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

Nhóm 08

TRỰC QUAN HÓA DỮ LIỆU

BÁO CÁO ĐỒ ÁN GIỮA KỲ

GIÁO VIÊN HƯỚNG DẪN
Thầy Bùi Tiến Lên

Tp. Hồ Chí Minh, tháng 4/2024

Lời cảm ơn

Báo cáo đồ án giữa kỳ là kết quả của quá trình cố gắng không ngừng của bản thân và được sự giúp đỡ của các thầy cô, bạn bè. Qua trang viết này nhóm xin gửi lời cảm ơn tới những người đã giúp đỡ cả nhóm trong thời gian học tập - nghiên cứu khoa học.

Xin tỏ lòng kính trọng và biết ơn sâu sắc đối với thầy Bùi Tiến Lên - giảng viên lý thuyết đã trực tiếp tận tình hướng dẫn cũng như cung cấp tài liệu thông tin khoa học cần thiết cho Đồ án này.

Nhóm xin chân thành cảm ơn!

Mục lục

Danh sách hình

Chương 1

Thông tin nhóm

1.1 Thông tin

Họ tên	MSSV
Phạm Nhật Duy	21120058
Hồ Sỹ Kiên	21120091
Hoàng Thành Nam	21120099
Trương Công Trung	21120158
Đinh Thị Thúy Hường	21120176

1.2 Đánh giá đồ án

- Xây dựng dashboard: 100%
- Các tiêu chí đánh giá: 100%
- Thuyết trình: 100%

Chương 2

Nội dung đồ án

2.1 Xây dựng dashboard

2.1.1 Thông tin chung về dữ liệu

Chủ đề được nhóm tham khảo từ *Chapter 16 (The Big Book of Dash-Boards): Sentiment Analysis: Showing overall distribution.* Qua đó, nhóm đã lựa chọn thống kê về đánh giá của người xem về những thể loại phim trên nền tảng Rotten Tomatoes.

Dữ liệu được thu thập một cách tỉ mỉ và hiệu quả từ trang web có thể truy cập công khai `https://www.rottentomatoes.com` kể từ ngày 12 tháng 4 năm 2023 bằng cách sử dụng các kỹ thuật Python và được tác giả công khai trên nền tảng Kaggle (`https://urlvn.net/rotten-tomatoes`) với giấy phép *CC0: Public Domain* - hợp pháp để sử dụng trong lĩnh vực nghiên cứu, học tập.

2.1.2 Xây dựng dashboard

Link dashboard: `https://urlvn.net/dashboard-rotten-tomatoes`.

Dashboard được xây dựng bằng cách kết hợp xử lý dữ liệu bằng phần mềm Microsoft Excel và trực quan hóa dữ liệu trên phần mềm Microsoft Power BI.

Để thống kê và phân tích những đánh giá của người xem, nhóm đã sử dụng kết hợp biểu đồ cột, biểu đồ boxplot để biểu diễn phân phối của dữ liệu và biểu đồ hình tháp để trực quan tương quan giữa 2 nhóm dữ liệu.

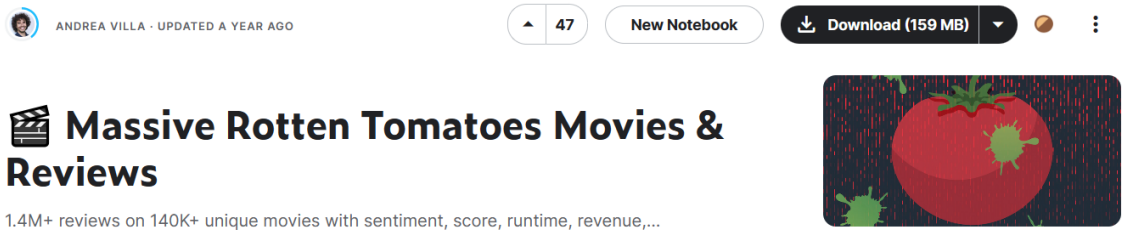
Mô tả encode dữ liệu:

Dữ liệu	Kiểu dữ liệu	Encode	Mô tả
tomatoMeter, audienceScore	Quantitative	Size, Color	Số lượng đánh giá trong khoảng càng nhiều thì cột biểu diễn càng cao. Nhóm điểm đánh giá tích cực sẽ có màu cam và ngược lại thì màu xanh.
genre	Categorical	Position	Mỗi thể loại phim được biểu diễn ở một dòng riêng biệt.

