ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo dự án cuối kỳ:

TRỰC QUAN & PHÂN TÍCH THỜI TIẾT CÁC VÙNG MIỀN TẠI VIỆT NAM

TỪ NĂM 2009 ĐẾN NĂM 2020

Giảng viên hướng dẫn:

TS. Bùi Tiến Lên

Môn học:

Tên môn học: Trực quan hóa dữ liệu (Data visualization)

Mã học phần: CSC10108

Sinh viên báo cáo:

21120179 Nguyễn Đặng Đăng Khoa

Thành phố Hồ Chí Minh, tháng 07 năm 2024

MỤC LỤC

MỤC LỤC 1
LỜI CẢM ƠN 2
BÁO CÁO DỰ ÁN 3
I. Giới thiệu về bộ dữ liệu 3
1. Lý do lựa chọn chủ đề 3
2. Nguồn dữ liệu 3
3. Giấy phép sử dụng 4
4. Mục tiêu trực quan hướng đến 4
II. Khám phá dữ liệu và tiền xử lý dữ liệu 4
1. Quan sát dữ liệu và khám phá dữ liệu 4
2. Tiền xử lý dữ liệu 6
III. Trực quan & phân tích dữ liệu 8
1. Trực quan dữ liệu 8
2. Phân tích dữ liệu 17
3. Tích hợp AI 22

LỜI CẢM ƠN

Xin gửi lời cảm ơn sâu sắc đến Quý Giảng viên môn học **Trực quan hóa dữ liệu** đã tận tình hướng dẫn và hỗ trợ nhóm trong suốt quá trình học tập và thực hiện dự án trực quan hóa và phân tích dữ liệu "**TRỰC QUAN & PHÂN TÍCH THỜI TIẾT CÁC VÙNG MIỀN TẠI VIỆT NAM**" từ năm 2009 đến năm 2020.

Sự tận tâm và nhiệt huyết của Quý Giảng viên đã truyền cảm hứng cho nhóm trong suốt quá trình học tập và nghiên cứu. Nhờ những bài giảng đầy đủ kiến thức và những lời khuyên bổ ích của Quý Giảng viên, các thành viên trong nhóm đã có được những hiểu biết sâu sắc về lĩnh vực trực quan hóa dữ liệu, từ đó áp dụng hiệu quả vào dự án của mình.

Bên cạnh đó, Quý Giảng viên cũng đã tạo điều kiện cho nhóm được tiếp cận với những nguồn tài liệu quý giá, giúp nhóm có thể thu thập, trực quan và phân tích dữ liệu một cách chính xác và hiệu quả. Nhờ sự hỗ trợ của Quý Giảng viên, nhóm đã có thể hoàn thành dự án đúng thời hạn và đạt được những kết quả như mong đợi.

Một lần nữa, xin bày tỏ lòng biết ơn sâu sắc đến Quý Giảng viên vì những đóng góp to lớn cho dự án của nhóm. Sự tận tâm, tâm huyết và những kiến thức chuyên môn quý giá của Quý Giảng viên sẽ là hành trang vô giá giúp thành viên trong nhóm tiếp tục học tập và phát triển trong tương lai.

Kính chúc Quý Giảng viên sức khỏe và thành công trong sự nghiệp!

Trân trọng.

BÁO CÁO DỰ ÁN

I. Giới thiệu về bộ dữ liệu

1. Lý do lựa chọn chủ đề

1.1. Tính cần thiết của thông tin thời tiết

Thời tiết có ảnh hưởng lớn đến đời sống hàng ngày và các hoạt động kinh tế, xã hội. Nắm bắt thông tin thời tiết kịp thời và chính xác giúp mọi người đưa ra quyết định phù hợp trong công việc và sinh hoạt, đồng thời hỗ trợ các cơ quan chức năng trong công tác quản lý và phòng chống thiên tai.

1.2. Sự đa dạng về khí hậu của Việt Nam

Việt Nam là quốc gia có khí hậu nhiệt đới gió mùa, nhưng lại có sự phân hóa đa dạng giữa các vùng miền. Điều này tạo ra những đặc điểm thời tiết riêng biệt, từ đó đặt ra nhu cầu tìm hiểu và phân tích khí hậu từng vùng để có thể đưa ra những dự báo chính xác và phù hợp.

1.3. Nhu cầu thực tế của người dân và các ngành kinh tế

Nông nghiệp, du lịch, giao thông và nhiều ngành kinh tế khác đều chịu tác động mạnh mẽ bởi thời tiết. Việc dự báo và phân tích thời tiết không chỉ giúp mọi người chủ động trong cuộc sống mà còn hỗ trợ các doanh nghiệp và tổ chức đưa ra quyết định kinh doanh hiệu quả.

1.4. Tính địa phương và sự gắn bó với quê hương

Việc lựa chọn phân tích thời tiết Việt Nam thể hiện sự quan tâm và gắn bó của mỗi thành viên trong nhóm với quê hương của mình. Với việc tìm hiểu về đặc trưng quê hương mình giúp nhóm có thêm kiến thức và hiểu biết.

Từ những lý do trên, nhóm cảm thấy dự án này mang tính thực tế cao trong cuộc sống, gần gũi với các thành viên trong nhóm, giúp nhóm có cơ hội được tiếp xúc với dữ liệu thực tế và có thể phần nhỏ nào đó đóng góp được cho cộng đồng mà các thành viên trong nhóm đang sinh sống.

2. Nguồn dữ liệu

Bộ dữ liệu của dự án được nhóm lấy từ trang web Kaggle. Tại đây, nhóm biết được rằng bộ dữ liệu này được thu thập từ nguồn www.accuweather.com

thuộc Accuweather - một công ty truyền thông Hoa Kỳ cung cấp dịch vụ dự báo thời tiết thương mại trên toàn cầu.

Tìm hiểu thêm, để có được bản ghi dữ liệu từ Accuweather cần phải đăng ký tài khoản cá nhân dành cho lập trình viên hoặc đăng ký dưới quyền doanh nghiệp bảo trợ, bị giới hạn số lần crawl dữ liệu trong một ngày để tránh vấn đề liên quan đến bảo mật theo chính sách của Accuweather.

3. Giấy phép sử dụng

Creative Commons Zero (CC0)

Giấy phép Creative Commons Zero (CC0) là một loại giấy phép bản quyền công khai cho phép người dùng sử dụng, sửa đổi và phân phối tác phẩm miễn phí mà không cần xin phép hoặc ghi công tác giả. Nói cách khác, tác phẩm được dâng hiến vào phạm vi công cộng và cộng đồng.

- *Miễn phí*: Có thể sử dụng tác phẩm cho bất kỳ mục đích nào, bao gồm cả mục đích thương mại, mà không cần phải trả phí hoặc xin phép;
- *Không ghi công*: Không bắt buộc ghi công tác giả khi sử dụng tác phẩm;
- *Miễn trừ trách nhiệm*: Tác giả không chịu trách nhiệm về bất kỳ thiệt hại nào phát sinh do việc sử dụng tác phẩm;
- *Không giới hạn*: Người dùng có thể sửa đổi, pha trộn và xây dựng dựa trên tác phẩm theo bất kỳ cách nào.

4. Mục tiêu trực quan hướng đến

Xử lí và hiển thị dữ liệu thông qua các biểu đồ để dễ dàng phân tích, rút trích thông tin hữu ích và dự đoán xu hướng:

- Dưới góc nhìn của một người dân có thể quan sát được kiểu thời tiết của từng giai đoạn hoặc từng mùa trong năm;
- Dưới góc nhìn của nhà nghiên cứu có thể quan sát sự phân bổ giá trị từng yếu tố thời tiết theo gian, từ đó đưa ra dự đoán xu hướng kiểu thời tiết, biến thiên của các yếu tố thời tiết và phát hiện sớm các trạng thái thời tiết bất thường có thể xảy ra trong tương lai của mỗi vùng miền.

II. Khám phá dữ liệu và tiền xử lý dữ liệu

1. Quan sát dữ liệu và khám phá dữ liệu

1.1. Kích thước và ý nghĩa dữ liệu

Đây là bộ dữ liệu tĩnh được tác giả thu thập từ ngày 01 tháng 01 năm 2009 đến ngày 18 tháng 06 năm 2021 tại 40 trạm đo ở Việt Nam mà tác giả quan tâm như thủ đô Hà Nội, thành phố Hồ Chí Minh, Huế…

Bộ dữ liệu gồm có:

- 181960 dòng dữ liệu ghi chép các giá trị từng yếu tố thời tiết tại thời điểm đo của trạm đo;
- 10 thuộc tính (biến độc lập):
  - `province` tỉnh/thành phố của trạm đo;
  - `max` nhiệt độ cao nhất trong ngày, đơn vị: °C;
  - `min` nhiệt độ thấp nhất trong ngày, đơn vị: °C;
  - `wind` tốc độ gió tại thời điểm đo, đơn vị: km/h;
  - `wind_d` hướng gió tại thời điểm đo;
  - `rain` tổng lượng mưa đo được trong ngày, đơn vị: mm;
  - `humidi` độ ẩm đo được trong ngày, đơn vị: %;
  - `cloud` độ phủ mây hay mật độ mây, đơn vị: %;
  - `pressure` áp suất không khí trong ngày, đơn vị: mbar;
  - `date` thời gian đo, đơn vị tính nhỏ nhất là ngày.

