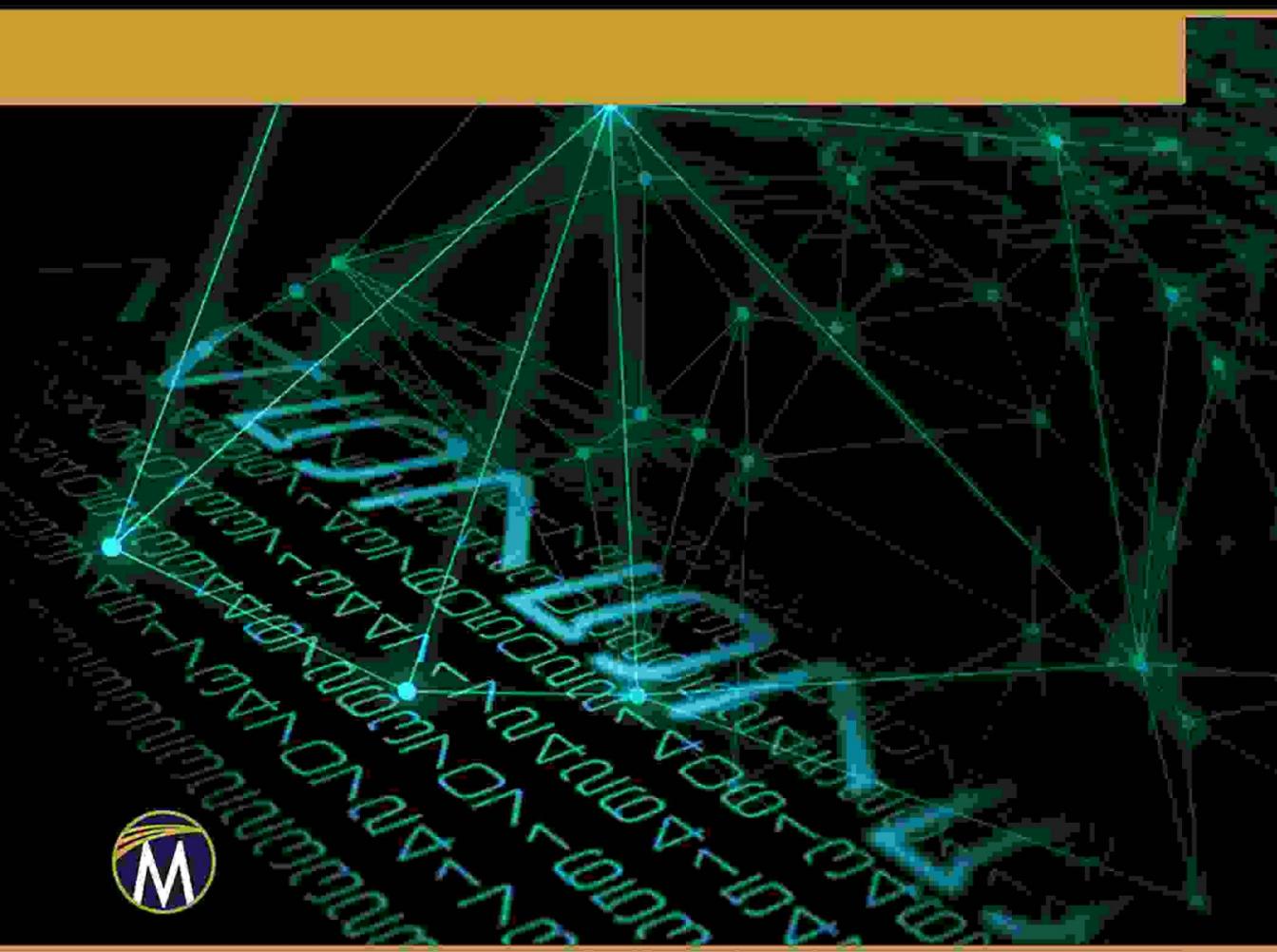


DATA SCIENCE TOOLS

R • Excel • KNIME • OpenOffice



C. GRECO

Khoa học dữ liệu

Công cụ

GIẤY PHÉP, TUYÊN BỐ TỪ CHỐI TRÁCH NHIỆM PHÁP LÝ VÀ BẢO HÀNH CÓ GIỚI HẠN

Bằng cách mua hoặc sử dụng cuốn sách này ("Tác phẩm"), bạn đồng ý rằng giấy phép này cấp quyền sử dụng nội dung có trong tài liệu này, nhưng không cấp cho bạn quyền sở hữu đối với bất kỳ nội dung văn bản nào trong cuốn sách hoặc quyền sở hữu đối với bất kỳ thông tin hoặc sản phẩm có trong đó. Giấy phép này không cho phép tải Tác phẩm lên Internet hoặc trên mạng (dưới bất kỳ hình thức nào) mà không có sự đồng ý bằng văn bản của Nhà xuất bản. Việc sao chép hoặc phổ biến bất kỳ văn bản, mã, mô phỏng, hình ảnh, v.v. nào có trong tài liệu này bị giới hạn và tuân theo các điều khoản cấp phép cho các sản phẩm tương ứng và phải xin phép Nhà xuất bản hoặc chủ sở hữu nội dung, v.v. để sao chép hoặc nối mạng bất kỳ phần nào của tài liệu văn bản (trong bất kỳ phương tiện nào) có trong Tác phẩm.

Mercury Learning and Information ("MLI" hoặc "Nhà xuất bản") và bất kỳ ai tham gia vào việc tạo, viết hoặc sản xuất đĩa đồng hành, các thuật toán, mã hoặc chương trình máy tính đi kèm ("phần mềm") và bất kỳ trang web đi kèm nào hoặc phần mềm của Công việc, không thể và không đảm bảo hiệu suất hoặc kết quả có thể thu được bằng cách sử dụng nội dung của Công việc. Tác giả, nhà phát triển và Nhà xuất bản đã cố gắng hết sức để đảm bảo tính chính xác và chức năng của tài liệu văn bản và/hoặc chương trình có trong gói này; tuy nhiên, chúng tôi không đảm bảo dưới bất kỳ hình thức nào, rõ ràng hay ngụ ý, liên quan đến việc thực hiện các nội dung hoặc chương trình này. Tác phẩm được bán "nguyên trạng" mà không có bảo hành (ngoại trừ các vật liệu bị lỗi được sử dụng để sản xuất sách hoặc do lỗi của ta

Tác giả, nhà phát triển và nhà xuất bản của bất kỳ nội dung đi kèm nào cũng như bất kỳ ai tham gia vào quá trình sáng tác, sản xuất và sản xuất tác phẩm này sẽ không chịu trách nhiệm pháp lý về những thiệt hại dưới bất kỳ hình thức nào phát sinh từ việc sử dụng (hoặc không có khả năng sử dụng) thuật toán, mã nguồn, chương trình máy tính hoặc tài liệu văn bản có trong án phẩm này. Điều này bao gồm, nhưng không giới hạn, mất doanh thu hoặc lợi nhuận, hoặc các thiệt hại ngẫu nhiên, vật chất hoặc hậu quả khác phát sinh từ việc sử dụng Công việc này.

Biện pháp khắc phục duy nhất trong trường hợp có khiếu nại dưới bất kỳ hình thức nào được giới hạn rõ ràng ở việc thay thế sách và chỉ theo quyết định của Nhà xuất bản. Việc sử dụng "bảo hành ngụ ý" và một số "loại trừ" nhất định khác nhau giữa các tiểu bang và có thể không áp dụng cho người mua sản phẩm này.

Khoa học dữ liệu Công cụ

R, Excel, KNIME, & Mở văn phòng

Christopher Greco



TÌM HIỂU VÀ THÔNG TIN THỦY NGÂN

Dulles, Virginia
Boston, Massachusetts
New Delhi

Bản quyền ©2020 của Mercury Learning and Information LLC. Đã đăng ký Bản quyền.

Không được sao chép sản phẩm này, các phần của nó hoặc bất kỳ phần mềm đi kèm nào theo bất kỳ cách nào, được lưu trữ trong bất kỳ loại hệ thống truy xuất nào, hoặc được truyền tải bằng bất kỳ phương tiện, phương tiện, màn hình điện tử hoặc màn hình cơ học nào, bao gồm nhưng không giới hạn ở bản sao, ghi âm, đăng trên Internet hoặc quét mà không có sự cho phép trước bằng văn bản của nhà xuất bản.

Nhà xuất bản: David Pallai

Tìm hiểu và thông tin về thủy ngân

Ở đĩa bạc nhanh 22841

Dulles, VA 20166

info@merclearning.com

www.merclearning.com

(800) 232-0223

C. Hi Lập. Công cụ khoa học dữ liệu: R, Excel, & Mở văn phòng.

KNIME, ISBN: 978-1-68392-583-5

Nhà xuất bản công nhận và tôn trọng tất cả các nhãn hiệu được sử dụng bởi các công ty, nhà sản xuất và nhà phát triển như một phương tiện để phân biệt sản phẩm của họ. Tất cả tên thương hiệu và tên sản phẩm được đề cập trong cuốn sách này là thương hiệu hoặc nhãn hiệu dịch vụ của các công ty tương ứng. Bất kỳ sự thiếu sót hoặc sử dụng sai (dưới bất kỳ hình thức nào) nhãn hiệu dịch vụ hoặc thương hiệu, v.v. không phải là hành vi cố gắng xâm phạm tài sản của người khác.

Thư viện Quốc hội Số kiểm soát: 2020937123

202122321 Được in trên giấy không có axit ở Hoa Kỳ

Các tiêu đề của chúng tôi có sẵn để các tổ chức, tập đoàn, v.v. áp dụng, cấp phép hoặc mua số lượng lớn. Để biết thêm thông tin, vui lòng liên hệ với Phòng Dịch vụ Khách hàng theo số (800) 232-0223 (miễn phí). Các phiên bản kỹ thuật số của các đầu sách của chúng tôi có sẵn tại: www.academiccourseware.com và các nhà cung cấp điện tử khác.

Nghĩa vụ duy nhất của Mercury Learning and Information đối với người mua là thay thế sách và/hoặc đĩa, dựa trên nguyên liệu bị lỗi hoặc tay nghề sản xuất bị lỗi, nhưng không dựa trên hoạt động hoặc chức năng của sản phẩm.

NỘI DUNG

lời nói đầu	ix
Sự nhìn nhận	xi
Ghi chú về quyền	xiii
Chương 1: Các bước đầu	1
tiên 1.1 Giới thiệu về Công cụ dữ liệu	1
1.1.1 Phần mềm dễ sử dụng 1.1.2	2
Phần mềm có sẵn ở mọi nơi 1.1.3 Phần mềm	2
được cập nhật thường xuyên 1.1.4 Tóm	2
tắt 1.2 Tại sao	2
phải phân tích dữ liệu (Khoa học dữ liệu)?	3
1.3 Lấy dữ liệu ở đâu	3
Chương 2: Nhập dữ liệu	5
2.1 Excel	5
2.1.1 Excel Analysis ToolPak	7
2.2 OpenOffice	9
2.3 Nhập vào R và Rattle	11
2.4 Nhập vào RStudio	12
2.5 Nhập Rattle	18
2.6 Nhập vào KNIME 2.6.1	24
Phương pháp tiếp cận Stoplight	32

vi • Nội dung

Chương 3:	kiểm tra thống kê	35
	3.1 Thống kê mô tả	35
	3.1.1 Excel	35
	3.1.2 OpenOffice	39
	3.1.3 RStudio/Rattle	42
	3.1.4 KIẾN THỨC	48
	3.2 Biểu đồ xác suất tích lũy	52
	3.2.1 Excel	52
	3.2.2 OpenOffice	56
	3.2.3 R/RStudio/Rattle	67
	3.2.4 KIẾN THỨC	73
	3.3 Thủ nghiệm T (Tham số)	91
	3.3.1 Excel	91
	3.3.2 OpenOffice	93
	3.3.3 R/RStudio/Rattle	96
	3.3.4 KIẾN THỨC	97
Chương 4: Các kiểm định thống kê khác		103
	4.1 Tương quan	103
	4.1.1 Excel	103
	4.1.2 OpenOffice	105
	4.1.3 R/RStudio/Rattle	106
	4.1.4 KIẾN THỨC	108
	4.2 Hồi quy	109
	4.2.1 Excel	110
	4.2.2 OpenOffice	112
	4.2.3 R/RStudio/Rattle	113
	4.2.4 KIẾN THỨC	115
	4.3 Khoảng tin cậy	117
	4.3.1 Excel	119
	4.3.2 OpenOffice	121
	4.3.3 R/RStudio/Rattle	122
	4.3.4 KIẾN THỨC	124

4.4 Lấy mẫu ngẫu nhiên	127
4.4.1 Excel	128
4.4.2 OpenOffice	129
4.4.3 R/RStudio/Rattle	132
4.4.4 KIẾN THỨC	134
Chương 5:	
Phương pháp thống kê cho các công cụ cụ thể	137
5.1 Sức mạnh	137
5.1.1 R/RStudio/Rattle	138
5.2 Thủ nghiệm F	140
5.2.1 Excel	140
5.2.2 R/RStudio/Rattle	142
5.2.3 KIẾN THỨC	143
5.3 Hồi quy bội/Tương quan	145
5.3.1 Excel	145
5.3.2 OpenOffice	147
5.3.3 R/RStudio/Rattle	148
5.3.4 KIẾN THỨC	150
5.4 Định luật Benford	151
5.4.1 Tiếng lạch cách	151
5.5 thang mây	157
5.5.1 KIẾN THỨC	157
5.6 Đám mây từ	160
5.6.1 R/RStudio	160
5.6.2 KIẾN THỨC	162
5.7 Lọc	170
5.7.1 Excel	171
5.7.2 OpenOffice	173
5.7.3 R/RStudio/Rattle	174
5.7.4 KIẾN THỨC	174
Chương 6: Tóm tắt	
6.1 Các	177
6.2 Phân	177
tích ToolPak	179

Chương 7:	Thông tin bổ sung 7.1 Bài	181
	tập một - Lốc xoáy và các quốc gia	181
	7.1.1 Trả lời bài tập 7.1	182
	7.1.2 Bài tập ghép	194
	nội Tham khảo	202
Mục lục		203

LỜI NÓI ĐẦU

Khoa học dữ liệu là tất cả các cơn thịnh nộ. Có khả năng cao là mọi cuốn sách bạn đọc, mọi trang Web bạn truy cập, mọi quảng cáo bạn nhận được đều là kết quả của khoa học dữ liệu và cùng với nó là phân tích dữ liệu. Những gì từng là "thống kê" hiện được gọi là phân tích dữ liệu hoặc khoa học dữ liệu. Các khái niệm đãng sau khoa học dữ liệu là vô số và phức tạp, nhưng khái niệm cơ bản là các khái niệm thống kê cơ bản rất quan trọng để hiểu dữ liệu. Cuốn sách này thực sự có hai mục đích. Đầu tiên là xem xét ngắn gọn một số khái niệm mà người đọc có thể gặp phải khi tham gia một khóa học (hoặc các khóa học) về thống kê, trong khi thứ hai là trình bày cách sử dụng các công cụ để trực quan hóa các khái niệm thống kê đó.

Có một số cảnh báo phải đi kèm với cuốn sách này. Đầu tiên là các công cụ thuộc một phiên bản nhất định, sẽ được mô tả bên dưới. Điều này có nghĩa là chắc chắn sẽ có các phiên bản trong tương lai của những công cụ này có thể hoạt động khác trên máy tính của bạn. Tôi muốn nói rõ rằng hiệu suất này không có nghĩa là những công cụ này sẽ hoạt động tốt hơn. Ba trong số này là các công cụ mã nguồn mở và miễn phí, và do đó, hoạt động tốt như mong muốn của nhóm nhà phát triển trong các phiên bản mới nhất của chúng. Trong hầu hết các trường hợp, công cụ này sẽ được cải tiến trong phiên bản mới hơn, nhưng có thể có một "nút bấm" khác sẽ được liên kết với các chức năng mới hơn.

Bạn sẽ thấy từ "nút bấm" xuyên suốt cuốn sách này dưới dạng cơ học của chính công cụ này. Tôi không ở đây để dạy người đọc về số liệu thống kê hoặc các khái niệm khác nhau tạo nên các chủ đề của cuốn sách này. Tôi ở đây để chỉ cho bạn cách các công cụ mã nguồn mở và miễn phí được áp dụng cho các khái niệm này.

Bây giờ là lúc đi vào trọng tâm của văn bản, các công cụ của khoa học dữ liệu. Sẽ có bốn công cụ sẽ bao gồm nội dung của cuốn sách này.

Ba là công cụ nguồn mở (FOSS hoặc Nguồn mở và Miễn phí), với một là phần mềm COS (Common Off the Shelf), nhưng cả bốn sẽ yêu cầu một số hướng dẫn khi sử dụng. Những điều này không phải lúc nào cũng trực quan hoặc dễ hiểu,

x • Lời nói đầu

vì vậy sẽ có nhiều trang màn hình cho từng chức năng cơ học. Tôi cảm thấy rằng làm quen với hình ảnh hơn hẳn tường thuật, vì vậy bạn sẽ không thấy nhiều văn bản, chủ yếu là mô tả và cơ chế từng bước. Một số bạn có thể băn khoăn về cách thực hành những kỹ năng này và đối với những đặc điểm đó, có một chương cuối cùng có một số tình huống cho phép người đọc áp dụng những gì họ đã học được từ những công cụ này.

Bố cục của cuốn sách này sẽ dựa trên khái niệm thống kê chứ không phải công cụ, nghĩa là mỗi chương sẽ bao gồm phần giải thích về khái niệm thống kê và sau đó là cách áp dụng từng công cụ cho khái niệm đó. Bằng cách sử dụng phương pháp trình bày này, người đọc có thể đi đến khái niệm được quy định và sử dụng công cụ được áp dụng một cách thoải mái nhất. Mỗi phần sẽ được gắn nhãn tương ứng, vì vậy chúng sẽ có cả trong mục lục và chỉ mục. Điều này làm cho các cá nhân thấy sự lựa chọn công cụ của họ và các khái niệm mà họ phải áp dụng cho các công cụ đó đơn giản hơn.

C. Hy Lạp

tháng 4 năm 2020

SỰ NHÌN NHẬN

Trước đây, khi làm những việc này, tôi luôn nhắc đến vợ, con và cháu của mình, điều mà đối với tôi không chỉ cần thiết mà còn là bắt buộc, bởi vì họ là những người tác động đến tôi hàng ngày. Cảm ơn các anh chị em của tôi, những người luôn đặt tiêu chuẩn đủ cao để đạt được sự xuất sắc, nhưng không quá cao đến mức tôi sẽ tự làm hại mình khi vượt qua nó. Tất cả các bạn luôn cung cấp cho tôi động lực để làm tốt hơn. Bây giờ, tôi phải thêm một vài người đã giúp tôi in cuốn sách này và trên các phương tiện truyền thông điện tử. Người đầu tiên và quan trọng nhất là Jim Walsh của Mercury Learning, người đã mạo hiểm để tôi viết một cuốn sách về các ứng dụng mã nguồn mở và miễn phí. Tôi thực sự tin tưởng vào cuốn sách này, và anh ấy tin tưởng tôi sẽ nỗ lực hết mình, nhưng ngoài ra, anh ấy còn đưa ra những gợi ý giúp tôi trở thành một nhà văn giỏi hơn và đóng góp cho bức tranh xuất bản lớn hơn. Tôi thực sự đánh giá cao tất cả sự giúp đỡ của bạn, Jim.

Các biên tập viên và nhà văn khác tại Mercury Learning giống như đang nhìn vào Đại sảnh Danh vọng Khoa học, Công nghệ, Kỹ thuật và Toán học (STEM). Tôi thực sự vinh dự và vinh dự được đặt tên sách cho nhóm quý tộc này. Cảm ơn tất cả các hướng dẫn.

Cuối cùng, cha tôi, người đã nói với tôi một cách dứt khoát rằng tôi không bao giờ nên cố gắng học "khoa học cứng" mà hãy gắn bó với "khoa học mềm", vì tôi thực sự rất đốt toán. Cảm ơn bố vì đã cho tôi động lực để theo đuổi thống kê và phân tích dữ liệu. Tôi nợ tất cả cho bạn.

LƯU Ý VỀ PHÉP

- Ảnh chụp màn hình của Tập đoàn Microsoft tuân theo các nguyên tắc được xem tại đây: <https://www.microsoft.com/en-us/legal/intellectualproperty/permissions/macking.aspx>.
- Các ảnh chụp màn hình của OpenOffice tuân theo các nguyên tắc có thể thấy ở đây: <https://www.openoffice.org/license.html>.
- Ảnh chụp màn hình R / RStudio được cho phép thông qua giấy phép RStudio và quyền <https://rstudio.com/about/software-license-descriptions/>.
- Quy R: <http://www.r-project.org>.
- Ảnh chụp màn hình rattle được sử dụng với sự cho phép và cũng được trích dẫn trong: Graham Williams. (2011). Khai thác dữ liệu với Rattle và R: Nghệ thuật khai thác dữ liệu để khám phá tri thức. Người dùng! New York, NY: Mùa xuân.
- Ảnh chụp màn hình KNIME được cho phép thông qua cấp phép và cấp phép KNIME: <https://www.knime.com/downloads/full-license>.

CHƯƠNG 1

NHỮNG BƯỚC ĐẦU TIÊN

1.1 GIỚI THIỆU CÔNG CỤ DỮ LIỆU

Mỗi người có những động lực khác nhau để theo đuổi những gì họ quan tâm. Hỏi ai đó về ô tô và họ có thể nói rằng họ ghét xe sedan, hoặc thích SUV, hoặc sẽ không bao giờ mua bất cứ thứ gì khác ngoài ô tô điện, hoặc có thể không mua ô tô nào cả! Mọi người có những sở thích khác nhau và điều này không thay đổi với các công cụ khoa học dữ liệu (thống kê). Một số người yêu thích Excel đến mức họ sẽ không sử dụng gì khác ngoài phần mềm đó cho bất kỳ việc gì, từ lập ngân sách đến phân tích dữ liệu. Có nhiều lý do để duy trì sự cống hiến, nhưng lý do chính từ kinh nghiệm của tôi là sự quen thuộc với đối tượng. Một người mới chỉ lái côn sẽ yêu thích bộ ly hợp, trong khi những người chưa bao giờ lái côn sẽ không có xu hướng thích một chiếc có lẫy chuyển số bằng tay.

Có những lý do nào để thích một ứng dụng phần mềm này hơn một ứng dụng phần mềm khác? Từ kinh nghiệm của tôi, có ba điểm chính:

1. Phần mềm rất dễ sử dụng
2. Phần mềm có sẵn ở mọi nơi
3. Phần mềm được cập nhật thường xuyên

Thông thường, có thể nói rằng phần mềm không đắt, nhưng với thời đại đăng ký, giấy phép phần mềm không còn vĩnh viễn, vì vậy, việc thanh toán hàng tháng là tất cả những gì cần thiết để đảm bảo rằng người đọc có quyền truy cập vào phần mềm miễn là đăng ký còn hiệu lực. Hãy khám phá từng điểm và xây dựng.

1.1.1 Phần mềm dễ sử dụng

Nếu một nhà phân tích có thể chọn một vài nút và – thì đây – kết quả xuất hiện, thì nó dễ hơn nhiều so với từ "p". Từ "p" là gì? Lập trình! Nếu một nhà phân tích phải lập trình, sẽ rất khó để có được kết quả. Tuy nhiên, các nhà phân tích không nhận ra rằng một khi thử gì đó đã được lập trình thì việc áp dụng cách lập trình đó sẽ dễ dàng hơn, nhưng đó là chuyện của một cuốn sách khác vào một thời điểm khác. Điểm chính để đạt được ở đây là phần mềm Giao diện người dùng đồ họa (GUI) dường như được ưa thích hơn phần mềm lập trình. Phần mềm COS nổi tiếng và cũng được biết đến là dễ sử dụng. Một số phần mềm FOSS sẽ yêu cầu chuẩn bị nhiều hơn.

1.1.2 Phần mềm có sẵn ở mọi nơi

Trong thời đại điện toán đám mây này, việc có thể truy cập phần mềm dường như là chuyện nhỏ. Sau khi nói chuyện với các đồng nghiệp, họ thích thực tế là họ có thể thực hiện và lưu công việc của mình trực tuyến để không bị mất. Họ cũng thích thực tế là các bản cập nhật minh bạch và được thực hiện trong khi họ đang sử dụng công cụ. Cuối cùng, họ thích thực tế là họ không phải lo lắng về việc cài đặt phần mềm và sử dụng bộ nhớ hoặc dung lượng ổ đĩa.

1.1.3 Phần Mềm Được Cập Nhật Thường Xuyên

Phần trước đề cập đến điều này, vì vậy chúng tôi sẽ không giải thích chi tiết. Tuy nhiên, điều quan trọng cần lưu ý là các công cụ sẽ được đề cập trong cuốn sách này được cập nhật thường xuyên. Thật không may, nhà phân tích sẽ phải là người chọn tham gia các bản cập nhật.

1.1.4 Tóm tắt

Bây giờ chúng ta đã tìm hiểu lý do tại sao các nhà phân tích thích một số công cụ nhất định, phần mô tả về những công cụ được đề cập trong cuốn sách này sẽ được đưa ra dưới dạng bảng để đơn giản hóa việc trình bày và (như đã nêu trước đây) giảm thiểu từ viết.

Phần mềm	Dễ dàng (1=DỄ DÀNG, 5=Khó)	Có sẵn	cập nhật
Excel	1	24/7	Công ty
R(RStudio / lách cách)	3	24/7	nha phan tich
KNIME	4	24/7	nha phan tich
Mở văn phòng	2	24/7	nha phan tich

1.2 TẠI SAO PHÂN TÍCH DỮ LIỆU (DATA KHOA HỌC) HAY KHÔNG?

Thế giới ngày nay là một bản tóm tắt dữ liệu. Dữ liệu tồn tại trong mọi việc chúng ta làm, cho dù đó là mua hàng tạp hóa hay tìm hiểu để mua nhà. Có rất nhiều ứng dụng và ứng dụng miễn phí có sẵn cho chúng tôi đến nỗi chúng tôi khó có thể nói không với bất kỳ ứng dụng nào trong số này. Như một tài liệu tham khảo đã nói, và tác giả này đã khái quát hóa, nếu những gì bạn đang tải xuống là miễn phí, thì bạn chính là sản phẩm (Poundstone, 2019). Điều này thật sâu sắc, bởi vì nguồn mở và miễn phí (FOSS) là thứ thường được truy cập và sẵn có cho tất cả chúng ta. Tuy nhiên, tại sao chúng ta cần khoa học dữ liệu để phân tích tất cả các thông tin này? Theo hiểu biết của tôi, có một số lý do khiến khoa học dữ liệu tồn tại. Đầu tiên, nó tồn tại để thu thập hàng nghìn tỷ byte thông tin được các công ty và cơ quan chính phủ thu thập để xác định mọi thứ, từ giá thành sửa đến lượng khí thải carbon trong không khí. Bốn mươi năm trước, hầu hết dữ liệu được thu thập, truy xuất và lưu trữ bằng giấy. Máy tính cá nhân là một giấc mơ và khoa học dữ liệu được gọi là lưu trữ hoặc một cái gì đó tương tự. Hướng tới phương tiện điện tử, cơ sở dữ liệu đã biến những đồng giấy thành kilo-, mega-, giga- và thậm chí là petabyte. Nhưng với lượng dữ liệu đó, việc phân tích đã chuyển từ bút chì và giấy sang máy tính cá nhân hoặc bất kỳ máy tính nào. Các nhà phân tích bắt đầu nhận ra rằng phần mềm động là phương tiện để chuyển phân tích dữ liệu thành một dạng dễ sử dụng hơn.

Khoa học dữ liệu phát triển từ nỗ lực phân tích dữ liệu này và sử dụng các phương pháp thống kê thông thường kết hợp với sức mạnh của máy tính để cung cấp khoa học dữ liệu cho tất cả các tổ chức tư nhân và công cộng. Với khả năng phân tích dữ liệu tiếp thị, kỹ thuật và nhân sự, giờ đây các công ty có khả năng tính toán xác suất thành công của sản phẩm hoặc doanh thu của họ sẽ tăng trong năm tới. Với sự phát triển của khoa học dữ liệu, nhiều công cụ giúp phân tích dữ liệu trở nên khả thi.

1.3 LẤY DỮ LIỆU Ở ĐÂU

Bây giờ chúng ta đã có phần giới thiệu về “tại sao” của khoa học dữ liệu, chủ đề tiếp theo là “ở đâu”. Bạn lấy dữ liệu ở đâu để sử dụng với các công cụ khoa học dữ liệu? Câu trả lời cho câu hỏi đó, đặc biệt là hiện nay, là dữ liệu có sẵn trên nhiều trang web để phân tích (Williams, 2011). Một số trang web này bao gồm:

1. www.data.gov, chứa các trang dữ liệu từ các cơ quan chính phủ khác nhau. Nếu bạn muốn biết về dữ liệu khí hậu, điều tra dân số hoặc kiểm soát dịch bệnh, thì đây là nơi dành cho bạn.

2. www.kaggle.com, không chỉ chứa dữ liệu mà còn có các cuộc thi với dữ liệu hiện có mà bất kỳ ai cũng có thể tham gia. Một bộ dữ liệu chứa các dữ liệu khác nhau được thu thập từ Titanic, bao gồm số người chết hoặc sống sót và tất cả thông tin nhân khẩu học để phân tích và tương quan.

3. Gần như bất kỳ cơ quan chính phủ liên bang nào. Nếu bạn không muốn truy cập một trang web chung, hãy truy cập www.cdc.gov, www.census.gov, www.noaa.gov hoặc bất kỳ trang web riêng biệt nào của chính phủ để biết dữ liệu liên quan đến những thứ như An sinh xã hội (www.ssa.gov) hoặc Thám chí là tình báo (www.nsa.gov) đối với một số dữ liệu lịch sử.

Bây giờ bạn đã có “tại sao” và “ở đâu” liên quan đến công cụ và khoa học dữ liệu, bây giờ bạn chuyển sang bước tiếp theo—thực sự sử dụng các công cụ với dữ liệu thực. Bên cạnh đó, bạn chắc chắn đã có đủ bối cảnh sân khấu này.

Dữ liệu cho cuốn sách này được lấy tại trang web, <https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/>, có dữ liệu theo dõi cơn lốc xoáy của Hoa Kỳ từ năm 1951 đến năm 2018. Cơ quan chính phủ NOAA là viết tắt của Cơ quan Khí quyển và Đại dương Quốc gia. Khuyến nghị là tải xuống các tệp này (bao nhiêu tùy thích) và sử dụng riêng chúng cho các ví dụ trong sách. Cuốn sách này sẽ tập trung vào quá trình theo dõi cơn lốc xoáy năm 1951 để làm cho nó tương đối đơn giản. Sau khi bạn tải xuống dữ liệu, bước tiếp theo là nhập dữ liệu vào công cụ thống kê yêu thích của bạn.

CHƯƠNG 2

NHẬP DỮ LIỆU

Bước đầu tiên để phân tích dữ liệu là nhập dữ liệu vào công cụ thích hợp.

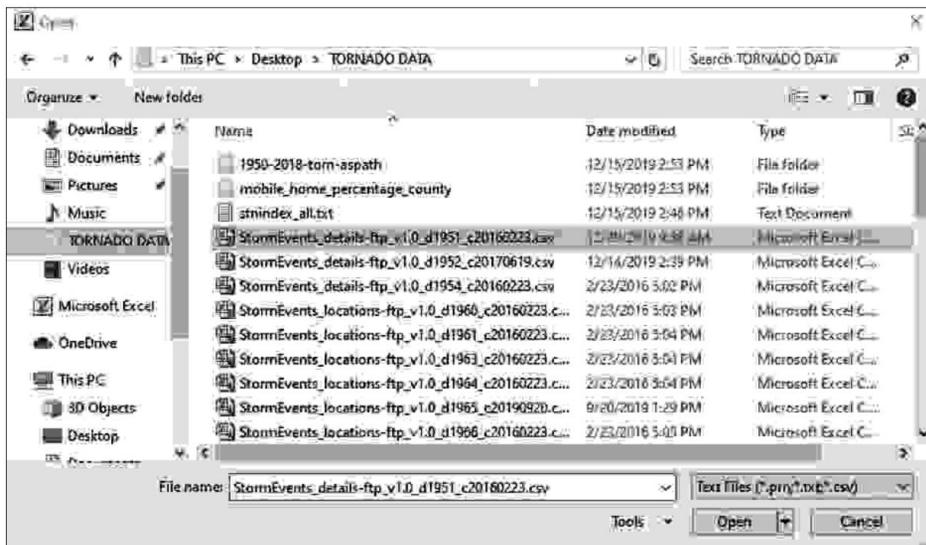
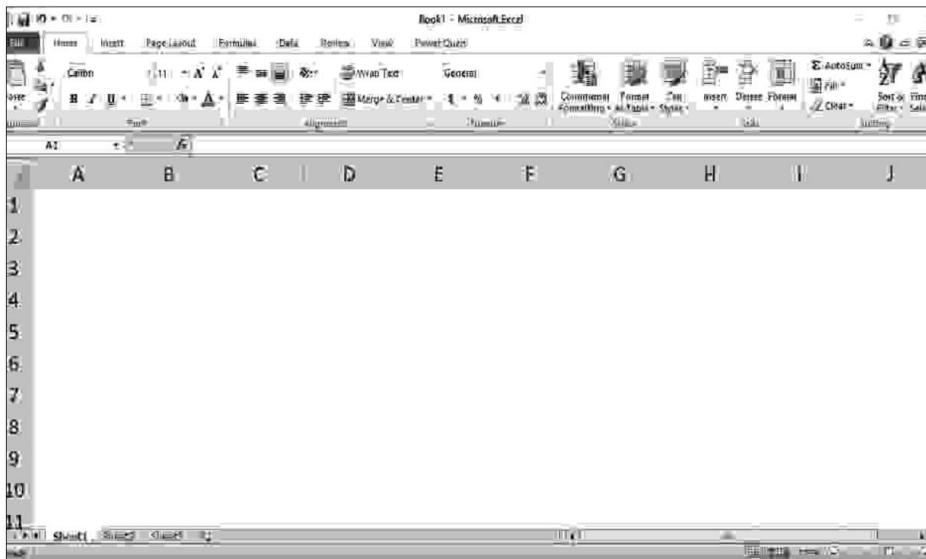
Phần đầu tiên này sẽ trình bày cách nhập dữ liệu bằng cách sử dụng từng công cụ—Excel, R, KNIME và OpenOffice. Vì hầu hết các nhà phân tích đều quen thuộc với Excel, nên Excel sẽ là ứng dụng đầu tiên được đề cập và sau đó là OpenOffice, vì nó rất gần với Excel về chức năng, để giới thiệu tốt về nhập dữ liệu.

2.1 XUẤT SẮC

Phiên bản cho văn bản này sẽ là Microsoft Excel 2016, vì đó là phiên bản xuất hiện trong nhiều cơ quan chính phủ liên bang. Tại thời điểm viết cuốn sách này, Excel 2019 đã có sẵn nhưng chưa được sử dụng trong dịch vụ công vào thời điểm này.

Nhập dữ liệu vào Excel không thể dễ dàng hơn. Tệp đã được tải xuống là tệp Giá trị được phân tách bằng dấu phẩy (CSV), vì vậy, để nhập tệp vào Excel, hãy chuyển đến vị trí tệp và bấm đúp vào tệp. Tệp sẽ xuất hiện trong Excel nếu máy tính mặc định tất cả các bảng tính sẽ vào Excel. Nếu không, hãy mở Excel và chọn “Tệp” và “Mở” để đến vị trí tệp và mở tệp. Các màn hình sau đây minh họa hoạt động.

6 • Công cụ khoa học dữ liệu



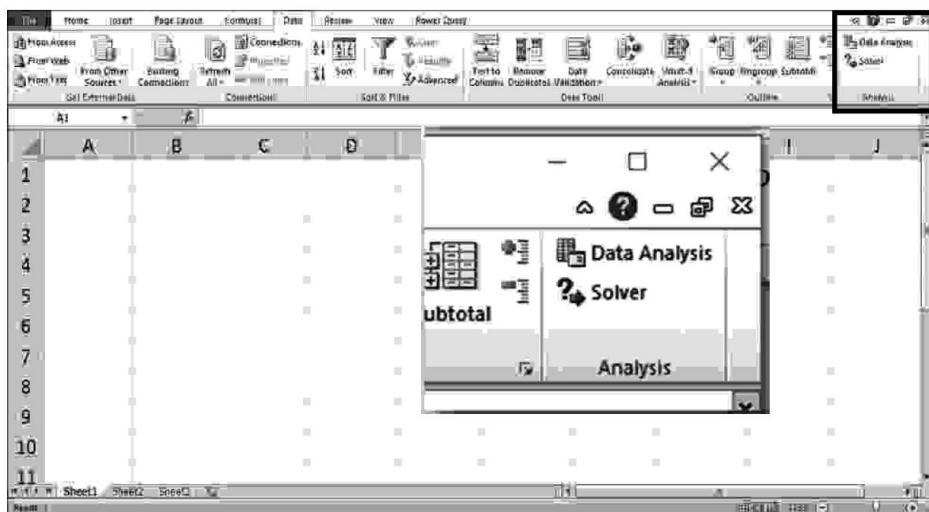
Một cảnh báo vào thời điểm này với Excel. Khi mở tệp, phần mở rộng mặc định cho Excel là phần mở rộng trang tính hoặc ".xlsx". Nếu trang tính là một CSV, thì mặc định đó phải được thay đổi, như đã trình bày trong quy trình trước. Sau khi tiện ích mở rộng được thay đổi, hãy nhấp vào "Mở" và bảng tính

sẽ xuất hiện trong Excel. Nếu mục đích là để duy trì dưới dạng CSV thì hãy lưu nó như vậy khi bạn hoàn thành công việc trên bảng tính. Nếu không, hãy lưu nó dưới dạng tệp "XLSX" để tất cả chức năng của Excel vẫn nằm trong bảng tính khi quá trình phân tích tiếp tục.

Đây có lẽ là cách nhập dễ dàng nhất cho bất kỳ ứng dụng nào được trình bày do tính chất trực quan của Excel.

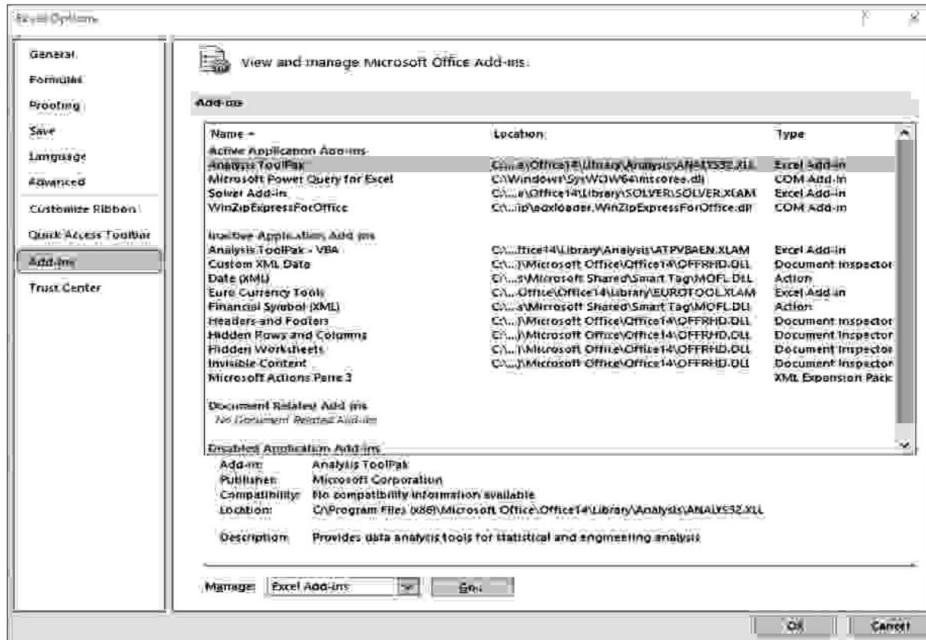
2.1.1 Công cụ phân tích Excel

Từ thời điểm này trở đi, đối với bất kỳ phân tích thống kê nào với Excel, chúng tôi sẽ sử dụng ToolPak Phân tích, công cụ này sẽ cần được cài đặt dưới dạng tiện ích bổ sung thông qua Excel. Nếu ToolPak Phân tích đã được cài đặt, nó sẽ hiển thị trong tab "Dữ liệu" của Excel như được hiển thị ở đây.



Nếu ToolPak Phân tích không hiển thị trên thanh công cụ Dữ liệu, nhà phân tích có thể thêm nó đơn giản bằng cách chuyển đến tab "Tệp" và chọn "Tùy chọn" ở cuối cột bên trái. Một màn hình sẽ xuất hiện hiển thị tất cả các khả năng ở cột bên trái. Nhà phân tích chọn "Phần bổ trợ" và màn hình bên dưới sẽ xuất hiện, hiển thị tất cả các phần bổ trợ có sẵn hoặc không có sẵn. Hãy dành một chút thời gian và xem xét các phần bổ trợ có sẵn như một phần của quá trình cài đặt Excel. Có một số trong số chúng và chúng rất hữu ích trong phân tích dữ liệu. Hãy dành thời gian khám phá những phần bổ trợ này để xem cách chúng có thể cải thiện phân tích của bạn, nhưng trong thời gian chờ đợi, hãy hoàn tất cài đặt phần bổ trợ ToolPak Phân tích để hoàn thành phân tích này.

8 • Công cụ khoa học dữ liệu



Khi chọn Tùy chọn, màn hình tiếp theo sẽ hiển thị một số lựa chọn ở cột bên trái. Chọn “Add-Ins” và sẽ có một danh sách các add-in khả dụng cho Excel. Chọn “Analysis ToolPak”, lúc này sẽ nằm trong “Phần bổ trợ ứng dụng không hoạt động” và đi xuống cuối màn hình có ghi “Quản lý:” để đảm bảo rằng “Phần bổ trợ Excel” có trong văn bản hộp. Nhấp vào nút “Go.” và màn hình sau sẽ xuất hiện.



Nhấp vào hộp kiểm bên cạnh “Analysis ToolPak” để kích hoạt phần bổ trợ và phần bổ trợ sẽ xuất hiện trên thanh công cụ Excel. Nếu không, hãy thử thoát khỏi Excel và thử lại quy trình. Nó sẽ hoạt động tại thời điểm đó. Nếu nó không hoạt động sau nhiều lần thử và máy tính là máy tính của chính phủ, thì có thể có một tường lửa tại chỗ sẽ ngăn việc sử dụng phần bổ trợ này. Nếu quản trị viên hệ thống không thể cung cấp cho máy tính quyền truy cập, có một mô tả ở cuối cuốn sách này sẽ minh họa nút bấm để thay thế cho ToolPak Phân tích.

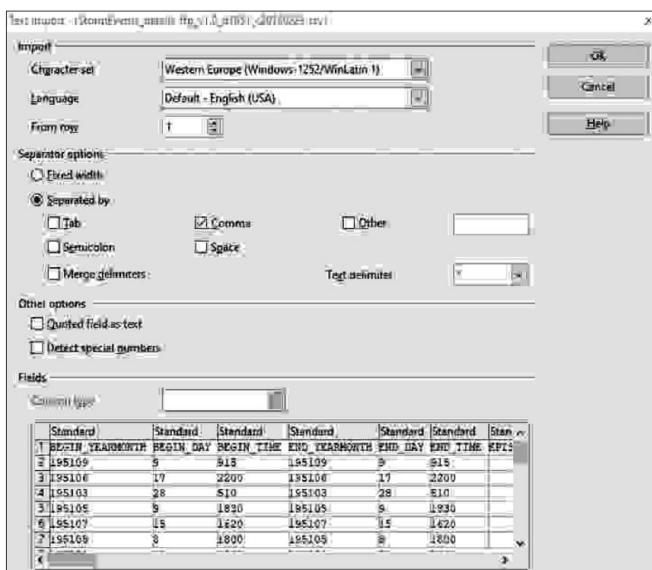
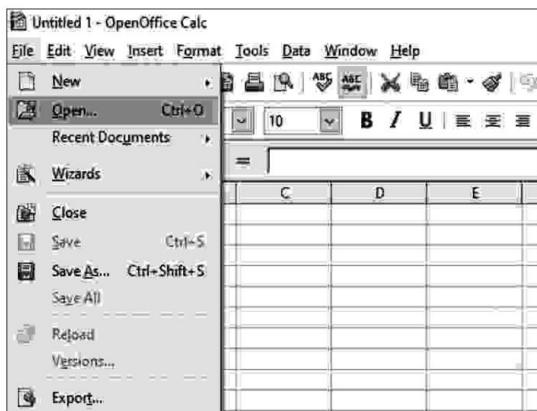
2.2 KHAI TRƯƠNG

Bước đầu tiên để sử dụng OpenOffice là tải xuống phần mềm từ trang web OpenOffice (www.openoffice.org), tương đối đơn giản. Phiên bản hiện tại của phần mềm là 4.1.7, đây sẽ là phiên bản mà chúng tôi sẽ sử dụng trong cuốn sách này. Khi bạn cài đặt OpenOffice, bạn không phải cài đặt tất cả các chức năng khác nhau và trong trường hợp này, bạn chỉ cần chương trình bảng tính, vì vậy khi mở màn hình giới thiệu, bạn sẽ thấy như sau:



Tại thời điểm này, hãy chọn Bảng tính và màn hình này sẽ xuất hiện, trông rất giống Excel. Trên thực tế, nếu bạn đã sử dụng Excel từ năm 1998 đến 2000, nó sẽ trông rất giống các phiên bản đó. Điều này có nghĩa là tính chất chức năng không hoàn toàn giống nhau, nhưng nó sẽ là mọi thứ bạn cần cho các khái niệm thống kê trong cuốn sách này.

Nhiệm vụ đầu tiên sẽ là nhập dữ liệu lấy từ Internet. Trong trường hợp này, đó sẽ là dữ liệu từ một trang web theo dõi các cơn lốc xoáy xảy ra ở Hoa Kỳ từ năm 1950-2018. Dữ liệu này sẽ được nhập bằng cách sử dụng kỹ thuật tương tự như trong Excel thông qua lệnh "mở" trong Menu Tệp như được mô tả ở đây:



The screenshot shows a Microsoft Excel spreadsheet with a large dataset. The columns represent various metadata fields such as ID, Title, Year, Month, Name, Month Number, and others. The data consists of approximately 31 rows of mostly identical or very similar entries, likely representing TV show episodes. The 'Properties' pane on the right indicates the active cell is A1.

Bây giờ đến phần làm sạch và chuyển đổi dữ liệu để chuẩn bị cho việc phân tích. Tuy nhiên, để cung cấp tệp này cho các công cụ khác, có thể thuận lợi khi lưu tệp dưới dạng tệp Excel hoặc thậm chí là tệp văn bản. Đối với những người thích tệp Giá trị được phân tách bằng dấu phẩy (CSV), hầu hết dữ liệu được tìm thấy trên nhiều trang web dữ liệu dường như được mặc định thành tệp CSV, do đó, để tệp này trong tiện ích mở rộng CSV sẽ ổn.

2.3 NHẬP VÀO R VÀ RATTLE

Nhập dữ liệu vào ứng dụng thống kê R tương đối dễ dàng nếu người đọc tải xuống cả ứng dụng R và RStudio. Có thể tìm thấy R trong trang web Mạng lưu trữ R toàn diện (CRAN) cho ứng dụng R (<https://cran.r-project.org/>), trong khi RStudio có thể được tìm thấy tại <https://rstudio.com/products/rstudio/>. Cả hai sẽ cần được cài đặt để làm cho R ít tập trung vào chương trình hơn và giao diện người dùng đồ họa (GUI) nhiều hơn một chút. Với mục đích của cuốn sách này, R sẽ đề cập đến phiên bản 3.6.2 và RStudio là phiên bản 1.2.5019. Điều này sẽ cung cấp một số tiêu chuẩn hóa cho các màn hình và chức năng khác nhau, nhưng chúng tôi nhận thấy rằng chức năng có thể khác nhau nhưng chưa bao giờ giảm với các phiên bản sau này. Chẳng hạn, "GGobi" là một chức năng dường như không hoạt động với các phiên bản Rattle gần đây, nhưng chúng tôi cũng nhận thấy rằng

"GGRaptr" cũng hoạt động tốt, vì vậy GGobi đã được thay thế và có một số việc phải làm từ phía các nhà phân tích để đi đến kết luận đó. Khi thực hiện các tham chiếu này, có một điểm quan trọng mà bất kỳ ai sử dụng R đều phải hiểu.

GGRaptr và GGobi là một phần của hàng nghìn "gói" theo đúng nghĩa đen có sẵn để hoạt động với R. Các gói này nằm trên mạng CRAN hoặc các mạng được liên kết là một phần của nỗ lực nguồn mở này. Cuốn sách sẽ chỉ cho bạn cách cài đặt các gói này và cung cấp chúng cho phân tích của bạn.

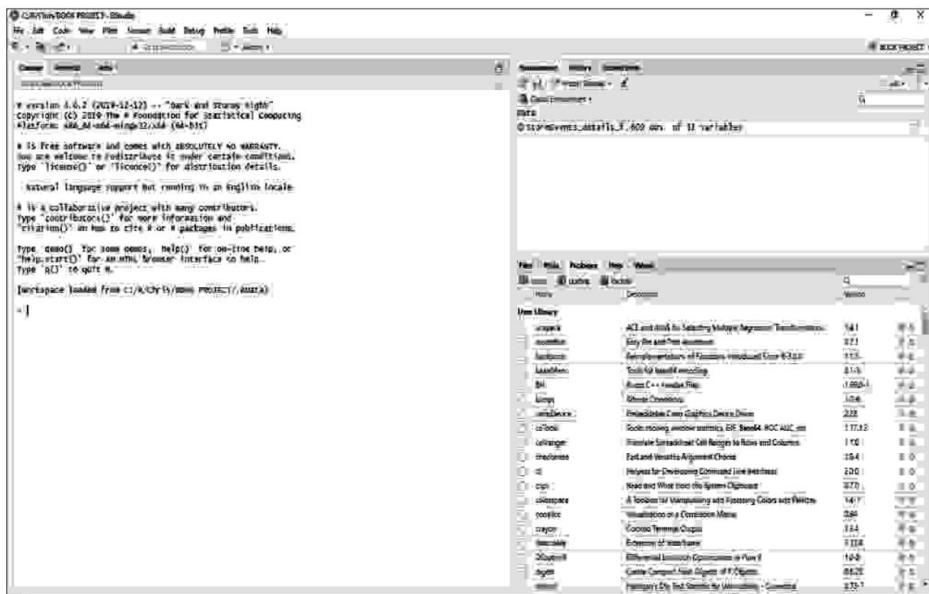
Các gói này mạnh mẽ và năng động đến mức một số trong số chúng được tạo riêng cho một số bài kiểm tra thống kê trong cuốn sách này. Tuy nhiên, như nhà phân tích sẽ thấy với R, không phải mọi thứ đều diễn ra như một bữa tiệc tự chọn; một số mặt hàng phải được nấu chín.

2.4 NHẬP VÀO RSTUDIO

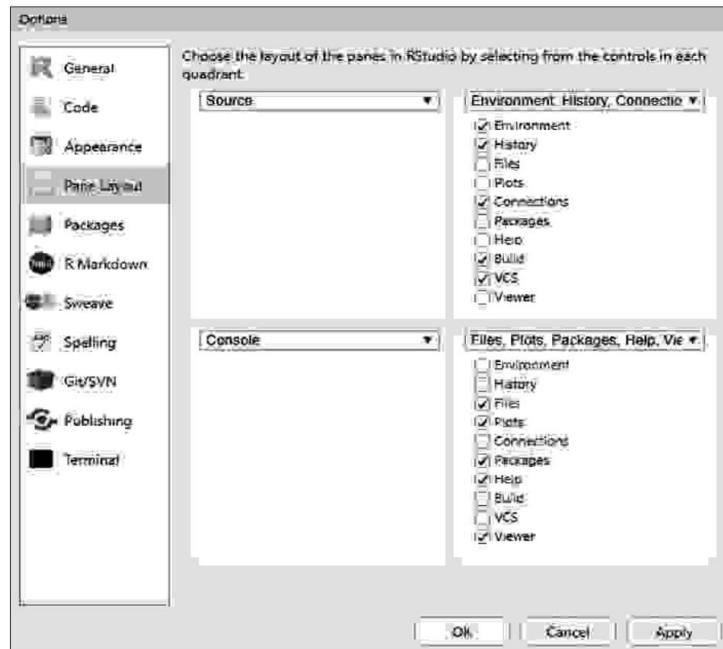
Khi RStudio được cài đặt và mở lần đầu tiên, màn hình môi trường làm việc mặc định này sẽ xuất hiện. Có một số điều quan trọng cần biết trước khi thực hiện bất kỳ nỗ lực nhập nào. Đầu tiên, cài đặt RStudio có vào thư mục "tài liệu" hay ở đĩa "C" không? Điều này có thể tạo ra sự khác biệt trong cách RStudio phản hồi một số lệnh và "gói". Để loại bỏ mọi sự cố có thể xảy ra với R hoặc RStudio, bạn nên khởi động ứng dụng với tư cách quản trị viên nếu đó là Hệ điều hành Windows. Bằng cách này, ứng dụng sẽ tự động có quyền truy cập vào các tệp nằm trên các thư mục và tệp được bảo vệ.

Khi tải xuống một sản phẩm nguồn mở, vui lòng đảm bảo rằng có phần mềm chống vi-rút đang hoạt động trên máy của bạn. Ngoài ra, hãy quét tệp thực thi đã được tải xuống trước khi kích hoạt sản phẩm. Cuối cùng, nếu kế hoạch là thực hiện phân tích trực tuyến, hãy đảm bảo có một Mạng riêng ảo (VPN) đang hoạt động được mua và hoạt động trên máy. Có rất nhiều VPN có sẵn trực tuyến, vì vậy hãy chọn một và sử dụng nó. Điều này sẽ ngăn chặn bất kỳ sự xâm nhập tích cực nào có thể xảy ra trong khi làm việc với ứng dụng nguồn mở. Mọi người sẽ tránh nguồn mở vì những lý do này, nhưng hiểu rằng một số ứng dụng thống kê đã có một số vấn đề về bảo mật, vì vậy hãy chuẩn bị sẵn sàng và điều đó sẽ ngăn chặn bất kỳ rủi ro nào có thể xảy ra với các sản phẩm phần mềm này.

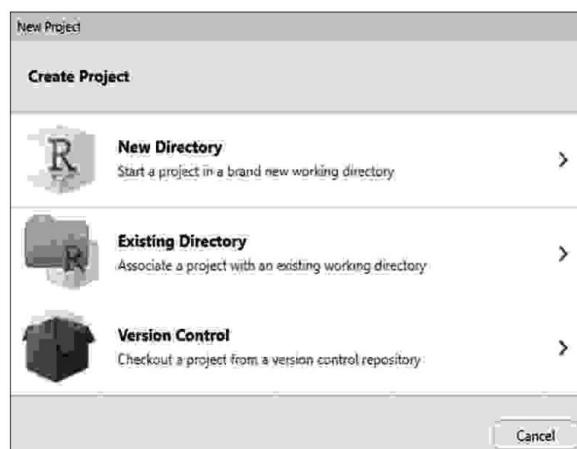
Bây giờ, hãy tiếp tục với quá trình nhập cho R và RStudio. Khi cài đặt xong R và RStudio, lần đầu tiên mở RStudio lên, màn hình sẽ xuất hiện như sau:



Mỗi khu vực này trên màn hình đại diện cho một "khung". Việc tùy chỉnh các ngăn này được thực hiện bằng cách nhấp vào "Chế độ xem" trên thanh công cụ trên cùng. Hãy giải thích từng khung một cách riêng biệt. Cái bên trái là ngăn "Bảng điều khiển" nơi thực hiện kết hợp chương trình. Mặc dù cuốn sách này không tập trung vào lập trình, nhưng đôi khi người phân tích phải nhập một số lệnh nhất định để thực hiện một nhiệm vụ. Khung này là nơi nó sẽ xảy ra. Ngăn bên trái này hoạt động như hai ngăn khi tệp được nhập. Tại thời điểm tệp được nhập, một ngăn khác sẽ xuất hiện được gọi là ngăn "Nguồn", ngăn này sẽ hiển thị toàn bộ tập dữ liệu. Thông tin thêm về điều này sau khi nhập. Hai khung bên phải hiển thị lịch sử của các lệnh được nhập (trên cùng) và các gói khác nhau đã được cài đặt (dưới cùng). Có các tab ở mỗi ngăn này áp dụng cho chức năng của từng ngăn. Điều tuyệt vời về RStudio (và có rất nhiều tính năng tuyệt vời về RStudio) là nếu bạn nhấp vào "Chế độ xem" và chọn "Bộ cục ngăn", bạn sẽ thấy màn hình sau, màn hình này có thể giúp bạn quyết định nơi bạn muốn từng phần của sơ đồ phát triển.



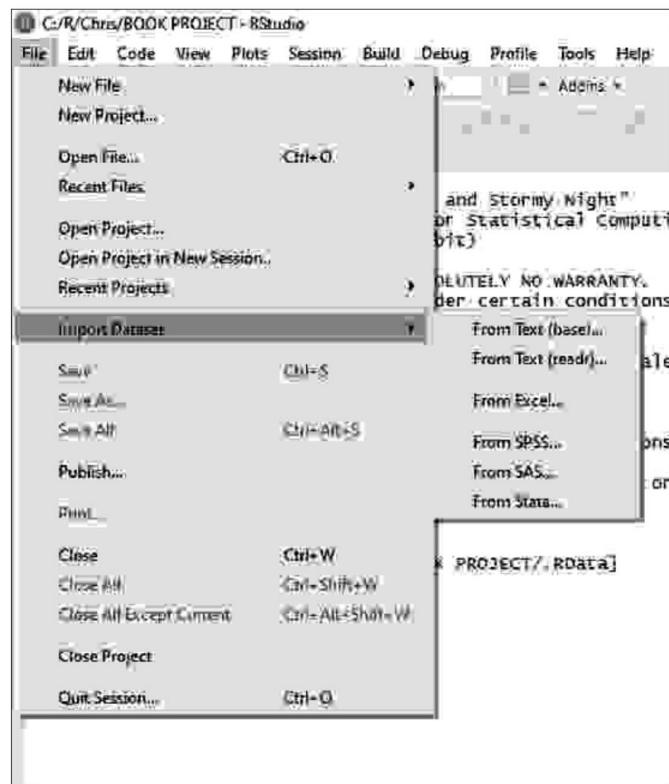
Một báo trước vào thời điểm này, nhưng báo trước này là tùy chọn. Trong khi sử dụng RStudio, bạn có thể đặt nơi bạn muốn dự án của mình được lưu trữ. Theo kinh nghiệm, một số nhà phân tích không lưu dự án của họ hoặc thậm chí không tạo dự án mà thay vào đó dựa vào RStudio để thực hiện việc đó một cách tự động. RStudio sẽ lưu các tệp vào thư mục R chính, nhưng bạn có thể lưu chúng vào một thư mục cụ thể hơn sẽ chứa tài liệu dự án của bạn. Phương pháp để mở một dự án mới và lưu dự án đó là chọn "Dự án mới" từ menu Tệp và bạn sẽ nhận được màn hình này.



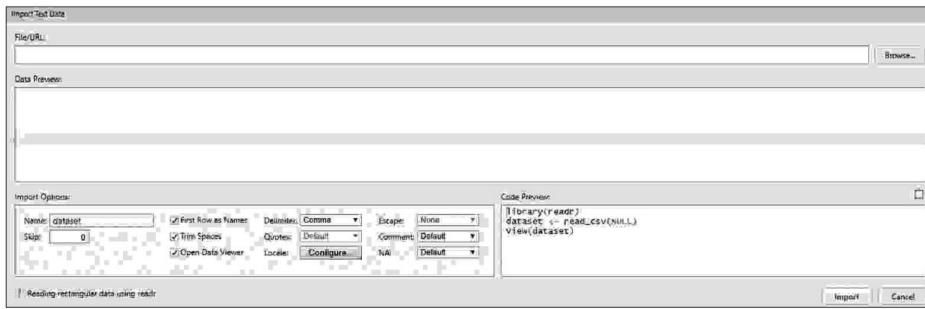
Các lựa chọn đều rõ ràng, vì vậy chúng tôi sẽ cho phép bạn khám phá nơi bạn muốn đặt các tệp dự án của mình. Khi bạn làm điều đó, RStudio sẽ mở trong dự án đó. Nếu bạn muốn nó mở một dự án khác, bạn đoán xem, bạn sử dụng lựa chọn “Mở” trong menu Tệp.

Những người đã sử dụng R trước đây có thể thích màn hình R “cơ bản” mà không có sự hỗ trợ của RStudio, điều này được đánh giá cao. RStudio sẽ hiển thị chương trình được tích hợp vào các lần nhấp chuột khác nhau, phần này sẽ được hiển thị sau. Đầu tiên, nhập dữ liệu là bước tiếp theo để RStudio (và Rattle) hoạt động.

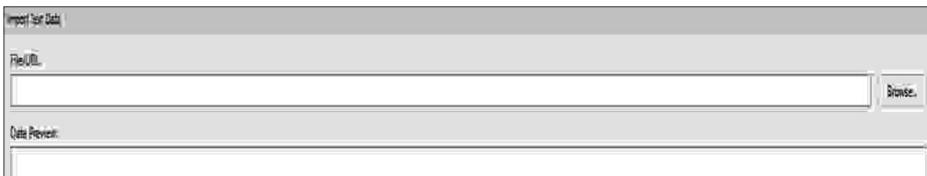
Để nhập dữ liệu vào RStudio, bạn chọn “nhập tập dữ liệu” từ Menu Tệp. Điều này được hiển thị trong ảnh chụp màn hình sau. Đảm bảo rằng “Tùy chỉnh (readr)” được chọn để bao gồm các tệp CSV đang được nhập.



Khi lựa chọn được chọn, màn hình sau sẽ xuất hiện với nhiều hộp văn bản trống. Việc xem xét những điều này một cách riêng biệt sẽ giúp hiểu được ý nghĩa của những hộp văn bản đó.



Có nhiều thành phần cho màn hình này, nhưng thành phần chính là hộp văn bản trên cùng nơi đặt tên tệp để truy xuất nó từ Internet hoặc máy tính của bạn. Đối với mục đích của cuốn sách này, trọng tâm sẽ là các tệp đã tải xuống tồn tại trên máy tính. Cùng một tệp được sử dụng trong các ví dụ trước, đó là Theo dõi Lốc xoáy năm 1951, cũng sẽ được sử dụng ở đây.



Sau khi tệp được chèn vào hộp “Tệp/URL”, thường thông qua việc sử dụng nút “Duyệt.”, thì tệp sẽ xuất hiện dưới dạng bản xem trước trong hộp văn bản lớn đang mở. Một nhà phân tích R khao khát có thể muốn biết chương trình cơ sở ming, và đó là trong hộp văn bản dưới cùng bên phải. Nếu một người có R, người đó có thể cắt và dán mã và nhận được kết quả tương tự, ngoại trừ việc tệp sẽ được lưu trong tệp R của bạn thay vì khu vực RStudio (hầu hết thời gian chúng giống nhau, do nhà phân tích cài đặt cả hai R và RStudio trong cùng một thư mục).

Sau khi tệp được nhập vào RStudio, nhà phân tích sẽ thấy tệp trong Ngăn tệp, trong trường hợp này là ở phía trên cùng bên trái của màn hình, được hiển thị như sau với màn hình chính trước và ngăn tệp thứ hai.

The screenshot shows the RStudio interface with the 'StormEvents_details_tidy_v1_0_d1951_c2...' dataset loaded into a data frame. The data frame contains the following columns:

BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EVENT_ID	STATE
1 195101	9	915	195101	9	915	10047282	MISS
2 195101	17	2200	195101	17	2200	10028729	KANS
3 195103	28	510	195103	28	510	10120421	TEXA
4 195105	9	1830	195105	9	1830	10088711	OKLA
5 195105	15	1620	195105	15	1620	10088712	OKLA
6 195106	8	1800	195106	8	1800	10018811	KANS
7 195107	16	1600	195107	16	1600	10141821	PENN
8 195107	11	1700	195107	11	1700	10141824	PENN
9 195107	27	2214	195107	27	2204	10094012	PENN
10 195107	30	1330	195107	30	1330	10094013	PENN
11 195108	29	1814	195108	28	1814	10088681	NEWY
12 195108	19	1830	195108	19	1830	10094012	OKLA
13 195108	3	1330	195108	3	1330	10093256	MICH

The screenshot shows the RStudio interface with the 'StormEvents_details_tidy_v1_0_d1951_c2...' dataset loaded into a data frame. The data frame contains the following columns:

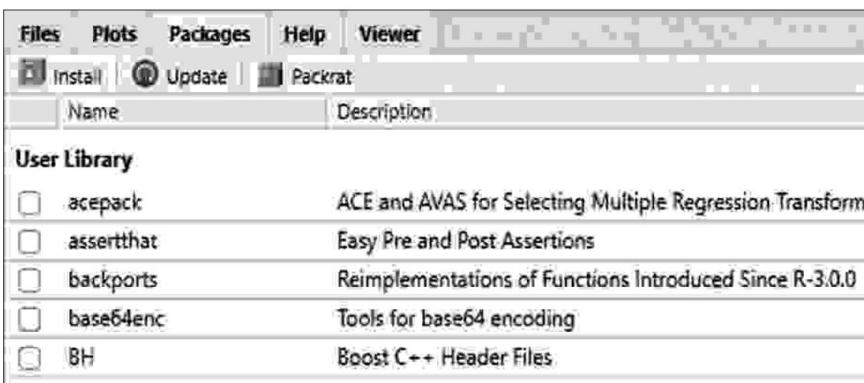
BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EVENT_ID	STATE
1 195101	9	915	195101	9	915	10047282	MISS
2 195101	17	2200	195101	17	2200	10028729	KANS
3 195103	28	510	195103	28	510	10120421	TEXA
4 195105	9	1830	195105	9	1830	10099717	OKLA
5 195107	15	1620	195107	15	1620	10099742	OKLA
6 195105	8	1800	195105	8	1800	10028691	KANS
7 195103	30	1500	195103	30	1500	10104933	PENN
8 195105	11	1330	195105	11	1330	10104934	PENN
9 195106	27	2204	195106	27	2204	10104935	PENN
10 195107	21	1100	195107	21	1100	10104936	PENN
11 195104	29	1815	195104	29	1815	10082587	NEWY
12 195102	19	1830	195102	19	1830	10099495	OKLA
13 195105	3	1335	195105	3	1330	10039190	MICH

Có một cảnh báo ở đây rất quan trọng khi sử dụng RStudio. Khi một tệp được nhập vào RStudio, tệp đó sẽ trở thành một "tibble". Đây là một thuật ngữ có nghĩa là tập dữ liệu thuộc một loại cụ thể và do đó sẽ cần một số gói R nhất định để nhanh chóng phân tích dữ liệu. Đừng lo lắng, vì tibble cũng được phân tích bằng các công cụ R thông thường, có thể được sử dụng thông qua RStudio.

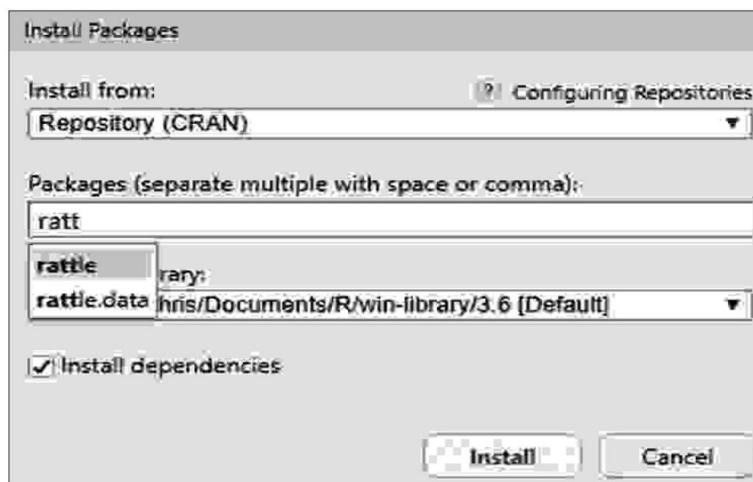
2.5 NHẬP KHẨU RATTLE

R có một gói đặc biệt mạnh mẽ gọi là Rattle, hữu ích đến mức nó phải được tách ra khỏi R trong khi mô tả việc nhập (hoặc bất kỳ chức năng nào khác cho vấn đề đó). Quá trình cài đặt Rattle bắt đầu với ngăn RStudio có tên là “Tệp, Lô, Gói và Trợ giúp” (ngăn phía dưới bên phải). Như được mô tả trong màn hình tiếp theo, phần này chứa một số gói đã được cài đặt trong ứng dụng R và sau đó là RStudio. Khi cài đặt R và RStudio lần đầu tiên, số lượng gói sẽ được giới hạn ở những gói được bao gồm trong bản cài đặt đó. Các gói khác được cài đặt riêng biệt hoặc kết hợp với các gói khác để kích hoạt gói chính đang được cài đặt. Tất cả điều này nghe có vẻ khó hiểu, vì vậy, việc mô tả quy trình cài đặt Rattle sẽ nhanh chóng giải quyết vấn đề này.

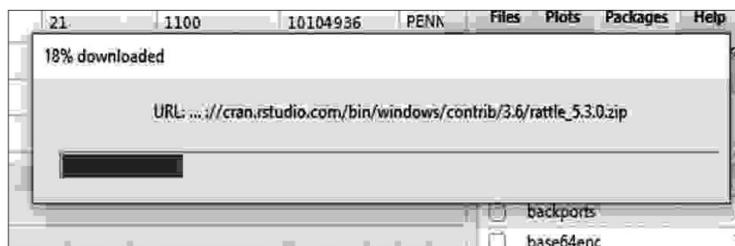
Bước đầu tiên khi cài đặt gói là đảm bảo tab “Gói” được chọn như minh họa trong phần sau. Hãy nhớ rằng ngăn này nằm ở dưới cùng bên phải của môi trường làm việc RStudio. Lưu ý nút “cài đặt” ở phía trên bên trái màn hình. Đây là cái chúng tôi sẽ sử dụng để cài đặt các gói.



Khi chọn “cài đặt”, cửa sổ bật lên sau sẽ xuất hiện, hiển thị máy chủ CRAN nơi lưu trữ gói (và có thể tải xuống và cài đặt) cùng với hộp văn bản trống cho gói. Hãy cẩn thận: máy tính phải được kết nối với Internet nếu không phần này sẽ bị lỗi. Bắt đầu nhập Rattle vào hộp văn bản trống và không cần kết thúc từ, “rattle” sẽ xuất hiện. Lưu ý rằng “Cài đặt phụ thuộc” được chọn. Điều này rất quan trọng vì nhiều gói có các gói con độc lập, nhưng gói này có các liên kết để hoạt động. Để lại điều này ở chế độ mặc định của nó bây giờ.



Nhấp vào nút “Cài đặt” và sẽ có một loạt hoạt động ở ngăn dưới cùng bên trái của RStudio. Điều này tốt vì điều đó có nghĩa là RStudio đã tìm thấy máy chủ chứa gói và đang tải xuống cũng như cài đặt gói.



```

Console | Terminal | Jobs |
C:/R/Chris/BOOK.RPROJECT>
downloaded 5.1 MB

package 'rattle' successfully unpacked and MD5 sums checked.

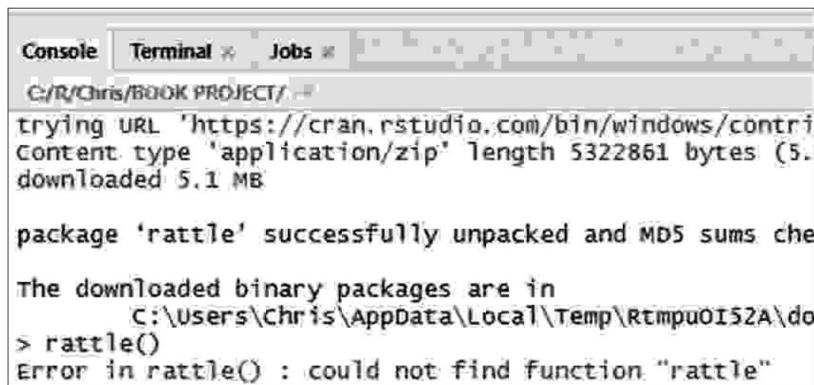
The downloaded binary packages are in
  C:\Users\Chris\AppData\Local\Temp\Rtmpu0152A\downloaded_packages
> install.packages("rattle")
Installing package into 'C:/users/chris/documents/R/win-library/3.6'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/rattle_5.3.0.zip'
content type 'application/zip' length 5322861 bytes (5.1 MB)
downloaded 5.1 MB

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\Chris\AppData\Local\Temp\Rtmpu0152A\downloaded_packages
> |

```

Bây giờ Rattle đã được cài đặt, vẫn còn một bước nữa phải hoàn thành, đó là kích hoạt gói trên R và RStudio. Nếu nhà phân tích gõ "Rattle" trên màn hình mà không tải gói, thông báo sẽ rõ ràng.



```

Console Terminal * Jobs *
C/R/Chris/BOOK PROJECT/ ...
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/rattle_1.3.10.zip'
Content type 'application/zip' length 5322861 bytes (5.1 MB)
downloaded 5.1 MB

package 'rattle' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
  C:/Users/Chris/AppData/Local/Temp/RtmpuOIS2A/down
> rattle()
Error in rattle() : could not find function "rattle"

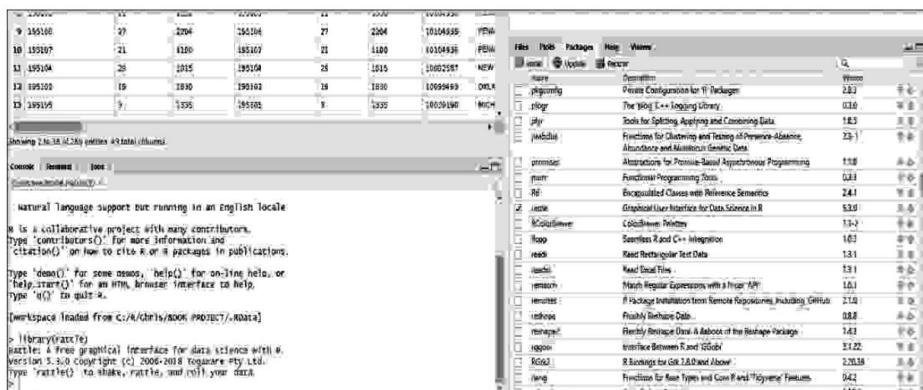
```

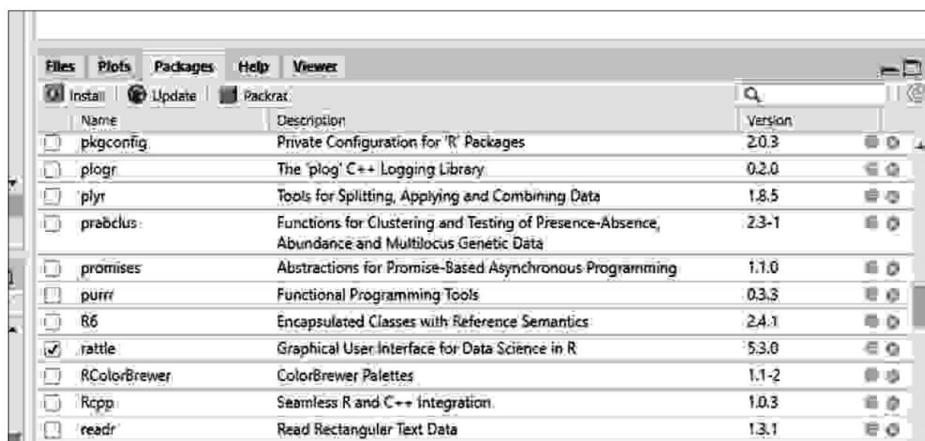
Rattle phải được tải vào R để nó hoạt động. Để làm điều này là đơn giản.

