



National University of Computer and Emerging Sciences



Weather Data Analysis

Team

<Muhammad Qasim>.....<21L-5246>

<Syed Usman Ali>.....<21L-5405>

Supervised by

<Ms. Maimoona Akram>

FAST School of Computing

National University of Computer and Emerging Sciences

[Lahore], Pakistan

Month Year

1. Introduction

The project focuses on weather data analysis to predict precipitation (rainfall). The dataset, sourced from NASA POWER, contains weather conditions recorded over time and includes parameters such as temperature, humidity, and wind speed. The primary objective is to preprocess, visualize, and analyze the dataset for predictive modelling. Key outcomes include identifying patterns and relationships between variables to enhance understanding and accuracy in precipitation prediction.

2. Data Preparation

2.1 Data Loading

The dataset, comprising 1094 rows and 11 columns, was loaded using the panda's `read_csv()` function. Methods like `info()`, `head()`, `tail()`, and `describe()` provided initial insights into data types and summary statistics.

2.2 Data Exploration

The dataset includes features such as temperature at different levels, humidity, precipitation, and wind speed. Visual inspection and summary statistics helped in understanding the distributions and ranges of these features.

- `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1094 entries, 0 to 1093
Data columns (total 11 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   YEAR                                     1094 non-null   int64
1   MO                                       1094 non-null   int64
2   DY                                       1094 non-null   int64
3   Temperature at 2 Meters @              1094 non-null   float64
4   Dew/Frost Point at 2 Meters (C)        1094 non-null   float64
5   Temperature at 2 Meters Maximum (C)    1094 non-null   float64
6   Temperature at 2 Meters Minimum @      1094 non-null   float64
7   Specific Humidity at 2 Meters (g/kg)   1094 non-null   float64
8   Relative Humidity at 2 Meters (%)       1094 non-null   float64
9   Precipitation Corrected (mm/day)       1094 non-null   float64
10  Wind Speed at 10 Meters                 1094 non-null   float64
dtypes: float64(8), int64(3)
memory usage: 94.1 KB
```

- `df.describe()`

	YEAR	MO	DY	Temperature at 2 Meters @	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum @	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
count	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000	1094.000000
mean	2021.999086	6.521024	15.706581	24.815777	12.788940	31.724680	18.856243	10.728995	53.420064	2.580137	2.306892
std	0.816683	3.447026	8.792141	8.070966	8.224657	7.615615	8.061764	5.761845	18.171100	7.764667	0.793325
min	2021.000000	1.000000	1.000000	7.540000	-5.410000	13.830000	0.320000	2.560000	9.500000	0.000000	0.570000
25%	2021.000000	4.000000	8.000000	17.562500	5.900000	25.650000	11.922500	5.920000	40.440000	0.000000	1.730000
50%	2022.000000	7.000000	16.000000	26.280000	11.535000	32.655000	19.355000	8.670000	54.155000	0.000000	2.170000
75%	2023.000000	10.000000	23.000000	31.050000	20.940000	37.047500	25.975000	15.975000	67.690000	1.197500	2.760000
max	2023.000000	12.000000	31.000000	40.560000	27.190000	47.820000	33.510000	23.250000	95.060000	96.170000	7.020000

- `df.head ()`

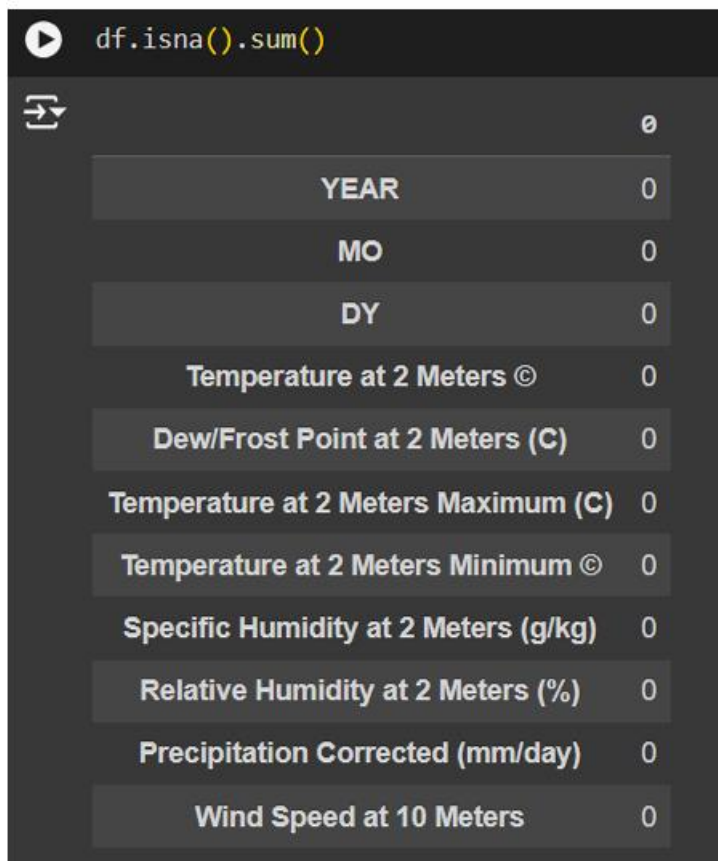
	YEAR	MO	DY	Temperature at 2 Meters @	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum @	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
0	2021	1	1	11.09	-1.05	18.75	5.82	3.54	45.00	0.00	1.43
1	2021	1	2	10.99	1.71	18.87	5.91	4.52	56.12	0.41	1.82
2	2021	1	3	12.78	10.72	17.98	8.79	8.18	87.69	1.54	2.55
3	2021	1	4	14.45	13.18	18.49	11.69	9.58	92.12	3.36	2.38
4	2021	1	5	14.19	13.42	17.01	11.78	9.70	95.06	32.72	3.84

- `df.tail ()`

	YEAR	MO	DY	Temperature at 2 Meters @	Dew/Frost Point at 2 Meters (C)	Temperature at 2 Meters Maximum (C)	Temperature at 2 Meters Minimum @	Specific Humidity at 2 Meters (g/kg)	Relative Humidity at 2 Meters (%)	Precipitation Corrected (mm/day)	Wind Speed at 10 Meters
1089	2023	12	26	15.95	4.68	23.08	9.92	5.37	48.56	0.0	1.18
1090	2023	12	27	14.60	4.86	21.62	9.49	5.43	53.31	0.0	1.45
1091	2023	12	28	13.03	4.43	21.74	7.23	5.25	58.12	0.0	2.27
1092	2023	12	29	13.35	4.33	20.53	7.79	5.25	56.56	0.0	1.25
1093	2023	12	30	12.95	3.88	19.33	8.33	5.07	55.50	0.0	1.09

2.3 Data Cleaning

No missing or duplicate values were found. Outliers were identified using skewness and the Z-score/IQR method. While wind speed outliers were removed, precipitation outliers were retained for further analysis.