Trong thuộc tính hướng gió `wind_w` sử dụng các ký hiệu là viết tắt của hướng gió bằng tiếng Anh, những ký hiệu này có ý nghĩa như sau:

Ký hiệu	Ý nghĩa tiếng Anh	Ý nghĩa tiếng Việt
N	North	Bắc
NNE	North-Northeast	Bắc Đông Bắc
NE	Northeast	Đông Bắc
ENE	East-Northeast	Đông Đông Bắc
E	East	Đông
ESE	East-Southeast	Đông Đông Nam
SE	Southeast	Đông Nam
SSE	South-Southeast	Nam Đông Nam
S	South	Nam
SSW	South-Southwest	Nam Tây Nam
SW	Southwest	Tây Nam
WSW	West-Southwest	Tây Tây Nam
W	West	Tây
WNW	West-Northwest	Tây Tây Bắc
NW	Northwest	Tây Bắc
NNW	North-Northwest	Bắc Tây Bắc

1.2. Dữ liệu trùng lặp

Nhóm quan sát, ngoại trừ trạm đo Hà Nội, mỗi trạm đều có 4549 dòng dữ liệu, chúng đều phản ánh cho một giá trị duy nhất tại thời điểm đo. Trạm đo Hà Nội có 9098 dòng dữ liệu, nhưng chỉ có 4548 dòng dữ liệu phản ánh giống hoàn toàn dữ liệu đã có trước đó, vậy tồn tại một thời điểm đo có phản ánh 2 giá trị khác nhau.

1.3. Dữ liệu bị thiếu

Nhóm quan sát thống kê theo trạm đo, tất cả các trạm đều không có dữ liệu bị thiếu.

2. Tiền xử lý dữ liệu

2.1. Bổ sung thông tin thuộc tính

Mục tiêu hướng đến quan sát xu hướng thời tiết ở từng vùng miền tại Việt Nam, cũng như loại bỏ yếu tố chủ quan trong việc lựa chọn trạm đo của tác giả, nhóm chỉnh sửa và bổ sung thông tin thuộc tính province:

- Thuộc tính `province` chưa thống nhất chung cùng cấp tỉnh/thành phố trực thuộc Trung Ương, nhóm sẽ đồng thời chỉnh sửa và chuyển về ngôn ngữ tiếng Việt;
- Bổ sung thông tin `Vùng miền` tương ứng với `Tỉnh/Thành phố` bằng ngôn ngữ tiếng Việt để quan sát yếu tố thời tiết bám sát đặc điểm vùng miền Việt Nam, loại bỏ sự lựa chọn chủ quan từ tác giả.

2.2. Chuyển đổi ngôn ngữ

Ngoài chuyển đổi dữ liệu sang tiếng Việt, nhóm còn chuyển đổi tên thuộc tính, cụ thể:

- `province` Trạm đo;
- `max` Nhiệt độ cao nhất (°C);
- `min` Nhiệt độ thấp nhất (°C);
- `wind` Tốc độ gió (km/h);
- `wind_d` Hướng gió;
- `rain` Lượng mưa (mm);
- `humidi` Độ ẩm (%);
- `cloud` Mật độ mây (%);
- `pressure` Áp suất (mbar);
- `date` Thời gian đo.

2.3. Xử lý dữ liệu trùng lặp

Đầu tiên, nhóm xóa các dòng trùng lặp dữ liệu hoàn toàn của trạm đo Hà Nội. Tiếp theo, sau khi sử dụng kỹ thuật trích xuất, nhóm thấy tại thời điểm ngày 12-12-2017, trạm Hà Nội có giá trị đo tốc độ gió khác nhau; tuy nhiên phần chênh lệch chỉ bằng đơn vị tính nhỏ nhất, thêm vào đó nhóm quan sát các ngày lân cận có sự biến thiên không quá lớn xung quanh giá trị này. Vì vậy, nhóm đánh giá xóa bỏ dòng dữ liệu bị lặp thời điểm đo tại trạm Hà Nội không ảnh hưởng quá nhiều đến dữ liệu, và nhóm lựa chọn xử lý bằng phương pháp này.

2.4. Loại bỏ dữ liệu dang dở

Dữ liệu được ghi chép từ ngày 01 tháng 01 năm 2009 đến ngày 18 tháng 06 năm 2021. Tại năm 2021 nhóm nhận xét kích thước mẫu dữ liệu chưa đủ để so sánh yếu tố thời tiết với các năm còn lại. Vì vậy nhóm sẽ loại bỏ các mẫu dữ liệu ghi chép của năm 2021.

Sau khi xử lý, bộ dữ liệu có kích thước: 170820 mẫu dữ liệu của 39 trạm đo từ năm 2009 đến năm 2020, biểu diễn thông tin qua 12 thuộc tính. Có nghĩa rằng mỗi trạm đo có 4380 mẫu dữ liệu.

2.5. Xác định kiểu dữ liệu thuộc tính

Nhóm xác định các nhóm thuộc tính sau:

- Thuộc tính định tính không thứ tự: `Trạm đo`, `Tỉnh/Thành phố`, `Vùng miền`, `Hướng gió`;
- Thuộc tính định tính có thứ tự: `Thời gian đo`;
- Thuộc tính định lượng rời rạc: `Nhiệt độ cao nhất (°C)`, `Nhiệt độ thấp nhất (°C)`, `Tốc độ gió (km/h)`, `Độ ẩm (%)`, `Mật độ mây (%)`, `Áp suất (mbar)`;
- Thuộc tính định lượng liên tục: `Lượng mưa (mm)`.

III. Trực quan & phân tích dữ liệu

1. Trực quan dữ liệu

1.1. Hướng dẫn truy cập

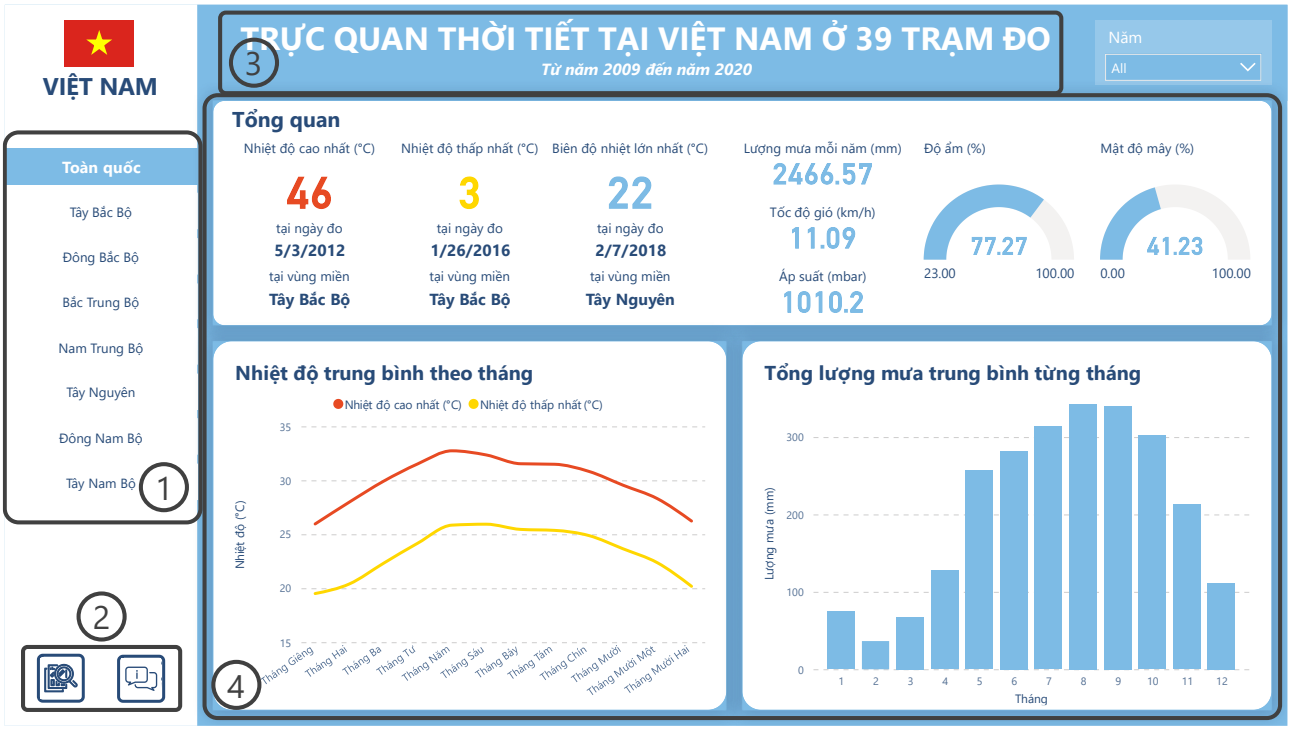
Giảng viên có thêm thao tác trực tiếp với dashboard của nhóm bằng cách truy cập đường link này, hoặc quét mã QR dưới đây:

Power BI

Lưu ý: Vì nhóm thực hiện với bản dùng thử của Power BI nên: Quý giảng viên vui lòng sử dụng email của trường học có miền hcmus.edu.vn để có quyền truy cập thao tác trực tiếp.