Một cách là gõ như sau trong R:

```
> thư viện (lúc lắc)
```

Một cách khác là sử dụng tab "gói" trong màn hình ở dưới cùng bên phải (trong cấu hình khung xem của cuốn sách này) và đánh dấu vào hộp kiểm bên cạnh Rattle (như minh họa trong màn hình sau). Vì RStudio được gắn vào R, nên mã sẽ xuất hiện như thẻ bằng phép thuật trong R.





Màn hình trước hiển thị tab Gói có tiếng kêu được chọn. Thời điểm một nhà phân tích thực hiện lựa chọn này, ngăn lập trình sẽ hoạt động như sau và tải gói Rattle, cùng với mọi phụ thuộc có thể đi kèm với gói chưa được cài đặt lần đầu tiên. Theo một số cách, R và RStudio dự đoán những gì nhà phân tích sẽ yêu cầu trước khi họ cần.

```

Console Terminal Jobs
C:/R/Chris/BOOK PROJECT

Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from C:/R/Chris/BOOK PROJECT/.RData]

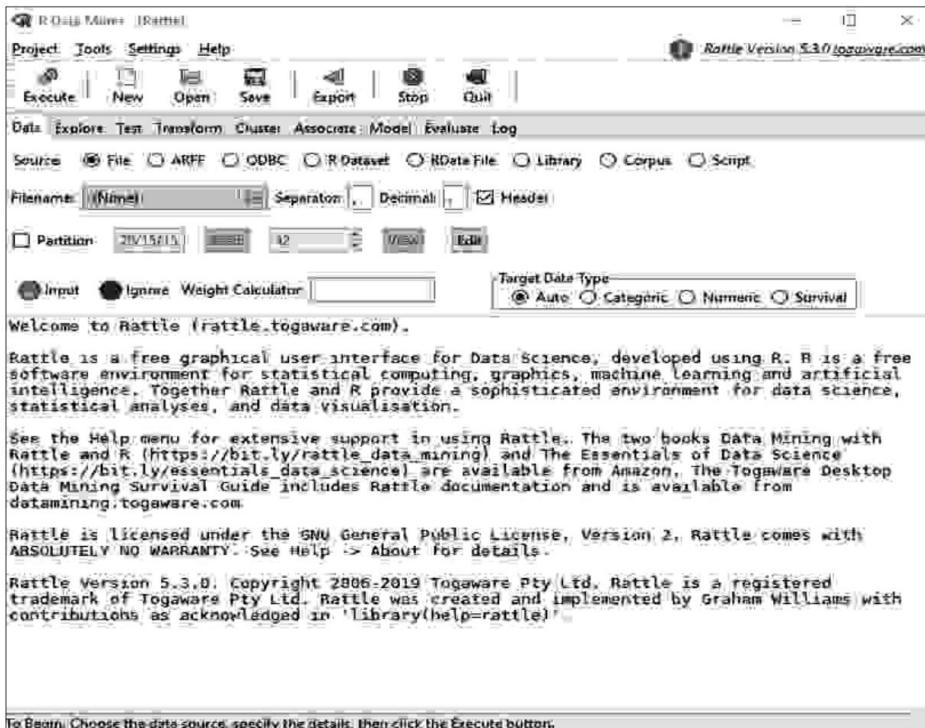
> library(rattle)
Rattle: A free graphical interface for data science with R.
Version 5.3.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data!
>

```

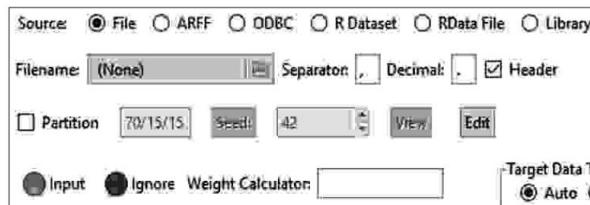
Để kích hoạt Rattle, hãy nhập thông tin sau vào ngăn lập trình:

```
> lach cach()
```

Tại thời điểm này, nhà phân tích đã cài đặt và tải gói, vì vậy Rattle sẽ hiển thị màn hình đầu tiên trong một cửa sổ riêng biệt sẽ xuất hiện dưới dạng:

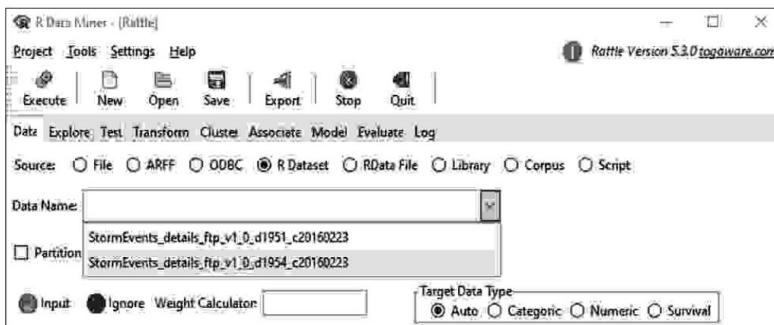


Màn hình này là màn hình chính của Rattle và là nơi thực hiện các chức năng của công cụ. Bước đầu tiên là nhập dữ liệu vào công cụ này. Đây là nơi R và Rattle được liên kết. Sau khi dữ liệu được nhập vào R (hoặc RStudio trong trường hợp này), thì dữ liệu đó sẽ được cung cấp cho tất cả các công cụ khác, trong trường hợp này là Rattle. Để nhập dữ liệu vào Rattle, hãy sử dụng hộp “Tên tệp” trong màn hình chính như sau:

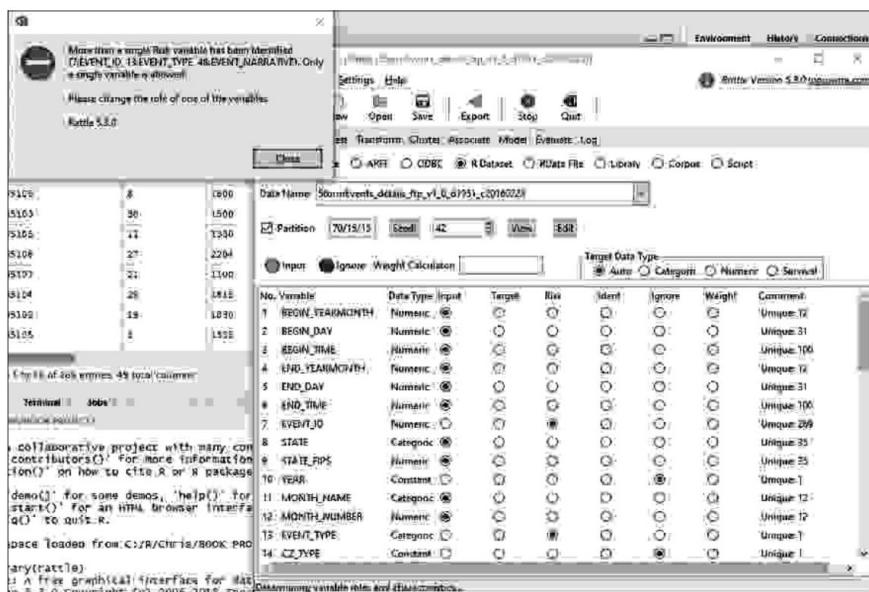


Trong trường hợp này, sử dụng lựa chọn nút radio của “Tệp” buộc bạn phải tiết lộ tên tệp để nhập dữ liệu. Sử dụng cùng một vị trí của dữ liệu bạn đã làm cho OpenOffice và đảm bảo rằng “Đầu phân cách” là dấu phẩy (vì đây là tệp CSV). Ngoài ra, hãy đảm bảo rằng “Tiêu đề” được chọn, vì dữ liệu này có tiêu đề.

Tuy nhiên, vì dữ liệu đã được tải vào R, nhà phân tích có thể chọn nút radio “Bộ dữ liệu R” như sau để hiển thị tập dữ liệu đã có trong R. Chọn tệp đầu tiên và nhấp vào “Thực thi” trên thanh công cụ biểu tượng đầu tiên và dữ liệu sẽ được nhập vào Rattle.



Cho dù nó có vẻ dễ dàng đến đâu, thì luôn có một số thay đổi về cấu hình đối với tập dữ liệu để làm cho nó dễ sử dụng hơn đối với Rattle. Trong trường hợp này, khi dữ liệu được nhập (thực thi), một thông báo cảnh báo sẽ xuất hiện như sau. Điều này có thể dễ dàng khắc phục bằng cách chỉ chọn một biến rủi ro thay vì có nhiều biến mà Rattle đã chọn.



Vì việc chọn một biến là đầu vào, mục tiêu, rủi ro, nhận dạng, bỏ qua hoặc trọng số được thực hiện bằng một cú chạm chuột, nên rất dễ khắc phục điều này bằng cách chỉ chọn một biến làm rủi ro. Nhưng trước khi điều này được thực hiện, cần phải giải thích một chút để mô tả các loại biến khác nhau mà nhà phân tích sẽ liên kết với từng loại dữ liệu này.

Bảng sau đây mô tả ngắn gọn về từng loại dữ liệu này được lấy từ CRAN. Khi từ “động” được sử dụng, điều đó có nghĩa là những điều này có thể được thay đổi bởi nhà phân tích bất cứ khi nào một mô hình hoặc đánh giá khác được thực hiện. Nhà phân tích chỉ cần thay đổi lựa chọn nút radio và nhấp lại vào “Thực thi”. Bộ dữ liệu được tự động thay đổi. Điều quan trọng ở phần cuối cùng này là bất cứ lúc nào nhà phân tích muốn thay đổi các biến số mục tiêu hoặc rủi ro, một thao tác quay lại đơn giản đổi với tập dữ liệu và Execute sẽ thay đổi các biến của tập dữ liệu. Trong một số trường hợp, như sẽ được giải thích sau trong văn bản này, một số biểu đồ cho phép thay đổi tương tác sẽ tự động thay đổi mục tiêu hoặc các yếu tố rủi ro trong tập dữ liệu. Điều tốt nhất vẫn chưa đến.

Kiểu	Sự miêu tả
Đầu vào	Các biến độc lập
Mục tiêu	Các biến phụ thuộc
Rủi ro	Giá trị động được sử dụng trong biểu đồ rủi ro
nhận dạng	Có một giá trị duy nhất cho mỗi bản ghi
Phớt lờ	Các biến bị loại bỏ
Cân nặng	Các biến được đánh giá cao hơn các biến khác để thể hiện tầm quan trọng

Bây giờ tệp đã được nhập vào RStudio và Rattle, công cụ tiếp theo được sử dụng để nhập sẽ là KNIME.

2.6 NHẬP VÀO KNIME

KNIME là một công cụ phân tích dữ liệu được phát triển ở Châu Âu và, theo kinh nghiệm của tác giả này, kết hợp phân tích thống kê thông thường với quy trình kỹ thuật hệ thống. Công cụ này có các “nút” hoặc mô-đun là các công cụ nhỏ chuyên đổi và phân tích khép kín để chia nhỏ tập dữ liệu và phân tích tập dữ liệu đó thành các thành phần mong muốn. Trước khi nhập, nhà phân tích phải tải xuống ứng dụng KNIME.

Hãy nhớ rằng cảnh báo tương tự cho công cụ mã nguồn mở này cũng được áp dụng. Đảm bảo rằng phần mềm chống vi-rút đang hoạt động và được kích hoạt, đồng thời bạn quét tệp exe có thể cát được trước khi kích hoạt phần mềm KNIME. Thận trọng là rất quan trọng ở đây, nhưng công cụ này sẽ xứng đáng với tất cả nỗ lực này.

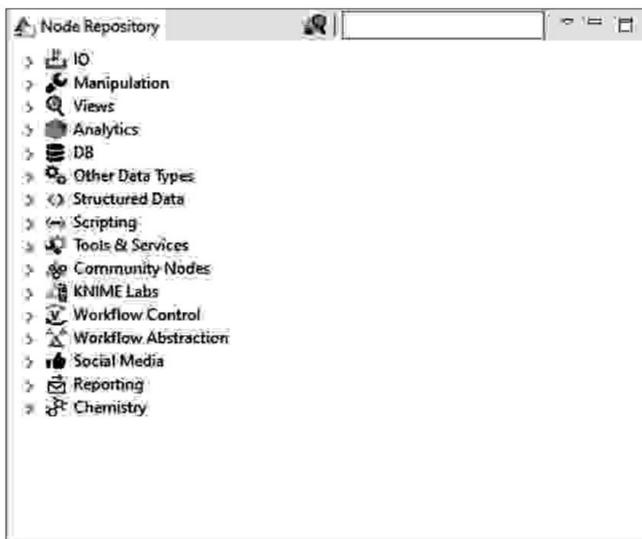
Sau hơn hai mươi năm thực hiện phân tích dữ liệu và giảng dạy tương tự, tác giả này chưa bao giờ thấy một công cụ nào như KNIME. Điều này không nên được hiểu là sự chứng thực kiên định của KNIME, nhưng đối với phân tích dữ liệu, công cụ này cung cấp nhiều biến thể của các "nút", khi được đặt trong một quy trình giống như biểu đồ dòng chảy, có thể cung cấp cho nhà phân tích một luồng chuyển đổi tập dữ liệu liên tục và phân tích. Chúng tôi thực sự khuyên rằng bất kỳ nhà phân tích nào ít nhất cũng nên dùng thử phần mềm này để xem nó có phù hợp với nhu cầu của bạn hay không.

Việc cài đặt tương đối dễ dàng. Truy cập trang web KNIME (www.knime.com) và tải xuống phiên bản mới nhất của ứng dụng. Đối với mục đích của cuốn sách này, phiên bản sẽ là 4.1.0, nhưng một lần nữa ứng dụng này có một số chức năng rất cơ bản sẽ không thay đổi trong tương lai gần, vì vậy đừng lo lắng về việc cập nhật văn bản này trong một thời gian.

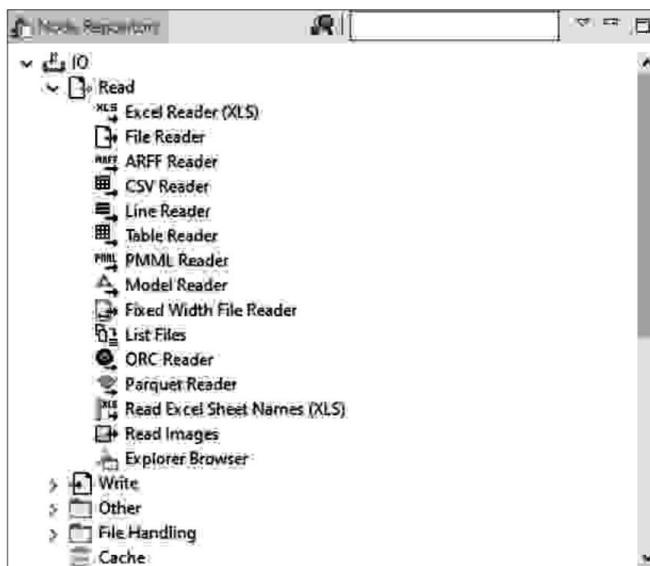
Một lưu ý: Đây là một ứng dụng rất "ngốn bộ nhớ", vì vậy nên có ít nhất 8 GB RAM và nếu máy tính có 12-16 GB, ứng dụng sẽ hoạt động nhanh hơn. Khi bạn tải xuống ứng dụng và mở nó, màn hình giật gân này sẽ xuất hiện. Có thể mất vài giây để điều này xuất hiện và nó có thể không chứa tất cả các biểu tượng có trong biểu tượng này, nhưng điều này sẽ được giải thích sau.

Sau khi mở ứng dụng KNIME sẽ xuất hiện màn hình sau với nhiều ngăn khác nhau. Hiện tại, ngăn "kho lưu trữ nút" sẽ là trọng tâm chính. Các ảnh chụp màn hình sau đây hiển thị màn hình giật gân đầu tiên, toàn bộ khu vực làm việc KNIME theo sau là ngăn kho lưu trữ nút. Một khuyến nghị là khám phá kỹ lưỡng công cụ KNIME, vì có nhiều cách để kết hợp các nút hoặc mô-đun này thành một luồng quy trình, như sẽ được trình bày sau chỉ với một hoặc hai trong số các nút này.



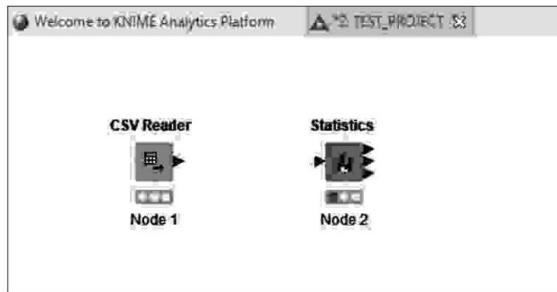


Phần “IO” của kho lưu trữ nút sẽ là phần sẽ nhập tập dữ liệu để phân tích. Vì tệp gốc là tệp CSV nên đó là tệp sẽ được nhập vào KNIME. Các màn hình sau hiển thị quá trình chọn và nhập dữ liệu. Vui lòng đặc biệt chú ý đến các tùy chọn khác có sẵn để nhập để xem vô số lựa chọn từ KNIME.

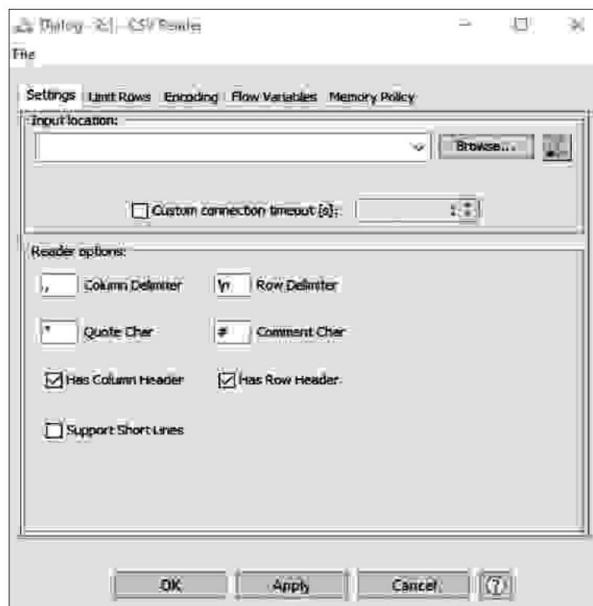


Như bạn có thể thấy, “Trình đọc CSV” là một trong các nút KNIME. Nút đang hoạt động trong không gian làm việc bằng cách nhấp và kéo nút vào đó

không gian làm việc. Sau khi hoàn thành, không gian làm việc sẽ xuất hiện như thế này (với một nút được thêm vào để tạo hiệu ứng). Nút “Trình đọc CSV” (hoặc nút 1) có chỉ báo màu vàng, có nghĩa là dữ liệu không có sẵn hoặc chưa được “làm sạch”, nghĩa là cần phải cấu hình dữ liệu để đèn chuyển sang màu xanh lục.



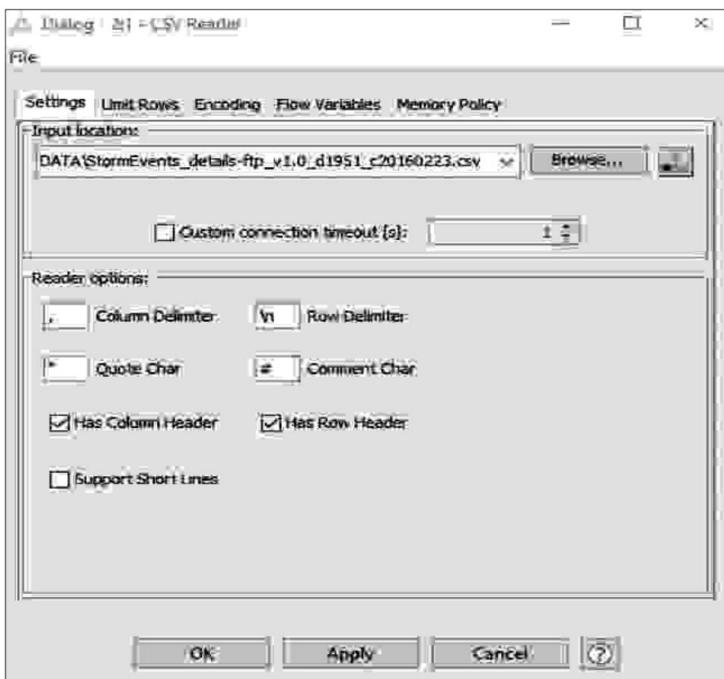
Cũng giống như một lưu ý, nhà phân tích cũng có thể đặt tên cho không gian làm việc như khi mở một dự án mới trong R. Điều này sẽ được thảo luận sau; hiện tại, nhập đúp vào nút “Trình đọc CSV” sau khi đặt nó vào không gian làm việc chính và màn hình này sẽ xuất hiện. Lưu ý rằng có một số lựa chọn mặc định trong cấu hình, bao gồm cả những lựa chọn trong “Tùy chọn đầu đọc” và hiện tại những lựa chọn đó đều ổn. Điều cần thiết là “Có tiêu đề cột” và “Có tiêu đề hàng” được chọn và “Đầu phân cách cột” hiển thị dấu phẩy.



Tại thời điểm này, bước tiếp theo sẽ là sử dụng nút “Duyệt.” để tìm kiếm tệp được nhập trên máy tính của bạn. Khi đã xong, màn hình sẽ xuất hiện tương tự như hình ảnh bên dưới.

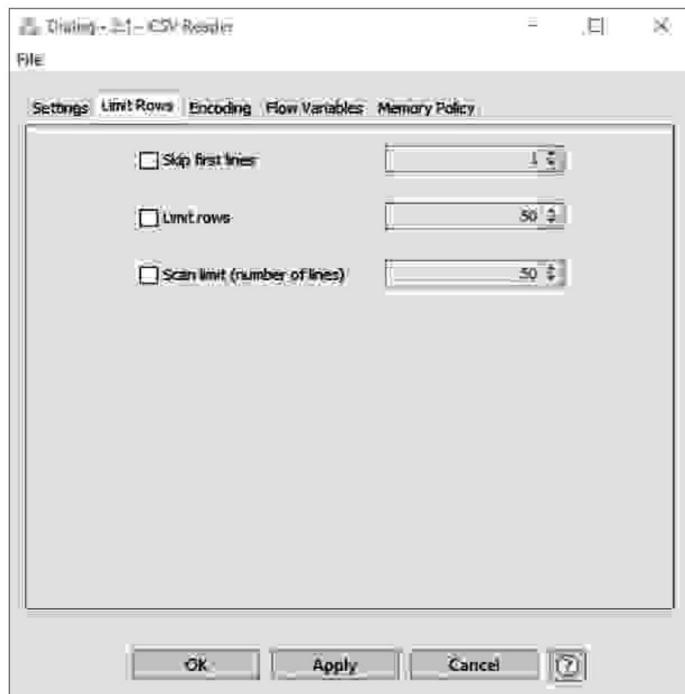
Sau khi hoàn tất, hãy nhấp vào nút “Áp dụng” hoặc nút “OK” để nhập dữ liệu. Tại thời điểm này, dữ liệu được nhập vào KNIME. Tuy nhiên, trước khi nhấp vào “Áp dụng”, hãy khám phá các tab khác để xem chúng là một phần của cấu hình tổng thể của tập dữ liệu này như thế nào.

Tab đầu tiên là tab “Giới hạn hàng”, được hiển thị như sau. Tab này sẽ hỗ trợ nhà phân tích xác định những hàng nào sẽ được đưa vào tập dữ liệu. Đây là một trong những khái niệm cơ bản của khoa học dữ liệu, đó là “hiểu dữ liệu” (một phần của CRISP-DM). Nếu nhà phân tích không hiểu những yêu cầu nào liên quan đến dữ liệu, sẽ rất khó để xác định dữ liệu nào là hữu ích. Trong trường hợp này, nhà phân tích có thể bỏ qua hoặc nhiều dòng đầu tiên như được xác định bởi nội dung dữ liệu, cùng với việc giới hạn số lượng hàng để quét.

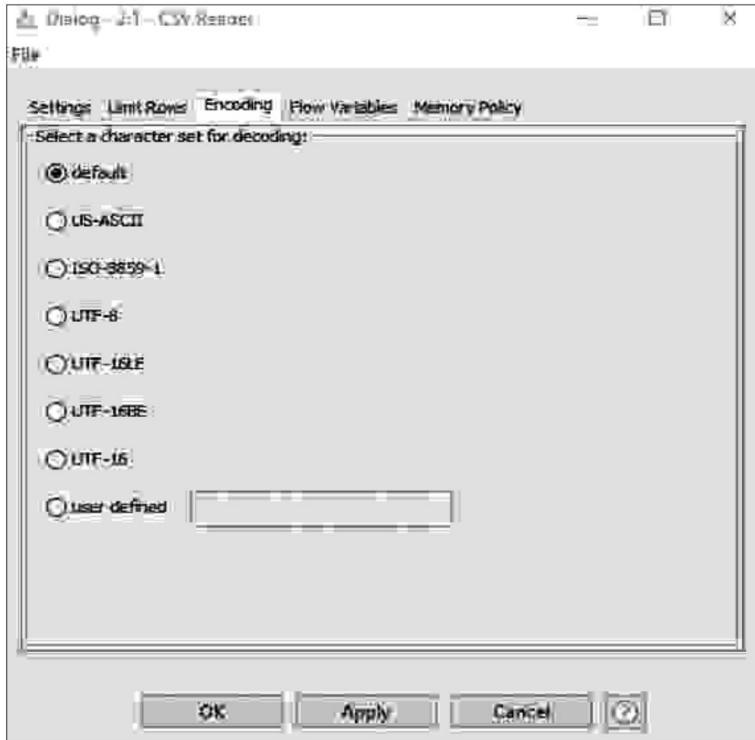


Tại sao một nhà phân tích sẽ làm điều này? Có những thuật ngữ liên quan đến việc chỉ lấy một phần dữ liệu để phân tích (chẳng hạn như “dữ liệu huấn luyện”), nhưng đủ để nói rằng việc thực hiện phân tích trên một phần dữ liệu nhanh hơn nhiều so với thực hiện trên tất cả dữ liệu và đánh giá có thể diễn ra sau đó trên một phần lớn hơn của dữ liệu. Tab này giúp thực hiện chức năng này mà không cần nỗ lực nhiều.

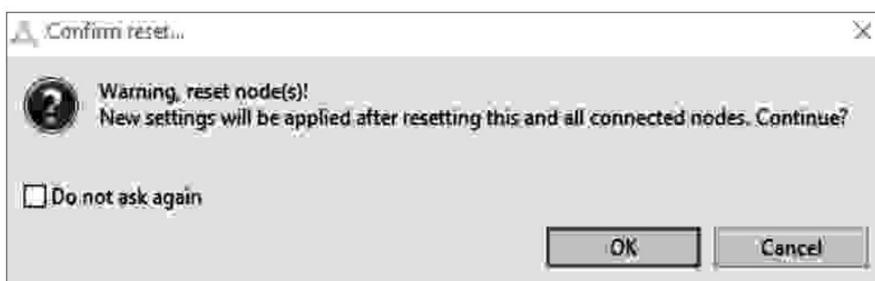
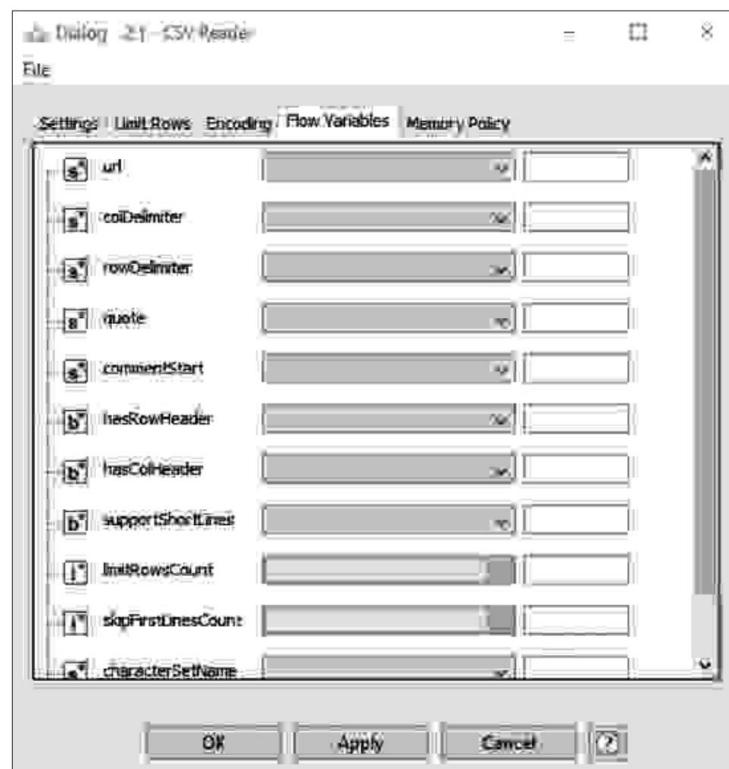
Loại tùy chỉnh cấu hình này chỉ là một tính năng giúp KNIME rất linh hoạt với tư cách là một sản phẩm phần mềm. May mắn thay, vì công cụ này là mã nguồn mở nên có nhiều trang web cộng đồng và nỗ lực cộng tác giúp mô tả các màn hình này và chức năng của chúng. Lý do cho sự chi tiết ở đây là có những lúc nhà phân tích cần liên hệ ngay lập tức giữa chức năng và khái niệm thống kê. Điều này được cung cấp ở đây.



Tab tiếp theo là tab “Mã hóa” được hiển thị như sau. Đôi khi mã hóa văn bản nhất định là quan trọng đối với lập trình hoặc các nỗ lực phân tích khác. Nhà phân tích có thể không bao giờ áp dụng các cài đặt này, nhưng điều quan trọng là phải biết vị trí của chúng trong trường hợp mã hóa cần được áp dụng cho tập dữ liệu. Trong hầu hết các trường hợp, nút radio “Mặc định” là nút được áp dụng, vì vậy tại thời điểm này không cần thay đổi lựa chọn đó. Tuy nhiên, trong trường hợp nhập văn bản thô, một số lựa chọn khác có thể giúp quá trình chuyển đổi sang KNIME mượt mà hơn và hữu ích hơn cho nhà phân tích.



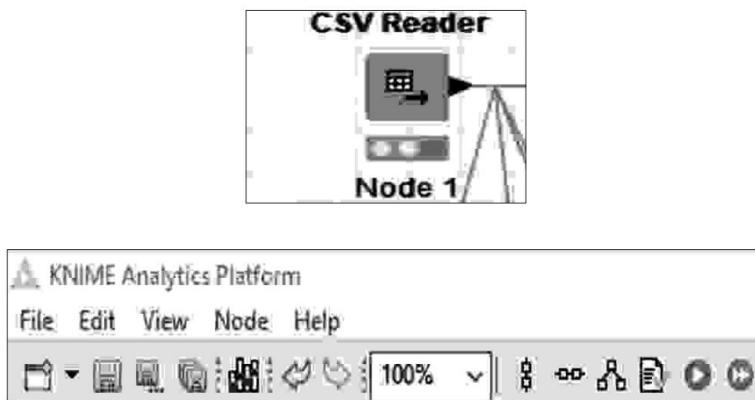
Tab tiếp theo là tab tồn tại trong ứng dụng KNIME có tên là "Biến dòng chảy". Mặc dù điều này sẽ không được sử dụng trong các ứng dụng thống kê này, nhưng chúng có thể được sử dụng trong các luồng nút trong tương lai. Đây là những giá trị do nhà phân tích xác định tồn tại để mỗi nút không phải được đặt riêng lẻ khi chúng tham chiếu đến tập dữ liệu. Cuốn sách này sẽ không đi sâu vào những điều này, nhưng các trang web KNIME đưa ra lời giải thích đầy đủ hơn về việc sử dụng các biến này. Những gì sẽ được nói về các biến lưu lượng là nhà phân tích có thể đặt chúng tại hộp văn bản bên cạnh mỗi biến. Điều này sẽ "buộc" biến trên từng nút tiếp theo trong sơ đồ luồng. Nhà phân tích cần nhớ rằng các biến luồng, sau khi được đặt và áp dụng, phải được xóa và tập dữ liệu được làm mới mỗi khi luồng được thực thi. Điều này được thực hiện đơn giản bằng cách nhấp vào nút "Áp dụng" sau mỗi lần cấu hình lại. Một thông báo sẽ xuất hiện cho nhà phân tích biết rằng nút đã được thay đổi, được hiển thị như sau.

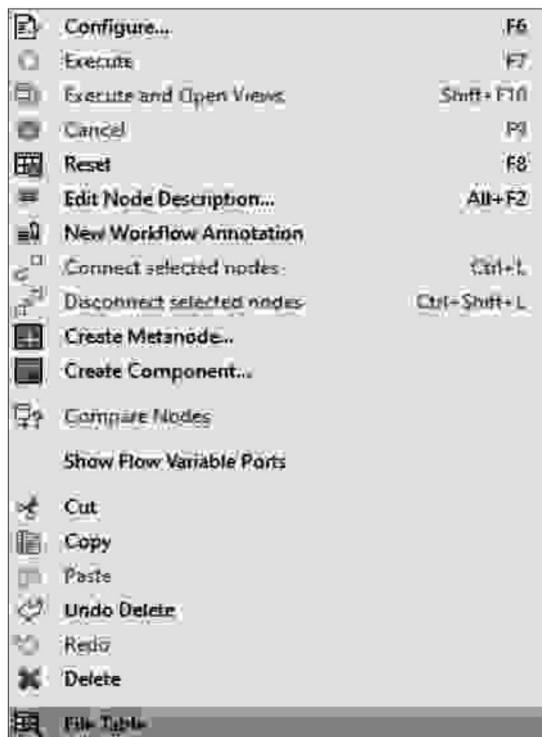


Sau khi nhập vào "OK", tập dữ liệu sẽ có cấu hình mới và quy trình phải được "Thực hiện lại". Tại thời điểm này, việc đánh giá là cần thiết đối với phương pháp "Đèn dừng" đối với KNIME. Nếu người đọc muốn biết thêm thông tin ngoài những gì tìm thấy trong cuốn sách này, thì các tài liệu tham khảo về tất cả các công cụ mã nguồn mở đều nằm trong phần Tài liệu tham khảo của cuốn sách này.

2.6.1 Tiếp cận đèn giao thông

Mỗi nút trong KNIME được quản lý bằng cách nhìn vào nút để xem trạng thái của nút đó. Nếu nút có đèn “xanh lục”, nút đó đã được thực thi hoặc kích hoạt trong luồng. Nút “Trình đọc CSV” sau đây có đèn xanh, cho biết nút này đã được thực thi. Nếu có bất kỳ thay đổi nào trong “Cấu hình” của nút này, thì đèn sẽ chuyển sang màu “vàng” hoặc có “tam giác thận trọng” bên dưới nút. Nếu đèn màu vàng, thì nhấp chuột phải vào nút và nhấp vào lựa chọn “Thực thi” sẽ thực thi nút cục bộ (điểm là tâm điểm của nút tại thời điểm đó). Để thực hiện TẤT CẢ các nút được gắn vào một tiêu điểm, hãy chuyển đến thanh công cụ chính và nhấp vào biểu tượng “mũi tên kép” được hiển thị như sau để thực hiện TẤT CẢ các nút được gắn vào nút đã chọn. Nếu một nút có hình tam giác thận trọng, thì điều đó có nghĩa là có điều gì đó không ổn với nút “cung cấp” cho nút thận trọng hoặc có gì đó không ổn với cấu hình. Cách tốt nhất để giảm hoặc loại bỏ các nút thận trọng này là đảm bảo cấu hình trên nút cấp liệu là chính xác bằng cách kiểm tra nút thông qua nhấp chuột phải vào nút cấp liệu và xem kết quả của nút đó. Một ví dụ về điều này sau đây.





Trong màn hình trước, “Bảng tệp” được chọn, vì đó là kết quả của nút. Điều này sẽ hiển thị bảng là kết quả của việc thực hiện nút đó. Sau khi được xác nhận, nút sẽ hiển thị cấu hình chính xác cho dòng quy trình. Nếu không chính xác, nhấp đúp vào nút để hiển thị màn hình cấu hình hoặc nhấp chuột phải và chọn “Configure.” để vào màn hình cấu hình.

CHƯƠNG 3

KIỂM TRA THỐNG KÊ

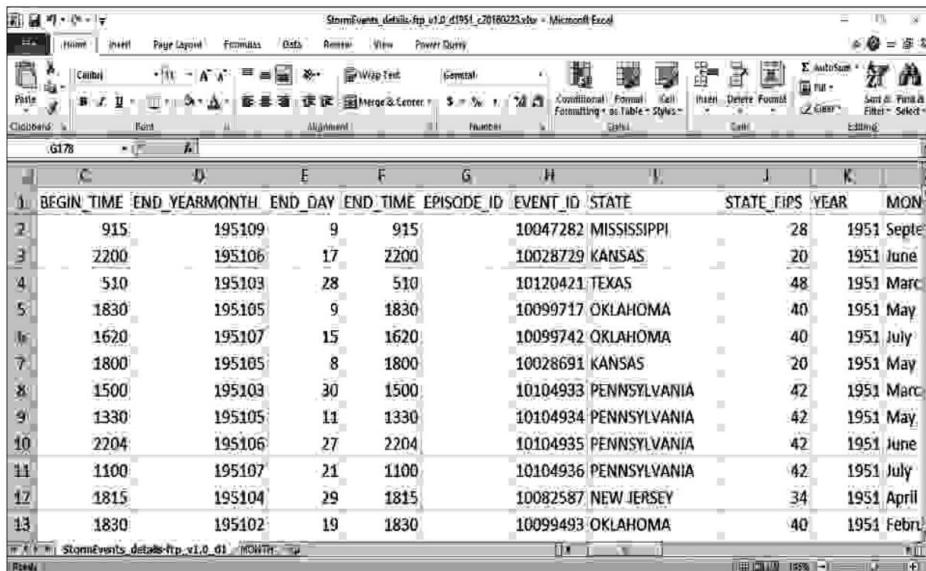
3.1 THỐNG KÊ MÔ TẢ

Chủ đề thường được giới thiệu trong thống kê là thống kê mô tả.

Nhiều học sinh đã được tiếp xúc với nhiều trong số này, bao gồm giá trị trung bình, trung bình, chê độ, phương sai và độ lệch chuẩn. Như đã hứa, cuốn sách này sẽ không đi sâu vào các công thức cho những điều này hoặc buộc học sinh phải làm chúng bằng tay. Lý do chính để nêu chúng ở đây là để áp dụng từng công cụ khoa học dữ liệu để hiển thị các thống kê mô tả này, hy vọng là với một chức năng trong công cụ.

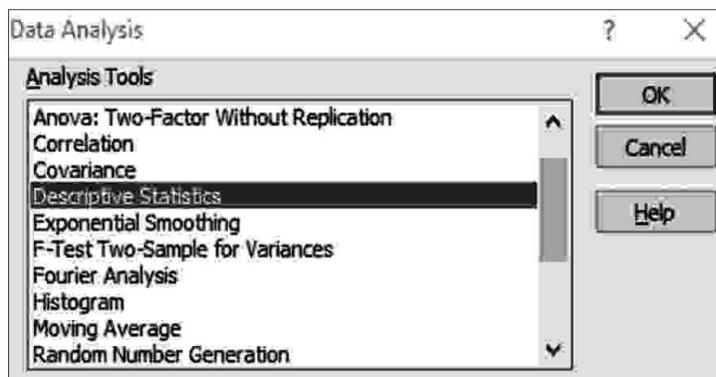
3.1.1 Excel

Excel, như đã thảo luận trước đây, có chức năng kỳ diệu này được gọi là ToolPak Phân tích, sẽ cung cấp cho nhà phân tích các số liệu thống kê mô tả mà không cần nhập từng phép tính vào ứng dụng. Bước đầu tiên là mở bộ dữ liệu và sau đó mở ToolPak Phân tích, như được mô tả trong phần sau hai màn hình.



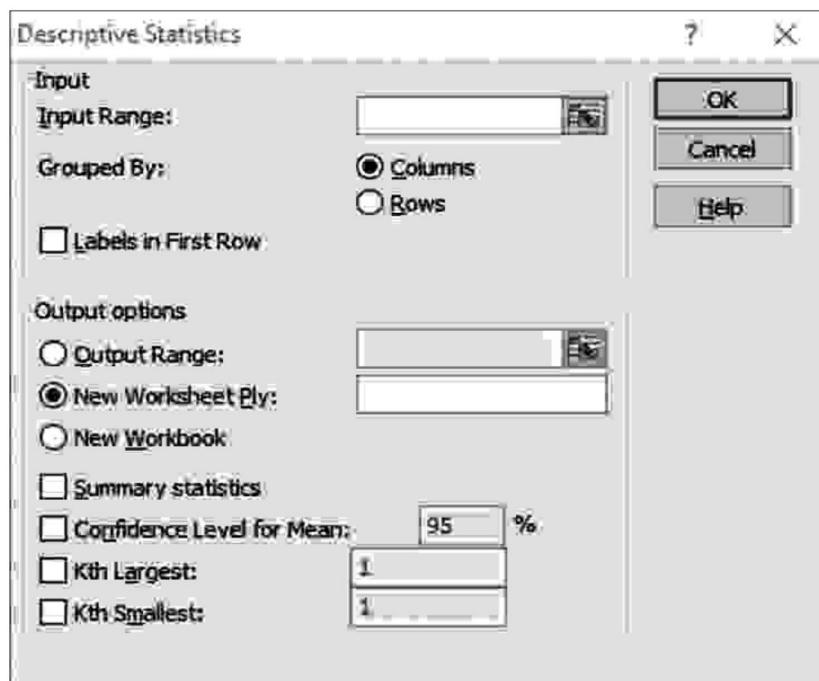
The screenshot shows a Microsoft Excel spreadsheet titled "StormEvents_details-ftp.v1.0_1951_20180223.xlsx". The data is organized into columns labeled C through K. Column C contains dates in the format "DDMMYY", while columns D through K contain various event details such as year, month, day, time, episode ID, event ID, state, state FIPS code, year, and month. The data spans from row 2 to row 13, showing events from September 1950 to February 1951.

	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MON
2	915	195109	9	915		10047282	MISSISSIPPI		28	1951 Septe
3	2200	195106	17	2200		10028729	KANSAS		20	1951 June
4	510	195103	28	510		10120421	TEXAS		48	1951 Marc
5	1830	195105	9	1830		10099717	OKLAHOMA		40	1951 May
6	1620	195107	15	1620		10099742	OKLAHOMA		40	1951 July
7	1800	195105	8	1800		10028691	KANSAS		20	1951 May
8	1500	195103	30	1500		10104933	PENNSYLVANIA		42	1951 Marc
9	1330	195105	11	1330		10104934	PENNSYLVANIA		42	1951 May
10	2204	195106	27	2204		10104935	PENNSYLVANIA		42	1951 June
11	1100	195107	21	1100		10104936	PENNSYLVANIA		42	1951 July
12	1815	195104	29	1815		10082587	NEW JERSEY		34	1951 April
13	1830	195102	19	1830		10099493	OKLAHOMA		40	1951 Febr

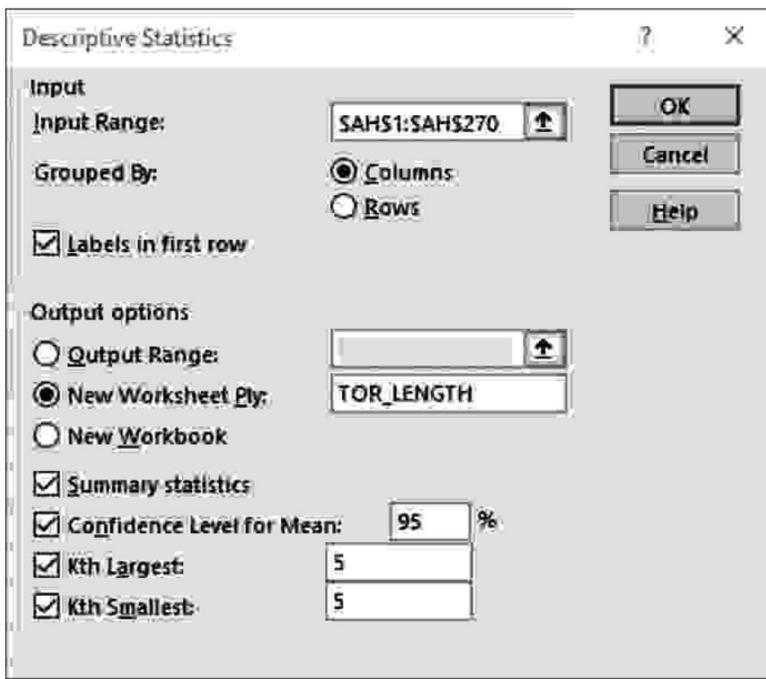


Khi nhà phân tích chọn biểu tượng “Phân tích dữ liệu” ở ngoài cùng bên phải của màn hình trước đó (sau khi chọn tab “Dữ liệu”), màn hình bật lên sẽ xuất hiện với nhiều tùy chọn để sử dụng các chức năng phân tích dữ liệu.

Trọng tâm của chương này là lựa chọn “Thống kê mô tả” (màu xanh lam). Người phân tích sẽ chọn tùy chọn này và nhấp vào nút OK, màn hình tiếp theo sẽ xuất hiện. Ngay lập tức, nhà phân tích sẽ nhận thấy rằng có nhiều khoảng trống vẫn bắn cần điền, nhưng đây không phải là vấn đề miễn là có tập dữ liệu để áp dụng chức năng này.



Ô trống đầu tiên cần điền là cột hoặc nhiều cột cần phân giải thành thống kê mô tả. Nhà phân tích có thể thực hiện việc này theo cách thủ công hoặc bằng cách chọn các cột từ tập dữ liệu. Có một lưu ý ở đây, chủ yếu liên quan đến cột so với hàng. Nếu tập dữ liệu có tên cột và dữ liệu đi xuống cột và ĐÂY là dữ liệu bạn cần giải quyết, thì hãy đảm bảo lựa chọn "Nhóm theo:" có nút radio "Cột" được chọn. Nếu "Hàng" được chọn, thì phân tích sẽ được thực hiện theo hàng thay vì cột. Hầu hết các bộ dữ liệu được định cấu hình hoặc sắp xếp theo cột, vì vậy đó là lý do tại sao nút radio cột đã được chọn. Sau khi hoàn thành, bước tiếp theo là điền vào màn hình như sau để nhận thống kê mô tả cho TOR_LENGTH, về cơ bản là độ dài của cơn lốc xoáy từ lần đầu tiên nhìn thấy cho đến khi tan biến. Đảm bảo rằng "Thống kê tóm tắt" được chọn; nếu không, kết quả sẽ không như những gì nhà phân tích mong đợi.



Một lời cảnh báo là cần thiết vào thời điểm này. Lựa chọn mặc định trong “Tùy chọn đầu ra” là để chỉ định phạm vi đầu ra. Nếu nhà phân tích thực hiện việc này, kết quả sẽ nằm trong cùng một trang tính với tập dữ liệu. Điều này có thể chứng minh là lần át trang tính để nhà phân tích sẽ phải cuộn ra ngoài các ô của tập dữ liệu để xem kết quả chức năng. Bạn nên luôn sử dụng tùy chọn “New Worksheet Ply:” và đặt tên cho trang tính tương tự như tiêu đề trước đó. Điều này sẽ đảm bảo rằng bảng dữ liệu sẽ vẫn chỉ là tập dữ liệu, thay vì thêm các cột không cần thiết vào bảng tính đó.

Một mục khác cần thiết là hộp kiểm “Cấp ở hàng đầu tiên”, đây là mục quan trọng. Thông thường, tiêu đề cột (hoặc nhãn) rất quan trọng để đặt tên cho mỗi cột. Không chọn hộp này sẽ cho Excel biết rằng có dữ liệu, không phải tên, ở hàng đầu tiên. Điều đó có thể gây nguy hiểm nếu có tiêu đề cột trong hàng đó.

Lưu ý rằng có “95%” trong hộp văn bản bên cạnh “Mức độ tin cậy cho giá trị trung bình:” cho biết rằng, nếu tập dữ liệu này là một mẫu của tập dữ liệu lớn hơn, hàm này sẽ hiển thị cho bạn một phạm vi trong đó sẽ có một 95% khả năng nó sẽ chứa giá trị trung bình. Điều này sẽ được đề cập trong phần “Khoảng tin cậy” trong phần tiếp theo, nhưng cũng đủ để nói rằng điều quan trọng là phải chọn hộp này và đặt nó ở mức 95% (vì đây là mức tin cậy thông thường cho Thống kê).

"Lớn nhất thứ K" và "Nhỏ nhất thứ K" đã được kiểm tra và đánh dấu "5" để minh họa cách sử dụng tùy chọn này. Về cơ bản, điều này sẽ hiển thị các giá trị "Lớn thứ 5" và "Nhỏ thứ 5" trong tập dữ liệu. Điều này có thể hữu ích nếu nhà phân tích muốn tìm hiểu xem một giá trị nhất định được xếp hạng như thế nào trong số tất cả các giá trị khác.

Khi nhấp vào "OK", phần sau đây sẽ xuất hiện trong một trang tính bổ sung trong sổ làm việc Excel của bạn. Nhà phân tích sẽ ngay lập tức nhận thấy rằng có nhiều thuật ngữ dễ nhận biết và những thuật ngữ không.

	A	B
1		TOR_LENGTH
2		
3	Mean	4.443494424
4	Standard Error	0.623786189
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.23085418
8	Sample Variance	104.6703773
9	Kurtosis	25.67453191
10	Skewness	4.376062845
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.228144665
19		
20		

Kết quả này sẽ cung cấp cho nhà phân tích mô tả về dữ liệu, giống như việc nhìn thấy một người sẽ cung cấp mô tả về người đó. Xin nhắc lại, cuốn sách này không phải là sách sơ lược về thống kê nên sẽ không đi sâu vào chi tiết cụ thể của từng tựa sách này. Quan trọng nhất, hãy nhớ rằng chức năng này sẽ cung cấp cho bạn bản xem trước tốt ở một biến tập dữ liệu ngay cả trước khi bắt kỳ đồ thị hoặc biểu đồ nào được tạo.

3.1.2 Văn phòng mở

OpenOffice, mặc dù giống như Excel, nhưng không có ToolPak Phân tích sẵn có cho phần mềm đó. Như vậy, sẽ tốn nhiều công sức hơn để có kết quả như Excel.

Bước đầu tiên là mở Bảng tính OpenOffice và mở tập dữ liệu đã được nhập ở bước trước. Kết quả sẽ giống như sau màn hình:

The screenshot shows a LibreOffice Calc spreadsheet titled "195105". The data is organized into columns labeled A through I. Column A contains row numbers from 1 to 16. Columns B through I contain specific data points. The data spans from 195105 to 195106. The last row (row 16) has a formula: =SUM(B2:B15). The bottom status bar indicates "Sum: 19179".

	A	B	C	D	E	F	G	H	I
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE
2	195105	9	915	195105	9	915		10047282	MISSISSIPPI
3	195105	17	2200	195105	17	2200		10028729	KANSAS
4	195105	28	510	195105	28	510		10120421	TEXAS
5	195105	9	1830	195105	9	1830		10099717	OKLAHOMA
6	195105	15	1620	195105	15	1620		10099742	OKLAHOMA
7	195105	8	1800	195105	8	1800		10028891	KANSAS
8	195105	30	1500	195105	30	1500		10104933	PENNSYLVANI
9	195105	11	1330	195105	11	1330		10104934	PENNSYLVANI
10	195105	27	2204	195105	27	2204		10104935	PENNSYLVANI
11	195107	21	1100	195107	21	1100		10104936	PENNSYLVANI
12	195104	29	1815	195104	29	1815		10082587	NEW JERSEY
13	195102	19	1830	195102	19	1830		10099493	OKLAHOMA
14	195105	3	1335	195105	3	1335		10039190	MICHIGAN
15	195106	1	1800	195106	1	1800		10039191	MICHIGAN
16	195106	26	1800	195106	26	1800		10039192	MICHIGAN
17	195105	18	1730	195105	18	1730		10099725	OKLAHOMA
18	195105	19	1915	195105	19	1915		10099726	OKLAHOMA
19	195105	16	1800	195105	16	1800		10099727	OKLAHOMA

Lúc này, thống kê mô tả sẽ bao gồm các mục sau:

1. Ý nghĩa
2. Trung vị
3. Chênh độ
4. Độ lệch chuẩn
5. Bệnh gai
6. Xiên
7. Tối thiểu
8. Tối đa
9. Mức độ tin cậy cho phương tiện

Tất cả các mục trước đây là một phần của thống kê mô tả có trong Excel Analysis ToolPak. Đây là trường hợp của OpenOffice. Tuy nhiên, hãy làm theo các công thức và nó làm cho nó có thể lặp lại.

Công thức đầu tiên sẽ dành cho giá trị trung bình (hoặc trung bình). Công thức sẽ xuất hiện như sau khi được đặt trong thanh công thức của OpenOffice:

=AVERAGE(AH2:AH270)

Công thức này phải được đặt sau ô AH270 để tránh mọi phép tính vòng. Các công thức tiếp theo nên được đặt sau (thấp hơn) phép tính TRUNG BÌNH. Một lời cảnh báo là cần thiết vào thời điểm này. Đảm bảo rằng bạn đặt AH2:AH270 làm tham chiếu tuyệt đối (ký hiệu đô la trước cả AH2 và AH270 để trông như thế này-\$AH\$2 và \$AH\$270). Điều này sẽ ngăn kết quả TRUNG BÌNH được đưa vào phép tính bên dưới nó, v.v. Các công thức sẽ xuất hiện như sau:

=MEDIAN(\$AH\$2:\$AH\$270)
 =MODE(\$AH\$2:\$AH\$270)
 =STDEV(\$AH\$2:\$AH\$270)
 =KURT(\$AH\$2:\$AH\$270)
 =SKEW(\$AH\$2:\$AH\$270)
 =MIN(\$AH\$2:\$AH\$270)
 =MAX(AH2:AH270)

=BẢO MẬT(0.95;AH274;269)

Kết quả của các tính toán này như sau (ở bên trái). Bên cạnh những kết quả đó là kết quả từ Excel (màn hình bên phải). Cái này quan trọng!

Làm thế nào để kết quả so sánh? Chúng khác nhau cơ bản hay tương đối giống nhau? Điều này bây giờ là để xác minh độ chính xác của công cụ của bạn.

Mean	4.4434944238
Median	0.5
Mode	0
Standard Dev	10.2308541821
Kurtosis	25.6745319148
Skew	4.376062845
Minimum	0
Maximum	92.6
Confidence Level for Means	0.039115622

A	B
TOR_LENGTH	
1 Mean	4.443494424
2 Standard Error	0.623786189
3 Median	0.5
4 Mode	0
5 Standard Deviation	10.23085418
6 Sample Variance	104.6703773
7 Kurtosis	25.67453191
8 Skewness	4.376062845
9 Range	92.6
10 Minimum	0
11 Maximum	92.6
12 Sum	1195.3
13 Count	269
14 Largest(5)	44.8
15 Smallest(5)	0
16 Confidence Level(95.0%)	1.228144665
17	
18	
19	
20	

Đáng chú ý nhất là sự khác biệt lớn giữa Mức tin cậy của OpenOffice và Excel. Sau khi xem xét các công thức, "alpha" mong muốn cho OpenOffice là sự khác biệt giữa "1" và Mức độ tin cậy, có nghĩa là, đối với OpenOffice, số thích hợp là "0,05" (hoặc 1 - 0,95), KHÔNG ".95" như đã gửi ban đầu. Sau khi thay đổi công thức này (được minh họa như sau), kết quả Mức tin cậy cũng nằm dưới công thức. Sự khác biệt không còn tồn tại. Đây là điều quan trọng cần nhớ–không phải mọi công thức đều hoàn toàn giống nhau giữa Excel và OpenOffice.

=CONFIDENCE(0,05;AH274;269)

18	Confidence Level(95.0%)	1.228144665	
19			Confidence Level for Means
20			1.222598464

TOR_LENGTH StormEw

Một điểm khác biệt nữa giữa Excel và OpenOffice là bất kỳ sự phân tách nào trong công thức phải được thực hiện bằng dấu "," trong Excel và dấu ";" trong OpenOffice. Nhà phân tích sẽ nhận được thông báo lỗi khi sử dụng sai biểu tượng trong OpenOffice. Cảnh báo này nhằm giảm bớt sự lo lắng cho các nhà phân tích thường cảm thấy có vấn đề với phần mềm nếu lỗi xuất hiện. Trong trường hợp này, thật đơn giản để thay đổi từ dấu phẩy thành dấu chấm phẩy.

Với Excel và OpenOffice xây dựng thống kê mô tả, RStudio/Rattle và KNIME sẽ khó khăn hơn một chút, nhưng chắc chắn không thể vượt qua được. RStudio là đầu tiên và sau đó là KNIME.

3.1.3 RStudio/Rattle

Bước đầu tiên rất đơn giản–hãy mở ứng dụng RStudio và gói Rattle như đã đề cập trong phần Nhập dữ liệu để đến màn hình sau.

No.	Variable	Data-Type	Input	Target	Risk	Ident.	Ignore	Weight	Comment
20.	INJURIES_DIRECT	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:17
21.	INJURIES_INDIRECT	Constant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:1
22.	DEATHS_DIRECT	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:6
23.	DEATHS_INDIRECT	Constant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:1
24.	DAMAGE_PROPERTY	Categorical	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:8
25.	DAMAGE_CROPS	Constant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:1
26.	MAGNITUDE	Constant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:1
27.	MAGNITUDE_TYPE	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:0 Missing:26
28.	FLOOD_CAUSE	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:0 Missing:26
29.	CATEGORY	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:0 Missing:26
30.	TOR_E_SCALE	Categorical	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:3 Missing:29
31.	TOR_LENGTH	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique:7
32.	TOR_WIDTH	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:39
33.	TOR_OTHER_WFO	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique:0 Missing:26

Rules noted: 269 observations and 15 input variables. No targets found for predictive modelling nor sampling.

Lưu ý “Các vai trò được lưu ý” ở cuối màn hình này. Nhà phân tích có thể chọn bất kỳ mục tiêu nào để xác định số liệu thống kê. Trong trường hợp này, “TOR_LENGTH” được chọn để nhất quán với cùng một biến được sử dụng trong các phần/công cụ trước đó.

Vui lòng lưu ý phần màn hình được đánh dấu là “Phân vùng” được chọn.

Điều này có nghĩa là Rattle sẽ tự động phân tách dữ liệu thành phần trăm để đào tạo và xác nhận; trong trường hợp này, 70% dữ liệu sẽ được sử dụng để đào tạo. Tập dữ liệu được lấy mẫu này không phải là toàn bộ tập dữ liệu nhưng được chọn ngẫu nhiên tự động để nhà phân tích có thể kiểm tra các chức năng khác nhau trên một mẫu của tập dữ liệu thay vì sử dụng toàn bộ tập dữ liệu. Đối với tập dữ liệu nhỏ, điều này là không cần thiết, nhưng đối với tập dữ liệu lớn hơn, điều này không chỉ có lợi mà còn cần thiết để tiết kiệm thời gian tính toán. Trong trường hợp này, hộp được chọn, nhưng bỏ chọn hộp này

hộp (hiển thị như sau) để thống kê mô tả chỉ được thực hiện đối với toàn bộ tập dữ liệu, nhất quán với các phần khác.

The screenshot shows the Rattle software interface version 5.2.0. The main window displays a table of variables with the following columns: No., Variable, Data Type, Input, Target, Risk, Ident, Ignore, Weight, and Comment. The 'Input' column has radio buttons for 'missing' (selected), 'constant', and 'formula'. The 'Target' column has radio buttons for 'none' (selected), 'categorical', and 'numerical'. The 'Weight' column has radio buttons for 'none' (selected), 'auto', and 'category'. The 'Comment' column provides statistics for each variable, such as 'Unique: 0 Missing: 26' for 'MAGNITUDE_TYPE'.

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
27	MAGNITUDE_TYPE	missing	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
28	FLOOD_CAUSE	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
29	CATEGORY	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
30	TORF_SCALE	Categorical	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 5 Missing: 35
31	TOR_LENGTH	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 71
32	TOR_WIDTH	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 39
33	TOR_OTHER_WFO	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
34	TOR_OTHER_CZ_STATE	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
35	TOR_OTHER_CZ_RPS	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
36	TOR_OTHER_CZ_NAME	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
37	BEGIN_RANGE	Constant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1
38	BEGIN_AZIMUTH	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
39	BEGIN_LOCATION	missing	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 26
40	END_RANGE	Constant	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1

Sau khi đạt được màn hình này, bước tiếp theo sẽ là đảm bảo rằng "Exe cute" được nhấp để khóa tập dữ liệu và các biến. Sau khi hoàn tất, hãy chọn nút radio bên cạnh tab "Dữ liệu" có tên là tab "Khám phá". Màn hình sau sẽ xuất hiện:

Rattle Version 5.3.0 (log: rattle.log)

Data **Explore** **Test** **Transform** **Cluster** **Associate** **Model** **Evaluations** **Log**

Type: Summary Distribution Correlation Principal Components Interactive

Summary Descriptive Basics Kurtosis Skewness Show Missing Cross Tab

Rattle timestamp: 2019-12-31 08:18:58 Chris

Basic statistics for each numeric variable of the dataset.

VTR LENGTH	(X.., X..)
None	269.000000
NAs	0.000000
Minimum	0.000000
Maximum	92.500000
1. Quartile	0.000000
3. Quartile	4.100000
Mean	4.443494
Median	0.500000
Sum	1105.300000
SE Mean	0.623788
LCL Mean	3.215350
UCL Mean	5.671639
Variance	184.870377
StDev	13.230854
Skewness	-4.327398
Kurtosis	24.968456

Rattle timestamp: 2019-12-31 08:18:59 Chris

Kurtosis for each numeric variable of the dataset.
Larger values mean sharper peaks and flatter tails.

Find: End

Data summary generated.

Lưu ý từ màn hình trước các nút radio được chọn và các hộp được chọn. Những gì sự kết hợp này làm là cung cấp cho nhà phân tích nhiều thông tin nhất từ việc nhấp vào biểu tượng "Thực thi". Trong trường hợp này, nó cung cấp nhiều kết quả khác nhau tương tự như kết quả do cả Excel và OpenOffice tạo ra. Xem từng màn hình sẽ hiển thị những kết quả này theo cách có tổ chức hơn.

Một vài phần đầu tiên của màn hình thống kê mô tả trong Rattle hiển thị một cột có tất cả các thống kê tóm tắt khác nhau, gần giống với cột

Màn hình Excel và OpenOffice. Màn hình Excel được đặt bên cạnh màn hình kết quả tổng hợp của Rattle để hiển thị những điểm tương đồng. Một lần nữa, có vẻ như các con số khớp với kết quả ban đầu từ Excel, giúp xác minh thuật toán trong (cho đến nay) tất cả các công cụ khác nhau. Đây là một điều tốt! Tính nhất quán là chìa khóa cho số liệu thống kê, do đó, việc có các kết quả giống nhau cho thấy tính nhất quán. Một kết quả hơi khác một chút là độ nhọn, nhưng điều này có thể là do làm tròn trong phép tính và không liên quan đến nhà phân tích dữ liệu. Định nghĩa của từng tên hàng trong kết quả Rattle được bao gồm trong bảng sau. Các khía cạnh rất thú vị và được quan tâm đặc biệt là "LCL" và "UCL", được bao gồm trong kết quả Rattle. Điều này chỉ định "Mức độ tin cậy thấp hơn" và "Mức độ tin cậy cao hơn", giống như "Mức độ tin cậy 95%" trong Excel. Vì giá trị này có vẻ khác, hãy lấy giá trị trung bình và cộng "Mức tin cậy 95%" từ Excel, sau đó lấy giá trị trung bình và trừ "Mức tin cậy 95%" từ Excel, và bạn sẽ nhận được UCL và LCL tương ứng. Rattle trình bày cùng một kết quả theo một cách khác. Xin đừng để điều đó ném bạn như một nhà phân tích. Các công cụ khác nhau có thể đưa ra kết quả khác nhau, nhưng điều đó không có nghĩa là chúng không nhất quán về kết quả, chỉ khác nhau về định dạng.

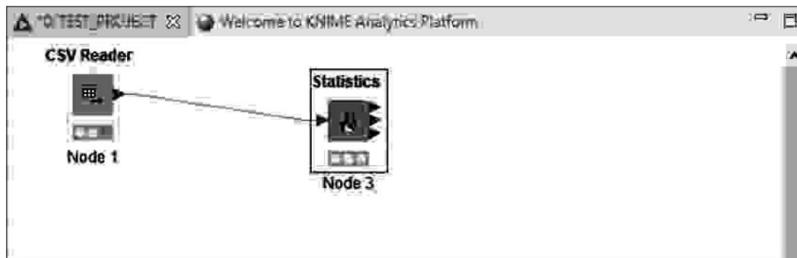
Basic statistics for each numeric variable of the dataset.		
STOR_LENGTH	X...X.i	
nobs	269.000000	
NAs	0.000000	
Minimum	0.000000	
Maximum	92.600000	
1. Quartile	0.000000	
3. Quartile	4.100000	
Mean	4.443494	
Median	0.500000	
Sum	1195.300000	
SE Mean	0.623786	
LCL Mean	3.215350	
UCL Mean	5.671639	
Variance	104.670377	
Stdev	10.230854	
Skewness	4.327380	
Kurtosis	24.968456	
A		
TOR_LENGTH		
1		
2		
3	Mean	4.443494424
4	Standard Error	0.623786189
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.23085418
8	Sample Variance	104.6703773
9	Kurtosis	25.67453191
10	Skewness	4.376062845
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.228144665
19		
20		
	TOR_LENGTH	StormEven

Tên hàng	Sự định nghĩa
quý tộc	Số đối tượng
NA	Dữ liệu bị mất
tối thiểu	Giá trị tối thiểu
tối đa	Giá trị lớn nhất
1. Phần tư	Phần tư thứ nhất (Phần trăm thứ 25)
3. Phần tư	Phần tư thứ ba (Phần trăm thứ 75)
Nghĩa là	Trung tâm số học của dữ liệu
Trung bình	Trung tâm dữ liệu vật lý
Tổng	Giá trị cộng của tất cả dữ liệu
SE có nghĩa là	Lỗi chuẩn (độ lệch chuẩn/căn bậc hai của đối tượng)
LCL có nghĩa là	Mức độ tự tin thấp hơn của ý nghĩa
Ý nghĩa của UCL	Mức độ tin cậy cao hơn của ý nghĩa
phương sai	Tổng bình phương chênh lệch giữa mỗi giá trị và giá trị trung bình
stdev	Độ lệch chuẩn (căn bậc hai của phương sai)
độ lệch	Tích cực có nghĩa là lệch phái, tiêu cực có nghĩa là lệch trái, 0 có nghĩa là phân phối bình thường (hoặc đóng)
gai nhọn	"Định" của dữ liệu (giá trị cao hơn có nghĩa là định sắc nét hơn)

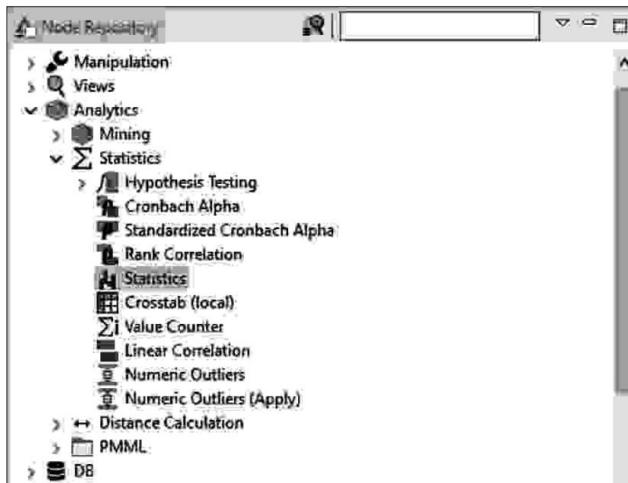
Cho đến nay, việc so sánh tất cả các kết quả dường như chỉ ra một số khác biệt rất nhỏ có thể do loại công thức hoặc do làm tròn trong công thức đó gây ra. May mắn thay, các công cụ nhất quán trong các số liệu chính của chúng, điều này cho thấy rằng việc sử dụng một số công cụ để xác minh kết quả là điều cần khám phá thêm, điều này sẽ xảy ra trong cuốn sách này.

3.1.4 KIẾN THỨC

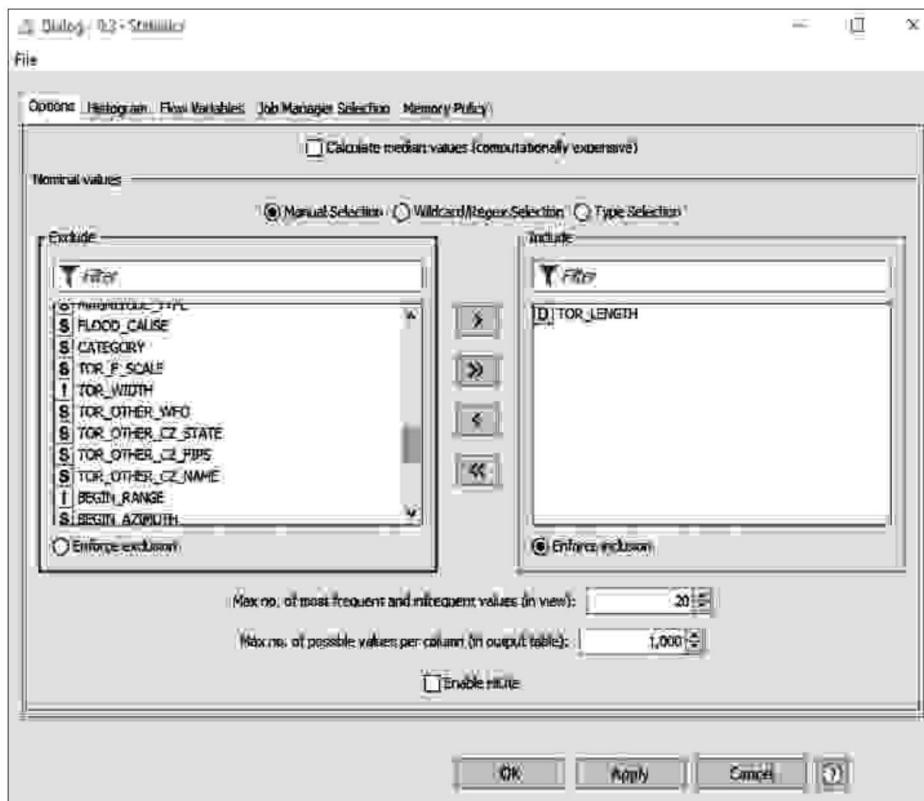
KNIME có bản chất là mô-đun, vì vậy sử dụng nút sẽ cung cấp cho bạn số liệu thống kê mô tả là cách tốt nhất để thực hiện chức năng này dễ dàng nhất có thể. Có một số chuẩn bị dữ liệu phải được thực hiện trước bước này. Bước đầu tiên là mở KNIME cho một dự án mới hoặc dự án mà bạn đã lưu như được hiển thị.