2.2 Các tiêu chí đánh giá

2.2.1 Tiêu chí 1: Nguồn dữ liệu đáng tin cậy

Nguồn dữ liệu được lấy từ Kaggle - một trong những nền tảng cung cấp dữ liệu lớn nhất trên thế giới. Tập dữ liệu được lấy từ một bài nghiên cứu uy tín với số lượt upvote cao. Dữ liệu được tác giả thu thập từ nền tảng Rotten Tomatoes - một trang đánh giá phim trung thực, uy tín và chuyên nghiệp.

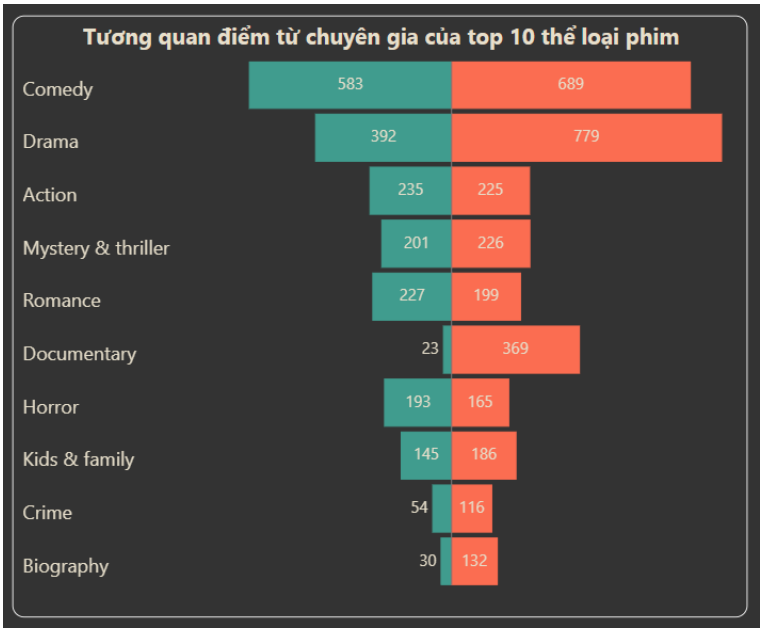


Hình 2.1: Dữ liệu được công khai trên Kaggle

2.2.2 Tiêu chí 2: Phù hợp với mục đích

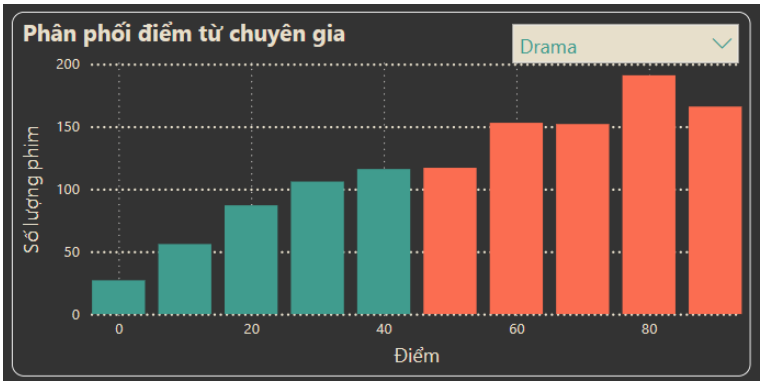
Các biểu đồ được sử dụng chính xác với mục đích và thể hiện đầy đủ ý nghĩa tương quan khi trực quan hóa dữ liệu.

Biểu đồ **hình tháp** (hình 2.2): vừa quan sát được tổng thể, vừa nhận biết tương quan tại từng điểm dữ liệu.



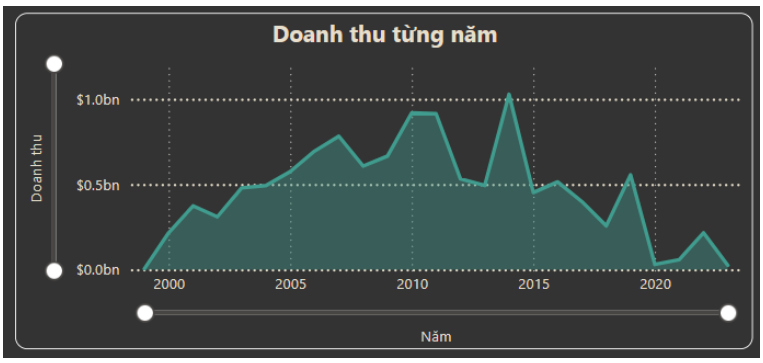
Hình 2.2: Biểu đồ hình tháp

Biểu đồ **tần suất** (hình 2.3): nhìn nhận được phân phối dữ liệu theo từng mức điểm và phân biệt được giữa 2 nhóm điểm tích cực và tiêu cực.



