Fulbright
SCHOOL OF PUBLIC POLICY AND MANAGEMENT

Disruptive Technologies
for Public Asset Governance

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Embassy of Switzerland
Swiss Cooperation Office in Vietnam

Cơ bản về dữ liệu và phân tích dữ liệu

Huỳnh Nhật Nam

Fulbright School of Public Policy and Management
Ho Chi Minh City, 16/08/2024

Outline

- Hiểu biết về dữ liệu - Một số khái niệm
- Phân tích khám phá dữ liệu
- Thảo luận

Hiểu biết về dữ liệu (Data literacy)

Một vài định nghĩa

> Hiểu biết về dữ liệu là năng lực **đọc, hiểu, tạo ra dữ liệu, và giao tiếp nội dung thông tin của dữ liệu.**

https://en.wikipedia.org/wiki/Data_literacy

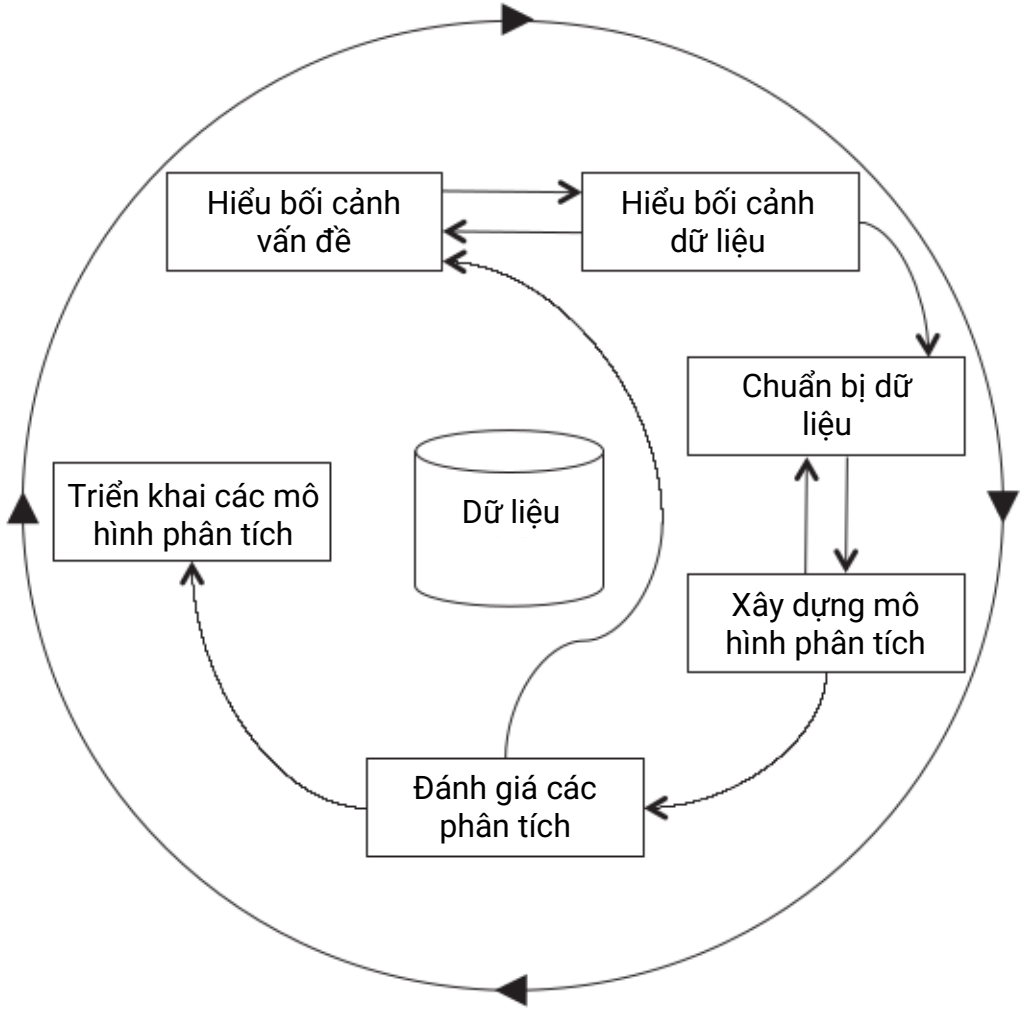
> Hiểu biết về dữ liệu mô tả mức độ một cá nhân hoặc tổ chức **hiểu, làm việc, phân tích, hiển thị, và ứng dụng dữ liệu để đạt được các mục tiêu của mình**.

https://www.dataversity.net/what-is-data-literacy/

> Hiểu biết về dữ liệu là năng lực **hiểu dữ liệu và các cách thực hành về dữ liệu** một cách đầy đủ để **diễn giải dữ liệu một cách có ý nghĩa và truyền đạt ý nghĩa đó một cách hiệu quả**.

https://www.nnlm.gov/guides/data-glossary/data-literacy#:~:text=Definition,and%20effectively%20communicate%20that%20meaning.

Vòng đời phân tích khai phá dữ liệu CRISP-DM



Provost & Fawcett, 2013, *Data Science for Business*, O'Reilly

- Một vấn đề được xác định rõ ràng đã được giải quyết một nửa.

- Cần hiểu vấn đề (và các giải pháp cũng như các tác động dự kiến) trước khi đi sâu vào các phân tích.
  - Vì sao vấn đề này quan trọng?
  - Vấn đề này tác động những ai?
  - Nếu chúng ta không có đủ dữ liệu (phù hợp) thì sẽ như thế nào?
  - Khi nào dự án kết thúc?
  - Nếu chúng ta không thích các kết quả phân tích thì sẽ như thế nào?

