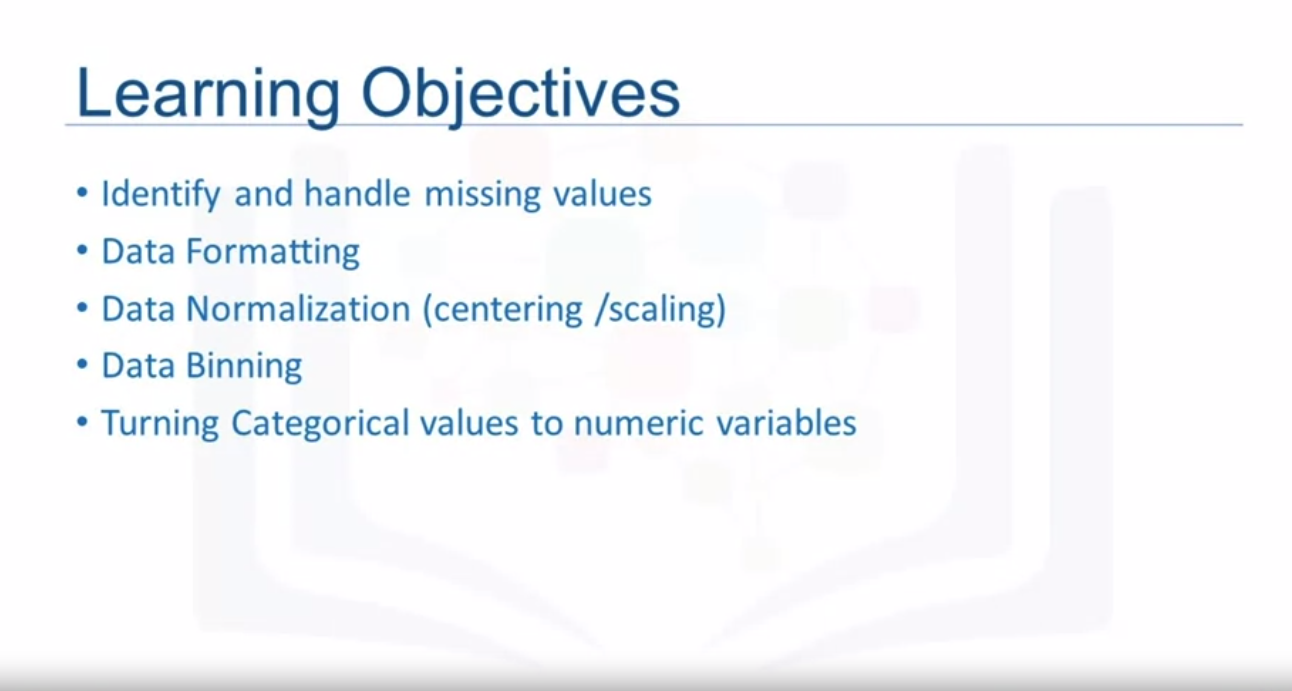
Sắp xếp dữ liệu (data wrangling)

Data pre-processing (tiền xử lý dữ liệu) là một bước cần thiết trong phân tích dữ liệu. Đây là quá trình chuyển đổi hoặc ánh xạ dữ liệu từ một dạng thô trở thành một dạng khác phù để chúng ta có thể làm việc với nó trong tương lai.

