Weather Trend Forecasting

This notebook analyzes the "Global Weather Repository.csv" dataset to forecast future weather trends. Includes EDA, anomaly detection, forecasting with Ridge Regression, and interactive geospatial visualizations.

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."

Dataset Overview

The dataset includes over 60,000 weather records globally. It includes over 40 features such as temperature, precipitation, wind, and air quality metrics. This cell loads the data and checks its structure.

Parsing and Sorting Dates

We convert the 'last_updated' column to datetime format and sort the dataframe by date to enable proper time series analysis.

	latitude	longitude	last_updated_epoch	temperature_celsius	temperature_fahrenh
count	60411.000000	60411.000000	6.041100e+04	60411.000000	60411.000
mean	19.137636	22.183645	1.729311e+09	22.155538	71.881
std	24.474532	65.808862	7.802518e+06	9.628106	17.330
min	-41.300000	-175.200000	1.715849e+09	-24.900000	-12.800
25%	3.750000	-6.836100	1.722688e+09	16.900000	62.400
50%	17.250000	23.316700	1.729330e+09	25.000000	77.000
75%	40.400000	50.580000	1.736072e+09	28.400000	83.100
max	64.150000	179.220000	1.742723e+09	49.200000	120.600

8 rows × 31 columns

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 60411 entries, 2024-05-16 01:45:00 to 2025-03-23 22:45:00
Data columns (total 41 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   country                        60411 non-null  object 
 1   location_name                  60411 non-null  object 
 2   latitude                       60411 non-null  float64
 3   longitude                      60411 non-null  float64
 4   timezone                       60411 non-null  object 
 5   last_updated_epoch             60411 non-null  int64  
 6   temperature_celsius            60411 non-null  float64
 7   temperature_fahrenheit         60411 non-null  float64
 8   condition_text                 60411 non-null  object 
 9   wind_mph                       60411 non-null  float64
 10  wind_kph                       60411 non-null  float64
 11  wind_degree                    60411 non-null  int64  
 12  wind_direction                 60411 non-null  object 
 13  pressure_mb                    60411 non-null  float64
 14  pressure_in                    60411 non-null  float64
 15  precip_mm                      60411 non-null  float64
 16  precip_in                      60411 non-null  float64
 17  humidity                       60411 non-null  int64  
 18  cloud                          60411 non-null  int64  
 19  feels_like_celsius             60411 non-null  float64
 20  feels_like_fahrenheit          60411 non-null  float64
 21  visibility_km                  60411 non-null  float64
 22  visibility_miles               60411 non-null  float64
 23  uv_index                       60411 non-null  float64
 24  gust_mph                       60411 non-null  float64
 25  gust_kph                       60411 non-null  float64
 26  air_quality_carbon_monoxide    60411 non-null  float64
 27  air_quality_ozone              60411 non-null  float64
 28  air_quality_nitrogen_dioxide   60411 non-null  float64
 29  air_quality_sulphur_dioxide    60411 non-null  float64
 30  air_quality_pm2.5              60411 non-null  float64
 31  air_quality_pm10               60411 non-null  float64
 32  air_quality_us-epa-index       60411 non-null  int64  
 33  air_quality_gb-defra-index     60411 non-null  int64  
 34  sunrise                        60411 non-null  object 
 35  sunset                         60411 non-null  object 
 36  moonrise                       60411 non-null  object 
 37  moonset                        60411 non-null  object 
 38  moon_phase                     60411 non-null  object 
 39  moon_illumination              60411 non-null  int64  
 40  target                         60411 non-null  float64
dtypes: float64(24), int64(7), object(10)
memory usage: 19.4+ MB
```

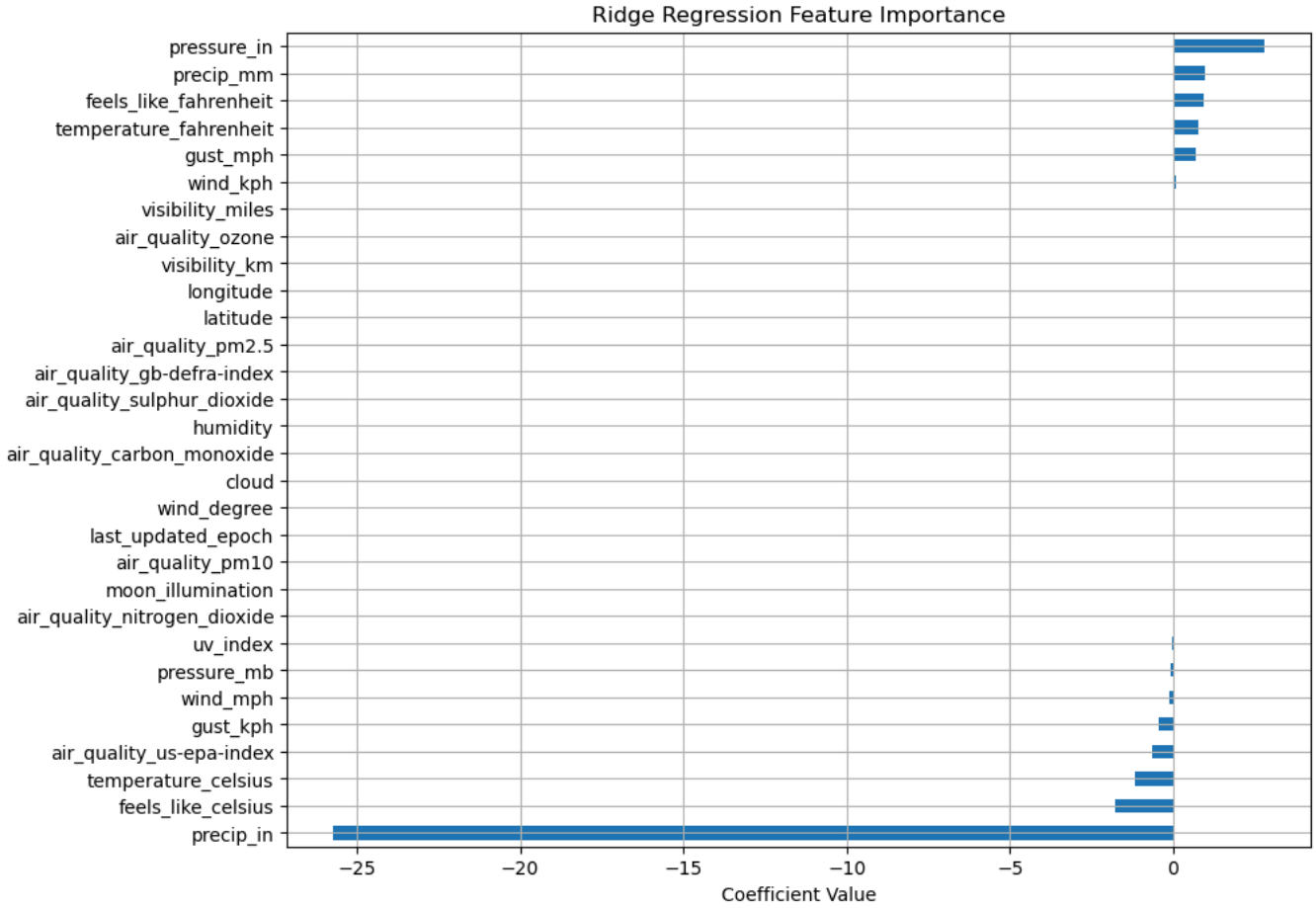
last_updated	country	location_name	latitude	longitude	timezone	last_updated_ep
2024-05-16 01:45:00	United States of America	Washington Park	46.60	-120.49	America/Los_Angeles	171584
2024-05-16 02:45:00	El Salvador	San Salvador	13.71	-89.20	America/El_Salvador	171584
2024-05-16 02:45:00	Costa Rica	San Juan	9.97	-84.08	America/Costa_Rica	171584
2024-05-16 02:45:00	Guatemala	Guatemala City	14.62	-90.53	America/Guatemala	171584
2024-05-16 02:45:00	Nicaragua	Managua	12.15	-86.27	America/Managua	171584

5 rows × 41 columns

Initial Data Inspection

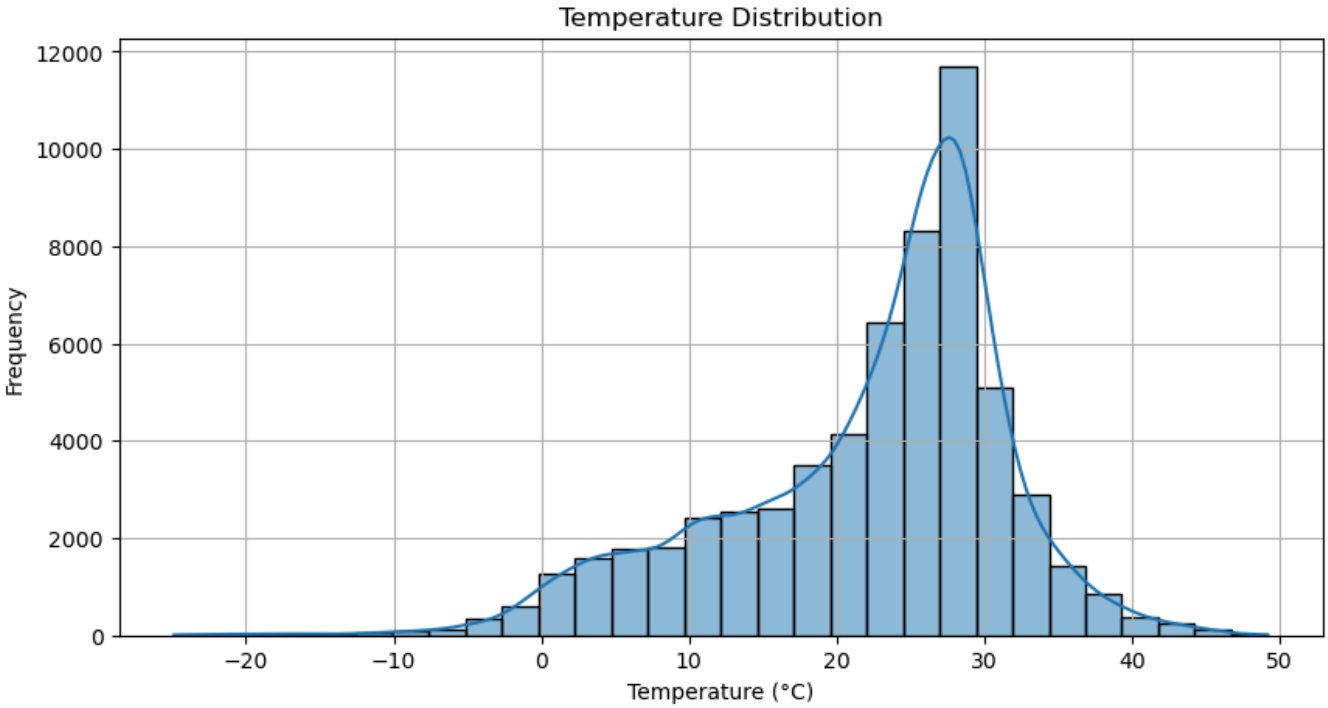
Here we examine data types, missing values, and summary statistics to understand feature distributions and potential issues.