Hình 2.3: Biểu đồ tần suất dữ liệu

Biểu đồ **vùng** (hình 2.4): quan sát được sự biến động doanh thu từng phim theo thời gian.



Hình 2.4: Biểu đồ vùng

2.2.3 Tiêu chí 3: Rõ ràng và dễ hiểu

Mỗi biểu đồ được lựa chọn tỉ mỉ, phù hợp và phản ánh cụ thể ý nghĩa của dữ liệu giúp người đọc dễ hiểu, không bị hoang mang khi tiếp cận.

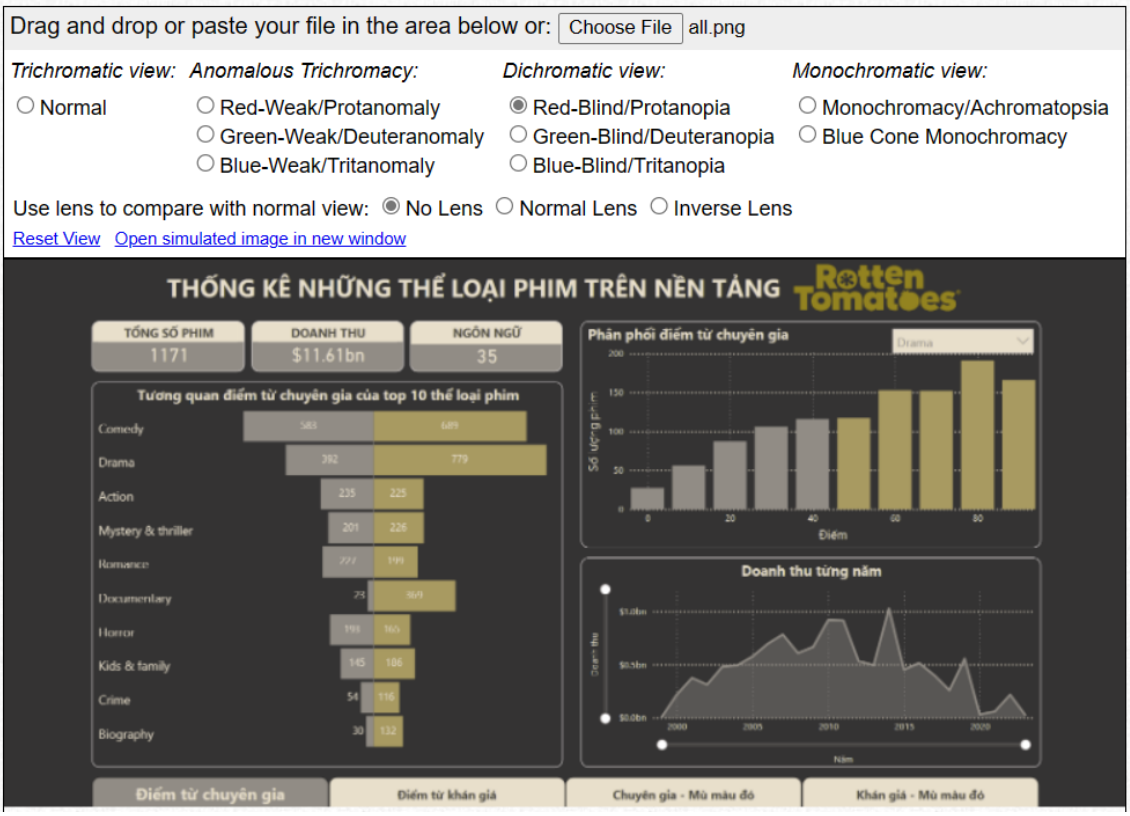
Đồng thời, mỗi biểu đồ đều có **Tên biểu đồ, nhãn cột x và y** góp phần có cái nhìn rõ ràng không bị mơ hồ. Bên cạnh đó, người đọc có thể tương tác, lựa chọn thông tin cần tìm hiểu khi cần thiết với những tiện ích được tích hợp sẵn trong Power BI.

