WEATHER FORECASTING & ANALYSIS

Abstract

A detailed analysis of weather forecasting using advanced statistics and machine learning models. The project aims to deliver accurate temperature predictions across diverse geographies. By implementing models such as SARIMA, Linear Regression, Random Forest, XGBoost, and an Ensemble approach, the study evaluates performance using metrics such as MAE, RMSE, MAPE, and R2. The resulting analysis, insights and visualizations demonstrates the forecasting accuracy.

PM Accelerator Mission

"By making Industry-leading tools and education available to individuals from all backgrounds. We level the playing field for future leaders. This is the PM Accelerator motto, as we grant aspiring and experienced PMs what they need most - Access. We introduce industry leaders, surround you with the right PM ecosystem, and discover the new world of AI product management skills."

Objective

The primary objectives of this project are :

- Analyzing historical weather data across the world and predicting future temperature trends.
- Comparing multiple forecasting models to identify the best model.

Data

- The dataset is extracted from the kaggle website.
- https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository/data

Methodology

1. Preprocessing

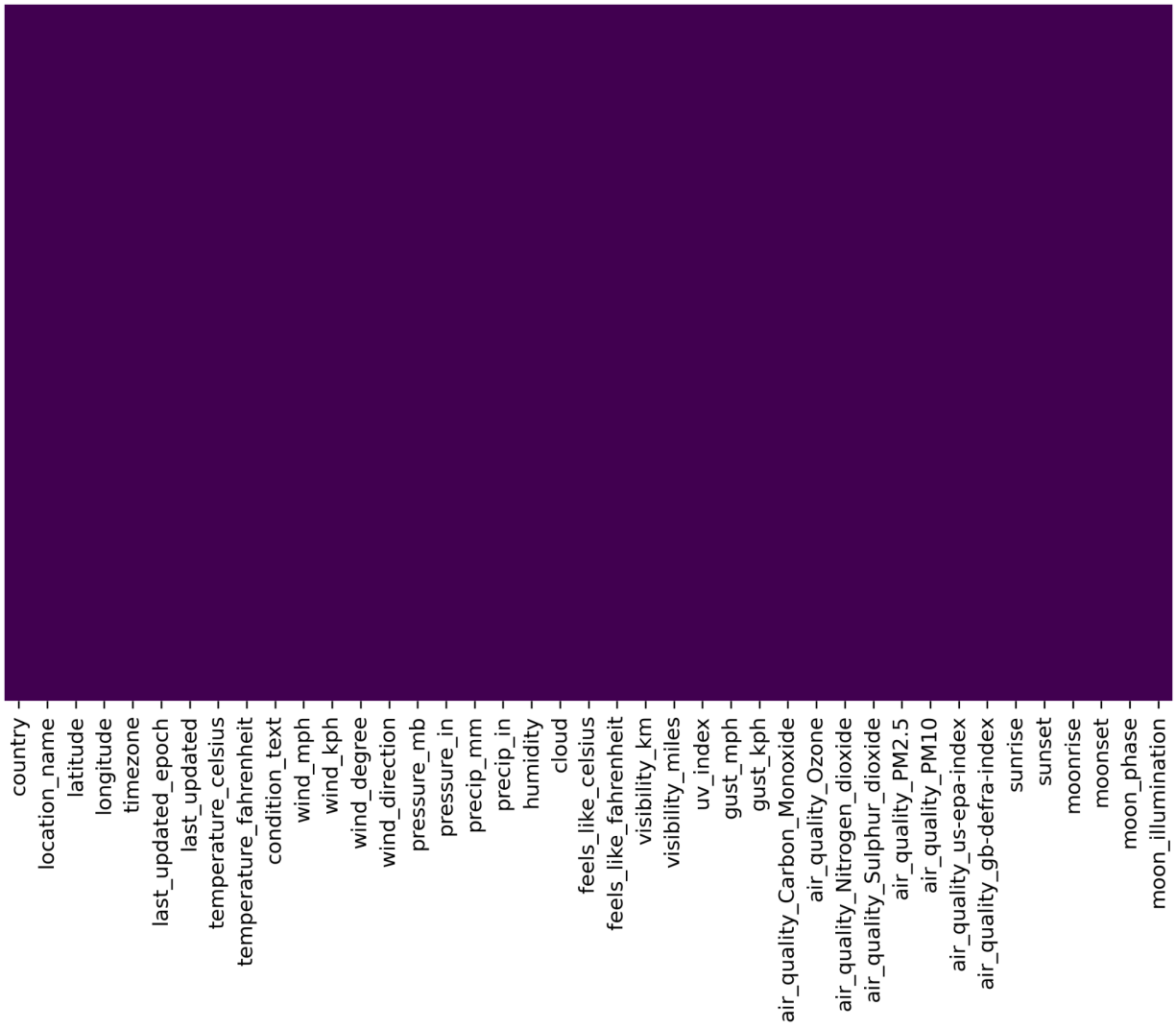
a. **Handling Missing Values :** The dataset has 0 missing values

Handling Numeric Columns

- The hanle_missing_values function is designed to process missing values in a dataset based on the data type and importance of each column.
- The columns with numerical data (float64 or int64) that are not specifically identified as special cases, the function calculates the median and fills missing values using the median. This approach is chosen to mitigate the effect of skewed data or outliers.
- When processing geographic data (latitude, longitude) the again uses the median values to fill missing entries.

