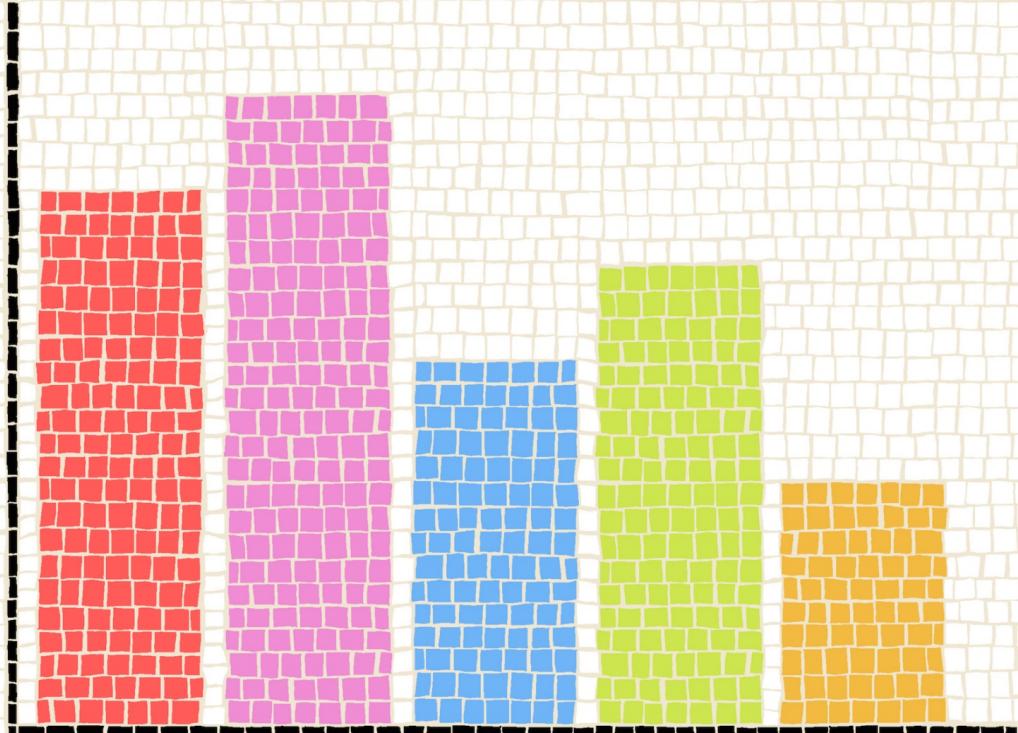


The Art of Data Science

A Guide for Anyone Who Works with Data



Roger D. Peng & Elizabeth Matsui

Nghệ thuật Khoa học Dữ liệu

Hướng dẫn cho bất kỳ ai làm việc với dữ liệu

Roger D. Peng và Elizabeth Matsui

Cuốn sách này được bán
tại <http://leanpub.com/artofdatascience>

Phiên bản này đã được xuất bản vào ngày 20-07-2016



Đây là một [Leanpub](#) sách. Leanpub trao quyền cho các tác giả và nhà xuất bản bằng quy trình Xuất bản Tinh gọn. [xuất bản tinh gọn](#) là hành động xuất bản sách điện tử đang trong quá trình sử dụng các công cụ nhẹ và nhiều lần lặp lại để nhận phản hồi của người đọc, xoay vòng cho đến khi bạn có cuốn sách phù hợp và tạo lực kéo sau khi bạn thực hiện.

© 2015 - 2016 Skybrude Consulting, LLC

Cũng bởi Roger D. Peng

Lập trình R cho Khoa học dữ liệu

Phân tích dữ liệu khám phá với R

Khoa học dữ liệu điều hành

Viết báo cáo cho Khoa học dữ liệu trong R

Cuộc trò chuyện về Khoa học dữ liệu

Đặc biệt cảm ơn Maggie Matsui, người đã tạo ra tất cả các tác phẩm nghệ thuật cho cuốn sách này.

nội dung

1. Phân tích dữ liệu như một nghệ thuật	1
2. Chu kỳ Phân tích	4
2.1 Đặt bối cảnh 2.2	5
Chu kỳ phân tích	6
2.3 Đặt kỳ vọng	10
2.4 Thu thập thông tin 2.5	9
So sánh kỳ vọng với dữ liệu 2.6 Áp dụng quy trình phân tích sử thi	11
3. Nêu và Hoàn thiện Câu hỏi	16
3.1 Các loại câu hỏi	16
3.2 Áp dụng Vòng tuần hoàn để Phát biểu và Tinh chỉnh Câu hỏi của Bạn.	20
3.3 Đặc điểm của một câu hỏi hay	20
3.4 Chuyển câu hỏi thành vấn đề dữ liệu 23	
3.5 Nghiên cứu tình huống . 26	
3.6 Suy nghĩ Kết luận . 30	
4. Phân tích dữ liệu khám phá	31
4.1 Danh sách kiểm tra phân tích dữ liệu khám phá: Một trường hợp Học . 33	
4.2 Xây dựng câu hỏi của bạn. 33	
4.3 Đọc dữ liệu của bạn 35	
4.4 Kiểm tra Bao bì . 36	
4.5 Nhìn vào phần trên cùng và phần dưới cùng của dữ liệu của bạn 39	

NỘI DUNG

4.6 ABC: Luôn kiểm tra các chữ "n" của bạn.	40
4.7 Xác thực với ít nhất một dữ liệu ngoài Nguồn	45
4.8 Tạo một cốt	46
truyện 4.9 Trước tiên hãy thử giải pháp dễ dàng	49
4.10 Các câu hỏi tiếp theo	53
5. Sử dụng các Mô hình để Khám phá Dữ liệu	55
của Bạn 5.1 Các Mô hình như Kỳ vọng.	57
5.2 So sánh Kỳ vọng của Mô hình với Thực tế .	60
5.3 Phản ứng với dữ liệu: Tinh chỉnh các kỳ vọng của chúng ta	64
5.4 Kiểm tra các mối quan hệ tuyến tính .	67
5.5 Khi Nào Chúng Ta Dừng Lại? .	73
5.6 Tóm tắt .	77
6. Suy luận: A Primer	78
6.1 Xác định dân số	78
6.2 Mô tả quá trình lấy mẫu.	79
6.3 Mô tả một mô hình cho dân số	79
6.4 Ví dụ nhanh	80
6.5 Các yếu tố ảnh hưởng đến chất lượng suy luận	84
6.6 Ví dụ: Sử dụng Apple Music	86
6.7 Quần thể có nhiều dạng	89
7. Mô hình hóa chính thức	92
7.1 Mục tiêu của mô hình hóa chính thức là gì? .	92
7.2 Khuôn khổ chung	93
7.3 Phân tích liên kết	95
7.4 Phân tích dự đoán	104
7.5 Tóm tắt	111
8. Suy luận so với Dự đoán: Ý nghĩa đối với Chiến lược điều chỉnh .	112
8.1 Ô nhiễm không khí và Tử vong ở Thành phố New York	113
8.2 Suy ra một Hiệp hội .	115
8.3 Dự Đoán Kết Quả	121

NỘI DUNG

8.4 Tóm tắt	123
9. Diễn giải kết quả của bạn	124
9.1 Nguyên tắc diễn giải	124
9.2 Nghiên cứu điển hình: Tiêu thụ soda không ăn kiêng và Chỉ số khói cơ thể	125
10. Giao tiếp	144
10.1 Giao tiếp thông thường	144
10.2 Đói tượng 10.3	146
Nội dung 10.4	148
Phong cách	151
10.5 Thái độ	151
11. Suy Nghĩ Kết Luận	153
12. Về Tác Giả	155

1. Phân tích dữ liệu như một nghệ thuật

Phân tích dữ liệu rất khó và một phần của vấn đề là ít người có thể giải thích cách thực hiện. Không phải là không có người thường xuyên phân tích dữ liệu. Đó là những người thực sự giỏi về nó vẫn chưa khai sáng cho chúng ta về quá trình suy nghĩ diễn ra trong đầu họ.

Hãy tưởng tượng bạn đang hỏi một nhạc sĩ cách cô ấy viết các bài hát của mình. Có rất nhiều công cụ mà cô ấy có thể vẽ. Chúng ta có một sự hiểu biết chung về cấu trúc của một bài hát hay: nó dài bao nhiêu, bao nhiêu câu, có thể có một câu tiếp theo là điệp khúc, v.v. Nói cách khác, có một khuôn khổ trừu tượng cho các bài hát nói chung.

Tương tự như vậy, chúng ta có lý thuyết âm nhạc cho chúng ta biết rằng sự kết hợp nhất định giữa các nốt và hợp âm hoạt động tốt với nhau và những sự kết hợp khác nghe không hay. Cuối cùng, dù những công cụ này có tốt đến đâu thì kiến thức về cấu trúc bài hát và nhạc lý thôi cũng không thể tạo nên một bài hát hay. Một số thứ khác là cần thiết.

Trong bài tiểu luận nổi tiếng năm 1974 của Donald Knuth [Lập trình máy tính coi như](#), Knuth nói về sự khác biệt giữa nghệ thuật [một nghệ thuật](#)¹ và khoa học. Trong bài luận đó, anh ấy đã cố gắng truyền đạt ý tưởng rằng mặc dù lập trình máy tính liên quan đến các máy phức tạp và kiến thức rất kỹ thuật, nhưng hành động viết một chương trình máy tính cũng có một yếu tố nghệ thuật. Trong bài luận này, ông nói rằng

Khoa học là kiến thức mà chúng ta hiểu rõ đến mức có thể dạy nó cho máy tính.

¹ <http://www.paulgraham.com/knuth.html>

Mọi thứ khác là nghệ thuật.

Tại một thời điểm nào đó, nhạc sĩ phải đưa một tia sáng tạo vào quy trình để tập hợp tất cả các công cụ sáng tác lại với nhau để tạo ra thứ gì đó mà mọi người muốn nghe. Đây là một phần quan trọng của nghệ thuật sáng tác bài hát. Tia sáng tạo đó rất khó diễn tả, càng khó viết ra, nhưng rõ ràng nó rất cần thiết để viết nên những bài hát hay. Nếu không, thì chúng ta sẽ có các chương trình máy tính thường xuyên viết các bài hát ăn khách. Dù tốt hay xấu, điều đó vẫn chưa xảy ra.

Giống như sáng tác nhạc (và lập trình máy tính), điều quan trọng là phải nhận ra rằng phân tích dữ liệu là một nghệ thuật. Nó vẫn chưa phải là thứ mà chúng ta có thể dạy cho máy tính.

Các nhà phân tích dữ liệu có nhiều công cụ tùy ý sử dụng, từ hồi quy tuyến tính đến cây phân loại và thậm chí cả học sâu, và những công cụ này đều đã được dạy cẩn thận cho máy tính.

Nhưng cuối cùng, một nhà phân tích dữ liệu phải tìm cách tập hợp tất cả các công cụ và áp dụng chúng vào dữ liệu để trả lời một câu hỏi có liên quan—một câu hỏi mà mọi người quan tâm.

Thật không may, quá trình phân tích dữ liệu không phải là quá trình mà chúng tôi có thể viết ra một cách hiệu quả. Đúng là có rất nhiều sách giáo khoa thống kê ngoài kia, nhiều cuốn nằm trên kệ của chúng tôi. Nhưng theo ý kiến của chúng tôi, không có cái nào trong số này thực sự giải quyết các vấn đề cốt lõi liên quan đến việc tiến hành phân tích dữ liệu trong thế giới thực. Năm 1991, Daryl Pregibon, một nhà thống kê nổi tiếng trước đây của AT&T Research và bây giờ là của Google, [đã nói về quá trình phân tích dữ liệu](#)² rằng “các nhà thống kê có một quy trình mà họ tán thành nhưng không hiểu đầy đủ”.

Mô tả phân tích dữ liệu trình bày một câu hỏi hóc búa khó khăn.

Một mặt, việc phát triển một khuôn khổ hữu ích liên quan đến việc mô tả đặc điểm của các yếu tố phân tích dữ liệu bằng cách sử dụng tóm tắt

²<http://www.nap.edu/catalog/1910/the-future-of-statistical-software-process-of-a-forum>

ngôn ngữ để tìm ra những điểm tương đồng giữa các loại phân tích khác nhau. Đôi khi, ngôn ngữ này là ngôn ngữ của toán học. Mặt khác, chính các chi tiết của một bài phân tích thường làm cho mỗi bài trở nên khó khăn và thú vị. Làm thế nào một người có thể khái quát hóa một cách hiệu quả qua nhiều phân tích dữ liệu khác nhau, mỗi phân tích đều có những khía cạnh đặc đáo quan trọng?

Những gì chúng tôi đặt ra trong cuốn sách này là viết ra quá trình phân tích dữ liệu. Những gì chúng tôi mô tả không phải là một “công thức” cụ thể để phân tích dữ liệu—đại loại như “áp dụng phương pháp này và sau đó chạy thử nghiệm đó”—mà là một quy trình chung có thể được áp dụng trong nhiều tình huống khác nhau. Thông qua kinh nghiệm sâu rộng của chúng tôi trong cả việc quản lý các nhà phân tích dữ liệu và tiến hành phân tích dữ liệu của riêng mình, chúng tôi đã quan sát cẩn thận điều gì tạo ra kết quả nhất quán và điều gì không tạo ra thông tin chi tiết hữu ích về dữ liệu. Mục tiêu của chúng tôi là viết ra những gì chúng tôi đã học được với hy vọng rằng những người khác có thể thấy nó hữu ích.

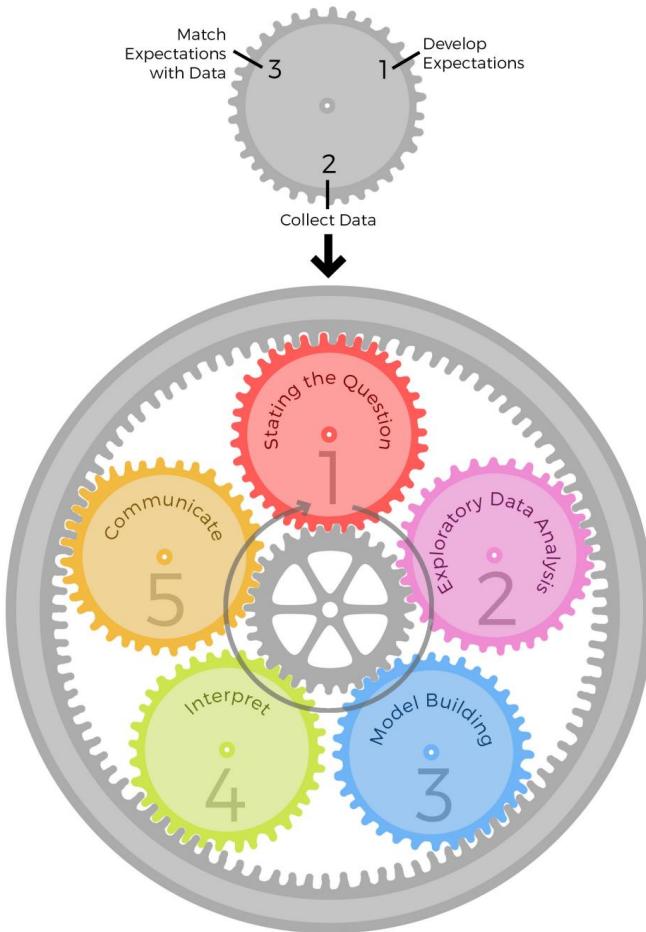
2. Chu kỳ Phân tích

Đối với những người không quen biết, một phân tích dữ liệu có thể tuân theo một quy trình tuyến tính, từng bước một mà cuối cùng, sẽ cho ra một kết quả mạch lạc và được đóng gói độc đáo. Trên thực tế, phân tích dữ liệu là một quá trình lặp đi lặp lại nhiều lần và phi tuyến tính, được phản ánh tốt hơn bởi một loạt các ngoại luân (xem Hình), trong đó thông tin được học ở mỗi bước, sau đó cho biết liệu (và cách thức) tinh chỉnh và làm lại, bước vừa được thực hiện hoặc liệu (và cách thức) tiến hành bước tiếp theo.

Ngoại luân là một vòng tròn nhỏ có tâm di chuyển xung quanh chu vi của một vòng tròn lớn hơn. Trong phân tích dữ liệu, quy trình lặp được áp dụng cho tất cả các bước của phân tích dữ liệu có thể được coi là một ngoại luân được lặp lại cho mỗi bước dọc theo chu vi của toàn bộ quy trình phân tích dữ liệu. Một số phân tích dữ liệu đường như là cố định và tuyến tính, chẳng hạn như thuật toán được nhúng vào các nền tảng phần mềm khác nhau, bao gồm cả ứng dụng. Tuy nhiên, các thuật toán này là sản phẩm phân tích dữ liệu cuối cùng xuất hiện từ công việc phi tuyến tính là phát triển và tinh chỉnh phân tích dữ liệu để nó có thể được “thuật toán hóa”.

Chu kỳ phân tích

5



Chu kỳ phân tích

2.1 Đặt bối cảnh

Trước khi đi sâu vào “vòng tròn phân tích”, nó rất hữu ích để tạm dừng và xem xét ý nghĩa của chúng tôi đối với “chì phân tích dữ liệu”. Mặc dù nhiều khái niệm chúng ta sẽ thảo luận trong phần này

Chu kỳ phân tích

6

cuốn sách này có thể áp dụng để tiến hành một nghiên cứu, khuôn khổ và các khái niệm trong cuốn sách này và các chương tiếp theo được điều chỉnh cụ thể để tiến hành phân tích dữ liệu. Trong khi một nghiên cứu bao gồm việc phát triển và thực hiện một kế hoạch thu thập dữ liệu, thì một phân tích dữ liệu cho rằng dữ liệu đã được thu thập. Cụ thể hơn, một nghiên cứu bao gồm việc phát triển một giả thuyết hoặc câu hỏi, thiết kế quy trình thu thập dữ liệu (hoặc giao thức nghiên cứu), thu thập dữ liệu, phân tích và giải thích dữ liệu. Bởi vì phân tích dữ liệu già định rằng dữ liệu đã được thu thập, nên nó bao gồm việc phát triển và sàng lọc câu hỏi cũng như quá trình phân tích và diễn giải dữ liệu. Điều quan trọng cần lưu ý là mặc dù phân tích dữ liệu thường được thực hiện mà không tiến hành nghiên cứu, nhưng nó cũng có thể được hình thành như một phần của nghiên cứu.

2.2 Chu kỳ phân tích

Có 5 hoạt động cốt lõi của phân tích dữ liệu:

1. Nêu và hoàn thiện câu hỏi
2. Khám phá dữ liệu
3. Xây dựng các mô hình thống kê chính thức
4. Diễn giải kết quả
5. Truyền đạt kết quả

5 hoạt động này có thể xảy ra ở các khoảng thời gian khác nhau: ví dụ: bạn có thể thực hiện tất cả 5 hoạt động trong vòng một ngày, nhưng cũng có thể xử lý từng hoạt động đối với một dự án lớn trong nhiều tháng. Trước khi thảo luận về các hoạt động cốt lõi này, sẽ diễn ra trong các chương sau, điều quan trọng trước tiên là phải hiểu khuôn khổ tổng thể được sử dụng để tiếp cận từng hoạt động này.

Chu kỳ phân tích

7

Mặc dù có nhiều loại hoạt động khác nhau mà bạn có thể tham gia khi thực hiện phân tích dữ liệu, mọi khía cạnh của toàn bộ quy trình đều có thể được tiếp cận thông qua một quy trình tương tác mà chúng tôi gọi là “vòng lặp phân tích dữ liệu”. Cụ thể hơn, đối với từng hoạt động trong số năm hoạt động cốt lõi, điều quan trọng là bạn phải tham gia vào các bước sau:

1. Đặt kỳ vọng, 2. Thu thập

thông tin (dữ liệu), so sánh dữ liệu với kỳ vọng của bạn và nếu kỳ vọng không khớp, 3. Sửa đổi kỳ vọng của bạn hoặc sửa dữ liệu để dữ liệu và kỳ vọng của bạn khớp với nhau.

Việc lặp đi lặp lại quy trình 3 bước này là cái mà chúng tôi gọi là “vòng tuần hoàn phân tích dữ liệu”. Khi bạn trải qua mọi giai đoạn của một phân tích, bạn sẽ cần trải qua chu trình ngoại vi để liên tục tinh chỉnh câu hỏi, phân tích dữ liệu khám phá, mô hình chính thức, diễn giải và giao tiếp của bạn.

Chu trình lặp đi lặp lại qua từng hoạt động trong số năm hoạt động cốt lõi này được thực hiện để hoàn thành phân tích dữ liệu tạo thành vòng phân tích dữ liệu lớn hơn (Xem Hình). Trong chương này, chúng tôi đi vào chi tiết về quá trình 3 bước tuần hoàn này là gì và đưa ra các ví dụ về cách bạn có thể áp dụng nó vào phân tích dữ liệu của mình.

Chu kỳ phân tích

16

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Chu kỳ phân tích

2.3 Đặt kỳ vọng

Phát triển kỳ vọng là quá trình suy nghĩ có chủ ý về những gì bạn mong đợi trước khi bạn làm bất cứ điều gì, chẳng hạn như như kiểm tra dữ liệu của bạn, thực hiện một thủ tục hoặc nhập lệnh. Đối với các nhà phân tích dữ liệu có kinh nghiệm, trong một số trường hợp, phát triển kỳ vọng có thể là một quá trình tự động, gần như có ý thức phụ, nhưng đó là một hoạt động quan trọng để trau dồi và được cân nhắc về.

Ví dụ, bạn có thẻ đi ăn tối với bạn bè tại một cơ sở chỉ dùng tiền mặt và cần dừng lại ở máy ATM để rút tiền trước khi gấp mặt. Đưa ra quyết định về số tiền bạn sẽ rút, bạn phải phát triển một số kỳ vọng về chi phí của bữa tối. Đây có thể là một kỳ vọng tự động bởi vì bạn ăn cơm trưa tại cơ sở này thường xuyên để bạn biết những gì

chi phí điển hình của một bữa ăn là ở đó, đó sẽ là một ví dụ về kiến thức tiên nghiệm . Một ví dụ khác về kiến thức tiên nghiệm là biết giá một bữa ăn điển hình tại một nhà hàng trong thành phố của bạn hoặc biết giá một bữa ăn tại những nhà hàng đắt nhất trong thành phố của bạn. Sử dụng thông tin đó, có lẽ bạn có thể đặt giới hạn trên và dưới cho bữa ăn sẽ có giá bao nhiêu.

Bạn cũng có thể đã tìm kiếm thông tin bên ngoài để phát triển kỳ vọng của mình, điều này có thể bao gồm việc hỏi bạn bè của bạn, những người sẽ tham gia cùng bạn hoặc những người đã từng ăn ở nhà hàng trước đó và/hoặc Google nhà hàng để tìm thông tin chi phí chung trực tuyến hoặc thực đơn có giá cả . Quá trình tương tự này, trong đó bạn sử dụng bất kỳ thông tin tiên nghiệm nào bạn có và/hoặc các nguồn bên ngoài để xác định những gì bạn mong đợi khi kiểm tra dữ liệu của mình hoặc thực hiện quy trình phân tích, áp dụng cho từng hoạt động cốt lõi của quy trình phân tích dữ liệu.

2.4 Thu thập thông tin

Bước này đòi hỏi phải thu thập thông tin về câu hỏi của bạn hoặc dữ liệu của bạn. Đối với câu hỏi của bạn, bạn thu thập thông tin bằng cách thực hiện tìm kiếm tài liệu hoặc hỏi các chuyên gia để đảm bảo rằng câu hỏi của bạn là một câu hỏi hay. Trong chương tiếp theo, chúng ta sẽ thảo luận về các đặc điểm của một câu hỏi hay. Đối với dữ liệu của bạn, sau khi bạn có một số kỳ vọng về kết quả sẽ như thế nào khi bạn kiểm tra dữ liệu của mình hoặc thực hiện quy trình phân tích, sau đó bạn thực hiện thao tác. Kết quả của hoạt động đó là dữ liệu bạn cần thu thập, sau đó bạn xác định xem dữ liệu bạn đã thu thập có phù hợp với mong đợi của mình hay không. Để mở rộng phép ẩn dụ về nhà hàng, khi bạn đến nhà hàng, nhận xét là thu thập dữ liệu.

2.5 So sánh Kỳ vọng với Dữ liệu

Bây giờ bạn đã có dữ liệu trong tay (việc kiểm tra tại nhà hàng), bước tiếp là so sánh kỳ vọng của bạn với dữ liệu. Có hai kết quả có thể xảy ra: hoặc kỳ vọng của bạn về chi phí khớp với số tiền trên séc hoặc không. Nếu kỳ vọng của bạn và dữ liệu phù hợp, thật tuyệt vời, bạn có thể chuyển sang hoạt động tiếp theo. Mặt khác, nếu kỳ vọng của bạn là chi phí 30 đô la, nhưng séc là 40 đô la, thì kỳ vọng của bạn và dữ liệu không khớp.

Có hai cách giải thích cho sự không phù hợp: thứ nhất, kỳ vọng của bạn đã sai và cần được sửa đổi, hoặc thứ hai, kiểm tra đã sai và có lỗi. Bạn xem lại tấm séc và thấy rằng bạn đã bị tính phí cho hai món tráng miệng thay vì món tráng miệng mà bạn đã có và kết luận rằng có lỗi trong dữ liệu, vì vậy hãy yêu cầu sửa tấm séc đó.

Một chỉ số quan trọng cho thấy việc phân tích dữ liệu của bạn đang diễn ra tốt như thế nào là mức độ dễ dàng hay khó khăn của việc đổi sảnh dữ liệu bạn đã thu thập với kỳ vọng ban đầu của mình. Bạn muốn thiết lập kỳ vọng và dữ liệu của mình sao cho việc kết hợp cả hai trở nên dễ dàng. Trong ví dụ về nhà hàng, kỳ vọng của bạn là 30 đô la và dữ liệu cho biết bữa ăn có giá 40 đô la, vì vậy dễ dàng nhận thấy rằng (a) kỳ vọng của bạn thấp hơn 10 đô la và (b) bữa ăn đắt hơn bạn nghĩ. Khi quay lại nơi này, bạn có thể mang thêm \$10. Nếu kỳ vọng ban đầu của chúng tôi là bữa ăn sẽ có giá từ 0 đô la đến 1.000 đô la, thì đúng là dữ liệu của chúng tôi nằm trong phạm vi đó, nhưng không rõ chúng tôi đã học được thêm bao nhiêu. Ví dụ: bạn có thay đổi hành vi của mình vào lần tới khi quay lại không? Kỳ vọng về một bữa ăn trị giá 30 đô la đôi khi được gọi là một giả thuyết sắc bén bởi vì nó nói lên một điều gì đó rất cụ thể có thể được xác minh bằng dữ liệu.

2.6 Áp dụng Epicycle của Quy trình Phân tích

Trước khi thảo luận về một vài ví dụ, hãy xem lại ba bước để sử dụng cho từng hoạt động phân tích dữ liệu cốt lõi. Những cái này là :

1. Đặt kỳ vọng, 2. Thu

thập thông tin (dữ liệu), so sánh dữ liệu với kỳ vọng của bạn và nếu kỳ vọng không khớp, 3. Sửa đổi kỳ vọng của bạn hoặc sửa dữ liệu để kỳ vọng của bạn và dữ liệu khớp với nhau.

Ví dụ: Tỷ lệ mắc bệnh hen suyễn ở Hoa Kỳ

Hãy áp dụng “vòng tuần hoàn phân tích dữ liệu” cho một ví dụ rất cơ bản. Giả sử câu hỏi ban đầu của bạn là xác định tỷ lệ mắc bệnh hen suyễn ở người lớn, bởi vì công ty của bạn muốn hiểu thị trường có thể lớn như thế nào đối với một loại thuốc điều trị hen suyễn mới. Bạn có một câu hỏi chung đã được sếp xác định, nhưng cần: (1) làm sắc nét câu hỏi, (2) khám phá dữ liệu, (3) xây dựng mô hình thống kê, (4) giải thích kết quả và (5) thông báo kết quả. Chúng ta sẽ áp dụng “vòng tuần hoàn” cho từng hoạt động trong số năm hoạt động cốt lõi này.

Đối với hoạt động đầu tiên, tinh chỉnh câu hỏi, trước tiên bạn sẽ phát triển những mong đợi của mình về câu hỏi, sau đó thu thập thông tin về câu hỏi và xác định xem thông tin bạn thu thập có phù hợp với mong đợi của bạn hay không, và nếu không, bạn sẽ sửa lại câu hỏi. Kỳ vọng của bạn là câu trả lời cho câu hỏi này là không rõ và câu hỏi có thể trả lời được. Tuy nhiên, một tài liệu và tìm kiếm trên internet cho thấy rằng câu hỏi này đã được trả lời (và được trả lời liên tục bởi Trung tâm Kiểm soát Dịch bệnh (CDC)), vì vậy bạn

Chu kỳ phân tích

12

xem xét lại câu hỏi vì bạn có thể chỉ cần truy cập trang web của CDC để lấy dữ liệu về tỷ lệ mắc bệnh hen suyễn gần đây.

Bạn thông báo cho sép của mình và bắt đầu một cuộc trò chuyện tiết lộ rằng bất kỳ loại thuốc mới nào được phát triển sẽ giúp ích cho những người mắc bệnh hen suyễn không được kiểm soát bằng thuốc hiện có, vì vậy bạn xác định một câu hỏi hay hơn, đó là "có bao nhiêu người ở Hoa Kỳ mắc bệnh hen suyễn". Bệnh hen suyễn hiện không được kiểm soát và các yếu tố dự đoán nhân khẩu học của bệnh hen suyễn không được kiểm soát là gì?" Bạn lặp lại quy trình thu thập thông tin để xác định xem câu hỏi của mình có thể trả lời được và có phải là câu hỏi hay không, đồng thời tiếp tục quy trình này cho đến khi bạn hài lòng rằng bạn đã tinh chỉnh câu hỏi của mình để có một câu hỏi hay có thể trả lời bằng dữ liệu có sẵn.

Giả sử rằng bạn đã xác định được một nguồn dữ liệu có thể tải xuống từ một trang web và là một mẫu đại diện cho dân số trưởng thành ở Hoa Kỳ, từ 18 tuổi trở lên.

Hoạt động tiếp theo là phân tích dữ liệu khám phá và bạn bắt đầu với kỳ vọng rằng khi bạn kiểm tra dữ liệu của mình, sẽ có 10.123 hàng (hoặc bản ghi), mỗi hàng đại diện cho một cá nhân ở Hoa Kỳ vì đây là thông tin được cung cấp trong tài liệu hoặc số mã , đi kèm với tập dữ liệu. Sách mã cũng cho bạn biết rằng sẽ có một biến cho biết tuổi của từng cá nhân trong tập dữ liệu.

Tuy nhiên, khi kiểm tra dữ liệu, bạn nhận thấy rằng chỉ có 4.803 hàng, vì vậy hãy quay lại số mã để xác nhận rằng kỳ vọng của bạn về số lượng hàng là chính xác và khi bạn xác nhận rằng kỳ vọng của mình là chính xác, bạn quay lại trang web nơi bạn đã tải xuống các tệp và phát hiện ra rằng có hai tệp chứa dữ liệu bạn cần, với một tệp chứa 4.803 bản ghi và tệp thứ hai chứa 5.320 bản ghi còn lại.

Bạn tải xuống tệp thứ hai và đọc nó vào sổ liệu thống kê của bạn

Chu kỳ phân tích

13

gói phần mềm và nới tệp thứ hai vào tệp đầu tiên.

Bây giờ bạn đã có số hàng chính xác, vì vậy, bạn chuyển sang xác định xem kỳ vọng của bạn về độ tuổi của dân số có khớp với kỳ vọng của bạn hay không, tức là mọi người đều từ 18 tuổi trở lên. Bạn tóm tắt biến tuổi để có thể xem các giá trị tối thiểu và tối đa và thấy rằng tất cả các cá nhân đều từ 18 tuổi trở lên, phù hợp với mong đợi của bạn. Mặc dù có nhiều việc bạn phải làm để kiểm tra và khám phá dữ liệu của mình, nhưng hai nhiệm vụ này là những ví dụ về cách tiếp cận cần thực hiện. Cuối cùng, bạn sẽ sử dụng bộ dữ liệu này để ước tính tỷ lệ mắc bệnh hen suyễn không kiểm soát được ở người trưởng thành ở Hoa Kỳ.

Hoạt động thứ ba là xây dựng một mô hình thống kê, cần thiết để xác định các đặc điểm nhân khẩu học giúp dự đoán chính xác nhất một người nào đó mắc bệnh hen suyễn không kiểm soát được.

Các mô hình thống kê phục vụ để tạo ra một công thức chính xác cho câu hỏi của bạn để bạn có thể thấy chính xác cách bạn muốn sử dụng dữ liệu của mình, cho dù đó là để ước tính một tham số cụ thể hay để đưa ra dự đoán. Các mô hình thống kê cũng cung cấp một khuôn khổ chính thức trong đó bạn có thể thách thức các phát hiện của mình và kiểm tra các giả định của mình.

Bây giờ bạn đã ước tính tỷ lệ mắc bệnh hen suyễn không được kiểm soát ở người trưởng thành Hoa Kỳ và xác định rằng tuổi, giới tính, chủng tộc, chỉ số khói cơ thể, tình trạng hút thuốc và thu nhập là những yếu tố dự đoán tốt nhất về bệnh hen suyễn không được kiểm soát hiện có, bạn chuyển sang hoạt động cốt lõi thứ tư, hoạt động này là diễn giải kết quả. Trên thực tế, diễn giải các kết quả xảy ra cùng với việc xây dựng mô hình cũng như sau khi bạn xây dựng xong mô hình của mình, nhưng về mặt khái niệm, chúng là các hoạt động riêng biệt.

Giả sử bạn đã xây dựng mô hình cuối cùng của mình và vì vậy bạn đang chuyển sang diễn giải những phát hiện của mô hình. Khi bạn kiểm tra mô hình dự đoán cuối cùng của mình, ban đầu, kỳ vọng của bạn được đổi chiều theo độ tuổi, chủng tộc người Mỹ gốc Phi/da đen,

chỉ số khôi cơ thể, tình trạng hút thuốc và thu nhập thấp đều có liên quan tích cực đến bệnh hen suyễn không được kiểm soát.

Tuy nhiên, bạn nhận thấy rằng giới tính nữ có mối quan hệ tỷ lệ nghịch với bệnh hen suyễn không kiểm soát được, khi nghiên cứu và thảo luận của bạn với các chuyên gia chỉ ra rằng ở người lớn, giới tính nữ có mối liên hệ tỷ lệ thuận với bệnh hen suyễn không kiểm soát được. Sự không phù hợp giữa kỳ vọng và kết quả này khiến bạn phải tạm dừng và thực hiện một số khám phá để xác định xem kết quả của bạn có thực sự chính xác hay không và bạn cần điều chỉnh kỳ vọng của mình hoặc liệu có vấn đề gì với kết quả hơn là kỳ vọng của bạn hay không. Sau khi tìm hiểu kỹ, bạn phát hiện ra rằng bạn đã nghĩ rằng biến giới tính được mã hóa 1 cho nữ và 0 cho nam, nhưng thay vào đó, số mã chỉ ra rằng biến giới tính được mã hóa 1 cho nam và 0 cho nữ. Vì vậy, việc giải thích kết quả của bạn là không chính xác, không phải mong đợi của bạn. Nay giờ bạn đã hiểu mã hóa cho biến giới tính là gì, cách giải thích của bạn về kết quả mô hình phù hợp với mong đợi của bạn, vì vậy bạn có thể chuyển sang truyền đạt những phát hiện của mình.

Cuối cùng, bạn truyền đạt những phát hiện của mình, và vâng, bản hùng ca cũng áp dụng cho giao tiếp. Đối với mục đích của ví dụ này, giả sử bạn đã lập một báo cáo không chính thức bao gồm một bản tóm tắt ngắn gọn về những phát hiện của bạn. Kỳ vọng của bạn là báo cáo của bạn sẽ truyền đạt thông tin mà sép của bạn muốn biết. Bạn gấp sép của mình để xem xét các phát hiện và có ấy hỏi hai câu hỏi: (1) dữ liệu trong bộ dữ liệu được thu thập gần đây như thế nào và (2) những thay đổi về mô hình nhân khẩu học dự kiến sẽ xảy ra trong 5-10 năm tới sẽ ảnh hưởng như thế nào sự phổ biến của bệnh hen suyễn không kiểm soát được. Mặc dù có thể đáng thất vọng khi báo cáo của bạn không đáp ứng đầy đủ nhu cầu của sép, nhưng việc nhận phản hồi là một phần quan trọng trong quá trình phân tích dữ liệu và trên thực tế, chúng tôi cho rằng một phân tích dữ liệu tốt cần có thông tin liên lạc, phản hồi và sau đó

Chu kỳ phân tích

15

hành động để đáp ứng với thông tin phản hồi.

Mặc dù bạn biết câu trả lời về số năm dữ liệu được thu thập, nhưng bạn nhận ra rằng mình đã không đưa thông tin này vào báo cáo của mình, vì vậy bạn sửa lại báo cáo để đưa thông tin đó vào. Bạn cũng nhận ra rằng câu hỏi của sép về tác động của việc thay đổi nhân khẩu học đối với tỷ lệ mắc bệnh hen suyễn không kiểm soát được là một câu hỏi hay vì công ty của bạn muốn dự đoán quy mô thị trường trong tương lai, vì vậy giờ đây bạn có một phân tích dữ liệu mới để giải quyết. Bạn cũng nên cảm thấy hài lòng khi phân tích dữ liệu của mình đưa thêm các câu hỏi lên hàng đầu, vì đây là một đặc điểm của phân tích dữ liệu thành công.

Trong các chương tiếp theo, chúng ta sẽ sử dụng rộng rãi khuôn khổ này để thảo luận về cách mỗi hoạt động trong quy trình phân tích dữ liệu cần được lặp đi lặp lại liên tục. Mặc dù ban đầu việc thực hiện ba bước có vẻ tẻ nhạt, nhưng cuối cùng, bạn sẽ hiểu rõ về nó và chu kỳ của quá trình sẽ diễn ra một cách tự nhiên và trong tiềm thức. Thật vậy, chúng tôi cho rằng hầu hết các nhà phân tích dữ liệu giỏi nhất thậm chí không nhận ra rằng họ đang làm điều này!

3. Nêu rõ và hoàn thiện câu hỏi

Thực hiện phân tích dữ liệu đòi hỏi khá nhiều suy nghĩ và chúng tôi tin rằng khi bạn đã hoàn thành tốt việc phân tích dữ liệu, bạn đã dành nhiều thời gian để suy nghĩ hơn là thực hiện. Việc suy nghĩ bắt đầu trước khi bạn nhìn vào tập dữ liệu và bạn nên suy nghĩ cẩn thận về câu hỏi của mình. Điểm này không thể được nhấn mạnh quá mức vì có thể tránh được nhiều cạm bẫy “chết người” của phân tích dữ liệu bằng cách sử dụng năng lượng tinh thần để trả lời đúng câu hỏi của bạn. Trong chương này, chúng ta sẽ thảo luận về các đặc điểm của một câu hỏi hay, các loại câu hỏi có thể được đặt ra và cách áp dụng quy trình lặp đi lặp lại theo chu kỳ để nêu rõ và tinh chỉnh câu hỏi của bạn sao cho khi bắt đầu xem xét dữ liệu, bạn sẽ có một cái nhìn sắc bén. , câu hỏi có thể trả lời được.

3.1 Các loại câu hỏi

Trước khi chúng tôi đi sâu vào việc nêu rõ câu hỏi, sẽ rất hữu ích nếu bạn xem xét các loại câu hỏi khác nhau là gì. Có sáu loại câu hỏi cơ bản và phần lớn các cuộc thảo luận sau đó xuất phát từ một bài báo¹ được xuất bản trong Khoa học bởi Roger và Jeff Leek². Nhiều loại câu hỏi mà bạn đang hỏi có thể là bước cơ bản nhất bạn có thể thực hiện để đảm bảo rằng, cuối cùng, cách giải thích kết quả của bạn là chính xác. Sáu loại câu hỏi là:

¹<http://www.sciencemag.org/content/347/6228/1314.short>

²<http://jteek.com>

1. Mô tả 2.

Khám phá 3. Suy

luận 4. Dự

đoán 5.

Nguyên

nhân 6. Cơ chế

Và loại câu hỏi bạn đang hỏi trực tiếp thông báo cách bạn diễn giải kết quả của mình.

Một câu hỏi mô tả là một câu hỏi tìm cách tóm tắt một đặc điểm của một tập hợp dữ liệu. Các ví dụ bao gồm xác định tỷ lệ nam giới, số khẩu phần trái cây và rau tươi trung bình mỗi ngày hoặc tần suất mắc bệnh do vi-rút trong một tập hợp dữ liệu được thu thập từ một nhóm cá nhân. Bản thân kết quả không thể giải thích được vì kết quả là một thực tế, một thuộc tính của tập hợp dữ liệu mà bạn đang làm việc.

