National University of Computer and Emerging Sciences

Weather Data Analysis (Rain Prediction)

Team

Suleiman Asif……21L-1818
Jazib Zafar……21L-1803

Supervised by

Ms. Maimoona Akram

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

December 2024

Table of Contents

Introduction....2
Data Preparation....3
Introduction....3
Data Loading and Exploration....3
Data Cleaning....5
Data Transformation....5
Data Analysis....6
Introduction....6
Univariate Analysis....6
Bivariate and Multivariate Analysis....18
Feature Analysis....24
Model Training....27
Introduction....27
Feature Selection....27
Model Selection....29
Model Training....31
Model Evaluation....34
Comparison....36

Project Vision: To Predict Rain accurately

Introduction

Weather prediction plays a crucial role in various sectors, including agriculture, disaster management, transportation, and daily decision-making. Accurate forecasting can significantly mitigate risks and help in planning resources effectively. This project focuses on analyzing and predicting precipitation levels, an essential component of weather forecasting, using a dataset containing key meteorological variables such as temperature, humidity, wind speed, and precipitation.

The dataset comprises weather data collected over multiple years, including numerical features like temperature, specific humidity, and wind speed, alongside categorical features like the year of observation. These features are critical in understanding the patterns and factors influencing precipitation levels, the primary target variable in this study.

The main objectives of this project are:

1. To clean and prepare the dataset for robust analysis and modeling.
2. To explore and analyze key weather features to understand their relationships and importance for precipitation prediction.
3. To build and compare machine learning and deep learning models for accurate precipitation prediction.
4. To provide actionable insights based on the analysis, aiding in future improvements in weather forecasting systems.

The expected outcomes include identifying the most significant features affecting precipitation, understanding the relationships between weather variables, and selecting the best predictive model for precipitation forecasting. These insights and models could be a foundation for developing more advanced weather prediction systems.

Detailed Dataset Overview

This dataset is scraped from “NASA POWER” and it includes weather conditions recorded over time to predict precipitation.

- **Dataset Name:** Weather Data
- **Number of Rows:** The dataset contains 1,094 rows.
- **Number of Columns:** The dataset contains 11 columns.
- **Features:**
    - YEAR, MO (Month), DY (Day): Date of the weather record.
    - Temperature at 2 Meters (°C): Average daily temperature measured at 2 meters above the ground.
    - Dew/Frost Point at 2 Meters (°C)
    - Temperature at 2 Meters Maximum (°C): Daily maximum.
    - Temperature at 2 Meters Minimum (°C): Daily minimum.
    - Specific Humidity at 2 Meters (g/kg): Humidity at 2 meters above the ground.
    - Relative Humidity at 2 Meters (%): Relative humidity as a percentage.
    - Precipitation Corrected (mm/day): Amount of precipitation in millimeters.
    - Wind Speed at 10 Meters (m/s): Wind speed at 10 meters above the ground.

Data Preparation

Introduction

This phase of the project aims to conduct data wrangling on weather data, an essential first step before performing any detailed analysis. Data wrangling ensures that the dataset is clean, structured, and ready for further exploratory data analysis and model training. The weather dataset used in this project provides detailed information such as temperature, humidity, and wind speed, which will later be used for analyzing weather patterns and trends which are useful especially for rain prediction.

Data Loading and Exploration

The data was loaded into the environment using “pandas.read_csv()”.

First few records of the dataset using head() and tail():

- df.head():

```
[27] 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():

```
[ ] 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

Summary of the dataset:

- df.describe():

Project Vision: To Predict Rain accurately

```
[17] 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	2.590137	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.info():

