TECH ASSESSMENT FOR DATA SCIENTISTS/ANALYST

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.

Introduction

In the analysis, we have used a GlobalWeatherRespository.csv dataset that contains weather and air quality data for various global regions between two time period - May 2024 and March 2025. The goal of the analysis was to clean the dataset appropriately and preprocess it to analyze various climate patterns. Eventually using predictive ML models to forecast the weather. In this particular report we are focusing on forecasting temperature.

Further analysis is also done to identify climate patterns and investigate the relationship between air quality and weather patterns.

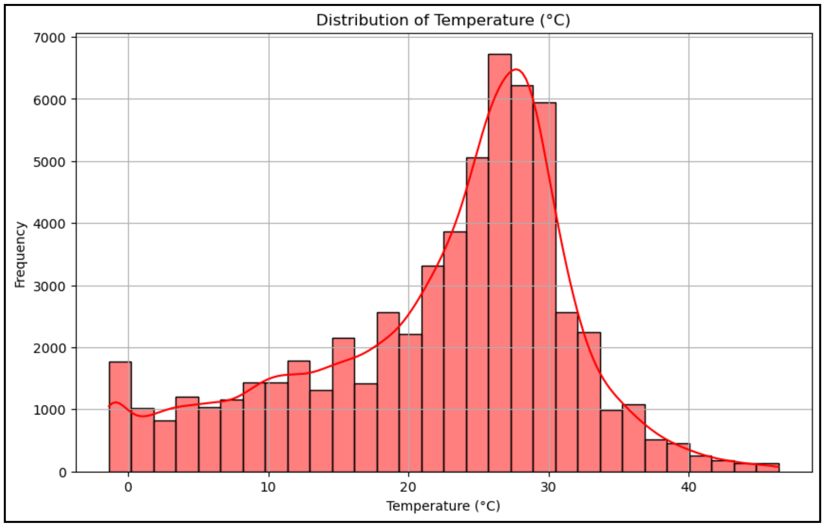
Methodology

Data Cleaning and Preprocessing

- The dataset includes 41 columns such as country, location_name, latitude, timezone, temperature (Celsius), humidity, wind speed, air quality and some more..
- Initially the goal was to identify the missing values int he dataset. We used isnull().sum but no null values were found. So we used a custom function with all the potential missing value indicators like N/A, NA, missing, unknown, ? Etc. and then this were replaced by NaN to maintain uniformity. Finally numeric columns were replaced with the median and categorical columns were swapped with the most frequent values using the SimpleImputer function.
- Next, the outliers were handled using the IQR method, Values below the 25th Percentile and above 75th Percentile were clipped.
- At the end, normalization was performed, using the standard scaler function to convert the values into z-score, and the cleaned dataset was stored as GlobalWeatherRepositiory_cleaned.csv.

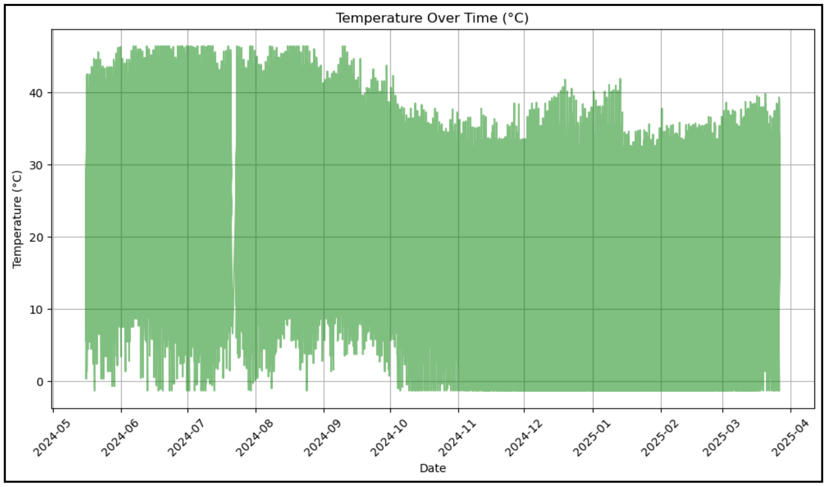
Exploratory Data Analysis(EDA)

