HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Nhập môn
Khoa học dữ liệu
(IT4142)

PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải
Team lecturers

ONE LOVE. ONE FUTURE.

2

Contents

- Lecture 1: Tổng quan về Khoa học dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- **Lecture 4: Phân tích và khám phá dữ liệu**
- Lecture 5: Trực quan hoá dữ liệu
- Lecture 6: Trực quan hoá dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiến độ bài tập lớn và hướng dẫn
- Lecture 10+11: Phân tích một số kiểu dữ liệu
- Lecture 12: Đánh giá kết quả phân tích

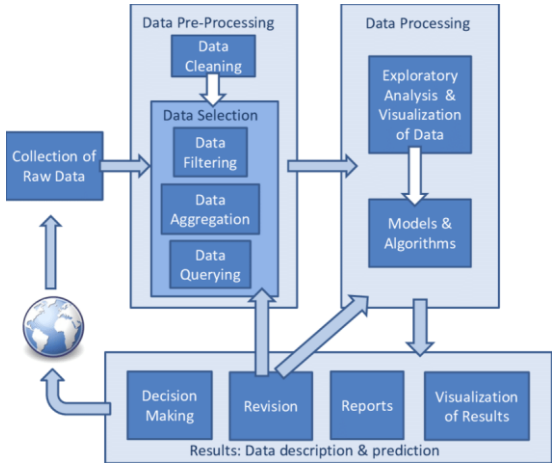
Mục tiêu bài giảng

- Hiểu các vấn đề cốt lõi trong phân tích thăm dò dữ liệu (EDA)
- Diễn giải và sử dụng các công cụ thống kê cho EDA
- Biểu diễn và diễn giải các đồ thị và biểu đồ cho EDA

Đặt vấn đề

- Muốn khai thác được dữ liệu, trước hết cần phải hiểu dữ liệu đang có
- Tại sao ?
  - Nhận biết các sai sót trong dữ liệu
  - Nhận biết các đặc trưng pattern của dữ liệu
  - Nhận biết nếu các giả định thống kê hiện tại không phù hợp với dữ liệu
  - Có căn cứ để đưa ra các giả thiết về dữ liệu
  - … nếu không hiểu dữ liệu, sẽ gặp nhất nhiều khó khăn để khai thác được giá trị từ dữ liệu

Quy trình làm khoa học dữ liệu



Trong quy trình này, EDA rất quan trọng nhưng thường đượcxem nhẹ, theo John Tukey

Trọng tâm của EDA

- EDA quan tâm tới cấu trúc, các ngoại lệ, và các mô hình từ dữ liệu
- EDA quan tâm tới tất cả các điểm dữ liệu trong tập dữ liệu
  - Các thống kê
  - Trực quan hóa
  - Phân cụm và phát hiện bất thường
  - Giảm chiều dữ liệu

EDA là gì

- EDA không phải là một tập các kỹ thuật, mà là một triết lý về cách mà chúng ta nên làm khi muốn hiểu về dữ liệu
  - Hỗ trợ lựa chọn đúng đắn các công cụ để tiền xử lý và phân tích dữ liệu
  - Cho phép sử dụng kinh nghiệm con người trong việc phát hiện và nhìn nhận các đặc trưng pattern của dữ liệu

Các câu hỏi chính khi phân tích EDA

- Giá trị tiêu chuẩn trong dữ liệu là bao nhiêu?
- Tính nhiễu của dữ liệu như thế nào?
- Dữ liệu có tuân theo phân bố nào không?
- Đặc trưng nào trong dữ liệu quan trọng với bài toán cần phân tích?
- Đặc trưng này của dữ liệu có quan trọng với bài toán cần phân tích không?
- Có thể mô tả mối tương quan giữa đặc trưng của dữ liệu và bài toán phân tích như thế nào?
- Có thể phân rã tín hiệu đúng và nhiễu trong dữ liệu hay không?
- Có thể trích xuất cấu trúc từ dữ liệu hay không?
- Dữ liệu có ngoại lệ outliers hay không?
- ...

EDA là một quá trình lặp

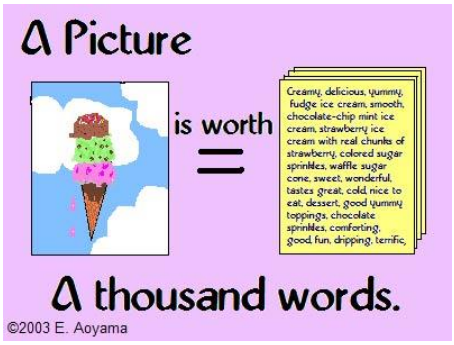
- Repeat...
  - Nhận diện và ưu tiên các câu hỏi liên quan theo thứ tự giảm dần độ quan trọng
  - Đặt câu hỏi
  - Xây dựng các đồ thị, biểu đồ để có thể trả lời câu hỏi đặt ra
  - Xem xét các kết quả và đặt ra các câu hỏi mới

Chiến lược khi phân tích EDA

- Kiểm tra từng biến (đặc trưng) của dữ liệu một cách lần lượt, sau đó mới tiến hành xem xét các mối liên hệ giữa các biến
- Bắt đầu bằng các đồ thị, sau đó tính toán các thống kê về một khía cạnh nhất định có liên quan của dữ liệu
- Chú ý tới kiểu dữ liệu
  - Kiểu số vs. Kiểu phân loại

Các kỹ thuật có thể dùng trong EDA

- Kỹ thuật đồ họa
  - Biểu đồ scatter plots, character plots, box plots, histograms, probability plots, residual plots, and mean plots.
- Kỹ thuật định lượng



