

R for babies

Thinh Ong

2024-09-02

Table of contents

Tổng quan	3
1 Giới thiệu	4
1.1 R	4
1.1.1 Ngôn ngữ lập trình	5
1.1.2 R và SPSS, Stata, SAS	6
1.2 Tại sao học R?	7
1.3 R và RStudio	8
1.3.1 Cài đặt R	10
1.3.2 Cài đặt RStudio	11
2 Chuẩn bị làm việc	12
2.1 File data đang ở đâu?	12
2.1.1 Windows	12
2.1.2 MacOS	14
2.1.3 Absolute path và relative path	14
2.2 R project	14
2.3 R packages	17
2.4 Đọc data vào R	17
2.5 Data gọn gàng (tidy data)	17
2.6 R code và R markdown	17
References	18

Tổng quan

Bài giảng được tham khảo từ các tài liệu sau:

Beckerman, A. P., Childs, D. Z., & Petchey, O. L. (2017). Getting started with R: an introduction for biologists. Oxford University Press.

1 Giới thiệu

1.1 R

R là một ngôn ngữ lập trình được phát triển bởi GS. Robert Gentleman và GS. Ross Ihaka tại Đại học Auckland¹. Tên gọi R được đặt theo tên của 2 tác giả (Robert và Ross).



Robert Gentleman: “Let’s write some software.”

Ross Ihaka: “Sure, that sounds like fun.”²

1.1.1 Ngôn ngữ lập trình

Ngôn ngữ lập trình là một tập hợp các hướng dẫn để yêu cầu máy tính thực hiện một số tác vụ nhất định.³

Ngôn ngữ là phương tiện để con người giao tiếp với con người. Ngôn ngữ lập trình là phương tiện để con người giao tiếp với máy tính.³ Vì vậy, học ngôn ngữ lập trình cũng giống như học ngoại ngữ, bao gồm từ vựng, ngữ pháp, cụm từ, mệnh đề... để viết thành một câu văn mà máy tính có thể hiểu được và làm đúng những gì con người muốn.



Một số ngôn ngữ lập trình trong phân tích dữ liệu [[Photo credit](#)]

1.1.2 R và SPSS, Stata, SAS



- R là ngôn ngữ lập trình. Người dùng giao tiếp, đối thoại với máy tính bằng cách nhập những câu văn (code) giống như chat với máy tính, để máy hiểu và làm đúng những gì con người muốn.
- SPSS, Stata, SAS là các gói phần mềm thống kê (software package) thương mại, được thiết kế giao diện người dùng kéo thả, click chọn để dễ dàng giao tiếp với máy tính hơn. Người dùng cũng có thể viết code (SPSS Syntax, Stata command, SAS program) để lưu lại các bước phân tích, nhưng đây không phải mục tiêu chính của các gói thương mại này. Người dùng không thể yêu cầu máy tính làm gì khác với những chức năng đã được quy định sẵn trong gói phần mềm. Mỗi gói phần mềm được viết bằng một ngôn ngữ lập trình:
 - SPSS: Java
 - Stata: C
 - SAS: C