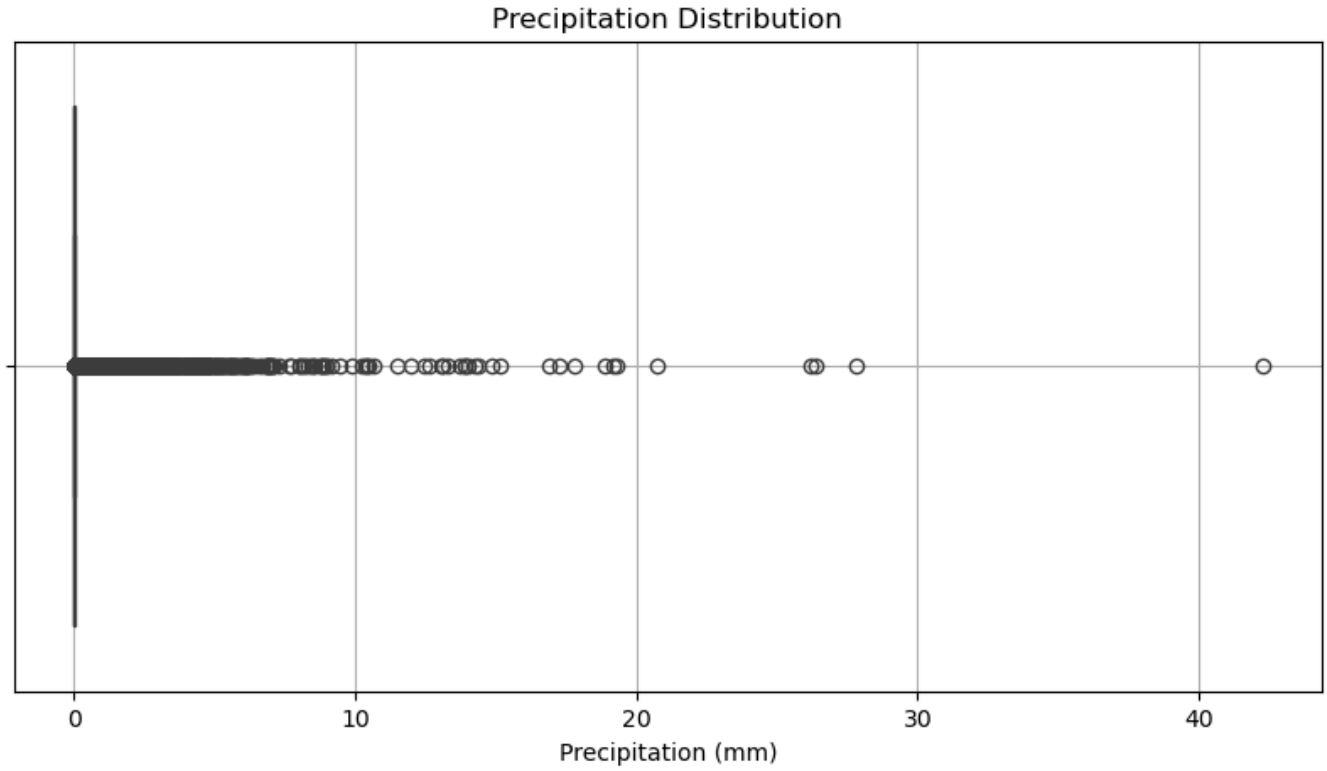
```
Ridge
Ridge(alpha=0.1)
```

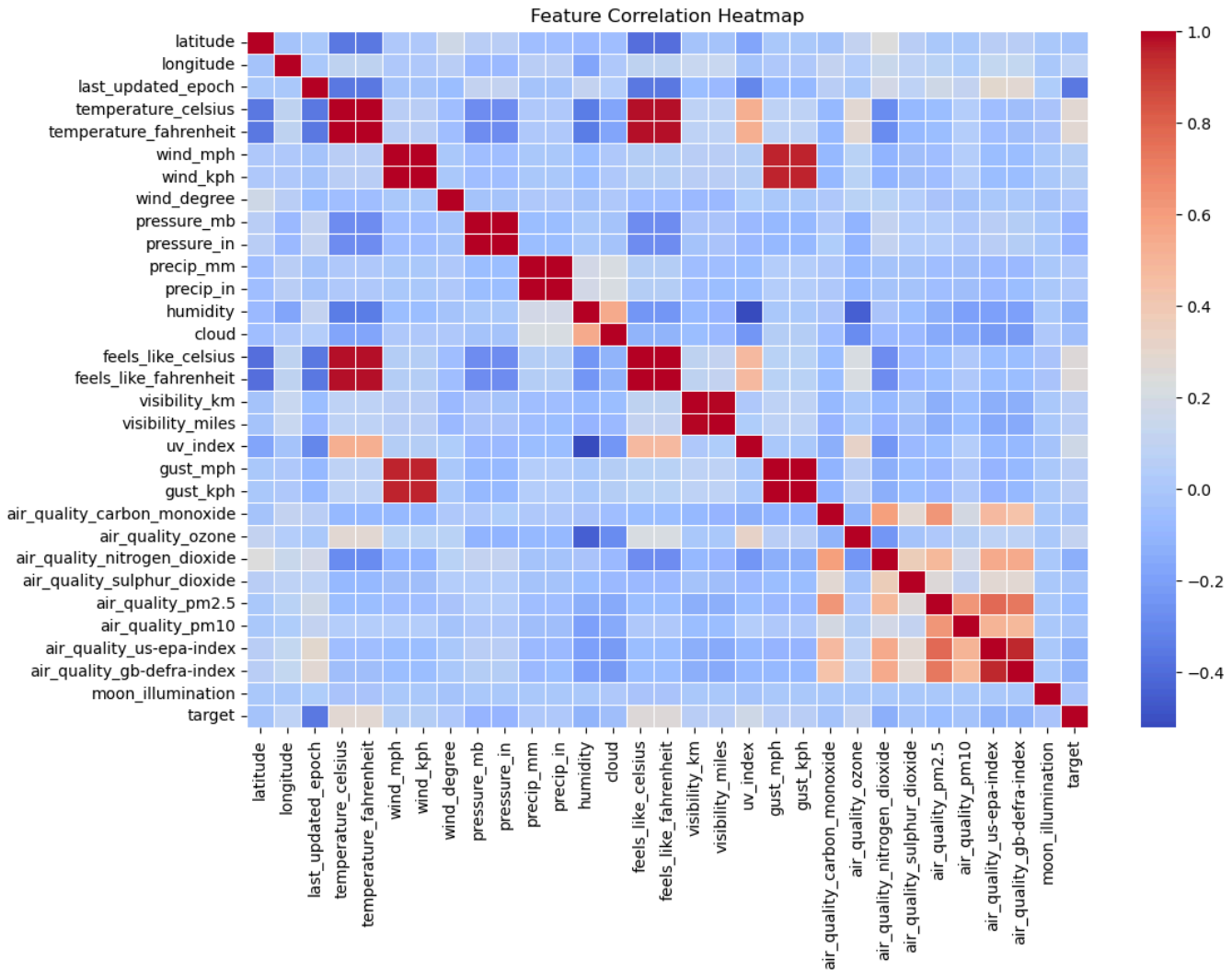


Basic Exploratory Data Analysis (EDA)

We begin with standard EDA to understand temperature distribution, wind patterns, and basic weather conditions. This helps us detect initial outliers, get a sense of data trends, and choose suitable features for further modeling.