Một câu hỏi khám phá là một trong đó bạn phân tích dữ liệu để xem liệu có các mẫu, xu hướng hoặc mối quan hệ giữa các biến hay không. Các loại phân tích này còn được gọi là phân tích "tạo ra giả thuyết" bởi vì thay vì kiểm tra một giả thuyết như sẽ được thực hiện với một câu hỏi suy luận, nhận quả hoặc cơ học, bạn đang tìm kiếm các mẫu có thể hỗ trợ đề xuất một giả thuyết. Nếu bạn có suy nghĩ chung rằng chế độ ăn uống có liên quan nào đó đến các bệnh do vi-rút, thì bạn có thể khám phá ý tưởng này bằng cách kiểm tra mối quan hệ giữa một loạt các yếu tố chế độ ăn uống và các bệnh do vi-rút. Bạn nhận thấy trong phân tích thăm dò của mình rằng những người ăn chế độ ăn nhiều thực phẩm nhất định ít mắc bệnh do vi-rút hơn những người có chế độ ăn không giàu những thực phẩm này, vì vậy bạn đề xuất giả thuyết rằng ở người trưởng thành, ăn ít nhất 5 phần trái cây tươi mỗi ngày và rau có liên quan đến ít bệnh do virus hơn mỗi năm.

Nêu và tinh chỉnh câu hỏi

18

Một câu hỏi suy luận sẽ là sự trình bày lại giả thuyết được đề xuất này dưới dạng một câu hỏi và sẽ được trả lời bằng cách phân tích một tập hợp dữ liệu khác, trong ví dụ này, là một mẫu đại diện cho người trưởng thành ở Hoa Kỳ. Bằng cách phân tích tập hợp dữ liệu khác nhau này, cả hai bạn đang xác định liệu mối liên hệ mà bạn quan sát được trong phân tích khám phá của mình có đúng với một mẫu khác hay không và liệu mối liên hệ đó có đúng trong một mẫu đại diện cho dân số Hoa Kỳ trưởng thành hay không, điều này cho thấy mối liên hệ đó có thể áp dụng cho tất cả người lớn ở Mỹ. Nói cách khác, bạn sẽ có thể suy ra điều gì là đúng, tính trung bình, đối với dân số trưởng thành ở Hoa Kỳ từ phân tích mà bạn thực hiện trên mẫu đại diện.

Một câu hỏi dự đoán sẽ là câu hỏi mà bạn hỏi những kiểu người nào sẽ ăn chế độ ăn nhiều trái cây và rau quả tươi trong năm tới.

Trong loại câu hỏi này, bạn ít quan tâm đến nguyên nhân khiến ai đó ăn một chế độ ăn kiêng nhất định, mà chỉ quan tâm đến điều gì dự đoán liệu ai đó sẽ ăn chế độ ăn kiêng nhất định đó hay không. Ví dụ: thu nhập cao hơn có thể là một trong những yếu tố dự đoán cuối cùng và bạn có thể không biết (hoặc thậm chí không quan tâm) tại sao những người có thu nhập cao hơn lại có nhiều khả năng ăn chế độ ăn nhiều trái cây và rau quả tươi hơn, nhưng điều quan trọng nhất là rằng thu nhập là một yếu tố dự đoán hành vi này.

Mặc dù một câu hỏi suy luận có thể cho chúng ta biết rằng những người ăn một loại thực phẩm nào đó có xu hướng ít mắc bệnh do virus hơn, nhưng câu trả lời cho câu hỏi này không cho chúng ta biết liệu việc ăn những thực phẩm này có làm giảm số ca mắc bệnh do virus hay không. Trưởng hợp cho một câu hỏi nhân quả. Một câu hỏi nhân quả hỏi về việc liệu việc thay đổi một yếu tố có làm thay đổi một yếu tố khác, tính trung bình, trong tổng thể hay không. Đôi khi, thiết kế cơ bản của việc thu thập dữ liệu, theo mặc định, cho phép câu hỏi mà bạn đặt ra là quan hệ nhân quả. Một ví dụ về điều này là dữ liệu được thu thập trong bối cảnh của một thử nghiệm ngẫu nhiên, trong đó mọi người được chỉ định ngẫu nhiên để ăn chế độ ăn nhiều trái cây và rau quả tươi hoặc chế độ ăn kiêng

Nêu và tinh chỉnh câu hỏi

19

ít trái cây và rau quả tươi. Trong các trường hợp khác, ngay cả khi dữ liệu của bạn không phải từ một thử nghiệm ngẫu nhiên, bạn có thể áp dụng phương pháp phân tích được thiết kế để trả lời câu hỏi nguyên nhân.

Cuối cùng, không có câu hỏi nào được mô tả cho đến nay sẽ dẫn đến câu trả lời cho chúng ta biết, nếu chế độ ăn kiêng thực sự làm giảm số lượng bệnh do vi-rút gây ra, thì chế độ ăn kiêng dẫn đến giảm số lượng bệnh do vi-rút như thế nào . Một câu hỏi hỏi làm thế nào một chế độ ăn nhiều trái cây và rau quả tươi dẫn đến giảm số ca bệnh do virus sẽ là một câu hỏi máy móc .

Có một vài điểm bổ sung về các loại câu hỏi quan trọng. Đầu tiên, do cần thiết, nhiều phân tích dữ liệu trả lời nhiều loại câu hỏi. Ví dụ: nếu phân tích dữ liệu nhằm trả lời một câu hỏi suy luận, thì các câu hỏi mô tả và khám phá cũng phải được trả lời trong quá trình trả lời câu hỏi suy luận. Để tiếp tục ví dụ của chúng ta về chế độ ăn uống và các bệnh do vi-rút, bạn sẽ không chuyển thẳng sang một mô hình thống kê về mối quan hệ giữa chế độ ăn nhiều trái cây và rau quả tươi với số ca bệnh do vi-rút mà không xác định tần suất của loại chế độ ăn này và các bệnh do vi-rút gây ra. và mối quan hệ của chúng với nhau trong mẫu này. Điểm thứ hai là loại câu hỏi bạn hỏi được xác định một phần bởi dữ liệu có sẵn cho bạn (trừ khi bạn định tiến hành nghiên cứu và thu thập dữ liệu cần thiết để thực hiện phân tích). Ví dụ: bạn có thể đặt câu hỏi nhân quả về chế độ ăn uống và các bệnh do vi-rút để biết liệu chế độ ăn nhiều trái cây và rau quả tươi có làm giảm số ca bệnh do vi-rút hay không và loại dữ liệu tốt nhất để trả lời câu hỏi nhân quả này là một trong đó chế độ ăn uống của mọi người thay đổi từ chế độ ăn nhiều trái cây và rau quả tươi sang chế độ ăn ít hoặc ngược lại. Nếu loại tập dữ liệu này không tồn tại, thì điều tốt nhất bạn có thể làm là áp dụng nguyên nhân

phương pháp phân tích dữ liệu quan sát hoặc thay vào đó trả lời một câu hỏi suy luận về chế độ ăn uống và các bệnh do vi-rút.

3.2 Áp dụng vòng tuần hoàn để phát biểu và tinh chỉnh câu hỏi của bạn

Giờ đây, bạn có thể sử dụng thông tin về các loại câu hỏi và đặc điểm của những câu hỏi hay làm hướng dẫn để tinh chỉnh câu hỏi của mình. Để thực hiện điều này, bạn có thể lặp lại qua 3 bước sau:

1. Thiết lập kỳ vọng của bạn về câu hỏi 2. Thu thập thông tin về câu hỏi của bạn 3. Xác định xem kỳ vọng của bạn có khớp với thông tin bạn đã thu thập hay không, sau đó tinh chỉnh câu hỏi (hoặc kỳ vọng) của bạn nếu kỳ vọng của bạn không khớp với thông tin bạn đã thu thập

3.3 Đặc điểm của một câu hỏi hay

Có năm đặc điểm chính của một câu hỏi hay để phân tích dữ liệu, bao gồm từ đặc điểm rất cơ bản là câu hỏi lẽ ra chưa được trả lời cho đến đặc điểm trừu tượng hơn là mỗi câu trả lời có thể có cho câu hỏi nên có một cách diễn giải duy nhất và có ý nghĩa. Chúng tôi sẽ thảo luận làm thế nào để đánh giá điều này một cách chi tiết hơn dưới đây.

Khi bắt đầu, câu hỏi nên được khán giả quan tâm, danh tính của câu hỏi sẽ phụ thuộc vào ngữ cảnh và môi trường mà bạn đang làm việc với dữ liệu. Nếu bạn đang ở trong giới học thuật, khán giả có thể là cộng tác viên của bạn, cộng đồng khoa học, cơ quan quản lý của chính phủ, của bạn.

các nhà tài trợ, và/hoặc công chúng. Nếu bạn đang làm việc cho một công ty mới thành lập, khán giả của bạn là sếp của bạn, ban lãnh đạo công ty và các nhà đầu tư. Ví dụ: việc trả lời câu hỏi liệu ô nhiễm hạt vật chất ngoài trời có liên quan đến các vấn đề phát triển ở trẻ em hay không có thể được những người liên quan đến việc kiểm soát ô nhiễm không khí quan tâm, nhưng có thể không được chuỗi cửa hàng tạp hóa quan tâm. Mặt khác, việc trả lời câu hỏi liệu doanh số bán pepperoni có cao hơn khi nó được trưng bày bên cạnh nước sốt pizza và vỏ bánh pizza hay khi nó được trưng bày cùng với các loại thịt đóng gói khác sẽ được chuỗi cửa hàng tạp hóa quan tâm, nhưng không phải vậy. người trong các ngành khác.

Bạn cũng nên kiểm tra xem câu hỏi chưa được trả lời chưa. Với sự bùng nổ dữ liệu gần đây, số lượng dữ liệu có sẵn công khai ngày càng tăng và tài liệu khoa học và các nguồn tài nguyên khác dường như vô tận, không có gì lạ khi phát hiện ra rằng câu hỏi quan tâm của bạn đã được trả lời. Một số nghiên cứu và thảo luận với các chuyên gia có thể giúp giải quyết vấn đề này và cũng có thể hữu ích vì ngay cả khi câu hỏi cụ thể mà bạn nghĩ đến chưa được trả lời, các câu hỏi liên quan có thể đã được trả lời và câu trả lời cho những câu hỏi liên quan này là thông tin để quyết định xem hoặc cách bạn tiến hành với câu hỏi cụ thể của mình.

Câu hỏi cũng nên xuất phát từ một khuôn khổ hợp lý .

Nói cách khác, câu hỏi ở trên về mối quan hệ giữa doanh số bán pepperoni và vị trí của nó trong cửa hàng là một câu hỏi hợp lý vì những người mua sắm mua nguyên liệu làm bánh pizza có nhiều khả năng quan tâm đến pepperoni hơn những người mua sắm khác và có thể có nhiều khả năng mua nó hơn nếu họ thấy điều đó cùng lúc với việc họ đang chọn các nguyên liệu làm bánh pizza khác. Một câu hỏi ít hợp lý hơn là liệu doanh số bán xúc xích tiêu có tương quan với doanh số bán sữa chua hay không, trừ khi bạn đã có một số kiến thức trước đó gợi ý rằng những điều này nên tương quan với nhau.

Nếu bạn hỏi một câu hỏi mà khuôn khổ của nó không hợp lý, bạn có thể sẽ nhận được một câu trả lời khó diễn giải hoặc khó tin tưởng. Trong câu hỏi về sửa chữa pepperoni, nếu bạn thấy chúng có mối tương quan với nhau, thì nhiều câu hỏi sẽ được đặt ra về tự kết quả: nó có thực sự đúng không?, tại sao những thứ này lại tương quan với nhau- liệu có cách giải thích nào khác không?, và những thứ khác. Bạn có thể đảm bảo rằng câu hỏi của mình dựa trên một khuôn khổ hợp lý bằng cách sử dụng kiến thức của riêng bạn về lĩnh vực chủ đề và thực hiện một nghiên cứu nhỏ, những điều này cùng nhau có thể giúp ích rất nhiều trong việc giúp bạn phân loại xem câu hỏi của bạn có dựa trên một khuôn khổ hợp lý hay không .

Tất nhiên, câu hỏi cũng nên có thể trả lời được. Mặc dù có lẽ điều này không cần nêu rõ, nhưng cần chỉ ra rằng một số câu hỏi hay nhất không thể trả lời được - vì dữ liệu không tồn tại hoặc không có phương tiện thu thập dữ liệu vì thiếu nguồn lực, tính khả thi, hoặc vấn đề đạo đức. Ví dụ, rất có thể là có những khiếm khuyết trong hoạt động của một số tế bào trong não gây ra bệnh tự kỷ, nhưng không thể thực hiện sinh thiết não để thu thập các tế bào sống để nghiên cứu, điều cần thiết để trả lời câu hỏi này.

Tính cụ thể cũng là một đặc điểm quan trọng của một câu hỏi hay. Một ví dụ về câu hỏi chung là: Ăn theo chế độ lành mạnh có tốt hơn cho bạn không? Làm việc theo hướng cụ thể sẽ tinh chỉnh câu hỏi của bạn và thông báo trực tiếp những bước cần thực hiện khi bạn bắt đầu xem dữ liệu. Một câu hỏi cụ thể hơn xuất hiện sau khi tự hỏi bản thân ý nghĩa của chế độ ăn "lành mạnh hơn" và khi bạn nói điều gì đó "tốt hơn cho bạn"? Quá trình tăng độ đặc hiệu sẽ dẫn đến một câu hỏi cuối cùng, tinh tế, chẳng hạn như: "Liệu ăn ít nhất 5 phần trái cây và rau quả tươi mỗi ngày có dẫn đến ít bị nhiễm trùng đường hô hấp trên (cảm lạnh) hơn không?" Với mức độ cụ thể này, kế hoạch tấn công của bạn rõ ràng hơn nhiều và câu trả lời bạn sẽ nhận được khi kết thúc phân tích dữ liệu sẽ là

dễ hiểu hơn vì bạn sẽ khuyên nghị hoặc không khuyên nghị hành động cụ thể là ăn ít nhất 5 phần trái cây và rau quả tươi mỗi ngày như một biện pháp bảo vệ chống nhiễm trùng đường hô hấp trên.

3.4 Chuyển câu hỏi thành vấn đề dữ liệu

Một khía cạnh khác cần xem xét khi bạn phát triển câu hỏi của mình là điều gì sẽ xảy ra khi bạn chuyển nó thành một bài toán dữ liệu. Mỗi câu hỏi phải được vận hành như một phân tích dữ liệu dẫn đến kết quả. Tạm dừng để suy nghĩ xem kết quả phân tích dữ liệu sẽ như thế nào và chúng có thể được diễn giải như thế nào là rất quan trọng vì nó có thể giúp bạn tránh lãng phí nhiều thời gian bắt tay vào phân tích mà kết quả không thể diễn giải được. Mặc dù chúng ta sẽ thảo luận về nhiều ví dụ về các câu hỏi dẫn đến những kết quả có ý nghĩa và có thể hiểu được trong suốt cuốn sách, nhưng có thể dễ dàng nhất để bắt đầu trước tiên bằng cách suy nghĩ về những loại câu hỏi nào không dẫn đến những câu trả lời có thể hiểu được .

Loại câu hỏi điển hình không đáp ứng tiêu chí này là câu hỏi sử dụng dữ liệu không phù hợp. Ví dụ, câu hỏi của bạn có thể là liệu uống thuốc bổ sung vitamin D có giúp giảm đau đầu hay không và bạn dự định trả lời câu hỏi đó bằng cách sử dụng số lần một người dùng thuốc giảm đau làm chỉ số cho số lần đau đầu của họ. Bạn có thể tìm thấy mối liên quan giữa việc bổ sung vitamin D và uống ít thuốc giảm đau hơn, nhưng sẽ không rõ ràng về cách giải thích kết quả này. Trên thực tế, có thể những người bổ sung vitamin D cũng có xu hướng ít dùng các loại thuốc không kê đơn khác chỉ vì chúng “tránh dùng thuốc” chứ không phải vì chúng thực sự

Nêu và tinh chỉnh câu hỏi

24

bớt đau đầu hơn. Cũng có thể là họ đang sử dụng ít thuốc giảm đau hơn vì họ có ít thuốc giảm đau hơn.

đau khớp, hoặc các loại đau khác, nhưng đau đầu không ít hơn.

Tất nhiên, một cách giải thích khác là chúng thực sự ít đau đầu hơn, nhưng vấn đề là bạn không thể xác định xem đây là cách giải thích chính xác hay một của các giải thích khác là chính xác. Về bản chất, vấn đề với câu hỏi này là đối với một câu trả lời duy nhất có thể, có nhiều cách hiểu. Kịch bản này của nhiều diễn giải phát sinh khi ít nhất một trong các biến bạn sử dụng (trong trường hợp này là sử dụng thuốc giảm đau) không phải là biện pháp tốt để khai niệm mà bạn thực sự theo đuổi (trong trường hợp này là đau đầu). ĐẾN giải quyết vấn đề này, bạn sẽ muốn đảm bảo rằng dữ liệu có sẵn để trả lời câu hỏi của bạn cung cấp hợp lý các biện pháp cụ thể của các yếu tố cần thiết để trả lời của bạn câu hỏi.

Một vấn đề liên quan cản trở việc giải thích kết quả là gây nhiễu. Gây nhầm lẫn là một vấn đề tiềm ẩn khi câu hỏi của bạn hỏi về mối quan hệ giữa các yếu tố, chẳng hạn như uống vitamin D và tần suất nhức đầu. Một mô tả ngắn gọn về khái niệm gây nhiễu là nó xuất hiện khi một yếu tố mà bạn không nhất thiết phải xem xét trong câu hỏi của bạn có liên quan đến cả hai mức độ quan tâm của bạn (ví dụ, uống vitamin D bổ sung) và kết quả quan tâm của bạn (giảm đau thuốc cắt cơ). Ví dụ, thu nhập có thể là một kẻ lừa đảo, bởi vì nó có thể liên quan đến cả việc uống vitamin. Bổ sung D và tần suất đau đầu, vì mọi người với thu nhập cao hơn có thể có xu hướng có nhiều khả năng để có một bổ sung và ít có khả năng mắc các vấn đề sức khỏe mãn tính, chẳng hạn như nhức đầu. Nói chung, miễn là bạn có thu nhập dữ liệu có sẵn cho bạn, bạn sẽ có thể điều chỉnh cho kẻ lừa đảo này và giảm số lượng các diễn giải có thể của câu trả lời cho câu hỏi của bạn. Khi bạn tinh chỉnh câu hỏi của mình,

dành thời gian xác định các yếu tố gây nhiều tiêm ẩn và suy nghĩ xem liệu tập dữ liệu của bạn có bao gồm thông tin về những yếu tố gây nhiều tiêm ẩn này hay không.

Một loại vấn đề khác có thể xảy ra khi dữ liệu không phù hợp được sử dụng là kết quả không thể diễn giải được vì cách cơ bản mà dữ liệu được thu thập dẫn đến kết quả sai lệch. Ví dụ: hãy tưởng tượng rằng bạn đang sử dụng tập dữ liệu được tạo từ cuộc khảo sát những phụ nữ đã có con. Cuộc khảo sát bao gồm thông tin về việc liệu con họ có mắc chứng tự kỷ hay không và liệu họ có báo cáo việc ăn sushi khi mang thai hay không và bạn thấy mối liên hệ giữa báo cáo ăn sushi khi mang thai và việc sinh con mắc chứng tự kỷ. Tuy nhiên, vì những người phụ nữ có con bị bệnh nhớ lại các phơi nhiễm, chẳng hạn như cá sống, xảy ra trong thời kỳ mang thai khác với những người có con khỏe mạnh, nên mối liên quan được quan sát thấy giữa phơi nhiễm sushi và bệnh tự kỷ có thể chỉ là biểu hiện của một bệnh tự kỷ. Xu hướng của người mẹ tập trung vào nhiều sự kiện hơn trong thời kỳ mang thai khi cô ấy sinh con với tình trạng sức khỏe. Đây là một ví dụ về sai lệch nhớ lại, nhưng có nhiều loại sai lệch có thể xảy ra.

Một khuynh hướng chính khác cần hiểu và cân nhắc khi tinh chỉnh câu hỏi của bạn là thiên kiến lựa chọn, xảy ra khi dữ liệu bạn đang phân tích được thu thập theo cách làm tăng tỷ lệ những người có cả hai đặc điểm trên những gì tồn tại trong dân số nói chung. Nếu một nghiên cứu quảng cáo rằng đó là nghiên cứu về bệnh tự kỷ và chế độ ăn uống khi mang thai, thì rất có thể những phụ nữ vừa ăn cá sống vừa có con mắc chứng tự kỷ sẽ có nhiều khả năng trả lời khảo sát hơn những người có một trong những điều trên. Điều kiện hoặc không có điều kiện nào trong số này. Kịch bản này sẽ dẫn đến một câu trả lời sai lệch cho câu hỏi của bạn về việc các bà mẹ ăn sushi khi mang thai và nguy cơ mắc chứng tự kỷ ở con cái của họ. Một nguyên tắc nhỏ là nếu

bạn đang kiểm tra mối quan hệ giữa hai yếu tố, sự thiên vị có thể là một vấn đề nếu bạn có nhiều (hoặc ít) khả năng quan sát các cá nhân có cả hai yếu tố do cách chọn dân số hoặc cách một người có thể nhớ lại quá khứ khi trả lời khảo sát. Sẽ có nhiều thảo luận hơn về sự thiên vị trong các chương tiếp theo về ([Suy luận: Cơ sở và diễn giải kết quả của bạn](#)), nhưng thời điểm tốt nhất để xem xét tác động của nó đối với phân tích dữ liệu của bạn là khi bạn đang xác định câu hỏi mà bạn sẽ trả lời và suy nghĩ về tình trạng của bạn. sẽ trả lời câu hỏi với dữ liệu có sẵn cho bạn.

3.5 Nghiên cứu điển hình

Joe làm việc cho một công ty sản xuất nhiều loại thiết bị và ứng dụng theo dõi hoạt động thể chất và tên của công ty là Fit on Fleek. Mục tiêu của Fit on Fleek, giống như nhiều công ty khởi nghiệp công nghệ, là sử dụng dữ liệu họ thu thập được từ người dùng thiết bị của họ để thực hiện tiếp thị có mục tiêu cho các sản phẩm khác nhau. Sản phẩm mà họ muốn tiếp thị là một sản phẩm mới mà họ vừa phát triển và chưa bắt đầu bán, đó là ứng dụng và theo dõi giấc ngủ theo dõi các giai đoạn khác nhau của giấc ngủ, chẳng hạn như giấc ngủ REM, đồng thời đưa ra lời khuyên để cải thiện giấc ngủ.

Trình theo dõi giấc ngủ có tên là Sleep on Fleek.

Sép của Joe yêu cầu anh phân tích dữ liệu mà công ty có về người dùng thiết bị và ứng dụng theo dõi sức khỏe của họ để xác định người dùng cho các quảng cáo Sleep on Fleek được nhắm mục tiêu. Fit on Fleek có các dữ liệu sau từ mỗi khách hàng của họ: thông tin nhân khẩu học cơ bản, số bước đi bộ mỗi ngày, số lần leo cầu thang mỗi ngày, số giờ tinh táo mỗi ngày, số giờ ngủ mỗi ngày (nhưng không có thông tin chi tiết hơn về giấc ngủ mà trình theo dõi giấc ngủ sẽ theo dõi).

Mặc dù Joe có một mục tiêu trong đầu, nhưng lợm lặt được từ môt

thảo luận với sép của mình và anh ấy cũng biết những loại dữ liệu nào có sẵn trong cơ sở dữ liệu Fit on Fleek, anh ấy vẫn chưa có câu hỏi nào. Tình huống này, trong đó Joe được đưa ra một mục tiêu, nhưng không phải là một câu hỏi, là phổ biến, vì vậy nhiệm vụ đầu tiên của Joe là chuyển mục tiêu thành một câu hỏi và điều này sẽ yêu cầu một số giao tiếp qua lại với sép của anh ấy. Cách tiếp cận thông tin liên lạc không chính thức diễn ra trong quá trình dự án phân tích dữ liệu, được trình bày chi tiết trong **chương Thông tin liên lạc**. Sau một vài cuộc thảo luận, Joe giải quyết được câu hỏi sau: "Người dùng Fit on Fleek nào không ngủ đủ giấc?" Anh ấy và ông chủ của mình đồng ý rằng những khách hàng có nhiều khả năng quan tâm đến việc mua thiết bị và ứng dụng Sleep on Fleek nhất là những người thường như có vấn đề với giấc ngủ và vấn đề dễ theo dõi nhất và có lẽ là vấn đề phổ biến nhất là không nhận được ngủ đủ.

Bạn có thể nghĩ rằng vì Joe hiện có một câu hỏi nên anh ấy nên chuyển sang tải dữ liệu xuống và bắt đầu thực hiện các phân tích khám phá, nhưng Joe vẫn còn một số việc phải làm để tinh chỉnh câu hỏi. Hai nhiệm vụ chính mà Joe cần giải quyết là: (1) suy nghĩ về cách câu hỏi của anh ấy đáp ứng hoặc không đáp ứng các đặc điểm của một câu hỏi hay và (2) xác định loại câu hỏi mà anh ấy đang hỏi để anh ấy có hiểu rõ về những loại kết luận nào có thể (và không thể) được rút ra khi anh ta hoàn thành việc phân tích dữ liệu.

Joe xem xét các đặc điểm của một câu hỏi hay và những kỳ vọng của anh ấy là câu hỏi của anh ấy có tất cả các đặc điểm sau: -quan tâm -chưa có câu trả lời -dựa trên khuôn khổ hợp lý -có thể trả lời -cụ thể Câu trả lời mà anh ấy sẽ nhận được khi kết thúc câu hỏi của mình phân tích (khi anh ấy dịch câu hỏi của mình thành một vấn đề dữ liệu) cũng có thể hiểu được.

Sau đó, anh ta suy nghĩ về những gì anh ta biết về câu hỏi và theo đánh giá của anh ta, câu hỏi đó rất đáng quan tâm khi sép của anh ta bày tỏ sự quan tâm.

Anh ta cũng biết rằng câu hỏi không thể đã được trả lời vì sép của anh ta chỉ ra rằng nó không có và việc xem xét các phân tích dữ liệu trước đây của công ty cho thấy không có phân tích nào trước đó được thiết kế để trả lời câu hỏi.

Tiếp theo, anh ta đánh giá xem câu hỏi có được đặt trong một khuôn khổ hợp lý hay không. Câu hỏi, Người dùng Fit on Fleek nào ngủ không đủ giấc?, dường như có cơ sở hợp lý vì có nghĩa là những người ngủ quá ít sẽ quan tâm đến việc cố gắng cải thiện giấc ngủ của họ bằng cách theo dõi nó. Tuy nhiên, Joe tự hỏi liệu thời lượng của giấc ngủ có phải là dấu hiệu tốt nhất cho việc một người có cảm thấy rằng họ ngủ không đủ giấc hay không. Anh ấy biết một số người thường ngủ ít hơn 5 tiếng mỗi đêm và họ có vẻ hài lòng với giấc ngủ của mình. Joe liên hệ với một chuyên gia về thuốc ngủ và biết rằng thước đo tốt hơn để xác định xem ai đó có bị ảnh hưởng bởi tình trạng thiếu ngủ hoặc giấc ngủ kém chất lượng hay không là tình trạng buồn ngủ vào ban ngày. Hóa ra kỳ vọng ban đầu của anh ấy rằng câu hỏi được đặt cơ sở trong một khuôn khổ hợp lý đã không khớp với thông tin anh ấy nhận được khi

anh ấy đã nói chuyện với một chuyên gia nội dung. Vì vậy, anh ấy sửa lại câu hỏi của mình để nó phù hợp với kỳ vọng về tính hợp lý của anh ấy và câu hỏi được sửa đổi là: Người dùng Fit on Fleek nào bị buồn ngủ vào ban ngày?

Joe dừng lại để đảm bảo rằng câu hỏi này thực sự là một câu hỏi có thể trả lời được với dữ liệu mà anh ấy có sẵn và xác nhận rằng nó đúng như vậy. Anh ấy cũng dừng lại để suy nghĩ về tính đặc thù của câu hỏi. Anh ấy tin rằng nó là cụ thể, nhưng thực hiện bài tập thảo luận câu hỏi với đồng nghiệp để thu thập thông tin về tính cụ thể của câu hỏi.

Khi anh ấy nêu ý tưởng trả lời câu hỏi này, đồng nghiệp của anh ấy

giải đáp hỏi anh ấy nhiều câu hỏi về ý nghĩa của các phần khác nhau của câu hỏi: "người dùng nào" có nghĩa là gì? Điều này có nghĩa là: Đặc điểm nhân khẩu học của những người dùng buồn ngủ là gì? Hay cái gì khác? Còn về "buồn ngủ trong ngày" thì sao? Nên cụm từ này có nghĩa là bất kỳ buồn ngủ vào bất kỳ ngày nào? Hoặc buồn ngủ kéo dài ít nhất một khoảng thời gian nhất định trong ít nhất một số ngày nhất định?

Cuộc trò chuyện với các đồng nghiệp rất nhiều thông tin và chỉ ra rằng câu hỏi không cụ thể lắm. Joe sửa lại câu hỏi của mình sao cho câu hỏi trở nên cụ thể: "Đặc điểm nhân khẩu học và sức khỏe nào xác định những người dùng có nhiều khả năng bị buồn ngủ kinh niên nhất, được định nghĩa là ít nhất một đợt buồn ngủ ít nhất hai ngày một lần?"

Bây giờ Joe chuyển sang suy nghĩ về những câu trả lời có thể có cho câu hỏi của anh ấy là gì và liệu chúng có thể hiểu được không. Joe xác định hai kết quả có thể xảy ra trong phân tích của mình: (1) không có đặc điểm nào xác định những người mắc chứng buồn ngủ ban ngày kinh niên hoặc (2) có một hoặc nhiều đặc điểm xác định những người mắc chứng buồn ngủ kinh niên vào ban ngày. Hai khả năng này đều có thể giải thích được và có ý nghĩa. Đầu tiên, Joe sẽ kết luận rằng không thể đưa quảng cáo về thiết bị theo dõi Sleep on Fleek cho những người được dự đoán là mắc chứng buồn ngủ kinh niên vào ban ngày và đối với lần thứ hai, anh ấy kết luận rằng có thể nhầm mục tiêu quảng cáo và anh ấy 'muốn biết (những) đặc điểm nào sẽ sử dụng để chọn người cho các quảng cáo được nhầm mục tiêu.

Bây giờ Joe đã có trong tay một câu hỏi hay, sau khi lặp đi lặp lại 3 bước của chu kỳ ngoại truyện khi cân nhắc liệu câu hỏi của mình có đáp ứng từng đặc điểm của một câu hỏi hay hay không, bước tiếp theo là để tìm ra loại câu hỏi mà anh ấy đặt ra. có. Anh ta trải qua một quá trình suy nghĩ tương tự như quá trình mà anh ta đã sử dụng cho từng đặc điểm ở trên. Anh ấy bắt đầu nghĩ rằng câu hỏi của mình là một câu hỏi thăm dò, nhưng khi anh ấy xem lại phần mô tả và ví dụ về

một câu hỏi khám phá, anh ấy nhận ra rằng mặc dù một số phần phân tích mà anh ấy sẽ thực hiện để trả lời câu hỏi sẽ mang tính thăm dò, nhưng cuối cùng thì câu hỏi của anh ấy còn hơn cả khám phá vì câu trả lời của nó sẽ dự đoán người dùng nào có khả năng mặc chứng buồn ngủ kinh niên vào ban ngày, vì vậy câu hỏi của anh ấy là một câu hỏi dự đoán. Việc xác định loại câu hỏi rất hữu ích bởi vì, cùng với một câu hỏi hay, giờ đây anh ta biết rằng anh ta cần sử dụng phương pháp dự đoán trong các phân tích của mình, đặc biệt là trong giai đoạn xây dựng mô hình (xem chương Mô hình chính thức) .

3.6 Suy nghĩ Kết luận

Bây giờ, bạn nên sẵn sàng áp dụng 3 bước của chu kỳ ngoại truyện để nêu và hoàn thiện một câu hỏi. Nếu bạn là một nhà phân tích dữ liệu thông minh, phần lớn quy trình này có thể diễn ra tự động, do đó bạn có thể không hoàn toàn ý thức được một số phần của quy trình dẫn bạn đến một câu hỏi hay. Cho đến khi bạn đạt đến điểm này, chương này có thể phục vụ như một nguồn tài nguyên hữu ích cho bạn khi bạn đối mặt với nhiệm vụ phát triển một câu hỏi hay. Trong các chương tiếp theo, chúng ta sẽ thảo luận về những việc cần làm với dữ liệu khi bạn đã có trong tay câu hỏi hay.

4. Phân tích dữ liệu khám phá

Phân tích dữ liệu khám phá là quá trình khám phá dữ liệu của bạn và nó thường bao gồm việc kiểm tra cấu trúc và các thành phần của tập dữ liệu, sự phân bố của các biến riêng lẻ và mối quan hệ giữa hai hoặc nhiều biến. Công cụ được dùng nhiều nhất để phân tích dữ liệu khám phá là trực quan hóa dữ liệu bằng cách sử dụng biểu diễn đồ họa của dữ liệu. Trực quan hóa dữ liệu được cho là công cụ quan trọng nhất để phân tích dữ liệu khám phá vì thông tin được truyền tải bằng màn hình đồ họa có thể được tiếp thu rất nhanh và vì nhìn chung dễ dàng nhận ra các mẫu trong màn hình đồ họa.

Có một số mục tiêu của phân tích dữ liệu thăm dò, trong đó là:

1. Để xác định xem có bất kỳ vấn đề nào với tập dữ liệu.
2. Để xác định xem câu hỏi bạn đang hỏi có thể được trả lời bằng dữ liệu mà bạn có hay không.
3. Để phát triển một bản phác thảo câu trả lời cho câu hỏi của bạn.

Ứng dụng phân tích dữ liệu khám phá của bạn sẽ được hướng dẫn bởi câu hỏi của bạn. Câu hỏi ví dụ được sử dụng trong chương này là: "Các quận ở miền đông Hoa Kỳ có nồng độ ôzôn cao hơn các quận ở miền tây Hoa Kỳ không?" Trong trường hợp này, bạn sẽ khám phá dữ liệu để xác định xem có vấn đề gì với tập dữ liệu hay không và để xác định xem bạn có thể trả lời câu hỏi của mình với tập dữ liệu này hay không.

Tất nhiên, để trả lời câu hỏi, bạn cần dữ liệu về ôzôn, hạt và khu vực của Hoa Kỳ. Bước tiếp theo là sử dụng phân tích dữ liệu thăm dò để bắt đầu trả lời câu hỏi của bạn, có thể bao gồm việc hiển thị các ô ôzôn theo vùng của Hoa Kỳ. Khi kết thúc quá trình phân tích dữ liệu khám phá, bạn nên hiểu rõ câu trả lời cho câu hỏi của mình là gì và được trang bị đầy đủ thông tin để chuyển sang các bước phân tích dữ liệu tiếp theo.

Điều quan trọng cần lưu ý là ở đây, một lần nữa, khái niệm về chu kỳ phân tích ngoại luân được áp dụng. Bạn nên kỳ vọng tập dữ liệu của mình sẽ trông như thế nào và liệu câu hỏi của bạn có thể được trả lời bằng dữ liệu bạn có hay không. Nếu nội dung và cấu trúc của tập dữ liệu không phù hợp với kỳ vọng của bạn, thì bạn sẽ cần quay lại và tìm hiểu xem kỳ vọng của mình có đúng không (nhưng có vấn đề với dữ liệu) hoặc kỳ vọng của bạn không chính xác, vì vậy bạn không thể sử dụng tập dữ liệu để trả lời câu hỏi và sẽ cần tìm một tập dữ liệu khác.

Bạn cũng nên có một số kỳ vọng về mức độ ôzôn cũng như liệu ôzôn của một khu vực sẽ cao hơn (hoặc thấp hơn) so với khu vực khác. Khi bạn chuyển sang bước 3 để bắt đầu trả lời câu hỏi của mình, bạn sẽ lại áp dụng chu kỳ phân tích để nếu, ví dụ, mức ôzôn trong bộ dữ liệu thấp hơn mức bạn mong đợi khi xem dữ liệu đã xuất bản trước đó, thì bạn sẽ cần để tạm dừng và tìm hiểu xem có vấn đề với dữ liệu của bạn hay liệu kỳ vọng của bạn có sai không. Kỳ vọng của bạn có thể không chính xác, ví dụ: nếu nguồn thông tin để đặt kỳ vọng của bạn về mức độ ôzôn là dữ liệu được thu thập từ 20 năm trước (khi mức độ có thể cao hơn) hoặc chỉ từ một thành phố duy nhất ở Hoa Kỳ. Chúng tôi sẽ đi vào chi tiết hơn với nghiên cứu điển hình bên dưới, nhưng điều này sẽ cung cấp cho bạn cái nhìn tổng quan về cách tiếp cận và mục tiêu của phân tích dữ liệu khám phá.

4.1 Danh sách kiểm tra phân tích dữ liệu khám phá: Một nghiên cứu điển hình

Trong phần này, chúng ta sẽ xem qua một “danh sách kiểm tra” không chính thức về những việc cần làm khi bắt tay vào phân tích dữ liệu khám phá. Để làm ví dụ đang chạy, tôi sẽ sử dụng bộ dữ liệu về mức ôzôn hàng giờ ở Hoa Kỳ cho năm 2014. Các yếu tố của danh sách kiểm tra là

1. Đặt câu hỏi của bạn 2. Đọc dữ liệu của bạn 3. Kiểm tra bao bì 4. Nhìn vào phần trên cùng và phần dưới cùng của dữ liệu 5. Kiểm tra các chữ “n” của bạn
6. Xác thực bằng ít nhất một nguồn dữ liệu bên ngoài 7. Tạo một cốt truyện 8. Hãy thử giải pháp để dàng đầu tiên 9. Theo dõi

Trong suốt ví dụ này, chúng tôi sẽ mô tả một phân tích đang diễn ra với mã R và dữ liệu thực. Một số ví dụ và đề xuất ở đây sẽ dành riêng cho môi trường phân tích thống kê R, nhưng hầu hết nên áp dụng cho bất kỳ hệ thống phần mềm nào. Thông thạo R là không cần thiết để hiểu ý chính của ví dụ. Vui lòng bỏ qua các phần mã.

4.2 Xây dựng câu hỏi của bạn

Trước đây trong cuốn sách này, chúng ta đã thảo luận về tầm quan trọng của việc đặt câu hỏi một cách đúng đắn. Xây dựng một câu hỏi có thể là một cách hữu ích để hướng dẫn phân tích dữ liệu khám phá

xử lý và để giới hạn số lượng đường dẫn theo cấp số nhân có thể được thực hiện với bất kỳ tập dữ liệu lớn nào. Cụ thể, một câu hỏi hoặc giả thuyết sắc bén có thể đóng vai trò là công cụ giảm kích thước có thể loại bỏ các biến không liên quan ngay đến câu hỏi.

Ví dụ, trong chương này, chúng ta sẽ xem xét bộ dữ liệu về ô nhiễm không khí từ Cơ quan Bảo vệ Môi trường Hoa Kỳ (EPA). Một câu hỏi chung người ta có thể như là

Mức độ ô nhiễm không khí ở bờ biển phía đông cao hơn ở bờ biển phía tây?

Nhưng một câu hỏi cụ thể hơn có thể là

Mức ozone trung bình hàng giờ ở Thành phố New York có cao hơn ở Los Angeles không?

Lưu ý rằng cả hai câu hỏi đều có thể được quan tâm và không đúng hay sai. Nhưng câu hỏi đầu tiên yêu cầu xem xét tất cả các chất gây ô nhiễm trên toàn bộ bờ biển phía đông và phía tây, trong khi câu hỏi thứ hai chỉ yêu cầu xem xét một chất gây ô nhiễm duy nhất ở hai thành phố.

Bạn nên dành vài phút để tìm ra câu hỏi mà bạn thực sự quan tâm và thu hẹp câu hỏi đó sao cho càng cụ thể càng tốt (mà không trở nên nhảm chán).

Đối với chương này, chúng ta sẽ xem xét câu hỏi sau:

Các quận ở miền đông Hoa Kỳ có mức ozone cao hơn các quận ở miền tây Hoa Kỳ không?

Xin lưu ý thêm, một trong những câu hỏi quan trọng nhất mà bạn có thể trả lời bằng phân tích dữ liệu khám phá là “Tôi có dữ liệu phù hợp để trả lời câu hỏi này không?” Thường ban đầu câu hỏi này khó trả lời, nhưng có thể trở nên rõ ràng hơn khi chúng ta sắp xếp và xem xét dữ liệu.

4.3 Đọc dữ liệu của bạn

Nhiệm vụ tiếp theo trong bất kỳ phân tích dữ liệu khám phá nào là đọc một số dữ liệu. Đôi khi dữ liệu sẽ có định dạng rất lộn xộn và bạn sẽ cần thực hiện một số thao tác đơn dẹp. Những lần khác, người khác sẽ đơn dẹp dữ liệu đó cho bạn, vì vậy bạn sẽ không phải bận tâm đến việc đơn dẹp.

