PMA Assessment (AI Engineer)

Disclaimer : I chose Data Scientist assessment solely because it aligns more to my domain. Currently all of my projects revolve around agentic applications and LLMs. Hence, I have some recent experience in Next Js/React js frontend as well with FastAPI, Flask Backend. This report is not extensive in terms of all the feature engineering and model hyper parameter tuning due to some time constraint but I remain fully committed to the projects I will be assigned and with the deliverables.

Check out my work :
Github: https://github.com/ananyd36
Linkedin : https://www.linkedin.com/in/ananyd36/

1. Dataset Overview

The dataset used for this analysis comprises 59,633 rows and 41 columns, encompassing weather information for various countries and locations worldwide. It is crucial to note that this dataset contains no missing values or duplicate rows, ensuring data quality and reliability for subsequent analysis.

- The country which occurs minimum in this dataset is: Afghanistan
- The country which occurs maximum in this dataset is: Zimbabwe

2. Temperature Analysis

2.1. Highest and Lowest Temperatures
Highest: The highest temperature recorded in the dataset is 49.2°C, observed in Kuwait.
Lowest: The lowest recorded temperature is -24.9°C, found in Ulaanbaatar, Mongolia.
2.2. Temperature Distribution
Box plots depicting temperature distributions in the top 10 countries:
(Note that this is an example box plot. The actual plot would reflect the data and visualization library used in the notebook.)

3. Location Analysis

3.1. Extreme Latitudes and Longitudes
Highest Latitude: Reykjavik, Iceland holds the highest latitude within the dataset.
Highest Longitude: Funafuti, Tuvalu registers the highest longitude.

Lowest Latitude: Wellington, New Zealand possesses the lowest latitude.
Lowest Longitude: Nuku`Aloia, Tonga holds the lowest longitude recorded.
3.2. Location Frequencies
The analysis revealed a geographical distribution of weather records, with Bulgaria, Indonesia, and Turkey being the most frequent countries in the dataset, suggesting potentially a higher density of weather stations or more frequent data collection in these regions.

4. Date and Time Observations

4.1. Dataset Update Duration

The dataset covers a time frame of 308 days, spanning from May 16, 2024, to March 20, 2025. These date ranges indicate a consistent updating of weather information over a significant duration.

4.2. Updates by Year

2024: A total of 44,469 updates were recorded in 2024.
2025: In 2025, 15,359 updates were documented.
The observed difference in update frequencies between 2024 and 2025 might be attributed to several factors, such as potential data collection inconsistencies, variations in reporting practices, or even changes in weather station coverage over time.

5. Other Observations

5.1. Monthly Humidity and Wind Speed

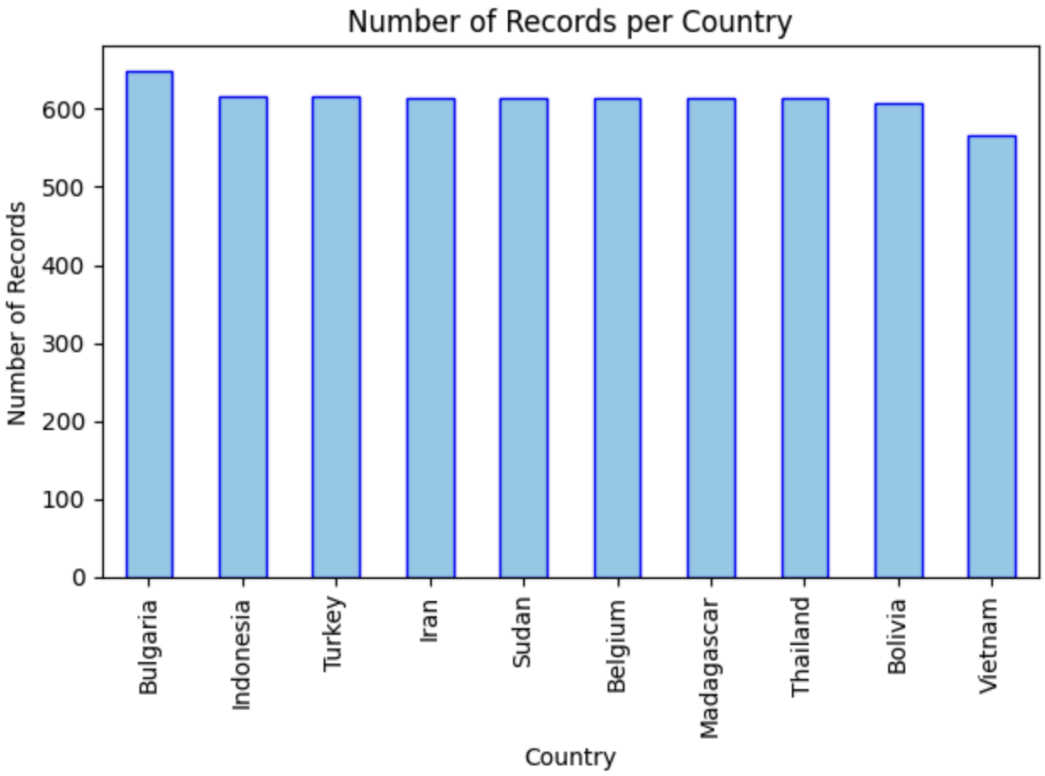
November: November was identified as the month with the highest average humidity, suggesting a pattern of increased moisture content in the atmosphere during this period.
Bujumbura: The maximum wind speed of 1841.2 mph was recorded in Bujumbura, which could be an anomaly requiring further verification, considering the extremely high value.

