**PM Accelerator**

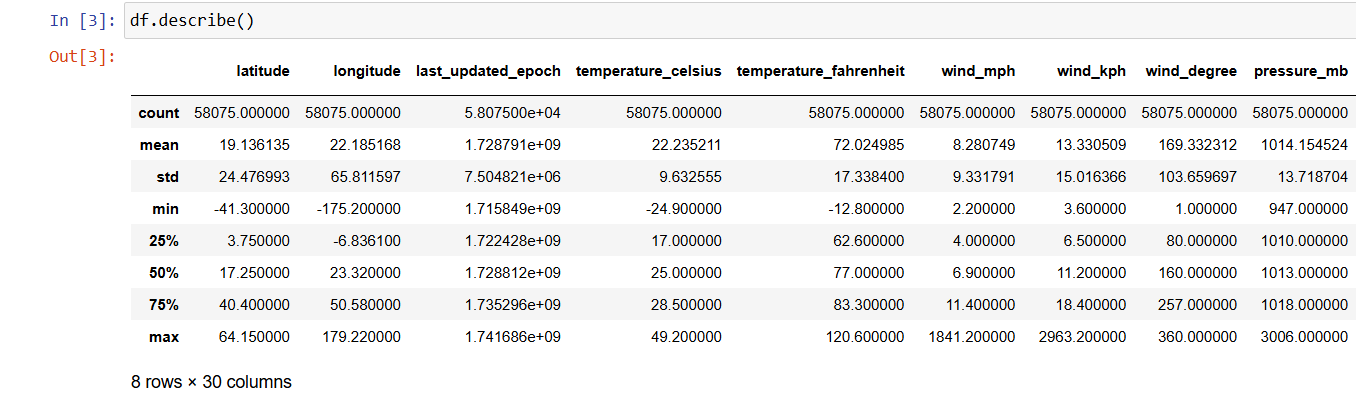
**Mission**: Product Management Accelerator (PM Accelerator) is dedicated to empowering international professionals to swiftly transition into product management roles, secure top offers from leading companies, and develop as product leaders. Our unique job search framework, developed by founder Dr. Nancy Li, helps to solidify PM skills, create product portfolios, overcome various challenges in the PM job search, accelerate career development, build a global network of product managers, and secure high salaries during job changes and searches.

**Objective:** Analyse the “Global Weather Repository.csv” dataset to forecast future weather trends and showcase data science skills through both basic and advanced techniques. This dataset provides Daily weather information for cities around the world. This dataset offers a comprehensive set of features that reflect the weather conditions worldwide. It includes over 40 features.

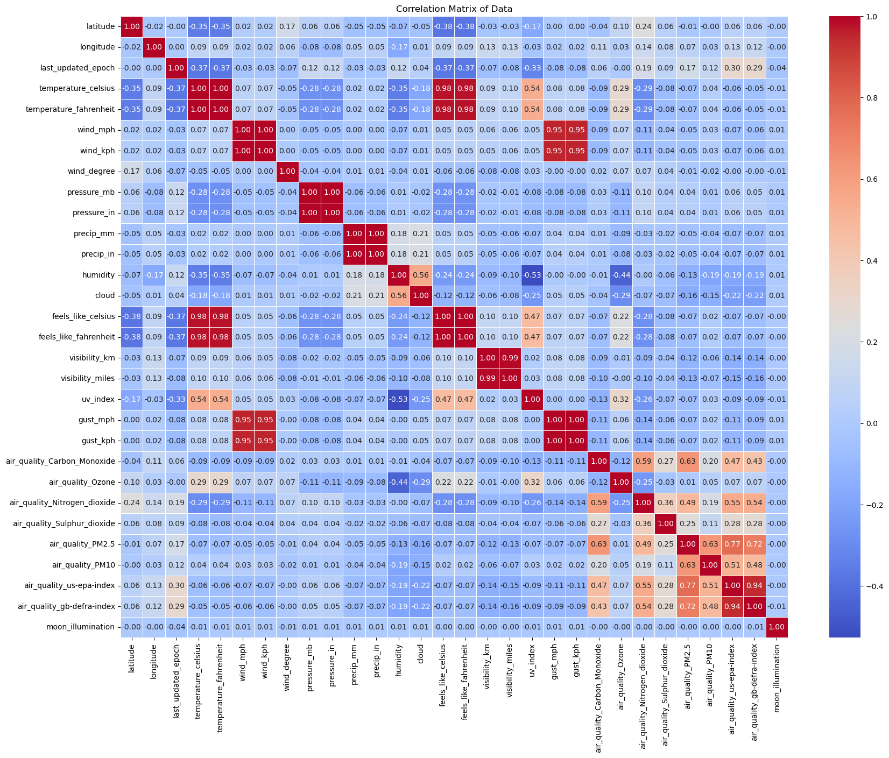
**Introduction:** The notebook performs a comprehensive analysis of weather data, including EDA, outlier handling, time series forecasting, anomaly detection, and predictive modelling. This document explains the code cells, insights gained and the rationale behind computations.

