CẨM NANG: R DÀNH CHO MỌI NGƯỜI

Le Duc Huy

2022-07-01

Contents

Chapter 1

Lời nói đầu

> Một ngôn ngữ lập trình tốt là ngôn ngữ giúp chúng ta có thể dễ dàng giao tiếp với máy và đặt được mục tiêu làm việc trong khoảng thời gian ngắn nhất.

Cuốn giáo trình hướng dẫn này dành cho tất cả những người đang thực hiện các công việc liên quan đến phân tích số liệu. Người học không cần phải có kiến thức về R trước đó vì các bài hướng dẫn sẽ đi từ những phần đơn giản nhất và kết hợp với các bài tập thực tiễn.

Hi vọng bạn đọc sẽ tìm thấy những điều thú vị và bổ ích từ cuốn cẩm nang hướng dẫn **R4ALl** này.

–Trân trọng!–

Tác giả

Chapter 2

Tại sao lại chọn R ?

2.1 Những lợi ích khi học R

Cái lợi đầu tiên nhất – Đó là nó hoàn toàn miễn phí. Nhiều người nói *"của cho là của ôi"* nhưng câu này hoàn toàn không đúng với R nhé. Của cho miễn phí hoàn toàn nhưng chất lượng thì miễn bàn. Vấn đề ở đây là bạn có đủ sức để khai thác hết tiềm năng của phần mềm này hay không.

Cộng đồng học R rất mạnh và hùng hậu trên thế giới – trong phân tích thống kê, có rất nhiều cộng đồng R ở các nước khác nhau. Nếu bạn có chút tiếng anh, hoặc không cần cũng được, cứ để anh google dịch lo bạn chỉ cần copy cái lỗi và tìm kiếm bài đăng. Bạn sẽ thấy hàng loạt bài đăng giải đáp thắc mắc trên mạng.

R đã rất phổ biến trong nghiên cứu khoa học.Nếu bạn đọc giả các tập san uy tín, bạn sẽ phát hiện rất nhiều bài báo trong số đó dùng R. Nói đúng ra, học R sẽ giúp các bạn giao tiếp ngôn ngữ thống kê và hội nhập nhanh hơn với các phương pháp phân tích nghiên cứu khoa học trên toàn thế giới.

Thêm một điểm nhỏ nữa, vì R cũng là một ngôn ngữ lập trình, nếu bạn thuần thục nó trong tương lai bạn sẽ tiếp thu các ngôn ngữ lập trình khác nhanh hơn như Python hay sử dụng syntax trong SPSS, STATA.

R cực kì mạnh trong thống kê so với các phần mềm khác như STATA, SPSS. Nhìn vào bảng so sánh này, bạn sẽ thấy R tiềm năng như thế nào?

2.2 Một số rào càn khi học R

Vậy có bạn sẽ hỏi tại sao, một phần mềm thống kê hay, xịn, miễn phí như R lại không phổ biến tại VN. Mình nghĩ có một số lý do sau: - Nhiều người quan

niệm, ngôn ngữ lập trình quá khó, khó đến mức không dám thử sức với nó. Nhưng tin mình đi nếu bạn đã học xong môn lập trình tin học cấp 2, R thật sự còn dễ hơn Pascal. Chỉ quan trọng là bạn có mở lòng ra với nó hay không. - Nhiều người đã quen với SPSS, STATA nên ngại học 1 ngôn ngữ mới. Tùy vào quan điểm mỗi người nhưng mình thấy R khắc phục được những hạn chế của 2 phần mềm trên: mở nhiều data cùng 1 lúc, xây dựng bản đồ tương tác, bản đồ phức tạp.

Nói tóm lại, khó hay không là do chính các bạn! Nếu các bạn tin mình, bạn sẽ cảm thấy mọi thứ chỉ khó khi mới bắt đầu dần dần phần mềm R sẽ đem lại nhiều điều thú vị. Hãy thử mở lòng và học hỏi thêm 1 phần mềm mới. Nó thật sự giúp các bạn nhiều lợi ích hơn những gì bạn nghĩ ^^ Nói đã khá nhiều, bây giờ nếu các bạn đã sẵn sàng hãy cùng bắt đầu hành trình chinh phục R.