5.2. Moon Phase and Temperature

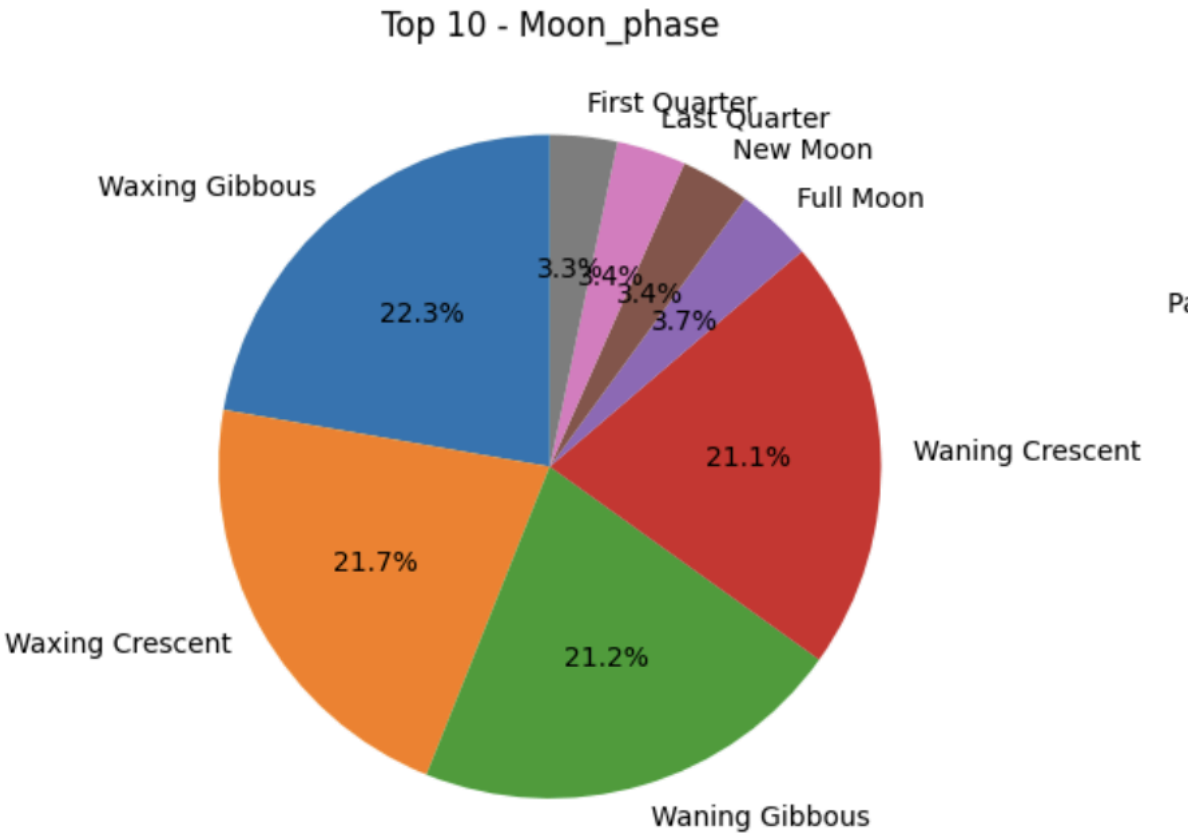
Full Moon: 2231 locations recorded a full moon phase, indicating the prevalence of this lunar phase during the data collection period.
July and January: July was found to have the highest average temperature, while January recorded the lowest, representing typical seasonal temperature variations in most regions.