Handling Textual Columns :

- The columns with categorical data the function fills missing values with most frequent value (mode)



-

b. Handling Outliers with IQR method

- The function performs outlier detection on all numeric columns grouped by country column and counts the number of outliers in each country and accumulates at total each numeric column.
- For each numeric columns, the function computes the first quartile (Q1) and the third quartile (Q3) for each country
- The Interquartile Range (IQR) is calculates as
    - IQR = Q3 - Q1
- Lower and upper bounds are established using the standard IQR multiplier
    - Lower Bound = Q1 - 1.5*IQR
    - Upper Bound = Q3 + 1.5*IQR
- Values outside these bounds are considered outliers.

- The highest outliers are in column
    1. **precip_mm** column with **9832** outliers which is 15.61%
    2. **Precip_in** column with **6693** outliers which is 10.63 %
    3. **Visibility_km** with **6334** outliers which is 9.88%.
    4. **Visibility_miles** with **6205** outliers which is 9.85%

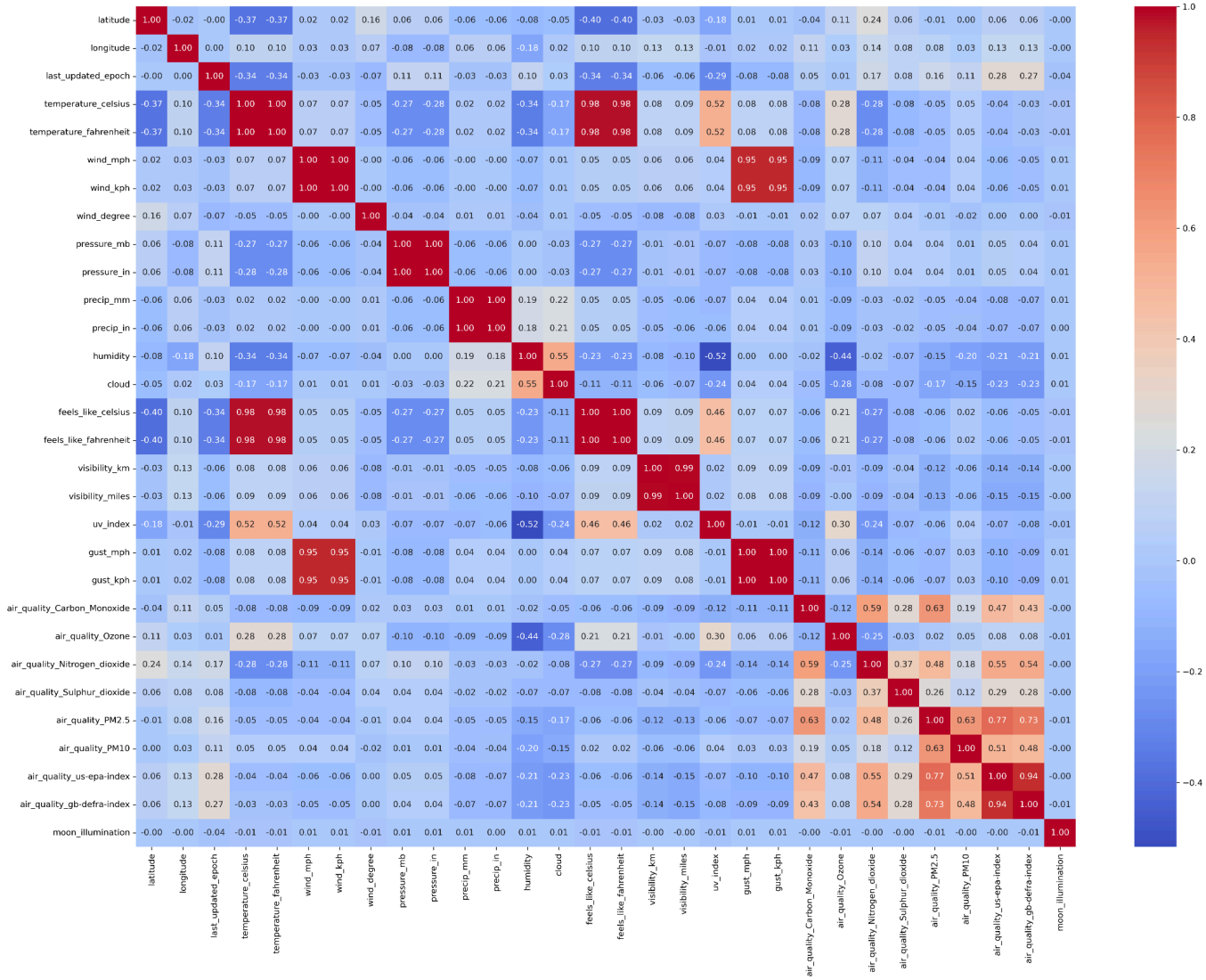
c. Normalizing Data

- The scalar is fitted on the specified numeric columns and transformed using StandardScalar.
- The StandardScaler transformation is on the z-score formula

