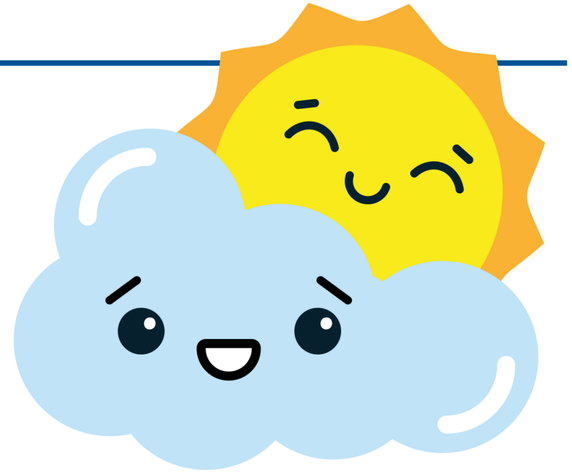
CapyData

Ho Chi Minh City Weather & Air Quality Analysis

Intro to Data Science

PROJECT MANAGEMENT

CapyData TEAM MEMBERS



Cao Hoang Loc
22127012

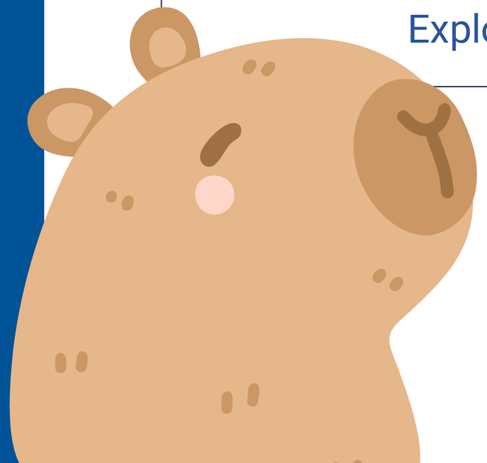
Data Pre-processing
Data Modeling - Classification

Vo Nguyen Phuong Quynh
22127255

Team leader
Exploring data Analysis

Phan Vo Minh Tue
22127255

Data Collecting
Exploring data Analysis



Pham Anh Van
21127039

Data Pre-processing
Data Modeling - Predict

CapyData

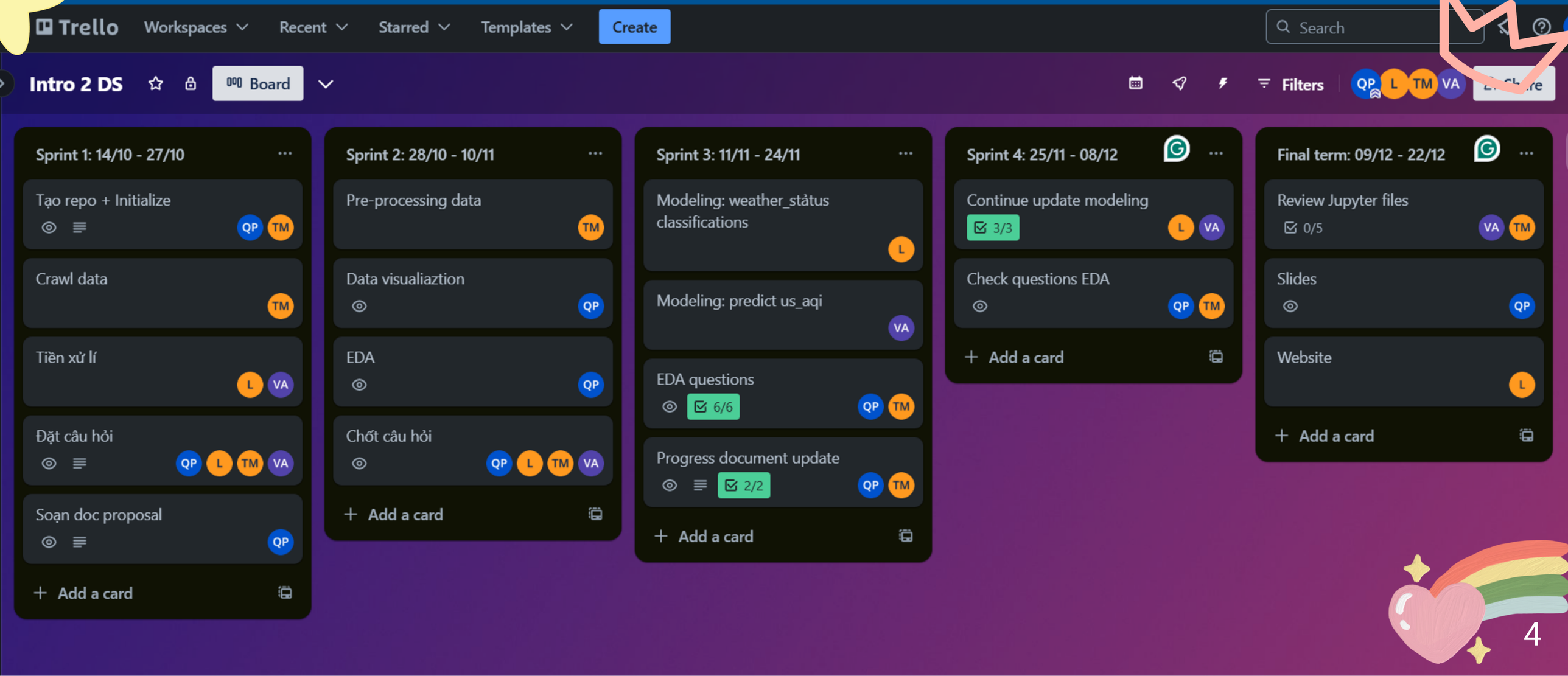
TOPIC

HCMC Weather & Air Quality Analysis

Ho Chi Minh City is dealing with worsening air quality due to its tropical monsoon climate. Our project will examine the connection between weather and air pollution from October 2022 to September 2024.

CapyData

PLAN FOR PROJECT

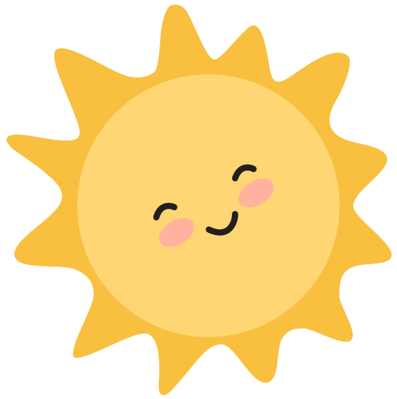


ACHIEVEMENT

Criteria for grading	Proportion
Apply suitable data science processes (collection, preprocessing, analysis, data modelling)	40%
Create questions and derive important, useful insights from the data	20%
Implement and explicitly explain the development process	15%
Compare to other modelling methods and point out the advantages and disadvantages	5%
Presentation	10%
Questions and Answers	10%
Bonus (interesting problem, impressive solution or further studies)	10%
Total	**110%**

CapyData

DATA COLLECTING



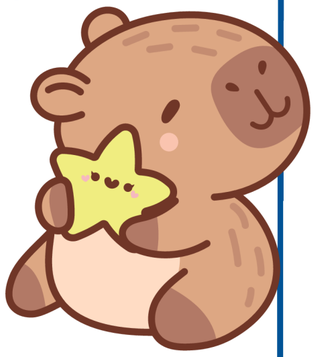
Data source

Weather and air quality were collected using the **Open-Meteo API**. Data was collected into 2 datasets (Weather & AQI), from***October 1, 2022***, to ***September 30, 2024***, providing hourly data points with **17,544 records** each dataset.

hcmc_weather_data.csv

10 features:

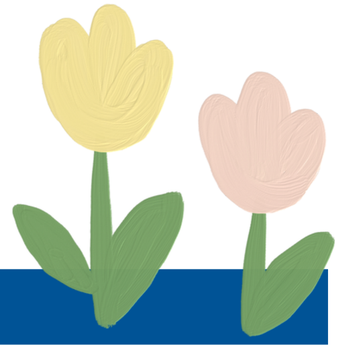
- Temperature and humidity at 2m
- Dew point and apparent temperature
- Precipitation and cloud cover
- Wind speed and direction at 10m
- Weather condition codes
- Vapour pressure deficit



hcmc_air_quality_data.csv

8 features:

- PM10 and PM2.5 concentrations
- Carbon monoxide (CO)
- Nitrogen dioxide (NO2)
- Sulphur dioxide (SO2)
- Ozone (O3)
- US Air Quality Index



Pre-processing

CapyData

Description

The Data Combination Process:

- Handling missing values
- Mapping weather_code to weather_status
- Validation of values
- Time continuity
- Outlier detection and handling
- Relationship validity
- Time zone adjustment

DATA Pre-processing

Handling missing values

The dataset contains 17544 non-null entries, so no missing values were found or handled.

