

Hướng dẫn thực hành & phân tích số liệu với phần mềm



ThS. BS. Lê Đức Huy

R Trước khi bắt đầu

- **Link cuộc họp:** <https://meet.google.com/xfo-fkkw-qac>
- Để tiếp cận với **bộ số liệu** và các **tài liệu khóa học**, sử dụng link sau đây: [Link](#).
- Trong quá trình học, khuyến khích người học thực hiện đầy đủ các câu hỏi và bài tập:
 - Link truy cập bài tập thảo luận: [online doc link](#).
 - Cài đặt Viewer để dễ dàng hỗ trợ trực tuyến. (Click vào [Link](#))

R Cài đặt phần mềm R và RStudio

- **B1. Tải và cài đặt R base:**

- Đối với Window: <https://cran.r-project.org/bin/windows/base/>
- Đối với MacOS: <https://cran.r-project.org/bin/macosx/>

- **B2. Tải và cài đặt Rstudio:**

<https://support--rstudio-com.netlify.app/products/rstudio/download/#download>

- Để xem hướng dẫn chi tiết cách tải phần mềm có thể tham khảo video sau:
<https://www.youtube.com/watch?v=NZxSA80IF1I>

Lưu ý:

- Chọn đúng phiên bản phần mềm để tương thích với window 32x hay 64x
- Tải và cài đặt R base trước khi cài đặt R studio.
- Trong một số trường hợp cần cập nhật **Java** để tránh bị lỗi trong khi cài đặt.

Bài 1: Giới thiệu về phần mềm R





Anh/ chị đã từng sử dụng những phần mềm thống kê nào?

