

PHẦN 3

TRỰC QUAN HÓA DỮ LIỆU

(Data Visualization)

ĐỒ THỊ & BIỂU ĐỒ

(GRAPHS AND CHARTS)¹

Đồ thị (graphs) và biểu đồ (charts) có thể sắp xếp và trình bày dữ liệu phức tạp giúp mọi người dễ hiểu hơn. Biết được sự khác biệt giữa các loại biểu đồ và biểu đồ khác nhau có thể giúp chọn loại biểu đồ tốt nhất cho dự án của mình.

10.1. TỔNG QUAN

Biểu đồ là một phần thiết yếu khi làm việc với dữ liệu vì chúng là cách để cô đọng lượng lớn dữ liệu thành một định dạng dễ hiểu. Trực quan hóa dữ liệu có thể mang lại hiểu biết sâu sắc cho người xem dữ liệu lần đầu tiên, cũng như truyền đạt kết quả cho những người khác không nhìn thấy dữ liệu thô. Có vô số loại biểu đồ, mỗi loại có trường hợp sử dụng khác nhau. Thông thường, phần khó nhất trong việc tạo trực quan hóa dữ liệu là tìm ra loại biểu đồ nào phù hợp nhất cho nhiệm vụ hiện tại.

Sự lựa chọn loại biểu đồ của bạn sẽ phụ thuộc vào nhiều yếu tố. Các loại số liệu, tính năng hoặc các biến khác mà bạn dự định vẽ biểu đồ là gì? Đối tượng mà bạn dự định trình bày là ai - đó chỉ là sự khám phá ban đầu của bạn hay bạn đang trình bày cho nhiều đối tượng hơn? Loại kết luận mà người trình bày muốn người đọc rút ra là gì?

10.1.1. Sự khác nhau giữa graphs và charts?

Mặc dù nhiều người sử dụng graph và chart thay thế cho nhau nhưng chúng là những hình ảnh trực quan khác nhau. Chart là bảng, sơ đồ hoặc hình ảnh sắp xếp lượng lớn dữ liệu một cách rõ ràng và chính xác. Mọi người sử dụng chart để giải thích dữ liệu hiện tại và đưa ra dự đoán. Còn graphs tập trung vào dữ liệu thô và hiển thị xu hướng theo thời gian.

Nắm vững các loại biểu đồ giúp truyền đạt thông tin hiệu quả, tóm tắt dữ liệu phức tạp, so sánh và phân tích, dự đoán xu hướng, và làm dữ liệu hấp dẫn hơn. Điều này quan trọng và ứng dụng rộng rãi trong nhiều lĩnh vực.

10.1.2. Lưu ý chung khi sử dụng graph hoặc chart

- Luôn bao gồm tiêu đề (title), nhãn trục (axis labels) và tỷ lệ (scale) được xác định rõ ràng.
- Biểu đồ dạng cột (bar, histogram, stacked column/bar graph) cần sử dụng thang đo có ý nghĩa đối với dữ liệu được sử dụng.
- Bao gồm chú thích (legends) và khóa (keys) cho biểu đồ pie, bar, histogram, stacked column/bar graph.
- Khi cần sử dụng các thùng (bin) cần phải có kích thước đồng đều trong biểu đồ (ví dụ: 10 đến 20, 20 đến 30, KHÔNG nên phân chia dạng 10 đến 15, 15 đến 25, 25 đến 40);
- Đối với dot plot và box plot, hãy đảm bảo bao gồm các giá trị tối thiểu và tối đa.

10.1.3. Phân loại biểu đồ

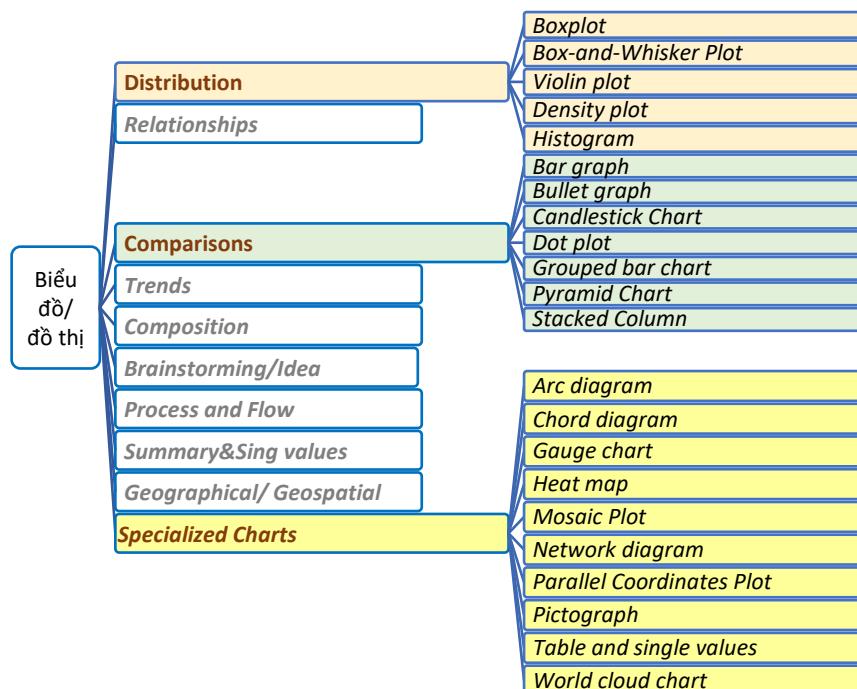
Trong trực quan hóa dữ liệu, biểu đồ có thể được phân loại thành nhiều loại dựa trên loại thông tin chúng truyền tải và cách chúng thể hiện dữ liệu.

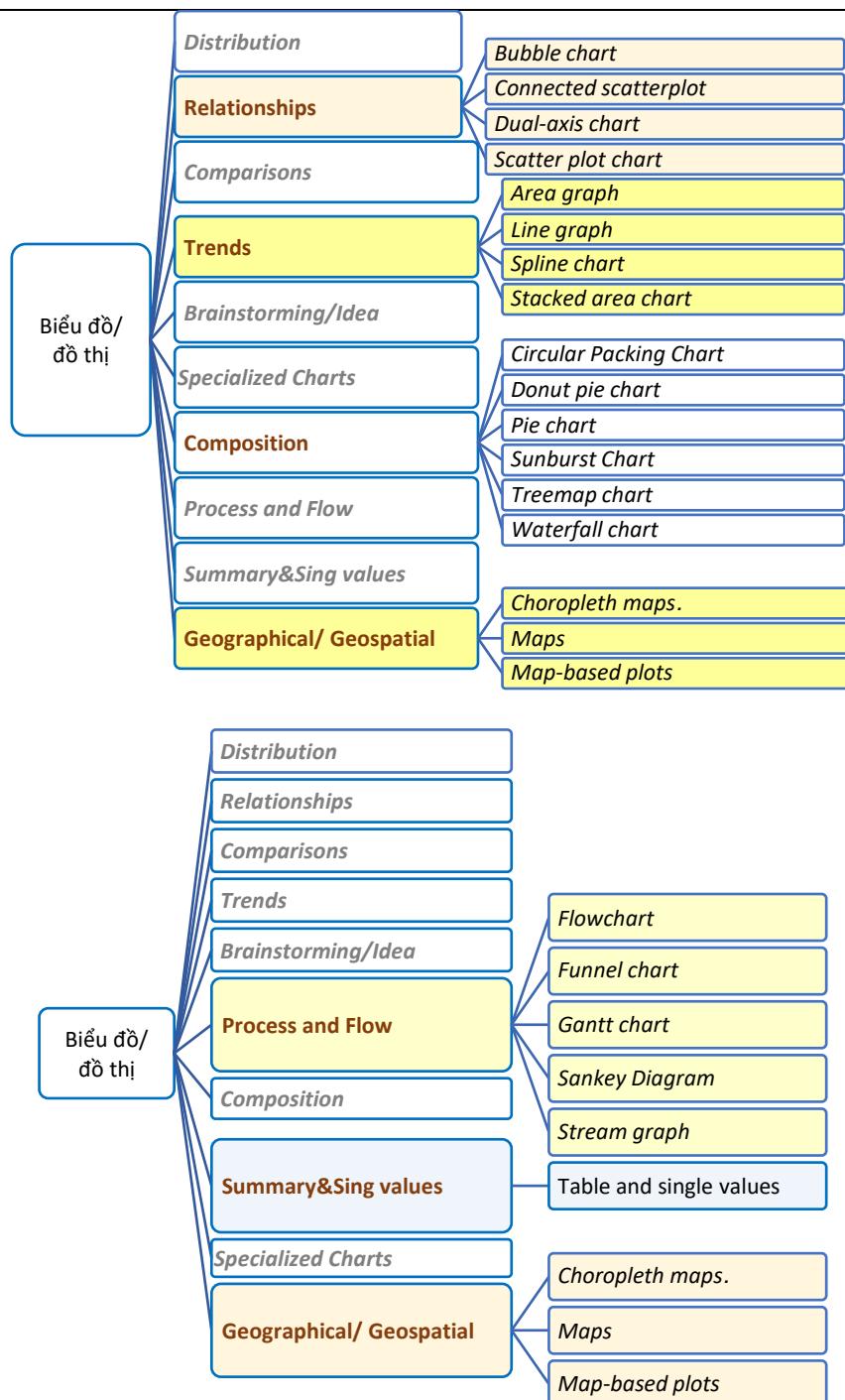
¹ <https://www.indeed.com/career-advice/career-development/types-of-graphs-and-charts>

Các danh mục này giúp chọn loại biểu đồ phù hợp tùy thuộc vào câu chuyện bạn muốn kể bằng dữ liệu của mình.



Hình 10-1. Phân loại đồ thị/biểu đồ theo công dụng/chức năng





10.2. CÁCH CHỌN ĐỒ THỊ/BIỂU ĐỒ PHÙ HỢP

Trực quan hóa dữ liệu là một thành phần quan trọng của phân tích dữ liệu vì chúng có khả năng tóm tắt lượng lớn dữ liệu một cách hiệu quả ở định dạng đồ họa. Có nhiều loại biểu đồ có sẵn, mỗi loại có điểm mạnh và trường hợp sử dụng riêng. Một trong những phần khó nhất của quá trình phân tích là chọn đúng cách để thể hiện dữ liệu bằng một trong những hình ảnh trực quan này.

Phần này sẽ tiếp cận việc chọn trực quan hóa dữ liệu dựa trên loại nhiệm vụ cần thực hiện. Có thể chia các nhiệm vụ thành 10 nhóm như sau:

- (i). ***Distribution Charts***: hiển thị sự phân bố của các điểm dữ liệu.
- (ii). ***Relationship Charts***: cho thấy các biến khác nhau có liên quan với nhau như thế nào.

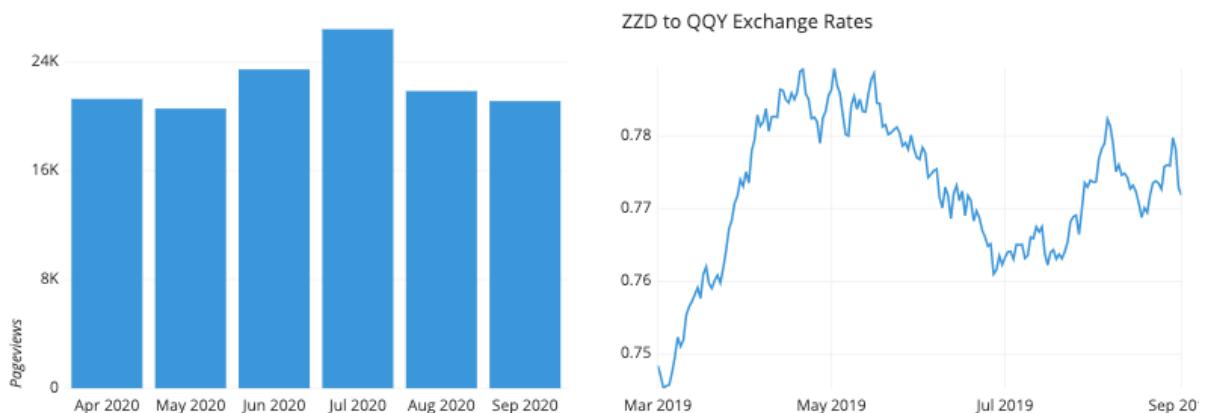
- (iii). *Comparison Charts*: so sánh các tập dữ liệu hoặc danh mục khác nhau.
- (iv). *Trend Charts*: hiển thị những thay đổi theo thời gian.
- (v). *Composition Charts*: hiển thị các phần của tổng thể.
- (vi). *Process and Flow Charts*: hiển thị các bước trong một quy trình.
- (vii). *Specialized Charts*: các loại biểu đồ đặc đáo/phức tạp dành cho các trường hợp sử dụng cụ thể.
- (viii). *Brainstorming/Idea Charts*: sử dụng để sắp xếp và trình bày ý tưởng.
- (ix). *Summary and Single Values*: sử dụng cho các bản tóm tắt đơn giản và các điểm dữ liệu riêng lẻ.
- (x). *Geographical/ Geospatial charts*: trực quan hóa dữ liệu liên quan đến vị trí địa lý.

Ngoài ra, các loại biến đang phân tích và đổi tượng xem trực quan cũng có thể ảnh hưởng đến biểu đồ nào sẽ hoạt động tốt nhất trong từng vai trò. Một số hình ảnh trực quan nhất định cũng có thể được sử dụng cho nhiều mục đích tùy thuộc vào các yếu tố này.

Lưu ý: Việc chọn biểu đồ phù hợp cho công việc phụ thuộc vào loại biến số đang xem xét và những gì bạn muốn thu được từ chúng. Dưới đây chỉ là các gợi ý chung, có thể việc thoát ra khỏi các chế độ tiêu chuẩn sẽ giúp có thêm những hiểu biết sâu sắc hơn. Thử nghiệm không chỉ với các loại biểu đồ khác nhau mà còn cả cách mã hóa các biến trong mỗi biểu đồ. Cũng nên nhớ rằng bạn không bị giới hạn trong việc hiển thị mọi thứ chỉ trong một biểu đồ. Thông thường, tốt hơn là nên giữ từng biểu đồ đơn giản và rõ ràng nhất có thể, thay vào đó hãy sử dụng nhiều biểu đồ để so sánh, thể hiện xu hướng và thể hiện mối quan hệ giữa nhiều biến số.

10.2.1. Biểu đồ thể hiện sự thay đổi theo thời gian (Trend charts)

- Một trong những ứng dụng phổ biến nhất để trực quan hóa dữ liệu là xem sự thay đổi giá trị của một biến theo thời gian. Các biểu đồ này thường có thời gian trên trục hoành, di chuyển từ trái sang phải, với các biến giá trị quan tâm trên trục tung. Có nhiều cách mã hóa các giá trị này:



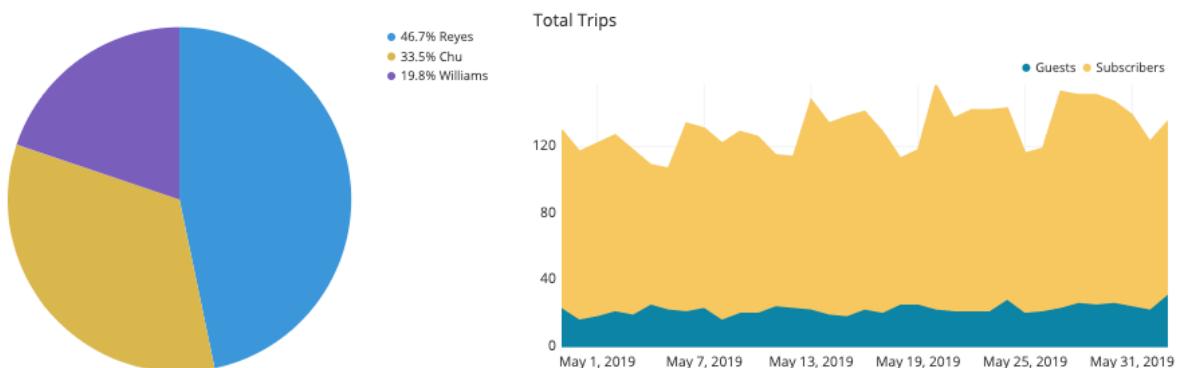
Hình 10-2. Sử dụng biểu đồ thể hiện sự thay đổi theo thời gian

- Các biểu đồ thường dùng trong nhóm này:
 - Bar charts: mã hóa giá trị theo chiều cao của thanh tính từ đường cơ sở.
 - Line charts: mã hóa giá trị theo vị trí dọc của các điểm được kết nối bằng các đoạn đường. Điều này hữu ích khi đường cơ sở không có ý nghĩa hoặc nếu số lượng thanh (bar) quá nhiều để vẽ đồ thị.

- Box plot: có thể hữu ích khi cần vẽ biểu đồ phân bố các giá trị cho từng khoảng thời gian; mỗi bộ hộp (set of box) và râu (whiskers) có thể hiển thị vị trí của các giá trị dữ liệu phổ biến nhất.
- Có một số loại biểu đồ chuyên biệt dành cho lĩnh vực tài chính, như candlestick hoặc Kagi chart.

10.2.2. Biểu đồ để hiển thị từng thành phần trong tổng thể (Component charts)

- Đôi khi, không chỉ cần biết tổng số mà còn cần biết các thành phần tạo nên tổng đó. Trong khi các biểu đồ khác như Bar chart tiêu chuẩn có thể được sử dụng để so sánh giá trị của các thành phần, các biểu đồ sau đây đặt sự phân tách từng phần thành toàn bộ lên hàng đầu:



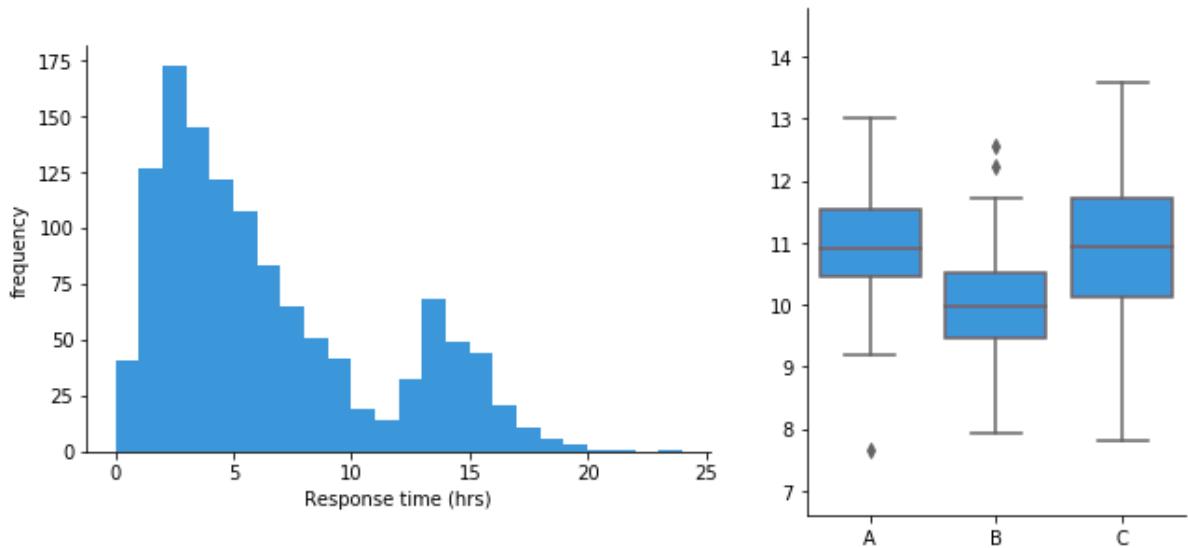
Hình 10-3. Biểu đồ để hiển thị từng thành phần trong tổng thể

- Các biểu đồ thường dùng trong nhóm này:
 - Pie chart và donut chart đều thể hiện tổng thể bằng một vòng tròn, được phân chia thành các phần thành đã tạo nên vòng tròn (100%).
 - Stacked bar chart: sửa đổi bar chart bằng cách chia mỗi thanh chính ra thành nhiều thành phần con.
 - Tương tự, stacked area chart: sửa đổi Line chart bằng cách sử dụng các bóng phía dưới đường để chia tổng thành các giá trị nhóm phụ.
 - Một loạt các loại biểu đồ phức tạp hơn khác cũng đã được phát triển để thể hiện mối quan hệ phân cấp. Chúng bao gồm Marimekko plot và treemap.

10.2.3. Biểu đồ để xem dữ liệu được phân phối như thế nào (Distribution Charts)

- Một cách sử dụng quan trọng của trực quan hóa là hiển thị cách phân phối giá trị của điểm dữ liệu. Điều này đặc biệt hữu ích trong quá trình khám phá, khi cố gắng xây dựng sự hiểu biết về các thuộc tính của các đối tượng dữ liệu.
- Các biểu đồ thường dùng trong nhóm này:
 - Bar chart được sử dụng khi một biến có tính chất định tính và nhận một số giá trị riêng biệt.
 - Histogram được sử dụng khi một biến mang tính định lượng, lấy giá trị số.
 - Density curve có thể được sử dụng thay cho histogram, như một ước tính được làm mịn của phân bố cơ bản.
 - Violin plot so sánh sự phân bố giá trị số giữa các nhóm bằng cách vẽ density curve cho mỗi nhóm.

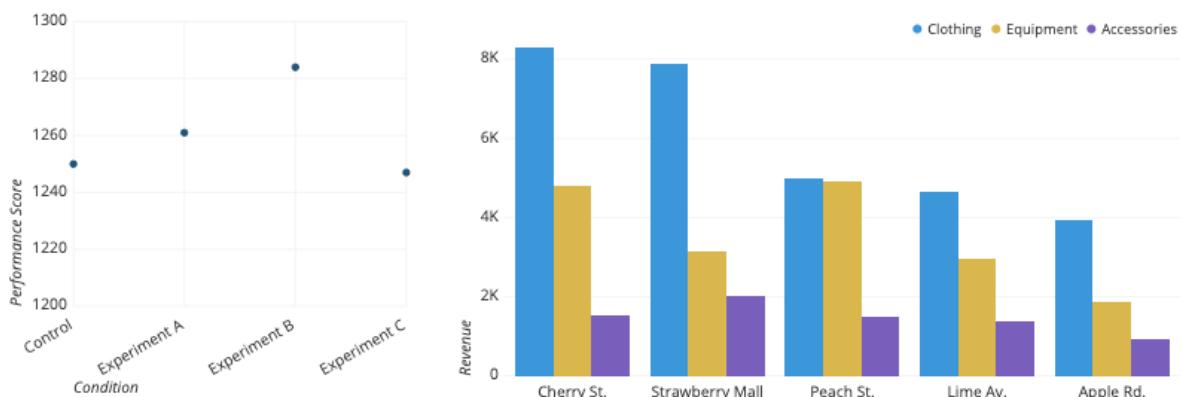
- Box plot là một cách khác để so sánh sự phân bố giữa các nhóm, nhưng với một bản tóm tắt số liệu các thống kê (min, tứ phân vị thứ 1, median, tứ phân vị thứ 3, max) thay vì biểu diễn hình dạng phân phối ước tính.



Hình 10-4. Biểu đồ để xem dữ liệu được phân phối như thế nào

10.2.4. Biểu đồ so sánh giá trị giữa các nhóm (Comparison Charts)

- Một ứng dụng rất phổ biến khác để trực quan hóa dữ liệu là so sánh các giá trị giữa các nhóm riêng biệt. Điều này thường được kết hợp với các vai trò khác để trực quan hóa dữ liệu, như hiển thị sự thay đổi theo thời gian hoặc xem cách phân phối dữ liệu.

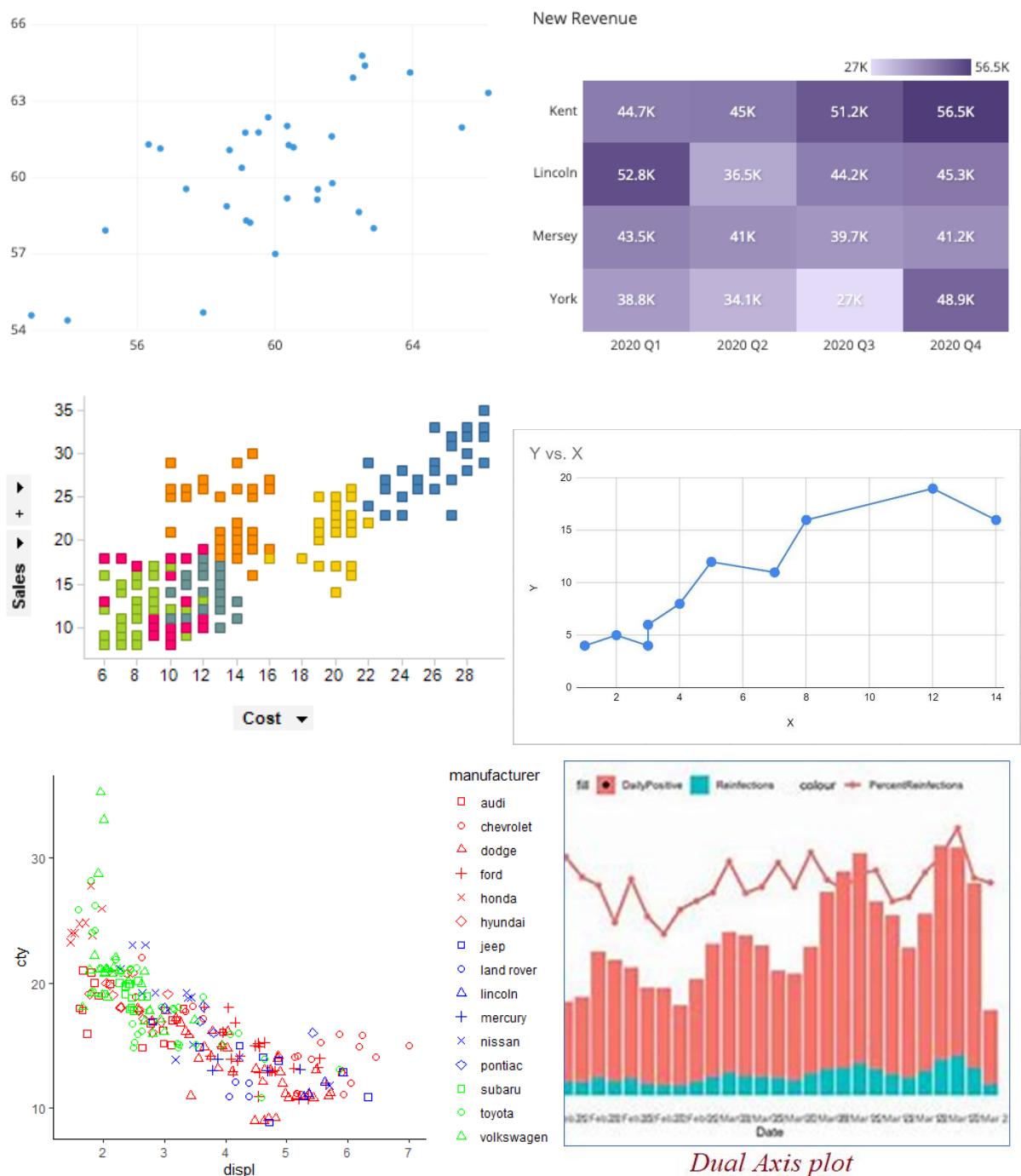


Hình 10-5. Biểu đồ so sánh giá trị giữa các nhóm

- Các biểu đồ thường dùng trong nhóm này:
 - Bar chart so sánh các giá trị giữa các nhóm bằng cách gán một thanh cho mỗi nhóm.
 - Dot plot có thể được sử dụng tương tự, ngoại trừ giá trị được biểu thị bằng vị trí điểm thay vì độ dài thanh. Điều này giống như một Line chart với các đoạn đường bị loại bỏ đi ‘sự kết nối’ giữa các điểm tuần tự. Cũng giống như Line chart, Dot plot rất hữu ích khi việc nối các đường kẻ giữa các điểm không có ý nghĩa.
 - Line chart có thể được sử dụng để so sánh giá trị giữa các nhóm theo thời gian bằng cách vẽ một đường cho mỗi nhóm.
 - Grouped Bar chart cho phép so sánh dữ liệu giữa hai biến nhau bằng cách vẽ nhiều thanh ở mỗi vị trí chứ không chỉ một.

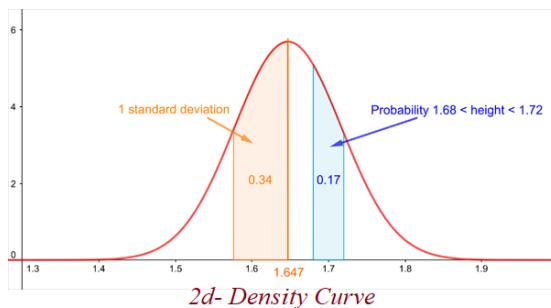
- Violin plot và box plot được sử dụng để so sánh sự phân bố dữ liệu giữa các nhóm.
- Funnel chart là biểu đồ chuyên dụng để hiển thị số lượng di chuyển như thế nào trong một quy trình, như theo dõi số lượng khách truy cập nhận được từ khi xem quảng cáo cho đến cuối cùng là mua hàng.
- Bullets chart là một biểu đồ chuyên dụng khác để so sánh giá trị thực với một hoặc nhiều điểm chuẩn.
- Một loại biểu đồ phụ xuất phát từ việc so sánh các giá trị giữa các nhóm đối với nhiều thuộc tính. Ví dụ về các biểu đồ này bao gồm parallel coordinates plot (và trường hợp đặc biệt của nó là slop plot) và dumbbell plot.

10.2.5. Biểu đồ để quan sát mối quan hệ giữa các biến (Relationship Charts)



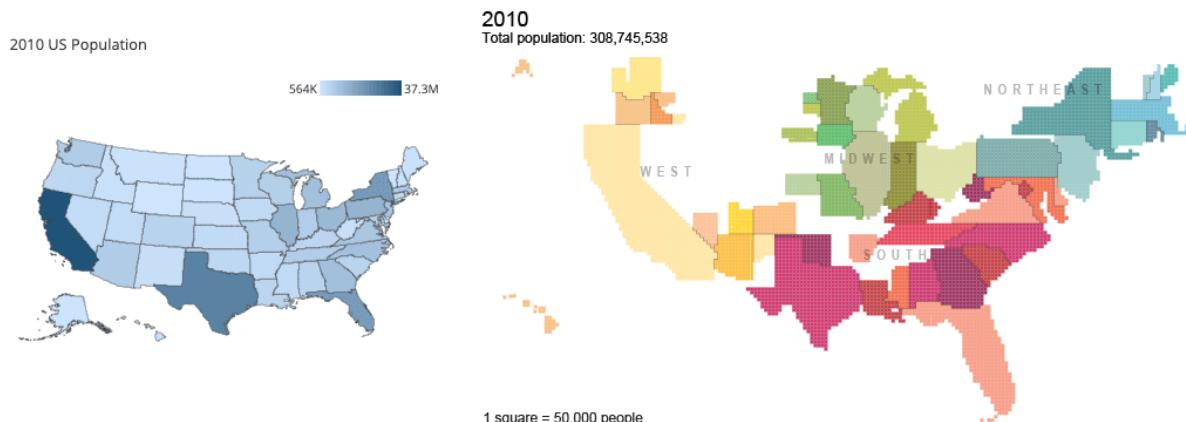
Hình 10-6. Nhóm các biểu đồ quan sát mối quan hệ giữa các biến

- Một nhiệm vụ khác xuất hiện trong quá trình khám phá dữ liệu là tìm hiểu mối quan hệ giữa các tính năng của dữ liệu. Các loại biểu đồ bên dưới có thể được sử dụng để vẽ hai hoặc nhiều biến số với nhau nhằm quan sát các xu hướng và mô hình giữa chúng.
- Các biểu đồ thường dùng trong nhóm này:
 - Scatter plot:
 - Là cách tiêu chuẩn để thể hiện mối quan hệ giữa hai biến.
 - Cũng có thể mở rộng Scatter plot thành các biến bổ sung bằng cách thêm màu sắc, hình dạng hoặc kích thước cho từng điểm làm chỉ báo, như trong bubble chart.
 - Khi biến thứ ba biểu thị thời gian, các điểm trong Scatter plot có thể được kết nối bằng các đoạn thẳng, tạo ra Scatter plot được kết nối.
 - Một lựa chọn thay thế khác cho biến thứ ba theo thời gian là dual axis plot (biểu đồ trực kép), chẳng hạn như vẽ Line chart và Bar chart có trục ngang chung.
 - Heatmap (bản đồ nhiệt):
 - Có thể hiển thị mối quan hệ giữa các nhóm khi một hoặc cả hai biến được so sánh không phải là số.
 - Heatmap cũng có thể được sử dụng cho dữ liệu số thuần túy, như trong biểu đồ 2-d histogram hoặc 2-d density curve (đường cong mật độ 2 chiều).



10.2.6. Biểu đồ để xem dữ liệu địa lý (Geographical/ Geospatial charts)

- Đôi khi, dữ liệu bao gồm dữ liệu địa lý như vĩ độ và kinh độ hoặc các khu vực như quốc gia hoặc tiểu bang. Mặc dù việc vẽ biểu đồ dữ liệu này có thể chỉ là mở rộng hình ảnh trực quan hiện có lên nền bản đồ (ví dụ: vẽ các điểm giống như Scatter plot trên đầu bản đồ), nhưng có các loại biểu đồ khác có tính đến miền ánh xạ. Hai trong số này được nêu bật dưới đây:



Hình 10-7. Nhóm các biểu đồ để xem dữ liệu địa lý

- Các biểu đồ thường dùng trong nhóm này:
 - Choropleth giống như một bản đồ nhiệt tô màu theo các khu vực địa chính trị thay vì một mạng lưới nghiêm ngặt.
 - Cartograms có cách tiếp cận khác bằng cách sử dụng kích thước của từng vùng để mã hóa giá trị. Cách tiếp cận này đòi hỏi một số biến dạng về hình dạng và cấu trúc liên kết.

10.2.7. Biểu đồ về quy trình (Process and Flow Charts)

Process and flow charts được sử dụng trong nhiều tình huống khác nhau để trực quan hóa và cải thiện quy trình và quy trình làm việc. Dưới đây là một số cách sử dụng phổ biến:

- (i).- Hiểu các quy trình (*Understanding Processes*): phát triển sự hiểu biết rõ ràng về cách thực hiện một quy trình.
- (ii).- Cải tiến quy trình (*Process Improvement*): Bằng cách nghiên cứu các bước liên quan, có thể xác định những điểm kém hiệu quả và các lĩnh vực cần cải tiến.
- (iii).- Giao tiếp (*Communication*): rất hữu ích trong việc truyền đạt các quy trình với người khác, đảm bảo mọi người tham gia đều hiểu được quy trình làm việc.
- (iv).- Tài liệu hóa (*Documentation*): giúp duy trì tính nhất quán và đóng vai trò tham khảo.
- (v).- Lập kế hoạch dự án (*Project Planning*): để phác thảo các bước, mốc thời gian và nguồn lực cần thiết.

Flowcharts đặc biệt hữu ích trong việc trực quan hóa các điểm quyết định và trình tự các bước trong một quy trình, trong khi sơ đồ quy trình có thể cung cấp thông tin chi tiết hơn, bao gồm đầu vào, đầu ra và mốc thời gian.

Có nhiều loại biểu đồ thuộc nhóm Process và Flow Charts, mỗi loại có mục đích và cách sử dụng riêng. Dưới đây là một số loại phổ biến:

1. **Flowchart:** Biểu đồ này mô tả các bước của một quy trình hoặc hệ thống bằng các hình dạng và mũi tên.
2. **BPMN (Business Process Model and Notation):** Đây là một tiêu chuẩn để mô hình hóa các quy trình kinh doanh phức tạp với nhiều ký hiệu chuyên dụng.
3. **Activity Diagram:** Thường được sử dụng trong UML (Unified Modeling Language) để mô tả luồng công việc hoặc hoạt động trong một hệ thống.
4. **Workflow Diagram:** Biểu đồ này thể hiện luồng công việc giữa các phòng ban hoặc bộ phận trong một tổ chức.
5. **State Diagram:** Biểu đồ này mô tả các trạng thái khác nhau của một thực thể và cách chúng chuyển đổi từ trạng thái này sang trạng thái khác.

Mỗi loại biểu đồ có cách sử dụng và ký hiệu riêng, phù hợp với các nhu cầu khác nhau trong việc mô hình hóa và cải thiện quy trình.

10.2.8. Biểu đồ dùng cho các chuyên ngành (Specialized Charts)

Specialized charts được sử dụng để cung cấp những hiểu biết sâu sắc hơn và trực quan hóa cụ thể hơn cho các tập dữ liệu phức tạp. Các biểu đồ này giúp làm cho dữ liệu phức tạp trở nên dễ hiểu và dễ thực hiện hơn bằng cách làm nổi bật các mẫu và mối quan hệ chính.

Dưới đây là một số loại phổ biến và công dụng của chúng:

- (i).- **Gantt Chart**: lập kế hoạch và quản lý dự án, thể hiện các nhiệm vụ theo thời gian.
- (ii).- **Waterfall Chart**: trực quan hóa dữ liệu tuần tự, chẳng hạn như báo cáo tài chính, để cho thấy giá trị ban đầu bị ảnh hưởng như thế nào bởi một loạt các giá trị dương hoặc âm trung gian.
- (iii).- **Radar Chart**: so sánh nhiều biến và hiển thị số liệu hiệu suất giữa các danh mục khác nhau.
- (iv).- **Tree Map**: Trực quan hóa dữ liệu phân cấp bằng cách sử dụng các hình chữ nhật lồng nhau, hữu ích để hiển thị tỷ lệ trong hệ thống phân cấp.
- (v).- **Funnel Chart**: Thường được sử dụng trong bán hàng và tiếp thị để thể hiện các giai đoạn trong một quy trình và xác định các điểm bỗn đờ tiềm năng.

10.2.9. Biểu đồ dùng cho ý tưởng sáng tạo (Brainstorming/Idea Charts)

- Brainstorming and idea charts là những công cụ mạnh mẽ để tạo ra và sắp xếp ý tưởng. Dưới đây là một số cách sử dụng phổ biến:
 - (i).- Tạo ý tưởng (*Idea Generation*): giúp tạo ra số lượng lớn ý tưởng một cách nhanh chóng, khuyến khích sự sáng tạo và tư duy tự do.
 - (ii).- Giải quyết vấn đề (*Problem Solving*): Những biểu đồ này có thể chia nhỏ các vấn đề phức tạp thành các phần quản lý, giúp tìm giải pháp dễ dàng hơn.
 - (iii).- Lập kế hoạch dự án (*Project Planning*): hỗ trợ phác thảo các yêu cầu, nhiệm vụ và tiến độ của dự án, đảm bảo tất cả các khía cạnh đều được đề cập.
 - (iv).- Ra quyết định (*Decision Making*): Bằng cách trực quan hóa các lựa chọn khác nhau và kết quả tiềm năng của chúng, những biểu đồ này hỗ trợ đưa ra quyết định sáng suốt.
 - (v).- Cộng tác nhóm (*Team Collaboration*): cung cấp không gian trực quan chung để các thành viên trong nhóm đóng góp ý tưởng, thúc đẩy sự hợp tác và giải quyết vấn đề tập thể.
- Một số loại *Brainstorming and idea charts* phổ biến bao gồm:
 - *Mind Maps* (Sơ đồ tư duy): Bắt đầu với ý tưởng trọng tâm và phân nhánh thành các chủ đề phụ liên quan.
 - *Flowcharts* (Lưu đồ): Phác thảo một loạt các bước hoặc quy trình từ đầu đến cuối.
 - *Bubble Maps* (Bản đồ bong bóng): Hữu ích cho việc lập kế hoạch giai đoạn đầu và sắp xếp các ý tưởng cốt lõi.
 - Starbursting: Tập trung vào khám phá ai, cái gì, ở đâu, khi nào, tại sao và như thế nào của ý tưởng trung tâm.

10.2.10. Biểu đồ với giá trị đơn (Single Values chart)

Sử dụng cho các bản tóm tắt đơn giản và các điểm dữ liệu riêng lẻ.



10.3. CHỌN MÀU SẮC ĐỂ TRỰC QUAN HÓA DỮ LIỆU

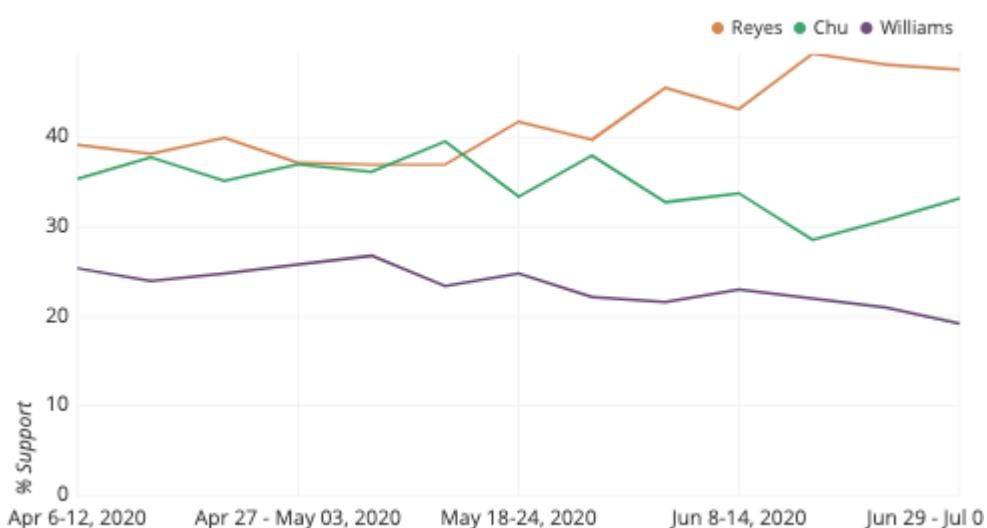
Lựa chọn màu sắc là yếu tố chính trong việc tạo ra biểu đồ hiệu quả. Một bộ màu sắc tốt sẽ làm nổi bật câu chuyện mà dữ liệu muốn kể, trong khi một bộ màu kém sẽ che giấu hoặc làm xao lâng mục đích của hình ảnh trực quan.

10.3.1. Các loại bảng màu

Có ba loại bảng màu chính để trực quan hóa dữ liệu:

- Bảng màu định tính (*Qualitative palettes*)
- Bảng màu tuần tự (*Sequential palettes*)
- Bảng màu phân kỳ (*Diverging palettes*)

Loại bảng màu sử dụng trong trực quan hóa phụ thuộc vào bản chất của dữ liệu được ánh xạ tới màu.



Hình 10-8. Bảng màu sử dụng cho đồ thị

10.3.1.1. Bảng màu định tính (*Qualitative palettes*)

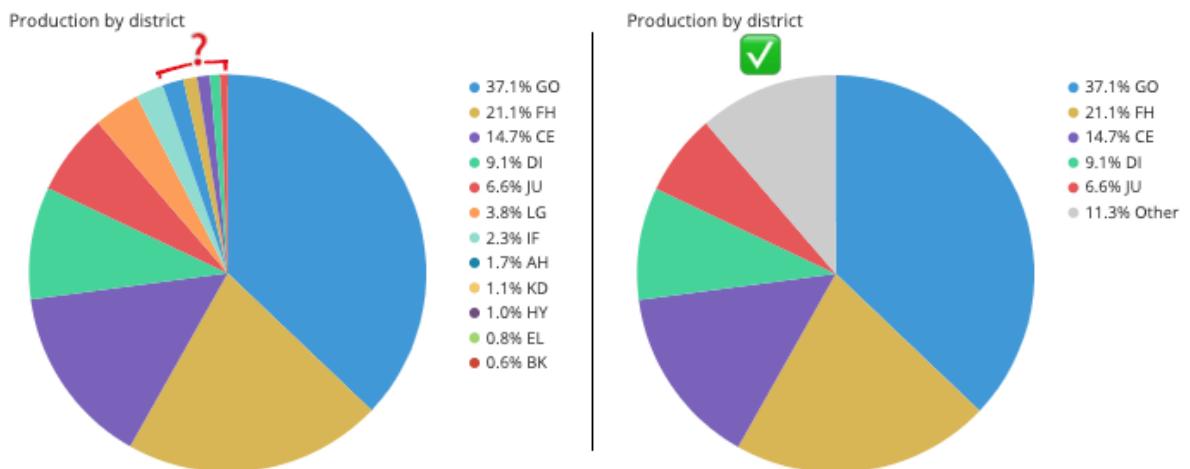
Bảng màu định tính được sử dụng khi biến có tính chất phân loại. Các biến phân loại là những biến mang nhãn riêng biệt (định danh) mà không có thứ tự cố hữu. Ví dụ bao gồm quốc gia hoặc tiểu bang, chủng tộc, giới tính của hàng, loại sản phẩm, Mỗi giá trị có thể có của biến được gán một màu từ bảng màu định tính.



Hình 10-9. Bảng màu định tính

Trong bảng màu định tính, màu sắc được gán cho mỗi nhóm cần phái khác biệt. Theo nguyên tắc chung, nên có gắng giới hạn kích thước bảng màu tối đa ở mức mười màu trở xuống. Với nhiều màu sắc hơn, sẽ gây khó khăn trong việc phân biệt giữa các nhóm. Nếu số lượng định danh nhiều hơn số lượng màu sắc định dùng, khi đó nên cố gắng gộp các giá trị lại với nhau, chẳng hạn như đặt các danh mục nhỏ nhất thành một danh mục “khác”.

Việc lặp lại các màu nhiều lần là một ý tưởng tồi vì điều này có thể gây nhầm lẫn.



Hình 10-10. Minh họa đồ thị dùng bảng màu định tính

Trong minh họa trên, những lát cắt nhỏ nhất bên trái không chỉ lặp lại các màu trong bảng màu mà còn khá khó phân biệt với nhau.

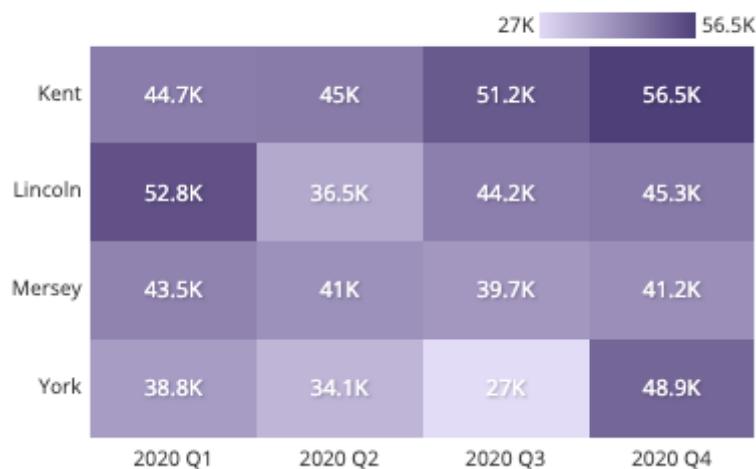
Cách chính để tạo ra sự khác biệt giữa các màu sắc là thông qua màu sắc của chúng. Có thể đạt được sự khác biệt bổ sung giữa các màu thông qua việc điều chỉnh độ sáng (lightness) và độ bão hòa (saturation), nhưng không nên tạo ra sự khác biệt quá lớn. Quá nhiều sự khác biệt có thể cho thấy rằng một số màu quan trọng hơn những màu khác – mặc dù đây có thể là một đặc tính hữu ích khi được sử dụng có chủ đích. Tránh sử dụng hai màu có cùng sắc độ nhưng có độ sáng và độ bão hòa khác nhau, trừ khi các giá trị liên quan đến những màu đó có liên quan với nhau. Ví dụ: có thể có Line chart với các số liệu hàng ngày có màu sáng và đường trung bình động hàng tuần có màu tối hơn.



Hình 10-11. Minh họa việc sử dụng sắc độ sáng của màu

10.3.1.2. Bảng màu tuần tự (Sequential palettes)

New Revenue



Hình 10-12. Minh họa việc sử dụng bảng màu tuần tự

Khi kiểu của dữ liệu được gán màu là kiểu số hoặc các giá trị được sắp xếp vốn có (kiểu thứ tự như xuất sắc, giỏi, khá, trung bình,...) khi đó có thể được mô tả bằng bảng màu tuần tự. Màu sắc được gán cho các giá trị dữ liệu một cách liên tục, thường dựa trên độ sáng, màu sắc hoặc cả hai.



Hình 10-13. Minh họa một dạng của bảng màu tuần tự

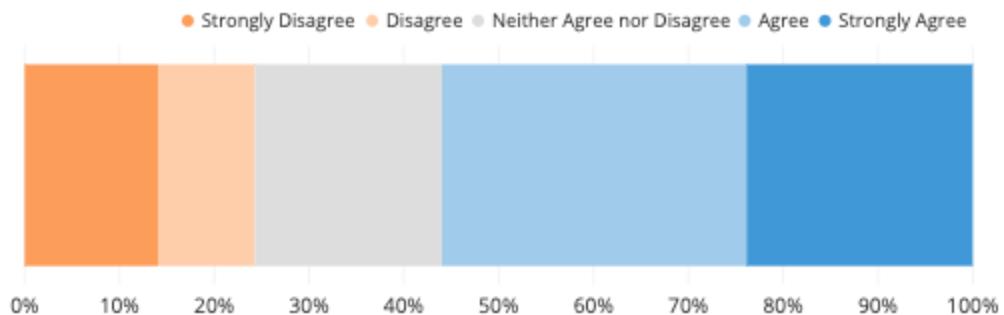
Kích thước nổi bật nhất của màu sắc đối với bảng màu tuần tự là độ sáng của nó. Thông thường, giá trị thấp hơn gắn liền với màu sáng hơn và giá trị cao hơn gắn liền với màu tối hơn. Tuy nhiên, điều này là do các ô có xu hướng nằm trên nền trắng hoặc nền sáng tương tự. Trên nền tối, thông thường có trường hợp ngược lại, trong đó các giá trị cao hơn được biểu thị bằng màu sáng hơn, nhạt hơn.



Kích thước phụ của bảng màu tuần tự là màu sắc của nó. Sẽ tốt hơn nếu chỉ sử dụng một màu duy nhất cho bản đồ màu của bạn, chủ yếu là thay đổi độ sáng để biểu thị giá trị. Tuy nhiên, đáng để cân nhắc việc trải rộng giữa hai màu như một sự trợ giúp bổ sung cho việc mã

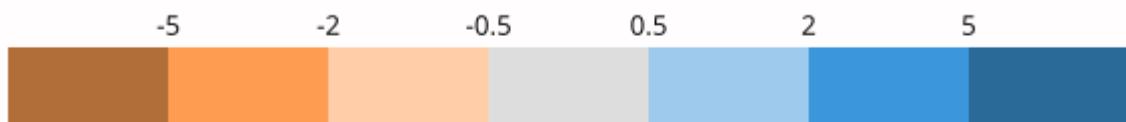
hóa. Thông thường, màu ám hơn (về phía đỏ hoặc vàng) sẽ ở đầu nhạt hơn, với màu lạnh hơn (về phía xanh lá cây, xanh lam hoặc tím) ở đầu tối hơn.

10.3.1.3. Bảng màu phân kỳ (Diverging palettes)



Hình 10-14. Minh họa bảng màu phân kỳ không có giá trị trung tâm

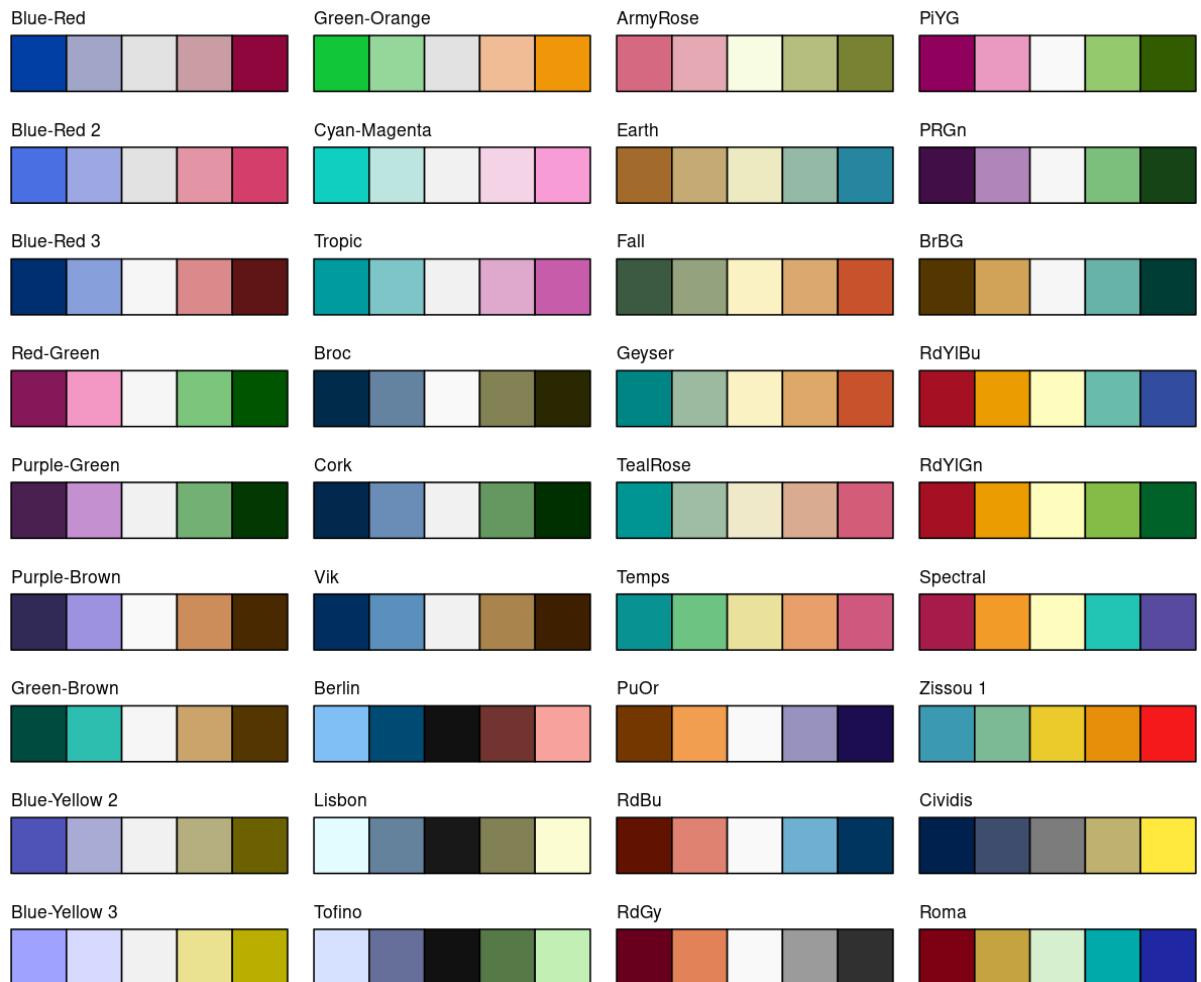
Nếu biến số có giá trị trung tâm có ý nghĩa, chẳng hạn như số 0, thì có thể áp dụng bảng phân kỳ. Bảng màu phân kỳ về cơ bản là sự kết hợp của hai bảng màu tuần tự với điểm cuối dùng chung nằm ở giá trị trung tâm. Các giá trị lớn hơn tâm được gán cho các màu ở một bên của tâm, trong khi các giá trị nhỏ hơn được gán cho các màu ở phía đối diện.



Hình 10-15. Minh họa bảng màu phân kỳ có giá trị trung tâm

Thông thường, một màu sắc đặc biệt được sử dụng cho từng bảng màu tuần tự thành phần để giúp dễ dàng phân biệt giữa các giá trị dương và âm so với tâm. Giống như các bảng màu tuần tự, giá trị trung tâm thường được gán một màu sáng, do đó các màu tối hơn biểu thị khoảng cách lớn hơn từ tâm.

Diverging (hcl.colors)



10.3.1.4. Bảng màu phân kỳ và bảng màu tuần tự

Các bảng màu tuần tự và phân kỳ có thể được liên kết với các giá trị dữ liệu theo hai cách khác nhau dưới dạng:

- (i). Một tập hợp màu phân kỳ (riêng biệt, rời rạc), mỗi màu được liên kết với một phạm vi số.
- (ii). Một tập hợp liên tục giữa giá trị số và màu.

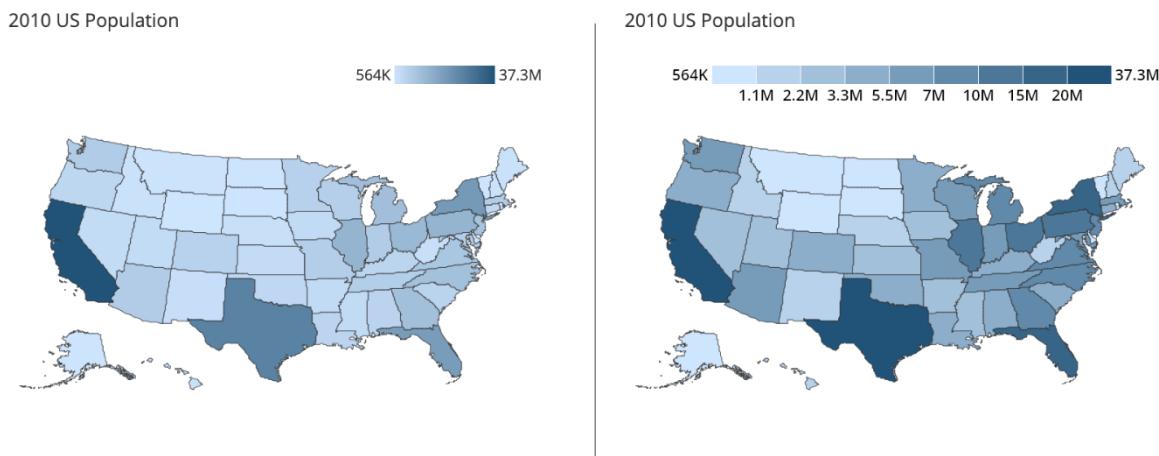


Hình 10-16. Minh họa bảng màu phân kỳ (trên) và bảng màu liên tục (dưới)

Các công cụ tạo bảng màu nói chung sẽ tuân theo bảng màu phân kỳ, trong khi các công cụ tạo trực quan hóa thường có khả năng xây dựng dựa trên bảng màu liên tục. Mặc dù việc có một hàm liên tục giữa giá trị và màu sắc có vẻ tự động tốt hơn, nhưng bảng màu phân kỳ vẫn có giá trị riêng của nó.

Khả năng phân biệt sự khác biệt về màu sắc của con người yếu hơn khả năng phân biệt vị trí hoặc chiều dài, vì vậy sẽ gặp bất lợi trong việc liên kết màu sắc với các giá trị chính xác.

- Việc phân kỳ hóa các giá trị có thể làm giảm tải nhận thức bằng cách đưa ra các mẫu rộng trong dữ liệu. Sử dụng bảng màu phân kỳ có nghĩa là ta có thể tạo các phạm vi có kích thước không đồng đều để thể hiện rõ hơn sự khác biệt trong dữ liệu. Ngoài ra, có thể đặt phạm vi giá trị cho một bảng màu riêng biệt theo cách thể hiện dữ liệu tốt hơn.
- Nếu dữ liệu bao gồm các giá trị ngoại lệ thì bảng màu liên tục có thể buộc hầu hết dữ liệu vào phạm vi giá trị hẹp hơn.



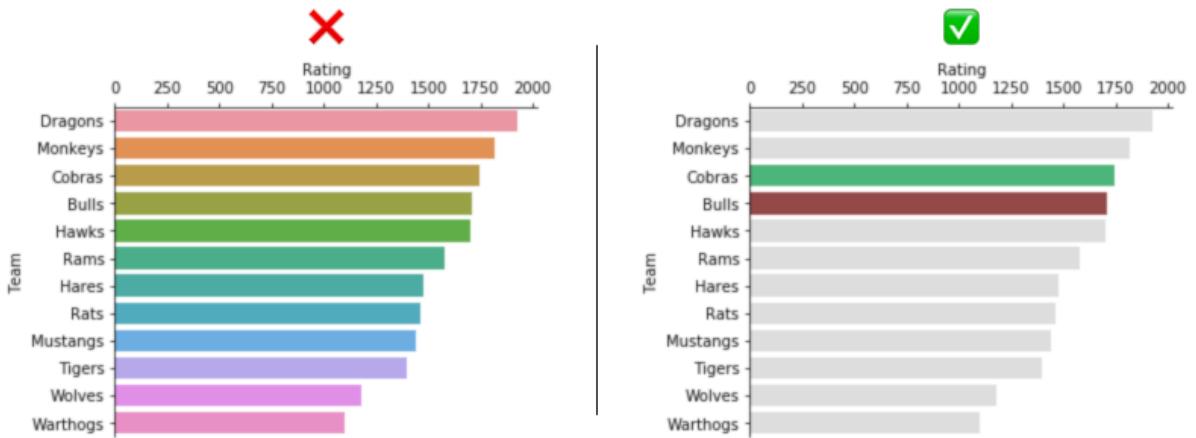
Hình 10-17. Bảng màu phân kỳ giúp dễ “lượng hóa” dữ liệu hơn khi dùng bảng màu liên tục

Một nhược điểm lớn của bảng màu phân kỳ là mất khả năng so sánh các phần tử nằm trong cùng một khoảng giá trị (cùng 1 bin, như từ 1,1M đến 2,2M). Trong trường hợp có một số khác biệt về màu sắc giữa các giá trị gần với bảng màu liên tục (chẳng hạn như giữa Texas và California trong ví dụ trên), thì không thấy được sự khác biệt giữa 2 bang này.

10.3.2. Một số gợi ý bổ sung đối với việc sử dụng màu

10.3.2.1. Tránh sử dụng màu sắc không cần thiết

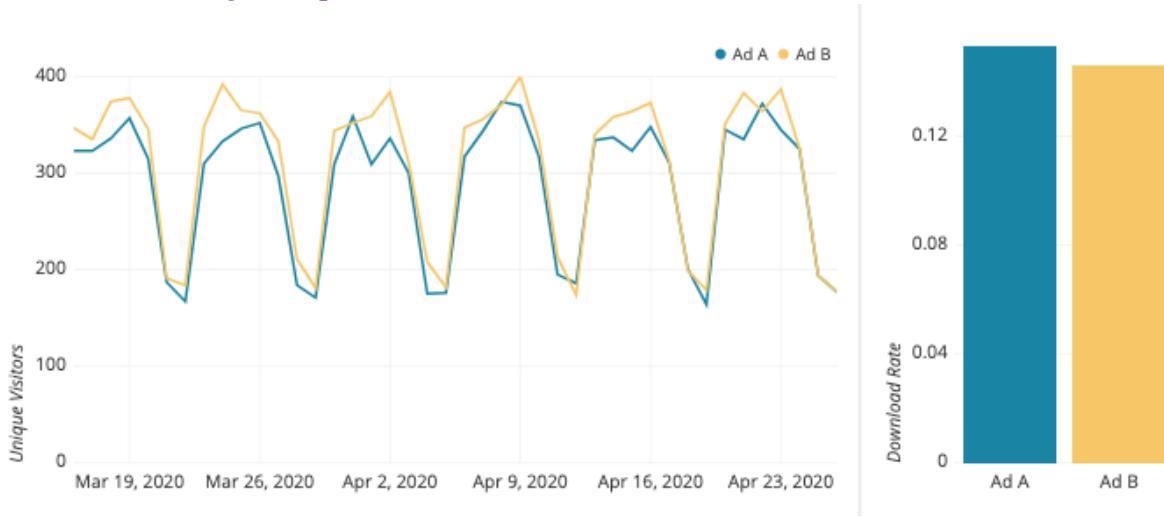
Mặc dù màu sắc là một phần quan trọng trong việc trực quan hóa dữ liệu nhưng cũng nên hạn chế và chỉ sử dụng màu sắc ở những nơi phù hợp. Không phải mọi biểu đồ tạo ra đều yêu cầu nhiều màu. Nếu bạn chỉ có hai biến để vẽ, chúng có thể sẽ được mã hóa theo vị trí hoặc độ dài dọc và ngang. Màu thường chỉ xuất hiện khi biến thứ ba cần được mã hóa thành biểu đồ hoặc nếu đó là một thành phần của biểu đồ chuyên dụng như Pie chart. Tuy nhiên, có những trường hợp màu sắc có thể được thêm vào để nhấn mạnh một phát hiện cụ thể hoặc như một cách mã hóa làm nổi bật thêm.



Hình 10-18. Tránh sử dụng quá nhiều màu nếu không cần thiết

Trong minh họa trên, việc sử dụng màu của thanh cầu vòng (rainbow bar) ở bên trái không có ý nghĩa và nên tránh. Ở bên phải, hầu hết các thanh có màu xám trung tính để làm nổi bật sự so sánh giữa hai thanh màu.

10.3.2.2. Sử dụng nhất quán màu sắc cho các biểu đồ

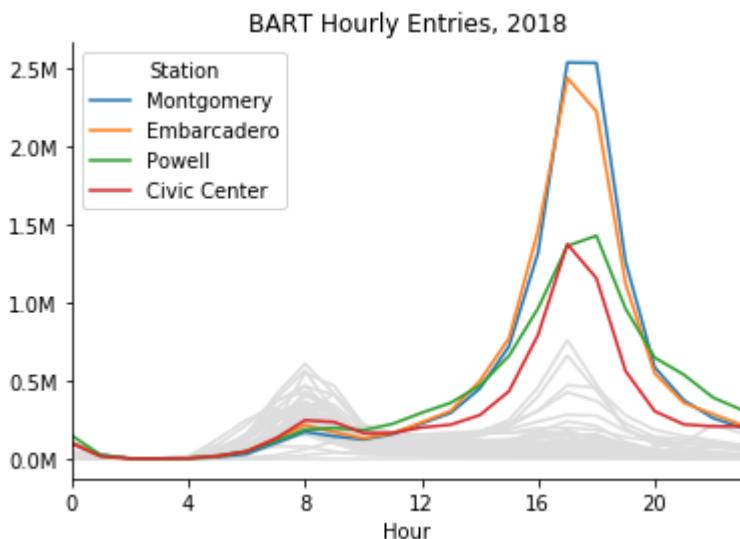


Hình 10-19. Sử dụng nhất quán màu sắc cho các biểu đồ

Nếu có trang tổng quan hoặc báo cáo bao gồm nhiều biểu đồ, nên sử dụng nhất quán màu giữa các biểu đồ khi chúng tham chiếu đến cùng một nhóm hoặc thực thể. Nếu màu sắc thay đổi ý nghĩa giữa các biểu đồ, điều này có thể khiến người đọc khó hiểu biểu đồ hơn.

10.3.2.3. Tận dụng ý nghĩa của màu sắc

- Đôi khi, có thể tận dụng cách cảm nhận màu sắc để nâng cao hiệu quả trực quan hóa. Nếu các nhóm đang vẽ đồ thị có những quy ước cố hữu về màu sắc, chẳng hạn như với các đội thể thao và các đảng chính trị, thì việc chỉ định màu thích hợp có thể giúp người đọc dễ dàng theo dõi hình ảnh hơn. Thậm chí có thể thử tạo các bảng màu tùy chỉnh xung quanh màu sắc thương hiệu vừa nêu.
- Nguyên tắc chung là tránh mức độ bão hòa màu và độ sáng quá cao để giảm mỏi mắt. Điều này cũng cho phép có chỗ để làm nổi bật các yếu tố quan trọng bằng cách tạo cho chúng một cái nhìn táo bạo hơn so với các yếu tố khác. Tương tự, không thể đánh giá thấp tầm quan trọng của màu xám khi đặt những dữ liệu không quan trọng ở chế độ nền, cùng với các mục đích khác.

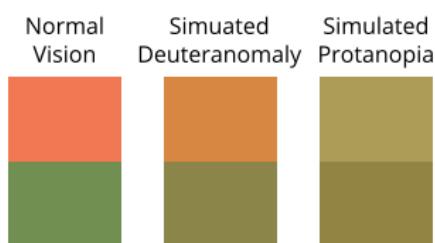


Hình 10-20. Sử dụng màu xám ở chế độ nền

- Ngoài ra, cần lưu ý rằng các nền văn hóa khác nhau có thể liên kết những ý nghĩa khác nhau với mỗi màu sắc. Ví dụ, màu đỏ có thể gắn liền với niềm đam mê hoặc sự nguy hiểm trong một số nền văn hóa phương Tây, nhưng sự thịnh vượng và may mắn ở một số nền văn hóa phương Đông. Điều này có thể không đặc biệt quan trọng trừ khi những phát hiện được trình bày cho nhiều đối tượng, nhưng đó là một công cụ khác cần ghi nhớ để giúp dễ dàng nắm bắt hình ảnh trực quan hơn.

10.3.2.4. Trực quan hóa và mù màu (color blindness)

Khoảng 4% dân số (hầu hết là nam giới) mắc một dạng mù màu nào đó. Các dạng mù màu phổ biến nhất gây nhầm lẫn giữa một số sắc thái nhất định của màu đỏ và xanh lá cây, mặc dù cũng có những dạng mù màu khiến các sắc thái màu xanh lam và vàng trông giống nhau. Vì những lý do này, nên thử và thay đổi một kích thước khác ngoài màu sắc để biểu thị giá trị liên quan đến màu sắc, như độ sáng và độ bão hòa. Cũng có thể sử dụng trình mô phỏng mù màu như Coblis để biết liệu hình ảnh cuối cùng của đồ thị đang xây dựng có dễ hiểu đối với người khác hay không và liệu có khả năng xảy ra sự mơ hồ hay không.

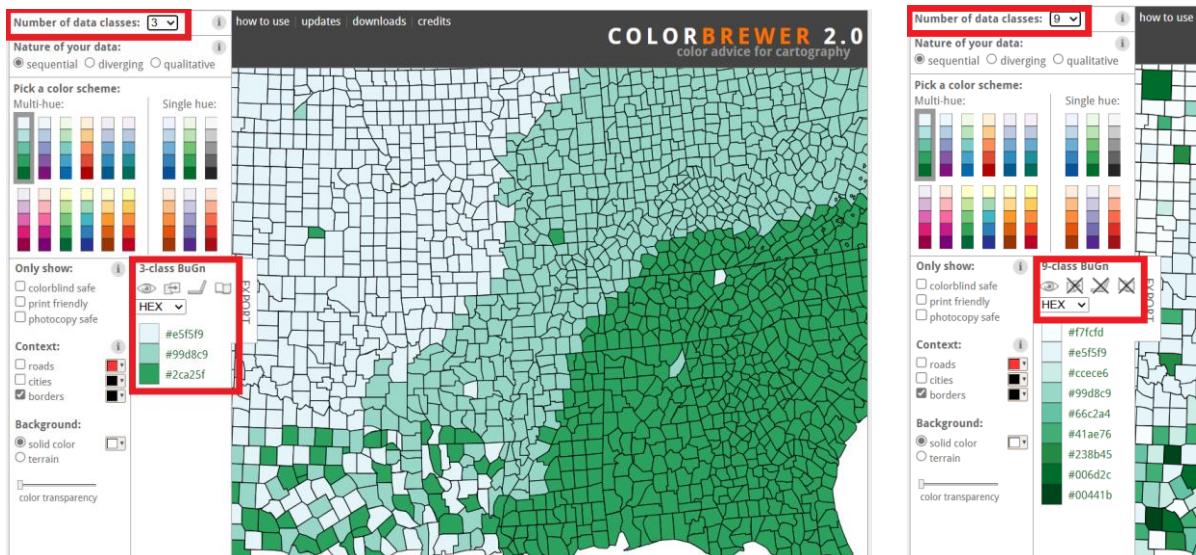


Hình 10-21. Trực quan hóa và mù màu (color blindness)

10.3.3. Một số công cụ trực tuyến hỗ trợ việc dùng màu sắc

Có rất nhiều công cụ trực tuyến giúp chọn và kiểm tra màu sắc cho việc trực quan hóa dữ liệu. Ở đây, chỉ nêu một số công cụ đơn giản để giúp việc lựa chọn màu sắc.

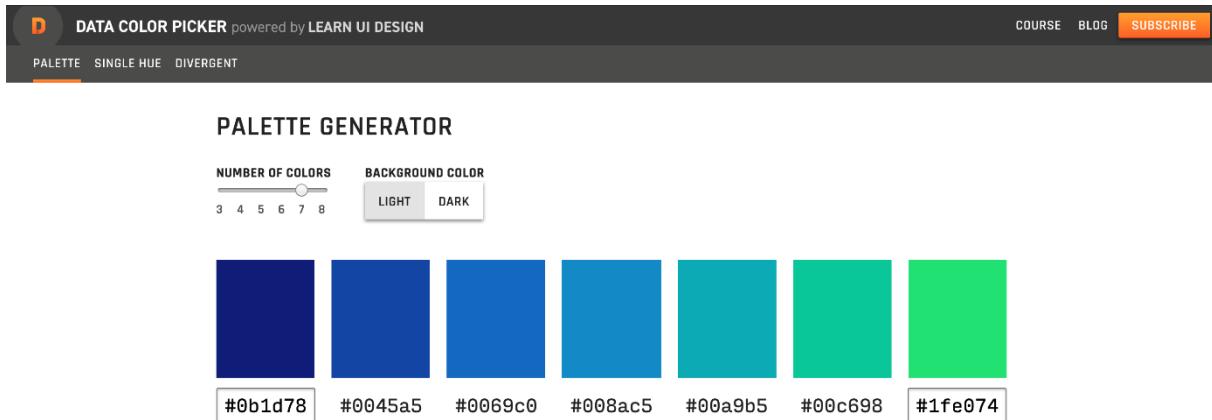
10.3.3.1. ColorBrewer¹



Hình 10-22. T ColorBrewer

ColorBrewer là tài liệu tham khảo cổ điển về bảng màu và cung cấp một số bảng màu khác nhau cho từng loại. Một số bảng màu có thể bị nghi ngờ về độ an toàn của mù màu, vì vậy hãy nhớ kiểm tra biểu tượng con mắt phía trên ngăn mă màu để kiểm tra xem bộ màu có tiềm ẩn nguy cơ cao gấp khó khăn về nhận thức hay không (được biểu thị bằng dấu ? và X tương ứng).

10.3.3.2. Data Color Picker



Hình 10-23. Data Color Picker

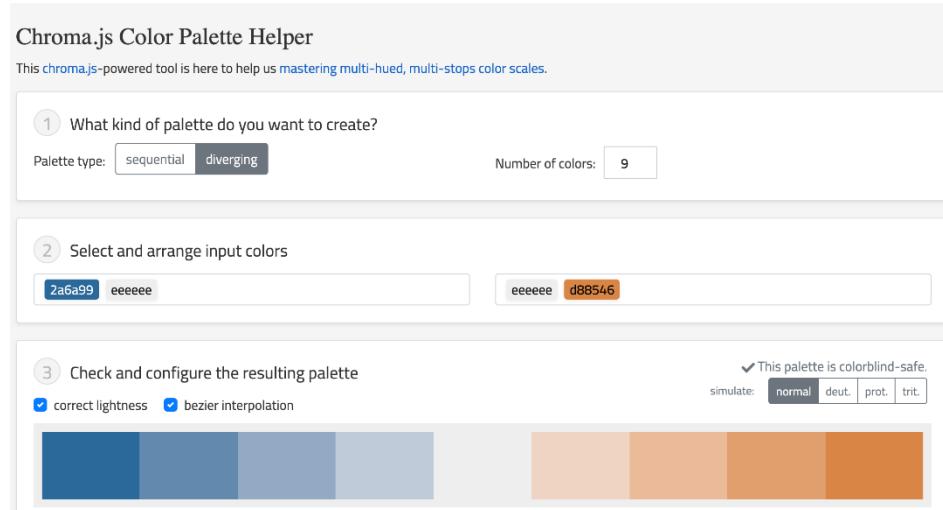


Hình 10-24. The color wheel trong Data Color Picker

¹ [ColorBrewer: Color Advice for Maps \(colorbrewer2.org\)](http://colorbrewer2.org)

Data Color Picker là một công cụ nhanh chóng và dễ sử dụng để tạo các bảng màu tuần tự và phân kỳ. Tab “Palette” mặc định được sử dụng tốt nhất để tạo các bảng màu tuần tự nhiều màu thay vì các bảng màu định tính, vì phép nội suy giữa các điểm cuối nhất thiết sẽ loại bỏ một số phân đoạn màu sắc trong bánh xe màu (the color wheel).

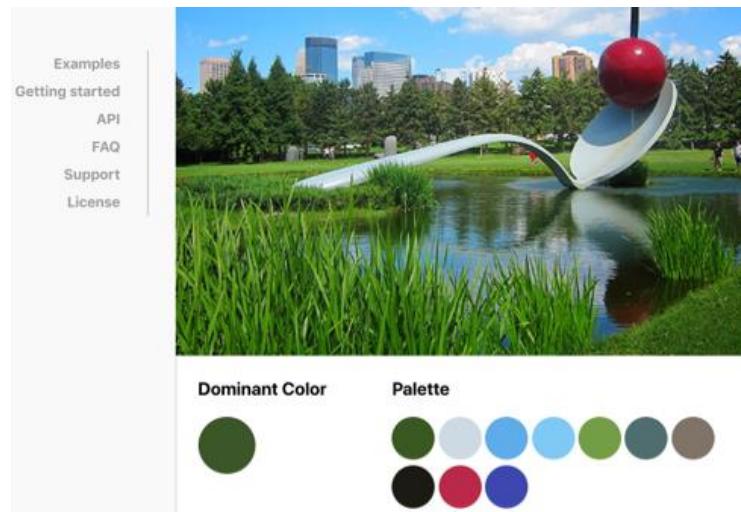
10.3.3.3. Chroma.js Color Palette Helper



Hình 10-25. Chroma.js Color Palette Helper

Trình chroma.js Color Palette Helper phức tạp hơn một chút so với Data Color Picker với các tùy chọn để điều chỉnh độ sáng, sử dụng phép nội suy bezier và việc nhập các giá trị màu khó hơn một chút. Tuy nhiên, nó cũng cho phép có thêm một số quyền tự do trong việc thiết lập nhiều điểm dừng để thuật toán thử và điều chỉnh bảng màu cho phù hợp. Là một phần thường bổ sung, ứng dụng này còn bao gồm trình mô phỏng mù màu trên cùng một trang, nêu bật các loại thiếu sót phổ biến nhất mà các vấn đề có thể phát sinh.

10.3.3.4. Color Thief

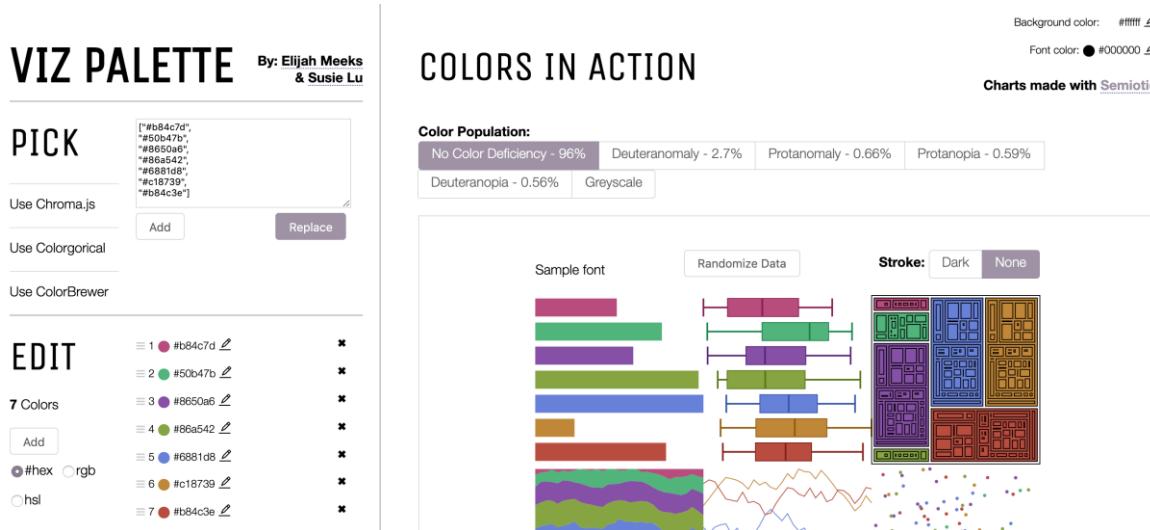


Hình 10-26. Color Thief

Một cách khác để tạo bảng màu chất lượng là lấy cảm hứng từ hình ảnh và ảnh chụp màn hình với bảng màu tự nhiên hấp dẫn. Hiện có một số công cụ giúp thực hiện việc này, nhưng Color Thief là một trong những công cụ dễ sử dụng nhất, tự động trích xuất bảng màu có kích thước phù hợp từ các ảnh được tải lên. Điều này không nhất thiết có nghĩa là bạn có thể sử dụng các màu được trích xuất một cách trực tiếp và theo thứ tự làm bảng màu trực quan.

Mặc dù chúng có thể là điểm khởi đầu đầy cảm hứng để các màu trông đẹp mắt khi kết hợp với nhau, nhưng có thể cần thực hiện một số chỉnh sửa và sửa đổi để đảm bảo rằng các màu chọn có hiệu quả trong bối cảnh trực quan hóa.