1.2. Giới thiệu dashboard

Nhóm tạo ra 3 dashboard chính phục vụ dự án cuối kỳ. Mỗi dashboard được thiết kế thành 4 phần chính:



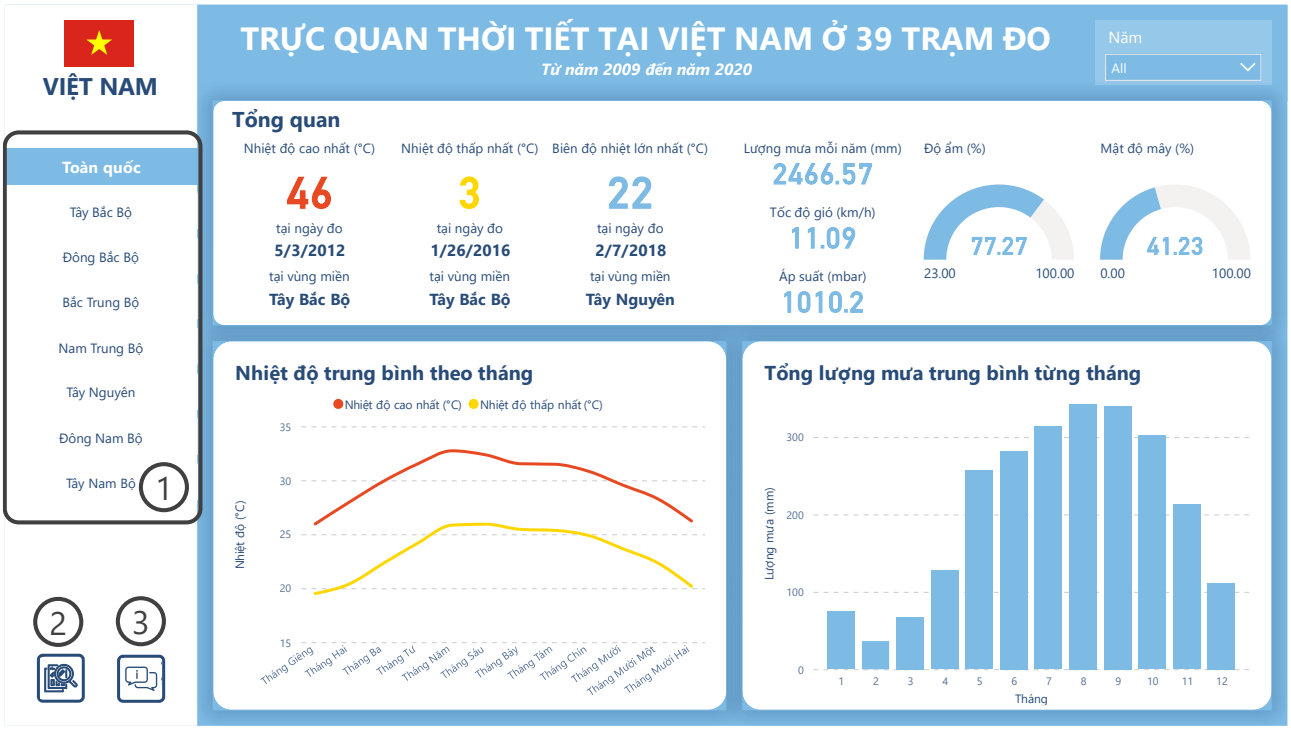
(1). Thanh điều hướng giữa các trang trong 7 vùng miền thuộc nhóm trực quan các vùng miền tại Việt Nam, giúp người dùng lựa chọn các vùng miền mong muốn quan sát. Thanh điều hướng chỉ lựa chọn duy nhất một vùng miền tại thời điểm xem;

(2). Nút điều hướng, tương tự thanh điều hướng, để người dùng lựa chọn một trong các trang trực quan "Mối quan hệ giữa các yếu tố thời tiết" và "Hỗ trợ hỏi đáp từ AI";

(3). Tiêu đề mà trang trực quan hướng đến;

(4). Trình bày nội dung trực quan.

Dựa vào mục tiêu trực quan, nhóm thiết kế dashboard thành 3 loại khác nhau:

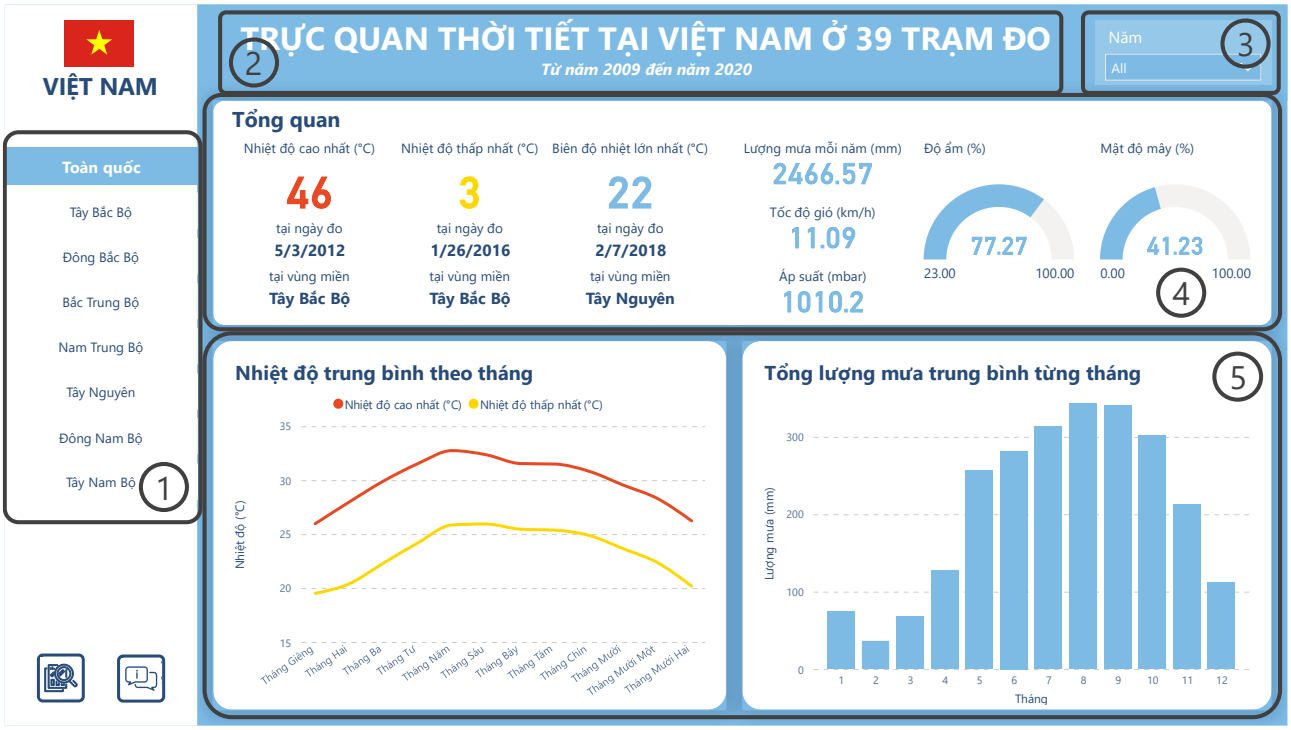


(1). Dashboard trực quan đơn giản các yếu tố thời tiết ở mỗi vùng miền, cung cấp nhanh chóng các thông tin và biểu đồ cơ bản, hướng đến mức độ tiếp cận rộng rãi;

(2). Dashboard quan sát các mối quan hệ giữa các yếu tố thời tiết. Nhóm sử dụng kinh nghiệm dân gian để kiểm chứng mối quan hệ của chúng dựa trên dữ liệu thực tế, từ đó kiểm định độ tin cậy trong các kinh nghiệm dân gian;

(3). Dashboard hỗ trợ bởi AI giúp người dùng trích lọc thông tin, trực quan biểu đồ mà nhóm chưa trực quan trong 2 dashboard trước đó.

