CHƯƠNG 2

QUY TRÌNH KHÁM PHÁ – PHÂN TÍCH - XỬ LÝ DỮ LIỆU

THS. TRẦN THỊ KIM CHI

NỘI DUNG

1. Tổ chức dữ liệu
2. Dữ liệu định tính và dữ liệu định lượng
3. Thang đo tỷ lệ
4. EDA – Khám phá phân tích dữ liệu

Data (dữ liệu) và information (thông tin)

- **Data:** sự biểu diễn của các đối tượng và sự kiện (văn bản, hình ảnh, âm thanh,…) được ghi nhận, có ý nghĩa không rõ ràng và được lưu trữ trên các phương tiện của máy tính.
- **Information:** dữ liệu đã được xử lý để làm tăng sự hiểu biết của người sử dụng.

Sự khác biệt giữa Data và Information

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

Dữ liệu

Baker, Kenneth D.	324917628
Doyle, Joan E.	476193248
Finkle, Clive R.	548429344
Lewis, John C.	551742186
McFerran, Debra R.	409723145

Thông tin

Class Roster

Course: MGT 500 Business Policy Semester: Spring 2010

Section: 2

Name	ID	Major	GPA
Baker, Kenneth D.	324917628	MGT	2.9
Doyle, Joan E.	476193248	MKT	3.4
Finkle, Clive R.	548429344	PRM	2.8
Lewis, John C.	551742186	MGT	3.7
McFerran, Debra R.	409723145	IS	2.9
Sisneros, Michael	392416582	ACCT	3.3

Business Injection

Tập dữ liệu

- ✓ **Một tập dữ liệu (dataset)** là một tập hợp các đối tượng (objects) và các thuộc tính của chúng
- ✓ Mỗi thuộc tính (attribute) mô tả một đặc điểm của một đối tượng
- ✓ **Ví dụ:** Các thuộc tính Refund, Marital, Status, Taxable Income, Cheat
- ✓ Một tập các giá trị của các thuộc tính sẽ mô tả một đối tượng
- ✓ Khái niệm đối tượng còn được tham chiếu đến các tên gọi khác như: bản ghi (record), điểm dữ liệu (data point), trường hợp (case), mẫu (sample), thực thể (entity) hoặc thể hiện – ví dụ (instance)

Các thuộc tính

Các đối tượng

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(Tan, Steinbach, Kumar - Introduction to Data Mining)

Các kiểu tập dữ liệu

Bản ghi (Record)

- Các bản ghi trong csdl quan hệ
- Ma trận dữ liệu
- Biểu diễn văn bản (document)
- Dữ liệu giao dịch

Đồ thị (Graph)

- World Wide Web
- Mạng thông tin, hoặc mạng xã hội
- Các cấu trúc phân tử (Molecular structures)

Có thứ tự (Ordered)

- Dữ liệu không gian (vd: bản đồ)
- Dữ liệu thời gian (vd: time-series data)
- Dữ liệu chuỗi (vd: chuỗi giao dịch)
- Dữ liệu chuỗi di truyền (genetic sequence data)

	team	coach	play	ball	score	game	win	lost	timeout	season
Document1	3	0	5	0	2	6	0	2	0	2
Document2	0	7	0	2	1	0	0	3	0	0
Document3	0	1	0	0	1	2	2	0	3	0

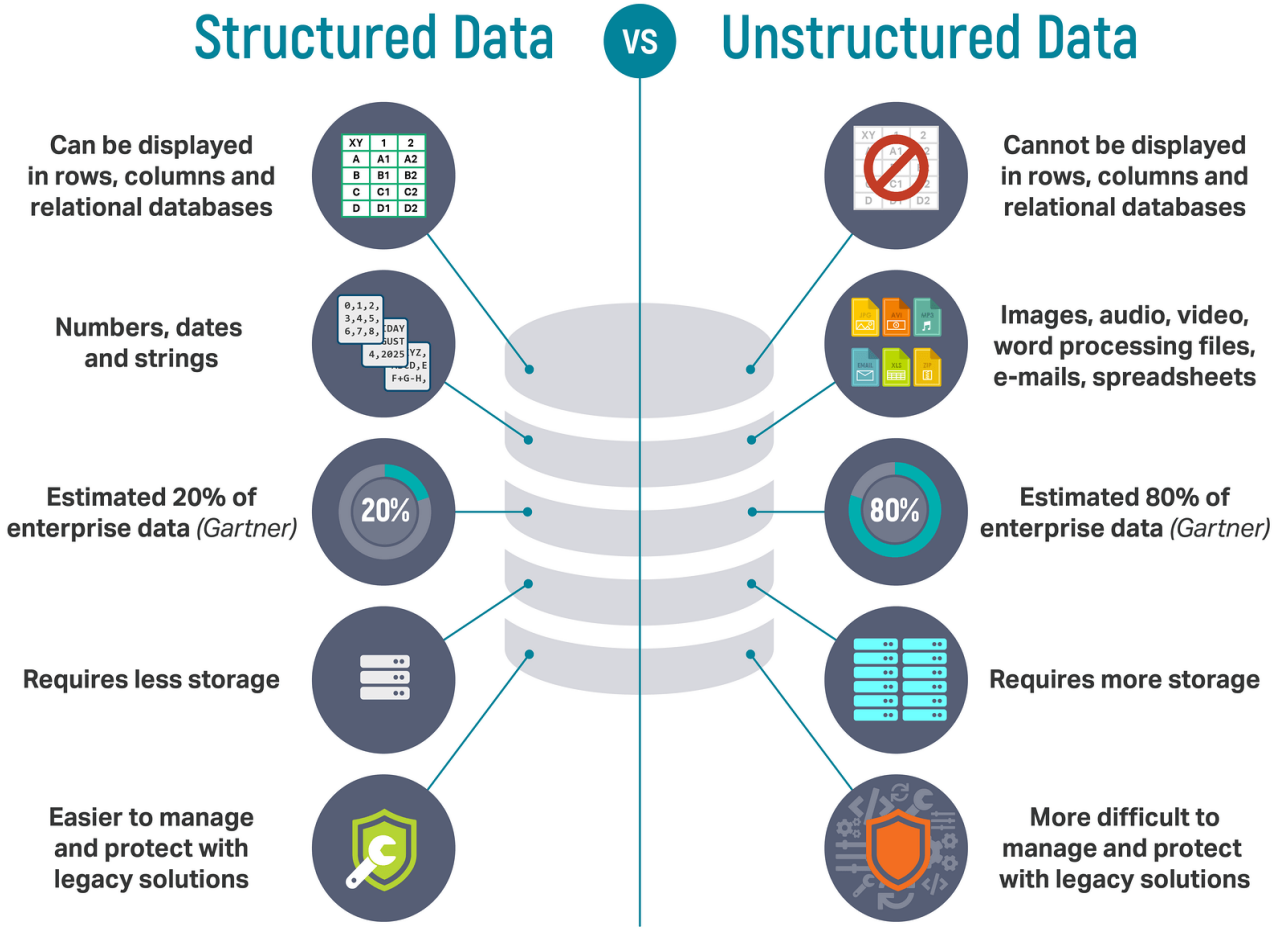
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

(Han, Kamber - Data Mining: Concepts and Techniques)

Các loại hình dữ liệ

Dữ liệu có cấu trúc: Số, ngày, chuỗi ký tự, …

Date	Gold price / ounce
04/09/2020	1500
05/09/2020	1750
06/09/2020	2000
07/09/2020	2250



Dữ liệu không cấu trúc:

- Hình ảnh, âm thanh, đoạn phim…

Dữ liệu bán cấu trúc

Structured data	Semi-structured data	Unstructured data
Databases	XML / JSON data Email Web pages	Audio Video Image data Natural language Documents

TỔ CHỨC DỮ LIỆU

3) Dữ liệu có những dạng cơ bản nào?

Dữ liệu có cấu trúc (structured data)

- Dữ liệu có cấu trúc là loại dữ liệu có tổ chức rõ ràng, là các dạng dữ liệu mà các thành phần của chúng được tổ chức thành các bảng, các trường và các cột.
- **Ví dụ:** các bảng dữ liệu trong các hệ quản trị cơ sở dữ liệu, tập tin Excel hoặc các tài liệu XML có cấu trúc.

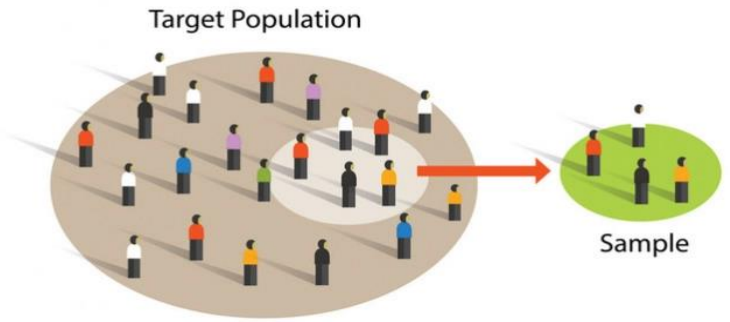
Dữ liệu không có cấu trúc (unstructured data)

- Dữ liệu không có cấu trúc là loại dữ liệu không tuân theo các quy tắc và tiêu chuẩn cụ thể, không có cấu trúc rõ ràng hoặc không tổ chức theo bất kỳ cấu trúc nào.
- **Ví dụ:** các tài liệu văn bản tự do, tài liệu HTML, email, tài liệu PDF, hình ảnh và video.

Tổng thể - Mẫu

- **Tổng thể** là tập hợp tất cả các đối tượng mà ta nghiên cứu.
- **Các đơn vị** (hay phần tử) tạo thành tổng thể được gọi là **đơn vị tổng thể**.
- **Mẫu** là một bộ phận lấy ra từ tổng thể hay là một tập con của tổng thể.

- **Ví dụ:** Để nghiên cứu điểm trung bình môn Toán của sinh viên Trường ĐH IUH, người ta đã xét bảng điểm ngẫu nhiên của 250 sinh viên. Hãy chỉ ra
  - Tổng thể ? Toàn bộ SV của trường ĐH IUH
  - Đơn vị tổng thể ? Sinh viên
  - Mẫu ? 250



Ví dụ

Công ty Gallup khảo sát 1013 người trưởng thành trong 241,742,385 người trưởng thành ở Mỹ. Kết quả có 66% người phản hồi lo lắng về hành vi đánh cắp thông tin cá nhân.

Tổng thể bao gồm:

241,742,385 người trưởng thành ở Mỹ

Mẫu gồm:

1013 người được khảo sát

Đơn vị tổng thể:

Người trưởng thành

Mục tiêu của khảo sát dùng từ dữ liệu thu nhập được để rút ra kết luận về toàn bộ quần thể

Bài tập – Xác định đơn vị tổng thể

Ví dụ 1: Tổng thể các công ty có cổ phiếu niêm yết trên sàn giao dịch chứng khoán TP. Hồ Chí Minh.

Ví dụ 2: Tổng thể các cổ đông của công ty A.

Ví dụ 3: Tổng thể những doanh nghiệp có hoạt động làm ăn phi pháp.

Ví dụ 4: Tổng thể những người thích xem phim truyền hình.

Biến – Dữ liệu

Biến (tiêu thức – statistical criteria): là khái niệm dùng để chỉ các đặc điểm của đơn vị tổng thể được lựa chọn để phục vụ cho mục đích nghiên cứu.

Dữ liệu: là kết quả, giá trị quan sát được của các biến.

- **Ví dụ:** Để nghiên cứu về chi tiêu của sinh viên trường ĐH Công Nghiệp, ta cần nghiên cứu các ***biến*** (hay các ***tiêu thức***) như: giới tính, tuổi, dân tộc, ngành học, số tiền chi tiêu trong một tháng và tiền làm thêm.
- Hãy xác định
  - Biến
  - Dữ liệu

Quan sát

Quan sát: tập hợp tất cả các dữ liệu thu thập được của một đơn vị tổng thể hay mẫu.

- Ví dụ: Quan sát mẫu
  - Sinh viên 1 (quan sát 1): giới tính: nam ; tuổi:20 ; dân tộc:Kinh ; ngành học:401 ; tiền chi tiêu trong tháng: 2,5 triệu đồng, lương: 5 triệu
  - Sinh viên 2 (quan sát 2): giới tính: nữ ; tuổi:21 ; dân tộc:Tày ; ngành học:402 ; tiền chi tiêu trong tháng: 2 triệu đồng, lương: 6 triệu
- Hãy biểu diễn bản ghi (records) cho các quan sát trên: Biến số -> attribute, column ; Quan sát -> object, row

Row x Column = 2 x 6

Column , variable, attribute

Giới tính	Tuổi	Dân tộc	Ngành học	Tiền chi tiêu	Lương
nam	20	Kinh	401	2,500,000	5,000,000
nữ	21	Tày	402	2,000,000	6,000,000

Header (tiêu đề)

Row, object

DỮ LIỆU ĐỊNH LƯỢNG (Quantitative data)

Dữ liệu định lượng (quantitative data or numerical data): là dữ liệu có thể đo đếm được, thường sử dụng thang đo khoảng hay thang đo tỷ lệ.

Ví dụ: tuổi, cân nặng, chiều cao, thu nhập…

Dữ liệu định lượng có thể phân biệt thành 2 loại đó là dữ liệu có giá trị **rời rạc** hay dữ liệu có giá trị **liên tục**.

- **Dữ liệu rời rạc (discrete data):** giá trị dữ liệu là các số nguyên.
  - Ví dụ: Số trẻ em trong trường tiểu học ABC
- **Dữ liệu liên tục (continuous data):** giá trị dữ liệu là các số thực
- Ví dụ: từ 50 đến 72 inch, thực sự có hàng triệu chiều cao có thể có: 52.04762 inch, 69.948376 inch, v.v…

DỮ LIỆU ĐỊNH LƯỢNG (Quantitative data)

Cơ sở để so sánh	Dữ liệu rời rạc	Dữ liệu liên tục
Ý nghĩa	Dữ liệu rời rạc đó là dữ liệu có khoảng cách rõ ràng giữa các giá trị.	Dữ liệu liên tục là dữ liệu rơi vào một chuỗi liên tục.
Xác định	Đếm	Đo lường được
Giá trị	Chứa những giá trị riêng biệt hoặc riêng biệt	Bao gồm bất kỳ giá trị nào trong phạm vi.
Đại diện đồ họa	Thanh biểu đồ	Biểu đồ
Tabulation (việc lập bảng dữ liệu) được gọi là	Phân phối tần số chưa được nhóm.	Phân phối tần số theo nhóm.
Phân loại	Bao gồm lẫn nhau	Độc quyền lẫn nhau, phân loại chồng chéo hoặc là loại trừ lẫn nhau
Biểu đồ chức năng	Hiển thị những điểm bị cô lập	Hiển thị những điểm được kết nối
Thí dụ	Các ngày trong tuần, Số lượng công nhân trong một công ty.	Chiều cao của trẻ em, Tốc độ của ô tô.

