# Tuần 4 Phân tích và trực quan hóa dữ liệu thăm dò

## 4a Đồ thị so sánh

Trong mô hình này, ta sẽ thảo luận về phân tích và trực quan hóa dữ liệu thăm dò. Phân tích dữ liệu thăm dò là hiển thị dữ liệu để có thể khám phá các khuynh hướng mà chúng ta không thể tự tìm ra. Trực quan là các kỹ thuật mà chúng ta sử dụng để xem dữ liệu hoặc thông tin số.

Phân tích dữ liệu thăm dò gồm 3 vấn đề: cách lập biểu đồ dữ liệu, cách hiển thị thông tin số một cách rõ ràng. Rất nhiều kỹ thuật đã được phát triển cho phép chúng ta hiển thị nhiều thông tin trong một không gian nhỏ. Nếu chúng ta có thể hiển thị thông tin một cách rõ ràng và ngắn gọn trong một không gian nhỏ, chúng ta có thể tối đa số lượng biểu đồ cũng như giải thích. Điều này thể cho thấy thay đổi mà ta sẽ không thể nhìn ra nếu chỉ đơn thuần sử dụng một bảng số hoặc một biểu đồ rất lớn mà ít thông tin.

Hiển thị dày đặc thông tin cho phép xem và so sánh nhiều hơn. Vấn đề đầu tiên của phân tích dữ liệu thăm dò là hiển thị dữ liệu ở độ phân giải cao một cách hiệu quả, thay thế biểu đồ tròn hay cột bằng biểu đồ có thể cho phép hiển thị nhiều thông tin hơn trong một không gian nhỏ hơn.

Vấn đề thứ hai của phân tích dữ liệu thăm dò là giải thích kết quả đạt được. Mục tiêu của việc phân tích dữ liệu là để khám phá những thứ trong dữ liệu mà ta không thể thấy. Vì vậy chúng ta phải tự rèn luyện đôi mắt và bộ não để có thể đọc biểu đồ và rút ra thông tin. Sau đó, quay lại dữ liệu và kiểm tra các điểm dữ liệu riêng lẻ hoặc các xu hướng để hiểu tại sao những gì ta thấy xảy ra và tại sao nó không được tính đến.

Vấn đề cuối cùng của phân tích dữ liệu thăm dò là mô hình hóa. Các mô hình thống kê, chẳng hạn như hồi quy, có thể được sử dụng để tóm tắt những gì chúng ta biết về dữ liệu của mình. Mô hình càng tốt, ta càng nắm bắt được nhiều hơn khi thông qua phân tích dữ liệu thăm dò. Mọi hiển thị dữ liệu là một phép so sánh, rõ ràng hoặc ẩn ý.

Tóm lại, đồ thị hoặc phép so sánh, dù đang lập biểu đồ của riêng mình hay đang đọc biểu đồ do người khác tạo ra, ta nên biết những phép so sánh nào đã được sử dụng bởi công cụ vẽ biểu đồ và những phép so sánh nào khó thấy hơn.

## 4b Sử dụng dữ liệu để trả lời câu hỏi

Ta sẽ nói về cách chúng ta sử dụng biểu đồ để học từ dữ liệu, cách đặt câu hỏi và hình thành các giả thuyết. Dữ liệu không chỉ để sao lưu các quyết định đã đưa ra mà còn để tìm hiểu và khám phá. Ta sẽ có thể tự vẽ biểu đồ dữ liệu và giải thích các biểu đồ do người khác thực hiện.

Ta phải kiểm tra dữ liệu của mình, xem xét dữ liệu của và hiểu các mô hình đang được sử dụng. Ta sẽ dùng nhiều hơn một biểu đồ, đi sâu hơn trong việc kiểm tra dữ liệu theo các cách khác nhau. Trong bất kỳ loại tóm tắt dữ liệu nào, luôn có một nhóm kiểm soát rõ ràng hoặc ngầm định.

Thế giới thực rất phức tạp nhưng chúng ta có thể học hỏi bằng cách nghiên cứu kỹ lưỡng dữ liệu. Chúng ta có thể xem xét tất cả các khuynh hướng trong dữ liệu. Dữ liệu là không có giới hạn. Ta có thể tải dữ liệu, xem các con số và lấy ra các biến khác nhau. Ta có thể vẽ các kết quả khác nhau dưới dạng hàm của ngày, giờ, trạng thái và các biến khác.

## 4d Quy trình ra quyết định trực quan hóa dữ liệu 1

Ta sẽ nói về quy trình đưa ra quyết định cho việc trực quan hóa dữ liệu, cách chọn biểu đồ phù hợp nhất cho dữ liệu muốn trực quan hóa.

Chúng ta có một bảng ở đây về hoạt động hàng ngày của nam giới, từ 15 tuổi trở lên, đến từ sáu quốc gia khác nhau. Trong mỗi cột là số giờ mỗi ngày mà mỗi quốc gia tham gia. Chúng ta có thể so sánh dữ liệu của Mỹ với các quốc gia khác và tìm ra các khuynh hướng trong dữ liệu này.

Nhưng chỉ nhìn vào dữ liệu này, thật khó để tìm ra các mẫu. Vì vậy, chúng ta cần trực quan dữ liệu này.

