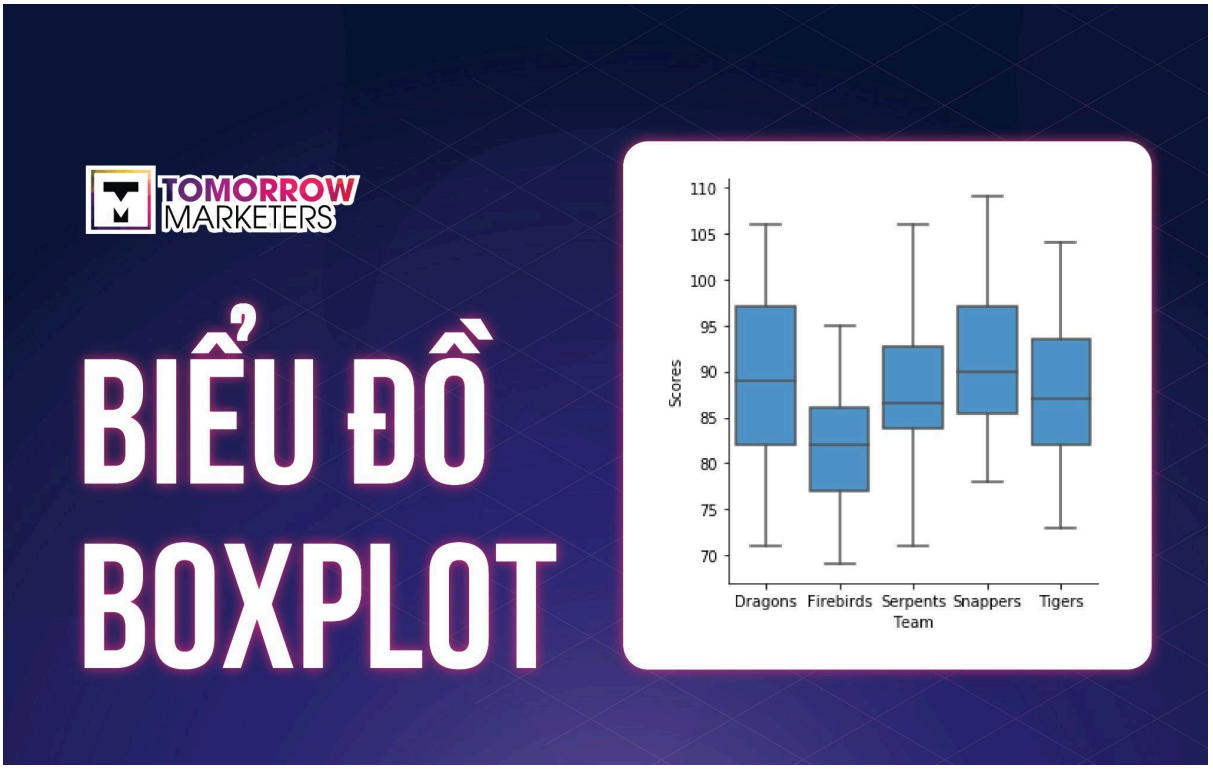
Title: Biểu đồ boxplot là gì và đọc hiểu biểu đồ này như nào?



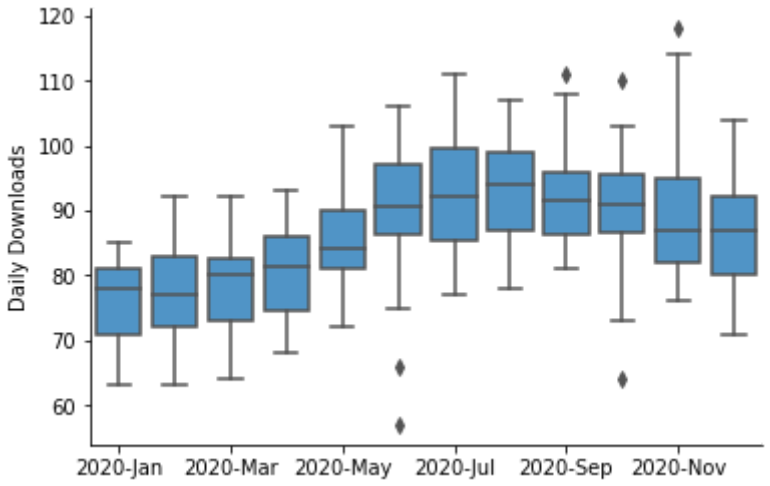
Tomorrow Marketers - Một trong những loại biểu đồ hay được sử dụng trong Data Visualization là Boxplot. Dạng biểu đồ này có ưu điểm là biểu diễn và tóm tắt rất nhiều thông tin mô tả tập dữ liệu, giúp người đọc báo cáo có thể rút ra nhiều insight khác nhau. Nhưng cũng bởi vậy, dạng biểu đồ này đôi khi có thể khó hiểu đối với một số người dùng báo cáo so với các biểu đồ quen thuộc như biểu đồ đường (line chart) hoặc biểu đồ cột (bar chart). Vậy biểu đồ boxplot là gì, đọc hiểu biểu đồ boxplot như nào và ứng dụng biểu đồ này trong phân tích dữ liệu ra sao? Cùng TM trả lời những câu hỏi này trong bài viết sau nhé!