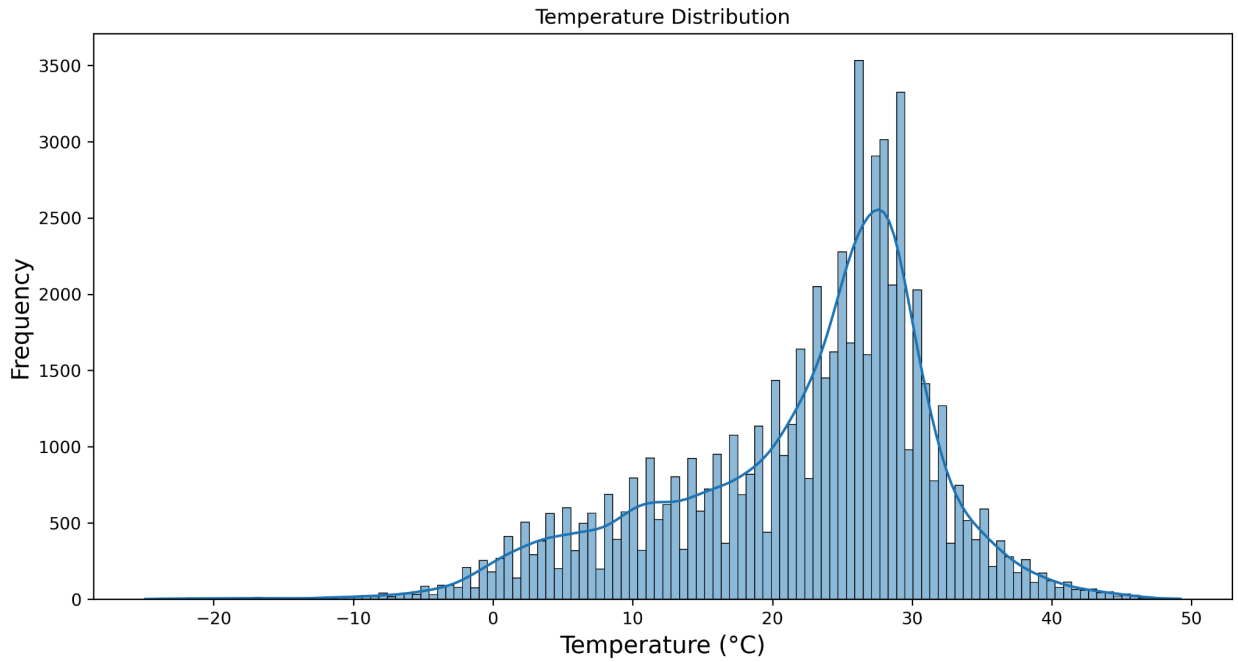
    Z = (x-mu)/sigma

2. Exploratory Data Analysis

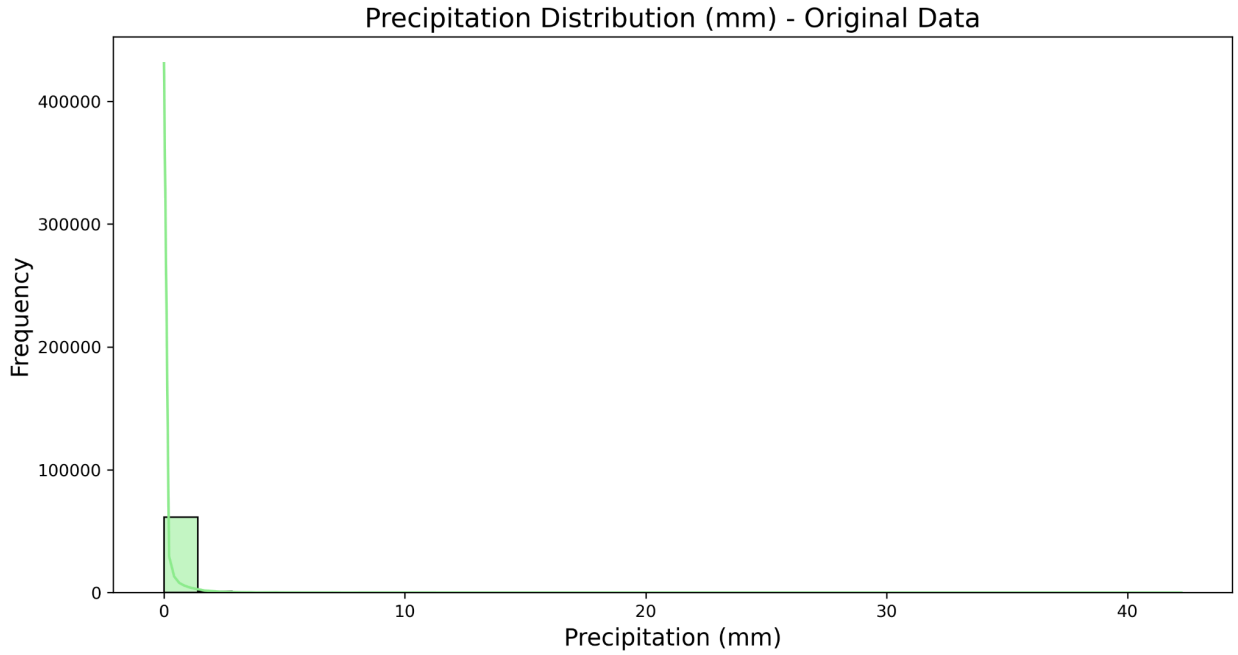
- **Strong Positive Correlations**
    - temperature_celsius & temperature_fahrenheit
    - Feels_like_celsis & feels_like_fahrenheit
    - Wind_mph & wind_kph
    - Gust_mph & gust_kph
    - Precip_mm & precip_in
    - visibility _km & visibility_miles
    - Pressure_mb & pressure_in
    - air_quality_PM2.5, air_quality_PM10, air_quality_Nitrogen_dioxide & air_quality_Sulphar_dioxide
- **Notable Correlations**
    - Temperature & humidity
    - Temperature & precipitation
- **Negative Correlations**
    - Wind_degree, temperature
    - Wind_degree, humidity
    - Last_upated_epoch
    - Moon_illumination



