

# PHÂN TÍCH DỮ LIỆU THÔNG MINH

ỨNG DỤNG CÔNG CỤ AI TRONG  
QUY TRÌNH PHÂN TÍCH DỮ LIỆU THỰC TẾ

Phân tích thời tiết tại thành phố Hồ Chí Minh

**Nhóm QuQuBeTA**



01

## Giới thiệu



# Nội dung

01

Giới thiệu

03

Kết quả và đề xuất

02

Quy trình  
phân tích dữ liệu

04

Ứng dụng hỗ trợ phân  
tích dữ liệu



## Thành viên nhóm

22120029 - Nguyễn Hữu Bền

22120300 - Nguyễn Tiến Quốc

22120301 - Nguyễn Trung Quốc

22120324 - Võ Thành Tâm



# 1. Giới thiệu



## Mục tiêu đề án

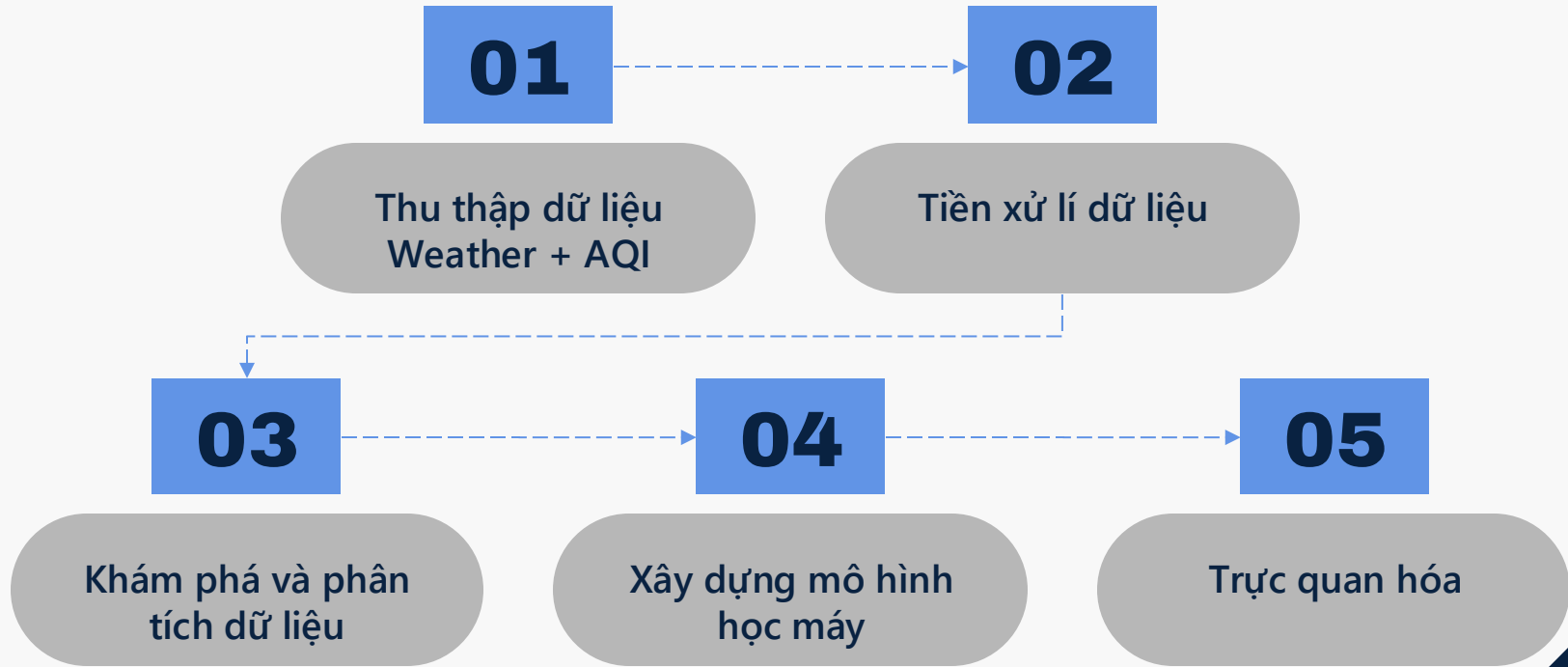
- Xây dựng pipeline phân tích dữ liệu thời tiết hoàn chỉnh.
- Vận dụng các công cụ AI (ChatGPT, Grok, Deepseek) để hỗ trợ các bước phân tích dữ liệu về thời tiết tại Thành Phố Hồ Chí Minh.
- Làm việc nhóm.

02

## Quy trình phân tích dữ liệu



## 2. Quy trình phân tích dữ liệu





## 2. Quy trình phân tích dữ liệu

Nhóm quyết định sử dụng 2 mô hình ngôn ngữ lớn (bản có phí và bản miễn phí):

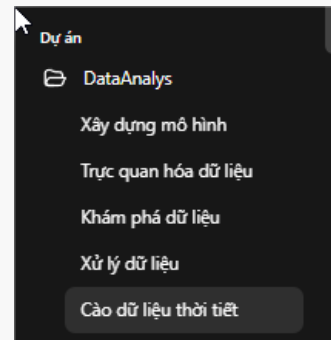
- ChatGPT Plus của Open AI.
- Grok của xAI.





## 2. Quy trình phân tích dữ liệu

- Đối với ChatGPT Plus, nhóm thực hiện đưa ra các prompt cho 5 khung chat trong 1 project ( ứng với 5 công đoạn chính trong quá trình phân tích dữ liệu).
- Ở từng khung chat, ta lần lượt thực hiện prompt theo các công thức: T-A-G (Task, Action, Goal: mục tiêu)



cho tôi các trang web mà tôi có thể cào dữ liệu về thời tiết ở tp Hồ Chí Minh 10 năm trở lại đây, mỗi dòng dữ liệu là 1 ngày



## 2. Quy trình phân tích dữ liệu

Ví dụ:

- Thu thập dữ liệu: prompt đưa ra nhằm tìm kiếm các trang nguồn có thể phục vụ cho công đoạn cào dữ liệu thời tiết: "Cho tôi các trang web mà tôi có thể cào dữ liệu về thời tiết ở tp Hồ Chí Minh 10 năm trở lại đây, mỗi dòng dữ liệu là 1 ngày".

cho tôi các trang web mà tôi có thể cào dữ liệu về thời tiết ở tp Hồ Chí Minh 10 năm trở lại đây, mỗi dòng dữ liệu là 1 ngày

Dưới đây là danh sách các trang web cho phép truy cập hoặc cào dữ liệu thời tiết tại TP. Hồ Chí Minh trong 10 năm qua, với mỗi dòng tương ứng 1 ngày, phù hợp để bạn trích xuất dạng CSV:

- 1. Time & Date
  - URL: <https://www.timeanddate.com/weather/vietnam/ho-chi-minh/historic>
  - Miễn phí, không cần tài khoản
  - Có dữ liệu theo ngày, giờ, nhiệt độ, độ ẩm, gió, v.v.
  - Có thể dùng Python + `requests` + `BeautifulSoup` để cào
  - Không có PM2.5 hay AQI
- 2. Visual Crossing Weather
  - URL: <https://www.visualcrossing.com/weather-data>
  - Có API REST miễn phí (đăng ký tài khoản)
  - Truy xuất được dữ liệu lịch sử theo ngày, đến hơn 10 năm
  - Có nhiệt độ, độ ẩm, gió, mưa, chỉ số UV, v.v.
  - Dữ liệu chất lượng cao hơn cần gọi trả phí
- 3. Open-Meteo API
  - URL: <https://open-meteo.com/en/docs/historical-weather-api>
  - Miễn phí, không cần token
  - Có thể lấy dữ liệu theo từng ngày (daily), từng giờ (hourly)
  - Có API JSON, dễ dùng với Python
  - Không có PM2.5 hoặc AQI
- 4. IQAir (AirVisual)

có thể cào thêm được các chỉ số AQI càng tốt, như chỉ số O<sub>3</sub>, bụi mịn

Tuyệt vời. Để cào được các chỉ số AQI, bụi mịn (PM2.5, PM10), khí độc (NO<sub>2</sub>, O<sub>3</sub>, CO, SO<sub>2</sub>) của TP.HCM trong 10 năm trở lại đây, bạn có thể dùng 2 nguồn sau:

- 1. [OpenAQ API (miễn phí & public)]
  - Website: <https://openaq.org>
  - Dữ liệu PM2.5, PM10, CO, NO<sub>2</sub>, SO<sub>2</sub>, O<sub>3</sub>
  - Có dữ liệu từ 2017 trở lại đây cho TP.HCM
  - Truy xuất được theo từng ngày hoặc giờ
- Ví dụ API:

```
bash
https://api.openaq.org/v2/measurements?city=HoChiMinh&parameter=pm25&date_from=2017-01-01&date_to=2027-01-01
```
- 2. [IQAir / AirVisual (dữ liệu chất lượng cao, HTML cào được)]
  - <https://www.iqair.com/vietnam/ho-chi-minh-city>
  - Dữ liệu đẹp, có biểu đồ AQI lịch sử (PM2.5, O<sub>3</sub>, CO, NO<sub>2</sub>, SO<sub>2</sub> ...)
  - Nhưng không có API miễn phí
  - Cào được bằng Selenium (web động, cần mô phỏng trình duyệt)
- Gợi ý cách làm:
  - Dùng API OpenAQ để lấy dữ liệu bụi mịn và khí độc từ 2017 → nay
  - Nếu cần dữ liệu cũ hơn (2014–2016), dùng Selenium + IQAir để cào HTML biểu đồ (hơi chậm)

## 2. Quy trình phân tích dữ liệu

Ví dụ Thu thập dữ liệu::

- ["https://www.timeanddate.com/weather/vietnam/ho-chi-minh/historic?month=5&year=2018"](https://www.timeanddate.com/weather/vietnam/ho-chi-minh/historic?month=5&year=2018) Cào dữ liệu trong này để lấy các thuộc tính thời tiết với mỗi dòng dữ liệu là một giờ (có thể thử trước với chỉ ngày trong ảnh".
- "Dùng request để lấy dữ liệu từ bảng sau (ảnh đính kèm). Chỉ nên thử với link này trước, sau muốn mở rộng bộ dữ liệu thì sẽ lặp để thay đổi tháng và năm trên link. Lấy ngày ở ảnh 2, giờ thì lấy bắt đầu từ 00h30 (ảnh 3) để khỏi có thêm thông tin ngày bên dưới. còn lại thông tin các chỉ số thời tiết thì lấy như bình thường".

