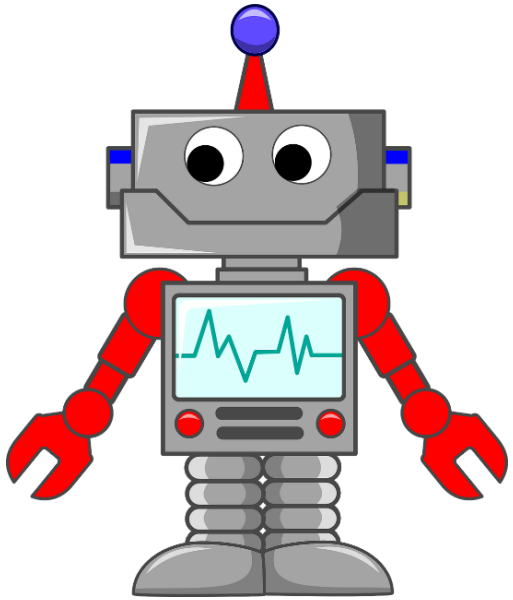
CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

BÀI 03

MACHINE LEARNING PIPELINE & PHÂN TÍCH DỮ LIỆU

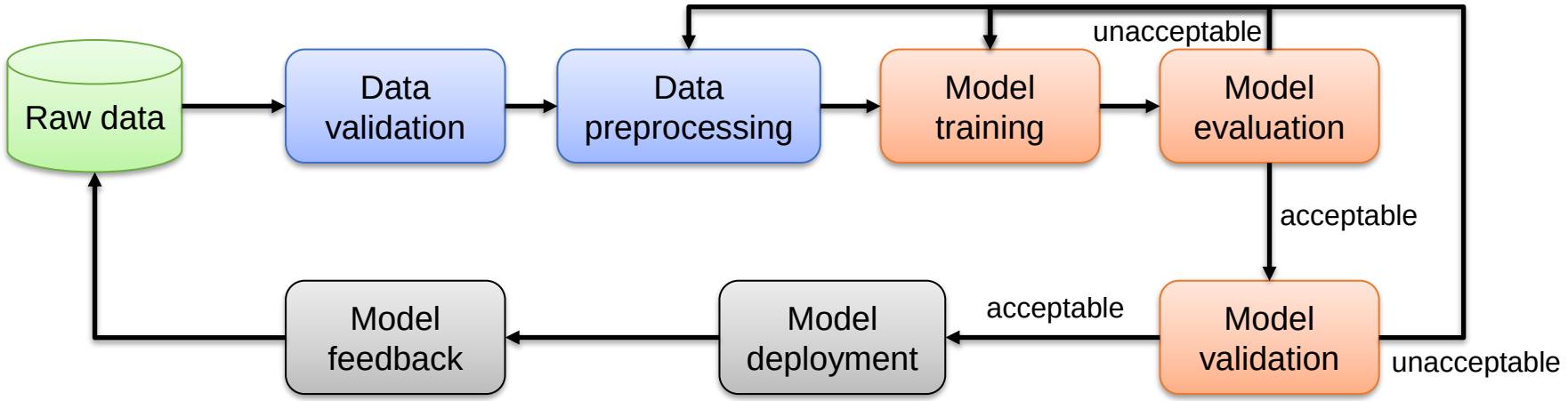


TS. Nguyễn Vinh Tiệp

NỘI DUNG

Machine Learning Pipeline

- **Machine Learning Pipeline** (Quy trình học máy):
  - Là chuỗi các bước xử lý dữ liệu, xây dựng mô hình được liên kết với nhau
  - Để chuẩn hoá, tối ưu hoá quá trình xây dựng, huấn luyện, đánh giá và triển khai mô hình máy học



Tại sao Machine Learning Pipeline

Tính hiệu quả	Khả năng tái lập (reproducibility)	Đơn giản hóa triển khai	Khả năng tái sử dụng (reusable)	Dễ kiểm soát	Dễ cộng tác
• Quy trình hóa và tự động hóa • Tiết kiệm thời gian, nguồn lực	• Lặp lại được kết quả thực hiện trước đó	• Mô hình triển luôn sẵn sàng chuyển giao từ thí nghiệm sang thực tế	• Mã nguồn, module có thể được dùng cho dự án khác	• Kiểm soát phiên bản • Đảm bảo bản mới nhất	• Mỗi người làm một module • Làm việc đồng thời

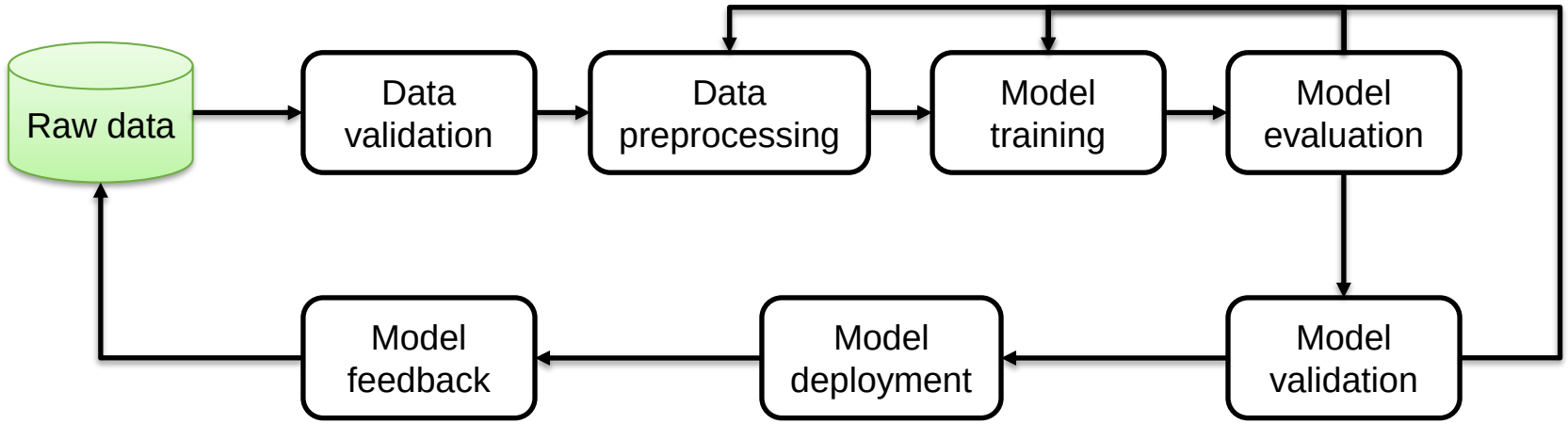
NỘI DUNG

1. GIỚI THIỆU MACHINE LEARNING PIPELINE

2. CÁC THÀNH PHẦN CỦA MACHINE LEARNING PIPELINE

3. PHÂN TÍCH DỮ LIỆU – EXPLORATORY DATA ANALYSIS

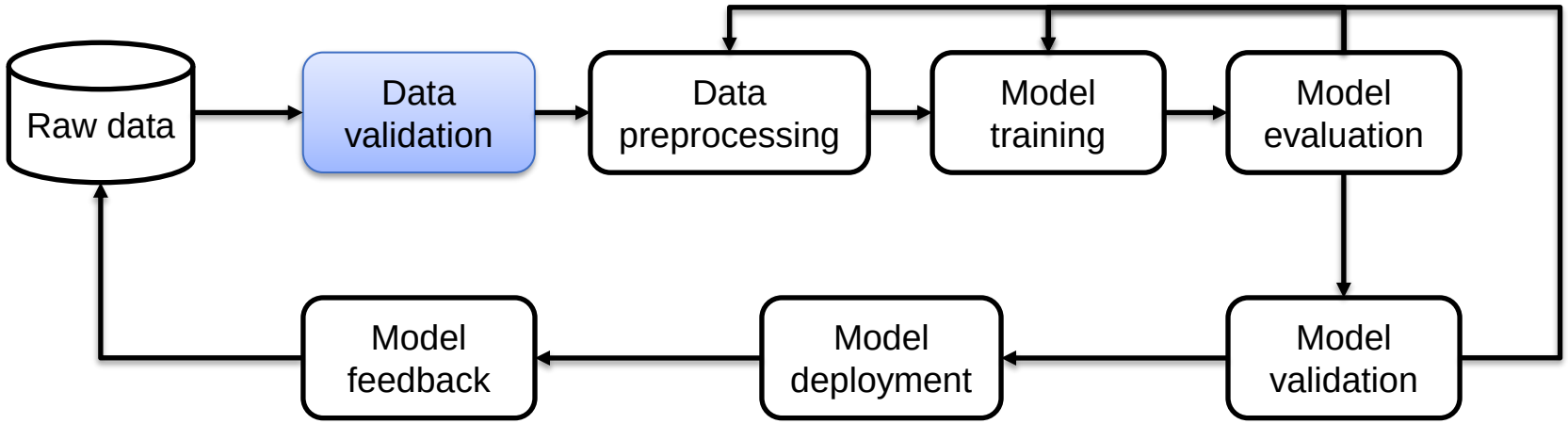
Thành phần của Machine Learning Pipeline



