Data Science Assessment Report

Shaad Fazal

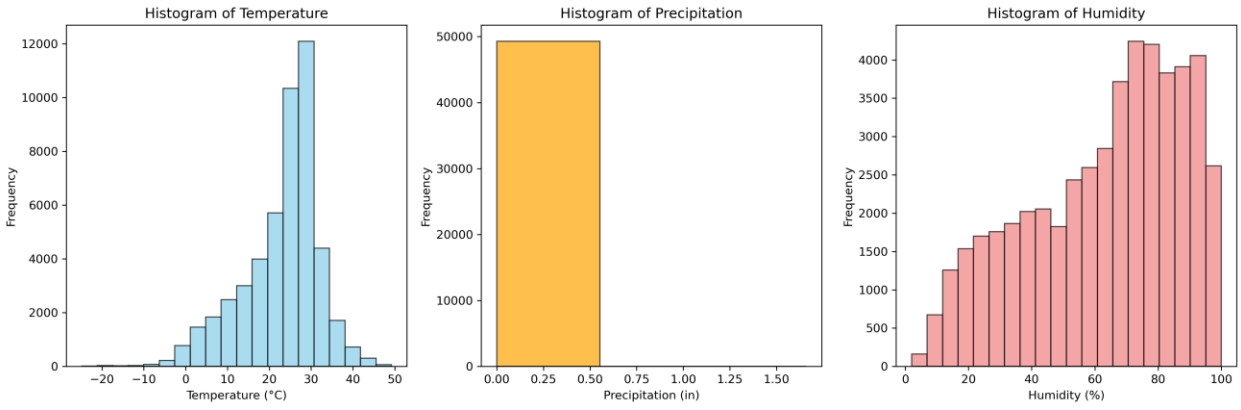
Email: shaad.fazal@gmail.com

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.

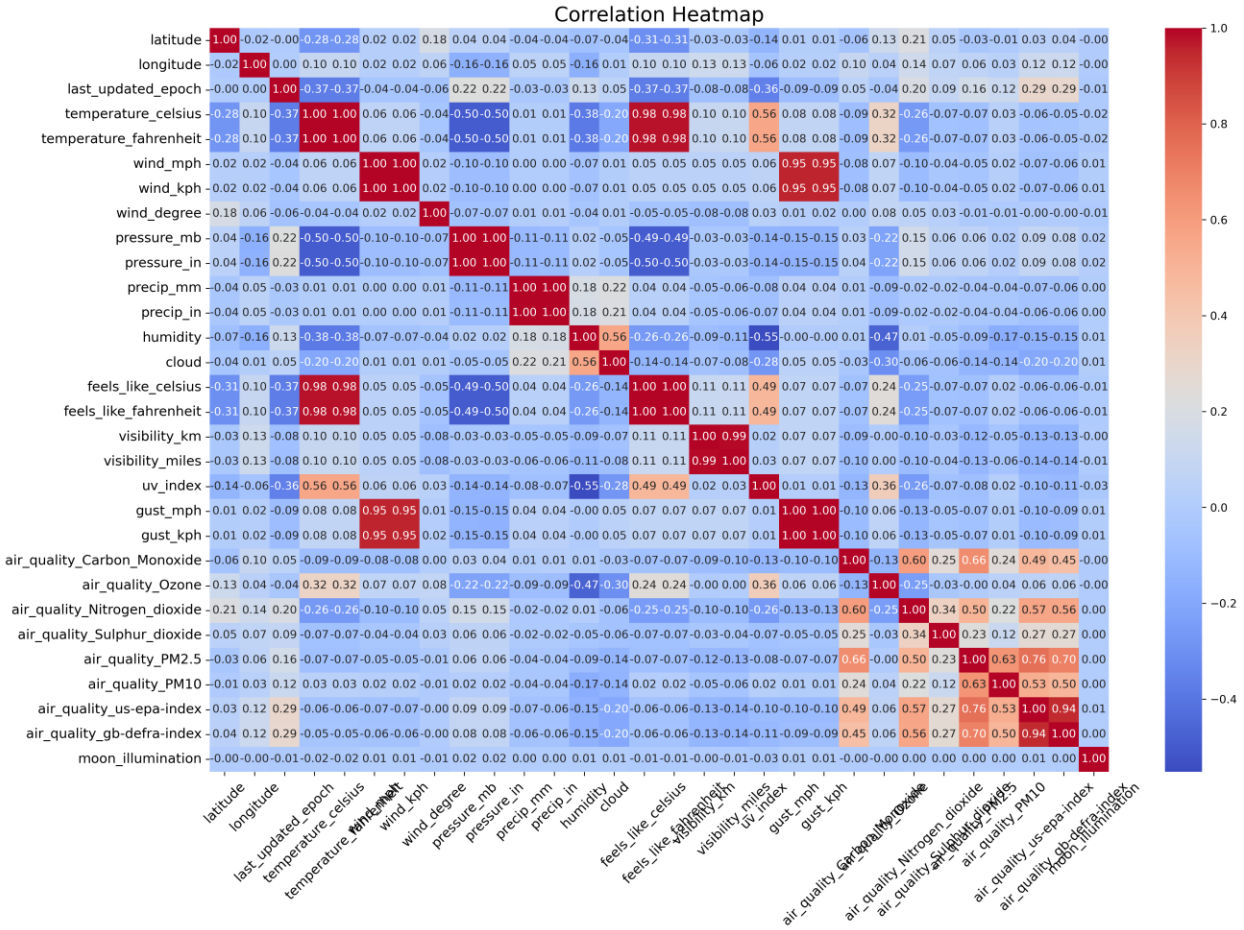
Data Cleaning and Processing: The dataset is checked for null or missing values which are found to be none.

Exploratory Data Analysis:

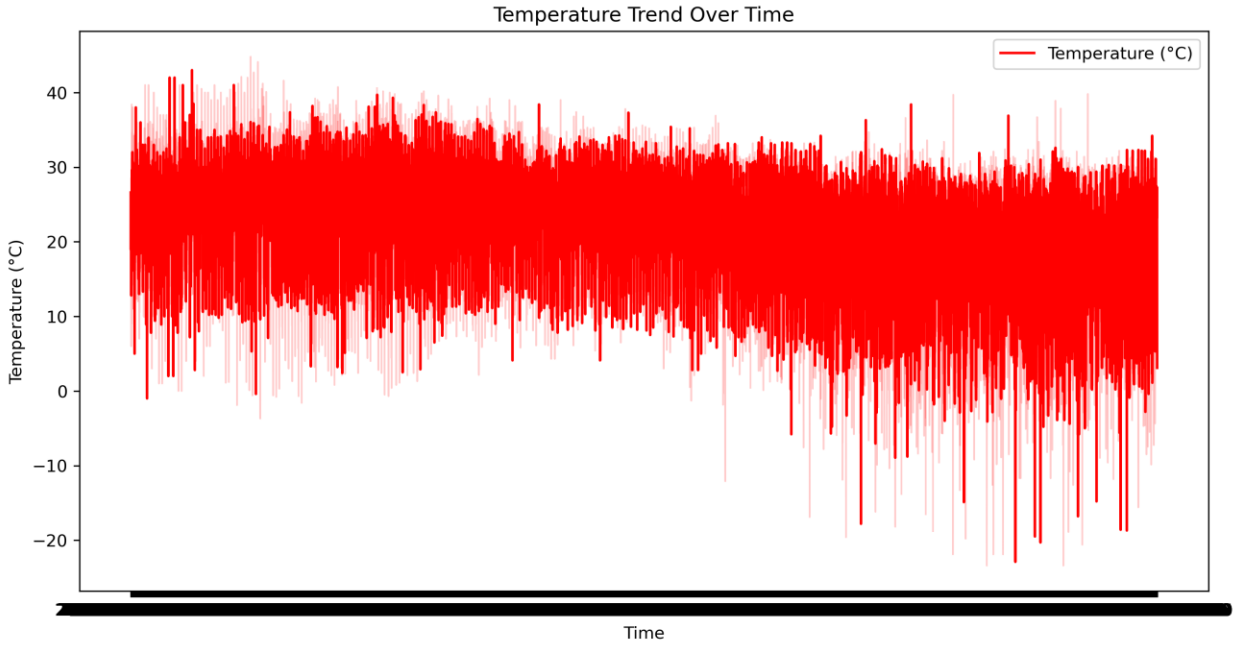
We generate a visual representation of the distributions of three environmental variables: temperature, precipitation, and humidity. It creates a single figure containing three histograms, each displaying the frequency distribution of one variable.