Một cách để điều tra lượng thời gian của một quốc gia và những người đàn ông ở quốc gia đó sẽ là sử dụng biểu đồ hình tròn. Đây là một cách tốt để hình dung dữ liệu này, nhưng nó có thể dễ dàng hơn bằng mắt nếu chúng ta có thể so sánh nhanh hơn tỷ lệ trong mỗi lần.

Chúng ta có thể muốn sử dụng biểu đồ cột. Đây là một cách đơn giản để minh họa lượng thời gian dành cho mỗi hoạt động của một quốc gia nơi chúng ta có thể quan sát nhanh chóng so sánh giữa các hoạt động hành vi khác nhau.

Chúng ta phải làm gì nếu muốn so sánh hai quốc gia? Chúng ta có thể đặt hai biểu đồ hình tròn cạnh nhau. Nhưng khi chúng ta có hai biểu đồ hình tròn này cạnh nhau, thực sự khá khó để rút ra so sánh nhanh giữa tất cả các phần khác nhau. Vì vậy, một cách thích hợp hơn để thể hiện những dữ liệu này là sử dụng biểu đồ độ dốc.

Biểu đồ độ dốc là một cách rất hiệu quả để so sánh hai quốc gia, tỷ lệ một cách nhanh chóng và dễ dàng.

Điều gì sẽ xảy ra nếu chúng ta muốn so sánh nhiều quốc gia hơn? Đây là những biểu đồ hình tròn mà chúng ta đã thấy trước đây. Thoạt nhìn, rất khó để có thể hiểu được quốc gia nào có tỷ lệ cá nhân dành nhiều thời gian hơn, chẳng hạn như các hoạt động ngủ, hoạt động giải trí hoặc các hoạt động làm việc được trả lương. Nó không rõ ràng ngay lập tức chỉ từ sáu biểu đồ hình tròn này. Một điều mà chúng ta cũng có thể nghĩ, tại sao chúng ta không vẽ một biểu đồ thanh tương tự như cách chúng ta vẽ dữ liệu khi chúng ta chỉ có một biểu đồ hình tròn?

Đây là những gì sẽ xảy ra nếu chúng ta vẽ biểu đồ cột. Chúng ta có sáu quốc gia ở phía dưới. Mỗi cột, màu sắc thể hiện hành vi của sở thích. Vì vậy, thanh nhạt đầu tiên là công việc được trả lương, công việc nhà, chăm sóc con cái, ăn ngủ, thời gian giải trí, và những thứ khác. Nhưng một lần nữa, khá khó để thực sự nhanh chóng nhận ra sự khác biệt riêng lẻ.

Tuy nhiên, có một cách khác mà chúng ta có thể vẽ các dữ liệu này. Đây được gọi là biểu đồ khác biệt. Thanh giữa này, đối với mỗi quốc gia, là lượng thời gian trung bình dành cho hành vi đó. Vì vậy, nếu một thanh nằm ngay trên dòng đó, điều đó có nghĩa là quốc gia đó dành lượng thời gian trung bình cho hành vi đó. Chúng ta có thể làm điều này cho từng nhóm khác nhau. Đây là một cách nhanh chóng và hiệu quả hơn nhiều để trực quan hóa bảng dữ liệu để nhanh chóng thu hút sự chú ý của chúng ta đến những gì giống hoặc khác nhau giữa các nhóm khác nhau.

Ta cần chọn biểu đồ dữ liệu hiệu quả nhất và truyền đạt câu chuyện với người đọc một cách dễ dàng và trực quan nhất để nó không mất quá nhiều thời gian trước khi họ có thể thấy điều mà chúng ta đang cố gắng chỉ ra với dữ liệu của mình.

Chúng ta có thể làm cho điều này hiệu quả hơn nữa. Đây là biểu đồ xem xét sự chênh lệch thu nhập theo giới 10 năm sau khi nhập học đại học. Chúng ta có những thanh màu đỏ đậm hơn này là nam giới và chúng ta có các thanh màu nhạt hơn là nữ giới. Trục y là mức lương hàng năm, sau 10 năm đăng ký. Nhưng cũng có sự chênh lệch lớn về giới tính. Biểu đồ cho thấy dữ liệu đó, và chúng ta có thể xem một số thông báo khá nhanh.

Nhưng cũng có một số hạn chế. Có quá nhiều thanh trên trục x. Vì vậy, một cách thay thế để vẽ những dữ liệu này có thể là thực hiện cái được gọi là biểu đồ chấm hoặc biểu đồ hình quả tạ. Ở đây, chúng ta đang dọc theo trục y, lần lượt hiển thị từng trường đại học. Nó dễ đọc hơn nhiều, vì văn bản nằm nghiêng và chúng ta không phải quay đầu lại để đọc. Mức lương hàng năm hiện nằm trên trục x. Mỗi dấu chấm là mức lương trung bình của nam giới và phụ nữ, và kết nối họ với đường thẳng này, thể hiện sự chênh lệch giới tính của họ. Những dữ liệu này được vẽ theo thứ tự mức lương trung bình của nam giới, giảm dần. Đây là một hình dung rõ ràng hơn nhiều. Sẽ dễ dàng hơn nhiều để nhanh chóng biết được trường đại học nào có sự chênh lệch giới tính lớn nhất. Ngoài ra, chúng ta có thể so sánh dễ dàng hơn nhiều. Đây là một cách rất nhanh chóng và dễ dàng để vẽ cùng một dữ liệu, sử dụng ít mực hơn trên trang và thu hút người đọc vào câu chuyện nhanh hơn nhiều.