10.3.3.5. Viz Palette



Hình 10-27. Viz Palette

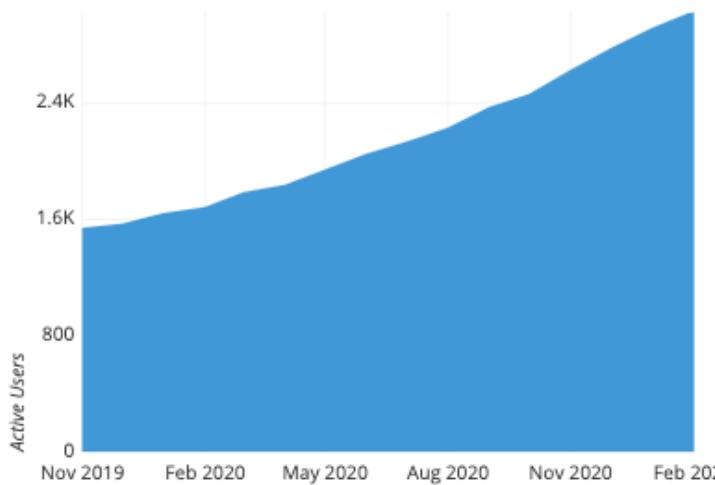
Trong phần trước, Coblis được liên kết như một nguồn tài nguyên để kiểm tra xem hình ảnh cuối cùng sẽ trông như thế nào đối với những người bị khuyết tật về nhận thức màu sắc. Viz Palette là một công cụ bảng màu rộng hơn có thể sử dụng để kiểm tra các bảng màu của mình trước khi kết hợp các hình ảnh trực quan của mình lại với nhau. Ngoài khả năng xem các tập hợp màu trong bối cảnh của các ô mẫu và trong trường hợp thiếu hụt nhận biết màu được mô phỏng, cũng có thể sửa đổi và thay đổi màu của bảng màu ngay lập tức.

10.4. MỘT SỐ ĐỒ THỊ DÙNG TRONG TRỰC QUAN HÓA DỮ LIỆU

10.4.1. Area graph (Area chart)

10.4.1.1. Giới thiệu

- Là sự điều chỉnh của Line chart trong đó khu vực dưới đường được điền vào để nhấn mạnh tầm quan trọng của nó.
- Tương tự như *line graph*, *area graph* sử dụng các điểm được nối với nhau bằng một đường.

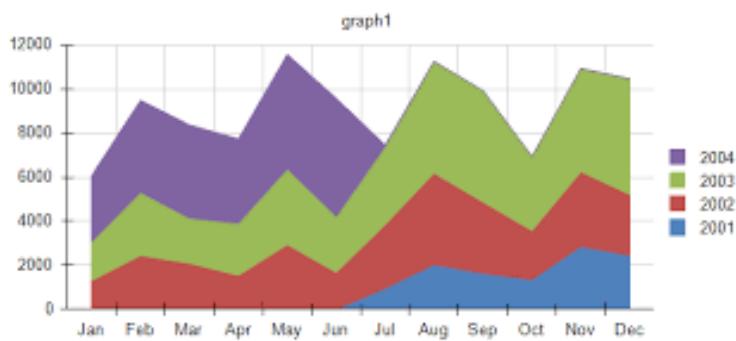


- Area chart kết hợp Line chart và Bar chart để hiển thị cách một hoặc nhiều giá trị số của nhóm thay đổi theo tiến trình của biến thứ hai, thường là theo thời gian.
- Màu tô cho khu vực dưới mỗi đường để nhấn mạnh tầm quan trọng của nó hơi trong suốt để có thể thấy được các khu vực chồng lấp. Dựa vào màu tô này để phân biệt Area chart với Line chart .Có thể sử dụng nhiều đường và màu sắc giữa mỗi dòng để cho biết số lượng cộng lại thành một tổng thể như thế nào. Tuy nhiên, *area graph* liên quan đến việc tô màu giữa đường vẽ và trực ngang hoặc giữa 2 đường vẽ liền kề nhau.

Area chart ở trên hiển thị số lượng người dùng đang hoạt động của một công ty dựa trên web, được tính theo tháng. Các giá trị cho mỗi tháng có thể được đo không chỉ từ vị trí thẳng đứng của phần trên cùng của hình mà còn từ chiều cao được tô màu giữa đường cơ sở và phần trên cùng. Trong biểu đồ này, có thể thấy số lượng người dùng đang hoạt động đã tăng gấp đôi từ tháng 11 năm 2019 đến tháng 2 năm 2020 và tỷ lệ thu hút người dùng đã tăng lên theo thời gian.

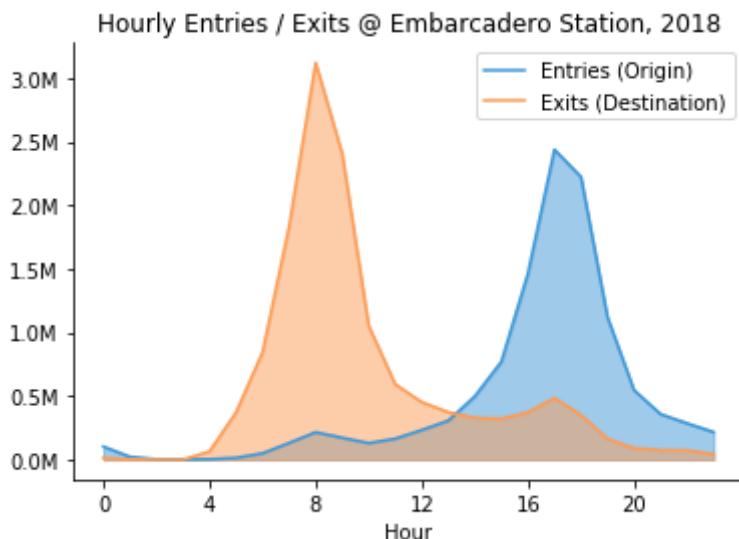
10.4.1.2. Sử dụng

- *Area graph* cho thấy sự thay đổi về một hoặc nhiều số lượng trong một khoảng thời gian nhất định.
- Giống như *Line charts*, *Area graph* được sử dụng để hiển thị sự phát triển của các giá trị định lượng trong một khoảng thời gian.
- Chúng được sử dụng phổ biến để hiển thị các xu hướng, thay vì truyền tải các giá trị cụ thể.
- Mặc dù ví dụ trên chỉ vẽ một đường duy nhất với vùng được tô bóng, Area chart thường được sử dụng với nhiều đường và chia thành hai loại Area chart khác nhau là:
 - So sánh giữa các nhóm (còn gọi là chuỗi - series)
 - Hiển thị cách chia tổng thể thành các phần thành phần.



10.4.1.2.1. Area chart chồng lấn lên nhau (Overlapping Area chart)

Trong trường hợp muốn so sánh giá trị giữa các nhóm, khi đó Area chart sẽ trở thành Overlapping Area chart. Trong Overlapping Area chart, bắt đầu bằng Line chart chuẩn. Đối với mỗi nhóm, một điểm được vẽ ở mỗi giá trị ngang với chiều cao biểu thị giá trị của nhóm trên biến trục tung; một đường nối tất cả các điểm của nhóm từ trái sang phải. Area chart thêm bóng giữa mỗi đường tới đường cơ sở bằng 0. Vì phần tô bóng cho các nhóm thường sẽ chồng lên nhau ở một mức độ nào đó nên độ trong suốt được bao gồm trong phần tô bóng để có thể dễ dàng nhìn thấy đường nét của tất cả các nhóm. Việc tô bóng giúp nhấn mạnh nhóm nào có giá trị lớn nhất dựa trên màu thuần khiết của nhóm nào được hiển thị.



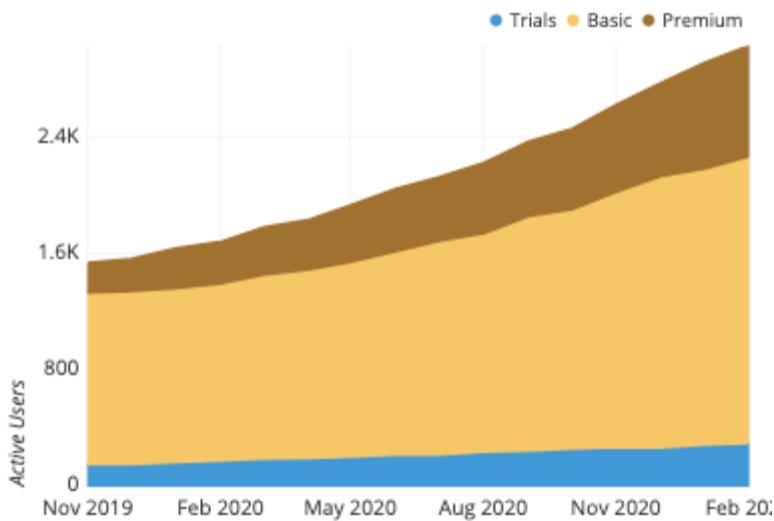
Hãy cẩn thận để các series này không phải lúc nào cũng cao hơn series kia, nếu không biểu đồ có thể bị nhầm lẫn với loại Area chart khác là Stacked area chart. Trong những trường hợp đó, chỉ cần sử dụng Line chart tiêu chuẩn sẽ là lựa chọn tốt hơn.

10.4.1.2.2. Stacked area chart

Nói chung, khi thuật ngữ 'Area chart' được sử dụng, điều thực sự được ngũ ý là Stacked area chart. Trong Overlapping Area chart, mỗi đường được tô bóng từ giá trị dọc của nó đến đường cơ sở chung. Trong Stacked area chart, các đường được vẽ lần lượt, với chiều cao của nhóm được vẽ sau sẽ đóng vai trò là đường cơ sở di chuyển. Như vậy, chiều cao xếp chồng đầy đủ của dòng trên cùng sẽ tương ứng với tổng chiều cao khi tính tổng của tất cả các nhóm.

Stacked area chart được sử dụng khi không chỉ muốn theo dõi tổng giá trị mà còn muốn hiểu sự phân tích tổng giá trị đó theo nhóm. So sánh độ cao của từng đoạn của đường cong cho

phép có được ý tưởng chung về cách mỗi nhóm con so sánh với nhóm khác trong đóng góp của chúng vào tổng số.



Biểu đồ cho thấy: Hầu hết người dùng đang hoạt động đều đến từ các tài khoản “Basic” (cơ bản), nhưng người dùng “Premium” (cao cấp) dường như đang tăng nhanh hơn một cách tương ứng.

10.4.1.2.3. Minh họa cấu trúc dữ liệu để vẽ đồ thị

- Dữ liệu được mô tả bằng Area chart thường được tổng hợp thành một bảng có hai cột trỏ lên. Cột đầu tiên cho biết các vị trí trên trực hoành nơi mỗi đường sẽ được vẽ. Mỗi cột tiếp theo sẽ chỉ ra sự đóng góp theo chiều dọc cho mỗi điểm, một cột trên mỗi chuỗi sẽ được vẽ. Định dạng này có thể áp dụng cho cả Overlapping Area chart và Stacked Area chart, với sự khác biệt chính giữa các biểu đồ là cách diễn giải các giá trị để hiển thị.

Month	Trials	Basic	Premium
2019-11	154	1180	201
2019-12	157	1186	219
2020-01	170	1195	270
2020-02	180	1213	285
...

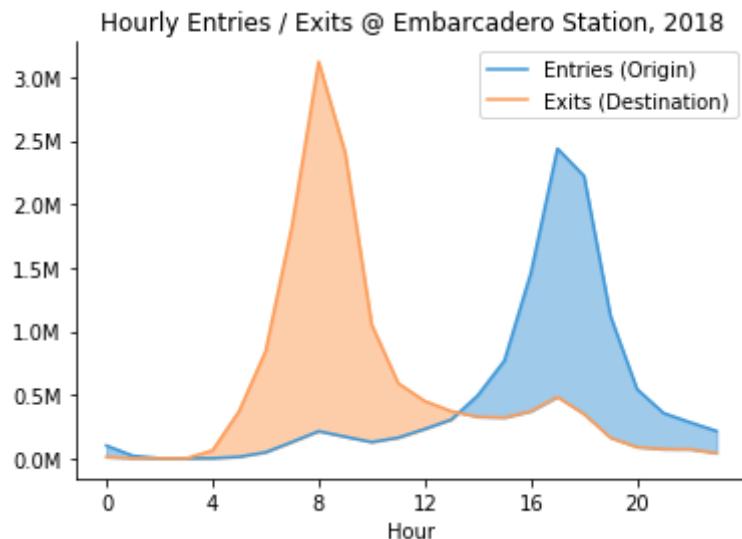
- Đối với Stacked Area chart, một số công cụ trực quan yêu cầu các cột liệt kê không phải những đóng góp riêng lẻ mà thay vào đó là những đóng góp tích lũy. Trong trường hợp này, các cột chỉ định trực tiếp chiều cao của dòng và sự đóng góp của mỗi nhóm được ngụ ý bằng sự khác biệt về giá trị giữa các cột.

Month	Trials	+ Basic	+ Premium
2019-11	154	1334	1535
2019-12	157	1343	1562
2020-01	170	1365	1635
2020-02	180	1393	1678
...

10.4.1.3. Sử dụng Area chart hiệu quả

10.4.1.3.1. Bao gồm đường cơ sở bằng 0

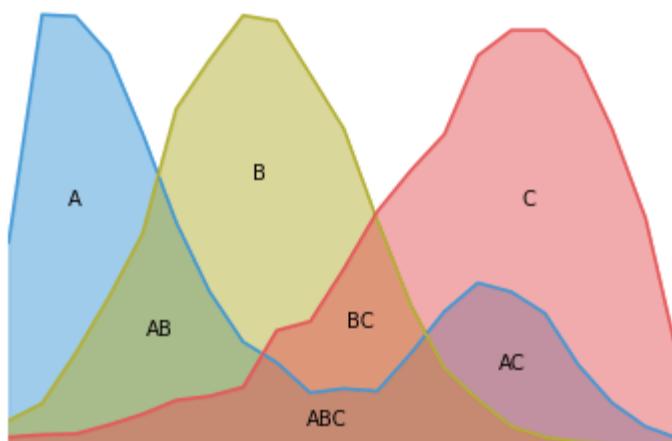
Mặc dù Line chart không bị hạn chế yêu cầu đường cơ sở bằng 0, việc bổ sung bóng có nghĩa là độ cao của các vùng được tô màu sẽ được sử dụng để so sánh kích thước của các giá trị của mỗi nhóm. Vì vậy, giống như Bar chart, bắt buộc phải có đường cơ sở bằng 0 để thực hiện tô bóng. Việc cắt trực sẽ dẫn đến tỷ lệ thực tế trong các giá trị nhóm không khớp với những gì được ngũ ý trong biểu đồ được tạo.



Một ngoại lệ đối với quy tắc này có thể xảy ra khi so sánh hai chuỗi trong Area chart chồng chéo với sự thay đổi đối với quy tắc tô bóng. Nếu giới hạn bóng ở giữa các đường, thay vì từ cả hai đường đến một đường cơ sở chung, khi đó có thể phóng to các giới hạn trực dọc thành hiệu ứng quan tâm mà không cần đường cơ sở. Bóng bây giờ mang một ý nghĩa khác, với màu sắc cho biết nhóm nào có giá trị lớn hơn và lượng màu cho biết kích thước của sự khác biệt.

10.4.1.3.2. Giới hạn số lượng series trong Overlapping Area chart

Khi càng có nhiều series trong Overlapping Area chart thì càng có nhiều sự kết hợp màu sắc khi chúng chồng lên nhau. Thực tế là hầu hết các màu sẽ không được liên kết với một nhóm duy nhất có thể gây ra một số khó khăn trong việc giải thích. Ngay cả khi chỉ có ba chuỗi, điều này đôi khi có thể là quá nhiều để theo dõi: ba màu riêng lẻ, ba màu chồng lên nhau và một màu cho cả ba nhóm chồng lên nhau, tổng cộng có bảy màu.



Việc so sánh hai series thường là an toàn, tuy nhiên nếu một serie luôn lớn hơn serie kia thì biểu đồ có thể dễ bị nhầm lẫn với Stacked Area chart. Người đọc cũng có thể bị nhầm lẫn khi giải thích những màu sắc chồng chéo sẽ không có trong phần chú thích.

Theo nguyên tắc chung, nếu đang nghĩ đến việc sử dụng Overlapping, hãy:

- Giới hạn ở hai series
- Suy nghĩ xem liệu việc sử dụng Line chart có thể hiện sự so sánh giữa các nhóm rõ ràng hơn hay không?

10.4.1.3.3. Sắp xếp thứ tự các dòng trong Stacked area chart

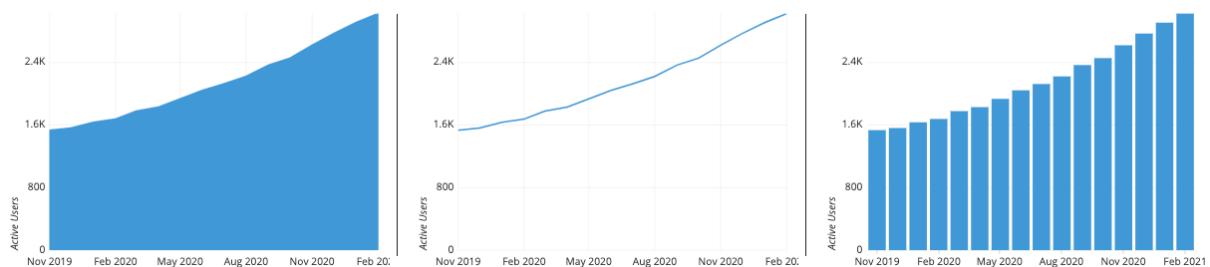
- Mặc dù hình dạng tổng thể của biểu đồ sẽ giống nhau bất kể thứ tự các đường vẽ của nhóm được vẽ như thế nào, nhưng việc đọc trực quan có thể được hỗ trợ thông qua việc lựa chọn tốt thứ tự đường kẻ.
- Một nguyên tắc nhỏ là đặt các nhóm lớn nhất hoặc ổn định nhất ở dưới cùng, sau đó lần lượt đến các nhóm có nhiều thay đổi nhất hoặc nhỏ nhất được sắp xếp giảm dần. Như vậy, nhóm ở trên cùng chính là nhóm có nhiều thay đổi nhất hoặc nhỏ nhất.

10.4.1.4. Các vấn đề thường gặp khi sử dụng Area chart

10.4.1.4.1. Sử dụng Area chart để vẽ duy nhất một series

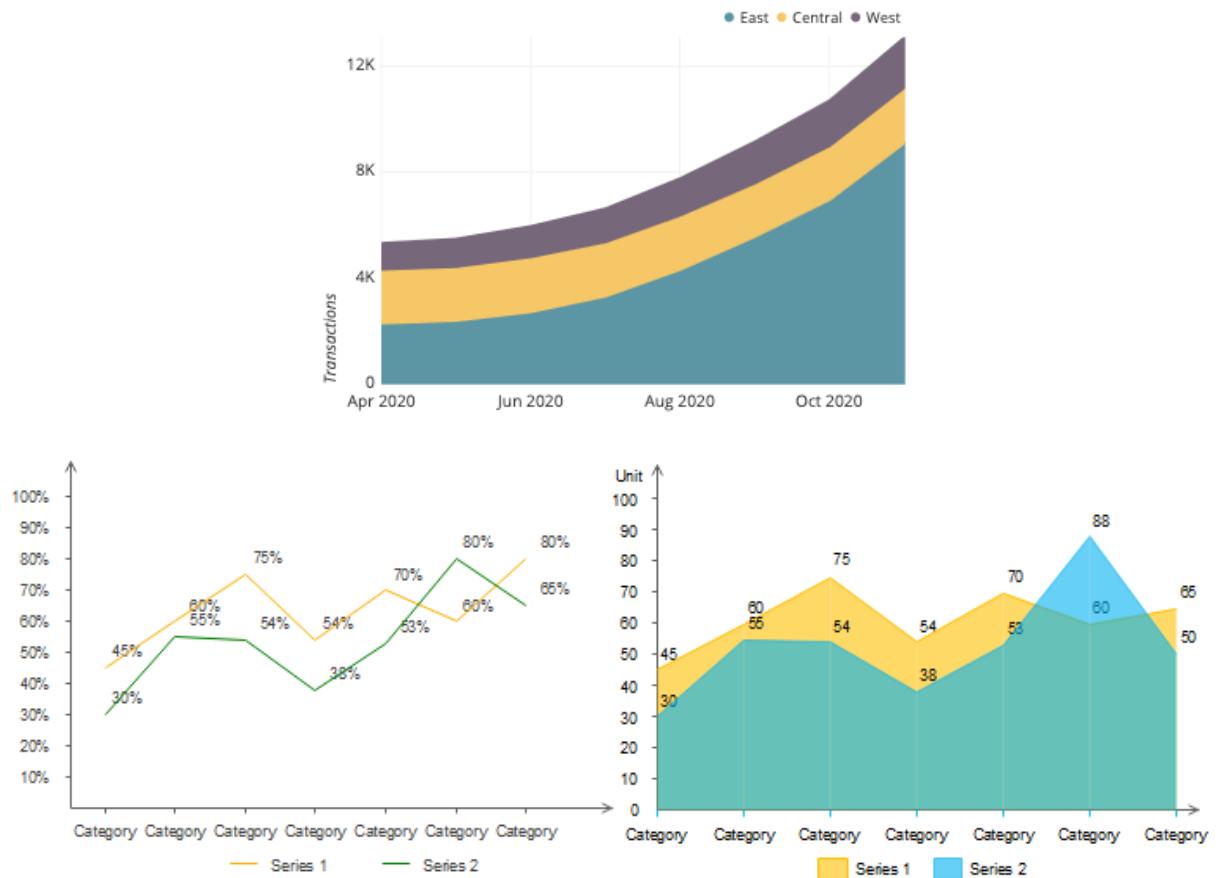
Nhu đã lưu ý ở trên, Area plot được dùng để so sánh hoặc hiển thị sự phân chia số lượng giữa hai hoặc nhiều chuỗi. Khi chỉ có một chuỗi giá trị để vẽ, việc sử dụng Area chart thường là sai lầm. Tùy chọn tốt hơn là chỉ chọn Bar chart hoặc Line chart, tùy thuộc vào nội dung muốn tìm hiểu hoặc truyền đạt về dữ liệu.

Nếu muốn hiểu sự tiến triển của các giá trị chính xác theo thời gian và khi không có quá nhiều giá trị để vẽ trên trục ngang thì Bar chart là một lựa chọn tốt. Nếu không, biểu đồ dạng đường sẽ là lựa chọn tốt hơn. Các dòng có tỷ lệ dữ liệu trên mực hiệu quả hơn và giao diện rõ ràng hơn với nhiều giá trị để vẽ so với các thanh. Ngoài ra, khi có nhiều giá trị, có thể quan tâm nhiều hơn đến hướng và độ dốc của xu hướng hơn là các giá trị chính xác, khi đó Line chart hoạt động tốt hơn.



10.4.1.4.2. Giải thích các giá trị trên các nhóm riêng lẻ trong Stacked area chart

Trong Stacked area chart, việc đo các giá trị chính xác chỉ thực sự dễ dàng đối với hai trường hợp: đối với tổng thể và đối với nhóm dưới cùng. Đối với các nhóm trung gian, để có được giá trị chính xác về đóng góp của một nhóm đòi hỏi phải tìm chiều cao của dòng của nhóm đó và trừ đi chiều cao của dòng bên dưới nó.

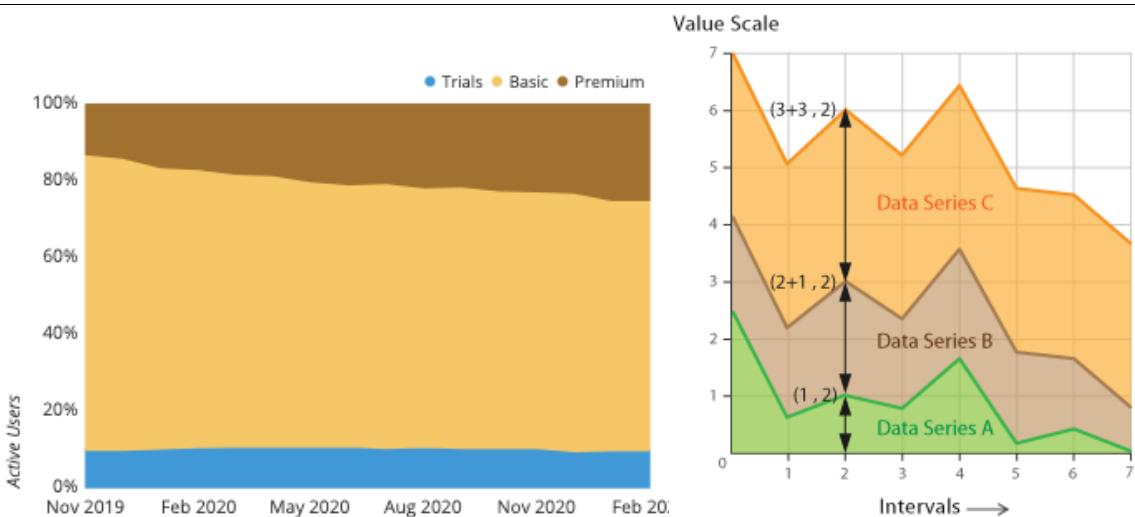


Nhiệm vụ càng trở nên tồi tệ hơn khi muốn theo dõi những thay đổi theo thời gian. Điều này dễ dàng đối với nhóm tổng và nhóm dưới cùng. Tuy nhiên, các nhóm trung gian bị cản trở bởi đường cơ sở thay đổi, gây khó khăn cho việc nhận biết chính xác sự khác biệt theo chiều dọc. Ví dụ minh họa trong hình sau cho thấy: mặc dù cường độ của nhóm màu vàng ở giữa dường như thay đổi theo thời gian, nhưng trên thực tế, độ cao tại mỗi điểm là nhất quán.

Nếu muốn biết về các giá trị nhóm chính xác và những thay đổi của chúng theo thời gian thì việc chọn Line chart chuẩn sẽ là lựa chọn tốt hơn.

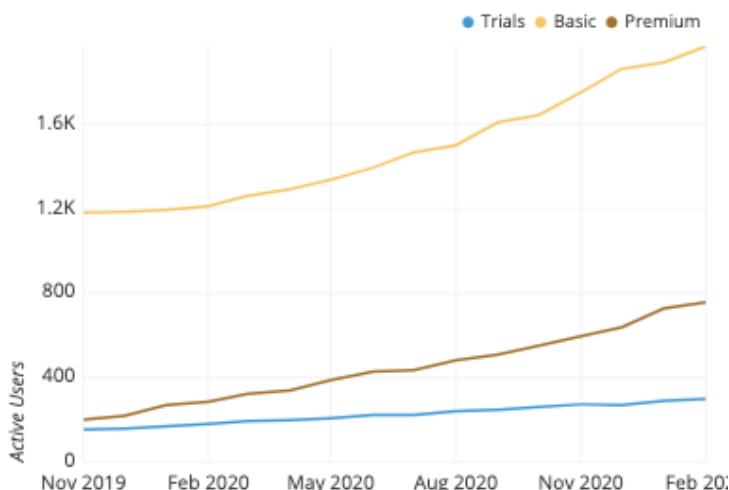
10.4.1.5. Tùy chọn phổ biến được dùng trên Stacked Area chart

Một tùy chọn phổ biến cho Area chart là Stacked Area chart theo tỷ lệ phần trăm hoặc tần suất tương đối. Thay vì xếp chồng các giá trị tuyệt đối của từng nhóm tại mỗi lát cắt dọc, xếp chồng phần đóng góp tương đối hoặc phần trăm của từng nhóm vào tổng, sao cho chiều cao tổng thể luôn là 100%. Loại biểu đồ này mất thông tin về xu hướng của tổng tuyệt đối (và do đó sẽ cần Line chart riêng) nhưng giúp đưa ra sự so sánh về đóng góp tương đối giữa các nhóm. Loại biểu đồ này có được đường cơ sở thứ hai ở đầu biểu đồ để có thể đánh giá sự đóng góp của một nhóm riêng lẻ.



10.4.1.6. Các đồ thị liên quan

10.4.1.6.1. Line chart

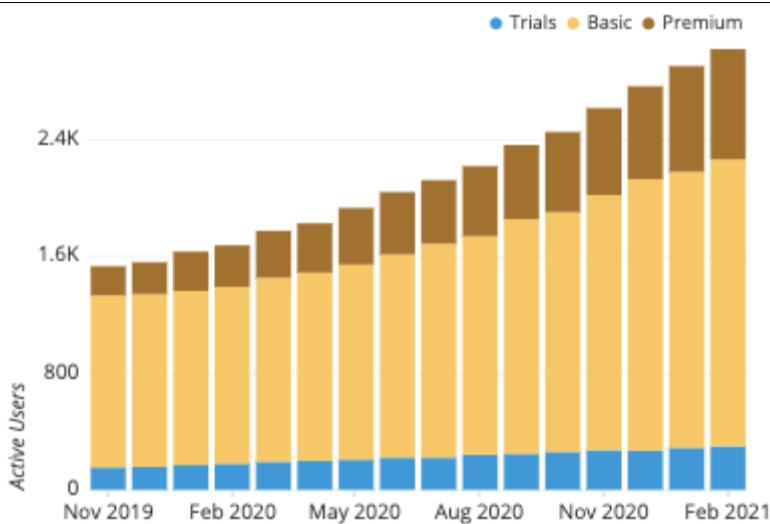


Line chart là tiền thân chính của loại Area chart. Nếu không chắc chắn liệu Area chart có phải là thứ cần quan tâm hay không thì việc sử dụng Line chart có thể sẽ là sai lầm. Điều này đặc biệt đúng đối với Overlapping Area chart, trong đó các vùng chồng chéo có thể nhanh chóng vượt khỏi tầm kiểm soát. Một cân nhắc khác, luôn có thể tạo nhiều biểu đồ hơn nếu muốn thực hiện nhiều so sánh trong dữ liệu của mình thay vì chỉ cảm thấy cần phải chọn chỉ một loại biểu đồ duy nhất để mang theo thông tin càng nhiều càng tốt.

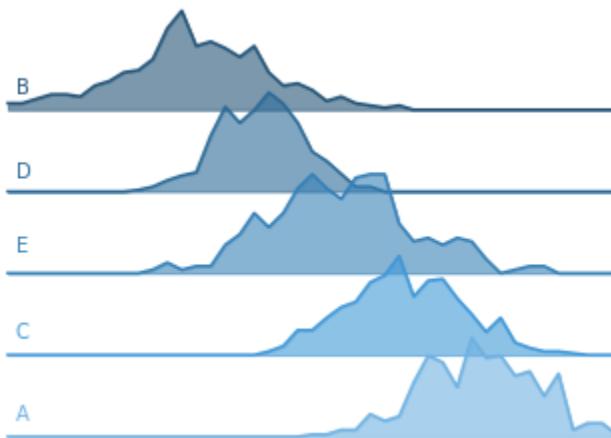
10.4.1.6.2. Bar chart và stacked bar chart

Tiền thân khác của Area chart là Bar chart. Tuy nhiên, chính xác hơn, mối liên kết gần nhất là giữa Stacked Area chart và Stacked bar chart; thực sự không có sự tương đồng tốt giữa Overlapping area chart và bất kỳ Bar chart nào.

Stacked bar chart rất giống với Stacked Area chart, chỉ có thanh thay vì đường. Vì vậy, nhiều hạn chế của Stacked Area chart cũng áp dụng cho Stacked bar chart. Tuy nhiên, một ưu điểm của các thanh xếp chồng lên nhau là việc đưa ra đánh giá nhất quán về giá trị trong mỗi bin trực ngang sẽ dễ dàng hơn nhiều. Các vùng được tô bóng trong Area chart có thể bị biến dạng như đã thấy ở trên, đặc biệt khi một đường thay đổi hướng. Vì mỗi vùng là hình chữ nhật trong Stacked bar chart nên loại biến dạng này sẽ tránh được.



10.4.1.6.3. Ridgeline plot

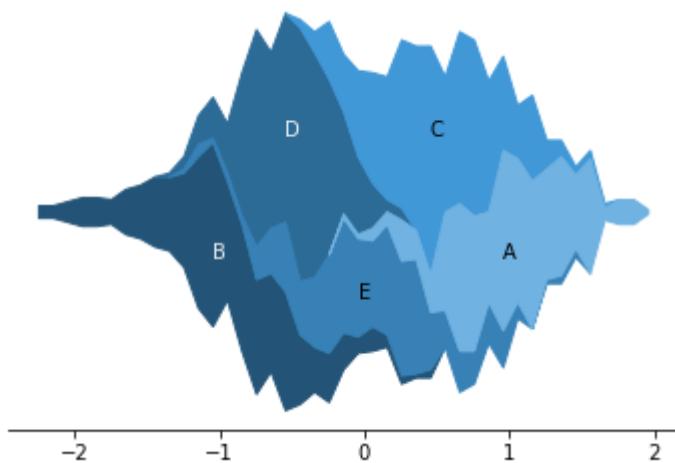


Một lựa chọn bổ sung cho Overlapping Area chart là Ridgeline plot. Thay vì vẽ tất cả các đường và vùng màu trên cùng một trục, Ridgeline plot đặt mỗi đường trên một trục khác nhau, mỗi đường có độ lệch dọc một phần so với các đường khác. Do tính chất lệch của các đường, các dấu dọc thường bị loại bỏ khỏi Ridgeline plot. Điều này có nghĩa là các Ridgeline plots sẽ hữu ích nhất khi có một mẫu rõ ràng trong các giá trị của từng chuỗi riêng lẻ chỉ dựa trên hình dạng của chúng.

10.4.1.6.4. Stream graph

Một họ hàng tương tự của Area chart là Stream graph. Trong Stacked Area chart, tất cả các đường được xếp chồng lên nhau trên đường cơ sở thẳng ở cuối ngăn xếp. Với Stream graph, đường cơ sở được đặt qua tâm của biểu đồ và các khu vực tập trung đối xứng xung quanh đường trung tâm. Vì điều này, rất khó để đánh giá các giá trị chính xác cho bất kỳ nhóm nào hoặc thậm chí cho tổng thể.

Stream graph được sử dụng tốt nhất ở dạng tương tác khi có nhiều dữ liệu được trình bày cho nhiều đối tượng. Tính tương tác rất quan trọng để cho phép người đọc tìm hiểu sâu hơn về hình ảnh và hình thành những phát hiện của riêng họ. Tuy nhiên, khi cần đưa ra những đánh giá chính xác hoặc cần thực hiện một bản trình bày tĩnh, tốt hơn hết sử dụng một hình ảnh trực quan thông thường hơn, tổng hợp dữ liệu theo cách rút ra tốt nhất những điểm muốn trình bày.



Biểu đồ này sử dụng cùng dữ liệu với ridgeline plot ở trên.

10.4.2. Arc diagram (biểu đồ cung)

10.4.2.1. Giới thiệu

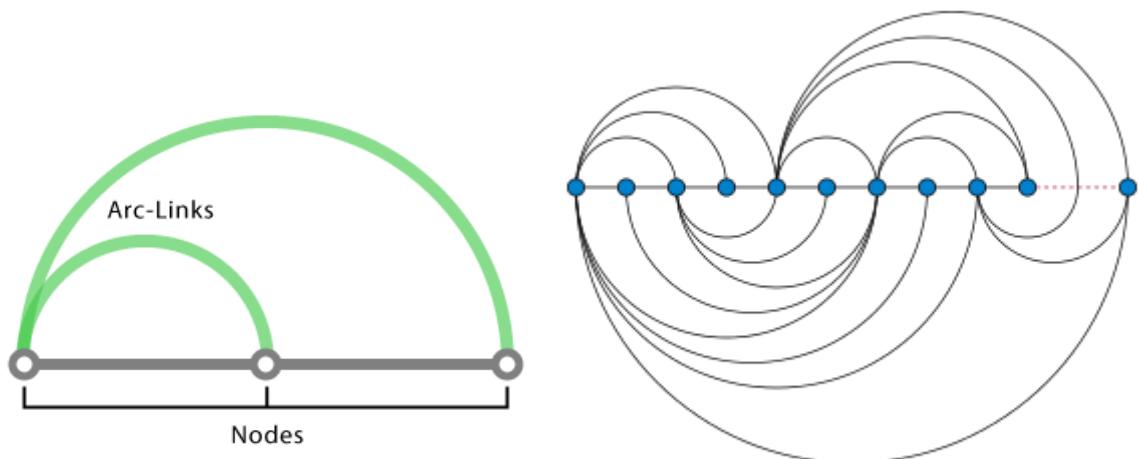
Arc diagram là một kiểu vẽ đồ thị, trong đó các nút (node) của đồ thị được đặt dọc theo một đường thẳng trong mặt phẳng Euclide, với các arcs được vẽ dưới dạng hình bán nguyệt trong một hoặc cả hai nửa mặt phẳng được giới hạn bởi đường thẳng hoặc dưới dạng các đường cong tròn. Trong một số trường hợp, các đoạn thẳng của chính đường đó cũng được phép làm các cạnh, miễn là chúng chỉ nối các đỉnh liên tiếp dọc theo đường đó.

Các biến thể của phong cách vẽ này trong đó các hình bán nguyệt được thay thế bằng các đường cong lồi thuộc loại khác.

Việc sử dụng cụm từ *Arc diagram* cho loại hình vẽ này tuân theo việc sử dụng loại sơ đồ tương tự của Wattenberg (2002) để trực quan hóa các mẫu lặp lại trong chuỗi, bằng cách sử dụng các cung để kết nối các cặp chuỗi con bằng nhau. Tuy nhiên, phong cách vẽ đồ thị này lâu đời hơn nhiều so với tên gọi của nó, bắt nguồn từ tác phẩm của Saaty (1964) và Nicholson (1968), những người đã sử dụng sơ đồ cung để nghiên cứu các số giao nhau của đồ thị. Một tên cũ hơn nhưng ít được sử dụng hơn cho sơ đồ cung là linear embeddings (nhưng tuy nhiên). Gần đây hơn, sơ đồ cung đã được sử dụng trong khuôn khổ cấu trúc liên kết mạch của các nút thắt và đám rối, trong đó chúng được gọi là circuit diagrams (sơ đồ mạch).

Heer, Bostock & Ogievetsky (2010) viết rằng sơ đồ cung "có thể không truyền tải cấu trúc tổng thể của đồ thị một cách hiệu quả như bố cục hai chiều", nhưng cách bố trí của chúng giúp dễ dàng hiển thị dữ liệu đa biến liên quan đến các đỉnh của đồ thị. Các ứng dụng của sơ đồ cung bao gồm sơ đồ Farey, trực quan hóa các kết nối lý thuyết số giữa các số hữu tỷ và sơ đồ biểu thị cấu trúc thứ cấp RNA trong đó các giao điểm của sơ đồ biểu thị các nút giả trong cấu trúc.

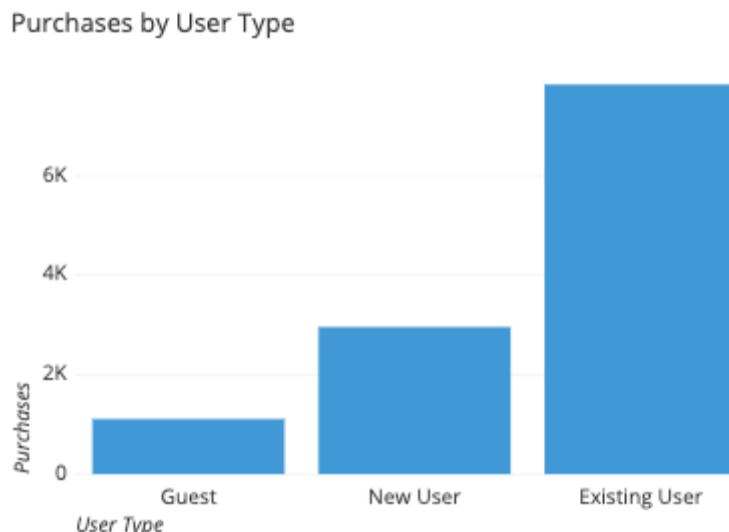
10.4.2.2. Minh họa



10.4.3. Bar graph (biểu đồ thanh – Column graph)

10.4.3.1. Giới thiệu

Bar graph (còn gọi là Bar chart, Column chart) vẽ các giá trị số cho các cấp độ của một thuộc tính phân loại dưới dạng thanh. Các giá trị phân loại được vẽ trên một trục biểu đồ và các giá trị được vẽ trên trục kia. Mỗi giá trị phân loại yêu cầu một thanh và độ dài của mỗi thanh tương ứng với giá trị của thanh. Các thanh được vẽ trên một đường cơ sở chung để cho phép dễ dàng so sánh các giá trị.



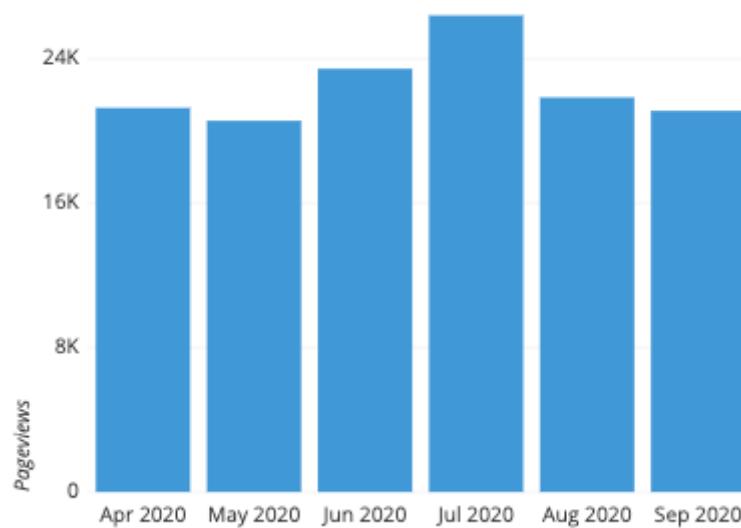
Trong Bar chart ở trên, mô tả số lượng giao dịch mua được thực hiện trên một trang web bởi các loại người dùng khác nhau. Đặc điểm phân loại là loại người dùng, được biểu thị trên trực hoành và chiều cao của mỗi thanh tương ứng với số lượng giao dịch mua được thực hiện theo từng loại người dùng. Có thể thấy từ biểu đồ này rằng mặc dù số lần mua hàng từ người dùng mới tạo tài khoản người dùng (New User) nhiều gấp ba lần so với những người không tạo tài khoản người dùng (Guest), nhưng cả hai đều bị giảm đi so với số lượng mua hàng được thực hiện bởi người dùng cũ (Existing User).

10.4.3.2. Sử dụng

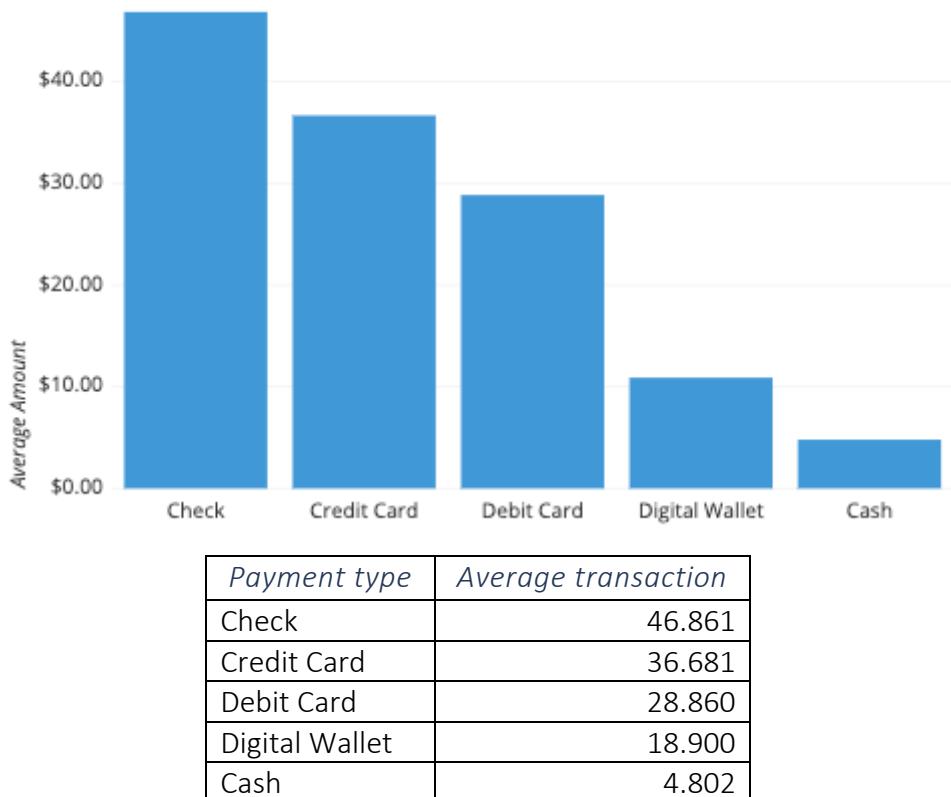
Bar chart được sử dụng khi bạn muốn hiển thị sự phân bố các điểm dữ liệu hoặc thực hiện so sánh các giá trị số liệu giữa các nhóm con khác nhau của dữ liệu. Từ Bar chart, có thể biết nhóm nào cao nhất hoặc phổ biến nhất và so sánh các nhóm với nhau như thế nào. Vì đây là một nhiệm vụ khá phổ biến nên Bar chart là loại biểu đồ thông dụng.

- Những điểm chính cần quan tâm khi sử dụng *bar graph*:
 - Được sử dụng cho dữ liệu định tính (danh mục).
 - Các cột không được kết nối.
 - Có thể hiện sự so sánh giữa các danh mục (các cột).
- *Biến chính của Bar chart*: là biến phân loại của nó, thường thuộc 1 trong các dạng sau:
 - Danh nghĩa (Nominal Attributes): có thể được coi là nhãn (label) như: quốc gia, tỉnh – thành phố, loại ngành nghề, phương thức truy cập trang web (máy tính để bàn, thiết bị di động), loại khách truy cập (miễn phí, cơ bản, cao cấp)...
 - Thứ tự (Ordinal Attributes): như về kích thước (nhỏ, trung bình, lớn), về học lực (giỏi, khá, trung bình, yếu), về BMI (UnderWeight, Normal, OverWeight, Obese), ...
 - Ngoài ra, một số biến không phân loại có thể được chuyển đổi thành các nhóm như: ngày, tháng, quý, năm, ... Điểm quan trọng đối với trường hợp này là các nhóm khác biệt.
- *Biến phụ của Bar chart*:
 - Giá trị của biến phụ xác định độ dài của mỗi thanh.
 - Sẽ có bản chất là số: số nguyên, số thực. Các giá trị có thể gấp là tần suất hoặc tỷ lệ, giá trị trung bình, tổng hoặc một số thước đo tóm tắt khác được tính riêng cho từng nhóm

Ví dụ: biểu đồ sau đây tính số lượt xem trang trong khoảng thời gian sáu tháng. Qua biểu đồ, có thể thấy có một đỉnh nhỏ vào tháng 6 và tháng 7 trước khi quay trở lại đường cơ sở trước đó.



Ví dụ mô tả quy mô giao dịch trung bình theo phương thức thanh toán. Lưu ý rằng mặc dù các khoản thanh toán trung bình bằng séc là cao nhất nhưng sẽ cần một biểu đồ khác để cho thấy tần suất khách hàng thực sự sử dụng chúng.



Dữ liệu được hiển thị dưới dạng Bar chart có thể ở dạng thu gọn như bảng trên, với một cột dành cho danh mục và cột thứ hai dành cho giá trị của chúng. Đôi khi, dữ liệu có thể ở dạng chưa được tổng hợp như hình minh họa trong bảng bên dưới, với công cụ trực quan hóa tự động thực hiện tổng hợp tại thời điểm tạo trực quan hóa.

...	payment_type	amount	(other data columns ...)
...	credit	48.81	...
...	debit	32.10	...
...	debit	26.48	...
...	cash	4.99	...
...	digital_wallet	12.57	...

Đối với Bar chart dựa trên số lượng, chỉ cần cột đầu tiên. Đối với Bar chart dựa trên tóm tắt, hãy nhóm theo cột đầu tiên, sau đó tính số đo tóm tắt trên cột thứ hai.

10.4.3.3. Sử dụng bar charts hiệu quả

10.4.3.3.1. Sử dụng chung đường cơ sở có giá trị bằng 0 (zero)

Hãy đảm bảo rằng tất cả các thanh đang được vẽ đều dựa trên đường cơ sở có giá trị bằng 0. Đường cơ sở đó không chỉ giúp người đọc dễ dàng so sánh độ dài thanh hơn mà còn duy trì tính trung thực của việc trực quan hóa dữ liệu của bạn. Bar chart có đường cơ sở khác 0

hoặc một số khoảng trống khác trong thang trực có thể dễ dàng trình bày sai sự so sánh giữa các nhóm vì tỷ lệ về độ dài thanh sẽ không khớp với tỷ lệ trong các giá trị thanh thực tế.

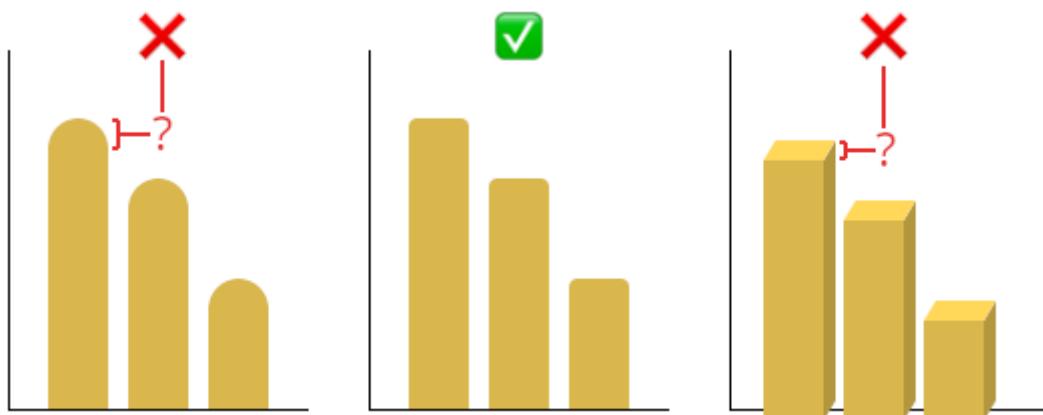


Trong hình trên, bằng cách cắt 90 điểm ra khỏi trực tung, chênh lệch giữa 2 thanh có thể được phóng đại để trông giống Our Product lớn gấp 3 lần Their Product.

10.4.3.3.2. Duy trì dạng hình chữ nhật cho thanh

Một điều không nên làm nữa là làm sai lệch hình dạng của các thanh cần vẽ. Một số công cụ sẽ cho phép làm tròn các nắp thanh thay vì chỉ có các cạnh thẳng. Việc làm tròn này có nghĩa là người đọc khó biết được giá trị thực tế cần đọc ở đâu: từ đầu hình bán nguyệt hay đâu đó ở giữa? Có thể bo tròn các góc một chút nhưng hãy đảm bảo mỗi thanh đủ phẳng để phân biệt giá trị thực của nó và giúp bạn dễ dàng so sánh giữa các thanh.

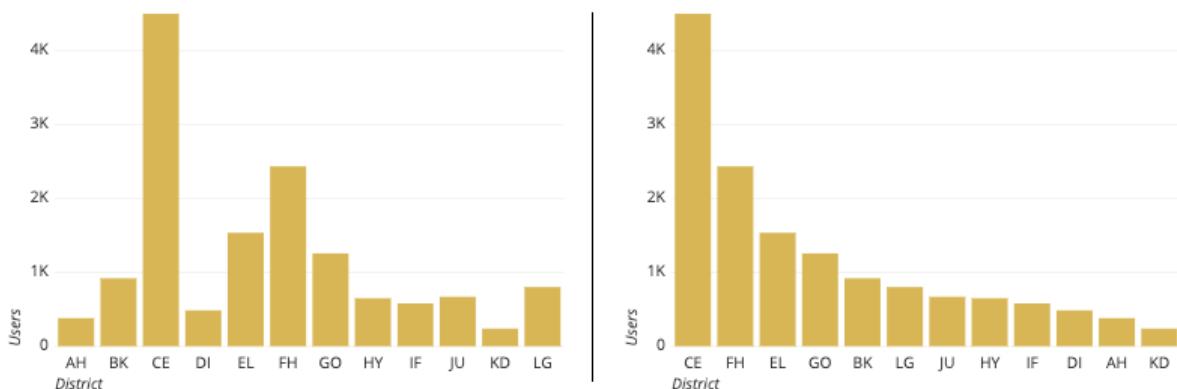
Tương tự, nên tránh đưa các hiệu ứng 3-d vào thanh của mình. Giống như làm tròn nhiều, điều này có thể khiến việc biết cách đo độ dài thanh trở nên khó khăn hơn và như một phần thường, có thể khiến đường cơ sở không được căn chỉnh.



10.4.3.3.3. Xem xét thứ tự của các cấp danh mục

Một điều cũng nên cân nhắc khi lập Bar chart là bạn sẽ vẽ các thanh theo thứ tự nào. Một quy ước tiêu chuẩn cần thực hiện là sắp xếp các thanh từ dài nhất đến ngắn nhất: mặc dù luôn có thể so sánh độ dài các ô bất kể thứ tự nào, nhưng điều này có thể giảm bớt gánh nặng

cho người đọc khi tự mình thực hiện những so sánh đó. Ngoại lệ chính đối với điều này là nếu các nhãn danh mục vốn đã được sắp xếp theo một cách nào đó. Trong những trường hợp như vậy, thứ tự vốn có thường được ưu tiên.

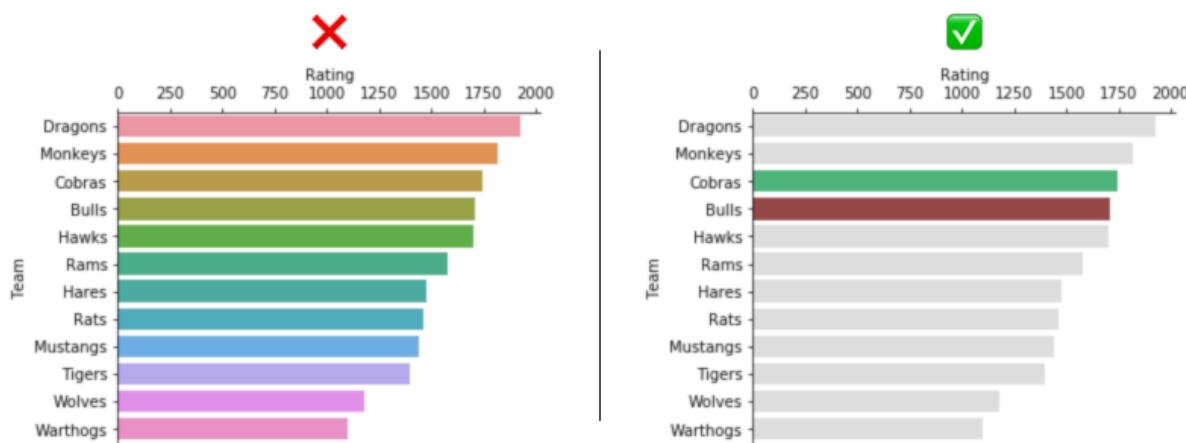


Mã quận nếu trước đây nếu trước đây không thường được sắp xếp theo thứ tự, thì cách thể hiện tốt hơn là sắp xếp theo giá trị.

10.4.3.3.4. Sử dụng màu sắc một cách khôn ngoan

Một điều cần cân nhắc khác là nên sử dụng màu sắc như thế nào trong Bar chart của mình. Cụ thể là:

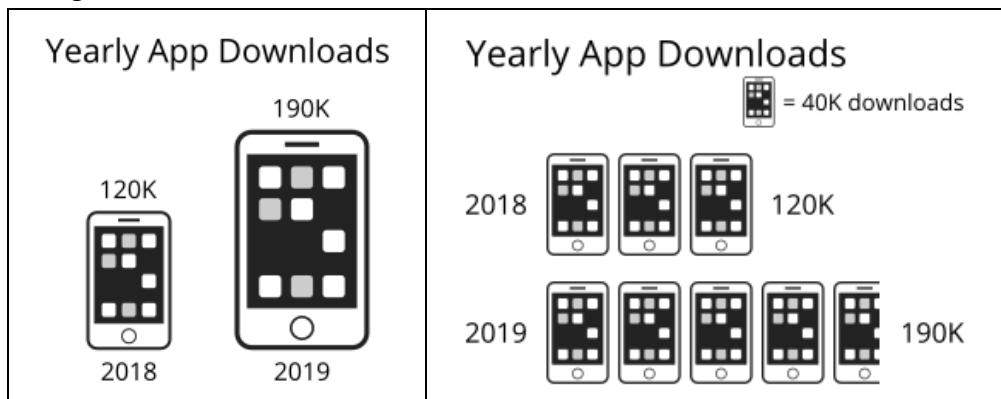
- Một số công cụ nhất định sẽ tô màu mỗi thanh khác nhau theo mặc định, nhưng điều này có thể khiến người đọc mất tập trung bằng cách ngụ ý thêm ý nghĩa nếu không có ý nghĩa nào tồn tại. Thay vào đó, màu sắc nên được sử dụng có mục đích. Ví dụ: có thể sử dụng màu sắc để đánh dấu các cột cụ thể nhằm nhấn mạnh nội dung cần trình bày.
- Màu sắc cũng có thể được sử dụng nếu chúng có ý nghĩa đối với các danh mục được đăng (ví dụ: để phù hợp với màu của công ty hoặc nhóm).



Hình bên trái cho thấy việc sử dụng màu sắc cầu vòng không thêm được bất cứ điều gì có ý nghĩa vào việc giải thích nội dung cần trình bày. Trong khi hình bên phải, hầu hết các thanh đều có màu xám trung tính để làm nổi bật sự so sánh giữa hai thanh màu.

10.4.3.3.5. Cách sử dụng sai phổ biến: Thay thế thanh bằng hình ảnh

- Có thể muốn thay thế các thanh bằng hình ảnh mô tả những gì đang được đo (ví dụ: túi tiền cho số tiền), hãy cẩn thận để không trình bày sai dữ liệu của mình theo cách này. Nếu lựa chọn ký hiệu của bạn chia tỷ lệ cả chiều rộng và chiều cao với giá trị, thì sự khác biệt sẽ trông lớn hơn nhiều so với thực tế vì cuối cùng mọi người sẽ so sánh diện tích của các thanh thay vì chỉ chiều rộng hoặc chiều cao của chúng. Trong ví dụ hình bên trái bên dưới, số lượt download tăng 58% từ năm 2018 đến năm 2019. Tuy nhiên, mức tăng trưởng này bị phóng đại với cách trình bày dựa trên biểu tượng, vì diện tích bề mặt của biểu tượng năm 2019 lớn hơn 2,5 lần kích thước của biểu tượng năm 2018.

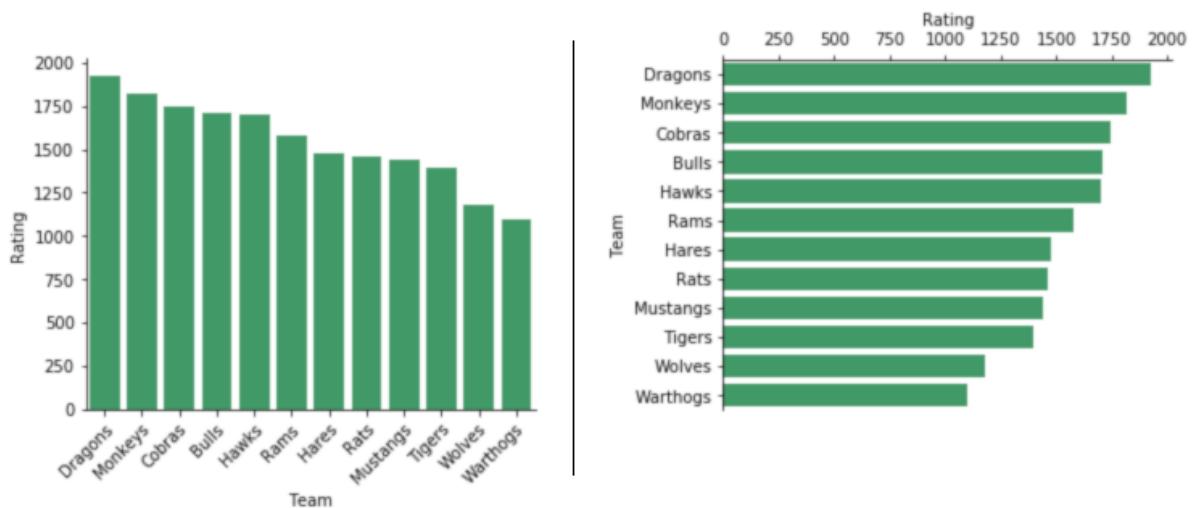


- Nếu cảm thấy cần phải sử dụng các biểu tượng để mô tả giá trị thì tùy chọn tốt hơn - mặc dù vẫn không tuyệt vời - là sử dụng loại biểu đồ tượng hình để thay thế. Trong biểu đồ chữ tượng hình, giá trị của mỗi danh mục được biểu thị bằng một loạt biểu tượng, mỗi biểu tượng đại diện cho một số lượng nhất định. Theo một nghĩa nào đó, điều này giống như thay đổi kết cấu của thanh tượng ứng thành một hình ảnh lặp lại. Một lưu ý lớn đối với loại biểu đồ này là nó có thể làm cho các giá trị khó đọc hơn vì người đọc cần thực hiện một số phép tính nhẩm để đánh giá các giá trị tương đối của từng danh mục. Lúc này đồ thị chuyển thành Pictograph.

10.4.3.4. Các tùy chọn phổ biến được dùng với Bar chart

10.4.3.4.1. Thanh ngang so với thanh dọc

Một biến thể phổ biến của Bar chart là liệu Bar chart có nên được định hướng theo chiều dọc (với các danh mục trên trực hoành) hay theo chiều ngang (với các danh mục trên trực tung). Mặc dù Bar chart dọc thường là biểu đồ mặc định nhưng khi gấp tên các nhãn danh mục dài, nên chuyển sang sử dụng Bar chart ngang. Trong biểu đồ dọc, các nhãn này có thể chồng lên nhau và cần được xoay hoặc dịch chuyển để vẫn dễ đọc; hướng ngang tránh được vấn đề này.



Nếu các thanh trong ví dụ trước được định hướng theo chiều đọc thì nhãn đánh dấu Nhóm sẽ cần phải được xoay để có thể đọc được.

10.4.3.4.2. Bao gồm chú thích giá trị



Một bổ sung phổ biến cho Bar chart là chú thích giá trị. Mặc dù người đọc khá dễ dàng so sánh độ dài thanh và đánh giá các giá trị gần đúng từ Bar chart, nhưng các giá trị chính xác nhất thiết phải dễ dàng nêu ra. Chú thích có thể cho biết những giá trị này ở những vị trí quan trọng và thường được đặt ở giữa thanh hoặc ở cuối thanh.

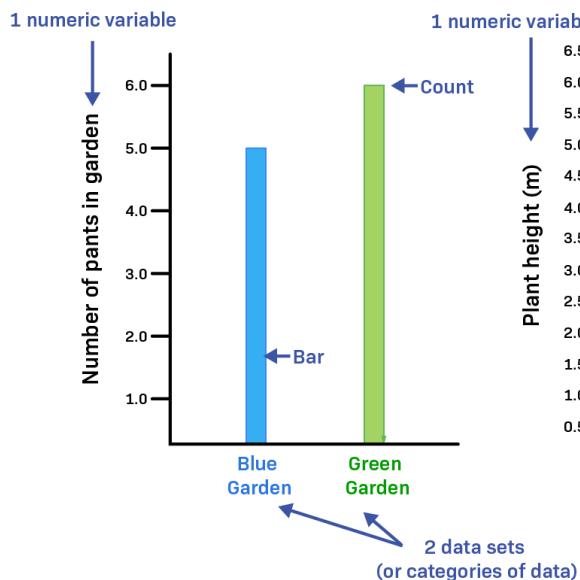
10.4.3.4.3. Thêm Error bars để bổ sung thông tin

Khi các giá trị số là thước đo tóm tắt, điều cần cân nhắc thường xuyên là có nên đưa các thanh lỗi (*error bars*) vào biểu đồ hay không. *Error bar* là các râu bổ sung được thêm vào cuối mỗi thanh để biểu thị sự thay đổi trong các điểm dữ liệu riêng lẻ góp phần vào thước đo tóm tắt. Vì có nhiều lựa chọn cho thước đo độ không đảm bảo (ví dụ: độ lệch chuẩn, khoảng tin cậy, phạm vi liên tú phân vị), điều quan trọng là khi hiển thị các *Error bar*, phải ghi chú trong chú thích hoặc nhận xét những gì các *Error bars* thể hiện.

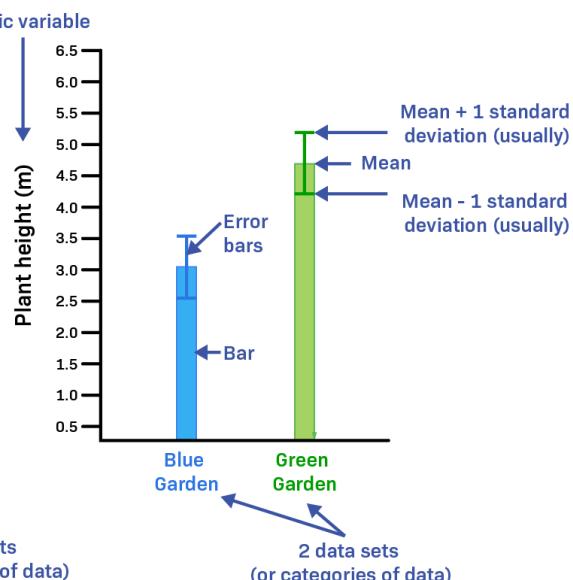
Ngoài ra, có thể muốn mô tả sự khác biệt trong mỗi danh mục bằng một loại biểu đồ khác, chẳng hạn như Violin plot. Mặc dù những biểu đồ này sẽ có nhiều yếu tố hơn để người

đọc phân tích, nhưng chúng cung cấp sự hiểu biết sâu sắc hơn về sự phân bố các giá trị trong mỗi nhóm.

(A) Barplot representing amounts



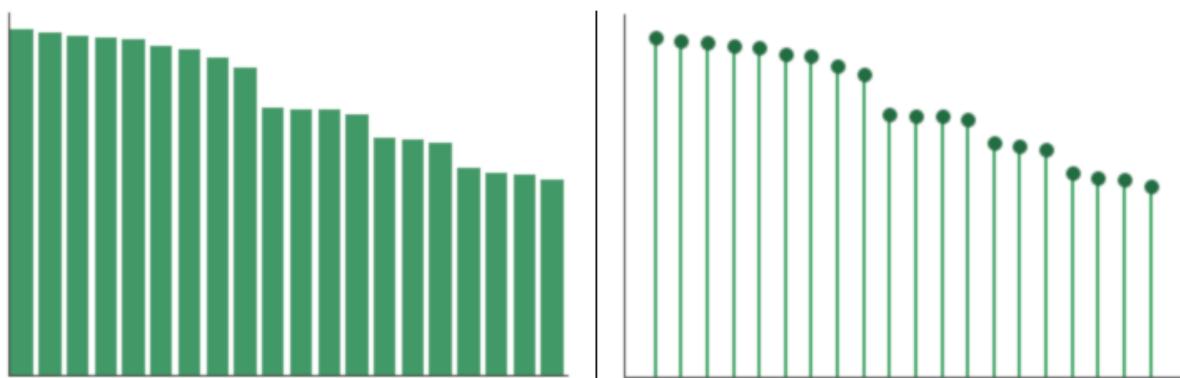
(B) Barplot summarizing data sets



Error bars cho biết độ lệch chuẩn của số tiền giao dịch đối với từng loại thanh toán. Sự thay đổi của thẻ tín dụng (credit) và thẻ ghi nợ (debit) thấp hơn so với các thẻ khác.

10.4.3.4.4. Lollipop chart

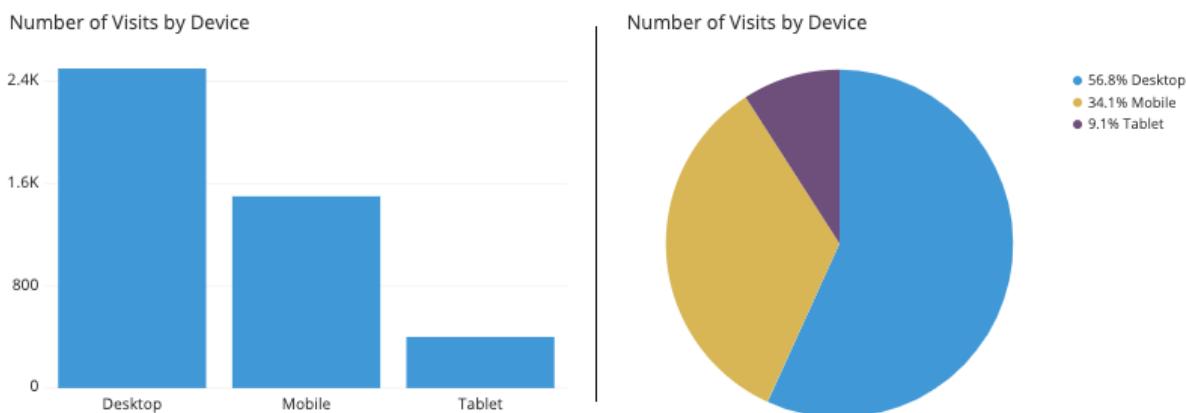
Một biến thể của Bar chart là Lollipop chart. Nó trình bày chính xác thông tin giống như Bar chart, nhưng có tính thẩm mỹ khác. Thay vì các thanh, có các đường có các dấu chấm tròn ở điểm cuối của chúng. Lollipop chart hữu ích nhất khi có nhiều danh mục và giá trị của chúng khá gần nhau. Bằng cách thay đổi hình thức thẩm mỹ của các giá trị được vẽ, nó có thể làm cho biểu đồ dễ đọc hơn nhiều.



10.4.3.5. Các đồ thị liên quan

10.4.3.5.1. Pie chart

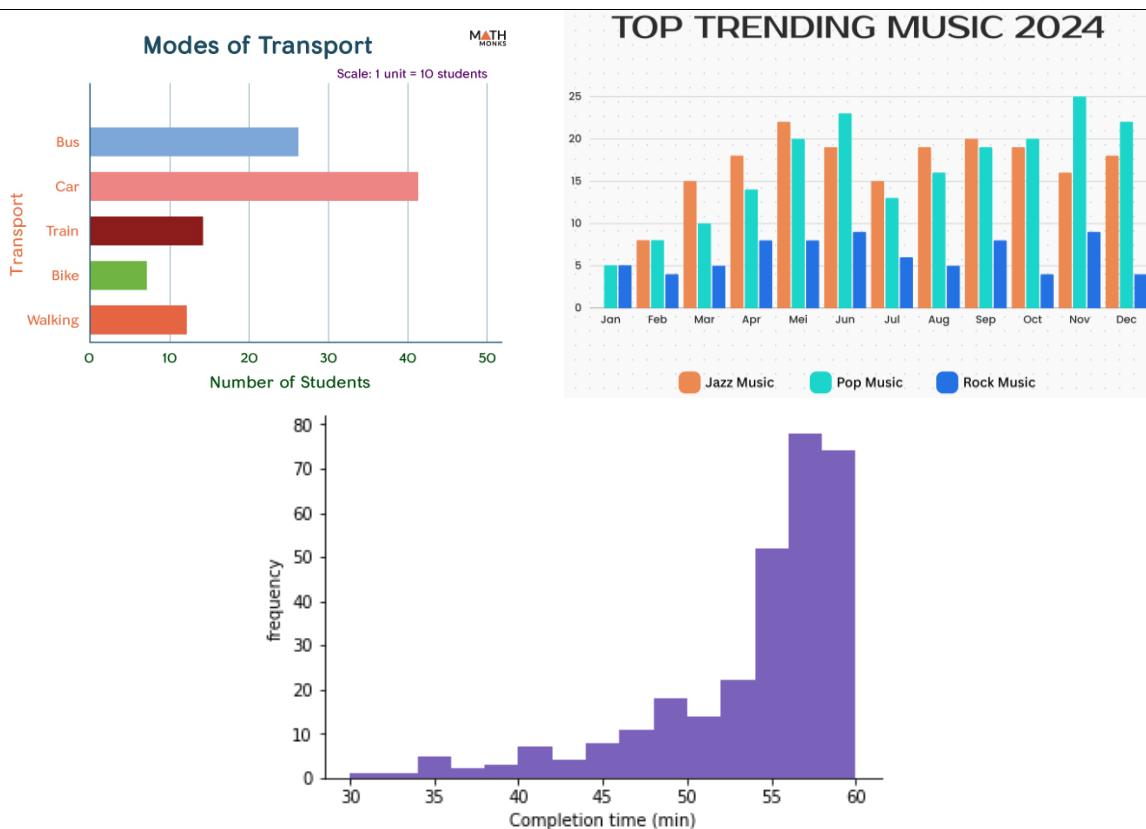
Nếu các giá trị trong Bar chart biểu thị các phần của tổng thể (tổng chiều dài thanh bằng tổng số điểm dữ liệu hoặc 100%), thì loại biểu đồ có thể thay thế là Pie chart. Mặc dù Pie chart chủ yếu chỉ hiển thị tỷ lệ % nhưng nếu cần mô tả việc phân chia từ bộ phận đến toàn bộ thì cũng nên suy nghĩ để chọn đồ thị này. Tuy nhiên, nếu cần so sánh về giá trị thì Bar chart vẫn giúp so sánh dễ dàng hơn.



10.4.3.5.2. Histogram

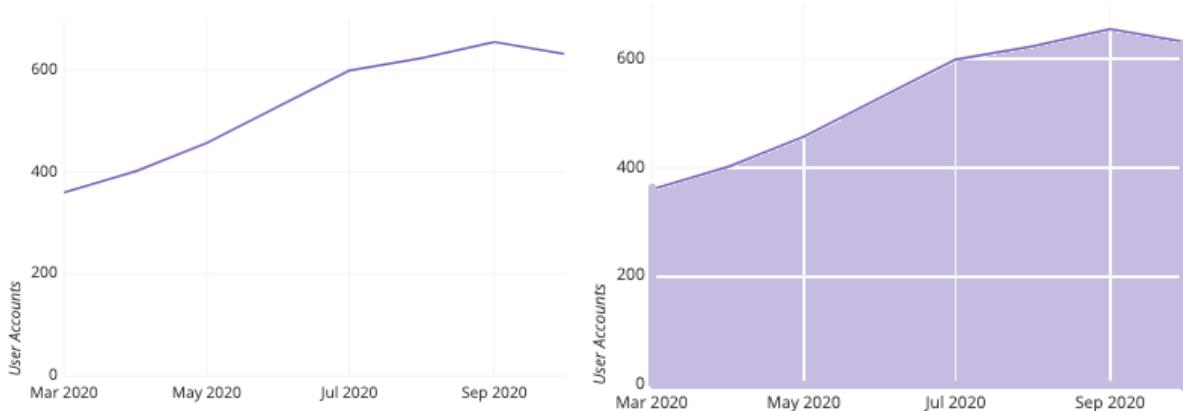
Histogram là họ hàng gần với Bar chart mô tả các giá trị tần suất. Mặc dù biến chính của Bar chart có tính chất phân loại, nhưng biến chính của Histogram lại liên tục và là số. Các thanh trong Bar chart thường được đặt ngay cạnh nhau để nhấn mạnh tính chất liên tục này: Bar chart thường có một khoảng cách giữa các thanh để nhấn mạnh tính chất phân loại của biến chính.

	<i>Bar chart</i>	<i>Histogram / Line</i>
<i>Kiểu dữ liệu của biến chính</i>	Danh mục (phân loại)	Liên tục (số)
<i>Khoảng cách giữa các bar</i>	Thường có để phân biệt các danh mục hoặc các khoảng thời gian	Không có để nhấn mạnh tính chất liên tục
<i>Màu sắc của bar</i>	Màu khác nhau cho mỗi danh mục	Dùng đồng nhất 1 màu



10.4.3.5.3. Line chart

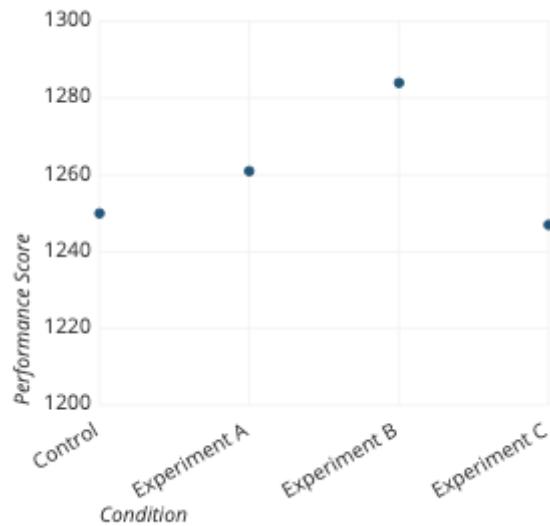
Đối với Bar chart mô tả số liệu thống kê tóm tắt, Line chart là biểu đồ tương đối gần nhất. Giống như mối quan hệ từ Bar chart đến Histogram, biến chính của biểu đồ đường thường là liên tục và là số, được nhấn mạnh bằng đường liên tục giữa các điểm. Việc tô bóng vùng giữa đường và đường cơ sở bằng 0 sẽ tạo ra Area chart, có thể được coi là sự kết hợp giữa Bar chart và Line chart.



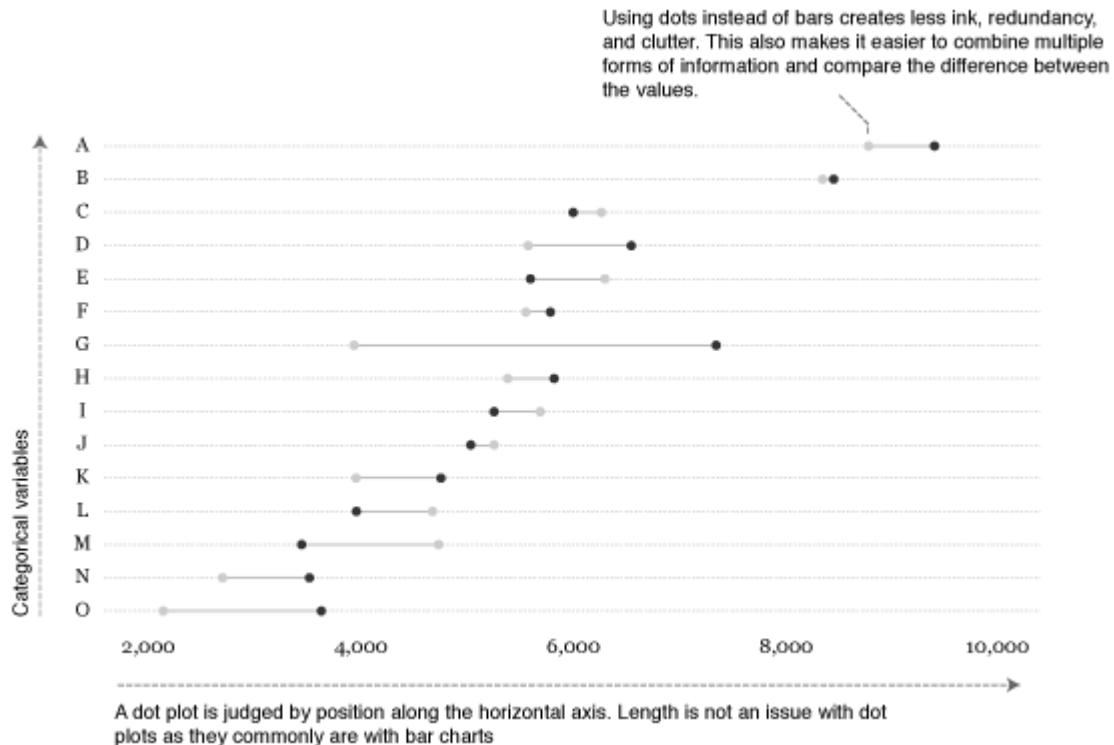
10.4.3.5.4. Dot plot

Ngoài ra, khi có số liệu thống kê tóm tắt về một biến chính được phân loại, có thể chọn Dot plot hoặc Cleveland Dot plot thay vì Bar chart. Dot plot về cơ bản là một Line chart không có đoạn thẳng nối giữa các điểm liền kề. Điều này giải phóng nó để sử dụng với các cấp độ phân loại, thay vì thể hiện dạng tiến triển liên tục. Ưu điểm lớn nhất của Dot plot so với Bar chart là các giá trị được biểu thị theo vị trí thay vì độ dài, vì vậy không nhất thiết cần đường cơ

sở bằng 0. Khi đường cơ sở cần thiết trên Bar chart cần trở nhận thức về những thay đổi hoặc khác biệt giữa các thanh thì Line chart hoặc Dot plot có thể là lựa chọn thay thế tốt.