Đọc thêm: Data Visualiz.”-*i90lư=”hm8uu hvb*ation là gì?

1. Biểu đồ hộp (boxplot) là gì?

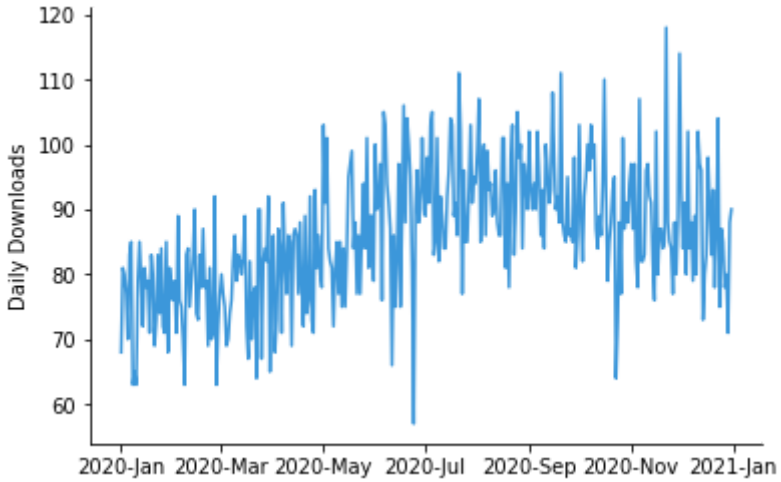
Box plot (hay còn được gọi là Box and Whisker plot) là dạng biểu đồ mô tả sự phân bố giá trị định lượng của một hoặc nhiều nhóm dữ liệu dạng phân loại (category data). Trong đó:

- Độ dài của hộp biểu thị phạm vi của 50% dữ liệu trung tâm;
- Đường kẻ giữa hộp là giá trị trung vị của tập dữ liệu (median);
- Các đường kẻ bên ngoài (còn được gọi là râu của biểu đồ) mô tả phạm vi của những giá trị dữ liệu ngoài khoảng 25% và 75% dữ liệu trung tâm;i
- Các dấu chấm bên ngoài là các giá trị ngoại lai.



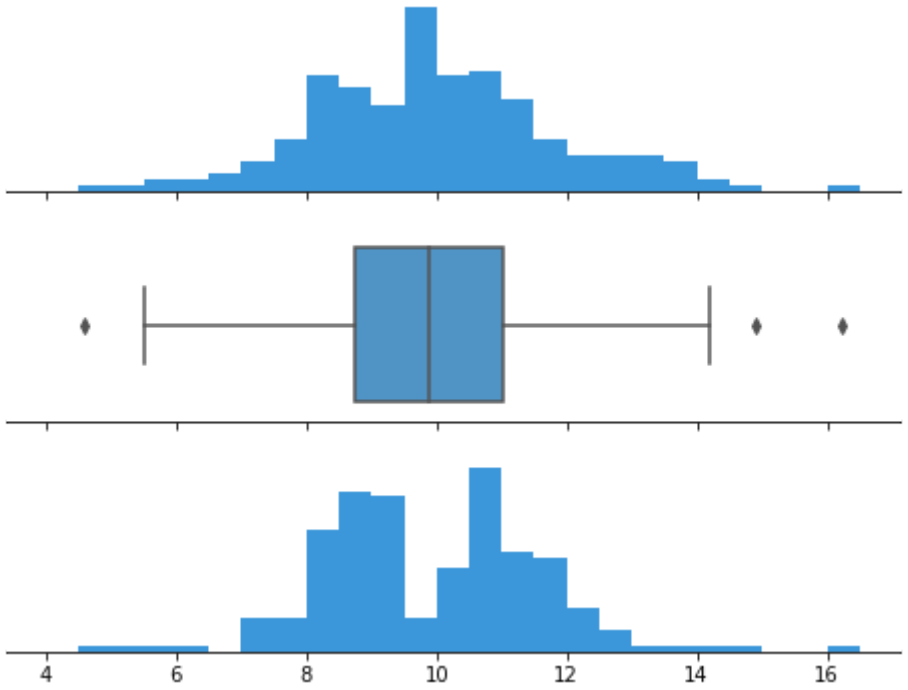
Biểu đồ ví dụ trên đây mô tả số lượt tải xuống hàng ngày của một digital app, được nhóm lại theo từng tháng. Từ biểu đồ này, chúng ta có thể thấy rằng số lượt tải xuống tăng dần từ khoảng 75 lượt/ngày trong tháng 1 lên khoảng 95 lượt/ngày trong tháng 8, trong khi số lượt tải xuống trung bình giảm nhẹ trong tháng 11 và tháng 12. Bên cạnh đó, các dấu chấm thì cho thấy 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.

