Analysis of a Weather Dataset

Exploratory Data Analysis and Visualization

Date: 4th August 2024

Name: Ishita Singh

The objective of this project was to analyze the weather dataset to understand trends, handle data inconsistencies, and build predictive models for weather parameters. I took the dataset online from a Github repository.

The dataset contains real-time measurements of outdoor temperature, relative humidity, diffuse solar radiation, and direct solar radiation for city Islamabad. The data was collected using a weather station installed in the city and includes 6-hour, 12-hour, and 24-hour predictions for each measurement.

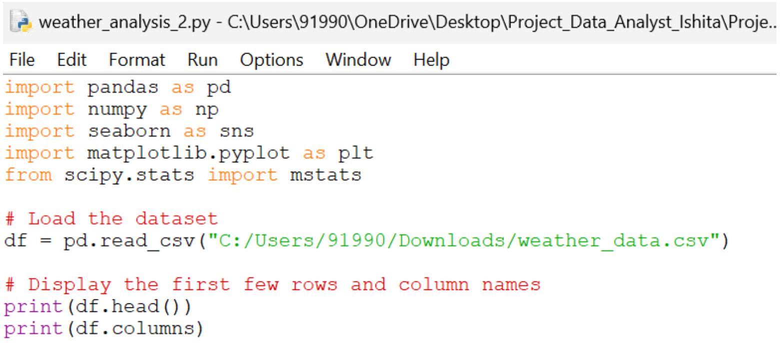
The data has 8761 rows of data, with several values marked 0 (missing values) that had to be dealt with. It has total 16 columns namely:

1. Outdoor Drybulb Temperature [C]
    - 6h Prediction Outdoor Drybulb Temperature [C]
    - 12h Prediction Outdoor Drybulb Temperature [C]
    - 24h Prediction Outdoor Drybulb Temperature [C]
2. Outdoor Relative Humidity [%]
    - 6h Prediction Outdoor Relative Humidity [%]
    - 12h Prediction Outdoor Relative Humidity [%]
    - 24h Prediction Outdoor Relative Humidity [%]
3. Diffuse Solar Radiation [W/m2]
    - 6h Prediction Diffuse Solar Radiation [W/m2]
    - 12h Prediction Diffuse Solar Radiation [W/m2]
    - 24h Prediction Diffuse Solar Radiation [W/m2]
4. Direct Solar Radiation [W/m2]
    - 6h Prediction Direct Solar Radiation [W/m2]
    - 12h Prediction Direct Solar Radiation [W/m2]
    - 24h Prediction Direct Solar Radiation [W/m2]

Following is the source of the data: https://figshare.com/articles/dataset/weather_data_csv/22579948

Working on Python:

I downloaded Python and installed the required libraries for project 1 and had to download a few more libraries like scipy and numpy. I am not experienced with Python and thus, used ChatGPT to help learn. I went onto YouTube as well to work on certain steps. Using the following code, I started the python script:

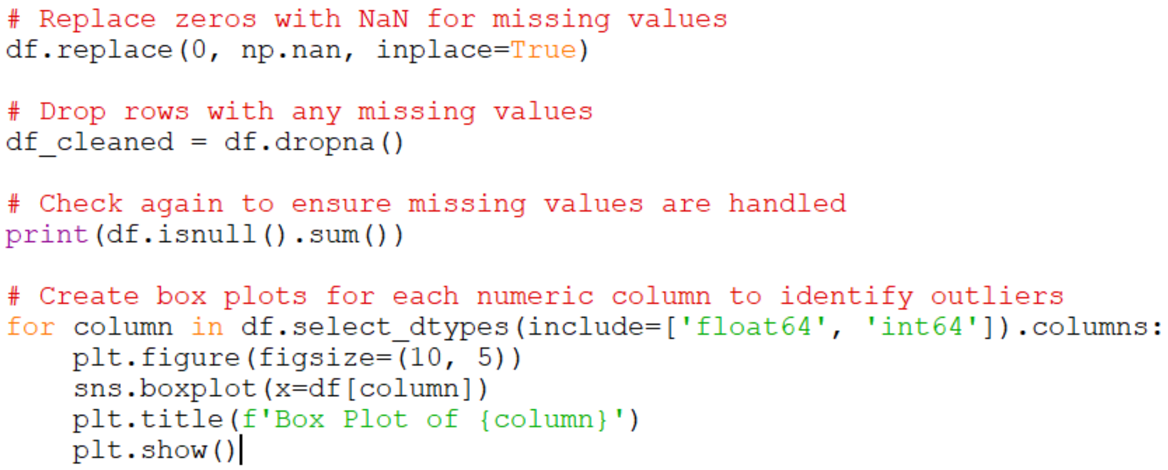

weather_analysis_2.py - C:\Users\91990\OneDrive\Desktop\Project_Data_Analyst_Ishita\Proje..
File
Edit
Format
Run
Options
Window
Help

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import mstats

# Load the dataset
df = pd.read_csv("C:/Users/91990/Downloads/weather_data.csv")

# Display the first few rows and column names
print(df.head())
print(df.columns)
```

Loading the dataset using the regular read_csv command. Then I used the print command to display the first few rows and columns to know if the data had been loaded properly.

Now, when I opened the dataset in excel, I noticed there were missing values but they were mentioned as 0 and when I first ran the find missing values command, it showed no missing values. So, I moved onto checking and handling outliers, for which I used box-plot visualization. But the plots were unusually odd and had huge number of outliers. There were some plots where the mean was as close as the 25th percentile or the 75th one and all other values were outliers.

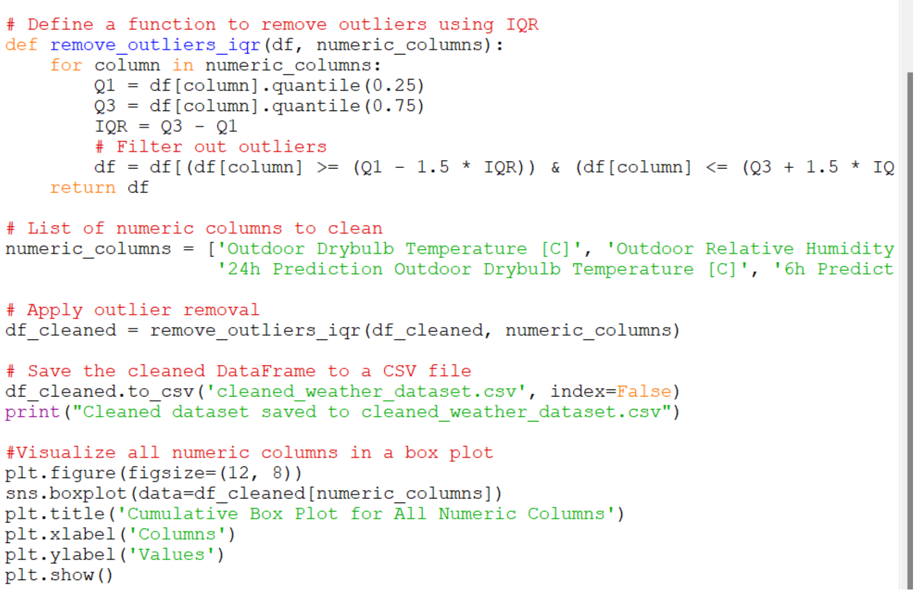
Then I was facing difficulty handling those outliers. After that I looked at the excel dataset, saw that the missing values were labelled 0 instead of a blank cell or NA. So, I run the following command to change the 0's to NA and then check for missing values:



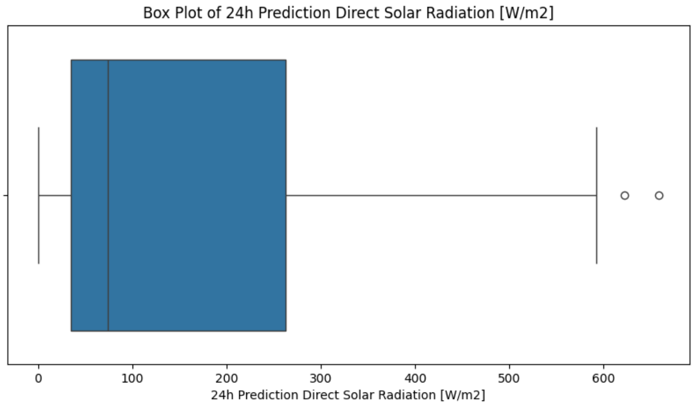
```
# Replace zeros with NaN for missing values
df.replace(0, np.nan, inplace=True)