1. **Loading Data**: Numpy, Pandas, Seaborn and Mtplotlib are those libraries essential for hanlding, processing and visualizing data efficiently. Pandas efficiently loads structured data for analysis. It ensures correct formatting before further processing. It allows easy manipulation of data for further analysis.
2. **Understanding Data & Performing EDA**:

* df.describe() summarises numerical features such as mean, median, and standard deviation. It helps detect the skewness in data and a quick insight into data distributions.

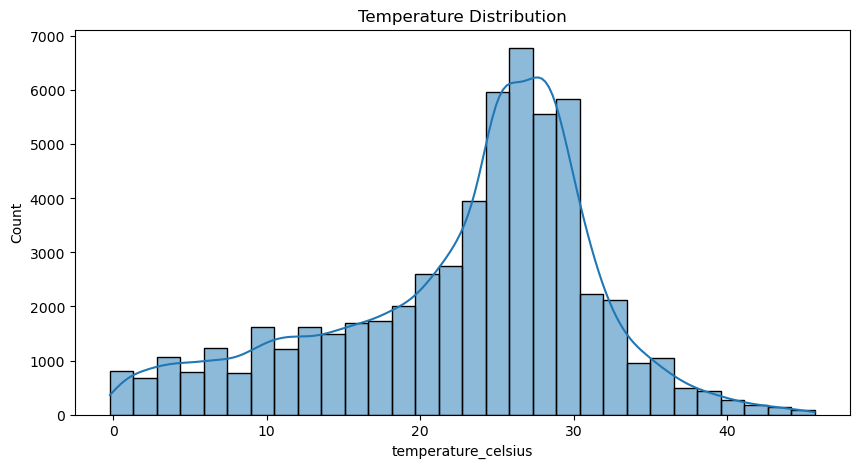


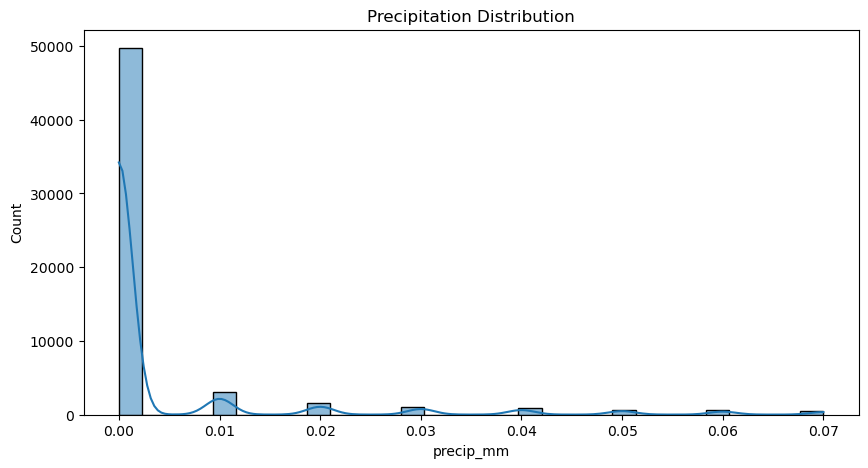
* The data do not have any missing values. Therefore it does not require any imputation.
* The heatmap shows how variables are related to each other. It helps in feature selection by removing redundant variables.