Nhìn chung, biểu đồ boxplot tổng hợp nhiều insight và biểu diễn một cách trực quan về xu hướng chung của dữ liệu, so với biểu đồ đường tương tự.



Biểu đồ boxplot được sử dụng để biểu thị phân phối giá trị định lượng của các dữ liệu định dạng (category data). Nhìn vào biểu đồ boxplot , chúng ta có thể biết được mức độ phân phối, tính đối xứng và độ lệch (skewness), phương sai (variance), giá trị trung vị (median), các phân vị Q1, Q3 và những giá trị ngoại lai của tập dữ liệu. Nhờ vậy, người dùng biểu đồ có thể biết được phần lớn dữ liệu chính nằm ở đâu và so sánh phạm vi này giữa các nhóm giá trị phân loại khác nhau.

Tuy nhiên, nhược điểm của dạng biểu đồ này chính là bạn không thể quan sát được phân phối chi tiết của tập dữ liệu như histogram.

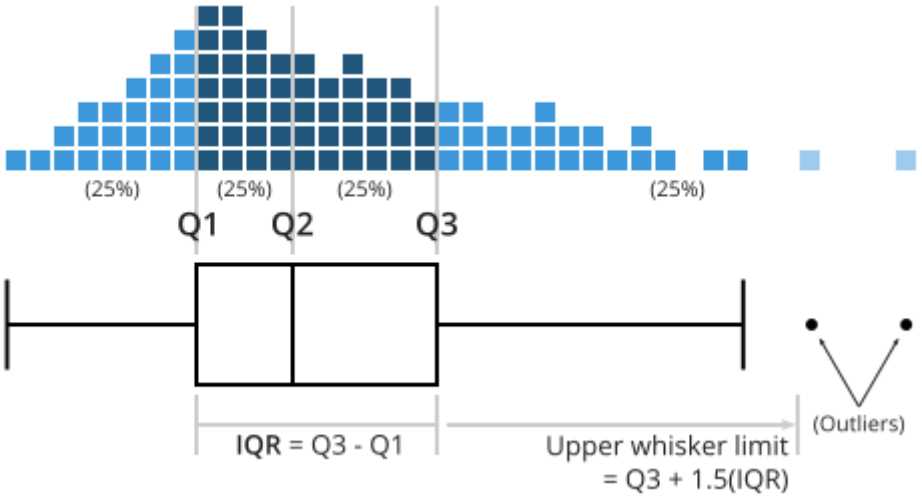


Một tập dữ liệu có thể được trực quan hóa dưới dạng biểu đồ histogram hoặc boxplot như hình trên đây

2. Đọc hiểu biểu đồ boxplot như nào?

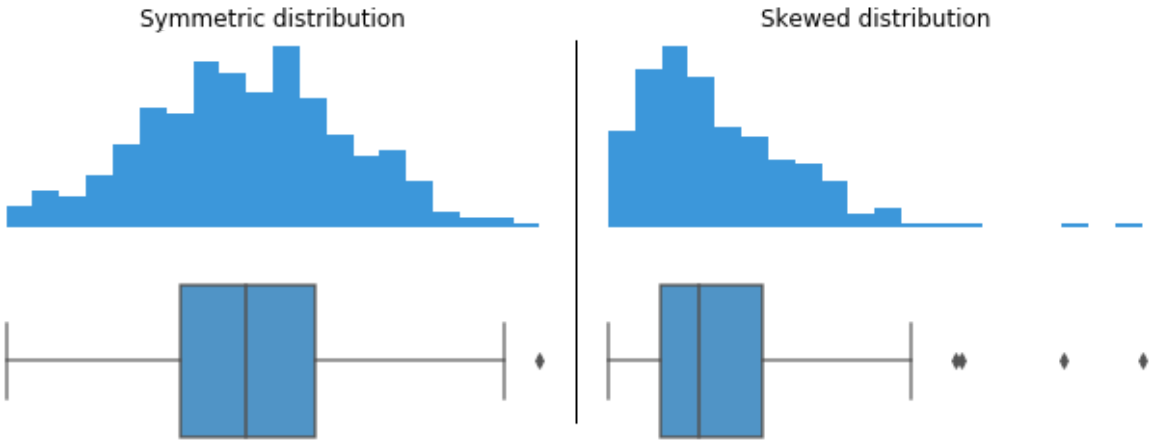
Để hiểu về boxplot, đầu tiên bạn cần làm quen với khái niệm percentile và quartile.