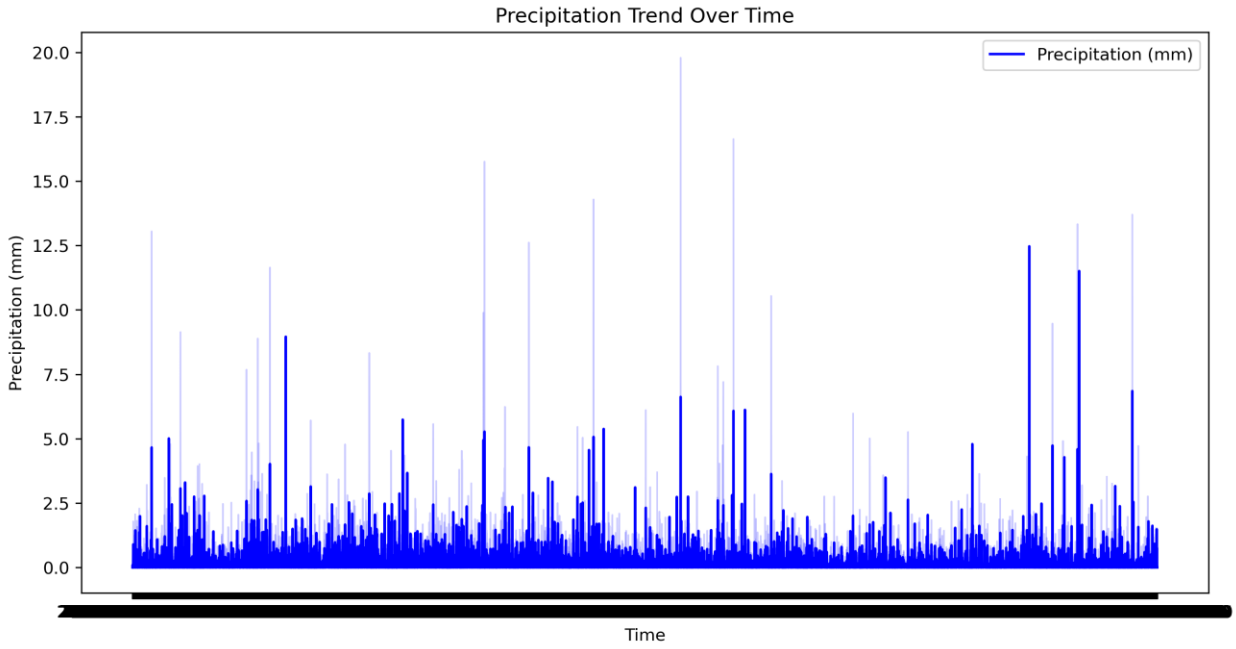


We also generate a **correlation heatmap** to visualize the relationships between numerical features in the dataset. This visualization helps identify strong positive or negative correlations between features, providing insights into relationships within the dataset.



We then generate **two line plots** to visualize trends in temperature and precipitation over time to create plot of time series visualization of these features.



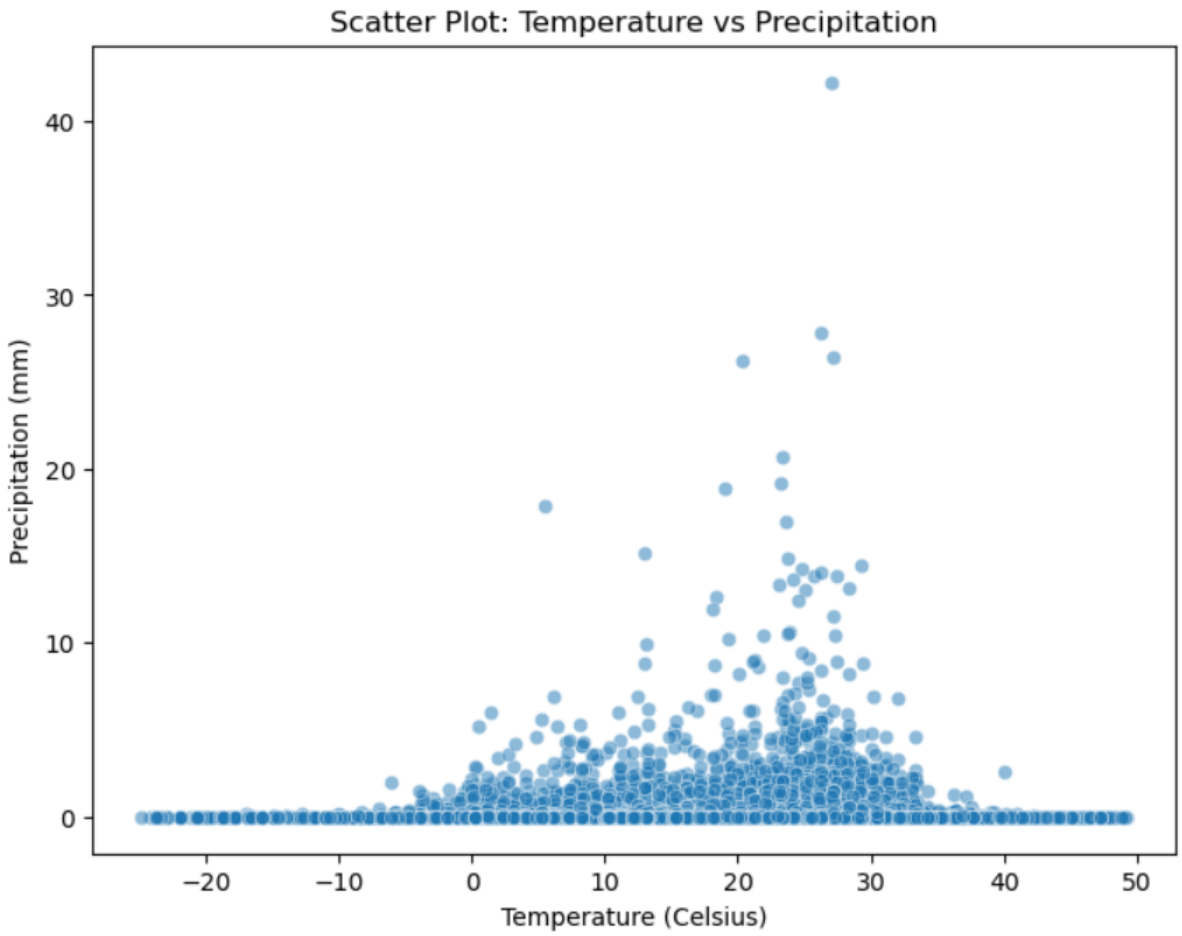


To get a clearer idea for feature correlation we compute for every feature the feature in dataset that possesses the highest correlation with it. Here are the results of it.

```
Feature: latitude | Highest Correlated Feature: air_quality_Nitrogen_dioxide |
Correlation: 0.21
Feature: longitude | Highest Correlated Feature: air_quality_Nitrogen_dioxide |
Correlation: 0.14
Feature: last_updated_epoch | Highest Correlated Feature: air_quality_us-epa-index |
Correlation: 0.29
Feature: temperature_celsius | Highest Correlated Feature: temperature_fahrenheit |
```

```
Correlation: 1.00
Feature: temperature_fahrenheit | Highest Correlated Feature: temperature_celsius |
Correlation: 1.00
Feature: wind_mph | Highest Correlated Feature: wind_kph | Correlation: 1.00
Feature: wind_kph | Highest Correlated Feature: wind_mph | Correlation: 1.00
Feature: wind_degree | Highest Correlated Feature: latitude | Correlation: 0.18
Feature: pressure_mb | Highest Correlated Feature: pressure_in | Correlation: 1.00
Feature: pressure_in | Highest Correlated Feature: pressure_mb | Correlation: 1.00
Feature: precip_mm | Highest Correlated Feature: precip_in | Correlation: 1.00
Feature: precip_in | Highest Correlated Feature: precip_mm | Correlation: 1.00
Feature: humidity | Highest Correlated Feature: cloud | Correlation: 0.56
Feature: cloud | Highest Correlated Feature: humidity | Correlation: 0.56
Feature: feels_like_celsius | Highest Correlated Feature: feels_like_fahrenheit |
Correlation: 1.00
Feature: feels_like_fahrenheit | Highest Correlated Feature: feels_like_celsius |
Correlation: 1.00
Feature: visibility_km | Highest Correlated Feature: visibility_miles | Correlation: 0.99
Feature: visibility_miles | Highest Correlated Feature: visibility_km | Correlation: 0.99
Feature: uv_index | Highest Correlated Feature: temperature_fahrenheit | Correlation:
0.56
Feature: gust_mph | Highest Correlated Feature: gust_kph | Correlation: 1.00
Feature: gust_kph | Highest Correlated Feature: gust_mph | Correlation: 1.00
Feature: air_quality_Carbon_Monoxide | Highest Correlated Feature: air_quality_PM2.5 |
Correlation: 0.66
Feature: air_quality_Ozone | Highest Correlated Feature: uv_index | Correlation: 0.36
Feature: air_quality_Nitrogen_dioxide | Highest Correlated Feature:
air_quality_Carbon_Monoxide | Correlation: 0.60
Feature: air_quality_Sulphur_dioxide | Highest Correlated Feature:
air_quality_Nitrogen_dioxide | Correlation: 0.34
Feature: air_quality_PM2.5 | Highest Correlated Feature: air_quality_us-epa-index |
Correlation: 0.76
Feature: air_quality_PM10 | Highest Correlated Feature: air_quality_PM2.5 | Correlation:
0.63
Feature: air_quality_us-epa-index | Highest Correlated Feature: air_quality_gb-defra-
index | Correlation: 0.94
Feature: air_quality_gb-defra-index | Highest Correlated Feature: air_quality_us-epa-
index | Correlation: 0.94
Feature: moon_illumination | Highest Correlated Feature: pressure_mb | Correlation: 0.02
```