Temperature Distribution
Frequency
12000
10000
8000
6000
4000
2000
0
−20
−10
0
10
20
30
40
50
Temperature (°C)

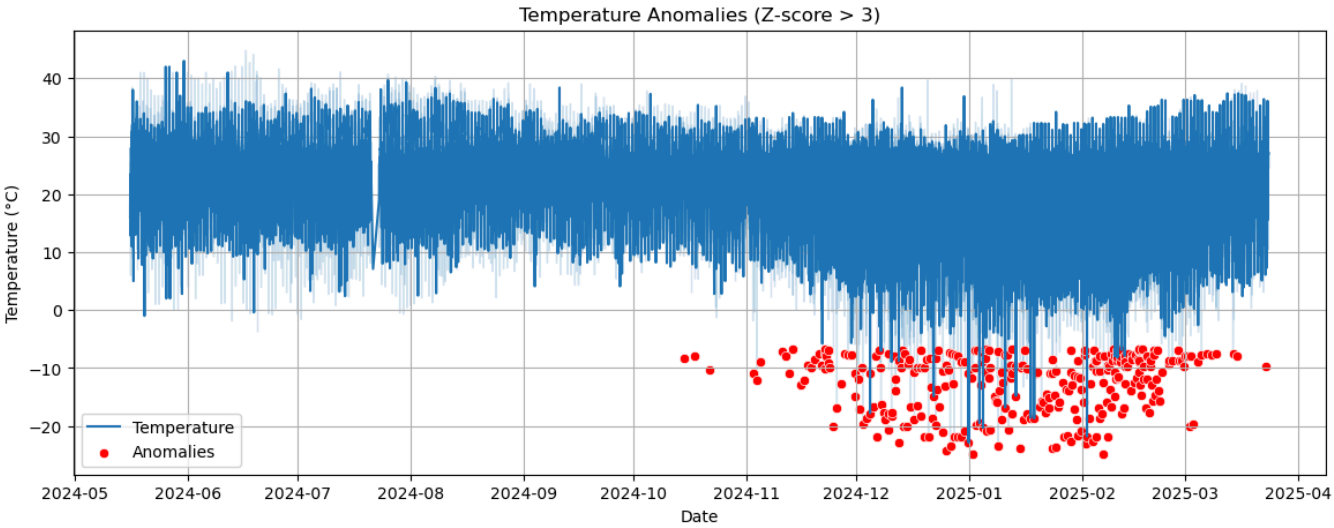

Precipitation Distribution
0
10
20
30
40
Precipitation (mm)



Correlation Heatmap

The heatmap helps visualize relationships between numeric features, revealing which variables are most closely associated. This heatmap shows that the humidity has strong negative correlation to uv index and air quality ozone which will be demonstrated by plots further in this report. Another interesting correlation showed by the heatmap is the fairly strong negative correlation between latitude and temperature, but not longitude.

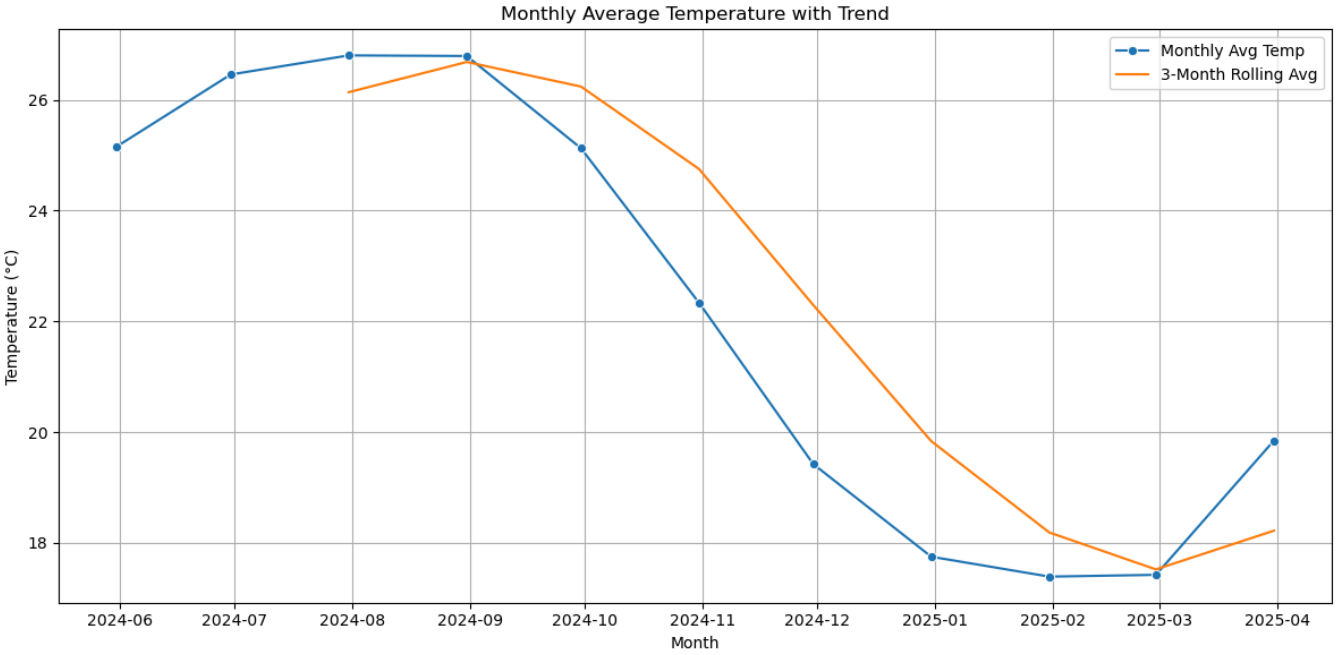
Temperature Anomaly Detection

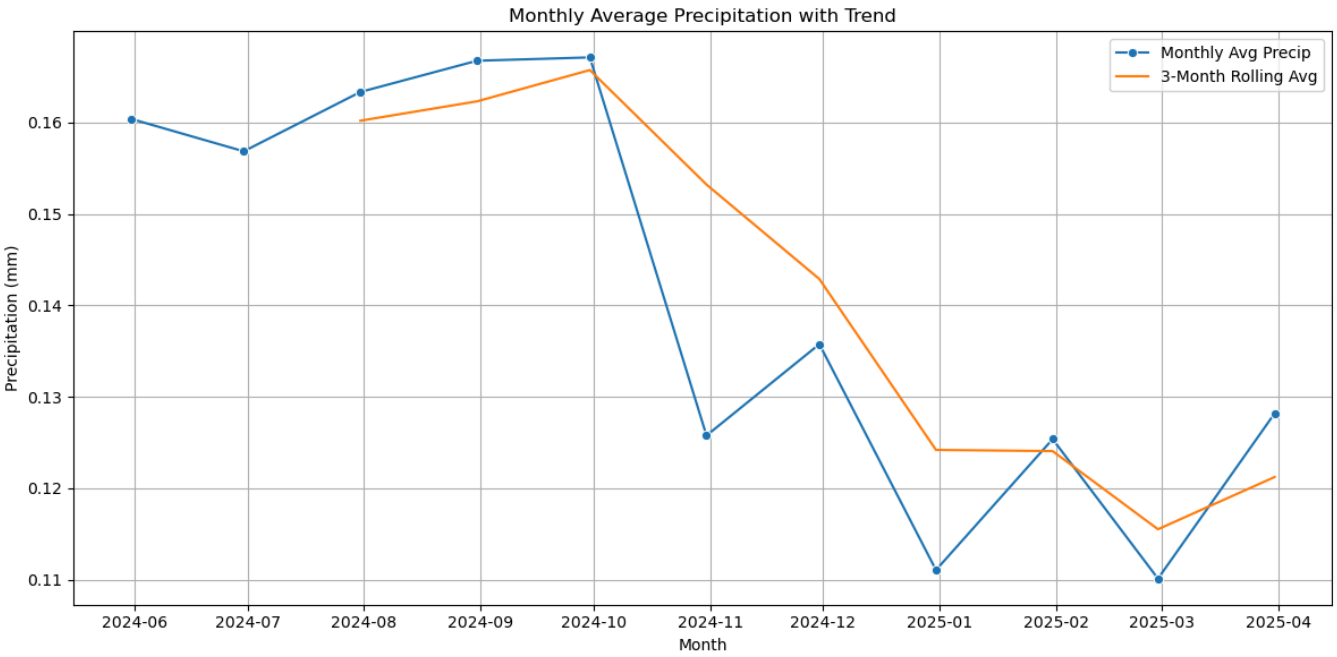


Temperature Anomaly Detection

We apply Z-score thresholding to detect extreme deviations in temperature. These anomalies may indicate measurement errors, natural disasters, or exceptional weather events. Removing or flagging them ensures our models are not skewed.

