BÁO CÁO
Đồ Án Cuối Kỳ

NHẬP MÔN KHOA HỌC DỮ LIỆU

OUR TEAM

20120307

Phạm Gia Khiêm

20120210

Trần Thị Kim Tiến

20120334

Lý Thành Nam

20120258

Lâm Quốc Chung

THÔNG TIN VỀ DỮ LIỆU

Tên dữ liệu: WORLD MOST POLLUTED CITIES 2017 – 2022

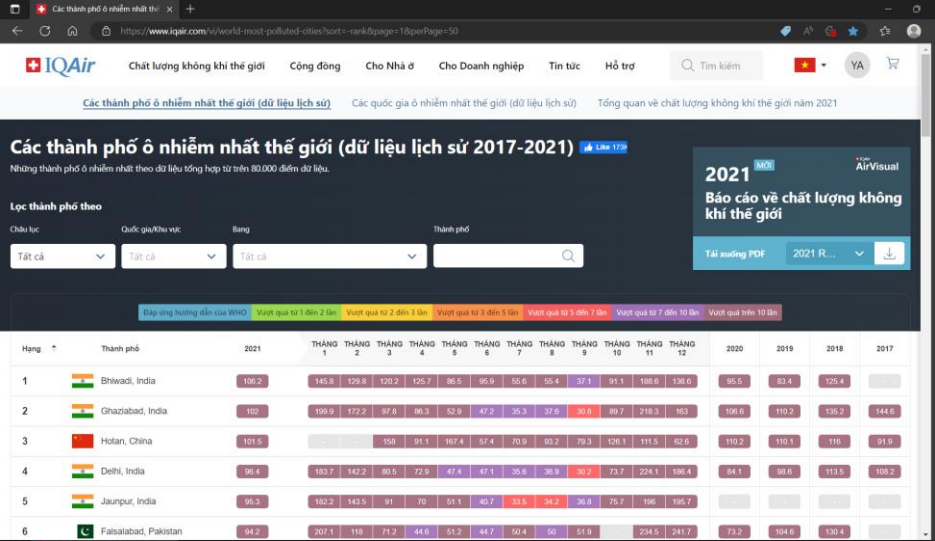
Thu thập dữ liệu, trên trang web:

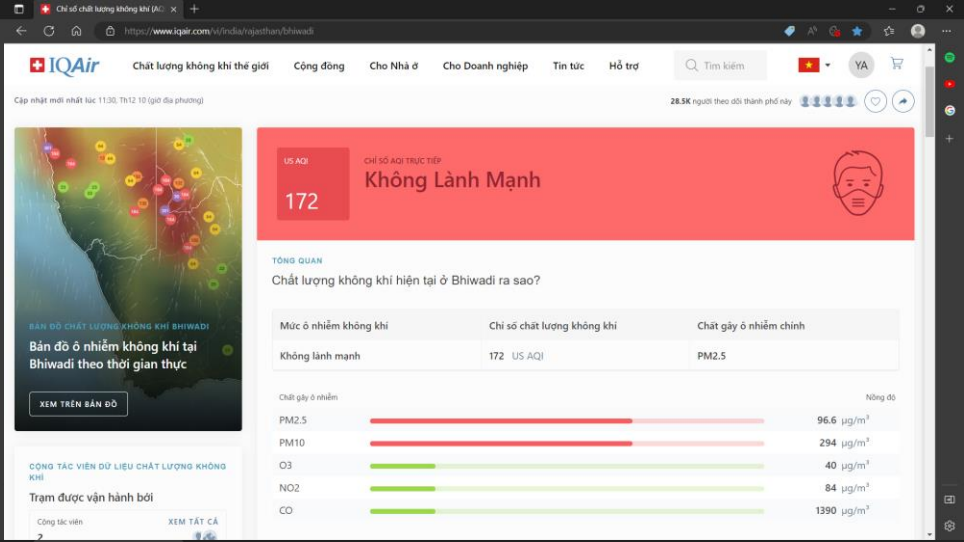
https://www.iqair.com/vi/world-most-polluted-cities?sort=-rank&page=1&perPage=50

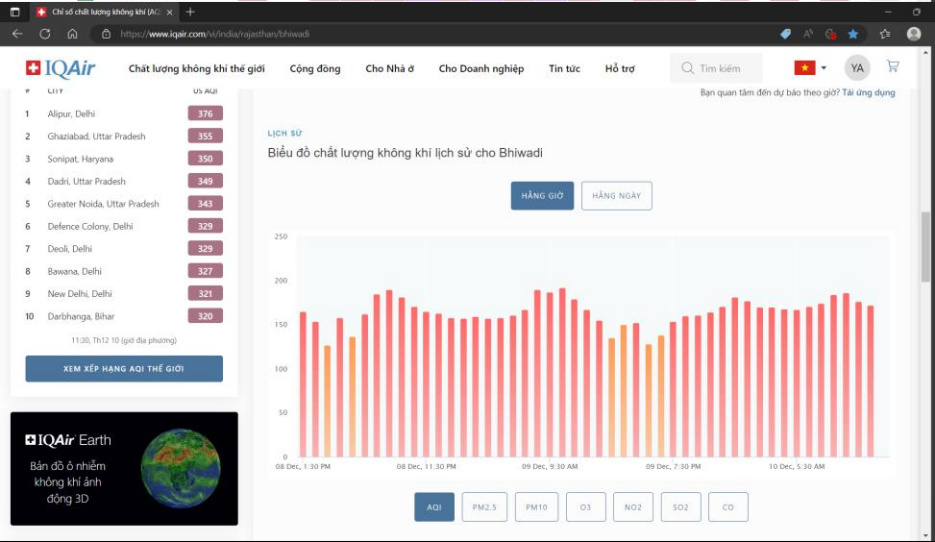
Cách thức crawl dữ liệu:

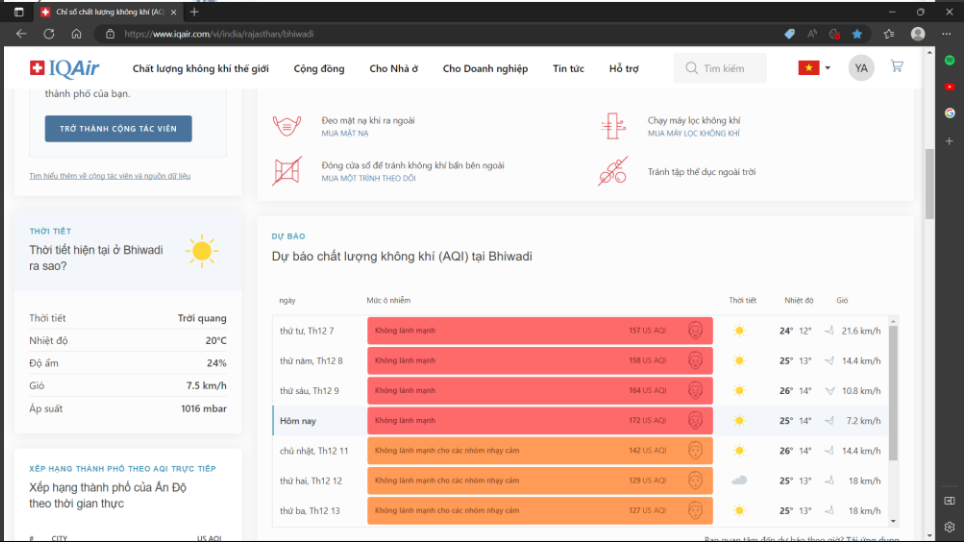
Sử dụng công cụ Scapy để lấy dữ liệu ô nhiễm của các địa điểm (city) của một quốc gia (country) từ 2017 đến 2021 và 12 tháng năm 2022.