Mapping `weather_code` to `weather_status`

Replaced numerical `weather_code` with descriptive `weather_status` using WMO Weather Interpretation Codes for better interpretability.

Validation of values

Verified that key variables fall within realistic ranges.:

- Temperature must in 20 - 55*C.
- Humidity must in 0-100.
- Other numeric value must be possitive.

--> All values passed validation checks.

Time continuity

No missing time intervals were detected; the dataset maintains hourly continuity.

CapyData

DATA Pre-processing

Outlier detection and handling

Identified: some example metrics:

- `temperature_2m: 185 outliers (1.05%)`
- `precipitation: 2183 outliers (12.44%)`
- `dew_point_2m: 738 outliers (4.21%)`

Kept outliers: Represent real phenomena, such as extreme weather or high pollution levels (pm10, pm2.5,...)

Handled outliers: Outliers deemed unrealistic (e.g., `temperature` or humidity values outside HCMC-specific ranges) were corrected by **IQR method.**

Relationship validity

No issues detected between variable relationships.

Time zone adjustment

Converted timestamps from UTC to UTC+7 (HCMC local time) for proper contextual analysis.

```
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   date_time                17544 non-null  datetime6
 1   temperature_2m           17544 non-null  float64
 2   relative_humidity_2m     17544 non-null  float64
 3   dew_point_2m             17544 non-null  float64
 4   apparent_temperature     17544 non-null  float64
 5   precipitation            17544 non-null  float64
 6   cloud_cover              17544 non-null  float64
 7   vapour_pressure_deficit  17544 non-null  float64
 8   wind_speed_10m           17544 non-null  float64
 9   wind_direction_10m       17544 non-null  float64
 10  pm10                     17544 non-null  float64
 11  pm2_5                    17544 non-null  float64
 12  carbon_monoxide          17544 non-null  float64
 13  nitrogen_dioxide         17544 non-null  float64
 14  sulphur_dioxide          17544 non-null  float64
 15  ozone                    17544 non-null  float64
 16  us_aqi                   17544 non-null  float64
 17  weather_status           17544 non-null  object
dtypes: datetime64[ns](1), float64(16), object(1)
```

CapyData

Pre-processing

The Data Combination Process:

- Handling missing values
- Mapping weather_code to weather_status
- Validation of values
- Time continuity
- Outlier detection and handling
- Relationship validity
- Time zone adjustment

clean_hcmc_waq.csv

- 17,544 records (rows)
- 18 attributes (columns): 16 numerical and 2 objective columns.

EDA

Statistics

Use df.describe() to calculate the numerical analysis of the numerical features:

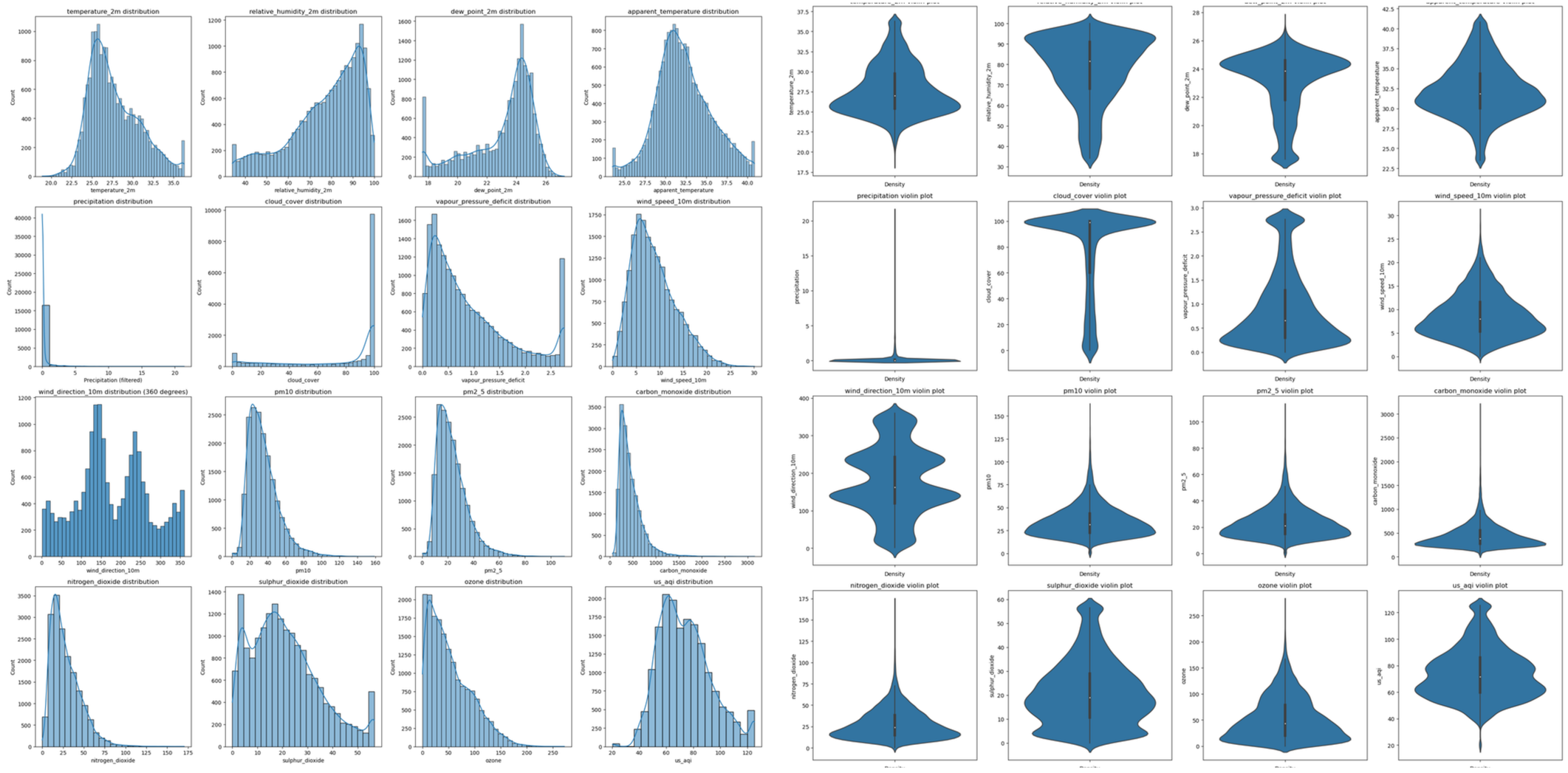
- Mean, range
- IQR value: 25%, 50%, 75%

- This suggests that the climate is consistently warm and humid, typical for tropical regions like Ho Chi Minh City.
- This indicates that air quality fluctuates, potentially due to factors like traffic or industrial activity.