Như mọi người có thể thấy, một nhà phân tích có thể thêm bất kỳ nút nào có sẵn trong KNIME vào không gian làm việc. Nhưng cái nào để thêm? Thông tin về nút đó ở đâu? Câu trả lời có ngay trong không gian làm việc của ứng dụng KNIME. Chọn nút "Thống kê" nằm trong phần "Thống kê" của danh mục nút "Phân tích". Vị trí được hiển thị như sau. Sau khi định vị nút, nhấp chuột trái và giữ, kéo nút vào không gian làm việc. Sau khi bạn định vị lại nút, hãy kết nối các nút bằng cách nhấp và giữ vào "hình tam giác màu đen" nằm ở phía bên phải của nút "Trình đọc CSV" và kết nối nó với nút "Thống kê". Không gian làm việc hiện đã sẵn sàng để thực hiện chức năng của nút.



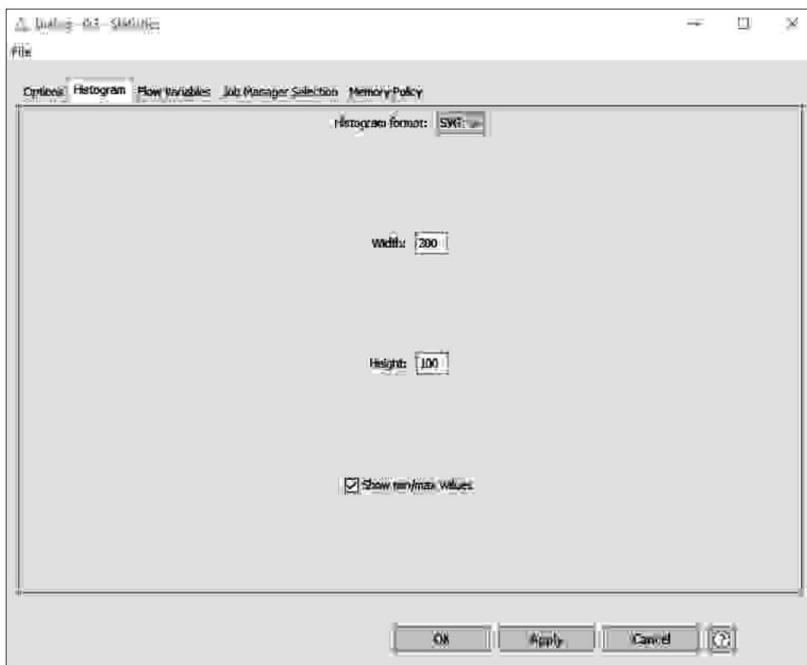
Sau khi đặt và kết nối các nút, nhấp đúp vào nút “Thông kê” và màn hình này sẽ xuất hiện. Những gì màn hình này thực hiện là thực hiện thống kê mô tả về biến hoặc các biến mà nhà phân tích chọn. Trong trường hợp này, chỉ “TOR_LENGTH” sẽ được chọn để nhất quán với các chức năng khác của công cụ. Cách để xóa phía bên phải của màn hình (nơi tồn tại các biến mà nhà phân tích muốn đánh giá) là nhấp vào mũi tên trái kép (<<) rồi ở phía bên trái của màn hình, chỉ chọn TOR_LENGTH, di chuyển nó sang bên phải cạnh màn hình bằng một lần bấm vào mũi tên phải (>). Bây giờ nút được định cấu hình với biến, nhưng cần thiết lập sự chuẩn bị khác trước khi nhấp vào “Thực thi”.



Lưu ý hai hộp văn bản bên dưới màn hình lựa chọn. Những gì chúng biểu thị là số lượng tối đa các giá trị khác nhau và số lượng giá trị tối đa trong các biến mà nhà phân tích đã chọn. Nếu có hơn 20 giá trị khác nhau, KNIME sẽ bỏ qua chúng. Nó có thể có lợi để thay đổi điều này thành một

số cao hơn để giải thích cho số lượng thay đổi giá trị cao hơn trong trường hợp có các giá trị duy nhất trong biến hoặc cột. Trong trường hợp này, thay đổi thành 100 sẽ đủ cho TOR_LENGTH. Khối thứ hai là đủ vì số hàng trong tập dữ liệu này là khoảng 300, do đó, số được đặt trong khối này phải nhiều hơn cột này.

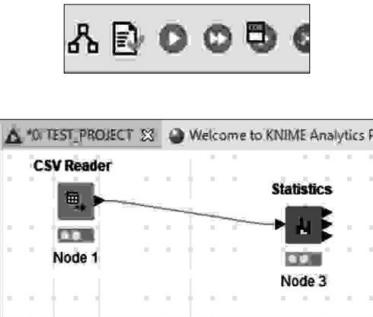
Sau khi hoàn thành, hãy chuyển sang tab tiếp theo, "Histogram", trông giống như màn hình sau. Không có nhiều thay đổi ở đây trừ khi nhà phân tích cần nhiều không gian pixel hơn hoặc hình ảnh lớn hơn. Màn hình sẽ mặc định là SVG, nhưng có thể đổi thành PNG nếu muốn. Khám phá cả hai để xem cái nào sẽ phù hợp với bản trình bày hoặc bài viết của bạn.



Khi tắt cả các màn hình đáp ứng nhu cầu của nhà phân tích, hãy nhấp vào OK hoặc Áp dụng và một thông báo có thể xuất hiện cho biết rằng nút đã được thay đổi và hỏi xem nhà phân tích có muốn thay đổi này không. Bấm OK lần nữa nếu thông báo này xuất hiện và bạn muốn thay đổi nút.

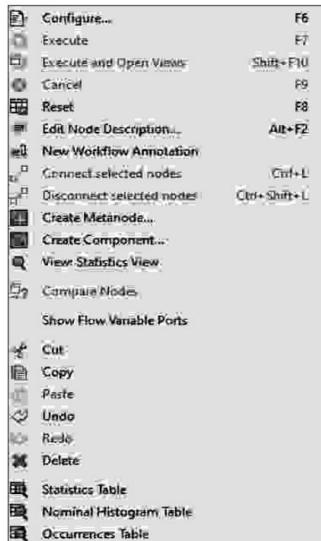
Một lưu ý thận trọng là cần thiết ở đây. Trên màn hình đầu tiên, nhà phân tích sẽ nhận thấy rằng khôi có nhãn "Tính giá trị trung bình" không được chọn. Điều này là do thiết kế, vì việc tính toán các giá trị trung bình rất khó khăn về mặt tính toán, đặc biệt là với nhiều biến số. Thật không may, với mục đích của cuộc biểu tình này, các giá trị trung bình là một phần của thống kê mô tả cần thiết, vì vậy khôi này sẽ phải được kiểm tra.

Sau khi cấu hình hoàn tất, nhà phân tích sẽ thấy các nút “Trình đọc CSV” và “Thống kê” trong cấu hình luồng được kết nối với nhau ra của Trình đọc CSV thành Thống kê. Bằng cách nhấp vào “mũi tên kép màu xanh lá cây” (được minh họa như sau), nhà phân tích sẽ thực hiện tất cả các nút và đèn xanh lục sẽ xuất hiện trên tất cả các nút như trong màn hình minh họa mũi tên kép màu xanh lá cây.



Nếu nhà phân tích nhấp chuột phải vào nút “Thống kê”, họ sẽ thấy lựa chọn xem bảng tóm tắt được hiển thị như sau. Điều này sẽ trình bày kết quả cho người phân tích chức năng vừa được thực hiện. Sau đây là cả màn hình này và màn hình cho kết quả.

Khi bảng xuất hiện, nhà phân tích sẽ nhận thấy rằng bảng chứa tất cả các biến khác nhau. Nếu nhà phân tích cuộn để xem TOR_LENGTH, nó sẽ hiển thị kết quả giống như các công cụ khác.



File Home Navigation View										
Table "default" - Rows: 25 Spec - Columns: 16 Properties - Row Variables										
Row ID	Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	D. Outlier	D. Outlier
TOR_LENGTH	TOR_LENGTH	0	92.6	4.443	10.231	104.67	4.376	25.675	1,195.3	

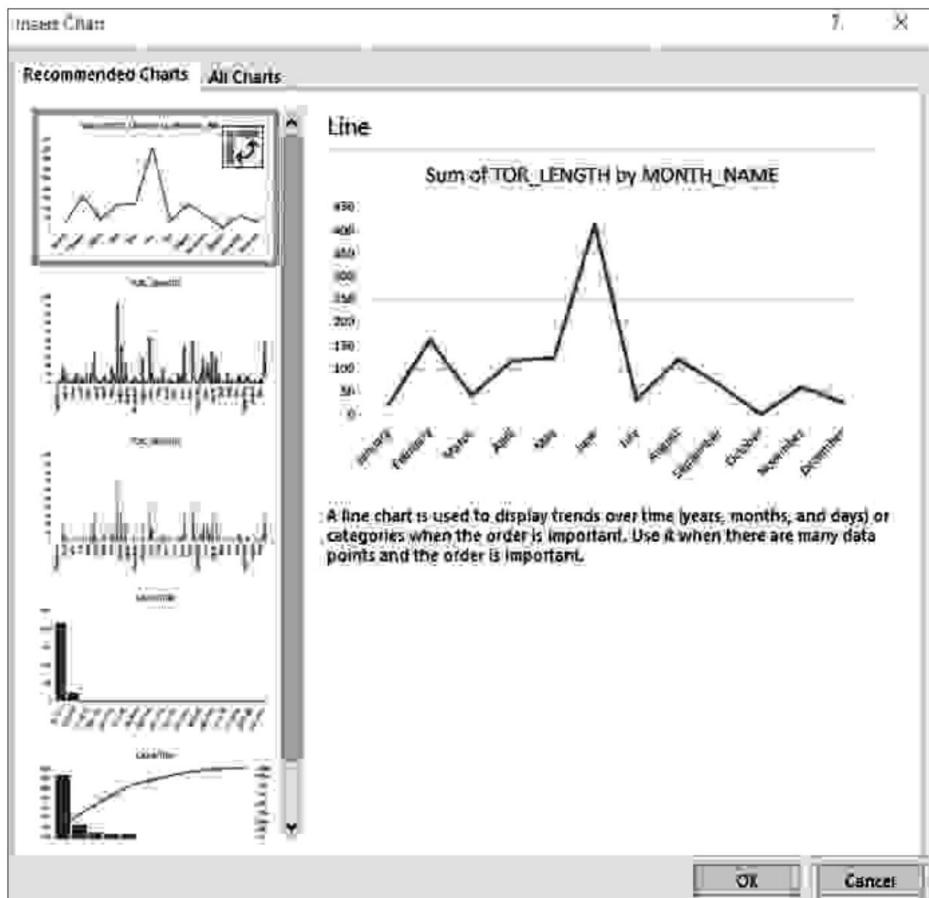
3.2 BIỂU ĐỒ XÁC SUẤT TÍCH LŨY

Mặc dù được đề cập một cách rất thô sơ trong hầu hết các lớp thống kê, một biểu đồ xác suất tích lũy là một cách trình bày dữ liệu rất hữu ích để chỉ ra nơi phát sinh các vấn đề chính. Ví dụ: nếu người quản lý muốn xem bộ phận nào đang có thời gian nghỉ phép được trả lương nhiều nhất (PTO) một năm, theo tháng, người quản lý có thể sử dụng biểu đồ này để xác định bộ phận nào (và tháng nào) thường như có thiên hướng nhất đối với PTO cho nhân viên. Tôi luôn ngạc nhiên rằng, khi thực hiện chức năng thống kê này, các tháng chính của PTO không phải là tháng 12 hay tháng 1, mà nhiều hơn vào mùa xuân và mùa thu. Điều này tương quan với việc tốt nghiệp (đại học và trung học), cùng với các trận bóng đá (cụ thể là các trận sân khách). Nói chung, điều này hữu ích cho nhiều ngành công nghiệp, từ ngân hàng đến chế tạo kim loại. Cuốn sách này sẽ đề cập đến từng công cụ với cùng một bộ dữ liệu đã được sử dụng từ trước đến nay và sẽ tập trung vào hai biến số, TOR_LENGTH và MONTH_NAME, để xem liệu lốc xoáy có xảy ra nhiều nhất trong những tháng nhất định hay không và sử dụng biểu đồ xác suất để hiển thị dữ liệu này.

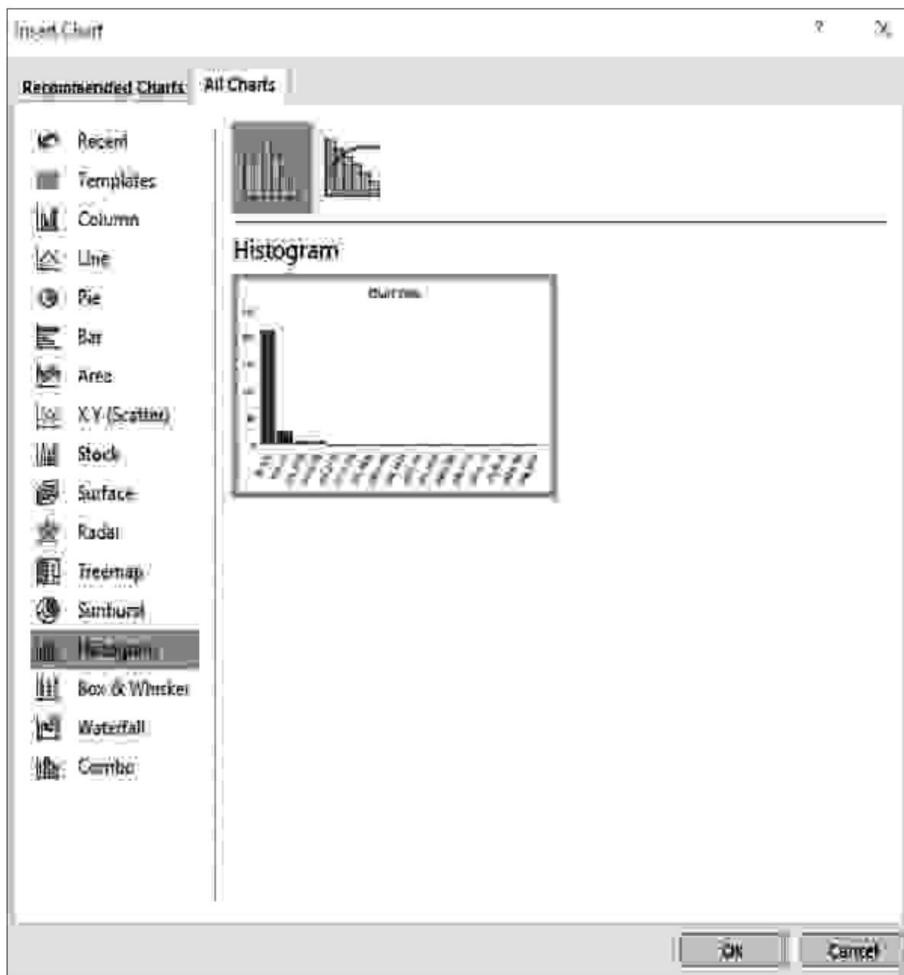
3.2.1 Excel

Excel có sẵn một chức năng được chuẩn bị cho biểu đồ xác suất tích lũy, còn được gọi là biểu đồ Pareto. Nhà phân tích không phải tiến hành thông qua bảng tổng hợp hoặc biểu đồ và có thể tự động tạo biểu đồ từ lựa chọn tập dữ liệu.

Như đã nêu trước đó, nhà phân tích nên chọn hai cột sẽ được lập biểu đồ, đó là TOR_LENGTH và MONTH_NAME. Để thực hiện việc này, hãy chọn cột đầu tiên và giữ phím "CTRL" trong khi chọn cột thứ hai. Sau khi hoàn tất, hãy chọn "Chèn" từ thanh công cụ chính và chọn "Biểu đồ được đề xuất". Điều này sẽ tạo ra màn hình sau.

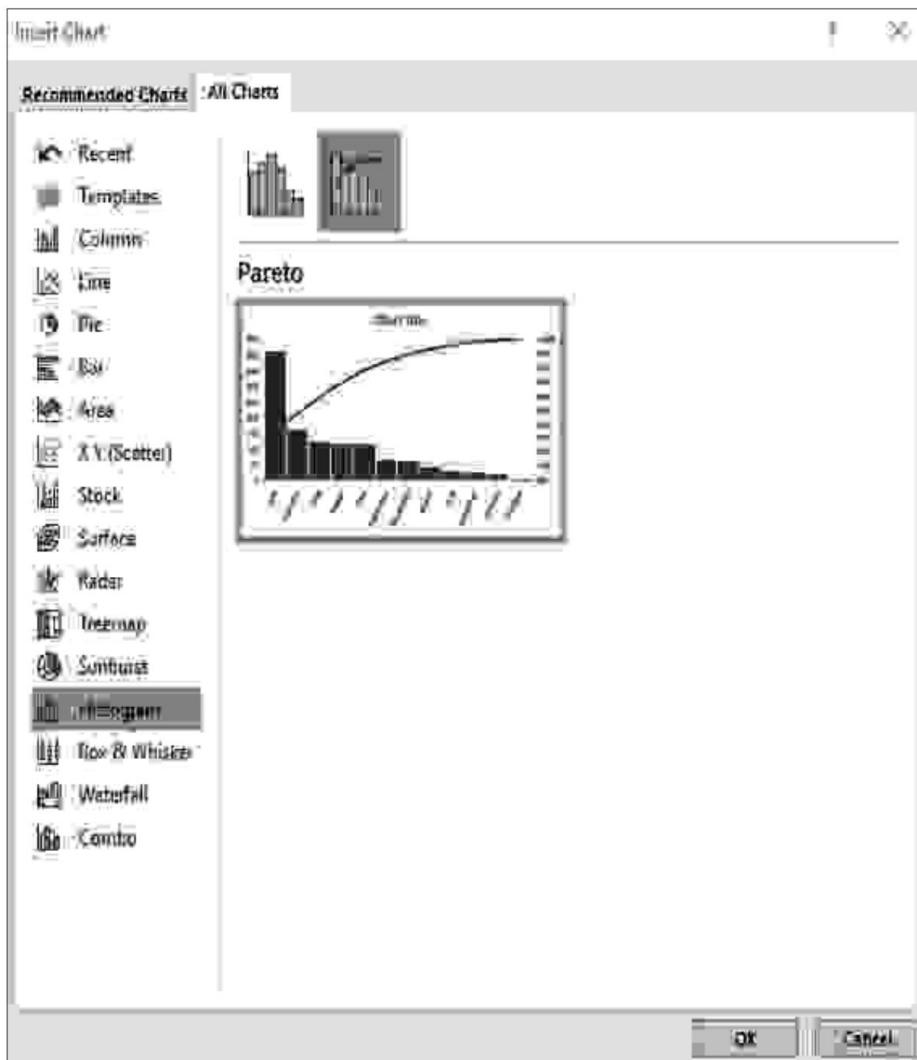


Tại thời điểm này, hãy chọn tab “Tất cả biểu đồ” (ở bên phải của tab đã chọn) và bạn sẽ thấy màn hình sau.



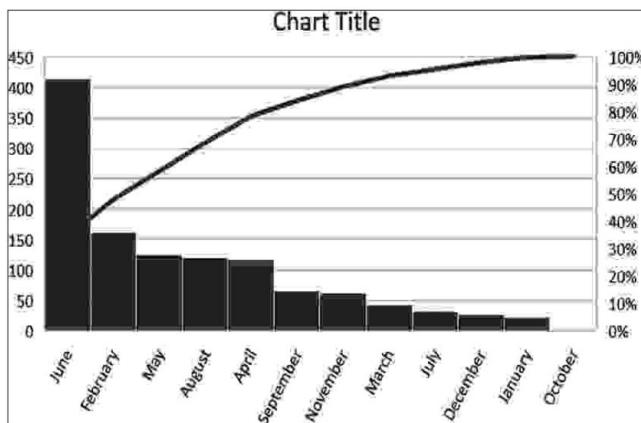
Sau khi bạn chọn "Histogram" từ menu bên trái, hai lựa chọn phụ sẽ xuất hiện ở màn hình bên phải ở cột phía trên bên trái. Biểu đồ bên trái là biểu đồ thông thường và biểu đồ bên phải là biểu đồ xác suất tích lũy.

Chọn cái bên phải và biểu đồ đã hoàn thành sẽ xuất hiện trong bản xem trước như được hiển thị.



Nhập vào OK và biểu đồ sẽ xuất hiện, nhưng nó có nghĩa là gì? Giải thích là có những cơn lốc xoáy kéo dài hơn (về cơ bản là những cơn lốc xoáy dài hơn) trong một số tháng so với những tháng khác. Để thực hiện phép đo Pareto thông thường, hãy tiếp tục nhìn sang bên phải cho đến khi đạt được mốc 80% và điều đó sẽ cho thấy những tháng tạo ra 80% số cơn lốc xoáy dài hơn. Các tháng sẽ là tháng sáu, tháng hai, tháng năm, tháng tám và tháng tư. Hãy nhớ rằng điều này chỉ tính đến năm 1951.

Giờ đây, nhà phân tích có thể lập một biểu đồ tương tự với các tiểu bang và xem các tiểu bang ở Hoa Kỳ nơi xảy ra 80% các cơn lốc xoáy dài hơn. Hãy thử và bạn sẽ bị sốc bởi những bang có lốc xoáy dài hơn. Không phải những người bạn mong đợi! Sự kỳ diệu của phân tích tập dữ liệu.



3.2.2 Văn phòng mở

OpenOffice không có nút “ma thuật” như Excel sở hữu, nhưng nó có khả năng tạo biểu đồ xác suất tích lũy trong khả năng bảng tổng hợp của nó.

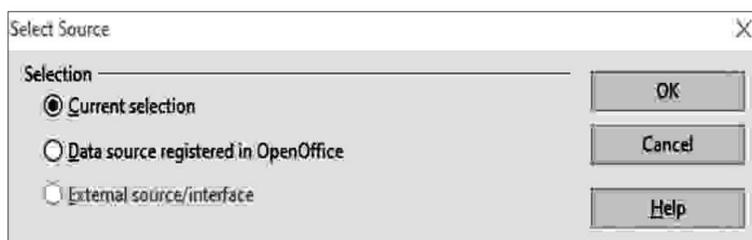
Bước đầu tiên là mở OpenOffice và đảm bảo rằng cùng một bộ dữ liệu đã được tải vào Excel cũng được tải trong OpenOffice. Màn hình tại thời điểm này sẽ giống như sau:

BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STA
195109	9	916	195109	9	916	10047282	MISI	
195106	17	2200	195106	17	2200	10028729	KAN	
195103	28	510	195103	28	510	10120421	TEX	
195105	9	1830	195105	9	1830	10088717	OKL	
195107	15	1620	195107	15	1620	10099742	OKL	
195105	8	1800	195105	8	1800	10028681	KAN	
195103	30	1500	195103	30	1500	10104933	PEN	
195105	11	1330	195105	11	1330	10104934	PEN	
195106	27	2204	195106	27	2204	10104935	PEN	
195107	21	1100	195107	21	1100	10104936	PEN	
195104	29	1815	195104	29	1815	10062587	NEV	
195102	19	1830	195102	19	1830	10089493	OKL	
195105	3	1335	195105	3	1335	10039190	MICI	
195106	1	1800	195106	1	1800	10039181	MICI	
195106	26	1800	195106	26	1800	10039182	MICI	
195105	18	1730	195105	18	1730	10099725	OKL	

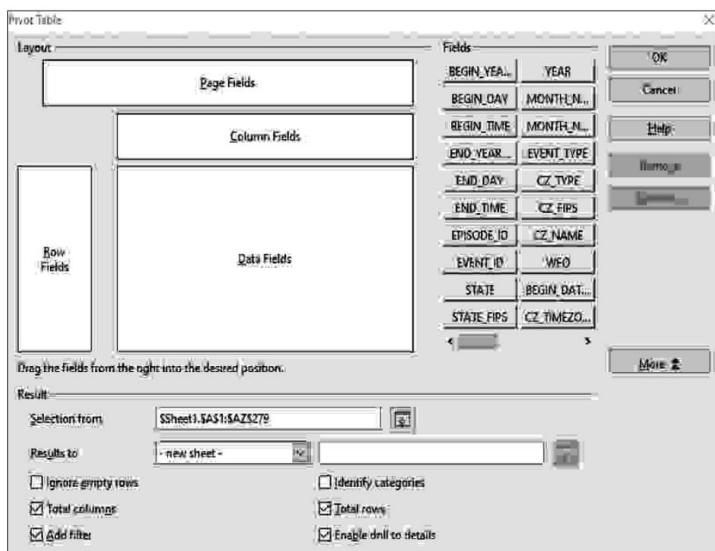
Bước tiếp theo cần làm là chèn biểu đồ trực (giống như với Excel) để sử dụng cách tiếp cận “theo nhóm” đối với dữ liệu và bật chức năng khả năng thăm dò tích lũy. Điều này nằm trong khu vực “Dữ liệu” của thanh công cụ chính và được hiển thị như sau. Chọn “Pivot Table” rồi chọn “Create”, lúc đó màn hình tiếp theo sẽ xuất hiện.

The screenshot shows the OpenOffice Calc interface with a CSV file named "StormEvents_details-ftp_v1_0_d1951_c20160223.csv" loaded. The Data menu is open, and the "Create..." option under the Pivot Table section is highlighted. The main window displays a table with columns labeled BEGIN_YEARMONTH, N_TIME, END_YE, and various numerical values.

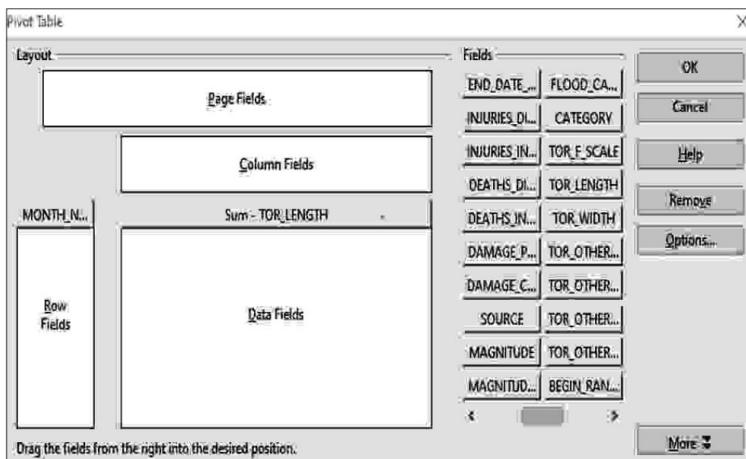
BEGIN_YEARMONTH	N_TIME	END_YE
195109	915	
195108	2200	
195103	610	
195105		
195107		
195106		
195103	1500	
195105	1330	
195106	2204	
195107	1100	
195104	1815	
195102	1830	
195105	1335	
195106	1800	
195106	1800	
195105	1730	



Nút radio “Lựa chọn hiện tại” sẽ là mặc định trong tình huống này; nhấn OK để hiển thị bảng như sau.

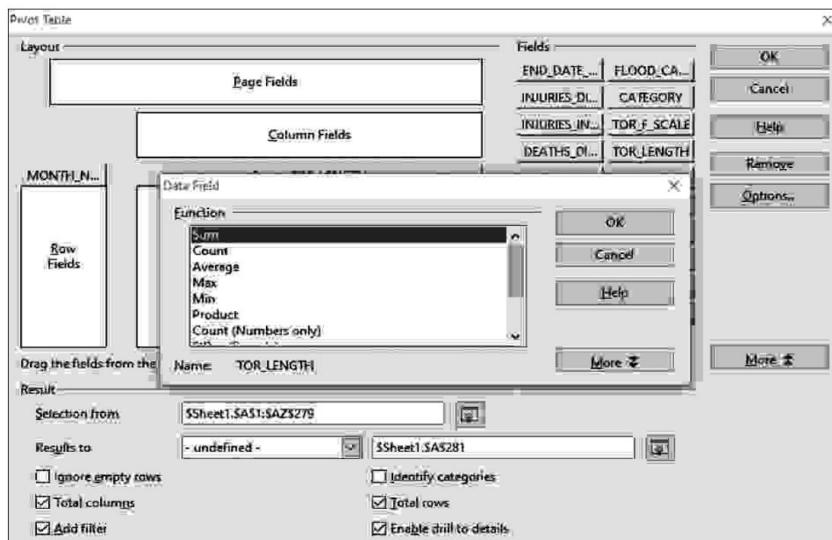


Có rất nhiều hoạt động trong màn hình này, nhưng bước đầu tiên là đảm bảo rằng thông tin bên dưới nút “Thêm” là chính xác. Đảm bảo rằng “Lựa chọn từ” đại diện cho dữ liệu bạn muốn trong biểu đồ trực và “Kết quả đến” là một trang tính mới. Bằng cách này, dữ liệu sẽ không được “xếp gọn” vào cùng một bảng tính với dữ liệu khác. Khám phá các “hộp kiểm” sao cho chúng khớp với cấu hình của nhà phân tích để nhận được nhiều phân tích nhất cho chức năng. Sau đó, cấu hình Pivot Table như sau cho ví dụ này. Lý do là yêu cầu là phải hiểu chiều dài cơn lốc xoáy có liên quan như thế nào với tháng trong năm. Nhà phân tích sẽ đặt “MONTH_NAME” vào Trường hàng và “TOR_LENGTH” vào Trường dữ liệu như sau.



Lưu ý rằng TOR_LENGTH có "Tổng" ở bên trái tiêu đề cột.

Tổng số được đánh giá cao, nhưng chiều dài cơn lốc xoáy trung bình là nơi đặt yêu cầu thực sự. Để thay đổi "Tổng" thành "Trung bình", nhấp chuột trái vào thanh màu xám được đánh dấu "Sum-TOR_LENGTH" và nhìn sang bên phải để thấy có lựa chọn "Tùy chọn." có màu xám đậm. Nhấp vào lựa chọn thay thế đó và màn hình sau sẽ xuất hiện. Chọn "Trung bình" và nhấp vào OK để thay đổi nhãn trên TOR_LENGTH. Tại thời điểm này, hãy nhấp vào OK và dữ liệu sau đó sẽ xuất hiện dưới dạng tập dữ liệu thông thường chỉ với tháng và độ dài cơn lốc xoáy.

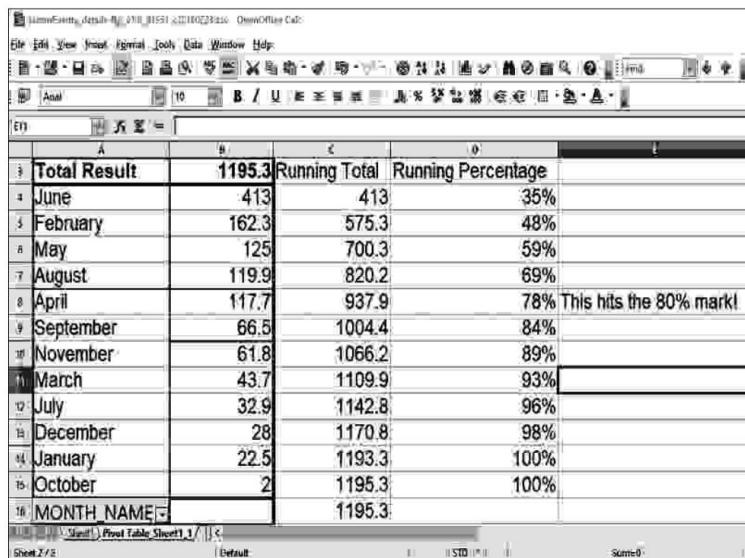


Giờ đây, nhà phân tích phải quay trở lại kiến thức về Excel trước khi Excel có khả năng đặt cùng một biến trong các hàng cho các mục đích khác nhau. OpenOffice không cho phép điều này, nhưng một phần công việc đã được thực hiện. Nhà phân tích phải sắp xếp dữ liệu số theo thứ tự giảm dần để hiển thị chuỗi tháng giống như trong Excel. Màn hình đó như sau:

	MONTH NAME	Value
4	June	413
5	February	162.3
6	May	125
7	August	119.9
8	April	117.7
9	September	66.5
10	November	61.8
11	March	43.7
12	July	32.9
13	December	28
14	January	22.5
15	October	2
16	Total Result	1195.3

Những gì đã đạt được cho đến nay chỉ là một phần của yêu cầu đối với biểu đồ xác suất tích lũy. Khi điều này được thực hiện, hãy thêm hai cột nữa và thực hiện tổng số đang chạy; các công thức được hiển thị trong màn hình tiếp theo cho tổng số hoạt động và tỷ lệ phần trăm hoạt động. Khi điều này hoàn tất, bước tiếp theo sẽ là chèn biểu đồ để hiển thị đúng kết quả.

	A	B	C	D
3	Total Result	1195.3		
4	June	413	=B4	=C4/\$C\$16
5	February	162.3	=C4+B5	=C5/\$C\$16
6	May	125	=C5+B6	=C6/\$C\$16
7	August	119.9	=B7+C6	=C7/\$C\$16
8	April	117.7	=C7+B8	=C8/\$C\$16
9	September	66.5	=C8+B9	=C9/\$C\$16
10	November	61.8	=C9+B10	=C10/\$C\$16
11	March	43.7	=C10+B11	=C11/\$C\$16
12	July	32.9	=C11+B12	=C12/\$C\$16
13	December	28	=C12+B13	=C13/\$C\$16
14	January	22.5	=C13+B14	=C14/\$C\$16
15	October	2	=C14+B15	=C15/\$C\$16
16	MONTH NAME		=C15	



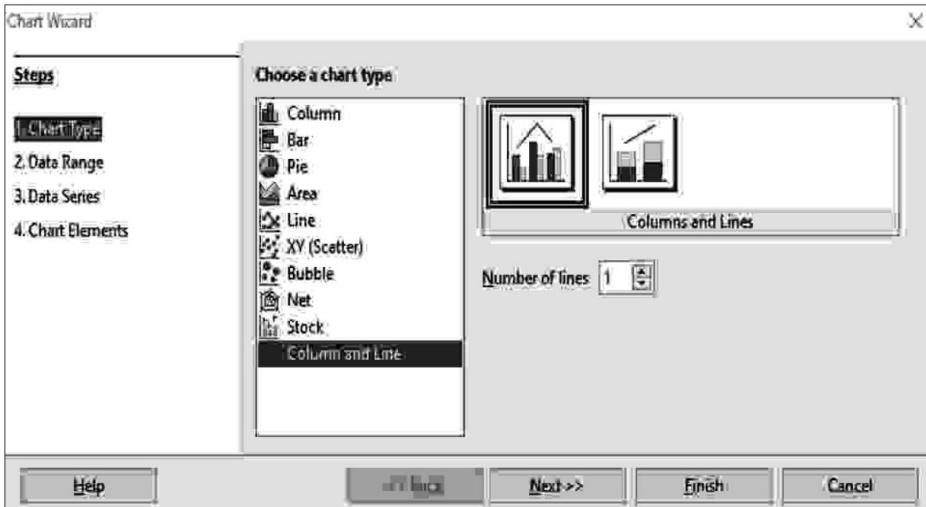
The screenshot shows a Pivot Table in OpenOffice Calc. The table has three columns: 'Total Result' (containing values like 1195.3, 413, etc.), 'Running Total' (containing values like 413, 575.3, etc.), and 'Running Percentage' (containing values like 35%, 48%, etc.). Row 16 contains the formula '=MONTH_NAME'. A note 'This hits the 80% mark!' is visible in the cell next to the percentage for March. The status bar at the bottom indicates 'Sheet1!_Pivot Table_Sheet1.1 / 1 / 1'.

	A	B	C	D	E
	Total Result	1195.3	Running Total	Running Percentage	
1	June	413	413	35%	
2	February	162.3	575.3	48%	
3	May	125	700.3	59%	
4	August	119.9	820.2	69%	
5	April	117.7	937.9	78%	This hits the 80% mark!
6	September	66.5	1004.4	84%	
7	November	61.8	1066.2	89%	
8	March	43.7	1109.9	93%	
9	July	32.9	1142.8	96%	
10	December	28	1170.8	98%	
11	January	22.5	1193.3	100%	
12	October	2	1195.3	100%	
13	MONTH_NAME		1195.3		

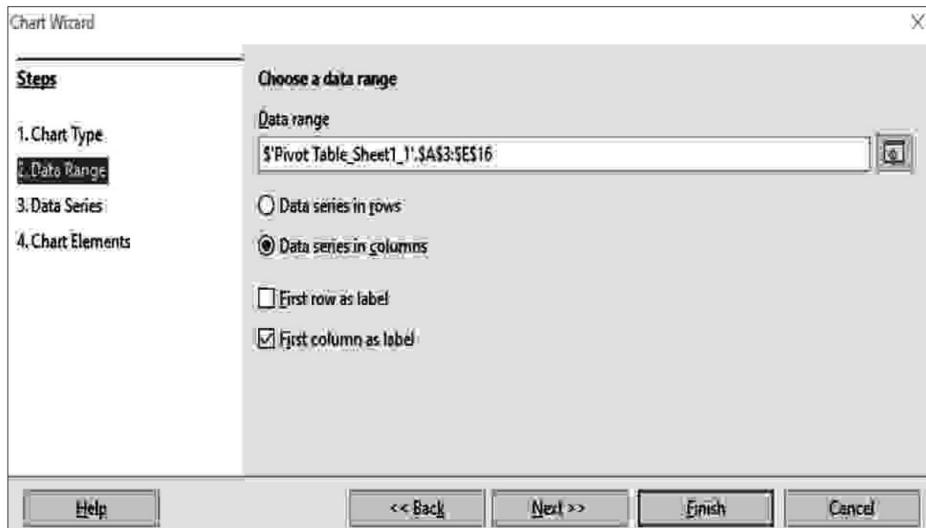
Có một số bước cấu hình phải được hoàn thành trước khi nhà phân tích nhận được kết quả tương tự như với Excel. Bây giờ, hãy xem xét kỹ các bước cần thiết.

Bước đầu tiên là sử dụng thanh công cụ “Chèn” để chèn biểu đồ; trong trường hợp này, biểu đồ thanh bình thường vẫn ổn, nhưng có một biểu đồ thanh kết hợp được mô tả như sau hoạt động rất tốt trong tình huống này. Khi điều này được chọn, màn hình tiếp theo sẽ xuất hiện để hiển thị các lựa chọn biểu đồ khác nhau.

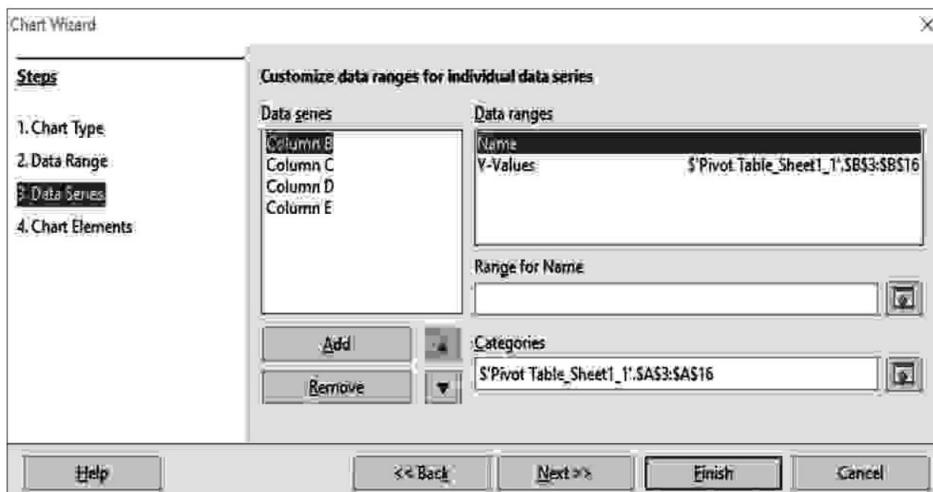




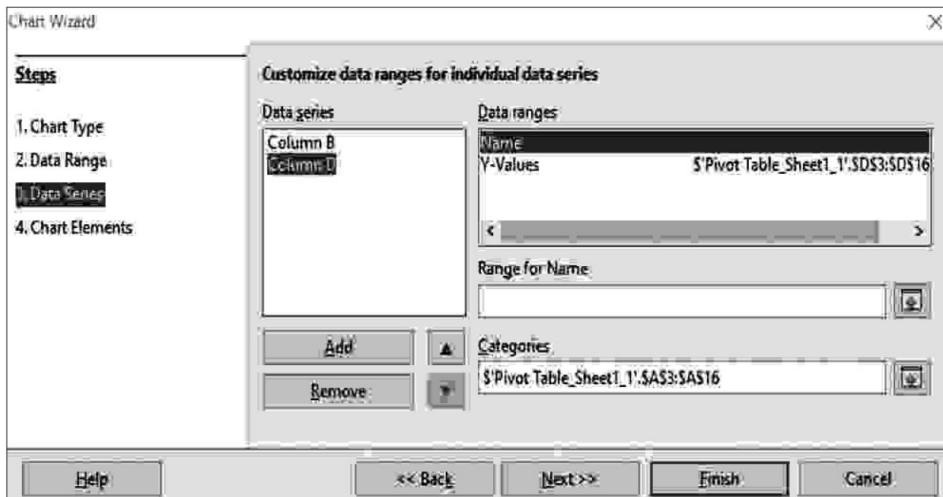
Khi biểu đồ Cột và Đường được chọn, hãy chuyển đến "2. Phạm vi dữ liệu" để xem phạm vi dữ liệu có trong biểu đồ.



Có vẻ như dữ liệu như mong muốn, nhưng cần phải loại bỏ một số dữ liệu để làm cho bảng "sạch hơn". Trong trường hợp này, hãy chuyển sang "3. Chuỗi dữ liệu" để xem tất cả các chuỗi có trên biểu đồ.



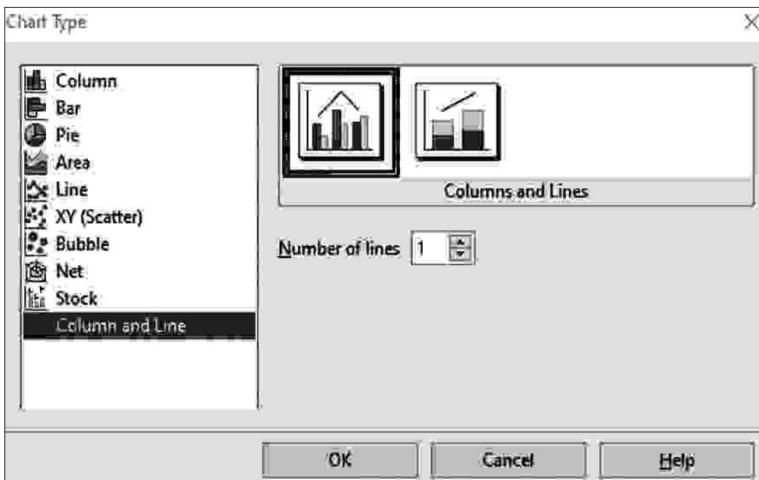
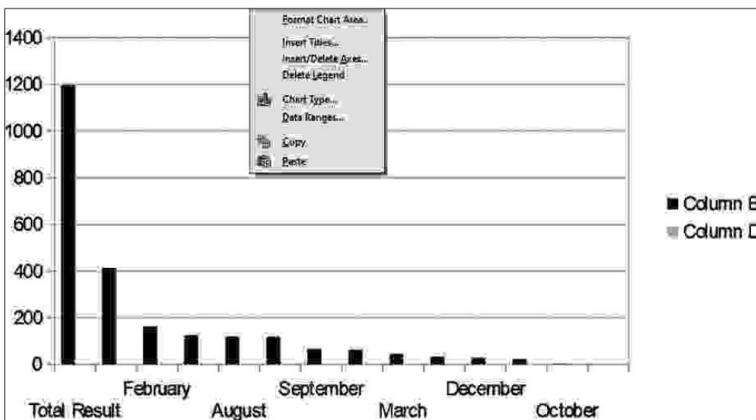
Khi nhà phân tích xem xét biểu đồ và bảng, nên xóa Cột E, cột không đóng góp vào bảng và Cột C, chỉ là tổng số đang chạy, để lại Cột B, hiển thị theo thứ tự giảm dần độ dài cơn lốc xoáy theo tháng và Cột D, là tỷ lệ phần trăm của các độ dài đó. Việc xóa chúng rất đơn giản-chọn cột và nhấp vào nút "Xóa". Màn hình hoàn thành như sau:



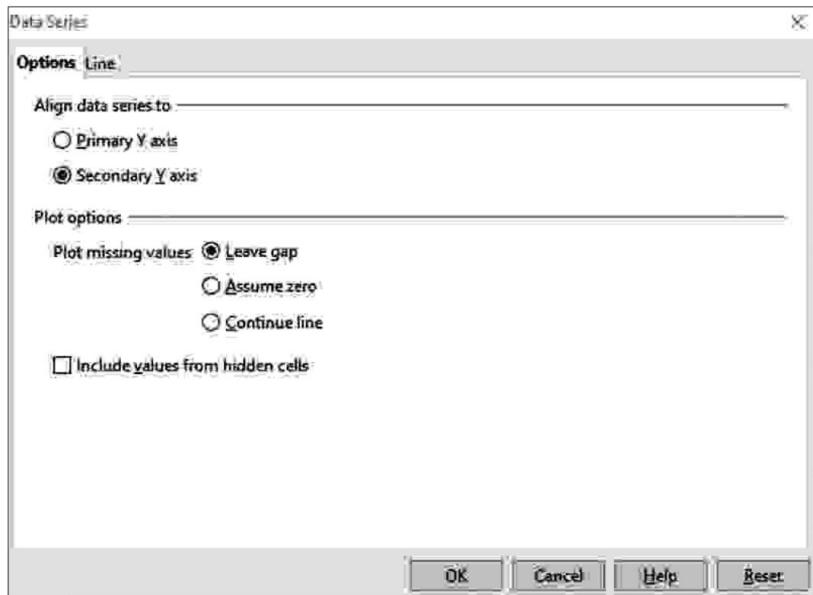
Tại thời điểm này, nhà phân tích cần chuyển sự chú ý của họ sang biểu đồ, biểu đồ hiện cho thấy những gì dường như chỉ là một yếu tố. Lý giải cho vấn đề này là

bởi vì giá trị cao nhất của phần tử khác là 100 (100%) và không xuất hiện trong phạm vi của các số. Nhà phân tích sẽ cần tạo một trục phụ cho tỷ lệ phần trăm này và điều đó sẽ cho phép dữ liệu xuất hiện.

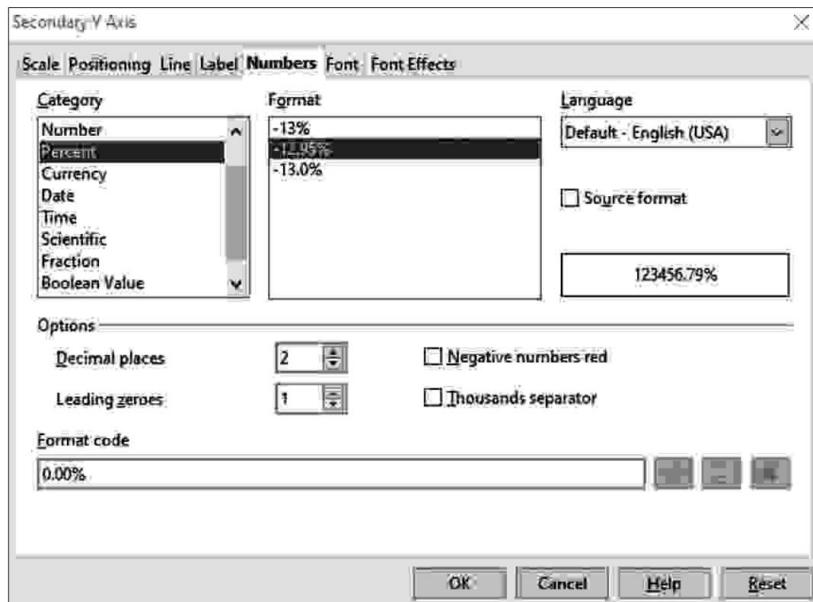
Bước tiếp theo sẽ là thiết lập trục thứ cấp. Điều này được thực hiện bằng cách đầu tiên thiết lập một dòng cho tổng số phần trăm chạy (Cột D). Nhấp chuột phải vào biểu đồ và chọn "Chart Type" như hình:



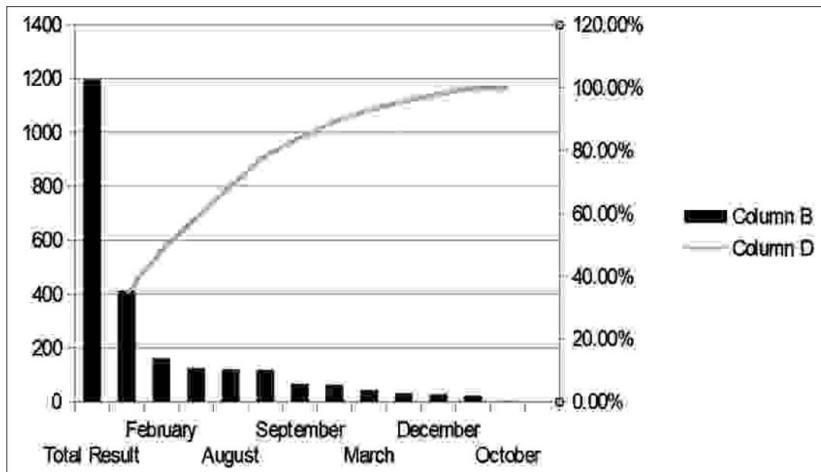
Bằng cách thực hiện chức năng này, nhà phân tích sau đó sẽ hiển thị một đường màu vàng đi qua trục x. Dòng này hiển thị tổng số phần trăm đang chạy. Mục tiêu là để dòng này hiển thị trên cùng một biểu đồ với các cột. Nhấp đúp vào dòng màu vàng và màn hình sau sẽ xuất hiện. Đảm bảo rằng bạn gán dòng này cho "Trục y phụ", vì mục tiêu là hiển thị cả số nguyên và tỷ lệ phần trăm trên cùng một biểu đồ.



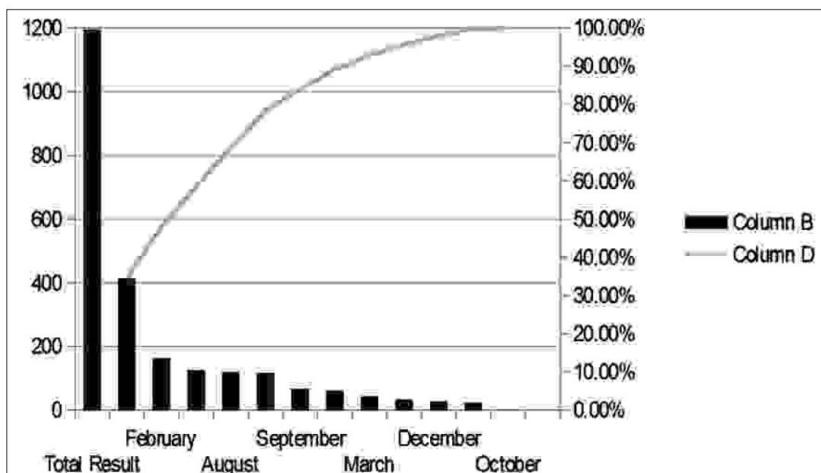
Điều này sẽ tạo ra biểu đồ sau, biểu đồ này không hiển thị tỷ lệ phần trăm thực nhưng có nhiều xác suất hơn trong khoảng từ 0 đến 1. Để thay đổi biểu đồ này thành tỷ lệ phần trăm tuổi, hãy nhập đúp vào các số ở bên phải và màn hình này sẽ xuất hiện.



Nhấp vào tab được đánh dấu là “Số” và xóa dấu kiểm trong hộp kiểm có nhãn “Định dạng nguồn”, sau đó chọn “Phản trǎm” từ danh sách bên trái. Nhấn OK và biểu đồ sau sẽ xuất hiện.



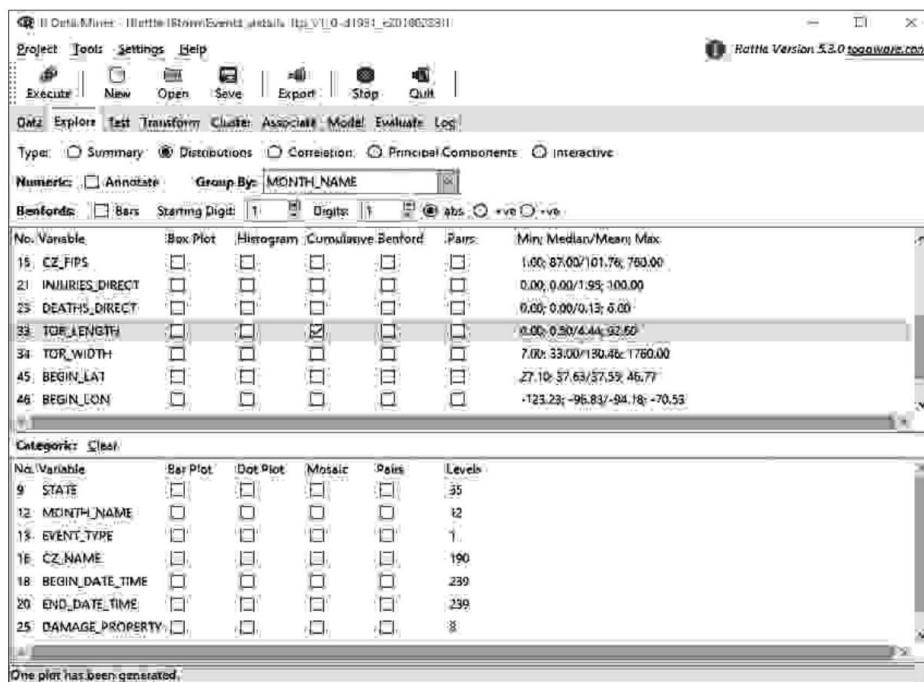
Tại thời điểm này, hãy khám phá một số chức năng khác trong biểu đồ để giảm các con số nhằm loại bỏ khoảng trống bổ sung ở đầu các giá trị sao cho biểu đồ trông giống như biểu đồ tiếp theo. Một lượng công việc đáng kể, nhưng kết quả tương tự. Khi nhà phân tích thực hành chức năng này, nó sẽ trở thành bản chất thứ hai.



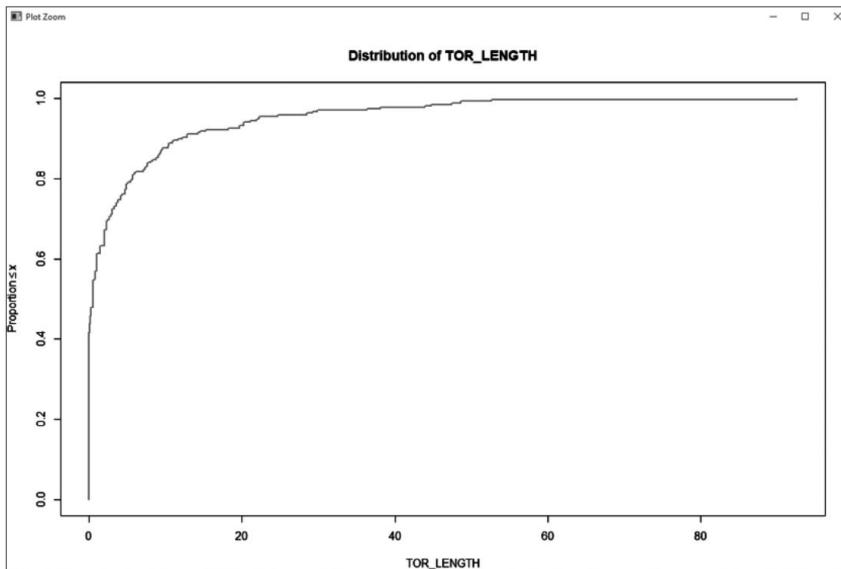
3.2.3 R/RStudio/Rattle

Quy trình tạo biểu đồ trong Rattle rất đơn giản. Tuy nhiên, quy trình để tạo ra một biểu đồ tương tự như những biểu đồ đã được trình bày phức tạp hơn nhiều. Vì quy trình tạo biểu đồ thông thường đơn giản hơn nên đối tượng sẽ hơi chêch hướng khỏi con đường chính cho công cụ này trong quy ước này.

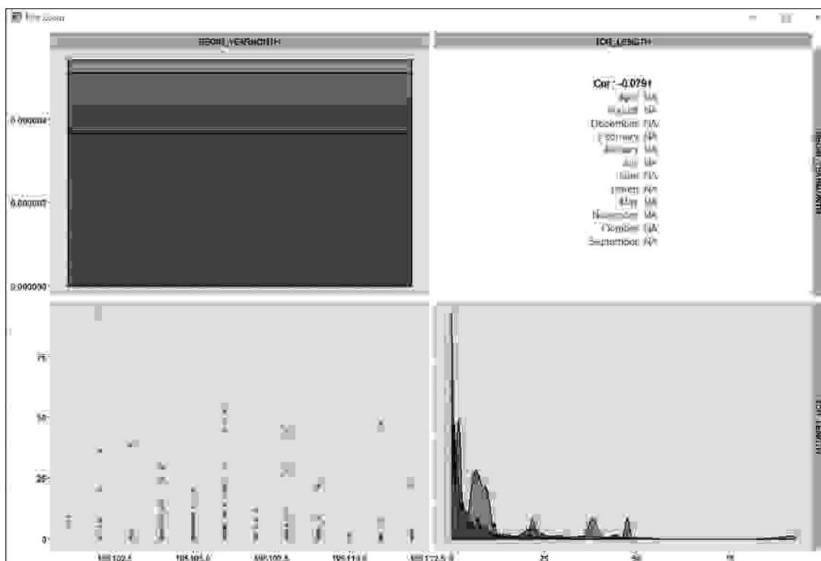
Trong Rattle, có một tab “Khám phá” cung cấp nhiều chức năng khác nhau cho nhà phân tích. Một trong số đó là “Phân phối”, cung cấp cho nhà phân tích một loạt các trực quan hóa dữ liệu và cùng nhau, có thể thực hiện tương tự như biểu đồ xác suất tích lũy được trình bày trong các phần trước. Phân phối đầu tiên là biểu đồ “Tích lũy”, đủ đơn giản để chọn bằng cách sử dụng hộp kiểm trên khu vực “Phân phối”, được hiển thị như sau.



Trên màn hình này, có hộp thả xuống “Nhóm theo:”, nơi nhà phân tích có thể chọn biến theo đó biến phân loại chính được nhóm, giống như trong bảng tổng hợp. Trong trường hợp này, biến số TOR_LENGTH không được ghép nối với bất kỳ biến nào khác. Nói cách khác, biểu đồ kết quả được vẽ như sau không có liên kết với MONTH_NAME như trong các phần trước.



Biểu đồ trước đó cho thấy khoảng 80% cơn lốc xoáy có chiều dài dưới 15, nhưng nó không liên quan đến chiều dài của MONTH_NAME. Có một biểu đồ mới quan hệ mô tả điều này, nhưng không giống như trong các phần trước. Đây là hộp kiểm “Cặp”, nơi nhà phân tích có thể chọn cả TOR_LENGTH và MONTH_NAME như sau:



Để có biểu đồ xác suất tích lũy thực sự, quá trình chuyển đổi dữ liệu sẽ phải được hoàn thành và các biểu đồ dữ liệu theo chương trình sẽ phải diễn ra để tạo ra kết quả tương tự như kết quả trước đó. Để làm được điều này, nhà phân tích sẽ phải dựa vào RStudio để thực hiện phần này chứ không phải Rattle. Chúng tôi khuyến khích nhà phân tích khám phá thêm các bản phân phối Rattle và các chức năng tương tự để xem riêng.

Trong thời gian chờ đợi, hãy quay lại RStudio và biểu đồ xác suất tích lũy. Bước đầu tiên cần làm là nhập tập dữ liệu vào RStudio, được trình bày trong phần Nhập dữ liệu. Bước tiếp theo là tách riêng các biến mà chúng tôi muốn sử dụng cho biểu đồ xác suất tích lũy của mình, đó là MONTH_NAME và TOR_LENGTH. RStudio cung cấp một Môi trường Phát triển Tích hợp (IDE) rất đẹp, được mặc định ở phía dưới bên trái của màn hình.

Màn hình sau đây cho thấy một số chương trình được thực hiện để cô lập các biến và tạo tổng thô tích lũy của độ dài cơn lốc xoáy. Bảng kết quả như sau, cùng với chương trình.

	MONTH_NAME	TOR_LENGTH
12	June	413.0
11	February	162.3
10	May	125.0
9	August	119.9
8	April	117.7
7	September	66.5
6	November	61.8
5	March	43.7
4	July	32.9
3	December	28.0
2	January	22.5
1	October	2.0

```
xyz<-xy%>%group_by(MONTH_NAME)%>%
```

```
tóm tắt(TOR_LENGTH=sum(TOR_LENGTH))
```

Điều mà chương trình trước đạt được là nhóm dữ liệu theo MONTH_NAME và TOR_LENGTH, tổng hợp độ dài cơn lốc xoáy thô và nhóm chúng theo tháng, giống như đã được thực hiện trong Bảng tổng hợp ở các phần trước. Bây giờ là nhiệm vụ thực hiện tỷ lệ phần trăm tích lũy và sau đó vẽ biểu đồ đó trong RStudio. Kết quả phụ thuộc vào việc thêm một cột

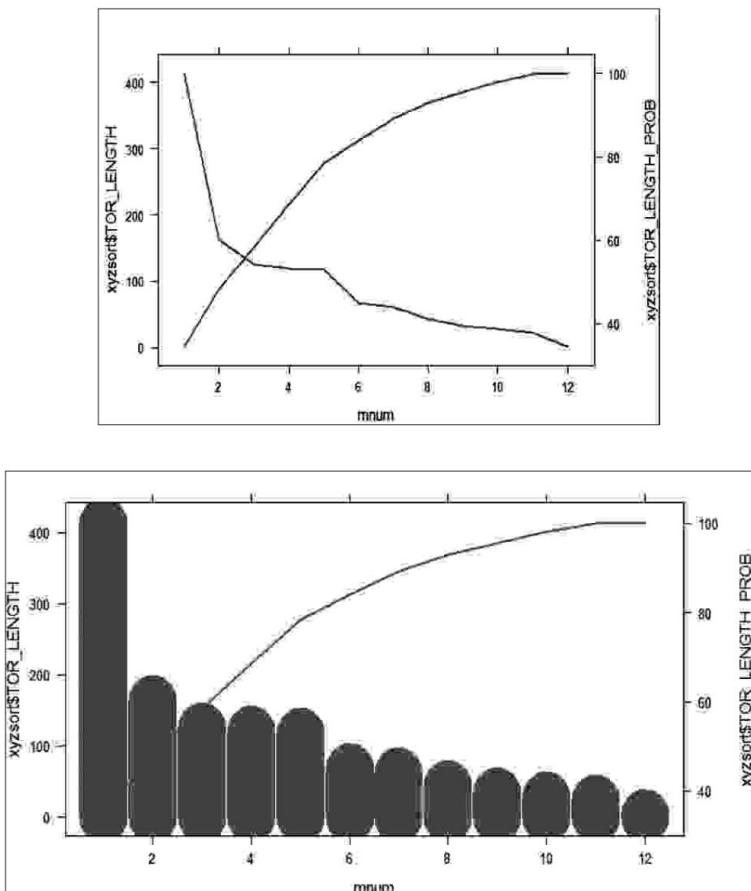
với các khoản tiền tích lũy và sau đó lấy các khoản tiền đó và thay đổi chúng thành tỷ lệ phần trăm. Màn hình và lập trình hiển thị bảng cần thiết để tạo biểu đồ xác suất tích lũy.

Bây giờ đến phần thử thách. Mọi thứ trong bảng chính xác như khi tạo biểu đồ trong các ứng dụng như Excel hoặc OpenOffice.

Thật không may, R không được tạo cho các biểu đồ có hai trục y như phải được thực hiện trong trường hợp cụ thể này. Do đó, đã đến lúc thực hiện thêm một số chương trình để hiển thị đúng biểu đồ này. Thách thức là làm điều đó để kết quả sẽ phản ánh kết quả từ các công cụ khác.

	MONTH_NAME	TOR_LENGTH	TOR_LENGTH_CUM	TOR_LENGTH_PROB
1	June	413.0	413.0	34.55200
2	February	162.3	575.3	48.13018
3	May	125.0	700.3	58.58780
4	August	119.9	820.2	68.61876
5	April	117.7	937.9	78.45566
6	September	66.5	1004.4	84.02911
7	November	61.8	1066.2	89.19936
8	March	43.7	1109.9	92.85535
9	July	32.9	1142.8	95.60780
10	December	28.0	1170.8	97.95031
11	January	22.5	1193.3	99.83268
12	October	2.0	1195.3	100.00000

```
> xyzsort[, "TOR_LENGTH_CUM"]<-cumsum(xyzsort$TOR_LENGTH)
> xyzsort[, "TOR_LENGTH_PROB"]<-cumsum(xyzsort$TOR_
CHIỀU DÀI/1195.3)
> xyzsort[, "TOR_LENGTH_PROB"]<-cumsum((xyzsort$TOR_
CHIỀU DÀI/1195.3)*100)
> install.packages("latticeExtra")
```



Biểu đồ trước khá gần với biểu đồ xác suất tích lũy trước đó, nhưng vẫn có chỗ cho một số định dạng. Có rất nhiều chỗ để khám phá R, RStudio và Rattle và chúng tôi sẽ để nhà phân tích tiếp tục tinh chỉnh ứng dụng này. Lập trình để có được kết quả này như sau và tài nguyên cho các mẹo lập trình này nằm trong Tài liệu tham khảo phần phát sinh.

```
xy<-StormEvents_details_ftp_v1_0_d1951_c20160223_FIXED
```

NHẬN XÉT: Sử dụng "xy" làm biến cho tập dữ liệu đã nhập

```
xyz<-xy%>%group_by(MONTH_NAME)%>%summarize(length=sum(TOR_LENGTH))
```

NHẬN XÉT: Điều này hoạt động giống như Bảng tổng hợp, lấy độ dài cơn lốc xoáy (TOR_LENGTH) và nhóm chúng theo tháng (MONTH_TÊN)

```
xyzsort<-arrange(xyz,desc(xyz$TOR_LENGTH))
```

NHẬN XÉT: Điều này sắp xếp độ dài cơn lốc xoáy trong bộ dữ liệu

```
xyzsort[, "TOR_LENGTH_CUM"]<-cumsum(xyzsort$TOR_LENGTH)
```

NHẬN XÉT: Điều này cho phép tổng tích lũy của cơn lốc xoáy chiều dài sau khi sắp xếp

```
mnum<-c(1,2,3,4,5,6,7,8,9,10,11,12)
```

NHẬN XÉT: Điều này cung cấp số hàng được liên kết với MONTH_NAME (thận trọng: Điều này có nghĩa là "1" = Tháng 6, KHÔNG PHẢI Tháng 1 vì trình tự hàng tạo ra sự khác biệt.)

```
obj1<-xyplot(xyzsort$TOR_LENGTH ~ mnum, xyzsort,type="h",lwd=50)
```

NHẬN XÉT: Điều này yêu cầu phải cài đặt gói "LATTICEEXTRA" và thiết lập chuỗi đầu tiên cho biểu đồ. Biểu đồ trung bình loại "h"

```
obj2<-xyplot(xyzsort$TOR_LENGTH_PROB ~ mnum, xyzsort,type  
= "l",lwd=2,col="steelblue")
```

NHẬN XÉT: Điều này cũng yêu cầu gói LATTICEEXTRA và thiết lập chuỗi thứ hai cho biểu đồ. Loại "l" (chữ thường L) là dòng.

```
doubleYScale(obj1,obj2, add.ylab2=TRUE,use.style=FALSE)
```

NHẬN XÉT: Điều này cung cấp trực y kép mà bạn cần cho biểu đồ và vẽ biểu đồ một dưới dạng biểu đồ thanh và biểu đồ thứ hai (tích lũy) dưới dạng biểu đồ đường

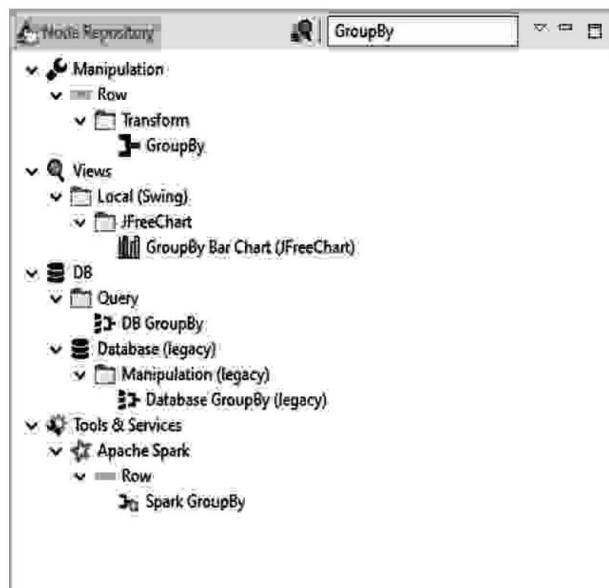
Vâng, điều này rất phức tạp, nhưng cho đến khi R cung cấp biểu đồ xác suất tích lũy tự động, đây chỉ là một trong nhiều cách lập trình để thực hiện ứng dụng này. Thật không may, lập trình không chỉ cần thiết trong trường hợp này mà còn là bắt buộc. May mắn thay, có rất nhiều tài liệu tham khảo cho hầu hết các loại ứng dụng này. Vui lòng khám phá chúng bằng công cụ tìm kiếm yêu thích của bạn.

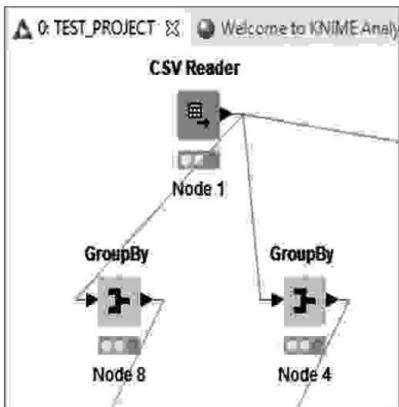
3.2.4 KIẾN THỨC

Công cụ KNIME cũng tương đối phức tạp khi nói đến biểu đồ xác suất tích lũy, nhưng có sẵn các nút để chuyển tập dữ liệu sang biểu đồ, sau khi thực hiện xong, có thể được điều chỉnh cho phù hợp với các tập dữ liệu khác có cấu trúc tương tự. Giống như biểu đồ luồng, các nút tạo thành cách tiếp cận từng bước để chuyển đổi dữ liệu thành biểu đồ. Quá trình này sẽ tiến hành từng nút một cho rõ ràng.

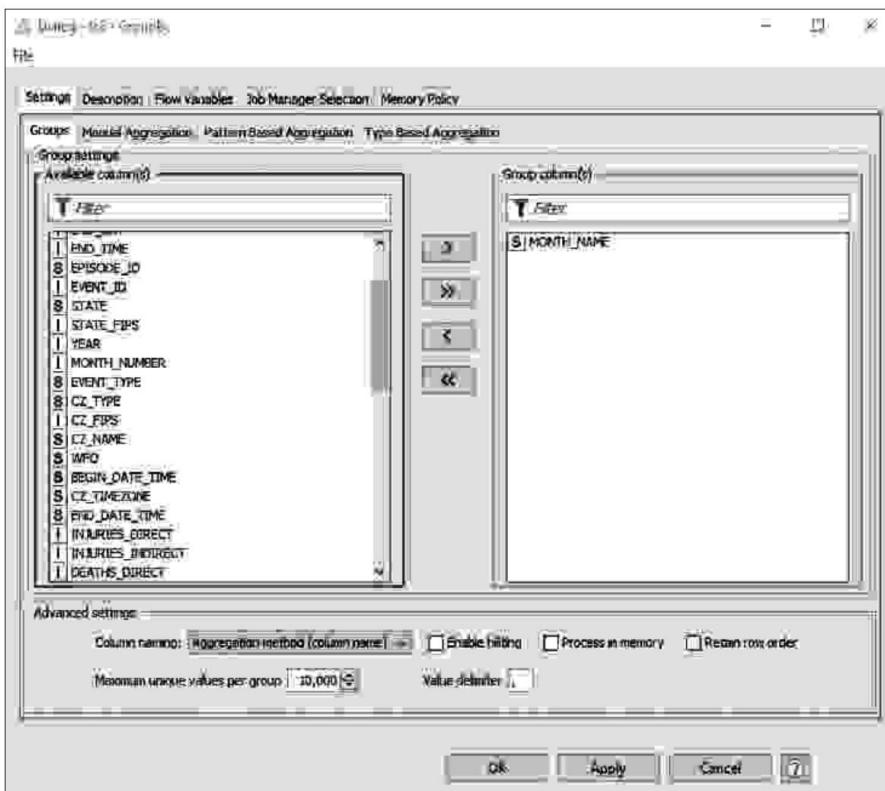
Trong phần cuối cùng trên KNIME, nút thống kê được gắn vào nút Trình đọc CSV để có được số liệu thống kê mô tả. Trong trường hợp này, sẽ có một số nút để có được một bảng phản ánh các nút trong các phần trước. Mỗi nút này sẽ chuyển đổi dữ liệu thành một sản phẩm cuối cùng mà sau đó nhà phân tích có thể xuất sang Excel, OpenOffice hoặc R. Trong trường hợp này, hãy xuất bảng đã hoàn thành sang OpenOffice như minh họa trong phần sau và sử dụng phần trên OpenOffice để tạo đầu ra CSV vào biểu đồ xác suất tích lũy. Đôi khi, việc xuất và sử dụng một công cụ khác sẽ dễ dàng hơn và sau đó có gắng làm phức tạp kết quả biểu đồ. Trong trường hợp này, xuất khẩu dễ dàng hơn.