Sau đó sẽ dùng Beautiful Soup để lấy dữ liệu về thời tiết và khí hậu tại một địa điểm của một quốc gia tính theo realtime.


IQAir
Chất lượng không khí thế giới
Cộng đồng
Cho Nhà ở
Cho Doanh nghiệp
Tin tức
Hỗ trợ
Tìm kiếm
YA
Các thành phố ô nhiễm nhất thế giới (dữ liệu lịch sử)
Các quốc gia ô nhiễm nhất thế giới (dữ liệu lịch sử)
Tổng quan về chất lượng không khí thế giới năm 2021
Các thành phố ô nhiễm nhất thế giới (dữ liệu lịch sử 2017-2021)
Like 17/3
Những thành phố ô nhiễm nhất theo dữ liệu tổng hợp từ trên 80.000 điểm dữ liệu.
2021 MỚI
AirVisual
Báo cáo về chất lượng không khí thế giới
Tải xuống PDF
2021 R...
Lọc thành phố theo
Châu lục
Quốc gia/Khu vực
Bang
Thành phố
Tất cả
Đáp ứng hướng dẫn của WHO
Vượt quá từ 1 đến 2 lần
Vượt quá từ 2 đến 3 lần
Vượt quá từ 3 đến 5 lần
Vượt quá từ 5 đến 7 lần
Vượt quá từ 7 đến 10 lần
Vượt quá trên 10 lần
Hạng
Thành phố
2021
THÁNG 1
THÁNG 2
THÁNG 3
THÁNG 4
THÁNG 5
THÁNG 6
THÁNG 7
THÁNG 8
THÁNG 9
THÁNG 10
THÁNG 11
THÁNG 12
2020
2019
2018
2017
1 Bhiwadi, India 106.2 145.8 129.8 120.2 125.7 86.5 95.9 55.6 55.4 37.1 91.1 188.6 136.6 95.5 83.4 125.4
2 Ghaziabad, India 102 199.9 172.2 97.8 86.3 52.9 47.2 35.3 37.6 30.8 89.7 218.3 163 106.6 110.2 135.2 144.6
3 Hotan, China 101.5 158 91.1 167.4 57.4 70.9 93.2 79.3 126.1 111.5 62.6 110.2 110.1 116 91.9
4 Delhi, India 99.4 183.7 142.2 80.5 72.9 47.4 47.1 35.6 36.9 30.2 73.7 224.1 186.4 84.1 98.6 113.5 108.2
5 Jaunpur, India 95.3 182.2 143.5 91 70 51.1 40.7 33.5 34.2 36.8 75.7 196 195.7
6 Faisalabad, Pakistan 94.2 207.1 118 71.2 44.6 51.2 44.7 50.4 50 51.9 234.5 241.7 73.2 104.6 130.4


IQAir
Chất lượng không khí thế giới
Cộng đồng
Cho Nhà ở
Cho Doanh nghiệp
Tin tức
Hỗ trợ
Tìm kiếm
YA
Cập nhật mới nhất lúc 11:30, Th12 10 (giờ địa phương)
28.5K người theo dõi thành phố này
BẢN ĐỒ CHẤT LƯỢNG KHÔNG KHÍ BHIWADI
Bản đồ ô nhiễm không khí tại Bhiwadi theo thời gian thực
XEM TRÊN BẢN ĐỒ
US AQI
172
CHỈ SỐ AQI TRỰC TIẾP
Không Lành Mạnh
TỔNG QUAN
Chất lượng không khí hiện tại ở Bhiwadi ra sao?
Mức ô nhiễm không khí
Chỉ số chất lượng không khí
Chất gây ô nhiễm chính
Không lành mạnh
172 US AQI
PM2.5
Chất gây ô nhiễm
Nồng độ
PM2.5 96.6 µg/m³
PM10 294 µg/m³
O3 40 µg/m³
NO2 84 µg/m³
CO 1390 µg/m³
CỘNG TÁC VIÊN DỮ LIỆU CHẤT LƯỢNG KHÔNG KHÍ
Trạm được vận hành bởi
Cộng tác viên
XEM TẤT CẢ


IQAir
Chất lượng không khí thế giới
Cộng đồng
Cho Nhà ở
Cho Doanh nghiệp
Tin tức
Hỗ trợ
Tìm kiếm
YA
Bạn quan tâm đến dự báo theo giờ? Tải ứng dụng
CITY US AQI
1 Alipur, Delhi 376
2 Ghaziabad, Uttar Pradesh 355
3 Sonipat, Haryana 350
4 Dadri, Uttar Pradesh 349
5 Greater Noida, Uttar Pradesh 343
6 Defence Colony, Delhi 329
7 Deoli, Delhi 329
8 Bawana, Delhi 327
9 New Delhi, Delhi 321
10 Darbhanga, Bihar 320
11:30, Th12 10 (giờ địa phương)
XEM XẾP HẠNG AQI THẾ GIỚI
IQAir Earth
Bản đồ ô nhiễm không khí ảnh động 3D
LỊCH SỬ
Biểu đồ chất lượng không khí lịch sử cho Bhiwadi
HÀNG GIỜ
HÀNG NGÀY
250
200
150
100
50
0
08 Dec, 1:30 PM
08 Dec, 11:30 PM
09 Dec, 9:30 AM
09 Dec, 7:30 PM
10 Dec, 5:30 AM
AQI
PM2.5
PM10
O3
NO2
SO2
CO


IQAir
Chất lượng không khí thế giới
Cộng đồng
Cho Nhà ở
Cho Doanh nghiệp
Tin tức
Hỗ trợ
Tìm kiếm
YA
thành phố của bạn.
TRỞ THÀNH CỘNG TÁC VIÊN
Tìm hiểu thêm về cộng tác viên và nguồn dữ liệu
Đeo mặt nạ khi ra ngoài
MUA MẶT NẠ
Chạy máy lọc không khí
MUA MÁY LỌC KHÔNG KHÍ
Đóng cửa sổ để tránh không khí bẩn bên ngoài
MUA MỘT TRÌNH THEO DÕI
Tránh tập thể dục ngoài trời
THỜI TIẾT
Thời tiết hiện tại ở Bhiwadi ra sao?
Thời tiết Trời quang
Nhiệt độ 20°C
Độ ẩm 24%
Gió 7.5 km/h
Áp suất 1016 mbar
XẾP HẠNG THÀNH PHỐ THEO AQI TRỰC TIẾP
Xếp hạng thành phố của Ấn Độ theo thời gian thực
CITY US AQI
DỰ BÁO
Dự báo chất lượng không khí (AQI) tại Bhiwadi
ngày
Mức ô nhiễm
Thời tiết
Nhiệt độ
Gió
thứ tư, Th12 7 Không lành mạnh 157 US AQI 24° 12° 21.6 km/h
thứ năm, Th12 8 Không lành mạnh 158 US AQI 25° 13° 14.4 km/h
thứ sáu, Th12 9 Không lành mạnh 164 US AQI 26° 14° 10.8 km/h
Hôm nay Không lành mạnh 172 US AQI 25° 14° 7.2 km/h
chủ nhật, Th12 11 Không lành mạnh cho các nhóm nhạy cảm 142 US AQI 26° 14° 14.4 km/h
thứ hai, Th12 12 Không lành mạnh cho các nhóm nhạy cảm 129 US AQI 25° 13° 18 km/h
thứ ba, Th12 13 Không lành mạnh cho các nhóm nhạy cảm 127 US AQI 25° 13° 18 km/h

THÔNG TIN VỀ DỮ LIỆU

Đối với dữ liệu chính:

- Mỗi dòng là một địa điểm của quốc gia. Và không có dữ liệu trùng.
- Mỗi cột mang ý nghĩa là dữ liệu về chỉ số mức độ ô nhiễm ở địa điểm của một quốc gia.
- Mỗi cột hiện đang có kiểu dữ liệu là string. Sau khi tiền xữ lý thì sẽ trở thành dạng float64.
- Với các cột năm 2017 – 2020 thì dữ liệu của mỗi cột phân bố khá thưa, có thể do có ít cộng tác viên thực hiện đo độ ô nhiễm ở các địa điểm vào thời điểm đó.