CapyData

Distribution

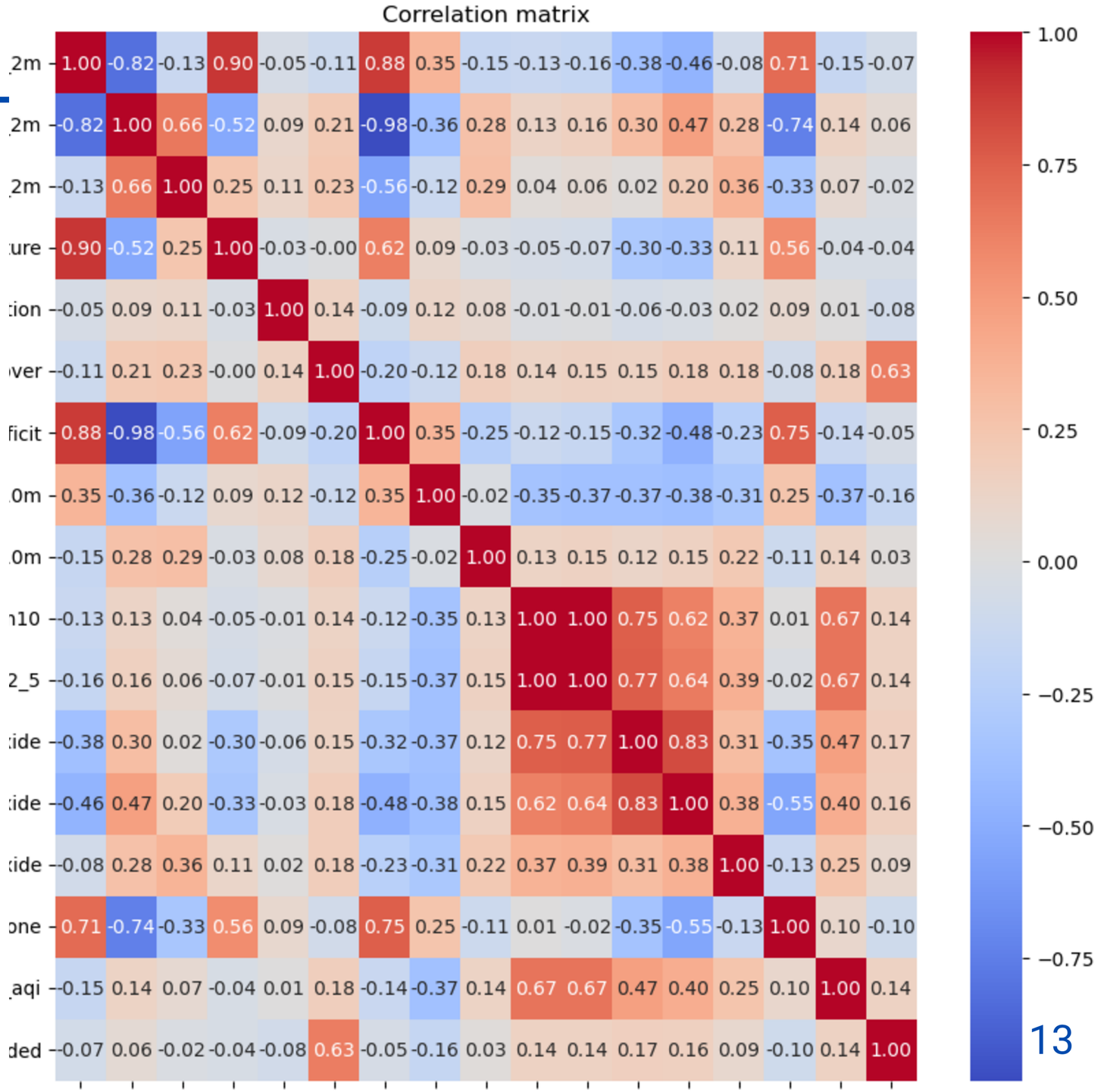
Use Histogram, violin chart, box plot



CapyData

Correlation matrix

Based on the correlation matrix, we assign that Weather and air quality in HCMC are linked together. Humidity inversely relates to vapour pressure deficit, while temperature and humidity impact ozone levels, and higher humidity may lower it.