Dot plot



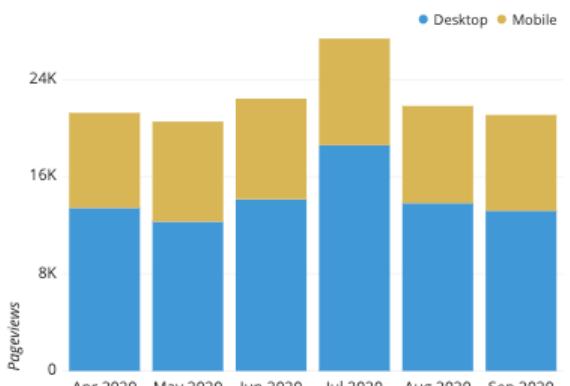
Cleveland Dot plot

10.4.3.5.5. Stacked bar chart và Grouped bar chart

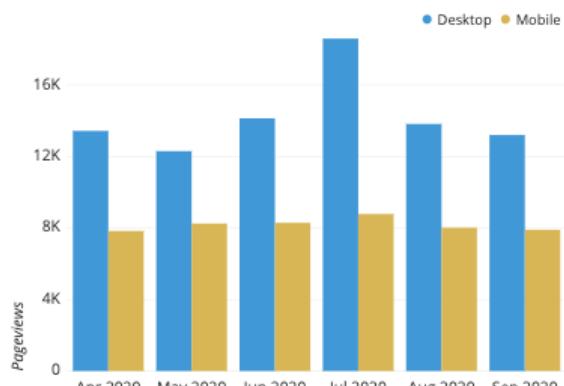
Bar chart có thể được mở rộng khi muốn thể hiện biến phân loại thứ hai để phân chia từng nhóm trong biến phân loại ban đầu. Nếu giá trị thanh mô tả tần suất nhóm thì biến phân loại thứ hai có thể chia số lượng của mỗi thanh thành các nhóm nhỏ. Có 2 biến thể:

- Stacked bar chart: các nhóm con trở thành các thành phần tạo nên nhóm chính.

- Grouped Bar chart: di chuyển các thanh của các nhóm con khác nhau về đường cơ sở.



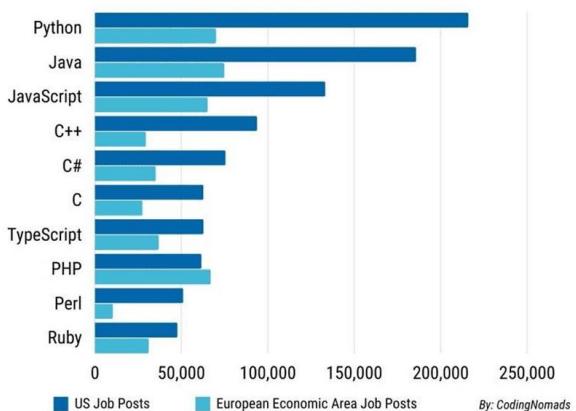
Stacked bar chart



Grouped Bar chart

Most in-demand programming languages of 2022

Based on LinkedIn job postings in the USA & Europe

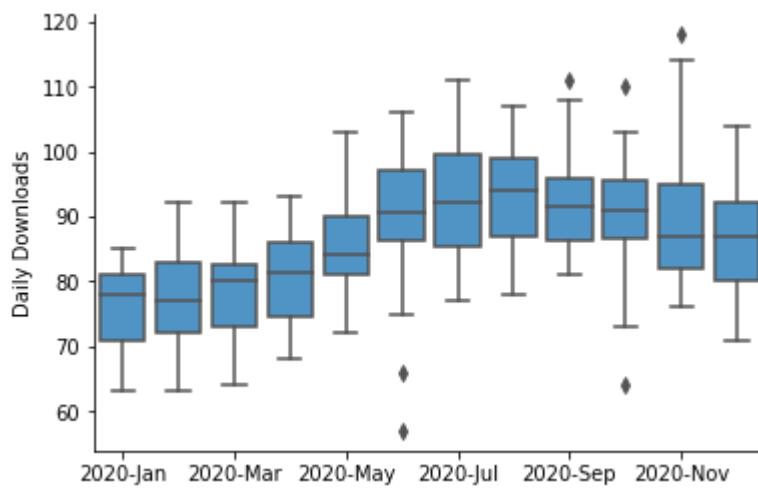


By: CodingNomads

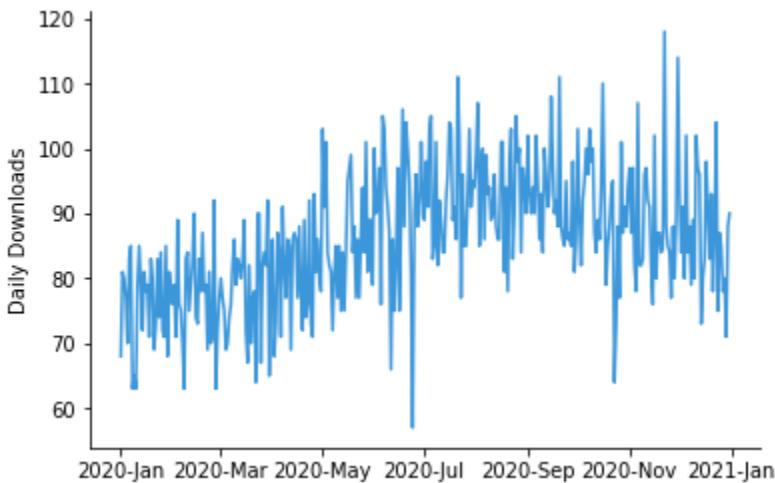
10.4.4. Boxplot (Box and Whisker Plot - biểu đồ hộp)

10.4.4.1. Giới thiệu

- Box plot sử dụng các hộp và đường để mô tả sự phân bố của một hoặc nhiều nhóm dữ liệu số.
- Là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu dựa trên tóm tắt 5 số (five summary number): giá trị tối thiểu (mean), tứ phân vị thứ 1 (Q1), trung vị (median), tứ phân vị thứ ba (Q3) và tối đa (max). Giới hạn hộp cho biết phạm vi của 50% dữ liệu ở giữa, với đường ở giữa đánh dấu giá trị trung bình. Các đường kéo dài từ mỗi hộp để nắm bắt phạm vi dữ liệu còn lại, với các dấu chấm được đặt qua các cạnh của đường để biểu thị các giá trị ngoại lệ (outliers).
- Mặc dù Box plot còn khá mới so với Histogram và Density plot, nhưng chúng có lợi thế là chiếm ít không gian hơn. Điều này rất hữu ích khi so sánh phân phối giữa nhiều nhóm trong datasets.

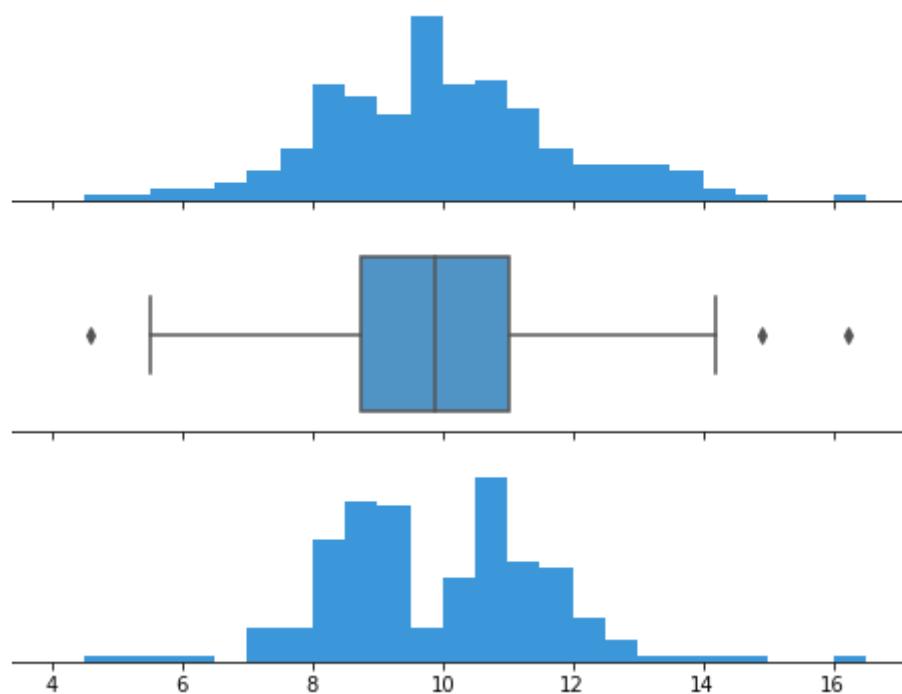


Box plot ví dụ ở trên hiển thị số lượt download hàng ngày cho một ứng dụng kỹ thuật số, được nhóm lại với nhau theo tháng. Từ biểu đồ này, có thể thấy rằng số lượt download đã tăng dần từ khoảng 75 lượt mỗi ngày trong tháng 1 lên khoảng 95 lượt mỗi ngày trong tháng 8. Có vẻ như số lượt tải xuống trung bình cũng giảm nhẹ trong tháng 11 và tháng 12. Điểm hiển thị những ngày có số lượt download khác thường: có hai ngày trong tháng 6 và một ngày trong tháng 10 có số lượt tải xuống thấp so với các ngày khác trong tháng. Boxplot cung cấp sự thể hiện rõ ràng hơn về xu hướng chung của dữ liệu so với Line chart tương đương.



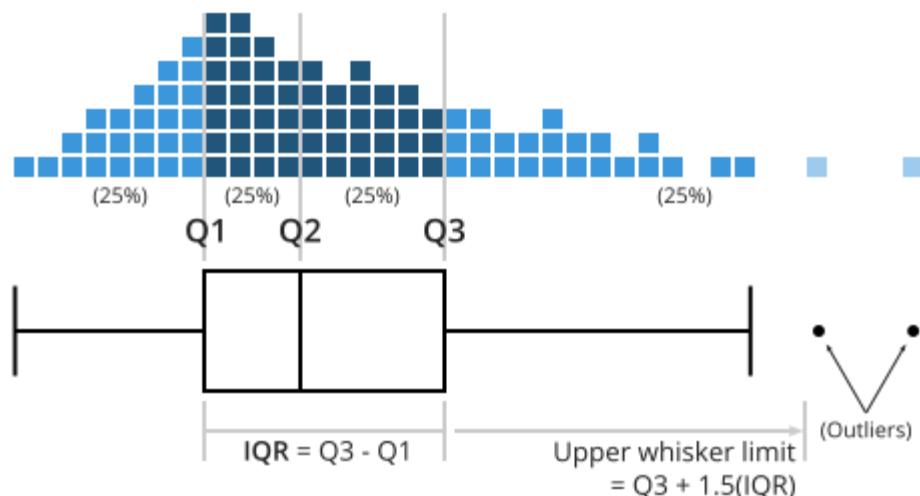
10.4.4.2. Sử dụng

- Những điểm chính khi sử dụng boxplot:
 - Dùng cho dữ liệu định lượng (thường là giá trị rời rạc), đặc biệt khi muốn so sánh chúng giữa nhiều nhóm
 - Khi cần theo dõi độ tập trung của các giá trị: thông tin chung về tính đối xứng, độ lệch, phương sai và ngoại lệ của một nhóm dữ liệu.
- Những hạn chế: Box plot
 - Hạn chế về mật độ dữ liệu mà nó có thể hiển thị.
 - Bỏ lỡ khả năng quan sát hình dạng chi tiết của phân phối, chẳng hạn như liệu có những điểm kỳ lạ trong phương thức phân phối hay không (số lượng 'bướu' -humps- hoặc đỉnh - peaks) và độ lệch (skew).



Các bộ dữ liệu đăng sau cả hai histogram tạo ra cùng một Box plot.

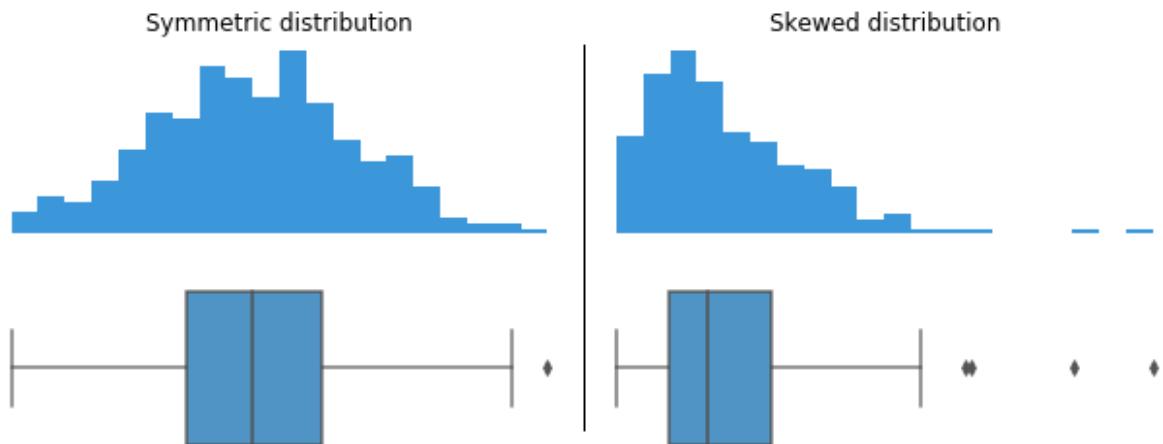
Việc xây dựng Box plot dựa trên các quartiles của tập dữ liệu hoặc các giá trị chia tập dữ liệu thành các phần bằng nhau. Tứ phân vị đầu tiên (Q1) lớn hơn 25% dữ liệu và nhỏ hơn 75% còn lại. Tứ phân vị thứ hai (Q2) nằm ở giữa, chia dữ liệu làm đôi. Q2 còn được gọi là trung vị. Tứ phân vị thứ ba (Q3) lớn hơn 75% dữ liệu và nhỏ hơn 25% còn lại. Trong Box plot, các đầu của hình hộp và đường tâm của nó đánh dấu vị trí của ba phần tư này.



Khoảng cách giữa Q3 và Q1 được gọi là phạm vi liên tứ phân vị (IQR) và đóng vai trò quan trọng trong việc các sợi râu (whiskers) kéo dài ra khỏi hộp dài bao nhiêu. Mỗi whisker kéo dài đến điểm dữ liệu xa nhất trong mỗi cánh nằm trong khoảng 1,5 lần IQR. Bất kỳ điểm dữ liệu nào xa hơn khoảng cách đó đều được coi là điểm ngoại lệ và được đánh dấu bằng dấu chấm. Có nhiều cách khác để xác định độ dài của whiskers, sẽ được thảo luận dưới đây.

Khi phân phối dữ liệu đối xứng, bạn có thể mong đợi trung vị nằm ở chính giữa hộp: khoảng cách giữa Q1 và Q2 phải giống như giữa Q2 và Q3. Các ngoại lệ phải xuất hiện đều ở

hai bên của hộp. Nếu phân phối bị lệch thì số trung vị sẽ không nằm ở giữa hộp mà thay vào đó sẽ lệch sang một bên. Cũng có thể nhận thấy sự mất cân bằng về độ dài của whiskers, trong đó một bên ngắn và không có ngoại lệ, còn bên kia có đuôi dài với nhiều ngoại lệ hơn.



- Cấu trúc dữ liệu để vẽ đồ thị

Date	...	Month	downloads
2020-01-30	...	2020-01	81
2020-01-31	...	2020-01	78
2020-02-01	...	2020-02	76
2020-02-02	...	2020-02	79
...

Các công cụ trực quan hóa thường có khả năng tạo ra các ô hình hộp từ một cột dữ liệu thô, chưa tổng hợp làm đầu vào; số liệu thống kê về các đầu hộp, râu và các phần ngoại lệ được tính toán tự động như một phần của quá trình tạo biểu đồ. Khi cần vẽ Box plot cho nhiều nhóm, các nhóm thường được biểu thị bằng các cột từ cột thứ hai trở đi, chẳng hạn như trong bảng trên.

10.4.4.3. Sử dụng box plot hiệu quả

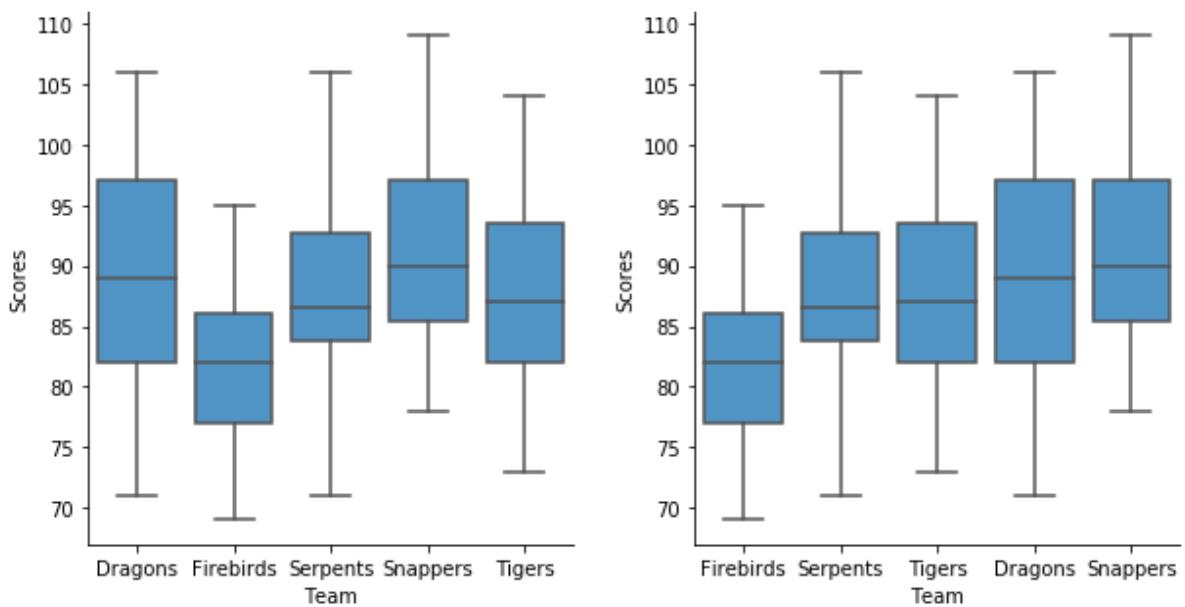
10.4.4.3.1. So sánh các nhóm

Box plot hoạt động tốt nhất khi cần thực hiện so sánh về phân phối giữa các nhóm. Chúng rất nhỏ gọn trong việc tóm tắt dữ liệu và dễ dàng so sánh các nhóm thông qua vị trí của hộp và vị trí của whiskers.

Sẽ ít dễ dàng hơn để chứng minh cho Box plot khi chỉ có phân phối của một nhóm để vẽ biểu đồ. Box plot chỉ cung cấp bản tóm tắt dữ liệu ở mức độ cao và thiếu khả năng hiển thị chi tiết về hình dạng phân phối dữ liệu. Nếu chỉ với một nhóm, nên chọn loại biểu đồ chi tiết hơn như Histogram hoặc Density curve.

10.4.4.3.2. Xem xét thứ tự các nhóm

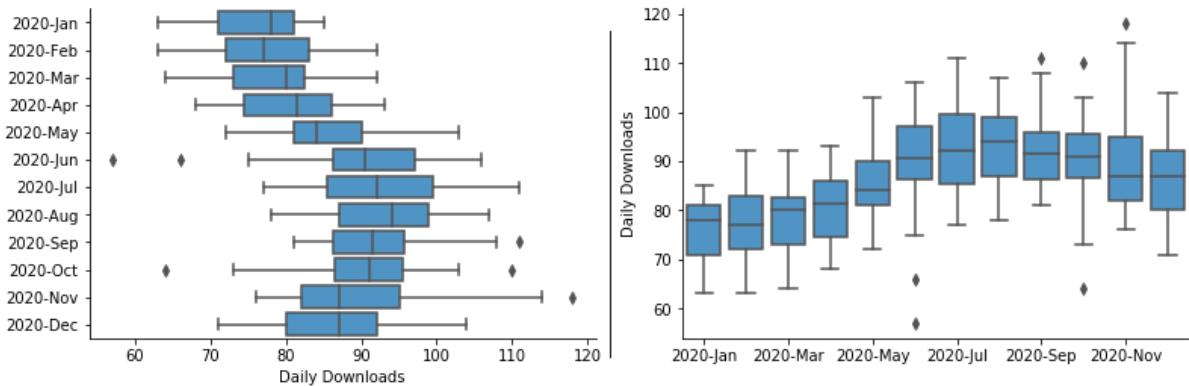
Nếu các nhóm được vẽ trong Box plot không có thứ tự cố hữu thì nên cân nhắc việc sắp xếp chúng theo thứ tự làm nổi bật các mô hình và cung cấp thêm hiểu biết cho người đọc. Một thứ tự chung cho các nhóm là sắp xếp chúng theo giá trị trung bình.



10.4.4.4. Các tùy chọn phổ biến được dùng với box plot

10.4.4.4.1. Sử dụng Vertical hay Horizontal box plot

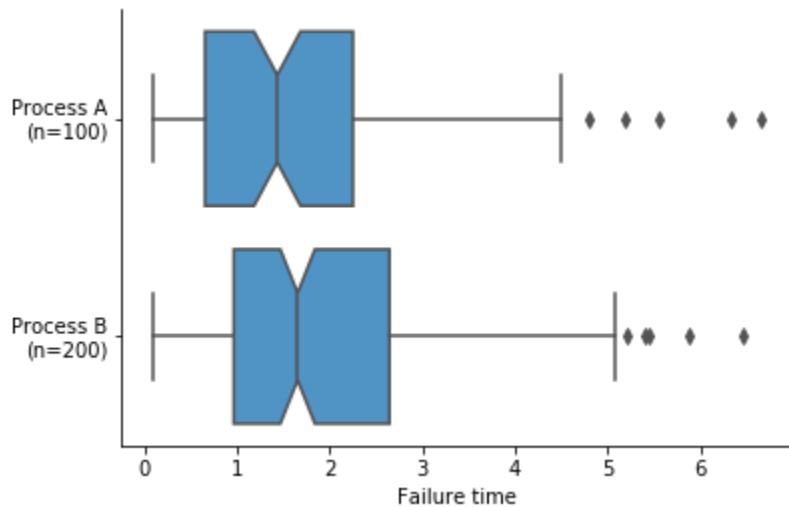
Có thể căn chỉnh ô vuông sao cho các hộp được đặt theo chiều dọc (với các nhóm được căn chỉnh theo trục ngang) hoặc theo chiều ngang (với các nhóm được căn chỉnh theo chiều dọc). Hướng ngang có thể là một định dạng hữu ích khi có nhiều nhóm cần vẽ hoặc nếu các tên nhóm đó dài (giúp hiển thị các tên danh mục dài mà không cần xoay hoặc cắt bớt). Mặt khác, hướng dọc có thể là định dạng tự nhiên hơn khi biến nhóm dựa trên đơn vị thời gian.



10.4.4.4.2. Chiều rộng (width) và rãnh (notches) của hộp có thể thay đổi

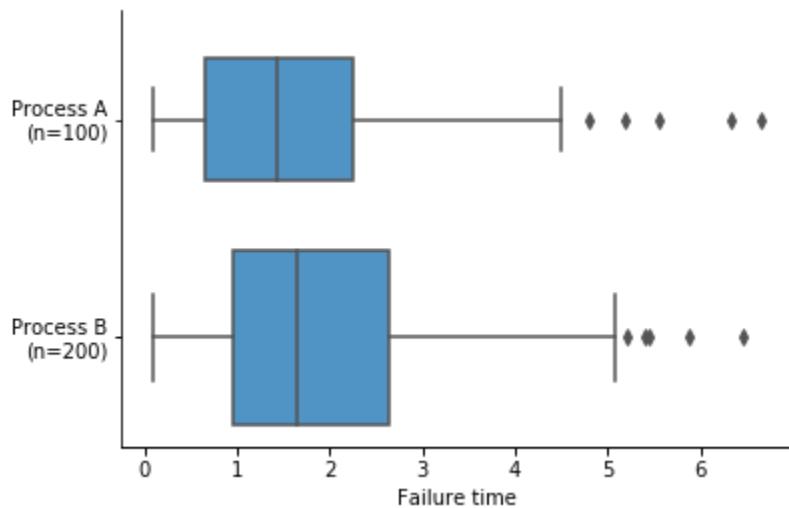
Một số công cụ trực quan nhất định bao gồm các tùy chọn để mã hóa thông tin thống kê bổ sung vào các ô hình hộp. Điều này rất hữu ích khi dữ liệu được thu thập đại diện cho các quan sát được lấy mẫu từ một quần thể lớn hơn.

Các khía được sử dụng để hiển thị các giá trị có khả năng xảy ra nhất đối với trung vị khi dữ liệu đại diện cho một mẫu. Khi thực hiện so sánh giữa các nhóm, có thể biết liệu sự khác biệt giữa các giá trị trung vị có ý nghĩa thống kê hay không dựa trên việc phạm vi của chúng có trùng nhau hay không. Nếu bất kỳ vùng khía nào trùng nhau thì không thể nói rằng các trung vị khác nhau về mặt thống kê; nếu chúng không trùng nhau thì có thể tin tưởng rằng các giá trị trung vị thực sự khác nhau.



Biểu đồ này gợi ý rằng Process B tạo ra các thành phần có thời gian sai sót tốt hơn (cao hơn), nhưng các khía cạnh chòng chéo cho thấy sự khác biệt về giá trị trung bình không có ý nghĩa thống kê.

Chiều rộng của hộp có thể được sử dụng làm chỉ báo về số lượng điểm dữ liệu thuộc mỗi nhóm. Chiều rộng của hộp thường được chia tỷ lệ theo căn bậc hai của số điểm dữ liệu, vì căn bậc hai tỷ lệ với độ không đảm bảo (tức là sai số chuẩn - error standard) về các giá trị thực. Vì chiều rộng của hộp diễn giải không phải lúc nào cũng trực quan nên một giải pháp thay thế khác là thêm chú thích với mỗi tên nhóm để ghi chú có bao nhiêu điểm trong mỗi nhóm.

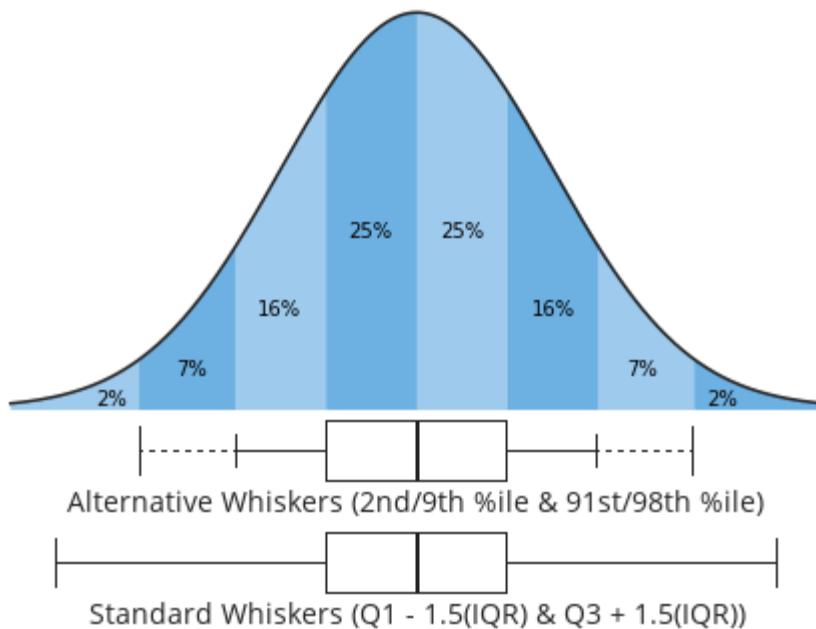


10.4.4.4.3. Phạm vi của Whisker và các giá trị ngoại lệ

Có nhiều cách để xác định độ dài tối đa của whisker kéo dài từ đầu các ô trong Box plot. Như đã lưu ý ở trên, cách truyền thống để kéo dài whisker là đến điểm dữ liệu xa nhất trong phạm vi 1,5 lần IQR tính từ mỗi đầu hộp. Ngoài ra, có thể đặt các dấu mốc ở các phân vị dữ liệu khác, chẳng hạn như cách các thành phần hộp nằm ở phân vị thứ 25, 50 và 75.

Các vị trí whisker thay thế phổ biến bao gồm phân vị thứ 9 và 91 hoặc phân vị thứ 2 và 98. Chúng dựa trên các đặc tính của phân phối chuẩn, liên quan đến ba phần tư trung tâm. Theo phân phối chuẩn, khoảng cách giữa phần trăm thứ 9 và thứ 25 (hoặc thứ 91 và thứ 75) phải có cùng kích thước với khoảng cách giữa phần trăm thứ 25 và thứ 50 (hoặc thứ 50 và thứ 75), trong khi khoảng cách giữa thứ 2 và thứ 25 (hoặc phần trăm thứ 98 và 75) phải gần bằng

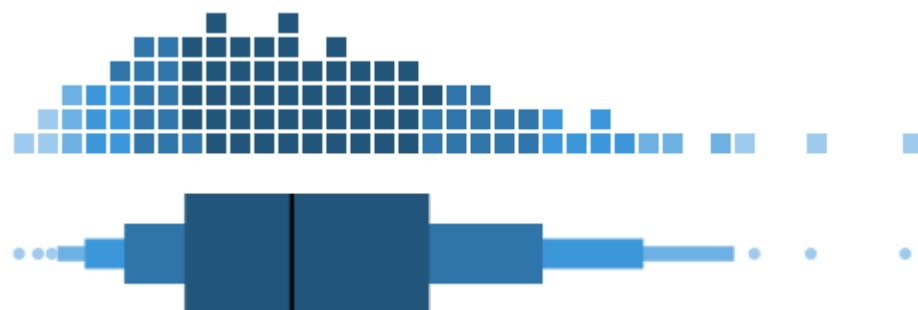
khoảng cách giữa phần trăm thứ 25 và 75. Điều này có thể giúp hỗ trợ khía cạnh tổng quan của Box plot, để biết liệu dữ liệu có đối xứng hay bị lệch hay không.



Khi sử dụng một trong những thông số kỹ thuật thay thế của whisker, nên ghi chú điều này trên hoặc gần ô để tránh nhầm lẫn với công thức chiều dài whisker truyền thống.

10.4.4.4.4. Letter-value plots

Được phát triển bởi Hofmann, Kafadar và Wickham, letter-value plots là phần mở rộng của Box plot tiêu chuẩn. Các ô có giá trị chữ cái sử dụng nhiều hộp để bao gồm các tỷ lệ ngày càng lớn hơn của tập dữ liệu. Hộp đầu tiên vẫn bao phủ 50% phần trung tâm, và hộp thứ hai kéo dài từ hộp thứ nhất đến một nửa diện tích còn lại (75% tổng thể, còn lại 12,5% ở mỗi đầu). Hộp thứ ba bao gồm một nửa diện tích còn lại (87,5% tổng thể, 6,25% còn lại ở mỗi đầu), v.v. cho đến khi quy trình kết thúc và các điểm còn lại được đánh dấu là ngoại lệ.



Letter-value plots được thúc đẩy bởi thực tế là khi thu thập được nhiều dữ liệu hơn, có thể đưa ra các ước tính ổn định hơn về các đuôi. Ngoài ra, nhiều điểm dữ liệu hơn có nghĩa là nhiều điểm dữ liệu hơn sẽ bị coi là ngoại lệ, dù hợp pháp hay không. Mặc dù letter-value plots vẫn còn thiếu một số chi tiết về phân phối như phương thức, nhưng đây có thể là cách kỹ lưỡng hơn để so sánh giữa các nhóm khi có sẵn nhiều dữ liệu.

10.4.4.5. Related plots

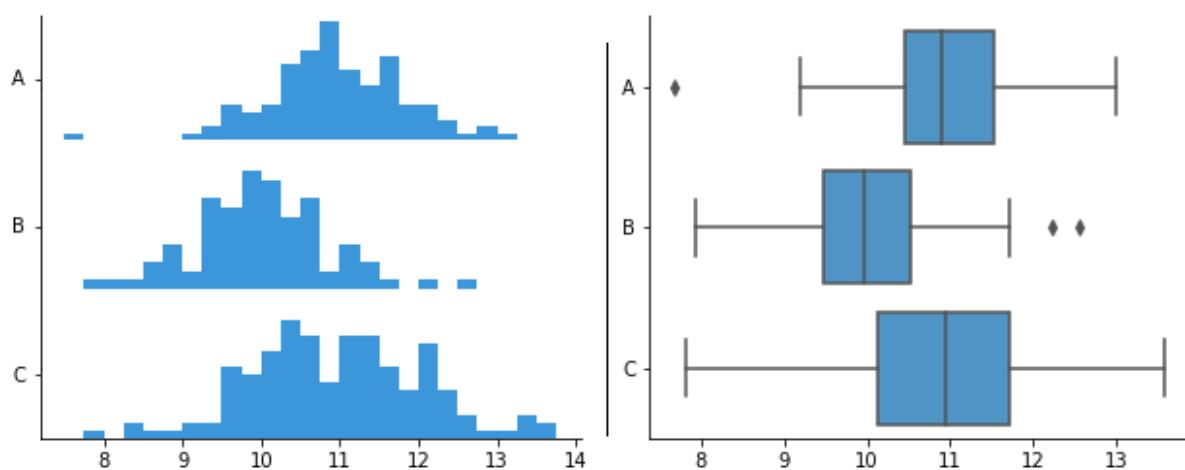
10.4.4.5.1. Histogram

As noted above, when you want to only plot the distribution of a single group, it is recommended that you use a [histogram](#) rather than a box plot. While a histogram does not include direct indications of quartiles like a box plot, the additional information about distributional shape is often a worthy tradeoff.

With two or more groups, multiple histograms can be stacked in a column like with a horizontal box plot. Note, however, that as more groups need to be plotted, it will become increasingly noisy and difficult to make out the shape of each group's histogram. In addition, the lack of statistical markings can make a comparison between groups trickier to perform. For these reasons, the box plot's summarizations can be preferable for the purpose of drawing comparisons between groups.

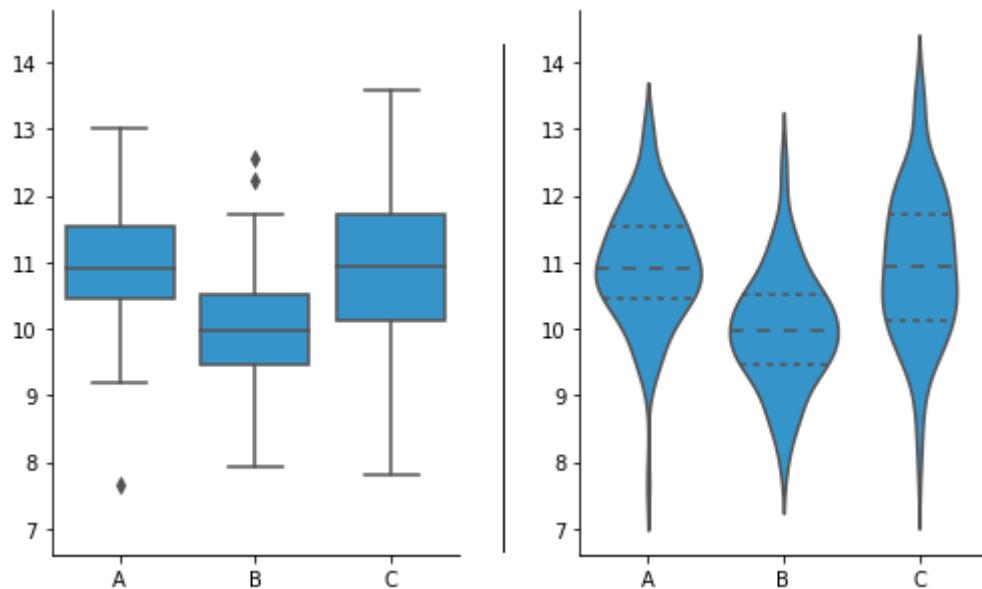
Như đã lưu ý ở trên, khi chỉ muốn vẽ biểu đồ phân bố của một nhóm duy nhất, nên sử dụng Histogram hơn là Box plot. Mặc dù Histogram không bao gồm các chỉ dẫn trực tiếp về các phần tử như Box plot, nhưng thông tin bổ sung về dạng phân bố thường là một sự đánh đổi xứng đáng.

Với hai hoặc nhiều nhóm, nhiều Histogram có thể được xếp chòng lên nhau trong một cột giống như Box plot ngang. Tuy nhiên, lưu ý rằng khi cần vẽ Histogram có nhiều nhóm hơn, việc xác định hình dạng biểu đồ của mỗi nhóm sẽ ngày càng trở nên “chật chội” và khó khăn hơn. Ngoài ra, việc thiếu các dấu hiệu thống kê có thể khiến việc so sánh giữa các nhóm trở nên khó thực hiện hơn. Vì những lý do này, việc tóm tắt bằng Box plot có thể phù hợp hơn cho mục đích đưa ra sự so sánh giữa các nhóm.



10.4.4.5.2. Violin plot

Một thay thế cho Box plot là Violin plot. Trong Violin plot, sự phân bố của mỗi nhóm được biểu thị bằng đường cong mật độ (density curve). Trong density curve, mỗi điểm dữ liệu không rơi vào một bin duy nhất như trong Histogram mà thay vào đó đóng góp một lượng diện tích nhỏ vào tổng phân bố. Violin plot là một cách nhỏ gọn để so sánh sự phân bố giữa các nhóm. Thông thường, các dấu hiệu bổ sung được thêm vào Violin plot cũng để cung cấp thông tin về Box plot tiêu chuẩn, nhưng điều này có thể làm cho biểu đồ kết quả trở nên “rõ ràm” hơn khi đọc.



10.4.4.5.3. Box-and-Whisker Plot

Box-and-Whisker Plot (Biểu đồ hình hộp và râu): Tương tự như biểu đồ hình hộp, nhưng thường được sử dụng thay thế cho nhau để mô tả cùng một biểu đồ.

10.4.5. Bullet graph

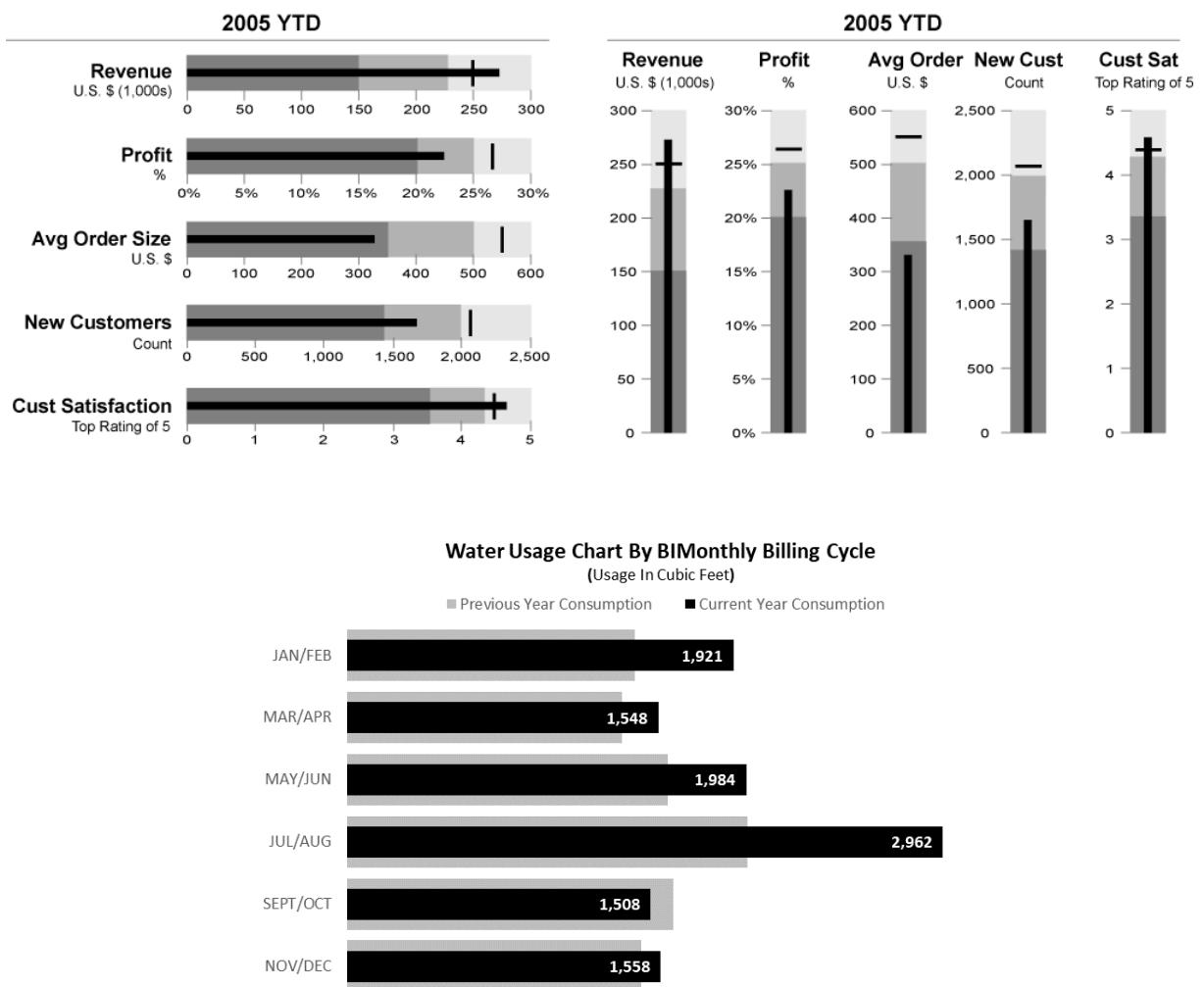
10.4.5.1. Giới thiệu

- Bullet Graph là một biến thể của bar graph được phát triển bởi Stephen few. Dường như lấy cảm hứng từ biểu đồ nhiệt kế (thermometer charts) truyền thống và thanh tiến trình được tìm thấy trong nhiều dashboards (bảng thông tin), Bullet Graph đóng vai trò thay thế cho gauge chart. Bullet Graph được phát triển để khắc phục các vấn đề cơ bản của gauge chart: do chúng thường hiển thị quá ít thông tin, cần quá nhiều không gian và chưa đầy những trang trí vô dụng và còn gây mất tập trung.
- Bullet Chart gồm 3 phần:
 - Một dòng hiển thị giá trị mục tiêu
 - Thanh giữa hiển thị giá trị thực tế
 - Thanh màu hiển thị các chỉ số hiệu suất

10.4.5.2. Sử dụng

Được sử dụng để hiển thị dữ liệu hiệu suất, Bullet Graph có một thước đo chính, duy nhất (ví dụ: doanh thu tính từ đầu năm hiện tại), so sánh thước đo đó với một hoặc nhiều thước đo khác để làm phong phú thêm ý nghĩa của nó (ví dụ: so với mục tiêu) bằng cách sử dụng màu (chẳng hạn như dùng màu sáng cho mức độ kém và độ đậm của màu tăng dần cho các mức tăng khác như đạt yêu cầu, khá, tốt, ...).

10.4.5.3. Minh họa



10.4.6. Brainstorm – (hay Sơ đồ tư duy – mind map)

10.4.6.1. Giới thiệu

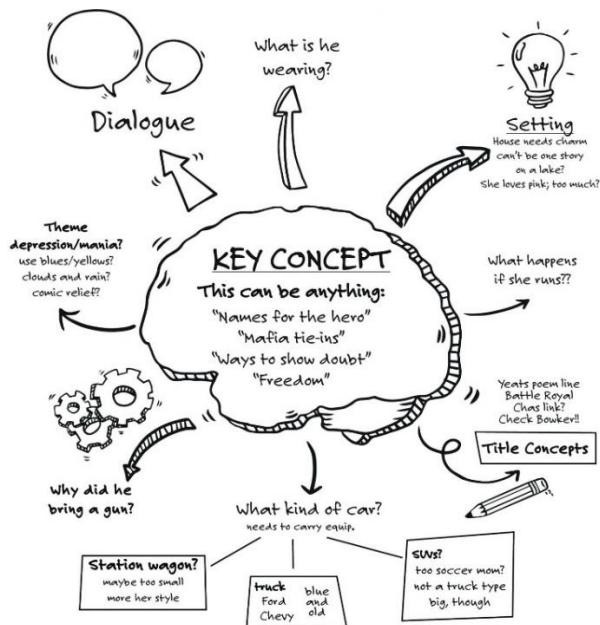
- Brainstorm là sơ đồ được sử dụng để thể hiện các ý tưởng, từ ngữ, hình ảnh và các khái niệm liên kết lại với nhau. Bên cạnh đó, Brainstorm cũng là một công cụ để hình thành việc lên ý tưởng, tìm kiếm sự liên kết, phân loại ý tưởng, tổ chức thông tin,...

10.4.6.2. Sử dụng

- Brainstorm thường được sử dụng ở giai đoạn đầu của dự án và hoạt động như một hình thức ghi chú. Chúng cũng hữu ích trong việc cộng tác và xây dựng tinh thần nhóm.
- Thông thường, người ta thường tạo Brainstorm theo các bước:
 - (i). Viết tiêu đề của dự án và gói gọn chúng lại trong một hình tròn hoặc một đám mây ở ngay giữa trang.
 - (ii). Ghi các từ khóa chính có liên quan đến chủ đề bao quanh đám mây chứa chủ đề chính. 3.
 - (iii). Đối với mỗi từ khóa chính, có thể vẽ nhiều nhánh con khác nhau tượng trưng cho các từ khóa phụ cấp 1 có liên quan/ảnh hưởng đến chủ đề chính

(iv). Có thể lặp lại bước 3 nhiều lần cho các từ khóa phụ để có các từ khóa phụ cấp 2, cấp 3,

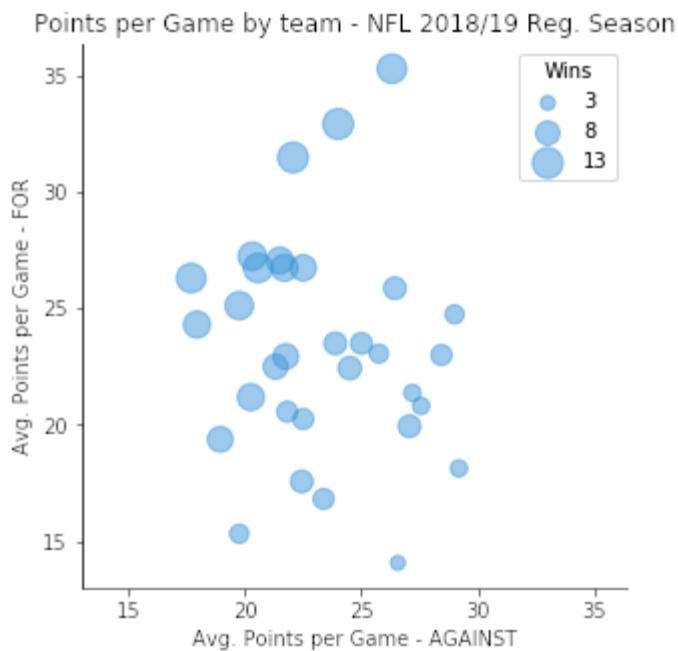
10.4.6.3. Minh họa

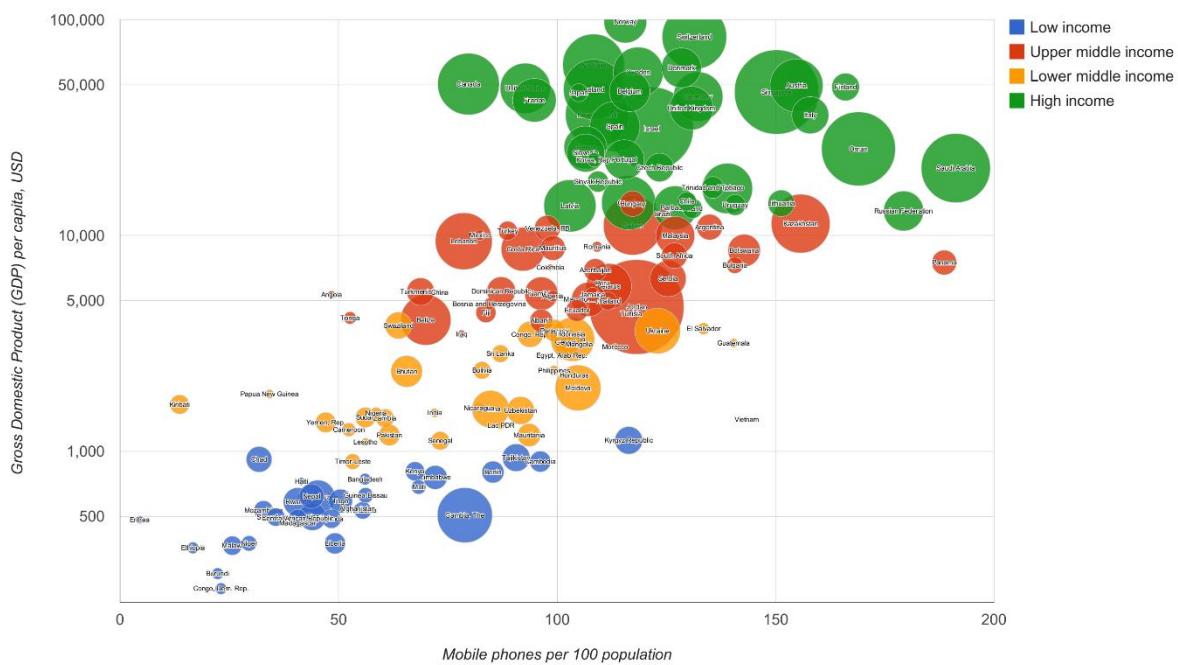


10.4.7. Bubble chart

10.4.7.1. Giới thiệu

Bubble chart (còn gọi là bubble plot) là phần mở rộng của Scatter plot được sử dụng để xem xét mối quan hệ giữa ba biến số là giá trị x, giá trị y và giá trị z (kích cỡ). Mỗi dấu chấm trong Bubble chart tương ứng với một điểm dữ liệu duy nhất và giá trị của các biến cho mỗi điểm được biểu thị bằng vị trí ngang, vị trí dọc và kích thước điểm.





Ví dụ ở trên mô tả số điểm ghi được trong mỗi trận đấu của các đội trong mùa giải thường lệ của Giải bóng đá quốc gia năm 2018. Mỗi bong bóng thể hiện thành tích của một đội. Vị trí nằm ngang của bong bóng ghi lại số điểm trung bình ghi được của đội đó trong mỗi trận đấu và vị trí thẳng đứng ghi lại số điểm trung bình mà đội đó ghi được trong mỗi trận đấu. Kích thước của mỗi bong bóng cho biết số trận thắng mà mỗi đội giành được, bong bóng lớn hơn tương ứng với tỷ lệ thắng cao hơn. (Hòa có giá trị bằng một nửa chiến thắng.)

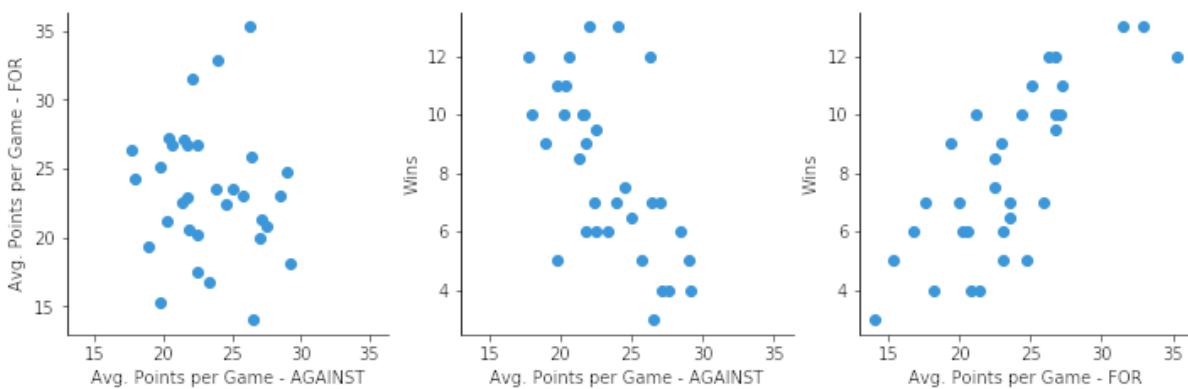
Từ biểu đồ trên, có thể thấy rằng có nhiều sự khác biệt về số điểm ghi được của các đội so với đội thủ của họ, nhưng không có mối tương quan đặc biệt chặt chẽ giữa hai điều này. Thay vào đó, điểm rút ra chính của biểu đồ đến từ biến số thứ ba: khi các đội ghi được nhiều điểm hơn và để đội thủ của họ ghi ít điểm hơn (về phía trên bên trái), họ sẽ giành được nhiều chiến thắng hơn, như người ta có thể mong đợi một cách tự nhiên.

The name “Bubble chart” is sometimes used to refer to a different chart type, the packed circle chart. This is a completely different chart type that will be discussed briefly towards the end of the article.

Tên “Bubble chart” đôi khi được dùng để chỉ một loại biểu đồ khác là Packed circle chart. Đây là một loại biểu đồ hoàn toàn khác sẽ được thảo luận ngắn gọn ở phần sau.

10.4.7.2. Sử dụng

- Giống như Scatter plot, Bubble chart chủ yếu được sử dụng để mô tả và hiển thị mối quan hệ giữa các biến số. Tuy nhiên, việc bổ sung kích thước điểm đánh dấu làm thay đổi nguyên tắc so sánh giữa ba biến thay vì chỉ hai biến.
- Trong một Bubble chart, có thể thực hiện ba so sánh theo cặp khác nhau (X so với Y, Y so với Z, X so với Z), cũng như so sánh ba chiều tổng thể. Nó sẽ yêu cầu nhiều Scatter plot hai biến để có được cùng một số thông tin chi tiết; thậm chí sau đó, việc suy ra mối quan hệ ba chiều giữa các điểm dữ liệu sẽ không trực tiếp như trong Bubble chart.



Ba biểu đồ phân tán ở trên hiển thị cùng dữ liệu với Bubble chart mẫu ban đầu. Mặc dù việc lấy số trận thắng cụ thể cho mỗi đội từ chuỗi biểu đồ này dễ dàng hơn nhưng mối quan hệ giữa cả ba biến số không được nêu rõ ràng như trong Bubble chart.

- Nhược điểm: khi có quá nhiều bong bóng sẽ làm cho biểu đồ trở nên khó đọc hơn.
- Cấu trúc dữ liệu dùng để vẽ biểu đồ

Avg_Points_Against	Avg_Points_For	Wins
26.56	14.06	3
26.44	25.88	7
17.94	24.31	10
23.38	16.81	6
...

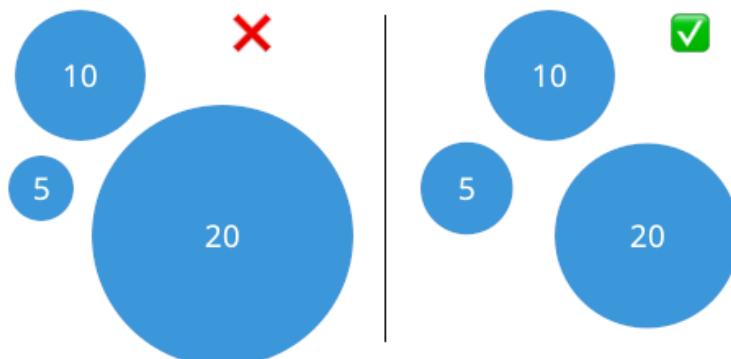
Bubble chart được tạo từ bảng dữ liệu có ba cột. Hai cột sẽ tương ứng với vị trí ngang và dọc của từng điểm, trong khi cột thứ ba sẽ cho biết kích thước của từng điểm. Một điểm (bubble) sẽ được vẽ cho mỗi hàng trong bảng.

10.4.7.3. Sử dụng bubble chart một cách hiệu quả

10.4.7.3.1. Chia tỷ lệ diện tích bong bóng theo giá trị

Một sai lầm dễ mắc phải là chia tỷ lệ đường kính hoặc bán kính của điểm theo giá trị của biến thứ ba. Khi thực hiện kiểu chia tỷ lệ này, một điểm có giá trị gấp đôi điểm khác sẽ có diện tích gấp bốn lần, làm cho giá trị của nó trông lớn hơn nhiều so với giá trị thực tế được đảm bảo.

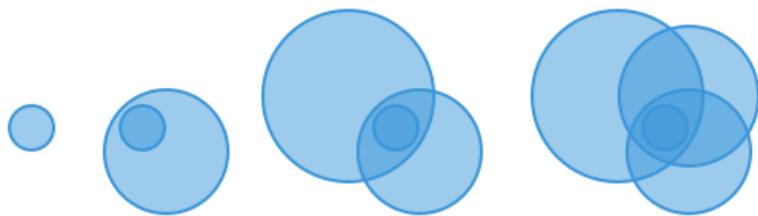
Thay vào đó, hãy đảm bảo rằng diện tích của bong bóng tương ứng với giá trị của biến thứ ba. Trong trường hợp tương tự như trên, một điểm có giá trị gấp đôi điểm khác sẽ có $\sqrt{2} = 1,41$ lần đường kính hoặc bán kính để diện tích của nó gấp đôi điểm nhỏ hơn.



Tùy thuộc vào cách tạo Bubble chart, có thể cần phải chia tỷ lệ dữ liệu của mình để tính toán cách các giá trị dữ liệu được ánh xạ tới kích thước điểm. Nhiều công cụ trực quan sẽ tự động khớp giá trị với diện tích, nhưng hãy cẩn thận với những trường hợp trong đó giá trị được khớp với đường kính hoặc bán kính.

10.4.7.3.2. Giới hạn số điểm để vẽ

Bubble chart thường được vẽ với độ trong suốt trên các điểm vì sự chồng chéo dễ xảy ra hơn nhiều so với khi tất cả các điểm đều có kích thước nhỏ. Sự chồng chéo này cũng có nghĩa là có những hạn chế về số lượng điểm dữ liệu có thể được vẽ trong khi vẫn giữ cho biểu đồ có thể đọc được.



Nếu không có màu trong suốt, điểm dữ liệu nhỏ hơn sẽ không thể nhìn thấy được so với điểm dữ liệu lớn hơn.

Không có bất kỳ hướng dẫn cứng nào về việc liệu tập dữ liệu có phù hợp với Bubble chart hay không, nhưng đó là điểm cần lưu ý khi tạo Bubble chart. Nếu có vẻ như có quá nhiều biểu đồ thì nên suy nghĩ về cách tóm tắt dữ liệu hoặc chọn một loại biểu đồ khác để thể hiện dữ liệu của mình. Việc giảm kích thước bong bóng có thể giúp tạo ra sự tách biệt vật lý giữa các điểm, nhưng làm như vậy cũng sẽ khiến việc đọc giá trị từ kích thước bong bóng trở nên khó khăn hơn.

10.4.7.3.3. Bao gồm chú thích

Nên đưa chú thích vào biểu đồ để hiển thị các kích thước bong bóng khác nhau tương ứng với các giá trị của biến thứ ba như thế nào. Khá dễ dàng để đánh giá và so sánh các giá trị dựa trên độ dài và vị trí ngang hoặc dọc nhau các dấu tích trên các trục. Chìa khóa dành cho kích thước bong bóng có cùng mục đích như các dấu kiểm dành cho biến thứ ba.

Nếu đang sử dụng một ứng dụng trực quan hóa có khả năng tương tác, nên bật tính năng này để các giá trị hiển thị khi các điểm riêng lẻ được chọn hoặc khi di chuột qua. Khi in, nên dán nhãn các điểm chính để cải thiện khả năng giao tiếp của Bubble chart.

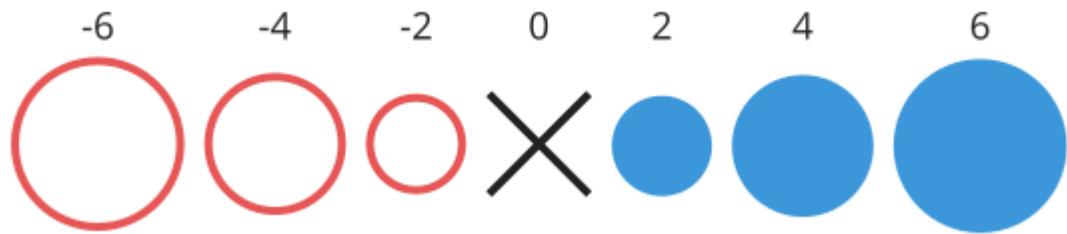
10.4.7.3.4. Trình bày một xu hướng rõ ràng

Nếu sử dụng Bubble chart để trình bày thông tin cho người khác, hãy đảm bảo rằng biểu đồ đó có thể thể hiện xu hướng rõ ràng bằng cách sử dụng kích thước điểm làm chỉ báo giá trị. Khi phát triển biểu đồ, hãy thử nghiệm thử tự các biến được vẽ. Hai biến quan trọng nhất hoặc mối quan hệ quan trọng nhất sẽ kết thúc trên trục dọc và trục ngang. Tránh sử dụng Bubble chart nếu biến thứ ba không đóng góp đáng kể vào câu chuyện được biểu đồ kể và thay vào đó hãy sử dụng các biểu đồ bổ sung, đơn giản hơn.

10.4.7.3.5. Kết hợp các giá trị âm

Nếu một biến nhận giá trị âm thì nó không thể được gán trực tiếp cho kích thước điểm dưới dạng mã hóa: xét cho cùng, làm sao một hình dạng có thể có vùng âm? Thông tin bổ sung cần được mã hóa thành kích thước hình dạng để biểu thị các giá trị âm. Ví dụ: có thể có các

vòng tròn được tô đầy biểu thị các giá trị dương và các vòng tròn không được lấp đầy biểu thị các giá trị âm. Một cách khác, bạn có thể có các điểm dương ở một màu và các điểm âm ở một màu khác, riêng biệt.



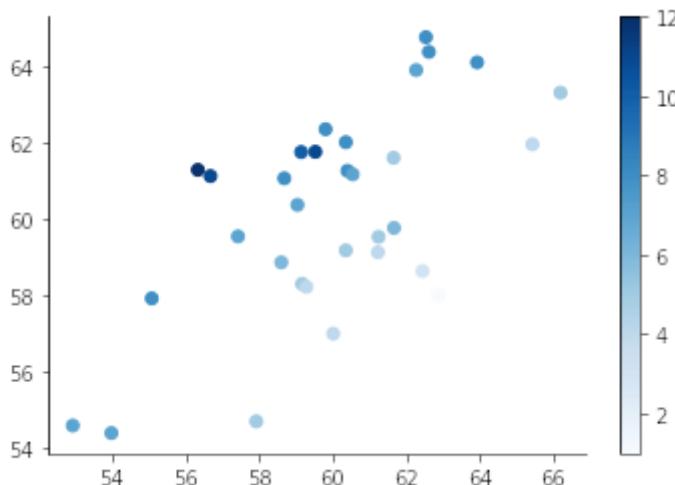
Tốt hơn hết nên kiểm tra xem các mã hóa như vậy có hợp lý ngay từ đầu hay không: thay vào đó, biến có thể tốt hơn nên được gán cho một trong các trục vị trí! Cố gắng tránh mã hóa các giá trị âm bằng bong bóng trừ khi nó thực sự tăng thêm giá trị cho biểu đồ.

10.4.7.4. Những biểu đồ liên quan

10.4.7.4.1. Scatter plot

Bubble chart được xây dựng dựa trên Scatter plot làm cơ sở, chỉ với việc bổ sung biến thứ ba thông qua kích thước điểm. Tuy nhiên, điều đáng nói là các biến thứ ba có thể được thêm vào Scatter plot thông qua các mã hóa điểm khác. Phổ biến nhất trong số này là màu sắc. Khi biến thứ ba mang tính phân loại (lấy các giá trị riêng biệt có thể được sắp xếp theo thứ tự hoặc không), có thể chỉ định màu sắc riêng biệt cho từng loại điểm. Trên thực tế, có thể sử dụng màu sắc làm biến số thứ tư kết hợp với kích thước điểm, nhưng điều này nên được sử dụng cẩn thận vì nó có thể dẫn đến tình trạng quá tải thông tin – những cảnh báo trước đó về việc trình bày một xu hướng rõ ràng sẽ được tăng cường đáng kể với biến số thứ tư.

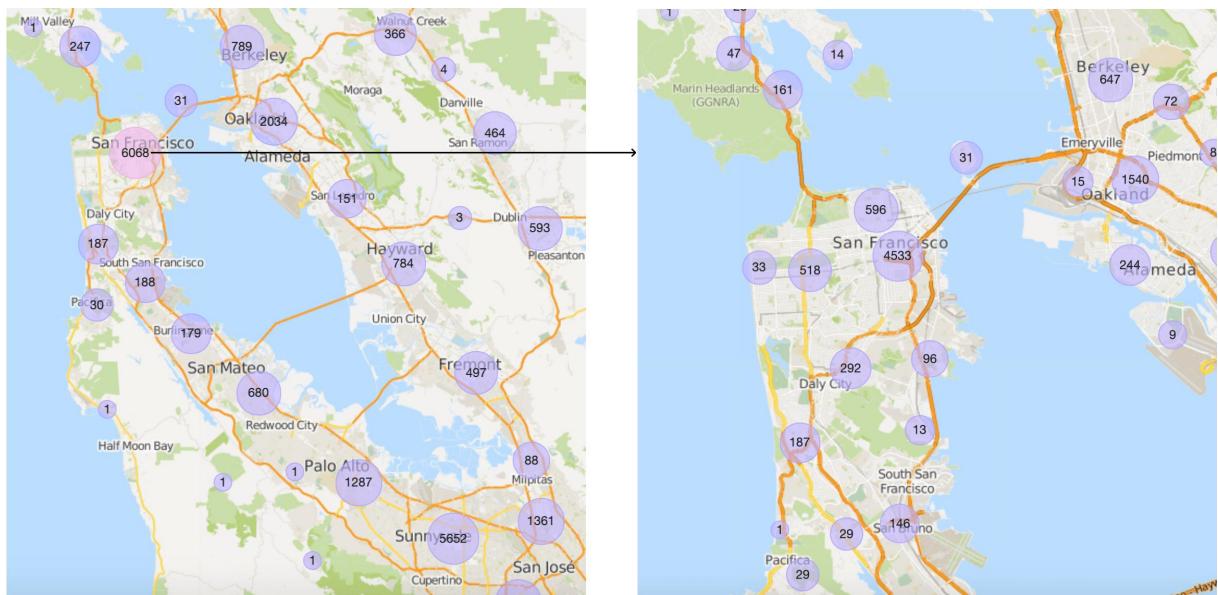
Màu sắc cũng có thể được sử dụng làm mã hóa cho các biến số. Nếu có bảng màu trong đó các màu có mối quan hệ liên tục (ví dụ: sáng và tối), khi đó có thể sử dụng màu để biểu thị giá trị cho biến thứ ba, thay vì kích thước. Lưu ý rằng nhận thức về giá trị dựa trên màu sắc có những hạn chế tương tự như việc sử dụng kích thước, do đó, khi sử dụng màu sắc cũng như đối với kích thước điểm cũng cần bổ sung thêm chú thích cho biểu đồ.



10.4.7.4.2. Bubble map

Nếu hai biến vị trí biểu thị tọa độ địa lý (tức là vĩ độ và kinh độ), có thể phủ các bong bóng lên bản đồ ở chế độ nền và có được bản đồ bong bóng (Bubble map). Bubble map là một

phản mở rộng thú vị của Scatter map có thể giúp giải quyết các vấn đề tiềm ẩn sau này liên quan đến việc vẽ quá mức. Nếu Scatter map có quá nhiều điểm trong một khu vực đến mức không thể dễ dàng nhìn thấy số lượng của chúng, có thể hoán đổi chúng bằng một bong bóng duy nhất báo cáo tổng số điểm trong khu vực đó.

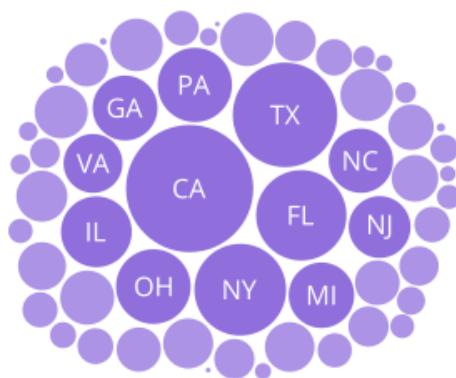


Mặc dù ví dụ này từ tìm kiếm cho thuê của Craigslist không có tỷ lệ bong bóng tiêu chuẩn nhưng nó vẫn chứng minh cách bong bóng có thể giúp tóm tắt các điểm có thể có mật độ dày đặc.

10.4.7.4.3. Packed circle chart

Packed circle charts (còn gọi là Circular packing, Bubble cloud) là loại biểu đồ có thể trông giống như Bubble chart trên bề mặt của nó. Mặc dù các bong bóng trong Packed circle charts biểu thị các giá trị số hoặc tần suất như trước đây nhưng đây là biến số duy nhất hiện diện: các bong bóng được nhóm lại với nhau theo một cách sắp xếp dày đặc mà không có bất kỳ trục vị trí thực nào.

2010 US Population by State / Territory



Theo một cách nào đó, có thể coi Packed circle charts như một Bar chart được tạo thành từ các đĩa (discs). Tuy nhiên, điều này bộc lộ điểm yếu của Packed circle charts: giống như Bubble chart, rất khó để có được giá trị chính xác hoặc thứ hạng từ các kích thước bong bóng không có thứ tự. Thông thường, khi đó tốt hơn hết nên sử dụng Bar chart, Lollipop chart hoặc Dot plot để truyền tải thông tin do chúng sử dụng vị trí để mã hóa giá trị. Một lợi thế mà các vòng tròn đóng gói có là nếu có nhiều nhóm để vẽ, việc đóng gói hình tròn có thể nhỏ gọn hơn

nhiều so với việc hiển thị từng danh mục trong một hàng dài. Tuy nhiên, cũng có thể gộp các giá trị nhỏ hơn vào nhóm “khác” (others) để giảm khoảng trống trong biểu đồ.

Thông thường, việc đóng gói hình tròn có xu hướng xuất hiện trong bối cảnh phân cấp, trong đó các vòng tròn nhỏ hơn được đặt bên trong các vòng tròn lớn hơn để cho thấy cách tổng thể được chia thành các phần ở nhiều cấp độ phân chia. Ngay cả ở đây, dạng hình tròn cho tỷ lệ có phần kém hiệu quả so với các loại biểu đồ khác như Treemap, vì vậy ưu điểm của Packed circle charts là ở tính thẩm mỹ hơn là tính thực tế.

10.4.8. Candlestick Chart

Candlestick Chart: Chủ yếu được sử dụng trong thị trường tài chính để mô tả biến động giá của chứng khoán, công cụ phái sinh hoặc tiền tệ.

10.4.9. Circular Packing Chart

Circular Packing Chart (Biểu đồ đóng gói hình tròn): Trực quan hóa dữ liệu phân cấp bằng cách sử dụng các vòng tròn; kích thước của mỗi vòng tròn tỷ lệ thuận với giá trị của nó.

10.4.10. Chord diagram

10.4.10.1. Giới thiệu

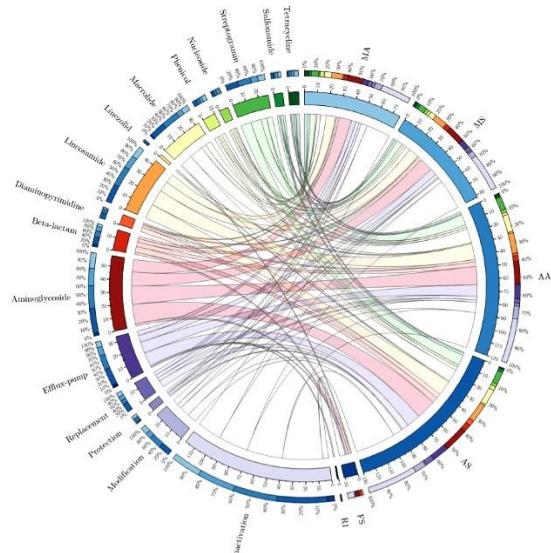
Chord diagram là một phương pháp đồ họa để hiển thị mối quan hệ giữa các dữ liệu trong ma trận. Dữ liệu được sắp xếp triệt để xung quanh một vòng tròn với mỗi quan hệ giữa các điểm dữ liệu thường được vẽ dưới dạng các vòng cung kết nối dữ liệu.

Định dạng này có thể mang tính thẩm mỹ, khiến nó trở thành một lựa chọn phổ biến trong thế giới trực quan hóa dữ liệu.

10.4.10.2. Sử dụng

Công dụng chính của Chord diagram là hiển thị các luồng hoặc kết nối giữa một số thực thể (được gọi là nút). Mỗi thực thể được thể hiện bằng một mảnh (thường có màu hoặc có hoa văn) đọc theo chu vi của vòng tròn. Các cung được vẽ giữa các thực thể để thể hiện các dòng chảy (và trao đổi trong kinh tế). Độ dày của dòng chảy tỷ lệ thuận với tầm quan trọng của dòng chảy.

10.4.10.3. Minh họa

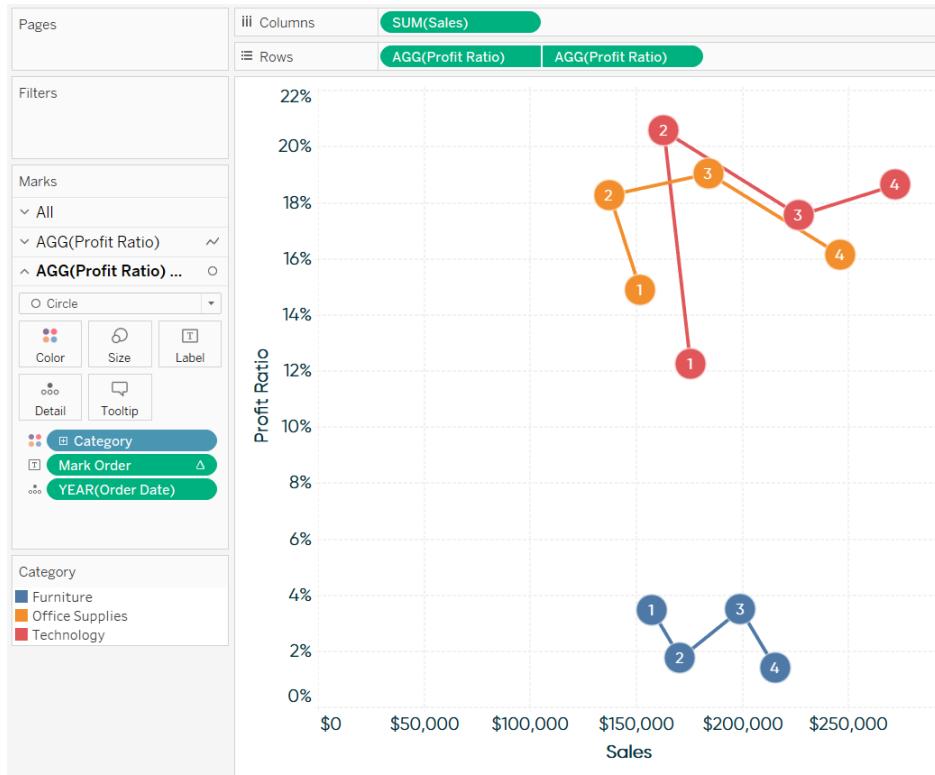


10.4.11. Connected scatterplot

10.4.11.1. Giới thiệu

Là sự kết hợp giữa biểu đồ phân tán (Scatter plot chart) và biểu đồ đường (Line graph), các dấu chấm phân tán sẽ được kết nối lại với nhau tạo thành một đường thẳng.

10.4.11.2. Minh họa



10.4.12. Density plot - Biểu đồ mật độ¹

10.4.12.1. Giới thiệu

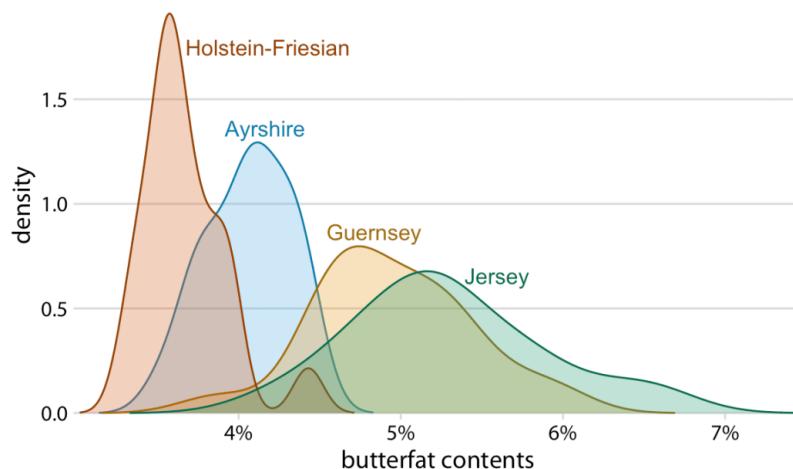
- Density plot lấy một biến số để biểu thị đường cong phân phối tron tru theo thời gian. Đinh của biểu đồ mật độ cho thấy mức độ tập trung tối đa của dữ liệu số. Biểu đồ này tương tự như histogram, trong đó sử dụng đường cong thay vì các thanh như trong histogram.
- Có thể vẽ tối đa năm biến trên cùng một biểu đồ để so sánh. Nếu vẽ nhiều biến hơn thì density chart sẽ trở nên nặng nề.

10.4.12.2. Sử dụng

- Density plot rất dễ hiểu và có thể so sánh giữa hai hoặc nhiều biến. Nó có điểm giống như histogram, nhưng cũng không giống như histogram vì nó không bị ảnh hưởng bởi một số bins (thùng) trên biểu đồ.
- Loại dữ liệu nào có thể hình dung với Density plot? Density plot phù hợp nhất cho dữ liệu liên quan đến số. Giả sử với tập dữ liệu về khoảng thời gian trên một trục và tần suất liên quan của nó trên một trục khác. Density plot sẽ hiển thị khoảng thời gian nào có tần suất tối đa, chỉ bằng cách nhìn vào mức cao nhất.

¹ [Density Plot - Data For Visualization](#)

10.4.12.3. Minh họa

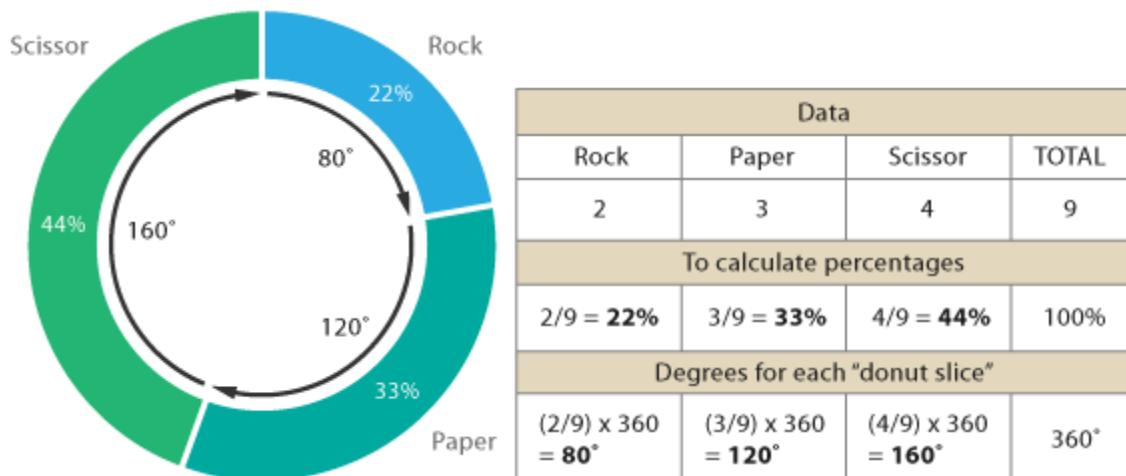


10.4.13. Donut pie chart

10.4.13.1. Giới thiệu

Là sự kết hợp giữa biểu đồ phân tán (Scatter plot chart) và biểu đồ đường (Line graph), Là một biến thể của Pie chart (Pie chart), dạng này không khác gì Pie chart ngoại trừ phần rỗng bên trong. Donut pie chart giúp người đọc tập trung nhiều hơn vào chiều dài của các cung tròn thay vì chỉ tập trung vào so sánh tỷ lệ giữa các lát cắt như Pie chart. Ngoài ra, donut pie chart sẽ tiết kiệm không gian hơn Pie chart vì phần trống bên trong có thể được sử dụng để hiển thị thông tin bên trong nó.