Gutman & Goldmeier, 2021, Becoming a *Data Head*, Wiley



Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA)

EDA - Một vài định nghĩa

> Phân tích khám phá dữ liệu không thể kể toàn bộ câu chuyện, nhưng không gì khác có thể đóng một vai trò nền tảng như vậy - **là bước đầu tiên**.
>
> John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

> Phân tích khám phá dữ liệu là **một thái độ**, **một tư duy linh hoạt**, sự sẵn sàng tìm kiếm những thông tin chúng ta tin là không có, cũng như những thông tin chúng ta tin là có trong dữ liệu.
>
> John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

> Mục tiêu chính của phân tích khám phá dữ liệu là kiểm tra các phân phối, các điểm dữ liệu bất thường, … **tạo ra các giả thuyết** bằng cách trực quan hóa và hiểu dữ liệu.
>
> https://link.springer.com/chapter/10.1007/978-3-319-43742-2_15

EDA - Các bước chính

- Các đặc điểm tổng quan của tập dữ liệu
- Phân tích mô tả
  - Phân tích và hiển thị dữ liệu đơn biến
  - Phân tích và hiển thị dữ liệu của cặp biến
- Phân tích suy diễn
- Các bất thường trong dữ liệu
  - Các giá trị ngoại lệ và các giá trị inliers
  - Các giá trị bị mất

EDA - Các đặc điểm tổng quan của tập dữ liệu

- **Dữ liệu phi cấu trúc**, ví dụ hình ảnh, chữ viết, mạng lưới, âm thanh.

- **Dữ liệu có cấu trúc**, ví dụ các bộ dữ liệu có dạng bảng như điều tra khảo sát và chuỗi thời gian.
  - số hàng dữ liệu và mỗi hàng dữ liệu đại diện thông tin gì, ví dụ một cá nhân, một hộ gia đình, một sự kiện ngập đường.
  - Số lượng và định dạng của các cột dữ liệu
    - *Dữ liệu định danh (nominal data).* Loại trừ lẫn nhau, không có ý nghĩa về số lượng và thứ tự, ví dụ Nam và Nữ.
    - *Dữ liệu thứ tự (ordinal data).* Các giá trị có thứ tự lớn nhỏ, nhưng sự khác nhau giữa các giá trị không được định nghĩa, ví dụ các giá trị của thang đo Likert.
    - *Dữ liệu khoảng (interval data).* Các giá trị có thứ tự, sự khác nhau giữa các giá trị được định nghĩa, nhưng không có khái niệm 'giá trị không', ví dụ nhiệt độ, giờ (theo đồng hồ)
    - *Dữ liệu tỷ lệ (ratio data).* Tương tự như dữ liệu khoảng nhưng có khái niệm "giá trị không", ví dụ thu nhập, độ sâu ngập.

EDA - Phân tích mô tả

- Phân tích và hiển thị dữ liệu đơn biến
  - Các đại lượng đo lường sự định tâm: trung bình (mean), trung vị (median), yếu vị (mode)
  - Các đại lượng đo lường sự biến thiên: khoảng giá trị (range), phương sai (variance), độ lệch chuẩn (standard deviation), khoảng tứ phân vị (interquartile range)

Loại dữ liệu	Các phép toán	Đại lượng đo lường sự định tâm	Các đại lượng đo lường sự biến thiên
Định danh (nominal)	● So sánh giống nhau (=, ≠)	● Giá trị yếu vị	● Không có
Thứ tự (ordinal)	● So sánh giống nhau (=, ≠) ● So sánh lớn nhỏ (>, <)	● Giá trị yếu vị ● Giá trị trung vị	● Khoảng giá trị ● Khoảng tứ phân vị
Khoảng (interval)	● So sánh giống nhau (=, ≠) ● So sánh lớn nhỏ (>, <) ● Cộng, trừ (+, -)	● Giá trị yếu vị ● Giá trị trung vị ● Trung bình số học	● Khoảng giá trị ● Khoảng tứ phân vị ● Độ lệch chuẩn ● Phương sai
Tỷ lệ (ratio)	● So sánh giống nhau (=, ≠) ● So sánh lớn nhỏ (>, <) ● Cộng, trừ (+, -) ● Nhân, chia (x, ÷)	● Giá trị yếu vị ● Giá trị trung vị ● Trung bình số học ● Trung bình hình học	● Khoảng giá trị ● Khoảng tứ phân vị ● Độ lệch chuẩn ● Phương sai ● Độ lệch chuẩn tương đối

