BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC THĂNG LONG**

Logo

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN MÔN**

**BIG DATA**

**(Học kỳ I nhóm 2 năm học 2022 – 2023)**

|  |  |
| --- | --- |
| **GIÁO VIÊN HƯỚNG DẪN**  **Vũ Minh Hoàng** | **SINH VIÊN THỰC HIỆN**  A41461 Phương Hải Đăng  A42048 Nguyễn Đức Thái  A38146 Nguyễn Minh Hiếu  A40985 Bùi Tiến Dũng |

Hà Nội - 2023

**Mục Lục**

[CHƯƠNG 1. Mục Đích 1](#_Toc148424941)

[CHƯƠNG 2. Các công cụ sử dụng 2](#_Toc148424942)

[2.1 Ngôn ngữ lập trình python 2](#_Toc148424943)

[2.2 Thư viện pyspark 3](#_Toc148424944)

[2.3 Thư viện Matplotlib 4](#_Toc148424945)

[2.4 Cartopy 5](#_Toc148424946)

[CHƯƠNG 3. Phân Tích dữ Liệu 6](#_Toc148424947)

[3.1 Nhập các thư viện cần sử dụng và dữ liệu 6](#_Toc148424948)

[3.2 Thu thập dữ liệu : 7](#_Toc148424949)

[3.3 Tổng quan về dữ liệu 9](#_Toc148424950)

[3.4 Xử lý dữ liệu 11](#_Toc148424951)

[3.5 Khám phá dữ liệu : 14](#_Toc148424952)

[3.6 Xây dựng giả thuyết và kiểm định giả thuyết 23](#_Toc148424953)

[3.6.1 Giả thuyết 1 : Nhiệt độ trung bình toàn cầu tăng cao hơn từ 2020 đến nay 23](#_Toc148424954)

[3.6.2 Giả thuyết 2 : Nhiệt độ bán cầu bắc nóng hơn bán cầu nam 26](#_Toc148424955)

[3.6.3 Giả thuyết 3: Các thành phố gần biển có nhiệt độ trung bình thấp hơn, tốc độ gió cao hơn so với các thành phố sâu trong đất liền. 28](#_Toc148424956)

[3.6.4 Giả thuyết 4 : Mùa hè nhiều mưa, bão nhất 29](#_Toc148424957)

[3.6.5 Giả thuyết 5: Sáng sớm (0h-6h) lạnh hơn so với buổi tối(18h-23h59) 33](#_Toc148424958)

[CHƯƠNG 4. Kết luận 36](#_Toc148424959)

**Lời Mở Đầu**

Trong bối cảnh ngày càng phức tạp và thách thức của thời tiết hiện nay, không thể xem thường vai trò quan trọng của dữ liệu thời tiết. Dữ liệu thời tiết, với sự tích lũy không ngừng và sự phong phú ngày càng tăng, đã trở thành một nguồn tài nguyên vô cùng quý giá. Nó giống như một kho tàng không giới hạn, đang đợi chúng ta khám phá và sử dụng. Cách chúng ta khai thác dữ liệu thời tiết này sẽ quyết định sự chuẩn bị và ứng phó của chúng ta trước các tình huống khó khăn và biến đổi trong điều kiện khí hậu.

Dữ liệu thời tiết không chỉ là nguồn thông tin mà còn là một công cụ biến đổi cách chúng ta dự đoán và ứng phó với thời tiết, quản lý nguồn lực và tương tác với môi trường xung quanh. Điều này đặt ra một loạt thách thức phức tạp, từ việc dự báo thời tiết chính xác, quản lý sự biến đổi trong điều kiện khí hậu và tìm hiểu cách tận dụng thông tin thời tiết trong nhiều khía cạnh của cuộc sống, từ nông nghiệp đến an toàn hàng hải và nguồn nước.

Trong ngữ cảnh này, việc thu thập, phân tích và tận dụng dữ liệu thời tiết là quan trọng để đảm bảo an toàn và hiệu suất cho nhiều lĩnh vực. Dữ liệu thời tiết không chỉ ảnh hưởng đến việc dự đoán mưa rào ngày mai mà còn quyết định sự thành công của ngành nông nghiệp trong việc lên kế hoạch cho mùa vụ, quản lý tài nguyên nước và điều hành an toàn hàng hải trong điều kiện biển động.

Hy vọng rằng việc nghiên cứu và sử dụng dữ liệu thời tiết này sẽ giúp cung cấp thông tin chi tiết và giá trị cho các tổ chức và cá nhân để chuẩn bị và ứng phó tốt hơn với biến đổi thời tiết và các tình huống liên quan. Đồng thời, nó cũng có tiềm năng để cải thiện hiệu suất và quản lý tài nguyên, đóng góp vào sự bền vững của nhiều lĩnh vực trong xã hội. Dữ liệu thời tiết không chỉ là một khía cạnh kỹ thuật mà còn là một phần không thể thiếu của cuộc sống hàng ngày và quản lý toàn cầu.

Cuối cùng, chúng tôi muốn bày tỏ lòng biết ơn đối với sự hỗ trợ và hướng dẫn quý báu từ thầy giáo Vũ Minh Hoàng, người đã đồng hành cùng chúng tôi trong việc tạo ra báo cáo này.

# Mục Đích

Trong thời đại số hóa hiện nay, việc thu thập và phân tích dữ liệu thời tiết đã trở nên vô cùng quan trọng đối với nhiều lĩnh vực của cuộc sống, từ nông nghiệp đến du lịch và ngành công nghiệp năng lượng. Phân tích dữ liệu thời tiết không chỉ giúp dự đoán và ứng phó với các biến đổi khí hậu, mà còn mang lại lợi ích kinh tế và an toàn cho xã hội.

Mục tiêu của phân tích dữ liệu thời tiết rất đa dạng và hữu ích. Trước hết, nó giúp dự báo thời tiết một cách chính xác hơn. Bằng cách sử dụng dữ liệu về áp suất không khí, nhiệt độ, độ ẩm, và gió, chúng ta có thể dự đoán được các biến đổi trong thời tiết, bao gồm cả cơn bão, lũ lụt, và biến đổi khí hậu dài hạn. Điều này rất quan trọng cho việc cảnh báo và ứng phó kịp thời, giúp bảo vệ cuộc sống và tài sản của mọi người.

Ngoài ra, phân tích dữ liệu thời tiết còn giúp nâng cao hiệu suất trong các lĩnh vực như nông nghiệp. Thông qua việc theo dõi dữ liệu thời tiết, nông dân có thể quyết định thời điểm tốt nhất để gieo trồng, tưới nước, và thu hoạch sản phẩm. Điều này giúp tối ưu hóa sản xuất nông nghiệp, giảm thiểu lãng phí tài nguyên và tăng cường năng suất.

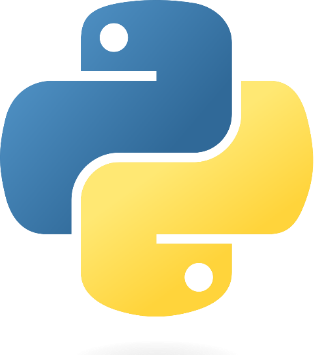
Trong ngành du lịch, phân tích dữ liệu thời tiết cũng đóng một vai trò quan trọng. Bằng cách dự đoán thời tiết, các nhà tổ chức du lịch có thể điều chỉnh lịch trình và hoạt động dựa trên dự báo. Điều này giúp du khách tránh mất thời gian và tiền bạc vào những ngày mưa, và tận hưởng những điểm đến tốt nhất dựa trên thời tiết.

Trong lĩnh vực năng lượng, phân tích dữ liệu thời tiết hỗ trợ việc quản lý và tối ưu hóa nguồn cung cấp năng lượng. Dự báo thời tiết chính xác giúp điều chỉnh việc sản xuất năng lượng điện, quản lý mạng lưới điện hiệu quả, và giảm thiểu nguy cơ cảnh báo sự cố.

Tóm lại, phân tích dữ liệu thời tiết không chỉ cung cấp thông tin về dự báo thời tiết, mà còn hỗ trợ quyết định và tối ưu hóa trong nhiều lĩnh vực khác nhau. Nó giúp bảo vệ cuộc sống và tài sản, tối ưu hóa sản xuất nông nghiệp, cải thiện trải nghiệm du lịch, và quản lý nguồn cung cấp năng lượng một cách hiệu quả. Đây là một ví dụ rõ ràng về cách dữ liệu và phân tích dữ liệu đã trở thành một phần không thể thiếu của cuộc sống hiện đại.

# Các công cụ sử dụng

## Ngôn ngữ lập trình python



Python, một ngôn ngữ lập trình đa năng, đã ra đời dưới sự sáng tạo của Guido van Rossum và chính thức ra mắt vào năm 1991. Từ đó, Python đã nhanh chóng trở thành một trong những ngôn ngữ lập trình phổ biến nhất trên toàn cầu, với một cộng đồng lập trình đa dạng và đông đảo. Dưới đây, chúng ta sẽ điểm qua một số điểm đáng chú ý về Python:

* Cú Pháp Dễ Đọc: Python được thiết kế với cú pháp sáng sủa và dễ đọc, giống như ngôn ngữ tự nhiên. Điều này giúp người mới học lập trình nắm bắt kiến thức một cách nhanh chóng. Cú pháp của Python sử dụng dấu tab để xác định các khối code thay vì dấu ngoặc như trong nhiều ngôn ngữ lập trình khác.
* Đa Năng: Python có khả năng sử dụng cho nhiều mục đích khác nhau. Nó phù hợp cho việc phát triển ứng dụng web, ứng dụng di động, ứng dụng máy học, phân tích dữ liệu, và nhiều ứng dụng khác. Điều này làm cho Python trở thành một công cụ linh hoạt cho các nhà phát triển.
* Thư Viện Phong Phú: Python có một hệ sinh thái thư viện phong phú và đa dạng. Các thư viện như NumPy cho tính toán số học, Pandas cho xử lý dữ liệu, Matplotlib cho vẽ biểu đồ, và TensorFlow/PyTorch cho máy học và trí tuệ nhân tạo giúp việc phát triển các ứng dụng phức tạp trở nên dễ dàng hơn.
* Hỗ Trợ Cộng Đồng: Python có một cộng đồng lớn và nhiệt tình. Bất kỳ khiếu nại hay thắc mắc nào có thể được giải quyết thông qua diễn đàn, blog, và các nguồn học tập trực tuyến. Cộng đồng này thường giúp các lập trình viên giải quyết vấn đề và chia sẻ kiến thức.

Python đã tỏ ra không chỉ là một ngôn ngữ lập trình mạnh mẽ, mà còn là một ngôn ngữ thân thiện với người dùng, giúp họ dễ dàng thực hiện các dự án và nghiên cứu trong nhiều lĩnh vực khác nhau. Điều này đã đóng góp một phần quan trọng vào sự phổ biến và thành công của Python trong cộng đồng lập trình toàn cầu.

## Thư viện pyspark

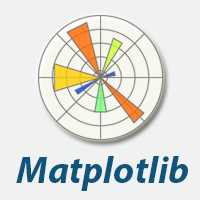


Pyspark, viết tắt của "Python for Apache Spark," là một thư viện mạnh mẽ dựa trên Apache Spark, một framework mã nguồn mở hàng đầu trong việc xử lý và phân tích dữ liệu trực tuyến lớn (big data). Dưới đây, chúng ta sẽ khám phá một số đặc điểm quan trọng của Pyspark:

* Xử Lý Dữ Liệu Trực Tuyến Lớn: Pyspark được đặc biệt thiết kế để đối phó với việc xử lý dữ liệu lớn. Nó cho phép bạn thực hiện các phân tích phức tạp, biến đổi và xử lý dữ liệu với dung lượng lớn, vượt trội so với các thư viện và công cụ thông thường.
* Tích Hợp Dễ Dàng với Python: Một điểm mạnh lớn của Pyspark là khả năng tích hợp mạnh mẽ với ngôn ngữ lập trình Python. Điều này có nghĩa là bạn có thể sử dụng các thư viện Python phổ biến như NumPy, Pandas và Matplotlib trong quá trình làm việc với dữ liệu trong môi trường Spark.
* Xử Lý Dữ Liệu Đa Dạng: Pyspark hỗ trợ xử lý dữ liệu đa dạng, bao gồm dữ liệu cấu trúc, bán cấu trúc và dữ liệu văn bản. Điều này làm cho nó trở thành lựa chọn ưu việt cho các dự án liên quan đến dữ liệu có đa dạng định dạng.
* Tối Ưu Hóa Tự Động: Apache Spark, và do đó Pyspark, có khả năng tự động tối ưu hóa quá trình xử lý dữ liệu. Tối ưu hóa này bao gồm việc phân phối dữ liệu và tính toán trên nhiều máy tính, giúp cải thiện hiệu suất xử lý dữ liệu một cách đáng kể.
* Hỗ Trợ Đa Nền Tảng: Pyspark có khả năng hoạt động trên nhiều nền tảng khác nhau, bao gồm Hadoop, Apache Mesos và Kubernetes. Điều này cho phép bạn triển khai ứng dụng Spark trên các hệ thống khác nhau một cách linh hoạt.