Biểu đồ boxplot chia tập dữ liệu thành các phần tư có số lượng giá trị bằng nhau (quartile). Phần tư đầu tiên (Q1) lớn hơn 25% dữ liệu và nhỏ hơn 75% còn lại. Phần tư thứ hai (Q2) nằm ở giữa, chia đôi dữ liệu. Q2 còn được gọi là trung vị (median). Phần tư thứ ba (Q3) lớn hơn 75% dữ liệu và nhỏ hơn 25% còn lại. Trong một biểu đồ boxplot , các hai giới hạn của hộp và đường trung tâm của nó đánh dấu vị trí của ba phần tư này.



Khoảng cách giữa Q3 và Q1 được gọi là khoảng tứ phân vị (IQR), đóng vai trò quan trọng trong việc xác định độ dài của râu của biểu đồ. Độ dài của râu kéo dài đến điểm dữ liệu xa nhất với giá trị bằng 1,5 lần IQR. Những điểm dữ liệu bên ngoài phạm vi này sẽ được coi là giá trị ngoại lai và được biểu thị bằng dấu chấm.

Khi dữ liệu phân phối đối xứng nhau, râu của hộp thể hiện giá trị trung bình sẽ nằm ở vị trí trung tâm của hộp và khoảng cách giữa Q1 và Q2 sẽ bằng khoảng cách giữa Q2 và Q3. Các giá trị ngoại lai sẽ có mặt đối xứng ở hai bên của hộp. Nếu một phân phối bị lệch, thì đường kẻ Q2 sẽ không nằm ở chính giữa hộp mà lệch vì một phía. Độ dài râu của biểu đồ cũng sẽ mất cân đối với bên dài hơn có nhiều giá trị ngoại lai hơn.



Đọc thêm: [10 khái niệm thống kê cơ bản cần biết khi làm việc với dữ liệu]

3. Ứng dụng biểu đồ boxplot trong phân tích dữ liệu

So sánh các nhóm khác nhau

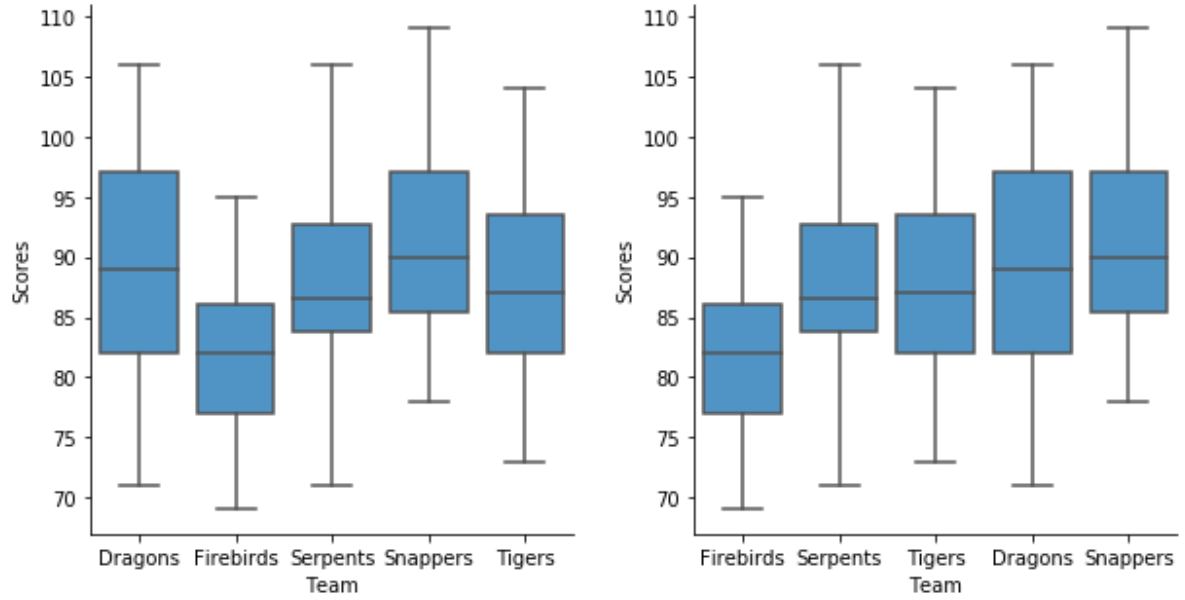
Box plot thường được sử dụng nhằm mục đích so sánh (comparison) phân phối của các nhóm dữ liệu khác nhau. Biểu đồ này được đánh giá là bao quát trong việc tóm tắt các chỉ số thống kê mô tả tập dữ liệu, giúp người đọc báo cáo dễ dàng so sánh các nhóm dữ liệu thông qua phạm vi của hộp và râu của biểu đồ.

Nếu chỉ có dữ liệu của một giá trị định dạng (category data), biểu đồ boxplot sẽ chỉ giúp tóm tắt dữ liệu và thiếu khả năng hiển thị các chi tiết về hình dạng của phân phối dữ liệu. Trong trường hợp này, bạn nên lựa chọn loại biểu đồ thể hiện mức độ phân phối chi tiết hơn như histogram hoặc đường cong mật độ (density curve).

