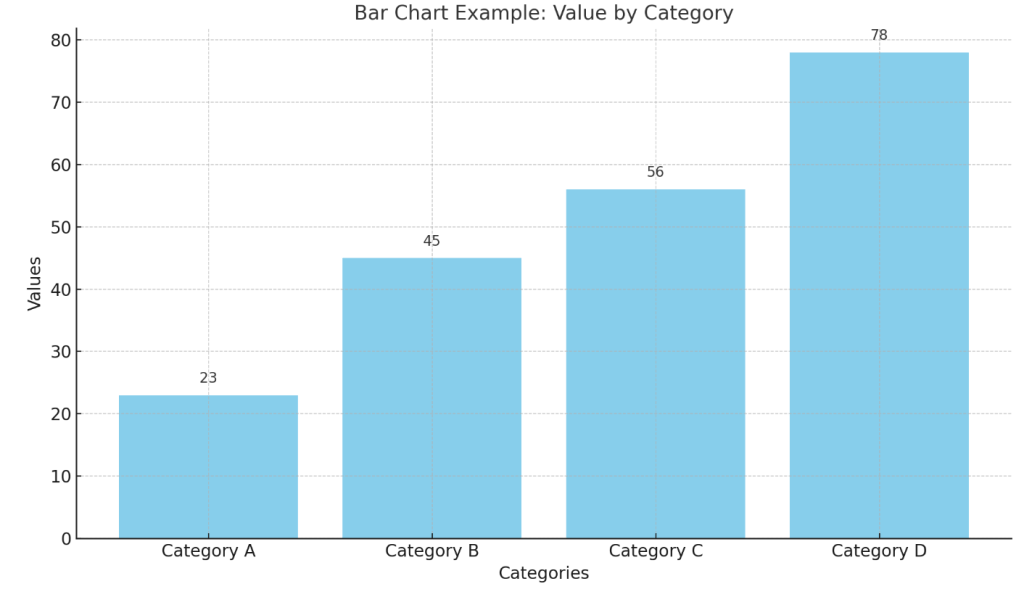
# **Phân tích và giải thích biểu đồ trực quan hóa dữ liệu**

Một trong những kỹ năng quý giá nhất của Nhà khoa học/Nhà phân tích dữ liệu là khả năng phân tích và giải thích các biểu đồ trực quan hóa. Nếu bạn muốn phân tích và diễn giải bất kỳ biểu đồ trực quan hóa dữ liệu nào, hãy dành thời gian để hiểu từng thành phần của biểu đồ. Chính các thành phần của biểu đồ sẽ vẽ dữ liệu và kể câu chuyện đằng sau những con số. Vì vậy, trong bài viết này, tôi sẽ hướng dẫn bạn đầy đủ về tất cả các thành phần của một số biểu đồ trực quan hóa dữ liệu thường được sử dụng và cách diễn giải chúng bằng cách hiểu các thành phần.

## Dưới đây là cách diễn giải đồ thị trực quan hóa dữ liệu

Chúng ta hãy xem qua các thành phần cốt lõi của biểu đồ trực quan hóa dữ liệu thường được sử dụng để hiểu cách diễn giải chúng.