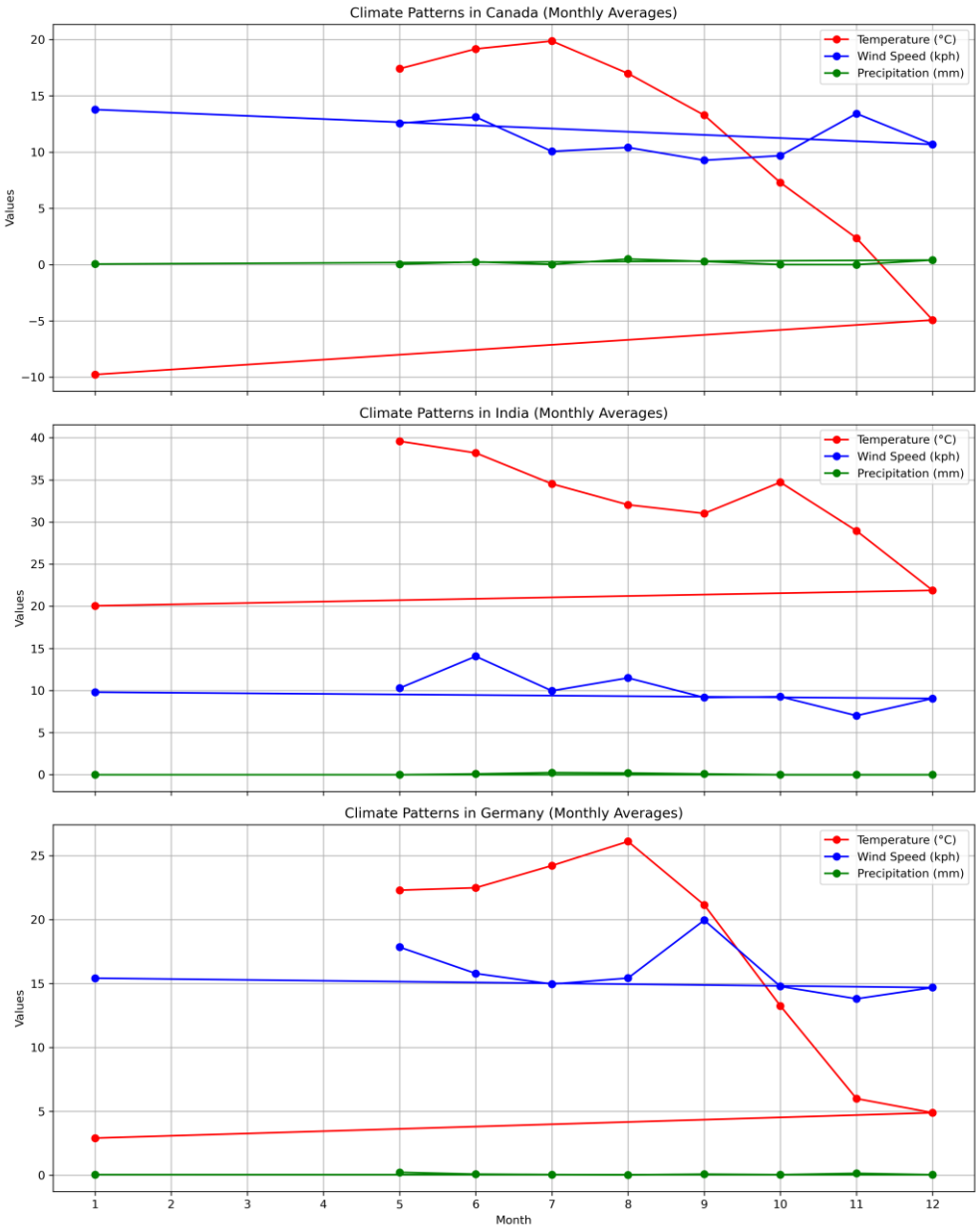
We also explore the relationship between temperature and precipitation using a scatter plot of their feature values which shows dependency to some extent though low correlation.



We also compute statistical metrics of the feature values to depict their distribution numerically.

- We perform a **long-term climate analysis** to study patterns and variations in temperature, wind speed, and precipitation across different regions (countries). Three countries (Canada, India, and Germany) are selected for analysis.
- For each country, line plots are generated to show monthly trends in temperature, wind speed, and precipitation.

- Each variable is represented by a distinct color (red for temperature, blue for wind speed, and green for precipitation) and marked with data points for clarity.



We also conduct an environmental impact analysis by examining air quality and its relationship with weather parameters. Here's a concise description of what it achieves and does:

1. Data Preparation

Relevant columns for air quality (e.g., Carbon Monoxide, Ozone, PM2.5, etc.) and weather (e.g., humidity, temperature, precipitation) are selected.

The last_updated column is converted to a datetime format, and the month is extracted for time-based analysis.

2. Correlation Analysis

A correlation matrix is created to analyze relationships between air quality and weather features.

A heatmap visualizes these correlations, with annotations showing correlation coefficients. This helps identify strong positive or negative relationships between variables.

3. Air Quality Analysis by Country

The average air quality metrics are calculated for each country, with a focus on PM2.5 levels.

A bar chart highlights the top 10 countries with the highest average PM2.5 levels, providing insights into regions with poor air quality.

4. Air Quality vs. Humidity

Scatter plots are created to explore the relationship between humidity and each air quality metric.

These plots help identify potential patterns or trends, such as whether higher humidity correlates with changes in air quality.

5. Monthly Trends in Air Quality and Humidity

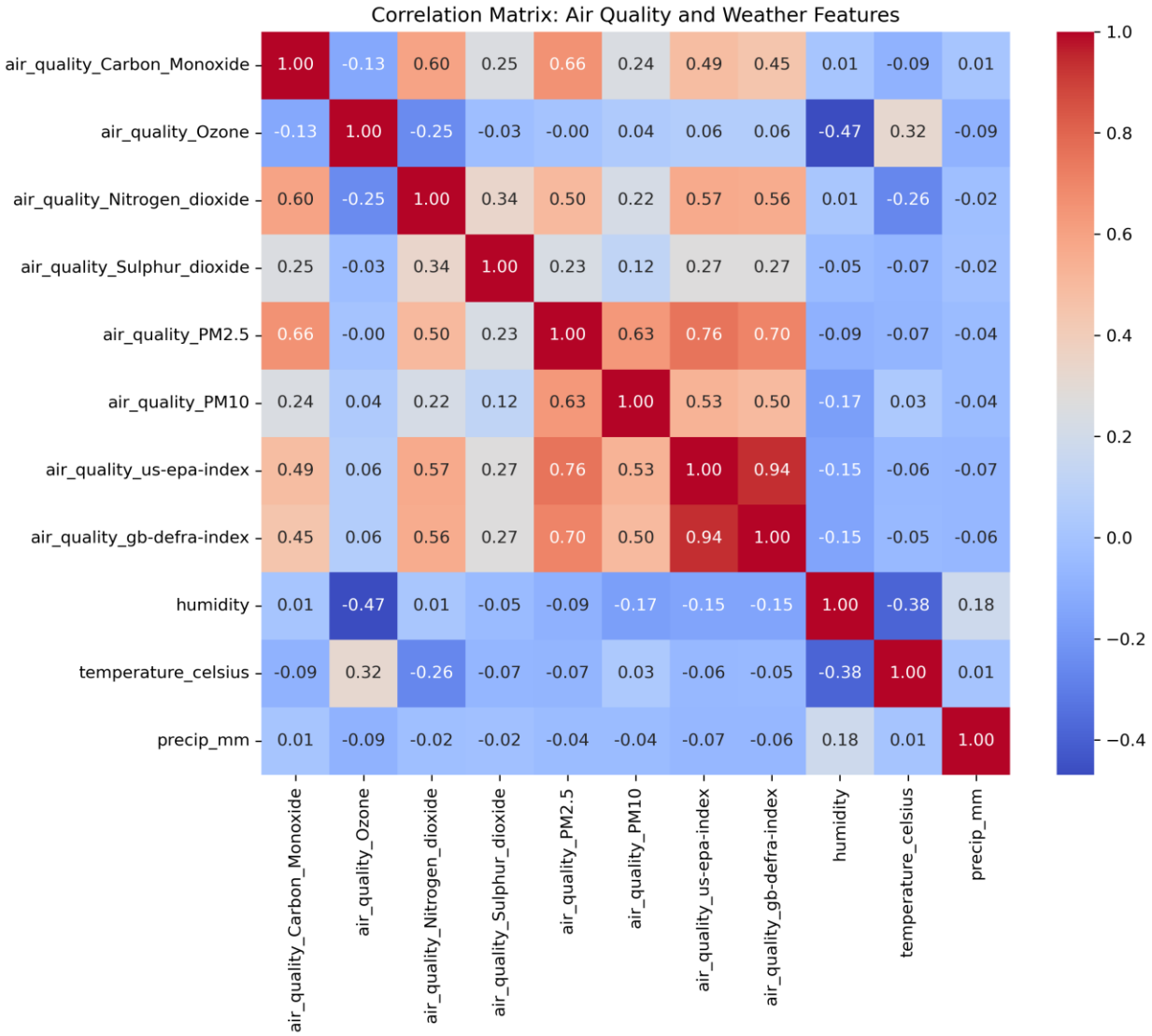
Monthly averages for air quality metrics and humidity are calculated to analyze seasonal trends.