# Drop rows with any missing values
df_cleaned = df.dropna()

# Check again to ensure missing values are handled
print(df.isnull().sum())

# Create box plots for each numeric column to identify outliers
for column in df.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 5))
    sns.boxplot(x=df[column])
    plt.title(f'Box Plot of {column}')
    plt.show()
```

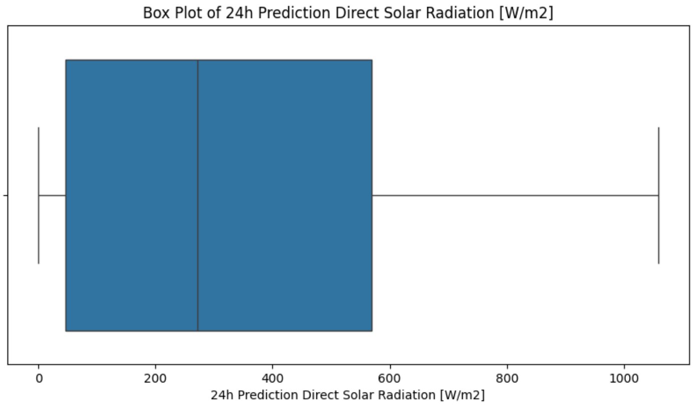
After getting the results for the missing values, I dropped those rows. After this, I ran the command to create box plot for each column to now visualize it better and check for outliers.

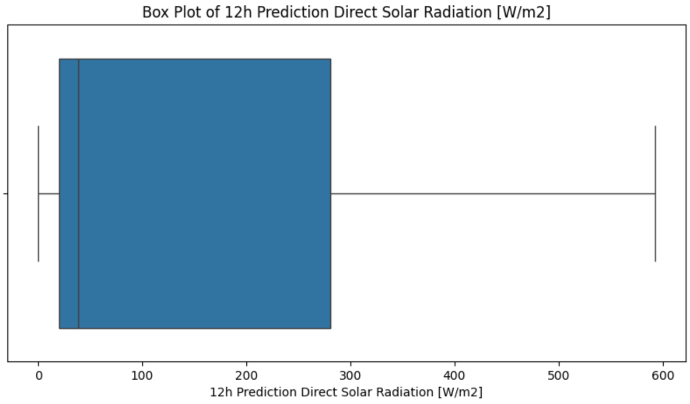
On a separate note, I noted all the columns that had outliers and run the following command to either filter them using the IQR (Inter-Quartile Range) Method, where we set a bar along the inner whiskers and remove all values that fall beyond it called the mild outliers. After which I saved the cleaned data to cleaned_weather_dataset.csv, and visualized the clean plots attached below, compared with the unclean on the left. I left some of the outliers because they were contributing to the dataset favourably.

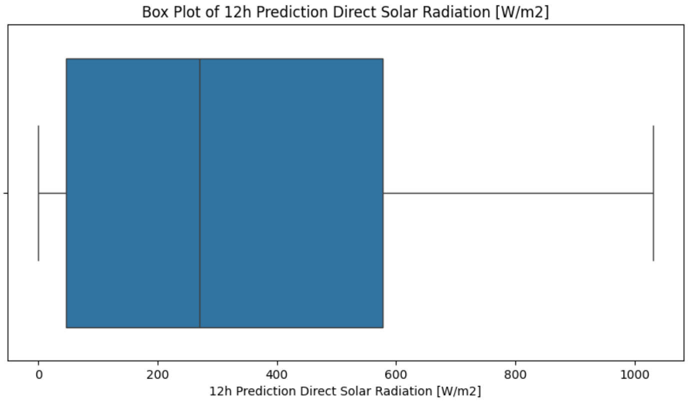


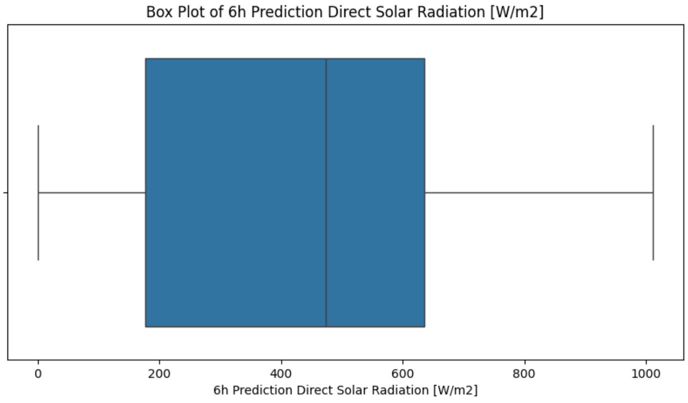
```
# Define a function to remove outliers using IQR
def remove_outliers_iqr(df, numeric_columns):
    for column in numeric_columns:
        Q1 = df[column].quantile(0.25)
        Q3 = df[column].quantile(0.75)
        IQR = Q3 - Q1
        # Filter out outliers
        df = df[(df[column] >= (Q1 - 1.5 * IQR)) & (df[column] <= (Q3 + 1.5 * IQ
    return df

# List of numeric columns to clean
numeric_columns = ['Outdoor Drybulb Temperature [C]', 'Outdoor Relative Humidity
                   '24h Prediction Outdoor Drybulb Temperature [C]', '6h Predict

# Apply outlier removal
df_cleaned = remove_outliers_iqr(df_cleaned, numeric_columns)

# Save the cleaned DataFrame to a CSV file
df_cleaned.to_csv('cleaned_weather_dataset.csv', index=False)
print("Cleaned dataset saved to cleaned_weather_dataset.csv")

