TRỰC QUAN HÓA DỮ LIỆU SỬ DỤNG R

*Bùi Ngọc Tân**

1. Tổng quan về trực quan hóa dữ liệu

Trực quan hóa dữ liệu là việc biểu diễn đồ họa các thông tin trừu tượng nhằm 2 mục đích: phân tích dữ liệu và truyền thông. Dữ liệu thường chứa đựng trong nó nhiều câu chuyện quan trọng và trực quan hóa dữ liệu là một một công cụ mạnh mẽ để khám phá và thấu hiểu những câu truyện này, và sau đó là truyền đạt tới những người khác. Thông tin thường trừu tượng vì nó miêu tả những thứ không hiện hữu. Thông tin thống kê lại càng trừu tượng hơn. Dù có liên quan tới doanh số bán hàng, tỷ lệ mắc bệnh, hoạt động thể chất hoặc bất kỳ thứ gì khác, thông tin không gắn với thế giới vật chất. Chúng ta chỉ có thể hiển thị nó một cách trực quan thông qua các việc diễn giải thông tin trừu tượng thành các đặc tính vật lý của thị giác (độ dài, vị trí, kích thước, hình dạng, màu sắc…) sử dụng quá trình tiếp nhận trực quan và nhận thức.

Trực quan hóa dữ liệu có hiệu quả bởi vì nó chuyển dịch sự bằng giữa nhận thức và tri thức nhằm tận dụng triệt để khả năng của não bộ con người. Nhận thức thị giác được xử lý bởi vỏ não thị giác nằm ở phía sau não bộ, cực kỳ nhanh nhạy và hiệu quả. Chúng ta nhìn thấy các sự vật hiện tượng ngay lập tức mà không mất một chút nỗ lực nào. Tri thức chủ yếu do vỏ não trước xử lý, thường chậm và kém hiệu quả hơn nhiều. Thông thường, các thức trình bày và suy diễn dữ liệu truyền thống chủ yếu dựa vào hoạt động nhận thức của não bộ. Trực quan hóa dữ liệu đã chuyển dịch sự cân bằng theo hướng tăng cường nhận thức thị giác, và tận dụng sức mạnh của mắt người bất cứ khi nào có thể.

Tuy nhiên, câu ngạn ngữ "một bức tranh bằng cả nghìn lời nói" chỉ đúng với một bức tranh được thiết kế tốt. Chúng ta có thể nghiền ngẫm một bảng số liệu cả ngày mà vẫn chẳng thể rõ ràng bằng một cái nhìn thoáng qua một bức tranh mô tả cùng các con số. Cụ thể hơn, những con số không thể giao tiếp khi được trình bày dưới dạng văn bản trong bảng số liệu, khi đó bộ não của chúng ta sẽ diễn giải sử dụng bộ xử lý tri thức và lời nói. Tuy nhiên, dữ liệu sẽ trở nên rõ ràng và dễ hiểu hơn khi được truyền thông trực quan sử dụng sức mạnh của "trực quan hóa dữ liệu".

Ngoài việc làm nổi bật mối quan hệ giữa các giá trị định lượng, trực quan hóa dữ liệu cũng được dùng để biểu diễn các mối quan hệ định tính. Ví dụ, mối quan hệ giữa mọi người trong một mạng xã hội như Facebook hoặc mạng lưới các nghi can khủng bố…

Lịch sử trực quan hóa dữ liệu

Con người đã biết sắp xếp dữ liệu thành các bảng (cột và dòng) ít nhất từ thế kỷ thứ 2, nhưng ý tưởng về việc biểu diễn các thông tin định lượng dưới dạng đồ họa mới chỉ xuất hiện trong thế kỷ thứ 17 theo sáng kiến của nhà triết học và toán học người Pháp Rene Decaster. Ông đã phát triển hệ thống tọa độ 2

* *Vụ Thống kê nước ngoài và Hợp tác quốc tế*

chiều gồm một trục hoành và trục tung dùng để biểu diễn trực quan các phép toán. Đến thế kỷ thứ 18, Scotsman William Playfair là người tiên phong trong việc khai thác tiềm năng của đồ họa trong việc truyền thông dữ liệu định lượng. Ông đã sáng tạo ra nhiều loại đồ thị mà ngày nay chúng ta vẫn sử dụng chẳng hạn như sử dụng đường kẻ đi lên và đi xuống theo chiều từ trái sang phải để biểu diễn sự thay đổi giá trị theo thời gian, biểu đồ cột, biểu đồ hình tròn.