R Nội dung bài học

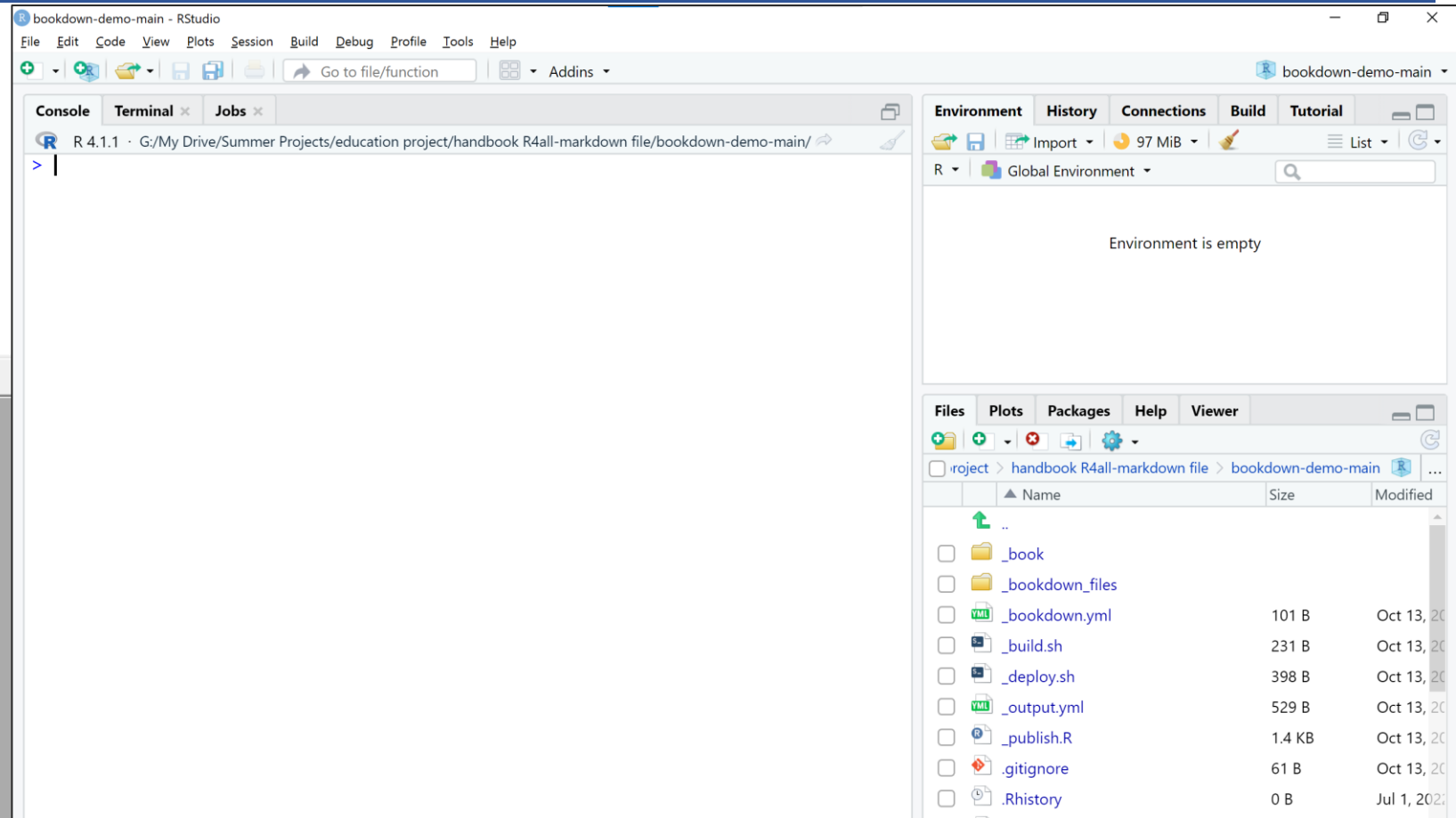
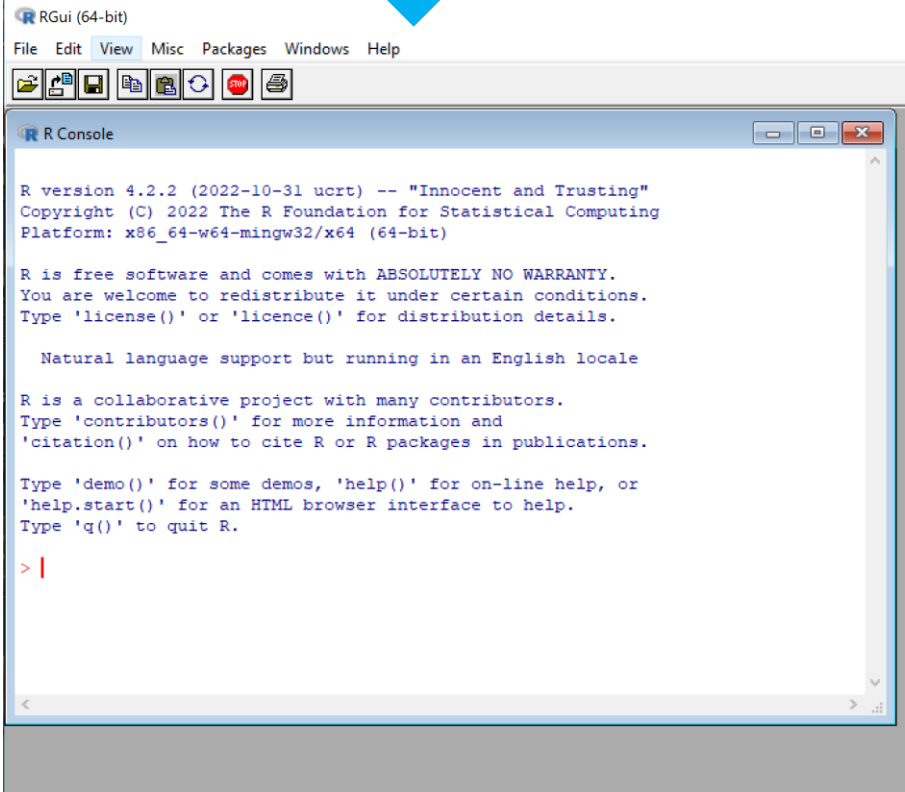
1. Tại sao cần học R?
2. Giao diện của phần mềm R và **R studio**
 1. Kí tự viết code và mẫu cấu trúc câu lệnh cơ bản
 2. Các loại dữ liệu và cấu trúc dữ liệu trong R
 3. Giới thiệu một số packages và libraries

R Tại sao cần học R?