10.4.13.2. Minh họa



10.4.14. Dot plot (biểu đồ điểm)

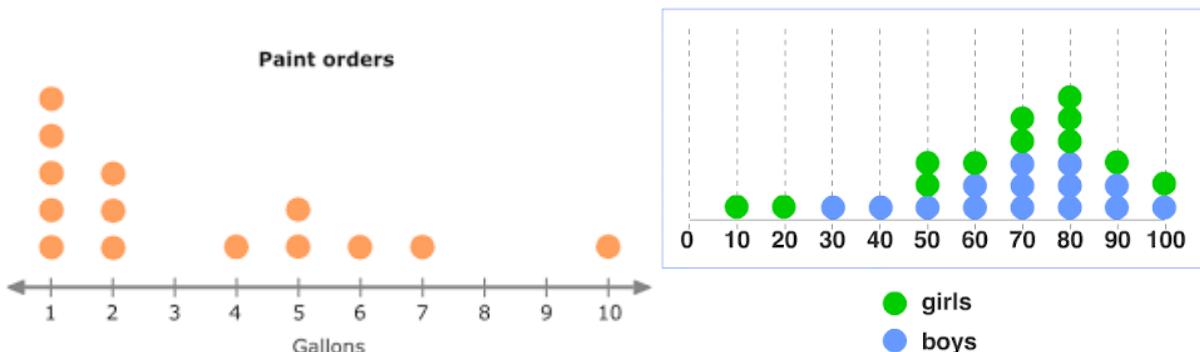
10.4.14.1. Giới thiệu

Được sử dụng trong thống kê cho các tập dữ liệu tương đối nhỏ, trong đó các giá trị rơi vào một số giá trị rời rạc so với dãy giá trị liên tục.

10.4.14.2. Minh họa

dot plot

A dot plot of weekly test result averages.



10.4.15. Dual-axis chart (Biểu đồ trục kép)

Overall CAC over time



Dual-axis charts phủ hai biểu đồ khác nhau có trục ngang chung lên nhau, nhưng có thể có tỷ lệ trục dọc khác nhau (một biểu đồ cho mỗi biểu đồ thành phần). Điều này có thể hữu ích khi hiển thị so sánh trực tiếp giữa hai tập hợp giá trị dọc, đồng thời bao gồm bối cảnh của biến trục hoành. Người ta thường sử dụng các loại biểu đồ cơ sở khác nhau, như kết hợp thanh và đường, để giảm sự nhầm lẫn về các tỷ lệ trục khác nhau cho từng biểu đồ thành phần.

10.4.16. Flowchart (lưu đồ)

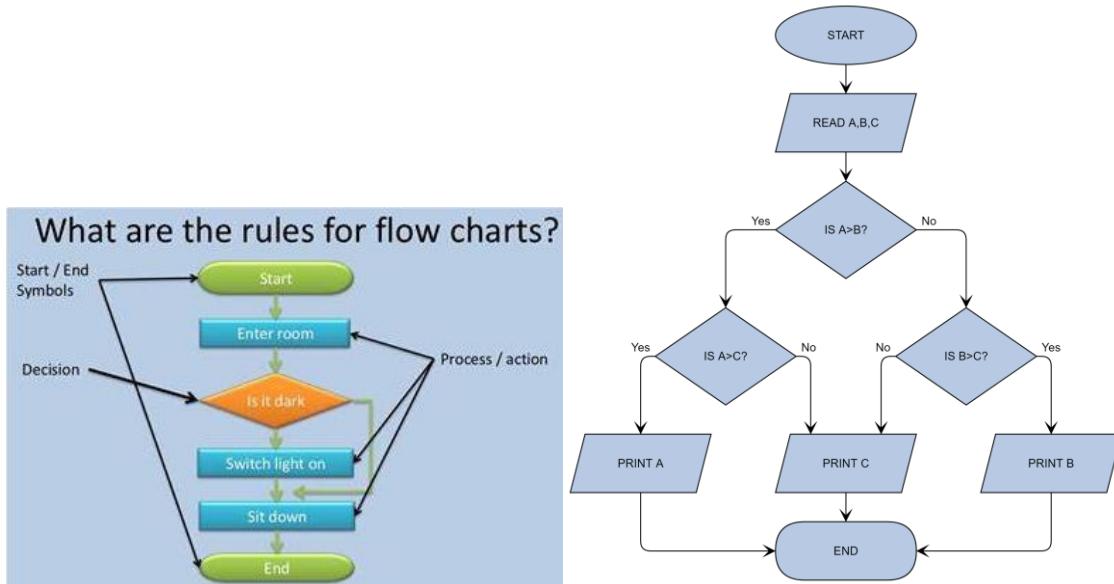
10.4.16.1. Giới thiệu

Flowcharts giúp tổ chức các bước, quyết định hoặc hành động trong một quy trình từ đầu đến cuối. Chúng thường bao gồm nhiều điểm bắt đầu hoặc điểm cuối, hiển thị các đường dẫn khác nhau mà có thể thực hiện trong một quy trình từ đầu đến cuối.

10.4.16.2. Sử dụng

Mọi người thường sử dụng Flowcharts để mô tả các tình huống phức tạp. Flowcharts sử dụng các hình dạng (*shapes*) được quy định trước để minh họa các phần khác nhau của quy trình và thường bao gồm chú thích để giải thích ý nghĩa của từng hình dạng.

10.4.16.3. Minh họa

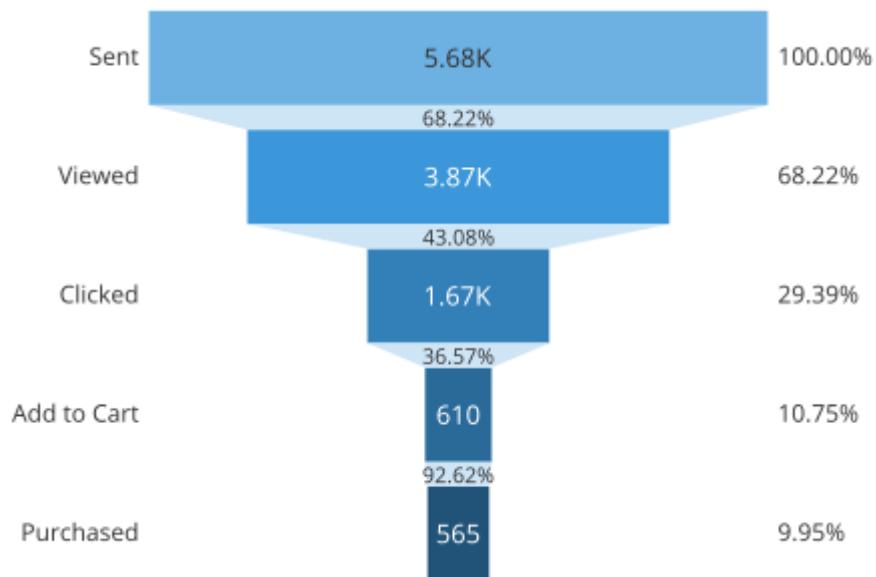


10.4.17. Funnel chart (biểu đồ kên)

10.4.17.1. Giới thiệu

Funnel charts minh họa cách giá trị tiến triển qua các giai đoạn khác nhau. Biểu đồ lấy tên từ hình dạng của nó, chúng rộng nhất ở phía trên và hẹp nhất ở phía dưới.. Số lượng người dùng ở mỗi giai đoạn của quy trình được biểu thị từ chiều rộng của kên khi nó thu hẹp.

Funnel Chart thường được thấy trong bối cảnh kinh doanh nơi khách truy cập hoặc người dùng cần được theo dõi trong luồng quy trình. Biểu đồ cho biết có bao nhiêu người dùng thực hiện từng giai đoạn của quy trình được theo dõi tính từ chiều rộng của kên ở mỗi phân chia giai đoạn. Việc thu hẹp kên giúp bán hàng theo cách tương tự, nhưng có thể làm xáo trộn tỷ lệ chuyển đổi thực sự là bao nhiêu.



Ví dụ về Funnel chart ở trên mô tả phản hồi cho chiến dịch khuyến mãi qua email đối với 1 sản phẩm đặc biệt. Mỗi giai đoạn trong số năm giai đoạn của quy trình được liên kết với một thanh có độ dài tương ứng với số lượng người dùng đã hoàn thành từng giai đoạn. Ngoài ra, bên cạnh mỗi thanh là tỷ lệ người dùng còn lại ở giai đoạn đầu tiên.

Từ biểu đồ này, có thể thấy rằng khoảng cách tuyệt đối lớn nhất là giữa việc xem email và nhấp vào liên kết quảng cáo. Có một sự sụt giảm tương đối khác giữa việc xem trang được liên kết và thêm sản phẩm vào giỏ hàng. Mặt khác, khi sản phẩm đã có trong giỏ hàng, hầu hết người dùng sẽ tiến hành mua hàng.

10.4.17.2. Sử dụng

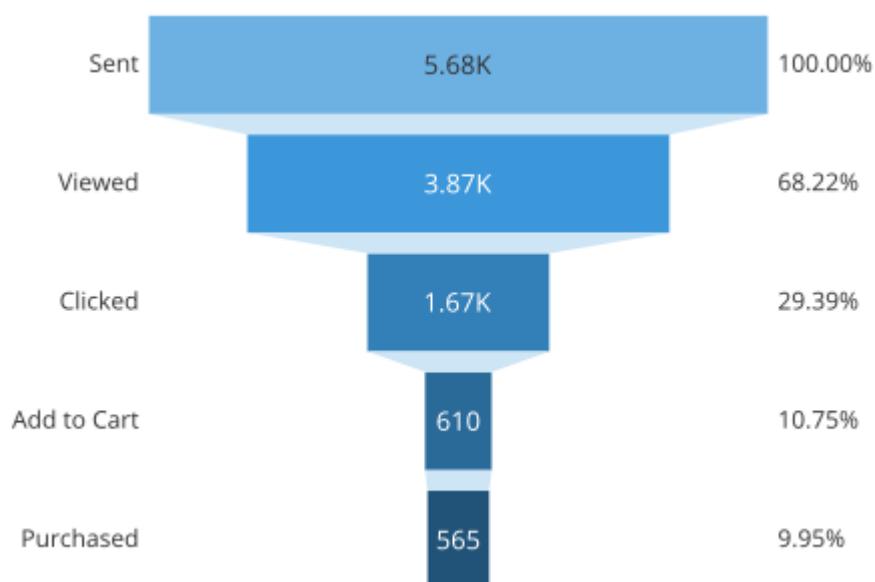
Funnel charts đặc biệt hữu ích khi theo dõi quá trình bán hàng. Chúng cũng hoạt động tốt để mô tả lưu lượng truy cập trang web, bao gồm số lượng khách truy cập vào một trang web, các trang đã xem, số lượt tải xuống được thực hiện và số lượng (hoặc tỷ lệ) người dùng ban đầu rời khỏi quy trình (hoặc luồng). Việc thực hiện đơn hàng là một cách sử dụng phổ biến khác của Funnel charts vì chúng có thể dễ dàng hiển thị “số lượng đơn hàng đã đặt, đã hủy và đã giao.

Loại biểu đồ này cho thấy phần mở đầu được chia thành các phần lũy tiến như thế nào.

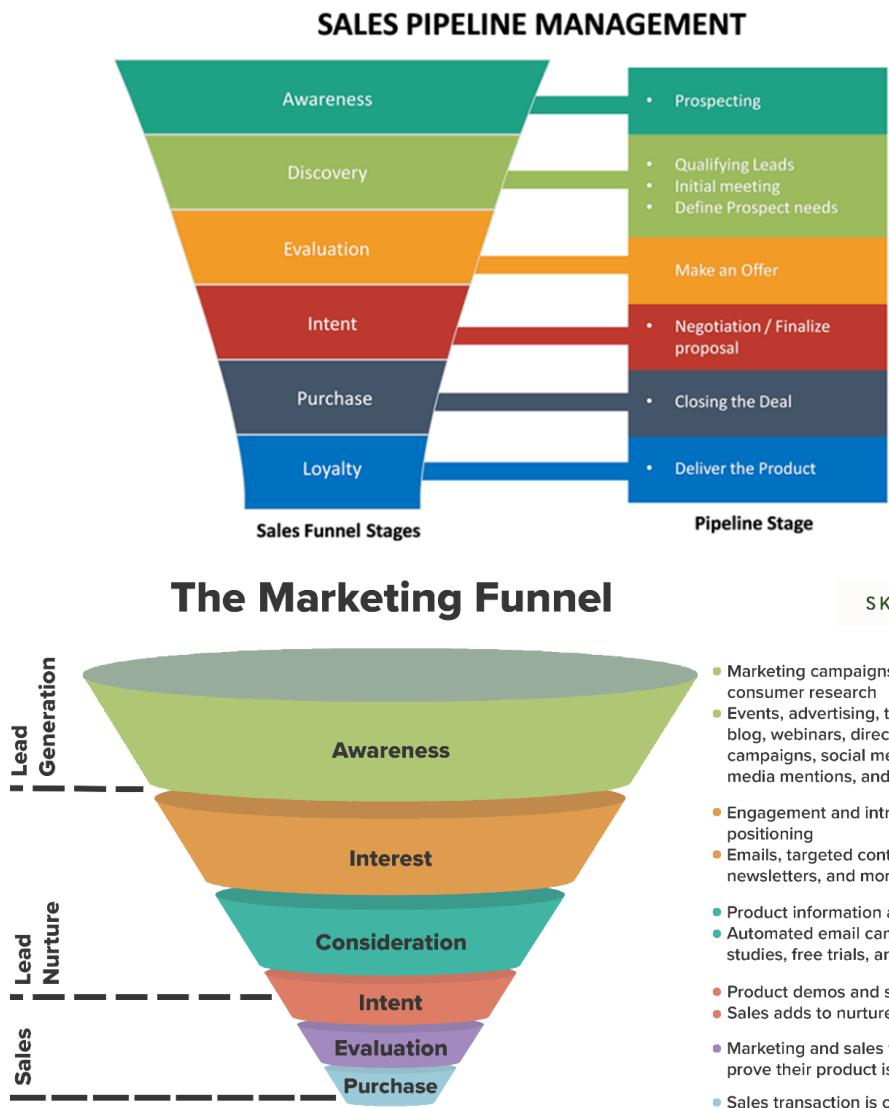
Bằng cách trực quan hóa số lượng người dùng thực hiện từng giai đoạn của quy trình, doanh nghiệp có thể hiểu được mức độ sụt giảm đáng kể ở đâu và cố gắng thực hiện các thay đổi để mang lại trải nghiệm tốt hơn cho người dùng. Lưu ý rằng sẽ không có đủ chi tiết trong Funnel chart để cho biết lý do xảy ra sự sụt giảm không mong muốn, vì vậy tốt nhất nên sử dụng Funnel chart làm hình ảnh hóa cấp cao trước khi chuyển sang điều tra sâu hơn.

10.4.17.2.1. Thống kê cần làm nổi bật trong Funnel chart

Trong Funnel chart, các tùy chọn chú thích tiêu chuẩn cho từng giai đoạn thường bao gồm hiển thị số lượng người dùng thô hoặc tỷ lệ người dùng so với giai đoạn mở đầu. Số lượng tuyệt đối có giá trị khi bắt đầu và kết thúc quá trình nhằm xây dựng sự hiểu biết về tổng số lượng người dùng. Tỷ lệ tương đối có thể cung cấp những hiểu biết nhanh hơn về mức độ hiệu quả của từng giai đoạn của quy trình. Nếu có thể, tốt nhất là hiển thị cả hai giá trị miễn là nó không làm rối loạn quá trình hiển thị.



Đối với một số công cụ trực quan, có thể có tùy chọn hiển thị các chủ thích khác cho từng giai đoạn. Một thông kê bổ sung đáng kể là tỷ lệ giữa các giai đoạn. Việc đưa chủ thích này vào giữa các giai đoạn sẽ cung cấp thêm thông tin cụ thể về tỷ lệ rời khỏi quy trình (hoặc luồng) mà không cần phải tính toán hoặc ước tính bên ngoài.



10.4.17.2.2. Minh họa cấu trúc dữ liệu dùng để vẽ biểu đồ

Stage	Users
Sent	5676
Viewed	3872
Clicked	1668
Add to Cart	610
Purchased	565

Dữ liệu để xây dựng Funnel chart thường được tóm tắt dưới dạng danh sách các giai đoạn và số lượng người dùng trong một bảng có hai cột. Mặc dù đầu ra của Funnel chart có thể hiển thị theo tỷ lệ của tổng số ban đầu, nhưng việc tính toán này thường không cần phải được thực hiện trước khi gửi dữ liệu đến ứng dụng trực quan hóa.

10.4.17.3. Những vấn đề thường gặp khi tạo funnel charts

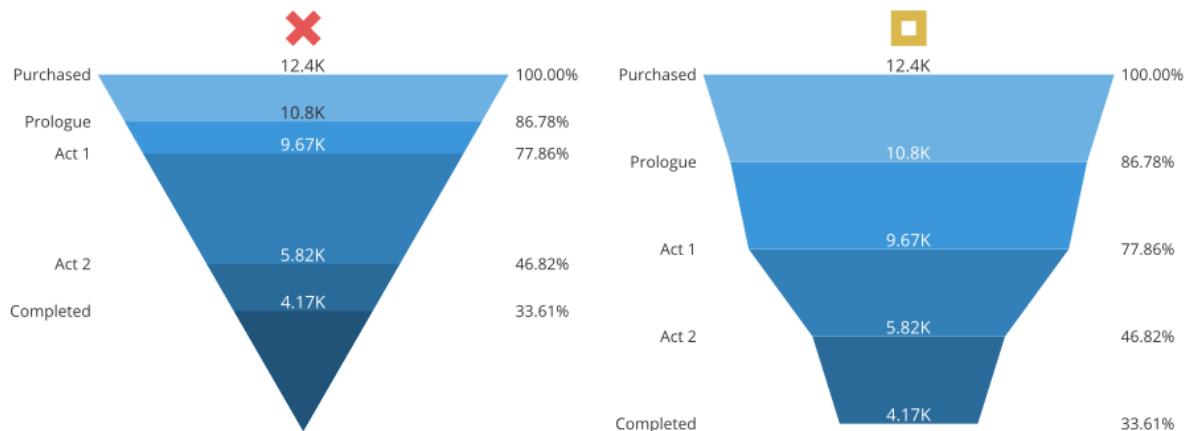
Trường hợp sử dụng chuyên biệt và hình dạng độc đáo của Funnel chart mang theo những vấn đề riêng liên quan đến các phương pháp hay nhất về trực quan hóa. Có nhiều cách triển khai Funnel chart khiến chúng khó hiểu hơn mức cần thiết.

10.4.17.3.1. Duy trì nhất quán về khoảng cách của các giai đoạn

Một cách có thể tạo Funnel chart là tạo biểu đồ hình tam giác thẳng. Các giai đoạn được đánh dấu trên hình tam giác có chiều rộng phù hợp với tỷ lệ người dùng đến từng giai đoạn. Tuy nhiên, điều này có thể làm biến dạng kích thước của nó và tầm quan trọng của từng giai đoạn.

Với cách chia tam giác thành các vùng, việc liên kết các giá trị giai đoạn với các vùng thay vì theo chiều rộng của ranh giới vùng có thể là điều hấp dẫn. Mặc dù thực tế là sự sụt giảm lớn hơn sẽ tương ứng với những khu vực lớn hơn, nhưng diện tích thực tế sẽ phụ thuộc vào nơi xảy ra sự sụt giảm đó. Việc thu nhỏ hình tam giác có nghĩa là việc mất đi một số lượng người dùng nhất định sớm trong quá trình sẽ có diện tích lớn hơn nhiều so với những người bị mất ở giai đoạn cuối của quá trình. Điều này làm cho những khoản “mất mát” ban đầu có vẻ quan trọng hơn những khoản “mất mát” sau đó, điều này có thể không thực sự đúng như vậy.

Trong ví dụ bên trái ở dưới, mức giảm tuyệt đối giữa hai giai đoạn đầu tiên Purchased (Mua hàng) – Prologue (Mở đầu) giống như giữa hai giai đoạn cuối Act2-Complete (Hoạt động 2 - Đã hoàn thành). Tuy nhiên, diện tích của điểm rơi đầu tiên lớn hơn nhiều do nằm ở vị trí cao hơn trong hình tam giác.



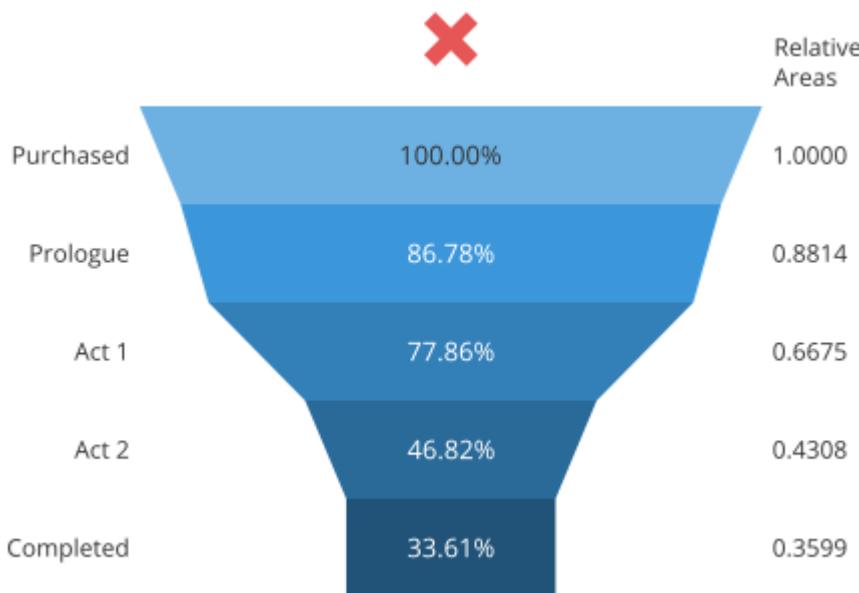
Thay vào đó, nên có các giai đoạn cách đều nhau trên trực tiếp đứng. Điều sẽ thay đổi bây giờ là độ dốc của phễu để kết nối các ranh giới giữa các giai đoạn.

10.4.17.3.2. Ghép nhãn với vị trí đó

Một lỗi phổ biến khác khi tạo Funnel chart là thêm vùng hình chữ nhật bổ sung vào kẽm ở trên cùng hoặc dưới cùng, sau đó đặt nhãn ở giữa mỗi vùng. Thoạt nhìn, có vẻ như các khu vực sẽ khớp chính xác với giá trị của từng giai đoạn, nhưng điều này nhìn chung không đúng.

Khoảng cách giữa các giai đoạn càng lớn thì độ dốc giữa các ranh giới vùng càng lớn và càng có ít diện tích trong vùng phễu tương ứng. Vùng hình chữ nhật sẽ không hiển thị phần giảm xuống và do đó được coi là lớn hơn giá trị giai đoạn tương ứng.

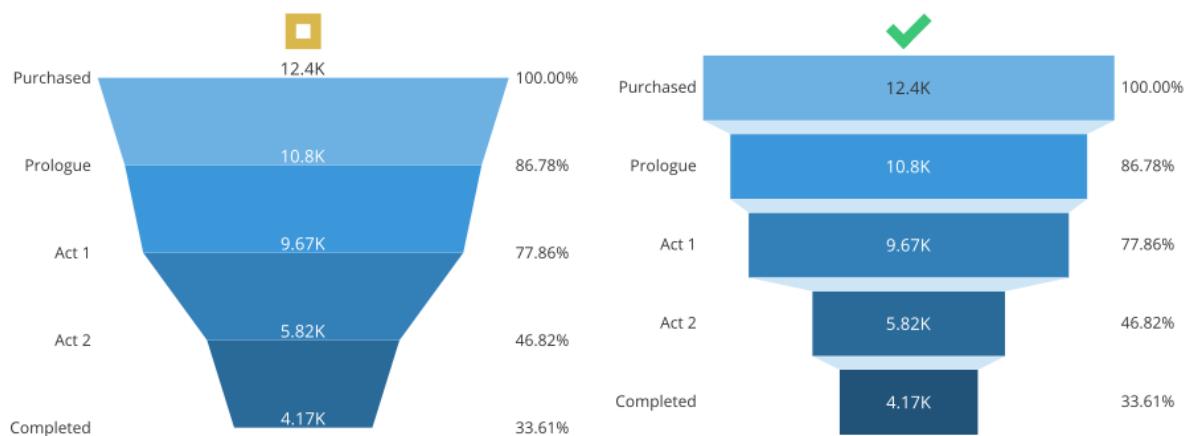
Trong ví dụ bên dưới, giai đoạn thứ hai và thứ năm có diện tích lớn hơn giá trị thực của chúng, trong khi giai đoạn thứ ba và thứ tư có diện tích nhỏ hơn giá trị của chúng.



Để có Funnel chart trong đó các khu vực khớp chính xác với các giá trị giai đoạn, cần một chiến lược thay thế để mô tả dữ liệu.

10.4.17.3.3. Sử dụng biểu đồ phễu kiểu thanh

Phương pháp xây dựng Funnel chart ở trên, trong đó các giá trị giai đoạn được đánh dấu trên ranh giới vùng với độ rộng phù hợp, là chính xác nhưng không lý tưởng. Thực tế là có những khu vực rộng lớn, được lấp đầy thu hút sự chú ý của người xem, tuy nhiên chúng không thể diễn giải được dưới dạng dữ liệu. Sẽ tốt hơn nhiều nếu có thể thay đổi mọi thứ bằng cách đẽ kích thước vùng tương ứng với giá trị giai đoạn, được phân tách bằng ranh giới vùng mỏng hơn.



Để thực hiện điều này, có thể tưởng tượng việc tạo độ dày cho các ranh giới vùng ban đầu để chúng trở thành các thanh, đẽ bẹp các khu vực dốc ban đầu. Vì mỗi thanh giai đoạn có các cạnh thẳng, điều này loại bỏ sự mơ hồ về giá trị: diện tích và chiều rộng giờ đây hoàn toàn tỷ lệ với nhau. Để tập trung vào các khu vực thực sự tương ứng với các giai đoạn của quy trình, nên tạo cho các vùng kết nối một màu bão hòa nhẹ, không nổi bật. Khi tạo sơ đồ hình phễu, đây là kiểu thiết kế nên hướng tới.

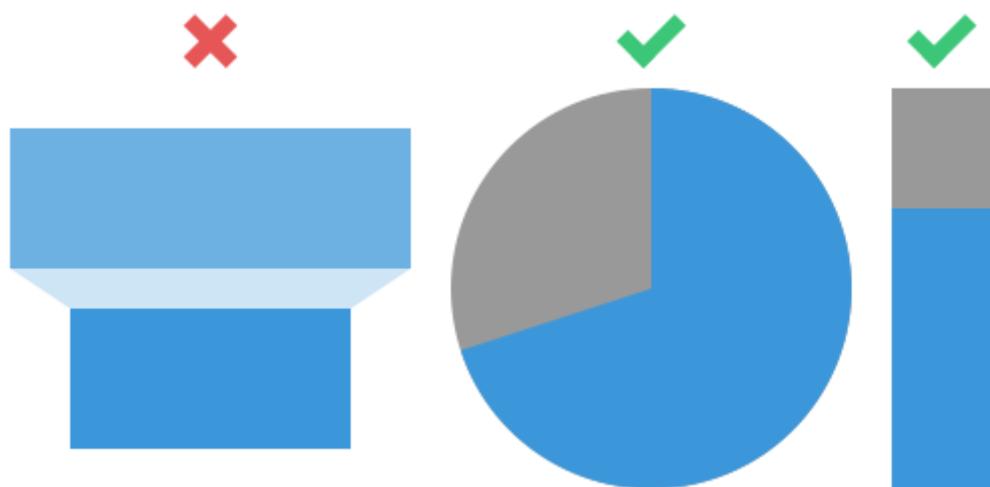
10.4.17.4. Sử dụng hiệu quả Funnel chart

10.4.17.4.1. Include at least three stages

In order to use a funnel chart, you should have at least three stages to plot. When there are only two stages in a process, we only have a single ratio to comprehend. A simpler part-to-whole representation like a [pie chart](#) or single [stacked bar](#) will work better in this case.

10.4.17.4.2. Bao gồm ít nhất ba giai đoạn

Để sử dụng Funnel chart, bạn phải có ít nhất ba giai đoạn để vẽ. Khi chỉ có hai giai đoạn trong một quy trình, chỉ có một tỷ lệ duy nhất để hiểu. Cách trình bày từng phần đơn giản hơn như Pie chart hoặc Stacked bar chart đơn sẽ hoạt động tốt hơn trong trường hợp này.



10.4.17.5. Các biểu đồ liên quan

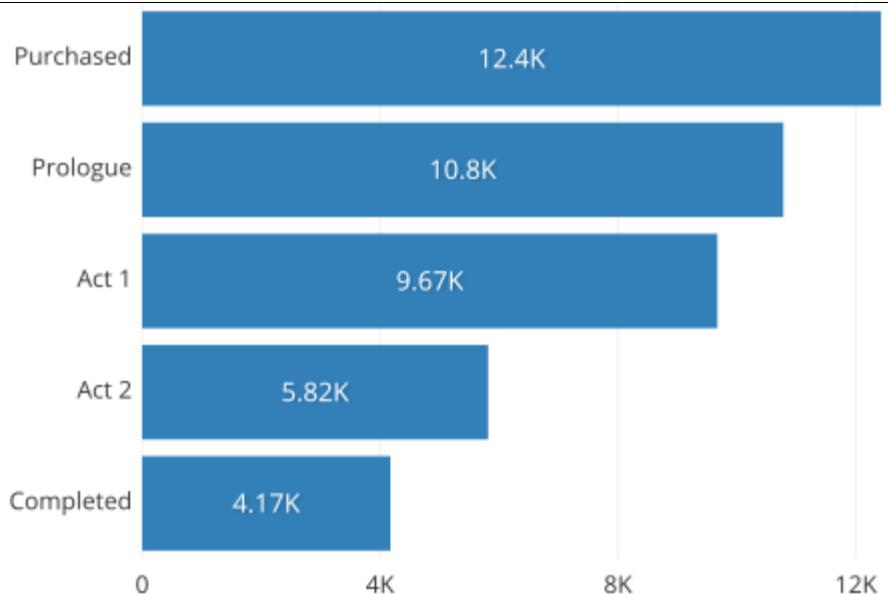
10.4.17.5.1. Bar chart

At its core, the funnel chart is really just a fancy-looking [bar chart](#). The funnel chart's "bars" are aligned to a center line, and connected by additional graphical elements. Meanwhile, a bar chart's bars are usually aligned to the left or bottom axis (depending on orientation) without additional connecting elements. This common baseline can make it easier to directly compare proportions without needing to look at any annotations.

While the best way to create a funnel chart is to essentially make a center-aligned bar chart, it is often a better idea overall to just use the bar chart instead.

Về cốt lõi, Funnel chart thực chất chỉ là một Bar chart có hình thức bắt mắt. Các "thanh" của Funnel chart được căn chỉnh theo đường trung tâm và được kết nối bằng các phần tử đồ họa bổ sung. Trong khi đó, các thanh của Bar chart thường được căn chỉnh theo trực trái hoặc trực dưới (tùy theo hướng) mà không có các yếu tố kết nối bổ sung. Đường cơ sở chung này có thể giúp việc so sánh trực tiếp các tỷ lệ trở nên dễ dàng hơn mà không cần xem bất kỳ chú thích nào.

Mặc dù cách tốt nhất để tạo Funnel chart về cơ bản là tạo Bar chart căn giữa, nhưng về tổng thể, tốt hơn hết là bạn chỉ nên sử dụng Bar chart.



10.4.17.5.2. Stacked bar chart

Để thực sự bán được thông tin chi tiết từng phần của quy trình kênh (*funnel process*), loại biểu đồ thích hợp nhất là Stacked bar chart. Thay vì vẽ các thanh giai đoạn theo một đường như trong Bar chart tiêu chuẩn, Stacked bar chart sẽ phủ tất cả các thanh ở cùng một vị trí. Mỗi vùng của Stacked bar chart sẽ hiển thị tỷ lệ người dùng dừng ở mỗi giai đoạn của quy trình. Hãy đảm bảo rằng nếu sử dụng cách trình bày này thì nó sẽ phù hợp với cách diễn giải dữ liệu.



Lưu ý rằng thứ tự của các giai đoạn được đặt sao cho có thể đọc được luồng quy trình từ trái sang phải.

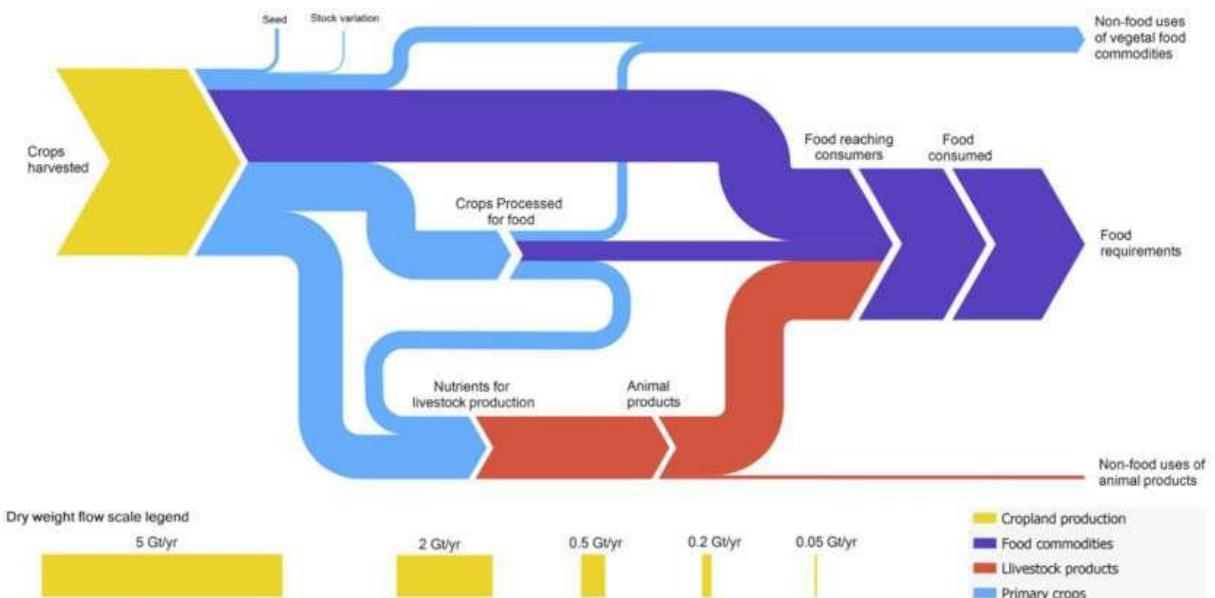
Một câu hỏi có thể được đặt ra là tại sao khoảng cách không đồng đều của Stacked bar chart ở đây lại ổn, nhưng lại không ổn với dạng hình tam giác thẳng của phễu. Trong trường hợp phễu hình tam giác có chiều rộng khác nhau đọc theo chiều dài của nó thì thanh xếp chồng lên nhau có cùng chiều rộng đọc theo chiều dài của nó. Vì điều này, khu vực tương ứng với mỗi người dùng sẽ luôn nhất quán, bất kể họ bị mất số lượng ở đâu trong quá trình.

10.4.17.5.3. Sankey diagram

Other flow diagrams like the Sankey diagram can depict more complicated processes than the funnel chart. While a funnel chart expects a simple, linear process, a Sankey diagram can depict multiple sources of input and output. Like a funnel chart, value is encoded in the width of the chart on each segment of the visualization, but those segments may not be regularly spaced out.

Các sơ đồ luồng khác (flow diagrams) như Sankey diagram có thể mô tả các quy trình phức tạp hơn Funnel chart. Trong khi Funnel chart yêu cầu một quy trình tuyến tính, đơn giản thì sơ đồ Sankey có thể mô tả nhiều nguồn đầu vào và đầu ra. Giống như Funnel chart, giá trị

được mã hóa theo chiều rộng của biểu đồ trên mỗi phân đoạn của hình ảnh trực quan nhưng các phân đoạn đó có thể không được giãn cách đều đặn.



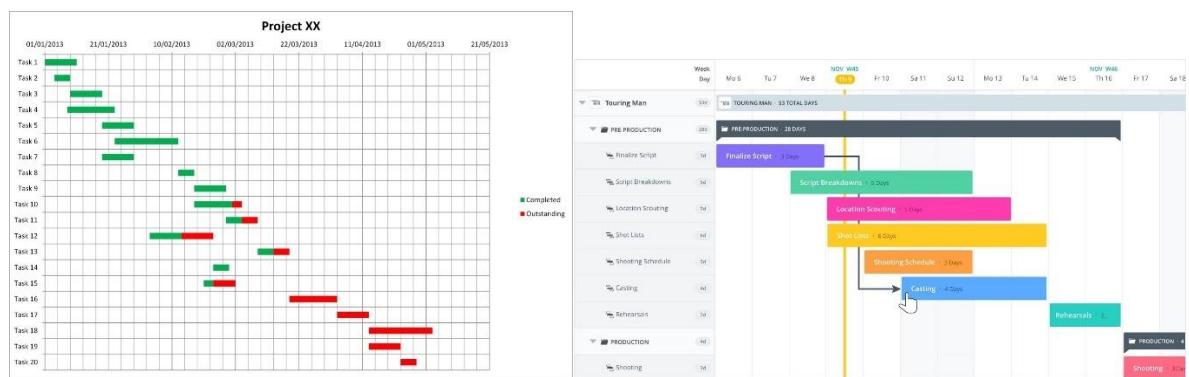
Sơ đồ Sankey ở trên cho thấy cây trồng đã thu hoạch được sử dụng như thế nào trên quy mô toàn cầu¹.

10.4.18. Gantt chart (đồ thị tiến độ)

10.4.18.1. Giới thiệu

Gantt chart minh họa lịch trình dự án. Trục ngang biểu thị khung thời gian của dự án theo ngày, tuần, tháng hoặc năm. Biểu đồ hiển thị từng nhiệm vụ dự án dưới dạng thanh trên trục tung. Độ dài của thanh tùy thuộc vào ngày bắt đầu và ngày kết thúc của nhiệm vụ, nhưng đôi khi cũng có một đường thẳng đứng cho ngày hiện tại. Người quản lý dự án sử dụng Gantt chart để theo dõi tiến độ và trạng thái hoàn thành của từng nhiệm vụ.

10.4.18.2. Minh họa



10.4.19. Gauge chart (biểu đồ đo)

10.4.19.1. Giới thiệu

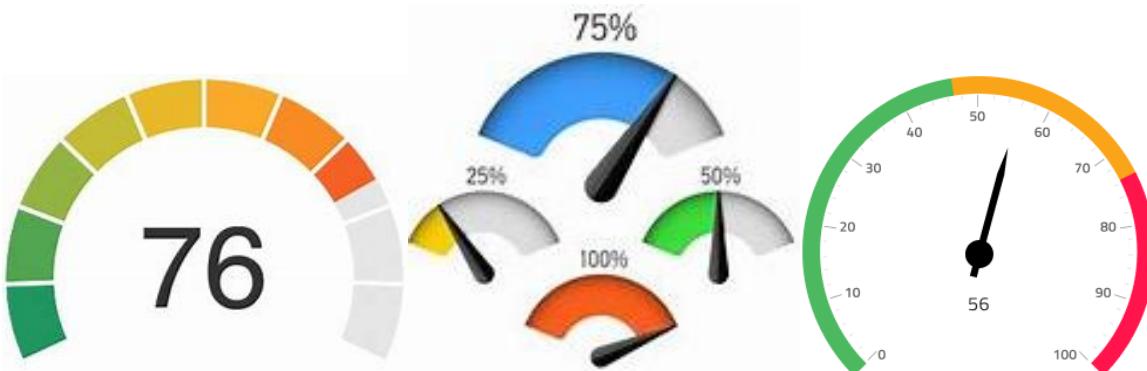
Gauge chart hiển thị dữ liệu dưới dạng số đọc trên mặt số. Gauge chart có một cung tròn, vạch kim trong biểu đồ thể hiện mục tiêu hoặc giá trị của mục tiêu được đề ra nhằm đo

¹ Nguồn: [Losses, inefficiencies and waste in the global food system](http://sankey-diagrams.com), qua sankey-diagrams.com

lường giá trị đã đạt được so với mục tiêu đã đề ra. Giá trị ngoài cùng bên trái vòng cung là giá trị nhỏ nhất, giá trị ngoài cùng bên phải vòng cung là giá trị lớn nhất.

Gauge Chart thường được dùng để minh họa tốc độ, mục tiêu doanh thu, nhiệt độ, ...

10.4.19.2. Minh họa



10.4.20. Geospatial Charts

10.4.20.1. Công dụng

- Biểu đồ không gian địa lý là công cụ mạnh mẽ cung cấp bối cảnh trực quan để phân tích dữ liệu, giúp hiểu thông tin địa lý phức tạp dễ dàng hơn.
- Biểu đồ không gian địa lý cực kỳ hữu ích để trực quan hóa dữ liệu liên quan đến vị trí địa lý.

10.4.20.2. Một số ứng dụng chính của biểu đồ không gian địa lý

- (i).- *Ánh xạ dữ liệu không gian*: được sử dụng để vẽ các điểm dữ liệu trên bản đồ, giúp trực quan hóa sự phân bố không gian của các yếu tố khác nhau, chẳng hạn như mật độ dân số, vị trí của các cơ sở hoặc phân bố tài nguyên.
- (ii).- *Phân tích các mô hình địa lý*: giúp xác định các mô hình và mối quan hệ trong dữ liệu cụ thể cho các vùng địa lý. Điều này có thể bao gồm việc phân tích các xu hướng như tỷ lệ tội phạm ở các khu vực khác nhau, những thay đổi về môi trường hoặc hoạt động kinh tế.
- (iii).- *Ra quyết định*: Các doanh nghiệp và chính phủ sử dụng biểu đồ không gian địa lý để đưa ra quyết định sáng suốt. Ví dụ: họ có thể xác định vị trí tốt nhất cho các cửa hàng mới, tối ưu hóa tuyến giao hàng hoặc lập kế hoạch cho các dự án phát triển đô thị.
- (iv).- *Theo dõi những thay đổi theo thời gian*: Những biểu đồ này có thể theo dõi những thay đổi theo thời gian, chẳng hạn như nạn phá rừng, mở rộng đô thị hoặc sự lây lan của dịch bệnh, cung cấp những hiểu biết sâu sắc có giá trị cho việc lập kế hoạch và giảm thiểu.
- (v).- *Theo dõi những thay đổi theo thời gian*: Những biểu đồ này có thể theo dõi những thay đổi theo thời gian, chẳng hạn như nạn phá rừng, mở rộng đô thị hoặc sự lây lan của dịch bệnh, cung cấp những hiểu biết sâu sắc có giá trị cho việc lập kế hoạch và giảm thiểu.
- (vi).- *Tiếp thị và bán hàng*: Các công ty sử dụng biểu đồ không gian địa lý để phân tích thị trường, xác định nhân khẩu học mục tiêu và điều chỉnh chiến lược tiếp thị cho phù hợp với các khu vực cụ thể.

- (vii).- *Ứng phó và quản lý thiên tai*: Trong các tình huống khẩn cấp, biểu đồ không gian địa lý rất quan trọng để điều phối các nỗ lực ứng phó, chẳng hạn như lập bản đồ các khu vực bị ảnh hưởng, lập kế hoạch tuyến đường sơ tán và phân bổ nguồn lực.
- (viii).- *Du lịch và Lữ hành*: giúp giới thiệu các địa điểm du lịch, lập kế hoạch lộ trình du lịch và cung cấp bản đồ chi tiết cho khách du lịch.

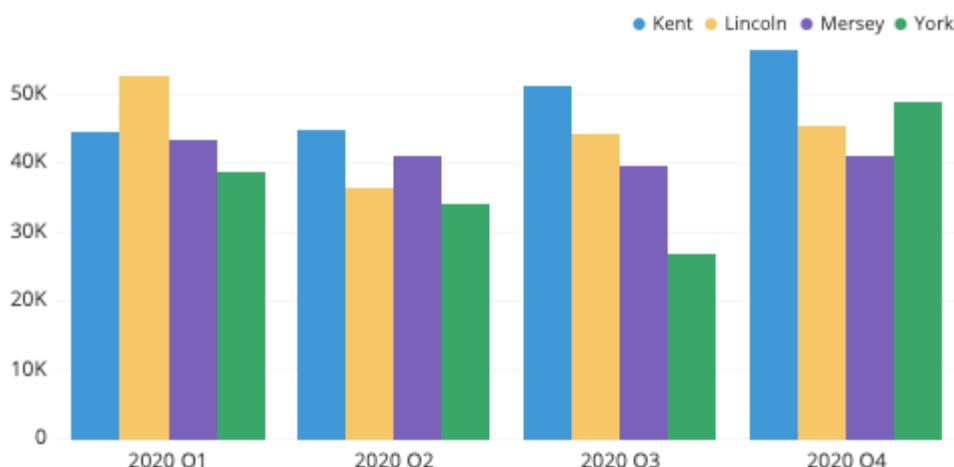
10.4.20.3. Một số dạng biểu đồ không gian địa lý

- *Bản đồ nhiệt (Heat Maps)*: Hiển thị cường độ dữ liệu tại các điểm địa lý cụ thể.
- *Bản đồ Choropleth (Choropleth Maps)*: Sử dụng các màu khác nhau để thể hiện phạm vi dữ liệu giữa các vùng.
- *Bản đồ phân phối điểm (Dot Distribution Maps)*: Sử dụng các dấu chấm để thể hiện sự xuất hiện của một biến trong các khu vực khác nhau.
- *Bản đồ ký hiệu tỷ lệ (Proportional Symbol Maps)*: Sử dụng các ký hiệu có kích thước khác nhau để biểu thị số lượng dữ liệu.

10.4.21. Grouped bar chart

10.4.21.1. Giới thiệu

Grouped bar chart (còn gọi là *clustered bar chart*, *multi-series bar chart*) vẽ các giá trị số cho mức của hai biến phân loại thay vì một. Các thanh được nhóm theo vị trí cho các cấp độ của một biến phân loại, với màu sắc biểu thị cấp độ danh mục phụ trong mỗi nhóm.



Grouped bar chart ở trên so sánh doanh thu hàng quý mới của bốn đại diện bán hàng trong một năm. Một cụm thanh được vẽ cho mỗi quý và trong mỗi cụm, một thanh cho mỗi đại diện. Màu sắc và vị trí nhất quán trong mỗi cụm: ví dụ: có thể thấy Kent luôn có màu xanh lam và được vẽ đầu tiên. Từ biểu đồ, có thể thấy rằng Lincoln có thành tích tốt nhất trong Q1 và Kent tốt nhất trong tất cả các quý còn lại. Cũng có thể kiểm tra các hoạt động riêng lẻ chẳng hạn như hoạt động tương đối ổn định của Mersey trong suốt cả năm hoặc cú tăng mạnh của York trong Quý 4 sau khi trượt dốc từ Quý 1 đến Quý 3.

10.4.21.2. Sử dụng

Giống như Bar chart tiêu chuẩn, Grouped bar chart được xây dựng để hiển thị sự phân bổ các điểm dữ liệu hoặc so sánh giữa các danh mục dữ liệu khác nhau. Điểm khác biệt của Grouped bar chart là việc phân chia các điểm dữ liệu thành hai biến phân loại khác nhau chứ không chỉ một. Grouped bar chart được sử dụng khi muốn xem biến danh mục thứ cấp thay đổi như thế nào trong từng cấp độ của biến danh mục chính hoặc khi muốn xem biến danh mục đầu

tiên thay đổi như thế nào qua các cấp độ của biến danh mục thứ hai. So sánh loại thứ nhất được gọi là so sánh “trong nhóm” (*within-group*) và so sánh loại thứ hai được gọi là “giữa các nhóm” (*between-group*). Trong ví dụ trên, so sánh trong nhóm sẽ tập trung vào các thanh trong một quý, trong khi so sánh giữa các nhóm sẽ tập trung vào các thanh cho một đại diện duy nhất trong các quý.

Để tạo điều kiện thuận lợi cho việc so sánh này, các thanh trong Grouped bar chart lại được vẽ một cách có hệ thống. Để so sánh trong nhóm, các cấp độ của biến phân loại chính sẽ xác định vị trí cho một cụm thanh được vẽ. Số lượng thanh được vẽ trong mỗi nhóm bằng với số cấp của biến phân loại thứ cấp. Việc so sánh giữa các nhóm được hỗ trợ bằng cách chọn màu sắc và thứ tự quán cho từng cấp độ của biến phụ được vẽ trong mỗi nhóm.

Grouped bar chart không được trang bị đầy đủ để so sánh tổng số giữa các cấp độ của các biến phân loại riêng lẻ. Vì không có bất kỳ yếu tố gốc nào cho tổng số nhóm trong Grouped bar chart, nên người đọc sẽ mất rất nhiều công sức để ước tính tổng số cho bất kỳ cấp độ phân loại, chính hoặc phụ nào. Nếu việc so sánh tổng số cho một biến phân loại là quan trọng thì loại biểu đồ khác như Bar chart tiêu chuẩn hoặc Grouped bar chart sẽ thực hiện nhiệm vụ tốt hơn.

10.4.21.2.1. Thứ tự của các biến phân loại

Một điều quan trọng cần cân nhắc khi tạo Grouped bar chart là quyết định biến nào trong hai biến phân loại sẽ là biến chính (cho biết vị trí trực cho mỗi cụm thanh) và biến nào sẽ là biến phụ (cho biết số lượng thanh cần vẽ trong mỗi cụm). Kiến thức về lĩnh vực của dữ liệu sẽ giúp trong việc quyết định xem biến nào quan trọng (bao trùm) hơn và do đó được chọn làm biến chính.

Các biến phân loại mô tả dữ liệu thời gian (ví dụ: các bản tóm tắt hàng tháng từ 20XX-tháng 1, 20XX-Tháng 2, 20XX-Tháng 3, v.v.) thường sẽ là lựa chọn rõ ràng cho biến phân loại chính. Đối với các biến phân loại thuần túy như giới tính hoặc quốc gia, nên ưu tiên đặt chúng làm biến phụ nếu chúng có số lượng cấp độ nhỏ để vẽ: càng có nhiều cấp độ, càng cần nhiều màu sắc khác biệt và càng khó phân biệt giữa chúng. Mặt khác, các biến số khác như độ tuổi (18-24, 25-34, 35-44, v.v.) hoặc điểm xếp hạng (thỏa thuận theo thang điểm từ 1-7) có thể hoạt động tốt như các biến phụ vì có thể chỉ quan tâm nhiều hơn đến việc phân phối liên tục các giá trị hơn là xác định chính xác các cấp độ riêng lẻ và giá trị của chúng.

Không thể tránh khỏi, sẽ có trường hợp không có lựa chọn rõ ràng về cách thiết lập hệ thống phân cấp danh mục, ngay cả sau khi xem xét các mục tiêu trực quan và kiến thức về lĩnh vực của dữ liệu. Không có hại gì khi chỉ thử nghiệm và thử cả hai thứ tự biến để xem cái nào truyền tải dữ liệu tốt nhất.

10.4.21.2.2. Cấu trúc dữ liệu để vẽ biểu đồ

Quarter	Kent	Lincoln	Mersey	York
2020-Q1	44 700	52 800	43 500	38 800
2020-Q2	45 000	36 500	41 000	34 100
2020-Q3	51 200	44 200	39 700	27 000
2020-Q4	56 500	45 300	41 200	48 900

Dữ liệu cho các Grouped bar chart thường ở dạng bảng giống như biểu đồ ở trên. Cột đầu tiên biểu thị các cấp độ của biến phân loại chính, trong khi cột thứ hai và các cột tiếp theo tương ứng với từng cấp độ của biến phân loại phụ. Các biến số trong các ô cho biết chiều cao của mỗi thanh; các thanh được vẽ theo hàng để tạo ra các nhóm thanh.

10.4.21.3. Sử dụng hiệu quả Grouped bar chart

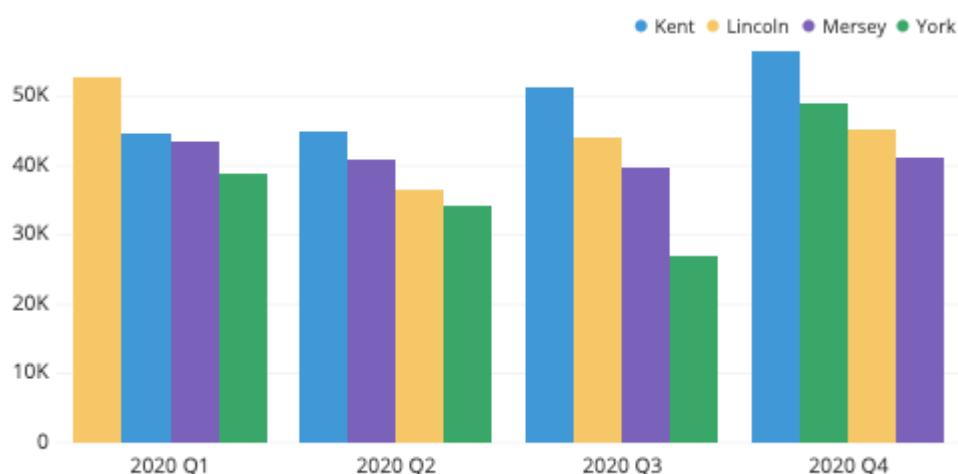
Các nguyên tắc sử dụng hiệu quả trong Grouped bar chart phù hợp với các nguyên tắc dành cho Bar chart tiêu chuẩn, nhưng có một số điều chỉnh do sự hiện diện của biến phân loại thứ cấp.

10.4.21.3.1. Duy trì đường cơ sở bằng 0

Việc bổ sung các thanh phân cụm thực sự không ảnh hưởng đến nguyên tắc đưa đường cơ sở bằng 0 vào Bar chart. Giống như trong biểu đồ cơ sở, đường cơ sở làm cho độ dài thanh đúng với giá trị của chúng.

10.4.21.3.2. Thứ tự các cấp độ danh mục

Nguyên tắc sắp xếp các thanh từ lớn nhất đến nhỏ nhất trừ khi chúng có thứ tự cố hữu cũng áp dụng cho Grouped bar chart giống như đối với Bar chart cơ bản, nhưng có một chút cân nhắc về cách xác định 'lớn nhất' đến 'nhỏ nhất'. Các đánh giá về quy mô nên được thực hiện trên từng biến phân loại riêng lẻ, bỏ qua sự phân chia của biến phân loại quan tâm khác. Điều này đặc biệt quan trọng đối với biến phân loại thứ cấp: việc sắp xếp các thanh nhât quán giũa các nhóm thường sẽ hữu ích hơn việc sắp xếp các thanh từ lớn nhất đến nhỏ nhất trong mỗi nhóm. Tuy nhiên, cách sắp xếp sau trong nhóm này có các trường hợp sử dụng, chẳng hạn như khi quan tâm đến việc xếp hạng theo một biến chính tạm thời.



10.4.21.3.3. Lựa chọn màu sắc hiệu quả

Bảng màu định tính (Qualitative Palette, e.g. Action type)



Bảng màu tuần tự (Sequential, e.g. Age group)



Bảng màu phân kỳ (Diverging Palette, e.g. Likert scale)



Mặc dù nguyên tắc chung là giữ tất cả các thanh có cùng màu cho Bar chart tiêu chuẩn, việc lựa chọn màu sắc trở thành một phần quan trọng của Grouped bar chart để phân biệt các

cấp độ của biến phân loại phụ. Lựa chọn quan trọng cần thực hiện ở đây là chọn bảng màu phù hợp với loại biến phụ mà bạn có: bảng màu định tính cho biến phân loại thuần túy hoặc bảng màu tuân tự hoặc phân kỳ cho các biến phân loại có thứ tự cố hữu.

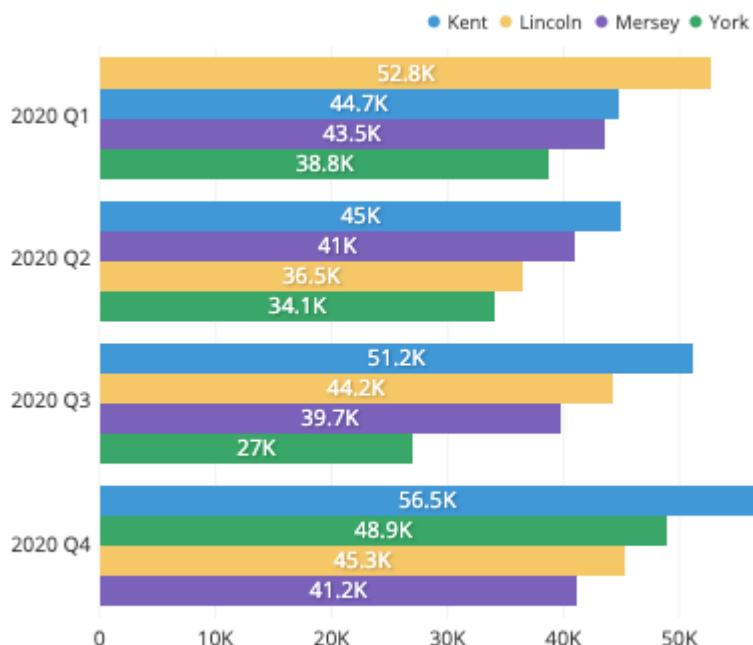
10.4.21.4. Các tùy chọn thường dùng kèm với Grouped bar chart

10.4.21.4.1. Horizontal grouped bar chart

Giống như Bar chart tiêu chuẩn, Grouped bar chart có thể được tạo bằng các thanh dọc (danh mục chính trên trực hoành) hoặc thanh ngang (danh mục chính trên trực tung). Hướng ngang mang lại những lợi ích tương tự như trong Bar chart tiêu chuẩn, cung cấp thêm chỗ cho các nhãn danh mục chính dài mà không cần phải xoay hoặc cắt bớt.

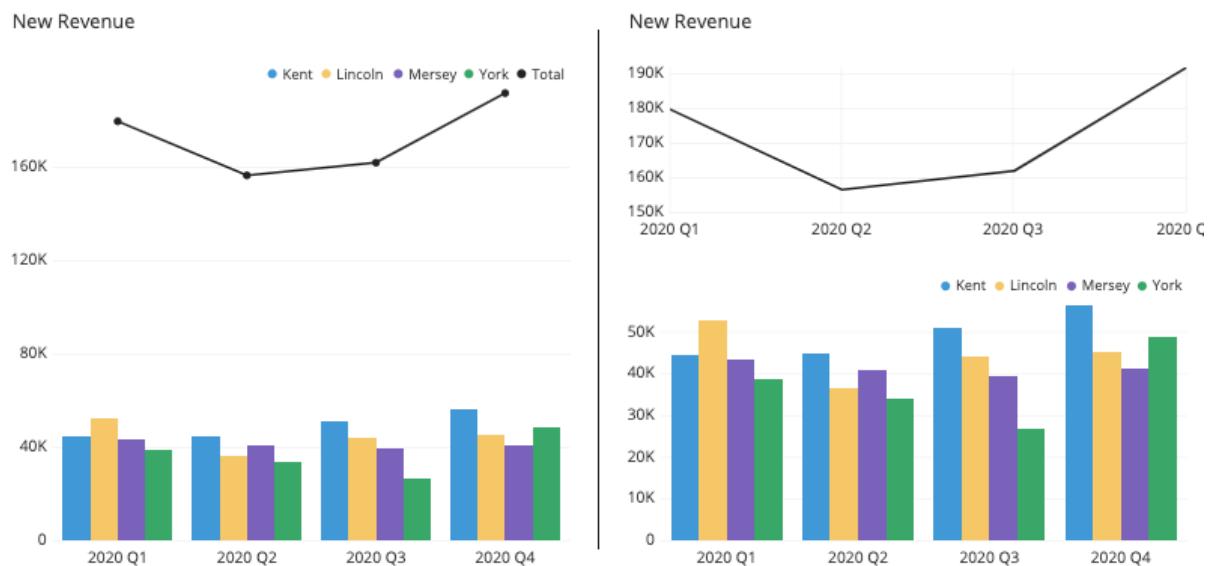
10.4.21.4.2. Chú thích giá trị

Các chú thích về độ dài thanh có thể được thêm vào Grouped bar chart gần giống như đối với Bar chart tiêu chuẩn. Mặc dù các chú thích có thể giúp người đọc xác định chính xác các giá trị, nhưng thực tế là thường sẽ có nhiều thanh hơn để vẽ có nghĩa là sự lộn xộn về mặt hình ảnh của các chú thích sẽ tăng lên đối với Grouped bar chart.



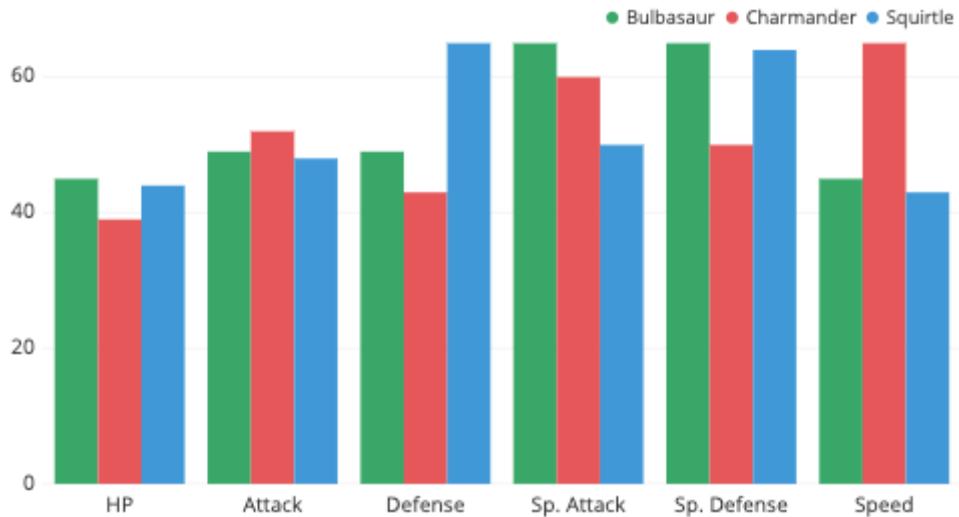
10.4.21.4.3. Các thành phần bổ sung để hiển thị tổng số

Như đã lưu ý trước đó, Grouped bar chart thường sẽ không bao gồm bất kỳ phần tử nào hiển thị tổng giá trị cho các biến phân loại chính hoặc phụ. Một cách để cộng tổng cho biến phân loại chính có thể là thêm một thanh lớn phía sau mỗi nhóm hoặc thành phần biểu đồ đường phía trên mỗi nhóm. Tuy nhiên, điều này có thể kéo dài đáng kể chiều cao của ô, đặc biệt khi có nhiều thanh phụ. Khuyến nghị chung vẫn là chỉ sử dụng một ô riêng nếu tổng số được quan tâm thay vì cố gắng ép mọi thứ vào một ô duy nhất.



10.4.21.4.4. Faceted bar charts (Biểu đồ thanh nhiều mặt)

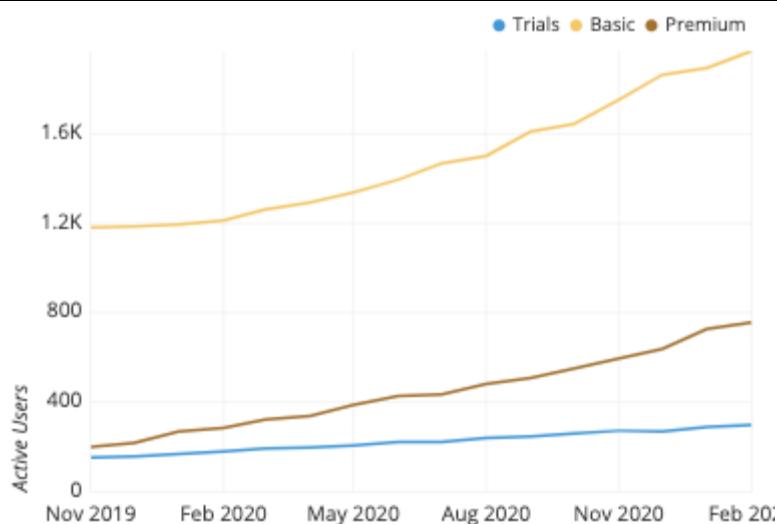
Một trường hợp sử dụng trông giống như Grouped bar chart là việc thay thế biến phân loại chính bằng nhiều số liệu khác nhau. Vì mỗi số liệu có thể có thang đo trực khác nhau nên mỗi số liệu có xu hướng có trực riêng. Trên thực tế, loại biểu đồ này chỉ đặt một số Bar chart tiêu chuẩn cạnh nhau (các mặt tiêu chuẩn), nhưng màu sắc của các thanh mới là thứ mang lại sức mạnh cho biểu đồ. Màu sắc và thứ tự thanh được chọn từ loại Grouped bar chart nhấn mạnh việc thực hiện so sánh trong nhóm tốt hơn nếu mỗi ô phụ chỉ được coi là độc lập với các ô khác.



10.4.21.5. Các đồ thị liên quan

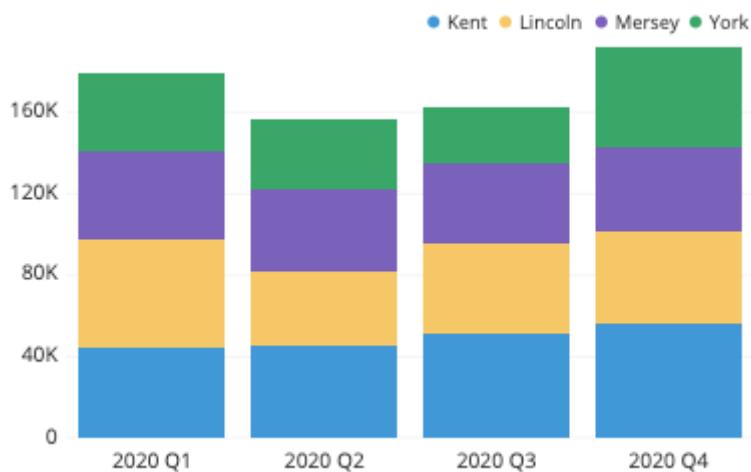
10.4.21.5.1. Line chart

Khi biến phân loại chính có tính chất liên tục, đặc biệt nếu nó liên quan đến thời gian thì loại biểu đồ thay thế hữu ích cần xem xét là Line chart. Line chart đặc biệt hữu ích khi có nhiều cấp độ trong biến phân loại chính: nhu cầu phân cụm nhiều thanh xung quanh mỗi vị trí có thể khiến biểu đồ khó đọc. Line chart giúp giải quyết vấn đề này bằng cách căn chỉnh từng nhóm phụ theo chiều dọc và đường kết nối giữa các điểm giúp dễ dàng theo dõi cách mỗi nhóm phụ thay đổi.



10.4.21.5.2. Stacked bar chart

Nếu sửa đổi một Grouped bar chart trong đó, đối với mỗi nhóm chính, ta xếp các thanh chồng lên nối tiếp nhau thay vì nằm cạnh nhau, thì kết quả sẽ là một Stacked bar chart. Tổng chiều dài của mỗi thanh chính sẽ giống như khi không có danh mục phụ và do đó, Stacked bar chart nhấn mạnh vào tổng các cấp danh mục chính và sự đóng góp tương đối từng phần của từng cấp danh mục phụ. Sự cân bằng với loại biểu đồ này là việc so sánh các danh mục phụ giữa các cấp danh mục chính trở nên khó khăn hơn nhiều.



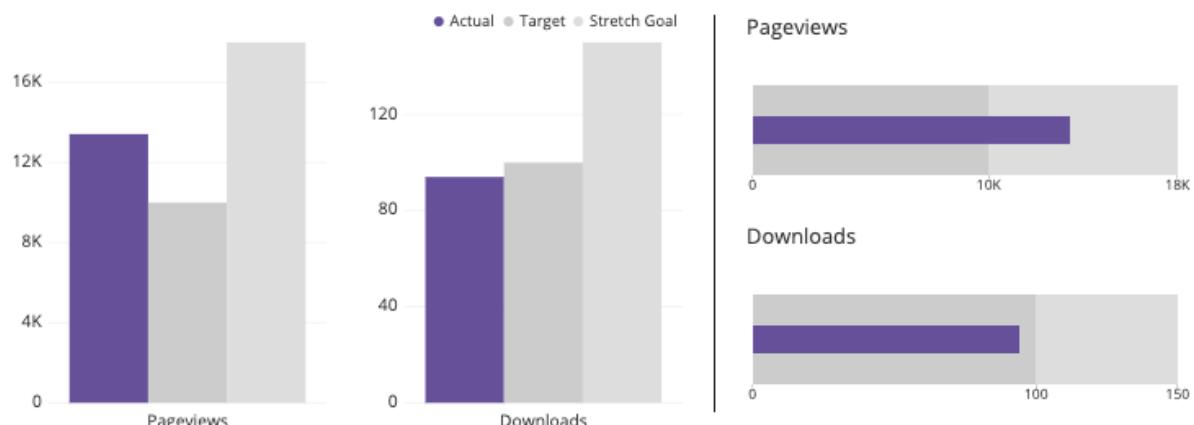
10.4.21.5.3. Heatmap

Nếu tưởng tượng các thanh có chiều sâu, có thể tưởng tượng việc thay đổi góc nhìn của mình để nhìn chúng từ trên cao. Nếu căn chỉnh các nhóm để chúng tạo thành một ma trận các thanh, thì về cơ bản sẽ thu được một Heatmap. Heatmap giống như các bảng được bổ sung màu sắc để nâng cao khả năng phát hiện các mẫu và xu hướng. Mặc dù Heatmap yêu cầu chú thích để có thể đọc các giá trị số dễ dàng như Grouped bar chart, nhưng chúng cũng rất nhỏ gọn và có một số cách sử dụng tổng quát khác.

	27K	56.5K	
Kent	44.7K	45K	51.2K
Lincoln	52.8K	36.5K	44.2K
Mersey	43.5K	41K	39.7K
York	38.8K	34.1K	27K
	2020 Q1	2020 Q2	2020 Q3
			2020 Q4

10.4.21.5.4. Bullet chart

Bullet chart (hoặc Bullet graph) là Bar chart chuyên dụng được sử dụng trong bối cảnh kinh doanh để theo dõi các số liệu hiệu suất so với mục tiêu của họ. Một thanh mỏng duy nhất biểu thị giá trị số liệu thực tế (*actual metric value*), trong khi các thanh lớn hơn và các dấu hiệu khác biểu thị giá trị mục tiêu (*goal value*) và các điểm chuẩn (*benchmarks*) khác. Theo một cách nào đó, đây giống như một Bar chart nhóm trong đó các cấp phân loại phụ là giá trị thực, mục tiêu và điểm chuẩn nhưng được vẽ theo kiểu chồng chéo cụ thể. Vì Bullet chart chỉ có một giá trị dữ liệu 'thực' nên đây là một cách tốt, gọn nhẹ để đưa ra đánh giá so sánh trong nháy mắt.

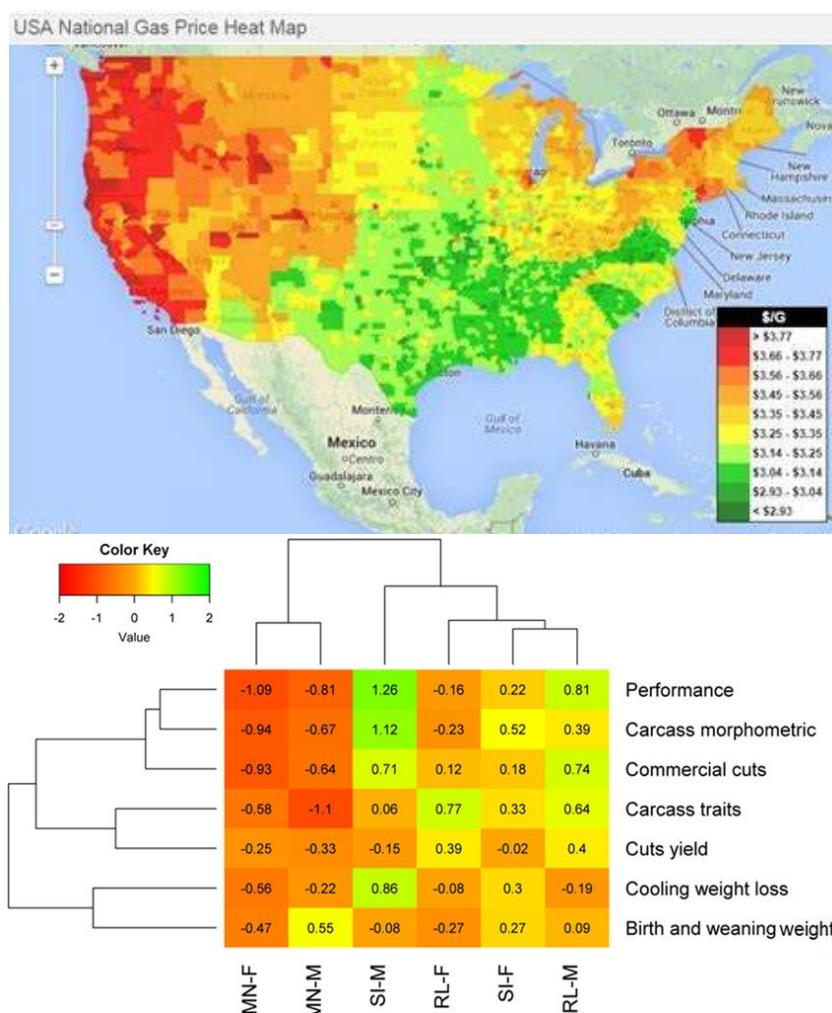
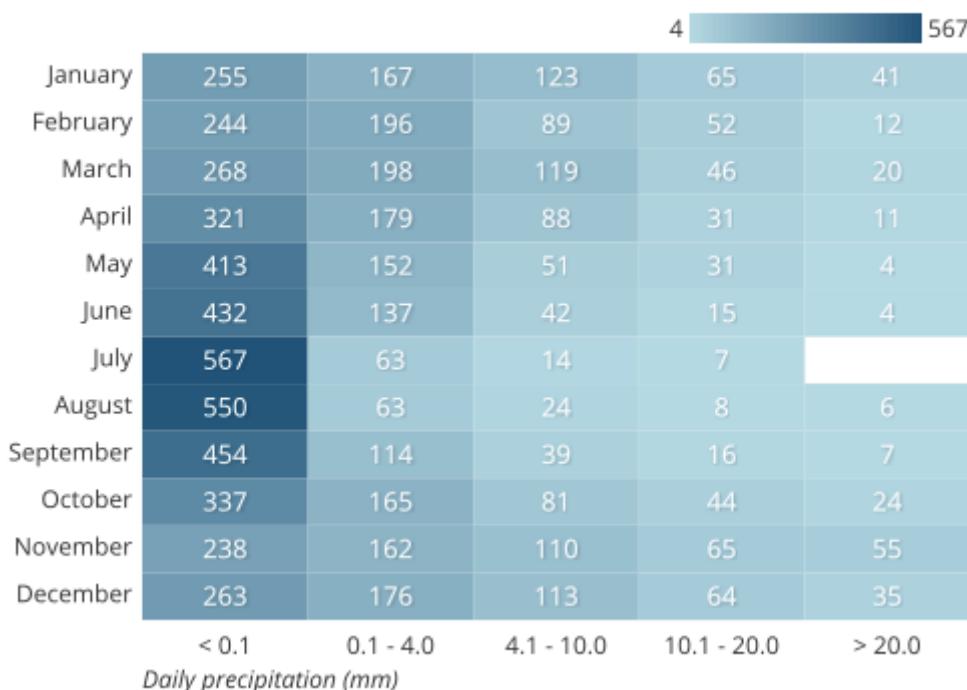


10.4.22. Heat map - Biểu đồ nhiệt

10.4.22.1. Giới thiệu

- Heat map (hoặc heatmap, bản đồ nhiệt) là kỹ thuật trực quan hóa dữ liệu 2 chiều mô tả các giá trị cho biến quan tâm chính trên hai biến trực dưới dạng lưới các ô vuông màu. Sự thay đổi màu sắc cho biết độ lớn hay cường độ của giá trị cần đo.
- Heat map là một thuật ngữ tương đối mới, nhưng việc sử dụng ma trận tô bóng (shading matrix) đã tồn tại hơn một thế kỷ.

Seattle precipitation by month, 1998-2018



Heatmap ví dụ ở trên mô tả sự phân bố lượng mưa hàng ngày, được nhóm theo tháng và được ghi lại trong 11 năm ở Seattle, Washington. Mỗi ô là một số đếm, giống như trong bảng dữ liệu tiêu chuẩn, nhưng số đếm đi kèm với một màu, với số lượng lớn hơn liên quan đến màu tối hơn. Từ Heatmap này, có thể thấy từ các màu tối nhất ở cột ngoài cùng bên trái rằng hầu hết các ngày không có mưa trong cả năm. Mô hình màu sắc ô qua các tháng cũng cho thấy mưa phổ biến hơn vào mùa đông từ tháng 11 đến tháng 3 và ít phổ biến nhất vào các tháng hè như tháng 7 và tháng 8.

2-d density plots (Đồ thi mật độ 2-d)

Thuật ngữ Heatmap cũng được sử dụng theo nghĩa tổng quát hơn, trong đó dữ liệu không bị ràng buộc vào lưới. Ví dụ: các công cụ theo dõi cho trang web có thể được thiết lập để xem cách người dùng tương tác với trang web, như nghiên cứu nơi người dùng nhấp vào hoặc người đọc có xu hướng cuộn xuống trang bao xa.



Mỗi lần người dùng click chuột (hoặc sự kiện theo dõi khác) được liên kết với một vị trí, tạo ra một lượng nhỏ giá trị số xung quanh vị trí của nó. Các giá trị này được cộng lại với

¹ Ví dụ về heatmap từ [Google Maps documentation](#)

nhau trên tất cả các sự kiện và sau đó được vẽ bằng bản đồ màu liên quan. Ngôn ngữ trực quan của đầu ra của các công cụ này, liên kết giá trị với màu sắc, tương tự như loại bản đồ nhiệt được xác định ở trên cùng, chỉ là không có cấu trúc dựa trên lưới. Heatmap loại này đôi khi còn được gọi là 2-d density plots (biểu đồ mật độ 2 chiều).

10.4.22.2. Sử dụng

- Heat maps rất hữu ích trong việc kiểm tra chéo dữ liệu thông qua việc đặt các biến vào hàng, cột và tô màu các ô trong bảng.
- Bên cạnh đó, chúng cũng rất tốt trong việc hiển thị phương sai trên nhiều biến, hiển thị bất kỳ biến nào tương tự với nhau để phát hiện xem có tồn tại bất kỳ mối tương quan nào giữa chúng hay không. Heatmaps là biểu đồ phù hợp để có cái nhìn tổng quan về dữ liệu số do sự phụ thuộc của chúng vào màu sắc để giao tiếp các giá trị.

Heatmap được sử dụng để hiển thị mối quan hệ giữa hai biến, một biến được vẽ trên mỗi trục. Bằng cách quan sát cách màu sắc của các ô thay đổi trên mỗi trục, giúp người xem có thể quan sát xem có bất kỳ mẫu nào có giá trị cho một hoặc cả hai biến hay không.

Các biến được vẽ trên mỗi trục có thể thuộc bất kỳ loại nào, cho dù chúng có kiểu dữ liệu là nhãn phân loại hay giá trị số. Trong trường hợp là số, giá trị số phải được đánh dấu giống như trong biểu đồ để tạo thành các ô lưới nơi các màu liên quan đến biến quan tâm chính sẽ được vẽ.

Màu sắc của ô có thể tương ứng với tất cả các loại số liệu, chẳng hạn như tần suất trong mỗi bin hoặc số liệu thống kê tóm tắt như giá trị trung bình hoặc trung vị cho biến thứ ba. Một cách nghĩ về việc xây dựng Heatmap là dưới dạng bảng hoặc ma trận, với mã hóa màu ở đầu các ô. Trong một số ứng dụng nhất định, các ô cũng có thể được tô màu dựa trên các giá trị không phải là số (ví dụ: mức chất lượng chung là thấp, trung bình, cao).

10.4.22.2.1. Một số ứng dụng của Heatmap trong thực tế

- Trong một số ứng dụng như phân tích tội phạm hoặc theo dõi lượt click vào trang web, màu sắc được sử dụng để biểu thị mật độ điểm dữ liệu thay vì giá trị liên quan đến từng điểm.
- Là một công cụ thể hiện dữ liệu trực quan hành vi của người dùng truy cập vào website thông qua màu sắc. Màu nóng là nơi được tương tác nhiều nhất và màu lạnh là nơi tương tác ít nhất.

10.4.22.2.2. Ví dụ về cấu trúc dữ liệu dùng để vẽ Heatmap

USERNAME	< 0.01	0.1 - 4.0	4.1 - 10.0	...
January	255	167	123	...
February	244	196	89	...
March	268	198	119	...
April	321	179	88	...
...

Các ứng dụng trực quan hóa khác nhau có thể có những cách chấp nhận dữ liệu khác nhau để vẽ đồ thị dưới dạng Heatmap. Ở một dạng chính, dữ liệu có thể được cung cấp giống như cách dữ liệu được hiển thị tự nhiên dưới dạng bảng. Cột đầu tiên sẽ giữ các giá trị cho một trục của bản đồ nhiệt, trong khi tên của các cột còn lại sẽ tương ứng với các bin cho trục còn lại. Các giá trị trong các cột đó sẽ được mã hóa vào Heatmap.