Đánh giá thứ tự của các nhóm

Bạn có thể sắp xếp thứ tự của các biểu đồ để biểu diễn phạm vi của hộp, giúp đánh giá phạm vi giá trị của các dữ liệu định dạng.

Thông thường, các hộp sẽ được sắp xếp theo mức độ tăng dần của giá trị trung vị Q2.



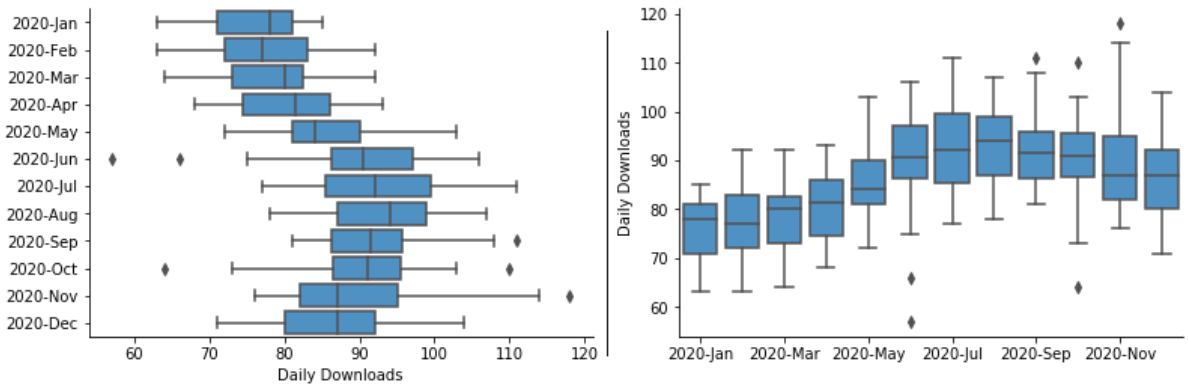
4. Các thành phần của biểu đồ boxplot

Chiều của biểu đồ boxplot

Nhằm mục đích trực quan hóa dữ liệu một cách rõ ràng và dễ đọc hơn, bạn có thể linh hoạt trong việc căn chỉnh một biểu đồ boxplot sao cho hộp được đặt theo chiều dọc (với các dữ liệu định dạng nằm trên trục ngang) hoặc theo chiều ngang (với các dữ liệu định dạng được sắp xếp theo chiều dọc).

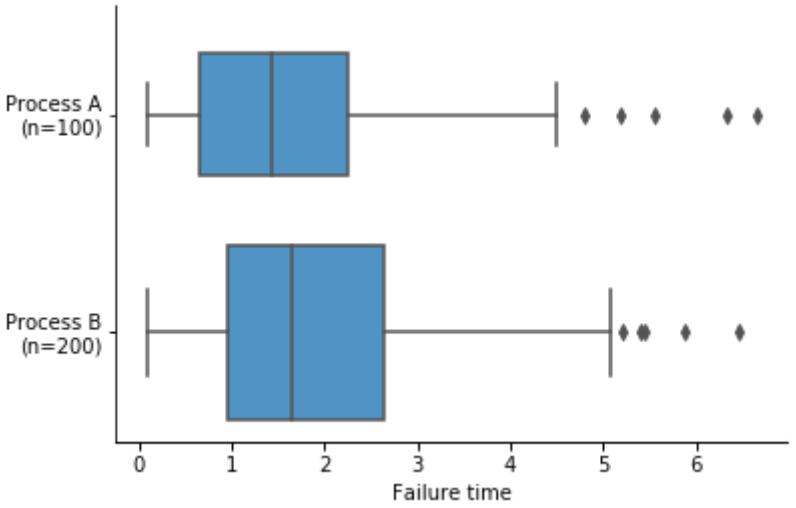
Biểu đồ boxplot ngang sẽ phù hợp hơn nếu bạn có nhiều dữ liệu định dạng hoặc các giá trị đó có độ dài không thích hợp để biểu thị theo chiều dọc, hạn chế việc phải xoay biểu đồ để đọc hoặc cắt bớt tên của các nhóm giá trị đó.

Trong khi đó, biểu đồ boxplot dọc thường được sử dụng hơn khi các biến định dạng là đơn vị thời gian.