Nút đầu tiên cần thiết để tạo bảng là nút "GroupBy", được tìm thấy bằng cách sử dụng khôi tìm kiếm trong nhóm menu phía dưới bên trái như được hiển thị. Nhà phân tích cũng có thể vào Thao tác -> Hàng -> Chuyển đổi -> NhómBy để đến nút. Nhấp chuột trái vào nút và kéo nút đó vào không gian làm việc và kết nối nút đó với nút Trình đọc CSV mà nhà phân tích đã sử dụng trước đó. Màn hình bây giờ sẽ xuất hiện như màn hình sau.





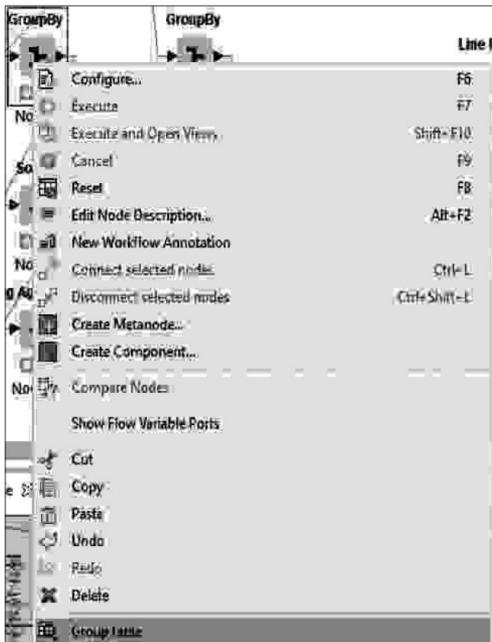
Màn hình hiển thị hai nút GroupBy cạnh nhau kết nối với nút Trình đọc CSV. Điều này là có chủ ý vì nhà phân tích sẽ cần cả hai nút này để kết hợp vào bảng mới cuối cùng. Nút đầu tiên sẽ là nút bên phải (có nhãn là Nút 4). Nhấp đúp vào nút này để hiển thị màn hình.



Sử dụng mũi tên đơn (>) để di chuyển MONTH_NAME qua để đưa nó vào Cài đặt nhóm. Điều này làm là nhóm các cột hoặc cột tiếp theo theo tên Nhóm, là MONTH_NAME. Về bản chất, bạn đang làm giống như di chuyển MONTH_NAME vào không gian "Cột" của Bảng Pivot trong Excel hoặc OpenOffice. Sau khi hoàn thành, vui lòng khám phá màn hình này để xem một số tùy chọn khác. Không có điều nào trong số này sẽ bị thay đổi với ứng dụng, nhưng trong tương lai, việc khám phá các tùy chọn này sẽ mang đến cho nhà phân tích nhiều dẫn xuất khác của công cụ và nút này.

Tiếp theo, di chuyển từ tab "Nhóm" sang tab "Tổng hợp thủ công" để xem màn hình sau. Giờ đây, việc di chuyển TOR_LENGTH vào cột như được hiển thị sẽ nhóm TOR_LENGTH theo MONTH_NAME. Đảm bảo rằng tùy chọn "tổng" được chọn từ trình đơn thả xuống cho biến này.

Nhấp vào OK hoặc Áp dụng và OK để định cấu hình nút và nhớ thực hiện để thương nút bằng cách sử dụng mũi tên màu xanh lục đơn hoặc mũi tên kép màu lục. Sau khi hoàn tất, nhấp chuột phải vào Nút 8 và chọn tùy chọn cuối cùng ở cuối menu phụ cho nút đó như được mô tả trong màn hình sau.



Khi tùy chọn Bảng nhóm được chọn, bảng kết quả sẽ xuất hiện như trong màn hình sau. Đây là bảng "xem trước" và sẽ cần được xuất nếu quá trình kết thúc tại đây. Hiện tại, điều này sẽ phục vụ như một trình giữ chỗ để đảm bảo kết quả là những gì nhà phân tích mong đợi.

Group table - 0:4 - GroupBy

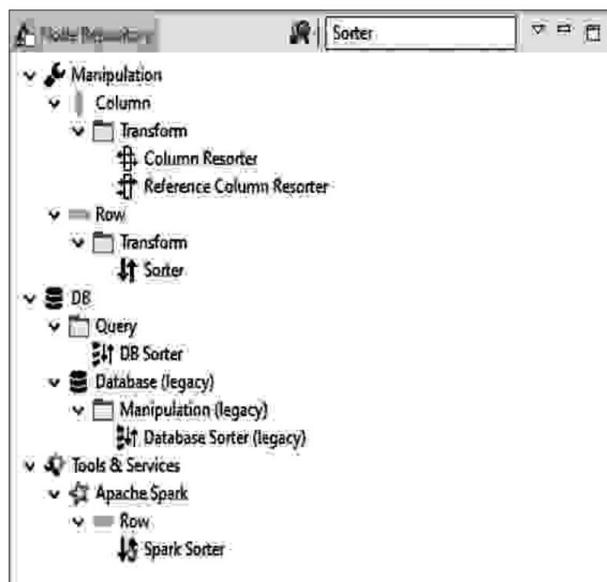
File Hilit Navigation View

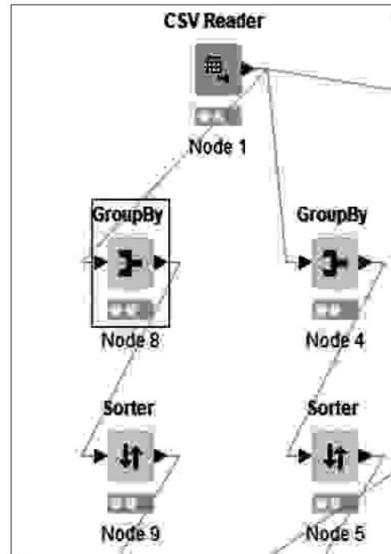
Table "default" - Rows: 12 Spec - Columns: 2 Properties

Row ID	MONTH...	Sum(T...)
Row0	April	117.7
Row1	August	119.9
Row2	December	28
Row3	February	162.3
Row4	January	22.5
Row5	July	32.9
Row6	June	413
Row7	March	43.7
Row8	May	125
Row9	November	61.8
Row10	October	2
Row11	September	66.5

Càng xa càng tốt. Những gì bảng này hiển thị là tổng chiều dài lốc xoáy mỗi tháng, nhưng lưu ý rằng các tháng theo thứ tự bảng chữ cái. Độ dài cơn lốc xoáy phải được sắp xếp theo thứ tự giảm dần để phản biến đồ thanh của biểu đồ xác suất tích lũy được hoàn chỉnh.

Nút tiếp theo để triển khai là nút Trình sắp xếp, nằm ở đây trong menu phụ phía dưới bên trái. Khi đã được chọn, kéo và thả và kết nối với nút GroupBy, nhấp đúp vào nút Sorter để hiển thị màn hình hiển thị các tùy chọn cấu hình khác nhau.





Một lưu ý trước khi tiếp tục với quá trình này. Xin lưu ý rằng GroupBy và bây giờ là Sorter được đặt trong danh mục "Hàng". Lúc đầu, điều này có thể gây nhầm lẫn vì nhà phân tích đang tìm cách sắp xếp cột, nhưng KNIME đặt Trình sắp xếp này vào danh mục Hàng vì mỗi hàng đang được sắp xếp như một phần của cột. Điều này nghe có vẻ không trực quan, nhưng nhà phân tích phải nghĩ về điều này trong suốt quá trình sử dụng KNIME. Điều đó không sai, chỉ là một góc nhìn khác.

Khi nút Bộ sắp xếp được kéo, thả và kết nối, bấm đúp vào nút để hiển thị màn hình này. Một lần nữa, bắt cứ khi nào nhà phân tích nhấp đúp vào nút, nó sẽ yêu cầu ứng dụng định cấu hình nút đó. Đây là cách dễ nhất để chuyển sang chế độ cấu hình. Ngoài ra, hãy nhớ rằng tại thời điểm này chúng ta đang làm việc trên Nút 5, phía bên tay phải của các nút trùng lặp.



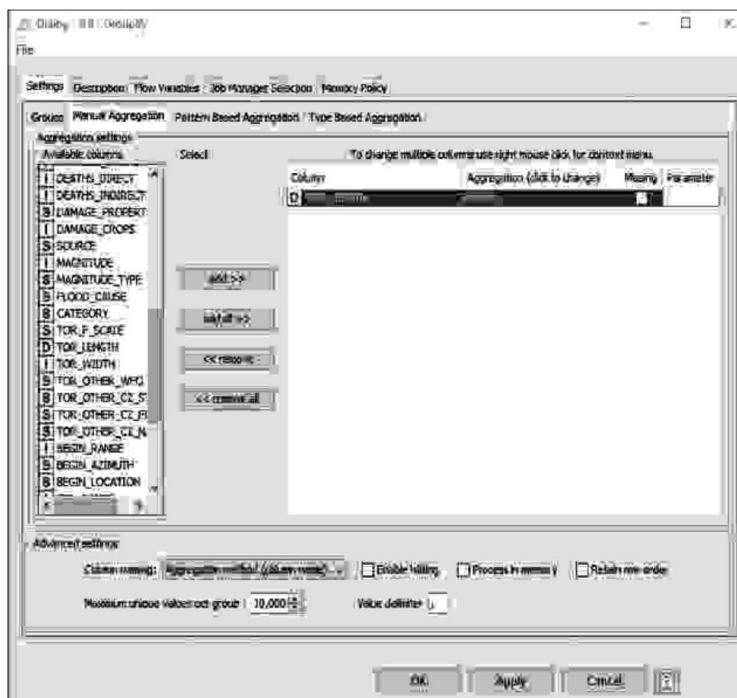
Màn hình cấu hình này rất đơn giản. Trước tiên, hãy đặt "Sắp xếp theo:" với "Tổng(TOR_LENGTH)" từ mũi tên xuống và đảm bảo rằng "Giảm dần" được chọn từ phía bên phải. Đừng lo lắng về khu vực được đánh dấu là "Thêm cột", vì điều này sẽ thêm các cột sẽ được đặt trong kho phân cấp thứ bậc, giống như việc chọn các cột sắp xếp bổ sung trong các ứng dụng khác.

Đừng lo lắng về các lựa chọn còn lại; nhấp vào Áp dụng và OK hoặc chỉ OK để định cấu hình nút. Tại thời điểm này, nút sẽ có màu "vàng" và sẽ cần được thực thi để kích hoạt luồng. Nhấp vào mũi tên đơn màu xanh lá cây trong khi nút này được chọn hoặc mũi tên kép màu xanh lá cây để thực hiện tất cả các nút. Bảng kết quả được khôi phục bằng cách nhấp chuột phải vào nút và chọn Bảng được sắp xếp. Màn hình đó được hiển thị như sau:

Row ID	MONTH...	Sum(T...
Row6	June	413
Row3	February	162.3
Row8	May	125
Row1	August	119.9
Row0	April	117.7
Row11	September	66.5
Row9	November	61.8
Row7	March	43.7
Row5	July	32.9
Row2	December	28
Row4	January	22.5
Row10	October	2

Nếu nhà phân tích tham khảo lại các công cụ khác vào thời điểm này, họ sẽ thấy rằng các con số đều giống nhau. Như đã nêu trước đây, đây là một cách tốt để xác minh quá trình phân tích. Bây giờ bảng đã được tính tổng và sắp xếp, bước tiếp theo là tính xác suất sao cho bảng có thể được sử dụng để theo dõi cả chiều dài cơn lốc xoáy và xác suất của những chiều dài đó (thực tế là tỷ lệ phần trăm, nhưng về cơ bản là giống nhau) so với tổng chiều dài trong suốt nhiều tháng.

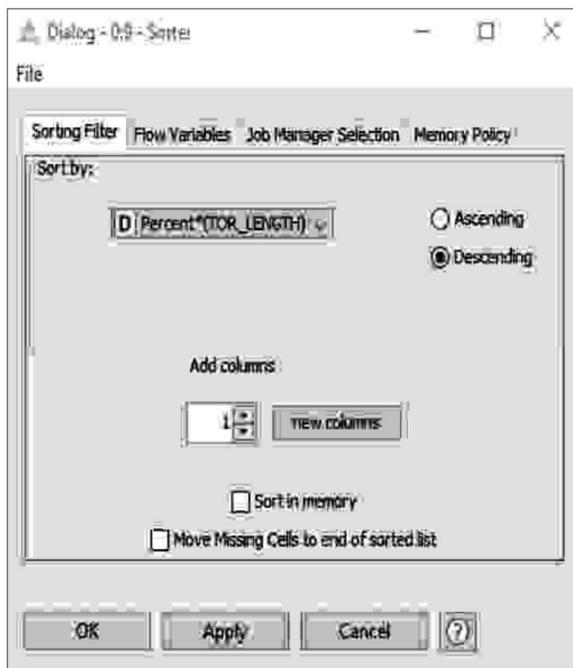
Vui lòng nhấp đúp vào Nút 8 để đến màn hình sau, đây là màn hình cấu hình cho phần trăm (hoặc xác suất) của độ dài cơn lốc xoáy.



Nhà phân tích sẽ nhận thấy rằng "Tổng hợp" hiện là "Phần trăm" thay vì "Tổng" để cột bây giờ sẽ có phần trăm. Tab "Nhóm" sẽ vẫn chọn cột MONTH_NAME vì đây là cột đầu tiên mà nhà phân tích muốn xem trong bảng. Khi Node 8 được định cấu hình và thực thi, hãy nhấp chuột phải và chọn "Bảng nhóm" để xem bảng kết quả từ luồng đó.

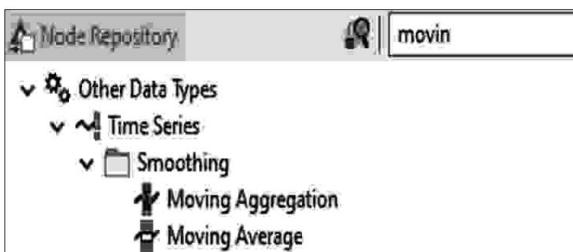
Group table - 08 - GroupBy		
File Hilitc Navigation View		
Table *default* - Rows: 12 Spec - Columns: 2 Prop		
Row ID	S MONTH...	D Percent...
Row0	April	10.037
Row1	August	10.037
Row2	December	3.717
Row3	February	4.461
Row4	January	1.115
Row5	July	8.55
Row6	June	28.996
Row7	March	2.23
Row8	May	21.561
Row9	November	4.461
Row10	October	0.743
Row11	September	4.089

Như nhà phân tích có thể thấy, tỷ lệ phần trăm không được sắp xếp và sẽ cần được sắp xếp sao cho bảng khớp với các số thô đã được sắp xếp trước đó. Điều này sẽ yêu cầu một nút "Sorter" khác (Node 9 trong trường hợp này) và nút Sorter sẽ gần giống như Node 5, ngoại trừ thay vì tính tổng, nhà phân tích sẽ sắp xếp theo thứ tự giảm dần tỷ lệ phần trăm của độ dài cơn lốc xoáy.



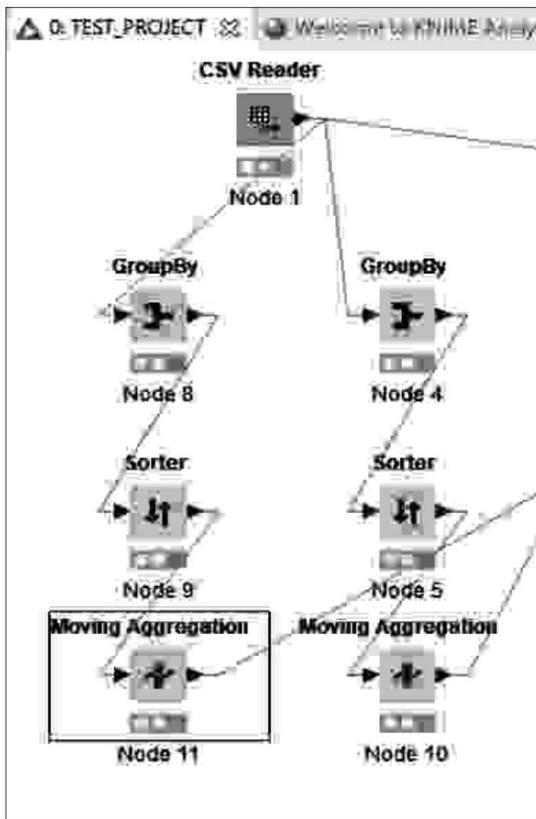
Như nhà phân tích có thể thấy, tab “Bộ lọc sắp xếp” trong nút này trông giống hệt như trong Nút 5, vì vậy không có sự khác biệt thực sự. Như đã nêu trước đây, vui lòng khám phá các tab khác và các tùy chọn khác trong các tab đó, vì chắc chắn có một số tab sẽ nâng cao trải nghiệm KNIME với phân tích dữ liệu.

Sau khi tính toán và sắp xếp xong, đó là lúc để tích lũy độ dài cơn lốc xoáy. Điều này được thực hiện thông qua một nút có tên là “Moving Aggregation,” nằm ở menu phụ bên trái như minh họa ở đây:

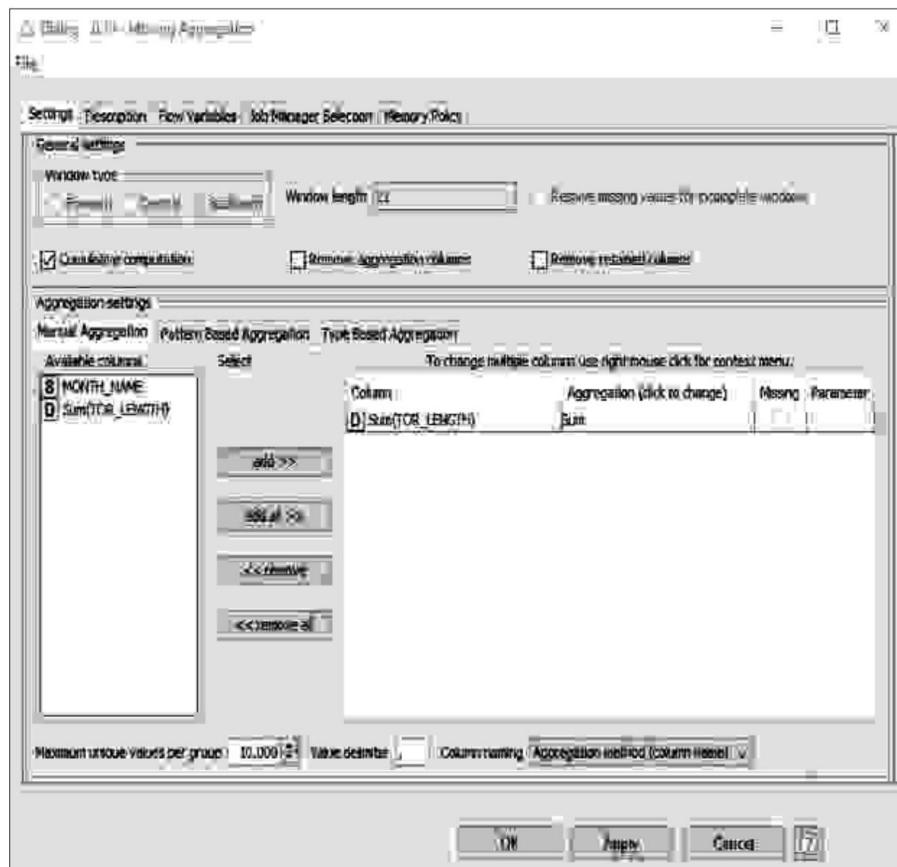


Như nhà phân tích có thể thấy, tên “movin” nằm trong hộp tìm kiếm, có nghĩa là không cần tiêu đề nút đầy đủ để đến nút mà phân tích phân tích cần.

Khi nút Tập hợp di chuyển được kéo, thả và kết nối, quy trình sẽ trông giống như quy trình tiếp theo. Xin lưu ý rằng nhà phân tích có thể di chuyển các nút xung quanh màn hình và đặt chúng vào bất kỳ cấu hình nào. Điều này giúp đọc sơ đồ quy trình làm việc cho nhà phân tích và những người có thể sử dụng quy trình sau khi nhà phân tích kết thúc cũng như chuyển đổi quy trình sang những người khác làm mẫu để nâng cao hơn nữa.



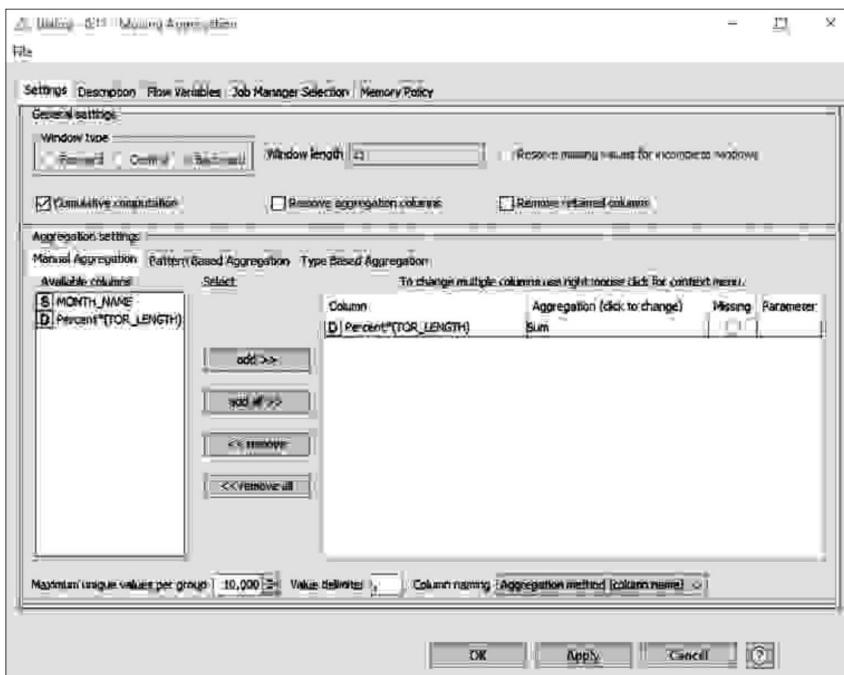
Nút đầu tiên sẽ được thảo luận là Nút 10 (phía dưới bên phải của màn hình trước). Nút này sẽ cung cấp một tính toán tích lũy về tổng chiều dài cơn lốc xoáy được sắp xếp ở trên nó (Nút 5). Nhấp đúp vào Nút 10 sẽ tạo ra màn hình này. Lưu ý rằng có một số tab này màn hình.



Xin hãy nhìn kỹ vào màn hình. Có một số lựa chọn cấu hình quan trọng trong bước này. Trước tiên, hãy đảm bảo rằng hộp “Tính toán tích lũy” được chọn, hộp này sẽ tô xám lựa chọn “Độ dài cửa sổ”. Đồng thời đảm bảo rằng cột chính xác được thêm vào bằng cách sử dụng nút “thêm >>”; trong trường hợp này, “Sum(TOR_LENGTH)” là cột mà nhà phân tích muốn tích lũy. Sau khi nhấp vào nút OK, nhà phân tích phải nhớ thực hiện nút này để kích hoạt quy trình luồng và hoàn thành các tính toán. Bảng được tạo ra từ toàn bộ quá trình này là màn hình sau. Vui lòng lưu ý cột bổ sung do quá trình này tạo ra.

Moving average values - (x10) - Moving Aggregation			
File Edit View			
Table "default" - Rows: 12 Spec Columns: 3 Properties Flow Variables			
Row ID	S MONTH_NAME	D Sum(TOR_LENGTH)	D Sum(SUM_TOR_LENGTH)
Row6	June	413	413
Row3	February	162.3	575.3
Row8	May	125	700.3
Row1	August	119.9	820.2
Row0	April	117.7	937.9
Row11	September	66.5	1,004.4
Row9	November	61.8	1,066.2
Row7	March	43.7	1,109.9
Row5	July	32.9	1,142.8
Row2	December	28	1,170.8
Row4	January	22.5	1,193.3
Row10	October	2	1,195.3

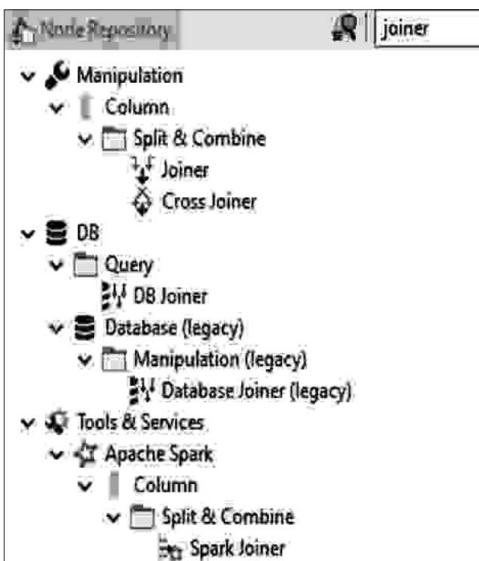
Bây giờ việc tích lũy số lượng đã hoàn tất, bước tiếp theo là tích lũy tỷ lệ phần trăm, được thực hiện theo cách tương tự (với các nút cùng loại) như phần trước. Lần này, hãy tập trung vào Nút 11, được minh họa như sau. Sự khác biệt duy nhất là nhà phân tích hiện đang tập trung vào cột phần trăm thay vì cột tổng, nhưng phép tính được yêu cầu vẫn sẽ là "tổng" như được mô tả ở đây.



Sau khi nút được định cấu hình và nút OK được nhấp, vui lòng nhớ nhấp vào Thực thi để kích hoạt quy trình. Việc hoàn thành bước này sẽ tạo ra bảng sau đây, là tỷ lệ phần trăm tích lũy (tổng số phải là 100).

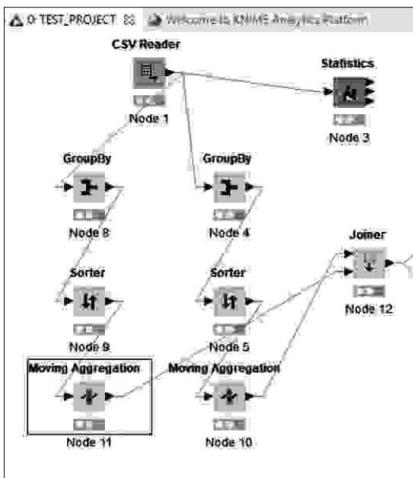
Row ID	MONTH...	Percent...	Sum(Pe...)
Row6	June	28.996	28.996
Row8	May	21.561	50.558
Row0	April	10.037	60.595
Row1	August	10.037	70.632
Row5	July	8.55	79.182
Row3	February	4.461	83.643
Row9	November	4.461	88.104
Row11	September	4.089	92.193
Row2	December	3.717	95.911
Row7	March	2.23	98.141
Row4	January	1.115	99.257
Row10	October	0.743	100

Bây giờ, các bảng kết quả từ cả số và tỷ lệ phần trăm phải được nối thành một bảng. Có một nút dành cho mọi thứ trong KNIME và việc tham gia cũng không có gì khác biệt. Tìm nút “Người tham gia” trong menu phụ ở phía dưới bên trái của màn hình KNIME bằng cách nhập tên vào hộp tìm kiếm.

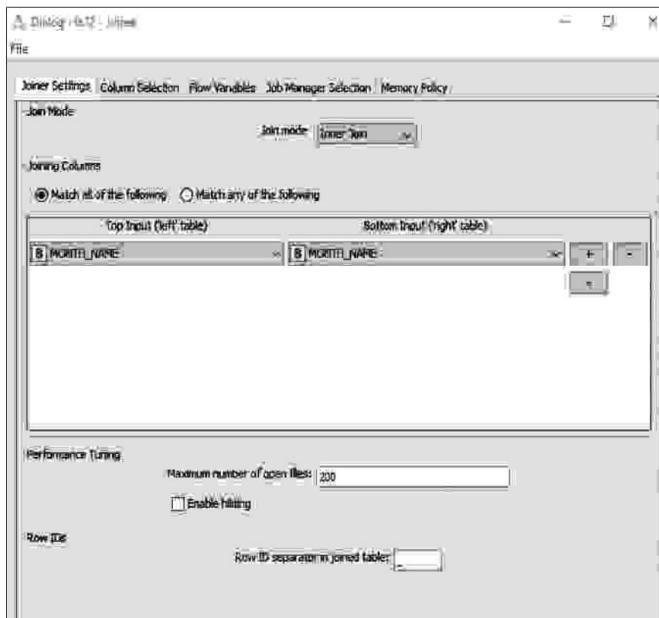


Trong trường hợp này, nút Joiner nằm trong danh mục "Cột".

Mục đích của Joiner là nối các cột, đây chính là điều mà nhà phân tích muốn trong trường hợp này. Sau khi nhà phân tích kéo và kết nối nút Joiner, quá trình kết quả sẽ giống như màn hình này.

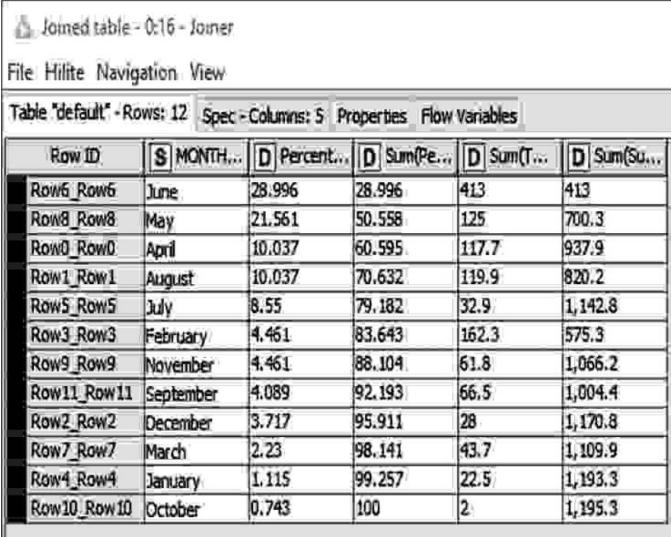


Vui lòng đảm bảo rằng cả hai nút đang Di chuyển tập hợp được kết nối với nút Tham gia. Có hai kết nối đầu vào với nút Joiner để có thể nối hai cột. Bấm đúp vào nút Tham gia sẽ hiển thị màn hình cấu hình.



Tab đầu tiên là “Cài đặt liên kết”, giúp nhà phân tích quyết định cột nào để “chốt” liên kết. Trong trường hợp này, cột “MONTH_NAME” tồn tại (và giống nhau) giữa hai nhóm cột được sử dụng, do đó, đó là cột khóa. “Chế độ tham gia” có một số tùy chọn, nhưng tùy chọn mặc định là tùy chọn được sử dụng lần này. Hãy nhớ rằng cả “Đầu vào trên cùng” và “Đầu vào dưới cùng” phải có cùng một cột để khóa nếu đó là mục đích của nút. Để giữ các cột có cùng phép tính và theo đúng thứ tự được sắp xếp, nhà phân tích không cần phải làm gì khác vào thời điểm này.

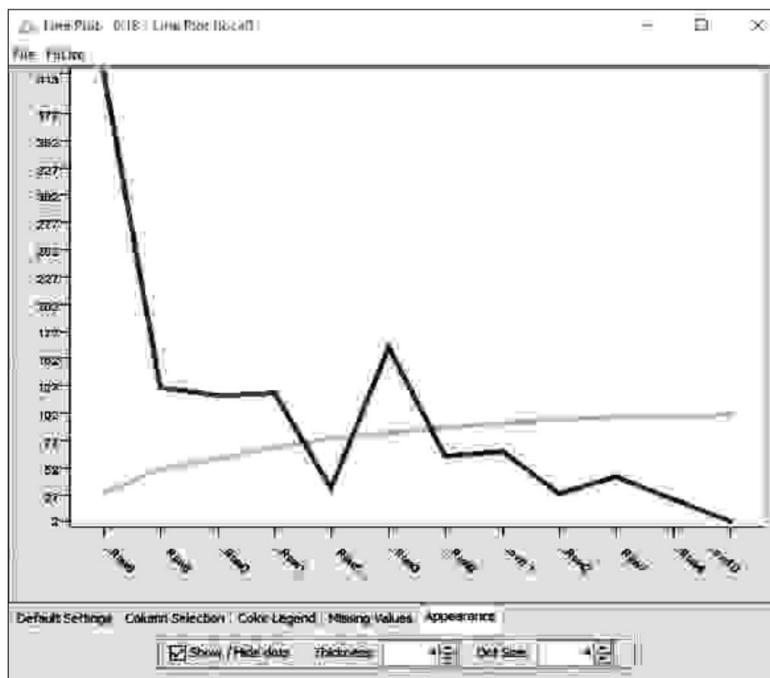
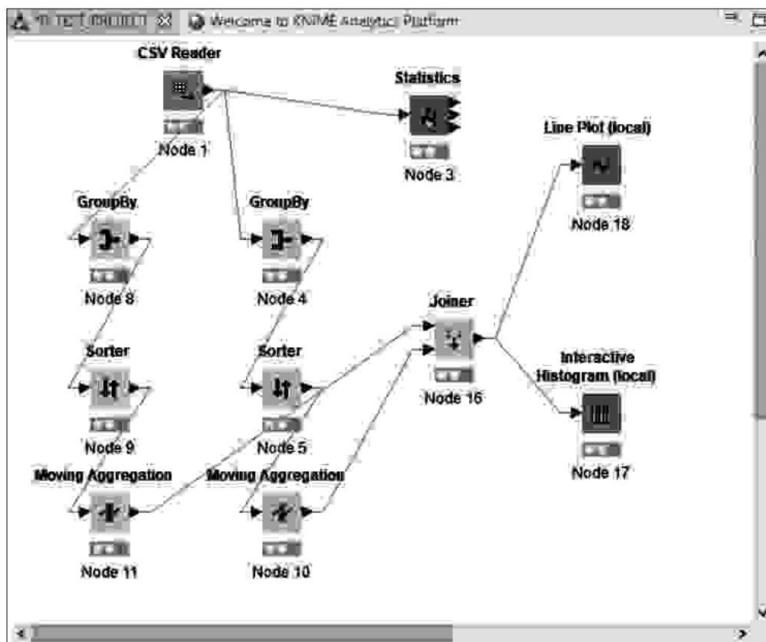
Bảng kết quả được tìm thấy bằng cách nhấp chuột phải vào nút Trình tham gia và chọn tùy chọn ở cuối menu phụ có tên là “Bảng đã tham gia”. Bảng xuất hiện như trong màn hình sau.



The screenshot shows a KNIME interface window titled "Joined table - 0:16 - Joiner". The window has tabs at the top: File, Help, Navigation, View, Spec, Columns, Properties, and Flow Variables. The "Spec" tab is selected. Below the tabs is a table with the following data:

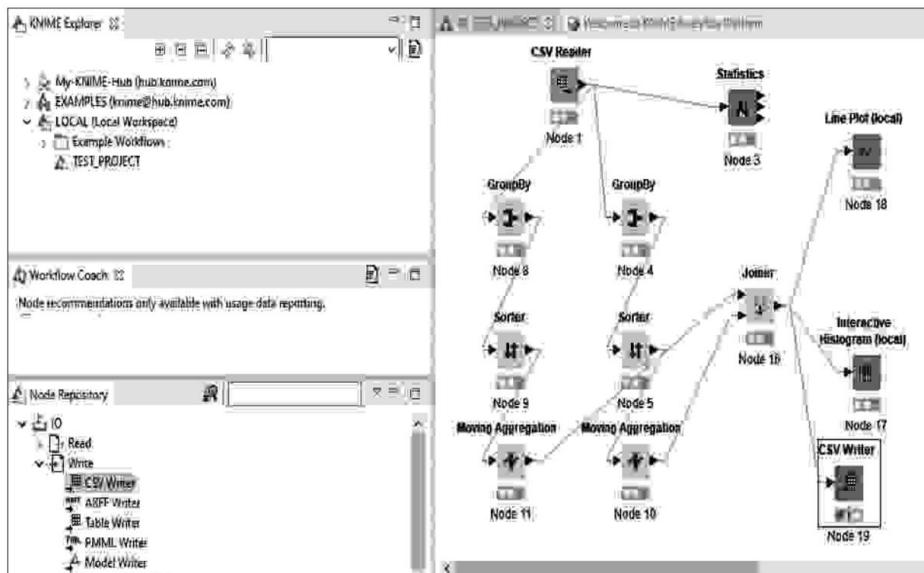
Row ID	MONTH...	D Percent...	D Sum(Pe...	D Sum(T...	D Sum(Su...
Row6_Row6	June	28.996	28.996	413	413
Row8_Row8	May	21.561	50.558	125	700.3
Row0_Row0	April	10.037	60.595	117.7	937.9
Row1_Row1	August	10.037	70.632	119.9	820.2
Row5_Row5	July	8.55	79.182	32.9	1,142.8
Row3_Row3	February	4.451	83.643	162.3	575.3
Row9_Row9	November	4.451	88.104	61.8	1,066.2
Row11_Row11	September	4.089	92.193	66.5	1,004.4
Row2_Row2	December	3.717	95.911	28	1,170.8
Row7_Row7	March	2.23	98.141	43.7	1,109.9
Row4_Row4	January	1.115	99.257	22.5	1,193.3
Row10_Row10	October	0.743	100	2	1,195.3

Đây là bảng sẽ được hoàn thành với các biểu đồ. Thật không may, KNIME không có biểu đồ xác suất tích lũy và số lượng lập trình cần thiết để tạo ra biểu đồ này nằm ngoài phạm vi của cuốn sách này. KNIME có các nút cần thiết cho biểu đồ đường hoặc biểu đồ nằm trong menu phụ. Một ví dụ về nút được mô tả như sau, với kết quả sau màn hình đó. Điều này gần đến mức bạn có thể nhận được biểu đồ xác suất tích lũy với các nút KNIME có sẵn.

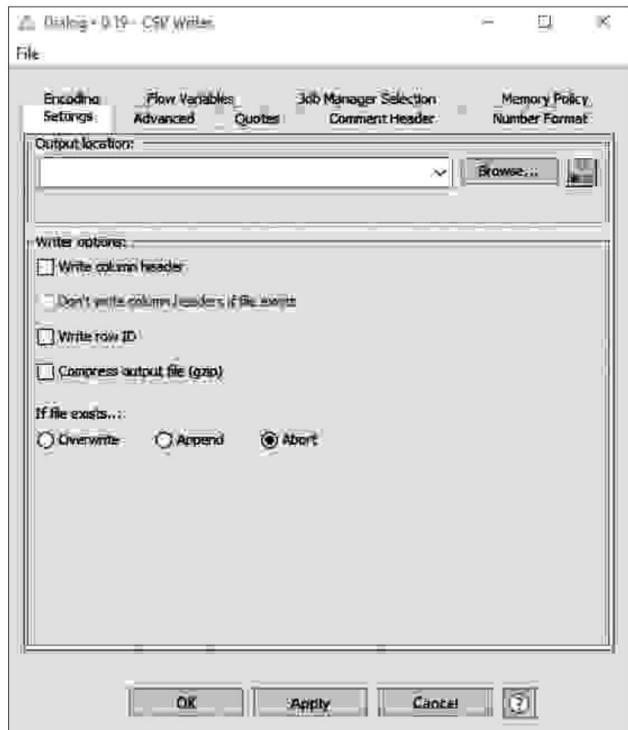


May mắn thay, KNIME có khả năng xuất (ghi) tệp vào nhiều ứng dụng thực hiện phân tích dữ liệu, một số trong đó có trong cuốn sách này. Tệp đủ cho phần này là tệp loại CSV.

Để xuất hoặc ghi tệp, có một nút được gọi là Trình ghi CSV, sau khi được kéo, đặt và kết nối, sẽ cung cấp khả năng xuất bảng đã hoàn thành sang tệp CSV. Sơ đồ quy trình làm việc sau cùng với vị trí của trình ghi CSV.



Bây giờ, nhấp đúp vào nút Trình ghi CSV để hiển thị màn hình cấu hình. Trong màn hình này, mục nhập chính là vị trí và tên tệp của tệp xuất. Trong trường hợp này, nhà phân tích có thể đặt vị trí này ở bất kỳ đâu từ máy tính cục bộ đến máy chủ mạng. Ngoài ra, hãy đảm bảo rằng hộp "Ghi tiêu đề cột" được chọn. Nếu không, các cột sẽ không có tiêu đề. Khi đã xong, tệp có thể được mở bằng công cụ mở tệp CSV. Trong trường hợp này, đây sẽ là OpenOffice. Khi đã xong, phần trên các biểu đồ xác suất tích lũy của OpenOffice có thể được triển khai.



Screenshot of Microsoft Excel showing a table of data. The table has columns labeled A through I. Column A contains month names from June to October. Columns B through E contain numerical values. The formula bar shows =MONTHNAME.

A	B	C	D	E
MONTH NAME	Percent(TOR LENGTH)	SumPercent(TOR LENGTH)	Sum(TOR LENGTH)	Sum(Sum(TOR LENGTH))
June	28.9862205279	28.9862205279	413	413
May	21.56133529	50.557209178	125	700.3
April	10.0371747212	60.594785539	117.7	937.9
August	10.0371747212	76.6315797602	119.9	820.2
July	8.5501858736	79.1821561338	32.9	1142.8
February	4.460965426	83.6431226756	182.3	575.3
November	4.460965426	88.1040852193	51.8	1066.2
September	4.0892153308	92.1933085502	86.5	1004.4
December	3.717472119	95.3107806591	28	1170.8
March	2.2304832714	98.1412639405	43.7	1109.9
January	1.152446357	99.2560055762	22.5	1133.3
October	0.7434944238	100	2	1195.3
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				

Nhớ là nếu trong tool chưa có chart thì xuất file ra và dùng tool có sẵn để làm chart nhé. Nếu một tính năng không có sẵn trong một công cụ, nó sẽ có sẵn trong công cụ kia (và có lẽ cũng dễ dàng hơn).

3.3 THỦ NGHIỆM T (Tham số)

Thủ nghiệm t là thử so sánh dữ liệu từ góc độ phương tiện. Bài kiểm tra rất có giá trị khi so sánh các mặt hàng như điểm kiểm tra, hàng tồn kho hoặc thậm chí nếu số lượng sản phẩm được đặt trong túi đáp ứng tiêu chuẩn cho số lượng đó (chẳng hạn như kẹo hoặc dinh). Nền tảng của bài kiểm tra t rất thú vị, nhưng điều đó tốt nhất nên để cho người hướng dẫn thống kê, vì những loại câu chuyện này giúp xây dựng sự hiểu biết tốt hơn về lý do tại sao khái niệm này được khởi xướng. Đối với những người quan tâm, tốt nhất là sử dụng công cụ tìm kiếm và nhập "Lịch sử Kiểm tra T của Học sinh". Sẽ có quá nhiều kết quả để hiểu rõ ràng về khái niệm này. Chỉ cần nói rằng kiểm tra t được sử dụng khi nhà phân tích có một mẫu dữ liệu và không biết độ lệch chuẩn tổng thể. Điều này đúng trong nhiều trường hợp phân tích dữ liệu. Không phải lúc nào cũng có thể tìm được độ lệch chuẩn của quần thể.

Một số người có thể đặt câu hỏi tại sao từ "Tham số" lại được đặt trong ngoặc đơn bên cạnh tiêu đề của phần này. Từ tham số khi được kết hợp với thống kê có nghĩa là phương pháp này có liên quan đến thuật toán như một phần của bảng hoặc phân phối chuẩn. Có những bài kiểm tra phi tham số chẳng hạn như lớp kiểm tra Wilcoxon, nhưng điều đó nằm ngoài phạm vi của cuốn sách này. Chỉ cần nói rằng các bài kiểm tra tham số là những bài kiểm tra mà hầu hết các nhà phân tích đã sử dụng trong quá khứ, cho dù chúng là kiểm tra chi bình phương, kiểm tra t hay kiểm tra Z. Vui lòng khám phá điều này để trở nên quen thuộc hơn với từ vựng phân tích dữ liệu.

3.3.1 Excel

Excel, thông qua ToolPak Phân tích, cung cấp một nền tảng hoàn hảo cho kiểm tra t. Quá trình thực hiện kiểm tra thống kê này tương đối đơn giản.

Đầu tiên, nhà phân tích mở ứng dụng Excel và tập dữ liệu, trong trường hợp này là cùng một ứng dụng được sử dụng cho các khái niệm và công cụ khác. Màn hình kết quả như sau, nhưng hãy hiểu rằng trang tính này chứa hai phần dữ liệu.

Đầu tiên là từ năm 1951 và thứ hai là từ năm 1954. Điều chúng tôi đang cố gắng xem là liệu chiều dài cơn lốc xoáy trung bình vào năm 1954 có lớn hơn năm 1951 hay không, mặc dù có nhiều hoạt động lốc xoáy hơn được ghi nhận vào năm 1954. Nếu đây là một giả thuyết, thì giả thuyết không sẽ là chiều dài cơn lốc xoáy trung bình

of 1951 = chiều dài cơn lốc xoáy trung bình của năm 1954; trong khi giả thuyết thay thế sẽ là chiều dài cơn lốc xoáy trung bình của năm 1951 nhỏ hơn ($<$) chiều dài cơn lốc xoáy trung bình của năm 1954.

Một giả định sẽ được đưa ra trong phần này là phương sai giữa hai bộ dữ liệu này là không bằng nhau. Sử dụng F-Test được cung cấp trong ToolPak Phân tích sẽ chứng minh điều này, điều này sẽ được thảo luận trong phần tiếp theo.

Bước đầu tiên cần làm là kết hợp các bộ dữ liệu lốc xoáy từ năm 1954 và từ năm 1951. Các bộ dữ liệu này có sẵn từ trang web được đề cập trong phần về nơi lấy dữ liệu. Đảm bảo rằng cả hai bộ dữ liệu này đều nằm trong cùng một trang tính. Điều tiếp theo mà nhà phân tích phải làm là mở Công cụ phân tích Pak, nằm trong tab Dữ liệu. Sau khi kích hoạt ToolPak, hãy chọn "T-Test: Hai mẫu giả định phương sai không bằng nhau," và tại thời điểm này, hãy di chuyển vào hai hộp văn bản các cột "TOR_LENGTH" từ năm 1951 (đầu tiên) và sau đó là năm 1954 (cột bên dưới cột đầu tiên hộp văn bản). Đảm bảo rằng "Nhân" được chọn để giải thích cho tiêu đề cột và chọn vị trí làm trang tính mới.

Sau khi hoàn thành việc này và bạn bấm OK, bạn sẽ nhận được kết quả như sau.

t-Test: Two-Sample Assuming Unequal Variances		
	TOR_LENGTH	TOR_LENGTH
Mean	4.443494424	5.322003284
Variance	104.6703773	114.6427058
Observations	269	609
Hypothesized Mean Difference	0	
df	535	
t Stat	-1.156176268	
P(T<t) one-tail	0.124062547	
t Critical one-tail	1.647706762	
P(T<t) two-tail	0.248125093	
t Critical two-tail	1.964408014	

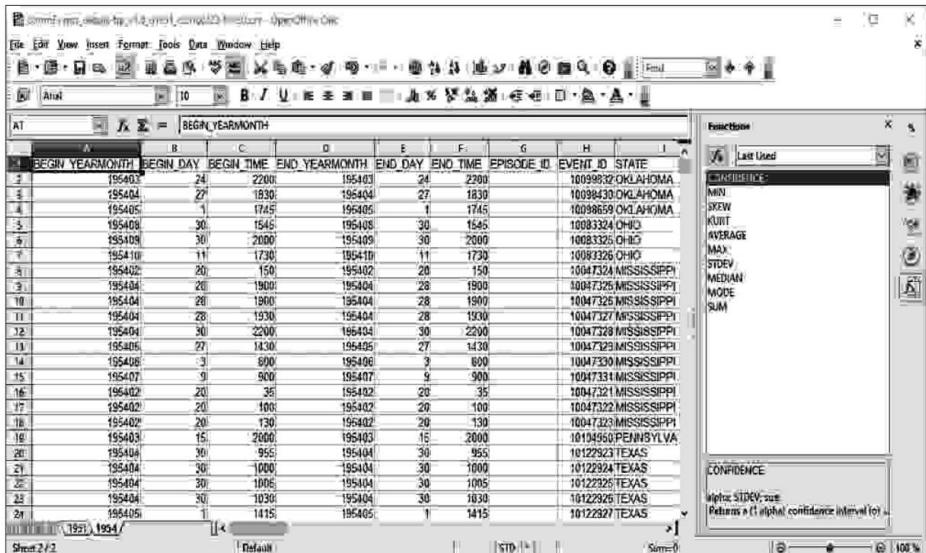
Tại thời điểm này, điều quan trọng là phải hiểu kết quả này mô tả điều gì. Về cơ bản, thử nghiệm đã được hoàn thành ở mức độ tin cậy 95%, có nghĩa là nếu " $P(T \leq t)$ một dưới" là $0,05$ hoặc thấp hơn ($<0,05$), thì giả thuyết không bị bác bỏ và chiều dài trung bình của hai cơn lốc xoáy là như nhau, trong khi nếu " $P(T \leq t)$ một dưới" lớn hơn $0,05$ ($>0,05$), thì giả thuyết không bị bác bỏ (trong một số vòng thống kê, từ này được chấp nhận, nhưng có nhiều tranh luận về từ này hoặc không bị bác bỏ). Điều này có nghĩa là, theo kiểm định t, nhà phân tích đã chỉ ra rằng chiều dài cơn lốc xoáy trung bình không khác nhau giữa năm 1951 và 1954 ở mức độ tin cậy 95%. Điều này một lần nữa già định phương sai không bằng nhau.

3.3.2 Văn phòng mở

Trong khi Excel cung cấp một ứng dụng tuyệt vời để thực hiện các bài kiểm tra t, thì OpenOffice lại dựa vào các công thức. Phần tiếp theo này sẽ thực hiện thử nghiệm tương tự trên cùng một bộ dữ liệu và nhận được kết quả tương tự.

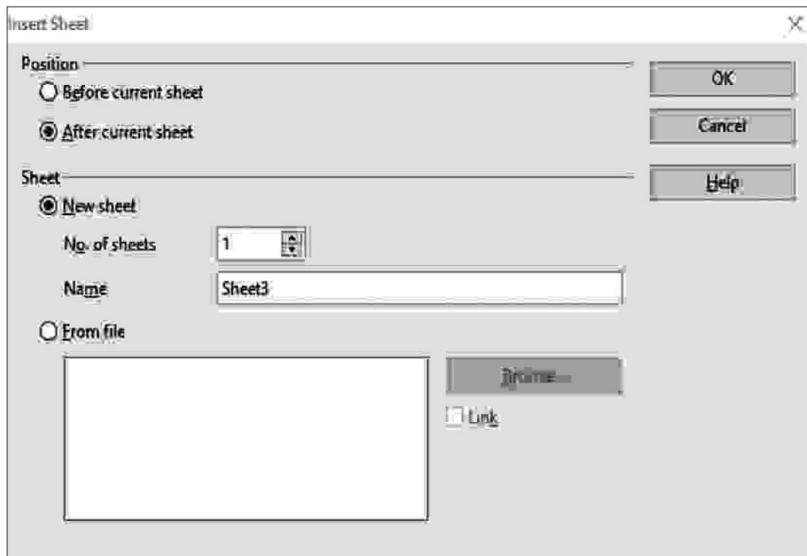
Bước đầu tiên, như với tất cả các phân tích dữ liệu, là nhập các bộ dữ liệu thích hợp. Điều này được thực hiện với chức năng nhập của OpenOffice, đã được trình bày trong phần nhập dữ liệu. Phương pháp kết hợp các trang tính vào một số làm việc cũng giống như với Excel, dẫn đến màn hình sau.

Có một cách khác để chèn các trang tính, như được hiển thị trong màn hình sau số làm việc OpenOffice. Trong trường hợp này, nhà phân tích đang chèn một trang tính từ một tệp, điều này cũng có thể được thực hiện trong Excel.

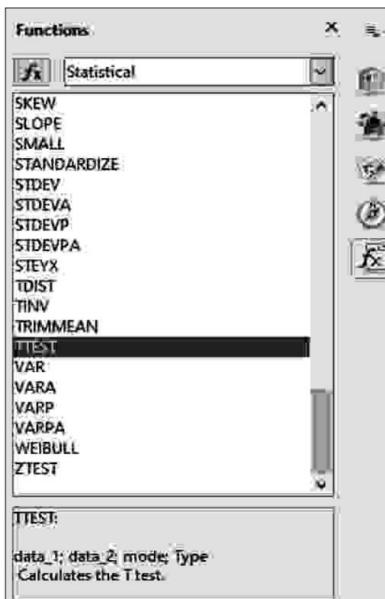


The screenshot shows a LibreOffice Calc spreadsheet with data in columns A through H. The data represents monthly events from 1954-01 to 1954-12, with columns for BEGIN YEAR/MONTH, BEGIN DAY, BEGIN TIME, END YEAR/MONTH, END DAY, END TIME, EPISODE ID, EVENT ID, and STATE. The Function Chooser dialog box is open on the right, showing the MIN function selected under the AVERAGE category. Other functions listed include MAX, STDEV, MEDIAN, MODE, and SUM.

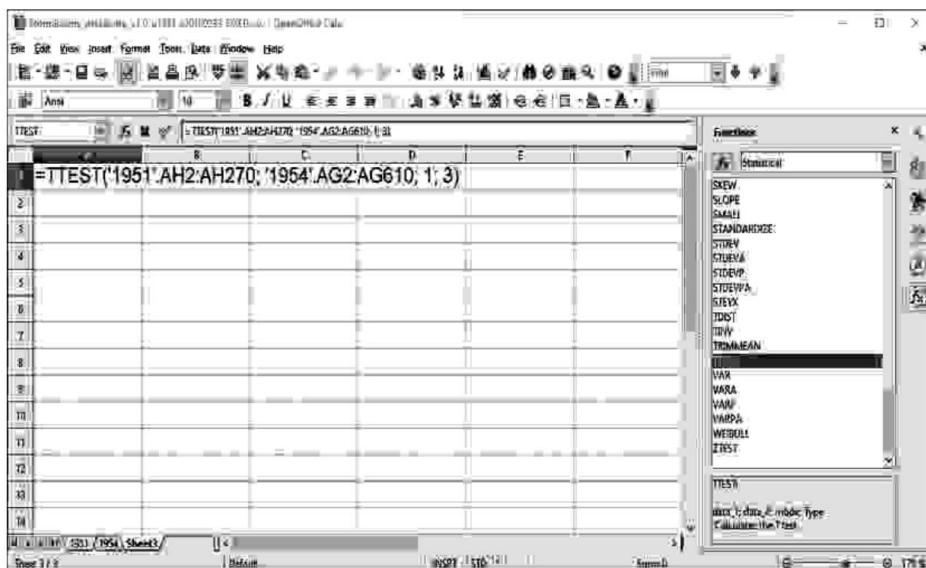
	A	B	C	D	E	F	G	H	
	BEGIN YEAR/MONTH	BEGIN DAY	BEGIN TIME	END YEAR/MONTH	END DAY	END TIME	EPISODE ID	EVENT ID	STATE
1	195401	24	2200	195401	24	2300	10099832	OKLAHOMA	
2	195404	27	1830	195404	27	1830	10099833	OKLAHOMA	
3	195405	1	1745	195405	1	1745	10099839	OKLAHOMA	
4	195408	30	1545	195408	30	1545	10083324	OHIO	
5	195409	30	2000	195409	30	2000	10083326	OHIO	
6	195410	11	1730	195410	11	1730	10083326	OHIO	
7	195402	20	150	195402	20	150	10047324	MISSISSIPPI	
8	195404	20	1900	195404	28	1900	10047325	MISSISSIPPI	
9	195404	28	1930	195404	28	1930	10047325	MISSISSIPPI	
10	195404	30	2200	195404	30	2200	10047228	MISSISSIPPI	
11	195401	28	1930	195404	28	1930	10047327	MISSISSIPPI	
12	195404	30	2200	195404	30	2200	10047229	MISSISSIPPI	
13	195405	27	1430	195405	27	1430	10047329	MISSISSIPPI	
14	195406	3	800	195406	3	800	10047330	MISSISSIPPI	
15	195407	9	900	195407	9	900	10047331	MISSISSIPPI	
16	195402	20	35	195402	20	35	10047321	MISSISSIPPI	
17	195402	20	100	195402	20	100	10047322	MISSISSIPPI	
18	195402	20	130	195402	20	130	10047323	MISSISSIPPI	
19	195403	15	2000	195403	15	2000	10114950	PENNSYLVANIA	
20	195404	30	955	195404	30	955	10122923	TEXAS	
21	195404	30	1000	195404	30	1000	10122924	TEXAS	
22	195404	30	1005	195404	30	1005	10122925	TEXAS	
23	195405	1	1415	195405	1	1415	10122927	TEXAS	



Tại thời điểm này, sẽ cần phải sử dụng một số công thức để có được kết quả kiểm định t. Ở ngoài cùng bên phải của màn hình là năm biểu tượng, dưới cùng là trình hướng dẫn công thức. Vui lòng kích hoạt trình hướng dẫn đó, trình hướng dẫn này sẽ hiển thị một ngăn màn hình khác có tên là "Chức năng".



Trước khi TTEST được chọn, bạn nên chèn một trang tính mới và chọn ô A1. Điều đó sẽ cung cấp một nơi để kết quả của TTEST cư trú. Màn hình sau đây hiển thị công thức đã hoàn thành với chú thích bên dưới để cho biết một số tham số trong công thức cung cấp những gì.



Nhìn giữa các dấu chấm phẩy, giống như dấu phẩy để phân tách giữa các tham số trong Excel, hai tham số đầu tiên là nội dung ô của độ dài cơn lốc xoáy 1951 và 1954. Hai cái cuối cùng biểu thị chế độ và loại thử nghiệm t. Số "1" có nghĩa đây là bài kiểm tra một đầu. Vâng, điều đó có nghĩa là "2" có nghĩa là thử nghiệm hai đầu. Tham số cuối cùng là "3", có nghĩa đây là thử nghiệm hai mẫu với các phương sai không bằng nhau. Số "1" có nghĩa là thử nghiệm mẫu được ghép đôi và số "2" có nghĩa là thử nghiệm hai mẫu có phương sai bằng nhau. Một trang web tuyệt vời để lấy thông tin chi tiết về các bài kiểm tra t cho OpenOffice nằm ở đây: https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_TTEST_function.

Khi công thức được kích hoạt bằng cách nhấn ENTER, số sau sẽ xuất hiện: - 0,1240626151. Nếu nhà phân tích nhìn lại kết quả đầu ra của Excel, thì con số này là giá trị p cho thử nghiệm một đầu. Nó có nghĩa tương tự ở đây cũng như trong phần phân tích trước. Có vẻ như độ dài của cơn lốc xoáy được coi là bằng nhau theo thử nghiệm thống kê này.

3.3.3 R/RStudio/Rattle

Công cụ R, và cụ thể là Rattle, rất khó khăn khi tiến hành kiểm tra t, nhưng với một chút kiên nhẫn và một chút lập trình, mọi thứ sẽ ổn.

Để bắt đầu với Rattle, trước tiên hãy mở gói theo phương pháp được thảo luận trong phần tham chiếu nhập dữ liệu bằng Rattle, nhưng sẽ có một chút thay đổi đối với quá trình nhập này. Trước tiên, chuyển đổi dữ liệu bằng R để bạn có ba cột—MONTH_NAME, TOR_LENGTH_1951 và TOR_LENGTH_1954. Theo cách này, t-test sẽ được thiết lập giống như các phần trước trong OpenOffice và Excel.

Để thiết lập ba cột, trước tiên hãy nhập tập dữ liệu năm 1951, tập dữ liệu này đã được hoàn thành và thêm tập dữ liệu năm 1954, thao tác này được thực hiện giống như cách nhập tập dữ liệu năm 1951. Sau khi hoàn thành, nhà phân tích sẽ muốn tách riêng và tạo một bảng có ba cột.

Màn hình hiển thị cả hai bộ dữ liệu trong ngăn nguồn RStudio được hiển thị như sau. Hãy nhớ rằng nhà phân tích muốn rút ngắn các tên tệp này bằng các chữ cái và số. Sẽ dễ dàng hơn nhiều trong lập trình nếu chúng được rút ngắn.

```
tor1951<-StormEvents_details_ftp_v1_0_d1951_c20160223_
ĐÃ SỬA
```

```
tor1954<-StormEvents_details_ftp_v1_0_d1954_c20160223
```

Sau khi hoàn thành, nhà phân tích có thể thực hiện kiểm tra t từ các lệnh sau, một lần nữa giả định các phương sai không bằng nhau và sử dụng mức độ tin cậy 95%. Sau đây là các lệnh, tiếp theo là kiểm tra t kết quả. Các kết quả hoàn toàn giống như trong các phần trước, ngoại trừ khoảng tin cậy âm, điều này sẽ được giải thích trong phần sau.

```
t.test(tor1951$TOR_LENGTH,tor1954$TOR_LENGTH,alternative=
"less",paired=FALSE,var.equal=FALSE,conf.level=0.95)
```

Thử nghiệm t mẫu Welch Two

dữ liệu: tor1951\$TOR_LENGTH và tor1954\$TOR_LENGTH
 $t = -1,1562$, $df = 534,86$, giá trị $p = 0,1241$
 giả thuyết thay thế: sự khác biệt thực sự về phương tiện nhỏ hơn 0

Khoảng tin cậy 95 phần trăm:

-Inf 0,3734851

Ước tính mẫu:

trung bình của x trung bình
 của y 4,443494 5,322003

Kết quả trước khớp với giá trị p (.1241) từ các phần trước.

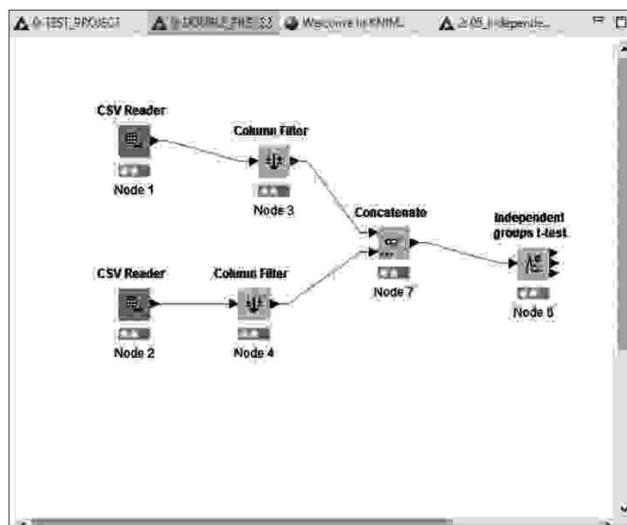
Ứng dụng R thậm chí còn viết ra giả thuyết thay thế cho bạn, điều này thật tiện lợi. Cho đến nay, tất cả các công cụ đều nhất trí với nhau, điều này chứng tỏ giá trị khi thuyết phục ai đó rằng kết quả đã được xác minh.

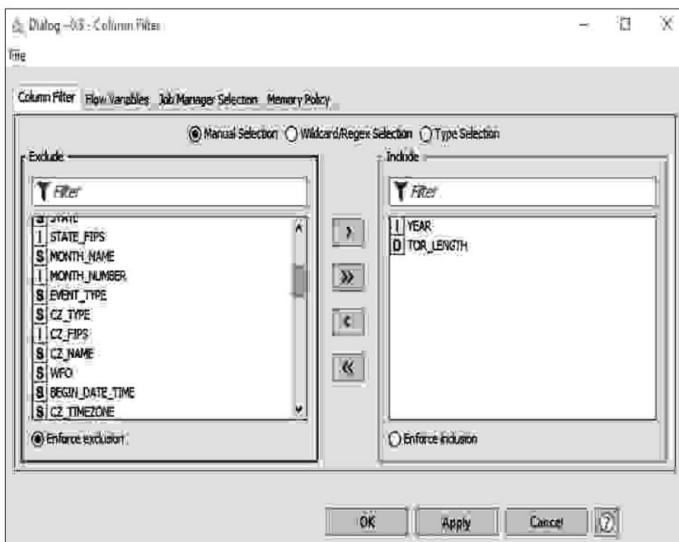
3.3.4 KIẾN THỨC

Ưu điểm chính mà KNIME có so với các công cụ khác là, nếu tồn tại một nút thực hiện chức năng kiểm tra, thì việc đặt nút đó trong dòng quy trình sẽ cho phép chuyển đổi và kiểm tra tập dữ liệu đó. Có một nút cho bài kiểm tra t tồn tại trong KNIME. Tuy nhiên, nhà phân tích phải đối mặt với việc kết hợp hai bộ dữ liệu để có thể so sánh độ dài cơn lốc xoáy với độ chính xác tương tự như trong các phần trước.

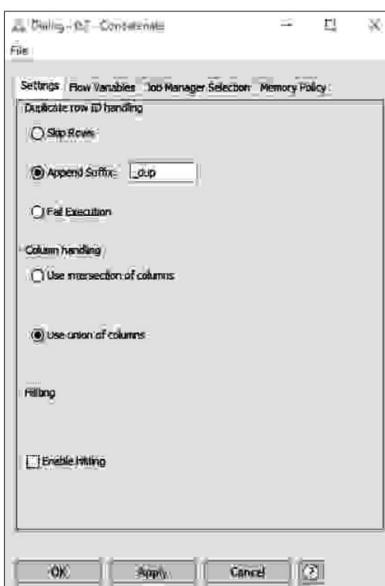
Trong KNIME, nếu có hai bộ dữ liệu được sử dụng, thì chỉ cần thêm một nút Trình đọc CSV khác vào đầu quy trình và đặt tệp thứ hai vào nút đó để sử dụng tiếp. Sau khi cả hai tệp được nhập qua các nút, thách thức sẽ đến để đảm bảo kiểm tra t được tìm thấy, kéo, đặt và kết nối đúng cách. Nút t-test nằm trong menu phụ bên trái (sử dụng hộp tìm kiếm và nhập "t-test" vào hộp đó). Nút được đặt và kết nối cho kiểm tra t được hiển thị trên màn hình sau cùng với các nút cần thiết để thực hiện kiểm tra t.

Bước đầu tiên là cài đặt các cột sẽ được sử dụng trong thử nghiệm, trong trường hợp này là chiều dài cơn lốc xoáy năm 1951 và 1954. Việc cách ly này sẽ được thực hiện với các nút Bộ lọc cột, các nút này sẽ được đặt đối diện với mỗi CSV như được hiển thị. Màn hình cho điều này được hiển thị sau quy trình làm việc.



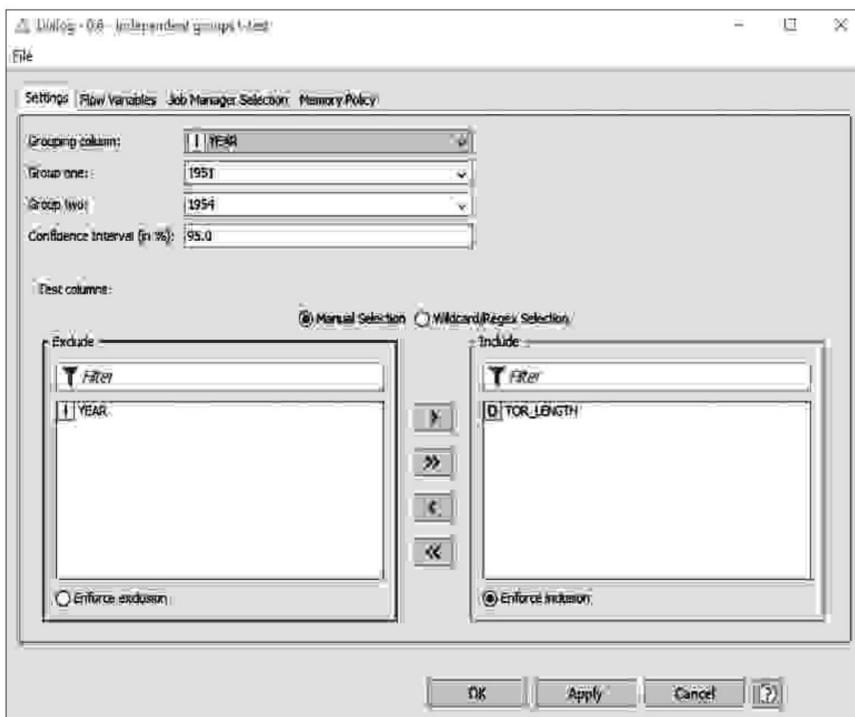


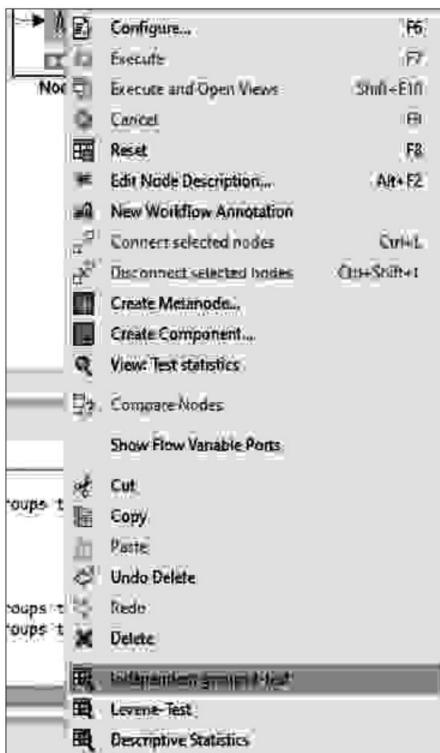
Trong cả hai nút của Bộ lọc cột, màn hình sẽ giống nhau, nhưng điều quan trọng cần lưu ý là chúng sẽ có hai năm riêng biệt, 1951 và 1954. Đó là lý do tại sao việc chọn cột NĂM sẽ giúp phân biệt các hàng sau khi hai cột được nối với nhau. Việc nối được thực hiện bằng cách sử dụng nút CONCATENATE, được hiển thị như sau. Trong trường hợp này, nhà phân tích sẽ muốn hợp nhất các hàng, vì điều đó sẽ thêm các hàng từ năm 1954 vào các hàng của năm 1951. Điều này rất quan trọng vì bước tiếp theo sẽ sử dụng các năm khác nhau.



Sản phẩm cuối cùng của nút này được minh họa như sau. Lưu ý rằng hiện có một cột-TOR_LENGTH-với năm 1951 cho đến khi cột đó cạn kiệt và sau đó là năm 1954. Đây sẽ là một điểm khác biệt quan trọng khi phân tích phân tích sẽ thêm nút kiểm tra t.

Tại thời điểm này, đã đến lúc thêm nút thực sự sẽ thực hiện kiểm tra thống kê-nút kiểm tra t. Trong trường hợp này, tên của nút là kiểm tra t của các nhóm độc lập mà nhà phân tích có thể tìm thấy bằng cách nhập tên đó vào hộp tìm kiếm. Màn hình cấu hình cho nút này (sau khi được kết nối) như sau. Lưu ý các cài đặt khác nhau trong hộp cấu hình, vì điều này là cần thiết để có được phản hồi chính xác nhất. Ngoài ra, hãy nhớ đặt nhóm 1951 vào ô đầu tiên và nhóm 1954 vào ô thứ hai. Tại thời điểm này, giả thuyết cũng giống như trong các phần trước, giả thuyết không bao gồm cả năm 1951 và 1954 có cùng độ dài cơn lốc xoáy trung bình, với giải pháp thay thế là năm 1951 có độ dài cơn lốc xoáy trung bình ít hơn năm 1954. Kết quả đối với nút này được xem bằng cách nhấp chuột phải vào nút và chọn tùy chọn đầu tiên từ dưới cùng của cửa sổ con đó như được hiển thị.

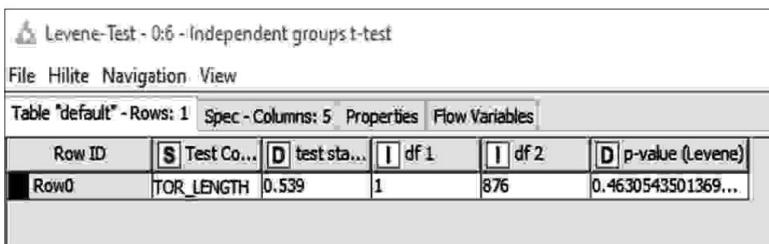




Sau khi nhà phân tích chọn “Kiểm tra t nhóm độc lập” từ menu, màn hình sau sẽ xuất hiện. Vui lòng so sánh các kết quả này với kết quả từ các phần khác. Sẽ có một số khác biệt nhỏ giữa cái này và ba cái khác. Tuy nhiên, các kết quả đều giống nhau và giả thuyết không bị bác bỏ, có nghĩa là không có ý nghĩa thống kê nào đối với điểm năm 1951 nhỏ hơn năm 1954 về độ dài lốc xoáy.

Independent groups t-test - 0.6 - Independent groups t-test						
File: Hilite Navigation View						
Table "default" - Rows: 2 Spec - Columns: 10 Properties Flow Variables						
Row ID	Test Co...	Variance Assumption	I	df	p-value (2-tailed)	
Row0	TOR_LENGTH	Equal variances assumed	-1.136	876	0.2562790899191453	
Row1	TOR_LENGTH	Equal variances not assumed	-1.156	534.858	0.2481252302253254	

Đây chỉ là một trong ba đầu ra từ nút này. Hai kết quả đầu ra khác bao gồm một "thử nghiệm F" (thử nghiệm Levene) có thể cung cấp cho nhà phân tích xác suất rằng hai mẫu có phương sai bằng nhau hoặc không bằng nhau. Màn hình đó được minh họa như sau. Kết quả là phép thử lớn hơn alpha ($0,05$ nhỏ hơn kết quả), điều đó có nghĩa là các phương sai không bằng nhau. Cho rằng tổng số cơn lốc xoáy năm 1954 gấp ba lần so với năm 1951, điều này có vẻ hợp lý. Tuy nhiên, thử nghiệm giúp xác nhận quan sát. Tất nhiên, hãy nhớ rằng không phải tất cả các công cụ đều được tạo ra như nhau. Nếu nhà phân tích nghi ngờ về tính xác thực của kết quả này, hãy tham khảo các công cụ khác và tiến hành các thử nghiệm tương tự để đảm bảo tính chính xác và nhất quán trong câu trả lời của bạn.



