



Bài giảng môn học:  
**Khai phá dữ liệu (7080508)**

## **CHƯƠNG 2: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU (Phần 3)**

**08/2024**



# Nội dung chương 2 – Phần 3

---

**1.1 Tại sao phải xử lý, làm sạch dữ liệu**

**1.2 Thể nào là một tập dữ liệu sạch**

**1.3 Một số kỹ thuật làm sạch quan trọng**

**1.4 Quy trình tổng quan để làm sạch và nâng cao chất lượng dữ liệu**

**1.5 Bài tập tổng hợp**

## 1.1 Tại sao phải xử lý, làm sạch dữ liệu

# Tại sao phải tiền xử lý, làm sạch dữ liệu?





# Tại sao phải tiền xử lý, làm sạch dữ liệu?

“Rác vào – Rác ra (Garbage in, garbage out)” là nguyên lý cơ bản của phân tích dữ liệu. Chúng ta không thể mong muốn tìm ra được các thông tin quan trọng, có ích khi đầu vào của nó là các dữ liệu không đầy đủ, chính xác, dữ liệu thiếu đồng bộ, nhất quán và chứa nhiều nhiễu



Dữ liệu sạch và chính xác có vai trò đặc biệt quan trọng trong phân tích dữ liệu, vì việc sử dụng các tập dữ liệu kém chất lượng có thể dẫn tới các thông tin, dự đoán bị sai lệch.

Đây là lý do chính khiến các nhà phân tích dành phần lớn thời gian (80%) trong bất kỳ một dự án phân tích dữ liệu nào.



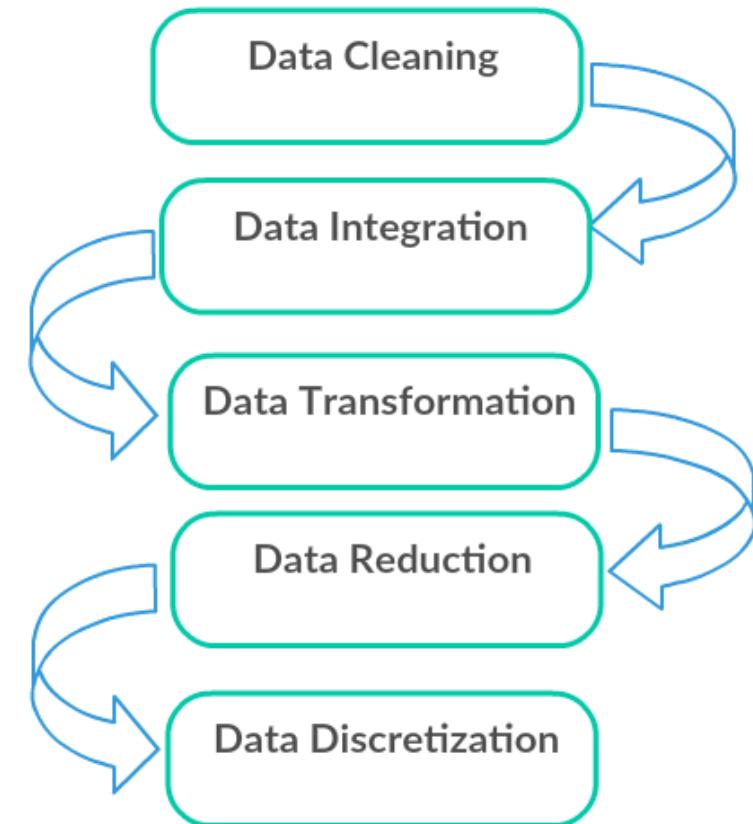
# Tại sao phải tiền xử lý, làm sạch dữ liệu?

**Làm sạch dữ liệu (Data cleaning)** là quá trình chuẩn bị dữ liệu để phân tích bằng cách loại bỏ các thông tin không liên quan, hoặc không chính xác, không đầy đủ có thể làm sai lệch kết quả và gây ra các quyết định sai lầm, không thực tế.

Con đường để biến một tập dữ liệu thô (Raw Data) thành một tập dữ liệu sẵn sàng cho phân tích (Clean Data) là một chặng đường dài.



Không có phương pháp, kỹ thuật làm sạch nào chung áp dụng cho mọi loại dữ liệu, mọi bài toán. Với mỗi một loại dữ liệu, bài toán cụ thể sẽ có các phương pháp xử lý phù hợp.





1.2 Thế nào là một tập dữ liệu  
sạch



# Thế nào là “Dữ liệu sạch”



Đích cuối cùng của quá trình làm sạch và chuẩn bị dữ liệu là một tập dữ liệu “sạch” sẵn sàng cho phân tích. Câu hỏi đặt ra là: **Thế nào là dữ liệu sạch?**

**Dữ liệu sạch** là phần thông tin đáp ứng các yêu cầu về chất lượng của dữ liệu và sẽ góp phần khám phá những hiểu biết có giá trị hỗ trợ việc ra quyết định, thúc đẩy hiệu quả sản xuất, kinh doanh...

Một tập dữ liệu sạch là tập dữ liệu có chất lượng đáp ứng được các thuộc tính:



# Thế nào là “Dữ liệu sạch”



- **Tính chính xác (Accuracy):** Được đo bằng mức độ tương thích giữa dữ liệu sau xử lý với dữ liệu thật từ các nguồn thu thập được. Đây là “**Tiêu chuẩn vàng – Golden standard**” của tập dữ liệu sạch chất lượng.
- **Tính đầy đủ (Completeness):** Đáp ứng đầy đủ các dữ liệu, thuộc tính cho bài toán, vấn đề cần giải quyết.
- **Tính nhất quán (Consistency):** Phản ánh khả năng tương thích của dữ liệu sạch với toàn bộ hệ thống.
- **Tính liên quan (Relevance):** Xác định mức độ liên quan mật thiết của các thông tin đến vấn đề đang giải quyết

# Thế nào là “Dữ liệu sạch”



- **Tính kịp thời (Timeliness):** Thể hiện mức độ “mới” và cập nhật kịp thời các dữ liệu khi có thay đổi
- **Tính hiệu lực (Validity):** Dữ liệu phải tuân theo quy tắc hoặc ràng buộc đã xác định.
- **Tính đồng nhất (Uniformity):** Thể hiện sự nhất quán của các đơn vị đo lường trong toàn bộ hệ thống.

## **1.3 Một số kỹ thuật làm sạch dữ liệu, nâng cao chất lượng dữ liệu quan trọng**

### **1.3.1 Kiểm tra mức độ tương thích dữ liệu (Data Comability)**



# Data Comability

Dữ liệu được thu thập từ rất nhiều nguồn, nhiều định dạng khác nhau. Do đó, trong thực tế sẽ có nhiều vấn đề phát sinh đối với dữ liệu.

Chúng ta chỉ so sánh các đối tượng trong cùng một hệ quy chiếu:



So sánh số lượng máy tính Macbook với máy tính Dell bán được trong tháng, khi đó sự so sánh được coi là công bằng, vì 2 mặt hàng có sự tương thích với nhau



Sự so sánh số lượng máy tính Macbook với số lượng xe ô tô điện Vinfast bán được trong tháng trở nên vô nghĩa



# Data Comability

Các đối tượng xem xét phải có sự tương thích:

- Khi so sánh trọng lượng 123.5 (**pound**) với 78.9 (**kg**).
- Khi so sánh doanh thu của bộ phim Cuốn theo chiều gió (**1939**) với bộ phim Avatar (**2009**) khi mà giá trị Đô la năm 2009 gấp 15.43 lần năm 1939.
- Khi so sánh giá vàng vào buổi trưa hôm nay tại **New York** và **London**, khi mà 2 thành phố này lệch nhau 5 tiếng và giá vàng còn bị ảnh hưởng bởi nhiều yếu tố khác.



Vấn đề mấu chốt là phải hiểu tập dữ liệu đang có, ý nghĩa và đơn vị tính của từng thuộc tính



# Data Comability

**Một số lưu ý để đảm bảo tính tương thích của dữ liệu:**

- **Chuẩn hóa đơn vị đo:** Cần phải kiểm tra và thống nhất các đơn vị đo với cùng một thuộc tính trong tập dữ liệu như: đơn vị chiều dài, trọng lượng, diện tích, tiền tệ...
- **Chuẩn hóa dữ liệu số:** Dữ liệu số có thể ở nhiều dạng khác nhau như số nguyên – 123 (integers), số thực 123.5 (decimals), phân số 123 ½ (fractions), thậm chí có thể được biểu diễn ở dạng văn bản ...Do vậy, trong cùng một thuộc tính cần phải thống nhất cùng một kiểu dữ liệu số.
- **Chuẩn hóa định danh:** Cùng một giá trị nhưng trong các tập dữ liệu khác nhau có thể được thể hiện ở những dạng khác nhau, ví dụ như giới tính có thể được thể hiện: Male – Female, M – F, Nam – Nữ, 0 – 1, MALE – FEMALE, male – female...hoặc địa chỉ: Hà Nội, HN, hà nội, HÀ NỘI...
- **Chuẩn hóa dữ liệu thời gian:** Cần thống nhất múi giờ (GMT, UTC) khi xử lý dữ liệu thời gian, chuẩn hóa định dạng dữ liệu thời gian YYYYMMDD – DDMMYYYY – MMDDYYYY – DDMMYY...