Pyspark là một công cụ mạnh mẽ cho việc xử lý và phân tích dữ liệu lớn, và sự kết hợp giữa tính dễ sử dụng của Python và khả năng xử lý mạnh mẽ của Apache Spark đã tạo ra một lựa chọn hấp dẫn cho các dự án dữ liệu và big data.

## Thư viện Matplotlib

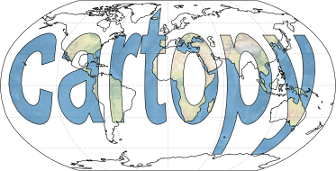


Matplotlib, thư viện mã nguồn mở mạnh mẽ cho việc tạo ra đồ họa và biểu đồ trong Python, là một công cụ quan trọng cho việc trực quan hóa dữ liệu và phân tích. Dưới đây, chúng ta sẽ điểm qua một số điểm quan trọng về thư viện này:

* Đa Dạng Về Biểu Đồ: Matplotlib cung cấp khả năng tạo ra nhiều loại biểu đồ khác nhau, bao gồm biểu đồ đường, biểu đồ cột, biểu đồ phân tán, biểu đồ hình ảnh, biểu đồ 3D, và nhiều loại biểu đồ khác. Điều này cho phép bạn hiển thị dữ liệu một cách linh hoạt và thích hợp cho nhiều nguồn dữ liệu khác nhau.
* Tích Hợp Với Python: Matplotlib tích hợp tốt với Python và hoạt động tương thích với nhiều thư viện và framework khác trong hệ sinh thái khoa học dữ liệu của Python như NumPy, Pandas và SciPy. Điều này giúp bạn làm việc với dữ liệu và tạo biểu đồ dễ dàng.
* Linh Hoạt và Tùy Chỉnh: Matplotlib cho phép bạn tùy chỉnh mọi khía cạnh của biểu đồ, từ màu sắc, kích thước, đánh dấu trục, đến tiêu đề và chú thích. Điều này giúp bạn tạo ra biểu đồ theo ý muốn và phù hợp với nhu cầu của dự án.
* Khả Năng Xuất Bản: Bạn có thể xuất biểu đồ Matplotlib dưới dạng hình ảnh chất lượng cao hoặc PDF để sử dụng trong báo cáo, trang web hoặc sách. Biểu đồ tạo ra bằng Matplotlib thích hợp cho mục đích xuất bản và trình bày dữ liệu một cách chuyên nghiệp.
* Cộng Đồng Lớn và Hỗ Trợ: Matplotlib có một cộng đồng đông đảo và nhiệt tình. Bạn có thể tìm kiếm sự hỗ trợ, tài liệu học tập và ví dụ trực tuyến dễ dàng khi gặp vấn đề hoặc cần sự hướng dẫn. Điều này giúp bạn giải quyết các thách thức và tận dụng tối đa tiềm năng của Matplotlib.

Matplotlib là một công cụ quan trọng cho việc biểu diễn và trực quan hóa dữ liệu trong Python, và sự đa dạng, tích hợp dễ dàng và khả năng tùy chỉnh đã làm cho nó trở thành lựa chọn phổ biến trong cộng đồng khoa học dữ liệu và phân tích dữ liệu.

## Cartopy



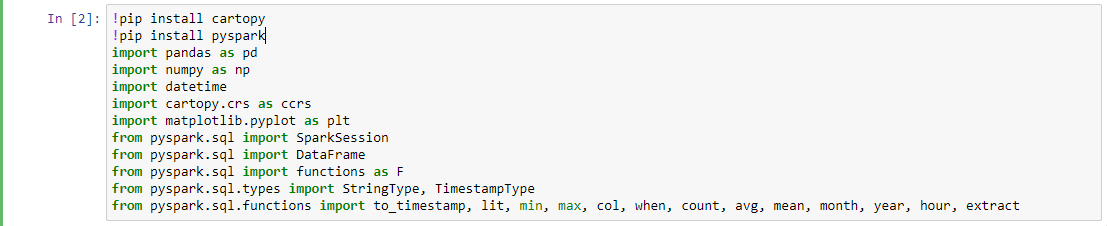
Cartopy là một thư viện Python mạnh mẽ được sử dụng để xử lý và vẽ dữ liệu địa lý. Dưới đây, chúng ta sẽ điểm qua một số điểm đáng chú ý về thư viện Cartopy:

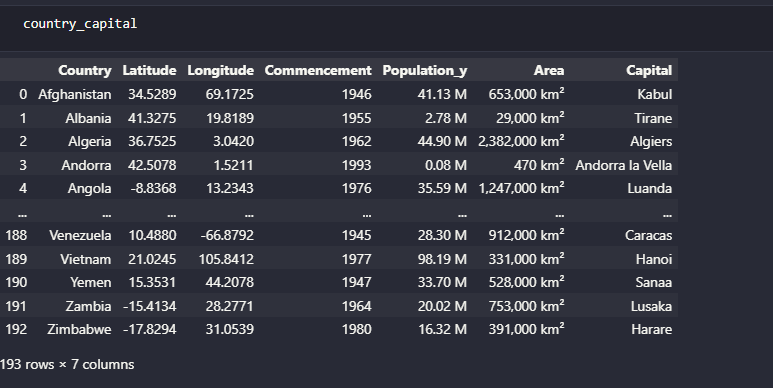
* Xử lý Dữ liệu Địa lý: Cartopy cung cấp các công cụ mạnh mẽ cho việc xử lý dữ liệu địa lý. Nó cho phép bạn dễ dàng thao tác với dữ liệu liên quan đến bản đồ, hệ tọa độ địa lý, và các đối tượng địa lý như hình dạng, ranh giới, và lưới mạng.
* Vẽ Bản Đồ Địa lý: Cartopy cho phép bạn vẽ bản đồ địa lý theo nhiều kiểu khác nhau. Bạn có thể vẽ bản đồ tọa độ, biểu đồ địa lý, ranh giới quốc gia, và nhiều loại biểu đồ địa lý khác. Điều này rất hữu ích cho việc hiển thị dữ liệu địa lý một cách trực quan.
* Hỗ trợ Hệ tọa độ Địa lý Đa dạng: Cartopy hỗ trợ nhiều hệ tọa độ địa lý khác nhau như Mercator, Plate Carrée, và Lambert Conformal. Điều này giúp bạn làm việc với dữ liệu địa lý từ nhiều nguồn khác nhau một cách dễ dàng.
* Hệ sinh thái Thư viện Mạnh mẽ: Cartopy tích hợp chặt chẽ với các thư viện khác như Matplotlib, NumPy, và SciPy. Điều này giúp bạn tạo ra các biểu đồ địa lý phức tạp và thực hiện các phân tích địa lý một cách hiệu quả.
* Hỗ Trợ Cộng Đồng: Cartopy cũng có một cộng đồng lập trình đang phát triển và hỗ trợ. Bất kỳ khiếu nại hay thắc mắc nào có thể được giải quyết thông qua diễn đàn, blog, và các nguồn học tập trực tuyến. Cộng đồng này thường giúp các lập trình viên giải quyết vấn đề và chia sẻ kiến thức.

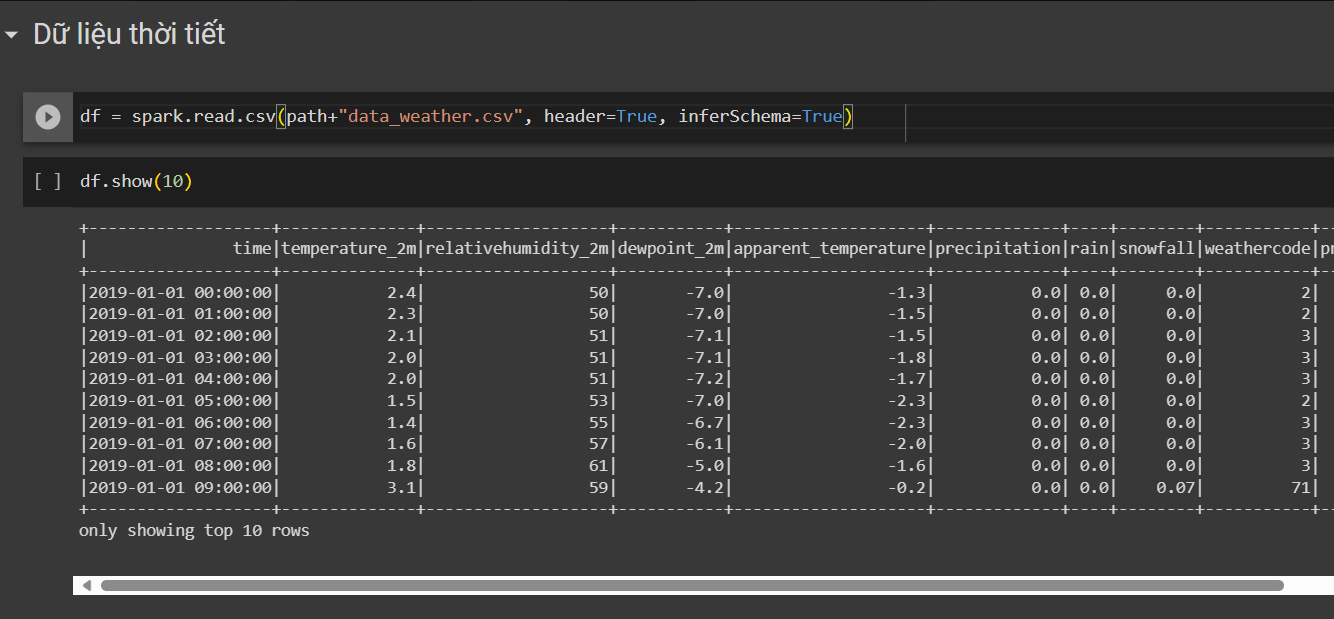
Cartopy đóng góp quan trọng vào việc xử lý và hiển thị dữ liệu địa lý một cách dễ dàng và hiệu quả trong ngôn ngữ lập trình Python. Thư viện này giúp các nhà khoa học, nhà phân tích dữ liệu, và lập trình viên thực hiện các dự án liên quan đến địa lý một cách hiệu quả và chất lượng.