2. Code mẫu lấy dữ liệu từ trang web:

```
python
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.common.by import By
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.service import Service

def scrape_weather(date="20180501"):
    url = f"https://www.timeanddate.com/weather/vietnam/ho-chi-minh/historic?m={date}"

    options = webdriver.ChromeOptions()
    options.add_argument('--headless') # chạy ngầm không hiển thị cửa sổ browser
    driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=options)

    driver.get(url)
    time.sleep(3) # chờ load trang

    rows = driver.find_elements(By.CSS_SELECTOR, "table#wat-his tbody tr")

    data = []

    current_date = pd.to_datetime(date, format='%Y%m%d').date()

    for row in rows:
        columns = row.find_elements(By.TAG_NAME, "td")
        if len(columns) == 0:
            continue

        time_of_day = row.find_element(By.TAG_NAME, "th").text.strip()

        # Bỏ qua khung giờ 00:00 như yêu cầu của bạn
        if time_of_day == "00:00":
            continue

        # Lấy dữ liệu các cột tương ứng
        temp = columns[1].text.replace("\u00b0C", "").strip()
        weather = columns[2].text.strip()
        wind = columns[3].text.strip()
        humidity = columns[4].text.strip()
        barometer = columns[5].text.strip().replace(" mbar", "")
```

## 2. Quy trình phân tích dữ liệu

Ví dụ Phân tích dữ liệu và trực quan hóa dữ liệu:  
Nêu rõ yêu cầu mong muốn AI thực hiện

Dựa trên dữ liệu đã tải lên, hãy thực hiện các trực quan hóa sau:

- Vẽ biểu đồ đường cho sự thay đổi AQI theo thời gian.
- Vẽ scatter plot để kiểm tra mối quan hệ giữa các đặc trưng như temperature, humidity, wind\_speed, pressure và AQI.
- Vẽ biểu đồ boxplot để kiểm tra phân phối của các đặc trưng temperature, humidity, wind\_speed, pressure, và AQI.
- Vẽ seasonal decomposition cho AQI nếu có xu hướng theo mùa. Sau mỗi biểu đồ, vui lòng đưa ra nhận xét về các đặc điểm nổi bật từ kết quả trực quan.

Có vẻ như cột `timestamp` không tồn tại do đã được xử lý trước đó (có thể đã được chuyển sang index hoặc đổi tên). Để giải quyết vấn đề này, tôi sẽ kiểm tra lại cấu trúc dữ liệu và thực hiện trực quan hóa lại cho bạn. [-]

Dữ liệu hiện tại không có cột `timestamp`, mà thay vào đó có cột `datetime`. Tôi sẽ tiến hành cập nhật lại đoạn code EDA dựa trên cột này và thực hiện các bước trực quan hóa cho bạn. [-]



**Kết quả và nhận xét từ các biểu đồ trực quan hóa:**

### 1. Biểu đồ đường (Line plot) AQI theo thời gian:

- **Nhận xét:**
  - Chỉ số AQI biến động mạnh theo thời gian.
  - Có các đợt tăng hoặc giảm AQI đáng chú ý theo từng thời điểm cụ thể.
  - AQI không ổn định, có khả năng chịu ảnh hưởng mạnh từ các điều kiện thời tiết và ô nhiễm.

### 2. Scatter plot giữa AQI và các yếu tố thời tiết:

- **Nhận xét:**
  - Temperature (temp):** Mối quan hệ yếu với AQI, biểu hiện không rõ ràng.
  - Humidity:** Phân bố rộng, nhưng ít thấy tương quan mạnh và rõ ràng với AQI.
  - Wind:** Có xu hướng nhẹ khi tốc độ gió thấp thì AQI cao hơn, nhưng tương quan không mạnh.
  - Pressure:** Không thấy mối tương quan rõ ràng với AQI.

Các yếu tố thời tiết thể hiện tương quan khá thấp với AQI. Điều này gợi ý rằng AQI bị ảnh hưởng mạnh hơn bởi các yếu tố ô nhiễm như PM2.5, PM10, SO2, NOx, CO...

## 2. Quy trình phân tích dữ liệu

### 3. Boxplot về phân phối các đặc trưng:

#### • Nhận xét:

- Temperature và Pressure có phân phối tương đối đối xứng, ít ngoại lệ rõ.
- Humidity và Wind cho thấy nhiều giá trị ngoại lệ hơn, cho thấy biến động này.
- AQI có độ biến thiên khá rõ, thể hiện qua khoảng giá trị rộng và một số r

### 4. Seasonal decomposition của AQI:

Quá trình phân rã mùa vụ thất bại do dữ liệu AQI hàng ngày vẫn còn chứa giá trị bị thực hiện phân tích này, cần xử lý tiếp tục dữ liệu thiếu:

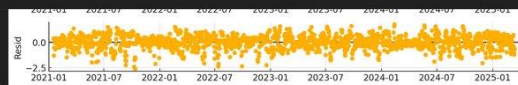
#### 🔧 Xử lý lại dữ liệu và Seasonal decomposition:

- Tôi sẽ điền giá trị thiếu bằng phương pháp nội suy tuyến tính (interpolation decomposition). Để tôi tiến hành tiếp bước này giúp bạn. [↩]

#### Matplotlib Chart



Ask anything



#### 📊 Nhận xét từ biểu đồ Seasonal Decomposition sau xử lý dữ liệu:

Biểu đồ phân rã chuỗi thời gian AQI gồm các thành phần chính:

##### 1. Xu hướng (Trend):

- AQI thể hiện xu hướng biến động theo từng giai đoạn thời gian, có thời điểm tăng mạnh và giảm rõ rệt.
- Xu hướng dài hạn này cần được phân tích sâu hơn để hiểu rõ nguyên nhân (ví dụ, thời tiết theo mùa hoặc tác động ô nhiễm từ con người).

##### 2. Mùa vụ (Seasonal):

- Có dấu hiệu rõ ràng về tính chu kỳ theo tháng (period = 30 ngày).
- Biên độ dao động mùa vụ rõ rệt, cho thấy AQI thường xuyên lặp lại theo từng tháng, có thể liên quan đến các yếu tố như điều kiện thời tiết định kỳ, lượng mưa, độ ẩm hoặc nhiệt độ.

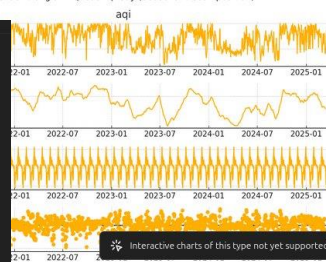
##### 3. Nhiễu (Residual):

- Phần residual (nhiều) khá lớn, thể hiện rằng ngoài xu hướng và tính chu kỳ còn có nhiều yếu tố ngẫu nhiên ảnh hưởng đến AQI.
- Điều này gợi ý rằng mô hình dự đoán cần chú ý đến các yếu tố nhiễu và xây dựng các đặc trưng bổ sung để giảm thiểu sai số.

decomposition. Để tôi tiến hành tiếp bước này giúp bạn. [↩]

#### Matplotlib Chart

Phân rã chuỗi thời gian AQI sau nội suy (Seasonal Decomposition)



#### 📊 Biểu đồ Seasonal Decomposition sau xử lý dữ liệu:

Biểu đồ phân rã chuỗi thời gian AQI gồm các thành phần chính:

Xu hướng biến động theo từng giai đoạn thời gian, có thời điểm tăng mạnh và giảm

n này cần được phân tích sâu hơn để hiểu rõ nguyên nhân (ví dụ, thời tiết theo mùa hoặc tác động ô nhiễm từ con người).

Biên độ dao động mùa vụ rõ rệt, cho thấy AQI thường xuyên lặp lại theo từng tháng (period = 30 ngày).

## 2. Quy trình phân tích dữ liệu

Ví dụ yêu cầu AI tự động phân tích rút ra kết luận từ những thông tin và biểu đồ.

### ✓ Kết luận và đề xuất:

- Dữ liệu AQI có tính chu kỳ và xu hướng rõ nét, do đó các mô hình chuỗi thời gian như ARIMA, SARIMA hoặc LSTM rất phù hợp để dự đoán AQI.
- Các yếu tố thời tiết (như nhiệt độ, độ ẩm, áp suất) có vai trò hạn chế trong việc dự đoán AQI, nên xem xét chúng như là các đặc trưng bổ sung chứ không phải là đặc trưng chính.
- Các yếu tố ô nhiễm (PM2.5, PM10, CO, NO, SO2, NH3) nên được ưu tiên hàng đầu trong mô hình.

Nếu bạn cần hỗ trợ tiếp theo để xây dựng và huấn luyện mô hình dự đoán AQI cụ thể, hãy cho tôi biết nhé!



Ask anything

### 📌 1. Nhận xét tổng quan về dữ liệu và mối quan hệ giữa AQI và đặc trưng thời tiết:

- Dữ liệu thu thập về AQI và các yếu tố thời tiết như nhiệt độ (temperature), độ ẩm (humidity), tốc độ gió (wind), và áp suất (pressure) khá đầy đủ nhưng vẫn có lượng đáng kể dữ liệu thiếu ở các đặc trưng thời tiết.
- Các đặc trưng thời tiết (`temp`, `humidity`, `wind`, `pressure`) có mức tương quan tương đối thấp với AQI. Đặc biệt:
  - Nhiệt độ (`temp`) và độ ẩm (`humidity`) không thể hiện mối quan hệ rõ ràng và mạnh mẽ với AQI.
  - Tốc độ gió (`wind`) có tương quan âm nhẹ với AQI, cho thấy AQI thường cao hơn khi tốc độ gió thấp (do khả năng phân tán các chất ô nhiễm kém hơn).
  - Áp suất (`pressure`) gần như không thể hiện mối tương quan rõ ràng với AQI.
- Trong khi đó, các yếu tố liên quan trực tiếp đến ô nhiễm như `PM2.5`, `PM10`, `CO`, `NO`, `SO2`, `NH3` lại có tương quan rất mạnh với AQI. Điều này hợp lý vì các chất này trực tiếp định nghĩa nên AQI.