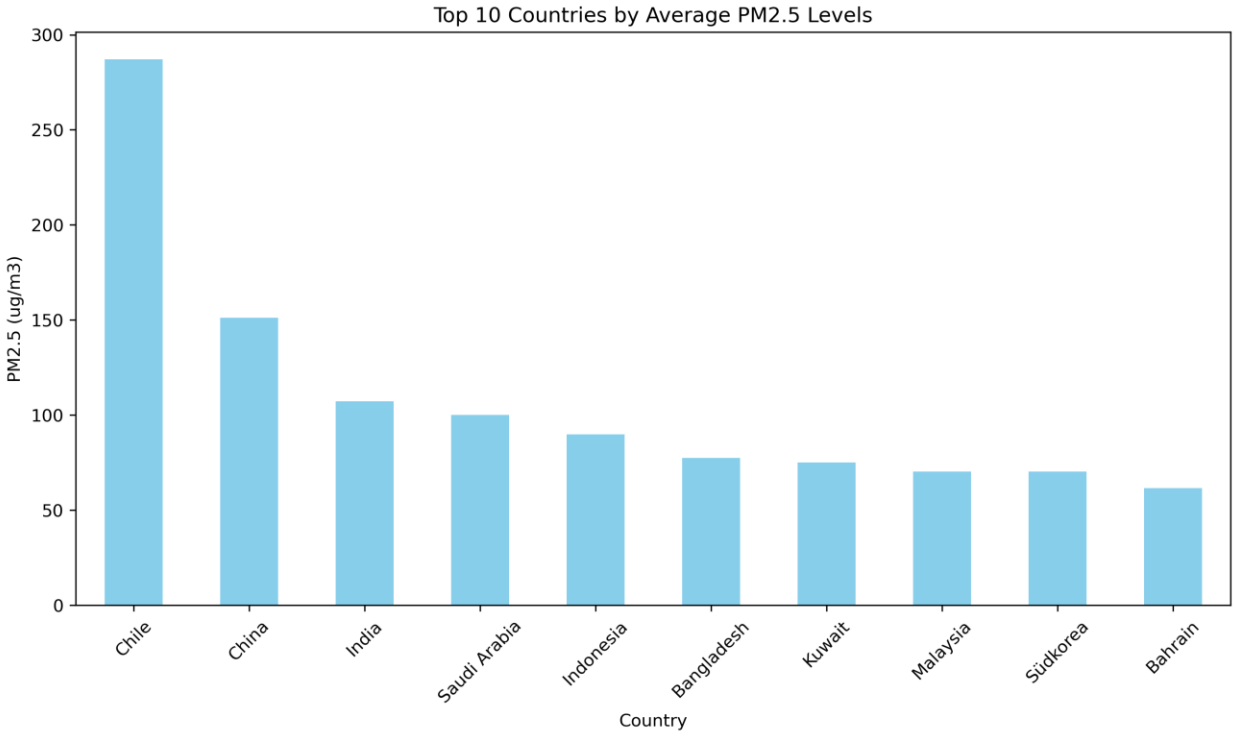
A line plot visualizes these trends, showing how air quality and humidity vary throughout the year.

