TS. Nguyễn Thế Vịnh
TS. Quách Xuân Trưởng
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/

Chương 1 : TỔNG QUAN VỀ PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU 8
Nội dung chính của chương 8
Mục tiêu cần đạt được của chương 8
Một số ví dụ điển hình của phân tích và trực quan dữ liệu 8
1.1 Cuộc xâm lược nước Nga của Napoléon 8
1.2 Phân tích dữ liệu thống kê 9
1.3 Bản đồ bùng phát dịch tả năm 1854 trên đường Broad, London 10
1.4 Trực quan hoá dữ liệu số trong học máy 11
1.5 Trực quan hoá thông tin trên mạng xã hội 13
1.6 Các con số biết nói 14
Giới thiệu về phân tích và trực quan hóa dữ liệu 15
1.8 Phân tích dữ liệu 15
1.9 Trực quan dữ liệu 17
Các công cụ cho việc phân tích và trực quan dữ liệu 28
1.10 Công cụ và phần mềm trực quan 28
1.11 Nguồn dữ liệu phục vụ cho việc học tập 31
Tổng kết chương 1 32
Câu hỏi ôn tập 33
Bài tập thực hành 33
Chương 2 : TRỪU TƯỢNG HÓA DỮ LIỆU 40
Nội dung chính của chương 40
Mục tiêu cần đạt được của chương 40
Trừu tượng hoá dữ liệu 40
2.1 Tại sao lại trừu tượng hoá dữ liệu? 40
2.2 Mô tả dữ liệu 41
2.3 Các kiểu dữ liệu 51
Chuyển đổi dữ liệu 54
2.4 Tại sao lại chuyển đổi dữ liệu? 54
2.5 Làm sạch dữ liệu 55
2.6 Chuẩn hoá dữ liệu 55
2.7 Tóm tắt dữ liệu (Xem mục 2.2) 57
2.8 Xây dựng đặc trưng 57
2.9 Giảm chiều dữ liệu 58

Tổng kết chương 2 59
Câu hỏi ôn tập 59
Bài tập thực hành 59
Chương 3 : TRỪU TƯỢNG HÓA TÁC VỤ 61
Nội dung chính của chương 61
Mục tiêu cần đạt được của chương 61
Trừu tượng hóa tác vụ 61
3.1 Tại sao phải trừu tượng hóa các tác vụ? 61
3.2 Trừu tượng hoá tác vụ thông qua hàm 62
3.3 Trừu tượng hoá tác vụ thông qua gói 63
3.4 Trừu tượng hoá tác vụ thông qua hướng đối tượng 65
3.5 Trừu tượng hoá tác vụ thông qua điều khiển hàm 67
Phân loại tác vụ 67
3.6 Khám phá sự thay đổi 67
3.7 Khám phá sự phân phối 68
3.8 Khám phá sự tương quan 68
3.9 Khám phá sự tụ hợp, nhóm 68
3.10 Khám phá độ lệch 68
3.11 Khám phá xếp hạng 69
Tổng kết chương 3 69
Câu hỏi ôn tập 69
Bài tập thực hành 70
Chương 4 : BỐN CẤP ĐỘ VỀ HỢP LỆ DỮ LIỆU 71
Nội dung chính của chương 71
Mục tiêu cần đạt được của chương 71
Các cấp độ thế kế trực quan 71
4.1 Tại sao cần phải hợp lệ? 71
4.2 Tình huống miền 72
4.3 Trừu tượng dữ liệu, tác vụ 75
4.4 Hình thức mã hoá/tương tác trực quan 75
Thuật toán 89
Tổng kết chương 4 97
Câu hỏi ôn tập 98
Bài tập thực hành 98

Chương 5 : CÁC THÀNH PHẦN DỮ LIỆU TRỰC QUAN 100
Nội dung chính của chương 100
Mục tiêu cần đạt được của chương 100
Các thành phần của trực quan dữ liệu 100
5.1 Phân loại trực quan dữ liệu 100
5. 2 Màu sắc 104
5.3 Phân tích dữ liệu trực quan 107
Cốt truyện về trực quan dữ liệu 110
5.4 Câu chuyện về Challenger 110
5.5 Tạo một bài thuyết trình hiệu quả 113
5.6 Bức tranh lớn của trực quan hóa dữ liệu 114
Tổng kết chương 5 116
Câu hỏi ôn tập 116
Bài tập thực hành 117
Chương 6 : DỮ LIỆU TRỰC QUAN THEO KHÔNG GIAN 120
Nội dung chính của chương 120
Mục tiêu cần đạt được của chương 120
Dữ liệu trực quan theo không gian 120
6.1 Giới thiệu về dữ liệu không gian 120
6.2 Cấu trúc dữ liệu không gian 123
6.3 Tạo bản đồ 125
Tổng kết chương 6 132
Câu hỏi ôn tập 132
Bài tập thực hành 133
Chương 7 : DỮ LIỆU TRỰC QUAN THEO THỜI GIAN 134
Nội dung chính của chương 134
Mục tiêu cần đạt được của chương 134
Khám phá dữ liệu thời gian 134
7.1 Giới thiệu 134
7.2 Hiểu dữ liệu thời gian 135
7.3 Trực quan dữ liệu theo xu hướng 138
Tổng kết chương 7 145
Câu hỏi ôn tập 145
Bài tập thực hành 145

Chương 8 : DỮ LIỆU TRỰC QUAN THEO MẠNG LƯỚI 148
Nội dung chính của chương 148
Mục tiêu cần đạt được của chương 148
Giới thiệu về mạng lưới 148
8.1 Giới thiệu chung 148
8.2 Phần mềm trực quan hoá mạng 148
8.3 Thư viện lập trình phân tích và trực quan hóa mạng 149
Đồ thị vô hướng 149
8.4 Đồ thị ngẫu nhiên 149
8.5 Màu sắc và kích thước nút 150
8.6 Vị trí nút 151
8.7 Chú thích nút 156
8.8 Nhãn nút 157
8.9 Hình dạng nút và độ trong suốt 160
Đồ thị có hướng 162
8.10 Gán nhãn các cạnh 164
8.11 Kích thước và màu sắc cạnh 166
8.12 Kiểu đường viền 167
8.13 Mũi tên cạnh 168
8.14 Tô màu các cạnh từ thuộc tính nút 169
Tổng kết chương 8 171
Câu hỏi ôn tập 172
Bài tập thực hành 172
Chương 9 : PHÂN TÍCH DỮ LIỆU TRỰC QUAN QUA CÁC CASE STUDIES 173
Nội dung chính của chương 173
Mục tiêu cần đạt được của chương 173
Các ví dụ điển hình về phân tích dữ liệu trực quan 174
9.1 Tại sao phải phân tích Case Study? 174
9.2 Scagnostics 174
9.3 VisDB 177
9.4 Trình khám phá cụm phân cấp 180
9.5 PivotGraph 184
9.6 InterRing 186
9.7 Hệ thống chòm sao 188

Tổng kết chương 9 190
Câu hỏi ôn tập 190
Bài tập thực hành 190

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 đô 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 Tableu 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 toạ độ Descartes 101
Hình 5-2 Hệ toạ độ cực 102
Hình 5-3 Hệ toạ 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 ..... **Error! Bookmark not defined.**
Hình 6-4 Phép chiếu bản đồ ..... **Error! Bookmark not defined.**
Hình 6-5 Bản đồ ban đầu của John Snow tái hiện không chính xác, các ca tử vong cách nhau khá xa ..... **Error! Bookmark not defined.**
Hình 6-6 Tái tạo lại bản đồ gốc của John Snow về đợt bùng phát dịch tả năm 1854 ..... **Error! Bookmark not defined.**
Hình 6-7 Khoảng cách giữa 2 điểm trên bản đồ ..... **Error! Bookmark not defined.**

Hình 6-8 Hình ảnh tĩnh sử dụng Leaflet .....Error! Bookmark not defined.
Hình 7-1 Phân phối dữ liệu theo năm.....138
Hình 7-2 Phân phối dữ liệu theo ngày.....139
Hình 7-3 Số lượng chuyến bay theo ngày.....140
Hình 7-4 Độ trễ chuyến bay theo phút.....141
Hình 7-5 Thời gian khởi hành theo lịch trình.....142
Hình 7-6 Làm tròn thời gian.....142
Hình 7-7 Làm tròn thời gian theo ngày.....143
Hình 7-8 sự phân bổ các chuyến bay trong ngày cho mọi ngày trong năm.....144
Hình 8-1 Đồ ngẫu nhiên Bernoulli vô hướng.....150
Hình 8-2 Cấu hình kích thước và màu sắc.....151
Hình 8-3 Đặt màu khác nhau cho các nút.....152
Hình 8-4 Mã hóa cứng màu sắc.....153
Hình 8-5 Mã hóa bằng bảng màu.....153
Hình 8-6 Đặt kích thước nút khác nhau.....154
Hình 8-7 Thay đổi kích thước nút theo dữ liệu.....155
Hình 8-8 Giới hạn kích thước của nút.....155
Hình 8-9 Thêm chú giải vào sơ đồ.....157
Hình 8-10 Thêm nhãn vào sơ đồ.....158
Hình 8-11 Thay đổi kích thước nhãn.....158
Hình 8-12 Màu sắc của nhãn nút.....159
Hình 8-13 Mức độ đổ màu.....159
Hình 8-14 Đặt màu nền.....160
Hình 8-15 Hình dạng nút đồng nhất.....160
Hình 8-16 Hình dạng nút không đồng nhất.....161
Hình 8-17 Kiểm soát độ trong suốt.....161
Hình 8-18 Tăng độ trong suốt và kích cỡ.....162
Hình 8-19 Đồ thị có hướng/vô hướng mặc định.....163
Hình 8-20 Tinh chỉnh màu sắc cho tác nhân và sự kiện.....164
Hình 8-21 Gán nhãn các nút và cạnh.....164
Hình 8-22 Thay đổi độ trong suốt của các cạnh.....165
Hình 8-23 Thay đổi kích thước và màu sắc cạnh.....166
Hình 8-24 Ánh xạ màu cạnh với các thuộc tính.....167
Hình 8-25 Thay đổi đường viền.....167
Hình 8-26 Đặt hướng cho cạnh (mũi tên).....168
Hình 8-27 Thay đổi vị trí mũi tên trước nút.....168
Hình 8-28 Tô màu cạnh và thuộc tính nút.....169
Hình 8-29 Nhóm các mạng.....170
Hình 8-30 Tô màu cạnh theo mạng.....171
Hình 9-1 Sáu nghiên cứu điển hình về trực quan dữ liệu bằng các phương pháp và công cụ khác nhau.....173
Hình 9-2 Ma trận biểu đồ phân tán (SPLOM) với dữ liệu bào ngư. 175
Hình 9-3 Chín phép đo scagnostics mô tả hình dạng biểu đồ phân tán.....176
Hình 9-4 Biểu đồ Scagnostics SPLOM phân tích tập dữ liệu bào ngư, hiển thị biểu đồ phân tán chi tiết trong cửa sổ bật lên.....177

Hình 9-5 Bố cục VisDB hiển thị các thuộc tính riêng biệt hoặc được kết hợp trong một dạng xem. 178
Hình 9-6 Bố cục VisDB cho tám thuộc tính của 1000 dữ liệu 179
Hình 9-7 Hướng và màu sắc của bố cục VisDB tuân theo thứ tự xoắn ốc. 180
Hình 9-8 HCE hiển thị hoạt động của gen trong các bảng đa chiều 182
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 183
Hình 9-10 Chi tiết về chế độ xem HCE theo thứ tự thuộc tính. 184
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 184
Hình 9-12 PivotGraph cho thấy giao tiếp giữa các giới tính chủ yếu xảy ra ở vị trí B 186
Hình 9-13 InterRing hierarchy vis mã hóa trực quan theo hình thức lấp đầy không gian theo hướng tâm và tương tác tập trung+văn bản dựa trên biến dạng. 187
Hình 9-14 Biểu đồ dạng chòm sao 189

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 vằn 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 hỉnh.

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:

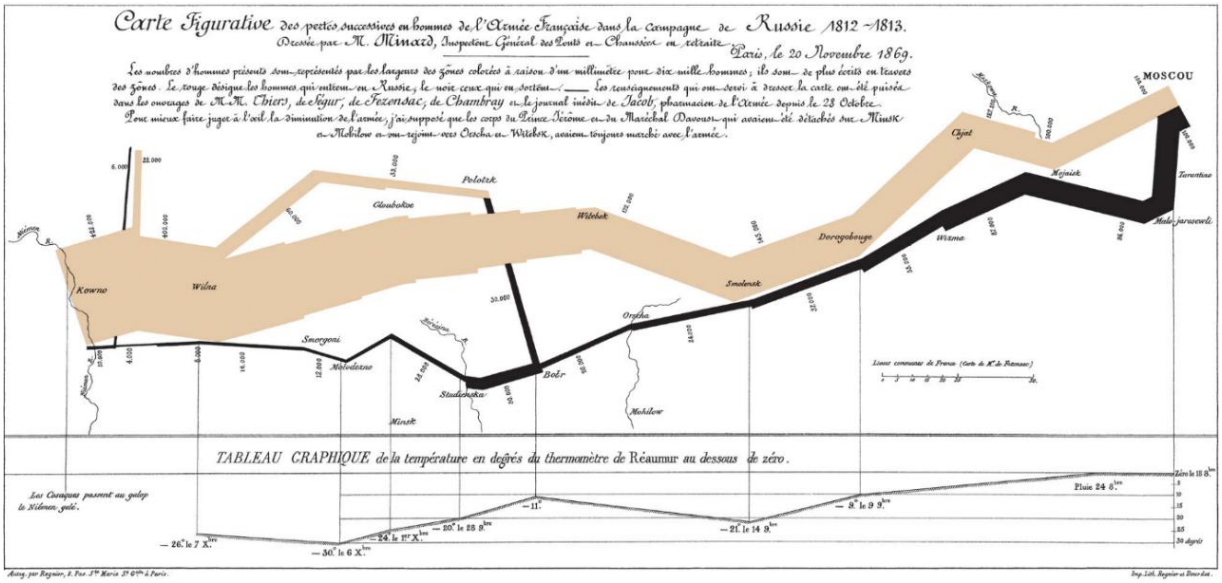
> "...Chiến dịch nước Nga là bước ngoặt trong các cuộc chiến tranh của Napoléon. Trong vòng 10 năm trước đó, quân Pháp thắng trận liên tục và xâm chiếm phần lớn châu Âu, nhưng tại Nga thì họ đã chịu thất bại lớn với hơn 550.000 quân thương vong hoặc đào ngũ. Thất bại này làm sụt giảm nghiêm trọng sức 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"**

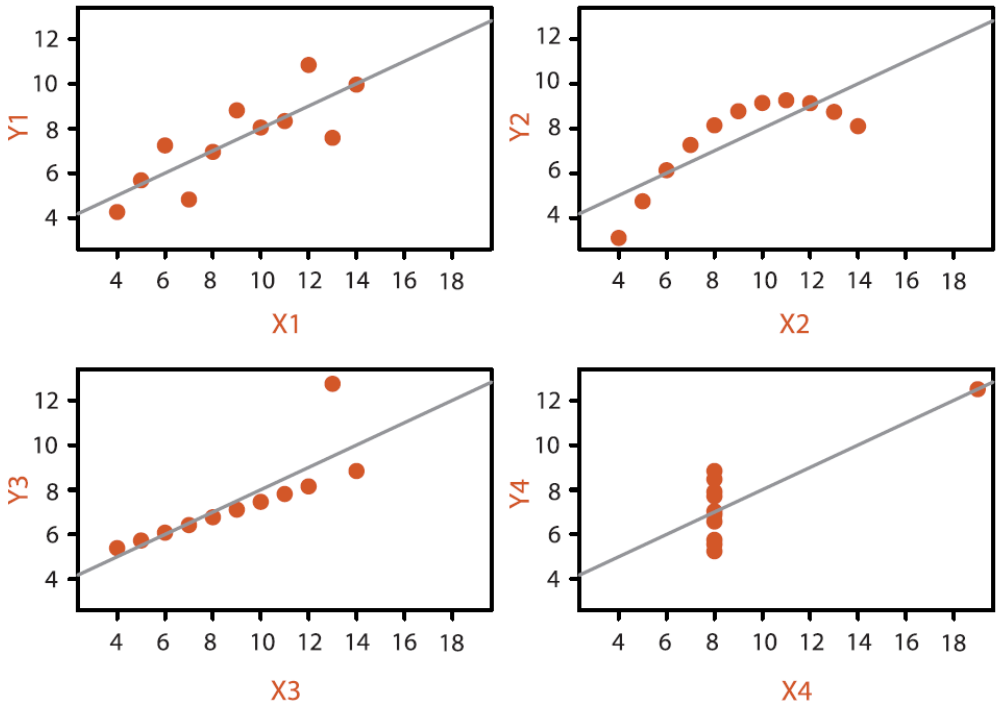
1.2 Phân tích dữ liệu thống kê

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	

Nếu chúng ta chỉ nhìn vào các con số thống kê ở trên, liệu chúng ta sẽ đưa ra kết luận gì khi mà giá trị trung bình, phương sai và số liệu về tương quan của bốn nhóm dữ liệu trên là đồng nhất? Vẫn cùng dữ liệu trên, chúng ta hãy xem xét dưới lăng kính của trực quan (xem Hình 1-2):