```
[ ] 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
```

Data Cleaning

- **Handling Missing Data:**
  The dataset was checked for missing values using “df.isnull().sum()”. It was found that there were no missing values, so no imputation or deletion was required.
- **Duplicate Removal:**
  Duplicate records were checked using “df.duplicated().sum()”. No duplicate records were found, so no rows were removed.
- **Data Type Conversion:**
  No need for a change of datatypes of any column as datatypes are in the correct format as required.
- **Outlier Detection:**
  Outliers were identified in the numerical columns using the IQR, Interquartile Range, method. Outliers were found in two columns: 'Precipitation Corrected (mm/day)' and 'Wind Speed at 10 Meters.' These outliers were removed by applying the lower and upper thresholds derived from the IQR method, and the dataset was updated accordingly. Although they are removed at this level but considering the fact that these outliers may affect weather patterns and can be important indicators that is why a dataset with the inclusion of these outliers is used for future tasks like model training and further analysis.

Data Transformation

- **Scaling:**
  The numerical columns were normalized using the “MinMaxScaler” technique from “sklearn”. “MinMaxScaler” was chosen as the preferred method because weather data typically falls within specific ranges, and this technique preserves the distribution while scaling the values to a consistent range. The scaled dataset was saved as “transformed_weather_data.csv”.
- **Categorical Encoding:**
  The dataset did not contain categorical variables that required label encoding or one-hot encoding. The YEAR, MO, and DY columns, while representing categorical information, were treated as ordinal features, preserving their natural order without applying encoding. This decision was based on the need to maintain the temporal nature of the data.

This concludes our first phase of the project which focuses on preparing the dataset for further analysis by ensuring it is clean, free from outliers, and properly scaled.

Data Analysis

Introduction

The primary objective of this Iteration of the project is to analyze weather data with a focus on predicting rainfall. In this iteration, we conducted Exploratory Data Analysis (EDA), a critical step to uncover patterns, relationships, and anomalies within the dataset. EDA provides insights into the structure of the data and guides the feature selection process for predictive modeling.
The dataset used in this analysis consists of historical weather records, including continuous features such as temperature, humidity, and wind speed, and categorical features such as year, month, and day. These variables are analyzed to understand their distributions and potential influences on precipitation.