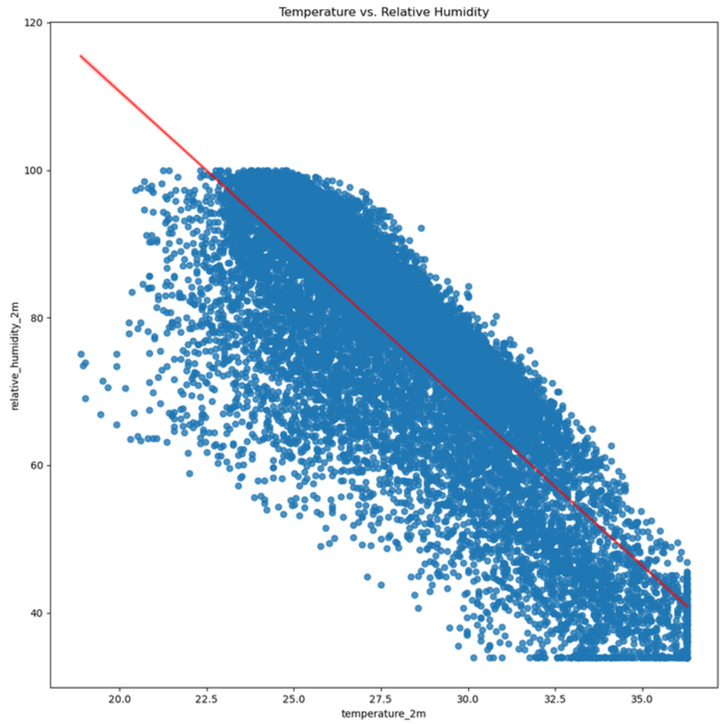


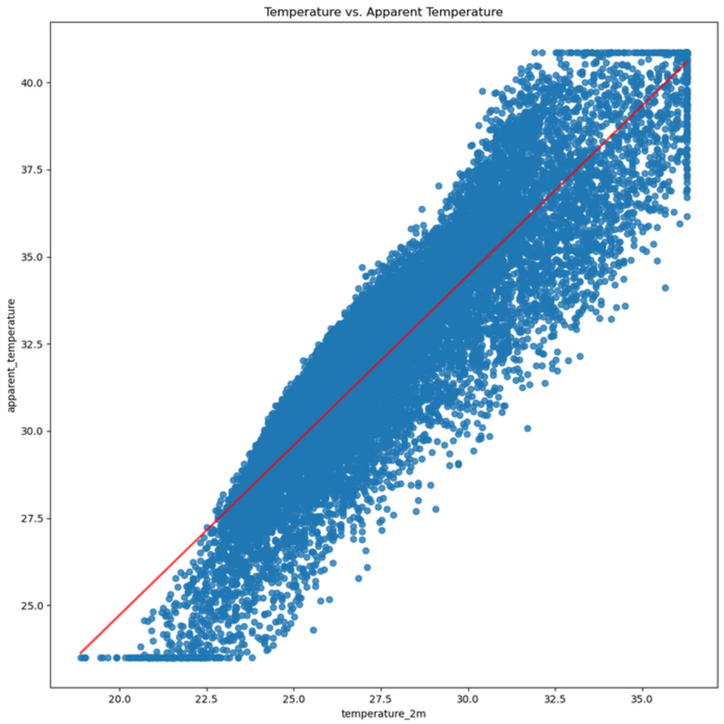
CapyData

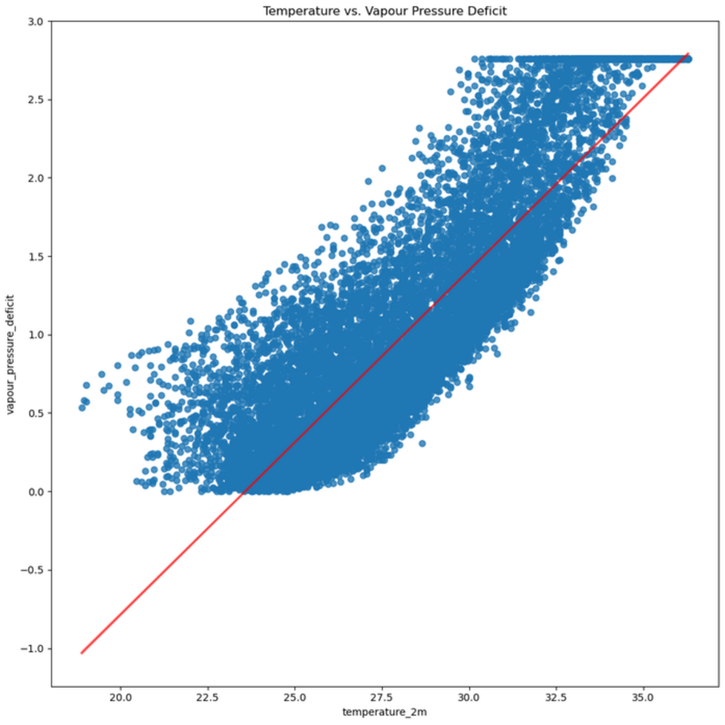
Relationship

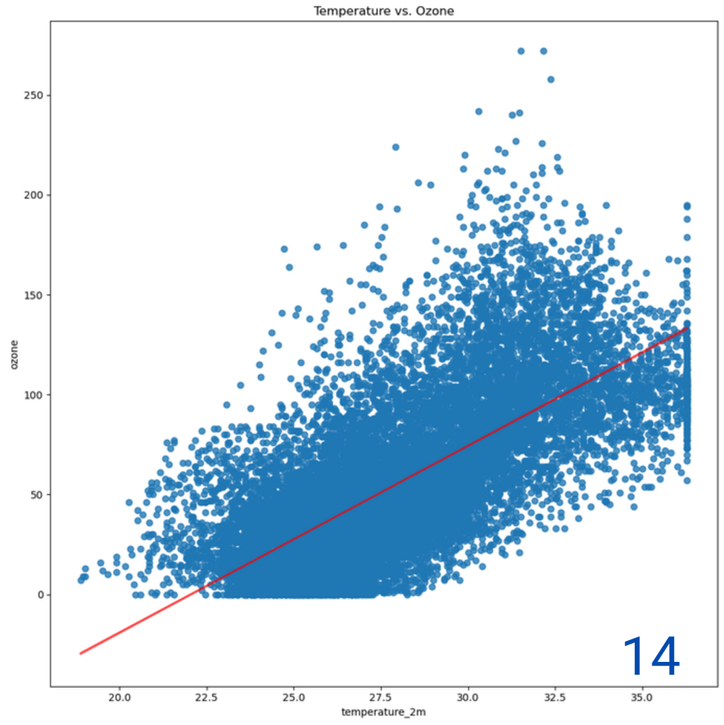
Temperature vs features

- Humidity
- Vapour Pressure Deficit