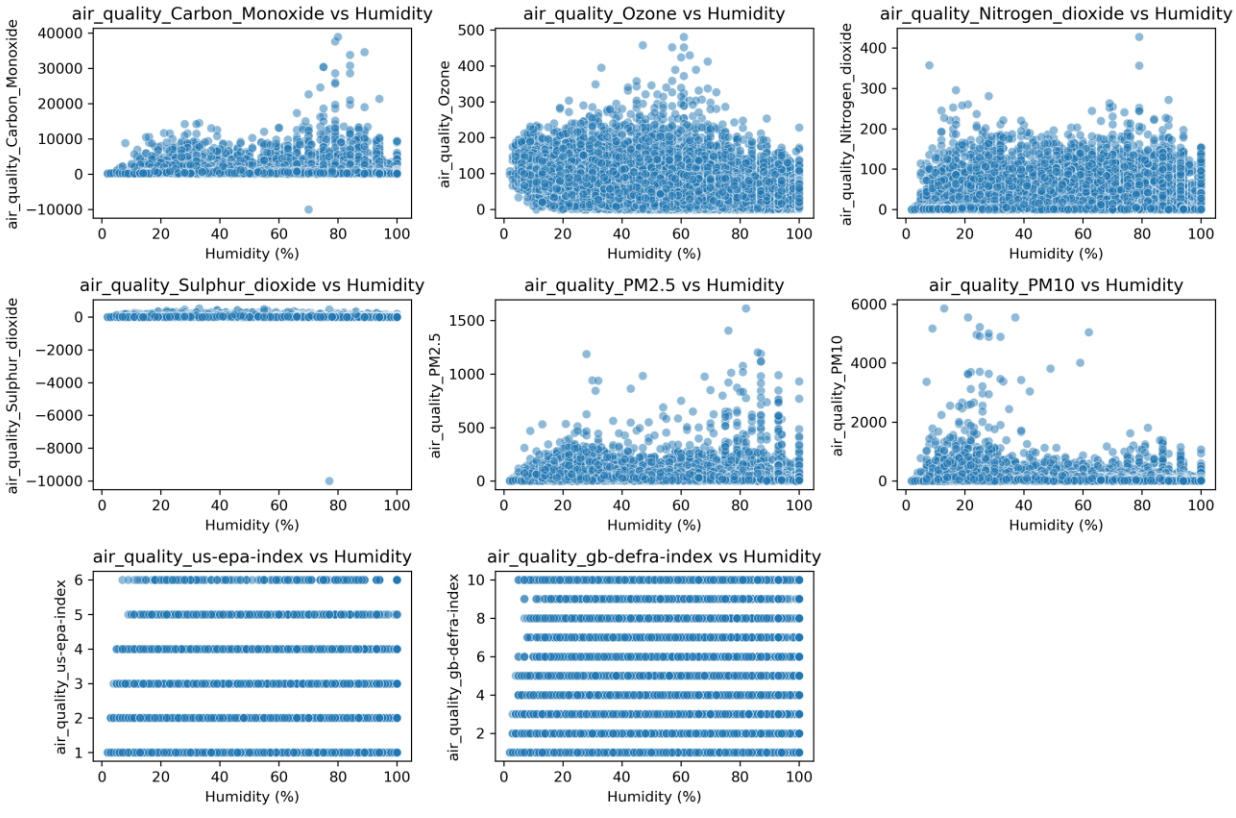
Purpose and Insights

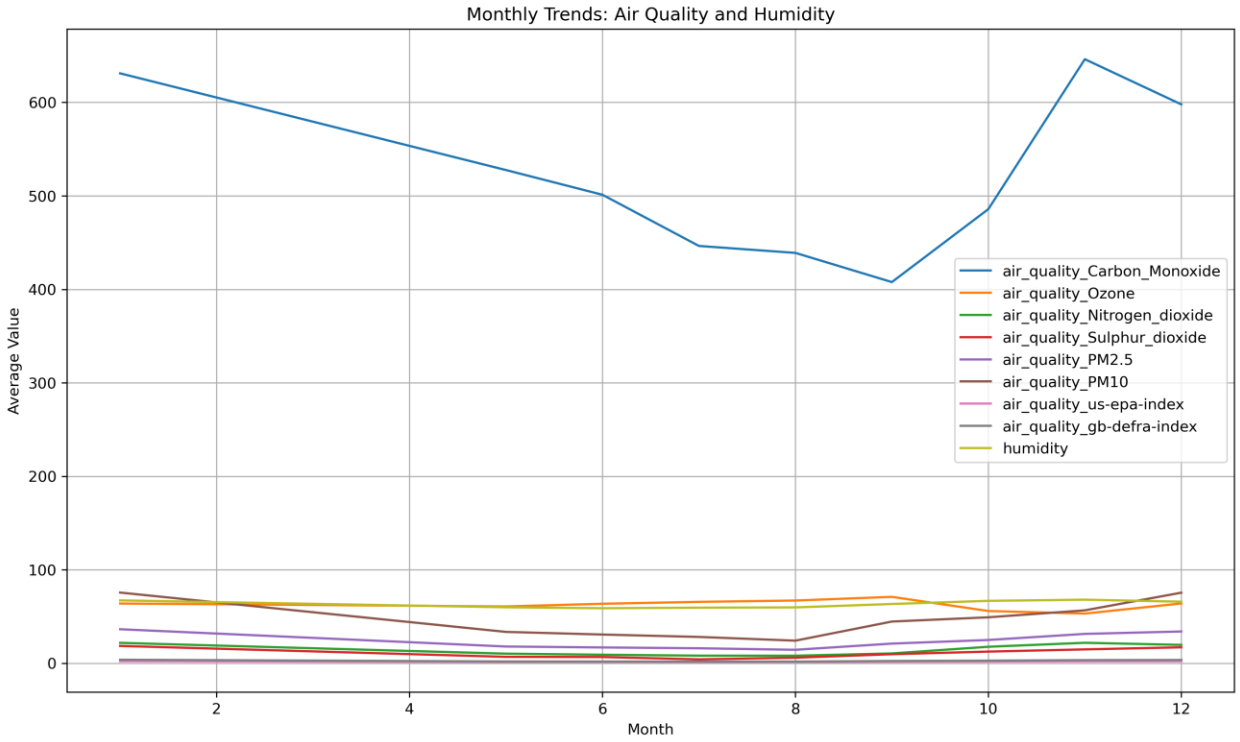
This analysis provides a comprehensive understanding of:

- How air quality metrics correlate with weather parameters.
- Which countries have the highest levels of air pollution (e.g., PM2.5)
- How air quality and humidity vary monthly, revealing potential seasonal patterns.



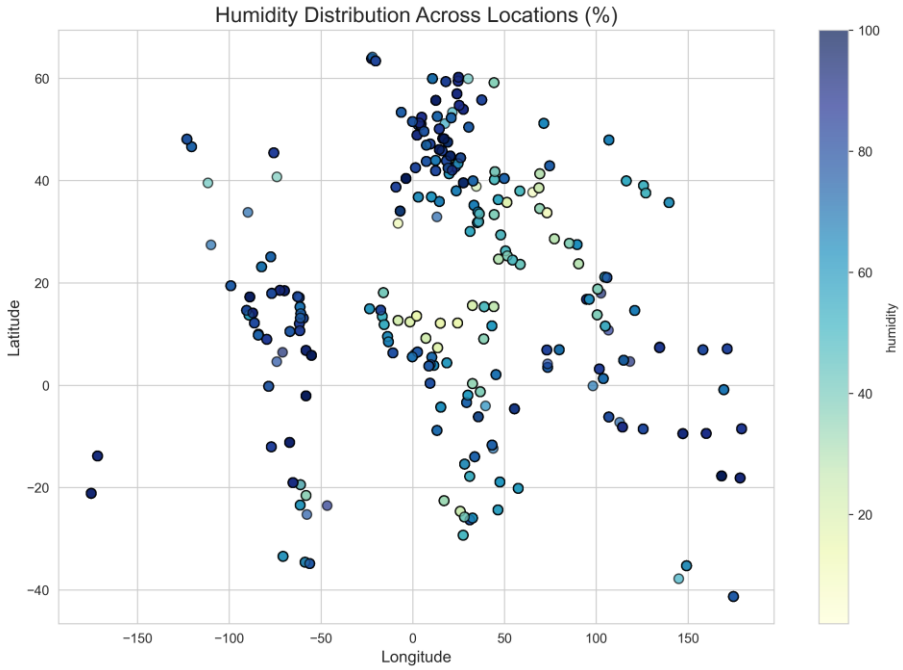

Top 10 Countries by Average PM2.5 Levels
PM2.5 (ug/m3)
Country
Chile
China
India
Saudi Arabia
Indonesia
Bangladesh
Kuwait
Malaysia
Südkorea
Bahrain

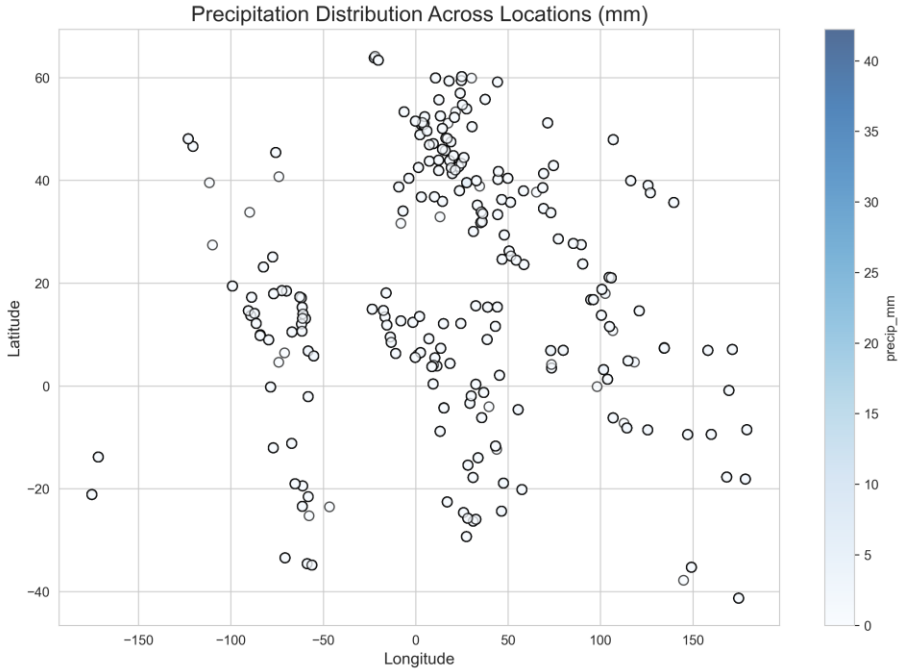

air_quality_Carbon_Monoxide vs Humidity
air_quality_Carbon_Monoxide
Humidity (%)
air_quality_Ozone vs Humidity
air_quality_Ozone
Humidity (%)
air_quality_Nitrogen_dioxide vs Humidity
air_quality_Nitrogen_dioxide
Humidity (%)
air_quality_Sulphur_dioxide vs Humidity
air_quality_Sulphur_dioxide
Humidity (%)
air_quality_PM2.5 vs Humidity
air_quality_PM2.5
Humidity (%)
air_quality_PM10 vs Humidity
air_quality_PM10
Humidity (%)
air_quality_us-epa-index vs Humidity
air_quality_us-epa-index
Humidity (%)
air_quality_gb-defra-index vs Humidity
air_quality_gb-defra-index
Humidity (%)

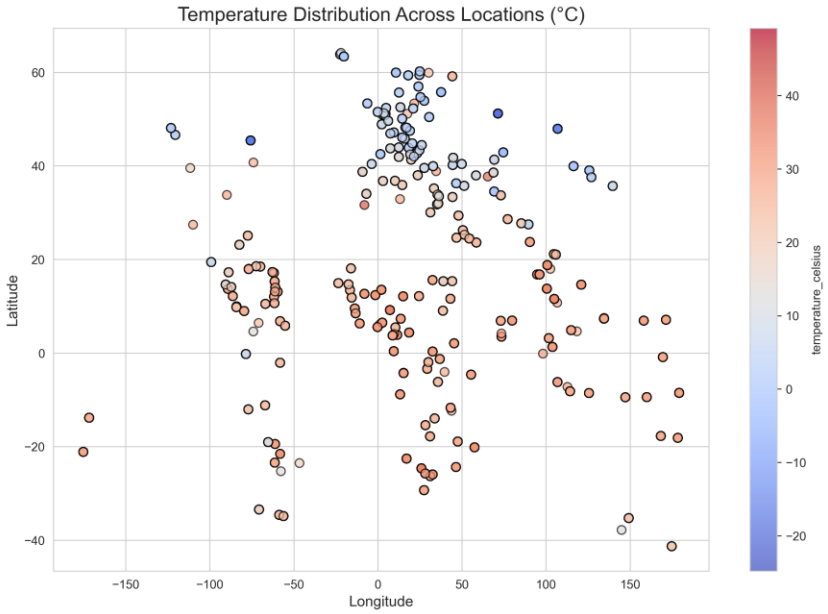


We also performs a **spatial analysis** to visualize geographical patterns in the data across different locations. The function creates scatter plots using latitude and longitude as the x and y axes, with the selected feature (e.g., temperature, humidity, precipitation) represented by color intensity.This spatial analysis helps identify geographical patterns and trends in environmental variables such as temperature, humidity, and precipitation. By visualizing these patterns on a map, it becomes easier to:

- Identify regions with extreme or unusual conditions.
- Compare environmental conditions across different locations.
- Gain insights into how geographical factors influence climate and weather.