Chúng ta sẽ không vất vả đơn dẹp tập dữ liệu ở đây, không phải vì nó không quan trọng, mà vì thường không có nhiều kiến thức tổng quát để thu được từ việc xem qua nó. Mỗi tập dữ liệu đều có những điểm kỳ quặc riêng và vì vậy, có lẽ tốt nhất hiện tại là không nên sa lầy vào các chi tiết.

Ở đây, chúng tôi có một bộ dữ liệu tương đối rõ ràng từ EPA Hoa Kỳ về các phép đo ôzôn hàng giờ trên toàn nước Mỹ trong năm 2014. Dữ liệu có sẵn từ trang web Hệ thống [Chất lượng Không khí](#) của EPA¹. Tôi chỉ cần tải xuống tệp zip từ trang web, giải nén kho lưu trữ và đặt tệp kết quả vào một thư mục có tên là “dữ liệu”. Nếu bạn muốn chạy mã này, bạn sẽ phải sử dụng cùng một cấu trúc thư mục.

Tập dữ liệu là một tệp giá trị (CSV) được phân tách bằng dấu phẩy, trong đó mỗi hàng của tệp chứa một phép đo ozone hàng giờ tại một số địa điểm trong quốc gia.

LƯU Ý: Việc chạy mã bên dưới có thể mất vài phút.
Có 7.147.884 hàng trong tệp CSV. Nếu mất quá nhiều thời gian,

¹http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

bạn có thể đọc trong một tập hợp con bằng cách chỉ định một giá trị cho đối số n_max thành read_csv() lớn hơn 0.

```
> library(readr)
> ozone <- read_csv("data/hourly_44201_2014.csv",
+ col_types = "ccccinncnnccnccnnccnn")
```

Gói trình đọc của Hadley Wickham là một gói tuyệt vời để đọc các tệp phẳng (như tệp CSV) rất nhanh hoặc ít nhất là nhanh hơn nhiều so với các chức năng tích hợp sẵn của R. Nó tạo ra một số sự đánh đổi để đạt được tốc độ đó, vì vậy các chức năng này không phải lúc nào cũng phù hợp, nhưng chúng phục vụ mục đích của chúng tôi ở đây.

Chuỗi ký tự được cung cấp cho đối số col_types chỉ định loại của từng cột trong tập dữ liệu. Mỗi chữ cái đại diện cho lớp của một cột: "c" cho ký tự, "n" cho số và "i" cho số nguyên. Không, tôi không biết một cách kỳ diệu các lớp của từng cột-tôi chỉ xem nhanh tệp để xem các lớp cột là gì. Nếu có quá nhiều cột, bạn không thể chỉ định col_types và read_csv() sẽ cố gắng tìm ra nó cho bạn.

Để thuận tiện cho sau này, chúng ta có thể viết lại tên của các cột để loại bỏ bất kỳ khoảng trắng nào.

```
> tên(ozone) <- make.names( tên(ozone))
```

4.4 Kiểm tra Bao bì

Bạn đã bao giờ nhận được món quà trước khi bạn được phép mở nó chưa? Chắc chắn, tất cả chúng ta đều có. Vấn đề là món quà đã được bọc kín, nhưng bạn lại rất muốn biết bên trong có gì. Một người phải làm gì trong những tình huống đó? Chà, bạn có thể lắc cái hộp một chút, có thể gõ nó bằng đốt ngón tay để xem nó có phát ra âm thanh rõ ràng không, hoặc thậm chí

cân xem nặng bao nhiêu. Đây là cách bạn nên suy nghĩ về tập dữ liệu của mình trước khi bắt đầu phân tích thực tế.

Giả sử bạn không nhận được bất kỳ cảnh báo hoặc lỗi nào khi đọc tập dữ liệu, thì bây giờ bạn sẽ có một đối tượng trong không gian làm việc của mình có tên là ozone. Thường thì chọc vào đồ vật đó một chút trước khi xé giấy gói là một ý kiến hay.

Ví dụ: bạn nên kiểm tra số lượng hàng

```
> nrow(ôzôn)
[1] 7147884
```

và cột.

```
> ncol(ôzôn)
[1] 23
```

Hãy nhớ khi chúng tôi nói có 7.147.884 hàng trong tệp? Làm thế nào mà phù hợp với những gì chúng ta đã đọc trong? Tập dữ liệu này cũng có tương đối ít cột, vì vậy bạn có thể kiểm tra tệp văn bản gốc để xem số cột được in ra (23) ở đây có khớp với số cột bạn thấy trong tệp gốc hay không.

Một điều khác bạn có thể làm trong R là chạy str() trên tập dữ liệu. Đây thường là một thao tác an toàn theo nghĩa là ngay cả với tập dữ liệu rất lớn, việc chạy str() cũng không mất quá nhiều thời gian.

```
> str(ozon)
Các lớp 'tbl_df', 'tbl' và 'data.frame': 7147884 theo dõi. của 23 biến \
les:
$ Bang.Mã          : chr "01" "01" "01" "01" ...
$ County.Code $    : chr "003" "003" "003" "003" ...
$ Site.Num         : chr "0010" "0010" "0010" "0010" ...
$ Thông số. Mã     : chr "44201" "44201" "44201" "44201" ...
$ POC              : int 1 1 1 1 1 1 1 1 1 ...
$ Vĩ độ            : số 30.5 30.5 30.5 30.5 30.5 ...
$ kinh độ          : số -87.9 -87.9 -87.9 -87.9 -87.9 ...
$ dữ liệu          : chr "NAD83" "NAD83" "NAD83" "NAD83" ...
$ Thông số.Tên     : chr "Ôzôn" "Ôzôn" "Ôzôn" "Ôzôn" ...
$ Ngày.Địa phương: chr "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
$ Thời gian.Địa phương: chr "01:00" "02:00" "03:00" "04:00" ...
$ Ngày.GMT          : chr "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
$ Time.GMT $        : chr "07:00" "08:00" "09:00" "10:00" ...
Sample.Measurement : num 0,047 0,047 0,043 0,038 0,035 0,035 0,0 \
34 0,037 0,044 0,046 ...
$ Units.of.Measure : chr "Phần triệu" "Phần triệu" \
"Phần triệu" "Phần triệu" ...
$ MDL               : số 0,005 0,005 0,005 0,005 0,005 0,005 0,0 \
05 0,005 0,005 0,005 ...
$ Uncertainty $     : num NA NA NA NA NA NA NA NA NA ...
$ Qualifier $       : chr "*****" ...
$ Method.Type $     : chr "FEM" "FEM" "FEM" "FEM" ...
$ Method.Name        : chr "INSTRUMENTAL - ULTRA VIOLET" "INSTRUIME \
NTAL - ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL - U \
LTRA TÍM" ...
$ Bang.Tên          : chr "Alabama" "Alabama" "Alabama" "Alabama" \
...
$ Quận.Tên         : chr "Baldwin" "Baldwin" "Baldwin" "Baldwin" \
...
$ Date.of.Last.Change: chr "2014-06-30" "2014-06-30" "2014-06-30" \
"30-06-2014" ...
```

Đầu ra cho str() trùng lặp một số thông tin mà chúng tôi đã có, như số hàng và số cột. Hơn quan trọng là bạn có thể kiểm tra các lớp của từng cột để đảm bảo rằng chúng được chỉ định chính xác (tức là num-

bers là số và chuỗi là ký tự, v.v.). Vì chúng tôi đã chỉ định trước tất cả các lớp cột trong `read_csv()` nên tất cả chúng phải khớp với những gì chúng tôi đã chỉ định.

Thông thường, chỉ với những thao tác đơn giản này, bạn có thể xác định các vấn đề tiềm ẩn với dữ liệu trước khi lao đầu vào phân tích dữ liệu phức tạp.

4.5 Nhìn vào Trên cùng và Dưới cùng của bạn Dữ liệu

Việc xem xét phần “bắt đầu” và “kết thúc” của tập dữ liệu ngay sau khi bạn kiểm tra bao bì thường rất hữu ích. Điều này cho bạn biết liệu dữ liệu đã được đọc đúng cách chưa, mọi thứ đã được định dạng đúng chưa và mọi thứ đều ở đó. Nếu dữ liệu của bạn là dữ liệu chuỗi thời gian, thì hãy đảm bảo ngày ở đầu và cuối của tập dữ liệu khớp với khoảng thời gian bắt đầu và kết thúc mà bạn mong đợi.

Trong R, bạn có thể xem qua phần trên cùng và dưới cùng của dữ liệu bằng các hàm `head()` và `tail()`.

Đây là hàng đầu.

```
> đầu(ozon[, c(6:7, 10)])
Vĩ độ Kinh độ Ngày.Địa phương 1
30.498 -87.88141 2014-03-01
2 30.498 -87.88141 2014-03-01
3 30.498 -87.88141 2014-03-01
4 30.498 -87.88141 2014-03-01
5 30.498 -87.88141 2014-03-01
6 30.498 -87.88141 2014-03-01
```

Để cho ngắn gọn, tôi chỉ lấy một vài cột. Và đây là đây.

```
> đuôi(ozon[, c(6:7, 10)])
Vĩ độ Kinh độ Ngày.Địa phương
7147879 18.17794 -65.91548 2014-09-30
7147880 18.17794 -65.91548 2014-09-30
7147881 18.17794 -65.91548 2014-09-30
7147882 18.17794 -65.91548 2014-09-30
7147883 18.17794 -65.91548 2014-09-30
7147884 18.17794 -65.91548 2014-09-30
```

Hàm tail() có thể đặc biệt hữu ích vì thường sẽ có một số vấn đề khi đọc phần cuối của tập dữ liệu và nếu bạn không kiểm tra cụ thể thì bạn sẽ không bao giờ biết được.

Đôi khi có định dạng kỳ lạ ở cuối hoặc một số dòng chú thích cũ mà ai đó đã quyết định dán ở cuối.

Điều này đặc biệt phổ biến với dữ liệu được xuất từ bảng tính Microsoft Excel.

Đảm bảo kiểm tra tất cả các cột và xác minh rằng tất cả dữ liệu trong mỗi cột trông đúng như mong muốn.

Đây không phải là một cách tiếp cận hoàn hảo, bởi vì chúng tôi chỉ xem xét một vài hàng, nhưng đó là một khởi đầu tốt.

4.6 ABC: Luôn kiểm tra các chữ "n" của bạn

Nói chung, đêm mọi thứ thường là một cách hay để tìm xem có gì sai hay không. Trong trường hợp đơn giản nhất, nếu bạn đang mong đợi có 1.000 quan sát và hóa ra chỉ có 20 quan sát, thì bạn biết chắc chắn đã xảy ra sự cố ở đâu đó. Nhưng có những lĩnh vực khác mà bạn có thể kiểm tra tùy thuộc vào ứng dụng của bạn. Để làm điều này đúng cách, bạn cần xác định một số điểm mốc có thể được sử dụng để kiểm tra dữ liệu của bạn. Ví dụ: nếu bạn đang thu thập dữ liệu về mọi người, chẳng hạn như trong một cuộc khảo sát hoặc thử nghiệm lâm sàng, thì bạn nên biết có bao nhiêu người trong nghiên cứu của mình.

Đó là điều bạn nên kiểm tra trong tập dữ liệu của mình, để thực hiện

chắc chắn rằng bạn có dữ liệu về tất cả những người mà bạn nghĩ rằng bạn sẽ có dữ liệu trên.

Trong ví dụ này, chúng tôi sẽ sử dụng thực tế là tập dữ liệu có ý chứa dữ liệu hàng giờ cho toàn bộ quốc gia. Đây sẽ là hai mốc của chúng tôi để so sánh.

Ở đây, chúng tôi có dữ liệu ozone hàng giờ đến từ màn hình trên toàn quốc. Các màn hình nên được theo dõi liên tục trong ngày, vì vậy tất cả các giờ nên được gửi đi. Chúng ta có thể xem biến Time.Local để xem

các phép đo thời gian được ghi lại là được thực hiện.

> đầu(bảng(ozone\$Time.Local))

	00:00	00:01	01:00	01:02	02:00	02:02
288698	2	290871	2	283709	2	

Một điều chúng tôi nhận thấy ở đây là trong khi hầu hết tất cả các phép đo trong tập dữ liệu được ghi lại là được thực hiện trên giờ, một số được thực hiện vào những thời điểm hơi khác nhau. nhỏ như vậy số lượng bài đọc được thực hiện vào những thời điểm tắt mà chúng tôi có thể không muốn quan tâm. Nhưng nó có vẻ hơi kỳ lạ, vì vậy nó có thể có giá trị kiểm tra nhanh chóng.

Chúng ta có thể xem những quan sát nào được đo tại thời gian "00:01".

```
> thư viện (dplyr)
> bộ lọc (ôzôn, Time.Local == "13:14") %>%
+     select(Bang.Name, County.Name, Date.Local,
+            Time.Local, Sample.Measurement)
Nguồn: khung dữ liệu cục bộ [2 x 5]
```

```
Tiêu bang.Tên Quận.Tên Ngày.Giờ địa phương.Địa phương
      (chr)          (chr) (chr) (chr)
1 NewYork        Franklin 2014-09-30 13:14
2 Newyork        Franklin 2014-09-30           13:14
Các biến không được hiển thị: Sample.Measurement (dbl)
```

Chúng ta có thể thấy rằng đó là một màn hình ở Quận Franklin, New York và rằng các phép đo đã được thực hiện vào ngày 30 tháng 9, 2014. Điều gì sẽ xảy ra nếu chúng ta chỉ kéo tất cả các phép đo đã thực hiện tại màn hình này vào ngày này?

```
> bộ lọc (ôzôn, State.Code == "36"
+     & County.Code == "033"
+     & Date.Local == "2014-09-30") %>%
+     select(Date.Local, Time.Local,
+            Sample.Measurement) %>%
+     as.data.frame

Date.Local Time.Local Sample.Measurement
1 2014-09-30      00:01       0,011
2 2014-09-30      01:02       0,012
3 2014-09-30      02:03       0,012
4 2014-09-30      03:04       0,011
5 2014-09-30      04:05       0,011
6 2014-09-30      05:06       0,011
7 2014-09-30      06:07       0,010
8 2014-09-30      07:08       0,010
9 2014-09-30      08:09       0,010
10 2014-09-30     09:10       0,010
11 2014-09-30     10:11       0,010
12 2014-09-30     11:12       0,012
13 2014-09-30     12:13       0,011
14 2014-09-30     13:14       0,013
15 2014-09-30     14:15       0,016
16 2014-09-30     15:16       0,017
17 2014-09-30     16:17       0,017
```

Phân tích dữ liệu thăm dò

43

18	2014-09-30	17:18	0,015
19	2014-09-30	18:19	0,017
20	2014-09-30	19:20	0,014
21	2014-09-30	20:21	0,014
22	2014-09-30	21:22	0,011
23	2014-09-30	22:23	0,010
24	2014-09-30	23:24	0,010
25	2014-09-30	00:01	0,010
26	2014-09-30	01:02	0,011
27	2014-09-30	02:03	0,011
28	2014-09-30	03:04	0,010
29	2014-09-30	04:05	0,010
30	2014-09-30	05:06	0,010
31	2014-09-30	06:07	0,009
32	2014-09-30	07:08	0,008
33	2014-09-30	08:09	0,009
34	2014-09-30	09:10	0,009
35	2014-09-30	10:11	0,009
36	2014-09-30	11:12	0,011
37	2014-09-30	12:13	0,010
38	2014-09-30	13:14	0,012
39	2014-09-30	14:15	0,015
40	2014-09-30	15:16	0,016
41	2014-09-30	16:17	0,016
42	2014-09-30	17:18	0,014
43	2014-09-30	18:19	0,016
44	2014-09-30	19:20	0,013
45	2014-09-30	20:21	0,013
46	2014-09-30	21:22	0,010
47	2014-09-30	22:23	0,009
48	2014-09-30	23:24	0,009

Bây giờ chúng ta có thể thấy rằng màn hình này chỉ ghi lại các giá trị của nó tại thời gian lẻ, thay vì vào giờ. Có vẻ như, từ việc tìm kiếm ở đầu ra trước đó, rằng đây là màn hình duy nhất trong quốc gia làm điều này, vì vậy nó có lẽ không phải là điều chúng tôi nên lo lắng về.

Bởi vì EPA giám sát ô nhiễm trên toàn quốc, nên có một đại diện tốt của các quốc gia. Có lẽ chúng ta sẽ thấy chính xác có bao nhiêu trạng thái được thể hiện trong này

tập dữ liệu.

```
> select(ozone, State.Name) %>% duy nhất %>% nrow
[1] 52
```

Vì vậy, có vẻ như đại diện là một chút quá tốt-có
52 tiểu bang trong tập dữ liệu, nhưng chỉ có 50 tiểu bang ở Hoa Kỳ!

Chúng ta có thể xem xét các yếu tố đặc đáo của State.Name
biến để xem những gì đang xảy ra.

```
> duy nhất(ozone$State.Name)
[1] "Alabama"           "Alasca"
[3] "Arizona"            "Arkansas"
[5] "California"         "Colorado"
[7] "Connecticut"         "Delaware"
[9] "Đặc khu Columbia"   "Florida"
[11] "Gruzia"              "Hawaii"
[13] "Idaho"                "Illinois"
[15] "Ấn Độ"                  "Iowa"
[17] "Kansas"                "Kentucky"
[19] "Louisiana"             "Maine"
[21] "Maryland"               "Massachusetts"
[23] "Michigan"                "Minnesota"
[25] "Mississippi"             "Missouri"
[27] "Montana"                 "Nebraska"
[29] "Nevada"                  "Mới Hampshire"
[31] "Áo mới"                  "Mexico mới"
[33] "New York"                 "Bắc Carolina"
[35] "Bắc Dakota"                "Ôi"
[37] "Oklahoma"                  "Oregon"
[39] "Pennsylvania"              "Đảo Rhode"
[41] "Nam Carolina"                "Nam Dakota"
[43] "Tennessee"                  "Texas"
[45] "Utah"                      "Vermont"
[47] "Virginia"                  "Washington"
[49] "Tây Virginia"                "Wisconsin"
[51] "Wyoming"                    "Puerto Rico"
```

Bây giờ chúng ta có thể thấy rằng Washington, DC (Quận Columbia)
và Puerto Rico là các bang "thêm" có trong bộ dữ liệu.

Vì họ rõ ràng là một phần của Hoa Kỳ (nhưng không phải là các quốc gia chính thức của liên minh) nên tất cả đều ổn.

Phản phân tích cuối cùng này đã sử dụng một thứ mà chúng ta sẽ thảo luận trong phần tiếp theo: dữ liệu bên ngoài. Chúng tôi biết rằng chỉ có 50 tiểu bang ở Hoa Kỳ, vì vậy việc nhìn thấy tên của 52 tiểu bang là một dấu hiệu ngay lập tức cho thấy có thể có điều gì đó không ổn. Trong trường hợp này, tất cả đều ổn nhưng việc xác thực dữ liệu của bạn bằng nguồn dữ liệu bên ngoài có thể rất hữu ích. Điều gì mang lại cho chúng ta ĐÉN..

4.7 Xác thực với ít nhất một nguồn dữ liệu ngoài

Đảm bảo dữ liệu của bạn khớp với thứ gì đó bên ngoài tập dữ liệu là rất quan trọng. Nó cho phép bạn đảm bảo rằng các phép đo gần như phù hợp với những gì chúng nên có và nó đóng vai trò kiểm tra xem những thứ khác có thể sai trong tập dữ liệu của bạn. Xác thực bên ngoài thường có thể đơn giản như kiểm tra dữ liệu của bạn dựa trên một số duy nhất, như chúng tôi sẽ thực hiện ở đây.

Ở Hoa Kỳ, chúng tôi có các tiêu chuẩn quốc gia về chất lượng không khí xung quanh và đối với ôzôn, [tiêu chuẩn hiện hành²](#) được đặt ra vào năm 2008 là “nồng độ tối đa hàng ngày trong 8 giờ cao thứ tư hàng năm, trung bình trong 3 năm” không được vượt quá 0,075 phần triệu (ppm). Các chi tiết chính xác về cách tính toán giá trị này không quan trọng đối với phân tích này, nhưng nói một cách đại khái, nồng độ trung bình trong 8 giờ không được cao hơn 0,075 ppm quá nhiều (có thể cao hơn do cách diễn đạt tiêu chuẩn).

Chúng ta hãy nhìn vào các phép đo hàng giờ của ozone.

² http://www.epa.gov/ttn-naaqs/standards/ozone/s_o3_history.html

```
> tóm tắt(ozone$Sample.Measurement)
tối thiểu 1 Qu. Trung bình Nghĩa thứ 3 Qu. tối đa.
0.00000 0.02000 0.03200 0.03123 0.04200 0.34900
```

Từ bản tóm tắt, chúng ta có thể thấy rằng nồng độ tối đa theo giờ là khá cao (0,349 ppm) nhưng nhìn chung, phần lớn phân phối thấp hơn nhiều so với 0,075.

Chúng ta có thể biết thêm một chút chi tiết về phân phối bằng cách xem xét các bộ dữ liệu.

```
> lượng tử(ozone$Sample.Measurement, seq(0, 1, 0,1))
0% 10% 20% 30% 40% 50% 60% 70%
0.000 0.010 0.018 0.023 0.028 0.032 0.036 0.040
80% 90% 100%
0,044 0,051 0,349
```

Biết rằng tiêu chuẩn quốc gia về ozone là khoảng 0,075, chúng ta có thể thấy từ dữ liệu rằng

- Dữ liệu ít nhất có thứ tự độ lớn phù hợp (tức là đơn vị chính xác)
- Phạm vi phân phối đại khái là những gì chúng tôi mong đợi, dựa trên quy định về mức độ ô nhiễm xung quanh
- Một số mức hàng giờ (dưới 10%) cao hơn 0,075 nhưng điều này có thể hợp lý dựa trên cách diễn đạt của tiêu chuẩn và mức trung bình liên quan.

4.8 Tạo một cốt truyện

Tạo một biểu đồ để trực quan hóa dữ liệu của bạn là một cách tốt để hiểu thêm về câu hỏi và dữ liệu của bạn.

Vẽ sơ đồ có thể xảy ra ở các giai đoạn khác nhau của phân tích dữ liệu. Vì

ví dụ, biểu đồ có thể xảy ra ở giai đoạn khám phá hoặc sau đó trong giai đoạn trình bày/giao tiếp.

Có hai lý do chính để tạo biểu đồ dữ liệu của bạn.

Họ đang tạo ra những kỳ vọng và kiểm tra những sai lệch so với kỳ vọng.

Ở giai đoạn đầu của quá trình phân tích, bạn có thể được trang bị một câu hỏi/giả thuyết, nhưng bạn có thể hiểu rất ít về những gì đang diễn ra trong dữ liệu. Bạn có thể đã xem qua một số trong số đó để thực hiện một số kiểm tra về độ chính xác, nhưng nếu tập dữ liệu của bạn đủ lớn, sẽ rất khó để xem tất cả dữ liệu.

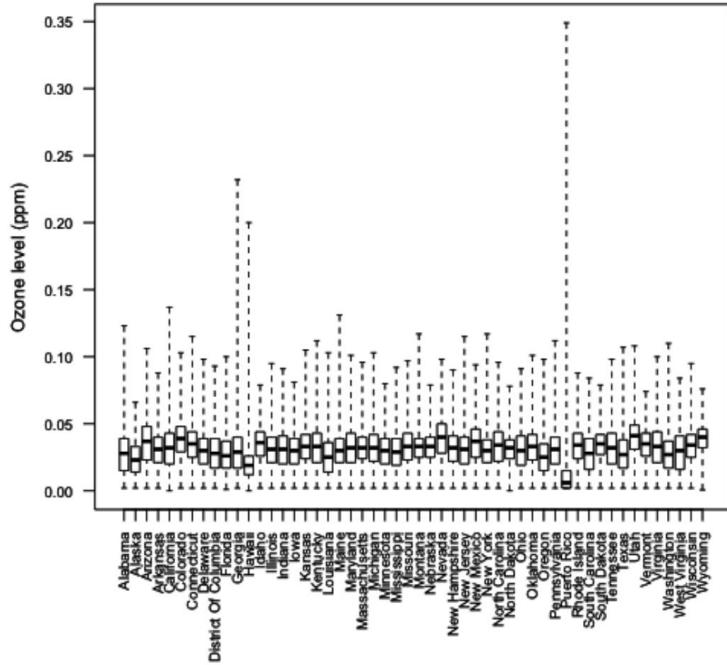
Vì vậy, việc tạo ra một số loại biểu đồ, phục vụ như một bản tóm tắt, sẽ là một công cụ hữu ích để đặt kỳ vọng về dữ liệu sẽ trông như thế nào.

Khi bạn đã hiểu rõ về dữ liệu, một câu hỏi/giả thuyết hay và một loạt các kỳ vọng về những gì dữ liệu nên nói trước câu hỏi của bạn, thì việc tạo biểu đồ có thể là một công cụ hữu ích để xem dữ liệu phù hợp với kỳ vọng của bạn đến mức nào. Cốt truyện đặc biệt tốt trong việc cho bạn thấy những sai lệch so với những gì bạn có thể mong đợi. Các bảng thường rất tốt trong việc tóm tắt dữ liệu bằng cách trình bày những thứ như phương tiện, trung vị hoặc các số liệu thống kê khác. Tuy nhiên, đồ thị có thể cho bạn thấy những thứ đó, cũng như cho bạn thấy những thứ khác xa với giá trị trung bình hoặc trung bình, vì vậy bạn có thể kiểm tra xem liệu có thứ gì đó được cho là ở xa như vậy hay không. Thông thường, những gì rõ ràng trong cốt truyện có thể được ẩn đi trong một bảng.

Đây là một **boxplot3** đơn giản của dữ liệu ozone, với một boxplot cho mỗi trạng thái.

³https://en.wikipedia.org/wiki/Box_plot

```
> par(las = 2, mar = c(10, 4, 2, 2), cex.axis = 0.8) >
boxplot(Sample.Measurement ~ State.Name, ozone, range = 0, ylab = \ "Mức
ozone ( trang/phút)")
```



Boxplot giá trị ôzôn theo trạng thái

Từ đồ thị, chúng ta có thể thấy rằng đối với hầu hết các trạng thái, dữ liệu nằm trong phạm vi khá hẹp dưới 0,05 ppm. Tuy nhiên, đối với Puerto Rico, chúng tôi thấy rằng các giá trị điển hình rất thấp, ngoại trừ một số giá trị cực kỳ cao. Tương tự như vậy, Georgia và Hawaii đôi khi có giá trị rất cao. Đây có thể là giá trị khám phá thêm, tùy thuộc vào câu hỏi của bạn.

4.9 Hãy thử giải pháp dễ dàng trước

Nhớ lại rằng câu hỏi ban đầu của chúng tôi là

Các quận ở miền đông Hoa Kỳ có mức ozone cao hơn các quận ở miền tây Hoa Kỳ không?

Câu trả lời đơn giản nhất mà chúng ta có thể cung cấp cho câu hỏi này là gì? Hiện tại, đừng lo lắng về việc liệu người trả lời có đúng hay không, nhưng vấn đề là làm thế nào bạn có thể cung cấp bằng chứng sơ bộ cho giả thuyết hoặc câu hỏi của mình. Bạn có thể bác bỏ bằng chứng đó sau khi phân tích sâu hơn, nhưng đây là bước đầu tiên. Điều quan trọng là, nếu bạn không tìm thấy bằng chứng về tín hiệu trong dữ liệu chỉ bằng cách sử dụng một đồ thị hoặc phân tích đơn giản, thì thường thì bạn sẽ khó tìm thấy điều gì đó bằng cách sử dụng phân tích phức tạp hơn.

Đầu tiên, chúng ta cần xác định ý nghĩa của "phương đông" và "phương tây". Điều đơn giản nhất cần làm ở đây là chỉ cần chia đất nước thành đông và tây bằng cách sử dụng một giá trị kinh độ cụ thể. Hiện tại, chúng tôi sẽ sử dụng -100 làm điểm cắt. Bất kỳ màn hình nào có kinh độ nhỏ hơn -100 sẽ là "tây" và bất kỳ màn hình nào có kinh độ lớn hơn hoặc bằng -100 sẽ là "đông".

```
> library(maps)
> map("state")
> abline(v = -100, lwd = 3)
> text(-120, 30, "West")
> text(-75, 30, "East")
```



Bản đồ khu vực Đông và Tây

Ở đây, chúng tôi tạo một biến mới có tên là vùng mà chúng tôi sử dụng để cho biết liệu một phép đo nhất định trong tập dữ liệu được ghi ở “phía đông” hay “phía tây”.

```
> ozone$region <- factor(ifelse(ozone$Longitude < -100, "tây", "east"))
```

Bây giờ, chúng ta có thể lập một bản tóm tắt đơn giản về nồng độ ôzôn ở phía đông và phía tây của Hoa Kỳ để xem mức độ có xu hướng ở đâu cao hơn.

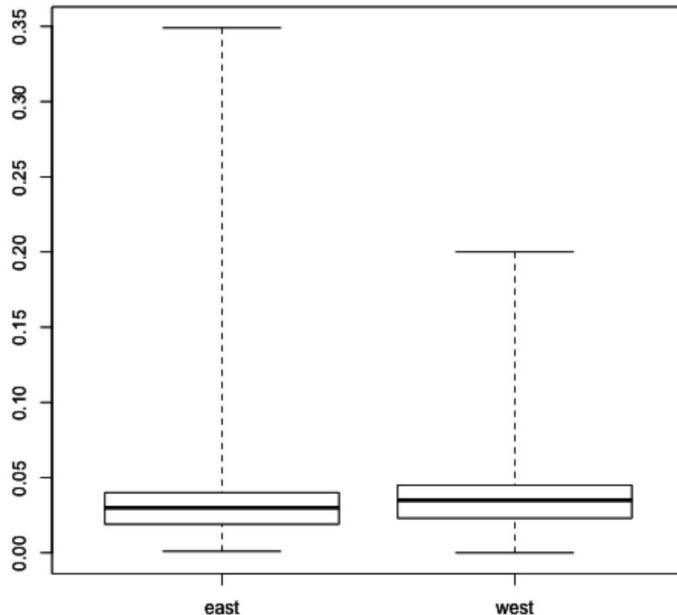
```
> group_by(ozone, khu vực) %>%
+     tóm tắt (mean = mean(Sample.Measurement, na.rm = TRUE),
+               median = median(Sample.Measurement, na.rm = TRUE\)
+ )
Nguồn: khung dữ liệu cục bộ [2 x 3]
```

```
vùng          có nghĩa là trung bình
(fctr) (dbl) (dbl)
1 đông 0,02995250 0,030
2 tây 0,03400735 0,035
```

Cả mức ôzôn trung bình và trung bình đều cao hơn ở tây Hoa Kỳ hơn ở miền đông Hoa Kỳ, khoảng 0,004 ppm.

Chúng ta cũng có thể tạo một biểu đồ ôzôn ở hai khu vực để xem họ so sánh như thế nào.

```
> boxplot(Sample.Measurement ~ vùng, ozon, phạm vi = 0)
```



Boxplot của Ozone cho khu vực Đông và Tây

Chúng ta có thể thấy từ các ô vuông rằng sự biến đổi của ôzôn ở phía đông có xu hướng cao hơn nhiều so với sự biến đổi ở phía tây.

Thách thức giải pháp của bạn

Giải pháp dễ dàng là tốt bởi vì nó dễ dàng, nhưng bạn không bao giờ nên để những kết quả đó tồn tại trong ngày. Bạn phải luôn nghĩ cách thách thức các kết quả, đặc biệt nếu những kết quả đó phù hợp với kỳ vọng trước đó của bạn.

Nhớ lại rằng trước đây chúng tôi nhận thấy rằng ba tiểu bang có một số

giá trị cao bất thường của ozone. Chúng tôi không biết nếu những các giá trị có thực hay không (hiện tại, giả sử chúng là có thật), nhưng có thể thú vị để xem liệu mô hình đông/tây có giống nhau không giữ vững nếu chúng tôi loại bỏ các trạng thái có hoạt động bất thường này.

```
> lọc(ozone, State.Name != "Puerto Rico"
+      & State.Name != "Georgia"
+      & State.Name != "Hawaii") %>%
+      group_by(khu vực) %>%
+      tóm tắt (mean = mean(Sample.Measurement, na.rm = TRUE),
+              median = median(Sample.Measurement, na.rm = TRUE))
))

Nguồn: khung dữ liệu cục bộ [2 x 3]
```

khu vực	có nghĩa là trung bình
(fctr)	(dbl) (dbl)
1 đông	0,03003692 0,030
2 tây	0,03406880 0,035

Thật vậy, có vẻ như mô hình giống nhau ngay cả với 3 các tiêu bang bị loại bỏ.

4.10 Câu hỏi tiếp theo

Trong chương này, chúng tôi đã trình bày một số bước đơn giản để thực hiện khi bắt đầu phân tích thăm dò. ví dụ phân tích được thực hiện trong chương này còn lâu mới hoàn hảo, nhưng nó khiến chúng tôi suy nghĩ về dữ liệu và câu hỏi quan tâm. Nó cũng cung cấp cho chúng tôi một số điều cần theo dõi trong trường hợp chúng tôi tiếp tục quan tâm đến câu hỏi này.

Tại thời điểm này, thật hữu ích khi xem xét một vài câu hỏi tiếp theo.

1. Bạn có dữ liệu phù hợp không? Đôi khi ở phần kết luận của một phân tích dữ liệu khám phá, kết luận là tập dữ liệu không thực sự phù hợp cho việc này

câu hỏi. Trong trường hợp này, bộ dữ liệu dường như hoàn toàn phù hợp để trả lời câu hỏi liệu các quận ở miền đông Hoa Kỳ có cấp độ cao hơn ở miền tây Hoa Kỳ hay không.

2. Bạn có cần dữ liệu khác không? Mặc dù dữ liệu có vẻ phù hợp để trả lời câu hỏi được đặt ra, nhưng cần lưu ý rằng bộ dữ liệu chỉ bao gồm một năm (2014).

Có thể đáng để kiểm tra xem liệu mô hình đông/tây có tồn tại trong những năm khác hay không, trong trường hợp đó, chúng tôi phải ra ngoài và thu thập dữ liệu khác.

3. Bạn có câu hỏi đúng không? Trong trường hợp này, không rõ ràng rằng câu hỏi mà chúng tôi đã có gắng trả lời có liên quan ngay lập tức và dữ liệu không thực sự chỉ ra bất cứ điều gì để tăng mức độ liên quan của câu hỏi. Ví dụ, có thể sẽ thú vị hơn khi đánh giá quận nào vi phạm tiêu chuẩn chất lượng không khí xung quanh quốc gia, bởi vì việc xác định điều này có thể có ý nghĩa pháp lý. Tuy nhiên, đây là một phép tính phức tạp hơn nhiều, yêu cầu dữ liệu từ ít nhất 3 năm trước.

Mục tiêu của phân tích dữ liệu khám phá là giúp bạn suy nghĩ về dữ liệu và suy luận về câu hỏi của mình. Tại thời điểm này, chúng tôi có thể tinh chỉnh câu hỏi của mình hoặc thu thập dữ liệu mới, tất cả trong một quy trình lặp đi lặp lại để tìm ra sự thật.

5. Sử dụng các mô hình để khám phá dữ liệu của bạn

Mục tiêu của chương này là mô tả khái niệm mô hình nói chung hơn là gì, giải thích mục đích của mô hình đối với một tập hợp dữ liệu là gì và cuối cùng là mô tả quy trình mà nhà phân tích dữ liệu tạo ra, đánh giá , và tinh chỉnh một mô hình. Theo nghĩa rất chung, một mô hình là thứ chúng ta xây dựng để giúp chúng ta hiểu thế giới thực. Một ví dụ phổ biến là việc sử dụng một con vật bắt chước bệnh ở người để giúp chúng ta hiểu và hy vọng đầy đủ về việc ngăn ngừa và/hoặc điều trị bệnh. Khái niệm tương tự cũng áp dụng cho một tập hợp dữ liệu - có lẽ bạn đang sử dụng dữ liệu để hiểu thế giới thực.

Trong thế giới chính trị, người thăm dò ý kiến có một bộ dữ liệu về một mẫu cử tri có khả năng và công việc của người thăm dò ý kiến là sử dụng mẫu này để dự đoán kết quả bầu cử. Nhà phân tích dữ liệu sử dụng dữ liệu bỏ phiếu để xây dựng một mô hình nhằm dự đoán điều gì sẽ xảy ra vào ngày bầu cử. Quá trình xây dựng một mô hình liên quan đến việc áp đặt một cấu trúc cụ thể lên dữ liệu và tạo ra một bản tóm tắt dữ liệu. Trong ví dụ về dữ liệu bỏ phiếu, bạn có thể có hàng nghìn quan sát, do đó, mô hình là một phương trình toán học phản ánh hình dạng hoặc mẫu của dữ liệu và phương trình này cho phép bạn tóm tắt hàng nghìn quan sát bằng một số chặng hạn. có thể là tỷ lệ cử tri sẽ bỏ phiếu cho ứng cử viên của bạn. Ngay bây giờ, những khái niệm cuối cùng này có thể hơi mơ hồ, nhưng chúng sẽ trở nên rõ ràng hơn nhiều khi bạn đọc tiếp.

Một mô hình thông kê phục vụ hai mục đích chính trong phân tích dữ liệu, đó là cung cấp một bản tóm tắt định lượng về

dữ liệu và áp đặt một cấu trúc cụ thể đối với dân số mà dữ liệu được lấy mẫu. Đôi khi rất hữu ích để hiểu mô hình là gì và tại sao nó có thể hữu ích thông qua minh họa của các ví dụ cục đoan. "Mô hình" thường đơn giản là không có mô hình nào cả.

Hãy tưởng tượng bạn muốn thực hiện một cuộc khảo sát với 20 người để hỏi xem họ săn sàng chi bao nhiêu cho một sản phẩm mà bạn đang phát triển. Mục tiêu của cuộc khảo sát này là gì? Tóm lại, nếu bạn đang dành thời gian và tiền bạc để phát triển một sản phẩm mới, bạn tin rằng có một lượng lớn người săn sàng mua sản phẩm này.

Tuy nhiên, quá tốn kém và phức tạp để hỏi mọi người trong cộng đồng đó xem họ săn sàng trả bao nhiêu. Vì vậy, bạn lấy một mẫu từ dân số đó để biết dân số đó sẽ trả bao nhiêu.

Một trong số chúng tôi (Roger) gần đây đã xuất bản một cuốn sách có tựa đề [R Programming for Data Science¹](#). Trước khi cuốn sách được xuất bản, những độc giả quan tâm có thể gửi tên và địa chỉ email quảng cáo của họ tới trang web của cuốn sách để được thông báo về việc xuất bản cuốn sách. Ngoài ra, có một tùy chọn để chỉ định số tiền họ săn sàng trả cho cuốn sách. Dưới đây là mẫu ran dom gồm 20 phản hồi từ những người tình nguyện cung cấp thông tin này.

25 20 15 5 30 7 5 10 12 40 30 30 10 25 10 20 10 10 25 5

Bây giờ, giả sử rằng ai đó đã hỏi bạn, "Dữ liệu nói gì?" Một điều bạn có thể làm chỉ đơn giản là bàn giao dữ liệu-tất cả 20 số. Vì tập dữ liệu không quá lớn nên đây không phải là một gánh nặng lớn. Cuối cùng, câu trả lời cho câu hỏi của họ nằm trong tập dữ liệu đó, nhưng có tất cả dữ liệu không phải là một bản tóm tắt dưới bất kỳ hình thức nào. Có tất cả dữ liệu là quan trọng, nhưng

¹ <https://leanpub.com/rprogramming>

thường không hữu ích lắm. Điều này là do mô hình tóm tắt thường không làm giảm dữ liệu.

Yếu tố quan trọng đầu tiên của một mô hình thống kê là giảm thiểu dữ liệu. Ý tưởng cơ bản là bạn muốn lấy tập hợp số ban đầu bao gồm tập dữ liệu của mình và biến đổi chúng thành một tập hợp số nhỏ hơn. Nếu ban đầu bạn bắt đầu với 20 số, thì mô hình của bạn sẽ tạo ra một bản tóm tắt có ít hơn 20 số. Quá trình giảm dữ liệu thường kết thúc bằng một thống kê. Nói chung, một thống kê là bất kỳ bản tóm tắt dữ liệu nào. Giá trị trung bình của mẫu, hoặc trung bình, là một thống kê. Trung vị, độ lệch chuẩn, mức tối đa, mức tối thiểu và phạm vi cũng vậy. Một số thống kê ít nhiều hữu ích hơn những thống kê khác nhưng tất cả chúng đều là bản tóm tắt của dữ liệu.

