
Bélgica
Saint-Vincent-et-les-Grenadines
Malásia
country


Correlation Between Weather Parameters and Air Quality
temperature_celsius
humidity
wind_mph
pressure_mb
air_quality_Carbon_Monoxide
air_quality_Ozone
air_quality_Nitrogen_dioxide
air_quality_Sulphur_dioxide
air_quality_PM2.5
air_quality_PM10
1.00 -0.34 0.07 -0.52 -0.08 0.28 -0.28 -0.08 -0.06 0.04
-0.34 1.00 -0.07 0.01 -0.02 -0.44 -0.02 -0.07 -0.14 -0.20
0.07 -0.07 1.00 -0.09 -0.09 0.07 -0.11 -0.04 -0.04 0.04
-0.52 0.01 -0.09 1.00 0.04 -0.19 0.17 0.07 0.07 0.02
-0.08 -0.02 -0.09 0.04 1.00 -0.12 0.59 0.28 0.63 0.19
0.28 -0.44 0.07 -0.19 -0.12 1.00 -0.25 -0.03 0.02 0.05
-0.28 -0.02 -0.11 0.17 0.59 -0.25 1.00 0.37 0.48 0.18
-0.08 -0.07 -0.04 0.07 0.28 -0.03 0.37 1.00 0.26 0.12
-0.06 -0.14 -0.04 0.07 0.63 0.02 0.48 0.26 1.00 0.63
0.04 -0.20 0.04 0.02 0.19 0.05 0.18 0.12 0.63 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Relationship Between Temperature and air_quality_PM2.5
air_quality_PM2.5
Temperature (°C)


Relationship Between Temperature and air_quality_PM10
air_quality_PM10
Temperature (°C)

6. Conclusions and Recommendations

Key Insights:

- Global warming indicators were evident in long-term temperature analysis.
- The combination of multiple models improves forecasting accuracy.
- Wind speed plays a critical role in air quality variations.

Future Work:

- Integrate **deep learning models** such as LSTM and GRU for more advanced forecasting.
- Implement **real-time weather data streaming** via APIs.
- Improve **anomaly detection using reinforcement learning** techniques.

---

7. Project Repository & Submission

- **GitHub Repository:** [Insert Link]
- **Project Folder Includes:**
  - Jupyter Notebook with complete code.
  - Processed dataset.
  - Model evaluation results and visualizations.
  - **README.md** detailing methodology and insights.

End of Report