Đối với dữ liệu xét mô hình:

- Mỗi dòng là một địa điểm bất kì của một quốc gia.
- Mỗi cột là thông tin của các input truyền vào mô hình.

THƯ VIỆN VÀ KỸ THUẬT

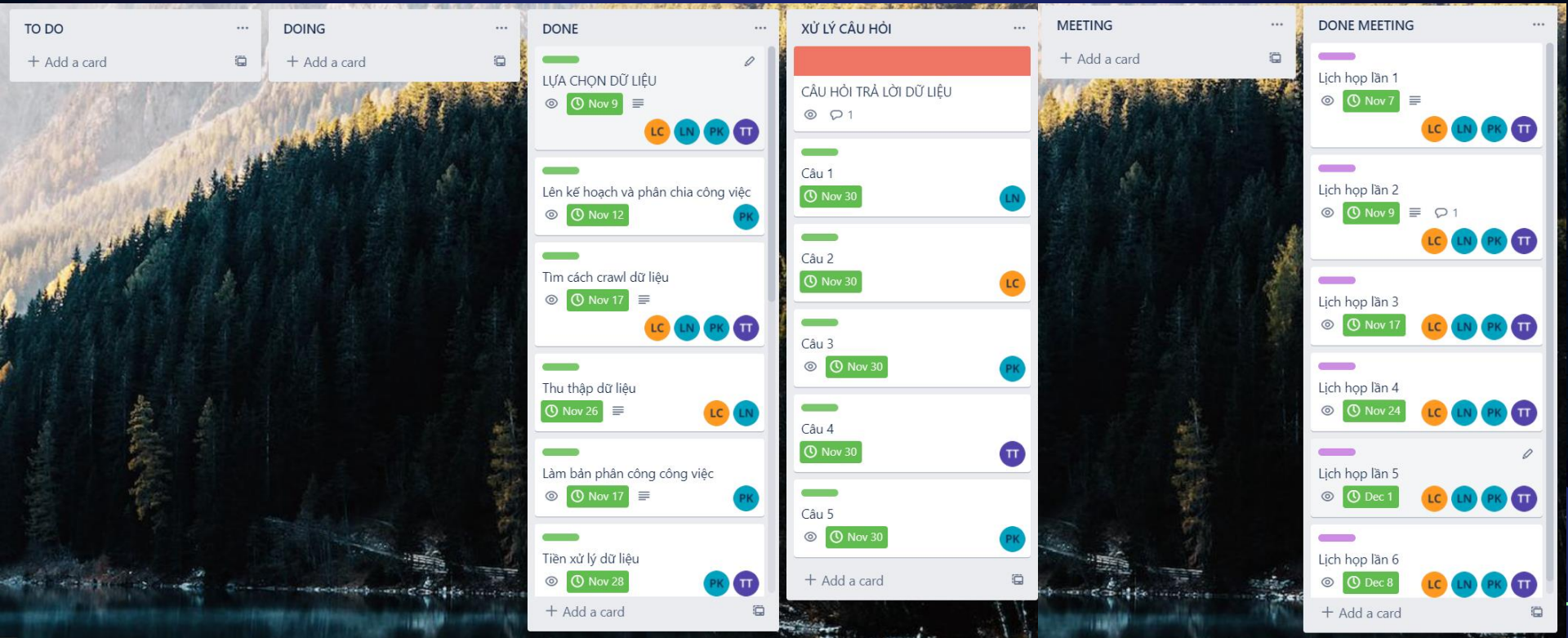
Sử dụng thư viện json, re, numpy, pandas và scrapy để cào dữ liệu từ trang web.

Thư viện

Kỹ thuật

- Dùng jupyter notebook (khuyến cáo dung linux) để tạo một folder để lưu trữ, xử lý dữ liệu cào về.
- Xữ lí, tạo hàm trong các file .py foder vừa tạo.

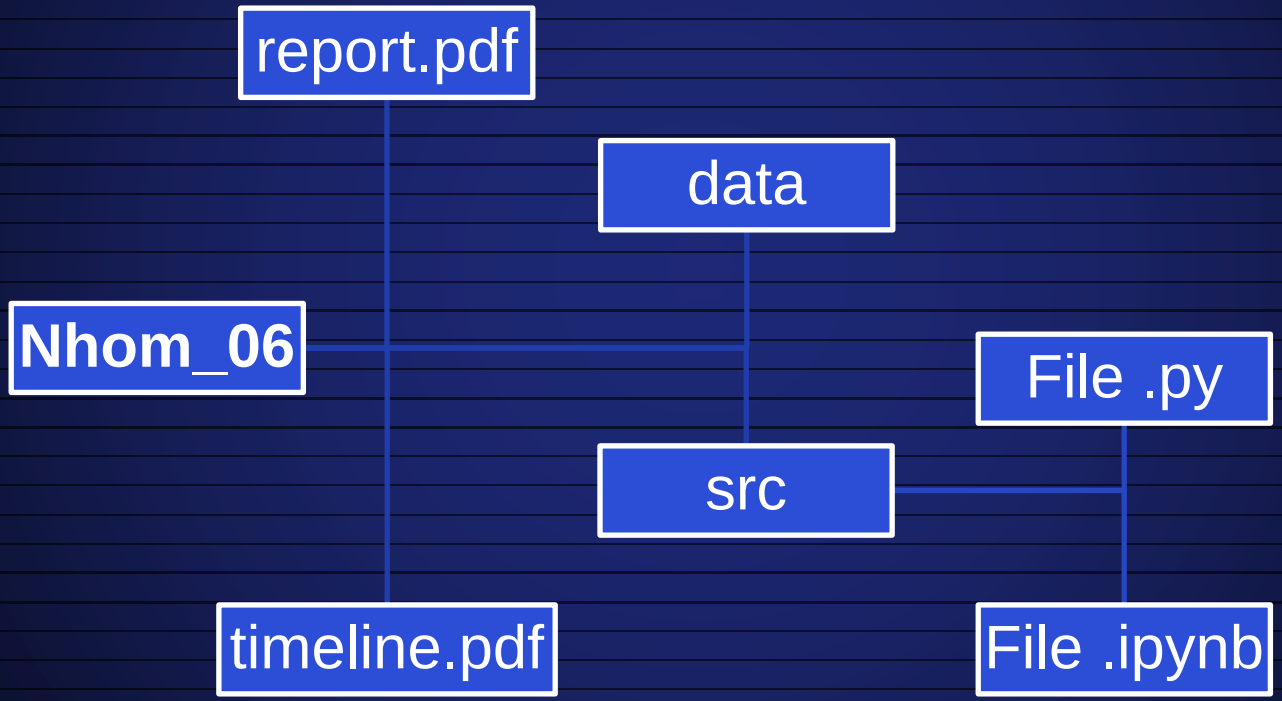
TRELLO QUẢN LÝ CÔNG VIỆC



TIMELINE

Họ và tên	Thời gian bắt đầu	DEADLINE 23:59											
		26.11		28.11		2.12		6.12	9.12		11.12		14.12	
Phạm Gia Khiêm	24.11			Tiền xử lý dữ liệu	☑	Câu hỏi 3	☑	Câu hỏi 5		☑	Phân tích đến xây dựng mô hình	☑	Đánh giá mô hình. Hoàn thành báo cáo và slide	☑
Trần Thị Kim Tiến						Câu hỏi 4				☑				
Lý Thành Nam		Thu thập dữ liệu	☑	Khám phá dữ liệu	☑	Câu hỏi 1	☑	Lấy dữ liệu để phân tích, xây dựng mô hình		☑	So sánh các phương pháp, mô hình			☑
Lâm Quốc Chung						Câu hỏi 2	☑							

CÁCH LƯU TRỮ



01

CITY

Top 5 thành phố, quốc gia có mật độ ô nhiễm cao nhất, thấp nhất.