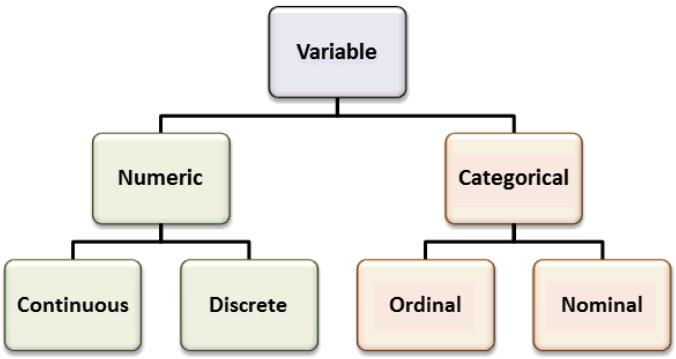
Phân tích thăm dò đơn biến

Quan sát và biến

- Dữ liệu là một tập hợp các quan sát (observations)
- Một thuộc tính là tập các giá trị mô tả một khía cạnh trên toàn bộ các quan sát, được gọi là một biến (variable)

HR Information		Contact	
Position	Salary	Office	Extn.
Accountant	$162,700	Tokyo	5407
Chief Executive Officer (CEO)	$1,200,000	London	5797
Junior Technical Author	$86,000	San Francisco	1562
Software Engineer	$132,000	London	2558

Các kiểu Variables



Số chiều của tập dữ liệu

- Đơn biến univariate: Mỗi quan sát chỉ gồm một biến
- Hai biến bivariate: Mỗi quan sát thực hiện với 2 biến
- Đa biến Multivariate: Quan sát thực hiện với nhiều biến

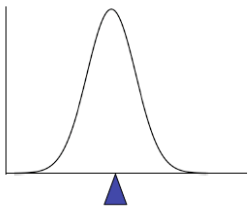
Đo bình quân - central tendency

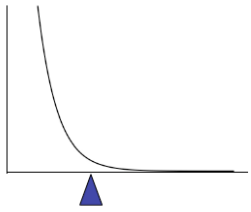
- Đo vị trí - Measures of Location: đánh giá tham số vị trí cho một phân bố. Vd., tìm số bình quân hay giá trị trung bình của dữ liệu
- Đo quy mô - Measures of Scale: đánh giá độ phân tán, độ biến thiên của một tập dữ liệu.
- Phép đo Skewness và Kurtosis

Số bình quân - Mean

- Đánh giá giá trị trung bình của một tập các quan sát, lấy tổng các giá trị chia cho số lượng các quan sát

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$





Số trung vị - Median

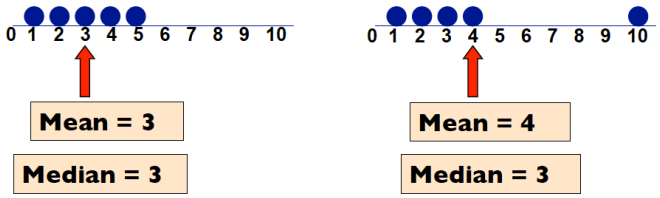
- Trung vị là giá trị của điểm dữ liệu mà một nửa số điểm có giá trị nhỏ hơn và một nửa số điểm dữ liệu còn lại lớn hơn giá trị của nó
- Cách tính
  - Nếu có một số lẻ các quan sát, tìm điểm dữ liệu có giá trị ở giữa
  - Nếu có một số chẵn các quan sát, tìm 2 điểm dữ liệu ở giữa và lấy trung bình
- Ví dụ
  - Tuổi của người chơi: 17 19 21 22 23 23 23 38
  - **Median = (22+23)/2 = 22.5**

Số yếu vị - Mode

- Mode là giá trị của phần tử có số lần xuất hiện lớn nhất trong các quan sát
  - Vd. 3, 4, 5, 6, 7, 7, 7, 8, 8, 9. Mode = 7
  - Vd. 3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9. Mode = {7, 8} = 7.5

Độ đo vị trí nào phù hợp

- Mean phù hợp với phân bố đối xứng và không có ngoại lệ
- Median phù hợp với phân bố lệch tâm hoặc dữ liệu có ngoại lệ



Đo quy mô: Phương sai và độ lệch chuẩn

- Phương sai - Variance:
  - là một độ đo sự phân tán thống kê
  - là giá trị kỳ vọng của bình phương của độ lệch của X so với giá trị trung bình của nó

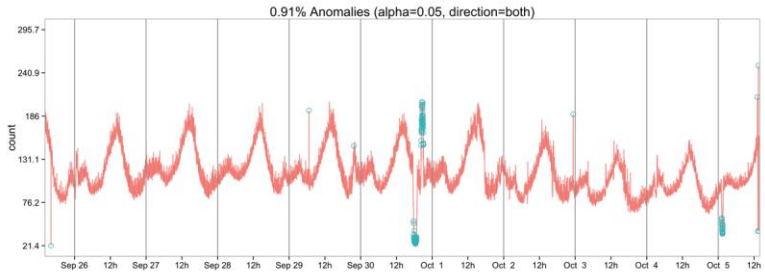
$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

- Độ lệch chuẩn - Standard Deviation:
  - căn bậc hai của phương sai

$$\hat{\sigma} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$$

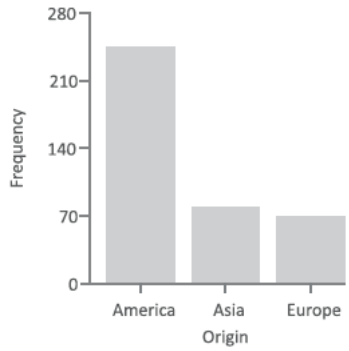
Run sequence plot

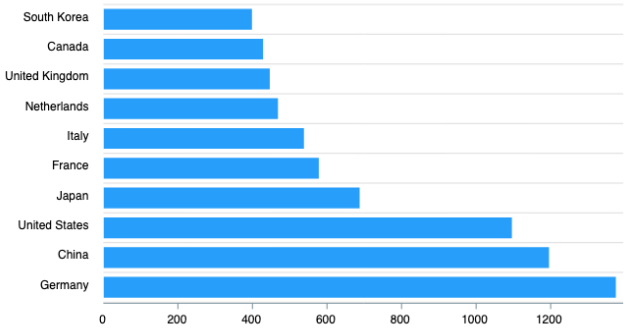
- Hiển thị các quan sát theo chuỗi thời gian
- Có thể được sử dụng để trả lời các câu hỏi
  - Có sự dịch chuyển nào về vị trí hay không?
  - Có sự dịch chuyển nào của phương sai hay không?
  - Có ngoại lệ nào không?