### 📌 2. Những sự bất thường, xu hướng, và mô hình phát hiện được:

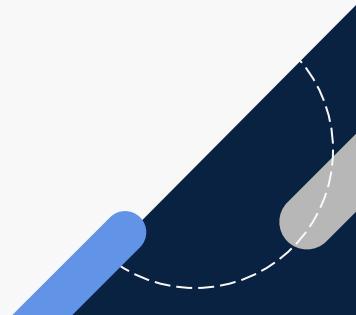
- Xu hướng AQI:**
  - AQI biến động mạnh mẽ theo thời gian. Có các giai đoạn AQI tăng đột biến, thể hiện những đợt ô nhiễm cao bất thường.
  - Xu hướng dài hạn cho thấy có các chu kỳ rõ ràng theo từng tháng, có thể do các yếu tố khí hậu theo mùa hoặc do tác động nhân tạo (giao thông, công nghiệp, cháy rừng...).
- Chu kỳ mùa vụ:**
  - Có dấu hiệu rất rõ về chu kỳ mùa vụ của AQI theo từng tháng. AQI thường đạt đỉnh vào một số thời điểm nhất định, có thể trùng với những thay đổi về thời tiết hoặc hoạt động con người (ví dụ: đốt rơm rạ vào mùa thu, cao điểm giao thông).
- Dữ liệu bất thường (ngoại lệ):**
  - Có một số điểm ngoại lệ rõ nét xuất hiện trong biểu đồ boxplot, đặc biệt ở các đặc trưng như độ ẩm, tốc độ gió. Các ngoại lệ này cần được xem xét kỹ để xác định xem là lỗi do đặc hay thực sự là những biến động tự nhiên.



## 2.1. Thu thập dữ liệu

Lựa chọn nguồn dữ liệu dựa trên các tiêu chí:

- Dữ liệu phải quan sát theo từng giờ
- Dữ liệu bao gồm: AQI (chỉ số chất lượng không khí), các yếu tố thời tiết như nhiệt độ, độ ẩm,...



## 2.1. Thu thập dữ liệu

Nhận lựa nhận thấy rằng dữ liệu chất lượng không khí và thời tiết không đi cùng nhau trong tập dữ liệu. Do đó quyết định tiến hành thu thập dữ liệu tại 2 nguồn khác nhau:

- Dữ liệu về AQI được thu thập bằng cách gọi API tại: <https://openweathermap.org/>
- Dữ liệu về các yếu tố thời tiết được thu thập tại <https://www.timeanddate.com/>

Ho Chi Minh Weather History for 1 January 2025

Show weather for: 1 January 2025

Time	Conditions	Temp	Weather	Wind	Humidity	Barometer	Visibility
00:00 Wed, 1 Jan		26 °C	Clear	11 km/h	70%	1011 mbar	16 km
00:30		26 °C	Passing clouds	11 km/h	65%	1011 mbar	N/A
01:00		26 °C	Passing clouds	9 km/h	70%	1010 mbar	N/A
01:30		25 °C	Clear	7 km/h	69%	1010 mbar	16 km
02:00		25 °C	Clear	9 km/h	69%	1010 mbar	16 km
02:30		25 °C	Clear	9 km/h	69%	1009 mbar	16 km
03:30		24 °C	Passing clouds	9 km/h	74%	1009 mbar	N/A
04:00		25 °C	Mostly cloudy	6 km/h	74%	1009 mbar	N/A
04:30		25 °C	Mostly cloudy	7 km/h	74%	1009 mbar	N/A
05:00		24 °C	Passing clouds	7 km/h	74%	1009 mbar	N/A
06:00		24 °C	Passing clouds	7 km/h	78%	1009 mbar	N/A

OpenWeather

Weather in your city

Guide API Dashboard Marketplace Pricing Maps

weather index maps

- Use as layers in Direct Tiles, OpenLayers, Leaflet, and Google Maps
- Please [contact us](#) to get a quote

Other weather API's collection

Air Pollution API

[API doc](#) [Subscribe](#)

- Current, forecast and historical air pollution data
- Forecast for 4 days ahead with 1-hour step
- Air Pollution API includes both Air Quality Index and indices for CO, NO, NO<sub>2</sub>, O<sub>3</sub>, SO<sub>2</sub>, NH<sub>3</sub>, PM<sub>2.5</sub>, PM<sub>10</sub>
- Included in both free and paid subscriptions

Fire Weather Index API

[API doc](#) [Get a quote](#)

- Current fire index and daily forecast for 5 days are available for any geo location (lat/lon)
- JSON format
- Please [contact us](#) to get a quote



## 2.2. Tiền xử lí dữ liệu

	Date	Time	Temperature (°C)	Weather	Wind	Wind Direction	Humidity	Barometer (mbar)	Visibility (km)
0	2019-02-01	12:00 SAT6 01 Tháng hai p.	26	Passing clouds.	9 km/h	Wind blowing from 130° Southeast to Northwest	70%	1013 mbar	NaN
1	2019-02-01	12:30 SA	26	Passing clouds.	9 km/h	Wind blowing from 130° Southeast to Northwest	74%	1013 mbar	NaN
2	2019-02-01	1:00 SA	26	Passing clouds.	11 km/h	Wind blowing from 130° Southeast to Northwest	74%	1013 mbar	NaN
3	2019-02-01	1:30 SA	25	Passing clouds.	9 km/h	Wind blowing from 140° Southeast to Northwest	79%	1012 mbar	NaN
4	2019-02-01	2:00 SA	25	Passing clouds.	11 km/h	Wind blowing from 130° Southeast to Northwest	79%	1012 mbar	NaN

- Đối với các yếu tố thời tiết tiến hành loại bỏ đơn vị, xử lí lại thời gian(AM,PM -> 24h).
- Cột Visibility phần lớn giá trị đều bị lỗi nên sẽ tiến hành loại bỏ cột.

	datetime_utc	aqi	co	no	no2	o3	so2	pm2_5	pm10	nh3
0	2020-11-25 01:00:00	5	1922.61	51.86	22.96	1.61	53.88	66.82	88.41	14.69
1	2020-11-25 02:00:00	5	2243.04	62.59	34.27	4.87	73.43	72.41	96.33	18.75
2	2020-11-25 03:00:00	5	2296.45	59.90	56.21	13.41	94.41	78.21	104.83	23.31
3	2020-11-25 04:00:00	4	1188.28	22.13	55.52	65.09	82.02	47.59	62.52	13.05
4	2020-11-25 05:00:00	5	1001.36	10.51	52.09	123.02	82.02	56.25	70.02	11.40

- Đối với chất lượng không khí, nhận thấy thời gian lưu trữ chỉ từ 25-11-2020 nên nhóm quyết định bắt đầu lấy từ 01-01-2021
- Sau đó tiến hành gộp 2 bộ dữ liệu dựa trên thời gian

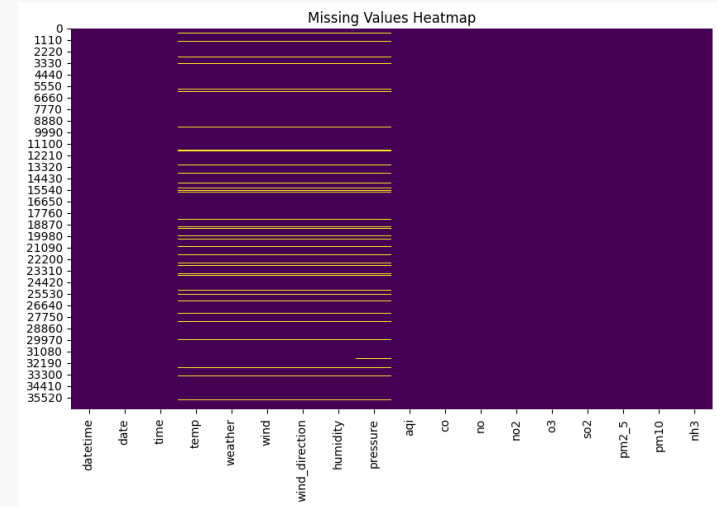
## 2.2. Tiền xử lí dữ liệu

Thu được bộ dữ liệu gồm 18 cột và 36624 mẫu dữ liệu

Tên Cột	Kiểu Dữ Liệu	Giải thích ý nghĩa
datetime	datetime	Thời gian đầy đủ (ngày + giờ)
date	date	Ngày quan sát
time	time	Giờ quan sát
temp	numerical	Nhiệt độ (°C)
weather	categorical	Tình trạng thời tiết
wind	numerical	Tốc độ gió km/h hoặc
wind_direction	numerical	Hướng gió (từ 0°-360°)
humidity	numerical	Độ ẩm tương đối (%)
pressure	numerical	Áp suất khí quyển (mbar)
aqi	numerical	Chất lượng không khí (Air Quality Index)
co	numerical	Nồng độ khí CO $\mu\text{g}/\text{m}^3$
no	numerical	Nồng độ NO $\mu\text{g}/\text{m}^3$
no2	numerical	Nồng độ NO <sub>2</sub> $\mu\text{g}/\text{m}^3$
o3	numerical	Nồng độ O <sub>3</sub> $\mu\text{g}/\text{m}^3$
so2	numerical	Nồng độ SO <sub>2</sub> $\mu\text{g}/\text{m}^3$
pm2_5	numerical	Nồng độ bụi mịn PM2.5 $\mu\text{g}/\text{m}^3$
pm10	numerical	Nồng độ bụi mịn PM10 $\mu\text{g}/\text{m}^3$
nh3	numerical	Nồng độ NH <sub>3</sub> $\mu\text{g}/\text{m}^3$