Source: https://www.scribbr.com/statistics/levels-of-measurement/

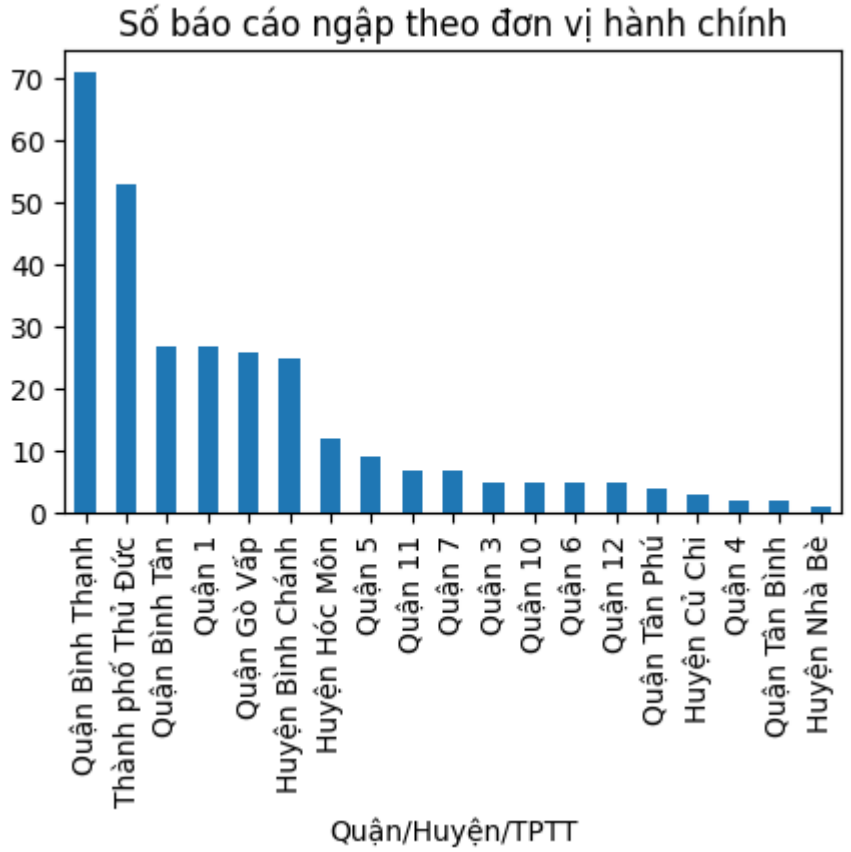
EDA - Phân tích mô tả

Dữ liệu mẫu từ nền tảng FEDS

Số hiệu báo cáo	Loại hình học	Quận/Huyện/TP TT	Phường/Xã	Đường/Phố	Vị trí ngập đầu	Vị trí ngập cuối	Mức ngập	Nguyên nhân ngập	Thời điểm tạo báo cáo	Thời điểm bắt đầu ngập	Thời điểm kết thúc ngập	Thời lượng ngập (phút)	Thời gian tiêu thoát (phút)	Độ sâu ngập	Chiều dài ngập (m)	Chiều rộng ngập (m)	Diện tích ngập (m²)	Vũ lượng (mm)	Thời điểm bắt đầu mưa	Thời điểm kết thúc mưa
2875	Đường	Quận Gò	Phường 1	Đường Ph	SN 520	Lê Đức T	Chưa xác	Do mưa	45511.80	45511.78	45511.81	45		0.1	435.58	9	3920.22			
2866	Đường	Huyện Bì	Thị Trấn	Đường Bi	22	26	Chưa xác	Do mưa	45504.66	45504.60	45504.66	77		0.2	117.77	10	1177.7			
1260	Điểm	Huyện Bì	Thị Trấn	Đường Bi	33	55	Chưa xác	Do mưa	45504.65	45503.64	45503.67	56		1	100	20	2000			
2865	Đường	Huyện Bì	Thị Trấn	Đường K	22	56	Chưa xác	Do mưa	45503.70	45502.70	45502.71	19		0.51	199.15	10	1991.5			
1258	Điểm	Quận Bìn	Phường A	Đường Số	48	128	Chưa xác	Do mưa	45503.62	45502.62	45502.70	120		0.1	160	16	2560			
1257	Điểm	Quận 4	Phường 3	Hẻm 256	1	5	Chưa xác	Do mưa	45503.62	45503.62	45503.62	3		1	66	22	1452			
2864	Đường	Quận Bìn	Phường A	Hẻm 532/	2	48	Chưa xác	Do mưa	45503.61	45491.37	45491.44	101		0.1	121.55	8	972.4			
1256	Điểm	Huyện Bì	Thị Trấn	Quốc Lộ	12	123	Chưa xác	Do mưa	45503.60	45502.60	45502.64	62		0.3	106.66	10	1066.6			
2863	Đường	Thành phố	Phường A	Đường Tr	Sn300	Sn500	Chưa xác	Do mưa	45503.59	45503.55	45503.58	43		0.2	119.43	12	1433.16			
2862	Đường	Thành phố	Phường A	Đường Hi	Kha Vạn	CA phườn	Chưa xác	Do mưa	45503.59	45502.59	45502.63	62		0.1	258.5	8	2068			
2861	Đường	Thành phố	Phường A	Đường Cà	100	50	Chưa xác	Do mưa	45503.59	45503.59	45503.60	12		0.1	128.57	150	19285.5			
1255	Điểm	Quận Bìn	Phường 1	Đường Lé	12	123	Chưa xác	Do mưa	45502.43	45495.43	45495.47	60		0.5	106.63	10	1066.3			
1254	Điểm	Quận 1	Phường B	Cầu Thị N	12a	112b	Chưa xác	Do mưa	45499.65	45498.59	45498.63	60		0.5	100	10	1000			
1253	Điểm	Huyện Bì	Thị Trấn	Đường Ng	1	88	Chưa xác	Do mưa	45498.66	45496.54	45496.55	20		0.15	106.66	20	2133.2			
1252	Điểm	Huyện Bì	Thị Trấn	Đường Ng	12	123	Chưa xác	Do mưa	45498.66	45496.54	45496.55	20		0.1	100	20	2000			
1251	Điểm	Quận 6	Phường 1	Đường Ly	3	88	Ngập tron	Do mưa	45497.73	45496.52	45496.55	45	25	0.1	100	20	2000	1	45496.25	45496.54
1250	Điểm	Quận 1	Phường B	Đường Lé	3	123	Ngập tron	Do mưa	45497.73	45496.52	45496.55	55	25	0.1	22	10	220	1	45496.25	45496.54
2860	Đường	Quận Gò	Phường 1	Đường Qu	1	90	Ngập vừa	Do mưa	45497.73	45496.52	45496.56	70	40	0.14	1331.38	10	13313.8	1	45496.25	45496.54
2859	Đường	Quận Bìn	Phường 1	Đường Dị	12	123	Ngập vừa	Do mưa	45497.72	45496.51	45496.56	75	40	0.15	1652.39	10	16523.9	1	45496.25	45496.54
1249	Điểm	Thành phố	Phường A	Đường M	2	123	Chưa xác	Do mưa	45497.72	45496.51	45496.56	69		0.15	10	10	100			

EDA - Phân tích mô tả

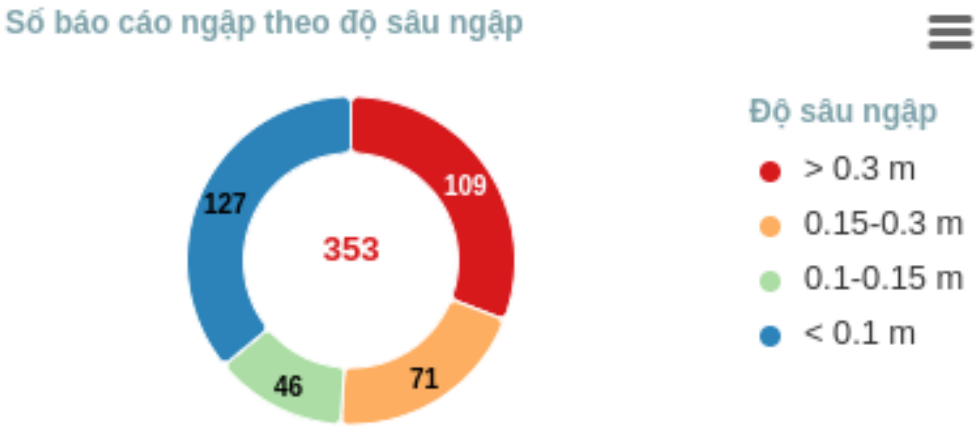
Phân tích và hiển thị dữ liệu đơn biến (cột giá trị định danh) - dữ liệu mẫu từ nền tảng FEDS



