# Cách thiết kế trực quan hóa hiệu quả

Phạm Ngọc Hải - Tổng hợp sửa đổi biên soạn
Nguyễn Văn Thắng - Dịch phần 2: Các bước
Nguyễn Đức Nam - Dịch phần 3: Các lỗi

October 2024

## Mục lục

# 1 Mở đầu

Mục tiêu của chương này là cung cấp một số cách để thiết kế hình ảnh trực quan thành công vì nó là hình ảnh truyền tải thông tin mong muốn đến đối tượng mục tiêu một cách hiệu quả, chính xác. Sẽ có vô số phương pháp khả thi để ánh xạ các thành phần dữ liệu thành các ảnh. Tương tự cũng có rất nhiều công cụ tương tác có thể được cung cấp cho người xem. Việc lựa chọn các kết hợp kỹ thuật hiệu quả nhất không phải là một quá trình đơn giản.

Một hình ảnh trực quan có thể không hiệu quả vì một số lý do:

- quá khó hiểu hoặc phức tạp để đối tượng người nghe mục tiêu có thể hiểu được
- một số dữ liệu bị bóp méo, che khuất
- thiếu hỗ trợ việc sửa đổi chế độ xem hoặc kiểm soát bản đồ màu
- một bản trình bày không hấp dẫn về mặt thị giác

Chương này trước tiên trình bày các cân nhắc về [1] cách thiết kế cho các thành phần mà tác giả cảm thấy cần thiết cho một hình ảnh trực quan tốt. Sau đó, khám phá [2] vấn đề thường gặp trong hình ảnh trực quan và một số kỹ thuật để tránh những vấn đề này.

# 2 Các bước thiết kế trực quan

Việc tạo hình ảnh hóa bao gồm quyết định cách ánh xạ các trường dữ liệu thành các hình ảnh, lựa chọn và triển khai các phương pháp để sửa đổi chế độ xem và chọn lượng dữ liệu cần thiết cho trực quan hóa.

Thông tin bổ sung liên quan đến dữ liệu được hiển thị (ví dụ: nhãn) và ánh xạ (ví dụ: khóa màu) cũng rất cần thiết cho việc diễn giải và phải được tích hợp vào hình ảnh.

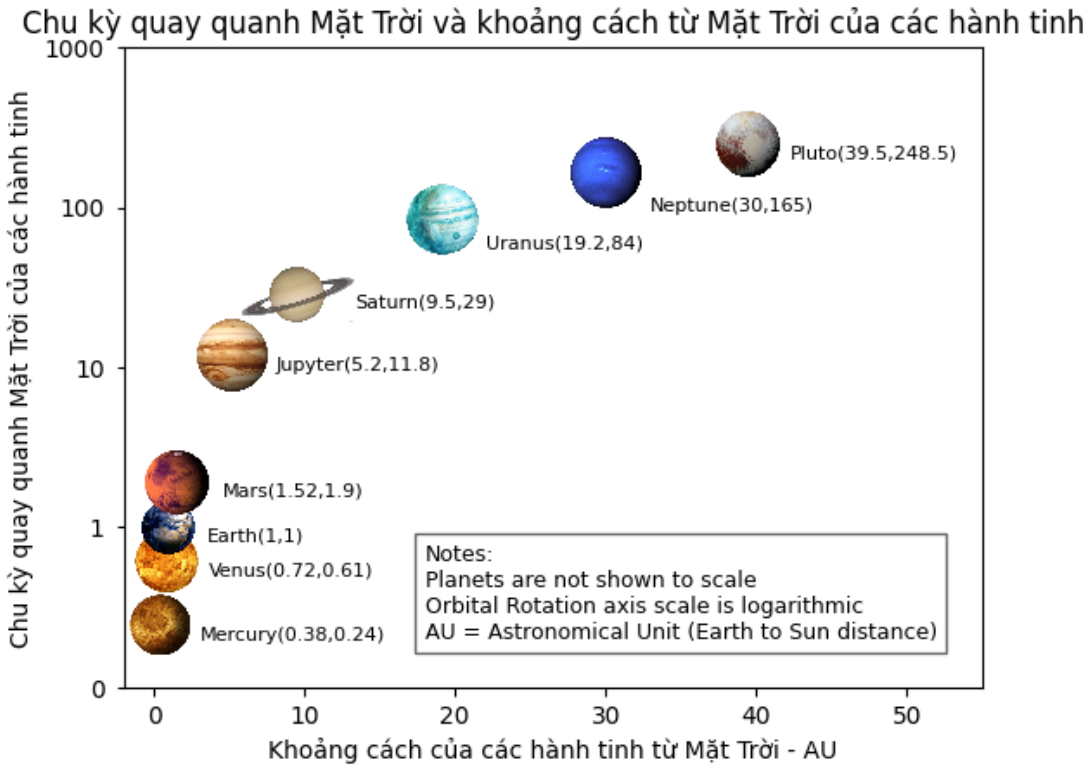
Cuối cùng, ít hữu hình hơn, cần cân nhắc đến tính thẩm mỹ tổng thể của màn hình hiển thị kết quả. Các phần nhỏ được trình bày sau đây sẽ làm rõ hơn các bước trong quy trình đã đề cập.

## 2.1 Bước 1. Ánh xạ dữ liệu sang biểu diễn trực quan

Để tạo ra hình ảnh trực quan hiệu quả, cần hiểu rõ ngữ nghĩa của dữ liệu và bối cảnh của người xem. Việc chọn cách hiển thị phù hợp với tư duy của người xem sẽ giúp họ dễ dàng hiểu hình ảnh hơn. Ngoài ra, nhà thiết kế nên nhất quán để tránh sự nhầm lẫn. Ánh xạ dữ liệu trực quan tốt sẽ giúp người xem diễn giải nhanh hơn vì không phải mất thời gian để hiểu.

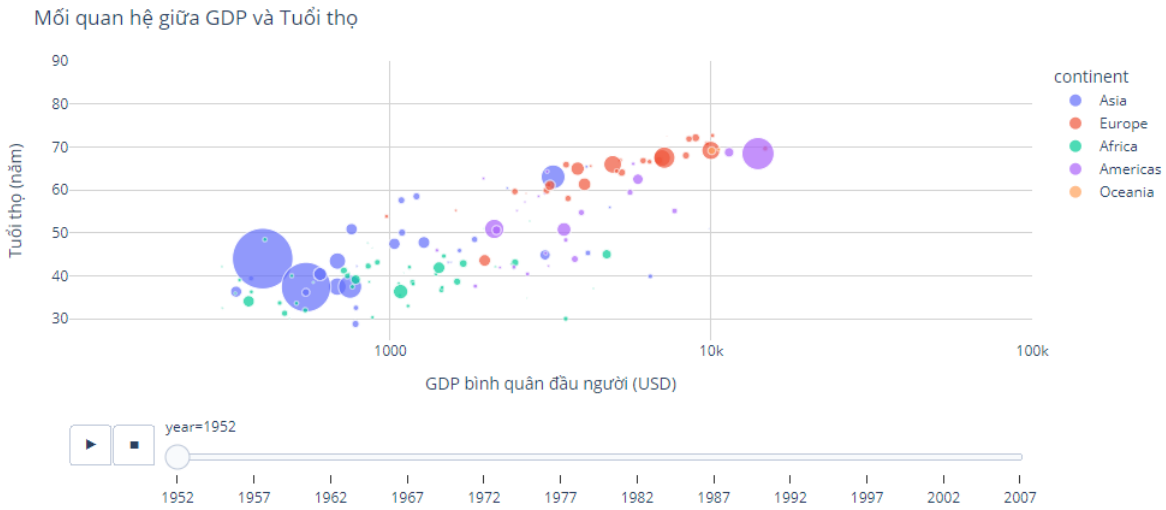
Ví dụ, trong Hình 1, hình ảnh các hành tinh được sử dụng để vẽ mối quan hệ giữa khoảng cách từ hành tinh đến Mặt Trời và thời gian quỹ đạo của nó.

Việc ánh xạ các thuộc tính dữ liệu không gian, để diễn giải bề mặt cầu 3 chiều của Trái Đất thành 1 mặt phẳng 2D là cách ánh xạ phổ biến và trực quan nhất được tìm thấy trong các hình ảnh trực quan. Ví dụ này là sớm nhất để tận dụng khả năng của con người trong việc liên hệ vị trí trên mặt phẳng 2D với vị trí trong thế giới ba chiều.




Hình 1: Sử dụng các ký hiệu biểu đồ phân tán trực quan để hiển thị khoảng cách từ các hành tinh đến mặt trời so và thời gian quỹ đạo của nó.

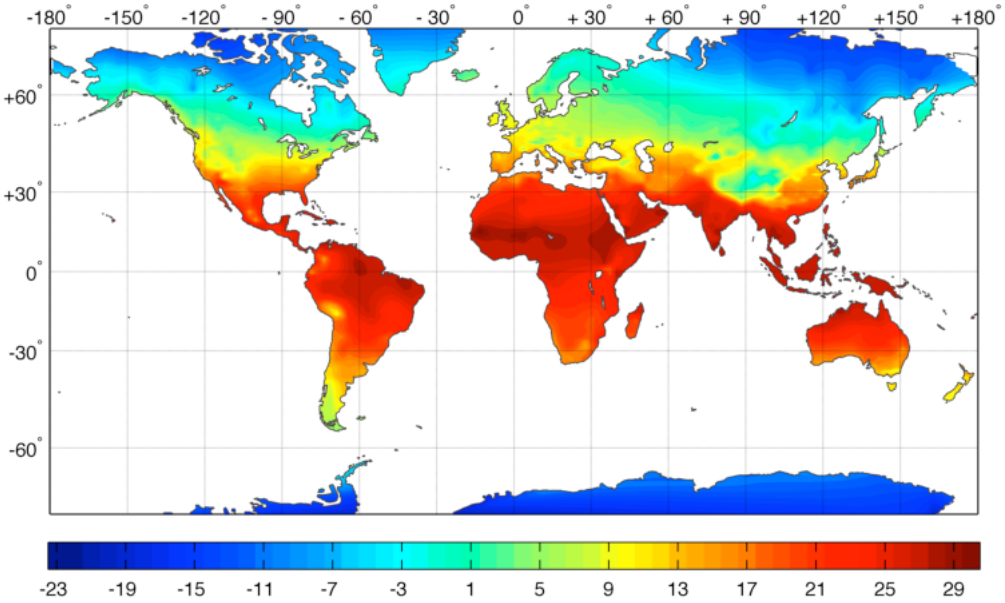
Tương tự như vậy, với sự ra đời của hoạt hình, rõ ràng là việc hiển thị các tập dữ liệu chuỗi thời gian thông qua hoạt hình là khá trực quan, với lợi thế bổ sung là cho phép thời gian thay đổi cả về tốc độ và hướng.




Hình 2: Sử dụng hoạt họa để thể hiện mối quan hệ giữa GDP bình quân đầu người (gdpPercap) và tuổi thọ trung bình (lifeExp) của các quốc gia qua các năm.

Một số ánh xạ dữ liệu trở nên trực quan hơn khi kết hợp với ngữ cảnh cụ thể:

- Ví dụ, ánh xạ nhiệt độ sang màu sắc rất phổ biến, ví dụ xanh đỏ đại diện cho nhiệt độ thấp và nhiệt độ cao.

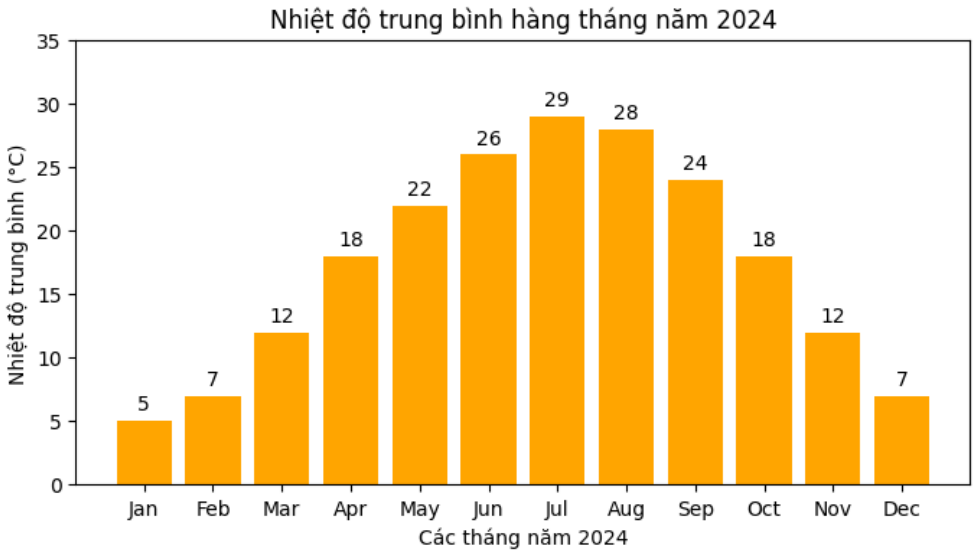



Hình 3: Sử dụng màu sắc để thể hiện mức độ của nhiệt độ.