- Raw data:
  - Tổng hợp từ nhiều nguồn: log, database, survey,…
  - Chưa qua kiểm tra hoặc tiền xử lý
  - Có thể không đầy đủ, nhiễu, chứa thông tin nhạy cảm

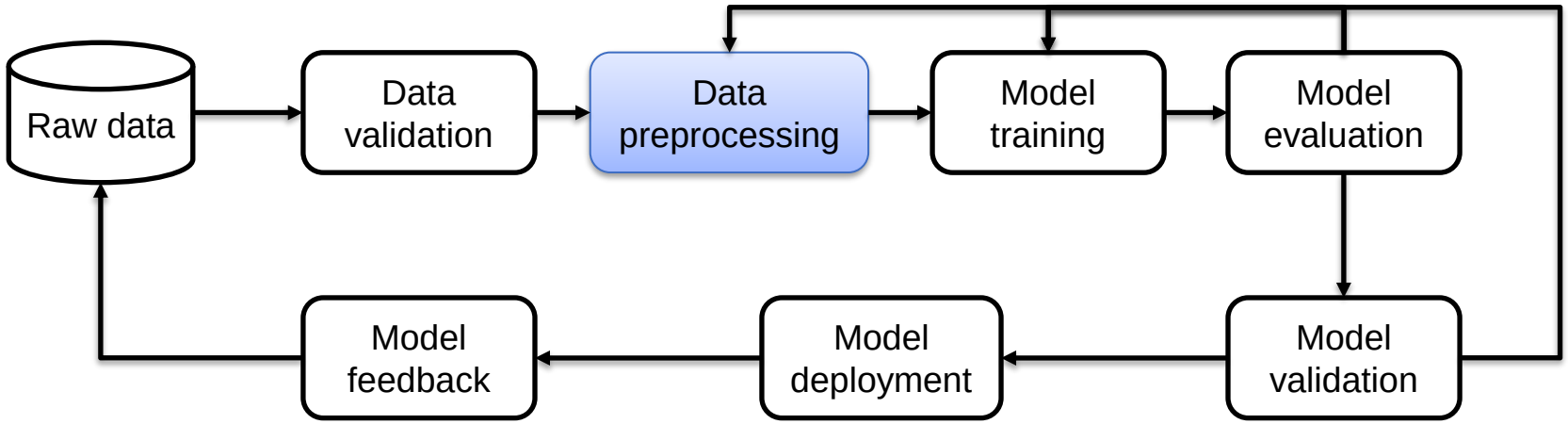
rò rỉ dữ liệu trong quá trình train thì lúc dùng cũng có thể mamats thông tin cá nhân

Thành phần của Machine Learning Pipeline



- Data validation: có vai trò rất quan trọng
  - Đảm bảo dữ liệu đầu vào phù hợp và chất lượng
  - **Kiểm tra**: tính nhất quán, tính đầy đủ, dữ liệu ngoại lệ
  - Thường được thực hiện với EDA (Exploratory Data Analysis)

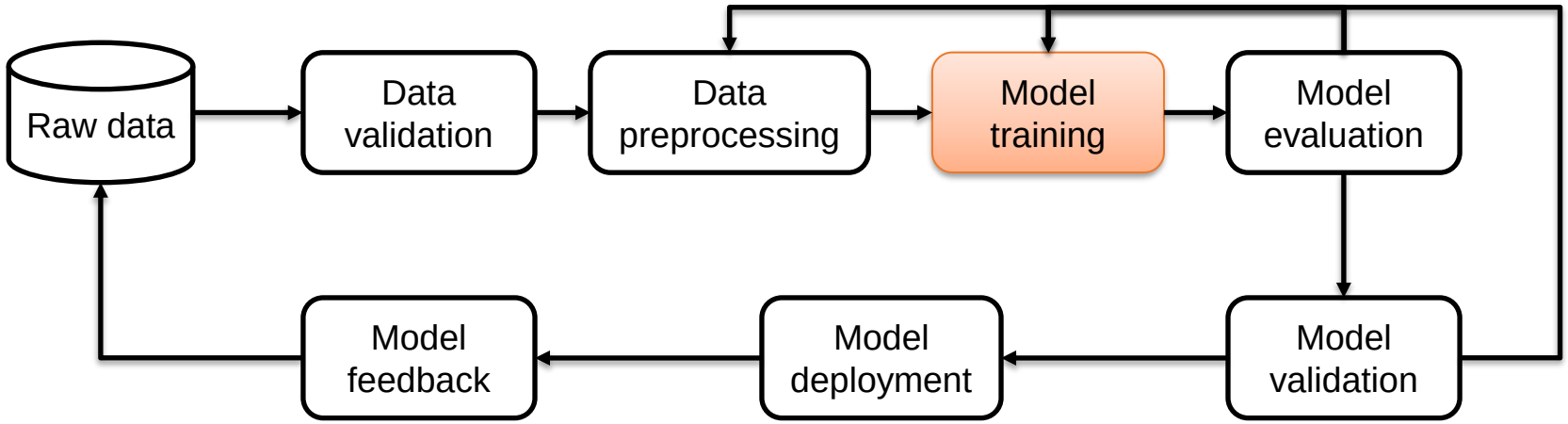
Thành phần của Machine Learning Pipeline



- Data preprocessing: thực hiện các công việc sau
  - Làm sạch dữ liệu, chuẩn hóa dữ liệu
  - Xử lý dữ liệu bị thiếu – Handle missing values
  - Xử lý dữ liệu ngoại lệ – Handle outlier values
  - Tạo mới đặc trưng – Feature extraction and transformation
  - Chọn lựa đặc trưng – Feature selection

Ví dụ: độ tuổi có thể là NAN hoặc 100
(nhiễu hay đúng là thật)

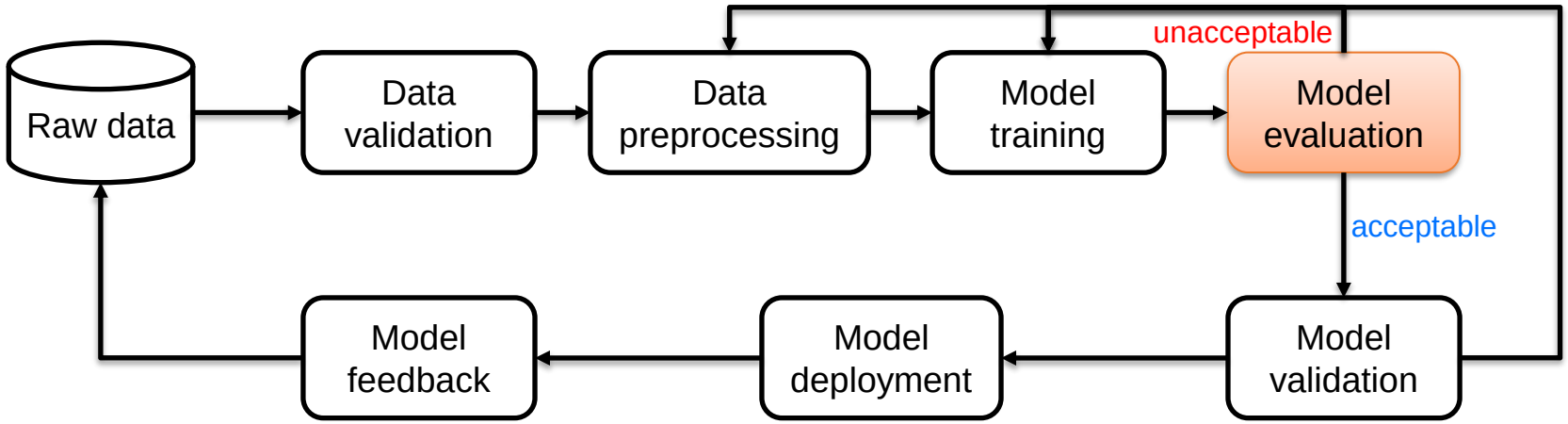
Thành phần của Machine Learning Pipeline