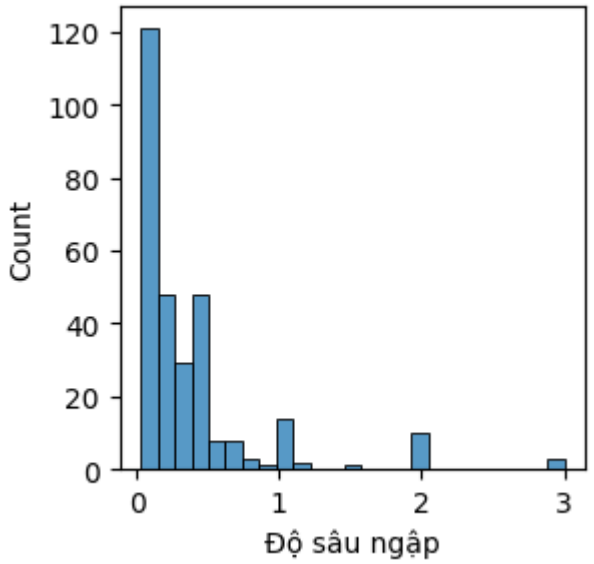
Quận/Huyện/TPTT	
Quận Bình Thạnh	71
Thành phố Thủ Đức	53
Quận Bình Tân	27
Quận 1	27
Quận Gò Vấp	26
Huyện Bình Chánh	25
Huyện Hóc Môn	12
Quận 5	9
Quận 11	7
Quận 7	7
Quận 3	5
Quận 10	5
Quận 6	5
Quận 12	5
Quận Tân Phú	4
Huyện Củ Chi	3
Quận 4	2
Quận Tân Bình	2
Huyện Nhà Bè	1

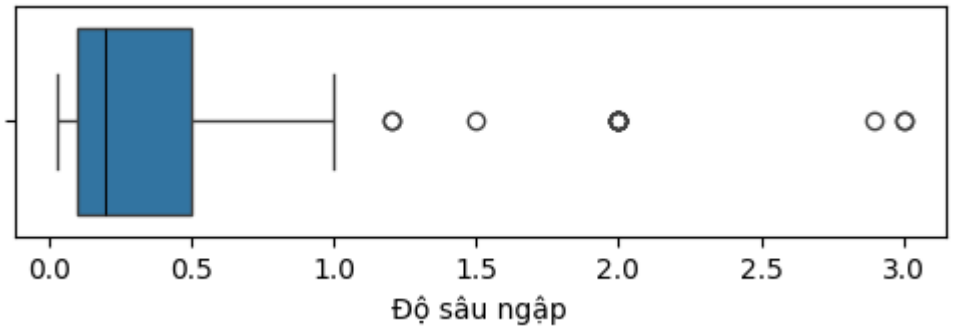
EDA - Phân tích mô tả

Phân tích và hiển thị dữ liệu đơn biến (cột giá trị số) - dữ liệu mẫu từ nền tảng FEDS



	Độ sâu ngập
count	296.000000
mean	0.378649
std	0.482299
min	0.030000
25%	0.100000
50%	0.200000
75%	0.500000
max	3.000000

