Data Science Tech Assessment: Weather Trend Forecasting

Mariam Tmane

TABLE OF CONTENTS

01 **Introduction**

02 **Data Cleaning and Preprocessing**

03 **EDA**

04 **Observations**

05 **Model and Evaluation**

06 **Conclusion**

01

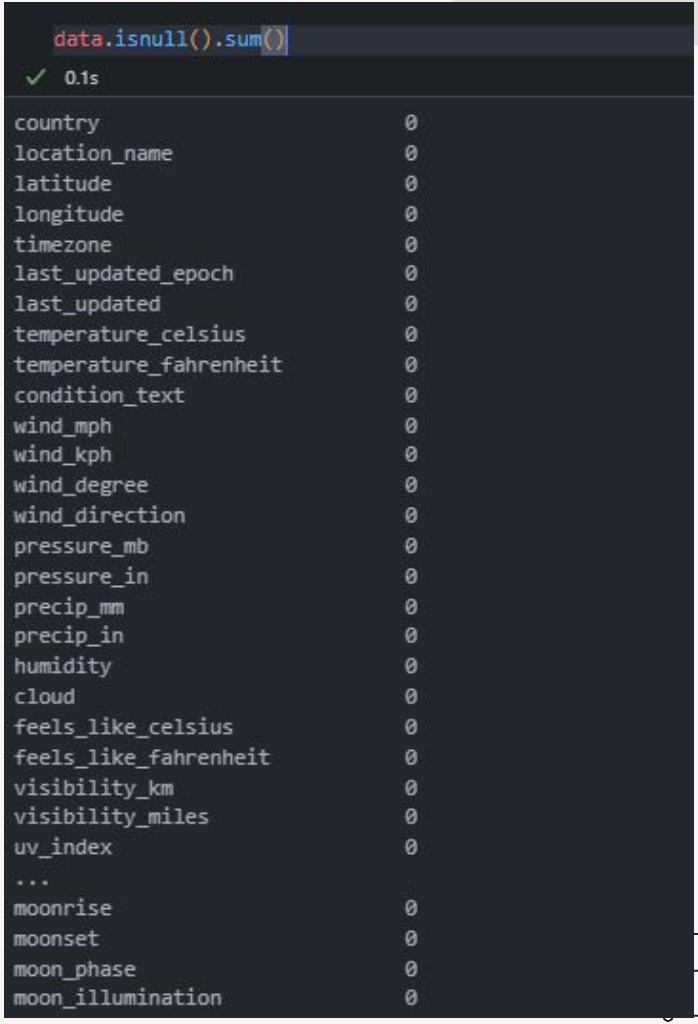
Data Cleaning and Preprocessing

Objective?

Ensuring that the data is clean, consistent, and ready for analysis by handling missing values, outliers, and scaling.

1. Handling Missing Values

There are no missing values in any of the columns in the data.



```
data.isnull().sum()
```

✓ 0.1s

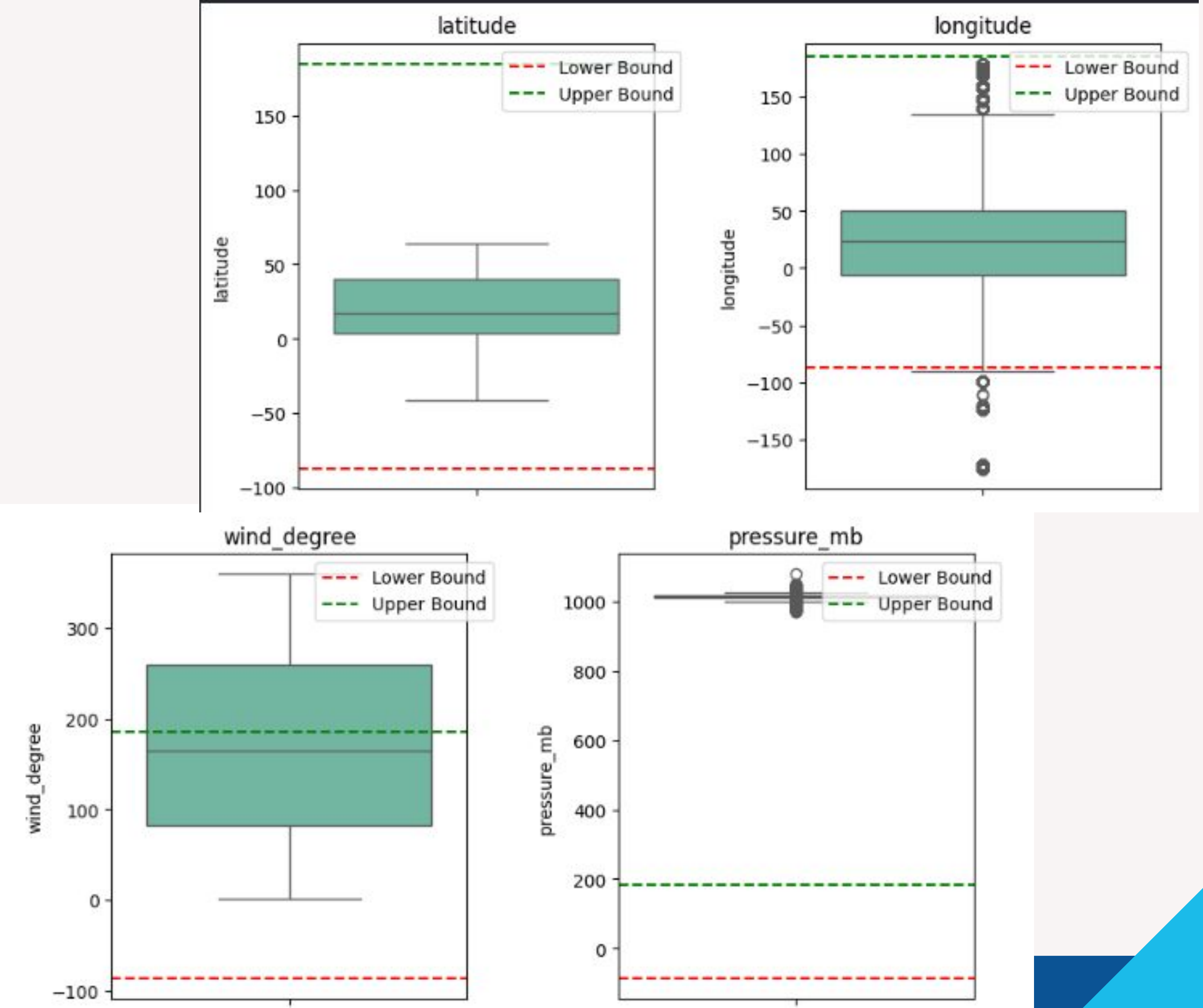
```
country                         0
location_name                   0
latitude                        0
longitude                       0
timezone                        0
last_updated_epoch              0
last_updated                    0
temperature_celsius             0
temperature_fahrenheit          0
condition_text                  0
wind_mph                        0
wind_kph                        0
wind_degree                     0
wind_direction                  0
pressure_mb                     0
pressure_in                     0
precip_mm                       0
precip_in                       0
humidity                        0
cloud                           0
feels_like_celsius              0
feels_like_fahrenheit           0
visibility_km                   0
visibility_miles                0
uv_index                        0
...
moonrise                        0
moonset                         0
moon_phase                      0
moon_illumination               0
```