- Model training:
  - Khởi tạo mô hình với các tham số (nếu có)
  - Huấn luyện mô hình với dữ liệu (đặc trưng) đã chuẩn bị

KNN và NaiveBayes không có tham số

Thành phần của Machine Learning Pipeline

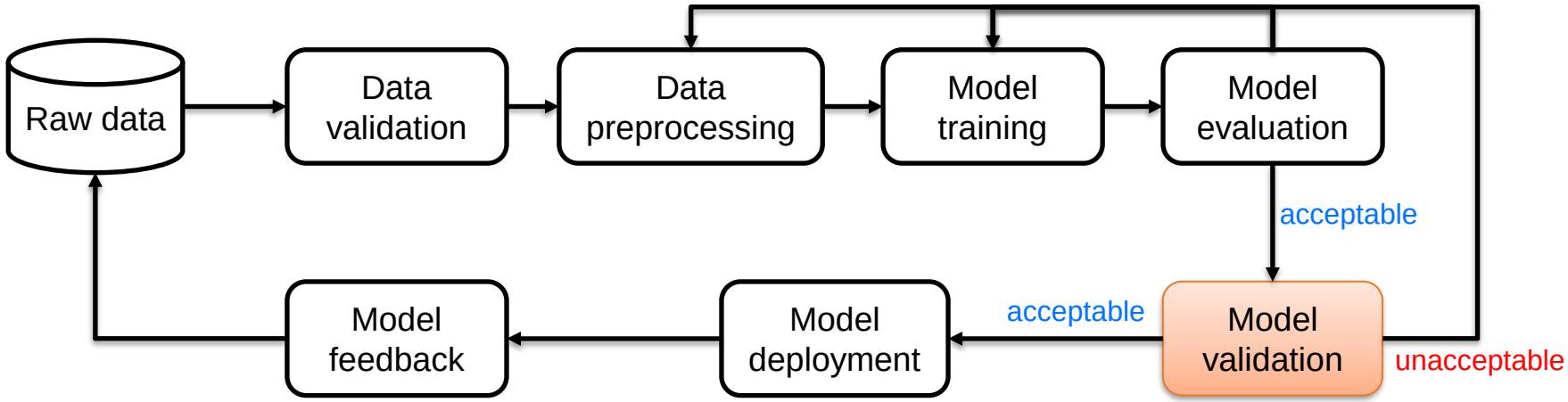


- Model evaluation: tiến hành đánh giá mô hình một cách định lượng
  - Dùng bộ dữ liệu thử nghiệm khách quan
  - Dùng độ đo (metric) cụ thể ra các con số
  - Nếu kết quả tốt: qua bước tiếp theo
  - Nếu kết quả không tốt: quay lại các pha trước

Đánh giá để đủ đi ra thực tế hay không, phải dùng nhiều độ đo quy trình khách quan và hoàn thiện

Dùng bộ dữ liệu chưa được thấy từ trước

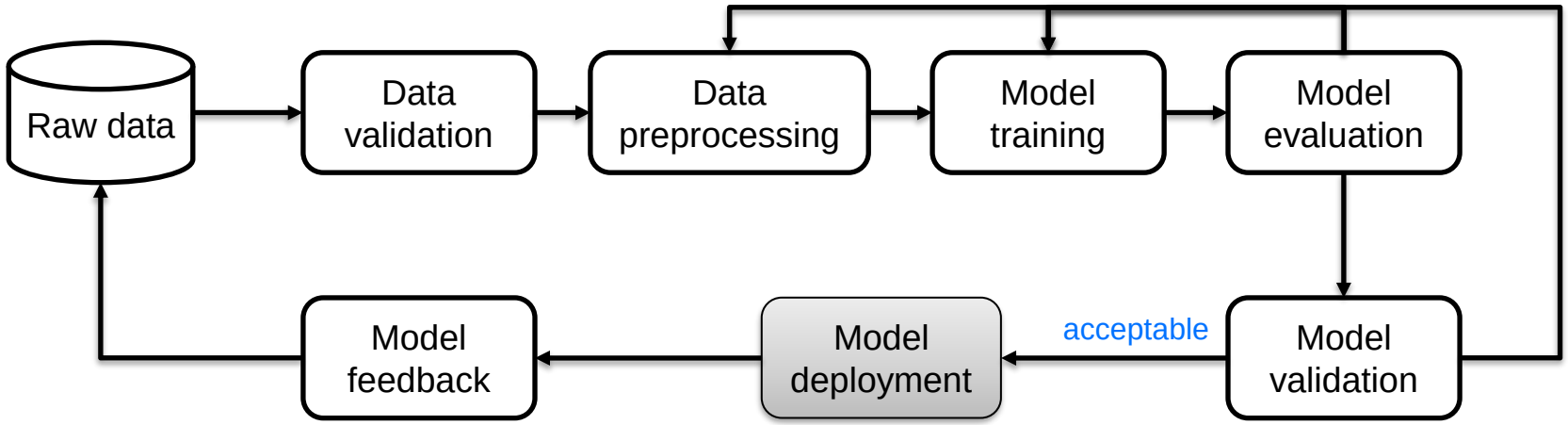
Thành phần của Machine Learning Pipeline



- Model validation: tiến hành kiểm định mô hình mang tính hệ thống
  - Thường được thực hiện độc lập với đội phát triển mô hình
  - Có chính xác trên dữ liệu chưa từng thấy?
  - Tốc độ thực thi được đảm bảo?
  - Có bị rò rỉ thông tin dữ liệu huấn luyện?

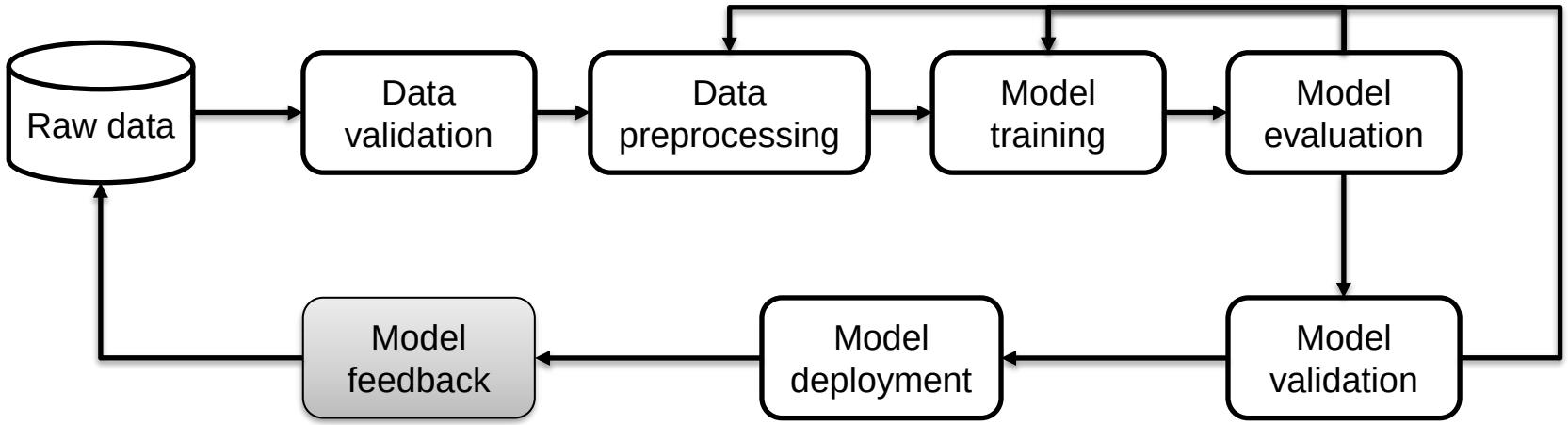
không chỉ đánh giá độ chính xác mà còn xem có đủ để đưa ra enduser không, thường là phía trên là ML engineer hoặc data engineer , AI,... còn ở dưới là đội ngũ khác để đo lường tiếp acc, tốc độ, rò rỉ dữ liệu, đáp ứng được hạ tầng hiện có (cost) hay không