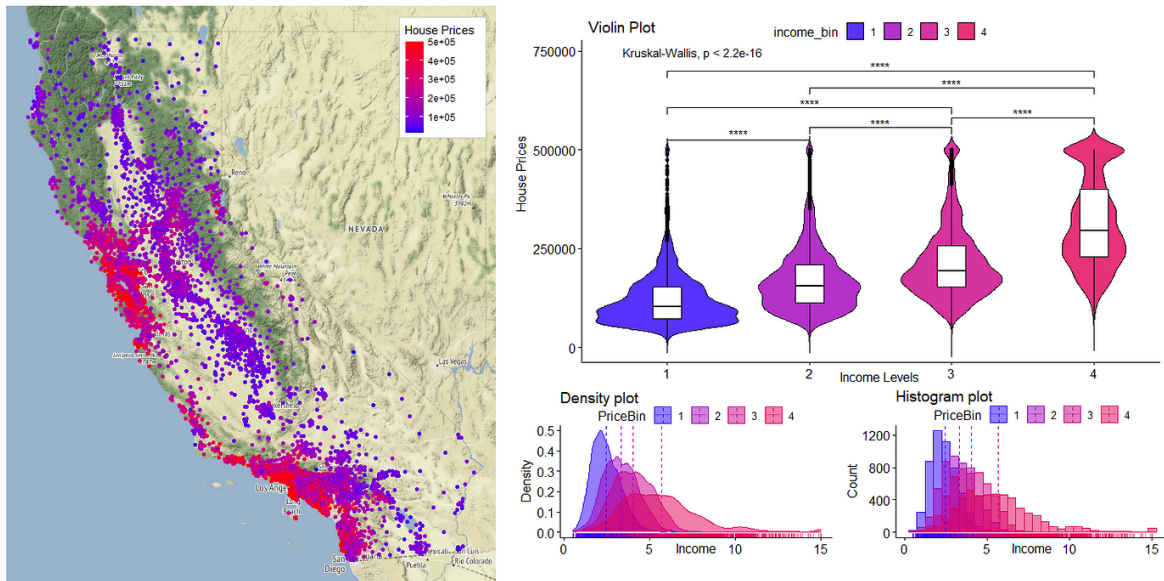
1.2 Tại sao học R?

1. Miễn phí và sử dụng được trên mọi hệ điều hành thông dụng (Windows, Macs, Linux).
2. R là ngôn ngữ truyền thống cho phân tích dữ liệu trong nghiên cứu y sinh (biostatistics), tin sinh (bioinformatics), dịch tễ, mô hình dự báo... Chuyên gia ở các ngành này liên tục phát triển các gói phần mềm (packages) viết bằng R cập nhật các phương pháp mới nhất, machine learning, thiết kế web...



Các gói phần mềm phổ biến trong R [Photo credit]