Có lẽ cách giảm dữ liệu đơn giản nhất mà bạn có thể tạo ra là giá trị trung bình hoặc trung bình số học đơn giản của dữ liệu, trong trường hợp này là 17,2 đô la. Đi từ 20 con số về 1 con số là mức giảm tối đa bạn có thể làm trong trường hợp này, vì vậy nó chắc chắn đáp ứng yếu tố tóm tắt của một mô hình.

5.1 Mô hình như Kỳ vọng

Nhưng một thống kê tóm tắt đơn giản, chẳng hạn như giá trị trung bình của một tập hợp số, là không đủ để xây dựng một mô hình. Một mô hình thống kê cũng phải áp đặt một số cấu trúc trên dữ liệu. Về cốt lõi, một mô hình thống kê cung cấp một mô tả về cách thế giới hoạt động và cách dữ liệu được tạo ra. Mô hình về cơ bản là kỳ vọng về mối quan hệ giữa các yếu tố khác nhau trong thế giới thực và trong tập dữ liệu của bạn.

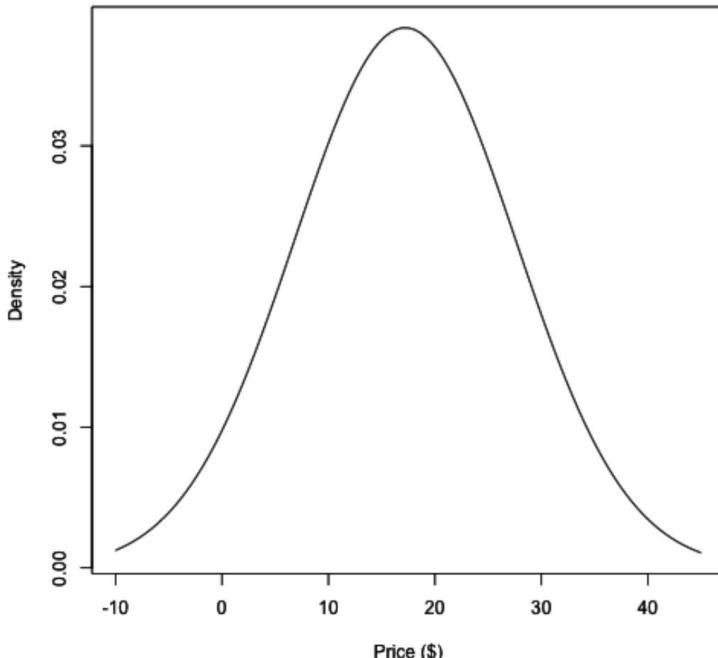
Điều làm cho một mô hình trở thành một mô hình thống kê là nó cho phép một số tính ngẫu nhiên trong việc tạo dữ liệu.

Áp dụng mô hình bình thường

Có lẽ mô hình thống kê phổ biến nhất trên thế giới là mô hình Bình thường. Mô hình này nói rằng tính ngẫu nhiên trong một tập hợp dữ liệu có thể được giải thích bằng phân phối Chuẩn hoặc đường cong hình chuông. Phân phối chuẩn được xác định đầy đủ bởi hai tham số—giá trị trung bình và độ lệch chuẩn.

Lấy dữ liệu mà chúng tôi đã mô tả trong phần trước - số tiền mà 20 người săn sàng trả cho một sản phẩm mới giả định. Hy vọng là 20 người này là mẫu đại diện cho toàn bộ những người có thể mua sản phẩm mới này. Nếu đúng như vậy, thì thông tin chứa trong bộ dữ liệu có thể cho bạn biết điều gì đó về mọi người trong quần thể.

Để áp dụng mô hình Bình thường cho tập dữ liệu này, chúng ta chỉ cần tính giá trị trung bình và độ lệch chuẩn. Trong trường hợp này, giá trị trung bình là \$17,2 và độ lệch chuẩn là \$10,39. Với những tham số đó, kỳ vọng của chúng tôi trong mô hình Bình thường là sự phân bổ giá mà mọi người săn sàng trả sẽ giống như thế này.



Mô hình bình thường cho giá

Theo mô hình, khoảng 68% dân số sẽ sẵn sàng trả khoảng từ 6,81 đô la đến 27,59 đô la cho sản phẩm mới này. Đó có phải là thông tin hữu ích hay không tùy thuộc vào các chi tiết cụ thể của tình huống mà chúng tôi sẽ đề cập đến vào lúc này.

Bạn có thể sử dụng mô hình thống kê để trả lời các câu hỏi phức tạp hơn nếu muốn. Ví dụ: giả sử bạn muốn biết "Tỷ lệ dân số sẵn sàng trả hơn 30 đô la cho cuốn sách này là bao nhiêu?" Sử dụng các thuộc tính của phân phối Chuẩn (và một chút trợ giúp tính toán từ R), chúng ta có thể dễ dàng thực hiện phép tính này.

```
pnorm(30, mean = mean(x), sd = sd(x), Lower.tail = FALSE)
```

```
[1] 0.1089893
```

Vì vậy, khoảng 11% dân số sẽ sẵn sàng trả hơn 30 đô la cho sản phẩm. Một lần nữa, điều này có hữu ích cho bạn hay không tùy thuộc vào các mục tiêu cụ thể của bạn.

Lưu ý rằng trong hình trên có một thứ quan trọng bị thiếu-dữ liệu! Điều đó không hoàn toàn đúng, bởi vì chúng tôi đã sử dụng dữ liệu để vẽ bức tranh (để tính giá trị trung bình và độ lệch chuẩn của phân phối Chuẩn), nhưng cuối cùng dữ liệu không xuất hiện trực tiếp trong biểu đồ. Trong trường hợp này, chúng tôi đang sử dụng phân phối Chuẩn để cho chúng tôi biết dân số trông như thế nào chứ không phải dữ liệu trông như thế nào.

Điểm mấu chốt ở đây là chúng tôi đã sử dụng phân phối Chuẩn để thiết lập hình dạng của phân phối mà chúng tôi mong đợi dữ liệu tuân theo. Phân phối chuẩn là kỳ vọng của chúng tôi về dữ liệu sẽ như thế nào.

5.2 So sánh Kỳ vọng của Mô hình với Thực tế

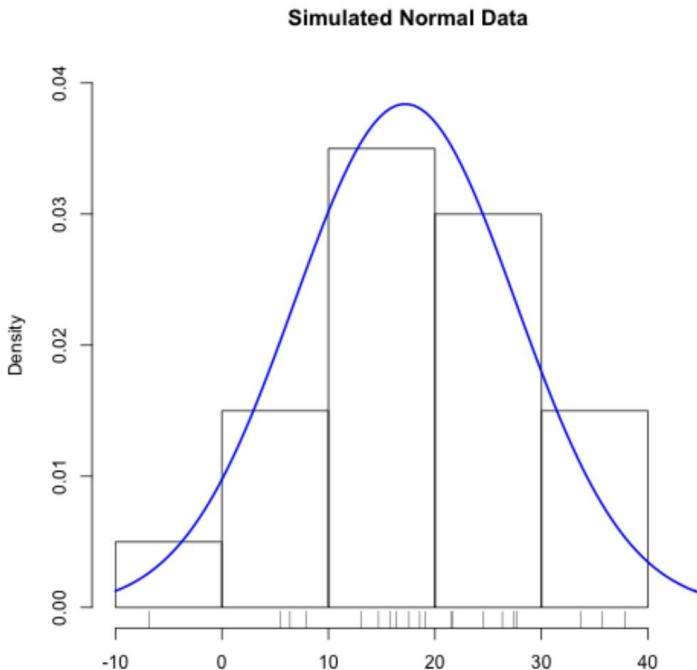
Chúng tôi có thể rất tự hào về việc phát triển mô hình thống kê của mình, nhưng cuối cùng thì tính hữu ích của nó sẽ phụ thuộc vào mức độ phản ánh của nó đối với dữ liệu chúng tôi thu thập trong thế giới thực. Làm thế nào để chúng ta biết nếu kỳ vọng của chúng tôi phù hợp với thực tế?

Vẽ một bức tranh giả

Để bắt đầu, chúng ta có thể tạo một số hình ảnh, chẳng hạn như biểu đồ dữ liệu. Nhưng trước khi chúng ta lấy dữ liệu, hãy tìm hiểu

những gì chúng tôi mong đợi để xem từ dữ liệu. Nếu dân số tuân theo một phân phối Chuẩn và dữ liệu là một mẫu ngẫu nhiên từ dân số đó, thì phân phối được tính bởi biểu đồ sẽ giống như mô hình lý thuyết được cung cấp bởi phân phối Chuẩn.

Trong hình bên dưới, tôi đã mô phỏng 20 điểm dữ liệu từ phân phối Chuẩn và phủ đường cong Chuẩn lý thuyết lên trên biểu đồ.



Biểu đồ dữ liệu bình thường mô phỏng

Lưu ý mức độ khớp của các thanh biểu đồ và đường cong màu xanh lam. Đây là những gì chúng tôi muốn thấy với dữ liệu. Nếu chúng ta thấy

này, thì chúng ta có thể kết luận rằng phân phối Chuẩn là một mô hình thống kê tốt cho dữ liệu.

Mô phỏng dữ liệu từ một mô hình giả thuyết, nếu có thể, là một cách hay để thiết lập các kỳ vọng trước khi bạn xem dữ liệu.

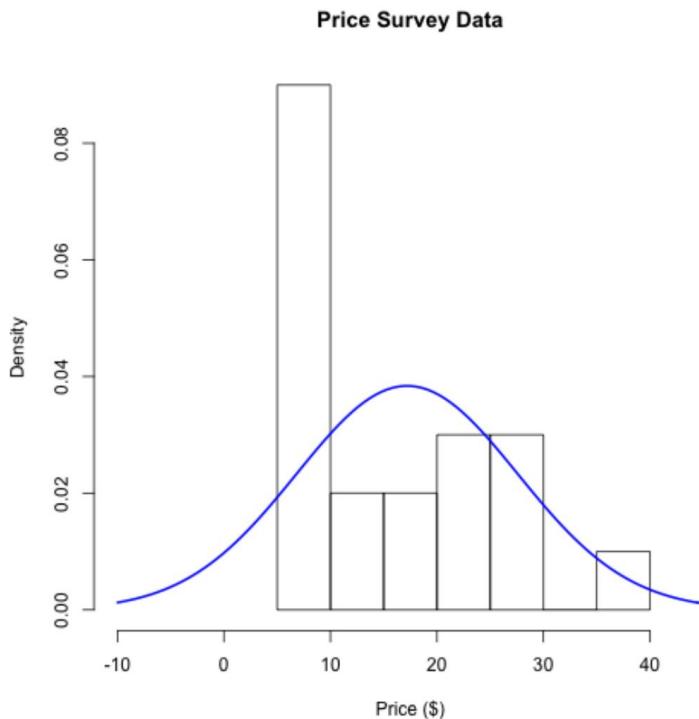
Vẽ một bức tranh giả (thậm chí bằng tay, nếu bạn phải làm) có thể là một công cụ rất hữu ích để bắt đầu các cuộc thảo luận về mô hình và những gì chúng ta mong đợi từ thực tế.

Ví dụ, thậm chí trước khi xem xét dữ liệu, chúng ta có thể nghĩ ngờ rằng mô hình Bình thường có thể không cung cấp một biểu hiện hoàn hảo về tổng thể. Cụ thể, phân phối Chuẩn cho phép các giá trị âm, nhưng chúng tôi không thực sự mong đợi rằng mọi người sẽ nói rằng họ sẵn sàng trả số tiền âm cho một cuốn sách.

Vì vậy, chúng tôi đã có một số bằng chứng rằng mô hình Bình thường có thể không phải là một mô hình hoàn hảo, nhưng không có mô hình nào là hoàn hảo. Câu hỏi đặt ra là liệu mô hình thống kê có cung cấp một xấp xỉ hợp lý có thể hữu ích theo một cách nào đó không?

hình thật

Dưới đây là biểu đồ dữ liệu từ mẫu của 20 người trả lời. Trên đầu biểu đồ, tôi đã phủ Đường cong thông thường lên trên biểu đồ gồm 20 điểm dữ liệu về số tiền mà mọi người nói rằng họ sẵn sàng trả cho cuốn sách.



Biểu đồ dữ liệu khảo sát giá

Những gì chúng tôi mong đợi là biểu đồ và đường màu xanh sẽ gần như theo sát nhau. Làm thế nào để so sánh mô hình và thực tế?

Thoạt nhìn, có vẻ như biểu đồ và phân phối Chuẩn không khớp lắm. Biểu đồ có mức tăng đột biến lớn khoảng 10 đô la, một tính năng không có ở đường cong màu xanh lam. Ngoài ra, phân phối Chuẩn cho phép các giá trị âm ở phía bên trái của biểu đồ, nhưng không có điểm dữ liệu nào trong vùng đó của biểu đồ.

Cho đến nay, dữ liệu cho thấy rằng mô hình Bình thường không thực sự là một đại diện tốt cho dân số, dựa trên dữ liệu

mà chúng tôi đã lấy mẫu từ dân số. Có vẻ như 20 người được khảo sát rất thích trả mức giá gần 10 đô la, trong khi có một số ít người sẵn sàng trả nhiều hơn thế. Các tính năng này của dữ liệu không được đặc trưng bởi phân phối Chuẩn.

5.3 Phản ứng với dữ liệu: Tinh chỉnh các kỳ vọng của chúng tôi

Được rồi, do đó, mô hình và dữ liệu không khớp lắm, như được biểu thị bằng biểu đồ ở trên. Vì vậy, những gì làm gì? Vâng, chúng ta có thể

1. Lấy một mô hình khác; hoặc
2. Lấy dữ liệu khác

Hoặc chúng ta có thể làm cả hai. Những gì chúng tôi làm để đáp lại phụ thuộc một chút vào niềm tin của chúng tôi về mô hình và sự hiểu biết của chúng tôi về quy trình thu thập dữ liệu. Nếu chúng ta cảm thấy mạnh mẽ rằng tổng số mức giá mà mọi người sẵn sàng trả nên tuân theo phân phối Chuẩn, thì chúng ta có thể ít có khả năng thực hiện các sửa đổi lớn đối với mô hình. Chúng tôi có thể kiểm tra quá trình thu thập dữ liệu để xem liệu nó có thể dẫn đến một số sai lệch trong dữ liệu hay không. Tuy nhiên, nếu quy trình thu thập dữ liệu hợp lý, thì chúng tôi có thể buộc phải kiểm tra lại mô hình của mình đối với dân số và xem có thể thay đổi điều gì. Trong trường hợp này, có khả năng là mô hình của chúng ta không phù hợp, đặc biệt là do khó hình dung một quy trình thu thập dữ liệu hợp lệ có thể dẫn đến các giá trị âm trong dữ liệu (như phân phối Chuẩn cho phép).

Để kết thúc vòng lặp ở đây, chúng tôi sẽ chọn một mô hình thống kê khác để đại diện cho dân số, phân phối Gamma.

Bản phân phối này có đặc điểm là nó chỉ cho phép tích cực

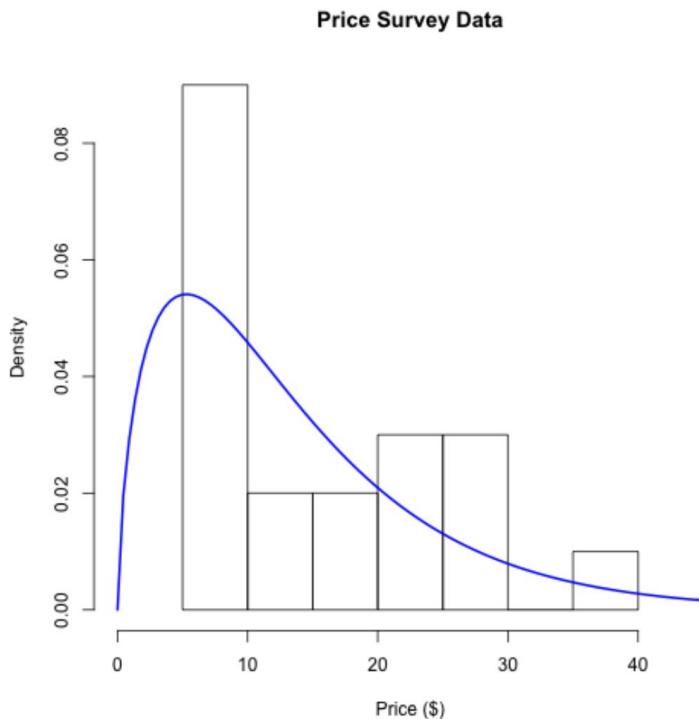
các giá trị, vì vậy nó loại bỏ vấn đề chúng tôi gặp phải với các giá trị âm với phân phối Chuẩn.

Bây giờ, chúng ta nên quay lại phần đầu của bước lặp và thực hiện như sau:

1. Phát triển kỳ vọng: Vẽ một bức tranh giả-chúng ta mong đợi thấy gì trước khi xem dữ liệu?
2. So sánh kỳ vọng của chúng tôi với dữ liệu 3.

Tinh chỉnh kỳ vọng của chúng tôi, dựa trên những gì dữ liệu cho thấy

Để bạn tham khảo, đây là biểu đồ của cùng một dữ liệu với phân phối Gamma (được ước tính bằng dữ liệu) được phủ lên.



Dữ liệu khảo sát giá với phân phối Gamma

Làm thế nào để dữ liệu phù hợp với mong đợi của bạn bây giờ?

Bạn có thể hỏi mô hình mà tôi sử dụng để biểu thị dân số mà dữ liệu được tạo ra có gì khác biệt? Chà, đối với những người mới bắt đầu, nó có thể ảnh hưởng đến các loại dự đoán mà bạn có thể thực hiện khi sử dụng mô hình. Ví dụ, nhớ lại trước đó họ quan tâm đến tỷ lệ dân số có thể sẵn sàng trả ít nhất 30 đô la cho cuốn sách. Mô hình mới của chúng tôi nói rằng chỉ có khoảng 7% dân số sẵn sàng trả ít nhất số tiền này (mô hình Bình thường tuyên bố 11% sẽ trả 30 đô la trở lên). Vì vậy, các mô hình khác nhau có thể mang lại những dự đoán khác nhau dựa trên

cùng một dữ liệu, có thể ảnh hưởng đến các quyết định được đưa ra đường.

5.4 Kiểm tra mối quan hệ tuyến tính

Việc xem xét dữ liệu và cố gắng hiểu tuyến tính là điều bình thường mối quan hệ giữa các biến quan tâm. Kỹ thuật thống kê phổ biến nhất để trợ giúp với nhiệm vụ này là tuyến tính hồi quy. Chúng ta có thể áp dụng các nguyên tắc đã thảo luận ở trên-phát triển kỳ vọng, so sánh kỳ vọng của chúng tôi với dữ liệu, tinh chỉnh kỳ vọng của chúng tôi-để áp dụng tuyến tính hồi quy là tốt.

Đối với ví dụ này, chúng ta sẽ xem xét một bộ dữ liệu chất lượng không khí đơn giản chứa thông tin về mức ozone tầng đối lưu trong

Thành phố New York vào năm 1999 trong các tháng từ tháng 5 đến 1999. Đây là một vài hàng đầu tiên của tập dữ liệu.

	khí quyển	tháng tạm thời
1	25.37262	55.33333
2	32.83333	57.66667
3	28.88667	56.66667
4	12.06854	56.66667
5	11.21920	63.66667
6	13.19110	60.00000

Dữ liệu chứa mức ozone trung bình hàng ngày (trong các phần trên một tỷ [ppb]) và nhiệt độ (tính bằng độ F).

Một câu hỏi quan tâm có thể thúc đẩy bộ sưu tập của tập dữ liệu này là “Nhiệt độ xung quanh liên quan như thế nào đến mức ôzôn xung quanh ở New York?”

kỳ vọng

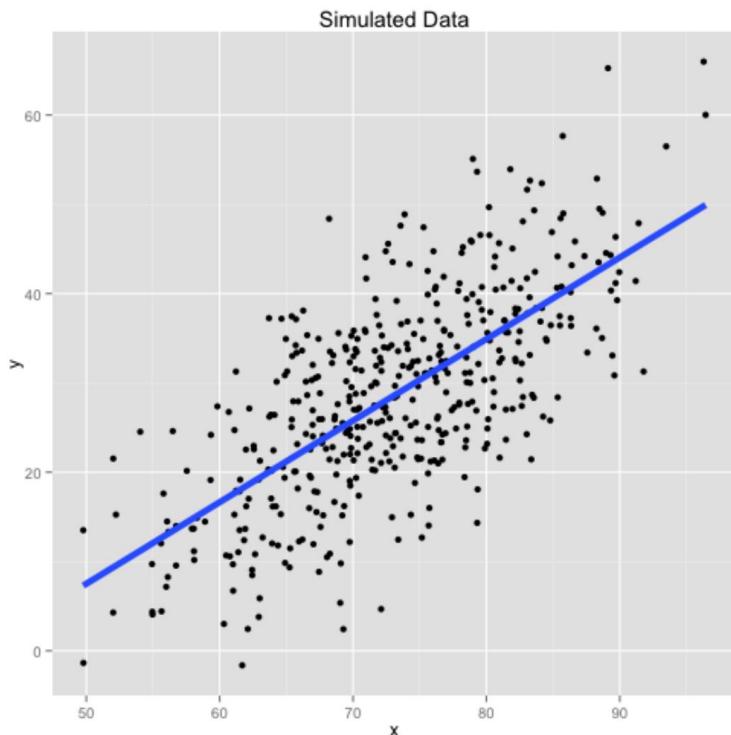
Sau khi đọc một chút về **sự hình thành ôzôn trong quả cầu atmo²**, chúng ta biết rằng sự hình thành của ôzôn phụ thuộc rất nhiều vào sự hiện diện của ánh sáng mặt trời. Ánh sáng mặt trời cũng liên quan đến nhiệt độ theo nghĩa là vào những ngày có nhiều ánh sáng mặt trời, chúng ta sẽ mong đợi nhiệt độ trung bình của ngày hôm đó sẽ cao hơn. Những ngày nhiều mây có cả nhiệt độ trung bình thấp hơn và ít ôzôn hơn. Vì vậy, có lý do để tin rằng vào những ngày có nhiệt độ cao hơn, chúng ta sẽ mong đợi mức ôzôn cao hơn. Đây là một mối quan hệ gián tiếp—chúng ta đang sử dụng nhiệt độ ở đây về cơ bản như một đại diện cho lượng ánh sáng mặt trời.

Mô hình đơn giản nhất mà chúng ta có thể xây dựng để mô tả đặc điểm của mối quan hệ giữa nhiệt độ và ôzôn là một mô hình tuyến tính. Mô hình này nói rằng khi nhiệt độ tăng, lượng ôzôn trong khí quyển tăng tuyến tính với nó. Chúng ta mong đợi điều này trông như thế nào?

Chúng ta có thể mô phỏng một số dữ liệu để tạo ra một bức tranh giả về mối quan hệ giữa ôzôn và nhiệt độ sẽ như thế nào dưới một mô hình tuyến tính. Đây là một mối quan hệ tuyến tính đơn giản cùng với dữ liệu mô phỏng trong biểu đồ phân tán.

²https://en.wikipedia.org/wiki/Tropospheric_ozone

Sử dụng các mô hình để khám phá dữ liệu của bạn



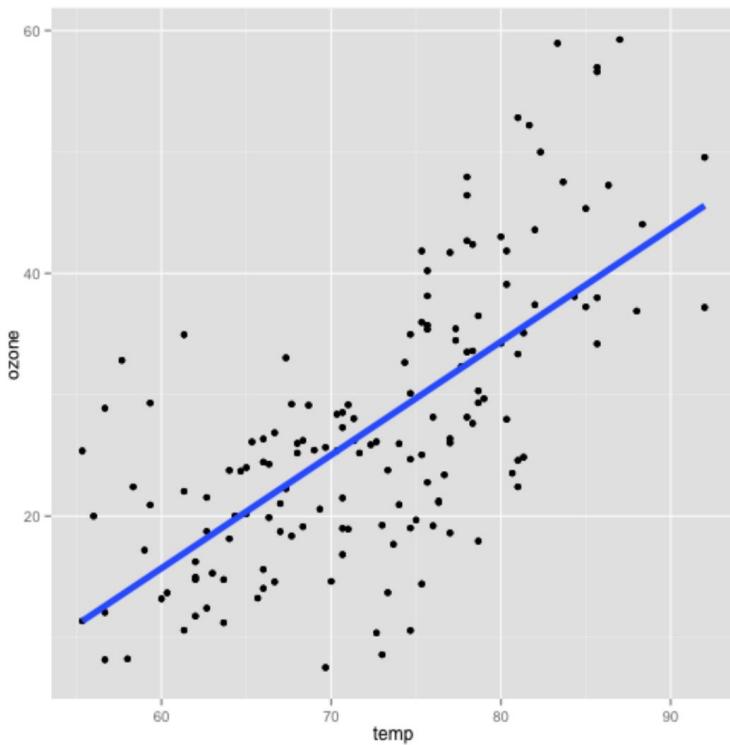
Dữ liệu mô phỏng với mô hình tuyến tính

Lưu ý rằng nếu bạn chọn bất kỳ điểm nào trên đường màu xanh lam, thì số điểm phía trên đường kẻ và số điểm phía dưới đường kẻ gần như bằng nhau (điều này còn được gọi là lỗi không thiên vị). Ngoài ra, các điểm trên biểu đồ phân tán đường như tăng tuyến tính khi bạn di chuyển về phía bên phải trên trục x, ngay cả khi có khá nhiều nhiễu/phân tán dọc theo đường.

Nếu chúng ta đúng về mô hình tuyến tính của mình, và đó là mô hình đặc trưng cho dữ liệu và mối quan hệ giữa ôzôn và nhiệt độ, thì nói một cách đại khái, đây là bức tranh chúng ta sẽ thấy khi vẽ biểu đồ dữ liệu.

So sánh kỳ vọng với dữ liệu

Đây là hình ảnh về dữ liệu nhiệt độ và ôzôn thực tế ở Thành phố New York trong năm 1999. Trên biểu đồ phân tán của dữ liệu, chúng tôi đã vẽ đồ thị đường hồi quy tuyến tính phù hợp được ước tính bằng dữ liệu.



Mô hình tuyến tính cho Ozone và nhiệt độ

Làm thế nào để hình ảnh này so sánh với hình ảnh mà bạn đang mong đợi để xem?

Có một điều rõ ràng: Dường như có một xu hướng gia tăng về ôzôn khi nhiệt độ tăng, như chúng ta đưa ra giả thuyết-

có kích thước. Tuy nhiên, có một vài sai lệch so với bức tranh giả đẹp mà chúng tôi đã thực hiện ở trên. Các điểm đường như không cân bằng xung quanh đường hồi quy màu xanh lam.

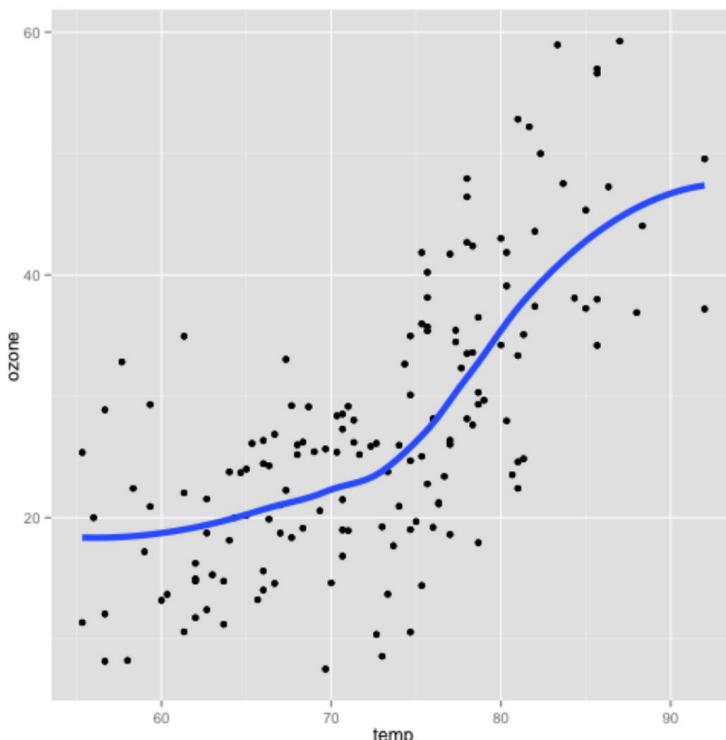
Nếu bạn vẽ một đường thẳng đứng xung quanh nhiệt độ 85 độ, bạn sẽ nhận thấy rằng hầu hết các điểm đều nằm trên đường thẳng. Vẽ một đường thẳng đứng quanh 70 độ cho thấy hầu hết các điểm nằm bên dưới đường thẳng. Điều này ngũ ý rằng ở nhiệt độ cao hơn, mô hình của chúng tôi thiên về phía dưới (nó đánh giá thấp ôzôn) và ở nhiệt độ vừa phải, mô hình của chúng tôi thiên về phía trên. Đây không phải là một tính năng tuyệt vời-trong tình huống này, chúng tôi có thể muốn mô hình của chúng tôi không bị sai lệch ở bất kỳ đâu.

Mô hình hồi quy tuyến tính đơn giản của chúng tôi đường như nắm bắt được mối quan hệ ngày càng tăng chung giữa nhiệt độ và ôzôn, nhưng nó đường như bị sai lệch trong các phạm vi nhiệt độ nhất định. Có vẻ như mô hình này cần cải thiện nếu chúng ta muốn mô tả rõ hơn mối quan hệ giữa nhiệt độ và ôzôn trong bộ dữ liệu này.

tinh chỉnh kỳ vọng

Từ hình trên, có vẻ như mối quan hệ giữa nhiệt độ và ozone có thể không tuyến tính. Thật vậy, các điểm dữ liệu cho thấy rằng có thể mối quan hệ không thay đổi cho đến khoảng 70 độ và sau đó nồng độ ôzôn tăng nhanh theo nhiệt độ sau đó. Điều này cho thấy một mối quan hệ phi tuyến tính giữa nhiệt độ và ozone.

Cách dễ nhất để chúng ta có thể nắm bắt được kỳ vọng đã sửa đổi này là làm mượt mà hơn, trong trường hợp này là mượt mà hơn.



Loess Smoother cho Ozone và Nhiệt độ

Cốt truyện này cho thấy một bức tranh khác - mối quan hệ đang dần tăng lên cho đến khoảng 75 độ, và sau đó gấp nếp rõ rệt sau đó.

Khoảng 90 độ, có một gợi ý rằng mối quan hệ lại giảm xuống.

Chất làm mịn (như hoàng thổ) là những công cụ hữu ích vì chúng nhanh chóng nắm bắt các xu hướng trong tập dữ liệu mà không tạo ra bất kỳ cấu trúc nào giả định về dữ liệu. Về cơ bản, chúng là một cách tự động hoặc được vi tính hóa để vẽ một đường cong trên một số dữ liệu. Tuy nhiên, smooths hiếm khi cho bạn biết bất cứ điều gì về cơ chế của mối quan hệ và do đó có thể bị hạn chế trong ý nghĩa đó. Để hiểu thêm về mối quan hệ

giữa nhiệt độ và ôzôn, chúng ta có thể cần phải dùng đến một mô hình chi tiết hơn so với mô hình tuyến tính đơn giản mà chúng ta có trước đây.

5.5 Khi Nào Chúng Ta Dừng Lại?

Trong các ví dụ trên, chúng tôi đã hoàn thành một lần lặp lại quy trình phân tích dữ liệu. Trong một số trường hợp, một lần lặp có thể là đủ, nhưng trong hầu hết các trường hợp thực tế, bạn sẽ cần lặp lại ít nhất một vài lần. Từ các ví dụ trên, vẫn còn một số việc phải làm:

- **Dữ liệu khảo sát giá:** Chúng tôi đã kết thúc ví dụ bằng cách điều chỉnh mô hình phân phối Gamma. Nhưng làm thế nào để điều đó phù hợp với dữ liệu? Chúng ta sẽ mong đợi điều gì từ dữ liệu nếu chúng thực sự tuân theo phân phối Gamma (chúng tôi chưa bao giờ thực hiện âm mưu đó)? Có cách nào tốt hơn để nắm bắt mức tăng đột biến trong phân phối ngay khoảng 10 đô la không?
- **Ôzôn và Nhiệt độ:** Càng mịn gợi ý về mối quan hệ phi tuyến tính giữa nhiệt độ và ôzôn, nhưng lý do cho điều này là gì? Sự phi tuyến tính có thật hay chỉ là một sự ngẫu nhiên xảy ra trong dữ liệu? Có một quá trình vật lý đã biết giải thích sự gia tăng mạnh mẽ nồng độ ôzôn vượt quá nhiệt độ nhất định và chúng ta có thể mô hình hóa quá trình đó không?

Cuối cùng, bạn có thể lặp đi lặp lại nhiều lần.

Mỗi câu trả lời thường sẽ đặt ra nhiều câu hỏi hơn và yêu cầu đào sâu hơn vào dữ liệu. Khi chính xác thì bạn dừng quá trình sau đó? Lý thuyết thống kê gợi ý một số cách tiếp cận khác nhau để xác định khi nào một mô hình thống kê là “đủ tốt” và phù hợp với dữ liệu. Đây không phải là điều chúng ta sẽ thảo luận ở đây mà thay vào đó, chúng ta sẽ thảo luận về một số tiêu chí cấp cao để xác định thời điểm bạn có thể cân nhắc dừng lặp lại phân tích dữ liệu.

Bạn hết dữ liệu?

Phân tích dữ liệu lặp đi lặp lại cuối cùng sẽ bắt đầu đặt ra những câu hỏi mà đơn giản là không thể trả lời bằng dữ liệu có sẵn.

Ví dụ, trong phân tích ôzôn/nhiệt độ, mô hình đề xuất rằng không chỉ có một mối quan hệ đơn giản giữa hai biến, nó có thể là phi tuyến tính. Nhưng dữ liệu không thể giải thích chính xác tại sao một mối quan hệ phi tuyến tính như vậy có thể tồn tại (mặc dù chúng có thể gợi ý một số các giả thuyết). Ngoài ra, bạn có thể cần thu thập thêm dữ liệu để xác định xem những gì bạn quan sát là có thật hay chỉ đơn giản là một may mắn hoặc tai nạn thống kê. Dù bằng cách nào, bạn cần phải quay trở lại ra thế giới và thu thập dữ liệu mới. Phân tích dữ liệu nhiều hơn không chắc sẽ mang lại những câu trả lời này.

Một tình huống khác mà bạn có thể thấy mình đang tìm kiếm ra thêm dữ liệu là khi bạn đã thực sự hoàn thành dữ liệu phân tích và đi đến kết quả khả quan, thường là một số phát hiện thú vị. Sau đó, nó có thể rất quan trọng để cố gắng sao chép bất cứ điều gì bạn đã tìm thấy bằng cách sử dụng khác, có thể độc lập, tập dữ liệu. Trong ví dụ về ôzôn/nhiệt độ, nếu chúng ta kết luận rằng có một mối quan hệ phi tuyến tính giữa nhiệt độ và ozone, kết luận của chúng tôi có thể là trở nên mạnh mẽ hơn nếu chúng ta có thể chỉ ra rằng mối quan hệ này đã có mặt ở các thành phố khác ngoài New York. Xác nhận độc lập như vậy có thể làm tăng sức mạnh của bằng chứng và có thể đóng một vai trò mạnh mẽ trong việc ra quyết định.

Bạn có đủ bằng chứng để đưa ra quyết định?

Phân tích dữ liệu thường được tiến hành để hỗ trợ việc ra quyết định, dù là trong kinh doanh, học viện, chính phủ hay ở những nơi khác, chúng tôi thường thu thập dữ liệu phân tích để cung cấp thông tin cho một số loại quyết định. Điều quan trọng là phải nhận ra rằng việc phân tích mà bạn thực hiện để đưa mình đến điểm mà bạn

có thể đưa ra quyết định về điều gì đó có thể rất khác so với phân tích mà bạn thực hiện để đạt được các mục tiêu khác, chẳng hạn như viết báo cáo, xuất bản bài báo hoặc đưa ra sản phẩm hoàn chỉnh.

Đó là lý do tại sao điều quan trọng là phải luôn ghi nhớ mục đích của phân tích dữ liệu khi bạn thực hiện vì bạn có thể đầu tư quá nhiều hoặc dưới mức các nguồn lực vào phân tích nếu phân tích không phù hợp với mục tiêu cuối cùng. Mục đích của phân tích dữ liệu có thể thay đổi theo thời gian và trên thực tế có thể có nhiều mục đích song song. Câu hỏi liệu bạn có đủ bằng chứng hay không tùy thuộc vào các yếu tố cụ thể đối với ứng dụng hiện có và hoàn cảnh cá nhân của bạn liên quan đến chi phí và lợi ích. Nếu bạn cảm thấy mình không có đủ bằng chứng để đưa ra quyết định, đó có thể là do bạn không có dữ liệu hoặc do bạn cần tiến hành phân tích thêm.

Bạn có thể đặt kết quả của mình trong bất kỳ bối cảnh lớn hơn nào không?

Một cách khác để đặt câu hỏi này là "Kết quả có hợp lý không?" Thông thường, bạn có thể trả lời câu hỏi này bằng cách tìm kiếm tài liệu có sẵn trong khu vực của mình hoặc xem liệu những người khác trong hoặc ngoài tổ chức của bạn có đưa ra kết luận tương tự hay không. Nếu kết quả phân tích của bạn chật chẽ với những gì người khác đã tìm thấy, đó có thể là một điều tốt, nhưng đó không phải là kết quả mong muốn duy nhất. Những phát hiện mâu thuẫn với kết quả trong quá khứ có thể dẫn đến một con đường khám phá mới. Trong cả hai trường hợp, thường rất khó để đi đến câu trả lời đúng nếu không điều tra thêm.

Bạn phải cẩn thận một chút với cách bạn trả lời câu hỏi này. Thông thường, đặc biệt là với các bộ dữ liệu rất lớn và phức tạp, thật dễ dàng để đạt được kết quả "có ý nghĩa" và phù hợp với hiểu biết của chúng ta về cách thức hoạt động của một quy trình nhất định. Trong tình huống này, điều quan trọng là phải cực kỳ chỉ trích những phát hiện của chúng ta và thách thức chúng càng nhiều càng tốt. Trong cựu của chúng tôi

kinh nghiệm, khi dữ liệu rất khớp với kỳ vọng của chúng tôi, đó có thể là kết quả của sai lầm hoặc hiểu lầm trong quá trình phân tích hoặc trong quá trình thu thập dữ liệu. Điều quan trọng là phải đặt câu hỏi về mọi khía cạnh của quá trình phân tích để đảm bảo mọi thứ được thực hiện một cách thích hợp.

Nếu kết quả của bạn không có ý nghĩa hoặc dữ liệu không phù hợp với mong đợi của bạn, thì đây là lúc mọi thứ trở nên thú vị.

Có thể đơn giản là bạn đã làm sai điều gì đó trong quá trình phân tích hoặc thu thập dữ liệu. Rất có thể, đó chính xác là những gì đã xảy ra. Đối với mỗi viên kim cương thô, có 99 mảnh than. Tuy nhiên, trong trường hợp bạn phát hiện ra điều gì đó bất thường mà những người khác chưa thấy, bạn sẽ cần (a) đảm bảo rằng phân tích được thực hiện đúng cách và (b) sao chép phát hiện của bạn trong một tập dữ liệu khác.

Kết quả đáng ngạc nhiên thường được đáp ứng với nhiều sự xem xét kỹ lưỡng và bạn sẽ cần chuẩn bị để bảo vệ công việc của mình một cách nghiêm ngặt.

Cuối cùng, nếu phân tích của bạn dẫn bạn đến một nơi mà bạn có thể trả lời dứt khoát câu hỏi "Kết quả có hợp lý không?" thì bắt kể bạn trả lời câu hỏi đó như thế nào, bạn có thể cần phải dừng phân tích của mình và kiểm tra cẩn thận mọi phần của nó.

Bạn đã hết thời gian?

Tiêu chí này có vẻ tùy ý nhưng lại đóng một vai trò lớn trong việc xác định thời điểm dừng phân tích trong thực tế.

Một câu hỏi liên quan có thể là "Bạn hết tiền rồi à?" Cuối cùng, sẽ có cả ngân sách thời gian và ngân sách tiền tệ xác định có bao nhiêu nguồn lực có thể được cam kết cho một phân tích nhất định. Nhận thức được những ngân sách này là gì, ngay cả khi bạn không nhất thiết phải kiểm soát chúng, có thể rất quan trọng để quản lý phân tích dữ liệu. Đặc biệt, bạn có thể cần tranh luận để có thêm nguồn lực và thuyết phục người khác đưa chúng cho bạn. Trong một tình huống như vậy,

thật hữu ích khi biết khi nào nên dừng lặp lại phân tích dữ liệu và chuẩn bị bất kỳ kết quả nào bạn có thể thu được cho đến nay để trình bày lập luận mạch lạc để tiếp tục phân tích.