## 2.3. Khám phá và phân tích dữ liệu

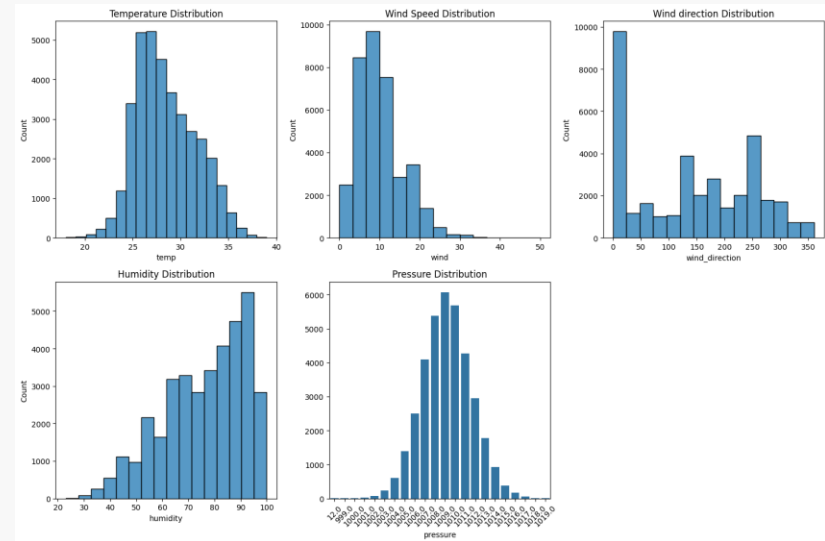
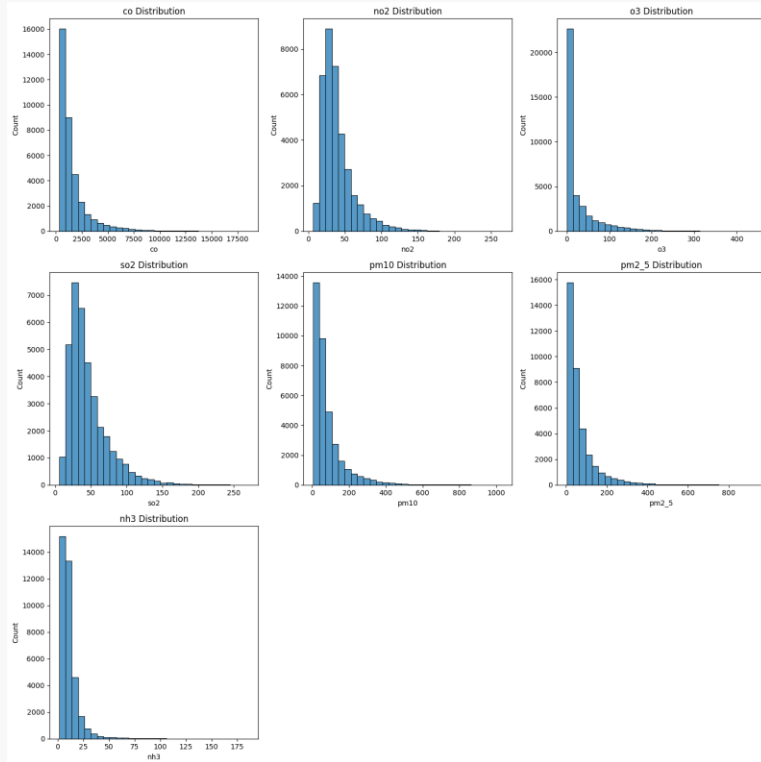
- Kiểm tra thấy có 4 giá trị lỗi -9999 tại các chỉ số không khí nên thay thành các giá trị NaN
- Kiểm tra thấy bộ dữ liệu bị thiếu những khoảng nhỏ (và tỉ lệ bị thiếu cũng không cao 7%) nên tiến hành sử dụng phương pháp ForwardFill
- Dữ liệu không có sự trùng lặp



```
[11] # Xem dữ liệu trùng lặp
      print("Duplicate rows:\n", data.duplicated().sum())

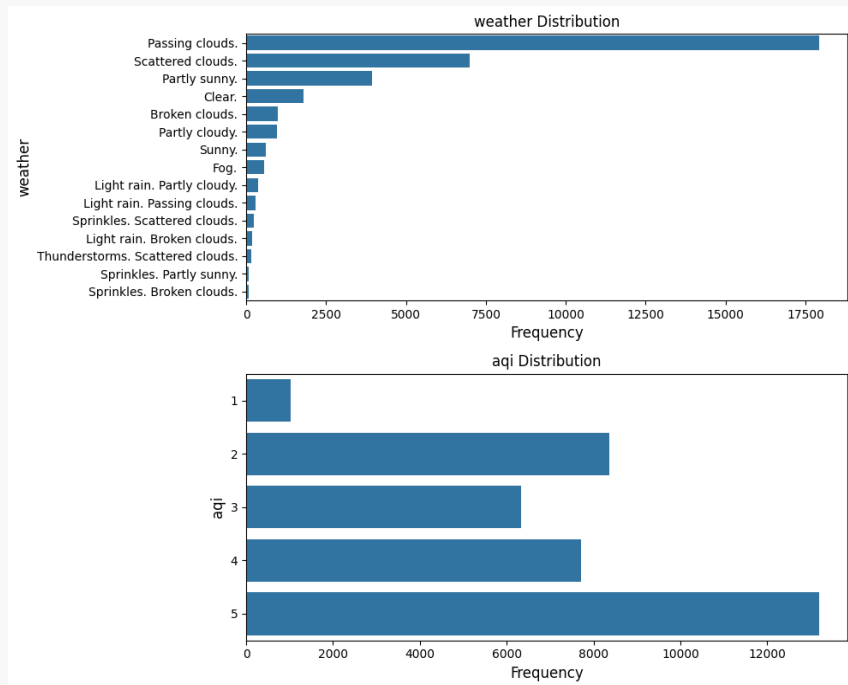
... Duplicate rows:
      0
```

## 2.3. Khám phá và phân tích dữ liệu



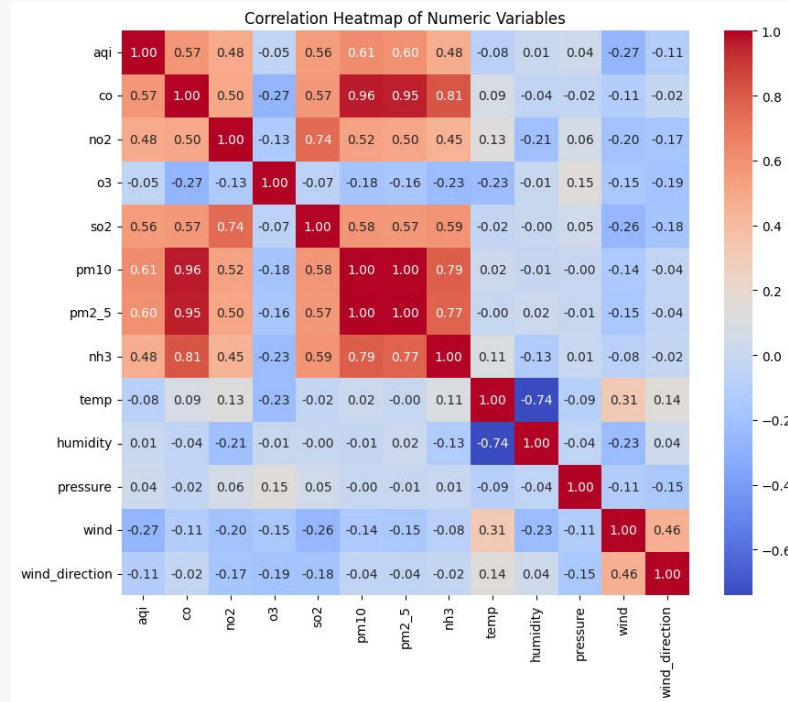
## 2.3. Khám phá và phân tích dữ liệu

- Hầu hết các đặc trưng Numerical đều có phân phối chuẩn.
- Riêng đặc trưng Categorical:
  - Thời tiết chỉ có một vài kiểu thời tiết chiếm ưu thế.
  - Còn đối với chỉ số AQI thì cho thấy được mức độ ô nhiễm không khí của thành phố đang ở mức báo động.



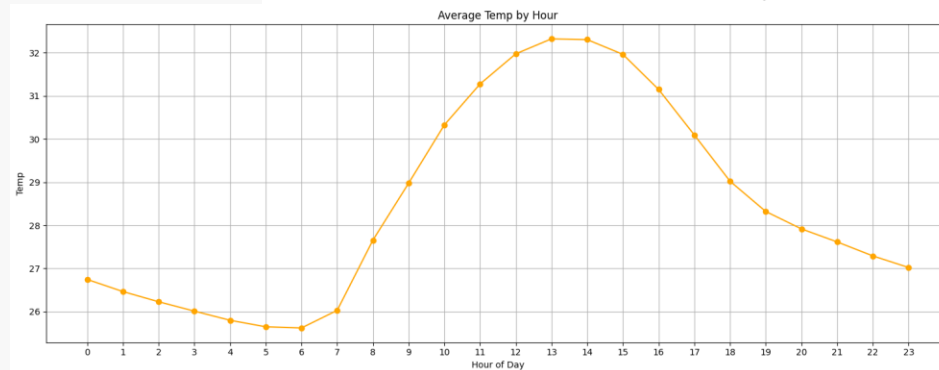
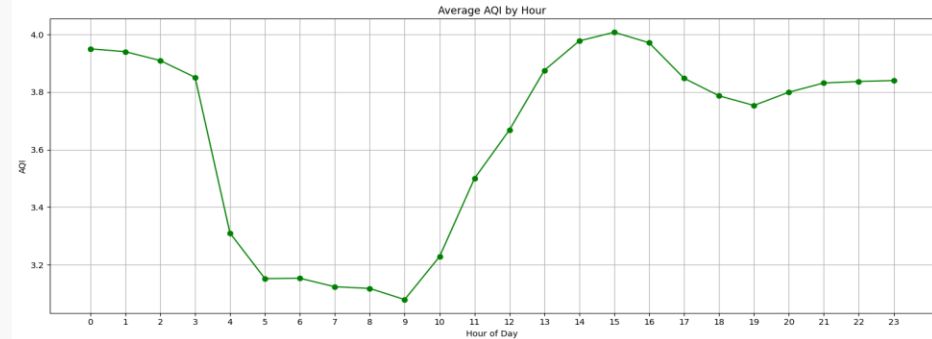
## 2.3. Khám phá và phân tích dữ liệu

- Các chất ô nhiễm trong không khí như CO, PM2.5, PM10 và NH3 **liên quan chặt chẽ với nhau**, thường cùng xuất hiện trong môi trường ô nhiễm nặng (giao thông, công nghiệp).
- Humidity – Temp:** có mối tương quan âm. Độ ẩm giảm khi nhiệt độ tăng – mối quan hệ thời tiết thường thấy.
- AQI – Wind:** gió giúp phát tán khí ô nhiễm, làm giảm AQI.