- Trong các lĩnh vực như bản đồ học và địa chất, màu sắc thường được dùng để phân loại đất đai hoặc lớp địa chất, vì vậy việc chọn màu phải phù hợp với bối cảnh ứng dụng.

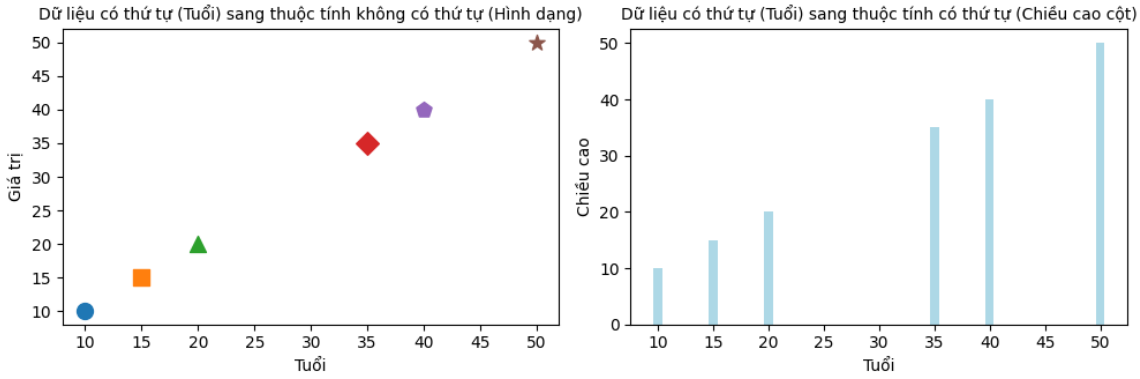

Hình 4: Sử dụng màu sắc để phân loại đất đai.

- Chiều cao hoặc độ dài cũng có thể dùng để biểu diễn nhiệt độ, giống như cách chúng ta đọc nhiệt kế. Đối với các bác sĩ, việc dùng độ dài để hiển thị áp lực hoặc các giá trị liên quan có thể rất tự nhiên.




Hình 5: Sử dụng chiều cao để biển diễn nhiệt độ trung bình hàng tháng năm 2024.

Khi chọn ánh xạ, cần xem xét tính tương thích giữa thang dữ liệu và thuộc tính. Dữ liệu có thứ tự (như tuổi) không nên ánh xạ sang thuộc tính không có thứ tự (như hình dạng). Tương tự, dữ liệu không có thứ tự (như quốc gia) không nên ánh xạ sang thuộc tính có thứ tự (như độ dài).




Hình 6: Dữ liệu có thứ tự (Tuổi) không nên ánh xạ sang thuộc tính không có thứ tự (Hình dạng).

Tuy nhiên, đôi khi cũng thú vị khi kiểm tra dữ liệu bằng các ánh xạ không trực quan, vì hình ảnh kết quả có thể tiết lộ một thuộc tính thú vị trong dữ liệu. Ví dụ, ánh xạ thời gian để tô màu dọc theo một đường vạch có thể tiết lộ các biến thể về tốc độ hạt mà nếu không thì có thể khó phát hiện. Do đó, một nguyên tắc chung hữu ích là thiết lập các ánh xạ mặc định dựa trên lựa chọn trực quan nhất theo người xem thông thường, nhưng, đặc biệt đối với các tác vụ khám phá, cho phép người xem tùy chỉnh nhiều ánh xạ trực quan khác nhau.

## 2.2 Bước 2. Chọn và sửa đổi chế độ xem cho phù hợp

Ngoại trừ các tập dữ liệu khá đơn giản, một chế độ xem hiếm khi đủ để truyền tải tất cả thông tin chứa trong dữ liệu. Như vậy điều quan trọng là phải có thể dự đoán các loại chế độ xem mà được sử dụng nhiều nhất bởi người xem thông thường và sau đó cung cấp

trực quan cách điều khiển cài đặt, tùy chỉnh các dạng xem mà người xem cảm thấy phù hợp.

Cần ghi nhớ chế độ xem phù hợp phụ thuộc vào:

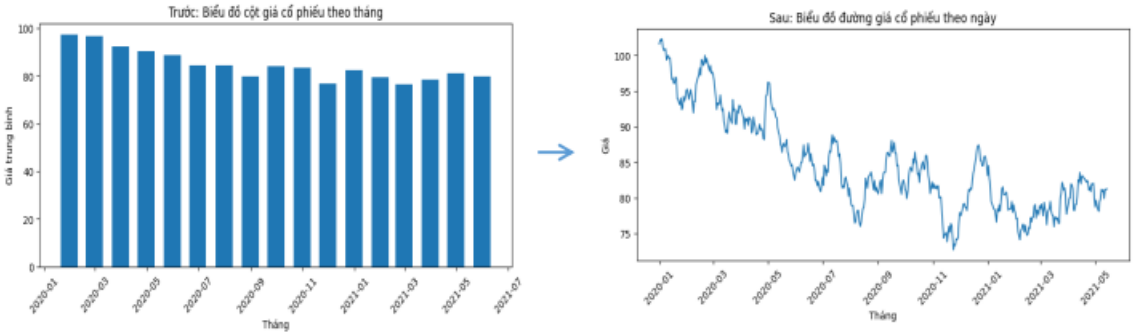
- loại dữ liệu được trình bày
- nhiệm vụ gắn liền với sự trực quan hóa.
- Mỗi chế độ xem phải có nhãn, cung cấp thông tin rõ ràng, đầy đủ
- và người xem cần ít hành động nhất có thể để sửa sang được chế độ xem mà họ thấy phù hợp.

Sau đây là ví dụ một số hành động của người xem để sửa đổi chế độ xem:

- Cuộn và phóng to: Cần thiết khi không thể hiển thị toàn bộ dữ liệu ở độ phân giải mong muốn.

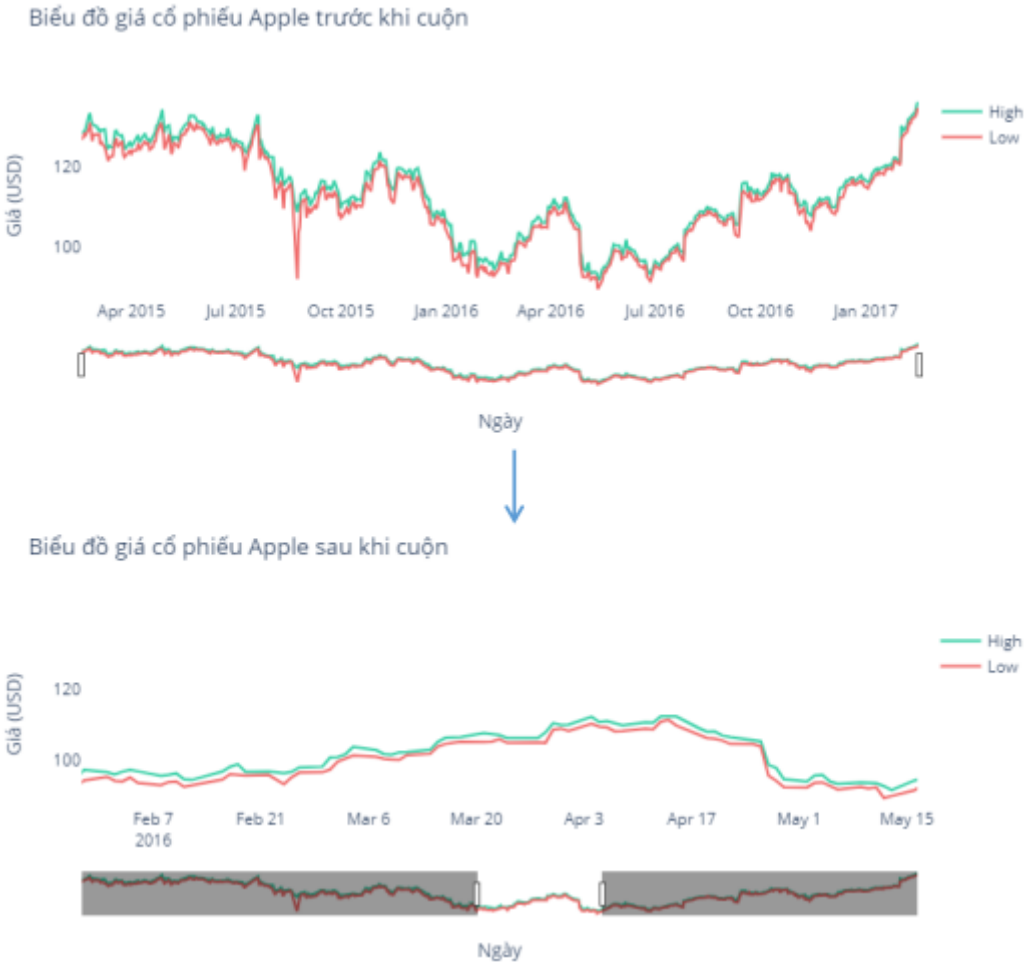

Hình 7: Bản đồ trước và sau khi phóng to để thấy được các chi tiết của dữ liệu.

- Điều khiển bảng màu: Luôn cần thiết, ít nhất nên hỗ trợ một số bảng màu khác nhau để cho phép người xem điều chỉnh màu sắc riêng lẻ hoặc toàn bộ bảng màu.
- Điều khiển ánh xạ: Cho phép người xem chuyển đổi giữa các cách hiển thị khác nhau của cùng một dữ liệu (trong các tool BI đã tích hợp). Lý do vì một số đặc điểm có thể bị ẩn trong ánh xạ này nhưng lại nổi bật trong ánh xạ khác.




Hình 8: Ánh xạ dữ liệu từ biểu đồ cột sang biểu đồ đường.

- Điều khiển tỷ lệ: Cho phép người xem thay đổi phạm vi và phân phối giá trị cho một trường dữ liệu trước khi ánh xạ. Việc lọc và cắt dữ liệu cũng giúp người xem tập trung vào các tập con cụ thể.




Hình 9: Biểu đồ thể hiện giá trị cổ phiếu Apple trước và sau khi thay đổi phạm vi giá trị của Ngày.

- Điều khiển mức độ chi tiết: Cung cấp khả năng loại bỏ hoặc làm nổi bật chi tiết, hỗ trợ các góc nhìn ở các mức độ khác nhau.

**Lưu ý** cho việc thiết kế chế độ xem và tùy chọn sửa đổi cho người xem:

- Trong mọi trường hợp, điều cần thiết là các thao tác xem phải đơn giản, dễ nhớ cho người xem và cung cấp những thông tin phù hợp, chính xác cho nhiệm vụ.
- Nếu có thể, thao tác sửa đổi chế độ xem trực tiếp (trực tiếp có thể thay đổi chế độ xem bằng các thao tác đơn giản: click, gõ phím, ...) thường được ưa thích.

## 2.3 Bước 3. Xác định mật độ thông tin phù hợp

Khi thiết kế trực quan hóa, quan trọng là xác định lượng thông tin cần hiển thị. Có hai hệ quả:

- Đồ họa thừa: Có quá ít thông tin để trình bày. Ví dụ, việc chỉ cần hiển thị tỷ lệ nam và nữ có thể chỉ dùng một con số. Một số đồ họa có thể cố gắng "làm đầy" thêm thông tin bằng cách hiển thị nhiều giá trị hơn, nhưng trong những trường hợp này, chỉ cần hiển thị các giá trị dưới dạng văn bản sẽ hiệu quả hơn.
- Thông tin thừa: Việc trực quan hóa chứa quá nhiều thông tin → nó có thể gây nhầm lẫn và khó hiểu. Thông tin quan trọng có thể bị mất trong một giao diện lộn xộn, khiến người xem khó xác định nơi cần chú ý (Hình 15).