Climate Trends Over Time





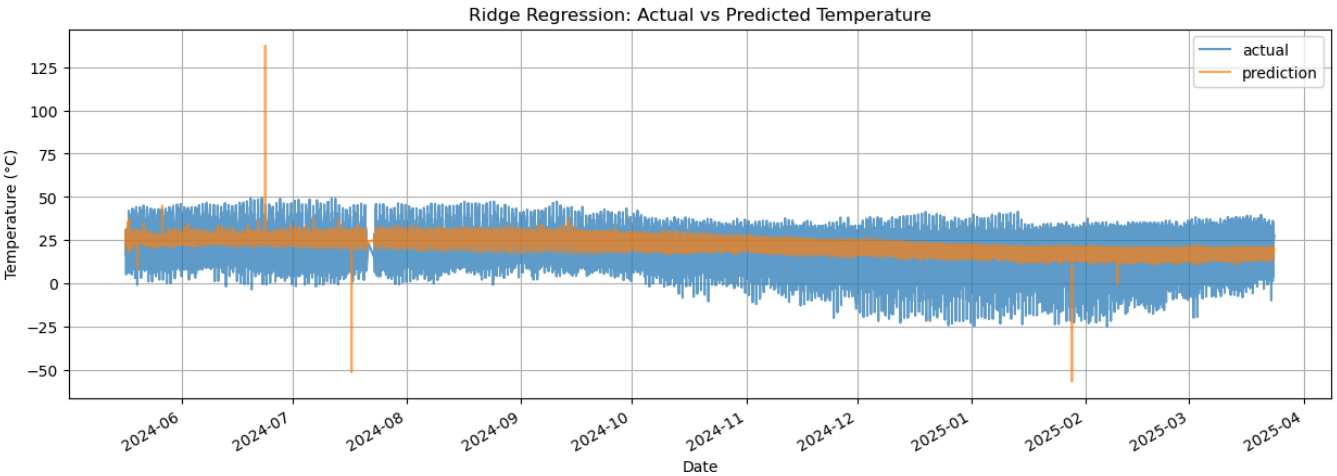
Long-Term Temperature Trend

By resampling temperature data monthly, we observe long-term warming or cooling trends globally. This chart helps visualize whether temperatures are generally increasing over the months available.

Ridge Regression Forecasting

We train a Ridge Regression model to forecast temperature. Ridge helps mitigate multicollinearity and is suitable for linear relationships.

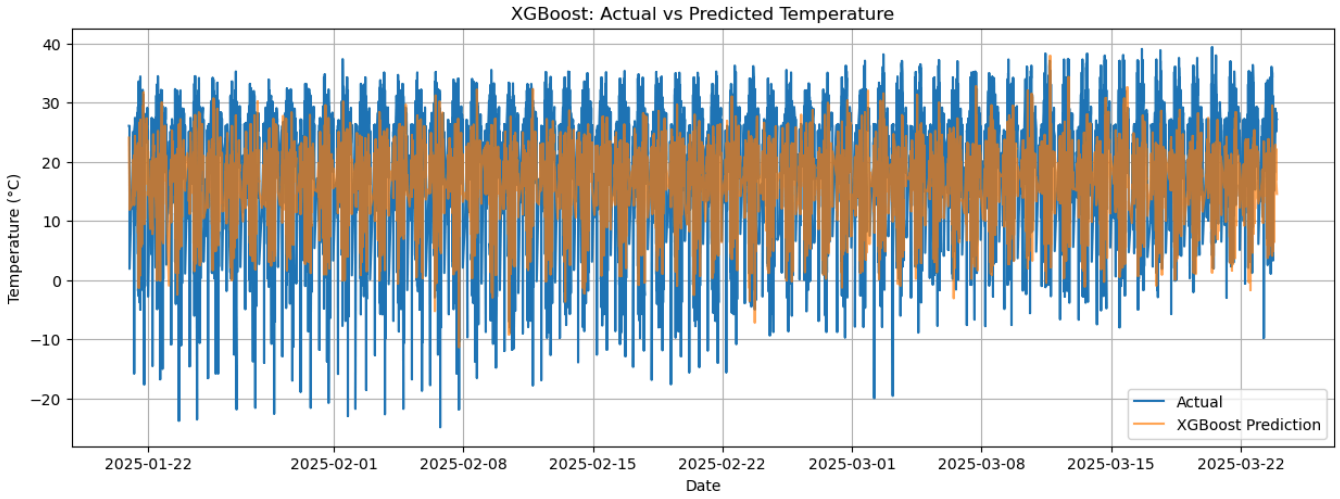
```
MAE: 6.98
RMSE: 8.85
```



Forecasting with XGBoost Regressor

XGBoost Model

XGBoost is a gradient boosting model that captures nonlinear relationships. It may improve accuracy compared to linear models.




```
XGBoost MAE: 9.26
XGBoost RMSE: 11.54
```

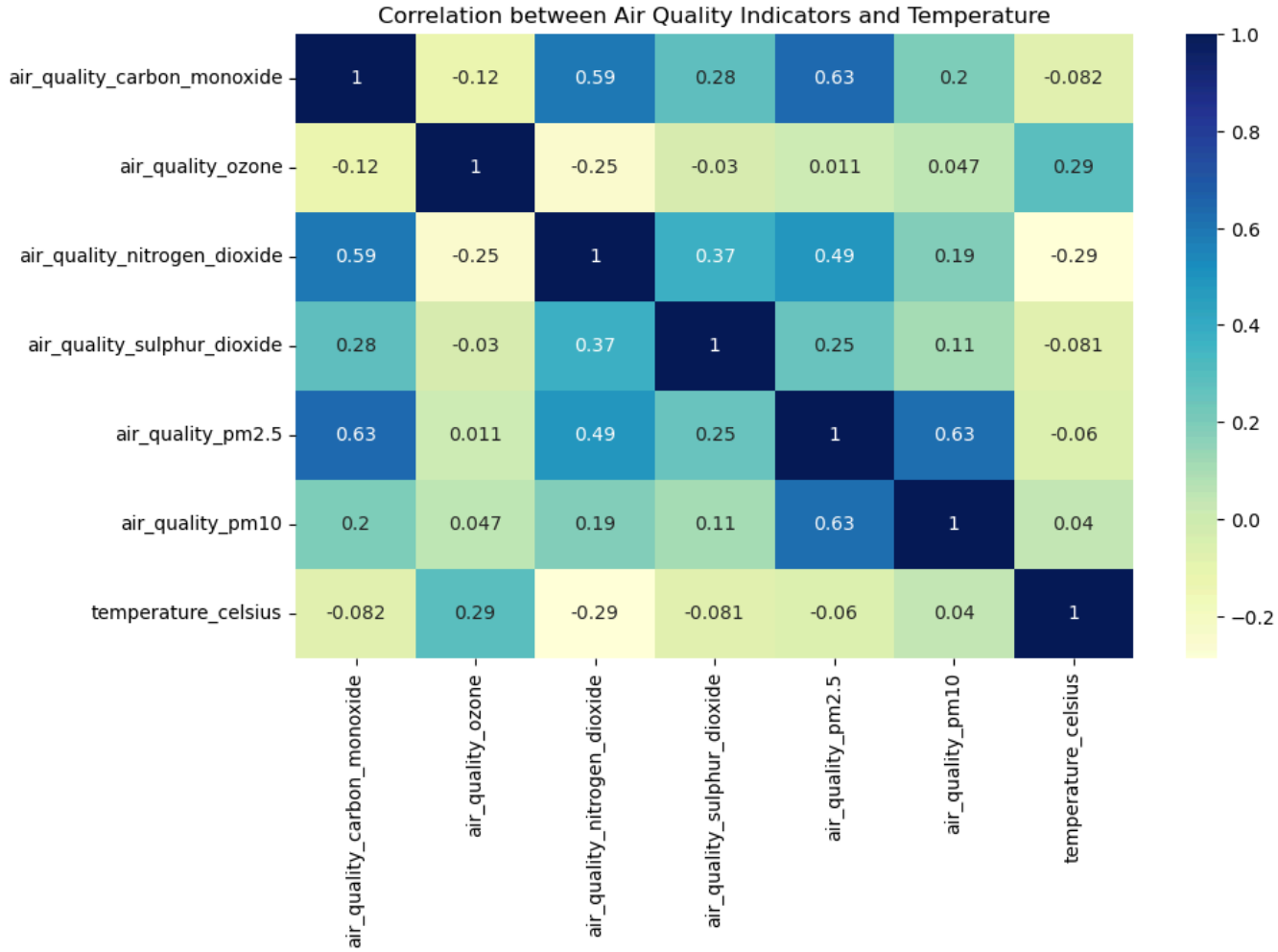
Ridge Regression Forecasting Performance

We backtested the Ridge Regression model across multiple windows and measured its Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results suggest the model performs reasonably well given the short-term nature of the data.