Dashboard được nhóm xây dựng phù hợp với cả đối tượng bị mù màu bằng cách điều chỉnh những phần màu sắc thành những sắc độ của một màu. Điều này góp phần giúp dashboard có thể tiếp cận chính xác nhất với tất cả mọi người.

Hình 2.5: Các phiên bản dành cho đối tượng mù màu

Nhóm đã kiểm tra dashboard bằng trang web **Coblis Color Blindness Simulator** để chắc chắn sản phẩm thu được phù hợp với mọi đối tương (không mù màu, mù màu đỏ, mù màu xanh, ...)

Link trang web: `https://urlvn.net/color-blind-test`.



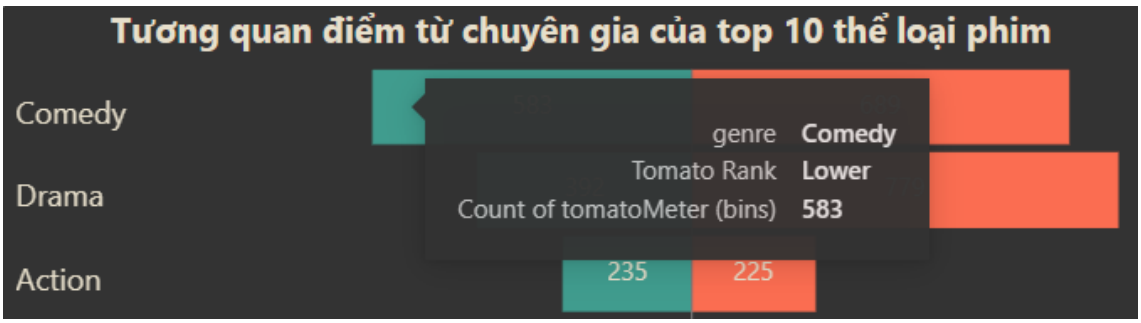
Hình 2.6: Kiểm tra với loại mù màu đỏ

2.2.4 Tiêu chí 4: Sự tích hợp và liên kết

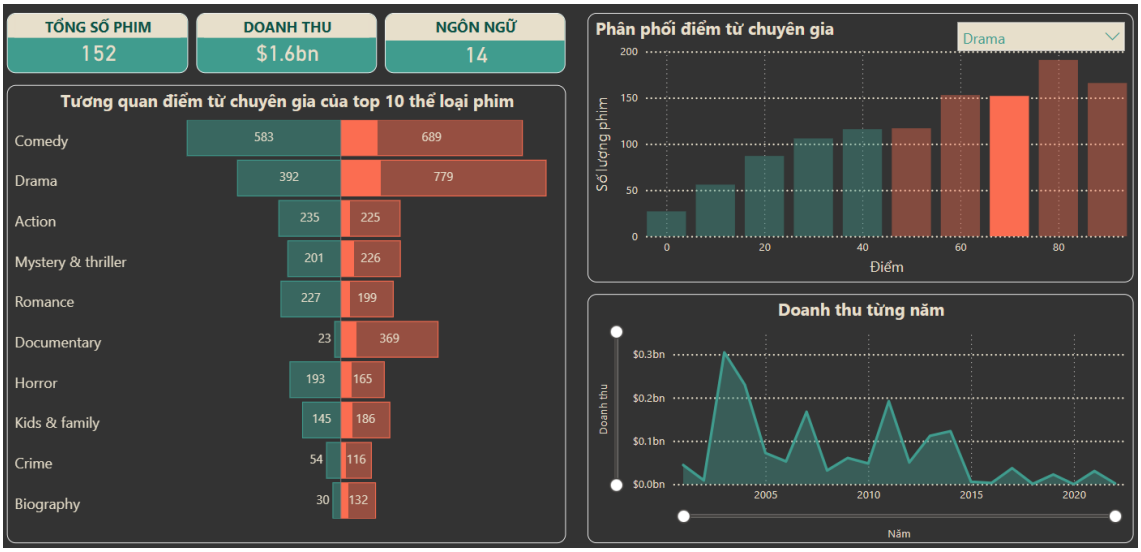
Mức độ tích hợp cao: kết hợp nhiều dạng trực quan hóa (biểu đồ, đồ thị, số liệu, bảng) để thống kê từng thể loại phim dựa trên dữ liệu). Ngoài ra, có bộ lọc để người dùng dễ dàng chọn lọc dữ liệu theo từng tiêu chí phù hợp với nhu cầu.