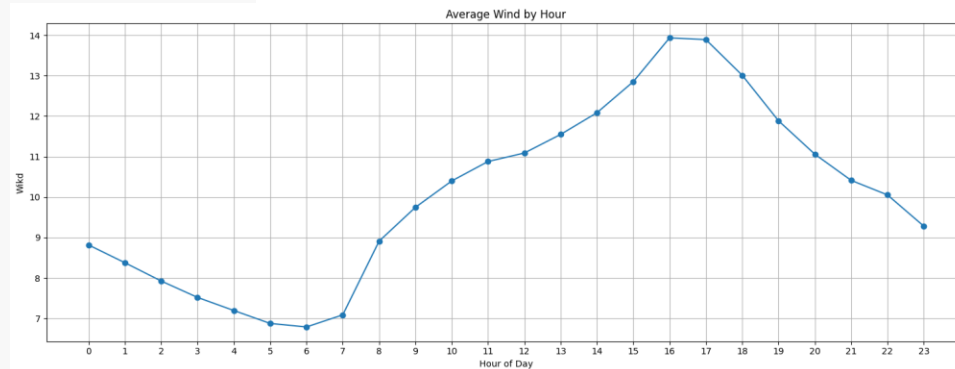
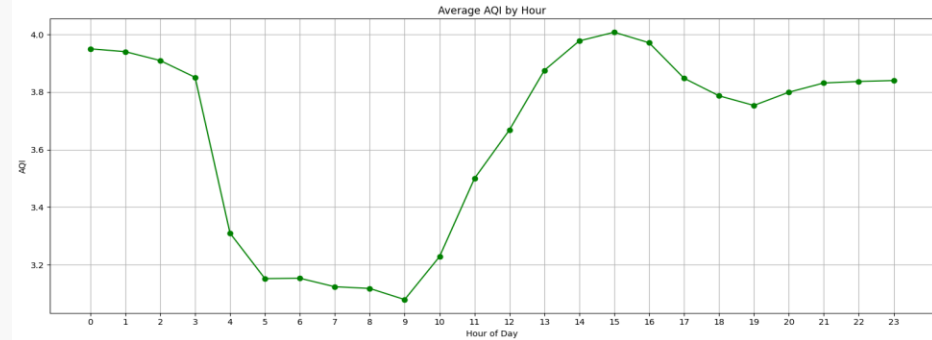
## 2.3. Khám phá và phân tích dữ liệu

- Mặc dù có hình dạng tương tự với **Nhiệt độ** nhưng đường như **Nhiệt độ** không có sự liên quan tới chỉ số **AQI** trung bình trong 1 ngày.



## 2.3. Khám phá và phân tích dữ liệu

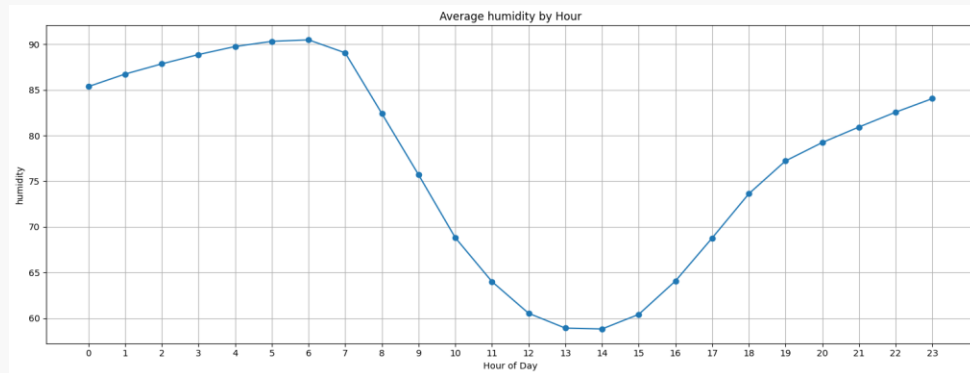
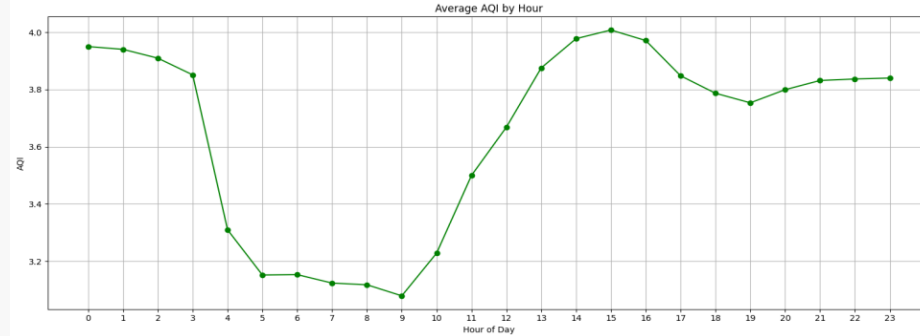
- Tương tự với **Nhiệt độ**, **Gió** dường như không có sự liên quan tới chỉ số **AQI** trung bình trong 1 ngày.





## 2.3. Khám phá và phân tích dữ liệu

- Hai biểu đồ cho thấy xu hướng **ngược chiều rõ rệt** giữa **độ ẩm** và **AQI** trong ngày.
- Điều này phù hợp với nhận định rằng độ ẩm cao giúp **giảm bụi mịn**

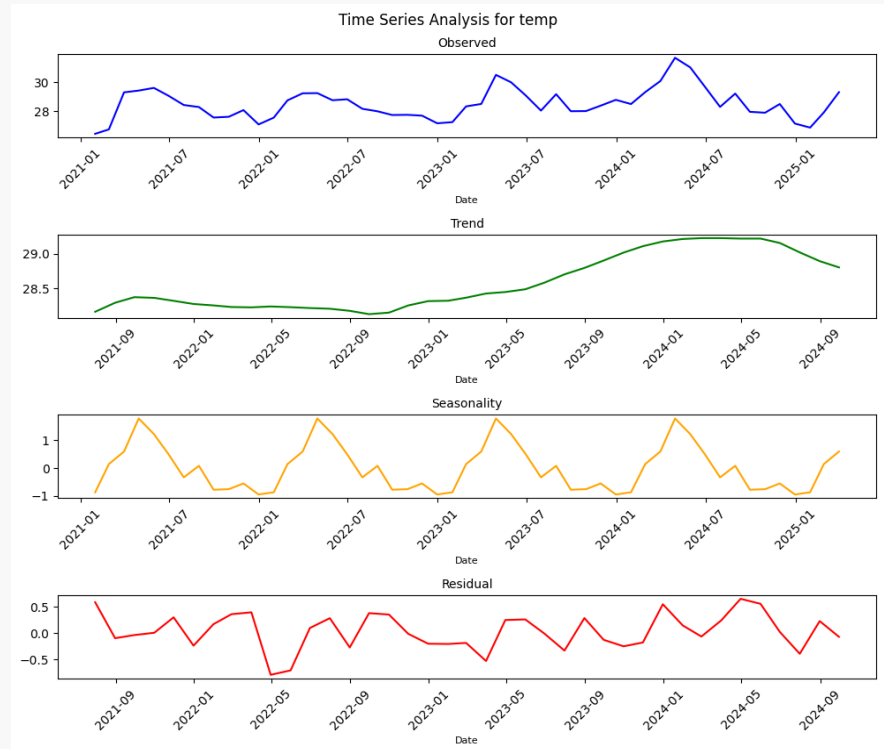


## 2.3. Khám phá và phân tích dữ liệu

Đối với Nhiệt độ:

- Trend: Nhiệt độ **tăng nhẹ** từ 2021 đến giữa 2024, sau đó **giảm** ở 2025, phản ánh biến đổi khí hậu hoặc thay đổi thời tiết.
- Seasonality: Nhiệt độ có chu kỳ 12 tháng, **cao** vào giữa năm vì mùa hè, **thấp** vào đầu/cuối năm (mùa đông), phù hợp với khí hậu nhiệt đới/ôn đới.
- Residual: Dao động nhỏ, dữ liệu ổn định.

-> Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình time-series (SARIMA, LSTM).

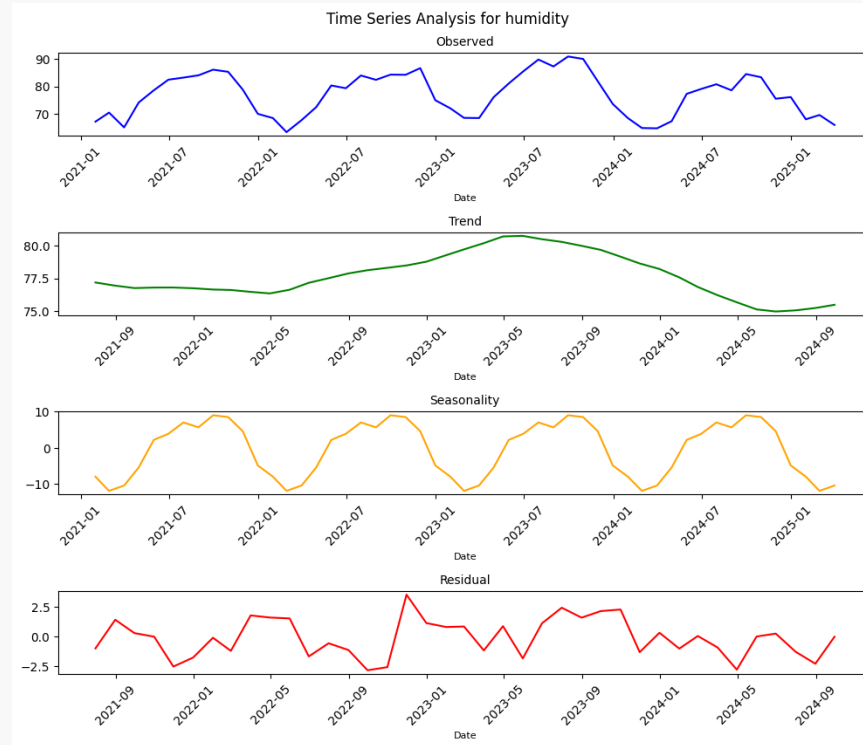


## 2.3. Khám phá và phân tích dữ liệu

Đối với Độ ẩm:

- Trend: Độ ẩm trung bình **giảm nhẹ** từ 2021 đến giữa 2023, sau đó **tăng** trở lại lên khoảng vào 2025.
- Seasonality: Độ ẩm có chu kỳ 12 tháng, **cao** vào giữa năm (mùa mưa), **thấp** vào đầu/cuối năm (mùa khô).
- Residual: Dao động **nhỏ**, dữ liệu **ổn định, ít nhiễu**.

-> Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo

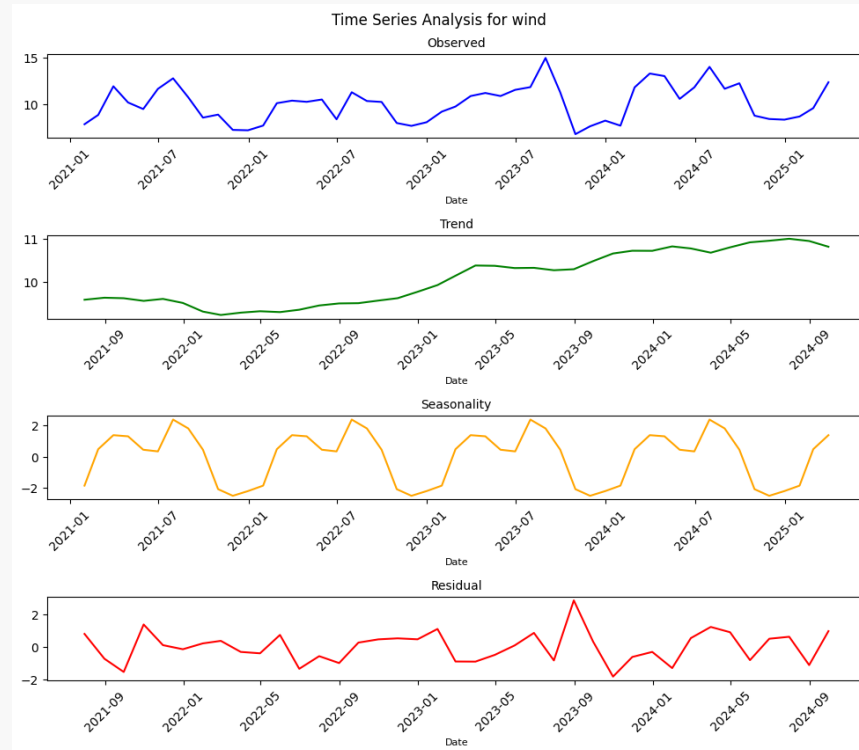


## 2.3. Khám phá và phân tích dữ liệu

Đối với Tốc độ gió:

- Trend: Tốc độ gió trung bình **tăng nhẹ** từ 2021 đến giữa 2024, sau đó **giảm nhẹ** về 10 vào 2025
- Seasonality: Tốc độ gió có chu kỳ 12 tháng, **cao** vào giữa năm (mùa hè), **thấp** vào đầu/cuối năm (mùa đông)
- Residual: Dao động nhỏ, dữ liệu **ổn định**, ít nhiễu.

-> Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo

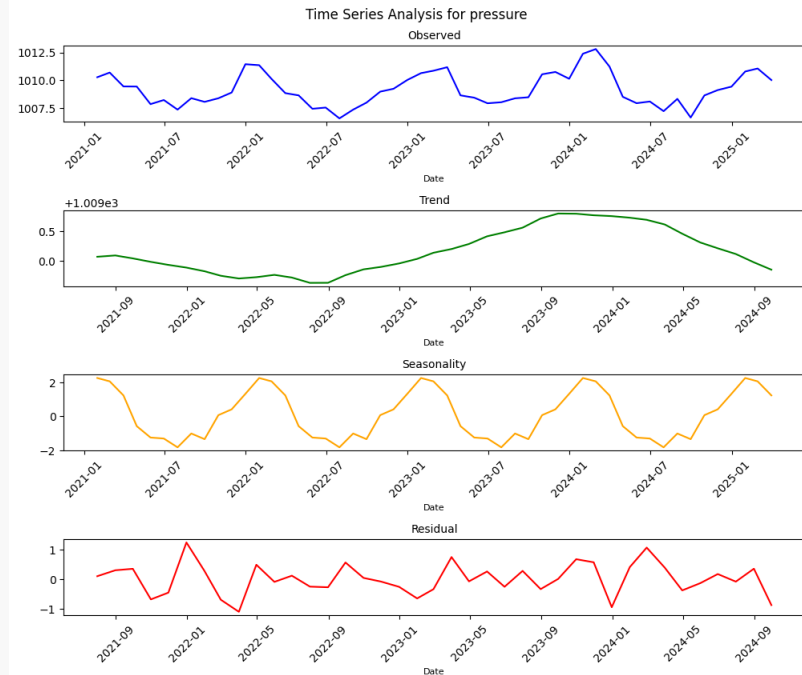


## 2.3. Khám phá và phân tích dữ liệu

Đối với Áp suất:

- Trend: Áp suất trung bình **giảm nhẹ** từ 2021 đến giữa 2022, sau đó tăng lên 1011 mBar vào giữa 2024, rồi **giảm nhẹ** về 1009 mBar vào 2025, phản ánh biến động khí hậu.
- Seasonality: Áp suất có chu kỳ 12 tháng, cao vào đầu/cuối năm (mùa đông), **thấp** vào giữa năm (mùa hè)
- Residual: Dao động **nhỏ**, dữ liệu **ổn định**, ít nhiễu.

-> Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình





## 2.3. Khám phá và phân tích dữ liệu

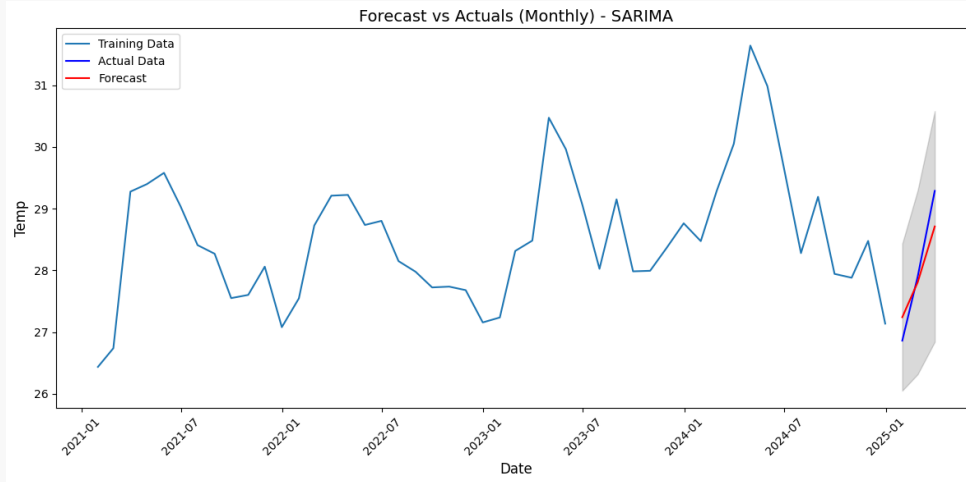
- Tiến hành sử dụng mô hình SARIMA để dự đoán các đặc trưng thời tiết (Nhiệt độ, độ ẩm, tốc độ gió và áp suất khí quyển).
- Trước khi huấn luyện mô hình, dữ liệu đầu vào được kiểm tra **tính dừng (stationarity)** nhằm đảm bảo điều kiện áp dụng phù hợp cho mô hình chuỗi thời gian. Cụ thể, nhóm đã thực hiện **kiểm định Augmented Dickey-Fuller (ADF)**.
- Các đặc trưng thời tiết (Nhiệt độ, độ ẩm, tốc độ gió và áp suất khí quyển) sau khi kiểm tra đều có tính dừng. Cụ thể:

```
# kiểm tra tính dừng của đặc trưng 'temp'
is_stationary = check_stationarity(data, 'temp', freq='ME')
✓ 0.0s
```

ADF Test for Temp:  
ADF Statistic: -4.0295961359064485  
p-value: 0.0012632246327010648  
Chuỗi temp ổn định (stationary) với p-value < 0.05.

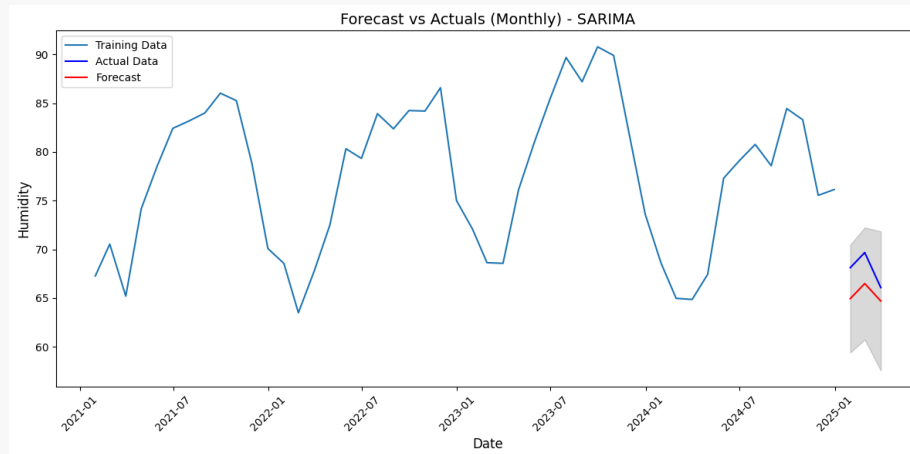
## 2.3. Khám phá và phân tích dữ liệu

- Đối với **Nhiệt độ**:
  - **Đường dự báo (Forecast)** có xu hướng **bám sát dữ liệu thực tế**, đặc biệt trong những tháng đầu tiên của giai đoạn dự báo. Điều này cho thấy mô hình **dự báo tương đối chính xác trong ngắn hạn**.
  - **Khoảng tin cậy (vùng xám)** mở rộng dần về phía cuối chuỗi cho thấy **mức độ không chắc chắn tăng theo thời gian**, điều này là **đặc điểm thường thấy** trong các mô hình chuỗi thời gian như SARIMA