2. Identifying Outliers

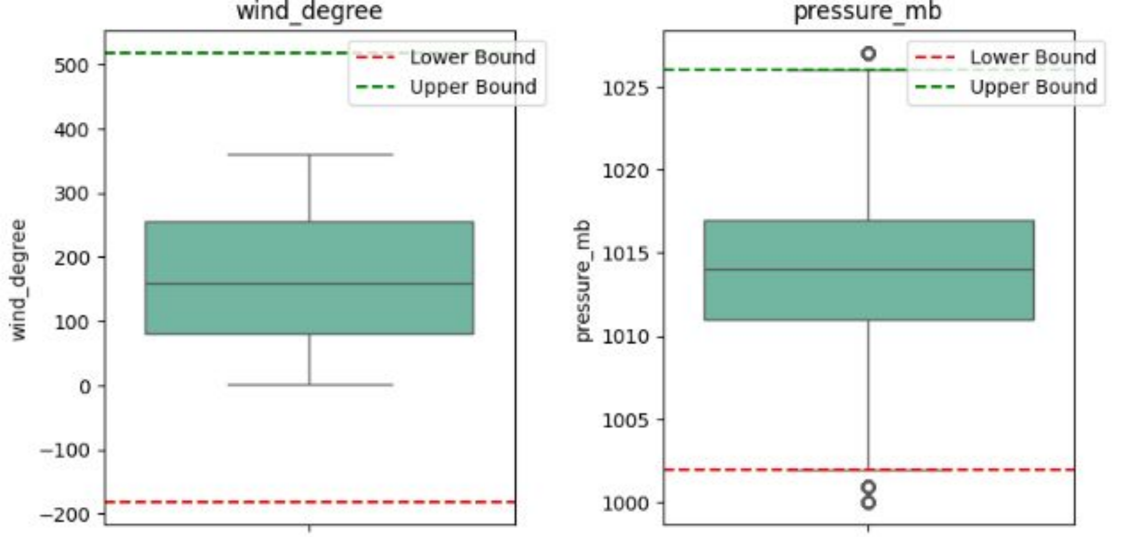
Outliers were detected using the Interquartile Range (IQR) method.