Bar charts

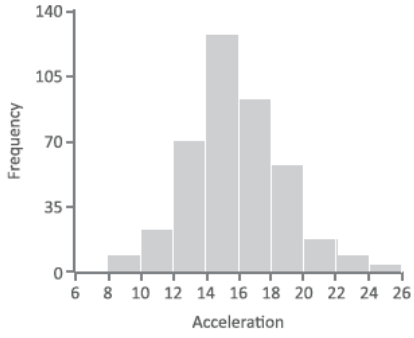
- Biểu đồ cột hiển thị mối tương quan về các giá trị đối với các biến



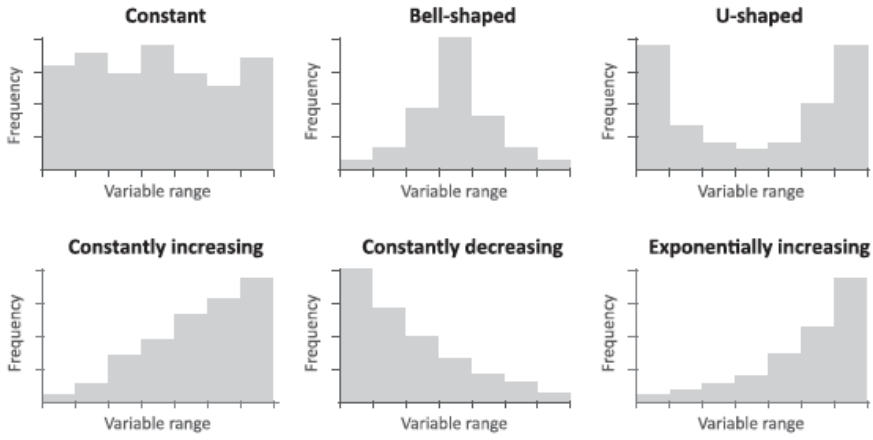


Histogram plot

- Biểu diễn thông tin tóm lược dưới dạng hình ảnh về phân bố của một tập dữ liệu đơn biến
- Histogram có thể dùng cho
  - Xem xét phân bố của tập các quan sát
  - Xem xét độ tập trung của dữ liệu
  - Xem xét sự phân tán của dữ liệu
  - Phân bố của dữ liệu là đối xứng hay lệch
  - Có ngoại lệ trong dữ liệu không?

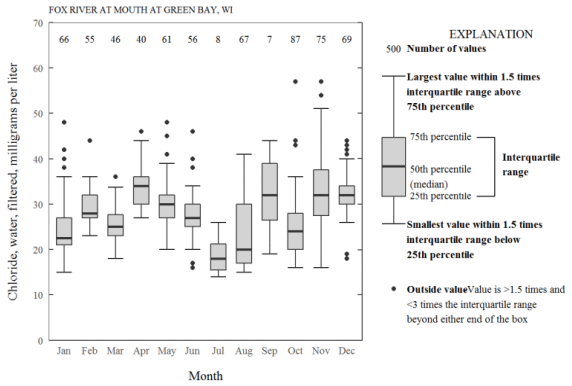


Ví dụ về các phân bố ứng với tần xuất các giá trị



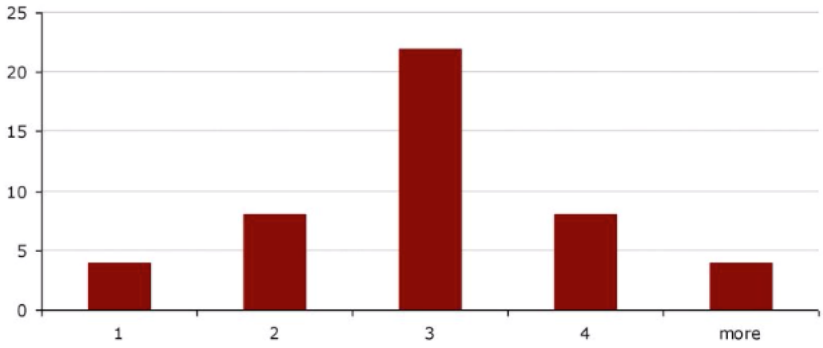
Box plot (2)

- Hiển thị giá trị nhỏ nhất, giá trị lớn nhất và các tứ phân vị
- Box plot có thể trả lời các câu hỏi
  - Có đặc trưng (biến) nào quan trọng ?
  - Độ tập trung vị trí có khác nhau giữa các nhóm con không?
  - Độ biến thiên có khác nhau giữa các nhóm con không?
  - Có ngoại lệ không?



Đo độ lệch - Skewness

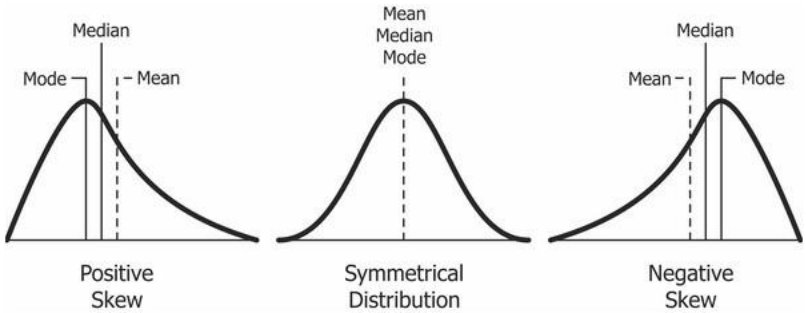
- Skewness đo sự bất đối xứng. Một phân bố hay một tập dữ liệu là đối xứng nếu nó là như nhau ở cả 2 phía từ vị trí trung tâm
- Dưới đây là phân bố đối xứng