Chapter 3

Hướng dẫn cài đặt phần mềm

Bài hướng dẫn này sẽ giúp anh chị tải và cài đặt phần mềm R. Về cơ bản, chúng ta cần cài đặt 2 phần mềm bao gồm **R base** và **R studio**.

- R base là phần mềm gốc xuất hiện đầu tiên, có thể thực hiện các câu lệnh phân tích. Có thể xem **R base** là một cái lõi.
- Tuy nhiên, để tăng hiệu suất làm việc với R, các nhà nghiên cứu đã thiết ra một phần mềm bổ sung với giao diện trực quan, hỗ trợ cú pháp dòng lệnh, đề xuất tên biến,... Đó là **R Studio**.

Vậy để tải và cài đặt phần mềm này, anh chị vui lòng thực hiện các bước cơ bản sau:

3.1 Tải phần mềm.

Để tải R base và R Studio anh chị vào địa chỉ trang web sau:

- R base: https://www.r-project.org/
- R Studio: https://www.rstudio.com/products/rstudio/download/#download

3.2 Cài đặt.

Anh chị lưu ý là **cần cài đặt R base trước và R studio sau.** Khi khởi động cài đặt, anh chị có thể chọn theo mặc định của phần mềm. Sau này nếu anh chị thấy đã thuần thục về R có thể cài đặt lại và thay đổi các tùy chọn cài đặt nâng cao của R base hay R Studio.

3.3 Khởi động phần mềm.

Lúc này chúng ta sẽ kiểm tra giao diện cụ thể.

!Một số lưu ý:

Khi tải phần mềm cần chọn đúng phiên bản để tương thích với hệ điều hành window bạn đang sử dụng (32x hay 64x). Để kiểm tra phiên bản window, anh chị bấm biểu tượng my computer, sau đó click chuột phải chọn properties để kiểm tra.

Chapter 4

GIAO DIỆN PHẦN MỀM

Nếu anh chị đã thấy được giao diện như trên hình (), tuyệt vời! Chúc mừng anh chị đã cài đặt thành công phần mềm R. Và bây giờ, em sẽ giới thiệu giao diện của R. Có thể xem R là phần cốt lõi còn Rstudio là phần vỏ giúp việc sử dụng R trở nên thân thiện và dễ dàng hơn. Do bài giảng của em xoay quanh sử dụng R studio nên em sẽ tập trung giới thiệu R studio: Như trên hình, mọi người có thể thấy R studio bao gồm 4 khu vực: 1. Khu vực ghi các dòng lệnh 2. Khu vực thực hiện câu lệnh và hiển thị kết quả 3. Khu vực môi trường làm việc bao gồm các đối tượng, biến số và list dữ liệu được nhập vào phân tích hay lưu trữ tạm 4. Khu vực cửa sổ chức năng bao gồm các cửa sổ phụ nhỏ theo các tab.

Hiện tại, các anh chị đã hiểu được các loại cửa sổ và giao diện làm việc với R. Bây giờ chúng ta sẽ cùng tìm hiểu kĩ hơn chức năng của từng cửa sổ: ## Cửa sổ dòng lệnh là nơi hiển thị của file R scripts hoặc file R mark down. Để tạo file R scripts hoặc file R markdown có một số cách sau: * Cách 1. Click vào biểu tượng New => Chọn file mới * Cách 2. Click vào File > chọn New > …

Vậy R scripts và R mark down (Rmd) có gì khác biệt nhau? Cả 2 dạng file này đều có chức năng lưu trữ câu lệnh. Tuy nhiên Rmd cung cấp một số tính năng giúp việc quản lý và thực hiện câu lệnh dễ dàng, rõ ràng hơn. * R md tiết kiệm số lần click chuột chạy dòng lệnh * Hiển thị kết quả theo từng nhóm câu lệnh. * Giúp xuất câu lệnh sang định dạng HTML và có khả năng xuất thành các tệp đầu ra khác (PDF, Word, Powerpoint, v.v.)