#Visualize all numeric columns in a box plot
plt.figure(figsize=(12, 8))
sns.boxplot(data=df_cleaned[numeric_columns])
plt.title('Cumulative Box Plot for All Numeric Columns')
plt.xlabel('Columns')
plt.ylabel('Values')
plt.show()
```

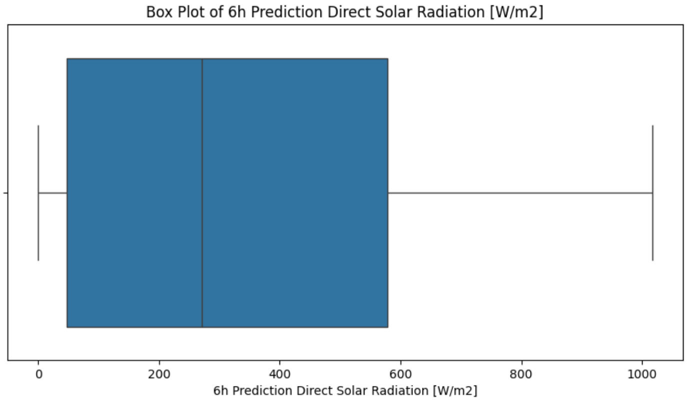

Box Plot of 24h Prediction Direct Solar Radiation [W/m2]
0
100
200
300
400
500
600
24h Prediction Direct Solar Radiation [W/m2]

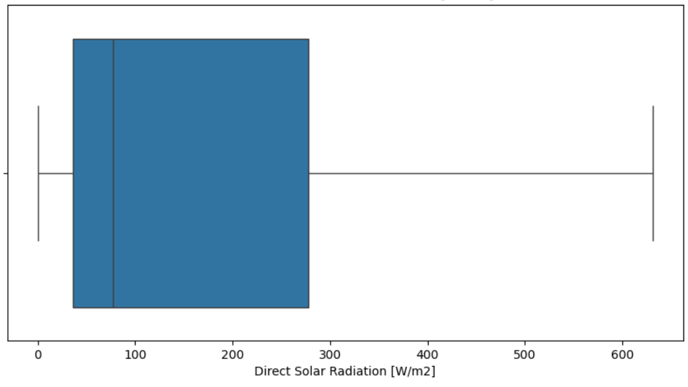

Box Plot of 24h Prediction Direct Solar Radiation [W/m2]
0
200
400
600
800
1000
24h Prediction Direct Solar Radiation [W/m2]

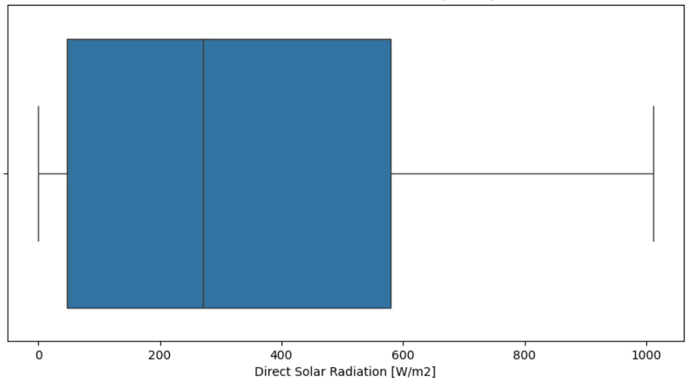

Box Plot of 12h Prediction Direct Solar Radiation [W/m2]
0
100
200
300
400
500
600
12h Prediction Direct Solar Radiation [W/m2]

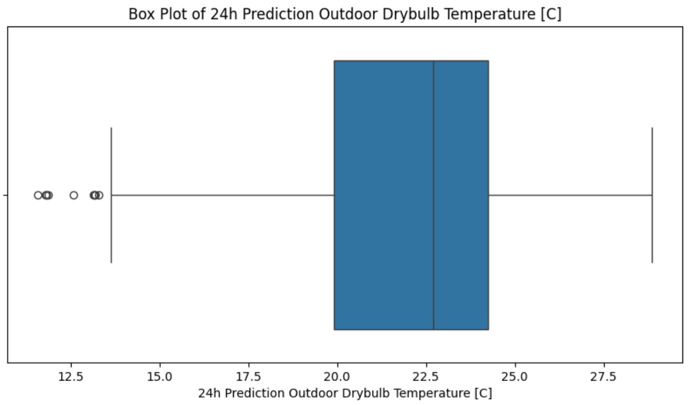

Box Plot of 12h Prediction Direct Solar Radiation [W/m2]
0
200
400
600
800
1000
12h Prediction Direct Solar Radiation [W/m2]

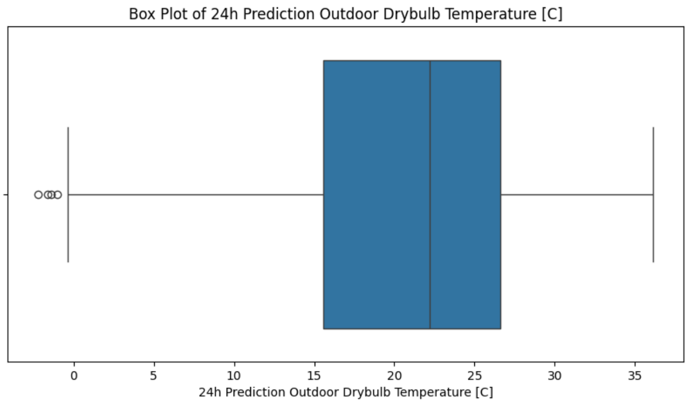

Box Plot of 6h Prediction Direct Solar Radiation [W/m2]
0
200
400
600
800
1000
6h Prediction Direct Solar Radiation [W/m2]

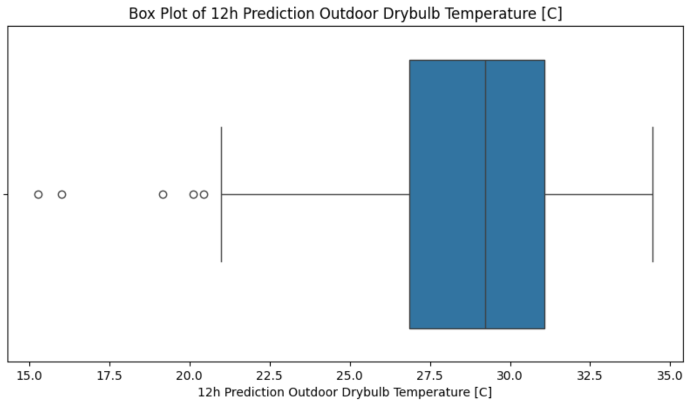

Box Plot of 6h Prediction Direct Solar Radiation [W/m2]
0
200
400
600
800
1000
6h Prediction Direct Solar Radiation [W/m2]

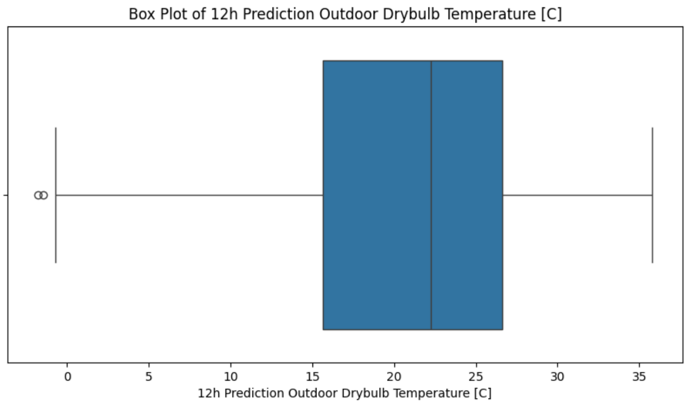

Box Plot of Direct Solar Radiation [W/m2]
0
100
200
300
400
500
600
Direct Solar Radiation [W/m2]

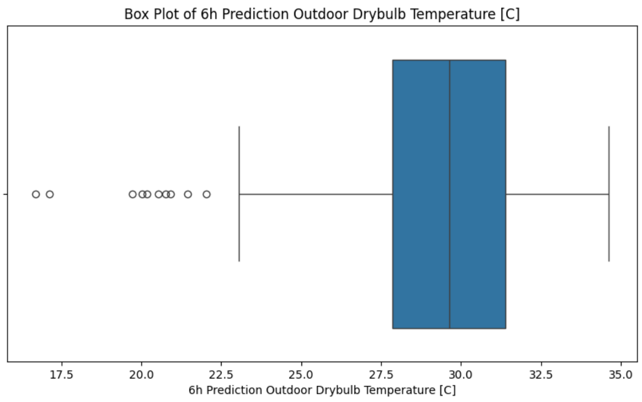

Box Plot of Direct Solar Radiation [W/m2]
0
200
400
600
800
1000
Direct Solar Radiation [W/m2]

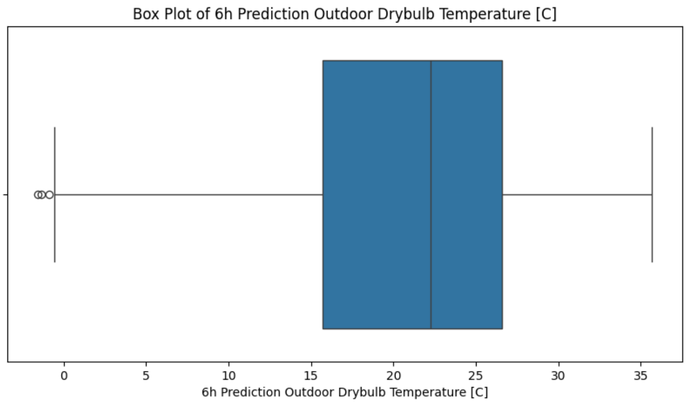

Box Plot of 24h Prediction Outdoor Drybulb Temperature [C]
12.5
15.0
17.5
20.0
22.5
25.0
27.5
24h Prediction Outdoor Drybulb Temperature [C]

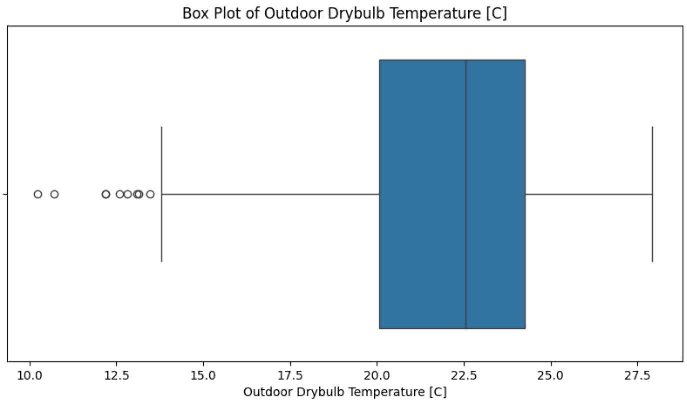

Box Plot of 24h Prediction Outdoor Drybulb Temperature [C]
0
5
10
15
20
25
30
35
24h Prediction Outdoor Drybulb Temperature [C]

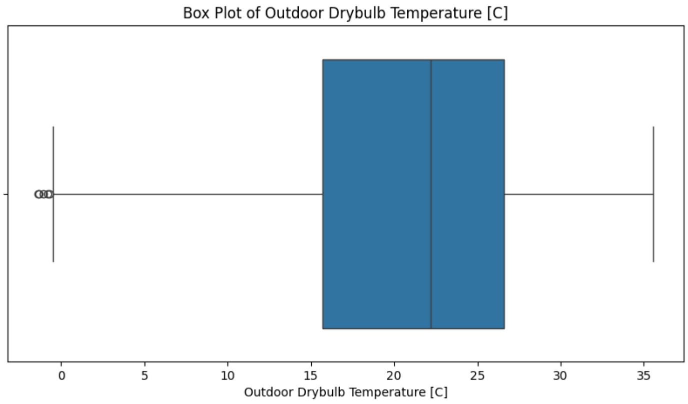

Box Plot of 12h Prediction Outdoor Drybulb Temperature [C]
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
12h Prediction Outdoor Drybulb Temperature [C]

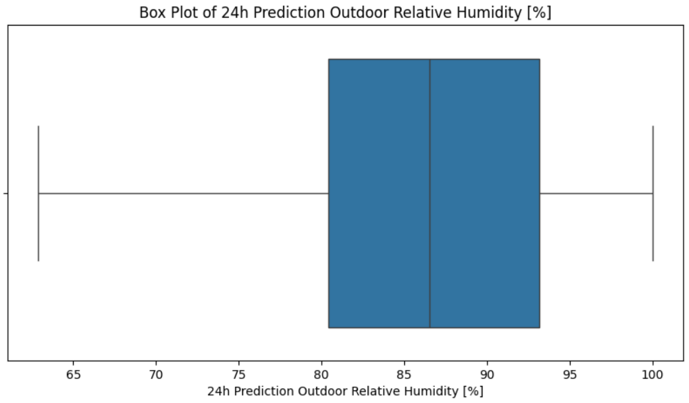

Box Plot of 12h Prediction Outdoor Drybulb Temperature [C]
0
5
10
15
20
25
30
35
12h Prediction Outdoor Drybulb Temperature [C]

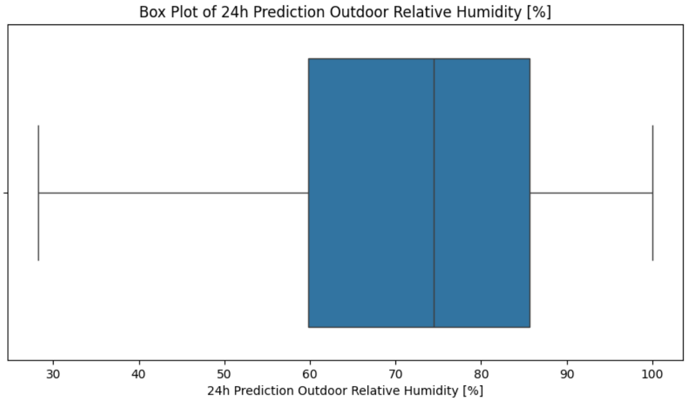

Box Plot of 6h Prediction Outdoor Drybulb Temperature [C]
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
6h Prediction Outdoor Drybulb Temperature [C]

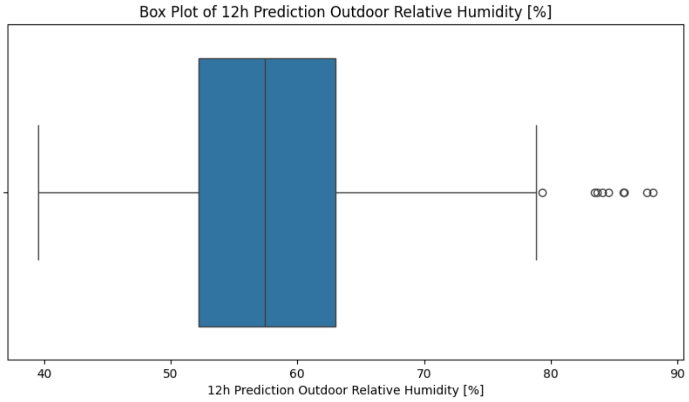

Box Plot of 6h Prediction Outdoor Drybulb Temperature [C]
0
5
10
15
20
25
30
35
6h Prediction Outdoor Drybulb Temperature [C]

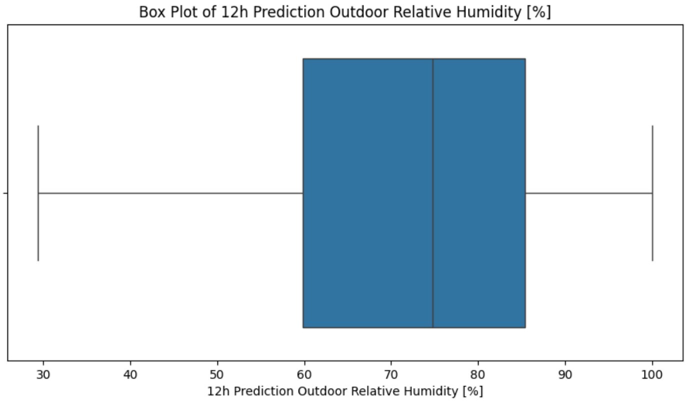