Temperature Distribution



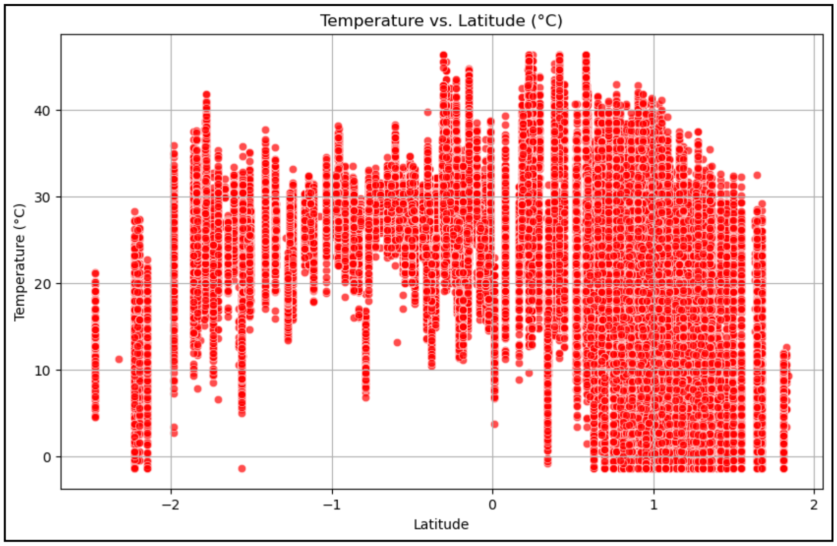
The Temperature distribution was done by plotting histogram. This shows a rough mean of around 22C. The temperature mostly ranges from 10C to 35C.

Temperature Over Time



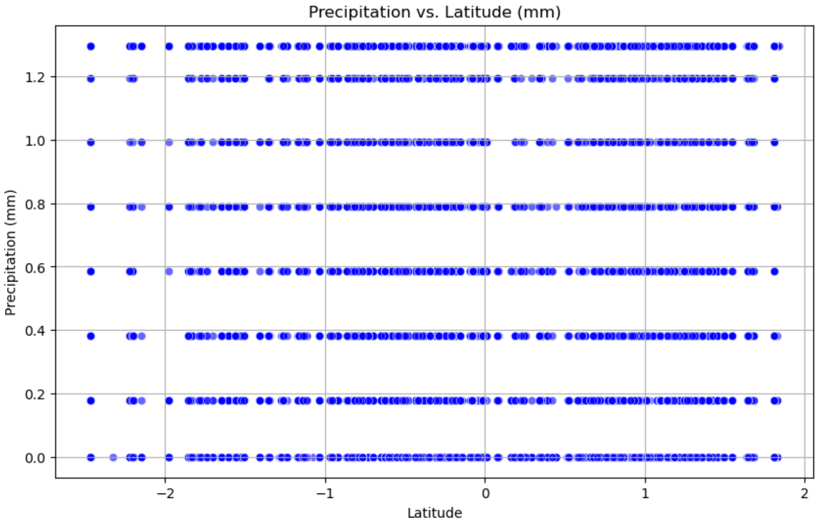
There seem to be some fluctuations between May 2024 and March 2025, with temperatures in May 2024 are higher unto 40C compared to early spring March 2024 for the Northern Hemisphere.