Environmental Impact: Air Quality Analysis

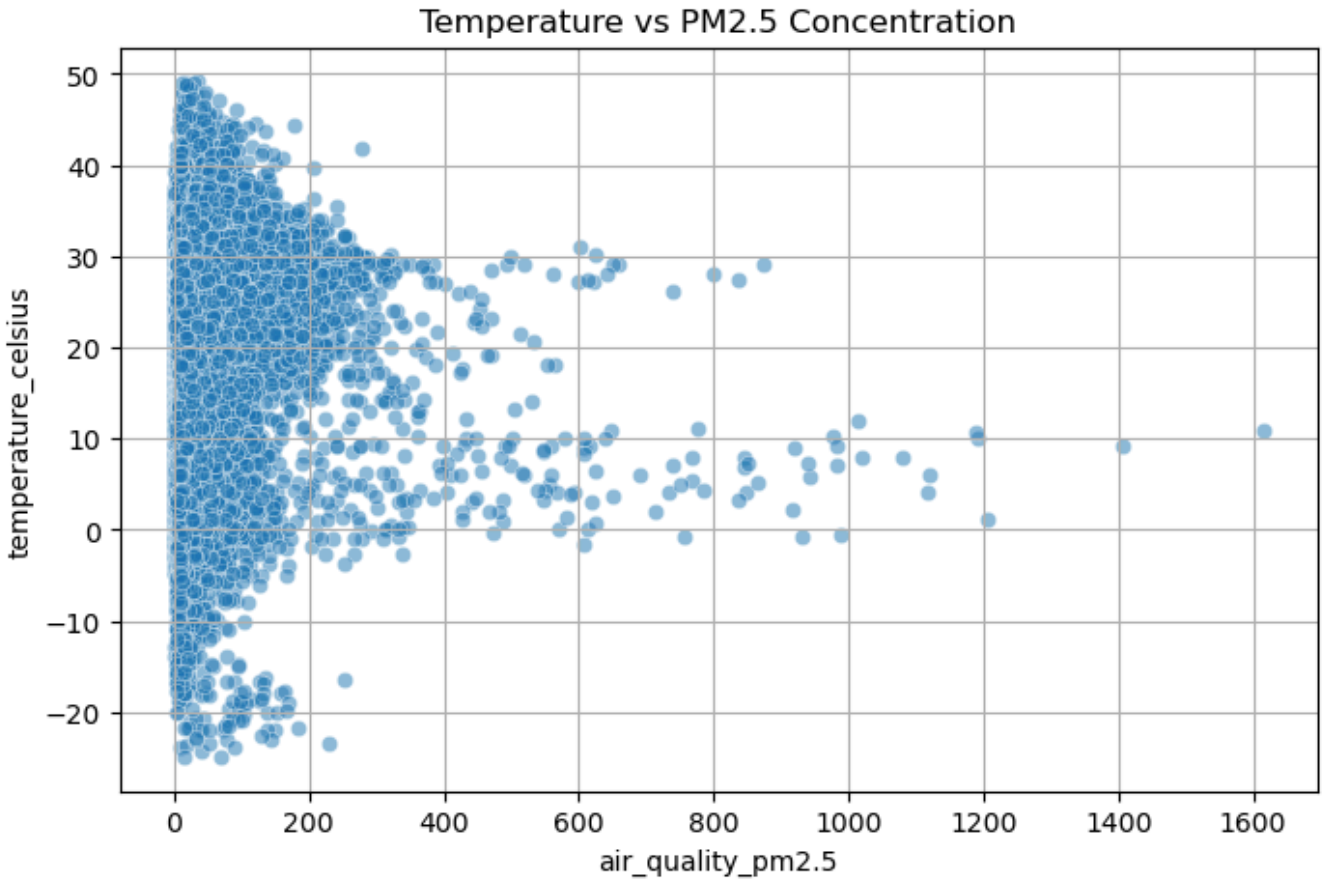
Feature Importance Visualization

This chart highlights the most influential features in the Ridge model, offering interpretability and insight into key drivers.

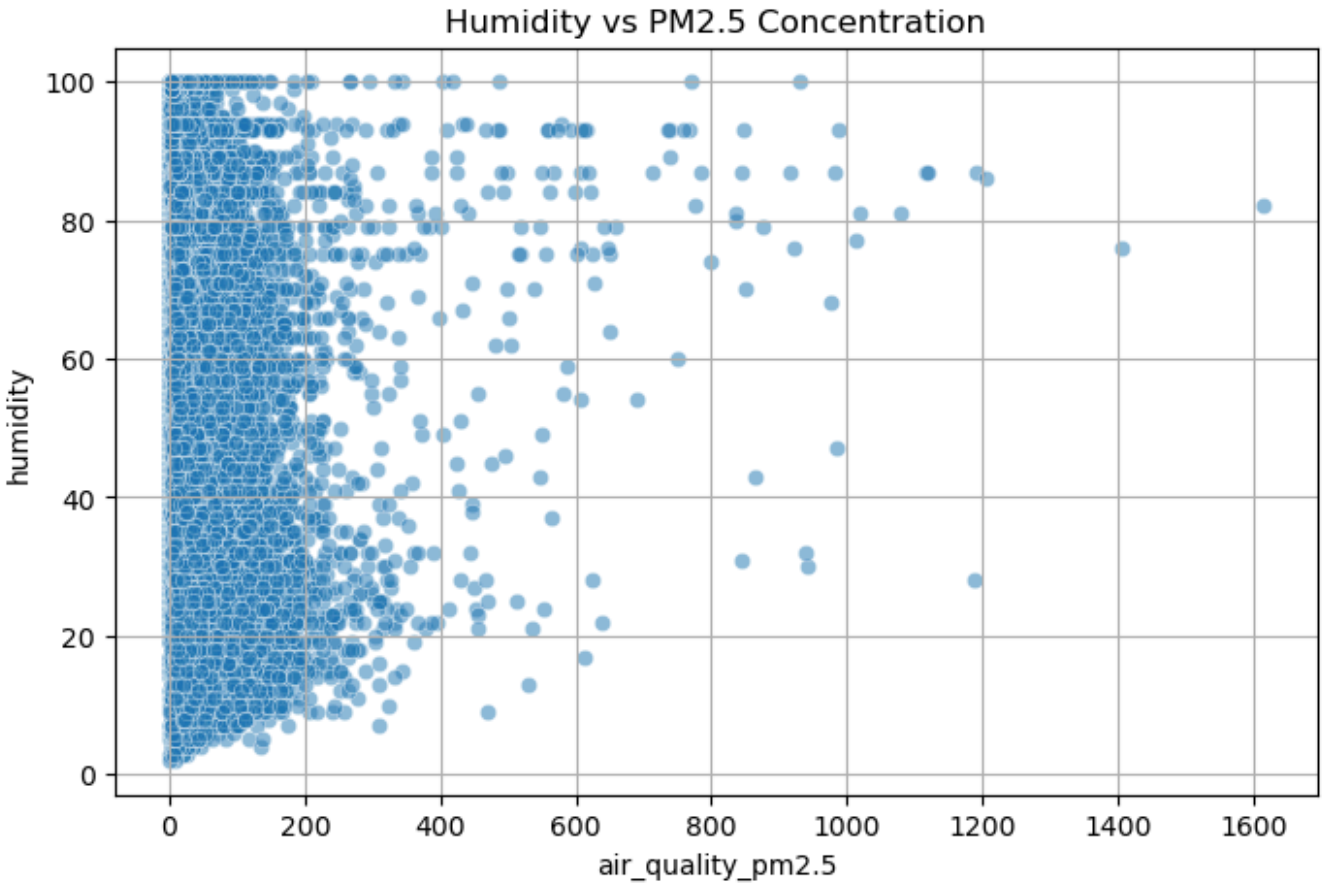


Environmental Impact Analysis

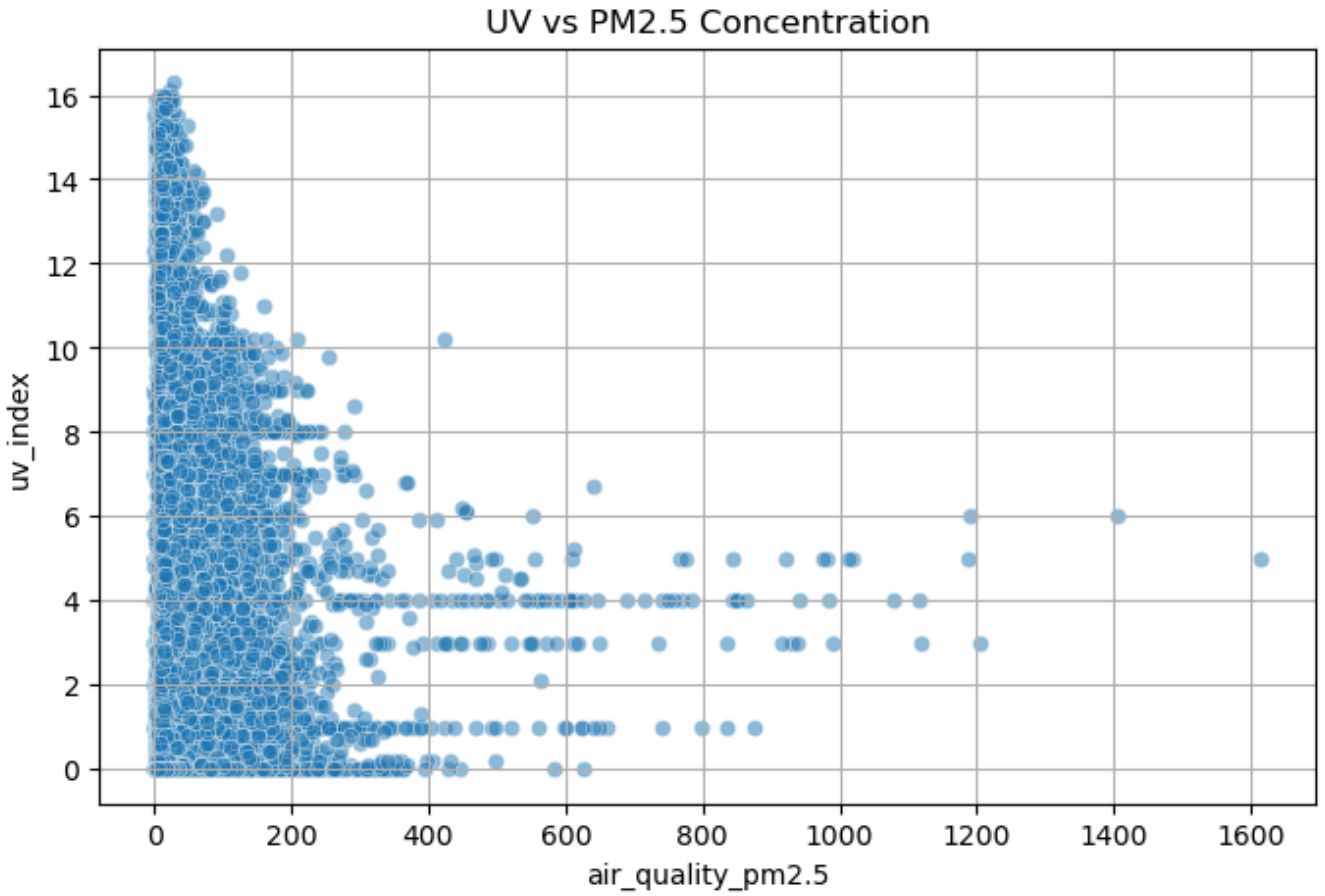
Here we explore the relationship between temperature and air quality metrics (e.g., PM2.5, NO2) to understand environmental correlations.