## 2.3. Khám phá và phân tích dữ liệu

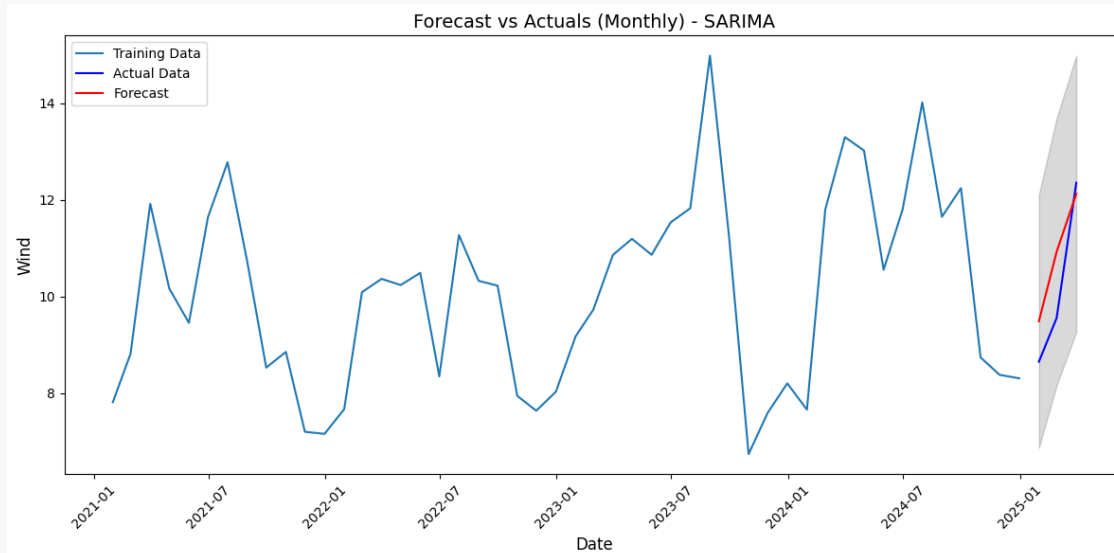
- Đối với **Độ ẩm**:
  - **Dự báo từ mô hình (Forecast)** có xu hướng **dưới mức thực tế**, đặc biệt ở tháng đầu tiên của giai đoạn dự đoán. Tuy nhiên, **khoảng cách giữa dữ liệu thực tế và dự báo vẫn nằm trong vùng tin cậy**.





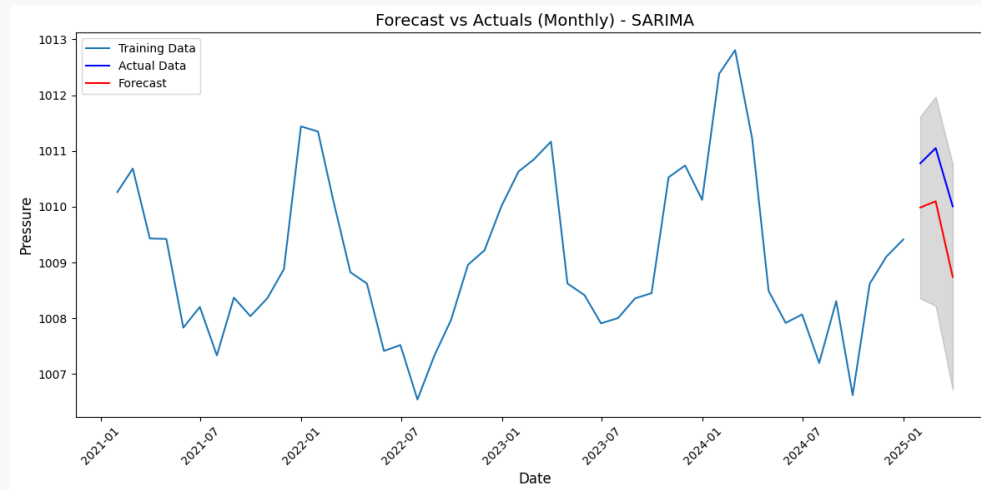
## 2.3. Khám phá và phân tích dữ liệu

- Đối với **Tốc độ gió**:
  - **Dự báo từ mô hình SARIMA** nhìn chung bám khá sát xu hướng của dữ liệu thực tế trong các tháng gần đây.
  - **Vùng xám mở rộng đáng kể** so với hai biểu đồ trước (nhiệt độ và độ ẩm), phản ánh rằng **tính không ổn định** của chuỗi gió cao hơn, khiến dự báo trở nên **ít chắc chắn hơn**.



## 2.3. Khám phá và phân tích dữ liệu

- Đối với áp suất không khí:
  - **Dự báo từ mô hình (Forecast)** có xu hướng **trên mức thực tế**, đặc biệt ở tháng đầu tiên của giai đoạn dự đoán. Tuy nhiên, **khoảng cách giữa dữ liệu thực tế và dự báo vẫn nằm trong vùng tin cậy**.



## 2.4. Xây dựng mô hình học máy

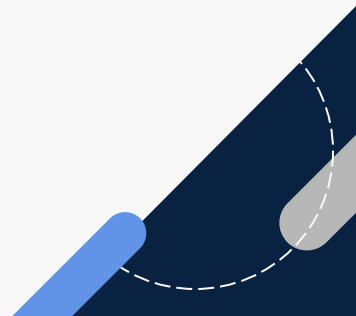
- Chỉ số **AQI** được phân loại thành **5 mức độ**.
- Nhóm quyết định xây dựng mô hình để dự đoán phân loại AQI dựa theo các chỉ số thời tiết và chỉ số không khí.

Qualitative name	Index
Good	1
Fair	2
Moderate	3
Poor	4
Very Poor	5



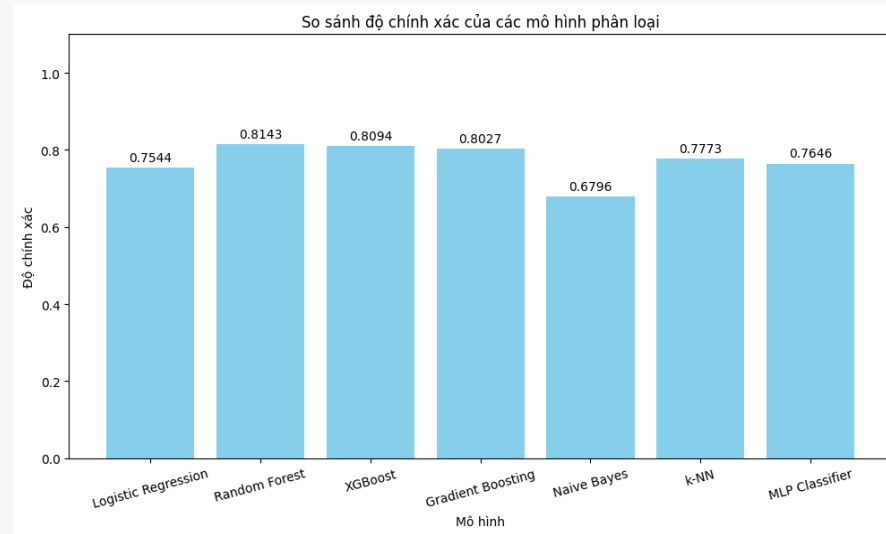
## 2.4. Xây dựng mô hình học máy

- Vì chỉ số được phân loại gán thành **5 nhãn**. Do đó nhóm chọn các phương pháp máy học có giám sát:
  - Logistic Regression: Mô hình phân loại hồi quy.
  - Random Forest, XGBoost, GradientBoost.
  - Naive Bayes.
  - k-NN.
  - Multilayer Perceptron: Mô hình Neural Network cơ bản (2 hidden layer).