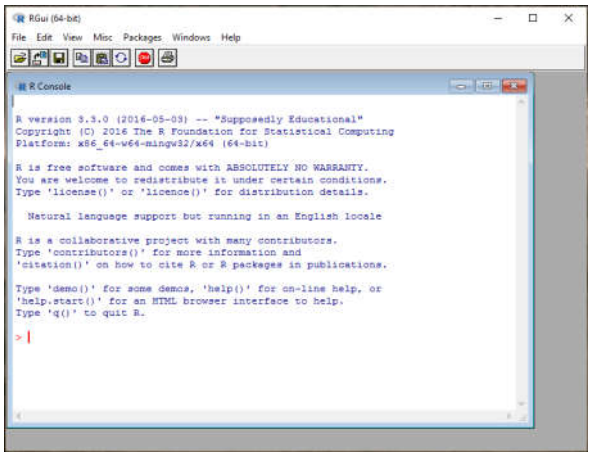
Việc sử dụng các biểu đồ trực quan hóa dữ liệu định lượng ngày càng phổ biến, nhưng cách thức và hiệu quả vẫn còn hạn chế. Trong nửa cuối thế kỷ 20, Jacques Bertin đã đặt nền móng cho rất nhiều tiến bộ trong lĩnh vực trực quan hóa bằng việc xuất bản cuốn sách "Triệu chứng học đồ họa" (The Semiology of Graphics) vào năm 1967. Cho đến năm 1983, Edward Tufte người được xem như cha để của trực quan hóa dữ liệu hiện đại đã xuất bản một cuốn sách mang tính đột phá "Biểu diễn trực quan thông tin định lượng" (The Visual Display of Quantitative Information). Trong đó ông đã chỉ ra rằng có những cách hiệu quả biểu diễn dữ liệu một cách trực quan tuy nhiên cách thức mà hầu hết mọi người thường làm không mang lại nhiều hiệu quả. Cũng phải kể đến công việc cải thiện các thực hành trực quan hóa dữ liệu của William Cleveland, người đã mở rộng và hoàn thiện các kỹ thuật trực quan hóa dữ liệu cho các nhà thống kê.

Kể từ khi bước sang thế kỷ 21, trực quan hóa dữ liệu đã được phổ biến rộng rãi, thông qua rất nhiều các phần mềm máy tính, chủ yếu là phần mềm thương mại. Tuy nhiên cũng có rất nhiều các phần mềm và nền tảng miễn phí giúp cải thiện trực quan hóa dữ liệu bằng cách tăng cường các tính năng thẩm mỹ và hiệu quả khai phá dữ liệu, truyền tải thông tin. Trong phần dưới đây chúng ta cùng xem xét R, một công cụ miễn phí và mạnh mẽ dùng để phân tích thống kê và trực quan hóa dữ liệu.

2. Sử dụng R trong phân tích và trực quan hóa dữ liệu

R là gì? Đây là một câu hỏi đơn giản nhưng không dễ trả lời. Theo định nghĩa rộng nhất thường được sử dụng để mô tả về các ngôn ngữ máy tính, R là một ngôn ngữ máy tính cho phép người sử dụng lập trình các thuật toán và sử dụng các công cụ đã được lập trình bởi những người khác. Cụ thể hơn, R là một ngôn ngữ lập trình và môi trường phần mềm dành cho tính toán và đồ họa thống kê.