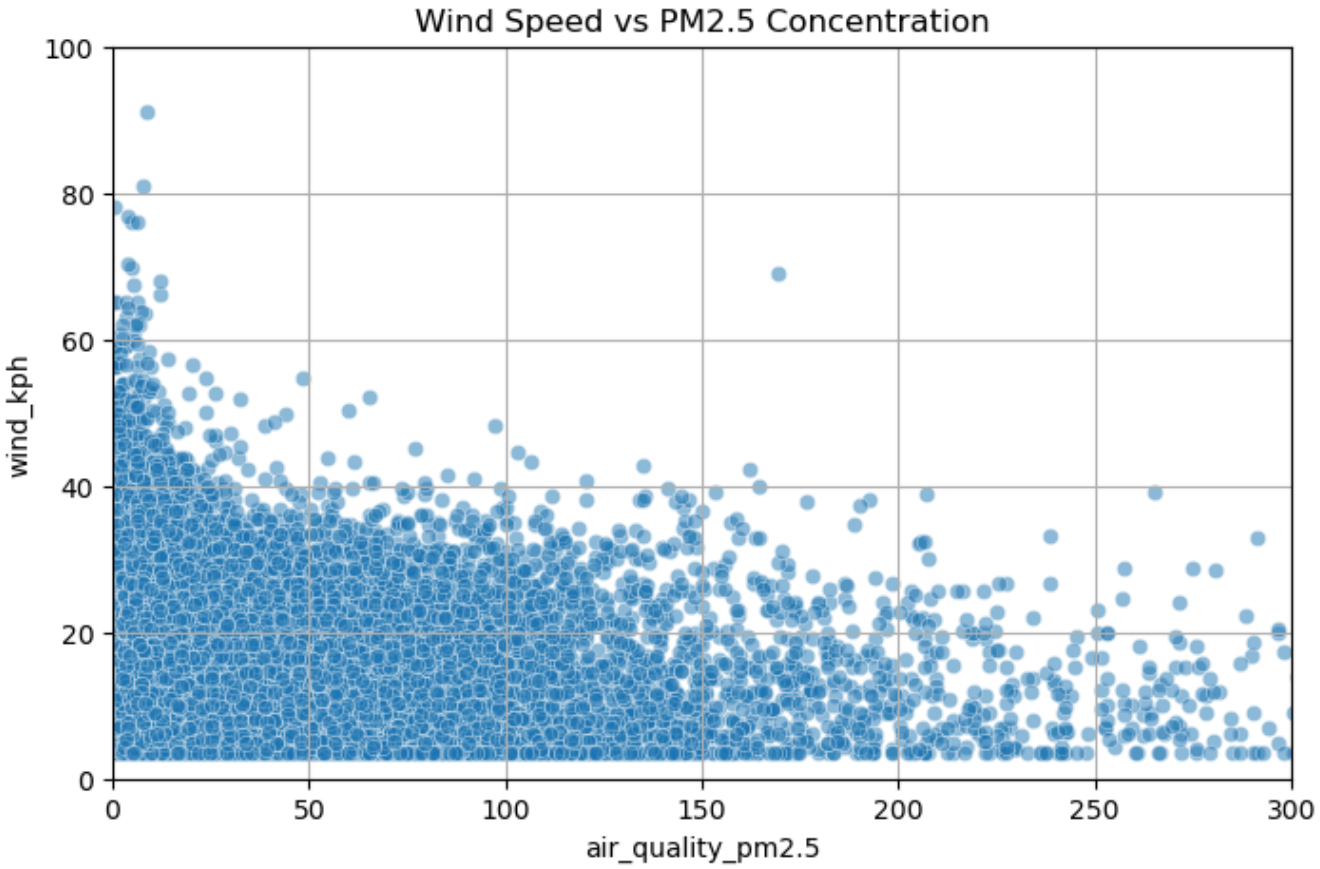
This chart shows peak PM2.5/poor air quality when temperatures are between 0-10C with another peak around 30C



This plot shows higher peaks of pm2.5 (poorer air quality) with higher humidity



This plot shows higher pm2.5 levels with lower UV index, indicating higher uv indexes lead to better air quality.

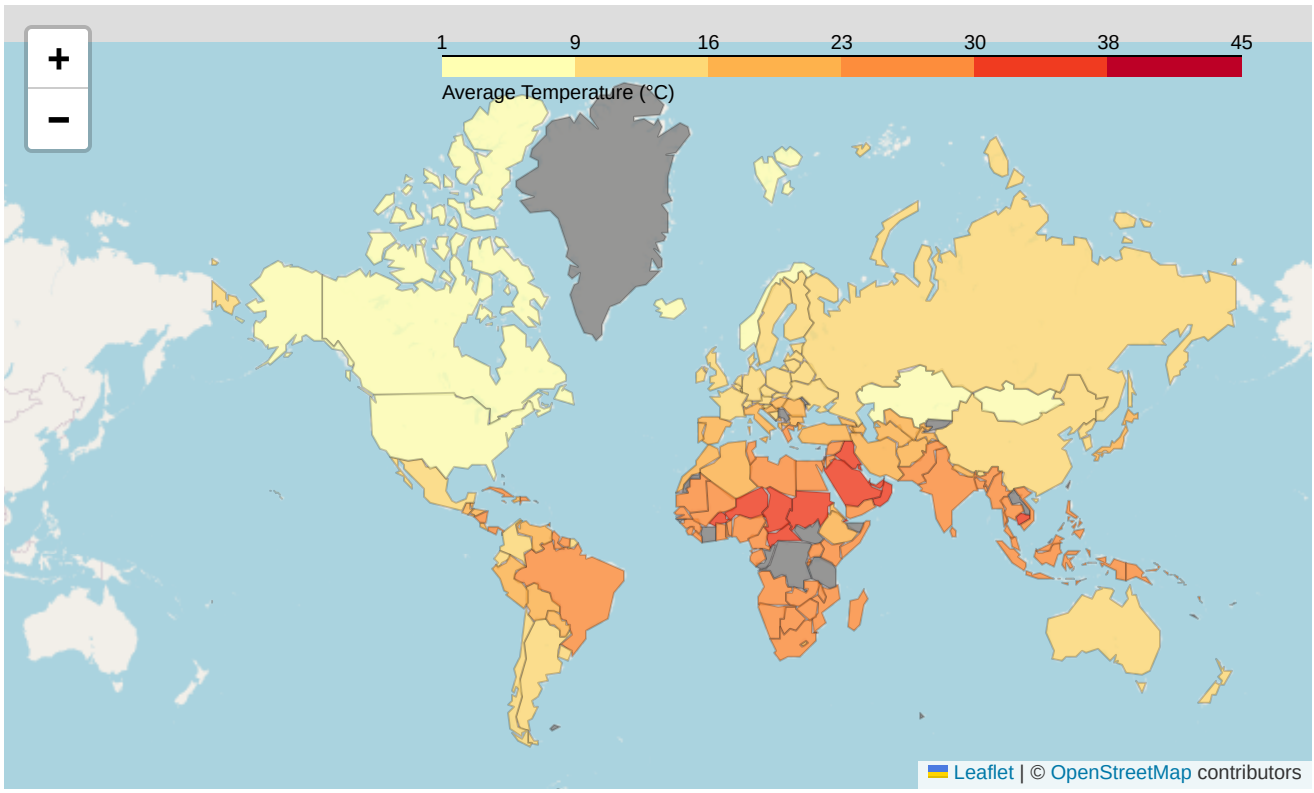


This plot shows better air quality with higher wind speeds, with perfect air quality with wind speeds above about 90km/h