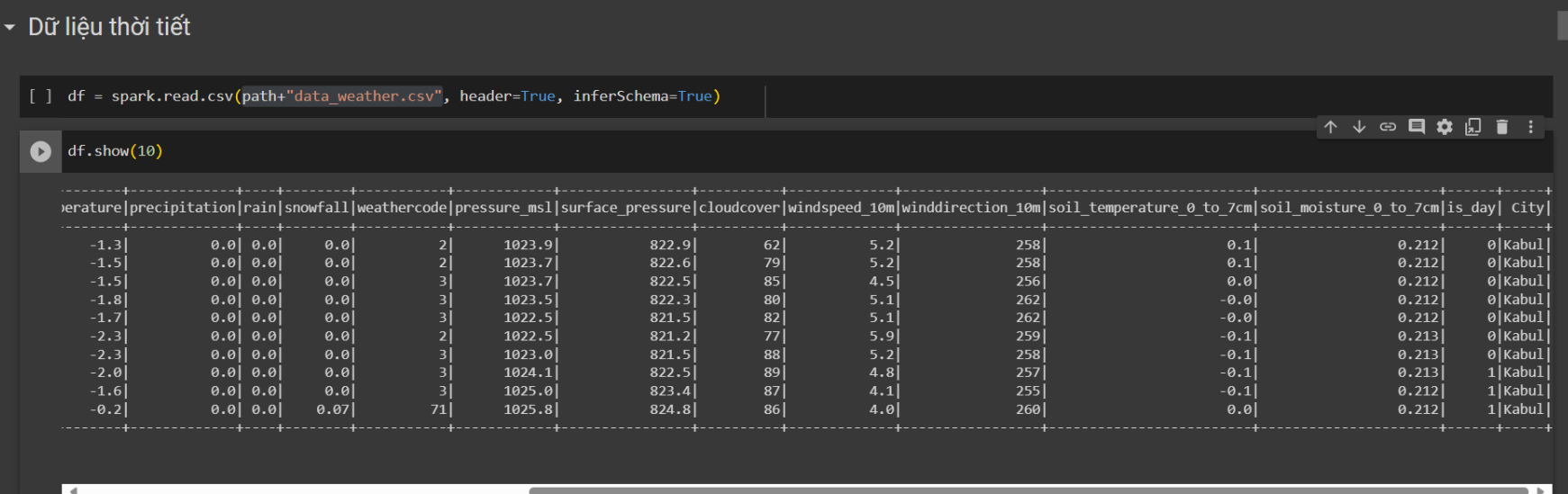
# Phân Tích dữ Liệu

## Nhập các thư viện cần sử dụng và dữ liệu







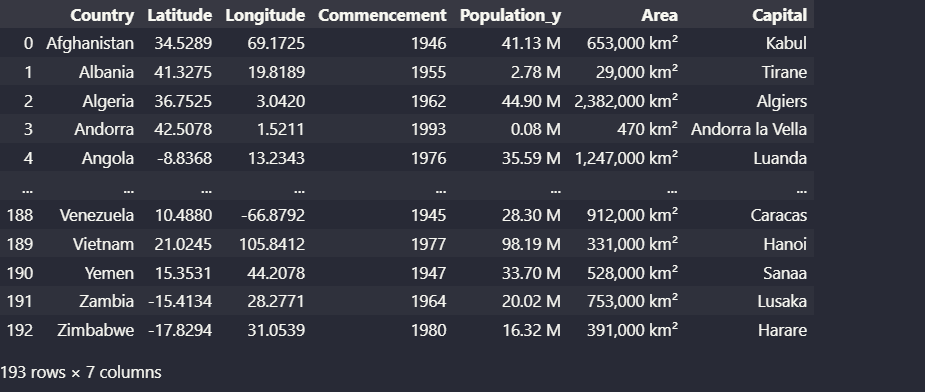


## Thu thập dữ liệu :

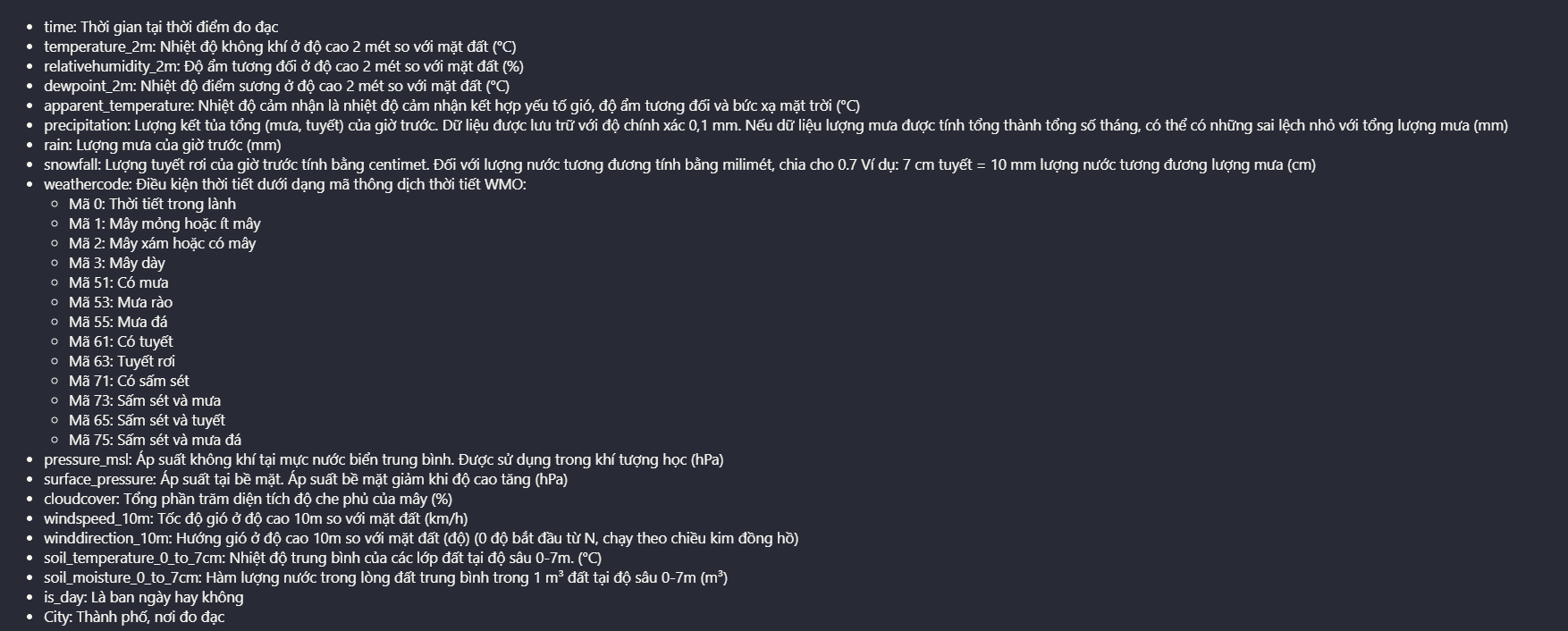
**Nhận xét** : Vì tên thủ đô nhiều nước bị sai nên tìm lại dữ liệu chứa tên thủ đô chính xác và tiếp tục nối

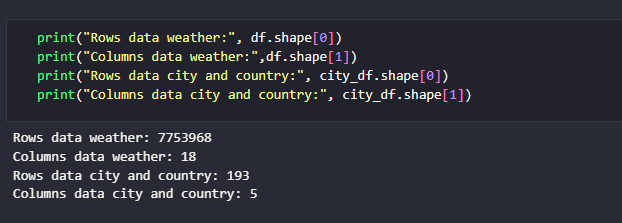




* Country: Tên quốc gia thuộc Liên Hợp Quốc
* Latitude: Vĩ độ của thủ đô
* Longitude: Kinh độ của thủ đô
* Commencement: Năm gia nhập tổ chức
* Population\_y: Dân số quốc gia (triệu người)
* Area: Diện tích lãnh thổ (km^2)
* Capital: Thủ đô của quốc gia đó

## Tổng quan về dữ liệu





Dữ liệu là một bảng dữ liệu có cấu trúc chứa thông tin về thời tiết. Dữ liệu bao gồm 7753968 dòng, 18 cột và 11424 dòng chứa giá trị bị thiếu.

**Sơ lược về thông tin các cột:**

- time: Thời gian tại thời điểm đo đạc

- temperature\_2m: Nhiệt độ không khí ở độ cao 2 mét so với mặt đất (°C)

- relativehumidity\_2m: Độ ẩm tương đối ở độ cao 2 mét so với mặt đất (%)

- dewpoint\_2m: Nhiệt độ điểm sương ở độ cao 2 mét so với mặt đất (°C)

- apparent\_temperature: Nhiệt độ cảm nhận là nhiệt độ cảm nhận kết hợp yếu tố gió, độ ẩm tương đối và bức xạ mặt trời (°C)

- precipitation: Lượng kết tủa tổng (mưa, tuyết) của giờ trước. Dữ liệu được lưu trữ với độ chính xác 0,1 mm. Nếu dữ liệu lượng mưa được tính tổng thành tổng số tháng, có thể có những sai lệch nhỏ với tổng lượng mưa (mm)

- rain: Lượng mưa của giờ trước (mm)

- snowfall: Lượng tuyết rơi của giờ trước tính bằng centimet. Đối với lượng nước tương đương tính bằng milimét, chia cho 0.7 Ví dụ: 7 cm tuyết = 10 mm lượng nước tương đương lượng mưa (cm)

- weathercode: Điều kiện thời tiết dưới dạng mã thông dịch thời tiết WMO:

+ Mã 0: Thời tiết trong lành

+ Mã 1: Mây mỏng hoặc ít mây

+ Mã 2: Mây xám hoặc có mây

+ Mã 3: Mây dày

+ Mã 51: Có mưa

+ Mã 53: Mưa rào

+ Mã 55: Mưa đá

+ Mã 61: Có tuyết

+ Mã 63: Tuyết rơi

+ Mã 71: Có sấm sét

+ Mã 73: Sấm sét và mưa

+ Mã 65: Sấm sét và tuyết

+ Mã 75: Sấm sét và mưa đá

- pressure\_msl: Áp suất không khí tại mực nước biển trung bình. Được sử dụng trong khí tượng học (hPa)

- surface\_pressure: Áp suất tại bề mặt. Áp suất bề mặt giảm khi độ cao tăng (hPa)

- cloudcover: Tổng phần trăm diện tích độ che phủ của mây (%)

- windspeed\_10m: Tốc độ gió ở độ cao 10m so với mặt đất (km/h)

- winddirection\_10m: Hướng gió ở độ cao 10m so với mặt đất (độ) (0 độ bắt đầu từ N, chạy theo chiều kim đồng hồ)

- soil\_temperature\_0\_to\_7cm: Nhiệt độ trung bình của các lớp đất tại độ sâu 0-7m. (°C)

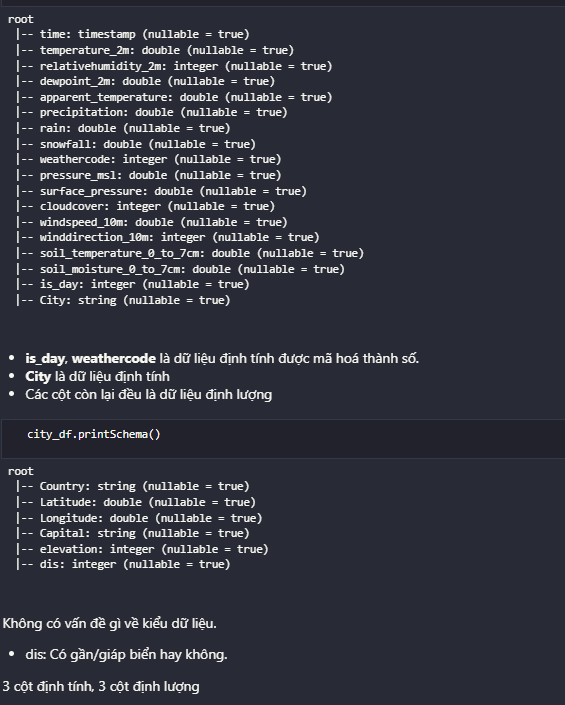
- soil\_moisture\_0\_to\_7cm: Hàm lượng nước trong lòng đất trung bình trong 1 m³ đất tại độ sâu 0-7m (m³)

- is\_day: Là ban ngày hay không

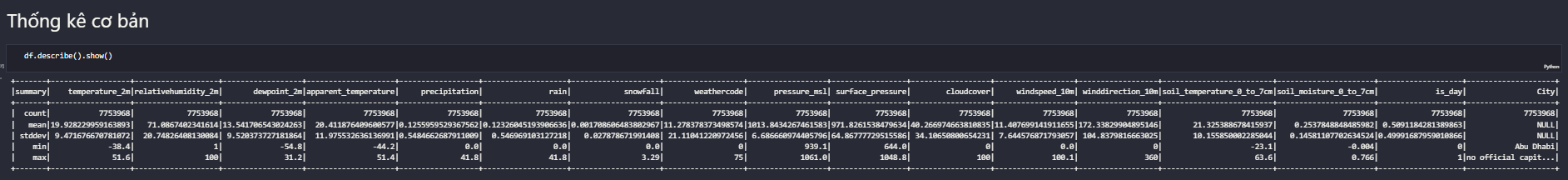
- City: Thành phố, nơi đo đạc

## Xử lý dữ liệu

**Kiểu dữ liệu :**



**Thống kê cơ bản :**



**Nhận xét :**

- Dựa vào \*count\*, không có giá trị ở cột nào bị thiếu

- Dựa vào \*min,max\*, không có thuộc tính đơn vị % nào có giá trị âm và giá trị vượt quá 100%. Tuy nhiên cột \*\*soil\_moisture\_0\_to\_7cm\*\* có min mang giá trị âm. Cần tìm hiểu giá trị này, liệu có phải bất thường?