Đây là một ví dụ khác về cách sử dụng các kỹ thuật tối thiểu để loại bỏ thông tin vô nghĩa và chỉ giữ lại thông tin có ý nghĩa trên biểu đồ. Đây là một ví dụ từ The Economist. Những gì họ đang vẽ trong biểu đồ dọc theo trục x này là các khu vực khác nhau trên thế giới. Trục y là tỷ lệ tham gia lực lượng lao động. Có một chút khó khăn để nói, vì nó nằm ở phía bên tay phải, điều này không bình thường. Và ngoài ra, tiêu đề ở trên cùng. Nó không theo trục. Sau đó, khi chúng ta nhìn vào biểu đồ chính, phải mất một chút thời gian trước khi chúng ta thực sự có thể định hướng những gì đang được hiển thị. Chúng ta có thể thấy rằng đây là nam giới. Và dưới đây được dán nhãn phụ nữ. Vì vậy, bây giờ chúng ta nhận ra rằng các đường màu đỏ đại diện cho dữ liệu cho phụ nữ và các đường màu xanh, dữ liệu cho nam giới. Sau đó, chúng ta nhận ra rằng họ đã chú thích một số ngày ở đây. Đây là năm 1990 và đây là năm 2013. Vì vậy, bây giờ chúng ta có thể nhận ra rằng điều này thực sự thể hiện tỷ lệ tham gia lực lượng lao động từ năm 1990 đến năm 2013 ở các khu vực khác nhau đối với phụ nữ và nam giới. Có tất cả bóng mờ và màu sắc bổ sung này, và không dễ dàng để nhanh chóng nhận ra câu chuyện mà họ đang cố gắng kể cho chúng ta.

Vì vậy, ta đã vẽ lại dữ liệu này bằng một biểu đồ chấm khác. Ở đây, trên trục y, ta có khu vực, 1990 hoặc 2013, các nước phát triển, Trung Đông, Đông Á, Mỹ Latinh. Dọc theo trục x, ta có tỷ lệ tham gia lực lượng lao động từ 0% đến 100% và ta đã vẽ biểu đồ với người lớn về tỷ lệ tham gia lực lượng lao động của nam giới và tỷ lệ tham gia lực lượng lao động của phụ nữ. Ở đây, dễ dàng hơn nhiều để nhanh chóng nhận thấy sự thay đổi từ năm 1990 đến năm 2013, hoặc thực tế, sự khác biệt giữa tỷ lệ tham gia của phụ nữ và nam giới giữa hai khoảng thời gian này giống nhau như thế nào.

Thật khó để thấy điều đó trên biểu đồ trước, nơi chúng ta phải nhìn và điều hướng để xem năm 1990 là gì và năm 2013 là gì. Một điều khác mà chúng ta có thể nhanh chóng nhận thấy ở đây là có sự khác biệt giữa các vùng khác nhau. Nó rõ ràng hơn nhiều trong định hướng của biểu đồ này.

Chúng ta có thể thấy rằng sự chênh lệch lớn nhất về sự tham gia của lực lượng lao động là ở Trung Đông và Bắc Phi, và tương tự, ở Mỹ Latinh so với Đông Á và Thái Bình Dương. Vì vậy, điều ta sẽ nhấn mạnh với những ví dụ này là chúng ta muốn loại bỏ thông tin vô nghĩa, nhưng giữ lại thông tin có ý nghĩa.

Hãy thử và giảm lượng mực mà chúng ta sử dụng trên biểu đồ, làm cho mọi người có thể đọc được và minh họa dữ liệu của chúng ta để người đọc có thể dễ dàng truy cập.

## 4e Quy trình ra quyết định về trực quan hóa dữ liệu 2

Một số kỹ thuật cho phép minh họa một cách hiệu quả các điểm trong dữ liệu sẽ thu hút sự chú ý của độc giả một cách nhanh chóng để họ có thể hiểu ngay thông điệp mà chúng ta đang cố gắng truyền tải.

Điều đầu tiên ta là làm nổi bật và chú thích các biểu đồ. Ở đây chúng ta có một đồ thị. Nó ghi trong tiêu đề Bàn thắng của Câu lạc bộ bóng đá Liverpool từ năm 1893 đến năm 2015. Nó cho biết mỗi dòng đại diện cho số bàn thắng tích lũy ghi được theo số trận trong một mùa giải. Vì vậy, những gì chúng ta có ở đây là tổng số bàn thắng mà Câu lạc bộ bóng đá Liverpool, một đội bóng ở Anh, ghi được so với số trận đấu trong một mùa giải. Đôi khi số trận sẽ là 40 mỗi mùa, đôi khi là 38. Một mùa giải ở đây chỉ có 30 trận trong mùa giải. Và đây là tổng số bàn thắng. Mỗi dòng đại diện cho một mùa khác nhau.