The screenshot shows a SPSS output window titled "Levene-Test - 0:6 -Independent groups t-test". The window includes a menu bar with "File", "Hilite", "Navigation", and "View". Below the menu is a toolbar with buttons for "Table", "Spec", "Columns", "Properties", and "Flow Variables". A table titled "Table \"default\" - Rows: 1" is displayed. The table has columns: Row ID, Test Co..., test sta..., df1, df2, and p-value (Levene). The first row, labeled "Row0", contains the value "TOR_LENGTH" in the "Test Co..." column, "0.539" in the "test sta..." column, "1" in the "df1" column, "876" in the "df2" column, and "0.4630543501369..." in the "p-value (Levene)" column.

Row ID	Test Co...	test sta...	df1	df2	p-value (Levene)
Row0	TOR_LENGTH	0.539	1	876	0.4630543501369...

CHƯƠNG 4

KIỂM TRA THỐNG KÊ THÊM

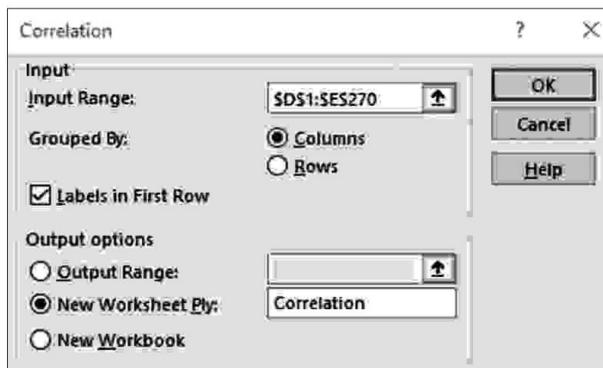
4.1 TƯƠNG QUAN

Theo ý kiến của tác giả này, mối tương quan có lẽ là một trong những khái niệm thống kê dễ nhận biết nhất. Bất cứ khi nào ai đó nghe thấy mối tương quan, họ có thể nghĩ rằng yếu tố này gây ra yếu tố kia, nhưng như nhiều nhà thống kê và nhà phân tích dữ liệu sẽ nói—sự tương quan không có nghĩa là quan hệ nhân quả. Tuy nhiên, tương quan vẫn là một khái niệm mạnh mẽ có thể dễ dàng thực hiện với các công cụ này theo một cách đơn giản. Trong cuốn sách này, mối tương quan sẽ không được hiển thị trên biểu đồ phân tán (có thể xuất hiện sau), nhưng nó sẽ được hiển thị bằng một ma trận hiển thị các biến và cách chúng được liên kết với nhau thông qua một số tương quan. Con số này nằm trong khoảng từ 0 đến 1, thể hiện mối quan hệ giữa hai biến này. Chẳng hạn, nếu có một mối tương quan là 0,90, thì đó được coi là một mối tương quan tích cực rất cao. Điều đó có nghĩa là khi một biến tăng lên thì biến kia cũng tăng theo. Nếu hệ số tương quan là -0,90, thì đây là hệ số tương quan âm rất cao, có nghĩa là khi một biến tăng thì biến kia giảm. Một ví dụ về mối tương quan tiêu cực sẽ là số năm sử dụng ô tô và giá của nó. Mối tương quan là thứ tồn tại trong tất cả các công cụ và sẽ được giải quyết từng công cụ một, giống như các phần khác.

4.1.1 Excel

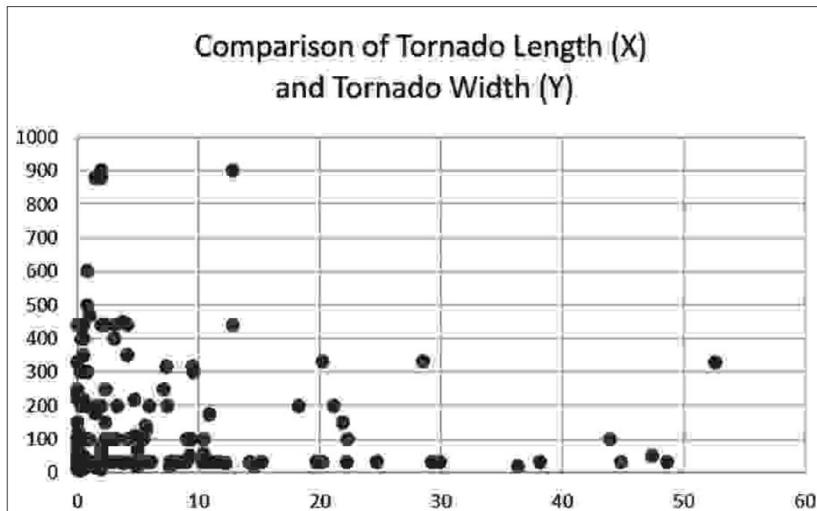
Tương quan rất dễ dàng trong Excel, đặc biệt là khi sử dụng ToolPak Phân tích. Trong phần này, người phân tích sẽ làm phép thử về mối tương quan giữa chiều dài cơn lốc xoáy và chiều rộng cơn lốc xoáy, hay cơn lốc xoáy chiếm bao nhiêu diện tích. Để làm điều này, trước tiên, nhà phân tích sẽ sử dụng cùng một tệp đã được sử dụng trong các phần trước, chủ yếu là khảo sát cơn lốc xoáy năm 1951. Sau khi tệp được nhập (hoặc mở

sau khi đã lưu), nhà phân tích sẽ mở Analysis ToolPak theo cách tương tự như trong các phần khác. Sau đó, nhà phân tích sẽ chọn tùy chọn Tương quan từ Công cụ phân tích như trong màn hình sau. Màn hình đã hoàn tất hiển thị hai cột phải được chọn và thực tế là chúng phải nằm cạnh nhau. Điều này có thể yêu cầu di chuyển một số cột cạnh nhau, nhưng với Excel, điều này cũng hơi đơn giản.



Kết quả tương quan cho sự kết hợp này như sau, nhưng nó có ý nghĩa gì? Ý nghĩa của kết quả này là có ít hơn một mối quan hệ .10 giữa chiều dài cơn lốc xoáy và chiều rộng cơn lốc xoáy (chỉ nhớ cho năm 1951), vì vậy bắt kể chiều dài của cơn lốc xoáy là bao nhiêu, việc dự đoán nó sẽ rộng bao nhiêu là gần như không thể.

Nếu nhà phân tích vẽ biểu đồ này trên biểu đồ phân tán, sử dụng lại Excel, thì biểu đồ sẽ trông giống như biểu đồ này, điều này cho thấy không có khả năng dự đoán giữa chiều dài cơn lốc xoáy và chiều rộng cơn lốc xoáy.



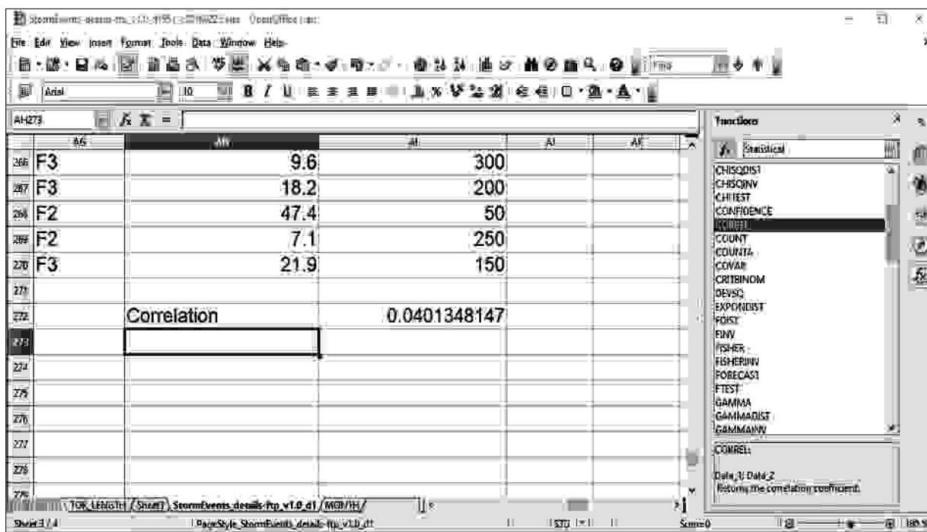
Miễn là nhà phân tích có các cột nằm cạnh nhau, một mối quan hệ tương quan có thể được tiến hành để thực hiện cái gọi là đa tương quan. Về cơ bản, nó được thực hiện giống như cách tương quan đã thực hiện trước đó, nhưng chỉ với nhiều mối tương quan hơn trong ma trận. Có một số thông tin về điều này được nêu trong Chương 7 của cuốn sách này.

4.1.2 Văn phòng mở

OpenOffice rất giống với Excel trong khu vực tương quan, nhưng thay vì sử dụng ToolPak Phân tích, các công thức thông thường là phương pháp thông thường cho OpenOffice. Trong trường hợp này, tệp sẽ giống nhau với hai biến giống nhau. Sự khác biệt lớn là bước đầu tiên sẽ là chọn một ô trống để đặt công thức và sau đó sử dụng công thức sau cho mối tương quan giữa chiều dài tor nado và chiều rộng lốc xoáy:

$$=\text{CORREL}(\text{AH2:AH270;AI2:AI270})$$

Khi nhà phân tích nhấn nút ENTER, màn hình sau xuất hiện. Điều này cho thấy mối tương quan giữa hai biến này.



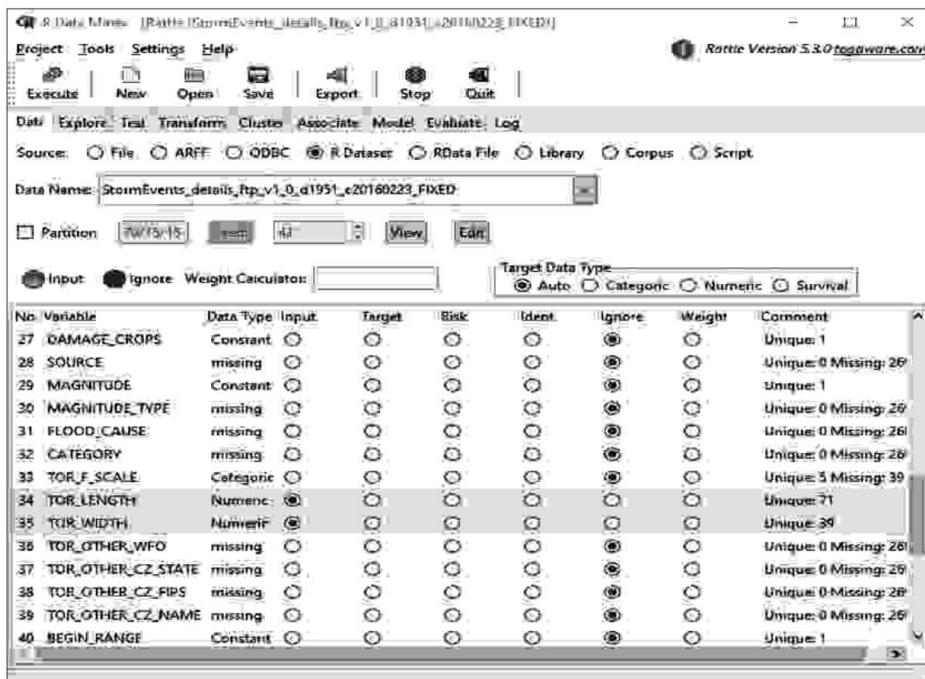
Như nhà phân tích có thể thấy, kết quả tương quan khớp với kết quả từ Excel. Nếu một bộ số tương quan là cần thiết, thì điều đó sẽ được đề cập trong phần tiếp theo về hồi quy, vì có chức năng trong OpenOffice để thực hiện hồi quy bộ số và do đó, nhiều tương quan.

4.1.3 R/RStudio/Rattle

Ứng dụng R rất linh hoạt vì nó áp dụng cho thử nghiệm thông thường và tương quan cũng không ngoại lệ. Quá trình thực hiện kiểm tra tương quan trực quan hơn so với nhiều công cụ khác. Trong phần này, Rattle sẽ được sử dụng để thực hiện chức năng tương quan, nhưng cuộc thảo luận cũng sẽ kéo theo việc sử dụng RStudio trong các chức năng lập trình đằng sau quá trình tương quan.

Trước tiên, hãy đảm bảo tệp theo dõi cơn lốc xoáy năm 1951 được nhập vào gói Rattle sẽ được kích hoạt sau khi mở ứng dụng RStudio.

Điều này được thể hiện trong phần trước khi nhập dữ liệu. Sau khi nhập, hãy đảm bảo rằng bạn nhấp vào Thực thi để dữ liệu được tải và nếu thông báo lỗi xuất hiện (trong trường hợp này sẽ xuất hiện), hãy gán từng biến cho “đầu vào” hoặc “bỏ qua”. Trong trường hợp này, hãy gán tất cả các biến theo TOR_LENGTH và TOR_RỘNG cho các nút radio “bỏ qua” để hạn chế mối tương quan. Nhà phân tích có thể chọn thực hiện tương quan trên tất cả các biến, được gọi là nhiều quan hệ tương quan, nhưng điều này có thể gây cồng kềnh và tốn bộ nhớ. Màn hình hoàn thành sẽ giống như hình ảnh sau:



Sau khi nhập vào biểu tượng "Thực thi" trên thanh công cụ, hãy chuyển đến tab "Khám phá" để sử dụng chức năng tương quan. Tại thời điểm này, nhà phân tích sẽ làm theo các lựa chọn trong màn hình này để nhận kết quả được hiển thị. Nếu kết quả xuất hiện khác với màn hình này, vui lòng đảm bảo rằng phương pháp "Pearson" được chọn trong hộp thả xuống. Nếu các phương pháp khác được chọn, các kết quả khác nhau sẽ xuất hiện và hơi khác nhau đáng kể, vì vậy hãy thận trọng và kiểm tra công việc.

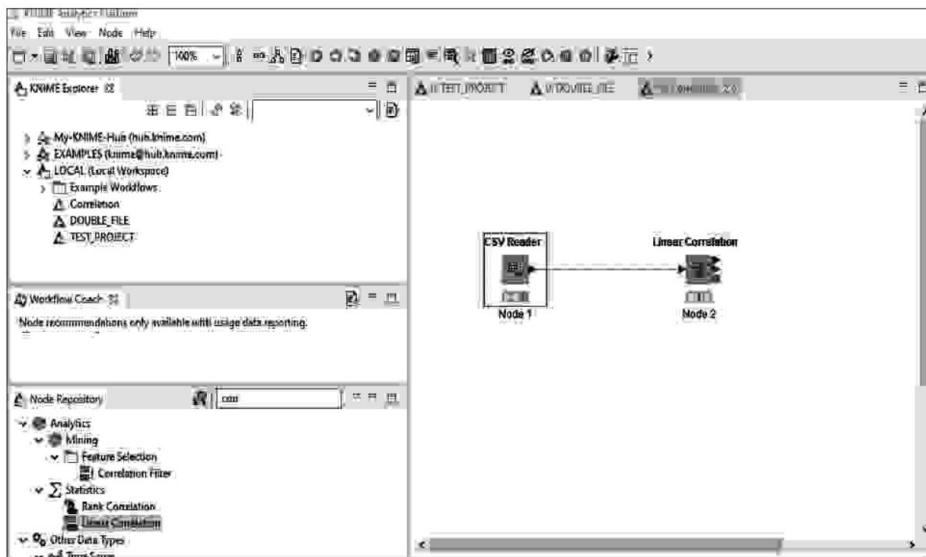


Như được hiển thị, số tương quan (.04013) giống như số đã được trình bày trong các phần khác. Sự dễ dàng mà điều này đã được thực hiện là khá ấn tượng. Đối với các chức năng tệp đơn lẻ, Rattle là một tùy chọn tốt để phân tích dữ liệu bằng cách sử dụng tương quan.

4.1.4 KIẾN THỨC

Tính linh hoạt của KNIME dựa trên chức năng của nút và công cụ KNIME có khả năng thực hiện mỗi tương quan với một nút. Để thực hiện việc này, hãy mở KNIME cho dự án đang thực hiện hoặc thậm chí tạo một dự án mới và bắt đầu nhập theo dõi cơn lốc xoáy năm 1951 như đã thực hiện trong phần nhập.

Khi quá trình này hoàn tất, hãy thêm nút tương quan vào nút Trình đọc CSV như được mô tả trong màn hình sau.



Cấu hình của nút Tương quan tuyến tính bao gồm việc xác định hai biến sẽ được kiểm tra và trong trường hợp này, nó sẽ là TOR_LENGTH và TOR_WIDTH. Chúng nên được đặt trong màn hình cấu hình như được hiển thị theo cách sau. Vui lòng đảm bảo rằng các cột hoặc biến mà nhà phân tích xác định là chính xác, bởi vì công cụ này, giống như bất kỳ công cụ nào khác, sẽ cung cấp cho nhà phân tích kết quả đã được nhập vào, vì nó không thể dự đoán những gì nhà phân tích muốn, chỉ những gì họ đã chọn. Nhà phân tích sẽ muốn nhấp chuột phải vào nút Tương quan và chọn tùy chọn ở cuối menu phụ có tên là Đo lường tương quan. Điều này sẽ tiết lộ màn hình sau

và cung cấp cho nhà phân tích một dấu hiệu về mối tương quan giữa hai biến số này. Hãy nhớ rằng một phần của phân tích dữ liệu không phải lúc nào cũng chọn các biến tương quan tốt nhất. Có một thử nghiệm liên tục và cải cách các giả thuyết để phân tích chính xác. Vui lòng xem xét điều đó khi màn hình tiếp theo này được trình bày.

Correlation measure - 2:2 - Linear Correlation			
File Hilitc Navigation View			
Table "default" - Rows: 1 Spec - Columns: 5 Properties Flow Variables			
Row ID	S First col...	S Second...	D Correlation value
Row0	TOR_LENGTH	TOR_WIDTH	0.04013481465483142

Nhà phân tích sẽ ngay lập tức nhận thấy rằng "Giá trị tương quan" giống với giá trị trong các phần trước. Lý do chính của việc sử dụng các giá trị và biến giống nhau là để chứng minh cho nhà phân tích rằng công cụ được sử dụng cho chức năng sẽ không đưa ra các kết quả khác nhau. Trong trường hợp này, tất cả các công cụ đều cho kết quả chính xác như nhau. Nguyên nhân chính của sự không nhất quán có thể là vấn đề làm tròn hoặc sử dụng một công thức khác, nhưng trong trường hợp này, tất cả các công cụ đều sử dụng phương pháp tương quan Pearson và công thức cho phương pháp đó rất nhất quán trong nhiều văn bản thống kê. Một số văn bản này được đưa vào làm tài liệu tham khảo cho cuốn sách này.

4.2 HỒI QUY

Một khái niệm thống kê mà các nhà phân tích dữ liệu thường như hiểu, ít nhất là những khái niệm mà tác giả này đã dạy, là hồi quy. Trên thực tế, tác giả này đã thấy hồi quy được áp dụng cho các bộ dữ liệu không cần loại phân tích này.

Tuy nhiên, hồi quy tuyến tính có phần quan trọng và cần được giải quyết đối với các công cụ này. Việc xem xét nhanh khái niệm này là cần thiết để tạo tiền đề cho các cuộc trình diễn tiếp theo.

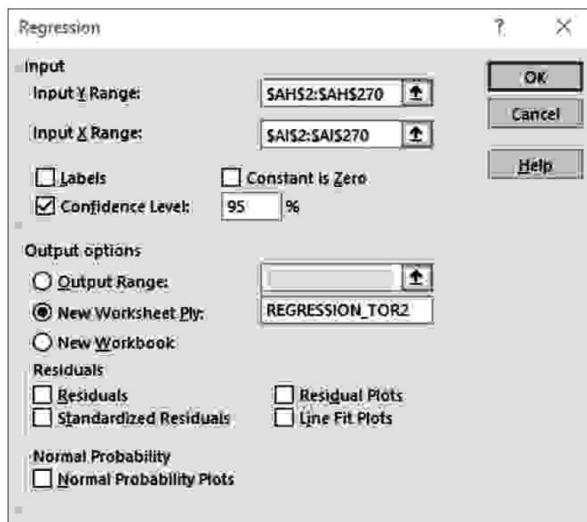
Hồi quy tuyến tính đang sử dụng một phương trình tuyến tính (xin lỗi về điều đó, cần phải hồi tưởng lại cơ bản của môn toán thời trung học) để vẽ ra một dự đoán có thể có về các giá trị trong tương lai dựa trên các kết quả trong quá khứ. Về bản chất, các điểm tọa độ XY được vẽ bằng cách sử dụng hai biến trong tập dữ liệu mà nhà phân tích đã chọn và một phương trình được lập từ đồ thị của các điểm đó. Phương trình được xây dựng là một phương trình tuyến tính (thông thường) tạo thành một đường cố gắng

chia các điểm dữ liệu trong đó một nửa số điểm nằm ở một bên của đường và một nửa ở bên kia (xấp xỉ). Trong phần này, nhà phân tích sẽ áp dụng các công cụ để xây dựng phương trình dự đoán này. Giải thích thêm về phương trình sẽ được đưa ra sau công cụ đầu tiên, trong trường hợp này là Excel.

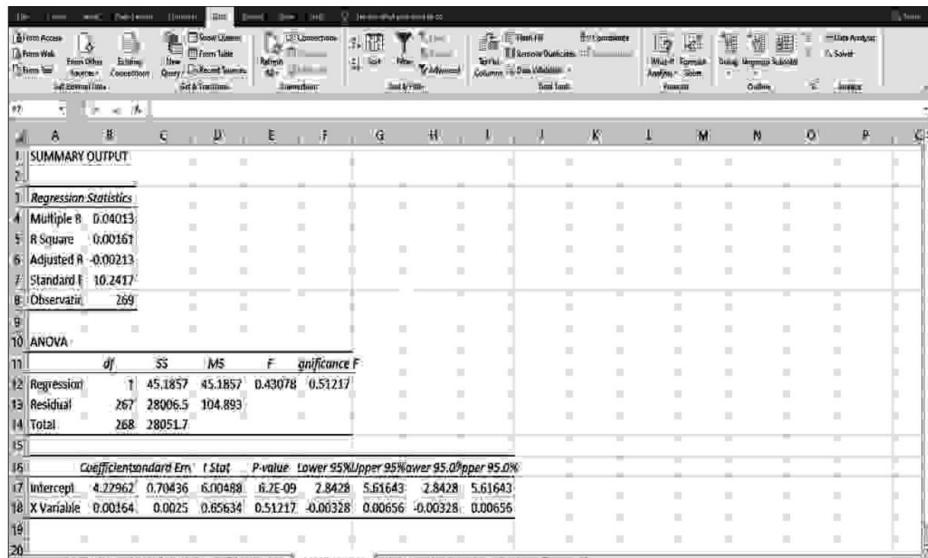
4.2.1 Excel

Excel, thông qua ToolPak Phân tích, có chức năng Hồi quy thực hiện công việc trình bày kết quả Hồi quy một cách nhanh chóng có thể được sử dụng trong các bài thuyết trình. Bước đầu tiên là nhập dữ liệu theo dõi cơn lốc xoáy năm 1951 để nhà phân tích có thể thực hiện hồi quy đối với hai biến trở lên. Trong trường hợp này, nhà phân tích sẽ sử dụng hồi quy theo hai biến số, chiều dài cơn lốc xoáy (TOR_LENGTH) và chiều rộng cơn lốc xoáy (TOR_WIDTH), cho mục đích trình diễn.

Sau khi kích hoạt Analysis ToolPak và chọn "Regression" từ menu, nhà phân tích sẽ chọn hai biến để kiểm tra. Biến đầu tiên sẽ là biến trực y, được gọi là biến "phản hồi" hoặc biến "phụ thuộc" và biến thứ hai sẽ là biến trực x, là biến có thể "dự đoán". Về bản chất, khi hoàn thành nhiệm vụ này, những gì nhà phân tích sẽ đặt vào "x" sẽ dẫn đến "y". Trong trường hợp này, nhà phân tích có thể đặt chiều rộng vào dấu "x" và kết quả sẽ là chiều dài cơn lốc xoáy. Một lần nữa, đây là mục đích trình diễn và không nên được hiểu là phân tích dự đoán thực sự để dự đoán độ dài của cơn lốc xoáy. Hãy nhớ rằng đây chỉ là dữ liệu của một năm.



Kết quả phân tích hồi quy như sau. Vui lòng lưu ý tất cả các số khác nhau trên màn hình này. Những thứ quan trọng đối với nhà phân tích ngay lập tức sẽ là những thứ bao gồm phương trình hồi quy.



The screenshot shows a Microsoft Access query results grid with the following data:

SUMMARY OUTPUT									
Regression Statistics									
1	Multiple R	0.04013							
2	R Square	0.00161							
3	Adjusted R	-0.00213							
4	Standard E	10.2417							
5	Observations	269							
ANOVA									
6	df	SS	MS	F	Significance F				
7	Regression	1	45.1857	45.1857	0.43078	0.51217			
8	Residual	267	28006.5	104.893					
9	Total	268	28051.7						
Coefficients Standard Error t Stat P-value Lower 95% Upper 95% Lower 95% Upper 95%									
10	Intercept	4.22962	0.70436	6.00488	6.7E-09	2.8428	5.61643	2.8428	5.61643
11	X Variable	0.00164	0.0025	0.85634	0.51217	-0.00328	0.00656	-0.00328	0.00656

Các số mà nhà phân tích quan tâm bao gồm các số nằm trong cột "Hệ số" cho cả "Giao đoạn chặn" và "Biến số X" (trong trường hợp này là TOR_LENGTH). Phương trình hồi quy thu được sẽ là $y=0,0164x + 4,22962$. Điều này có nghĩa là nếu nhà phân tích muốn biết chiều rộng cơn lốc xoáy mà họ mong đợi, thì nhà phân tích đặt chiều dài cơn lốc xoáy cho biến "x", nhân nó với 0,0164 và cộng 4,33962 để tìm chiều rộng cơn lốc xoáy gần đúng. Đây là nơi hồi quy có thể bị lạm dụng. Đầu tiên, đây chỉ là một năm, có tính đến 12 tháng, không phải tất cả các tháng đều có lốc xoáy hoặc lốc xoáy có độ dài bất kỳ. Thứ hai, mối tương quan, như đã đề cập trong phần trước, rất mong manh-là 0,040-có nghĩa là có ít hơn 0,1 mối tương quan, một mối tương quan cực kỳ thấp giữa độ dài ngày và cơn lốc xoáy. Nếu nhà phân tích muốn thực hiện hồi quy bội, điều đó có thể được thực hiện thông qua công cụ này, nhưng các cột chứa dữ liệu phải nằm cạnh nhau, do đó phần minh họa sẽ được thảo luận sau. Mục đích chính của các biến này là để đảm bảo rằng các công cụ cho kết quả nhất quán. Điểm chính ở đây là hai biến này không phải là sự kết hợp tốt cho mục đích hồi quy do có mối tương quan và thiếu dữ liệu theo chiều dọc.

4.2.2 Văn phòng mở

Hàm hồi quy OpenOffice có thể so sánh với công thức "Pre-Analysis ToolPak" cho Excel. Hàm công thức được gọi là "linest" và được gọi là "công thức mảng". Điều này có nghĩa là kết quả của công thức được thực hiện trên một số ô. Để biến một công thức thành công thức mảng, trước khi nhấn Enter, hãy kết hợp các phím CTRL-SHIFT-ENTER để chuyển đổi công thức thành công thức mảng. Công thức sẽ được đặt trong dấu ngoặc nhọn ({}) thay vì dấu ngoặc đơn.

Bước đầu tiên để sử dụng công thức hồi quy trong OpenOffice là mở tệp đã được sử dụng trong các phần trước và đảm bảo rằng các biến được chọn là TOR_LENGTH và TOR_WIDTH cho minh họa này.

Bước tiếp theo là đặt công thức hồi quy vào một ô trống (giống như một nhà phân tích thực hiện trong Excel). Hãy nhớ rằng dấu chấm phẩy phân tách các tham số của công thức-chữ không phải dấu phẩy. Công thức hồi quy sẽ trông tương tự như thế này (đối với các cột/biến cụ thể đã đề cập trước đó).

$$=LINEST(AH2:AH270;AI2:AI270;1;1)$$

Thoạt nhìn, công thức này trông rất giống công thức LINEST trong Excel, ngoại trừ dấu chấm phẩy. Sự khác biệt giữa công thức này và các công thức khác trong OpenOffice là ghi nhớ CTRL-SHIFT-ENTER để xem toàn bộ kết quả của phân tích hồi quy. Màn hình sau đây cho biết điều gì sẽ xảy ra khi bạn hoàn thành công thức theo cách này.

	AG	AH	AI	AJ	AK
270	F3	21.9	150		
271					
272		TOR_LENGTH	TOR_WIDTH		
273		Regression	0.001639459	4.229615	
274			0.002497892	0.704363	
275			0.001610803	10.24174	
276			0.430778393	267	
277			45.18570962	28006.48	
278					
279					
280					

Tất cả những con số này có ý nghĩa gì? Làm thế nào để chúng liên quan đến kết quả trong Excel? Bảng sau đây sẽ làm sáng tỏ một chút về các ô trước đó trong OpenOffice. Trang web để nhận bản dịch đầy đủ nằm ở đây: https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_LINEST_function.

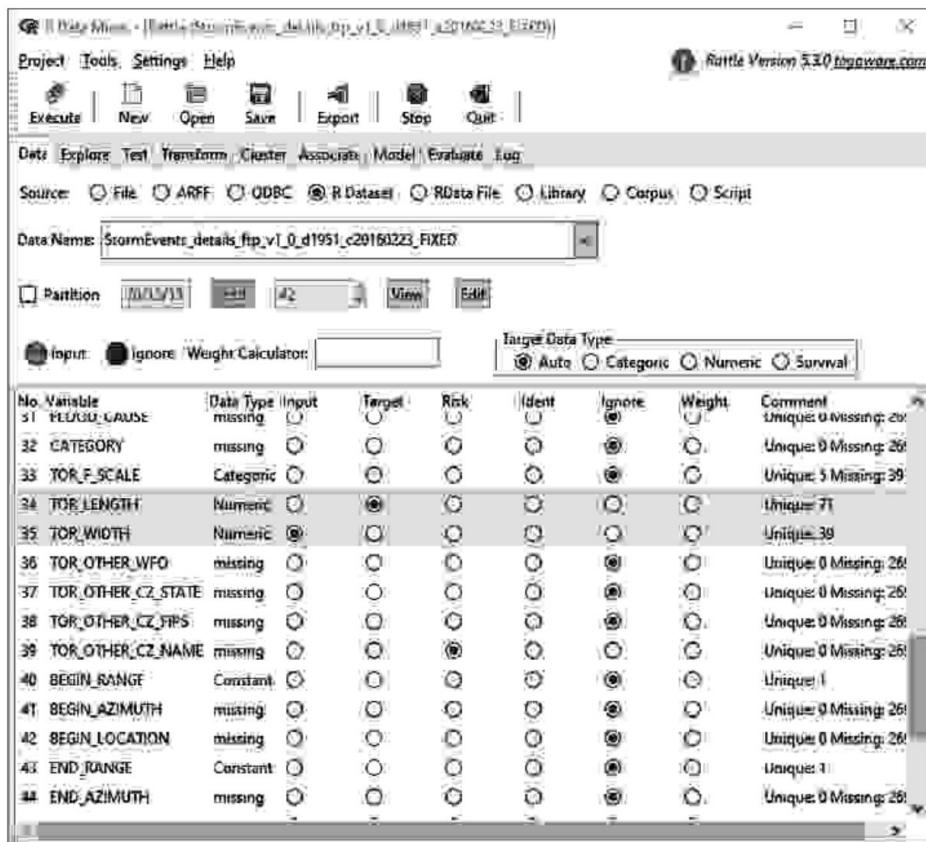
Đối với cuốn sách này, chỉ một vài ô trong bảng được bao gồm.

giá trị "x"	"y" đánh chặn
R2 (giải thích trong phần này)	
"F" (Bài kiểm tra Levene được giải thích trong cuốn sách này)	Mức độ Tự do (ngoài phạm vi của cuốn sách này)

Bảng này cho chúng ta biết điều gì? Ô đầu tiên bên trái giống với "Biến X" từ kết quả Excel (AI 273 trong trường hợp này). Ô thứ hai ("Y" biến) giống như "đoạn chặn" (AJ 273). "r2" là bình phương của giá trị tương quan "r" và sẽ phản ánh giá trị được đánh dấu "R Square" trong bảng Excel. Những ô trong bảng trước là quan trọng nhất tại thời điểm này và có giá trị nhất đối với nhà phân tích dữ liệu sử dụng công cụ. Bằng cách sử dụng các ô của bảng được giải thích, giờ đây có thể lập công thức cho một phương trình tuyến tính và có thể tính toán được mối tương quan. Không "đẹp" như Excel nhưng cũng cho kết quả tương tự.

4.2.3 R/RStudio/Rattle

Gói Rattle trong R cung cấp một phương pháp phân tích hồi quy hoàn toàn phù hợp. Các bước để thiết lập dữ liệu được phân tích theo cách này giống như trước đây-nhập dữ liệu giống như đã được sử dụng trong các phần trước và sử dụng TOR_LENGTH và TOR_WIDTH cho phân tích hồi quy. Màn hình Rattle cho dữ liệu sẽ xuất hiện như sau:



Vui lòng lưu ý rằng tất cả các biến khác đã được đặt trong "Bỏ qua" vì nhà phân tích sẽ không cần đến chúng vào thời điểm này. TOR_LENGTH đã được đặt trong cột "Mục tiêu", trong khi TOR_WIDTH sẽ được đặt trong cột "Đầu vào". Điều này giống như đặt TOR_LENGTH trong "trục y" và TOR_WIDTH trong "trục X". Đôi với việc sử dụng hồi quy bội trong tương lai, các biến khác có thể được nhập lại vào cột "Đầu vào" để bao gồm chúng. Nếu có bất kỳ thay đổi nào đối với tập dữ liệu, hãy đảm bảo rằng nhà phân tích nhập vào biểu tượng "Thực thi". Nếu không, bảng sẽ giống như trước khi thay đổi. Thực hiện lưu các thay đổi.

Sau khi dữ liệu được đặt và biểu tượng Thực thi được nhập, hãy chuyển đến tab "Mô hình" và chọn loại "Tuyến tính" và "Số" bên dưới lựa chọn đó. Sau đó, nhập vào biểu tượng "Execute" và màn hình sau sẽ xuất hiện. Rõ ràng là kết quả ở đây của Rattle khớp với kết quả từ các công cụ khác.

```

Rattle Version 3.1.0 (http://www.commbio.org)
Project Tools Settings Help
Execute New Open Save Export Stop Close
Data Explore Test Structure Cluster Associate Model Evaluate Log
Type: Tree (Forest) SVM (Linear) Neural Net Survival All
Response: Numeric Generalized Poisson Logistic Wald Multinomial Model Builder Im
File
Summary of the Linear Regression model (built using lm):
Call:
lm(formula = TOR_LENGTH ~ ., data = crssdataset, offset = crssinput,
    offset = crsstarget)

Residuals:
    Min      1Q  Median      3Q     Max
-6.115 -4.284 -4.084 -0.884  87.824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.229615  0.704383  6.005 0.00000000623 ***
TOR_WIDTH   0.001639  0.002498  0.656  0.512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
Residual standard error: 10.24 on 267 degrees of freedom
Multiple R-squared:  0.001611, Adjusted R-squared:  -0.002128
F-statistic: 0.4388 on 1 and 267 DF, p-value: 0.5122

-- ANOVA --
Analysis of Variance Table

Response: TOR_LENGTH
          Df  Sum Sq Mean Sq F Value Pr(>F)
TOR_WIDTH  1  45.2  45.206  0.4308 0.5122
Residuals 267 28006.5 104.853
[1] "\n"
Time taken: 0.01 secs
Rattle timestamp: 2020-01-09 10:36:29 China

```

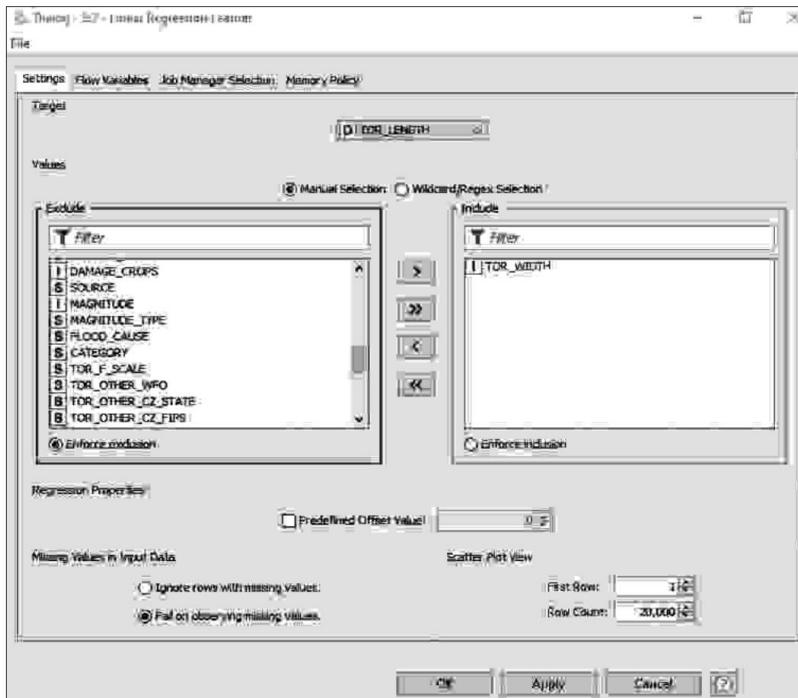
Linear model evaluation has been plotted.

Hiện tại, có vẻ như mọi công cụ đã đồng ý với những công cụ khác liên quan đến khái niệm này, vì vậy đây sẽ là một cách tuyệt vời để xác minh kết quả từ hồi quy. Với chức năng sẵn có và tương đối đơn giản để sử dụng, không có lý do gì mà việc xác minh kết quả lại không được thực hiện trong tình huống này.

4.2.4 KIẾN THỨC

KNIME là công cụ cuối cùng sẽ được đề cập trong khái niệm hồi quy này. Như trong các trường hợp khác, ứng dụng KNIME có sẵn một nút để phân tích hồi quy. Sau khi dữ liệu được nhập thông qua Trình đọc CSV, nhà phân tích có thể kết nối dữ liệu với nút hồi quy, được gọi là Tuyến tính

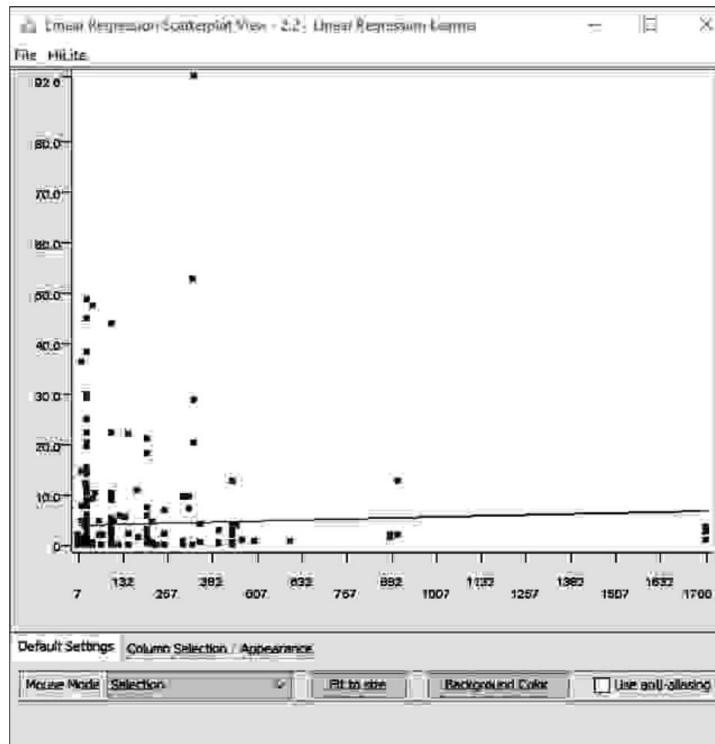
Nút Người học hồi quy, được tìm thấy bằng cách nhập từ "hồi quy" vào hộp tìm kiếm. Sau khi nút được kéo, đặt và kết nối, nhấp đúp vào nút đó để mở nút, hiển thị màn hình cấu hình sau.



Định cấu hình màn hình chính xác như trong hình trước, sử dụng TOR_LENGTH làm mục tiêu (giống như Rattle) và TOR_WIDTH làm cột để bao gồm đối với mục tiêu này. Nhấn OK và thực hiện nút (nhớ mũi tên màu xanh lá cây). Sau khi nút được thực thi, hãy nhấp chuột phải vào nút đó và chọn "Ché độ xem: Ché độ xem kết quả hồi quy tuyến tính". Thao tác này sẽ tạo ra cửa sổ này, hiển thị các kết quả tương tự như trong các phần trước.

Variable	Coeff.	Std. Err.	t-value	P> t
TOR_WIDTH	0.0016	0.0025	0.6563	0.5122
Intercept	4.2296	0.7044	6.0049	6.23E-9
Multiple R-Squared: 0.0016				
Adjusted R-Squared: -0.0021				

Nếu nhà phân tích mong muốn, họ cũng có thể chọn lựa chọn menu bên dưới lựa chọn tạo ra cửa sổ này để hiển thị kết quả biểu đồ phân tán, kết quả này sẽ khớp với biểu đồ phân tán trong Excel.



Điều rất quan trọng là nhà phân tích hiểu rằng hồi quy là một mô hình và phải được xác minh và kiểm tra thông qua các kỹ thuật đánh giá. Điều này nằm ngoài phạm vi của cuốn sách này, nhưng các công cụ được trình bày ở đây có rất nhiều chức năng kiểm tra cho mục đích này. Như đã nêu trước đây, việc khám phá từng hoặc tắt cả các công cụ này sẽ nâng cao kiến thức về cả phân tích và thống kê dữ liệu.

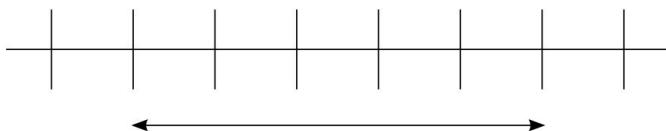
4.3 KHOẢNG TIN CẬY

Khoảng tin cậy đã và đang quay trở lại trong lĩnh vực thống kê.

Trong cuốn sách của mình, Thống kê đã sai, Alex Reinhart giải thích rằng khoảng tin cậy là một phương pháp thống kê đơn giản không được sử dụng thường xuyên khi cần thiết, dẫn đến một số kết quả thú vị, nếu không muốn nói là không chính xác.

(Reinhart, 2015). Vì khoảng tin cậy là một phương pháp tương đối đơn giản để đánh giá tác động của một giá trị đối với một giá trị khác, nên cần có một lời giải thích ngắn gọn trước khi bắt đầu sử dụng công cụ.

Khi một nhà phân tích dữ liệu giải quyết một khoảng tin cậy, họ đang sử dụng quan điểm một chiều của mức độ tin cậy. Mặc dù điều này nghe có vẻ khó hiểu, nhưng một minh họa đơn giản có thể giúp làm rõ khái niệm này. Nếu một nhà phân tích loại bỏ đường cong hình chuông khỏi đường cong thông thường tiêu chuẩn, thì kết quả sẽ như sau:



Đường thẳng đứng ở giữa biểu thị giá trị trung bình và các đường bên phải và bên trái biểu thị các độ lệch chuẩn khác nhau cộng hoặc trừ giá trị trung bình đó. Các mũi tên biểu thị khoảng tin cậy 95% dựa trên mức độ tin cậy 95%. Về bản chất, điều này có nghĩa là, khi đưa ra một mẫu dân số, khoảng tin cậy sẽ cho nhà phân tích biết xác suất mà giá trị trung bình nằm giữa các khoảng đó là bao nhiêu. Nói cách khác, ở khoảng tin cậy 95%, có 95% khả năng giá trị trung bình nằm ở đâu đó giữa Mức tin cậy trên (UCL) và Mức tin cậy dưới (LCL).

Làm thế nào điều này giúp với phân tích dữ liệu? Câu trả lời một lần nữa bắt nguồn từ Reinhart, người đã đề cập đến khoảng tin cậy với giọng điệu gần như tôn kính trong cuốn sách của mình. Xuyên suốt cuốn sách, ông đưa ra rất nhiều ví dụ, hầu hết từ các nghiên cứu thực tế, giúp bảo vệ rõ ràng việc sử dụng khoảng tin cậy.

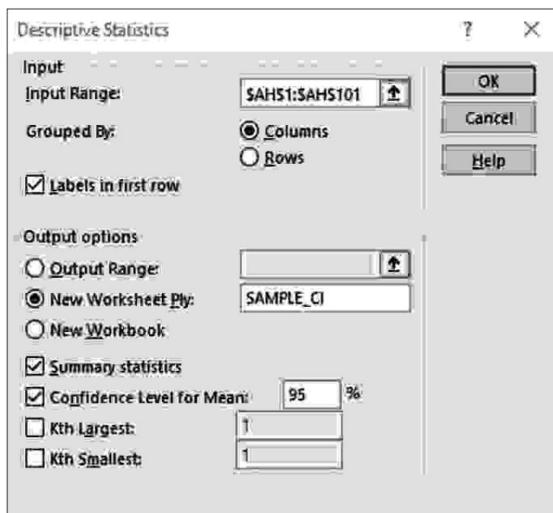
Điểm mấu chốt trong khoảng tin cậy là nó giúp cung cấp xác minh kết quả nghiên cứu. Ví dụ: nếu một mẫu cho thấy thời gian làm việc trung bình của một người là 8 năm và sử dụng khoảng tin cậy 95%, phạm vi nằm trong khoảng từ 6 đến 10, nhà phân tích có thể tuyên bố rằng với độ chắc chắn 95%, rằng trung bình của một người việc làm trong dân số là khoảng từ 6 đến 10 năm. (Reinhart, 2015). Toàn bộ ý tưởng về khoảng tin cậy là cung cấp cho nhà phân tích một bài đọc về nghiên cứu được tiến hành tốt như thế nào. Reinhart tuyên bố rằng nếu khoảng quá rộng, như trong ví dụ của chúng tôi, kết quả là 1-20, điều đó có nghĩa là không có đủ mẫu được thực hiện và cần phải thu thập thêm mẫu để thu hẹp khoảng (Reinhart, 2015).

Lý do học khoảng tin cậy là để kết hợp nó vào bất kỳ nghiên cứu hoặc dự án phân tích nào mà nhà phân tích phải thực hiện và để chứng minh rằng phương pháp thực tế là đơn giản, hiệu quả và có sẵn trong nhiều công cụ khác nhau.

4.3.1 Excel

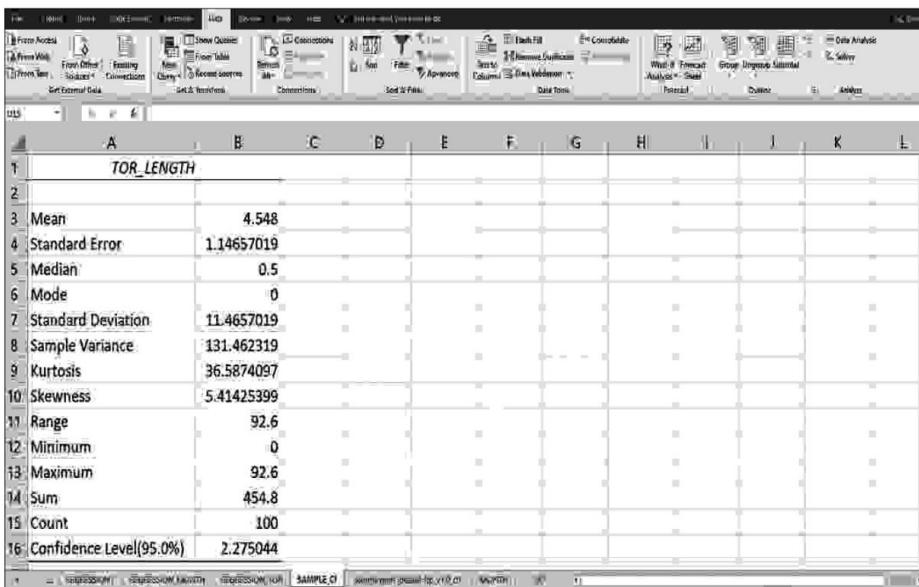
Excel, thông qua ToolPak Phân tích, có khả năng trình bày khoảng tin cậy như một phần của phần thống kê mô tả tổng thể của ToolPak. Quy trình này bắt đầu bằng việc nhập dữ liệu, sử dụng lại dữ liệu theo dõi cơn lốc xoáy năm 1951 và mở ToolPak Phân tích. Chọn Thống kê mô tả từ menu ToolPak và sử dụng TOR_LENGTH làm cột sẽ được sử dụng, nhưng lần này chỉ chọn 100 hàng đầu tiên. Đây sẽ là mẫu sẽ được sử dụng để so sánh khoảng tin cậy với trung bình dân số thực tế ở cuối. Xin lưu ý rằng chỉ chọn 100 hàng đầu tiên không phải là một mẫu ngẫu nhiên thực sự, nhưng điều đó sẽ được thảo luận trong phần sau. Đối với cuộc biểu tình này, điều này sẽ đủ.

Khi ToolPak Phân tích được mở và cột và hàng được chọn sen, màn hình sẽ xuất hiện như sau:



Một lời cảnh báo liên quan đến màn hình này. Vui lòng lưu ý rằng "Nhãn ở hàng đầu tiên" được chọn. Nếu không có nhãn trong lựa chọn, Excel sẽ tự động lấy hàng đầu tiên và sử dụng nó làm nhãn, không báo cáo trong phân tích dữ liệu. Điều này sẽ dẫn đến các vấn đề với bất kỳ phân tích nào được thực hiện trên dữ liệu. Chỉ là một lời cảnh báo cho những người sử dụng cài đặt hoặc cấu hình mặc định. Vui lòng kiểm tra các cài đặt này trước khi nhấp vào OK. Ngoài ra, vui lòng lưu ý rằng Thống kê tổng hợp và Mức độ tin cậy cho Phương tiện đều được chọn. Nếu ana lyst không kiểm tra khỏi Thống kê Tóm tắt, đó sẽ là một vấn đề, nhưng

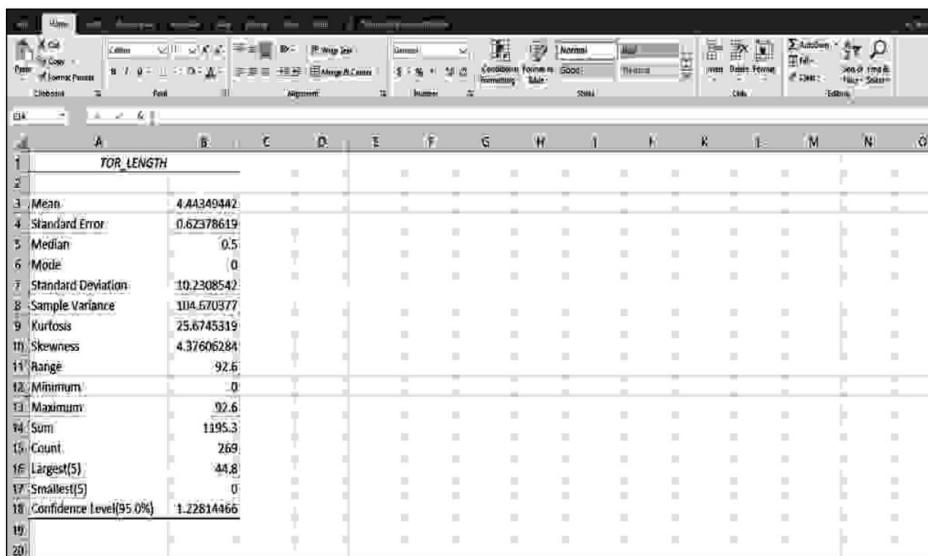
nếu họ không kiểm tra khói Mức độ Tin cậy cho Phương tiện, kết quả mong muốn trong tình huống này sẽ không xuất hiện. Như đã nêu trước đây, vui lòng đảm bảo rằng cấu hình theo cách mong muốn trước khi thực hiện. Kết quả của mức độ tin cậy là trong màn hình sau đây. Nhưng nó có nghĩa gì?



The screenshot shows a Microsoft Excel spreadsheet with the Data Analysis ribbon tab selected. The data is organized into two columns: 'Metric' and 'Value'. The metrics listed are Mean, Standard Error, Median, Mode, Standard Deviation, Sample Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, Count, and Confidence Level(95.0%). The values are: Mean (4.548), Standard Error (1.14657019), Median (0.5), Mode (0), Standard Deviation (11.4657019), Sample Variance (131.462319), Kurtosis (36.5874097), Skewness (5.41425399), Range (92.6), Minimum (0), Maximum (92.6), Sum (454.8), Count (100), and Confidence Level(95.0%) (2.275044).

	TOR_LENGTH
Mean	4.548
Standard Error	1.14657019
Median	0.5
Mode	0
Standard Deviation	11.4657019
Sample Variance	131.462319
Kurtosis	36.5874097
Skewness	5.41425399
Range	92.6
Minimum	0
Maximum	92.6
Sum	454.8
Count	100
Confidence Level(95.0%)	2.275044

Hàng Mức độ tin cậy (95,0%) trong kết quả cho biết mức độ tin cậy là 2,275044. Điều này có nghĩa là, với độ tin cậy 95%, trung bình tổng thể nằm trong khoảng $\pm 2,275044$, có nghĩa là, có tính đến việc trung bình mẫu là 4,548 theo kết quả tóm tắt, trung bình tổng thể nằm trong khoảng từ 2,272956 đến 6,823044. Để tiếp tục hiển thị kết quả này, hãy lấy toàn bộ cột TOR_LENGTH làm giá trị trung bình cho thống kê tóm tắt, sử dụng ToolPak với kết quả sau.



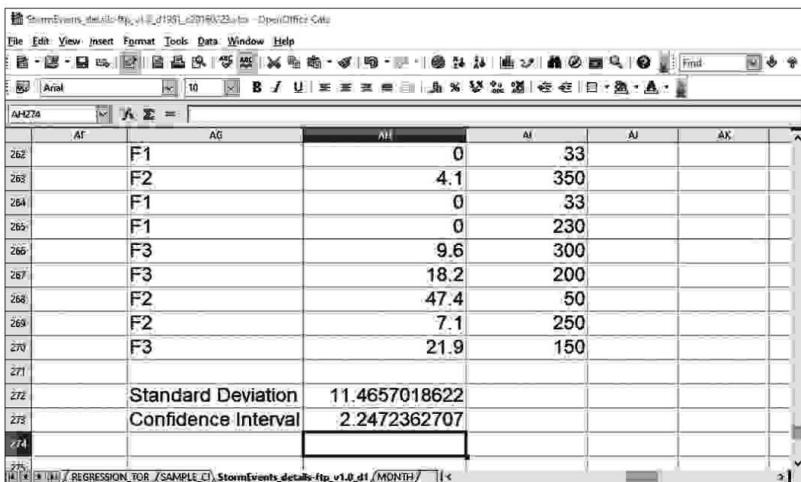
The screenshot shows a Microsoft Excel spreadsheet with the title 'TOR_LENGTH' in cell A1. The data consists of 18 rows of statistical measures:

	TOR_LENGTH
3 Mean	4.44349442
4 Standard Error	0.62378619
5 Median	0.5
6 Mode	0
7 Standard Deviation	10.2308542
8 Sample Variance	114.670377
9 Kurtosis	25.6745319
10 Skewness	4.37606284
11 Range	92.6
12 Minimum	0
13 Maximum	92.6
14 Sum	1195.3
15 Count	269
16 Largest(5)	44.8
17 Smallest(5)	0
18 Confidence Level(95.0%)	1.22814466
19	
20	

Giá trị trung bình của dân số là 4,44349442, nằm trong khoảng từ 2,27 đến 6,82. Nếu mức độ tin cậy thấp hơn, phạm vi cũng sẽ thấp hơn, vì vậy mức độ tin cậy 80% sẽ tạo ra phạm vi hẹp hơn.

4.3.2 Văn phòng mở

Chức năng OpenOffice của khoảng tin cậy không có cùng sự tiện lợi của ToolPak phân tích của Excel, nhưng nó thực hiện công việc. Bước đầu tiên cũng giống như các phần khác; nhập dữ liệu cơ bản xoay năm 1951 và chọn một ô trống trong trang tính đó để hiển thị khoảng tin cậy như được hiển thị.



The screenshot shows an OpenOffice Calc spreadsheet with the title 'StormEvents_details_1951_2010V23.xls - OpenOffice Calc'. The data consists of 27 rows of statistical measures:

	A1	A6	A11	A16	A21	A26
262	F1		0	33		
263	F2		4.1	350		
264	F1		0	33		
265	F1		0	230		
266	F3		9.6	300		
267	F3		18.2	200		
268	F2		47.4	50		
269	F2		7.1	250		
270	F3		21.9	150		
271						
272	Standard Deviation		11.4657018622			
273	Confidence Interval		2.2472362707			
274						
275						

Trong ví dụ này, kết quả Khoảng tin cậy được đặt ở cuối cột TOR_LENGTH. Bằng cách này, việc chọn cột sẽ dễ dàng hơn nhiều. Hãy nhớ rằng chỉ 100 hàng đầu tiên sẽ được chọn, nhưng trong trường hợp này, hãy bắt đầu từ AH2, không phải AH1. OpenOffice không có cùng mức độ quan tâm đối với các tiêu đề, vì vậy hãy đảm bảo rằng chúng không được bao gồm trong quá trình lấy dữ liệu. Công thức for cho khoảng tin cậy phải bao gồm độ lệch chuẩn, là STDEVA hoặc độ lệch chuẩn của một mẫu. Công thức này nên được đặt trong AH272.

=STDEVA(AH2:AH101)

Bây giờ điều này làm là cung cấp khả năng của khoảng tin cậy để sử dụng độ lệch chuẩn vào công thức khoảng tin cậy, công thức này sẽ được đặt trong AH273 và sẽ giống như sau:

=BẢO MẬT(0,05;AH272;100)

Một lời giải thích là cần thiết trên công thức trước đó. "0,05" ở đầu được gọi là "giá trị alpha". Khi số liệu thống kê đề cập đến "alpha", điều đó có nghĩa là xác suất có "dương tính giả" từ kết quả. Đây còn được gọi là lỗi Loại 1, nhưng phép tính thực sự để xác định alpha đang lấy $1 - \text{Mức độ tin cậy}$. Trong trường hợp này, mức độ tin cậy là 95% hoặc 0,95. Nếu nhà phân tích lấy 1,95 thì kết quả là 0,05. Một lời giải thích dài, nhưng đó là một lời giải thích cần thiết với OpenOffice, vì công cụ này sử dụng alpha nhiều hơn mức độ tin cậy. Kết quả không khớp chính xác với kết quả của Excel, nhưng một lần nữa, đây có thể là một tình huống làm tròn hoặc một tình huống trong đó công thức cho khoảng tin cậy có phép tính chính xác hơn kết quả khác. Sự khác biệt không lớn (chênh lệch 0,03), vì vậy những kết quả này có thể được sử dụng để kiểm chứng lẫn nhau.

Độ lệch chuẩn hoàn toàn giống nhau, điều này mang lại nhiều sự tin cậy cho tính xác thực và nhất quán của các công cụ này.

Một điểm nữa trước khi rời OpenOffice. Hãy nhớ rằng, lấy mẫu càng nhiều thì phạm vi của khoảng tin cậy càng hẹp. Điều này sẽ đúng trong tất cả các công cụ khác và trong bất kỳ công cụ thống kê nào khác liên quan đến phân tích dữ liệu. Như với bất kỳ nghiên cứu nào, lấy mẫu càng nhiều thì kết quả càng chính xác, miễn là việc lấy mẫu được thực hiện ngẫu nhiên.

4.3.3 R/RStudio/Rattle

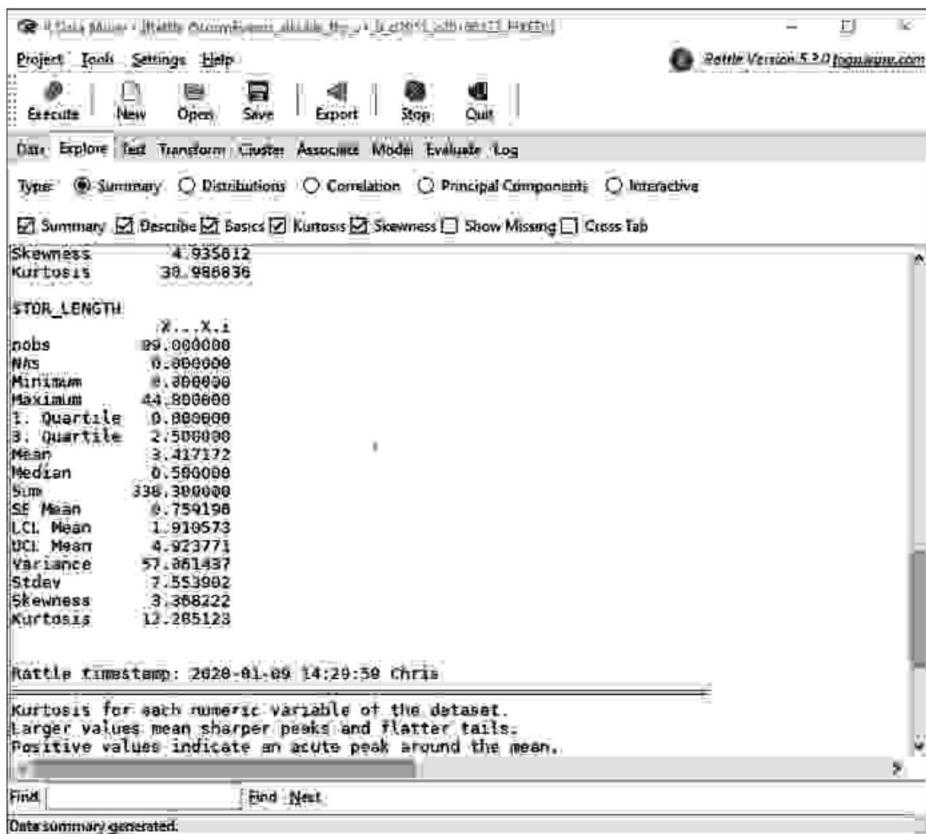
Việc sử dụng Rattle cho khoảng tin cậy gần giống như với ToolPak Phân tích trong Excel. Bước đầu tiên là nhập dữ liệu và chuẩn bị cho

phương pháp. Trong trường hợp này, đầu vào hoạt động duy nhất sẽ là TOR_LENGTH. Cũng lưu ý rằng khối "phân vùng" được chọn. Ký hiệu "70/15/15" có nghĩa là dữ liệu được chia thành 70% đào tạo, 15% xác thực và 15% kiểm tra. Điều đó có nghĩa là 70% được lấy mẫu, tức là khoảng 140 hàng. Con số này nhiều hơn một chút so với những gì thử nghiệm này đòi hỏi, vì vậy hãy thay đổi khối phân vùng thành 37/39/24, điều này sẽ làm cho dữ liệu huấn luyện ở mức 37%, tạo ra các hàng 99 và rất gần với các phần trước. Màn hình mới được cấu

No.	Variable	Data Type	Input:	Target	Risk	Ident	Ignore	Weight	Comment
29	MAGNITUDE	Constant	○	○	○	○	○	○	Unique: 1
30	MAGNITUDE_TYPE	missing	○	○	○	○	○	○	Unique: 0 Missing: 26
31	FLOOD_CAUSE	missing	○	○	○	○	○	○	Unique: 0 Missing: 26
32	CATEGORY	missing	○	○	○	○	○	○	Unique: 0 Missing: 26
33	TOR_F_SCALE	Category	○	○	○	○	○	○	Unique: 3 Missing: 39
34	TOR_LENGTH	Numeric	○	●	○	○	○	○	Unique: 71
35	TOR_WIDTH	Numeric	●	○	○	○	○	○	Unique: 39
36	TOR_OTHER_WFO	missing	○	○	○	○	●	○	Unique: 0 Missing: 26
37	TOR_OTHER_CZ_STATE	missing	○	○	○	○	●	○	Unique: 0 Missing: 26
38	TOR_OTHER_CZ_IPS	missing	○	○	○	○	●	○	Unique: 0 Missing: 26
39	TOR_OTHER_CZ_NAME	missing	○	○	○	○	●	○	Unique: 0 Missing: 26
40	BEGIN_RANGE	Constant	○	○	○	○	●	○	Unique: 1
41	BEGIN_AZIMUTH	missing	○	○	○	○	●	○	Unique: 0 Missing: 26
42	BEGIN_LOCATION	missing	○	○	○	○	●	○	Unique: 0 Missing: 26

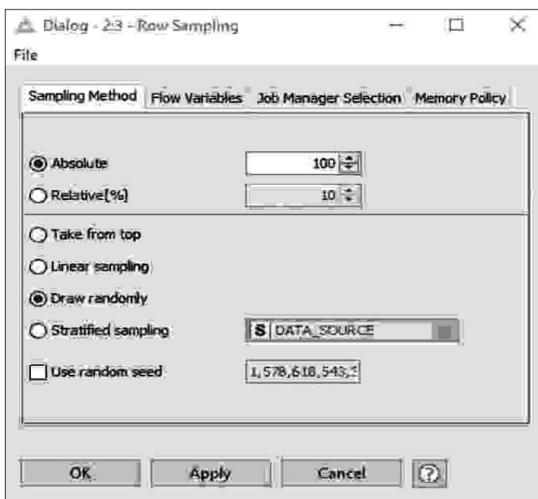
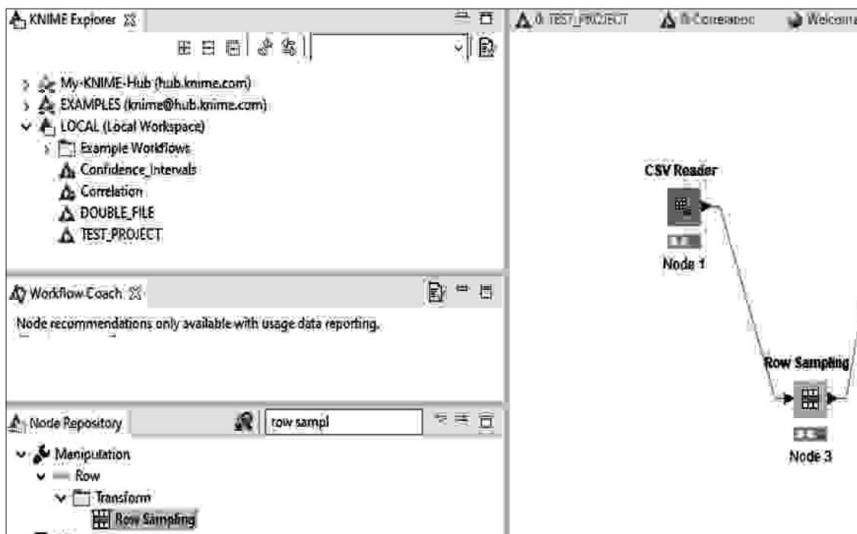
Roles noted: 269 observations and 1 input variables. The target is TOR_LENGTH, Numeric, Regression model is enabled.

Đây là lúc Rattle tự động đi xa hơn. Lưu ý rằng bên cạnh khối phân vùng có một nút "hạt giống". Điều này làm là lấy mẫu ngẫu nhiên tập dữ liệu bằng cách sử dụng giá trị gốc làm điểm bắt đầu. Điều này có nghĩa là dữ liệu này sẽ được lấy mẫu ngẫu nhiên và dữ liệu đó sẽ được kiểm tra với màn hình tiếp theo để thống kê mô tả, được minh họa như sau. Đảm bảo rằng các ô được đánh dấu gần như toàn bộ ngoại trừ hai ô cuối cùng.

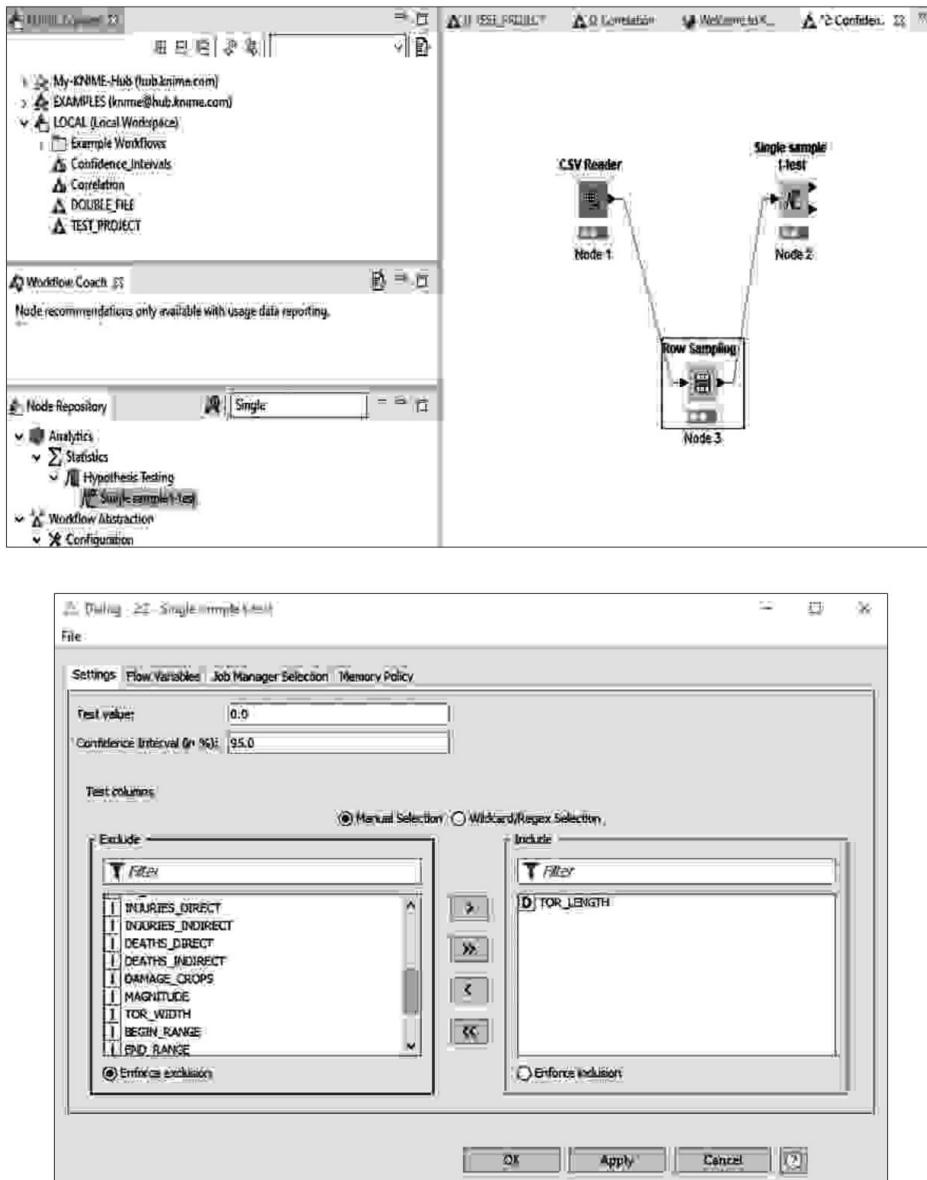


4.3.4 KIẾN THỨC

KNIME cung cấp một nút cho khoảng tin cậy, nhưng nó là một phần của một nút khác, vì vậy điều quan trọng là sử dụng nghiên cứu để xem cách các nút này có thể được sử dụng trong một số loại phương pháp. Đầu tiên, điều quan trọng là sử dụng lấy mẫu TOR_LENGTH như chúng tôi đã làm trong các phần khác. Trong trường hợp này, nút cần đưa vào được gọi là "Lấy mẫu hàng". Sau khi được kéo, đặt và kết nối, cấu hình sẽ giống như thế này để lấy mẫu ngẫu nhiên. Tuy nhiên, xin lưu ý rằng có nhiều kết hợp lấy mẫu có sẵn với nút này. Cấu hình được sử dụng trong phần này như sau:

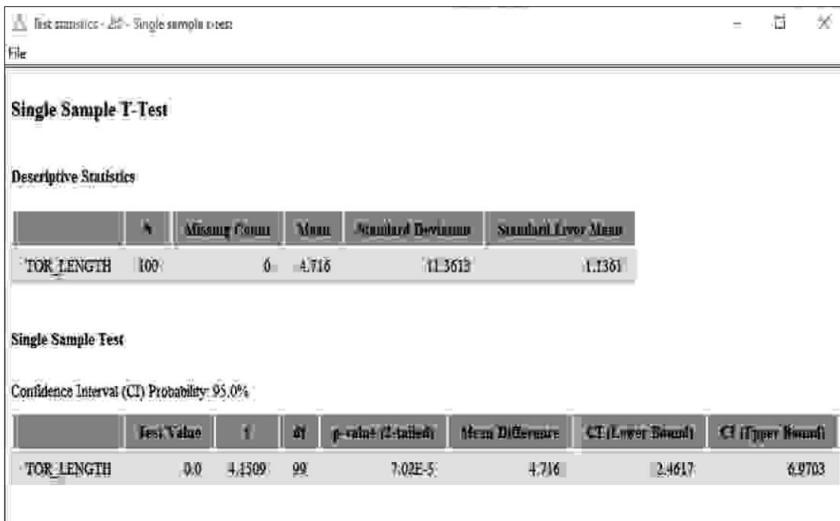


Như nhà phân tích có thể thấy từ màn hình cấu hình, mẫu hiện tại dành cho 100 giá trị được rút ngẫu nhiên. KNIME cung cấp một nút cho việc lấy mẫu này và theo kinh nghiệm, đó là một cách hay để lấy mẫu chỉ với một bước. Quá trình lấy mẫu đã hoàn tất, nhưng vẫn cần phải có một nút cung cấp khoảng tin cậy và nút đó đến từ nút Kiểm tra T mẫu đơn. Sau khi nút đó được kéo, đặt và kết nối, các màn hình sau sẽ hiển thị quy trình cuối cùng và màn hình cấu hình cho nút này.