Giải pháp cho vấn đề quá tải thông tin:

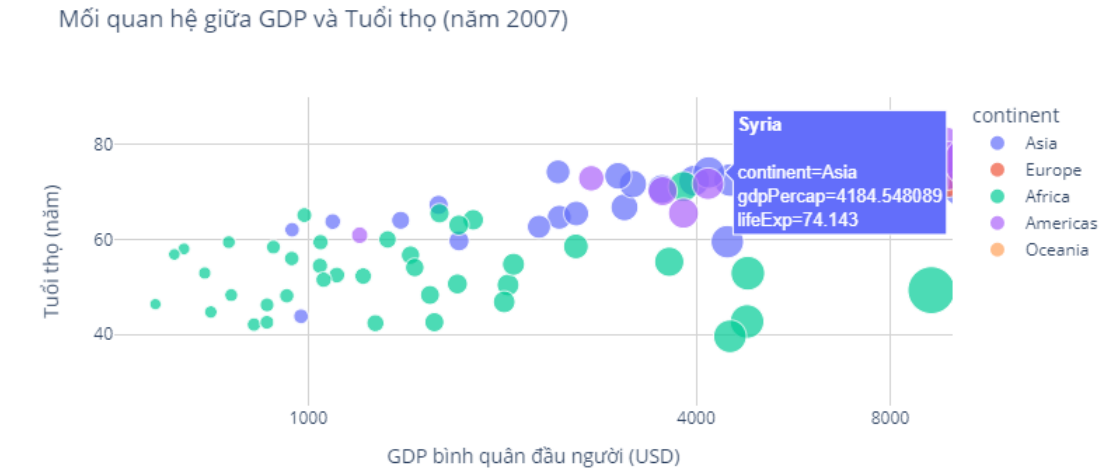
- Tùy chọn hiển thị: Cho phép người xem bật hoặc tắt các thành phần khác nhau, giúp họ tập trung vào thông tin quan trọng.
- Nhiều màn hình: Sử dụng các màn hình riêng biệt để hiển thị 1 nhóm thông tin cụ thể mà không gây lộn xộn.
- Lọc dữ liệu: Loại bỏ các điểm dữ liệu không quan trọng để người xem chỉ tập trung vào các phần có ý nghĩa.
- Tỉ lệ: Điều chỉnh kích thước của một số dữ liệu cho phân bổ trên không gian màn hình hợp lý hơn.

Những giải pháp này giúp tối ưu hóa trực quan hóa, đảm bảo rằng người xem dễ dàng hiểu và tương tác với thông tin.

## 2.4 Bước 4. Thêm vào các từ khóa, nhãn và chú thích

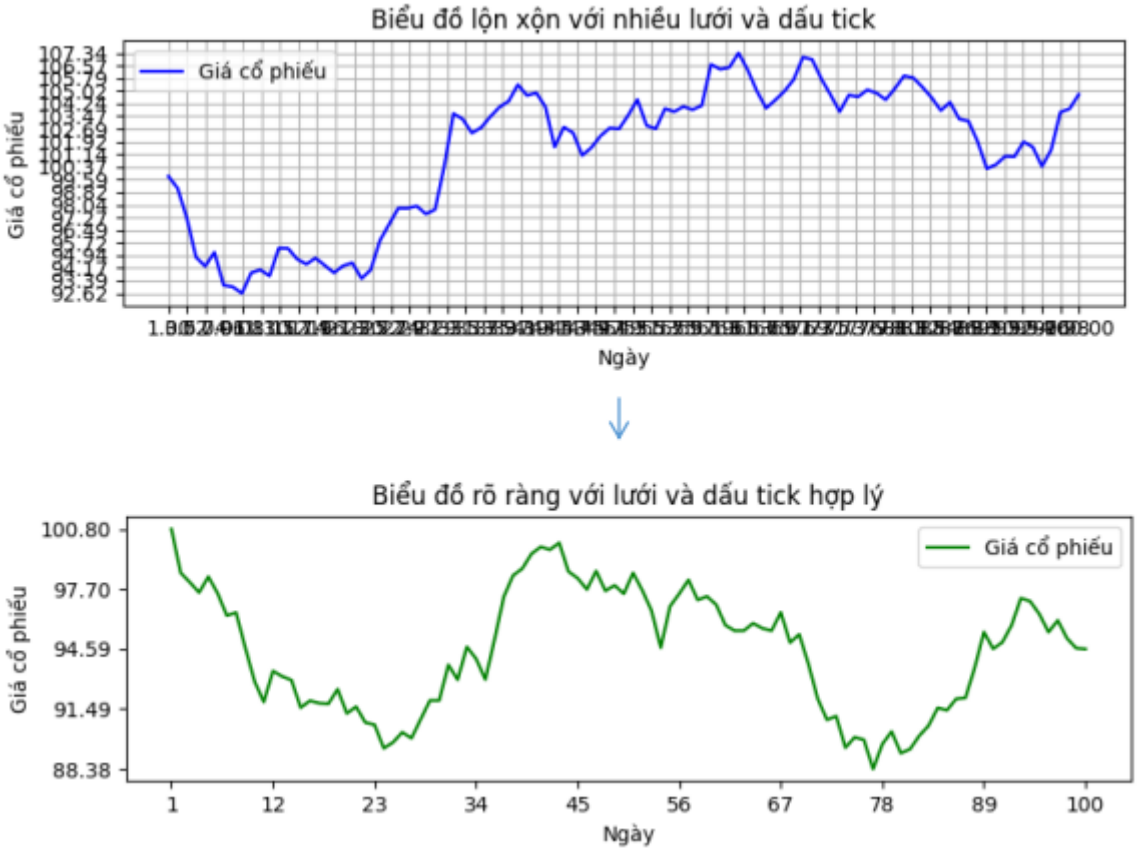
Một vấn đề phổ biến trong trực quan hóa là thiếu thông tin hỗ trợ để người xem có thể hiểu rõ và chính xác thông tin truyền tải. Để giải quyết vấn đề đó, chúng ta cần cung cấp các yếu tố sau:

- Chú thích chi tiết: Giải thích dữ liệu và phép ánh xạ dữ liệu sang biểu đồ trực quan được sử dụng.
- Nhãn trục: Ghi rõ đơn vị đo lường trên các trục.
- Chú giải ký hiệu: Cung cấp từ khóa cho các ký hiệu, nằm ở viền hoặc trong một widget riêng.
- Giải thích màu sắc: Ý nghĩa của màu sắc sử dụng trong phép trực quan, như thanh màu có nhãn.




Hình 10: Biểu đồ thể hiện mối quan hệ giữa GDP và tuổi thọ của các quốc gia thuộc các châu lục năm 2007.

- Lưới và dấu tick: Hiển thị các giá trị và phạm vi của các trường số khi cần đánh giá chính xác.




Hình 11: Biểu đồ thể hiện lộn xộn và rõ ràng khi sử dụng lưới và dấu tick.

## 2.5 Bước 5. Điều chỉnh màu sắc được sử dụng cho phù hợp

Màu sắc thường bị lạm dụng trong các biểu đồ, dẫn đến sự nhầm lẫn hoặc diễn giải sai.

Thêm nữa, việc chọn sai bảng màu hoặc cố gắng truyền tải quá nhiều thông tin qua màu sắc có thể làm giảm hiệu quả trực quan hóa.

Đặc biệt, do màu sắc có thể bị ảnh hưởng bởi môi trường quan sát và nhiều người mắc chứng mù màu, điều này càng làm phức tạp quá trình thiết kế.

Dưới đây là hướng dẫn sử dụng màu sắc hiệu quả trong biểu đồ:

- Giới hạn số lượng màu sắc, tránh làm rối mắt người xem.
- Sử dụng thêm phương pháp ánh xạ bổ sung, ví dụ như kết hợp cả màu sắc và kích thước, để dễ dàng truyền đạt thông tin.
- Ngoài ra, khi tạo bảng màu cho dữ liệu số, có thể thay đổi sắc độ (hue) và độ sáng (lightness) (về cơ bản khi thay đổi độ sắc và độ sáng sẽ tạo ra các màu khác nhau) để giúp dễ phân biệt các đối tượng.

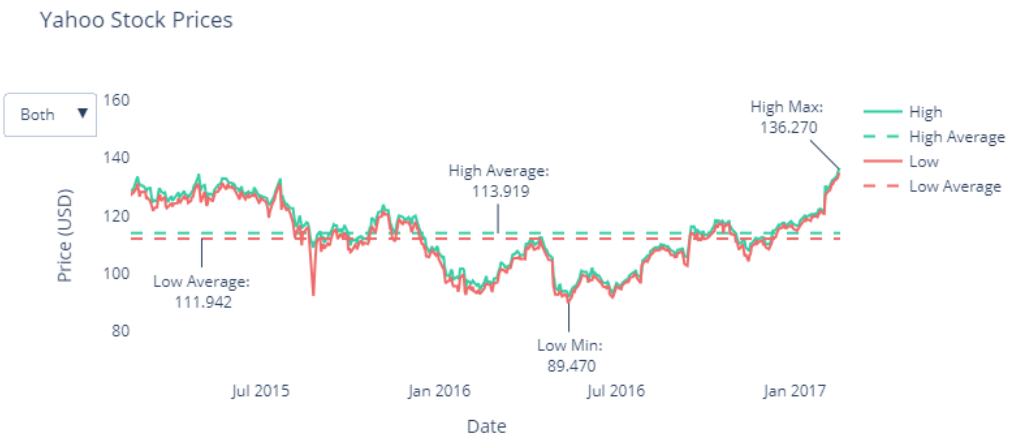

Hình 12: Biểu đồ thể hiện số lượng dân cư (các điểm dữ liệu lớn thì có lượng dân cư lớn và ngược lại) của các quốc gia thuộc các châu lục năm 2002.

## 2.6 Bước 6. Bước cuối cùng, đảm bảo thẩm mỹ

Đây là bước chúng ta thực hiện cân bằng giữa chức năng và hình thức. Một biểu đồ tốt nên vừa cung cấp thông tin vừa hấp dẫn về mặt hình ảnh.

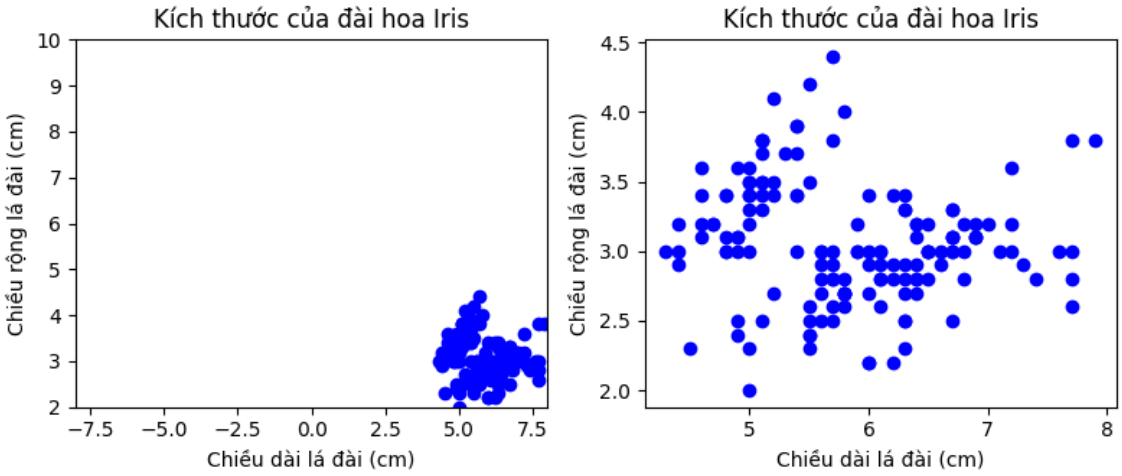
Dưới đây là hướng dẫn nâng cao tính thẩm mỹ:

- Tập trung: Nên làm nổi bật những phần quan trọng nhất của biểu đồ để người xem biết nên chú ý vào đâu. Nếu không có sự nhấn mạnh thích hợp, thông tin quan trọng có thể bị bỏ qua.



