



Ho Chi Minh City Weather & Air Quality Analysis



- › **Intro to Data Science**

PROJECT MANAGEMENT

TEAM MEMBERS

CapyData TEAM MEMBERS



Cao Hoang Loc
22127012

Data Pre-processing
Data Modeling - Classification

Phan Vo Minh Tue
22127255

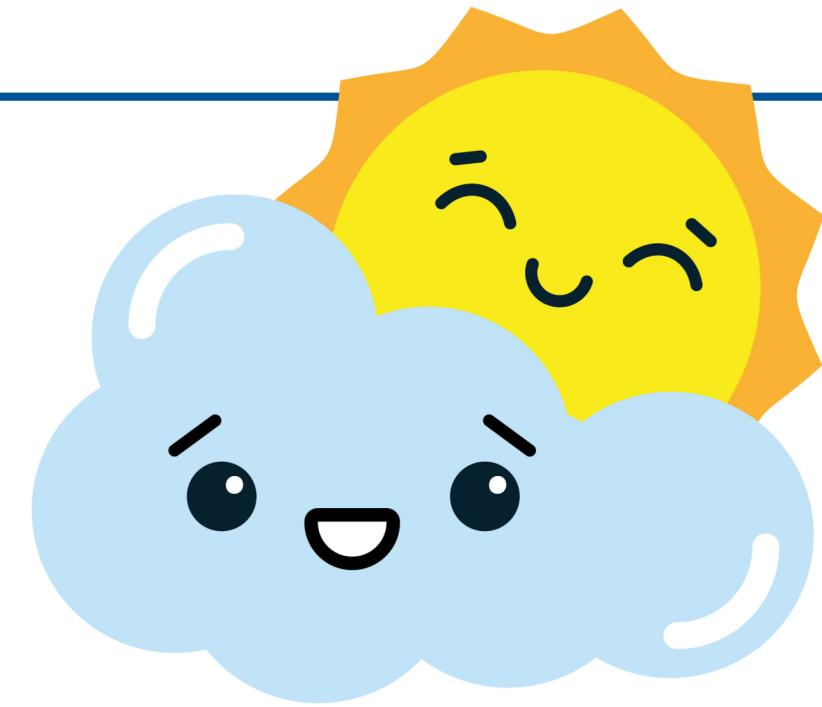
Data Collecting
Exploring data Analysis

Vo Nguyen Phuong Quynh
22127255

Team leader
Exploring data Analysis

Pham Anh Van
21127039

Data Pre-processing
Data Modeling - Predict



TOPIC



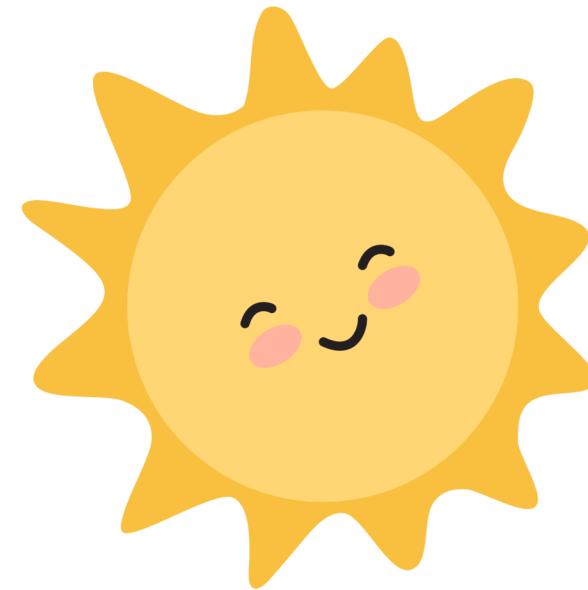
HCMC Weather & Air Quality Analysis

Ho Chi Minh City is dealing with worsening air quality due to its tropical monsoon climate. Our project will examine the connection between weather (temperature, humidity, rainfall) and air pollution (particulate matter) from October 2022 to September 2024. By identifying seasonal trends and pollution sources, we aim to support better environmental policies and public health in the city.



DATA COLLECTING

Completed



➤ Data source

Weather and air quality data for Ho Chi Minh City (latitude 10.823099, longitude 106.629664) were collected using the **Open-Meteo API**. Data was collected into 2 datasets (Weather & AQI), spans *October 1, 2022*, to *September 30, 2024*, providing hourly data points with **17,544 records** each dataset.