Hình 1: Màn hình làm việc của R, hay còn gọi là cửa sổ lệnh



Ưu điểm của R trong thống kê và trực quan hóa dữ liệu

Hiện tại, có rất nhiều phần mềm có sẵn để phân tích dữ liệu: phần mềm bảng tính như Excel, các hệ thống dựa trên thủ tục như SAS, các hệ thống dựa trên giao diện người dùng như SPSS, và rất nhiều các hệ thống khai thác dữ liệu khác. Điều gì làm cho R trở nên khác biệt?

Thứ nhất, R miễn phí. Là một dự án mã nguồn mở, người sử dụng có thể tải về và sử dụng R miễn phí và không phải lo lắng về phí thuê bao, quản lý giấy phép, hoặc giới hạn người dùng. Nhưng quan trọng nhất, R là một hệ thống mở, trong đó bất kỳ người sử dụng

nào cũng có thể kiểm tra mã và đóng góp để hoàn thiện R. Hàng nghìn chuyên gia trên khắp thế giới đã và đang làm điều này, và những đóng góp của họ mang lại lợi ích cho hàng triệu người sử dụng R ngày hôm nay.

Thứ hai, R là một ngôn ngữ. Trong R, người sử dụng phân tích dữ liệu bằng cách viết các hàm và kịch bản, không phải bằng cách trỏ và nhấn chuột. Tưởng chừng đây có vẻ là một điểm yếu, nhưng thực sự R là một ngôn ngữ dễ học, và là một cách tự nhiên và mạnh mẽ dùng cho phân tích dữ liệu. Khi đã học và biết về ngôn ngữ này, người sử dụng sẽ cảm nhận được rất nhiều tiện ích. Ngoài ra các kịch bản phân tích có thể được lưu lại và sử dụng nhiều lần cũng như tự động hóa một chuỗi các tác vụ, và được lồng ghép trong những quá trình khác.

Thứ ba, R rất mạnh trong hoạt động đồ họa và trực quan dữ liệu. Một trong những nguyên tắc đầu tiên khi thiết kế R là các biểu đồ và đồ thị là một phần thiết yếu của quá trình phân tích dữ liệu. Kết quả là, R có các công cụ đồ họa tuyệt vời, từ những biểu đồ được sử dụng phổ biến như biểu đồ cột và biểu đồ tán xạ cho tới các loại đồ họa phức tạp tích hợp ma trận các biểu đồ, hoặc các loại đồ họa mới do người dùng tự sáng tạo. Kết quả là các loại hình đồ họa dựa trên R thường xuyên xuất hiện trong những ấn phẩm phổ thông như Thời báo New York, the Economist v.v.

Thứ tư, R là một bộ công cụ phân tích thống kê linh hoạt. Tất cả các công cụ phân tích dữ liệu tiêu chuẩn đều được xây dựng bằng chính ngôn ngữ R: từ việc truy cập dữ liệu trong nhiều định dạng khác nhau, cho tới các thao tác trên dữ liệu (biến đổi biến, trộn và tổng hợp biến v.v…), tới các mô hình thống kê truyền thống và hiện đại như (hồi quy, ANOVA, GLM, mô hình cây v.v…). Tất cả các kỹ thuật nói trên đều được xây dựng dựa trên khuôn khổ hướng đối tượng, giúp dễ dàng trích xuất và kết hợp thông tin từ các kết quả phân tích, chứ không chỉ đơn thuần là cắt và dán từ một kết quả đơn lẻ.

Thứ năm, R luôn được cập nhật các kỹ thuật phân tích và đồ họa mạnh mẽ, tiên tiến nhất. Các học giả và các nhà nghiên cứu hàng đầu từ khắp nơi trên thế giới sử dụng R để phát triển các phương pháp mới nhất trong thống kê, máy học, và mô hình dự đoán. R còn được sử dụng rộng rãi trong lĩnh vực tài chính, phân tích di truyền và hiều lĩnh vực khác. Hiện tại có hơn 2000 gói mở rộng trong R áp dụng trong mọi lĩnh vực, có sẵn để tải về. Ngoài ra, R còn có một cộng đồng người dùng mạnh mẽ và tích cực. Với hàng ngàn người đóng góp mã nguồn và hơn hai triệu người dùng trên khắp thế giới, bất kỳ câu hỏi nào về R đều sẽ được giải đáp cặn kẽ. Có thể nói các nguồn lực cộng đồng cho R luôn sẵn có trên Internet và hỗ trợ mọi lĩnh vực.