DỮ LIỆU ĐỊNH LƯỢNG (Quantitative data)

Bài tập: Xác định các loại dữ liệu

+ Màu mắt có thể thuộc một trong các loại sau: xanh lam, xanh lục, nâu.

+ Số lượng học sinh trong một lớp học.

+ Khoảng thời gian cần thiết để hoàn thành một dự án.

+ Chiều cao của trẻ em.

+ Lượng mưa, tính bằng inch, rơi trong một cơn bão.

+ Số lượng công nhân trong một công ty.

+ Số lượng các bộ phận bị hư hỏng trong quá trình vận chuyển.

+ Kích cỡ giày.

+ Các diện tích vuông của một ngôi nhà hai phòng ngủ.

+ Trọng lượng của một chiếc xe tải.

+ Số lượng ngôn ngữ mà một cá nhân nói.

+ Số lần chạy nhà trong một trận đấu bóng chày.

+ Số câu hỏi kiểm tra bạn trả lời đúng.

+ Tốc độ của ô tô.

+ Thời gian thức dậy.

+ Dụng cụ trong kệ.

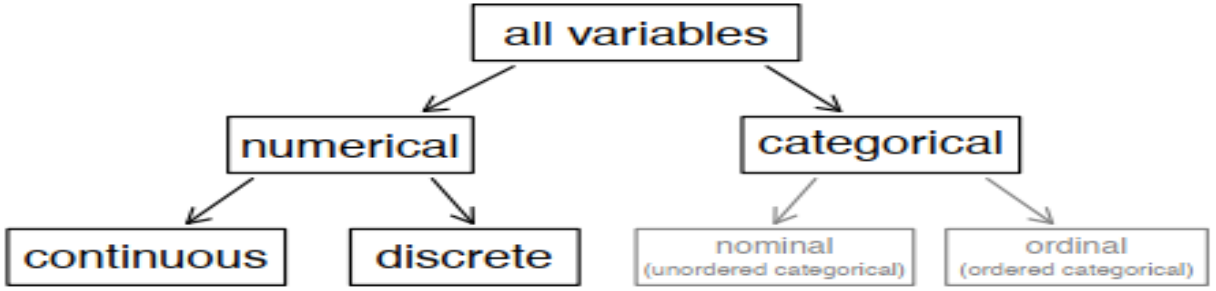
+ Số anh chị em mà một cá nhân được chọn ngẫu nhiên có.

DỮ LIỆU ĐỊNH TÍNH (Qualitative data)

Dữ liệu định tính hay dữ liệu phân loại (qualitative data or categorical data): là dữ liệu sử dụng để phân loại, giá trị của dữ liệu được sử dụng để đại diện cho một phân loại nào đó. Thường sử dụng thang đo định danh hay thang đo thứ bậc.

Ví dụ: giới tính, màu sắc, xếp hạng…

Dữ liệu định tính có thể chia làm hai loại đó là: dữ liệu định tính **có thứ tự** và dữ liệu định tính **không có thứ tự**



Biến định tính và biến định lượng

Biến định tính (qualitative variable) còn gọi là biến phân loại (categorical variable) : phản ánh tính chất, loại hình, không thể hiện trực tiếp bằng các con số (nhưng có thể biểu diễn bằng số).

Ví dụ: phân loại kỹ năng làm việc nhóm của sinh viên 1: Giao tiếp , 2. Làm việc nhóm.

Các nhóm tuổi: (1) dưới 22 tuổi, (2) từ 22 đến 30 tuổi, (3) từ 31 đến 50 tuổi, (4) trên 50 tuổi. Đây

Biến định lượng (quantiative variable) hay *tiêu thức số lượng*: biểu hiện trực tiếp bằng con số. Biến định lượng chia làm hai loại là liên tục và rời rạc.

Ví dụ: Phân loại nhóm biến định tính và biến định lượng trong các biến sau: *giới tính, tuổi, dân tộc, ngành học, số tiền chi tiêu trong tháng và lương*

So sánh biến định tính và biến định lượng

Dữ liệu định tính	Dữ liệu định lượng
- Phản ánh tính chất, sự hơn kém - Không tính được giá trị trung bình - Được thể hiện dưới nhiều cách thức khác nhau. VD • Xếp loại học tập : Giỏi – Khá – Trung bình – Yếu • Giới tính : Nam – Nữ • Giới tính : • Nữ → 0 hoặc F • Nam → 1 hoặc M	- Phản ánh mức độ của sự hơn kém - Tính được giá trị trung bình - Được thể hiện bằng các con số cụ thể VD • Tuổi tác, thu nhập, điểm số…

Biến độc lập và biến phụ thuộc

Biến độc lập (**Independent variable)**. *Biến độc lập* là biến số tác động tới biến số khác (biến phụ thuộc) trong một mô hình kinh tế. Được biểu diễn trên trục abscissa (x) trong biểu đồ.

Ví dụ 1: Biến độc lập có thể là liều lượng phân bón, loại phân bón, lượng nước tưới, thời gian chiếu sáng khác nhau,… (hay còn gọi là các nghiệm thức khác nhau).

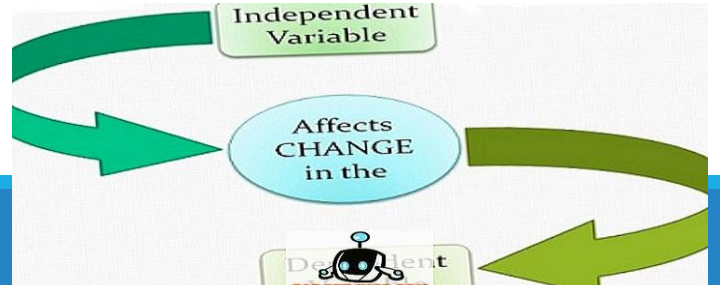
Ví dụ 2: Giá hàng hoá là biến số độc lập ảnh hưởng tới lượng cầu về hàng hóa đó.

Biến phụ thuộc (**Dependent variable)**. *Biến phụ thuộc* là biến số chịu ảnh hưởng của một biến số khác trong mô hình. Được biểu diễn trên trục abscissa (y) trong biểu đồ.

Ví dụ, nhu cầu về một hàng hoá bị ảnh hưởng bởi giá cả của nó.

Ví dụ: Khi nghiên cứu sự sinh trưởng của cây mía, các biến phụ thuộc ở đây có thể bao gồm: chiều cao cây, số lá, trọng lượng cây… và kết quả đo đạc của biến phụ thuộc ở các nghiệm thức khác nhau có thể khác nhau.

Hãy cho biết biến nào là độc lập, biến nào là phụ thuộc: Doanh số bán hàng, chi phí quảng cáo, lợi nhuận gộp, chi phí đầu tư.



Biến độc lập và biến phụ thuộc

Bài tập: xác định các biến độc lập và phụ thuộc

1. Nghiên cứu để nghiên cứu ảnh hưởng của việc tiêu thụ rượu đối với huyết áp. Lượng rượu tiêu thụ hàng ngày sẽ là biến số độc lập (nguyên nhân) ĐL và huyết áp sẽ là biến số phụ thuộc (tác động). PT
2. Tác dụng của thuốc lá đối với sức đề kháng vật lý. Sử dụng thuốc lá ĐL , kháng thể chất . PT
3. Ảnh hưởng của tiêu thụ đường đến trọng lượng. Tiêu thụ đường ĐL , trọng lượng PT
4. Đo lường tác động của chiều cao lên cân nặng. Chiều cao là ĐL và cân nặng là PT
5. Một nghiên cứu, "bình luận tích cực" và "lòng tự trọng“

   Nhận xét tích cực gây ra thay đổi lòng tự trọng và lòng tự trọng không thể gây ra thay đổi trong bình luận tích cực.

6. Trong "Tiếp xúc với ánh sáng mặt trời nhiều hơn làm tăng mức độ hạnh phúc ở những người lao động cả ngày trong văn phòng kín",

   Nhận xét: phơi nắng sẽ dẫn đến biến số độc lập và mức độ hạnh phúc sẽ phụ thuộc.

7. Trong câu hỏi "Lợi ích hoặc tác hại của các mạng xã hội ở trẻ em là gì?",

   Mạng xã hội rõ ràng có thể được xác định là biến độc lập, bởi vì nó được cho là gây ra tác dụng có lợi hoặc tác hại đối với trẻ em.

So sánh biến độc lập và phụ thuộc

CƠ SỞ SO SÁNH	BIẾN ĐỘC LẬP	BIẾN PHỤ THUỘC
Ý nghĩa	Biến độc lập là một biến có giá trị được nhà nghiên cứu thay đổi có chủ ý nhằm đạt được kết quả mong muốn.	Biến phụ thuộc đề cập đến một biến thay đổi giá trị của nó để thay đổi qua lại các giá trị của biến độc lập.
Nó là gì?	Tiền đề	Hệ quả
Mối quan hệ	Nguyên nhân được cho là	Hiệu quả quan sát
Giá trị	Thao tác bởi các nhà nghiên cứu.	Đo bởi nhà nghiên cứu.
Thường được ký hiệu là	x	y, f(x)

Lượng biến

Kiểu thuộc tính rời rạc (Discrete-valued attributes) hay còn gọi là lượng biến rời rạc:

- Giá trị có thể nhận một trị số cụ thể, có thể đếm được.
- Tập các giá trị là một tập hữu hạn
- Bao gồm cả các thuộc tính có kiểu giá trị là các số nguyên
- Bao gồm cả các thuộc tính nhị phân (binary attributes)
- Ví dụ: Số lượng học sinh trong 1 lớp, số người trong gia đình...

Kiểu thuộc tính liên tục (Continuous-valued attributes) hay còn gọi là lượng biến liên tục

- Giá trị có thể có của nó có thể lấp đầy một khoảng trên trục số.
- Các giá trị là các số thực (real numbers)
- Ví dụ: Chiều cao, cân nặng của các bạn trong lớp...

Lượng biến

Kiểu định danh/chuỗi (nominal): không có thứ tự

- Lấy giá trị từ một tập không có thứ tự các giá trị (định danh)
- Vd: Các thuộc tính như: Name, Profession, …

Kiểu nhị phân (binary): là một trường hợp đặc biệt của kiểu định danh

- Tập các giá trị chỉ gồm có 2 giá trị (Y/N, 0/1, T/F)

Kiểu có thứ tự (ordinal):

- Lấy giá trị từ một tập có thứ tự các giá trị
- Vd1: Các thuộc tính lấy giá trị số như: Age, Height,…
- Vd2: Thuộc tính Income lấy giá trị từ tập {low, medium, high}

Loại hình phân tích dữ liệu

Trong lĩnh vực tiếp thị, kinh doanh, xã hội học, tâm lý học, khoa học & công nghệ, kinh tế, v.v ... có hai cách tiến hành nghiên cứu tiêu chuẩn, đó là nghiên cứu định tính hoặc nghiên cứu định lượng.

Nghiên cứu định tính dựa vào tường thuật bằng lời nói như dữ liệu nói hoặc viết

Nghiên cứu định lượng sử dụng các quan sát logic hoặc thống kê để đưa ra kết luận.

Loại hình phân tích dữ liệu

Phân tích dữ liệu trong nghiên cứu định tính

- Phân tích và nghiên cứu dữ liệu thông tin chủ quan (subjective information) tốt hơn thông tin số. Bởi vì thông tin bao gồm từ ngữ, sự mô tả, hình ảnh, đồ vật. Thu thập kiến thức từ dữ liệu vướng víu như vậy rất khó khăn; do đó, nó thường được sử dụng để nghiên cứu khám phá cũng như phân tích dữ liệu.

Tìm kiếm các mẫu trong dữ liệu định tính

- Mặc dù có một số cách khác nhau để khám phá các mẫu trong dữ liệu in (printed data), nhưng chiến lược dựa trên từ ngữ là phương pháp được sử dụng rộng rãi và phụ thuộc nhất để nghiên cứu và phân tích dữ liệu.

Đặc biệt, quy trình phân tích dữ liệu trong nghiên cứu định tính được thực hiện thủ công. Ở đây, các chuyên gia đọc thông tin có thể truy cập và tìm các từ đơn điệu hoặc thường được sử dụng.

Loại hình phân tích dữ liệu

Phân tích dữ liệu trong nghiên cứu định lượng

- **Chuẩn bị dữ liệu để phân tích**
  - Giai đoạn đầu tiên trong nghiên cứu và phân tích dữ liệu được thực hiện để kiểm tra với mục tiêu rằng thông tin định danh (nominal information) có thể được thay đổi thành một thứ quan trọng. Việc chuẩn bị dữ liệu bao gồm những bước sau đây.
  - Xác thực dữ liệu (Data Validation)
  - Chỉnh sửa dữ liệu (Data Editing)
  - Mã hóa dữ liệu (Data Coding)
- Đối với nghiên cứu thống kê định lượng, việc phân tích mô tả thường đưa ra những con số tối ưu. Tuy nhiên, phân tích không bao giờ đủ để chỉ ra lý do ẩn sau những con số này.

So sánh nghiên cứu định tính và định lượng

CƠ SỞ SO SÁNH	PHÂN TÍCH ĐỊNH TÍNH	PHÂN TÍCH ĐỊNH LƯỢNG
Ý nghĩa	Nghiên cứu định tính là một phương pháp tìm hiểu phát triển sự hiểu biết về khoa học xã hội và con người, để tìm ra cách mọi người suy nghĩ và cảm nhận.	Nghiên cứu định lượng là một phương pháp nghiên cứu được sử dụng để tạo ra dữ liệu số và dữ kiện cứng, bằng cách sử dụng kỹ thuật thống kê, logic và toán học.
Thiên nhiên	Toàn diện	Đặc biệt
Tiếp cận	Chủ quan	Mục tiêu
Loại nghiên cứu	Thăm dò	Kết luận
Lý luận	Cảm ứng	Khấu trừ
Lấy mẫu	Mục đích	Ngẫu nhiên
Dữ liệu	Bằng lời nói	Đo lường được
Thắc mắc	Định hướng quy trình	Định hướng kết quả
Giả thuyết	Tạo	Thử nghiệm

So sánh nghiên cứu định tính và định lượng

CƠ SỞ SO SÁNH	PHÂN TÍCH ĐỊNH TÍNH	PHÂN TÍCH ĐỊNH LƯỢNG
Các yếu tố phân tích	Từ ngữ, hình ảnh và đồ vật	Dữ liệu số
Mục tiêu	Để tìm hiểu và khám phá những ý tưởng được sử dụng trong các quy trình đang diễn ra.	Để kiểm tra mối quan hệ nguyên nhân và kết quả giữa các biến.
Phương pháp	Các kỹ thuật phi cấu trúc như phỏng vấn sâu, thảo luận nhóm, v.v.	Các kỹ thuật có cấu trúc như khảo sát, bảng câu hỏi và quan sát.
Kết quả	Phát triển sự hiểu biết ban đầu	Đề xuất khóa học hành động cuối cùng

Các cấp bậc đo lường và thang đo

- Với các biến số thuộc vào các nhóm Định Tính hoặc Định Lượng, thì giá trị dữ liệu trong các quan sát lại được phân loại vào các thang đo lường trong thống kê nhằm thể hiện mức độ thông tin có được.
- Trong thống kê người ta sử dụng bốn cấp bậc đo lường theo mức độ thông tin tăng dần, đó là thang đo:
  - ***định danh (nominal),***
  - ***phân loại,***
  - ***thứ bậc (ordinal),***
  - ***khoảng (interval)***
  - ***tỉ lệ (ratio)***.

