Những sai lầm trực quan hóa dữ liệu bạn nên tránh

Trực quan hóa dữ liệu đúng cách là rất quan trọng cho việc phân tích và giải thích. Có một số lỗi trực quan hóa dữ liệu phổ biến mà người mới bắt đầu và các chuyên gia khoa học dữ liệu thường mắc phải có thể cản trở việc giải thích, tính toàn vẹn và hiệu quả của dữ liệu được trực quan hóa. Vì vậy, nếu bạn muốn biết về một số sai lầm bạn nên tránh để cải thiện kỹ năng trực quan hóa dữ liệu của mình thì bài viết này là dành cho bạn. Trong bài viết này, tôi sẽ hướng dẫn bạn một số lỗi trực quan hóa dữ liệu phổ biến mà bạn nên tránh.

Những sai lầm trực quan hóa dữ liệu bạn nên tránh

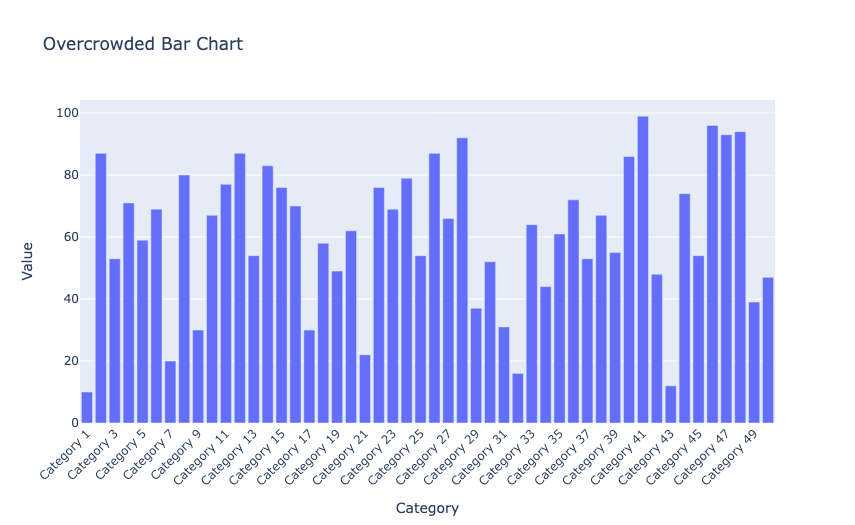
Dưới đây là một số lỗi trực quan hóa dữ liệu mà nhiều người mới bắt đầu và các chuyên gia thường mắc phải khi trực quan hóa dữ liệu:

1. Quá nhiều trong biểu đồ thanh (Bar Chart)
2. Thiếu đường xu hướng (TrendLines) trong biểu đồ phân tán (Scatter Plots).
3. Sử dụng quá nhiều lát (Slices) trong biểu đồ hình tròn (Pie Charts)
4. Kích thước thùng (Bin) không phù hợp trong biểu đồ Histogram
5. Bỏ các ngoại lệ trong biểu đồ hộp (Box Plots).

Chúng ta hãy xem xét chi tiết những lỗi trực quan hóa dữ liệu này và cách tránh chúng.

## **Quá nhiều trong biểu đồ thanh**

Tình trạng quá tải xảy ra khi có quá nhiều thanh bị dồn vào biểu đồ thanh, gây khó khăn cho việc phân biệt giữa các thanh riêng lẻ, nhãn đọc hoặc các mẫu phân biệt. Nó thường xảy ra khi cố gắng hiển thị quá nhiều thông tin trong một biểu đồ. Nó dẫn đến hình ảnh lộn xộn, trong đó thông điệp của dữ liệu bị che khuất và các chi tiết quan trọng có thể bị bỏ qua.

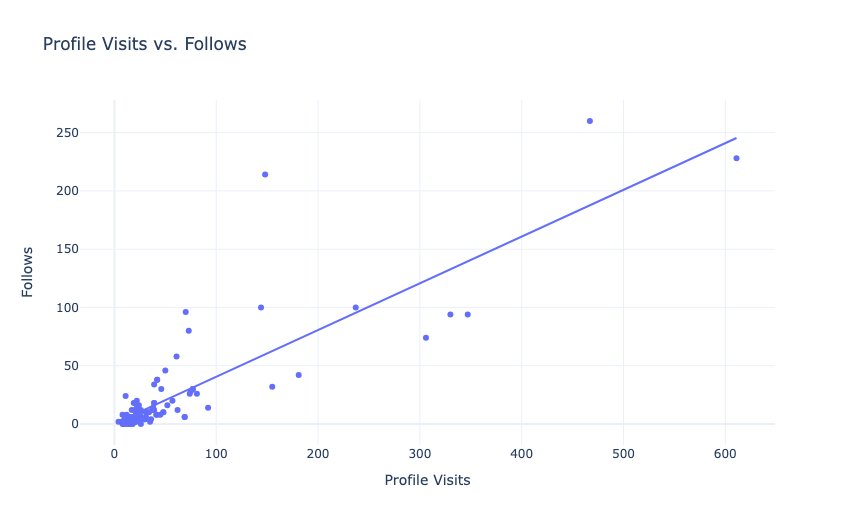


Để tránh tình trạng quá tải trong biểu đồ thanh:

1. Tập trung vào các điểm dữ liệu quan trọng nhất. Nếu tất cả các điểm dữ liệu đều cần thiết, hãy cân nhắc việc chia biểu đồ thành nhiều biểu đồ nhỏ hơn.
2. Điều chỉnh khoảng cách giữa các thanh và sử dụng bố cục theo chiều ngang nếu nó giúp phù hợp với các nhãn dài hơn.
3. Để so sánh nhiều danh mục, hãy cân nhắc sử dụng biểu đồ đường (Line Chart), biểu đồ chấm (Dot Plot) hoặc biểu đồ thanh xếp hạng (Rank Bar Chart.).

## **Thiếu đường xu hướng trong biểu đồ phân tán**

Bỏ qua các đường xu hướng khỏi biểu đồ phân tán khi khám phá mối quan hệ giữa hai biến là một lỗi trực quan hóa dữ liệu khác mà bạn nên tránh. Đường xu hướng giúp tóm tắt mô hình cơ bản hoặc mối quan hệ giữa các biến số, giúp bạn dễ hiểu hơn trong nháy mắt. Nếu không có đường xu hướng, việc xác định bất kỳ mối tương quan, xu hướng hoặc ngoại lệ nào trong biểu đồ phân tán có thể gặp khó khăn, làm giảm giá trị phân tích của nó.

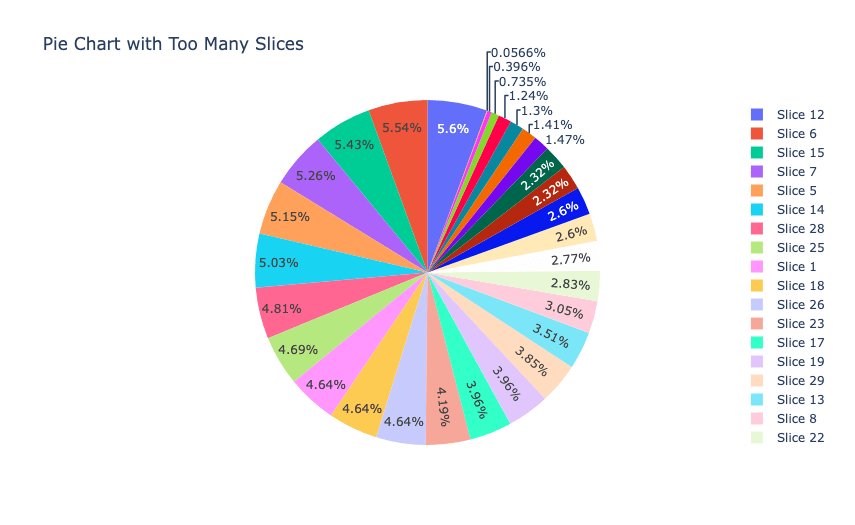


Vì vậy hãy luôn đảm bảo:

1. Thêm một đường xu hướng để làm nổi bật hướng chung hoặc mối tương quan giữa hai biến số.
2. Việc thêm đường xu hướng sẽ có ý nghĩa đối với dữ liệu của bạn và không đơn giản hóa quá mức hoặc trình bày sai mối quan hệ.

## **Sử dụng quá nhiều lát trong biểu đồ hình tròn**

Tạo biểu đồ hình tròn có quá nhiều lát, đặc biệt khi nhiều lát có kích thước tương tự nhau là một lỗi trực quan hóa dữ liệu khác mà bạn nên tránh. Việc phân biệt giữa các lát cắt, đọc nhãn hoặc trích xuất những hiểu biết có ý nghĩa trở nên khó khăn. Biểu đồ hình tròn trở nên quá tải về mặt trực quan và sự khác biệt giữa các danh mục không rõ ràng, điều này khiến việc truyền đạt dữ liệu trở nên kém hiệu quả hơn.