```
df.isna().sum()
```

	0
YEAR	0
MO	0
DY	0
Temperature at 2 Meters ©	0
Dew/Frost Point at 2 Meters (C)	0
Temperature at 2 Meters Maximum (C)	0
Temperature at 2 Meters Minimum ©	0
Specific Humidity at 2 Meters (g/kg)	0
Relative Humidity at 2 Meters (%)	0
Precipitation Corrected (mm/day)	0
Wind Speed at 10 Meters	0

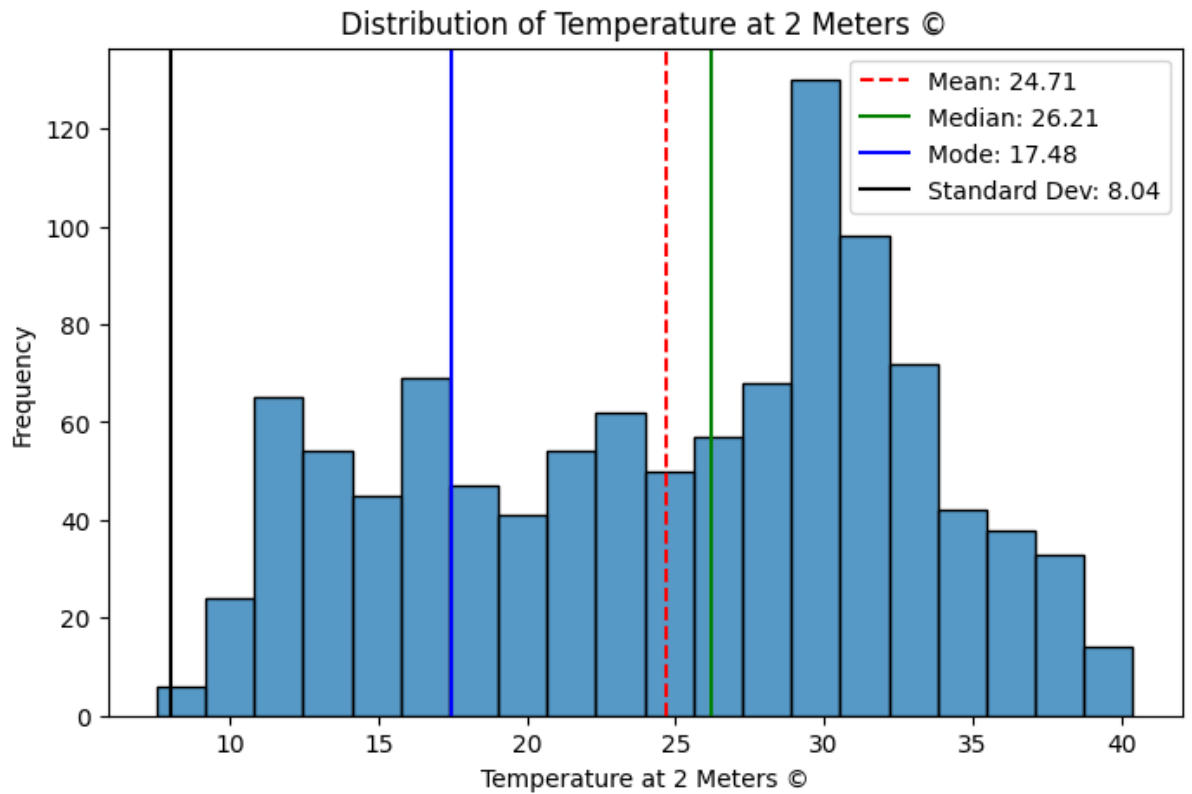
2.4 Data Transformation

Standardization (Z-score normalization) was applied to ensure consistent feature scaling, as variables had differing units and scales.

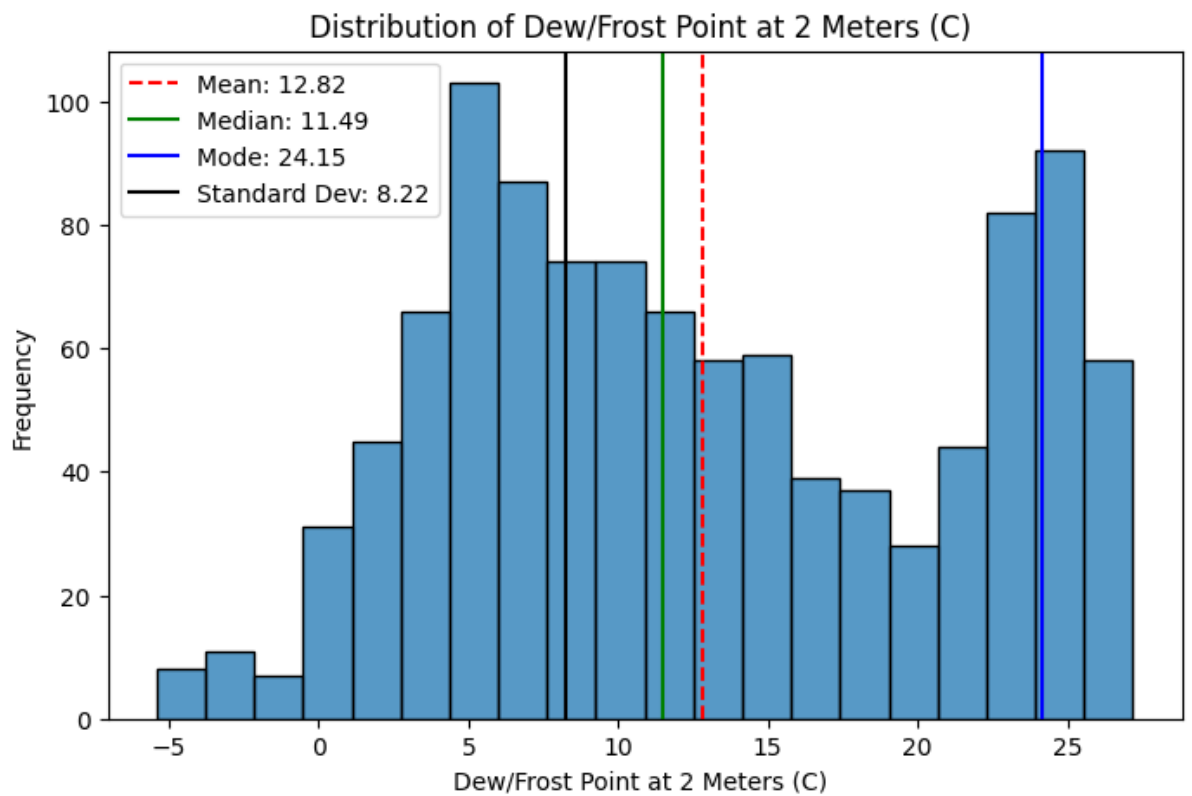
3. Data Analysis

3.1 Univariate Analysis

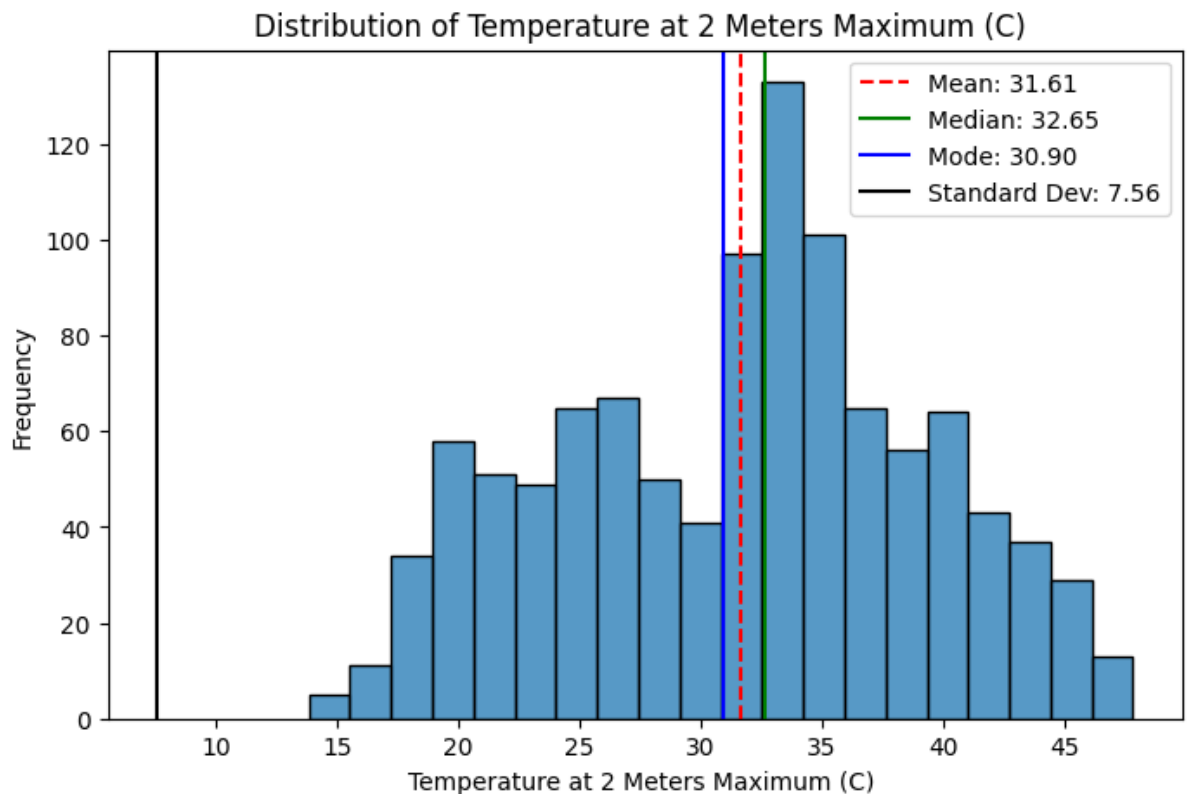
- Temperature at 2 Meters: The distribution was slightly left-skewed, with a mean of 24.71°C and a peak frequency around 30°C.



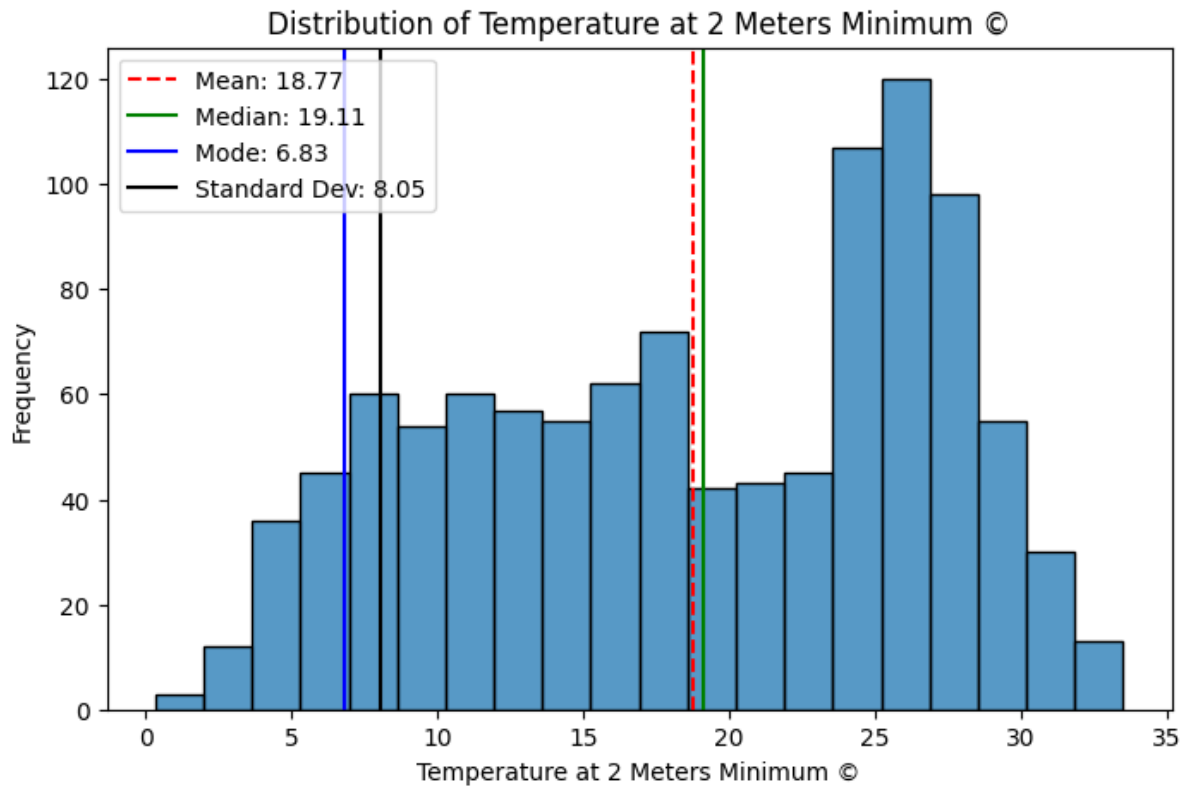
- Dew/Frost Point at 2 Meters: Slightly right-skewed with a mean of 12.82°C and values mostly concentrated between 5°C and 15°C.



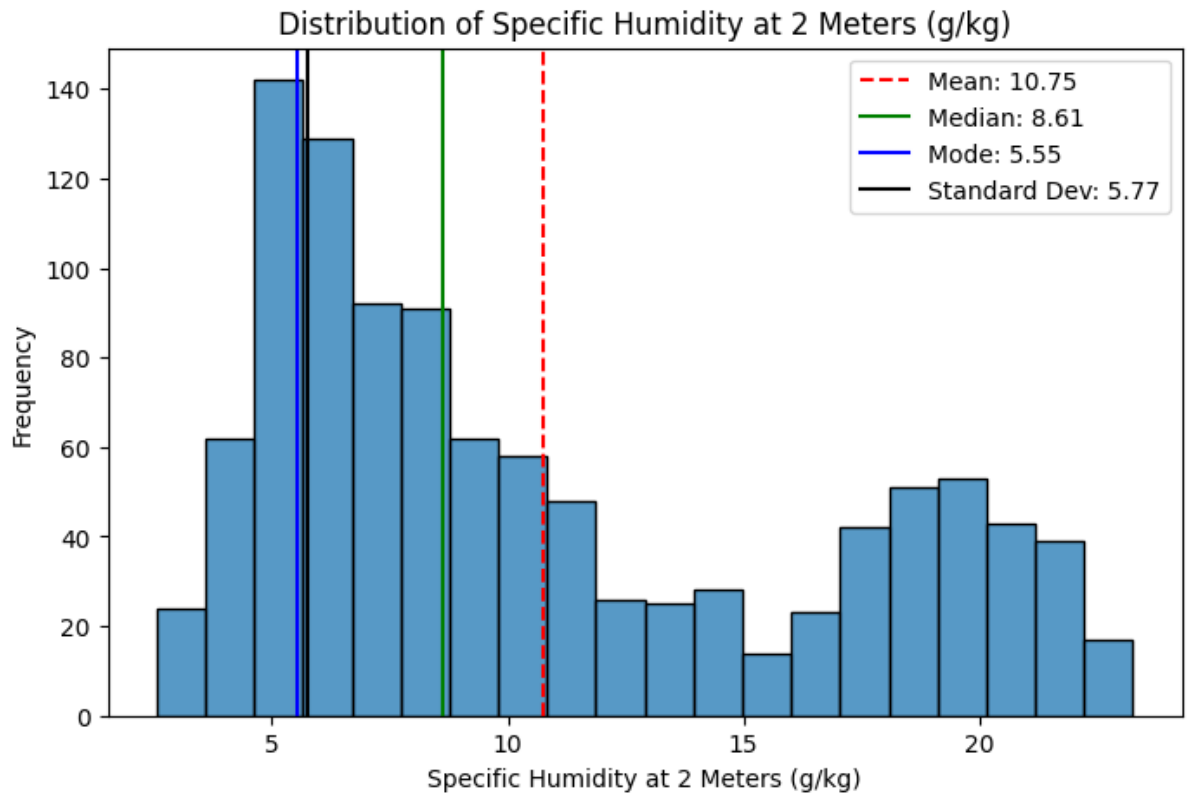
- Temperature at 2 Meters Maximum: Close proximity of mean, mode, and median suggests fairly normal distribution, with a slight skew towards higher temperatures. In addition to this the histogram also shows that the most common temperature range is between 30°C and 35°C, which aligns with the mode of 30.90°C. Rare occurrences occur on 10°C to 15°C indicating cooler days.



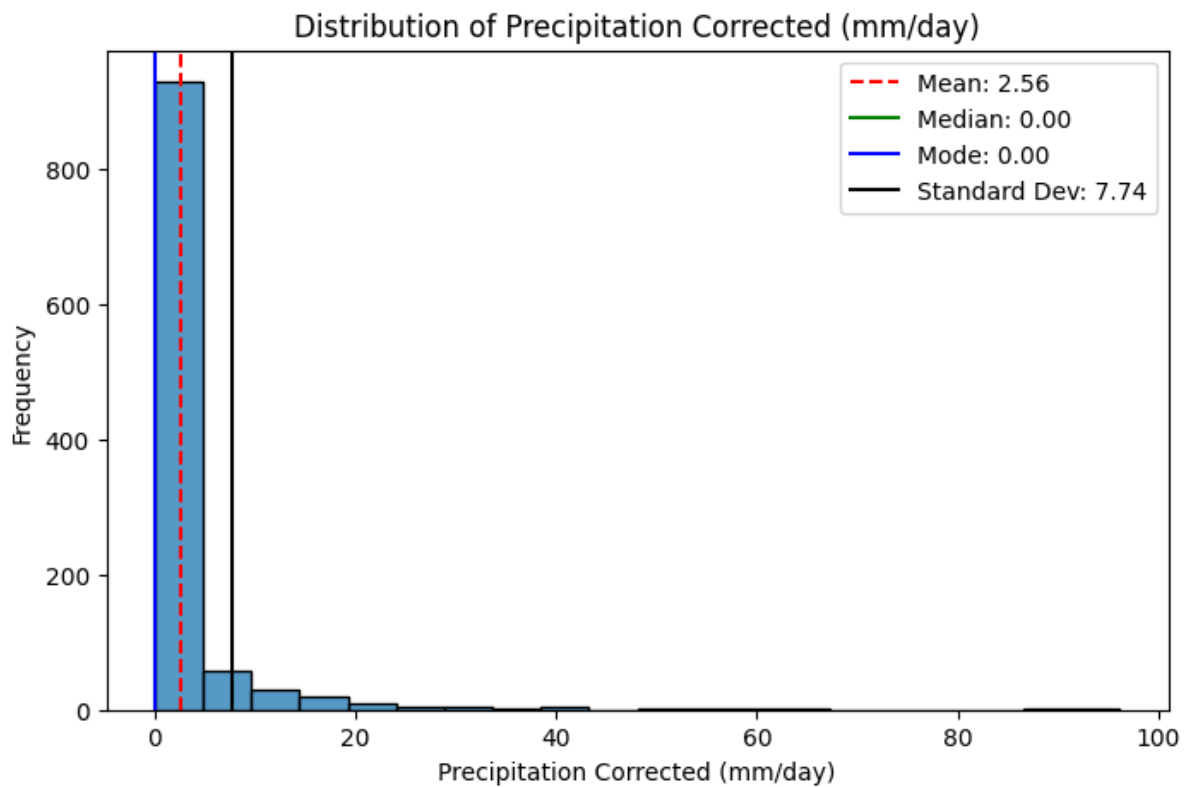
- Temperature at 2 Meters minimum: Histogram shows that the frequency distribution peaks around 25°C indicating that this is the most common temperature range. The mean temperature is 18.77°C, the median is 19.11°C, and the mode is 6.83°C. The standard deviation is 8.05°C, indicating variability around the mean.