Hình 1-2 Phân phối của các nhóm dữ liệu

Dựa vào hình trên, chúng ta có thể thấy mặc dù dữ liệu có nhiều điểm tương đồng giữa bốn nhóm, nhưng rõ ràng ***phân bố*** và ***xu hướng*** dữ liệu là **khác nhau**, và điều này sẽ quyết định việc lựa chọn mô hình hay phương pháp phân tích dữ liệu sau này. Trong ví dụ trên, tập dữ liệu đầu tiên không có gì đặc biệt và phù hợp với trực giác. Biểu đồ phân tán thứ hai cho thấy một mẫu phi tuyến tính rõ ràng, cho thấy rằng việc sử dụng hồi quy tuyến tính không đủ để hiểu hoàn toàn dữ liệu. Tập dữ liệu thứ ba cho thấy một ngoại lệ đơn lẻ có thể làm sai lệch hồi quy vì nó không khớp hoàn toàn với phần còn lại của dữ liệu. Cuối cùng, tập dữ liệu thứ tư cho thấy một trường hợp đặc biệt khi các đo lường gây hiểu lầm lớn, với đường hồi quy gần như vuông góc với mẫu dữ liệu thực tế. Bộ tứ của Anscombe minh họa nguyên tắc rằng một bản tóm tắt đơn lẻ thường che giấu cấu trúc thực sự của dữ liệu và không đủ để hiểu rõ.

1.3 Bản đồ bùng phát dịch tả năm 1854 trên đường Broad, London

Vào những năm 1850, người ta tin rằng dịch tả lan truyền thông qua miasma (khí độc) trong không khí. Vi khuẩn chưa được hiểu rõ và sự bùng phát đột ngột và nghiêm trọng của dịch tả tại khu Soho ở London là một bí ẩn lớn. Nhưng làm thế nào chúng ta có thể tìm hiểu nguyên nhân gây ra dịch tả?



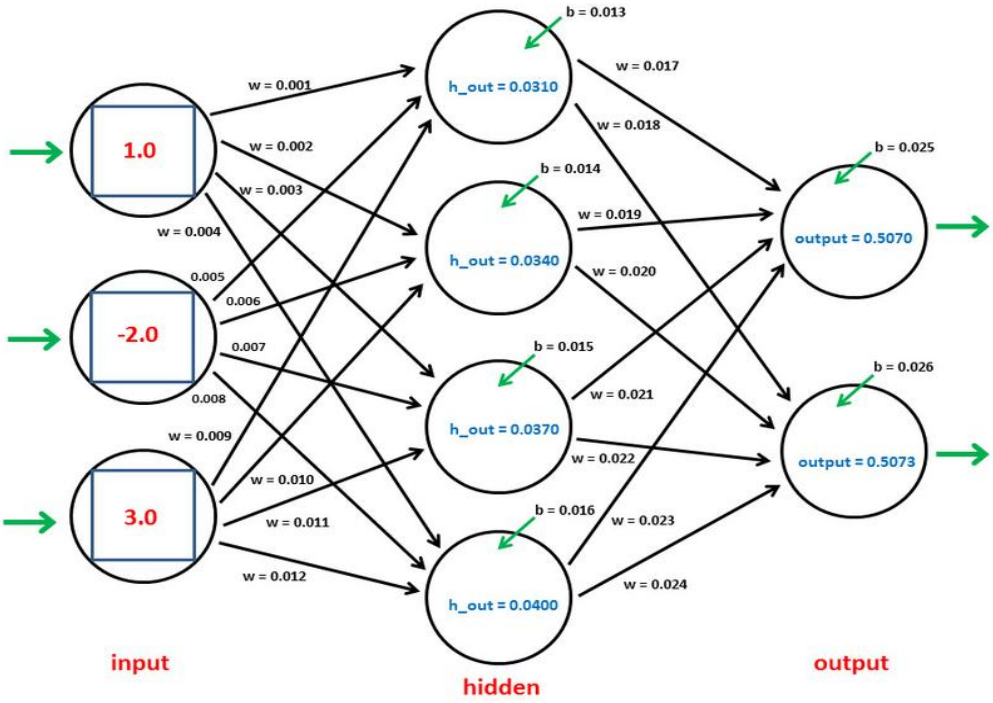
Hình 1-3 Bản đồ bùng phát dịch tả năm 1854 trên đường Broad, London

Và sản phẩm về trực quan hoá dữ liệu của bác sĩ Doctor John Snow ra đời [3]. Bức hình ở trên được xem là một trong những kiệt tác của trực quan hoá dữ liệu trong mọi thời đại (xem Hình 1-3)

Tư duy trực quan của Snow cũng giống như các nhà báo: Các thanh màu đen biểu diễn các ca tử vong, độ dài thanh thể hiện số lượng. Mức độ tập trung và độ dài của các thanh này cho thấy mộ số các khu phố có ca tử vong cao hơn những nơi khác. Dựa vào bản đồ trên, Snow phát hiện ra rằng "*Rõ ràng là các ca bệnh tập trung xung quanh máy bơm ở phố Broad (nay là Broadwick).*" Máy bơm (giếng) bị ô nhiễm bởi nước thải được đề cập đại diện cho khu vực có mật độ bùng phát dịch tả cao. Sau đó, giải pháp để ngăn chặn là xây dựng hệ thống thoát nước thải bảo vệ giếng khỏi sự ô nhiễm. Đây cũng là một hình ảnh trực quan cực kỳ thành công khi nó phát hiện ra nguyên nhân cốt lõi của vấn đề và truyền cảm hứng cho một sáng kiến hiệu quả.

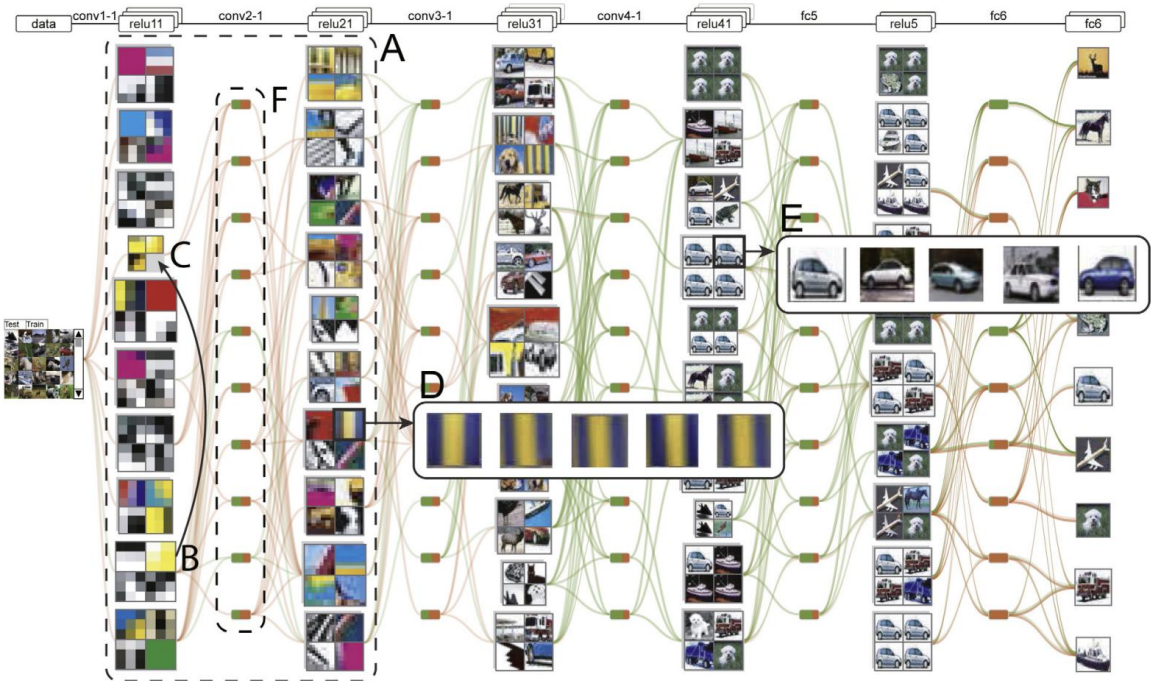
1.4 Trực quan hoá dữ liệu số trong học máy

Trong kỷ nguyên của dữ liệu lớn và trí tuệ nhân tạo, các mô hình (thuật toán) trong học máy, học sâu được xem là các "hộp đen" - bởi lẽ mô hình sẽ tự điều chỉnh các tham số trong mỗi lớp. Việc tìm hiểu sâu và chi tiết cách thức mô hình "học" từ dữ liệu như thế nào là một vấn đề thách thức nếu chúng ta chỉ đơn thuần nhìn vào các con số (tự điều chỉnh) trong mô hình như ví dụ dưới đây (xem Hình 1-4).



Hình 1-4 Dữ liệu số trong học máy

Câu hỏi mà chúng ta đang đặt là liệu có cách nào để hiểu được mô hình đang cố gắng học hoặc rút ra thông tin từ dữ liệu đầu vào không? Để tiếp cận vấn đề này, chúng ta có thể xem xét mô hình từ một góc nhìn khác, đó là trực quan hóa dữ liệu bằng cách chuyển đổi các con số thành hình ảnh, như trong Hình 1-5.

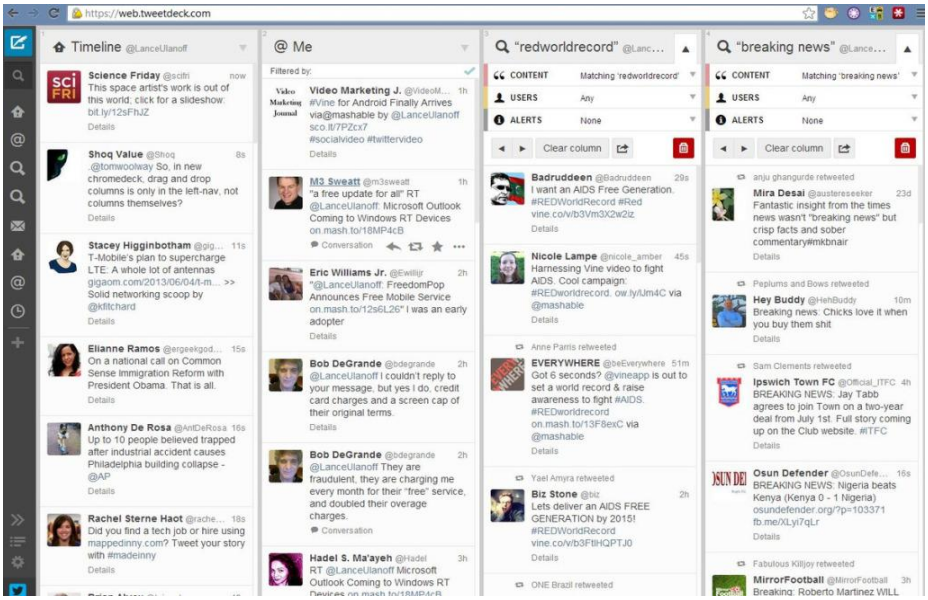


Hình 1-5 Trực quan hoá dữ liệu số trong học máy

Rõ ràng, nhìn vào hình trên, chúng ta có thể thấy mô hình đang cố gắng "học" những loại thông tin gì từ dữ liệu giữa các lớp.

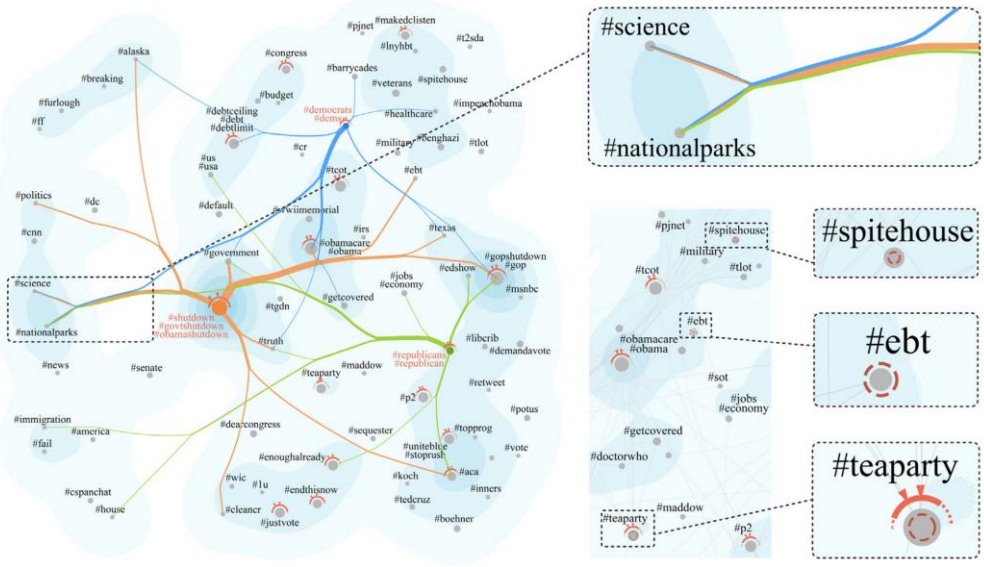
1.5 Trực quan hoá thông tin trên mạng xã hội

Trong thập kỷ qua, với sự bùng nổ của các trang web tin tức và mạng xã hội như Facebook, Zalo, Twitter, việc hiểu và kết nối thông tin với nhau đã trở thành một chủ đề quan trọng. Điều này đặt ra một thách thức lớn đối với các nhà phân tích và những người phải đưa ra quyết định, đặc biệt khi có một lượng lớn thông tin đang được hiển thị, tương tự như trong Hình 1-6.



Hình 1-6 Dữ liệu văn bản trên mạng xã hội

Việc phân tích dữ liệu sẽ có cơ sở hơn rất nhiều khi chúng ta trực quan hoá thông tin như trong ví dụ dưới đây. Rõ ràng, thông tin lúc này đã được phân nhóm (thể hiện qua đường bao), và kết nối (các đường kết nối) lại với nhau như trong Hình 1-7



Hình 1-7 Trực quan hoá dữ liệu mạng xã hội

Và còn rất nhiều những ví dụ khác nữa…. Thông qua 05 ví dụ cụ thể trên, ta có thể tự trả lời được câu hỏi có nên tìm hiểu về trực quan dữ liệu hay không.