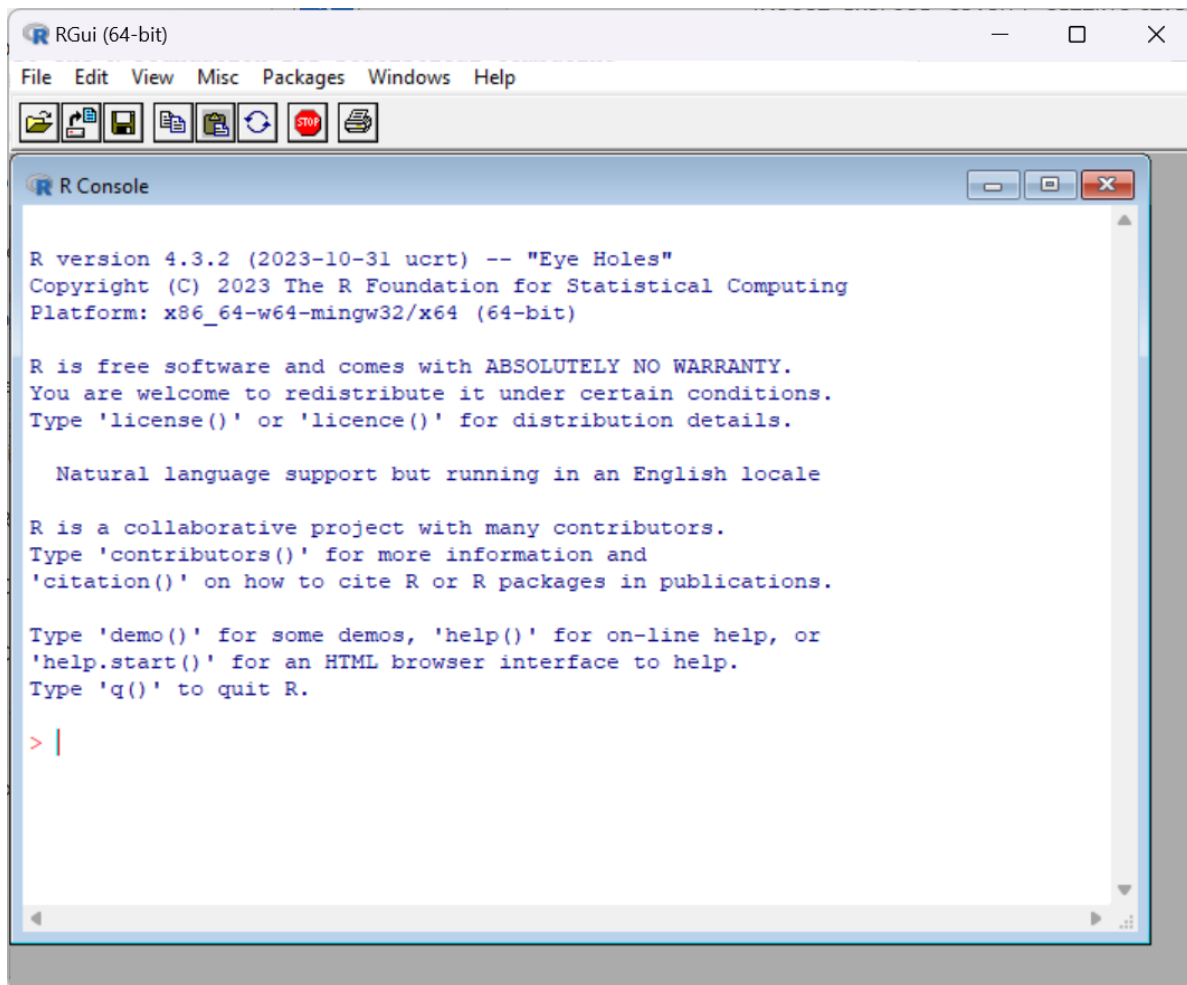
3. Vẽ biểu đồ chất lượng cao



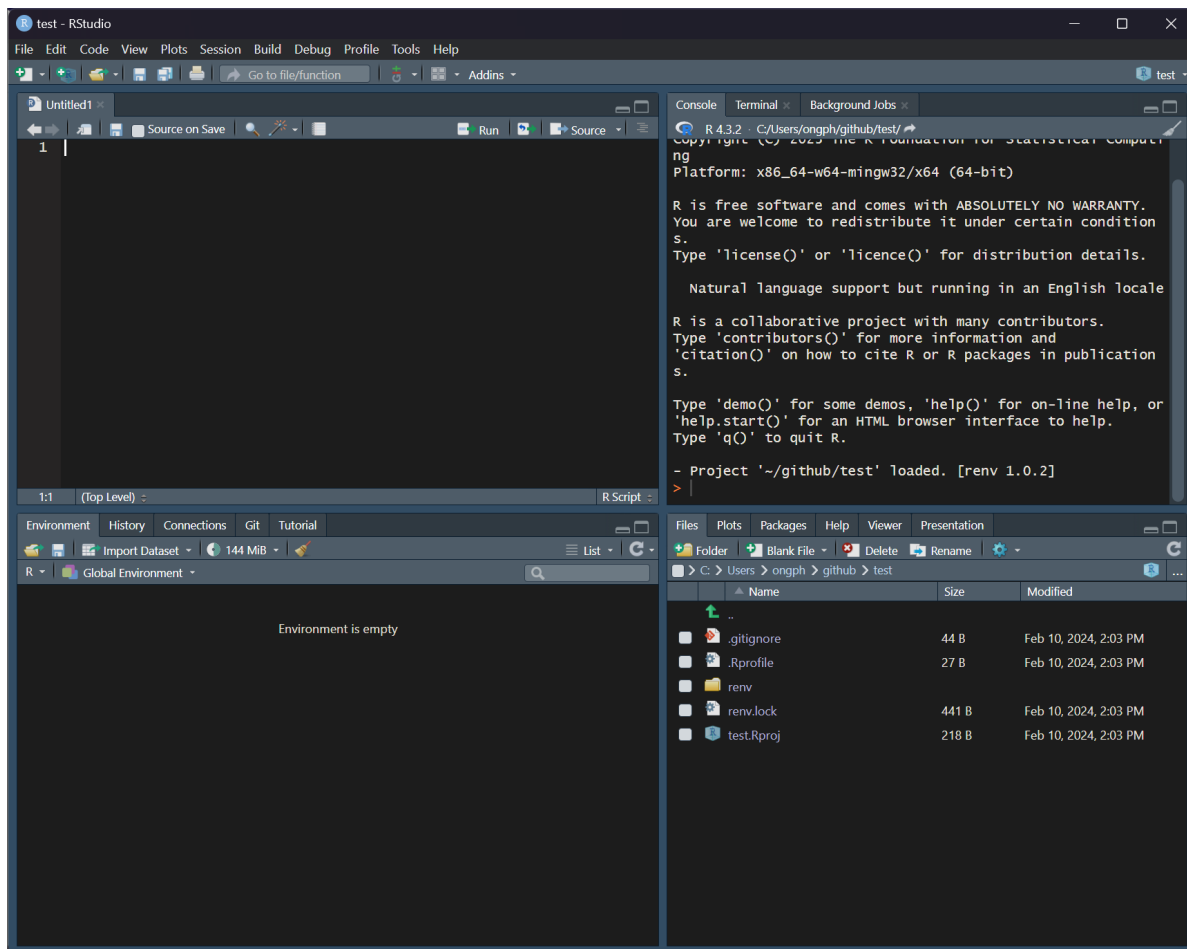
Một số biểu đồ vẽ bằng R [[Photo credit](#)]