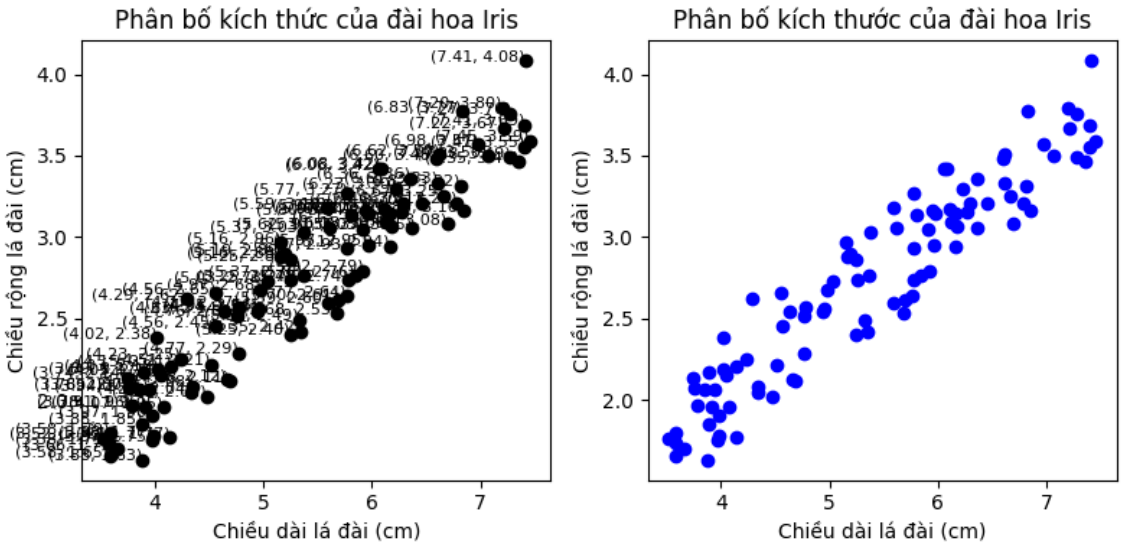

Hình 13: Làm nổi bật giá trị min và max của giá cổ phiếu.

- Cân bằng: Sử dụng không gian màn hình hợp lý. Đặt các thành phần chính ở trung tâm và tránh nổi bật đường viền hoặc khu vực ít quan trọng.




Hình 14: Sử dụng không gian không hợp lý và hợp lý khi thể hiện dữ liệu.

- Đơn giản: Giữ lượng thông tin thể hiện trực quan ở mức vừa đủ bằng những thiết kế tối giản. Nên dùng những hình ảnh đơn giản và tránh dùng các thiết kế phức tạp nếu những thiết kế đơn giản hơn có thể truyền đạt thông điệp tương tự. Một kỹ thuật hữu ích là loại bỏ từng yếu tố và xem xét liệu việc mất thông tin có chấp nhận được không. Ghi nhớ "tối giản" - đơn giản mà vẫn đạt yêu cầu đặt ra.




Hình 15: Thông tin được thể hiện vừa đủ và thừa thãi khi cùng một mục đích thể hiện sự phân bố của kích thước của đài hoa Iris.

# 3 Các lỗi sai hay gặp khi thiết kế trực quan

Ngay cả khi tuân thủ quy trình các bước thiết kế trực quan khoa học nêu trên, vẫn có thể có vài lỗi sai, vấn đề gặp phải.

Những vấn đề này do nhiều lý do:

- Liên quan đến quyết định **nên trực quan hóa điều gì**
- **phương pháp ánh xạ trực quan nào là phù hợp** nhất để sử dụng.
- Một số vấn đề liên quan đến việc **bóp méo dữ liệu**, dù cố ý hay vô tình, có thể dẫn đến sự hiểu sai.
- Các vấn đề khác bao gồm việc **che giấu dữ liệu thực sự** đằng sau các phiên bản "đã được làm sạch" hoặc các đồ họa phụ trợ quá thừa thãi đi kèm.

Trong tất cả các trường hợp này, vẫn có thể thực hiện các bước để cải thiện chất lượng và tính "trung thực" của hình ảnh trực quan. Sau đây chúng ta sẽ bắt đầu.

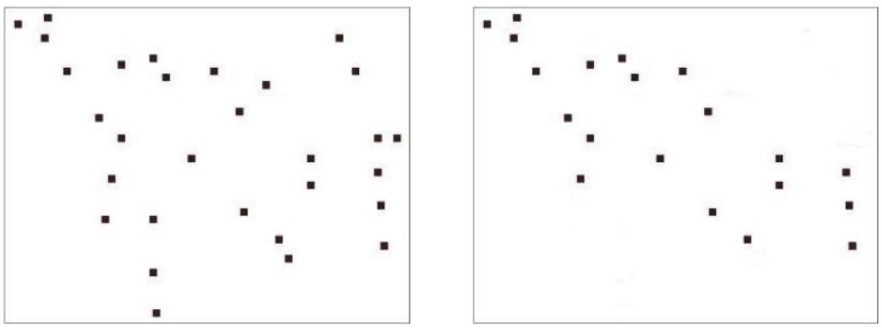
## 3.1 Hình ảnh dễ gây hiểu lầm

Một trong những quy tắc quan trọng nhất của trực quan hóa là nó phải mô tả chính xác dữ liệu. Tuy nhiên, trong suốt lịch sử, có nhiều ví dụ về việc trực quan hóa dữ liệu bị bóp méo được sử dụng để thay đổi quan điểm và lừa dối khán giả. Những "ảnh giả" này có thể xuất hiện ở khắp mọi nơi, từ các tạp chí danh tiếng cho đến các danh mục của công ty.

Trong phần này, chúng ta sẽ xác định một số chiến lược phổ biến trong việc tạo ra các hình ảnh trực quan gây hiểu lầm, không phải để người đọc áp dụng, mà là để tránh!

### 3.1.1 Làm sạch dữ liệu

Dữ liệu thô thường có nhiều yếu tố "lộn xộn", chẳng hạn như những điểm dữ liệu bất thường (outliers) hoặc lỗi thu thập dữ liệu. Trong quá trình làm trực quan, người làm trực quan hóa có thể bị cám dỗ loại bỏ những điểm dữ liệu không hợp lý để giúp hình ảnh trông dễ nhìn hơn hoặc để loại bỏ các dữ liệu không phù hợp với quan điểm mà họ muốn thể hiện.



(a) (b)

Hình 16: (a) Dữ liệu thô cho thấy thiếu mối tương quan (b) Dữ liệu đã làm sạch tiết lộ mối tương quan giả

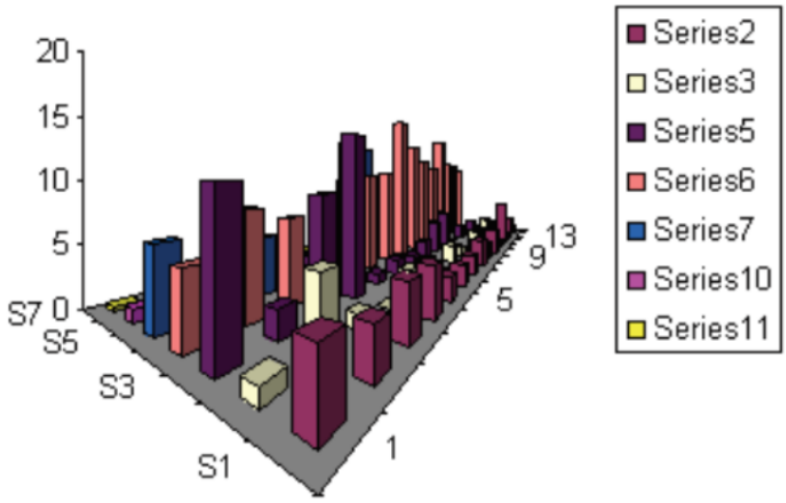
Loại bỏ các điểm ngoại lệ là một cách phổ biến trong trường hợp này. Các điểm ngoại lệ là những giá trị không bình thường trong dữ liệu, có thể là do lỗi kỹ thuật hoặc là thực tế hiếm hoi nhưng hợp lệ.

Khi những điểm ngoại lệ này bị loại bỏ không có lý do hợp lý hoặc không thông báo cho người xem, hình ảnh trực quan có thể trở nên thiên lệch và gây hiểu lầm.

**Giải pháp**: Trừ khi có bằng chứng rằng các điểm ngoại lệ là kết quả của lỗi kỹ thuật, chúng không nên bị loại bỏ một cách tùy tiện. Đồng thời, người thiết kế cần cung cấp cho người xem tùy chọn hiển thị hoặc ẩn những điểm ngoại lệ này để đảm bảo tính minh bạch của dữ liệu.

### 3.1.2 Tỷ lệ không cân đối

Tỷ lệ là một công cụ mạnh mẽ trong hình ảnh trực quan, vì việc lựa chọn kỹ lưỡng các yếu tố tỷ lệ có thể giúp làm lộ ra các mẫu và cấu trúc không thể nhìn thấy ở các góc nhìn không được tỷ lệ hóa.




Hình 17: Vis Lies: Sự bóp méo kích thước do phối cảnh (Phối cảnh là một kỹ thuật giúp tái hiện các đối tượng trong không gian ba chiều (3D) lên một bề mặt hai chiều (2D). Mục đích của phối cảnh là tạo cảm giác về chiều sâu và khoảng cách trong hình ảnh, làm cho các đối tượng trông giống như chúng tồn tại trong không gian thực tế.)

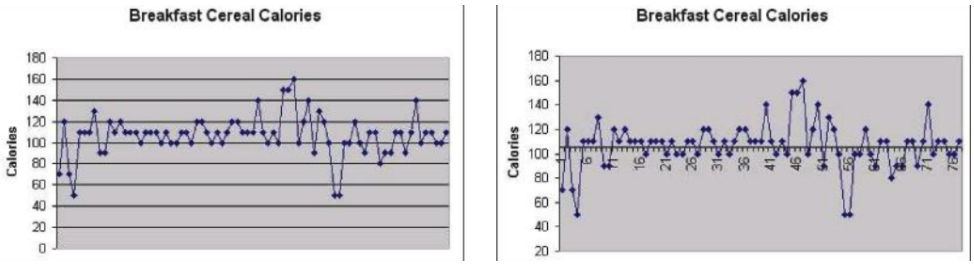
Tuy nhiên, tỷ lệ cũng có thể bị lợi dụng để lừa dối người xem. Ví dụ, thay đổi tỷ lệ của các trục hay kích thước các yếu tố đồ họa có thể làm cho một xu hướng dường như mạnh hơn hoặc yếu hơn so với thực tế. Điều này tạo ra một sự sai lệch, khiến người xem đánh giá không chính xác sự thay đổi trong dữ liệu.

Edward Tufte, một chuyên gia về hình ảnh trực quan, đưa ra khái niệm "**lie factor**", chỉ tỷ lệ giữa sự thay đổi thực tế trong dữ liệu và sự thay đổi trong hình ảnh. Nếu tỷ lệ này quá cao hoặc thấp, hình ảnh đó có thể bị coi là lừa dối. Ví dụ (Hình 15), trong biểu đồ mà tỷ lệ các vật thể ở tiền cảnh và hậu cảnh bị bóp méo bởi phối cảnh, người xem có thể nghĩ rằng các đối tượng ở gần lớn hơn đáng kể so với thực tế.

**Giải pháp**: Cung cấp các công cụ như đường kẻ, điểm tham chiếu, hoặc chỉ báo điểm gốc để người xem có thể hiểu rõ hơn về cách dữ liệu được tỷ lệ hóa và có thể so sánh chính xác hơn.

### 3.1.3 Biến dạng phạm vi

Người xem thường có kỳ vọng về phạm vi của một chiều dữ liệu nhất định. Bằng cách điều chỉnh phạm vi này khác xa so với kỳ vọng, người dùng có thể bị lừa dẫn đến sự hiểu sai. Điều này thường xảy ra khi trục được di chuyển sao cho không còn tương ứng với giá trị "mốc 0" như mong đợi (xem Hình 18).



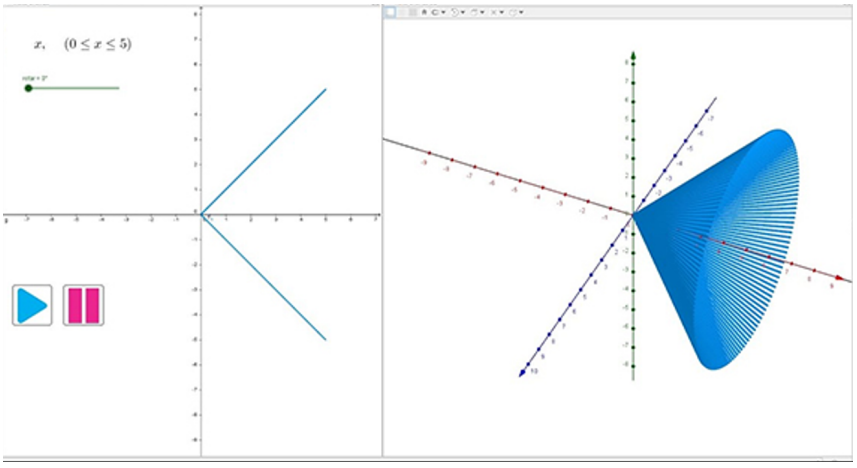

Hình 18: Đồ thị với 2 đường cơ sở khác nhau

Do con người có khả năng phán đoán tương đối rất mạnh mẽ, tức là chúng ta thường so sánh các đối tượng với nhau thay vì chỉ đánh giá từng đối tượng một cách độc lập, việc thay đổi điểm gốc – tức là giá trị mà từ đó chúng ta bắt đầu so sánh – có thể ảnh hưởng cách chúng ta hiểu và diễn giải thông tin.