The correlation matrix provides insights into the relationships between various weather and environmental factors. Strong positive correlations exist between temperature (Celsius & Fahrenheit) and the "feels like" temperature, as expected, while wind speed and gust speed are also highly correlated, indicating that stronger winds generally accompany higher gusts. Air quality indicators such as Carbon Monoxide, PM2.5, and PM10 show moderate to strong correlations, suggesting that poor air quality conditions often involve multiple pollutants simultaneously. On the other hand, strong negative correlations are observed between temperature and humidity, implying that hotter weather is typically associated with lower humidity levels. Atmospheric pressure negatively correlates with precipitation, reinforcing the idea that low-pressure systems often lead to rainfall, while higher pressure is linked to clearer weather. A notable inverse relationship is also seen between UV index and cloud cover, where increased cloudiness reduces direct sunlight exposure. Interestingly, geographic coordinates (latitude and longitude) do not exhibit strong correlations with weather variables, suggesting that location alone does not determine climate variations in this dataset. These insights are useful for feature selection in predictive models, weather forecasting, and understanding environmental impacts on air quality.

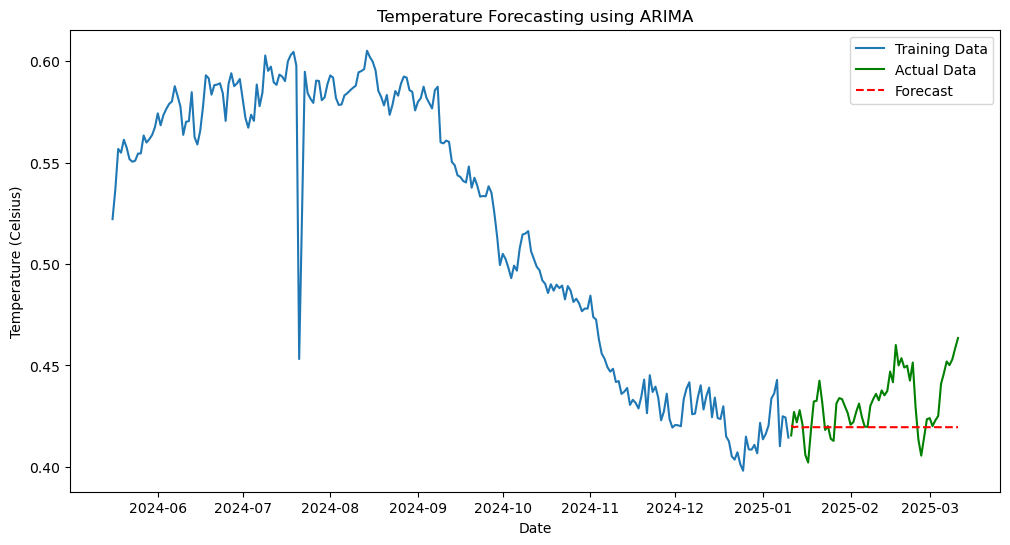
1. **Handling Outliers:** Boxplots help visualize data distribution and detect outliers. The initial boxplots showed significant outliers in several numerical features, including temperature, wind speed, precipitation, and air quality indicators, which could distort statistical analysis and bias machine learning models. By replacing these outliers with the median, the distribution became more balanced, reducing extreme values while preserving meaningful variations. This transformation ensures that the dataset remains realistic without being overly influenced by anomalies, leading to improved model stability and more accurate predictions. The refined boxplots demonstrate a cleaner dataset, making it more suitable for analysis and forecasting.
2. **Visualizing the temperature and Precipitation distribution.**