- Ozone





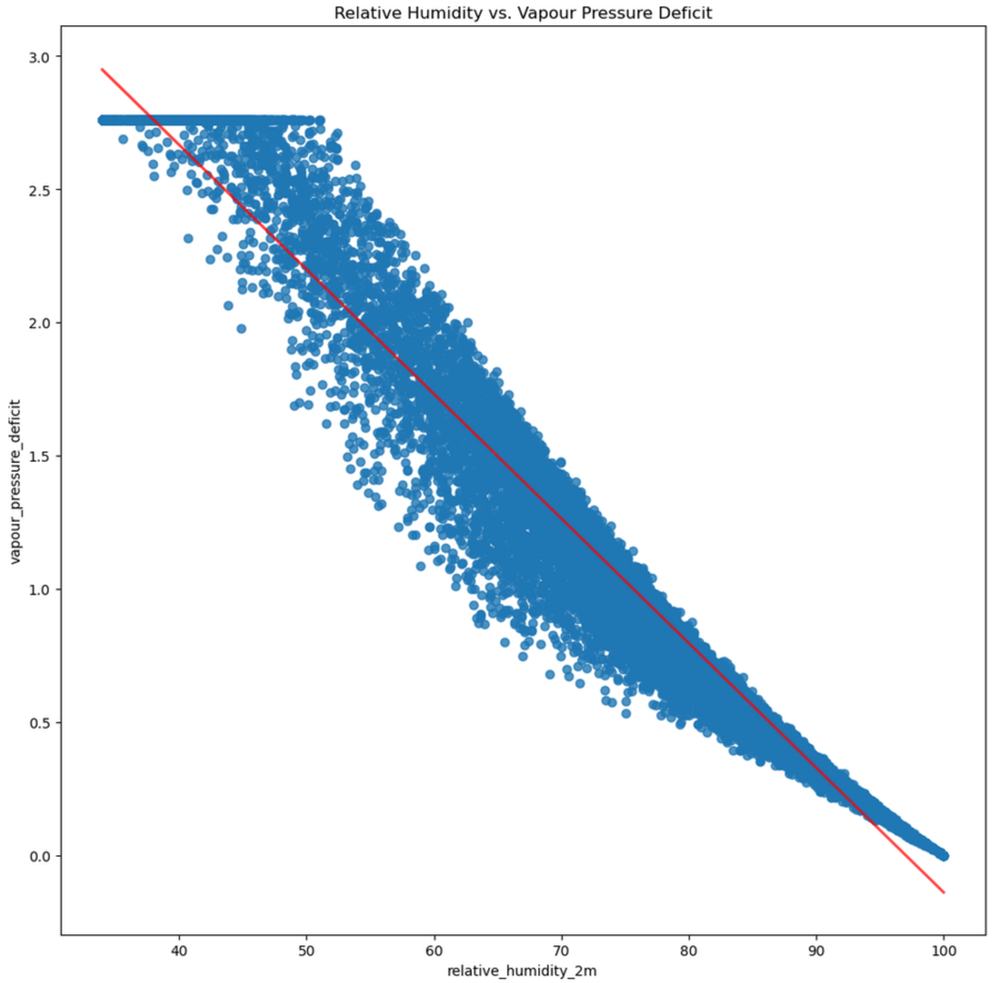


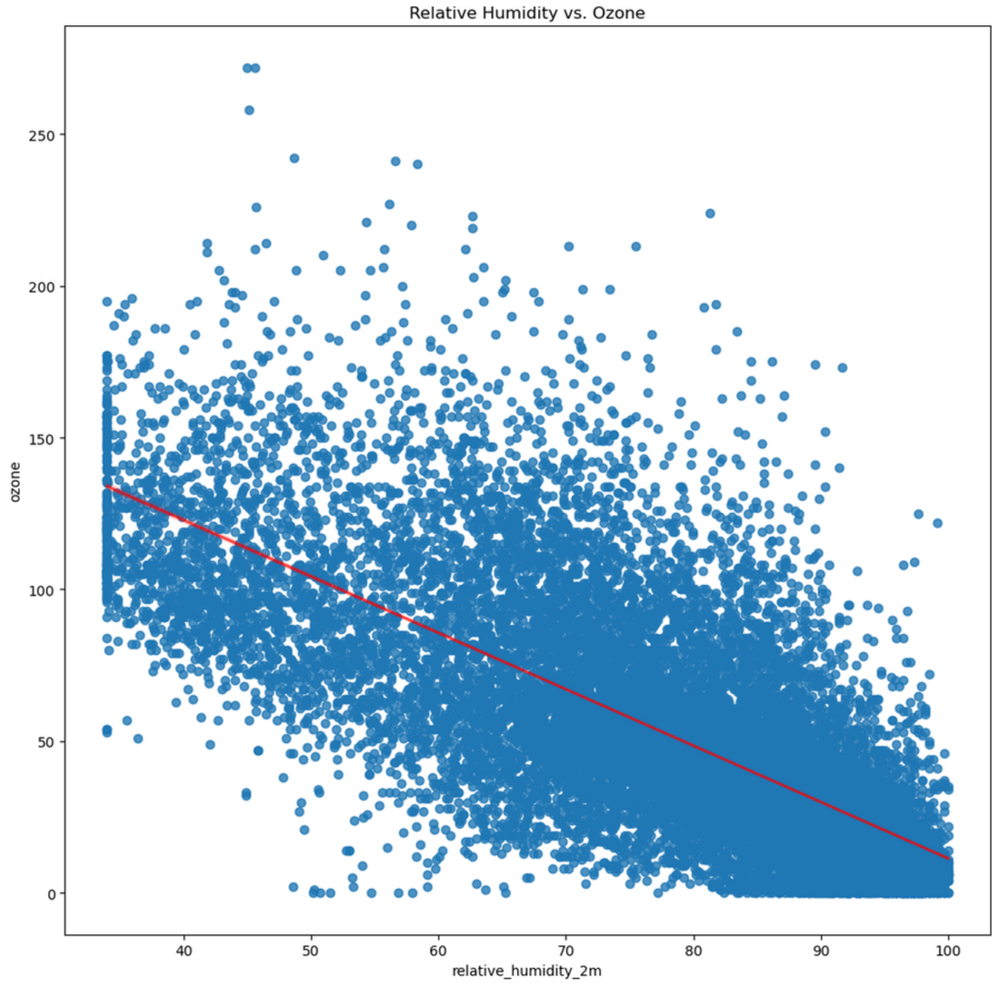


CapyData

Relationship

Humidity vs Vapour pressure Deficit and Ozone

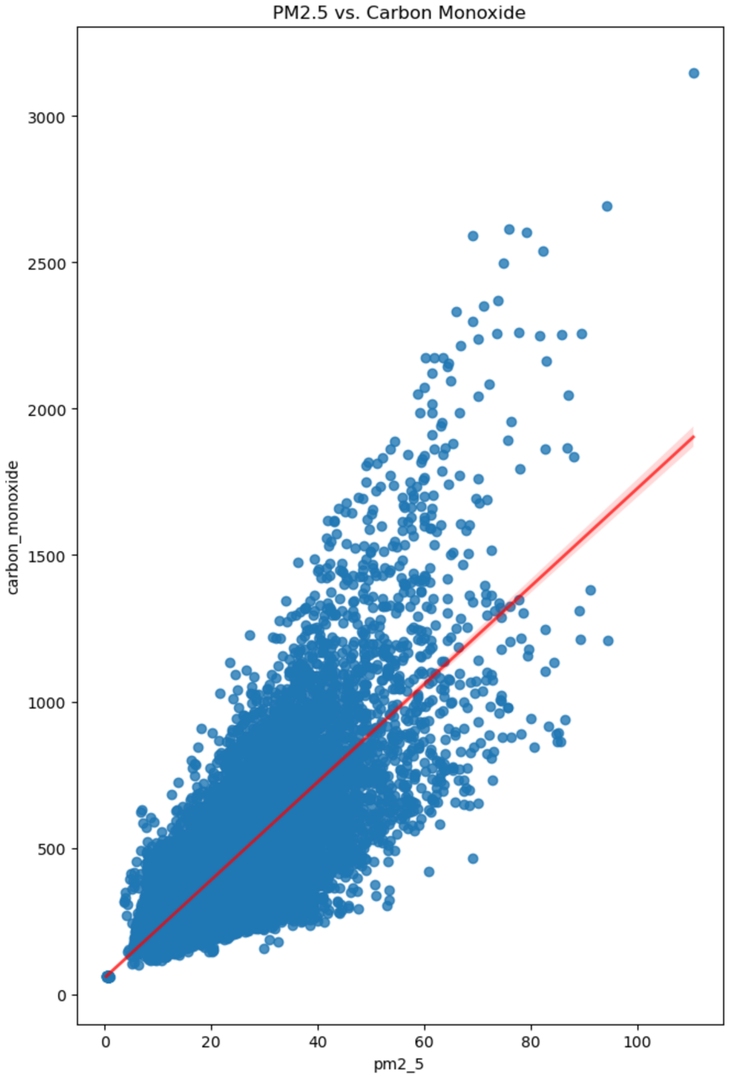


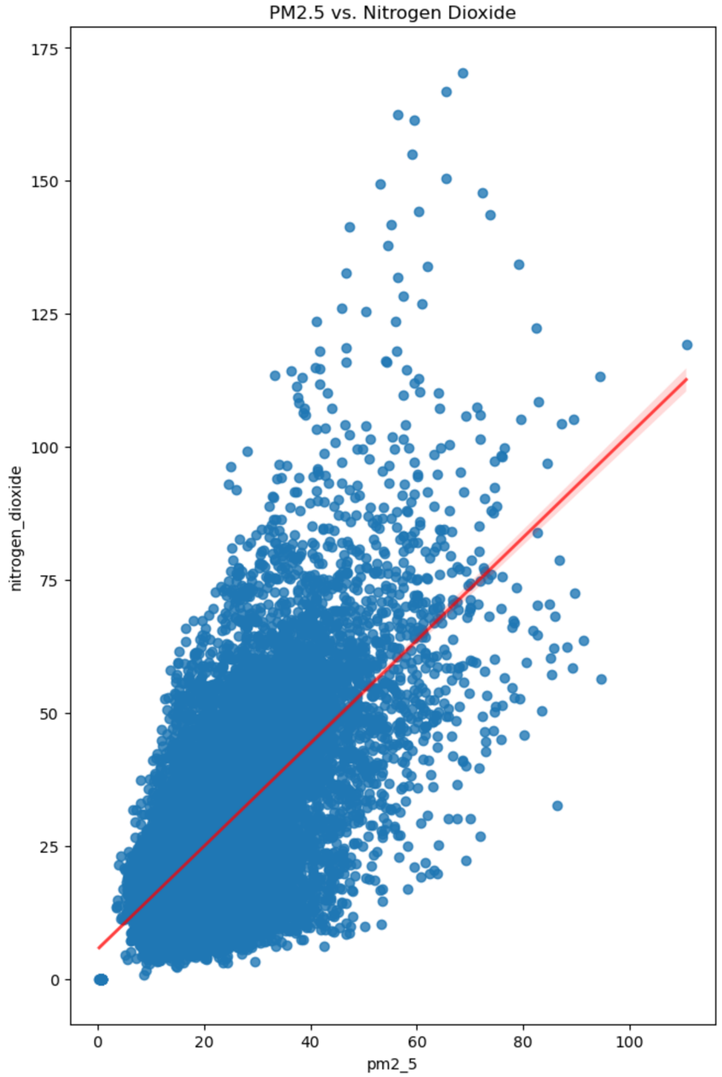


