

TS. Nguyễn Thế Vịnh
TS. Quách Xuân Trường
TS. Trần Quang Quý
ThS. Nguyễn Văn Việt

GIÁO TRÌNH PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU

2023

Trường Đại học
CNTT & TT, ĐHTN
Quyết Thắng,
Thái nguyên
<https://ictu.edu.vn/>

DANH MỤC HÌNH VẼ

Hình 1-1 Cuộc xâm lược nước Nga của Napoléon	9
Hình 1-2 Phân phối của các nhóm dữ liệu.....	10
Hình 1-3 Bản đồ bùng phát dịch tả năm 1854 trên đường Broad, London	11
Hình 1-4 Dữ liệu số trong học máy.....	12
Hình 1-5 Trực quan hoá dữ liệu số trong học máy.....	12
Hình 1-6 Dữ liệu văn bản trên mạng xã hội.....	13
Hình 1-7 Trực quan hoá dữ liệu mạng xã hội.....	13
Hình 1-8 Một số loại đồ thị để trực quan dữ liệu	19
Hình 1-9 Số người chết vì súng ở Florida.....	20
Hình 1-10 Sơ đồ lựa chọn cách trực quan dữ liệu	22
Hình 1-11 Biểu đồ choropleths (a) và isopleths (b).....	22
Hình 1-12 Ví dụ về biểu đồ cột.....	23
Hình 1-13 Ví dụ về biểu đồ hộp và râu	23
Hình 1-14 Ví dụ về biểu đồ dấu đầu dòng (Bullet chart)	24
Hình 1-15 Ví dụ về biểu đồ Gantt.....	24
Hình 1-16 Ví dụ về bản đồ nhiệt về thể hiện thời gian vị trí con trỏ chuột trên web ...	25
Hình 1-17 Ví dụ về bảng nổi bật.....	25
Hình 1-18 Ví dụ về biểu đồ tần suất.....	26
Hình 1-19 Ví dụ về biểu đồ hình tròn.....	26
Hình 1-20 Ví dụ về biểu đồ tree map	27
Hình 1-21 Ví dụ biểu đồ vĩ cầm.....	27
Hình 1-22 Công cụ trực quan hoá dữ liệu của Microsoft Office	28
Hình 1-23 Phần mềm Tableau cho việc trực quan hóa dữ liệu.....	29
Hình 1-24 Trực quan hóa dữ liệu với Microsoft Power BI.....	29
Hình 1-25 Google Data Studio	30
Hình 2-1 Trừu tượng hoá dữ liệu	40
Hình 2-2 Biểu đồ boxplot trên tập dữ liệu iris	43
Hình 2-3 Sử dụng ggplot để tạo biểu đồ boxplot.....	44
Hình 2-4 Biểu đồ rung lắc (Jitter chart)	45
Hình 2-5 Biểu đồ phân phối cho tập dữ liệu iris.....	46
Hình 2-6 Mối tương quan giữa các biến (pair)	47
Hình 2-7 Xem chi tiết mối quan hệ giữa 2 biến.....	48
Hình 2-8 Trực quan hóa số liệu mối tương quan	49
Hình 2-9 Phân loại dữ liệu	51
Hình 3-1 Trừu tượng hóa tác vụ thông qua hàm.....	62
Hình 3-2 Trừu tượng hóa tác vụ thông qua gói.....	63
Hình 3-3 Biểu đồ cột trong gói.....	64
Hình 3-4 Kết hợp các biểu đồ trong gói.....	64
Hình 3-5 Phân tích trực quan với gói.....	65
Hình 4-1 Bốn cấp độ hợp lệ dữ liệu	71
Hình 4-2 Cửa hàng vs doanh số.....	76