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

Sử dụng thư viện pandas, numpy, matplotlib.pyplot và seaborn

Kỹ thuật

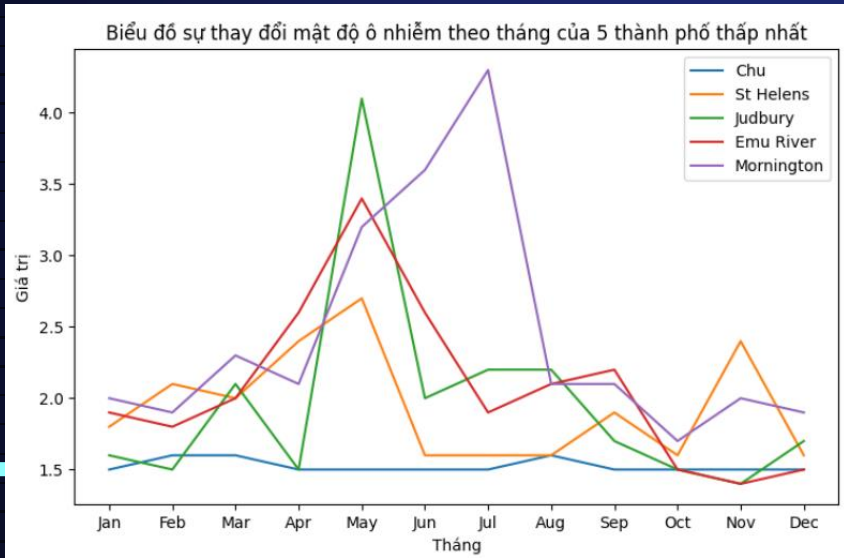
- Đầu tiên sử dụng fillna và med() để lấp những giá trị nan trong df bằng giá trị trung vị median sau đó thực hiện trả lời câu hỏi
- Sử dụng sort_values để sắp xếp cột tương ứng (City, Country)
Với City chỉ cần dùng index để lấy 5 hàng đầu tiên sau khi sort từ đó lấy dữ liệu đúng yêu cầu
- Với Country phức tạp hơn, sử dụng groupby và means() để tính ra dataframe đúng yêu cầu và thực hiện như với City

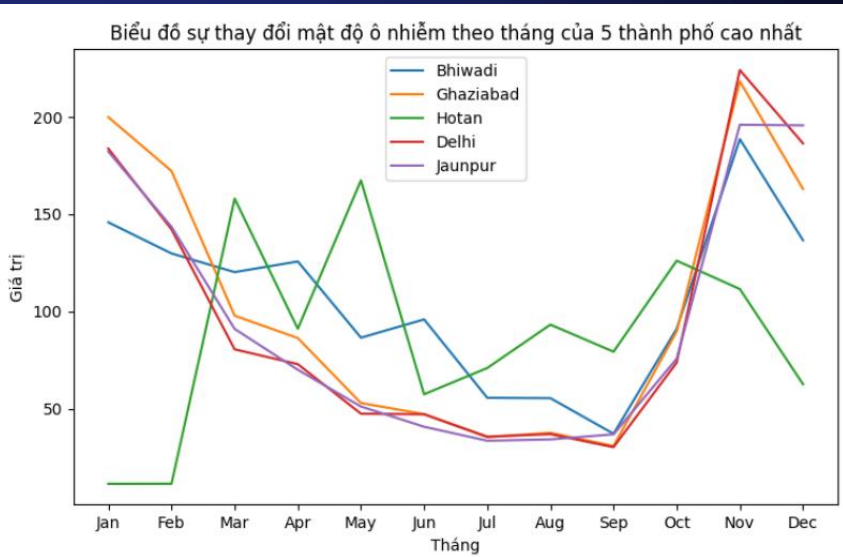
BIỂU ĐỒ 2 NHÓM THÀNH PHỐ

Cách thực hiện

- Sử dụng thư viện pandas, numpy, matplotlib.pyplot
- Thực hiện sort và lấy ra 5 dòng đầu tiên để tìm 5 thành phố chỉ số cao nhất / nhỏ nhất
- Sử dụng dữ liệu theo 12 tháng của 5 thành phố này sau đó vẽ đồ thị bằng plt.plot()