- 2 cột \*\*temperature\*\* có độ trải lớn do sự khác biệt về mùa và về vị trí của thành phố vùng cực, vùng xích đạo. Cần kiểm tra xem liệu có phải là giá trị outlier. Trung bình, trung vị gần tương đương nhau. Tuy nhiên độ lệch chuẩn của \*\*apparent\_temperature\*\* cao hơn \*\*temperature\_2m\*\*, cho thấy dữ liệu của \*\*apparent\_temperature\*\* tập trung vào điểm trung bình ít hơn so với \*\*temperature\_2m\*\*

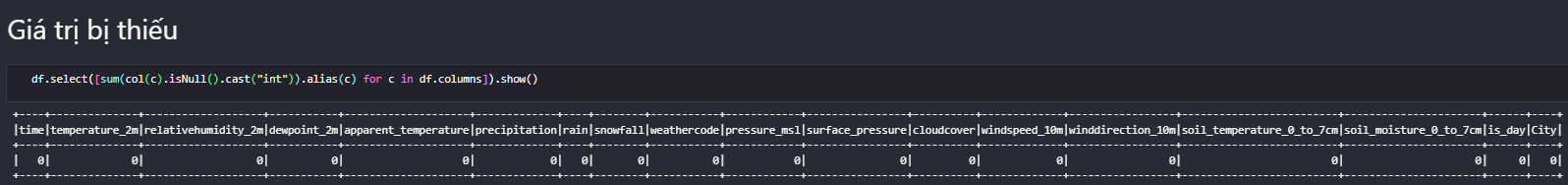
- \*\*relativehumidity\_2m\*\*: Dựa vào tứ phân vị thứ nhất (60%) và mean (71), std (20.7), các thành phố đa phần ở những khó xảy ra hạn hán (mức đề phòng hạn hán là ~< 39%).

- Dựa theo tứ phân vị thứ 3 của các cột \*\*precipitation, rain, snowfall\*\* cho thấy thời tiết đa phần không có mưa hay tuyết.

- \*\*weathercode, is\_day\*\* trên thực tế là giá trị định tính nên các giá trị thống kê trên không có ý nghĩa.

- \*\*windspeed\_10m\*\* có max cao bất thường. Cần kiểm tra liệu có phải outlier hay do bão.

- Chỉ có giá trị max ở cột Capital bất thường. Cần kiểm tra tên các thành phố, thủ đô ở cả 2 bảng dữ liệu



**Nhận xét :**

- Có thể thấy không có giá trị nào bị null.

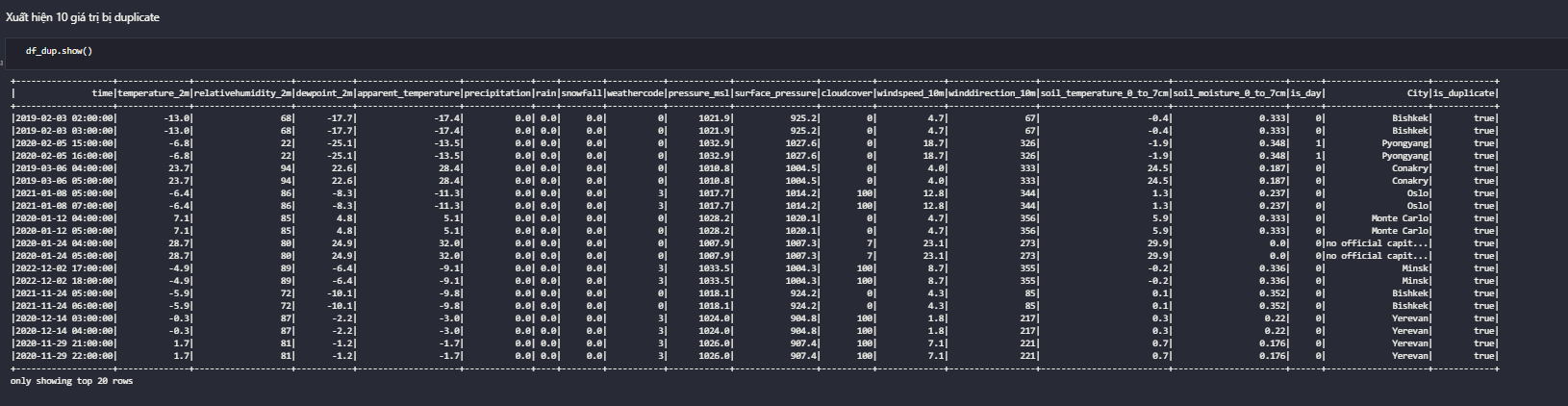
- Vì kiểu dữ liệu đa phần là kiểu số nên có thể dữ liệu null được thay thế bằng giá trị thay thế hoặc 0. Tuy nhiên khả năng gán 0 cho giá trị null khó có thể xảy ra do 0 cũng là giá trị có ý nghĩa, vì vậy người hoặc máy nhập liệu sẽ không gán 0 thay cho giá trị null. Còn lại thông thường sẽ chọn 1 giá trị ngoài khoảng giá trị của các cột (chẳng hạn cột về % từ 0-100, null sẽ thay thế bằng -1), tuy nhiên khi describe không thấy xuất hiện giá trị min hay max nào bất thường.

- Cần lưu ý \*\*time\*\* liệu có mang 1 giá trị nào thay thế giá trị null (Vì đang là kiểu object) Nếu có lỗi trong quá trình chuyển thành datetime thì đó chính là giá trị thay thế null

- Cột \*\*City\*\* không thể null do trong quá trình nối bảng đều có thêm giá trị của các thành phố

**Trùng lặp** :



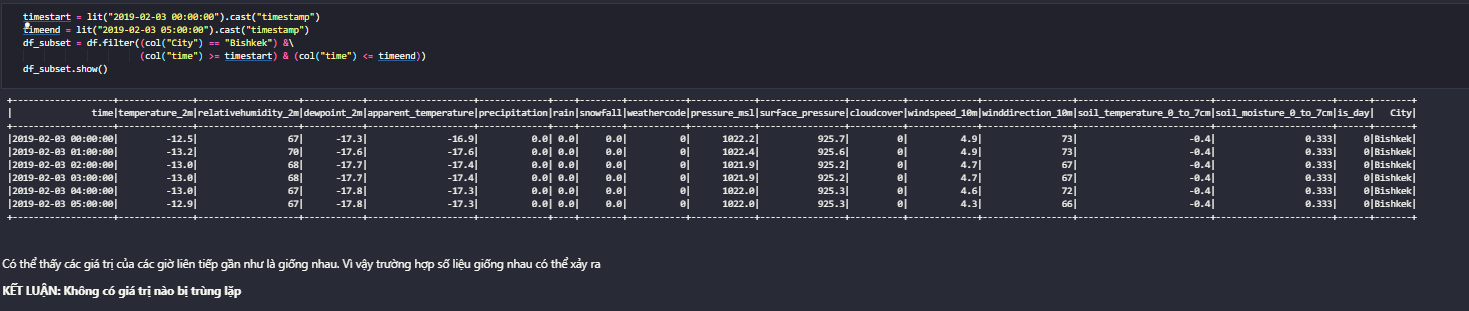


Có các cặp thời gian gần nhau (cách nhau 1h) và cùng thành phố. Số liệu có thể giống nhau do:

+ Số liệu 2 giờ thực sự giống nhau

+ Bản ghi có thời gian muộn hơn chưa kịp update số liệu nên đã fill method = "bfill"

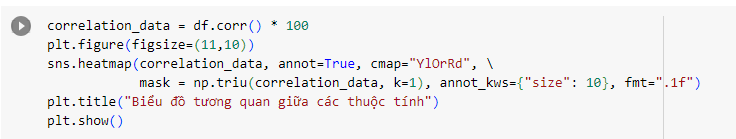
Kiểm tra trường hợp đầu tiên: Số liệu giống nhau thật hay không

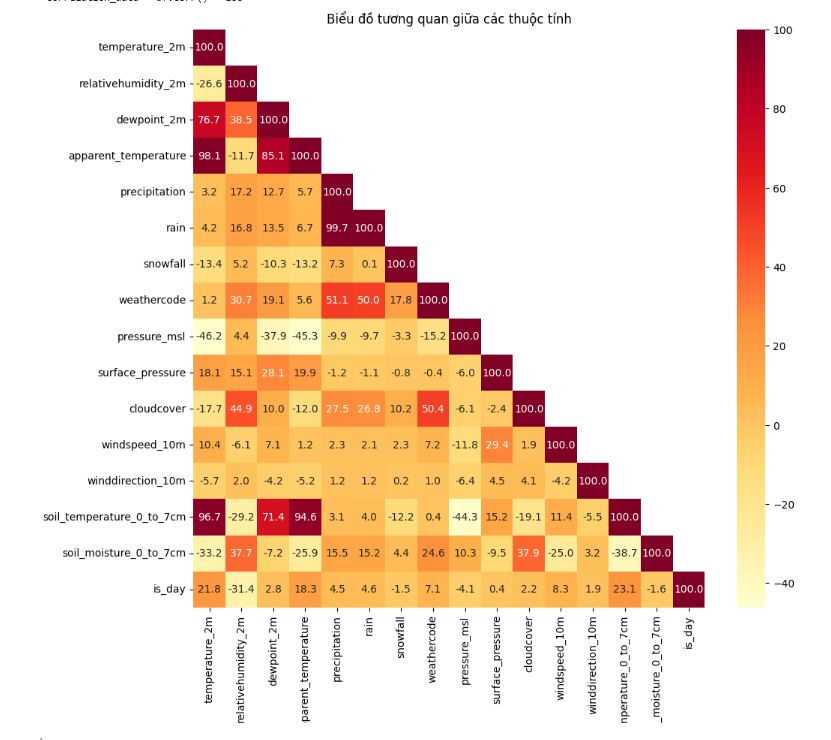


Nhận xét : Có thể thấy các giá trị của các giờ liên tiếp gần như là giống nhau. Vì vậy trường hợp số liệu giống nhau có thể xảy ra

KẾT LUẬN: Không có giá trị nào bị trùng lặp

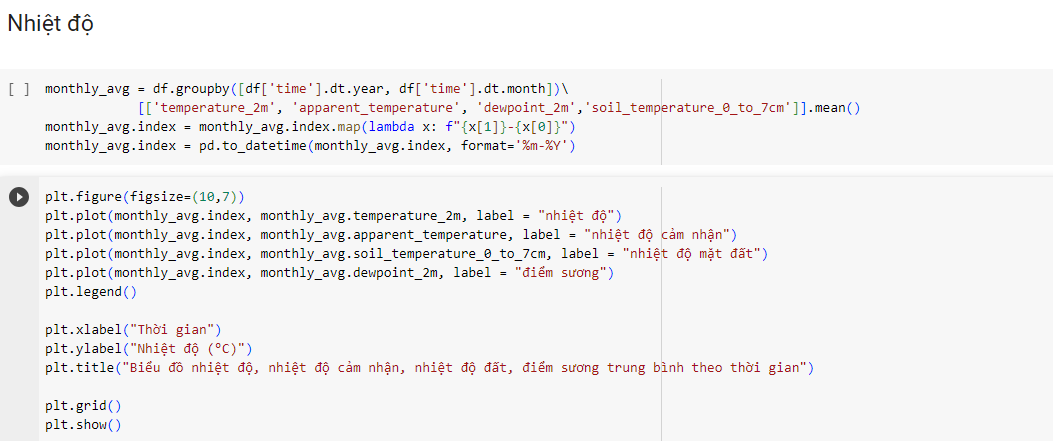
## Khám phá dữ liệu :