Như nhà phân tích có thể thấy từ các màn hình này, cấu hình cho nút thử nghiệm t rất đơn giản. Khoảng tin cậy là 95%, có thể thay đổi được và biến là TOR_LENGTH. Lưu ý rằng "Giá trị thử nghiệm" là 0. Có một lý do cho điều này trong trường hợp này. Trong các bài kiểm tra t bình thường với một

mẫu, giả thuyết không sẽ dựa trên giá trị trung bình của mẫu bằng một giá trị. Giá trị thử nghiệm là giá trị đó, vì vậy thật tốt khi biết lý do tại sao khói đó có sẵn trong nút này. Điều này sẽ không được sử dụng trong trường hợp này. Sau khi thực hiện tất cả các nút, kết quả sau đây có sẵn bằng cách nhấp chuột phải vào nút kiểm tra t và chọn "Chế độ xem: Kiểm tra thống kê". Nhà phân tích có thể thấy CI (Giới hạn dưới) và CI (Giới hạn trên). Điều đó có nghĩa là trung bình dân số nằm ở đâu đó giữa hai giá trị này. Có 95% khả năng nó nằm giữa hai giá trị này. Đừng lo lắng về bất kỳ số nào khác trong hàng này, vì chúng liên quan đến bài kiểm tra t. Tuy nhiên, như nhà phân tích có thể thấy, các con số rất gần với các phần khác, điều này cho thấy rằng kết quả ít nhất cũng phần nào phù hợp với mẫu 100.



4.4 LẤY MẪU NGẪU NHIÊN

Qua nhiều năm giảng dạy về thống kê và phân tích dữ liệu, lấy mẫu ngẫu nhiên là một khái niệm thường bị sinh viên hiểu sai. Học sinh thường cần nhắc chọn 10 hoặc 20 giá trị đầu tiên trong một tập dữ liệu lấy mẫu ngẫu nhiên khi nó không ngẫu nhiên. Ngẫu nhiên đang xem xét tất cả các giá trị như nhau (Reinhart, 2015). Đây là điều mà một số nhà nghiên cứu không làm vì lý do thuận tiện hoặc cần thiết (Reinhart, 2015). Tuy nhiên, để đo chính xác trung tâm của tập dữ liệu hoặc để dự đoán chính xác tác động của một sự kiện đối với một sự kiện khác trong dân số, thì cần phải có một mẫu chính xác. Có một số phương pháp

thực hiện chức năng này và tất cả các công cụ được đề cập đều có cách lấy mẫu ngẫu nhiên. Một số trong số này đã được đề cập trong các phần trước, nhưng đây là phần ôn lại các phương pháp đó.

4.4.1 Excel

Excel có ToolPak Phân tích, do đó có chức năng lấy mẫu, nhưng khi nhà phân tích thực hiện việc này, họ sẽ nhận được các bản sao trong quy trình. Một cách để ngăn trùng lặp là chỉ định một giá trị duy nhất cho mọi sự kiện, do đó ngăn chặn trùng lặp. Bước đầu tiên trong quy trình này là tài tập dữ liệu đã được sử dụng trong các phần trước—theo dõi cơn lốc xoáy năm 1951. Khi đã hoàn tất, hãy chèn một cột vào đầu dữ liệu và đặt tên là "Số ngẫu nhiên", vì đây sẽ là số duy nhất sẽ được gán cho mỗi hàng.

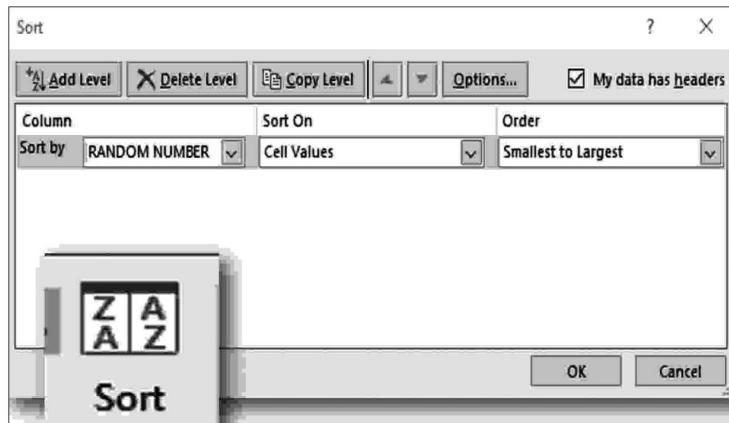
Trong hàng đầu tiên của cột đó, đặt công thức sau:

$$= \text{RAND}()$$

Sau bước đó, đảm bảo rằng tất cả cột đó có cùng một công thức bằng cách bấm đúp vào "điều khiển điền" nằm ở dưới cùng bên phải của ô công thức ("điều khiển điền" sẽ giống như một dấu cộng in đậm). Bằng cách nhấp đúp vào ô điều khiển điền, tất cả các ô trống trong cột công thức sẽ được điền bằng một số ngẫu nhiên. Trên thực tế, mỗi khi nhà phân tích thực hiện bất kỳ phép tính nào hoặc nhấn phím ENTER, các hàng sẽ xáo trộn lại. Điều này sẽ làm cho bất kỳ mẫu dữ liệu nào thực sự ngẫu nhiên. Tập dữ liệu đã hoàn thành, trước khi xáo trộn và sau khi xáo trộn, được hiển thị như sau:

	RANDOM NUMBER	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEAR	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE
2	=RAND()	195109	9	915	195109	9	915		10047282	MISSISSIPP
3		195106	17	2200	195106	17	2200		10028729	KANSAS
4		195103	28	510	195103	28	510		10120421	TEXAS
5		195105	9	1830	195105	9	1830		10099717	OKLAHOM
6		195107	15	1620	195107	15	1620		10099742	OKLAHOM
7		195105	8	1800	195105	8	1800		10028691	KANSAS
8		195103	30	1500	195103	30	1500		10104933	PENNSYLV
9		195105	11	1330	195105	11	1330		10104934	PENNSYLV
10		195106	27	2204	195106	27	2204		10104935	PENNSYLV
11		195107	21	1100	195107	21	1100		10104936	PENNSYLV
12		195104	29	1815	195104	29	1815		10082587	NEW JERSE
13		195102	19	1830	195102	19	1830		10099493	OKLAHOM
14		195105	3	1335	195105	3	1335		10039190	MICHIGAN
15		195106	1	1800	195106	1	1800		10039191	MICHIGAN

Để sắp xếp nhằm đảm bảo rằng số ngẫu nhiên bao gồm tất cả các cột, hãy sử dụng tùy chọn sắp xếp trong tab Dữ liệu của màn hình, tương tự như màn hình sau. Bằng cách nhấp vào biểu tượng này trên thanh công cụ, menu sau sẽ xuất hiện. Sử dụng mũi tên xuống để chọn cột có số ngẫu nhiên và nhấp vào OK.

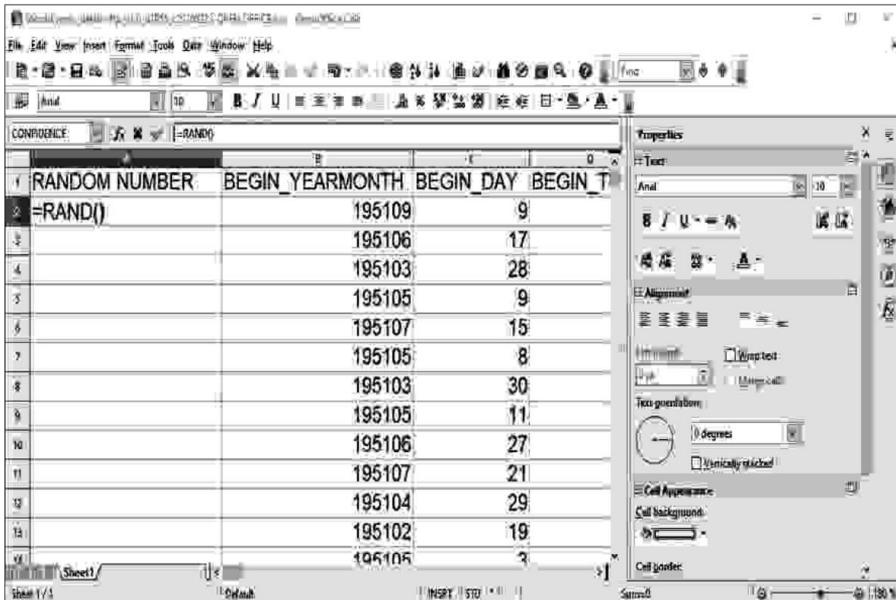


Một lời cảnh báo tại thời điểm này cho chức năng này. Vui lòng đảm bảo ô được chọn là ô trong cột SỐ NGẪU NHIÊN. Nếu không, lyst ana sẽ chỉ được sắp xếp dựa trên cột được chọn. Nếu cột được chọn là cột tháng, thì việc sắp xếp sẽ theo tên tháng-chứ không phải theo số ngẫu nhiên. Nếu ô được chọn là số ngẫu nhiên, nhà phân tích sẽ thấy các cột được xáo trộn tương ứng. Đây là một cách đơn giản để xáo trộn bộ bài mà không sợ dữ liệu bị sai lệch thông qua lấy mẫu có hệ thống mà không có tính ngẫu nhiên.

4.4.2 Văn phòng mở

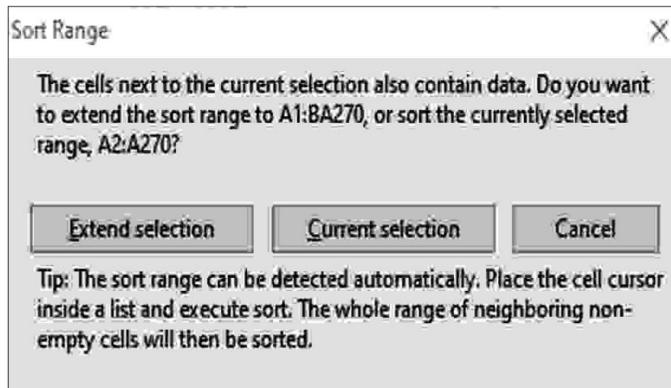
Hãy coi OpenOffice rất giống với Excel, với một số khác biệt rất nhỏ trong các công thức. Do đó, việc lấy mẫu ngẫu nhiên trong OpenOffice sẽ rất giống với quy trình trong Excel.

Tất nhiên, bước đầu tiên là nhập tập dữ liệu, sau đó ana lyst thực hiện các chức năng tương tự như Excel bằng cách chèn một cột và đặt tên cho nó là SỐ NGẪU NHIÊN và đặt cùng một công thức như Excel vào ô đầu tiên của cột đó. Màn hình sẽ giống như thế này khi hoàn thành bước đó:



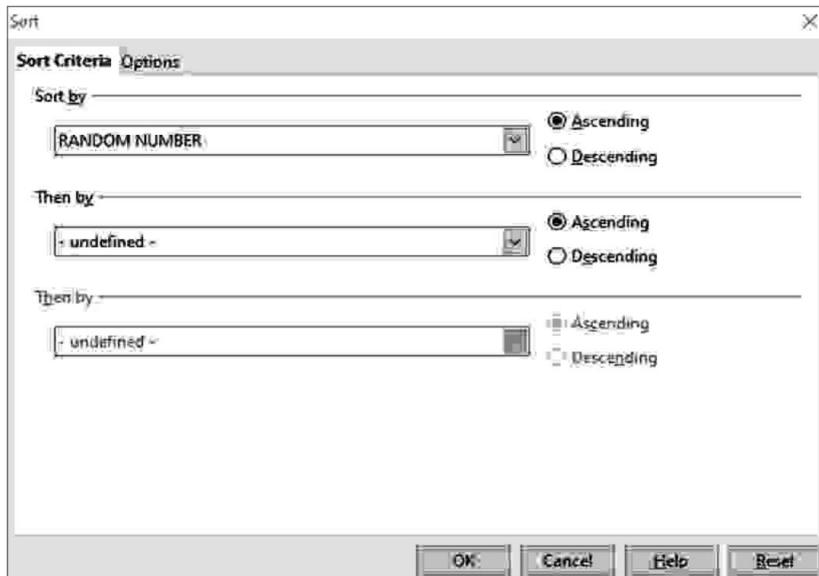
Nếu màn hình này trông quen thuộc, thì nó phải như vậy, vì định dạng rất giống với phần trước. Ô, và việc nhấp đúp vào “fill handle” cũng hoạt động với công cụ này. Điều này giúp việc sao chép các công thức xuống một cột dễ dàng hơn nhiều, bao gồm tất cả các hàng. Một chút thận trọng ở đây là cần thiết. Nếu bất kỳ hàng nào trống hoặc chứa các giá trị trống bên cạnh cột mà bạn đang bấm đúp, thì có thể có một điểm dừng tại hàng đó, vì vậy sẽ có lợi cho nhà phân tích khi kiểm tra cột để đảm bảo rằng tất cả các hàng đều có một số ngẫu nhiên.

Sau khi hoàn thành bước đó, hãy chuyển đến tab Dữ liệu trên thanh công cụ và chọn tùy chọn Sắp xếp. Sau khi hoàn thành, màn hình sau sẽ xuất hiện:

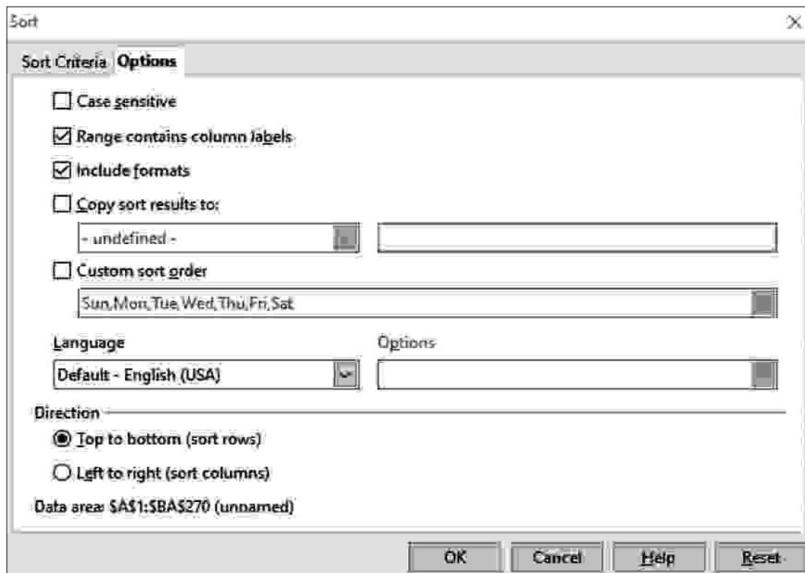


Chọn nút “Mở rộng lựa chọn” để đặt toàn bộ tập dữ liệu dưới tiêu điểm của sắp xếp số ngẫu nhiên. Theo cách này, toàn bộ tập dữ liệu (với các hàng nguyên vẹn, điều này rất quan trọng), sẽ được sắp xếp dựa trên cột số ngẫu nhiên. Sau đó, chỉ cần tiếp tục sắp xếp để xáo trộn tập dữ liệu.

Một khuyến nghị về sắp xếp tập dữ liệu là không lưu tập dữ liệu cho đến khi nhà phân tích biết rằng tất cả các hàng vẫn còn nguyên vẹn.



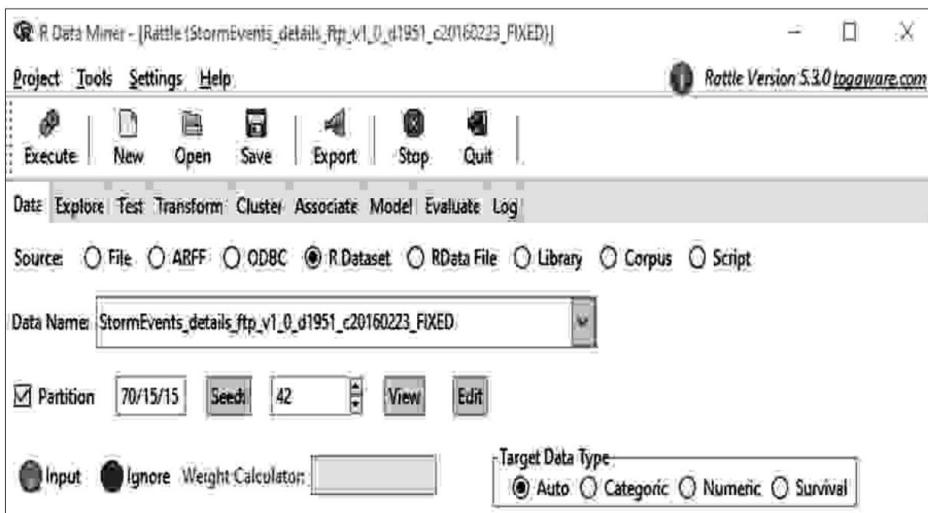
Màn hình sau đây hiển thị tab “tùy chọn” của cùng một menu như trước đây. Mặc dù nhà phân tích có thể lướt qua tab này, nhưng điều quan trọng là phải xem các tùy chọn khác nhau được cung cấp với chức năng sắp xếp này để đảm bảo rằng chúng phù hợp với ý định của nhà phân tích. Một hộp kiểm như “Phạm vi chứa các nhãn cột” rất quan trọng khi sắp xếp vì nếu bỏ chọn hộp này, các tiêu đề sẽ được sắp xếp cùng với dữ liệu. Đây có thể là một kết quả lộn xộn sẽ ảnh hưởng đến các thử nghiệm khác nhau do nhà phân tích thực hiện. Nói cách khác, hãy cẩn thận với những màn hình khác nhau này và khám phá chúng bất cứ khi nào có thể.



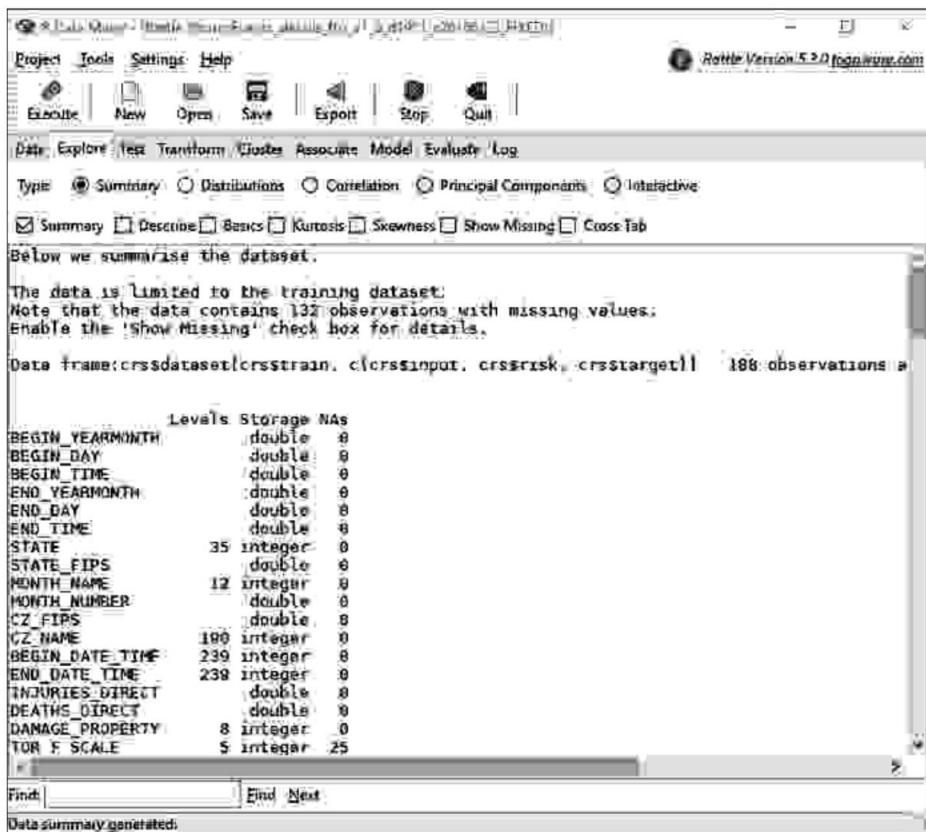
4.4.3 R/RStudio/Rattle

Việc lấy mẫu được thực hiện với Rattle đã được đề cập trong các phần trước nhưng sẽ được làm mới với chủ đề cụ thể này. Khi gói Rattle được kích hoạt trong RStudio và bộ dữ liệu được tải, bước tiếp theo sẽ là lấy mẫu dữ liệu ngẫu nhiên. Trong trường hợp này, toàn bộ tập dữ liệu sẽ được đưa vào lấy mẫu, trong khi ở phần trước, các cột cụ thể được xác định trong tab Dữ liệu.

Tab Dữ liệu sẽ trông như thế này nếu nhà phân tích muốn lấy mẫu 50% tập dữ liệu (được gọi là tập dữ liệu huấn luyện) và sử dụng dữ liệu đó như một cách kiểm tra các phương pháp và chức năng khác nhau. Tại thời điểm này, nhà phân tích phải đảm bảo rằng chỉ có một biến "Rủi ro" (công cụ sẽ cảnh báo bạn nếu có nhiều hơn) và các giá trị "Phân vùng" cộng lại bằng 100 (công cụ cũng sẽ cảnh báo bạn về điều này). Màn hình sau đây sẽ tổng hợp nó để lấy mẫu. Một lưu ý ở đây rất quan trọng-giá trị "hạt giống" là một giá trị thường dựa trên đồng hồ máy tính, nhưng bằng cách đặt cùng một hạt giống, các giá trị ngẫu nhiên giống nhau sẽ được tạo lại, điều này có thể hữu ích khi so sánh thử nghiệm trên cùng một tập dữ liệu. Tuy nhiên, nhà phân tích có thể đặt lại hạt giống bằng cách nhấn nút "hạt giống". Mặc định là 42 để bắt đầu quá trình này.



Làm thế nào để nhà phân tích biết nếu việc lấy mẫu thực sự xảy ra? Sử dụng một chức năng rất đơn giản trong Rattle (như số liệu thống kê tóm tắt trên tab "Khám phá") và màn hình sau đây cho thấy có 188 quan sát, chiếm 50% toàn bộ tập dữ liệu. Việc lấy mẫu đã hoạt động và có số lượng giá trị thích hợp trong thử nghiệm. Nếu nhà phân tích có một con số cụ thể mà họ cần lấy mẫu (điều đó sẽ được thảo luận trong phần bổ sung), thì việc tính toán con số đó tương đối đơn giản. Tổng số hàng là 279, vì vậy hãy lấy số cần lấy mẫu, chẳng hạn như 140 và chia số đó cho tổng số hàng. Phần trăm đó là những gì bạn đưa vào khía "phân vùng" của tab Dữ liệu trong Rattle. Trong trường hợp này, nó sẽ là 140/279, tức là xấp xỉ 50%. Tuy nhiên, hãy nhớ rằng việc lấy mẫu có thể được điều chỉnh theo bất kỳ cách lấy mẫu nào cần thiết để tăng "sức mạnh" của kiểm tra thống kê (sẽ được thảo luận sau) hoặc bất kỳ yếu tố nào khác có thể giúp tăng độ chính xác của phương pháp thống kê hoặc kiểm tra.



4.4.4 KIẾN THỨC

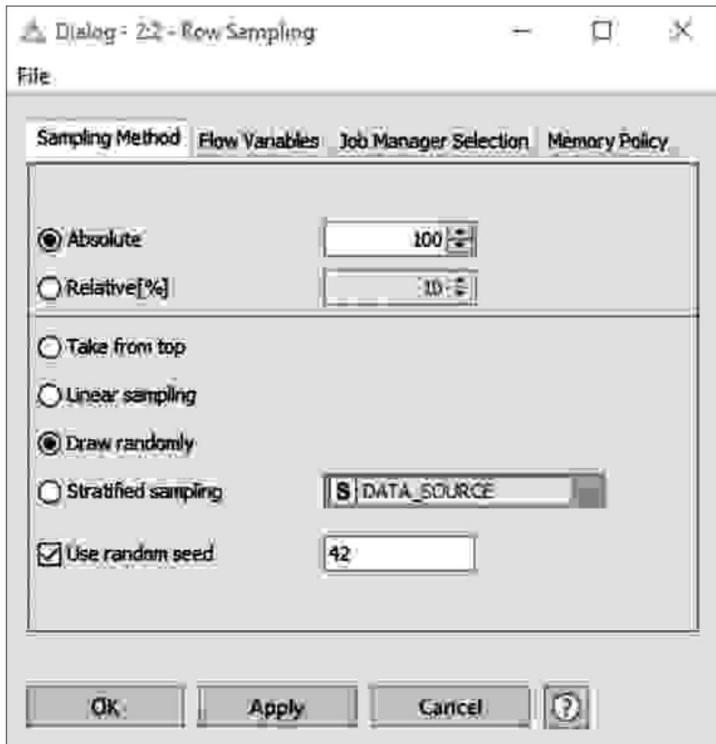
KNIME có khả năng lấy mẫu thông qua, đợi nó, một nút cho mục đích này.

Có một điều cần được giải thích về các nút trong KNIME.

Không giống như các công cụ khác, KNIME liên kết rõ ràng các nút với các hàng trong tập dữ liệu. Điều này rất quan trọng khi sử dụng KNIME, vì điều này thực sự rất chính xác. Khi việc lấy mẫu được thực hiện trên bất kỳ tập dữ liệu nào, nếu các tiêu đề cột xử lý các biến, thì việc lấy mẫu được thực hiện với các hàng. Đây là một trong những lĩnh vực mà KNIME khác với các công cụ khác, nhưng sự khác biệt đó không làm cho nó sai, chỉ là một góc nhìn khác với một từ vựng khác.

Bước đầu tiên với KNIME là kéo và đặt nút Trình đọc CSV với tập dữ liệu đã nhập, sau đó kéo và đặt nút Lấy mẫu hàng, kết nối nó với nút Trình đọc CSV. Lời nhắc nhanh: mỗi nút phải được định cấu hình và thực thi trước khi quá trình có thể hoàn tất.

Bước tiếp theo là thiết lập việc lấy mẫu theo mong muốn của nhà phân tích. Đây là được thực hiện bằng cách nhấp đúp vào nút Lấy mẫu hàng để hiển thị màn hình này.



Xin lưu ý rằng nhà phân tích có thể chọn tỷ lệ phần trăm của tập dữ liệu để lấy mẫu hoặc một số tuyệt đối. Ngoài ra, hãy lưu ý khỏi "Sử dụng hạt giống ngẫu nhiên", mà nhà phân tích có thể đặt thành cùng một hạt giống đã được đặt trong Rattle. Cùng với đó, nếu phân vùng Rattle được đặt thành 50%, nhà phân tích cũng có thể đặt tỷ lệ phần trăm đó trong màn hình cấu hình này thành 50% để xem liệu việc lấy mẫu có tạo ra kết quả giống hay khác không. Đây là một cách tuyệt vời để xác định các kết quả khác nhau bằng cách sử dụng hai công cụ khác nhau, cả hai đều kiểm tra tính nhất quán và độ chính xác của các phương pháp. Trong trường hợp này, nhà phân tích đặt hạt giống ngẫu nhiên giống như với Rattle và số thành 188, giống như với Rattle. Nút được thực hiện tạo ra các kết quả sau:

TOR_LENGTH	TOR_LENGTH	0	92.6	4.874	10.944	119.769	4.446	25.933	916.3
------------	------------	---	------	-------	--------	---------	-------	--------	-------

Nhà phân tích có thể kiểm tra điều này với kết quả Rattle để xem mức độ gần nhau của chúng về giá trị. Điểm mấu chốt là các phương pháp lấy mẫu trong một số công cụ dễ dàng hơn nhiều và rất ít nếu có bất kỳ sự trùng lặp nào được hoàn thành với việc lấy mẫu ngẫu nhiên. Một số công cụ có các chức năng được tạo để lấy mẫu, trong khi những công cụ khác cần cấu hình nhiều hơn một chút. Tuy nhiên, rõ ràng là việc lấy mẫu sẽ tiếp tục với nhà phân tích dữ liệu trong tương lai, vì việc phân tích dân số hơi khó khăn. Việc lấy mẫu là quan trọng và nhất quán nếu được thực hiện ngẫu nhiên.

CHƯƠNG 5

PHƯƠNG PHÁP THỐNG KÊ CHO CÔNG CỤ CỤ THỂ

5.1 CÔNG SUẤT

Sức mạnh là thứ mà một giảng viên đại học có kinh nghiệm về thống kê sẽ đề cập đến với một số sở thích thoáng qua, nhưng chắc chắn không phải là chi tiết tuyệt vời. Theo một tài liệu tham khảo, sức mạnh không chỉ là một lựa chọn mà còn là một yêu cầu (Reinhart, 2015). Để hiểu rõ hơn về quyền lực, việc xem xét các loại lỗi là cần thiết. Trọng tâm của việc kiểm tra giả thuyết trong cuốn sách này là lỗi Loại 1 (hoặc dương tính giả). Bằng cách nêu rõ "alpha" là 0,05, nhà phân tích đang thực hiện lỗi Loại 1, có nghĩa là chỉ có 5% khả năng xảy ra kết quả dương tính giả. Dương tính giả có nghĩa là xét nghiệm cho thấy một kết quả có thể không chính xác, chẳng hạn như kết quả xét nghiệm cúm cho thấy một người bị cúm nhưng không mắc bệnh này. Trong bài kiểm tra nguồn điện, lỗi Loại 2 được thực hiện, đó là âm tính giả. Về cơ bản, điều này có nghĩa là, nếu sức mạnh là 80% (hoặc 0,8), thì có 80% khả năng xét nghiệm cho biết ai đó không bị cúm và thực tế sẽ không bị cúm. Vẫn có 20% khả năng âm tính giả hoặc một người nào đó đã được xét nghiệm cúm cho kết quả âm tính nhưng thực sự bị cúm. Trong thế giới thống kê, 80% sức mạnh là chấp nhận được và thông thường. Thách thức thực sự đằng sau điều này là có một số sự kiện bắt buộc phải được lấy mẫu để tạo ra kết quả 80% sức mạnh này. Điều sẽ được chứng minh ở đây là quá trình để có được kết quả lấy mẫu đó, và do đó, một kết quả thống kê chính xác hơn.

5.1.1 R/RStudio/Rattle

Excel không có khả năng thực hiện quyền lực ngoại trừ bằng cách nhập công thức theo cách thủ công và điều tương tự cũng xảy ra với OpenOffice. Để làm cho điều này trở nên dễ dàng nhất có thể đổi với nhà phân tích, phần này sẽ chỉ tập trung vào công cụ trong văn bản này có thể thực hiện chức năng nguồn trực tiếp từ một chức năng hiện có. Đây sẽ là công cụ R/RStudio/Rattle.

Bước đầu tiên để thực hiện quy trình này là nhập bộ dữ liệu cần thiết, đó sẽ là dữ liệu theo dõi cơn lốc xoáy năm 1951 và 1954, tập trung vào các biến TOR_LENGTH như trong phần trước. Một khi điều này được hoàn thành, hãy xác định chức năng sẽ cần thiết; trong trường hợp này, nó sẽ là gói "pwr", có thể được cài đặt giống như bất kỳ gói nào khác trong R hoặc RStudio, như được mô tả trong phần trước. Khi điều này được hoàn thành, chỉ cần điền vào các tham số của công thức với các giá trị để có được giá trị còn thiếu.

Chẳng hạn, nếu nhà phân tích muốn biết họ cần bao nhiêu mẫu để có công suất 80% (tương tự như có "alpha" là 0,05), khi họ có hai biến có trung bình là 5 và 4, với độ lệch chuẩn tổng thể là 5, nhà phân tích cần tìm xem họ cần bao nhiêu mẫu để đạt được công suất 80%. Trong R/RStudio, sau khi cài đặt gói "pwr", nhà phân tích cần đưa công thức sau vào không gian làm việc của RStudio.

```
> pwr.norm.test(d=.2,sig.level=.05,power=.8,alternative=
  "hai mặt")
```

Tính toán công suất trung bình cho phân phối bình thường với phương sai đã biết

```
d = 0,2
n = 196,2215
sig.level = 0,05
sức mạnh = 0,8
thay thế = hai mặt
```

Lý do sử dụng "hai mặt" là nhà phân tích không quan tâm liệu một ý nghĩa nhỏ hơn hay lớn hơn ý nghĩa kia, chỉ cần chúng bằng nhau hay không bằng nhau. Số lượng giá trị cần thiết để có được công suất 80% sẽ là 197, vì các sự kiện thường là số nguyên nên 196,2 được làm tròn lên. Điều gì sẽ xảy ra nếu nhà phân tích muốn xem liệu một giá trị trung bình có lớn hơn giá trị trung bình kia không? Cần bao nhiêu sự kiện sau đó để có được 80% sức mạnh? Đây là một thay đổi đơn giản trong công thức, do đó công thức bây giờ sẽ được đọc như sau, nhưng phương án thay thế sẽ được thay đổi thành "lớn hơn" để giải quyết đúng phương án thay thế

giả thuyết. Bạn đọc có thể nhớ rằng kiểm định giả thuyết đã được đề cập trong phần trước, và một khía cạnh của kiểm định giả thuyết vẫn còn tồn tại là giả thuyết không luôn luôn có một giá trị bằng với giá trị kia (chẳng hạn như $mean1=mean2$, v.v.). Giả thuyết thay thế sẽ là một trong ba giả thuyết sau: "một giá trị nhỏ hơn giá trị kia", "một giá trị lớn hơn giá trị kia" hoặc "một giá trị không bằng giá trị kia". Cả hai.

"hai bên" có nghĩa là tùy chọn thứ ba hoặc giá trị trung bình này không bằng giá trị trung bình kia. Tùy chọn "lớn hơn" có nghĩa là một giá trị trung bình lớn hơn giá trị trung bình kia. Nếu nhà phân tích thay đổi phương án thành "lớn hơn", thì kết quả sẽ thay đổi như sau:

```
> pwr.norm.test(d=.2,sig.level=.05,power=.8,alternative=
  "lớn hơn")
```

Tính toán công suất trung bình cho phân phối bình thường với phuơng sai đã biết

```
d = 0,2
n = 154,5639
sig.level = 0,05
sức mạnh = 0,8
```

thay thế = lớn hơn

Như nhà phân tích có thể thấy, mẫu thay đổi từ 197 thành 155. Điều này có nghĩa là, để có được 80% sức mạnh, sẽ tốn ít nỗ lực lấy mẫu hơn nếu giả thuyết thay thế lớn hơn. Tại thời điểm này, tùy chọn cuối cùng hoặc "ít hơn" chưa được chọn, nhưng điều này là không thể với "d" là số dương. Lý do là một phần của phép tính đi vào "d" là $(mean1-mean2)/\text{độ lệch chuẩn}$. Nếu "d" là dương, thì "less" không phải là một tùy chọn vì $mean1-mean2$ là dương. Nhà phân tích sẽ phải thay đổi "d" thành số âm để sử dụng tùy chọn "ít hơn". Cảnh báo spoiler: giá trị sau khi thực hiện việc này sẽ bằng với số "lớn hơn". Lý do cho điều này là vì nhà phân tích đang kiểm tra phân phối bình thường (hoặc những gì nhà phân tích nghĩ có thể là phân phối bình thường).

Đó là câu trả lời của R/RStudio cho phép tính công suất. Như mọi người có thể thấy, nhà phân tích cần một số nỗ lực, nhưng nó vẫn đơn giản hơn nhiều so với việc thực hiện chức năng tương tự trong bất kỳ công cụ nào khác. Do đó, phần này sẽ chỉ đề cập đến công cụ R/RStudio để tính toán công suất. Thông tin thêm về dụng cụ điện cầm tay và tầm quan trọng của nó có sẵn bằng cách xem các tài liệu tham khảo ở cuối cuốn sách này.

5.2 THỦ NGHIỆM F

F-Test là một cách kiểm tra xem hai biến đang được kiểm tra có phương sai bằng nhau hay không bằng nhau. Điều này rất quan trọng bất cứ khi nào tiến hành thử nghiệm t hai mẫu, vì phép tính đối với Thông kê T là khác nhau đối với các phương sai bằng nhau hoặc không bằng nhau. Thủ nghiệm này còn được gọi là Thủ nghiệm Levene, được đặt tên theo tác giả của một bài tiểu luận về phương pháp này (Levene, 1960), phát hiện với cơ hội thông thường (thường là 95%) liệu các phương sai có bằng nhau giữa các biến khác nhau giữa các bộ dữ liệu hay trong phạm vi một tập dữ liệu. Hầu hết các công cụ đều đã có sẵn chức năng này, nhưng điều thú vị là chúng thường như không được sử dụng để kiểm tra phương sai trước khi sử dụng kiểm định t. Có một trang web tuyệt vời có thể giúp nhà phân tích giải thích và khái niệm Levene, cùng với các công thức và công cụ (Technology, 2013).

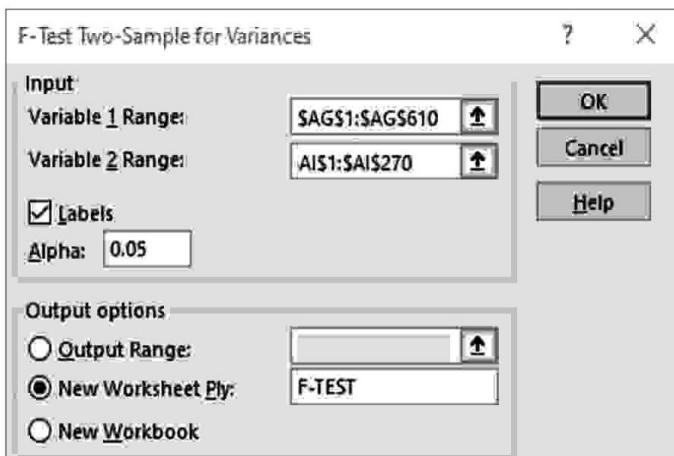
Đây được coi là một kỹ thuật khác thường, bởi vì nhà phân tích cần sử dụng kỹ thuật này trước khi tiến hành các thử nghiệm khác.

5.2.1 Excel

Excel có ToolPak phân tích, có thể dễ dàng thực hiện Levene F-Test và thực sự có một lựa chọn cho điều này trong ToolPak. Quá trình sử dụng này rất đơn giản.

Bước đầu tiên là nhập dữ liệu, trong trường hợp này sẽ là theo dõi cơn lốc xoáy năm 1951 và 1954 như đã được thực hiện cho thử nghiệm t.

Sau khi nhập, hãy mở Analysis ToolPak và chọn F-Test Two Sample for Variance, nằm ngay bên dưới Làm mìn theo cấp số nhân. Lúc này, hãy chọn các cột (hai cột) dữ liệu cần so sánh, giống như xuất hiện trên màn hình sau.



Nhà phân tích sẽ nhận thấy rằng khối "Nhân" đã được chọn và một trang tính mới đang được tạo để lưu giữ kết quả. Vì mục đích của phần này, các cột năm 1951 và 1954 được đánh dấu "TOR_LENGTH," vốn là phần chính của các minh họa này, đang được sử dụng lại để đảm bảo tính nhất quán.

Nhà phân tích có thể đặt câu hỏi về lý do sử dụng bất kỳ loại thử nghiệm nào để xem liệu các phương sai có khác nhau hay không, vì chúng sẽ khác nếu nhà phân tích thực hiện thống kê tổng hợp của các bộ dữ liệu. Tuy nhiên, do số lượng sự kiện trong mỗi mẫu và sự khác biệt giữa chúng (279 so với hơn 600), nên việc đưa ra đánh giá về các phương sai dựa trên quan sát là không thực sự phù hợp để kiểm tra thống kê. Bằng cách thực hiện Kiểm tra Levene, kết quả cho nhà phân tích biết khả năng các phương sai là không bằng nhau do sự chênh lệch về số lượng mẫu. Điều này rất quan trọng đối với thử nghiệm t tiếp theo. Kết quả từ Bài kiểm tra Levene Excel trước đó sẽ theo sau. Tất cả nó có nghĩa gì?

	TOR_LENGTH	TOR_LENGTH
Mean	5.322003284	4.443494424
Variance	114.6427058	104.6703773
Observations	609	269
df	608	268
F	1.095273647	
P(F<=f) one-tail	0.195331243	
F Critical one-tail	1.190155543	

Khu vực mà nhà phân tích muốn tập trung vào là ba hàng cuối cùng, cho chúng ta biết liệu giả thuyết không (rằng cả hai phương sai đều bằng nhau) có đúng hay không. "F" là 1,09 và "F Critical one-tail" là 1,19. Vì F nhỏ hơn F tới hạn, nhà phân tích sẽ không bác bỏ giả thuyết không, nghĩa là hai phương sai bằng nhau. Nếu nhà phân tích muốn xác nhận, hãy xem biểu đồ một đuôi " $P(F \leq f)$ ", hiển thị giá trị là 0,193. Giá trị này lớn hơn "alpha" được đặt ở màn hình cấu hình, là 0,05. Nếu giá trị p từ Thủ nghiệm F lớn hơn alpha, thì giả thuyết không bị bác bỏ và

chúng ta có thể suy ra rằng hai phương sai bằng nhau. Trong trường hợp này, giá trị p lớn hơn alpha, cho thấy khả năng các phương sai bằng nhau. Tại thời điểm này, nhà phân tích sau đó có thể chọn lựa chọn kiểm tra t chính xác để chạy quy trình đó.

5.2.2 R/RStudio/Rattle

Levene Test for Rattle là một quy trình tương đối đơn giản. Tuy nhiên, và điều này rất quan trọng, Rattle chỉ chứa một tập dữ liệu tại một thời điểm. Để thực hiện kết hợp các biến, nhà phân tích sẽ phải chuẩn bị dữ liệu trước khi đưa dữ liệu vào Rattle, nếu không thì sử dụng tính năng lập trình vốn có của RStudio. Trong trường hợp này, tính năng lập trình RStudio là sự lựa chọn, đơn giản vì nó chỉ là một hoặc hai dòng mã.

Bước đầu tiên là thông thường-nhập dữ liệu, việc này đã được hoàn thành. Sau đó, có bước tiên quyết cần thiết để đảm bảo rằng gói phù hợp đã được cài đặt. Trong trường hợp này, hàm var.test nằm trong gói STATS, gói này đã được cài đặt trong R và sau đó được cài đặt tự động trong RStudio. Cách mà nhà phân tích có thể tìm thấy này sẽ được giải thích trong thông tin bổ sung.

Sau khi gói thích hợp được cài đặt và kích hoạt, sẽ cần một dòng mã để đảm bảo rằng cả hai tệp đều được so sánh đầy đủ. Chỉ cần nhớ rằng cả hai tệp phải được nhập vào RStudio để quá trình so sánh diễn ra.

Các dòng mã sau đây và kết quả được đưa vào để xem xét. Một lần nữa, hãy nhớ rằng những kết quả này có thể không giống như với Excel. Lý do chính là thuật toán cơ bản có thể hơi khác một chút, nhưng kết quả sẽ giống nhau.

```
> tor1951<- StormEvents_details_ftp_v1_0_d1951_c20160223
> tor1954<- StormEvents_details_ftp_v1_0_d1954_c20160223
> var.test(tor1951$TOR_LENGTH,tor1954$TOR_LENGTH)
```

Phép thử F để so sánh hai phương sai

dữ liệu: tor1951\$TOR_LENGTH và tor1954\$TOR_LENGTH

F = 0,91301, số df = 268, mẫu số df = 608, giá trị p =

0,3907

giả thuyết thay thế: tỷ lệ phương sai thực sự không bằng 1

Khoảng tin cậy 95 phần trăm:

0,7478819 1,1237249

Ước tính mẫu:

tỷ lệ phương sai

0,9130138

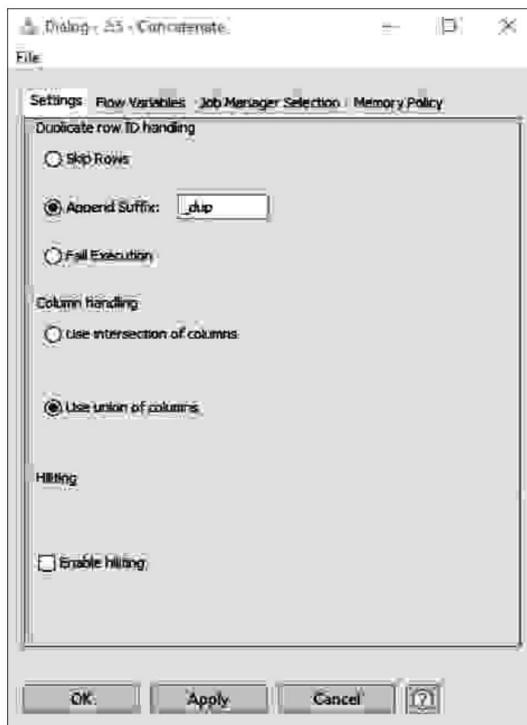
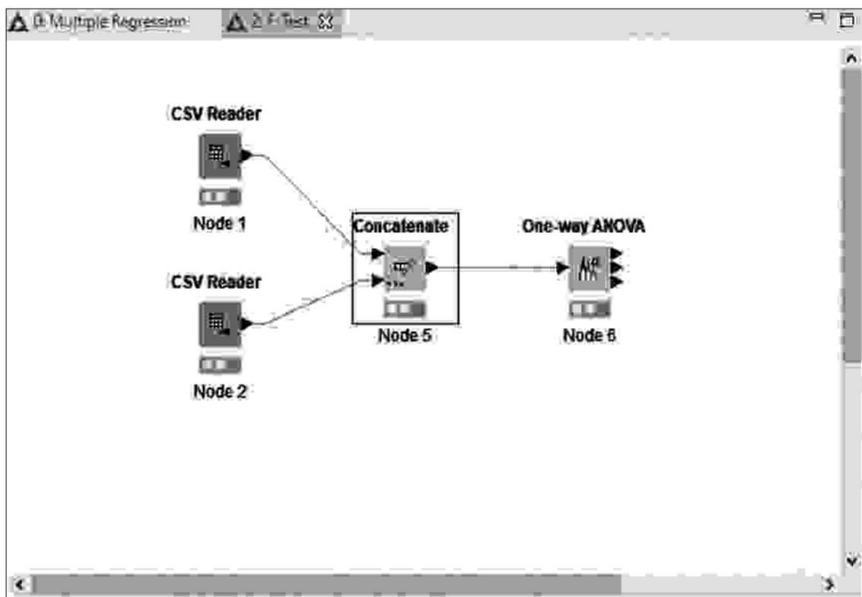
Nhà phân tích sẽ nhận thấy rằng gói RStudio bao gồm giả thuyết thay thế, điều này rất hữu ích. Về cơ bản, điều này có nghĩa là nếu giá trị p nhỏ hơn alpha (như đã được thảo luận, là 0,05), thì giả thuyết vô hiệu có thể bị bác bỏ. Tuy nhiên, trong trường hợp này, giá trị p lớn hơn giá trị alpha, vì vậy giả thuyết không bị bác bỏ, nghĩa là có xác suất thống kê để hai phương sai bằng nhau.

5.2.3 KIẾN THỨC

KNIME có một nút để thực hiện Bài kiểm tra F Levene (thật bất ngờ!), nhưng nó là một phần của một nút khác gọi là ANOVA một chiều, do đó, thực hiện tìm kiếm trên Bài kiểm tra F Levene sẽ không tiết lộ nút thích hợp. Có một số chuẩn bị cần thiết trước khi có thể thực hiện Bài kiểm tra F.

Bước đầu tiên sẽ là nhập hai tệp (theo dõi cơn lốc xoáy năm 1951 và 1954) thông qua nút Trình đọc CSV. Sẽ có hai nút Trình đọc CSV để chứa hai tệp. Sau khi hoàn tất, hãy kéo và kết nối nút Concatenate như minh họa trong màn hình sau. Định cấu hình nút như được hiển thị sau màn hình quy trình làm việc. Có một cái gì đó để nhớ về kết quả sắp được tiết lộ. Chúng có thể khác với các công cụ khác, nhưng đừng lo lắng. Một lần nữa, sự khác biệt thường là do hoạt động bên trong của các công cụ và kết quả sẽ giống với lựa chọn giả thuyết. Ngoài ra, nếu có bất kỳ sự khác biệt nào giữa dữ liệu ban đầu và dữ liệu được chọn để phân tích thì kết quả sẽ có sự khác biệt.

Một khía cạnh của tính nhất quán là hiểu dữ liệu và đảm bảo rằng tất cả các khía cạnh của dữ liệu đều giống nhau cho mọi thử nghiệm. Như trong thử nghiệm, nếu các đối tượng không giống nhau ở một khía cạnh quan trọng đối với bài kiểm tra, thì bài kiểm tra sẽ bị sai lệch.



Levene Test - 26- One-way ANOVA					
File Hilitc Navigation View					
Table "default"- Rows: 1 Spec - Columns: 5 Properties Flow Variables					
Row ID	Test Column	test statistic (Levene)	df,1	p-value	p-value (Levene)
Row0	TOR_LENGTH	0,539	1	876	0,4630543501369...

Điều quan trọng cần nhấn mạnh là kết quả này từ KNIME dẫn đến kết luận giống như các công cụ khác. Vì giá trị p là 0,463 và alpha là 0,05, nên giá trị p lớn hơn alpha, điều này sẽ dẫn đến cùng một kết quả – rằng hai phương sai giữa độ dài cơn lốc xoáy năm 1951 và 1954 có xác suất rất cao là bằng nhau.

5.3 HỒI QUY/TƯƠNG QUAN NHIỀU LẦN

Có những trường hợp khi phân tích yêu cầu một số biến được kiểm tra mối quan hệ. Trong trường hợp này, một số công cụ cung cấp phương thức trực tiếp để thực hiện chức năng này. Tuy nhiên, có một số thận trọng khi sử dụng hồi quy bội. Theo một nguồn, điều quan trọng là phải hiểu được hậu quả của hồi quy bội hoặc tương quan. Một trong số đó được gọi là ghi đè dữ liệu (Reinhart, 2015). Về bản chất, trong bất kỳ tập dữ liệu nào, việc sử dụng nhiều mối tương quan thường có thể dẫn đến ít nhất một biến liên quan đến một biến khác. Méo để làm điều này là đảm bảo rằng nhà phân tích có các yêu cầu trước khi tiến hành thử nghiệm này, do đó giảm thiểu để loại bỏ tình trạng này. Có cả một cuốn sách về các mối tương quan giả mạo, đó là kết quả của việc nhà phân tích tìm kiếm một mối quan hệ thay vì duy trì sự khách quan. Tất cả các nhà phân tích dữ liệu trong tương lai nên đọc cuốn sách này (Vigen, Tyler, Spurious Correlations: Correlation Does not Equal Causation, Hachette Books, New York, 2015.).

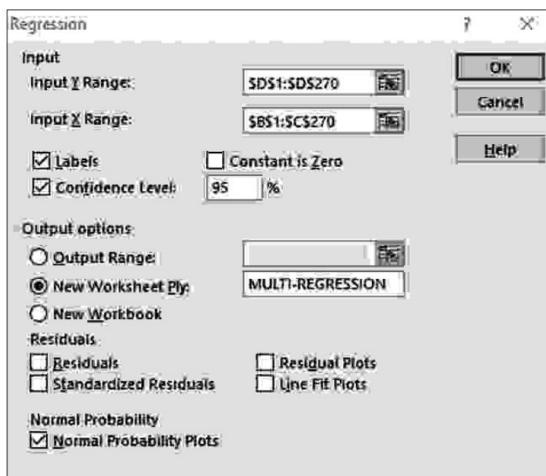
5.3.1 Excel

Để thực hiện hồi quy bội hoặc tương quan trong Excel rất đơn giản với ToolPak Phân tích. Thay vì chỉ chọn một cột cho "Y", hãy chọn nhiều cột. Tuy nhiên, và điều này rất quan trọng, tất cả các cột được chọn phải liền kề nhau. Không thể có một cột giữa những cột mà nhà phân tích muốn kiểm tra. Do đó, nhà phân tích sẽ phải đảm bảo rằng dữ liệu được định dạng và làm sạch đúng cách trước khi tiến hành thử nghiệm này. Quy trình thực hiện

một hồi quy bội trước tiên là xem xét các biến mà nhà phân tích sẽ hồi quy. Trong trường hợp này, nó sẽ là TOR_LENGTH hoặc chiều dài cơn lốc xoáy và BEGIN_DAY và BEGIN_TIME, tương ứng là ngày và giờ khi cơn lốc xoáy xảy ra. Nhà phân tích muốn biết liệu họ có thể dự đoán độ dài cơn lốc xoáy từ ngày và giờ cơn lốc xoáy bắt đầu hay không. Để thực hiện việc này, tệp cơn lốc xoáy năm 1951, sẽ được sử dụng trong trường hợp này, phải được nhập và tệp ToolPak Phân tích phải được tải. Có một mục khác cần được xem xét. Các cột đang được xem xét phải ở cùng nhau, vì vậy bộ phân tích phân tích sẽ phải đảm bảo rằng điều đó được hoàn thành trước khi thực hiện hồi quy bội. Việc xem xét tiếp theo là sử dụng biến nào làm biến phụ thuộc và biến độc lập. Biến phụ thuộc là "y" và biến độc lập là "x". Điều này rất quan trọng vì nó sẽ xác định cột nào sẽ được sử dụng trong hàm.

Sau khi hoàn thành các bước này, nhà phân tích có thể sử dụng Công cụ phân tích Pak và chọn "Hồi quy". Khi đã chọn, màn hình sau sẽ xuất hiện và nó được định cấu hình sao cho "trục y" hoặc biến phụ thuộc là TOR_LENGTH và "trục x" hoặc biến độc lập là BEGIN_TIME và BEGIN_DAY. Kết quả sẽ là cẩm thời gian và ngày và nhận được độ dài cơn lốc xoáy ước tính dựa trên hai biến đó. Xin lưu ý rằng đây là gần đúng và cần được xác thực thông qua sử dụng thực tế.

Tuy nhiên, với mục đích của văn bản này, ví dụ này sẽ hoạt động tốt.



Kết quả là màn hình sau. Phương trình hiển thị phần chặn (y) và hai giá trị x (BEGIN_TIME và BEGIN_DAY). Mặc dù mối quan hệ mong manh nhưng có một công thức khả thi từ hàm này.

Một khía cạnh nữa của hồi quy bội là mối tương quan, rất thấp

đến mức không tồn tại. Nếu nhà phân tích đang cố gắng dự đoán độ dài cơn lốc xoáy từ hai biến độc lập, nó sẽ tạo ra kết quả, nhưng mối liên hệ của hai biến này rất mong manh với độ dài cơn lốc xoáy.

Regression Statistics							
Multiple R	0.061961094						
R Square	0.003839177						
Adjusted R Square	-0.00365075						
Standard Error	10.24951233						
Observations	269						
ANOVA							
	df	SS	MS	F	Significance F		
Regression	2	107.6952964	53.8476	0.51257844	0.599539607		
Residual	266	27943.96582	105.053				
Total	268	28051.66112					
Coefficients							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%
Intercept	4.587879354	2.357912284	1.94574	0.05273944	-0.05466678	9.23042549	0.05466678
BEGIN_DAY	0.062088764	0.074515009	0.8324	0.40545785	-0.0846255	0.20880303	-0.0846255
BEGIN_TIME	-0.000072515	0.00120338	-0.60259	0.0309451	0.00164421	0.00309451	0.00164421

5.3.2 Văn phòng mở

Cũng như các chức năng khác, OpenOffice không có ToolPak Phân tích để tạo ra kết quả một cách hiệu quả, nhưng nó có thể thực hiện nhiều hồi quy dựa trên các công thức. Sau khi nhập dữ liệu giống như được sử dụng với Excel, công thức "linest" được sử dụng với các cột liền kề, sau đó công thức được chuyển đổi thành công thức mảng bằng cách sử dụng CTRL-SHIFT-ENTER, tạo ra kết quả sau:

A	B	C	D	E	F	G	H
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	TOR_LENGTH	TOR_WIDTH		
2	195109	9	915	0.1	100		
3	195106	17	2200	0.7	33	-0.00072515	0.062088764
4	195103	28	510	0.5	17	0.00120338	0.074515009
5	195105	9	1830	0	33	0.003839177	10.24951233#N/A
6	195107	15	1620	0	100	0.512578441	2.357912284#N/A
7	195105	8	1800	0	33	107.6952964	27943.96582#N/A
8	195103	30	1500	0.1	20	#N/A	#N/A
9	195105	11	1330	8	33	#N/A	#N/A
10	195106	27	2204	19.7	33		
11	195107	21	1100	0.1	33		

Cách đọc kết quả này là xem F3 (đã chọn) và nó giống với số trong BEGIN_TIME trong phần đọc Excel. Nói cách khác, nó là một trong những giá trị "x". Giá trị "x" khác, BEGIN_DAY, nằm ở G3 và phần chẵn nằm ở H3. Về cơ bản, điều này có nghĩa là công thức sẽ được đọc như sau:

$$0,00072515x+ .062088764x^2+4,587879354.$$

Điều này có nghĩa là nếu một nhà phân tích muốn biết chiều dài cơn lốc xoáy sẽ là bao nhiêu vào ngày 9 của tháng lúc 09:00, thì nhà phân tích sẽ cắm những con số này vào "x" và "x2" và thêm phần chẵn để có được chiều dài cơn lốc xoáy. Xin nhắc lại, đây chỉ là một ví dụ và không cho thấy mối quan hệ giữa các yếu tố này. Đây là chỉ để trình diễn.

5.3.3 R/RStudio/Rattle

Trong hồi quy bội, Rattle là một lựa chọn tốt cho chức năng này vì nó có sẵn trong gói. Cấu hình giống như trong các phần khác, bước đầu tiên là nhập và chỉ định các biến cù thể theo chỉ định thích hợp. Như được hiển thị trong màn hình sau, phải có một biến "mục tiêu" được chỉ định, nếu không hàm hồi quy sẽ không hoạt động. Trong trường hợp này, hàm mục tiêu sẽ là độ dài cơn lốc xoáy hoặc biến TOR_LENGTH, vì đó là biến sẽ là biến phụ thuộc. Các yếu tố khác, thời gian và ngày, sẽ là các biến độc lập, tương tự như cấu hình trước đó trong OpenOffice. Sau khi hoàn tất, nhấp vào biểu tượng "Execute" và kết quả sau sẽ xuất hiện.

Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1. BEGIN_YEARMONTH	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 12
2. BEGIN_DAY	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 31
3. BEGIN_TIME	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 100
4. END_YEARMONTH	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 12
5. END_DAY	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 31
6. END_TIME	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 100
7. EPISODE_ID	missing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 0 Missing: 205
8. EVENT_ID	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 209
9. STATE	Category	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 35
10. STATE_RPS	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 35
11. YEAR	Constant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1
12. MONTH_NAME	Category	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 12
13. EVENT_TYPE	Category	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1
14. CZ_TYPE	Constant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1

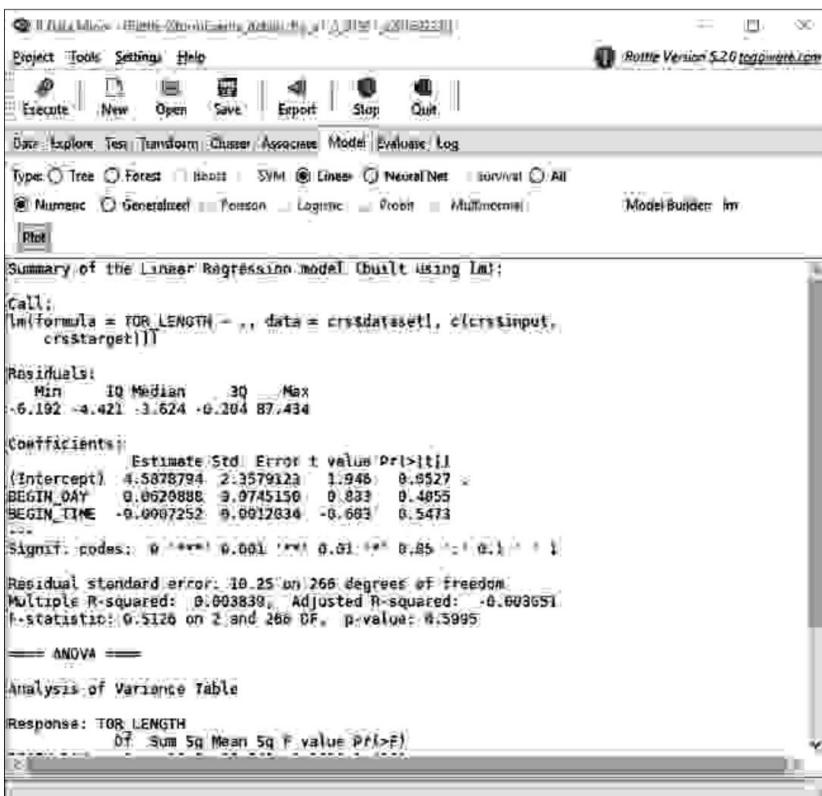
Sau khi dữ liệu được định cấu hình, hãy chuyển đến tab "Mô hình" để thực hiện hồi quy. Màn hình được cấu hình giống như màn hình sau và khi nhấn biểu tượng "Execute", kết quả sẽ hiển thị như hình minh họa.

Toàn bộ màn hình có thẻ hơi choáng ngợp, nhưng kết quả rất giống với các phần trước, trong đó các số đối diện từ BEGIN_

TIME và BEGIN_DAY giống như những giờ được cung cấp bởi các công cụ khác và phản ánh gần nhất kết quả đọc của Excel.

Một lời cảnh báo khi định cấu hình các màn hình này. Nó đã được lưu ý trong các phần trước của Rattle rằng nhà phân tích phải chú ý đến các tùy chọn dữ liệu để đảm bảo rằng tập dữ liệu sẽ bao gồm tất cả các hàng. Ghi nhớ các tùy chọn

"Phân vùng"? Điều này rất quan trọng, vì việc thực hiện bất kỳ chức năng nào trong Rattle sẽ tạo ra các kết quả khác nhau với các cài đặt khác nhau trong tùy chọn Phân vùng. Nếu nhà phân tích chỉ sử dụng Rattle, thì có những bước chuẩn bị không chỉ quan trọng mà còn quan trọng để đảm bảo kết quả nhất quán.



The screenshot shows the Rattle software interface version 5.20. The window title is "Rattle Version 5.20 (rattle.rattle.com)". The menu bar includes Project, Tools, Settings, Help, and a toolbar with Execute, New, Open, Save, Export, Stop, and Quit. Below the toolbar is a navigation bar with tabs: Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, Log. The "Model" tab is selected. A dropdown menu "Type:" is open, showing options: Tree, Forest, Boost, SVM, Linear, Neural Net, Survival, All, Numeric, Generalized, Poisson, Logistic, Probit, Multinomial, and Model Builder: lm. The "Plot" button is also visible. The main pane displays the output of a linear regression model:

```

Summary of the Linear Regression model (built using lm):
call:
lm(formula = TOR_LENGTH ~ ., data = crsdataset1, offset = crsinput,
    crstarget))

Residuals:
    Min      1Q  Median      3Q     Max
-6.192 -4.421 -3.624 -0.204  87.434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.5878794 2.3579123  1.946 0.0527 *
BEGIN_DAY   0.0620888 0.9745150  0.883 0.4855
BEGIN_TIME  -0.0007252 0.0012834 -0.683 0.5473
...
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.25 on 266 degrees of freedom
Multiple R-squared:  0.003839, Adjusted R-squared:  -0.003651
F-statistic: 0.5126 on 2 and 266 DF, p-value: 0.5995

==== ANOVA ====
Analysis of Variance Table

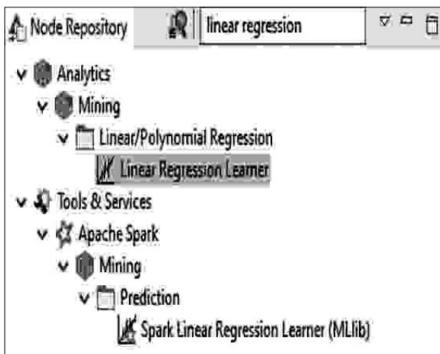
Response: TOR_LENGTH
          Df Sum Sq Mean Sq F value Pr(>F)

```

5.3.4 KIẾN THỨC

Nút hồi quy bội nằm trong KNIME không hiển thị ngay lập tức.

Nút nằm ở vị trí trong màn hình này. Một lần nữa, đặt "hồi quy" trong khối tìm kiếm sẽ xác định vị trí của nút như sau:



Hồi quy bội KNIME tương đối đơn giản. Nút Trình học hồi quy tuyến tính được đặt và kết nối với nút Trình đọc CSV.

Bằng cách nhấp đúp vào nút Trình học hồi quy tuyến tính, một màn hình sẽ xuất hiện cho phép người dùng chọn cột nào là cột mục tiêu và cột nào là cột "độc lập". Điều này được thực hiện giống như các nút khác thuộc loại này.

Khi nút được định cấu hình, hãy thực thi cả hai nút và sau khi người dùng nhận được "đèn xanh", sau đó nhấp chuột phải vào nút Trình học hồi quy tuyến tính và chọn bảng hệ số và thống kê để xem kết quả. Người dùng nên sử dụng cùng một phương pháp đọc các kết quả này như được mô tả trong các phần trước.

Điều quan trọng cần nhớ là mọi nút mà nhà phân tích cần có thẻ có sẵn trong KNIME, nhưng đôi khi phải mất một số lần tìm kiếm để có được các nút đó. Một lưu ý bổ sung là có rất nhiều thông tin liên lạc từ cộng đồng nếu nhà phân tích cần hỗ trợ trong KNIME. Trong một số trường hợp, các cộng đồng này bao gồm các quy trình thực tế đã hoàn tất và có sẵn để tải xuống nhằm kiểm tra và xem quy trình hoạt động như thế nào. Nghiên cứu các nút này là thực tế và góp phần vào quá trình học tập với các công cụ này. Cùng với nghiên cứu là thực hành, điều này vô cùng quý giá.

5.4 ĐỊNH LUẬT BENFORD

Định luật Benford được phát triển để phát hiện sự bất thường trong dữ liệu số, cụ thể là đầu vào kế toán. Lý thuyết cơ bản rằng sau nó là những con số được “phân phối bình thường” phản ánh một đường cong giảm dần từ “1” đến “9”. Bằng cách áp dụng Định luật Benford, nhà phân tích có thể phát hiện xem có sự bất thường nào về số lượng hay không, điều này có thể dẫn đến việc tiết lộ các nội dung gian lận. Điều này được các kế toán viên và nhà phân tích tài chính sử dụng để giúp hạn chế gian lận và các vấn đề về kế toán (Statistical Consultants Limited, 2011). Một công cụ đường như thực hiện điều này mà không tốn nhiều công sức là Rattle, vì nó có nó như một tùy chọn trong các chức năng được cung cấp cùng với công cụ. Một lưu ý thận trọng ở đây là Rattle (và R nói chung) hoạt động trên các gói, điều đó có nghĩa là đôi khi sẽ cần một gói để hoàn thành một chức năng trong Rattle. Khi điều này xảy ra, Rattle sẽ nói với nhà phân tích rằng cần có một gói và hỏi xem nhà phân tích có muốn cài đặt gói đó không. Nếu nhà phân tích chọn tùy chọn không, gói sẽ không được cài đặt và quá trình sẽ kết thúc. Một điểm về các nhà phân tích phụ thuộc vào Rattle là họ tin tưởng công cụ này để tải xuống các mục và không cài đặt phần mềm độc hại hoặc các tệp ngốn bộ nhớ khác. Trong những năm sử dụng Rattle, điều này chưa xảy ra với tác giả này, nhưng không có gì đảm bảo về vấn đề an toàn 100% với công cụ này. Tuy nhiên, điều tương tự cũng xảy ra đôi với các công cụ nổi tiếng hơn được các công ty và chính phủ liên bang tin tưởng và sử dụng, những công cụ đã cài đặt các tệp mà tin tặc sử dụng cho phần mềm độc hại. Trong mọi trường hợp, bất kỳ nhà khoa học dữ liệu nào cũng nên kích hoạt và tiếp tục bất kỳ phần mềm chống vi-rút nào mà họ đã cài đặt trên máy tính của mình.

5.4.1 Tiếng lách cách

Cấu hình của Benford's Law for Rattle hơi phức tạp, nhưng nó vẫn rất mạnh so với việc sử dụng các công thức trong một số công cụ khác. Rattle đặt Định luật Benford trong tab “Khám phá” bên dưới “Phân phối”. Khi nhà phân tích chọn nút radio Phân phối, màn hình sau sẽ hiển thị và việc chọn Định luật Benford dễ dàng như đánh dấu vào ô. Tuy nhiên, có một số chuẩn bị đi kèm với việc kiểm tra đó.

Trước tiên, hãy quay lại tab “Dữ liệu” và đảm bảo rằng TOR_LENGTH được chọn làm “Mục tiêu”, vì đó là biến mà nhà phân tích muốn sử dụng để xem nó có tuân theo Định luật Benford hay không. Đảm bảo rằng, sau khi TOR_LENGTH được chọn, nhà phân tích nhấp vào biểu tượng “Thực thi” để kích hoạt biểu tượng đó trong tập dữ liệu. Cũng nên nhớ rằng không cần phải “bỏ qua” tất cả các biến khác, vì mục tiêu là biến sẽ là biến chính được xem xét trong phần

nút Định luật Benford. Xin lưu ý rằng tab "Khám phá" có rất nhiều chức năng có sẵn để phân tích dữ liệu, vì vậy vui lòng thử các cách kết hợp khác nhau này để xem liệu có chức năng nào phù hợp với nhu cầu của nhà phân tích hay không.

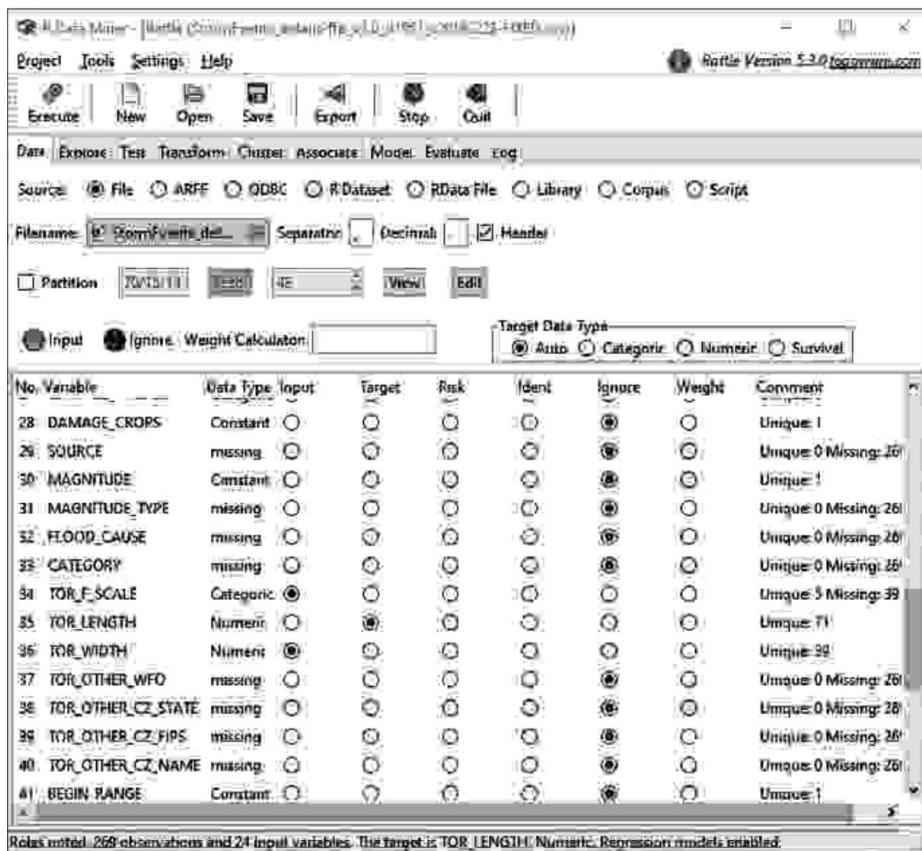
Sau khi dữ liệu được chọn phù hợp như minh họa trong phần sau, bước tiếp theo sẽ là đảm bảo rằng cấu hình của tab "Khám phá" là chính xác để chức năng hoạt động. Một lời cảnh báo ở đây là đảm bảo rằng tập dữ liệu phù hợp với chức năng. Khi nhà phân tích chọn dữ liệu, có thể có xu hướng sử dụng tùy chọn "Bộ dữ liệu R" trong tab "Dữ liệu". Mặc dù điều này sẽ tốt cho nhiều chức năng, nhưng tùy chọn Định luật Benford cần phải đọc một "khung dữ liệu", đây không phải là loại tập dữ liệu do lựa chọn "Bộ dữ liệu R". Tập dữ liệu do lựa chọn đó tạo ra được gọi là "tibble", đây là một loại tập dữ liệu rất linh hoạt với nhiều gói có sẵn trong R và Rattle. Tuy nhiên, nó không tương thích với chức năng Định luật Benford. Do đó, thay vì sử dụng tùy chọn "Bộ dữ liệu R", tốt nhất bạn nên chọn "Tệp" làm nguồn. Theo cách này, bằng cách sử dụng tệp trực tiếp từ máy tính, không có sự chuyển đổi từ R để biến nó thành tibble.