1.3 R và RStudio

R là ngôn ngữ lập trình. Sau khi cài R, chúng ta mở lên sẽ thấy giao diện giống như một khung chat trống. Khung chat này là nơi chúng ta viết code để giao tiếp với máy tính.



RStudio là một môi trường phát triển tích hợp (integrated development environment hay IDE) hay nói đơn giản là một phần mềm để viết code R hiệu quả hơn.



Vì vậy, chúng ta cần cài đặt riêng R (ngôn ngữ lập trình) và RStudio (IDE).

1.3.1 Cài đặt R

Truy cập <https://cran.r-project.org/> và tải R cho hệ điều hành của mình.

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) ([Debian](#), [Fedora/Redhat](#), [Ubuntu](#))
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

1.3.2 Cài đặt RStudio

Truy cập <https://posit.co/download/rstudio-desktop/> và tải RStudio cho hệ điều hành của mình.

OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2023.12.1-402.EXE ↓	215.66 MB	D3C03C42
macOS 12+	RSTUDIO-2023.12.1-402.DMG ↓	382.66 MB	C8D9185D
Ubuntu 20/Debian 11	RSTUDIO-2023.12.1-402-AMD64.DEB ↓	149.27 MB	81F221BE
Ubuntu 22/Debian 12	RSTUDIO-2023.12.1-402-AMD64.DEB ↓	149.96 MB	75542CC2

2 Chuẩn bị làm việc

💡 Mục tiêu

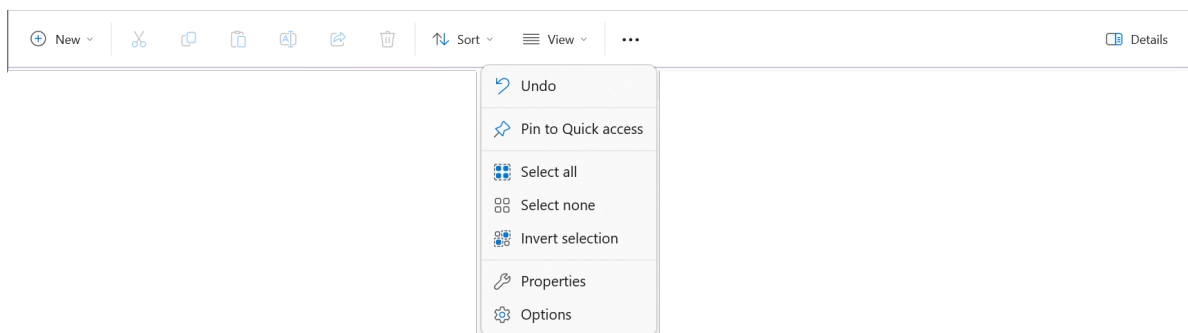
1. Biết cách tìm path tới nơi lưu file data
2. Biết cách cài đặt và gọi các packages trong R
3. Biết cách đọc data vào R
4. Hiểu cấu trúc tidy data