1.2.1. Dashboard trực quan các vùng miền tại Việt Nam



Trên đây là hình ảnh nhóm dashboard trực quan thời tiết các vùng miền tại Việt Nam. Trong dashboard được thiết kế thành 5 phần:

(1). Thanh điều hướng giữa các trang trong 7 vùng miền để người dùng lựa chọn các vùng miền mong muốn quan sát. Thanh điều hướng chỉ lựa chọn duy nhất một vùng miền tại thời điểm xem;

(2). Tiêu đề mà trang trực quan hướng đến;

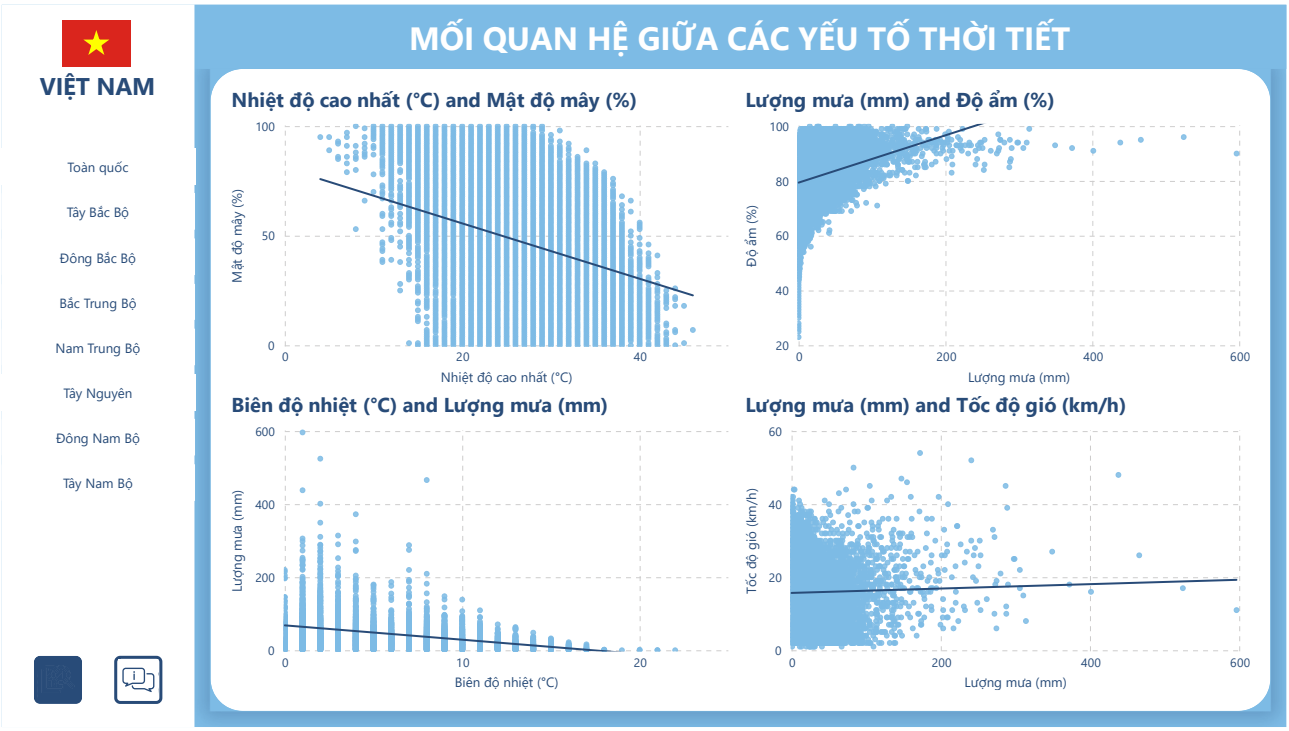
(3). Slicer xuất hiện ở các trang trong 7 vùng miền thuộc nhóm trực quan các vùng miền tại Việt Nam, giúp người dùng lựa chọn thời gian quan sát mong muốn. Slicer này được phép lọc nhiều lựa chọn;

(4). Trực quan các thông tin và số liệu cơ bản;

(5). Trực quan bằng biểu đồ theo xu hướng thời gian.

Trong dashboard nhóm có sử dụng biến hình ảnh màu sắc cho biến nhiệt độ để người xem dễ nhận biết giữa các yếu tố, đồng thời sử dụng màu đỏ cho nhiệt độ cao nhất, màu vàng cam cho nhiệt độ thấp nhất tượng trưng mức thang nhiệt độ trong các bản tin dự báo thời tiết vẫn hay dùng.

1.2.2. Dashboard mối quan hệ giữa các yếu tố thời tiết



Tại dashboard này người xem mối quan hệ ảnh hưởng lẫn nhau giữa các yếu tố thời tiết. Dưới góc nhìn của nhóm liệt kê ra một số kinh nghiệm dân gian để kiểm định bằng dữ liệu thực tế.

1.2.3. Dashboard hỗ trợ trực quan từ AI

Với công cụ Copilot trong hệ sinh thái Microsoft, việc tích hợp công cụ hỗ trợ thông minh trở nên dễ dàng. Tuy nhiên sự đánh đổi phiên bản miễn phí là chỉ

hỗ trợ ngôn ngữ tiếng Anh. Vấn đề này cũng không gây khó khăn khi người dùng muốn trực quan thêm một số thông tin khác.

1.3. Đánh giá trực quan

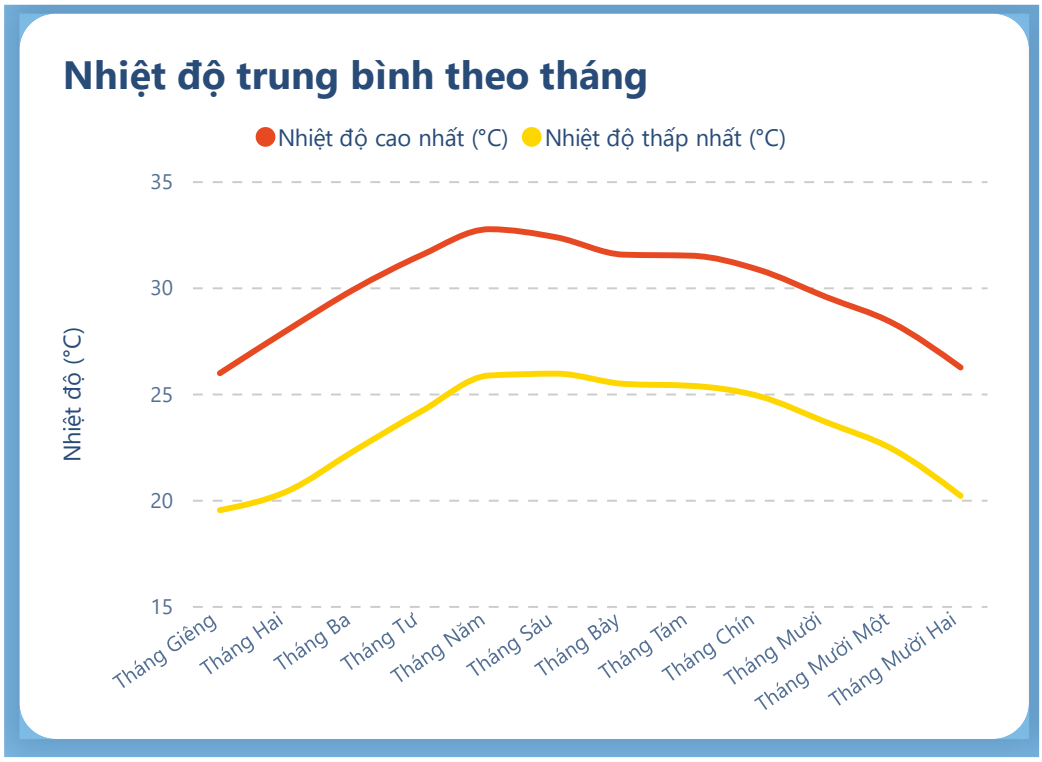
1.3.1. Kết hợp nguồn dữ liệu đáng tin cậy

Đây là nguồn dữ liệu đáng tin cậy vì được thu thập từ Accuweather, công ty truyền thông Hoa Kỳ cung cấp dịch vụ dự báo thời tiết thương mại trên toàn cầu, được nhiều hãng thiết bị điện tử sử dụng.