## **Biểu đồ thanh (Bar Chart)**

****

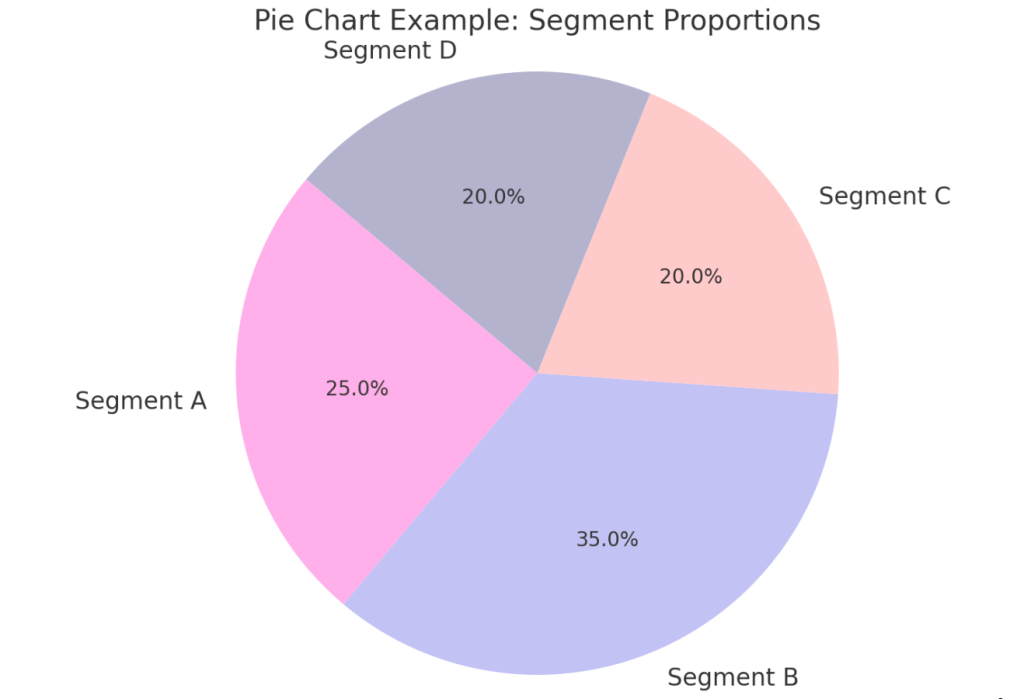
Biểu đồ thanh là công cụ nền tảng trong việc trực quan hóa dữ liệu, được thiết kế để so sánh các giá trị số giữa các danh mục khác nhau. Cấu trúc của biểu đồ thanh rất đơn giản nhưng nó cung cấp những hiểu biết sâu sắc thông qua các thành phần chính của nó:

* **Trục** (**Axes)**: Trục ngang (trục x) thường hiển thị các danh mục, trong khi trục dọc (trục y) định lượng các giá trị. Sự sắp xếp này giúp nhanh chóng xác định cách các danh mục khác nhau xếp chồng lên nhau.
* **Thanh (Bars)** : Chiều cao hoặc chiều dài của mỗi thanh tỷ lệ thuận với giá trị mà nó đại diện, cung cấp sự so sánh trực quan giữa các danh mục.
* **Nhãn/Tiêu đề (Labels/Titles)** : Những phần tử này cung cấp ngữ cảnh thiết yếu, đảm bảo người xem hiểu được những gì đang được trình bày mà không mơ hồ.

Để phân tích và diễn giải biểu đồ thanh một cách hiệu quả, hãy bắt đầu bằng cách so sánh trực quan kích thước của các thanh, đại diện cho giá trị của các danh mục khác nhau, để đánh giá độ lớn tương đối của chúng và xác định danh mục nào nổi bật nhờ giá trị cao hơn hoặc thấp hơn. Sự so sánh này tạo điều kiện cho việc đánh giá nhanh chóng bối cảnh của dữ liệu.

Tiếp theo, kiểm tra cách sắp xếp các thanh cho bất kỳ mẫu nào có thể quan sát được, chẳng hạn như xu hướng tăng dần hoặc giảm dần, có thể làm sáng tỏ động lực và định hướng cơ bản trong tập dữ liệu, từ đó cho thấy các giá trị thay đổi như thế nào giữa các danh mục. Cuối cùng, đặc biệt chú ý đến bất kỳ thanh nào khác biệt đáng kể so với các thanh còn lại, vì những giá trị ngoại lệ này có thể chỉ ra các trường hợp ngoại lệ hoặc điểm bất thường cần được điều tra thêm.