Liên kết trực quan: màu sắc được sử dụng thống nhất cho các nhóm dữ liệu trong mọi biểu đồ của dashboard (tức là màu cam sẽ đại diện cho nhóm đánh giá tích cực và màu xanh sẽ đại diện cho nhóm đánh giá tiêu cực), từ đó người dùng dễ dàng so sánh dữ liệu.

Liên kết thông tin: chọn dữ liệu theo thể loại phim trên một biểu đồ thì dashboard sẽ tự động cập nhật dữ liệu liên quan trên các biểu đồ khác. Hơn thế nữa khi di chuột qua các điểm dữ liệu, biểu đồ sẽ hiện thị thông tin chi tiết về tên thể phim, thể loại, doanh thu và đánh giá từ người dùng.



Hình 2.7: Thông tin chi tiết thể loại Comedy



Hình 2.8: Sự thay đổi với mức điểm 70

2.2.5 Tiêu chí 5: Sự thay đổi và xu hướng

Mối quan hệ giữa các biến rõ ràng, dễ hiểu. Qua quan sát, nhóm rút ra được một số kết luận như sau:

- **Doanh thu và điểm đánh giá**: phim có điểm đánh giá cao thường sẽ mang lại danh thu cao hơn.
- **Thể loại và điểm đánh giá**: phim tài liệu có điểm đánh giá cao nhiều hơn so với điểm đánh giá thấp, phim kinh dị có số điểm đánh giá thấp khá nhiều so với những đánh giá cao.
- **Doanh thu và năm**: dựa trên biểu đồ đường, nhóm thể hiện sự thay đổi doanh thu theo thời gian.

2.2.6 Tiêu chí 6: Tương tác và điều hướng

Tính tương tác:

- Biểu đồ có khả năng phả hồi dựa trên tương tác của người dùng để lọc thông tin, hoặc hiện thị thông tin chi tiết khi người dùng di chuột qua để cung cấp thêm thông tin, vì vậy tổng thể dashboard vẫn đầy đủ và không gây hoang mang, cồng kềnh cho người xem.
- Bảng điều khiển ở bên dưới cung cấp khả năng di chuyển tới những trang dashboard được đánh giá theo tiêu chí khác hoặc một phiên bản khác dành cho đối tượng bị mù màu.

Hình 2.9: Bảng điều khiển

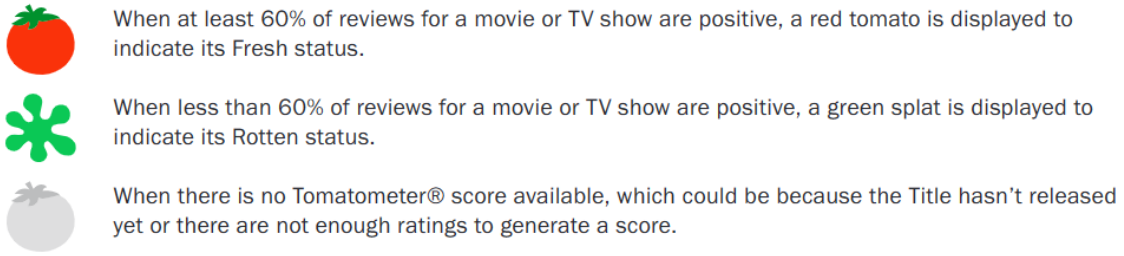
Tính điều hướng:

- Giao diện được thiết kế rõ ràng và dễ sử dụng. Từ menu, biểu đồ, nút lệnh được tổ chức một các logic và có thứ tự giúp người dùng dễ dàng quan sát, tìm kiếm và sử dụng những chức năng mong muốn. Cụ thể: những thông tin tổng quan sẽ được đặt để phía bên trái dashboard, dần sang phải sẽ là những biểu đồ phân tích chi tiết hơn về những thuộc tính dữ liệu.
- Ngoài ra, các biểu đồ được sắp xếp và gom nhóm theo từng chức năng cũng như ý nghĩa mà bản thân biểu đồ thể hiện. Ví dụ: nhóm biểu đồ thể hiện phân phối dữ liệu, nhóm số liệu tổng quát của thể loại phim, ...