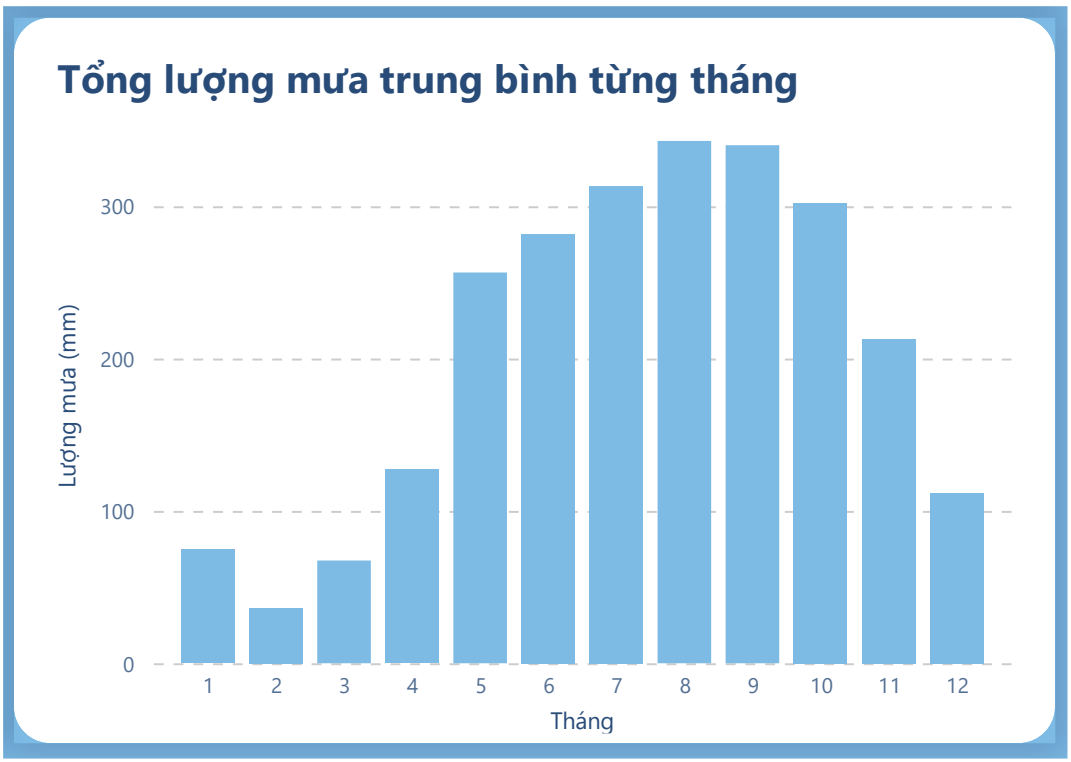
1.3.2. Phù hợp với mục đích

Dashboard sử dụng các loại biểu đồ linh hoạt và phù hợp với mục đích trực quan dữ liệu:

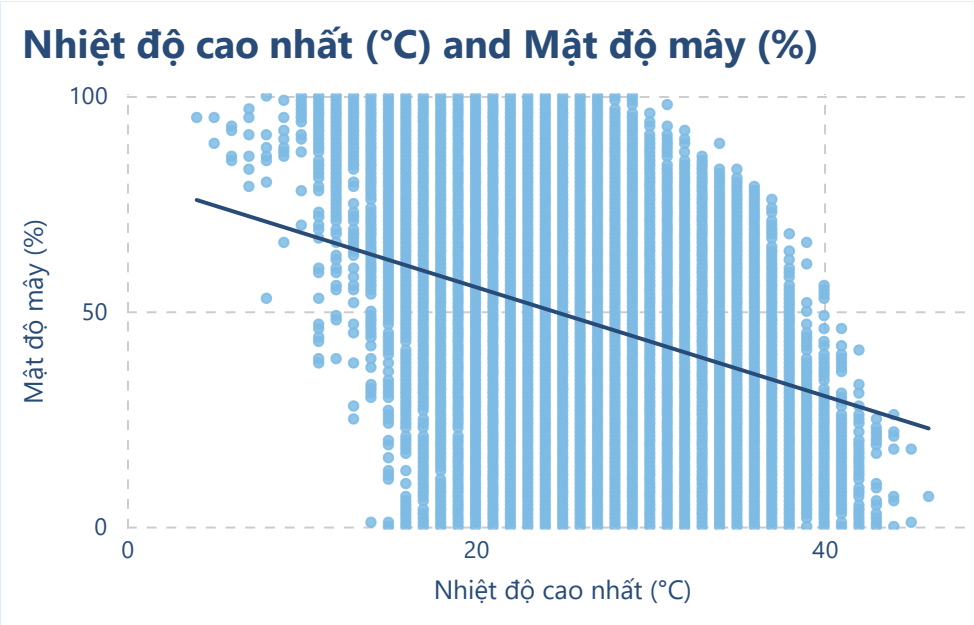
- Biểu đồ đường: Theo dõi xu hướng theo thời gian của đối tượng.



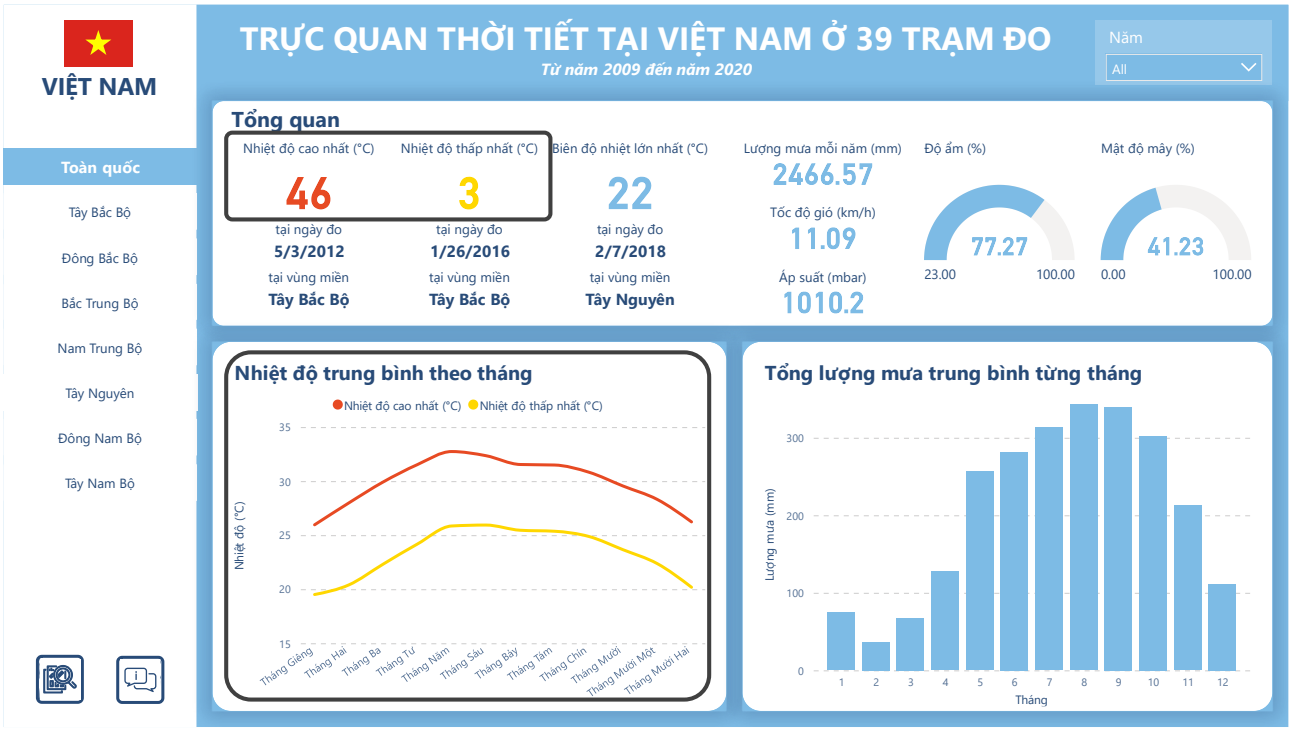
- Biểu đồ cột: So sánh số lượng giữa các đối tượng, đồng thời cũng có thể quan sát xu hướng của yếu tố thời tiết.



- Biểu đồ phân tán: Quan sát sự phân bố và mối quan hệ giữa các yếu tố thời tiết.



Yếu tố Nhiệt độ cao nhất (°C) và Nhiệt độ thấp nhất (°C) khi trực quan độc lập được sử dụng biến màu sắc tương ứng đỏ và vàng để phân biệt với nhau, phân biệt với các yếu tố thời tiết khác, gần gũi với màu sắc các bản tin thời tiết hay sử dụng.



1.3.3. Rõ ràng và dễ hiểu

Dashboard được chia thành các trang rõ ràng dựa theo mục đích trực quan.

Các biểu đồ và đồ thị có số liệu rõ ràng, có đầy đủ tiêu đề, chú thích giúp người xem dễ dàng tiếp cận dữ liệu được trực quan.

Các số liệu, đối tượng quan trọng, cần thu hút sự chú ý được làm nổi bật bằng màu sắc nổi bật, sáng hơn.

1.3.4. Sự tích hợp và liên kết

Giữa các biểu đồ, số liệu và bộ lọc có tính liên kết với nhau khi người dùng lựa chọn đại lượng theo ý muốn của một biến.

Giữa các bảng cơ sở dữ liệu được khai báo mối quan hệ giữa các biến một cách rõ ràng và đầy đủ góp phần làm tính liên kết giữa các biểu đồ trở nên linh hoạt.