1.6 Các con số biết nói

1. Theo nghiên cứu của Snejana Dineva, học sinh ghi nhớ được **65%** thông tin hình ảnh, so với 10-20% thông tin bằng văn bản hoặc nói. "*Dineva, S. (2019). The importance of visualization in e-learning courses. In dalam The 14 International Conference on Virtual Learning ICVL.*"
2. Hình ảnh minh họa hiệu quả hơn **83%** so với văn bản thuần tuý. "*Dineva, S. (2019). The importance of visualization in e-learning courses. In dalam The 14 International Conference on Virtual Learning ICVL.*"
3. Hình ảnh cải thiện việc học lên tới **400%** so với học trực tuyến dựa trên văn bản. "*Yarbrough, J. R. (2019). Infographics: in support of online visual learning. Academy of Educational Leadership Journal, 23(2), 1-15.*"
4. Bộ não của chúng ta truyền tải **90%** thông tin dưới dạng hình ảnh. (https://news.mit.edu/2014/in-the-blink-of-an-eye-0116)
5. Bộ não của chúng ta có thể xử lý thông tin hình ảnh nhanh hơn **60,000 lần**. "*Mitsea, Eleni, Athanasios Drigas, and Charalampos Skianis. "ICTs and Speed Learning in Special Education: High-Consciousness Training Strategies for High-Capacity Learners through Metacognition Lens." Technium Soc. Sci. J. 27 (2022): 230.*"
6. Trực quan có hiệu quả hơn **43%** trong việc thuyết phục khán giả. https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/
7. Nghiên cứu cho thấy rằng khi các nhóm sử dụng ngôn ngữ hình ảnh, họ có thể đạt được sự đồng thuận thường xuyên hơn **21%** so với các nhóm không sử dụng hình ảnh. https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/
8. Kết quả của một nghiên cứu cho thấy thông tin bằng văn bản, khi kết hợp với thông tin trực quan và hành động, sẽ dễ nhớ hơn **70%** so với chỉ có văn bản. https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/
9. Việc sử dụng ngôn ngữ hình ảnh cải thiện hiệu quả giải quyết vấn đề lên 19%, tạo ra kết quả cao hơn **22%** trong thời gian ít hơn **13%** và tăng tính thuyết phục của người thuyết trình lên **43%**. https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/
10. Một nghiên cứu được thực hiện tại Đại học Cornell cho thấy khi một tuyên bố khoa học được trình bày bằng con số hoặc bằng từ ngữ thuần túy thì 68% người đọc tin vào điều đó. Nhưng khi tuyên bố được kèm theo một biểu đồ đơn giản, con số đã tăng lên **97%**. "*Levie, W. H., & Lentz, R. (1982). Effects of text illustrations: A review of research. Ectj, 30(4), 195-232.*"

Danh sách trên chỉ là một trong rất nhiều các công bố, nghiên cứu về tính hữu ích của việc trực quan hoá dữ liệu.

Giới thiệu về phân tích và trực quan hóa dữ liệu

1.8 Phân tích dữ liệu

Trong thế giới hiện đại dựa trên dữ liệu, khả năng rút ra những thông tin có giá trị từ lượng thông tin khổng lồ trở nên ngày càng quan trọng. Đó là lý do tại sao phân tích dữ liệu đóng vai trò quan trọng. Phân tích dữ liệu là quá trình điều tra các tập dữ liệu lớn và phức tạp để khám phá các mẫu, đưa ra kết luận và đưa ra những quyết định thông minh. Nó liên quan đến việc sử dụng các kỹ thuật thống kê và toán học, cùng với các công cụ và công nghệ tiên tiến, để chuyển đổi dữ liệu thô thành những thông tin có ý nghĩa và giá trị.

Phân tích dữ liệu bao gồm một loạt các phương pháp và cách tiếp cận, bao gồm phân tích mô tả, phân tích chuẩn đoán, phân tích dự đoán và phân tích đề xuất. Phân tích mô tả tập trung vào việc tóm tắt và hiểu dữ liệu lịch sử để có cái nhìn rõ ràng hơn về những gì đã xảy ra trong quá khứ. Nó giúp chúng ta khám phá các đặc điểm, xu hướng và biến đổi trong dữ liệu. Phân tích chuẩn đoán tiến xa hơn bằng cách xác định nguyên nhân của các sự kiện và xu hướng trong quá khứ. Chúng ta sẽ điều tra các mối quan hệ và tương quan giữa các yếu tố khác nhau để hiểu rõ hơn về sự tương tác và ảnh hưởng của chúng. Phân tích dự đoán sử dụng các mô hình thống kê và thuật toán học máy để dự báo kết quả trong tương lai dựa trên những mẫu dữ liệu đã xảy ra trong quá khứ. Điều này giúp chúng ta đưa ra những dự đoán và ước tính về những gì có thể xảy ra và đưa ra quyết định thông minh. Cuối cùng, phân tích đề xuất đưa ra các hành động chính để đạt được kết quả mong muốn. Dựa trên phân tích trước đó, chúng ta có thể đề xuất các biện pháp và chiến lược tối ưu nhằm đạt được mục tiêu và cải thiện hiệu quả của quyết định và hành động..

Lĩnh vực phân tích dữ liệu dựa trên nhiều nguyên tắc khác nhau như thống kê, toán học, khoa học máy tính và chuyên môn về miền. Nó liên quan đến việc thu thập và lưu trữ dữ liệu từ nhiều nguồn, bao gồm cơ sở dữ liệu có cấu trúc, tài liệu văn bản phi cấu trúc, dữ liệu cảm biến, phương tiện truyền thông xã hội, v.v. Với sự ra đời của dữ liệu lớn, các tổ chức hiện có quyền truy cập vào khối lượng thông tin khổng lồ, tạo ra cả cơ hội và thách thức trong việc phân tích và trích xuất những hiểu biết có thể hành động.

Phân tích dữ liệu có nhiều ứng dụng trong các ngành và lĩnh vực. Trong kinh doanh, nó giúp các tổ chức cải thiện hiệu quả hoạt động, tối ưu hóa các chiến lược tiếp thị, xác định sở thích của khách hàng và phát hiện gian lận hoặc bất thường. Trong chăm sóc sức khỏe, phân tích dữ liệu hỗ trợ theo dõi bệnh nhân, dự đoán bệnh và tối ưu hóa điều trị. Về tài chính, nó hỗ trợ quản lý rủi ro, phân tích đầu tư và phát hiện gian lận. Ngoài ra, các chính phủ, đội thể thao, tổ chức phi lợi nhuận và nhiều lĩnh vực khác được hưởng lợi từ phân tích dữ liệu để thúc đẩy quá trình ra quyết định và đạt được lợi thế cạnh tranh.

Để thực hiện phân tích dữ liệu một cách hiệu quả, các chuyên gia sử dụng kết hợp các công cụ và công nghệ. Chúng bao gồm các ngôn ngữ lập trình như Python và

R, gói phần mềm thống kê, công cụ trực quan hóa dữ liệu và hệ thống quản lý cơ sở dữ liệu. Ngoài ra, các kỹ thuật máy học và trí tuệ nhân tạo được sử dụng để tự động hóa các quy trình phân tích dữ liệu và trích xuất thông tin chuyên sâu từ các bộ dữ liệu phức tạp.

Cân nhắc về đạo đức và quyền riêng tư dữ liệu là các khía cạnh không thể thiếu của phân tích dữ liệu. Vì phân tích dữ liệu liên quan đến việc xử lý thông tin nhạy cảm, chẳng hạn như dữ liệu cá nhân và tài chính, điều quan trọng là phải đảm bảo tuân thủ các quy định về quyền riêng tư và nguyên tắc đạo đức. Các tổ chức phải ưu tiên bảo mật dữ liệu, áp dụng các thực hành dữ liệu có trách nhiệm và đảm bảo tính minh bạch trong việc thu thập, phân tích và sử dụng dữ liệu.

1.8.1 Các phương pháp phân tích dữ liệu

Các phương pháp nghiên cứu bao gồm nhiều cách tiếp cận khác nhau, mỗi cách tiếp cận phù hợp với tính chất cụ thể của nghiên cứu và loại dữ liệu được thu thập. Những phương pháp này có thể được phân loại thành nghiên cứu định tính và định lượng. ***Nghiên cứu định tính*** tập trung vào việc hiểu và diễn giải kinh nghiệm, hành vi và hiện tượng xã hội của con người thông qua các phương pháp như phỏng vấn, nhóm tập trung, quan sát và phân tích tài liệu. Nó nhằm mục đích khám phá các quan điểm chủ quan, khám phá ý nghĩa cơ bản và tạo ra dữ liệu mô tả phong phú. Mặt khác, ***nghiên cứu định lượng*** liên quan đến việc thu thập và phân tích dữ liệu số thông qua khảo sát, thí nghiệm, phân tích thống kê và mô hình toán học. Nó nhằm mục đích thiết lập các mẫu, mối quan hệ và kết luận có thể khái quát hóa bằng cách sử dụng các kỹ thuật thống kê và đo lường nghiêm ngặt.

1.8.2 Các phương pháp thu thập dữ liệu

Quá trình thu thập dữ liệu thu thập thông tin có hệ thống phù hợp với mục tiêu nghiên cứu. Điều cần thiết là đảm bảo rằng dữ liệu được thu thập là chính xác, đáng tin cậy và đại diện cho dân số hoặc hiện tượng đang được điều tra. Dữ liệu có thể được thu thập thông qua các phương pháp khác nhau, bao gồm khảo sát, phỏng vấn, quan sát, thí nghiệm, cơ sở dữ liệu hiện có và hồ sơ lưu trữ. Mỗi phương pháp đều có những điểm mạnh và hạn chế, và chúng ta phải lựa chọn cẩn thận phương pháp thích hợp nhất dựa trên các câu hỏi nghiên cứu, các tài nguyên sẵn có và các cân nhắc về đạo đức.

Khảo sát: Phương pháp khảo sát là quá trình thu thập dữ liệu bằng cách đặt câu hỏi hoặc yêu cầu thông tin từ một mẫu dân số hoặc một nhóm người cụ thể. Các câu hỏi có thể được đặt trong một cuộc khảo sát trực tiếp, điện thoại, trực tuyến hoặc bằng cách gửi phiếu điền thông tin. Phương pháp này thường được sử dụng để thu thập dữ liệu về ý kiến, hành vi, hoặc đặc điểm của một nhóm người cụ thể.

Phỏng vấn: Phương pháp phỏng vấn bao gồm việc tiếp xúc trực tiếp với các cá nhân hoặc nhóm người để thu thập thông tin. Có thể có các loại phỏng vấn khác nhau, bao gồm phỏng vấn cá nhân, phỏng vấn nhóm, phỏng vấn cấu trúc hoặc phỏng vấn

không cấu trúc. Phương pháp này cho phép nghiên cứu viên thu thập thông tin sâu hơn về các khía cạnh như ý kiến, kinh nghiệm và suy nghĩ của người được phỏng vấn.

Quan sát: Phương pháp quan sát liên quan đến việc theo dõi và ghi lại thông tin về hành vi, sự kiện hoặc hiện tượng trong một môi trường cụ thể. Quan sát có thể được thực hiện theo hai hình thức: quan sát chủ quan, khi người quan sát tham gia vào tình huống, và quan sát khách quan, khi người quan sát không tác động đến môi trường nghiên cứu. Phương pháp này thường được sử dụng để thu thập dữ liệu về hành vi tự nhiên và tương tác xã hội.

Thí nghiệm: Phương pháp thí nghiệm là quá trình nghiên cứu có mục đích xác định mối quan hệ nguyên nhân và kết quả bằng cách kiểm tra các giả định hoặc giả định trong một môi trường kiểm soát. Trong một thí nghiệm, các biến được kiểm soát và giả định được thử nghiệm để xem liệu chúng có ảnh hưởng đến kết quả hay không. Các phương pháp thí nghiệm cung cấp một cách mạnh mẽ để xác định mối quan hệ nguyên nhân và hiệu ứng giữa các biến.

Cơ sở dữ liệu hiện có: Phương pháp này liên quan đến việc sử dụng dữ liệu đã được thu thập trước đó và được lưu trữ trong cơ sở dữ liệu. Các cơ sở dữ liệu có thể bao gồm thông tin từ các nguồn như báo cáo thống kê, nghiên cứu trước đây, dữ liệu trực tuyến hoặc hồ sơ bệnh án. Bằng cách sử dụng cơ sở dữ liệu hiện có, nghiên cứu viên có thể phân tích và tìm hiểu về các vấn đề và mẫu lặp lại đã được ghi nhận trước đó.

Hồ sơ lưu trữ: Phương pháp này liên quan đến việc thu thập dữ liệu từ các hồ sơ, tài liệu hoặc hồ sơ lưu trữ có sẵn. Đây có thể là các tài liệu lịch sử, bản ghi, báo cáo, hồ sơ bệnh án, văn bản pháp lý hoặc bất kỳ nguồn thông tin chứa dữ liệu quan trọng nào khác. Phương pháp này cho phép nghiên cứu viên tiếp cận dữ liệu đã tồn tại để tìm hiểu thông tin lịch sử, xu hướng hoặc mô hình.

1.9 Trực quan dữ liệu

1.9.1 Định nghĩa về trực quan hóa dữ liệu

Trực quan hóa dữ liệu là quá trình biến đổi dữ liệu phức tạp thành các biểu đồ hình ảnh dễ hiểu và dễ giải thích hơn. Với sự gia tăng vượt bậc của dữ liệu trong những năm gần đây, tầm quan trọng của trực quan hóa dữ liệu trở nên ngày càng rõ ràng. Nó cho phép chúng ta nhanh chóng và dễ dàng hiểu được lượng dữ liệu lớn, xác định các thông tin quan trọng và truyền đạt những phát hiện cho người khác một cách hiệu quả hơn.

Ở mức độ cơ bản, trực quan hóa dữ liệu bao gồm việc sử dụng các yếu tố đồ họa và hình ảnh như biểu đồ, đồ thị và bản đồ để trình bày dữ liệu một cách dễ tiếp cận và dễ hiểu. Các biểu đồ và hình ảnh này có thể được sử dụng để thể hiện các xu hướng, mô hình và mối quan hệ mà không thể thấy ngay trong dữ liệu gốc. Bằng cách trình bày dữ liệu dưới dạng hình ảnh, việc xác định các điểm ngoại lệ, biểu hiện không

đúng và những phát hiện chủ chốt khác trở nên dễ dàng hơn, điều này là rất quan trọng để đưa ra các quyết định có căn cứ.

Trực quan hóa dữ liệu có thể được áp dụng trong nhiều lĩnh vực khác nhau, từ kinh doanh và tài chính đến y tế và khoa học. Trong thế giới kinh doanh, trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu bán hàng, hành vi của khách hàng và các chỉ số quan trọng khác. Nó có thể giúp doanh nghiệp xác định các xu hướng và mô hình có thể ảnh hưởng đến hiệu quả kinh doanh của họ, cũng như đưa ra các quyết định thông minh về tiếp thị, phát triển sản phẩm và các lĩnh vực quan trọng khác. Trong lĩnh vực tài chính, trực quan hóa dữ liệu được sử dụng để theo dõi giá cổ phiếu, giám sát các chỉ số kinh tế và phân tích các chỉ số hiệu suất tài chính. Nó có thể giúp nhà đầu tư đưa ra các quyết định thông minh về việc đầu tư tiền của họ, cũng như xác định các cơ hội phát triển và mở rộng. Trong lĩnh vực y tế, trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu bệnh nhân, theo dõi dịch bệnh và giám sát xu hướng sức khỏe công cộng. Nó có thể giúp các bác sĩ và nhà cung cấp dịch vụ y tế xác định các mô hình và xu hướng trong dữ liệu bệnh nhân, cũng như đưa ra các quyết định thông minh về điều trị và chăm sóc. Trong lĩnh vực khoa học, trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu nghiên cứu, theo dõi các xu hướng khoa học và truyền đạt các phát hiện cho người khác. Nó có thể giúp các nhà khoa học xác định các mô hình và mối quan hệ trong dữ liệu của họ, cũng như truyền đạt các phát hiện của họ đến khán giả rộng hơn.

Có nhiều kỹ thuật trực quan hóa dữ liệu khác nhau, bao gồm biểu đồ, đồ thị, bản đồ, bản đồ nhiệt, biểu đồ phân tán,.... Mỗi loại trực quan hóa có điểm mạnh và điểm yếu riêng, và việc lựa chọn loại trực quan hóa phụ thuộc vào dữ liệu cụ thể đang được phân tích và những phát hiện cần truyền đạt. Ngoài loại trực quan hóa được sử dụng, còn có nhiều công cụ và công nghệ khác nhau dành cho việc trực quan hóa dữ liệu. Chúng có thể từ phần mềm bảng tính đơn giản đến các nền tảng trực quan hóa dữ liệu tiên tiến cho phép phân tích và cộng tác thời gian thực.