Thang đo định danh (Nomial)

Thang đo định danh là thang đo gồm các nhãn hay tên để phân biệt các thuộc tính của phần tử, không thể hiện sự hơn kém.

Thang đo này được sử dụng cho các dữ liệu định tính.

Ví dụ

- Mã số
- Số thứ tự (mục đích để đếm)
- Tên

Baker, Kenneth D.

Doyle, Joan E.

Finkle, Clive R.

Lewis, John C.

McFerran, Debra R.

Thang đo định danh (Nomial)

Thang đo phân loại (Categorical)

- Là thang đo định danh. κhông thể hiện sự hơn kém.
- Thang đo này được sử dụng cho các dữ liệu định tính.
- Người ta thường sử dụng các số để phân loại các đối tượng, đây là các mã số dùng để đếm số lần xuất hiện, không phải để so sánh hơn kém

Câu hỏi điều tra: Sinh viên hiện đang sống ở đâu? (Chọn từ 1 đến 4)

1. Sống cùng gia đình
2. Ký túc xá
3. Nhà trọ
4. Trường hợp khác

Refund	Marital Status
Yes	Single
No	Married
No	Single
Yes	Married
No	Divorced
No	Married
Yes	Divorced
No	Single
No	Married
No	Single

Thang đo thứ bậc (Ordinal)

- Là thang đo định danh nhưng thể hiện sự hơn kém của dữ liệu, không biết chính xác mức độ hơn kém đó. Dữ liệu thể hiện tính chất của dữ liệu danh nghĩa và thứ tự hoặc xếp hạng có ý nghĩa.
- Thang đo này được sử dụng cho các dữ liệu định tính và cả định lượng.

Đo thái độ về hành vi nào đó:

1. Hoàn toàn đồng ý
2. Đồng ý
3. Chưa quyết định
4. Hoàn toàn không đồng ý.

Thành tích huy chương trong Olympic:

1. Vàng
2. Bạc
3. Đồng
4. Khuyến khích – ghi nhận

Thang đo thứ bậc (Ordinal)

Thang đo khoảng (Interval)

- Thang đo khoảng là thang đo thứ bậc có khoảng cách đều nhau.
- Thang đo này đánh giá chính xác mức độ hơn kém cụ thể
- Thang đo này được sử dụng cho các dữ liệu định tính và cả định lượng.

Ví dụ: Thu nhập bình quân hàng tháng của bạn là:

1. Từ 1,5 triệu đến 2 triệu
2. Từ 2 triệu đến 2,5 triệu
3. Từ 2,5 triệu đến 3 triệu

→ Khoảng cách đều nhau bằng 500 ngàn đồng

→ Thực hiện được các phép toán cộng trừ.

Thang đo khoảng (Interval)

Examples

Temperature

IQ score

Income ranges

How is interval data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

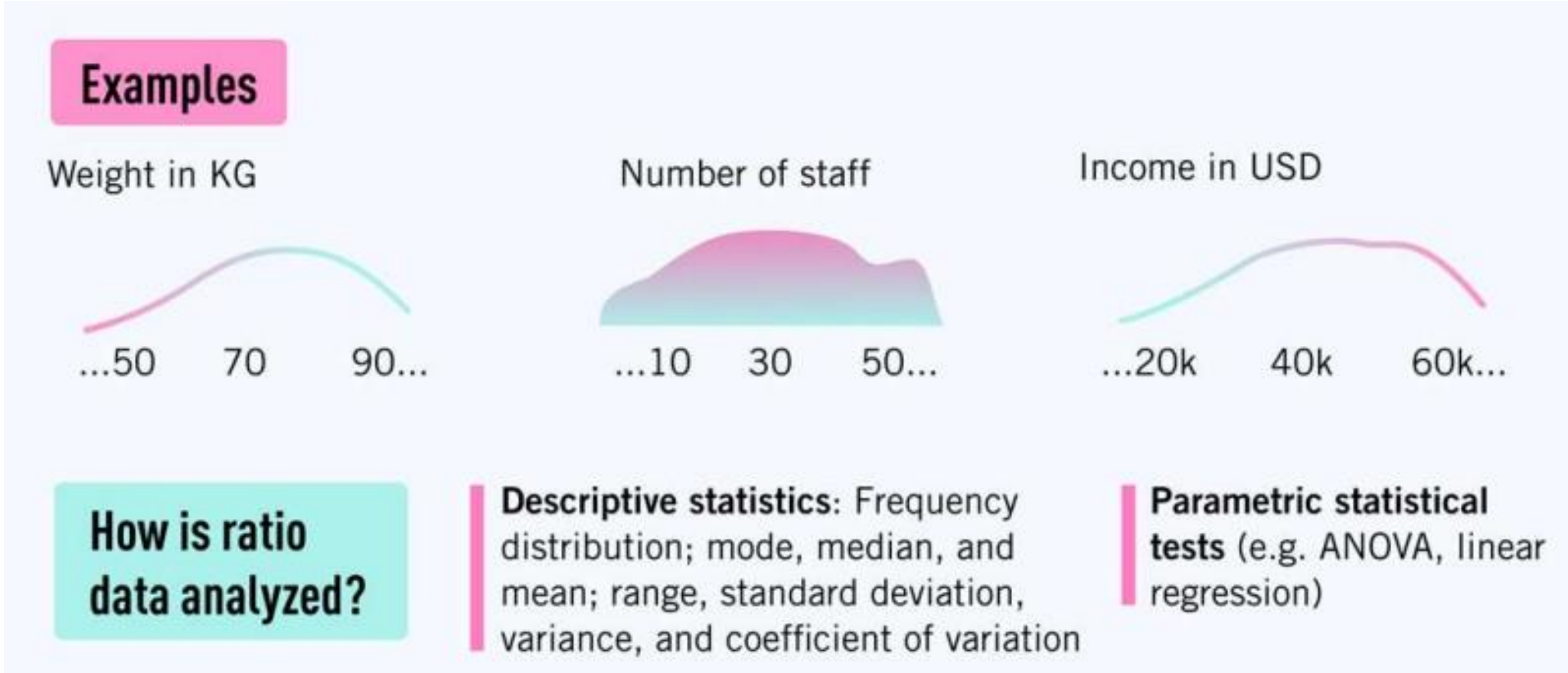
Parametric statistical tests (e.g. t-test, linear regression)

Thang đo tỉ lệ (Ratio)

- Là thang đo mà dữ liệu có tất cả đặc tính của thang đo khoảng và tỷ lệ có ý nghĩa.
- Là loại thang đo dùng cho các dữ liệu định lượng. Đây là thang đo ở bậc cao nhất trong hệ thống thang đo.

Bạn A nặng 80kg. Bạn B nặng 40kg => bạn A nặng gấp đôi bạn B (dù đổi ở bất cứ đơn vị nào).

Thang đo tỉ lệ (Ratio)



Nhận diện các loại thang đo

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Bài tập nhận diện các loại thang đo

Cho biết loại thang đo sử dụng?

1. Vui lòng cho biết thu nhập hàng tháng của anh/chị? .....

2. Vui lòng cho biết mức thu nhập hàng tháng của anh/chị?

1. Không có	2. Dưới 2 triệu	3. Từ 2 - 5 triệu	4. Trên 5 triệu
❑	❑	❑	❑

3. Nghề nghiệp của anh/chị là gì?

1. Công chức	❑	4. Chủ doanh nghiệp	❑	7. Nội trợ	❑
2. Cán bộ quản lý	❑	5. Buôn bán nhỏ	❑	8. Về hưu	❑
3. Nhân viên VP	❑	6. Lao động phổ thông	❑	9. Khác	❑

4. Lần sau anh/chị có chọn siêu thị này nữa không?

Hoàn toàn không chắc chắn	Không chắc chắn	Chưa biết	Chắc chắn	Hoàn toàn chắc chắn
1. ☐	2. ☐	3. ☐	4. ☐	5. ☐

Bài tập nhận diện các loại thang đo

1. Loại điện thoại di động mà bạn sử dụng chính?

☐ Nokia
☐ Samsung
☐ Motorola
☐ Khác

2. Mức độ hài lòng chung của bạn khi sử dụng loại điện thoại trên?

Rất không hài lòng 1 2 3 4 5 Rất hài lòng

3. Chi tiêu trung bình một tháng cho việc gọi điện thoại di động .....ngàn đ

4. Bạn theo dõi thông tin về các loại điện thoại mới như thế nào?

☐ Không bao giờ ☐ Ít khi ☐ Thỉnh thoảng ☐ Thường xuyên

5. Bạn thường sử dụng tính năng nào

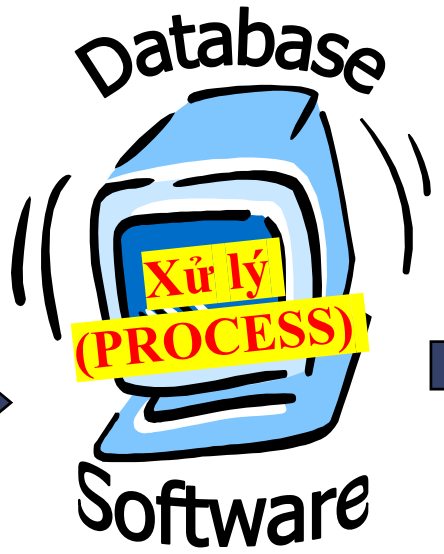
☐ Nghe - gọi ☐ Tin nhắn ☐ Nghe nhạc ☐ Quay phim, chụp hình ☐ Games ☐ Khác

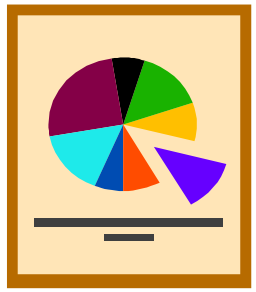
6. Giới tính: ☐Nam ☐Nữ

Qúa trình cơ bản khai thác thông tin từ dữ liệu

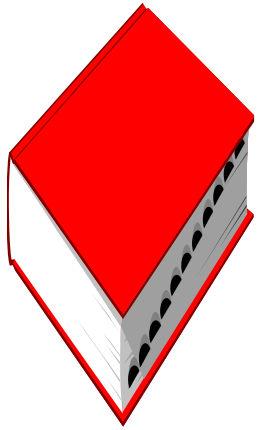
How you can transform meaningless data into business intelligence.

Dữ liệu thô





Dữ liệu tĩnh



Mô tả & Trình bày

Khai thác dữ liệu (EDA) và dự báo

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

- **EDA (viết tắt của Exploratory Data Analyst) – Phân tích khám phá dữ liệu:** là một phương pháp khám phá dữ liệu, tìm ra các xu hướng, mẫu thử hoặc kiểm tra các giả định trong dữ liệu nhằm mục đích hiểu rõ về cấu trúc và tính chất của dữ liệu. Khi áp dụng các thuật toán học máy hoặc xây dựng các mô hình dự đoán.
- EDA góp phần quan trọng trong quá trình xử lý dữ liệu, giúp giải quyết các điều kiện ngoại lệ, giá trị thiếu và những vấn đề ảnh hưởng đến kết quả cuối cùng.

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

Mục đích phân tích khám phá dữ liệu:

- Tìm hiểu về cấu trúc dữ liệu, nắm rõ đặc điểm, cấu trúc và mô hình dữ liệu.
- Hỗ trợ làm sạch dữ liệu với các kỹ thuật xác định các giá trị bị thiếu, sai sót hoặc các điểm dữ liệu bất thường.
- Xác định mối tương quan giữa các biến
- Xây dựng cơ sở dữ liệu quan hệ.
- Phát triển và kiểm chứng các giả thuyết và giả định.
- Xây dựng data model.
- Xác định phạm vi sai lệch.
- Xác định các công cụ thống kê và kỹ thuật phân tích thích hợp nhất.
- Phát hiện các mẫu (pattern), xu hướng thay đổi các biến.
- Hiểu rõ hơn về đặc điểm mô tả của các biến và tập dữ liệu.
- Chuẩn bị cho bước phân tích tiếp theo

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

Các kỹ thuật phân tích chủ yếu được dùng trong EDA:

1. Phân tích đơn biến

Phân tích đơn biến được thực hiện với mục đích là hiểu được sự phân bổ của các giá trị cho một biến duy nhất. Dữ liệu đơn biến không theo loại dữ liệu cụ thể mà được phân theo mục đích sử dụng hoặc bản chất riêng. Để phân tích một tập dữ liệu, các loại kỹ thuật phân tích đơn biến sẽ được sử dụng tùy thuộc vào các loại biến đề cập.

2. Phân tích hai biến

Phân tích hai biến là phương pháp kiểm tra sự liên quan giữa hai dữ liệu khác nhau, cách thức để xác định xem có mối liên hệ nào giữa hai biến hay không, nếu có thì mối liên hệ đó mạnh đến mức nào và thể hiện theo hướng nào. Đây là một kỹ thuật phân tích giúp xác định cách kết nối giữa hai biến và tìm ra xu hướng trong dữ liệu.

3. Phân tích đa biến

Phân tích đa biến kỹ thuật phân tích ở cấp độ phức tạp hơn, được sử dụng khi có nhiều hơn hai biến trong tập dữ liệu. Phân tích đa biến giúp giảm thiểu và đơn giản hóa dữ liệu mà không làm mất bất kỳ chi tiết quan trọng nào trong tập dữ liệu. Điều quan trọng nhất trong phương pháp này là phải hiểu mối quan hệ giữa các biến dự đoán hành vi của các biến dựa trên quan sát.