5.6 Tóm tắt

Xây dựng mô hình, giống như toàn bộ quá trình phân tích dữ liệu, là một quá trình lặp đi lặp lại. Các mô hình được sử dụng để giảm thiểu dữ liệu và cung cấp cho bạn một số thông tin chi tiết về dân số mà bạn đang cố gắng suy luận. Điều quan trọng trước tiên là đặt kỳ vọng của bạn về cách một mô hình sẽ mô tả đặc điểm của tập dữ liệu trước khi bạn thực sự áp dụng một mô hình cho dữ liệu. Sau đó, bạn có thể kiểm tra xem mô hình của bạn phù hợp với mong đợi của bạn như thế nào. Thông thường, sẽ có các tính năng của tập dữ liệu không phù hợp với mô hình của bạn và bạn sẽ phải tinh chỉnh mô hình của mình hoặc kiểm tra quy trình thu thập dữ liệu.

6. Suy luận: Sơ lược

Suy luận là một trong nhiều mục tiêu có thể đạt được trong phân tích dữ liệu và do đó, đáng để thảo luận về hành động suy luận chính xác là gì. Nhớ lại trước đây chúng tôi đã mô tả một trong sáu loại câu hỏi bạn có thể hỏi trong phân tích dữ liệu là một câu hỏi suy luận . Vậy suy luận là gì?

Nói chung, mục tiêu của suy luận là có thể đưa ra một tuyên bố về điều gì đó không được quan sát và lý tưởng nhất là có thể mô tả bất kỳ sự không chắc chắn nào mà bạn có về tuyên bố đó. Suy luận rất khó vì có sự khác biệt giữa những gì bạn có thể quan sát và những gì cuối cùng bạn muốn biết.

6.1 Xác định quần thể

Ngôn ngữ suy luận có thể thay đổi tùy thuộc vào ứng dụng, nhưng thông thường nhất, chúng tôi đề cập đến những thứ chúng tôi không thể quan sát (nhưng muốn biết) dưới dạng dân số hoặc các đặc điểm của dân số và dữ liệu mà chúng tôi quan sát dưới dạng mẫu. Mục tiêu là sử dụng mẫu để bằng cách nào đó đưa ra tuyên bố về dân số. Để làm được điều này, chúng ta cần xác định một số điều.

Xác định dân số là nhiệm vụ quan trọng nhất. Nếu bạn không thể xác định hoặc mô tả dân số một cách mạch lạc, thì bạn không thể suy luận. Chỉ cần dừng lại. Khi bạn đã tìm ra dân số là gì và bạn muốn đưa ra tuyên bố về đặc điểm nào của dân số (ví dụ: giá trị trung bình), thì sau đó bạn có thể dịch nó thành một cụm từ cụ thể hơn.

báo cáo bằng cách sử dụng một mô hình thống kê chính thức (được đề cập ở phần sau của cuốn sách này).

6.2 Mô tả quy trình lấy mẫu

Làm thế nào mà dữ liệu đi từ quần thể đến máy tính của bạn? Có thể mô tả quá trình này là rất quan trọng để xác định liệu dữ liệu có hữu ích cho việc suy luận về các đặc điểm của dân số hay không. Ví dụ điển hình, nếu bạn quan tâm đến độ tuổi trung bình của phụ nữ trong dân số, nhưng quy trình lấy mẫu của bạn bằng cách nào đó được thiết kế sao cho chỉ tạo ra dữ liệu về nam giới, thì bạn không thể sử dụng dữ liệu để suy luận về độ tuổi trung bình của phụ nữ. Phụ nữ. Hiểu quy trình lấy mẫu là chìa khóa để xác định xem mẫu của bạn có đại diện cho dân số quan tâm hay không. Lưu ý rằng nếu bạn gặp khó khăn trong việc mô tả dân số, bạn sẽ gặp khó khăn trong việc mô tả quá trình lấy mẫu dữ liệu từ dân số. Vì vậy, việc mô tả quy trình lấy mẫu phụ thuộc vào khả năng mô tả tổng thể một cách mạch lạc của bạn.

6.3 Mô tả một mô hình cho dân số

Chúng ta cần có một biểu diễn trừu tượng về cách thức các phần tử của quần thể có quan hệ với nhau. Thông thường, điều này xuất hiện dưới dạng một mô hình thống kê mà chúng ta có thể biểu diễn bằng cách sử dụng ký hiệu toán học. Tuy nhiên, trong các tình huống phức tạp hơn, chúng ta có thể sử dụng các biểu diễn thuật toán không thể viết gọn gàng trên giấy (nhiều phương pháp học máy phải được mô tả theo cách này). Mô hình đơn giản nhất có thể là một mô hình tuyến tính đơn giản, chẳng hạn như

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Ở đây, x và y là các đặc điểm của tổng thể và β_0 và β_1 mô tả mối quan hệ giữa các đặc điểm đó (nghĩa là chúng có liên quan tích cực hay tiêu cực?). Phần tử cuối cùng ϵ là phần tử tổng hợp nhằm nắm bắt tất cả các yếu tố góp phần tạo nên sự khác biệt giữa y và giá trị mà chúng ta mong đợi của y , đó là $\beta_0 + \beta_1 x$. Phần cuối cùng này làm cho mô hình trở thành một mô hình thống kê vì chúng ta thường cho phép ϵ là ngẫu nhiên.

Một đặc điểm khác mà chúng ta thường cần đưa ra giả định là các đơn vị khác nhau trong quần thể tương tác với nhau như thế nào. Thông thường, không có bất kỳ thông tin bổ sung nào, chúng tôi sẽ giả định rằng các đơn vị trong tổng thể là độc lập, nghĩa là các phép đo của một đơn vị không cung cấp bất kỳ thông tin nào về các phép đo trên một đơn vị khác. Tốt nhất, giả định này là gần đúng, nhưng nó có thể là một xấp xỉ hữu ích. Trong một số tình huống, chẳng hạn như khi nghiên cứu những thứ được kết nối chặt chẽ trong không gian hoặc thời gian, giả định rõ ràng là sai và chúng ta phải sử dụng các phương pháp mô hình hóa đặc biệt để giải thích cho sự thiếu độc lập.

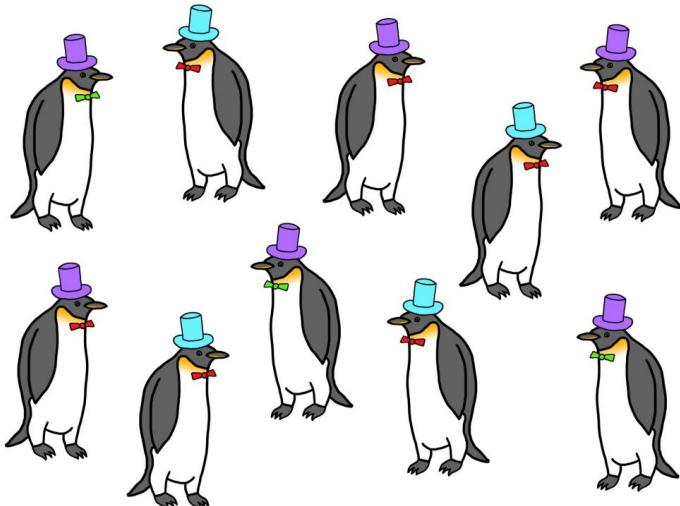
George Box, một nhà thống kê, từng nói rằng¹ "tất cả các mô hình đều sai, nhưng một số là hữu ích". Có khả năng là bất kỳ mô hình nào bạn nghĩ ra để mô tả các đặc điểm của dân số đều sai về mặt kỹ thuật. Nhưng bạn không nên cố định vào việc phát triển một mô hình đúng đắn ; thay vào đó, bạn nên xác định một mô hình hữu ích cho mình và kể một câu chuyện về dữ liệu cũng như về các quy trình cơ bản mà bạn đang cố gắng nghiên cứu.

6.4 Ví dụ nhanh

Hãy xem xét nhóm chim cánh cụt dưới đây (vì chim cánh cụt rất tuyệt vời), mỗi con đội một chiếc mũ màu tím hoặc màu ngọc lam.

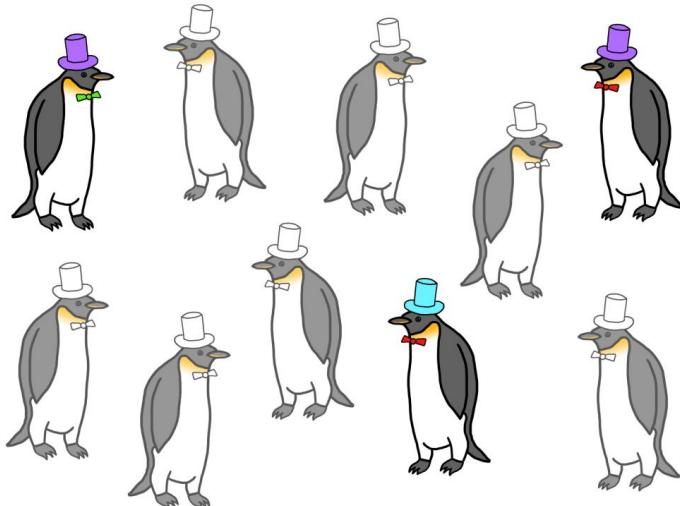
¹ https://en.wikipedia.org/wiki/All_models_are_wrong

Có tổng cộng 10 con chim cánh cụt trong nhóm này. Chúng tôi sẽ gọi chúng là quần thể.



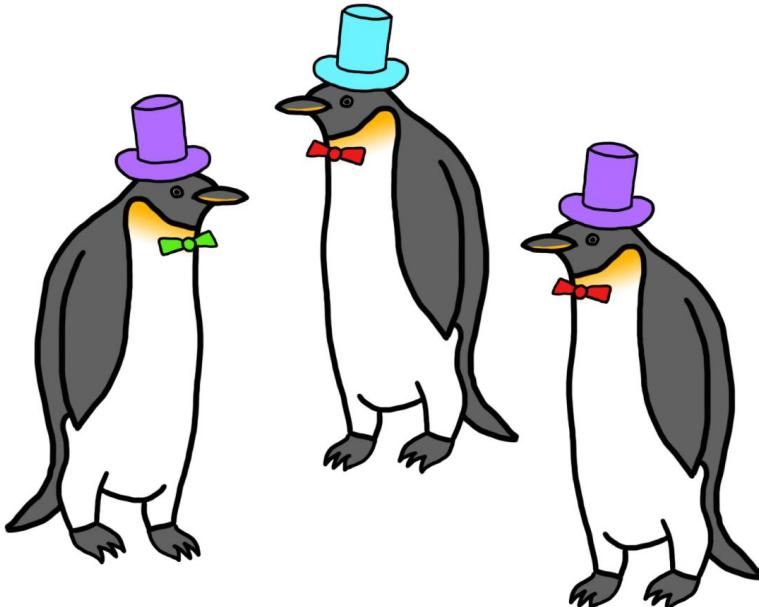
Quần thể chim cánh cụt có mõ màu ngọc lam và màu tím

Bây giờ, giả sử bạn muốn biết tỷ lệ dân số chim cánh cụt đội mũ màu ngọc lam là bao nhiêu. Nhưng có một nhược điểm là bạn không có thời gian, tiền bạc hoặc khả năng chăm sóc 10 chú chim cánh cụt. Ai làm? Bạn chỉ có đủ khả năng để chăm sóc ba con chim cánh cụt, vì vậy bạn lấy mẫu ngẫu nhiên ba trong số 10 con chim cánh cụt này.



Mẫu 3 con chim cánh cụt từ quần thể

Điểm mấu chốt là bạn không bao giờ quan sát được toàn bộ quần thể chim cánh cụt. Vậy giờ, kết quả cuối cùng của bạn là tập dữ liệu chỉ chứa ba chú chim cánh cụt.



Bộ dữ liệu của Penguins

Tại thời điểm này, một câu hỏi dễ đặt ra là "Tỷ lệ chim cánh cụt trong bộ dữ liệu của tôi đội mũ màu ngọc lam là bao nhiêu?". Từ hình trên, rõ ràng là $1/3$ số chim cánh cụt đội đội mũ màu ngọc lam. Chúng tôi không chắc chắn về tỷ lệ đó vì dữ liệu đang ở ngay trước mặt chúng tôi.

Câu hỏi khó đặt ra là "Dựa trên dữ liệu tôi có, tỷ lệ chim cánh cụt trong quần thể ban đầu đội mũ màu ngọc lam là bao nhiêu?" Tại thời điểm này, chúng tôi chỉ có mẫu gồm ba con chim cánh cụt và không quan sát toàn bộ quần thể.

Chúng ta có thể làm gì? Chúng ta cần suy luận về dân số bằng cách sử dụng dữ liệu chúng ta có trong tay.

Ba điều chúng ta cần làm để suy luận
là:

1. Định nghĩa dân số. Ở đây, dân số là 10 con chim cánh cụt ban đầu mà từ đó chúng tôi đã lấy mẫu bộ dữ liệu gồm ba con chim cánh cụt.
2. Mô tả quy trình lấy mẫu. Chúng tôi chưa đề cập rõ ràng về điều này, nhưng giả sử bây giờ "quy trình lấy mẫu" của chúng tôi bao gồm việc lấy ba con chim cánh cụt đầu tiên đến chỗ chúng tôi.
3. Mô tả một mô hình cho quần thể. Chúng ta sẽ giả định rằng những chiếc mũ mà chim cánh cụt đội là độc lập với nhau, vì vậy việc một con chim cánh cụt đội mũ màu tím không ảnh hưởng đến việc liệu một con chim cánh cụt khác có đội mũ màu ngọc lam hay không. Vì chúng tôi chỉ muốn ước tính một tỷ lệ đơn giản của những chú chim cánh cụt đội mũ màu ngọc lam, nên chúng tôi không cần đưa ra bất kỳ giả định phức tạp nào về mối quan hệ của chim cánh cụt với nhau.

Với ba thành phần trên, chúng ta có thể ước tính tỷ lệ chim cánh cụt đội mũ màu ngọc lam là $1/3$. Làm thế nào tốt của một ước tính này là? Cho rằng chúng ta biết sự thật ở đây $2/5$ số chim cánh cụt có mũ màu ngọc lam trong quần thể - chúng ta có thể đặt câu hỏi liệu $1/3$ có phải là ước tính hợp lý hay không.

Câu trả lời cho câu hỏi đó phụ thuộc vào nhiều yếu tố sẽ được thảo luận trong phần tiếp theo.

6.5 Các yếu tố ảnh hưởng đến chất lượng suy luận

Các yếu tố chính ảnh hưởng đến chất lượng của suy luận mà bạn có thể đưa ra liên quan đến những vi phạm trong suy nghĩ của chúng tôi về quy trình lấy mẫu và mô hình cho tổng thể. Rõ ràng, nếu chúng ta không thể định nghĩa dân số một cách mạch lạc, thì bất kỳ "suy luận" nào mà chúng ta đưa ra đối với dân số cũng sẽ được định nghĩa mơ hồ tương tự.

Việc vi phạm hiểu biết của chúng tôi về cách thức hoạt động của quy trình lấy mẫu sẽ dẫn đến việc chúng tôi đã thu thập dữ liệu không đại diện cho dân số theo cách mà chúng ta nghĩ nó sẽ. Điều này sẽ ảnh hưởng đến suy luận của chúng ta trong đó suy luận mà chúng ta đưa ra sẽ không áp dụng cho toàn bộ dân số, mà cho một nhóm dân số được lựa chọn cụ thể. Cái này hiện tượng đôi khi được gọi là sai lệch lựa chọn bởi vì số lượng mà bạn ước tính bị sai lệch việc lựa chọn dân số mà bạn đã lấy mẫu.

Vi phạm mô hình mà chúng tôi đặt ra cho dân số có thể dẫn đến việc chúng tôi ước tính sai mối quan hệ giữa các đặc điểm của dân số hoặc đánh giá thấp tính không chắc chắn của các ước tính của chúng tôi. Ví dụ, nếu đúng là chim cánh cụt có thể ảnh hưởng đến màu mău mà những con chim cánh cụt khác đội, thì điều đó sẽ vi phạm giả định về độc lập giữa những chú chim cánh cụt. Điều này sẽ dẫn đến sự gia tăng trong sự không chắc chắn của bất kỳ ước tính nào mà chúng tôi thực hiện từ dữ liệu. TRONG chung, sự phụ thuộc giữa các đơn vị trong quần thể giảm "kích thước mẫu hiệu quả" của tập dữ liệu của bạn vì các đơn vị bạn quan sát không thực sự độc lập với nhau và làm không đại diện cho các bit thông tin độc lập.

Một lý do cuối cùng cho sự khác biệt giữa ước tính của chúng tôi từ dữ liệu và sự thật trong dân số là biến thiên lấy mẫu.

Bởi vì chúng tôi đã lấy mẫu ngẫu nhiên những con chim cánh cụt từ quần thể, nên có khả năng là nếu chúng tôi tiến hành lại thí nghiệm và lấy mẫu ba con chim cánh cụt khác, chúng tôi sẽ có được một ước tính khác nhau về số lượng chim cánh cụt với mău màu ngọc lam, đơn giản là do sự thay đổi ngẫu nhiên trong quá trình lấy mẫu. Điều này sẽ xảy ra ngay cả khi mô tả của chúng tôi về quá trình lấy mẫu là chính xác và mô hình của chúng tôi cho dân số là hoàn hảo.

Trong hầu hết các trường hợp, sự khác biệt giữa những gì chúng ta có thể ước tính với dữ liệu và sự thật trong dân số có thể là gì

được giải thích bởi sự kết hợp của cả ba yếu tố. Đôi khi, mỗi yếu tố đóng vai trò lớn như thế nào trong một vấn đề nhất định có thể khó xác định do thiếu thông tin, nhưng thường đáng để suy nghĩ về từng yếu tố trong số những yếu tố này và quyết định xem yếu tố nào có thể đóng vai trò chủ đạo. Bằng cách đó, người ta có thể khắc phục vấn đề, ví dụ, trong các nghiên cứu hoặc thí nghiệm trong tương lai.

6.6 Ví dụ: Sử dụng Apple Music

Vào ngày 18 tháng 8 năm 2015, công ty nghiên cứu thị trường người tiêu dùng MusicWatch đã công bố một nghiên cứu² về một dịch vụ âm nhạc mới do Apple, Inc. ra mắt có tên là Apple Music. Dịch vụ này là một dịch vụ phát nhạc trực tuyến mới được thiết kế để cung cấp cho người dùng quyền truy cập trực tuyến vào một danh mục nhạc lớn với giá 9,99 đô la mỗi tháng. Tuy nhiên, có một thời gian dùng thử miễn phí kéo dài trong 3 tháng. Vào thời điểm đó, có nhiều suy đoán về việc có bao nhiêu người dùng cuối cùng sẽ tiếp tục trả 9,99 đô la mỗi tháng sau khi thời gian dùng thử miễn phí kết thúc.

Nghiên cứu của MusicWatch đã tuyên bố, trong số những thử khác, rằng

Trong số những người đã dùng thử Apple Music, 48% cho biết họ hiện không sử dụng dịch vụ này.

Điều này có nghĩa là gần một nửa số người đã đăng ký thời gian dùng thử miễn phí của Apple Music không quan tâm đến việc sử dụng thêm và có khả năng sẽ không trả tiền sau khi thời gian dùng thử kết thúc. Nếu đó là sự thật, nó sẽ là một đòn giáng mạnh vào dịch vụ mới ra mắt.

²<http://www.businesswire.com/news/home/20150818005755/en#.VddbR7Scy6F>

Nhưng làm thế nào mà MusicWatch đến được con số của nó? Nó tuyên bố đã khảo sát 5.000 người trong nghiên cứu của mình. Ngay trước khi cuộc khảo sát của MusicWatch được công bố, Apple tuyên bố rằng khoảng 11 triệu người đã đăng ký dịch vụ Apple Music mới của họ (vì dịch vụ này mới ra mắt nên tất cả những người đăng ký đều đang trong thời gian dùng thử miễn phí). Rõ ràng, 5.000 người không chiếm toàn bộ dân số, vì vậy chúng tôi chỉ có một mẫu nhỏ người dùng.

Mục tiêu mà MusicWatch đang cố gắng trả lời là gì? Có vẻ như họ muốn biết tỷ lệ phần trăm tất cả những người đã đăng ký Apple Music vẫn đang sử dụng dịch vụ. Bởi vì sẽ rất tốn kém nếu khảo sát toàn bộ 11 triệu người, nên họ phải sử dụng một mẫu nhỏ hơn nhiều là 5.000 người. Họ có thể suy luận về toàn bộ dân số từ mẫu 5.000 người không?

Hãy xem xét ba thành phần để suy luận:

1. Dân số: Chúng tôi quan tâm đến hành vi của toàn bộ cơ sở người dùng Apple Music, theo Apple, khoảng 11 triệu người.
2. Quy trình lấy mẫu: Thông cáo báo chí không nêu rõ nghiên cứu được tiến hành như thế nào và dữ liệu được thu thập như thế nào. Có khả năng đây là một cuộc khảo sát qua điện thoại và vì vậy mọi người được chọn ngẫu nhiên để được gọi và hỏi về việc sử dụng dịch vụ của họ. Bạn có nghĩ rằng quá trình này dẫn đến một mẫu người trả lời đại diện cho toàn bộ người dùng Apple Music không?
3. Mẫu cho tổng thể: Với kích thước tương đối nhỏ của mẫu so với toàn bộ tổng thể, có khả năng các cá nhân trong cuộc khảo sát có thể được coi là độc lập với nhau. Nói cách khác, không có khả năng một người trả lời trong cuộc khảo sát có thể ảnh hưởng đến người trả lời khác.

Nếu mẫu là đại diện và các cá nhân độc lập, chúng ta có thể sử dụng con số 48% để ước tính tỷ lệ phần trăm dân số không còn sử dụng dịch vụ.

Thông cáo báo chí từ MusicWatch không chỉ ra bất kỳ biện pháp đo lường sự không chắc chắn nào, vì vậy chúng tôi không biết con số này đáng tin cậy đến mức nào.

Thật thú vị, ngay sau khi cuộc khảo sát MusicWatch được phát hành, Apple đã đưa ra một tuyên bố cho ấn phẩm The Verge, nói rằng 79% người dùng đã đăng ký vẫn đang sử dụng dịch vụ (tức là chỉ có 21% đã ngừng sử dụng, trái ngược với 48% % được báo cáo bởi MusicWatch). Giờ đây, sự khác biệt giữa Apple và MusicWatch là Apple có quyền truy cập dễ dàng vào toàn bộ người dùng Apple Music.

Nếu họ muốn biết bao nhiêu phần trăm dân số vẫn đang sử dụng nó, họ chỉ cần đếm số người dùng đang hoạt động của dịch vụ và chia cho tổng số người đã đăng ký. Không có gì chắc chắn về con số cụ thể đó vì không cần lấy mẫu để ước tính (tôi cho rằng Apple đã không sử dụng lấy mẫu để ước tính tỷ lệ phần trăm).

Nếu chúng tôi tin rằng Apple và MusicWatch đang đo lường cùng một thứ trong các phân tích của họ (và không rõ liệu họ có đúng như vậy không), thì điều đó có nghĩa là ước tính của MusicWatch về tỷ lệ phần trăm dân số (48%) khác xa so với giá trị thực (21%). Điều gì sẽ giải thích sự khác biệt lớn này?

1. Biến dị ngẫu nhiên. Đúng là cuộc khảo sát của MusicWatch là một mẫu nhỏ so với toàn bộ dân số, nhưng mẫu vẫn lớn với 5.000 người. Ngoài ra, phân tích khá đơn giản (chỉ lấy tỷ lệ người dùng vẫn đang sử dụng dịch vụ), do đó, độ không chắc chắn liên quan đến ước tính đó có thể không lớn.

2. Xu hướng lựa chọn. Hãy nhớ lại rằng không rõ Mu sicWatch đã lấy mẫu những người trả lời như thế nào nhưng có thể cách họ làm đã khiến họ thu hút được một nhóm người trả lời ít có khuynh hướng sử dụng Apple Music hơn. Ngoài điều này, chúng tôi thực sự không thể nói nhiều hơn nếu không biết chi tiết về quy trình khảo sát.
3. Sự khác biệt trong đo lường. Một điều chúng tôi không biết là MusicWatch hoặc Apple định nghĩa "vẫn đang sử dụng dịch vụ" như thế nào. Bạn có thể hình dung ra nhiều cách khác nhau để xác định xem một người có còn sử dụng dịch vụ hay không. Bạn có thể hỏi "Bạn đã sử dụng nó trong tuần trước chưa?" hoặc có lẽ "Bạn đã sử dụng nó ngày hôm qua?" Các câu trả lời cho những câu hỏi này sẽ khá khác nhau và có khả năng dẫn đến tỷ lệ phần trăm sử dụng tổng thể khác nhau.
4. Người trả lời không độc lập. Có thể những người trả lời khảo sát không độc lập với nhau. Điều này chủ yếu ảnh hưởng đến sự không chắc chắn về ước tính, làm cho nó lớn hơn chúng ta có thể mong đợi nếu những người trả lời đều độc lập. Tuy nhiên, vì chúng tôi không biết MusicWatch không chắc chắn về ước tính của họ ngay từ đầu nên rất khó để biết liệu sự phụ thuộc giữa những người trả lời có thể đóng một vai trò nào đó hay không.

6.7 Quần thể có nhiều dạng

Có rất nhiều chiến lược mà người ta có thể sử dụng để thiết lập một khuôn khổ chính thức cho việc đưa ra các tuyên bố suy luận. Thông thường, theo đúng nghĩa đen, có một quần thể gồm các đơn vị (ví dụ: người, chim cánh cụt, v.v.) mà bạn muốn đưa ra tuyên bố. Trong những trường hợp đó, rõ ràng sự không chắc chắn đến từ đâu (lấy mẫu từ dân số) và chính xác bạn đang cố gắng ước tính điều gì (một số tính năng của dân số).

Tuy nhiên, trong các ứng dụng khác, nó có thể không rõ ràng

chính xác là dân số và chính xác những gì bạn đang cố gắng ước tính. Trong những trường hợp đó, bạn sẽ phải xác định dân số rõ ràng hơn vì có thể có nhiều hơn một khả năng.

Chuỗi thời gian

Một số quy trình được đo theo thời gian (mỗi phút, mỗi ngày, v.v.). Ví dụ: chúng tôi có thể quan tâm đến dữ liệu phân tích bao gồm giá cổ phiếu đóng cửa hàng ngày của Apple cho năm dương lịch 2014. Nếu chúng tôi muốn suy luận từ tập dữ liệu này, dân số sẽ là bao nhiêu? Có một vài khả năng.

1. Chúng tôi có thể lập luận rằng năm 2014 được lấy mẫu ngẫu nhiên từ tổng thể của tất cả các năm có thể có dữ liệu, do đó những suy luận mà chúng tôi đưa ra áp dụng cho các năm khác của giá cổ phiếu.
2. Chúng ta có thể nói rằng cổ phiếu của Apple đại diện cho một mẫu từ toàn bộ thị trường chứng khoán, để chúng ta có thể suy luận về các cổ phiếu khác từ bộ dữ liệu này.

Bất kể bạn chọn gì, điều quan trọng là phải làm rõ dân số mà bạn đang đề cập đến trước khi bạn cố gắng suy luận từ dữ liệu.

quá trình tự nhiên

Các hiện tượng tự nhiên, chẳng hạn như động đất, hỏa hoạn, bão, các hiện tượng liên quan đến thời tiết và các sự kiện khác xảy ra trong tự nhiên, thường được ghi lại theo thời gian và không gian. Đôi với các phép đo thuận túy theo thời gian, chúng ta có thể xác định dân số giống như cách chúng ta đã xác định dân số ở trên với ví dụ về chuỗi thời gian. Tuy nhiên, chúng tôi có thể có dữ liệu

chỉ được đo trong không gian. Ví dụ: chúng ta có thể có bản đồ tâm chấn của tất cả các trận động đất đã xảy ra trong một khu vực. Vậy thì dân số là gì? Một cách tiếp cận phổ biến là nói rằng có một quá trình ngẫu nhiên không quan sát được gây ra các trận động đất ngẫu nhiên cho khu vực và dữ liệu của chúng tôi đại diện cho một mẫu ngẫu nhiên từ quá trình này. Trong trường hợp đó, chúng tôi đang sử dụng dữ liệu để cố gắng tìm hiểu thêm về quy trình không được quan sát này.

Dữ liệu dưới dạng dân số

Một kỹ thuật luôn khả thi, nhưng không được sử dụng phổ biến, là coi tập dữ liệu là một tổng thể. Trong trường hợp này, không có suy luận vì không có lấy mẫu. Bởi vì tập dữ liệu của bạn là dân số nên không có gì không chắc chắn về bất kỳ đặc điểm nào của dân số. Điều này nghe có vẻ không phải là một chiến lược hữu ích nhưng có những trường hợp có thể sử dụng nó để trả lời những câu hỏi quan trọng. Đặc biệt, có những lúc chúng tôi không quan tâm đến những thứ bên ngoài tập dữ liệu.

Ví dụ, các tổ chức thường phân tích dữ liệu tiền lương để đảm bảo rằng phụ nữ không bị trả lương thấp hơn nam giới cho công việc tương đương hoặc không có sự mất cân bằng lớn giữa nhân viên thuộc các nhóm dân tộc khác nhau. Trong cài đặt này, sự khác biệt về tiền lương giữa các nhóm khác nhau có thể được tính toán trong bộ dữ liệu và người ta có thể xem liệu sự khác biệt có đủ lớn để được quan tâm hay không. Vấn đề là dữ liệu trả lời trực tiếp một câu hỏi được quan tâm, đó là "Có sự khác biệt lớn về lương nào cần được giải quyết không?" Trong trường hợp này, không cần phải đưa ra suy luận về nhân viên bên ngoài tổ chức (theo định nghĩa là không có) hoặc nhân viên tại các tổ chức khác mà bạn không có bất kỳ quyền kiểm soát nào. Tập dữ liệu là dân số và câu trả lời cho bất kỳ câu hỏi nào liên quan đến dân số đều nằm trong tập dữ liệu đó.

7. Người mẫu chính thức

Chương này thường là một phần của sách giáo khoa hoặc khóa học thống kê mà mọi người có xu hướng đậm đầu vào tường. Đặc biệt, thường có rất nhiều toán học. Toán học thì tốt, nhưng toán vô cớ thì không tốt. Chúng tôi không ủng hộ điều đó.

Điều quan trọng là phải nhận ra rằng việc biểu diễn một mô hình bằng cách sử dụng ký hiệu toán học thường rất hữu ích vì đây là một ký hiệu nhỏ gọn và có thể dễ dàng diễn giải khi bạn đã quen với nó. Ngoài ra, viết ra một mô hình thống kê bằng cách sử dụng ký hiệu toán học, thay vì chỉ sử dụng ngôn ngữ tự nhiên, buộc bạn phải chính xác trong mô tả mô hình và trong tuyên bố của bạn về những gì bạn đang cố gắng đạt được, chẳng hạn như ước tính một tham số.

7.1 Mục tiêu của mô hình hóa chính thức là gì?

Một mục tiêu chính của mô hình chính thức là phát triển một đặc điểm kỹ thuật chính xác cho câu hỏi của bạn và cách dữ liệu của bạn có thể được sử dụng để trả lời câu hỏi đó. Các mô hình chính thức cho phép bạn xác định rõ ràng những gì bạn đang cố gắng suy luận từ dữ liệu và những gì hình thành mối quan hệ giữa các đặc điểm của tổng thể. Có thể khó đạt được mức độ chính xác này nếu chỉ sử dụng từ ngữ.

Các tham số đóng một vai trò quan trọng trong nhiều mô hình thống kê chính thức (trong ngôn ngữ thống kê, chúng được gọi là các mô hình thống kê tham số). Đây là những con số mà chúng tôi sử dụng để đại diện cho các tính năng hoặc liên kết tồn tại trong dân số.

Bởi vì chúng đại diện cho các tính năng dân số, các tham số thường được coi là ẩn số và mục tiêu của chúng tôi là ước tính chúng từ dữ liệu chúng tôi thu thập.

Ví dụ: giả sử chúng ta muốn đánh giá mối quan hệ giữa số ounce soda mà một người con trai tiêu thụ mỗi ngày và chỉ số BMI của người đó. Độ dốc của một đường mà bạn có thể vẽ biểu đồ trực quan hóa mối quan hệ này là thông số bạn muốn ước tính để trả lời câu hỏi của mình: "BMI sẽ tăng bao nhiêu khi tiêu thụ thêm mỗi ounce soda?" Cụ thể hơn, bạn đang sử dụng mô hình hồi quy tuyến tính để hình thành vấn đề này.

Một mục tiêu khác của mô hình hóa chính thức là phát triển một khuôn khổ nghiêm ngặt mà bạn có thể thách thức và kiểm tra các kết quả chính của mình. Tại thời điểm này trong phân tích dữ liệu của bạn, bạn đã nêu và tinh chỉnh câu hỏi của mình, bạn đã khám phá dữ liệu một cách trực quan và có thể tiến hành một số mô hình khám phá. Điều quan trọng là bạn có thể hiểu rõ câu trả lời cho câu hỏi của mình là gì, nhưng có thể có một số nghi ngờ về việc liệu những phát hiện của bạn có phù hợp với sự giám sát chặt chẽ hay không. Giả sử bạn vẫn quan tâm đến việc tiếp tục với kết quả của mình, đây là lúc mô hình hóa chính thức có thể đóng một vai trò quan trọng.

7.2 Khuôn khổ chung

Chúng ta có thể áp dụng chu kỳ phân tích cơ bản cho phần lập mô hình chính thức của phân tích dữ liệu. Chúng tôi vẫn muốn đặt kỳ vọng, thu thập thông tin và điều chỉnh kỳ vọng của mình dựa trên dữ liệu. Trong bài đặt này, ba giai đoạn này trông như sau.

1. Đặt kỳ vọng. Đặt kỳ vọng xuất hiện dưới hình thức phát triển một mô hình chính đại diện cho

cảm giác tốt nhất của bạn về những gì cung cấp câu trả lời cho câu hỏi của bạn. Mô hình này được chọn dựa trên bất kỳ thông tin nào bạn hiện có.

2. Thu thập thông tin. Sau khi mô hình chính được thiết lập, chúng tôi sẽ muốn tạo một tập hợp các mô hình phụ thách thức mô hình chính theo một cách nào đó. Chúng tôi sẽ thảo luận các ví dụ về ý nghĩa của điều này bên dưới.
3. Xem xét lại các kỳ vọng. Nếu các mô hình thứ cấp của chúng tôi thành công trong việc thách thức mô hình chính của chúng tôi và khiến các kết luận của mô hình chính bị nghi ngờ, thì chúng tôi có thể cần phải điều chỉnh hoặc sửa đổi mô hình chính để phản ánh tốt hơn những gì chúng tôi đã học được từ giây mô hình ondary.

Mô hình chính

Nó thường hữu ích để bắt đầu với một mô hình chính. Mô hình này có thể sẽ được bắt nguồn từ bất kỳ phân tích thăm dò nào mà bạn đã tiến hành và sẽ đóng vai trò là ứng cử viên chính cho điều gì đó tóm tắt ngắn gọn kết quả của bạn và phù hợp với mong đợi của bạn. Điều quan trọng là phải nhận ra rằng tại bất kỳ thời điểm nào trong quá trình phân tích dữ liệu, mô hình chính không nhất thiết phải là mô hình cuối cùng. Nó chỉ đơn giản là mô hình mà bạn sẽ so sánh với các mô hình thứ cấp khác. Quá trình so sánh mô hình của bạn với các mô hình thứ cấp khác thường được gọi là phân tích độ nhạy, bởi vì bạn quan tâm đến việc xem mức độ nhạy cảm của mô hình với các thay đổi, chẳng hạn như thêm hoặc xóa các yếu tố dự đoán hoặc loại bỏ các giá trị ngoại lai trong dữ liệu.

Qua quy trình lặp đi lặp lại của mô hình chính thức, bạn có thể quyết định rằng một mô hình khác phù hợp hơn để làm mô hình chính. Điều này hoàn toàn bình thường và là một phần của quá trình thiết lập các kỳ vọng, thu thập thông tin và tinh chỉnh các kỳ vọng dựa trên dữ liệu.

Mô hình phụ

Khi bạn đã quyết định chọn một mô hình chính, thì thông thường bạn sẽ phát triển một loạt các mô hình thứ cấp. Mục đích của các mô hình này là để kiểm tra tính hợp pháp và tính vững chắc của mô hình chính của bạn và có khả năng tạo ra bằng chứng chống lại mô hình chính của bạn. Nếu các mô hình thứ cấp thành công trong việc tạo ra bằng chứng bác bỏ các kết luận của mô hình chính của bạn, thì bạn có thể cần phải xem lại mô hình chính và xem kết luận của nó có còn hợp lý hay không.

7.3 Phân tích liên kết

Các phân tích liên kết là những phân tích mà chúng tôi đang xem xét mối liên hệ giữa hai hoặc nhiều tính năng với sự có mặt của các yếu tố gây nhiều tiềm ẩn khác. Có ba loại biến quan trọng cần xem xét trong một phân tích liên kết.

1. Kết quả. Kết quả là tính năng của tập dữ liệu của bạn được cho là sẽ thay đổi cùng với trình xác định chính của bạn. Ngay cả khi bạn không đặt câu hỏi về nguyên nhân hoặc cơ học, vì vậy bạn không nhất thiết phải tin rằng kết quả phản ứng với những thay đổi trong yếu tố dự đoán chính, thì kết quả vẫn cần được xác định một cách chính thức nhất.
các phương pháp mô hình hóa.
2. Dự đoán chính. Thông thường, đối với các phân tích liên kết, có một yếu tố dự báo chính được quan tâm (có thể có một vài trong số chúng). Chúng tôi muốn biết kết quả thay đổi như thế nào với yếu tố dự đoán chính này. Tuy nhiên, sự hiểu biết của chúng tôi về mối quan hệ đó có thể bị thách thức bởi sự hiện diện của các yếu tố gây nhiều tiềm ẩn.

3. Các yếu tố gây nhiễu tiềm ẩn. Đây là một lớp lớn các yếu tố dự đoán có liên quan đến yếu tố dự đoán chính và kết quả. Điều quan trọng là phải hiểu rõ chúng là gì và liệu chúng có sẵn trong tập dữ liệu của bạn hay không. Nếu một yếu tố gây nhiễu chính không có sẵn trong tập dữ liệu, đôi khi sẽ có một proxy có liên quan đến yếu tố gây nhiễu chính đó có thể được thay thế.

Khi bạn đã xác định được ba loại biến này trong tập dữ liệu của mình, bạn có thể bắt đầu nghĩ về việc lập mô hình chính thức trong môi trường liên kết.

Dạng cơ bản của một mô hình trong phân tích liên kết sẽ là

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

Ở đây

- y là kết quả • x
là yếu tố dự đoán chính
- z là yếu tố gây nhiễu tiềm
ẩn • ε là sai số ngẫu nhiên độc lập • α là phần chặn, nghĩa là giá trị y khi $x = 0$ và $z = 0$ • β là sự thay đổi của y liên quan đến tăng 1 đơn vị x , điều chỉnh theo z
- y là sự thay đổi của y liên quan đến việc tăng 1 đơn vị trong z , điều chỉnh theo x

Đây là một mô hình tuyến tính và mối quan tâm chính của chúng tôi là ước tính hệ số β , hệ số định lượng mối quan hệ giữa yếu tố dự đoán chính x và kết quả y .

Mặc dù chúng ta sẽ phải ước lượng α và γ như một phần của quá trình ước lượng β , nhưng chúng ta không thực sự quan tâm đến

giá trị của các α và γ đó. Trong tài liệu thống kê, các hệ số như α và γ đôi khi được gọi là các tham số phiền toái vì chúng ta phải sử dụng dữ liệu để ước tính chúng để hoàn thành đặc tả mô hình, nhưng chúng ta không thực sự quan tâm đến giá trị của chúng.

Mô hình hiển thị ở trên có thể được coi là mô hình chính. Có một yếu tố dự đoán chính và một yếu tố gây nhiễu trong mô hình mà có lẽ ai cũng biết rằng bạn chỉ nên quảng cáo cho yếu tố gây nhiễu đó. Mô hình này có thể tạo ra kết quả hợp lý và tuân theo những gì thường được biết đến trong khu vực.