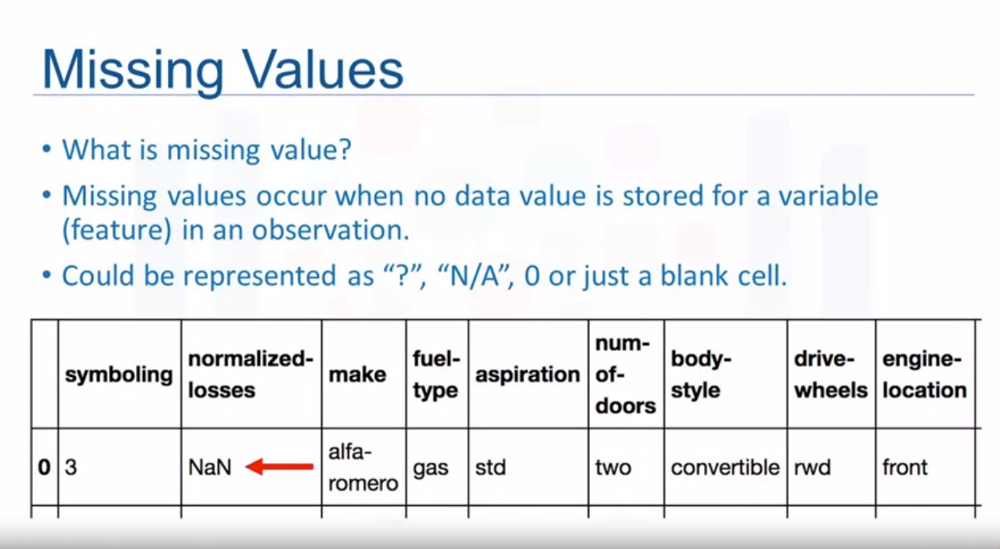
Tiền xử lý dữ liệu còn thường được gọi là data cleaning (làm sạch dữ liệu) hoặc data wrangling (sắp xếp dữ liệu) hoặc những thuật ngữ khác.



Vậy Data pre-processing bao gồm:

**Xác địng và xử lý giá trị còn thiếu (identify & handle missing values)**

Mising value (giá trị bị thiếu) xuất hiện bất cứ khi nào một mục nhập dữ liệu được để trống.



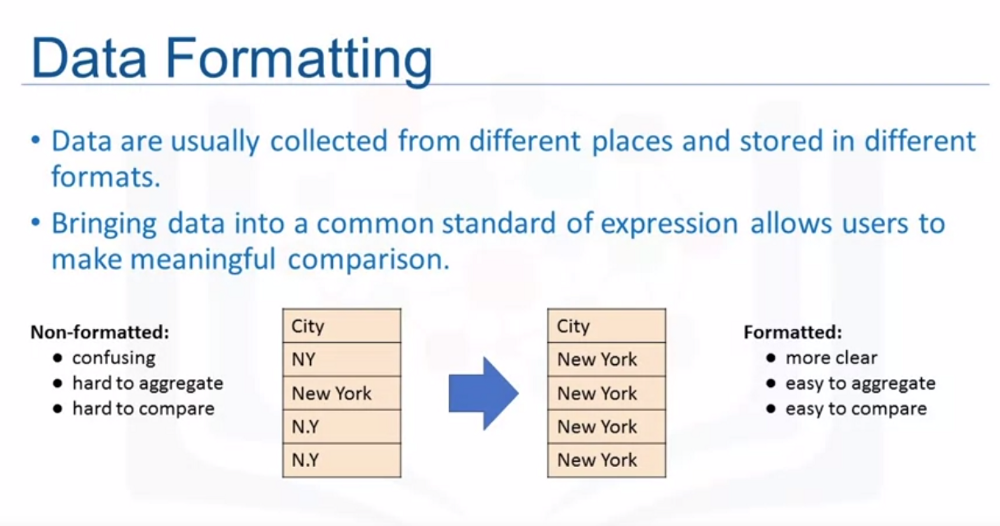
Khi không có giá trị dữ liệu được lưu trữ cho đặc trưng hoặc một quan sát cụ thể, chúng tôi nói đặc trưng này có một missing value trong tập dữ liệu xuất hiện dưới dạng dấu hỏi, số không hoặc chỉ một ô trống. Trong ví dụ ở đây, normalized loss có một mising value được biểu thị bằng NaN.