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

Quy trình các bước thực hiện EDA:

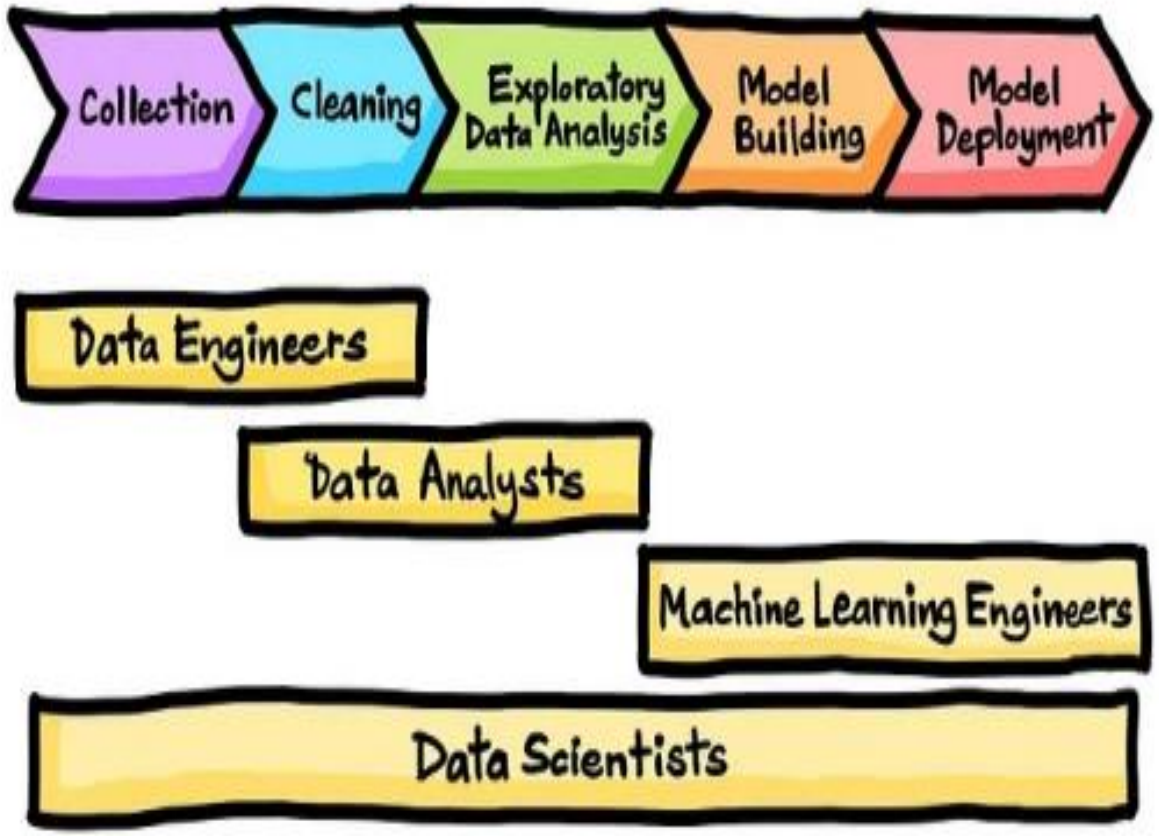
- **Bước 1 - Thu thập dữ liệu**: Thu thập dữ liệu từ các nguồn, sau đó lưu trữ và tổ chức một cách chính xác để các bước tiếp theo được thực hiện một cách nhanh chóng.
- **Bước 2 - Kiểm tra dữ liệu**: Kiểm tra sơ bộ về tệp dữ liệu, xem số lượng, kiểu dữ liệu, thuộc tính dữ liệu và các đặc điểm khác. Quá trình này sẽ giúp các nhà phân tích dữ liệu định hình được các phương án xử lý dữ liệu tiếp theo.
- **Bước 3 - Xử lý dữ liệu**: Ở bước này, các nhà phân tích dữ liệu sẽ thực hiện các phần việc như bổ sung các giá trị thiếu, xóa các giá trị trùng lặp, xử lý các dữ liệu ngoại lệ và chuyển đổi định dạng dữ liệu.

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

Quy trình các bước thực hiện EDA:

- **Bước 4** - **Trực quan dữ liệu**: Sử dụng các kỹ thuật phân tích kết hợp với các biểu đồ để hiểu về các mẫu, xu hướng và mối tương quan giữa các dữ liệu. Tùy vào mối quan hệ giữa các biến để ứng dụng các kỹ thuật phân tích để khai thác điểm đặc trưng của tệp dữ liệu.
- **Bước 5** - **Đúc kết**: Dựa trên các bước đã thực hiện, phân tích và đưa ra kết luận về các dữ liệu đã xử lý. Ghi nhận các mẫu quan trọng đã tìm thấy, trình bày các xu hướng và khía cạnh khác của dữ liệu.
- **Bước 6** - **Báo cáo kết quả**: Sử dụng các biểu đồ phân tích, hình ảnh và các mô tả liên quan để báo cáo kết quả dữ liệu một cách chi tiết và rõ ràng.

EDA (Exploratory Data Analysis)
Phân tích Khám phá Dữ liệu)

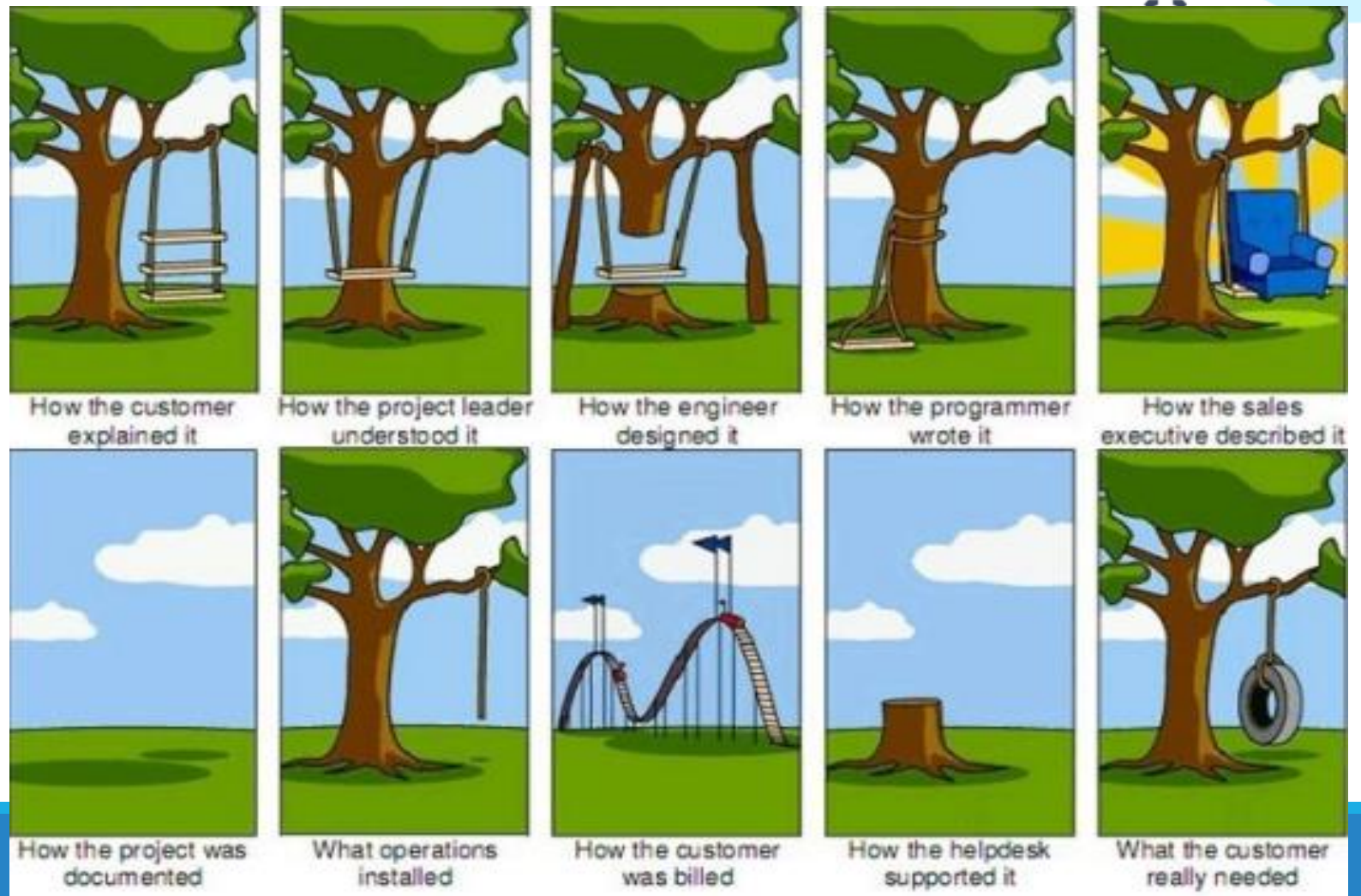


Thu thập dữ liệu

- Thu thập dữ liệu là một quá trình tổng hợp tất cả các thông tin từ nhiều nguồn khác nhau và lưu trữ chúng lại trong một hệ thống đã được thiết lập sẵn, sau đó cho phép một cá nhân hay tổ chức có thể trả lời câu hỏi có liên quan đến dữ liệu và đánh giá kết quả.
- Mục đích của việc thu thập dữ liệu đó là phục vụ cho việc phân tích, nghiên cứu, quản lý, kinh doanh hoặc đưa ra quyết định liên quan đến các lĩnh vực như khoa học, xã hội, kinh doanh...
- Các nguồn dữ liệu có thể bao gồm các hình ảnh, văn bản, video, âm thanh, dữ liệu từ mạng xã hội, website hay các nguồn dữ liệu khác. Mục tiêu là thu thập những bằng chứng cho phép phân tích, nghiên cứu để từ đó đưa ra các câu trả lời đáng tin cây với những câu hỏi được đặt ra.

Thu thập dữ liệu

Mục tiêu:
Đối tượng là ai?
Yêu cầu là gì?
Mục tiêu là gì?
Thời gian dữ liệu dùng cho phân tích?



Thu thập dữ liệu

Dữ liệu đã có sẵn chưa?

- **Nếu có:** xác định các cơ sở dữ liệu, trường và bảng đang chứa dữ liệu mà mình cần.
- **Nếu chưa:**
  - Xác định nguồn lấy.
  - Xác định cách lấy: lấy qua API hay cần setup để xuất dữ liệu thô? Lấy dạng streaming hay batching?
  - Xác định dạng dữ liệu: Dữ liệu có cấu trúc, dữ liệu bán cấu trúc hay dữ liệu phi cấu trúc

QUÁ TRÌNH THU THẬP DỮ LIỆU

(0.1): Xác định bài toán
(0.2): Xây dựng phương pháp thu thập dữ liệu (đặt câu hỏi….)
(0.3): Tiến hành thu thập dữ liệu (khảo sát ….)

XÁC ĐỊNH BÀI TOÁN & THU THẬP DỮ LIỆU

1. Giá trị hóa dữ liệu

2. Mã hóa các câu trả lời

3. Nhập dữ liệu vào máy tính

Chuyển hóa dữ liệu

4. Lưu trữ dữ liệu

5. Truy xuất dữ liệu

Lưu trữ

Tầm quan trọng của việc thu thập dữ liệu

Cung cấp thông tin

Dữ liệu được thu thập từ nhiều nguồn khác nhau từ đó cung cấp các thông tin hữu ích cho việc ra quyết định của người dùng. Đây có thể là dữ liệu về thị trường, khách hàng, kiến thức kinh tế, xã hội, văn hóa...

Xác định xu hướng và mối quan hệ

Việc thu thập dữ liệu giúp cho mọi người có thể xác định được mối quan hệ và xu hướng giữa các yếu tố với nhau từ đó giúp cho nhà nghiên cứu có thể hiểu rõ về mối tương quan của những yếu tố này.

Đưa ra dự đoán

Từ những dữ liệu thu thập được thông qua quá trình phân tích, nghiên cứu, các nhà phân tích có thể đưa ra dự đoán về các xu hướng của nhiều lĩnh vực trong tương lai từ đó giúp các nhà quản lý lập kế hoạch và đưa ra quyết định chính xác hơn.

Cải thiện hiệu quả

Thông qua quá trình thu thập dữ liệu, các nhà quản trị sẽ hiểu rõ hơn về các hoạt động, công việc của họ từ đó cải thiện hiệu suất làm việc đạt kết quả tối ưu hơn.

Các phương pháp thu thập dữ liệu

Các phương pháp thu thập dữ liệu

1. Phương pháp lắng nghe mạng xã hội
2. Phương pháp phỏng vấn bằng thư
3. Phương pháp phỏng vấn cá nhân trực tiếp
4. Phương pháp quan sát
5. Phương pháp thu thập dữ liệu điều tra thăm dò
6. Phương pháp thảo luận nhóm
7. Phương pháp thử nghiệm
8. Phương pháp phỏng vấn điện thoại
9. Phương pháp hỏi ý kiến chuyên gia
10.

Các phương pháp thu thập dữ liệu

VD: Một khảo sát về thời gian sử dụng facebook của sinh viên IUH được tiến hành online cho kết quả rằng thời gian sử dụng trung bình facebook là 3.2 giờ.

→ Đây là dữ liệu thu được từ quan sát, vì người được khảo sát không chịu sự tác động nào

VD: Trong một thử nghiệm y tế cộng đồng, người ta tiêm cho 200.745 trẻ em loại vaccine X, và tiêm cho 201.229 trẻ em một loại vaccine giả dược (vaccine giả dược này không gây ảnh hưởng gì đến sức khỏe)
→ Trong ví dụ này, người ta đã chia đối tượng cần nghiên cứu ra làm hai nhóm cho nên đây là dữ liệu thu được từ thử nghiệm

VD: Một khảo sát về thời gian sử dụng điện thoại của sinh viên IUH trong 1 tiết học

Các phương pháp thu thập dữ liệu

Ví dụ: Điều tra về sự ảnh hưởng của điều kiện sinh hoạt đến mức độ cận thị của sinh viên trường ĐH Công Nghiệp TP.HCM nhằm phục vụ cho chiến lược mở cửa hàng kính thuốc xung quanh trường ĐH Công Nghiệp TP.HCM.