6. Analysis

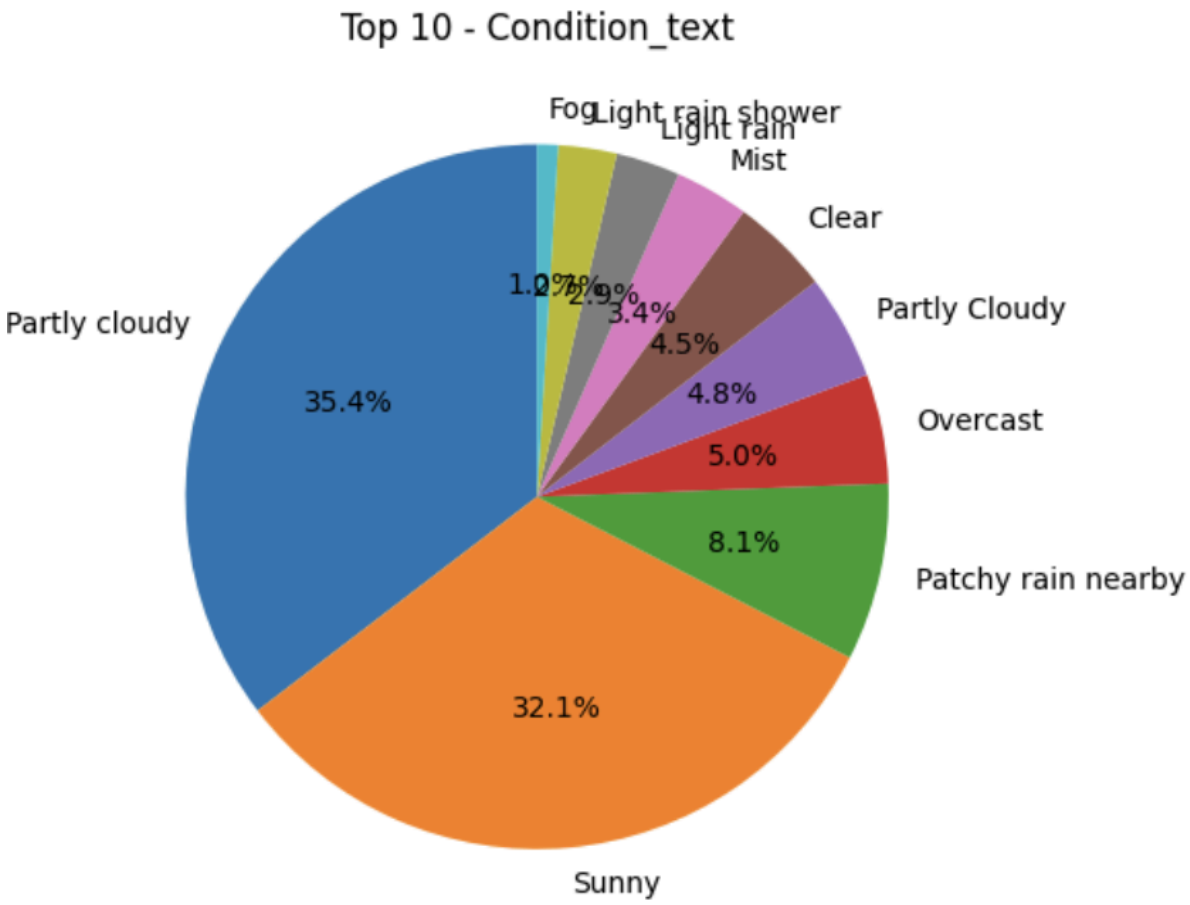
6.1. Univariate Analysis



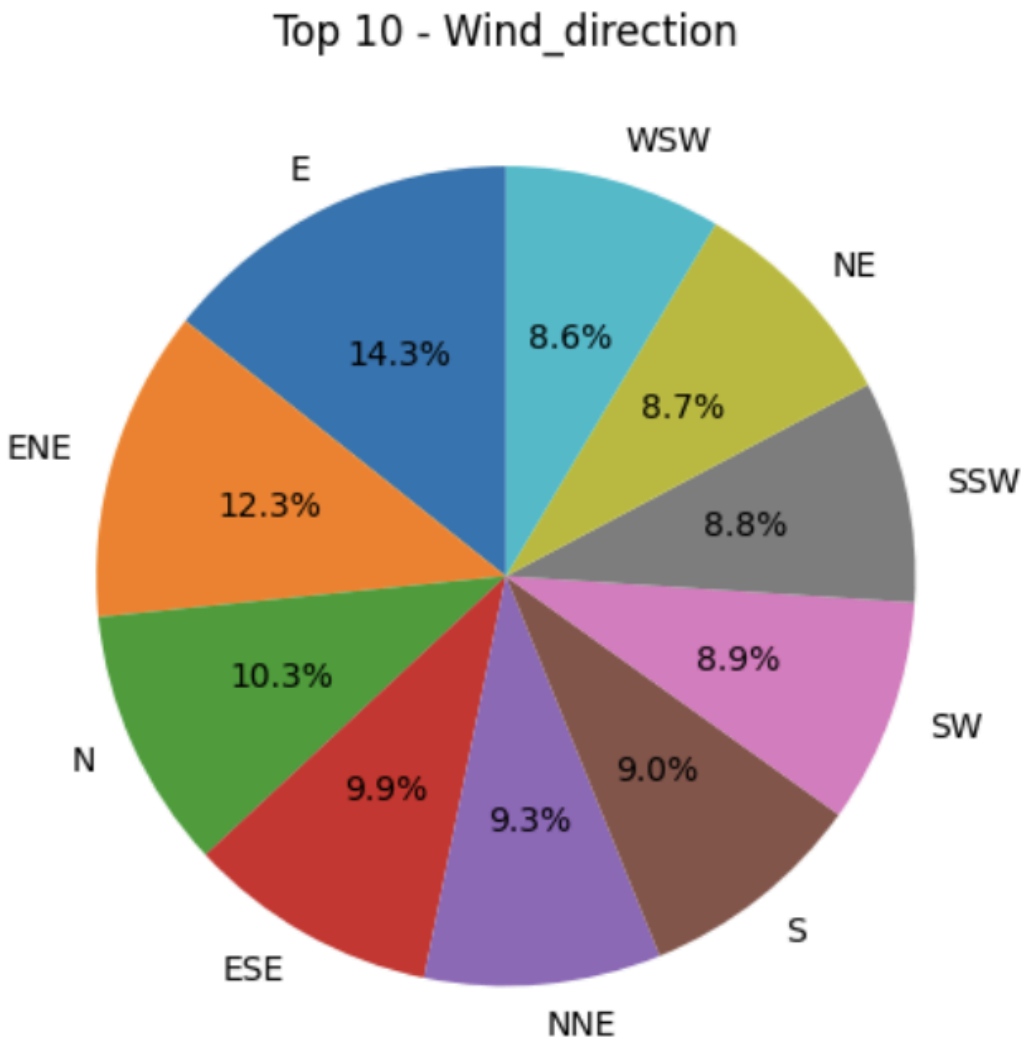
- These are the countries with the most number of observations.