Hình 4-3 Doanh số của hàng số 14.....	78
Hình 4-4 Tăng trưởng Q3 (2012) của cửa hàng số 7.....	80
Hình 4-5 Doanh số theo ngày lễ giáng sinh.....	83
Hình 4-6 Doanh số theo ngày lễ lao động.....	84
Hình 4-7 Doanh số theo ngày lễ tạ ơn.....	85
Hình 4-8 Doanh số theo ngày super bowl.....	85
Hình 4-9 Trực quan hóa theo khoảng thời gian.....	87
Hình 4-10 Dữ liệu theo mùa.....	88
Hình 4-11 Sử dụng vòng lặp tạo biểu đồ.....	90
Hình 4-12 Xử lý ngoại lệ nhiệt độ vs doanh số tuần.....	91
Hình 4-13 Xử lý ngoại lệ CPS vs doanh số tuần.....	92
Hình 4-14 Xử lý ngoại lệ tỉ lệ thất nghiệp vs doanh số tuần.....	92
Hình 4-15 Xử lý ngoại lệ giá nhiên liệu vs doanh số tuần.....	93
Hình 4-16 Xử lý ngoại lệ ngày lễ vs doanh số tuần.....	94
Hình 4-17 Xử lý dữ liệu ngoại lệ doanh số vs tháng.....	94
Hình 4-18 Doanh số theo quý.....	95
Hình 4-19 Biểu đồ nhiệt tương quan.....	97
Hình 5-1 Hệ tọa độ Descartes.....	101
Hình 5-2 Hệ tọa độ cực.....	102
Hình 5-3 Hệ tọa độ địa lý.....	102
Hình 5-4 Thành phần nhỏ (small multiples).....	104
Hình 5-5 Các lớp dữ liệu.....	104
Hình 5-6 Màu sắc không phù hợp với người bị mù màu.....	105
Hình 5-7 ColorBrewer (và gói R).....	106
Hình 5-8 Bảng màu có trong RColorBrewer.....	107
Hình 5-9 Điểm trung bình SAT.....	107
Hình 5-10 tiến trình đạt kỷ lục thế giới trong bộ môn bơi tự do.....	108
Hình 5-11 Đánh giá Sức khỏe và Liên kết với Chăm sóc Ban đầu (HELP).....	109
Hình 5-12 Phân bố dân số năm 2010 của Massachusetts.....	109
Hình 5-13 Biểu đồ phân tán giữa nhiệt độ và độ hư hỏng của vòng O.....	111
Hình 5-14 Phiên bản phóng to giữa nhiệt độ và mức độ thiệt hại vòng O.....	111
Hình 5-15 Dữ liệu cho hai biến khác nhau.....	112
Hình 5-16 Đối tượng kỹ thuật số 3D.....	114
Hình 5-17 Before Us is the Salesman's House.....	115
Hình 5-18 Máy Shakespeare.....	116
Hình 5-19 Top 10 thương hiệu thuốc lá bán chạy nhất từ 2004 - 2007.....	117
Hình 5-20 Phân bố nông nghiệp tại Hoa Kỳ.....	118
Hình 5-21 Mối liên hệ giữa các công việc.....	118
Hình 6-1 Bản đồ dịch tả của John Snow.....	121
Hình 6-2 Bản đồ dịch tả sử dụng R.....	122
Hình 6-3 Bản đồ ngữ cảnh.....	126
Hình 6-4 Phép chiếu bản đồ.....	128
Hình 6-5 Bản đồ ban đầu của John Snow tái hiện không chính xác, các ca tử vong cách nhau khá xa.....	130
Hình 6-6 Tái tạo lại bản đồ gốc của John Snow về đợt bùng phát dịch tả năm 1854.....	131