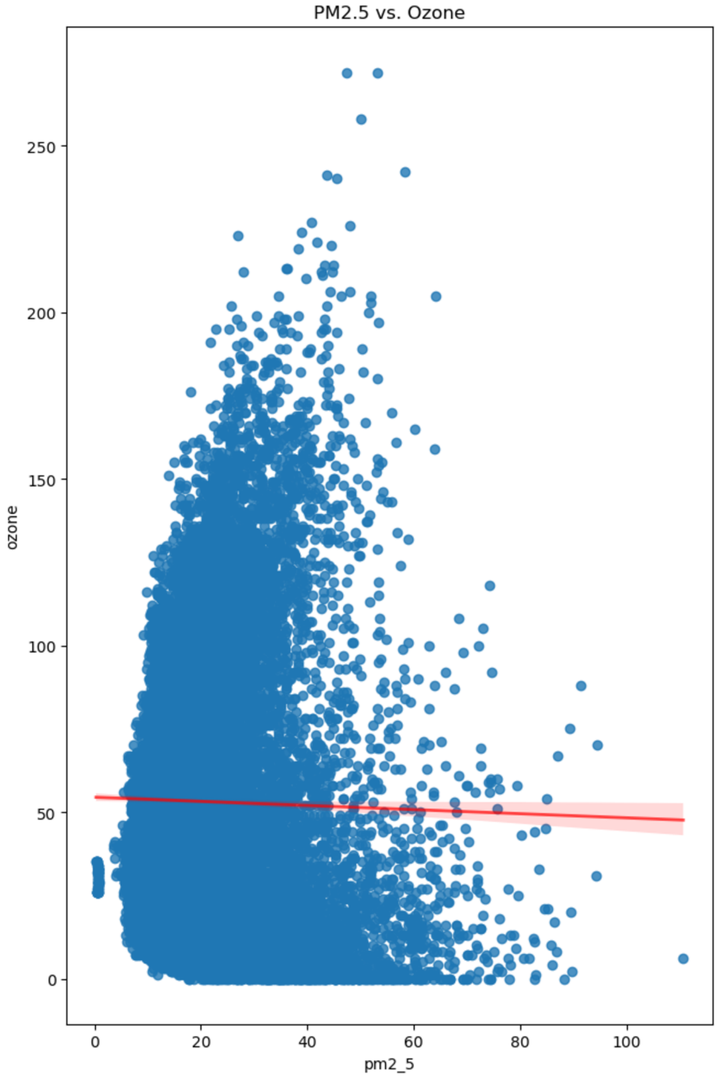
CapyData

Relationship

PM2.5 vs CO2, NO2, Ozone



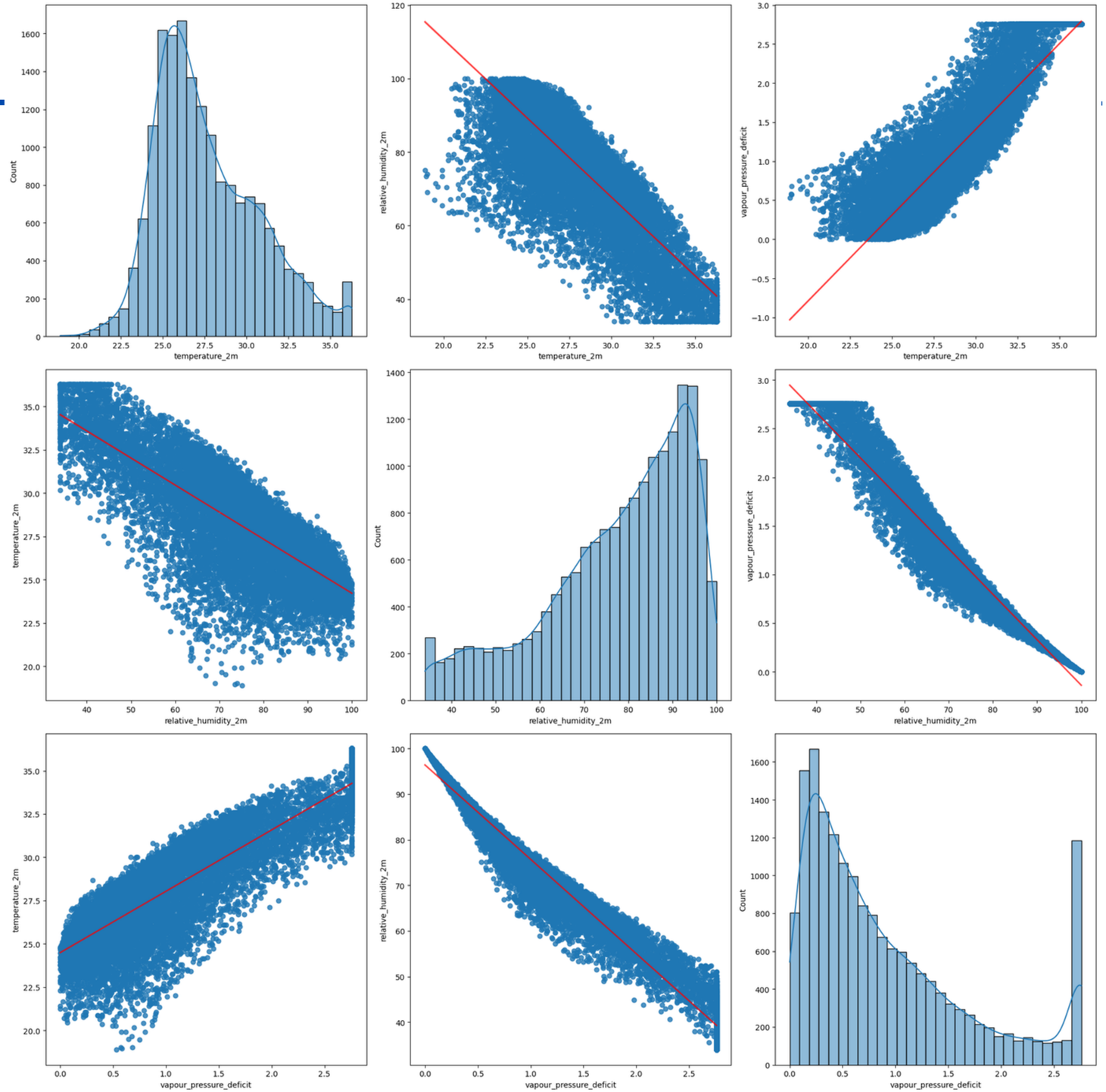




CapyData

Relationship

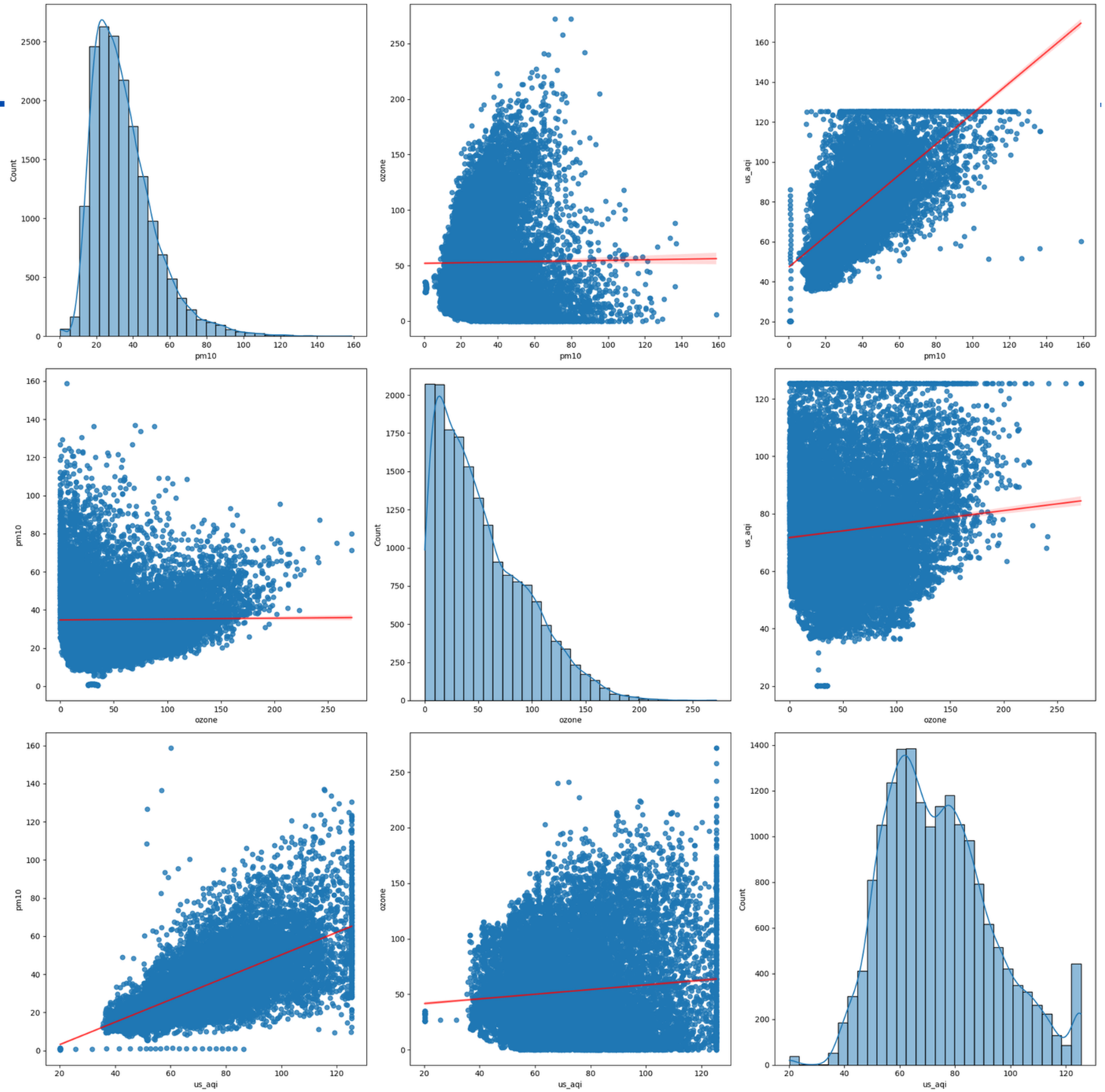
Temperature vs Humidity vs Vapour pressure deficit



