# **Phân tích và trực quan hóa dữ liệu thăm dò**

## **Phân tích dữ liệu:**

Phân tích dữ liệu khám phá dùng để hiển thị bất kỳ một thông số theo một cách cho phép. Trực quan hoá là sử dụng các kỹ thuật để xem số liệu hoặc xem thông tin.

Phân tích dữ liệu thăm dò gồm 3 vấn đề:

* Đầu tiên là cách lập biểu đồ dữ liệu,hiển thị thông tin một cách ngắn gọn,dễ hiểu. Biểu đồ giúp hiển thị thông tin và dễ dàng so sánh dữ liệu. Do đó, hiển thị càng nhiều thông tin càng tốt. Cần hiển thị dữ liệu một cách hiệu quả.
* Thứ hai, mục tiêu của EDA là xem các dữ liệu để khám phá ra những điều mà bạn không dự đoán được. Do đó, phải rèn luyện cách đọc các đồ thị để lấy thông tin. Sau đó, quay lại kiểm tra các dữ kiệu riêng lẻ hoặc xu hướng để hiểu cách mà nó diễn ra.
* Cuối cùng chính là làm mô hình. Mô hình của cùng một dữ liệu có thể khác nhau tuỳ thuộc vào mục tiêu khám phá của chúng ta.

Phân tích dữ liệu khám phá không có một hình dạng, đồ thị hay kịch bản nhất định. Chúng ta có thể xem xét tấ cả các mô hình trong dữ liệu, có thể biểu diễn các thông tin khác nhau. Bạn có thể tự quyết định những gì muốn xem và cách khám phá. Chọn mô hình/ đồ thị mà có thể biểu diễn dữ kiệu một cách hiệu quả nhất (dễ đọc, so sánh,…)

## **Sử dụng dữ liệu để trả lời câu hỏi:**

Ta phải kiểm tra dữ liệu của mình, xem xét dữ liệu của và hiểu các mô hình đang được sử dụng. Ta sẽ dùng nhiều hơn một biểu đồ, đi sâu hơn trong việc kiểm tra dữ liệu theo các cách khác nhau. Trong bất kỳ loại tóm tắt dữ liệu nào, luôn có một nhóm kiểm soát rõ ràng hoặc ngầm định.

Thế giới thực rất phức tạp nhưng chúng ta có thể học hỏi bằng cách nghiên cứu kỹ lưỡng dữ liệu. Chúng ta có thể xem xét tất cả các khuynh hướng trong dữ liệu. Dữ liệu là không có giới hạn. Ta có thể tải dữ liệu, xem các con số và lấy ra các biến khác nhau. Ta có thể vẽ các kết quả khác nhau dưới dạng hàm của ngày, giờ, trạng thái và các biến khác.

## **Tại sao phải trực quan hoá dữ liệu:**

* Lý do đầu tiên là vì tập dữ liệu lớn rất khó để giải thích chỉ bằng cách kiểm tra trực quan dữ liệu thô. Có thể rất khó để thử và giải thích ý nghĩa từ dữ liệu nếu chỉ xem xét các con số thô, hoặc thậm chí trong nhiều trường hợp là bản tóm tắt của dữ liệu,báo cáo.
* Giúp giao tiếp và truyền dặt thông tin một cách dễ dàng và hiệu quả. Tuy nhiên, cần phải chọn biểu đồ hợp lý, nhiều thông tin để việc truyền đạt hiệu quả.
* Có thể khám phá ra được các thông tin mới, các mối quan hệ giữa các biến trong dữ liệu.
* Tạo tương tác với dữ liệu, kết nối nhiều dữ liệu lại với nhau.

## **Quy trình ra quyết định trực quan hóa dữ liệu :**

Ta có dữ liệu thô => Đặt câu hỏi với dữ liệu => Trực quan hoá dữ liệu dựa trên các câu hỏi =>Dựa trên mô hình trực quan, sàng lọc => Thông tin để giao tiếp

## **Trang tổng quan:**

Trang tổng quan là một cách sắp xếp dữ liệu để chúng ta có thể thấy nhiều biểu đồ và cách tất cả chúng được liên kết với nhau. Thông thường, bảng điều khiển được tạo bằng một số biểu đồ trong một bảng điều khiển. Những biểu đồ đó đến từ một tập dữ liệu nhưng tập dữ liệu đó có thể thay đổi theo thời gian. Chúng ta có thể sử dụng các tập dữ liệu khác nhau để tạo ra các loại biểu đồ giống nhau.