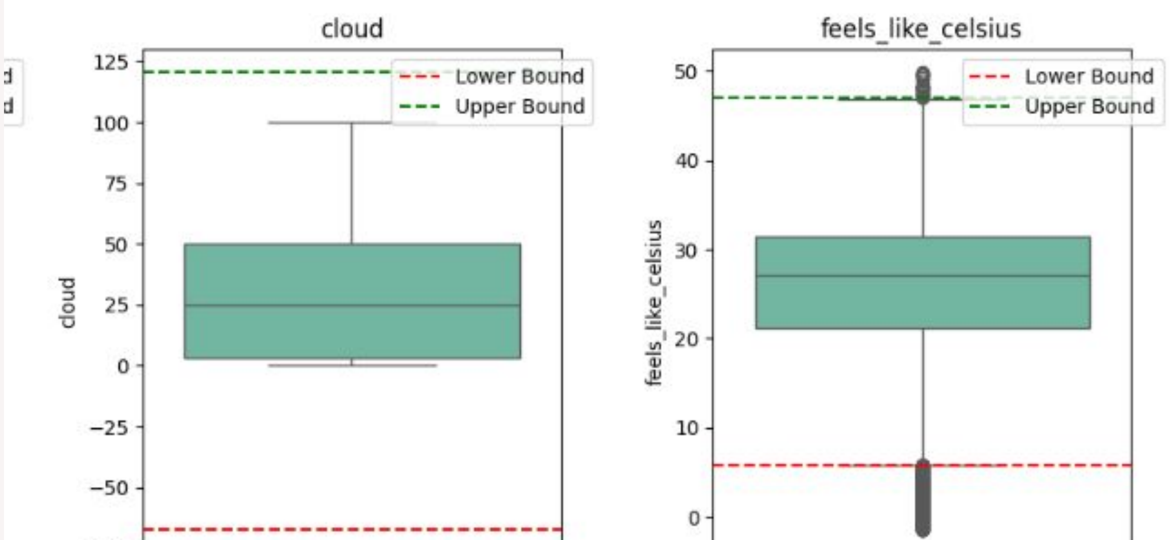
- Lower Bound: Calculated as Q1 - 1.5(Q3 - Q1)
- Upper Bound: Calculated as Q3 +1.5(Q3 - Q1)

- **Box plots were also used to visually identify the outliers and confirm their impact on the data.**



- **Box plots after handling the outliers using IQR method**

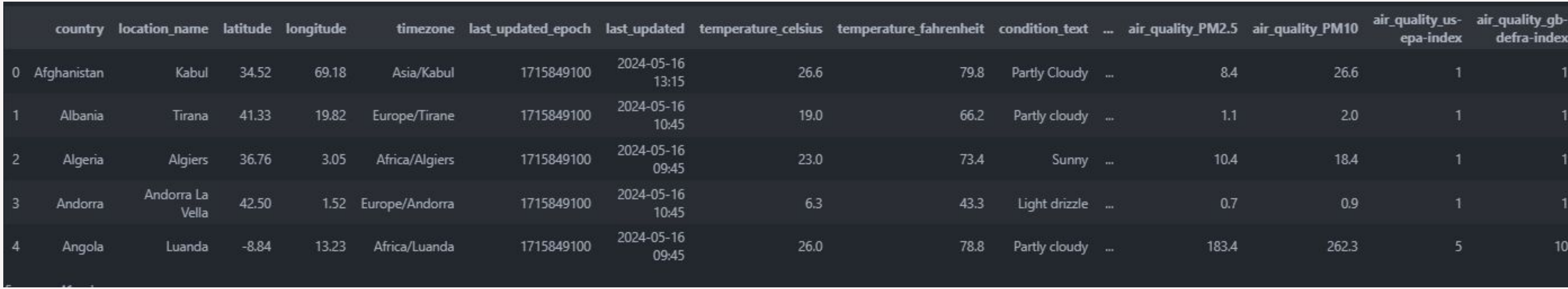




3. Normalizing the Data

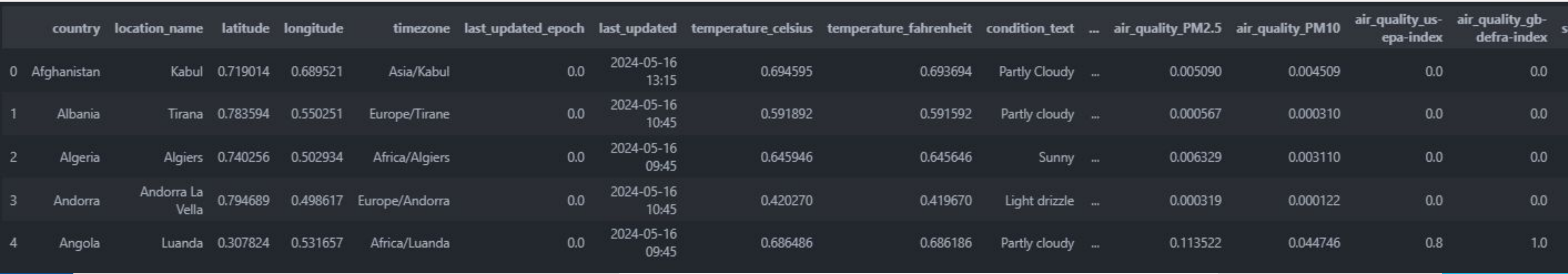
Method 1: Sickitlearn MinMaxScalar()

Before Normalization:



	country	location_name	latitude	longitude	timezone	last_updated_epoch	last_updated	temperature_celsius	temperature_fahrenheit	condition_text	...	air_quality_PM2.5	air_quality_PM10	air_quality_us-epa-index	air_quality_gb-defra-index
0	Afghanistan	Kabul	34.52	69.18	Asia/Kabul	1715849100	2024-05-16 13:15	26.6	79.8	Partly Cloudy	...	8.4	26.6	1	1
1	Albania	Tirana	41.33	19.82	Europe/Tirane	1715849100	2024-05-16 10:45	19.0	66.2	Partly cloudy	...	1.1	2.0	1	1
2	Algeria	Algiers	36.76	3.05	Africa/Algiers	1715849100	2024-05-16 09:45	23.0	73.4	Sunny	...	10.4	18.4	1	1
3	Andorra	Andorra La Vella	42.50	1.52	Europe/Andorra	1715849100	2024-05-16 10:45	6.3	43.3	Light drizzle	...	0.7	0.9	1	1
4	Angola	Luanda	-8.84	13.23	Africa/Luanda	1715849100	2024-05-16 09:45	26.0	78.8	Partly cloudy	...	183.4	262.3	5	10