Vì vậy, chúng ta có thể thấy từ biểu đồ này rằng có một số mùa giải họ không ghi được nhiều bàn thắng, có một số mùa giải họ ghi được rất nhiều bàn thắng và có nhiều mùa giải họ ghi được số bàn thắng trung bình.

Ta đánh dấu hai dòng màu đỏ tươi là năm 2013, dòng màu đỏ đậm là năm 2014. Rõ ràng ngay lập tức rằng hai mùa giải này, chỉ cách nhau một năm, có sự khác biệt rất lớn về số lượng bàn thắng của Câu lạc bộ bóng đá Liverpool trong hai thời điểm này. Hơn nữa, chúng ta có thể thấy với một đường màu đỏ này bằng cách đánh dấu nó và cũng chú thích đường khác này, đường cao nhất, rằng năm nay là một mùa giải ghi bàn thắng cao trong lịch sử cho Câu lạc bộ bóng đá Liverpool.

Vì vậy, chúng ta có thể trình bày thông tin này, ví dụ, nếu chúng ta cố gắng giải thích tại sao họ có tỷ lệ ghi bàn cao trong lịch sử ở đây so với mùa giải này. Nó thực sự minh họa sự khác biệt giữa hai mùa. Câu chuyện thực tế ở đây là vào năm 2013, cầu thủ xuất sắc nhất của họ, người ghi bàn thắng của họ, Luis Suarez, đã bị treo giò 9 trận đầu tiên của mùa giải. Chúng ta có thể thấy điều này ở đây, tỷ lệ ghi bàn của họ rất, rất giống nhau trong chín trận đầu tiên. Chỉ khi anh ấy trở lại vào năm 2013, tỷ lệ ghi bàn mới tăng cao, trong khi vào năm 2014 sau khi anh ấy rời câu lạc bộ để đến Barcelona, ​​tỷ lệ ghi bàn của họ thấp hơn nhiều.

Vì vậy, chỉ đơn giản bằng cách đánh dấu hai dòng, chú thích các biểu đồ, chúng ta có thể thấy rằng có một câu chuyện đằng sau dữ liệu mà nếu không chúng ta sẽ không thể thấy được. Một cách khác để truyền tải rất nhiều thông tin rất nhanh chóng và có thể tiếp cận được với người đọc, người xem, là sử dụng các bội số nhỏ.

Bội số nhỏ là một biểu đồ mà chúng ta có các bảng điều khiển, nơi chúng ta có cùng một biểu đồ được lặp lại nhiều lần. Đây là một ví dụ mà chúng ta có cùng một biểu đồ được lặp lại ba lần. Ở đây, bạn đang xem tỷ lệ thất nghiệp từ năm 2005 đến năm 2014. Và bạn đang xem xét tỷ lệ đó trên bốn nhóm khác nhau là chủng tộc và sắc tộc - người châu Á, người da đen hoặc người Mỹ gốc Phi, người gốc Tây Ban Nha hoặc người Latinh và người da trắng.

Điều hiển nhiên trong bảng điều khiển này, mà khi chúng ta nhìn vào phần trên cùng, chúng ta có thể thấy những cá nhân đã có bằng đại học, là những người Mỹ gốc Phi hoặc những cá nhân báo cáo là người da đen hoặc người Mỹ gốc Phi, có tỷ lệ thất nghiệp trung bình cao hơn so với ba sắc tộc khác.

Khi chuyển sang bảng tiếp theo, chúng ta có thể thấy rằng tỷ lệ thất nghiệp ở bảng này cao hơn bảng trước. Và đây là những cá nhân chỉ nhận bằng tốt nghiệp trung học phổ thông. Và trong bảng thứ ba, đây là những cá nhân chưa học hết cấp ba.

Vì vậy, có một số điều mà chúng ta có thể thấy trong biểu đồ này. Trước hết, mọi bảng điều khiển đều có cùng một mô hình là nó đi xuống và lên xuống. Và không có gì ngạc nhiên khi điều này liên quan đến suy thoái kinh tế. Khi đỉnh tăng lên trong mỗi bảng, đây là khi cuộc suy thoái xảy ra vào năm 2008, 2009. Và những gì chúng ta có thể thấy ngay lập tức trên cả ba nhóm là điều này rất rõ ràng, nhưng sự phục hồi đã chậm hơn nhiều. Đó là một sự suy giảm chậm hơn. Những gì chúng ta cũng có thể thấy là bất kể trình độ học vấn - ít hơn bằng cấp ba, bằng tốt nghiệp trung học hoặc bằng đại học - những người da đen hoặc người Mỹ gốc Phi có tỷ lệ thất nghiệp trung bình cao hơn nhiều so với ba chủng tộc và sắc tộc còn lại.

Vì vậy, bội số nhỏ này là một cách rất, rất hiệu quả để nhận được nhiều thông tin qua một cách rất nhanh chóng. Chúng ta có thể thấy ảnh hưởng của trình độ học vấn, nhưng bằng cách sử dụng các đường khác nhau, chúng ta có thể thấy ảnh hưởng của chủng tộc và dân tộc và tác động của điều đó đến tỷ lệ thất nghiệp.