Hình thức phổ biến khác cho dữ liệu Heatmap thiết lập nó ở định dạng ba cột. Mỗi ô trong Heatmap được liên kết với một hàng trong bảng dữ liệu. Hai cột đầu tiên chỉ định 'tọa độ' của ô bản đồ nhiệt, trong khi cột thứ ba cho biết giá trị của ô.

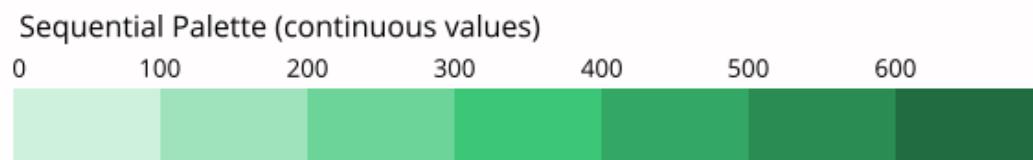
MONTH	PRCP_BUCKET	COUNT
March	10.1 - 20.0	46
March	> 20.0	20
April	< 0.1	321
April	0.1 - 4.0	179
...

10.4.22.3. Sử dụng heatmap hiệu quả

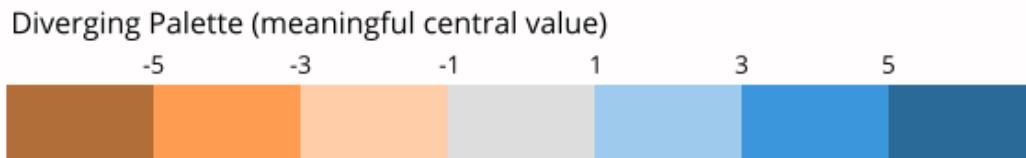
10.4.22.3.1. Chọn bảng màu phù hợp

Màu sắc là thành phần cốt lõi của loại biểu đồ này, vì vậy, cần đảm bảo rằng chọn bảng màu thích hợp để khớp với dữ liệu. Thông thường nhất, sẽ có một dải màu tuần tự giữa giá trị và màu sắc, trong đó màu nhạt hơn tương ứng với giá trị nhỏ hơn và màu tối hơn tương ứng với giá trị lớn hơn hoặc ngược lại. Tuy nhiên, bảng màu phân kỳ có thể được sử dụng khi các giá trị có điểm 0 có ý nghĩa.

Bảng màu tuần tự (sequential palette) mô tả giá trị liên tục

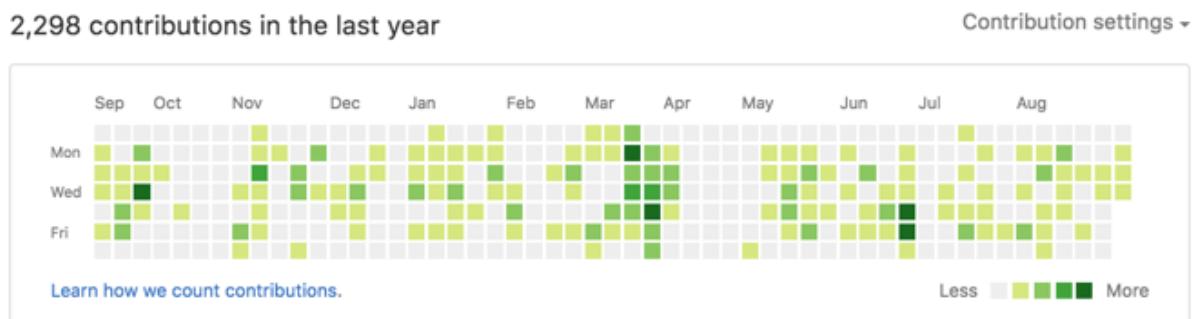


Bảng màu phân kỳ (Diverging palette) mô tả giá trị hướng tâm.



10.4.22.3.2. Nên bao gồm chú thích (legend) cho đồ thị

Heatmap thường phải bao gồm chú thích về cách màu sắc ánh xạ tới các giá trị số. Vì bản thân màu sắc không có mối liên hệ có hữu nào với giá trị nên chìa khóa rất quan trọng để người xem nắm bắt được các giá trị trong Heatmap. Một ngoại lệ đối với việc bao gồm chú thích có thể xảy ra khi sự liên kết tuyệt đối của giá trị với màu sắc không quan trọng mà chỉ vẽ các mẫu dữ liệu tương đối.



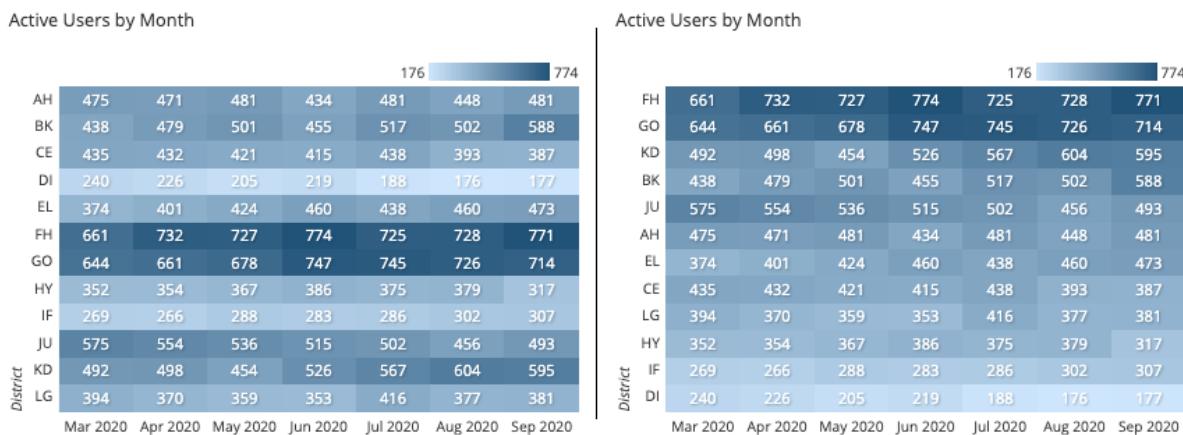
Lịch đóng góp của GitHub sử dụng chú thích chung để hiển thị số lượng đóng góp theo ngày.
(Nguồn: [GitHub](#))

10.4.22.3.3. Hiển thị giá trị trong ô

Việc ánh xạ màu thành giá trị thiếu độ chính xác, đặc biệt là so với các mã hóa khác như vị trí hoặc độ dài. Nếu có thể, nên thêm chú thích giá trị ô vào Heatmap dưới dạng mã hóa kép giá trị.

10.4.22.3.4. Sắp xếp các cấp độ theo mức độ tương tự hoặc giá trị

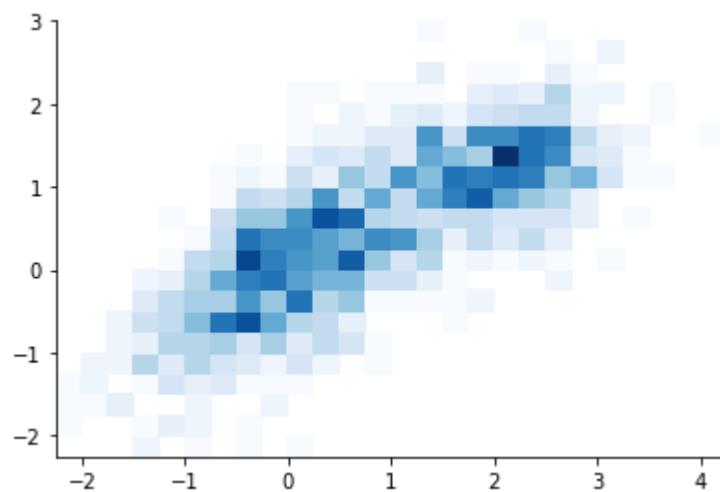
Khi một hoặc cả hai biến trực trong biểu đồ có bản chất là phân loại, có thể đáng cân nhắc việc thay đổi thứ tự mà các mức biến trực đó được biểu thị. Nếu các danh mục không có thứ tự cố hữu, có thể muốn chọn thứ tự giúp người đọc nắm bắt tốt nhất các mẫu trong dữ liệu. Một tùy chọn phổ biến là sắp xếp các danh mục theo giá trị ô trung bình của chúng từ lớn nhất đến nhỏ nhất.



Bản đồ nhiệt bên phải được sắp xếp theo giá trị cột cuối cùng.

Một kỹ thuật tiên tiến hơn bao gồm việc nhóm và phân cụm các giá trị danh mục bằng cách đo lường mức độ tương tự. Điều này thường thấy trong trường hợp sử dụng Clustered heatmap (bản đồ nhiệt theo cụm).

10.4.22.3.5. Sử dụng dấu tích 1 cách hữu ích



Đối với các biến trực số, có thể lựa chọn cách thiết lập các bin và cách chúng được biểu thị trong biểu đồ. Nếu có ít bin, có thể giữ dấu tích trên mỗi bin giống như đối với biến trực phân loại. Tuy nhiên, khi có nhiều bin, lựa chọn tốt hơn là đánh dấu các dấu tích giữa các nhóm

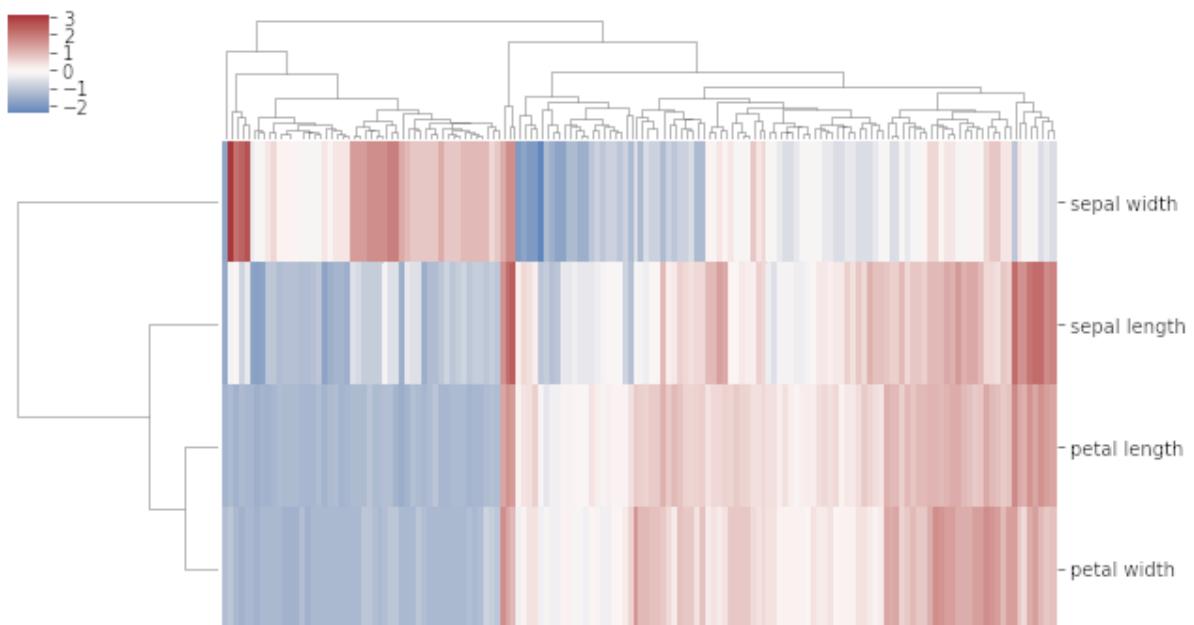
bin để tránh tình trạng quá nhiều dấu tích. Số lượng bin nên sử dụng và kích thước của chúng sẽ phụ thuộc vào bản chất của dữ liệu, vì vậy, nên thử nghiệm với các cài đặt khác nhau.

10.4.22.4. Các tùy chọn phổ biến được dùng trên heatmap

10.4.22.4.1. Clustered heatmap (Bản đồ nhiệt cụm)

Thay vì để trực hoành biểu thị các mức hoặc giá trị của một biến duy nhất, một biến thể phổ biến là nó biểu thị các phép đo của các biến hoặc số liệu khác nhau. Nếu đặt trực tung làm các quan sát riêng lẻ, sẽ thu được thứ gì đó giống như một bảng dữ liệu tiêu chuẩn, trong đó mỗi hàng là một quan sát và các cột là giá trị của thực thể trên mỗi biến được đo.

Loại Heatmap này đôi khi được gọi là Clustering Heatmap hoặc Clustered, vì mục tiêu của loại biểu đồ này là xây dựng mối liên kết giữa cả điểm dữ liệu và tính năng của chúng. Từ đó xem những cá thể nào giống hoặc khác nhau với mục tiêu tương tự về các biến số. Các công cụ phân tích xây dựng loại Heatmap này thường sẽ triển khai phân cụm như một phần của quy trình của chúng. Trường hợp sử dụng này được tìm thấy trong các lĩnh vực như khoa học sinh học, chẳng hạn như khi nghiên cứu những điểm tương đồng trong biểu hiện gen giữa các cá nhân.

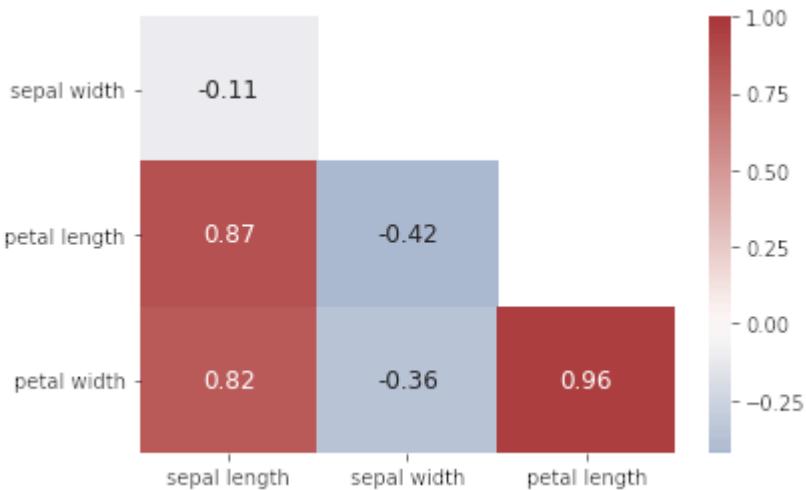


Trong bản đồ clustered heatmap ở trên, mỗi cột biểu thị một mẫu hoa riêng lẻ và mỗi hàng là một phép đo từ mẫu đó.

10.4.22.4.2. Correlogram (Biểu đồ tương quan)

Correlogram là một biến thể của Heatmap thay thế từng biến trên hai trục bằng danh sách các biến số trong tập dữ liệu. Mỗi ô mô tả mối quan hệ giữa các biến giao nhau, chẳng hạn như mối tương quan tuyến tính. Đôi khi, những mối tương quan đơn giản này được thay thế bằng những biểu diễn phức tạp hơn, như Scatter plots.

Correlograms thường được thấy trong vai trò thăm dò, giúp các nhà phân tích hiểu được mối quan hệ giữa các biến nhằm phục vụ việc xây dựng các mô hình thống kê mô tả hoặc dự đoán.

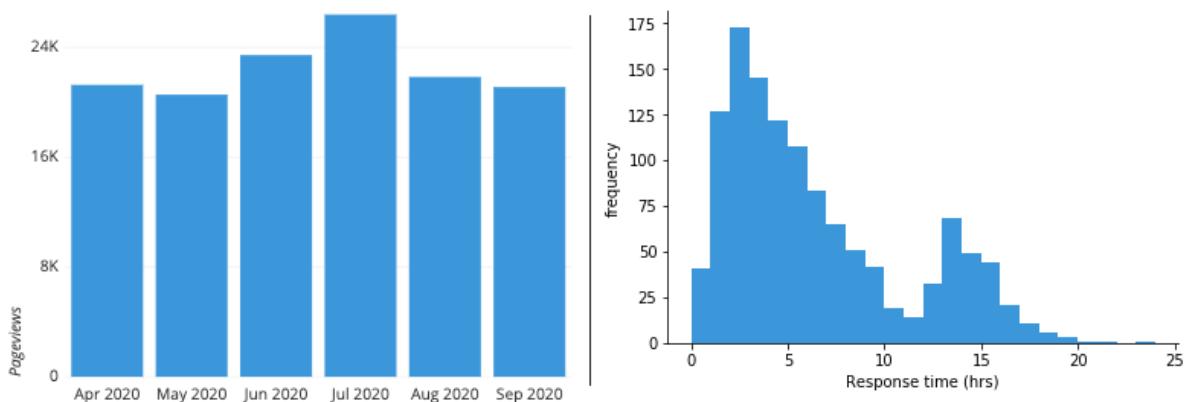


Chiều dài cánh hoa (petal length) có mối tương quan chặt chẽ với chiều rộng cánh hoa (petal width) và chiều dài đài hoa (sepal length); chiều dài đài hoa có tương quan nghịch với ba biến còn lại.

10.4.22.5. Các đồ thị liên quan

10.4.22.5.1. Bar chart and histogram

Các điểm tương tự một chiều gần nhất của Heatmap là Bar chart và Histogram, tương ứng với dữ liệu phân loại và số. Đối với những biểu đồ này, độ dài thanh là chỉ báo về giá trị, thay vì màu sắc. (Mặc dù cần lưu ý rằng các thanh biểu đồ có xu hướng chỉ mô tả thông tin tần suất – khi số liệu tóm tắt được tính toán trên mỗi bin, thay vào đó, chúng tôi có xu hướng sử dụng Line chart). Các ghi chú thực tiễn tốt nhất về mức thứ tự và đặt dấu kiểm ở trên đèn từ những ghi chú này các loại biểu đồ cơ bản hơn.

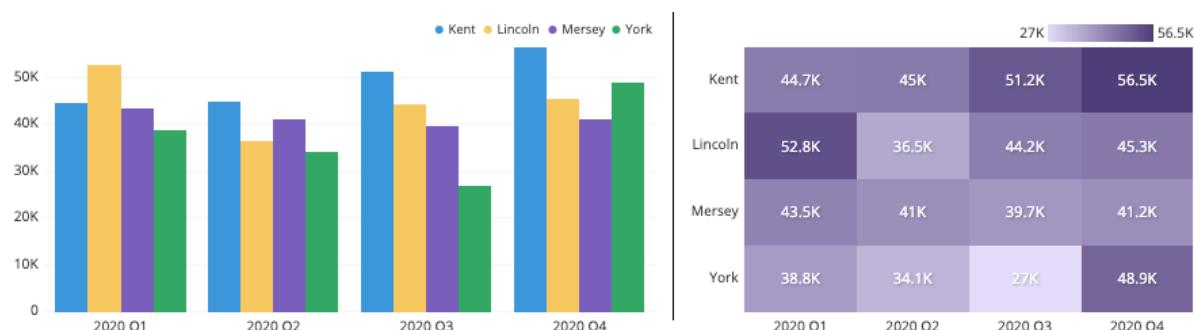


10.4.22.5.2. Grouped bar chart

Một cách khác để hiển thị dữ liệu trong Heatmap là thông qua Grouped bar chart. Mỗi hàng của Heatmap trở thành một cụm thanh và chiều cao của mỗi thanh biểu thị giá trị của ô tương ứng. Thay vào đó, màu được sử dụng để đảm bảo rằng các giá trị cột có thể được theo dõi giữa các cụm.

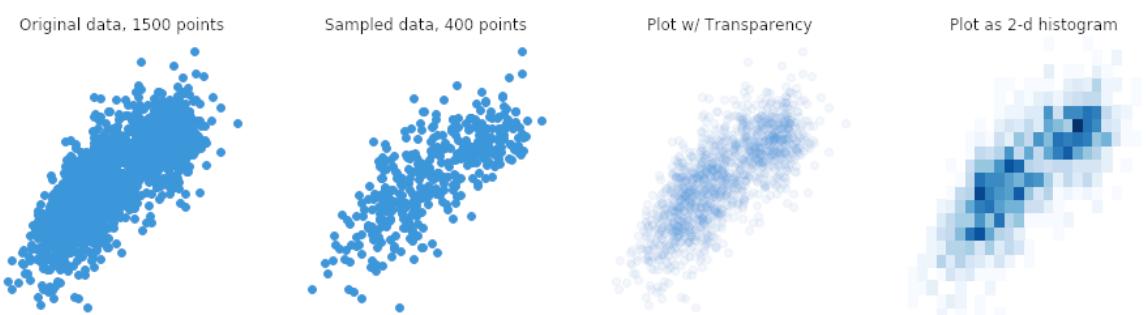
Grouped bar chart được sử dụng khi muốn so sánh chính xác hơn giữa các giá trị ô. Tuy nhiên, chúng là một lựa chọn không tốt khi có nhiều thanh cần được vẽ và khi cả hai biến trực đều có bản chất là số. Trong trường hợp đó, tốt nhất nên sử dụng Heatmap, bản đồ này nhỏ gọn

hơn và thực hiện công việc hiển thị tổng quan rộng hơn trên cả hai biến trực cùng một lúc tốt hơn.



10.4.22.5.3. Scatter plot

Scatter plots dường như không liên quan đến Heatmap vì chúng biểu thị các điểm dữ liệu riêng lẻ theo vị trí thay vì màu sắc. Tuy nhiên, khi có quá nhiều điểm dữ liệu có mức độ trùng lặp cao, điều này có thể che khuất mối quan hệ giữa các biến, một vấn đề được gọi là vẽ biểu đồ quá mức (*an issue called overplotting*). Một trong những lựa chọn để khắc phục tình trạng vẽ đồ thị quá mức là sử dụng bản đồ nhiệt để đếm số điểm rơi vào mỗi bin. Việc sử dụng bản đồ nhiệt này còn được gọi là 2-d histogram (biểu đồ 2 chiều).



10.4.22.5.4. Choropleth

Ngôn ngữ liên kết màu sắc với giá trị không chỉ là lĩnh vực của Heatmap. Một ví dụ cụ thể về loại mã hóa này có thể được thấy trong Choropleth. Choropleth giống như một Heatmap trong đó các giá trị số được mã hóa bằng các vùng màu nhưng những giá trị này được liên kết với các vùng địa lý chứ không phải theo một lưới nghiêm ngặt.

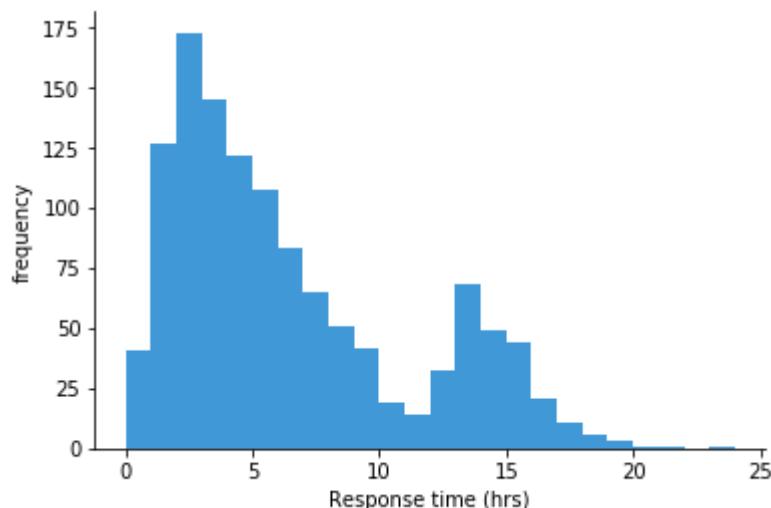
2010 US Population

564K 37.3M

**10.4.23. Histogram****10.4.23.1. Giới thiệu**

- *Histogram* là một dạng biểu đồ thể hiện tần suất dạng cột, hiển thị sự phân bố của một biến. Trục x thể hiện phạm vi và trục y hiển thị tần suất. *Histograms* giúp đưa ra các ước tính về vị trí các giá trị tập trung, các điểm cực trị là gì, có khoảng trống hoặc giá trị bất thường nào không. Bên cạnh đó, chúng cũng hữu ích để đưa ra cái nhìn sơ bộ về phân phối xác suất.
- Thoạt nhìn, histogram trông giống như bar graph. Tuy nhiên, có một sự khác biệt chính giữa chúng là:
 - *Bar graph* biểu thị dữ liệu phân loại
 - *Histogram* biểu thị dữ liệu liên tục.

Histogram là biểu đồ biểu thị sự phân bố các giá trị của một biến số dưới dạng một chuỗi các thanh. Mỗi thanh thường bao gồm một phạm vi giá trị số được gọi là thùng (bin) hoặc lớp (class); chiều cao của thanh cho biết tần suất của các điểm dữ liệu có giá trị trong thùng tương ứng.



Biểu đồ ở trên hiển thị phân bố tần suất theo thời gian phản hồi đối với các yêu cầu về việc đăng ký vé. Chiều ngang của mỗi thanh tương ứng với thời gian là một giờ và chiều cao

cho biết số lượng vé được đăng ký trong thời gian đó. Có thể thấy rằng tần suất phản hồi lớn nhất là trong khoảng 2-3 giờ, với phần đuôi ở bên phải dài hơn bên trái. Ngoài ra còn có một khoảng thời gian (khoảng 13-14 giờ) có số lượng phản hồi tăng vọt hơn những giờ lân cận. Nếu chỉ nhìn vào số liệu thống kê như giá trị trung bình và độ lệch chuẩn, có thể bỏ lỡ thực tế rằng có hai đỉnh này đã đóng góp vào số liệu thống kê tổng thể.

10.4.23.2. Sử dụng

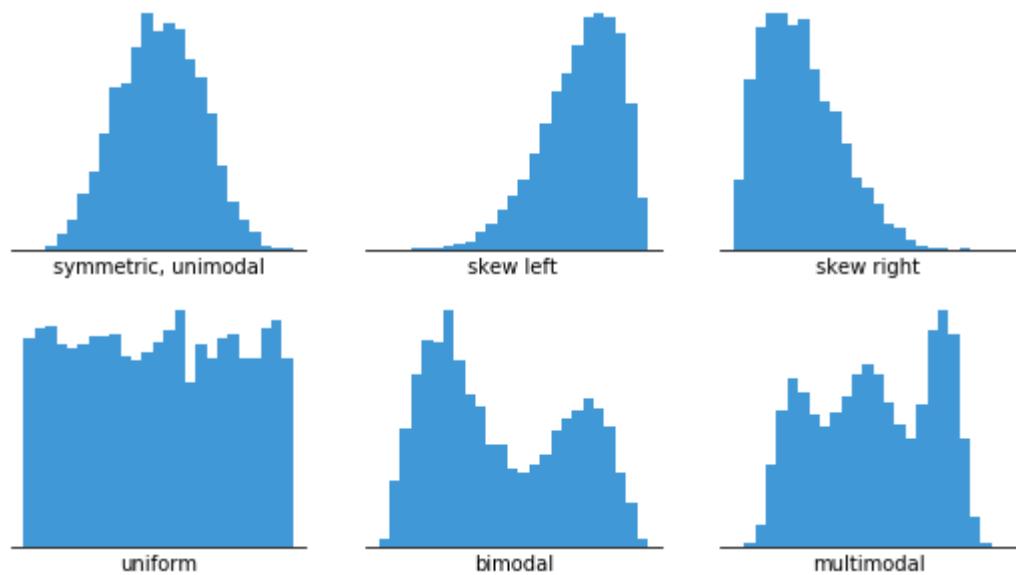
10.4.23.2.1. Những điểm chính khi sử dụng histogram

- Dùng cho dữ liệu định lượng.
- Thường được sử dụng để đếm chứ không hiển thị theo dạng phần trăm.
- Các cột được kết nối
- Phải bỏ dữ liệu vào thùng (bin), tức là trước tiên thực hiện sắp xếp dữ liệu, sau đó chuyển đổi giá trị các đoạn về 1 giá trị đồng nhất.

10.4.23.2.2. Các trường hợp thường dùng của histogram

- Khi dữ liệu có tính liên tục.
- Khi muốn thể hiện hình dạng phân phối của dữ liệu để minh họa số liệu thống kê.

Histogram rất tốt để hiển thị các đặc điểm phân phối chung của các biến số liệu. Có thể thấy gần đúng vị trí của các đỉnh của phân bố, phân bố có bị lệch (skewed) hay đối xứng (symmetric) hay không và liệu có bất kỳ giá trị ngoại lệ (outliers) nào không.



Để sử dụng Histogram, chỉ cần yêu cầu một biến nhận các giá trị số liên tục sẽ được đưa vào trực hoành. Dựa vào đó, các giá trị này sẽ được chia thành các khoảng đều nhau (bất kể giá trị tuyệt đối của chúng là bao nhiêu). Khoảng cách này chính là độ rộng của mỗi bin. Ví dụ: với thang điểm từ 0-100, có thể chia thành 20 thanh (bin), mỗi bin có độ rộng cùng là 5.

10.4.23.2.3. Xác định số lượng bin cho Histogram

Thông tin về số lượng bin và ranh giới của chúng để kiểm đếm các điểm dữ liệu không phải là bản thân dữ liệu. Thay vào đó, việc thiết lập các bin là một quyết định riêng biệt phải

thực hiện khi xây dựng Histogram. Cách chỉ định các bin sẽ có ảnh hưởng lớn đến cách diễn giải biểu đồ, như sẽ thấy bên dưới.

Khi một giá trị nằm trên ranh giới bin, cần nhất quán trong việc xét giá trị đó sẽ được gán cho bin ở bên phải hoặc bên trái của nó (hoặc vào các bin đầu tiên nếu nó nằm ở ngay điểm đầu tiên và vào bin cuối nếu nó nằm ở ngay điểm cuối). Bên nào được chọn tùy thuộc vào công cụ trực quan; một số công cụ có tùy chọn ghi đè tùy chọn mặc định của chúng. Trong tài liệu này, giả định rằng các giá trị trên ranh giới giữa 2 bin sẽ được gán cho bin bên phải.

10.4.23.2.4. Ví dụ về cấu trúc dữ liệu dùng để vẽ đồ thị

- *Cách thông thường:* từ dữ liệu ban đầu, cần thực hiện việc tóm tắt dữ liệu để có bảng gồm 2 cột như minh họa sau, trong đó, cột đầu tiên biểu thị ranh giới của thùng và cột thứ hai cho biết số lượng quan sát trong mỗi thùng.

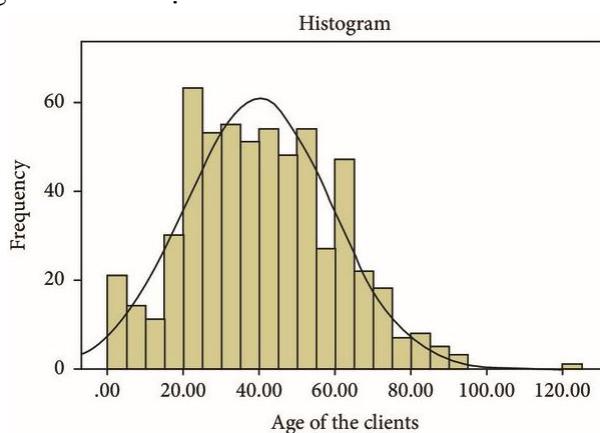
Bin Edges	Frequency
0	41
1	127
2	173
...	...
...	...
23	0
24	1

- *Trong một số công cụ nhất định:* chỉ có thể hoạt động với cột dữ liệu gốc, chưa được tổng hợp, sau đó chỉ định các tham số để tạo nhóm (như first_bin_edge, last_bin_edge, bin_size) cho dữ liệu trước khi tạo histogram.

...	response_delta	(other data columns ...)
	1.6414	...
	5.3931	...
	14.3728	...
	7.3064	...
	2.0908	...

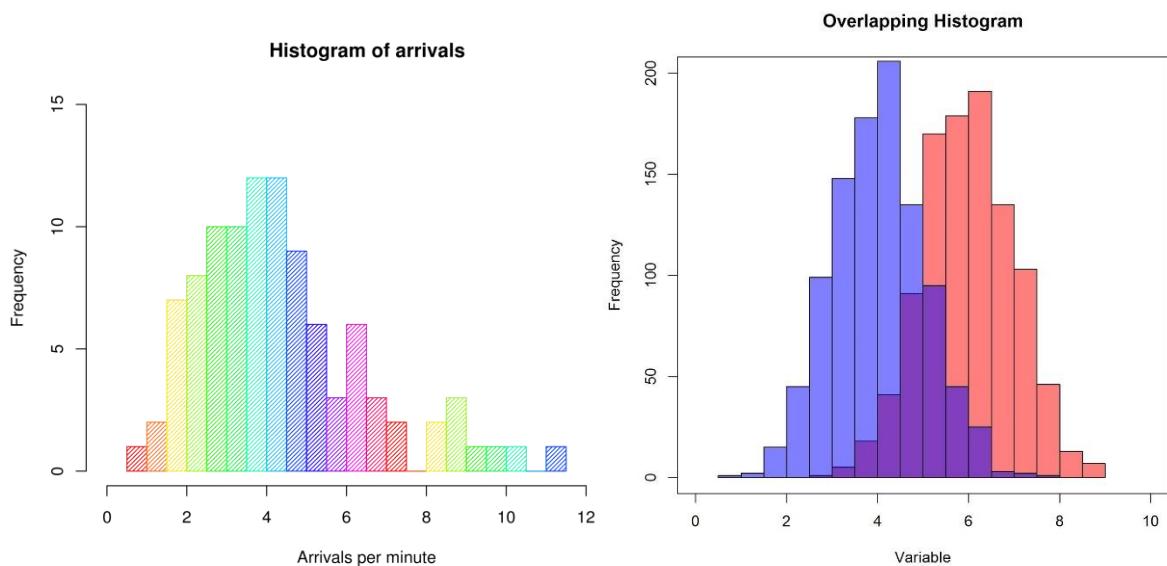
First_bin_edge	0
Last_bin_edge	24
Bin_size	1

- Ví dụ: biểu đồ có thể hiển thị số lượng người thuộc một độ tuổi nhất định trong dân số. Chiều cao hoặc chiều dài của mỗi thanh trong biểu đồ cho biết có bao nhiêu người trong mỗi danh mục.



Mean = 39.79
Std. dev. = 19.44
N = 592

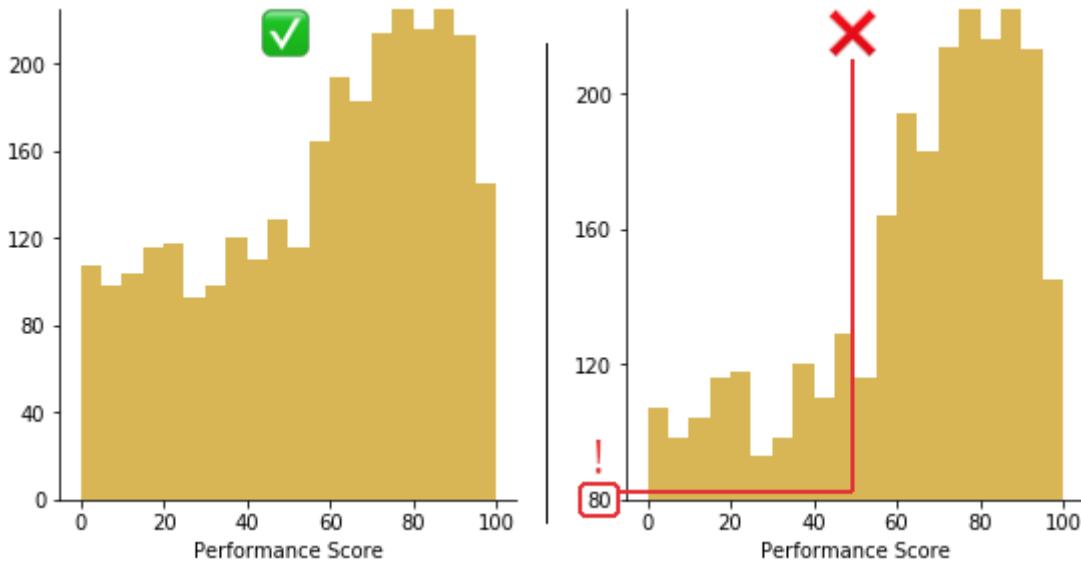
- Một số minh họa khác của Histogram



10.4.23.3. Tăng hiệu quả sử dụng cho histogram

10.4.23.3.1. Sử dụng đường cơ sở với giá trị thấp nhất tính từ 0 (zero)

Một khía cạnh quan trọng của biểu đồ là chúng phải được vẽ với đường cơ sở có giá trị bằng 0. Vì tần suất của dữ liệu trong mỗi bin được biểu thị bằng chiều cao của mỗi thanh nên việc thay đổi đường cơ sở hoặc tạo ra một khoảng trống trong thang đo sẽ làm sai lệch nhận thức về phân bố dữ liệu.

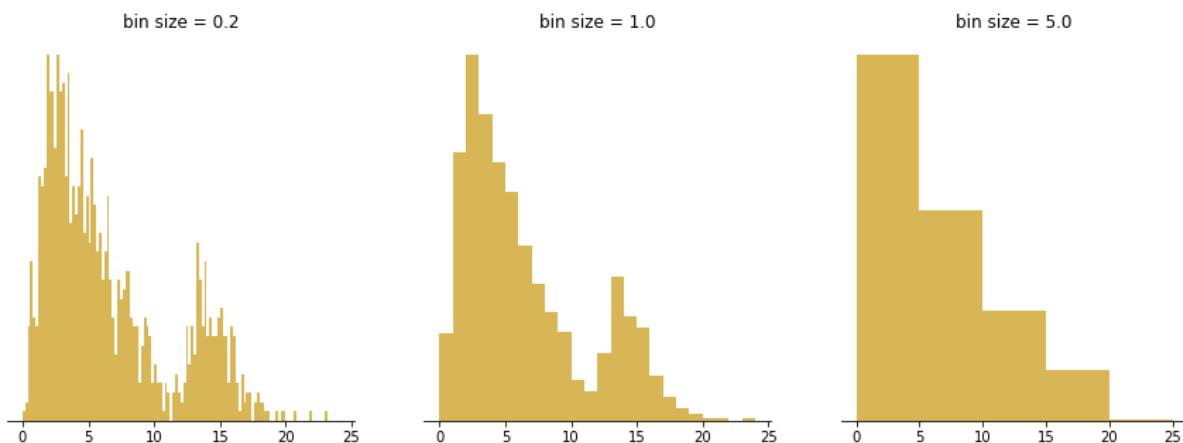


10.4.23.3.2. Chọn số lượng bin phù hợp

Mặc dù các công cụ có thể tạo biểu đồ thường có một số thuật toán mặc định để chọn ranh giới cho các bin, nhưng có thể cần sử dụng kiến thức về lĩnh vực của dữ liệu để thử nghiệm nhiều tùy chọn khác nhau từ đó sẽ tìm ra giá trị phù hợp với mục đích của việc trừu tượng hóa dữ liệu.

Việc lựa chọn kích thước bin có mối quan hệ nghịch đảo với số lượng thùng. Kích thước thùng càng lớn thì càng có ít thùng để bao phủ toàn bộ phạm vi dữ liệu. Với kích thước bin càng nhỏ thì càng cần nhiều bin hơn. Nên dành chút thời gian để kiểm tra các kích thước bin khác nhau để xem cách phân phối trong mỗi bin, sau đó chọn biểu đồ thể hiện dữ liệu tốt nhất.

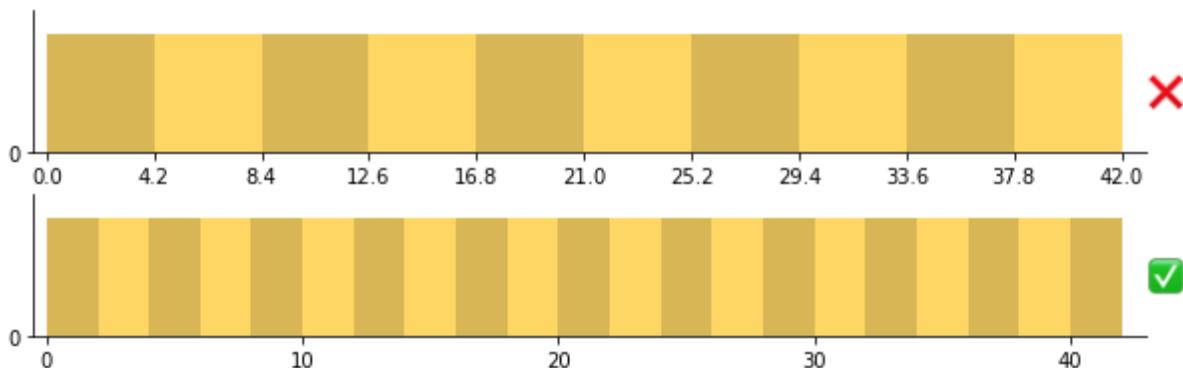
Nếu có quá nhiều bin thì việc phân phối dữ liệu sẽ trông thô và sẽ khó phân biệt tín hiệu với nhiễu. Ngược lại, với quá ít thùng, biểu đồ sẽ thiếu các chi tiết cần thiết để phân biệt bát kỳ mẫu hữu ích nào từ dữ liệu.



Các bin bên trái có kích thước quá nhỏ, hàm ý có rất nhiều đỉnh và đáy giả. Ngược lại các bin bên phải lại có kích thước quá lớn, che giấu mọi dấu hiệu về đỉnh thứ hai.

10.4.23.3.3. Chọn giá trị làm ranh giới cho các bin

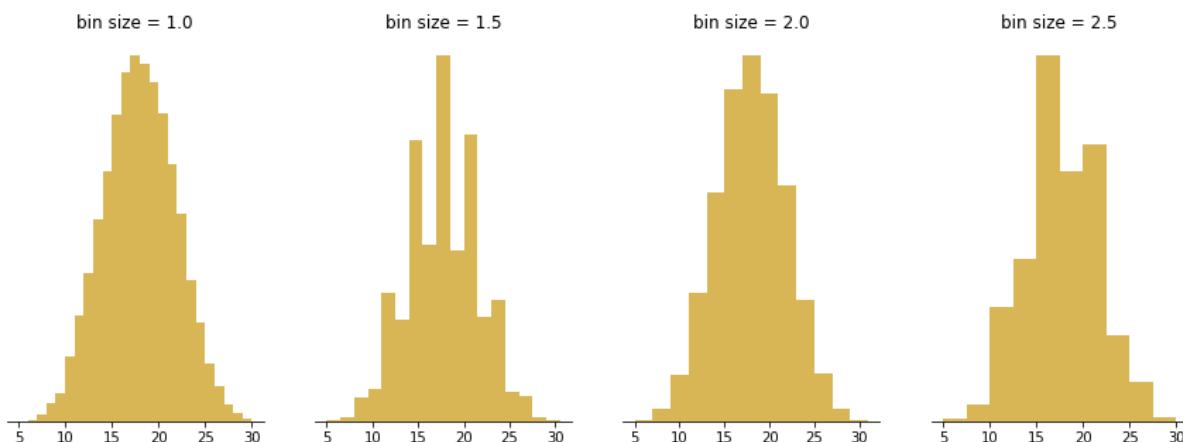
- Các dấu kiểm và nhãn thường phải nằm trên ranh giới bin để thông báo rõ nhất giới hạn của mỗi thanh nằm ở đâu. Không cần phải đặt nhãn cho mỗi thanh nhưng việc đặt chúng ở giữa một số thanh sẽ giúp người đọc theo dõi giá trị. Ngoài ra, sẽ rất hữu ích nếu nhãn là các giá trị chỉ có một số ít chữ số có nghĩa để dễ đọc.
- Điều này gợi ý rằng các bin có kích thước 1, 2, 2,5, 4 hoặc 5 (chia đều cho 5, 10 và 20) hoặc lũy thừa 10 là những kích thước bin tốt. Điều này cũng có nghĩa là các bin có kích thước 3, 7 hoặc 9 có thể sẽ khó đọc hơn và không nên sử dụng trừ khi ngữ cảnh có ý nghĩa đối với chúng.



Hình minh họa trên cho thấy việc chia dữ liệu một cách bát cẩn thành 10 bin từ tối thiểu đến tối đa có thể dẫn đến một số cách chia bin rất kỳ quặc, khó hiểu. Hình bên dưới sử dụng cách chia kích thước bin theo thang 10 giúp dễ theo dõi hơn rất nhiều.

- Khi giá trị của biến là số nguyên, nhưng kích thước bin lại được phân chia thành số lẻ (ví dụ 2,5): Đây có thể là vấn đề vì bin chạy từ 0 đến 2,5 có cơ hội thu thập ba giá trị nguyên khác nhau (0, 1, 2) nhưng bin từ 2,5 đến 5 chỉ có thể thu thập hai giá trị khác nhau (3, 4) do giá trị 5 sẽ rơi vào thùng sau. Điều này có nghĩa là biểu đồ

có thể trông “gập ghèn” một cách bất thường chỉ do số lượng giá trị mà mỗi thùng có thể chứa.



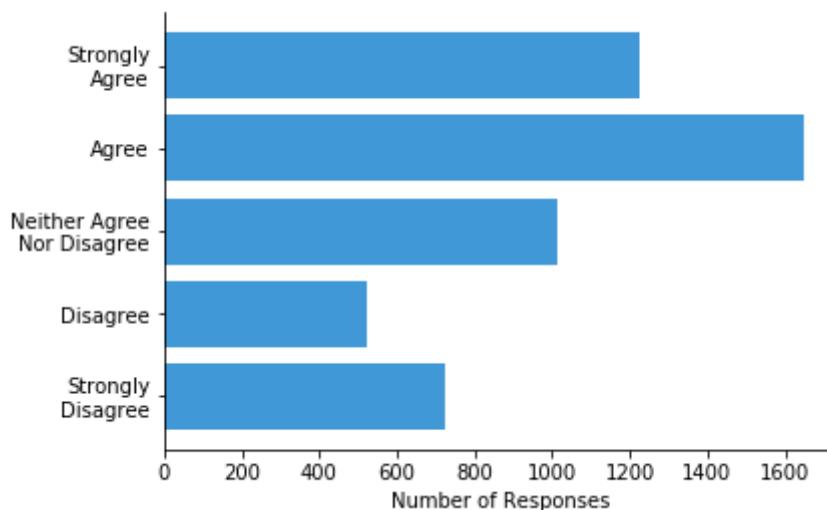
Hình trên minh họa: cùng 1 dữ liệu nhưng với kích thước bin thay đổi sẽ dẫn đến sự phân bố kết quả trên biểu đồ rất khác nhau.

10.4.23.4. Những trường hợp sử dụng sai thường gặp (Common misuses)

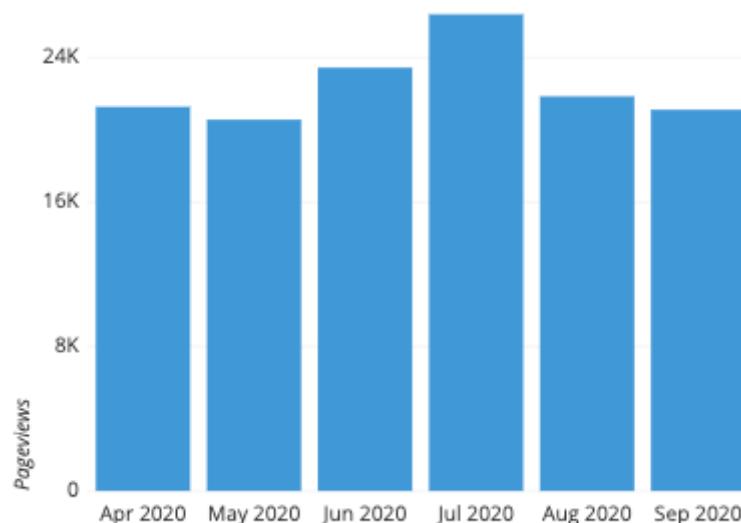
10.4.23.4.1. Sử dụng biến đo không phải là số liên tục

Như đã lưu ý trong phần mở đầu, histogram được dùng để mô tả sự phân bố tần suất của một biến số liên tục. Khi biến quan tâm không phù hợp với thuộc tính này, thay vào đó cần sử dụng một loại biểu đồ khác: Bar chart.

- Đối với biến kiểu phân loại (định danh): như loại người dùng (ví dụ: khách hàng, nhà cung cấp, người dùng, ...) hoặc vị trí (khu vực, lớp, hàng sản xuất, ...) rõ ràng không phải là số và do đó nên sử dụng Bar chart.
- Đối với biến kiểu phân loại nhưng có thứ tự: ví dụ nếu có các câu trả lời khảo sát theo thang điểm từ 1 đến 5, để mã hóa các giá trị từ “hoàn toàn không đồng ý” đến “hoàn toàn đồng ý” thì phân bố tần suất sẽ được hiển thị dưới dạng Bar chart. Lý do là sự khác biệt giữa các giá trị riêng lẻ có thể không nhất quán: thực sự không biết rằng sự khác biệt có ý nghĩa giữa điểm 1 và 2 (“rất không đồng ý” đến “không đồng ý”) cũng giống như sự khác biệt giữa điểm 2 và 3 (“không đồng ý” đến “không đồng ý cũng không phản đối”). Khi đó nên sử dụng Bar chart.



- *Đối với biến kiểu số rời rạc:* (ví dụ: số nguyên 1, 2, 3, v.v.) có thể được vẽ bằng Bar chart hoặc Histogram, tùy thuộc vào ngữ cảnh. Việc sử dụng Histogram sẽ có nhiều khả năng hơn khi có nhiều giá trị khác nhau cần vẽ. Khi phạm vi (range) của các giá trị số là lớn và khoảng cách giữa các giá trị xa nhau và không đều thì việc nhóm chúng thành liên tục sẽ là một ý tưởng hay.
- *Đối với biến kiểu thời gian:*
 - Khi các giá trị tương ứng với các khoảng thời gian tương đối (ví dụ: 30 giây, 20 phút), thì việc gộp các khoảng thời gian cho biểu đồ sẽ có ý nghĩa.
 - Khi các giá trị tương ứng với thời gian là giá trị tuyệt đối (ví dụ: ngày 10 tháng 1, 12:15) thì sự khác biệt sẽ trở nên mờ nhạt. Khi các điểm dữ liệu mới được ghi lại, các giá trị thường sẽ được đưa vào các bin mới được tạo thay vì nằm trong phạm vi các bin hiện có.
 - Ngoài ra, một số lựa chọn nhóm tự nhiên nhất định, chẳng hạn như theo tháng hoặc quý, đưa ra kích thước bin hơi không đồng đều. Vì những lý do này, không quá bất thường khi thấy một loại biểu đồ khác như Bar chart hoặc Line chart được sử dụng.

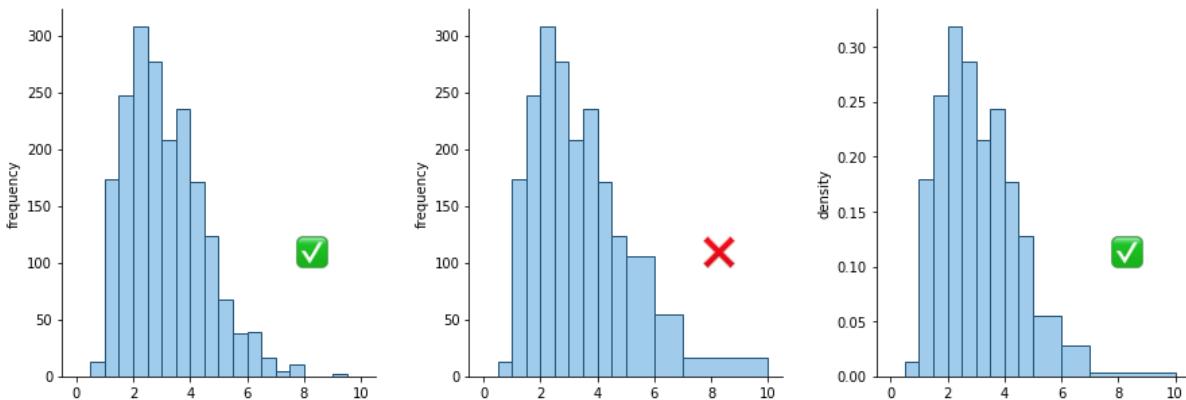


10.4.23.4.2. Sử dụng kích thước của bin không đồng đều

Mặc dù tất cả các ví dụ cho đến nay đều hiển thị histogram bằng cách sử dụng các bin có kích thước bằng nhau nhưng đây thực sự không phải là yêu cầu kỹ thuật. Khi dữ liệu thưa thớt, chẳng hạn như khi có đuôi dữ liệu dài, bạn có thể này ra ý tưởng sử dụng chiều rộng bin lớn hơn để che phủ không gian đó. Tuy nhiên, việc tạo biểu đồ với các bin có kích thước không bằng nhau không hẳn là một sai lầm, nhưng làm như vậy đòi hỏi một số thay đổi lớn trong cách tạo biểu đồ và có thể gây ra nhiều khó khăn trong việc diễn giải.

- Kích thước bin nhất quán: Điểm kỹ thuật về histogram là tổng diện tích của các thanh biểu thị toàn bộ dữ liệu và diện tích mà mỗi thanh chiếm giữ biểu thị tỷ lệ của mỗi thùng trong toàn bộ dữ liệu. Khi kích thước bin nhất quán, điều này làm cho diện tích thanh đo và chiều cao tương đương nhau.
- Kích thước bin thay đổi: chiều cao không còn tương ứng với tổng tần suất xuất hiện. Làm như vậy sẽ làm sai lệch nhận thức về số lượng điểm trong mỗi thùng, vì việc tăng kích thước thùng sẽ chỉ khiến nó trông lớn hơn. Trong đồ thị ở giữa của hình bên dưới, các thùng từ 5-6, 6-7 và 7-10 trông có vẻ như chứa nhiều điểm hơn thực tế.

Thay vào đó, trực tung cần mã hóa mật độ tần suất trên một đơn vị kích thước thùng. Ví dụ, ở khung bên phải của hình trên, thùng từ 2-2,5 có chiều cao khoảng 0,32. Nhận với chiều rộng thùng là 0,5 và có thể ước tính khoảng 16% dữ liệu trong thùng đó. Chiều cao của các ngăn rộng hơn đã được thu nhỏ lại so với khung trung tâm: hãy lưu ý hình dạng tổng thể trông giống với biểu đồ ban đầu như thế nào với các kích thước ngăn bằng nhau. Mật độ không phải là một khái niệm dễ nắm bắt, và một biểu đồ như vậy được trình bày cho những người không quen với khái niệm này sẽ gặp khó khăn trong việc diễn giải nó.



Hình bên trái: biểu đồ với các bin có kích thước bằng nhau; Hình giữa: biểu đồ có các bin không bằng nhau nhưng đơn vị trực đọc không đúng; Hình phải: biểu đồ với các thùng không bằng nhau với chiều cao mật độ

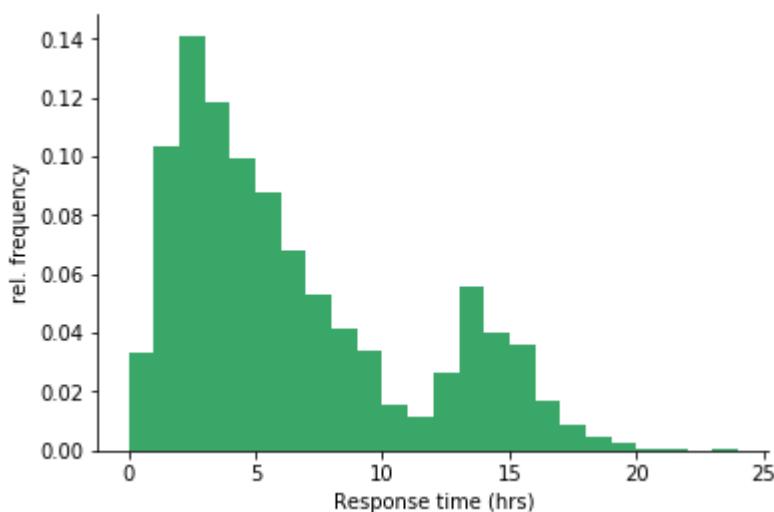
Vì tất cả những lý do vừa nêu ở trên, lời khuyên tốt nhất là hãy thử và chỉ sử dụng các kích thước bin hoàn toàn bằng nhau. Sự hiện diện của các bin trống và một số nhiễu tăng lên trong các phạm vi có dữ liệu thưa thớt thường sẽ góp phần làm tăng khả năng diễn giải của biểu đồ. Mặt khác, nếu có các khía cạnh có hữu của biến được vẽ biểu thị

cho thấy kích thước bin không đồng đều thì thay vì sử dụng biểu đồ bin không đồng đều (*uneven-bin histogram*), nên sử dụng Bar chart.

10.4.23.5. Các tùy chọn thường dùng kèm với histogram

10.4.23.5.1. Tần suất tuyệt đối và tần suất tương đối (Absolute frequency vs. relative frequency)

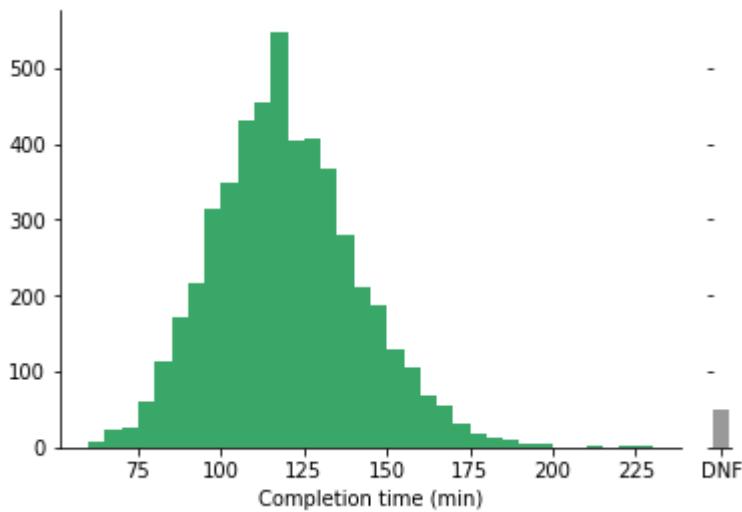
Tùy thuộc vào mục tiêu trực quan hóa, có thể muốn thay đổi các đơn vị trên trục tung của biểu đồ theo tần suất tuyệt đối hoặc tần suất tương đối. Tần suất tuyệt đối chỉ là số lần xuất hiện tự nhiên trong mỗi bin, trong khi tần suất tương đối là tỷ lệ số lần xuất hiện trong mỗi thùng. Việc lựa chọn đơn vị trực sẽ phụ thuộc vào loại so sánh mà người trình bày muốn nhấn mạnh về phân bố dữ liệu.



Chuyển đổi ví dụ đầu tiên thành tần suất tương đối, việc cộng năm thanh đầu tiên sẽ dễ dàng hơn nhiều để thấy rằng khoảng một nửa số yêu cầu được phản hồi trong vòng năm giờ.

10.4.23.5.2. Hiển thị thông tin về các dữ liệu chưa biết hoặc thiếu (unknown or missing data)

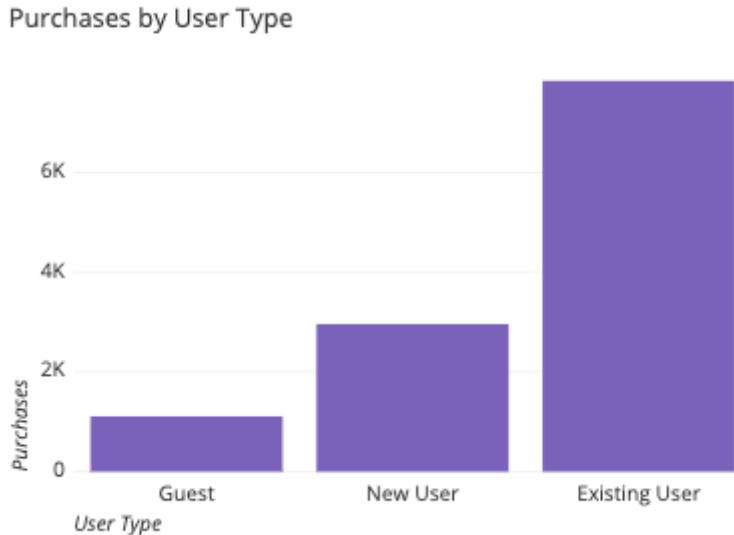
Đây thực sự không phải là một lựa chọn đặc biệt phổ biến, nhưng đáng để cân nhắc khi tùy chỉnh các ô. Nếu một hàng dữ liệu thiếu giá trị cho biên quan tâm thì giá trị đó thường sẽ bị bỏ qua trong bảng kiểm cho mỗi bin. Nếu việc hiển thị số lượng giá trị bị thiếu hoặc chưa biết là quan trọng thì có thể kết hợp biểu đồ với một thanh bổ sung mô tả tần suất của những giá trị chưa biết này. Khi vẽ Bar chart này, nên đặt nó trên một trục song song với biểu đồ chính và có màu trung tính khác để các điểm được thu thập trong thanh đó không bị nhầm lẫn với việc có giá trị số.



10.4.23.6. Các đồ thị liên quan

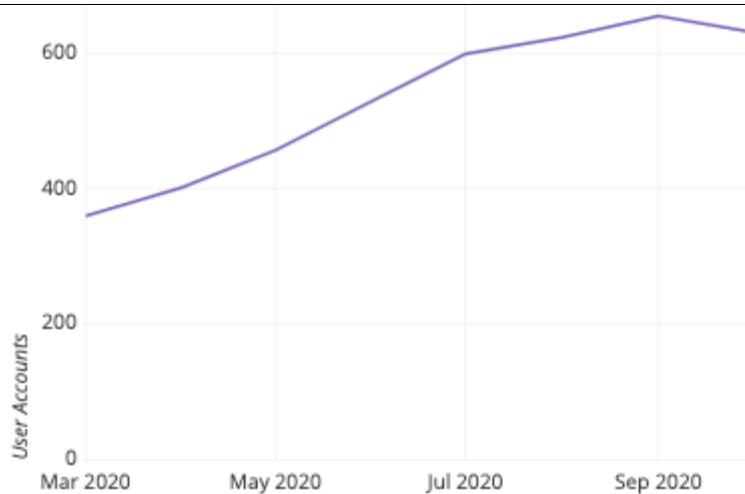
10.4.23.6.1. Bar chart

Như đã lưu ý ở trên, nếu biến quan tâm không phải là biến liên tục và dạng số mà thay vào đó là biến rời rạc hoặc phân loại, thì cần sử dụng Bar chart. Ngược lại với Histogram, các thanh trên Bar chart thường có một khoảng cách nhỏ với nhau: điều này nhấn mạnh tính chất rời rạc của biến được vẽ.



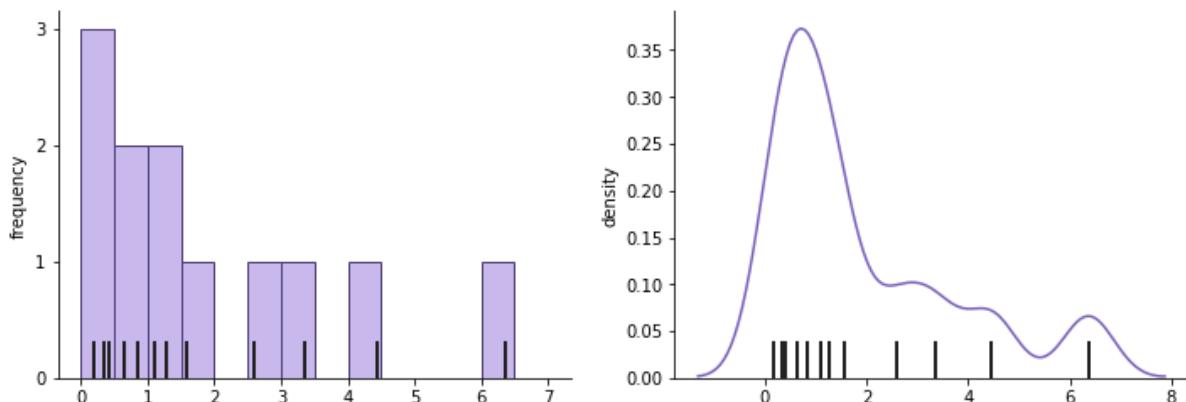
10.4.23.6.2. Line chart

Nếu dữ liệu số được gộp nhưng muốn trực tung của biểu đồ truyền tải nội dung nào đó ngoài thông tin tần suất thì nên hướng tới việc sử dụng Line chart. Vị trí thẳng đứng của các điểm trong Line chart có thể mô tả các giá trị hoặc tóm tắt thống kê của biến thứ hai. Khi Line chart được sử dụng để mô tả sự phân bố tần suất như Histogram thì biểu đồ này được gọi là đa giác tần suất (frequency polygon).



10.4.23.6.3. Density curve

Đường cong mật độ (Density curve) hoặc ước tính mật độ hạt nhân (Kernel Density Estimate - KDE), là một thay thế cho Histogram cung cấp cho mỗi điểm dữ liệu sự đóng góp liên tục vào phân phối. Trong Histogram, có thể coi mỗi điểm dữ liệu như việc đổ chất lỏng từ giá trị của nó vào một loạt hình trụ bên dưới (các bin). Trong KDE, mỗi điểm dữ liệu sẽ thêm một khối lượng nhỏ xung quanh giá trị thực của nó, được xếp chồng lên nhau trên các điểm dữ liệu để tạo đường cong cuối cùng. Hình dạng của khối lượng là 'hạt nhân' (kernel) và có vô số lựa chọn. Do có rất nhiều tùy chọn khi chọn hạt nhân và các tham số của nó, density curves thường là miền của các công cụ trực quan hóa theo chương trình.

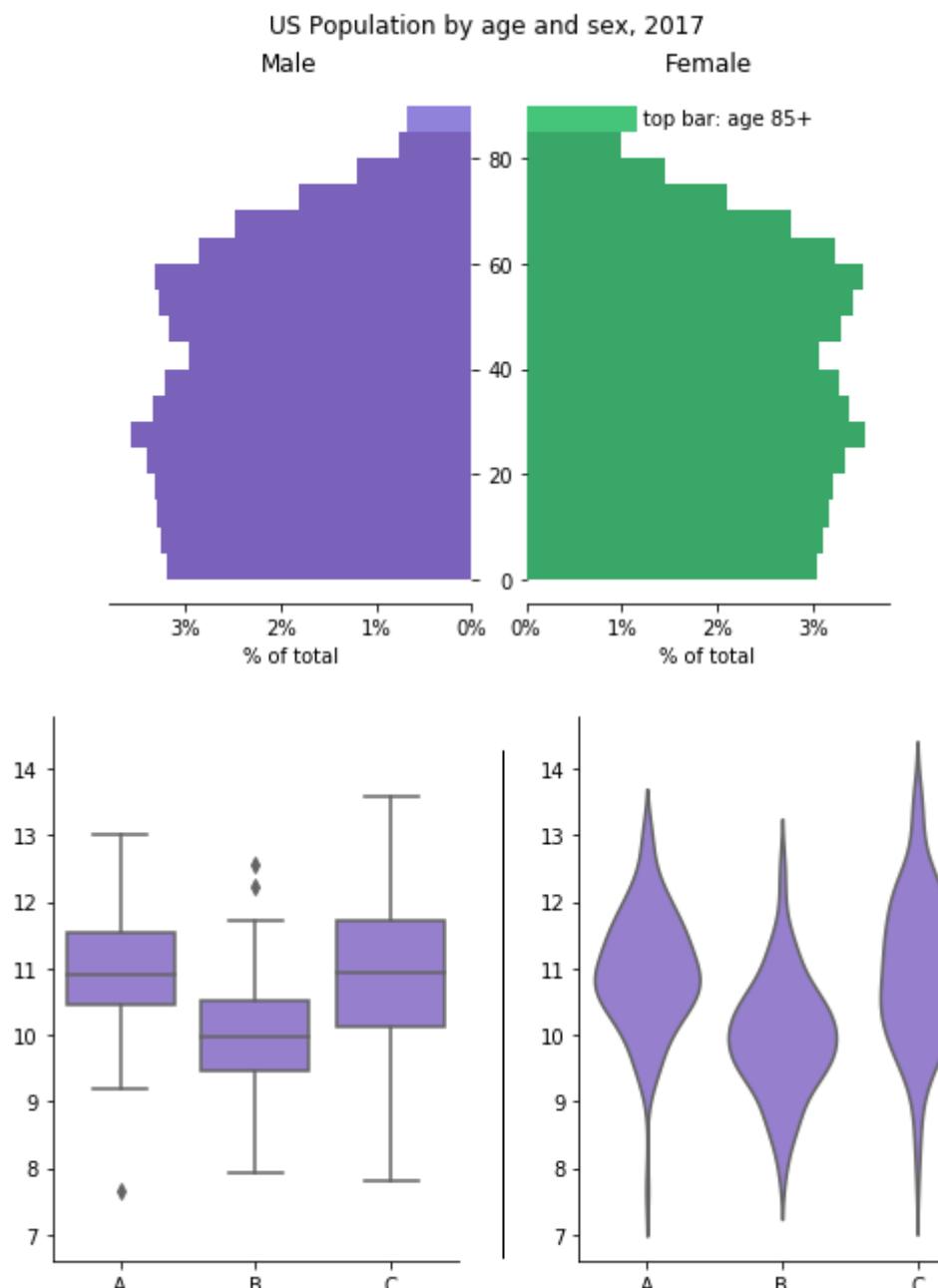


Các dấu gạch ngang màu đen dày biểu thị các điểm dữ liệu góp phần vào Histogram (trái) và density curve (phải). Lưu ý mỗi điểm đóng góp một đường cong hình chuông nhỏ vào hình dạng tổng thể như thế nào.

10.4.23.6.4. Box plot và violin plot

Histograms rất hữu ích trong việc hiển thị sự phân bố của một biến duy nhất, nhưng sẽ hơi khó để so sánh giữa các Histogram nếu muốn so sánh cùng biến đó nhưng giữa các nhóm khác nhau.

Với hai nhóm, một giải pháp khả thi là vẽ Histograms của hai nhóm đối lưng nhau (*back-to-back*). Một phiên bản theo miền cụ thể của loại biểu đồ này là Population pyramid (kim tự tháp dân số), biểu thị sự phân bố độ tuổi của một quốc gia hoặc khu vực khác đối với nam và nữ dưới dạng biểu đồ dọc đối lưng với nhau.



Tuy nhiên, khi có ba nhóm trở lên, giải pháp giáp lùng sẽ không hiệu quả. Một giải pháp có thể là tạo biểu đồ các mặt, vẽ biểu đồ cho mỗi nhóm trong một hàng hoặc cột. Một cách khác là sử dụng một loại đồ thị khác, chẳng hạn như Box plot hoặc Violin plot. Cả hai loại biểu đồ này thường được sử dụng khi muốn so sánh sự phân bố của một biến số giữa các cấp độ của một biến phân loại. So với biểu đồ nhiều mặt, các biểu đồ này đánh đổi sự mô tả chính xác về tần suất tuyệt đối để so sánh các phân bố tương đối nhỏ gọn hơn.

10.4.24. Line graph (Biểu đồ đường)

10.4.24.1. Giới thiệu

Line chart (còn gọi là Line plot, Line graph) sử dụng các điểm được kết nối bằng các đoạn đường từ trái sang phải để thể hiện sự thay đổi về giá trị. Trục hoành mô tả một tiến trình liên tục, thường là theo thời gian, trong khi trục tung báo cáo các giá trị cho số liệu quan tâm trong tiến trình đó.

ZZD to QQY Exchange Rates



Line charts trên cho thấy tỷ giá hối đoái giữa hai loại tiền tệ trong khoảng thời gian sáu tháng. Khi thời gian trôi qua từ trái sang phải, các điểm sẽ kết nối tỷ giá hối đoái hàng ngày. Có thể đọc từ độ dốc chung của đường và các vị trí thẳng đứng của nó rằng tỷ lệ này đã cải thiện từ khoảng 0,75 lên 0,78 trong khoảng thời gian từ tháng 3 đến đầu tháng 4, sau đó giảm dần xuống khoảng 0,765 vào cuối tháng 5 và tháng 6.

- *Line charts* là một loại biểu đồ minh họa cách dữ liệu liên quan thay đổi trong một khoảng thời gian cụ thể.
- *Line charts* thường có hai trục:
 - Trục X: (trục hoành) Trục này thường dùng để hiển thị dòng thời gian, như ngày, tháng hoặc năm. Các giá trị âm có thể được hiển thị bên dưới trục x. Trục X là trục độc lập vì các giá trị không phụ thuộc vào bất kỳ thứ gì (vì thời gian luôn chuyển động tiến về phía trước bất kể các yếu tố khác).
 - Trục Y: (trục tung) Trục này thường liên quan đến số lượng hay giá trị (hàng bán/lợi nhuận/ ...). Trục Y là trục phụ thuộc vì các giá trị của nó phụ thuộc vào các giá trị trên trục X. Mỗi giá trị X chỉ có một giá trị Y, dẫn đến một đường tiến triển theo chiều ngang. Ví dụ: Xét 1 công ty, vào một ngày nhất định (X) sẽ luôn có duy nhất 1 giá trị về lợi nhuận (Y).
- *Line charts* có thể có nhiều đường trên cùng một trục mà người đọc thường so sánh với nhau. Ví dụ: có thể tạo *Line graph* hiển thị sự khác biệt về lợi nhuận mà mỗi bộ phận tạo ra trong một công ty. Mỗi bộ phận có thể có một dòng riêng với màu sắc riêng và một chú thích trên biểu đồ có thể giải thích màu nào tương ứng với đường kẻ nào nào. Tuy nhiên, tránh sử dụng nhiều hơn 3-4 dòng trên mỗi biểu đồ, vì điều này làm cho biểu đồ trở nên lộn xộn và khó đọc.

10.4.24.2. Sử dụng

10.4.24.2.1. Khi nào nên dùng *Line charts*?

- *Diễn dịch thông tin từ số liệu sang hình ảnh*: Một trong những lý do chính khiến mọi người sử dụng *Line charts* là vì chúng dễ đọc. Các công ty thường nắm bắt rất

nhiều thông tin trong cơ sở dữ liệu hoặc bảng tính. Bằng cách tạo Line charts, giúp trình bày thông tin theo cách mới, giúp mọi người hiểu dữ liệu dễ dàng hơn.

- So sánh các điểm dữ liệu: Line charts cũng có thể hữu ích để biểu thị và giải thích nhiều điểm dữ liệu.

Ví dụ: cần phân tích 6 hoạt động cơ bản của những người lớn tuổi (giả sử chỉ quan tâm khoảng tuổi từ 65-90) gồm ăn, tắm, mặc quần áo, vào hoặc ra khỏi giường, sử dụng nhà vệ sinh và đi bộ. Người ta cần lập biểu đồ thể hiện mức độ khó khi thực hiện các hoạt động hàng ngày dựa trên độ tuổi của một cá nhân. Khi đó, độ tuổi (từ 60-90) được đặt trên trục x, trục y có thể chứa phần trăm độ khó của từng hoạt động. Có sáu đường kẻ mỗi đường kẻ (có màu khác nhau) thể hiện các hoạt động khác nhau.

- Line chart thường được sử dụng thường xuyên để hiển thị xu hướng và phân tích dữ liệu đã thay đổi như thế nào theo thời gian. Độ dốc của biểu đồ hướng lên cho biết nơi giá trị đã tăng và độ dốc hướng xuống cho biết nơi giá trị đã giảm.
- Line chart rất hữu ích để minh họa các xu hướng như thay đổi nhiệt độ trong những ngày nhất định, số lượng hàng bán theo ngày, ...

10.4.24.2.2. Bố cục của Line chart

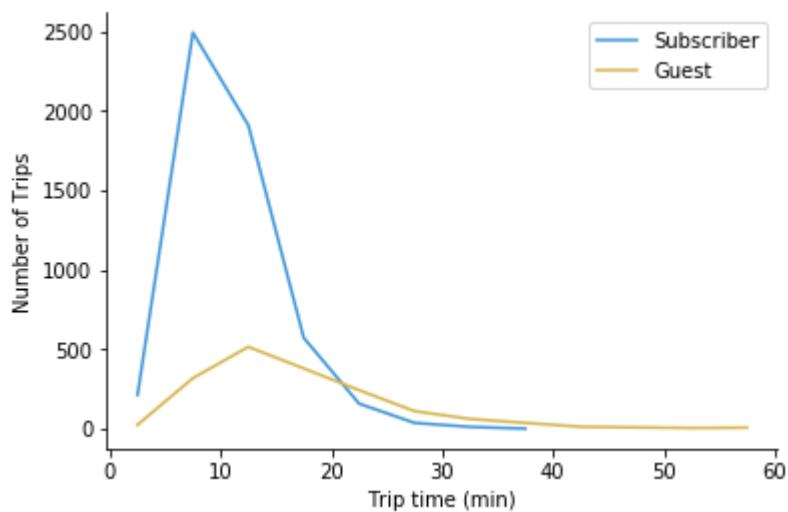
Line chart được sử dụng khi muốn nhấn mạnh những thay đổi về giá trị của một biến (được vẽ trên trục tung) cho các giá trị liên tục của biến thứ hai (được vẽ trên trục ngang). Sự nhấn mạnh vào các mô hình thay đổi này được thể hiện bằng các đoạn đường di chuyển nhất quán từ trái sang phải và quan sát độ dốc của các đường di chuyển lên hoặc xuống.

Trên trục hoành, cần một biến mô tả các giá trị liên tục có khoảng đo đều đặn. Trường hợp phổ biến là biến này đại diện cho thời điểm quan sát (mỗi phút, giờ, ngày, tuần hoặc tháng). Việc lựa chọn kích thước khoảng hoặc thùng (bin) là một quyết định mà nhà phân tích thường cần đưa ra đối với dữ liệu, thay vì nó là một đặc tính vốn có của dữ liệu.

Trục tung sẽ cho biết giá trị của biến số thứ hai cho các điểm nằm trong mỗi khoảng được xác định bởi biến trục hoành. Thông thường, đây sẽ là bản tóm tắt thống kê như tổng giá trị hoặc giá trị trung bình của các sự kiện trong mỗi bin.

10.4.24.2.3. Line chart với nhiều đồ thị đơn

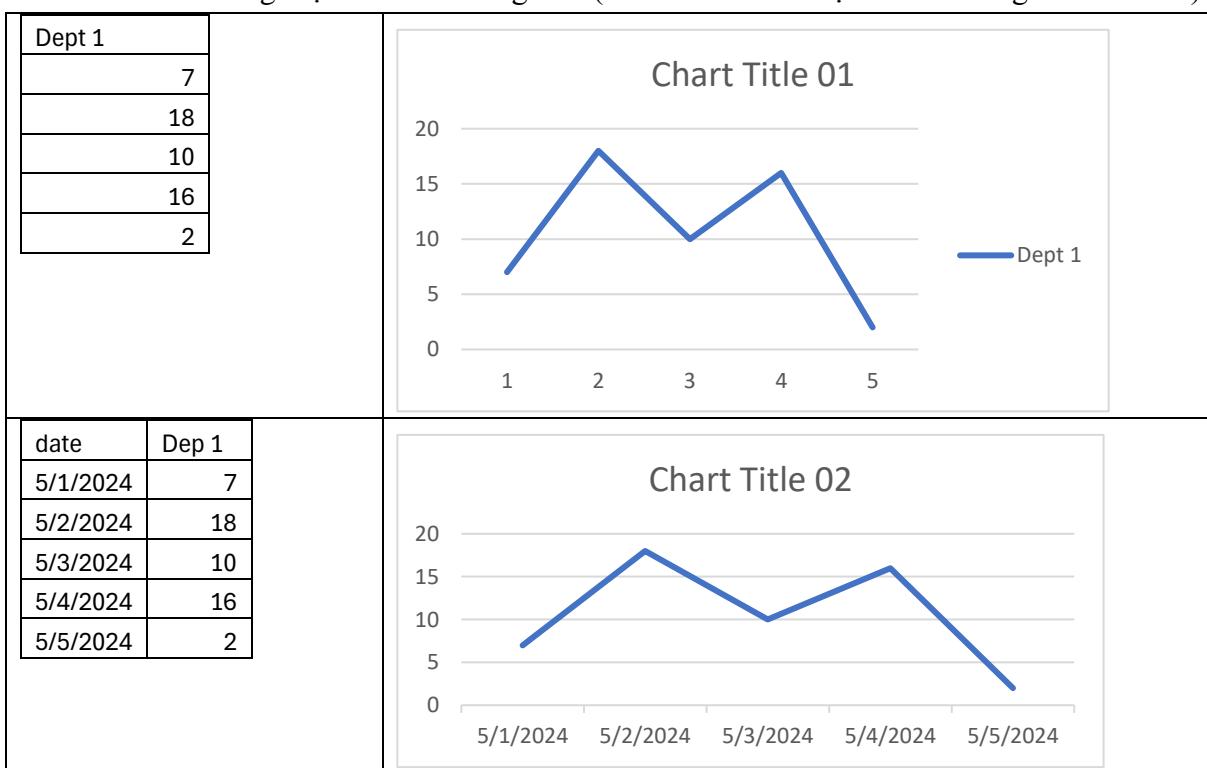
Nhiều đường cũng có thể được vẽ trong một biểu đồ đường đơn (single line chart) để so sánh xu hướng giữa các chuỗi. Trường hợp sử dụng phổ biến cho việc này là quan sát sự phân tích dữ liệu giữa các nhóm con khác nhau. Khả năng vẽ nhiều đường cũng mang lại cho biểu đồ đường một trường hợp sử dụng đặc biệt mà biểu đồ đường thường không được chọn. Thông thường, người ta sẽ sử dụng Histogram để mô tả phân bố tần suất của một biến số. Tuy nhiên, vì rất khó để vẽ hai Histogram trên cùng một bộ trục nên Line chart đóng vai trò là một phương thức so sánh tốt để thay thế. Line chart dùng để mô tả sự phân bố tần suất thường được gọi là đa giác tần suất (Frequency polygons).