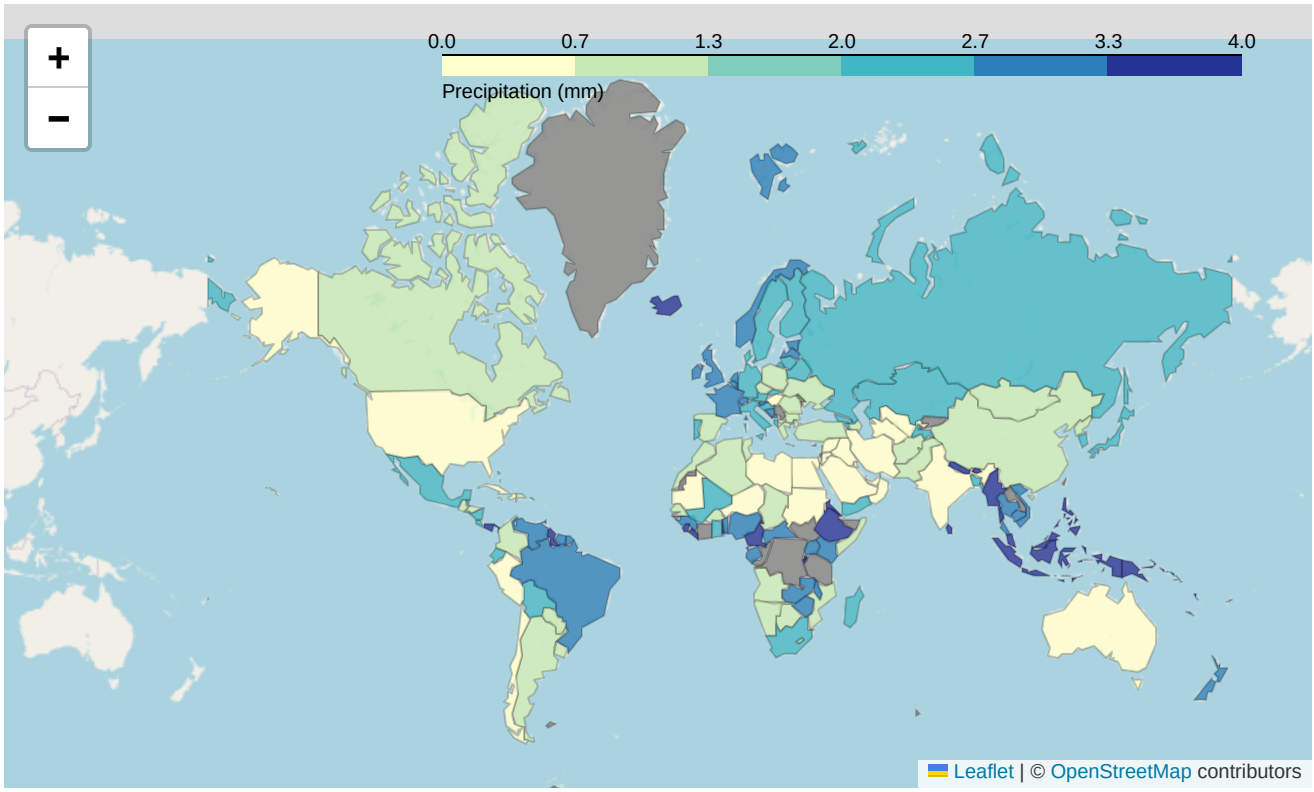
Global Climate Choropleth Maps

Geospatial Weather Visualization

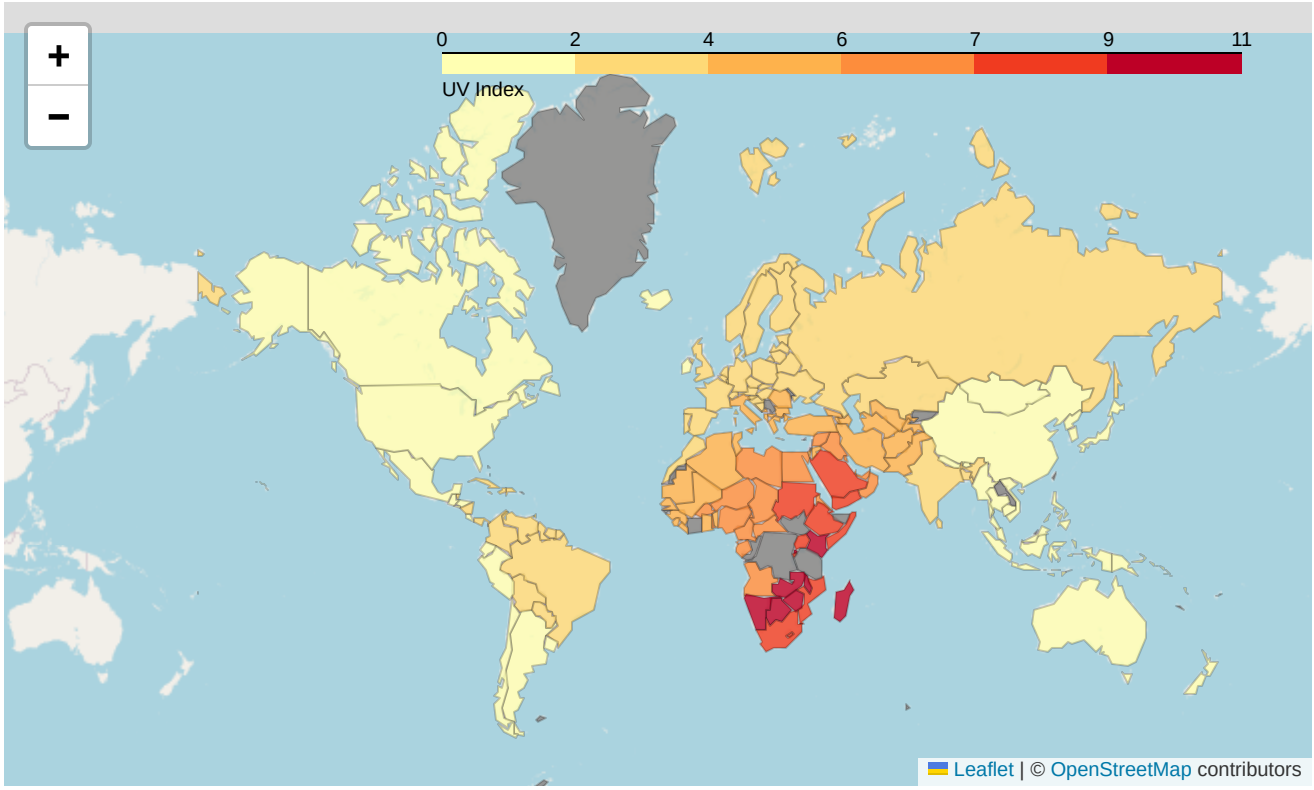
We use Folium to plot country-level average temperature and precipitation. This reveals geographic weather patterns globally.



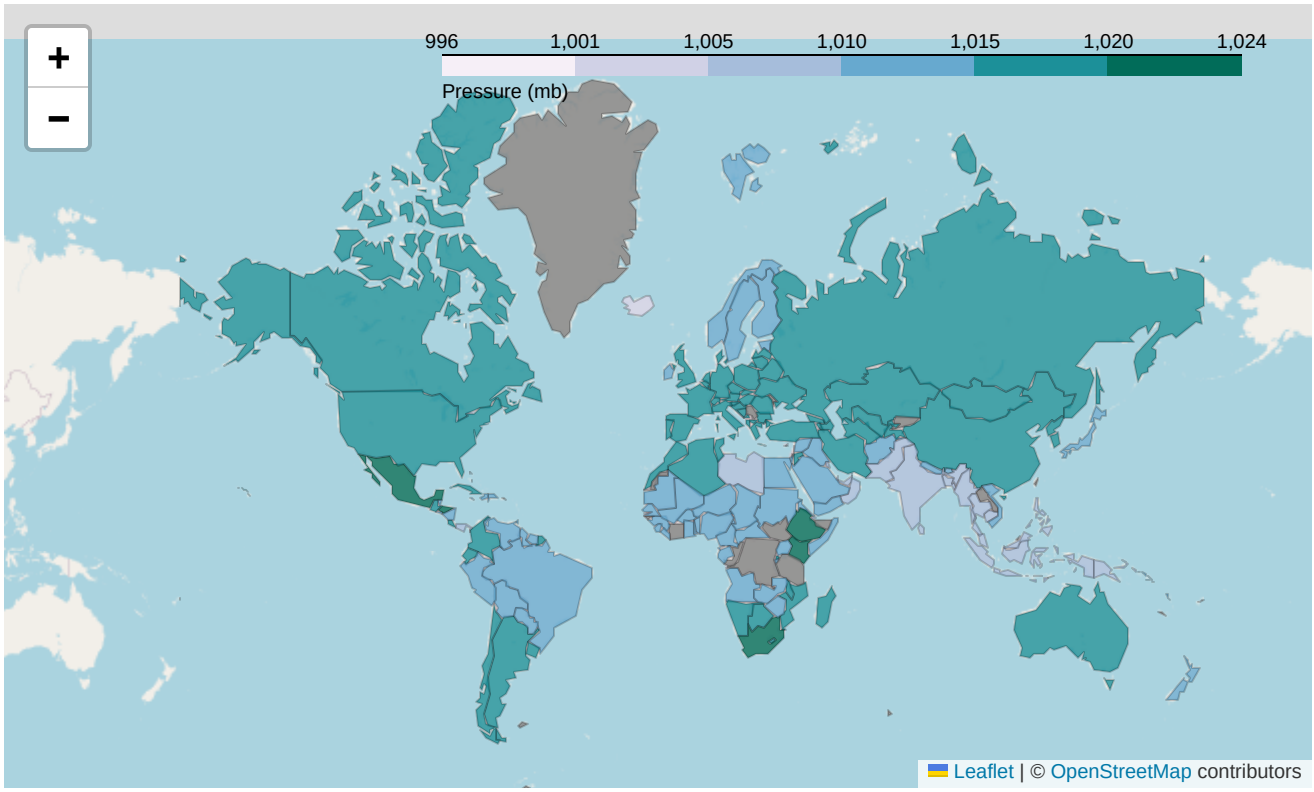
Areas with the highest temperatures are close to the equator and/or are in arid (dry, desert) climates.