Mean = median = mode = 3

Negative, positive skewness

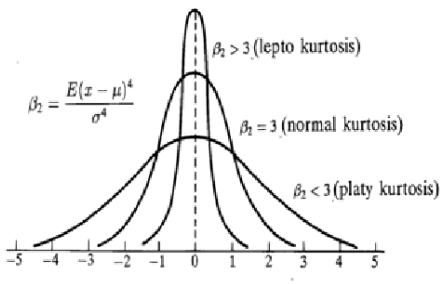
$$s_k = \frac{\sum_{i=1}^{T}(x_i - \bar{x})^3}{\sigma^3}$$



Độ nhọn - Kurtosis

- Kurtosis là chỉ số đo nếu một phân bố là nhọn hay phẳng so với một phân bố chuẩn
- Kurtosis càng cao, phân bố càng nhọn

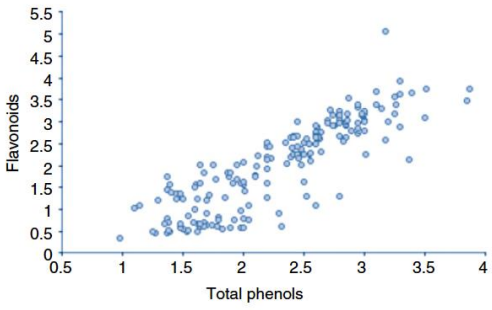
$$k = \frac{\sum_{i=1}^{T}(x_i - \bar{x})^4}{\sigma^4}$$



Phân tích hiểu mối quan hệ giữa các biến

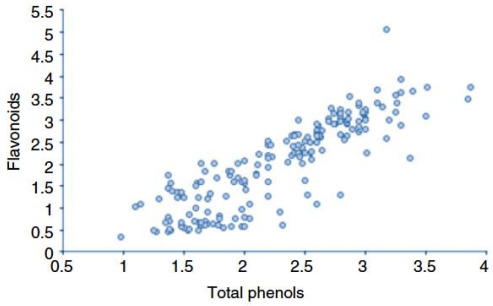
Scatter plot

- Cho phép nhận diện có hay không mối quan hệ giữa 2 biến
  - Mỗi biến được biểu diễn trên một trục x hoặc y
  - Mỗi điểm là một quan sát

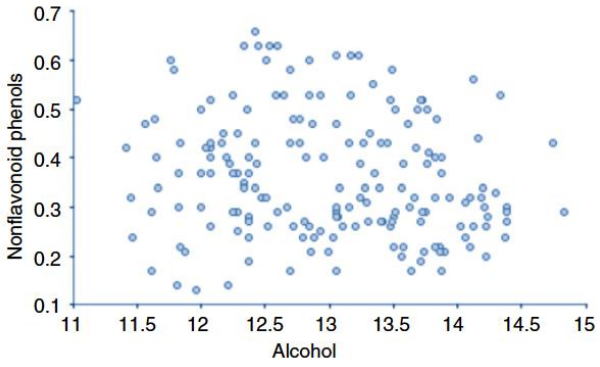


Scatter plot

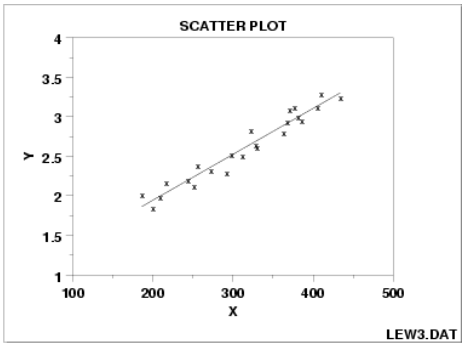
- Statter plot cho phép trả lời các câu hỏi sau
  - Có mối quan hệ giữa biến X và Y hay không?
  - Mối liên hệ có phải là tuyến tính hay không?
  - Mối liên hệ này là phi tuyến hay không?
  - Sự biến thiên của biến Y có phụ thuộc vào biến X hay không?
  - Có ngoại lệ hay không?

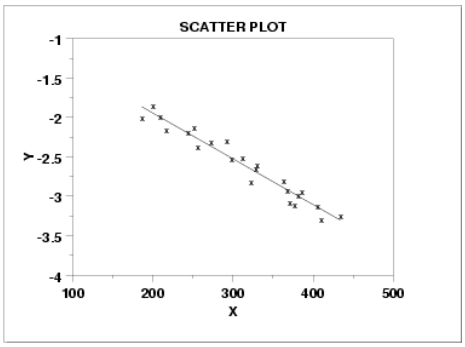


Scatter plot: Không có mối quan hệ

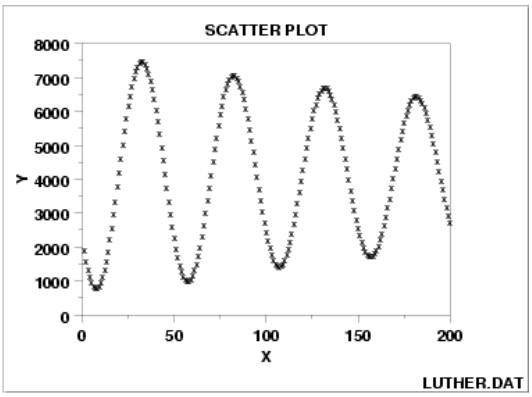


Scatter plot: Quan hệ tuyến tính (positive – negative correlation)