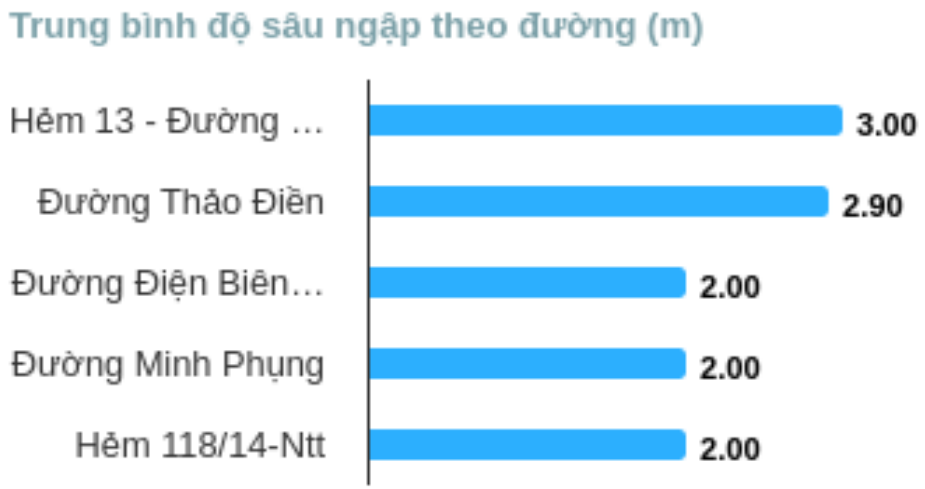


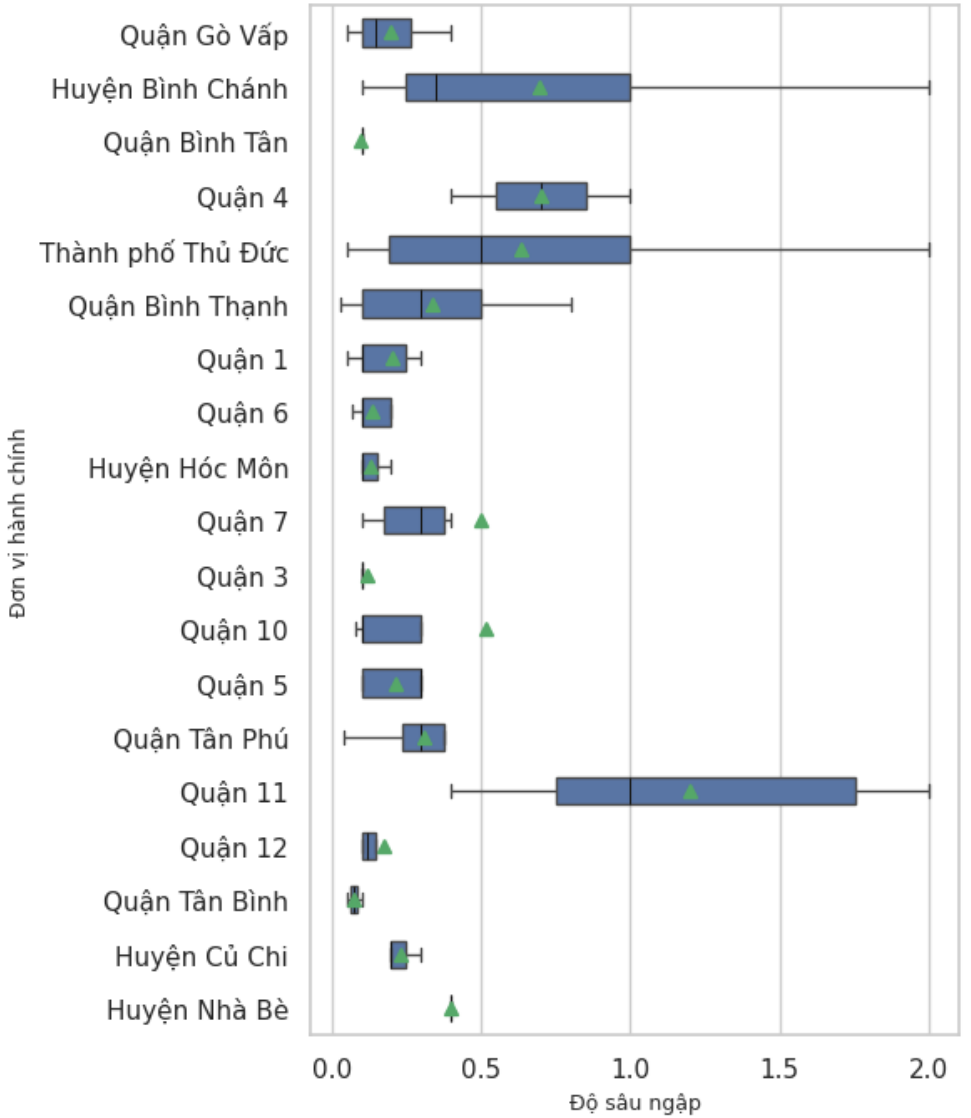


Loại đồ thị nào thể hiện đầy đủ thông tin hơn về (các) cột dữ liệu đang được tìm hiểu?