Thứ sáu, tính đa nền tảng. R chạy được trên rất nhiều các nền tảng hệ điều hành khác nhau gồm Windows, Unix, và Mac OS X. Như vậy người sử dụng có thể chạy phần mềm này trên bất kỳ máy tính sẵn có này.

Thứ bảy, R có khả năng vô hạn. Với R, người sử dụng có thể sử dụng các đoạn mã do người khác đóng góp trong một cộng đồng mã nguồn mở hoặc tự xây dựng các hàm, công cụ R của riêng mình. R cũng là một công cụ tuyệt vời để tương tác với các ứng dụng khác: kết hợp R với cơ sở dữ liệu MySQL, máy chủ web Apache, và với các giao diện lập trình API của dịch vụ Google Maps, giúp người dùng có các công cụ phân tích GIS theo thời gian thực v.v…

***Nhược điểm của R** trong thống kê và trực quan hóa dữ liệu:*

Thứ nhất, R tương đối khó nắm bắt đối với những người mới sử dụng do chỉ có giao diện dòng lệnh. Tuy đã có nhiều giao diện

người dùng đồ họa (GUI) dễ sử dụng cho R, chẳng hạn như RGUI, R Commander, RStudio, RKWard…, sử dụng các tương tác chỏ và nhấn, nhưng nhìn chung các giao điện này không thực sự đẹp mắt như các phần phềm thương mại.

Thứ hai, tài liệu hướng dẫn về R rất nhiều nhưng quá ngắn gọn, tương đối khó hiểu đối với những người không chuyên về thống kê. Tuy nhiên, ngày càng nhiều các tài liệu, sách hướng dẫn có chất lượng cao về R được xuất bản, giúp người sử dụng dễ dàng tiếp cận với R hơn.

Thứ ba, chất lượng của một số gói mở rộng trong R còn chưa thực sự hoàn hảo. Tuy nhiên, nếu một gói mở rộng được nhiều người sử dụng, nó sẽ nhanh chóng phát triển thành một sản phẩm mạnh mẽ thông qua các nỗ lực hợp tác trong cộng đồng người sử dụng R. Ngoài ra người sử dụng R còn có thể mua các gói hỗ trợ từ một số nhà cung cấp quốc tế.

Thứ tư, nhiều lệnh trong R không thực sự chú trọng vào việc quản lý bộ nhớ, do đó R có thể nhanh chóng chiếm dụng hết bộ nhớ trong máy tính. Đây được xem là một hạn chế khi thực hiện các hoạt động khai thác dữ liệu. Có nhiều giải pháp đối với vấn đề này, chẳng hạn việc sử dụng một hệ điều hành 64-bít có thể giúp truy cập nhiều bộ nhớ hơn.

3. Ví dụ về trực quan hóa dữ liệu sử dụng R

Dưới đây là một bảng đơn giản về doanh số bán hàng trong năm của một công ty phân theo khu vực nội địa và quốc tế. Bảng 1 đã làm rất tốt 2 việc: biểu thị giá trị doanh số bán một cách chính xác và cung cấp một phương tiện hiệu quả để tra cứu các giá trị theo khu vực và các tháng nhất định.