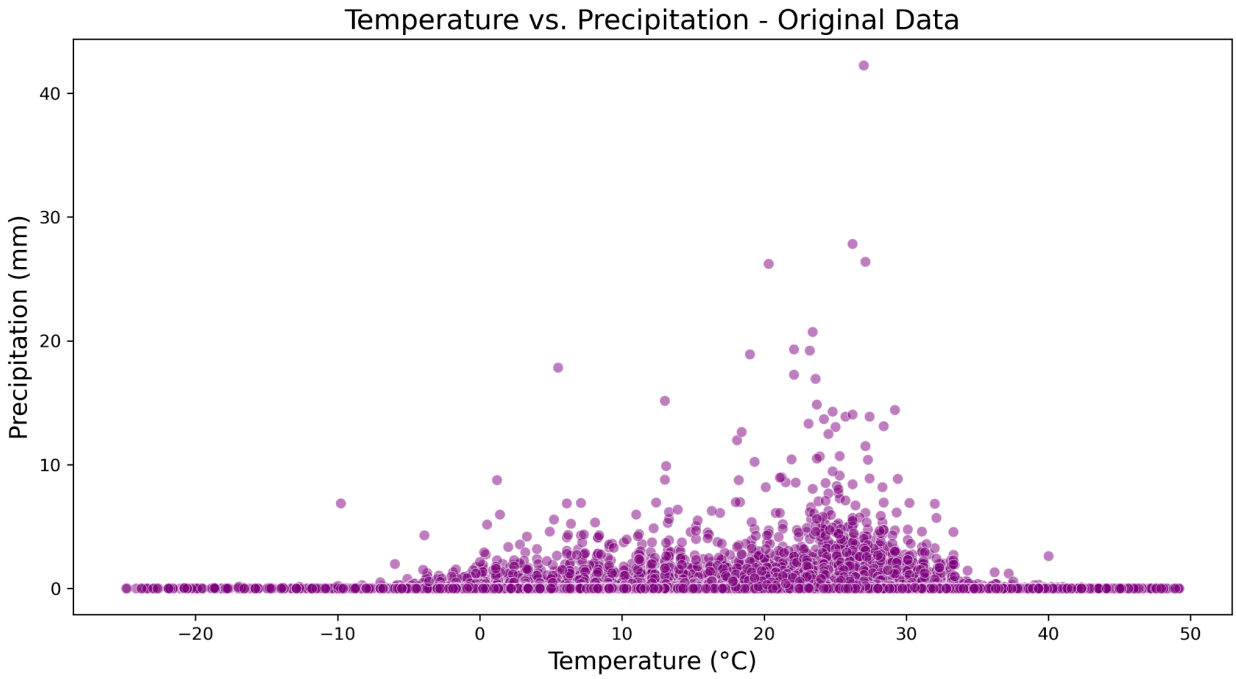
Visualization for temperature



Visualization for Precipitation



Visualization for Temperature vs Precipitation



3. Advance Exploratory Data Analysis - Anomaly Detection

- As if we have handled the outliers there were no new outliers for this anomaly detection.
- This EDA is based on the Rolling Statistics Calculation.
- A 7-day rolling window is used to calculate Rolling Mean and Rolling Standard Deviation

Rolling Mean Calculation:

$$\mu_t = \frac{1}{7} \sum_{i=t-3}^{t+3} x_i$$

Rolling Standard Deviation:

$$\sigma_t = \sqrt{\frac{1}{6} \sum_{i=t-3}^{t+3} (x_i - \mu_t)^2}$$

Z-Score Calculation:

$$z_t = \frac{x_t - \mu_t}{\sigma_t}$$

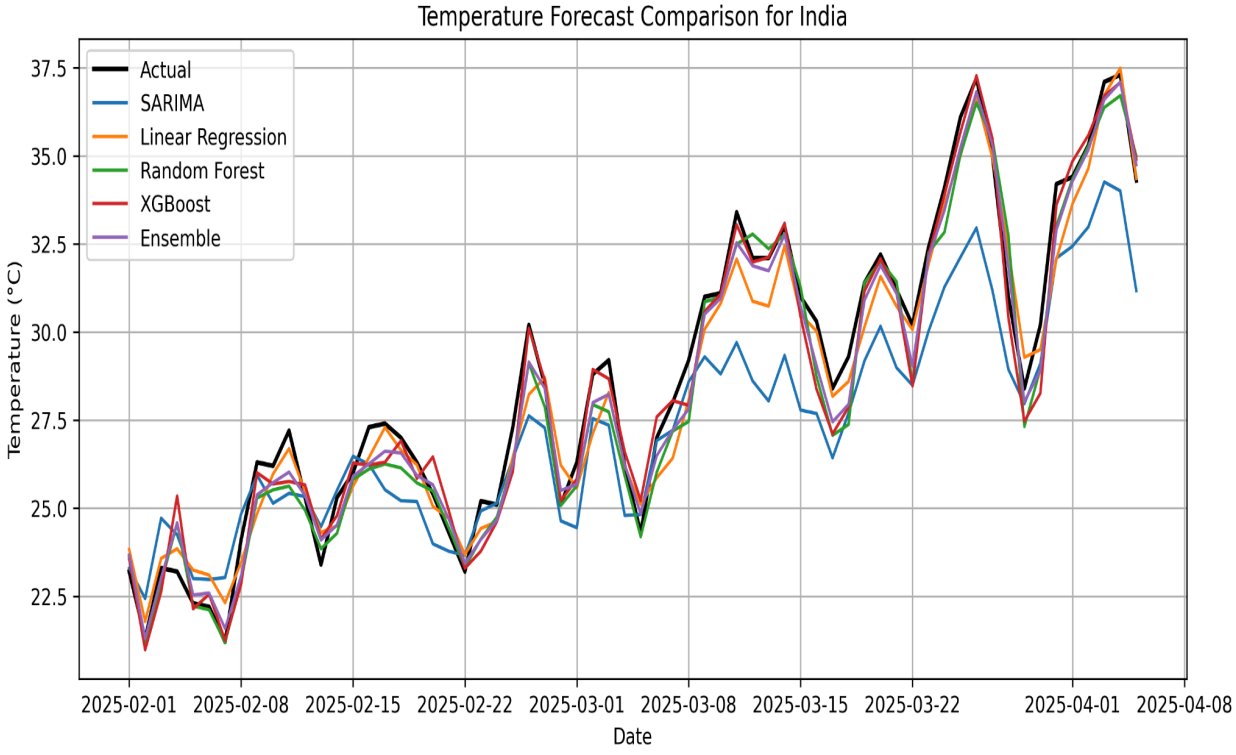
- If the absolute value of the z-score exceeds 3, that value is flagged as an anomaly for the respective columns.

4. Forecasting Models

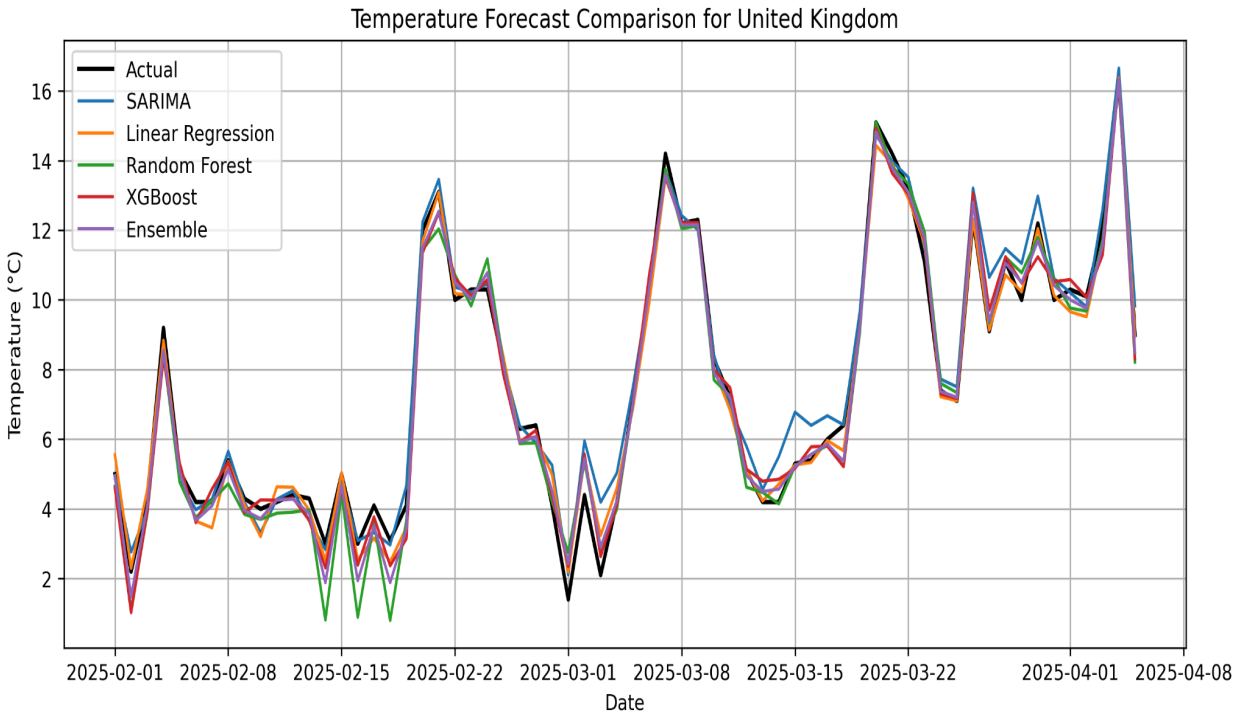
- The forecasting is done on grouping each country, resampling to daily means and creating lage features to capture temporal dependencies.