EDA - Phân tích mô tả

Phân tích và hiển thị dữ liệu cặp biến - dữ liệu mẫu từ nền tảng FEDS

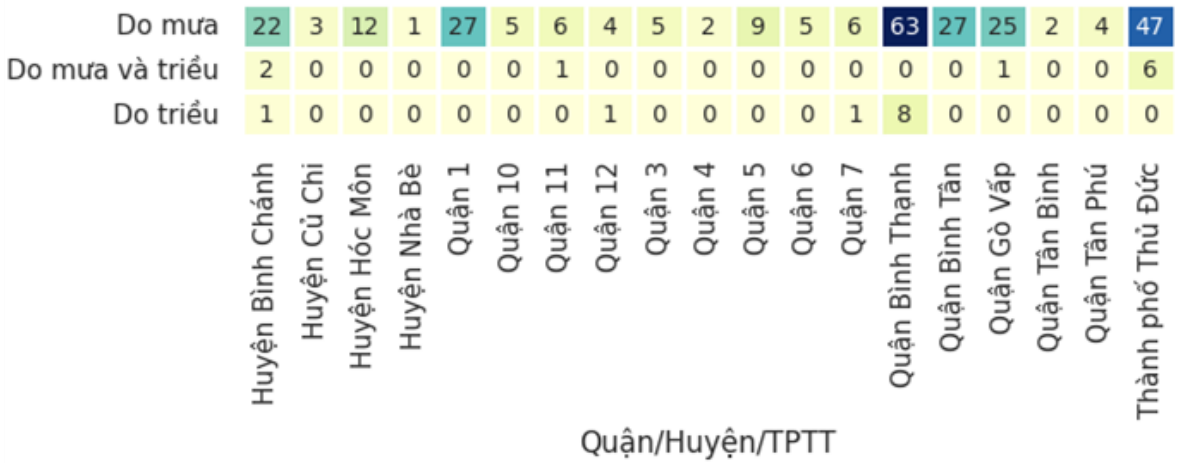


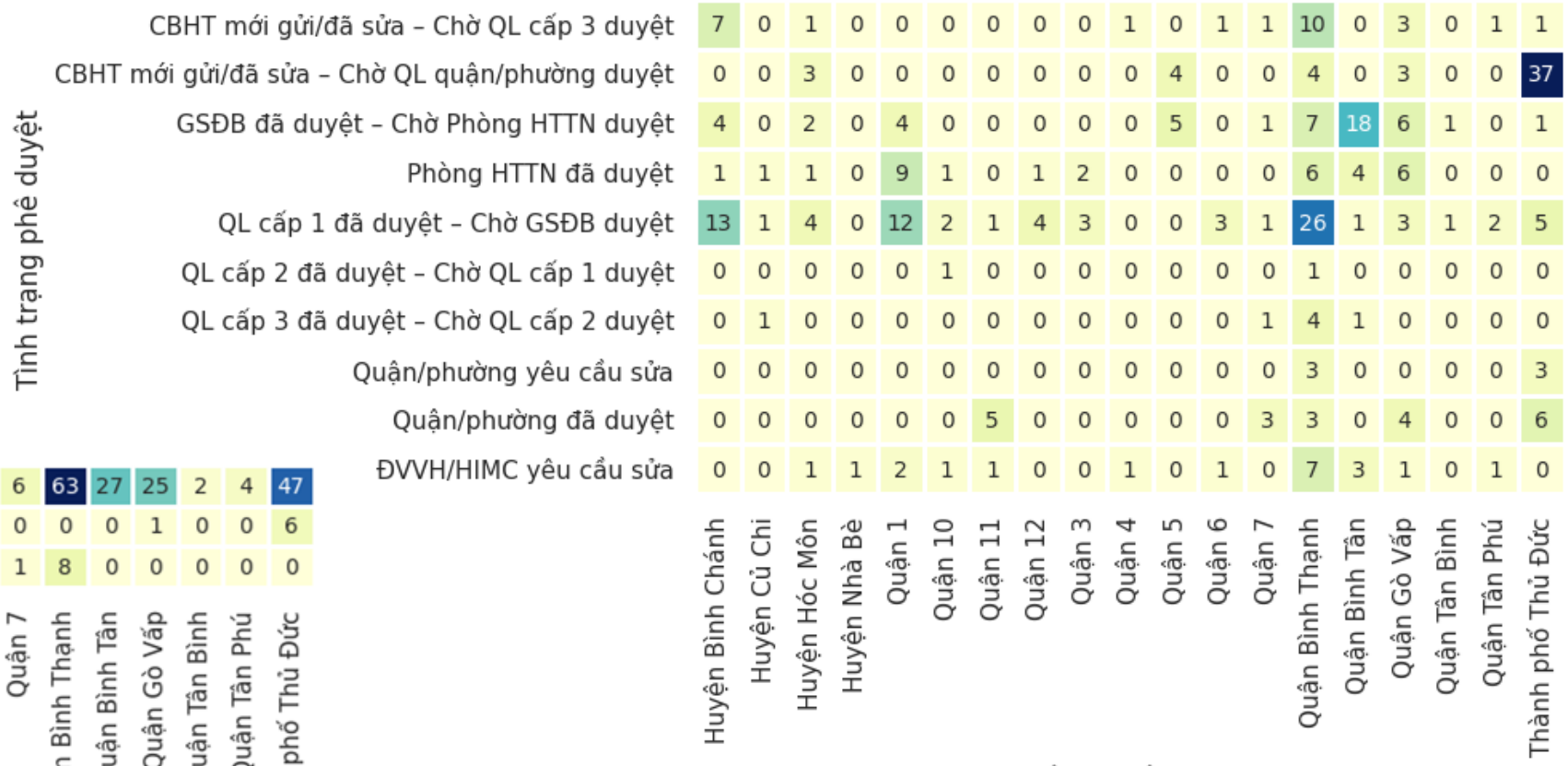


Loại đồ thị nào thể hiện đầy đủ thông tin hơn về (các) cột dữ liệu đang được tìm hiểu?

EDA - Phân tích mô tả

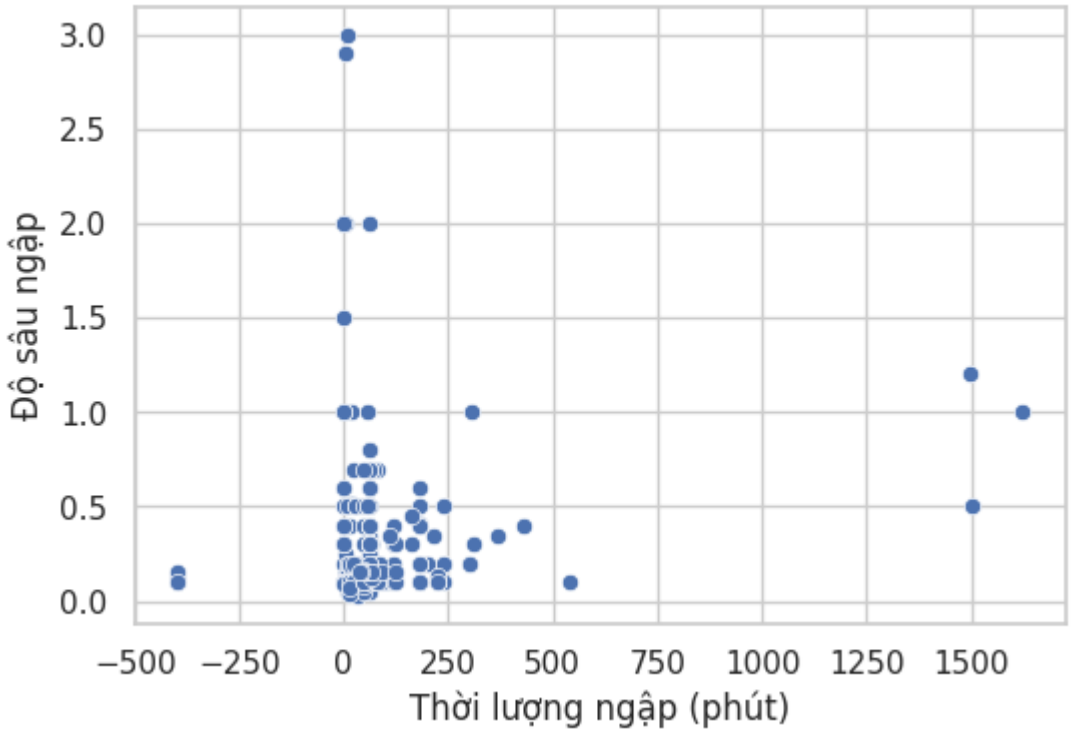
Phân tích và hiển thị dữ liệu cặp biến - dữ liệu mẫu từ nền tảng FEDS

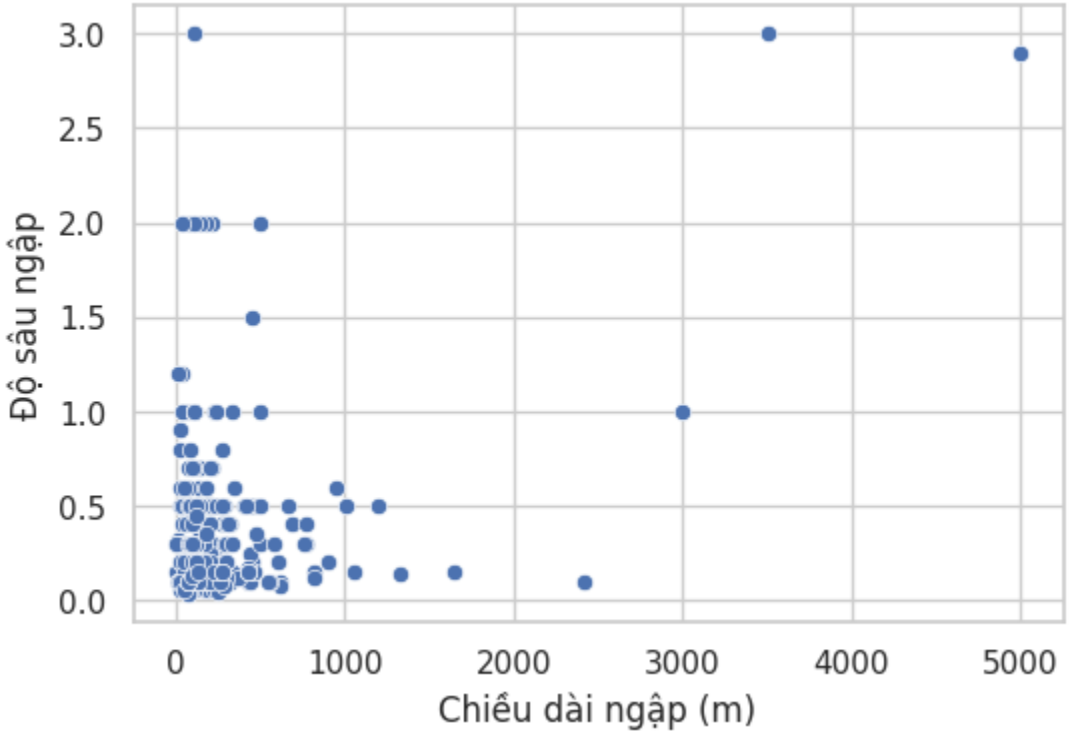




EDA - Phân tích mô tả

Phân tích và hiển thị dữ liệu cặp biến - dữ liệu mẫu từ nền tảng FEDS





Thông tin được thể hiện trong các đồ thị trên có tương ứng với các hiểu biết trước đây về tình hình ngập đường ở TPHCM không?