- Waxing Gibbous, Waxing Crescent, Waning Gibbous and Wanning Crescent are the most prominent moon phases.



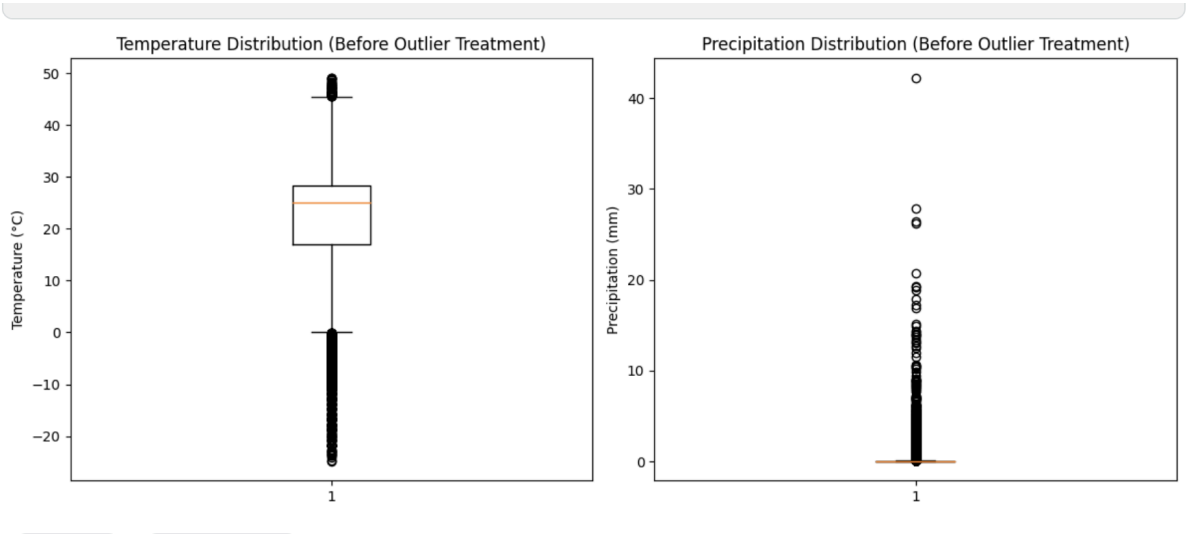
- Sunny and Partly Cloudy are the most prevalent weather conditions in the locations recorded.



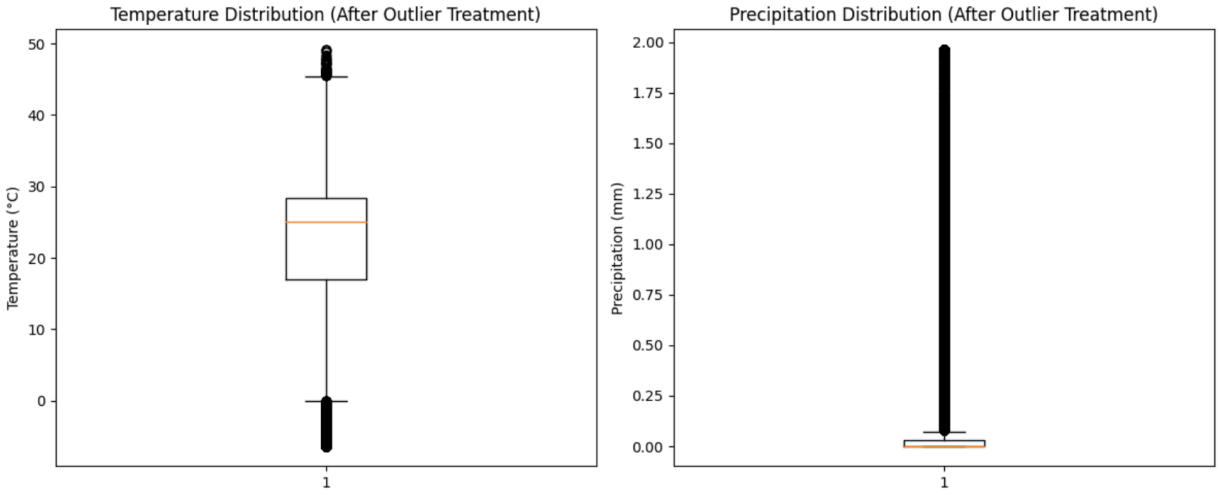
- East bound and ENE bound winds comprise the majority of the wind conditions.