Hình 6-7 Khoảng cách giữa 2 điểm trên bản đồ	132
Hình 6-8 Hình ảnh tĩnh sử dụng Leaflet	134
Hình 7-1 Phân phối dữ liệu theo năm.....	140
Hình 7-2 Phân phối dữ liệu theo ngày	141
Hình 7-3 Số lượng chuyến bay theo ngày.....	142
Hình 7-4 Độ trễ chuyến bay theo phút	143
Hình 7-5 Thời gian khởi hành theo lịch trình	144
Hình 7-6 Làm tròn thời gian	144
Hình 7-7 Làm tròn thời gian theo ngày	145
Hình 7-8 sự phân bố các chuyến bay trong ngày cho mọi ngày trong năm	146
Hình 8-1 Đồ ngẫu nhiên Bernoulli vô hướng	152
Hình 8-2 Cấu hình kích thước và màu sắc	153
Hình 8-3 Đặt màu khác nhau cho các nút.....	154
Hình 8-4 Mã hóa cứng màu sắc	155
Hình 8-5 Mã hóa bằng bảng màu.....	155
Hình 8-6 Đặt kích thước nút khác nhau	156
Hình 8-7 Thay đổi kích thước nút theo dữ liệu.....	157
Hình 8-8 Giới hạn kích thước của nút.....	157
Hình 8-9 Thêm chú giải vào sơ đồ	159
Hình 8-10 Thêm nhãn vào sơ đồ.....	160
Hình 8-11 Thay đổi kích thước nhãn.....	160
Hình 8-12 Màu sắc của nhãn nút.....	161
Hình 8-13 Mức độ đổ màu.....	161
Hình 8-14 Đặt màu nền	162
Hình 8-15 Hình dạng nút đồng nhất.....	162
Hình 8-16 Hình dạng nút không đồng nhất	163
Hình 8-17 Kiểm soát độ trong suốt.....	163
Hình 8-18 Tăng độ trong suốt và kích cỡ	164
Hình 8-19 Đồ thị có hướng/vô hướng mặc định	165
Hình 8-20 Tinh chỉnh màu sắc cho tác nhân và sự kiện.....	166
Hình 8-21 Gán nhãn các nút và cạnh.....	166
Hình 8-22 Thay đổi độ trong suốt của các cạnh	167
Hình 8-23 Thay đổi kích thước và màu sắc cạnh	168
Hình 8-24 Ánh xạ màu cạnh với các thuộc tính.....	169
Hình 8-25 Thay đổi đường viền.....	169
Hình 8-26 Đặt hướng cho cạnh (mũi tên)	170
Hình 8-27 Thay đổi vị trí mũi tên trước nút.....	170
Hình 8-28 Tô màu cạnh và thuộc tính nút	171
Hình 8-29 Nhóm các mạng	172
Hình 8-30 Tô màu cạnh theo mạng.....	173
Hình 9-1 Sáu nghiên cứu điển hình về hệ thống thị giác đầy đủ bằng các công cụ khác nhau	175
Hình 9-2 Ma trận biểu đồ phân tán (SPLoM) hiển thị dữ liệu bào ngư.....	177
Hình 9-3 Chín phép đo scagnostics mô tả hình dạng biểu đồ phân tán, với các ví dụ về tập dữ liệu trong thế giới thực.	178

Hình 9-4 Scagnostics SPLOM phân tích tập dữ liệu bào ngư, hiển thị biểu đồ phân tán chi tiết trong cửa sổ bật lên	179
Hình 9-5 Bố cục VisDB hiển thị các thuộc tính riêng biệt hoặc được kết hợp trong một dạng xem.....	180
Hình 9-6 Bố cục VisDB cho năm thuộc tính, tám thuộc tính, 1000 mục.	181
Hình 9-7 Hướng và màu sắc của bố cục VisDB tuân theo thứ tự xoắn ốc.....	181
Hình 9-8 HCE hiển thị hoạt động của gen trong các bảng đa chiều.....	183
Hình 9-9 HCE khám phá các kết hợp thuộc tính theo cặp bằng cách sử dụng tổng quan về ma trận và biểu đồ phân tán.....	184
Hình 9-10 Chế độ xem xếp hạng theo tính năng HCE với tổng quan về danh sách và biểu đồ/ô đồ.....	185
Hình 9-11 PivotGraph: chế độ xem liên kết nút của mạng nhỏ, giới tính và phân chia công ty được mã hóa theo hình dạng nút	185
Hình 9-12 PivotGraph cho thấy giao tiếp giữa các giới tính chủ yếu xảy ra ở vị trí B	186
Hình 9-13 Thành ngữ phân cấp InterRing sử dụng tiêu điểm dựa trên biến dạng + tương tác ngữ cảnh.....	187
Hình 9-14 Biểu đồ dạng chòm sao	190

DANH MỤC BẢNG BIỂU

Bảng 1-1 Trung bình, phương sai và tương quan của 4 nhóm dữ liệu	9
Bảng 2-1 Xem nhanh dữ liệu với lệnh head().....	41
Bảng 2-2 Tóm tắt dữ liệu với lệnh summary().....	42
Bảng 2-3 Mối tương quan (dạng số) giữa các biến.....	48
Bảng 4-1 Dữ liệu bán hàng của Walmart.....	73
Bảng 4-2 Tổ chức lại dữ liệu	81
Bảng 4-3 Khung dữ liệu mới	81
Bảng 4-4 Tính toán trên dữ liệu mới.....	82
Bảng 4-5 Ma trận tương quan.....	96
Bảng 6-1 Dữ liệu thô ban đầu của Snow.....	121

THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	Mô tả
vis	Là viết tắt của visualization (trực quan hóa).
SPLOM	SPLOM hoặc ma trận biểu đồ phân tán (ScatterPLOt Matrix) là một kỹ thuật được phát minh bởi John Hartigan vào năm 1975, được sử dụng để xác định mối tương quan giữa một chuỗi các biến.
HCE	Là viết tắt của từ Hierarchical Clustering Explorer. Đây là công cụ khám phá phân cụm phân cấp được thiết kế để nhập bất kỳ dự án microarray nào có sẵn và kiểm tra tương tác tác động của các định nghĩa do người dùng xác định.
Scagnostics	Là viết tắt của cụm từ "scatterplot diagnostics". Nó là một thuật ngữ trong lĩnh vực thống kê và trực quan hóa dữ liệu.
Monotonic	Một biểu đồ phân tán được coi là monotonic nếu có một xu hướng tăng dần (monotonic increasing) hoặc giảm dần (monotonic decreasing) giữa các giá trị của hai biến được biểu diễn trên trục x và trục y
Stringy	Một biểu đồ phân tán được xem là "stringy" khi có sự tập trung của các điểm dữ liệu thành các dải hoặc chuỗi kéo dài theo một hướng cụ thể
Skinny	Một biểu đồ phân tán được xem là "skinny" khi nó có hình dáng hẹp và dài, thường là do sự tập trung của các điểm dữ liệu thành một dải hoặc chuỗi kéo dài theo một hướng cụ thể
Convex	Một biểu đồ phân tán được xem là "convex" khi các điểm dữ liệu tạo thành một hình dạng lồi (convex shape) hoặc hình dạng hộp (box shape)
Striated	Một biểu đồ phân tán được coi là "striated" khi có sự xuất hiện của các vết sọc, đường kẻ hoặc mẫu vân trên biểu đồ, tạo thành các hình dạng đặc trưng
Sparse	Một biểu đồ phân tán được xem là "sparse" khi có sự thưa thớt và phân tán của các điểm dữ liệu trên biểu đồ.
Clumpy	Một biểu đồ phân tán được xem là "clumpy" khi có sự tập trung của các điểm dữ liệu thành các nhóm hoặc cụm.
Skewed	Một biểu đồ phân tán được coi là "skewed" khi có sự chênh lệch hoặc méo lệch của phân phối dữ liệu trên trục x hoặc trục y.
Outlying	Một biểu đồ phân tán được coi là "outlying" khi có sự xuất hiện của các điểm dữ liệu nằm xa hơn so với phần lớn các điểm dữ liệu khác.

Mở đầu

Giáo trình *Phân tích và trực quan dữ liệu* được tập thể giảng viên Khoa Công nghệ thông tin biên soạn nhằm phục vụ cho việc giảng dạy theo chương trình đào tạo mới, kết hợp xen kẽ giữa lý thuyết và thực hành.

Nội dung giáo trình cung cấp cho sinh viên các kiến thức cơ bản về phân tích và trực quan dữ liệu; trừu tượng hóa dữ liệu và tác vụ; hiểu các thành phần cơ bản của dữ liệu trực quan; thao tác với các loại dữ liệu khác nhau; trực quan hóa dữ liệu theo không gian, thời gian, và mạng lưới. Sau khi hoàn thành học phần này, sinh viên có thể thành thạo việc trực quan hóa và phân tích các loại dữ liệu khác nhau, qua đó làm tăng khả năng tư duy logic và nâng cao khả năng cơ hội việc làm sau khi tốt nghiệp. Nội dung giáo trình gồm 9 chương:

Chương 1. Tổng quan về phân tích và trực quan dữ liệu.

Chương 2. Trừu tượng hóa dữ liệu.

Chương 3. Trừu tượng hóa tác vụ.

Chương 4. Bốn cấp độ về hợp lệ dữ liệu.

Chương 5. Các thành phần dữ liệu trực quan.

Chương 6. Dữ liệu trực quan theo không gian.

Chương 7. Dữ liệu trực quan theo thời gian.