EDA - Phân tích suy diễn

- Các phân tích suy luận đưa ra các kết luận về mặt xác suất cho một tổng thể dựa trên các trị thống kê của một tập dữ liệu *được lấy mẫu ngẫu nhiên* từ tổng thể đó.

  - *Ước lượng*. Dùng các trị thống kê của tập dữ liệu mẫu để ước lượng (dự đoán) một thông số nào đó của tổng thể.
  - *Kiểm định giả thuyết*. Dùng các trị thống kê của tập dữ liệu mẫu để kiểm chứng một kết luận về một thông số nào đó của tổng thể.

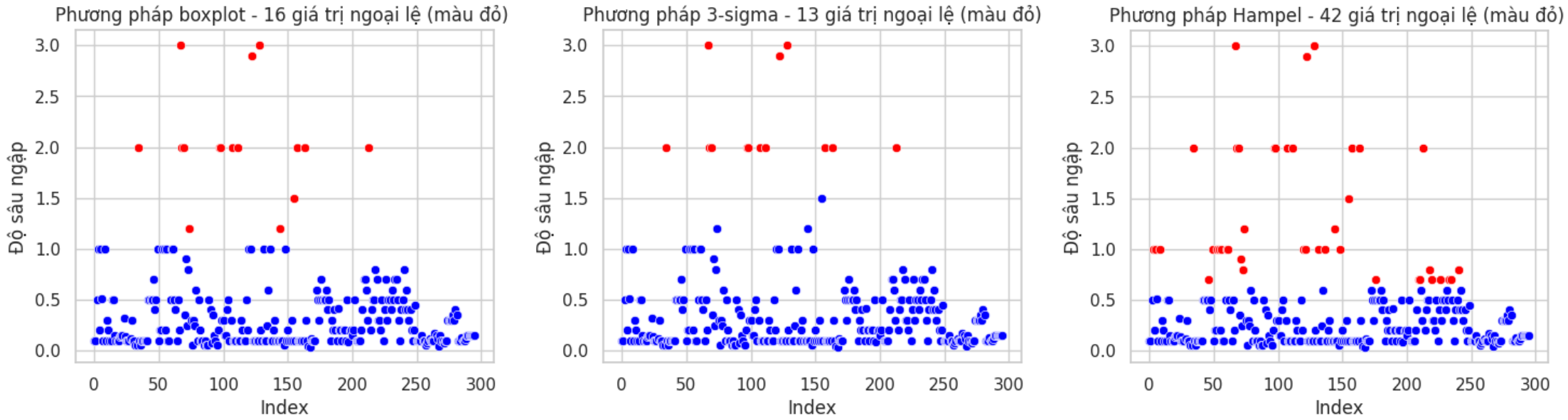
https://online.stat.psu.edu/stat500/lesson/5/5.1

Chúng ta có thể (hoặc có nên) thực hiện các phân tích suy diễn dựa trên dữ liệu về độ sâu ngập của 35 điểm ngập nặng của TPHCM không?

EDA - Các giá trị ngoại lệ

> ... là các quan sát có giá trị có vẻ không nhất quán với các quan sát còn lại của tập dữ liệu đó.
>
> V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 2nd edition, 1984

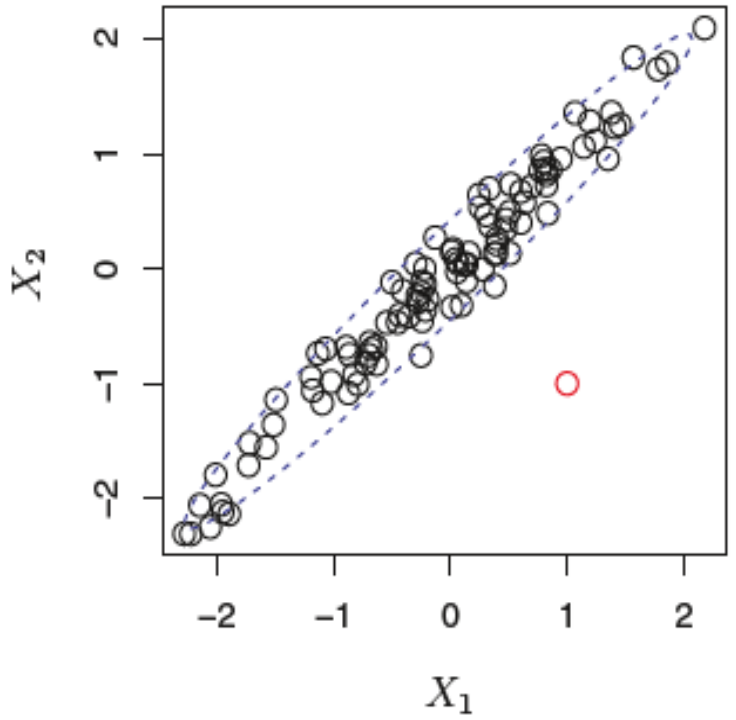
- Các giá trị lớn hoặc nhỏ bất thường làm biến dạng các đặc tính của một tập dữ liệu.
- Các giá trị ngoại lệ có thể do *tổng hợp các lỗi dữ liệu* hoặc là *các trường hợp đặc biệt*.
- Các phương pháp toán *có thể giúp xác định* nhưng *không thể giúp diễn giải* các giá trị ngoại lệ.



Nguyên nhân của các điểm giá trị "Độ sâu ngập" bất thường (màu đỏ trong hình trên) trong dữ liệu mẫu từ nền tảng FEDS là gì?

EDA - Các giá trị ngoại lệ

Xác định các điểm ngoại lệ trong tập dữ liệu nhiều biến?