After Normalization:

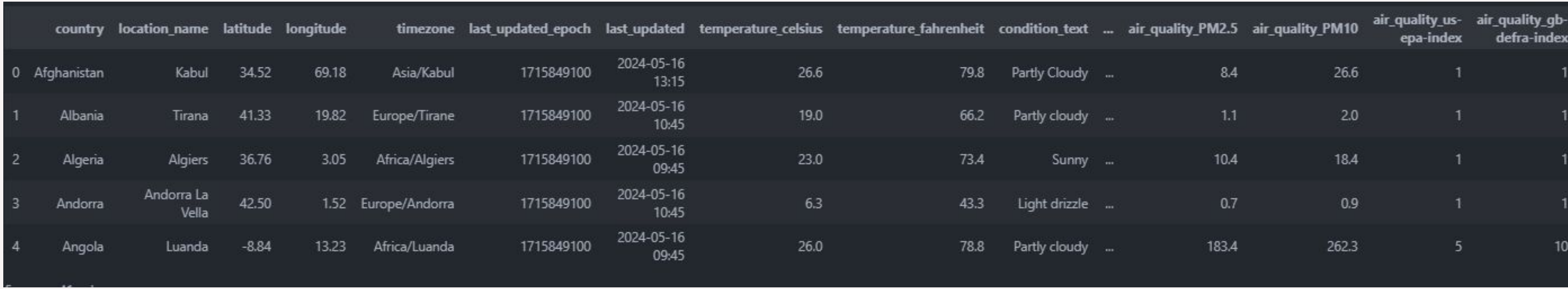


	country	location_name	latitude	longitude	timezone	last_updated_epoch	last_updated	temperature_celsius	temperature_fahrenheit	condition_text	...	air_quality_PM2.5	air_quality_PM10	air_quality_us-epa-index	air_quality_gb-defra-index	s
0	Afghanistan	Kabul	0.719014	0.689521	Asia/Kabul	0.0	2024-05-16 13:15	0.694595	0.693694	Partly Cloudy	...	0.005090	0.004509	0.0	0.0	
1	Albania	Tirana	0.783594	0.550251	Europe/Tirane	0.0	2024-05-16 10:45	0.591892	0.591592	Partly cloudy	...	0.000567	0.000310	0.0	0.0	
2	Algeria	Algiers	0.740256	0.502934	Africa/Algiers	0.0	2024-05-16 09:45	0.645946	0.645646	Sunny	...	0.006329	0.003110	0.0	0.0	
3	Andorra	Andorra La Vella	0.794689	0.498617	Europe/Andorra	0.0	2024-05-16 10:45	0.420270	0.419670	Light drizzle	...	0.000319	0.000122	0.0	0.0	
4	Angola	Luanda	0.307824	0.531657	Africa/Luanda	0.0	2024-05-16 09:45	0.686486	0.686186	Partly cloudy	...	0.113522	0.044746	0.8	1.0	

3. Normalizing the Data

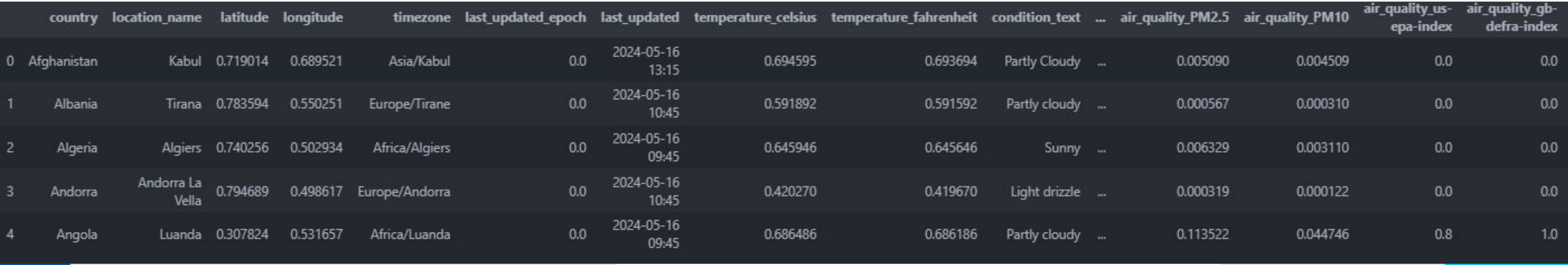
Method 2: Manually normalizing using min(), max()

Before Normalization:



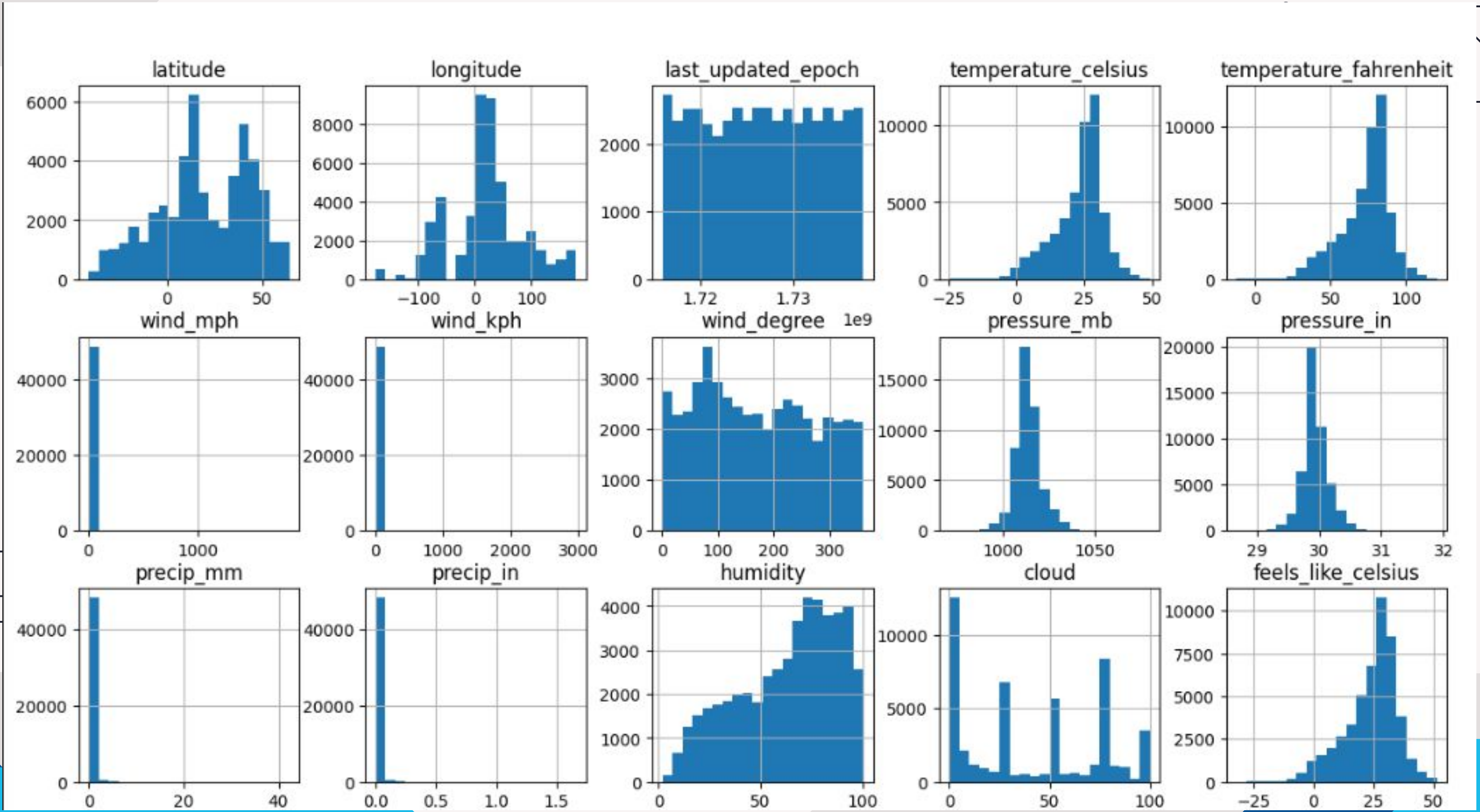
	country	location_name	latitude	longitude	timezone	last_updated_epoch	last_updated	temperature_celsius	temperature_fahrenheit	condition_text	...	air_quality_PM2.5	air_quality_PM10	air_quality_us-epa-index	air_quality_gb-defra-index
0	Afghanistan	Kabul	34.52	69.18	Asia/Kabul	1715849100	2024-05-16 13:15	26.6	79.8	Partly Cloudy	...	8.4	26.6	1	1
1	Albania	Tirana	41.33	19.82	Europe/Tirane	1715849100	2024-05-16 10:45	19.0	66.2	Partly cloudy	...	1.1	2.0	1	1
2	Algeria	Algiers	36.76	3.05	Africa/Algiers	1715849100	2024-05-16 09:45	23.0	73.4	Sunny	...	10.4	18.4	1	1
3	Andorra	Andorra La Vella	42.50	1.52	Europe/Andorra	1715849100	2024-05-16 10:45	6.3	43.3	Light drizzle	...	0.7	0.9	1	1
4	Angola	Luanda	-8.84	13.23	Africa/Luanda	1715849100	2024-05-16 09:45	26.0	78.8	Partly cloudy	...	183.4	262.3	5	10

After Normalization:



	country	location_name	latitude	longitude	timezone	last_updated_epoch	last_updated	temperature_celsius	temperature_fahrenheit	condition_text	...	air_quality_PM2.5	air_quality_PM10	air_quality_us-epa-index	air_quality_gb-defra-index
0	Afghanistan	Kabul	0.719014	0.689521	Asia/Kabul	0.0	2024-05-16 13:15	0.694595	0.693694	Partly Cloudy	...	0.005090	0.004509	0.0	0.0
1	Albania	Tirana	0.783594	0.550251	Europe/Tirane	0.0	2024-05-16 10:45	0.591892	0.591592	Partly cloudy	...	0.000567	0.000310	0.0	0.0
2	Algeria	Algiers	0.740256	0.502934	Africa/Algiers	0.0	2024-05-16 09:45	0.645946	0.645646	Sunny	...	0.006329	0.003110	0.0	0.0
3	Andorra	Andorra La Vella	0.794689	0.498617	Europe/Andorra	0.0	2024-05-16 10:45	0.420270	0.419670	Light drizzle	...	0.000319	0.000122	0.0	0.0
4	Angola	Luanda	0.307824	0.531657	Africa/Luanda	0.0	2024-05-16 09:45	0.686486	0.686186	Partly cloudy	...	0.113522	0.044746	0.8	1.0