**Giải pháp**: Người thiết kế cần minh bạch với người xem về điểm gốc của biểu đồ và lý do thay đổi nếu có. Việc cho phép người xem tùy chọn điều chỉnh điểm gốc cũng là một cách giúp họ không bị lừa dối bởi phạm vi dữ liệu.

### 3.1.4 Lạm dụng tính đa chiều

Hình ảnh trực quan càng nhiều chiều thì càng dễ gây hiểu lầm. Điều này là do con người gặp nhiều khó khăn hơn khi đánh giá thể tích (3 chiều) so với diện tích (2 chiều), và diện tích lại khó hơn chiều dài (1 chiều).



Hình 19: Giá trị vô hướng (scalar) được đánh giá thông qua đồ thị 2D và 3D

Ví dụ: Nếu một giá trị vô hướng (scalar) được đánh giá thông qua đồ thị 3D, người xem có thể gặp khó khăn trong việc đánh giá chính xác sự khác biệt giữa các yếu tố, dẫn đến khả năng hiểu sai về mối quan hệ dữ liệu.

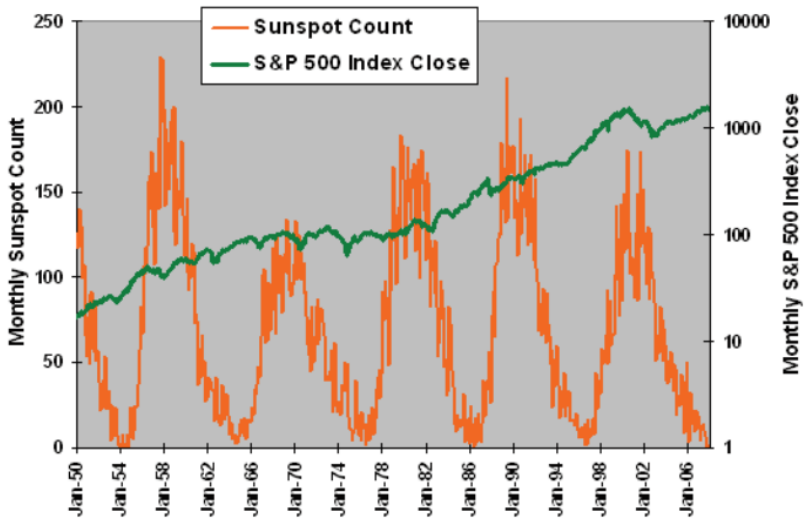
**Giải pháp**: Việc đơn giản hóa hình ảnh trực quan sẽ giúp người xem dễ dàng nắm bắt thông tin hơn. Chỉ nên sử dụng nhiều chiều dữ liệu khi thật sự cần thiết và đảm bảo rằng người xem có thể hiểu rõ ý nghĩa của chúng.

## 3.2 Hình ảnh không có nghĩa

Hình ảnh trực quan được thiết kế để truyền tải thông tin, và điều quan trọng là thông tin đó phải có ý nghĩa. Tuy nhiên khi thiết kế, một số lỗi mắc phải đã khiến hình ảnh trực quan trở nên vô nghĩa.

### 3.2.1 Kết hợp các quan hệ ngẫu nhiên như là một quan hệ nhân quả có tương quan

Các hình ảnh trực quan thường được tạo ra bằng cách kết hợp các tập dữ liệu từ nhiều nguồn khác nhau. Tuy nhiên, việc kết hợp các thành phần không liên quan vào một hình ảnh duy nhất rất dễ xảy ra, và từ đó có thể nhận thấy những gì dường như là một cấu trúc, ví dụ, vẽ biểu đồ giá trị thị trường chứng khoán so với sự xuất hiện của vết đen mặt trời (xem Hình 20).




Hình 20: Một biểu đồ vô nghĩa, thể hiện sự xuất hiện của vết đen mặt trời so với chỉ số SP 500

Trong trường hợp này, mối quan hệ ngẫu nhiên bị nhầm lẫn với mối quan hệ nhân quả. Khi quyết định kết hợp dữ liệu nào, điều quan trọng là phải đảm bảo rằng có một logic nhất định trong sự kết hợp đó.

Một trong những vấn đề được tìm thấy trong các quy trình nhận diện mẫu phân tích dữ liệu là những mối quan hệ không liên quan thường được phát hiện và báo cáo, sau đó phải được loại bỏ bởi các chuyên gia trong lĩnh vực.

### 3.2.2 So sánh trong phạm vi không gian và thời gian không tương quan

Một yếu tố khác cần được xem xét là sự tương thích giữa phạm vi thời gian và không gian của dữ liệu khi so sánh. Ví dụ, không nên so sánh doanh số bán hàng của một sản phẩm cụ thể trong một năm ở một khu vực nhất định của đất nước với doanh số bán hàng của cùng sản phẩm đó ở một khu vực và năm khác, trừ khi có giả thuyết về sự thay đổi trong xu hướng quan tâm đến sản phẩm đó.

### 3.2.3 So sánh trong phạm vi đơn vị không chuẩn hóa

Sự tương thích về đơn vị cũng cần được kiểm tra khi tạo tập dữ liệu cho hình ảnh trực quan. Ví dụ, các sản phẩm thực phẩm được đo bằng giá theo thể tích thường bị trộn lẫn với những sản phẩm được đo bằng giá theo trọng lượng. Một hình ảnh trực quan hiệu quả có thể chuẩn hóa cả hai đơn vị khác nhau về chung 1 thang đo về giá theo khẩu phần ăn.

### 3.2.4 Gán thứ tự cho các đối tượng dữ liệu không có quan hệ thứ tự

Dữ liệu phân loại là loại dữ liệu không có thứ tự tự nhiên, chẳng hạn như tên công ty hoặc loại sản phẩm. Khi các dữ liệu này được biểu diễn trên một đồ thị với vị trí cụ thể, có thể tạo ra cảm giác rằng chúng có thứ tự hoặc có thể so sánh như dữ liệu liên tục.

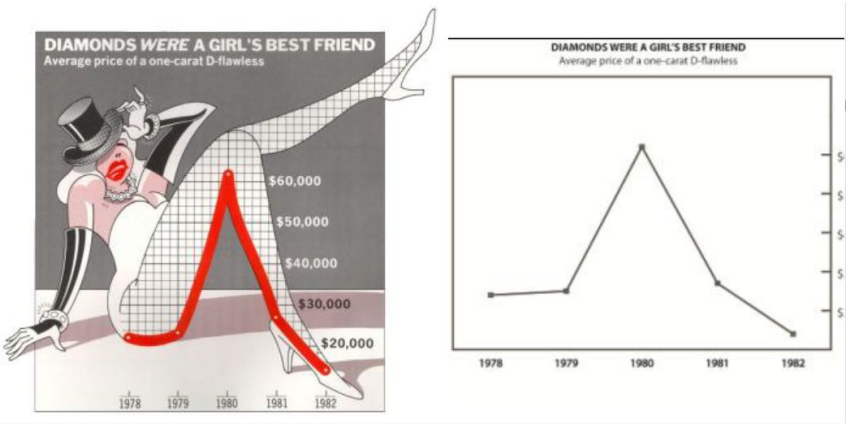
Ví dụ: Giả sử bạn có một danh sách các công ty và bạn muốn biểu diễn chúng trên một đồ thị. Nếu bạn gán mỗi công ty một vị trí trên trục x của đồ thị và cố gắng vẽ một đường thẳng hoặc đường cong qua các điểm này, bạn đang áp dụng một phép toán liên tục (vẽ đường) cho dữ liệu phân loại (tên công ty).

**Giải pháp:** Điểm mấu chốt là cần có sự cân nhắc về mặt ngữ nghĩa của hình ảnh trực quan để đảm bảo rằng nó có ý nghĩa logic.

## 3.3 Dữ liệu bị che mờ

"**Chart junk**" là thuật ngữ được Tufte sử dụng để chỉ các yếu tố đồ họa bổ sung không cần thiết, làm cho hình ảnh trở nên phức tạp hơn mà không hỗ trợ việc diễn giải dữ liệu. Những yếu tố này có thể gây rối mắt, làm người xem khó tập trung vào dữ liệu chính.

Việc xác định lượng đồ họa bổ sung cần đưa vào một hình ảnh trực quan có thể là một quá trình khó khăn, vì nhà thiết kế có thể không hiểu rõ nhu cầu của tất cả người xem tiềm năng.




Hình 21: Hình ảnh đồ họa được thêm vào không cần thiết làm người xem không tập trung được vào dữ liệu

Khác với các biểu đồ tĩnh của Tufte, các hình ảnh trực quan hiện đại có thể linh hoạt và tùy biến, cho phép người xem điều chỉnh loại và mật độ thông tin hỗ trợ. Trong một số nhiệm vụ, người xem có thể chuyển đổi giữa cái nhìn tổng quát định tính và phân tích định lượng. Đối với cái nhìn định tính, việc làm rõ dữ liệu là rất quan trọng, trong khi đối với phân tích định lượng, các công cụ giúp định lượng các yếu tố là rất cần thiết.

**Giải pháp:** Một quy tắc tốt là cung cấp đủ công cụ để đáp ứng nhu cầu định lượng của người xem, nhưng đồng thời cũng cho phép họ có tùy chọn tắt hoặc bật này để giảm bớt sự lộn xộn trong hình ảnh trực quan.

## 3.4 Dữ liệu thô so với dữ liệu suy diễn

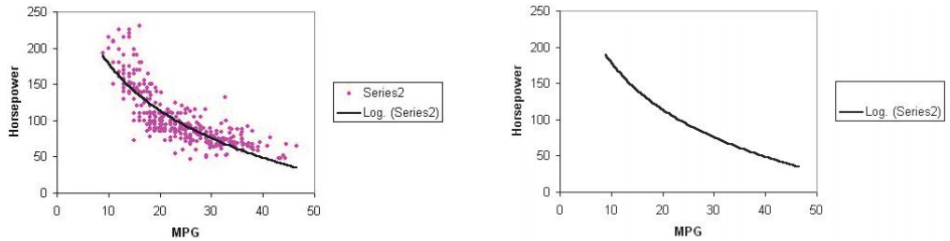
Một sai lầm khác khi trực quan hóa là việc xử lí dữ liệu thô và dùng dữ liệu suy diễn để thiết kế hình ảnh. Việc xử lí thường được thực hiện bằng các cách sau:

### 3.4.1 Áp đặt một mô hình khớp suy diễn

Một hình thức phổ biến trong hình ảnh trực quan là tính toán một mô hình phân tích của dữ liệu bằng cách sử dụng phép khớp đường cong hoặc bề mặt để đạt được kết quả có tính thẩm mỹ hơn. Tuy nhiên, đây là một hình thức bóp méo sự thật và có thể dẫn đến những giả định và kết luận sai lầm từ phía người quan sát.

Trong một số hình ảnh trực quan, thường loại bỏ tất cả dữ liệu thô và chỉ hiển thị kết quả xấp xỉ mượt mà từ dữ liệu đó. Điều này buộc người xem phải tin rằng phép xấp xỉ là sự phản ánh chính xác của dữ liệu, nhưng điều này không phải lúc nào cũng đúng khi nhà thiết kế áp dụng các thuật toán khớp số liệu một cách mù quáng.

**Giải pháp:** Hiển thị cả dữ liệu thô và mô hình đã khớp trước, đồng thời cho phép người xem giảm mức độ hiển thị hoặc lọc bỏ một trong hai tùy theo yêu cầu (xem hình 22).