1.9.2 Ưu điểm của trực quan hóa dữ liệu

Trực quan hóa dữ liệu là một công cụ mạnh mẽ cho phép chúng ta biểu diễn dữ liệu phức tạp dưới dạng hình ảnh. Đôi mắt của chúng ta tự nhiên được hút bởi màu sắc và mô hình, và chúng ta có thể nhanh chóng nhận ra sự khác biệt giữa chúng. Điều này giải thích tại sao trực quan hóa dữ liệu là một cách hiệu quả để truyền thông tin. Nó cho phép chúng ta nhanh chóng nhận ra các xu hướng, điểm ngoại lệ và mối quan hệ trong lượng lớn dữ liệu

Một trong những lợi ích lớn nhất của trực quan hóa dữ liệu là cho phép chúng ta dễ dàng chia sẻ thông tin. Thay vì phải xem qua hàng trang số liệu và văn bản, chúng ta có thể sử dụng biểu đồ, đồ thị và các biểu đồ trực quan khác để truyền tải thông điệp của mình. Điều này làm cho việc hiểu và giải thích dữ liệu dễ dàng hơn đối với người khác, đặc biệt là khi trình bày cho những đối tượng không chuyên về kỹ thuật (như Hình 1-8).



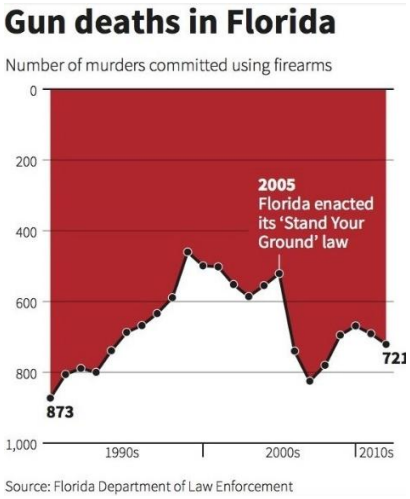
Hình 1-8 Một số loại đồ thị để trực quan dữ liệu

Lợi ích khác của trực quan hóa dữ liệu là cho phép chúng ta tương tác khám phá các cơ hội. Với các công cụ phù hợp, chúng ta có thể thao tác và phân tích dữ liệu trong thời gian thực, điều này vô cùng mạnh mẽ. Chúng ta có thể nhanh chóng xác định các xu hướng, mô hình và điểm ngoại lệ, và sử dụng thông tin này để đưa ra những quyết định có căn cứ hơn. Trực quan hóa dữ liệu cũng cho phép chúng ta hình dung các mô hình và mối quan hệ có thể không thể nhận ra ngay lập tức trong dữ liệu gốc. Bằng cách biểu diễn dữ liệu dưới dạng hình ảnh, chúng ta có thể xác định các mối quan hệ và mô hình tương quan mà nếu không sẽ bị bỏ qua. Điều này có thể vô cùng hữu ích trong các lĩnh vực như tài chính, nơi xác định xu hướng và mô hình có thể là yếu tố quyết định thành công. Có lẽ lợi ích quan trọng nhất của trực quan hóa dữ liệu là cho phép chúng ta kể câu chuyện với mục đích. Bằng cách trình bày dữ liệu dưới dạng hình ảnh, chúng ta có thể tạo ra một cốt truyện giúp người khác hiểu và giải thích dữ liệu. Điều này đặc biệt hữu ích khi trình bày cho các bên liên quan, nơi khả năng truyền đạt giá trị của dữ liệu có thể là sự khác biệt giữa thành công và thất bại. Ngoài những lợi ích này, trực quan hóa dữ liệu cũng mang lại nhiều lợi ích khác. Ví dụ, nó có thể giúp chúng ta xác định lỗi và không nhất quán trong dữ liệu, từ đó dẫn đến phân tích và đưa ra quyết định chính xác hơn. Nó cũng có thể giúp chúng ta xác định các cơ hội phát triển và mở rộng, cũng như các lĩnh vực cần cải tiến.

1.9.3 Nhược điểm của trực quan hóa dữ liệu

Mặc dù trực quan hóa dữ liệu mang nhiều lợi ích, nhưng cũng tồn tại một số nhược điểm không thể nhận thấy ngay lập tức. Một trong những nhược điểm chính của trực quan hóa dữ liệu là khả năng xuất hiện thông tin thiên vị hoặc không chính xác. Mặc dù đồ họa có thể là một công cụ mạnh mẽ để truyền thông, chúng cũng có thể dẫn đến hiểu lầm nếu không được thiết kế và diễn giải cẩn thận. Các biểu đồ và hình ảnh có thể bị điều chỉnh hoặc chênh lệch để trình bày một quan điểm cụ thể, hoặc chúng có thể dựa trên dữ liệu không đầy đủ hoặc không chính xác. Điều này có thể dẫn đến những giả định và kết luận không chính xác, gây hại cho quyết định. Hình 1-9

minh họa một trường hợp sử dụng dữ liệu trực quan để truyền tải thông tin mà dễ gây hiểu nhầm cho người đọc.



Hình 1-9 Số người chết vì súng ở Florida

Các quy ước thông thường sẽ khiến bạn nghĩ rằng số người chết vì súng giảm ở Florida từ năm 2005 đến năm 2012. Nhưng khi quan sát kỹ, bạn sẽ nhận ra rằng trục y bị lộn ngược. Số 0 ở trên cùng và giá trị lớn nhất ở dưới cùng. Điều này vi phạm quy ước đã được thiết lập rõ ràng rằng giá trị y tăng lên khi chúng ta đi lên.

Nhược điểm khác của trực quan hóa dữ liệu là tương quan không luôn đồng nghĩa với nguyên nhân. Việc hai biến có mối quan hệ không nhất thiết có nghĩa là một biến gây ra biến còn lại. Đây là một sự phân biệt quan trọng phải cẩn thận xem xét khi diễn giải các trực quan hóa. Thiếu hiểu biết sâu sắc về dữ liệu cơ bản, rất dễ dẫn đến kết luận sai lầm dựa trên tương quan rõ ràng. Một nhược điểm thứ ba của trực quan hóa dữ liệu là thông điệp cốt lõi có thể bị lạc trong quá trình dịch. Khi sử dụng trực quan hóa để truyền tải thông tin phức tạp, có nguy cơ thông điệp chính bị che giấu hoặc hiểu sai. Điều này có thể xảy ra khi trực quan hóa quá phức tạp hoặc thiết kế kém, làm cho người xem khó hiểu ý nghĩa dự định. Hơn nữa, trực quan hóa có thể bị hiểu sai bởi những người xem thiếu kiến thức nền tảng hoặc kỹ năng phân tích cần thiết để hiểu đúng. Nhược điểm khác của trực quan hóa dữ liệu là việc tạo ra các biểu đồ hiệu quả có thể gặp khó khăn. Tạo ra một biểu đồ hiệu quả đòi hỏi sự kết hợp giữa kỹ năng kỹ thuật, chuyên môn về thiết kế và kiến thức chuyên ngành cụ thể. Điều này có thể là một nhiệm vụ khó khăn, đặc biệt đối với những người không chuyên, có thể không có cơ hội tiếp cận các công cụ hoặc đào tạo cần thiết. Ngay cả đối với những người có kỹ năng tương ứng, việc thiết kế một trực quan hóa hiệu quả để truyền tải thông tin phức tạp mà vẫn hấp dẫn mắt và dễ hiểu là điều khó khăn. Cuối cùng, trực quan hóa dữ liệu có thể tốn thời gian và tài nguyên. Tạo ra các trực quan hóa chất lượng cao đòi hỏi lượng thời gian và công sức đáng kể, đặc biệt khi làm việc với các tập dữ liệu lớn và phức tạp. Ngoài ra, quá trình thiết kế và thử nghiệm trực quan hóa

thường liên quan đến nhiều lần lặp và sửa đổi, điều này có thể tốn thời gian và tốn kém.

1.9.4 Trực quan hóa dữ liệu và dữ liệu lớn

Trong thời đại dữ liệu lớn, trực quan hóa dữ liệu đã trở thành một công cụ ngày càng quan trọng để giải thích hàng nghìn hàng tỷ dòng dữ liệu được tạo ra mỗi ngày. Với lượng dữ liệu khổng lồ này, việc trích xuất thông tin hữu ích và đưa ra quyết định dựa trên dữ liệu trở nên khó khăn. Trực quan hóa dữ liệu giúp kể câu chuyện bằng cách sắp xếp dữ liệu thành một hình thức dễ hiểu hơn, làm nổi bật các xu hướng và điểm ngoại lệ mà không thể thấy ngay trong dữ liệu thô.

Tại bản chất của nó, trực quan hóa dữ liệu là về việc kể câu chuyện. Một trực quan hóa tốt loại bỏ những thông tin không cần thiết từ dữ liệu và làm nổi bật thông tin hữu ích, trình bày một cách hấp dẫn mắt và dễ hiểu. Tuy nhiên, tạo ra những trực quan hóa dữ liệu hiệu quả không đơn giản chỉ là việc trang trí biểu đồ hay thêm một số yếu tố thiết kế đẹp mắt. Điều này đòi hỏi sự cân nhắc tinh tế giữa hình thức và chức năng, với dữ liệu và hình ảnh hỗ trợ nhau để truyền tải một thông điệp ý nghĩa.

Một trong những thách thức của trực quan hóa dữ liệu là tìm sự cân bằng phù hợp giữa sự đơn giản và phức tạp. Một biểu đồ quá đơn giản có thể không thu hút sự chú ý của bất kỳ ai hoặc truyền tải một thông điệp mạnh mẽ. Ngược lại, một trực quan hóa phức tạp quá tải thông tin có thể không truyền tải đúng thông điệp hoặc quá phức tạp cho người xem hiểu. Chìa khóa là tìm sự cân bằng phù hợp giữa sự đơn giản và phức tạp, trình bày dữ liệu một cách hấp dẫn mắt và dễ hiểu, trong khi vẫn truyền tải một thông điệp ý nghĩa.

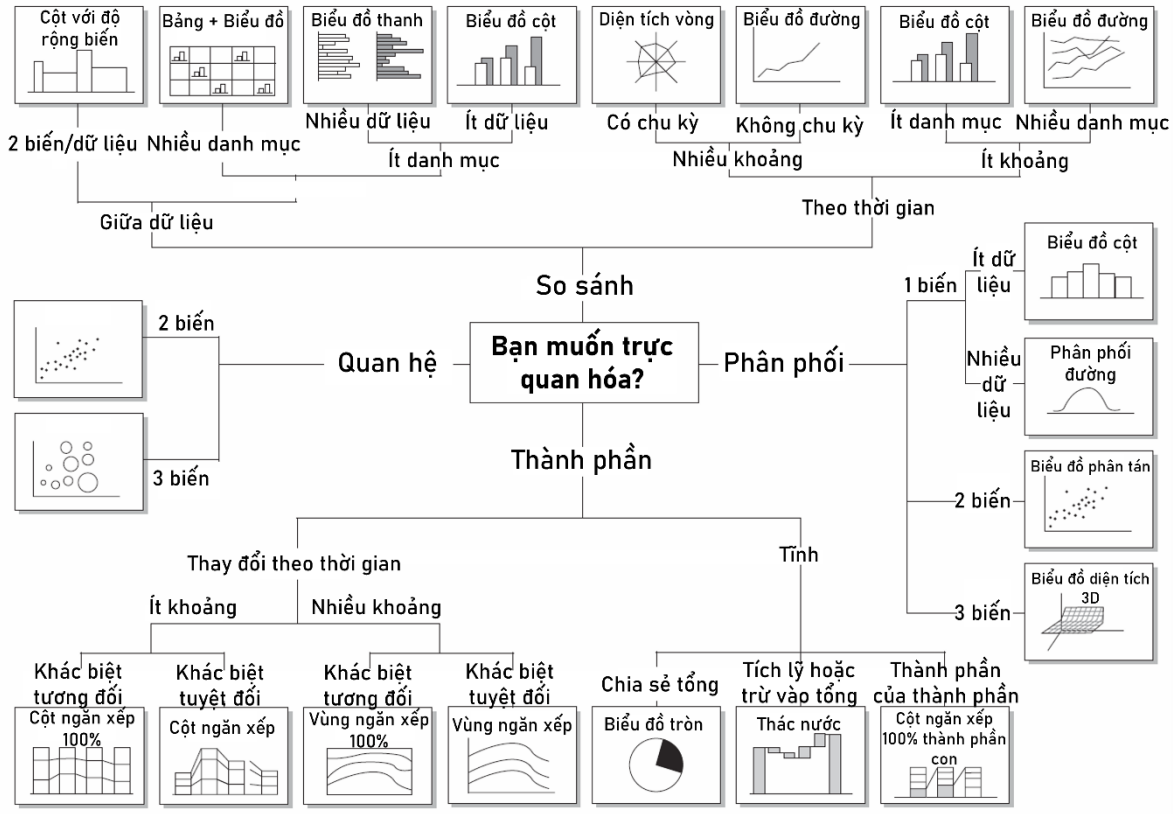
Có nhiều công cụ và phần mềm có sẵn để giúp tạo ra trực quan hóa dữ liệu hiệu quả. Một số công cụ phổ biến nhất bao gồm Microsoft Office, Tableau, Microsoft Power BI và Google Data Studio. Những công cụ này cho phép người dùng tạo các bảng điều khiển, biểu đồ và đồ thị tương tác có thể tùy chỉnh để đáp ứng các yêu cầu cụ thể. Chúng cũng cung cấp một loạt các lựa chọn trực quan hóa dữ liệu, từ biểu đồ cột đơn giản đến các sơ đồ mạng phức tạp.

Tất nhiên, một trong những cách tốt nhất để hiểu về trực quan hóa dữ liệu là thấy nó trong thực tế. Có nhiều bộ sưu tập trực quan hóa dữ liệu công cộng có sẵn trực tuyến, trưng bày những trực quan hóa được tạo ra bởi các nhà khoa học dữ liệu, nhà phân tích và nhà thiết kế. Ví dụ, bộ sưu tập công cộng của Tableau có nhiều loại trực quan hóa, từ bảng điều khiển kinh doanh, bản đồ và các loại trực quan hóa khác, và có thể được sử dụng như một nguồn cảm hứng cho những người muốn tạo ra trực quan hóa của riêng họ.