Univariate Analysis

Objective:

Univariate analysis examines individual variables to understand their statistical properties, distributions, and any anomalies like outliers or skewness.

Methodology:

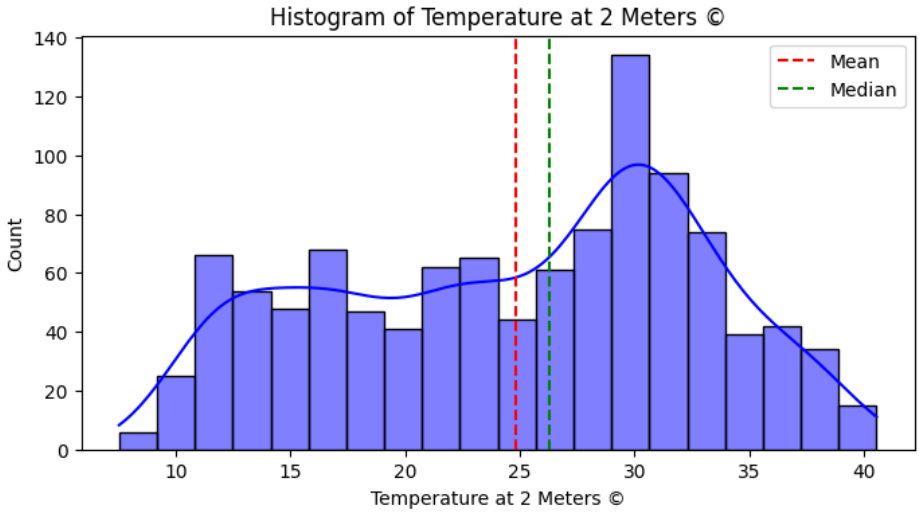
- **Continuous Variables**:
  - **Visualizations**:
    - Histograms were used to visualize frequency distributions.
    - Box plots were used to detect outliers and assess the spread of the data.
  - **Statistics**:
    - Mean, median, mode, and interquartile range (IQR) were calculated for each variable to summarize its central tendency and variability.
- **Categorical Variables**:
  - Frequency distributions were analyzed using count plots.
  - Bar plots were used to visually represent the counts for each category.

Findings Numerical Variables:

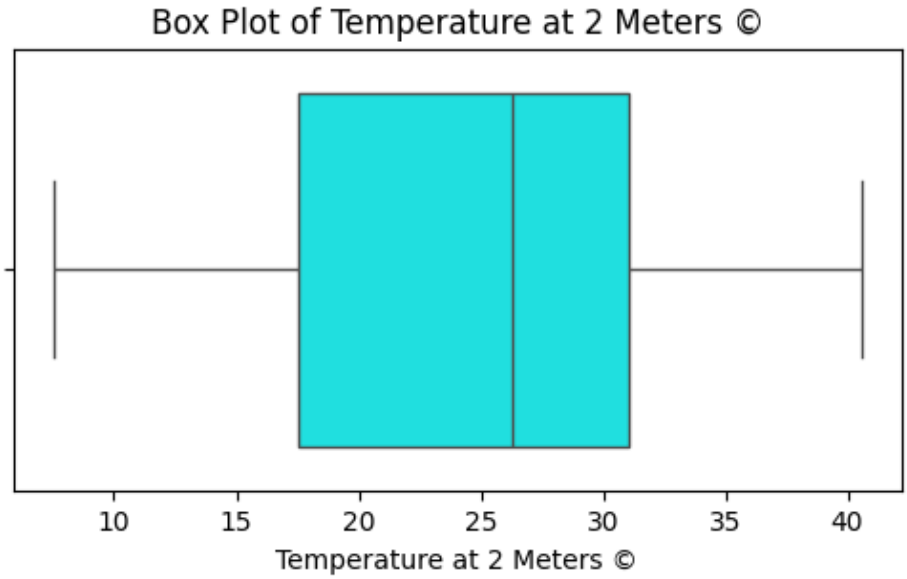
1. **Temperature at 2 Meters (°C):**

   - **Statistics**:

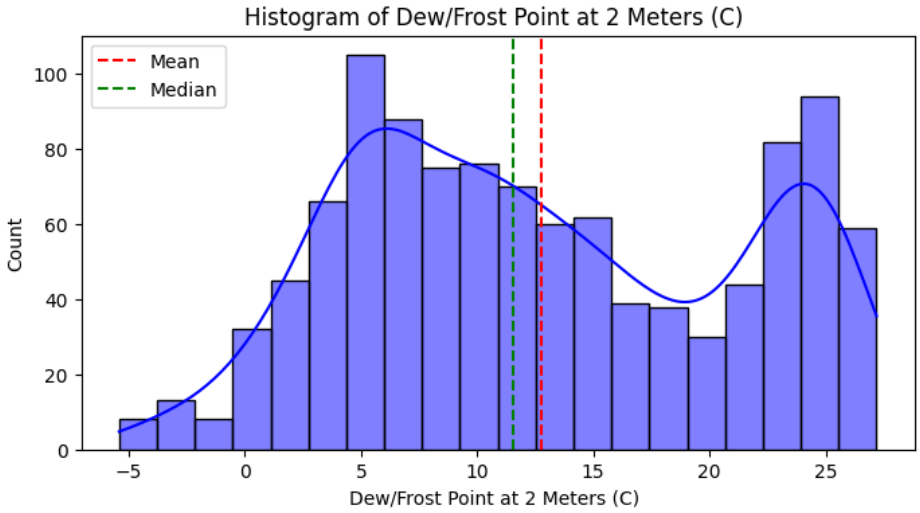
     - **Mean**: 24.82, **Median**: 26.28, **Mode:** 17.48, **IQR**: 13.49
     - **Min**: 7.54, **Max**: 40.56
     - **Std. Dev.**: 8.07
   - **Diagram (Observations):**
     - Histogram: The mean (red dashed line) is slightly smaller than the median (green dashed line), suggesting a slight left skew in the data. This might
       imply the presence of some smaller temperature values pushing the mean down.

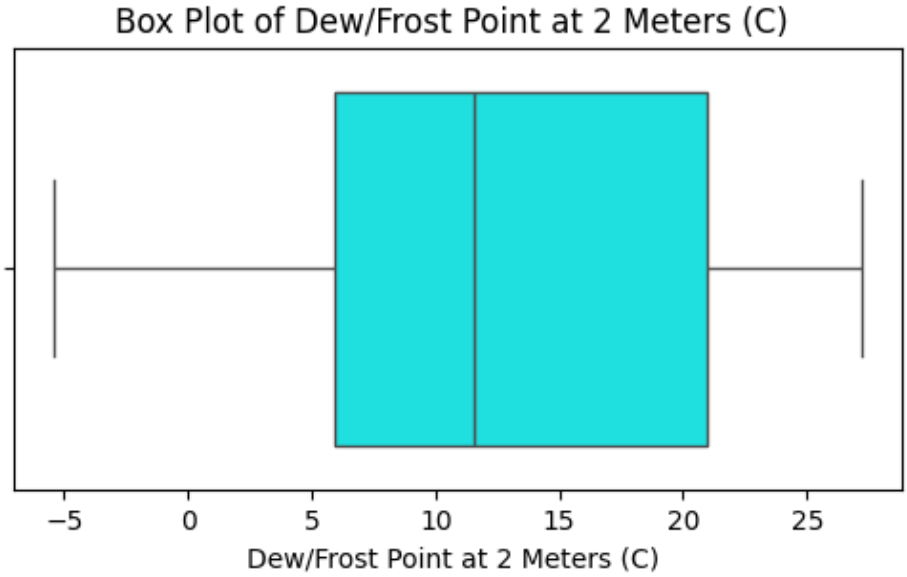


- Box Plot: The box extends from about 17°C to 32°C, indicating the interquartile range (IQR), which contains the middle 50% of the data. There are no visible outliers, as the whiskers extend to the minimum and maximum values without any points outside this range

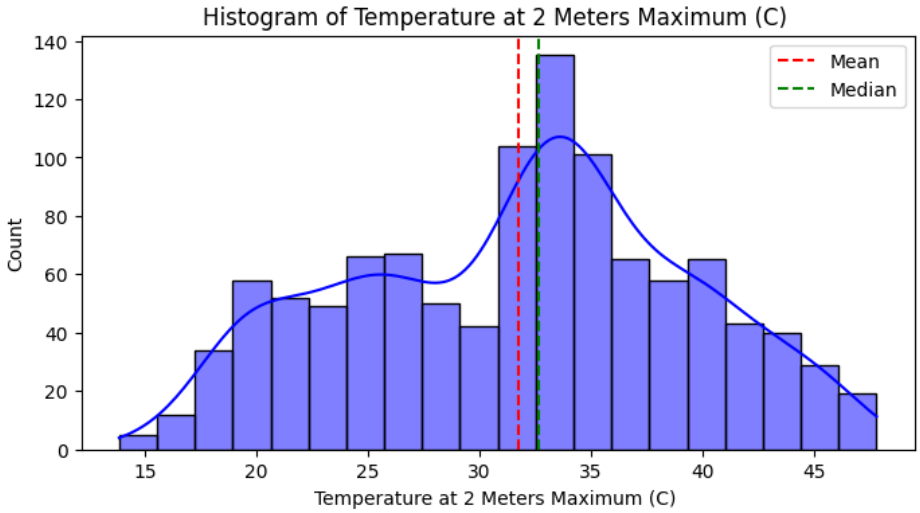


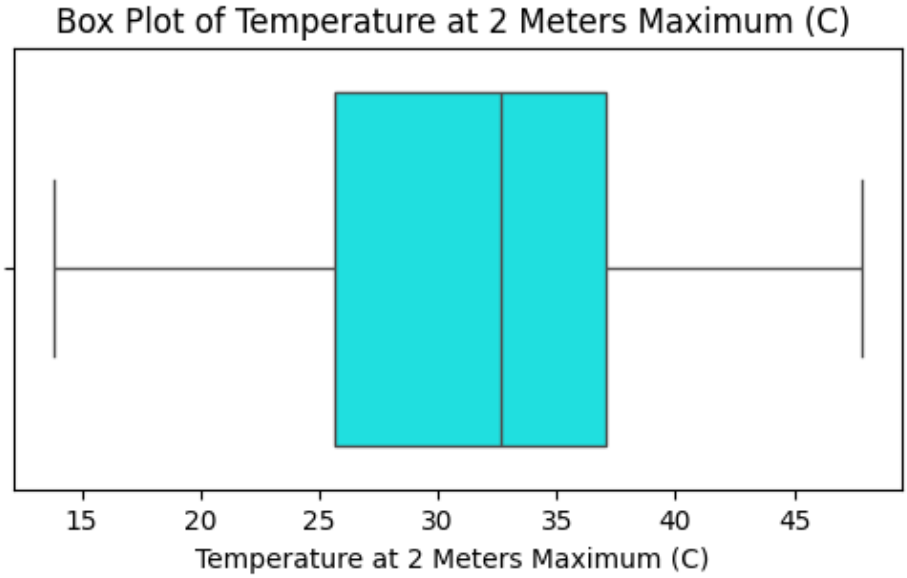
2. **Dew/Frost Point at 2 Meters (°C):**
   - **Statistics:**
     - **Mean**: 12.79, **Median**: 11.54, **Mode:** 24.15, **IQR**: 15.04
     - **Min:** -5.41**, Max:** 27.19
     - **Std. Dev.:** 8.22
   - **Diagram (Observations):**
     - Values range from below freezing to warm dew points.
     - Mild right skew.
     - No significant outliers observed.



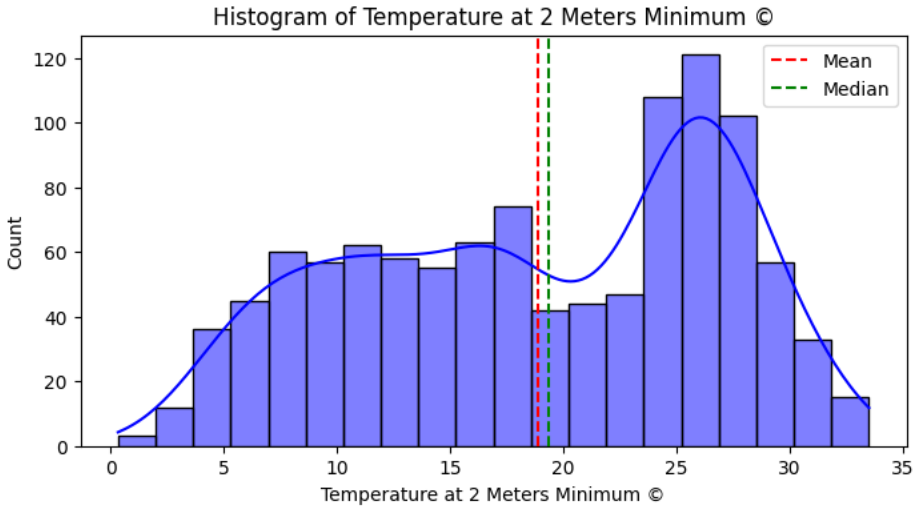


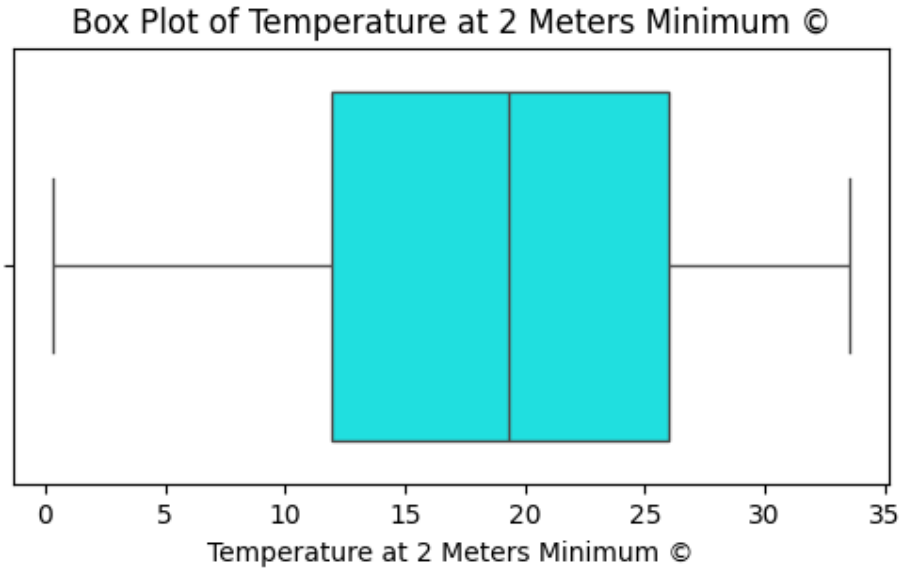
3. **Temperature at 2 Meters Maximum (°C):**
    - **Statistics:**
        - **Mean**: 31.72, **Median**: 32.66, **Mode**: 30.90, **IQR**: 11.40
        - **Min**: 13.83, **Max**: 47.82
        - **Std. Dev.**: 7.62
    - **Diagram (Observations):**
        - Higher variability compared to other temperature columns.
        - Left skew due to occasional extremely high maximum temperatures.



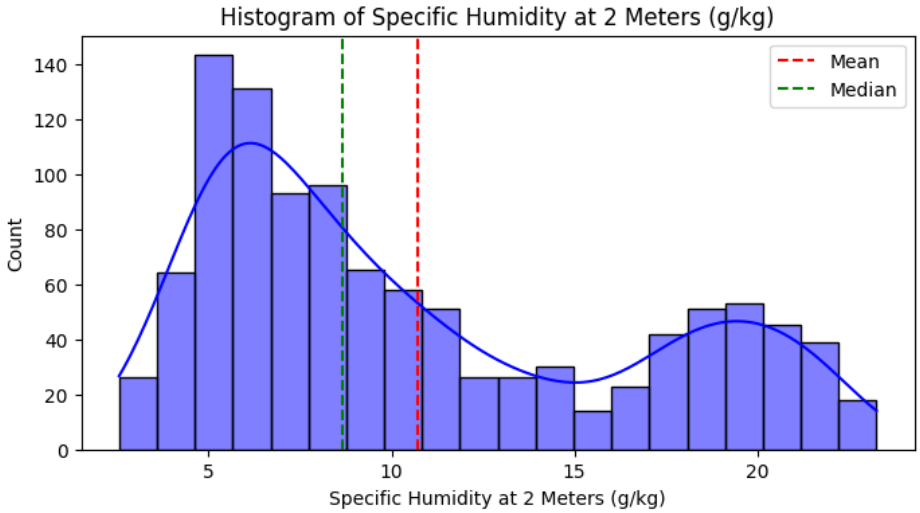


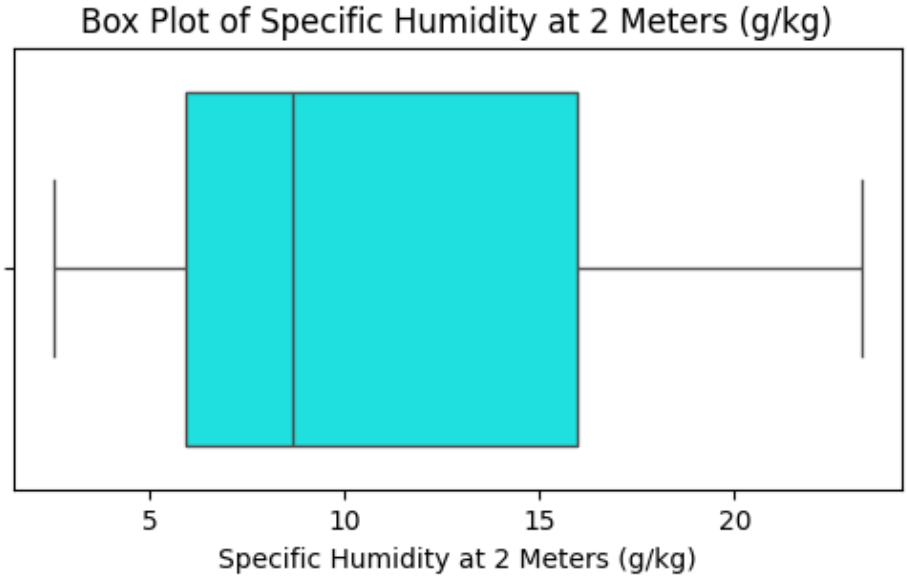
4. **Temperature at 2 Meters Minimum (°C):**
    - **Statistics:**