Chúng ta cũng có thể thấy thang đo thời gian này kéo dài từ năm 2005 và 2000 đến năm 2014. Điều ta muốn nhấn mạnh về biểu đồ này là chúng ta đã sử dụng các đường lưới rất, rất tối thiểu. Điều này cho phép người đọc tập trung vào chuỗi thời gian và không tập trung vào các chi tiết bổ sung của biểu đồ không quan trọng để hiểu những gì chúng ta đang cố gắng nói với người xem.

Những gì chúng ta có thể làm với biểu đồ này là sử dụng một trục nhỏ với các dấu tích. Điều đó có thể giúp chúng ta nhìn thấy những năm khác nhau. Nhưng thực sự những gì ta đang cố gắng thể hiện với biểu đồ này là mô hình này, đường cong này. Và điều đó khá rõ ràng khi chỉ sử dụng một trục rất đơn giản.

Đây là một ví dụ khác về bội số nhỏ. Những gì chúng ta có ở đây là chín quốc gia khác nhau. Và chúng ta thấy cùng một dữ liệu chín lần khác nhau, nhưng đối với mỗi quốc gia, chúng ta sử dụng một đường màu xanh lam dày hơn nhiều để làm nổi bật thông tin của quốc gia đó. Những gì chúng ta đang thực sự xem xét, nếu chúng ta chỉ tập trung vào một biểu đồ là tỷ lệ dân số theo từng nhóm tuổi. Dọc theo trục x, chúng ta có 0, 1, 2, 3, tất cả các cách lên đến 100. Và mỗi khoảng duy nhất trên trục x là tỷ lệ dân số từ 0 đến 1, 1 đến 2, v.v.

Bằng cách nhìn vào dữ liệu của Ý so với tất cả các quốc gia khác, được hiển thị bằng màu xám trên mọi biểu đồ, chúng ta có thể thấy rằng Ý có số lượng cá nhân rất trẻ trong khoảng 0 đến 25 rất thấp. Ngược lại, nếu chúng ta nhìn vào phần đuôi bên kia của biểu đồ này, chúng ta có thể thấy rằng Ý có một tỷ lệ rất lớn những người từ 75 đến 100 tuổi. Nếu chúng ta đi đến góc dưới cùng bên phải, chúng ta có thể thấy rằng Brazil có mô hình ngược lại. So với tất cả các quốc gia khác, Brazil có dân số đặc biệt trẻ, nói một cách tỷ lệ. Họ cao hơn nhiều so với tất cả những người khác, trong khi họ có nhiều, ít hơn nhiều cá thể trong phạm vi 75 đến 100. Vì vậy họ có sự phân bố dân số rất trẻ. Bằng cách sử dụng bội số nhỏ này, chúng ta có thể thấy ngay mô hình ở các quốc gia khác nhau, quốc gia nào giống nhau, quốc gia nào khác nhau?

Ví dụ, chúng ta có thể thấy rằng Đức có một mô hình rất, rất giống với Pháp. Chúng ta cũng có thể phác thảo ở đây rằng có một số khác biệt rất thú vị trong đường cong, có thể liên quan đến sự thay đổi nhân khẩu học sau Thế chiến thứ hai. Đối với một hiệu ứng thấp hơn, chúng ta thấy rằng với phân phối Ý. Ví dụ, nếu chúng ta nhìn vào biểu đồ của Trung Quốc, chúng ta có thể thấy rằng có những thay đổi rất, rất lớn về tỷ lệ cá nhân tại các thời điểm khác nhau. Một lần nữa, liên quan đến yếu tố văn hóa và yếu tố lịch sử.

Vì vậy, một lần nữa, bằng cách so sánh và vẽ cùng một dữ liệu trong nhiều bảng điều khiển, chúng ta có thể chọn ra các mẫu hiệu quả hơn nhiều.

## 4j Trang tổng quan

Trang tổng quan là một cách sắp xếp dữ liệu để chúng ta có thể thấy nhiều biểu đồ và cách tất cả chúng được liên kết với nhau. Thông thường, bảng điều khiển được tạo bằng một số biểu đồ trong một bảng điều khiển. Những biểu đồ đó đến từ một tập dữ liệu nhưng tập dữ liệu đó có thể thay đổi theo thời gian. Chúng ta có thể sử dụng các tập dữ liệu khác nhau để tạo ra các loại biểu đồ giống nhau.

Điều quan trọng của các bảng điều khiển này là nhiều hình ảnh trực quan được sắp xếp cùng nhau, sao cho chúng xây dựng một câu chuyện mà bản thân một biểu đồ không đủ để cho chúng ta biết. Bằng cách sắp xếp nhiều biểu đồ theo cách thích hợp, chúng ta có thể xây dựng một bức tranh rất đầy đủ về câu chuyện cơ bản mà chúng ta muốn truyền đạt.

Một tính năng tuyệt vời khác của trang tổng quan là chúng có thể động, có nghĩa là chúng thay đổi khi dữ liệu được nhập vào hệ thống, để chúng ta có thể xem các sự kiện diễn ra như thế nào theo thời gian. Chúng ta có thể thay đổi thủ công các tập dữ liệu để xem liệu các mối quan hệ mà chúng ta thấy có đúng với một tập dữ liệu cũng đúng với một tập dữ liệu khác hay không.