CapyData

Relationship

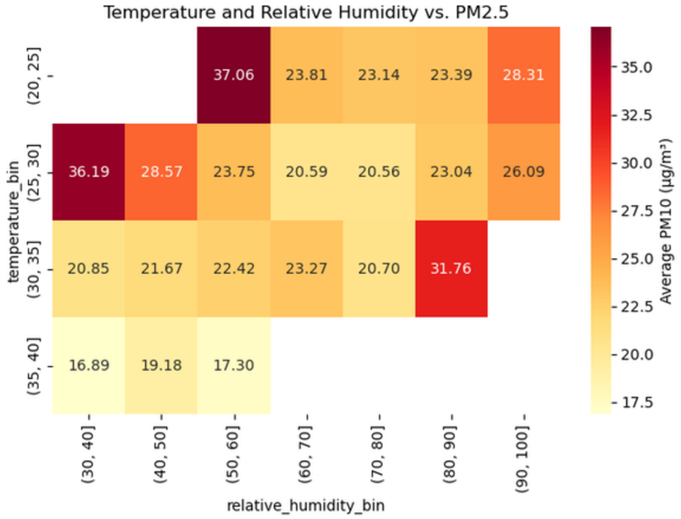
PM10 vs Ozone vs US AQI

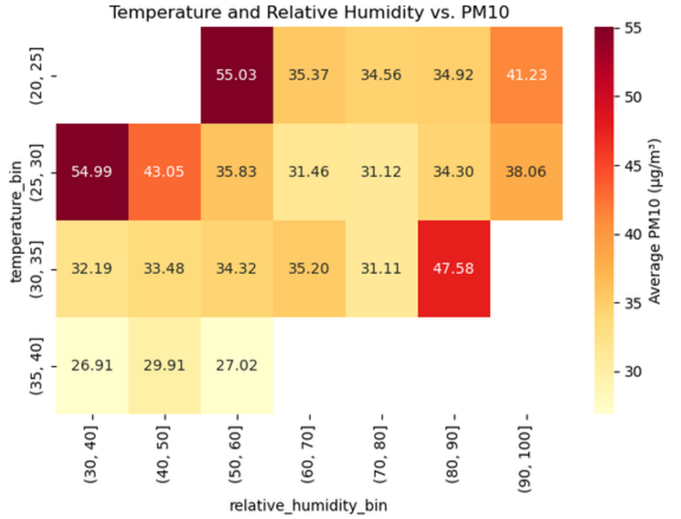


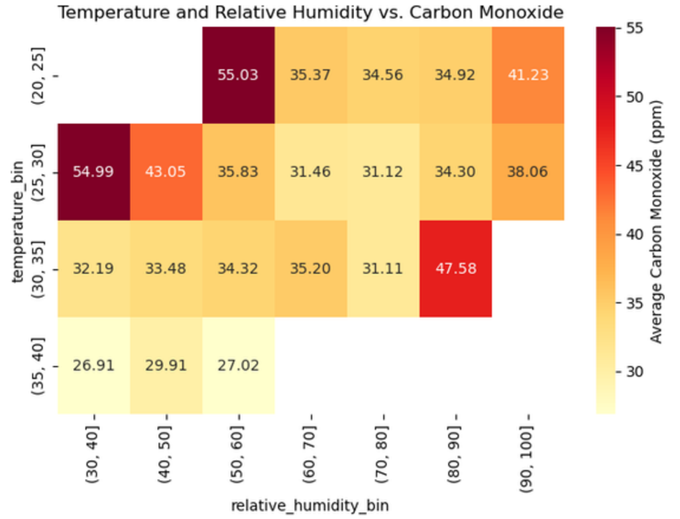
CapyData

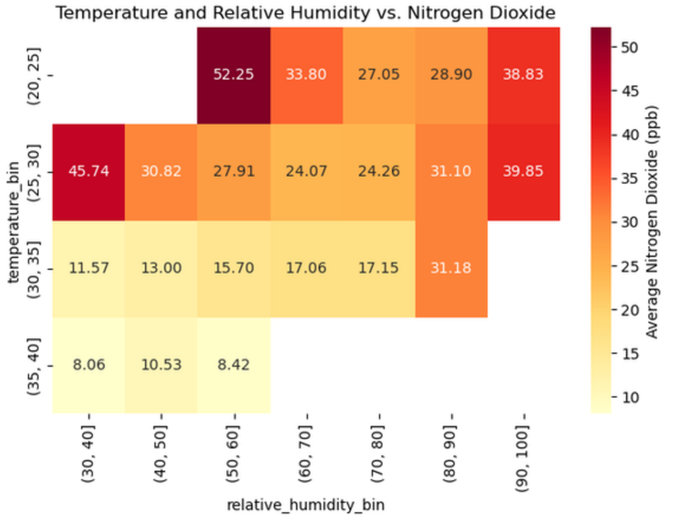
Relationship

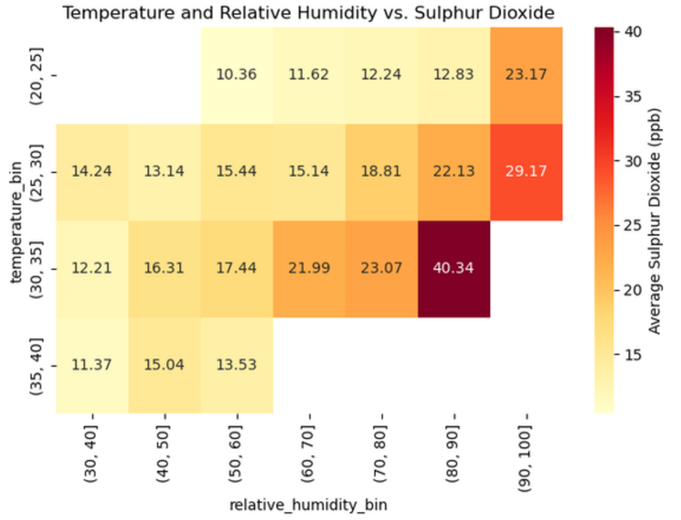
Temperature and Humidity vs Air Parameters