Có nhiều cách để xử lý các missing value và điều này không phụ thuộc vào Python, R, hay bất cứ công cụ nào bạn dùng. Tất nhiên mỗi trường hợp khác nhau nên được đánh giá khác nhau. Tuy nhiên, đây là những lựa chọn điển hình bạn có thể xem xét:

* Đầu tiên là kiểm tra xem người hoặc nhóm đã thu thập dữ liệu có thể quay trở lại và tìm giá trị thực tế. Một khả năng khác là loại bỏ dữ liệu nơi tìm thấy mising value đó. Khi bạn drop (loại bỏ) dữ liệu, bạn có thể drop toàn bộ biến hoặc chỉ dữ liệu có giá trị còn thiếu.
* Nếu bạn không có nhiều quan sát với dữ liệu bị thiếu, thường tốt nhất là nên loại bỏ chúng. Thay vì xoá dữ liệu, bạn sẽ muốn xem xét làm một điều gì khác có ít tác động nhất đến dữ liệu. Thay thế dữ liệu sẽ tốt hơn vì không có dữ liệu nào bị lãng phí. Tuy nhiên, nó ít chính xác hơn vì chúng ta cần phải thay thế dữ liệu bị thiếu bằng cách đoán xem dữ liệu sẽ là gì. Một tiêu chuẩn việc thay thế này là thay giá trị của missing value thiếu giá trị bởi giá trị trung bình của toàn bộ biến.

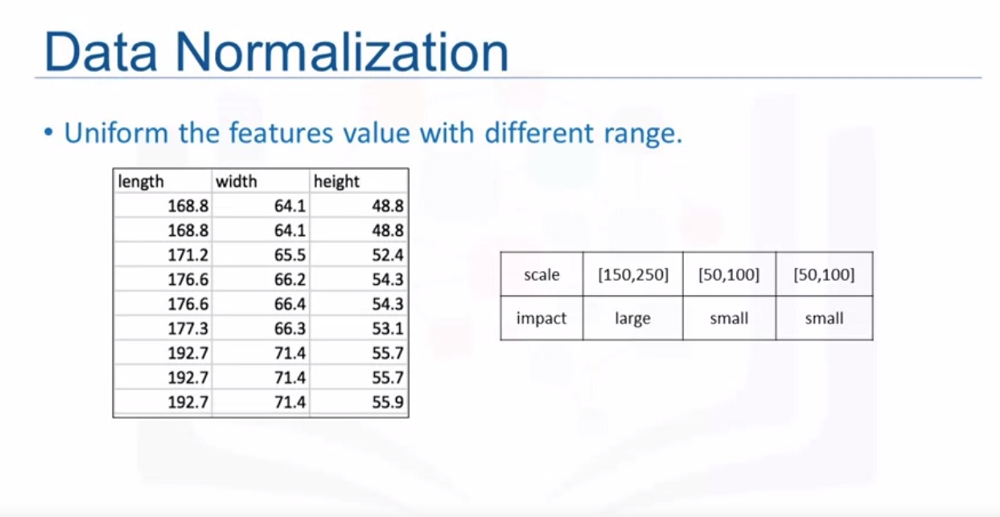
**Định dạng dữ liệu (Data formatting)**

Dữ liệu thường được thu thập từ nhiều nơi khác nhau, bởi nhiều người khác nhau, có thể được lưu trữ trong các định dạng khác nhau. Data formatting (định dạng dữ liệu) có nghiã là đưa dữ liệu vào một tiêu chuẩn chung cho phép người dùng thực hiện các phép so sánh có ý nghĩa. Là một phần của data cleaning (làm sạch dữ liệu), data formatting đảm bảo dữ liệu phù hợp và dễ hiểu.