1.9.5 Những kiểu trực quan hóa dữ liệu

Khi nghĩ về trực quan hóa dữ liệu, chúng ta thường liên tưởng ngay đến biểu đồ cột đơn giản hoặc biểu đồ hình tròn. Mặc dù chúng có vai trò quan trọng trong việc trực quan hóa dữ liệu và được sử dụng rộng rãi trong nhiều ngữ cảnh, nhưng để trực quan hóa dữ liệu một cách đúng đắn, chúng ta cần có bộ thông tin đúng và phù hợp. Các biểu đồ đơn giản chỉ là một phần nhỏ của hình dung toàn bộ khả năng trực quan hóa dữ liệu. Có một loạt các phương pháp trực quan hóa khác nhau để trình bày dữ

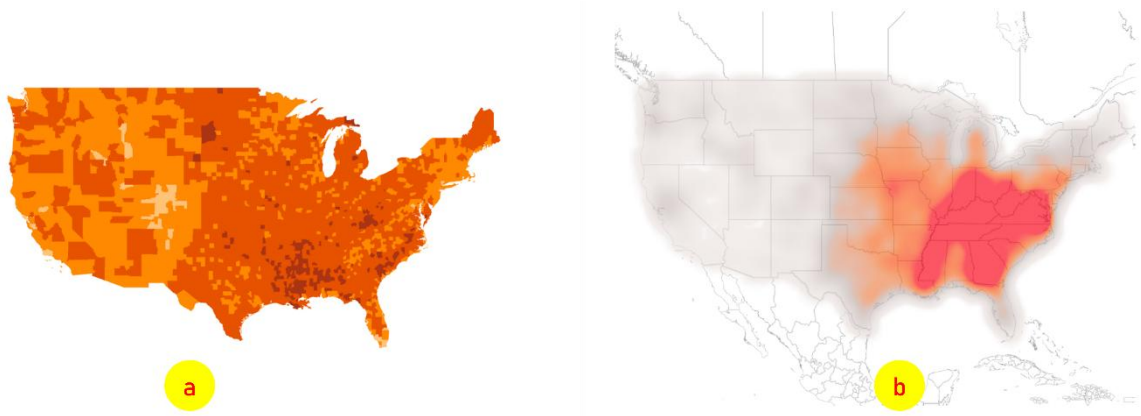
liệu một cách hiệu quả và thú vị (xem Hình 1-10). Dưới đây là một số ví dụ về các loại biểu đồ



Hình 1-10 Sơ đồ lựa chọn cách trực quan dữ liệu

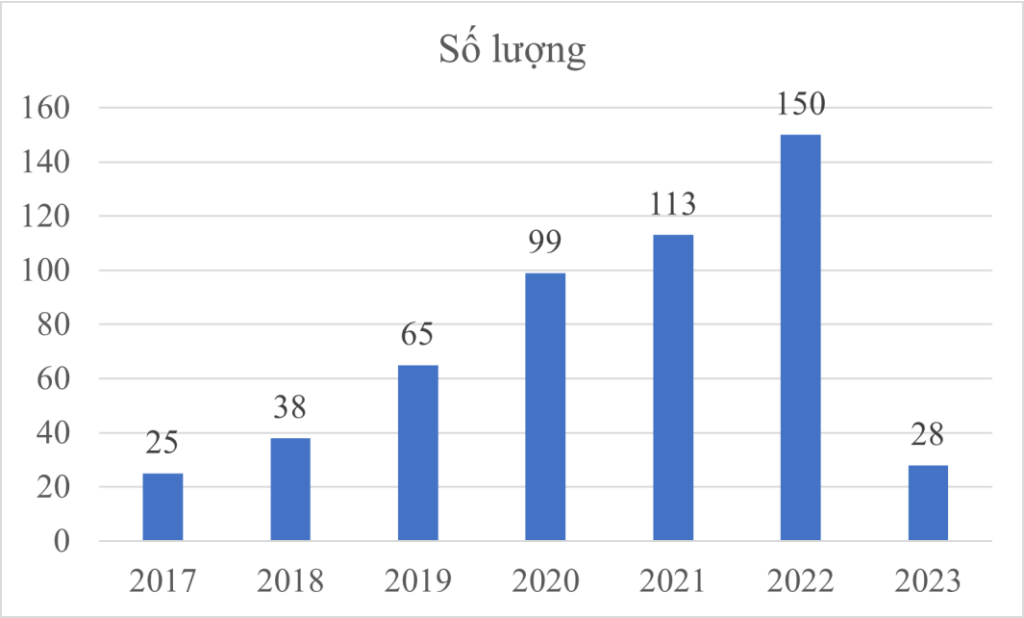
Ví dụ cụ thể hơn:

- **Biểu đồ khu vực (Area map)**: Là một dạng trực quan hóa địa lý, bản đồ khu vực được sử dụng để hiển thị các giá trị cụ thể trên bản đồ của một quốc gia, tiểu bang, hạt, hoặc bất kỳ vị trí địa lý nào khác. Hai loại bản đồ khu vực phổ biến là choropleths và isopleths (xem Hình 1-11).



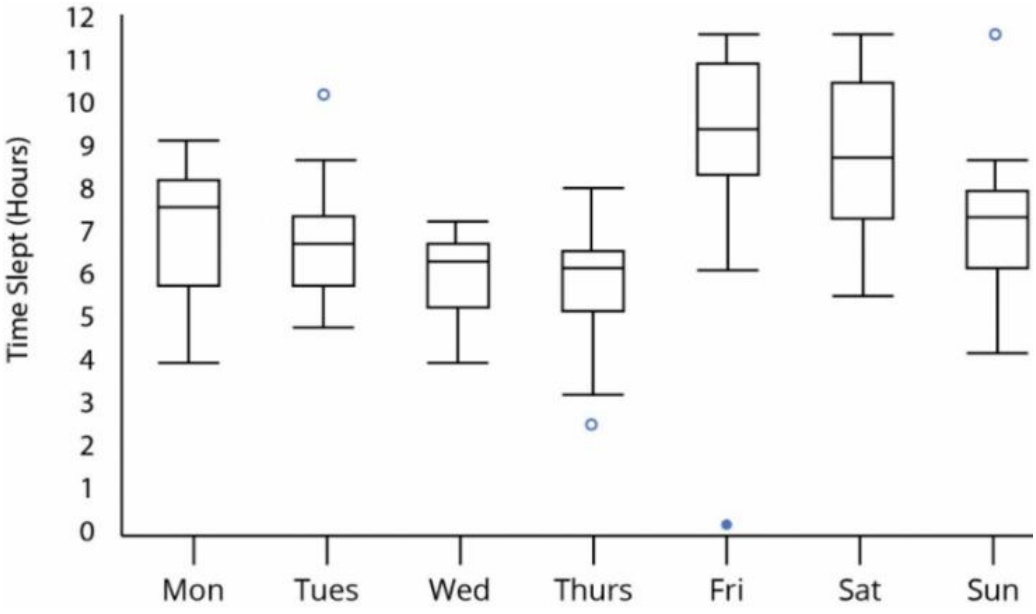
Hình 1-11 Biểu đồ choropleths (a) và isopleths (b)

- **Biểu đồ cột (Bar chart):** Biểu đồ cột đại diện cho các giá trị số liệu so sánh với nhau. Chiều dài của cột đại diện cho giá trị của mỗi biến số (xem Hình 1-12)



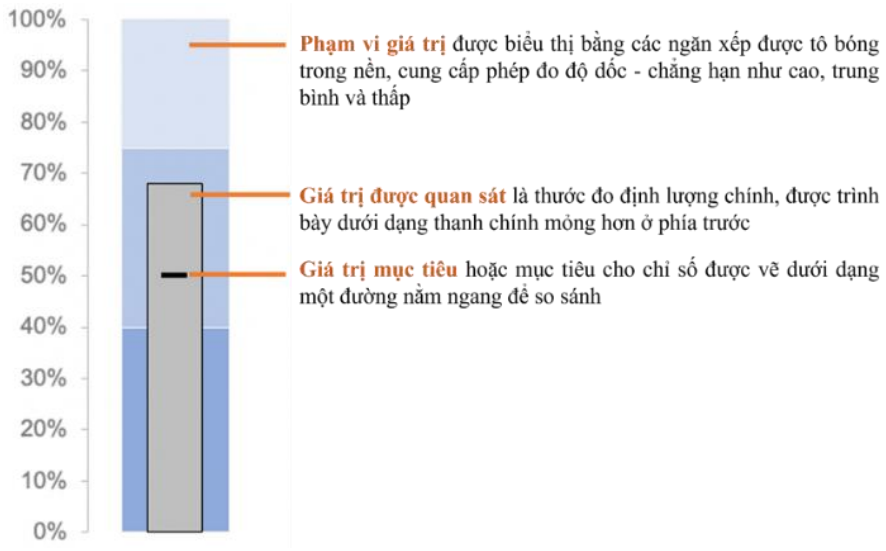
Hình 1-12 Ví dụ về biểu đồ cột

- **Biểu đồ hộp-và-râu (Box-and-whisker plots)**: Biểu đồ này hiển thị một phạm vi giá trị (hộp) qua một số đo lường cụ thể (thanh ngang) như được minh họa trong Hình 1-13



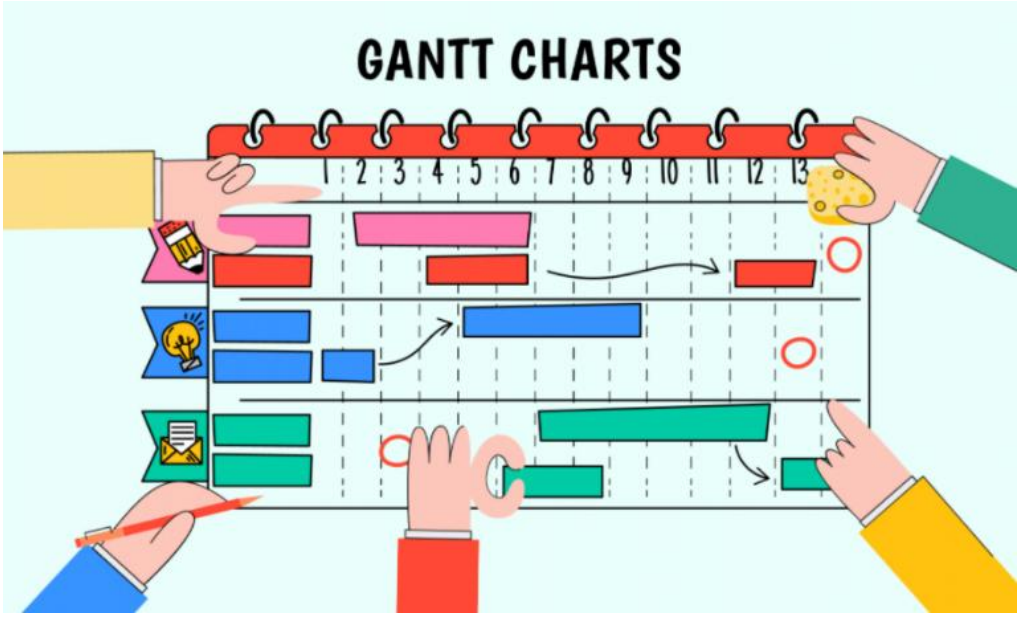
Hình 1-13 Ví dụ về biểu đồ hộp và râu

- **Biểu đồ dấu đầu dòng (Bullet graph):** Là một cột được đánh dấu trên nền để thể hiện tiến trình hoặc hiệu suất so với mục tiêu, được ký hiệu bằng một đường thẳng trên biểu đồ (xem Hình 1-14).



Hình 1-14 Ví dụ về biểu đồ dấu đầu dòng (Bullet chart)

- **Biểu đồ Gantt:** Thường được sử dụng trong quản lý dự án, biểu đồ Gantt là một biểu đồ cột thể hiện dòng thời gian và các nhiệm vụ (xem Hình 1-15).



Hình 1-15 Ví dụ về biểu đồ Gantt

- **Bản đồ nhiệt (Heat map):** Là một loại trực quan hóa địa lý dưới dạng bản đồ, hiển thị các giá trị dữ liệu cụ thể dưới dạng các màu sắc khác nhau (không nhất thiết là nhiệt độ, nhưng đó là một ứng dụng thông thường) như được thể hiện trong Hình 1-16.



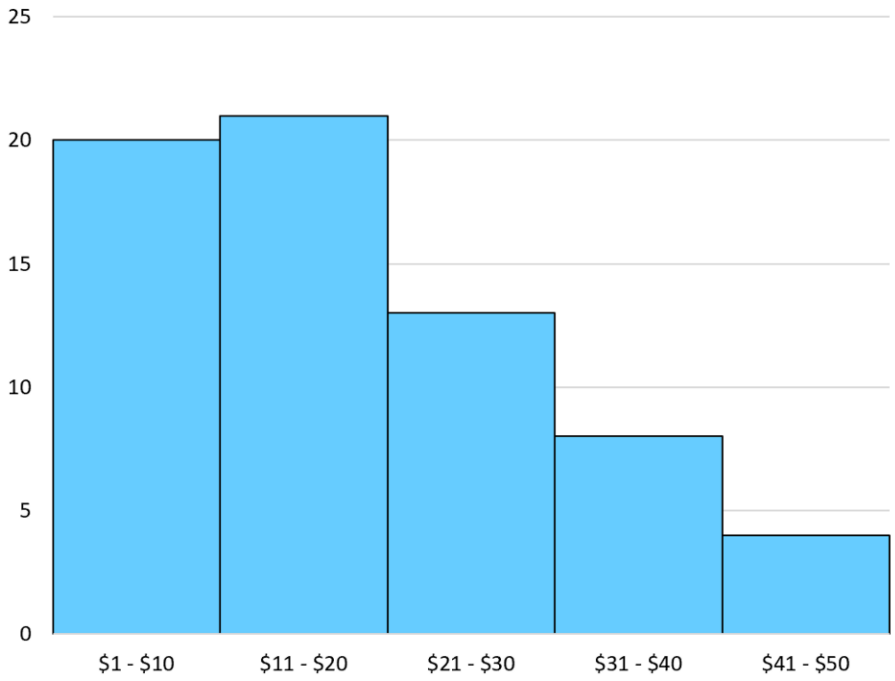
Hình 1-16 Ví dụ về bản đồ nhiệt về thể hiện thời gian vị trí con trỏ chuột trên web

- **Bảng nổi bật (Highlight table):** Là một dạng bảng sử dụng màu sắc để phân loại dữ liệu tương tự, cho phép người xem đọc nó dễ dàng và trực quan hơn (xem Hình 1-17).



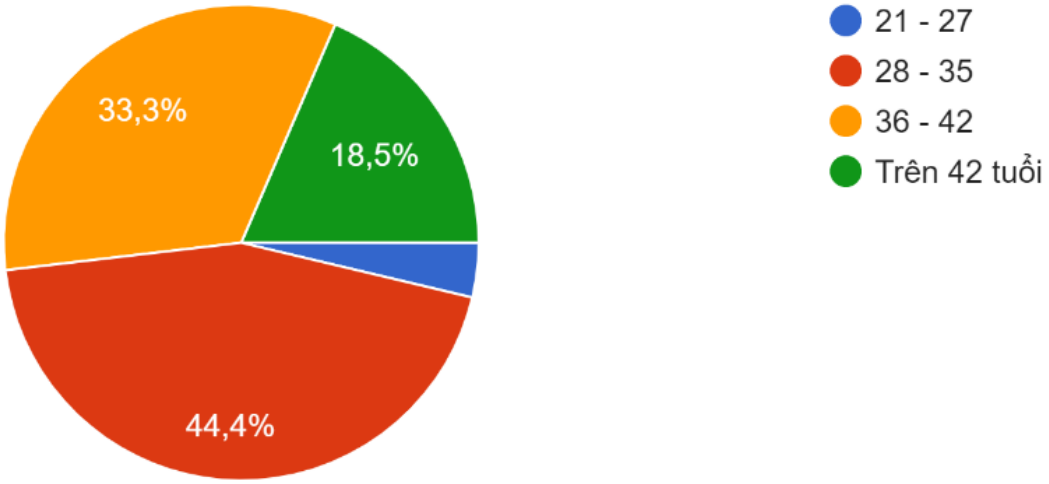
Hình 1-17 Ví dụ về bảng nổi bật

- **Biểu đồ tần suất (Histogram):** Là một loại biểu đồ cột chia các đo lường liên tục thành các nhóm để phân tích phân phối (xem Hình 1-18).