- Hoàn toàn miễn phí;
- Rất phổ biến;
- Dễ học và trực quan với những người đã có kinh nghiệm phân tích;
- Làm việc với nhiều bộ dữ liệu khác nhau cùng 1 thời điểm (SPSS, STATA chỉ làm việc với 1 bộ dữ liệu)
- Cấu trúc câu lệnh linh hoạt -> **Giúp tái sử dụng** (tái thực hiện câu lệnh) dễ dàng;
- Tính kế thừa và nâng cao hiệu suất qua thời gian – sử dụng các package được xây dựng sẵn.

R So sánh giao diện R base và R studio

R base



R studio



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Untitled1

Source on Save Run Source

1

Các cửa sổ trong R

1 Khu vực ghi dòng lệnh

2 Khu vực thực hiện lệnh

3 Khu vực môi trường làm việc

4 Khu vực cửa sổ chức năng

Environment History Connections Build Tutorial

Import 97 MiB

R Global Environment

Environment is empty

Files Plots Packages Help Viewer

project > handbook R4all-markdown file > bookdown-demo-main

	Name	Size	Modified
	..		
	_book		
	_bookdown_files		
	_bookdown.yml	101 B	Oct 13, 2020
	_build.sh	231 B	Oct 13, 2020
	_deploy.sh	398 B	Oct 13, 2020
	_output.yml	529 B	Oct 13, 2020
	_publish.R	1.4 KB	Oct 13, 2020
	.gitignore	61 B	Oct 13, 2020
	.Rhistory	0 B	Jul 1, 2020
	.travis.yml	209 B	Oct 13, 2020

Console Terminal Jobs

R 4.1.1 · G:/My Drive/Summer Projects/education project/handbook R4all-markdown file/bookdown-demo-main/