Thành phần của Machine Learning Pipeline



- Model deployment: triển khai mô hình trên môi trường thực tế
  - Đóng gói mô hình
  - Chuyển giao mô hình

Thành phần của Machine Learning Pipeline



- Model feedback: theo dõi phản hồi về mô hình
  - **Phản hồi chủ quan**: nhận xét, đánh giá trực tiếp của người dùng
  - **Phản hồi khách quan**: thông qua các chỉ báo trung gian

nghĩa là theo dõi hành vi người dùng , có thể xem người dùng hay dùng hơn không, có tiện hơn, nhanh hơn không

NỘI DUNG

1. GIỚI THIỆU MACHINE LEARNING PIPELINE

2. CÁC THÀNH PHẦN CỦA MACHINE LEARNING PIPELINE

3. PHÂN TÍCH DỮ LIỆU – EXPLORATORY DATA ANALYSIS

Tại sao cần Phân tích dữ liệu

có thể dữ liệu bị thiếu hoặc nhiễu

- Để hiểu rõ dữ liệu:
  - # đặc trưng, # mẫu, loại đặc trưng, phân bố dữ liệu
  - Mối quan hệ hoặc xu hướng trong dữ liệu

đánh giá đc vai trò của dữ liệu, tính chất của dữ liệu, phát hiện vài vấn đề đang có trong bộ dự liệu

phân bố: dải giá trị, các giá trị biên có phải là ngoại lệ hay nhiễu không, mối quan hệ giữa các dữ liệu có mật thiết, xu hướng với input đầu vào không

Loại đặc trưng: số hay là phân loại/danh mục

- Làm sạch dữ liệu:
  - Xác định một số vấn đề: thiếu dữ liệu, trùng lặp, nhiễu, dữ liệu lỗi hoặc không nhất quán
  - Từ đó: xóa / điền khuyết / giữ nguyên

2 hàng giống i chang nhau nếu trùng lắp nhiều thì gây ra hiện tượng bias (học thuộc trùng lắp dữ liệu) => mô hình không còn tổng quát

Không nhất quán: ví dụ cùng 1 người công ty viết khác nhau ghi khác nhau nhưng ý nghĩa lại giống

Tại sao cần Phân tích dữ liệu

- Là tiền đề để:
  - Chọn lựa đặc trưng tốt cho bài toán, xây dựng mô hình
  - Có khả năng tạo ra đặc trưng mới tốt hơn
  - Chuẩn hóa đặc trưng về dạng phù hợp mô hình máy học

- Lường trước một số tình huống:
  - Mất cân bằng, thiên lệch dữ liệu
  - Một số vấn đề khác của dữ liệu: rò rỉ thông tin, vi phạm dữ liệu

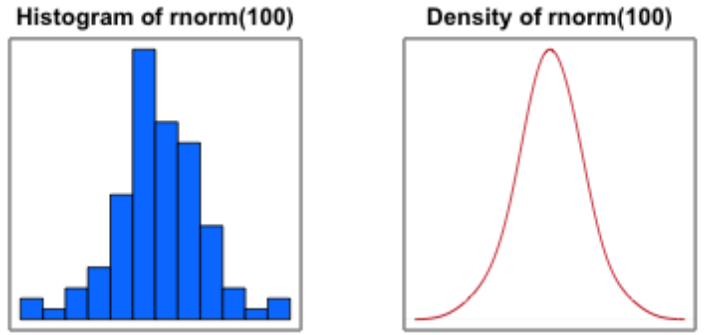
Ẩn danh hóa thông tin (anonymize)

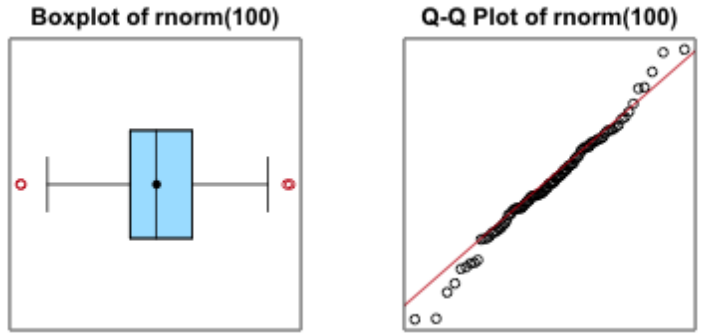
- **Là công cụ truyền thông**: trực quan hóa giúp giải thích vấn đề với người không chuyên môn

Công cụ thực hiện EDA

- Công cụ trực quan: histogram, box plot, ma trận tương quan

- Công cụ thống kê
  - Chỉ số thống kê: trung bình, trung vị, phương sai, độ lệch chuẩn
  - Phân tích tương quan: đơn biến, đa biến
  - Một số công cụ: hồi quy tuyến tính, hệ số tương đồng Pearson,…





Các kiểu phân tích dữ liệu

- Có 3 kiểu phân tích chính:
  - Phân tích đơn biến — Xem xét từng biến dữ liệu để phân tích độc lập
  - Phân tích hai biến — Xem xét tương quan theo từng cặp dữ liệu
  - Phân tích đa biến — lấy ra bộ các biến cùng phân tích 1 lúc

Phân tích đơn biến

- Chỉ quan tâm đến một đặc trưng / cột trong dữ liệu

- Phân tích bằng thống kê:
  - **Xu hướng tập trung (Central Tendency)**: các tham số ước lượng như trung bình, trung vị, mode
  - **Khoảng giá trị (Range)**: khoảng cách giữa giá trị **tối đa** và **tối thiểu** trong dữ liệu phân tích biên như các ý phía trên
  - **Phương sai và độ lệch chuẩn**

Phân tích đơn biến bằng thống kê

- Ví dụ về trung bình, độ lệch chuẩn, tứ phân vị thứ nhất, ba và trung vị của các biến trong dữ liệu

pandas

```
train[['Age', 'RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck']].describe()
```

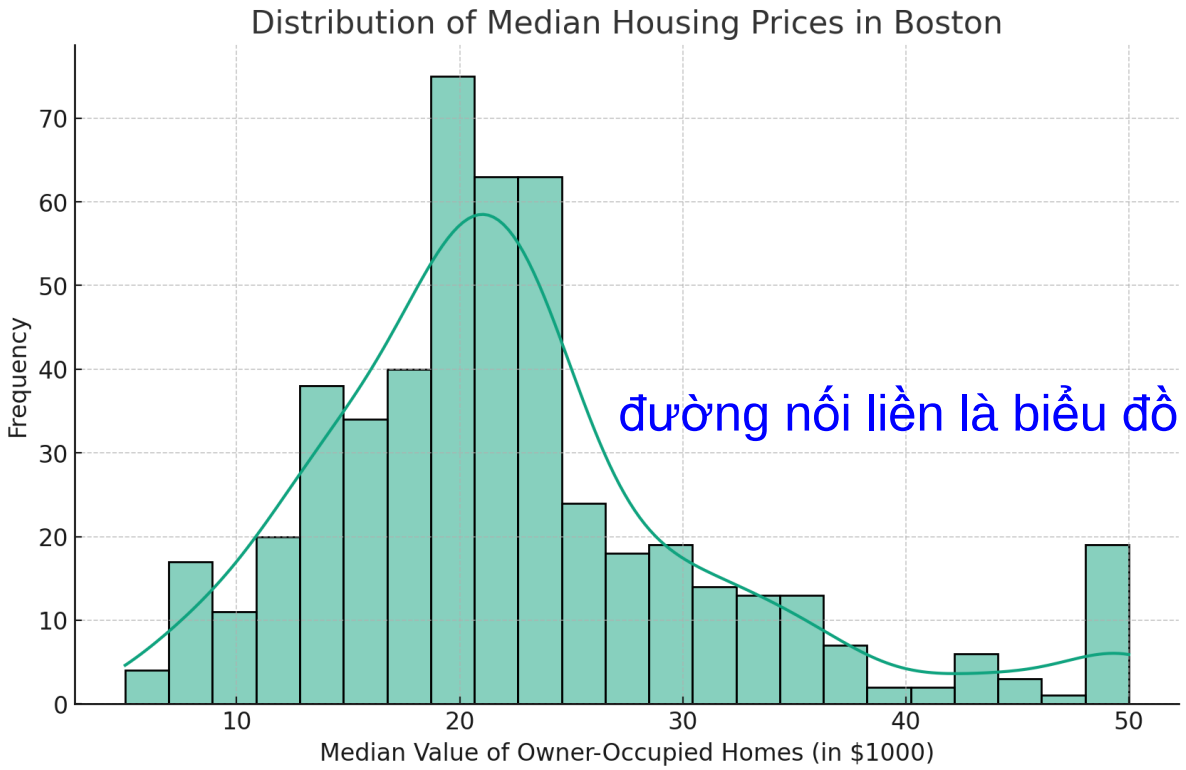
	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	8514.000000	8512.000000	8510.000000	8485.000000	8510.000000	8505.000000
mean	28.827930	224.687617	458.077203	173.729169	311.138778	304.854791
std	14.489021	666.717663	1611.489240	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000	27.000000	59.000000	46.000000
max	79.000000	14327.000000	29813.000000	23492.000000	22408.000000	24133.000000