Một tính năng rất hay về trang tổng quan là tính thẩm mỹ. Người dùng bị thu hút rất nhiều vào việc muốn sử dụng trang tổng quan. Khi chúng ta chỉ nhìn thấy một biểu đồ đôi khi có thể hơi nhàm chán. Nhưng khi chúng ta có nhiều biểu đồ sẽ thay đổi linh hoạt khi chúng ta làm điều gì đó thực sự thu hút sự chú ý của người dùng và khiến họ muốn tương tác với dữ liệu. Tuy nhiên, thường với bảng điều khiển, chúng ta có thể mất một chút thời gian và điều quan trọng cần nhớ là tính tiện dụng được đặt lên hàng đầu. Chúng cần đủ đơn giản để thông điệp không bị che khuất và nội dung chúng ta muốn truyền đạt cho người khác phải rất, rất rõ ràng. Nó không nên chỉ trông đẹp mà còn phải hiệu quả.

Để biết cách xây dựng trang tổng quan một cách thích hợp nhất, chúng ta luôn cần phải có trong đầu suy nghĩ của mình mục đích mà trang tổng quan này phục vụ. Ta đang cố gắng giao tiếp với dữ liệu này bằng những dữ liệu nào? Vì vậy, chúng ta hãy xem xét một chút về hai ví dụ nhanh về trang tổng quan chỉ để hiểu được ý nghĩa của trang tổng quan.

Đây là một ví dụ về một bảng điều khiển rất, rất đơn giản.

Nó cho biết 173.506 người hiện đang truy cập các trang web của chính phủ. Ngay lập tức chúng ta hiểu điều này đang muốn nói với chúng ta. Nhưng đây không chỉ là một biểu đồ. Chúng ta có thể thấy có nhiều bảng khác nhau ở đây và mỗi bảng cung cấp cho chúng ta thông tin khác nhau. Nhưng tất cả chúng đều được liên kết với cùng một tập dữ liệu cơ bản. Vì vậy, biểu đồ chính mà chúng ta thấy đang truyền đạt thông tin chính và chỉ đơn giản là tổng số lượt truy cập ngày hôm nay được chia nhỏ theo giờ. Dữ liệu đang được cập nhật từng phút theo thời gian thực. Đây là một trang tổng quan tự động thay đổi theo thời gian. Nó truyền tải một phần thông tin chính, tổng số cá nhân, cũng như chia nhỏ dữ liệu thành các phần phụ. Có bao nhiêu người đã ghé thăm theo giờ.

Nhưng thậm chí có nhiều tiểu mục hơn. Vì vậy, trên bảng điều khiển này ở phía bên tay phải, bạn có thể thấy tổng lượt truy cập vào mỗi trang. Và trang đang được truy cập thường xuyên nhất hiện nay là truyền hình NASA và sau đó là hai trang khác của NASA và sau đó là trang Dịch vụ Thời tiết Quốc gia. Và do đó, với số lượng giảm dần, chúng ta nhận được số lượng người trên mỗi trang.

Nó cũng tương tác, để chúng ta có thể xem qua các khoảng thời gian khác nhau. Nếu chúng ta đặt tab này, chúng ta sẽ thấy những trang nào được truy cập nhiều nhất trong bảy ngày qua. Và bạn có thể thấy người đứng đầu bây giờ không phải là NASA vì vậy sự kiện của NASA rõ ràng là một điều gì đó đang diễn ra ngày hôm nay.

Đó là cơ quan tài trợ chính cho Viện Y tế Quốc gia. Đây là hạn chót cho các đơn xin tài trợ. Đó có thể là lý do tại sao trong bảy ngày qua, đó là mức cao nhất. Và sau đó có thêm thông tin về thiết bị nào, trình duyệt nào, hệ điều hành nào mà các cá nhân truy cập vào các trang web của chính phủ.

Đây là bản chất của một bảng điều khiển. Đó là cùng một tập dữ liệu cơ bản nhưng cung cấp cho chúng ta nhiều phần thông tin khác nhau trong thời gian thực mà chúng ta có thể muốn biết. Có một phần thông tin chính và sau đó là các phần phụ.

Đây là ví dụ thứ hai về trang tổng quan là Google Xu hướng. Hầu hết mọi người có thể sẽ sử dụng điều này vào lúc này hay lúc khác. Bạn chỉ cần nhập cụm từ tìm kiếm của mình và bạn sẽ có được lịch sử của cụm từ tìm kiếm đó. Một cá nhân đã tìm kiếm một thuật ngữ cụ thể bao nhiêu lần? Hãy chỉ chọn New York Yankees làm một thuật ngữ. Chúng ta nhập nó vào và trang tổng quan sẽ tự động cập nhật. Và một lần nữa chúng ta có một biểu đồ chính ở giữa. Nó cho thấy theo thời gian, số lần truy cập cho cụm từ tìm kiếm New York Yankees, nó không phải là con số tuyệt đối, đây thực sự là một số liệu tỷ lệ. Chúng ta thực sự có thể thấy rằng từ năm đầu tiên được tính là năm 2005, đến năm 2015. Nhưng New York Yankees đã thực sự được tìm kiếm ngày càng ít lần, tương đối nhiều. Nếu chúng ta tắt nút này ở đây, chúng ta sẽ loại bỏ các tiêu đề tin tức mà Google sử dụng để thu hút sự chú ý đến các tính năng cụ thể. Điều rất thú vị về biểu đồ này là chúng ta có thể thấy rằng có tính tuần hoàn. Mô hình này giống nhau qua mỗi mùa giải và điều này rõ ràng là tương ứng với mùa bóng chày. Liệu nó có đạt đỉnh ở cuối khi New York Yankees lọt vào vòng loại trực tiếp. Mọi người quan tâm đến New York Yankees hơn họ nếu họ không lọt vào vòng loại trực tiếp.