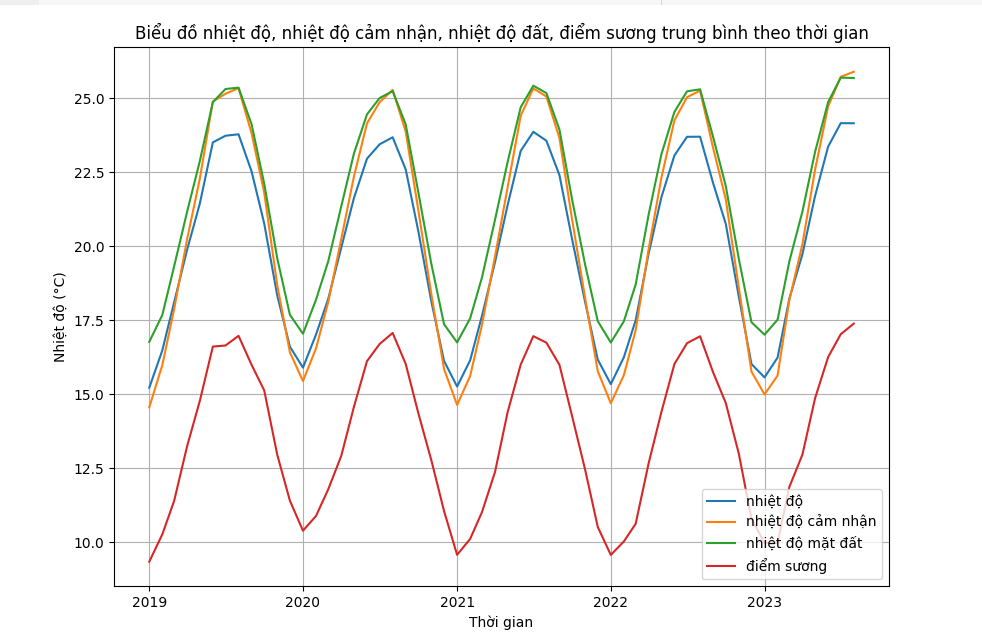


****

**Nhận xét :** Có thể thấy temperature\_2m, apparent\_temperature, dewpoint\_2m và soil\_temperature\_0\_to\_7cm có tương quan với nhau rất cao. Tất cả đều liên quan đến nhiệt độ và có mối quan hệ với nhau.

Độ che phủ mây cũng ảnh hưởng tới một phần thời tiết và độ ẩm**.**

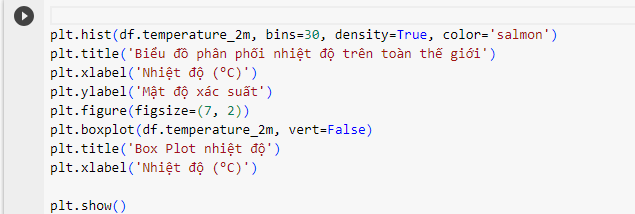
****

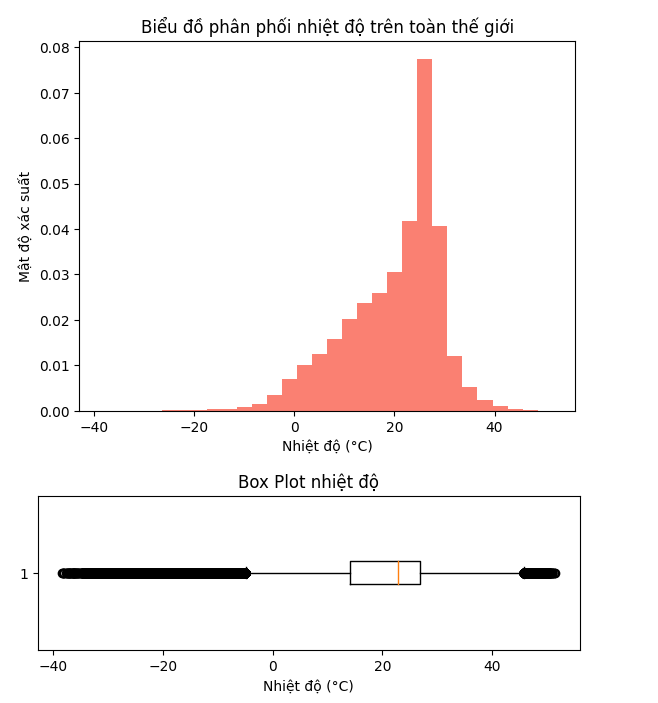
****

**Nhận xét :**

* Nhiệt độ theo từng năm đều có 1 chu trình là lạnh vào mùa xuân, ấm dần và nóng vào giữa năm (tháng 6- tháng 8) rồi lạnh dần đến mùa đông.
* Năm 2023 có sự chênh lệch về nhiệt độ, nhiệt độ cảm nhận cao nhất tức rõ rệt hơn hẳn so với các năm còn lại
* Nhiệt độ mặt đất đa phần nóng hơn so với nhiệt độ cảm nhận, nhiệt độ, điểm sương. Vào mùa đông, nhiệt độ có thể chênh lệch khoảng 2°C so với nhiệt độ ngoài trời. Vì vậy một vài con vật biết tận dụng việc ở trong hang vào kì ngủ đông để giữa được nhiệt tốt hơn.
* Nhiệt độ điểm sương lúc nào cũng thấp hơn so với 3 loại nhiệt độ trên khoảng 5-8 °C
* Có thể thấy nhiệt độ cảm nhận dao động mạnh hơn so với nhiệt độ thực tế, vì vậy:
* Vào mùa hè, chúng ta sẽ cảm thấy nóng hơn khoảng 1 - 1.5 °C so với nhiệt độ dự báo
* Vào mùa đông, chúng ta sẽ cảm thấy lạnh hơn khoảng 0.5°C so với nhiệt độ dự báo

**Lưu ý :** rằng điều này còn phụ thuộc vào lượng hơi ẩm trong không khí.

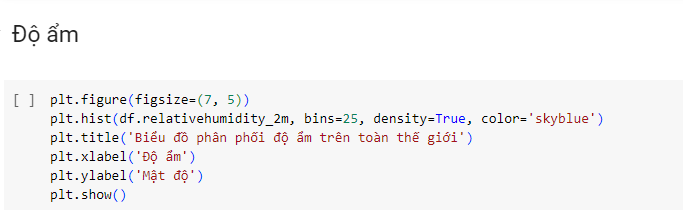


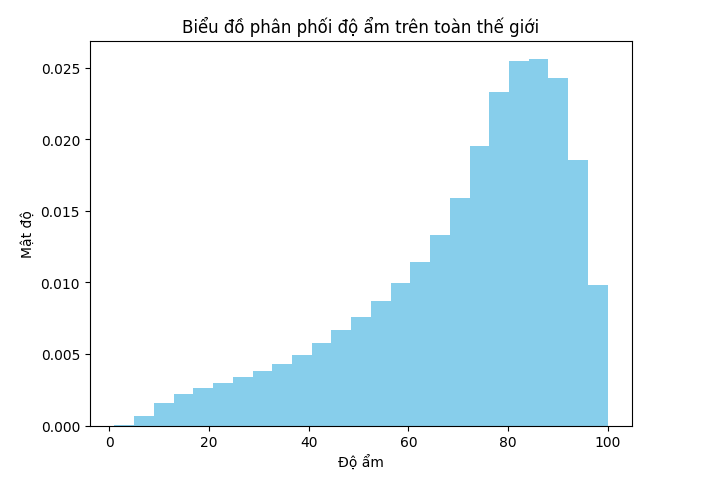


**Nhận xét :**

* Biểu đồ bị lệch trái. Mode khoảng 25 độ C
* Nhiệt độ cao đa phần trong khoảng 0 đến 38
* Nhiệt độ trên 42 và dưới -20 rất ít.

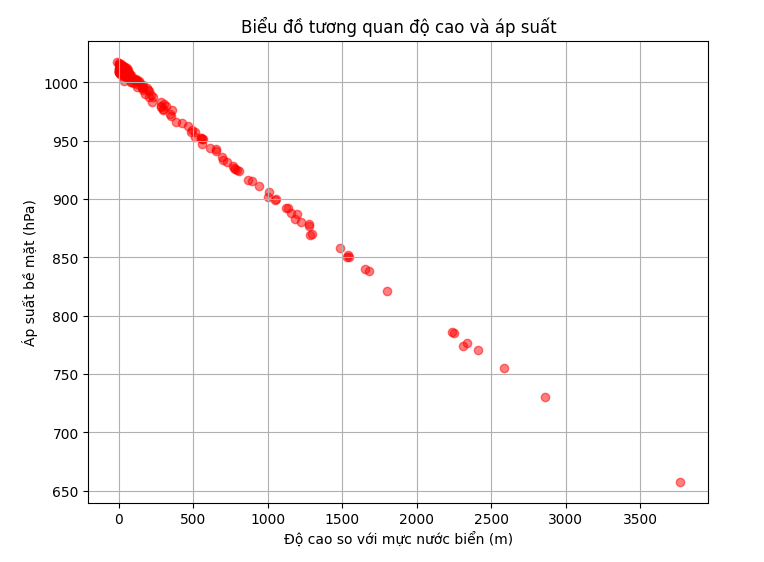
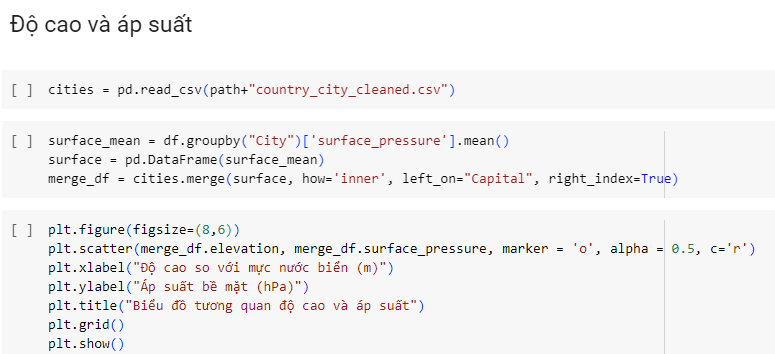
**Như vậy có thể** thấy các thành phố đa phần được xây dựng ở những nơi có điều kiện khí hậu tốt, có mức nhiệt phù hợp để sinh sống.