lệch chuẩn

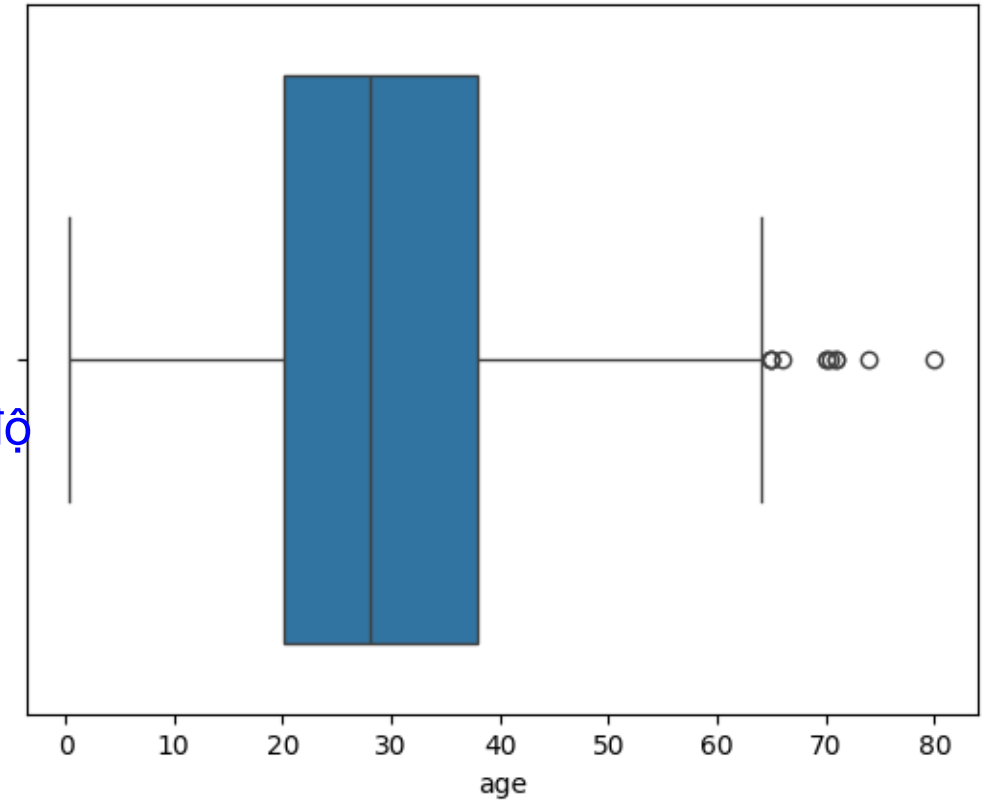
Phân tích đơn biến bằng biểu đồ

- Phân tích đơn biến sử dụng biểu đồ
  - **Biểu đồ Histogram**: biểu đồ trong đó tần số của dữ liệu được biểu thị bằng các thanh hình chữ nhật
  - **Biểu đồ mật độ**: một phiên bản liên tục của biểu đồ histogram
  - **Box-plot**: thông tin được thể hiện dưới dạng các hộp (giá trị nhỏ nhất, tứ phân vị thứ nhất (first quartile), trung vị (median), tứ phân vị thứ ba (third quartile), giá trị lớn nhất) vượt quá min- max thì là outlayer có thể là nhiễu

Histogram vs Box-plot



đường nối liền là biểu đồ mật độ

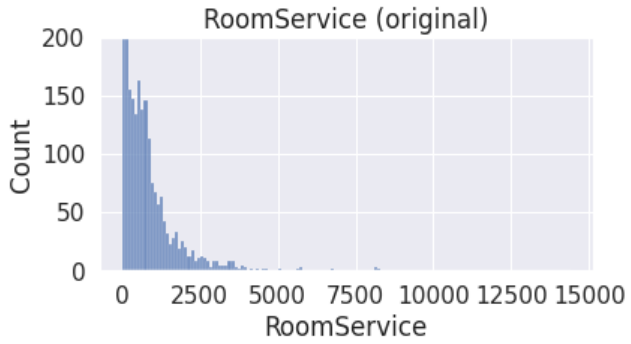


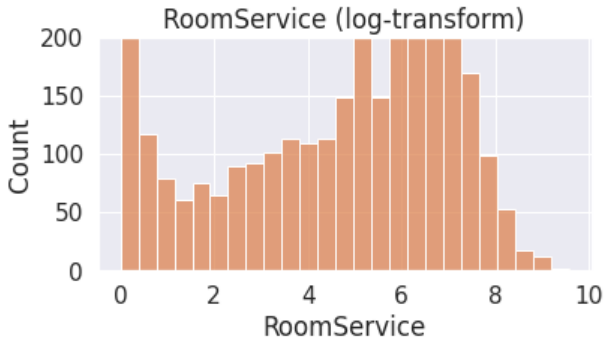
Biểu đồ tần suất về giá nhà

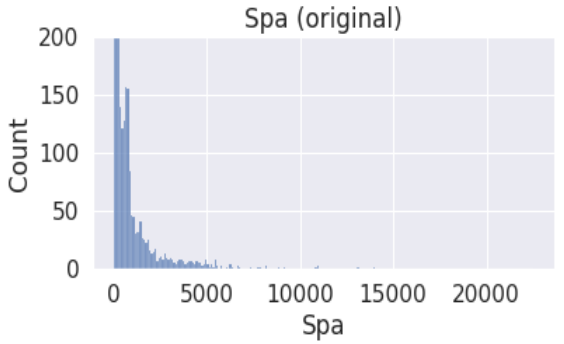
Box-plot của tuổi

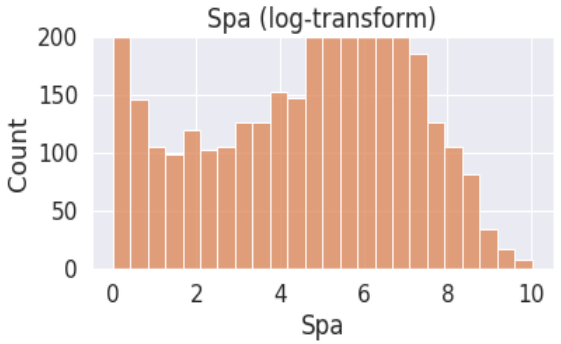
Ví dụ về phân tích đơn biến

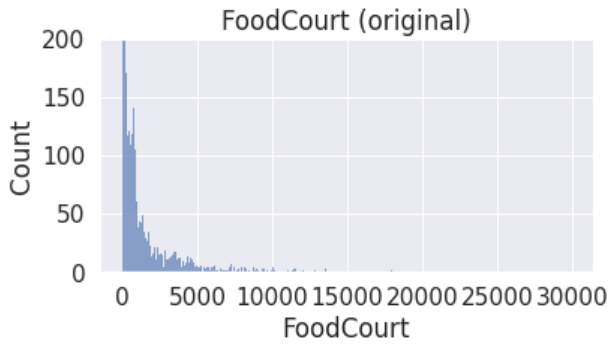
từ 1000 về sau thì giảm mạnh => thiên lệch