BIỂU ĐỒ 2 NHÓM THÀNH PHỐ



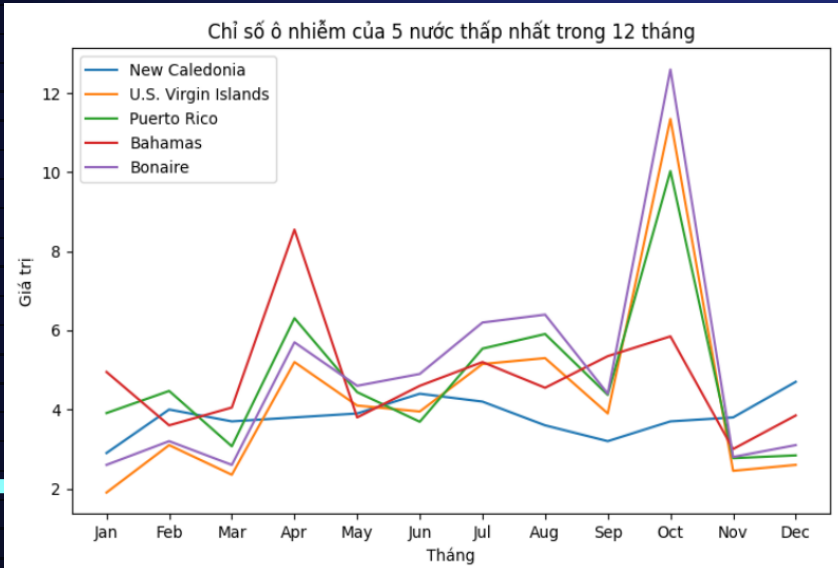


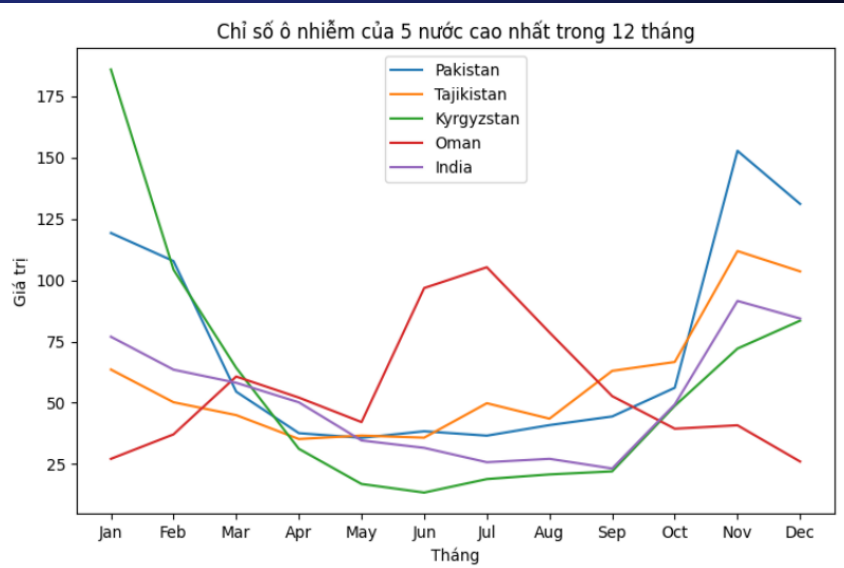
BIỂU ĐỒ 2 NHÓM QUỐC GIA

Cách thực hiện

- Sử dụng thư viện pandas, numpy, matplotlib.pyplot
- Thực hiện groupby theo cột ‘Country’ và tính giá trị mean() các cột còn lại (lưu ý làm tròn tới 2 chữ số sau dấu phẩy bằng round(*)
- Thực hiện sort df mới và lấy ra 5 dòng đầu tiên để tìm 5 thành phố chỉ số cao nhất / nhỏ nhất
- Sử dụng dữ liệu theo 12 tháng của 5 thành phố này sau đó vẽ đồ thị bằng plt.plot()

BIỂU ĐỒ 2 NHÓM QUỐC GIA





02

VIET NAM

Việt Nam nằm trong top bao nhiêu, với mật độ ô nhiễm là bao nhiêu qua từng năm (2017 - 2021).

THƯ VIỆN VÀ KỸ THUẬT

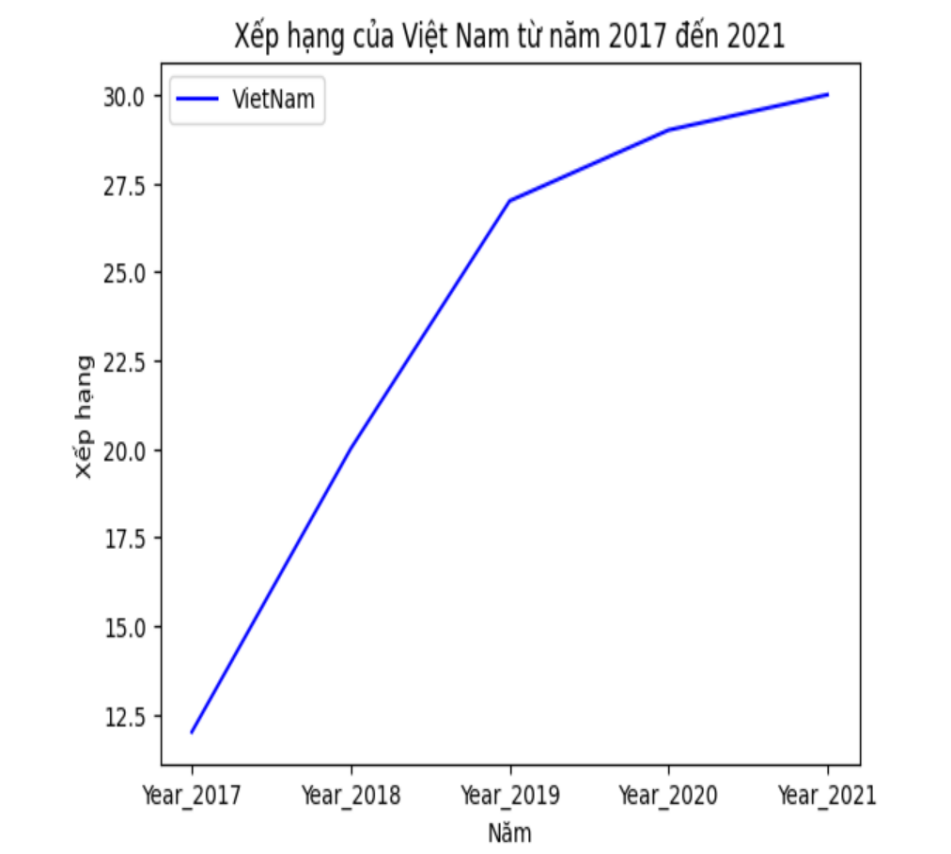
Thư viện

Sử dụng thư viện pandas, numpy, matplotlib.pyplot và seaborn

Kỹ thuật

- Dùng df.drop() để xóa đi các cột không dùng cho câu hỏi.
- Dùng (...).mean() để tính giá trị mean của 1 cột.
- Dùng df.fillna(mean) để thay giá trị nan thành mean.
- Dùng df_Country.groupby().mean() để nhóm dự liệu theo 'Country' và tính mean.
- Dùng pd.DataFrame() để nhóm dữ liệu đã tính toán thành bảng.
- Viết hàm get_rank() để tính thứ hạng của quốc gia.
- Trực quan hóa dữ liệu bằng plt.plot()

Thứ hạng Việt Nam giai đoạn 2017-2021



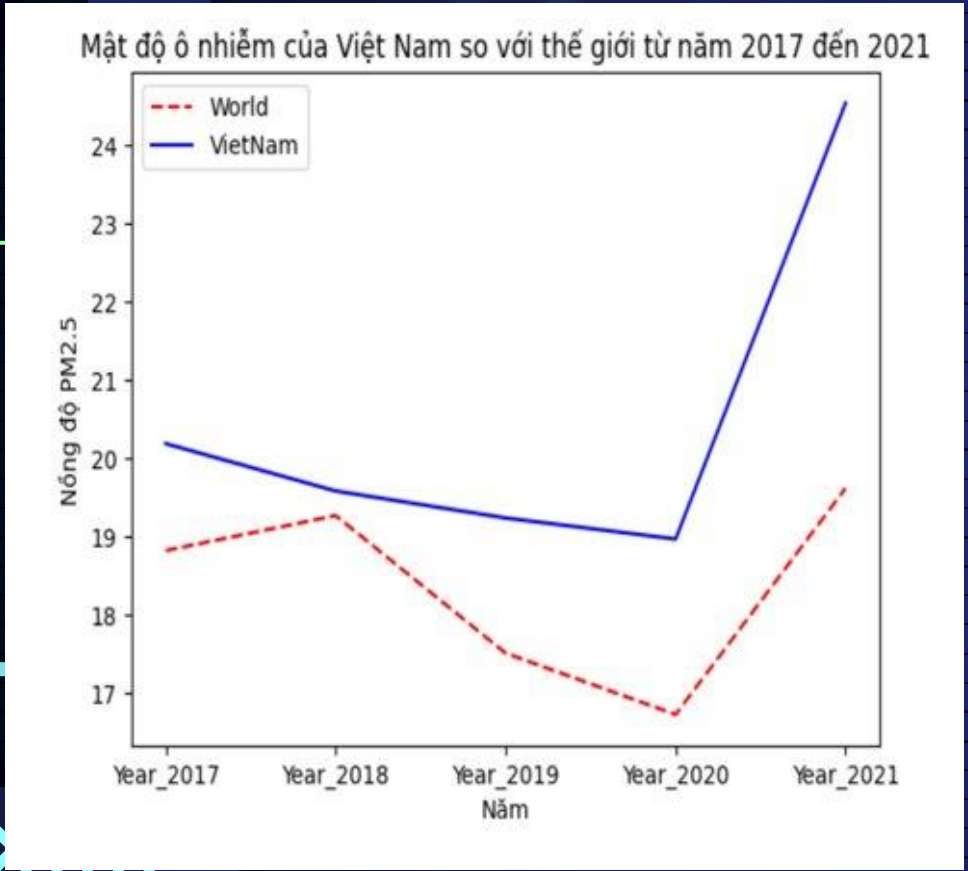
Cách thực hiện

- Nhóm dữ liệu theo 'Country' và tính mean theo country đó.
- Qua hàm get_rank để tính thứ hạng của Việt Nam qua từng năm.