****

****

**Nhận xét :**

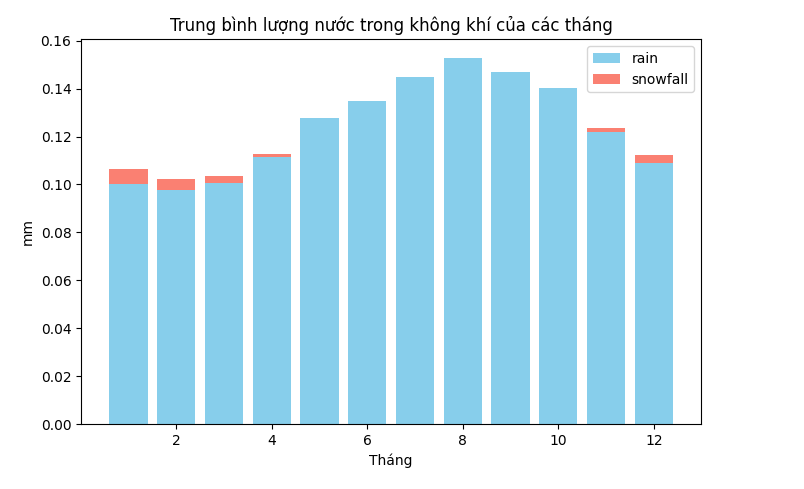
* Biểu đồ lệch trái. Đa phần độ ẩm đều cao, giá trị mode khoảng 85%
* Độ ẩm như trên phù hợp với điều kiện sống của con người. Tuy nhiên nếu cao hoặc thấp quá đều không tốt. Nên để mức ổn định 40-80%



**Nhận xét** :

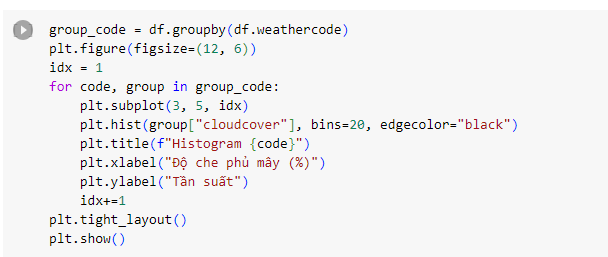
* Càng lên cao, áp suất càng giảm. Độ tương quan giữa 2 đại lượng này rất cao. Có thể thấy gần như là tuyến tính.
* Ngoài ra có thể thấy càng lên cao, số lượng thành phố càng ít đi. Các thành phố lớn đa phần tập trung ở đồng bằng. Có nhiều nguyên nhân khiến ít thành phố lớn ở trên cao, chủ yếu là do điều kiện khí hậu, con đường vận chuyển nhu yếu phẩm và các thànhh phố xung quanh, áp suất không khí thấp thì tức nồng độ oxi trong không khí cũng loãng hơn.

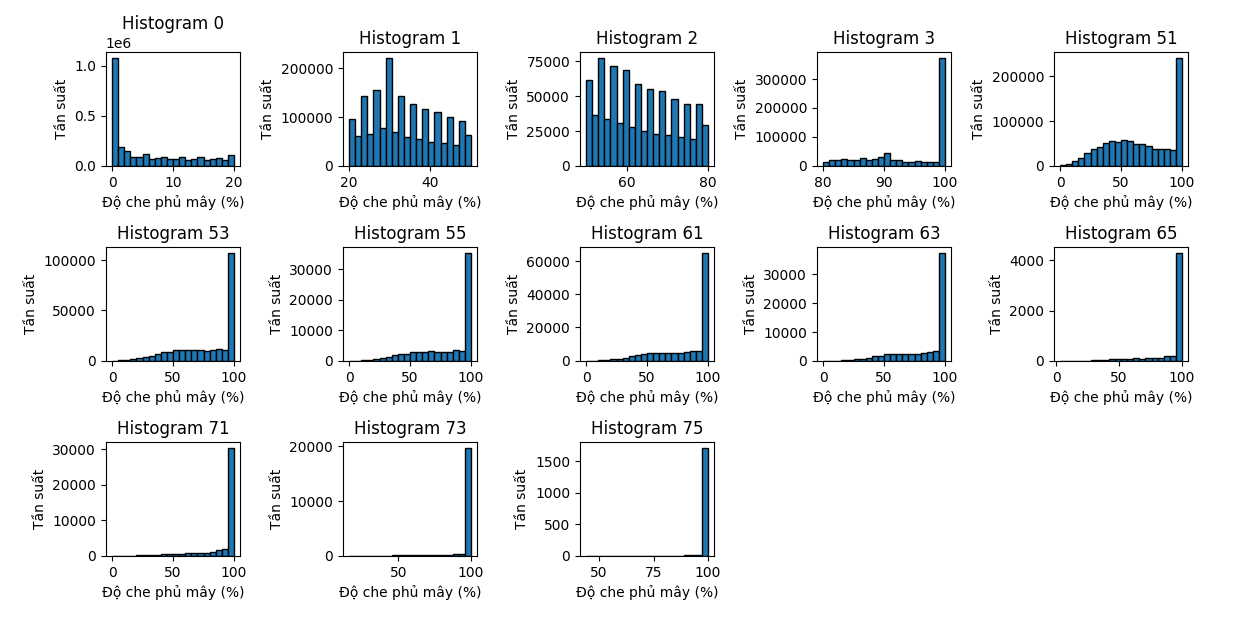




**Nhận xét** : Tháng 7- tháng 10 có lượng nước nhiều nhất trong năm. Đa phần vào mùa hè nên không có tuyết. Tuyết đa phần xuất hiện vào tháng 11 đến cuối tháng 4.

**Từ biểu đồ trên** cũng có thể thấy đa phần lượng nước trong không khí đến từ mưa là chủ yếu. Lượng tuyết rơi không nhiều.

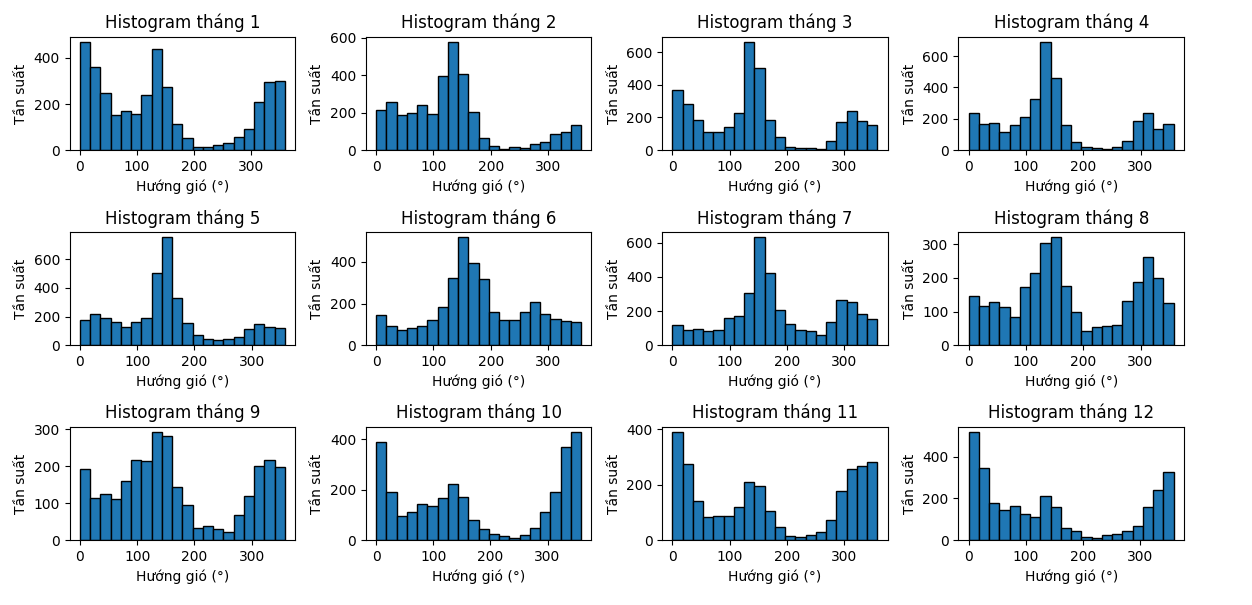




**Nhận xét** : Có thể thấy từ mã 3 trở đi, chủ yếu cloudcover đều là 100%. Tức lúc mưa, mây sẽ che phủ kín bầu trời.

Từ đây có thể biết rằng mỗi khi mây dày, khả năng cao sẽ mưa.





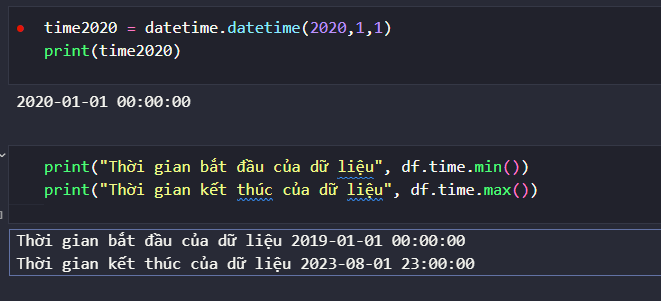
Nhận xét: Có sự thay đổi của hướng gió chủ yếu qua các tháng:

* Từ tháng 2 đến tháng 7 chủ yếu hướng gió ~150°, tức hướng gió ĐÔNG NAM
* Từ tháng 8, bắt đầu xuất hiện gió có hướng ~300° mạnh hơn, tức từ phía TÂY BẮC
* Đến tháng 10, gió ĐÔNG NAM suy yếu, gió hướng TÂY BẮC dần chuyển hướng thành gió từ phía Bắc, trong khoảng 350° và 0°.
* Từ tháng 11 đến tháng 12, xuất hiện thêm các hướng gió từ phía ĐÔNG BẮC đến hết tháng 1.
* Tháng 1 bắt đầu hình thành lại gió ĐÔNG NAM.

## Xây dựng giả thuyết và kiểm định giả thuyết

### Giả thuyết 1 : Nhiệt độ trung bình toàn cầu tăng cao hơn từ 2020 đến nay

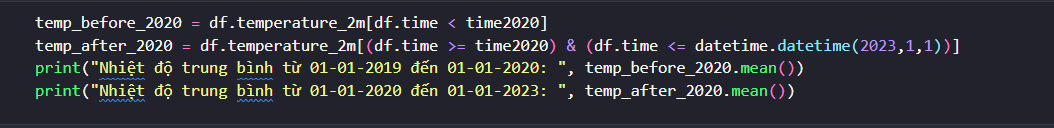
Nhiệt độ trung bình toàn cầu được đo bằng nhiệt độ thực tế (temperature\_2m) chứ không phải nhiệt độ cảm nhận (apparent\_temperature)



Thời gian bắt đầu của dữ liệu 2019-01-01 00:00:00

Thời gian kết thúc của dữ liệu 2023-08-01 23:00:00

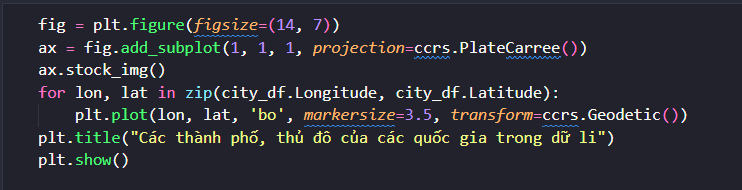
Nhận xét: Để chính xác hơn, chỉ cần kiểm tra đến 01/01/2023. Bởi đến 01/08/2023 vẫn chưa phải mùa đông. Mùa đông nhiệt độ thấp hơn sẽ khiến nhiệt độ trung bình sẽ giảm xuống 1 chút.