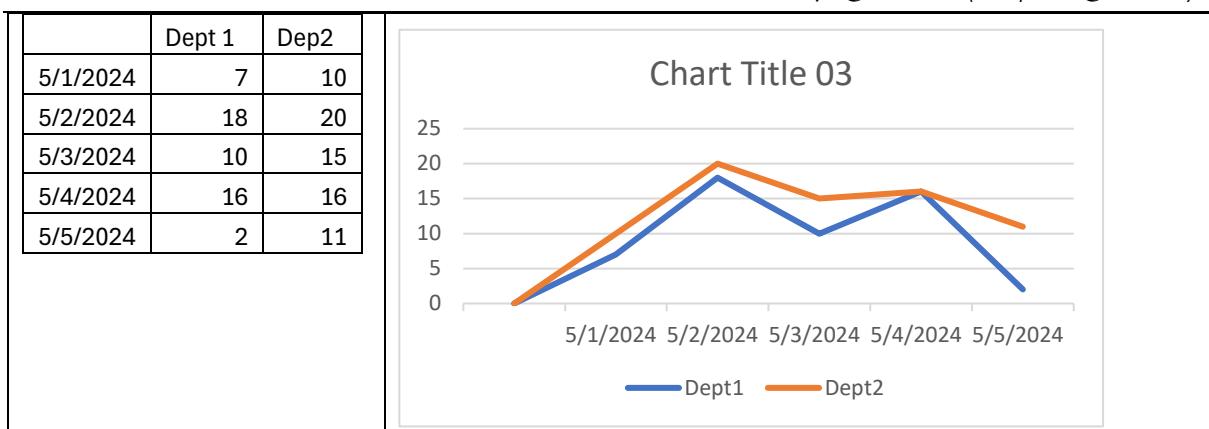


Line chart ở trên cho thấy số chuyến đi của người đăng ký nhiều hơn số lượng khách nhưng trung bình khách có xu hướng thực hiện các chuyến đi dài hơn.

10.4.24.2.4. Cấu trúc dữ liệu dùng để vẽ Line chart

- Để sử dụng Line chart, dữ liệu thường cần được tổng hợp vào một bảng có hai cột trở lên. Các giá trị trong cột đầu tiên cho biết vị trí của các điểm trên trực hoành đối với mỗi đường được vẽ. Mỗi cột tiếp theo cho biết vị trí thẳng đứng của các điểm trên cùng một biểu đồ đường đơn (khi có nhiều đồ thị đơn trên cùng 1 Line chart).





- Một số công cụ nhất định tạo biểu đồ dạng đường từ một định dạng dữ liệu khác, trong đó cần có ba cột bắt kẽ có bao nhiêu dòng được vẽ. Trong những trường hợp này, các cột chỉ định các giá trị theo chiều ngang, giá trị theo chiều dọc và dòng nào cho mỗi hàng sẽ được chỉ định.

Date	User type	Trips
2019-03-01	Guest	23
2019-03-01	Subscriber	102
2019-03-02	Guest	24
2019-03-03	Subscriber	77
...

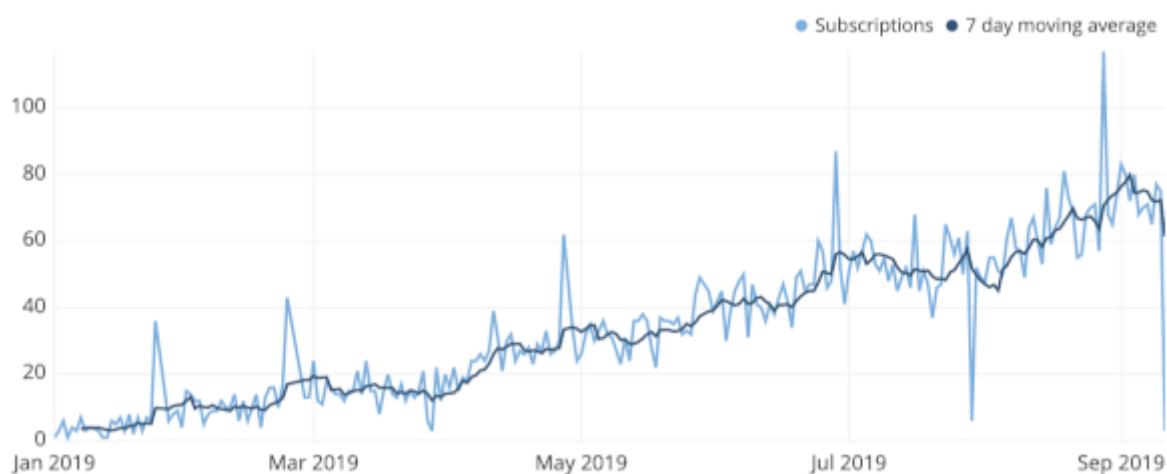
10.4.24.3. Sử dụng hiệu quả line chart

10.4.24.3.1. Chọn màu sắc hoặc kiểu nét vẽ cho đồ thị

Khi cần so sánh các mục khác nhau, có thể cho hiển thị mỗi đường kẻ của đồ thị bằng một màu khác nhau để dễ dàng phân biệt giữa các nhóm dữ liệu khác nhau. Chìa khóa giúp mọi người tham khảo màu sắc hoặc kiểu dáng của từng dòng để hiểu chúng đại diện cho điều gì. Ví dụ: có thể có các đường chấm dài cho một bộ phận và các đường chấm ngắn hơn cho bộ phận khác.

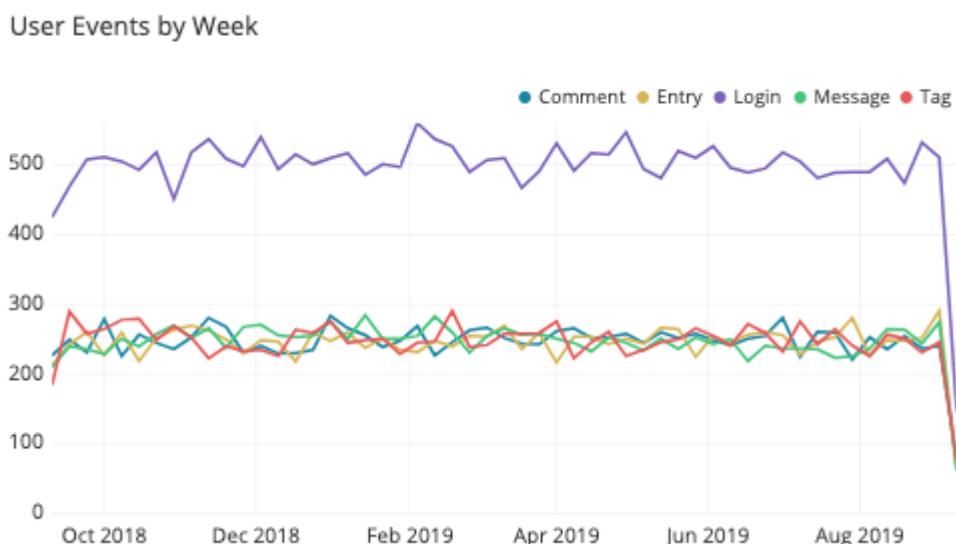
10.4.24.3.2. Chọn khoảng đo thích hợp

- Một khía cạnh quan trọng của việc tạo biểu đồ đường là chọn khoảng hoặc kích thước bin phù hợp. Đối với khoảng đo:
 - Quá rộng: có thể phải mất quá nhiều thời gian để xem xu hướng dữ liệu đang dẫn đầu ở đâu, làm mất đi tín hiệu hữu ích.
 - Quá ngắn chỉ có thể phát hiện ra nhiều hơn là tín hiệu.
- Việc kiểm tra các khoảng thời gian khác nhau hoặc dựa vào kiến thức về lĩnh vực của dữ liệu đang xét có thể cho biết lựa chọn phù hợp về kích thước bin.
- Cũng có thể sử dụng nhiều đường, với một đường cho khoảng thời gian chi tiết và sau đó là đường thứ hai cho xu hướng tổng thể, tính trung bình trên một cửa sổ cuộn.



10.4.24.3.3. Đừng vẽ quá nhiều dòng trên cùng 1 biểu đồ

Sức mạnh lớn đi kèm với trách nhiệm lớn, do đó, mặc dù có khả năng kỹ thuật để đưa nhiều đường vào một biểu đồ đường, nhưng cũng nên thận trọng về lượng dữ liệu mà mình vẽ. Một nguyên tắc nhỏ là hãy giới hạn ở năm dòng hoặc ít hơn, kẻo đồ thị sẽ trở thành một mớ hỗn độn không thể đọc được. Tuy nhiên, nếu các dòng được phân tách rõ ràng, vẫn có thể vẽ tất cả các giá trị muốn theo dõi.

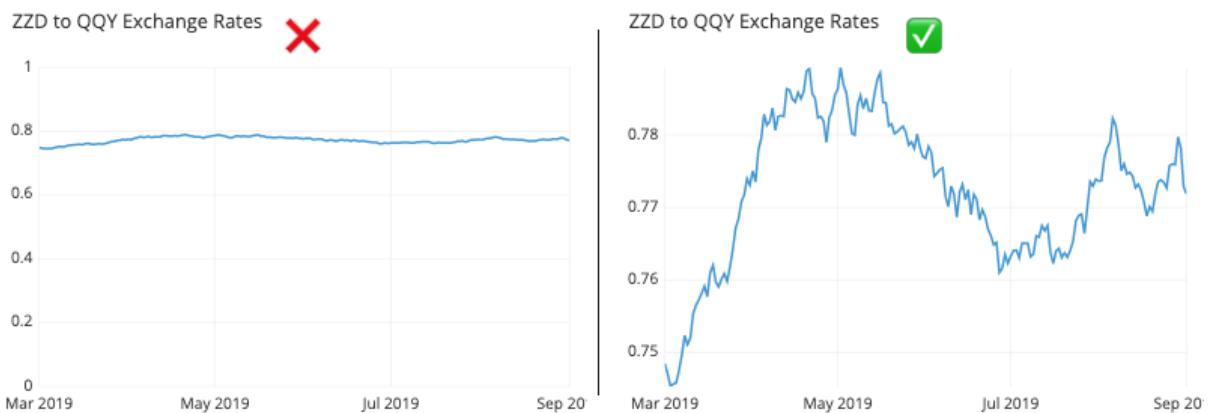


Nếu thấy cần phải vẽ nhiều đường hơn mức có thể đọc được trên một trục, có thể cân nhắc việc ghép các ô thành một lưới các biểu đồ đường nhỏ hơn. Sẽ khó xem chi tiết hơn trong các biểu đồ này, vì vậy, nên sắp xếp chúng theo một số đặc điểm quan trọng (như giá trị trung bình hoặc giá trị cuối cùng) để giúp rút ra những điểm quan trọng.

10.4.24.3.4. Trường hợp sử dụng sai thường gặp (Common misuses): Sử dụng nghiêm ngặt đường cơ sở có giá trị bằng 0

Mặc dù đường cơ sở bằng 0 cho trục tung là yêu cầu bắt buộc đối với Bar graph và Histogram, nhưng không cần bao gồm đường cơ sở bằng 0 cho biểu đồ đường. Hãy nhớ lại rằng mục tiêu chính của biểu đồ đường là nhấn mạnh những thay đổi về giá trị chứ không phải độ lớn của bản thân các giá trị. Trong trường hợp đường số 0 không có ý nghĩa hoặc hữu ích, có

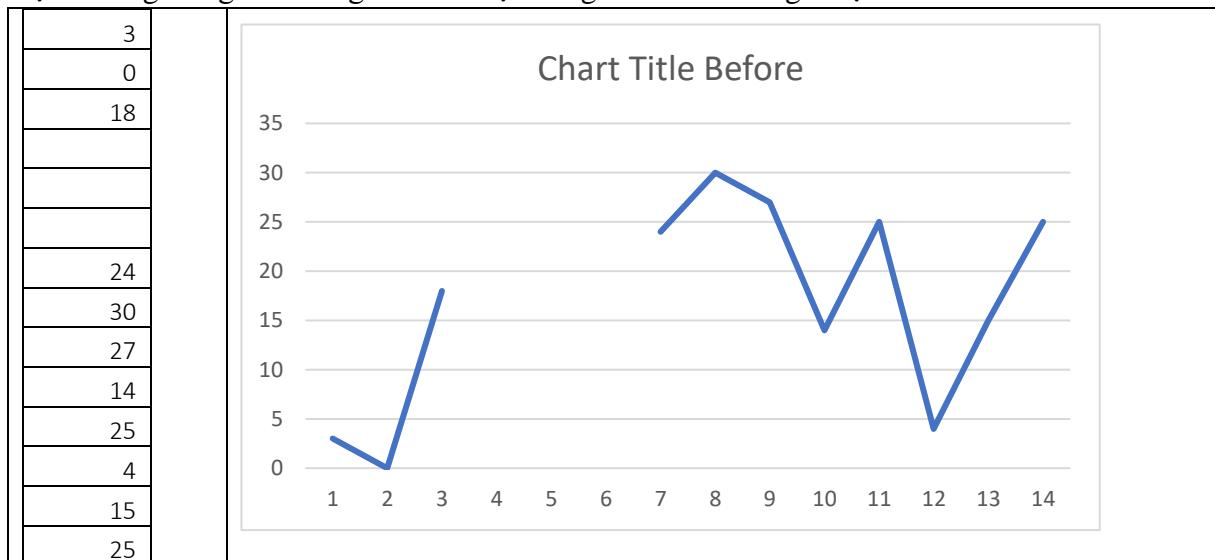
thể phóng to phạm vi trực tung đến mức sẽ tạo ra những thay đổi về giá trị mang tính thông tin nhất.

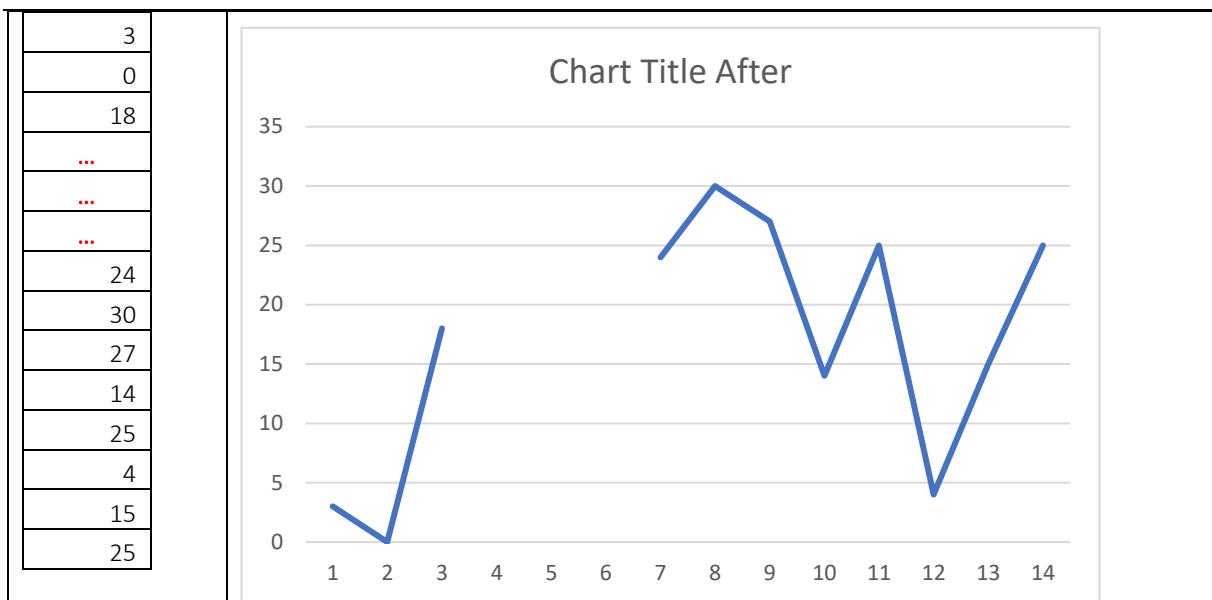


Tuy nhiên, có một trường hợp sử dụng mà đường cơ sở bằng 0 vẫn cần thiết. Khi Line chart được sử dụng để hiển thị phân bố tần suất thì nó được sử dụng tương đương với Bar graph và Histogram. Do đó, nó sẽ tuân theo yêu cầu tương tự là cần bao gồm đường cơ sở có giá trị bằng 0 làm điểm neo cho độ cao của biểu đồ đường.

10.4.24.3.5. Khi chart thiếu thông tin cho 1 số bin

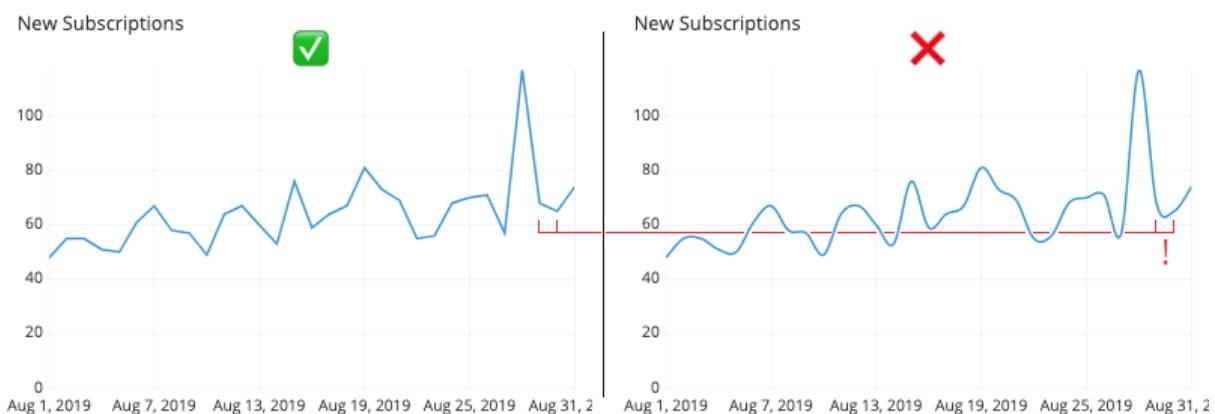
Khi Line chart thiếu thông tin cho một số bin nhất định (các khoảng trống trong bản ghi - records). Khi đó, hãy thử hiển thị tất cả các điểm chứ không chỉ đường. Nếu việc thêm các điểm sẽ làm xáo trộn khả năng diễn giải của biểu đồ, thì một giải pháp thay thế khác là thêm một khoảng trống trên dòng để hiển thị những chỗ còn thiếu giá trị.





10.4.24.3.6. Tự nội suy đường cong giữa các điểm

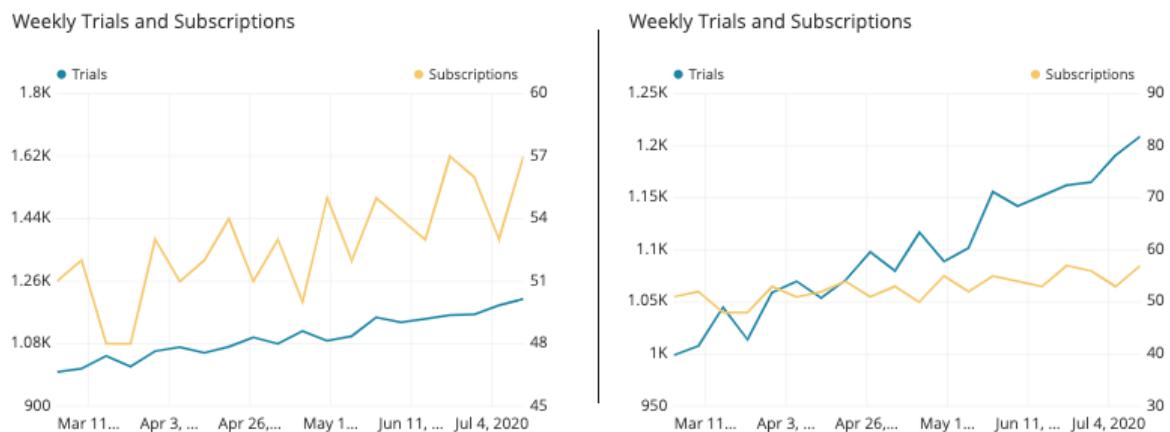
Trong Line chart tiêu chuẩn, mỗi điểm được kết nối với điểm tiếp theo bằng một đoạn thẳng, từ đầu đến cuối. Tuy nhiên, có thể có sự cảm dỗ về mặt thẩm mỹ khi cố gắng liên kết tất cả các điểm một cách tron tru, tạo thành một đường cong đi qua tất cả các điểm cùng một lúc. Nên tuyệt đối không làm như vậy! Trong ví dụ dưới đây, việc cố gắng thực hiện kiểu khớp này chắc chắn sẽ làm sai lệch nhận thức về xu hướng trong dữ liệu. Hướng và độ dốc của đường được cho là biểu thị sự thay đổi về giá trị và do đó, đường cong có thể ám chỉ sự hiện diện của các điểm dữ liệu bổ sung giữa các phép đo thực tế không tồn tại.



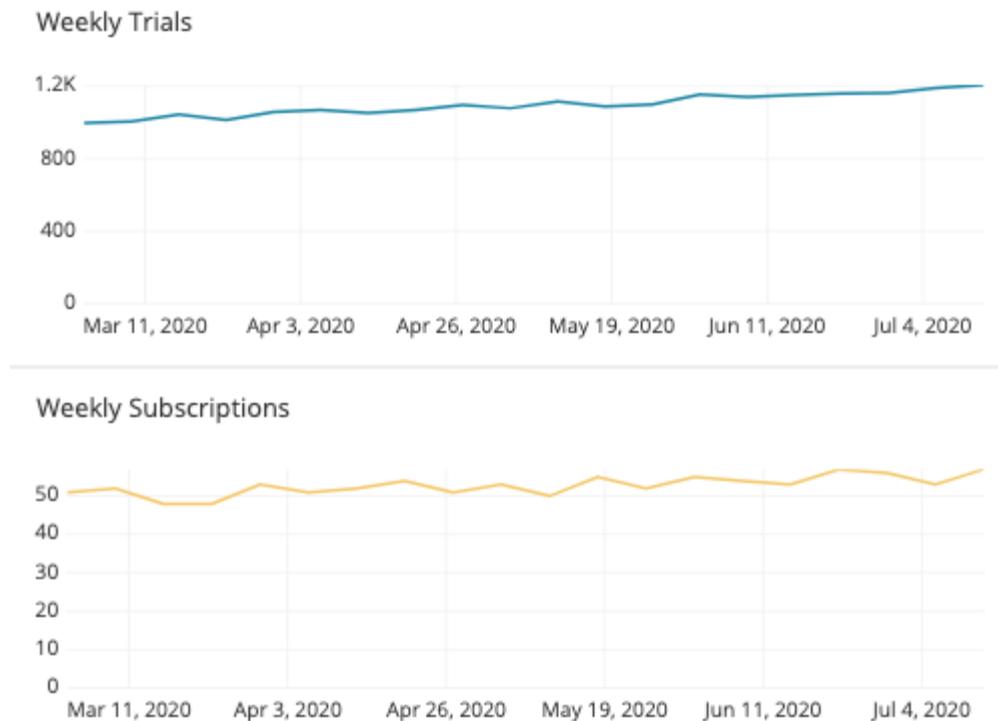
10.4.24.3.7. Sử dụng trực kép gây hiểu lầm

Ví dụ về biểu đồ đường có nhiều đường cho đến nay đều có mỗi đường là một phần của cùng một miền và do đó có thể vẽ được trên cùng một trực. Tuy nhiên, không có gì giới hạn mỗi dòng mô tả các giá trị trên cùng một đơn vị. Khi một biểu đồ đường bao gồm hai chuỗi, mỗi chuỗi mô tả tóm tắt của một biến khác nhau, thì sẽ có một biểu đồ trực kép.

Vấn đề với biểu đồ trực kép là nó có thể dễ dàng gây hiểu nhầm. Tùy thuộc vào cách chia tỷ lệ của từng trực, mối quan hệ nhận thức giữa hai đường có thể thay đổi. Trong hai biểu đồ bên dưới, số lượng bản dùng thử (Trials) và đăng ký (Subscriptions) hàng tuần được biểu thị trong các biểu đồ trực kép. Dữ liệu hoàn toàn giống nhau cho mỗi biến, nhưng do lựa chọn chia tỷ lệ dọc cho từng biến, mối quan hệ được suy ra giữa các biến sẽ thay đổi.



Mặc dù nhiều công cụ trực quan có khả năng tạo biểu đồ trực kép, nhưng các đề xuất phổ biến lại đề xuất chia riêng hai trục nằm trong cùng một miền hay các miền riêng biệt. Thay vào đó, việc ghép hai đường thẳng thành các biểu đồ riêng biệt vẫn cho phép quan sát được các mô hình thay đổi chung của cả hai biến, đồng thời làm giảm so sánh chúng theo những cách sai lầm.



10.4.24.4. Các tùy chọn thường dùng kèm với Line chart

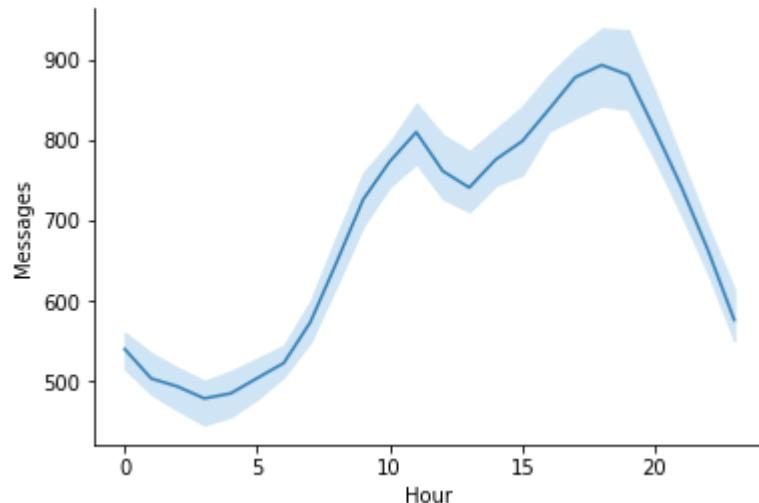
10.4.24.4.1. Sử dụng trend line

Line charts thường được dùng kèm với trend line, là công cụ hữu ích giúp người đọc hiểu các điểm dữ liệu hay nói cụ thể hơn là giúp hiểu xu hướng và mối quan hệ của dữ liệu. Chúng giúp dễ dàng xem xét các xu hướng trong dữ liệu mà không cần giải thích nhiều cho người đọc biểu đồ.

10.4.24.4.2. Bao gồm các dòng bổ sung để thể hiện sự không chắc chắn

Khi có một dòng mô tả tóm tắt thống kê như giá trị trung bình hoặc trung vị, cũng có thể có tùy chọn thêm vào biểu đồ để hiển thị độ không chắc chắn hoặc độ biến thiên của dữ

liệu tại mỗi điểm được vẽ. Một cách để thực hiện điều này là thông qua việc bổ sung các thanh lõi tại mỗi điểm để hiển thị độ lệch chuẩn hoặc một số thước đo độ không đảm bảo khác. Một cách khác là thêm các dòng hỗ trợ ở trên hoặc dưới dòng để hiển thị các giới hạn nhất định trên dữ liệu. Những dòng này có thể được hiển thị dưới dạng bóng để hiển thị các giá trị dữ liệu phổ biến nhất, như trong ví dụ bên dưới.



10.4.24.4.3. Biểu đồ thu nhỏ (Sparkline)

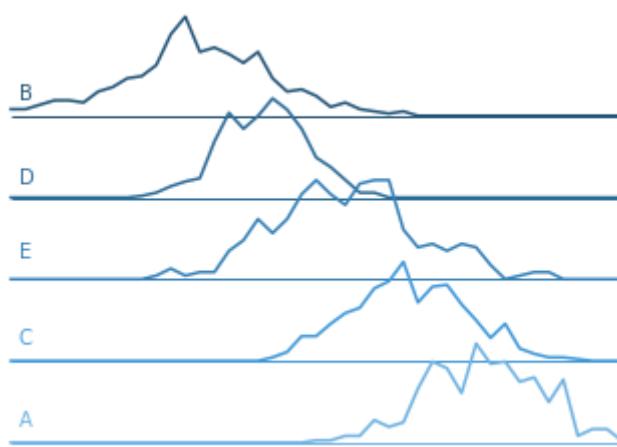
Một trường hợp sử dụng đặc biệt cho Line chart là biểu đồ thu nhỏ. Biểu đồ thu nhỏ về cơ bản là một Line chart nhỏ, được xây dựng để phù hợp với văn bản hoặc đọc theo nhiều giá trị trong bảng. Vì kích thước nhỏ nên nó sẽ không bao gồm bất kỳ nhãn nào. Thông kê có thể được đặt bên cạnh biểu đồ thu nhỏ để hiển thị giá trị bắt đầu và kết thúc hoặc có thể là giá trị tối thiểu hoặc tối đa. Điểm chính của biểu đồ thu nhỏ là thể hiện sự thay đổi trong một khoảng thời gian và thường thấy trong bối cảnh tài chính.

Symbol	Chart	Value	Change
XP		24.54	-1.25
YSK		31.39	+0.54
ZFR		16.78	-0.14

10.4.24.4.4. Ridgeline plot

Một loại biểu đồ biến thể cho Line chart có nhiều đường là Ridgeline. Trong Ridgeline, mỗi đường được vẽ trên một trục khác nhau, hơi lệch nhau theo chiều dọc. Sự chênh lệch nhỏ này có thể tiết kiệm không gian so với việc ghép mặt hoàn chỉnh các ô. Giống như biểu đồ thu nhỏ, việc đánh dấu trực tung thường bị tránh: sẽ khó đọc những giá trị đó trên các trục khác nhau.

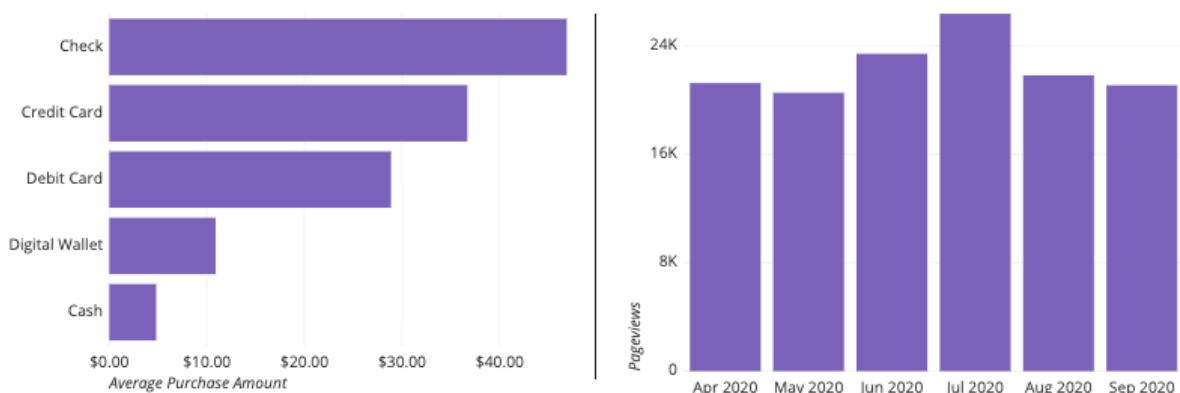
Biểu đồ Ridgeline chủ yếu được sử dụng để so sánh nhiều nhóm trên phân bố tần suất của chúng. Điều này hữu ích nhất khi có thể nhìn thấy một mẫu rõ ràng khi các dòng được sắp xếp theo một cách nào đó.



10.4.24.5. Các đồ thị liên quan

10.4.24.5.1. Bar chart

Nếu biến muốn hiển thị trên trục hoành không phải là số hoặc thứ tự mà thay vào đó là biến phân loại thì cần sử dụng Bar chart thay vì biểu đồ đường. Các thanh trong Bar chart thường được phân tách bằng những khoảng trống nhỏ, giúp nhấn mạnh tính chất riêng biệt của các danh mục được vẽ. Tuy nhiên, xin lưu ý rằng khi trục hoành là số hoặc có thứ tự, khi đó không bị hạn chế sử dụng Bar chart, như trong ví dụ bên dưới.

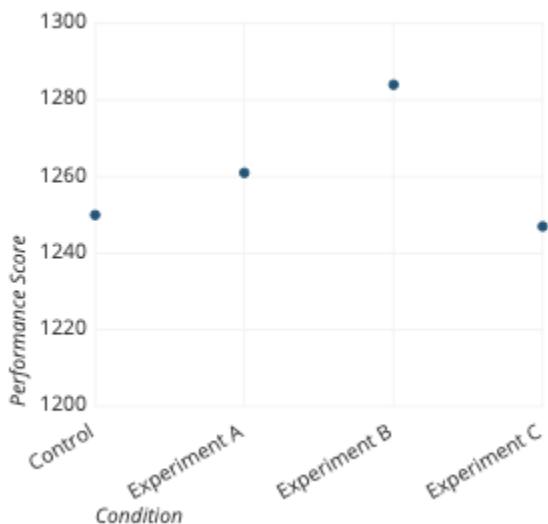


Bên trái: Bar chart trên các nhóm phân loại. Phải: Bar chart trên các nhóm thời gian.

10.4.24.5.2. Dot plot

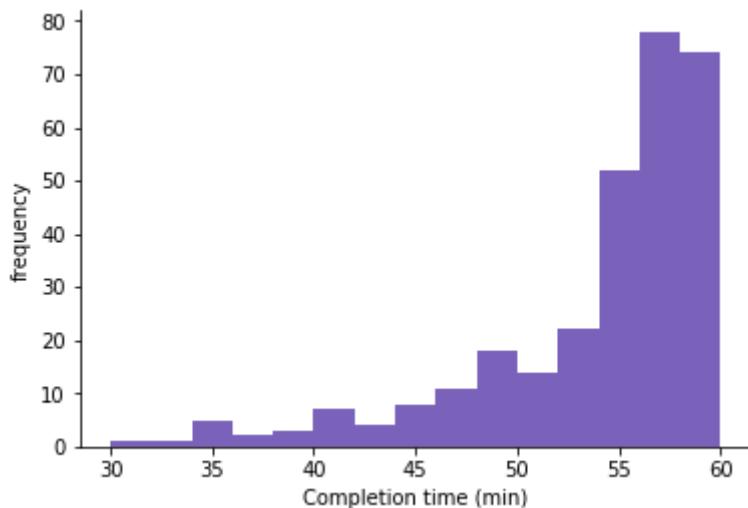
Một loại biểu đồ khác có thể sử dụng khi biến trục hoành được phân loại là Dot plot hoặc Cleveland dot plot. Dot plot giống như Line chart, ngoại trừ việc không có đoạn thẳng nào nối các điểm liên tiếp. Việc thiếu các đoạn đường này sẽ giải phóng các điểm khỏi tiến trình tuần tự của chúng và do đó thứ tự của các nhãn và điểm có thể được điều chỉnh tự do giống như Bar chart. Ưu điểm chính của việc sử dụng Dot plot trên Bar graph là Dot plot, giống như Line chart, không bao gồm đường cơ sở bằng 0.

Nếu có các giá trị trên các mức của một biến phân loại, nhưng các giá trị liên quan không có đường cơ sở bằng 0 có ý nghĩa thì Dot plot có thể là loại biểu đồ phù hợp.



10.4.24.5.3. Histogram

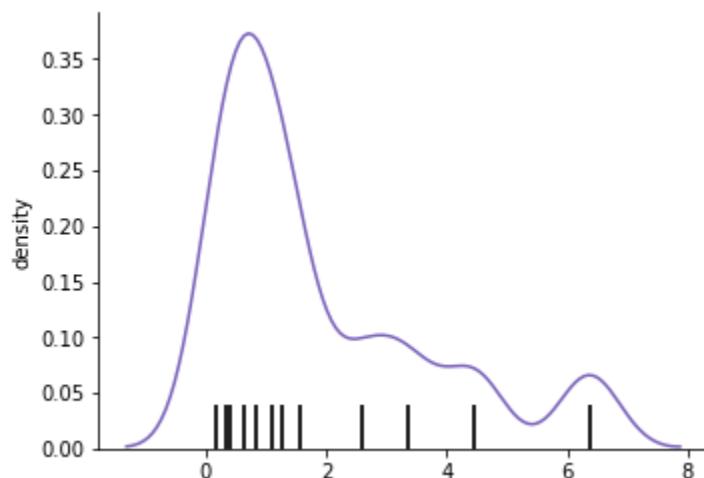
Khi trục tung của Line chart mô tả thông tin về phân bố tần suất, thay vào đó, có tùy chọn hiển thị dữ liệu dưới dạng Histogram. Một trong những lợi ích chính của Histogram là các thanh hiển thị tần suất nhấp nháy trong Line chart, đặc biệt là ở các đỉnh và đáy của một phân bố. Tuy nhiên, Line chart có một lợi thế trong việc hiển thị phân bố tần suất: nếu cần so sánh hai nhóm khác nhau thì điều này rất khó đối với Histogram. Như đã thấy trong phần trước khi sử dụng Line chart, có thể vẽ các đường của hai nhóm trên cùng một trục mà không gặp vấn đề gì.



10.4.24.5.4. Density curve

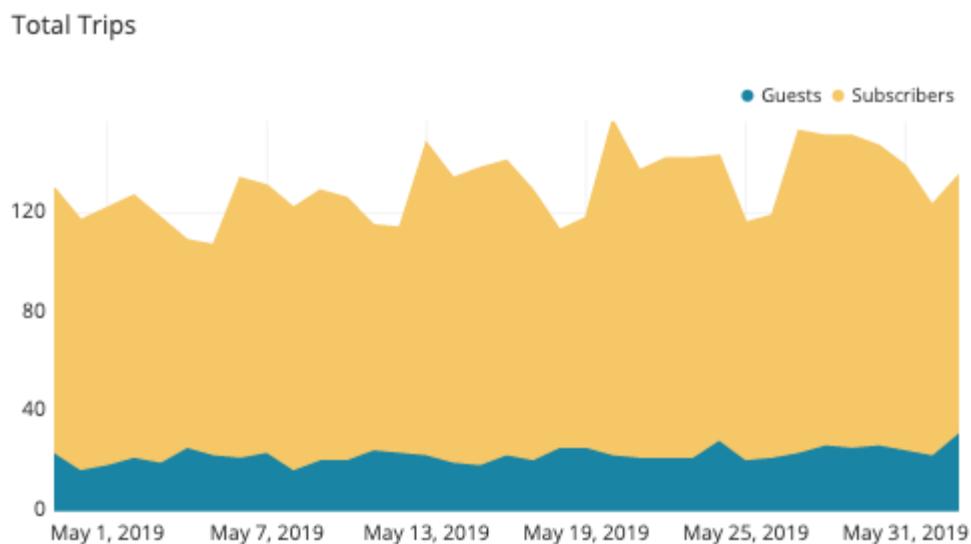
Một lựa chọn khác cho Line chart dựa trên tần suất là Density curve hoặc *kernel density estimate* (ước tính mật độ hạt nhân - KDE). Trong khi Line chart tổng hợp số lần đếm tần suất theo bin thành các điểm riêng lẻ thì KDE tổng hợp mức đóng góp của từng điểm một cách liên tục. Trong KDE, mỗi điểm đóng góp một khối lượng nhỏ xoay quanh giá trị thực của nó (hạt nhân chuẩn – titular kernel); tổng của tất cả các khối lượng cho density curve cuối cùng. Vì có rất nhiều tùy chọn về hình dạng của hạt nhân nên việc ước

tính mật độ hạt nhân thường được dành riêng cho các phương pháp tiếp cận theo chương trình để trực quan hóa dữ liệu.



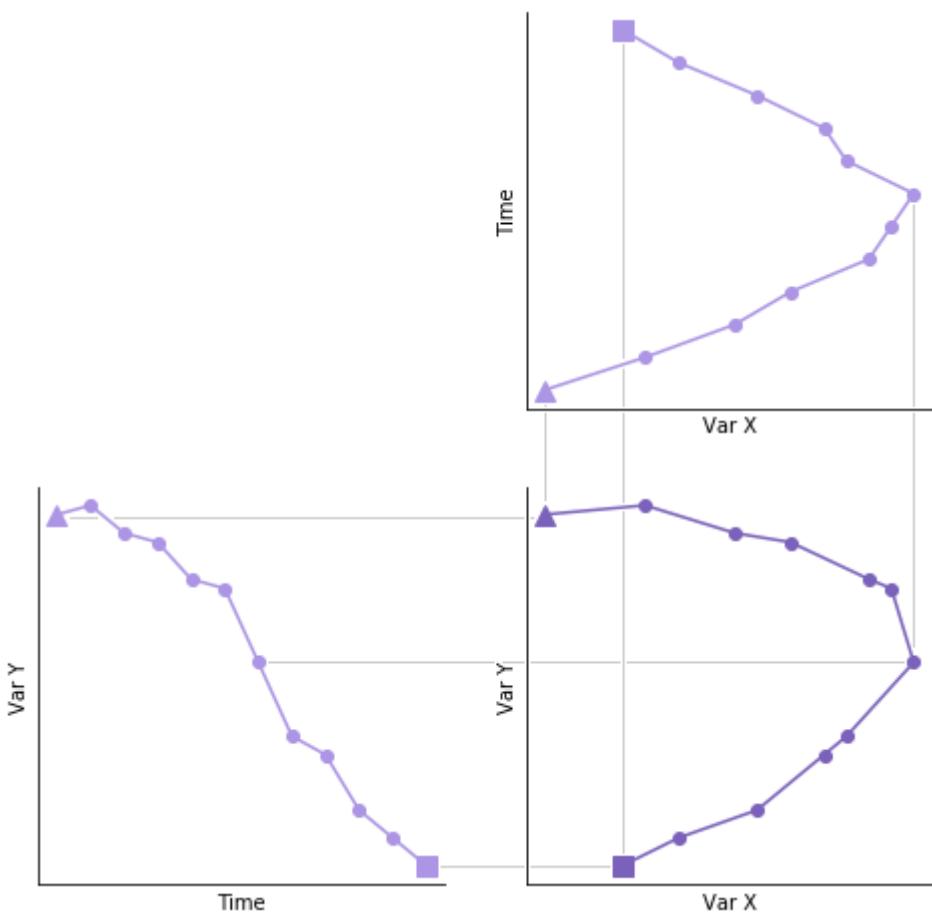
10.4.24.5.5. Area chart

Phần mở rộng của Line chart liên quan đến việc thêm bóng giữa đường và đường cơ sở bằng 0, được gọi là Area chart. Area chart có thể được coi là sự kết hợp giữa Line chart với Bar graph, vì các giá trị có thể được đọc không chỉ từ vị trí thẳng đứng của chúng mà còn từ kích thước của vùng bóng mờ giữa mỗi điểm và đường cơ sở.



10.4.24.5.6. Connected scatter plot

Nếu có hai chuỗi giá trị và muốn vẽ bằng biểu đồ đường, loại biểu đồ thay thế có thể sử dụng là Connected scatter plot (biểu đồ phân tán được kết nối). Trong scatter plot tiêu chuẩn, hai trục biểu thị hai biến quan tâm và các điểm được vẽ trên trục biểu thị giá trị trên các biến đó. Nếu kết nối các điểm theo thứ tự được chỉ định bởi biến thứ ba như thời gian, ta sẽ nhận được một Connected scatter plot. Connected scatter plot rất tốt để xem xét không chỉ mối quan hệ giữa hai biến mà còn xem chúng thay đổi như thế nào theo thời gian hoặc giá trị của biến thứ ba.



Connected scatter plot (phía dưới bên phải) là sự kết hợp của hai biểu đồ đường (phía trên bên phải, phía dưới bên trái). Lưu ý các trục hoán đổi cho biểu đồ phía trên bên phải.

10.4.25. Phân tích xu hướng

- Các chuyên gia trong lĩnh vực kinh doanh, tài chính và kinh tế có thể quan tâm đến việc nhận biết và xác định xu hướng thị trường.
- Có 3 loại xu hướng chính là xu hướng tăng, xu hướng giảm và xu hướng ngang.
- Phân tích xu hướng là một kỹ thuật mà các nhà phân tích tài chính có thể sử dụng để đánh giá hiệu quả hoạt động của một lĩnh vực kinh doanh và đưa ra dự đoán về những thay đổi sắp tới. Phân tích xu hướng có thể giúp hiểu mô hình bán hàng, báo cáo chi phí, dự báo ngân sách và theo dõi chi tiêu. Hiểu kỹ thuật này có thể giúp xác định chiến lược kinh doanh hoặc đầu tư.

10.4.25.1. Phân tích xu hướng là gì (What is trend analysis)?

Phân tích xu hướng là một kỹ thuật sử dụng báo cáo tài chính để nhận biết các mô hình trên thị trường và dự báo hiệu suất trong tương lai. Nó liên quan đến việc thu thập thông tin từ hồ sơ và vẽ dữ liệu trên biểu đồ để xác định các mô hình kinh tế. Các chuyên gia tài chính xác định các đường xu hướng, là những đường kết nối các điểm dữ liệu cho phép các nhà phân tích xác định các mô hình tăng và giảm trên thị trường. Đường xu hướng cho phép các chuyên gia tài chính phân tích dữ liệu trong quá khứ và đưa ra dự đoán về tương lai của thị trường trong một ngành cụ thể. Mục tiêu của kỹ thuật này là đánh giá sự thay đổi trong thị trường từ thời kỳ này sang thời kỳ khác. Điều này có thể giúp các nhà đầu tư đưa ra quyết định kinh doanh thông minh. Các kỹ thuật như phân tích xu hướng hàng tháng có thể cung cấp thông tin về một khoảng

thời gian ngắn hạn, trong khi các chiến lược thay thế như phân tích xu hướng hàng năm có thể giúp các nhà phân tích tài chính xem xét dữ liệu dài hạn. Sự thay đổi xu hướng có thể xảy ra vì những lý do sau:

- Tăng hoặc giảm nguồn cung cổ phiếu so với số lượng cổ phiếu hiện có
- Quy định mới của chính phủ
- Biến động thị trường

10.4.25.2. Lợi ích của phân tích xu hướng

Có nhiều lý do khiến việc phân tích xu hướng có thể quan trọng đối với các chuyên gia về tài chính, kinh tế và kinh doanh. Dưới đây là một số lợi ích cần xem xét:

- **Lợi nhuận tài chính (Financial profit):** Đưa ra những dự đoán có cơ sở về thời điểm đầu tư vào cổ phiếu và tài sản có thể dẫn đến thu được lợi nhuận trong thời kỳ kinh tế tăng trưởng. Thị trường tăng trưởng, được xác định bằng xu hướng tăng, thường xảy ra khi giá cổ phiếu hoặc tài sản tăng ít nhất 20% sau khi thị trường xuống thấp và có thể là thời điểm sinh lời để mọi người đầu tư.
- **Xác định các thành phần kinh doanh thành công (Identifying successful business components):** Việc nêu bật những hoạt động và chiến lược nào đang phục vụ tốt cho doanh nghiệp và những lĩnh vực nào có thể cần cải thiện có thể có ích. Việc thu thập thông tin có giá trị về giá cổ phiếu và tài sản có thể ảnh hưởng đến việc ra quyết định trong tương lai về các chiến lược ngắn hạn và dài hạn.
- **Hiểu biết về động lực thị trường (Understanding market dynamics):** Bằng cách theo dõi sự chuyển động lên xuống của thị trường, có thể làm quen với động lực thị trường và nâng cao hiểu biết của mình về các mô hình tài chính và kinh tế. Kiến thức này có thể giúp doanh nghiệp của phát triển mạnh hoặc cho phép kiếm lợi nhuận từ các khoản đầu tư của mình.

10.4.25.3. Phân loại xu hướng

Có 3 loại xu hướng:

(i). Xu hướng tăng (Uptrend hoặc xu hướng tăng trưởng)

Xu hướng tăng hoặc xu hướng thị trường tăng trưởng cho thấy thị trường tài chính đang đi lên. Điều này có ý nghĩa trong từng môi trường cụ thể như:

- Giá tài sản, cổ phiếu ngày càng tăng.
- Đang là thời kỳ tăng trưởng kinh tế.
- Số lượng việc làm sẵn có ngày càng tăng.
- Nền kinh tế đang chuyển sang thị trường tích cực khi chu kỳ đầu tư bắt đầu.
- ...

Xu hướng tăng có thể xảy ra cùng với những thay đổi tích cực trong mô hình kinh doanh hoặc an ninh của công ty liên quan đến kinh tế vĩ mô. Các nhà phân tích tài chính mô tả xu hướng tăng qua giá trị trên trục Y của điểm đầu thấp hơn giá trị trên trục Y của điểm cuối trong dữ liệu theo thời gian.

(ii). Xu hướng giảm (Downtrend hay xu hướng tăng trưởng giảm)

Khi có xu hướng giảm, tùy từng môi trường cụ thể các nhà phân tích có thể chỉ ra những điều sau:

- Thị trường tài chính đang đi xuống.
- Quy mô nền kinh tế và giá trị cổ phiếu, tài sản có thể giảm.
- Doanh số bán hàng giảm, các công ty có thể đóng cửa hoặc đánh giá lại mô hình kinh doanh của mình.
- Các doanh nghiệp có thể tìm kiếm những cách thức mới để duy trì tính cạnh tranh.

Mặc dù giá có thể tăng và giảm không liên tục, nhưng xu hướng giảm xảy ra khi giá trị trên trục Y của điểm đầu cao hơn giá trị trên trục Y của điểm cuối trong dữ liệu theo thời gian.

(iii). Xu hướng ngang (Horizontal trend)

Xu hướng ngang hoặc xu hướng đi ngang xảy ra khi giá cổ phiếu hoặc tài sản không tăng hoặc giảm đáng kể và tương đối ổn định. Loại xu hướng này có thể dẫn đến những điều sau:

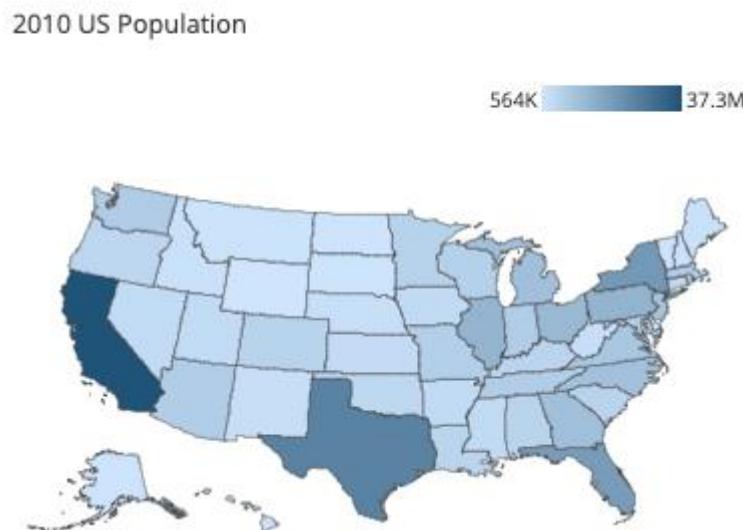
- Các nhà đầu tư có thể gặp khó khăn trong việc xác định hướng đi của xu hướng này và dự đoán liệu đây có phải là thời điểm tốt để khách hàng đầu tư hay không.
- Các chuyên gia tài chính có thể không dự báo được các diễn biến ngắn hạn hoặc dài hạn trên thị trường.
- Chính phủ có thể cần khuyến khích xu hướng đi lên và tăng trưởng của nền kinh tế.

10.4.25.4. Ví dụ về phân tích xu hướng

Dưới đây là một số ví dụ về phân tích xu hướng để giúp hiểu rõ hơn về khái niệm này:

- Mô hình bán hàng (*Sales patterns*): Một nhóm các nhà phân tích tài chính có thể xem xét các mô hình bán hàng để xác định xem chúng đang tăng hay giảm và nguồn gốc của những thay đổi này là gì. Mô hình bán hàng có thể thay đổi do sản phẩm mới, cơ sở khách hàng mới hoặc đặc điểm kỹ thuật của các khu vực bán hàng khác nhau.
- Báo cáo chi phí (*Expense reports*): Kế toán viên có thể kiểm tra báo cáo chi phí để đảm bảo rằng mọi khoản phí đều hợp pháp. Điều này có thể giúp xác nhận rằng dữ liệu đại diện cho các giao dịch hợp lệ trên thị trường, có thể giúp các nhà phân tích tài chính đưa ra dự đoán và ước tính hợp pháp.
- Dự báo ngân sách (*Budget forecasting*): Các nhà phân tích tài chính có thể đưa ra ước tính về các khoản thu và chi để lập ngân sách và dự đoán kết quả sắp tới. Các công ty có thể hưởng lợi từ việc dự báo ngân sách để tối ưu hóa doanh thu và đảm bảo họ còn đủ tiền trong trường hợp có xu hướng giảm trong tương lai.
- Theo dõi chi tiêu (*Expenditure tracking*): Các chuyên gia tài chính có thể xem xét các khoản mục chi phí trong kỳ báo cáo, tức là một tháng, quý hoặc năm mà kế toán viên lập báo cáo tài chính. Điều này có thể giúp các chuyên gia xác định xem có khoản chi tiêu bất thường nào mà họ có thể nghiên cứu thêm hay không.

10.4.26. Map-based plots



Có một số họ đồ thị chuyên biệt được nhóm theo cách sử dụng: đồ thị dựa trên bản đồ hoặc đồ thị không gian địa lý. Khi các giá trị trong tập dữ liệu tương ứng với các vị trí địa lý thực tế, việc vẽ biểu đồ chúng bằng một loại bản đồ nào đó có thể rất có giá trị. Một ví dụ phổ biến của loại bản đồ này là bản đồ hợp xướng giống như bản đồ trên. Điều này áp dụng cách tiếp cận bản đồ nhiệt để mô tả giá trị thông qua việc sử dụng màu sắc, nhưng thay vì các giá trị được vẽ dưới dạng lưỡi, chúng được điền vào các vùng trên bản đồ.

10.4.27. Mosaic Plot

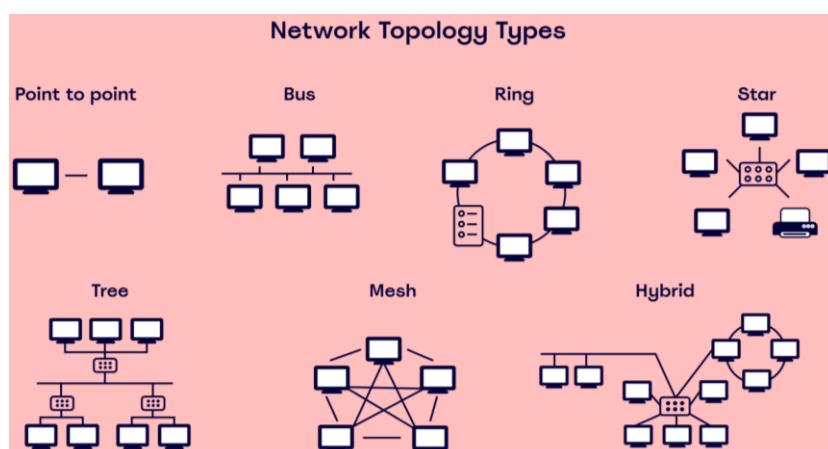
Mosaic Plot (Sơ đồ khám): Còn được gọi là biểu đồ Marimekko, được sử dụng để hiển thị dữ liệu từ hai hoặc nhiều biến phân loại.

10.4.28. Network diagram

10.4.28.1. Giới thiệu

Còn được gọi là Network Graph, Network Map, Node-Link Diagram. Chúng cho thấy mọi thứ được liên kết với nhau thông qua việc sử dụng nodes và các cạnh liên kết với nhau. Nó minh họa các mục khác nhau có mối quan hệ với nhau như thế nào. Thông thường, các nút sẽ được vẽ dưới dạng các chấm nhỏ, vòng tròn hoặc cũng có thể sử dụng các biểu tượng. Các liên kết thường được hiển thị dưới dạng các đường nối giữa các nodes.

10.4.28.2. Minh họa



10.4.29. Parallel Coordinates Plot

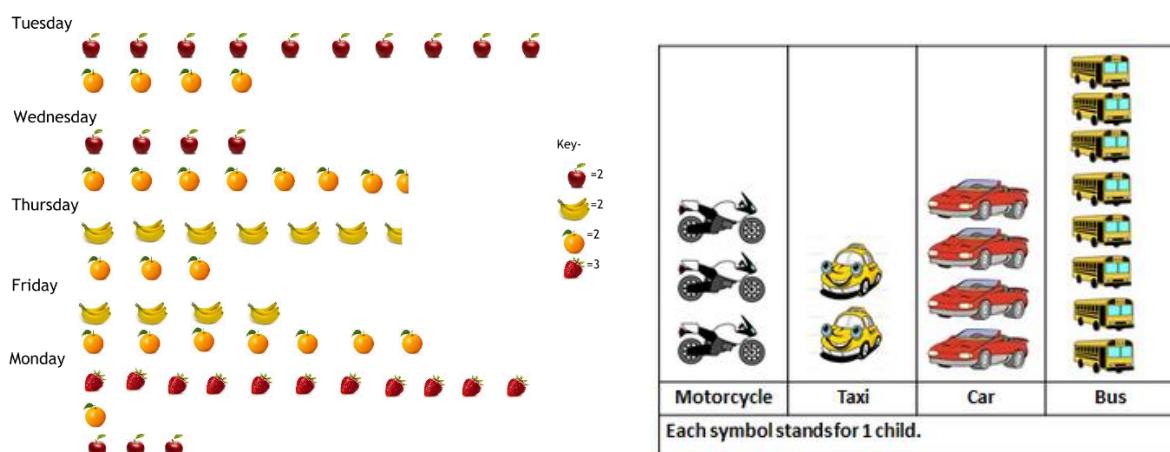
Parallel Coordinates Plot: Được sử dụng để vẽ biểu đồ dữ liệu đa biến bằng cách vẽ mỗi quan sát dưới dạng một đường thẳng trên các trục song song.

10.4.30. Pictograph(chữ tượng hình)

10.4.30.1. Giới thiệu

Pictograph sử dụng hình ảnh hoặc ký hiệu để hiển thị dữ liệu thay vì thanh. Mỗi hình ảnh đại diện cho một số mặt hàng nhất định. Pictograph có thể hữu ích khi muốn hiển thị dữ liệu dưới dạng bản trình bày có tính trực quan cao. Ví dụ: có thể sử dụng hình ảnh một cuốn sách để hiển thị số lượng sách mà một cửa hàng đã bán được trong khoảng thời gian vài tháng.

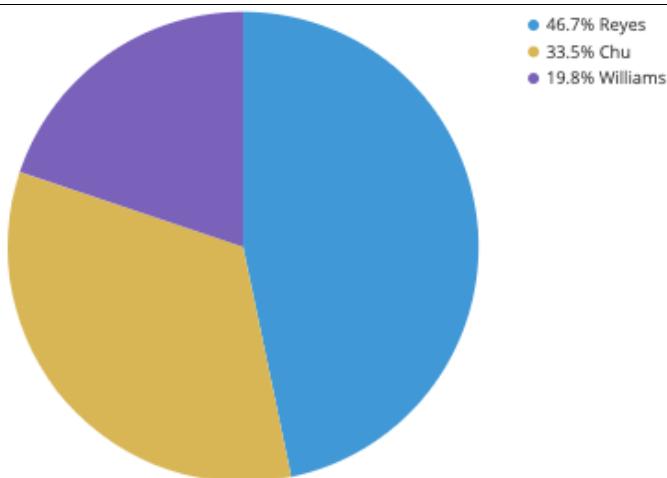
10.4.30.2. Minh họa



10.4.31. Pie chart (đồ thị hình tròn)

10.4.31.1. Giới thiệu

- Là một trong những biểu đồ phổ biến nhất thể hiện tỷ lệ của các thành phần trong toàn bộ dữ liệu. Pie chart thường được sử dụng theo tỷ lệ và tỷ lệ phần trăm giữa các danh mục, bằng cách chia một vòng tròn thành các phân đoạn theo tỷ lệ (giống như một chiếc bánh được cắt thành từng lát). Mỗi chiều dài cung tròn đại diện cho tỷ lệ của từng danh mục, tổng vòng tròn đại diện cho tổng dữ liệu, bằng 100%.
- Pie chart cho thấy tổng số (100%) được phân chia giữa các cấp của một biến phân loại dưới dạng một vòng tròn được chia thành các lát cắt xuyên tâm. Mỗi giá trị phân loại tương ứng với một lát cắt của vòng tròn và kích thước của mỗi lát cắt (cả về diện tích và chiều dài cung) cho biết tỷ lệ của toàn bộ mỗi cấp độ danh mục chiếm bao nhiêu.

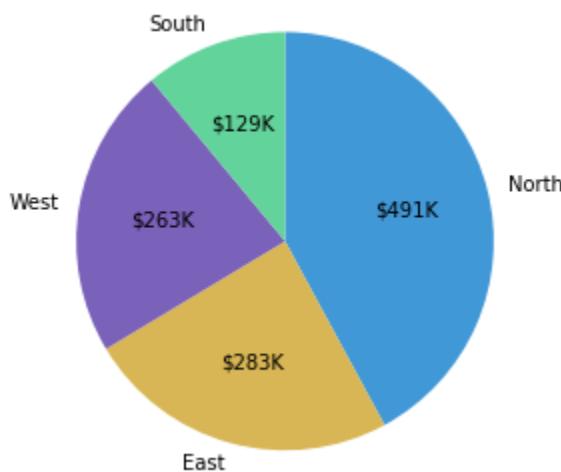


Pie chart ở trên mô tả việc phân phối phiếu bầu cho một cuộc bầu cử ở một thành phố nhỏ. Có thể thấy rằng Reyes, được đại diện bởi lát màu xanh đầu tiên, chỉ có ít hơn một nửa số phiếu bầu. Chu (màu vàng) đứng thứ hai với khoảng 1/3 số phiếu bầu, trong khi Williams (màu tím) đứng cuối cùng với khoảng 1/5 số phiếu bầu. Các chú thích ở phía bên phải cho đánh giá chính xác hơn về tỷ lệ, nhưng Pie chart cho thấy bao quát về vị trí số phiếu giảm.

- Nhược điểm của pie chart là không thể hiển thị nhiều hơn một vài giá trị, vì khi số lượng giá trị được hiển thị tăng lên, kích thước của mỗi phân đoạn trở nên nhỏ hơn. Điều này khiến chúng không phù hợp với lượng lớn dữ liệu.

10.4.31.2. Sử dụng

- Pie chart có trường hợp sử dụng khá hẹp được gói gọn đặc biệt tốt theo định nghĩa của nó. Để sử dụng Pie chart, phải có một số lượng nguyên được chia thành nhiều phần riêng biệt. Mục tiêu chính của Pie chart là so sánh sự đóng góp của từng nhóm với tổng thể, thay vì so sánh các nhóm với nhau. Nếu các điểm trên không được thỏa mãn thì Pie chart không phù hợp và nên sử dụng loại biểu đồ khác để thay thế.
- Các giá trị bao gồm một tổng thể và các phạm trù phân chia tổng thể thường có hai loại chính:
 - (i). Khi 'tổng bộ' đại diện cho tổng số: Ví dụ về điều này bao gồm số phiếu bầu trong một cuộc bầu cử chia theo ứng cử viên hoặc số lượng giao dịch chia theo loại người dùng (ví dụ: khách, người dùng mới, người dùng hiện tại).
 - (ii). Khi tổng là tổng của một biến dữ liệu thực tế: Ví dụ: có thể không quan tâm đến số lượng giao dịch mà quan tâm đến tổng số tiền từ tất cả các giao dịch. Việc chia tổng số này cho một thuộc tính như loại người dùng, độ tuổi hoặc vị trí có thể cung cấp thông tin chi tiết về nơi doanh nghiệp thành công nhất.



- Cấu trúc dữ liệu dùng để vẽ biểu đồ

Region	Total Revenue
North	491 064.51
East	283 445.43
South	128 753.87
West	263 391.13

- Dữ liệu cho Pie chart có thể được tóm tắt trong bảng như trên, trong đó cột đầu tiên biểu thị một danh mục và cột thứ hai biểu thị tỷ lệ, tần suất, giá trị hoặc số lượng của danh mục đó. Thông thường, tổng số không cần phải được chỉ định riêng trừ khi nó được liệt kê ở một nơi khác trên hình được tạo.
- Ngoài ra, một số công cụ chỉ có thể hoạt động với dữ liệu chưa được tổng hợp như trong bảng bên dưới, về cơ bản là thực hiện việc tổng hợp vào bảng trên tại thời điểm tạo Pie chart.

...	region	amount	(other data columns ...)
...	North	34.98	...
...	South	21.05	...
...	South	14.99	...
...	West	14.99	...
...	East	63.25	...

10.4.31.3. Sử dụng hiệu quả Pie chart

- Những điểm chính khi sử dụng pie:
 - So sánh 1 biến cho dữ liệu định tính
 - Cần thể hiện sự đóng góp của từng phần tử trong tổng thể.
- Các chuyên gia có thể sử dụng Pie chart trong các bài thuyết trình kinh doanh để thể hiện các phân khúc dân số, nghiên cứu thị trường, phân bổ ngân sách, ...

10.4.31.3.1. Bao gồm chú thích

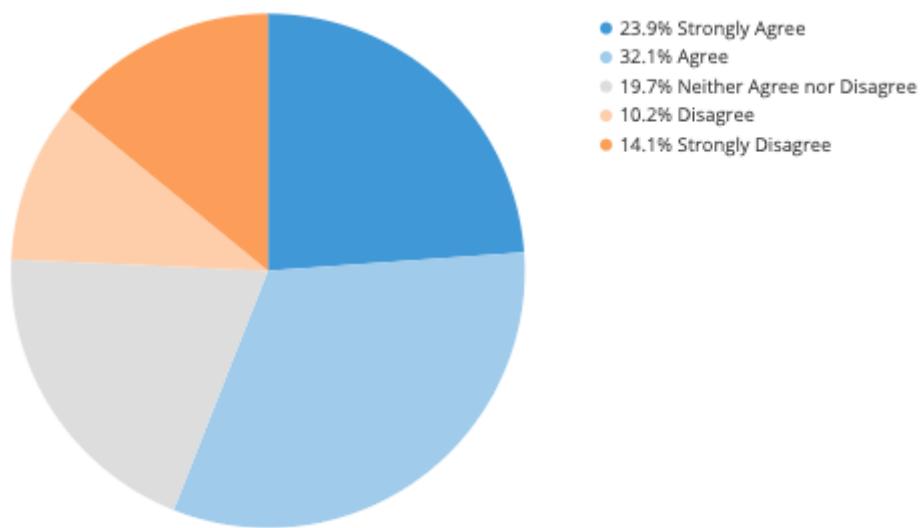
Thực sự rất khó để phân biệt tỷ lệ chính xác từ Pie chart, ngoài các phân số nhỏ như 1/2 (50%), 1/3 (33%) và 1/4 (25%). Hơn nữa, nếu các giá trị lát cắt nhằm mô tả số lượng chứ không phải tỷ lệ, thì Pie chart thường thiếu dấu kiểm để cho phép ước tính giá trị trực tiếp từ kích

thước lát cắt. Chính vì những lý do này mà chú thích là một phần tiêu chuẩn được đưa vào Pie chart.

10.4.31.3.2. Xem xét thứ tự của các lát

Một thứ tự tốt cho các lát cắt có thể giúp người đọc hiểu được nội dung đồ thị dễ dàng hơn nhiều. Thứ tự điển hình đi từ lát lớn nhất đến lát nhỏ nhất, rất hữu ích khi có các danh mục có giá trị giống nhau. Tuy nhiên, nếu các cấp độ danh mục có một thứ tự cố hữu thì việc vẽ các lát cắt theo thứ tự đó thường tốt hơn.

Đối với việc chọn điểm bắt đầu, bạn nên vẽ các lát cắt theo hướng định hướng theo bản số. Các công cụ trực quan hóa thường sẽ bắt đầu từ bên phải hoặc từ trên cùng. Mặc dù bắt đầu từ bên phải có cơ sở toán học liên quan đến các quy ước đo góc, nhưng bắt đầu từ trên cùng mang lại cảm giác trực quan hơn vì nó phù hợp với cách đọc từ trên xuống dưới và cách nghĩ về tiến trình của thời gian trên đồng hồ hoặc mặt đồng hồ.

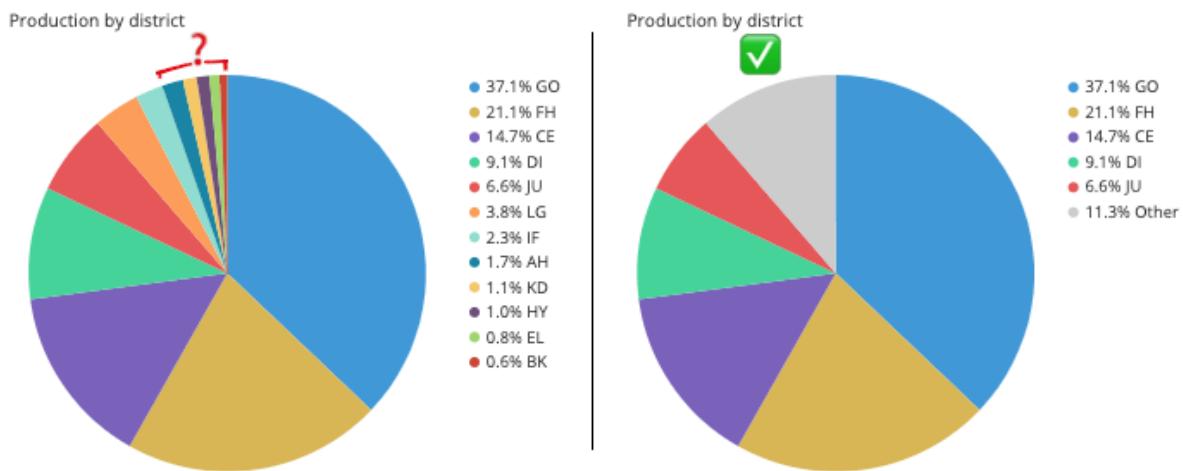


Đồ thị trên không sắp xếp theo kích thước vì các nhãn ở đây có ý nghĩa.

10.4.31.3.3. Limit the number of pie slices

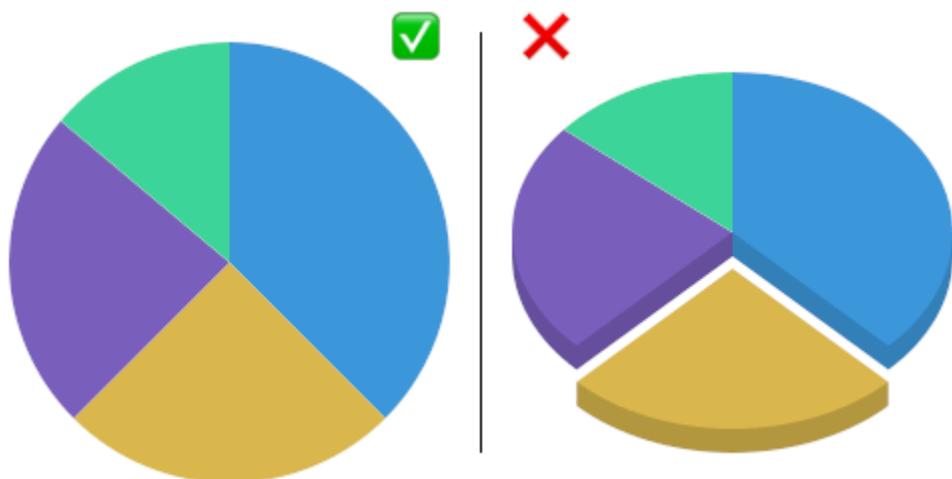
Pie chart có nhiều lát cắt có thể khó đọc. Có thể khó nhìn thấy những lát cắt nhỏ nhất và khó có thể chọn đủ màu sắc để làm cho tất cả các lát cắt trở nên khác biệt. Các đề xuất khác nhau nhưng nếu bạn có nhiều hơn khoảng năm danh mục, bạn có thể chọn 1 trong 2 cách sau:

- Sử dụng loại biểu đồ khác để thay thế cho Pie chart.
- Có thể cân nhắc việc gộp các lát nhỏ thành một lát 'khác', có màu xám trung tính.



10.4.31.3.4. Tránh các hiệu ứng bóp méo

- Việc đọc Pie chart một cách chính xác yêu cầu các diện tích, độ dài cung và góc của các lát cắt đều hướng đến sự thể hiện chính xác của dữ liệu. Mặc dù việc tránh các hiệu ứng 3-d là một ý tưởng hay cho bất kỳ biểu đồ nào nhưng điều này đặc biệt quan trọng đối với Pie chart. Việc bóp hoặc kéo dài hình tròn hoặc thêm độ sâu không cần thiết có thể dễ dàng làm sai lệch độ lớn của mỗi lát so với tổng thể.
- Một biến dạng khác có thể đến từ Pie chart là ‘exploded’ (bùng nổ), trong đó các lát cắt được kéo ra từ giữa để nhấn mạnh. Sự nhấn mạnh này đi kèm với một cái giá phải trả, trong đó những khoảng trống có thể khiến việc đánh giá sự so sánh từng phần với toàn bộ trở nên khó khăn hơn.

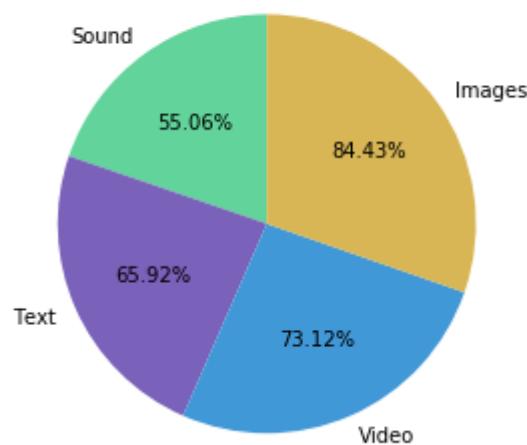


10.4.31.4. Các vấn đề thường gặp khi sử dụng Pie chart

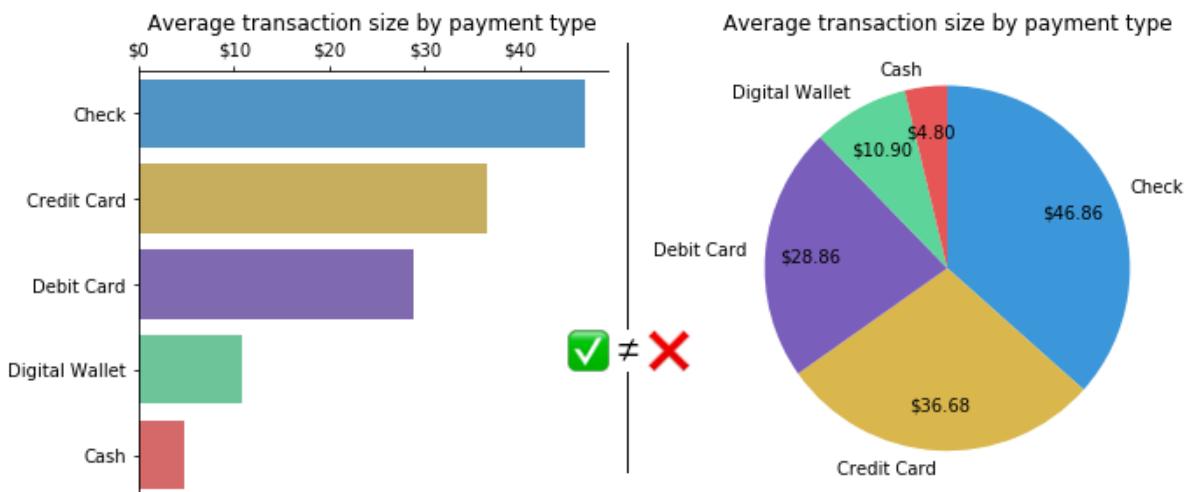
10.4.31.4.1. Lắp đầy Pie chart dựa trên dữ liệu không tương thích

- Một trong những sai lầm phổ biến nhất khi sử dụng Pie chart là làm cho Pie chart khớp với dữ liệu không thể hiện sự so sánh từng phần với toàn bộ. Sự nhầm lẫn này xảy ra thường xuyên nhất khi các giá trị được vẽ là tỷ lệ phần trăm hoặc tỷ lệ, nhưng không bao gồm một tổng thể hoàn chỉnh. Ví dụ bên dưới cho thấy tần suất những người được khảo sát sử dụng từng ứng dụng trong số bốn ứng dụng, nhưng vì nhiều người sử dụng nhiều ứng dụng nên tổng tỷ lệ lên tới hơn 100%.

X What apps do users use each week?

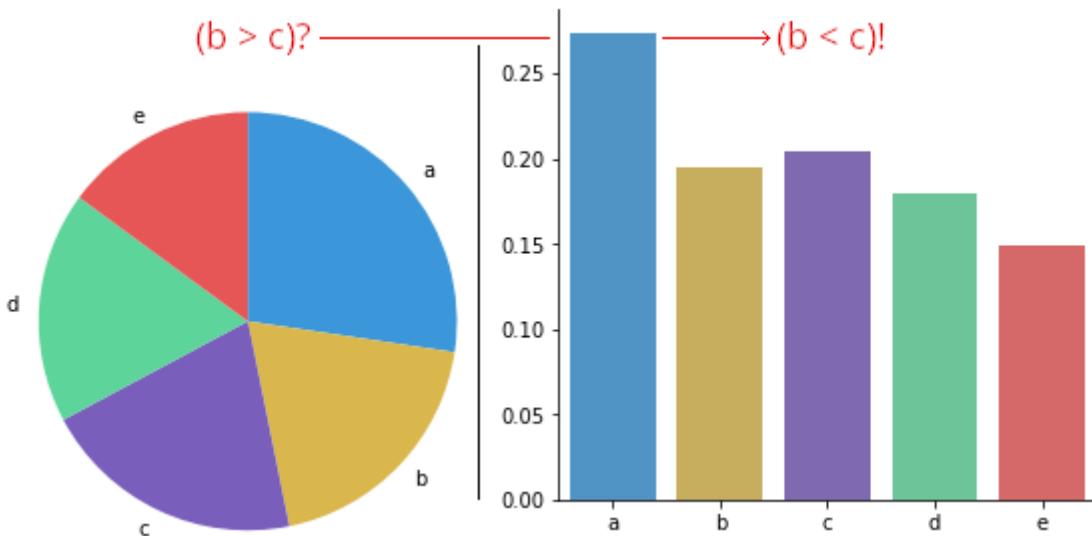


- Một trường hợp phức tạp khác là nếu các giá trị được sử dụng cho mỗi nhóm là một số liệu thống kê tóm tắt không phải là tổng số. Biểu đồ bên phải được xây dựng dựa trên số tiền giao dịch trung bình cho nhiều loại giao dịch (với tổng là \$128.1). Tuy nhiên, vì nó bỏ qua tần suất sử dụng từng loại giao dịch nên nó làm sai lệch doanh thu đến từ mỗi loại. Mặc dù séc có mức trung bình cao nhất nhưng chúng cũng có thể khá hiếm khi được sử dụng. Trong cả hai trường hợp, Bar chart là loại biểu đồ thích hợp để sử dụng.



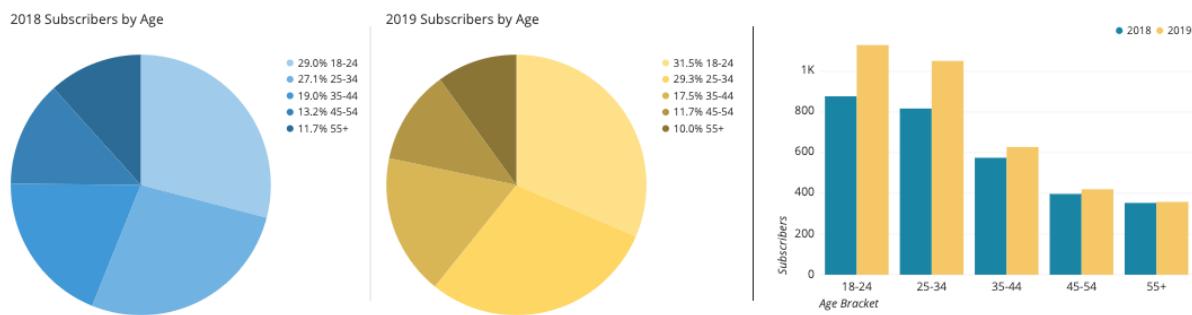
10.4.31.4.2. Sử dụng Pie chart để so sánh các nhóm với nhau

Nếu muốn so sánh giữa các nhóm thay vì so sánh từng nhóm với tổng thể thì tốt hơn hết nên sử dụng một loại biểu đồ khác. Ngay cả khi sắp xếp các lát cắt theo kích thước, cũng khó có thể biết được hai lát cắt khác nhau như thế nào, đặc biệt là khi chúng di chuyển ra xa điểm bắt đầu/kết thúc. Trong ví dụ dưới đây, có thể giả định rằng lát thứ hai lớn hơn lát thứ ba do thứ tự, nhưng Bar chart tương ứng thực sự cho thấy điều ngược lại. Điều chính có thể nói từ Pie chart là cả hai lát đều có tỷ lệ gần như giống nhau trên tổng thể.