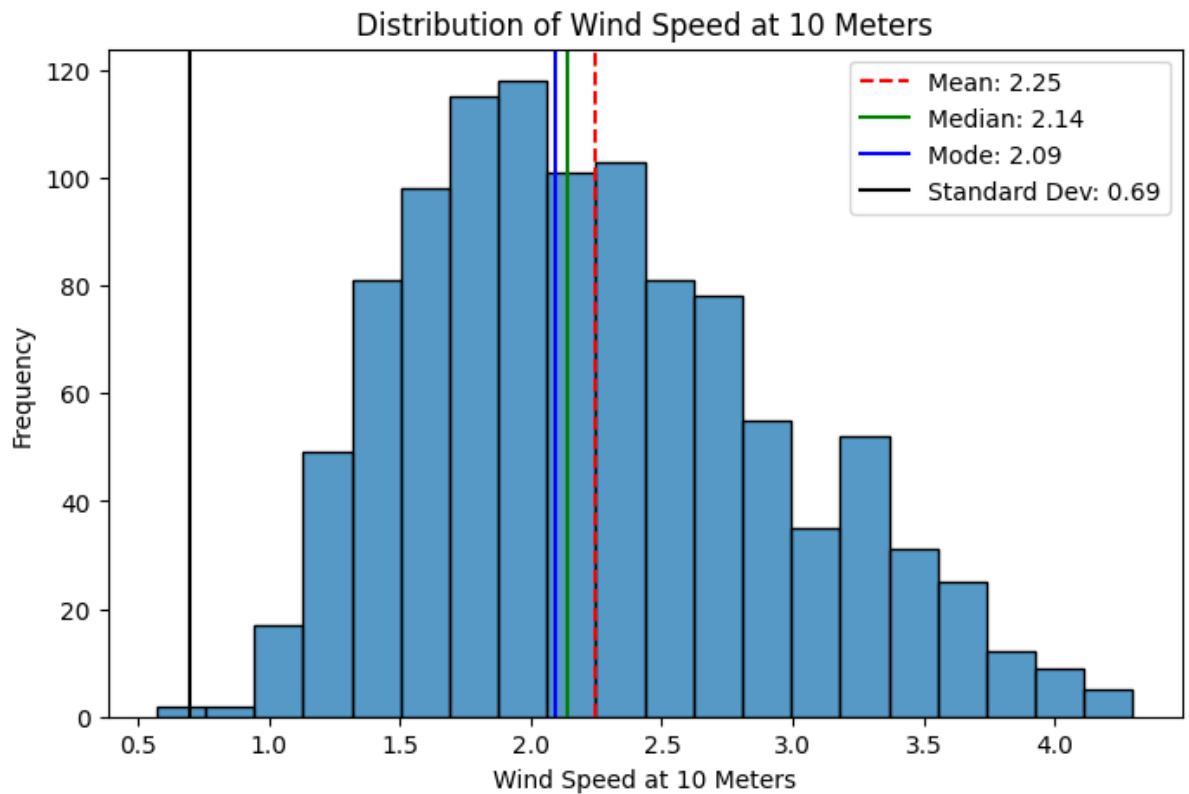
- Specific humidity at 2 meters: The frequency distribution of specific humidity at 2 meters features shows that the most common range is around 5.55 g/kg, indicating typical humidity levels. The mean specific humidity is 10.75 g/kg, the median is 8.61 g/kg, and the mode is 5.55 g/kg. The standard deviation is 5.77 g/kg, indicating variability around the mean.



- Precipitation Corrected: A positively skewed distribution with a mean of 2.56 mm/day, highlighting zero precipitation as the most frequent occurrence.



- Wind Speed at 10 Meters: A symmetric distribution with a mean of 2.25 m/s, showing clustering around the average value.



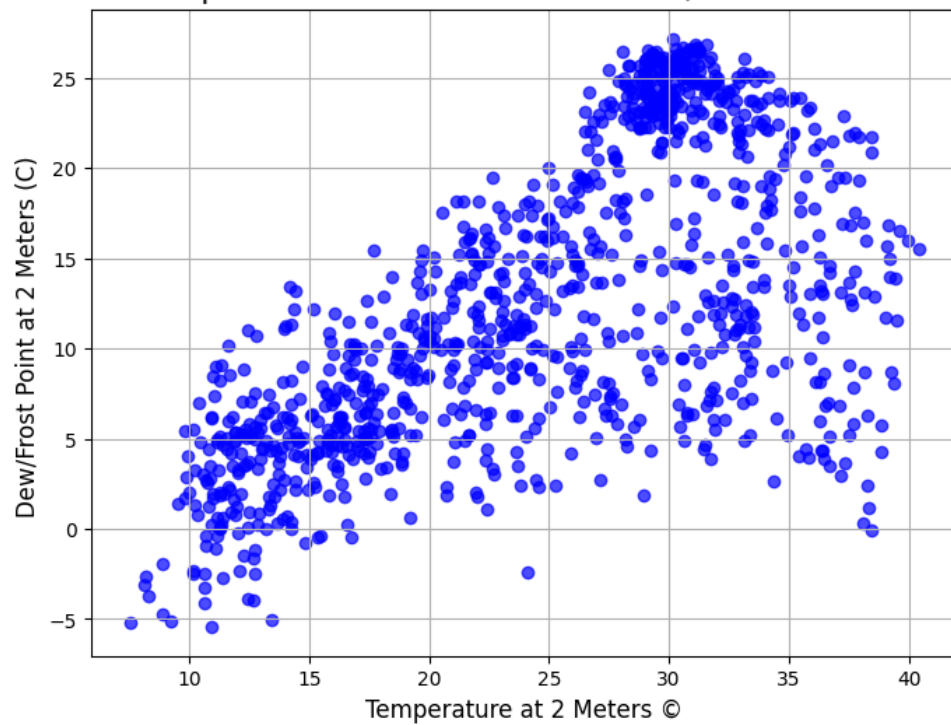
3.2 Bivariate and Multivariate Analysis

Scatter plots and correlation matrices revealed:

Temperature at 2 Meters vs. Dew/Frost Point:

- Positive correlation.
- Cluster around 25°C temperature and 15°C–20°C dew/frost point suggests common conditions.

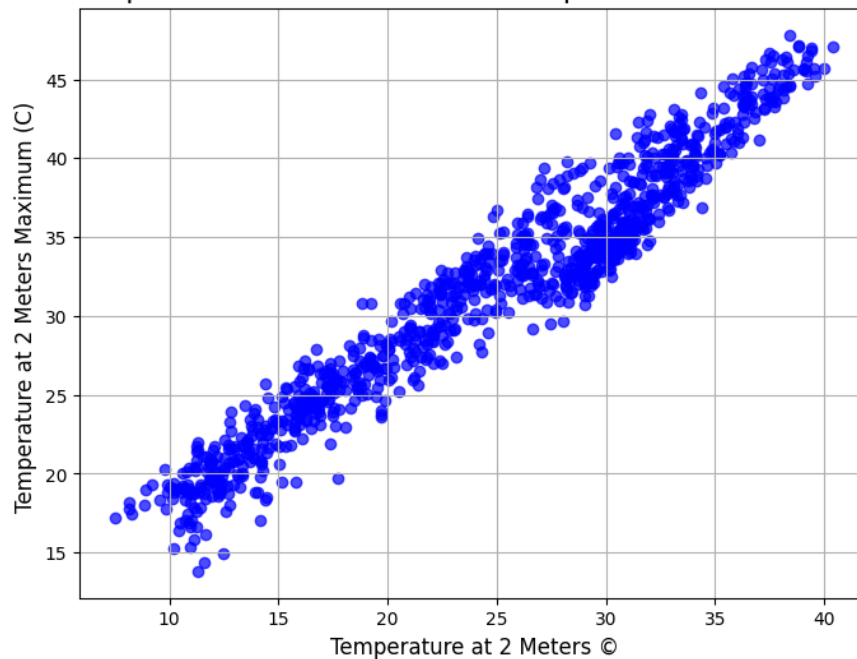
Scatter Plot: Temperature at 2 Meters © vs Dew/Frost Point at 2 Meters (C)



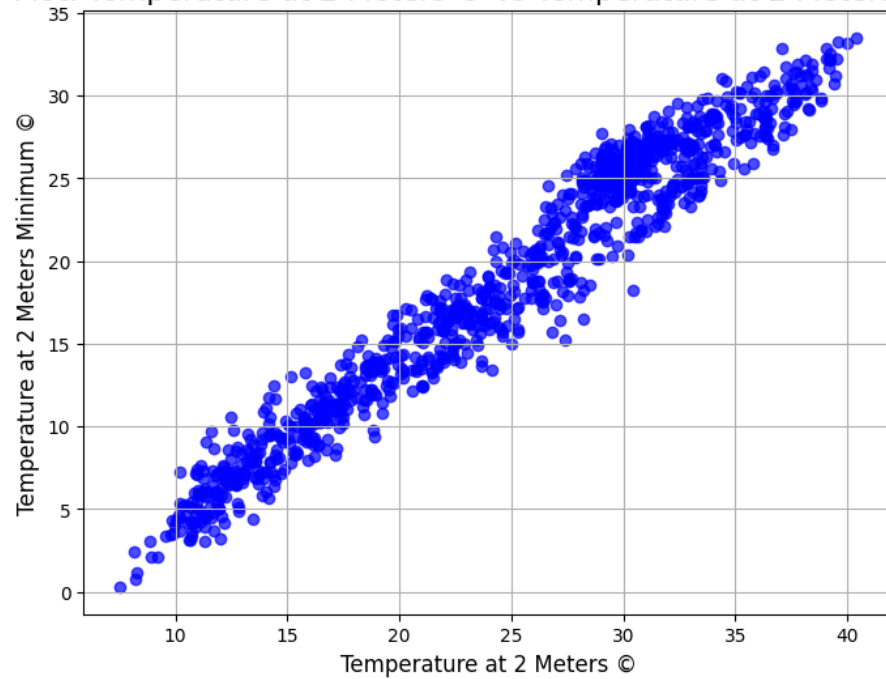
Temperature at 2 Meters vs. Max/Min Temperature:

- Strong positive correlation.
- Higher temperatures at 2 meters are associated with higher max/min temperatures.

Scatter Plot: Temperature at 2 Meters © vs Temperature at 2 Meters Maximum (C)



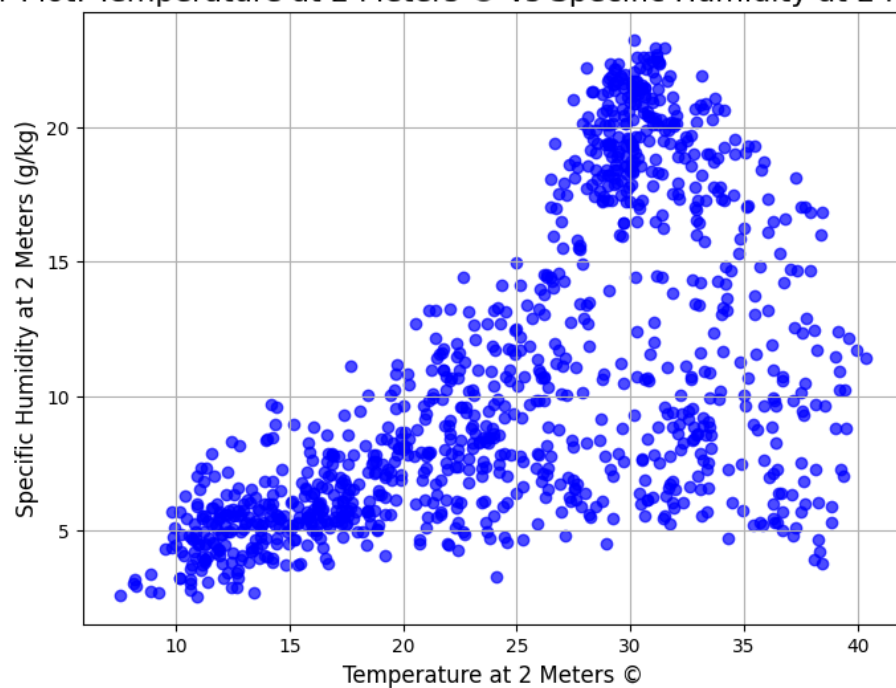
Scatter Plot: Temperature at 2 Meters © vs Temperature at 2 Meters Minimum ©



Temperature vs. Specific Humidity:

- Positive correlation.
- Data densely packed between 20°C–30°C temperatures and 5–15 g/kg specific humidity.