Box Plot of Outdoor Drybulb Temperature [C]
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
Outdoor Drybulb Temperature [C]

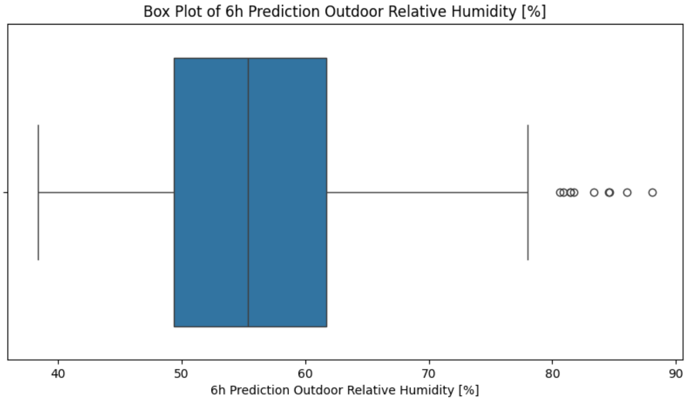

Box Plot of Outdoor Drybulb Temperature [C]
0
5
10
15
20
25
30
35
Outdoor Drybulb Temperature [C]

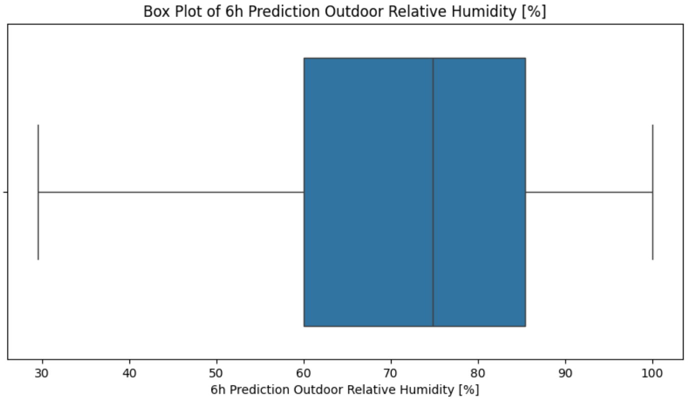

Box Plot of 24h Prediction Outdoor Relative Humidity [%]
65
70
75
80
85
90
95
100
24h Prediction Outdoor Relative Humidity [%]

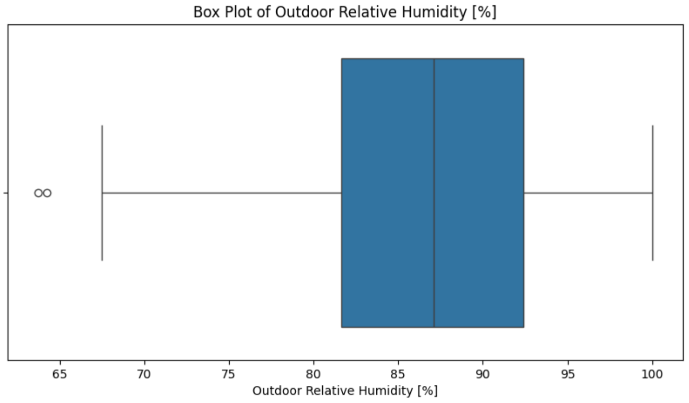

Box Plot of 24h Prediction Outdoor Relative Humidity [%]
30
40
50
60
70
80
90
100
24h Prediction Outdoor Relative Humidity [%]

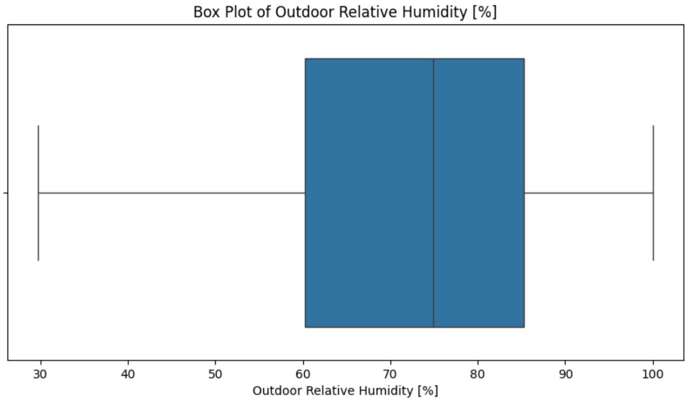

Box Plot of 12h Prediction Outdoor Relative Humidity [%]
40
50
60
70
80
90
12h Prediction Outdoor Relative Humidity [%]

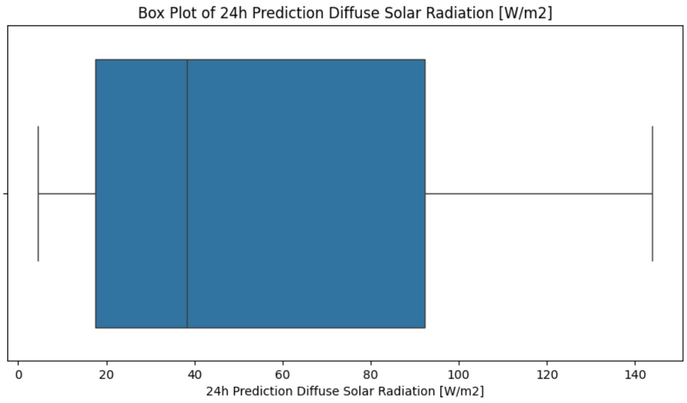

Box Plot of 12h Prediction Outdoor Relative Humidity [%]
30
40
50
60
70
80
90
100
12h Prediction Outdoor Relative Humidity [%]

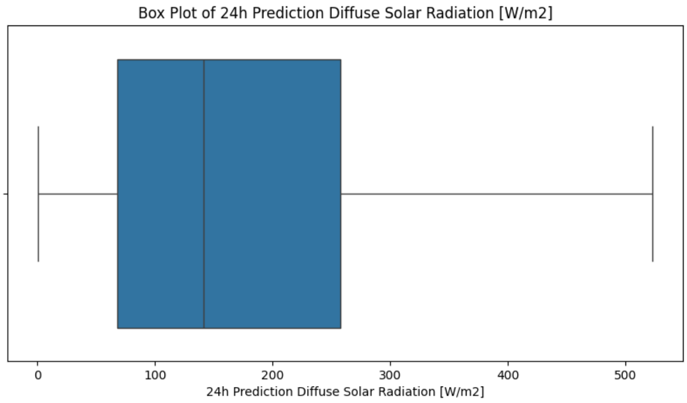

Box Plot of 6h Prediction Outdoor Relative Humidity [%]
40
50
60
70
80
90
6h Prediction Outdoor Relative Humidity [%]

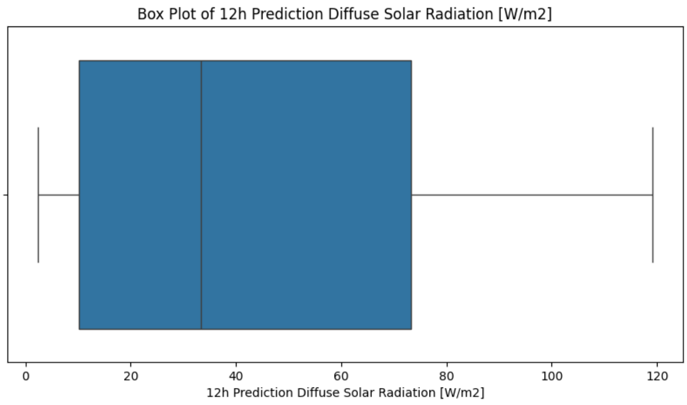

Box Plot of 6h Prediction Outdoor Relative Humidity [%]
30
40
50
60
70
80
90
100
6h Prediction Outdoor Relative Humidity [%]

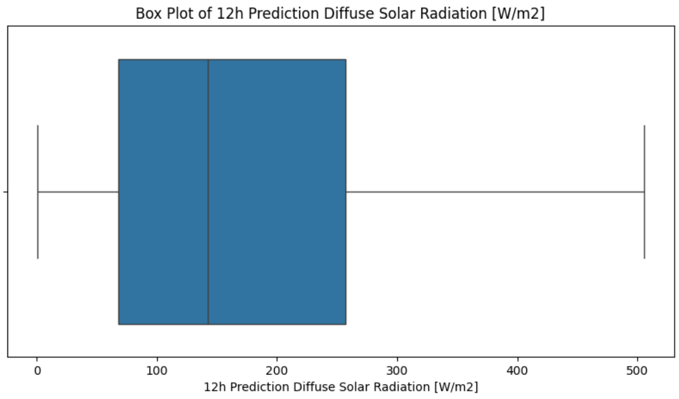

Box Plot of Outdoor Relative Humidity [%]
65
70
75
80
85
90
95
100
Outdoor Relative Humidity [%]

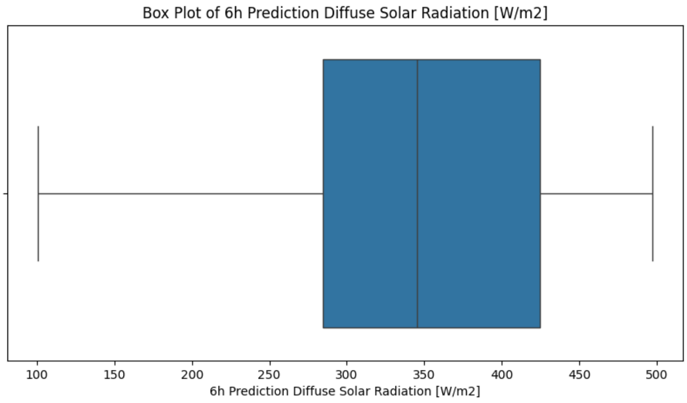

Box Plot of Outdoor Relative Humidity [%]
30
40
50
60
70
80
90
100
Outdoor Relative Humidity [%]

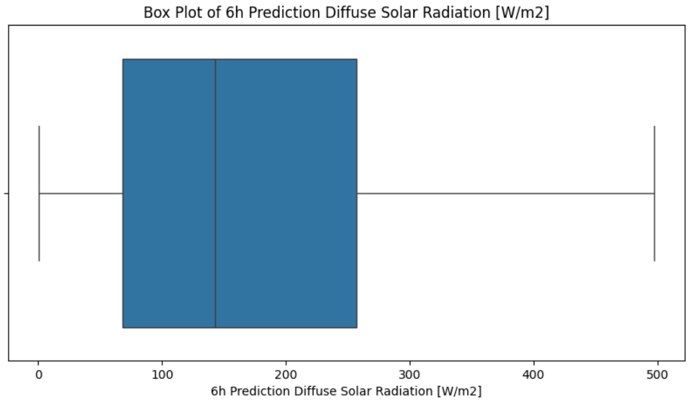

Box Plot of 24h Prediction Diffuse Solar Radiation [W/m2]
0
20
40
60
80
100
120
140
24h Prediction Diffuse Solar Radiation [W/m2]

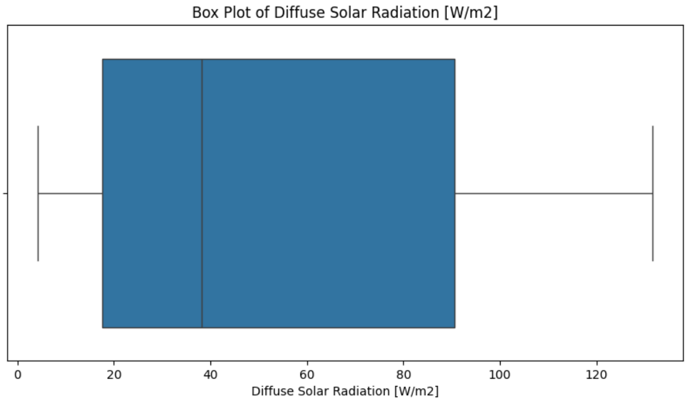

Box Plot of 24h Prediction Diffuse Solar Radiation [W/m2]
0
100
200
300
400
500
24h Prediction Diffuse Solar Radiation [W/m2]

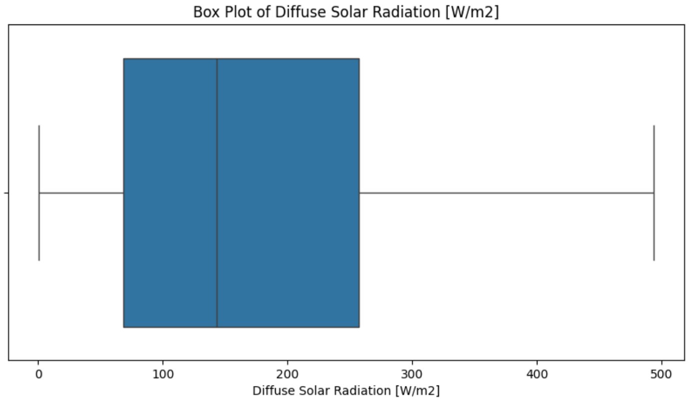

Box Plot of 12h Prediction Diffuse Solar Radiation [W/m2]
0
20
40
60
80
100
120
12h Prediction Diffuse Solar Radiation [W/m2]


Box Plot of 12h Prediction Diffuse Solar Radiation [W/m2]
0
100
200
300
400
500
12h Prediction Diffuse Solar Radiation [W/m2]


Box Plot of 6h Prediction Diffuse Solar Radiation [W/m2]
100
150
200
250
300
350
400
450
500
6h Prediction Diffuse Solar Radiation [W/m2]

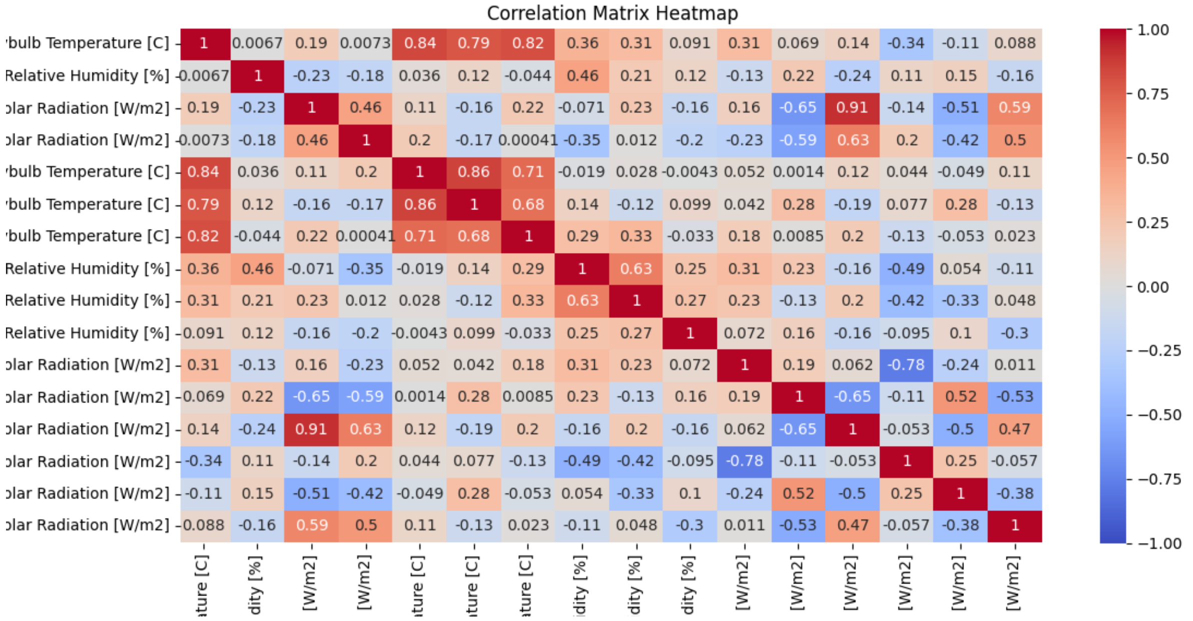

Box Plot of 6h Prediction Diffuse Solar Radiation [W/m2]
0
100
200
300
400
500
6h Prediction Diffuse Solar Radiation [W/m2]


Box Plot of Diffuse Solar Radiation [W/m2]
0
20
40
60
80
100
120
Diffuse Solar Radiation [W/m2]


Box Plot of Diffuse Solar Radiation [W/m2]
0
100
200
300