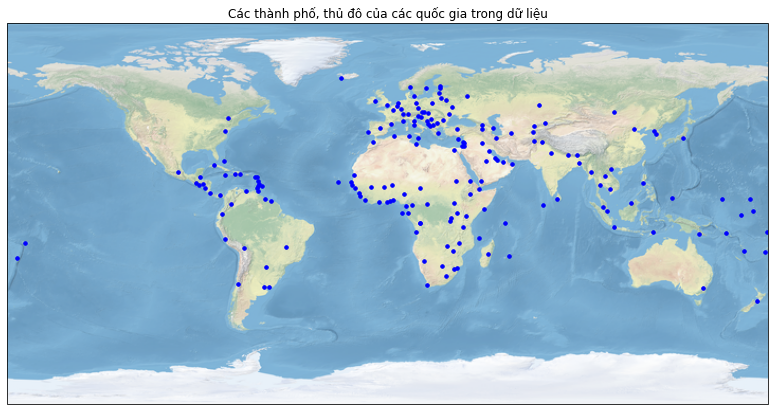


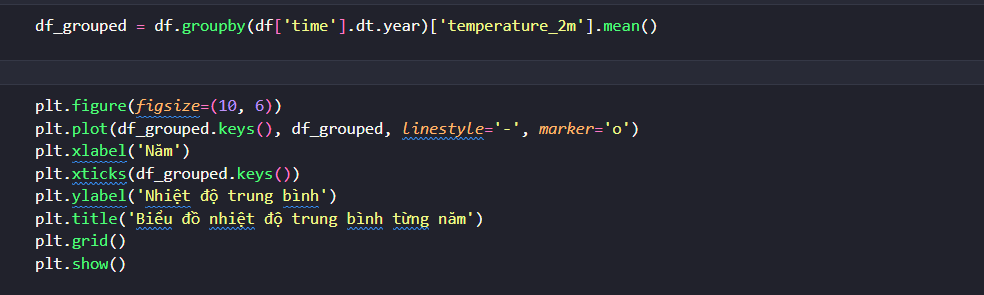
Nhiệt độ trung bình từ 01-01-2019 đến 01-01-2020: 20.045268176119503

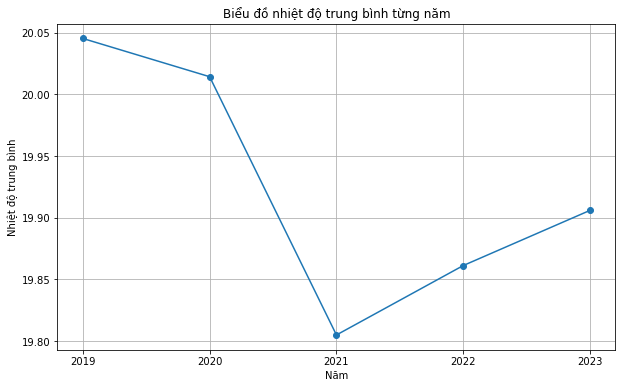
Nhiệt độ trung bình từ 01-01-2020 đến 01-01-2023: 19.893404788191134

**Nhận xét**: Nhiệt độ trung bình sau năm 2020 thấp hơn trước năm 2020. Giả thuyết có thể sai. Ngoài ra 2 nhiệt độ đều cao hơn so với nhiệt độ trung bình toàn cầu thực tế bởi chưa có dữ liệu của các vùng xung quanh vòng cực 2 bán cầu.



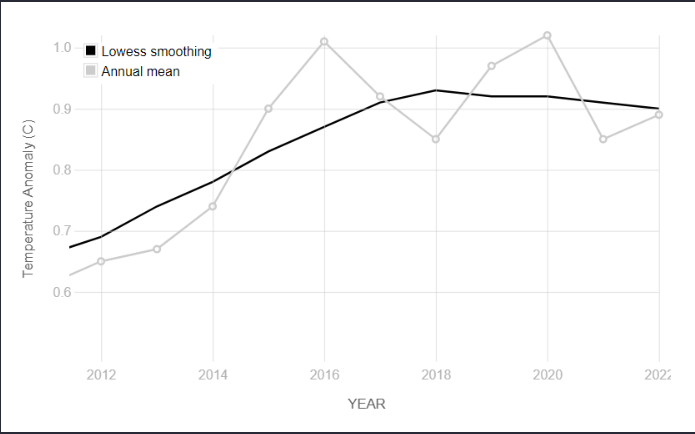






Nhận xét : Biểu đồ có xu hướng giảm từ năm 2020 trở đi và có xu hướng tăng trở lại

Theo Nasa, biểu đồ dưới đây cho thấy sự thay đổi về nhiệt độ bề mặt trung bình toàn cầu so với mức trung bình dài hạn từ 1951 tới 1980

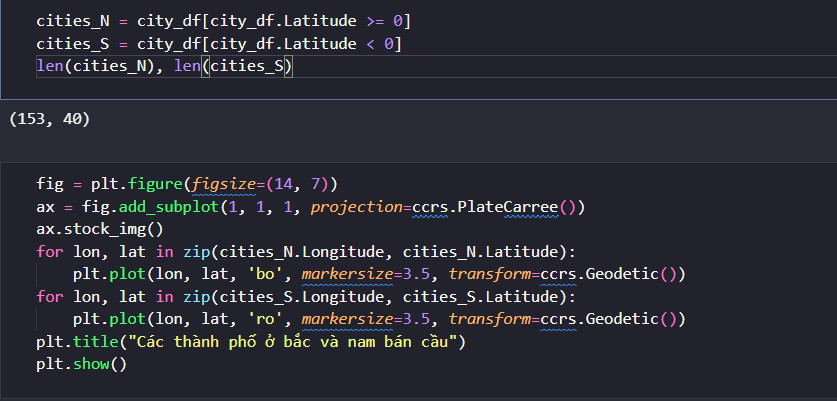


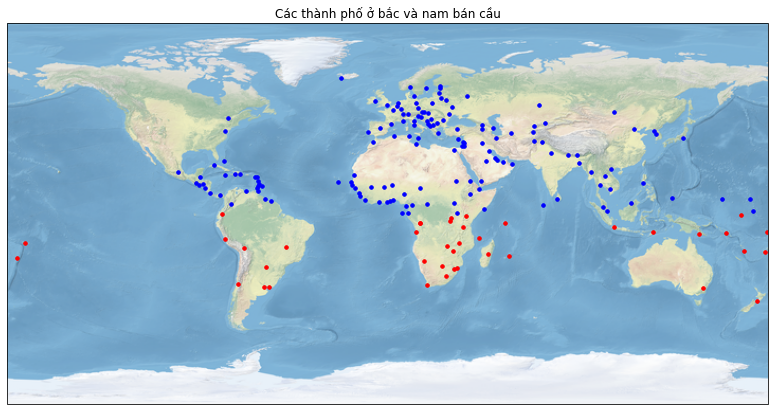
**Nhận xét** : Có thể giải thích bởi sự hoạt động công nghiệp, phương tiện giao thông chững lại do dịch Covid 19. Lượng khí thải giảm mạnh khiến nhiệt độ trung bình giảm đi một chút so với năm 2020.

* **GIẢ THUYẾT 1 SAI**

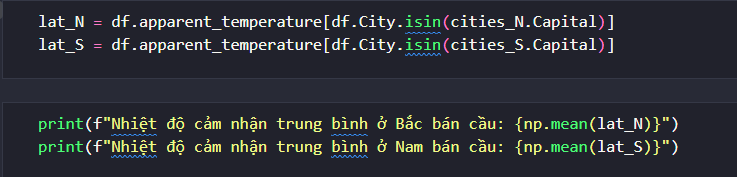
### Giả thuyết 2 : Nhiệt độ bán cầu bắc nóng hơn bán cầu nam

Cảm giác nóng lạnh dựa trên nhiệt độ cảm nhận sẽ chính xác hơn



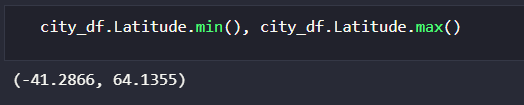


Nhận xét: Số thành phố ở Bắc bán cầu gấp gần 4 lần số thành phố ở Nam bán cầu.



Nhiệt độ cảm nhận trung bình ở Bắc bán cầu: 19.89575734415625

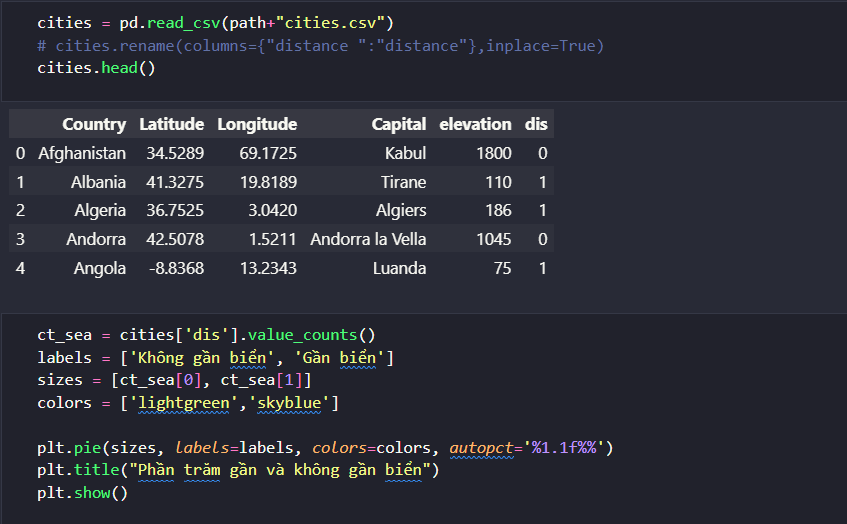
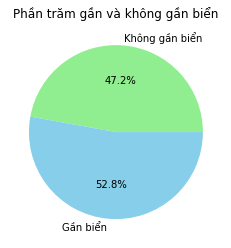
Nhiệt độ cảm nhận trung bình ở Nam bán cầu: 22.386031834925763



Giải thích: Do hầu hết lục địa tập trung về Bắc bán cầu, vì vậy dân cư chủ yếu tập trung ở Bắc bán cầu nhiều hơn. Vì vậy các thành phố cũng nhiều hơn so với phía Nam. Từ bản đồ, các thành phố gần cực Nam rất ít. Vì vậy ảnh hưởng nhiều đến nhiệt độ trung bình giữa các thành phố ở 2 bán cầu.

* **GIẢ THUYẾT 2 SAI**

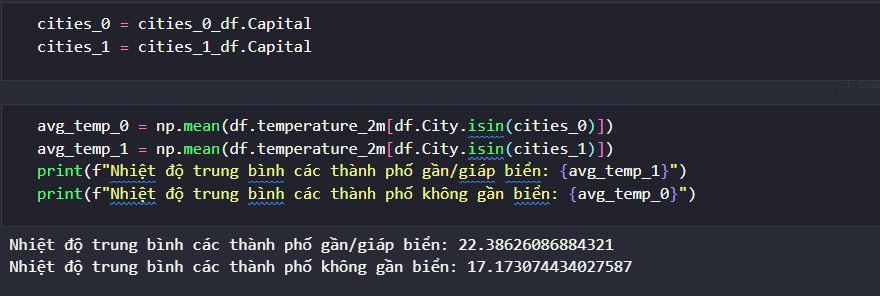
### Giả thuyết 3: Các thành phố gần biển có nhiệt độ trung bình thấp hơn, tốc độ gió cao hơn so với các thành phố sâu trong đất liền.

**Nhận xét** : Các thành phố đều giáp biển nhiều hơn 1 chút

**So sánh:**

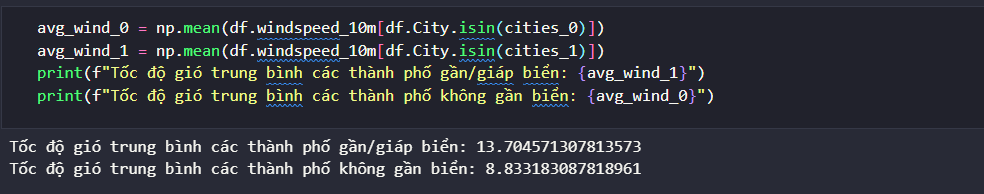
* **Nhiệt độ trung bình :**

****

**Nhận xét:** Nhiệt độ trung bình các thành phố gần/ giáp biển ấm hơn so với các thành phố sâu trong đất liền.

**Giải thích**: Do các vùng biển có các dòng hải lưu. Các dòng hải lưu có ảnh hưởng đáng kể đến nhiệt độ của các thành phố gần biển và là một yếu tố quan trọng trong việc làm cho các khu vực này có nhiệt độ ấm hơn. Vì vậy giúp biên độ nhiệt không thay đổi mạnh bằng các thành phố sâu trong đất liền.