2.1 File data đang ở đâu?

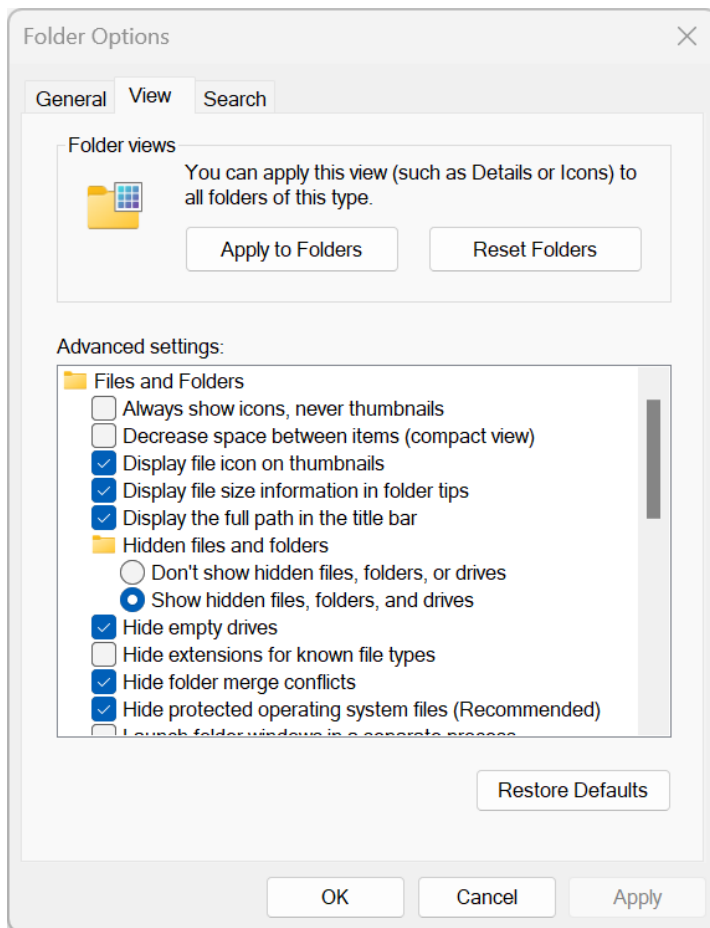
Bước đầu tiên chúng ta phải biết file data đang nằm ở đâu trong máy tính. Vị trí, hay địa chỉ, hay đường dẫn lưu file data trong máy tính gọi là *path*.

2.1.1 Windows

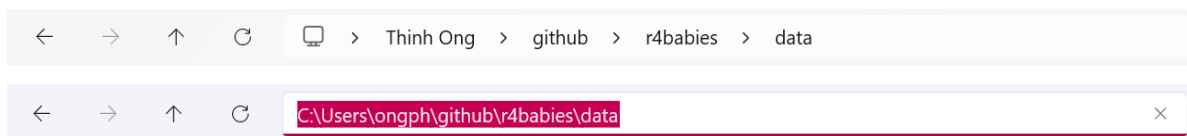
Để Windows hiển thị đường dẫn trong Explorer, trên thanh menu chọn dấu ... > **Options**.



Chọn tab View, tick vào ô `Display the full path in the title bar`.