R Giao diện của phần mềm R

Khu vực ghi dòng lệnh

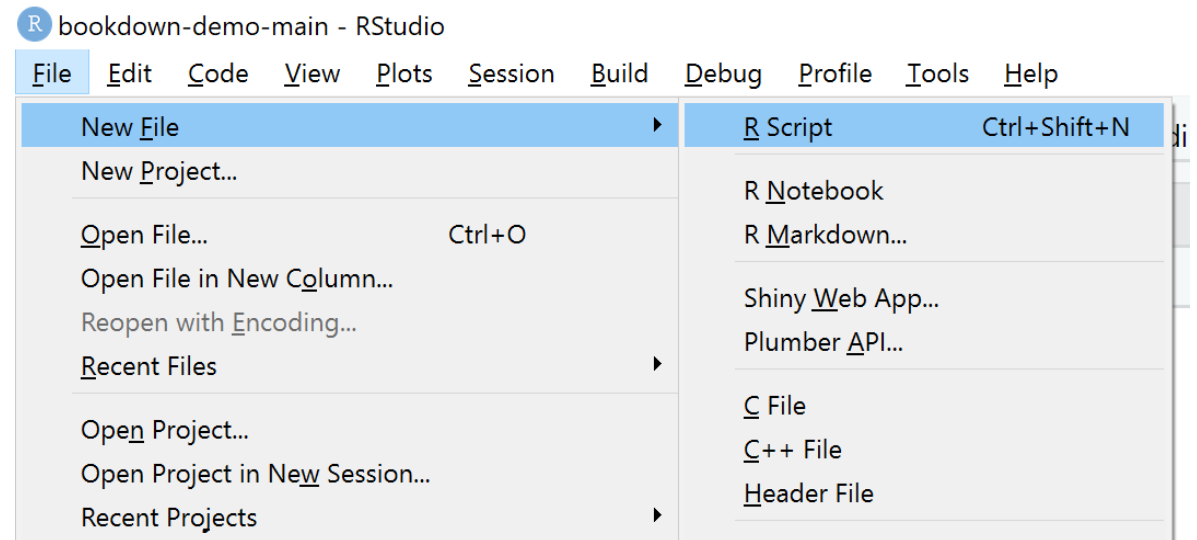
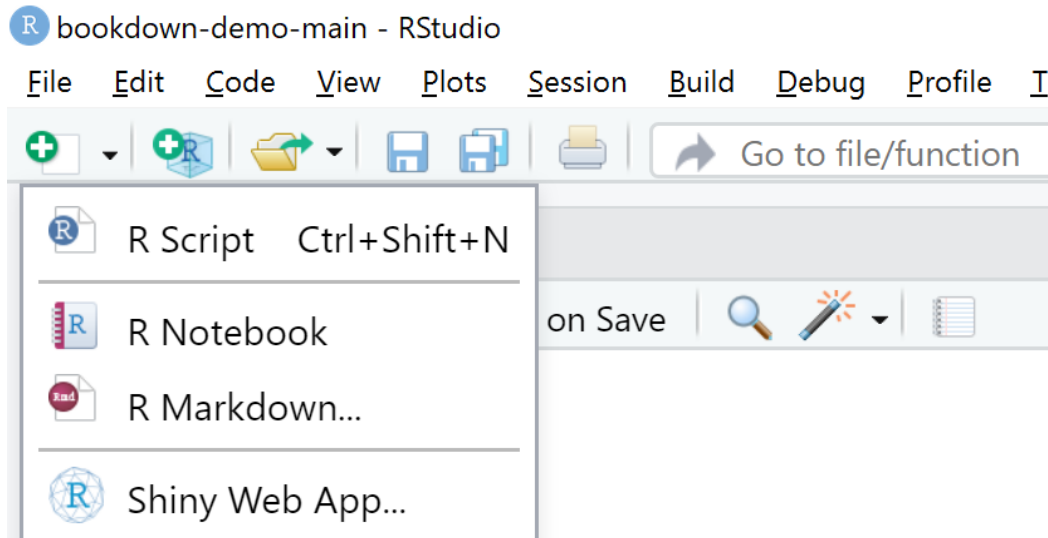
- Đây là nơi hiển thị của file **R scripts** hoặc file **R mark down (Rmd)**. Cả 2 dạng file này đều có **chức năng lưu trữ câu lệnh**. Tuy nhiên Rmd cung cấp một số tính năng giúp việc quản lý và thực hiện câu lệnh dễ dàng hơn ví dụ như:
 - Rmd tiết kiệm số lần click chuột chạy dòng lệnh
 - Rmd hiển thị kết quả ngay dưới đoạn 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.)

R Giao diện của phần mềm R

Khu vực ghi dòng lệnh

Tạo file lưu trữ câu lệnh trong R

- 2 dạng file lưu trữ file phổ biến nhất là R script và R markdown



R Giao diện của phần mềm R

Khu vực ghi dòng lệnh

Một số kí tự đặc biệt trong viết câu lệnh trong R:

- Dấu gán giá trị: "<-"
- Dấu tiếp cận giá trị: "\$"
- Dấu kết nối các câu lệnh phân tích (trong package dplyr): "%>%"
- Dấu bình luận/ chú thích: "#"

```
2 # Bai 1
3 ## VD1:
4 ```{r}
5 # Một số câu lệnh cơ bản
6 rm(list=ls())
7 k <- 5
8 data <- data.frame(x = c("a", "b", "b", "a", "c"),
9                    y = c(1,2,3,2,3))
10
11 data$x
12 data1 <- data %>% filter(x != "a") %>% mutate(new_column = "z")
13
```

R Giao diện của phần mềm R

Khu vực ghi dòng lệnh

Cấu trúc chung của một câu lệnh trong R

$$[A] <- [B]$$

A: Có thể là một bảng số liệu, một cột dữ liệu hay chỉ là một biến số để lưu trữ giá trị của [B] sau quá trình phân tích.