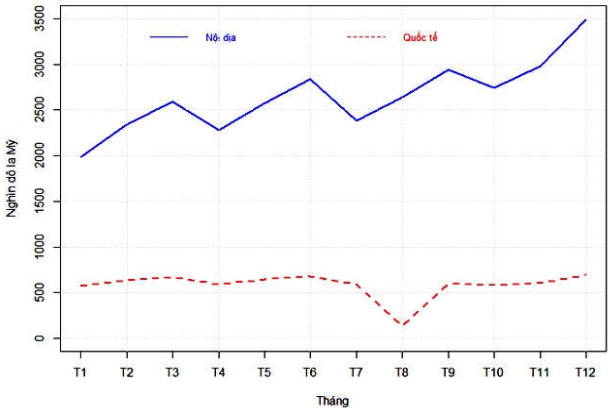
Bảng 1: Doanh số bán hàng năm 2015

Đơn vị: nghìn đô la Mỹ

Thời gian	Nội địa	Quốc tế	Tổng
Tháng 1	1983	574	2557
Tháng 2	2343	636	2979
Tháng 3	2593	673	3266
Tháng 4	2283	593	2876
Tháng 5	2574	644	3218
Tháng 6	2838	679	3571
Tháng 7	2382	593	2975
Tháng 8	2634	139	2773
Tháng 9	2938	599	3537
Tháng 10	2739	583	3322
Tháng 11	2983	602	3585
Tháng 12	3493	690	4183
Tổng	**31783**	**7005**	**38788**

Tuy nhiên, nếu chúng ta cần xem xét các hình mẫu, xu hướng hoặc sự khác biệt giữa các giá trị nói trên, chúng ta muốn có một cái nhìn thoáng qua về câu truyện chứa đựng trong các con số nói trên, hặc nếu chúng ta cần so sánh toàn bộ tập số liệu chứ không phải chỉ là so sánh 2 số cùng một lúc, bảng này không thể làm được những điều trên.

Hãy xem xét đồ thị đường kẻ dùng để biểu diễn dữ liệu trong Hình 2:



Hình 2: Biểu đồ đường thể hiện doanh số bán hàng năm 2015

Thông qua đồ thị ở trên, có thể thấy:

- Doanh số bán nội địa cao hơn nhiều và có xu hướng tăng vững chắc so với doanh số bán quốc tế.

- Nhìn tổng thể, doanh số bán nội địa có xu hướng tăng trong cả năm.

- Doanh số bán quốc tế, ngược lại, tương đối ổn định mặc dù có sụt giảm đáng kể trong tháng 8.

- Doanh số bán nội địa thể hiện hình mẫu tăng giảm theo chu kỳ quý, cụ thể luôn luôn tăng cao trong những tháng cuối của quý sau đó giảm mạnh trong tháng đầu của quý tiếp theo.

Ví dụ tiếp theo là bảng số liệu về các nguyên nhân tử vong tại Mỹ năm 2007:

Bảng 2: Nguyên nhân tử vong năm 2007 tại Mỹ

Nguyên nhân tử vong	Số ca tử vong
Bệnh tim	616067
Các bệnh khác	577181
Ung thư	562875
Đột quỵ	135952
Bệnh hô hấp mãn tính	127924
Tai nạn	123706
Alzheimer	74632
Đái đường	71382
Cúm và viêm phổi	52717
Hội chứng thận hư	46448
Nhiễm trùng máu	34828
Tổng	**2423712**

Với mục đích chuyển tải các thông tin vắn tắt thành các trình bày trực quan mà con người có thể giải mã một cách dễ dàng, hiệu quả, chính xác và có ý nghĩa, hãy cùng xem xét và phân tích việc sử dụng biểu đồ hình tròn để biểu diễn dữ liệu liên quan đến nguyên nhân tử vong (xem Hình 3 bên dưới).

Có thể thấy, trực quan hóa dữ liệu sử dụng biểu đồ hình tròn ở trên đã đạt được mục đích truyền tải thông tin tới người sử dụng. Cụ thể:

- *Chỉ rõ bản chất mối quan hệ.* Điểm mạnh của biểu đồ hình tròn là chỉ rõ mối quan hệ giữa từng phần với tổng thể giữa các giá trị.