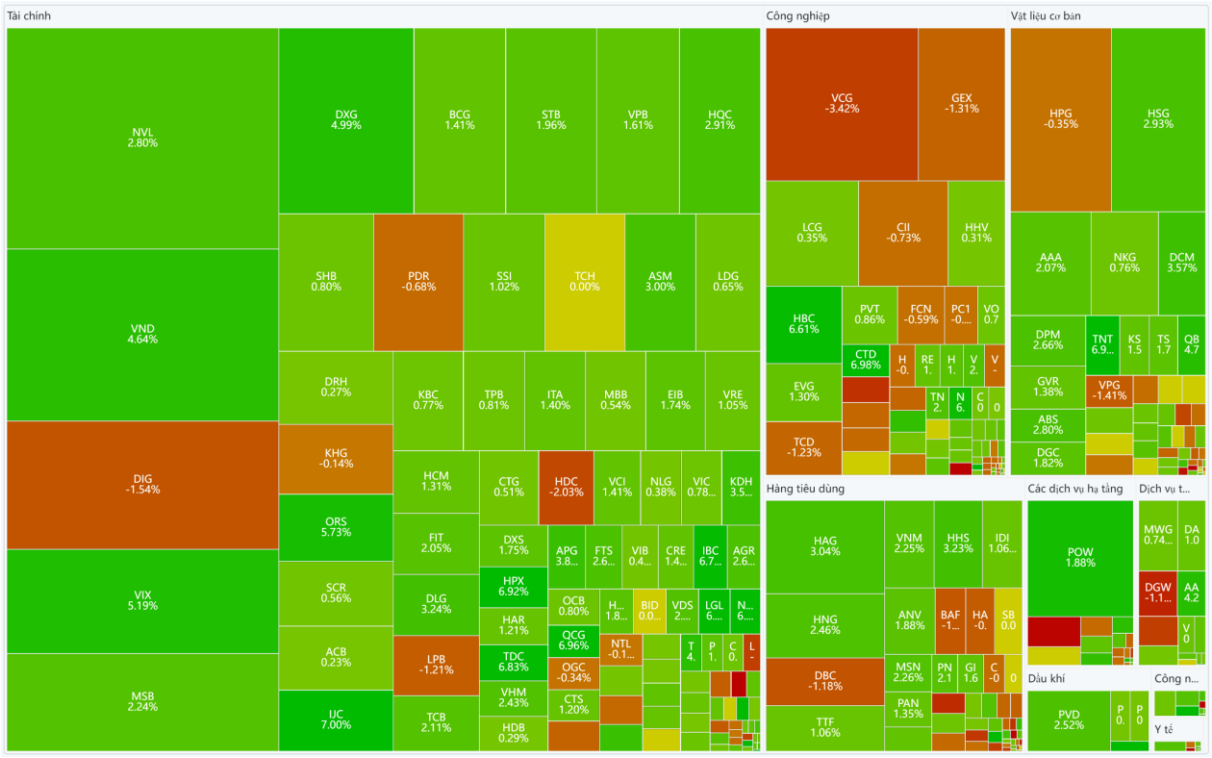
Hình 1-18 Ví dụ về biểu đồ tần suất

- **Biểu đồ hình tròn (Pie chart):** Là một biểu đồ tròn với các phần tam giác thể hiện dữ liệu dưới dạng tỷ lệ phần trăm của tổng thể (xem Hình 1-19).



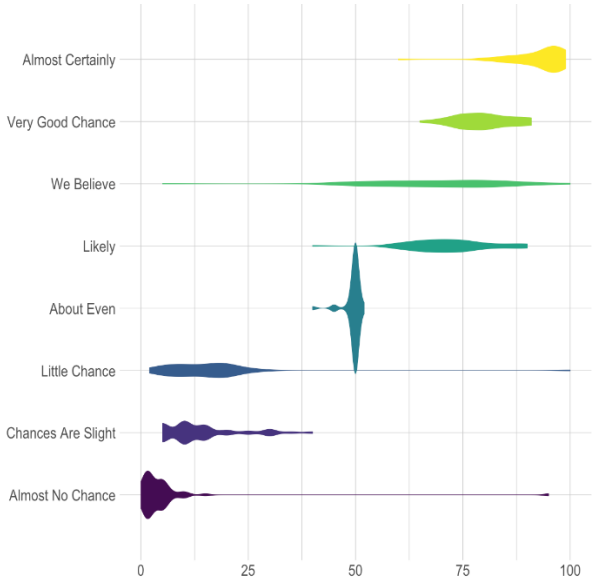
Hình 1-19 Ví dụ về biểu đồ hình tròn

- **Biểu đồ Treemap:** Là một loại biểu đồ hiển thị các giá trị khác nhau, có liên quan nhau dưới dạng các hình chữ nhật lồng nhau (xem Hình 1-20).



Hình 1-20 Ví dụ về biểu đồ tree map

- **Biểu đồ vĩ cầm (violin boxplot):** Biểu đồ vĩ cầm cho phép trực quan hóa phân phối của một biến số cho một hoặc một số nhóm. Mỗi 'vĩ cầm' đại diện cho một nhóm hoặc một biến. Hình dạng đại diện cho ước tính mật độ của biến: càng nhiều điểm dữ liệu trong một phạm vi cụ thể, vĩ cầm càng lớn cho phạm vi đó (xem Hình 1-21).

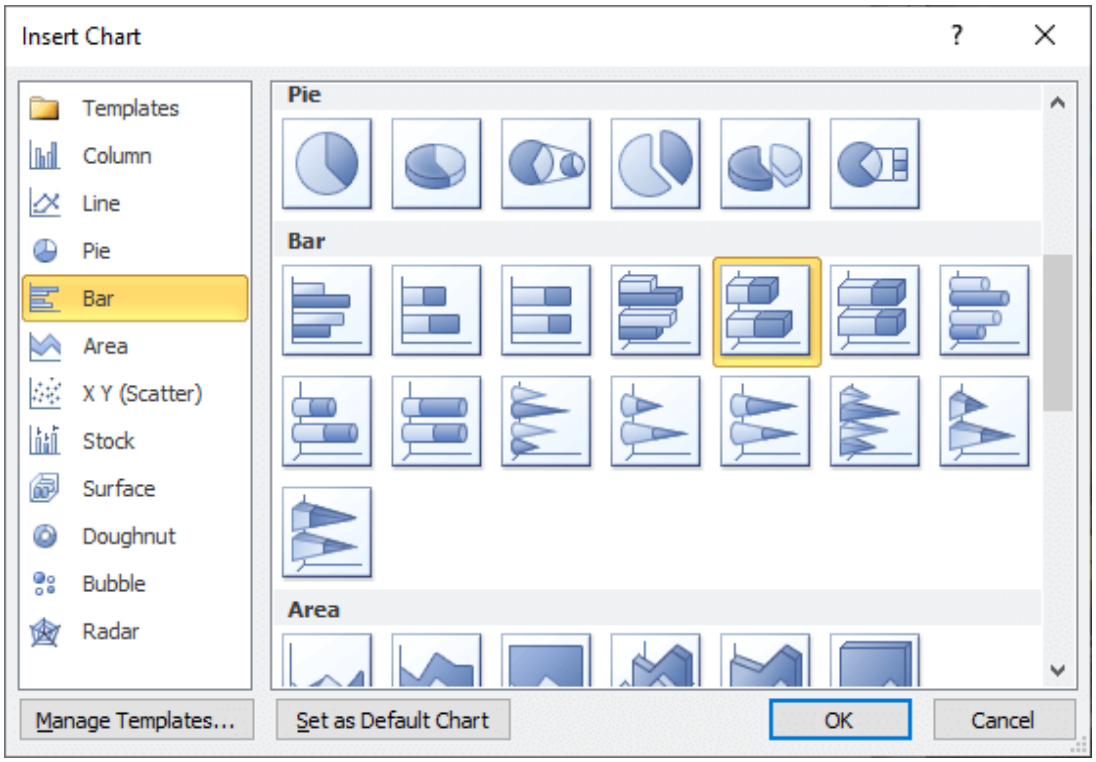


Hình 1-21 Ví dụ biểu đồ vĩ cầm

Các công cụ cho việc phân tích và trực quan dữ liệu

1.10 Công cụ và phần mềm trực quan

Một trong những công cụ đơn giản và dễ sử dụng nhất đó chính là bộ biểu đồ do **Microsoft Office** (xem Hình 1-22) tạo ra thông qua các công cụ như Word, PowerPoint, Excel,...Điểm mạnh của bộ trực quan hoá dữ liệu này là khả năng ***gợi ý*** các biểu đồ phù hợp với từng loại dữ liệu khác nhau.



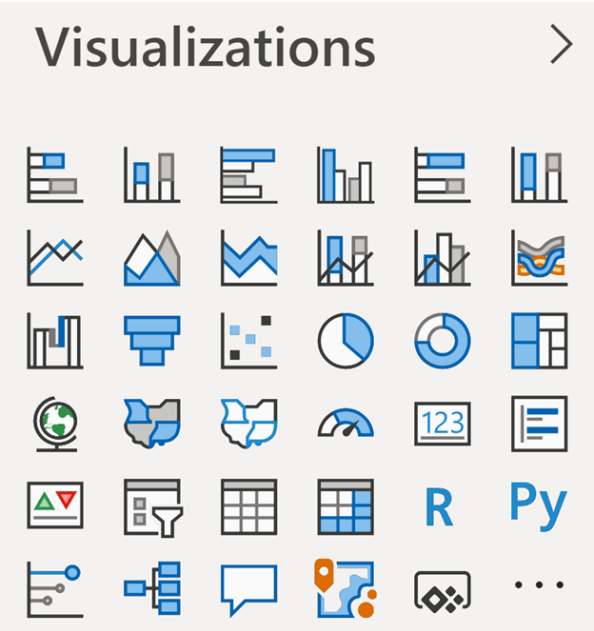
Hình 1-22 Công cụ trực quan hoá dữ liệu của Microsoft Office

Công cụ trực quan hóa mạnh khác là **Tableau**. Tableau là một phần mềm trực quan hóa dữ liệu mạnh mẽ cho phép người dùng tạo ra bảng điều khiển tương tác, biểu đồ và đồ thị có thể tùy chỉnh để đáp ứng các nhu cầu và yêu cầu cụ thể (xem Hình 1-23). Nó cung cấp một loạt các tùy chọn trực quan hóa dữ liệu, từ biểu đồ cột đơn giản và biểu đồ đường đến bản đồ nhiệt phức tạp và biểu đồ mạng. Với Tableau, người dùng có thể dễ dàng kết nối với một loạt các nguồn dữ liệu, bao gồm các bảng tính, cơ sở dữ liệu và các dịch vụ dựa trên đám mây, và tạo ra các trực quan hóa tương tác có thể được chia sẻ trực tuyến hoặc nhúng vào các ứng dụng khác.



Hình 1-23 Phần mềm Tableu cho việc trực quan hóa dữ liệu

Một công cụ trực quan hóa phổ biến khác là **Microsoft Power BI**. Power BI là một dịch vụ phân tích kinh doanh dựa trên đám mây cho phép người dùng tạo ra bảng điều khiển tương tác, báo cáo và trực quan hóa dữ liệu có thể chia sẻ với người khác (xem Hình 1-24). Nó cung cấp một loạt các tùy chọn trực quan hóa dữ liệu, bao gồm biểu đồ cột, biểu đồ đường và biểu đồ phân tán, cũng như các trực quan hóa phức tạp hơn như bản đồ nhiệt và bản đồ choropleth. Power BI cũng tích hợp với một loạt các nguồn dữ liệu, bao gồm các bảng tính Excel, cơ sở dữ liệu SQL Server và các dịch vụ dựa trên đám mây như Salesforce và Google Analytics.



Hình 1-24 Trực quan hóa dữ liệu với Microsoft Power BI

Google Data Studio là một công cụ trực quan hóa phổ biến khác cho phép người dùng tạo ra bảng điều khiển và báo cáo tương tác (xem Hình 1-25). Nó cung cấp một loạt các tùy chọn trực quan hóa dữ liệu, bao gồm biểu đồ cột đơn giản và biểu đồ đường, cũng như bản đồ nhiệt và bản đồ có thể tùy chỉnh với dữ liệu từ Google Maps. Data Studio cũng tích hợp với một loạt các nguồn dữ liệu, bao gồm Google Analytics, Google Sheets và Google Ads.



Hình 1-25 Google Data Studio

Trong lĩnh vực phân tích và trực quan hóa dữ liệu, hai ngôn ngữ lập trình đã nổi lên như một công cụ mạnh mẽ: R và Python. R và Python cung cấp nhiều loại thư viện và gói được thiết kế đặc biệt để thao tác dữ liệu, phân tích thống kê và khám phá trực quan. Những ngôn ngữ này đã trở nên phổ biến trong giới khoa học dữ liệu, nhà nghiên cứu và nhà phân tích nhờ tính linh hoạt, hỗ trợ cộng đồng rộng rãi và hệ sinh thái thư viện phong phú.

R là ngôn ngữ lập trình và môi trường được phát triển đặc biệt cho tính toán thống kê và đồ họa. Nó cung cấp một tập hợp đầy đủ các chức năng và gói cài sẵn cho phép thao tác dữ liệu, mô hình hóa thống kê và biểu diễn trực quan. Sức mạnh của R nằm ở bộ sưu tập phong phú các kỹ thuật thống kê và khả năng trực quan hóa. Nó cung cấp một loạt các gói như **ggplot2**, **dplyr** và **dirtyr**, cho phép người dùng tạo các biểu đồ cung cấp thông tin và hấp dẫn trực quan, thực hiện các chuyển đổi dữ liệu phức tạp và tiến hành phân tích thống kê tinh vi. Cú pháp của R đặc biệt phù hợp với mô hình thống kê và phân tích dữ liệu khám phá, khiến nó trở thành lựa chọn phổ biến của các nhà thống kê và nhà nghiên cứu.

Mặt khác, Python là một ngôn ngữ lập trình có mục đích chung đã đạt được sức hút đáng kể trong lĩnh vực phân tích và trực quan hóa dữ liệu. Tính linh hoạt và dễ sử dụng của Python khiến nó trở thành một lựa chọn hấp dẫn cho nhiều ứng dụng, bao gồm cả phân tích dữ liệu. Python cung cấp nhiều thư viện, chẳng hạn như NumPy, Pandas và Matplotlib, cung cấp các công cụ mạnh mẽ để thao tác dữ liệu, tính toán số và trực quan hóa. Đặc biệt, Pandas là một thư viện rất phổ biến để phân tích dữ liệu, cung cấp các chức năng và cấu trúc dữ liệu hiệu quả để xử lý dữ liệu có cấu trúc.

Matplotlib và Seaborn được sử dụng rộng rãi để tạo nhiều hình ảnh tĩnh và tương tác. Ngoài ra, tính linh hoạt của Python cho phép tích hợp liền mạch với các thư viện và công cụ khác, khiến nó trở thành ngôn ngữ ưa thích cho quy trình phân tích dữ liệu từ đầu đến cuối.

Cả R và Python đều có những điểm mạnh và đặc điểm riêng và việc lựa chọn giữa chúng phụ thuộc vào nhiều yếu tố khác nhau, bao gồm các yêu cầu cụ thể của dự án, sở thích cá nhân và hệ sinh thái hiện có trong một tổ chức. Sự nhấn mạnh của R vào mô hình thống kê và bộ sưu tập phong phú các gói chuyên dụng khiến nó trở thành một lựa chọn tuyệt vời cho các dự án định hướng nghiên cứu và phân tích thống kê. Tính linh hoạt, dễ đọc và khả năng ứng dụng rộng rãi hơn của Python trên các miền khiến nó trở thành lựa chọn phổ biến để phân tích dữ liệu đa năng và tích hợp vào các hệ thống phần mềm lớn hơn.