## **Biểu đồ hình tròn (Pie Chart)**

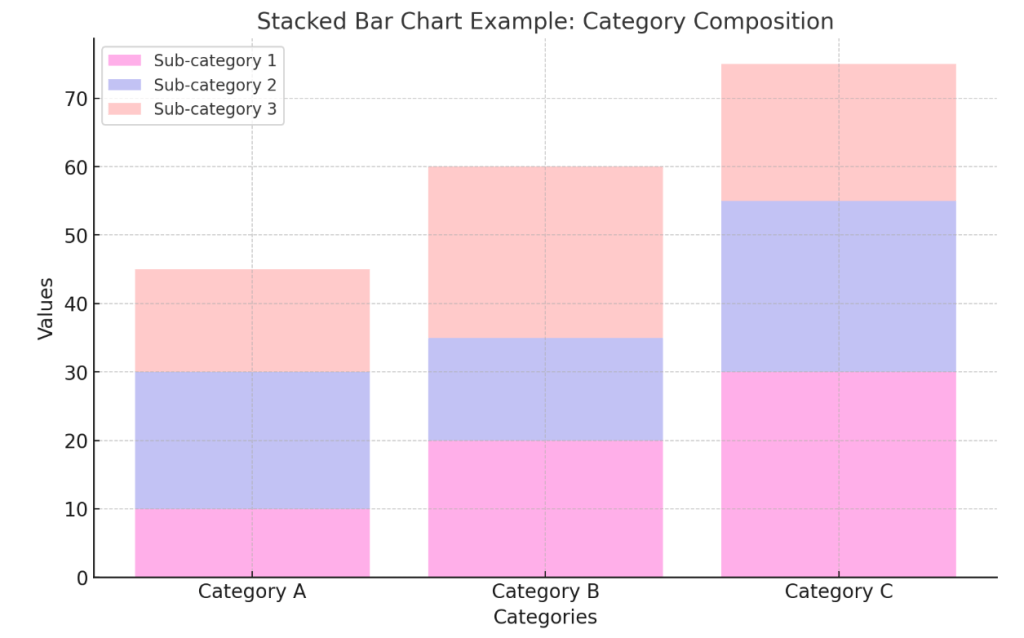
****

Biểu đồ hình tròn cung cấp một phương pháp trực quan để hiển thị tỷ lệ và tỷ lệ phần trăm nhằm chia tổng thể thành các phần cấu thành. Dưới đây là các thành phần của biểu đồ hình tròn:

* **Các lát cắt (Slices)**: Mỗi lát bánh đại diện cho một phần của tổng thể, với chiều dài cung của nó biểu thị tỷ lệ của từng loại.
* **Nhãn/Chú giải**(**Labels/Legends)**: Cung cấp sự rõ ràng về ý nghĩa của từng lát cắt, đảm bảo người xem có thể diễn giải dữ liệu một cách chính xác.

Để phân tích và diễn giải biểu đồ hình tròn, ban đầu hãy tập trung vào phân tích tỷ lệ của từng phần, điều này cho thấy kích thước tương đối và sự đóng góp của từng danh mục vào tổng thể, giúp hiểu được sự phân bổ các phần trong tổng số dữ liệu. Luôn tìm kiếm sự thống trị giữa các phần, trong đó phần lớn hơn biểu thị các danh mục có thị phần đáng kể. Nó nêu bật các lĩnh vực quan tâm hoặc ưu tiên chính trong tập dữ liệu.

## **Biểu đồ thanh xếp chồng (Stacked Bar Chart)**

****

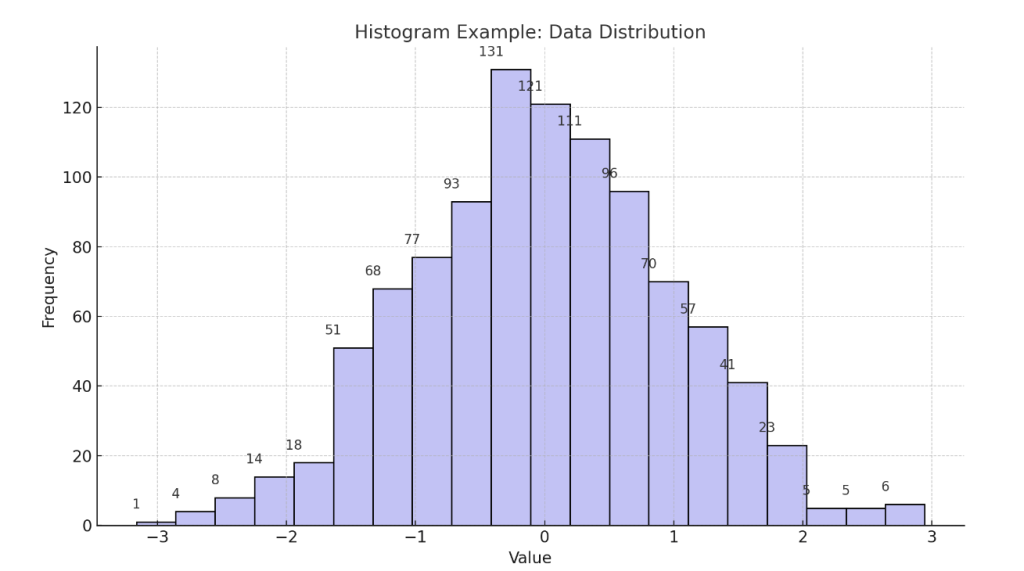
Biểu đồ thanh xếp chồng nâng cao biểu đồ thanh truyền thống bằng cách chia mỗi thanh thành các phân đoạn đại diện cho các danh mục phụ, mang lại hiểu biết theo lớp về thành phần dữ liệu. Dưới đây là các thành phần của biểu đồ thanh xếp chồng:

* **Thanh (Bars)**: Phản ánh tổng giá trị cho từng danh mục, với thanh được chia thành các phân đoạn được mã hóa bằng màu.
* **Phân đoạn (Segments)** : Mỗi phân đoạn đại diện cho một phần của tổng thể, biểu thị kích thước tương đối của các danh mục phụ.
* **Nhãn/Chú giải (Labels/Legends)** : Rất quan trọng để giải mã các phân đoạn, cung cấp thông tin về ý nghĩa của mỗi màu.

Để phân tích và giải thích biểu đồ thanh xếp chồng, hãy bắt đầu bằng cách so sánh tổng chiều cao của các thanh để hiểu giá trị tổng thể giữa các danh mục khác nhau, giúp nhanh chóng xác định danh mục nào có giá trị tích lũy cao nhất. Bước này cung cấp nền tảng để đánh giá các điểm dữ liệu tổng hợp.

Sau đó, đi sâu vào thành phần của từng thanh bằng cách kiểm tra các phân đoạn của nó, đại diện cho các danh mục phụ, để xem các thành phần này đóng góp như thế nào vào tổng số và tỷ lệ của chúng khác nhau như thế nào giữa các danh mục. Việc kiểm tra chi tiết này có thể làm nổi bật tầm quan trọng tương đối của các danh mục phụ và tiết lộ các mô hình hoặc sự mất cân bằng trong dữ liệu. Cuối cùng, hãy tập trung tìm hiểu thành phần của từng danh mục bằng cách phân tích tỷ lệ và cách sắp xếp các phân đoạn trong các thanh.

## **Histogram**

****

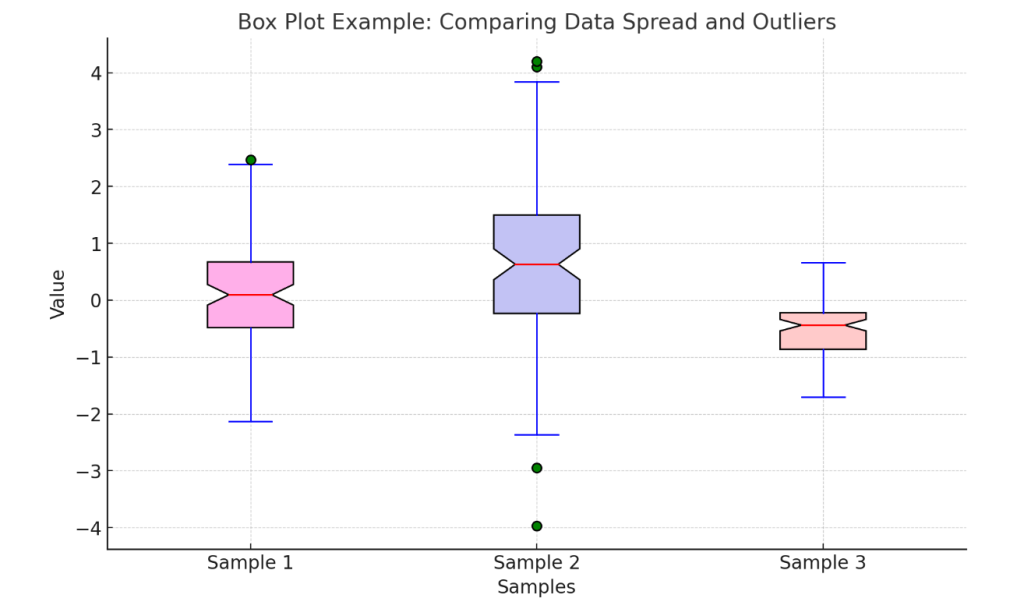
Histogram được sử dụng để trực quan hóa việc phân phối tập dữ liệu bằng cách nhóm các giá trị vào các nhóm và hiển thị tần suất dữ liệu trong các khoảng đó. Dưới đây là các thành phần của histogram bạn nên biết:

* **Thùng (Bins)** : Các khoảng này nắm bắt phạm vi dữ liệu, giúp sắp xếp tập dữ liệu thành các phân đoạn có thể quản lý được.
* **Tần suất (Frequency)** : Chiều cao của mỗi thanh cho biết có bao nhiêu quan sát nằm trong một thùng cụ thể, điều này cho thấy mô hình phân phối.