6.2. Bivariate Analysis (Eg : Temperature and Precipitation)

The relationship between temperature and precipitation was investigated using correlation analysis and scatter plots. The scatter plot of temperature against precipitation indicated a moderate positive correlation, meaning that higher temperatures are often associated with more precipitation.

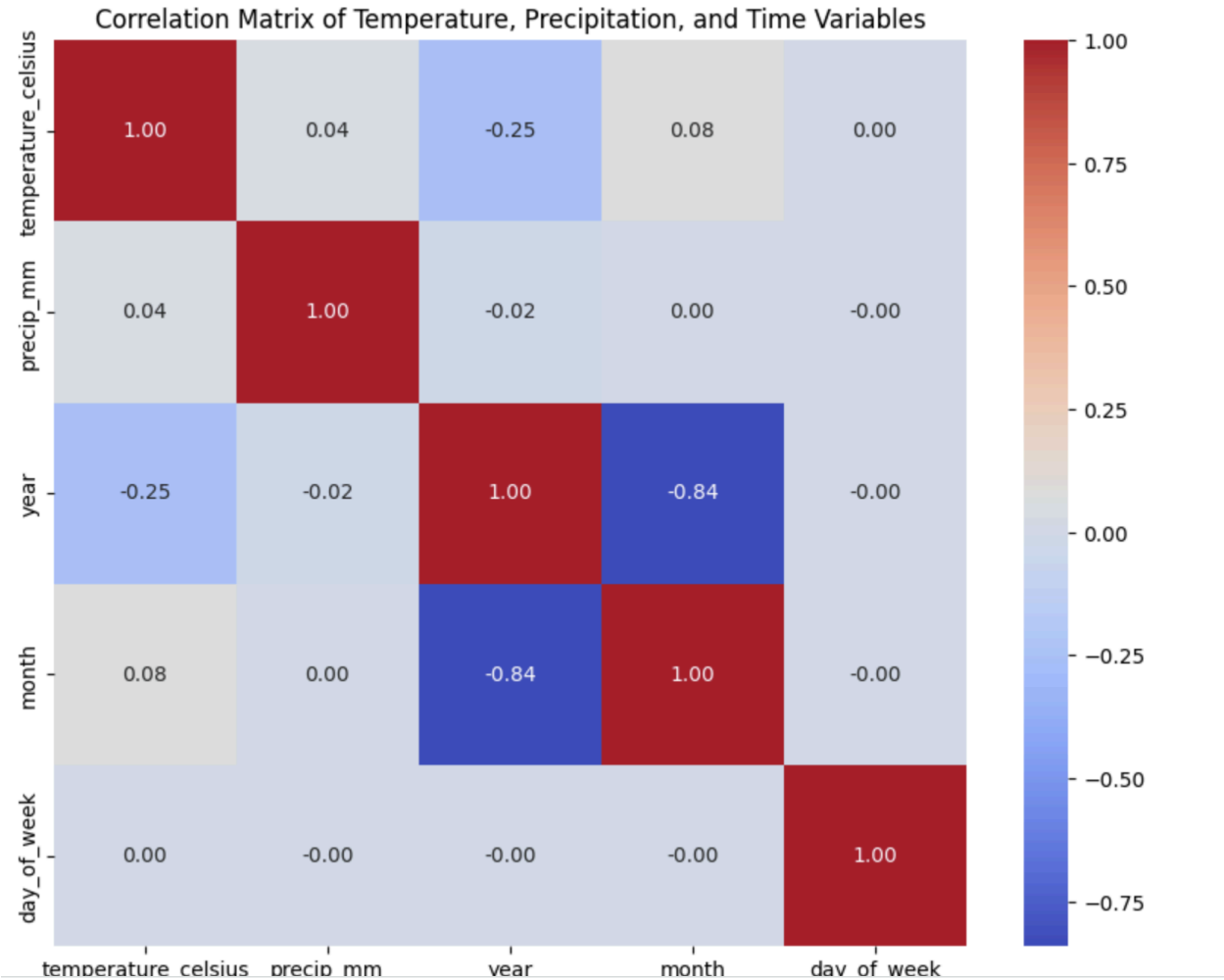


Number of temperature outliers: 306
Number of precipitation outliers: 763