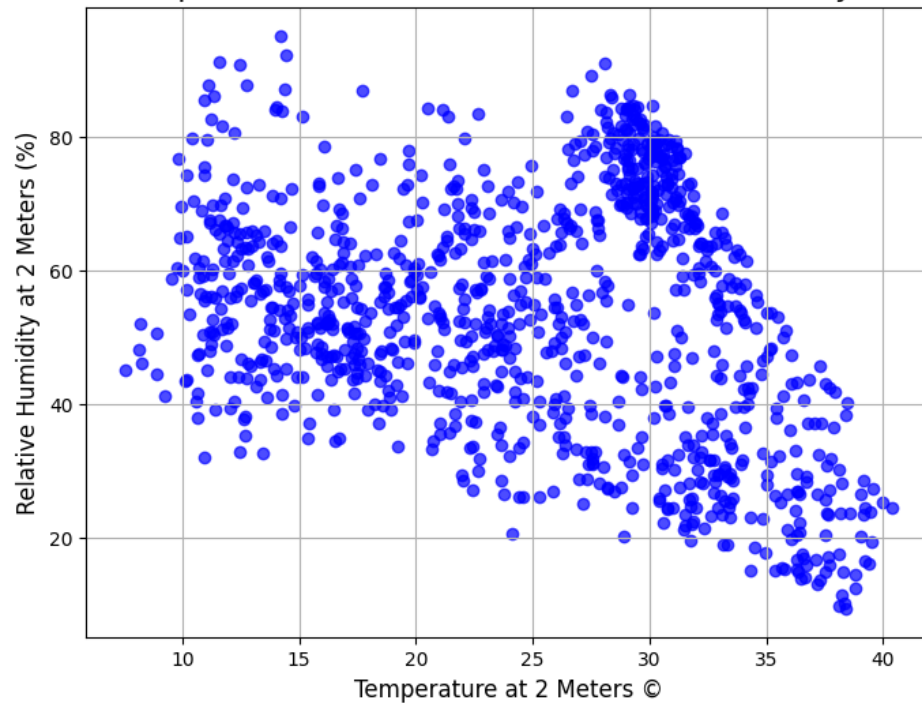
Scatter Plot: Temperature at 2 Meters © vs Specific Humidity at 2 Meters (g/kg)



Temperature vs. Relative Humidity:

- Negative correlation.
- As temperature increases, relative humidity decreases.

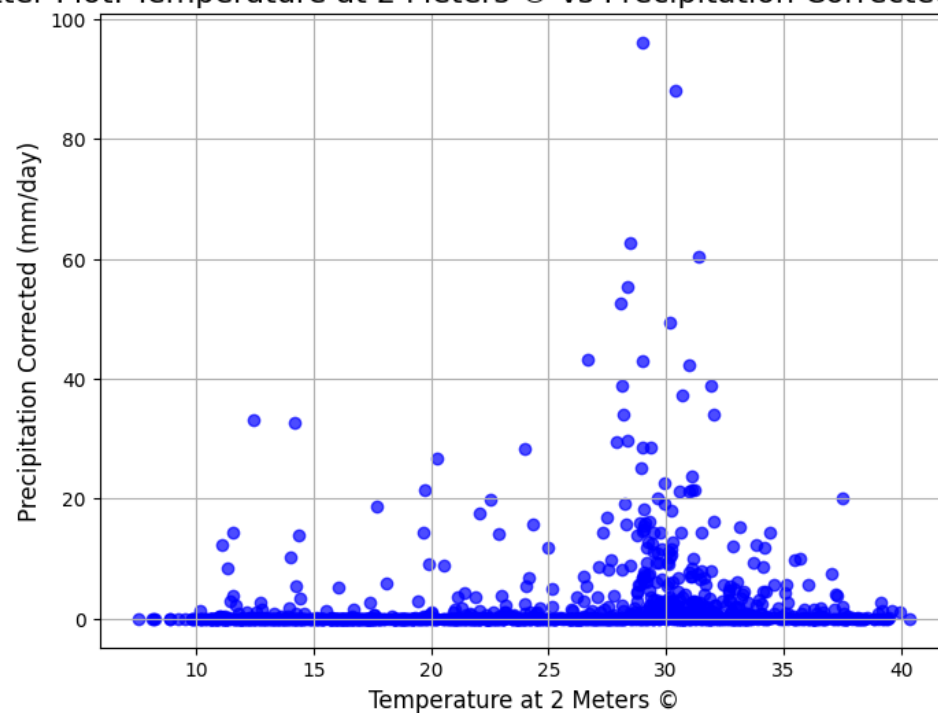
Scatter Plot: Temperature at 2 Meters © vs Relative Humidity at 2 Meters (%)



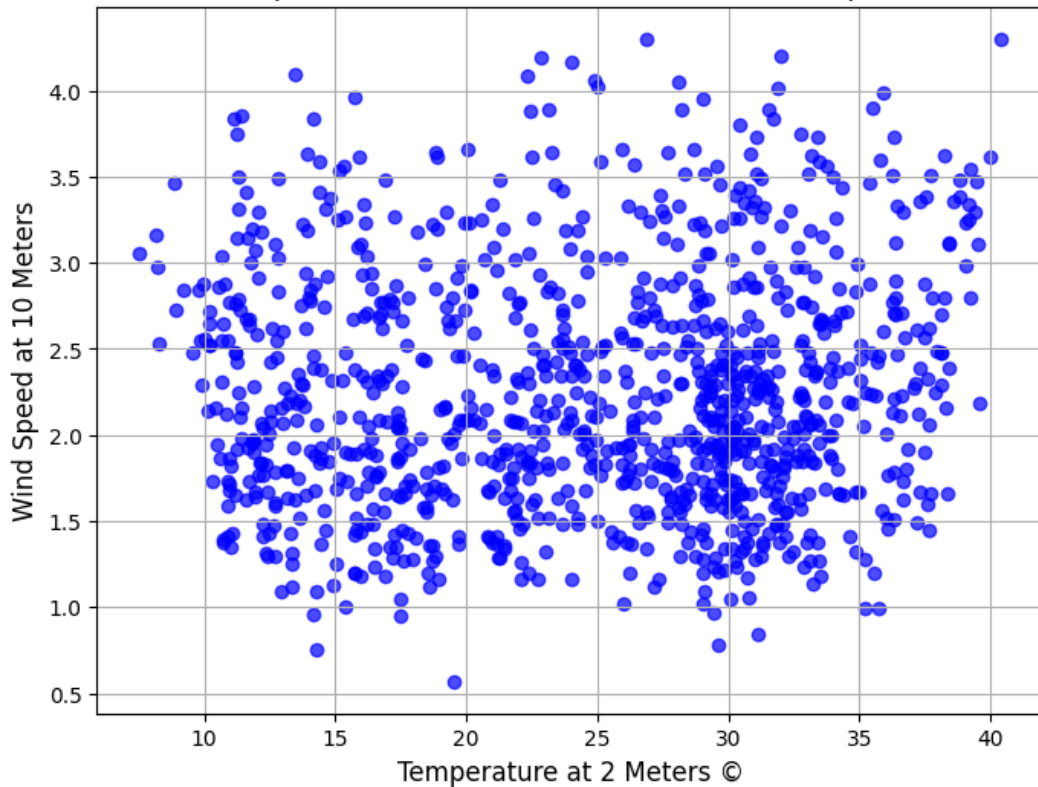
Temperature vs. Precipitation and Wind Speed:

- Very low/negligible correlation.

Scatter Plot: Temperature at 2 Meters © vs Precipitation Corrected (mm/day)



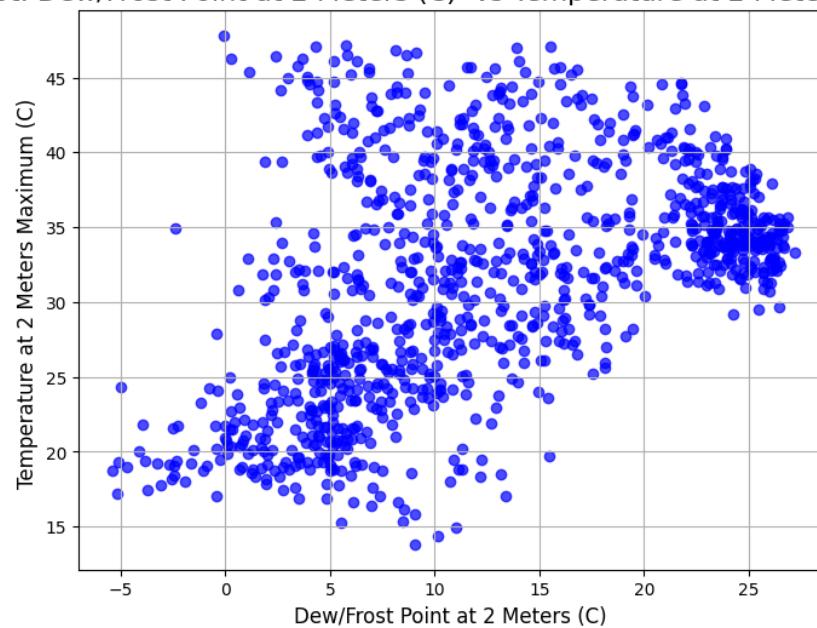
Scatter Plot: Temperature at 2 Meters © vs Wind Speed at 10 Meters



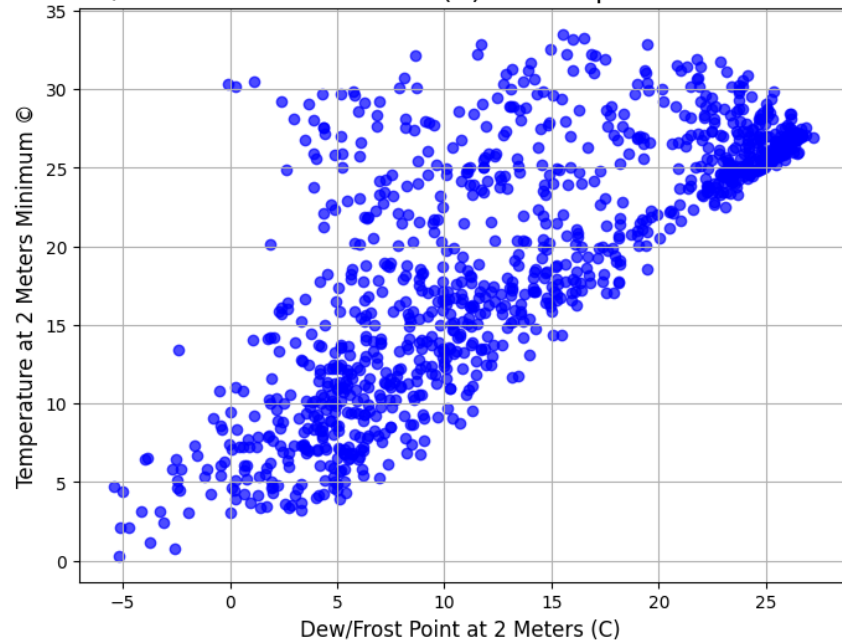
Dew/Frost vs. Temperature at 2 Meters (Max/Min):

- Positive correlation with both.
- Stronger with Temperature at 2 Meters Minimum (0.76).

Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Temperature at 2 Meters Maximum (C)



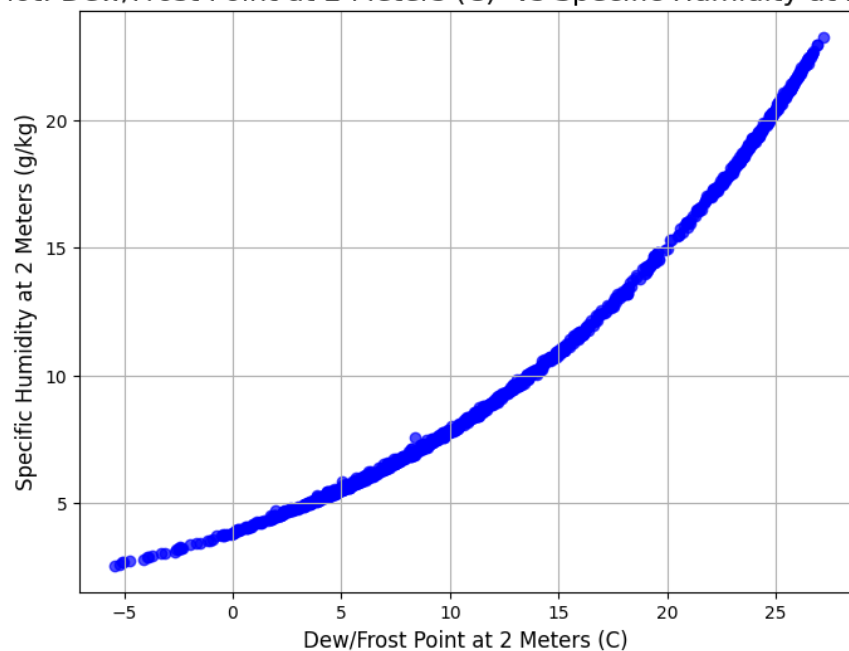
Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Temperature at 2 Meters Minimum ©



Dew/Frost vs. Specific Humidity:

- Highly positive correlation (0.98).

Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Specific Humidity at 2 Meters (g/kg)

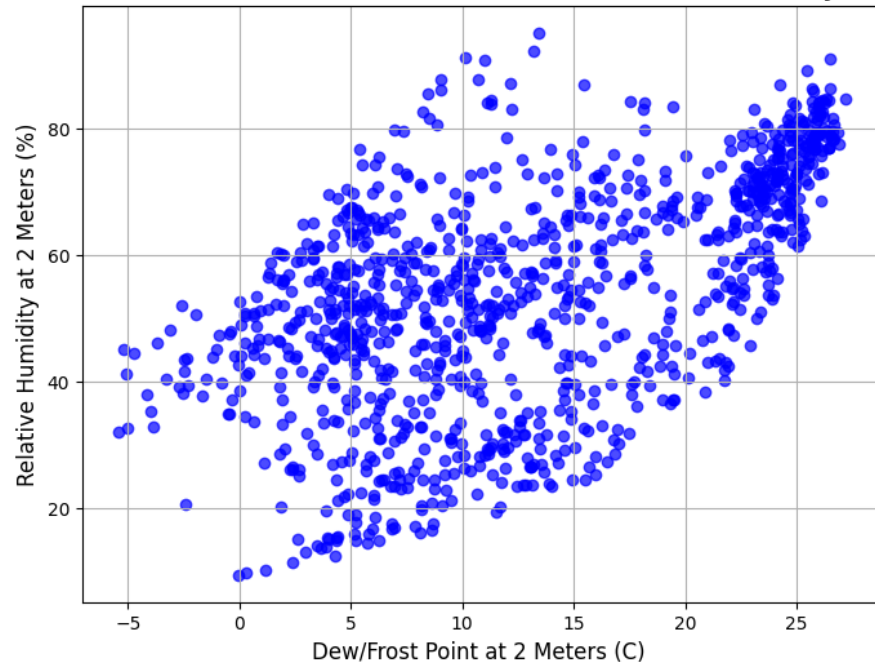


Dew/Frost vs. Relative Humidity, Precipitation, and Wind Speed:

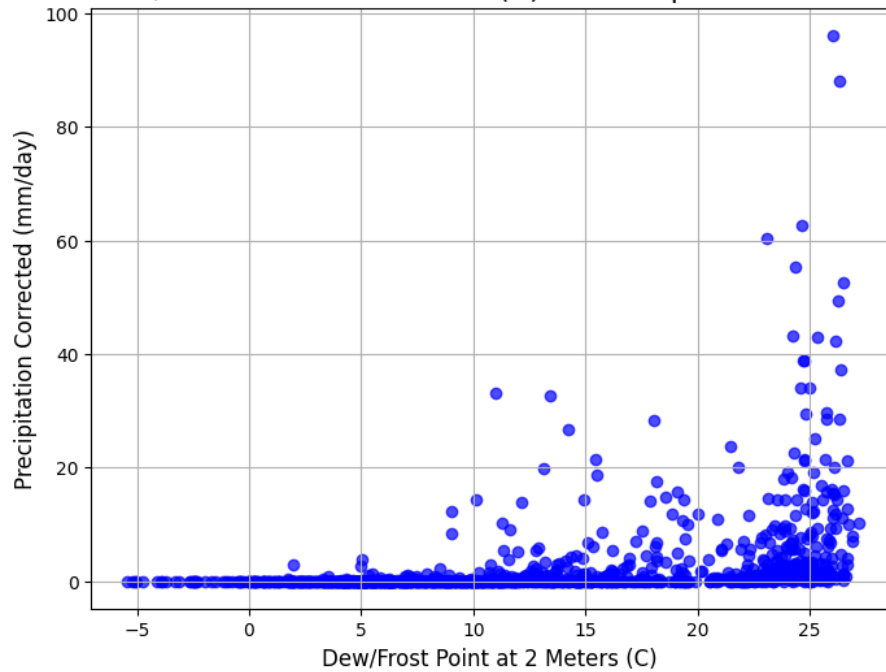
- Positive but weak correlation with relative humidity and precipitation.

- Negative correlation with wind speed.

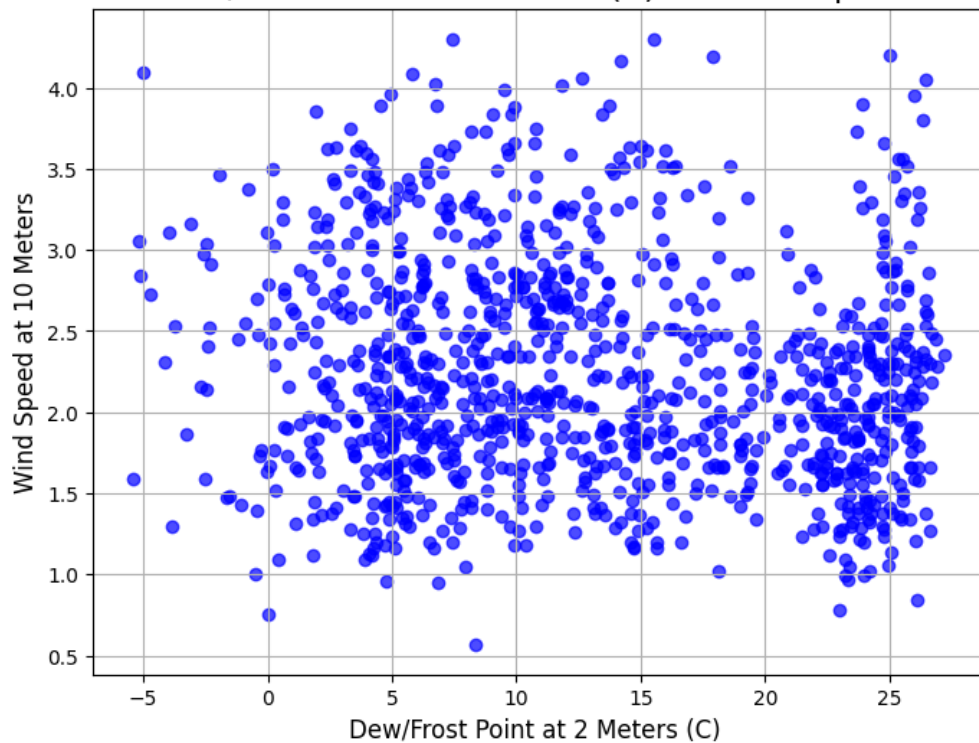
Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Relative Humidity at 2 Meters (%)



Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Precipitation Corrected (mm/day)



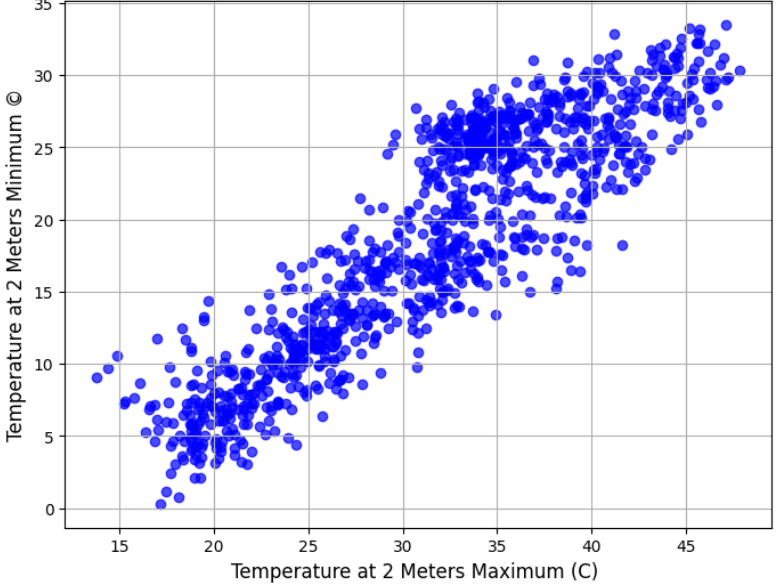
Scatter Plot: Dew/Frost Point at 2 Meters (C) vs Wind Speed at 10 Meters



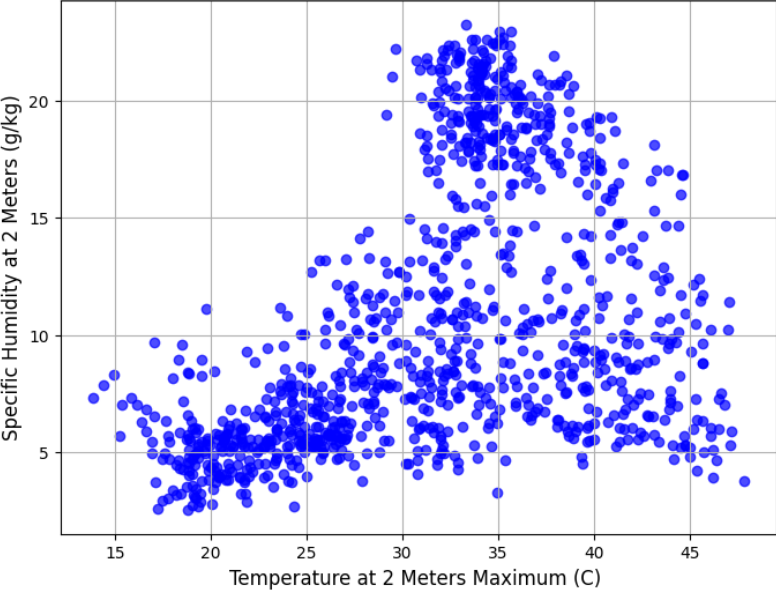
Temperature at 2 Meters Max vs. Min and Specific Humidity:

- Strong positive correlation with Min Temperature.
- Positive correlation with Specific Humidity.
- Negative correlation with Relative Humidity (-0.44).
- Weak positive correlation with Precipitation and Wind Speed.

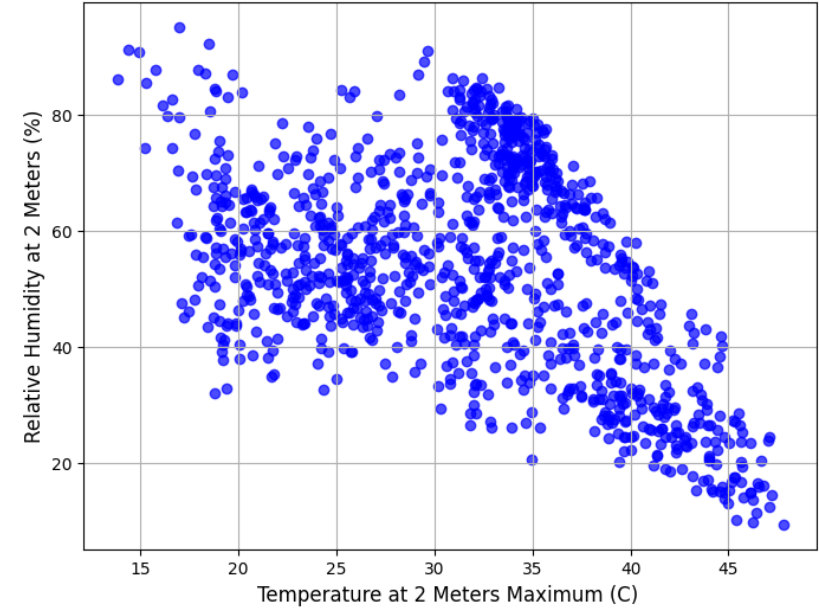
Scatter Plot: Temperature at 2 Meters Maximum (C) vs Temperature at 2 Meters Minimum ©



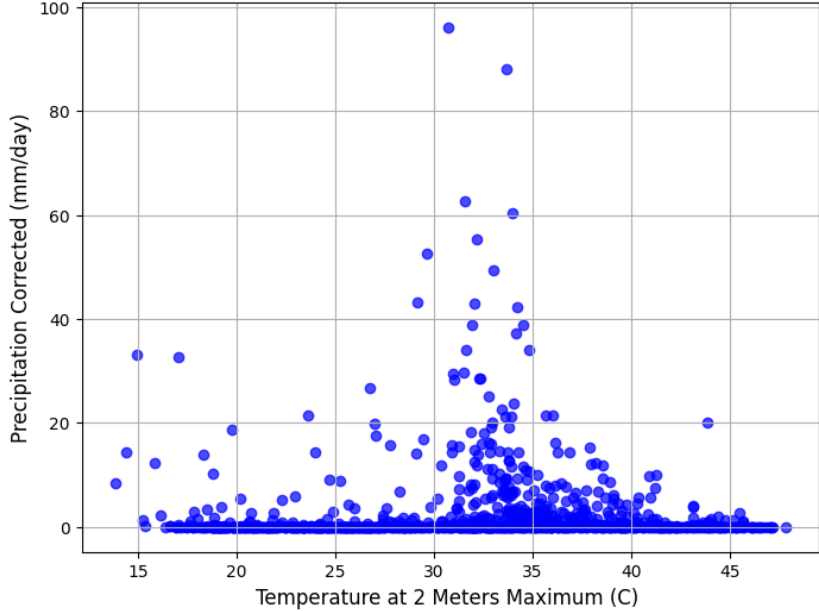
Scatter Plot: Temperature at 2 Meters Maximum (C) vs Specific Humidity at 2 Meters (g/kg)



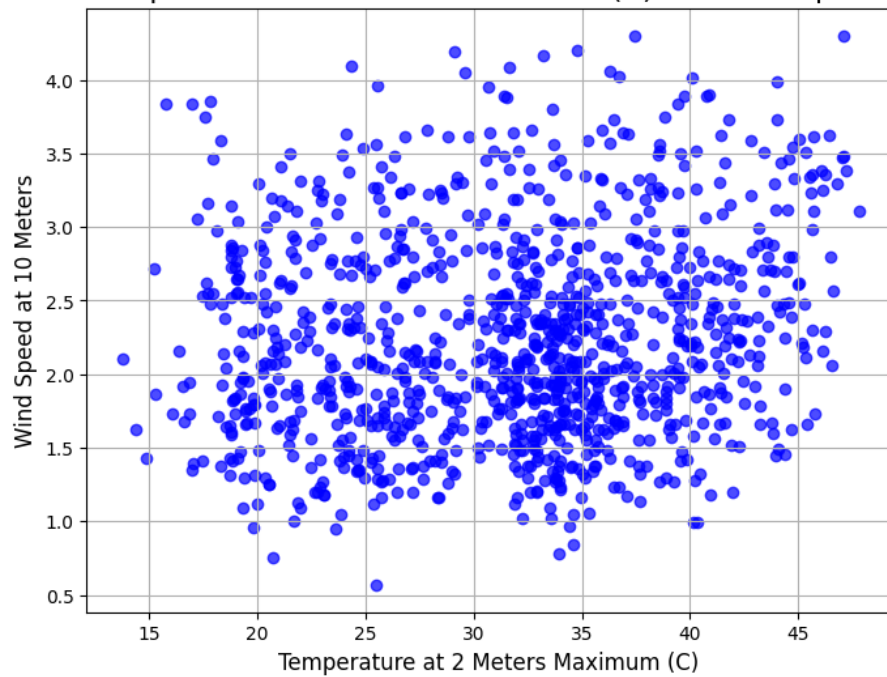
Scatter Plot: Temperature at 2 Meters Maximum (C) vs Relative Humidity at 2 Meters (%)



Scatter Plot: Temperature at 2 Meters Maximum (C) vs Precipitation Corrected (mm/day)



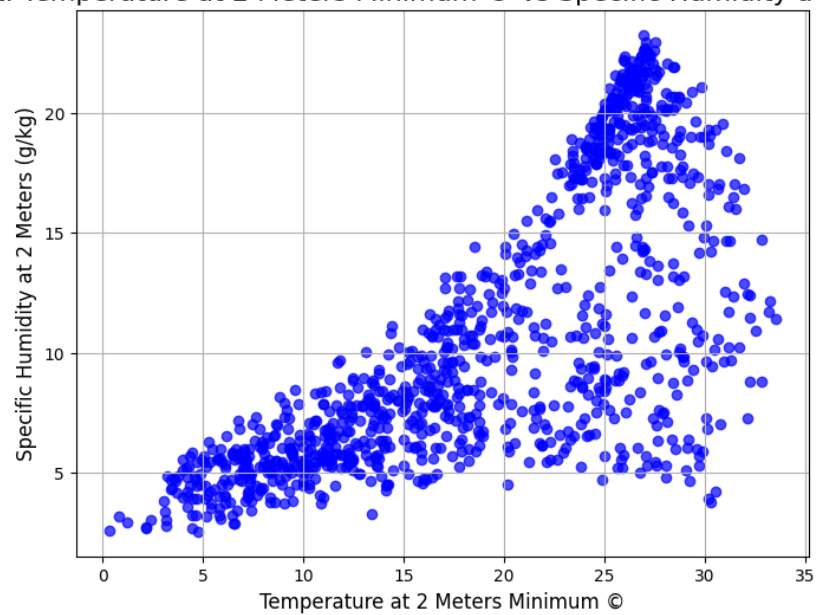
Scatter Plot: Temperature at 2 Meters Maximum (C) vs Wind Speed at 10 Meters



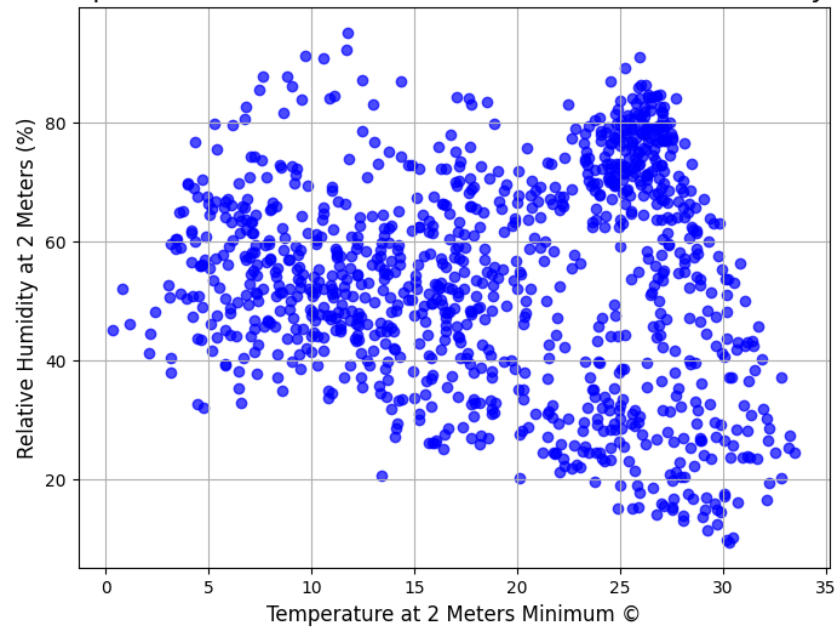
Temperature at 2 Meters Min vs. Specific Humidity and Relative Humidity:

- Positive correlation with Specific Humidity (0.72).
- Negative correlation with Relative Humidity.

Scatter Plot: Temperature at 2 Meters Minimum © vs Specific Humidity at 2 Meters (g/kg)



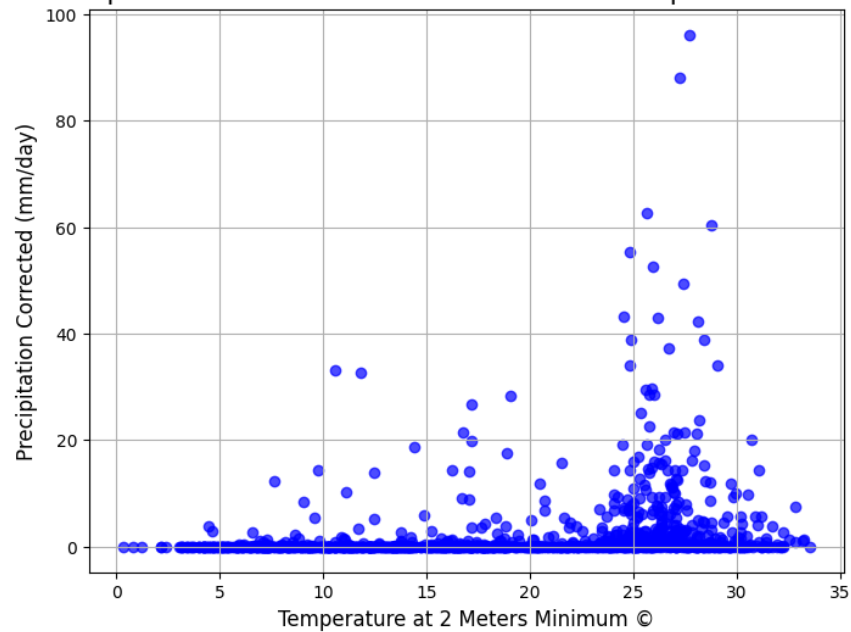
Scatter Plot: Temperature at 2 Meters Minimum © vs Relative Humidity at 2 Meters (%)



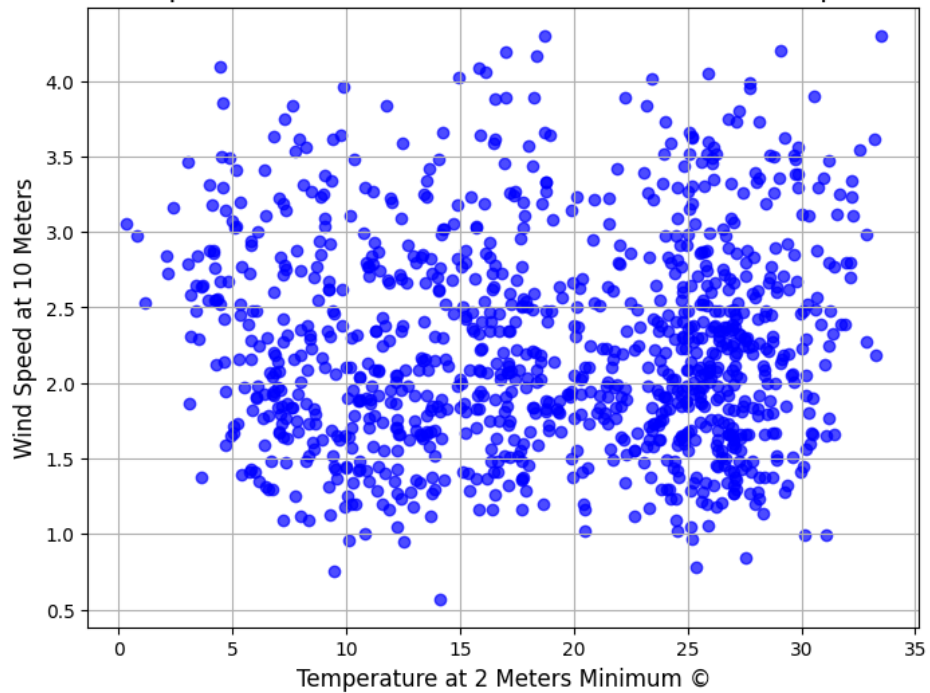
Temperature at 2 Meters Min vs. Precipitation and Wind Speed:

- Weak positive correlation.

Scatter Plot: Temperature at 2 Meters Minimum © vs Precipitation Corrected (mm/day)



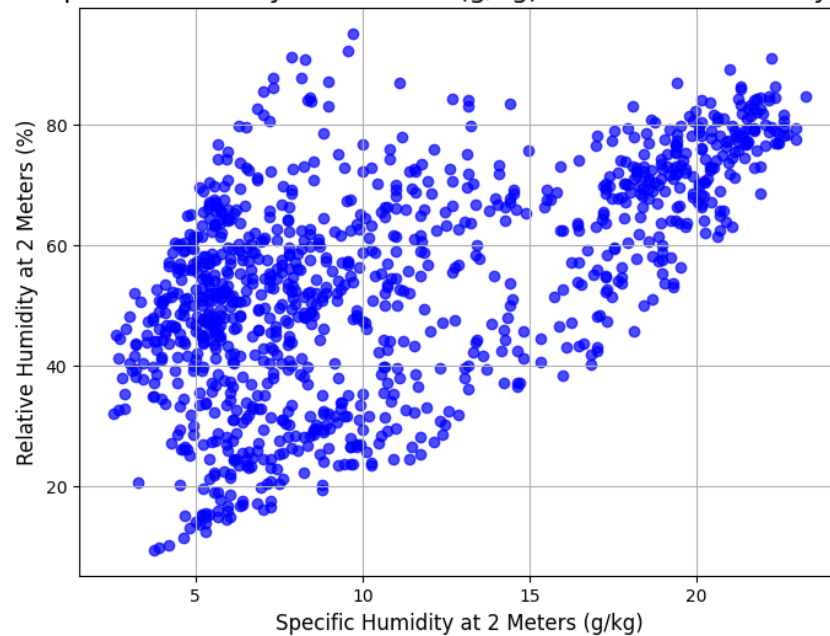
Scatter Plot: Temperature at 2 Meters Minimum © vs Wind Speed at 10 Meters



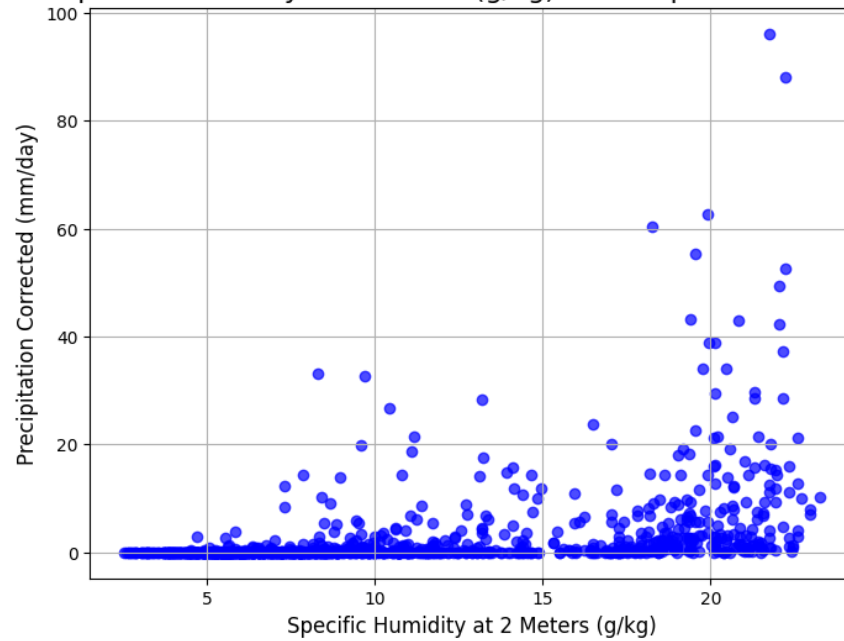
Specific Humidity vs. Relative Humidity and Precipitation:

- Positive correlation as both variables increases together.

Scatter Plot: Specific Humidity at 2 Meters (g/kg) vs Relative Humidity at 2 Meters (%)



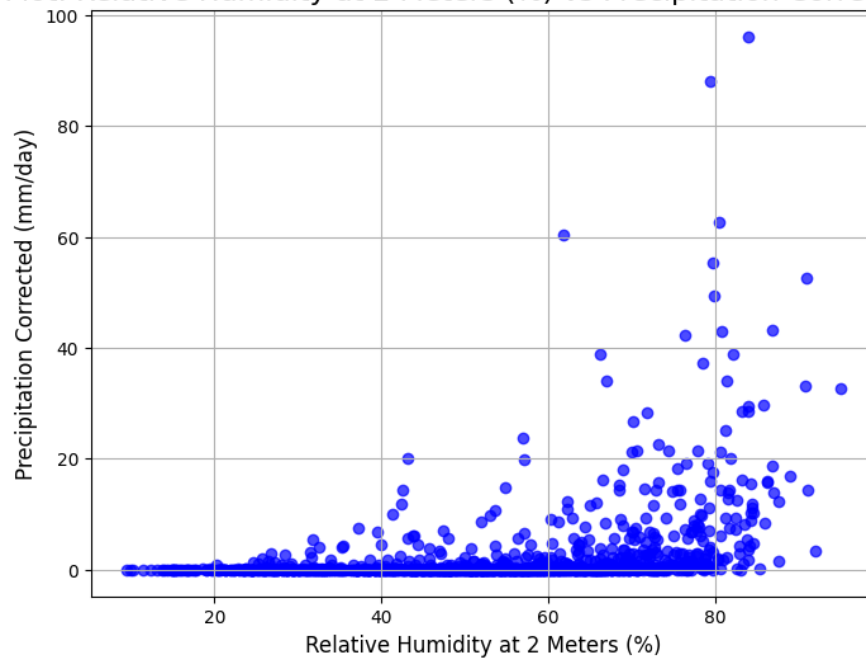
Scatter Plot: Specific Humidity at 2 Meters (g/kg) vs Precipitation Corrected (mm/day)



Relative Humidity vs. Precipitation:

- Positive correlation.

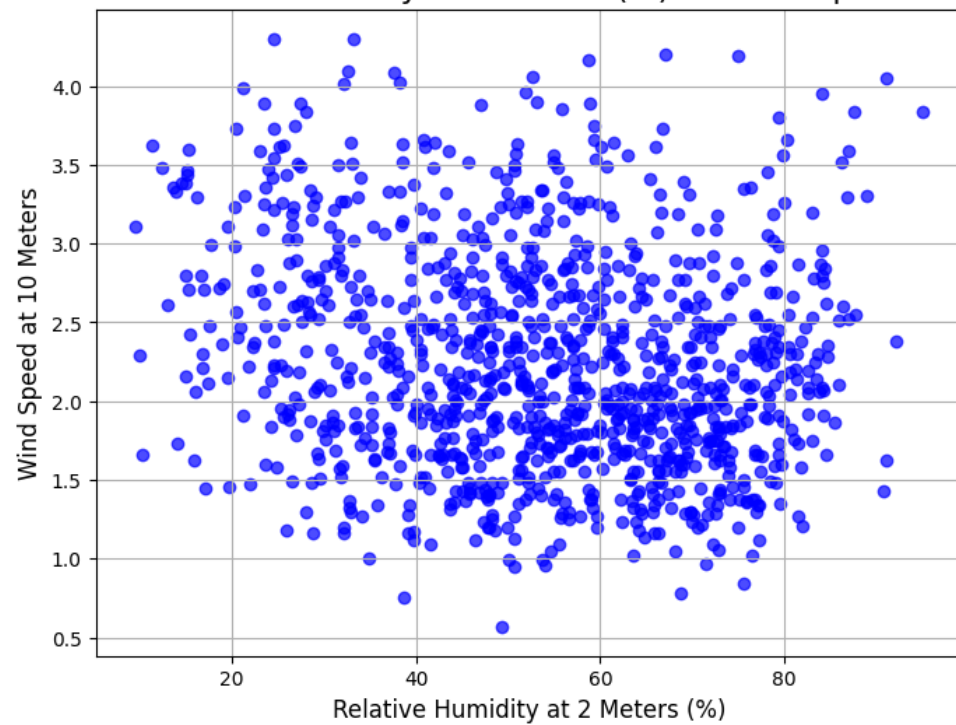
Scatter Plot: Relative Humidity at 2 Meters (%) vs Precipitation Corrected (mm/day)



Relative Humidity vs. Wind Speed:

- Negative correlation.

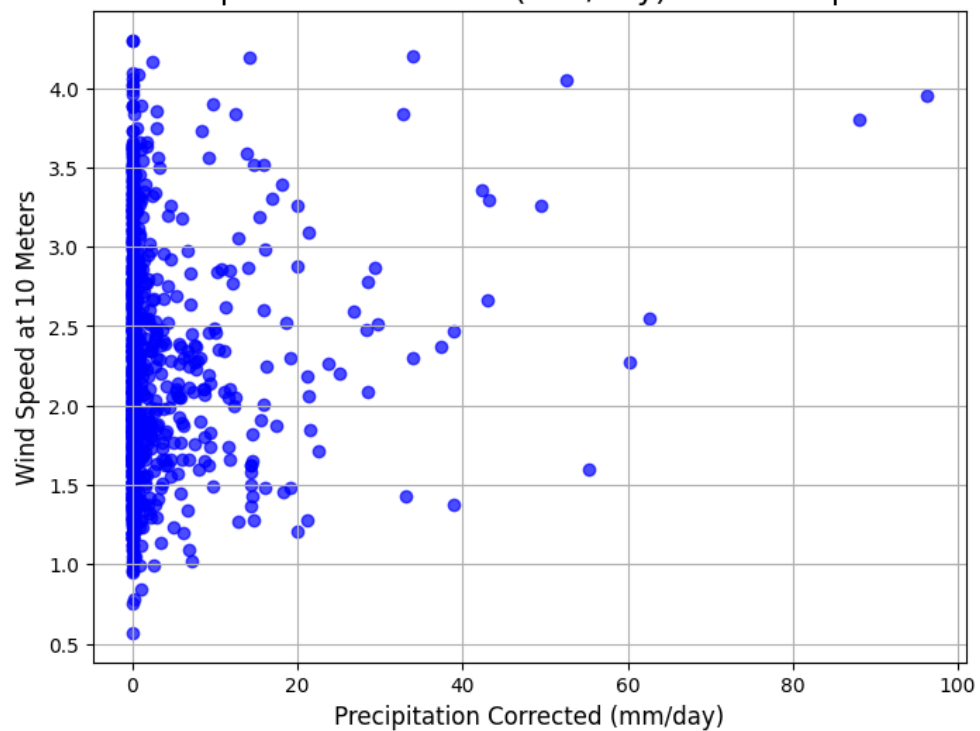
Scatter Plot: Relative Humidity at 2 Meters (%) vs Wind Speed at 10 Meters

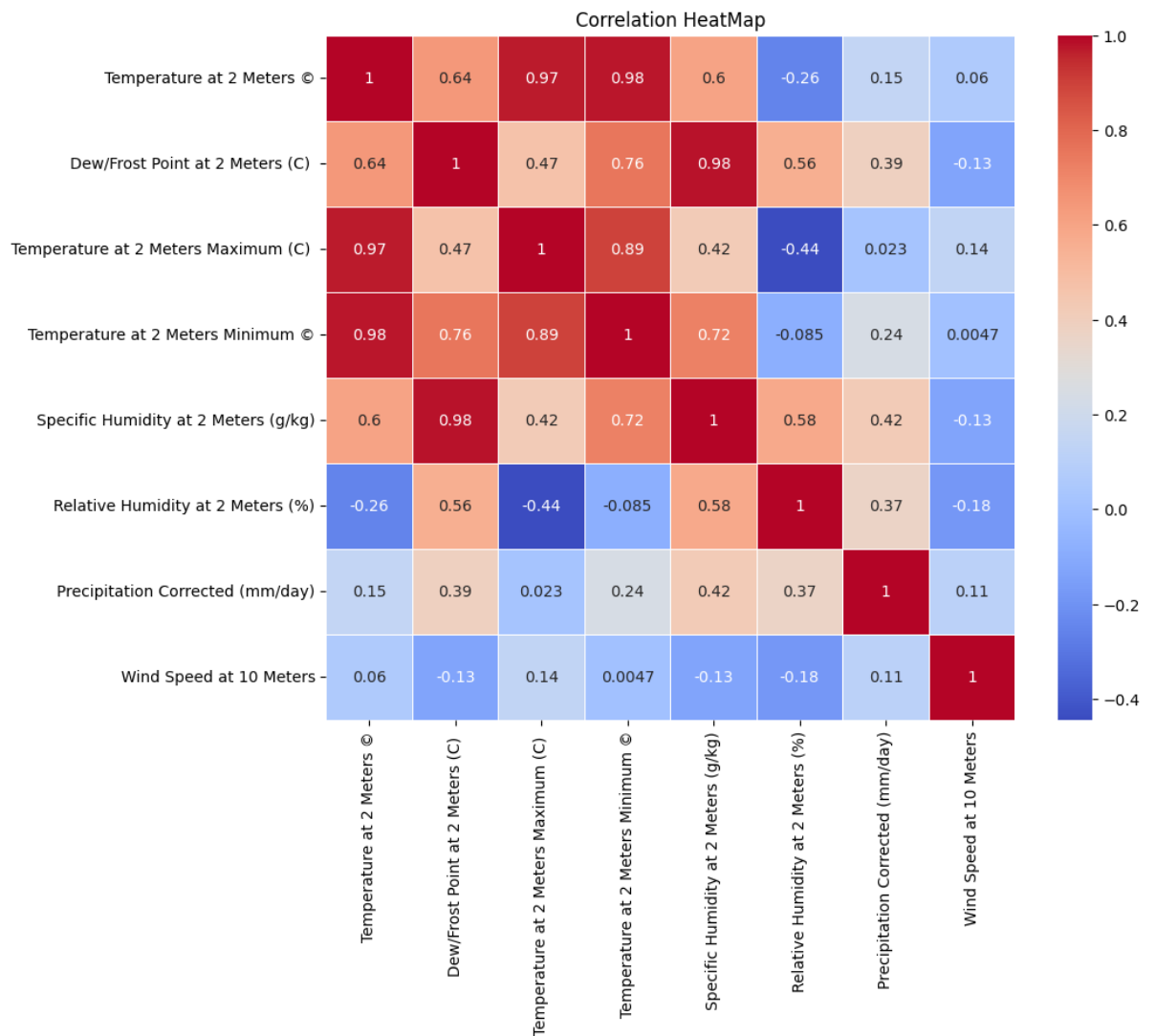


Precipitation vs. Wind Speed:

- Positive but weak correlation (0.11).

Scatter Plot: Precipitation Corrected (mm/day) vs Wind Speed at 10 Meters





3.3 Feature Analysis

Top Features:

- **Specific Humidity (2 M):** Highest importance (0.423), strong relationship with Precipitation Corrected.
- **Dew/Frost Point (2 M):** Second highest importance (0.389).
- **Relative Humidity:** Moderate importance.

Moderate Features:

- Temperature at 2 Meters Minimum and Temperature at 2 Meters.

Low Importance Features:

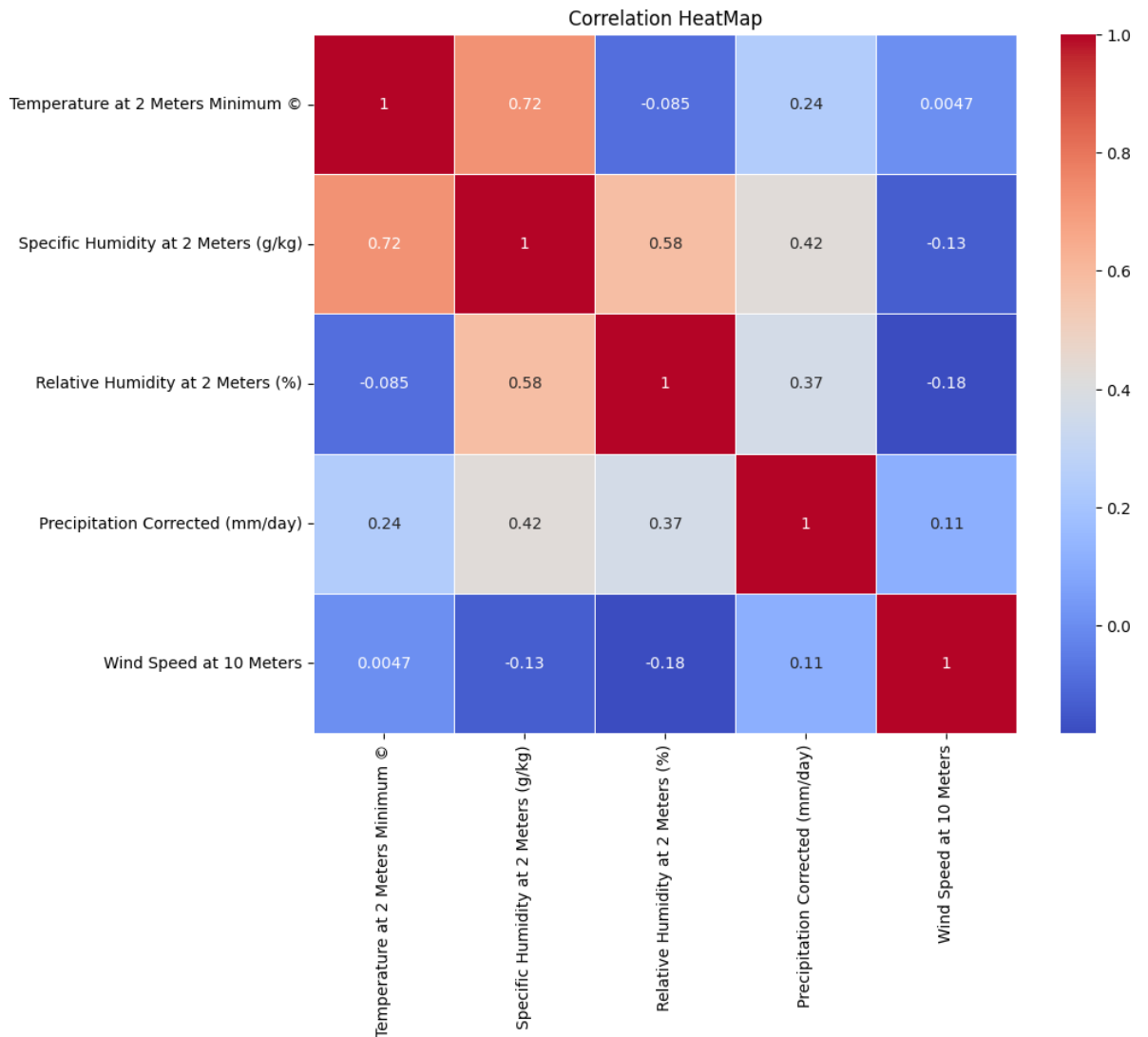
- Temperature at 2 Meters Maximum and Wind Speed at 10 Meters are irrelevant.