- We have used the following models for forecasting
    1. SARIMAX
    2. Linear Regression
    3. Random Forest
    4. XGBoost
    5. Ensemble

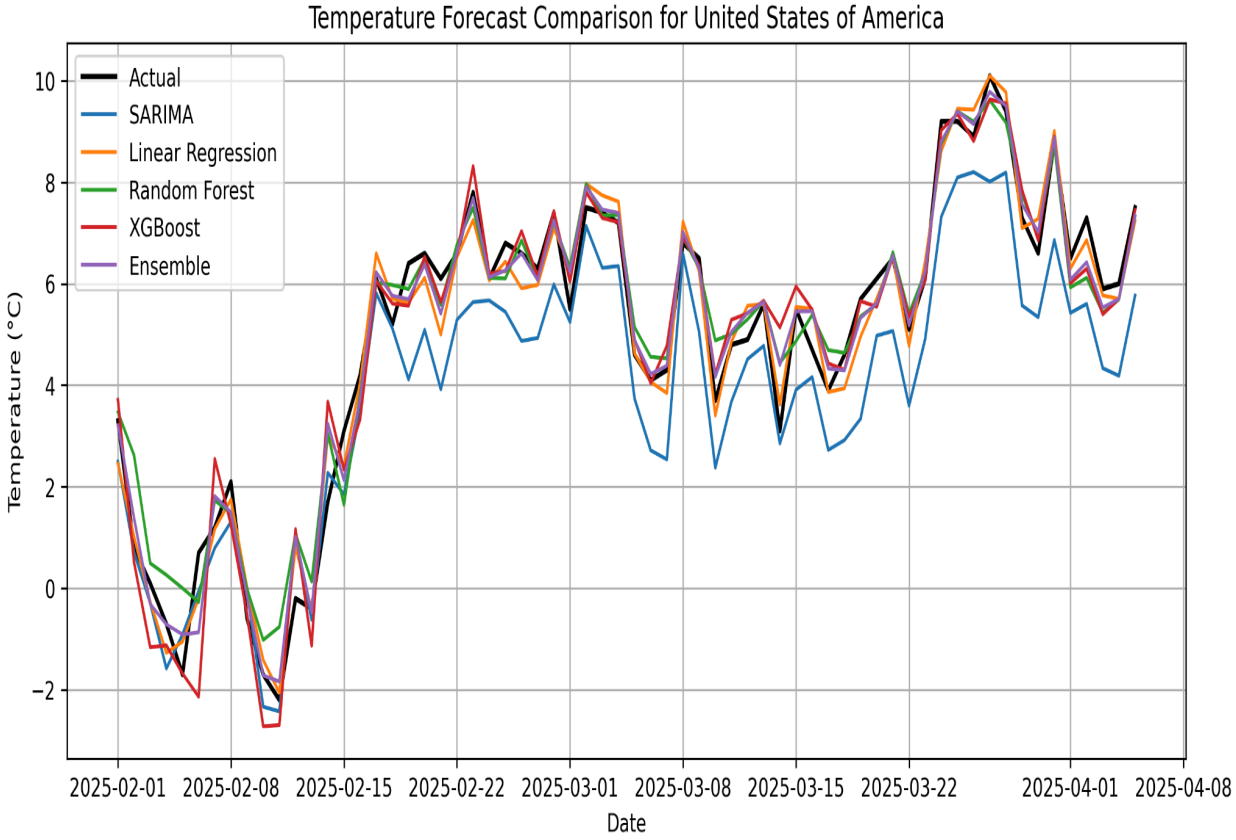
Visualization for Test data for Country 'INDIA'



Visualization for Test data for Country 'USA'



Visualization for Test data for Country 'UK'