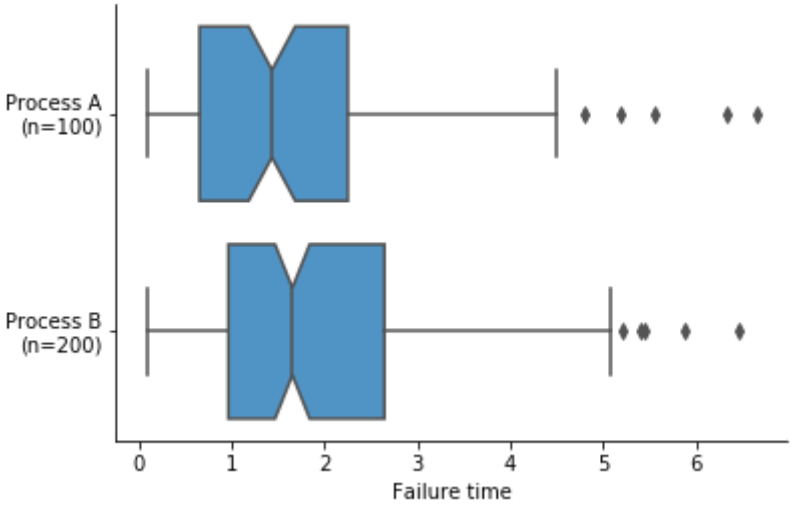


Chiều dài phạm vi hộp

Chiều rộng/phạm vi của hộp biểu diễn số lượng điểm dữ liệu của mỗi nhóm. Chiều rộng của hộp thường được chia tỷ lệ theo căn bậc hai của số điểm dữ liệu, vì căn bậc hai sẽ tỷ lệ thuận với sai số chuẩn (standard error) của các giá trị. Bạn cũng có thể bổ sung chú thích số lượng điểm dữ liệu với mỗi nhóm giá trị phân loại để người đọc báo cáo dễ dàng nắm được kích thước mẫu.



Khi sử dụng biểu đồ boxplot để so sánh phân phối dữ liệu giữa các nhóm giá trị phân loại, bạn cũng có thể so sánh các giá trị phân vị bằng cách so sánh đường kẻ bên trong hộp.



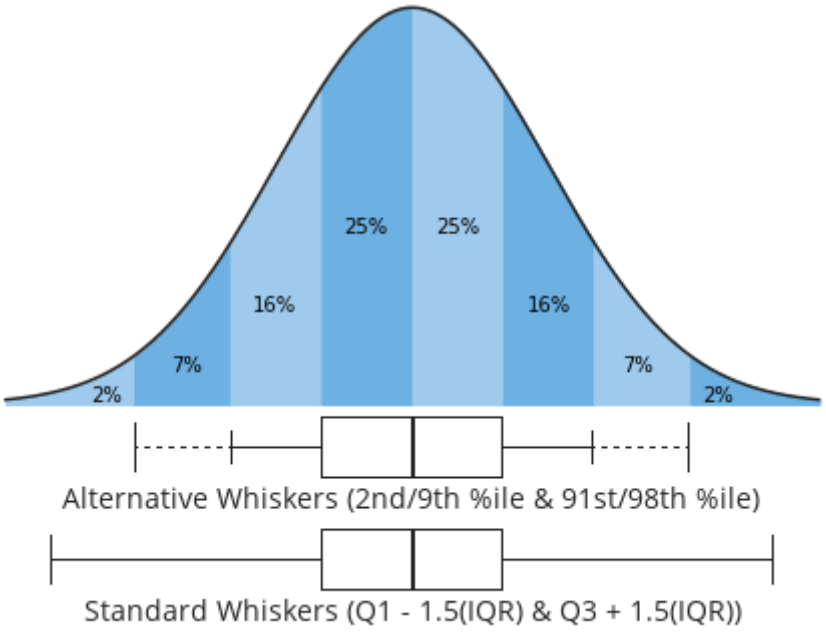
Biểu đồ boxplot phía trên cho thấy Quy trình B đang tạo ra các sản phẩm với các thời gian hỏng cao hơn.

Độ dài râu của biểu đồ và các giá trị ngoại lai

Có nhiều cách để xác định độ dài tối đa của râu của biểu đồ. Cách phổ biến nhất chính là lấy giá trị tương ứng 1,5 lần IQR từ mỗi đầu hộp.

Cách thứ hai, bạn có thể đánh dấu độ dài tối đa của râu của biểu đồ tại một giá trị phân vị khác tương tự như giá trị Q1, Q2 và Q3. Thông thường, điểm tối đa này là giá

trị ở phân vị 9% và 91%, hoặc giá trị ở phân vị 2% và 98%. Việc lựa chọn khoảng giá trị này sẽ phụ thuộc vào phân phối chuẩn (normal distribution) của tập dữ liệu. Theo phân phối chuẩn, khoảng cách giữa phân vị 9% và 25% (hoặc 91% và 75%) phải có cùng kích thước với khoảng cách giữa phân vị 25% và 50% (hoặc 50% và 75%), trong khi khoảng cách giữa phân vị 2% và 25% (hoặc 98% và 75%) phải bằng khoảng cách giữa phân vị 25% và 75%. Dựa vào đây, bạn có thể nhanh chóng nhận ra dữ liệu đối xứng (phân phối chuẩn) hay bất đối xứng (phân phối lệch).