Chương 8. Dữ liệu trực quan theo mạng lưới.

Chương 9. Phân tích dữ liệu trực quan qua các ví dụ điển hình.

Mặc dù tập thể tác giả đã dành nhiều thời gian và công sức để biên soạn, song khó tránh khỏi thiếu sót. Vì vậy, chúng tôi kính mong quý thầy cô và các bạn sinh viên đóng góp ý kiến để cuốn giáo trình được hoàn thiện hơn. Xin trân trọng cảm ơn !

Chương 1 : TỔNG QUAN VỀ PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU

Nội dung chính của chương

Chương này nhằm giới thiệu cho sinh viên vai trò và tầm quan trọng của trực quan hóa dữ liệu, các ví dụ điển hình cũng như các công cụ và tài liệu sẽ được sử dụng trong suốt quyển giáo trình. Bố cục của chương được chia thành 03 phần. Phần 01 giới thiệu một số các ví dụ điển hình khi áp dụng trực quan hóa dữ liệu so với phương pháp phân tích thuần túy, qua đó sinh viên có thể tự suy ngẫm và cân nhắc có cần phải trực quan hóa dữ liệu hay không. Phần 02 tập trung vào một số các khái niệm cơ bản, công cụ điển hình nhằm giúp cho sinh viên làm quen với chủ đề. Phần 03 hướng dẫn cài đặt một số công cụ, thư viện, dữ liệu để giúp cho sinh viên thực hành xuyên suốt quyển giáo trình.

Mục tiêu cần đạt được của chương

- Hiểu được tầm quan trọng của việc phân tích và trực quan dữ liệu
- Nhớ và liên tưởng tới một số các ví dụ điển hình của trực quan dữ liệu
- Nắm được một số công cụ, thư viện, dữ liệu hữu ích cho việc thực hành trực quan dữ liệu

Một số ví dụ điển hình của phân tích và trực quan dữ liệu

1.1 Cuộc xâm lược nước Nga của Napoléon

Việc học lịch sử có lẽ là vấn đề gây ra nhiều khó khăn đối với các bạn học sinh bởi lẽ nó liên quan đến rất nhiều câu chữ, dấu mốc thời gian, số lượng quân đội, địa điểm,... Chúng ta thường gặp những bài giảng lịch sử kiểu như:

“...Chiến dịch nước Nga là bước ngoặt trong các cuộc chiến tranh của Napoléon. Trong vòng 10 năm trước đó, quân Pháp thắng trận liên tục và xâm chiếm phần lớn châu Âu, nhưng tại Nga thì họ đã chịu thất bại lớn với hơn 550.000 quân thương vong hoặc đào ngũ. Thất bại này làm sụt giảm nghiêm trọng sức mạnh của Đế chế Pháp và lực lượng đồng minh, gây ra sự thay đổi lớn trong nền chính trị ở châu Âu và làm suy giảm đáng kể quyền bá chủ của Pháp ở châu Âu. Danh tiếng của Napoléon là một thiên tài quân sự bất khả chiến bại đã mất đi sau thất bại này. Các đồng minh của Pháp, đầu tiên là Phổ và sau đó là Áo, đã lần lượt phá vỡ liên minh và chuyển sang chống lại Pháp, gây nên chiến tranh Liên minh thứ sáu.

Chiến dịch bắt đầu ngày 24 tháng 6 năm 1812, khi Napoléon vượt sông Neman. Napoléon buộc Hoàng đế nước Nga là Aleksandr I ở lại trong Hệ thống phong tỏa Lục địa của Vương quốc Anh, mục đích chính là để tránh các đe dọa của Nga tới Ba Lan. Napoléon đặt tên cuộc xâm lăng này là chiến dịch Ba Lan lần thứ hai, người Nga tuyên bố phát động một cuộc chiến tranh Vệ quốc...