- 1. **Giới tính của bạn là**: Nam/Nữ
- 2. **Bạn đang sống ở**:
  - a. Gia đình, nhà người thân
  - b. Ký túc xá
  - c. Nhà trọ
- 3. **Một ngày bạn giành bao nhiêu thời gian cho việc tự học?**
  - a. Dưới 3 giờ
  - b. Khoảng 3-5 giờ
  - c. Trên 5 giờ
- 4. **Một ngày bạn sử dụng máy vi tính bao lâu?**
  - a. Dưới 1 giờ
  - b. 1-3 giờ
  - c. 3-5 giờ
  - d. Trên 5 giờ
- 5. **Hiện nay mắt của bạn bao nhiêu độ?**

Các kỹ thuật thu thập dữ liệu bằng python

1. Phát sinh tự động thông tin (Crawler and auto-generated nepowt)
2. Web Scraping
3. API
4. RSS
5. Phương pháp thu thập dữ liệu cho AI tạo sinh

Các kỹ thuật thu thập dữ liệu bằng python

- Phát sinh tự động thông tin (Crawler and auto-generated nepowt)

Các kỹ thuật thu thập dữ liệu bằng python

Thư viện Beautifulsoup để rút trích dữ liệu từ website:
Beautifulsoup là một thư viện Python được sử dụng để phân tích cú pháp HTML và XML, giúp cho việc trích xuất thông tin từ các trang web trở nên dễ dàng hơn.

Thư viện này cung cấp các công cụ để điều hướng, tìm kiếm và trích xuất dữ liệu từ cấu trúc HTML hoặc XML. Nó có thể giúp bạn tìm kiếm các thẻ HTML, tìm kiếm các thuộc tính của các thẻ đó, truy cập nội dung bên trong các thẻ và thậm chí có thể giúp bạn tìm kiếm các đoạn văn bản cụ thể trong tài liệu.

Sử dụng Beautifulsoup, bạn có thể trích xuất dữ liệu từ các trang web và lưu trữ nó trong các định dạng khác nhau như CSV, Excel, JSON hoặc SQL database để phân tích dữ liệu hoặc xây dựng các ứng dụng khác.

Trích xuất nội dung văn bản từ một trang web HTML

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

```python
import requests
from bs4 import BeautifulSoup

# Tạo yêu cầu đến trang web
url = 'https://www.example.com'
response = requests.get(url)

# Phân tích cú pháp HTML và lấy nội dung
soup = BeautifulSoup(response.content, 'html.parser')
content = soup.get_text()

# In ra nội dung văn bản
print(content)
```

Tìm các thẻ HTML và truy xuất các thuộc tính và giá trị của chúng


```python
import requests
from bs4 import BeautifulSoup

# Tạo yêu cầu đến trang web
url = 'https://www.example.com'
response = requests.get(url)

# Phân tích cú pháp HTML và tìm kiếm các thẻ
soup = BeautifulSoup(response.content, 'html.parser')
links = soup.find_all('a')

# Truy xuất thuộc tính href của các thẻ a và in ra màn hình
for link in links:
    href = link.get('href')
    print(href)
```

Các bước lấy dữ liệu từ Web

Bước 1: Cài đặt module (hướng dẫn cho cmd Window)

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

- Lấy dữ liệu
- Tách dữ liệu crawl được từ web
- Phân tích dữ liệu
- Lấy dữ liệu chi tiết của từng bài

Tạo hàm để tái sử dụng trong python

Các bước lấy dữ liệu từ Web

Bước 1: Cài đặt module (hướng dẫn cho cmd Window)

- Để lấy dữ liệu từ web bằng python, đầu tiên cần cài đặt Requests:
  pip install requests (hoặc python –m pip install requests)
- Cài đặt Pilow:
  pip install Pillow (hoặc python –m pip install Pillow)
- **Lưu ý:** Nếu bạn đang dùng PIP cũ thì hãy update lên pip mới trước khi cài Pillow với cú pháp như sau nhé: Update PIP:

pip install -–upgrade pip (hoặc python –m pip install -–upgrade pip)

Các bước lấy dữ liệu từ Web

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

Lấy dữ liệu: Module Request dùng để gửi HTTP request, điều này cũng giống như thao tác bạn thường làm khi tìm kiếm thứ gì đó trên mạng: Vào trình duyệt, gõ vào thanh tìm kiếm "mcivietnam" và enter, bạn sẽ nhận được giao diện của trang web MCI hoặc một dạng dữ liệu khác trả về.

Để lấy được dữ liệu trả về chúng ta phải sử dụng một module để hỗ trợ là Request:

requests.method(url, params, data, json, headers, cookies, files, auth, timeout, allow_redirects, proxies, verify, stream, cert)

Có thể sửa lại theo cách này để đoạn mã như sau:

```
import requests
response = requests.get("https://tuoitre.vn/tin-moi-nhat.htm")
print(response) →Kết quả: <Response [200]>
```

Các bước lấy dữ liệu từ Web

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

Lấy dữ liệu: Module Request dùng để gửi HTTP request, điều này cũng giống như thao tác bạn thường làm khi tìm kiếm thứ gì đó trên mạng: Vào trình duyệt, gõ vào thanh tìm kiếm "mcivietnam" và enter, bạn sẽ nhận được giao diện của trang web MCI hoặc một dạng dữ liệu khác trả về.

Để lấy được dữ liệu trả về chúng ta phải sử dụng một module để hỗ trợ là Request:

requests.method(url, params, data, json, headers, cookies, files, auth, timeout, allow_redirects, proxies, verify, stream, cert)

Có thể sửa lại theo cách này để đoạn mã như sau:

```
import requests
response = requests.get("https://tuoitre.vn/tin-moi-nhat.htm")
print(response)
```

Các bước lấy dữ liệu từ Web

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

Tách dữ liệu crawl được từ web

Dùng module **BeautifulSoup4 để** tách các dữ liệu ra thành dạng cây để thuận tiện hơn cho quá trình truy xuất dữ liệu.

Cú pháp cài đặt module:

pip install beautifulsoup4 (hoặc python –m pip install beautifulsoup4

BeautifulSoup4 sẽ giúp bạn phân tích dữ liệu HTML, XML thành dạng cây như sau:

```
import requests
from bs4 import BeautifulSoup
response = requests.get("https://tuoitre.vn/tin-moi-nhat.htm")
soup = BeautifulSoup(response.content, "html.parser")
print(soup)
```

Các bước lấy dữ liệu từ Web

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

Phân tích dữ liệu:

Ví dụ để lấy chi tiết về một bài viết cụ thể, bạn cần biết liên kết để truy cập đến bài viết đó. Bạn có thể tìm kiếm thông tin đó bằng cách ấn F12 và tìm xem link bài báo ở đâu. Ví dụ chúng ta tìm thấy link bài báo ở trong thẻ <a>, nằm trong thẻ h3 và có class là “title-news”. Vậy chúng ra cần phải lọc tất cả thẻ h3 có class “title-news” và lấy thẻ a trong đó, đoạn mã cần thiết được xây dựng là:

```
titles = soup.findAll('h3', class_='title-news')
print(titles)
```

Kết quả sẽ trả về là một mảng các thẻ h3 là tiêu đề của các bài báo. Tiếp theo chúng ta cần lấy link của các bài viết đó:

```
links = [link.find('a').attrs["href"] for link in titles]
print(links)
```

Các bước lấy dữ liệu từ Web

Bước 2: Crawl dữ liệu từ danh sách tin tức mới nhất

Phân tích dữ liệu: Lấy dữ liệu chi tiết của từng bài

Code để truy cập vào từng bài viết, lấy 1 ảnh đại diện và 1 đoạn trích trong bài viết đó:

```
for link in links:
    news = requests.get("https://tuoitre.vn" + link)
    soup = BeautifulSoup(news.content, "html.parser")
    title = soup.find("h1", class_="article-title").text
    abstract = soup.find("h2", class_="sapo").text
    body = soup.find("div", id="main-detail-body")
    content = body.findChildren("p", recursive=False)[0].text +      body.findChildren("p", recursive=False)[1].text
    image = body.find("img").attrs["src"]
    print("Tiêu đề: " + title)
    print("Mô tả: " + abstract)
    print("Nội dung: " + content)
    print("Ảnh minh họa: " + image)
    print("_____")
```

Tạo hàm để tái sử dụng trong python

INDUSTRIAL UNIVERSITY OF

```
def crawNewsData(baseUrl, url):
    response = requests.get(url)
    soup = BeautifulSoup(response.content,
        "html.parser")
    titles = soup.findAll('h3', class_='title-news')
    links = [link.find('a').attrs["href"] for link in titles]
    data = []
    for link in links:
        news = requests.get(baseUrl + link)
        soup = BeautifulSoup(news.content,
        "html.parser")
        title = soup.find("h1", class_="article-
            title").text
        abstract = soup.find("h2",
            class_="sapo").text
        body = soup.find("div", id="main-detail-body")
```

```
Content = “”
try:
    content = body.findChildren("p",
        recursive=False)[0].text +
        body.findChildren("p",
        recursive=False)[1].text
except:
    content = ""
image = body.find("img").attrs["src"]
data.append({
        "title": title,
        "abstract": abstract,
        "content": content,
        "image": image,        })
return data
```

Các kỹ thuật thu thập dữ liệu bằng python

Ví dụ 1:

```
import requests
from bs4 import BeautifulSoup
# Gửi yêu cầu HTTP đến trang web
url = 'https://www.example.com‘
response = requests.get(url)
# Kiểm tra trạng thái phản hồi
if response.status_code == 200:
# Phân tích cú pháp HTML của trang web
    soup = BeautifulSoup(response.content,
'html.parser')
# Lấy tiêu đề của trang web
    title = soup.title.string
    print('Tiêu đề:', title)
# Lấy toàn bộ nội dung của thẻ div có
class là "content"
    content_div = soup.find('div',
    class_='content')
    content = content_div.get_text()
    print('Nội dung:', content)
else:
    print('Không thể kết nối đến trang web')
```

Các kỹ thuật thu thập dữ liệu bằng python

```
pip3 install newspaper3k
```

Và để thu thập dữ liệu của một url bất kỳ, hãy dùng 5 dòng code sau đây:

```
from newspaper import Article
url = 'https://vnexpress.net/12-000-nguoi-do-ve-cua-lo-4092705.html'
article = Article(url)
article.download()
article.parse()
# Xong rồi đấy, giờ lấy data thôi
print(article.title)
> 12.000 người đổ về Cửa Lò - VnExpress
...
```

Các kỹ thuật thu thập dữ liệu bằng python

Web Scraping

- Web scraping là quá trình tự động trích xuất dữ liệu từ các trang web. Điều này thường được thực hiện bằng cách phân tích cú pháp mã HTML của trang web để tìm và trích xuất thông tin cần thiết.

Công cụ và Thư viện

- **Requests**: Thư viện này cho phép bạn gửi các yêu cầu HTTP đến một trang web và nhận về mã HTML của trang đó
- **BeautifulSoup**: Một thư viện giúp phân tích cú pháp và trích xuất dữ liệu từ HTML/XML một cách dễ dàng
- **Scrapy**: Một framework mạnh mẽ cho phép thu thập dữ liệu từ web một cách nhanh chóng và tự động

Các kỹ thuật thu thập dữ liệu bằng python

Ví dụ: Giả sử bạn muốn thu thập tiêu đề và liên kết của các bài viết mới nhất từ trang tin tức Tuổi Trẻ

```
import requests
from bs4 import BeautifulSoup

url = "https://tuoitre.vn/tin-moi-nhat.htm"
response = requests.get(url)
soup = BeautifulSoup(response.content, "html.parser")

for article in soup.find_all("article"):
    title = article.find("h3").text
    link = article.find("a")["href"]
    print(title, link)
```

Các kỹ thuật thu thập dữ liệu bằng python

API

- Nhiều trang tin tức cung cấp API cho phép truy cập dữ liệu một cách chính thức và hiệu quả. Sử dụng API giúp tránh được các vấn đề về pháp lý liên quan đến web scraping và thường cung cấp dữ liệu đã được cấu trúc sẵn.

Công cụ và Thư viện

- **Requests**: Để gửi yêu cầu đến API và nhận dữ liệu trả về
- **Python API Tutorials**: Hướng dẫn sử dụng các API phổ biến với Python

Các kỹ thuật thu thập dữ liệu bằng python

Ví dụ: Giả sử bạn muốn sử dụng API của trang tin tức để lấy dữ liệu:

```python
import requests

api_url = "https://newsapi.org/v2/top-headlines?country=us&apiKey=YOUR_API
response = requests.get(api_url)
articles = response.json()["articles"]

for article in articles:
    print(article["title"], article["url"])
```

Các kỹ thuật thu thập dữ liệu bằng python

RSS

- **RSS (Rich Site Summary) l**à một tập tin thường được biểu diễn dưới dạng XML có nhiệm vụ tóm tắt thông tin giúp cho người đọc dễ dàng tìm kiếm cũng như cập nhật nội dung. Nhiều trang tin tức cung cấp feed RSS, cho phép bạn dễ dàng thu thập dữ liệu mà không cần phải phân tích cú pháp HTML.

Công cụ và Thư viện

- **Feedparser**: Thư viện Python hỗ trợ đọc dữ liệu từ feed RSS

Các kỹ thuật thu thập dữ liệu bằng python

Ví dụ

Thu thập tiêu đề và liên kết từ feed RSS của trang Tuổi Trẻ:

```
import feedparser

NewsFeed = feedparser.parse("https://tuoitre.vn/rss/tin-moi-nhat.rss")

for entry in NewsFeed.entries:
    print(entry.title, entry.link)
```

Các kỹ thuật thu thập dữ liệu bằng python

- Một số đường link RSS của một số trong báo điện tử phía dưới:
  - RSS ictnews (vietnamnet.vn)
  - Document (24h.com.vn)
  - RSS – VnExpress
  - ……..

Các kỹ thuật thu thập dữ liệu bằng python

Ý tưởng của ứng dụng

- Ứng dụng sẽ có một số chức năng chính sau:
- Thu thập thông tin từ đường link: https://ictnews.vietnamnet.vn/rss/thong-tin-truyen-thong/toan-van-phat-bieu.rss
- Lưu trữ thông tin thu thập được vào một file có tên **thong-tin-truyen-thong.xml**
- Phân tích file thong-tin-truyen-thong.xml để tách ra các thông tin cần thiết bao gồm:
  Guid: Thông tin định dạng các item trong nội dung xml
  link: Đường dẫn chi tiết của bài báo
  PubDate: Ngày bài báo được publish
  title: Tiêu đề của bài viết
  description: Mô tả của bài viết
  media:content: File ảnh thumbnail của bài viết

Lưu trữ thông tin vừa phân tích được ra một file dạng csv làm đầu vào cho quá trình làm sạch dữ liệu hoặc đầu vào cho các ứng dụng khác.