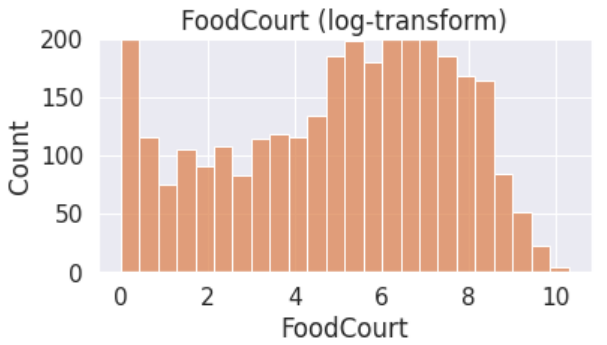


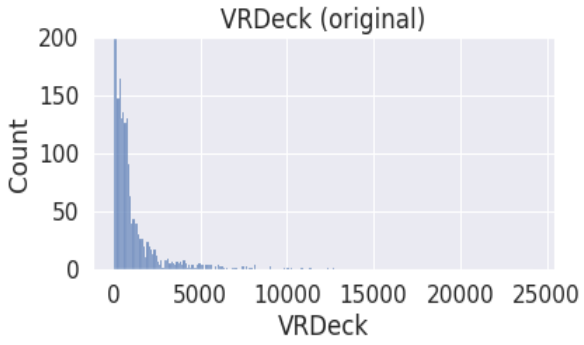


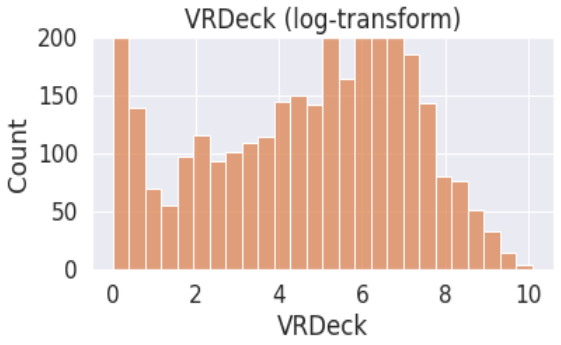


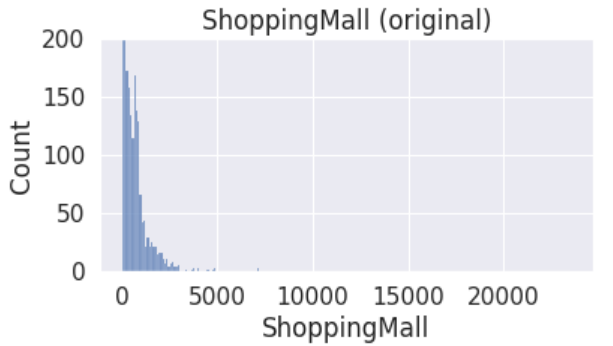


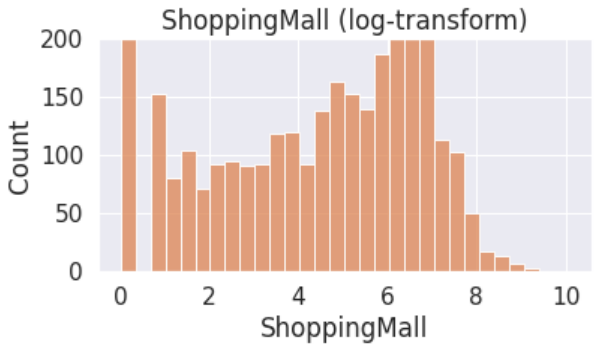


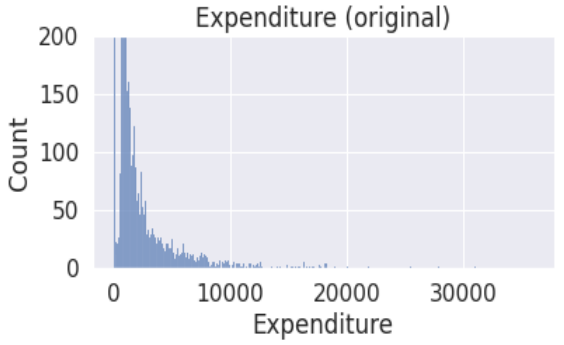


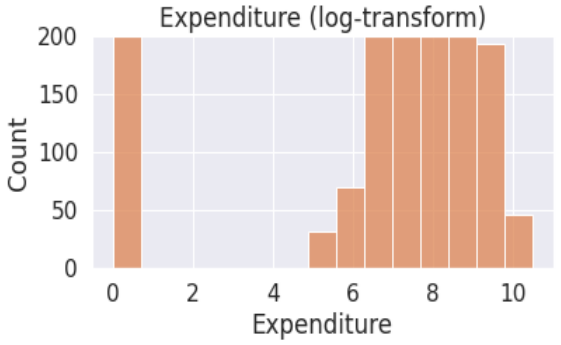












Phân tích hai biến

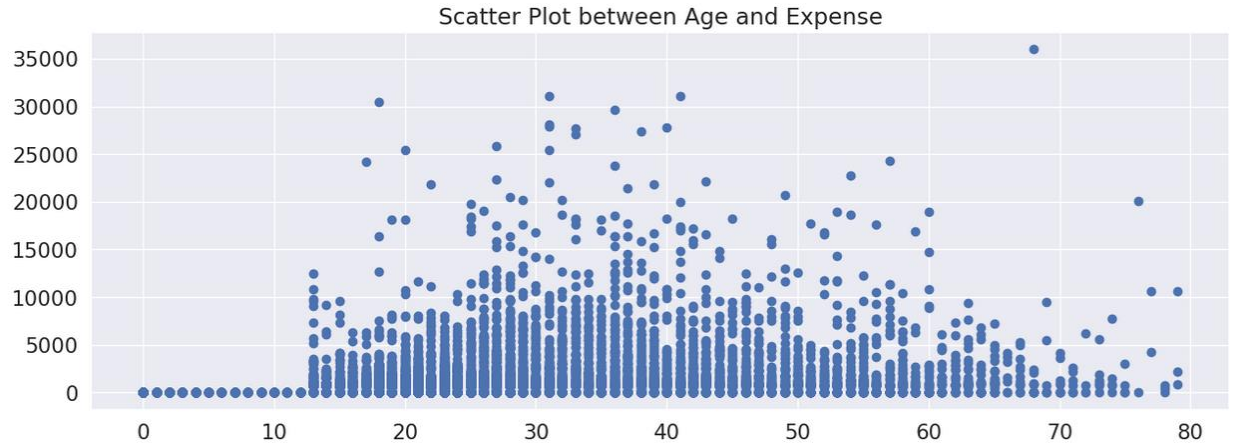
- Xác định xem có tồn tại mối liên hệ thống kê giữa hai biến hay không

- Có ba loại chính:
  - Phân tích dạng số-số (numeric – numeric)
  - Phân tích dạng số-phân loại (numeric – category)
  - Phân tích dạng phân loại-phân loại (category – category)