Ý nghĩa

So sánh thứ hạng ô nhiễm của Việt Nam từ 2017-2021

Việt Nam so với thế giới giai đoạn 2017-2021



Cách thực hiện

- Thế giới: Từ bảng dữ liệu đã nhóm theo Country, tính mean các cột năm từ 2017-2021.
- Việt Nam: Từ bảng dữ liệu đã nhóm theo Country, lấy ra country = 'Vietnam'.

Ý nghĩa

- Sự thay đổi mật độ ô nhiễm của Việt Nam và thế giới từ 2017-2021.
- So sánh sự tương quan giữa 2 đường.

03

VIET NAM

Sự thay đổi của không khí Việt Nam từ 2017 - 2022
Địa điểm có không khí tốt nhất trong 12 tháng năm 2022

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

Sử dụng thư viện pandas, numpy, matplotlib.pyplot và seaborn

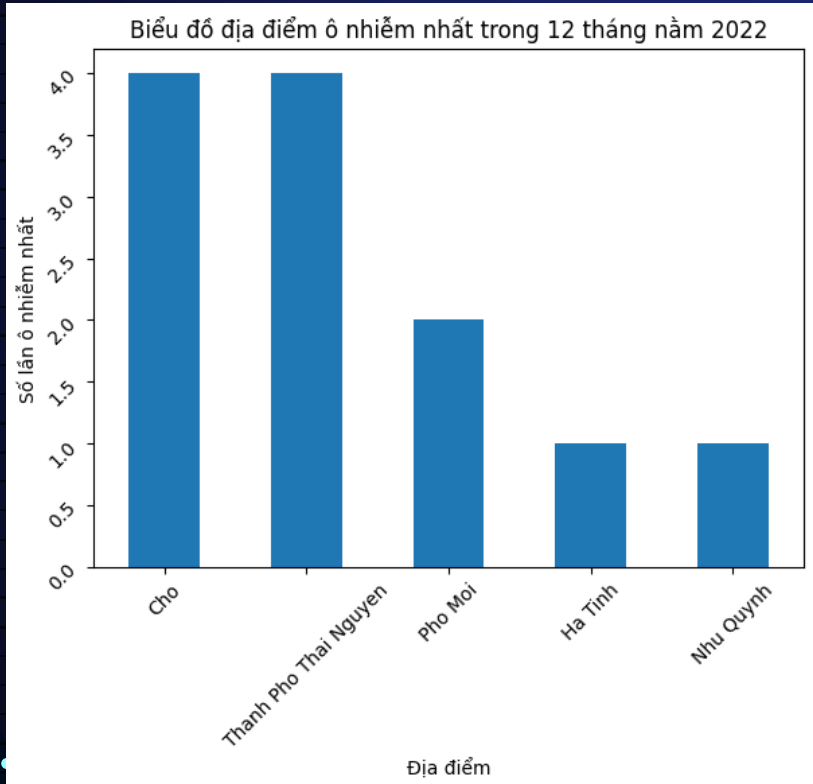
Ý tưởng

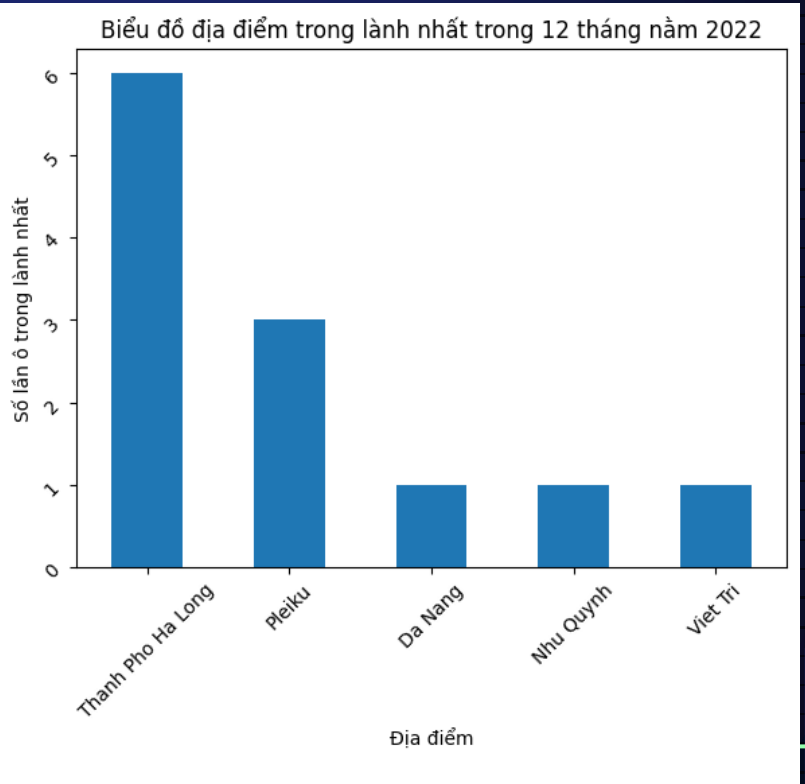
- Lấy dữ liệu của Việt Nam, thống kê theo 12 tháng của năm 2022.
- Thống kê số lần địa điểm ở Việt Nam có mức ô nhiễm lớn nhất, ít nhất.
- Tính mean 2022 và đưa ra sự thay đổi của không khí Việt Nam từ 2017 đến 2022

Kỹ thuật

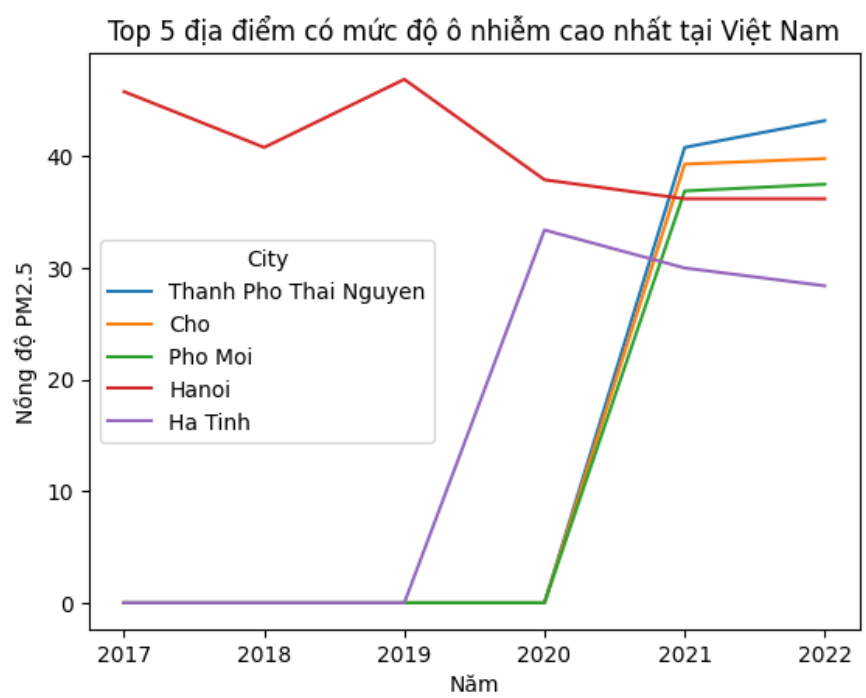
- Dùng np.percentile( ) để tính các giá trị min, max, median.
- Dùng df.drop( ) để xóa các cột dữ liệu thừa.
- Truy xuất vị trí phần tử bằng df.idmax và df.idmin.
- Vẽ trực tiếp bằng pandas, df.plot( ) thay vì vẽ bằng plt.