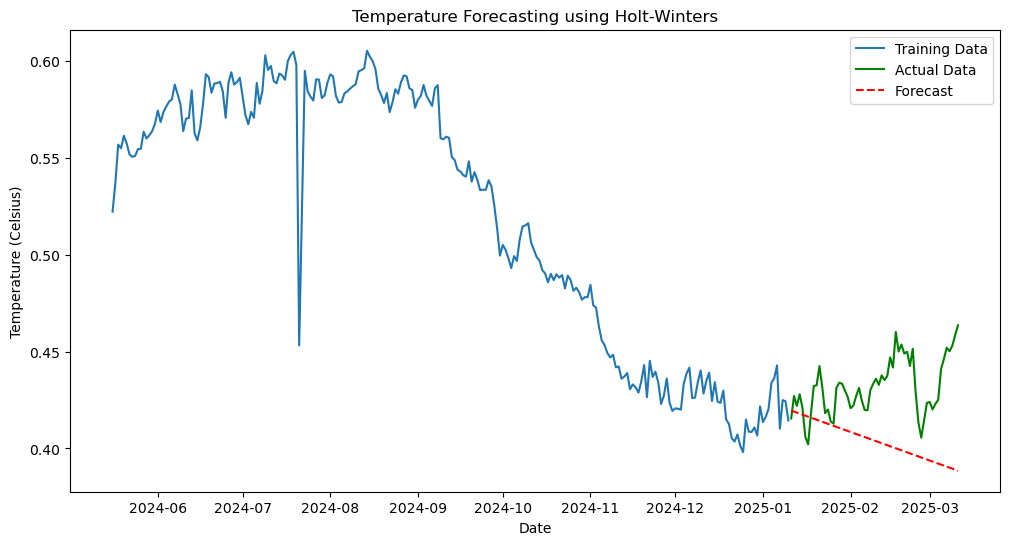




The temperature distribution graph shows a nearly normal distribution, with most values concentrated between **15°C and 35°C**, peaking around **30°C**, and a slight right skew indicating occasional extreme heat. This suggests a stable temperature pattern with fewer instances of very high temperatures. In contrast, the precipitation distribution is highly **right-skewed**, with the majority of days recording **0 mm** of rainfall, indicating frequent dry conditions. The sporadic peaks in precipitation suggest that rainfall occurs in **short bursts** rather than being evenly distributed.

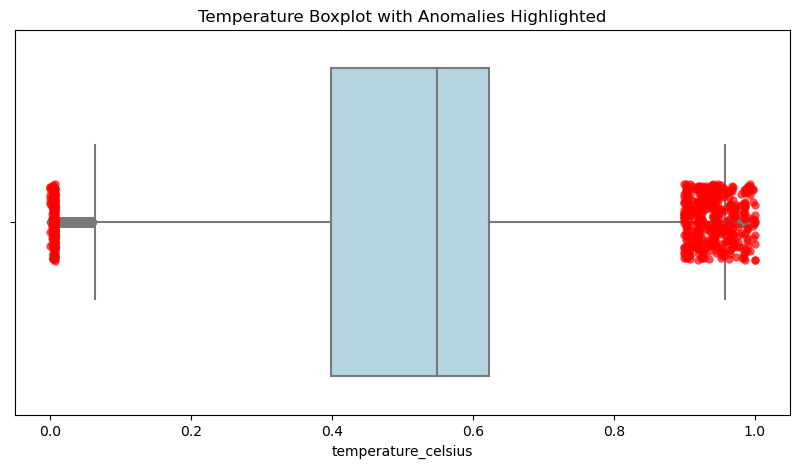
1. **Normalizing Data using MinMaxScaler:** It ensures all the numerical values are scaled between 0 and 1. It is suitable for weather data, which may have varied ranges among different features. That’s why we do not use StandardScaler which assumes normal distribution. MinMaxScaler preserves the original data distribution.
2. **Time Series Forecasting:** I used ARIMA and Holt Winters for time series forecasting. These are fficient and captures trends and autocorrelaltions. ARIMA is effective for non seasonal data while Holt winters is effective for seasonal time series forecasting. Below are the graphs showing the temperature forecasting using both the Models.