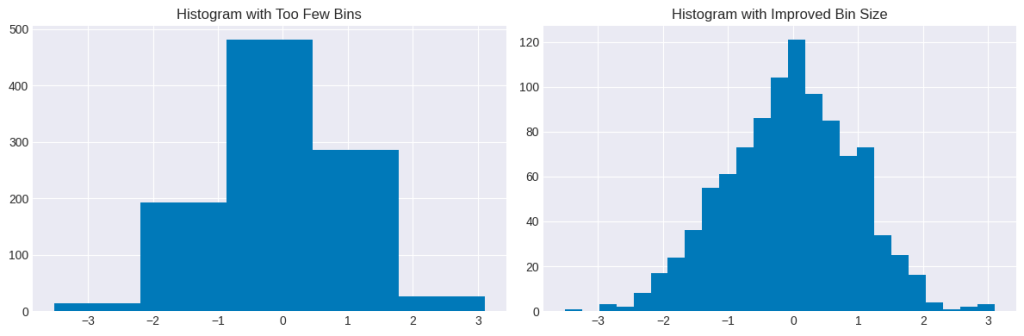
Vì vậy, để tránh sai lầm này:

1. Chỉ sử dụng biểu đồ hình tròn khi có một vài danh mục riêng biệt. Một nguyên tắc nhỏ là giới hạn ở khoảng 5-7 lát.
2. Hãy xem xét biểu đồ thanh hoặc biểu đồ thanh xếp chồng, có thể xử lý hiệu quả nhiều danh mục hơn và giúp việc so sánh dễ dàng hơn so với biểu đồ hình tròn.
3. Hoặc nhóm các danh mục nhỏ hơn thành danh mục “Khác” để giảm số lượng lát.

## **Kích thước thùng không phù hợp trong biểu đồ Histogram**

Kích thước thùng đề cập đến độ rộng của các khoảng mà dữ liệu được nhóm. Mỗi thùng đại diện cho một phạm vi giá trị và dữ liệu trong mỗi phạm vi được tính và biểu thị dưới dạng thanh trong biểu đồ. Việc chọn kích thước thùng quá lớn hoặc quá nhỏ cho biểu đồ có thể làm sai lệch đáng kể nhận thức về phân phối dữ liệu. Các ngăn quá lớn có thể đơn giản hóa dữ liệu quá mức, che giấu các biến thể, trong khi các ngăn quá nhỏ có thể làm dữ liệu trở nên phức tạp quá mức, tạo ra nhiễu.

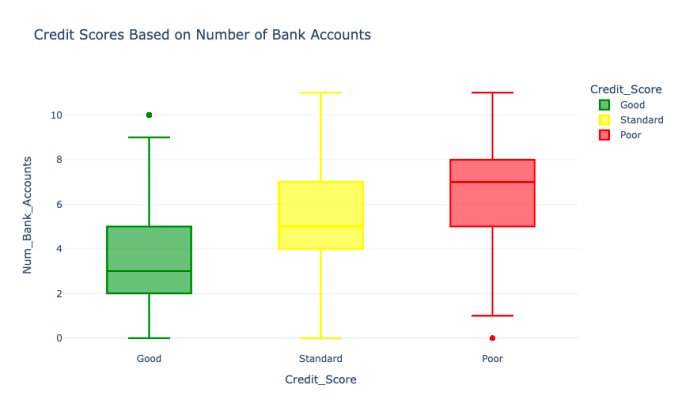
Nó có thể dẫn đến những diễn giải không chính xác về phân bổ dữ liệu, chẳng hạn như thiếu các chế độ hoặc phóng đại các ngoại lệ.



Để tránh chọn kích thước ngăn không phù hợp trong biểu đồ, hãy thử nghiệm bằng cách điều chỉnh kích thước ngăn để tìm sự cân bằng thể hiện chính xác việc phân bổ dữ liệu.

## **Nhìn ra các ngoại lệ trong biểu đồ hình hộp**

Việc không tính toán hoặc đánh dấu chính xác các ngoại lệ trong biểu đồ hình hộp là một lỗi trực quan hóa dữ liệu khác. Các ngoại lệ có thể ảnh hưởng đáng kể đến việc giải thích dữ liệu và các biểu đồ hình hộp được thiết kế đặc biệt để hiển thị các điểm dữ liệu này. Việc bỏ qua các giá trị ngoại lệ có thể gây nhầm lẫn về tính biến thiên thực sự và xu hướng trung tâm của dữ liệu.



Để tránh bỏ qua các giá trị ngoại lệ trong biểu đồ hình hộp:

1. Đảm bảo các giá trị ngoại lệ được đánh dấu và cân nhắc việc thêm chú thích để giải thích các lý do có thể gây ra những điểm bất thường này.
2. Trước khi hoàn tất việc trực quan hóa, hãy phân tích các ngoại lệ để xác định xem chúng có phải là do lỗi nhập dữ liệu, lỗi lấy mẫu hay chúng thể hiện sự biến đổi thực sự trong dữ liệu.
3. Nếu cần, hãy điều chỉnh thang đo trục y để đảm bảo hiển thị các ngoại lệ mà không làm biến dạng phần còn lại của phân phối dữ liệu.

## **Kết luận**

Đây là một số lỗi trực quan hóa dữ liệu phổ biến mà người mới bắt đầu và các chuyên gia thường mắc phải mà bạn nên tránh. Bằng cách nhận biết và giải quyết những lỗi phổ biến này, bạn có thể tạo trực quan hóa dữ liệu hiệu quả, chính xác và thân thiện với người dùng hơn.