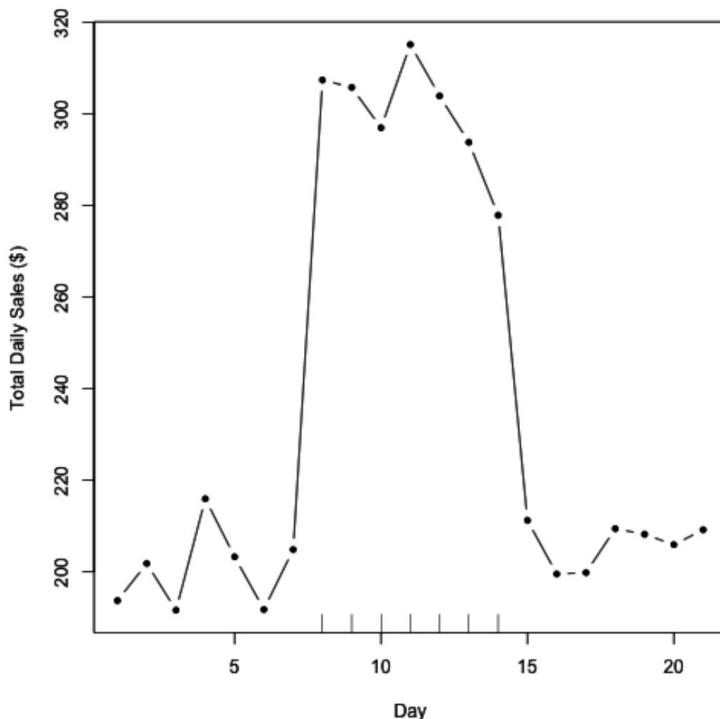
Ví dụ: Chiến dịch quảng cáo trực tuyến

Giả sử chúng tôi đang bán một sản phẩm mới trên web và chúng tôi quan tâm đến việc liệu việc mua quảng cáo trên Facebook có giúp tăng doanh số bán sản phẩm đó hay không. Để bắt đầu, chúng tôi có thể bắt đầu chiến dịch quảng cáo thử nghiệm kéo dài 1 tuần trên Facebook và đánh giá mức độ thành công của chiến dịch đó. Nếu thành công, chúng tôi có thể tiếp tục mua quảng cáo cho sản phẩm.

Một cách tiếp cận đơn giản có thể là theo dõi doanh số bán hàng hàng ngày trước, trong và sau chiến dịch quảng cáo (lưu ý rằng có nhiều cách chính xác hơn để thực hiện việc này với URL theo dõi và Google Analytics, nhưng bây giờ chúng ta hãy tạm gác việc đó sang một bên). Nói một cách đơn giản, nếu chiến dịch kéo dài một tuần, chúng ta có thể xem xét tuần trước, tuần trong và tuần sau để xem liệu có bất kỳ sự thay đổi nào trong doanh số hàng ngày hay không.

kỳ vọng

Trong một thế giới lý tưởng, dữ liệu có thể trông giống như thế này.



Chiến dịch quảng cáo giả định

Các dấu kiểm trên trục x cho biết khoảng thời gian chiến dịch hoạt động. Trong trường hợp này, khá rõ ràng ảnh hưởng của chiến dịch quảng cáo đối với doanh số bán hàng. Chỉ cần nhìn bằng mắt thường, bạn có thể biết rằng chiến dịch quảng cáo đã thêm khoảng 100 đô la mỗi ngày vào tổng doanh thu hàng ngày. Mô hình chính của bạn có thể trông giống như

$$y = \alpha + \beta x + \varepsilon$$

trong đó y là tổng doanh số bán hàng ngày và x là chỉ số cho biết liệu một ngày nhất định có rơi vào chiến dịch quảng cáo hay không. Các hypo-

dữ liệu lý thuyết cho cốt truyện trên có thể trông như sau.

ngày chiến dịch bán hàng	
1 193.7355	0 1
2 201.8364	0 2
3 191.6437	0 3
4 215.9528	0 4
5 203.2951	0 5
6 191.7953	0 6
7 204.8743	0 7
8 307.3832	1 8
9 305.7578	1 9
10 296.9461	1 10
11 315.1178	1 11
12 303.8984	1 12
13 293.7876	1 13
14 277.8530	1 14
15 211.2493	0 15
16 199.5507	0 16
17 199.8381	0 17
18 209.4384	0 18
19 208.2122	0 19
20 205.9390	0 20
21 209.1898	0 21

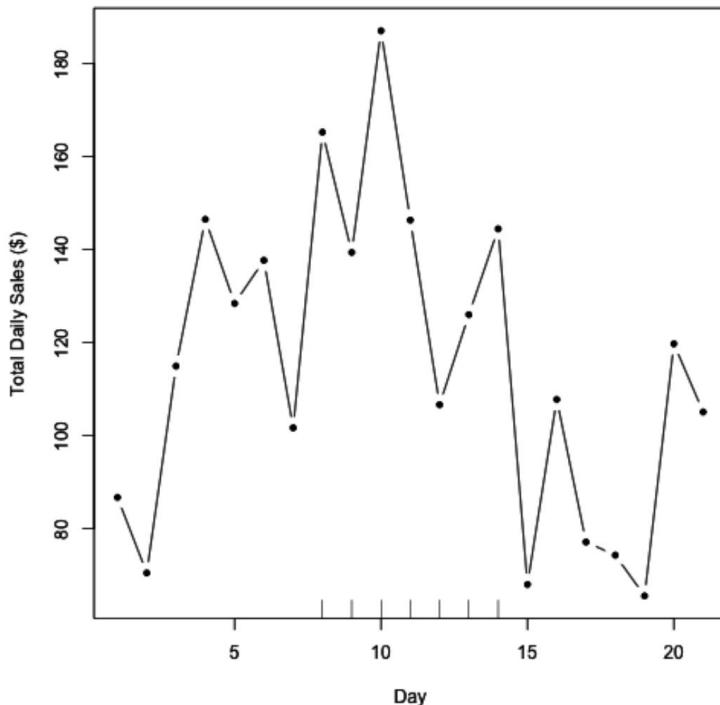
Với dữ liệu này và mô hình chính ở trên, chúng tôi ước tính là 96,78 đô la, không xa so với dự đoán ban đầu của chúng tôi của \$100.

Đặt kỳ vọng. Cuộc thảo luận về kịch bản lý tưởng này quan trọng không phải vì nó hoàn toàn có khả năng xảy ra, mà là bởi vì nó hướng dẫn về những gì chúng ta mong đợi để xem nếu thế giới hoạt động theo một khuôn khổ đơn giản hơn và làm thế nào chúng tôi sẽ phân tích dữ liệu theo những kỳ vọng đó.

Dữ liệu thực tế hơn

Thật không may, chúng tôi hiếm khi thấy dữ liệu như cốt truyện ở trên. TRONG thực tế, kích thước hiệu ứng có xu hướng nhỏ hơn, tiếng ồn có xu hướng

cao hơn, và có xu hướng có các yếu tố khác tác động. Thông thường, dữ liệu sẽ trông giống như thế này.



Dữ liệu bán hàng hàng ngày thực tế hơn

Mặc dù có vẻ như doanh số bán hàng tăng lên trong khoảng thời gian diễn ra chiến dịch quảng cáo (được biểu thị lại bằng dấu kiêm), nhưng hơi khó để tranh luận rằng doanh số bán hàng tăng lên là do chiến dịch. Thật vậy, trong những ngày trước khi chiến dịch bắt đầu, doanh số bán hàng dường như đã tăng nhẹ. Đó là tình cờ hay có những xu hướng khác đang diễn ra trong nền? Có thể có một xu hướng nền trơn tru để doanh số hàng ngày có xu hướng tăng và giảm trong suốt tháng. Do đó, ngay cả khi không có chiến dịch quảng cáo,

có thể chúng tôi đã thấy doanh số bán hàng tăng lên.

Câu hỏi bây giờ là liệu chiến dịch quảng cáo có tăng doanh số bán hàng hàng ngày dựa trên xu hướng cơ bản hiện có này hay không.

Hãy lấy mô hình chính của chúng ta, mô hình này chỉ bao gồm kết quả đầu ra và chỉ số của chiến dịch quảng cáo của chúng ta làm công cụ dự đoán chính. Sử dụng mô hình đó, chúng tôi ước tính β , mức tăng doanh thu hàng ngày do chiến dịch quảng cáo, là 44,75 đô la.

Tuy nhiên, giả sử chúng ta kết hợp một xu hướng nền vào mô hình của mình, vì vậy, thay vì mô hình chính, chúng ta phù hợp với những điều sau đây.

$$y = \alpha + \beta x + \gamma_1 t + \gamma_2 t^2 + \varepsilon$$

trong đó t bây giờ biểu thị số ngày (tức là 1, 2, . . . , 21).

Những gì chúng tôi đã làm là thêm một hàm bậc hai của t vào mô hình để cho phép một số độ cong trong xu hướng (trái ngược với một hàm tuyến tính chỉ cho phép một mẫu tăng hoặc giảm nghiêm ngặt). Sử dụng mô hình này, chúng tôi ước tính β là 39,86 đô la, thấp hơn một chút so với mô hình chính ước tính cho β .

Chúng tôi có thể khớp với một mô hình cuối cùng, cho phép xu hướng nền linh hoạt hơn nữa-chúng tôi sử dụng đa thức bậc 4 để biểu diễn xu hướng đó. Mặc dù chúng ta có thể thấy mô hình bậc hai của mình đủ phức tạp, nhưng mục đích của mô hình cuối cùng này là chỉ đẩy giới hạn ra ngoài một chút để xem mọi thứ thay đổi như thế nào trong những hoàn cảnh khắc nghiệt hơn. Mô hình này cho chúng tôi ước tính β là 49,1 đô la, trên thực tế lớn hơn ước tính từ mô hình chính của chúng tôi.

Tại thời điểm này, chúng tôi có một mô hình chính và hai mô hình phụ, đưa ra các ước tính hơi khác nhau về mối liên hệ giữa chiến dịch quảng cáo của chúng tôi và tổng doanh thu hàng ngày.

Người mẫu	Đặc trưng	Ước tính cho β
Mô hình 1 (chính)	Không có yếu tố gây nhiễu	\$44,75
Mô hình 2 (thứ cấp)	Xu hướng thời gian bậc	\$39,86
Mô hình 3 (thứ cấp)	hai xu hướng thời gian bậc 4	\$49,1

Sự đánh giá

Việc xác định nơi sẽ đi từ đây có thể phụ thuộc vào các yếu tố bên ngoài tập dữ liệu. Một số cân nhắc điển hình là

1. Kích thước hiệu ứng. Ba mô hình trình bày một loạt các thời gian ước tính từ \$39,86 đến \$49,1. Đây có phải là một phạm vi lớn? Có thể đối với tổ chức của bạn, phạm vi cường độ này không đủ lớn để thực sự tạo ra sự khác biệt và do đó, tất cả các mô hình có thể được coi là tương đương nhau. Hoặc bạn có thể coi 3 ước tính này khác biệt đáng kể với nhau, trong trường hợp đó, bạn có thể đặt nặng mô hình này hơn mô hình khác. Một yếu tố khác có thể là chi phí của chiến dịch quảng cáo, trong trường hợp đó bạn sẽ quan tâm đến lợi tức đầu tư của mình vào quảng cáo. Việc tăng 39,86 đô la mỗi ngày có thể đáng giá nếu tổng chi phí quảng cáo là 10 đô la mỗi ngày, nhưng có thể không phải nếu chi phí là 20 đô la mỗi ngày. Sau đó, bạn có thể cần mức tăng doanh số bán hàng cao hơn để làm cho chiến dịch có giá trị. Vấn đề ở đây là có một số bằng chứng từ mô hình chính thức của bạn rằng chiến dịch quảng cáo chỉ có thể tăng tổng doanh số hàng ngày của bạn lên 39,86, tuy nhiên, bằng chứng khác cho thấy nó có thể cao hơn. Câu hỏi đặt ra là liệu bạn có nghĩ rằng việc mua nhiều quảng cáo hơn có đáng để mạo hiểm hay không,

với nhiều khả năng, hoặc cho dù bạn nghĩ rằng ngay cả ở cấp cao hơn, nó có thể không xứng đáng.

2. **Tính hợp lý.** Mặc dù bạn có thể điều chỉnh một loạt mô hình nhằm mục đích thách thức mô hình chính của mình, nhưng có thể xảy ra trường hợp một số mô hình hợp lý hơn những mô hình khác, về mặt gần với bất kỳ "sự thật" nào về dân số. Ở đây, mô hình có xu hướng bậc hai có vẻ hợp lý vì nó có khả năng nắm bắt một mô hình tăng và giảm có thể có trong dữ liệu, nếu có. Mô hình với đa thức bậc 4 cũng có khả năng nắm bắt mẫu này tương tự, nhưng có vẻ quá phức tạp để mô tả một mẫu đơn giản như vậy. Việc một mô hình có thể được coi là hợp lý nhiều hay ít sẽ phụ thuộc vào kiến thức của bạn về chủ đề và khả năng của bạn trong việc ánh xạ các sự kiện trong thế giới thực vào công thức toán học của mô hình. Bạn có thể cần tham khảo ý kiến của các chuyên gia khác trong lĩnh vực này để đánh giá tính hợp lý của các mô hình khác nhau.

3. **Tiết kiệm.** Trong trường hợp các mô hình khác nhau đều kể cùng một câu chuyện (tức là các ước tính β đủ gần nhau để được coi là "giống nhau"), thì thường nên chọn mô hình đơn giản nhất.
Có hai lý do cho việc này. Đầu tiên, với một mô hình đơn giản hơn, có thể dễ dàng hơn để kể một câu chuyện về những gì đang diễn ra trong dữ liệu thông qua các tham số khác nhau trong mô hình. Ví dụ, việc giải thích một xu hướng tuy nhiên sẽ dễ dàng hơn là giải thích một xu hướng cấp số nhân. Thứ hai, các mô hình đơn giản hơn, từ góc độ thống kê, "hiệu quả" hơn, để chúng sử dụng tốt hơn dữ liệu trên mỗi tham số đang được ước tính. Độ phức tạp trong một mô hình thống kê thường đề cập đến số lượng tham số trong mô hình - trong ví dụ này, mô hình chính có 2 tham số, trong khi mô hình phức tạp nhất có 6 tham số. Nếu không có mô hình

tạo ra kết quả tốt hơn so với mô hình khác, chúng tôi có thể thích một mô hình chỉ chứa 2 tham số vì nó đơn giản hơn để mô tả và tiết kiệm hơn. Nếu các mô hình sơ cấp và thứ cấp tạo ra sự khác biệt đáng kể, thì có thể chọn một mô hình tiết kiệm một mô hình phức tạp hơn, nhưng không phải nếu mô hình phức tạp hơn kể một câu chuyện hấp dẫn hơn.

7.4 Phân tích dự đoán

Trong phần trước, chúng tôi đã mô tả các phân tích liên kết, trong đó mục tiêu là để xem liệu yếu tố dự báo chính x và kết quả y có liên quan với nhau hay không. Nhưng đôi khi mục tiêu là sử dụng tất cả thông tin bạn có để dự đoán y. Hơn nữa, sẽ không có vấn đề gì nếu các biến được coi là không có quan hệ nhân quả với kết quả mà bạn muốn dự đoán bởi vì mục tiêu là dự đoán, không phát triển sự hiểu biết về mối quan hệ giữa các tính năng.

Với các mô hình dự đoán, chúng tôi có các biến kết quả-các đặc điểm mà chúng tôi muốn đưa ra dự đoán-nhưng chúng tôi thường không phân biệt giữa “các yếu tố dự đoán chính” và các yếu tố dự đoán khác. Trong hầu hết các trường hợp, bất kỳ yếu tố dự đoán nào có thể được sử dụng để dự đoán kết quả sẽ được xem xét trong một phân tích và có thể, theo tiên nghiệm, được coi trọng như nhau về tầm quan trọng của nó trong việc dự đoán kết quả. Các phân tích dự đoán thường sẽ để thuật toán dự đoán xác định tầm quan trọng của từng yếu tố dự đoán và xác định dạng chức năng của mô hình.

Đối với nhiều phân tích dự đoán, không thể viết ra mô hình đang được sử dụng để dự đoán theo nghĩa đen vì nó không thể được biểu diễn bằng ký hiệu toán học tiêu chuẩn. Nhiều thói quen dự đoán hiện đại được cấu trúc như các thuật toán hoặc thủ tục lấy dữ liệu đầu vào và chuyển đổi

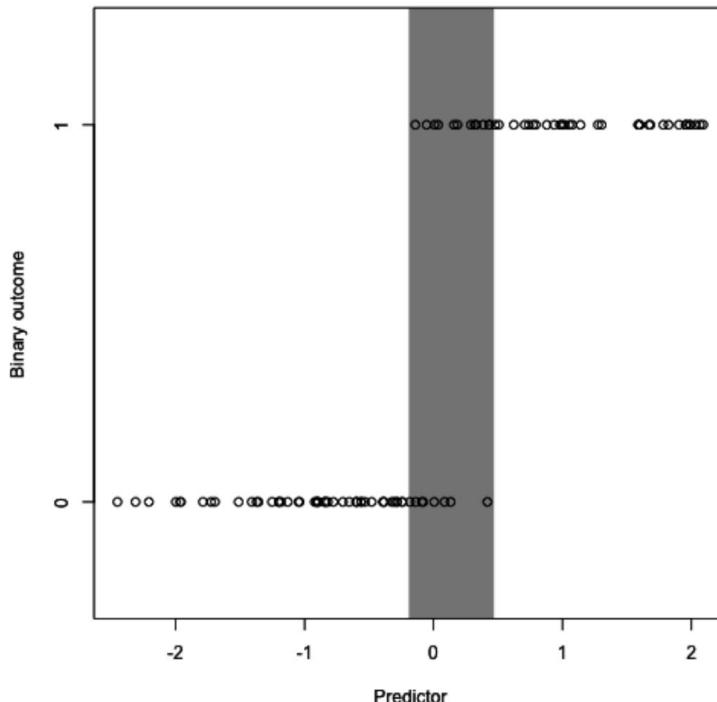
hình thành chúng thành đầu ra. Con đường mà đầu vào được chuyển đổi thành đầu ra có thể rất phi tuyến tính và các yếu tố dự đoán có thể tương tác với các yếu tố dự đoán khác trên đường đi. Thông thường, không có tham số quan tâm nào mà chúng tôi có gắng ước tính - trên thực tế, nhiều quy trình thuật toán hoàn toàn không có bất kỳ tham số có thể ước tính nào.

Điều quan trọng cần nhớ với các phân tích dự đoán là chúng ta thường không quan tâm đến các chi tiết cụ thể của mô hình. Trong hầu hết các trường hợp, miễn là phương pháp "hoạt động", có thể lặp lại và đưa ra dự đoán tốt với sai số tối thiểu, thì chúng ta đã đạt được mục tiêu của mình.

Với các phân tích dự đoán, loại phân tích chính xác mà bạn thực hiện phụ thuộc vào bản chất của kết quả (giống như với tất cả các phân tích). Các vấn đề dự đoán thường xuất hiện dưới dạng một vấn đề phân loại trong đó kết quả là nhị phân. Trong một số trường hợp, kết quả có thể có nhiều hơn hai cấp độ, nhưng trường hợp nhị phân cho đến nay là phổ biến nhất. Trong phần này, chúng ta sẽ tập trung vào vấn đề phân loại nhị phân.

kỳ vọng

Kịch bản lý tưởng trong một vấn đề dự đoán là gì? Nói chung, những gì chúng tôi muốn là một công cụ dự đoán hoặc một tập hợp các công cụ dự đoán để tạo ra sự phân tách tốt trong kết quả. Đây là một ví dụ về một yếu tố dự đoán duy nhất tạo ra sự phân tách hợp lý trong một kết quả nhị phân.



Kết quả nhận các giá trị 0 và 1, trong khi bộ dự đoán liên tục và nhận các giá trị trong khoảng từ -2 đến 2. Vùng màu xám được biểu thị trong biểu đồ làm nổi bật khu vực mà các giá trị của bộ dự đoán có thể nhận các giá trị 0 hoặc 1. Đối với bên phải của vùng màu xám, bạn sẽ nhận thấy rằng giá trị của kết quả luôn là 1 và bên trái của vùng màu xám, giá trị của kết quả luôn là 0. Trong các bài toán dự đoán, chính vùng màu xám này là nơi chúng ta không chắc chắn nhất về kết quả, đưa ra giá trị của yếu tố dự đoán.

Mục tiêu của hầu hết các vấn đề dự đoán là xác định một tập hợp các yếu tố dự đoán giúp giảm thiểu kích thước của vùng màu xám đó trong

cốt truyện trên. Ngược lại, người ta thường xác định các yếu tố dự đoán (đặc biệt là những yếu tố phân loại) có tỷ lệ kết quả hoàn toàn tách biệt, do đó vùng màu xám được giảm xuống bằng không.

Tuy nhiên, những tình huống như vậy thường chỉ ra một vấn đề suy biến không được quan tâm nhiều hoặc thậm chí là sai sót trong dữ liệu. Ví dụ, một biến liên tục đã được phân đôi sẽ được phân tách hoàn hảo bởi đối tác liên tục của nó. Một sai lầm phổ biến là đưa phiên bản liên tục làm yếu tố dự báo vào mô hình và phiên bản khác biệt làm kết quả. Trong dữ liệu thực tế, bạn có thể thấy sự tách biệt gần như hoàn hảo khi do lưỡng các tính năng hoặc đặc điểm được biết là được liên kết với nhau một cách máy móc hoặc thông qua một số quy trình xác định. Ví dụ: nếu kết quả là một chi báo về khả năng mắc ung thư buồng trứng của một người, thì giới tính của người đó có thể là một yếu tố dự báo rất tốt, nhưng có thể đó không phải là điều chúng ta quan tâm nhiều.

Dữ liệu thế giới thực

Đối với ví dụ này, chúng tôi sẽ sử dụng dữ liệu về giá trị tín dụng của các cá nhân. Bộ dữ liệu được lấy từ [UCI Machine Learning Repository](#)¹.

Bộ dữ liệu phân loại các cá nhân thành rủi ro tín dụng “Tốt” hoặc “Xấu” và bao gồm nhiều yếu tố dự đoán có thể dự đoán mức độ xứng đáng của tín dụng. Có tổng cộng 1000 quan sát trong bộ dữ liệu và 62 tính năng.

Vì mục đích của phần trình bày này, chúng tôi bỏ qua mã cho ví dụ này, nhưng các tệp mã có thể được lấy từ trang web của Sách.

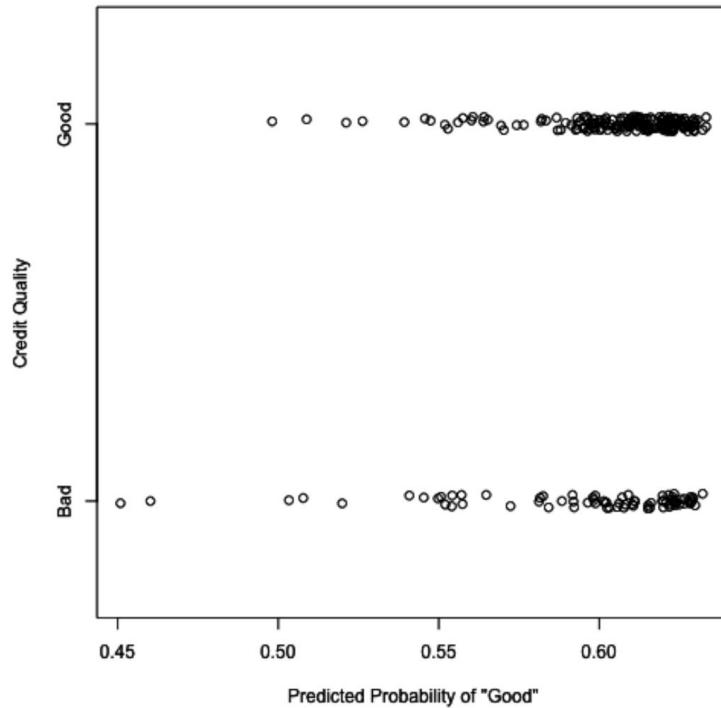
Điều đầu tiên chúng tôi làm cho một vấn đề dự đoán là chia dữ liệu thành tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm. Tập dữ liệu huấn luyện dùng để phát triển và điều chỉnh mô hình và tập dữ liệu thử nghiệm dùng để đánh giá mô hình đã phù hợp của chúng tôi và

¹[https://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))

ước tính tỷ lệ lỗi của nó. Trong ví dụ này, chúng tôi sử dụng ngẫu nhiên 75% các quan sát để phục vụ như tập dữ liệu huấn luyện. Các 25% còn lại sẽ đóng vai trò là tập dữ liệu thử nghiệm.

Sau khi khớp mô hình với tập dữ liệu huấn luyện, chúng ta có thể tính toán xác suất dự đoán là có "Tốt"

tín dụng từ tập dữ liệu thử nghiệm. Chúng tôi vẽ các xác suất dự đoán đó trên trục x cùng với tín dụng thực của từng cá nhân. trạng thái trên trục y bên dưới. (Tọa độ trực y có được xáo trộn ngẫu nhiên để hiển thị một số chi tiết hơn.)



Ở đây chúng ta có thể thấy rằng không có sự tách biệt hoàn toàn tốt

mà chúng ta đã thấy trong kịch bản lý tưởng. Trên phạm vi xác suất dự đoán, có những cá nhân có cả tín dụng "Tốt" và "Xấu". Điều này gợi ý rằng thuật toán dự đoán mà chúng tôi đã sử dụng có lẽ đang gặp khó khăn trong việc tìm kiếm sự kết hợp tốt các đặc điểm có thể phân biệt những người có rủi ro tín dụng tốt và xấu.

Chúng ta có thể tính toán một số thống kê tóm tắt về thuật toán dự đoán dưới đây.

Mã trộn nhằm lắn và thống kê

Thống kê

Dự đoán Xấu Tốt

Xấu	2	1
-----	---	---

Tốt	73	174
-----	----	-----

Độ chính xác : 0,704

Khoảng tin cậy 95% : (0,6432, 0,7599)

Tỷ lệ không có thông tin : 0,7

Giá trị P [Ac > NIR] : 0,4762

Kappa : 0,0289

Giá trị P thử nghiệm của McNemar : <2e-16

Độ nhạy: 0,99429

Độ đặc hiệu : 0,02667

Pos Giá trị dự đoán: 0,70445

Giá trị dự đoán âm: 0,66667

Tỷ lệ: 0,70000

Tỷ lệ phát hiện : 0,69600

Tỷ lệ phát hiện : 0,98800

Độ chính xác cân bằng: 0,51048

Lớp 'Tích cực' : Tốt

Chúng ta có thể thấy rằng độ chính xác là khoảng 70%, điều này không tốt đối với hầu hết các thuật toán dự đoán. Đặc biệt, tính đặc hiệu của thuật toán rất kém, nghĩa là nếu bạn là một

Rủi ro tín dụng “xấu”, xác suất bạn bị xếp vào loại như vậy chỉ khoảng 2,6%.

Sự đánh giá

Đối với các vấn đề về dự đoán, việc quyết định bước tiếp theo sau khi điều chỉnh mô hình ban đầu có thể phụ thuộc vào một số yếu tố.

1. Chất lượng dự đoán. Độ chính xác của mô hình có đủ tốt cho mục đích của bạn không? Điều này phụ thuộc vào mục tiêu cuối cùng và những rủi ro liên quan đến các hành động tiếp theo. Đối với các ứng dụng y tế, trong đó kết quả có thể là sự hiện diện của một căn bệnh, chúng tôi có thể muốn có độ nhạy cao để nếu bạn thực sự mắc bệnh, thuật toán sẽ phát hiện ra nó. Bằng cách đó chúng tôi có thể đưa bạn vào điều trị một cách nhanh chóng. Tuy nhiên, nếu phương pháp điều trị rất đau đớn, có thể có nhiều tác dụng phụ, thì chúng ta thực sự có thể thích tính đặc hiệu cao hơn, điều này sẽ đảm bảo rằng chúng ta không điều trị nhầm cho người không mắc bệnh. Đối với các ứng dụng tài chính, như ví dụ về giá trị tín dụng được sử dụng ở đây, có thể có chi phí bất đối xứng liên quan đến việc nhầm tín dụng tốt thành xấu so với nhầm tín dụng xấu thành tốt.
2. Điều chỉnh mô hình. Một dấu hiệu của các thuật toán dự đoán là nhiều tham số điều chỉnh của chúng. Đối khi các thông số tham số này có thể có ảnh hưởng lớn đến chất lượng dự đoán nếu chúng bị thay đổi và do đó, điều quan trọng là phải được thông báo về tác động của các tham số điều chỉnh đối với bất kỳ thuật toán nào bạn sử dụng. Không có thuật toán dự đoán nào mà một bộ tham số điều chỉnh duy nhất hoạt động tốt cho mọi vấn đề. Rất có thể, để phù hợp với mô hình ban đầu, bạn sẽ sử dụng các thông số “mặc định”, nhưng các thông số mặc định này có thể không đủ cho mục đích của bạn. Việc loyer hoay với các tham số điều chỉnh có thể làm thay đổi đáng kể chất lượng của

dự đoán của bạn. Điều rất quan trọng là bạn ghi lại các giá trị của các tham số điều chỉnh này để phân tích có thể được sao chép trong tương lai.

3. Sự sẵn có của Dữ liệu Khác. Nhiều thuật toán dự đoán khá tốt trong việc khám phá cấu trúc của tập dữ liệu lớn và phức tạp, đồng thời xác định cấu trúc có thể dự đoán tốt nhất kết quả của bạn. Nếu bạn tìm thấy rằng mô hình của bạn không hoạt động tốt, ngay cả sau một số điều chỉnh các thông số điều chỉnh, có khả năng là bạn cần dữ liệu bổ sung để cải thiện dự đoán của bạn.

7.5 Tóm tắt

Mô hình chính thức thường là khía cạnh kỹ thuật nhất của phân tích dữ liệu, và mục đích của nó là trình bày chính xác những gì là mục tiêu của phân tích và để cung cấp một khung làm việc chặt chẽ để thách thức những phát hiện của bạn và để kiểm tra những giả định của bạn. Cách tiếp cận mà bạn thực hiện có thể khác nhau tùy thuộc vào chủ yếu vào việc câu hỏi của bạn về cơ bản là về ước tính một hiệp hội phát triển một dự đoán tốt.

8. Suy luận so với Dự đoán:

Hàm ý cho việc lập mô hình Chiến lược

Hiểu được liệu bạn đang trả lời một câu hỏi suy luận hay một câu hỏi dự đoán là một khái niệm quan trọng vì loại câu hỏi bạn đang trả lời có thể ảnh hưởng lớn đến chiến lược lập mô hình mà bạn theo đuổi. Nếu bạn không hiểu rõ mình đang hỏi loại câu hỏi nào, bạn có thể sử dụng sai loại phương pháp lập mô hình và cuối cùng đưa ra kết luận sai từ dữ liệu của mình.

Mục đích của chương này là cho bạn thấy điều gì có thể xảy ra khi bạn nhầm lẫn câu hỏi này với câu hỏi khác.

Những điều quan trọng cần nhớ là

1. Đối với các câu hỏi suy luận, mục tiêu thường là ước tính mối liên hệ giữa yếu tố dự đoán quan tâm và kết quả. Thường chỉ có một số yếu tố dự báo đáng quan tâm (hoặc thậm chí chỉ một), tuy nhiên, thường có nhiều biến số gây nhiều tiềm ẩn cần xem xét. Mục tiêu chính của việc lập mô hình là ước tính mối liên hệ đồng thời đảm bảo bạn điều chỉnh thích hợp cho bất kỳ yếu tố gây nhiễu tiềm ẩn nào. Thông thường, các phân tích độ nhạy được tiến hành để xem liệu các mối liên hệ về lợi ích có mạnh mẽ đối với các nhóm yếu tố gây nhiễu khác nhau hay không.
2. Đối với các câu hỏi dự đoán, mục tiêu là xác định một mô hình dự đoán kết quả tốt nhất. Thông thường, chúng tôi không đặt bất kỳ tầm quan trọng ưu tiên nào lên các yếu tố dự đoán, miễn là chúng giỏi dự đoán kết quả.

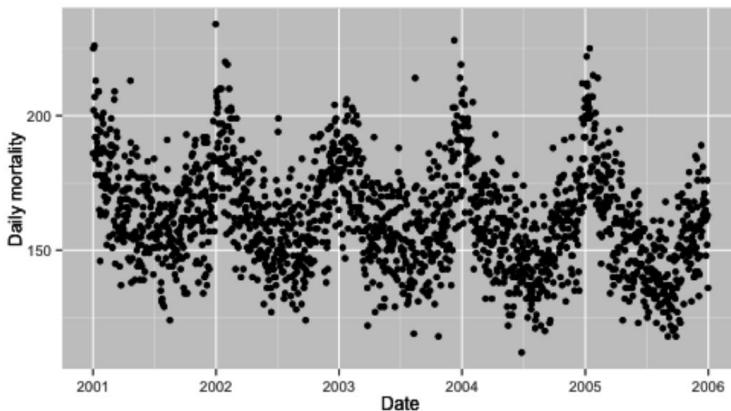
Không có khái niệm về “người gây nhiễu” hay “người dự đoán quan tâm” bởi vì tất cả những người dự đoán đều có khả năng hữu ích để dự đoán kết quả. Ngoài ra, chúng ta thường không quan tâm đến “cách thức hoạt động của mô hình” hay kể một câu chuyện chi tiết về các yếu tố dự đoán. Mục tiêu chính là phát triển một mô hình có kỹ năng dự đoán tốt và ước tính tỷ lệ lỗi hợp lý từ dữ liệu.

8.1 Ô nhiễm không khí và tử vong ở thành phố New York

Ví dụ sau đây cho thấy các loại câu hỏi khác nhau và cách tiếp cận mô hình hóa tương ứng có thể dẫn đến các kết luận khác nhau như thế nào. Ví dụ này sử dụng dữ liệu về ô nhiễm không khí và tử vong cho Thành phố New York. Dữ liệu ban đầu được sử dụng như một phần của Nghiên cứu [về bệnh tật, tử vong và ô nhiễm không khí quốc gia1](#) (NMMAP).

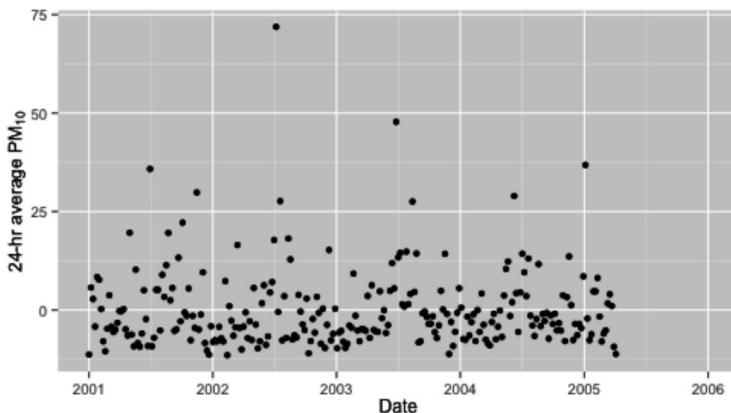
Dưới đây là biểu đồ về tỷ lệ tử vong hàng ngày do mọi nguyên nhân trong những năm 2001-2005.

¹<http://www.ihapss.jhsph.edu>



Tỷ lệ tử vong hàng ngày ở thành phố New York, 2001-2005

Và đây là đồ thị về mức trung bình trong 24 giờ của các hạt vật chất có đường kính khí động học nhỏ hơn hoặc bằng 10 micron (PM10).



PM10 hàng ngày tại thành phố New York, 2001-2005

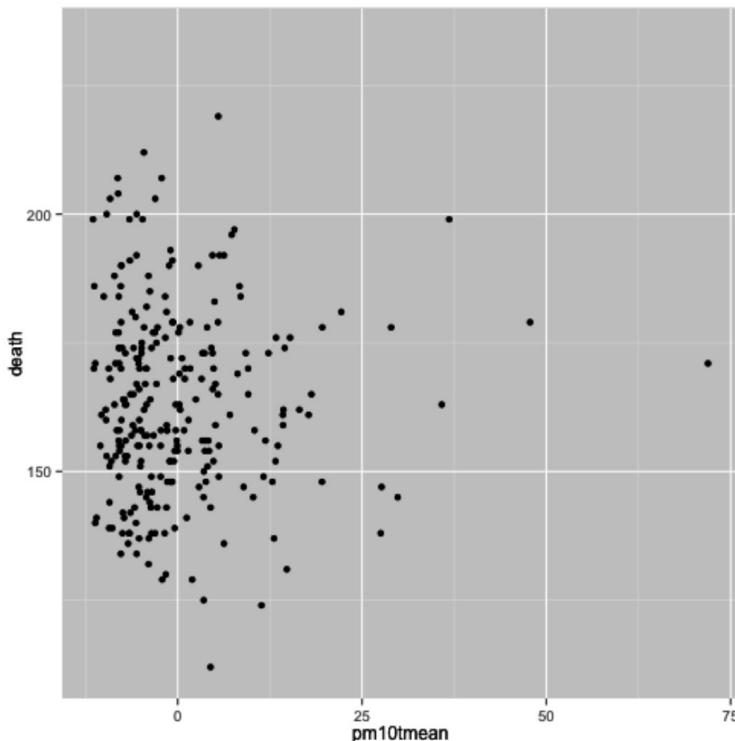
Lưu ý rằng có ít điểm hơn trên biểu đồ trên so với trên biểu đồ dữ liệu về tỷ lệ tử vong. Điều này là do PM10 không được đo hàng ngày. Cũng lưu ý rằng

có các giá trị âm trong biểu đồ PM10-điều này là do dữ liệu PM10 đã được trừ trung bình. Nói chung, giá trị âm của PM10 là không thể.

8.2 Suy ra một Hiệp hội

Cách tiếp cận đầu tiên mà chúng tôi sẽ thực hiện sẽ là hỏi “Có mối liên hệ nào giữa mức PM10 trung bình hàng ngày trong 24 giờ và tỷ lệ tử vong hàng ngày không?” Đây là một câu hỏi suy luận và chúng tôi đang cố gắng ước tính một hiệp hội. Ngoài ra, đối với câu hỏi này, chúng tôi biết có một số yếu tố gây nhiều tiềm ẩn mà chúng tôi sẽ phải giải quyết.

Chúng ta hãy xem mối liên hệ hai chiều giữa PM10 và tỷ lệ tử vong. Đây là một biểu đồ phân tán của hai biến.



PM10 và tỷ lệ tử vong ở thành phố New York

Có vẻ như không có nhiều thứ đang diễn ra ở đó, và một mô hình hồi quy tuyến tính đơn giản của nhật ký tử vong hàng ngày và PM10 dường như xác nhận điều đó.

Ước tính Std. Lỗi (Đang chặn) 5.0884308354 0.0069353779 733.75138151 0.0000000 pm10tmean 0,00004033446 0,0006913941 0,05833786 0,9535247	giá trị t Pr(> t) 5.0884308354 0.0069353779 733.75138151 0.0000000 0,00004033446 0,0006913941 0,05833786 0,9535247
--	---

Trong bảng hệ số trên, hệ số cho pm10tmean khá nhỏ và sai số chuẩn của nó tương đối lớn. Về mặt hiệu quả, ước tính này của hiệp hội bằng không.

Tuy nhiên, chúng tôi biết khá nhiều về cả PM10 và tỷ lệ tử vong hàng ngày, và một điều chúng tôi biết là mùa đóng một vai trò lớn trong cả hai biến số. Đặc biệt, chúng ta biết rằng tỷ lệ tử vong có xu hướng cao hơn vào mùa đông và thấp hơn vào mùa hè. PM10 có xu hướng thể hiện mô hình ngược lại, cao hơn vào mùa hè và thấp hơn vào mùa đông. Bởi vì mùa có liên quan đến cả PM10 và tỷ lệ tử vong, nó là một ứng cử viên sáng giá cho yếu tố gây nhiễu và sẽ hợp lý nếu điều chỉnh nó trong mô hình.

Đây là kết quả cho mô hình thứ hai, bao gồm cả PM10 và mùa. Mùa được bao gồm như một biến chỉ báo với 4 cấp độ.

	Ước tính Std. Lỗi	giá trị t	Pr(> t)
(Đang chặn)	5.166484285 0.0112629532 458.714886 0.000000e+00		
mùaQ2	-0,109271301 0,0166902948 -6,546996 3,209291e-10		
mùa quý 3	-0,155503242 0,0169729148 -9,161847 1,736346e-17		
mùaQ4	-0,060317619 0,0167189714 -3,607735 3,716291e-04		
pm10tmean	0,001499111 0,0006156902 2,434847 1,558453e-02		

Bây giờ hãy lưu ý rằng hệ số pm10tmean lớn hơn một chút so với trước đây và giá trị t của nó cũng lớn, cho thấy mối liên hệ chặt chẽ. Sao có thể như thế được?

Hóa ra chúng ta có một ví dụ kinh điển về [Nghịch lý Simpson](#)² đây. Mỗi quan hệ tổng thể giữa P10 và tỷ lệ tử vong là không có giá trị, nhưng khi chúng tôi tính đến sự thay đổi theo mùa về cả tỷ lệ tử vong và PM10, thì mối liên hệ này là tích cực. Kết quả đáng ngạc nhiên đến từ những cách ngược lại trong đó mùa có liên quan đến tỷ lệ tử vong và PM10.

Cho đến nay chúng tôi đã tính đến mùa, nhưng có những yếu tố gây nhiễu tiềm tàng khác. Đặc biệt, các biến thời tiết, chẳng hạn như nhiệt độ và nhiệt độ điểm sương, cũng liên quan đến sự hình thành và tỷ lệ chết của PM10.