# Data Comability

- Mã nhân viên (ID),
- Tên nhân viên (Name),
- Giới tính (Gender),
- Ngày sinh (Birthday),
- Lương tháng (Salary),
- Phòng ban (Department),
- Địa chỉ (Adress)



Company - 1

ID	Name	Gender	Birthday	Salary (USD)	Department	Adress
MK011	Jonh	Male	11/08/1984	1530	Marketing	New York
MK032	Roses	Female	02/05/1980	1250	Marketing	Los Angeles
SL010	Smith	Male	13/03/1999	1175	Sales	New York
SL010	Tom	Male	17/06/2000	1435	Sales	San Francisco
MK125	Daniel	Female	21/11/1995	1045	Marketing	Los Angeles

Company - 2

ID	Name	Gender	Birthday	Salary (1000 USD)	Department	Adress
mk001	alex	m	08/25/1981	1.345	mk	new york
sl123	blake	m	02/15/1998	1.75	sl	new york
sl203	susan	f	11/19/1990	1.015	sl	san francisco
mk734	Lisa	f	02/14/2001	0.98	mk	san francisco
sl066	Mary	f	12/12/1997	1.45	sl	los angeles

Cho biết các vấn đề cần xử lý để dữ liệu của 2 bảng này tương thích với nhau?

### **1.3.2 Xử lý dữ liệu khuyết thiếu (Missing Data)**



# Missing Data

Không phải tất cả các bộ dữ liệu đều đầy đủ, giá trị khuyết thiếu (missing values) là vấn đề rất thường gặp trong dữ liệu, việc phát hiện và xử lý chúng là yêu cầu quan trọng trong quá trình làm sạch dữ liệu

Complete data		MAR		MCAR		MNAR	
Age	IQ score	Age	IQ score	Age	IQ score	Age	IQ score
25	133	25		25		25	133
26	121	26		26	121	26	121
29	91	29		29	91	29	
30	105	30		30		30	
30	110	30		30	110	30	110
31	98	31		31		31	
44	118	44	118	44	118	44	118
46	93	46	93	46	93	46	
48	141	48	141	48		48	141
51	104	51	104	51		51	
51	116	51	116	51	116	51	116
54	97	54	97	54		54	

- **Dữ liệu thiếu ngẫu nhiên (Missing at Random – MAR):** Sự mất mát dữ liệu ở đây là ngẫu nhiên, tuy nhiên vẫn có mối quan hệ hệ thống giữa dữ liệu bị mất và dữ liệu được quan sát.
- **Dữ liệu thiếu hoàn toàn ngẫu nhiên (Missing Completely at Random – MCAR):** Sự mất mát dữ liệu tại các điểm quan sát là hoàn toàn ngẫu nhiên, và không có bất kỳ một mối quan hệ hay sự liên hệ nào giữa dữ liệu thiếu với các dữ liệu quan sát khác.
- **Dữ liệu thiếu không ngẫu nhiên (Missing Not at Random – MNAR):** Sự mất mát dữ liệu không phải là ngẫu nhiên mà có một mối quan hệ xu hướng giữa giá trị bị thiếu và giá trị không bị thiếu trong một biến.

# Missing Data

Các giá trị dữ liệu khuyết thiếu xuất hiện dưới nhiều dạng khác nhau, phổ biến nhất là dữ liệu trống (rỗng), hoặc có thể được ký hiệu là NA, N/A, -1, 99, 999... Do đó, để có thể phát hiện ra dữ liệu khuyết thiếu trong tập dữ liệu cần phải hiểu rõ dữ liệu đang xử lý, các thuộc tính và kiểu dữ liệu thiếu được biểu diễn ứng với từng thuộc tính

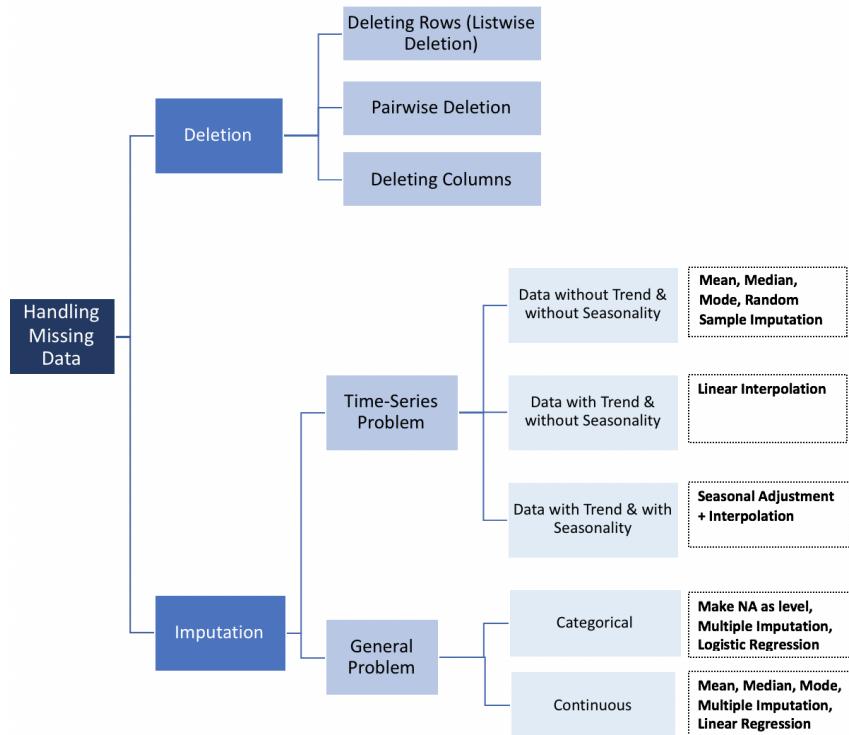
A	B	C	D	E	F	G	
1	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
2	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
3	01 15-9-2019		24.21	24.02	24.93	25.16	24.83
4	02 15-9-2019	25.05	23.73	23.89	24.79	24.8	24.55
5	03 15-9-2019	24.79	23.36	23.83		24.74	24.48
6	04 15-9-2019	24.59	23.05	23.69	24.82	24.8	24.38
7	05 15-9-2019	24.4		23.52	24.79	24.87	24.4
8	06 15-9-2019	24.38	22.79	23.68	25.1	24.71	24.41
9	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
10	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
11	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
12	10 15-9-2019		29.97			27.68	27.53
13	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
14	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
15	13 15-9-2019	30.95		27.83	27.44	28	30.66
16	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
17	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
18	16 15-9-2019	30.8	30.2		26.45	27.29	29.13
19	17 15-9-2019	29.94	29.36	25.8	26.67	26.69	28.72
20	18 15-9-2019	28.53	27.48	24.82	25.92	25.81	27.46
21	19 15-9-2019	28.89	27.03	24.93	25.88	25.93	27.07
22	20 15-9-2019	28.06	26.41	24.7		25.97	26.75
23	21 15-9-2019	27.43	26.2	24.41	25.62	25.94	26.32
24	22 15-9-2019	26.98	25.79	24.17	25.6	25.9	26.29
25	23 15-9-2019	26.68	25.31	23.81	25.53	25.8	26.36
26							
27							

A	DEPTH	GR	NPHI	RHOB	RT
~A	500.1000	64.5768	-999.0000	-999.0000	1.3877
	500.2524	63.8732	-999.0000	-999.0000	1.1355
	500.4048	63.1720	-999.0000	-999.0000	0.8908
	500.5572	62.5449	-999.0000	-999.0000	0.8785
	500.7096	61.8662	-999.0000	-999.0000	0.8731
	500.8620	60.4208	-999.0000	-999.0000	0.9706
	501.0144	59.0552	-999.0000	-999.0000	1.0608
	501.1668	58.4555	-999.0000	-999.0000	1.0807
	501.3192	57.9431	-999.0000	-999.0000	1.0818
	501.4716	58.0369	-999.0000	-999.0000	0.9526
	501.6240	58.2013	-999.0000	-999.0000	0.8494
	501.7764	58.7432	-999.0000	-999.0000	0.8856
	501.9288	59.5950	-999.0000	-999.0000	0.9073
	502.0812	61.7766	-999.0000	-999.0000	0.8668
	502.2336	63.8182	-999.0000	-999.0000	0.8575
	502.3860	65.3647	-999.0000	-999.0000	0.9586
	502.5384	66.5884	-999.0000	-999.0000	1.0321
	502.6908	66.8538	-999.0000	-999.0000	1.0236
	502.8432	67.2312	-999.0000	-999.0000	0.9932
	502.9956	67.8915	-999.0000	-999.0000	0.9071
	503.1480	68.2685	-999.0000	-999.0000	0.8446
	503.3004	68.0294	-999.0000	-999.0000	0.8334
	503.4528	67.1629	-999.0000	-999.0000	0.8970
	503.6052	65.1127	-999.0000	-999.0000	1.1015
	503.7576	63.1847	-999.0000	-999.0000	1.2438
	503.9100	61.4579	-999.0000	-999.0000	1.2833

Việc **phát hiện và thống kê dữ liệu thiếu không khó**, sử dụng các công cụ, hoặc các hàm xây dựng có thể dễ dàng phát hiện ra chúng.