Chạy những câu lệnh đầu tiên

Giả sử chúng ta bấm chạy câu lệnh để in hello Vietnam. Đầu tiên chúng ta đưa con trỏ chuột đến dòng chữ “Hello Vietnam!” và bấm **Ctrl + Enter** hoặc bấm biểu tượng Run trên màn hình.

```
print("Hello Vietnam!")
```

```
## [1] "Hello Vietnam!"
```

4.1 Cửa sổ thực thi lệnh

Đây là nơi sẽ xuất hiện các dòng lệnh chúng ta chạy từ cửa sổ R scripts/ Rmd. Đối với Rmd, kết quả thường sẽ được biểu thị trong cửa sổ dòng lệnh.

Một lưu ý nhỏ *là biểu đồ sẽ không được hiển thị ở khu vực này, thay vào đó nó sẽ được hiện thị ở cảu sổ chức năng (khi chạy R scripts) hay cửa sổ dòng lệnh khi chạy R md.*

4.2 Khu vực môi trường làm việc

Đây là nơi hiển thị các đối tượng, tệp dữ liệu biến số được nhập hoặc tạo ra trong quá trình xử lý. Anh chị có thể click vào các tệp để hiển thị đối tượng được lưu trữ. Nhờ cửa sổ môi trường này, R cho phép người dùng làm việc với nhiều bộ dữ liệu cùng 1 lúc. Lưu ý: Khi bắt đầu làm việc một dự án mới, hãy xóa hết các thư mục và dữ liệu cũ trong môi trường để tránh nhầm lẫn và lỗi khi thực hiện phân tích dữ liệu. ## Khu vực cửa sổ chức năng Khu vực này giúp hiển thị một số nội dung sau:

- Hiển thị biểu đồ nếu thực hiện câu lệnh trên file R scripts
- Biểu thị tài liệu hướng dẫn sử dụng câu lệnh. Ví dụ để tìm hiểu 1 câu lệnh read.csv -> bấm `?Read.csv`, thông tin về cách sử dụng câu lệnh sẽ được hiện thị tại cửa số này.
- Tiếp cận với các file dữ liệu Rmd, R scripts, R data, csv …;
- Xác định các package được được cài đặt. ### Giới thiệu sơ qua về packages: Packages là tập hợp các câu lệnh được phát triển bởi các lập trình hoặc nghiên cứu trước đó nhằm giúp việc tiến hành phân tích nhanh, gọn và hiệu quả hơn. Để cài đặt packages có nhiều cách

```
install.packages("tidyverse") # Cài đặt package cho lần sử dụng đầu tiên
library(tidyverse) # Để nhập package mỗi khi cần sử dụng
```

4.3 Tóm tắt bài học

Bài học hôm nay giúp chúng ta hiểu được giao diện của R Studio gồm 4 cửa sổ:

Cửa sổ	Chức năng
Dòng lệnh	Viết và lưu trữ câu lệnh Hiển thị file R scripts hoặc R md
Kết quả	Hiển thị kết quả và các câu lệnh đã thực hiện
Môi trường làm việc	Hiển thị các biến số, đối tượng và list dữ liệu được tạo ra trong quá trình phân tích hay nhập vào R.
Chức năng	Hiển thị trợ giúp Biểu đồ Các file thư mục

Tài liệu tham khảo

- R dành cho khoa học dữ liệu (R for data science)
- Cẩm nang dịch tễ học với R

Chapter 5

Quy trình phân tích dữ liệu cơ bản với R

(Source: R for Data science)

Chào mừng mọi người đã đến với bài giảng tiếp theo trong chuỗi bài giảng chia sẻ **R4All**. Bài giảng này sẽ hướng dẫn cho mọi người một cái nhìn khái quát về quy trình phân tích dữ liệu nói chung và cách lồng ghép R vào quy trình này để đạt được mục tiêu phân tích.

Qua quá trình làm việc với nhiều bạn sinh viên cũng như các anh chị khi mới bắt đầu vào phân tích, mọi người thường không nắm rõ một quy trình phân tích cụ thể. Lý do chủ quan là do các anh chị và các bạn này mới bắt đầu vào phân tích, lý do khách quan là không có nhiều tài liệu đề cập đến vấn đề này dù rằng đây là một vấn đề rất cơ bản và quan trọng cho bất cứ cá nhân nào thực hiện phân tích dữ liệu.