Câu lệnh trướcc chỉ là một gợi ý, vì có những lệnh có thể thay đổi tibble thành "khung dữ liệu" ngay trong R; nhưng nếu lập trình không thích hợp hơn, thì nhập tệp máy tính thông thường là tùy chọn phù hợp.

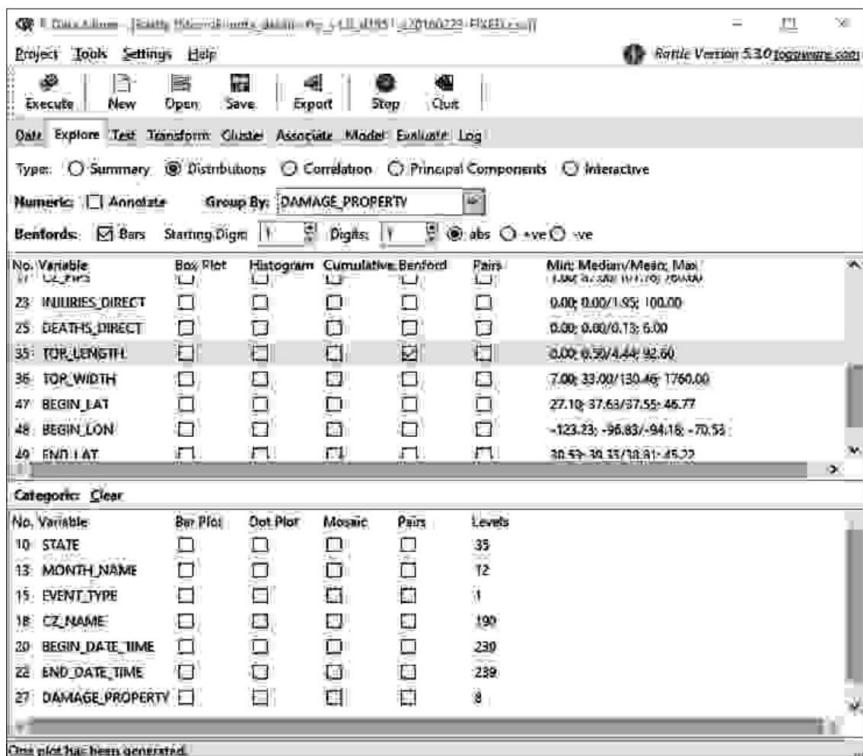
Khi dữ liệu đã được chọn và nhập, cần phải biến một biến thành "mục tiêu" và trong trường hợp này TOR_LENGTH đã được chọn. Điều này sẽ đảm bảo rằng chức năng thích hợp sẽ nhận ra và tập trung vào TOR_LENGTH là yếu tố được xem xét. Màn hình sau hiển thị các lựa chọn thích hợp. Một lưu ý là hộp kiểm "phân vùng" không được chọn.

Trong trường hợp này, tất cả các hàng sẽ được xem xét trong chức năng này, nhưng như trong phần trước về tập dữ liệu "đào tạo", nhà phân tích có thể chọn xác định tỷ lệ phần trăm cần thiết (lấy mẫu) để thực hiện kiểm tra và sau đó xác thực nó bằng một phần khác của toàn bộ tập dữ liệu.

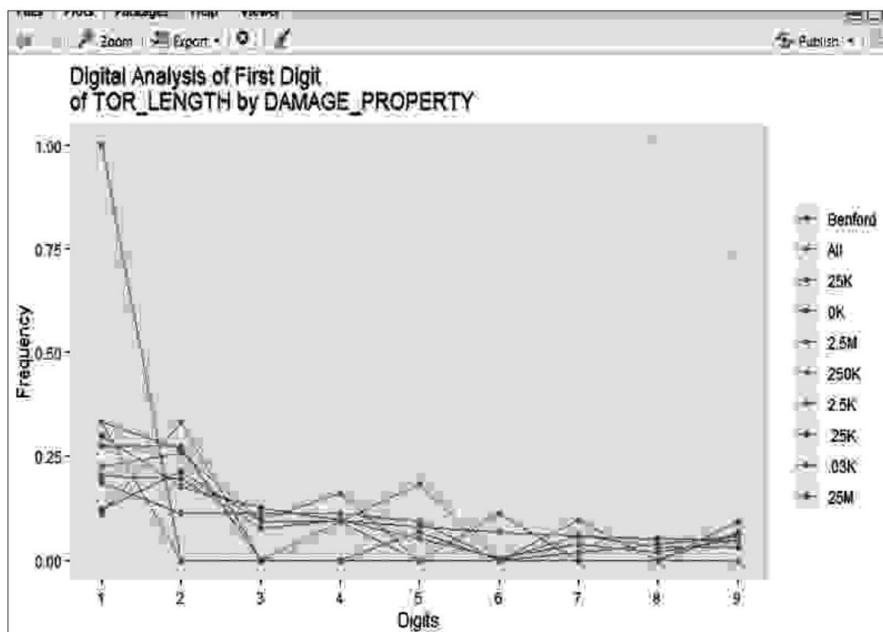
Khi tập dữ liệu được nhập và dữ liệu được định cấu hình, đã đến lúc chuyển đến tab "Khám phá" và xử lý các lựa chọn cần thiết để kích hoạt chức năng Định luật Benford trong Rattle.



Bước đầu tiên là chọn “Phân phối” từ các tùy chọn và một cửa sổ có hai màn hình sẽ xuất hiện. Hiện tại, phần trên cùng sẽ là trọng tâm của phân này. Chọn một biến cho tùy chọn Định luật Benford và sau đó đảm bảo rằng lựa chọn “nhóm theo” là một biến thể hiện mối liên hệ hợp lệ. Trong trường hợp này, DAMAGE_PROPERTY được hiển thị dưới dạng biến phân loại. Các lựa chọn sẽ xuất hiện như sau:



Sau khi các lựa chọn được thực hiện, hãy nhấp vào biểu tượng Thực thi và kiểm tra màn hình biểu đồ RStudio, theo mặc định, nằm ở phần tư dưới cùng bên phải của màn hình. Ở đó, nhà phân tích sẽ thấy màn hình sau đây, như một lời cảnh báo, có vẻ rất phức tạp. Dòng chính mà nhà phân tích nên tập trung vào là dòng được đánh dấu "Benford", cho biết xác suất các chữ số đầu tiên sẽ xuất hiện trong dữ liệu bình thường. Chẳng hạn, nhìn vào dòng Benford (màu đỏ), chữ số "1" xuất hiện khoảng 0,30 hoặc 30% thời gian. Nếu nhà phân tích đang nhìn vào chấm "đỏ" xuất hiện ở đầu biểu đồ, thì đây không phải là đường Benford mà là đường mô tả số chữ số "1" xuất hiện trong TOR_LENGTH khi giải quyết 25 triệu hoặc 25 triệu đô la hư hại. Tuy nhiên, với các số liệu DAMAGE_PROPERTY khác, chẳng hạn như 2,5M hoặc 25K, đường này hơi gần với đường Benford. Điều này có nghĩa là có một số điểm tương đồng giữa những số liệu đó và tính quy phạm của Luật Benford.



Tuy nhiên, mặc dù biểu đồ có thể không phân biệt đối xử, vẫn có một chương trình trong R/RStudio đưa ra đánh giá về việc liệu Định luật Benford có được tuân theo bởi dữ liệu hay không. Dòng lập trình cho điều này được hiển thị như sau khi nó xuất hiện trong R:

```
> benford(tor1951$TOR_LENGTH,number.of.digits=1,sign=
  "dương",rặc=SAI,vòng=3)
```

Từ dòng này, kết quả sau sẽ xuất hiện. Như nhà phân tích có thể thấy, nó đánh giá dữ liệu là không phù hợp với Định luật Benford, nhưng cuối cùng xác nhận tuyên bố đó bằng tuyên bố rằng không có dữ liệu trong thế giới thực nào sẽ hoàn toàn tuân theo Định luật Benford. Điều này rất quan trọng, vì việc phản ánh dữ liệu thực tế vào một lý thuyết không phải là một kết quả thực tế.

Đối tượng Benford:

Dữ liệu: tor1951\$TOR_LENGTH
 Số lượng quan sát được sử dụng = 157
 Số quan sát. cho thứ tự thứ hai = 69
 Chữ số đầu tiên được phân tích = 1

bộ ngựa:

Giá trị thống kê
trung bình 0,476
Biến 0,089
Ex.Kurtosis -1.098
Độ lệch -0,112

5 sai lệch lớn nhất:

chữ số tuyệt đối.diff
1 5 13,57
2 6 9,51
3 1 9,26
4 2 8,35
5 7 2.10

Thống kê:

Pearson's Chi-squared test

dữ liệu: tor1951\$TOR_LENGTH
X bình phương = 28,47, df = 8, giá trị p = 0,0003927

Kiểm tra vòng cung Mantissa

dữ liệu: tor1951\$TOR_LENGTH
L2 = 0,0039588, df = 2, giá trị p = 0,5371

Độ lệch tuyệt đối trung bình (MAD): 0,03218442
MAD Conformity - Nigrini (2012): Sự Không Phù Hợp
Hệ số biến dạng: -34.24091

Hãy nhớ rằng: Dữ liệu thực sẽ không bao giờ phù hợp hoàn hảo với
Định luật Benford. Bạn không nên tập trung vào giá trị p!

Điều này cho thấy chức năng này có thể được thực hiện tương đối dễ dàng
từ công cụ này mà không cần lập trình bổ sung. Một nhận xét bổ sung là, để
tính năng này hoạt động, nhà phân tích có thể phải cài đặt "benford.analysis"

gói là một phần của R nhưng không được cài đặt tự động với R hoặc RStudio cơ sở.

5.5 NÂNG

Thang máy là một phương pháp đánh giá một mô hình dự đoán. Nhiều lần, nhà phân tích sẽ tiến hành một mô hình hoặc thử nghiệm mà không đánh giá giá trị tiềm năng của thử nghiệm đó. Trong trường hợp này, Lift có thể đánh giá giá trị dự đoán trước khi chạy thử nghiệm thực tế. Hãy nghĩ về giá trị của một chức năng như vậy. Nếu thử nghiệm không có giá trị hoặc không phù hợp, tại sao lại chạy thử nghiệm? Trong cuốn sách Data Science for Business (được tìm thấy trong phần tham khảo), các tác giả trình bày một ví dụ về việc ai đó đi vào cửa hàng và mua kết hợp các sản phẩm. Mức tăng sẽ xác định tính khả thi của việc dự đoán liệu một người nào đó có mua sự kết hợp cụ thể của các sản phẩm đó hay không, cho dù đó là bia và trứng hay bia và khoai tây chiên. Điều này dựa trên xác suất mua một sản phẩm và sau đó là sản phẩm khác.

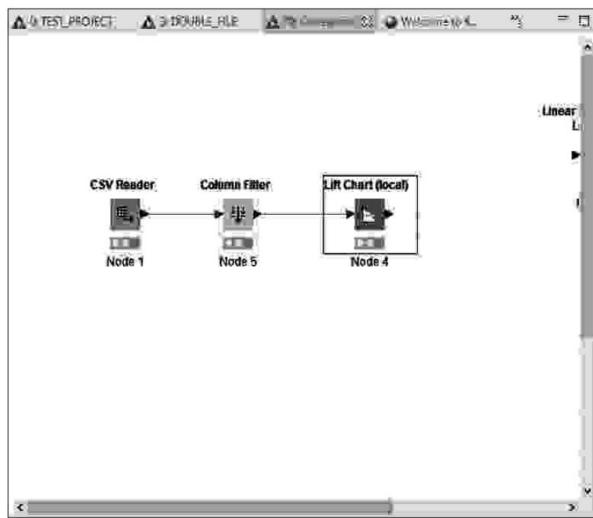
Công thức này được bao gồm trong cuốn sách mà tác giả này đặc biệt khuyên mọi nhà phân tích nên đọc, đặc biệt nếu họ là thành viên của một công ty lớn tạo ra doanh thu từ việc sản xuất và bán các sản phẩm, cụ thể là hàng tiêu dùng (Provost, 2013). Điều này đã làm được là cho thấy tính khả thi của việc dự đoán việc mua kết hợp các sản phẩm. Công cụ tốt nhất trong số những công cụ được mô tả trong văn bản này để thực hiện chức năng nâng là KNIME, vì nó có một nút để thực hiện phương pháp này.

5.5.1 KIẾN THỨC

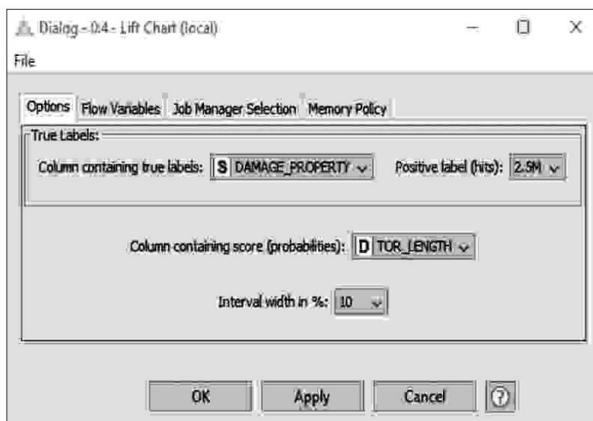
Cũng như nhiều chức năng khác, KNIME có một nút để tính toán mức tăng mà nhà phân tích có thể tìm thấy thông qua thanh tìm kiếm như trong màn hình sau.

Hãy nhớ rằng nhà phân tích không phải nhập toàn bộ tên nút, vì KNIME sẽ tìm kiếm khi nhà phân tích đang nhập.

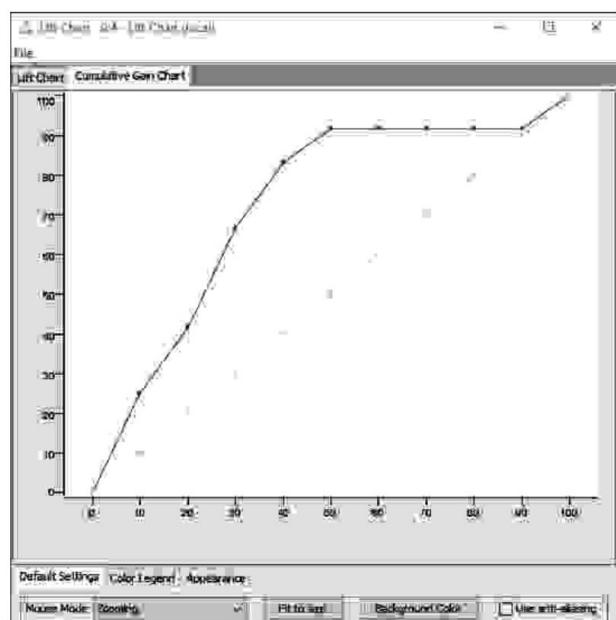
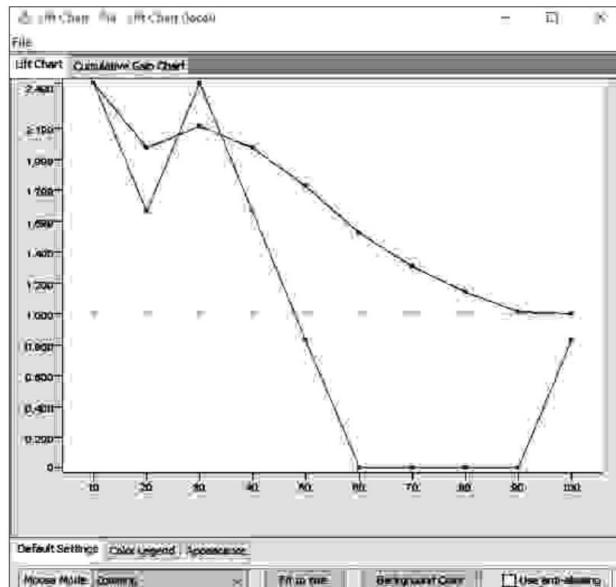
Sau khi dữ liệu được nhập thông qua Trình đọc CSV và Biểu đồ mức tăng được tìm thấy, kéo, đặt và kết nối với Trình đọc CSV, bước tiếp theo là định cấu hình nút Biểu đồ mức tăng. Màn hình sau đây là một cấu hình của màn hình này hiển thị TOR_LENGTH với DAMAGE_PROPERTY làm giá trị dự đoán. Điều này có nghĩa là phương pháp này đang đánh giá khả năng dự đoán chiều dài cơn lốc xoáy từ những thiệt hại do cơn lốc xoáy đó gây ra cho tài sản.



Màn hình sau đây hiển thị cấu hình cho nút Biểu đồ thang máy (cục bộ). Như nhà phân tích có thể thấy, DAMAGE_PROPERTY sẽ được đặt theo TOR_LENGTH để xem liệu điều này có tạo ra một mô hình dự đoán tốt hay không. Người dùng phải đặt "Nhận tích cực (lượt truy cập)" cụ thể cho danh mục 2,5 triệu (hoặc 2,5 triệu đô la thiệt hại) để xem liệu có đáng để có một mô hình dự đoán dựa trên con số này hay không. Nhà phân tích có thể sử dụng mũi tên xuống để chọn các mức độ thiệt hại khác, nhưng mức độ thiệt hại này phải được dự đoán dựa trên độ dài cơn lốc xoáy, cho thấy mối liên hệ giữa thiệt hại và độ dài. Kết quả, sau khi được thực thi, nằm bên dưới màn hình này. Nút này cho phép cả biểu đồ tăng và biểu đồ tăng tích lũy, cả hai đều hữu ích cho nhà phân tích. Biểu đồ thang máy cho thấy khoảng cách lớn giữa phép đo (màu đỏ) và đường cơ sở (màu xanh lá cây), đây là một chỉ số dự đoán tốt. Ngoài ra, biểu đồ mức tăng tích lũy cho thấy đường tăng trên đường cơ sở trong suốt biểu đồ, đây cũng là một chỉ báo dự đoán tốt.



Các màn hình sau đây là kết quả của chức năng nâng, như đã nêu trước đó. Xin lưu ý rằng có một số tùy chọn định dạng bao gồm màu chủ thích và các tùy chọn cụ thể khác. Vui lòng khám phá những điều này vì việc hiểu các tùy chọn này có thể thay đổi giao diện của biểu đồ và nâng cao trải nghiệm của nhà phân tích và người nhận với công cụ này luôn hữu ích như thế nào.



Sau khi xem xét các biểu đồ này, nhà phân tích sẽ biết liệu có đáng để xử lý một mô hình dự đoán dựa trên hai biến này hay không. Một lưu ý thận trọng vào thời điểm này. Nhà phân tích chỉ chọn một loại thiệt hại. Có thể đáng để kiểm tra các lượng thiệt hại khác để xem liệu có bất kỳ cách sử dụng nào để liên kết hai yếu tố này trong một mô hình hay không.

Ngoài ra, nhà phân tích có thể nhận thấy rằng có một nút ở giữa nút Trình đọc CSV và nút Biểu đồ thang máy. Nút đó đã được đặt trong quy trình để giới hạn số lượng cột được sử dụng cho Biểu đồ mức tăng. Điều quan trọng là chỉ những yếu tố đang được xem xét mới có sẵn; mặt khác, có một khả năng nhỏ là KNIME có thể cố gắng đưa các yếu tố khác vào hỗn hợp, nghĩ rằng các đặc điểm phân loại là mở. Điều này đã xảy ra với một số nút, nhưng cách để khắc phục điều đó là chỉ sử dụng những cột áp dụng hoặc dính vào các nút được đánh dấu trong ngoặc đơn (cyc bộ), vì đó là những nút đường như cung cấp những điều cơ bản nhưng cũng có vẻ là nhiều nhất ổn định. Tất nhiên, đây là ý kiến của tác giả này.

5.6 WORDCLOUD

Đôi khi, nhà phân tích nhận được dữ liệu toàn là văn bản và tự hỏi làm thế nào để chuyển từ dạng từ thành số để phân tích. Thật may mắn là nhà phân tích không còn phải bận tâm đến sự biến đổi này nữa. Nhờ các thuật toán và nghiên cứu của các nhà thống kê và phân tích khác, giờ đây có chức năng lấy các từ và phân tích các từ đó để tìm ra từ được sử dụng nhiều nhất và ít nhất.

Mặc dù điều này có vẻ chiêu lệ, nhưng chức năng này cung cấp cho người phân tích và người nhận dữ liệu được phân tích hình ảnh một màn hình của "kho văn bản" hoặc các từ trong văn bản. Đây là điều sẽ được thảo luận tiếp theo, cụ thể là với các công cụ cho phép hoàn thành phân tích này với ít bước lập trình hoặc lập trình gian khổ nhất.

5.6.1 R/RStudio

Sự kết hợp R/RStudio cho phép phương pháp nhanh nhất để phân tích các từ trong văn bản. Trong trường hợp này, dữ liệu sẽ khác một chút so với các phần trước.

Dữ liệu được nhập sẽ từ dữ liệu (chi tiết) theo dõi cơn lốc xoáy năm 1995 từ liên kết được mô tả trong một số phần đầu tiên của cuốn sách. Sau khi dữ liệu được mở, hãy xóa tất cả các cột ngoại trừ cột được đánh dấu là "Tường thuật sự kiện". Nhà phân tích sẽ sử dụng dữ liệu này dưới dạng văn bản để trích xuất các từ có thể đưa ra một số mẫu có giá trị cho quá trình phân tích.

Xin lưu ý thêm, có một số gói cần thiết để wordcloud hoạt động. Một số được cài đặt như một phần của gói wordcloud, nhưng bạn có thể cần cài đặt gói "tm" và "RColorBrewer" để tạo hình ảnh màu. Nếu một nhà phân tích muốn xem wordcloud trông như thế nào, thì có rất nhiều trang web có sẵn để xem một hoặc thực sự xem những trang nhỏ miễn phí. Chỉ cần đặt "wordcloud" vào công cụ tìm kiếm và có rất nhiều ví dụ có sẵn để xem. Ngoài ra còn có một ví dụ ở cuối phần này.

Sau khi tường thuật được tách biệt, hãy sao chép tất cả văn bản và lưu văn bản dưới dạng tệp ".txt" để giảm bớt mọi ký tự không liên quan có thể được đưa vào như một phần của câu hình xử lý văn bản. Tệp ".txt" tương đối đơn giản và có rất ít ký tự không liên quan để làm mờ phân tích.

Đặt tên tệp là "textanalysis.txt" để đơn giản hóa việc nhận dạng và nhập tệp vào R/RStudio. Tuy nhiên, lần này, để nhập tệp dưới dạng tệp văn bản sẽ cần một số lệnh lập trình để quá trình nhập có thể hoạt động. Các lệnh sau được lấy từ một trang web phân tích chuyên về các hàm R (Sankhar, 2018).

```
> thư viện (tm)
> thư viện (RColorBrewer)
> setwd("C:/")
> speech="CORPUS/textanalysis.txt"
> thư viện (wordcloud)
> speech_clean<-readLines(bài phát biểu)
> wordcloud(speech_clean)
```

Cần phải lập trình trước đó để chức năng wordcloud hoạt động bình thường. Hai dòng đầu tiên tải các gói R giúp làm sạch văn bản và nâng cao gói wordcloud. Dòng thứ ba đặt thư mục làm việc để toàn bộ vị trí tệp (có thể là một dòng dài) có thể được rút ngắn. Dòng thứ tư đặt biến "lời nói" thành tệp văn bản. Dòng thứ năm mở gói wordcloud, trong khi dòng thứ sáu sử dụng chức năng "readLines" để đọc từng dòng của văn bản "bài phát biểu" và sử dụng biến "speech_clean" có thể thay đổi để lưu trữ kết quả. Dòng cuối cùng kích hoạt chức năng wordcloud trên văn bản cuối cùng. Kết quả của dòng cuối cùng được minh họa như sau:

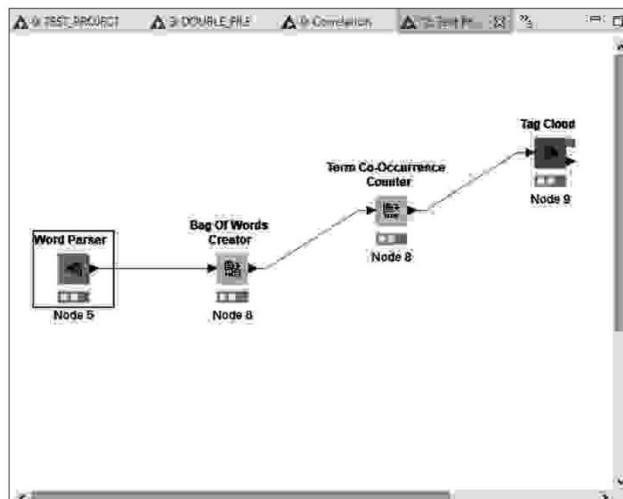


Wordcloud trước đó hiển thị những từ xuất hiện nhiều nhất trong câu chuyện dưới dạng lớn nhất và những từ xuất hiện ít hơn dưới dạng văn bản ngày càng nhỏ hơn.

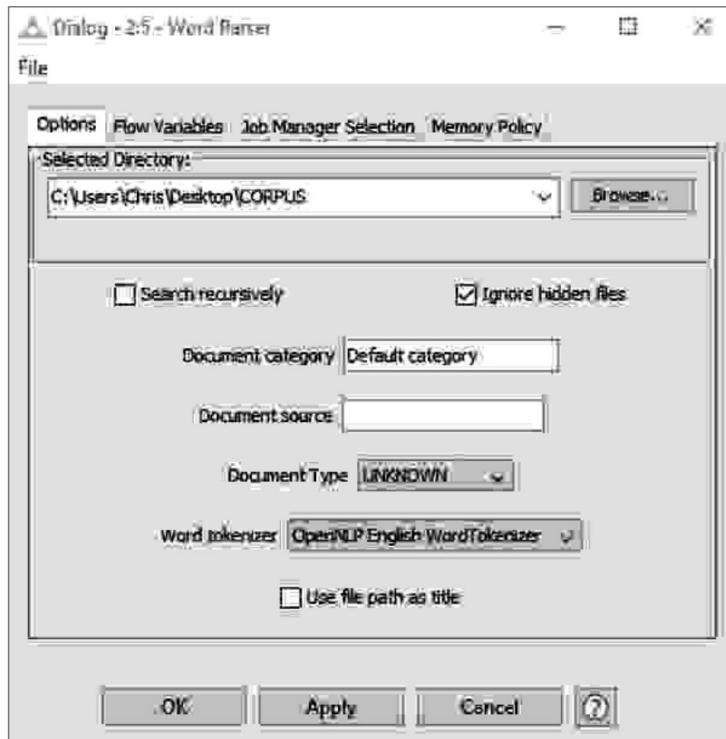
Từ đó, có vẻ như "thạch cao tuyết hoa" xuất hiện nhiều hơn là "winston" hoặc "scottsdale." Điều này có thể có nghĩa là gì, vì đây là tên của các quận, nên một quận có thể hứng chịu những cơn bão nghiêm trọng hơn quận khác. Điều này có thể được thực hiện mà không cần bất kỳ phân tích số nào. Điều này không có nghĩa là điều này có thể trả lời tất cả các câu hỏi mà nhà phân tích có thể có, nhưng nó có thể làm sáng tỏ một bộ dữ liệu khó hiểu trong một trường hợp khác. Sử dụng phân tích văn bản có thể giúp giảm bớt một số nhầm lẫn về dữ liệu.

5.6.2 KIẾN THỨC

KNIME cũng có chức năng wordcloud trong thư viện nút có tên là "đám mây thẻ", được minh họa trong màn hình quy trình làm việc. Lưu ý các nút khác nhau, vì mỗi nút sẽ được mô tả khi nó xuất hiện từ trái sang phải trong quy trình làm việc.

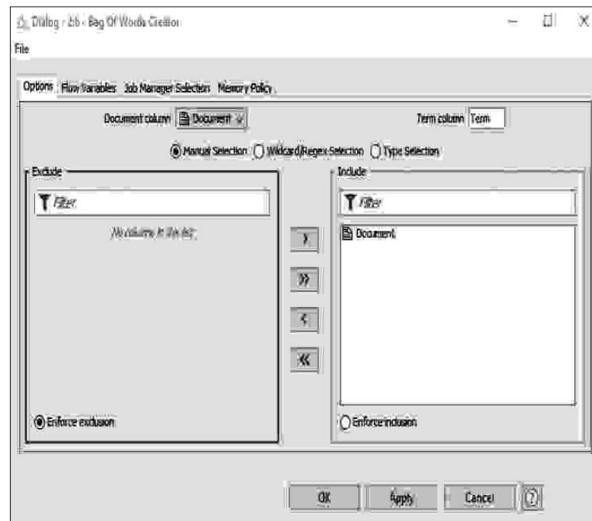


Nút đầu tiên sẽ là nút “Trình phân tích cú pháp từ”, lấy các tài liệu Microsoft Word và chuẩn bị chúng để phân tích văn bản. Cấu hình cho màn hình này như sau và hiển thị vị trí thư mục sẽ được phân tích. Một cảnh báo là đây không phải là tệp mà là thư mục. Nút sẽ tìm kiếm thư mục cho các tệp Word và sử dụng chúng để phân tích.



Lựa chọn “Loại tài liệu” không xác định, nhưng có một số lựa chọn cho mũi tên xuống này, bao gồm sách và tiền trình. Việc họ chọn phân tích cái nào là tùy thuộc vào nhà phân tích. Lựa chọn chưa biết rất phù hợp với ví dụ này. “Word Tokenizer” là mặc định và một lần nữa, nhà phân tích có thể chọn giữa một số loại chức năng phân tích cú pháp này. Sẽ có lợi cho nhà phân tích nếu thử một số trong số này để xem liệu chúng có tạo ra sự khác biệt trong khai thác văn bản hay không. Đối với ví dụ này, mặc định được chọn.

Nút tiếp theo trong quy trình làm việc là nút “Bag of Words Creator”, nút này chia văn bản thành các từ và sử dụng các chỉ số bằng số để tổng hợp số lần từ đó xuất hiện trong một câu hoặc một nhóm câu. Để thấy rõ điều này, màn hình cấu hình như sau (tab đầu tiên).

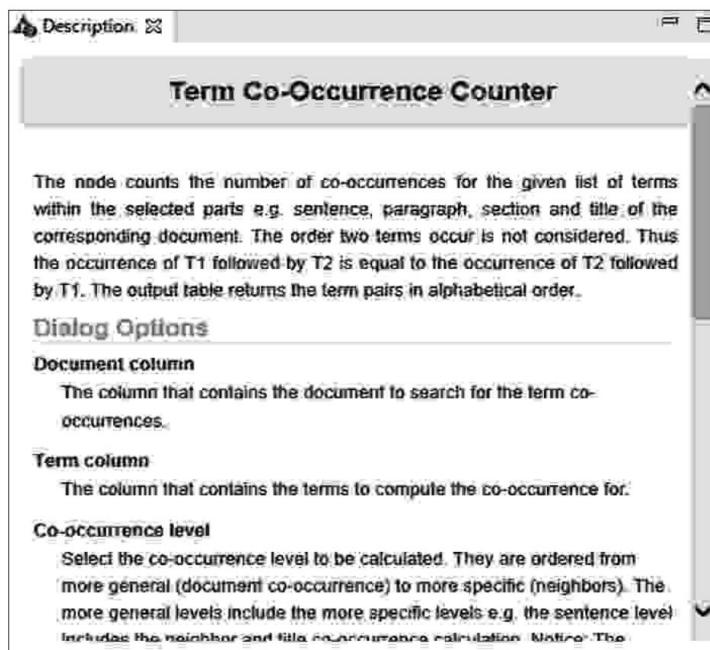


Chỉ có một cột được đưa vào phân tích có tên là "Tài liệu" và cột "Thuật ngữ" được đặt tên là "Thuật ngữ" theo mặc định. Điều này rất quan trọng vì các nút tiếp theo sẽ dựa vào kết quả của nút này để phân tích thêm. Sau khi hoàn thành, kết quả của nút đó trông giống như sau:

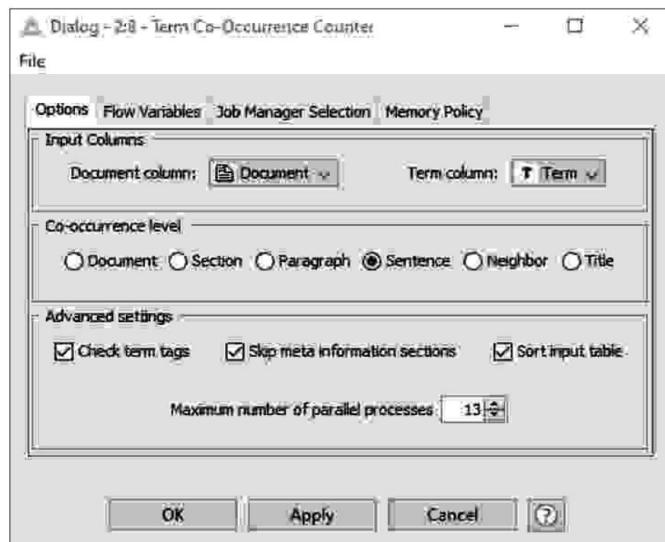
Row ID	Document	Term
Row0	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1]	[1]
Row1	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [Pinoo1]	[Pinoo1]
Row2	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [03..1638CS...]	[03..1638CS...]
Row3	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1..1649CST]	[1..1649CST]
Row4	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1]	[1]
Row5	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1]	[1]
Row6	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1]	[1]
Row7	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [Hall]	[Hall]
Row8	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [0]	[0]
Row9	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [1..75]	[1..75]
Row10	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [0]	[0]
Row11	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [Three-quar...	[Three-quar...
Row12	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [inch]	[inch]
Row13	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [hail]	[hail]
Row14	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [hail]	[hail]
Row15	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [reported]	[reported]
Row16	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [at]	[at]
Row17	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [Ommoor]	[Ommoor]
Row18	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [valley]	[valley]
Row19	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [golf]	[golf]
Row20	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [club]	[club]
Row21	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [in]	[in]
Row22	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [Homewood]	[Homewood]
Row23	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [0]	[0]
Row24	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [One]	[One]
Row25	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [the]	[the]
Row26	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [trace]	[trace]
Row27	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [crossings]	[crossings]
Row28	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [area]	[area]
Row29	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [off]	[off]
Row30	Pinson_03_1638CST_*,1649CST_*,?/?/?Hall (1..75) Three-quarters inch hail was reported at Ommoor Valley Golf Club in Homewood... [changes]	[changes]

Nút đã tách các câu theo từng từ và đặt từ đó thành một hàng. Điều này sẽ mất nhiều giờ nếu được thực hiện bằng tay, nhưng chức năng này làm cho nó có vẻ dễ dàng. Nút tiếp theo sẽ lấy những từ này và đếm số lần chúng xuất hiện với một từ khác trong văn bản. Điều này được sử dụng cho một số phân tích nâng cao, chẳng hạn như cách các từ được sử dụng trong sự kết hợp của các từ khác và như vậy.

Nút này được gọi là “Bộ đếm cùng xuất hiện thuật ngữ” và mô tả về nút này, cũng như với tất cả các nút, xuất hiện khi nhà phân tích nhấp vào nút, như minh họa trong phần sau. Thông thường những mô tả này là đủ để nhà phân tích biết liệu nút đó có hữu ích trong quy trình làm việc hay không hoặc thứ gì đó có thể hữu ích trong các quy trình sau này. Đây là một phần của KNIME làm cho nó rất thân thiện với nhà phân tích, trong đó mọi nút đều được kèm theo một mô tả.



Màn hình cấu hình cho nút được mô tả như sau. Nhà phân tích có thể chọn một số tùy chọn cho nút này và những tùy chọn được chọn cho ví dụ này hoạt động tốt với tập dữ liệu.

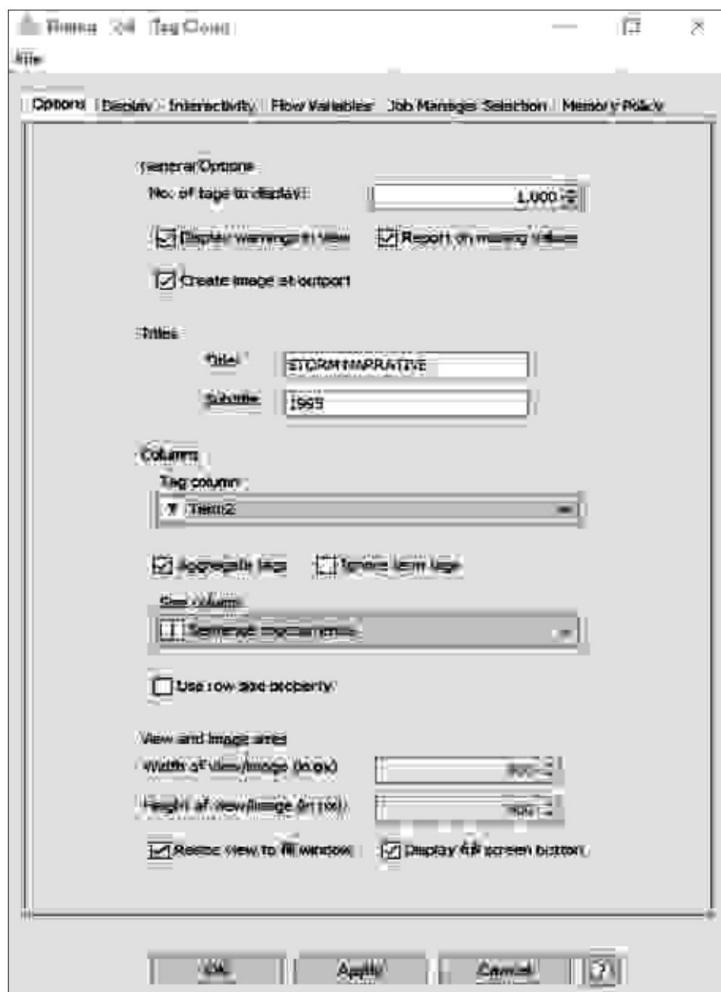


Lưu ý rằng “Mức độ cùng xuất hiện” được đặt ở “Câu”, nhưng có các lựa chọn khác mà phân tích có thể chọn, vì vậy vui lòng khám phá các nút này để xem liệu có sự kết hợp nào cung cấp kết quả hoàn chỉnh cần thiết cho phân tích hay không. Kết quả từ nút này là sự xuất hiện của từ này với từ khác trong câu xuyên suốt văn bản. Bảng được hiển thị như sau:

Row ID		▼ Term 1	▼ Term 2	▼ Sentence	▼ Neighbor	▼ This col...
Row0	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row1	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row2	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	*,1645CST[]	,1645CST[]	2	1
Row3	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	*,1645CST[]	1645CST[]	2	1
Row4	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row5	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	6	12	3
Row6	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row7	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row8	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row9	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	1,10]	2	1
Row10	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Floor"]	2	2	1
Row11	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	Three-quart...	2	2	1
Row12	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	Three-quart...	3	3	1
Row13	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	7	7	1
Row14	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	wac[]	24	1
Row15	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	reported[]	33	1
Row16	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	at[]	15	1
Row17	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Comoor[]	3	1
Row18	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Valley[]	5	1
Row19	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Golf[]	3	1
Row20	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Club[]	3	1
Row21	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Homewood[]	5	1
Row22	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	Homewood[]	3	1
Row23	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	,1645CST[]	2	1
Row24	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	*,1645CST[]	2	1
Row25	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	["Hall"]	2	1
Row26	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	["Hall"]	2	1
Row27	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	["Hall"]	2	1
Row28	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	["Hall"]	2	1
Row29	*_1645CST_,,,7,7,7,Hall (1.75)	Three-quarters inch hall was reported at Comoor Valley Golf Club in Homewood...	["Hall"]	["Hall"]	2	1

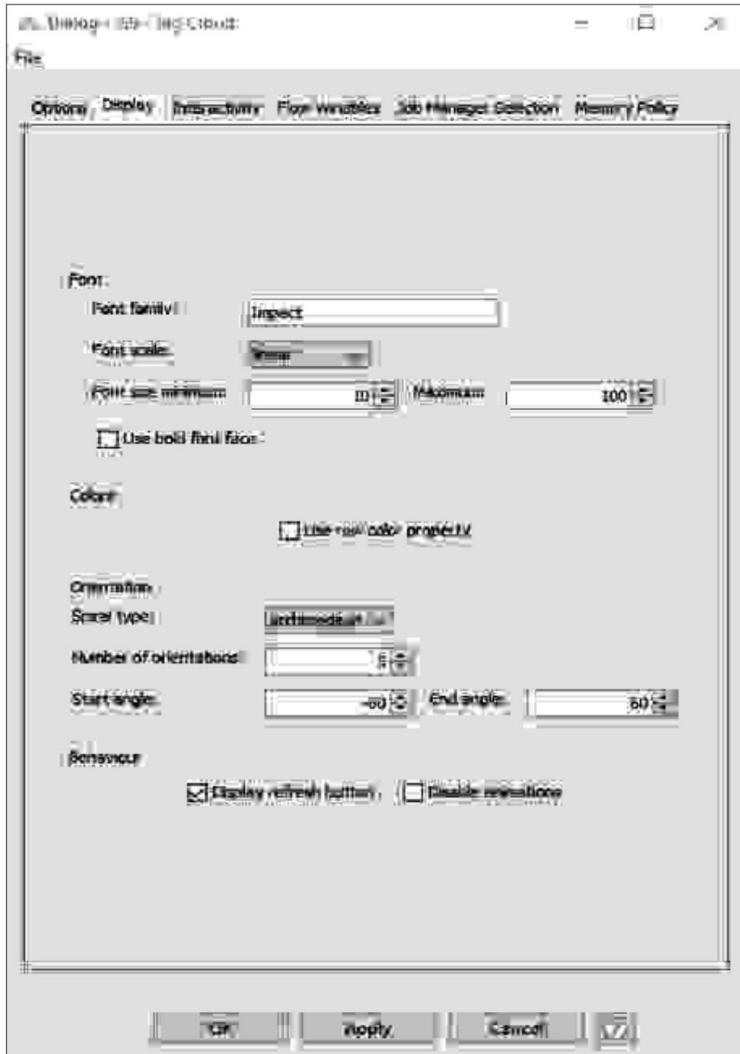
Mặc dù thú vị, nhưng bảng này không hữu ích lắm, nhưng về mặt trực quan, nó sẽ giúp nhà phân tích xác định những từ xuất hiện nhiều nhất và những từ xuất hiện ít nhất. Hình ảnh này được cung cấp bởi nút cuối cùng có tên là “Đám mây thẻ”, theo mô tả, là mã giống như được cung cấp trên một trang web có tên là “Wordle”, mà nhà phân tích có thể thấy trên trang web www.wordle.net.

net và được phát triển bởi Jonathan Feinberg (như được hiển thị trong phần tin dụng trên trang web). Màn hình cấu hình nút Tag Cloud được hiển thị, cùng với cấu hình được đặt cho ví dụ này. Một lần nữa, hãy khám phá các tùy chọn khác nhau cho tất cả các nút này, vì chúng có thể tạo ra một số bản trình bày trực quan rất thú vị để sử dụng trong phân tích dữ liệu.



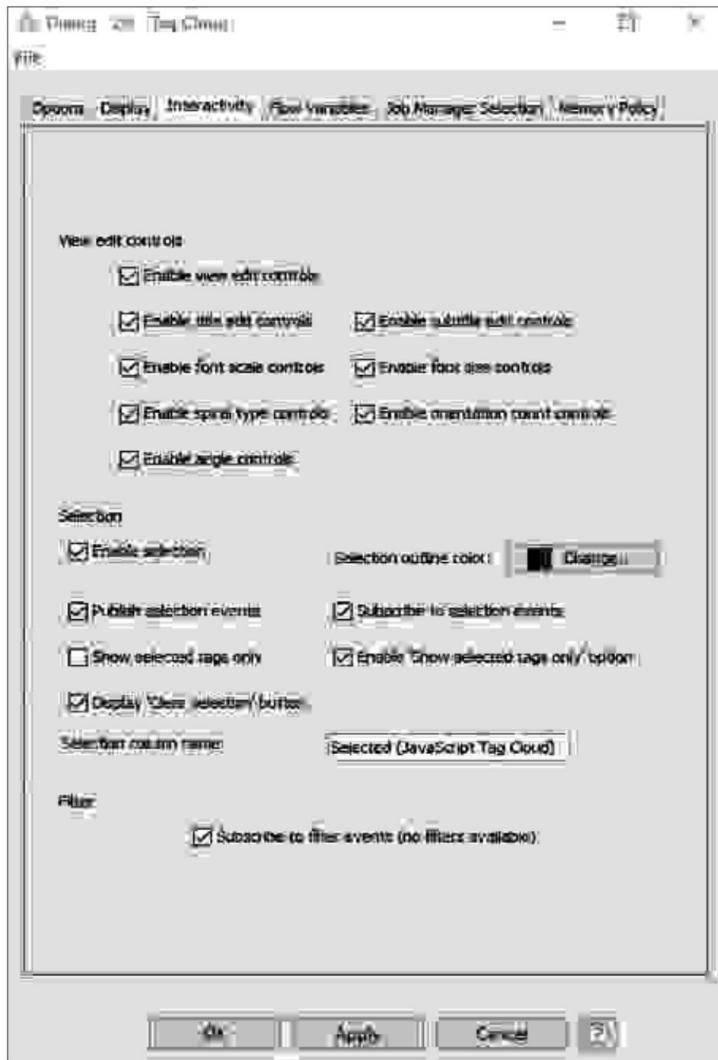
Trong tab này của màn hình cấu hình, nhà phân tích đã đặt tiêu đề và phụ đề của hình ảnh, cùng với cột thẻ và cột kích thước. Cột Kích thước sẽ xác định kích thước của từ dựa trên lần xuất hiện hoặc cùng xuất hiện.

Nhà phân tích có thể sử dụng các cột khác cho mục đích này và nhận được kết quả khác. Đối với ví dụ này, các cài đặt này sẽ thực hiện chức năng. Tab thứ hai hoặc "Hiển thị" tiếp theo với cấu hình sau.

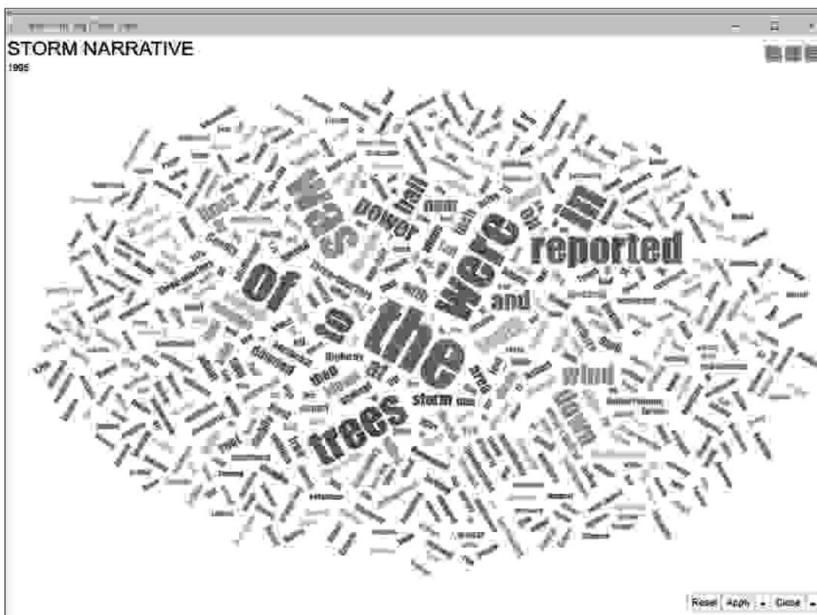


Lựa chọn duy nhất của "Tỷ lệ phóng chữ" dựa trên việc phát triển phóng chữ một cách tuyến tính là trường hợp này, nhưng có các tỷ lệ khác có sẵn từ mũi tên xuống và những tỷ lệ đó có thể hiển thị một số hình ảnh khác với hình ảnh mà nhà phân tích sẽ thấy trong phần sau.

Tab cuối cùng của "Tương tác" cung cấp cho nhà phân tích cách thao tác với hình ảnh cuối cùng, mặc dù đôi khi những cách này có cách cung cấp quá nhiều lựa chọn cho nhà phân tích, điều này làm "làm rối" phân tích. Tuy nhiên, những lựa chọn này ở đây để nhà phân tích đưa ra nêu họ quyết định như vậy.



Kết quả cuối cùng, sau khi quá trình hoàn tất và được kết nối (và executed), là hình ảnh sau. Lưu ý cách các từ hiển thị sự khác biệt về kích thước dựa trên sự xuất hiện của từ. Như trong phiên bản R đã giải thích trước đó, loại phân tích này có thể chỉ ra một số xu hướng và mẫu thú vị. Mặc dù ví dụ này có thể không phải là cách tốt nhất để sử dụng, nhưng các ví dụ như nhận xét về khảo sát hoặc nhận xét văn bản rất phù hợp để sử dụng cho loại phân tích này.



Đây chỉ là một số công cụ có sẵn để khai thác văn bản và nhà phân tích càng tham gia nhiều vào loại dữ liệu này thì càng có nhiều khả năng phân tích văn bản thực tế. Nếu nhà phân tích muốn kiểm tra các chức năng này, hãy chọn bài phát biểu từ nguồn trực tuyến và sử dụng bài phát biểu đó cho các loại phân tích này. Các khả năng là vô tận.

5.7 LỌC

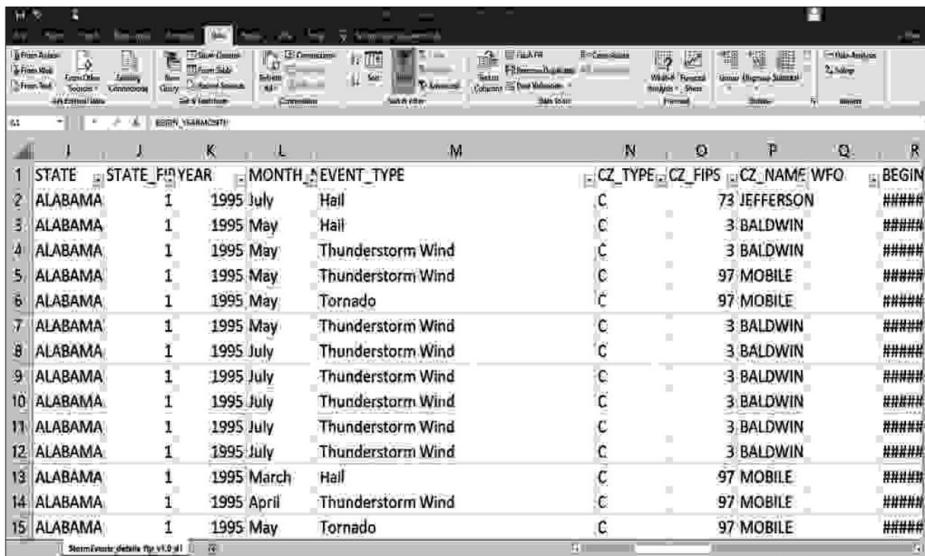
Có lẽ nhiệm vụ cơ bản nhất của bất kỳ nhà phân tích nào là làm sạch dữ liệu để chỉ những dữ liệu thực sự thích hợp mới được sử dụng trong kiểm tra thống kê. Một trong những cách chính để thực hiện chức năng này là lọc các biến để đảm bảo rằng chỉ những biến cần thiết mới hiển thị. Phần này sẽ đề cập đến chức năng đó và sử dụng các công cụ khác nhau để chỉ ra cách thực hiện chức năng đó.

5.7.1 Excel

Trong trường hợp với Excel, việc lọc được thực hiện bằng hai phương pháp. Phương pháp đầu tiên là chọn tùy chọn “Bộ lọc” từ tab “Dữ liệu”, trong khi phương pháp thứ hai là biến bảng tính thành bảng dữ liệu. Cả hai phương pháp đó sẽ được trình bày ở đây.

Trong phương pháp đầu tiên, nhà phân tích trước tiên sẽ nhập dữ liệu; trong trường hợp này, dữ liệu sẽ là dữ liệu lốc xoáy từ năm 1995, vì dữ liệu đó không chỉ chứa dữ liệu lốc xoáy và nhà phân tích chỉ muốn dữ liệu lốc xoáy. Quá trình lọc sẽ loại bỏ tất cả dữ liệu khác trừ dữ liệu lốc xoáy.

Sau khi nhập dữ liệu, nhà phân tích sẽ chuyển đến tab Dữ liệu và ở đó có lựa chọn “Bộ lọc” (trông giống như một cái phễu). Nhập vào kênh và mũi tên xuống bộ lọc sẽ xuất hiện bên cạnh tất cả các biến (tiêu đề cột trong dữ liệu như được hiển thị).

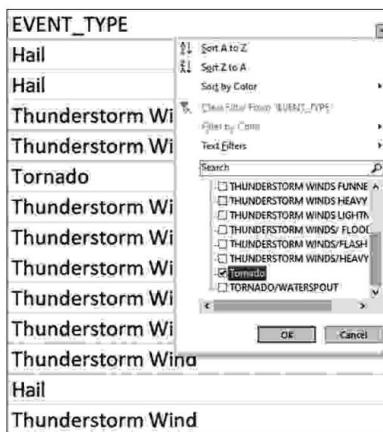


The screenshot shows an Excel spreadsheet titled "DataEvent_detail (F9 v1.0.xls)". The data is filtered to show only events where the "EVENT_TYPE" column contains "Tornado". The visible rows are 1 through 15, all corresponding to Alabama in 1995. The columns include STATE, STATE_FIPS, YEAR, MONTH, EVENT_TYPE, CZ_TYPE, CZ_FIPS, CZ_NAME, WFO, and BEGIN. The "EVENT_TYPE" column is highlighted with yellow, and the filter dropdown arrow is visible next to it. The rest of the columns are greyed out.

	STATE	STATE_FIPS	YEAR	MONTH	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME	WFO	BEGIN
1	ALABAMA	1	1995	July	Hail	C	73	JEFFERSON	#####	
2	ALABAMA	1	1995	May	Hail	C	3	BALDWIN	#####	
3	ALABAMA	1	1995	May	Thunderstorm Wind	C	3	BALDWIN	#####	
4	ALABAMA	1	1995	May	Thunderstorm Wind	C	97	MOBILE	#####	
5	ALABAMA	1	1995	May	Tornado	C	97	MOBILE	#####	
6	ALABAMA	1	1995	May	Thunderstorm Wind	C	3	BALDWIN	#####	
7	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
8	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
9	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
10	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
11	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
12	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN	#####	
13	ALABAMA	1	1995	March	Hail	C	97	MOBILE	#####	
14	ALABAMA	1	1995	April	Thunderstorm Wind	C	97	MOBILE	#####	
15	ALABAMA	1	1995	May	Tornado	C	97	MOBILE	#####	

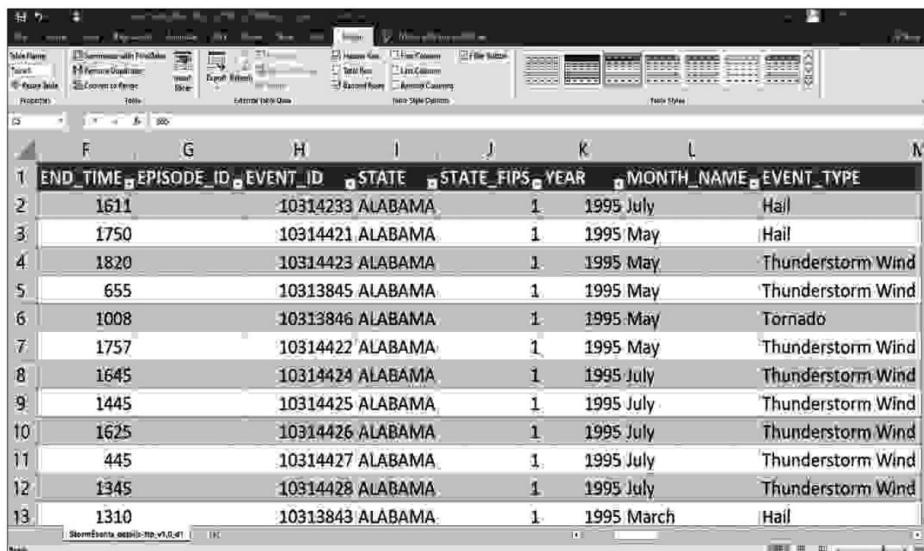
Cuộn cho đến khi cột “EVENT_TYPE” xuất hiện như hình minh họa và sử dụng mũi tên xuống để chỉ chọn cơn lốc xoáy từ các lựa chọn khác nhau có sẵn. Sau khi hoàn thành, chỉ các hàng lốc xoáy sẽ xuất hiện. Sau đó, nhà phân tích có thể chọn bảng tính và sao chép nó sang một trang tính khác để chỉ xử lý các trường hợp xảy ra lốc xoáy.

The screenshot shows the Microsoft Power BI Data Editor interface. On the left, there is a table with columns: STATE, STATE_FIS, YEAR, MONTH, and EVENT_TYPE. The data consists of 15 rows for Alabama in 1995, with various event types like Hail, Thunderstorms, and Tornadoes. On the right, another table is shown with columns: CZ_TYPE, CZ_FIPS, CZ_NAME, WFO, and BEGIN. Below these tables is a filter dialog for the EVENT_TYPE column. The dialog lists several event types and provides a checkbox-based filter for more specific categories. The checked filters include THUNDERSTORM WINDS/FUNNE, THUNDERSTORM WINDS/HEAVY, THUNDERSTORM WINDS/LIGHTN, THUNDERSTORM WINDS/FLOOD, THUNDERSTORM WINDS/FLASH, and TORNADO.



Cách thứ hai để lọc các cột là thay đổi trang tính thành bảng dữ liệu. Quá trình để làm điều này là tương đối đơn giản. Bước đầu tiên là nhập dữ liệu như trước, nhưng lần này hãy chuyển đến tab Chèn và chọn “Bảng” để thay đổi trang tính (hoặc phạm vi) thành bảng dữ liệu.

Kết quả của việc làm như vậy được mô tả trong màn hình sau:



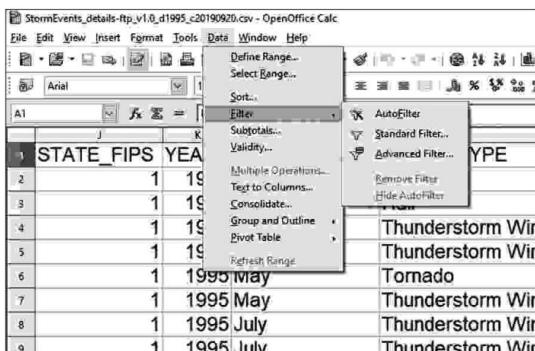
The screenshot shows a Microsoft Excel spreadsheet titled "StormEvents_details-ftp.v1.0_d1995_20190920.csv - Microsoft Excel". The table has columns labeled END_TIME, EPISODE_ID, EVENT_ID, STATE, STATE_FIPS, YEAR, MONTH_NAME, and EVENT_TYPE. The data consists of 13 rows of storm event details. The last row is highlighted in yellow.

	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE
2	1611		10314233	ALABAMA	1	1995	July	Hail
3	1750		10314421	ALABAMA	1	1995	May	Hail
4	1820		10314423	ALABAMA	1	1995	May	Thunderstorm Wind
5	655		10313845	ALABAMA	1	1995	May	Thunderstorm Wind
6	1008		10313846	ALABAMA	1	1995	May	Tornado
7	1757		10314422	ALABAMA	1	1995	May	Thunderstorm Wind
8	1645		10314424	ALABAMA	1	1995	July	Thunderstorm Wind
9	1445		10314425	ALABAMA	1	1995	July	Thunderstorm Wind
10	1625		10314426	ALABAMA	1	1995	July	Thunderstorm Wind
11	445		10314427	ALABAMA	1	1995	July	Thunderstorm Wind
12	1345		10314428	ALABAMA	1	1995	July	Thunderstorm Wind
13	1310		10313843	ALABAMA	1	1995	March	Hail

Như nhà phân tích có thể thấy, bảng dữ liệu đã được trang bị các mũi tên xuống bộ lọc như một phần của quá trình chuyển đổi, vì vậy nhà phân tích có thể sử dụng các mũi tên này như trong các đoạn trước. Có nhiều ưu điểm khi thay đổi trang tính thành bảng dữ liệu, nhưng chúng nằm ngoài phạm vi của cuốn sách này và không được đề cập nhiều hơn trong nhiều cuốn sách Excel hiện có. Điều này được bao gồm trong cuốn sách này chỉ để sử dụng như là một so sánh với các công cụ khác có sẵn ở đây.

5.7.2 Văn phòng mở

OpenOffice có cảm giác giống như các phiên bản Excel cũ hơn, do đó, nơi tự nhiên để bắt đầu quá trình lọc sẽ là truy cập Bảng tính OpenOffice, nhập dữ liệu và chuyển đến tab "Dữ liệu" như trong Excel. Như minh họa trong phần sau, nó hơi khác một chút vì bên dưới tùy chọn bộ lọc có một số lựa chọn.



Lựa chọn “AutoFilter” phù hợp với ví dụ này. Ngay sau khi lựa chọn đó được chọn, cùng loại mũi tên xuống sẽ xuất hiện bên cạnh các tiêu đề cột và nhà phân tích có thể chọn cách lọc dữ liệu. Trong trường hợp này, chọn “lốc xoáy” có vẻ phù hợp.

5.7.3 R/RStudio/Rattle

R có một gói gọi là “dplyr” có thể được tải và sử dụng trong RStudio. Thao tác này sẽ lọc cơ sở dữ liệu để chỉ những cột cần thiết mới được xem và có thể được tải vào Rattle dưới dạng cơ sở dữ liệu R. Trong trường hợp này, nhà phân tích muốn giới hạn các cột chỉ ở những hàng có “Lốc xoáy” trong đó, vì vậy đó là thuật ngữ được sử dụng. Tuy nhiên, để thiết lập cơ sở dữ liệu mới dưới dạng cơ sở dữ liệu được lọc và để tắt tên tệp dài, nhà phân tích quyết định lưu trữ cơ sở dữ liệu Cơn bão nghiêm trọng năm 1995 đã nhập vào cơ sở dữ liệu TORNADO_1995. Điều này cung cấp cho quá trình chuyển đổi dễ dàng hơn nhiều vào lĩnh vực lập trình. Các lệnh sau sẽ tạo ra kết quả cần thiết để tiếp tục với bất kỳ phân tích nào sau đó.

```
> thư viện (dplyr)
> TORNADO_1995<-StormEvents_details_ftp_v1_0_d1995_
c20190920
> TORNADO_1995<-bộ lọc(TORNADO_1995,EVENT_
TYPE=="Lốc xoáy")
> Xem(TORNADO_1995)
```

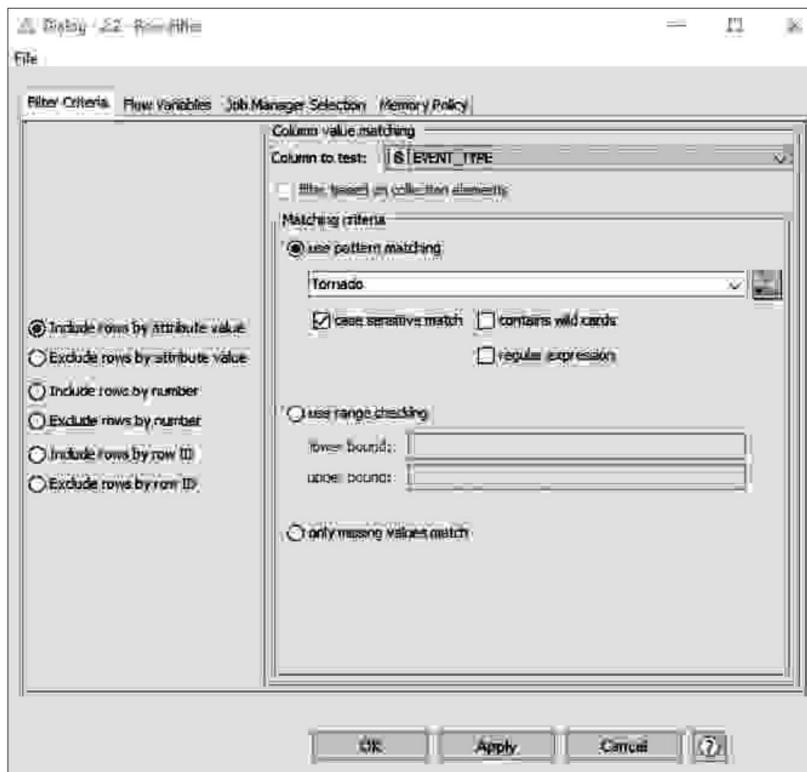
Dòng cuối cùng (bắt đầu bằng “Chế độ xem”) chỉ đơn giản là hiển thị dữ liệu trong ngăn để xem dữ liệu ở định dạng bảng. Điều này giúp nhà phân tích đảm bảo rằng quá trình lọc được thực hiện đúng cách.

5.7.4 KIẾN THỨC

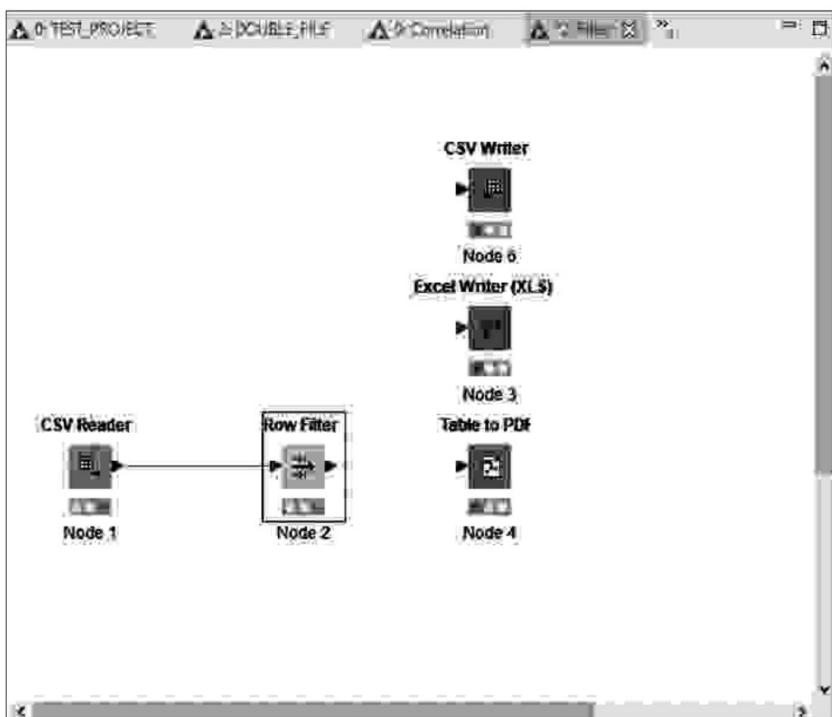
Công cụ KNIME có thể lọc bằng nút cho mục đích này. Bước đầu tiên sẽ là nhập dữ liệu cơ bản năm 1995 vào KNIME bằng nút Trình đọc CSV đã thử và đúng, sau đó lọc dữ liệu bằng nút “Bộ lọc hàng”. Màn hình cấu hình cho nút như sau và bạn cần khám phá cấu hình này để phù hợp nhất với nhu cầu phân tích.

Điều quan trọng cần lưu ý là các phần khác nhau của màn hình cấu hình này. Cột được chọn ở trên cùng bên phải và sau đó “Tiêu chí phù hợp” được chọn. Trong trường hợp này, nhà phân tích chỉ muốn phần lốc xoáy của dữ liệu, vì vậy phần đó được chọn, nhưng ngoài ra, hộp dành cho độ nhạy trường hợp được chọn để đảm bảo rằng

biến phù hợp. Lưu ý rằng "Bao gồm các hàng theo giá trị thuộc tính" được chọn, vì nó phải như vậy, vì nhà phân tích chỉ muốn những hàng được đánh dấu là "Lốc xoáy". Hãy cẩn thận bắt cứ khi nào chọn bất kỳ tùy chọn nào khác, bởi vì chọn sai sẽ loại bỏ tất cả các hàng mà nhà phân tích muốn sử dụng!



Màn hình quy trình làm việc đã hoàn thành cho quy trình làm việc của bộ lọc KNIME như sau và lưu ý đặc biệt về các nút được thêm vào. Các nút này ở đó để hiển thị các loại công cụ khác nhau mà KNIME có thể xuất, bao gồm hai công cụ được đề cập trong văn bản này. Các nút này có thể được tìm thấy trong danh mục "IO" và có thể là một cải tiến lớn đối với phân tích dữ liệu, vì cùng một dữ liệu có thể được phân tích bằng các công cụ khác nhau. Các nút này cũng có thể được sử dụng để xuất liên tiếp một số bộ dữ liệu vì sau khi các nút được đặt trong quy trình làm việc, đầu ra cũng có thể được xác định một cách nhất quán.



Lý do bao gồm tùy chọn “Bảng thành PDF” là vì có những lúc phân tích hoàn chỉnh phù hợp nhất cho một báo cáo và không gì bằng việc chuyển đổi bảng thành PDF để giúp đưa trang báo cáo đó vào kiểu tài liệu linh hoạt nhất. Ngoài ra, tài liệu PDF có thể được nhập vào một số công cụ và được sử dụng trong phân tích dữ liệu trong tương lai, vì vậy việc chuyển đổi sang PDF chỉ có ý nghĩa về lâu dài. Bất kể lý do là gì, việc sử dụng các nút đầu ra sẽ giúp nhà phân tích năng động trong các phân tích trong tương lai của họ.

CHƯƠNG 6

BẢN TÓM TẮT

Văn bản này dựa trên một số khái niệm thống kê khác nhau hiện đang tồn tại trong "nhà bánh xe" của nhà phân tích dữ liệu. Nếu nhà phân tích dữ liệu không quen thuộc với bất kỳ khái niệm nào đã nói ở trên, vui lòng đọc nhiều tài liệu thống kê và tài liệu tham khảo ở cuối cuốn sách này hoặc được tìm thấy ở nhiều hiệu sách trực tuyến và truyền thông. Bắt đầu với một số văn bản rất cơ bản và chuyển sang phức tạp hơn. Dù nhà phân tích làm gì theo cách phân tích dữ liệu, thì một nền tảng thống kê tốt là cần thiết và hiệu quả. Luôn có thêm thông tin về phân tích dữ liệu, khoa học dữ liệu và số liệu thống kê, vì vậy đừng bao giờ để cơ hội học tập có thể bỏ qua. Ngoài ra, đối với các công cụ này, chức năng rất "lớp đất mặt" đã được chứng minh với chúng. Chắc chắn có nhiều thông tin hơn và nhiều chức năng hơn có thể có với các ứng dụng này. Điều quan trọng là nhà phân tích sử dụng các công cụ này cho mục đích phân tích. Tránh sử dụng công cụ để hiển thị biểu đồ đầy màu sắc hoặc để trực quan hóa thứ gì đó có thể không hợp lệ. Chính danh tiếng của nhà phân tích đang bị đe dọa khi lấy các biến được kết nối lồng léo và cố gắng kết nối chúng. Đó không phải là mục đích của các công cụ phân tích. Mục đích của các công cụ là cung cấp cho nhà phân tích phương pháp nhanh nhất để tính toán thứ gì đó sẽ mất hàng giờ để thực hiện bằng các phương pháp thủ công như máy

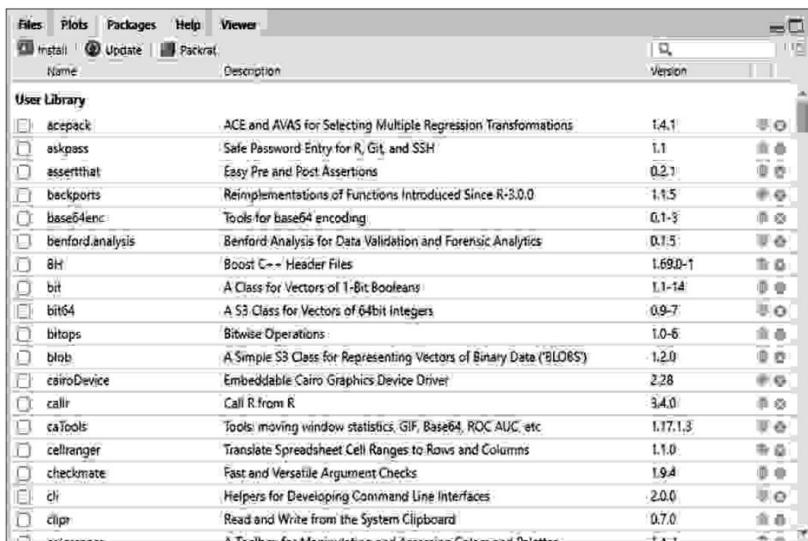
6.1 CÁC GÓI

Có một số lĩnh vực không được thảo luận ở phần đầu của văn bản cần được làm rõ ngay bây giờ. Đầu tiên là "gói" liên quan đến Rattle (và R nói chung). Một gói trong R đề cập đến một chức năng được lập trình cụ thể hoạt động như một quy trình "một bước" để thực hiện các thử nghiệm và mô hình nhất định. Đây là những yếu tố quan trọng giúp quá trình phân tích trở nên nhanh chóng và hiệu quả nhất có thể, nhưng có một số lưu ý cần đi kèm với các gói này. Đầu tiên

gói phải được cài đặt để được kích hoạt. Một số được cài đặt với bản cài đặt R cơ bản, nhưng có nhiều cái thì không. Rattle thực sự là một gói phải được cài đặt để hoạt động. Mỗi khi nhà phân tích đóng R (do đó sẽ đóng Rattle nếu nó đang mở), cơ sở R sẽ chuyển đổi trở lại thành không kích hoạt các gói trong R. Gói sẽ vẫn được cài đặt nhưng sẽ không được kích hoạt cho đến khi người dùng thực hiện điều này thông qua cửa sổ lập trình (đã được trình bày) hoặc bằng cách “đánh dấu vào ô” bên cạnh gói trong ngăn dưới cùng bên phải của IDE của RStudio, được mô tả như sau. Các gói được hiển thị chỉ là một phần nhỏ trong số các gói có sẵn thông qua Mạng lưu trữ R toàn diện (hoặc CRAN), từ đó có thể tìm thấy và cài đặt bất kỳ gói nào. Khi nhà phân tích cài đặt R, họ có thể chọn CRAN “mirror” (về cơ bản là máy chủ) để tải các gói này từ đó.