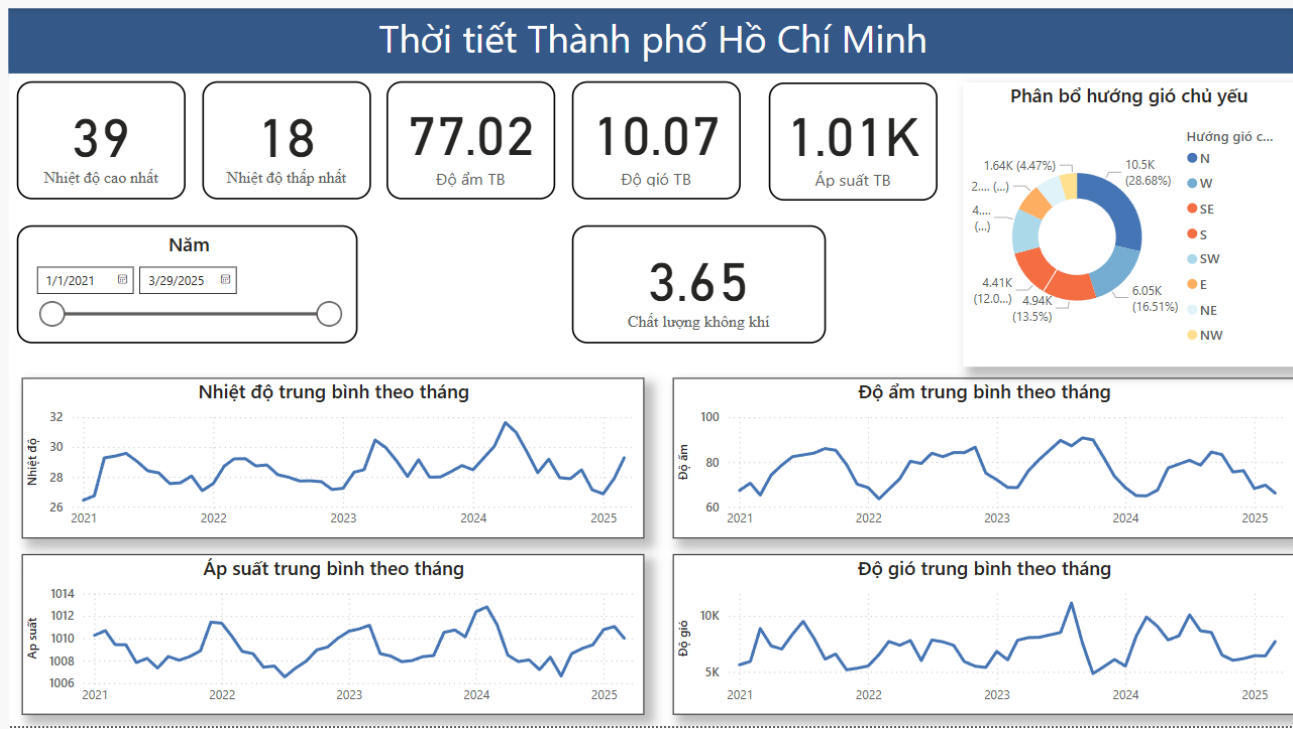
## 2.4. Xây dựng mô hình học máy

- **Random Forest, XGBoost, Gradient Boosting** là những thuật toán Tree-based algorithms. Có accuracy khá cao, cho thấy khả năng phân loại của phương pháp này tốt.
- **MLP Classifier** với **79.64%**: đây là mô hình mạng Neural cơ bản cũng cho thấy độ hiệu quả khi phân loại.
- k-NN đạt 76.50%.
- Logistic Regression với 74.45%.
- Naive Bayes đạt 66.79%.

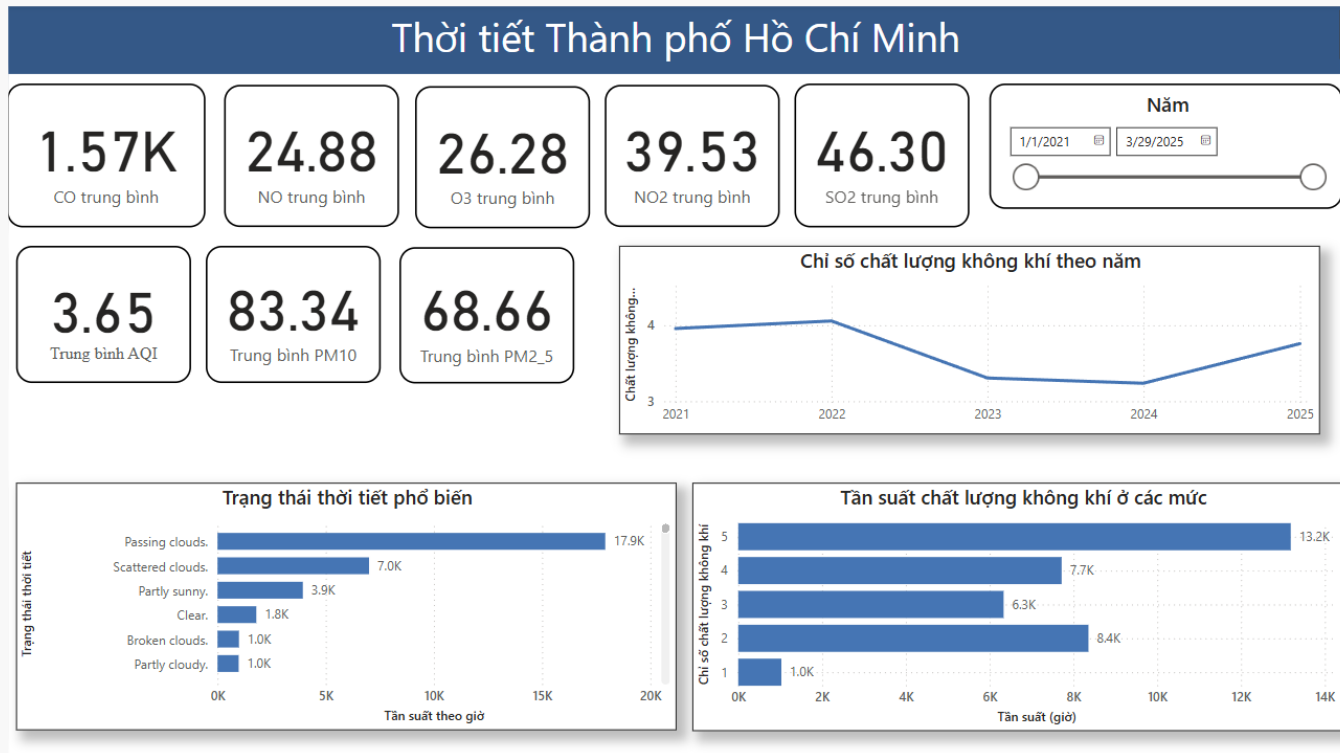


## 2.4. Trực quan hóa

- Trực quan hóa sử dụng PowerBI



## 2.4. Trực quan hóa



03

## Kết quả & Đề xuất







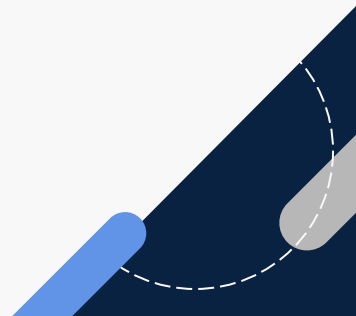
### 3. Kết quả và đề xuất

#### Kết quả:

- Phân tích chuỗi thời gian với mô hình **SARIMA** cho thấy khả năng dự báo tốt về xu hướng tổng thể và biến động mùa vụ của nhiệt độ, độ ẩm, vận tốc gió và áp suất trong 3 tháng đầu năm 2025.
- Sau khi phân tích ma trận tương quan và sử dụng các mô hình học máy cho thấy dường như yếu tố thời tiết không ảnh hưởng nhiều tới chỉ số chất lượng không khí.

#### Đề xuất:

- Khi xây dựng mô hình học máy, cần phải **fine-tune** mô hình để đưa ra các tham số tốt nhất của mô hình. Nhưng vì tài nguyên không cho phép nên chỉ thực hiện bằng những tham số tốt nhất của mô hình.
- Hoàn thiện thêm pipeline phân tích dữ liệu một cách tự động hóa.



04

## Ứng dụng hỗ trợ phân tích dữ liệu





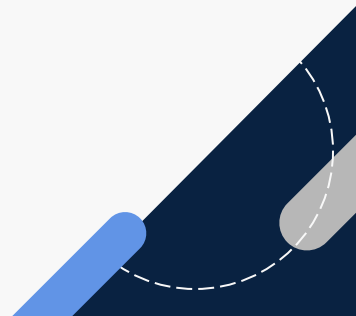
## 4. Ứng dụng hỗ trợ phân tích dữ liệu

Nhóm thực hiện build ứng dụng nhằm hỗ trợ việc phân tích dữ liệu file csv, bao gồm các tính năng như:

- Tải lên tệp CSV.
- Trả lời câu hỏi: Phân tích dữ liệu và đưa ra phản hồi thông minh thông qua giao diện trò chuyện. Tạo biểu đồ và hình ảnh minh họa (nếu có) dựa trên dữ liệu và các câu hỏi.
- Tương tác với dữ liệu: Công cụ trực quan hóa.

Ứng dụng này sử dụng:

- Streamlit
- Langchain
- Mô hình ngôn ngữ lớn (LLMs):
- PyGWalker: Công cụ mạnh mẽ để tạo trực quan hóa dữ liệu tương tác.
- API: (Tùy chọn) Sử dụng các mô hình ngôn ngữ lớn (DeepSeek)



## 4. Ứng dụng hỗ trợ phân tích dữ liệu

The screenshot shows the QuQuBeTa application interface. On the left, there's a sidebar with 'app' and 'visualization' tabs. The main area is titled 'QuQuBeTa - Your Data Analysis Assistant'. It includes a 'Select a model' dropdown set to 'deepseek', a 'Data Preview' table, and a 'Run Query' button. The table displays weather data for four time slots.

	datetime	date	time	temp	weather	wind	wind_direction	humidity	pressure
0	1/1/2023 0:00	1/1/2023	0:00	24	Passing clouds.	13	30	61	1012
1	1/1/2023 1:00	1/1/2023	1:00	24	Passing clouds.	11	40	61	1011
2	1/1/2023 2:00	1/1/2023	2:00	23	Passing clouds.	9	20	60	1011
3	1/1/2023 3:00	1/1/2023	3:00	23	Passing clouds.	7	10	60	1011
4	1/1/2023 4:00	1/1/2023	4:00	23	Passing clouds.	13	30	60	1011

The screenshot shows the QuQuBeTa application interface with the 'Output' section. It displays a list of statistics for the weather data, including total rows, average temperature, and humidity. Below the output, there's a 'Intermediate Steps' section showing the code used to load and process the data.

**Input:**  
Hãy phân tích dữ liệu bên

**Output:**  
Dữ liệu bao gồm thông tin thời tiết và chất lượng không khí theo giờ trong năm 2021 với các điểm chính:

- Tổng cộng 36,424 bản ghi với 18 cột
- Cả giá trị thiếu ở các cột thời tiết (nhiệt độ, gió, độ ẩm...) chiếm ~7.4%
- Cả giá trị bất thường cần xử lý trong cột độ ẩm và pm2.5
- Chỉ số lượng không khí (AQI) trung bình (AQI) trong các bản ghi đầu
- Nhiệt độ trung bình 18.56°C, độ ẩm 76.62%, áp suất ~1000 hPa
- Cần làm sạch dữ liệu (xử lý giá trị thiếu và bất thường) trước khi phân tích sâu hơn

**Intermediate Steps:**  
Step: python\_req\_ast  
Tool Input:  
`df = load()`  
Log: Thought: Tôi cần phân tích dữ liệu trong DataFrame bằng cách xem xét các thông tin cơ bản như cấu trúc dữ liệu, thông kê mô tả, và kiểm tra các giá trị thiếu.  
Action: python\_req\_ast Action Input: df.info()  
Result:  

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 36424 entries, 0 to 36423
```



Cảm ơn  
đã lắng nghe