- *Biểu diễn số lượng một cách chính xác.* Biểu đồ hình tròn mã hóa các giá trị sử dụng 3 tính chất hình ảnh: diện tích của mỗi miếng cắt, góc độ của mỗi miếng cắt tại tâm của hình tròn, và độ dài của mỗi miếng cắt theo chu vi đường tròn. Nhận thức hình ảnh của con người không hỗ trợ nhiều trong việc giải mã diện tích, góc độ và độ dài của các miếng cắt. Tuy nhiên, việc bổ sung các giá trị phần trăm tương ứng với mỗi miếng cắt sẽ giúp cảm nhận một cách chính xác hơn về dữ liệu được biểu diễn trong đồ thị.

- *Dễ dàng so sánh số lượng.* Do có thể cảm nhận một cách chính xác nên chúng ta có thể so sánh số lượng một cách dễ dàng và chính xác. Tuy nhiên, biểu đồ này sử dụng các ghi chú để ghi nhãn các miếng cắt và buộc người sử dụng phải nhìn kỹ các ghi chú, làm cho việc so sánh trở nên khó khăn hơn. Sử dụng biểu đồ cột trong trường hợp này sẽ giúp cho việc so sánh số lượng dễ dàng hơn do người sử dụng có thể dễ dàng so sánh chiều dài các thanh trong biểu đồ.

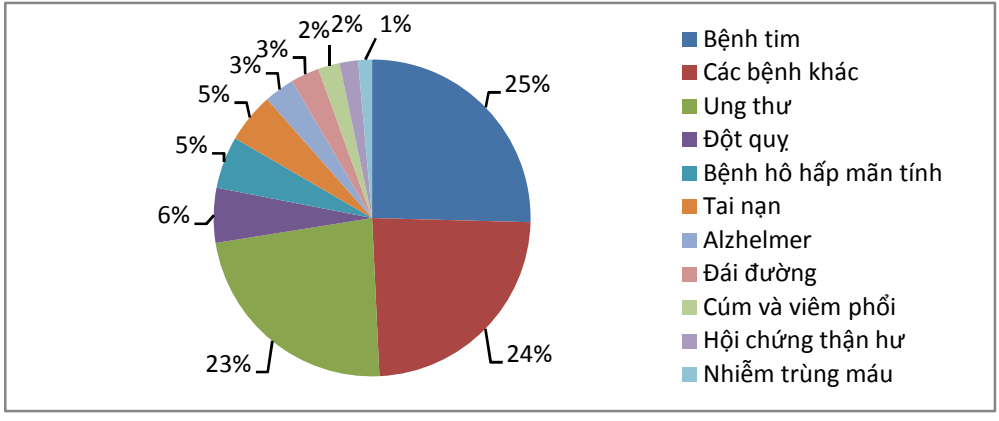
- *Thấy được thứ tự xếp hạng của các giá trị.* Trong biểu đồ hình tròn ở trên, sự khác biệt về diện tích, góc độ và chiều dài theo chu vi của các miếng cắt giúp người sử dụng có thể thấy được thứ tự xếp hạng các nguyên nhân gây ra tử vong từ cao đến thấp theo chiều kim đồng hồ.

- *Làm rõ cách thức sử dụng thông tin.* Thông qua việc so sánh các miếng cắt trong biểu đồ hình tròn, người sử dụng có thể hiểu rõ hơn về các nguyên nhân gây ra tử vong và tỷ lệ của các nguyên nhân so với tổng số ca tử vong tại Mỹ trong năm 2007.

(Xem tiếp trang 26)

(Tiếp theo trang 21)

Hình 3: Biểu đồ thể hiện nguyên nhân tử vong tại Mỹ năm 2007



Tài liệu tham khảo:

1. Winston Chang, R Graphics Cookbook, 2013;
2. Nguyen Van Tuan, Phân Tích Dữ Liệu Với R, 2014;
3. Data Journalism Handbook, Using Data Visualization to Find Insights in Data, 2011.