Humidity Distribution Across Locations (%)
Latitude
Longitude
humidity
60
40
20
0
−20
−40
−150
−100
−50
0
50
100
150
100
80
60
40
20


Precipitation Distribution Across Locations (mm)
Latitude
Longitude
precip_mm
60
40
20
0
−20
−40
−150
−100
−50
0
50
100
150
40
35
30
25
20
15
10
5
0



Then we perform analysis focusing on understanding geographical patterns and variations in weather conditions across selected countries as well as various continents. By examining key weather metrics—temperature, humidity, precipitation, and wind speed—it provides qualitative insights into how these factors differ across regions and what these differences might imply. Here's a breakdown of the purpose and insights. The analysis compares weather conditions (temperature, humidity, precipitation, and wind speed) across eight countries, providing a clear picture of how these variables vary geographically. It helps identify regions with extreme or unique weather patterns.

Trend Identification: By aggregating and visualizing data, the analysis reveals trends in average weather conditions, such as which countries are hotter, more humid, or experience higher precipitation.

Temperature Patterns: The bar plot and box plot for temperature reveal which countries have the highest and lowest average temperatures.

For example, countries like India and Nigeria might show higher average temperatures, while Canada and England might have lower averages.

The box plot provides additional insights into the variability of temperatures within each country, showing whether temperatures are consistent or fluctuate significantly.

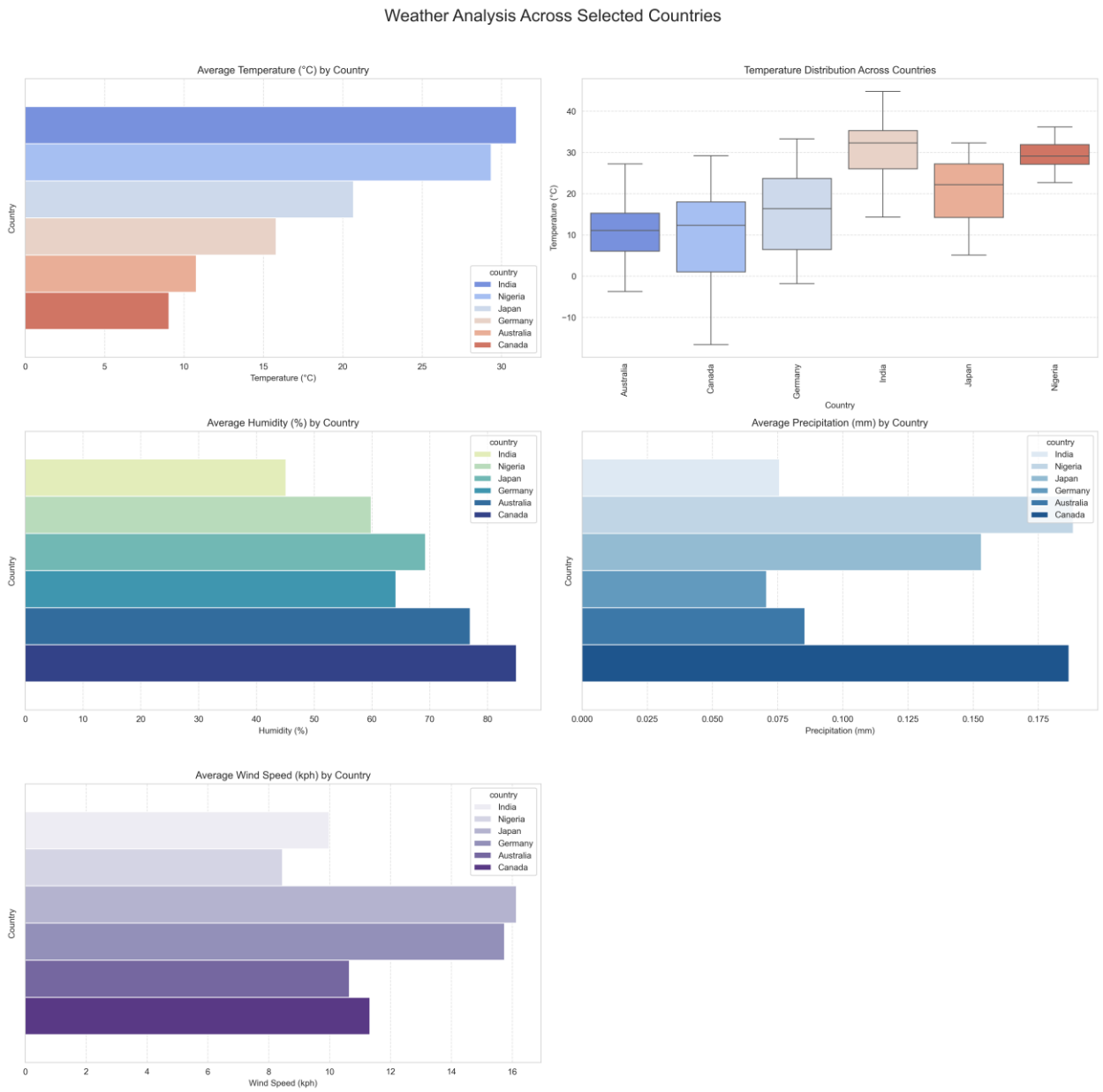
Humidity Levels: The humidity bar plot highlights countries with higher or lower average humidity.

Tropical or coastal countries (e.g., India, Nigeria) might show higher humidity levels, while arid or temperate regions (e.g., Australia, USA) might have lower humidity.

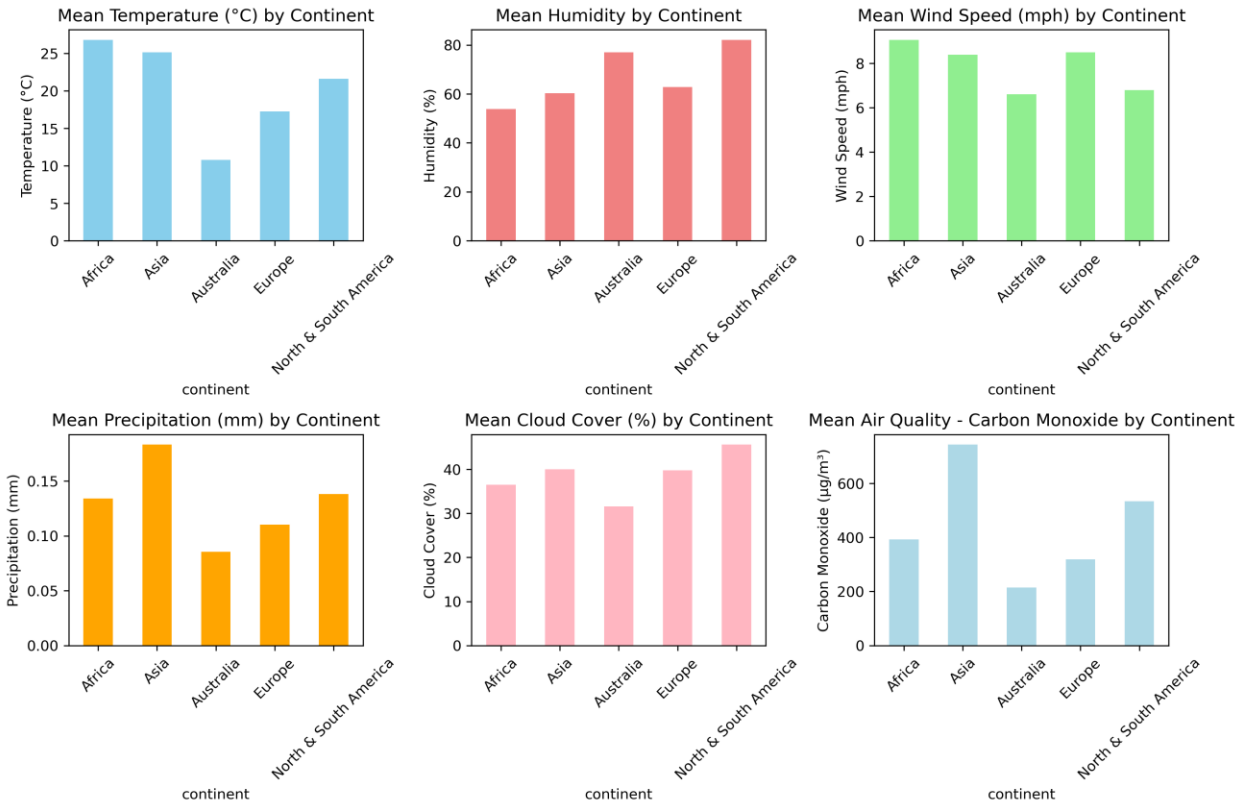
Precipitation Trends: The precipitation bar plot shows which countries receive the most and least rainfall on average.