Các kỹ thuật thu thập dữ liệu bằng python

Viết mã nguồn: Tạo file fetch-rss.py

Import các thư viện cần dùng

```
import requests import csv import xml.etree.ElementTree as et
```

Viết các hàm thu thập, phân tích và lưu trữ

Hàm Thu thập dữ liệu từ link RSS

```
# Truy cập nội dung link rss và tải nội dung xuống file thong-tin-truyen-thong.xml
def loadRSS(url, fileName):
        response = requests.get(url) with open(fileName, 'wb')
        as f: f.write(response.content)
```

Các kỹ thuật thu thập dữ liệu bằng python

Hàm đẩy dữ liệu thu thập được ra file CSV

```
def savetoCSV(newsitems, filename):
        fields = [ 'guid', 'link', 'pubDate', 'category', 'title',
        'description', 'media', ]
# Viết dữ liệu ra file csv with
open(filename, 'w', encoding='utf-8') as csvfile: writer =
csv.DictWriter(csvfile, fieldnames=fields)
writer.writerows(newsitems)
```

Các kỹ thuật thu thập dữ liệu bằng python

Hàm phân tích dữ liệu thu thập được

```
def parseRSS(xmlFileName):
        rssTree = et.parse(xmlFileName)
        root = rssTree.getroot()
        newsitems = []
        for item in root.findall('./channel/item'):
                news = {}
                for child in item:
                        if child.tag == '{http://search.yahoo.com/mrss/}content':
                                news['media'] = child.attrib['url']
                        else:
                                news[child.tag] = child.text
newsitems.append(news)
return newsitems
```

Các kỹ thuật thu thập dữ liệu bằng python

Hàm main để chạy ứng dụng

```
def main():
        loadRSS( "https://ictnews.vietnamnet.vn/rss/thong-tin-truyen-
        thong/toan-van-phat-bieu.rss", "thong-tin-truyen-thong.xml")
        newsitems = parseRSS("thong-tin-truyen-thong.xml")
        savetoCSV(newsitems, 'thong-tin-truyen-thong.csv')
if __name__ == "__main__": main()
```

Các kỹ thuật thu thập dữ liệu bằng python

Phương pháp thu thập dữ liệu từ AI tạo sinh:

- **Dữ liệu AI tạo sinh (Generative AI data)** đề cập đến kho thông tin khổng lồ được sử dụng để đào tạo các mô hình AI tạo sinh. Dữ liệu này có thể bao gồm văn bản, hình ảnh, âm thanh hoặc video.
- Các mô hình tạo sinh tìm hiểu các mẫu đặc trưng từ dữ liệu, từ đó tạo ra nội dung mới phù hợp với độ phức tạp và cấu trúc của dữ liệu đầu vào. Một số tác vụ này bao gồm tạo hình ảnh, tạo video, xử lý ngôn ngữ tự nhiên, v.v.

Các kỹ thuật thu thập dữ liệu bằng python

Phương pháp thu thập dữ liệu từ AI tạo sinh:

1. Crowdsourcing

- Crowdsourcing liên quan đến việc lấy dữ liệu từ một nhóm người quy mô lớn, thường là thông qua internet. Phương pháp này có thể cung cấp dữ liệu đa dạng và có chất lượng cao.

- Ví dụ, nếu đào tạo mô hình AI hội thoại, bạn có thể thu thập dữ liệu hội thoại từ cộng đồng người dùng trên khắp thế giới, giúp mô hình có thể hiểu và tương tác bằng nhiều ngôn ngữ và phong cách khác nhau.
- Tuy nhiên, thu thập dữ liệu từ cộng đồng đòi hỏi phải phát triển một nền tảng trực tuyến giúp doanh nghiệp thuê và quản lý nhóm thu thập dữ liệu.

Các kỹ thuật thu thập dữ liệu bằng python

2. Thu thập dữ liệu từ web

- Thu thập dữ liệu từ web liên quan đến việc trích xuất dữ liệu tự động từ internet.
- Ví dụ: một mô hình AI tạo sinh tập trung vào tính năng viết tin tức có thể thu thập các bài viết từ nhiều trang web tin tức khác nhau.

Các kỹ thuật thu thập dữ liệu bằng python

2. Thu thập dữ liệu từ web

Một số công cụ thu thập dữ liệu từ web bạn có thể tham khảo bao gồm:

- **Scrapy:** Khung thu thập dữ liệu web nguồn mở mạnh mẽ và được sử dụng rộng rãi trong Python. Ưu điểm của Scrapy là tính linh hoạt và khả năng mở rộng.
- **Selenium:** Không chỉ là một trình thu thập thông tin, Selenium còn là một khung thử nghiệm có thể được sử dụng để quét web bằng cách tự động hóa các trình duyệt. Selenium được sử dụng chủ yếu trong việc xử lý nội dung động được hiển thị bằng JavaScript.

Các kỹ thuật thu thập dữ liệu bằng python

2. Thu thập dữ liệu từ web

- **Beautiful Soup:** Đây là thư viện Python giúp thu thập dữ liệu từ các tệp HTML và XML. Beautiful Soup cung cấp các phương thức đơn giản bằng Python để điều hướng, tìm kiếm và sửa đổi cây phân tích cú pháp (parse tree).
- **Apache Nutch:** Đây là trình thu thập dữ liệu web nguồn mở được viết bằng Java. Apache Nutch có khả năng mở rộng cao, phù hợp để xây dựng và duy trì các kho lưu trữ web quy mô lớn.

Các kỹ thuật thu thập dữ liệu bằng python

2. Thu thập dữ liệu từ web

- **Crawler4j:** Crawler4j cũng là một trình thu thập dữ liệu web nguồn mở dựa trên Java cung cấp giao diện đơn giản để thu thập dữ liệu trên web và truy xuất các trang web.
- **Heritrix:** Trình thu thập dữ liệu web linh hoạt, có thể mở rộng và được thiết kế để lưu trữ web. Heritrix được sử dụng bởi nhiều tổ chức khác nhau cho mục đích lưu trữ.
- **ParseHub:** Công cụ trích xuất dữ liệu trực quan cho phép người dùng trích xuất dữ liệu từ các trang web mà không cần mã hóa.
- **Octoparse:** Tương tự như ParseHub, Octoparse là một công cụ quét web trực quan cho phép người dùng trích xuất dữ liệu mà không cần kỹ năng lập trình. Octoparse được biết đến vì tính dễ sử dụng và linh hoạt.

Các kỹ thuật thu thập dữ liệu bằng python

3. Tạo dữ liệu tổng hợp

- Với sự phát triển của các mô hình AI tạo sinh, việc tạo dữ liệu tổng hợp ngày càng thu hút sự chú ý của cộng đồng công nghệ. Theo cách tiếp cận này, một mô hình AI tạo sinh sẽ tạo ra dữ liệu tổng hợp để huấn luyện một mô hình khác.
- Ví dụ: có thể sử dụng ChatGPT hoặc Llama để sinh dữ liệu giả hội thoại nhằm huấn luyện kĩ năng hội thoại cho các mô hình ngôn ngữ nhỏ hơn; hoặc sử dụng Dall-E để sinh dữ liệu hình ảnh giúp tăng cường dữ liệu cho các mô hình phân loại, phát hiện thực thể, v.v.

Các kỹ thuật thu thập dữ liệu bằng python

4. Bộ dữ liệu mở

Nhiều tổ chức và cá nhân cung cấp công khai các bộ dữ liệu mở phục vụ mục đích nghiên cứu. Những bộ dữ liệu này hoàn toàn có thể được sử dụng để đào tạo mô hình AI tạo sinh.

Bạn có thể tham khảo một số nguồn dữ liệu như:

- Wikipedia đối với dữ liệu văn bản
- ImageNet đối với dữ liệu hình ảnh
- LibriSpeech đối với dữ liệu âm thanh
- Sách
- Báo chí thời sự
- Tạp chí khoa học

Các kỹ thuật thu thập dữ liệu bằng python

5. Tăng cường dữ liệu

- Dữ liệu hiện có có thể được sửa đổi hoặc kết hợp để tạo dữ liệu mới. Cách tiếp cận này được gọi là tăng cường dữ liệu.
- **Ví dụ:** hình ảnh có thể được xoay, thu nhỏ hoặc biến đổi theo nhiều cách khác, trong khi dữ liệu văn bản có thể được tổng hợp bằng cách thay thế, xóa hoặc sắp xếp lại các từ.

Các kỹ thuật thu thập dữ liệu bằng python

6. Dữ liệu người dùng

- Dữ liệu độc quyền, chẳng hạn như nhật ký cuộc gọi (call log) của khách hàng, cũng có thể được sử dụng để đào tạo các mô hình ngôn ngữ lớn, đặc biệt cho các nhiệm vụ liên quan đến dịch vụ khách hàng, chẳng hạn như tạo phản hồi tự động, phân tích cảm xúc hoặc nhận dạng ý định.

Các kỹ thuật thu thập dữ liệu bằng python

6. Dữ liệu người dùng

Tuy nhiên, một số yếu tố quan trọng phải được xem xét khi sử dụng dữ liệu này là:

- **Phiên âm:** Nhật ký cuộc gọi, thường là âm thanh, cần phiên âm thành văn bản để đào tạo các mẫu dựa trên văn bản như GPT-3 hoặc GPT-4.
- **Quyền riêng tư:** Đảm bảo nhật ký cuộc gọi được ẩn danh và tuân thủ các quy định về quyền riêng tư, cũng như có sự đồng thuận rõ ràng của khách hàng.
- **Thiên kiến:** Nhật ký cuộc gọi có thể chứa các ý kiến chủ quan, làm ảnh hưởng đến hiệu suất của mô hình trên các loại cuộc gọi hoặc thời gian khác nhau.
- **Làm sạch dữ liệu:** Dữ liệu cuộc gọi cần được làm sạch để loại bỏ nhiễu như cuộc trò chuyện không liên quan, tiếng ồn xung quanh hoặc lỗi phiên âm.

Kiểm tra dữ liệu

- **Kiểm tra dữ liệu**: Kiểm tra sơ bộ về tệp dữ liệu, xem số lượng, kiểu dữ liệu, thuộc tính dữ liệu và các đặc điểm khác. Quá trình này sẽ giúp các nhà phân tích dữ liệu định hình được các phương án xử lý dữ liệu tiếp theo.

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

Bài tập

Thu thập dữ liệu từ những nguồn sau đây:

URL1 : https://vingroup.net/en/investor-relations/info-disclosure/information-disclosure/1

URL2 : https://ir.vincom.com.vn/en/information-disclosure/information-disclosure-en/

URL3 : https://ir.vinhomes.vn/en/category/announcement/

URL4 : Top 10 largest: https://vn.tradingview.com/markets/stocks-vietnam/market-movers-large-cap/

URL5 : Các công bố của VRE: https://www.hsx.vn/Modules/Listed/Web/SymbolView/683

URL6 : Các công bố của VIC: https://www.hsx.vn/Modules/Listed/Web/SymbolView/100

URL7 : Các công bố của VHM: https://www.hsx.vn/Modules/Listed/Web/SymbolView/719

Các thư viên có thể cần sử dụng: `xlwings` , `requests` , `BeautifulSoup` , `pandas` , `json`

Phương pháp điều tra chọn mẫu dữ liệu

Khái niệm

Tổng thể chung là tổng thể bao gồm toàn bộ các đơn vị thuộc đối tượng điều tra.

Tổng thể mẫu là tổng thể bao gồm một số đơn vị nhất định được chọn ra từ tổng thể chung để điều tra thực tế.

Lấy mẫu có thay thế (Samping with replacement) hay chọn hoàn lại

- Khi một ví dụ (bản ghi) được lấy mẫu, nó không bị loại khỏi tập dữ liệu ban đầu (có thể được chọn nhiều hơn một lần). Như vậy số đơn vị của tổng thể chung là không thay đổi trong suốt quá trình lấy mẫu.

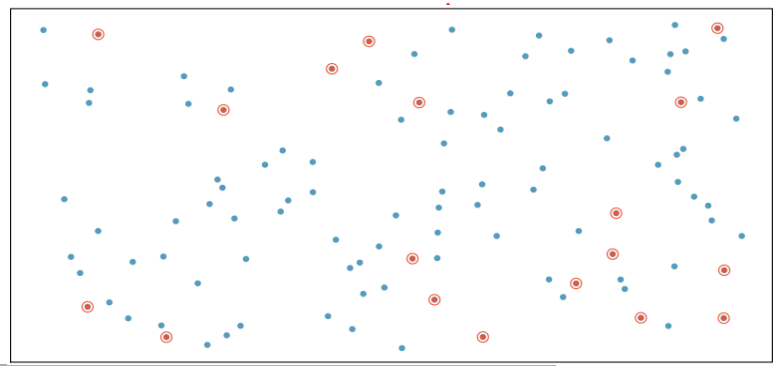
Lấy mẫu không thay thế (Sampling without replacement) hay chọn không hoàn lại

- Khi một ví dụ (bản ghi) được lấy mẫu, nó sẽ được loại khỏi tập dữ liệu ban đầu (sẽ không thể được chọn thêm một lần nào nữa)

Chọn mẫu xác xuất đều là đảm bảo mỗi đơn vị của hiện tượng đều có cơ hội được chọn và mẫu như nhau.

Chọn mẫu xác xuất không đều nghĩa là không cần đảm bảo khả năng được chọn vào mẫu của các đơn vị phải bằng nhau. Việc chọn mẫu không đều có khó khan và phức tạp.

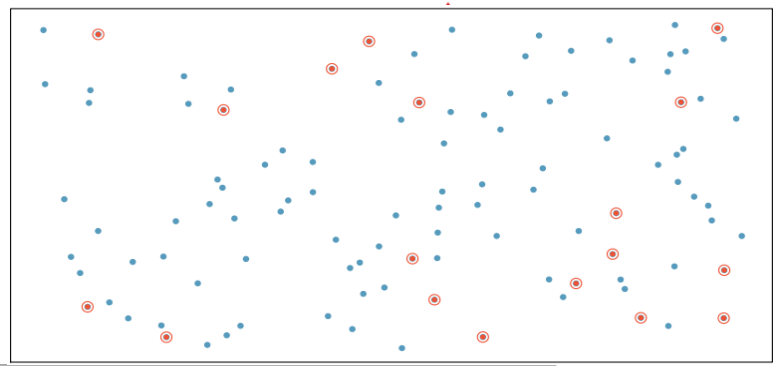
Các kỹ thuật chọn mẫu



Mục tiêu của các kỹ thuật chọn mẫu là bảo đảm cho mẫu được chọn thực sự phản ánh trung thực, đại diện cho toàn bộ tổng thể.
Các nhóm kĩ thuật chọn mẫu

1. Lấy mẫu ngẫu nhiên đơn giản
2. Lấy mẫu có hệ thống
3. Lấy mẫu tiện lợi
4. Lấy mẫu phân tầng
5. Lấy mẫu theo cụm
6. Lấy mẫu nhiều giai đoạn

Các kỹ thuật chọn mẫu



Lấy mẫu ngẫu nhiên đơn giản (simple random sample): là cách lấy mẫu mà mỗi giá trị dữ liệu được lấy từ quần thể theo *cùng một cách* và *cơ hội được chọn* của mỗi giá trị dữ liệu là như nhau.