1.11 Nguồn dữ liệu phục vụ cho việc học tập

Iris

WALMART_SALES_DATA

- **Sự chậm trễ của hãng hàng không** (Airline Delays). Cục Thống kê Vận tải Hoa Kỳ đã thu thập dữ liệu của hơn 169 triệu chuyến bay nội địa kể từ tháng 10 năm 1987. Các nhà nghiên cứu đã phát triển gói hàng không (**airlines** package) để cho phép người dùng R tải xuống và xử lý những dữ liệu này với ít rắc rối nhất. Những dữ liệu này ban đầu được sử dụng cho ASA Data Expo 2009. Gói *nycflights13* chứa một tập hợp con các dữ liệu này (chỉ các chuyến bay rời ba sân bay nổi bật nhất của Thành phố New York trong năm 2013). https://www.rdocumentation.org/packages/nycflights13/versions/1.0.1
- **Tên của em bé** (Baby names). Gói **babynames** cho R cung cấp dữ liệu về mức độ phổ biến của từng tên em bé từ Cơ quan An sinh Xã hội Hoa Kỳ. Ví dụ: những dữ liệu này có thể được sử dụng để theo dõi mức độ phổ biến của một số tên nhất định theo thời gian. https://www.rdocumentation.org/packages/babynames/versions/1.0.0
- **Iris**, được Ronald Fisher giới thiệu trong bài báo năm 1936 của ông. Việc sử dụng nhiều phép đo trong các vấn đề phân loại, bao gồm ba loài thực vật (setosa, virginica, versicolor) và bốn đặc điểm được đo cho mỗi mẫu. https://rpubs.com/moeransm/intro-iris
- **MacLeish**. Trạm thực địa Ada và Archibald MacLeish là một khu đất rộng 260 mẫu Anh do Cao đẳng Smith sở hữu và điều hành. Nó được sử dụng bởi các giảng viên, sinh viên và các thành viên của cộng đồng địa phương để nghiên cứu môi trường, các hoạt động ngoài trời và giải trí. Gói **macleish** R cho phép bạn tải xuống và xử lý dữ liệu thời tiết (dưới dạng chuỗi thời gian) từ Trạm thực địa MacLeish bằng khung **etl**. Nó cũng chứa các tệp hình dạng để bối cảnh hóa thông tin không gian. https://www.rdocumentation.org/packages/macleish/versions/0.3.9

- **MtCars**: Dữ liệu được lấy từ tạp chí Motor Trend Mỹ năm 1974 và bao gồm thông tin về tiêu thụ nhiên liệu cùng 10 yếu tố về thiết kế và hiệu suất của 32 loại xe (mô hình từ năm 1973 đến 1974).
- **Giá nhà ở Boston (Boston Housing):** Dữ liệu nhà ở Boston là tập dữ liệu trong gói MASS. Tập dữ liệu có 506 hàng và 14 cột. https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset
- Chất lượng không ký (Air quality): Các phép đo chất lượng không khí hàng ngày ở New York, từ tháng 5 đến tháng 9 năm 1973. Khung dữ liệu với 153 quan sát trên 6 biến số. https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/airquality.
- **Titanic**: Bộ dữ liệu titanic có sẵn ở cơ sở R. Dữ liệu có 5 biến và chỉ có 32 hàng. https://cran.r-project.org/web/packages/explore/vignettes/explore_titanic.html
- **Trọng lượng gà con (Chickweights)**: Trọng lượng so với tuổi của gà con ở các khẩu phần ăn khác nhau. Khung dữ liệu ChickWeight có 578 hàng và 4 cột từ một thí nghiệm về ảnh hưởng của khẩu phần ăn đến sự phát triển ban đầu của gà con. https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/ChickWeight
- **Dữ liệu bán hàng của Walmart (Walmart Sales Data):** Walmart tổ chức một số sự kiện giảm giá khuyến mại trong suốt cả năm. Những đợt giảm giá này xảy ra trước các ngày lễ nổi bật, bốn ngày lễ lớn nhất trong số đó là Super Bowl, Ngày lễ Lao động, Lễ tạ ơn và Giáng sinh. Những tuần bao gồm những ngày nghỉ lễ này được đánh giá cao hơn năm lần so với những tuần không nghỉ lễ. Một phần thách thức mà cuộc thi này đưa ra là mô hình hóa tác động của việc giảm giá trong những tuần nghỉ lễ này khi không có dữ liệu lịch sử hoàn chỉnh/lý tưởng. Dữ liệu bán hàng lịch sử của 45 cửa hàng Walmart ở các khu vực khác nhau. https://www.kaggle.com/datasets/yasserh/walmart-dataset

Tổng kết chương 1

Trong chương này, chúng ta đã khám phá về trực quan hóa dữ liệu và vai trò quan trọng của nó trong thế giới dữ liệu lớn. Trực quan hóa dữ liệu giúp biến thông tin phức tạp thành các biểu đồ, đồ thị và bản đồ dễ hiểu và tiện lợi. Nó cung cấp các công cụ trực quan để nhìn thấy và hiểu xu hướng, điểm ngoại lệ và mẫu trong dữ liệu. Điều này giúp các nhân viên và doanh nghiệp trình bày dữ liệu một cách dễ dàng và thông qua đối tượng không chuyên về kỹ thuật. Trong tương lai, vai trò của trực quan hóa dữ liệu sẽ ngày càng trở nên quan trọng khi dữ liệu tiếp tục phát triển và trở thành một phần không thể thiếu của hoạt động thông tin và phân tích.

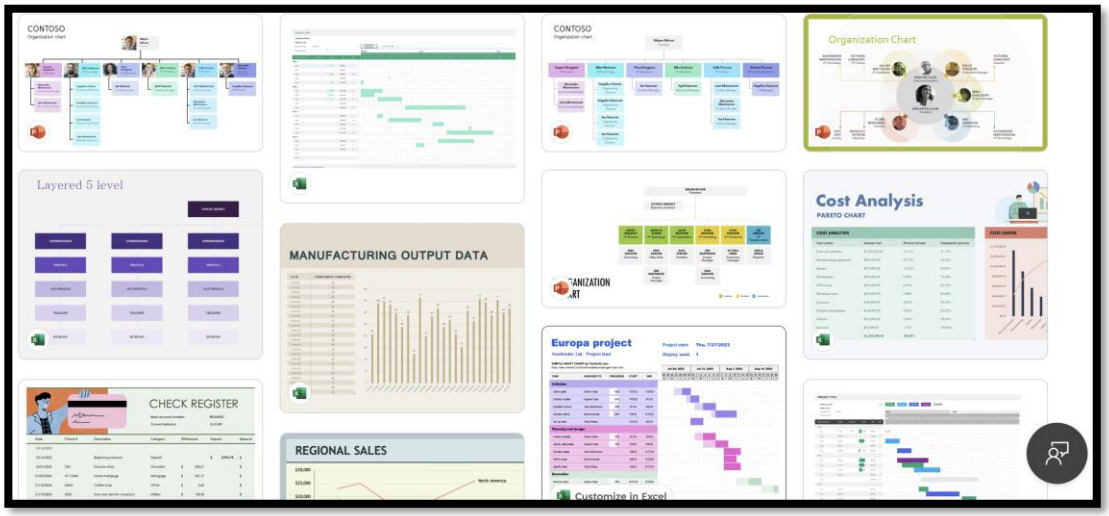
Câu hỏi ôn tập

1. Trực quan hóa dữ liệu là gì và nó giúp hiểu thông tin phức tạp như thế nào?
2. Làm cách nào để các công cụ trực quan hóa dữ liệu có thể trình bày thông tin một cách trực quan hấp dẫn và dễ hiểu?
3. Những lợi thế của trực quan hóa dữ liệu là gì? Đưa ra ví dụ về cách nó có thể hữu ích trong các ngành công nghiệp khác nhau.
4. Thảo luận về một số nhược điểm của trực quan hóa dữ liệu và cách tránh chúng.
5. Tại sao trực quan hóa dữ liệu lại quan trọng trong thời đại Dữ liệu lớn? Nó giúp ích như thế nào trong việc đưa ra quyết định dựa trên dữ liệu?
6. Giải thích khái niệm về nhà khoa học dữ liệu công dân và cách trực quan hóa dữ liệu đóng vai trò quan trọng trong bối cảnh này.
7. Làm thế nào để trực quan hóa dữ liệu giúp quản lý và trình bày dữ liệu ở dạng dễ hiểu và dễ phân tích hơn?
8. Cung cấp các ví dụ về các loại trực quan hóa khác nhau và giải thích cách chúng có thể được sử dụng để biểu thị dữ liệu một cách hiệu quả.
9. Một số công cụ và phần mềm trực quan hóa dữ liệu phổ biến hiện có?
10. Tại sao điều quan trọng đối với các chuyên gia là phát triển kỹ năng trực quan hóa dữ liệu trong thế giới dựa trên dữ liệu ngày nay? Làm thế nào các kỹ năng trực quan hóa dữ liệu có thể có lợi trong các ngành và nghề khác nhau?

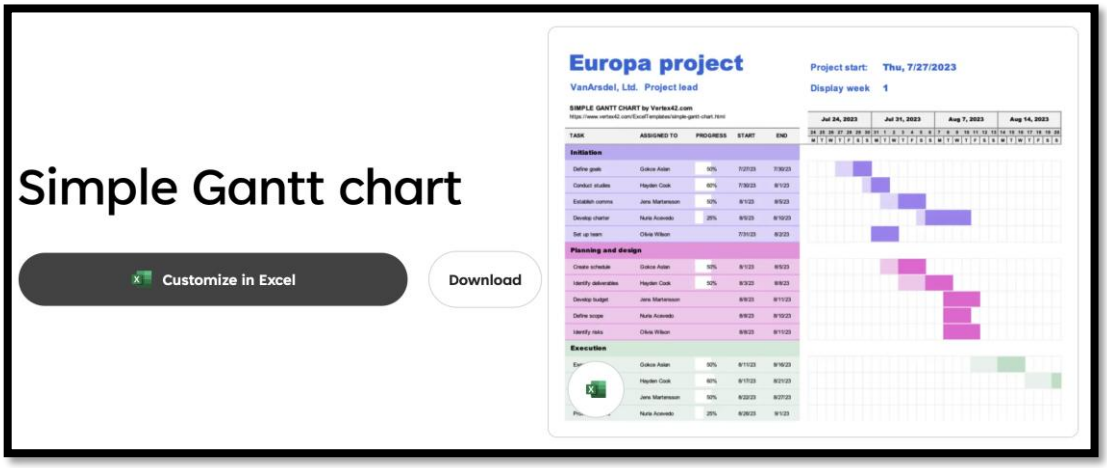
Bài tập thực hành

Bài 1: Tạo biểu đồ nhanh từ các bản mẫu của Microsoft

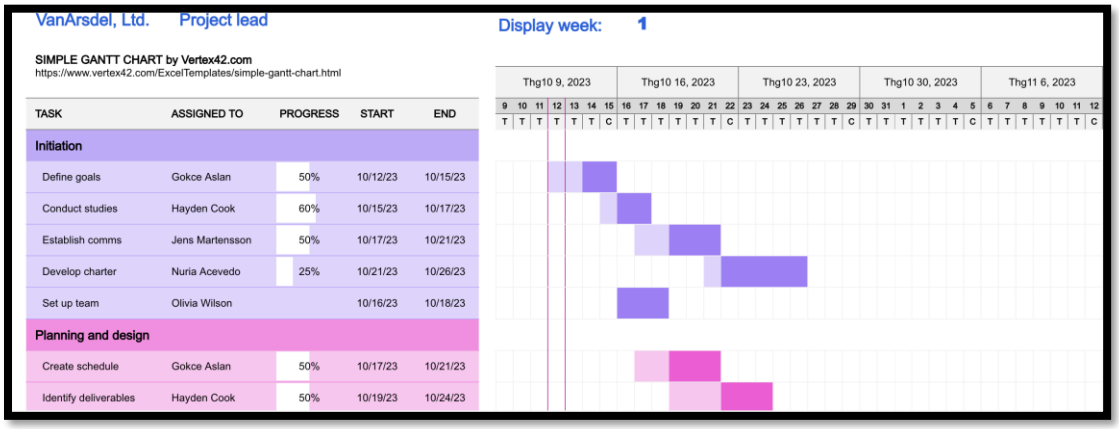
- Bước 1: Truy cập vào trang web: https://create.microsoft.com/en-us/search?query=Charts
- Bước 2: Chọn một biểu đồ bất kỳ (ví dụ Europa project)



- Bước 3: Chọn Customize in Excel

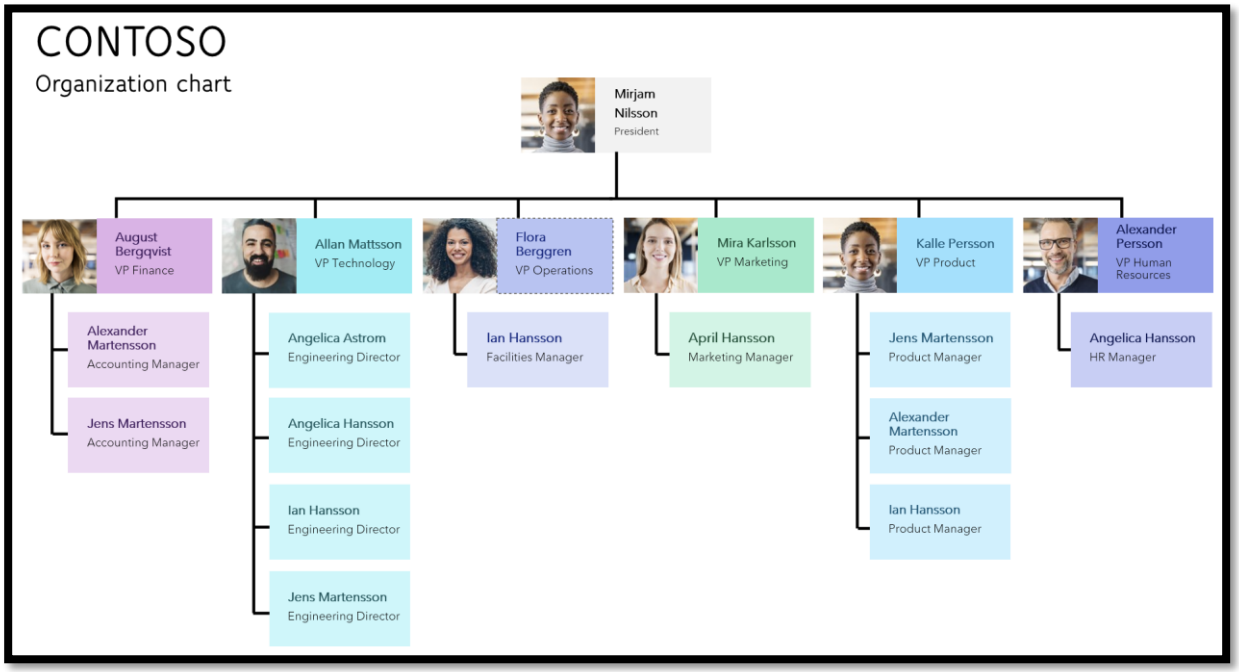


- Bước 4: Thay đổi dữ liệu ở bảng bên trái (bất kỳ dữ liệu nào bao gồm: initiation - khởi tạo, assigned to – người thực hiện, progress - tiến độ,…) để xem biểu đồ thay đổi ở cột bên phải



Bài 2: Tạo biểu đồ nhanh từ các bản mẫu của Microsoft

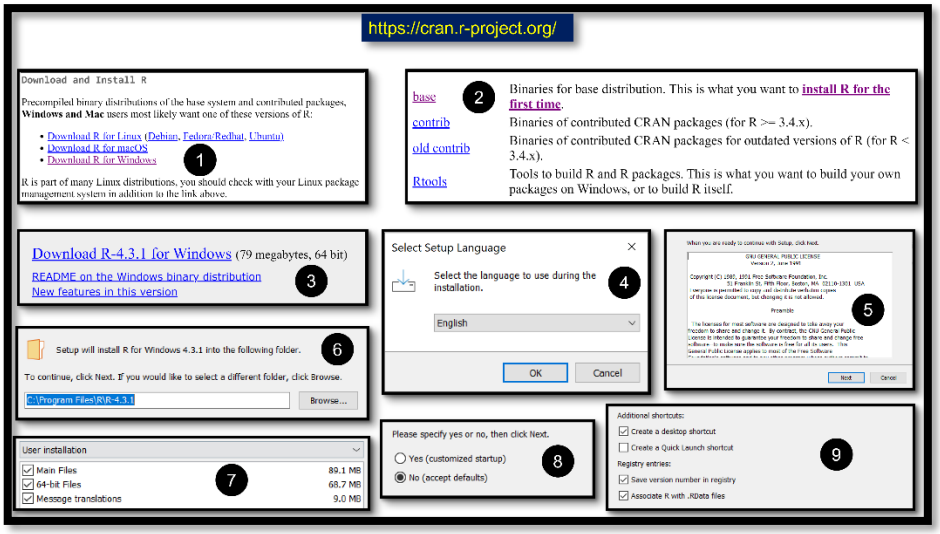
- Bước 1: Truy cập vào trang web: https://create.microsoft.com/en-us/search?query=Charts
- Bước 2: Chọn một biểu đồ Contoso Organization Chart
- Bước 3: Chọn Customize in PowerPoint
- Bước 4: Thay thế nội dung, hình ảnh trong bản mẫu bằng sơ đồ của ICTU theo cấu trúc sau



Bài 3: Cài đặt R và RStudio

Cài đặt R (https://cran.r-project.org/) rồi thực hiện theo các bước sau (xem Hình 1. 1)