➤ hcmc_weather_data.csv

10 features:

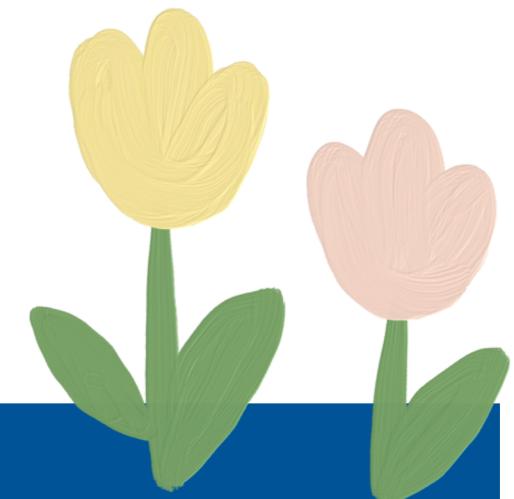
- Temperature and humidity at 2m
- Dew point and apparent temperature
- Precipitation and cloud cover
- Wind speed and direction at 10m
- Weather condition codes
- Vapour pressure deficit



➤ 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

Completed



Description

The merged dataset combining weather and air quality data has been fully processed to ensure data quality and readiness for analysis. The Data Combination Process:

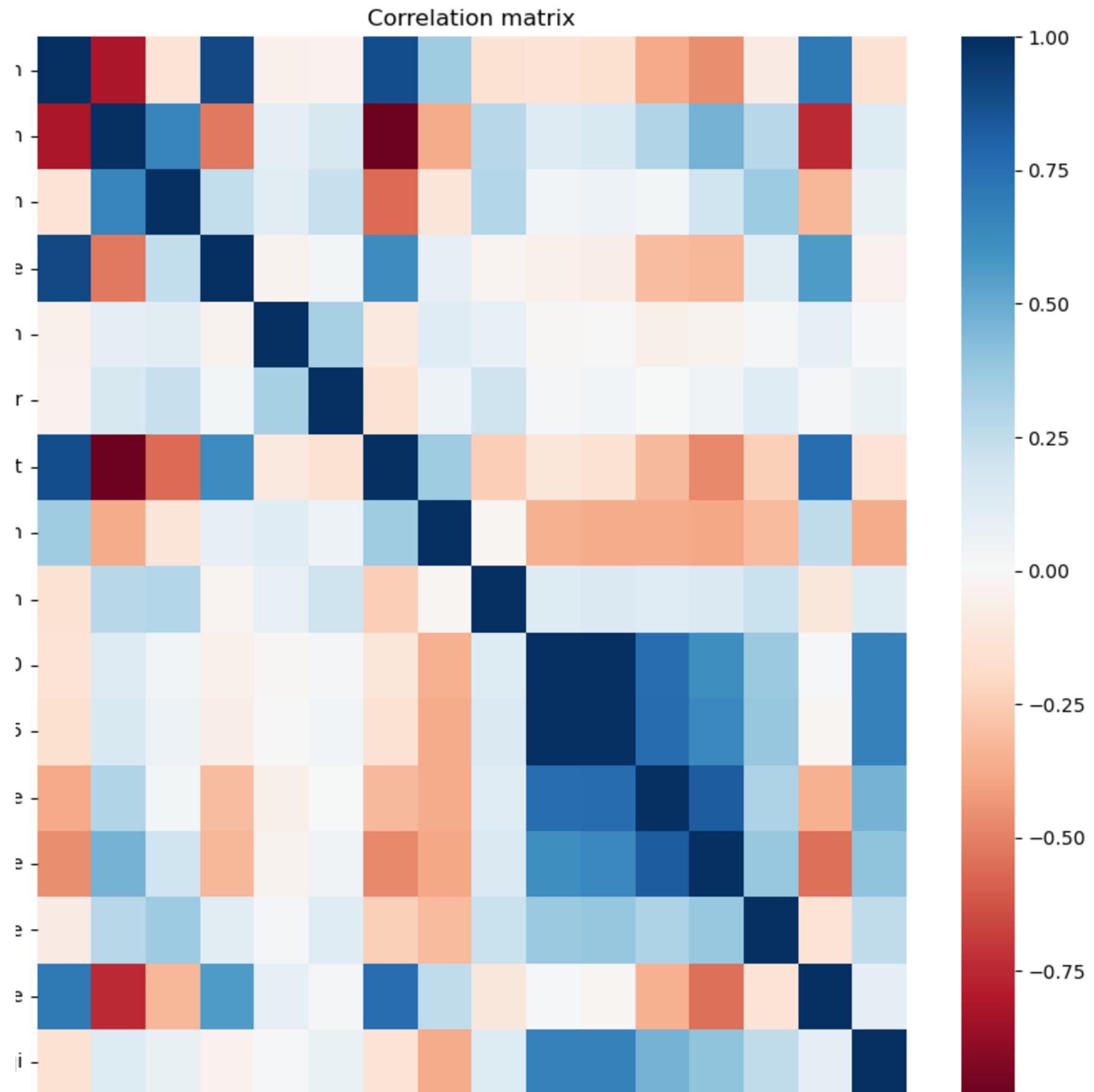
- Handling missing values
- Mapping weather_code to weather_status
- Validation of values
- Time continuity
- Outlier detection and handling
- Relationship validity
- Time zone adjustment

clean_hcmc_waq.csv

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

EDA

Completed



Correlations between features

- 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.