Điều tuyệt vời về trang tổng quan là chúng ta có thể thêm dữ liệu vào đó và xem điều đó ảnh hưởng đến thông tin như thế nào. Nếu chúng ta tìm kiếm New York Mets, nó sẽ cập nhật thông tin đó nhận được một dòng khác trên biểu đồ. Và chúng ta có thể thấy rằng họ thậm chí còn ít phổ biến hơn New York Yankees. Ngoại trừ một trường hợp ngoại lệ là vào năm 2015, nơi họ đã thực hiện World Series. Và chúng đột nhiên trở nên rất phổ biến về lịch sử cụm từ tìm kiếm.

Nhưng bảng điều khiển này một lần nữa không chỉ là một biểu đồ. Ngoài ra còn có các biểu đồ khác. Vì vậy, bên dưới Google thêm bản đồ thế giới này. Và chúng ta có thể thấy rằng bóng chày phổ biến nhất ở Bắc Mỹ, không có gì đáng ngạc nhiên. Chúng ta có thể chuyển qua và di chuột qua để biết thêm thông tin về mức độ phổ biến chính xác và chúng ta có thể xem xét các quốc gia cụ thể hoặc chúng ta có thể nhấn nút này ở đây và tìm kiếm một thành phố cụ thể mà New York Yankees là thú vị nhất. Và vì vậy chúng ta có thể tự động cập nhật thông tin của mình theo thời gian.

Và điều này thực chất là những gì trang tổng quan cung cấp cho chúng ta là một cách tương tác với dữ liệu, để chúng ta có thể nhận được câu trả lời cho các câu hỏi và so sánh và đối chiếu các nhóm khác nhau theo cách sẽ rất khó nếu chúng ta chỉ mãi làm một biểu đồ tại một thời điểm.

Điểm đầu tiên mà ta muốn làm là khả năng tái sử dụng. Chúng ta có thể sử dụng lại các bố cục trang tổng quan giống nhau với nhiều tập dữ liệu khác nhau và chúng ta không cần phải tiếp tục cập nhật số liệu và vẽ lại số liệu bằng cách nhập các bộ dữ liệu khác nhau và sử dụng chương trình của chúng ta hoặc phần mềm khác của chúng ta để tạo các biểu đồ đó. Khi chúng ta đã tạo xong mẫu cung cấp dữ liệu và chúng ta có thể sử dụng lại trang tổng quan để tạo biểu đồ nhiều lần.

Ưu điểm thứ hai mà ta nhấn mạnh là chúng ta có thể có những hình dung liên kết. Điều này có nghĩa là chúng ta có thể xem xét một đối tượng địa lý trên một biểu đồ và đồng thời đối tượng địa lý đó sẽ nổi bật trên biểu đồ khác. Ví dụ: nếu chúng ta muốn xem Google Xu hướng năm 2006 - có bao nhiêu lượt truy cập - chúng ta có thể liên kết điểm đó - di chuột qua điểm đó trên biểu đồ đường với bản đồ để xem điều đó thay đổi như thế nào trong năm 2006 so với toàn bộ khoảng thời gian. Hình ảnh hóa được liên kết có thể là một cách rất hiệu quả để mang lại ý nghĩa cho dữ liệu.

Ưu điểm thứ ba của trang tổng quan là chúng ta có thể điều chỉnh mọi thứ một cách nhanh chóng, chúng ta có thể điều chỉnh chúng khi chúng ta tiếp tục đối với các thông số đặc biệt. Ví dụ: nếu chúng ta đang thực hiện phân tích khám phá, chúng ta có thể đang xây dựng một mô hình, một mô hình thống kê về dữ liệu của chúng ta. Và chúng ta có thể muốn nghĩ rằng chúng ta nên bao gồm biến này hay loại trừ biến này. Nếu chúng ta có một hằng số trong một thuật toán thống kê, có lẽ chúng ta muốn điều chỉnh hằng số đó và xem nó ảnh hưởng như thế nào đến đầu ra. Nếu chúng ta làm điều đó lặp đi lặp lại, vẽ lại biểu đồ mỗi lần, điều đó sẽ rất tốn công sức. Nó sẽ làm chậm quá trình học tập của chúng ta. Nhưng bằng cách có một bảng điều khiển, chúng ta có thể chỉ có một thanh trượt và chúng ta có thể trượt giá trị đó qua. Hoặc chúng ta có thể có một menu để chúng ta chọn các giá trị để nhập hoặc loại trừ. Và chúng ta thấy kết quả một cách linh động. Tiết kiệm được nhiều thời gian.