²https://en.wikipedia.org/wiki/Simpson%27s_paradox

Trong mô hình sau đây, chúng tôi bao gồm nhiệt độ (tmpd) và nhiệt độ điểm sương (dptp). Chúng tôi cũng bao gồm ngày biến trong trường hợp có bất kỳ xu hướng dài hạn nào cần được hạch toán.

	Ước lượng	Tiêu chuẩn Lỗi	giá trị t	$\Pr(t)$
(Đang chặn)	5.62066568788	0.16471183741	34.1242365	1.851690e-96
ngày	-0.00002984198	0,00001315212	-2,2689856	2,411521e-02
mùa quý 2	-0,05805970053	0,02299356287	-2,5250415	1,218288e-02
mùa quý 3	-0,07655519887	0,02004104658	-2,6361033	8,006912e-03
mùa quý 4	-0,03154694305	0,01832712585	-1,7213252	8,041910e-02
tmpd	-0,00295931276	0,00128835065	-2,2969777	2,244054e-02
dptp	0,00068342228	0,00103489541	0,6603781	5,096144e-01
pm10tmean	0,00237049992	0,00065856022	3,5995189	3,837886e-04

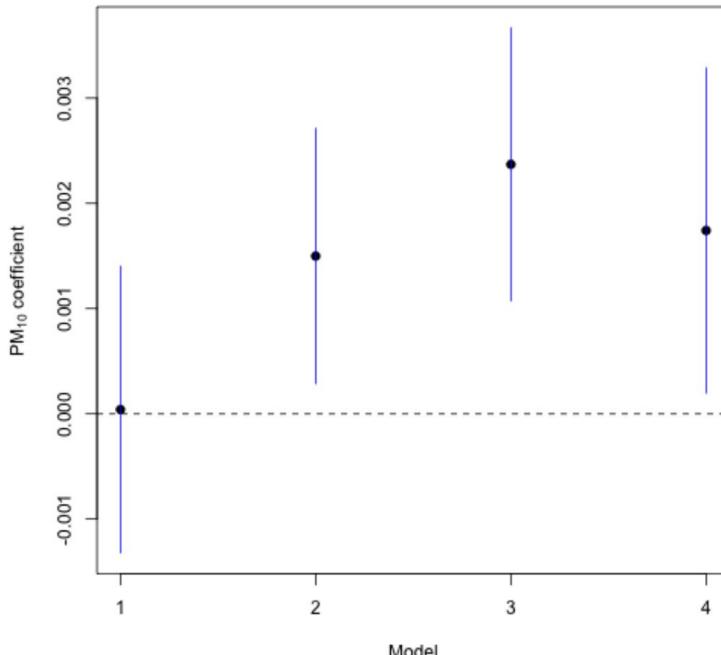
Lưu ý rằng hệ số pm10tmean thậm chí còn lớn hơn đó là trong mô hình trước đó. dường như vẫn còn một mối liên quan giữa PM10 và tỷ lệ tử vong. Kích thước hiệu ứng là nhỏ, nhưng chúng ta sẽ thảo luận về điều đó sau.

Cuối cùng, một loại yếu tố gây nhiều tiềm năng khác bao gồm các yếu tố khác ô nhiễm. Trước khi chúng ta đồ lỗi cho PM10 là một chất có hại gây ô nhiễm, điều quan trọng là chúng ta kiểm tra xem có có thể là một chất gây ô nhiễm khác có thể giải thích những gì chúng ta quan sát. NO2 là một ứng cử viên sáng giá vì nó chia sẻ một số nguồn với PM10 và được biết là có liên quan đến tử vong. Hãy xem điều gì sẽ xảy ra khi chúng ta đưa nó vào người mẫu.

	Ước lượng	Tiêu chuẩn Lỗi	giá trị t	$Pr(t)$
(Đánh chặn)	5.61378604085	0.16440280471	34.1465345	2.548704e-96
ngày	-0,00002973484	0,00001312231	-2,2659756	2,430503e-02
mùaQ2	-0,05143935218	0,02338034983	-2,2001105	2,871069e-02
mùa quý 3	-0,06569205605	0,02990520457	-2,1966764	2,895825e-02
mùaQ4	-0,02750381423	0,01849165119	-1,4873639	1,381739e-01
tmpd	-0,00296833498	0,00128542535	-2,3092239	2,174371e-02
dptp	0,00070306996	0,00103262057	0,6808599	4,965877e-01
no2tmean	0,00126556418	0,00086229169	1,4676753	1,434444e-01
pm10tmean	0,00174189857	0,00078432327	2,2208937	2,725117e-02

Lưu ý trong bảng các hệ số rằng hệ số trung bình no2t có độ lớn tương tự như hệ số pm10tmean , mặc dù giá trị t của nó không lớn bằng. Hệ số pm10tmean thường như có ý nghĩa thống kê, nhưng nó hơi độ lớn nhỏ hơn bấy giờ.

Dưới đây là đồ thị của hệ số PM10 từ cả bốn các mô hình mà chúng tôi đã thử.



Mối liên hệ giữa PM10 và tỷ lệ tử vong theo các mô hình khác nhau

Ngoài trừ Mô hình 1, không tính đến bất kỳ yếu tố gây nhiễu tiềm ẩn nào, dường như có mối liên hệ tích cực giữa PM10 và tỷ lệ tử vong trong các Mô hình 2-4. Điều này có nghĩa là gì và chúng ta nên làm gì với nó tùy thuộc vào mục tiêu cuối cùng của chúng ta là gì và chúng tôi không thảo luận chi tiết ở đây. Điều đáng chú ý là kích thước hiệu ứng nói chung là nhỏ, đặc biệt là so với một số yếu tố dự đoán khác trong mô hình. Tuy nhiên, cũng cần lưu ý rằng có lẽ mọi người ở Thành phố New York đều thở, và do đó, một tác động nhỏ có thể có tác động lớn.

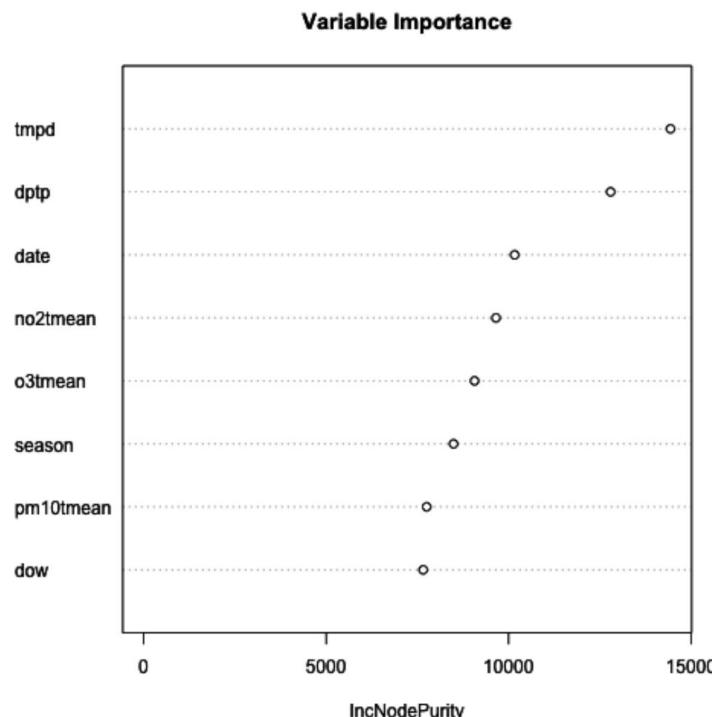
8.3 Dự Đoán Kết Quả

Một chiến lược khác mà chúng tôi có thể thực hiện là hỏi "Điều gì dự đoán tốt nhất về tỷ lệ tử vong ở Thành phố New York?" Đây rõ ràng là một câu hỏi dự đoán và chúng ta có thể sử dụng dữ liệu có sẵn để xây dựng một mô hình. Ở đây, chúng ta sẽ sử dụng **rừng ngẫu nhiên**³ chiến lược mô hình hóa, là một phương pháp học máy hoạt động tốt khi có một số lượng lớn các yếu tố dự đoán.

Một loại đầu ra mà chúng ta có thể thu được từ thủ tục ngẫu nhiên cho est là thước đo mức độ quan trọng của biến. Nói một cách đại khái, phép đo này chỉ ra tầm quan trọng của một biến nhất định đối với việc cải thiện kỹ năng dự đoán của mô hình.

Dưới đây là một biểu đồ có tầm quan trọng thay đổi, thu được sau khi điều chỉnh một mô hình rừng ngẫu nhiên. Các giá trị lớn hơn trên trục x biểu thị tầm quan trọng lớn hơn.

³https://en.wikipedia.org/wiki/Random_forest



Biểu đồ tầm quan trọng của biến rừng ngẫu nhiên để dự đoán tỷ lệ tử vong

Lưu ý rằng biến pm10tmean nằm gần cuối danh sách về mức độ quan trọng. Đó là bởi vì nó không đóng góp nhiều vào việc dự đoán kết quả, tỷ lệ tử vong.

Nhớ lại trong phần trước rằng quy mô ảnh hưởng có vẻ nhỏ, nghĩa là nó không thực sự giải thích được nhiều khả năng thay đổi về tỷ lệ tử vong. Các công cụ dự đoán như nhiệt độ và nhiệt độ điểm sương hữu ích hơn khi là công cụ dự đoán tỷ lệ tử vong hàng ngày. Ngay cả NO2 cũng là một chất dự đoán tốt hơn PM10.

Tuy nhiên, chỉ vì PM10 không phải là yếu tố dự báo chắc chắn về tỷ lệ tử vong không có nghĩa là nó không có mối liên hệ xác đáng với tỷ lệ tử vong. Với sự đánh đổi có

được thực hiện khi phát triển mô hình dự đoán, PM10 không cao trong danh sách các yếu tố dự đoán mà chúng tôi sẽ đưa vào-chúng tôi đơn giản là không thể bao gồm mọi yếu tố dự đoán.

8.4 Tóm tắt

Trong bất kỳ phân tích dữ liệu nào, bạn muốn tự hỏi mình "Tôi đang hỏi một câu hỏi suy luận hay một câu hỏi dự đoán?" Cái này nên được xóa sạch trước khi bắt kỳ dữ liệu nào được phân tích, vì câu trả lời cho câu hỏi có thể hướng dẫn toàn bộ chiến lược lập mô hình. Trong ví dụ ở đây, nếu chúng ta quyết định dự đoán cách tiếp cận, chúng ta có thể đã lầm tưởng rằng PM10 không liên quan đến tỷ lệ tử vong. Tuy nhiên, phương pháp áp suy luận cho thấy mối liên hệ có ý nghĩa thống kê với tử vong. Đặt câu hỏi đúng và áp dụng chiến lược mô hình thích hợp, có thể đóng một vai trò lớn trong các loại kết luận bạn rút ra từ dữ liệu.

9. Giải thích kết quả của bạn

Mặc dù chúng tôi đã dành cả một chương để diễn giải kết quả phân tích dữ liệu, nhưng việc diễn giải thực sự diễn ra liên tục trong suốt quá trình phân tích. Các nhà phân tích dữ liệu có kinh nghiệm thậm chí có thể không nhận thức được tần suất họ diễn giải các phát hiện của mình vì nó đã trở thành bản chất thứ hai đối với họ.

Đến giờ, quy trình 3 bước theo chu kỳ gồm: thiết lập kỳ vọng, thu thập thông tin (dữ liệu), sau đó đổi sánh kỳ vọng với dữ liệu, hẳn đã rất quen thuộc với bạn, vì vậy bạn sẽ nhận ra rằng bước thứ ba, đổi sánh kỳ vọng với dữ liệu, là diễn giải của chính nó. Theo một cách nào đó, chúng tôi đã đề cập đến chủ đề diễn giải các kết quả xuyên suốt cuốn sách. Tuy nhiên, nó xứng đáng có chương riêng vì có nhiều điều cần giải thích hơn là kết hợp kỳ vọng với kết quả và bởi vì bản thân nó là một bước quan trọng của phân tích dữ liệu. Bởi vì diễn giải xảy ra một cách tự do nhất sau khi hoàn thành các phân tích cơ bản và hỗ trợ của bạn, bao gồm cả **mô hình hóa chính thức**, nhưng trước khi **truyền đạt** kết quả, chúng tôi đã đặt chương này ở giữa các chương tương ứng này.

9.1 Nguyên tắc diễn giải

Có một số nguyên tắc diễn giải kết quả mà chúng tôi sẽ minh họa trong chương này. Những nguyên tắc này là:

1. Xem lại câu hỏi ban đầu của bạn

2. Bắt đầu với mô hình thống kê sơ cấp để xác định phương hướng và tập trung vào bản chất của kết quả thay vì đánh giá nhị phân kết quả (ví dụ: có ý nghĩa thống kê hay không). Bản chất của kết quả bao gồm ba đặc điểm: tính định hướng, độ lớn và tính không chắc chắn của nó. Sự không chắc chắn là một đánh giá về khả năng kết quả đạt được một cách tình cờ.
3. Phát triển một diễn giải tổng thể dựa trên (a) toàn bộ phân tích của bạn và (b) bối cảnh của những gì đã biết về chủ đề này.
4. Xem xét các hàm ý sẽ hướng dẫn bạn xác định (những) hành động nào, nếu có, nên được thực hiện do câu trả lời cho câu hỏi của bạn.

Điều quan trọng cần lưu ý là chu kỳ phân tích cũng áp dụng cho diễn giải. Ở mỗi bước diễn giải, bạn nên có kỳ vọng trước khi thực hiện bước đó, sau đó xem kết quả của bước đó có phù hợp với mong đợi của bạn hay không. Kỳ vọng của bạn dựa trên những gì bạn học được trong quá trình phân tích dữ liệu khám phá và lập mô hình chính thức, và khi diễn giải của bạn không phù hợp với kỳ vọng của bạn, thì bạn sẽ cần xác định xem chúng không phù hợp vì kỳ vọng của bạn không chính xác hay do bạn giải thích là không chính xác. Mặc dù bạn có thể đang ở một trong những bước cuối cùng của phân tích dữ liệu khi bạn chính thức diễn giải kết quả của mình, nhưng bạn có thể cần quay lại phân tích dữ liệu khám phá hoặc lập mô hình để khớp kỳ vọng với dữ liệu.

9.2 Nghiên cứu diễn hình: Mức tiêu

thụ soda không ăn kiêng và Chỉ số khói cơ thể

Có lẽ dễ dàng nhất là nhìn thấy các nguyên tắc diễn giải trong thực tế để học cách áp dụng chúng cho chính bạn.

phân tích dữ liệu, vì vậy chúng tôi sẽ sử dụng một nghiên cứu điển hình để minh họa từng nguyên tắc.

Xem lại câu hỏi

Nguyên tắc đầu tiên là nhắc nhở bản thân về câu hỏi ban đầu của bạn. Điều này có vẻ giống như một tuyên bố thiếu sót, nhưng không có gì lạ khi mọi người đi lạc hướng khi họ trải qua quá trình phân tích thăm dò và lập mô hình chính thức. Điều này thường xảy ra khi một nhà phân tích dữ liệu đi quá xa khỏi lộ trình theo đuổi một phát hiện tình cờ xuất hiện trong quá trình phân tích dữ liệu khám phá hoặc lập mô hình chính thức. Sau đó, (các) mô hình cuối cùng cung cấp câu trả lời cho một câu hỏi khác xuất hiện trong quá trình phân tích thay vì câu hỏi ban đầu.

Nhắc nhở bản thân về câu hỏi của bạn cũng giúp cung cấp một khuôn khổ cho việc giải thích của bạn. Ví dụ: câu hỏi ban đầu của bạn có thể là "Cứ uống một lon nước ngọt 12 ounce mỗi ngày, chỉ số BMI trung bình của người trưởng thành ở Hoa Kỳ lớn hơn bao nhiêu?". Cách diễn đạt câu hỏi cho bạn biết rằng ý định ban đầu của bạn là xác định xem chỉ số BMI của những người trưởng thành ở Hoa Kỳ uống trung bình hai lon nước ngọt 12 ounce mỗi ngày trung bình cao hơn bao nhiêu so với những người trưởng thành chỉ uống một lon 12 ounce. ounce soda trung bình mỗi ngày. Việc giải thích các phân tích của bạn sẽ mang lại một tuyên bố như: Cứ thêm 1 lon nước ngọt 12 ounce mà người lớn ở Hoa Kỳ uống, BMI tăng trung bình $X \text{ kg/m}^2$. Nhưng nó không nên đưa ra một tuyên bố như: "Cứ mỗi ounce soda bổ sung mà người lớn ở Hoa Kỳ uống, BMI tăng trung bình $X \text{ kg/m}^2$."

Một cách khác để xem lại câu hỏi của bạn cung cấp một khuôn khổ để diễn giải kết quả của bạn là việc nhắc nhở bản thân về loại câu hỏi mà bạn đã hỏi cung cấp

một khuôn khổ rõ ràng để giải thích (Xem [Nêu rõ và Tinh chỉnh Câu hỏi để xem xét các loại câu hỏi](#)).

Ví dụ: nếu câu hỏi của bạn là: "Trong số những người trưởng thành ở Hoa Kỳ, những người uống thêm 1 khẩu phần 12 ounce nước ngọt không ăn kiêng mỗi ngày có chỉ số BMI trung bình cao hơn không?", điều này cho bạn biết rằng câu hỏi của bạn là một câu hỏi suy luận câu hỏi và rằng mục tiêu của bạn là hiểu tác động trung bình của việc uống thêm 12 ounce nước ngọt không ăn kiêng mỗi ngày đối với chỉ số BMI của dân số trưởng thành ở Hoa Kỳ. Để trả lời câu hỏi này, bạn có thể đã thực hiện phân tích bằng cách sử dụng dữ liệu cắt ngang được thu thập trên một mẫu đại diện cho dân số trưởng thành ở Hoa Kỳ và trong trường hợp này, cách diễn giải kết quả của bạn được đóng khung theo mối liên hệ giữa một yếu tố bổ sung. Khẩu phần 12 ounce soda mỗi ngày và chỉ số BMI, tính trung bình trong dân số trưởng thành ở Hoa Kỳ.

Bởi vì câu hỏi của bạn không phải là câu hỏi nguyên nhân và do đó phân tích của bạn không phải là phân tích nguyên nhân, kết quả không thể được đóng khung theo điều gì sẽ xảy ra nếu dân số bắt đầu tiêu thụ thêm một lon nước ngọt mỗi ngày. Một câu hỏi nhân quả có thể là: "Uống thêm 12 ounce nước ngọt không ăn kiêng mỗi ngày có ảnh hưởng gì đến chỉ số BMI?", và để trả lời câu hỏi này, bạn có thể phân tích dữ liệu từ một thử nghiệm lâm sàng chỉ định ngẫu nhiên một nhóm uống thêm một lon soda và nhóm còn lại uống thêm một lon giả dược. Kết quả từ loại câu hỏi và phân tích này có thể được hiểu là tác động nhân quả của việc uống thêm 12 ounce lon nước ngọt mỗi ngày sẽ như thế nào đối với chỉ số BMI. Vì phân tích đang so sánh tác động trung bình lên BMI giữa hai nhóm (soda và giả dược), nên kết quả sẽ được hiểu là tác động nhân quả trung bình trong dân số.

Mục đích thứ ba của việc xem lại câu hỏi ban đầu của bạn là điều quan trọng là phải tạm dừng và xem xét liệu cách tiếp cận của bạn để trả lời câu hỏi có thể tạo ra một câu trả lời thiên vị hay không .

sult. Mặc dù chúng tôi đã đề cập đến sự thiên vị ở một mức độ nào đó trong chương về **Đặt ra và Tinh chỉnh Câu hỏi**, nhưng đôi khi thông tin mới thu được trong quá trình phân tích dữ liệu khám phá và/hoặc lập mô hình ảnh hưởng trực tiếp đến đánh giá của bạn về việc liệu kết quả của bạn có thể bị sai lệch hay không. Hãy nhớ lại rằng sự thiên vị là một vấn đề mang tính hệ thống đối với việc thu thập hoặc phân tích dữ liệu dẫn đến câu trả lời không chính xác cho câu hỏi của bạn.

Chúng tôi sẽ sử dụng ví dụ soda-BMI để minh họa một ví dụ đơn giản hơn về sự sai lệch. Giả sử rằng câu hỏi tổng thể của bạn về mối quan hệ soda-BMI đã bao gồm một câu hỏi ban đầu là: Mức tiêu thụ soda không ăn kiêng trung bình hàng ngày của người trưởng thành ở Hoa Kỳ là bao nhiêu? Giả sử rằng phân tích của bạn chỉ ra rằng trong mẫu bạn đang phân tích, là mẫu của tất cả người trưởng thành ở Hoa Kỳ, số lượng trung bình 12 ounce uống soda không ăn kiêng mỗi ngày là 0,5, vì vậy bạn suy ra rằng số lượng trung bình của 12 ounce khẩu phần soda uống mỗi ngày của người lớn ở Mỹ cũng là 0,5. Vì bạn phải luôn thách thức kết quả của mình, điều quan trọng là phải xem xét liệu phân tích của bạn có thiên kiến cố hữu hay không.

Vì vậy, làm thế nào để bạn làm điều này? Bạn bắt đầu bằng cách tưởng tượng rằng kết quả của bạn là không chính xác, sau đó suy nghĩ về những cách thức mà việc thu thập hoặc phân tích dữ liệu có thể đã có một vấn đề mang tính hệ thống dẫn đến ước tính không chính xác về số lượng trung bình của lon nước ngọt không ăn kiêng 12 ounce mỗi người uống hàng ngày của người lớn ở Mỹ. Mặc dù bài tập tưởng tượng rằng kết quả của bạn sai được thảo luận như một cách tiếp cận để đánh giá khả năng xảy ra sai lệch, nhưng đây là một cách tuyệt vời để thử thách kết quả của bạn ở mọi bước phân tích, cho dù bạn đang đánh giá rủi ro sai lệch, gây nhiễu hay một vấn đề kỹ thuật với phân tích của bạn.

Thí nghiệm tưởng tượng diễn ra như sau: hãy tưởng tượng rằng số lượng trung bình thực sự của 12 ounce khẩu phần không

Tỷ lệ uống soda ăn kiêng mỗi ngày của người trưởng thành ở Mỹ là 2. Bay giờ hãy tưởng tượng kết quả từ phân tích mẫu của bạn, là 0,5, có thể khác xa kết quả thực như thế nào: vì một lý do nào đó, mẫu dân số bao gồm tập dữ liệu của bạn không phải là một mẫu dân số ngẫu nhiên và thay vào đó có một số lượng không cân xứng những người không uống bất kỳ loại nước ngọt không dành cho người ăn kiêng nào, điều này làm giảm số lượng trung bình ước tính là 12 ounce khẩu phần nước ngọt dành cho người không ăn kiêng được tiêu thụ mỗi ngày. Bạn cũng có thể tưởng tượng rằng nếu kết quả mẫu của bạn là 4, cao hơn nhiều so với lượng thực sự người trưởng thành ở Hoa Kỳ uống mỗi ngày, thì mẫu của bạn có số lượng người tiêu thụ nhiều soda không ăn kiêng không tương xứng nên ước tính được tạo ra từ các phân tích của bạn cao hơn giá trị thực. Vì vậy, làm thế nào bạn có thể đánh giá xem mẫu của bạn có phải là ngẫu nhiên hay không?

Để tìm hiểu xem mẫu của bạn có phải là mẫu không ngẫu nhiên của dân số mục tiêu hay không, hãy nghĩ xem điều gì có thể xảy ra để thu hút thêm những người không tiêu thụ soda không dành cho người ăn kiêng (hoặc nhiều người tiêu thụ nhiều loại nước ngọt này) được đưa vào. Trong mẫu. Có lẽ nghiên cứu được quảng cáo cho sự tham gia của một tạp chí thể hình, và đặc giả của tạp chí thể hình ít có khả năng uống soda không ăn kiêng hơn. Hoặc có lẽ dữ liệu được thu thập bởi một cuộc khảo sát trên internet và những người trả lời khảo sát trên internet ít có khả năng uống nước ngọt không dành cho người ăn kiêng.

Hoặc có lẽ cuộc khảo sát đã thu thập thông tin về mức tiêu thụ soda không dành cho người ăn kiêng bằng cách cung cấp danh sách các loại nước ngọt không dành cho người ăn kiêng và yêu cầu những người tham gia khảo sát cho biết họ đã tiêu thụ loại nào, nhưng cuộc khảo sát đã bỏ qua Mountain Dew và Cherry Coke, để những người chủ yếu uống những loại này nước ngọt không dành cho người ăn kiêng được phân loại là không tiêu thụ nước ngọt dành cho người không ăn kiêng (hoặc tiêu thụ ít nước ngọt hơn mức họ thực sự tiêu thụ). Và như thế.

Mặc dù chúng tôi đã minh họa kịch bản đơn giản nhất cho sự thiên vị, mà

xảy ra khi ước tính tỷ lệ phổ biến hoặc giá trị trung bình, tất nhiên bạn cũng có thể nhận được kết quả sai lệch cho ước tính mối quan hệ giữa hai biến. Ví dụ: các phương pháp khảo sát có thể vô tình lấy mẫu quá mức những người không tiêu thụ nước ngọt không ăn kiêng và có chỉ số BMI cao (chẳng hạn như những người mắc bệnh tiểu đường loại 2), do đó kết quả sẽ chỉ ra (không chính xác) rằng tiêu thụ nước ngọt không ăn kiêng không liên quan đến việc có chỉ số BMI cao hơn. Vấn đề là việc tạm dừng để thực hiện một thử nghiệm suy nghĩ có chủ ý về các nguồn gây sai lệch là cực kỳ quan trọng vì đó thực sự là cách duy nhất để đánh giá khả năng dẫn đến một kết quả sai lệch. Thử nghiệm suy nghĩ này cũng nên được tiến hành khi bạn đang nêu và tinh chỉnh câu hỏi của mình cũng như khi bạn đang tiến hành phân tích và lập mô hình khám phá.

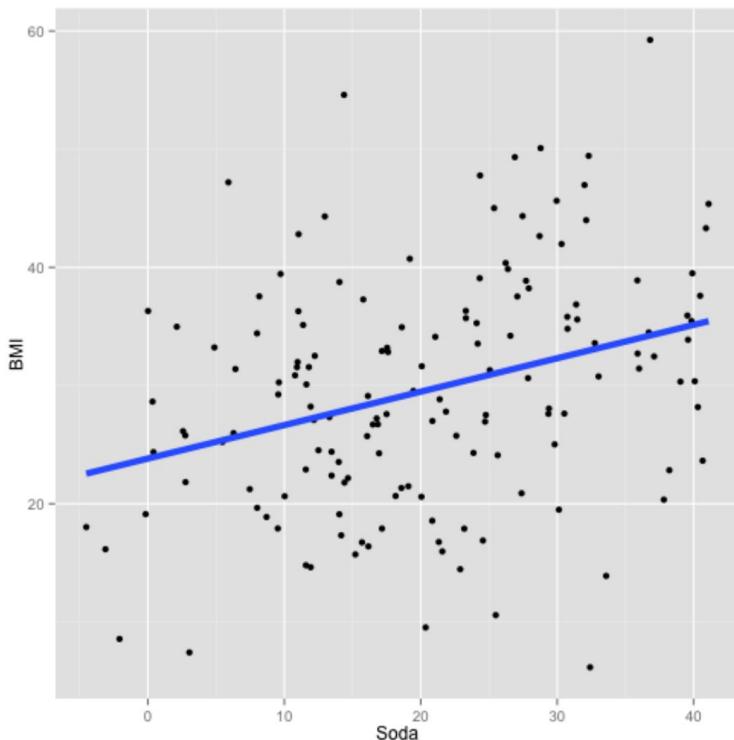
Bắt đầu với mô hình chính và đánh giá hướng, cường độ và độ không đảm bảo của kết quả

Nguyên tắc thứ hai là bắt đầu với một mô hình duy nhất và tập trung vào tính liên tục đầy đủ của kết quả, bao gồm cả hướng và độ lớn của nó, và mức độ chắc chắn (hoặc không chắc chắn) về việc liệu kết quả từ mẫu bạn phân tích có phản ánh đúng kết quả hay không cho người dân nói chung. Rất nhiều thông tin cần thiết để diễn giải kết quả của bạn sẽ bị bỏ sót nếu bạn phóng to một đặc điểm của kết quả (chẳng hạn như giá trị p), do đó bạn có thể bỏ qua hoặc che đậm thông tin quan trọng khác do mô hình cung cấp. Mặc dù diễn giải của bạn chưa hoàn chỉnh cho đến khi bạn xem xét toàn bộ các kết quả, nhưng trước tiên, điều hữu ích nhất là tập trung vào diễn giải các kết quả của mô hình mà bạn tin rằng sẽ trả lời tốt nhất câu hỏi của bạn và phản ánh (hoặc “phù hợp”) với dữ liệu của bạn, đó là mô hình chính của bạn (Xem Mô hình chính thức). Đừng dành nhiều thời gian lo lắng về việc bắt đầu với mô hình đơn lẻ nào, bởi vì trong

cuối cùng, bạn sẽ xem xét tất cả các kết quả của mình và bài tập diễn giải ban đầu này nhằm định hướng cho bạn và cung cấp một khuôn khổ cho diễn giải cuối cùng của bạn.

định hướng

Dựa trên ví dụ soda-BMI, hãy xem tập dữ liệu mẫu bên dưới với một mô hình phù hợp được phủ lên.



Dữ liệu mẫu cho ví dụ BMI-soda

Chúng ta sẽ tập trung vào những gì mô hình cho chúng ta biết về chiều hướng của mối quan hệ giữa lượng tiêu thụ soda và chỉ số BMI, mức độ của mối quan hệ và mức độ không chắc chắn.

mức độ mờ nhạt của mối quan hệ, hoặc khả năng mô tả của mô hình về mối quan hệ giữa mức tiêu thụ soda không dành cho người ăn kiêng và chỉ số BMI là có thật so với chỉ phản ánh sự thay đổi ngẫu nhiên bạn mong đợi khi lấy mẫu từ một quần thể lớn hơn.

Mô hình chỉ ra rằng tính định hướng của mối quan hệ là tích cực, nghĩa là khi mức tiêu thụ soda không dành cho người ăn kiêng tăng, chỉ số BMI tăng. Các kết quả tiềm năng khác có thể đã là một định hướng tiêu cực, hoặc không có định hướng (a giá trị xấp xỉ 0). Liệu hướng tích cực của kết quả phù hợp với mong đợi của bạn đã được phát triển từ phân tích dữ liệu khám phá? Nếu vậy, bạn đang trong tình trạng tốt và có thể chuyển sang giải thích tiếp theo hoạt động. Nếu không, có một vài lời giải thích có thể. Đầu tiên, kỳ vọng của bạn có thể không đúng bởi vì một trong hai phân tích thăm dò đã được thực hiện không chính xác hoặc cách giải thích của bạn về các phân tích thăm dò là không chính xác. Thứ hai, phân tích thăm dò và diễn giải của bạn về nó có thể đúng, nhưng mô hình chính thức có thể đã được thực hiện không chính xác. Lưu ý rằng với quá trình này, bạn đã từng một lần nữa áp dụng chu kỳ phân tích dữ liệu.

Kích cỡ

Khi bạn đã xác định và giải quyết bất kỳ sự khác biệt nào giữa kỳ vọng của bạn và diễn giải về chiều hướng của mối quan hệ, bước tiếp theo là xem xét độ lớn của mối quan hệ. Vì mô hình là tuyến tính hồi quy, bạn có thể thấy rằng độ dốc của mối quan hệ, được phản ánh bởi hệ số beta, là 0,28. phiên dịch các độ dốc yêu cầu biết các đơn vị của biến "soda". Nếu đơn vị là lon nước ngọt 12 ounce mỗi ngày, thì sự xen kẽ của độ dốc này là BMI tăng 0,28 kg/m² trên mỗi lon soda không ăn kiêng 12 ounce bổ sung đó là tiêu thụ mỗi ngày. Tuy nhiên, các đơn vị tính bằng ounce soda,

do đó, cách giải thích mô hình của bạn là chỉ số BMI tăng thêm 0,28 kg/m² cho mỗi ounce bổ sung soda không ăn kiêng được tiêu thụ mỗi ngày.

Mặc dù bạn cảm thấy thoải mái khi hiểu chính xác các đơn vị của biến soda và có cách diễn giải chính xác về mô hình, nhưng bạn vẫn chưa hoàn toàn có câu trả lời cho câu hỏi của mình, vốn được đóng khung theo mối liên hệ của từng biến bổ sung. Lon nước ngọt 12 ounce và chỉ số BMI, không phải thêm mỗi ounce nước ngọt không dành cho người ăn kiêng. Vì vậy, bạn sẽ cần chuyển đổi độ dốc 0,28 để nó tương ứng với mức tăng tiêu thụ soda là 12 ounce, thay vì 1 ounce. Vì mô hình này là mô hình tuyến tính nên bạn chỉ cần nhân hệ số góc hoặc hệ số beta với 12 để có 3,36, cho bạn biết rằng mỗi lon soda 12 ounce bổ sung được tiêu thụ mỗi ngày có liên quan đến chỉ số BMI là 3,36 kg/m² cao hơn.

Tất nhiên, tùy chọn khác là tạo một biến soda mới có đơn vị là 12 ounce thay vì 1 ounce, nhưng nhân hệ số góc là một phép toán đơn giản và hiệu quả hơn nhiều. Ở đây, một lần nữa, bạn nên có một số kỳ vọng, dựa trên phân tích dữ liệu khám phá mà bạn đã thực hiện, về mức độ của mối quan hệ giữa mức tiêu thụ soda không dành cho người ăn kiêng và chỉ số BMI, vì vậy bạn nên xác định xem cách diễn giải của mình về mức độ của mối quan hệ có phù hợp với mong đợi của bạn hay không. Nếu không, bạn sẽ cần xác định xem kỳ vọng của mình có sai hay không hoặc liệu cách giải thích của bạn có sai không và hành động phù hợp để phù hợp với kỳ vọng và kết quả của cách giải thích của bạn.

Một cân nhắc quan trọng khác về tầm quan trọng của mối quan hệ là liệu nó có ý nghĩa hay không. Ví dụ: chỉ số BMI tăng 0,01 cho mỗi 20 ounce bổ sung được tiêu thụ mỗi ngày có lẽ không có ý nghĩa đặc biệt vì một lượng lớn soda có liên quan đến mức tăng rất nhỏ chỉ số BMI. Mặt khác, nếu có 0,28 kg/m²

tăng chỉ số BMI cho mỗi 1 ounce tăng tiêu thụ soda, điều này trên thực tế sẽ khá có ý nghĩa. Bởi vì bạn biết chỉ số BMI thường dao động từ độ tuổi thanh thiếu niên đến 30 tuổi, thay đổi $0,01 \text{ kg/m}^2$ là nhỏ nhưng thay đổi $0,28 \text{ kg/m}^2$ có thể có ý nghĩa.

Khi xét đến các loại khói lượng soda mà mọi người có thể tiêu thụ, mức tăng $0,01 \text{ kg/m}^2$ cho mỗi 20 ounce lượng soda tiêu thụ là nhỏ vì mọi người (hy vọng) không uống 10 khẩu phần 20 ounce mỗi ngày, tức là bao nhiêu ai đó sẽ cần phải uống để quan sát chỉ số BMI tăng $0,1 \text{ kg/m}^2$. Mặt khác, chỉ số BMI tăng thêm $0,28 \text{ kg/m}^2$ đối với mỗi ounce soda bổ sung sẽ tăng lên nhanh chóng đối với những người tiêu thụ thêm 20 ounce soda không ăn kiêng mỗi ngày - điều này tương đương với mức tăng BMI dự kiến là $5,6 \text{ kg/m}^2$.

Sau đó, một phần quan trọng trong việc diễn giải tầm quan trọng của kết quả là hiểu mức độ của kết quả so với những gì bạn biết về loại thông tin này trong dân số mà bạn quan tâm.

Tính không chắc chắn

Bây giờ bạn đã nắm được những gì mô hình nói về định hướng và mức độ của mối quan hệ giữa mức tiêu thụ soda không ăn kiêng và chỉ số BMI, bước tiếp theo là xem xét mức độ không chắc chắn cho câu trả lời của bạn. Hãy nhớ lại rằng mô hình của bạn đã được xây dựng để phù hợp với dữ liệu được thu thập từ một mẫu của toàn bộ dân số và bạn đang sử dụng mô hình này để hiểu mức độ tiêu thụ soda không dành cho người ăn kiêng có liên quan như thế nào đến chỉ số BMI trong toàn bộ dân số trưởng thành ở Hoa Kỳ.

Hãy quay lại ví dụ soda-BMI của chúng ta, ví dụ này liên quan đến việc sử dụng các kết quả thu được trên mẫu để đưa ra suy luận về mối quan hệ soda-BMI thực sự là gì

tổng dân số trưởng thành ở Hoa Kỳ. Hãy tưởng tượng rằng kết quả từ phân tích dữ liệu mẫu của bạn chỉ ra rằng trong mẫu của bạn, những người uống thêm một ounce soda không ăn kiêng mỗi ngày có chỉ số BMI cao hơn 0,28 kg/m² so với những người uống ít hơn một ounce mỗi ngày.

Tuy nhiên, làm thế nào để bạn biết liệu kết quả này chỉ đơn giản là “tiếng ồn” của việc lấy mẫu ngẫu nhiên hay đó là một kết quả gần đúng về mối quan hệ thực sự giữa tổng thể dân số?

Để đánh giá xem kết quả từ mẫu có đơn giản là “nhiều” ngẫu nhiên hay không, chúng tôi sử dụng các phép đo độ không đảm bảo. Mặc dù một số người có thể kỳ vọng rằng tất cả các mẫu ngẫu nhiên đóng vai trò là đại diện thay thế tuyệt vời cho tổng thể dân số, nhưng điều này không đúng. Để minh họa ý tưởng này bằng một ví dụ đơn giản, hãy tưởng tượng rằng tỷ lệ phụ nữ trong tổng dân số trưởng thành của Hoa Kỳ là 51% và bạn lấy một mẫu ngẫu nhiên gồm 100 người trưởng thành. Mẫu này có thể có 45% nữ. Hãy tưởng tượng rằng bạn lấy một mẫu mới gồm 100 người trưởng thành và mẫu của bạn có 53% là nữ. Bạn có thể vẽ nhiều mẫu như thế này và thậm chí vẽ mẫu 35% hoặc 70% là nữ. Xác suất rút ra một mẫu có tỷ lệ nữ giới khác với tỷ lệ nữ giới chung của quần thể là rất nhỏ, trong khi xác suất lấy mẫu có gần 51% nữ giới lại cao hơn nhiều.

Chính khái niệm này – xác suất mà mẫu của bạn phản ánh câu trả lời cho toàn bộ tổng thể thay đổi tùy thuộc vào mức độ gần (hoặc xa) của kết quả mẫu của bạn với kết quả thực của tổng thể – đó là nền tảng của khái niệm về sự không chắc chắn. Bởi vì chúng tôi không biết câu trả lời cho tổng thể là gì (đó là lý do tại sao chúng tôi thực hiện phân tích ngay từ đầu!), nên không thể thể hiện sự không chắc chắn về khả năng hoặc không chắc chắn rằng kết quả mẫu của bạn phản ánh tổng thể dân số.

Vì vậy, có những cách tiếp cận khác để đo độ không đảm bảo

dựa trên khái niệm chung này và chúng tôi sẽ thảo luận về hai cách tiếp cận phổ biến bên dưới.

Một công cụ cung cấp thước đo liên tục hơn về độ không chắc chắn là khoảng tin cậy. Khoảng tin cậy là một dải giá trị chứa kết quả mẫu của bạn và bạn có chút tin tưởng rằng khoảng tin cậy đó cũng chưa kết quả thực cho toàn bộ tổng thể. Hầu hết các phần mềm lập mô hình thống kê thường cung cấp khoảng tin cậy 95%, do đó nếu 95% CI cho ước tính mẫu $0,28 \text{ kg/m}^2$ từ trên xuống là $0,15-0,42 \text{ kg/m}^2$, thì cách diễn giải gần đúng là bạn có thể tin tưởng 95% rằng kết quả thực sự cho toàn bộ dân số nằm trong khoảng từ $0,15$ đến $0,42 \text{ kg/m}^2$.

Một định nghĩa chính xác hơn về khoảng tin cậy 95% là trên các mẫu lặp lại, nếu chúng ta tiến hành thử nghiệm này nhiều lần (mỗi lần thu thập một tập dữ liệu có cùng kích thước) thì khoảng tin cậy được xây dựng theo cách này sẽ bao trùm sự thật 95% thời gian. Điều quan trọng là phải nhận ra rằng vì khoảng tin cậy được xây dựng từ dữ liệu, bản thân khoảng tin cậy là ngẫu nhiên. Do đó, nếu chúng tôi thu thập dữ liệu mới, khoảng thời gian chúng tôi xây dựng sẽ hơi khác một chút. Tuy nhiên, sự thật, nghĩa là giá trị tổng thể của tham số, sẽ luôn giữ nguyên.