Để phân tích và diễn giải histogram, hãy bắt đầu bằng cách đánh giá hình dạng của phân phối, chẳng hạn như phân bố đó là chuẩn, lệch hay hai đỉnh, điều này cho thấy những hiểu biết sâu sắc quan trọng về các đặc điểm và xu hướng cơ bản của tập dữ liệu. Hình dạng của phân bố có thể chỉ ra xu hướng trung tâm và tính biến đổi cũng như liệu các điểm dữ liệu được nhóm lại hay trải rộng trên phạm vi.

Tiếp theo, đánh giá độ phân tán bằng cách xem xét chiều rộng và phân bố của các khoảng (bins), giúp hiểu rõ phạm vi và độ biến thiên của dữ liệu, chỉ ra mức độ phân tán của các điểm dữ liệu xung quanh xu hướng trung tâm. Các khoảng rộng hơn có thể ám chỉ sự biến thiên lớn hơn trong giá trị dữ liệu, trong khi các khoảng hẹp hơn chỉ ra sự nhất quán nhiều hơn. Cuối cùng, hãy chú ý đến các khoảng thưa, đặc biệt là những khoảng ở hai đầu cực của biểu đồ histogram, vì chúng có thể biểu thị các ngoại lệ hoặc các điểm dữ liệu bất thường lệch lớn so với phần còn lại của bộ dữ liệu. Những ngoại lệ này có thể nêu bật các điểm bất thường hoặc các khía cạnh độc đáo của dữ liệu cần được khám phá thêm.

## **Biểu đồ hình hộp (Box Plot)**

****

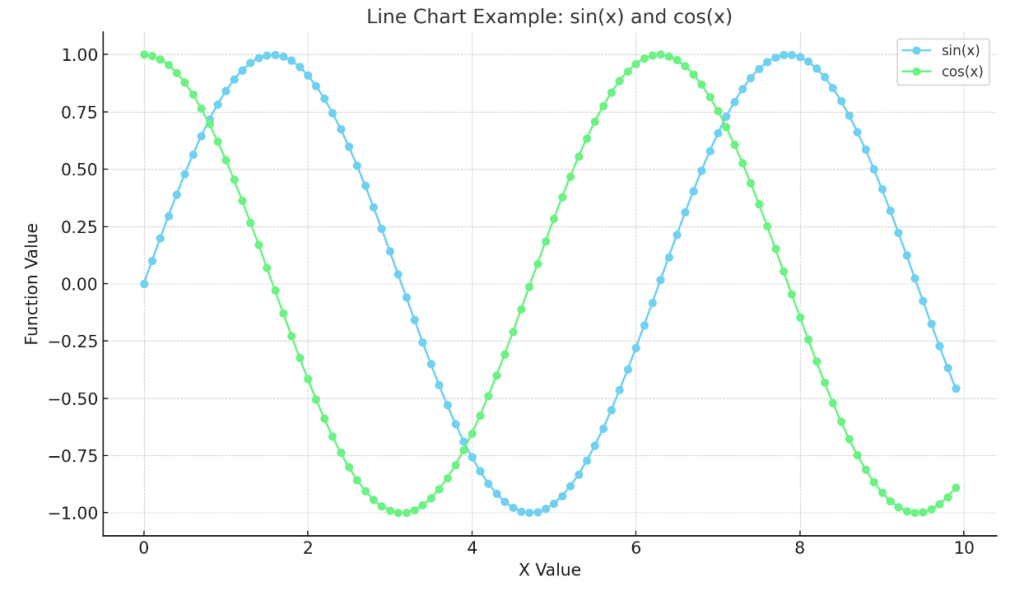
Biểu đồ hộp (hoặc biểu đồ ria) cung cấp một bản tóm tắt thống kê gọn gàng về phân phối, nổi bật lên xu hướng trung tâm và độ biến thiên. Dưới đây là các thành phần của biểu đồ hộp bạn cần biết:

* Hộp (**Box)** : Là phần cốt lõi của biểu đồ, đại diện cho 50% dữ liệu ở giữa của bộ dữ liệu (khoảng tứ phân vị), với một dòng chỉ ra giá trị trung vị.
* Ria (**Whiskers)**: Mở rộng từ hộp để bao trùm toàn bộ phạm vi phân tán của dữ liệu, loại trừ các ngoại lệ.
* Ngoại lệ (**Outliers)**: Các điểm dữ liệu được vẽ riêng lẻ nằm ngoài ria, đánh dấu các điểm dữ liệu lệch lớn so với phần còn lại.

Để phân tích và giải thích một biểu đồ hộp, ban đầu hãy tập trung vào hộp trung tâm, đại diện cho khoảng tứ phân vị (IQR) và chứa giá trị trung vị, đề xuất cái nhìn nhanh chóng vào xu hướng trung tâm và độ biến thiên của bộ dữ liệu. IQR tự thân nó cho thấy sự phân tán của 50% dữ liệu ở giữa, cung cấp một hình ảnh rõ ràng về sự phân tán dữ liệu.

Tiếp theo, đánh giá đối xứng của phân phối bằng cách xem xét cách hộp và ria được sắp xếp xung quanh giá trị trung vị. Một phân phối đối xứng sẽ có giá trị trung vị nằm ở vị trí trung tâm trong hộp, và ria có chiều dài xấp xỉ bằng nhau, trong khi sự không đối xứng có thể ám chỉ sự lệch trong dữ liệu, với phần lớn các điểm dữ liệu nằm ở trên hoặc dưới giá trị trung vị. Cuối cùng, quan sát bất kỳ điểm riêng lẻ nào được vẽ ngoài ria, vì những điểm này đại diện cho các ngoại lệ. Ngoại lệ có thể chỉ ra các trường hợp ngoại lệ trong bộ dữ liệu hoặc ám chỉ về sự biến thiên dữ liệu không phù hợp với mô hình chung.

## **Biểu đồ đường (Line Charts)**

****

Biểu đồ đường là công cụ không thể thiếu để theo dõi sự thay đổi theo thời gian bằng cách cung cấp một cái nhìn động về xu hướng và biến động. Dưới đây là các thành phần của một biểu đồ đường bạn cần biết:

* Đường (**Line)**: Kết nối các điểm dữ liệu liên tiếp để minh họa chuyển động hoặc xu hướng của dữ liệu theo thời gian hoặc qua các điều kiện khác nhau.
* Điểm (**Points)**: Đánh dấu các giá trị dữ liệu thực tế, neo đường vào quan sát thực tế.
* Trục (**Axes)**: Thông thường, trục x biểu thị thời gian, trong khi trục y định lượng biến số quan tâm.

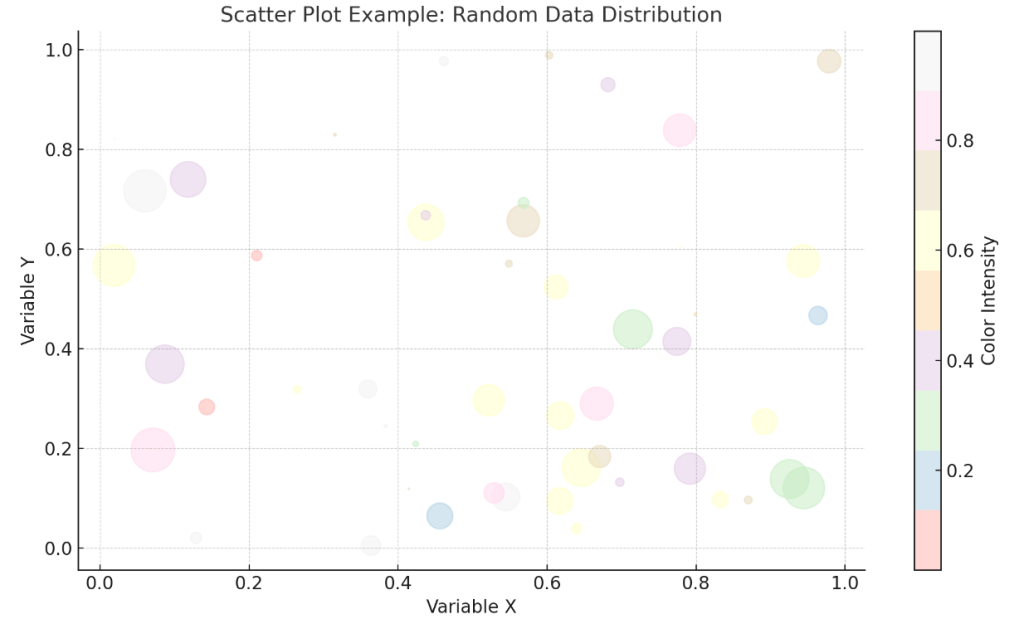
Để phân tích và giải thích một biểu đồ đường, bắt đầu bằng cách kiểm tra hướng tổng thể của đường(s), có thể chỉ ra xu hướng theo thời gian hoặc qua các điều kiện:

* xu hướng lên chỉ ra sự tăng
* xu hướng xuống chỉ ra sự giảm
* và đường phẳng biểu thị sự ổn định

Việc phân tích này rất quan trọng để hiểu rõ chuyển động và tiến trình tổng thể của bộ dữ liệu.

Tiếp theo, đánh giá sự biến động của đường bằng cách lưu ý đến các biến động của nó; những thay đổi thường xuyên và sắc nét cho thấy biến thiên cao, phản ánh các giai đoạn bất ổn hoặc thay đổi nhanh chóng trong dữ liệu, trong khi một đường mượt mà chỉ ra sự ổn định. Cuối cùng, khi có nhiều đường, so sánh các quỹ đạo và mẫu vẽ của chúng để đánh giá hiệu suất hoặc hành vi tương đối. Sự so sánh này có thể nêu bật cách các nhóm hoặc biến số khác nhau phản ứng với nhau qua cùng một khoảng thời gian, tiết lộ những hiểu biết về động lực và tương tác của chúng.

## **Biểu đồ phân tán (Scatter Plots)**

****

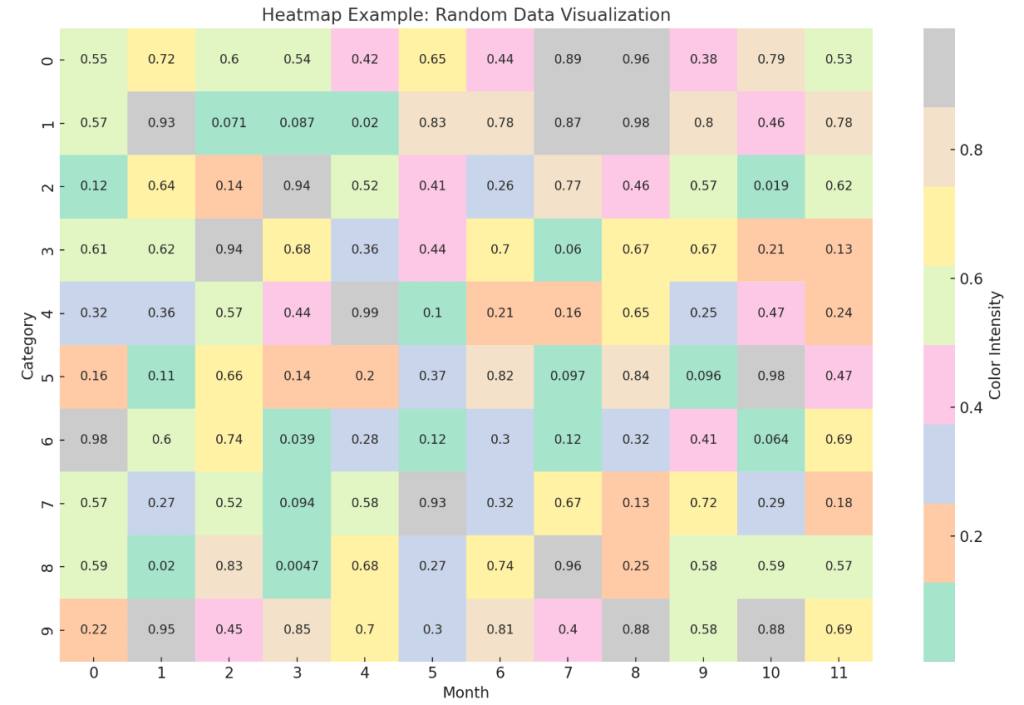
Biểu đồ phân tán là lý tưởng để khám phá mối quan hệ giữa hai biến, hiển thị các điểm dữ liệu trên một mặt phẳng hai chiều. Dưới đây là các thành phần của một biểu đồ phân tán bạn cần biết:

* Điểm Dữ Liệu (**Data Points)**: Mỗi điểm đại diện cho một quan sát, được vẽ theo hai biến số quan tâm.
* Trục (**Axes)**: Định nghĩa các chiều của phân tích, với mỗi trục đại diện cho một trong các biến số.