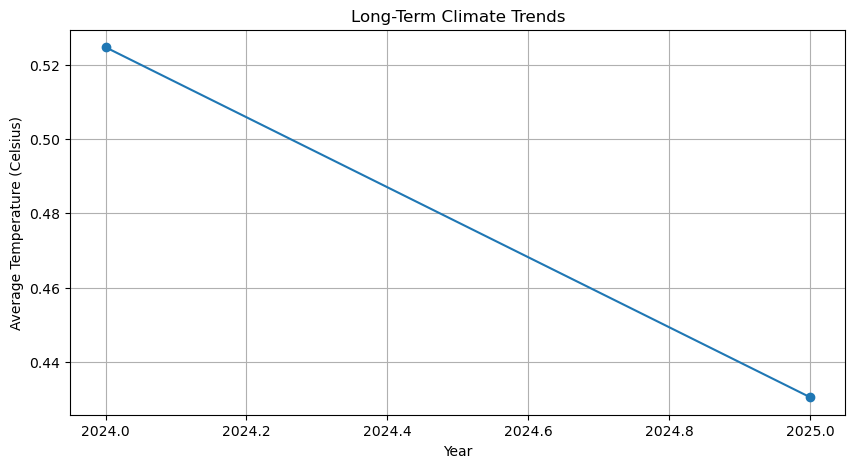


The ARIMA forecast (first graph) remains largely flat, indicating that the model struggles to capture seasonal variations and short-term fluctuations, leading to a static and less adaptive prediction. In contrast, the Holt-Winters forecast (second graph) shows a downward trend, demonstrating its ability to capture trends better than ARIMA. However, it appears to overemphasize the declining pattern and fails to adjust to the recent temperature increase observed in the actual data. While ARIMA is more suited for capturing long-term dependencies, it fails in dynamic forecasting, whereas Holt-Winters, though better at detecting trends, may exaggerate them. Both models have limitations in fully aligning with actual temperature variations, suggesting the need for further tuning or alternative forecasting approaches.

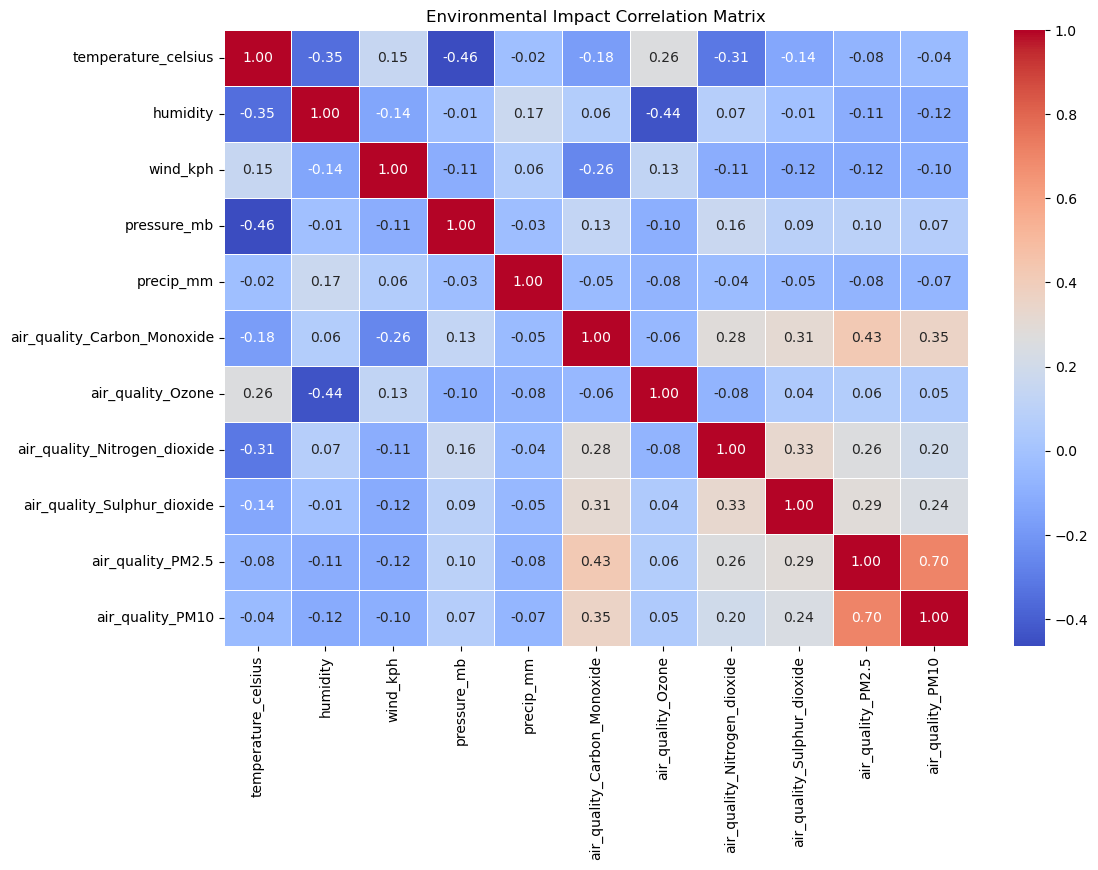
1. **Anomaly Detection:** For detecting anomalies, Isolation Forest works well by isolating rare values and Local Outlier Factor identifies anomalies based on density comparisons. The Red anomalies in the boxplot indicate extreme weather events.