400
500
Diffuse Solar Radiation [W/m2]

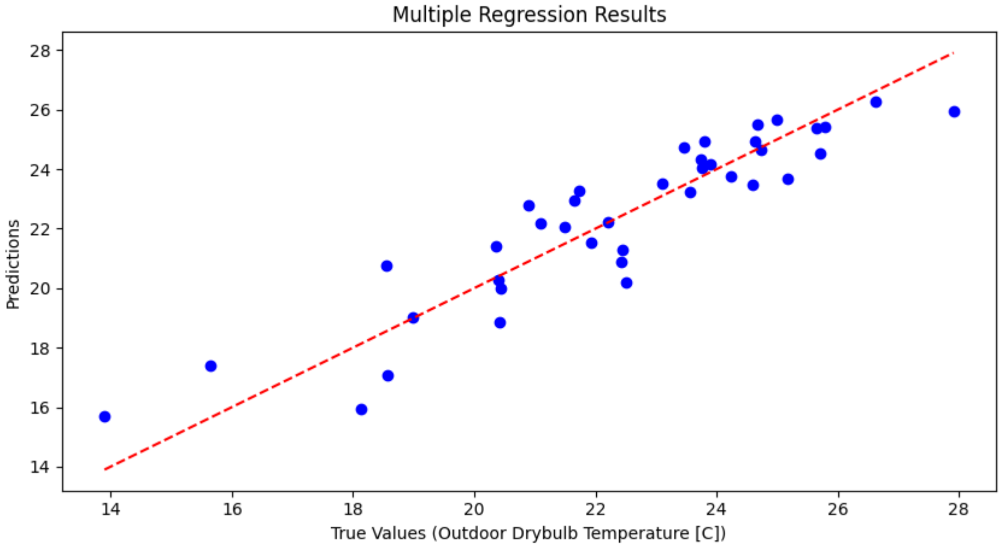
Correlation and Regression

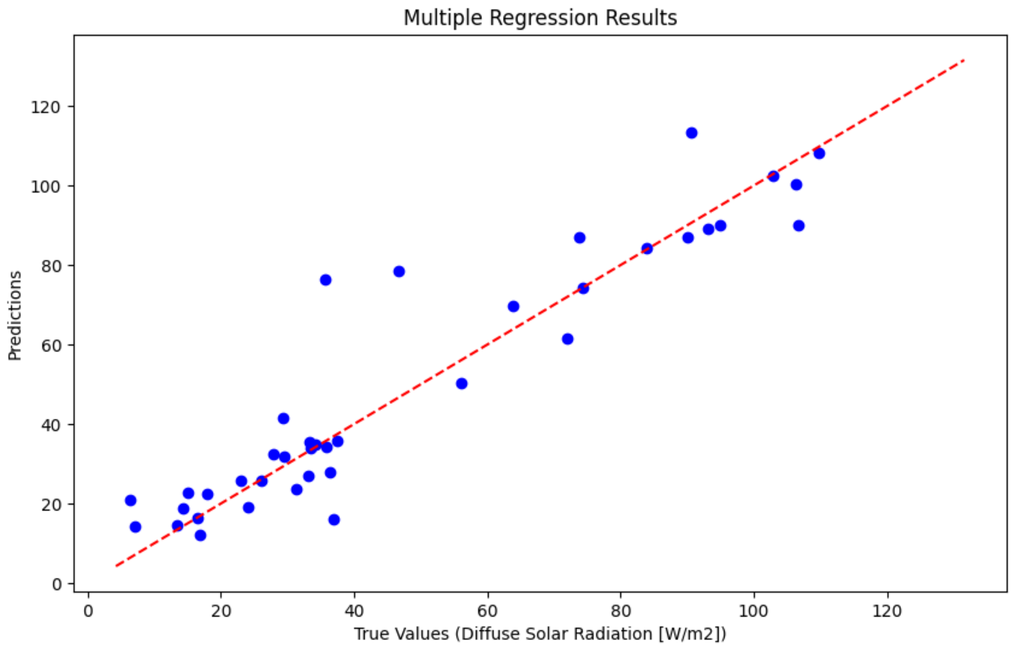
In a separate Python file, I completed the correlation matrix and regression analysis. For this file, I loaded the clean dataset. I also had to download more libraries like sklearn. The correlation matrix calculation and visualization were done in a similar manner as the iris dataset. The correlation heatmap is attached below, whose interpretation will be discussed later.

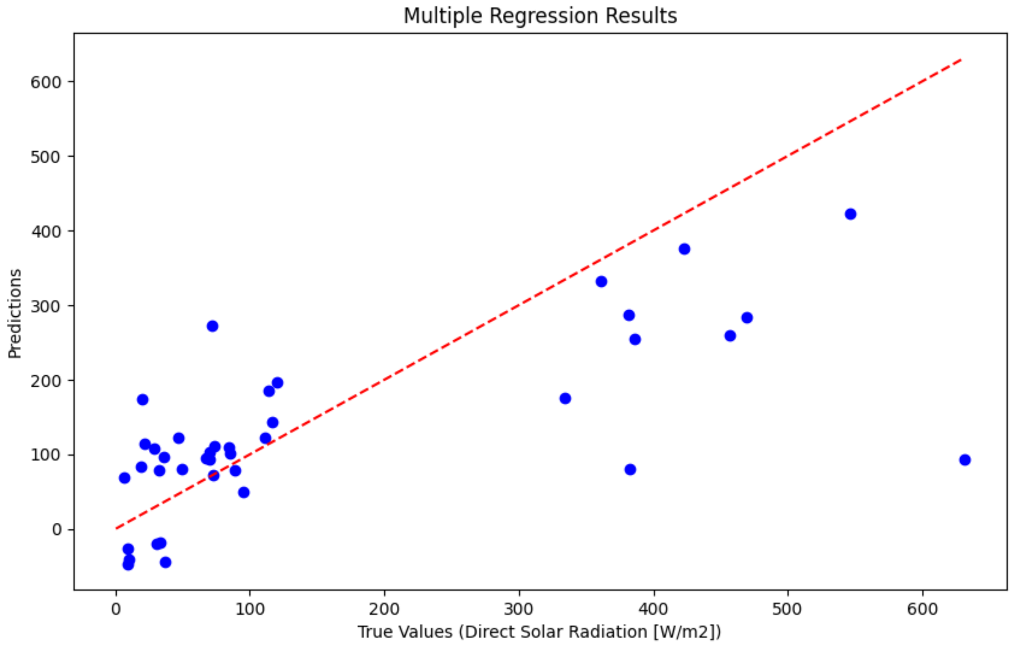


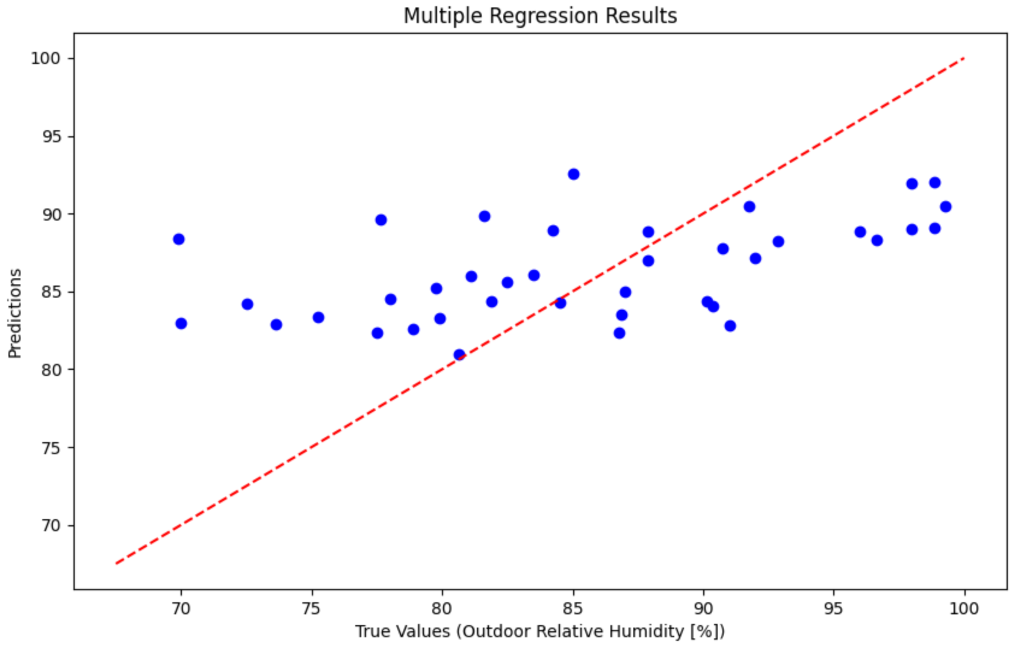
	ature [C]	dity [%]	[W/m2]	[W/m2]	ture [C]	ture [C]	ture [C]	idity [%]	dity [%]	dity [%]	[W/m2]	[W/m2]	[W/m2]	[W/m2]	[W/m2]	[W/m2]
'bulb Temperature [C]	1	0.0067	0.19	0.0073	0.84	0.79	0.82	0.36	0.31	0.091	0.31	0.069	0.14	-0.34	-0.11	0.088
Relative Humidity [%]	0.0067	1	-0.23	-0.18	0.036	0.12	-0.044	0.46	0.21	0.12	-0.13	0.22	-0.24	0.11	0.15	-0.16
olar Radiation [W/m2]	0.19	-0.23	1	0.46	0.11	-0.16	0.22	-0.071	0.23	-0.16	0.16	-0.65	0.91	-0.14	-0.51	0.59
olar Radiation [W/m2]	0.0073	-0.18	0.46	1	0.2	-0.17	0.00041	-0.35	0.012	-0.2	-0.23	-0.59	0.63	0.2	-0.42	0.5
'bulb Temperature [C]	0.84	0.036	0.11	0.2	1	0.86	0.71	-0.019	0.028	-0.0043	0.052	0.0014	0.12	0.044	-0.049	0.11
'bulb Temperature [C]	0.79	0.12	-0.16	-0.17	0.86	1	0.68	0.14	-0.12	0.099	0.042	0.28	-0.19	0.077	0.28	-0.13
'bulb Temperature [C]	0.82	-0.044	0.22	0.00041	0.71	0.68	1	0.29	0.33	-0.033	0.18	0.0085	0.2	-0.13	-0.053	0.023
Relative Humidity [%]	0.36	0.46	-0.071	-0.35	-0.019	0.14	0.29	1	0.63	0.25	0.31	0.23	-0.16	-0.49	0.054	-0.11
Relative Humidity [%]	0.31	0.21	0.23	0.012	0.028	-0.12	0.33	0.63	1	0.27	0.23	-0.13	0.2	-0.42	-0.33	0.048
Relative Humidity [%]	0.091	0.12	-0.16	-0.2	-0.0043	0.099	-0.033	0.25	0.27	1	0.072	0.16	-0.16	-0.095	0.1	-0.3
olar Radiation [W/m2]	0.31	-0.13	0.16	-0.23	0.052	0.042	0.18	0.31	0.23	0.072	1	0.19	0.062	-0.78	-0.24	0.011
olar Radiation [W/m2]	0.069	0.22	-0.65	-0.59	0.0014	0.28	0.0085	0.23	-0.13	0.16	0.19	1	-0.65	-0.11	0.52	-0.53
olar Radiation [W/m2]	0.14	-0.24	0.91	0.63	0.12	-0.19	0.2	-0.16	0.2	-0.16	0.062	-0.65	1	-0.053	-0.5	0.47
olar Radiation [W/m2]	-0.34	0.11	-0.14	0.2	0.044	0.077	-0.13	-0.49	-0.42	-0.095	-0.78	-0.11	-0.053	1	0.25	-0.057
olar Radiation [W/m2]	-0.11	0.15	-0.51	-0.42	-0.049	0.28	-0.053	0.054	-0.33	0.1	-0.24	0.52	-0.5	0.25	1	-0.38
olar Radiation [W/m2]	0.088	-0.16	0.59	0.5	0.11	-0.13	0.023	-0.11	0.048	-0.3	0.011	-0.53	0.47	-0.057	-0.38	1