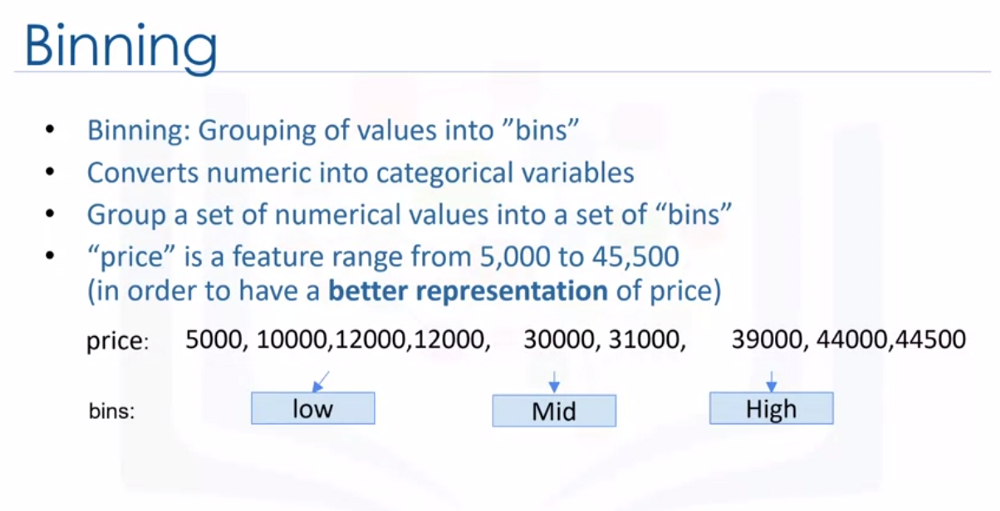
Ví dụ: mọi người có thể sử dụng các biểu thức khác nhau để đại diện cho thành phố New York, chẳng hạn như NY, nY, N.Y hoặc New York. Đôi khi, những dữ liệu không sach này là một điều đáng xem.

**Chuẩn hoá dữ liệu (Data normalization)**



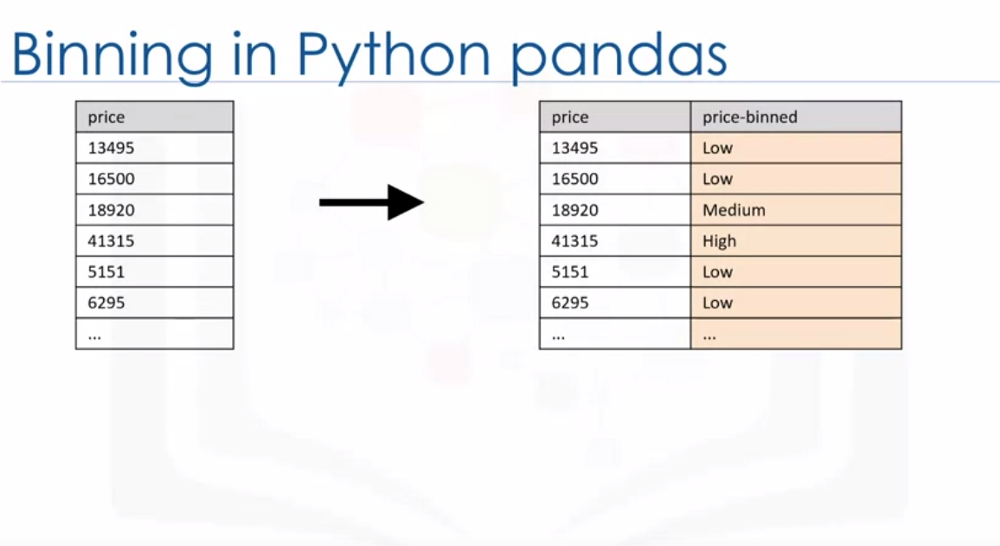
Khi chúng ta xem bộ dữ liệu xe đã sử dụng, chúng ta nhận thấy trong dữ liệu rằng đặc trưng về độ dài nằm trong khoảng từ 150-250, trong khi chiều rộng và chiều cao dao động từ 50-100. Chúng ta có thể chuẩn hoá các biến này để phạm vi của các giá trị nhất quán. Việc chuẩn hoá này làm cho một số phân tích thống kê dễ dàng hơn. Bằng cách làm cho phạm vi nhất quán giữa các biến, chuẩn hoá cho phép thực hiện các phép so sánh giữa các đặc trưng, đảm bảo chúng có ảnh hưởng tương tự nhau. Nó cũng quan trọng vì lý do tính toán.