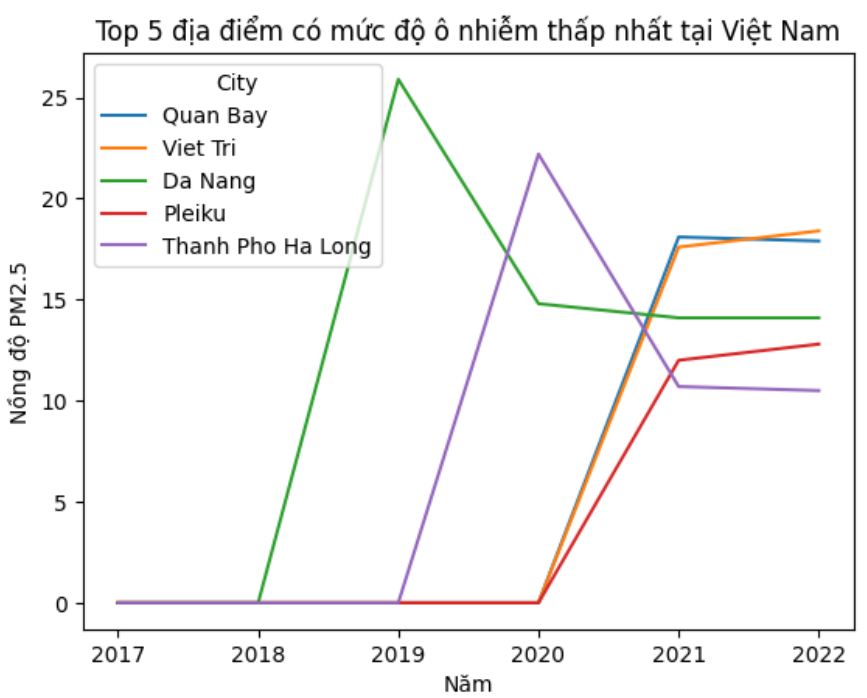
ĐỊA ĐIỂM Ô NHIỄM – TRONG LÀNH





BIẾN ĐỘNG QUA TỪNG NĂM





04

City and Country

Địa điểm, quốc gia nào có nhiều biến động về khí hậu lớn nhất trong 5 năm 2017 - 2021

ĐỊA ĐIỂM BIẾN ĐỘNG

Có thể gọi tắt là ĐIỂM BIẾN ĐỘNG, là các điểm mà tại đó có sự thay đổi đột ngột về mức độ ô nhiễm.
Ví dụ: Tháng 1, 2, 3 ở mức xanh tuy nhiên qua tháng 4 thì bỗng nhiên nhảy vọt lên mức báo động.
Từ năm 2017 – 2021, có những sự thay đổi lớn.

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

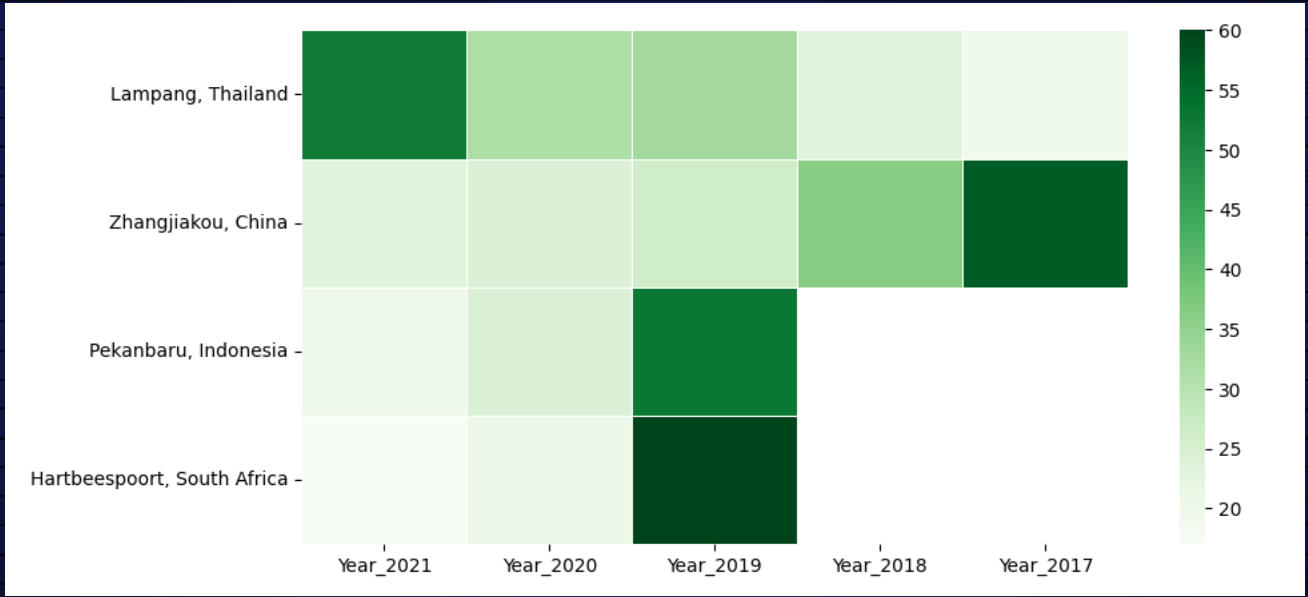
- Sử dụng thư viện pandas, numpy, matplotlib.pyplot và seaborn.

Kỹ thuật

- Lấy max, min của dữ liệu
- Trích xuất dữ liệu theo mục đích.
- Trực quan hóa dữ liệu bằng sns.heatmap( ) và plt.bar( )

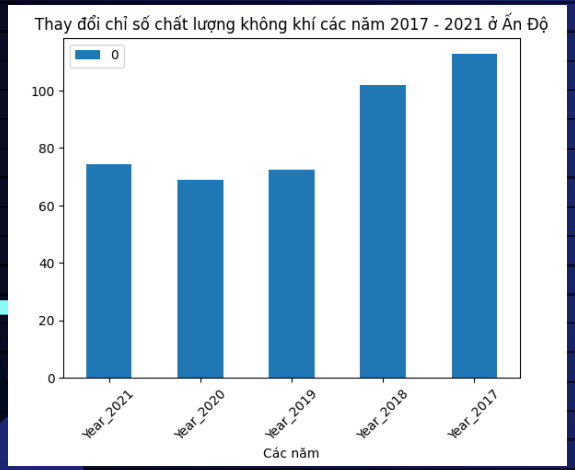
Ý tưởng và cách thực hiện

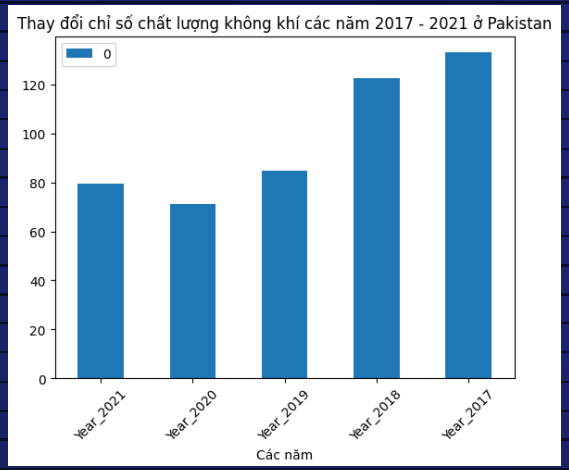
Ta lấy max, min của từng thành phố.
Sau đó, ưu tiên lấy những thành phố nào có max, min có sự thay đổi lớn. Trực quan hóa các thay đổi của thành phố đó qua các năm.