After this, for regression I though of creating a regression analysis between the 4 broad parameters and their 6h, 12h and 24h predictions to define a relationship that would help predict future values. I further visualized them using a scatter plot and the best fit line. And I also calculated each of their Mean Squared Error and R^2 values, whose analysis will be mentioned later. Following are the regression graphs:




Multiple Regression Results
Predictions
0
20
40
60
80
100
120
0
20
40
60
80
100
120
True Values (Diffuse Solar Radiation [W/m2])


Multiple Regression Results
Predictions
0
100
200
300
400
500
600
0
100
200
300
400
500
600
True Values (Direct Solar Radiation [W/m2])


Multiple Regression Results
Predictions
70
75
80
85
90
95
100
70
75
80
85
90
95
100
True Values (Outdoor Relative Humidity [%])

Analysis and Observations

Beginning with the Correlation Matrix:

1. **Outdoor Drybulb Temperature [C]:**

- **High Positive Correlation:**
  - With its 6h (0.84), 12h (0.79), and 24h (0.82) predictions.
  - With Direct Solar Radiation (0.46) and its predictions (0.91, 0.63).

- **Moderate Positive Correlation:**
  - With Diffuse Solar Radiation (0.19).

- **Possible Reasons:**
  - Temperature trends are consistent over short periods.
  - Solar radiation directly affects temperature.

2. **Outdoor Relative Humidity [%]:**

- **Low to Moderate Positive Correlation:**
  - With Drybulb Temperature and its predictions (0.067, 0.036, 0.12).
  - Higher within Relative Humidity predictions (0.63, 0.71).

- **Possible Reasons:**
  - Humidity influenced by temperature and other factors like precipitation.
  - Stable daily humidity patterns.

3. **Diffuse Solar Radiation [W/m2]:**

- **Moderate Positive Correlation:**
  - With Drybulb Temperature predictions (0.46).

- **Possible Reasons:**
  - Relationship with total solar radiation affecting temperature.

4. **Direct Solar Radiation [W/m2]:**

- **Moderate to High Positive Correlation:**
  - With Drybulb Temperature predictions (0.46, 0.63, 0.91).

- **Possible Reasons:**
  - Direct solar radiation significantly impacts surface temperature.

5. **Prediction Accuracy:**

- **High Correlation:**

  - Among predictions for the same parameters (e.g., temperature, humidity) at different intervals.
- **Decreasing Correlation Values:**
  - As prediction time increases, indicating higher uncertainty over longer periods.

Detailed Analysis

1. **Outdoor Drybulb Temperature:**
   - High Correlation Values:
     - Predictions for 6h, 12h, 24h (0.84, 0.79, 0.82) indicate consistent temperature trends.
   - High Correlation with Direct Solar Radiation:
     - 0.91 (6h), 0.63 (12h), reflecting the significant impact of solar radiation on temperature.

2. **Outdoor Relative Humidity**:
   - Moderate Correlations:
     - With temperature and solar radiation indicate complex influences.