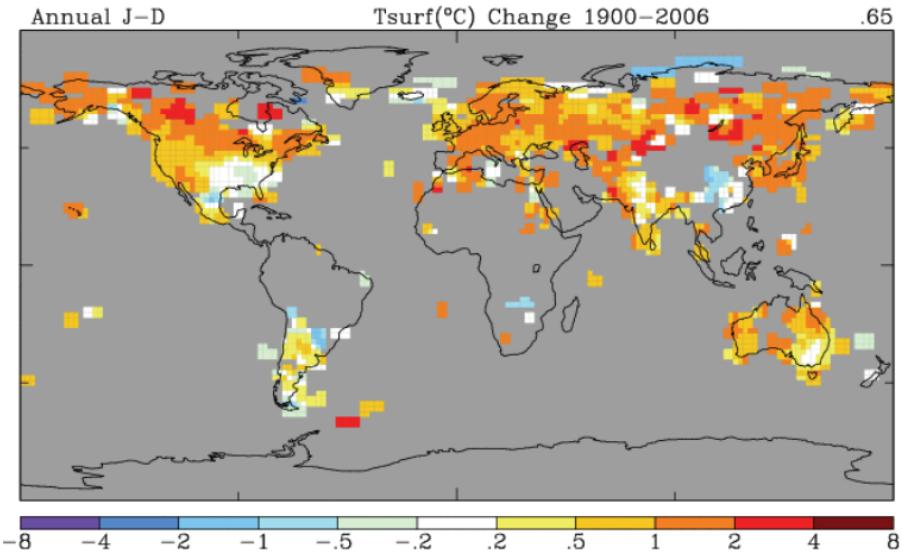



(a) (b)

Hình 22: (a)Dữ liệu với đường cong; (b)Chỉ đường cong

### 3.4.2 Áp đặt 1 tập được lấy lại mẫu

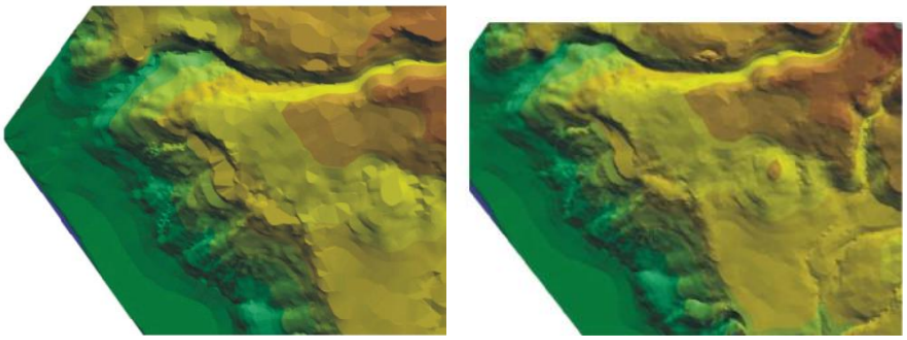
Một hình thức khác của việc làm sạch dữ liệu là quá trình lấy mẫu lại (resampling), trong đó dữ liệu thô, được định vị trên lưới thưa thớt hoặc ngẫu nhiên, được sử dụng để tạo ra dữ liệu dày hơn hoặc trên lưới có khoảng cách đều đặn. Điều này có thể mang lại một hình ảnh trực quan phong phú hơn, gần với mẫu liên tục, nhưng nó cũng đánh lừa người dùng khiến họ tin rằng tập dữ liệu lớn hơn nhiều so với thực tế. Mật độ lấy mẫu lại càng cao, người dùng càng dễ hiểu sai dữ liệu, trừ khi hiện tượng đang quan sát có ít biến động.




Hình 23: Dữ liệu thưa thớt về biến đổi nhiệt độ toàn cầu có thể tạo ra các giá trị sai lệch cho hầu hết các khu vực trên Trái Đất nếu được nội suy

Ví dụ, (Hình 23) cho thấy vị trí của các trạm theo dõi nhiệt độ toàn cầu. Rõ ràng, có những khoảng trống lớn nơi không có trạm nào tồn tại, vì vậy việc lấy mẫu lại có thể dẫn đến nhiều kết luận sai lầm, chẳng hạn như vùng phía bắc Nam Mỹ có thể bị nội suy từ dữ liệu của bốn hoặc năm trạm, dẫn đến kết luận rằng nhiệt độ của khu vực này đã giảm trong thế kỷ qua.

Lấy mẫu không đầy đủ cũng là một vấn đề khác. Như các hình ảnh trong (Hình 24) cho thấy, việc lấy mẫu mà không xem xét đặc điểm của dữ liệu có thể bỏ lỡ nhiều tính năng quan trọng. Hình bên trái là hình được lấy mẫu và nội suy một cách đồng đều, trong khi hình bên phải sử dụng thông tin đường viền để thêm các điểm mẫu tại những nơi có sự thay đổi đáng kể.



Hình 24: Sự khác biệt trong cách lấy mẫu và nội suy từ cùng một tập dữ liệu có thể tạo ra những hình ảnh rất khác nhau. Trong trường hợp này, một số chi tiết trong hình ảnh bên phải không được nhìn thấy trong hình ảnh bên trái.

Điều quan trọng là người dùng luôn có quyền truy cập vào dữ liệu thô và được thông báo về bất kỳ hoạt động làm sạch, làm mịn hoặc lấy mẫu lại nào đã được áp dụng. Trong một số lĩnh vực, chẳng hạn như chẩn đoán hình ảnh (radiology), các nhà phân tích kiên quyết phản đối bất kỳ dạng làm mịn hoặc lọc dữ liệu nào, vì có nguy cơ tín hiệu quan trọng trong dữ liệu có thể bị loại bỏ dưới dạng nhiễu.

**Giải pháp:** Cung cấp các chế độ hiển thị dữ liệu thô trước khi tạo ra các phiên bản mới, cho phép người dùng quyết định liệu sự dẫn xuất có phản ánh chính xác dữ liệu ban đầu hay không.

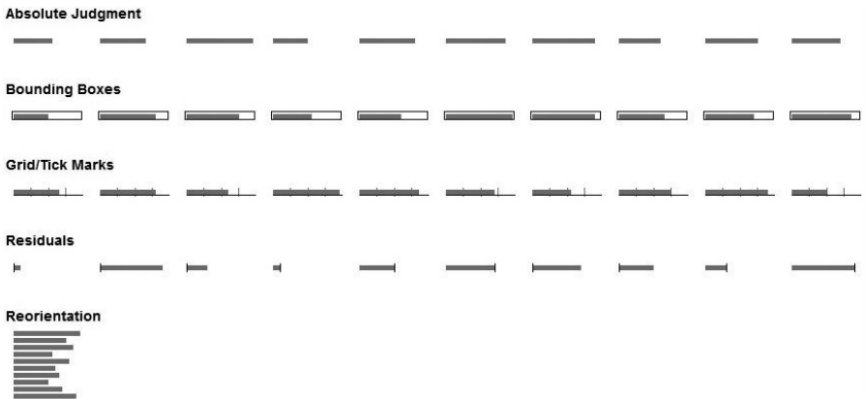
## 3.5 Phán đoán tuyệt đối so với Phán đoán tương đối

Phán đoán tuyệt đối là khi người dùng đưa ra một kết luận dựa trên một giá trị cụ thể, không bị ảnh hưởng bởi các yếu tố khác trong biểu đồ. Phán đoán tương đối là khi người dùng phải so sánh các giá trị giữa các phần tử trên biểu đồ để đưa ra nhận định.

Con người có khả năng tương đối hạn chế trong việc thực hiện các phán đoán tuyệt đối về kích thích thị giác. Điều này có nghĩa là những hình ảnh trực quan phụ thuộc quá nhiều vào việc người dùng thực hiện các phép đo chính xác các thuộc tính đồ họa như vị trí, chiều dài và màu sắc sẽ dẫn đến những vấn đề trong việc diễn giải. Một cách để khắc phục hạn chế này là thiết kế các hình ảnh trực quan dựa vào phán đoán tương đối hơn là tuyệt đối, hoặc chỉ sử dụng một số giá trị riêng biệt nhỏ cho mỗi thuộc tính đồ họa được sử dụng để truyền tải thông tin.

**Giải pháp:**

- Thiết kế các hình ảnh trực quan dựa vào phán đoán tương đối hơn là tuyệt đối, hoặc sử dụng thêm một vài số nhỏ riêng biệt cho mỗi đối tượng đồ họa để truyền tải thông tin.
- Các hộp giới hạn, lưới và dấu vạch đều là những công cụ tuyệt vời để giúp chuyển việc phán đoán tuyệt đối thành một nhiệm vụ phụ thuộc nhiều hơn vào phán đoán tương đối. Bằng cách so sánh chiều dài hoặc vị trí của một thực thể đồ họa cùng với một cấu trúc đã được định lượng, người xem có thể nhanh chóng xác định giá trị gần đúng so với các mức đã biết.
- Sử dụng các phần dư (ví dụ, trừ các giá trị từ giá trị trung bình của chúng) cũng có thể biến việc đo lường chính xác thành việc quyết định tương đối liệu một giá trị nằm trên hoặc dưới một mức cụ thể nào đó (xem hình 25).




Hình 25: Một số ví dụ về phán đoán tuyệt đối và tương đối

# 4 Tổng kết & Các ví dụ ứng dụng thực tế

Như vậy, trong chương này, một số quy tắc thiết kế cho việc tạo ra hình ảnh trực quan hiệu quả đã được trình bày, chúng bao gồm **quy trình đảm bảo phù hợp các yếu tố**:

- **Ánh xạ** dữ liệu
- **Chế độ** xem
- **Mật độ** thông tin
- Thông tin **chú thích**
- **Màu sắc**
- **Thẩm mỹ**

Ngoài ra, chúng tôi cung cấp thêm 1 số kỹ thuật để tránh các lỗi khác sau khi hoàn thành quy trình trên:

- **Tránh đánh lừa người xem** bằng các tỷ lệ không cân bằng và các chiêu trò hình ảnh khác
- Đảm bảo rằng **hình ảnh trực quan cần có ý nghĩa** ngữ nghĩa
- Sử dụng lưới một cách hợp lý để **tránh che khuất dữ liệu**. Quá nhiều "chart junk" có thể **gây mất tập trung sự chú ý của người xem**
- Cung cấp **quyền truy cập** vào **dữ liệu thô** cho người xem; thường thì việc làm sạch dữ liệu là chấp nhận được, nhưng người xem nên biết cách dữ liệu kết quả đã được tạo ra
- Thiết kế hình ảnh trực quan **ưu tiên phán đoán tương đối** hơn là phán đoán tuyệt đối

Sau đây chúng ta hãy cùng thử ứng dụng những điều đã được học này vào các bộ dữ liệu (bài toán) thực tế.

Chủ đề chính của nhóm liên quan đến việc phân tích các yếu tố kinh tế vĩ mô và vi mô để lựa chọn danh mục đầu tư phù hợp, bởi vậy em sẽ phân tích thành 2 bộ dữ liệu (bài toán) con để ứng dụng các kiến thức nêu trên:

- Bộ dữ liệu nhỏ: Gồm các thông tin về mã 1 cổ phiếu theo thời gian
- Bộ dữ liệu lớn: Gồm 32 chiều mô tả các tính chất kinh tế vĩ mô quan trọng

## 4.1 Bộ dữ liệu nhỏ

Trước khi bắt đầu quy trình 6 bước trực quan hóa đã đề cập thì ta cần trả lời câu hỏi muốn chỉ ra điều gì cho người xem. Ở đây, em mong muốn đưa ra các thông tin 1 mã cổ phiếu theo thời gian để người dùng quyết định có đầu tư hay không.

Như vậy, đã xác định được mục tiêu, giờ ta đến với bước đầu tiên - xác định ánh xạ phù hợp.

### Xác định ánh xạ

Do dữ liệu dạng chuỗi thời gian cổ phiếu nên ánh xạ nó theo biểu đồ nến Nhật (Candlestick chart): Phổ biến nhất trong phân tích kỹ thuật, thể hiện giá mở cửa, đóng cửa, cao nhất, thấp nhất dưới dạng các thanh hình nến, giúp phân tích hành vi của thị trường.

**Màu sắc** được sử dụng ở đây để phân biệt 1 cách trực quan các phiên giảm điểm hoặc tăng điểm. Trong khi đó **độ dài** 1 thanh cho biết biên độ dao động của giá cổ phiếu trong phiên đó.