# Missing Data

Sau khi phát hiện dữ liệu khuyết thiếu, cần phải xử lý chúng sao cho phù hợp. Việc lựa chọn phương pháp nào phụ thuộc vào từng thuộc tính, loại dữ liệu và bài toán cụ thể. Các phương pháp xử lý dữ liệu thiếu cơ bản được phân thành 2 nhóm chính:



- **Loại bỏ các dữ liệu thiếu (Deletion):** Xóa bỏ các bản ghi (rows) hoặc các thuộc tính (columns) chứa dữ liệu thiếu; Phương áp này áp dụng khi lượng dữ liệu thiếu ở một thuộc tính hoặc bản ghi nào đó lớn, thuộc tính chứa dữ liệu thiếu không quan trọng; dữ liệu thiếu là hoàn toàn ngẫu nhiên, ...Đây là phương pháp xử lý rất đơn giản và nhanh chóng thực hiện bởi các thư viện hỗ trợ.
- **Thay thế các giá trị thiếu bằng giá trị phù hợp (Imputation):** Trong nhiều trường hợp không thể hoặc không cần xóa bỏ các bản ghi, thuộc tính chứa dữ liệu khuyết thiếu, có thể thay thế nó bằng một giá trị phù hợp. Việc lựa chọn giá trị thay thế vào các vị trí dữ liệu thiếu được xác định sao cho khả năng gần đúng nhất (sai số nhỏ nhất) giữa giá trị thay thế với giá trị thật.



# Missing Data

Xử lý khuyết thiếu cho dữ liệu chuỗi thời gian  
(Time series Data)

	A	B	C	D	E	F	G
1	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
2	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
3	01 15-9-2019		24.21	24.02	24.93	25.16	24.83
4	02 15-9-2019	25.05	23.73	23.89	24.79	24.8	24.55
5	03 15-9-2019	24.79	23.36	23.83		24.74	24.48
6	04 15-9-2019	24.59	23.05	23.69	24.82	24.8	24.38
7	05 15-9-2019	24.4		23.52	24.79	24.87	24.4
8	06 15-9-2019	24.38	22.79	23.68	25.1	24.71	24.41
9	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
10	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
11	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
12	10 15-9-2019		29.97			27.68	27.53
13	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
14	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
15	13 15-9-2019	30.95		27.83	27.44	28	30.66
16	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
17	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
18	16 15-9-2019	30.8	30.2		26.45	27.29	29.13
19	17 15-9-2019	29.94	29.36	25.8	26.67	26.69	28.72
20	18 15-9-2019	28.53	27.48	24.82	25.92	25.81	27.46
21	19 15-9-2019	28.89	27.03	24.93	25.88	25.93	27.07
22	20 15-9-2019	28.06	26.41	24.7		25.97	26.75
23	21 15-9-2019	27.43	26.2	24.41	25.62	25.94	26.32
24	22 15-9-2019	26.98	25.79	24.17	25.6	25.9	26.29
25	23 15-9-2019	26.68	25.31	23.81	25.53	25.8	26.36
26							
27							

Xử lý khuyết thiếu cho dữ liệu rời rạc:

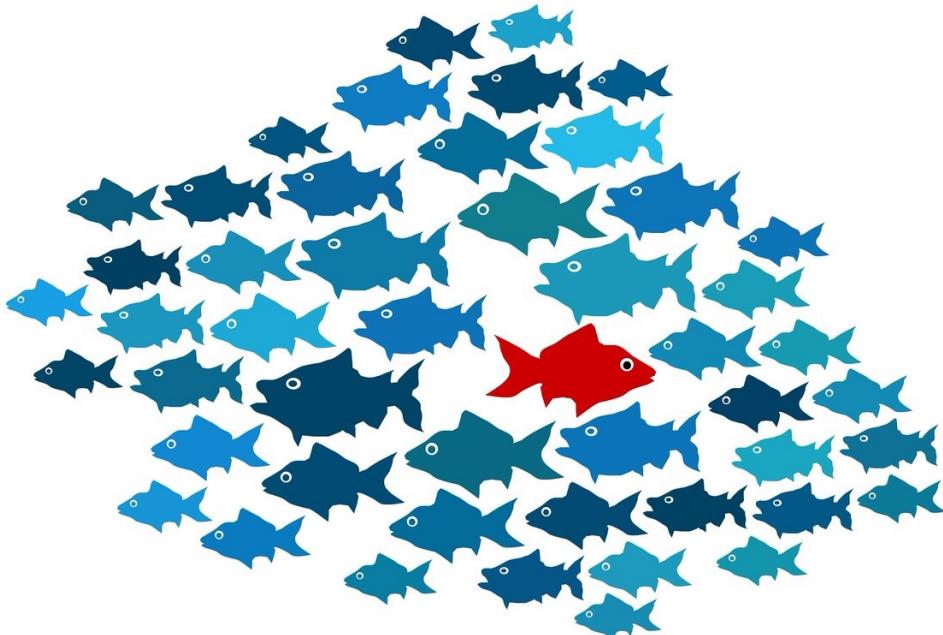
Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
male	22.0	1	0	A/5 21171	7.2500	NaN	S
female	38.0	1	0	PC 17599	71.2933	C85	C
female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
female	35.0	1	0	113803	53.1000	C123	S
male	35.0	0	0	373450	8.0500	NaN	S
male	NaN	0	0	330877	8.4583	NaN	Q
male	54.0	0	0	17463	51.8625	E46	S

Chi tiết các bước thực hiện trong file jupyter notebook...

### **1.3.3 Xử lý dữ liệu ngoại lai (Outliers Data)**

# Outliers Data

Một điểm ngoại lai (điểm bất thường) là một điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu. Ta thường xem các giá trị ngoại lai như là các mẫu dữ liệu đặc biệt, cách xa khỏi phần lớn dữ liệu khác trong tập dữ liệu



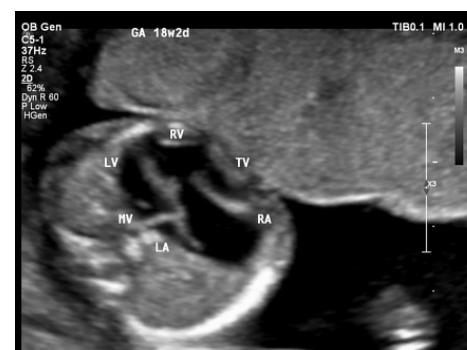
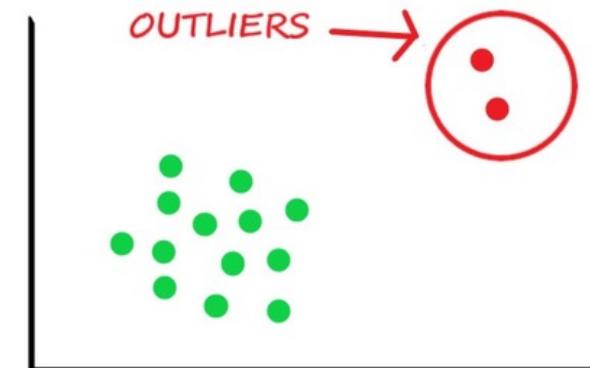
Có rất nhiều nguyên nhân chủ quan và khách quan dẫn tới sự xuất hiện của các điểm ngoại lai trong tập dữ liệu như:

- Các lỗi nhập dữ liệu do con người gây ra;
- Các lỗi đo lường do thiết bị, dụng cụ lấy mẫu, thí nghiệm gây ra;
- Do cố ý tạo ra để phục vụ việc kiểm tra các phương pháp phát hiện;
- Các lỗi xử lý dữ liệu phát sinh trong quá trình thao tác dữ liệu;
- Các lỗi do lấy mẫu được trích xuất hoặc trộn dữ liệu từ các nguồn sai khác nhau;
- Do tự nhiên gây ra, đây không phải là lỗi mà là các giá trị quan sát thật tuy nhiên rất hiếm khi xuất hiện;

# Outliers Data

Việc phát hiện dữ liệu ngoại lai giúp chúng ta có những hiểu biết sâu sắc về từng ứng dụng cụ thể. Một số ứng dụng của dữ liệu ngoại lai trong thực tế như:

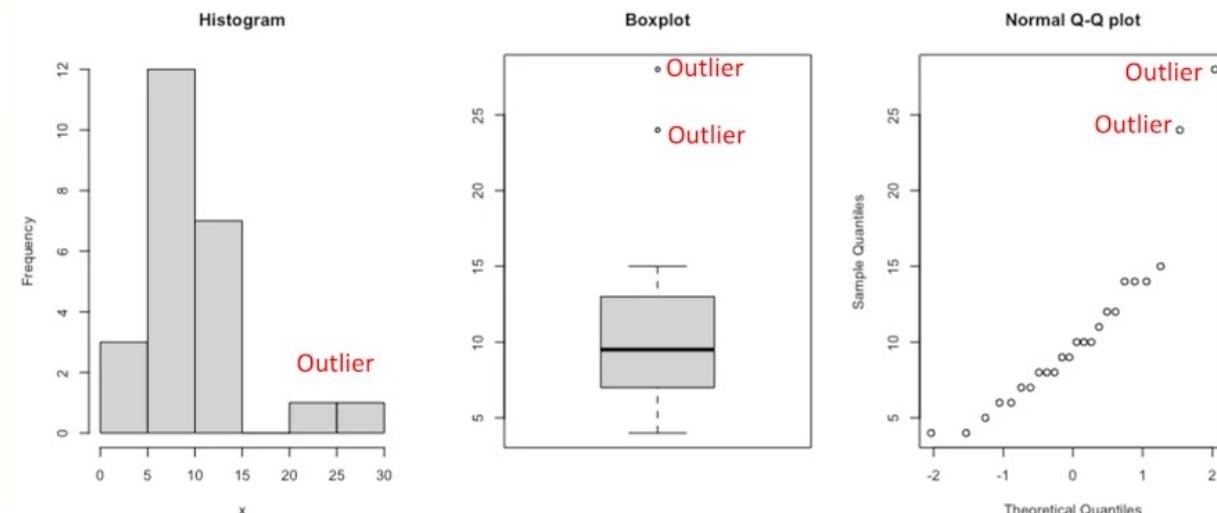
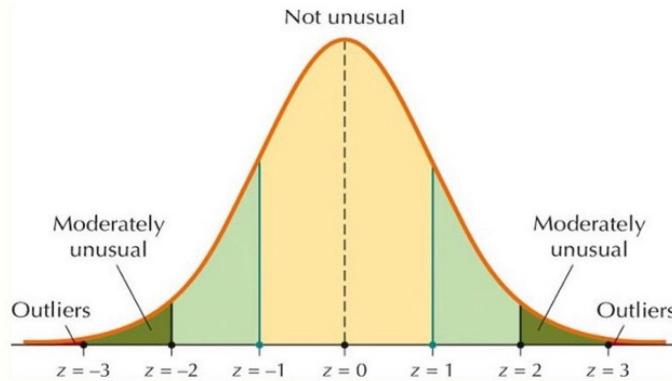
- Hệ thống phát hiện xâm nhập (Intrusion detection systems)
- Phát hiện gian lận tín dụng (Credit card fraud)
- Các sự kiện cảm biến quan tâm (Interesting sensor events)
- Trong chuẩn đoán y tế (Medical diagnosis)
- Trong thực thi pháp luật (Law enforcement)
- Trong khoa học trái đất (Earth science)



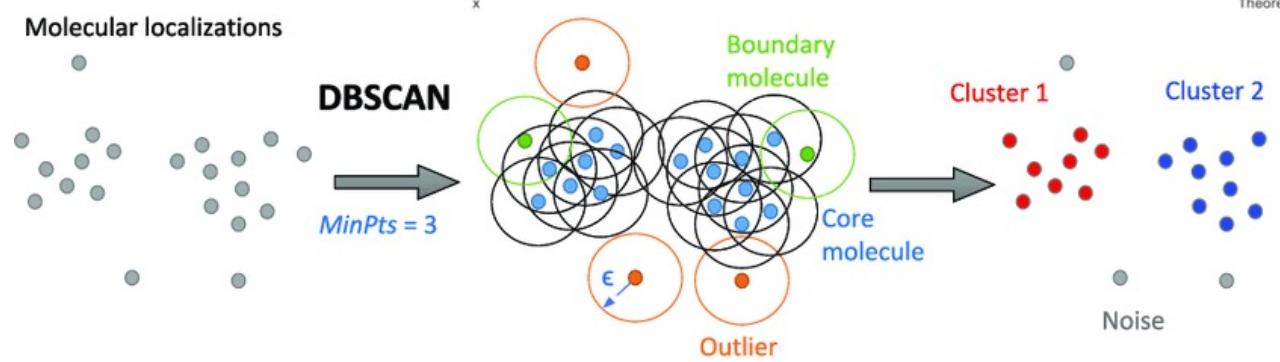
# Outliers Data

Có nhiều phương pháp để phát hiện các điểm dữ liệu ngoại lai:

## Detecting Outliers with z-Scores



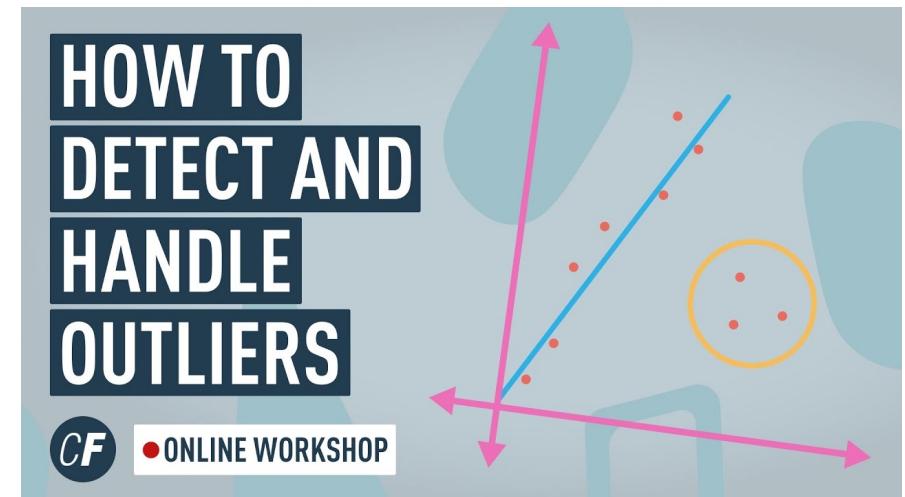
## Molecular localizations



# Outliers Data

Không có một phương pháp, cách thức xử lý ngoại lai chung nào áp dụng cho tất cả các bài toán, các kiểu dữ liệu khác nhau. Một số phương pháp thường sử dụng bao gồm:

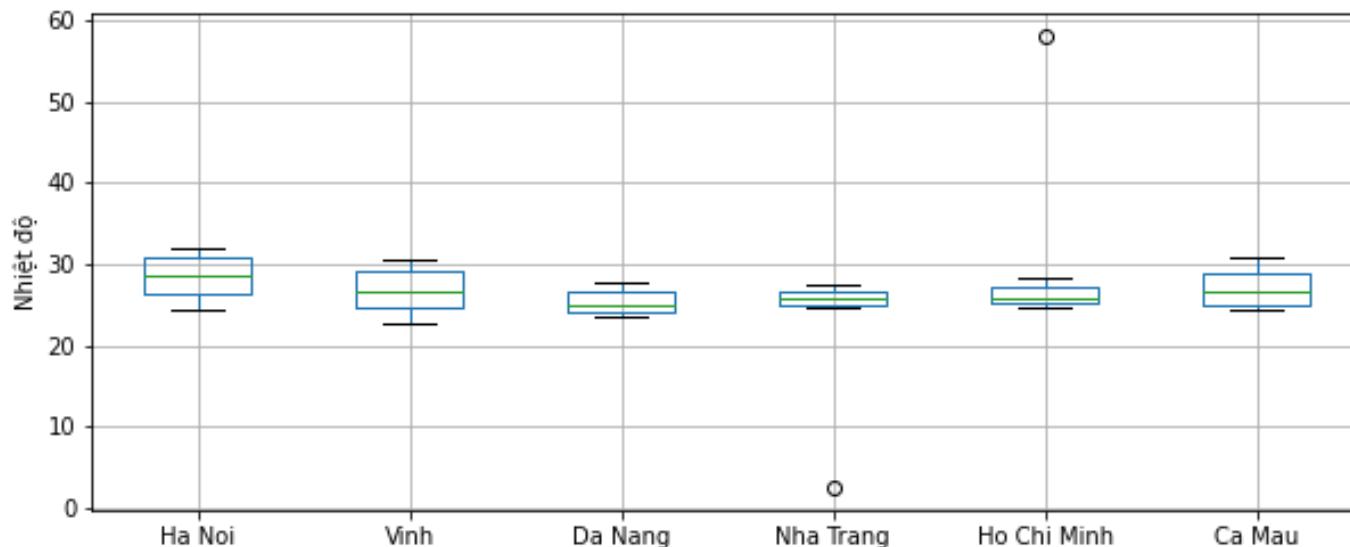
- Phát hiện và không làm gì với dữ liệu ngoại lai.
- Xóa bỏ/cắt bỏ các điểm dữ liệu ngoại lai
- Thay thế dữ liệu ngoại lai bằng giá trị phù hợp hơn
- Chuyển đổi các điểm ngoại lai về thành các điểm dữ liệu thiếu (Missing values) và áp dụng các phương pháp xử lý như một điểm dữ liệu khuyết thiếu.