        - **Mean**: 18.86, **Median**: 19.36, **Mode**: 6.83, **IQR**: 14.05
        - **Min**: 0.32, **Max**: 33.51
        - **Std. Dev.**: 8.06
    - **Diagram (Observations):**
        - Minimum temperatures are generally consistent, but outliers near freezing are rare.
        - Distribution is relatively symmetrical especially if we see a box plot.



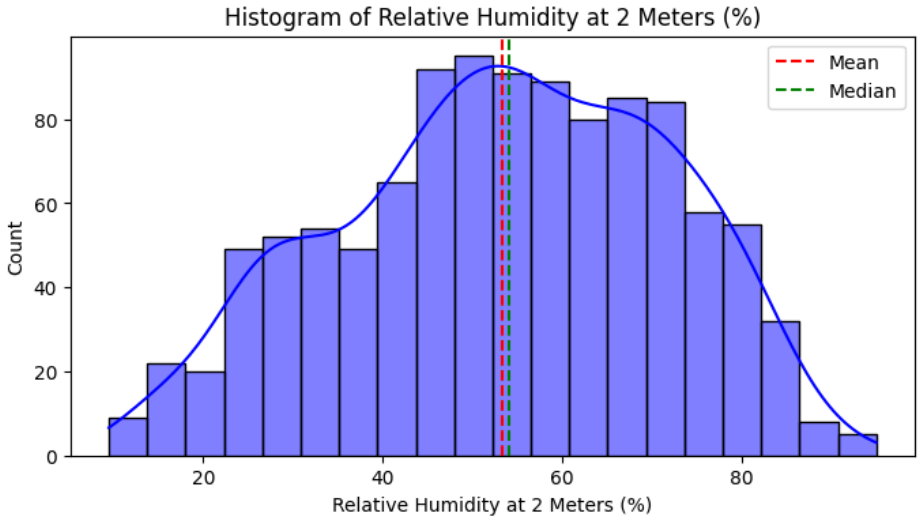


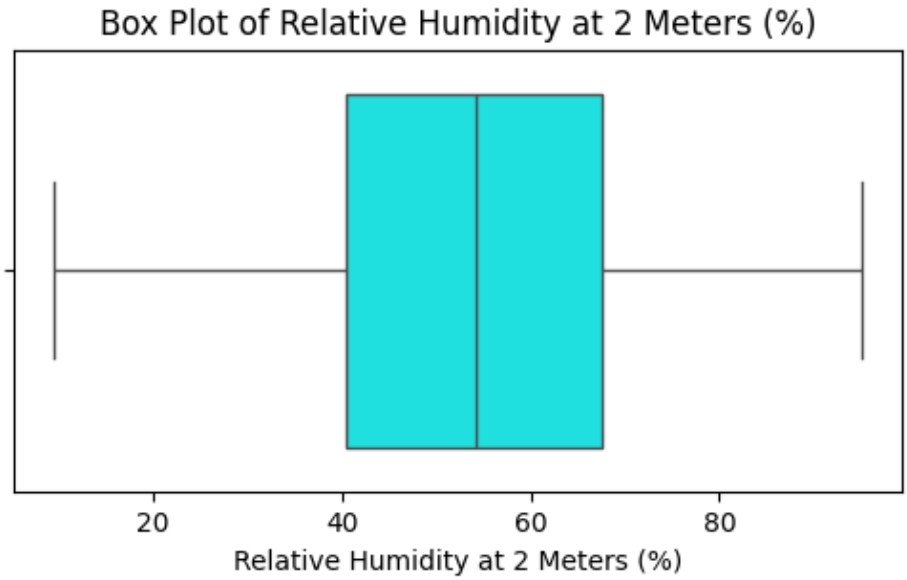
5. **Specific Humidity at 2 Meters (g/kg):**
    - **Statistics:**
        - **Mean**: 10.73, **Median**: 8.67, **Mode**: 5.55, **IQR**: 10.05
        - **Min**: 2.56, **Max**: 23.25
        - **Std. Dev.**: 5.76
    - **Diagram (Observations):**
        - Right skew due to some highly humid days.



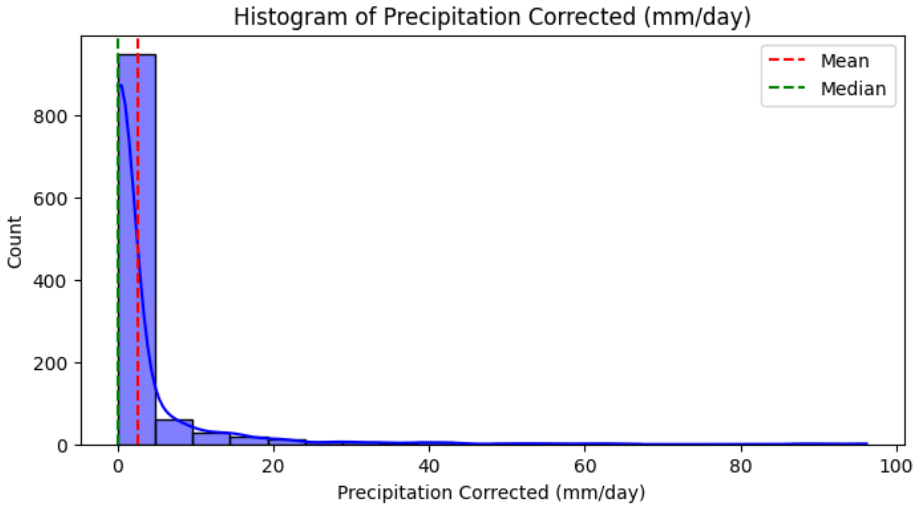


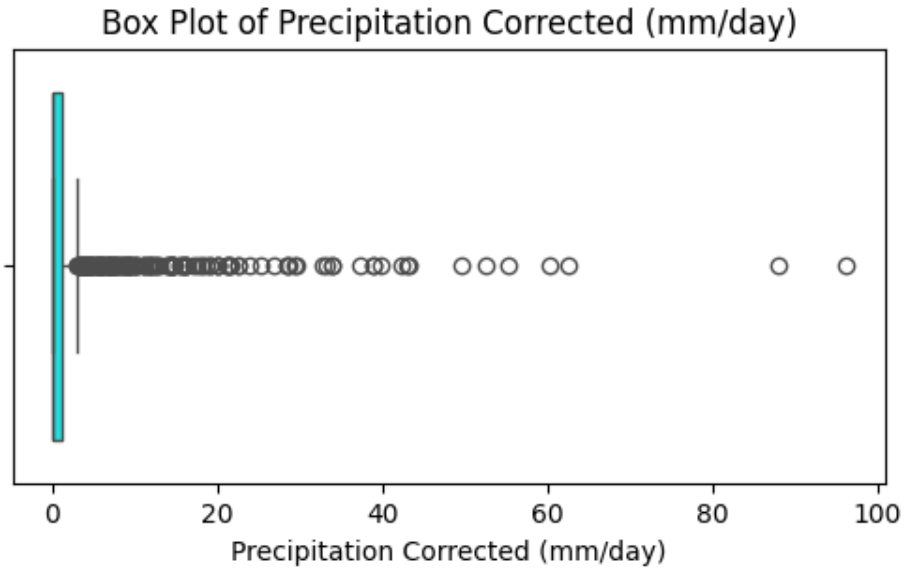
6. **Relative Humidity at 2 Meters (%):**
    - **Statistics:**
        - **Mean**: 53.42, **Median**: 54.16, **Mode**: 50.81, **IQR**: 27.25
        - **Min**: 9.50, **Max**: 95.06
        - **Std. Dev.**: 18.17
    - **Diagram (Observations):**
        - Fairly balanced distribution with a slight skew.



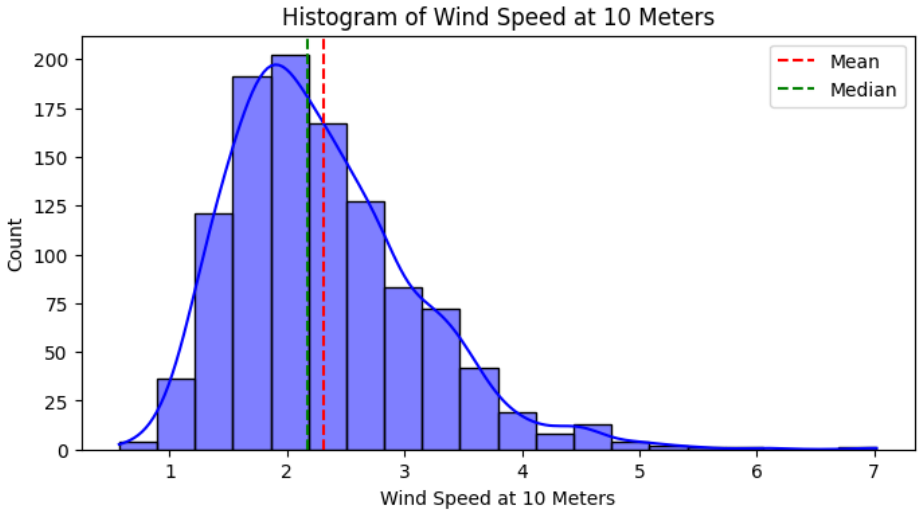


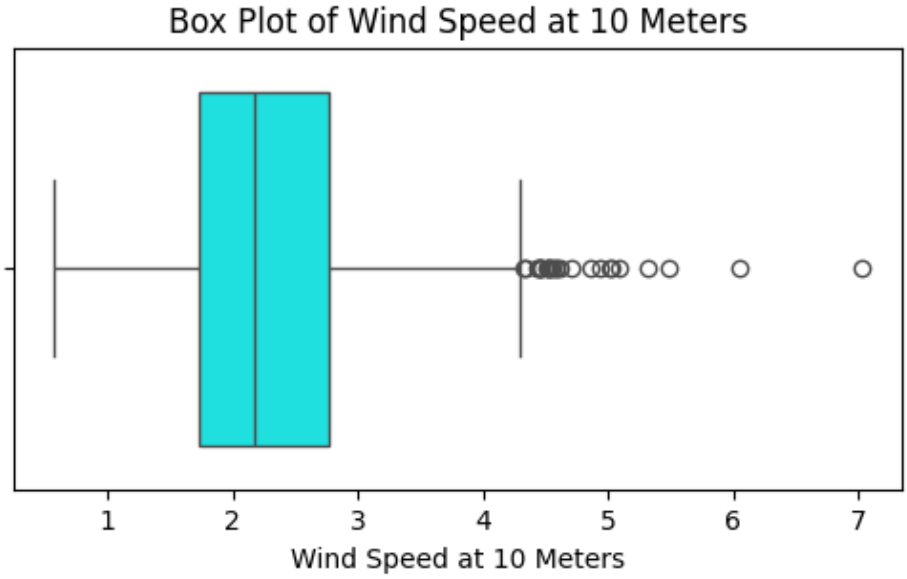
7. **Precipitation Corrected (mm/day):**
    - **Statistics:**
        - **Mean**: 2.59, **Median**: 0.00, **Mode**: 0.00, **IQR**: 1.20
        - **Min**: 0.00, **Max**: 96.17
        - **Std. Dev.**: 7.76
    - **Diagram (Observations):**
        - Highly skewed distribution with most days having no precipitation.
        - Extreme outliers reflect rare days of very high rainfall.





8. **Wind Speed at 10 Meters (m/s):**
    - **Statistics:**
        - **Mean**: 2.31, **Median**: 2.17, **Mode**: 2.09, **IQR**: 1.03
        - **Min**: 0.57, **Max**: 7.02
        - **Std. Dev.**: 0.79
    - **Diagram (Observations):**
        - Symmetrical distribution, indicating consistent wind speeds.
        - Few minor outliers for high wind speeds.



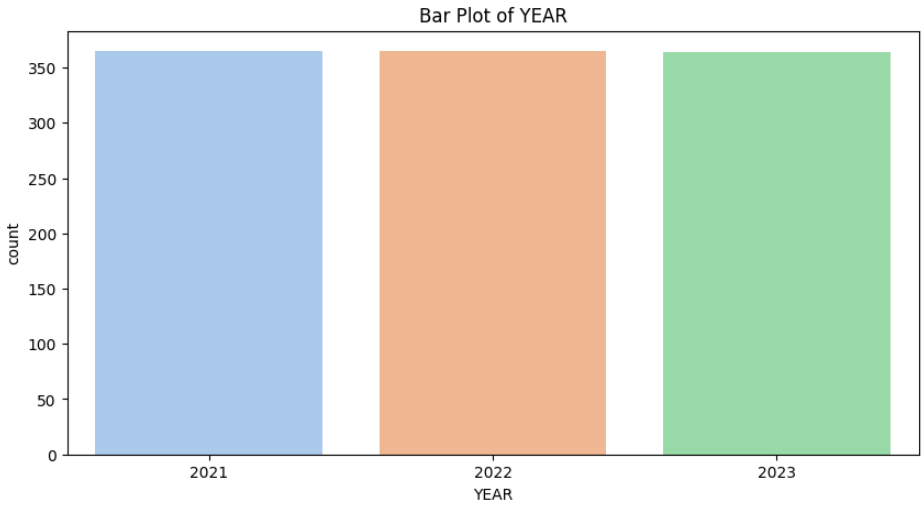


Key Insights from Findings of Numerical Columns/Variables:

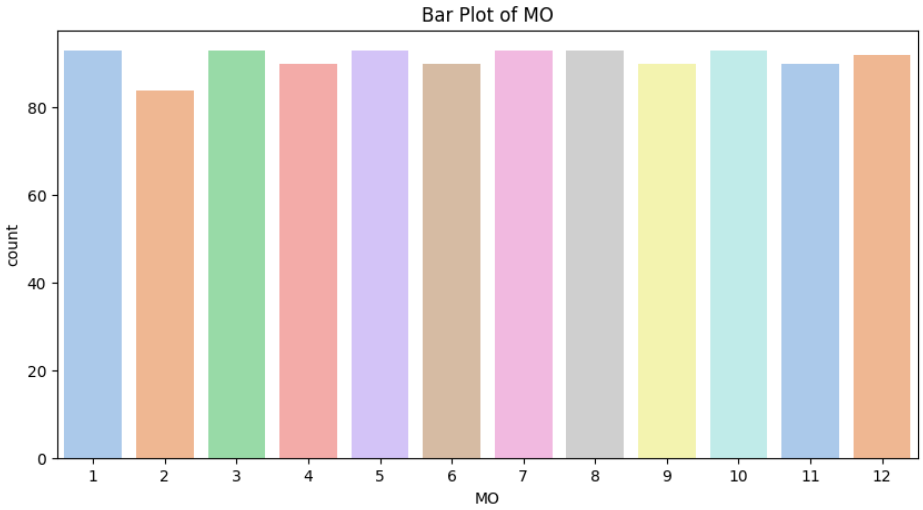
- **Skewness**:
  - Most features have relatively normal distributions, except for `Precipitation Corrected`, which is heavily skewed due to a large number of zero values and a few extreme outliers.
- **Outliers**:
  - Significant outliers exist for `Precipitation Corrected` (extreme rainfall days).
  - Minor outliers for `Wind Speed`.
- **Central Tendency**:
  - Features like `Temperature` and `Humidity` have consistent means and medians, suggesting reliability for modeling.

Categorical Variables:

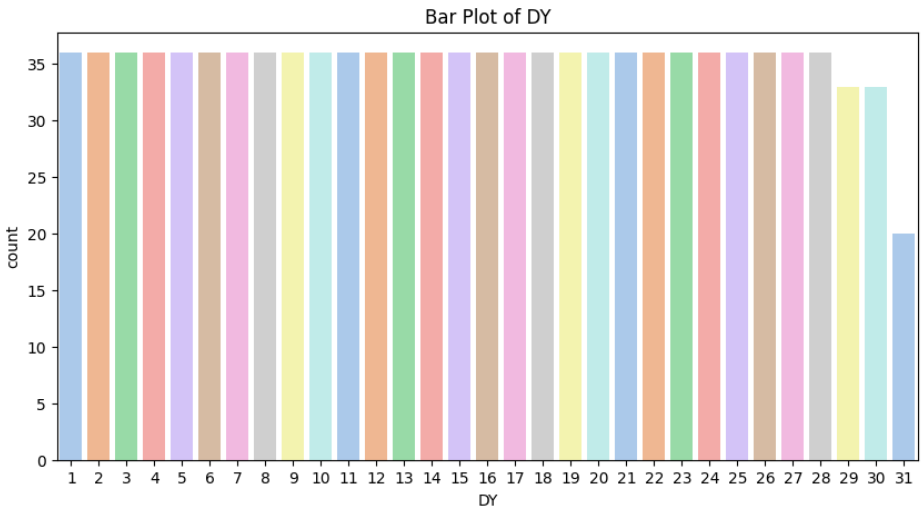
1. **YEAR**:
    - Observations:
        - Data spans three years (2021, 2022, and 2023), with nearly equal counts for each year.
        - The distribution is balanced, and no irregularities were observed.