Do vậy sau khi đọc xong bài giảng này, anh chị và các bạn sẽ hiểu được một quy trình cơ bản trong phân tích dữ liệu từ đó đặt ra mục tiêu để có những chiến thuật phân tích phù hợp. Theo kinh nghiệm bản thân và từ các tài liệu tham khảo, phân tích dữ liệu thường trải qua các bước sau: * Bước 1. Xác định câu hỏi phân tích/ mục tiêu phân tích. * Bước 2. Thiết lập môi trường quản lý phân tích dữ liệu * Bước 3. Chuẩn bị và nhập dữ liệu vào R * Bước 4. Kiểm tra và làm sạch dữ liệu * Bước 5. Xuất dữ liệu đã làm sạch * Bước 6: Phân tích mô tả

5.1 Xác định câu hỏi phân tích (mục tiêu phân tích)

Việc xác định mục tiêu phân tích cực kì quan trọng vì nó giúp anh chị tập trung hơn và tránh lạc hướng trong quá trình xử lý dữ liệu. Bên cạnh đó việc xác định mục tiêu phân tích cũng giúp anh chị xác định các phương pháp thống kê cần thiết cho phân tích dữ liệu và các nguồn số liệu cần thiết. Qua đó giúp chúng ta tiết kiệm thời gian và nguồn lực phân tích. Trong quá trình xác định mục tiêu phân tích, chúng ta cũng có thể tham khảo ý kiến góp ý từ những người xung quang, các giáo viên hướng dẫn, anh chị và bạn bè. Mục tiêu phân tích có thể được xây dựng dựa vào mục tiêu nghiên cứu, khung lý thuyết trong nghiên cứu. Trong mục tiêu phân tích chúng ta cần xác định được cơ bản các yếu tố sau: Cấp độ phân tích: mô tả, xây dựng mô hình, tìm kiếm mối quan hệ (Tìm hiểu các cấp độ phân tích trong nghiên cứu…)

- Xác định các biến số
- Biến số chính hay kết quả (Biến số phụ thuộc) và các biến số liên quan. Ví dụ: Phân tích số lượng ca mắc COVID-19 và các yếu tố liên quan ở các quốc gia châu Á.
- Từ mục tiêu trên, chúng ta có thể xác định được số lượng ca mắc COVID-19 là biến số chính cần quan tâm. Các biến số liên quan gồm: danh sách tên các quốc gia thuộc châu Á, Các yếu tố ảnh hưởng đến số lượng ca mắc: chính sách phòng dịch, tỷ lệ bao phủ vắc xin, tỷ lệ xét nghiệm COVID-19, đặc điểm sức khỏe dân số của quốc gia đó, …
- Cấp độ phân tích ở đây là mô tả và xây dựng mô hình để phát hiện các yếu tố liên quan.

5.2 Thiết lập môi trường quản lý phân tích dữ liệu

Nhiều người khi mới bắt đầu sử dụng có xu hướng bắt tay vào phân tích ngay lập tức và không chuẩn bị môi trường quản lý dữ liệu. Điều này dẫn đến tác hại lâu dài trong việc quản lý phân tích dữ liệu và làm tốn rất nhiều thời gian để truy tìm file sau này vì bị lạc mất file, xóa nhầm, quên mất phiên bản phân tích mới nhất, gặp trở ngại khi chia sẻ project với những người xung quanh… Do vậy hay dành chút thời gian thiết lập môi trường và tạo thói quen trước khi bắt đầu các bước phân tích dữ liệu đầu tiên. Đầu tiên chúng ta cần tạo một thư mục và đặt tên thư mục theo tên dự án ngắn gọn. Trong thư mục này, chúng ta sẽ tạo ra các thư mục con gồm: dataset, scripts (code), results, plots. Mọi người có thể tham khảo hình bên dưới để biết rõ hơn về cách xây dựng hệ thống thư mục trong project.