Hình 26: Biểu đồ Candlestick

### Cung cấp chế độ xem phù hợp

Ở đây hình 27, em cung cấp cho người xem các khả năng điều chỉnh phù hợp với góc nhìn mà họ mong muốn bằng thao tác cuộn thu phóng:

Bên cạnh khả năng cuộn, thu phóng; 1 biểu đồ trực quan hóa cần cung cấp khả năng **thay đổi nhãn màu sắc** để phù hợp với tư duy nhìn nhận của người xem. Hình 28 là ví dụ khả năng cho phép người dùng thay đổi các nhãn màu sắc cho phù hợp góc nhìn của họ.

Để có thể thay đổi ánh xạ dữ liệu cho phù hợp mong muốn người dùng, trong các phân tích kỹ thuật mã cổ phiếu, họ hoàn toàn có thể chuyển sang ánh xạ log như tại hình . Đây là ánh xạ giúp tập trung hơn vào tỷ lệ thay đổi của các phiên đánh giá so với phiên trước đó. Qua đó giúp đánh giá nên ”trade” hay không.

### Bốn bước còn lại

Các bước tiếp theo là xác định **mật độ thông tin** phù hợp, thêm các **chú thích**, điều chỉnh **màu sắc** và **thẩm mỹ** đều tương đối dễ dàng do dữ liệu có ít chiều.

Kết quả thu được qua 6 bước có thể quan sát ở nhóm các hình 30a hoặc hình 30.

## 4.2 Bộ dữ liệu lớn

Nhìn chung, với 1 bộ dữ liệu nhỏ. Sau khi đã xác định được mục tiêu muốn trình bày cho người xem. Kết hợp cùng việc tuân thủ nghiêm ngặt 6 bước trong quy trình thì cơ bản thành công và ít khả năng gặp lỗi.

Tuy nhiên khi bộ dữ liệu có nhiều chiều, nhiều bản ghi; khi người làm trực quan mơ hồ về ý định truyền tải, kể cả khi tuân thủ 6 bước quy trình, sai sót vẫn có thể đến.

Sau đây em sẽ trình bày cách khắc phục 1 số lỗi kinh điển đã đề cập trong báo cáo này.


(a) Cho phép người dùng xem toàn bộ lịch sử giá cổ phiểu


(b) Cho phép người dùng cuộn thu phóng

Hình 27: Khả năng cuộn, thu phóng

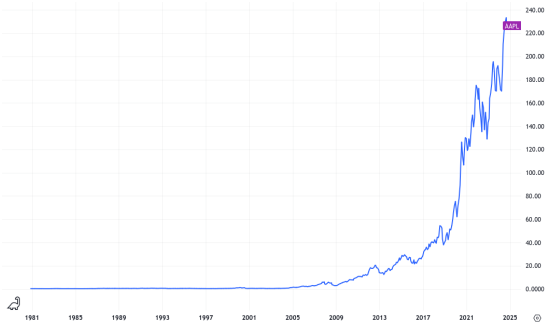

(a) Màu xanh:tím


(b) Màu xanh:đỏ

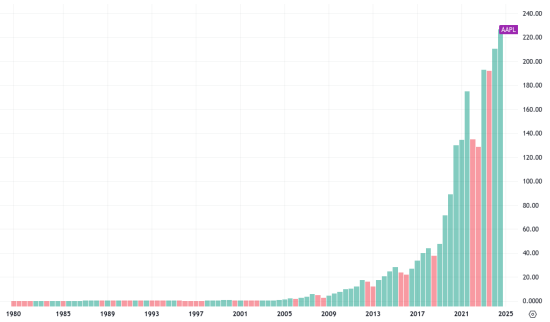
Hình 28: Khả năng thay đổi các nhãn màu


Hình 29: Ánh xạ log làm nổi bật tỷ lệ thay đổi qua các phiên

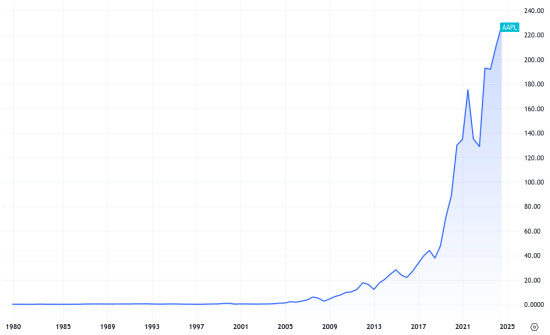
(a) Kết quả trực quan 1



(b) Kết quả trực quan 2



(c) Kết quả trực quan 3



(d) Kết quả trực quan 4

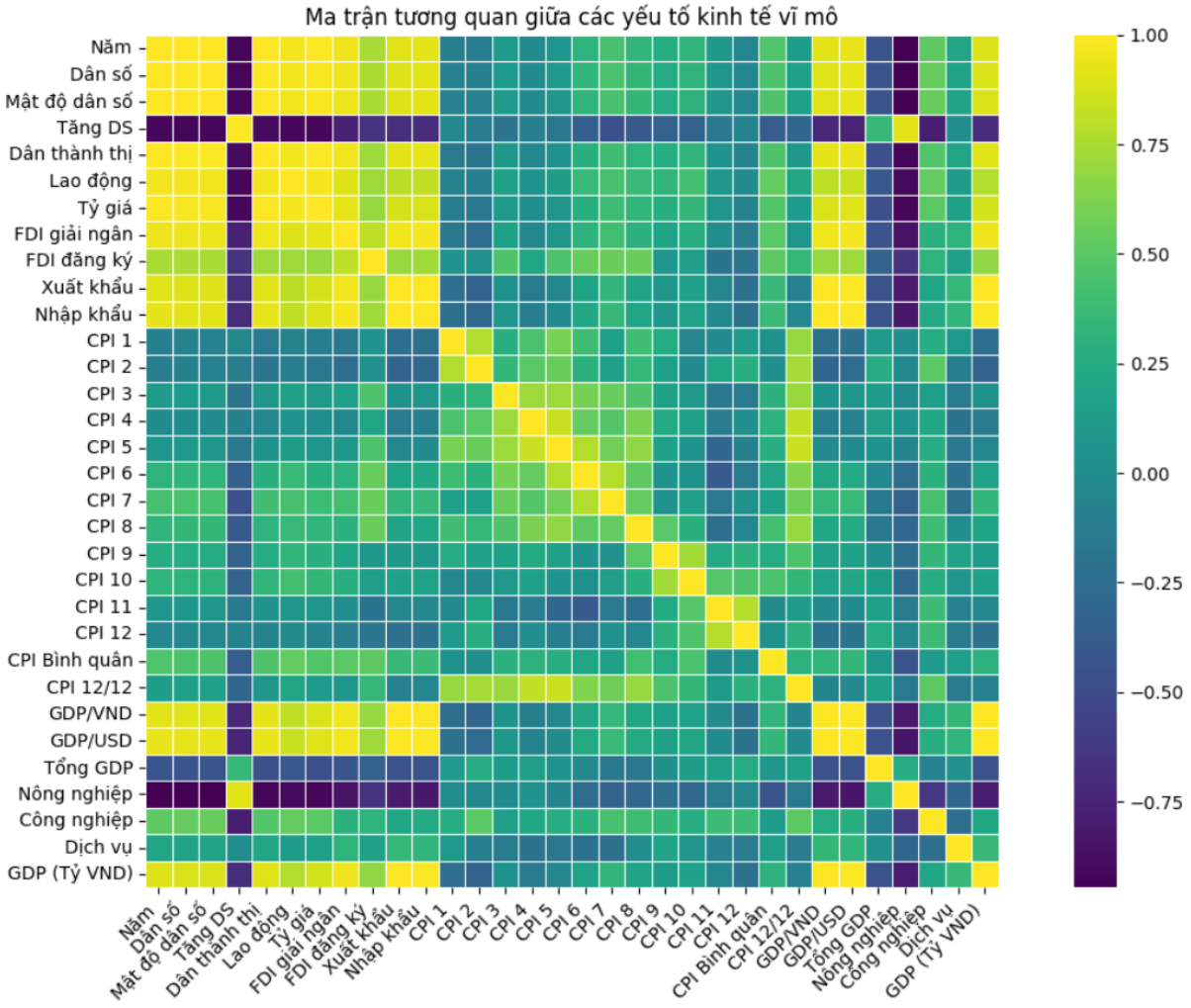
Hình 30: Bốn kết quả trực quan khác nhau

### Xác định mục tiêu truyền tải

Trước khi thực hiện 6 bước ta vẫn cần xác định rõ mục tiêu truyền tải qua trực quan hóa. Ở đây mục tiêu chính là đánh giá các yếu tố tương quan của nền kinh tế vĩ mô để người dùng nhìn ra quy luật và ra quyết định phù hợp.

Hầu hết mọi người làm trực quan hóa bỏ qua bước này và dễ tạo ra các sản phẩm trực quan vô nghĩa thiếu trọng tâm.

Hình 31 minh họa đánh giá sơ bộ giúp người làm trực quan biết mình cần lựa chọn các thông tin gì để chỉ ra cho người xem.

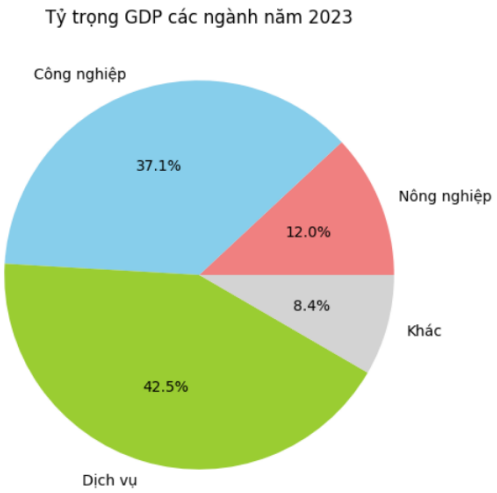



Hình 31: Tương quan các yếu tố kinh tế vĩ mô

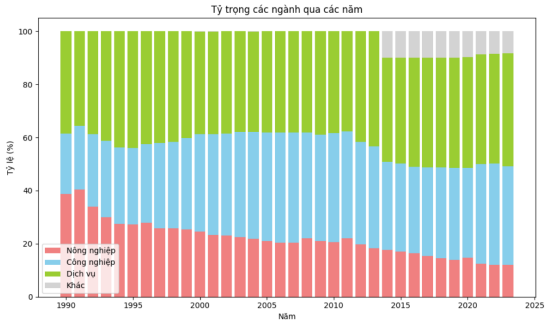
### Tránh tạo các trực quan vô nghĩa

Điểm mấu chốt chính là xác định được mục tiêu truyền tải. Với mỗi mục tiêu con phân ra ta có thể xác định được biểu diễn trực quan phù hợp, tránh các trực quan vô nghĩa.

Lấy ví dụ như nếu mục tiêu chỉ ra cho người xem quan điểm: ”Trong hầu hết thời gian lịch sử, Việt Nam là 1 quốc gia nghèo”. Ta hoàn toàn có thể dùng trực quan hình 32.