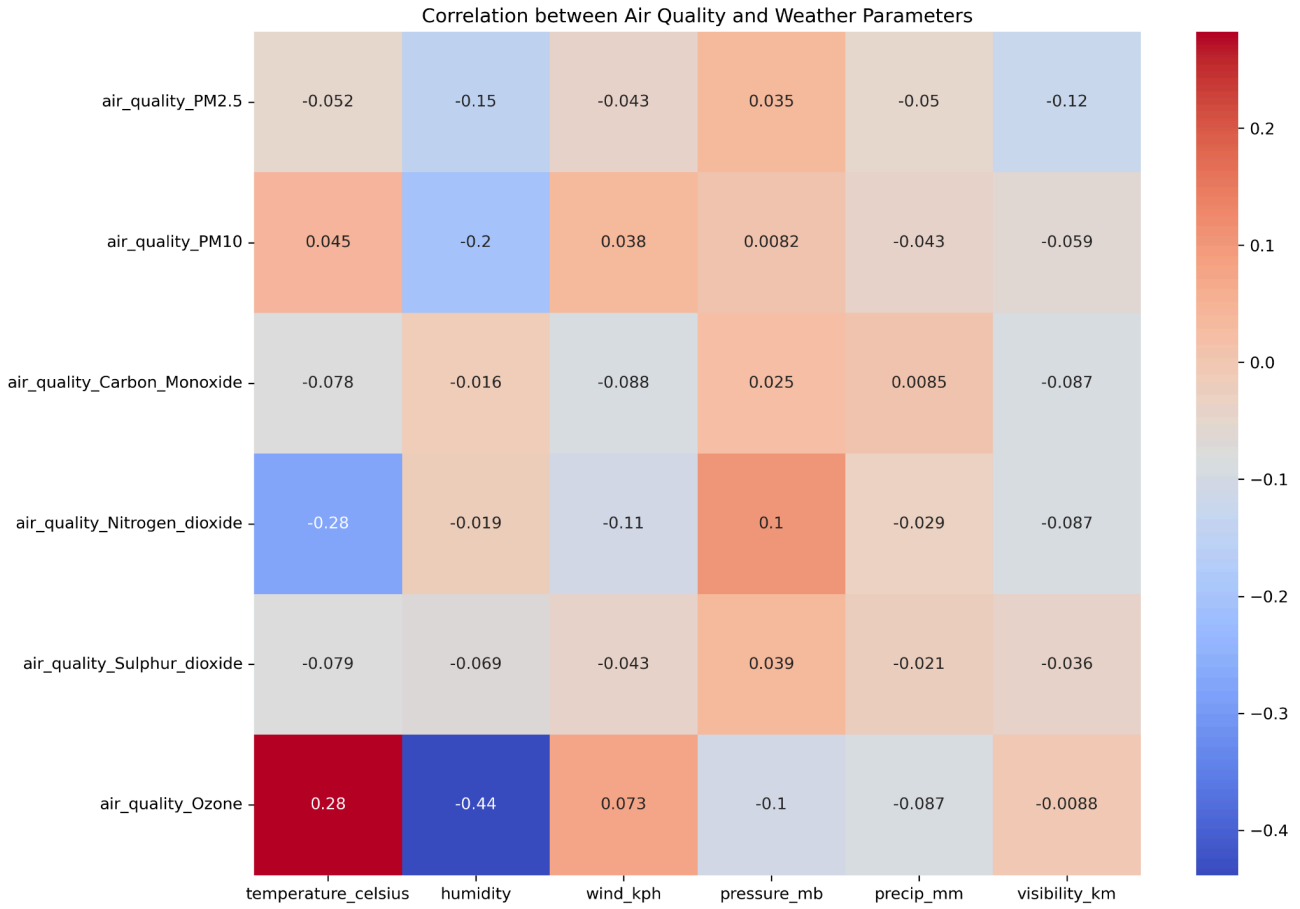
5. Evaluation Metrics

- We have used 4 evaluation metrics
    - Mean Absolute Error
    - Root Mean Squared Error
    - Mean Absolute Percentage Error
    - R Squared
- Forecasting Results India:
    - SARIMA: MAE=1.68, RMSE=2.02, MAPE=5.56%, R^2=0.77
    - Linear Regression: MAE=0.71, RMSE=0.86, MAPE=2.53%, R^2=0.96
    - Random Forest: MAE=0.70, RMSE=0.88, MAPE=2.44%, R^2=0.96
    - XGBoost: MAE=0.59, RMSE=0.79, MAPE=2.13%, R^2=0.97
    - Ensemble: MAE=0.59, RMSE=0.71, MAPE=2.08%, R^2=0.97
- Forecasting Results United States of America:
    - SARIMA: MAE=1.11, RMSE=1.26, MAPE=40.89%, R^2=0.82
    - Linear Regression: MAE=0.41, RMSE=0.50, MAPE=27.00%, R^2=0.97
    - Random Forest: MAE=0.53, RMSE=0.68, MAPE=39.61%, R^2=0.95
    - XGBoost: MAE=0.49, RMSE=0.72, MAPE=52.61%, R^2=0.94
    - Ensemble: MAE=0.39, RMSE=0.53, MAPE=30.68%, R^2=0.97
- Forecasting Results United Kingdom:
    - SARIMA: MAE=0.49, RMSE=0.65, MAPE=9.76%, R^2=0.97
    - Linear Regression: MAE=0.35, RMSE=0.45, MAPE=7.37%, R^2=0.99
    - Random Forest: MAE=0.52, RMSE=0.71, MAPE=11.53%, R^2=0.96
    - XGBoost: MAE=0.43, RMSE=0.53, MAPE=8.67%, R^2=0.98
    - Ensemble: MAE=0.38, RMSE=0.49, MAPE=8.28%, R^2=0.98

6. Unique Analyses

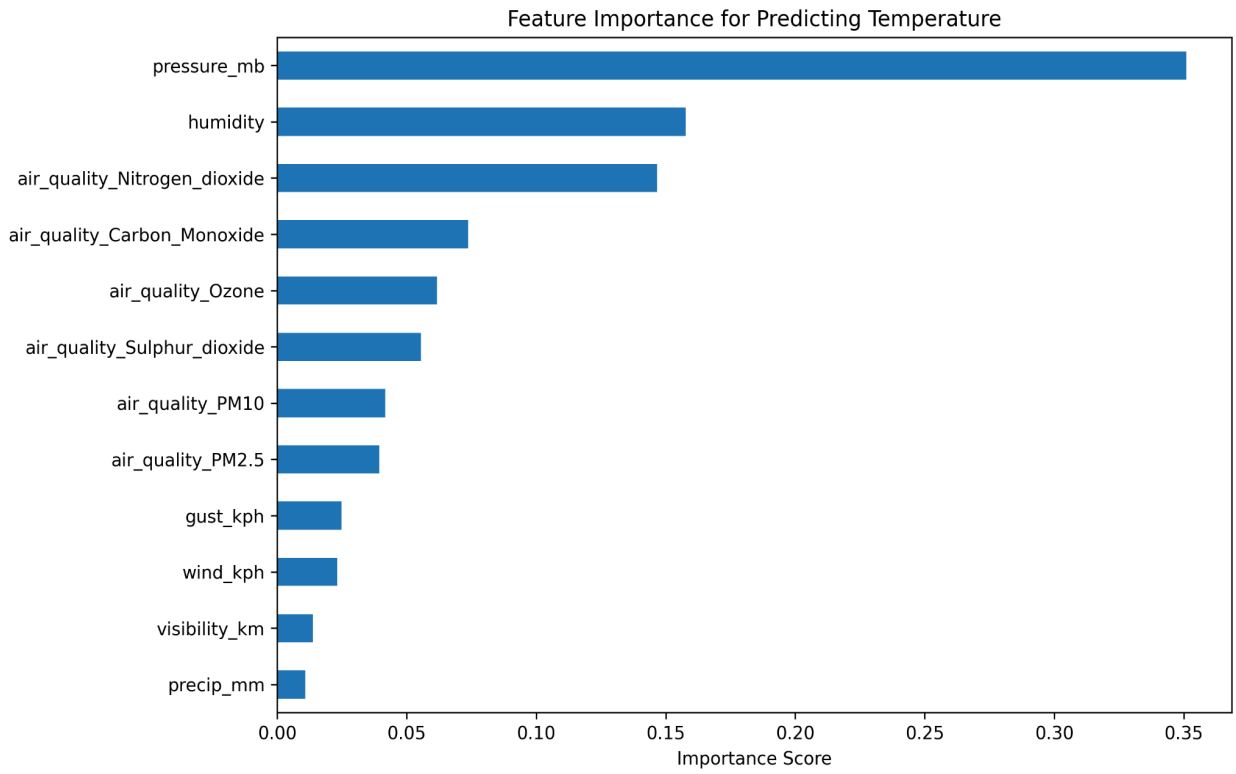
Air Qualities correlations with Weather Parameters