Điều quan trọng của các bảng điều khiển này là nhiều hình ảnh trực quan được sắp xếp cùng nhau, sao cho chúng xây dựng một câu chuyện mà bản thân một biểu đồ không đủ để cho chúng ta biết. Bằng cách sắp xếp nhiều biểu đồ theo cách thích hợp, chúng ta có thể xây dựng một bức tranh rất đầy đủ về câu chuyện cơ bản mà chúng ta muốn truyền đạt.

Một tính năng tuyệt vời khác của trang tổng quan là chúng có thể động, có nghĩa là chúng thay đổi khi dữ liệu được nhập vào hệ thống, để chúng ta có thể xem các sự kiện diễn ra như thế nào theo thời gian. Chúng ta có thể thay đổi thủ công các tập dữ liệu để xem liệu các mối quan hệ mà chúng ta thấy có đúng với một tập dữ liệu cũng đúng với một tập dữ liệu khác hay không.

Một tính năng rất hay về trang tổng quan là tính thẩm mỹ. Người dùng bị thu hút rất nhiều vào việc muốn sử dụng trang tổng quan. Khi chúng ta chỉ nhìn thấy một biểu đồ đôi khi có thể hơi nhàm chán. Nhưng khi chúng ta có nhiều biểu đồ sẽ thay đổi linh hoạt khi chúng ta làm điều gì đó thực sự thu hút sự chú ý của người dùng và khiến họ muốn tương tác với dữ liệu. Tuy nhiên, thường với bảng điều khiển, chúng ta có thể mất một chút thời gian và điều quan trọng cần nhớ là tính tiện dụng được đặt lên hàng đầu. Chúng cần đủ đơn giản để thông điệp không bị che khuất và nội dung chúng ta muốn truyền đạt cho người khác phải rất, rất rõ ràng. Nó không nên chỉ trông đẹp mà còn phải hiệu quả.

Điều tuyệt vời về trang tổng quan là chúng ta có thể thêm dữ liệu vào đó và xem điều đó ảnh hưởng đến thông tin như thế nào. Nếu chúng ta tìm kiếm New York Mets, nó sẽ cập nhật thông tin đó nhận được một dòng khác trên biểu đồ. Và chúng ta có thể thấy rằng họ thậm chí còn ít phổ biến hơn New York Yankees. Ngoại trừ một trường hợp ngoại lệ là vào năm 2015, nơi họ đã thực hiện World Series. Và chúng đột nhiên trở nên rất phổ biến về lịch sử cụm từ tìm kiếm.

Điểm đầu tiên mà ta muốn làm là khả năng tái sử dụng. Chúng ta có thể sử dụng lại các bố cục trang tổng quan giống nhau với nhiều tập dữ liệu khác nhau và chúng ta không cần phải tiếp tục cập nhật số liệu và vẽ lại số liệu bằng cách nhập các bộ dữ liệu khác nhau và sử dụng chương trình của chúng ta hoặc phần mềm khác của chúng ta để tạo các biểu đồ đó. Khi chúng ta đã tạo xong mẫu cung cấp dữ liệu và chúng ta có thể sử dụng lại trang tổng quan để tạo biểu đồ nhiều lần.

Ưu điểm thứ hai mà ta nhấn mạnh là chúng ta có thể có những hình dung liên kết. Điều này có nghĩa là chúng ta có thể xem xét một đối tượng địa lý trên một biểu đồ và đồng thời đối tượng địa lý đó sẽ nổi bật trên biểu đồ khác. Ví dụ: nếu chúng ta muốn xem Google Xu hướng năm 2006 - có bao nhiêu lượt truy cập - chúng ta có thể liên kết điểm đó - di chuột qua điểm đó trên biểu đồ đường với bản đồ để xem điều đó thay đổi như thế nào trong năm 2006 so với toàn bộ khoảng thời gian. Hình ảnh hóa được liên kết có thể là một cách rất hiệu quả để mang lại ý nghĩa cho dữ liệu.