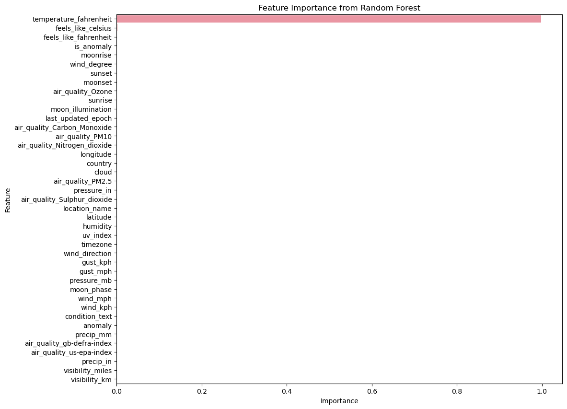
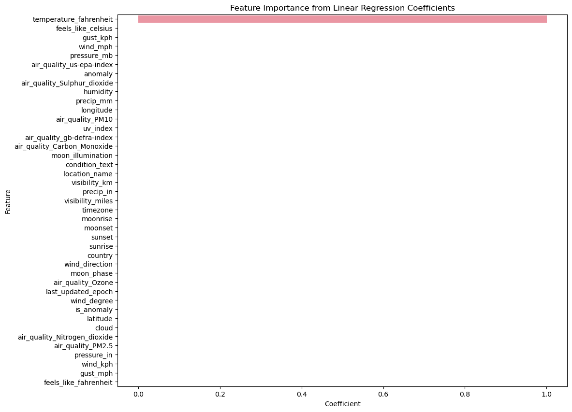
1. **Forecasting with Multiple Models:** Use of Random Forest, Linear Regression and Ridge Regression were chosen because they handle large datasets effectively. Achieving 0.99 r2 score, meant they captured almost all variance in the dataset. Thus there was no need for deep neural networks or dimensionality reduction. The r2 score was 0.99 even after performing grid search with cross validation.
2. **Climate Trend Analysis:** The graph represents long-term climate trends by showing the average temperature (in Celsius) over two years (2024 and 2025). The downward slope indicates a declining temperature trend, suggesting that 2025 is experiencing lower average temperatures compared to 2024. This could be due to seasonal shifts, climate variations, or external environmental factors. The steady decline implies a consistent cooling pattern rather than short-term fluctuations. However, further analysis over multiple years would be needed to determine if this is a temporary variation or part of a larger climate trend.



