Pre-processing Data in Python

1. Nội dung

* Tiền xử lí dữ liệu là quá trình chuyển đổi hoặc ánh xạ từ dạng thô sang một định dạng khác để làm cho nó sẵn sàng để phân tích thêm
* Tiền xử lí dữ liệu thường được gọi là làm sạch dữ liệu ( Data Cleaning) hoặc sắp xếp dữ liệu (Data Wrangling)
* Xác định và xử lí các dữ liệu còn thiếu
* Các định dạng dữ liệu: dữ liệu từ nhiều nguồn khác nhau có thể ở các định dạng khác nhau, quy ước khác nhau 🡪 chuẩn hóa các giá trị thành cùng định dạng, cùng quy ước.
* Chuẩn hóa dữ liệu: các cột dữ liệu số khác nhau có thể có phạm vi rất khác nhau và so sánh trực tiếp thường không có ý nghĩa 🡪 chuẩn hóa dữ liệu là một cách để đưa tất cả dữ liệu vào một phạm vi tương tự để so sánh hữu ích hơn.

Cụ thể sẽ tập trung vào các kĩ thuật : centering và scaling

* Binning dữ liệu : tạo ra các danh mục lớn hơn từ một tập hợp các giá trị số. 🡪 hữu ích khi so sánh giữa các nhóm dữ liệu
* Biến phân loại : chuyển đổi giá trị phân loại thành các biến số để làm cho mô hình thống kê dễ dàng hơn.

1. Slide