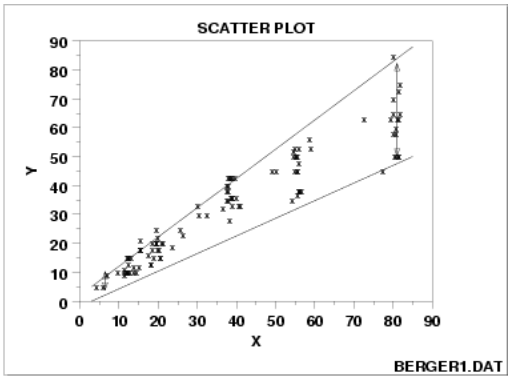




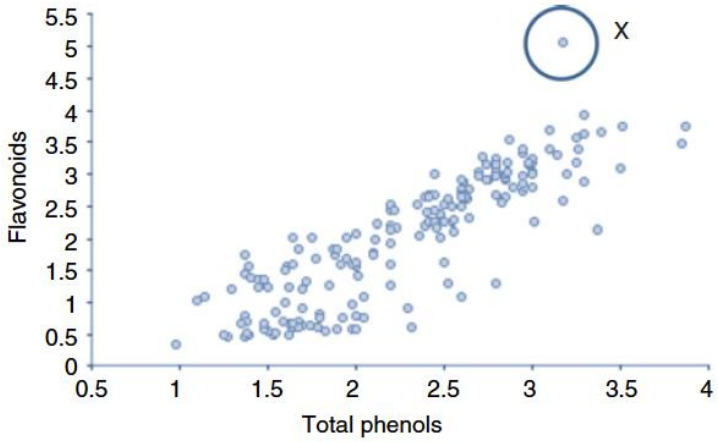
Scatter plot: Quan hệ hình sin



Scatter plot: Biến thiên của Y trong quan hệ với X

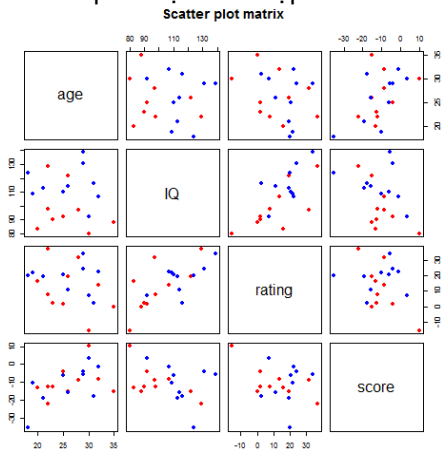


Scatter plot: Phát hiện ngoại lệ



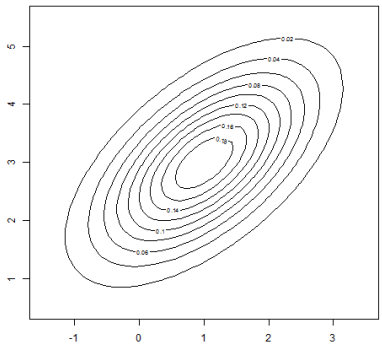
Scatterplot matrix

- Một tập hợp nhiều **scatter plots** tổ chức thành lưới hay matrix
- Mỗi scatter plot biểu diễn mối quan hệ của 1 cặp biến



Contour plots

- Hiển thị không gian đa chiều trên bề mặt không gian 2 chiều
- Contour line biểu diễn các giá trị cùng mức
- Contour plot thể hiện mối quan hệ Z thay đổi như thế nào với Y và X



Xác định các nhóm, phân cụm dữ liệu

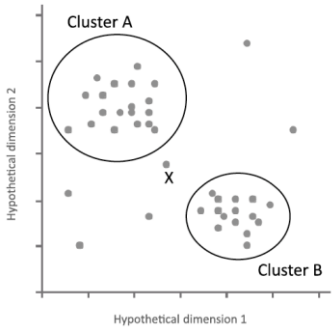
Clustering Methods in Exploratory Analysis

Đặt vấn đề

- Phân rã một tập dữ liệu thành các tập dữ liệu nhỏ hơn giúp hiểu cấu thành tập quan sát đầu vào
  - Làm rõ mối quan hệ gom cụm trong dữ liệu
  - Xác định các điểm quan sát mà khác biệt so với các cụm dữ liệu còn lại

Gom cụm - clustering

- Là cách thức nhóm các điểm dữ liệu mà tương tự nhau theo một cách nào đó – thường theo một vài tiêu chí đã xác định
- Đây là một dạng của học không giám sát
  - không có nhãn mô tả chúng ta nên gom cụm dữ liệu như thế nào



Làm thế nào để tìm các cụm dữ liệu?

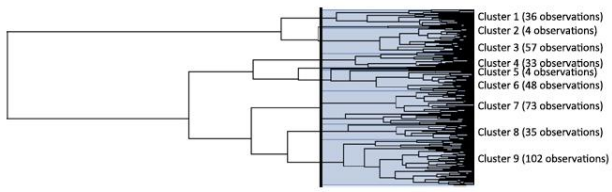
- Clustering tổ chức dữ liệu thành các nhóm/cụm
  - Thế nào là gần nhau?
  - Nhóm các điểm dữ liệu lại như thế nào?
  - Trực quan hóa các nhóm như thế nào?
  - Diễn dải kết quả gom cụm

Các kiểu clustering

- Gom cụm phân cấp - Hierarchical clustering
- Gom cụm phẳng - Flat clustering

Gom cụm phân cấp

- Sử dụng hướng tiếp cận theo cách thức kết tụ
  - Tìm những điểm ở gần nhau nhất
  - Đưa những điểm này vào nhóm
  - TÌm những điểm, cụm gần nhau tiếp theo
- Yêu cầu
  - Cần định nghĩa khoảng cách
  - Giải thuật theo hướng gộp từ nhỏ đến lớn
- Kết quả
  - Một cây biểu diễn các cụm được phân cấp (dendrogram)



Tính khoảng cách

- Một phương pháp gom cụm cần một cách để đo sự gần nhau giữa các quan sát
- Dữ liệu liên tục
  - Khoảng ách Euclidean
  - Độ đo tương quan (correlation similarity)
- Dữ liệu rời rạc
  - Khoảng cách Manhattan
- Cần phải lựa chọn hàm khoảng cách/tương đồng phù hợp với từng loại dữ liệu