# Outliers Data

Phát hiện và xử lý ngoại lai cho dữ liệu nhiệt độ:



*Chi tiết các bước thực hiện trong file jupyter notebook...*

time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
01 15-9-2019	25.31	24.21	24.02	24.93	25.16	24.83
02 15-9-2019	25.05	23.73	23.89	24.79	24.8	24.55
03 15-9-2019	24.79	23.36	23.83	24.84	24.74	24.48
04 15-9-2019	24.59	23.05	23.69	24.82	24.8	24.38
05 15-9-2019	30.8	22.8	23.52	24.79	24.87	24.4
06 15-9-2019	24.38	22.79	23.68	2.51	24.71	24.41
07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
10 15-9-2019	31.35	29.97	26.96	27.23	27.68	27.53
11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
13 15-9-2019	30.95	30.25	27.83	27.44	58	30.66
14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
16 15-9-2019	30.8	30.2	26.6	26.45	27.29	29.13
17 15-9-2019	29.94	29.36	25.8	26.67	26.69	28.72
18 15-9-2019	28.53	27.48	24.82	25.92	25.81	27.46
19 15-9-2019	28.89	27.03	24.93	25.88	25.93	27.07
20 15-9-2019	28.06	26.41	24.7	25.7	25.97	26.75
21 15-9-2019	27.43	26.2	24.41	25.62	25.94	26.32
22 15-9-2019	26.98	25.79	24.17	25.6	25.9	26.29
23 15-9-2019	26.68	25.31	23.81	25.53	25.8	26.36

### **1.3.4 Xử lý dữ liệu trùng lặp (Duplicated Data)**



# Duplicated Data

Quá trình tích hợp dữ liệu từ nhiều nguồn, có thể dẫn tới nhiều bản ghi bị trùng lặp. Việc phát hiện và xử lý các dữ liệu trùng lặp này cũng rất quan trọng để nâng cao chất lượng của dữ liệu trước khi phân tích.

	<b>id</b>	<b>first_name</b>	<b>last_name</b>	<b>email</b>
▶	1	Carine	Schmitt	carine.schmitt@verizon.net
	4	Janine	Labrune	janine.labrune@aol.com
	6	Janine	Labrune	janine.labrune@aol.com
	2	Jean	King	jean.king@me.com
	12	Jean	King	jean.king@me.com
	5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
	10	Julie	Murphy	julie.murphy@yahoo.com
	11	Kwai	Lee	kwai.lee@google.com
	3	Peter	Ferguson	peter.ferguson@google.com
	9	Roland	Keitel	roland.keitel@yahoo.com
	14	Roland	Keitel	roland.keitel@yahoo.com
	7	Susan	Nelson	susan.nelson@comcast.net
	13	Susan	Nelson	susan.nelson@comcast.net
	8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Để phát các bản ghi trùng lặp trong dữ liệu không khó, sử dụng bộ lọc hoặc các phương thức được cung cấp sẵn trong các gói thư viện có thể dễ dàng thống kê số lượng bản ghi trùng lặp trong dữ liệu và liệt kê chi tiết danh sách các bản ghi này.

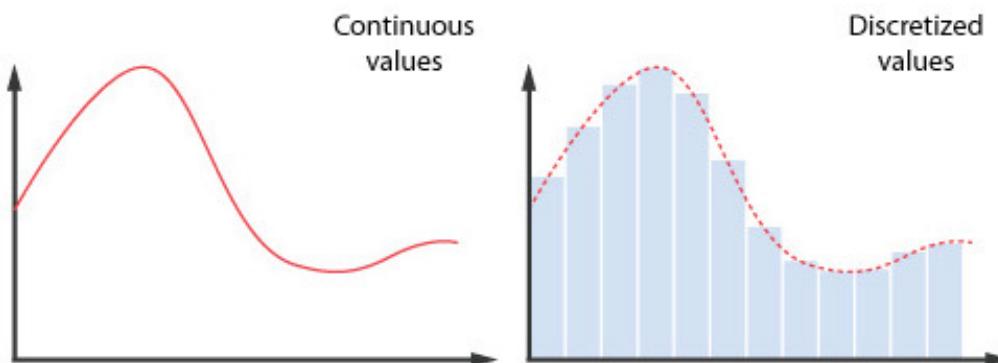
Với các bản ghi trùng lặp, chúng ta cũng sẽ có các phương án xử lý riêng phù hợp. Một số lựa chọn để xử lý dữ liệu trùng lặp bao gồm:

- Xóa tất cả các bản ghi trùng lặp trong dữ liệu
- Giữ lại bản ghi đầu tiên, xóa tất cả các bản ghi trùng lặp phía sau
- Giữ lại bản ghi cuối cùng, xoát tất cả các bản ghi trùng lặp phía trước

### **1.3.5 Rời rạc hóa dữ liệu (Discretize Data)**

# Discretize Data

Rời rạc hóa (discretize) thuộc tính chứa dữ liệu liên tục trong các bộ dữ liệu là một trong những kỹ thuật làm sạch dữ liệu thường gặp trong quá trình phân tích, khai phá dữ liệu. Mục tiêu là biến đổi các giá trị liên tục về các giá trị rời rạc.



Một số phương pháp rời rạc hóa dữ liệu tiêu biểu như:

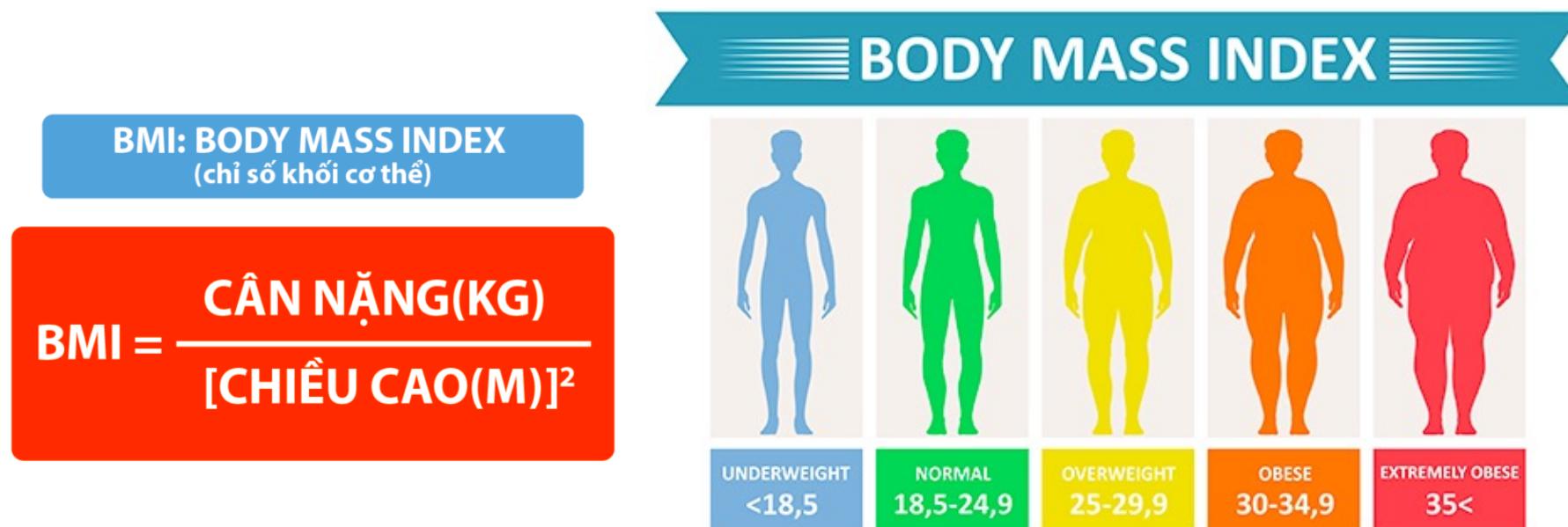
- Rời rạc hóa theo khoảng cách (Equal width discretization)
- Rời rạc hóa theo tần số (Equal frequency discretization)
- Rời rạc hóa dựa trên Entropy (Entropy-based discretization)

Các thuật toán rời rạc hóa đóng vai trò quan trọng trong khai phá dữ liệu và khám phá tri thức. Kết quả của quá trình rời rạc hóa tạo ra các giá trị rời rạc cho các thuộc tính liên tục giúp cho việc cảm nhận dữ liệu dễ dàng, cũng giúp cho việc huấn luyện nhiều mô hình học máy chính xác và nhanh hơn.