Ưu điểm thứ ba của trang tổng quan là chúng ta có thể điều chỉnh mọi thứ một cách nhanh chóng, chúng ta có thể điều chỉnh chúng khi chúng ta tiếp tục đối với các thông số đặc biệt. Ví dụ: nếu chúng ta đang thực hiện phân tích khám phá, chúng ta có thể đang xây dựng một mô hình, một mô hình thống kê về dữ liệu của chúng ta. Và chúng ta có thể muốn nghĩ rằng chúng ta nên bao gồm biến này hay loại trừ biến này. Nếu chúng ta có một hằng số trong một thuật toán thống kê, có lẽ chúng ta muốn điều chỉnh hằng số đó và xem nó ảnh hưởng như thế nào đến đầu ra. Nếu chúng ta làm điều đó lặp đi lặp lại, vẽ lại biểu đồ mỗi lần, điều đó sẽ rất tốn công sức. Nó sẽ làm chậm quá trình học tập của chúng ta. Nhưng bằng cách có một bảng điều khiển, chúng ta có thể chỉ có một thanh trượt và chúng ta có thể trượt giá trị đó qua. Hoặc chúng ta có thể có một menu để chúng ta chọn các giá trị để nhập hoặc loại trừ. Và chúng ta thấy kết quả một cách linh động. Tiết kiệm được nhiều thời gian.

Điều đáng giá thứ tư là tính tương tác. Ta đã chỉ ra một số ví dụ về tương tác, chẳng hạn như di chuột qua các điểm và nhận thông tin. Nhưng với bảng điều khiển, chúng ta nhận được nhiều tương tác hơn. Người dùng thực sự có thể khám phá. Và ngày càng có nhiều trường hợp các cá nhân đang đọc và xem trực quan dữ liệu ngay lập tức cảm thấy như họ có thể di chuột qua các hình ảnh trực quan và nhận được nhiều thông tin hơn chỉ là một hình ảnh trực quan tĩnh. Vì vậy, người dùng đang bắt đầu mong đợi sự tương tác. Họ có thể học nhanh hơn bằng cách tự tương tác với dữ liệu và biểu đồ.

Nhược điểm là công việc trả trước có nghĩa là chúng ta phải tùy chỉnh trang tổng quan của riêng mình và mất nhiều thời gian hơn một chút. Nhưng nó nhanh hơn nhiều về lâu dài. Một số ví dụ hoặc một ví dụ cụ thể về nơi có thể sử dụng trang tổng quan để phân tích khám phá nhằm phát triển những hiểu biết chính về mối quan hệ giữa các biến khác nhau và cách chúng ta có thể hình dung thông tin đó cũng như thực hiện một số dữ liệu phân tích về các mối quan hệ đó. Đi vào chi tiết hơn về các khía cạnh cụ thể của tập dữ liệu cơ bản. Đây là ý tưởng về hệ thống phân cấp thông tin thường có một số biểu đồ nên được rõ ràng nhất, nổi bật nhất. Chúng nên xuất hiện đầu tiên. Đây là những biểu đồ mà khi chúng ta có người dùng của mình lần đầu tiên nhìn vào hình ảnh trực quan, họ sẽ ngay lập tức hiểu được ý tưởng mà chúng ta đang cố gắng truyền đạt. Sau đó, các biểu đồ tiếp theo có thể là một tập hợp con của dữ liệu.

Lọc dữ liệu để hiển thị các mẫu phức tạp hơn. Chúng sẽ đến sau những biểu đồ ban đầu. Và vì vậy, khi chúng ta xây dựng một bảng điều khiển, chúng ta cần suy nghĩ xem người dùng của ta sẽ giao tiếp với thông tin này như thế nào. Chúng ta cần làm gì trước để họ hiểu những gì họ đang xem và điều gì sẽ xảy ra sau đó? Ví dụ, tất cả thông tin có nên xuất hiện trên một màn hình không? Một số người tranh luận rằng trang tổng quan chỉ nên xuất hiện trên một màn hình, nhưng những người khác sẽ tranh luận, chẳng hạn như ta, rằng có thông tin và biểu đồ trong các tab khác hoặc trên các menu có thể nhấp khác hoặc thông tin sẽ bật lên khi di chuột qua các biểu đồ cụ thể hoặc điểm dữ liệu là một cách rất hiệu quả để nhận được nhiều thông tin hơn. Và tất cả những điều này là để cân bằng giữa lựa chọn thiết kế với lựa chọn dữ liệu, điều gì trông hấp dẫn nhất? Điều gì sẽ lôi kéo mọi người tương tác với dữ liệu so với điều gì truyền đạt dữ liệu hiệu quả nhất? Và thực sự chúng ta nên luôn xem xét đặc điểm thứ hai đó một cách nổi bật nhất. Làm cách nào để ta có thể truyền đạt dữ liệu một cách hiệu quả nhất?