10.4.31.4.3. So sánh các giá trị trên nhiều Pie chart

Có thể có trường hợp muốn so sánh nhiều bánh với nhau: ví dụ: so sánh phân bổ nhân khẩu học của người dùng trong nhiều năm. Tuy nhiên, điều này gặp phải một vấn đề tương tự như phần trước, nơi muốn so sánh các nhóm với nhau. Tệ hơn nữa, đó là sự so sánh giữa các chiếc bánh, vì vậy không thể dễ dàng dựa vào thứ tự các lát bánh để so sánh. Thể hiện dữ liệu bằng cách sử dụng một biểu đồ khác, như Stacked bar chart, Grouped bar chart hoặc Line chart, thường là lựa chọn tốt hơn khi muốn so sánh giữa các nhóm này. Giống như những chiếc bánh nướng thực tế, tốt nhất nên thực hiện từng Pie chart một lần.

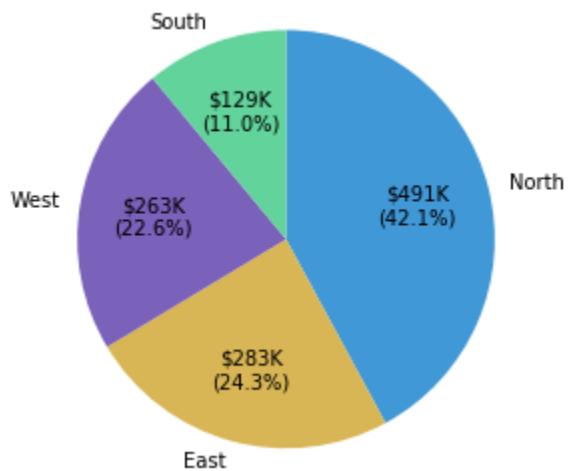


So sánh các Pie chart có thể hàm ý sự thu hẹp của các nhóm tuổi lớn hơn theo tỷ lệ, nhưng Grouped bar chart lại cho thấy sự tăng trưởng ở các nhóm trẻ hơn.

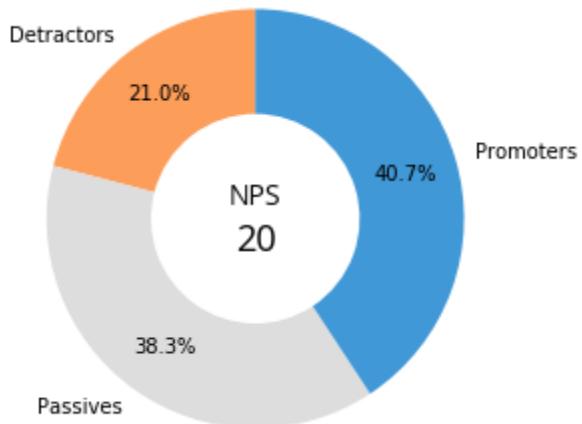
10.4.32. Các tùy chọn thường dùng kèm với Pie chart

10.4.32.1.1. Tần suất tuyệt đối so với tần suất tương đối

Pie chart có thể được dán nhãn theo giá trị tuyệt đối hoặc theo tỷ lệ. Việc dán nhãn các lát cắt với số lượng tuyệt đối và ngữ ý tỷ lệ với các kích thước lát cắt là thông thường, nhưng hãy xem xét cẩn thận các mục tiêu trực quan hóa để quyết định kiểu chú thích tốt nhất để sử dụng cho biểu đồ. Trong một số trường hợp, việc bao gồm cả hai số trong chú thích có thể có giá trị cho văn bản bổ sung.



10.4.32.1.2. Doughnut plot

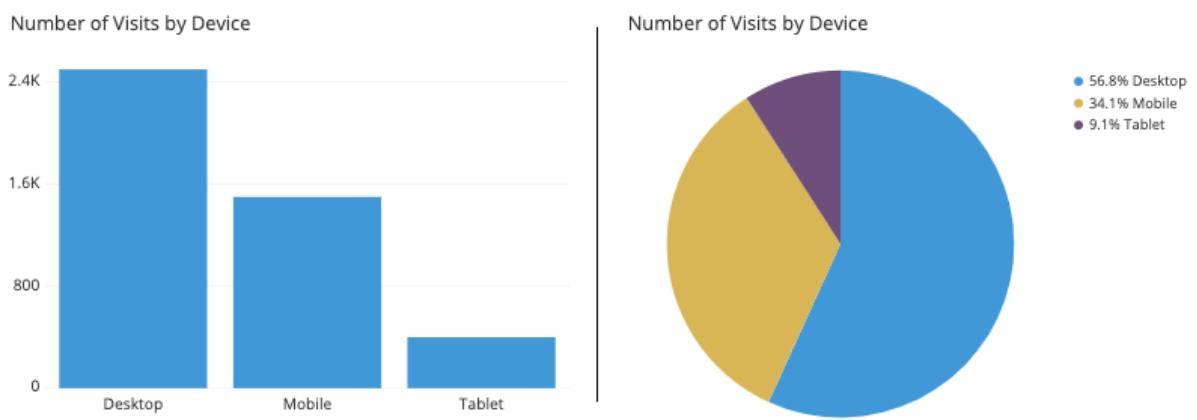


Doughnut (hay còn gọi là Donut plot) chỉ đơn giản là một Pie chart đã loại bỏ vòng tròn ở giữa. Trong hầu hết các trường hợp, không có sự khác biệt đáng kể về khả năng đọc giữa Pie chart và Doughnut, vì vậy việc lựa chọn Doughnut thay vì Pie chart tiêu chuẩn chủ yếu mang tính thẩm mỹ. Một lợi ích nhỏ cho hình dạng của Doughnut là khu vực trung tâm có thể được sử dụng để cung cấp thêm thông tin hoặc báo cáo số liệu thống kê.

10.4.32.1.3. Các đồ thị liên quan

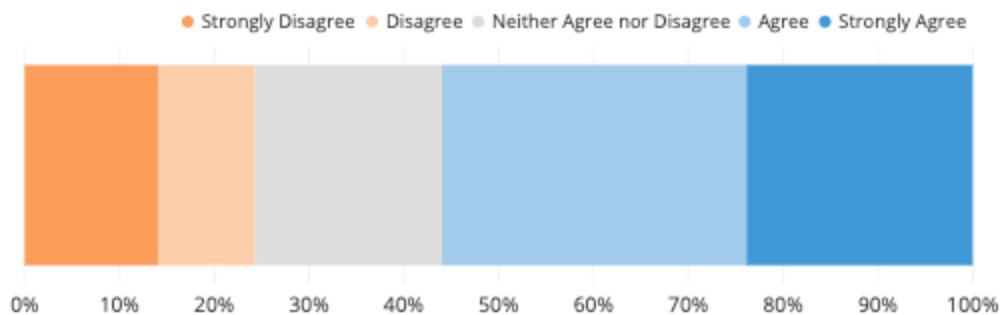
10.4.32.1.3.1 Bar chart

Sự cạnh tranh lớn nhất đối với Pie chart đến từ Bar chart. Trong hầu hết các trường hợp, sẽ không muốn sử dụng Pie chart – thay vào đó, Bar chart sẽ trình bày các điểm một cách gọn gàng và rõ ràng hơn. Nhiều vấn đề với Pie chart được giải quyết thông qua việc sử dụng Bar chart. Tuy nhiên, Bar chart không ngay lập tức thể hiện sự so sánh từng phần với toàn bộ, đây là lợi ích chính của Pie chart.



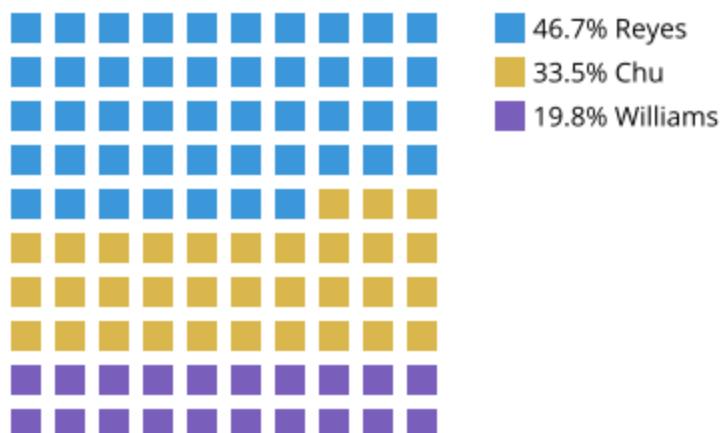
10.4.32.1.3.2 Stacked bar chart

Mặt khác, Stacked bar chart có thể chứng tỏ là đối thủ mạnh của Pie chart về khả năng truyền đạt sự so sánh từng phần với toàn bộ. Một thanh xếp chồng đơn có thể được coi là các lát của Pie chart được cuộn thành dạng hình chữ nhật. Dạng hình chữ nhật cũng giúp dễ dàng so sánh các phân tích phân loại giữa các nhóm khác nhau. Tuy nhiên, Pie chart vẫn có lợi thế về tính quen thuộc và tính thẩm mỹ, vì vậy, chúng vẫn cần được lưu ý trong trường hợp sử dụng so sánh từng phần với toàn bộ.



10.4.32.1.3.3 Waffle chart

Một lựa chọn thay thế khác cho Pie chart là Waffle chart, còn được gọi là Square chart hoặc Square pie. Doughnut bao gồm 100 biểu tượng, thường là các hình vuông được bố trí trong lưới 10 x 10. Mỗi biểu tượng đại diện cho 1% dữ liệu và các biểu tượng được tô màu dựa trên sự phân bổ dữ liệu theo phân loại. Mặc dù sẽ cần phải làm tròn số lượng danh mục để phù hợp với cấu trúc của biểu đồ – không bao giờ chia các biểu tượng trong biểu đồ này – đó có thể là một cách làm cho tỷ lệ tương đối của mỗi danh mục dễ đọc hơn.

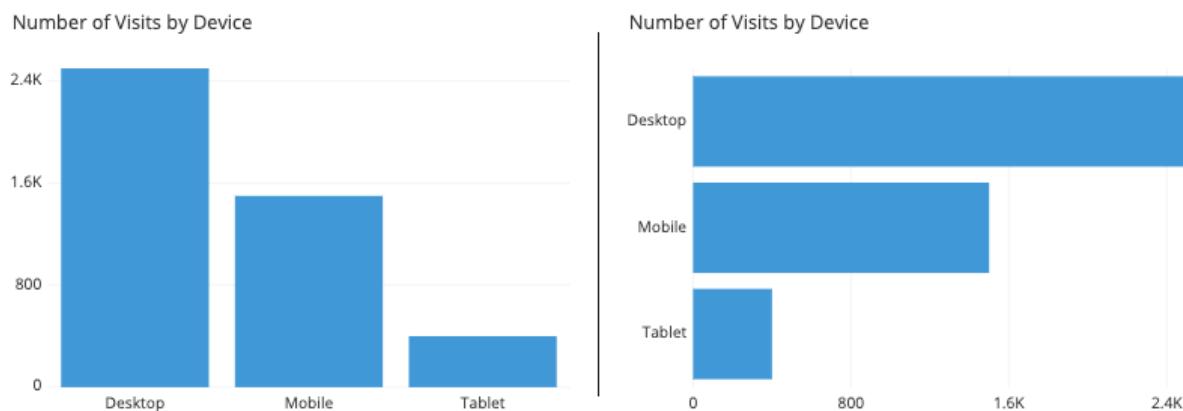


10.4.32.2. Chọn lựa giữa 2 biểu đồ Bar chart và Pie chart

10.4.32.2.1. So sánh cách biểu diễn của 2 biểu đồ trên cùng 1 dữ liệu

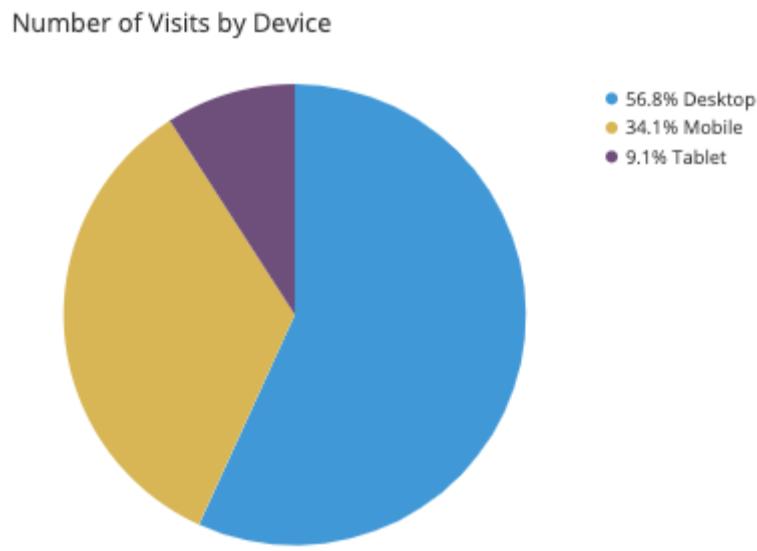
- Bar chart

Các Bar chart ví dụ bên dưới cho thấy cách phân chia cơ sở người dùng của ứng dụng giữa các loại thiết bị khác nhau. Lưu ý rằng Bar chart có thể được định hướng theo hai cách, với các thanh được định hướng theo chiều dọc hoặc chiều ngang.



- Pie chart

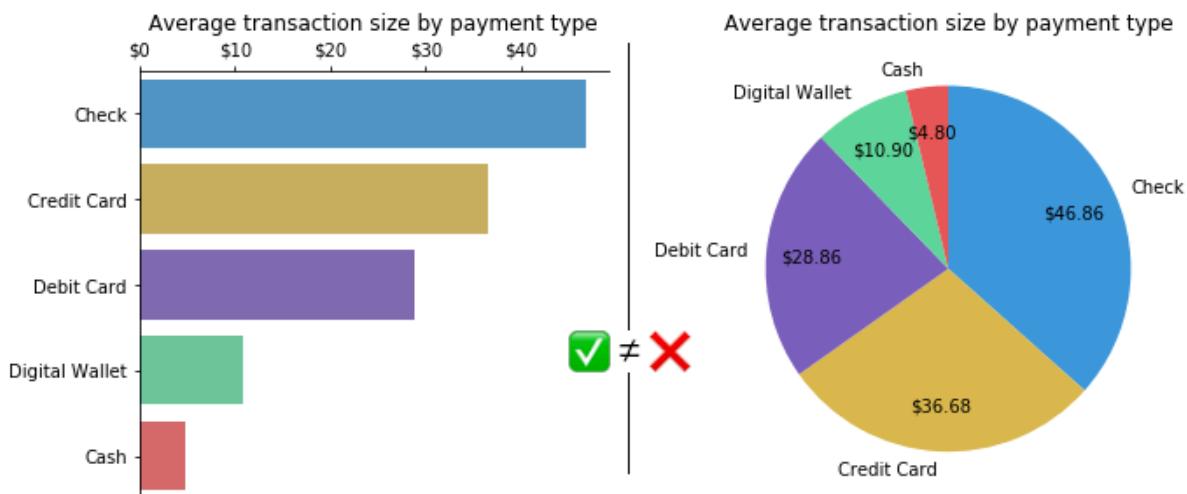
Pie chart cho thấy tổng số được phân chia như thế nào giữa các danh mục riêng biệt dưới dạng một vòng tròn được chia thành các lát xuyên tâm. Mỗi danh mục được liên kết với một lát duy nhất có kích thước tương ứng với tỷ lệ của danh mục trong tổng số. Hình bên dưới vẽ cùng một dữ liệu như trên nhưng thay vào đó sử dụng Pie chart.



10.4.32.2.2. Nhận xét

Mặc dù ví dụ trên cho thấy cách cùng một dữ liệu có thể được biểu thị theo nhiều cách, nhưng đừng phạm sai lầm khi nghĩ rằng chúng luôn có thể thay thế cho nhau.

- Với Bar chart, có thể tự chọn bất kỳ giá trị nào trên trục giá trị số mà mình muốn. Đối với các giá trị số biểu thị tổng số liệu hoặc số điểm dữ liệu, tổng giữa các nhóm sẽ có xu hướng bằng tổng trên toàn bộ dữ liệu.
 - Pie chart cũng có giá trị như Bar chart như một lựa chọn trực quan. Tuy nhiên, nếu các giá trị số biểu thị một số thống kê khác trong đó tổng giữa các nhóm không bằng các nhóm bỏ qua thống kê thì đó là lúc gặp vấn đề.
- ⇒ Bar chart phù hợp trong trường hợp này, nhưng Pie chart lại không phù hợp. Vì hình tròn ngụ ý rằng các lát cắt là một phần của một tổng thể nên người đọc rất dễ nhầm lẫn tổng các lát cắt là đại diện của một loại tổng thể nào đó.



Hình bên phải cho thấy việc cộng quy mô giao dịch trung bình sẽ không bằng mức trung bình của tổng số.

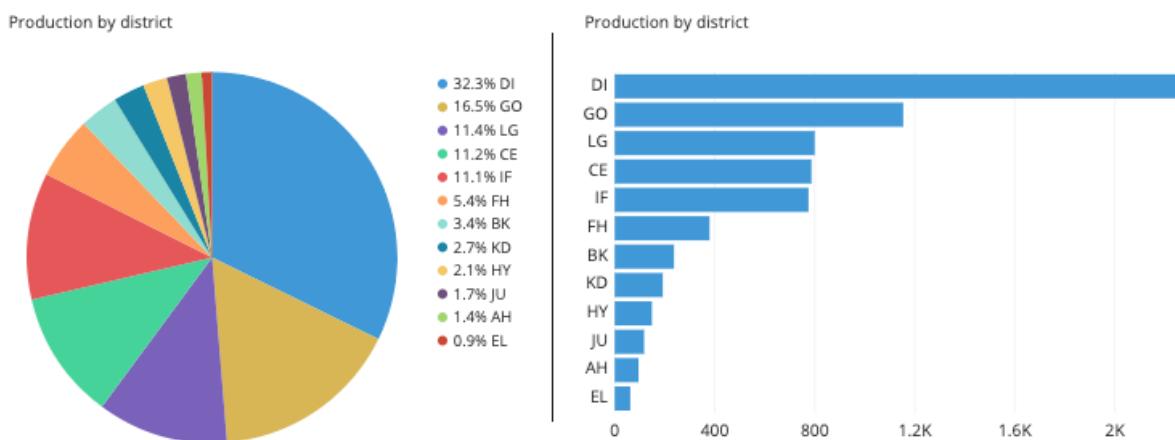
Tóm lại,

- Pie chart chỉ có thể được sử dụng nếu tổng của các phần riêng lẻ cộng lại thành một tổng thể có ý nghĩa và được xây dựng để trực quan hóa cách mỗi phần đóng góp vào tổng thể đó.
- Bar chart có thể được sử dụng cho nhiều loại dữ liệu hơn, không chỉ để chia nhỏ tổng thể thành các thành phần.

10.4.32.2.3. Nhược điểm của Pie chart

Ngay cả với dữ liệu có thể sử dụng Pie chart, Pie chart có thể không phải là lựa chọn trực quan tốt. Có nhiều trường hợp có thể gây ra vấn đề cho Pie chart.

- **Quan tâm đến sự đóng góp chính xác của mỗi nhóm:** Bỏ qua sự hiện diện của các chú thích bổ sung, khó có thể biết được mỗi lát chiếm tỷ lệ bao nhiêu trong toàn bộ. Mặc dù có thể dễ dàng đưa ra phán đoán khi một lát chiếm bội số của 1/3 hoặc 1/4, nhưng việc xác định một giá trị nhỏ hơn hoặc một giá trị ở giữa sẽ khó hơn nhiều. Điều này tốt nếu muốn loại bỏ những đánh giá như “hơn một nửa” hoặc “khoảng một phần ba”, nhưng đối với những thông điệp được điều chỉnh tốt hơn, hình ảnh trực quan không tự đứng vững được.
- **Nhiều lát cắt có giá trị tương tự nhau:** Vì Pie chart thường không có dấu xung quanh trung tâm nên khó có thể so sánh các nhóm có kích thước tương tự nhau. Mặc dù việc sắp xếp các lát cắt là một quy ước tốt nhưng đây không phải là một bước đảm bảo trong việc tạo Pie chart. Nếu không có chú thích, có thể nói rằng hai nhóm có quy mô tương tự nhau nhưng không biết nhóm nào lớn hơn.
- **Quá nhiều lát:** Nếu có quá nhiều lát thì có thể sẽ gặp phải vấn đề là có những lát có kích thước tương tự nhau (xem phần trên) hoặc những lát quá nhỏ. Những lát cắt nhỏ đó có thể khó đọc và khó tô màu.



Những hạn chế trên có thể được giảm bớt bằng cách sử dụng Bar chart. Việc đánh giá các giá trị chính xác từ chiều dài thanh sẽ dễ dàng hơn nhiều so với các khu vực hoặc góc cắt, đặc biệt vì Bar chart đương nhiên có một trục dành riêng cho việc đánh dấu giá trị – không cần chú thích. Nếu cần có tỷ lệ thì các giá trị trực có thể ở dạng tỷ lệ thay vì đơn vị tự nhiên. Việc phát hiện những khác biệt nhỏ so với chiều cao của thanh cũng dễ dàng hơn, ngay cả khi chúng được đặt không theo thứ tự. Khi có nhiều danh mục, việc tìm thêm không gian cho nhiều thanh hơn là tương đối dễ dàng, đặc biệt nếu chúng được vẽ theo chiều ngang.

Nhìn chung, Bar chart là một hình ảnh trực quan đậm đặc thông tin hơn nhiều so với Pie chart. Trên thực tế, lựa chọn mặc định có thể là Bar chart. Nếu không chắc chắn liệu Pie chart có phải là lựa chọn tốt để trực quan hóa hay không thì tốt nhất nên sử dụng Bar chart một cách an toàn.

10.4.32.2.4. Khi nào nên sử dụng Pie chart

Điều đó không có nghĩa là Pie chart không có chỗ trong trực quan hóa: nó có thể hiệu quả khi truyền đạt kết quả cho người khác. Lợi ích chính của Pie chart là nó ngay lập tức thể hiện ý tưởng so sánh từng phần với toàn bộ. Với Bar chart, có thể không thể hiện ngay được mỗi thanh đóng góp bao nhiêu vào tổng thể hoặc đó là loại so sánh đáng quan tâm, trừ khi đơn vị thanh là về tỷ lệ hoặc tỷ lệ phần trăm. Khi đó, dù sao cũng cần có các chú thích bổ sung để ghi chú cả giá trị tuyệt đối cũng như giá trị tương đối.

Mặt khác, Pie chart rất quen thuộc và phù hợp với khả năng cảm thụ thẩm mỹ chung. Đặc biệt nếu chỉ quan tâm đến một hoặc hai phần, Pie chart có thể giúp làm nổi bật câu chuyện xung quanh những phần đó. Khi các lát nằm xung quanh các phần số nhỏ ($1/3, 1/4$), những điều rút ra đó có thể được chuyển tải dễ dàng bằng một chiếc bánh. Việc kết hợp các phần nhỏ hoặc không thú vị vào nhóm 'khác' có thể làm sạch thông tin mà Pie chart cần hiển thị. Bar chart có thể tốt hơn trong trường hợp chung, nhưng nếu cần trình bày những phát hiện cho người khác, Pie chart có thể sẽ hiệu quả và hấp dẫn hơn.

10.4.32.2.5. Kết luận

Cả Bar chart và Pie chart đều là những lựa chọn phổ biến khi vẽ các giá trị số dựa trên các nhãn phân loại. Nói chung, tính linh hoạt của Bar chart và mật độ thông tin cao hơn khiến Bar chart trở thành một lựa chọn mặc định tốt. Tuy nhiên, Pie chart có phạm vi hẹp nếu nó là lựa chọn phù hợp để truyền tải những thông tin sau:

- Giá trị của các nhóm riêng lẻ phải cộng lại thành một tổng có ý nghĩa.
- Sự so sánh từng phần với toàn bộ phải được quan tâm hơn là so sánh giữa các nhóm.
- Số lát nên tương đối nhỏ, nhiều nhất là khoảng năm lát.
- Các phần quan tâm có thể xác định được của các vùng, là bội số của $1/4$ hoặc $1/3$.

Nếu Pie chart không thỏa 4 điều kiện này, Bar chart có lẽ là lựa chọn tốt hơn. Nhưng nếu có câu hỏi về việc sử dụng cái nào thì bạn nên thử cả hai để xem loại biểu đồ nào trình bày dữ liệu của bạn tốt nhất.

10.4.33. Pyramid Chart

Pyramid Chart (Biểu đồ kim tự tháp): Được sử dụng để trực quan hóa dữ liệu theo cách phân cấp, thường hiển thị tỷ lệ hoặc các giai đoạn trong một quy trình.

10.4.34. Sankey Diagram

10.4.34.1. Giới thiệu

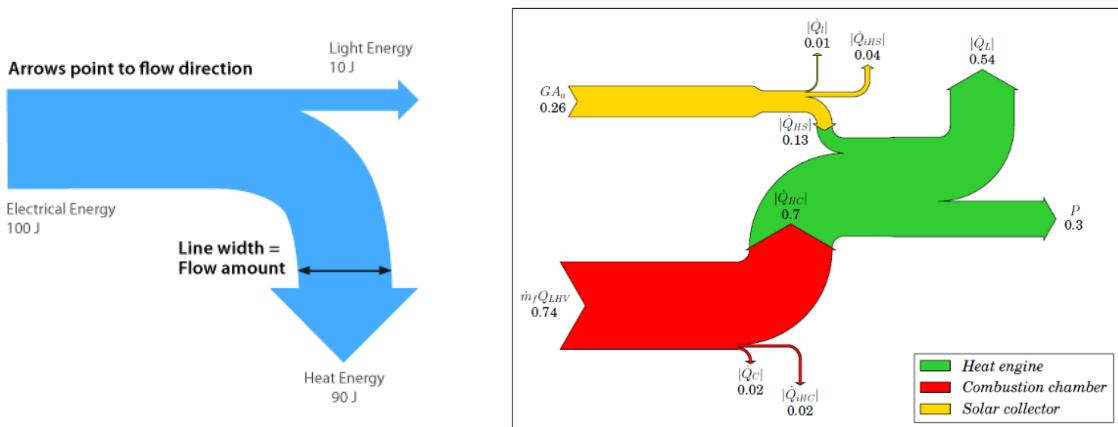
- Là một loại sơ đồ dòng trong đó chiều rộng của các mũi tên tỷ lệ với tốc độ dòng chảy. Dòng chảy này có thể là bất kỳ đại lượng nào có thể đo lường được, vì vậy mũi tên càng lớn thì lượng dòng chảy càng lớn. Màu sắc được sử dụng để chia sơ đồ thành các loại khác nhau hoặc để hiển thị sự chuyển đổi từ trạng thái này sang trạng thái khác của quá trình.

- Sankey diagram nhấn mạnh các chuyển giao hoặc luồng chính trong hệ thống. Chúng giúp xác định những đóng góp quan trọng nhất cho một luồng. Chúng thường hiển thị số lượng được bảo toàn trong ranh giới hệ thống xác định.

10.4.34.2. Sử dụng

Sankey diagram thường được dùng để trực quan hóa các tài khoản năng lượng, tài khoản dòng nguyên liệu ở cấp khu vực hoặc quốc gia và phân tích chi phí.

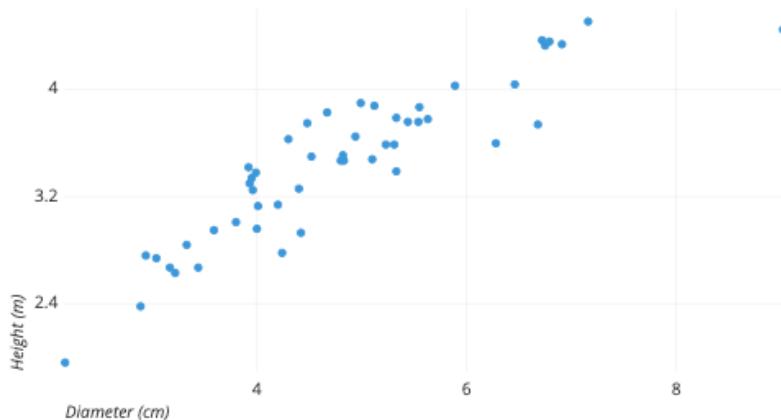
10.4.34.3. Minh họa



10.4.35. Scatter plots (hay Scatter chart hoặc Scatter graph-biểu đồ phân tán)

10.4.35.1. Giới thiệu

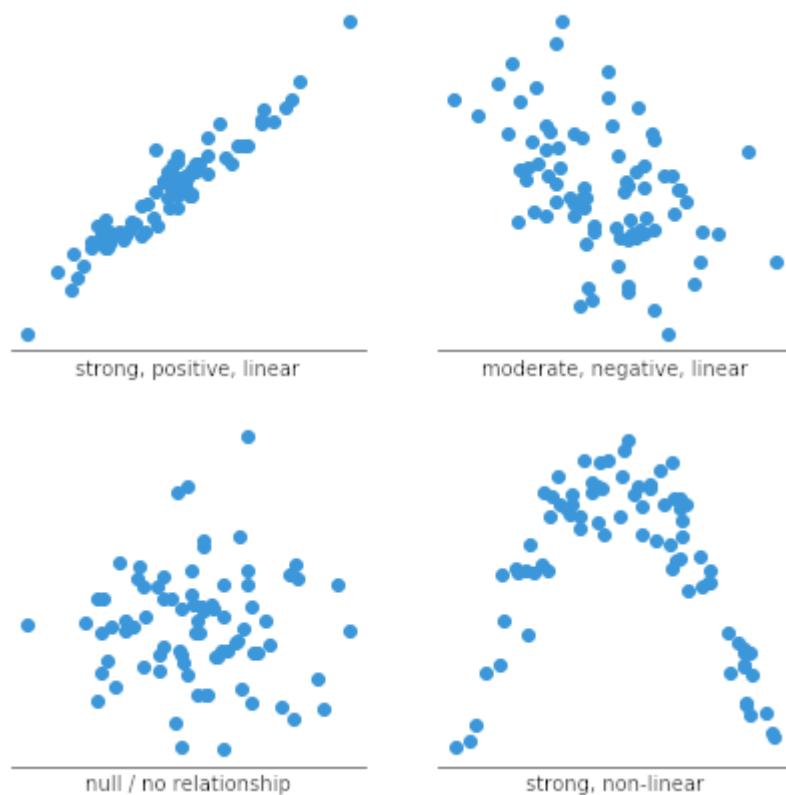
- Scatter plot tương tự như Line chart ở chỗ chúng sử dụng 2 trục ngang và dọc để vẽ các điểm dữ liệu. Scatter plot sử dụng các dấu chấm để biểu thị các giá trị cho hai biến số khác nhau. Vị trí của mỗi dấu chấm trên trục ngang và trục dọc biểu thị giá trị cho một điểm dữ liệu riêng lẻ.
- Scatter plot thường được sử dụng để quan sát các mối quan hệ giữa 2 biến (biến nguyên nhân hay biến độc lập và biến kết quả hay biến phụ thuộc vì giá trị của nó phụ thuộc vào biến đầu tiên). Mỗi tương quan này được biểu diễn dưới dạng các dấu chấm tròn đại diện cho 2 biến, với một biến phụ thuộc chạy cố định trên trục tung và một biến độc lập chạy cố định dựa vào trục hoành. Nếu không có mối tương quan, các dấu chấm sẽ xuất hiện ở những vị trí ngẫu nhiên trên biểu đồ. Nếu có mối tương quan chặt chẽ, các điểm sẽ nằm gần nhau và tạo thành một đường xuyên qua biểu đồ được gọi là đường xu hướng (trend line).
- Ví dụ: có thể sử dụng Scatter plot để hiển thị mối quan hệ giữa chiều cao và cân nặng của một người. Quá trình này bao gồm việc vẽ một biến dọc theo trục hoành và biến còn lại dọc theo trục tung.



Scatter plot trong minh họa trên cho thấy mối liên hệ giữa đường kính và chiều cao của các cây trồng. Mỗi dấu chấm tượng trưng cho một cây; vị trí nằm ngang của mỗi điểm biểu thị đường kính của cây đó (tính bằng cm) và vị trí thẳng đứng biểu thị chiều cao của cây đó (tính bằng mét). Từ đồ thị, có thể thấy mối tương quan dương nhìn chung chặt chẽ giữa đường kính và chiều cao của cây. Cũng có thể quan sát một điểm ngoại lệ, một cây có đường kính lớn hơn nhiều so với những cây khác. Cây này có vẻ khá ngắn so với chu vi của nó, điều này có thể cần được điều tra thêm.

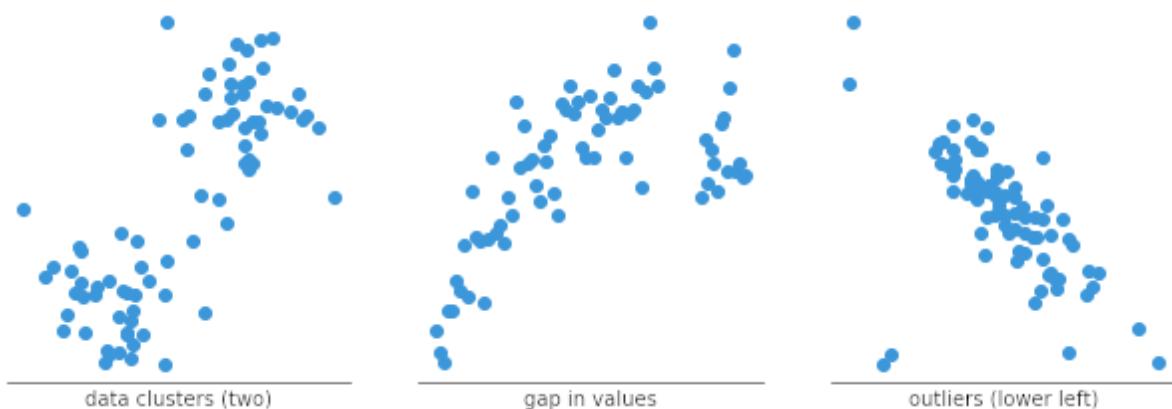
10.4.35.2. Sử dụng

- Công dụng chính của Scatter plot là quan sát và hiển thị mối quan hệ giữa hai biến số. Các dấu chấm trong Scatter plot không chỉ hiển thị giá trị của các điểm dữ liệu riêng lẻ mà còn hiển thị các mẫu khi dữ liệu được lấy tổng thể.



- Việc xác định các mối quan hệ tương quan là phổ biến với Scatter plots. Trong những trường hợp này, người dùng muốn biết, nếu được cung cấp một giá trị theo chiều ngang cụ thể thì dự đoán nào sẽ tốt cho giá trị theo chiều dọc.
- Thông thường, biến trên trục hoành biểu thị một biến độc lập và biến trên trục tung biểu thị biến phụ thuộc. Mối quan hệ giữa các biến có thể được mô tả theo nhiều cách: tăng hay giảm, mạnh hay yếu, tuyến tính hoặc phi tuyến.
- Những điểm chính cần quan tâm khi sử dụng scatter plot:
 - Dùng cho dữ liệu định lượng (số).
 - Hiển thị mối tương quan
 - Có thể tìm thấy sức mạnh và hướng đi của mối quan hệ.
- Scatter plot cũng có thể hữu ích trong các trường hợp:
 - Các điểm dữ liệu được phân chia thành các nhóm dựa trên mức độ tập hợp các điểm.
 - Cũng có thể hiển thị liệu có bất kỳ khoảng trống không mong muốn nào trong dữ liệu hay không
 - Phát hiện các điểm ngoại lệ.

Những điều này có thể hữu ích nếu muốn phân chia dữ liệu thành các phần khác nhau, chẳng hạn như trong quá trình phát triển tính cách người dùng.



10.4.35.3. Tạo Example of data structure

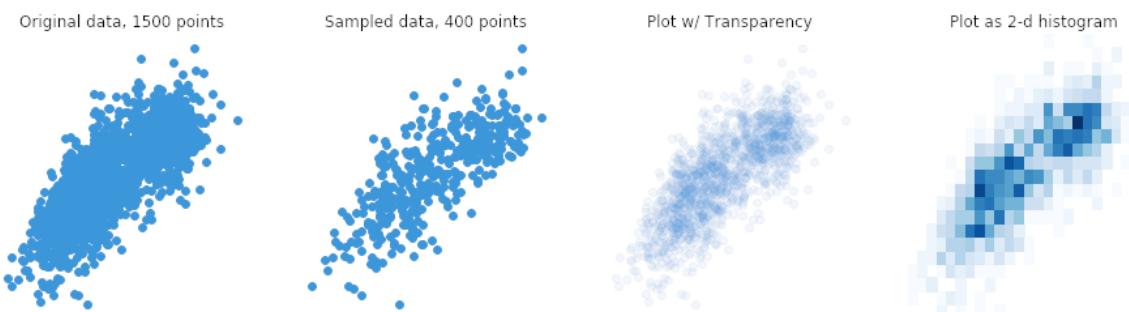
- Để tạo Scatter plot, cần chọn hai cột từ bảng dữ liệu, mỗi cột cho mỗi chiều của biểu đồ. Mỗi hàng của bảng sẽ trở thành một dấu chấm duy nhất trong ô với vị trí theo các giá trị cột.
- Ví dụ: để vẽ

Diameter	Height
4.20	3.14
5.55	3.87
3.33	2.84
6.91	4.34
...	...

10.4.35.4. Các vấn đề thường gặp khi sử dụng Scatter plot

10.4.35.4.1. Các điểm vẽ chồng lên nhau

- Khi có nhiều điểm dữ liệu để vẽ, có thể dẫn đến các điểm vẽ chồng lên nhau đôi khi dẫn đến mức gây khó khăn trong việc nhìn thấy mối quan hệ giữa các điểm và biến. Có thể khó để biết các điểm dữ liệu có mật độ dày đặc như thế nào khi nhiều điểm dữ liệu nằm trong một khu vực nhỏ.
- Có một số cách phổ biến để giảm bớt vấn đề này:
 - (i). Chỉ lấy mẫu một tập hợp con các điểm dữ liệu: việc lựa chọn các điểm ngẫu nhiên vẫn phải đưa ra ý tưởng chung về các mẫu trong dữ liệu đầy đủ.
 - (ii). Có thể thay đổi hình dạng của các dấu chấm, thêm độ trong suốt để cho phép hiển thị các phần chồng chéo hoặc giảm kích thước của điểm để ít xảy ra sự chồng chéo hơn.
 - (iii). Có thể chọn loại biểu đồ khác như heatmap, trong đó màu sắc biểu thị số điểm trong mỗi thùng (bin). Heatmap trong trường hợp sử dụng này còn được gọi là histogram 2 chiều (2-d histograms).



10.4.35.4.2. Cho rằng mối tương quan trên đồ thị là quan hệ nhân quả

- (i). Đây không phải là vấn đề lớn với việc tạo Scatter plot mà là vấn đề về cách diễn giải nó. Đơn giản vì khi quan sát mối quan hệ giữa hai biến trong Scatter plot, điều đó không có nghĩa là những thay đổi ở một biến sẽ dẫn đến những thay đổi ở biến kia (quan hệ nhân quả) mà có thể mối quan hệ được quan sát được điều khiển bởi một số biến thứ ba ảnh hưởng đến cả hai biến được vẽ, mối liên hệ nhân quả bị đảo ngược hoặc mô hình đó chỉ đơn giản là ngẫu nhiên.

Ví dụ, nhìn vào số liệu thống kê của 1 thành phố: lượng khói gian xanh tăng dần qua các năm, song song đó số lượng tệ nạn cũng tăng qua các năm. Sẽ là sai lầm khi kết luận rằng cái này gây ra cái kia. Mà có thể bỏ qua thực tế là các thành phố lớn hơn với nhiều người hơn sẽ có xu hướng có nhiều tệ nạn hơn.

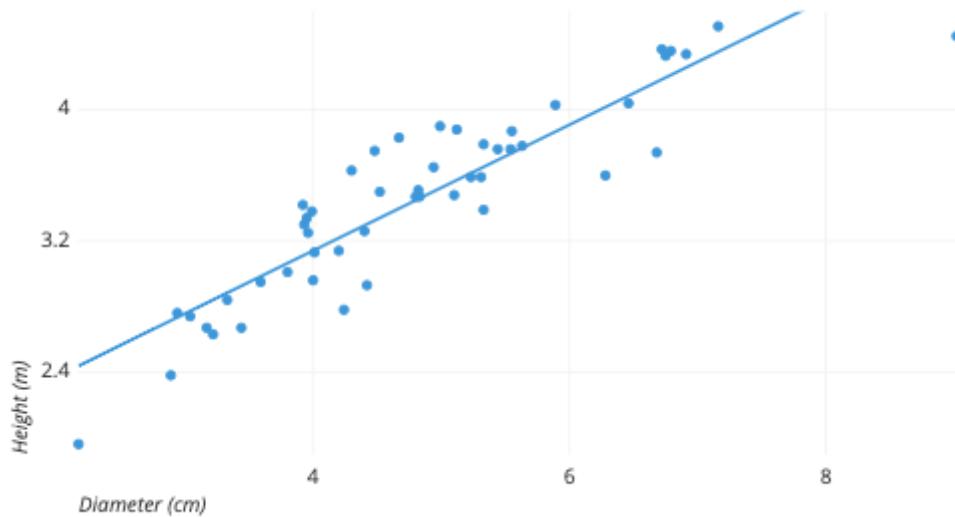
- (ii). Nếu cần thiết lập mối liên hệ nhân quả thì cần phải thực hiện phân tích sâu hơn để kiểm soát hoặc giải thích các tác động của các biến tiềm ẩn khác, nhằm loại trừ các giải thích có thể khác.

10.4.35.5. Các tùy chọn phổ biến được dùng trên Scatter plot

10.4.35.5.1. Thêm đường xu hướng (trend line)

Khi sử dụng Scatter plot để xem xét mối quan hệ dự đoán hoặc tương quan giữa các biến, người ta thường thêm một đường xu hướng vào biểu đồ để thể hiện sự phù hợp nhất về

mặt toán học với dữ liệu. Điều này có thể cung cấp một tín hiệu bổ sung về mức độ mạnh mẽ của mối quan hệ giữa hai biến số và liệu có bất kỳ điểm bất thường nào đang ảnh hưởng đến việc tính toán đường xu hướng hay không.

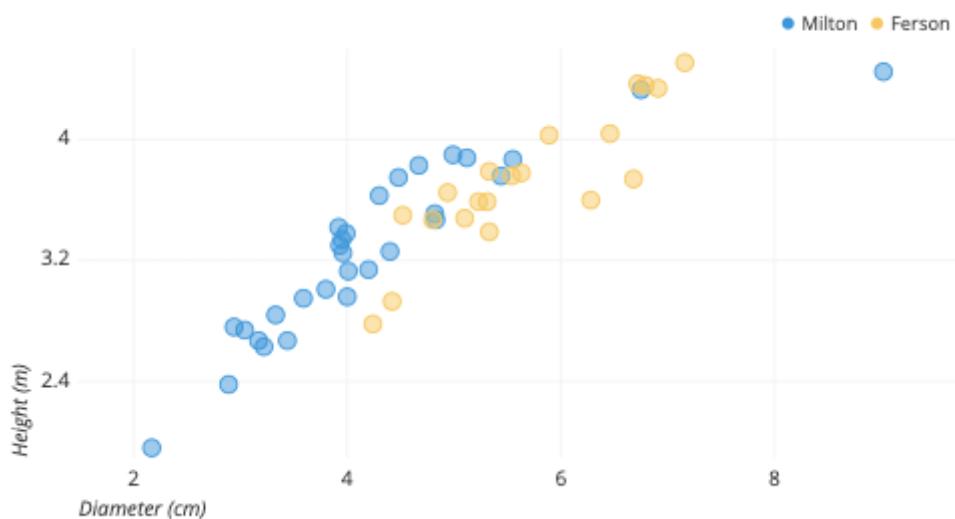


10.4.35.5.2. Sử dụng biến thứ ba cho dữ liệu phân loại

Một sửa đổi phổ biến cơ bản của scatter plot là việc bổ sung biến thứ ba. Giá trị của biến thứ ba có thể được mã hóa bằng cách sửa đổi cách vẽ các điểm.

(i). Đổi với thuộc tính thứ 3 thuộc kiểu phân loại (danh nghĩa):

- *Sử dụng màu:* Đổi với biến thứ ba biểu thị các giá trị phân loại (như khu vực địa lý hoặc giới tính), cách mã hóa phổ biến nhất là thông qua màu của điểm. Việc tạo cho mỗi điểm thuộc 1 nhóm sẽ có cùng một màu sắc riêng biệt giúp cho người xem dễ dàng hiểu rõ hơn về dữ liệu.



Qua hình trên, ta thấy việc tô màu các điểm theo loại cây cho thấy Fersons (màu vàng) nhìn chung rộng hơn Miltions (màu xanh), nhưng cũng ngắn hơn ở cùng đường kính.

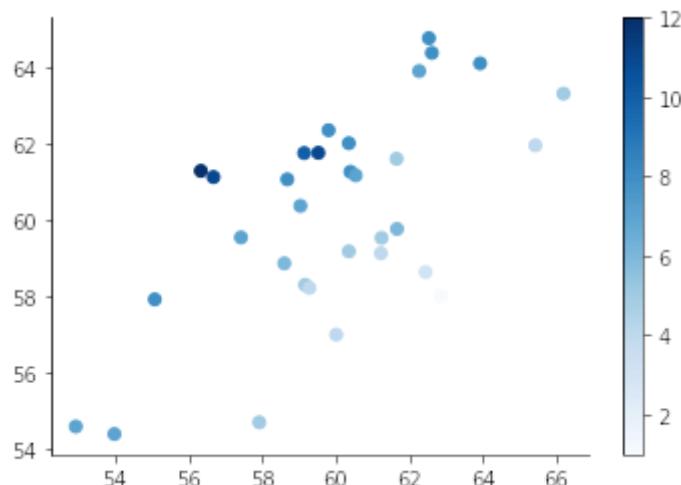
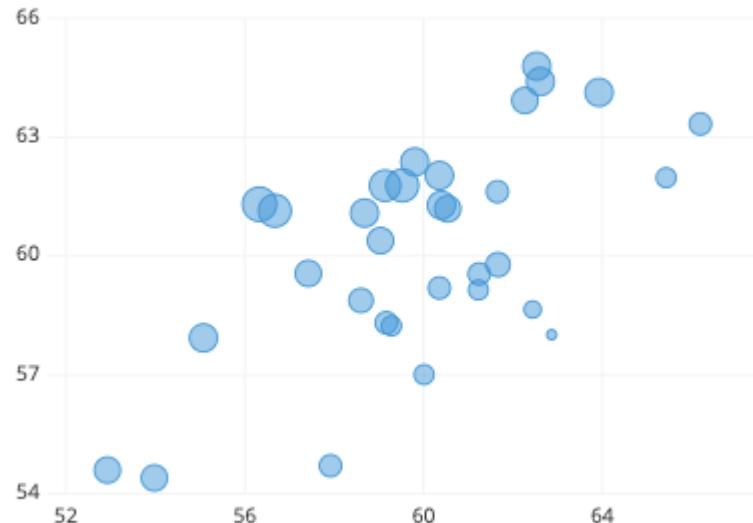
- *Sử dụng hình dạng:* Một tùy chọn khác đôi khi được thấy đối với mã hóa biến thứ ba là hình dạng. Một vấn đề tiềm ẩn với hình dạng là các hình dạng khác nhau có thể có kích thước và diện tích bề mặt khác nhau, điều này có thể ảnh hưởng đến cách cảm nhận các nhóm. Tuy nhiên, trong một số trường hợp không thể sử dụng màu sắc (như trong in ấn), hình dạng có thể là lựa chọn tốt nhất để phân biệt giữa các nhóm.



Các hình dạng trên đã được thu nhỏ để sử dụng cùng một lượng mực in.

(ii). Đối với các biến thứ ba có giá trị số

- *Sử dụng bong bóng (Bubble):* cách mã hóa phổ biến là thay đổi kích thước điểm. Biểu đồ phân tán có kích thước điểm dựa trên biến thứ ba thực sự có một tên riêng biệt là Bubble chart. Điểm có bong bóng lớn hơn cho thấy giá trị cao hơn.

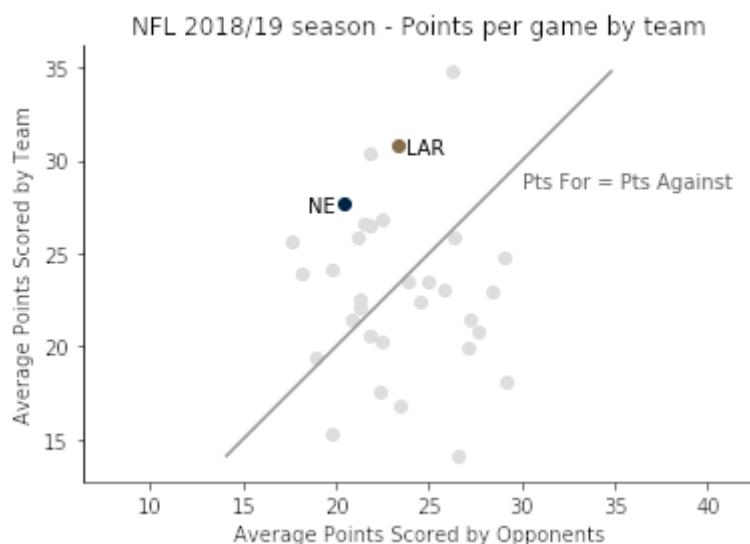


- *Sử dụng chuỗi màu liên tục:* cũng là một sự thay thế khác, thay vì sử dụng các màu riêng biệt cho các điểm như trong trường hợp phân loại, ở đây sẽ sử dụng

một chuỗi màu liên tục, chẳng hạn như các màu tối hơn biểu thị giá trị cao hơn. Lưu ý rằng, đối với cả kích thước và màu sắc, chú thích rất quan trọng để giải thích biến thứ ba, vì mắt khó có thể phân biệt được kích thước và màu sắc dễ dàng như khi phân biệt vị trí.

10.4.35.5.3. Đánh dấu bằng chú thích và màu sắc

Nếu muốn sử dụng Scatter plot để trình bày thông tin chi tiết, nên làm nổi bật các điểm quan tâm cụ thể thông qua việc sử dụng chú thích và màu sắc. Việc loại bỏ các điểm không quan trọng sẽ làm nổi bật các điểm còn lại và cung cấp thông tin tham khảo để so sánh các điểm còn lại.



10.4.35.6. Các đồ thị liên quan

10.4.35.6.1. Scatter map

Khi hai biến trong biểu đồ phân tán là tọa độ địa lý - vĩ độ và kinh độ - có thể chồng các điểm trên bản đồ để có được bản đồ phân tán (còn gọi là Dot map). Điều này có thể thuận tiện khi bối cảnh địa lý hữu ích cho việc rút ra những hiểu biết cụ thể và có thể được kết hợp với các mã hóa biến thứ ba khác như kích thước điểm và màu sắc.

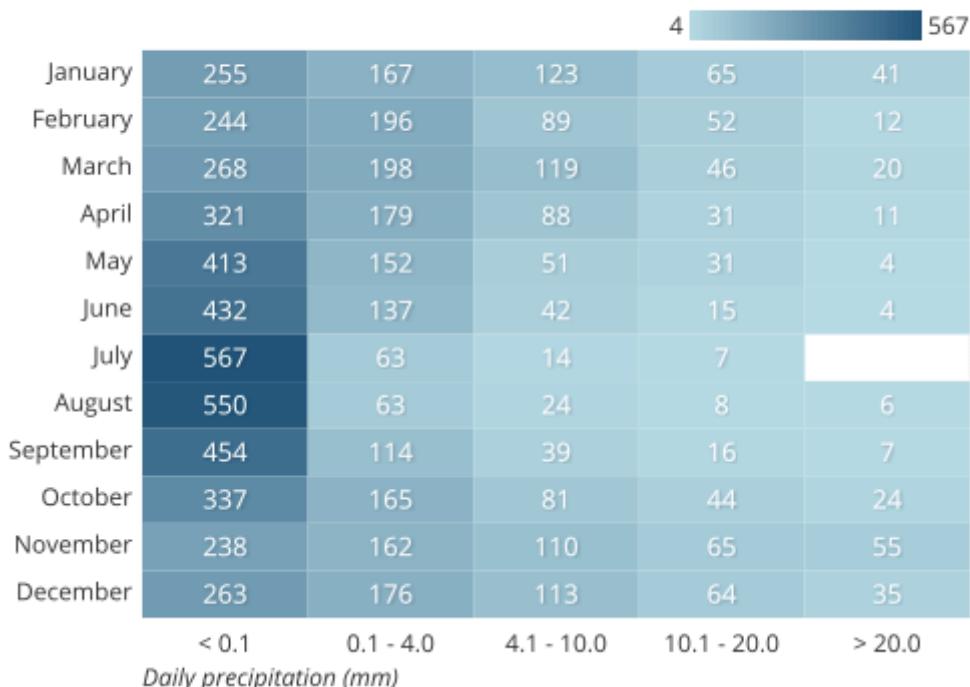
Một ví dụ nổi tiếng về bản đồ phân tán là bản đồ bùng phát dịch tả năm 1854 của John Snow¹, cho thấy các trường hợp dịch tả (thanh màu đen) tập trung xung quanh một máy bơm nước cụ thể trên Phố Broad (đáy chấm ở giữa).

¹ [Wikimedia Commons](#)



10.4.35.6.2. Heatmap

Seattle precipitation by month, 1998-2018



Minh họa lượng mưa ở Seattle theo tháng từ 1998-2018

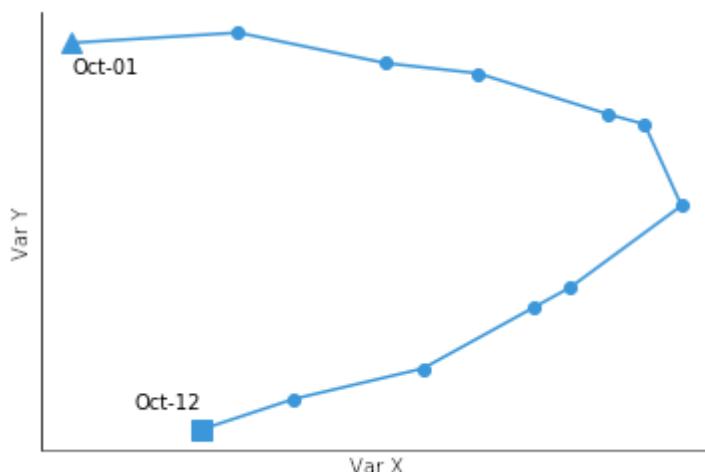
Như đã lưu ý ở trên, Heatmap có thể là một lựa chọn thay thế tốt cho biểu đồ phân tán khi có nhiều điểm dữ liệu cần được vẽ và mật độ của chúng gây ra vấn đề mật độ các điểm nhiều quá mức.

Tuy nhiên, bản đồ nhiệt cũng có thể được sử dụng theo cách tương tự để hiển thị mối quan hệ giữa các biến khi một hoặc cả hai biến không liên tục và là số.

Nếu cố gắng mô tả các giá trị rời rạc bằng Scatter plot, tất cả các điểm của một mức sẽ nằm trên một đường thẳng. Heatmap có thể khắc phục tình trạng vẽ đồ thị quá mức này thông qua việc gộp các giá trị của chúng vào các hộp đếm (boxes of counts).

10.4.35.6.3. Scatter plot với đường kết nối

Nếu muốn thêm vào Scatter plot biến thứ 3 biến thị dấu thời gian thì một loại biểu đồ có thể chọn là Connected scatter plot (biểu đồ phân tán được kết nối). Thay vì sửa đổi hình thức của các điểm để biểu thị ngày tháng, ở đây sử dụng các đoạn đường để kết nối các quan sát theo thứ tự. Điều này có thể giúp dễ dàng nhận thấy hai biến số chính không chỉ liên quan với nhau mà còn mối quan hệ đó thay đổi như thế nào theo thời gian. Nếu trực hoành cũng tương ứng với thời gian thì tất cả các đoạn thẳng sẽ nối các điểm một cách nhất quán từ trái sang phải và có một Line chart.

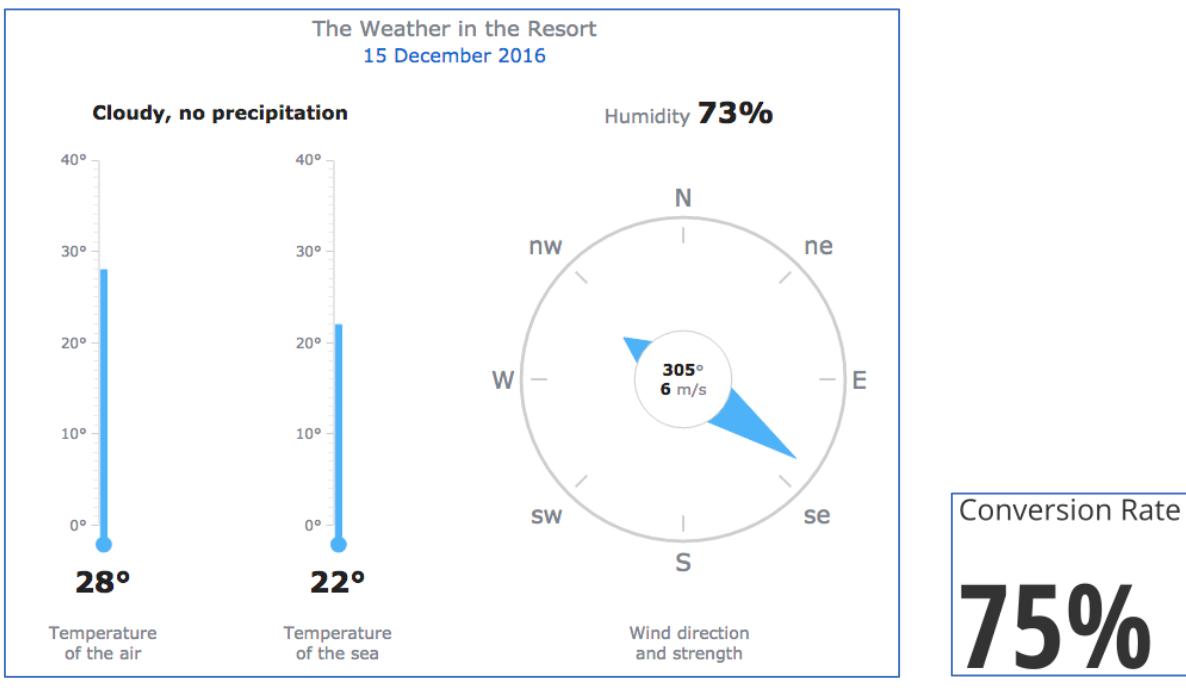


10.4.36. Single values chart (Biểu đồ chỉ gồm 1 giá trị)¹

- Single value chart dùng để hiển thị kết quả cho truy vấn tìm kiếm hoặc số liệu dưới dạng một giá trị duy nhất để phân tích nhanh. Được dùng khi chỉ có một số để hiển thị, hoặc muốn nhấn mạnh 1 số liệu quan trọng nào đó. Việc chỉ hiển thị giá trị là một cách tiếp cận hợp lý để mô tả dữ liệu. Khi các giá trị chính xác được quan tâm trong phân tích, có thể đưa chúng vào bảng đi kèm hoặc thông qua các chú thích trên hình ảnh trực quan.
- Single value chart hiển thị một bản ghi từ một tìm kiếm để làm nổi bật giá trị đó trong nháy mắt. Nếu truy vấn trả về nhiều giá trị trong tab Tổng hợp thì chỉ giá trị đầu tiên được hiển thị trong biểu đồ giá trị duy nhất.
- Các hình tượng thường dùng gồm: nhiệt kế, la bàn, đồng hồ tốc độ, v.v.:
- Các trường hợp thường dùng: sức tải của máy tải máy chủ so với công suất thiết kế, số lượng phòng đang được thuê, thể tích bể chứa, chỉ báo nhiệt độ, chỉ báo tốc độ, hiển thị số lượng lỗi do trang web do công ty tạo ra trong 24 giờ qua v.v...

¹ [Single-Value Data \(Indicators\): Choose Right Chart Type for Data Visualization \(Part 5\)](#)

- Một số minh họa:



Single Value Chart

Operations Cost

\$850,111

Used to draw quick attention to high-level metrics.

Single Value Indicator Chart

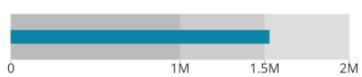
Yearly Subscriptions

8,519 4.22%

Used to draw quick attention to high-level metrics with an indication of movement.

Bullet Chart

Revenue



Used to draw quick attention to high-level metrics in relation to a goal.

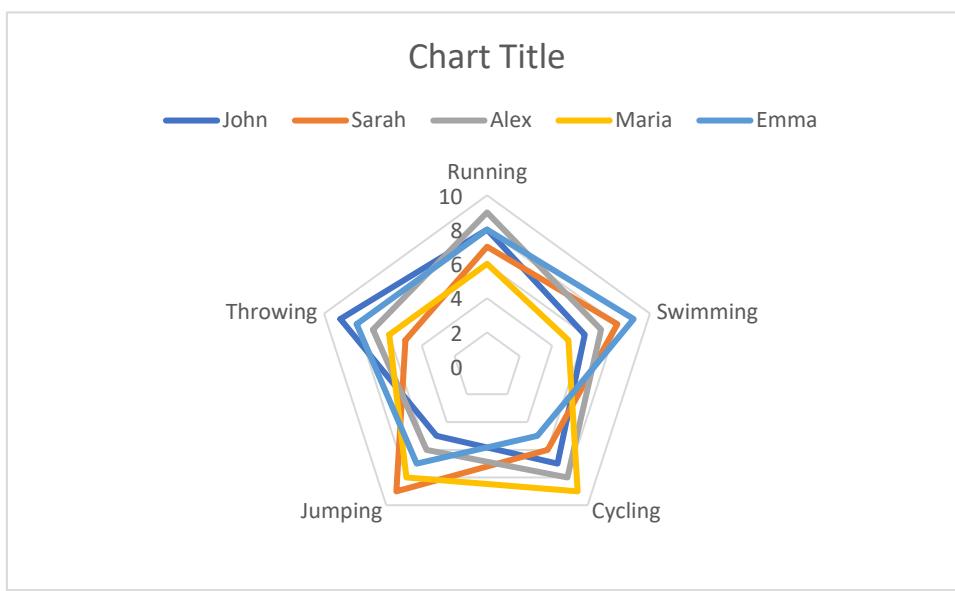
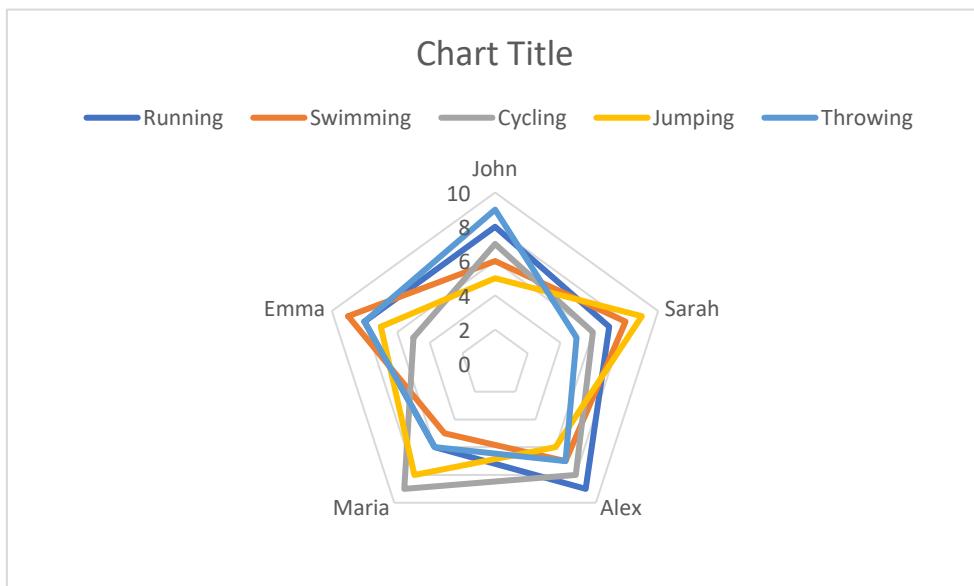
10.4.37. Spider Chart¹

- Radar Chart (hay Spider chart) dùng để hiển thị dữ liệu đa biến dưới dạng biểu đồ hai chiều gồm ba biến định lượng trở lên được biểu diễn trên các trục bắt đầu từ cùng một điểm.
- Radar Chart được chia thành nhiều loại tùy thuộc vào trường hợp sử dụng:
 - Biểu đồ so sánh: Khi được sử dụng để so sánh các danh mục hoặc nhóm khác nhau, Biểu đồ nhện giúp trực quan hóa hiệu suất trên nhiều chiều.
 - Biểu đồ đa biến: Biểu đồ này rất tuyệt vời để hiển thị dữ liệu đa biến, cho phép so sánh nhiều biến cùng một lúc.
 - Biểu đồ thành phần: Nó cũng có thể được sử dụng để hiển thị thành phần của một tổng thể, với mỗi trục đại diện cho một thành phần khác nhau.
 - Biểu đồ chuyên ngành: Với thiết kế hình tròn độc đáo và khả năng xử lý nhiều điểm dữ liệu, Biểu đồ Nhện có thể được coi là chuyên dụng cho các mục đích cụ thể của nó.

¹ <https://www.explor.co/blog/understanding-spider-charts-what-is-a-spider-chart-and-how-to-use-it>

- Spider chart đặc biệt hữu ích trong trường hợp muốn so sánh số liệu hiệu suất, cấp độ kỹ năng hoặc kết quả khảo sát trên nhiều danh mục bằng một hình ảnh trực quan, dễ hiểu.
- Ví dụ:

Athlete	Running	Swimming	Cycling	Jumping	Throwing
John	8	6	7	5	9
Sarah	7	8	6	9	5
Alex	9	7	8	6	7
Maria	6	5	9	8	6
Emma	8	9	5	7	8



10.4.38. Spline chart

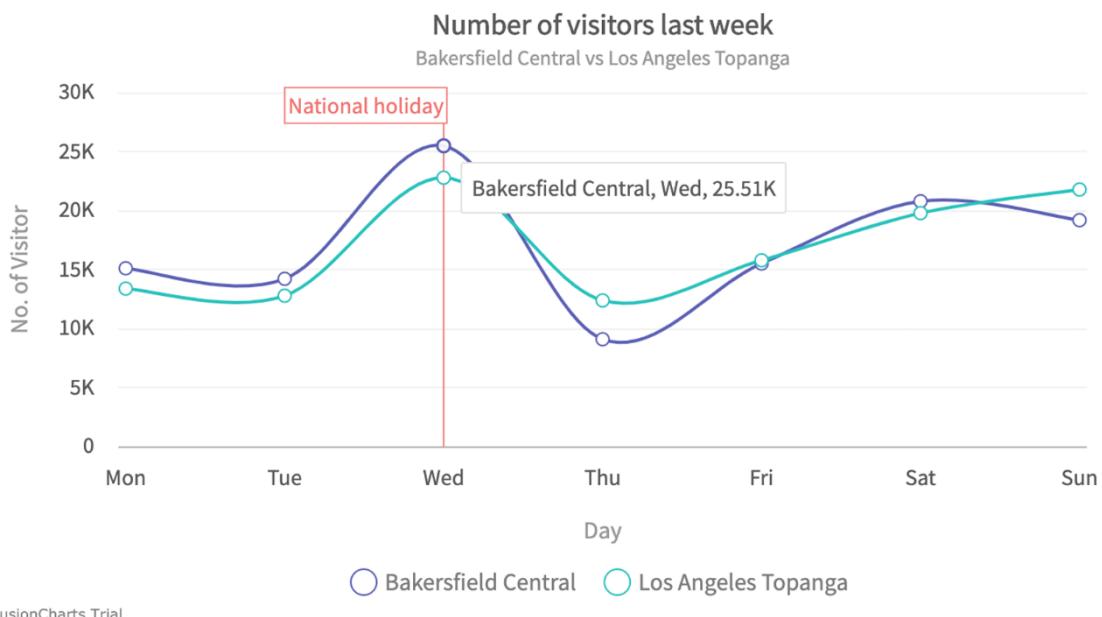
10.4.38.1. Giới thiệu

Là phiên bản của Line chart, chúng khác nhau ở chỗ, dữ liệu được kết nối với các dấu chấm tạo thành đường cong nhằm mục đích cải thiện thiết kế của biểu đồ để tính ra các giá trị bị thiếu, trái ngược với Line chart.

10.4.38.2. Sử dụng

Giống như Line chart thông thường, nó thường được sử dụng để nhấn mạnh xu hướng của dữ liệu trong các khoảng thời gian bằng nhau.

10.4.38.3. Minh họa

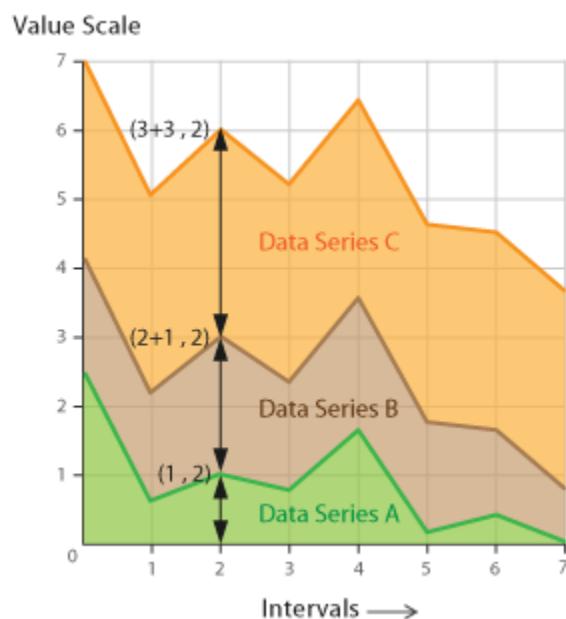


10.4.39. Stacked area chart

10.4.39.1. Giới thiệu

Là một dạng biểu đồ phức hợp của Area graphs, chúng hoạt động giống Area graphs. Stacked area chart được sử dụng để truyền đạt các số nguyên, vì chúng không hoạt động với các giá trị âm. Chúng hữu ích trong việc so sánh nhiều biến đã thay đổi như thế nào trong một khoảng thời gian.

10.4.39.2. Minh họa



10.4.40. Stacked Column/Bar charts (Biểu đồ thanh xếp chồng)

10.4.40.1. Giới thiệu

- Stacked bar chart (còn gọi là Stacked bar graph) mở rộng Bar graph tiêu chuẩn từ việc xem xét các giá trị số trên một biến phân loại thành hai biến. Mỗi thanh trong Bar graph được chia thành một số thanh phụ xếp chồng lên nhau, mỗi thanh tương ứng với một mức của biến phân loại thứ hai.
- Những điểm chính cần quan tâm khi sử dụng stacked column/bar graph:
 - Tương tự như Pie chart nhưng cho phép:
 - Giá trị biểu diễn là giá trị số chứ không phải phần trăm.
 - So sánh nhiều tổng thể hơn.
 - Có thể được dùng để đếm hoặc ước tính phần trăm của từng danh mục trong tổng thể.



Stacked bar chart ở trên mô tả doanh thu từ một nhà bán lẻ đồ nghề trong một khoảng thời gian cụ thể, qua hai biến phân loại: vị trí cửa hàng và loại hàng (Clothing, Equipment, Accessories). Biến phân loại chính là vị trí cửa hàng: có thể thấy từ chiều cao tổng thể của thanh được sắp xếp rằng vị trí Cherry St. có doanh thu cao nhất và Apple Rd. thấp nhất. Mỗi thanh được chia nhỏ dựa trên cấp độ của biến phân loại thứ hai là loại hàng. Có thể thấy rằng ở hầu hết các địa điểm, doanh số bán quần áo (clothing) lớn hơn một chút so với thiết bị (Equipment), và thiết bị lại lớn hơn phụ kiện (accessories). Vị trí của Strawberry Mall dường như có tỷ trọng doanh thu từ thiết bị thấp hơn, trong khi thiết bị có tỷ trọng doanh thu lớn hơn ở Peach St.

10.4.40.2. Sử dụng

Mục tiêu chính của Bar chart tiêu chuẩn là so sánh các giá trị số giữa các cấp của một biến phân loại. Một thanh được vẽ cho từng cấp độ của biến phân loại, độ dài của mỗi thanh biểu thị giá trị số. Stacked bar chart cũng đạt được mục tiêu này nhưng cũng hướng tới mục tiêu thứ hai.

Stacked bar chart được dùng khi muốn quan tâm đến sự phân tách tương đối của từng thanh chính dựa trên mức của biến phân loại thứ hai. Mỗi thanh chính sẽ bao gồm một số thanh phụ, mỗi thanh tương ứng với một cấp độ của biến phân loại thứ hai. Tổng chiều dài của mỗi thanh xếp chồng lên nhau vẫn giống như trước, nhưng cho thấy các nhóm phụ đóng góp vào tổng chiều dài đó như thế nào.

10.4.40.2.1. Thứ tự của các biến phân loại (Order of categorical variables)

Một điều quan trọng cần cân nhắc khi xây dựng Stacked bar chart là quyết định biến nào trong số hai biến phân loại sẽ là biến chính (cho biết vị trí trực chính và chiều dài thanh tổng thể) và biến nào sẽ là biến phụ (cho biết cách chia từng thanh chính). Biến 'quan trọng'

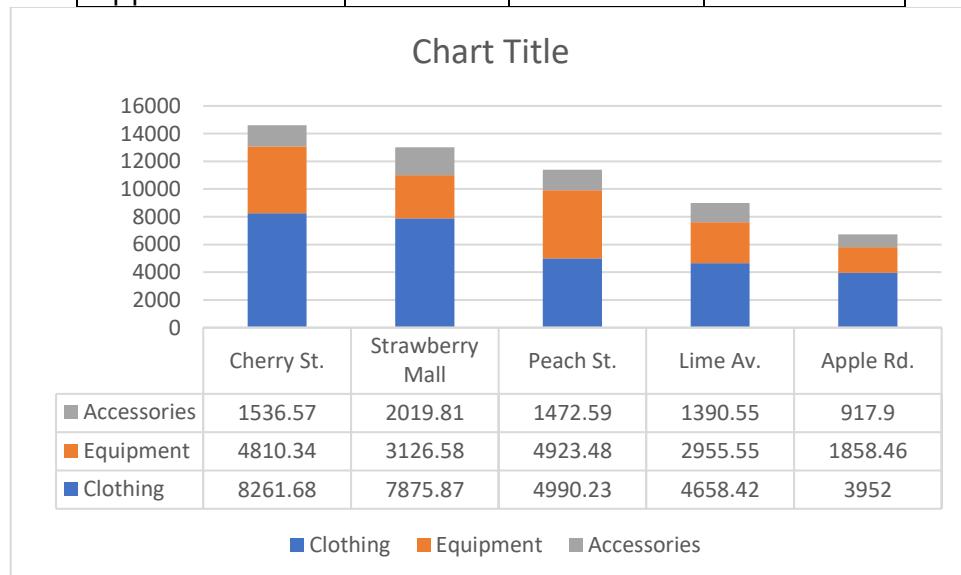
nhất phải là biến chính; cần sử dụng kiến thức về lĩnh vực của dữ liệu và loại biến phân loại cụ thể để đưa ra quyết định về cách gán các biến phân loại.

Ví dụ: nếu một biến phân loại mô tả dữ liệu thời gian (ví dụ: tháng 1-2024, tháng 2-2024, tháng 3-2024, v.v.) thì đó thường sẽ là lựa chọn rõ ràng cho biến phân loại chính. Tiếp theo trong hệ thống phân cấp chung là các biến số hoặc biến thứ tự khác, như độ tuổi (18-24, 25-34, 35-44, v.v.), giới tính, nghề nghiệp, khu vực (địa lý), ... Các biến phân loại theo kiểu nhẵn thuần túy (ví dụ: giới tính, bộ phận, khu vực địa lý) thường không có trọng số mạnh để được coi là biến chính. Một điểm cần nhắc khác là các biến có nhiều cấp độ hơn thường tốt hơn khi được dùng làm biến chính; chúng tôi muốn giới hạn số lượng cấp độ phụ ở một lượng khá nhỏ để làm cho bảng phân tích dễ đọc hơn.

Cuối cùng, những quy tắc kinh nghiệm vừa nêu ở trên chỉ là những hướng dẫn chung. Kiến thức về lĩnh vực của dữ liệu, mục tiêu của việc trực quan hóa và sẽ giúp xác định hệ thống phân cấp tốt nhất cho các biến phân loại cho từng trường hợp. Ví dụ: nếu muốn xem bảng phân tích độ tuổi theo bộ phận sản phẩm thì đây là lý do chính đáng để đặt biến phân loại thuần túy (bộ phận) làm biến chính.

10.4.40.2.2. Ví dụ về cấu trúc dữ liệu dùng cho việc vẽ đồ thị

Store	Clothing	Equipment	Accessories
Cherry St.	8261.68	4810.34	1536.57
Strawberry Mall	7875.87	3126.58	2019.81
Peach St.	4990.23	4923.48	1472.59
Lime Av.	4658.42	2955.55	1390.55
Apple Rd.	3952.00	1858.46	917.90



Dữ liệu cho Stacked bar chart thường được định dạng thành bảng có ba cột trỏ lên. Các giá trị ở cột đầu tiên biểu thị mức độ của biến phân loại chính. Mỗi cột sau cột đầu tiên sẽ tương ứng với một cấp của biến phân loại phụ. Các giá trị ô chính cho biết độ dài của mỗi thanh phụ trong biểu đồ. Các thanh được tạo trên các hàng: khi Stacked Bar chart được tạo, mỗi thanh chính sẽ có tổng chiều dài bằng tổng chiều dài trên hàng tương ứng của nó.

Đối với một số công cụ nhất định, bước trung gian để tạo Stacked bar chart có thể yêu cầu tính tổng tích lũy trên mỗi hàng. Cột ngoài cùng bên phải sẽ chứa độ dài của các thanh

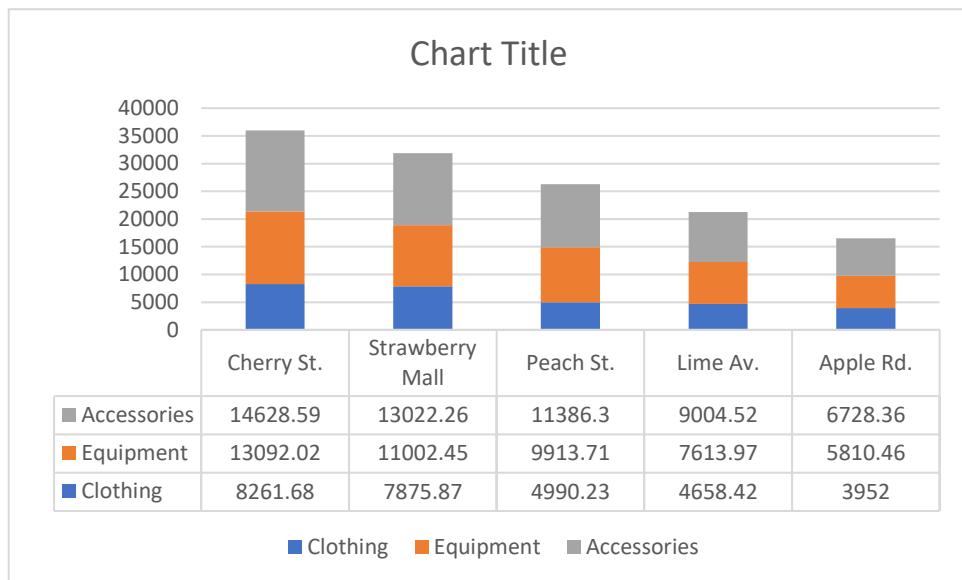
chính. Các thanh phụ được xác định bởi sự khác biệt về giá trị giữa các cột liên tiếp. Đối với các công cụ yêu cầu loại cấu trúc bảng dữ liệu này, hãy cẩn thận với các giá trị âm vì điều này có thể gây ra sự chồng chéo hoặc khoảng trống giữa các thanh khiến dữ liệu trình bày sai.