And 6 more question ...



EDA Question 1



Is there a correlation between wind speed/direction and PM10 levels?
Does wind from certain directions bring higher pollution levels?

› Correlation



Higher wind speeds (to the right on the x-axis) generally correlate with lower average PM10 concentrations. This trend suggests that as wind speed increases, it disperses particulate matter, lowering pollution concentrations in the area.

› Directions bring higher



Winds from the West and Northwest bring the highest PM10 pollution levels. Winds from the East and Southeast are associated with cleaner air, possibly due to fewer pollution sources or more effective pollutant dispersion in those directions.

EDA Question 2

How do extreme weather events affect air quality parameters, and what are the lag effects on pollutant concentrations?



➤ Effects

Extreme weather events have varying impacts on different pollutants, with SO_2 and O_3 being most significantly affected.

➤ Lag effects

The lag effects persist for considerable periods (1000-4000 hours), suggesting the need for extended monitoring and management strategies post-extreme weather events.



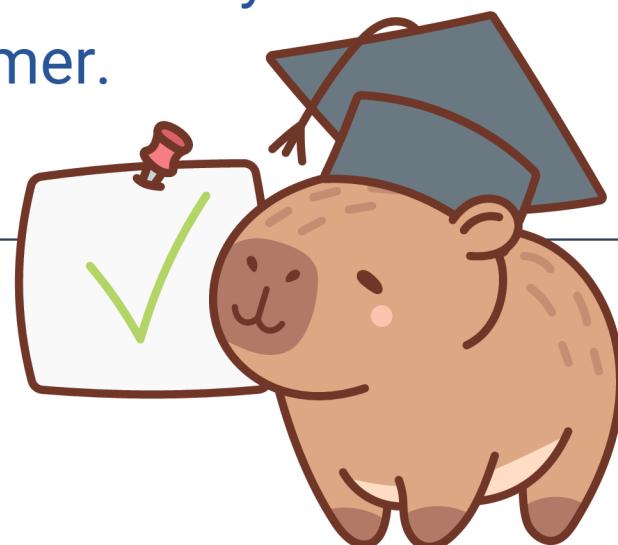
EDA Question 3

Are there distinct seasonal or monthly patterns in air quality metrics?



› Carbon Monoxide (CO)

This pollutant stands out with significantly higher concentrations than the others, fluctuating around 300–700 µg/m³. It shows a noticeable seasonal trend, with peaks around the last months of the year and lower concentrations in early summer.



› Other Pollutants

The rest of the pollutants (PM10, PM2.5, NO₂, SO₂, O₃, and US AQI) have relatively low concentrations compared to CO, all staying below 100 µg/m³ and stable throughout the year.



EDA Question 4

What is the relationship between precipitation and air quality?
Does rainfall help reduce pollutant concentrations, and if so, to what extent?



➤ Relationships

Rainfall seems to contribute slightly to the reduction of particulate pollutants (PM10 and PM2.5), but the effect is relatively minor and not strong enough to significantly impact the overall air quality (as reflected by AQI).

➤ Rain help reduce

This finding implies that while rain can help reduce pollutant levels, it may not be sufficient to substantially improve air quality on its own.



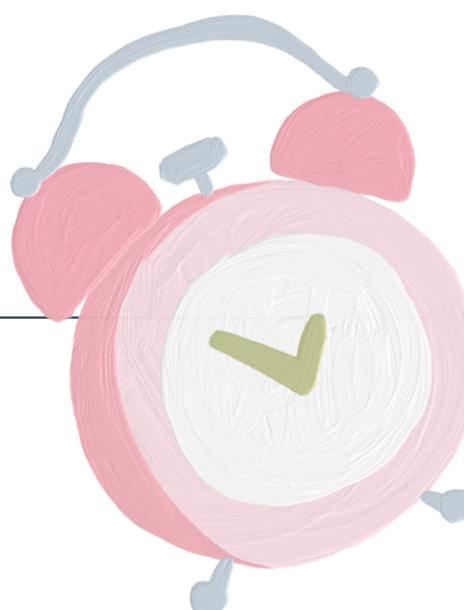
EDA Question 5

Are there specific times of day (morning, afternoon, evening) when pollution levels tend to be higher?