Khoảng cách Euclidean

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

$$d_{A-B} = \sqrt{(0.7-0.6)^2 + (0.8-0.8)^2 + (0.4-0.5)^2 + (0.5-0.4)^2 + (0.2-0.2)^2}$$
$$d_{A-B} = 0.17$$

$$d_{A-C} = \sqrt{(0.7-0.8)^2 + (0.8-0.9)^2 + (0.4-0.7)^2 + (0.5-0.8)^2 + (0.2-0.9)^2}$$
$$d_{A-C} = 0.83$$

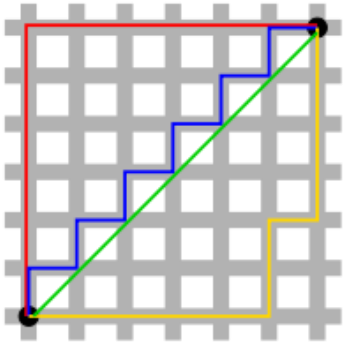
Khoảng cách Manhattan

- Tổng độ dài hình chiếu của các phân đoạn giữa 2 điểm trên trục tọa độ

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

where $(\mathbf{p}, \mathbf{q})$ are vectors

$$\mathbf{p} = (p_1, p_2, \ldots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \ldots, q_n)$$

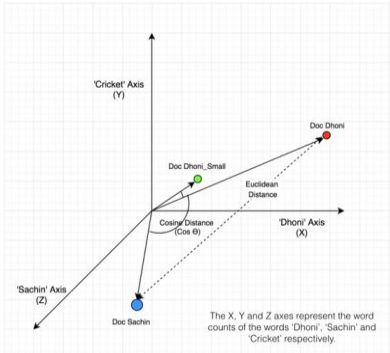


Khoảng cách Cosine

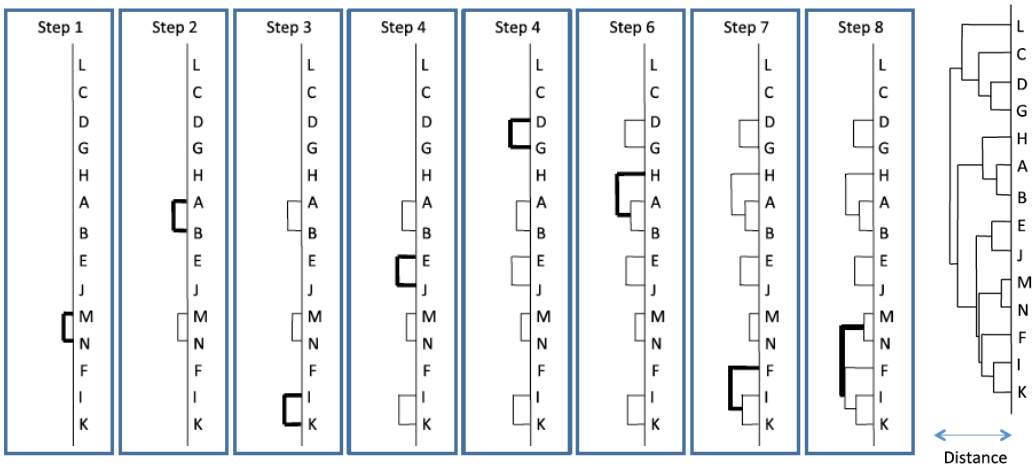
$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

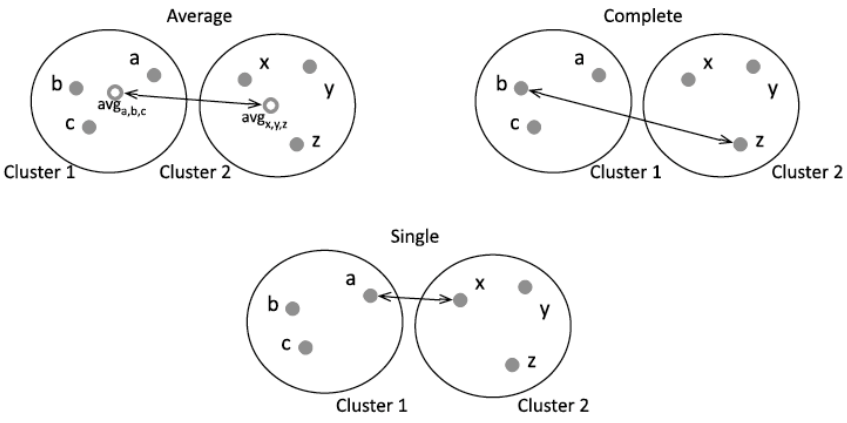
Projection of Documents in 3D Space



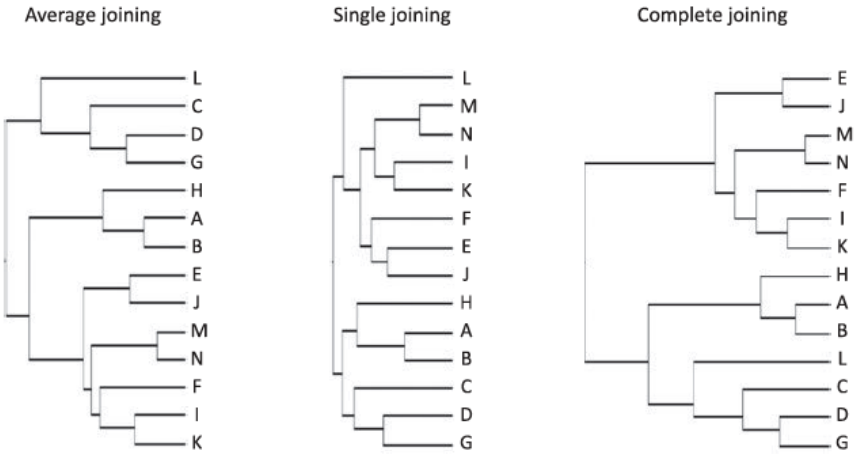
Giải thuật Agglomerative Hierarchical Clustering