   - High Correlations within Humidity Predictions:
     - Reflect stable humidity patterns over short-term forecasts.

3. Solar Radiation (Diffuse and Direct):
   - Moderate Correlation:
     - Between diffuse and direct solar radiation reflects their intrinsic relationship.
   - Temperature Correlation:
     - Higher correlation with direct solar radiation due to its significant impact.

4. Temporal Predictions:
   - Decreasing Correlation:
     - Higher uncertainty and variability in weather patterns over longer periods.

Conclusion

The correlation matrix provides insights into how weather parameters in Islamabad interact and how effectively their short-term predictions are correlated. High correlations among drybulb temperature predictions suggest reliable forecasting models, while the moderate correlations with relative humidity highlight the complexity of predicting humidity due to various influencing factors. Solar radiation's significant impact on temperature is evident from their strong correlations.

This analysis helps in understanding the dependencies and can aid in improving predictive models by considering additional factors for parameters with lower correlations.

Next we move onto the scatter plots:

Each scatter plot compares the true values (x-axis) with the predicted values (y-axis), accompanied by a red dashed line indicating the ideal fit where predictions perfectly match the true values. Additionally, the Mean Squared Error (MSE) and R^2 scores provide a quantitative measure of the model's performance.

Scatter Plot Analysis with MSE and R^2 Scores

1. Direct Solar Radiation (W/m²)

- Mean Squared Error: 16870.537
- R^2 Score: 0.447
- Description: The scatter plot depicts the relationship between the true and predicted values for Direct Solar Radiation.

Analysis:

- The high MSE indicates substantial prediction errors.
- The R^2 score of 0.447 suggests that the model explains about 44.7% of the variance in the true values, indicating moderate predictive power.
- The wide scatter of data points, especially for values above 100 W/m², confirms significant deviations from the ideal fit line, suggesting the model struggles with higher values of Direct Solar Radiation.

2. Diffuse Solar Radiation (W/m²)

- Mean Squared Error: 131.487
- R^2 Score: 0.869
- Description: The scatter plot shows the relationship between the true values and the predicted values for Diffuse Solar Radiation.

Analysis:

- The relatively lower MSE indicates better prediction accuracy compared to Direct Solar Radiation.