Sau khi làm các bước trên, click vào title bar của Explorer sẽ thấy hiển thị đầy đủ path của file hoặc folder hiện tại như sau:



! Lưu ý

Trong Windows, khi copy path vào R cần sửa lại theo 1 trong 2 cách sau:

1. Sửa tất cả dấu \ thành /:

C:/Users/ongph/github/r4babies/data

2. Sửa tất cả dấu \ thành \\:

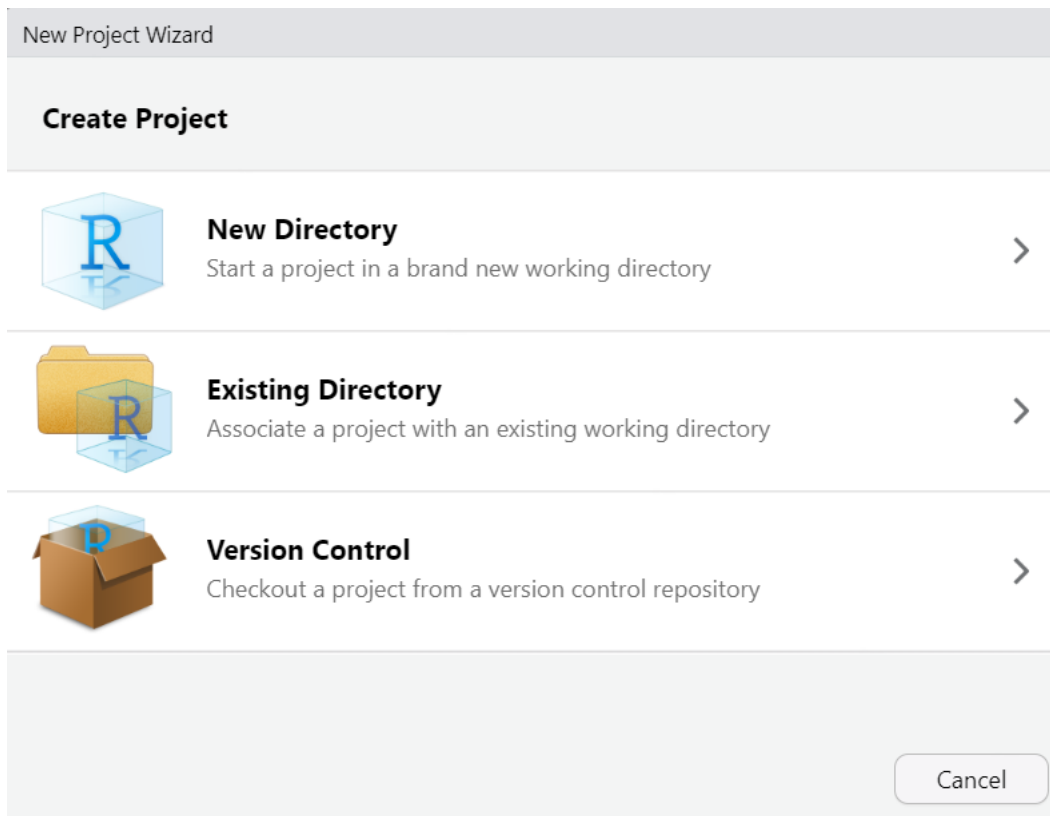
```
C:\\Users\\ongph\\github\\r4babies\\data
```