Luật ghép cụm

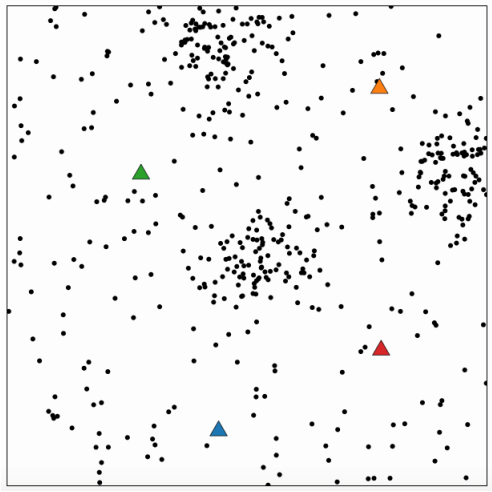


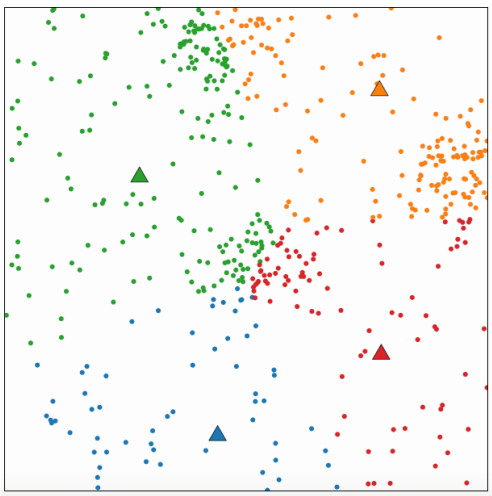
Kết quả AHC

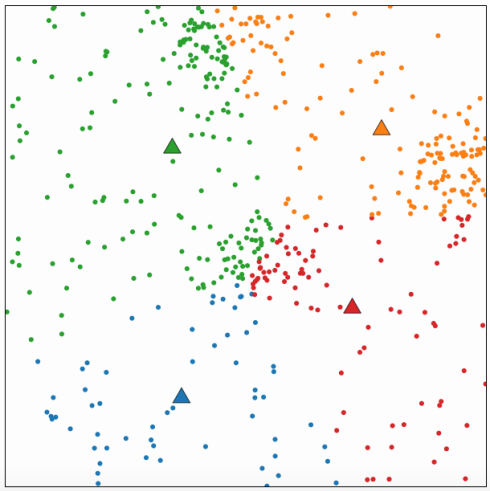


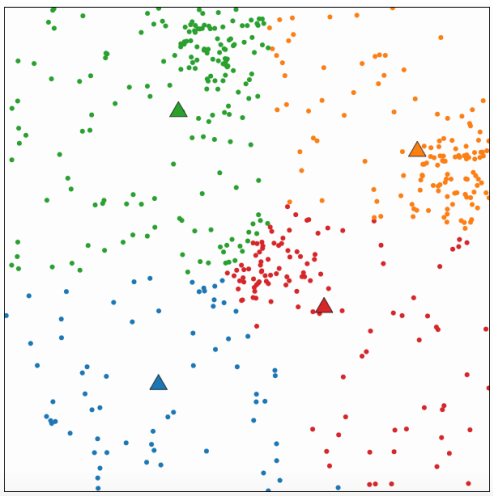
Phân cụm K-mean

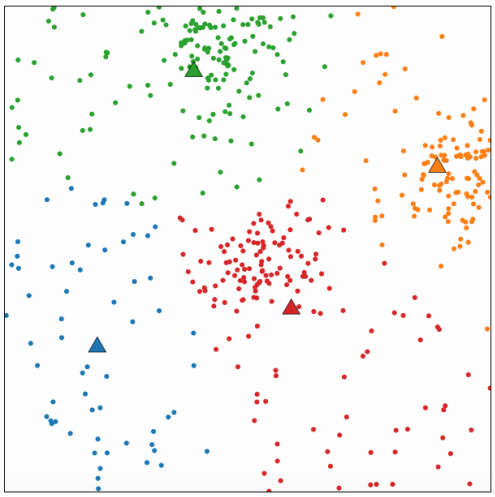
- Hướng tiếp cận phân tách
  - Cố định số lượng cụm
  - Tính toán "centroid" cho các cụm
  - Gán các điểm quan sát tới centroid gần nhất
  - Tính toán lại các centroid
- Yêu cầu
  - Một hàm khoảng cách
  - Số lượng các cụm
  - Một phép đoán khởi đầu cho các centroids
- Kết quả
  - Ước lượng các cụm theo centroid
  - Mỗi quan sát được gán theo 1 centroid











Giảm chiều dữ liệu

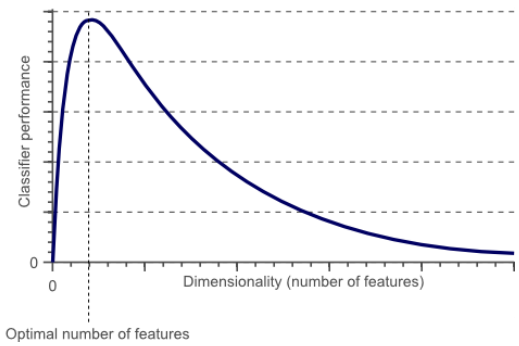
Principal Components Analysis and Singular Value Decomposition

Đặt vấn đề

- Phần lớn các giải thuật học máy và phân tích dữ liệu không hiệu quả với dữ liệu có số chiều quá lớn
  - Các đặc trưng không liên quan hoặc dư thừa có thể tạo nhiễu
  - Số lượng các chiều nguyên thủy của dữ liệu có thể nhỏ hơn thực tế