EDA - Các giá trị inliers

> ... là các giá trị lỗi nhưng nằm trong phạm vi ý nghĩa thống kê của một phân phối xác suất.
>
> D. DesJardins. Paper 169: Outliers, inliers and just plain liars – new eda+ techniques for understanding data. In *Proceedings SAS User's Group International Conference*, SUG126. Cary, NC, USA, 2001

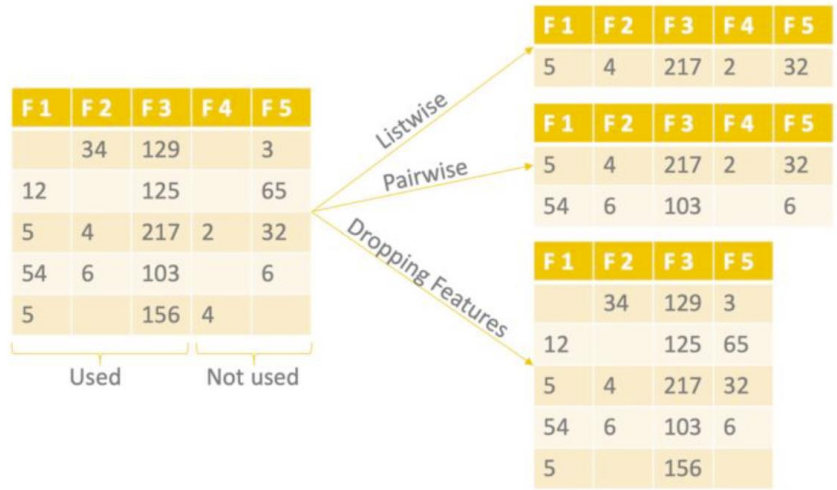
- Các giá trị inliers thường ở dưới dạng các giá trị có tần suất xuất hiện cao bất thường.
- Tương tự các giá trị ngoại lệ, các giá trị inliers không nhất thiết là các điểm dữ liệu xấu, nhưng cần được tìm hiểu thêm.

	Chiều dài ngập (m)	count
0	150.00	10
1	100.00	9
2	200.00	6
3	435.58	6
4	232.68	6

Nguyên nhân của các giá trị 'Chiều dài ngập (m)' có tần suất xuất hiện bất thường được xác định như trên là gì?

EDA - Các giá trị bị mất

- Mất giá trị có thể do lỗi con người, lỗi thiết bị, hoặc do tích hợp các cơ sở dữ liệu.
- Các giá trị ngoại lệ và inliers, nếu được xác định là giá trị xấu, cũng có thể xem bị mất.
- Các giá trị bị mất có thể được phân thành 3 loại
  - Mất ngẫu nhiên hoàn toàn (Missing completely at random)
  - Mất ngẫu nhiên (Missing at random)
  - Mất không ngẫu nhiên (Missing not at random)
- Các cách xử lý dữ liệu bị mất
  - xóa các hàng dữ liệu hoặc cột dữ liệu bị mất giá trị, hoặc
  - thay thế các dữ liệu bị mất
    - phụ thuộc vào loại hình giá trị bị mất.
    - không dùng 1 giá trị duy nhất để thay thế các giá trị bị mất.



https://www.kdnuggets.com/2020/09/missing-value-imputation-review.html

Cột dữ liệu 'Thời lượng ngập' trong tập dữ liệu mẫu từ nền tảng FEDS có 23 giá trị bị mất. Những câu hỏi chúng ta cần trả lời để giúp xử lý phù hợp các giá trị bị mất này là gì?

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.578.5816&rep=rep1&type=pdf
http://www.stat.columbia.edu/~gelman/arm/missing.pdf

Dữ liệu và phân tích dữ liệu phục vụ ra quyết định







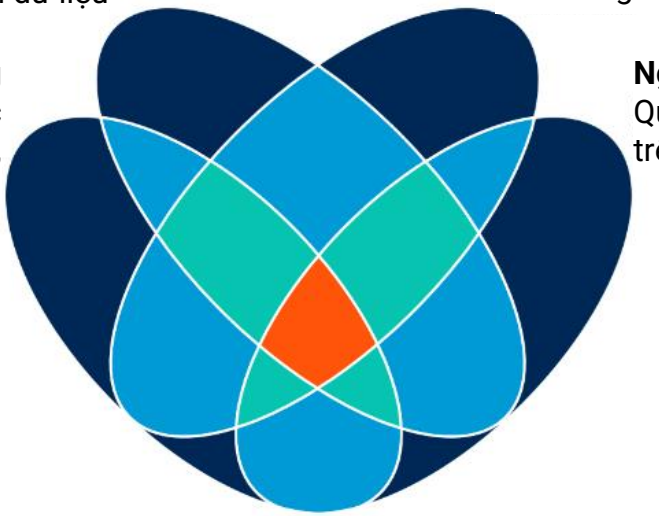
Dữ liệu và phân tích dữ liệu phục vụ ra quyết định

Hợp lý
Quyết định được thực hiện dựa trên dữ liệu

Hợp tác
Quyết định được thực hiện bởi hai người (hoặc hơn)

Bản năng
Quyết định được thực hiện dựa trên cảm tính, kiến thức, kinh nghiệm.

Nguyên tắc
Quyết định được thực hiện dựa trên các nguyên tắc có sẵn



© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Gartner

Source: Gartner

Dữ liệu và phân tích dữ liệu phục vụ ra quyết định

Con người + Máy > **Con người hoặc Máy**

Source: Fin(Legal) Tech lecture, Daniel Katz (2017)

- *Dữ liệu (và thông tin) thu được từ nền tảng FEDS nhằm phục vụ việc ra những quyết định nào?*
- *Những cá nhân hoặc sở ban ngành nào tham gia vào quá trình ra quyết định này?*
- *Việc ra những quyết định này được thực hiện như thế nào, trước và sau khi có (dữ liệu và thông tin từ) nền tảng FEDS?*

Thảo luận

Fulbright
SCHOOL OF PUBLIC POLICY AND MANAGEMENT

Disruptive Technologies
for Public Asset Governance

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Embassy of Switzerland
Swiss Cooperation Office in Vietnam

Cảm ơn

Contact: *Huỳnh Nhật Nam – Giảng viên, Trường Chính sách công và Quản lý Fulbright, nam.huynh@fulbright.edu.vn*
Vũ Thúy Vinh – Quản lý dự án, Trường Chính sách công và Quản lý Fulbright, vinh.vu@fulbright.edu.vn