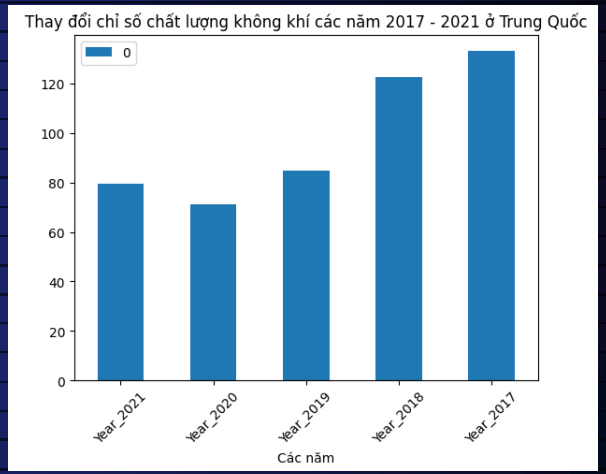


Ý tưởng và cách thực hiện

Lấy 3 quốc gia có nhiều thành phố xảy ra biến động nhất để trực quan hóa.







05

City and Country

Địa điểm, quốc gia nào có nhiều biến động về khí hậu lớn nhất trong 12 tháng năm 2022

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

- Sử dụng thư viện pandas, numpy, matplotlib.pyplot và seaborn.
- Dùng thư viện warnings để bỏ qua các cảnh báo từ hệ thống.

Kỹ thuật

- Trích xuất giá trị của df bằng thủ thuật so sánh >, <, =.
- Sắp xếp giá trị của dữ liệu theo 1 cột bằng df.sort_values( ).
- Trực quan hóa dữ liệu bằng sns.heatmap( ) và biểu diễn trực tiếp dữ liệu bằng từ df.plot.bar( )

- Dùng heatmap() để trực quan hóa sự khác biệt giữa các địa điểm trong 12 tháng.
- Ý tưởng: Lấy các địa điểm có mức độ ô nhiễm biến động đột ngột. Ví dụ như tháng [1, 2, 3] có mức độ ô nhiễm là [4, 5, 6] thì qua tháng 4, mức độ ô nhiễm tăng > 100.

OUR MODELS

NHÓM CHÚNG EM SẼ DÙNG 2 MÔ HÌNH CHÍNH

RANDOM FOREST

LOGISITC REGRESSION

DỮ LIỆU THU THẬP

	Thời tiết	Nhiệt độ	Độ ẩm	Gió	Áp suất	PM2.5	PM10	O3	NO2	SO2	CO	target
0	Trời quang	11°C	52%	8.8 km/h	1017 mbar	61.3	185.3	2	52.2	10.4	1160.0	Không lành mạnh
1	Trời quang	12°C	44%	9.4 km/h	1016 mbar	139.2	245.5	4.5	44.6	8.8	1150.0	Không lành mạnh
2	Nhiều mây	-4°C	45%	5 km/h	1036 mbar	157	416.5	9	59	27.5	NaN	Rất không lành mạnh
3	Sương mù	12°C	87%	7.4 km/h	1016 mbar	123.9	222.3	23	54.8	NaN	NaN	Không lành mạnh
4	Trời quang	12°C	56%	7.1 km/h	1016 mbar	83.3*	NaN	NaN	NaN	NaN	NaN	Không lành mạnh

Dữ liệu gồm các thông tin chi tiết về thời tiết, nhiệt độ, độ ẩm, gió, áp suất và các chất khí gây ô nhiễm ở một địa điểm nào đó được thu thập.
Cột target chính là kết quả của quá trình thu thập từ thực tế, dùng làm tập test.

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

Sử dụng thư viện pandas, numpy, datetime, time.
Đồng thời tạo một thư viện riêng là gets để xử lý các dữ liệu của trang web.

Kỹ thuật

- Tạo một thư viện riêng là gets, dùng thư viện Beautiful Soup để cào dữ liệu từ trang web về.
- Hàm main, crawl dữ liệu về và biến đổi dict thành một dataframe sau đó ghi vào file csv.
- Tên của file csv sẽ phụ thuộc vào ngay tháng mà người dùng crawl dữ liệu, bằng hàm today( ).
- Tạo file bash để người dùng chỉ cần click choột và chờ đợi dữ liệu crawl về cho mình.

RANDOM FOREST VS. LOGISTIC REGRESSION

	precision	recall	f1-score	support
1.0	1.00	0.99	0.99	629
2.0	0.97	0.99	0.98	347
3.0	0.87	0.99	0.93	90
4.0	0.98	0.97	0.97	115
5.0	0.67	0.15	0.25	13
6.0	0.71	0.50	0.59	10
accuracy			0.97	1204
macro avg	0.87	0.76	0.78	1204
weighted avg	0.97	0.97	0.97	1204

Random Forest

	precision	recall	f1-score	support
1.0	0.93	0.98	0.95	629
2.0	0.87	0.87	0.87	347
3.0	0.70	0.66	0.68	90
4.0	0.68	0.53	0.60	115
5.0	0.56	0.38	0.45	13
6.0	0.38	0.50	0.43	10
accuracy			0.87	1204
macro avg	0.69	0.65	0.66	1204
weighted avg	0.86	0.87	0.87	1204

Logistic Regression

Random Forest cho thấy độ hiệu quả cao hơn so với Logistic Regression.
Tuy nhiên với độ chính xác là 0.97/1.0 thì Random Forest có khả năng bị Over Fitting

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

Thư viện chung: pandas, numpy, warnings, matplotlib, seaborn.
Thư viện của marchine learning: sklearn. Cùng các class như: impute, model_selection, metrics

1. Random Forest: Sử dụng thư viện RandomForestClassifier của sklearn.
2. Logistic Regression: Sử dụng thư viện LogisticRegression của sklearn.

Kỹ thuật

- Dùng pd.concat để nối các file input là một df được crawl theo từng ngày.
- Thay thế các giá trị là string thành một con số cụ thể để máy học.
- Chỉnh sửa các giá trị dư thừa của dữ liệu, đưa toàn bộ dữ liệu về np.float64.
- Thay np.nan bằng các giá trị mean của từng cột.

THƯ VIỆN VÀ KỸ THUẬT

Thư viện

Thư viện chung: pandas, numpy, warnings, matplotlib, seaborn.
Thư viện của marchine learning: sklearn. Cùng các class như: impute, model_selection, metrics

1. Random Forest: Sử dụng thư viện RandomForestClassifier của sklearn.
2. Logistic Regression: Sử dụng thư viện LogisticRegression của sklearn.

Kỹ thuật

- Đối với các mô hình thì đều dùng tập test có tỉ lệ 70% so với bộ dữ liệu.
- Dùng Confusion Matrix để kiểm chứng tính đúng sai giữa 2 label.
- Dự đoán các giá trị precision, recall, f1-score và support nhờ hàm metrics.classification_report( )