Temperature Vs. Latitude



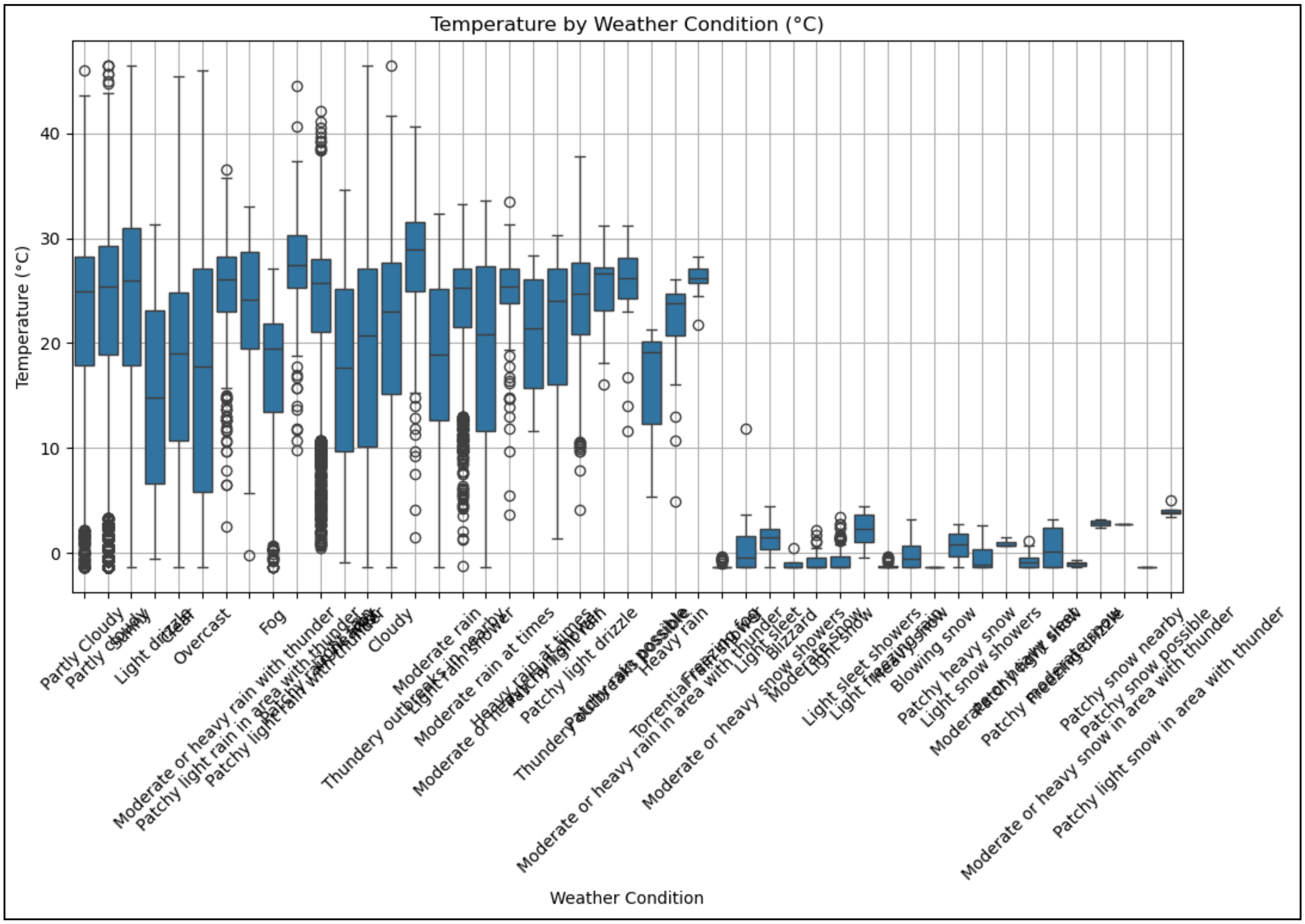
The Scatter plot clearly shows that the temperature decreases as we move away from the equator. Neear the equator the temperature often exceeds 30C and drops to 0C as we move higher reflecting an expected latitude based gradient .