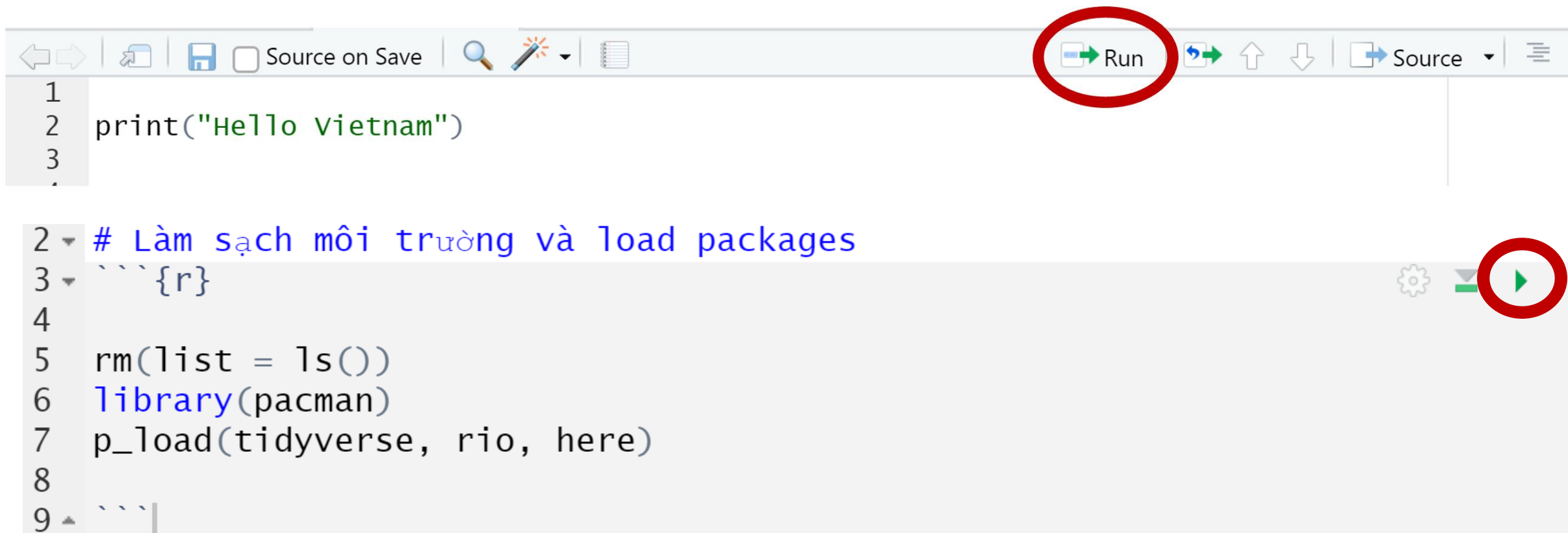
B: Có thể là một bảng số liệu hoặc một chuỗi giá trị được phân tích và tính toán bằng nhiều câu lệnh khác nhau.

```
12 data1 <- data %>% filter(x != "a") %>% mutate(new_column = "z")
```