Điều đáng giá thứ tư là tính tương tác. Ta đã chỉ ra một số ví dụ về tương tác, chẳng hạn như di chuột qua các điểm và nhận thông tin. Nhưng với bảng điều khiển, chúng ta nhận được nhiều tương tác hơn. Người dùng thực sự có thể khám phá. Và ngày càng có nhiều trường hợp các cá nhân đang đọc và xem trực quan dữ liệu ngay lập tức cảm thấy như họ có thể di chuột qua các hình ảnh trực quan và nhận được nhiều thông tin hơn chỉ là một hình ảnh trực quan tĩnh. Vì vậy, người dùng đang bắt đầu mong đợi sự tương tác. Họ có thể học nhanh hơn bằng cách tự tương tác với dữ liệu và biểu đồ.

Và một số cá nhân có thể thấy rằng việc xem hình ảnh trực quan ở một dạng không thực sự giúp họ học được khái niệm cơ bản, nhưng nếu họ có thể thao tác dữ liệu để tự tạo ra một biểu đồ khác, điều đó sẽ giúp ích cho việc học và tiếp thu của họ của kiến ​​thức. Và do đó tính năng tương tác này là một tính năng thực sự hữu ích của trang tổng quan. Vì vậy, tại sao chúng ta nên sử dụng trang tổng quan trong suy nghĩ của ta, có hai lợi ích thực sự mạnh mẽ của việc sử dụng trang tổng quan trong hai trường hợp mà chúng ta muốn sử dụng chúng.

Đầu tiên, như ta đã nói trước đây, là chúng ta có một câu chuyện mà chúng ta muốn giao tiếp. Và cách hiệu quả nhất là hiển thị nhiều biểu đồ sẽ truyền đạt điểm đó. Đây sẽ là hình dung như một phương tiện giao tiếp của một sản phẩm cuối cùng. Chúng ta biết tường thuật của mình, chúng ta biết thông tin chúng ta muốn truyền tải và chúng ta sẽ xây dựng một bảng điều khiển để truyền đạt điều đó. Đây là cách sử dụng trang tổng quan phổ biến nhất.

Cách thứ hai trong đó một bảng điều khiển rất, rất hữu ích sẽ dành cho phân tích khám phá. Bây giờ, có nhiều cách chúng ta có thể thực hiện phân tích khám phá. Chúng ta có thể tự tạo nhiều biểu đồ, thử các biến khác nhau theo kiểu tĩnh hoặc trong phần mềm thống kê của riêng mình. Tuy nhiên, về lâu dài, việc xây dựng một bảng điều khiển thường là một cách nhanh hơn để thực hiện phân tích khám phá hơn là chỉ thử nhiều cách khác nhau để xem dữ liệu trong một gói thống kê.

Nhược điểm là công việc trả trước có nghĩa là chúng ta phải tùy chỉnh trang tổng quan của riêng mình và mất nhiều thời gian hơn một chút. Nhưng nó nhanh hơn nhiều về lâu dài. Một số ví dụ hoặc một ví dụ cụ thể về nơi có thể sử dụng trang tổng quan để phân tích khám phá nhằm phát triển những hiểu biết chính về mối quan hệ giữa các biến khác nhau và cách chúng ta có thể hình dung thông tin đó cũng như thực hiện một số dữ liệu phân tích về các mối quan hệ đó. Đi vào chi tiết hơn về các khía cạnh cụ thể của tập dữ liệu cơ bản. Đây là ý tưởng về hệ thống phân cấp thông tin thường có một số biểu đồ nên được rõ ràng nhất, nổi bật nhất. Chúng nên xuất hiện đầu tiên. Đây là những biểu đồ mà khi chúng ta có người dùng của mình lần đầu tiên nhìn vào hình ảnh trực quan, họ sẽ ngay lập tức hiểu được ý tưởng mà chúng ta đang cố gắng truyền đạt. Sau đó, các biểu đồ tiếp theo có thể là một tập hợp con của dữ liệu.

Lọc dữ liệu để hiển thị các mẫu phức tạp hơn. Chúng sẽ đến sau những biểu đồ ban đầu. Và vì vậy, khi chúng ta xây dựng một bảng điều khiển, chúng ta cần suy nghĩ xem người dùng của ta sẽ giao tiếp với thông tin này như thế nào. Chúng ta cần làm gì trước để họ hiểu những gì họ đang xem và điều gì sẽ xảy ra sau đó? Ví dụ, tất cả thông tin có nên xuất hiện trên một màn hình không? Một số người tranh luận rằng trang tổng quan chỉ nên xuất hiện trên một màn hình, nhưng những người khác sẽ tranh luận, chẳng hạn như ta, rằng có thông tin và biểu đồ trong các tab khác hoặc trên các menu có thể nhấp khác hoặc thông tin sẽ bật lên khi di chuột qua các biểu đồ cụ thể hoặc điểm dữ liệu là một cách rất hiệu quả để nhận được nhiều thông tin hơn. Và tất cả những điều này là để cân bằng giữa lựa chọn thiết kế với lựa chọn dữ liệu, điều gì trông hấp dẫn nhất? Điều gì sẽ lôi kéo mọi người tương tác với dữ liệu so với điều gì truyền đạt dữ liệu hiệu quả nhất? Và thực sự chúng ta nên luôn xem xét đặc điểm thứ hai đó một cách nổi bật nhất. Làm cách nào để ta có thể truyền đạt dữ liệu một cách hiệu quả nhất?