1.3.5. Tương tác và điều hướng

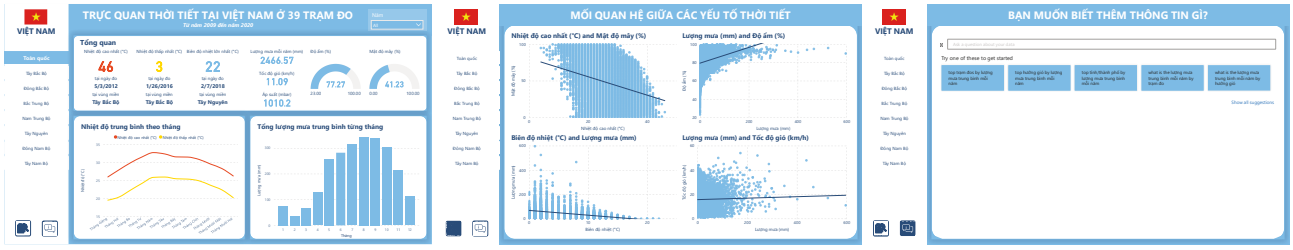
Các biểu đồ được biểu diễn rõ ràng, sạch sẽ; giữa các biểu đồ có tính liên kết dữ liệu với nhau. Vì vậy khi người dùng thăm dò đại lượng bất kỳ trong biểu đồ, các thông tin số liệu và biểu đồ liên quan cũng thay đổi phù hợp với mục đích thăm dò của người xem, giúp người xem có thêm nhiều thông tin.

Tính điều hướng cũng được thiết kế sử dụng các biểu tượng hoặc văn bản rõ ràng, dễ hiểu và linh hoạt, giúp người dùng dễ dàng điều hướng đến dashboard mong muốn.

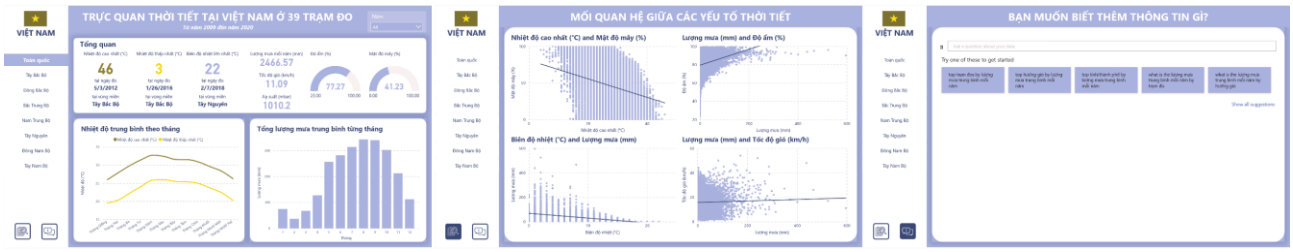
1.3.6. Thiết kế hấp dẫn

- Sử dụng gam màu tươi mát, dễ chịu, liên quan đến chủ đề thời tiết.
- Sử dụng kết hợp màu nền là màu trắng, viền màu đen và xám với màu chữ là màu xanh, màu xám và màu đen giúp cho việc quan sát dữ liệu nên dễ dàng.
- Những điểm dữ liệu, đối tượng được tô màu xanh biển và các màu sắc tươi tắn để làm nổi bật, thu hút sự tập trung.
- Đơn giản hóa gam màu, tránh sử dụng màu sắc lòe loẹt, sặc sỡ gây rối mắt và phân tán sự tập trung.

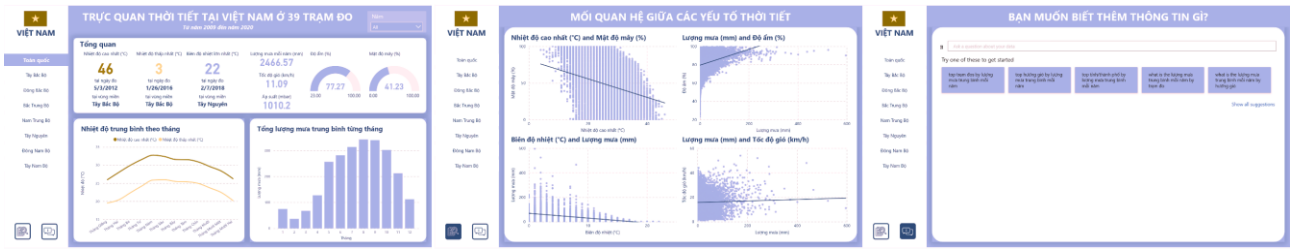
1.3.7. Sự thân thiện màu sắc



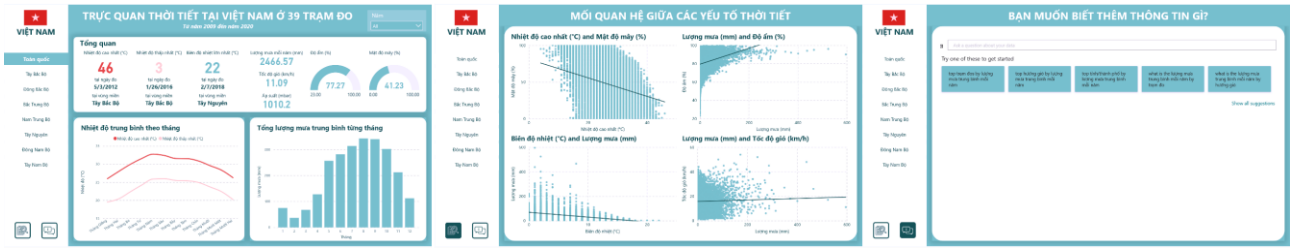
Dashboard dưới góc nhìn của người mù màu đỏ.



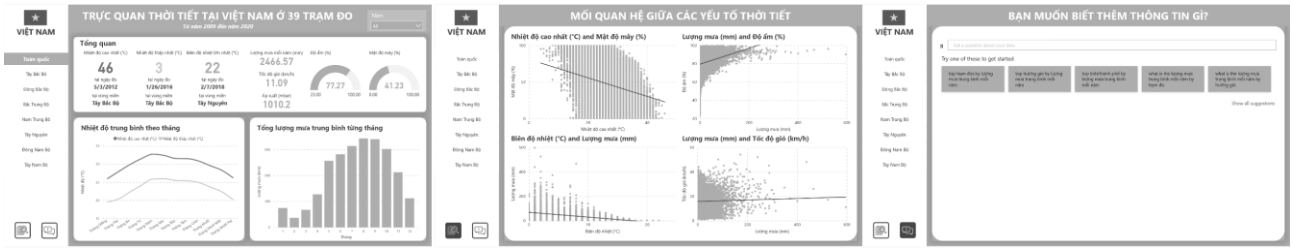
Dashboard dưới góc nhìn của người mù màu xanh lá.



Dashboard dưới góc nhìn của người mù màu xanh dương.



Dashboard dưới góc nhìn của người mù màu sắc hoàn toàn.



Nhận xét:

- Dashboard được thiết kế ban đầu không hướng đến đối tượng mù màu cụ thể nhưng tổng quan các biểu đồ vẫn hỗ trợ tốt với từng đối tượng mù màu.
- Người mù màu hoàn toàn có thể nhận biết và phân biệt giữa các đối tượng trong cùng dựa trên độ đậm nhạt của màu sắc trên biểu đồ nên dashboard hoàn toàn thân thiện với người dùng mù màu.

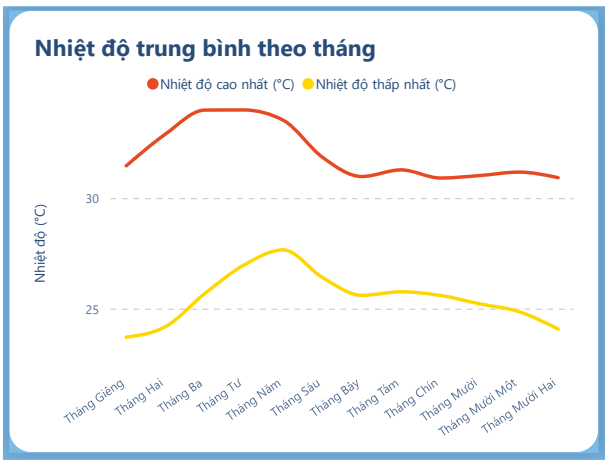
2. Phân tích dữ liệu

2.1. Mục tiêu hướng đến

- Yếu tố thời tiết mọi người thường quan tâm là nhiệt độ và lượng mưa, vì vậy nhóm tập trung trực quan và phân tích làm nổi bật các yếu tố này. Các yếu tố khác có thể nhờ hỗ trợ AI trực quan biểu đồ để tránh dashboard bị quá tải và rối mắt.
- Lựa chọn một vùng miền Đông Nam Bộ, khu vực mà nhóm đang sinh sống, là dashboard đại diện trong nhóm dashboard phân tích các vùng miền để tìm xu hướng nhiệt độ và lượng mưa, kiểm chứng bằng mô hình tuyến tính và phi tuyến.
- Nhóm dashboard mối quan hệ giữa các yếu tố thời tiết hướng đến tìm những yếu tố ảnh hưởng đến nhiệt độ và lượng mưa. Kiểm chứng bằng mô hình tuyến tính với biểu đồ đại diện là mối quan hệ nhiệt độ và mật độ mây.