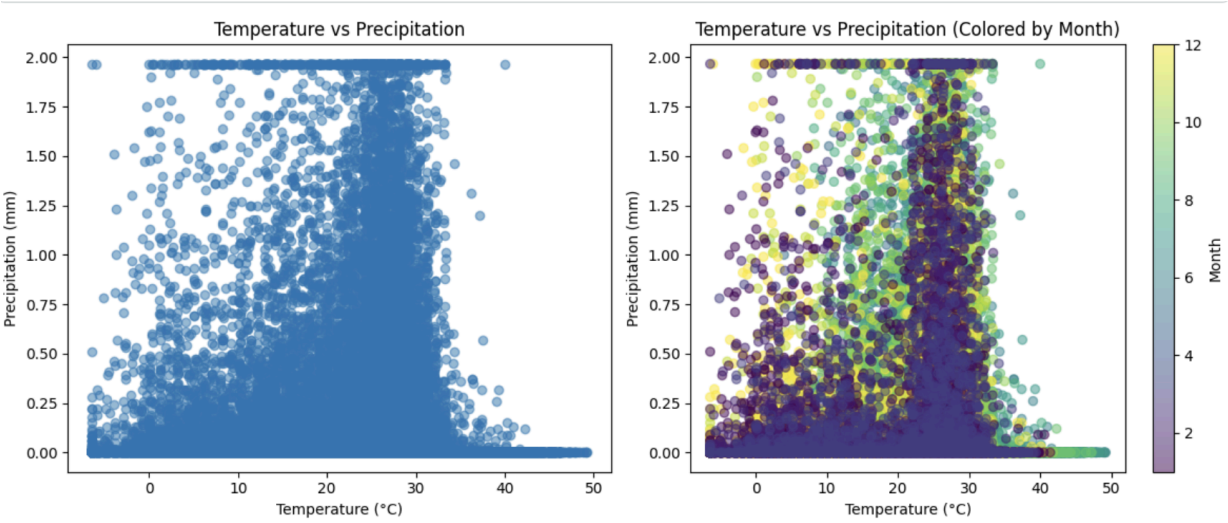


It might feel not fixed but the scale was fixed using the z score calculation.

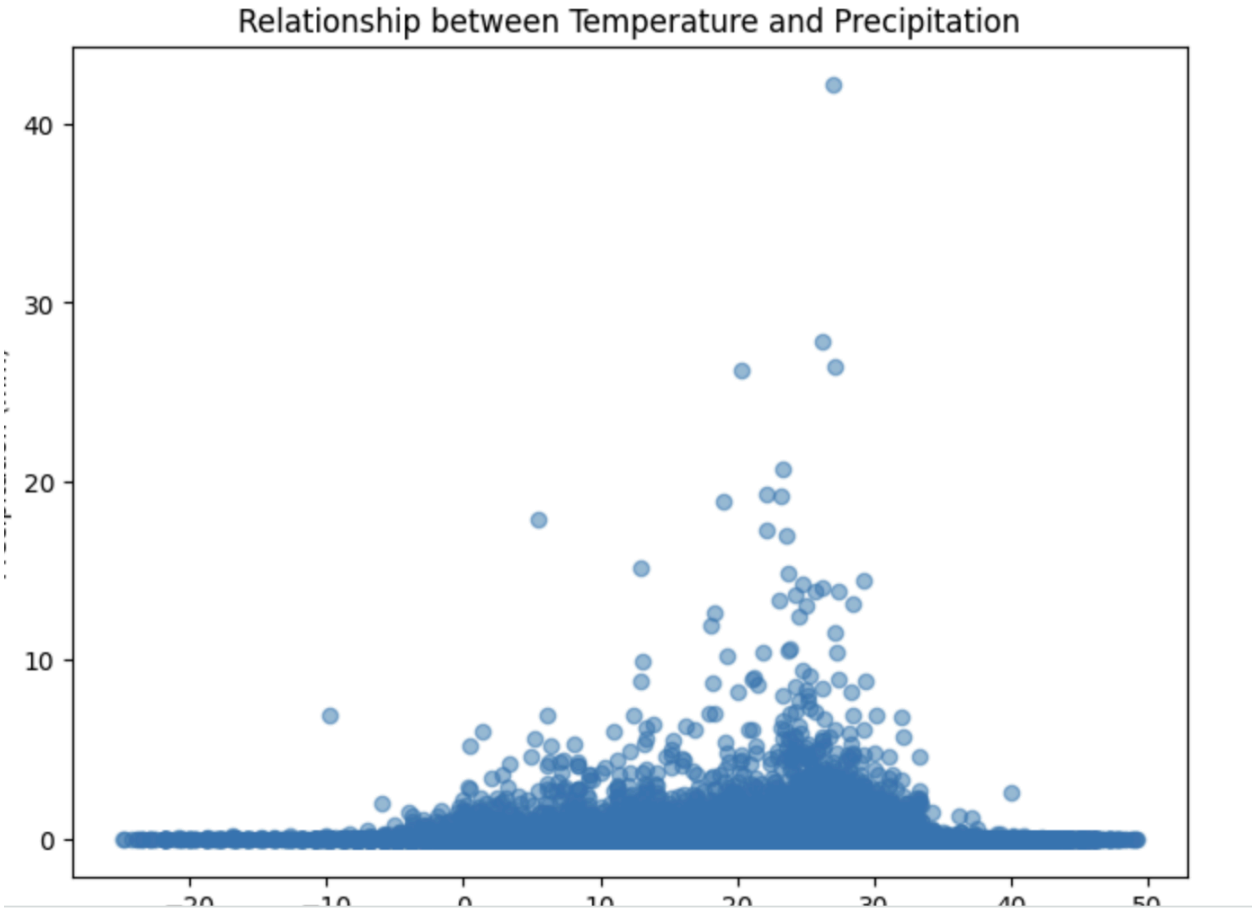
- We plotted violin plots for each of the continuous variables to check for distribution across these columns. As per the scope of this assessment , we will be comparing temperature and precipitation cols later in this report and you can also refer the notebook for more details.