Vấn đề về số lượng chiều của dữ liệu

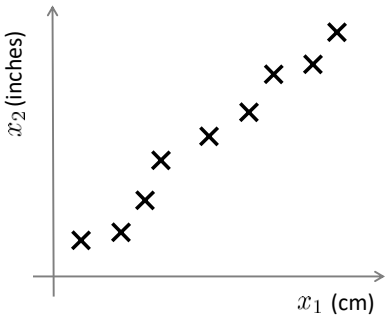
- Số lượng các mẫu để đạt được cùng chất lượng tăng lên theo cấp số nhân khi tăng số lượng các chiều của dữ liệu
- Trong thực tế, số lượng các mẫu (quan sát) là cố định
  - Hiệu năng hàm phân lớp thường giảm khi tăng số lượng các đặc trưng



Hướng tiếp cận

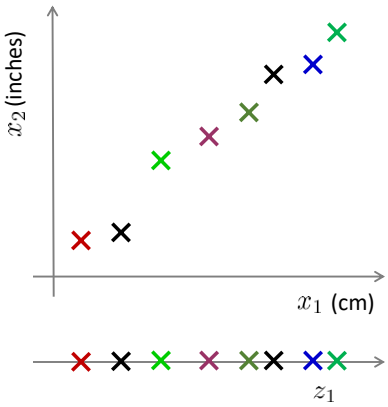
- Giảm chiều là một hướng tiếp cận để giảm kích thước dữ liệu
- Ưu điểm
  - Trực quan hóa : đưa không gian đa chiều về không gian 2, 3 chiều
  - Nén dữ liệu: tối ưu cho lưu trữ và truy xuất
  - Giảm nhiễu: tạo hiệu ứng tích cực tăng độ chính xác

Nén dữ liệu



Reduce data from
2D to 1D

Nén dữ liệu (2)

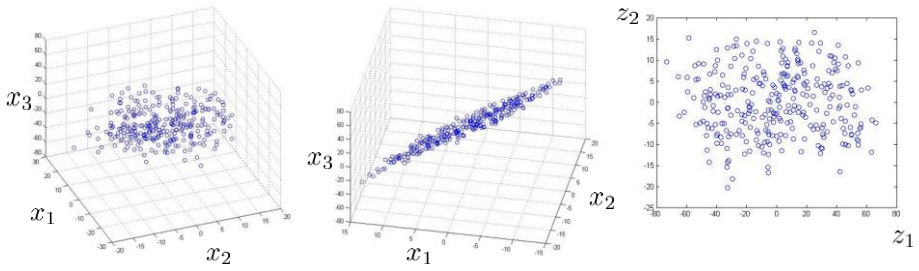


Reduce data from 2D to 1D

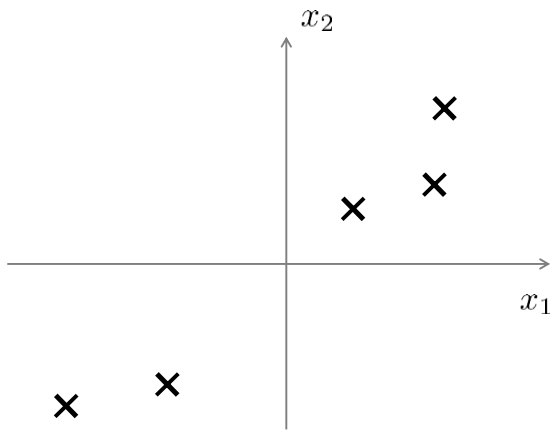
$$x^{(1)} \rightarrow z^{(1)}$$
$$x^{(2)} \rightarrow z^{(2)}$$
$$\vdots$$
$$x^{(m)} \rightarrow z^{(m)}$$

Nén dữ liệu (3)

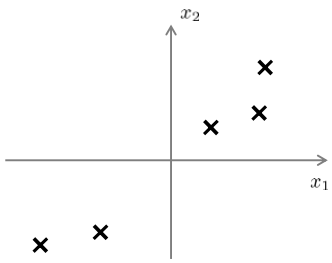
Reduce data from 3D to 2D

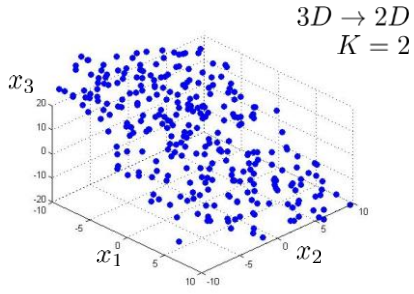


Phân tích thành phần chính - Principal Component Analysis (PCA)



Phát biểu sơ bộ về PCA





Reduce from 2-dimension to 1-dimension: Find a direction (a vector $u^{(1)} \in \mathbb{R}^n$) onto which to project the data so as to minimize the projection error.

Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, so as to minimize the projection error.

THANK YOU !

Exploratory data analysis in Tableau

CitiesExt.csv

- Ten countries with the highest population, bar chart showing populations
- Pie chart showing relative number of cities with negative longitude and positive longitude. Label the two slices “west” for west of the Prime Meridian (negative longitude), and “east” for east of the Prime Meridian (positive longitude)
- Is there is any relationship between the latitude of cities in a country (x-axis) and the population of that country (y-axis) (scatter plot)

PlayersExt.csv

- Create a bar chart showing the average number of minutes played by players in each of the four positions.
- Create a stacked bar chart for teams that played more than 4 games, showing their number of wins, draws, and losses.
- Create a pie chart showing the relative percentage of teams with 0, 1, and 2 red cards. Note: the pie should have three slices.
- Create a scatterplot of players showing passes (y-axis) versus minutes (x-axis). (Why are there some lines of dots?)
- Create a map of countries colored light to dark blue based on how many goals their team made (“goalsFor”).
- Create a pie chart showing the relative percentage of players making <= 0.25 passes per minute, >= 0.5 passes per minute, and between 0.25 and 0.5.

Lag plot

- Lag plots can provide answers to the following questions:
  - 1. Are the data random?
  - 2. Is there serial correlation in the data?
  - 3. What is a suitable model for the data?
  - 4. Are there outliers in the data?