- **Ví dụ 1:** Điều tra chi tiêu của người dân sống ở Tp Hồ Chí Minh → (ngẫu nhiên)

Lấy mẫu không ngẫu nhiên (lấy mẫu phi xác suất).

- **Ví dụ 2:** Điều tra chi tiêu của những người có thu nhập cao sống ở Tp Hồ Chí Minh → (phi ngẫu nhiên)

Các kỹ thuật chọn mẫu ngẫu nhiên và phi xác suất

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

Kỹ thuật chọn mẫu xác suất

- Lấy mẫu ngẫu nhiên đơn giản
- Lấy mẫu hệ thống
- Lấy mẫu cả khối/cụm
- Lấy mẫu phân tầng

Kỹ thuật chọn mẫu phi xác suất

- Lấy mẫu thuận tiện
- Lấy mẫu định mức
- Lấy mẫu phán đoán

Kỹ thuật chọn mẫu ngẫu nhiên đơn giản

Là loại mẫu được chọn trực tiếp và ngẫu nhiên từ tổng thể.

- **Tổng thể nhỏ**: Mẫu được chọn bằng cách bốc thăm, quay số,…

  Ví dụ: Lớp có 50 sinh viên, chọn ngẫu nhiên 10 bạn trong lớp bằng cách bốc thăm

- **Tổng thể lớn**: Mẫu được chọn bằng hàm ngẫu nhiên (random) trong các công cụ phần mềm.

→ Phương pháp này có thể cho 1 kết quả tốt và đảm bảo tính ngẫu nhiên.

Kỹ thuật Bootstrapping

Bootstrapping là một kỹ thuật lấy mẫu ngẫu nhiên mạnh mẽ

Hữu ích khi kích thước mẫu đang làm việc nhỏ.

Những loại kỹ thuật này không giả định gì về việc phân phối dữ liệu nghiên cứu.

Ví dụ: Cho mẫu nghiên cứu: 1, 2, 4, 4, 10.

- Sử dụng Bootstrapping để tạo ra 10 mẫu dùng nghiên cứu

2, 1, 10, 4, 2

4, 10, 10, 2, 4

1, 4, 1, 4, 4

4, 1, 1, 4, 10

4, 4, 1, 4, 2

4, 10, 10, 10, 4

2, 4, 4, 2, 1

2, 4, 1, 10, 4

1, 10, 2, 10, 10

4, 1, 10, 1, 10

Kỹ thuật chọn mẫu hệ thống (máy móc)

Lấy mẫu có hệ thống (Systematic Sampling): là lấy mẫu bằng cách chọn một điểm bắt đầu và điểm kết thúc, sau đó lần lượt chọn phần tử thứ k từ quần thể. Mỗi đơn vị được chọn vào mẫu căn cứ vào từng khoảng cách nhất định (khoảng thời gian,không gian,thứ tự bằng nhau).

Phương pháp:

- Đánh số thứ tự cho danh sách tổng thể để chọn mẫu.
  - Tổng số lượng tổng thể N
- Xác định cỡ mẫu muốn lấy.
  - Số lượng n
- Chia danh sách thành k nhóm
  - k=N/n, k gọi là khoảng cách chọn mẫu
  - **Lưu ý**: k là số nguyên được lấy làm tròn

Kỹ thuật chọn mẫu hệ thống (1)

Nếu N chia hết cho n (k nguyên): Chọn mẫu hệ thống theo đường thẳng: Chia tổng thể ra n nhóm, trong nhóm đầu tiên lấy ra ngẫu nhiên 1 phần tử, các phần tử tiếp theo được lấy cách phần tử này 1 khoảng là k, 2k, 3k,…

Ví dụ: Giả sử: ngân hàng ABC có danh sách khách hàng vay vốn có độ tuổi từ 25 đến 35 được đánh số từ 1 → 60. Ngân hàng chỉ cho phép nhân viên phân tích lấy ngẫu nhiên thông tin của 10 khách hàng.

Hướng dẫn:

- Chọn 10 số từ [1,60] số nguyên dương đầu tiên theo phương pháp chọn mẫu hệ thống.
- N=60, n=10, k=N/n=6
  - **VD 1**: Nếu phần tử được chọn đầu tiên là 4 thì ta được mẫu là: 4, 10, 16, 22, 28, 34, 40, 46, 52, 58
  - **VD 2**: Nếu phần tử được chọn đầu tiên là 6 thì ta được mẫu là: 6, 12, 18, 24, 30, 36, 42, 48, 54, 60

Kỹ thuật chọn mẫu hệ thống (2)

Nếu N không chia hết cho n (k thập phân): Ta dùng kỹ thuật chọn mẫu hệ thống quay vòng được mô tả như sau

- Chọn ngẫu nhiên 1 phần tử bất kì trong danh sách từ 1 đến N. Các phần tử tiếp theo được lấy cách phần tử này 1 khoảng là k, 2k, 3k…

Ví dụ: Giả sử: ngân hàng ABC có danh sách khách hang vay vốn có độ tuổi từ 40 đến 50 được đánh số từ 1 → 56. Ngân hàng chỉ cho phép nhân viên phân tích lấy ngẫu nhiên thông tin của 10 khách hàng

Hướng dẫn

- $N=56 \quad n=10$
- $k=N/n=5,6$, chọn $k=6$
- **VD 1**: Nếu phần tử được chọn đầu tiên là 6 thì ta được mẫu là: 6, 12, 18, 24, 30, 36, 42, 48, 54, 4
- **VD 2**: Nếu phần tử được chọn đầu tiên là 13 thì ta được mẫu là: 13, 19, 25, 31, 37, 43, 49, 55, 5, 11

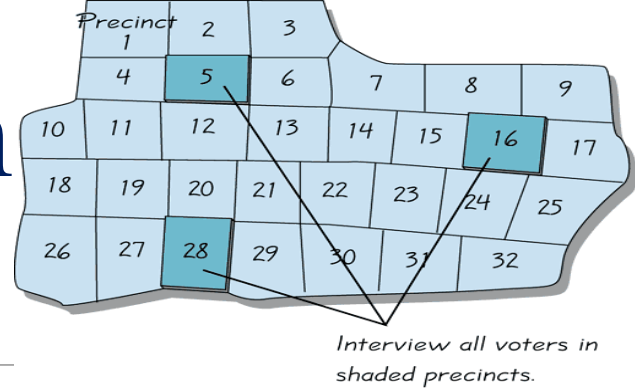
= 54+6-56

Bài tập

Thực hiện nghiên cứu về tình trạng hoạt động của doanh nghiệp trong quý 4 năm 2021, tiến hành chọn ra 293 doanh nghiệp trong địa phương, biết rằng địa phương có 3000 doanh nghiệp, theo phương pháp chọn mẫu ngẫu nhiên hệ thống. Hãy ghi ghi ra mã số của 5 doanh nghiệp đầu tiên và 5 doanh nghiệp cuối cùng trong danh sách doanh nghiệp được chọn, biết rằng mã số hạt giống là 6.

- *N=3000 n=293*
- *k=N/n=3000/293 =10.23 , chọn k=10*
- Nếu phần tử được chọn đầu tiên là 6 thì ta được:
- Mã số của 5 doanh nghiệp đầu tiên là : 6, 16, 26, 36, 46
- Mã số của 5 doanh nghiệp cuối cùng là : 286, 276, 266, 256,246

Kỹ thuật chọn mẫu khối /cụm và chọn mẫu nhiều giai đoạn



Lấy mẫu theo cụm (Cluster Sampling): chia quần thể thành nhiều cụm nhỏ, chọn ngẫu nhiên một số cụm và lấy tất cả các dữ liệu của các cụm được chọn.

Chọn mẫu khối

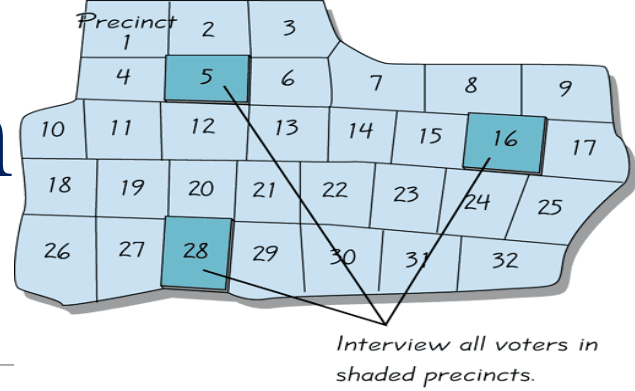
- Ví dụ 1: Quận Thủ Đức có khoảng 800 khu phố, điều tra mức sống của dân cư ở đây, ta có thể chọn ra ngẫu nhiên 10 khu phố, sau đó khảo sát toàn bộ hộ dân của 10 khu phố này

Lấy mẫu nhiều giai đoạn (Multisatage Sampling): là phương pháp lấy mẫu bằng cách kết hợp nhiều phương pháp lấy mẫu đơn giản với nhau

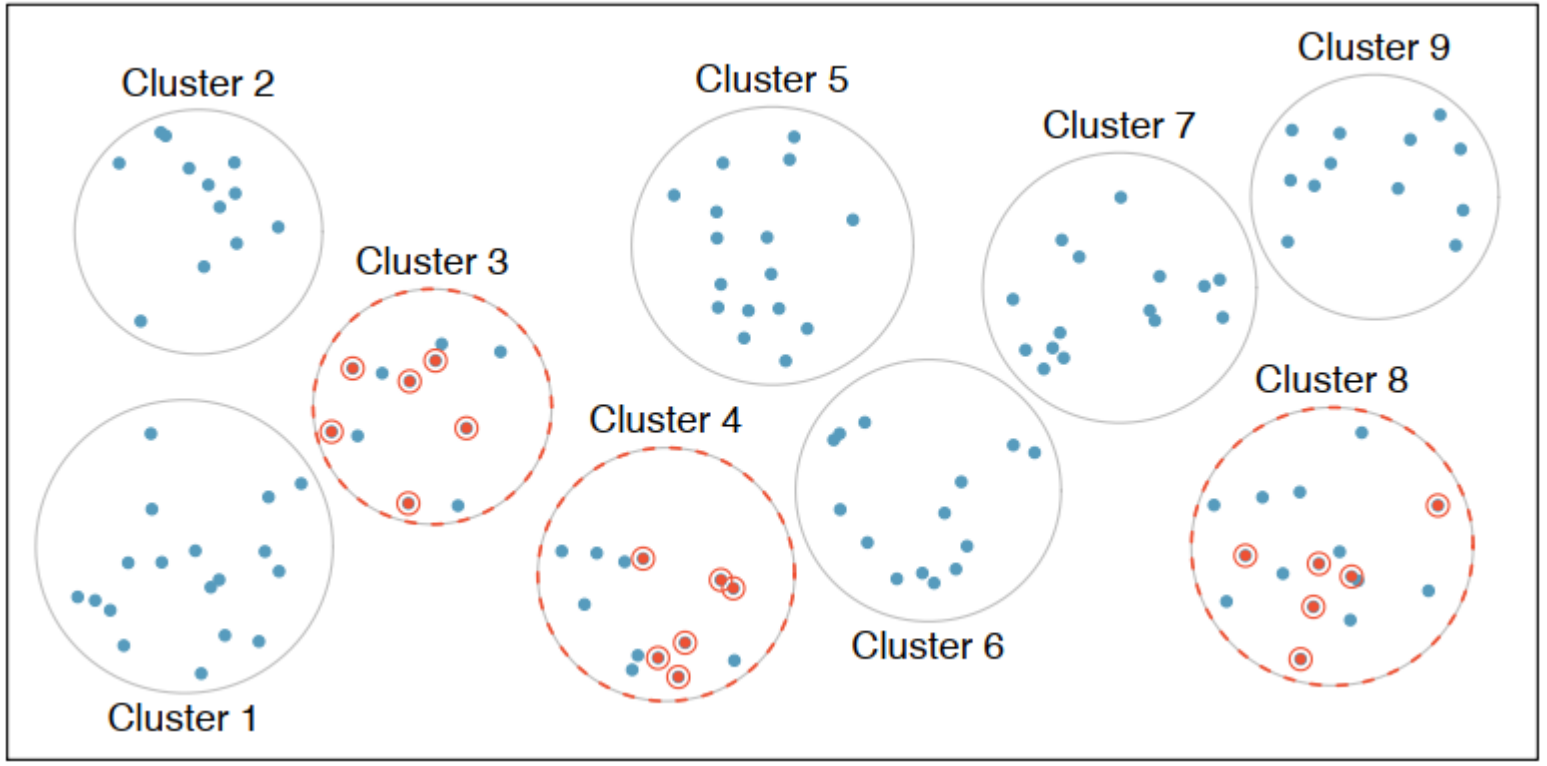
- Ví dụ 2: Chọn ra ngẫu nhiên 10 khu phố, trong mỗi khu phố chọn ra khoảng 10 hộ gia đình

Chú ý : Kỹ thuật này áp dụng khi ta không có sẵn một danh sách quan sát để chọn ra mẫu