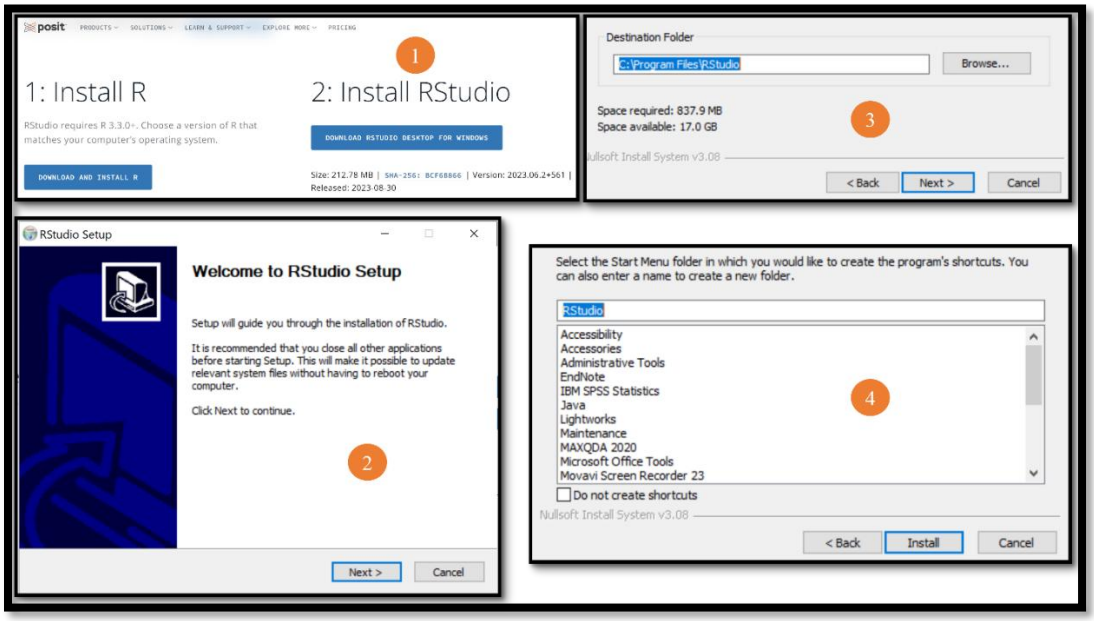
- Bước 1: Chọn “Download R for XXX” với XXX là phiên bản hệ điều hành đang sử dụng tương ứng (ví dụ: Ubuntu, macOS, hoặc Windows)
- Bước 2: Chọn "install R for the first time" (cài đặt R lần đầu tiên) để tải xuống tệp thực thi R (.exe) hoặc bấm vào “base”.
- Bước 3: Chọn “Dowload R-4.3.1 for Windows” với 4.3.1 là phiên bản (có thể khác khi đọc tài liệu này trong tương lai)
- Bước 4: Chọn ngôn ngữ cài đặt “English” (tiếng Anh)
- Bước 5: Đọc điều khoản sử dụng và chọn Next
- Bước 6: Chọn đường dẫn để cài đặt R
- Bước 7: Chọn các files để cài đặt (để mặc định)
- Bước 8: Chọn chi tiết các thành phần cài đặt (để mặc định)
- Bước 9: Tạo shortcut ngoài màn hình nền, đăng ký vào Registry và tự mở các file liên quan đến R (để mặc định). Chọn Next để tiến hành cài đặt



Hình 1. 1 Các bước thực hiện cài đặt R

Cài đặt RStudio (https://posit.co/download/rstudio-desktop) rồi thực hiện theo các bước sau (xem Hình 1.2). Lưu ý cài đặt R trước, sau đó mới cài RStudio

- Bước 1: Chọn "DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS"
- Bước 2: Chọn Next trong cửa sổ chào mừng cài đặt
- Bước 3: Chọn thư mục cài đặt (để mặc định)
- Bước 4: Chọn tên (để mặc định) và sau đó chọn "Install" để cài đặt



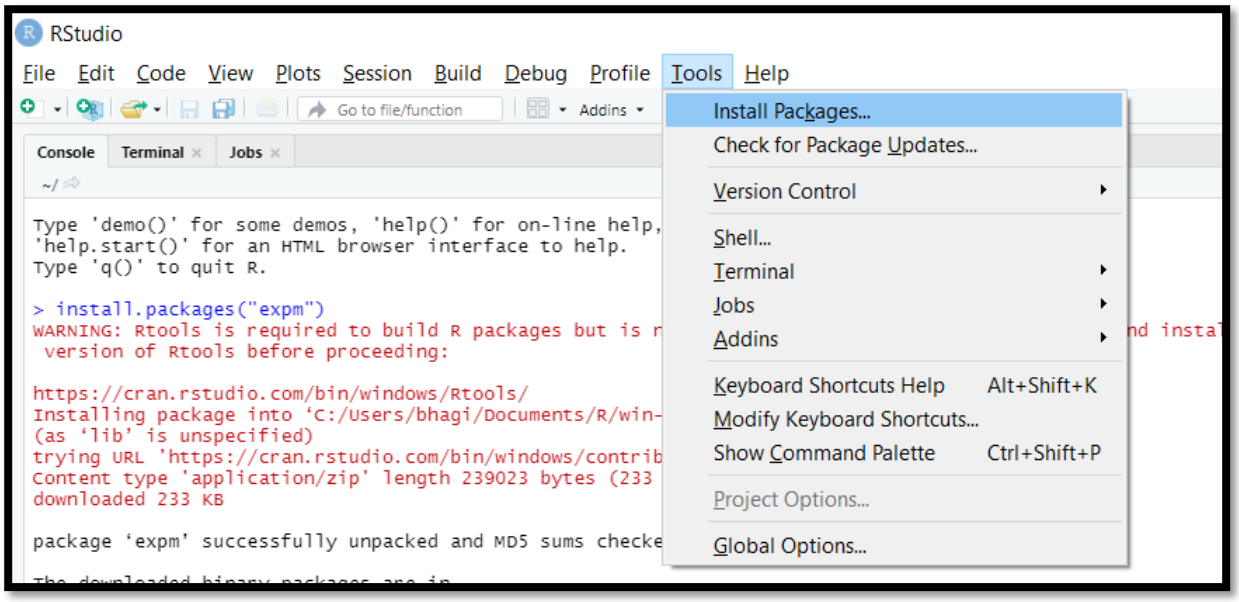
Hình 1.2 Cài đặt RStudio

Bài 4: Cài đặt thư viện trong Rstudio

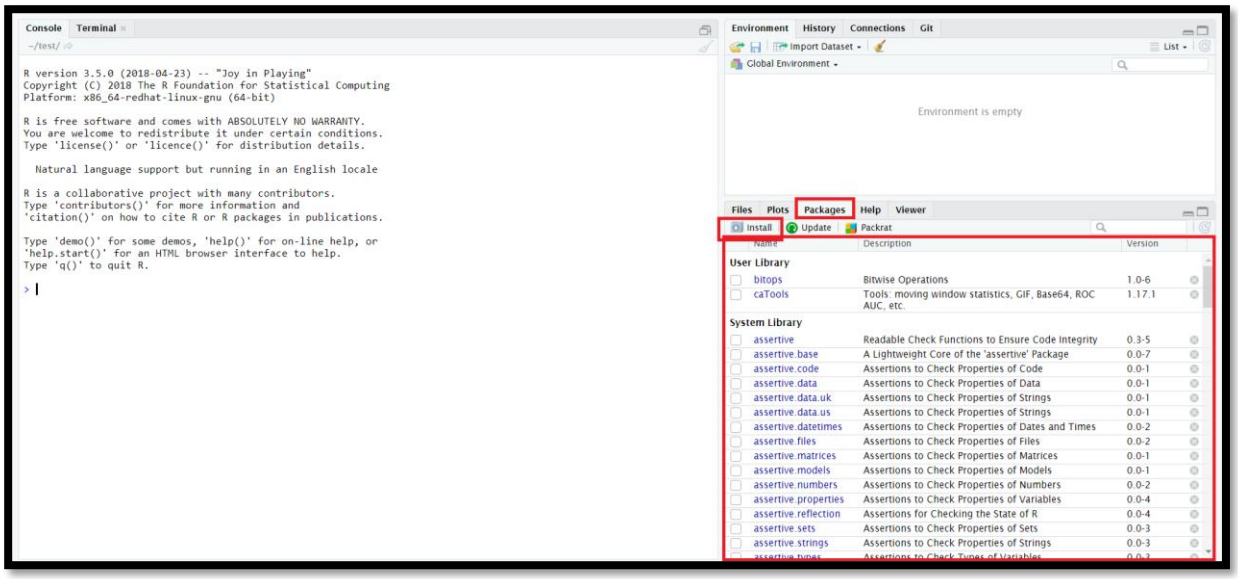
Để cài đặt thư viện trong Rstudio chúng ta có 2 cách tiếp cận cơ bản

Cách 1: Sử dụng tuỳ chọn ứng dụng

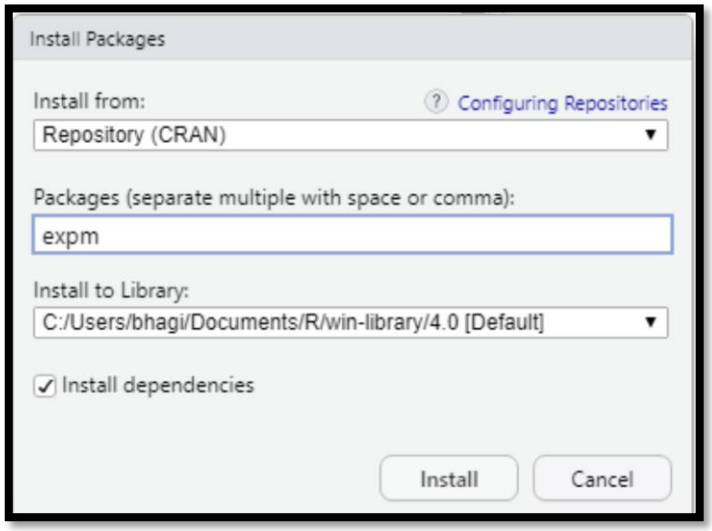
- Bước 1: Mở Rstudio
- Bước 2: Chọn *Tools/Install Packages…*



hoặc



- Bước 3: Chọn tên gói thư viện cần cài đặt → Install



Cách 2: Sử dụng câu lệnh

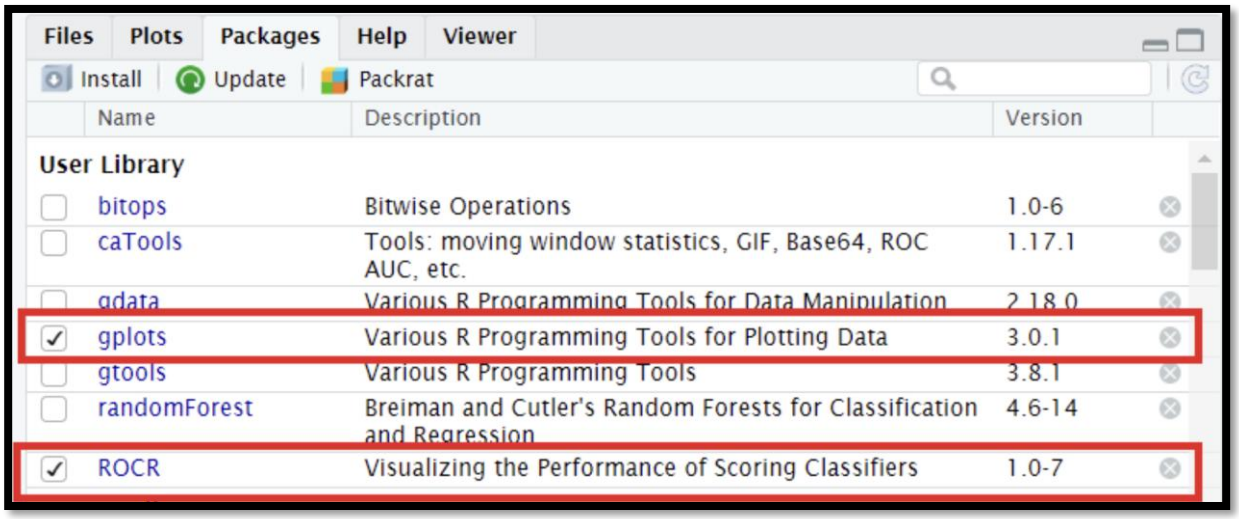
Cú pháp: install.packages("package name")
Ví dụ: install.packages("gplots")

Bài 5: Tải thư viện trong Rstudio

Cách 1: Sử dụng câu lệnh

Cú pháp: library("package name")
Ví dụ: library("gplots")

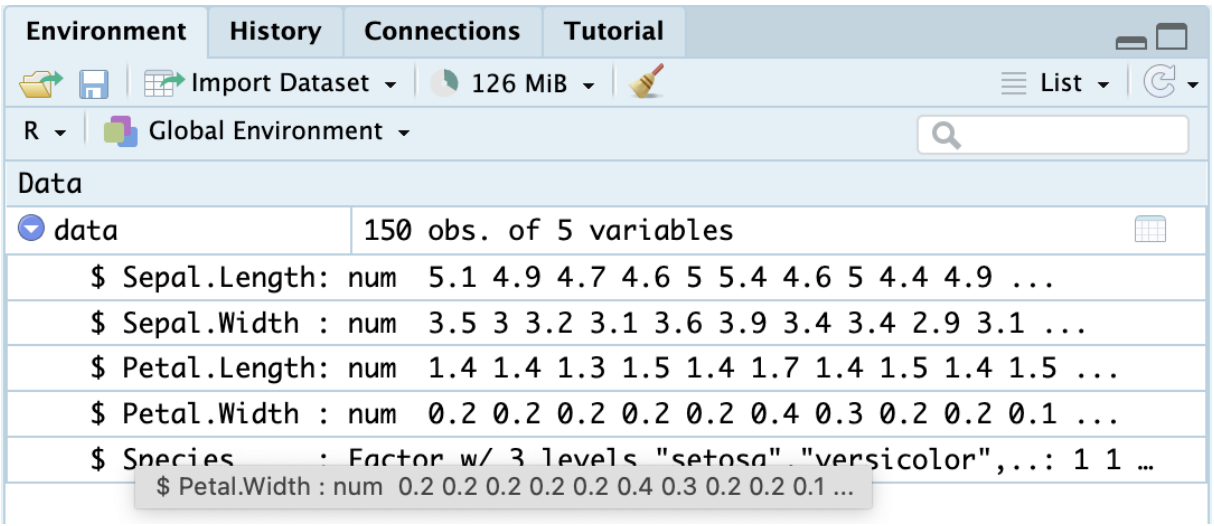
Cách 2: Sử dụng tuỳ chọn ứng dụng



Bài 6: Thực hành cài đặt các bộ thư viện sau: "ggplot2", "dplyr", "tidyr", "mosaicData", "carData"

Bài 7: Xem cấu trúc dữ liệu

Cách 1: Sử dụng tuỳ chọn

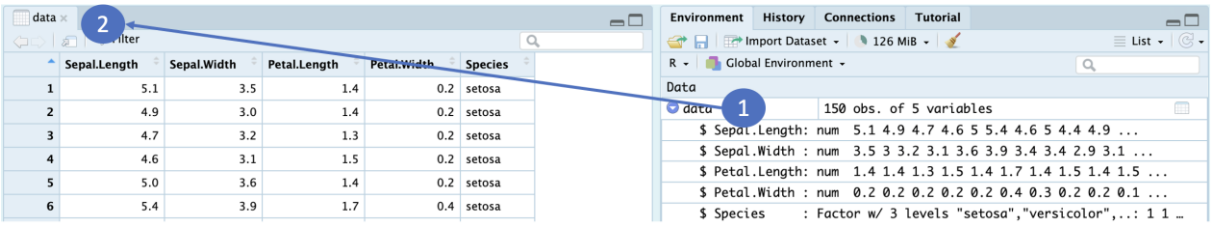


Cách 2: Sử dụng câu lệnh

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Bài 8: Xem nội dung dữ liệu

Cách 1: Sử dụng tuỳ chọn, bấm vào dữ liệu (1) và xem kết quả ở (2)



Cách 2: Sử dụng câu lệnh: View("Tên dữ liệu"). Ví dụ: View(iris)

Bài 9: Khám phá cấu trúc dữ liệu của các bộ dữ liệu thực hành trong mục 1.11

Bài 10: Khám phá nội dung dữ liệu của các bộ dữ liệu thực hành trong mục 1.11