Precipitation Vs. Latitude



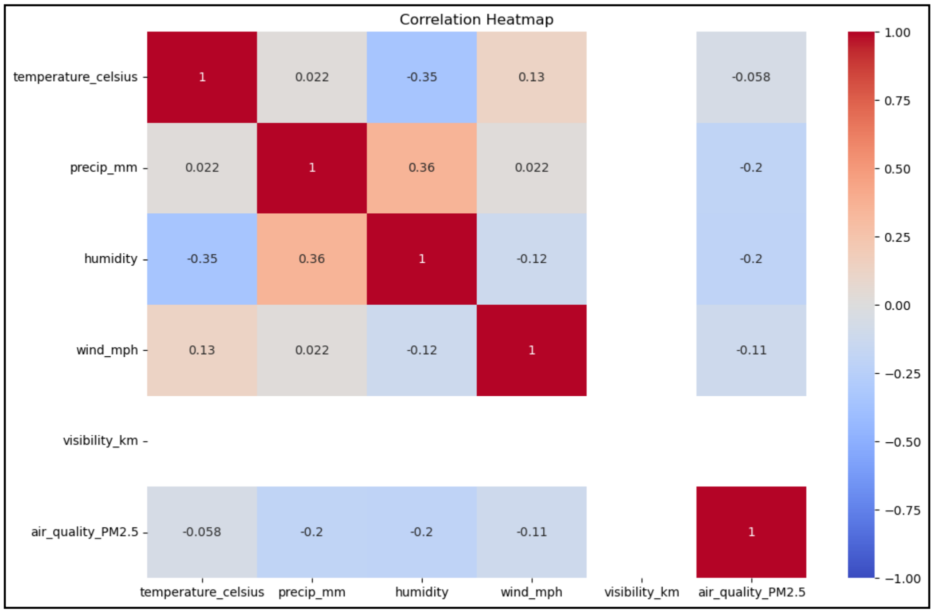
Precipitation doesn't show any clear pattern, which means that the precipitation has no ties with the latitude.

Temperature by Weather Condition



In this not so clear boxplot (there were many parameters), we can see the significant variation across weather conditions. Sunny and clear conditions have higher median temperature (around 30C), and Blizzard and Heavy snow have lowest temperature (around 0 C),

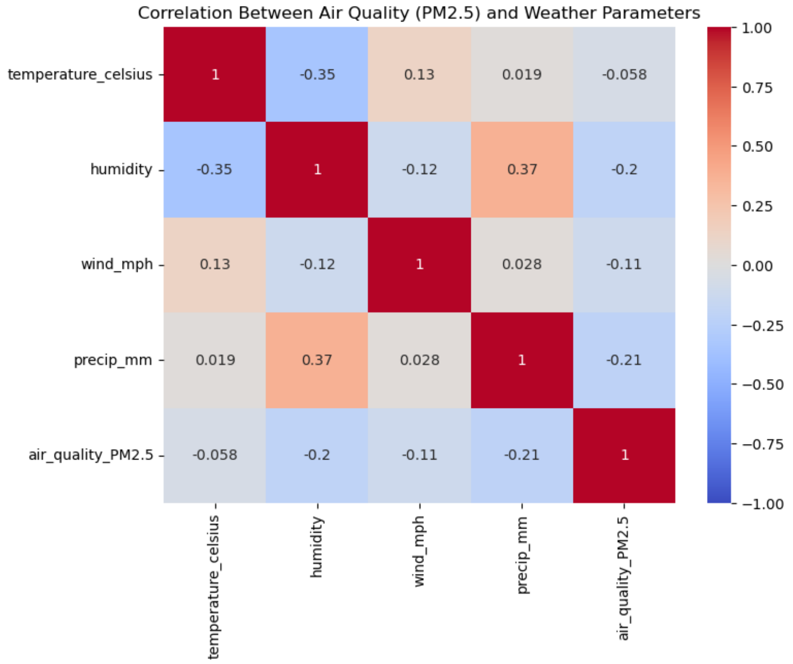
Correlation Heatmap



Temperature and humidity are negatively correlated (-0.35), meaning higher temperatures lower humidity.
Precipitation and humidity are positively correlated (0.36), indicating more rain increases humidity.
Precipitation reduces PM2.5 levels (-0.2).
Warmer temperatures slightly improve air quality (-0.058).