* **Tốc độ gió trung bình :**

****

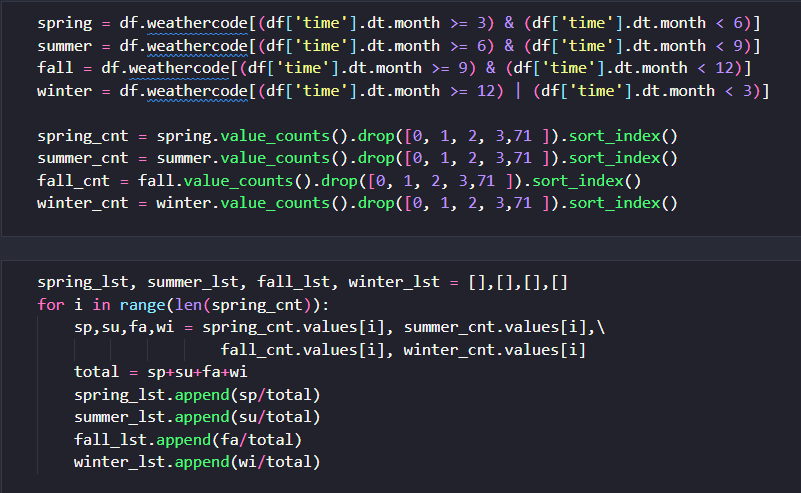
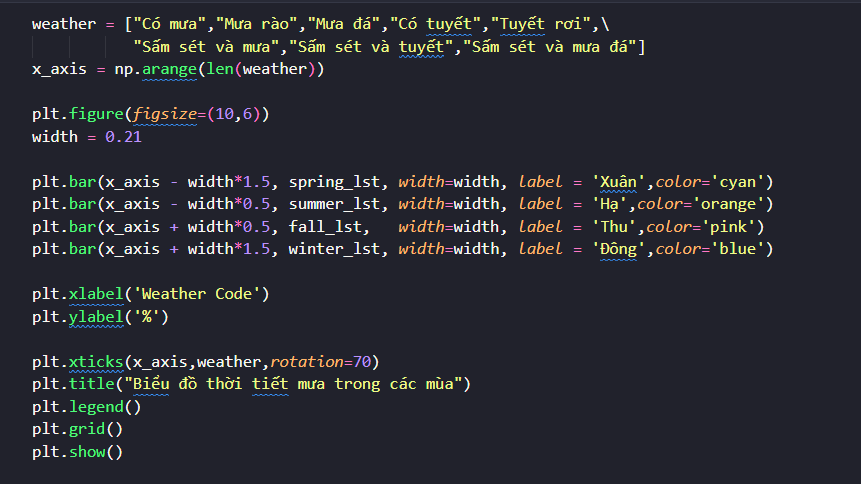
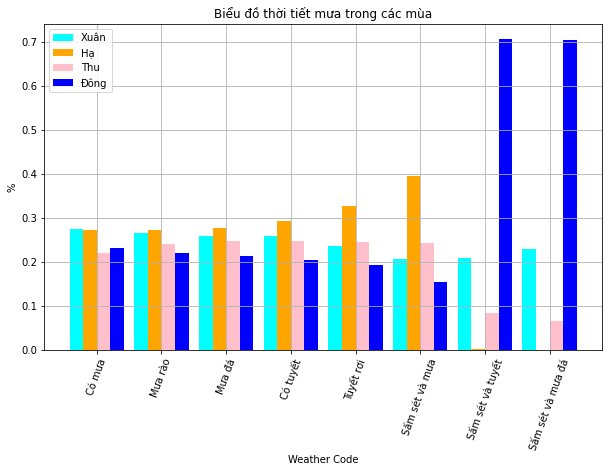
**Nhận xét:** Tốc độ gió các thành phố gần biển cao hơn.

**Giải thích:** Do yếu tố địa hình. Các thành phố gần biển không phải chịu ảnh hưởng của núi cao chặn gió nên tốc độ gió gần biển sẽ cao hơn trong đất liền.

* **Kết Luận** : Giả thuyết đúng phần tốc độ gió nhưng sai về phần nhiệt độ

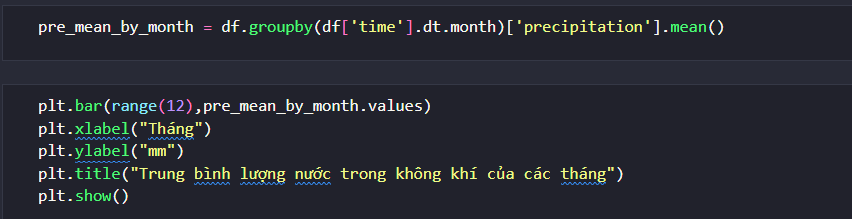
### Giả thuyết 4 : Mùa hè nhiều mưa, bão nhất

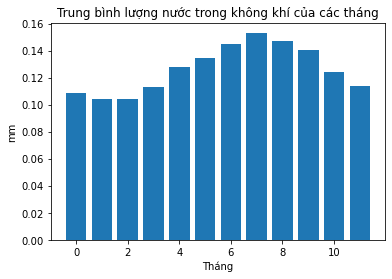
Dựa trên thời tiết, tổng lượng mưa và tốc độ gió

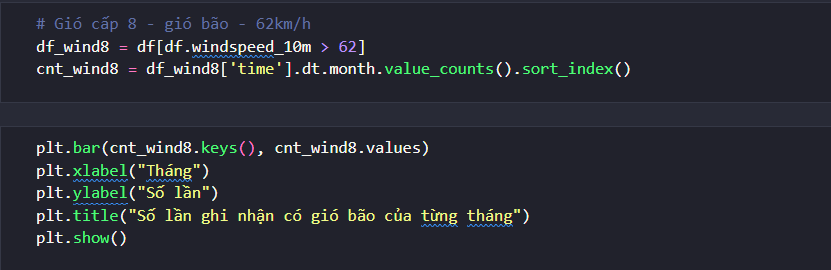
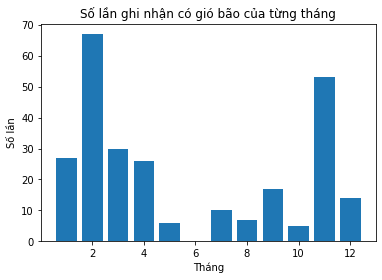
**Nhận xét**: Mùa hè vẫn có tuyết do một vài nơi gần cực. Mùa hè có nhiều mưa hơn các mùa khác.

**Cần dựa vào lượng mưa để chính xác hơn.**

****

****

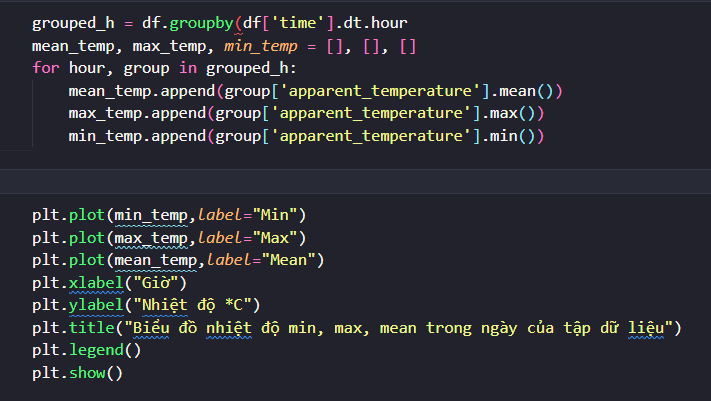
**Nhận xét:** Có thể thấy trong tháng 6-9 của mùa hè có lượng nước trong không khí nhiều hơn. Tức có lượng mưa nhiều hơn các tháng còn lại.

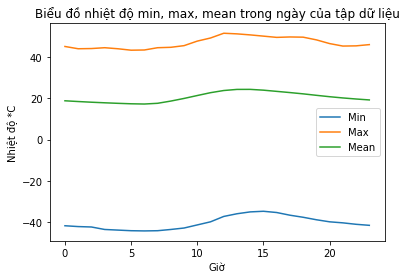
 

**Nhận xét** : Mùa hè ít gió mạnh hơn các mùa khác.Trên thực tế, mùa bão thường kết thúc vào tháng 11. Vì vậy có thể thấy tháng 11 ghi nhận nhiều gió mạnh.

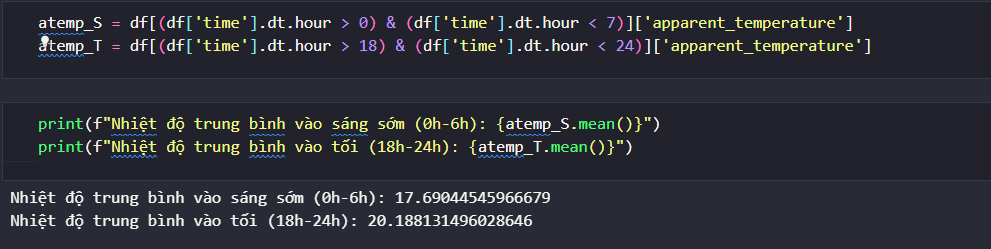
* **Kết luận: GIẢ THUYẾT MƯA ĐÚNG, GIẢ THUYẾT BÃO SAI**

### Giả thuyết 5: Sáng sớm (0h-6h) lạnh hơn so với buổi tối(18h-23h59)





Nhận xét: Có sự thay đổi về nhiệt độ trong ngày. Nhiệt độ xuống thấp nhất vào khoảng sáng sớm và tăng cao nhất vào chiều.



**Nhận xét** : Nhiệt độ trung bình vào buổi tối vẫn cao hơn vào sáng sớm.

**Giải thích:** Do nhiệt độ từ ánh sáng mặt trời. Buổi tối là thời điểm mặt trời mới lặn, nhiệt độ lúc này bắt đầu mới hạ dần. Còn vào sáng sớm, cách khoảng 6 tiếng sau khi mặt trời lặn, nhiệt độ lúc này đã hạ xuống mức thấp.

* **Kết luận: GIẢ THUYẾT ĐÚNG**

# Kết luận

Phân tích dữ liệu thời tiết là một công cụ cực kỳ quan trọng giúp chúng ta hiểu rõ hơn về các biến đổi trong thời tiết và khí hậu, đồng thời hỗ trợ nhiều khía cạnh của cuộc sống hàng ngày, từ dự đoán thời tiết cho việc quản lý tài nguyên tự nhiên và quyết định trong kinh doanh. Trong tương lai, việc phân tích dữ liệu thời tiết có thể đối mặt với các hướng mở rộng quan trọng sau:

* Dự Đoán Thời Tiết Nâng Cao: Phát triển mô hình dự đoán thời tiết ngày càng chính xác và đáng tin cậy là một ước mơ trong lĩnh vực này. Phân tích dữ liệu thời tiết có thể được cải thiện để giúp dự đoán cả biến đổi thời tiết ngắn hạn và dài hạn, giúp cảnh báo và ứng phó với các hiện tượng khí hậu cực đoan.
* Ứng Dụng Dữ Liệu Thời Tiết Trong Kinh Doanh: Nhiều ngành công nghiệp như nông nghiệp, năng lượng, và du lịch có thể tận dụng phân tích dữ liệu thời tiết để quản lý tài nguyên và thực hiện quyết định kinh doanh thông minh. Trong tương lai, việc tích hợp dữ liệu thời tiết vào quy trình sản xuất và quản lý sẽ ngày càng quan trọng.
* Mô Hình Hóa Biến Đổi Khí Hậu: Hiểu biết về biến đổi khí hậu và ảnh hưởng của nó lên môi trường và xã hội là một lĩnh vực nghiên cứu quan trọng. Phân tích dữ liệu thời tiết có thể được sử dụng để xây dựng mô hình và dự đoán xu hướng biến đổi khí hậu để đưa ra các biện pháp hỗ trợ và kiểm soát.
* Quản Lý Tình Huống Khẩn Cấp: Phân tích dữ liệu thời tiết có thể được cải tiến để hỗ trợ quản lý tình huống khẩn cấp như dự báo bão, lũ lụt, hoặc hạn hán. Điều này giúp chuẩn bị và phản ứng nhanh chóng trong các tình huống khẩn cấp.
* Tùy Chỉnh Dự Đoán Thời Tiết Cá Nhân Hóa: Trong tương lai, có thể phát triển hệ thống dự đoán thời tiết cá nhân hóa dựa trên dữ liệu cá nhân của người dùng, giúp họ có cái nhìn cụ thể về thời tiết trong khu vực của họ và điều này có thể được ứng dụng trong việc lập kế hoạch hoặc quyết định hàng ngày.

Phân tích dữ liệu thời tiết sẽ tiếp tục là một công cụ quan trọng trong cuộc sống hàng ngày và trong quá trình nghiên cứu và quản lý các khía cạnh của môi trường và tài nguyên tự nhiên.

Lưu ý bài : bài báo cáo và code được lưu trữ tại :

**--- Kết thúc ---**