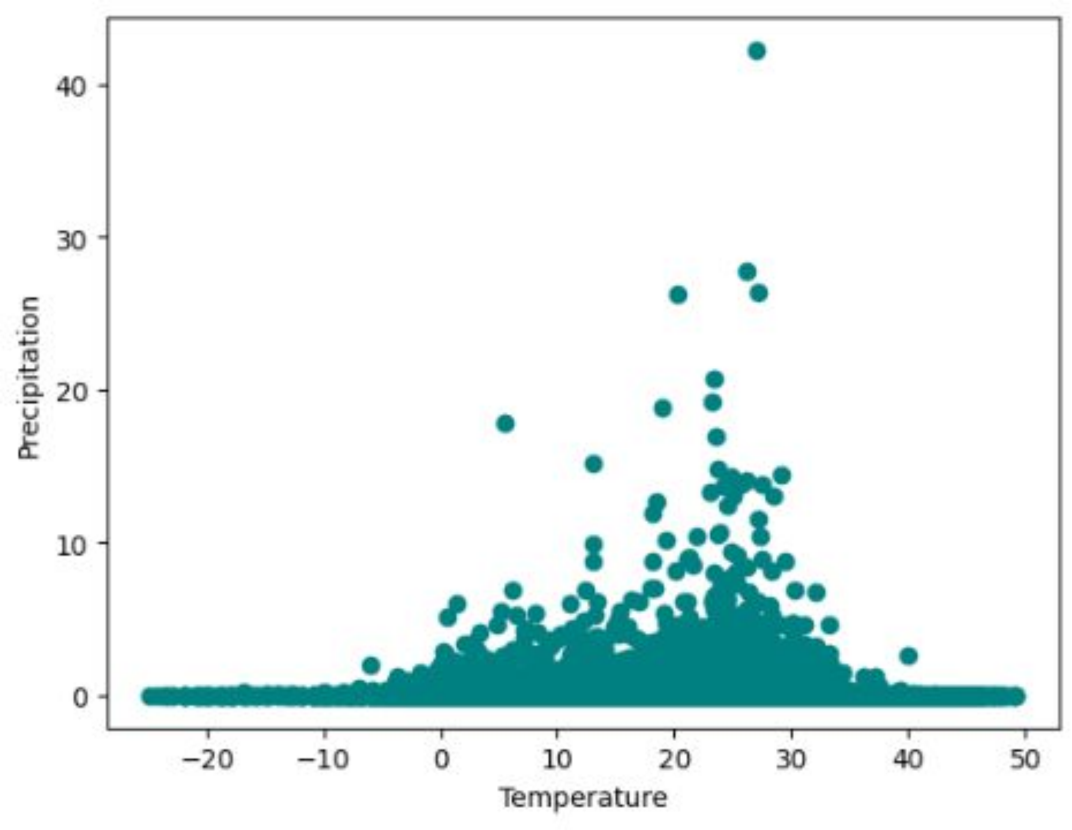
Histograms and Distributions:



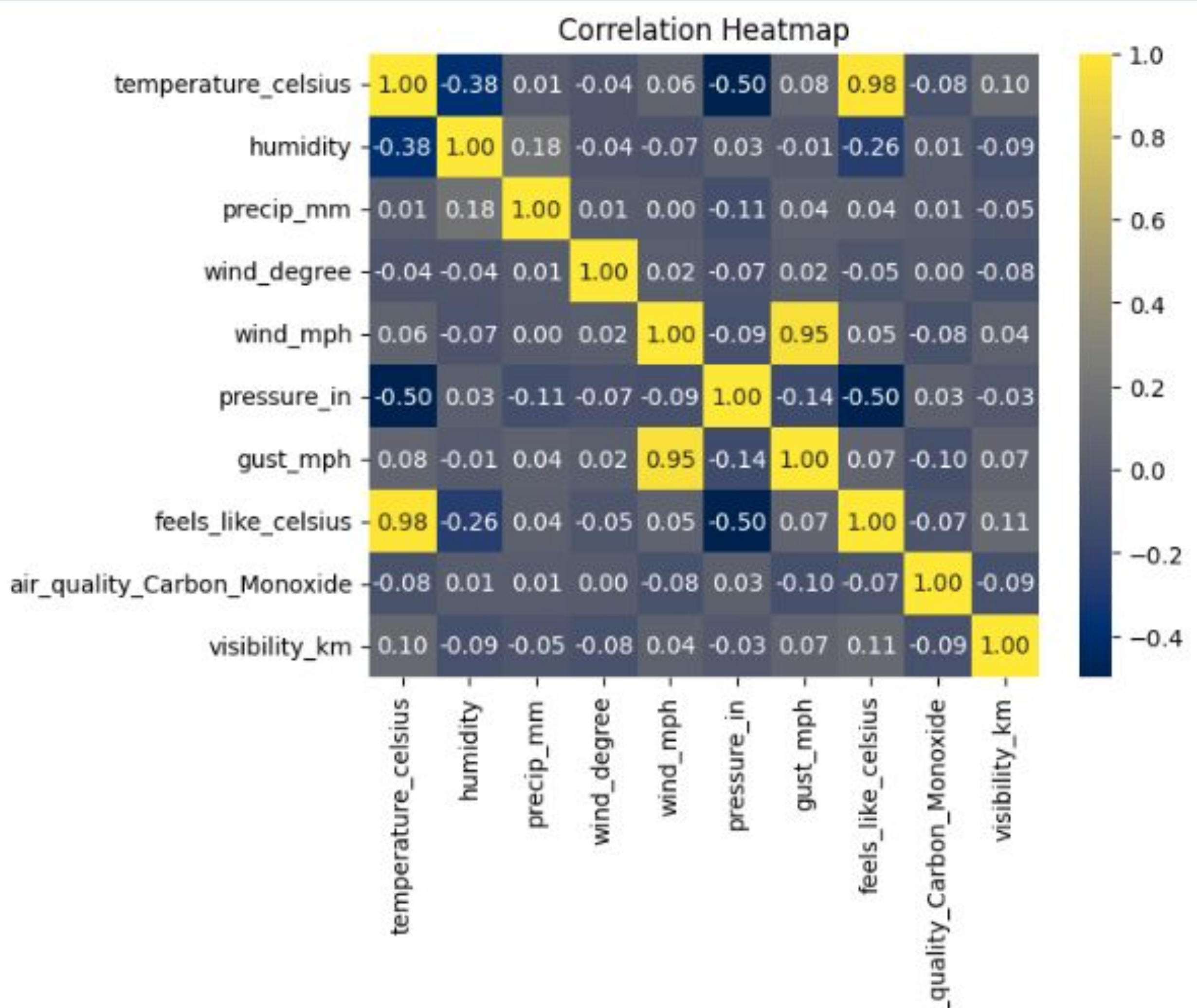
02

Exploratory Data Analysis (EDA):