    - Visuals:
        - The bar plot for YEAR shows equal heights, confirming the uniform distribution.



2. **MO (Month):**
    - Observations:
        - Most months have between 90 and 93 records, with February having slightly fewer (84).
        - No major imbalance is observed, but February's lower count aligns with its shorter duration.
    - Visuals:
        - The bar plot for MO shows slightly lower counts for February, while other months are nearly uniform.



3. **DY (Day):**
    - Observations:
        - Most days have 36 records, except for the 29th, 30th, and 31st, which have fewer counts.
        - These variations are due to the different lengths of months as February often lacks the 30th and 31st.
        - The 31st has the lowest count (20 records), as not all months have a 31st day.
    - Visuals:
        - The bar plot for days clearly shows the decreasing counts for the 29th, 30th, and 31st.



Conclusion:

1. Continuous Variables:

- **Histograms**:

    - Visualized the overall distribution of numerical features, showing patterns such as normality or skewness.
- **Box Plots**:
    - Identified the presence of outliers, especially for precipitation.

2. Categorical Variables:

- **Bar Plots**:
    - Visualized counts for `MO` and `DY`, confirming no missing categories and highlighting seasonal trends.

Bivariate and Multivariate Analysis

Objective:

This section explores the relationships between variables to identify trends, correlations, and patterns. Both bivariate (pairwise relationships) and multivariate (interactions among multiple variables) analyses were conducted.

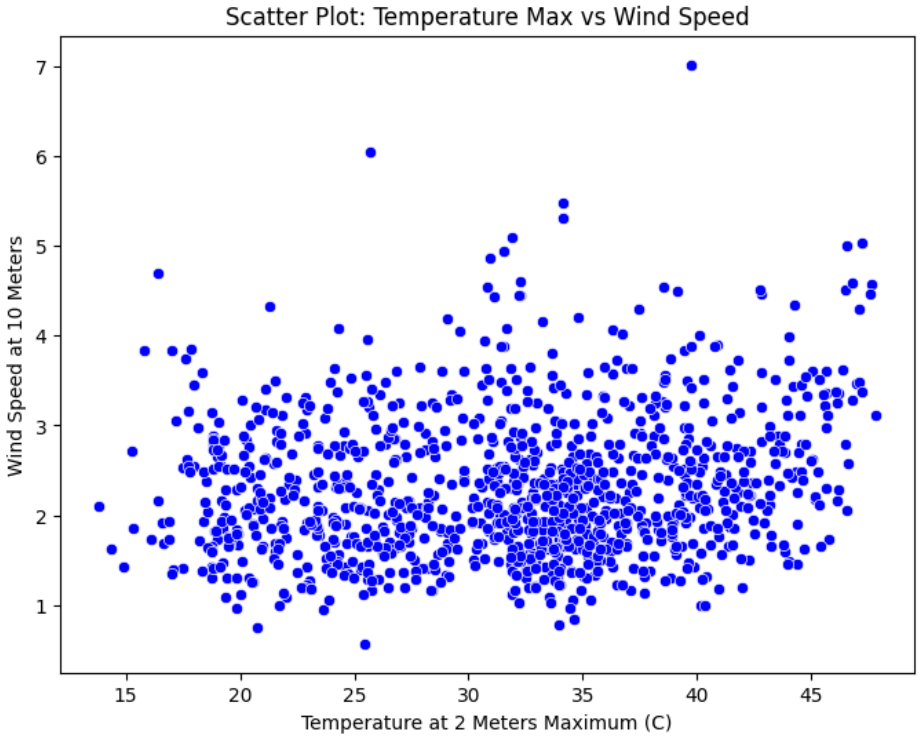
1. Bivariate Analysis

Scatter Plots:

Scatter plots were created to visualize pairwise relationships between selected continuous variables.

Findings:

1. **Temperature at 2 Meters Maximum (°C) vs. Wind Speed at 10 Meters**:
    - **Observation**:
        - A weak positive correlation is observed between maximum temperature and wind speed.
    - **Visualization**:
        - The scatter plot shows majority concentration below '3' of wind speed.



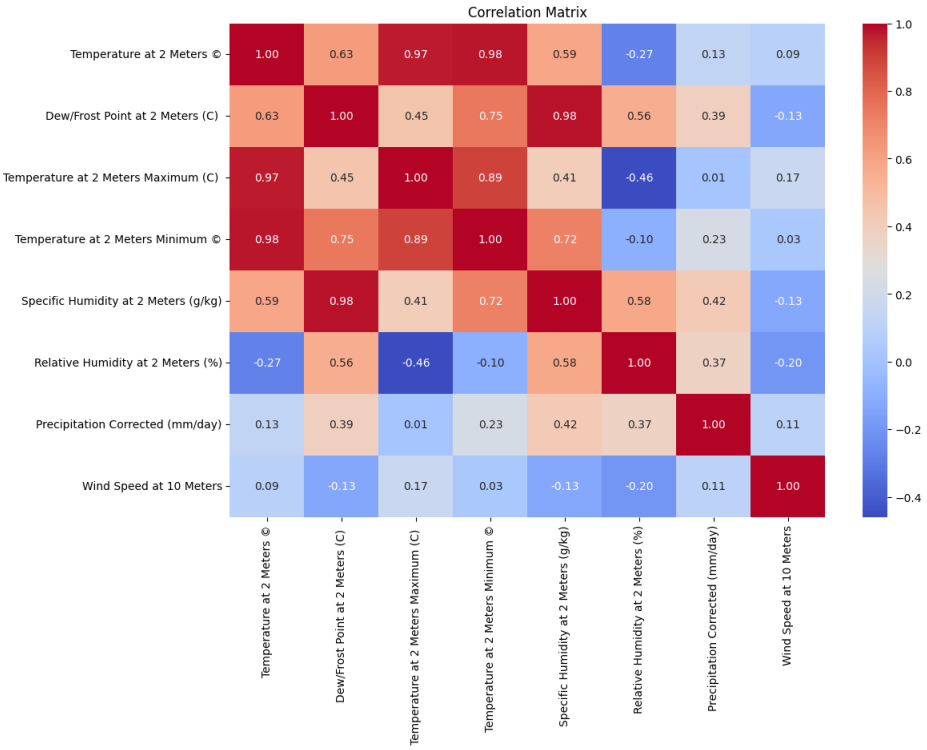
Correlation Matrix:

A correlation matrix was generated to quantify relationships between all numerical variables. A heatmap was used for visualization.

Weather Data Analysis (Rain Prediction)

Findings:

- **Strong Positive Correlations**:
  - **Temperature Max (°C)** and **Temperature Min (°C)**: 0.89 (strong positive correlation).
- **Moderate Positive Correlations**:
  - **Relative Humidity (%)** and **Precipitation Corrected (mm/day)**: 0.37 (moderate positive correlation).
  - **Dew/Frost Point (°C)** and **Relative Humidity (%)**: 0.56 (moderate positive correlation).
- **Negative Correlations**:
  - **Temperature at 2 Meters Minimum (°C)** and **Relative Humidity (%)**: -0.10(weak negative correlation).
- **Weak or No Correlation**:
  - **Wind Speed at 10 Meters** shows weak correlations with other variables.



Group-by Analysis:

Each numerical feature was grouped by month (`MO`) to compute the average value. This revealed seasonal patterns.

Findings:

1. **Precipitation Corrected (mm/day)**:
   - Highest precipitation occurs during the wet season (e.g., July, August, Sep).
   - Minimal precipitation during dry months (e.g., February, November, December).
2. **Temperature**:
   - **Maximum Temperature** peaks in the summer months (e.g., June).

- **Minimum Temperature** is lowest in the winter months (e.g., January, December).

3. **Relative Humidity**:
   - Humidity is highest in August, and September and lowest during (April to June).
4. **Wind Speed**:
   - Wind speeds are relatively consistent across months, with slight increases during (January to July).

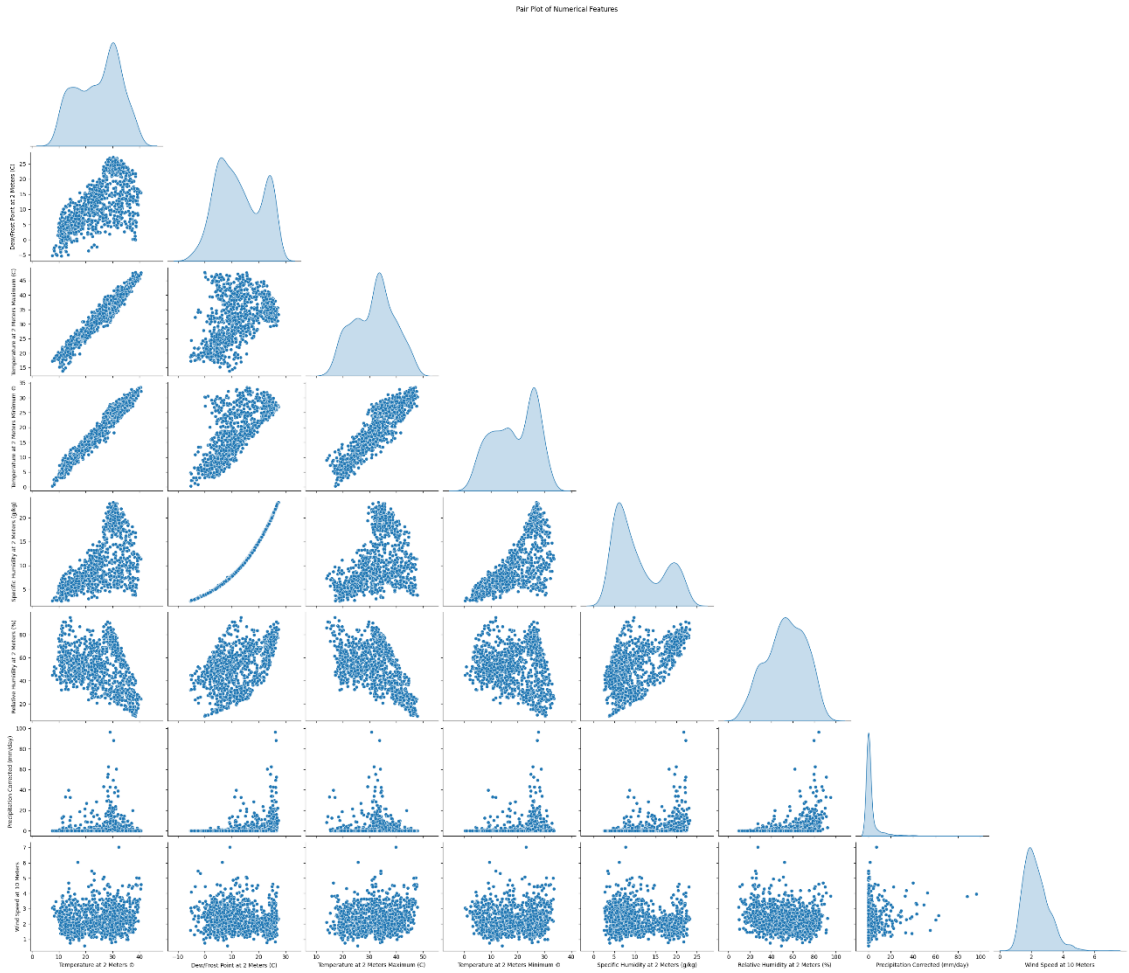
Notes: For all 8 Diagrams please refer to Python's Jupiter notebook.

2. Multivariate Analysis

Pair Plots**:** Pair plots were used to visualize the relationships between all numerical variables.

Findings:

1. **Temperature and Humidity**:
   - **Temperature Minimum (°C)** is negatively correlated with **Relative Humidity (%).**
2. **Clusters**:
   - Clusters in precipitation values are visible, particularly during specific months.
3. **Outliers**:
   - Significant outliers are present in **Precipitation Corrected (mm/day),** reflecting extreme rainfall events.



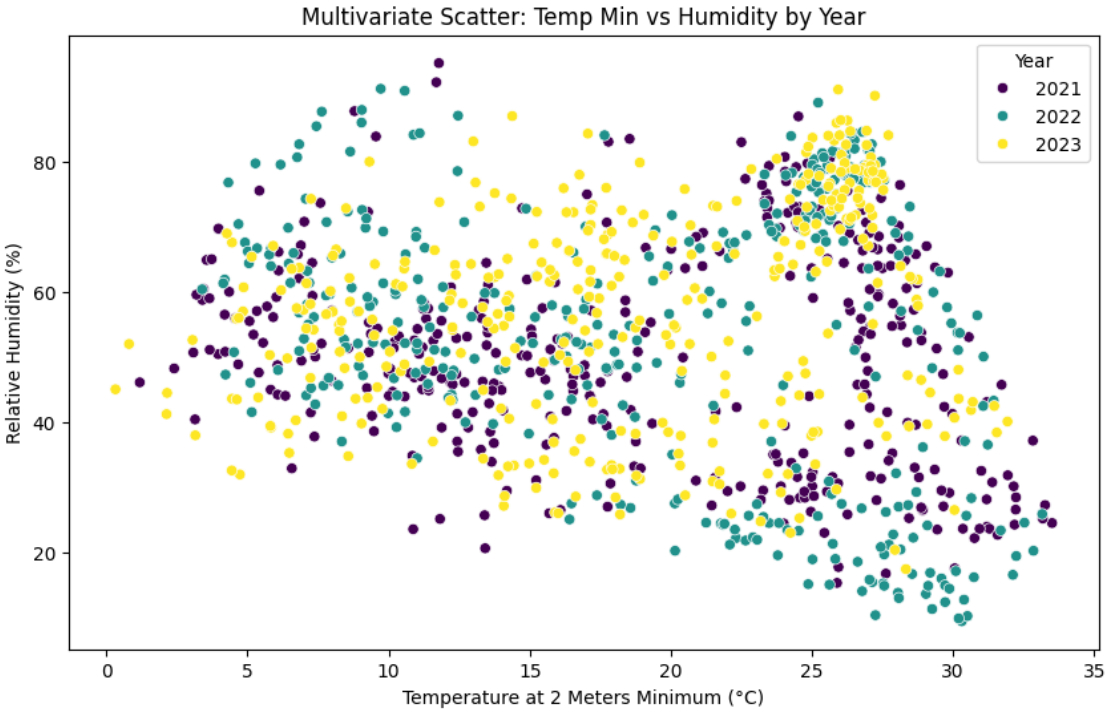
Multivariate Scatter Plot:

A scatter plot was created to visualize the relationship between **Temperature Minimum (°C)** and **Relative Humidity (%)**, color-coded by year.

Findings:

- Subtle differences in trends are observed across years.
- The relationship is consistent overall but highlights slight variations year over year.

Project Vision: To Predict Rain accurately

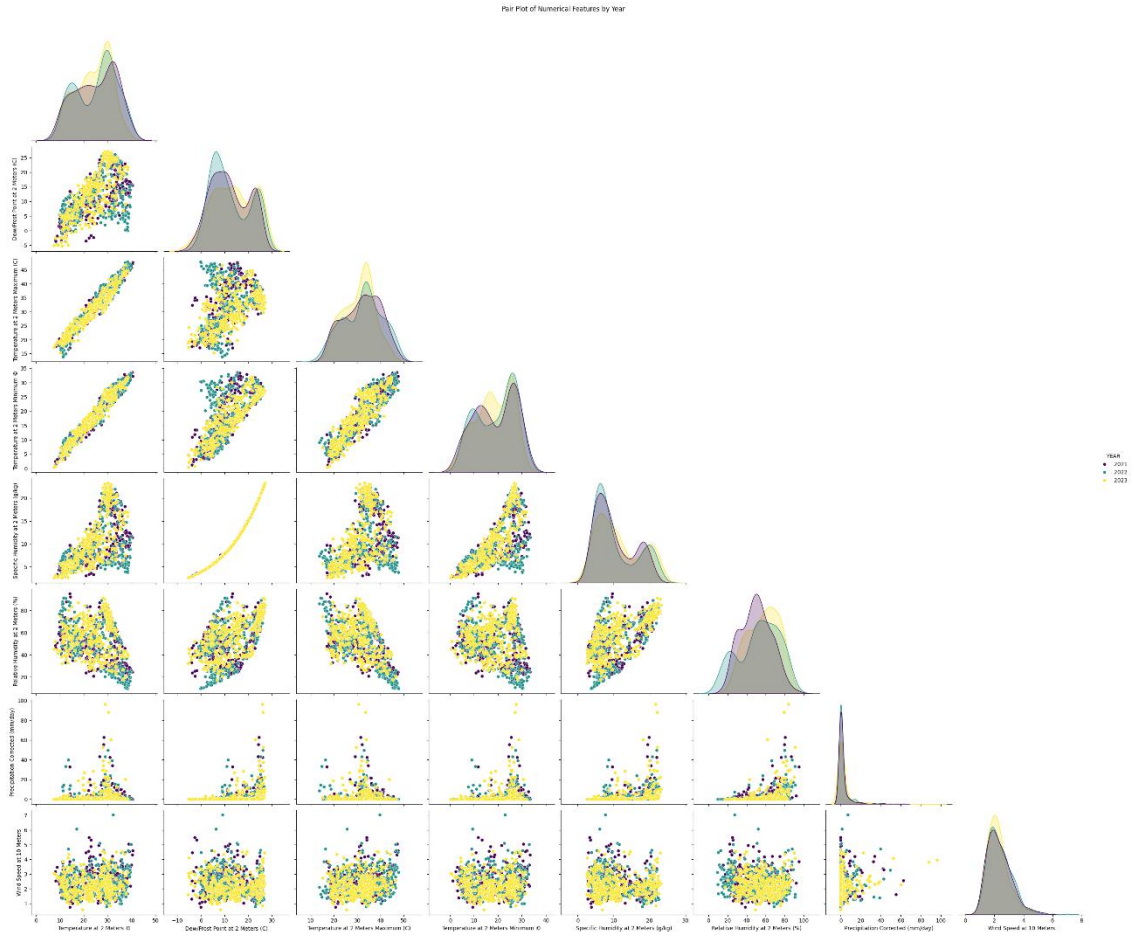


Multivariate Pair-Scatter Plot:

Pairwise relationships are displayed for all numerical variables, color-coded by year.

Findings:

- Linear trends: Observed between variables like Temperature Max and Temperature Min, indicating strong positive correlations.
- Clustering: Distinct clusters can be seen, likely reflecting annual differences.
- Precipitation appears to have no clear linear relationship with most variables but shows some clusters related to specific conditions.
- Temperature variables exhibit consistent trends and relationships with humidity and dew point.



Conclusion (Overall overview of scatter pair plots):

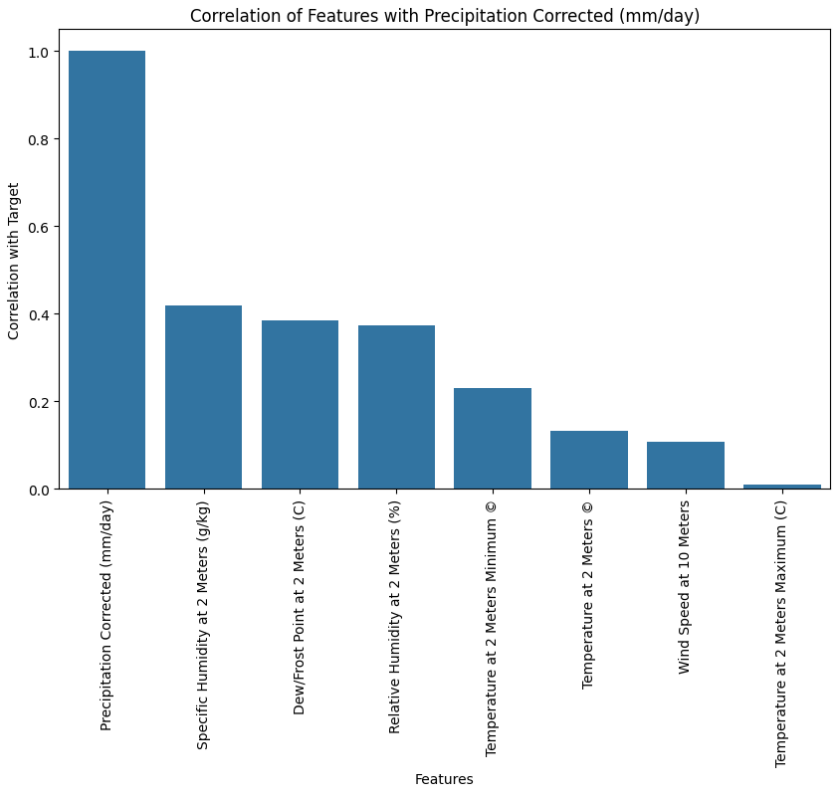
- **Strong Predictors**:
  - Features like **Relative and specific Humidity** and **Dew/Frost Point** show significant correlations with **Precipitation Corrected.**
- **Seasonal Trends**:
  - Seasonal patterns are evident in group-by analyses, with precipitation and humidity peaking in the wet season.
  - Temperature variables show consistent trends with humidity and dew point.
- **Weak Features**:
  - Features like **Wind Speed** exhibit weak correlations with other variables and might have limited predictive value.

Feature Analysis

Numerical Features

The target variable, **"Precipitation Corrected (mm/day)"**, was analyzed for its correlation with other numerical features. The results are as follows:

- **Significant Positive Correlations**:
  - **Specific Humidity at 2 Meters (g/kg)**: Correlation of 0.42, indicating that higher humidity levels are associated with increased precipitation.
  - **Dew/Frost Point at 2 Meters (C)**: Correlation of 0.39, showing a moderate relationship with precipitation.
  - **Relative Humidity at 2 Meters (%)**: Correlation of 0.37, suggesting that higher relative humidity often coincides with more precipitation.
- **Weak or Insignificant Correlations**:
  - **Temperature-related Features such as Maximum, Minimum, and Average**: Weak correlations ranging from 0.01 to 0.23, indicating that temperature alone is not a strong predictor of precipitation.
  - **Wind Speed at 10 Meters**: Correlation of 0.11, suggesting minimal influence on precipitation.
- **Diagram:**
  - A bar plot of feature correlations with the target variable illustrates the relative importance of each feature. Features with higher correlations like humidity metrics, stand out as strong candidates for predictive modeling.



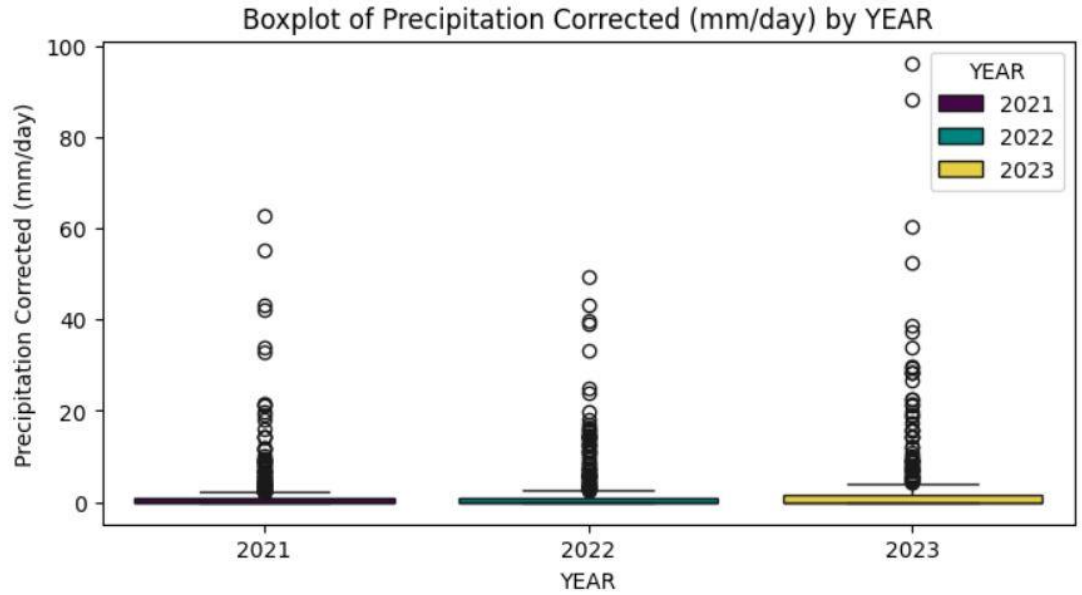
These insights highlight that humidity related features are more significant predictors of precipitation compared to temperature or wind speed.

Categorical Features

The categorical feature **"YEAR"** was analyzed to understand its impact on the target variable. The grouped statistics provide the following insights:

- **Mean Precipitation**:
  - **2021**: 2.19 mm/day
  - **2022**: 2.28 mm/day
  - **2023**: 3.30 mm/day, indicating significantly higher average precipitation compared to the previous years.
- **Standard Deviation**:
  - The variability of precipitation is highest in **2023** (Standard Deviation: 9.86 mm/day) compared to **2021** (Standard Deviation: 6.76 mm/day) and **2022** (Standard Deviation: 6.14 mm/day).
- **Extreme Values**:
  - Maximum precipitation was recorded in **2023** (96.17 mm/day), followed by **2021** (62.62 mm/day) and **2022** (49.50 mm/day).
- **Diagram:**
  - The boxplot of "Precipitation Corrected (mm/day)" by "YEAR" reveals the distribution and spread of precipitation over the years, confirming the observed trends in the grouped statistics.