```
Specific Humidity at 2 Meters (g/kg)    0.423028
Dew/Frost Point at 2 Meters (C)        0.389070
Relative Humidity at 2 Meters (%)       0.367329
Temperature at 2 Meters Minimum @      0.242463
Temperature at 2 Meters @              0.146545
Wind Speed at 10 Meters                 0.112458
Temperature at 2 Meters Maximum (C)    0.023439
Name: Precipitation Corrected (mm/day), dtype: float64
```

4. Model Training

4.1 Feature Selection

To reduce redundancy and prevent overfitting, features with high correlations among themselves were dropped initially. The correlation matrix was recalculated, and the most relevant features were identified based on their correlation with the target variable (precipitation). Linear regression was then applied to evaluate the selected features. The R^2 score and Mean Squared Error (MSE) were computed both with all features and with only the selected important features. The results indicated that the model's performance (in terms of R^2 and MSE) was better when all features were included, leading to the decision to retain all features for training.



4.2 Model Selection

Linear Regression was chosen as a baseline model because it is simple to implement and interpret. However, due to the presence of non-linear relationships in the dataset, Random Forest was also considered, as it is well-suited for handling non-linear interactions. Gradient Boosting was then used as it performs well on small datasets and is often superior to Random Forest in terms of accuracy and efficiency. For deep learning, a Simple Fully Connected Neural Network was implemented to explore the dataset's potential for complex pattern recognition.

4.3 Model Training

Each model was trained on the pre-processed dataset. In order to optimize the performance hyperparameter tuning was conducted for the Random Forest, Gradient Boosting, and Neural Network models to optimize their performance.

- Random Forest: Parameters such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf were fine-tuned.
- Gradient Boosting: Key parameters like learning rate, number of estimators, and maximum depth were adjusted for optimal results.
- Neural Network: The architecture (number of layers, neurons per layer), activation functions, batch size, and learning rate were tuned using randomized search techniques.

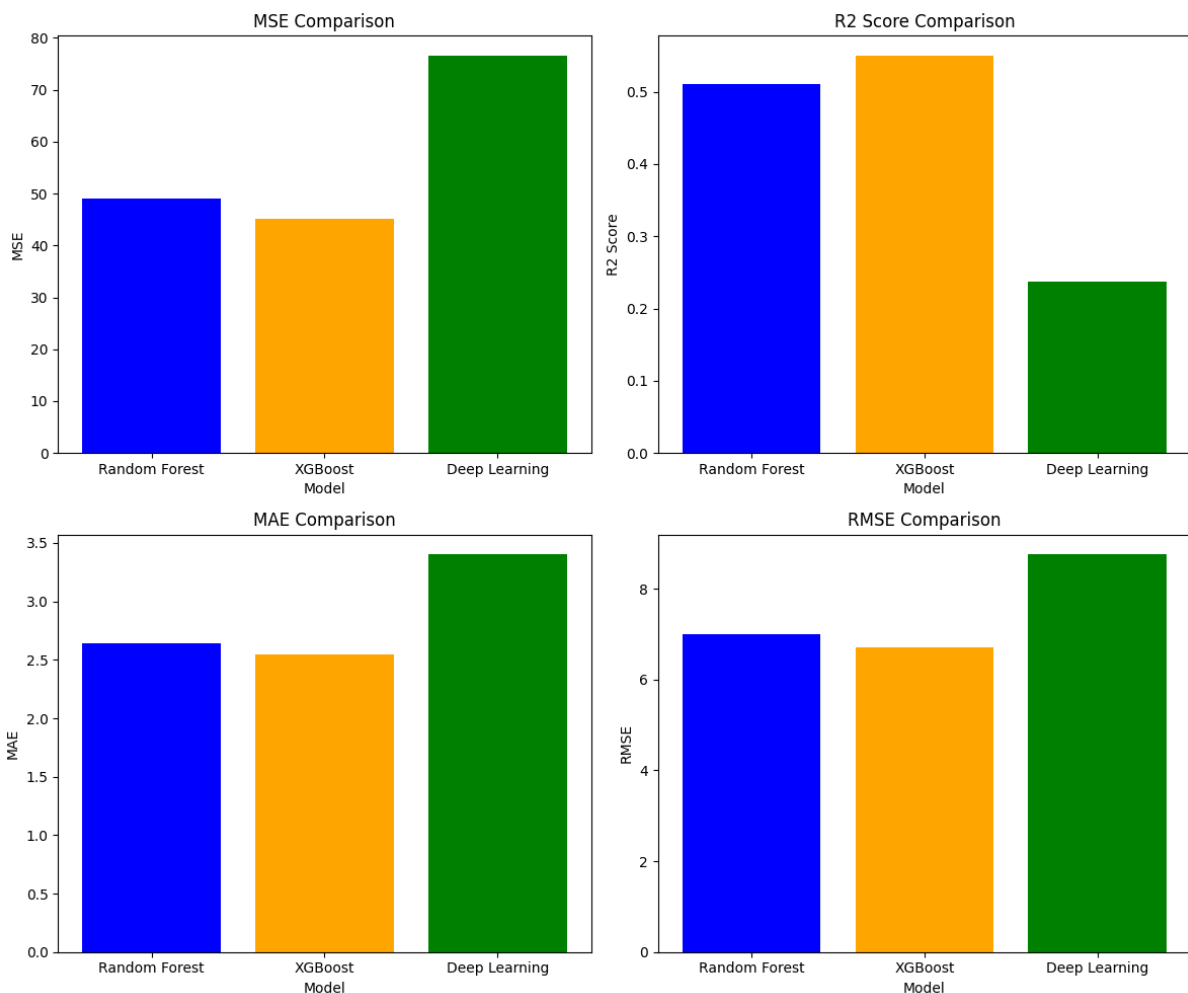
4.4 Model Evaluation

Gradient Boost emerged as the best performing model with the lowest errors like MSE, RMSE, MAE and highest R^2 score, which makes it most accurate for our dataset. Random forest also performs well capturing around 51% of the variance i.e. $R^2 = 0.51$. However, its errors are greater than Gradient Boost showing it provides reasonably accurate predictions but not the best. On the other hand, deep learning model performed the worst. Its higher MSE and RMSE indicates larger prediction errors and R^2 score of 0.24 shows that it explains only 24% of the variance. This poor performance is likely due to the small dataset size which is 1069 rows which is insufficient for training deep learning models effectively.