Điểm hệ 10	Điểm chữ	Điểm hệ 4
Từ 8,5 - 10,0	A	4
Từ 7,0 - 8,4	B	3
Từ 5,5 - 6,9	C	2
Từ 4,0 - 5,4	D	1
Dưới 4,0	F	0

# Discretize Data

Công ty có 500 nhân viên, cần đánh giá và kiểm tra sức khỏe: Tính toán chỉ số BMI và thống kê số lượng nhân viên theo từng nhóm



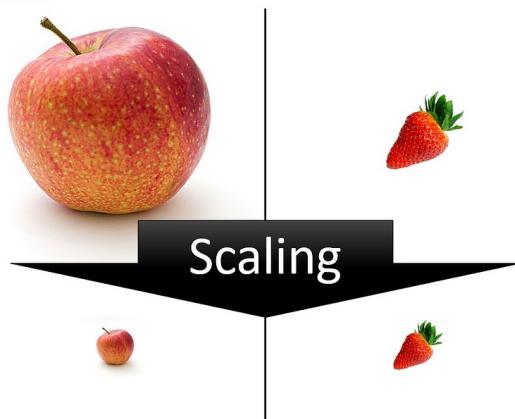
*Chi tiết các bước thực hiện trong file jupyter notebook...*

## 1.3.6 Chuẩn hóa dữ liệu (Scaling Data)



# Scaling Data

**Scaling data** là một quá trình trong khoa học dữ liệu và học máy, trong đó các giá trị của dữ liệu được chuyển đổi về một phạm vi nhất định để chuẩn bị cho các thuật toán phân tích hoặc mô hình học máy. Mục tiêu chính của việc scaling là đảm bảo rằng tất cả các đặc trưng (features) của dữ liệu có cùng tỷ lệ, giúp cải thiện độ chính xác và hiệu quả của các mô hình học máy.



#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Age Range: 27-48

Salary Range: 47000-78000

# Scaling Data

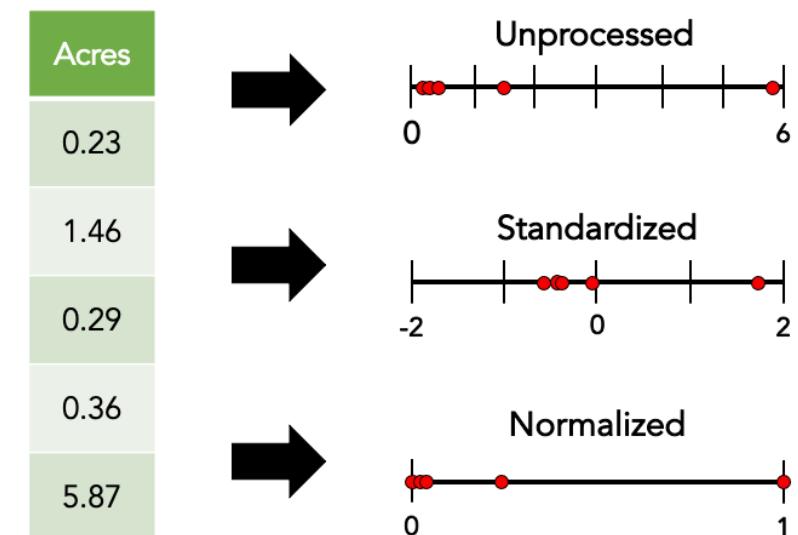
## Lý do cần Scaling Data:

- Cải thiện hiệu suất của thuật toán:** Nhiều thuật toán học máy như Hồi quy tuyến tính (Linear Regression), KNN, Kmeans, SVM... đều hoạt động tốt hơn khi các đặc trưng của dữ liệu có tỷ lệ giống nhau.
- Giảm độ lệch:** Nếu một đặc trưng có phạm vi giá trị lớn hơn, nó có thể chi phối mô hình và khiến mô hình không học được từ các đặc trưng khác.

**scaling data** là một bước quan trọng trong quá trình tiền xử lý dữ liệu để đảm bảo rằng các thuật toán học máy có thể học hiệu quả từ dữ liệu và cho kết quả chính xác.

## Các phương pháp Scaling Data:

- **Normalization**
- **Standardization**





# Scaling Data

**Min-Max Scaling (Normalization – Chuẩn hóa):** Phương pháp này chuyển đổi các giá trị dữ liệu về phạm vi từ 0 đến 1 (hoặc trong một phạm vi khác được xác định trước).

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization



Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1

Range 0-1

Range 0-1

How to calculate Normalized value?

X = 35, min = 27, max = 48 for column Age.

$$X_{\text{norm}}(\text{for } 35) = \frac{35-27}{48-27} = 0.3809$$



# Scaling Data

**Standardization (Z-Score):** Phương pháp này chuyển đổi các giá trị dữ liệu sao cho trung bình 0 và độ lệch chuẩn bằng 1.

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Mean = 37.42857  
 Mean = 60428.5714  
Std. Dev. = 6.883876  
 Std. Dev. = 10499.7570

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

Age	Standardized Age	Salary	Standardized Salary
44	0.954611636	73000	1.197306616
27	-1.514927162	47000	-1.278941158
30	-1.079126198	53000	-0.707499364
38	0.083009708	62000	0.149663327
40	0.373543684	57000	-0.326538168
35	-0.352791257	53000	-0.707499364
48	1.535679589	78000	1.673508111

**Mean = 0**  
**Std. dev. = 1**

How to calculate Standardized value?  
 $x = 35, \text{mean} = 37.42, \text{Std. Dev.} = 6.88$   
 for column Age.  
 $x_{\text{std}}(\text{for } 35) = \frac{35 - 37.42}{6.88} = -0.3527$

Standardization



## **1.3.7 Làm sạch và chuẩn hóa dữ liệu văn bản (Clean Text)**

---



# Clean Text

Dữ liệu văn bản là loại dữ liệu không có cấu trúc. Quá trình thu thập, tổng hợp dữ liệu từ các nguồn dẫn đến văn bản chứa rất nhiều nhiễu (noise).

"SORRY JB FAN &#65292; REALLY JUSTIN COME BACK". That's the top comment on a Justin Timberlake video with 10 million views . Son of a fucking bitch! I'm looking forward this video.  
Have you been seen it yet. ^.^ @john\_fanjustin@gmail.com; <https://allflim.com>

Với các bài toán về xử lý ngôn ngữ tự nhiên – NLP (Natural Language Processing) việc làm sạch và chuẩn hóa dữ liệu văn bản là rất quan trọng, giúp cho các thuật toán có thể trích xuất được những đặc trưng tốt nhất từ đó nâng cao hiệu quả, chất lượng của các mô hình.

Các kỹ thuật tiền xử lý dữ liệu văn bản cũng khác rất nhiều so với dữ liệu số. Để lựa chọn được phương án xử lý tốt và phù hợp phải có những hiểu biết sâu sắc về loại ngôn ngữ của văn bản đang xử lý, cũng như lĩnh vực mà văn bản đang đề cập.

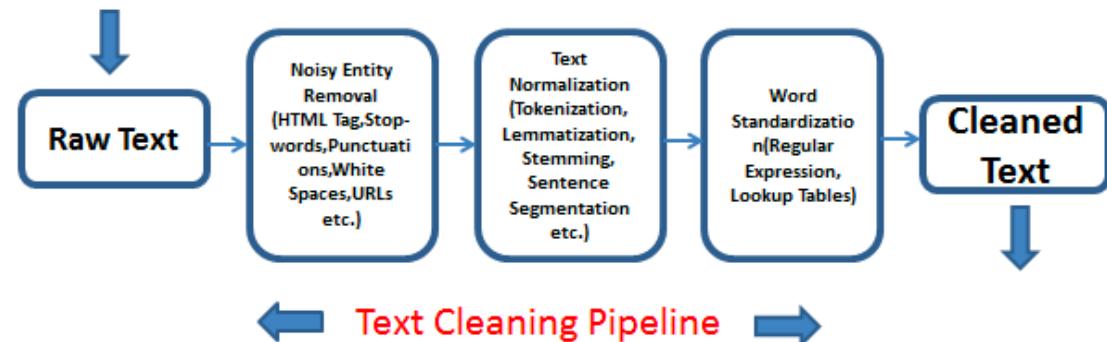
**CLEAN-TEXT  
FOR NLP**



# Clean Text

Một số kỹ thuật tiền xử lý dữ liệu với ngôn ngữ Tiếng Anh như:

- Chuyển đổi về chữ thường
- Chuẩn hóa các ký hiệu viết tắt
- Loại bỏ các đường liên kết (link), địa chỉ email, thẻ tag
- Loại bỏ tất cả các dấu câu, ký tự đặc biệt, số
- Loại bỏ các từ StopWords
- Chuẩn hóa từ (Stemming & Lemmantization)
- Xử lý các Emoji