2.2. Phân tích thời tiết tại vùng miền đại diện

2.2.1. Yếu tố nhiệt độ



Điểm sơ qua cách tính toán để có số liệu trực quan, giá trị nhiệt độ cao nhất và nhiệt độ thấp nhất là nhiệt độ được ghi chép giá trị lớn nhất và nhỏ nhất trong ngày; tính toán trung bình áp dụng cho nhóm theo tháng và năm.

- Nhiệt độ cao nhất xu hướng tăng từ tháng 1 đến tháng 4, có thể đạt đỉnh tại tháng 2, 3 và 4. Xu hướng giảm từ từ tháng 5 đến cuối năm. Tăng độ tin cậy cho dự đoán, nhóm tiến hành kiểm định mô hình hồi quy tuyến tính và phi tuyến:
  - Mô hình hồi quy tuyến tính:

    Nhiệt độ cao nhất (°C) = $-0.227634 * \text{Tháng} + 33.4756$

    Kết quả kiểm định cho độ tin cậy 98.2%.

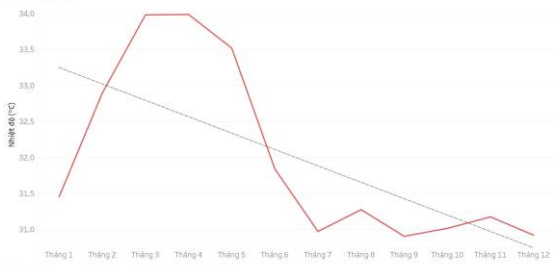
    P-value: 0.0188742

    Equation: Avg. Nhiệt độ cao nhất (°C) = -0.227634*Tháng + 33.4756

    Coefficients

Term	Value	StdErr	t-value	p-value
Tháng	-0.227634	0.0813695	-2.79754	0.0188742
intercept	33.4756	0.598864	55.8986	< 0.0001

    R-Squared: 0.4390

  - Mô hình hồi quy phi tuyến:

    Nhiệt độ cao nhất (°C) =

    $$-0.00450018 * \text{Tháng}^4 + 0.139723 * \text{Tháng}^3 - 1.46989 * \text{Tháng}^2 + 5.6317 * \text{Tháng} + 26.9365$$

    Kết quả kiểm định cho thấy mô hình phù hợp với dữ liệu R-squared

= 0.9403 và độ tin cậy 99%.

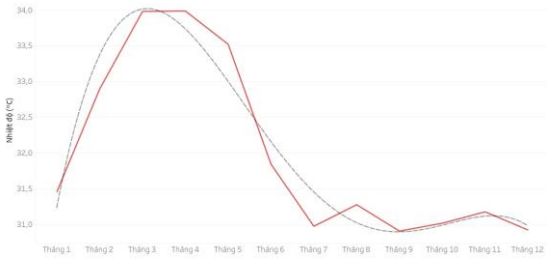
P-value: 0.0002232

Equation: Avg. Nhiệt độ cao nhất (°C) = -0.00450018*Tháng^4 + 0.139723*Tháng^3 + -1.46989*Tháng^2 + 5.6317*Tháng + 26.9365

Coefficients

Term	Value	StdErr	t-value	p-value
Tháng^4	-0.0045	0.0012367	-3.639	0.0082981
Tháng^3	0.139723	0.0323457	4.31967	0.0034816
Tháng^2	-1.46989	0.285088	-5.1559	0.0013154
Tháng	5.6317	0.962942	5.84843	0.0006316
intercept	26.9365	0.990369	27.1984	< 0.0001

R-Squared: 0.9403



- Nhiệt độ cao nhất xu hướng tăng từ tháng 1 đến tháng 5, có thể đạt đỉnh tại tháng 5. Xu hướng giảm từ từ tháng 5 đến cuối năm. Tăng độ tin cậy cho dự đoán, nhóm tiến hành kiểm định mô hình hồi quy tuyến tính và phi tuyến:
  - Mô hình hồi quy tuyến tính:

    Nhiệt độ thấp nhất (°C) = $-0.0228118 * \text{Tháng} + 25.6255$

    Kết quả kiểm định cho độ tin cậy khá thấp 17.1% và đánh giá mức độ phù hợp với dữ liệu cũng không cho khả quan.

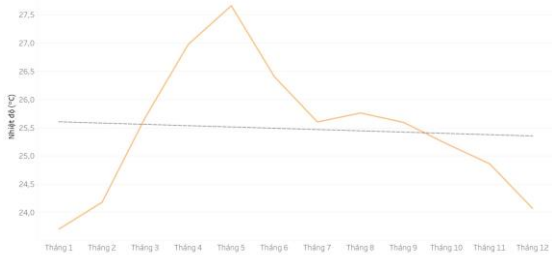
P-value: 0.82901

Equation: Avg. Nhiệt độ thấp nhất (°C) = -0.0228118*Tháng + 25.6255

Coefficients

Term	Value	StdErr	t-value	p-value
Tháng	-0.0228	0.102896	-0.221699	0.82901
intercept	25.6255	0.757291	33.8384	< 0.0001

R-Squared: 0.0049



  - Mô hình hồi quy phi tuyến:

    Nhiệt độ thấp nhất (°C) =

    $$0.000359489 * \text{Tháng}^6 - 0.0154949 * \text{Tháng}^5 + 0.25889 * \text{Tháng}^4 - 2.09055 * \text{Tháng}^3 + 8.16929 * \text{Tháng}^2 - 13.0386 * \text{Tháng} + 30.4654$$

    Kết quả kiểm định cho kết quả mô hình phi tuyến biểu diễn khá phức tạp để phù hợp với dữ liệu R-squared=0.9662, nhưng có độ tin

cậy 99%.

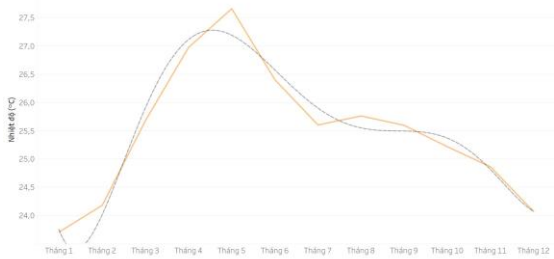
P-value: 0.0015793

Equation: Avg. Nhiệt độ thấp nhất (°C) = 0.000359489*Month of Thời gian đo^6 + -0.0154949*Month of Thời gian đo^5 + 0.25889*Month of Thời gian đo^4 + -2.09055*Month of Thời gian đo^3 + 8.16929*Month of Thời gian đo^2 + -13.0386*Month of Thời gian đo + 30.4654

Coefficients

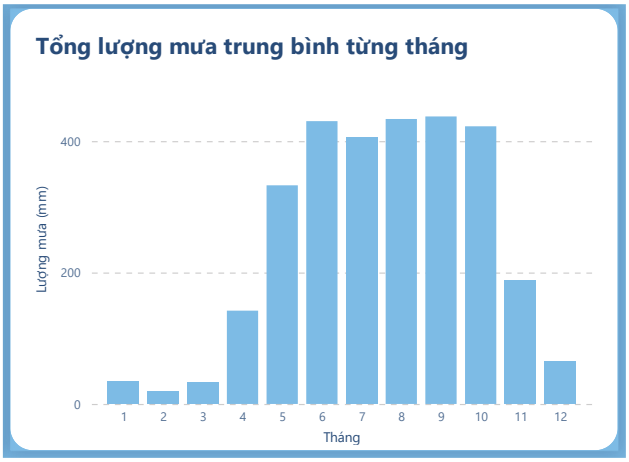
Term	Value	StdErr	t-value	p-value
Tháng^6	0.0003595	0.0001464	2.45634	0.0574863
Tháng^5	-0.0154949	0.0057204	-2.70868	0.0423428
Tháng^4	0.25889	0.0869761	2.97657	0.030922
Tháng^3	-2.09055	0.64786	-3.22685	0.0232873
Tháng^2	8.16929	2.4371	3.35205	0.0202834
Tháng	-13.0386	4.24432	-3.072	0.0277206
intercept	30.4654	2.53918	11.9981	< 0.0001

R-Squared: 0.9662



- ***Kết luận:***
  - Trong năm, từ tháng 1 đến tháng 12, nhiệt độ cao nhất trong ngày có xu hướng giảm, nhiệt độ thấp nhất trong ngày chưa đủ cơ sở chắc chắn sự biến thiên;
  - Nhiệt độ cao nhất từ mô hình cho thấy đạt đỉnh cao nhất vào tháng 3 và chạm đáy thấp nhất vào tháng 9;
  - Nhiệt độ thấp nhất từ mô hình cho thấy có sự biến đổi phức tạp và khó dự đoán. Nhiệt độ thấp nhất chạm đỉnh lớn nhất tại tháng 4 và chạm đáy lạnh nhất vào tháng 1 và 2;
  - Kết hợp dự nhiệt độ cao nhất và thấp nhất tại tháng 5 sẽ có thời tiết nóng nực cả ngày và đêm, tháng 1 sẽ có thời tiết lạnh cả ngày và đêm.