2.1.2 MacOS

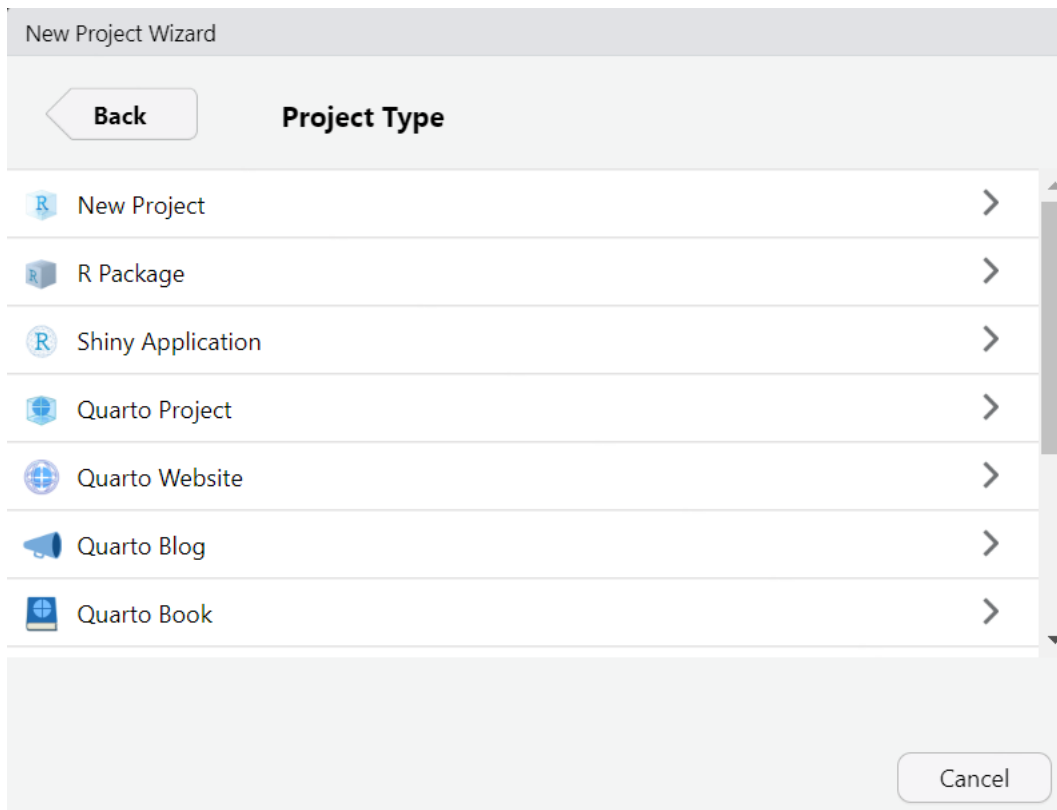
2.1.3 Absolute path và relative path

2.2 R project

Trên thanh menu vào **File > New Project... > New Directory**.




Chọn Project Type là **New Project**.



Đặt tên folder cho project này, và chọn đường dẫn nơi sẽ lưu project trong máy tính.

New Project Wizard

[Back](#) **Create New Project**



Directory name:

Create project as subdirectory of:
 [Browse...](#)

☒ Create a git repository
☒ Use renv with this project

☐ Open in new session

[Create Project](#) [Cancel](#)

Trong folder `/test` là một R project mới tạo này, chúng ta tạo thêm 2 folders `/data` và `/analysis`. Các thành phần trong folder này như sau:

```
test
|-- data
|   |-- data.xlsx
|-- analysis
|   |-- code.R
|   |-- analysis.Rmd
|-- test.Rproj
```

- Folder `/data`: chứa tất cả data
- Folder `/analysis`: chứa tất cả file code, markdown

Mục đích của việc chuẩn bị này là

```
getwd()
```

```
[1] "C:/Users/ongph/github/r4babies"
```


2.3 R packages

```
install.packages("readxl")
```

! Lưu ý

Khi cài đặt, tên package phải nằm trong dấu ngoặc kép ""

Sau khi cài đặt thì package sẽ trở thành một “thư viện” trong R. Để gọi thư viện này, dùng:

```
library(readxl)
```

! Lưu ý

Khi gọi library thì tên library không nằm trong dấu ngoặc kép nữa ""

2.4 Đọc data vào R

2.5 Data gọn gàng (tidy data)

Tidy data⁴

2.6 R code và R markdown

Lưu lại các đoạn hội thoại giữa chúng ta và R.

References

1. Ihaka, R. & Gentleman, R. [R: A Language for Data Analysis and Graphics](#). *Journal of Computational and Graphical Statistics* **5**, 299–314 (1996).
2. Ihaka, R. [The R Project: A Brief History and Thoughts About the Future](#). (1998).
3. Lenovo, Inc. [Programming Language: What is a programming language?](#) (2024).
4. Wickham, H. [Tidy Data](#). *Journal of Statistical Software* **59**, (2014).