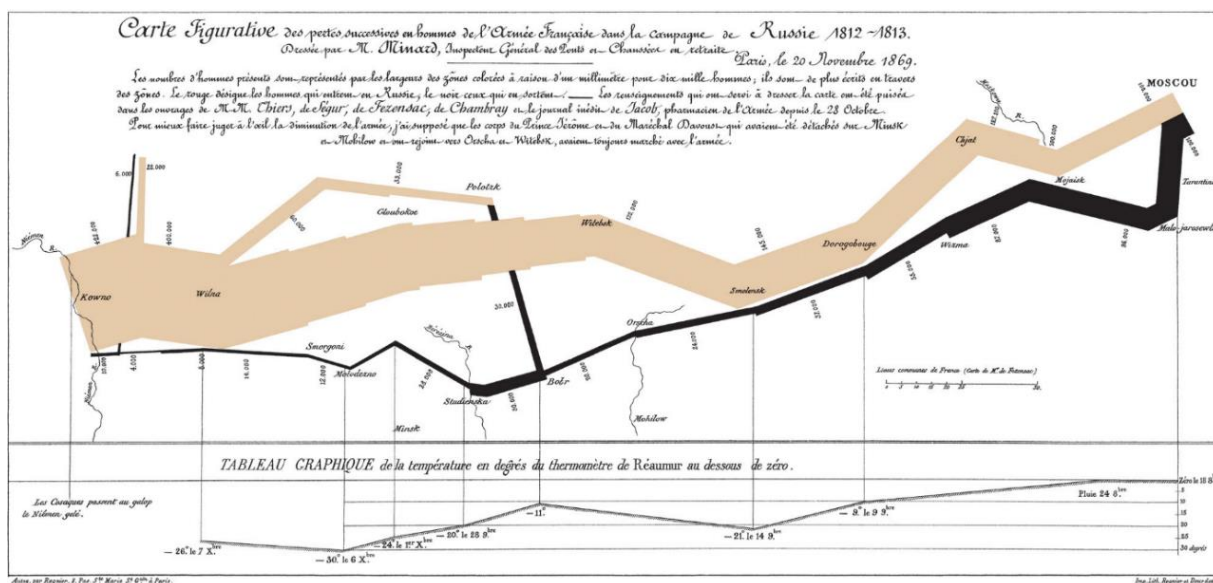
Hơn 680.000 tinh binh của đại quân Napoléon hành quân qua phía tây nước Nga, họ giành chiến thắng trong một số trận đánh nhỏ và một trận đánh lớn ở Smolensk vào 16-18 tháng 8. Tuy nhiên, trên cùng một ngày, cánh bắc của quân đội Nga do nguyên soái Pyotr Khristianovich Wittgenstein đã chặn cánh bắc của quân đội Pháp, dẫn đầu bởi Thống chế Nicolas Oudinot trong trận Polotsk. Điều này ngăn chặn cuộc hành quân của Pháp tới kinh thành Sankt-Peterburg; số phận của cuộc chiến

Câu hỏi đặt ra là:

Làm thế nào để hấp thụ những thông tin trên dễ dàng hơn?

Chúng ta hãy xem một cách tiếp cận khác về vấn đề lịch sử trên từ góc nhìn của trực quan dữ liệu. Biểu đồ của Minard [1] hiển thị sáu loại thông tin: địa lý, thời gian, nhiệt độ, hướng di chuyển của quân đội và số lượng quân còn lại (xem Hình 1-1). Chiều rộng của các con đường vàng (tiến quân) và đen (rút lui) thể hiện quy mô của lực lượng, một milimét cho 10.000 người. Các đặc điểm địa lý và các trận đánh lớn được

đánh dấu và đặt tên, đồng thời nhiệt độ giảm mạnh trên hành trình trở về được hiển thị dọc phía dưới.



Hình 1-1 Cuộc xâm lược nước Nga của Napoléon

Rõ ràng, chỉ với một bức hình nhưng nội dung mà nó có thể truyền đạt là **“đáng giá một nghìn từ”**

1.2 Phân tích dữ liệu thống kê

Chúng ta hãy cùng xem xét bốn nhóm dữ liệu sau [2], mỗi nhóm dữ liệu sẽ có 02 thuộc tính (X, Y). Ở đây, điểm trung bình là điểm nằm giữa của dữ liệu, phương sai là mức độ phân tán của dữ liệu, và tương quan là mối liên hệ giữa các thuộc tính (xem Bảng 1-1).

Bảng 1-1 Trung bình, phương sai và tương quan của 4 nhóm dữ liệu

	Nhóm 1		Nhóm 2		Nhóm 3		Nhóm 4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.23	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Trung bình	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Phương sai	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Tương quan	0.816		0.816		0.816		0.816	