- There are 6 air quality metrics : PM2.5, PM10, Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide, and Ozone
- There are 6 weather parameters : temperature_celsius, humidity, wind_kph, pressure_mb, precip_mm, and visibility_km



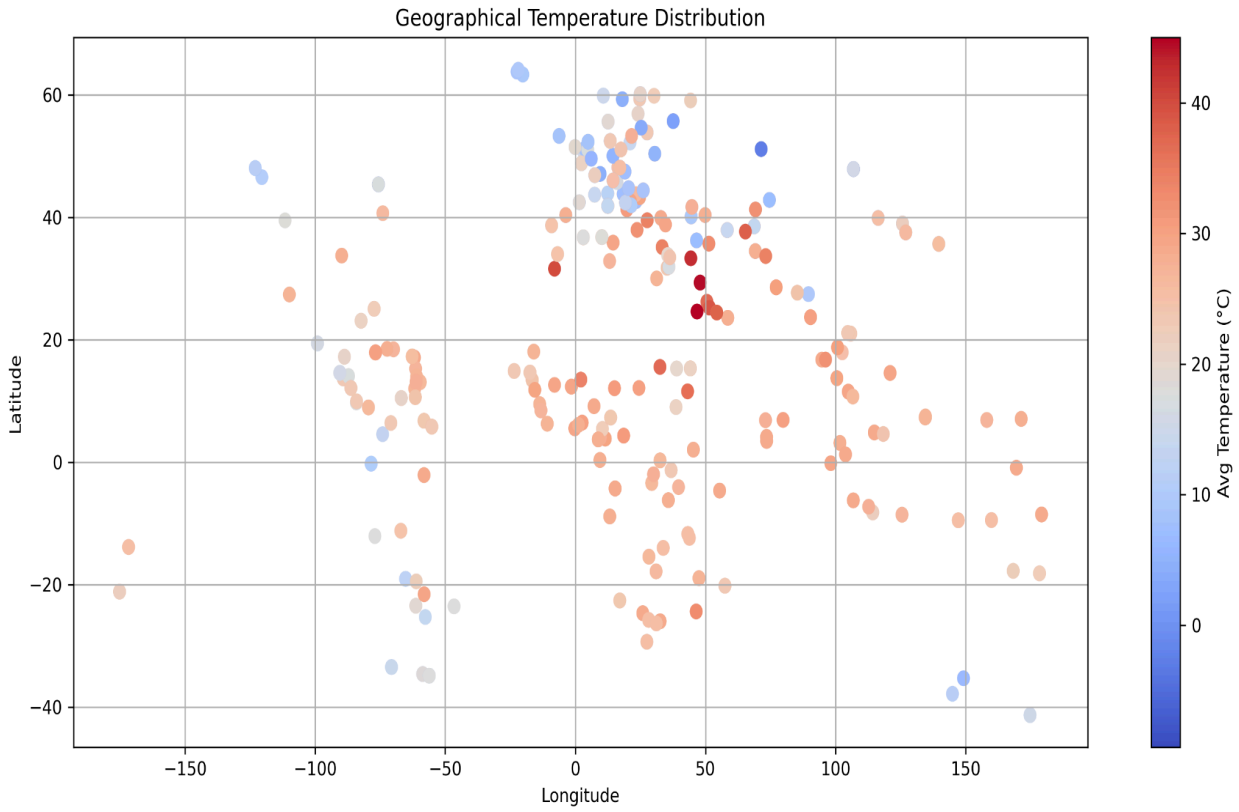
Important Features for Predicting Temperature

- The features priority is as follows :
  - Pressure_mb > humidity > air_quality_Nitrogen_dioxide > air_quality_Carbon_Monoxide > air_quality_Ozone > air_quality_Sulphar_dioxide > air_quality_PM10 > air_quality_PM2.5 > gust_kph



Visualization of Geographical patterns

- From the visualization we can conclude that
    a. -20 latitude to 20 latitude has high temperature
    b. Above 40 latitude and below -20 latitude has low temperature


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