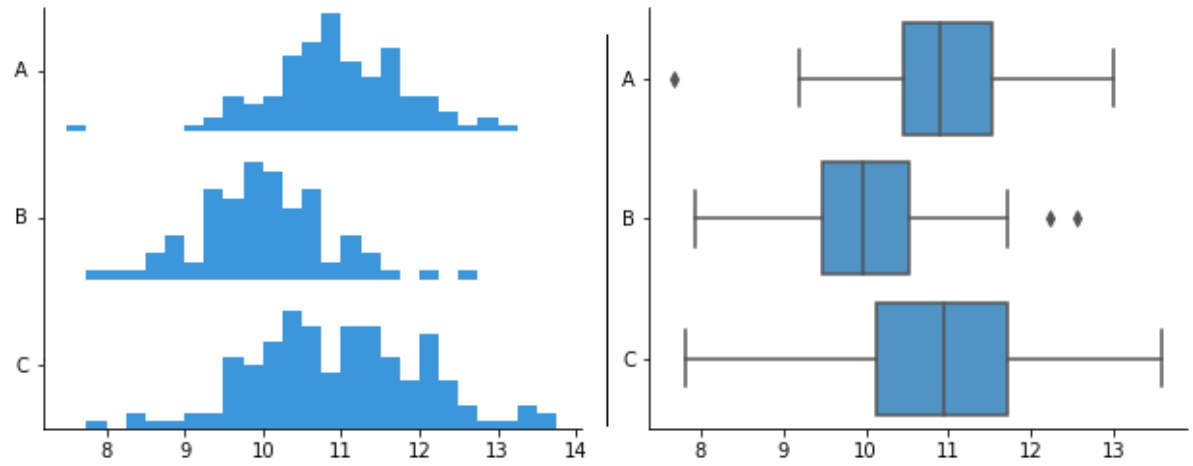
Nếu sử dụng cách hai để xác định phạm vi chiều dài của râu của biểu đồ, bạn nên có thêm ghi chú để tránh nhầm lẫn với công thức truyền thống của cách thứ nhất.

5. Các biểu đồ tương tự

Histogram

Histogram sẽ là dạng biểu đồ phù hợp hơn boxplot nếu bạn muốn phân tích phân phối của một giá trị định dạng duy nhất. Dạng biểu đồ này sẽ không mô tả chính xác các phần tư của dữ liệu như boxplot , nhưng với histogram, bạn có thể quan sát được chi tiết về phân phối của tập dữ liệu.

Nếu có hai hoặc nhiều nhóm dữ liệu định dạng, bạn có thể trực quan bằng nhiều biểu đồ histogram được xếp chồng lên nhau, giống như với biểu đồ boxplot ngang ở hình dưới đây. Tuy nhiên, cần lưu ý rằng, khi vẽ biểu đồ cho quá nhiều nhóm giá trị định dạng hơn, biểu đồ histogram của mỗi nhóm sẽ ngày càng nhiễu và khó xác định hình dạng của mỗi biểu đồ. Ngoài ra, việc thiếu các chỉ số thống kê có thể khiến việc so sánh giữa các nhóm trở nên khó khăn hơn. Vì vậy, boxplot sẽ phù hợp nếu mục đích của bạn là so sánh phân phối giữa các nhóm giá trị định dạng.