Countries with monsoon climates (e.g., India) or those near large water bodies might have higher precipitation, while arid regions (e.g., Australia) might have lower levels.

Wind Speed Variations: The wind speed bar plot identifies countries with higher or lower average wind speeds.



This continental-level analysis provides a clear and comprehensive view of how weather and air quality vary across different regions of the world. By visualizing these differences, it offers valuable insights into regional climate patterns and environmental conditions, making it a useful tool for researchers, policymakers, and other stakeholders. The visualizations are designed to be intuitive and accessible, ensuring that the insights can be easily communicated and acted upon.



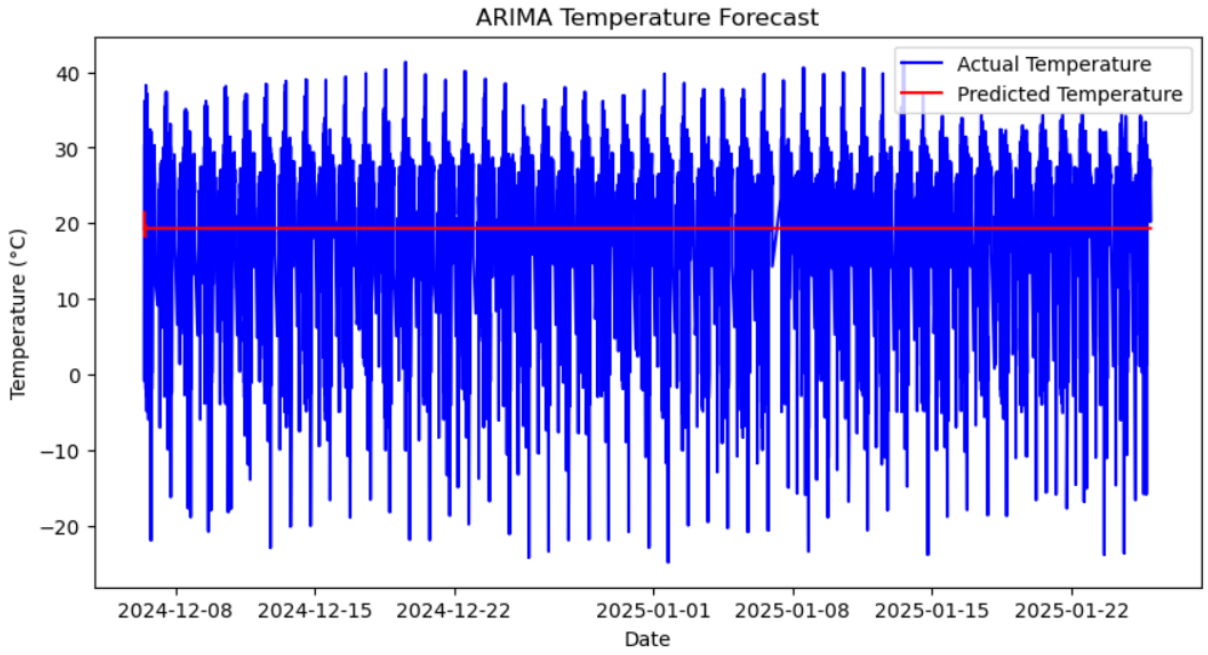
We then move on to forecasting models and their performance on the test dataset which is formed by splitting the original dataset. We present the quantitative and qualitative results for the forecasting models.

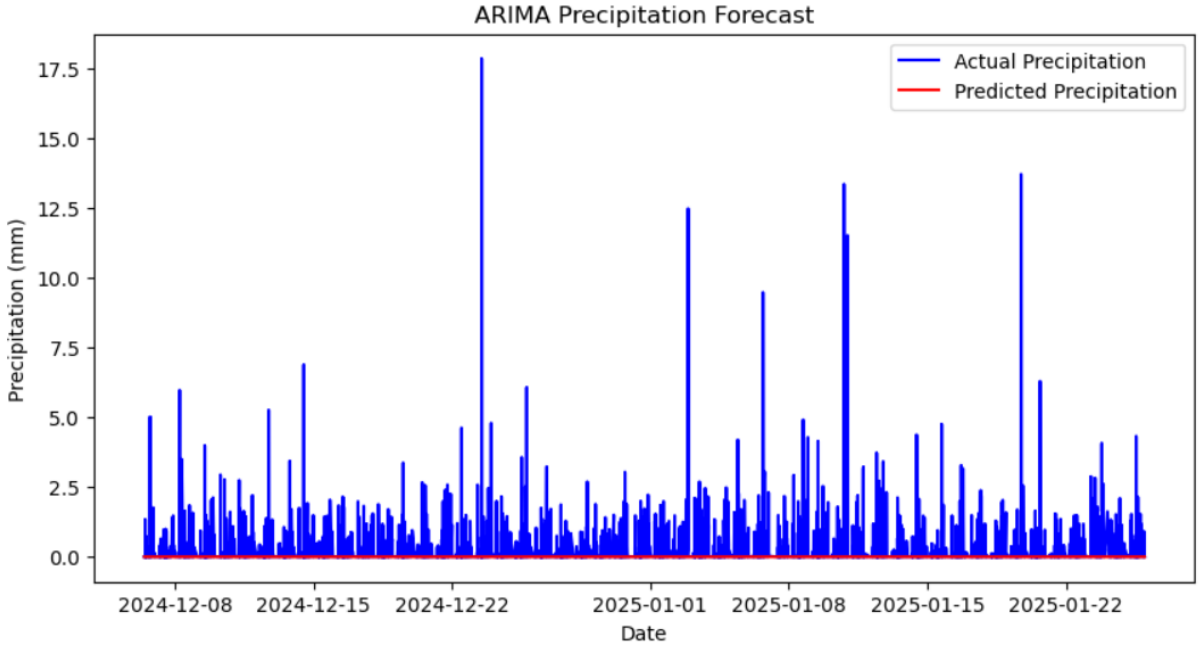
1. ARIMA (AutoRegressive Integrated Moving Average):

   ARIMA is a statistical model used for time series forecasting. It combines three components:
   AutoRegressive (AR): Models the relationship between an observation and its lagged values.
   Integrated (I): Differencing the data to make it stationary.
   Moving Average (MA): Models the relationship between an observation and residual errors from a moving average.

   ```
   For Temperature Forecasting:

   Mean Absolute Error (MAE): 9.2933
   Mean Squared Error (MSE): 125.4819
   Root Mean Squared Error (RMSE): 11.2019
   R-squared (R²): -0.0309
   ```

For Precipitation forecasting:
Mean Absolute Error (MAE): 0.1194
Mean Squared Error (MSE): 0.2807
Root Mean Squared Error (RMSE): 0.5298
R-squared (R^2): -0.0535