```
Random Forest Performance:
MSE: 49.06670340014337
R2 Score: 0.5107141691818763
Mean absolute error MAE : 2.644459770955416
RMSE: 7.004762908203487
-----
```

```
Gradient Boost (xgbost) Performance:
MSE: 45.13633780430232
R2 Score: 0.5499071873126692
Mean absolute error MAE : 2.544224936418218
RMSE: 6.7183582670398225
-----
```

```
Deep learning model Performance:
MSE: 76.56018550619504
R2 Score: 0.23655327590483988
Mean absolute error MAE : 3.400470018386841
RMSE: 8.749867742211595
```



5. Conclusion and Future Work

This project demonstrated the effective use of weather data for precipitation prediction. Key findings include the strong influence of features like specific humidity and relative humidity on precipitation, while variables like wind speed and temperature were less significant. However, we kept all the features in model training as it improves accuracy.

Gradient Boosting emerged as the most accurate model, achieving the lowest errors and the highest R^2 score, making it well-suited for this dataset. Random Forest provided reasonably accurate results but was outperformed by Gradient Boosting. The deep learning model, however, showed limited effectiveness due to the small dataset size, highlighting the importance of data quantity in training complex models.

Overall, the project achieved its primary objectives of analyzing weather data and identifying key factors influencing precipitation. Future work can include collecting more data over extended periods or include additional features to improve model performance, especially for deep learning approaches. Furthermore, include broader climatic variables, such as seasonal patterns or geographic data, to increase prediction accuracy.