2.2.7 Tiêu chí 7: Thiết kế hấp dẫn

Thiết kế đồ họa hấp dẫn: kết hợp nhiều biểu đồ ý nghĩa, các cột và đường đều thể hiện thông tin trực quan của dữ liệu.

Sử dụng màu sắc có ý nghĩa: màu sắc được sử dụng hài hòa, đồng nhất và có ý nghĩa như màu xanh đại diện cho điểm đánh giá thấp còn màu cam sẽ đại diện cho điểm đánh giá cao (lấy cảm hứng từng quả cà



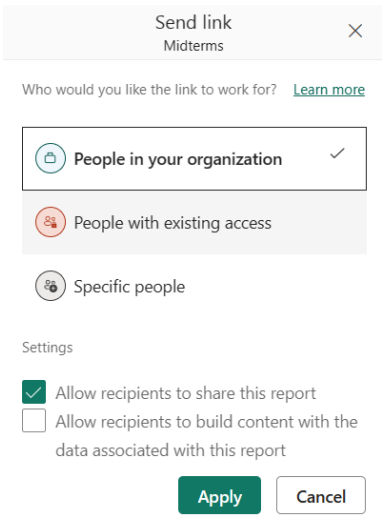
Hình 2.10: Cảm hứng lựa chọn màu sắc

chua - tinh thần của nền tảng Rotten Tomatoes). Bên cạnh đó, màu sắc được sử dụng tinh chỉnh hợp lý, không bị quá tải hay rối mắt, và đặc biệt là có cả phiên bản được tùy chỉnh cho đối tượng mù màu.

2.2.8 Tiêu chí 9: Khả năng tích hợp và chia sẻ

Tích hợp: được xây dựng bằng Power BI nên có thể liên kết và phần mềm Microsoft PowerPoint, giúp quá trình trình bày trở nên chuyên nghiệp và thuận tiện.

Chia sẻ: dễ dàng chia sẻ dưới dạng có thể chỉnh sửa giúp làm việc nhóm hiệu quả hơn hoặc dạng chỉ xem khi trình chiếu giúp tăng tính an toàn cho dashboard. Tuy nhiên, những người dùng không cùng chung tổ chức sẽ không thể truy cập dashboard mà nhóm đã chia sẻ (do sử dụng bản Standard), cần email của trường cung cấp để có thể truy cập thuận lợi hơn.



Hình 2.11: Tùy chọn chia sẻ

2.2.9 Tiêu chí 10: Hiệu suất

Dashboard có khả năng kết nối và lưu trữ lượng dữ liệu lớn và truy xuất liên tục. Do nền tảng Power BI hổ trợ tối ưu dữ liệu lớn (tối đa 8 - 10 triệu dòng), đồng thời hỗ trọ truy cập dữ liệu từ nhiều nguồn.

Bên cạnh đó, thời gian tải hay tốc độ phản hồi sẽ phụ thuộc vào đường truyền mạng người dùng sử dụng.

2.2.10 Nguyên lý trực quan hóa - Visual Perception

Similarity: các nhóm điểm đánh giá tích cực hay tiêu cực sử dụng 2 màu sắc tương phản để cho thấy sự khác biệt.

Proximity: những nhóm biểu đồ khác loại sẽ được phân bố riêng biệt và chia cột theo nội dung thể hiện (từ tổng quát đến chi tiết).

Enclosure: biểu đồ được bao viền lại để người xe có thể gom nhóm thông tin và phân biệt rõ ràng với các biểu đồ khác.

Connection: các nhóm dữ liệu mang ý nghĩa giống nhau sẽ có màu giống nhau để tạo sự liên kết giữa các biểu đồ. Ví dụ: nhóm dữ liệu thể hiện đánh giá tích cực sẽ luôn có màu cam và ngược lại thì sẽ màu xanh.

Tài liệu tham khảo

[1] The Big Book of Dashboard, e-Book. Online. Available. https://www.bigbookofdashboards.com/. Accessed: April 01, 2024

[2] Color Blindness. Online. Available. https://www.color-blindness.com//. Accessed: April 02, 2024

[3] Visual Perception. Online. Available. https://urlvn.net/visual-perception. Accessed: April 01, 2024