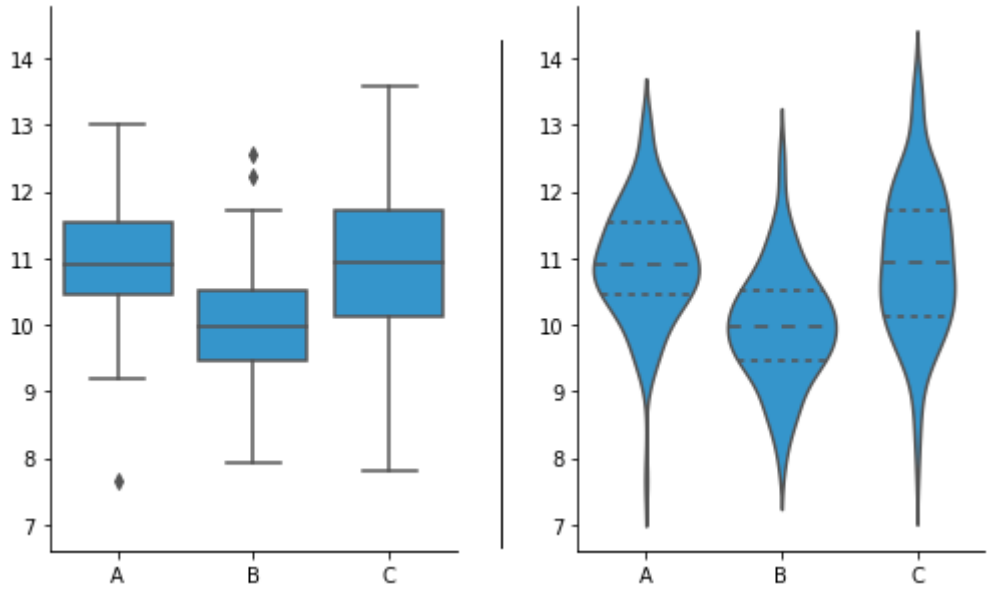
Violin plot

Một thay thế cho cốt truyện hình hộp là cốt truyện vĩ cầm. Trong một biểu đồ violon, phân phối của mỗi nhóm được biểu thị bằng một đường cong mật độ. Trong một đường cong mật độ, mỗi điểm dữ liệu không rơi vào một ngăn đơn lẻ như trong biểu đồ, mà thay vào đó đóng góp một phần diện tích nhỏ vào tổng phân phối. Biểu đồ vĩ cầm là một cách nhỏ gọn để so sánh phân phối giữa các nhóm. Thông thường, các dấu bổ sung được thêm vào biểu đồ vĩ cầm để cung cấp thông tin biểu đồ hộp tiêu chuẩn, nhưng điều này có thể làm cho biểu đồ kết quả trở nên ồn ào hơn khi đọc.

Một biểu đồ khác tương tự boxplot chính là violin plot. Với biểu đồ violin plot, phân phối của mỗi nhóm dữ liệu định dạng được biểu diễn bằng đường cong mật độ (density curve). Nói cách khác, đường cong này là phạm vi của các điểm dữ liệu. Như vậy, mỗi điểm dữ liệu sẽ là một phần tạo nên diện tích của đường cong này thay vì được tóm tắt trong phạm vi của hộp. Ngoài ra, tương tự boxplot , các giá trị ngoại lai cũng được biểu diễn dưới dạng dấu chấm trong biểu đồ violin plot.

Violin plot là dạng biểu đồ được sử dụng để so sánh mức độ phân phối dữ liệu giữa các nhóm dữ liệu định dạng.

Đọc thêm: Data Visualization - Cách chọn loại biểu đồ minh họa tốt nhất cho metrics của bạn?



Biểu đồ giá trị chữ cái (Letter-value plots)

Được phát triển bởi Hofmann, Kafadar và Wickham, biểu đồ letter-value plot là một dạng mở rộng của biểu đồ boxplot tiêu chuẩn. Dạng biểu đồ này sử dụng các hộp chữ nhật để biểu diễn mức độ phân phối của tập dữ liệu. Hộp đầu tiên chiếm 50% diện tích trung tâm. Hộp thứ hai mở rộng từ hộp thứ nhất và chiếm một nửa diện tích còn lại (75% tổng thể, 12,5% còn lại ở mỗi đầu), trong khi hộp thứ ba chiếm một nửa diện tích còn lại (87,5% tổng thể, 6,25% còn lại ở mỗi đầu), v.v. cho đến khi kết thúc và các điểm dữ liệu còn lại được đánh dấu là giá trị ngoại lai.

Dạng biểu đồ này được phát triển với ưu điểm giúp đưa ra các ước tính ổn định hơn ở các dữ liệu phần đuôi ngoài của tập dữ liệu khi bạn có nhiều dữ liệu hơn. Ngoài ra, khi càng có nhiều dữ liệu, sẽ càng có nhiều điểm dữ liệu có thể được coi là giá trị ngoại lai.

Mặc dù dạng biểu đồ này vẫn còn nhược điểm trong việc hiển thị một số giá trị thống kê mô tả phân phối (ví dụ: giá trị mode), nhưng dạng biểu đồ này có thể cung cấp nhiều insight chi tiết khi so sánh giữa các nhóm giá trị định dạng nếu bạn có quá nhiều dữ liệu.

Đọc thêm: Kể chuyện bằng dữ liệu (data storytelling) không khó chỉ với 4 bước

Tạm kết

Mọi dữ liệu đều ẩn chứa một câu chuyện, nhưng nếu chọn sai dạng biểu đồ, bạn có thể chỉ đang trình bày dữ liệu mà không thực sự kể chuyện thông qua dữ liệu, hoặc tệ hơn, việc này còn có thể khiến người đọc báo cáo gặp phải thiên kiến và đưa ra quyết định không chính xác.

Biết cách lựa chọn biểu đồ và áp dụng đúng các quy tắc trực quan hóa dữ liệu sẽ giúp bạn gỡ rối những khó khăn này. Những nội dung này được giảng dạy trong khóa học Data Analysis của Tomorrow Marketers. Nếu bạn đang gặp khó khăn trong việc khai thác triệt để sức mạnh của dữ liệu trong kinh doanh, đừng bỏ lỡ khóa học nhé!

Bài viết được biên dịch từ chartio, xin vui lòng không sao chép dưới mọi hình thức.