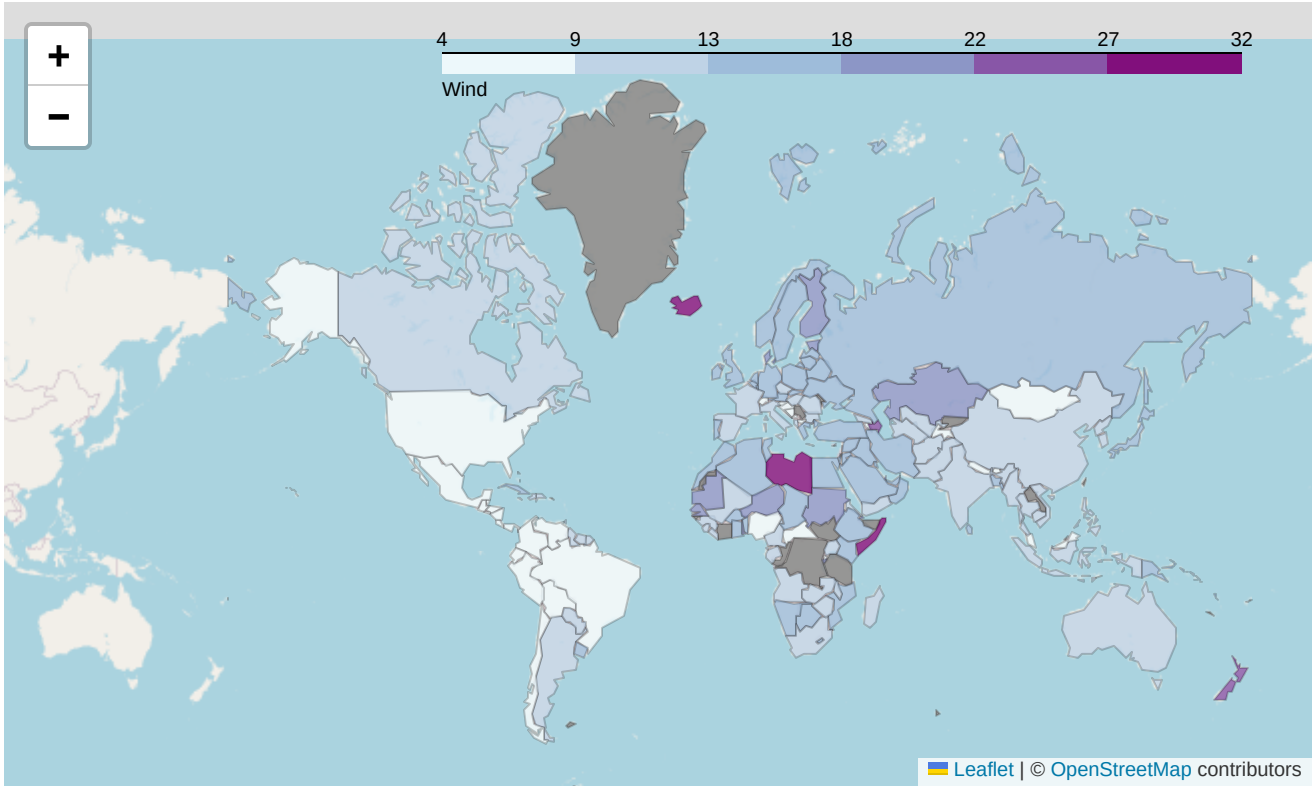
Proximity to equator seems to have an effect on average precipitation in an area.



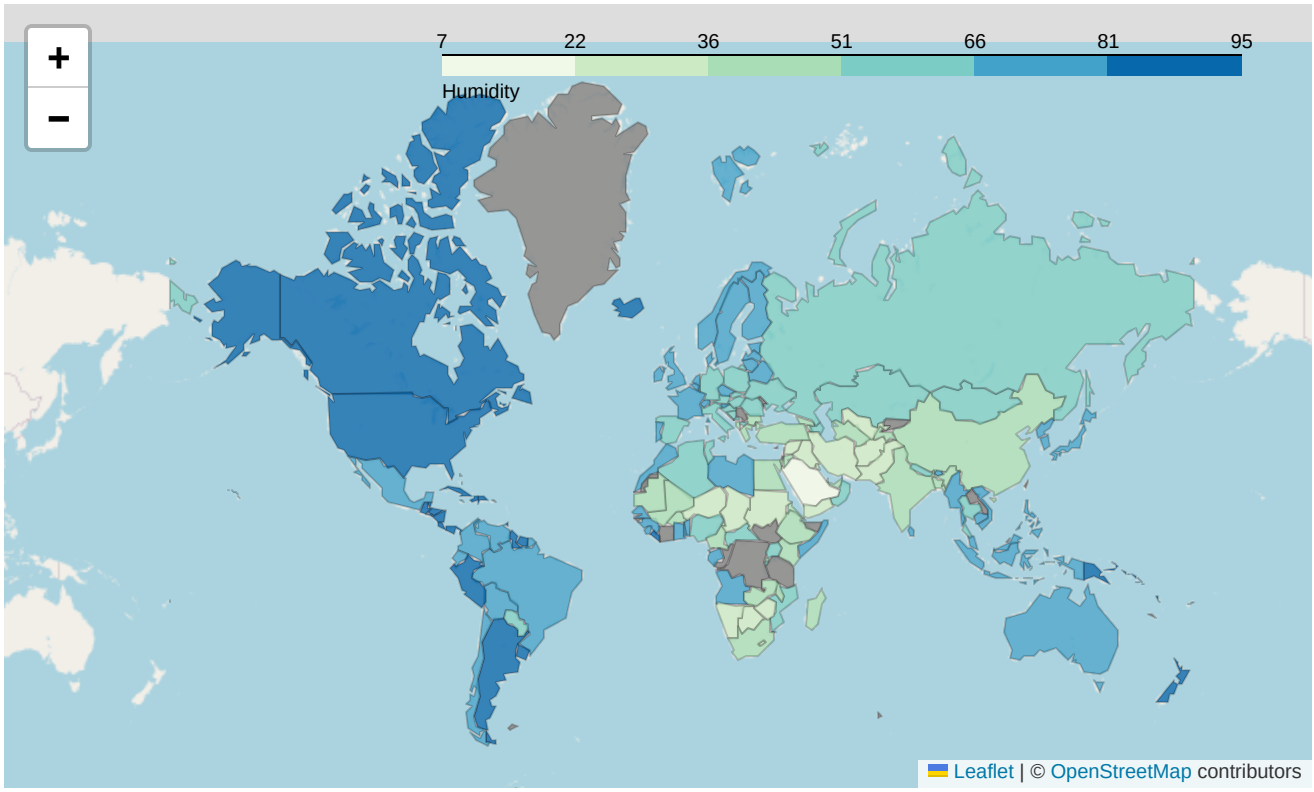
Africa has the highest UV index as it is closest to the sun due to the earth's rotational tilt. This map illustrates that effect.



This chart highlights zones of similar pressure that seem to cross over continental lines creating new zones.



This map shows a few countries with the highest wind speeds. It seems Africa has the highest wind speeds by continent.



This map shows that the americas may have highest humidity levels compared to the rest of the planet.

Conclusion

- Completed climate analysis through extensive data cleaning, feature engineering, and exploratory analysis.
- Developed and evaluated Ridge and XGBoost models for temperature prediction.
- Analyzed air quality and climate data to find environmental impact.
- Analyzed and visualized geographical patterns in the data through spatial analysis.
- Future improvements may include seasonal decomposition, deep learning, or ensemble modeling.