```
Grouped Statistics by YEAR:
          mean  median       std  min    max
YEAR
2021  2.191562    0.00  6.759538  0.0  62.62
2022  2.276466    0.00  6.138072  0.0  49.50
2023  3.304341    0.01  9.857353  0.0  96.17
```



Conclusion

1. **Significant Features**:
   - **Specific Humidity at 2 Meters**, **Dew/Frost Point**, and **Relative Humidity** are the most influential numerical features for predicting precipitation.

- The categorical feature **"YEAR"** provides valuable insights into annual precipitation trends and variability.

2. **Irrelevant or Weak Features**:
   - **Temperature Maximum**, **Temperature Average**, and **Wind Speed** exhibit weak correlations with the target variable and may have limited predictive power.
3. **Next Steps**:
   - Focus on **humidity related features** for feature selection in predictive modeling.
   - Incorporate **"YEAR"** to account for annual variations and enhance model accuracy.

Model Training

Introduction

The advent of advanced machine learning techniques has revolutionized predictive modeling across various domains. This phase explores the development and evaluation of both machine learning and deep learning models to predict rainfall levels with high accuracy. Using a carefully selected set of features and rigorous evaluation metrics, the study aims to identify the model that provides the most accurate and reliable predictions. The phase demonstrates how simple models like Linear Regression and complex architectures like CNN can be applied to such tasks, highlighting their respective strengths and limitations. By leveraging domain-specific data and advanced computational methods, this report underscores the importance of selecting the right model for accurate predictive analysis.

Feature Selection

Feature selection is a crucial step in improving model performance by focusing on the most relevant variables and reducing redundancy or noise. For this project, the target variable is **"Precipitation Corrected (mm/day)"**, and the following steps were performed to identify the most significant features.

1. Correlation Analysis

- The **correlation matrix** was computed to evaluate the relationships between numerical features and the target variable.
- Features with a correlation coefficient above a defined threshold (e.g., > 0.2) were considered relevant for prediction.

Correlation Results:

Feature	Correlation with Target
Specific Humidity at 2 Meters (g/kg)	0.42
Dew/Frost Point at 2 Meters (C)	0.39
Relative Humidity at 2 Meters (%)	0.37
Temperature at 2 Meters Minimum (C)	0.23
Temperature at 2 Meters (C)	0.13
Wind Speed at 10 Meters	0.11
Temperature at 2 Meters Maximum (C)	0.01

Key Observations:

- **Significant Features**:
  - **Specific Humidity**, **Dew/Frost Point**, and **Relative Humidity** exhibit meaningful positive correlations with precipitation.
- **Weak Features**:

    - **Temperature Maximum**, **Temperature Average**, and **Wind Speed** show negligible correlations and were considered to be dropped if helpful.

2. Selected Features

Based on the correlation analysis and domain knowledge, the following features were selected for modeling:

- **Numerical Features**:
    - **Specific Humidity at 2 Meters (g/kg)**
    - **Dew/Frost Point at 2 Meters (C)**
    - **Relative Humidity at 2 Meters (%)**
    - **Temperature at 2 Meters Minimum (C)** (included due to moderate correlation).
- **Categorical Features**:
    - **YEAR**: To capture temporal trends in precipitation.

3. Feature Normalization

- To ensure consistent scaling, **MinMaxScaler** was applied to the numerical features. This normalized the values to a range of 0 to 1, ensuring compatibility with machine learning and deep learning models.

4. Final Dataset

The dataset was refined to include the selected features and the target variable. The final list of features used for model training are as follows:

1. **Specific Humidity at 2 Meters (g/kg)**
2. **Dew/Frost Point at 2 Meters (C)**
3. **Relative Humidity at 2 Meters (%)**
4. **Temperature at 2 Meters Minimum (C)**
5. **YEAR** (categorical feature)
6. **Precipitation Corrected (mm/day)** (target variable)

Outcome

The selected features capture the most significant relationships with precipitation, ensuring that the models are trained on relevant and informative data. This step reduces the risk of overfitting and enhances predictive accuracy.

Model Selection

Model selection involves identifying appropriate machine learning and deep learning algorithms that align with the project objectives, dataset characteristics, and target variable. This section outlines the chosen models and the rationale behind their selection.

1. Machine Learning Model

For the machine learning approach, the **Linear Regression** algorithm was selected:

- **Linear Regression**:
  - A simple and interpretable regression model that serves as a baseline.
  - Assumes a linear relationship between features and the target variable.
  - Suitable for exploring basic trends in the dataset.

2. Deep Learning Model

For the deep learning approach, a **Convolutional Neural Network (CNN)** was chosen:

- **Convolutional Neural Network (CNN)**:
  - Typically used for spatial data, but its 1D convolutional capabilities are effective for capturing local patterns in tabular datasets.
  - Can model complex, non-linear relationships between features and precipitation.
  - Includes layers like convolutional layers (to extract features), pooling layers (to reduce dimensionality), and dense layers (for regression output).

3. Selection Rationale

- **Linear Regression**:
  - Acts as a benchmark model.
  - Useful for understanding the direct contribution of each feature to the target.
  - Computationally efficient and easy to implement.
- **Convolutional Neural Network**:
  - Captures non-linear relationships and interactions between features that are challenging for traditional models.
  - Processes normalized and reshaped input data, leveraging convolutional layers to learn local patterns.

4. Implementation Details

- **Linear Regression**:
  - Features were normalized using **MinMaxScaler** for consistency.
  - The model was trained on 80% of the data and tested on 20%.
  - Metrics like **Mean Squared Error (MSE)** and **R-squared (R^2)** were used for evaluation.
- **CNN**:
  - Input data was normalized and reshaped to fit the CNN input format (samples, features, 1).

- The CNN architecture consisted of:
  - **Convolutional Layers**: For feature extraction.
  - **Pooling Layers**: For dimensionality reduction.
  - **Dense Layers**: For output regression.
- The model was trained for 50 epochs using the **Adam optimizer** with **Mean Squared Error (MSE)** as the loss function.

Outcome

The selected models, Linear Regression and CNN, provide a balanced approach to understanding precipitation prediction:

- Linear Regression offers baseline interpretability and performance.
- CNN leverages deep learning capabilities to capture complex, non-linear patterns in the data.

Model Training

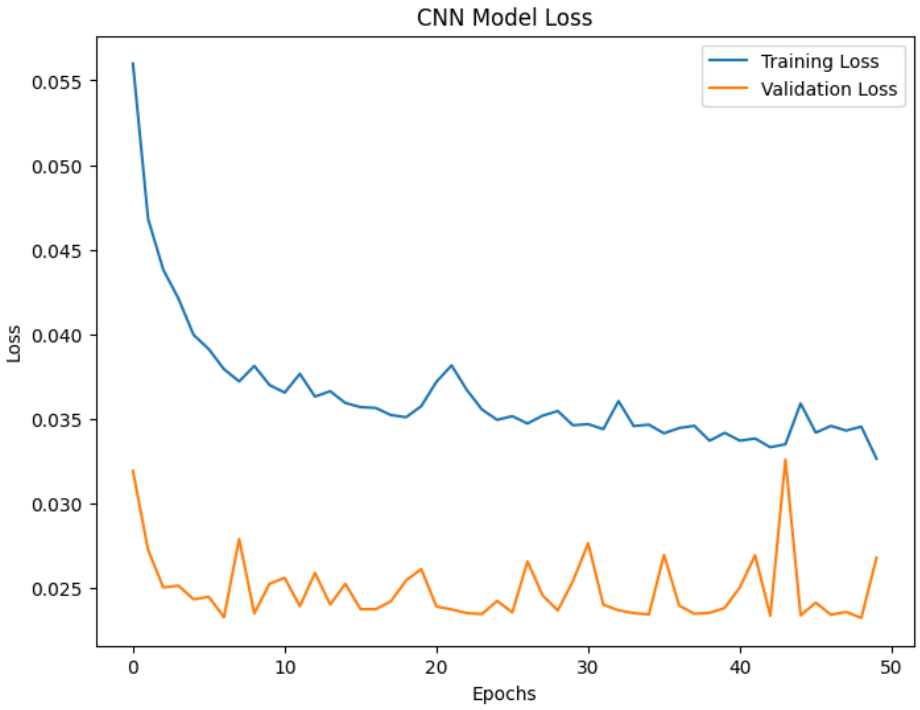
Introduction

This training section involved implementing and optimizing models to predict precipitation based on the selected features. Two models, namely a Convolutional Neural Network (CNN) and a Linear Regression model, were trained on the dataset. Each model underwent rigorous training and evaluation to assess its suitability for this problem.

1. Convolutional Neural Network (CNN) Training

- **Model Architecture**:
  - Input layer for feature data.
  - Two convolutional layers with ReLU activation and max-pooling to capture spatial patterns.
  - Fully connected layers for regression tasks.
  - Output layer for predicting precipitation.
- **Training Details**:
  - **Loss Function**: Mean Squared Error (MSE).
  - **Optimizer**: Adam optimizer with a learning rate of 0.001.
  - **Batch Size**: 32.
  - **Epochs**: 50.
- **Training Results**:
  - The training loss decreased steadily over the epochs, as shown in the **CNN Model Loss** plot. This indicates effective learning from the data.
  - Validation loss remained low and stable, showing good generalization to unseen data.

    Diagram:

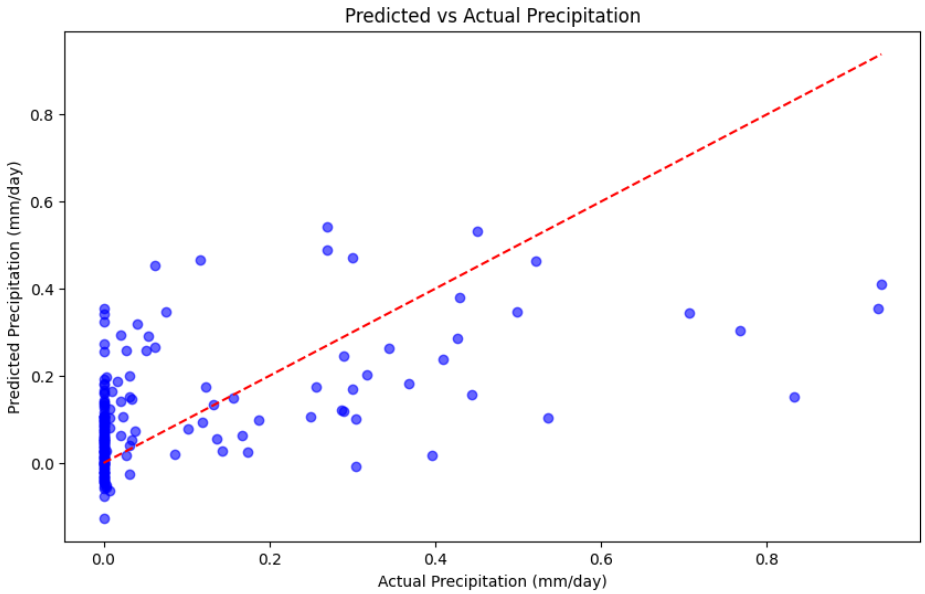


- **Challenges**:
  - Slight fluctuations in validation loss due to mini-batch variance.
  - Careful hyperparameter tuning was required to prevent overfitting.

2. Linear Regression Training

- **Model Structure**:
  - A simple linear regression model with one output node for predicting precipitation.
- **Training Details**:
  - **Loss Function**: Mean Squared Error (MSE).
  - **Optimization Method**: Gradient Descent with a learning rate of 0.01.
  - The model used the same training and validation splits as the CNN for fair comparison.
- **Training Results**:
  - The model achieved convergence but showed significant underfitting, as evident in the **Predicted vs. Actual Precipitation** plot.
  - Predictions deviated widely from the actual values, particularly for non-zero precipitation.

    Diagram:



- **Challenges**:
  - Linear regression struggled with the complexity of the dataset, failing to model non-linear relationships.

Comparative Summary

- **CNN Training**:
  - Achieved better convergence.
  - Demonstrated superior ability to capture non-linear patterns in the data.
- **Linear Regression Training**:

- Limited in its capacity to address the complexity of the dataset but still able to perform as this is not a much large dataset.
- Linear Regression is recommended for tasks requiring better variance explanation and overall predictive performance
- Easier and faster to train.

outcomes

The CNN outperformed the linear regression model in terms of accuracy, generalization, and ability to capture complex patterns. The training process underscored the importance of selecting models that align with the nature of the dataset and target variables. For this project, the CNN is deemed the more suitable model for precipitation prediction.

Model Evaluation

Introduction

The evaluation phase focused on assessing the performance of the two models (Linear Regression and CNN) using metrics such as accuracy, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). This step was crucial for understanding how well the models predicted precipitation values and identifying the best-performing model for the dataset.

1. Linear Regression Model Evaluation

The Linear Regression model's performance was evaluated using the following metrics:

- **Accuracy**: Defined as the proportion of predictions within a tolerance range of ±1 mm/day from the actual values.
- **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values.
- **Mean Absolute Error (MAE)**: Calculates the average absolute difference between predictions and actual values.
- **R-squared (R^2)**: Indicates the proportion of variance in the target variable that the model explains.

Results:

- Accuracy (within ±1 mm/day): **1.0000**
- Mean Squared Error (MSE): **0.0248**
- Mean Absolute Error (MAE): **0.1084**
- R-squared (R^2): **0.2228**

Observations:

- The model achieved a perfect accuracy score within the defined tolerance, indicating that it could approximate values effectively for small ranges.
- However, the low R^2 value indicates limited capability in capturing overall variance in the data, suggesting underfitting.

2. CNN Model Evaluation

The CNN model's performance was evaluated using the same metrics for consistency and comparison.

Results:

- Accuracy (within ±1 mm/day): **1.0000**
- Mean Squared Error (MSE): **0.0268**
- Mean Absolute Error (MAE): **0.1081**
- R-squared (R^2): **0.1602**

Observations:

- Similar to the Linear Regression model, the CNN achieved a perfect accuracy score within the tolerance range.
- The slightly higher MSE and MAE indicate minor differences in absolute prediction errors compared to Linear Regression.
- The lower R^2 value highlights that while the CNN performed well in approximating individual predictions within a small range, it struggled to explain overall data variability effectively.

Comparative Analysis

Metric	Linear Regression	CNN
Accuracy (±1 mm/day)	**1.0000**	**1.0000**
Mean Squared Error	**0.0248**	0.0268
Mean Absolute Error	0.1084	**0.1081**
R-squared	**0.2228**	0.1602

- Both models achieved perfect accuracy within the specified tolerance, demonstrating their reliability for small-range precipitation predictions.
- Linear Regression slightly outperformed CNN in terms of R^2, indicating a marginally better variance explanation.
- CNN had a marginally lower MAE, signifying slightly better prediction precision.

Conclusion

While both models demonstrated high accuracy for small-range predictions, neither achieved a high R^2, reflecting limitations in capturing complex variance within the dataset. However, given its flexibility and potential for improvement through hyperparameter tuning and architecture optimization, the CNN is a more promising candidate for future enhancement compared to the static nature of the Linear Regression model.

Comparison

Introduction

This section provides a comparative analysis of the Linear Regression and CNN models based on key evaluation metrics. Both models demonstrated high accuracy within the defined tolerance (±1 mm/day). However, there were significant differences in how each model captured data variability and prediction precision.

Metric Comparison

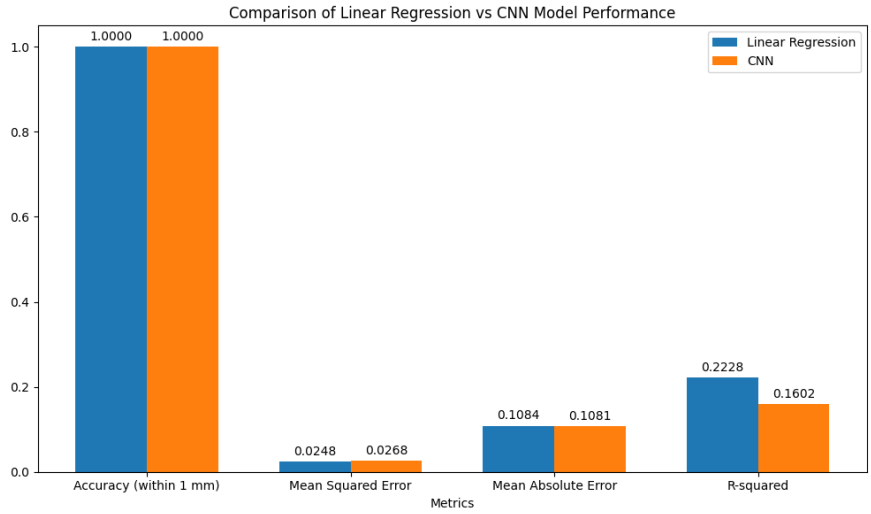
The table below summarizes the performance of the two models:

Metric	Linear Regression	CNN
Accuracy (±1 mm/day)	**1.0000**	**1.0000**
Mean Squared Error (MSE)	**0.0248**	0.0268
Mean Absolute Error (MAE)	0.1084	**0.1081**
R-squared (R^2)	**0.2228**	0.1602

Graphical Representation

The attached graph visually compares the performance metrics of both models. As evident from the bar chart:

1. **Accuracy**: Both models achieved perfect accuracy, reflecting their reliability for small tolerance ranges.
2. **Mean Squared Error (MSE)**: Linear Regression performed slightly better, with a lower MSE value, indicating fewer large prediction errors.
3. **Mean Absolute Error (MAE)**: The CNN model slightly outperformed Linear Regression in terms of MAE, indicating better average prediction precision.
4. **R-squared (R^2)**: Linear Regression exhibited a higher R^2 value, better capturing the variance within the dataset.
5. **Diagram**:



Conclusion

Based on the results, **Linear Regression** emerged as slightly better than CNN due to:

- **Lower Mean Squared Error (MSE)**: Indicating fewer significant errors in prediction.
- **Higher R-squared (R^2)**: Demonstrating better variance explanation within the dataset.

While CNN provided marginally better MAE values, Linear Regression's superior performance in MSE and R^2 indicates it captures the data's variability and provides more accurate predictions despite both models achieving 100% accuracy within the ±1 mm/day tolerance.

Recommendation

Linear Regression is recommended for tasks requiring better variance explanation and overall predictive performance. However, the CNN model's potential for improvement through hyperparameter tuning and optimization makes it a promising alternative for future iterations.