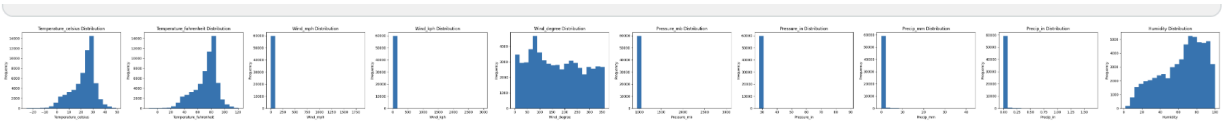
It doesn't give us much information as per correlation is concerned.



The scatter plot suggests that initial months with lower temperatures precipitation are not so much. Only the temperature range of 20-30 C shows variable precipitation ranges.



As concluded above only there is no clear conclusion but maximum precipitation happens in the temperature range of 20-30 degree Celsius.



Modeling Experiments in the Global Weather Data Analysis

1. ARIMA Modeling

Objective: To forecast temperature, precipitation, and humidity using the Autoregressive Integrated Moving Average (ARIMA) model.

Methodology:

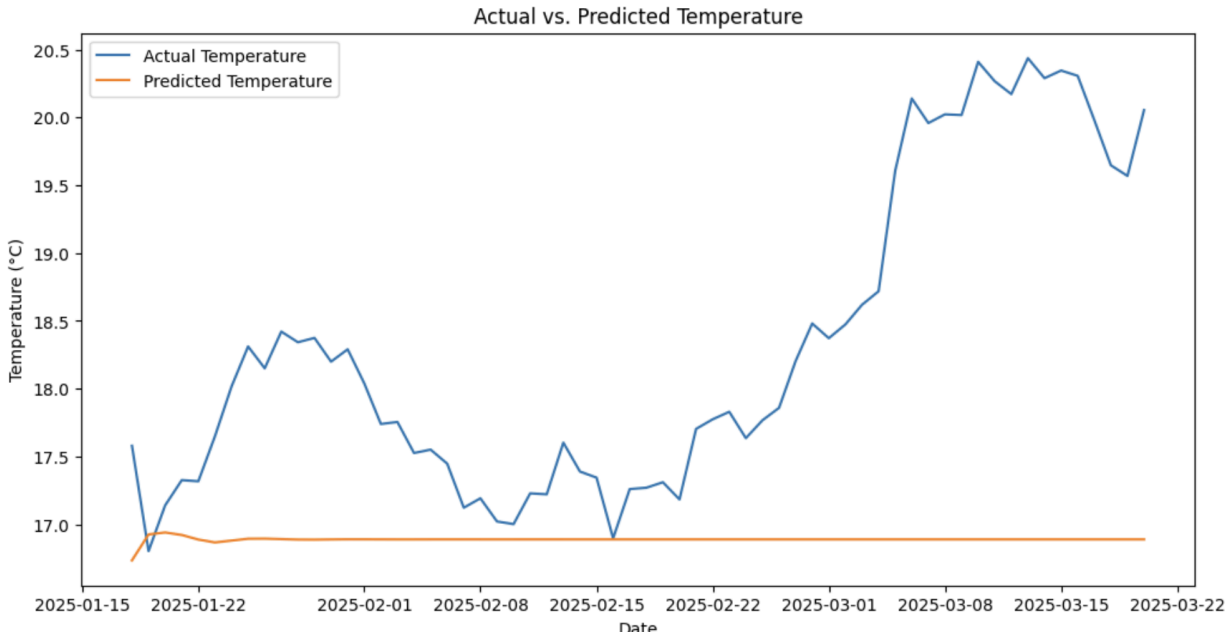
- Data Preparation:
  - The dataset was first preprocessed by converting date columns to datetime objects.
  - The data was then resampled to a daily frequency to ensure consistent time intervals for analysis.
  - Column last updated was used for indexing.
- Model Selection:
  - The ARIMA model was chosen due to its ability to capture temporal dependencies in time series data.
  - The model's order (p, d, q) was determined based on the autocorrelation and partial autocorrelation functions (ACF and PACF) of the time series.
- Model Training:
  - Separate ARIMA models were trained for temperature, precipitation, and humidity using the training data.
  - The models were fitted to the historical data to learn the underlying patterns and relationships.
- Model Evaluation:
  - The trained models were used to predict future values of temperature, precipitation, and humidity.
  - The model's performance was evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).
  - Lower values of these metrics indicate better forecasting accuracy.