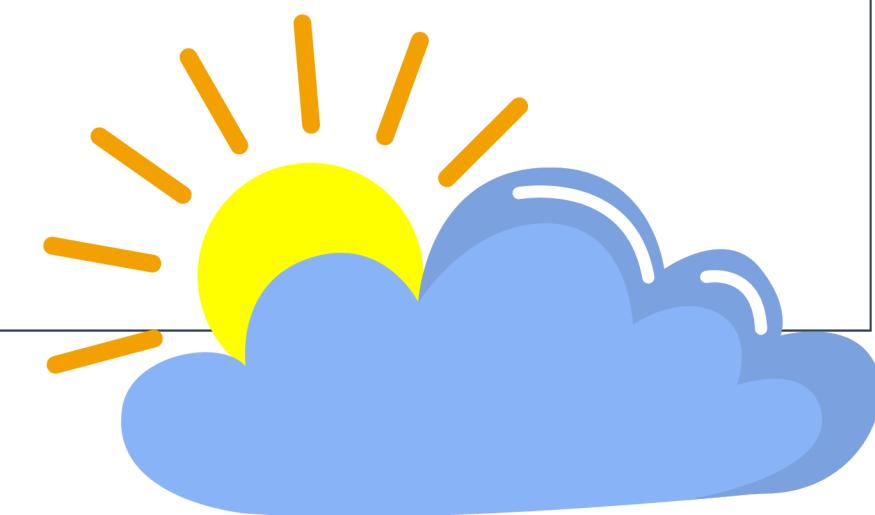
› Ozone

Ozone levels peak in the morning, likely due to favorable conditions for ozone formation, such as sunlight availability and less cloud cover



› PM10 & PM2.5

PM10 and PM2.5 are consistent with slightly higher concentrations in the afternoon.



EDA Question 6

How do temperature patterns vary across different time periods (daily, monthly)?
Are there significant anomalies in temperature trends?



➤ Temperature patterns

While actual temperatures show moderate fluctuations, apparent temperatures can exceed actual temperatures by up to 10°C during peak 4-5-6 months, indicating significant heat stress conditions.



➤ Anomalies

The presence of strong daily cycles and monthly patterns, combined with notable anomalies in early 2024, suggests a changing temperature regime that could have important implications for urban planning and public health considerations.

DATA MODELING

On-going



➤ **Weather status classification**

The goal is to develop a classification model with the accuracy of at least 98% to predict weather status.

From the data, we have 2 models:

➤ **Air US_AQI prediction**

The goal is to predict US-AQI (Air Quality Index).
The current aim is to evaluate the compatibility and performance of several popular time series prediction models with the dataset.



Air US_AQI Prediction:

Approach:

ARIMA

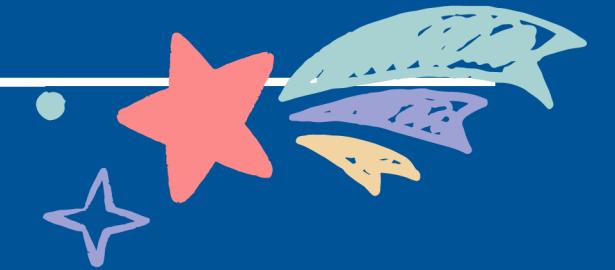
SARIMA

Random Forest Regressor

XGBoost Regressor

LSTM

Outcome: The models tested did not achieve the desired prediction accuracy, as the error rates remained high and did not improve with increasing model complexity.



Weather status classification

Approach:

Train DecisionTreeClassifier and RandomForestClassifier on full features

Train DecisionTreeClassifier and RandomForestClassifier on top 10 best features

Train following approach 2 but on balanced dataset

RandomForest with Hyperparameter Tuning

Outcome: The models tested did not achieve the desired prediction accuracy, as the error rates remained high and did not improve with increasing model complexity.



ISSUES



Challenging Question

How can we identify high-risk periods for air pollution based on the combination of weather and pollution factors?

Analogous Questions

- Is there a correlation between wind speed and direction and PM2.5 or PM10 levels?
- How do changes in wind direction and speed correlate with variations in pollutant levels across different areas of the city?

Data Complexity in Visualization

The dataset has a high number of rows and columns, making it visually overwhelming when attempting to interpret trends or key points.

IDEAS TO SOLVE

Resolving Redundant Analysis Questions

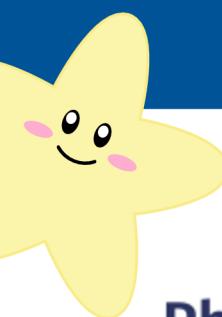
- Combine Q2 and Q6 into a single comprehensive inquiry: "Is there a correlation between wind speed/direction and PM2.5 or PM10 levels?"
- Structured Analysis Approach.

Managing Data Visualization Complexity

- Filtering Techniques
- Enhanced Visualization Methods: create hierarchical dashboards



PLAN FOR REMAINING TIME



Phase	14/10 - 27/10	28/10 - 10/11	11/11 - 24/11	25/11 - 8/12
Project Planning and Setup				
Data Collection and Processing				
Data Analysis and Modeling				
Prepare for midterm presentation				
Research & Question Answering				
Review & Report				
Prepare for final presentation				



THANK YOU

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