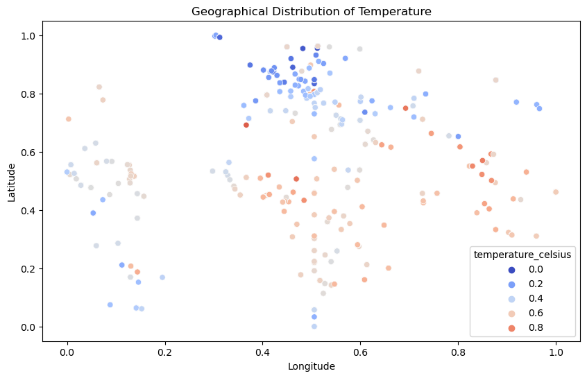
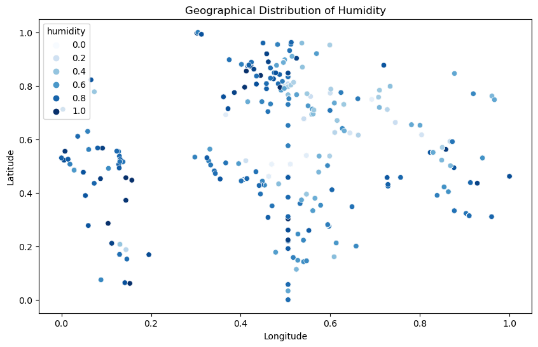
1. **Environmental impact**: The Environmental Impact Correlation Matrix highlights key relationships between climate variables and air quality indicators. Temperature shows a negative correlation with humidity (-0.35) and pressure (-0.46), indicating that higher temperatures are often associated with lower humidity and reduced atmospheric pressure, which can lead to unstable weather conditions. Air pollution indicators such as PM2.5 and PM10 (0.70 correlation) are strongly linked, suggesting that particulate pollutants often rise together, primarily from industrial emissions and vehicle exhaust. Carbon Monoxide also correlates with PM2.5 (0.43), reinforcing the impact of combustion-related activities on air pollution. Wind speed exhibits a weak negative correlation with pollutants, indicating that stronger winds help disperse pollution and improve air quality. Additionally, humidity and ozone (-0.44 correlation) suggest that higher humidity levels contribute to the breakdown of ozone, reducing its concentration.



1. **Feature Importance:** The feature importance analysis from the Random Forest model and Linear Regression (using Coefficients) shows that the model heavily relies on temperature\_fahrenheit, making it the most dominant feature while all other variables contribute almost nothing. This suggests feature redundancy, as temperature measurements (Fahrenheit, Celsius, and "feels like" temperature) are highly correlated, causing the model to ignore other important weather factors such as humidity, wind speed, air pressure, and air quality indices. This over-reliance may lead to overfitting and poor generalization. To improve the model, it is necessary to remove redundant features, re-evaluate the impact of environmental factors, and apply feature selection techniques to ensure a more balanced contribution from multiple weather variables, leading to a more accurate and reliable forecasting system**.**

1. **Geographical Weather Patterns:** The geographical distribution plots visualize temperature and humidity variations across different locations using latitude and longitude coordinates. The first plot represents temperature distribution, where warmer regions (red points) and cooler regions (blue points) are scattered, indicating spatial temperature variation. Higher temperatures seem to cluster in specific areas, while cooler regions are more widespread. The second plot shows the geographical distribution of humidity, with darker blue points indicating higher humidity levels. Humidity appears more concentrated in certain areas, suggesting that some locations experience significantly more moisture than others. Comparing both graphs, areas with higher humidity tend to have lower temperatures, reinforcing the inverse relationship between temperature and humidity.

1. **Spatial Analysis:** These spatial density plots show how temperature and data points are distributed across different locations using latitude and longitude. The first plot represents the density of recorded data points, where darker blue areas indicate lower data density, while red spots represent regions with higher concentration. This helps identify areas with more recorded observations. The second plot displays the spatial distribution of temperature, where red areas indicate higher temperatures, and blue areas represent cooler regions. These plots help visualize how temperature changes across different locations and highlight regions with extreme heat or cold.