- The R^2 score of 0.869 indicates that the model explains approximately 86.9% of the variance in the true values, showing strong predictive power.
- Despite the overall good performance, there is some spread in the data, particularly at higher values, indicating increasing prediction errors as the true values increase.

3. Outdoor Relative Humidity (%)

- Mean Squared Error: 49.972
- R^2 Score: 0.258

- Description: The scatter plot shows the true versus predicted values for Outdoor Relative Humidity.

Analysis:

- The MSE indicates moderate prediction errors.
- The R² score of 0.258 suggests that the model explains only 25.8% of the variance in the true values, indicating limited predictive power.
- The points are relatively closer to the ideal fit line, indicating better model performance compared to the solar radiation variables, but the spread suggests room for improvement in capturing the variability in outdoor relative humidity.

1. Overall Model Performance:

- The model performs variably across different weather parameters.
- Predictions for Diffuse Solar Radiation are more accurate compared to those for Direct Solar Radiation and Outdoor Relative Humidity.

2. Accuracy and Error Trends:

- **Direct Solar Radiation**: High MSE and moderate R² score indicate significant prediction errors, especially at higher values.
- **Diffuse Solar Radiation**: Moderate MSE and high R² score suggest good overall accuracy with increasing errors at higher values.
- **Outdoor Relative Humidity**: Moderate MSE and low R² score indicate reasonable accuracy but limited predictive power.

3. Potential Improvements:

- **Model Refinement:** Improve the model to handle higher values of solar radiation, possibly by incorporating more data or using advanced regression techniques.
- **Feature Engineering**: Adding more relevant features might help in capturing the complexities of solar radiation and humidity better.
- **Regularization Techniques:** Applying regularization might help in reducing overfitting, leading to better generalization.

4. Practical Implications:

- Accurate humidity predictions can be beneficial for various applications like agriculture and weather forecasting.
- Improved solar radiation predictions are crucial for solar energy projects and environmental studies.

Conclusion

The multiple regression model shows varying levels of accuracy for different weather parameters in Islamabad. While it performs well for Diffuse Solar Radiation, there is room for improvement in predicting Direct Solar Radiation and Outdoor Relative Humidity. Enhancing the model's capabilities through additional data, refined features, and advanced techniques can lead to more reliable weather predictions.