Phân tích dạng số - số

- Khi cả hai biến được so sánh đều là dữ liệu số
- Một số phương pháp trực quan có thể được sử dụng:
  - **Biểu đồ phân tán (Scatter plot):** thể hiện mọi điểm dữ liệu lên biểu đồ
  - **Biểu đồ cặp (Pair plot)**
  - **Ma trận tương quan (Correlation matrix)**

mỗi ô thể hiện tương quan của 2 thuộc tính



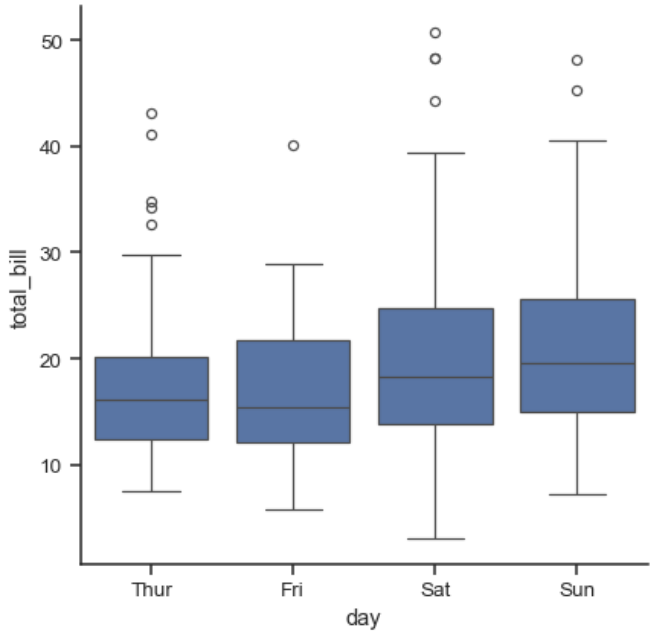
Biểu đồ phân tán giữa tuổi và chi tiêu

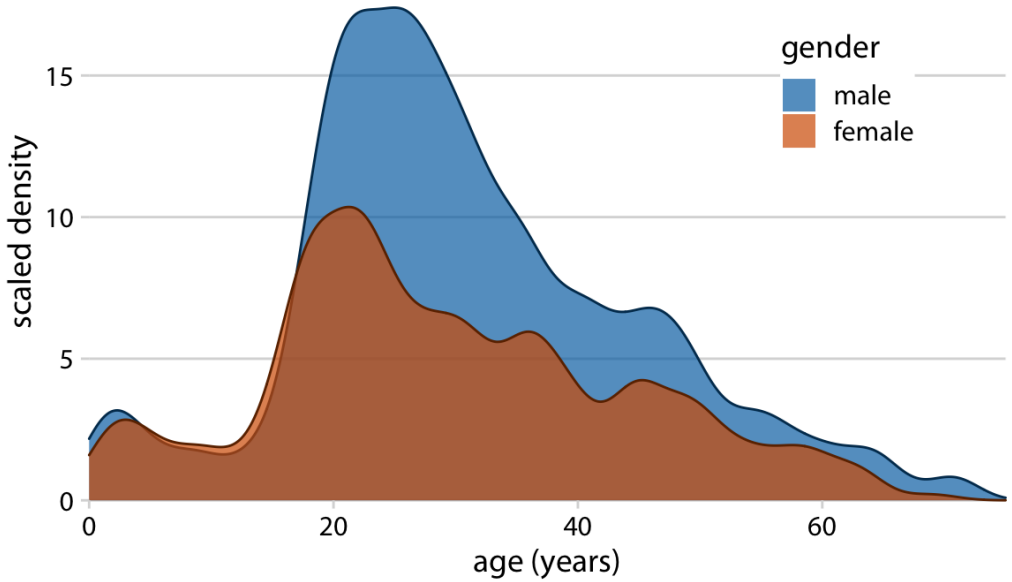
Ma trận tương quan kết hợp bản đồ nhiệt

Phân tích dạng số - phân loại

- Khi một biến kiểu số và biến khác kiểu phân loại
- Một số phương pháp trực quan có thể được sử dụng:
  - Multiple density / histogram
  - Box-plot

mỗi box là phân loại



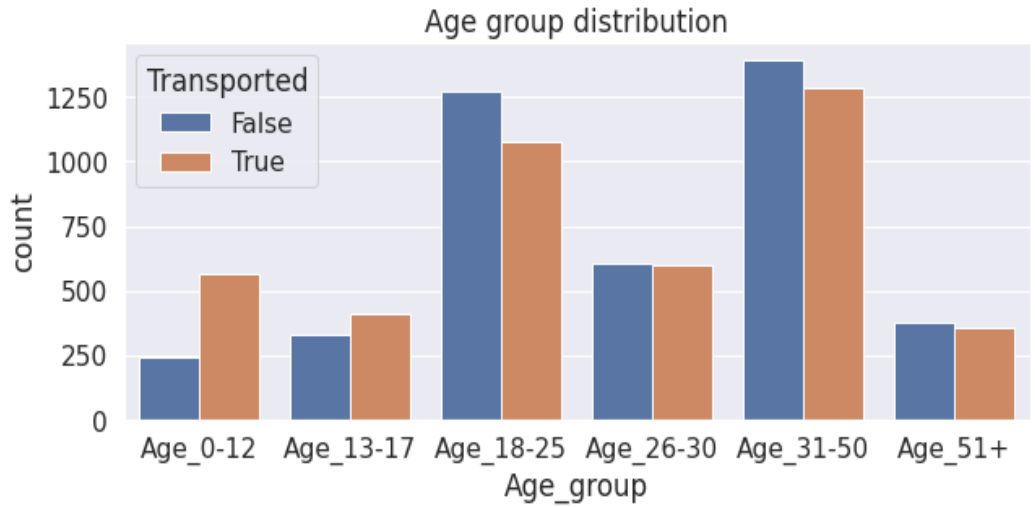


Phân tích dạng phân loại – phân loại

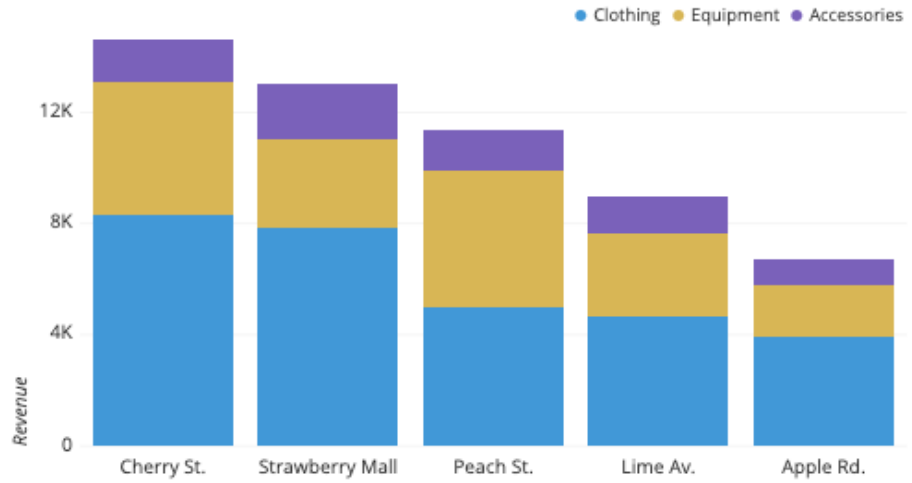
- Khi cả hai biến đều có kiểu phân loại

- Một số phương pháp trực quan có thể được sử dụng:
  - Biểu đồ cột xếp chồng (Stacked Bar Chart)
  - Biểu đồ cột theo cụm (Clustered Bar Chart)