Store	Clothing	+ Equipment	+ Accessories
Cherry St.	8261.68	13072.02	14608.59
Strawberry Mall	7875.87	11002.45	13022.26
Peach St.	4990.23	9913.71	11386.30
Lime Av.	4658.42	7613.97	9004.52
Apple Rd.	3952.00	5810.46	6728.36

Giải thích: về giá trị của 2 cột +Equipment và +Accessories

+Equipment: cộng dồn giá trị 2 cột Clothing và Equipment

+ Accessories: cộng dồn giá trị 3 cột Clothing, Equipment và Accessories

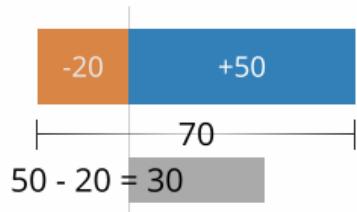


10.4.40.3. Một số phương pháp để sử dụng hiệu quả stacked bar chart

Về bản chất, Stacked bar charts đề xuất tuân theo các phương pháp hay nhất tương tự như Bar charts tiêu chuẩn mà chúng được xây dựng từ đó. Tuy nhiên, việc bổ sung biến phân loại thứ hai mang lại những cân nhắc bổ sung để tạo Stacked bar charts hiệu quả.

10.4.40.3.1. Duy trì đường cơ sở bằng 0 (a zero-baseline)

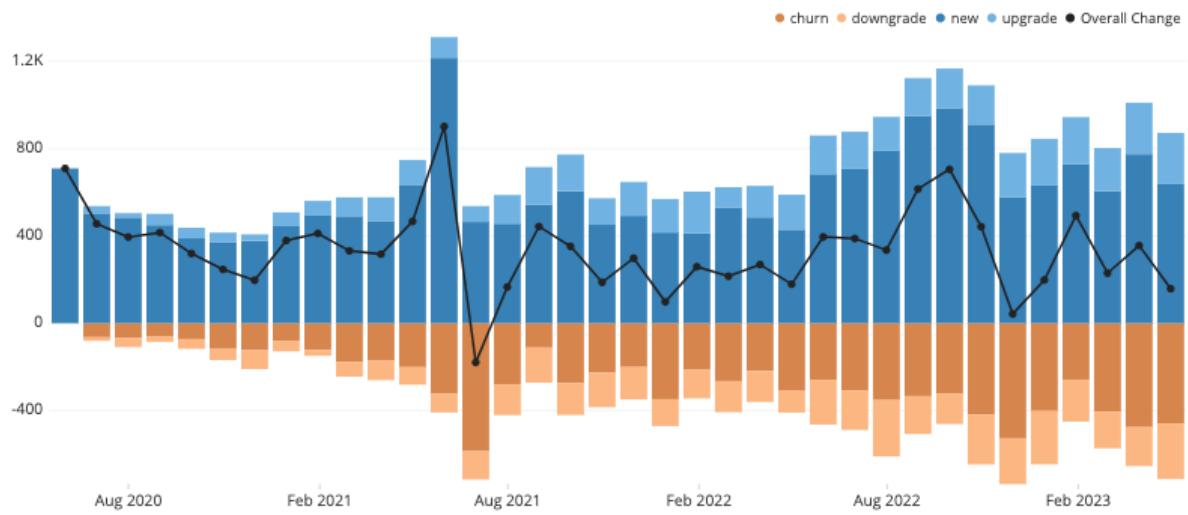
Khi Bar chart tiêu chuẩn gấp giá trị âm, thanh tương ứng sẽ được vẽ bên dưới hoặc bên trái của đường cơ sở (tùy thuộc vào việc các thanh được định hướng theo chiều dọc hay chiều ngang). Trong Stacked bar chart, có thể thực hiện cách biểu diễn tương tự bằng cách xếp chồng các thanh theo hướng âm.



Tuy nhiên, khi kết hợp các thanh dương và thanh âm thì không còn trường hợp chiều dài tổng thể của thanh tương ứng với tổng giá trị của thanh nữa. Khi điều này xảy ra, nên vẽ

thêm một đường hoặc một chuỗi điểm trên đầu các thanh để hiển thị tổng số thực: sự khác biệt giữa độ dài của thanh dương và thanh âm.

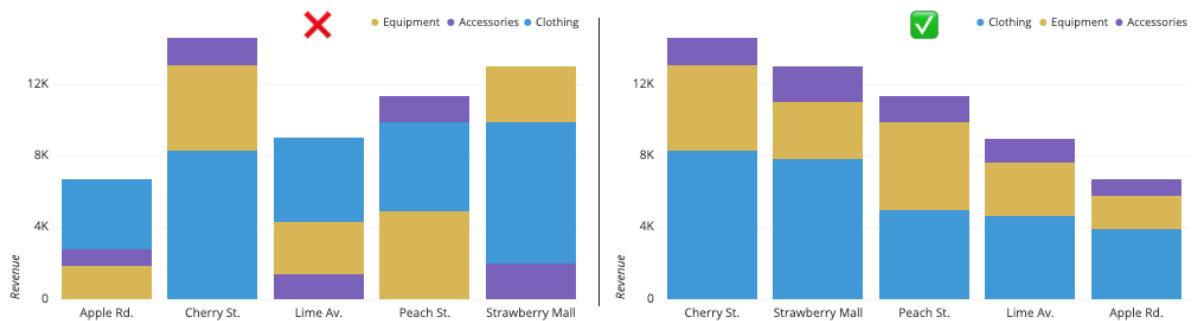
New MRR by Type



Khi các giá trị phụ luôn dương hoặc âm cho từng nhóm con, có thể dễ dàng duy trì thứ tự nhất quán của các thanh phụ trong mỗi thanh chính. Tuy nhiên, nếu nhiều nhóm con chuyển đổi giữa dương và âm vào các thời điểm khác nhau thì sẽ không thể sắp xếp thứ tự hợp lý vì các thanh chuyển đổi mức trên và dưới đường cơ sở. Trong những trường hợp này, tốt nhất nên xem xét loại biểu đồ khác cho dữ liệu. Line chart hoặc Bar chart có thể cung cấp cách hiển thị nhất quán hơn cho các nhóm riêng lẻ, mặc dù chúng mất khả năng xem tổng số chính. Nếu việc xem tổng số thực sự quan trọng thì điều đó luôn có thể được hiển thị trong một biểu đồ bổ sung khác chứ đừng mong hiển thị mọi thứ trong một biểu đồ duy nhất.

10.4.40.3.2. Sắp xếp thứ tự các cấp độ danh mục

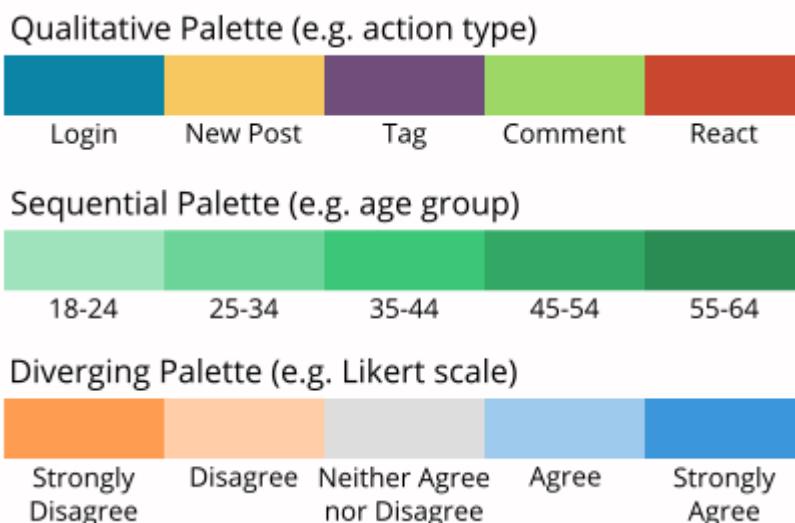
Với Stacked bar chart, sẽ cần xem xét thứ tự các cấp danh mục cho cả hai biến phân loại sẽ được vẽ. Quy tắc ngón tay cái cho Bar chart tiêu chuẩn có thể được áp dụng cho cả hai biến: sắp xếp các thanh từ lớn nhất đến nhỏ nhất trừ khi có thứ tự cấp độ nội tại.



Để làm rõ quy tắc này cho biến phân loại thứ cấp, quyết định này phải dựa trên quy mô tổng thể của từng cấp độ phân loại. Nên xếp từng thanh chính theo cùng một thứ tự. Việc duy trì tính nhất quán này giúp việc liên kết các thanh phụ với các cấp danh mục phụ trở nên dễ dàng hơn. Tính nhất quán này cũng có nghĩa là nhóm được vẽ đầu tiên luôn nằm trên đường cơ sở, làm cho kích thước của chúng dễ đọc. Do đó, nếu việc theo dõi các giá trị chính xác là quan trọng đối với một mức biến thứ cấp cụ thể thì thay vào đó, các thanh phụ của nó nên được đặt trên đường cơ sở.

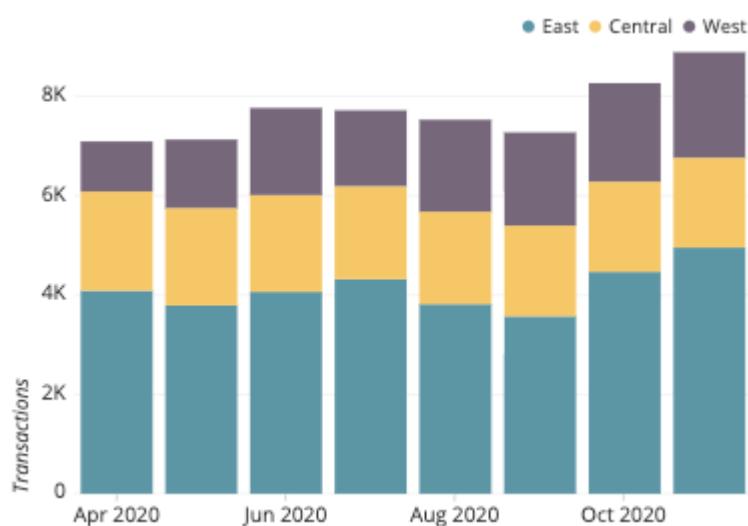
10.4.40.3.3. Lựa chọn màu sắc hiệu quả

Mặc dù khuyến nghị chung là chỉ sử dụng một màu duy nhất trong Bar chart tiêu chuẩn, việc sử dụng màu sắc để phân biệt các mức biến thứ cấp là điều tất yếu đối với Stacked bar chart. Điểm quan trọng là đảm bảo rằng việc lựa chọn bảng màu để gán cho từng cấp độ phân loại phù hợp với loại biến: bảng màu định tính cho các biến phân loại thuần túy (purely categorical) và tuần tự (sequential) hoặc phân kỳ (Diverging) cho các biến có thứ tự có ý nghĩa.



10.4.40.4. Khó khăn khi giải thích các thành phần phụ khi sử dụng Stacked bar chart

- Mặc dù việc so sánh tổng các giá trị số giữa các cấp độ của biến phân loại chính là đơn giản, nhưng việc đánh giá các phép chia hoặc so sánh khác bằng cách sử dụng biến phân loại phụ lại không đơn giản. Nếu muốn thấy sự thay đổi ở cấp độ thứ cấp trên biến phân loại chính, điều này chỉ có thể được thực hiện dễ dàng đối với cấp độ được vẽ theo đường cơ sở.
- Đối với tất cả các cấp độ thứ cấp khác, đường cơ sở của chúng sẽ trải qua những thay đổi, khiến việc đánh giá độ dài của thanh phụ thay đổi như thế nào trên các thanh chính trở nên khó khăn hơn. Trong ví dụ dưới đây, khó có thể nói rằng nhóm màu vàng ở giữa thực tế đang giảm nhẹ theo thời gian.

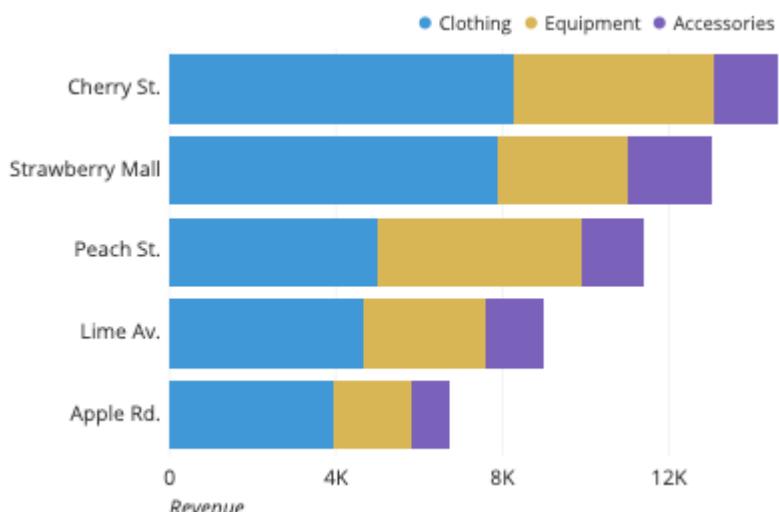


- Ngay cả việc cố gắng so sánh các thanh phụ trong mỗi thanh chính cũng có thể khó khăn. Ngay cả khi làm theo hướng dẫn để sắp xếp các cấp danh mục phụ theo kích thước tổng thể, điều này không đảm bảo rằng chúng sẽ được sắp xếp theo kích thước trong một thanh chính cụ thể. Trong cùng hình ảnh trên, thật khó để biết nhóm “West” màu tím vượt trội hơn nhóm “Center” màu vàng ở đâu về kích thước.
- Một trong những mục tiêu chuẩn của Stacked bar chart là đưa ra các đánh giá tương đối về các nhóm phụ. Nếu việc so sánh các nhóm phụ là quan trọng thì nên sử dụng loại biểu đồ khác như Line chart hoặc Grouped bar chart.

10.4.40.5. Các tùy chọn phổ biến được dùng trên stacked bar chart

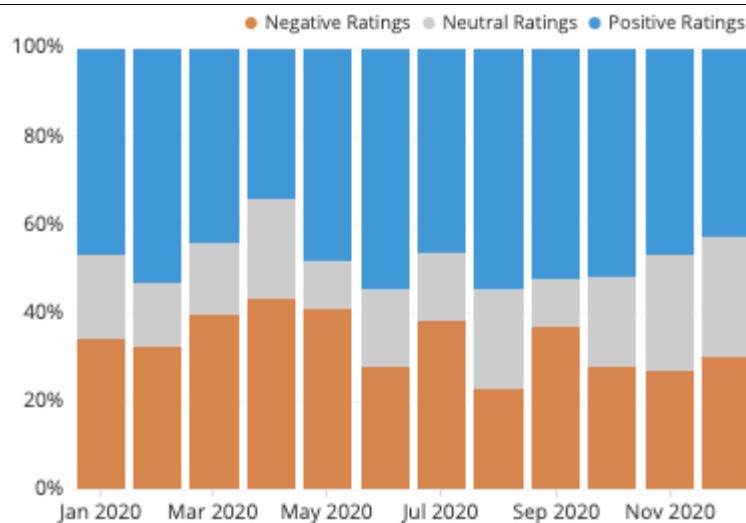
10.4.40.5.1. Horizontal stacked bar chart

Giống như Bar chart tiêu chuẩn, các thanh trong Stacked bar chart có thể được định hướng theo chiều ngang (với các danh mục chính trên trực tung) cũng như theo chiều dọc (với các danh mục chính trên trực hoành). Hướng ngang mang lại những lợi ích tương tự như trước đây, cho phép hiển thị dễ dàng các cấp độ danh mục dài mà không cần xoay hoặc cắt bớt.



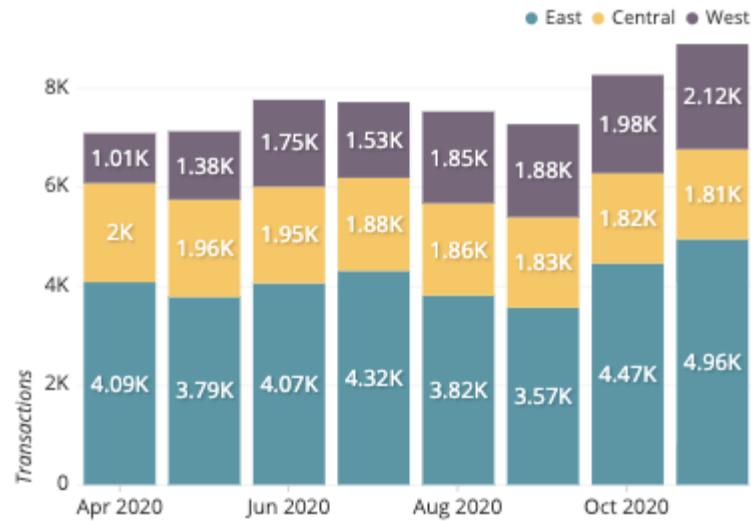
10.4.40.5.2. Stacked bar chart hiển thị giá trị theo phần trăm hoặc tần suất tương đối

Một tùy chọn phổ biến khác cho stacked bar charts là hiển thị theo tỷ lệ phần trăm hoặc tần suất tương đối. Ở đây, mỗi thanh chính được chia tỷ lệ để có cùng chiều cao, sao cho mỗi thanh phụ trở thành một giá trị phần trăm đóng góp cho tổng thể ở mỗi cấp danh mục chính. Điều này loại bỏ khả năng so sánh tổng số giữa cấp độ danh mục chính nhưng cho phép thực hiện phân tích tốt hơn về sự phân bố tương đối của các nhóm thứ cấp. Việc cố định độ cao của mỗi thanh chính giống nhau cũng tạo ra một đường cơ sở khác ở đầu biểu đồ, nơi có thể theo dõi nhau trên các thanh chính.



10.4.40.5.3. Chú thích giá trị

Một cách để giảm bớt vấn đề so sánh kích thước thanh phụ với độ dài của chúng là thêm chú thích vào từng thanh để cho biết giá trị của từng thành phần thứ cấp. Tuy nhiên, điều này làm tăng thêm sự lộn xộn về mặt hình ảnh, vì vậy hãy cẩn thận về việc có sử dụng nó hay không. Đảm bảo rằng Stacked bar charts phù hợp với mục tiêu chính của bạn về trực quan hóa hoặc cần chọn loại biểu đồ khác.

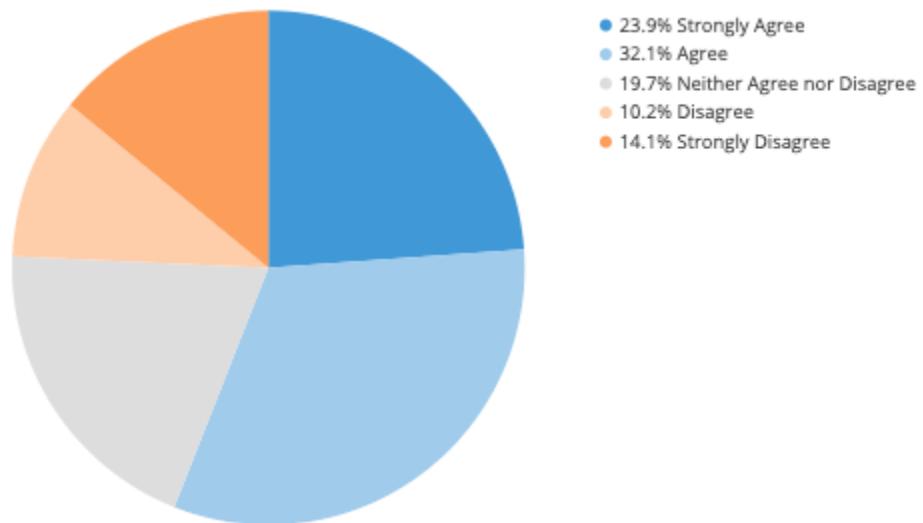


10.4.40.6. Các đồ thị liên quan

10.4.40.6.1. Pie chart

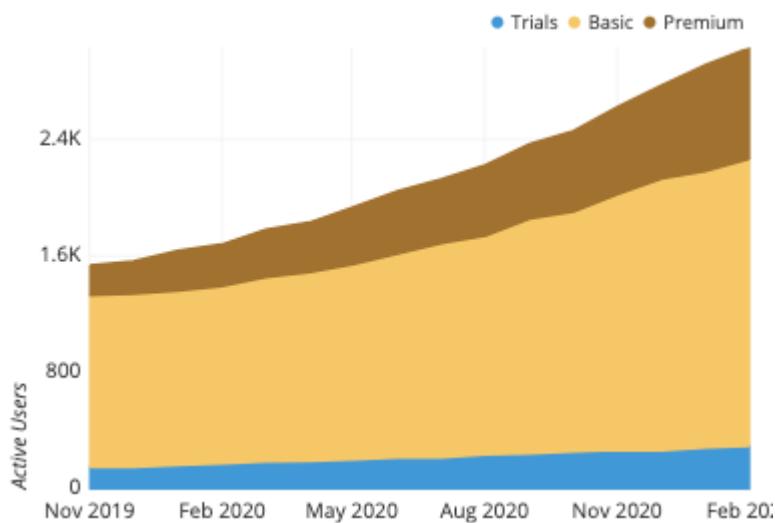
- Khi chỉ có một thanh được vẽ, Pie charts có thể được coi là một giải pháp thay thế cho Stacked bar charts.
- Một số hạn chế của Pie chart so với Stacked bar chart:
 - Không nên cố gắng sử dụng Pie chart khi muốn so sánh hai hoặc nhiều nhóm chính, như trường hợp thông thường của Stacked bar charts.
 - Vì Pie charts thường không có bất kỳ dấu tích nào nên việc đánh giá tỷ lệ chính xác cả bên trong (các thành phần thứ cấp của mỗi phần) và giữa các hình tròn có thể khó khăn hơn.

- Pie charts cũng bị giới hạn ở mức chỉ so sánh tương đối hoặc tỷ lệ phần trăm, thay vì giá trị tuyệt đối.
- Ngoài ra, nhiều Stacked bar charts cùng hiển thị bên nhau sẽ có xu hướng chiếm ít không gian hơn so với nhiều Pie charts, cho phép xem toàn bộ dữ liệu dễ dàng hơn.



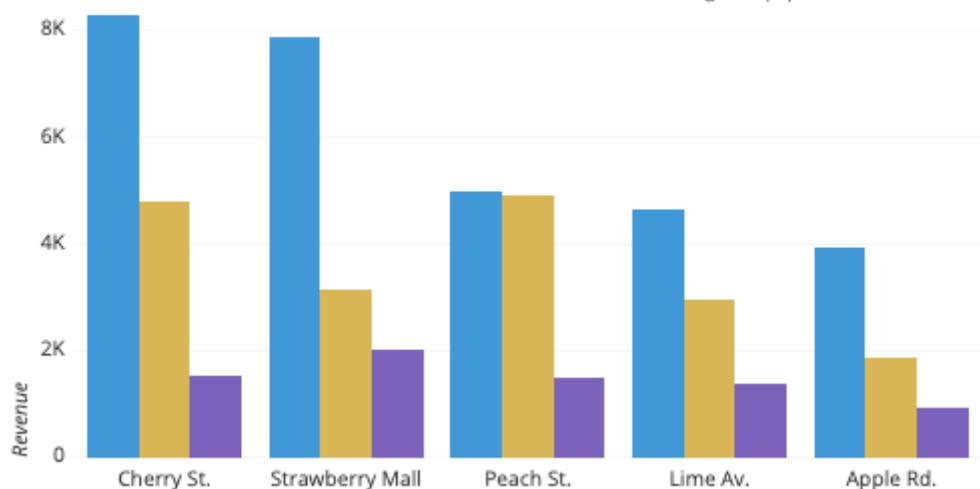
10.4.40.6.2. Area chart

Khi biến phân loại chính được lấy từ một đối tượng liên tục, chẳng hạn như các khoảng thời gian, ta có thể chọn sử dụng Stacked area chart thay vì các Stacked Bar chart. Các Stacked area chart có xu hướng nhấn mạnh những thay đổi và xu hướng hơn là những con số chính xác và sẽ dễ đọc hơn nhiều khi có nhiều thanh để vẽ. Ngoài ra, tính chất liên kết của Area chart giúp nhấn mạnh tính chất liên tục của biến chính.



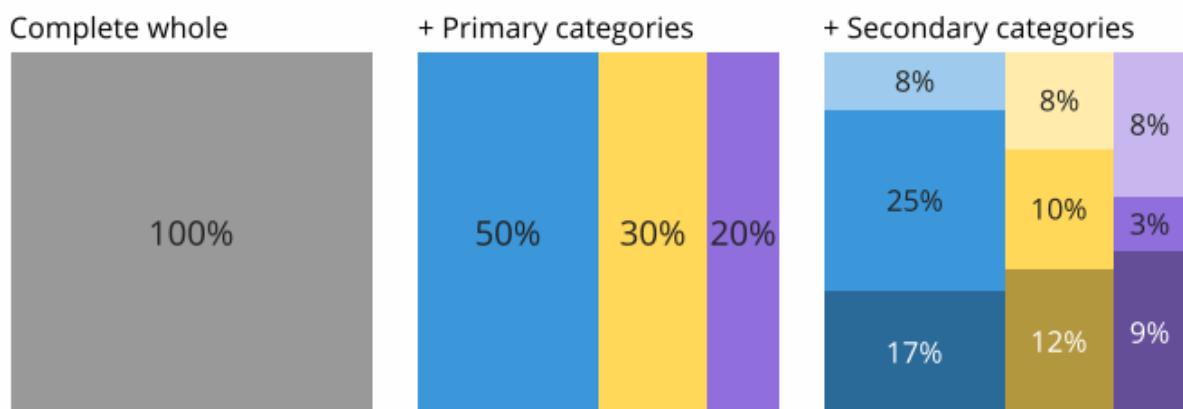
10.4.40.6.3. Grouped bar chart

Nếu bỏ ngăn xếp trên từng thanh chính và thay vào đó đặt các thanh phụ thành nhóm trên đường cơ sở thì sẽ có được Grouped bar chart, còn được gọi là clustered bar chart. Với Grouped bar chart, đã đánh đổi khả năng quan sát tổng số trong từng cấp danh mục chính và hiểu chính xác hơn về cách xếp hạng các danh mục phụ trong từng cấp danh mục chính.



10.4.40.6.4. Marimekko chart

Khi biến số đại diện cho một loại tổng thể nào đó đã được chia thành các phần của hai biến phân loại, thì loại biểu đồ khó hiểu hơn có thể chọn là biểu đồ Marimekko (còn gọi là Mekko chart, mosaic plot, matrix plot). Biểu đồ Marimekko về cơ bản là một hình vuông hoặc hình chữ nhật được chia thành Bar chart xếp chồng lên nhau theo hai hướng liên tiếp. So với Stacked bar chart có giá trị tuyệt đối tiêu chuẩn, giờ đây mỗi thanh chính sẽ có cùng chiều dài nhưng chiều rộng khác nhau. Lưu ý rằng điều này làm cho việc giải thích các thanh phụ thậm chí còn khó khăn hơn trong biểu đồ marimekko so với Stacked bar chart vì không thể nhìn vào độ dài thanh mà thay vào đó cần phải xem xét đến diện tích của các vùng hộp.



10.4.41. Sunburst Chart

Sunburst Chart: Tương tự như Pie chart nhưng có thể hiển thị dữ liệu phân cấp thông qua các vòng tròn đồng tâm.

10.4.42. Treemap chart¹

10.4.42.1. Giới thiệu

- **Treemap chart** là một dạng biểu đồ biểu diễn dữ liệu thành các hình chữ nhật, kích thước của mỗi hình chữ nhật thể hiện độ lớn của đối tượng. Trong khi các loại biểu đồ: Pie/Donut/Bar Chart chỉ hiệu quả khi biểu diễn cho dữ liệu nhỏ (10 thành phần trở lại), thì Treemaps là loại biểu đồ thay thế hiệu quả. Nếu có rất nhiều đối tượng dữ liệu

¹ <https://fastwork.vn/bieu-do-cay-treemap-chart-ung-dung-trong-bieu-dien-data-quan-ly-so-lieu-ban-hang/>

vậy muốn xem những đối tượng nào đang chiếm tỉ trọng cao, dùng Treemaps để xem điều này.

- Treemap chart là hình ảnh hóa cho dữ liệu phân cấp. Biểu đồ này bao gồm một loạt các hình chữ nhật lồng nhau có kích thước tỷ lệ với giá trị dữ liệu tương ứng. Một hình chữ nhật lớn đại diện cho một nhánh của cây dữ liệu và nó được chia thành các hình chữ nhật nhỏ hơn đại diện cho kích thước của mỗi thành phần trong nhánh đó.
- Treemap chart được thường xuyên sử dụng để hiển thị dữ liệu phân cấp. Thông tin sẽ được hiển thị dưới dạng một cụm hình chữ nhật khác nhau về kích thước và màu sắc, tùy thuộc vào giá trị dữ liệu của chúng và kích thước của hình chữ nhật mà màu sắc cũng khác nhau. Thông thường, kích thước của mỗi hình chữ nhật sẽ đại diện cho số lượng, tỷ trọng (%) trong khi màu sắc có thể đại diện cho một giá trị số hoặc một danh mục. Treemap chart có thể được sử dụng trong một không gian hạn chế mà vẫn hiển thị một số lượng lớn các mục thông tin đồng thời. Có thể chọn hiển thị toàn bộ nhánh dữ liệu ngang hàng hoặc chọn hiển thị theo nhóm. Treemap chart cũng có thể cho phép xem nhanh các xu hướng và đưa ra so sánh nhanh chóng.
- Treemap chart là một trong những tùy chọn nhỏ gọn và tiết kiệm không gian nhất để hiển thị phân cấp và cũng rất tốt để so sánh tỷ lệ giữa các danh mục thông qua kích thước của chúng. Khi có sự tương quan giữa màu sắc và kích thước trong cấu trúc cây, người dùng có thể đọc được rất nhiều ý nghĩa khác nhau của dữ liệu so với các biểu đồ khác.
- Tuy vậy, biểu đồ này không thể hiện được rõ ràng các cấp bậc từ cao nhất xuống thấp nhất.

10.4.42.2. Sử dụng

Nhờ khả năng biểu diễn đồng thời nhiều dữ liệu, Treemap chart thường xuyên được ứng dụng trong báo cáo kinh doanh thay cho các loại biểu đồ khác. Treemap chart có thể đồng thời diễn đạt được rất nhiều vấn đề như:

- Hiển thị doanh số bán hàng của từng team.
- Hiển thị doanh số bán hàng của từng nhân viên
- Xếp loại các team/các nhân viên có doanh số từ cao tới thấp
- Cho biết nhân viên/team nào đang có tỉ lệ bán hàng tốt nhất với tỷ trọng bao nhiêu %
- Tỷ trọng đóng góp vào doanh số bán hàng của từng team/nhân viên cũng được thể hiện thông qua kích thước hình vuông. Hình vuông càng lớn, tỷ trọng càng cao và ngược lại
- So sánh nhanh được mức độ chênh lệch doanh số giữa các nhân viên, giữa các team nhờ ánh xạ nhanh từ sự khác nhau giữa tỷ lệ các hình vuông
- Màu sắc đậm – nhạt cũng được hiển thị để biểu diễn dữ liệu từ cao tới thấp
- Màu sắc khác nhau có thể được dùng để phản ánh chiều hướng tốt-xấu của dữ liệu như: nhân viên nào đạt KPI cùng màu, nhân viên không đạt KPI có màu hiển thị khác...

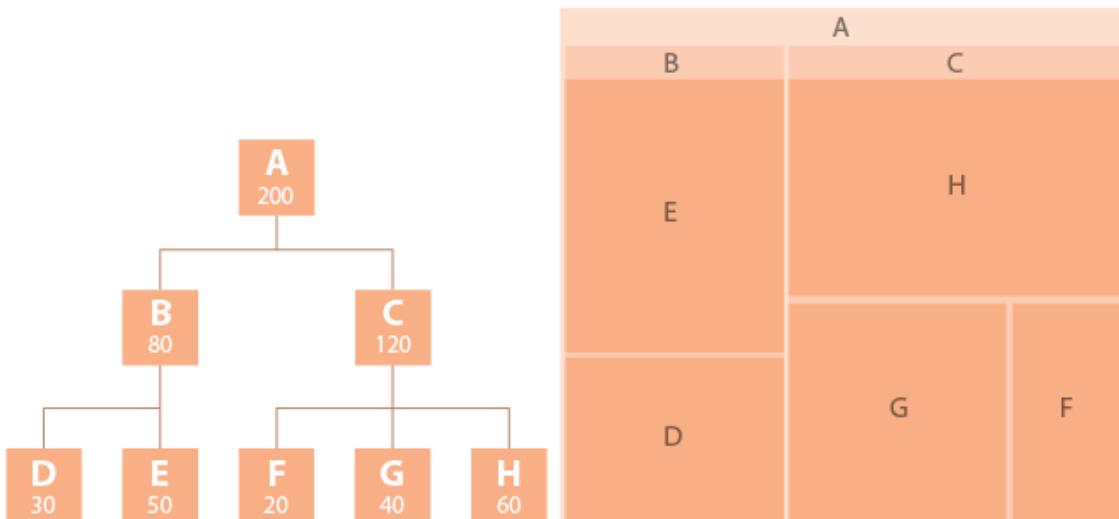
10.4.42.2.1. Các trường hợp nên dùng Treemap chart

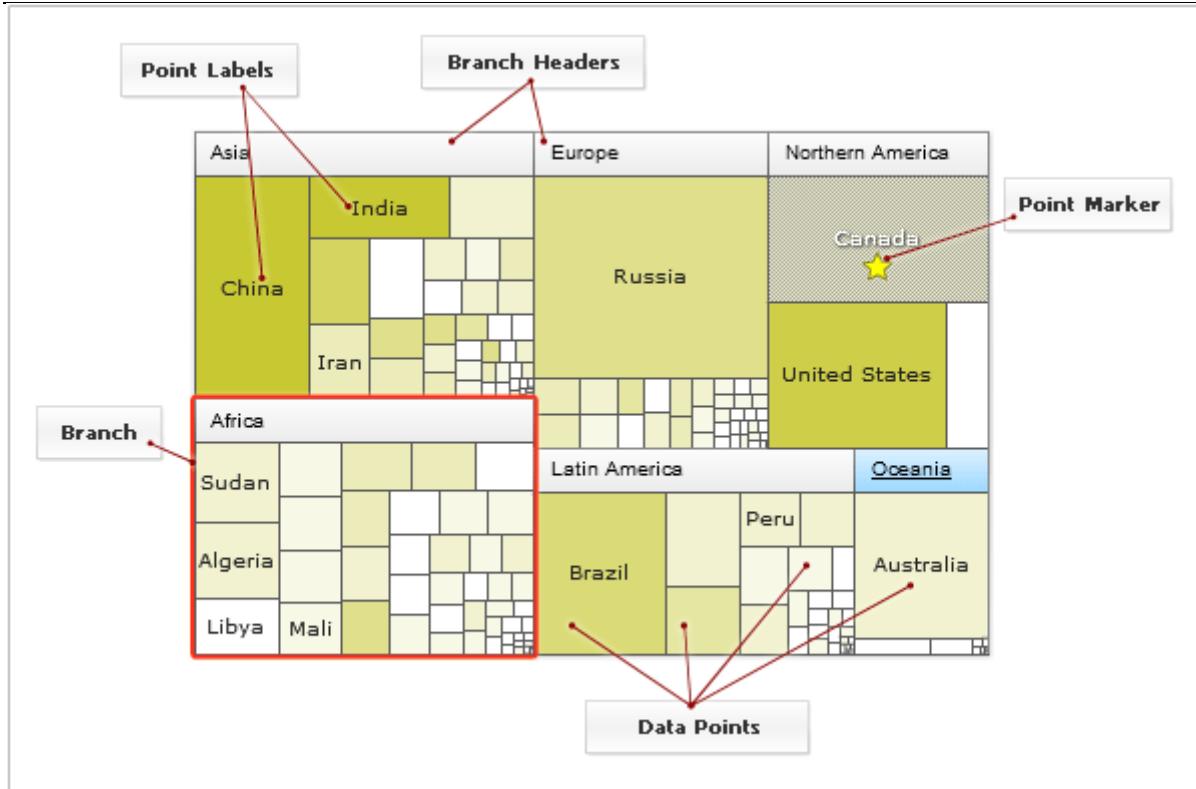
- Muốn hình dung mối quan hệ một phần đến toàn bộ giữa một số lượng lớn các danh mục => Hàng trăm nhân viên kinh doanh không thể biểu diễn bằng PIE CHART
- So sánh chính xác giữa các danh mục là không quan trọng => Giúp phát hiện so sánh nhanh những sự hiệu quả các KD chứ không phải so sánh chính xác hơn kém bao nhiêu
- Dữ liệu được phân cấp => Vừa biểu diễn dc số liệu giữa các Team KD và giữa các cá nhân
- Không gian có hạn và muốn cung cấp cho người dùng cái nhìn tổng quan về một lượng lớn dữ liệu phân cấp.
- Không thể sử dụng đồ thị thông thường, chẳng hạn như Bar chart, vì có quá nhiều mục để biểu diễn dưới dạng thanh trong một biểu đồ đơn lẻ hoặc trong một loạt biểu đồ trên một màn hình.
- Muốn cung cấp một bản tóm tắt nhanh, cấp cao về những điểm tương đồng và bất thường trong một danh mục, cũng như giữa nhiều danh mục.
- Muốn so sánh từng phần với toàn bộ.
- Muốn so sánh sơ bộ giữa các danh mục cấp cao nhất, cũng như so sánh trong các danh mục ở cấp thấp hơn.

10.4.42.2.2. Những trường hợp không nên sử dụng treemap

- Muốn so sánh định lượng chính xác. Trong trường hợp này, hãy sử dụng Bar chart để thay thế.
- Tập dữ liệu chỉ chứa một số lượng nhỏ các danh mục. Trong trường hợp này, chúng tôi khuyên nên sử dụng Bar chart.
- Có sự khác biệt lớn về độ lớn của các giá trị đo.
- Muốn hiển thị các giá trị âm. Chúng không thể được hiển thị trong treemaps.

10.4.42.3. Minh họa





10.4.42.4. Đặc tính quan trọng của Biểu đồ cây Treemap chart

10.4.42.4.1. Responsiveness – Tùy chỉnh tự động

Tính tùy chỉnh khi biểu diễn dữ liệu là một đặc điểm được người dùng đánh giá cao khi sử dụng Treemap chart. Treemap chart có thể đáp ứng được không giới hạn các nhánh dữ liệu. Khi kích thước của màn hình nhỏ hơn, các nhãn nhỏ nhất sẽ bắt đầu bị cắt bớt và ẩn đi nếu không có đủ chỗ để hiển thị cho tất cả (nếu tỷ trọng % quá nhỏ).

10.4.42.4.2. Color Palette – Dải màu phân cấp

Treemap là sự kết hợp giữa bar graph và heat map. Nếu đã làm quen với 2 dạng biểu đồ này sẽ thấy những lợi ích kép khi sử dụng Treemap chart để biểu diễn dữ liệu.

Heat map dùng để so sánh dữ liệu bằng màu sắc từ đậm tới nhạt để thể hiện giá trị từ cao tới thấp của dữ liệu. Các loại dữ liệu thường được thể hiện ở biểu đồ này có thể là các phân khúc của thị trường mục tiêu, mức độ sử dụng sản phẩm của người dùng tại nhiều khu vực. Với loại biểu đồ này, người làm cần phải thống nhất kích thước của các hình vuông/hình chữ nhật. Đặc biệt chỉ nên sử dụng một tone màu và chuyển màu sắc đậm/nhạt để thể hiện mức độ mạnh, yếu của dữ liệu.

Tương tự, khi sử dụng vào biểu đồ cây Treemap. Treemap chart hỗ trợ 2 bảng màu tuần tự và ngũ nghĩa để giải thích và biểu diễn dữ liệu.

- Sử dụng bảng màu tuần tự để biểu diễn các giá trị từ cao đến thấp bằng cách sử dụng cùng 1 màu sắc nhưng với sắc thái, cấp độ đậm nhạt khác nhau. Thông thường, màu đậm để diễn đạt các giá trị lớn, màu càng nhạt thì giá trị càng nhỏ
- Sử dụng bảng màu ngũ nghĩa để hiển thị các giá trị tốt, xấu và quan trọng. Trong trường hợp này, màu sắc không được sử dụng để biểu diễn giá trị của số liệu mà thường minh họa cho một xu hướng tăng/giảm, đánh giá mức độ tốt/xấu/... của số liệu đó.

10.4.42.4.3. 3. Legend – Chú thích

Treemap chart hỗ trợ cả chú thích giá trị và chú thích trên từng dữ liệu

- Sử dụng chú thích dữ liệu để thể hiện tên và giá trị của dữ liệu/nhãn
- Sử dụng chú thích dựa trên giá trị nếu đang sử dụng bảng màu tuần tự và muốn giải thích hệ màu được sử dụng có ý nghĩa như thế nào

10.4.42.4.4. 4. Drilldown – Phân cấp sâu

Một số ứng dụng hỗ trợ rất tốt cho người dùng khi xem Treemap chart trên máy tính, có thể tùy chọn xem tất cả danh mục hoặc xem danh mục theo nhóm bằng cách chọn một hình chữ nhật (1 nhóm) và nhấn nút Drill Down trên thanh công cụ biểu đồ.. Khi có quá nhiều data, có thể chọn cách hiển thị gộp nhóm. cũng có thể click vào từng nhóm để xem chi tiết các thành phần trong nhóm đó thay vì hiển thị toàn bộ dữ liệu trên cùng 1 màn hình.

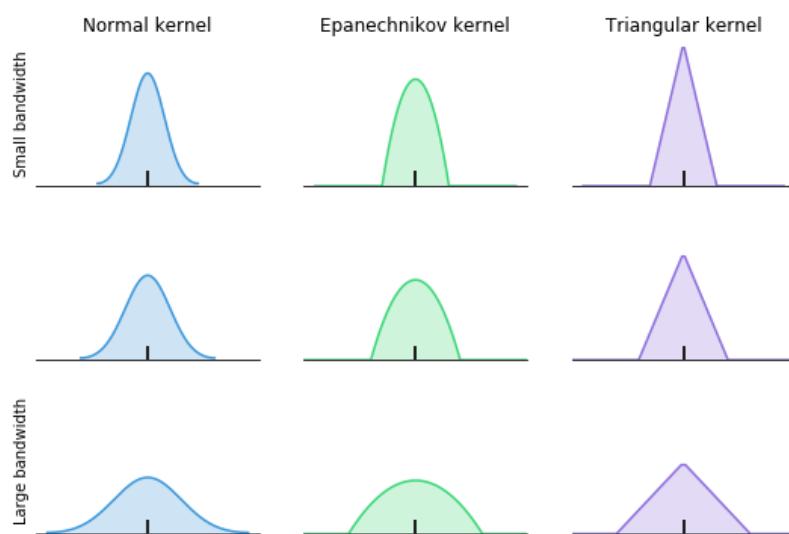
Khi người dùng click vào một hình chữ nhật đại diện cho 1 dữ liệu, tất cả các giá trị liên quan được hiển thị cũng có thể tùy chỉnh cửa sổ bật lên để hiển thị các thông tin và hoạt động khác.

10.4.43. Violin plot¹

10.4.43.1. Đường cong mật độ (density curve)

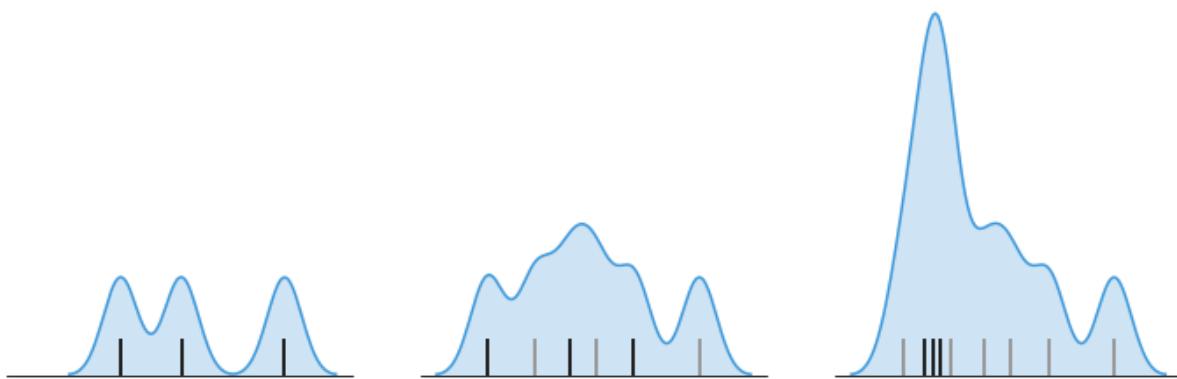
Đường cong mật độ (density curve), còn gọi là biểu đồ mật độ hạt nhân (Kernel Density Plot) hoặc Uớc tính mật độ hạt nhân (kernel density estimate - KDE), là một mô tả phân bố dữ liệu ít gấp hơn so với histogram.

Trong KDE, mỗi điểm dữ liệu đóng góp một vùng nhỏ xung quanh giá trị thực của nó. Hình dạng của vùng này được gọi là hàm kernel (the kernel function). Hạt nhân (kernel) có thể có nhiều hình dạng khác nhau, từ đường cong hình chuông nhẵn (smooth bell curves) cho đến đỉnh hình tam giác sắc nét (sharp triangular peaks). Ngoài ra, các kernels có thể có chiều rộng hoặc bao phủ khác nhau, ảnh hưởng đến mức độ ảnh hưởng của từng điểm dữ liệu riêng lẻ. Kích thước bao phủ thường được xác định bằng cách sử dụng các quy tắc toán học nhưng có thể được điều chỉnh tùy thuộc vào hình dạng và độ lệch của dữ liệu được vẽ.



¹ <https://www.atlassian.com/data/charts/violin-plot-complete-guide#:~:text=A%20violin%20plot%20depicts%20distributions,plot%2C%20to%20provide%20additional%20information.>

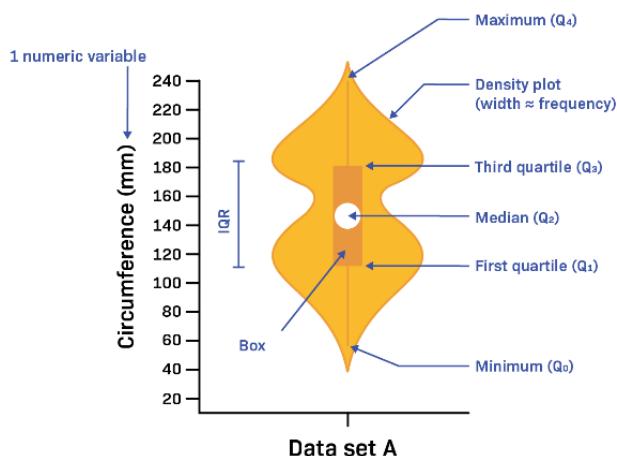
Để xây dựng density curve cuối cùng, các khu vực dành cho tất cả các điểm dữ liệu được xếp chồng lên nhau thành một tổng thể hoàn chỉnh. Mỗi điểm dữ liệu có ảnh hưởng tương đương đến phân phối cuối cùng. Vì có nhiều điểm dữ liệu hơn trong một vùng nên độ cao của density curve trong vùng đó sẽ tăng lên.



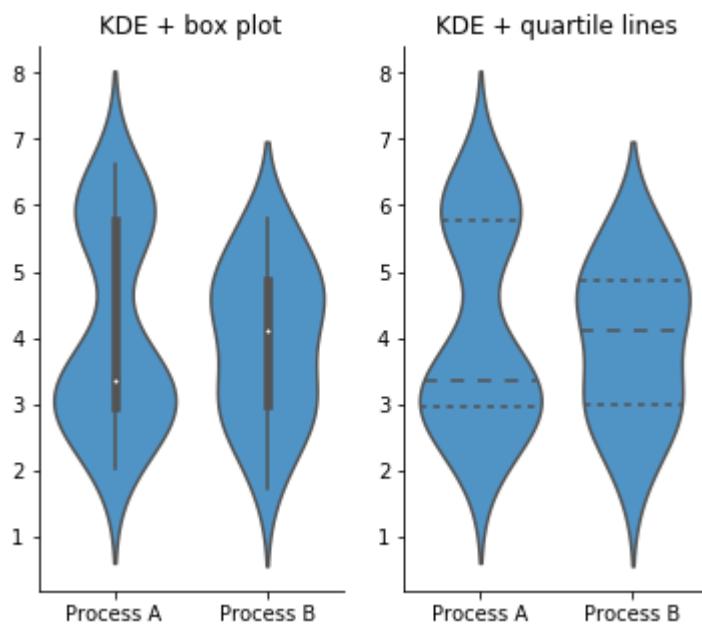
Ước tính mật độ hạt nhân (Kernel density estimation) được sử dụng tốt nhất khi có sẵn lượng dữ liệu hợp lý, dẫn đến ước tính mật độ ổn định hơn. Với ít điểm dữ liệu có sẵn, có thể dễ dàng bị nhầm lẫn bởi độ mượt của đường cong hoặc độ dài của đuôi đi qua các điểm lớn nhất và nhỏ nhất.

10.4.43.2. Giới thiệu

- Giống như box plot, violin plot được sử dụng để thể hiện sự so sánh của một phân phối thay đổi (hoặc phân phối mẫu) trên các "danh mục" khác nhau. Một violin plot có nhiều thông tin hơn một box plot đơn thuần vì violin plot hiển thị toàn bộ phân phối dữ liệu.

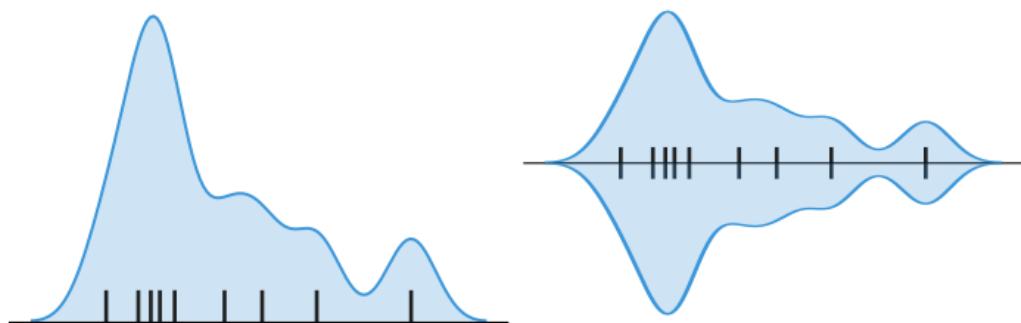


- Violin plot mô tả sự phân bố dữ liệu số cho một hoặc nhiều nhóm bằng cách sử dụng các đường cong mật độ (density curves). Độ rộng của mỗi đường cong tương ứng với tần suất gần đúng của các điểm dữ liệu trong từng vùng. Mật độ thường đi kèm với một loại biểu đồ chồng lên nhau, chẳng hạn như box plot, để cung cấp thêm thông tin.



Minh họa về violin plot ở trên mô tả kết quả của một thí nghiệm với một nhóm đối chứng và hai điều kiện thí nghiệm. Ở giữa mỗi density curve là một boxplot nhỏ, trong đó hình chữ nhật hiển thị các điểm cuối của tứ phân vị thứ nhất và thứ ba và dấu chấm ở giữa là điểm trung vị (median). Từ đồ thị, có thể thấy rằng hai kỹ thuật thử nghiệm mang lại những lợi ích khác nhau so với đối chứng. Tuy nhiên, điều kiện thí nghiệm thứ hai (B) có phân bố kéo dài hơn nhiều so với hai nhóm còn lại, không có đỉnh rõ rệt. Thực tế thứ hai có thể đã bị bỏ qua chỉ với box plot.

- Trong Violin plot, các density curve riêng lẻ được xây dựng xung quanh các đường trung tâm, thay vì xếp chồng lên nhau trên các đường cơ sở. Ngoài sự khác biệt về kiểu hiển thị này, các đường cong trong Violin plot tuân theo cách xây dựng và diễn giải giống nhau.



10.4.43.3. Sử dụng

10.4.43.3.1. Các trường hợp thường dùng

- Violin plot được sử dụng khi muốn quan sát sự phân bố của dữ liệu số và đặc biệt hữu ích khi muốn so sánh sự phân bố giữa nhiều nhóm. Các đỉnh, đáy và đuôi của density curve của mỗi nhóm có thể được so sánh để xem các nhóm giống nhau hay khác nhau ở đâu. Các yếu tố bổ sung, như tứ phân vị boxplot, thường được thêm vào Violin plot để cung cấp thêm các cách so sánh các nhóm.

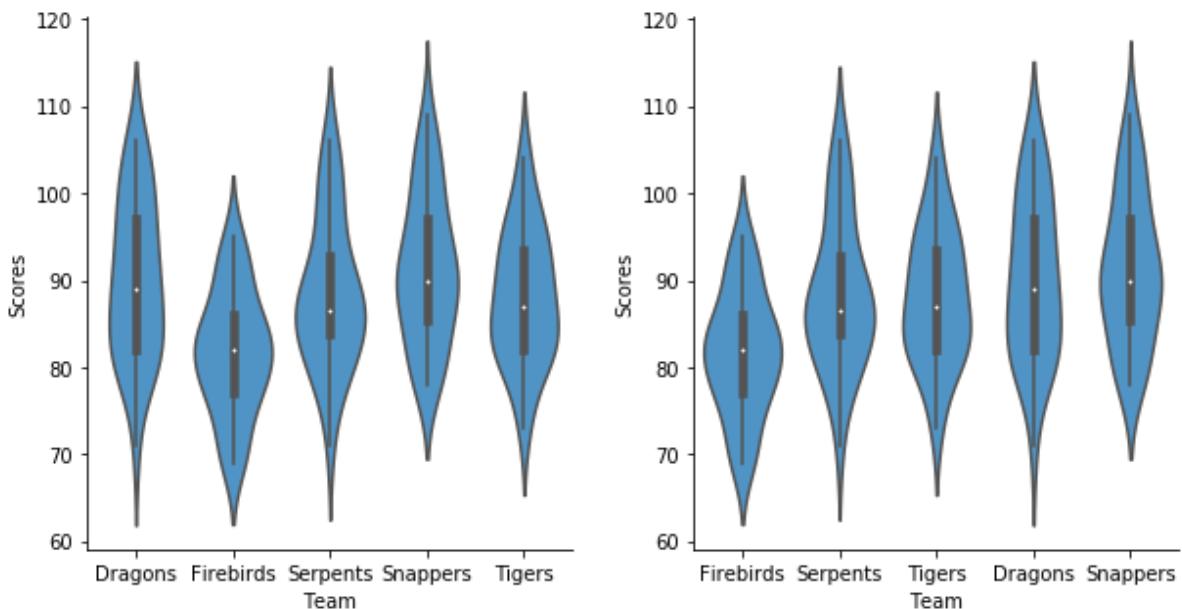
10.4.43.3.2. Sử dụng violin plot hiệu quả

- Ví dụ về cấu trúc dữ liệu sử dụng khi vẽ violin plot:

CONDITION	SCORE
Control	30
Exp. A	33
Exp. B	25
Exp. A	36
...	...

Cách phổ biến nhất để cấu trúc dữ liệu để tạo Violin plot là thông qua một bảng có hai cột. Mỗi hàng tương ứng với một điểm dữ liệu duy nhất, trong khi các giá trị ô biểu thị tư cách thành viên nhóm và giá trị số cho mỗi điểm. Tất cả các đặc điểm đặc trưng sẽ được tính toán tự động từ dữ liệu đầu vào thô này. Nếu tất cả dữ liệu nằm trong một nhóm duy nhất thì cột chỉ tên nhóm của các thành viên nhóm sẽ không cần thiết.

- *Xem xét thứ tự của các nhóm:* Khi các nhóm trong violin plot không có thứ tự cố hữu, có thể thay đổi thứ tự trong biểu đồ của các nhóm để dễ dàng hiểu rõ hơn về dữ liệu. Ví dụ: việc sắp xếp các nhóm theo giá trị trung bình giúp cho việc xếp hạng các nhóm được thể hiện rõ ràng ngay lập tức.



10.4.43.3.3. Mức độ phổ biến của violin plot

Các violin plot ít phổ biến hơn các biểu đồ khác như box plot do việc thiết lập hạt nhân và bảng thông phức tạp hơn. Chúng cũng có thể gây nhiều về mặt thị giác, đặc biệt là với loại biểu đồ được xếp chồng lên nhau. Nếu cần một biểu đồ để trình bày những phát hiện cho người xem chưa quen với violin plot, thì tốt hơn nên sử dụng một hình ảnh đơn giản và trực quan hơn như box plot.

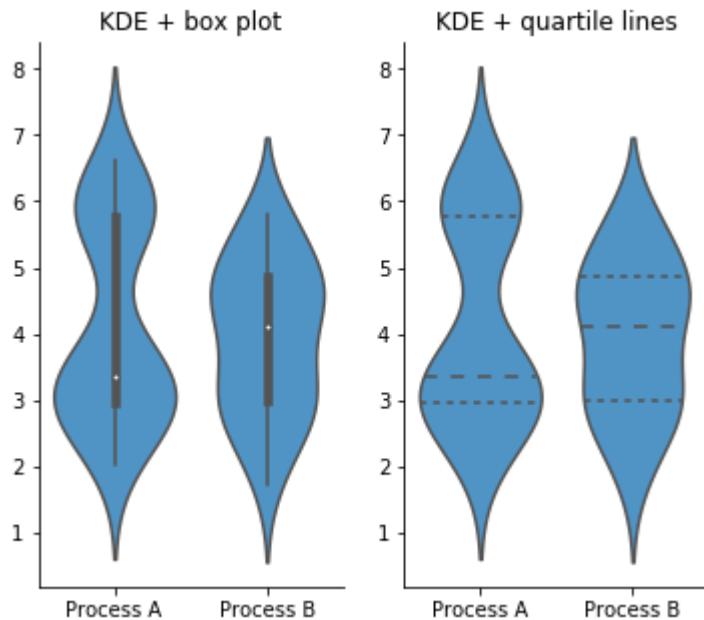
10.4.43.4. Các tùy chọn phổ biến được dùng với violin plot

10.4.43.4.1. Bổ sung lớp phủ lên violin plot với biểu đồ bô sung

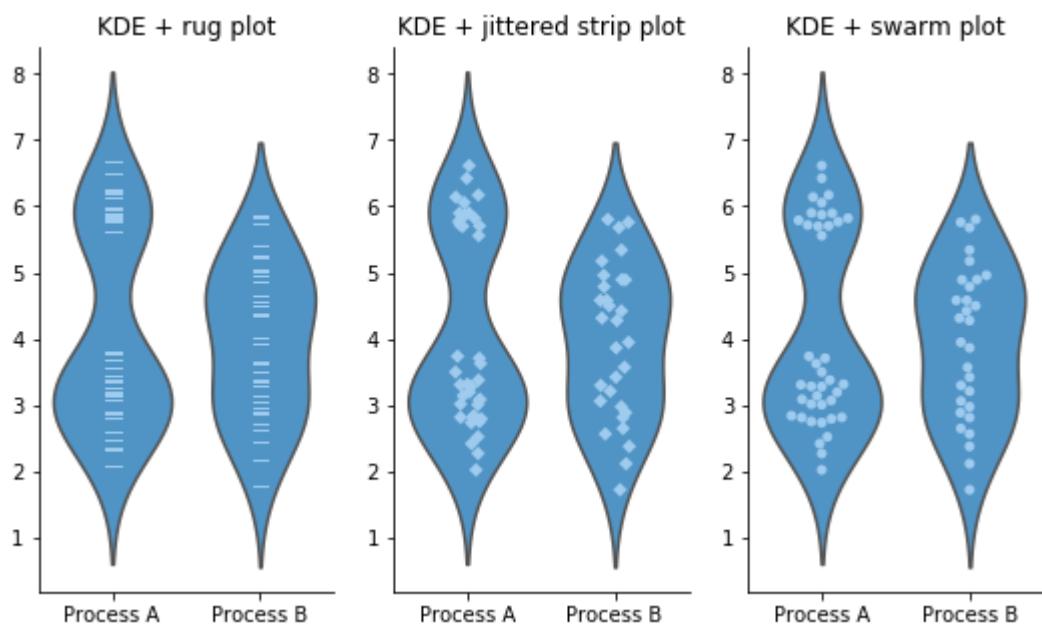
- Bản thân các violin plot thực sự có thể khá hạn chế. Nếu tính đối xứng, độ lệch hoặc hình dạng khác và các đặc điểm biến đổi khác nhau giữa các nhóm thì có thể khó

so sánh chính xác các density curve giữa các nhóm. Vì lý do này mà các Violin plot thường được hiển thị bằng một loại biểu đồ phủ khác.

- Phần bô sung phổ biến nhất cho violin plot là box plot. Thông thường, phần bô sung này được giả định theo mặc định; violin plot đôi khi được mô tả là sự kết hợp giữa KDE và box plot. Trong một số trường hợp nhất định, chỉ một tập hợp con các đặc điểm của box plot sẽ được vẽ để giảm nhiễu hình ảnh, chẳng hạn như ba dòng biểu thị vị trí tứ phân vị mà không có đường kẻ.



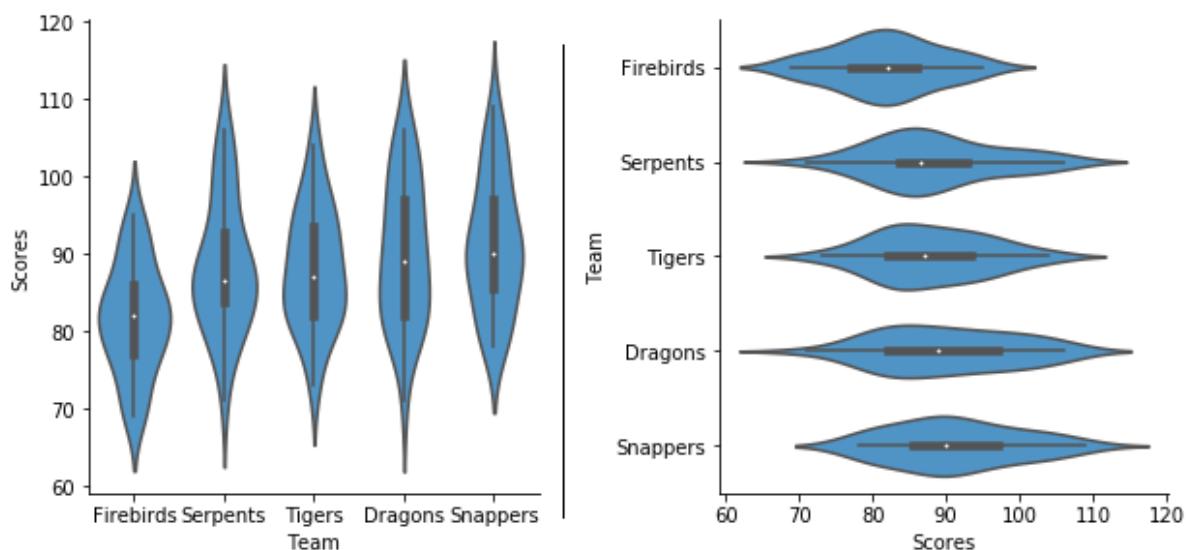
- Có những biểu đồ phân phối khác có thể được phủ lên thay vì Box plot.
 - Rug plot (biểu đồ thảm) hoặc strip plot (biểu đồ dài) thêm mọi điểm dữ liệu vào đường trung tâm dưới dạng dấu tích (tick mark) hoặc dấu chấm (dot), giống như 1-d scatter plot (biểu đồ phân tán 1-d).



- Swarm plot (biểu đồ bầy đàn) sẽ bù đắp các điểm dữ liệu từ đường trung tâm để tránh chồng chéo. Một chiến lược thay thế là tạo ngẫu nhiên các điểm dao động từ đường trung tâm; mặc dù nó không đảm bảo tránh được sự chồng chéo.
- Các lớp phủ biểu đồ thay thế này được sử dụng tốt nhất khi có số lượng điểm dữ liệu từ thấp đến trung bình trong mỗi nhóm. Mặc dù việc hiển thị các điểm dữ liệu riêng lẻ có thể làm rõ cách tạo ra các density curve và tiết lộ thông tin về quy mô nhóm thường không được thể hiện rõ ràng trong violin plot, nhưng sự hiện diện của chúng sẽ làm tăng thêm nhiều trên biểu đồ và có thể gây mất tập trung. Ngoài ra, khi quy mô nhóm đủ lớn, ước tính phân bố từ đường cong mật độ (density curve) và box plot sẽ đủ ổn định để cung cấp những hiểu biết hợp lý.

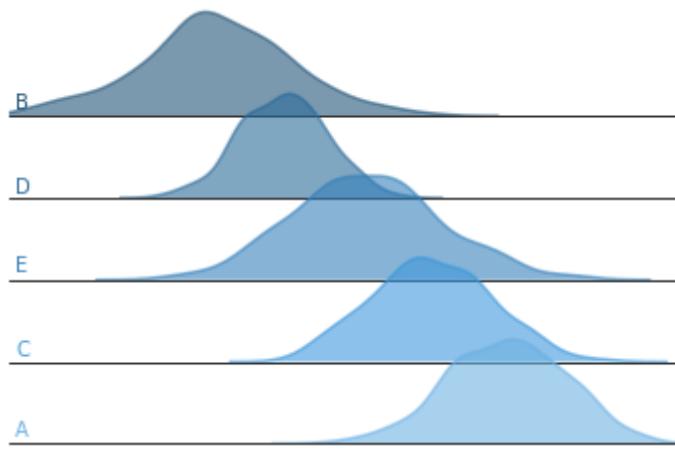
10.4.43.4.2. Violin plot dọc và ngang

- Violin plot có thể được định hướng bằng đường cong mật độ dọc (vertical density curves) hoặc đường cong mật độ ngang (horizontal density curves).
- Violin plot theo chiều ngang là lựa chọn tốt khi bạn cần hiển thị tên nhóm dài hoặc khi có nhiều nhóm để biểu thị.
- Việc mở rộng một biểu đồ trên trực thăng đứng thường dễ dàng hơn so với chiều ngang của nó; điều này rất quan trọng khi cần có đủ không gian để quan sát rõ ràng hình dạng của density curve.



10.4.43.4.3. Ridgeline plot

Một cách khác để so sánh sự phân bố giữa các nhóm bằng cách sử dụng density curve là sử dụng Ridgeline plot. Ridgeline plot bao gồm một chồng các density curve đều đặn theo chiều dọc. Thông thường, các đường cong sẽ được sắp xếp chồng chéo nhẹ, có thể tiết kiệm không gian so với việc tách hoàn toàn các trực. Sự chồng chéo này có nghĩa là các density curve có xu hướng được vẽ mà không có bất kỳ lớp phủ bổ sung nào. Ridgeline plot được sử dụng tốt nhất khi có một mẫu rõ ràng trong dữ liệu giữa các nhóm.

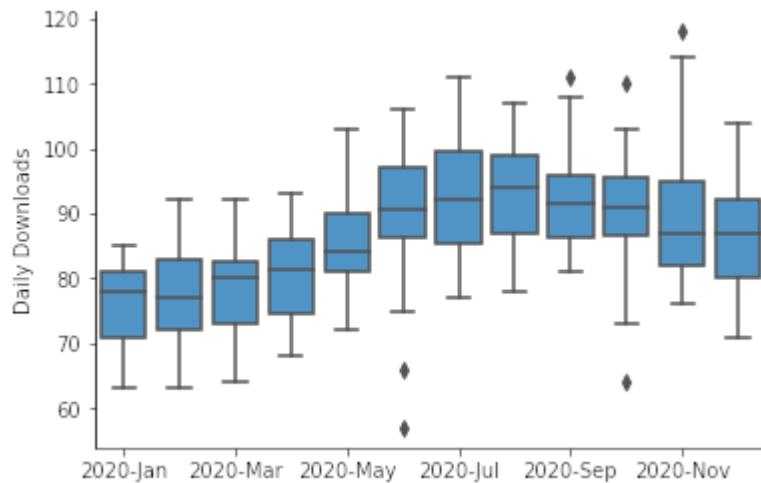


10.4.43.5. Các đồ thị liên quan

10.4.43.5.1. Density curves và boxplot

Như đã lưu ý trước đó, violin plot thường được thể hiện dưới dạng một loạt các density curve, hộp (boxes) và râu (whiskers) chồng lên nhau. Box plot bị hạn chế về lượng thông tin chúng có thể truyền tải, nhưng chúng dễ diễn giải hơn nhiều, đặc biệt là trong việc so sánh giữa các nhóm. Các density curve đều nhằm mô tả các chi tiết phân bố nhưng về mặt trực quan khó diễn giải hơn và nhiều hơn. Nhưng được kết hợp trong violin plot, cả hai bổ sung cho nhau để đạt được hiệu quả tốt nhất của cả hai loại biểu đồ.

Điều đó nói lên rằng, có những tình huống mà việc chỉ tạo ra một box plot là nổi bật. Nếu có nhiều nhóm để vẽ đồ thị thì tính đơn giản của box plot có thể mang lại lợi ích lớn. Bất kỳ box (hộp) và râu (whiskers) riêng lẻ nào cũng cần ít không gian hơn để có thể đọc được so với density curve. Khi không gian là vấn đề được quan tâm hoặc việc thể hiện bao gồm tất cả thông kê có tầm quan trọng hàng đầu thì boxplot có thể thích hợp hơn violin plot.

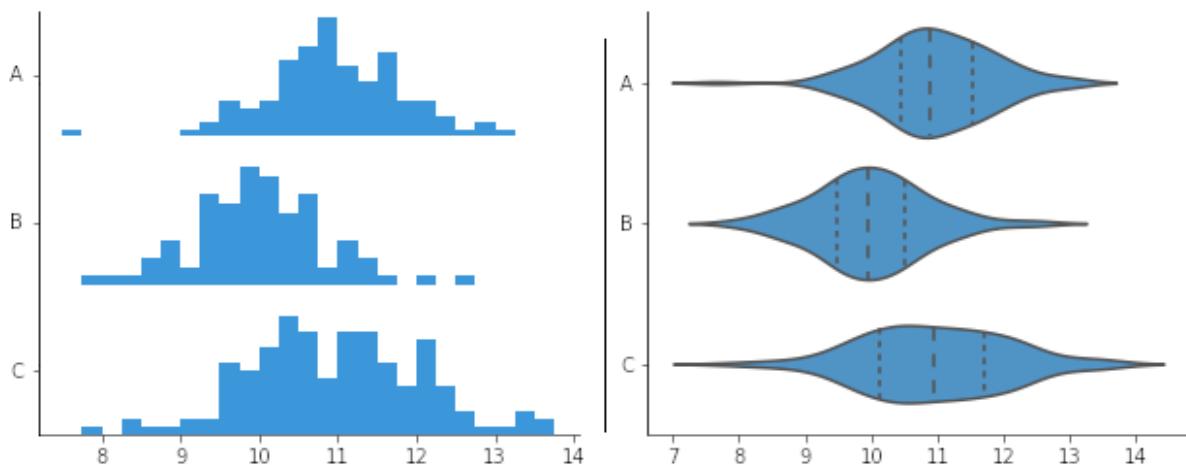


10.4.43.5.2. Histogram

So với density curves, Histogram là loại biểu đồ được biết đến rộng rãi hơn để mô tả sự phân bố. Trong khi việc thiết lập KDE đòi hỏi phải lo lắng về hình dạng hạt nhân (kernel shape) và băng thông (bandwidth), việc tạo Histogram đòi hỏi phải xem xét kích

thước thùng (bin sizes) và vị trí các cạnh sẽ được căn chỉnh. Đối với cả hai loại biểu đồ, việc lựa chọn các tham số này có thể ảnh hưởng đến giao diện của biểu đồ cuối cùng.

Nói chung, histogram được hiển thị theo chiều ngang với đường cơ sở phía dưới (bottom baseline). Có thể xây dựng violin plot bằng cách sử dụng histogram căn giữa (a center-aligned histogram) thay vì KDE cho phần thân chính, nhưng điều này có xu hướng yêu cầu bố cục tùy chỉnh của các yếu tố trực quan.

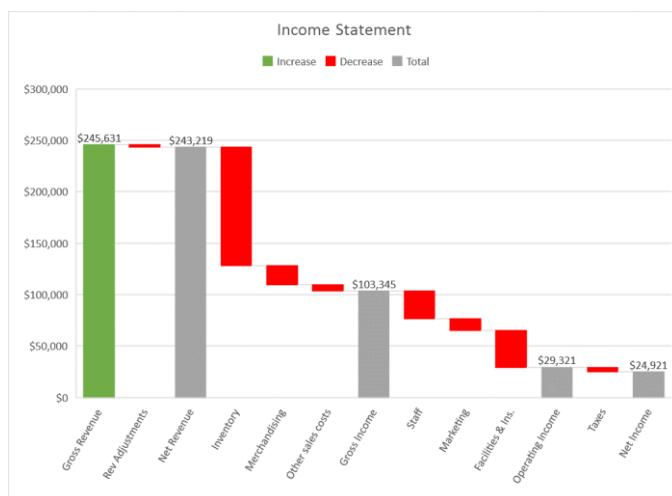


10.4.44. Waterfall chart (đồ thị thác nước)

10.4.44.1. Giới thiệu

Biểu đồ thác nước phản ánh sự khác biệt theo thời gian. Chúng chứng minh cả tác động tích cực và tiêu cực của các yếu tố khác nhau lên giá trị ban đầu, chẳng hạn như số dư đầu kỳ. Biểu đồ thác nước rất hữu ích khi minh họa báo cáo tài chính, phân tích lãi lỗ và so sánh thu nhập. Có thể sử dụng biểu đồ này để làm nổi bật ngân sách so với số tiền chi tiêu. Các giá trị dương và âm thường tuân theo môt màu để hiển thị giá trị tăng hoặc giảm như thế nào do một loạt thay đổi theo thời gian.

10.4.44.2. Minh họa





10.4.45. World cloud chart

10.4.45.1. Giới thiệu

- Còn được biết đến với tên gọi Tag Cloud, được hiểu là một hình ảnh trực quan, thể hiện các từ phổ biến nhất xuất hiện trong văn bản. Được sử dụng để hình dung mối quan hệ giữa các từ khác nhau hoặc để nắm bắt xu hướng về các từ phổ biến nhất.
- Màu sắc được sử dụng trên Word Clouds chủ yếu mang tính thẩm mỹ, nhưng nó cũng có thể được sử dụng để phân loại các từ hoặc để hiển thị một biến dữ liệu khác.
- Thông thường, Word Clouds được sử dụng trên các trang web hoặc blog để mô tả việc sử dụng từ khóa hoặc thẻ. Word Clouds cũng có thể được sử dụng để so sánh hai phần khác nhau của văn bản với nhau.
- Mặc dù đơn giản và dễ hiểu, nhưng Word Clouds có một số nhược điểm như:
 - Các từ dài được nhấn mạnh hơn các từ ngắn.
 - Những từ có chữ cái có thể nhận được nhiều sự chú ý hơn.

10.4.45.2. Minh họa

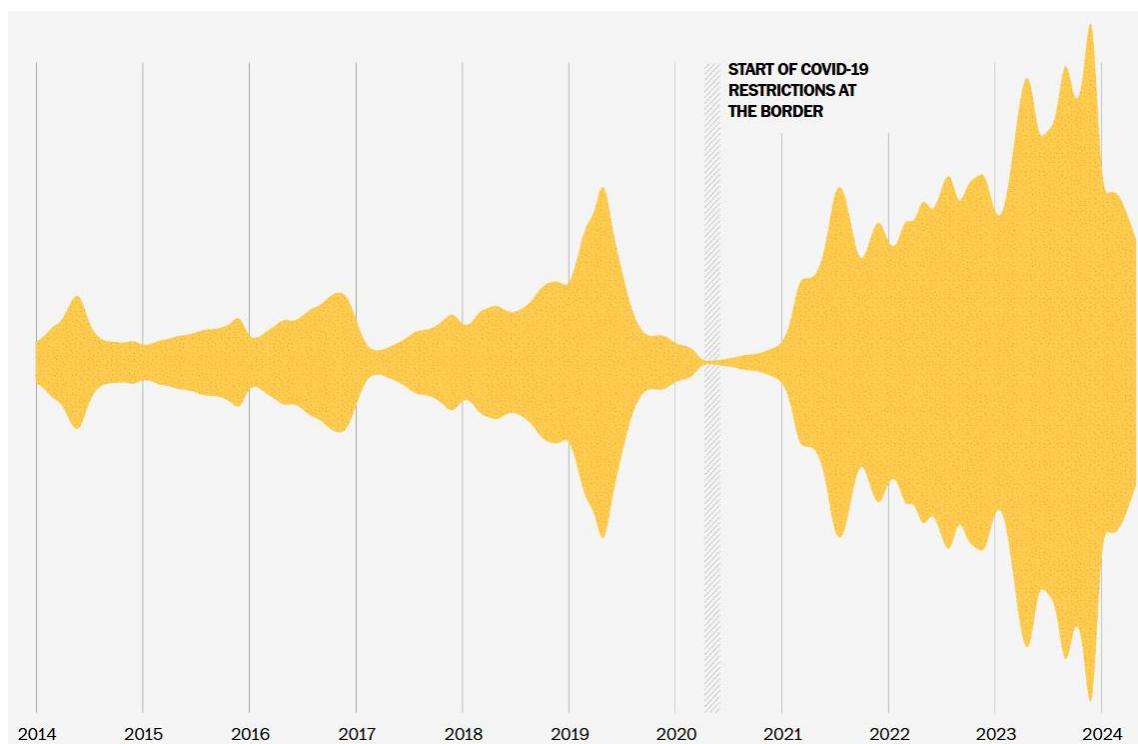
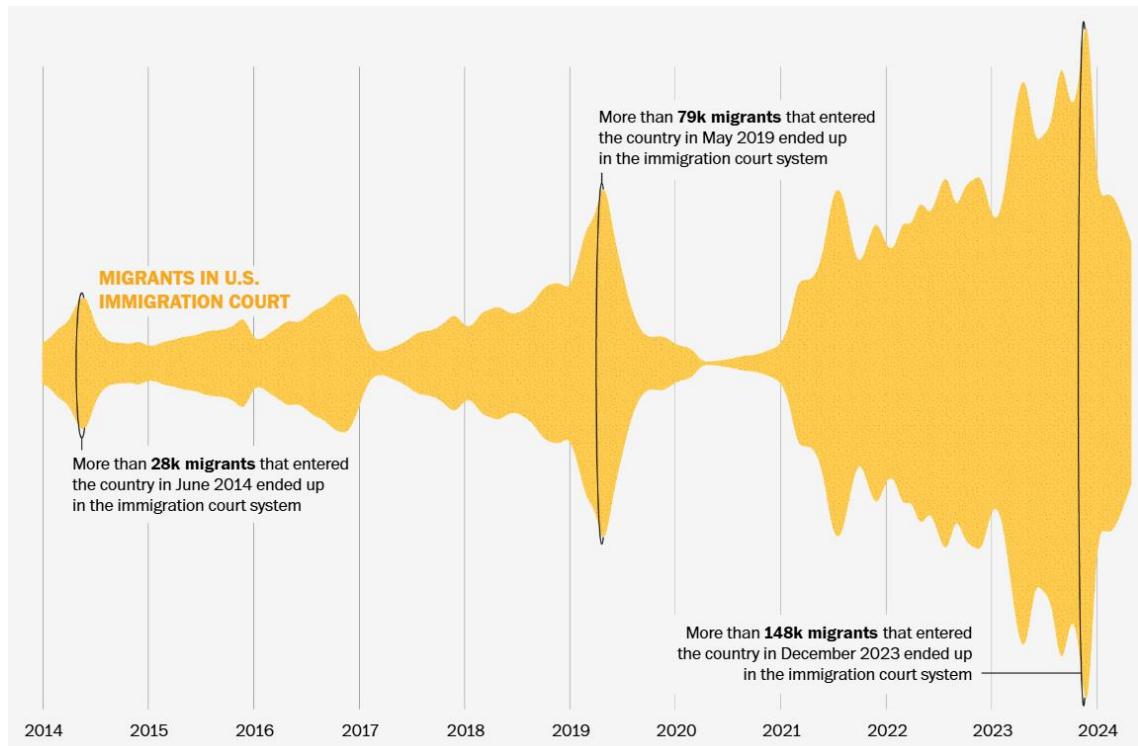


10.4.46. streamgraph¹

- Đây là biểu đồ luồng (streamgraph), một loại biểu đồ vùng xếp chồng (stacked area chart) được sử dụng để hiển thị các thay đổi về số lượng theo thời gian cho nhiều danh mục.
- Minh họa:

¹ <https://flourish.studio/blog/streamgraphs/>

- 4,1 triệu người di cư: Họ đến từ đâu, họ sống ở đâu ở Hoa Kỳ¹.
- Trong biểu đồ này, dữ liệu thể hiện các luồng di cư từ Guatemala, Honduras và El Salvador theo thời gian. Mỗi màu đại diện cho một quốc gia và chiều rộng của mỗi phần tại một thời điểm nhất định cho thấy lượng người di cư từ quốc gia đó. Biểu đồ nêu bật các sự kiện đáng chú ý, chẳng hạn như đợt di cư cao điểm vào tháng 5 năm 2019 và tác động của các hạn chế liên quan đến COVID-19 bắt đầu từ đầu năm 2020.



¹ [Where millions of immigrants in the U.S. came from and now live - Washington Post](#)