**Binning**



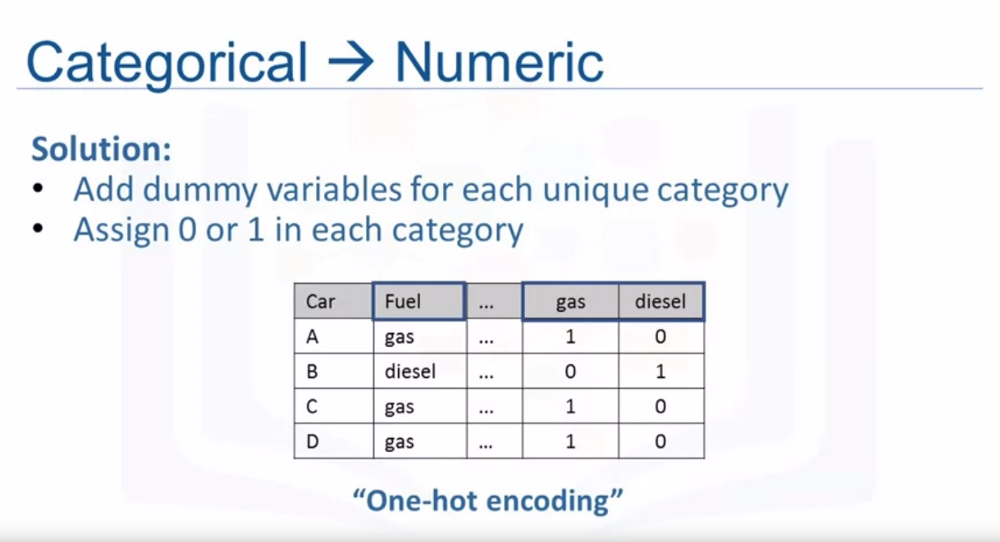
Binning là khi bạn nhóm các giá trị lại với nhau thành các bin. Ví dụ: bạn có thể bin “age” thành [0 đến 5], [6 đến 0], [11 đến 15] v.v. Đôi khi, binning có thể cải thiện độ chính xác của mô hình dự đoán.

Ngoài ra, đôi khi chúng tô sử dụng dữ liệu binning để nhóm một tập hợp các giá trị số thành một số lượng bin nhỏ hơn để có một sự hiểu biết tốt hơn về phân phối dữ liệu.



Bây giờ hãy quay lại với biểu đồ mà ta xem xét tính biến thiên trong giá kim cương. Chúng ta có thể phân loại chúng thành 3 bin: low, medium, high.

**Chuyển biến phân loại thành biến địng lượng (Turning categorical variables into quantitative variables)**



Hầu hết các mô hình thống kê không thể lấy các object hoặc string làm đầu vào và đối với việc training mô hình chỉ nhận các số làm đầu vào. Trong bộ dữ liệu xe hơi, đặc trưng fuel type như một biến categorical có hai giá trị, gas hoặc diesel, ở định dạng string.

Để phân tích sâu hơn, phai chuyển đổi các biến này thành mộ số dạng định dạng số, mã hoá các giá trị bằng cách thêm các đặc trưng mới tương ứng với mỗi giá trị trong đặc trưng ban đầu chúng ta muốn mã hoá.

Trong trường hợp fuel type có hai giá trị không trùng nhau, gas và diesel, vậy nên phải tạo ra hai đặc trưng mới, gas và diesel. Khi một fias trị xảy ra trong đặc trưng ban đầu, đặt giá trị tương tứng thành 1 trong đặct trưng mới Phần còn lại của các đặc trưng được đặt thành 0.

Trong ví dụ về fuel, đối với xe B, giá trị fuel là diesel. Do đó, đặc trưng diesel bằng 1 và đặc trưng gas bằng 0. Tương tự, đối với xe D, giá trị nhiên liệu là gas. Do đó, đặt đặc trưng gas bằng 1 và đặc trưng diesel bằng 0. Kỹ thuật này thường được gọi là “One-hot encoding”.