Tất nhiên, một công cụ khác để đo lường sự không chắc chắn là giá trị p, đơn giản là xác suất nhận được kết quả mẫu là $0,28 \text{ kg/m}^2$ (hoặc cực đoan hơn) khi mối quan hệ thực sự giữa mức tiêu thụ soda không ăn kiêng và BMI trong tổng thể dân số là 0. Mặc dù giá trị p là thước đo liên tục của sự không chắc chắn, nhưng nhiều người coi

giá trị $p <0,05$, cho biết rằng có ít hơn 5% xác suất quan sát kết quả mẫu (hoặc kết quả cực đoan hơn) khi không có mối quan hệ nào trong toàn bộ tổng thể, là "có ý nghĩa thống kê". Điểm cắt này là arbitrary và cho chúng ta biết rất ít về mức độ không chắc chắn hoặc về câu trả lời thực sự cho toàn bộ dân số nằm ở đâu. Tập trung chủ yếu vào giá trị p là một cách tiếp cận rủi ro để diễn giải sự không chắc chắn vì nó có thể dẫn đến việc bỏ qua những thông tin quan trọng hơn cần thiết để diễn giải kết quả của bạn một cách chu đáo và chính xác.

CI hữu ích hơn giá trị p , bởi vì nó đưa ra một phạm vi, cung cấp một số ước tính định lượng về kết quả tổng thể thực tế có thể là gì và nó cũng cung cấp một cách để thể hiện mức độ chắc chắn của phạm vi đó. Kết quả tổng thể dân số.

Hãy xem cách sử dụng giá trị p so với 95% CI để diễn giải sự không chắc chắn về kết quả từ phân tích soda-BMI. Giả sử kết quả của chúng tôi là chỉ số BMI trung bình cao hơn $0,28 \text{ kg/m}^2$ trong số mẫu của chúng tôi, những người uống nhiều hơn 1 ounce soda không ăn kiêng mỗi ngày và giá trị p liên quan đến kết quả này là $0,03$. Sử dụng giá trị p làm công cụ để đo lường độ không đảm bảo và đặt ngưỡng có ý nghĩa thống kê ở mức $0,05$, chúng tôi sẽ xử lý độ không đảm bảo như sau: có ít hơn 5% khả năng chúng tôi sẽ nhận được kết quả này ($0,28$) hoặc hơn thế nữa cực đoan nếu giá trị dân số thực là 0 (hay nói cách khác, thực sự không có mối liên hệ nào giữa mức tiêu thụ soda và BMI trong tổng thể dân số).

Bây giờ chúng ta hãy thực hiện bài tập tương tự với 95% CI. Khoảng tin cậy 95% cho phân tích này là $0,15-0,42$. Sử dụng CI làm công cụ để giải thích sự không chắc chắn, chúng tôi có thể nói rằng chúng tôi tin tưởng 95% rằng mối quan hệ thực sự giữa mức tiêu thụ soda và BMI ở dân số Hoa Kỳ trưởng thành nằm ở

ở đâu đó giữa mức tăng BMI trung bình từ 0,15 đến 0,42 kg/m² trên mỗi ounce bổ sung soda không ăn kiêng được tiêu thụ. Sử dụng phương pháp thứ hai này cho chúng ta biết điều gì đó về phạm vi ảnh hưởng có thể có của soda đối với BMI và cũng cho chúng ta biết rằng rất ít khả năng soda không có mối liên hệ nào với BMI trong toàn bộ dân số trưởng thành ở Hoa Kỳ.

Mặt khác, việc sử dụng giá trị p làm thước đo độ không chắc chắn ngữ ý rằng chúng ta chỉ có hai lựa chọn về mặt diễn giải kết quả: hoặc có một lượng lớn độ không chắc chắn về nó nên chúng ta phải kết luận rằng không có mối quan hệ nào. giữa mức tiêu thụ soda và BMI, hoặc có rất ít sự không chắc chắn về kết quả nên chúng ta phải kết luận rằng có mối quan hệ giữa mức tiêu thụ soda và BMI. Việc sử dụng giá trị p ràng buộc chúng ta theo cách không phản ánh quá trình cân nhắc độ mạnh của bằng chứng ủng hộ (hoặc chống lại) một giả thuyết.

Một điểm khác về sự không chắc chắn là chúng ta đã thảo luận về việc đánh giá sự không chắc chắn thông qua các cách tiếp cận thống kê cổ điển hơn, dựa trên mô hình Người theo chủ nghĩa thường xuyên, là cách tiếp cận phổ biến nhất. Công việc khung Bayesian là một cách tiếp cận thay thế trong đó bạn cập nhật niềm tin trước đây của mình dựa trên bằng chứng do phân tích cung cấp.

Trong thực tế, phương pháp Người theo chủ nghĩa thường xuyên mà chúng ta đã thảo luận ở trên được sử dụng phổ biến hơn và trong môi trường thực tế hiếm khi dẫn đến kết luận khác với kết luận thu được bằng cách sử dụng phương pháp Bayes.

Một lưu ý quan trọng là đôi khi việc đánh giá độ không chắc chắn là không cần thiết vì một số loại phân tích không nhằm đưa ra kết luận về một tổng thể lớn hơn. Ví dụ: nếu bạn muốn hiểu mối quan hệ giữa độ tuổi và số tiền chi tiêu mỗi tháng cho các sản phẩm của công ty bạn, thì bạn có thể có tất cả dữ liệu về toàn bộ hoặc "tổng thể" dân số mà bạn quan tâm là khách hàng của công ty bạn. Trong trường hợp này bạn không

phải dựa vào một mẫu, bởi vì công ty của bạn thu thập dữ liệu về độ tuổi và hoạt động mua hàng của TẤT CẢ khách hàng của họ. Trong trường hợp này, bạn sẽ không cần xem xét sự không chắc chắn rằng kết quả của bạn phản ánh sự thật đối với tổng thể vì kết quả phân tích của bạn là sự thật đối với tổng thể của bạn.

Phát triển một diễn giải tổng thể bằng cách xem xét toàn bộ các phân tích của bạn và thông tin bên ngoài

Bây giờ bạn đã dành nhiều nỗ lực để giải thích các kết quả của mô hình chính của mình, bước tiếp theo là phát triển cách diễn giải tổng thể các kết quả của bạn bằng cách xem xét cả tổng thể các phân tích và thông tin bên ngoài các phân tích của bạn. Việc diễn giải kết quả từ mô hình chính của bạn phục vụ để đặt kỳ vọng cho diễn giải tổng thể của bạn khi bạn xem xét tất cả các phân tích của mình. Dựa trên ví dụ về chỉ số BMI-soda, giả sử rằng cách giải thích của bạn về mô hình chính là chỉ số BMI trung bình cao hơn 0,28 kg/m² ở những người trưởng thành ở Hoa Kỳ tiêu thụ thêm trung bình 1 ounce soda mỗi ngày. Hãy nhớ lại rằng mô hình chính này được xây dựng sau khi thu thập thông tin thông qua các phân tích thăm dò và bạn có thể đã tinh chỉnh mô hình này khi bạn đang thực hiện quá trình diễn giải các kết quả của nó bằng cách đánh giá hướng, độ lớn và độ không chắc chắn của các kết quả của mô hình.

Như đã thảo luận trong [chương Lập mô hình chính thức](#), không có một mô hình đơn lẻ nào cung cấp câu trả lời cho câu hỏi của bạn. Thay vào đó, có những mô hình bổ sung phục vụ để thách thức kết quả thu được trong mô hình chính. Một loại mô hình thứ cấp phổ biến là mô hình được xây dựng để xác định mức độ nhạy cảm của các kết quả trong

mô hình chính là những thay đổi trong dữ liệu. Một ví dụ điển hình là loại bỏ các giá trị ngoại lệ để đánh giá mức độ thay đổi của kết quả mô hình chính của bạn. Nếu kết quả của mô hình sơ cấp phần lớn được thúc đẩy bởi một số ít, chẳng hạn như những người tiêu thụ nhiều nước ngọt có ga, thì phát hiện này sẽ gợi ý rằng có thể không có mối quan hệ tuyến tính giữa mức tiêu thụ nước ngọt có ga và chỉ số BMI và thay vào đó, lượng tiêu thụ nước ngọt có ga chỉ có thể ảnh hưởng đến chỉ số BMI ở những người có mức tiêu thụ soda rất cao. Phát hiện này sẽ dẫn đến việc sửa đổi mô hình chính của bạn.

Ví dụ thứ hai là đánh giá tác động của những kẻ lừa đảo tiềm ẩn đối với kết quả từ mô hình chính. Mặc dù mô hình chính đã chứa các yếu tố gây nhiễu chính, nhưng thường có các yếu tố gây nhiễu tiềm năng bỗng nhiên cần được đánh giá. Trong ví dụ soda-BMI, bạn có thể xây dựng một mô hình thứ cấp bao gồm thu nhập vì bạn nhận ra rằng mối quan hệ mà bạn quan sát thấy trong mô hình chính của mình có thể được giải thích hoàn toàn bằng tình trạng kinh tế xã hội: những người có tình trạng kinh tế xã hội cao hơn có thể uống ít hơn - diet soda và cũng có chỉ số BMI thấp hơn, nhưng không phải vì họ uống ít soda mà xảy ra trường hợp này. Thay vào đó, một số yếu tố khác liên quan đến tình trạng kinh tế xã hội có ảnh hưởng đến BMI. Vì vậy, bạn có thể chạy một mô hình thứ cấp trong đó thu nhập được thêm vào mô hình chính để xác định xem đây có phải là trường hợp không. Mặc dù có những ví dụ khác về việc sử dụng các mô hình thứ cấp, đây là hai ví dụ phổ biến.

Vì vậy, làm thế nào để bạn giải thích làm thế nào các kết quả mô hình phụ này ảnh hưởng đến kết quả chính của bạn? Bạn có thể rơi vào mô hình của: tính định hướng, độ lớn và sự không chắc chắn. Khi bạn thêm thu nhập vào mô hình soda-BMI, điều đó có làm thay đổi hướng của mối quan hệ ước tính của bạn giữa soda và BMI từ mô hình chính - thành mối liên hệ tiêu cực hoặc không có mối liên hệ nào không? Nếu nó đã làm, rằng

sẽ là một thay đổi đáng kể và gợi ý rằng có điều gì đó không đúng với dữ liệu của bạn (chẳng hạn như với biến thu nhập) hoặc mối liên hệ giữa mức tiêu thụ soda và BMI hoàn toàn được giải thích bằng thu nhập.

Giả sử rằng việc thêm thu nhập không làm thay đổi hướng và giả sử rằng nó đã thay đổi cường độ sao cho ước tính của mô hình chính là $0,28 \text{ kg/m}^2$ giảm xuống $0,12\text{kg/m}^2$.

Mức độ của mối quan hệ giữa soda và BMI đã giảm 57%, do đó, điều này sẽ được hiểu là thu nhập giải thích hơn một nửa, nhưng không phải tất cả, mối quan hệ giữa mức tiêu thụ soda và BMI.

Bây giờ bạn chuyển sang sự không chắc chắn. Khoảng tin cậy 95% cho ước tính với mô hình bao gồm thu nhập là $0,01-0,23$, do đó chúng tôi có thể tin tưởng 95% rằng mối quan hệ thực sự giữa soda và BMI trong dân số Hoa Kỳ trưởng thành, không phụ thuộc vào thu nhập, nằm ở đâu đó trong phạm vi này. Điều gì sẽ xảy ra nếu 95% CI cho ước tính là $-0,02-0,26$, nhưng ước tính vẫn là $0,12 \text{ kg/m}^2$? Mặc dù CI hiện bao gồm 0, kết quả từ mô hình chính, $0,12$, không thay đổi, cho thấy rằng thu nhập dường như không giải thích bất kỳ mối liên hệ nào giữa mức tiêu thụ soda và BMI, nhưng nó đã làm tăng sự không chắc chắn của kết quả. Một lý do khiến việc bổ sung thu nhập vào mô hình có thể làm tăng tính không chắc chắn là một số người trong mẫu đã thiếu dữ liệu thu nhập nên cỡ mẫu bị giảm. Kiểm tra n của bạn sẽ giúp bạn xác định xem đây có phải là trường hợp không.

Điều quan trọng nữa là xem xét kết quả tổng thể của bạn trong bối cảnh thông tin bên ngoài. Thông tin bên ngoài vừa là kiến thức chung mà bạn hoặc các thành viên trong nhóm của bạn có về chủ đề, kết quả từ các phân tích tương tự và thông tin về dân số mục tiêu. Một ví dụ đã thảo luận ở trên là việc có ý thức về lượng tiêu thụ soda điển hình và hợp lý ở những người trưởng thành trong

Hoa Kỳ rất hữu ích để hiểu liệu mức độ ảnh hưởng của việc tiêu thụ soda đối với BMI có ý nghĩa hay không. Cũng có thể hữu ích khi biết bao nhiêu phần trăm dân số trưởng thành ở Hoa Kỳ uống soda không ăn kiêng và tỷ lệ béo phì để hiểu quy mô dân số mà kết quả của bạn có thể phù hợp.

Một ví dụ thú vị về tầm quan trọng của việc suy nghĩ về quy mô dân số có thể bị ảnh hưởng là ô nhiễm không khí. Đối với mối liên hệ giữa ô nhiễm không khí ngoài trời và các hậu quả sức khỏe nghiêm trọng như các biến cố tim mạch (đột quỵ, đau tim), mức độ ảnh hưởng là nhỏ, nhưng do ô nhiễm không khí ảnh hưởng đến hàng trăm triệu người ở Hoa Kỳ, số lượng các biến cố tim mạch do ô nhiễm khá cao.

Ngoài ra, bạn có thể không phải là người đầu tiên thử và trả lời câu hỏi này hoặc các câu hỏi liên quan. Những người khác có thể đã thực hiện một phân tích để trả lời câu hỏi trong một nhóm dân số khác (ví dụ như thanh thiếu niên) hoặc để trả lời một câu hỏi khác, nhưng có liên quan, chẳng hạn như: "mối quan hệ giữa việc tiêu thụ soda không ăn kiêng và mức đường trong máu là gì?" Việc hiểu kết quả của bạn phù hợp như thế nào với bối cảnh của khái kiến thức về chủ đề này giúp bạn và những người khác đánh giá liệu có một câu chuyện hoặc mô hình tổng thể nào xuất hiện trên tất cả các nguồn kiến thức chỉ ra việc tiêu thụ soda không dành cho người ăn kiêng có liên quan đến lượng đường trong máu cao, insulin hay không sức đề kháng, chỉ số BMI và bệnh tiểu đường loại 2. Mặt khác, nếu kết quả phân tích của bạn khác với cơ sở tri thức bên ngoài, thì điều đó cũng quan trọng. Mặc dù hầu hết các trường hợp khi kết quả rất khác biệt so với kiến thức bên ngoài, có một lời giải thích như lỗi hoặc sự khác biệt trong phương pháp thu thập dữ liệu hoặc dân số được nghiên cứu, đôi khi một phát hiện khác biệt rõ ràng là một cái nhìn sâu sắc thực sự mới lạ.

Hàm ý

Bây giờ bạn đã giải thích các kết quả của mình và có trong tay các kết luận, bạn sẽ muốn suy nghĩ về ý nghĩa của các kết luận của mình. Xét cho cùng, mục đích của việc khám hậu môn thường là để thông báo một quyết định hoặc thực hiện một hành động. Đôi khi những hàm ý rất đơn giản, nhưng những lúc khác, những hàm ý cần phải suy nghĩ. Một ví dụ về hàm ý đơn giản là nếu bạn thực hiện phân tích để xác định xem việc mua quảng cáo có làm tăng doanh số bán hàng hay không và nếu có thì khoản đầu tư vào quảng cáo có mang lại lợi nhuận ròng hay không. Bạn có thể biết rằng có hoặc không có lợi nhuận ròng và nếu có lợi nhuận ròng, phát hiện này sẽ hỗ trợ việc tiếp tục quảng cáo.

Một ví dụ phức tạp hơn là ví dụ soda-BMI mà chúng ta đã sử dụng trong suốt chương này. Nếu tiêu thụ soda hóa ra có liên quan đến chỉ số BMI cao hơn, với khẩu phần bổ sung 20 ounce mỗi ngày có liên quan đến chỉ số BMI cao hơn 0,28 kg/m², thì phát hiện này có nghĩa là nếu bạn có thể giảm tiêu thụ soda, bạn có thể giảm chỉ số BMI trung bình của tổng thể dân số. Tuy nhiên, vì phân tích của bạn không phải là nguyên nhân và bạn chỉ chứng minh được mối liên hệ, nên bạn có thể muốn thực hiện một nghiên cứu trong đó bạn chỉ định ngẫu nhiên những người thay thế một trong số 20 ounce nước ngọt họ uống mỗi ngày bằng nước ngọt dành cho người ăn kiêng hoặc để không thay thế soda không ăn kiêng của họ. Tuy nhiên, trong môi trường y tế công cộng, nhóm của bạn có thể quyết định rằng mối liên hệ này là bằng chứng đầy đủ để khởi động một chiến dịch y tế công cộng nhằm giảm mức tiêu thụ soda và bạn không cần thêm dữ liệu từ một thử nghiệm lâm sàng. Thay vào đó, bạn có thể lên kế hoạch theo dõi chỉ số BMI của dân số trong và sau chiến dịch y tế công cộng như một phương tiện để ước tính tác động sức khỏe cộng đồng của việc giảm tiêu thụ soda không dành cho người ăn kiêng. Điểm rút ra ở đây là hành động xuất phát từ các hàm ý thường phụ thuộc vào nhiệm vụ của tổ chức đã yêu cầu phân tích.

10. Giao tiếp

Giao tiếp là nền tảng để phân tích dữ liệu tốt. Mục đích của chúng tôi là giải quyết trong chương này là vai trò của giao tiếp thông thường trong quá trình thực hiện phân tích dữ liệu của bạn và phô biến kết quả cuối cùng của bạn trong một môi trường trang trọng hơn, thường là cho đối tượng lớn hơn, bên ngoài. Có rất nhiều cuốn sách hay đề cập đến "cách thức" của việc thuyết trình trang trọng, dưới dạng bài nói chuyện hoặc bài viết, chẳng hạn như sách trắng hoặc bài báo khoa học. Tuy nhiên, trong chương này, chúng ta sẽ tập trung vào:

1. Cách sử dụng giao tiếp thông thường như một trong những công cụ cần thiết để thực hiện phân tích dữ liệu tốt; và
2. Làm thế nào để truyền đạt những điểm chính trong phân tích dữ liệu của bạn khi giao tiếp không chính thức và chính thức.

Giao tiếp vừa là một trong những công cụ phân tích dữ liệu, vừa là sản phẩm cuối cùng của phân tích dữ liệu: sẽ chẳng ích gì khi thực hiện phân tích dữ liệu nếu bạn không truyền đạt quy trình và kết quả của mình tới khán giả. Một nhà phân tích dữ liệu giỏi giao tiếp không chính thức nhiều lần trong quá trình phân tích dữ liệu và cũng suy nghĩ cẩn thận để truyền đạt kết quả cuối cùng sao cho phân tích hữu ích và cung cấp nhiều thông tin nhất có thể cho nhiều đối tượng hơn.

10.1 Giao tiếp thông thường

Mục đích chính của giao tiếp thông thường là thu thập dữ liệu, đây là một phần của quy trình tuần hoàn cho mỗi lõi

hoạt động. Bạn thu thập dữ liệu bằng cách truyền đạt kết quả của mình và phản hồi bạn nhận được từ khán giả sẽ thông báo các bước tiếp theo trong quá trình phân tích dữ liệu của bạn. Các loại phản hồi mà bạn nhận được không chỉ bao gồm câu trả lời cho các câu hỏi cụ thể mà còn cả bình luận và câu hỏi mà khán giả đặt ra để phản hồi báo cáo của bạn (bằng văn bản hoặc bằng lời nói). Hình thức giao tiếp thông thường của bạn phụ thuộc vào mục tiêu của giao tiếp là gì. Ví dụ: nếu mục tiêu của bạn là làm rõ cách một biến được mã hóa bởi vì khi bạn khám phá tập dữ liệu, nó có vẻ là một biến thứ tự, nhưng bạn đã hiểu rằng đó là một biến liên tục, thì thông tin liên lạc của bạn sẽ ngắn gọn và đi thẳng vào vấn đề. .

Mặt khác, nếu một số kết quả từ phân tích dữ liệu khám phá của bạn không như bạn mong đợi, thông tin liên lạc của bạn có thể ở dạng một cuộc họp nhỏ, không chính thức bao gồm việc hiển thị các bảng và/hoặc số liệu liên quan đến vấn đề của bạn. Loại giao tiếp thân mật thứ ba là loại giao tiếp mà bạn có thể không có câu hỏi cụ thể để hỏi khán giả của mình, nhưng thay vào đó, bạn đang tìm kiếm phản hồi về quy trình phân tích dữ liệu và/hoặc kết quả để giúp bạn tinh chỉnh quy trình và/hoặc để thông báo cho lần tiếp theo của bạn các bước.

Tóm lại, có ba loại giao tiếp thân mật chính và chúng được phân loại dựa trên mục tiêu giao tiếp của bạn: (1) để trả lời một câu hỏi rất tập trung, thường là câu hỏi kỹ thuật hoặc câu hỏi nhằm thu thập thông tin. , (2) để giúp bạn giải quyết một số kết quả khó hiểu hoặc không hoàn toàn như bạn mong đợi và (3) để nhận được các án tượng và phản hồi chung như một phương tiện xác định các sự cố chưa từng xảy ra với bạn để bạn có thể tinh chỉnh dữ liệu của mình Phân tích.

Tập trung vào một vài khái niệm cốt lõi sẽ giúp bạn đạt được mục tiêu của mình khi lập kế hoạch giao tiếp thông thường. Những khái niệm này là:

1. Khán giả: Biết khán giả của bạn và khi bạn kiểm soát được khán giả là ai, hãy chọn đúng đối tượng cho loại phản hồi mà bạn đang tìm kiếm.
2. Nội dung: Tập trung, ngắn gọn nhưng cung cấp đủ thông tin để người hiểu được thông tin bạn trình bày và vấn đề bạn đặt ra.
3. Phong cách: Tránh biệt ngữ. Trừ khi bạn đang truyền đạt về một vấn đề kỹ thuật cao tập trung cho đối tượng kỹ thuật cao, tốt nhất là sử dụng ngôn ngữ, số liệu và bảng biểu mà đối tượng chung chung hơn có thể hiểu được.
4. Thái độ: Có thái độ cởi mở, hợp tác để bạn sẵn sàng tham gia hoàn toàn vào cuộc đối thoại và để khán giả của bạn nhận được thông điệp rằng mục tiêu của bạn không phải là "bảo vệ" câu hỏi hoặc công việc của bạn, mà là để họ tiếp nhận ý kiến của họ một cách đúng đắn. rằng bạn có thể làm công việc tốt nhất của bạn.

10.2 Khán giả

Đối với nhiều loại giao tiếp thông thường, bạn sẽ có khả năng chọn đối tượng của mình, nhưng trong một số trường hợp, chẳng hạn như khi bạn gửi báo cáo tạm thời cho sép hoặc nhóm của mình, đối tượng có thể được xác định trước. Khán giả của bạn có thể bao gồm các nhà phân tích dữ liệu khác, (những) cá nhân đã đặt câu hỏi, sép của bạn và/hoặc những người quản lý khác hoặc thành viên nhóm điều hành, những nhà phân tích phi dữ liệu là chuyên gia nội dung và/hoặc ai đó đại diện cho công chúng nói chung .

Đối với loại giao tiếp thông thường đầu tiên, trong đó bạn chủ yếu tìm kiếm kiến thức thực tế hoặc làm rõ về tập dữ liệu hoặc thông tin liên quan, hãy chọn một người (hoặc nhiều người) có kiến thức thực tế để trả lời

câu hỏi và đáp ứng các truy vấn là thích hợp nhất.

Đối với câu hỏi về cách thu thập dữ liệu cho một biến trong tập dữ liệu, bạn có thể tiếp cận người đã thu thập dữ liệu hoặc người đã làm việc với tập dữ liệu trước đây hoặc chịu trách nhiệm biên soạn dữ liệu.

Nếu câu hỏi là về lệnh sử dụng trong ngôn ngữ lập trình thống kê để chạy một loại kiểm tra thống kê nhất định, thông tin này thường dễ dàng tìm thấy bằng cách tìm kiếm trên internet.

Nhưng nếu điều này không thành công, việc truy vấn một người sử dụng ngôn ngữ lập trình cụ thể sẽ là phù hợp. ăn.

Đối với loại giao tiếp thông thường thứ hai, trong đó bạn có một số kết quả và bạn không chắc liệu chúng có đúng như những gì bạn mong đợi hay chúng không như những gì bạn mong đợi, bạn có thể sẽ được giúp đỡ nhiều nhất nếu tham gia nhiều hơn một lần. người và họ đại diện cho một loạt các quan điểm.

Các cuộc họp hiệu quả và hữu ích nhất thường bao gồm những người có chuyên môn về lĩnh vực nội dung và phân tích dữ liệu. Theo nguyên tắc chung, bạn giao tiếp với càng nhiều loại bên liên quan trong khi thực hiện dự án phân tích dữ liệu thì sản phẩm cuối cùng của bạn sẽ càng tốt hơn. Ví dụ: nếu bạn chỉ giao tiếp với các nhà phân tích dữ liệu khác, bạn có thể bỏ qua một số khía cạnh quan trọng trong quá trình phân tích dữ liệu của mình mà lẽ ra bạn có thể phát hiện ra nếu bạn giao tiếp với sếp, chuyên gia nội dung hoặc những người khác.

Đối với loại giao tiếp thông thường thứ ba, thường xảy ra khi bạn đến một nơi tự nhiên để tạm dừng phân tích dữ liệu của mình. Mặc dù khi nào và ở đâu trong quá trình phân tích dữ liệu của bạn, những lần tạm dừng này xảy ra tùy thuộc vào phân tích cụ thể mà bạn đang thực hiện, nhưng một nơi rất phổ biến để tạm dừng và kiểm tra lại là sau khi hoàn thành ít nhất một số phân tích dữ liệu thăm dò. Điều quan trọng là tạm dừng và yêu cầu phản hồi vào thời điểm này vì bài tập này thường sẽ xác định các phân tích khám phá bổ sung quan trọng đối với

hình thành các bước tiếp theo, chẳng hạn như xây dựng mô hình, và do đó ngăn cản bạn dành thời gian và công sức để theo đuổi các mô hình không liên quan, không phù hợp hoặc cả hai. Hình thức giao tiếp này hiệu quả nhất khi nó diễn ra dưới hình thức gặp mặt trực tiếp, nhưng hội nghị truyền hình và trò chuyện qua điện thoại cũng có thể hiệu quả. Khi chọn đối tượng của bạn, hãy nghĩ xem ai trong số những người có mặt với bạn sẽ đưa ra phản hồi hữu ích nhất và quan điểm nào sẽ quan trọng để cung cấp thông tin cho các bước phân tích tiếp theo của bạn. Ở mức tối thiểu, bạn nên có cả chuyên môn về phân tích dữ liệu và nội dung, nhưng trong cuộc họp kiểu này, cũng có thể hữu ích khi lắng nghe ý kiến của những người chia sẻ hoặc ít nhất là hiểu quan điểm của đối tượng mục tiêu lớn hơn để giao tiếp chính thức về kết quả phân tích dữ liệu của bạn.

10.3 Nội dung

Nguyên tắc hướng dẫn quan trọng nhất là điều chỉnh thông tin bạn cung cấp cho phù hợp với mục tiêu của giao tiếp. Đối với một câu hỏi nhằm mục tiêu nhằm mục đích làm rõ về mã hóa của một biến, người nhận thông tin liên lạc của bạn không cần biết mục tiêu tổng thể của phân tích của bạn, những gì bạn đã làm cho đến thời điểm này hoặc xem bất kỳ số liệu hoặc bảng nào. Một câu hỏi cụ thể, rõ ràng như "Tôi đang phân tích bộ dữ liệu tội phạm mà bạn đã gửi cho tôi tuần trước và đang xem xét biến "giáo dục" và thấy rằng nó được mã hóa 0, 1 và 2, nhưng tôi không xem bất kỳ nhãn nào cho các mã đó. Bạn có biết những mã này cho biến "giáo dục" đại diện cho cái gì không?"

Đối với loại giao tiếp thứ hai, trong đó bạn đang tìm kiếm phản hồi vì một vấn đề khó hiểu hoặc bất ngờ với phân tích của mình, sẽ có thêm thông tin cơ bản.

cần thiết, nhưng có thể không có đầy đủ thông tin cơ bản cho toàn bộ dự án. Để minh họa khái niệm này, hãy giả sử rằng bạn đã kiểm tra mối quan hệ giữa chiều cao và chức năng phổi và bạn xây dựng một biểu đồ phân tán, điều đó cho thấy rằng mối quan hệ là phi tuyến tính vì có đường như là độ cong cho mối quan hệ. mặc dù bạn có một số ý tưởng về cách tiếp cận để xử lý phi tuyến tính các mối quan hệ, bạn tìm kiếm thông tin đầu vào từ những người khác một cách thích hợp. Sau khi suy nghĩ một chút về mục tiêu giao tiếp, bạn quyết định hai mục tiêu chính: (1) Để hiểu liệu có cách tiếp cận tốt nhất để xử lý tính phi tuyến tính của mối quan hệ hay không và nếu có, cách xác định cách tiếp cận nào là tốt nhất, và (2) Để hiểu thêm về mối quan hệ phi tuyến tính mà bạn quan sát, bao gồm cả việc liệu điều này được mong đợi và/hoặc đã biết và liệu tính phi tuyến tính có quan trọng để nắm bắt trong các phân tích của bạn hay không.

Để đạt được mục tiêu của mình, bạn sẽ cần cung cấp khán giả với một số bối cảnh và nền tảng, nhưng cung cấp nền tảng toàn diện cho dự án phân tích dữ liệu và xem lại tất cả các bước bạn đã thực hiện cho đến nay là không cần thiết và có khả năng sẽ tiêu tốn thời gian và công sức sẽ tốt hơn dành cho các mục tiêu cụ thể của bạn. Trong ví dụ này, ngữ cảnh và bối cảnh phù hợp có thể bao gồm những điều sau đây:

(1) mục tiêu tổng thể của phân tích dữ liệu, (2) chiều cao như thế nào và chức năng phổi phù hợp với mục tiêu tổng thể của dữ liệu phân tích, ví dụ, chiều cao có thể là một yếu tố gây nhiễu tiềm năng, hoặc yếu tố dự đoán chính được quan tâm và (3) những gì bạn có được thực hiện cho đến nay liên quan đến chiều cao và chức năng phổi và những gì bạn đã học được. Bước cuối cùng này sẽ bao gồm một số hiển thị dữ liệu trực quan, chẳng hạn như biểu đồ phân tán đã nói ở trên. Nội dung cuối cùng của bài thuyết trình của bạn, sau đó, sẽ bao gồm một tuyên bố về các mục tiêu cho cuộc thảo luận, một bản tóm tắt tổng quan về dự án phân tích dữ liệu, vấn đề cụ thể như thế nào bạn đang phải đối mặt với sự phù hợp với dự án phân tích dữ liệu tổng thể và

rồi cuối cùng là những phát hiện thích hợp từ phân tích của bạn liên quan đến chiều cao và chức năng phổi.

Nếu bạn đang phát triển một bản trình bày trang chiếu, thì nên dành ít trang chiếu hơn cho nền và ngữ cảnh hơn là phần trình bày kết quả phân tích dữ liệu về chiều cao và chức năng phổi. Một slide là đủ cho tổng quan phân tích dữ liệu và 1-2 slide là đủ để giải thích bối cảnh của vấn đề chức năng chiều cao phổi trong dự án phân tích dữ liệu lớn hơn. Phần chính của bài thuyết trình không nên yêu cầu nhiều hơn 5-8 slide, do đó tổng thời gian thuyết trình không quá 10-15 phút. Mặc dù các slide chắc chắn là không cần thiết, nhưng một công cụ trực quan để trình bày thông tin này là rất hữu ích và không nên ám chỉ rằng bài thuyết trình phải "dành cho người xấu". Thay vào đó, ý tưởng là cung cấp cho nhóm đầy đủ thông tin để tạo ra cuộc thảo luận tập trung vào các mục tiêu của bạn, điều này đạt được tốt nhất bằng một bài thuyết trình thân mật.

Những nguyên tắc tương tự này áp dụng cho loại giao tiếp thứ ba, ngoại trừ việc bạn có thể không có các mục tiêu tập trung và thay vào đó, bạn có thể đang tìm kiếm phản hồi chung về dự án phân tích dữ liệu của mình từ khán giả. Nếu đây là trường hợp, mục tiêu tổng quát hơn này nên được nêu rõ và phần còn lại của nội dung nên bao gồm tuyên bố về câu hỏi mà bạn đang tìm cách trả lời bằng phân tích, (các) mục tiêu của phân tích dữ liệu, tóm tắt các đặc điểm của tập dữ liệu (nguồn dữ liệu, số lượng quan sát, v.v.), tóm tắt các phân tích khám phá của bạn, tóm tắt về xây dựng mô hình của bạn, diễn giải kết quả của bạn và kết luận. Bằng cách cung cấp các điểm chính từ toàn bộ phân tích dữ liệu của bạn, khán giả của bạn sẽ có thể cung cấp phản hồi về dự án tổng thể cũng như từng bước phân tích dữ liệu. Một cuộc thảo luận được lên kế hoạch tốt mang lại phản hồi hữu ích, chu đáo và nên được coi là thành

bạn được trang bị các tinh chỉnh bổ sung để thực hiện phân tích dữ liệu và quan điểm chu đáo về những gì nên được đưa vào phần trình bày chính thức hơn về kết quả cuối cùng của bạn cho khán giả bên ngoài.

10.4 Phong cách

Mặc dù phong cách giao tiếp ngày càng trang trọng từ kiểu giao tiếp thông thường thứ nhất đến kiểu thứ ba, nhưng tất cả các giao tiếp này phần lớn nên ở dạng không chính thức và có lẽ ngoại trừ giao tiếp tập trung về một vấn đề kỹ thuật nhỏ, nên tránh sử dụng biệt ngữ. Bởi vì mục đích chính của giao tiếp thông thường là để nhận phản hồi, phong cách giao tiếp của bạn nên khuyến khích thảo luận. Một số cách tiếp cận để khuyến khích thảo luận bao gồm nêu rõ trước rằng bạn muốn phần lớn cuộc họp bao gồm thảo luận tích cực và bạn hoan nghênh các câu hỏi trong suốt bài thuyết trình của mình thay vì yêu cầu khán giả giữ chúng cho đến khi kết thúc bài thuyết trình của bạn. Nếu một khán giả đưa ra bình luận, việc hỏi những người khác trong khán giả nghĩ gì cũng sẽ thúc đẩy thảo luận. Về bản chất, để có được phản hồi tốt nhất, bạn muốn nghe khán giả của mình đang nghĩ gì và điều này rất có thể đạt được bằng cách thiết lập một giọng điệu thân mật và tích cực khuyến khích thảo luận.

10.5 Thái độ

Thái độ phòng thủ hoặc phản đối có thể phá hỏng mọi công việc bạn đã bỏ ra để lựa chọn khán giả một cách cẩn thận, suy nghĩ về việc xác định đầy đủ các mục tiêu và chuẩn bị nội dung của mình, đồng thời tuyên bố rằng bạn đang tìm kiếm cuộc thảo luận. Khán giả của bạn sẽ miễn cưỡng đưa ra phản hồi mang tính xây dựng nếu họ

cảm thấy rằng phản hồi của họ sẽ không được đón nhận nồng nhiệt và bạn sẽ rời cuộc họp mà không đạt được mục tiêu của mình và không sẵn sàng thực hiện bất kỳ cài tiến hoặc bổ sung nào đối với phân tích dữ liệu của mình. Và khi đến lúc trình bày chính thức trước khán giả bên ngoài, bạn sẽ không chuẩn bị kỹ lưỡng và sẽ không thể trình bày tác phẩm tốt nhất của mình.

Để tránh cảm bối này, hãy cố tình trau dồi thái độ tiếp thu và tích cực trước khi giao tiếp bằng cách đặt cái tôi và sự bất an của bạn sang một bên. Nếu bạn có thể làm điều này thành công, nó sẽ phục vụ bạn tốt. Trên thực tế, cả hai chúng ta đều biết những người có sự nghiệp thành công phần lớn dựa trên thái độ tích cực và hoan nghênh của họ đối với phản hồi, bao gồm cả những lời chỉ trích mang tính xây dựng.

11. Suy nghĩ kết luận

Bây giờ bạn sẽ được trang bị một phương pháp mà bạn có thể áp dụng cho các phân tích dữ liệu của mình. Mặc dù mỗi bộ dữ liệu là sinh vật độc nhất của riêng nó và mỗi phân tích có các vấn đề cụ thể riêng cần giải quyết, nhưng việc giải quyết từng bước bằng khung ngoại luân rất hữu ích cho bất kỳ phân tích nào. Khi bạn làm việc thông qua việc phát triển câu hỏi, khám phá dữ liệu, lập mô hình dữ liệu, diễn giải kết quả và truyền đạt kết quả, hãy nhớ luôn đặt kỳ vọng và sau đó so sánh kết quả của hành động với kỳ vọng của bạn.

Nếu chúng không khớp, hãy xác định xem vấn đề là do kết quả của hành động hay kỳ vọng của bạn và khắc phục sự cố để chúng khớp với nhau. Nếu bạn không thể xác định vấn đề, hãy tìm ý kiến đóng góp từ những người khác và sau đó khi bạn đã khắc phục vấn đề, hãy chuyển sang hành động tiếp theo. Khung ngoại luân này sẽ giúp bạn tiếp tục lộ trình mà sẽ kết thúc bằng câu trả lời hữu ích cho câu hỏi của bạn.

Ngoài khuôn khổ ngoại luân, còn có các hoạt động phân tích dữ liệu mà chúng ta đã thảo luận xuyên suốt cuốn sách.

Mặc dù tất cả các hoạt động phân tích đều quan trọng, nhưng nếu chúng tôi phải xác định những hoạt động quan trọng nhất để đảm bảo rằng phân tích dữ liệu của bạn cung cấp câu trả lời hợp lệ, có ý nghĩa và dễ hiểu cho câu hỏi của bạn, chúng tôi sẽ bao gồm những điều sau:

1. Hãy suy nghĩ kỹ về việc phát triển câu hỏi của bạn và sử dụng câu hỏi để hướng dẫn bạn trong suốt tất cả các bước phân tích.
2. Thực hiện theo

ABC: 1. Luôn kiểm tra

2. Luôn thử thách 3.
Luôn giao tiếp

Cách tốt nhất để khung ngoại luân và các hoạt động này trở thành bản chất thứ hai là thực hiện nhiều phân tích dữ liệu, vì vậy chúng tôi khuyến khích bạn tận dụng các cơ hội phân tích dữ liệu đến với mình. Mặc dù với thực tế, nhiều nguyên tắc trong số này sẽ trở thành bản chất thứ hai đối với bạn, nhưng chúng tôi nhận thấy rằng việc xem lại các nguyên tắc này đã giúp giải quyết một loạt vấn đề mà chúng tôi gặp phải trong các phân tích của riêng mình. Do đó, chúng tôi hy vọng rằng cuốn sách sẽ tiếp tục phục vụ như một nguồn tài nguyên hữu ích sau khi bạn đọc xong khi bạn chạm trán với những trở ngại xảy ra trong mọi phân tích.

12. Về Tác Giả

Roger D. Peng là Giáo sư thống kê sinh học tại Trường Y tế Công cộng Johns Hopkins Bloomberg. Ông cũng là người đồng sáng lập của [Johns Hopkins Data Science Specialization1](#), đã đăng ký hơn 1,5 triệu sinh viên, [Johns Hopkins Executive Data Science Specialization2](#), blog [Thống kê đơn giản3](#) nơi anh ấy viết về thống kê và khoa học dữ liệu cho công chúng và Độ lệch chuẩn không quá chuẩn4 tệp âm thanh. Roger có thể được tìm thấy trên Twitter và GitHub dưới tên người dùng [@rdpeng5](#).

Elizabeth Matsui là Giáo sư Nhi khoa, Dịch tễ học và Khoa học Sức khỏe Môi trường tại Đại học Johns Hopkins và là nhà dịch tễ/miễn dịch học nhi khoa thực hành.

Cô chỉ đạo một trung tâm quản lý và phân tích dữ liệu với Dr. Peng hỗ trợ các nghiên cứu dịch tễ học và thử nghiệm lâm sàng, đồng thời là người đồng sáng lập [Skybrude Consulting, LLC6](#), một công ty tư vấn/khoa học dữ liệu. Elizabeth có thể được tìm thấy trên Twitter [@ eliza687](#).

¹<http://www.coursera.org/specialization/jhudatascience/>

²<https://www.coursera.org/specializations/executive-data-science>

³<http://simplystatistics.org/>

⁴<https://soundcloud.com/nssd-podcast>

⁵<https://twitter.com/rdpeng>

⁶<http://skybrudeconsulting.com>

⁷<https://twitter.com/eliza68>