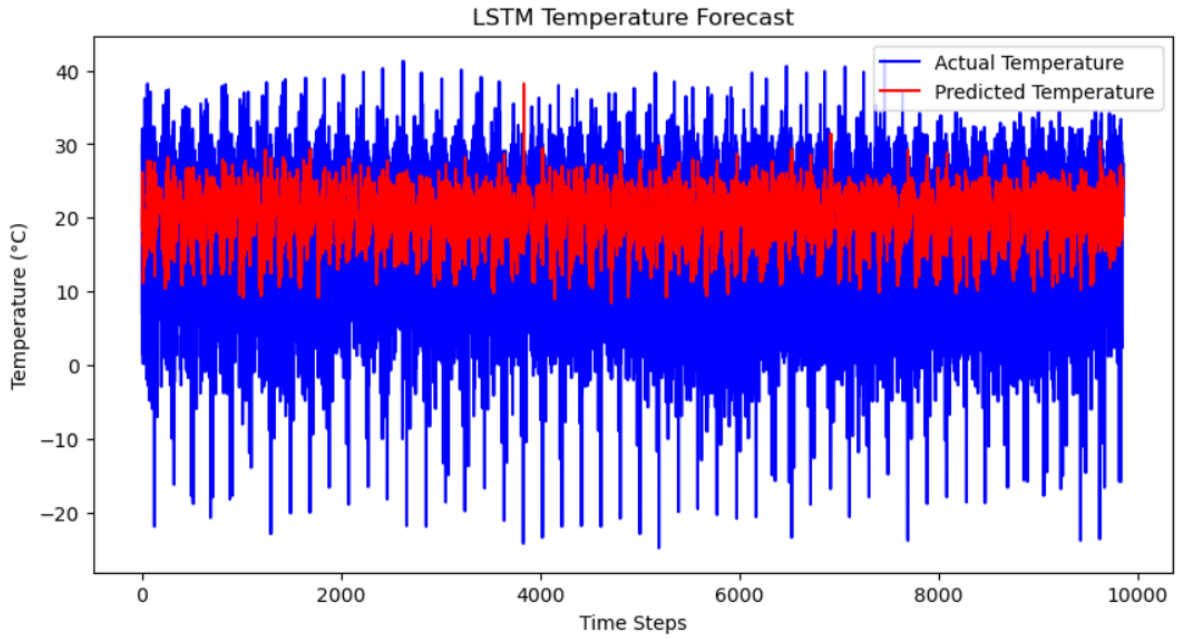


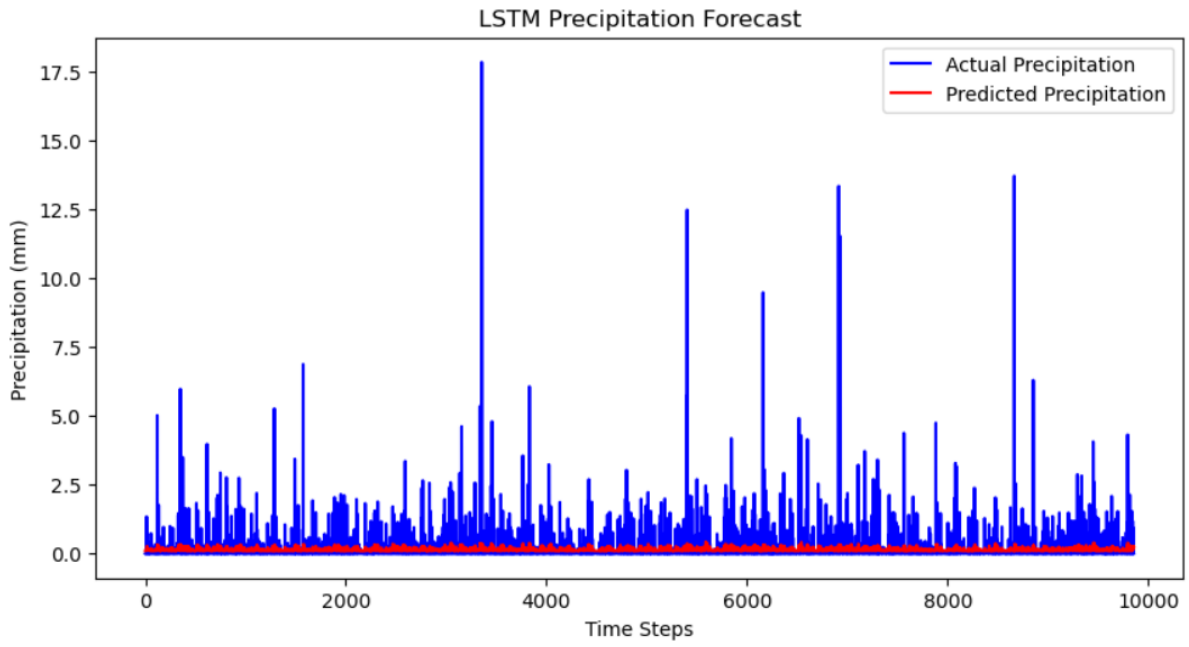
2. LSTM
   LSTM (Long Short-Term Memory)

LSTM is a type of Recurrent Neural Network (RNN) designed for sequence prediction tasks. It is particularly effective for time series data because it can capture long-term dependencies and patterns using memory cells and gates (input, forget, and output gates).

LSTM Temperature Forecasting Metrics:
Mean Absolute Error (MAE): 8.9519
Mean Squared Error (MSE): 131.4387
Root Mean Squared Error (RMSE): 11.4647
R-squared (R^2): -0.0801

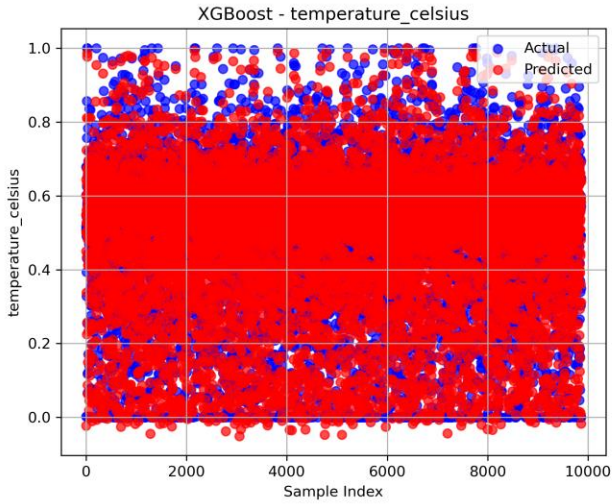
LSTM Precipitation Forecasting Metrics:
Mean Absolute Error (MAE): 0.2074
Mean Squared Error (MSE): 0.2663
Root Mean Squared Error (RMSE): 0.5160
R-squared (R^2): 0.0012

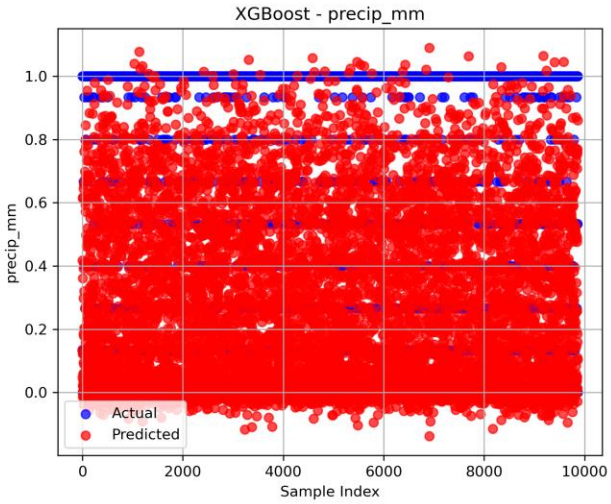


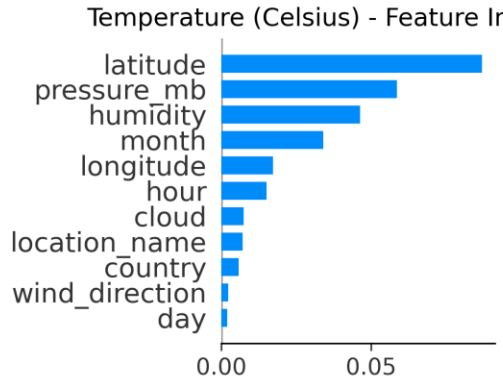


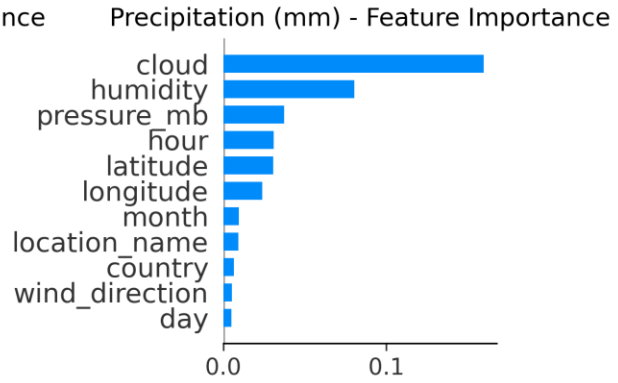
Then we build and train the ensemble machine learning model XGBoost which is based on Gradient boosted Decision Trees also we perform feature importance analysis part of the unique analysis in advanced assessment for this model and gives a plot of features in decreasing order of their contribution in correct forecasting .

```
XGBoost Model Performance:

Performance for temperature_celsius:
Mean Absolute Error (MAE): 0.04
Mean Squared Error (MSE): 0.00
Root Mean Squared Error (RMSE): 0.05
R-squared (R2): 0.94

Performance for precip_mm:
Mean Absolute Error (MAE): 0.19
Mean Squared Error (MSE): 0.08
Root Mean Squared Error (RMSE): 0.29
R-squared (R2): 0.48
```









an(|SHAP value|) (average impact on model output) mean(|SHAP value|) (average impact on model outpu