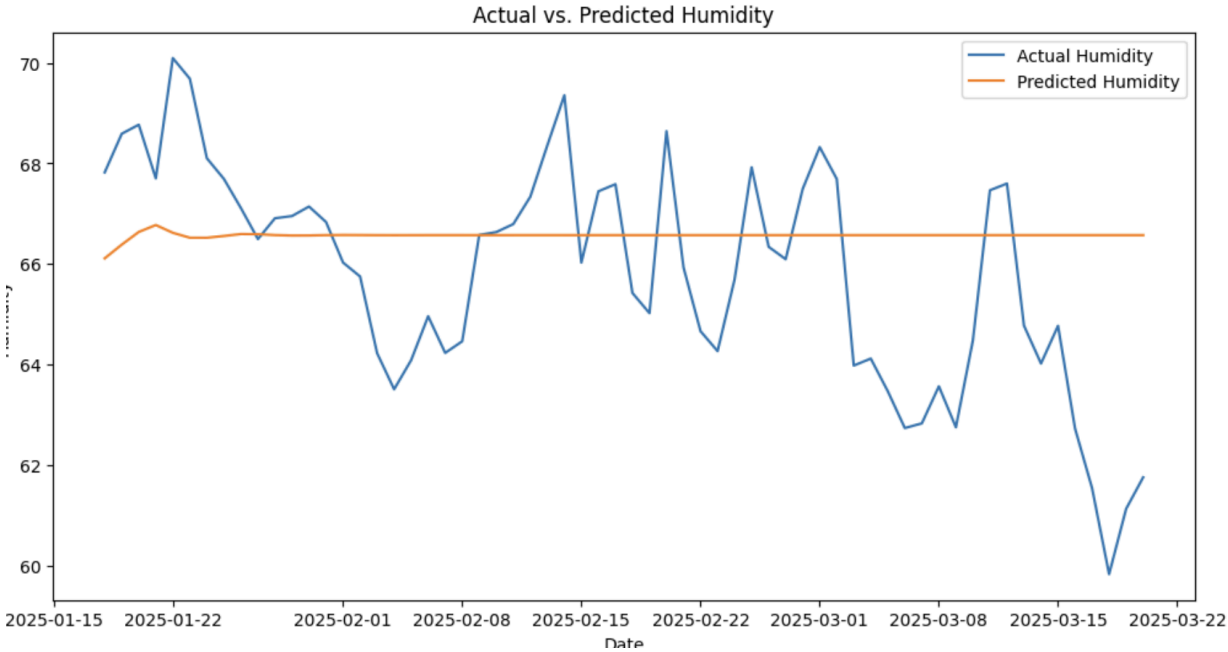
Results:

- The ARIMA models demonstrated reasonable performance in forecasting temperature, precipitation, and humidity.
- The RMSE, MAE, and MAPE values were relatively low, suggesting that the models captured the temporal patterns in the data effectively.

Visualization:

- Actual versus predicted values were visualized using line plots to illustrate the model's forecasting accuracy.





These plots showed a not so good agreement between the actual and predicted values. We need to go back and fine tune our data preprocessing and modeling.

2. LSTM Modeling

Objective: To forecast temperature using Long Short-Term Memory (LSTM) networks, a type of recurrent neural network.

Methodology:

- Data Preprocessing:

The temperature data was scaled using MinMaxScaler to ensure that all features have a similar range.
Sequences of data were created, where each sequence consisted of past temperature values (input) and the next temperature value (output).

- Model Architecture:
  - An LSTM model was constructed with multiple LSTM layers and dropout layers to prevent overfitting.
  - The output layer consisted of a single neuron to predict the next temperature value.
- Model Training:
  - The model was trained using the training data and the Adam optimizer.
  - The training process involved iteratively adjusting the model's weights to minimize the prediction error.
- Model Evaluation:
  - The trained model was used to make predictions on the test data.
  - The model's performance was evaluated using RMSE.
  - A lower RMSE value indicates better forecasting accuracy.

Results:

- The LSTM model achieved a competitive RMSE score, demonstrating its ability to capture complex temporal dependencies in temperature data.

```
Epoch 1/3
/usr/local/lib/python3.10/dist-packages/keras/src/layers/rnn/rnn
`input_shape`/`input_dim` argument to a layer. When using Sequen
hape)` object as the first layer in the model instead.
  super().__init__(**kwargs)
1489/1489 ━━━━━━━━━━━━━━━━━━━━ 292s 194ms/step - loss: 0.0319
Epoch 2/3
1489/1489 ━━━━━━━━━━━━━━━━━━━━ 284s 190ms/step - loss: 0.0237
Epoch 3/3
1489/1489 ━━━━━━━━━━━━━━━━━━━━ 282s 189ms/step - loss: 0.0228
373/373 ━━━━━━━━━━━━━━━━━━━━ 28s 73ms/step
RMSE: 18.046165328069694
```

Conclusion:
Both the ARIMA and LSTM modeling experiments provided valuable insights into forecasting weather variables. The ARIMA models showed bad performance for temperature, precipitation, and humidity prediction, while the LSTM model demonstrated its capability in capturing complex temporal patterns in temperature data in training data but not in testing data, indicating major overfitting.
These experiments highlight the importance of extensive feature engineering and modeling fine tuning and we will be deep diving into these aspects but for the scope of this assessment. I am submitting this as my report. I acknowledge the fact this notebook and model needs much more work than it's there and I will continue to work on this post submission.