Để phân tích và giải thích một biểu đồ phân tán, đầu tiên, quan sát tổng thể mô hình hình thành bởi các điểm dữ liệu để đánh giá bản chất của mối quan hệ giữa hai biến liên quan; một mô hình có xu hướng dốc lên cho thấy mối tương quan tích cực và dốc xuống chỉ ra mối tương quan tiêu cực, và không có mô hình rõ ràng ngụ ý ít hoặc không có tương quan. Bước này rất quan trọng để hiểu cách thức thay đổi trong một biến có thể ảnh hưởng đến biến kia.

Tiếp theo, tìm kiếm các cụm điểm được nhóm lại với nhau, vì những điều này có thể tiết lộ các tập con trong dữ liệu của bạn hoặc chỉ ra rằng các phân đoạn khác nhau có cách biểu hiện theo những cách riêng biệt, có khả năng cung cấp cái nhìn sâu sắc vào các mẫu hoặc mối quan hệ ẩn. Cuối cùng, xác định bất kỳ điểm nào nổi bật so với phần còn lại của cụm dữ liệu chính; những ngoại lệ này có thể đại diện cho các trường hợp bất thường hoặc độc đáo lệch khỏi xu hướng chung và có thể cần được điều tra thêm để hiểu nguyên nhân hoặc hậu quả của chúng.

## **Biểu đồ nhiệt (Heat Maps)**

****

Bản đồ nhiệt sử dụng các dải màu để biểu diễn giá trị dữ liệu trên một ma trận hai chiều, cung cấp một cách trực quan hóa dữ liệu phức tạp một cách dày đặc nhưng trực quan. Dưới đây là các thành phần của bản đồ nhiệt bạn cần biết:

* Thang Màu (**Colour Scale)**: Các màu sắc hoặc các sắc độ khác nhau biểu diễn các giá trị dữ liệu khác nhau, với thang màu thường được hiển thị bên cạnh để tham khảo.
* Ma Trận (**Matrix)**: Bố cục nơi giá trị dữ liệu được gán cho các tọa độ cụ thể, thường biểu diễn cho các hạng mục hoặc khoảng thời gian.

Để phân tích và giải thích bản đồ nhiệt, bắt đầu bằng cách tập trung vào cường độ của các màu sắc được sử dụng, vì những màu sắc này biểu diễn độ lớn của giá trị dữ liệu; các khu vực có màu sắc mãnh liệt hơn nổi bật các tập trung đáng kể của giá trị cao hoặc thấp, thu hút sự chú ý đến các khu vực quan tâm trong bộ dữ liệu.

Tiếp theo, quan sát các dải màu và sự chuyển đổi giữa các màu trên bản đồ nhiệt để xác định các mô hình, xu hướng, hoặc tương quan; các dải màu mượt mà có thể chỉ ra sự thay đổi dần dần hoặc xu hướng theo thời gian hoặc hạng mục, trong khi những thay đổi màu sắc rõ rệt có thể chỉ ra sự thay đổi đột ngột hoặc ranh giới giữa các phân đoạn dữ liệu khác nhau. Cuối cùng, sử dụng phân bố màu để thực hiện phân tích so sánh trên bản đồ nhiệt, so sánh các khu vực hoặc phần khác nhau để hiểu cách giá trị dữ liệu biến đổi theo không gian hoặc hạng mục.

## **Kết luận**

Đây là cách bạn có thể phân tích và giải thích các đồ thị trực quan hóa dữ liệu như một Nhà khoa học dữ liệu/Phân tích viên. Nếu bạn muốn phân tích và giải thích bất kỳ đồ thị trực quan hóa dữ liệu nào, hãy dành thời gian để hiểu các thành phần của mỗi đồ thị. Chính các thành phần của đồ thị vẽ dữ liệu và kể câu chuyện đằng sau các con số.