Trong một số trường hợp, bạn nên nhấp vào nút “cài đặt” trong màn hình sau và nhập chức năng cần kích hoạt. Trong hầu hết các trường hợp, có một chương trình con chức năng (gói) có thể thực hiện “công việc lonen” cho nhà phân tích. Đây là sức mạnh của R và Rattle—để giúp nhà phân tích giải quyết các vấn đề phân tích mà không cần chuyên môn lập trình đặc biệt.

Một điểm nữa về các gói. Các chương trình con này có thể dựa (phụ thuộc) vào các gói khác và có thể cài đặt các phụ thuộc này để tuổi gói hoạt động bình thường. Trong nhiều trường hợp, điều này sẽ được thông báo cho người dùng và quyền sẽ được yêu cầu cài đặt các gói phụ thuộc. Nếu người dùng có bất kỳ nghi ngờ nào về tính phù hợp của gói hoặc các phần phụ thuộc, họ có thể từ chối yêu cầu này. Tuy nhiên, điều đó có nghĩa là gói sẽ không hoạt động bình thường.



The screenshot shows the RStudio interface with the 'Packages' tab selected. The window title is 'Packages'. At the top, there are buttons for 'Install', 'Update', and 'Purify'. Below the buttons is a search bar and a 'Version' dropdown. The main area is titled 'User Library' and lists various R packages with their names, descriptions, and versions. Some packages have small icons next to them. A scroll bar is visible on the right side of the package list.

Name	Description	Version
User Library		
acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
askpass	Safe Password Entry for R, Git, and SSH	1.1
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.5
base64enc	Tools for base64 encoding	0.1-3
benford.analysis	Benford Analysis for Data Validation and Forensic Analytics	0.7.5
BH	Boost C++ Header Files	1.69.0-1
bit	A Class for Vectors of 1-Bit Booleans	1.1-14
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
bitops	Bitwise Operations	1.0-6
blob	A Simple S3 Class for Representing Vectors of Binary Data ("BLOBS")	1.2.0
cairoDevice	Embeddable Cairo Graphics Device Driver	2.28
callr	Call R from R	3.4.0
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc	1.17.1.3
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
checkmate	Fast and Versatile Argument Checks	1.9.4
cli	Helpers for Developing Command Line Interfaces	2.0.0
clipr	Read and Write from the System Clipboard	0.7.0
colorspace	A Toolkit for Manipulating and Accessing Colorspace Colours	1.4.1

Khía cạnh thú vị của RStudio là, bằng cách chọn hộp bên cạnh gói, chương trình R sẽ tự động được kích hoạt để đặt gói theo ý của người dùng. Không có chương trình nào khác mà người dùng phải hoàn thành, chỉ cần đảm bảo hộp được chọn. Đây chỉ là một lý do tại sao tải xuống và cài đặt RStudio rất đáng giá đối với bất kỳ ai muốn thực hiện phân tích dữ liệu bằng ứng dụng FOSS.

6.2 CÔNG CỤ PHÂN TÍCH

ToolPak Phân tích trong Excel có thể không khả dụng cho mọi người dùng, đặc biệt đối với những người có thể làm việc trong chính phủ liên bang, vì đó được gọi là phần bổ trợ và có thể nằm trong các thỏa thuận bổ sung khác ngoài Microsoft Office cơ bản. Do đó, phần bổ trợ có thể không sẵn dùng cho những người cần nó. Nếu đây là trường hợp, thì nhà phân tích luôn có thể thực hiện phương pháp "thủ công" để có được dữ liệu giống như khi sử dụng ToolPak phân tích. Điều này phức tạp hơn so với việc sử dụng sự tiện lợi của phần bổ trợ, nhưng với một chút kiên nhẫn và bền bỉ, kết quả tương tự sẽ xuất hiện.

Để làm điều này, điều đầu tiên là nhập dữ liệu như đã làm trước đó, nhưng lần này sử dụng phần dưới cùng của trang tính để liệt kê và sử dụng các công thức khác nhau để tạo ra bản tóm tắt mô tả như trong phần trước. Màn hình sau đây sẽ hiển thị tất cả các công thức cần thiết để cung cấp thông tin về độ dài cơn lốc xoáy (TOR_LENGTH) trên dữ liệu cơn lốc xoáy năm 1951.

Mỗi công thức này sẽ được thảo luận và kết quả được hiển thị. Một gợi ý về cách hiển thị công thức trong Excel: nếu cần hiển thị công thức trong bảng tính, hãy chuyển đến tab "Công thức" trên thanh công cụ chính và chọn "Hiển thị Công thức". Nếu bạn muốn sử dụng phím tắt, hãy giữ phím CTRL và nhấn nút "~" nằm ngay bên dưới phím "Esc". Đây là phím "chuyển đổi" mà nhà phân tích có thể nhấn liên tục để hiển thị các công thức hoặc kết quả.

	A1	BB	BC	BD	BE	BF	BG	BH
1	TOR_LENGTH							
271	4.443494424	MEAN						
272	0.5	MEDIAN						
273	0	MODE						
274	104.2812681	VARIANCE						
275	10.21182002	STANDARD DEVIATION						
276	0	MIN						
277	92.6	MAX						
278	92.6	RANGE						
279	4.376062845	SKEW						
280	25.67453191	KURTOSIS						
281	1.218061905	CONFIDENCE INTERVAL						

	A1	BB	BC
1	TOR_LENGTH		
271	=AVERAGE(A12:A1270)	MEAN	
272	=MEDIAN(A12:A1270)	MEDIAN	
273	=MODE(A12:A1270)	MODE	
274	=VAR.P(A12:A1270)	VARIANCE	
275	=STDEV.P(A12:A1270)	STANDARD DEVIATION	
276	=MIN(A12:A1270)	MIN	
277	=MAX(A12:A1270)	MAX	
278	=A1277-A1276	RANGE	
279	=SKEW(A12:A1270)	SKEW	
280	=KURT(A12:A1270)	KURTOSIS	
281	=CONFIDENCE.NORM(0.05,A1275,270)	CONFIDENCE INTERVAL	

So sánh các kết quả này với kết quả từ phần Thống kê mô tả và sẽ có rất ít sự khác biệt nếu có. Ngoài ra, phần hiển thị các công thức này trong OpenOffice cũng khác vì OpenOffice sử dụng ";" và Excel sử dụng "," vì vậy hãy nhớ những khác biệt này khi di chuyển giữa các công cụ.

CHƯƠNG 7

BỔ SUNG THÔNG TIN

Phần này sẽ chứa thông tin đã bị bỏ qua trong phần giải thích của một số phần khác cùng với một số bài tập để người đọc sử dụng nhằm tập trung tốt hơn vào các khái niệm khác nhau được trình bày trong các phần trước. Các câu trả lời sẽ bao gồm hầu hết các công cụ, sử dụng công cụ tốt nhất cho vấn đề và hướng tới những công cụ khác có thể thực hiện thủ thuật. Không phải tất cả các công cụ sẽ được đưa vào tất cả các câu trả lời một cách riêng biệt, nhưng mỗi công cụ riêng lẻ sẽ được hiển thị trong ít nhất một trong các câu trả lời. Vui lòng đi đến những nơi có thể truy cập dữ liệu và sử dụng dữ liệu cho các bài tập để có một số niềm vui phân tích. Nếu không, cuốn sách này sẽ kết thúc trên kệ, không bao giờ được sử dụng ngoài một cái chăn giấy.

7.1 BÀI TẬP MỘT - LÚC LÙI VÀ CÁC BANG

Bài tập đầu tiên sẽ khám phá việc sử dụng một số công cụ để phân tích những bang nào dường như có lốc xoáy nhiều hơn những bang khác. Nhà phân tích không bao giờ nên đi vào một phân tích vội vàng để kết luận. Có thể có một giả thuyết mà nhà phân tích muốn đưa ra. Đây không phải là một kết luận, mà thực sự là một khẳng định. Chẳng hạn, nhà phân tích có thể nói rằng có nhiều lốc xoáy ở Texas hơn ở Connecticut vào năm 2018. Đây là khẳng định có thể được xác minh bằng dữ liệu và phân tích cơ bản.

Phần này sẽ tập trung vào một bài kiểm tra thống kê cụ thể và cách bác bỏ hoặc không bác bỏ giả thuyết không dựa trên bài kiểm tra đó. Bước đầu tiên là nêu giả thuyết là giả thuyết không và sau đó đưa ra một giả thuyết thay thế. Để làm cho nó rõ ràng, đây không phải là một quá trình chính thức, nhưng

sử dụng công thức giả thuyết không chính thức giúp phân tích chính xác hơn, vì việc không nêu rõ một giả thuyết cho phép nhà phân tích "chơi linh vực" liên quan đến loại biến dữ liệu để kiểm tra và nhờ đó, mở rộng mối quan hệ giữa các trường này cho đến khi một hoặc nhiều hơn nữa có liên quan. Đây là một cách thực hiện phân tích thiên vị và sẽ dẫn đến các mối tương quan giả có thể xảy ra hoặc tệ hơn là xác định rằng một biến có mối quan hệ nhân quả với một biến khác trong khi thực tế không có mối quan hệ nào thuộc loại đó.

Trong trường hợp này, khẳng định (hoặc tuyên bố) là có nhiều lốc xoáy ở Texas hơn ở Connecticut vào năm 2018. Giả thuyết vô hiệu (không phải khẳng định trong trường hợp này) là có cùng số lượng lốc xoáy ở Texas và ở Con mực hoa. Có những nhà phân tích sẽ cố gắng thực hiện phân tích tương quan hoặc hồi quy để chứng minh khẳng định, nhưng trong trường hợp này, một phân tích mô tả đơn giản là quá đủ để đưa ra trường hợp.

Bước đầu tiên sẽ là nhập dữ liệu vào công cụ và sau đó thực hiện thống kê mô tả đối với dữ liệu đó. Sau khi thực hiện thử nghiệm đó, nhà phân tích có thể hiển thị mối quan hệ bằng biểu đồ thanh đơn giản hoặc hình ảnh tương tự. Hãy nhớ rằng có một số loại dữ liệu phù hợp hơn với một số loại bản trình bày trực quan nhất định. Các biến rời rạc (những biến số nguyên), không phụ thuộc vào thời gian, sẽ dễ thích ứng hơn với biểu đồ thanh. Các nghiên cứu theo chiều dọc, những nghiên cứu dựa trên các năm tiếp theo, phù hợp hơn với biểu đồ đường. Điều này rất quan trọng vì chính kiểu liên kết đó sẽ cho phép nhà phân tích trình bày rất hiệu quả mà không gây nhầm lẫn cho khán giả.

Tìm câu trả lời cho bài tập này bằng cách sử dụng bất kỳ công cụ nào được trình bày trong văn bản này cùng với bộ dữ liệu lốc xoáy năm 2018 (đảm bảo rằng nó có ghi "chi tiết" trên tên tệp) tại trang web được đề cập trong một số phần đầu tiên của văn bản này, cụ thể là tại <https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/>.

Hãy nhớ rằng tệp này sẽ có tất cả các sự kiện bão, vì vậy nhà phân tích sẽ phải lọc các cơn lốc xoáy khỏi các sự kiện bão còn lại để có được dữ liệu phù hợp để phân tích. Lọc đã được giải quyết trong một phần trước đó.

7.1.1 Trả lời bài tập 7.1

Câu trả lời cho bài tập sẽ yêu cầu lọc dữ liệu để đảm bảo rằng chỉ các hàng lốc xoáy mới được đưa vào dữ liệu. Sau đó, một phép so sánh đơn giản giữa Texas và Connecticut về số lượng (hoặc trung bình) lốc xoáy sẽ đủ để phân tích. Có một cái gì đó có thể được xem xét trong phân tích này.

Các yếu tố như dân số và diện tích đất có thể ảnh hưởng đến dữ liệu thực tế, cụ thể là về số lượng lốc xoáy. Điều này sẽ được chuẩn hóa ("chuan hua") bằng cách sử dụng trọng số, trong trường hợp này không được xem xét và không nằm trong phạm vi của cuốn sách này.

7.1.1.1 Trả lời Theo OpenOffice

Câu trả lời cho câu hỏi sử dụng OpenOffice sẽ rất giống với Excel, ngoại trừ việc OpenOffice không có ToolPak Phân tích.

Các bước để phân tích dữ liệu, xem xét giả thuyết rằng có ít lốc xoáy ở Connecticut hơn Texas, sẽ là nhập dữ liệu rồi lọc dữ liệu để chỉ xem xét các sự kiện lốc xoáy. Sau đó, trình bày dữ liệu dưới dạng hình ảnh sẽ cho người nhận câu trả lời cho câu hỏi (hoặc có xác nhận giả thuyết hay không).

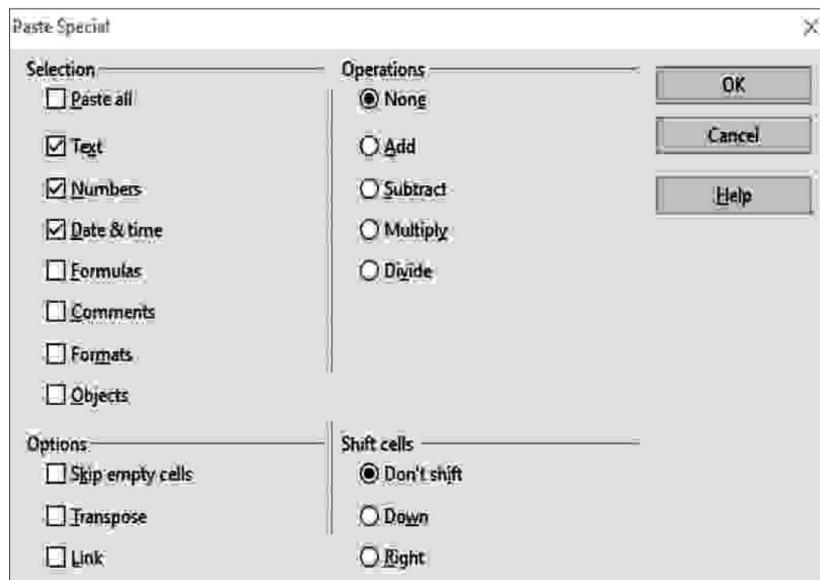
Dữ liệu trong OpenOffice sẽ giống như màn hình sau đây, hiển thị toàn bộ dữ liệu, bao gồm tất cả các cơn bão nghiêm trọng, cần được lọc. Màn hình tiếp theo hiển thị dữ liệu đã lọc; và màn hình cuối cùng hiển thị biểu đồ thanh hiển thị số lượng lốc xoáy trên mỗi tiểu bang.

	F	G	H	I	J	K	L	M	N	
	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH	NAME	EVENT_TYPE	CZ_TYP
1	1611	10314233	ALABAMA			1	1995	July	Hail	C
2	1750	10314421	ALABAMA			1	1995	May	Hail	C
3	1820	10314423	ALABAMA			1	1995	May	Thunderstorm Wind	C
4	655	10313845	ALABAMA			1	1995	May	Thunderstorm Wind	C
5	1008	10313846	ALABAMA			1	1995	May	Tornado	C
6	1757	10314422	ALABAMA			1	1995	May	Thunderstorm Wind	C
7	1645	10314424	ALABAMA			1	1995	July	Thunderstorm Wind	C
8	1445	10314425	ALABAMA			1	1995	July	Thunderstorm Wind	C
9	1625	10314426	ALABAMA			1	1995	July	Thunderstorm Wind	C
10	445	10314427	ALABAMA			1	1995	July	Thunderstorm Wind	C
11	1345	10314428	ALABAMA			1	1995	July	Thunderstorm Wind	C
12	1310	10313843	ALABAMA			1	1995	March	Hail	C
13	205	10313844	ALABAMA			1	1995	April	Thunderstorm Wind	C
14	40	10313847	ALABAMA			1	1995	May	Tornado	C
15	325	10313848	ALABAMA			1	1995	May	Hail	C
16	1355	10313850	ALABAMA			1	1995	May	Thunderstorm Wind	C
17	1402	10313851	ALABAMA			1	1995	May	Thunderstorm Wind	C
18	1005	10313849	ALABAMA			1	1995	April	Thunderstorm Wind	C
19	1449	10314040	ALABAMA			1	1995	October	Thunderstorm Wind	C

Bước tiếp theo là lọc dữ liệu để chỉ những cơn lốc xoáy hiển thị trong cột EVENT_TYPE. Điều này được thực hiện thông qua lựa chọn Dữ liệu trên thanh công cụ và chọn tùy chọn Lọc. với tùy chọn phụ Tự động Lọc. như được hiển thị. Điều này sẽ làm là đặt phễu bên cạnh tất cả các cột và sau đó nhà phân tích có thể chọn biến mong muốn.

	STATE	STATE_FIP	
1	ALABAMA		
2	ALABAMA		Tornado
3	ALABAMA		Thunderstorm Wind
4	ALABAMA		Thunderstorm Wind
5	ALABAMA		Thunderstorm Wind
6	ALABAMA		Thunderstorm Wind
7	ALABAMA		Thunderstorm Wind
8	ALABAMA		Thunderstorm Wind
9	ALABAMA		Thunderstorm Wind
10	ALABAMA	1995 July	Thunderstorm Wind
11	ALABAMA	1995 July	Thunderstorm Wind
12	ALABAMA	1995 July	Thunderstorm Wind
13	ALABAMA	1995 March	Hail
14	ALABAMA	1995 April	Thunderstorm Wind
15	ALABAMA	1995 May	Tornado
16	ALABAMA	1995 May	Hail
17	ALABAMA	1995 May	Thunderstorm Wind
18	ALABAMA	1995 May	Thunderstorm Wind
19	ALABAMA	1995 April	Thunderstorm Wind
20	ALABAMA	1995 October	Thunderstorm Wind
21	ALABAMA	1995 June	Hail
22	ALABAMA	1995 April	Thunderstorm Wind
23	ALABAMA	1995 July	Thunderstorm Wind
24	ALABAMA	1995 August	Thunderstorm Wind

Kết quả của bộ lọc này được hiển thị trong màn hình tiếp theo. Xin lưu ý rằng yếu tố lốc xoáy hiện là yếu tố duy nhất hiển thị. Bước tiếp theo sẽ là sao chép và dán dữ liệu đã lọc vào một trang tính khác. Đây là quy trình tương tự như quy trình được thực hiện trong Excel, vì vậy hãy chọn tất cả dữ liệu và dán vào một trang tính khác. Tuy nhiên, hãy cẩn thận với bước này vì mong muốn là dán dưới dạng các giá trị chứ không chỉ dán mọi thứ, vì thao tác đó sẽ bao gồm tất cả các giá trị chưa được lọc, làm cho trang tính đã sao chép chứa tất cả dữ liệu, không chỉ các hàng lốc xoáy. Để thực hiện thao tác này, bạn nhấp chuột phải vào dữ liệu cần sao chép rồi chọn ô A1 trong trang tính trống. Nhấp chuột phải vào trang tính trống và sẽ có một tùy chọn có tên là Paste Special. Khi bấm vào đó, màn hình sau sẽ xuất hiện:



Nếu hộp kiểm được chọn cho Dán tất cả, trang tính được sao chép sẽ giống với trang tính chưa được lọc. Nếu cái đó không được chọn và ba cái được chọn được hiển thị, thì trang tính đã sao chép sẽ dành cho tất cả các ý định và mục đích, một trang tính mới chỉ gồm các sự kiện lốc xoáy. Đó là kết quả mà nhà phân tích muốn đạt được.

	I	K	L	M	N	O	P	
1	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME
2	ALABAMA	1	1995	May	Tornado	C	97	MOBILE
3	ALABAMA	1	1995	May	Tornado	C	97	MOBILE
4	ALABAMA	1	1995	July	Tornado	C	75	LAMAR
5	ALABAMA	1	1995	April	Tornado	C	51	ELMORE
6	ALABAMA	1	1995	March	Tornado	C	103	MORGAN
7	FLORIDA	12	1995	February	Tornado	C	21	COLLIER
8	ALABAMA	1	1995	May	Tornado	C	77	LAUDERDALE
9	ALABAMA	1	1995	April	Tornado	C	75	LAMAR
10	ALABAMA	1	1995	April	Tornado	C	53	ESCAMBIA
11	ALABAMA	1	1995	March	Tornado	C	125	TUSCALOOSA
12	ALABAMA	1	1995	October	Tornado	C	53	ESCAMBIA
13	MISSOURI	29	1995	June	Tornado	C	35	CARTER
14	MISSOURI	29	1995	April	Tornado	C	141	MORGAN
15	MONTANA	30	1995	May	Tornado	C	0	MT2003 - 0C

Từ kết quả này, giờ đây, nhà phân tích có thể tạo một bảng tổng hợp với STATE làm trục x và Tornado làm trục y. Nhà phân tích sẽ muốn đảm bảo rằng hai tiểu bang được so sánh là Texas và Connecticut như minh họa dưới đây:

The screenshot shows a spreadsheet application window titled "StormEvents_details-tp_v1_0_41995_12019020.csv - OpenOffice Calc". The table contains data from 1995, including columns for STATE, MONTH_NAME, EVENT_TYPE, CZ_TYPE, CZ_FIPS, and CZ_NAME. A context menu is open over the table, with the "Pivot Table" option highlighted.

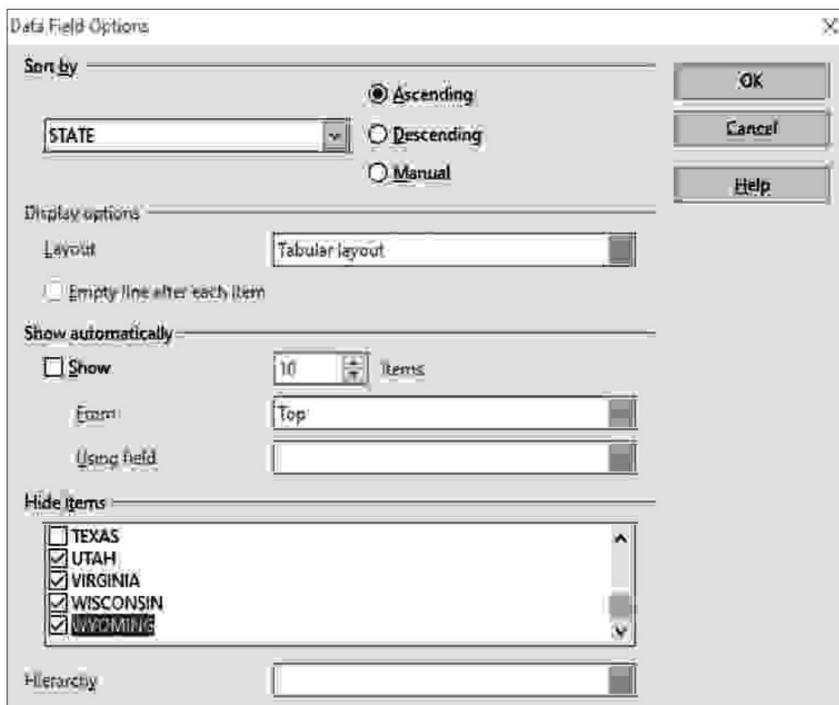
The screenshot shows the "Pivot Table" dialog box. In the "Layout" section, the "Row Fields" is set to "STATE" and the "Data Fields" is set to "Count - EVENT_TYPE". In the "Fields" section, various fields are listed: BEGIN_YEAR, YEAR, BEGIN_DAY, MONTH_NAM..., BEGIN_TIME, EVENT_TYPE, END_YEAR..., CZ_TYPE, END_DAY, CZ_FIPS, END_TIME, CZ_NAME, EPISODE_ID, WFO, EVENT_ID, BEGIN_DATE..., STATE, CZ_TIMEZONE..., STATE_FIPS, and END_DATE....

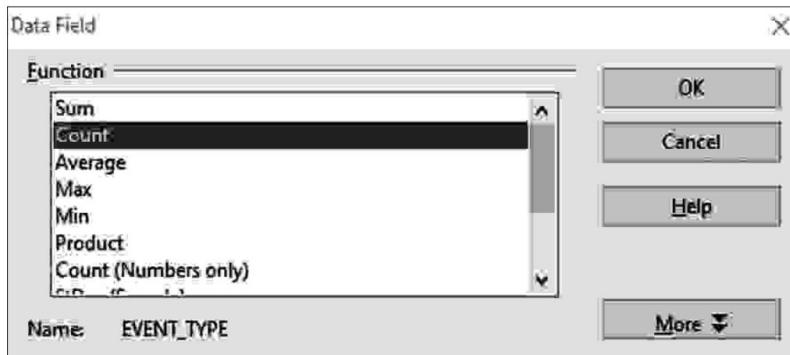
Result:

- Selection from: \$ Filtered Data - Tornado '\$A\$1:\$AV\$1218'
- Results to: - new sheet -
- Ignore empty rows:
- Total columns:
- Add filter:
- Identify categories:
- Total rows:
- Enable drill to details:

Vui lòng định cấu hình màn hình Bảng tổng hợp như trong màn hình trước, nhưng đảm bảo rằng bảng tổng hợp được đặt trong một trang tính khác. Nếu không, bảng tổng hợp sẽ tồn tại trong cùng một trang tính với dữ liệu gốc, điều này có thể gây ra sự cố nếu nhà phân tích không biết về điều này và chọn toàn bộ trang tính, có thể bao gồm bảng tổng hợp, làm ô nhiễm dữ liệu với các số bổ sung.

Kết quả của bảng tổng hợp, với STATE là các hàng và DATA là số lượng lốc xoáy, như sau, và rõ ràng là có nhiều lốc xoáy ở Texas hơn là ở Connecticut. Tuy nhiên, và điều này rất quan trọng, diện tích đất của Connecticut nhỏ hơn nhiều so với Texas và nếu điều này được tính trọng số, các con số sẽ gần hơn. Điều đó nằm ngoài phạm vi của cuốn sách này, nhưng việc khám phá những đặc điểm này sẽ chỉ giúp người phân tích trở thành người sử dụng dữ liệu chi tiết hơn nhiều.





StormEvents_details-ftp_v1_2_d1995_20190226.csv - OpenOffice Calc		
	A	B
1	Filter	
2		
3	STATE	
4	CONNECTICUT	3
5	TEXAS	229
6	Total Result	232
7		
8		
9		
10		

7.1.1.2 Trả lời Theo Rattle

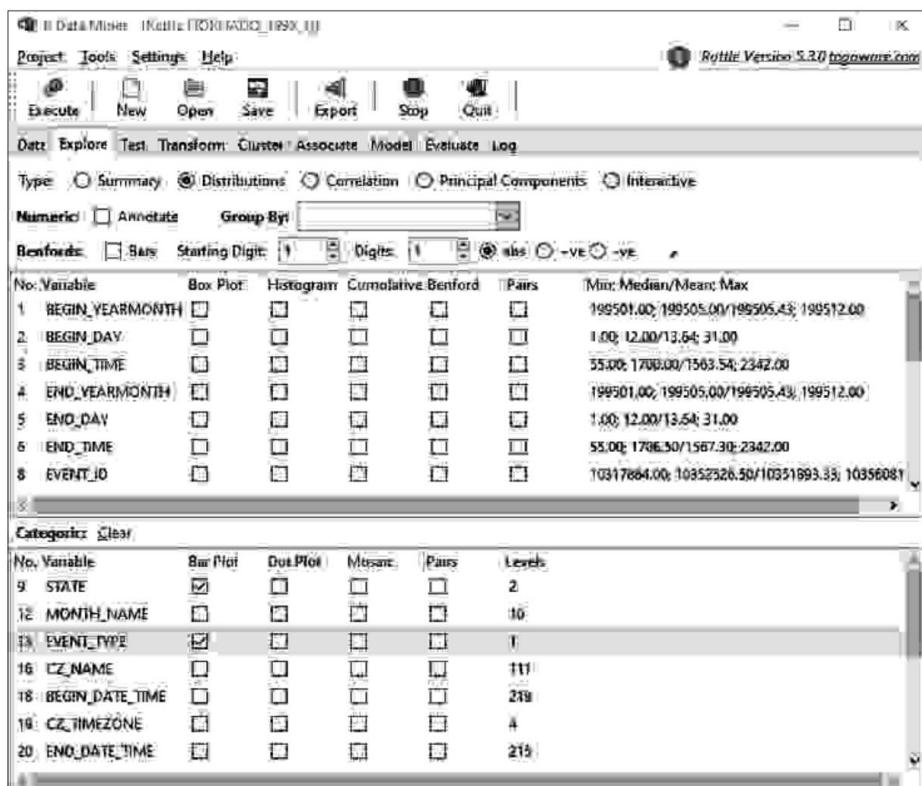
Hình thành một biểu đồ trong Rattle tương đối đơn giản. Phần khó khăn nhất của quá trình này là lọc dữ liệu, nhưng điều đó có thể được thực hiện trong cả R và Rattle. Quá trình lọc bằng R/RStudio đã được trình bày trong phần trước, vì vậy tệp TORNADO_1995 đã chỉ có bản ghi lốc xoáy và các trường được chuẩn bị. Để nhập phần này vào Rattle, hãy sử dụng chức năng nhập dữ liệu như được mô tả trong phần trước. Trước khi nhập dữ liệu, bạn nên loại bỏ tất cả các BIỂU TƯỢNG ngoại trừ những BIỂU TƯỢNG cần thiết. Điều này cũng được thực hiện thông qua lệnh lọc với dòng lệnh hiển thị ở đây.

```
TORNADO_1995_1<- filter(TORNADO_1995,STATE=="TEXAS" |  
BANG=="CONNECTICUT")
```

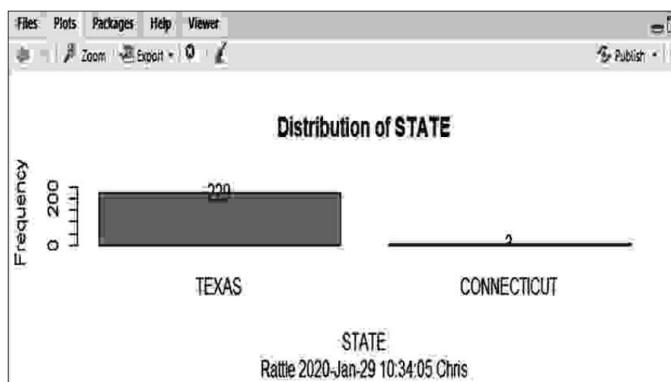
Lưu ý rằng có một đường ống ("|") được sử dụng cho hàm "hoặc", vì vậy dòng này cho biết rằng tôi muốn bộ dữ liệu gốc được lọc để chỉ hiển thị những cơn lốc xoáy đã xảy ra ở Texas hoặc Connecticut. Sau khi hoàn tất, hãy nhập dữ liệu vào Rattle và đảm bảo rằng biểu tượng Thực thi đã được nhấp. Người dùng có thể phải sử dụng nút radio được đánh dấu là "bỏ qua" để đảm bảo rằng chỉ hai trường STATE và EVENT_TYPE đó được chọn. Điều này được thể hiện như sau:

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
5	END_DAY	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 26
6	END_TIME	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 173
7	EPISODE_ID	missing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 9 Missing: 23
8	EVENT_ID	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 232
9	STATE	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique 2
10	STATEPIPE	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 2
11	YEAR	Constant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 1
12	MONTH_NAME	Categorical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 10
13	EVENT_TYPE	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 1
14	CZ_TYPE	Constant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 1
15	CZ_RPS	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 109
16	CZ_NAME	Categorical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 111
17	WFO	missing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique 0 Missing: 23
18	BEGIN_DATE_TIME	Categorical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 219

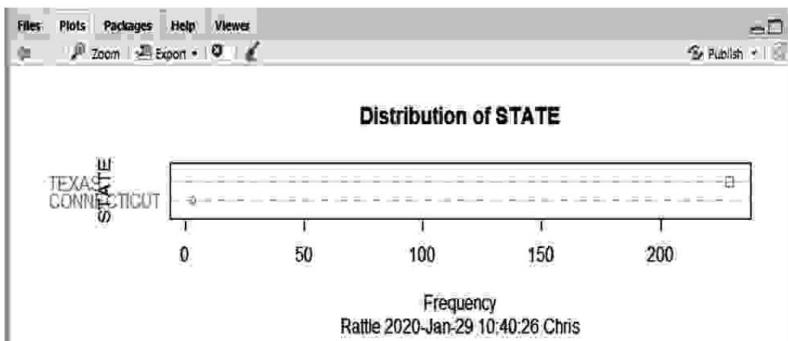
Sau khi hoàn tất, hãy chuyển đến tab "Khám phá" và sử dụng lựa chọn "Phân phối" của biểu đồ thanh để hiển thị hai kết quả. Màn hình câu hình như sau:



Kết quả của việc nhập vào biểu tượng Thực thi được hiển thị như sau. hình ảnh là tự giải thích và khớp với kết quả từ màn hình OpenOffice.



Chỉ cần một vài gợi ý. Đầu tiên, hình ảnh kết quả sẽ xuất hiện trong một ngăn trong RStudio, không phải trong Rattle, vì vậy đừng mong đợi kết quả ở vị trí đó. Thứ hai, con số cho số cơn lốc xoáy ở Connecticut bị cắt bỏ; nó phải là 3 và sẽ có một biểu đồ khác theo sau biểu đồ này sẽ hiển thị tổng số, biểu đồ này sẽ chứng minh rằng 3 xuất hiện ở đây. Nếu người dùng muốn một cách khác để hiển thị kết quả, hãy sử dụng chức năng “Dot Plot” trong tab Phân phối để tạo ra kết quả này.

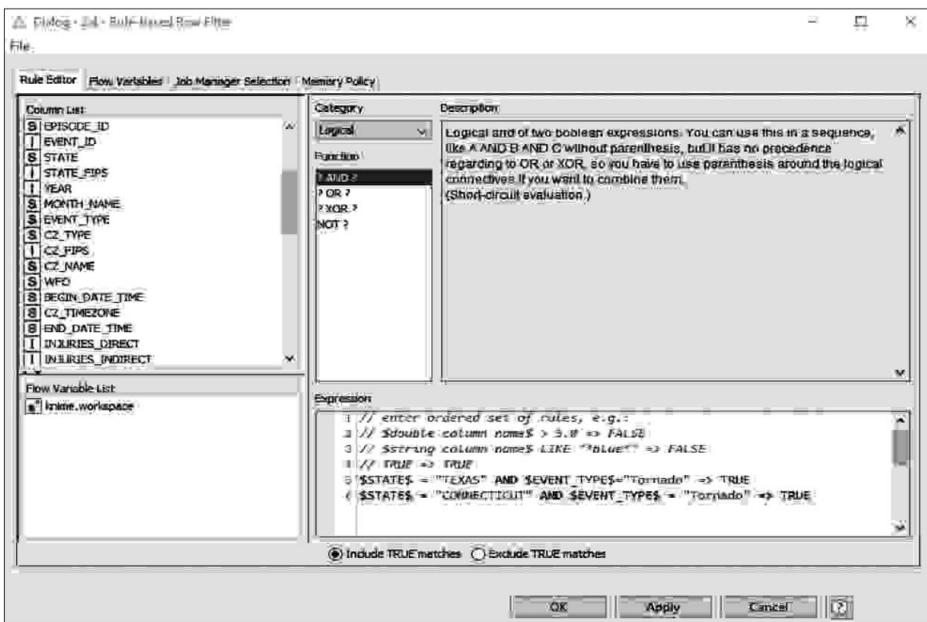


Điều này cho thấy sự chênh lệch lớn của các cơn lốc xoáy khi không sử dụng tổng số. Điều này một mình sẽ chứng minh rằng có ít lốc xoáy ở Connecticut hơn Texas (trên danh nghĩa, trong mọi trường hợp).

7.1.1.3 Trả lời Theo KNIME

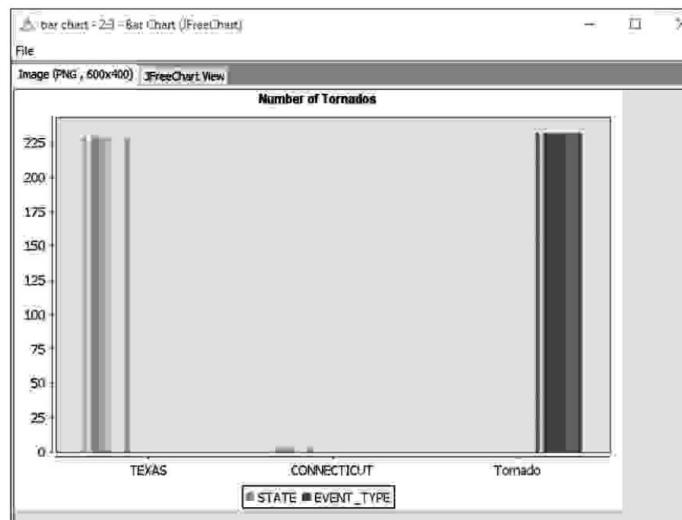
Công cụ KNIME có khả năng hiển thị kết quả phân tích với rất ít nút. Nút đầu tiên sẽ là Trình đọc CSV sẽ đọc toàn bộ tập dữ liệu vào công cụ. Nút thứ hai sẽ là nút Bộ lọc hàng để chỉ hiển thị các cột STATE và EVENT_TYPE và nút thứ ba sẽ hiển thị kết quả.

Có một nút bổ sung mà người dùng sẽ cần để thực hiện nhanh tác vụ lọc. Nút này được gọi là Bộ lọc hàng dựa trên quy tắc và sẽ yêu cầu lập trình một chút để làm cho bộ lọc hoạt động dưới dạng kết hợp của STATE và EVENT_TYPE. Cấu hình cho nút này được minh họa như sau. Việc lập trình sẽ được giải thích từng dòng một.

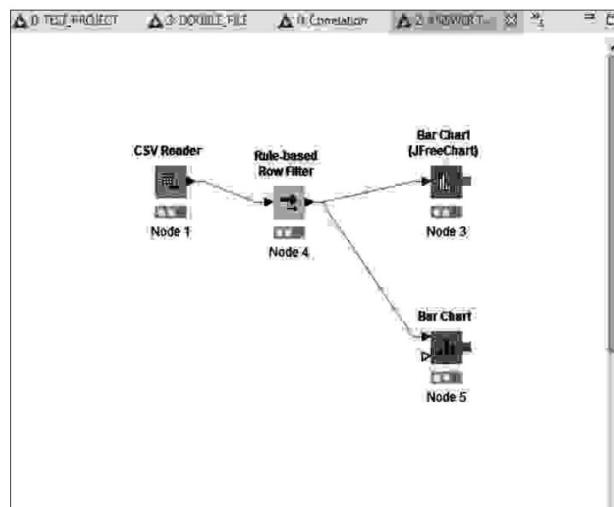


Dòng đầu tiên (dòng số 5 trong màn hình cấu hình) về cơ bản là đặt các hàng STATE thành TEXAS và EVENT_TYPE thành Tornado. Dòng thứ hai (dòng #6 trong màn hình cấu hình) đang đặt các hàng STATE thành CONNECTICUT và EVENT_TYPE thành Tornado. Bằng cách đặt các dòng này nối tiếp nhau, người dùng đang đặt AND giữa chúng và dữ liệu. Kết quả sẽ chỉ hiển thị những cơn lốc xoáy ở Texas và Connecticut.

Nút cuối cùng trong quy trình này là nút trực quan hóa để giúp trình bày kết quả. Trong trường hợp này, nút Biểu đồ thanh là hoàn toàn chấp nhận được. Có một số để lựa chọn, và hai được bao gồm trong quá trình này. Đây là kết quả của tùy chọn Biểu đồ thanh (JFree Chart). Bức ảnh đáng giá 1.000 từ.



Cột thứ ba trong biểu đồ này hiển thị EVENT_TYPE là Lốc xoáy, sẽ tính cả Texas và Connecticut. Hầu hết các công cụ hiển thị điều này như là một phần của chức năng này. Quy trình công việc đã hoàn thành được hiển thị cùng với các nút được đặt tên. Hãy nhớ rằng bộ lọc hàng sẽ chỉ yêu cầu lập trình một chút. Tuy nhiên, cũng nên nhớ rằng nhà phân tích có thể lấy bất kỳ năm theo dõi lốc xoáy nào và chạy nó qua quy trình công việc này để xác định sự khác biệt về tốc độ lốc xoáy giữa Texas và Connecticut. Khi quy trình công việc được thiết lập, không cần cấu hình thêm.



Một nhận xét cuối cùng về KNIME. Có hai nút Biểu đồ thanh, một nút được gắn nhãn là (Biểu đồ JFree) và nút còn lại đơn giản. Biểu đồ thanh đơn giản không thân thiện như Biểu đồ JFree, do đó, nhà phân tích nên sử dụng Biểu đồ JFree nếu có sự lựa chọn. Biểu đồ thanh đơn giản được bao gồm để chứng minh rằng một nút có thể cung cấp cho hai hoặc nhiều nút khác.

7.1.2 Bài tập ghép nối

Một lĩnh vực phân tích giúp xác định các thay đổi từ sự kiện này sang sự kiện khác có sẵn thông qua thử nghiệm t được gọi là mẫu ghép nối. Về bản chất, điều này đòi hỏi phải lấy một mẫu dữ liệu được ghép nối từng trường một giữa trường này và trường khác để xác định xem có thực sự có sự thay đổi giữa trường này và trường kia hay không, có tính đến cơ hội. Ví dụ: nếu nhà phân tích muốn biết liệu một cá nhân có thể nhìn rõ hơn trước và sau khi phẫu thuật đặc thù tinh thể hay không, phương pháp kiểm tra trước và sau kiểm tra sẽ giúp xác định xem điều này có đúng hay không bằng cách sử dụng kỹ thuật lấy mẫu theo cặp. Để làm điều này, bài tập sau đây trình bày dữ liệu mô phỏng cho thấy 100 sinh viên đã làm bài kiểm tra trước và sau giờ học. Mỗi học sinh được chỉ định một số (theo thứ tự) giống nhau cho mỗi bài kiểm tra. Cả hai bài kiểm tra đều có những câu hỏi giống hệt nhau, nhưng các câu trả lời được chọn ngẫu nhiên giữa bài kiểm tra trước và sau bài kiểm tra để loại bỏ việc học sinh chỉ ghi nhớ câu trả lời. Nhà phân tích phải sử dụng thử nghiệm t lấy mẫu cặp đôi để xác định câu hỏi sau: "Học sinh có thực hiện bài kiểm tra sau tốt hơn so với bài kiểm tra trước không?" Từ câu hỏi này, giả thuyết không sau đây được tạo ra:

H_0 : Điểm sau kiểm tra và trước kiểm tra của sinh viên không khác nhau

H_a : Điểm của học sinh trước kiểm tra và sau kiểm tra có sự khác biệt

Đây được gọi là bài kiểm tra "hai đầu", vì điểm sau bài kiểm tra nhỏ hơn hoặc lớn hơn điểm trước bài kiểm tra không quan trọng. Đây là một phương pháp thử nghiệm đơn giản hơn và đây sẽ là tùy chọn ưu tiên cho thử nghiệm cụ thể này.

Văn bản cho bài tập này nằm ở đây, bao gồm hai cột cần thiết cho bài kiểm tra. Nhà phân tích có thể sử dụng hướng dẫn nhập cho từng công cụ để đưa dữ liệu vào các ứng dụng phân tích khác nhau. Xin lưu ý rằng có các dấu phẩy được bao gồm để tạo một tệp giá trị được phân tách bằng dấu phẩy sẽ giúp nhập vào các công cụ khác nhau. Nếu nhà phân tích muốn tải xuống dữ liệu, họ có thể làm như vậy từ trang web của tác giả tại www.greectech.com.

Nhà phân tích sẽ điều hướng đến trang tải xuống và sẽ thấy tệp thích hợp được liệt kê trong "Sách Bài tập 2".

Sinh viên, Kiểm tra trước, Kiểm tra sau

	85,	82	39,	88,	82	77, 87, 78,	87
1,	92,	92	40,	61,	90	58, 79, 80,	82
2,	77,	82	41,	99,	99	80, 69, 81,	82
3,	64,	79	42,	63,	78	66, 82, 94,	73
4,	69,	70	43,	75,	91	97, 83, 84,	94
5,	57,	89	44,	75,	93	94, 85, 82,	89
	84,	79	45,	65,	88	86, 79,	95
6,	54,	70	46,	76,	99	87, 56, 88,	90
7,	65,	97	47,	95,	75	96, 89, 90,	94
8,	53,	88	48,	89,	88	89, 90,	87
9,	86,	80	49,	89,	77	50, 91, 72,	74
10,	54,	82	50,	53,	78	92, 93, 86,	94
11,	92,	90	51,	87,	97	67, 94, 95,	85
12,	78,	99	52,	53,	79	80, 96, 83,	72
13,	94,	89	53,	96,	89	63, 97,	98
14,	77,	95	54,	93,	83	99, 50, 98,	80
15,	70,	79	55,	73,	71	82, 100, 88,	88
16,	96,	98	56,	81,	93		99
17,	80,	86	57,	96,	98		78
18,	87,	81	58,	85,	81		75
19,	89,	99	59,	75,	92		84
20,	71,	81	60,	71,	94		90
21,	92, 98		61,	58,	91		97
22,	53, 97		62,	96,	72		76
23,	100, 99		63,	55,	77		
24,	54, 89		64,	90,	73		
25,	62,	94	65,	66,	74		
26,	65,	76	66,	87,	79		
27,	91,	92	67,	91,	90		
28,	76,	71	68,	85,	86		
29,	74,	84	69,	66,	95		
30,	95,	99	70,	80,	81		
31,	73,	83	71,	84,	89		
32,	50,	91	72,	55,	70		
33,	85,	81	73,	70,	88		
34,	73,	87	74,	59,	78		
35,	100, 91		75,	91,	83		
36, 37, 38,	52, 88		76,	95,	92		

7.1.2.1 Trả lời cho bài tập 2 - Rattle

Câu trả lời cho bài tập này có thể được giải trong Rattle chỉ với một chút nỗ lực.

Bước đầu tiên sẽ là nhập dữ liệu, là một tệp văn bản, vào Rattle bằng cách sử dụng tab "DATA" và sau đó chuyển sang "EXPLORE" để tiến hành kiểm tra t.

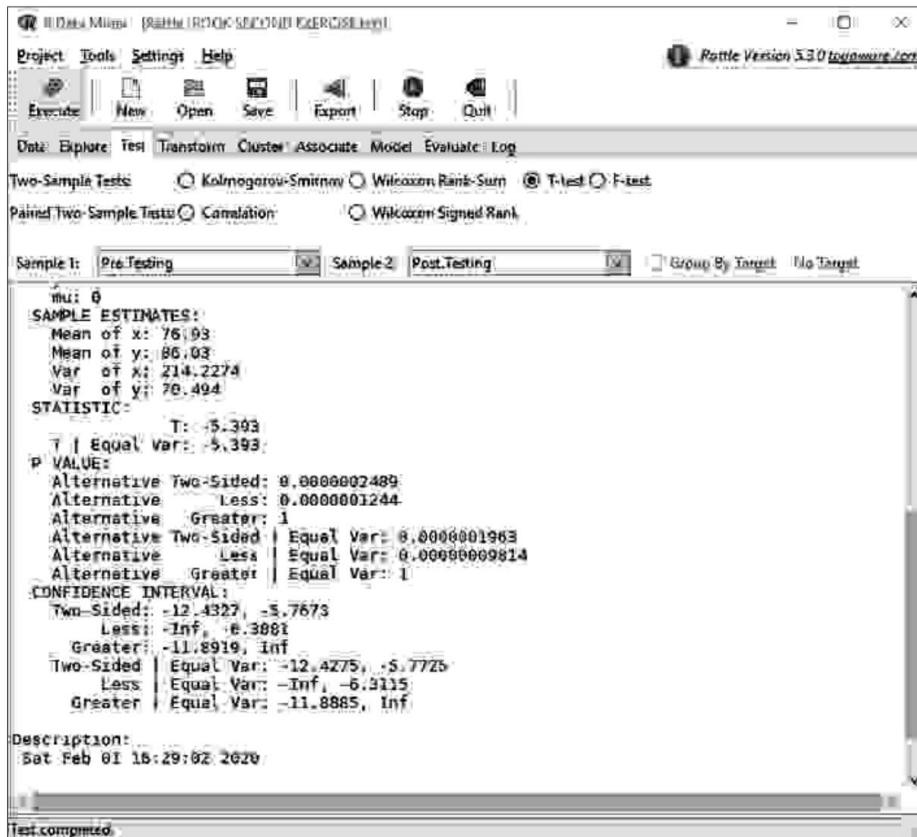
Cấu hình hoàn chỉnh của sự cố nằm trong hình minh họa tiếp theo. Quốc gia giải thích cho màn hình này sẽ lấy từng phần và mô tả từng phần cho nhà phân tích. Hãy nhớ rằng lý do chính của bài kiểm tra là để xác định xem trên thực tế có sự khác biệt nào giữa bài kiểm tra thứ nhất và thứ hai hay không. Ngoài ra, hãy nhớ rằng nhà phân tích muốn có một mẫu, trong trường hợp này được đặt bằng Rattle thông qua tab DỮ LIỆU, sử dụng cài đặt "phân vùng" ở mức 50/25/25, có nghĩa là 50% dữ liệu sẽ được lấy mẫu và kiểm tra.

Khu vực chính để tập trung vào kết quả này là ở giữa màn hình, nơi hiển thị giá trị p cho các thử nghiệm "đuôi" khác nhau. Đối với những người không quen thuộc với thống kê, các thử nghiệm "theo đuôi" để cập đến việc liệu nhà phân tích có đang thử nghiệm nếu giá trị trung bình của một mẫu nhỏ hơn hoặc lớn hơn giá trị trung bình của mẫu kia hay không. Trong trường hợp này, mẫu đầu tiên (kiểm tra trước) sẽ nhỏ hơn mẫu thứ hai (kiểm tra sau) nếu đó là kiểm tra bên phải và kiểm tra bên trái cho trường hợp ngược lại (kiểm tra đầu tiên). (mẫu lớn hơn mẫu thứ hai). Điều đó có nghĩa là gì? Nó biểu thị rằng giả thuyết thay thế khẳng định rằng, nếu trung bình mẫu thứ nhất không bằng trung bình mẫu thứ hai, thì mẫu thứ nhất nhỏ hơn mẫu thứ hai. Trong trường hợp này, nhà phân tích muốn biết liệu kết quả sau kiểm tra có lớn hơn kết quả trước kiểm tra hay không. Điều này sau đó sẽ khẳng định rằng có sự chuyển giao kiến thức và bài kiểm tra sau cho thấy học sinh thực sự đã học tài liệu mà họ không biết trong bài kiểm tra trước (nói chung). Để cụ thể hơn, mỗi câu hỏi giữa các sinh viên có thể được chạy qua bài kiểm tra này để xem liệu có sự khác biệt đáng kể giữa bài kiểm tra trước và bài kiểm tra sau hay không để xem liệu người hướng dẫn có thực sự nâng cao kiến thức của học sinh hay không.

Điều này rất quan trọng đối với người hướng dẫn, vì nó cho biết nội dung đó giúp ích hay cản trở sinh viên. Không người hướng dẫn nào muốn thấy rằng sinh viên học ít hơn trong lớp học của họ.

Nếu nhà phân tích muốn biết xác suất của bài kiểm tra trước ít hơn bài kiểm tra sau, tính đến một sự kiện ngẫu nhiên mà bài kiểm tra trước có thể ghi nhận nhiều hơn bài kiểm tra sau, thì họ cần xem xét "P -VALUE" của màn hình trong phần "Alternative" và "Less" cho số này: .0000001244. Điều này có nghĩa là về cơ bản không có khả năng kiểm tra trước ít hơn kiểm tra sau với mức độ tin cậy 95%. Điều đó cũng có nghĩa là, khi xem xét rằng cơ hội không phải là một yếu tố, thì các giá trị trước thử nghiệm tổng thể sẽ ít hơn so với các giá trị sau thử nghiệm.

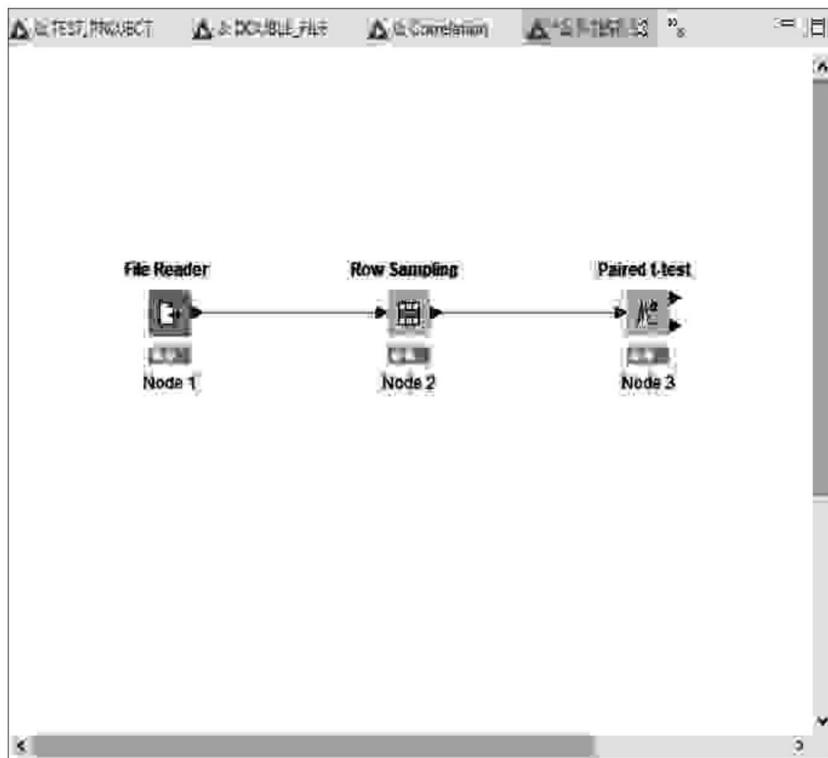
Một cách khác để xem xét điều này thông qua công cụ này là khoảng tin cậy, nằm trên màn hình tại khu vực “Khoảng tin cậy”. Nếu nhà phân tích xem “Ít hơn:” thì họ sẽ thấy rằng khoảng ở mức 95% là Infinity (Inf) đến 6,3081. Nếu nhà phân tích xem xét kỹ lưỡng, họ sẽ thấy rằng “0” không được bao gồm trong phạm vi đó, điều đó có nghĩa là các phương tiện trước và sau kiểm tra không bao giờ bằng 0, cho thấy thực tế là trước kiểm tra nhỏ hơn sau khi kiểm tra Bài kiểm tra. Đây là một yếu tố nữa trong bài kiểm tra thống kê tổng thể và một yếu tố, như đã nêu trước đó trong cuốn sách này, là yếu tố quan trọng trong phân tích dữ liệu tổng thể.



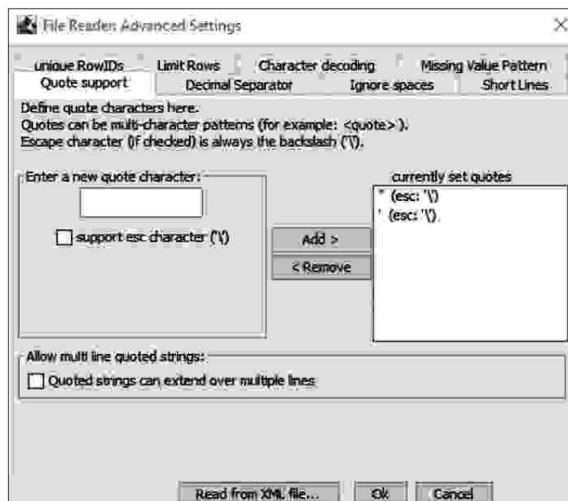
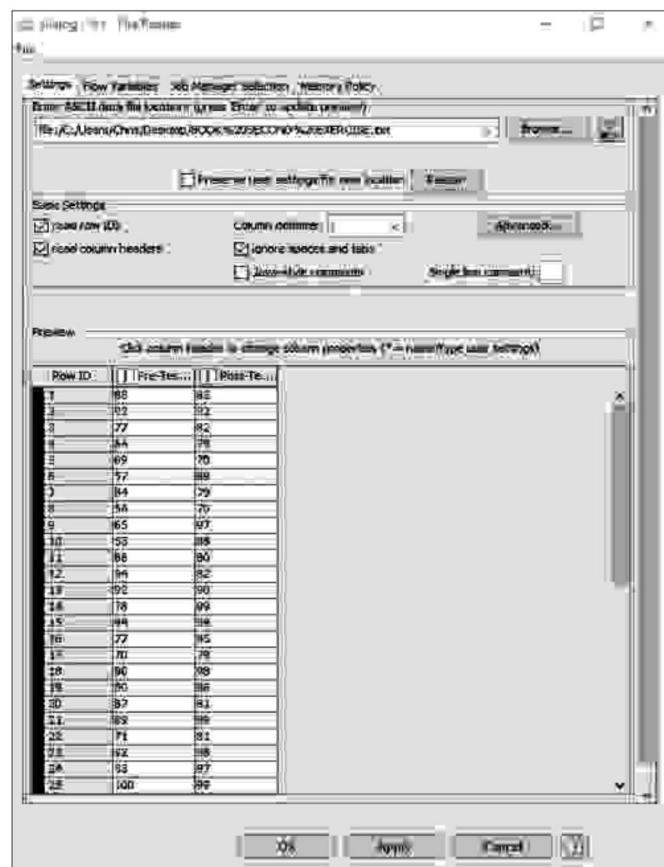
7.1.2.2 Đáp án bài tập 2 – KNIME

KNIME có một nút sẽ hỗ trợ vấn đề này, nhưng trước tiên, nhà phân tích phải nhập dữ liệu vào công cụ và làm việc với dữ liệu. Thao tác này được thực hiện giống như trong các phần trước, nhưng lần này thay vì sử dụng Trình đọc CSV

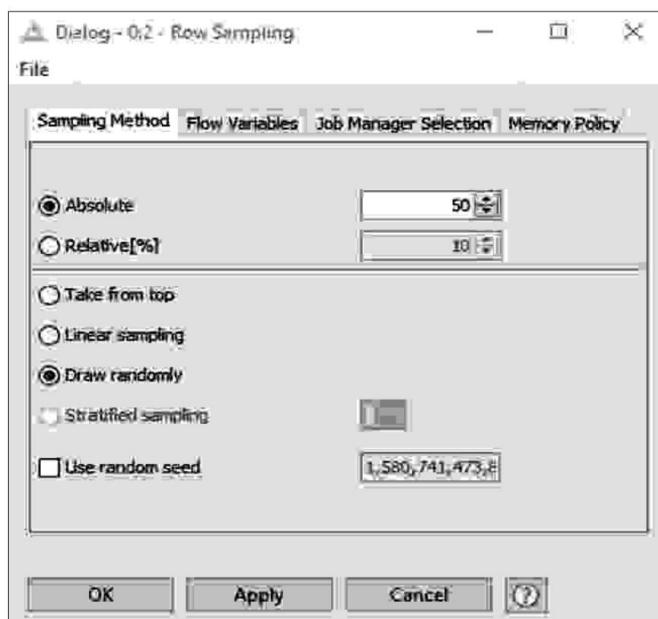
nút, nhà phân tích sử dụng nút Trình đọc tệp, nút này sẽ đọc văn bản từ tệp văn bản. Nhà phân tích có thể sử dụng Trình đọc CSV cho việc này, nhưng việc sử dụng Trình đọc tệp sẽ dễ dàng hơn vì khả năng dấu phân cách là thứ gì đó không phải là dấu phẩy. Quy trình công việc được hiển thị trong màn hình sau và sẽ được giải thích theo từng nút để đảm bảo nhà phân tích hiểu được cả quy trình và các nút khác nhau.



Nút đầu tiên là nút Trình đọc tệp, có màn hình cấu hình như hình. Lưu ý rằng các hộp kiểm bao gồm những hộp kiểm tương tự với các nút khác. Ngoài ra còn có một khôi "Nâng cao." có thể được chọn và nó đã được đưa vào để nhà phân tích khám phá khi cần thiết.

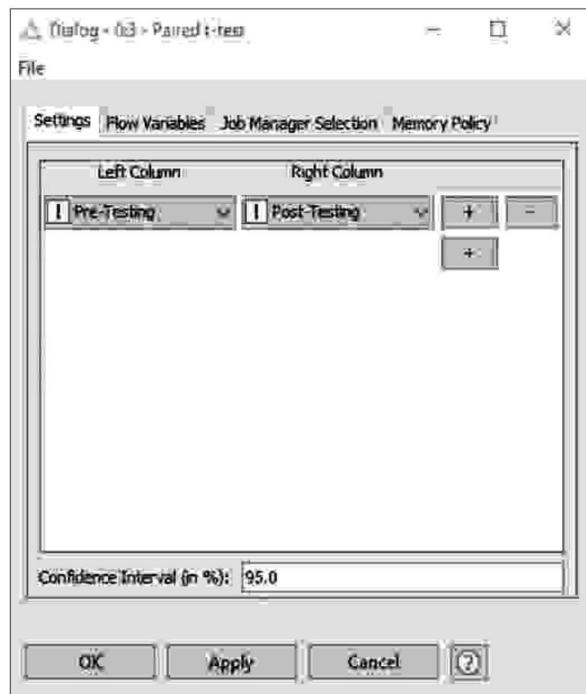


Nút thứ hai là nút **Lấy mẫu hàng**, đã được sử dụng trước đây trong phần lấy mẫu ngẫu nhiên. Nút này sẽ lấy mẫu các hàng cần thiết để kiểm tra hợp lệ. Công thức "sức mạnh" chưa được sử dụng trong trường hợp này, vì vậy 50 hàng ban đầu đã được lấy mẫu, điều đó có nghĩa là kết quả của thử nghiệm này không khớp với kết quả từ các câu trả lời trước đó (vì chúng là các hàng được lấy mẫu ngẫu nhiên). Cấu hình cho màn hình này như sau:



Nút cuối cùng là nút **kiểm tra t ghép nối**, nút này sẽ thực hiện các tính toán cần thiết cho kết quả. Hãy nhớ nhấp vào mũi tên màu xanh lá cây Thực thi sau mỗi lần thay đổi cấu hình. Đừng lo lắng nếu bạn quên, vì ứng dụng sẽ nhắc nhở nhà phân tích rằng có điều gì đó đã thay đổi sẽ ảnh hưởng đến quy trình làm việc.

Kết quả xuất hiện sau màn hình này và nhà phân tích nên lưu ý các giá trị p được đánh kèm cùng với khoảng tin cậy, như đã giải thích trong phần trước. Điều cần thiết là nhà phân tích phải xem xét các khoảng tin cậy, vì đó là những loại thử nghiệm đơn giản để thực hiện và tạo ra kết quả hiệu quả như các thử nghiệm thống kê phức tạp hơn. Hãy nhớ rằng phức tạp không có nghĩa là chính xác.



Paired Samples Statistics						
	Column	N	Missing Cases	Mean	Standard Deviation	Standard Error Mean
Pair 1	Pre-Testing	50	0	98.98	14.1506	2.0265
Pair 1	Post-Testing	50	0	84.54	8.505	1.2028

Paired Samples Test						
Confidence Interval (CI) Probability: 95.0%						
	Effect	t	Df	p-value (2-tailed)	Mean	Standard Deviation
Pair 1	Pre-Testing Post-Testing	3.9032	49	0.001	14.1506	15.2505

Xem nhanh các kết quả này cho thấy giá trị p là 0,001 và khoảng tin cậy (CI) nằm trong khoảng từ 11 đến 3. Điều này có nghĩa là "0" không được bao gồm trong phạm vi đó, biểu thị rằng các phương tiện không giống nhau. Nó cũng chỉ ra rằng giới hạn trên là 3, có nghĩa là kiểm tra trước nhỏ hơn kiểm tra sau, điều này cũng cực kỳ có giá trị đối với nhà phân tích và có thể gợi ý một số

kiểm tra bổ sung. Nhà phân tích có thể thực hiện các thử nghiệm khác giúp xác định bất kỳ sự khác biệt cụ thể nào, nhưng ít nhất nó cũng tạo ra kết quả rằng trên thực tế có sự khác biệt giữa các phương tiện trước và sau thử nghiệm.

NGƯỜI GIỚI THIỆU

Levene, H. (1960). Đóng góp cho Xác suất và Thống kê: Tiểu luận Vinh danh Harold Hotelling (Ingram Olkin, Ed.). Thành phố Redwood, CA: Nhà xuất bản Đại học Stanford.

Poundstone, W. (2019). Phép tính Ngày tận thế: Phương trình Dự đoán Tương lai đang Biến đổi Mọi thứ Chúng ta Biết về Sự sống và Vũ trụ như thế nào. New York, NY: Tập đoàn sách Hachette.

Provost, F., và Fawcett, T. (2013). Khoa học dữ liệu cho doanh nghiệp: Những điều bạn cần biết về khai thác dữ liệu và tư duy phân tích dữ liệu. Sebastopol: O'Reilly.

Reinhart, A. (2015). Số liệu thống kê đã thực hiện sai: Hướng dẫn đầy đủ đáng tiếc. San Francisco, CA: Không có máy ép tinh bột.

Sankhar, A. (2018, ngày 30 tháng 11). Cách tạo WordCloud trong Đào tạo R. Analytics. <https://analyticstraining.com/how-to-create-a-word-cloud-in-r/>.

Công ty TNHH tư vấn thống kê (2011, ngày 14 tháng 5). Phát hiện gian lận kế toán và luật của Benford. Công ty TNHH tư vấn thống kê www.statisticalconsultants.co.nz/blog/benfords-law-and-accounting-fraud-detection.html

Công nghệ, NI (2013, 30 tháng 10). Thử nghiệm Levene về sự bình đẳng của phương sai. Sổ tay thống kê kỹ thuật. <https://www.itl.nist.gov/div898/sotay/index.htm>.

Williams, G. (2011). Khai thác dữ liệu với Rattle và R. New York, NY: Springer Khoa học + Truyền thông kinh doanh.

MỤC LỤC

MỘT

- Giá trị alpha, 122
- ToolPak phân tích, 7-9, 35, 179-180
- Lưu trữ, 3
- Công thức mảng, 112

b

- Định luật Benford, Rattle, 151-157

c

- Tệp Giá trị được Phân tách bằng Đầu phẩy (CSV), 5
- Mạng lưu trữ R toàn diện (CRAN), 11-12, 178
- Khoảng tin cậy, 117-118
 - Excel, 119-121
 - KNIME, 124-127
 - OpenOffice, 121-122
 - R/RStudio/Rattle, 122-124
- Tương quan, 103
 - Excel, 103-105
 - KNIME, 108-109
 - OpenOffice, 105-106
 - R/RStudio/Rattle, 106-108
- Đo lường tương quan, 108
- Giá trị tương quan, 109
- CRAN. Xem Mạng lưu trữ R toàn diện

tệp CSV. Xem các tệp Giá trị được Phân tách bằng Đầu phẩy

Biểu đồ xác suất tích lũy, 52

- Excel, 52-56
- KNIME, 67-91
- OpenOffice, 56-66
- R/RStudio/Rattle, 67-72

đ.

Phân tích dữ liệu, 3

- Khoa học dữ liệu, 3
- Công cụ dữ liệu, 1-2
- Trang web dữ liệu, 3-4
- Biến phụ thuộc, 110
- Thống kê mô tả, 182
 - Excel, 35-39
 - KNIME, 48-52
 - OpenOffice, 39-42
 - RStudio/Rattle, 42-47
- Biến rỗi rạc, 182 dplyr, 174

e

Excel, 5-7, 35-39

- ToolPak phân tích, 7-9, khoảng tin cậy 179-180, tương quan 119-121, biểu đồ xác suất tích lũy 103-105, thống kê mô tả 52-56, lọc 35-39, 171-173

- Thử nghiệm F,
hồi quy bội/tương quan 140-142, lấy mẫu
ngẫu nhiên 145-147, hồi quy
128-129, thử nghiệm
t 110-111 (tham số), 91-93
- F
- Âm tính giả, 137
Dương tính giả, 137
"Tệp, Lô, Gói và Trợ giúp", 18
- Lọc, 170
Excel, 171-173
KNIME, 174-176
OpenOffice, 173-174
R/RStudio/Rattle, 174
- Nguồn mở và miễn phí (FOSS), 3
Kiểm tra F, 101
Excel, 140-142
KNIME, 143-145
R/RStudio/Rattle, 142-143
- g
- GGobi, 11-12
GGRaptr, 12
Giao diện người dùng đồ họa (GUI), 2, 11
-
- Nhập dữ liệu
KNIME, 24-32
Tiếng lạch cách, 18-24
R/Rattle, 11-12
RStudio, 12-17
- Nhóm độc lập t-test, 99, 100
- K
- KNIME
khoảng tin cậy, tương quan
124-127, biểu đồ xác
suất tích lũy 108-109, 67-91
- thống kê mô tả, lọc 48-52,
174-176
F-Test, 143-145
nhập dữ liệu, 24-32
llift, 157-
160 bội quy/tương quan, 150-151 lấy mẫu
ngẫu nhiên, 134-136 hồi
quy, 115-117 t-test
(tham số), 97-101
Wordcloud, 163-170
- l
- Kiểm tra Levene, 101, 140, 141, 142
thang máy, 157
KNIME, 157-160
- Hồi quy tuyến tính, 109
Đòng, 112
Mức tin cậy thấp hơn (LCL), 46, 118
- m
- Ý nghĩa, 41
Microsoft Excel 2016, 5
Hồi quy bội/tương quan
Excel, 145-147
KNIME, 150-151
OpenOffice, 147-148
R/RStudio/Rattle, 148-149
- N
- Tương quan nghịch, 103
Nút, 24
- ô
- OpenOffice, khoảng
tin cậy 9-11, tương quan 121-
122, biểu đồ xác suất
tích lũy 105-106, thống kê mô tả 56-
66, lọc 39-42, 173-174

hồi quy bội/tương quan, lấy mẫu ngẫu nhiên 147-148, hồi quy 129-132, 112-113

T-test (tham só), 93-95

P

Gói, 177-179

Thời gian nghỉ có lương (PTO), 52

Biểu đồ Pareto, 52

Phương pháp tương quan Pearson, 107, 109

Tương quan dương, 103

Sau khi kiểm tra, cho hiệu suất của học sinh, 194-201

Quyền lực, 137

R/RStudio/Rattle, 138-139

Kiểm tra trước, cho hiệu suất của học sinh, 194-202

PTO. Xem Thời gian nghỉ có lương

I

Lấy mẫu ngẫu nhiên, 127-128

Excel, 128-129

KNIME, 134-136

OpenOffice, 129-132

R/RStudio/Rattle, 132-134

Tiếng lạch cách, 18-24

Định luật Benford, 151-157

Nhập khẩu lực lạc, 18-24
nhập dữ liệu, gói 18-24,
178

Nhập khẩu lực lạc, 18-24

Gói "RColorBrewer", 161

Hồi quy, 109-110

Excel, 110-111

KNIME, 115-117

OpenOffice, 112-113

R/RStudio/Rattle, 113-115

Reinhart, Alex, 117-118

Biến phản hồi, 110

R/RStudio, nhập dữ liệu, 11-12

R/RStudio, Wordcloud, 160-162

R/RStudio/Rattle

khoảng tin cậy, tương quan 122-

124, biểu đồ xác suất

tích lũy 106-108, lọc 67-72, 174

Kiểm tra F, hồi

quy bội/tương quan 142-143, lũy thừa 148-149,

lấy mẫu ngẫu nhiên

138-139, hồi quy 132-134, kiểm
tra t 113-115 (tham
số), 96-97

RStudio

nhập dữ liệu, gói 12-17,
178-179

RStudio/Rattle, thống kê mô tả, 42-47

S

Phản mềm, 1-2

Thống kê sai (Reinhart), 117-118

Tiếp cận đèn giao thông, 32-33

t

Thé dám mây, 162, 167

Các bài kiểm tra "Tailed", 196, 197

Tibble, 152

Gói "Tm", 161

Lốc xoáy ở Texas và Connecticut

bài tập, câu trả lời

181-182 của KNIME, câu trả lời

191-194 của OpenOffice, câu trả lời

183-188 của Rattle, bài tập

lấy mẫu theo cặp 188-191, 194

câu trả lời của KNIME, 197-202

câu trả lời của Rattle, 196-

197 Theo dõi cơn lốc xoáy, 16, 106, 108, 119,
128, 138, 140, 143,

160 T-test, 95, 97, 140, 194-202. Xem thêm tham số

F-Test Excel,

91-93 KNIME,

97-101 OpenOffice,

93-95 R/RStudio/Rattle,

96-97

Nút kiểm tra T, 97, 99

V

Thử nghiệm “hai đuôi”, 194

Mạng riêng ảo (VPN), 12

Lỗi loại 1, 137

W

Lỗi loại 2, 137

bạn

Đám mây từ, 160

Mức tin cậy trên (UCL), 46, 118

KNIME, 163-170

R/RStudio, 160-163

Từ ngữ, 167