2.2.2. Yếu tố lượng mưa



Lượng mưa tính tổng theo nhóm tháng và năm, sau đó kết quả này được sử dụng để biểu diễn biểu đồ theo phương thức tính giá trị trung bình. Có nghĩa rằng giá trị mỗi cột trong biểu đồ là kết quả trung bình của tổng lượng mưa trong một tháng của mỗi năm.

Quan sát biểu đồ, tổng lượng mưa trong tháng đạt đỉnh nhiều nhất trong năm ở các tháng 6, 8 và 9; chạm đáy ít nhất trong năm ở tháng 2. Điều này dự đoán rằng vùng Đông Nam Bộ có mưa nhiều nhất trong năm vào thời điểm tháng 6 đến tháng 9, ít mưa nhất vào thời điểm từ tháng 1 đến tháng 3.

Kiểm định 2 điểm cực trị bằng mô hình hồi quy phi tuyến bậc 3 có kết quả như sau:

- Mô hình: Tổng lượng mưa trong tháng = $-2.17712 * \text{Tháng}^3 + 29.9168 * \text{Tháng}^2 - 44.5148 * \text{Tháng} + 17.9944$;
- Mức độ phù hợp mô hình với dữ liệu khá tốt, kết quả R-squared = 0.9345 và độ tin cậy 99%.

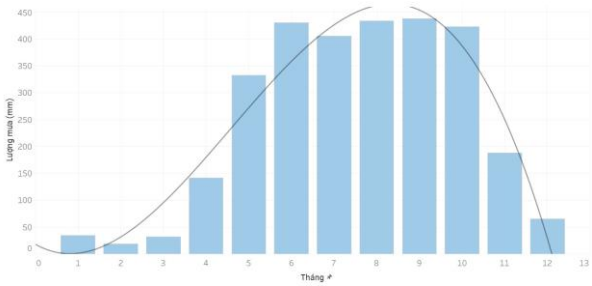
P-value: < 0.0001

Equation: Avg. Trung bình của Tổng lượng mưa trong tháng mỗi vùng miền (mm) = -2.17712*Tháng^3 + 29.9168*Tháng^2 + -44.5148*Tháng + 17.9944

Coefficients

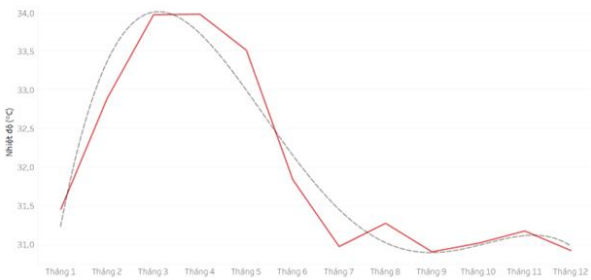
Term	Value	StdErr	t-value	p-value
Tháng^3	-2.17712	0.503269	-4.32596	0.0025257
Tháng^2	29.9168	9.92511	3.01425	0.0167051
Tháng	-44.5148	56.6663	-0.785561	0.454759
intercept	17.9944	88.6169	0.203059	0.844159

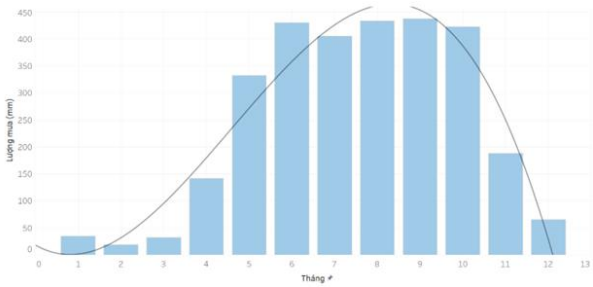
R-Squared: 0.9345



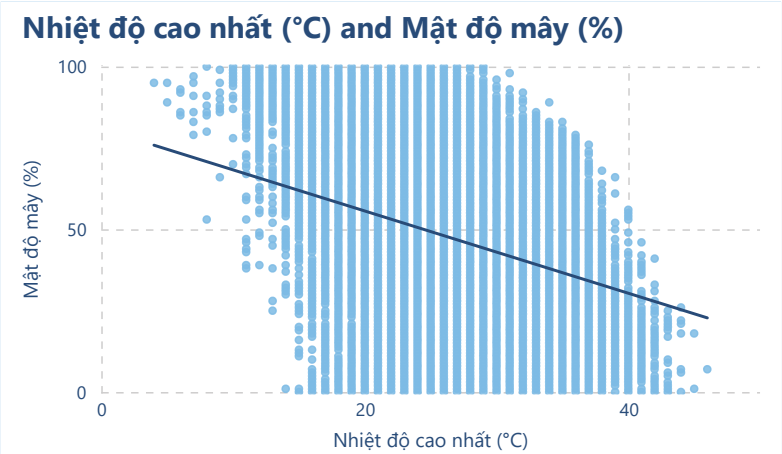
Kết luận:

- Tổng lượng mưa một tháng trong năm đạt đỉnh nhiều nhất tại tháng 8 và 9, chạm đáy ít nhất tại tháng 1 và 2. Kết quả mô hình kết hợp với dữ liệu ghi chép thực phản ánh được thời tiết thực tế trong cuộc sống, quý I có lượng mưa ít nhất, quý III có lượng mưa nhiều nhất;
- Kết hợp với biểu đồ mô hình dự đoán nhiệt độ cao nhất trong ngày cho thấy quý I có nhiệt độ cao nhất trong ngày đang ở giai đoạn sắp đạt đỉnh cực đại còn lượng mưa thì ở mức thấp nhất năm, quý III có lượng mưa nhiều nhất năm thì nhiệt độ cao nhất trong ngày thấp nhất năm. Có thể đưa ra cảnh báo rằng thời tiết ở Đông Nam Bộ ở quý I khá nóng bởi nhiệt độ cao và ít mưa để kìm hãm nhiệt độ, quý III mưa nhiều và nhiệt độ thấp nhất năm gây khó khăn trong việc phơi khô của các ngành nghề sản xuất và kinh tế cũng như đồ dùng cá nhân.





2.3. *Mối quan hệ giữa các yếu tố thời tiết*



Quan sát biểu đồ phân tán, ta thấy tồn tại mối quan hệ giữa nhiệt độ cao nhất trong ngày và mật độ mây.

Gọi H_0: Nhiệt độ cao nhất (°C) có sự tương quan với Mật độ mây (%);

H_1: Nhiệt độ cao nhất (°C) không tương quan với Mật độ mây (%).

Kiểm định giả thiết ta có kết quả:

```
P-value:   < 0.0001
Equation:  Mật độ mây (%) = -2.93935*Nhiệt độ cao nhất (°C) + 129.033
Coefficients
Term                        Value      StdErr     t-value    p-value
Nhiệt độ cao nhất (°C)      -2.93935   0.0105371  -278.952   < 0.0001
intercept                   129.033    0.31832    405.356    < 0.0001
```

Kết quả kiểm định cho thấy nhiệt độ cao nhất trong ngày tương quan tỷ lệ nghịch với mật độ mây với độ tin cậy 99%. Điều này đủ cơ sở để cảnh báo mọi người vùng Đông Nam Bộ nên mang vật dụng che chắn khi ở ngoài trời có nền nhiệt độ cao, bởi vì mật độ mây sẽ thưa thớt có thể làm mọi người bị say nắng và các bệnh lý khác.

3. Tích hợp AI

Công cụ Q&A (Hỏi đáp) trong Power BI là một tính năng cho phép người dùng tương tác với dữ liệu bằng ngôn ngữ tự nhiên thông qua dashboard. Người

dùng đặt câu hỏi bằng tiếng Anh, và Q&A sẽ tự động tạo ra các biểu đồ trực quan, bảng dữ liệu, số liệu chi tiết... để trả lời câu hỏi của người dùng.

Q&A được xây dựng dựa trên các công nghệ xử lý ngôn ngữ tự nhiên và học máy của Microsoft. Cụ thể, nó sử dụng các mô hình AI như LUIS (Language Understanding Intelligent Service) và Azure Cognitive Search để hiểu ý nghĩa của câu hỏi và tìm kiếm câu trả lời phù hợp trong dữ liệu. Như vậy để cho công cụ AI trở nên hiệu quả, cơ sở dữ liệu phải được định nghĩa truy vấn rõ ràng và chi tiết.

Các tính năng nổi bật của Q&A trong bản miễn phí bao gồm:

- Đặt câu hỏi bằng ngôn ngữ tự nhiên: Người dùng có thể đặt câu hỏi theo cách nói chuyện thông thường, không cần phải biết cú pháp truy vấn phức tạp.
- Tự động tạo biểu đồ: Q&A sẽ tự động chọn loại biểu đồ phù hợp nhất để hiển thị câu trả lời cho câu hỏi của bạn.
- Gợi ý câu hỏi: Q&A sẽ gợi ý các câu hỏi liên quan để giúp bạn khám phá dữ liệu của mình sâu hơn.