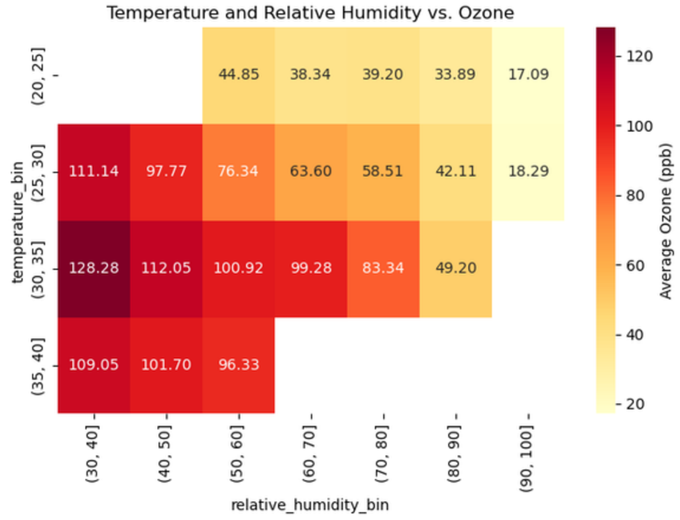






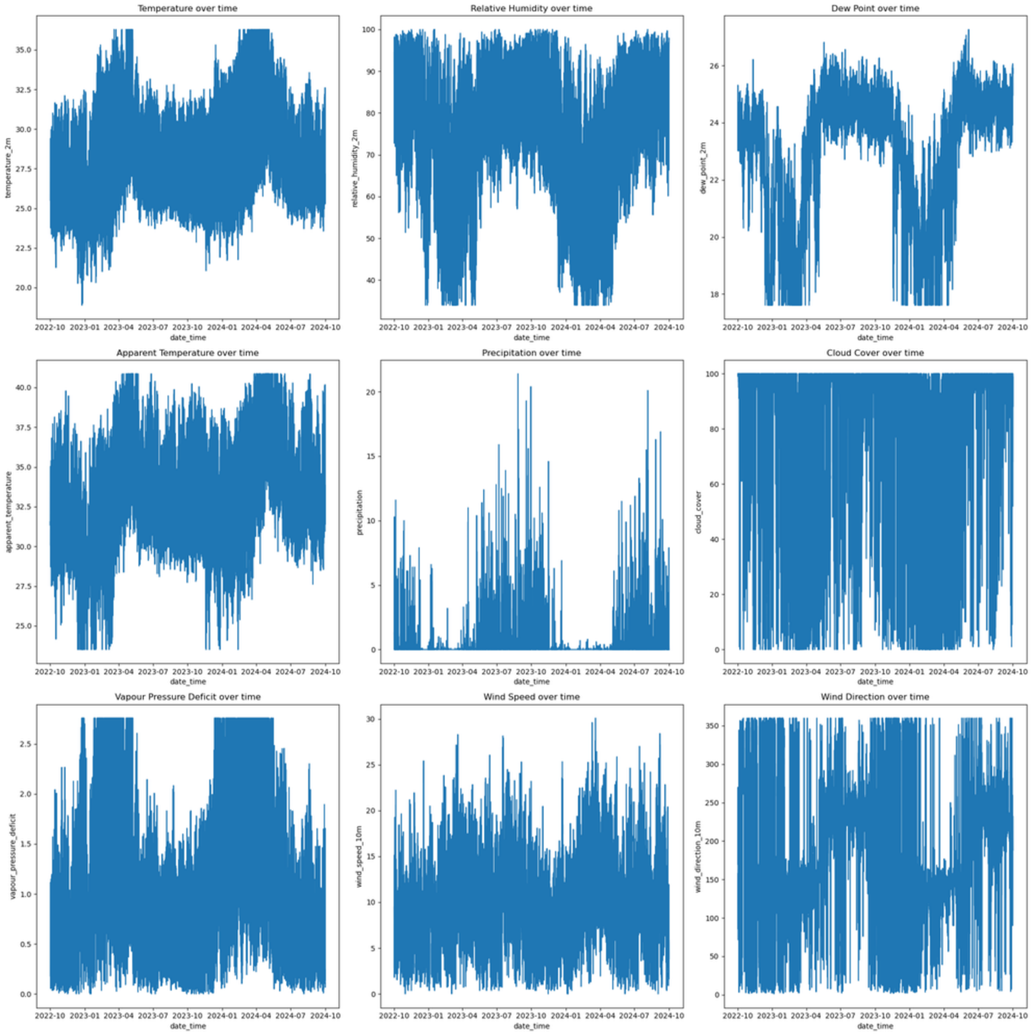




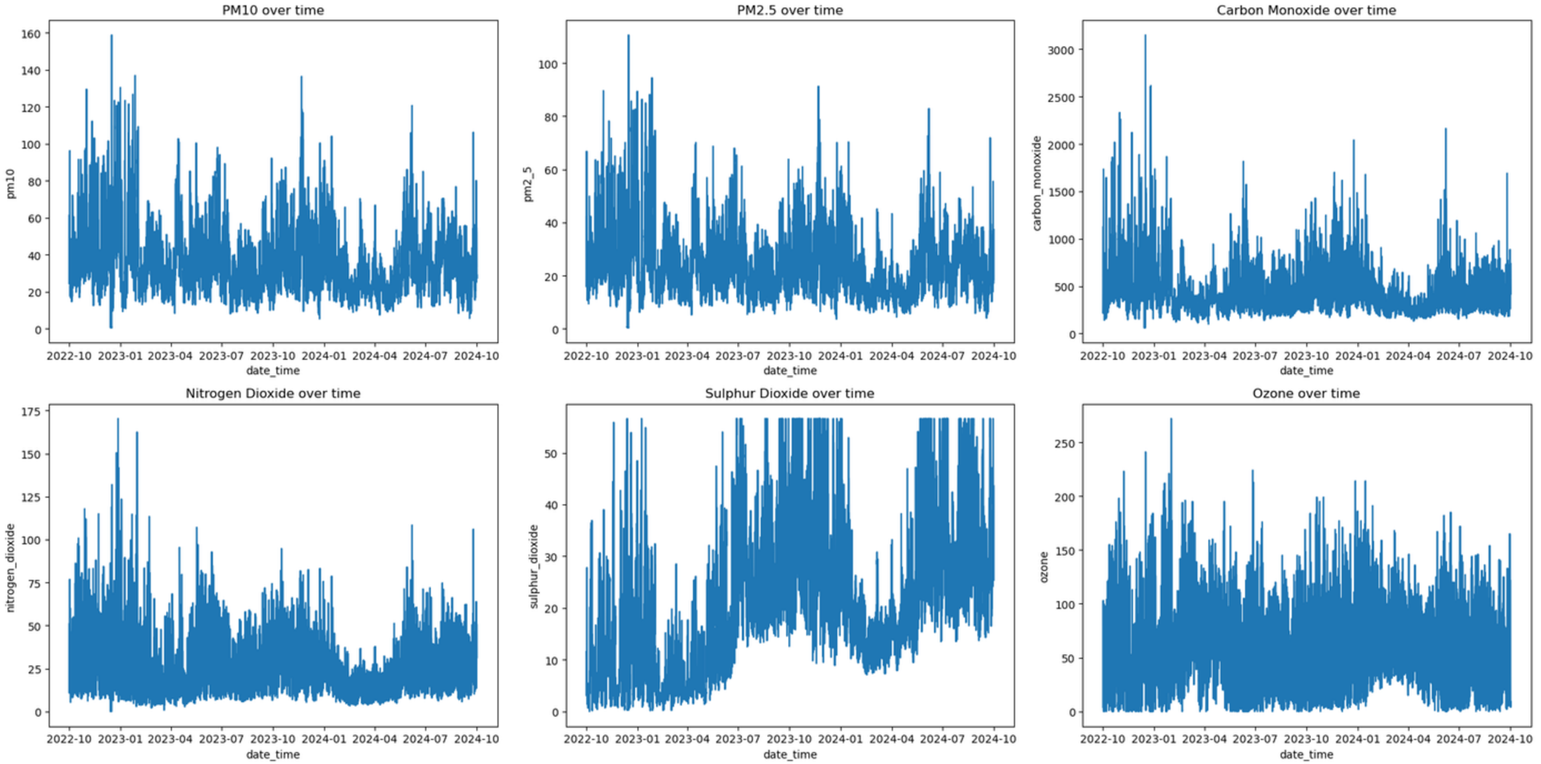


CapyData

Line plot



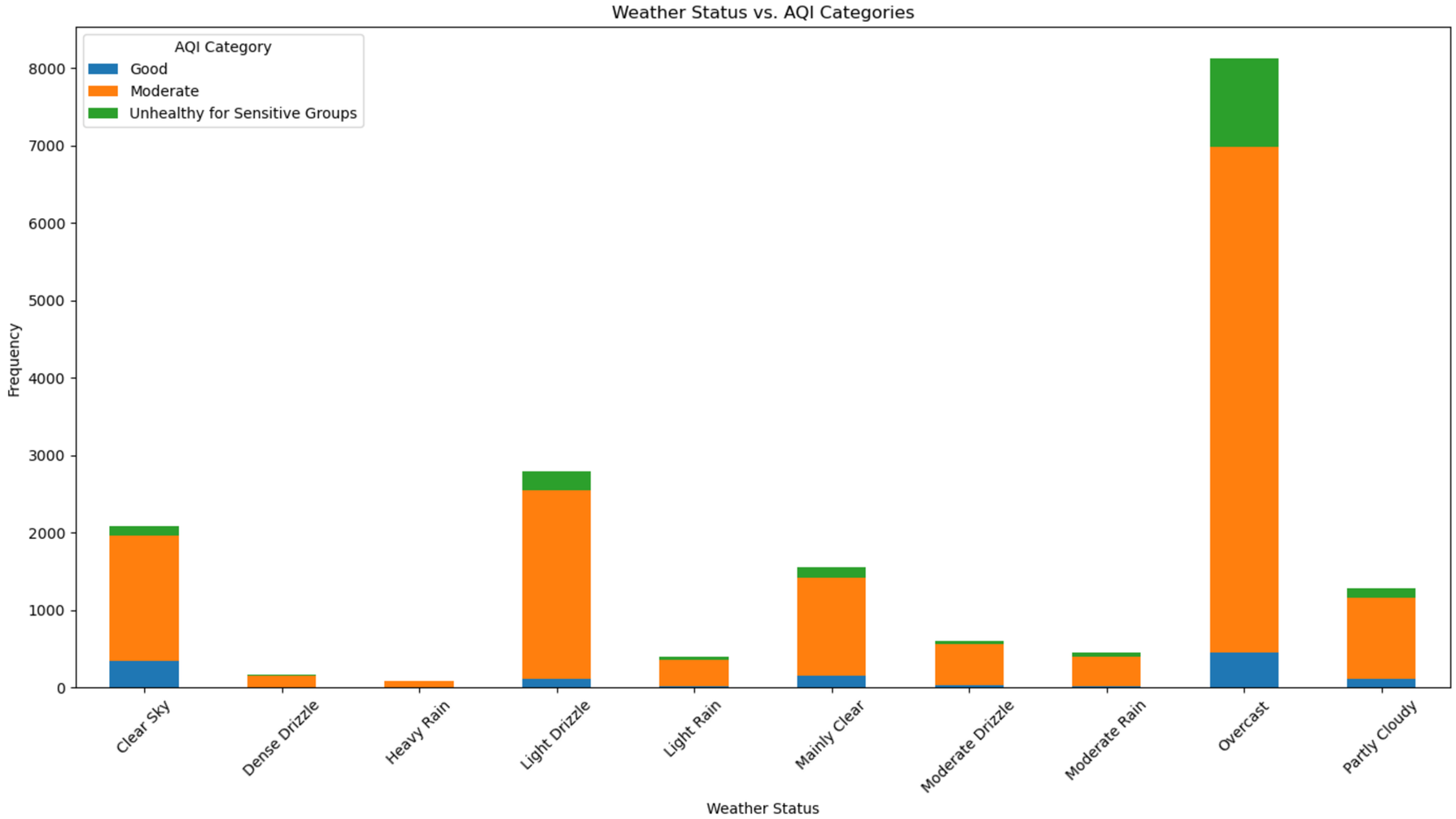
Weather and Air Parameters over time



CapyData

Categories

Weather & Air Quality categories



CapyData

Questioning Method

Base on the outcomes from EDA and real-life problem, we frame 5 questions:

- Is there a correlation between wind speed/direction and PM10 levels? Does wind from certain directions bring higher pollution levels?
- Are there distinct seasonal or monthly patterns in air quality metrics?
- What is the relationship between precipitation and air quality? Does rainfall help reduce pollutant concentrations, and if so, to what extent?
- Are there specific times of day (morning, afternoon, evening) when pollution levels tend to be higher?
- Does a significant increase or decrease in temperature impact pollutant levels such as NO2 and ozone?

=> To answer these, we're going to víualize all relevant data, then give comment and conclusion from the outcomes

Weather status classification

Approach:

Train Decision Tree and Random Forest
Train on full features
Train on top 10 best features
Use SMOTE to rebalance the dataset
Hyperparameter Tuning Random Forest with GridSearchCV

Outcome: The models have almost perfect prediction with 100% accuracy

CapyData

Air US_AQI Prediction:

Approach:

- ARIMA
- Random Forest
- XGBoost

Outcome: The best model is XGBoost with MAE = 13.4

OUTCOMES

VISIT OUR WEBSITE RIGHT NOW

THANK YOU

We sincerely appreciate your interest and willingness to support.
Thank you once again, and have a wonderful day!