# Clean Text

Văn bản gốc (Raw Text)	"SORRY JB FAN &#65292;REALLY JUSTIN COME BACK". That's the top comment on a Justin Timberlake video with 10 million views 🎈 . Son of a fucking bitch! I'm looking forward this video. Have you been seen it yet. ^.^😊 @john fanjustin@gmail.com; <a href="https://allflim.com">https://allflim.com</a>
1. Chuyển đổi về chữ thường	"sorry jb fan &#65292;really justin come back". that's the top comment on a justin timberlake video with 10 million views 🎈 . son of a fucking bitch! i'm looking forward this video. have you been seen it yet. ^.^😊 @john fanjustin@gmail.com; <a href="https://allflim.com">https://allflim.com</a>
2. Chuẩn hóa các ký hiệu viết tắt	"sorry jb fan &#65292;really justin come back". that is the top comment on a justin timberlake video with 10 million views 🎈 . son of a fucking bitch! i am looking forward this video. have you been seen it yet. ^.^😊 @john fanjustin@gmail.com; <a href="https://allflim.com">https://allflim.com</a>
3. Loại bỏ các link, email, tag	"sorry jb fan &#65292;really justin come back". that is the top comment on a justin timberlake video with 10 million views 🎈 . son of a fucking bitch! i am looking forward this video. have you been seen it yet. ^.^😊 ;
4. Loại bỏ các dấu câu, ý tự đặc biệt, số	sorry jb fan really justin come back that is the top comment on a justin timberlake video with million views 🎈 son of a fucking bitch i am looking forward this video have you been seen it yet 😊
5. Loại bỏ Stopwords	sorry jb fan really justin come back top comment justin timberlake video million views 🎈 son fucking bitch looking forward video seen yet 😊
6. Chuẩn hóa từ	sorry jb fan really justin come back top comment justin timberlake video million view 🎈 son fuck bitch look forward video see yet 😊
7. Xử lý Emoji	sorry jb fan really justin come back top comment justin timberlake video million view son fuck bitch look forward video see yet
Văn bản đã làm sạch (Cleaned Text)	<b>sorry jb fan really justin come back top comment justin timberlake video million view son fuck bitch look forward video see yet</b>



## **1.4 Quy trình làm sạch dữ liệu**





# Quy trình làm sạch dữ liệu

**DATA  
CLEANING  
CYCLE**

- 1 DATA AUDIT
- 2 WORKFLOW SPECIFICATION
- 3 WORKFLOW EXECUTION
- 4 VALIDATION
- 5 REPORTING

- **Bước 1 - Kiểm tra dữ liệu (Data audit):** Để bắt đầu quy trình làm sạch dữ liệu việc đầu tiên là cần kiểm tra, xem xét kỹ những dữ liệu đã thu thập và sẽ được sử dụng. Xác định các loại lỗi mà tập dữ liệu có thể chứa và vị trí của chúng. Việc này có thể được thực hiện thông qua các phương pháp thống kê và cơ sở dữ liệu để tìm ra những điểm bất thường, mâu thuẫn.
- **Bước 2 – Xây dựng luồng công việc (Workflow specification) :** Giai đoạn này sẽ xác định các thao tác, hoạt động, trình tự làm sạch tập dữ liệu. Quy trình làm sạch dữ liệu điển hình bao gồm một loạt thao tác được thực hiện lặp lại trên dữ liệu cho đến khi đạt được tập dữ liệu chất lượng.



# Quy trình làm sạch dữ liệu



- **Bước 3 – Thực thi luồng công việc (Workflow execution):** Thực hiện các kỹ thuật làm sạch dữ liệu theo như luồng công việc đã xây dựng ở bước 2. Quá trình làm sạch dữ liệu có thể có các phương thức, kỹ thuật khác nhau phụ thuộc vào bản chất của dữ liệu và từng dự án cụ thể. Nhưng mục tiêu cuối cùng luôn giống nhau đó là loại bỏ hoặc chỉnh sửa dữ liệu không ban đầu để thu được dữ liệu sạch.
- **Bước 4 – Kiểm tra, xác thực (Validation):** Sau khi thực hiện làm sạch cần kiểm tra lại dữ liệu và đảm bảo rằng tất cả các yêu cầu và ràng buộc đã được thực thi, và thực thi đầy đủ trên dữ liệu thực tế.
- **Bước 5 – Báo cáo (Reporting):** Giai đoạn cuối cùng, nhưng không thể bỏ qua đó là xây dựng các báo cáo. Việc tạo ra các báo cáo tổng kết và tóm tắt về quá trình làm sạch dữ liệu là điều cần thiết. Đặc biệt trong trường hợp xử lý nhiều tập dữ liệu và làm việc nhóm.

# Quy trình làm sạch dữ liệu

---



**IMPORTANT**

Hiểu biết sâu sắc dữ liệu mình có, bài toán mình đang giải quyết là yêu cầu bắt buộc trong quá trình làm sạch và chuẩn bị dữ liệu. Nó là yếu tố quan trọng ảnh hưởng tới chất lượng của dữ liệu sau khi làm sạch.

# Làm sạch dữ liệu

---



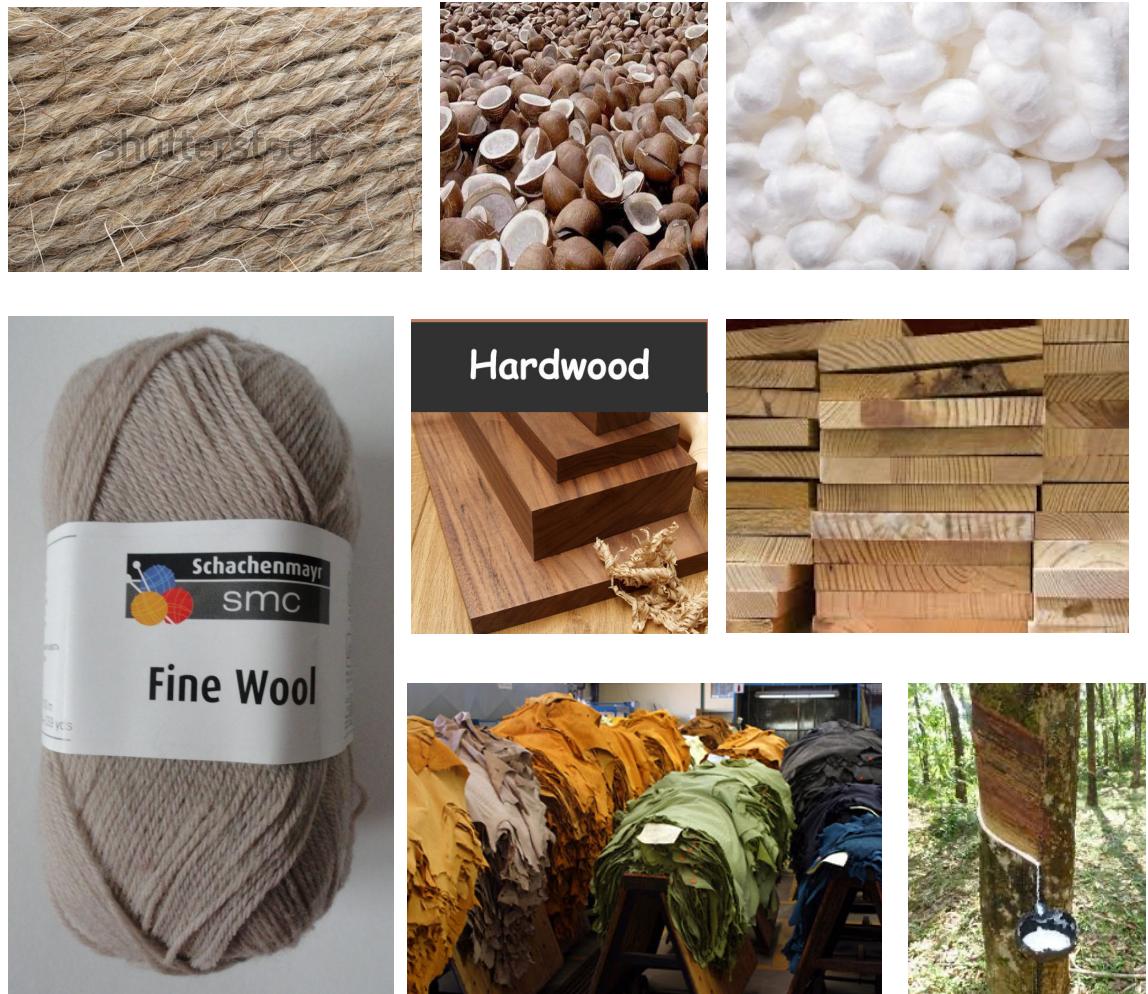
## Ví dụ



# Ví dụ quy trình làm sạch dữ liệu

+ Tập dữ liệu về giá bán một số mặt hàng nông sản của Mỹ theo từng tháng từ 04/1990 đến 04/2020. Bao gồm 8 loại sản phẩm:

1. Coarse wool (Len thô)
2. Copra (Cùi dừa)
3. Cotton (Bông)
4. Fine wool (Len min)
5. Hard log (Gỗ cứng)
6. Hard sawnwood (Gỗ xẻ cứng)
7. Hide (Da thú)
8. Rubber (Cao su)





# Ví dụ quy trình làm sạch dữ liệu

A	B	C	D	E	F	G	H	I	J	K
Month	Coarse wool Price	Coarse wool price % Change	Copra Price	Copra price % Change	Cotton Price	Cotton price % Change	Fine wool Price	Fine wool price % Change	Hard log Price	Hard log price
Apr-90	482.34	-	236	-	1.83	-	1,071.63	-	161.2	-
May-90	447.26	-7.27%	234	-0.85%	1.89	3.28%	1,057.18	-1.35%	172.86	
Jun-90	440.99	-1.40%	216	-7.69%	1.99	5.29%	898.24	-15.03%	181.67	
Jul-90	418.44	-5.11%	205	-5.09%	2.01	1.01%	895.83	-0.27%	187.96	
Aug-90	418.44	0.00%	198	-3.41%	1.79	-10.95%	951.22	6.18%	186.13	
Sep-90	412.18	-1.50%	196	-1.01%	1.79	0.00%	936.77	-1.52%	185.33	
Oct-90	394.64	-4.26%	198	1.02%	1.79	0.00%	901.85	-3.73%	189.76	
Nov-90	334.5	-15.24%	236	19.19%	1.82	1.68%	888.61	-1.47%	179.02	
Dec-90	328.24	-1.87%	237	0.42%	1.85	1.65%	870.55	-2.03%	171.13	
Jan-91	319.47	-2.67%	233	-1.69%	1.85	0.00%	887.41	1.94%	169.19	
Feb-91	323.23	1.18%	226	-3.00%	1.87	1.08%	596.02	-32.84%	176.93	
Mar-91	328.24	1.55%	236	4.42%	1.86	-0.53%	586.39	-1.62%	162.57	
Apr-91	365.82	11.45%	224	-5.08%	1.83	-1.61%	596.02	1.64%	175.59	
May-91	371.88	1.66%	226	0.89%	1.82	-0.55%	721	20.97%	174.04	
Jun-91	340.6	-8.41%	245	8.41%	1.78	-2.20%	777.51	7.84%	200.15	
Jul-91	337.48	-0.92%	303	23.67%	1.7	-4.49%	723.48	-6.95%	207.82	
Aug-91	337.22	-0.08%	299	-1.32%	1.62	-4.71%	680.64	-5.92%	210.57	
Sep-91	313.96	-6.90%	296	-1.00%	1.55	-4.32%	613.45	-9.87%	210.68	
Oct-91	308.39	-1.77%	353	19.26%	1.5	-3.23%	558.18	-9.01%	214.44	
Nov-91	307.57	-0.27%	385	9.07%	1.4	-6.67%	641.49	14.93%	195.5	
Dec-91	295.4	-3.96%	411	6.75%	1.36	-2.86%	652.19	1.67%	200.33	
Jan-92	297.04	0.56%	488	18.73%	1.31	-3.68%	619.37	-5.03%	202.85	
Feb-92	341.89	15.10%	471	-3.48%	1.24	-5.34%	655.83	5.89%	214.04	
Mar-92	341.18	-0.21%	429	-8.92%	1.22	-1.61%	667.93	1.84%	201.96	
Apr-92	352.12	3.21%	425	-0.93%	1.28	4.92%	667.7	-0.03%	199.67	

agricultural\_raw\_material





# Ví dụ quy trình làm sạch dữ liệu

Tập dữ liệu bao gồm 17 cột:

- **Month:** Tháng của năm (Tháng - Năm): Apr – 90 (Tháng: 3 ký tự đầu tiên – Năm: 2 số cuối của năm)
- Mỗi sản phẩm nông nghiệp bao gồm 2 thông tin, ví dụ:
  - **Coarse wool Price:** Giá bán Len thô của tháng (\$USD): 482.34
  - **Coarse wool price % Change:** Tỷ lệ % thay đổi mức giá bán len thô của tháng so với tháng liền trước đó: -7.27% (Mức giá giảm so với tháng trước là 7.27%) | 1.01% (Mức giá tăng so với tháng trước là 1.01 %) | 0.00% (Mức giá tháng này và tháng trước đó như nhau, không thay đổi giá)

*(Tương tự với 7 mặt hàng còn lại)*



**TARGET**

Tập dữ liệu thô (Raw data) về giá và tỷ lệ thay đổi giá của một số mặt hàng nông nghiệp của USA. File dữ liệu này cần phải được xử lý và làm sạch trước khi sử dụng cho phân tích hay bất kỳ mục đích gì.

**Áp dụng quy trình làm sạch dữ liệu để xử lý với tập dữ liệu này nhằm nâng cao chất lượng phục vụ cho việc phân tích?**



## 2.5 Bài tập



# Thực hành

---

Tập dữ liệu **Data\_Patient\_Heart\_Attack.xlsx** lưu trữ hồ sơ sức khỏe của các bệnh nhân khám bệnh đau tim, được thu thập từ các phòng khám chuyên khoa, Mỗi một bệnh nhân bao gồm 9 thông tin:

1. id: Mã của bệnh nhân (object)
2. Age: Tuổi của bệnh nhân (số)
3. Gender: Giới tính của bệnh nhân (chuỗi: Male – Female)
4. Type: Cho biết loại triệu chứng đau ngực mà bệnh nhân này mắc phải, với 4 giá trị: (Typical angina, Atypical angina, Non-anginal pain, Asymptomatic)
5. Blood\_pressure: Huyết áp của bệnh nhân – đơn vị: mmhg (số)
6. Cholesterol: Chỉ số cholesterol của bệnh nhân – đơn vị: mg/dl (số)
7. Heartbeat: Thông số nhịp tim của bệnh nhân – đơn vị: lần/phút (số)
8. Thalassemia: Chỉ số Thalassemia của bệnh nhân chỉ gồm 3 giá trị (3: Bình thường | 6: Khiếm khuyết cố định | 7: Kiếm khuyết có thể đảo ngược)
9. Result: Cho biết bệnh nhân có bị bệnh tim hay không? (No: Không bị bệnh tim mạch | Yes: Bị bệnh tim mạch)



# Thực hành

Screenshot of Microsoft Excel showing a dataset titled "Data\_Patient\_Heart\_Attack". The table contains 23 rows of patient data across 10 columns.

	A	B	C	D	E	F	G	H	I
1	<b>id</b>	<b>Age</b>	<b>Gender</b>	<b>Type</b>	<b>Blood_pressure</b>	<b>Cholesterol</b>	<b>Heartbeat</b>	<b>Thalassemia</b>	<b>Result</b>
2	Patient_01	63	Male	Typical angina	145	233	150	6	No
3	Patient_02	67	Male	Asymptomatic	160	286	108	Three	Yes
4	Patient_03	67	Male	Asymptomatic	120	229	129	7	Yes
5	Patient_04	37	Male	Non-anginal pain	130	250	187	Three	No
6	Patient_05	41	F	Atypical angina	130	204	172	6	No
7	Patient_16	56	Male	Atypical angina	120	236	178	Three	No
8	Patient_07	62	Female	Asymptomatic	140	268	0	Three	Yes
9	Patient_08	57	Female	Asymptomatic	120	354	163	Three	No
10	Patient_19	63	Male	Asymptomatic	130	254	147	7	Yes
11	Patient_10	53	Male	Asymptomatic	140	203	155	7	Yes
12	Patient_110	57	Male	Asymptomatic	140	192	148	6	No
13	Patient_120	56	Female	Atypical angina	140	294	153	Three	No
14	Patient_130	56	Male	Non-anginal pain	130	256	142	6	Yes
15	Patient_140	44	Male	Atypical angina	120	263	173	7	No
16	Patient_150	52	Male	Non-anginal pain	172	199	162	7	No
17	Patient_160	57	Male	Non-anginal pain	150	168	174	Three	No
18	Patient_170	48	Male	Atypical angina	110	229	168	7	Yes
19	Patient_180	54	Male	Asymptomatic	140	239	160	Three	No
20	Patient_190	48	Female	Non-anginal pain	130	275	139	Three	No
21	Patient_200	49	Male	Atypical angina	130	266	171	Three	No
22	Patient_21	64	Male	Typical angina	110	211	144	6	No
23	Patient_22	58	Female	Typical angina	150	283	162	Three	No



# Thực hành



## Yêu cầu:

1. Nghiên cứu dữ liệu **Data\_Patient\_Heart\_Attack.xlsx**, xác định các vấn đề cần làm sạch với tập dữ liệu này?
2. Đề xuất các phương pháp và Thực hiện làm sạch tập dữ liệu để thu được một tập dữ liệu chất lượng
3. Lưu tập dữ liệu đã làm sạch ra file Excel Data\_Patient\_OK.xlsx