nghĩa là bên đây là xếp theo chồng lên cột, còn hình trái là xếp bên cạnh



Ví dụ về biểu đồ theo cụm

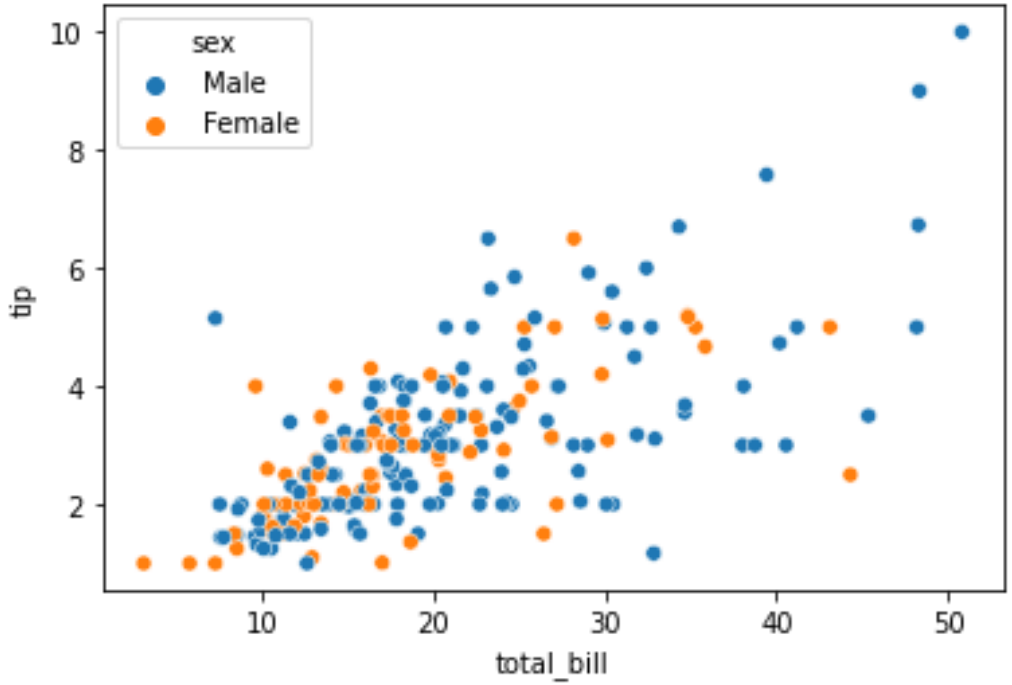


Ví dụ về biểu đồ xếp chồng

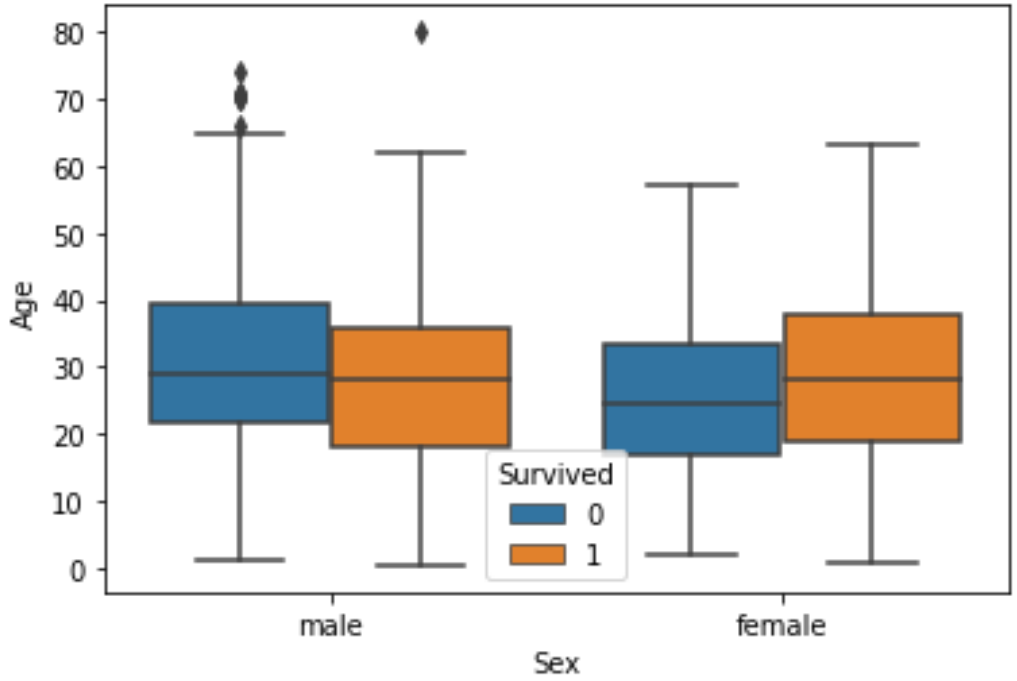
Phân tích đa biến

- Khi phân tích nhiều hơn 2 biến cùng lúc
- Để trực quan hóa cần sử dụng nhuần nhuyễn các loại biểu đồ tương ứng với các tổ hợp biến dạng số và dạng phân loại

số - số - phân loại



số - phân loại - phân loại



Phát hiện dữ liệu bị thiếu

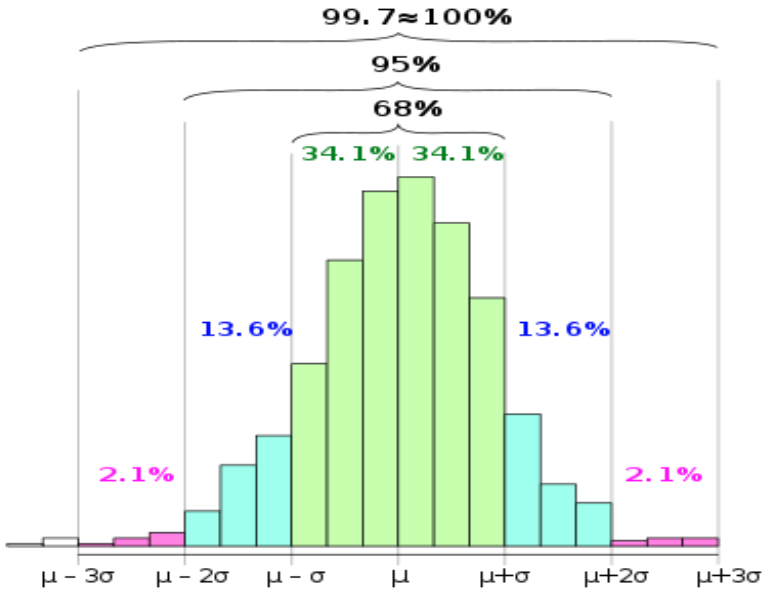
- Sử dụng hàm .isnull() của dataframe
- Có thể loại bỏ dòng có giá trị NaN hoặc rỗng với hàm drop_nan()

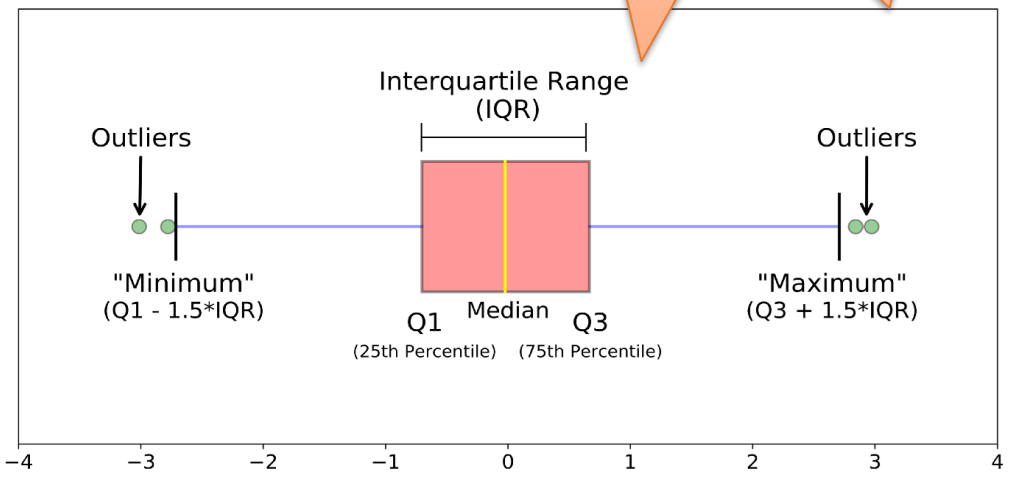
	Name	Age	Place	College
a	Ankit	22.0	Up	Geu
b	Ankita	NaN	Delhi	NaN
c	Rahul	16.0	Tokyo	Abes
d	Simran	41.0	Delhi	Gehu
e	Shaurya	NaN	Delhi	Geu
f	Shivangi	35.0	Mumbai	NaN
g	Swapnil	35.0	NaN	Geu
i	NaN	35.0	Uk	Geu
j	Jeet	35.0	Guj	Gehu
k	NaN	NaN	NaN	NaN

Phát hiện dữ liệu nhiễu

- Phương pháp thống kê:
  - **Phương pháp tính trung bình và độ lệch chu**: để xác h có a tri ngoại lệ (với dữ liệu dạng Gaussian hoặc tương tự
  - **Phương pháp Interquartile Range (IQR):** để xá với dữ liệu phân phối không phải Gaussian

Xem thêm phần Data preprocessing





BÀI QUIZ VÀ HỎI ĐÁP