Visualization (scatterplot) of correlation between precipitation and temperature



Heatmap: Correlation of several important features



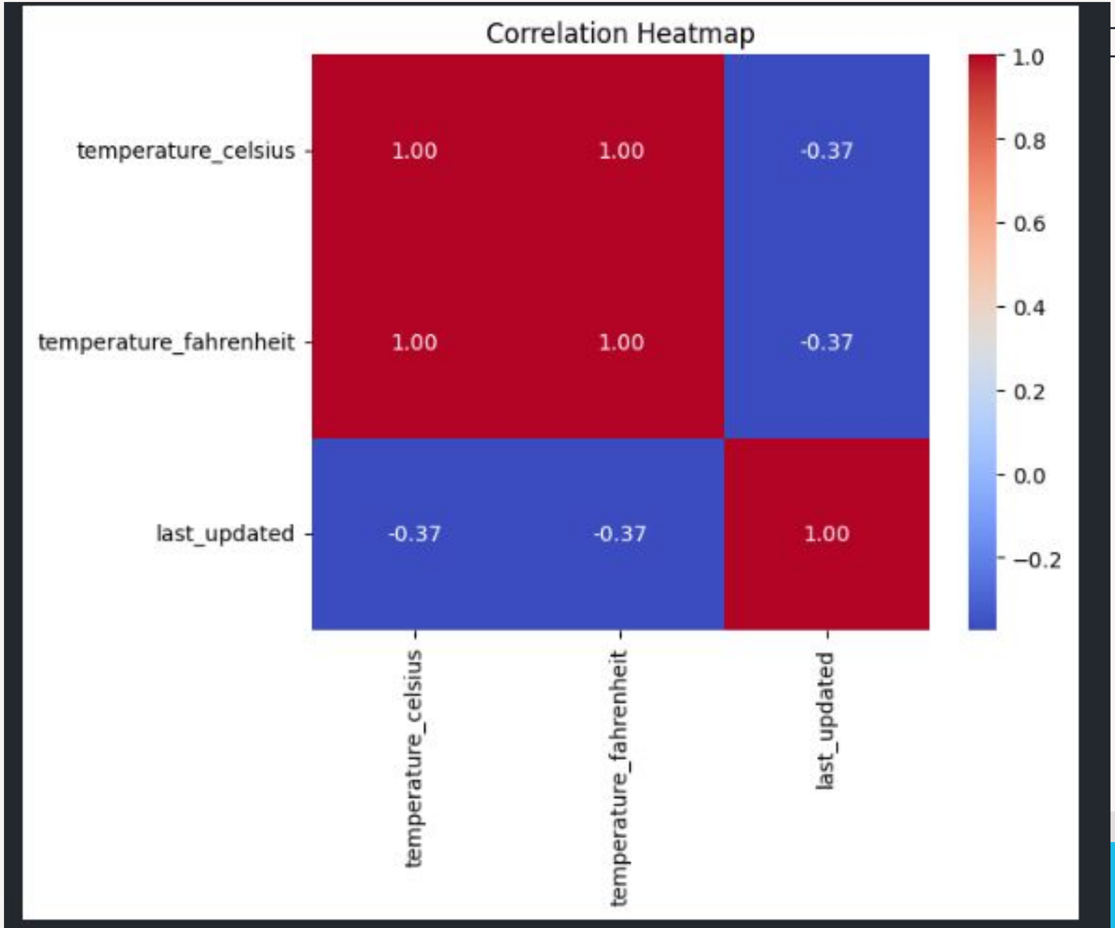
Correlation Heatmap

	temperature_celsius	humidity	precip_mm	wind_degree	wind_mph	pressure_in	gust_mph	feels_like_celsius	_quality_Carbon_Monoxide	visibility_km
temperature_celsius	1.00	-0.38	0.01	-0.04	0.06	-0.50	0.08	0.98	-0.08	0.10
humidity	-0.38	1.00	0.18	-0.04	-0.07	0.03	-0.01	-0.26	0.01	-0.09
precip_mm	0.01	0.18	1.00	0.01	0.00	-0.11	0.04	0.04	0.01	-0.05
wind_degree	-0.04	-0.04	0.01	1.00	0.02	-0.07	0.02	-0.05	0.00	-0.08
wind_mph	0.06	-0.07	0.00	0.02	1.00	-0.09	0.95	0.05	-0.08	0.04
pressure_in	-0.50	0.03	-0.11	-0.07	-0.09	1.00	-0.14	-0.50	0.03	-0.03
gust_mph	0.08	-0.01	0.04	0.02	0.95	-0.14	1.00	0.07	-0.10	0.07
feels_like_celsius	0.98	-0.26	0.04	-0.05	0.05	-0.50	0.07	1.00	-0.07	0.11
air_quality_Carbon_Monoxide	-0.08	0.01	0.01	0.00	-0.08	0.03	-0.10	-0.07	1.00	-0.09
visibility_km	0.10	-0.09	-0.05	-0.08	0.04	-0.03	0.07	0.11	-0.09	1.00

The high correlation between ‘feels_like_celsius’ and ‘temperature_celsius’, means they teach the model the same thing, and can be combined into one feature for dimensionality reduction purposes.

How much does the temperature vary between the last update made?

-0.37



Observations

Country with the highest precipitation:

- country 84.00
- precip_mm 42.24

Country with the lowest precipitation:

- country 84.00
- precip_mm 42.24

Average humidity:

- 63.23521684694485

What is the average wind speed in km:

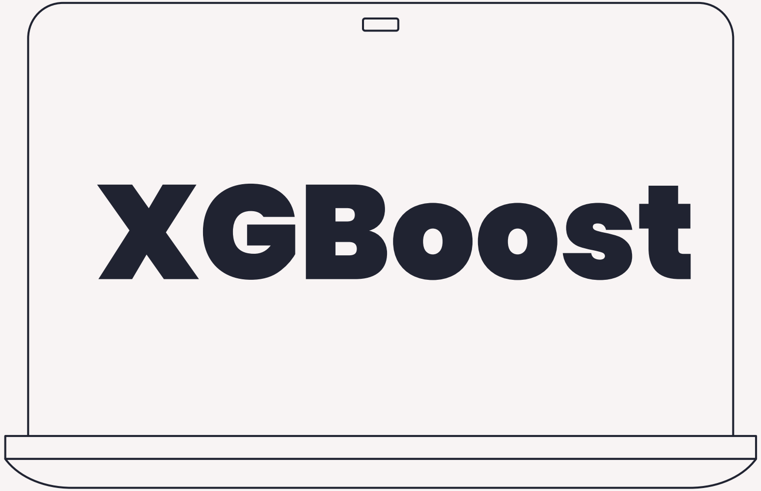
- 13.334174175406908

How many records have a wind speed > average wind speed? What's the percentage?

- Number of records with a wind speed greater than average wind speed (kph): 20026
- Percentage of records that have a wind speed greater than the average wind speed: 41%

03 Model Building:

- XGBoost (Extreme Gradient Boosting) is a machine learning algorithm based on decision trees.
- Steps:
  - Performed a **train-test split** to evaluate the model's performance on unseen data (80% train, 20% test)
  - Instantiated and trained the model
  - Evaluate the model on testing dataset using three different metrics:
    - MSE
    - RMSE
    - MAE



XGBoost

```
Test scores for testing data set:
MSE: 0.03897702586760852
RMSE: 0.19742600099178556
MAE: 0.0600163983106448

Test scores for training data set:
MSE: 0.004766101814151707
RMSE: 0.06903695976903754
MAE: 0.04254048546037757
```

Conclusion for XGBoost:

The results for MSE, RMSE, and MAE for both testing and training dataset indicate that the model generalizes relatively well, as indicated by the very low MSE, RMSE, and MAE values. This suggests that XGBoost is able to capture the relationships in the data efficiently, but could still improve, which can be done through:

- Cross validation or fine-tuning hyperparameters
- Further Analyze data for anomalies or outliers
- Feature engineering to combine features and dimensionality reduction.

Mission

I have completed this assessment and am eager to join the internship program at PM Accelerator. I am excited to contribute to PM Accelerator’s mission, which is ***to break down financial barriers and achieve educational fairness.***

I am committed to helping create accessible opportunities for all and look forward to making a meaningful impact through the program.

THANKS!