Kỹ thuật chọn mẫu khối /cụm và chọn mẫu nhiều giai đoạn

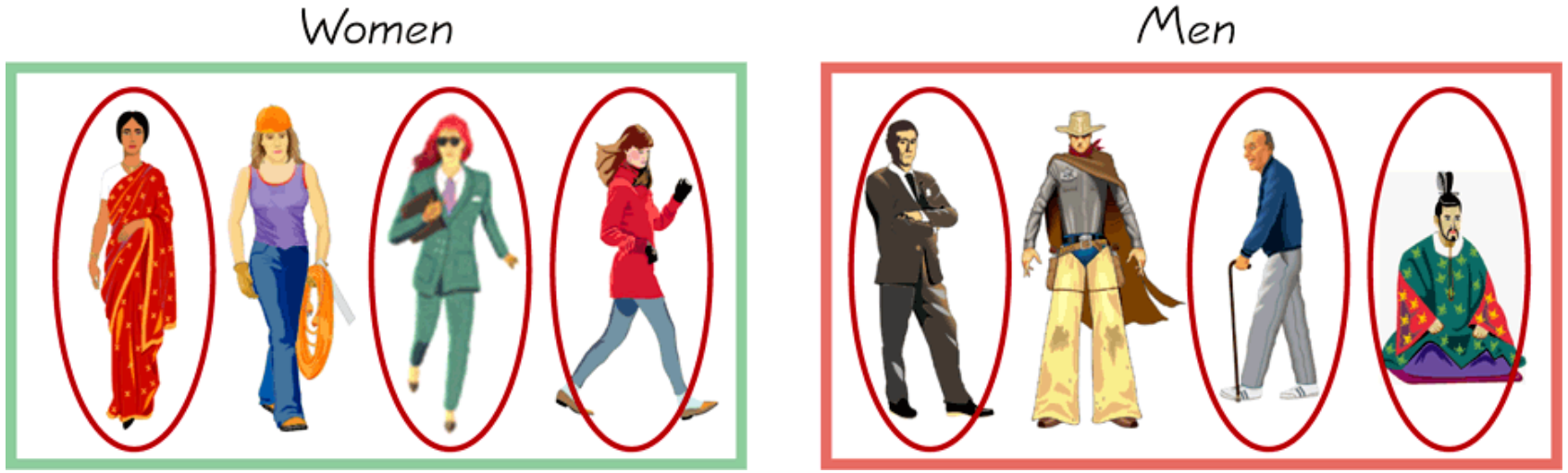


Ví dụ: Lấy mẫu nhiều giai đoạn (Multisatage Sampling)

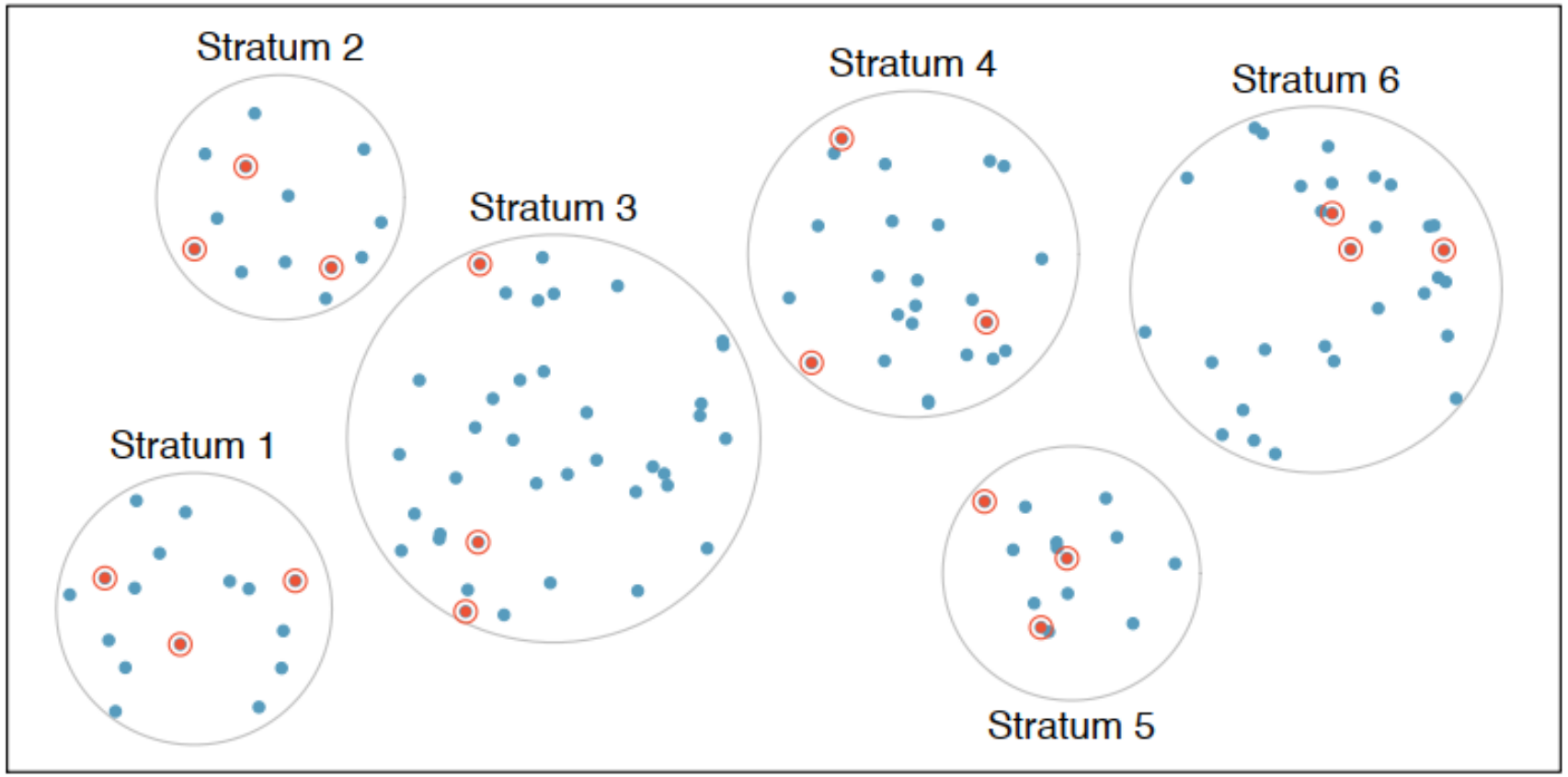


Lấy mẫu phân tầng (Stratified sampling)

Lấy mẫu phân tầng (Straified sampling): chia quần thể thành nhiều nhóm nhỏ, mỗi nhóm có cùng đặc tính, sau đó lấy mẫu bằng cách chọn ngẫu nhiên từ các nhóm nhỏ đó.



Lấy mẫu phân tầng (Stratified sampling)



Kỹ thuật chọn mẫu phân tầng

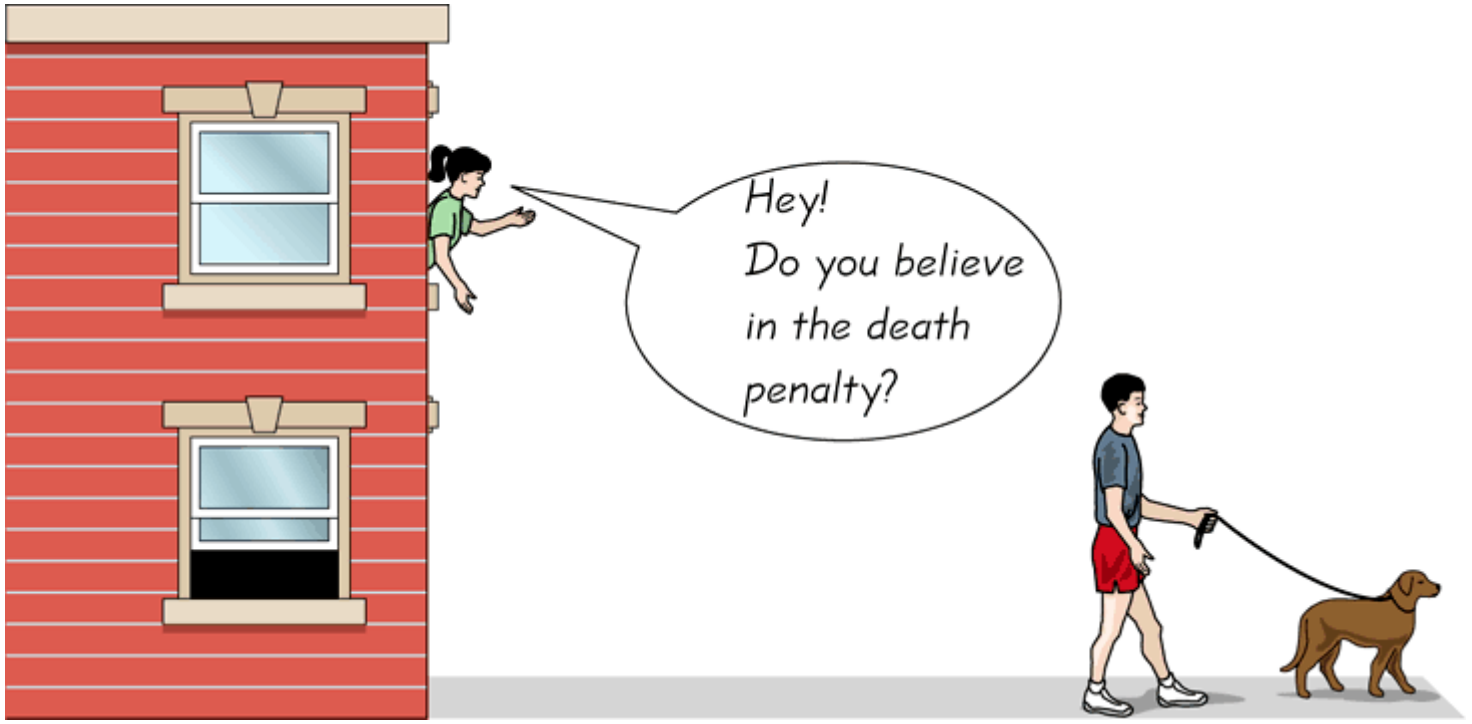
Bài toán: Điều tra sự yêu thích tham gia hoạt động Đoàn của sinh viên ĐHQG Tp HCM. Biết rằng ĐHQG Tp HCM là trường đại học quy mô lớn với hơn 100,000 sinh viên và có nhiều trường đại học thành viên như: ĐH KHTN, ĐH BK, ĐH KHXH & NV, ĐH KT-Luật,…

Giải pháp lấy mẫu phân tầng:

- Điều tra 6 trường, mỗi trường điều tra với số lượng SV khác nhau ngẫu nhiên…

Lấy mẫu tiện lợi (Convenience Sampling)

Lấy mẫu tiện lợi: là cách lấy mẫu mà kết quả được thu nhập một cách dễ dàng



Kỹ thuật chọn mẫu phi xác suất

Kỹ thuật chọn mẫu thuận tiện

- **Ví dụ:** Để mở spa thì điều tra đối tượng nào? Điều tra ngẫu nhiên ? Hay tập trung vào 1 nhóm đối tượng nào đó?

Kỹ thuật chọn mẫu định mức

- **Ví dụ:** Điều tra sự ảnh hưởng của giới tính và nơi ở đến vấn đề yêu thích hoạt động Đoàn của sinh viên trường đại học Công Nghiệp TP.HCM.
- **Giải pháp:** Sau khi thảo luận đưa ra giải pháp, nhóm nghiên cứu quyết định điều tra mẫu có kích thước 200 sinh viên, kèm theo yêu cầu:
  - Số sinh viên nữ chiếm ½ kích thước mẫu
  - Số sinh viên ở kí túc xá chiếm ½ kích thước mẫu

Kỹ thuật chọn mẫu phi xác suất

Kỹ thuật chọn mẫu phán đoán

- Chủ yếu dựa vào kinh nghiệm phỏng vấn
- **Ví dụ:** Phân tích điều kiện học tập ảnh hưởng đến điểm số của sinh viên chăm chỉ trong IUH
- **Giải pháp:**
  - Tổng thể sinh viên chăm chỉ của IUH
  - Chọn mẫu sinh viên chăm chỉ: phán đoán dựa trên điểm trung bình hoặc sinh viên dành nhiều thời gian vào thư viện. Sau đó tiến hành chọn mẫu ngẫu nhiên.

Sai số trong điều tra thống kê

Là chênh lệch giá trị số thu được qua điều tra so với giá trị số thực tế của hiện tượng

Phân loại

- **Sai số do đăng ký, ghi chép**
  - Vô ý khai báo, đăng ký, ghi chép sai
  - Cố tình khai báo, đăng ký, ghi chép sai
  - Đo lường
  - Hiểu sai nội dung câu hỏi
  - Ý thức xã hội…
- **Sai số do tính chất đại biểu**
  - Số lượng đơn vị mẫu không đủ lớn
  - Vi phạm nguyên tắc chọn mẫu ngẫu nhiên
  - Kết cấu tổng thể mẫu khác tổng thể chung…

Cách khắc phục sai số

Đối với sai số do đăng ký, ghi chép

- Làm tốt công tác chuẩn bị điều tra (soạn thảo câu bảng hỏi…)
- Làm tốt công tác kiểm tra, giám sát…

Đối với sai số do tính chất đại biểu

- Lựa chọn phương pháp tổ chức chọn mẫu phù hợp
- Tăng số đơn vị điều tra
- Đảm bảo nguyên tắc ngẫu nhiên

BÀI TẬP CÁ NHÂN VÀ NHÓM

Các đề tài nhóm

1. Khảo sát về thời gian sử dụng facebook, điện thoại của sinh viên IUH trong giờ học (7,2)
2. Các nhân tố ảnh hưởng đến việc ứng dụng thành công hệ thống ERp tại DOANH NGHIỆP (3)
3. Nghiên cứu ảnh hưởng của thuốc lá, rượu bia đến sức khoẻ của cộng đồng(1,8,4)
4. Điều tra sự yêu thích tham gia hoạt động Đoàn, thiện nguyện của sinh viên IUH (5)
5. Điều tra và phân tích ảnh hưởng của điều kiện học tập, sinh hoạt, ăn ở và môi trường sống đến kết quả học tập và rèn luyện của sinh viên IUH (10,9)

BÀI TẬP CÁ NHÂN VÀ NHÓM

1. Mỗi nhóm tìm hiểu các kỹ thuật chọn mẫu. Cho ví dụ minh hoạ cho mỗi kỹ thuật.
2. Mỗi nhóm chọn 1 đề tài phân tích áp dụng các kỹ thuật đã học để thu thập yêu cầu bao gồm Câu hỏi phỏng vấn, bảng câu hỏi khảo sát các câu hỏi đồng thời phải có câu trả lời.
3. Vẽ mỗi loại biểu đồ ít nhất 1 ví dụ minh hoạ theo kết quả thống kê của đề tài của mỗi nhóm.

https://www.amazon.com/dp/B0CRHHBFS7/ref=sspa_dk_detail_0?psc=1&pd_rd_w=UZCqi&content-id=amzn1.sym.eb7c1ac5-7c51-4df5-ba34-ca810f1f119a&pf_rd_p=eb7c1ac5-7c51-4df5-ba34-ca810f1f119a&pf_rd_r=86DEV3RMM7EZVK085C7C&pd_rd_wg=5vSa4&pd_rd_r=8bcd464a-6d63-4056-a9cd-d24ff8cc8385&sp_csd=d2lkZ2V0TmFtZT1zcF9kZXRhaWw

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

THANK YOU
Q & A

https://github.com/Hack-with-Github/Free-Security-eBooks