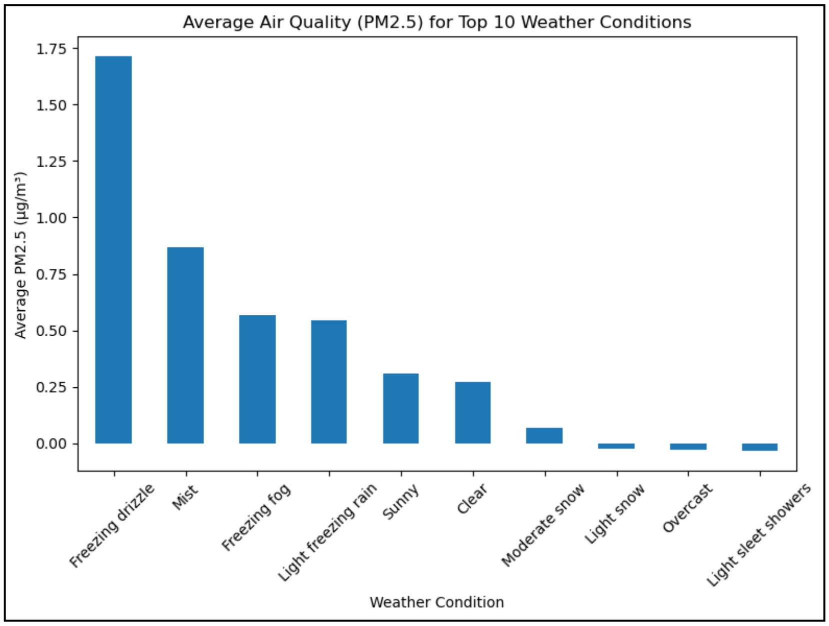
Air Quality

Analyze air quality and its correlation with various weather parameters.



Wind speed has a negative correlation with PM2.5 (-0.11) which indicates that higher wind speeds reduce pollution.
Humidity has a positive correlation with PM2.5 (0.2) which suggests that higher humidity traps pollutants.
Precipitation has a negative correlation with PM2.5 (-0.2), confirming that precipitation reduces pollution.

Air Quality by Weather Condition



Mist had the highest average PM2.5 (1.75 µg/m³), followed by Light freezing drizzle and Fog (around 0.75 µg/m³), due to high humidity and low wind.
Light sleet showers” and Overcast had the lowest PM2.5 levels (around 0.1 µg/m³) among the top 10.

Predictive Modelling

For the purpose of this analysis three different machine learning models were used to predict temperature based on various weather and environmental features. The three models were Linear Regression, Support Vector Regression (SVR), and Random Forest. These models were then evaluated and compared based on performance metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R² score.

Model Performance

Model	MSE (°C²)	MAE (°C)	R²
Linear Regression	0.94	0.75	0.19
SVR	0.96	0.74	0.17
Random Forest	0.16	0.26	0.86

Random Forest worked significantly better than Linear Regression and SVR, with an R^2 of 0.86, which means that it explained 86% of the variance in temperature. Its MAE of 0.26°C suggests extremely accurate predictions, and its MSE of 0.16 °C^2 guarantees minimal error

Random Forest achieved:
83% lower MSE than Linear Regression (0.94 → 0.16)
65% lower MAE than Linear Regression (0.75 → 0.26)
352% improvement in R^2 over Linear Regression (0.19 → 0.86)

Why did Random Forest performed best?

Captured Non-Linear Interactions: Unlike Linear Regression, Random Forest is able to capture intricate, non-linear interactions between features.
Feature Interactions: The model was able to learn interactions well, such as how humidity and weather condition together influence temperature.
Resilient to Overfitting: By the averaging of multiple decision trees, it maintained robust generalization without overfitting.
Improved Handling of Categorical Features: Random Forest was able to handle categorical features (e.g., weather condition) well in its prediction.

Conclusion

The dataset GlobalWeatherRepository.csv provided a robust framework to understand the climate patterns, air quality and temperature forecasting. The EDA revealed that mean temperature was around 22.14C and followed a latitudinal and seasonal pattern, with Asia and Africa being the warmest continents and North America and Europe being on the cooler side.
Precipitation helps in reducing PM2.5 levels h when present and also higher wind speeds helps in improving the air quality. Mist is linked with high PM2.5 levels.
Random Forest emerged as the best model for temperature forecasting achieving an R^2 of 0.86, an 83% reduction in MSE, and a 65% reduction in MAE compared to Linear Regression and performed better than SVR too.