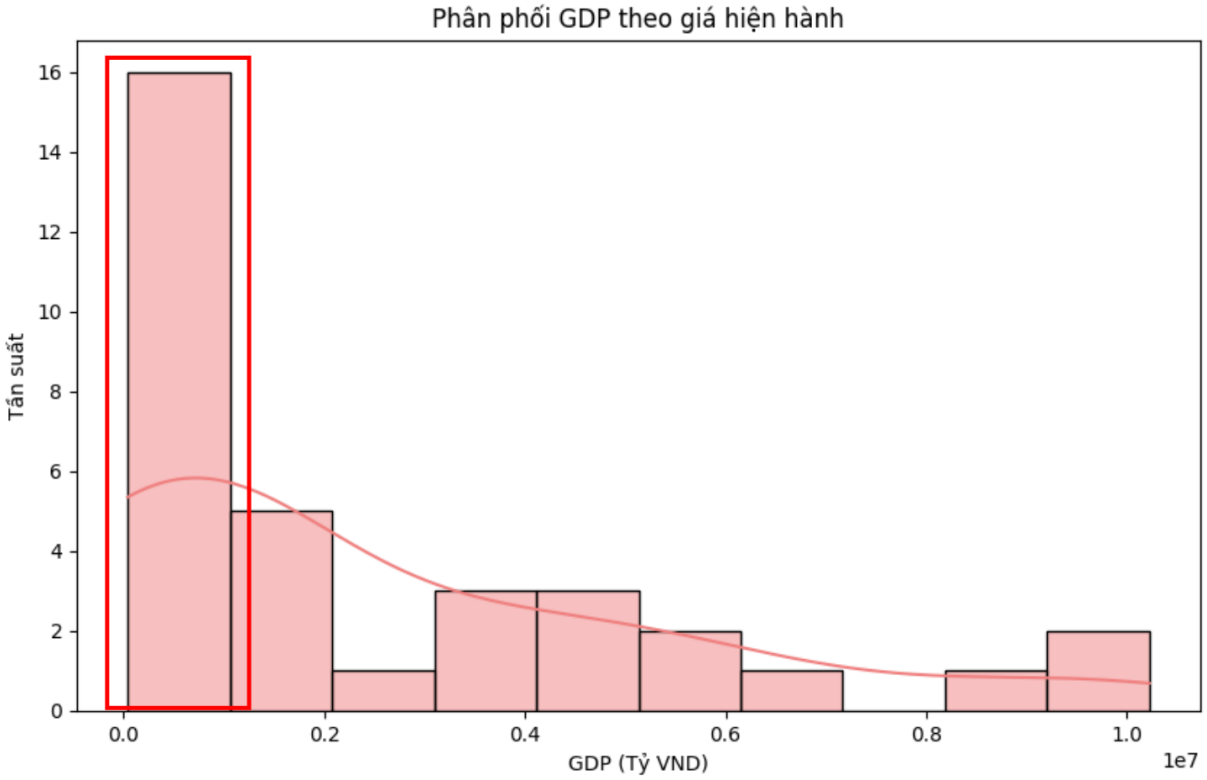



(a) Biểu đồ không hiển thị sự thay đổi tỷ trọng theo thời gian




(b) Biểu đồ hiển thị sự thay đổi tỷ trọng theo thời gian

Hình 33: So sánh hai cách trình bày dữ liệu. Việc không hiển thị sự thay đổi tỷ trọng theo thời gian (như trong Hình 33a) có thể gây hiểu nhầm và đánh lừa người xem. Chỉ có khi biểu đồ thể hiện đầy đủ thông tin về sự biến động theo thời gian (như trong Hình 33b), người xem mới có cái nhìn toàn diện và chính xác về dữ liệu.



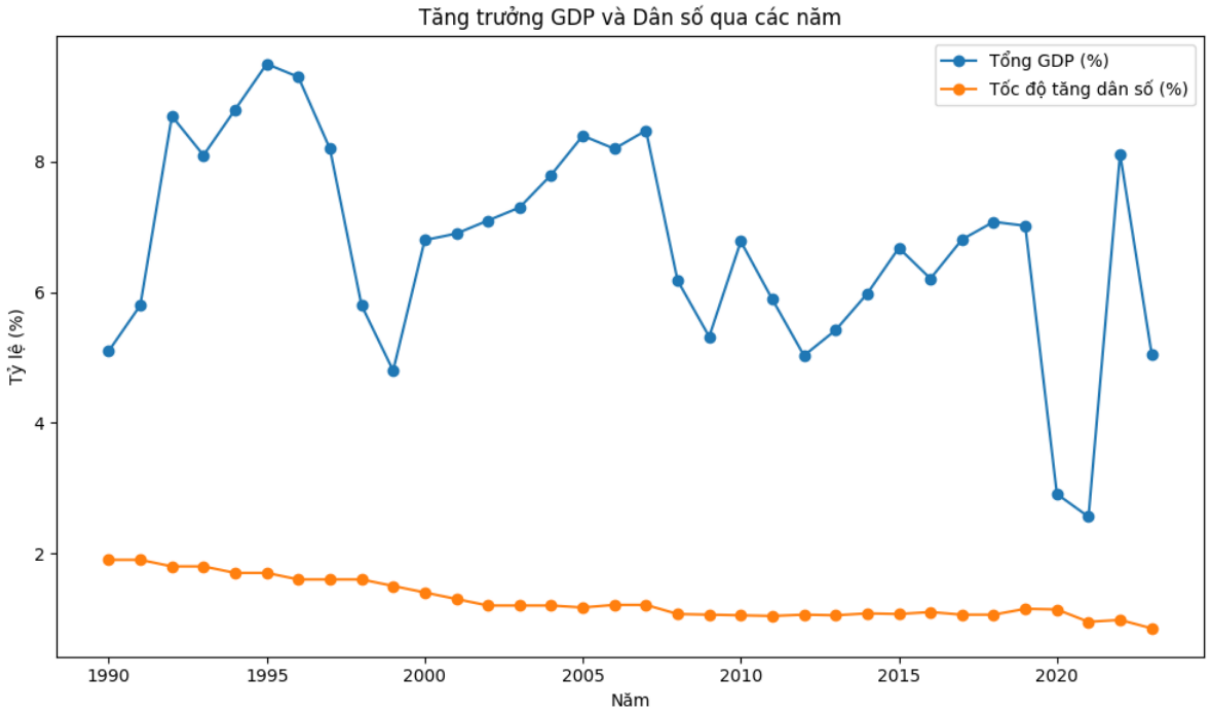

Hình 32: Phân phối giá trị GDP trong lịch sử thống kê

### Hình ảnh bị lợi dụng để làm người xem hiểu sai

Nếu muốn thuyết phục người xem về tỷ trọng của các nhóm ngành kinh tế (giả sử cho 1 quyết định đầu tư) mà không cho họ hiểu về xu hướng dịch chuyển tỷ trọng của các nhóm ngành theo thời gian thì chính là lợi dụng hình ảnh để làm người dùng hiểu sai.

### Ảo tưởng sự tương quan

Hình 34 chỉ ra sự giảm dần của tốc độ gia tăng dân số và sự giảm dần tốc độ tăng trưởng GDP. Đây là yếu tố phức tạp và cần nhiều nghiên cứu thống kê xác nhận trước khi kết luận cuối cùng.

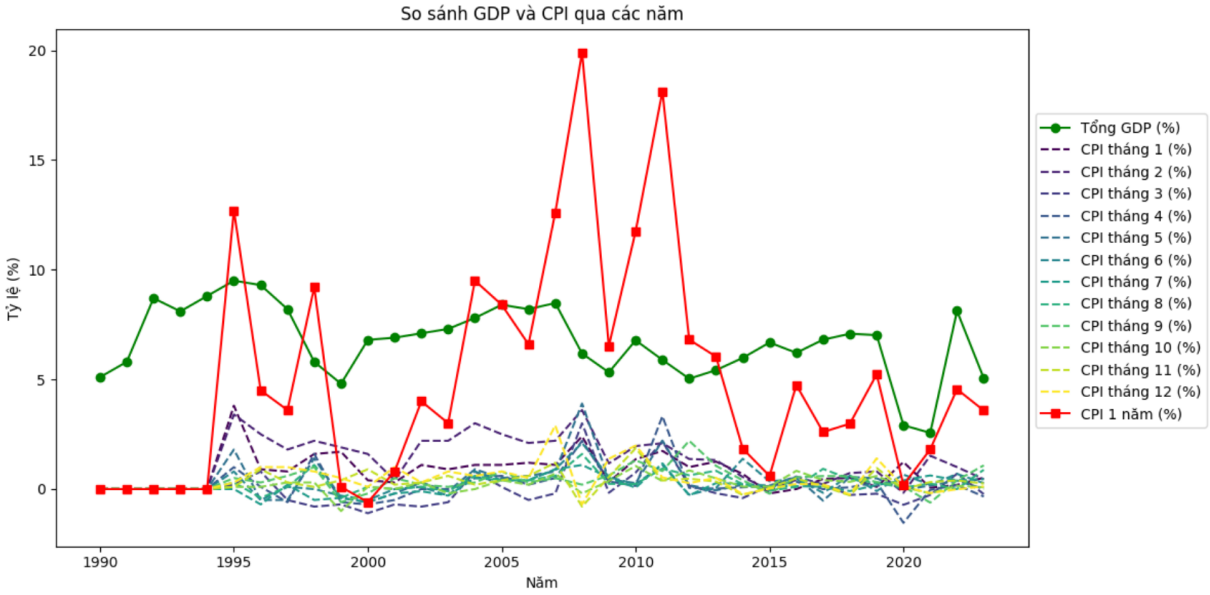



Hình 34: Cần xác thực yếu tố tương quan

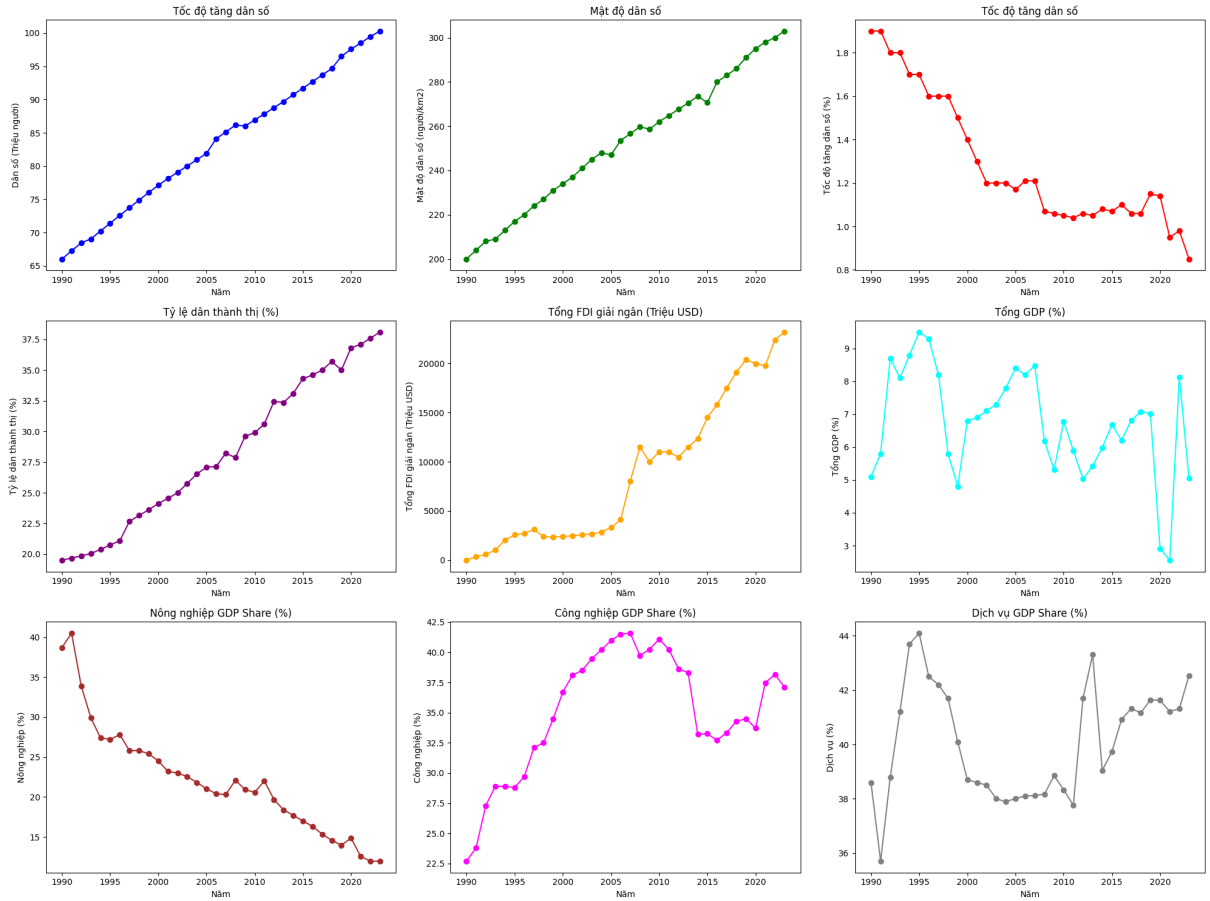
### Điều chỉnh lượng thông tin phù hợp cho dữ liệu lớn

Với những bộ dữ liệu phức tạp, các yếu tố cốt lõi cần được tô đậm như 1 cách để người xem tập trung như hình 35.

Hoặc phân tách ra thành nhiều góc độ nhìn như hình 36 với mỗi cửa sổ sẽ tập trung truyền tải 1 thông điệp cụ thể thay vì gộp chung vào 1 đồ thị duy nhất.




Hình 35: So sánh GDP và CPI qua các năm




Hình 36: Nhiều cửa sổ xem

Và đó là những gì để giúp tạo ra phiên bản trực quan phù hợp.

# 5 Tài liệu tham khảo

[1] https://www.researchgate.net/figure/Global-distribution-of-annual-mean-temperature-T-C
[2] https://www.instructables.com/Orrery-a-Mechanical-Solar-System-Model-From-Plywoo/

[3] https://plotly.com/python/time-series/