R Giao diện của phần mềm R

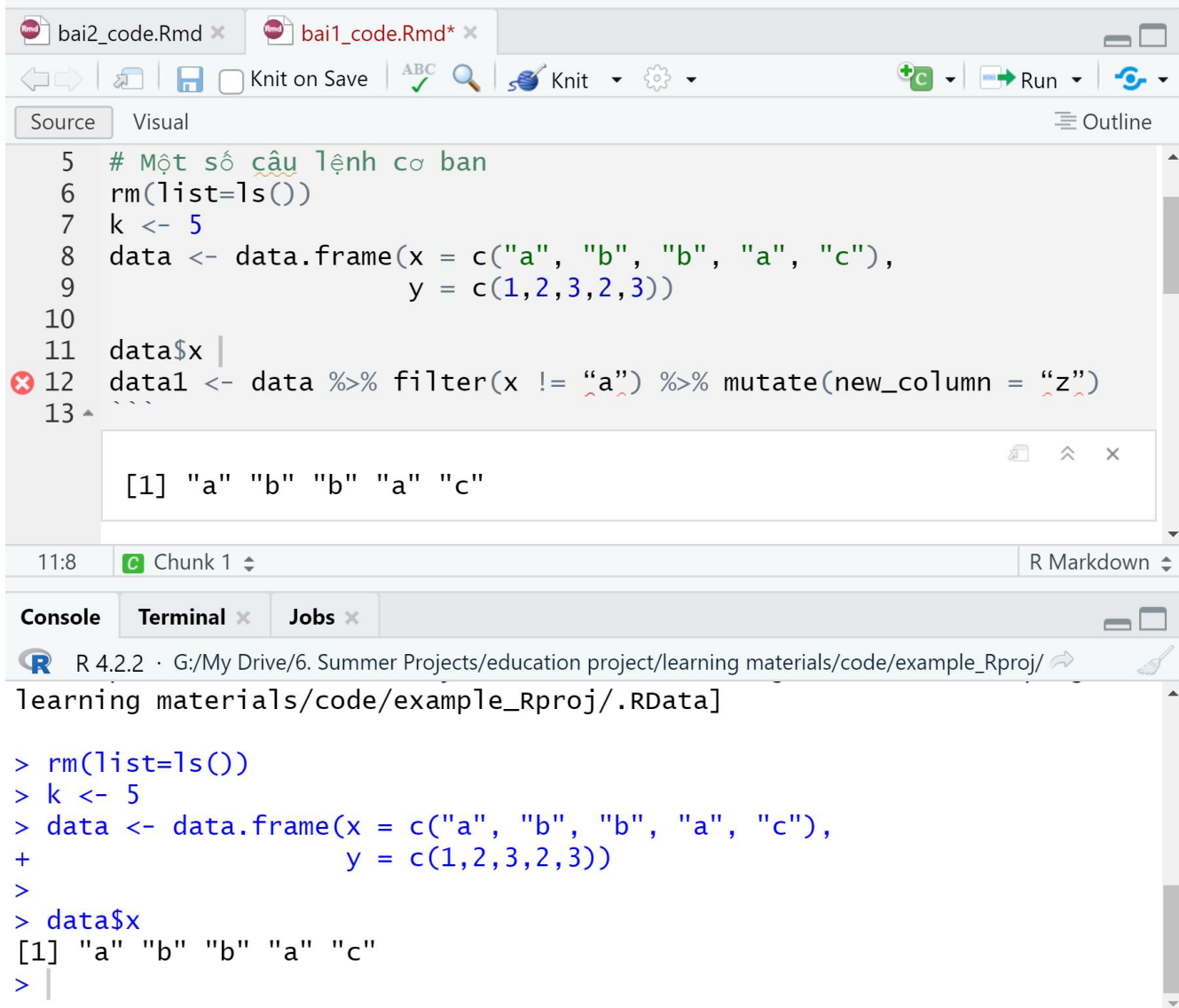
Khu vực ghi dòng lệnh

- Để thực hiện câu lệnh: đưa con trỏ chuột đến dòng lệnh cần chạy, bấm nút “Run” trên thanh toolbar hoặc bấm **Ctrl+Enter**.



R Giao diện của phần mềm R

Khu vực thực thi dòng lệnh



The screenshot shows the RStudio interface. The top pane (Source editor) displays R code in a file named 'bai2_code.Rmd'. The code includes comments in Vietnamese and R commands to create a data frame and filter it. The bottom pane (Console) shows the output of the code execution, including the creation of the data frame and the filtered result.

```
5 # Một số câu lệnh cơ bản
6 rm(list=ls())
7 k <- 5
8 data <- data.frame(x = c("a", "b", "b", "a", "c"),
9                   y = c(1,2,3,2,3))
10
11 data$x |
12 data1 <- data %>% filter(x != "a") %>% mutate(new_column = "z")
13
```

Output in the Console:

```
> rm(list=ls())
> k <- 5
> data <- data.frame(x = c("a", "b", "b", "a", "c"),
+                   y = c(1,2,3,2,3))
>
> data$x
[1] "a" "b" "b" "a" "c"
```

- 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 hiển thị ở cửa sổ chức năng (khi chạy R scripts) hay cửa sổ dòng lệnh khi chạy Rmd.

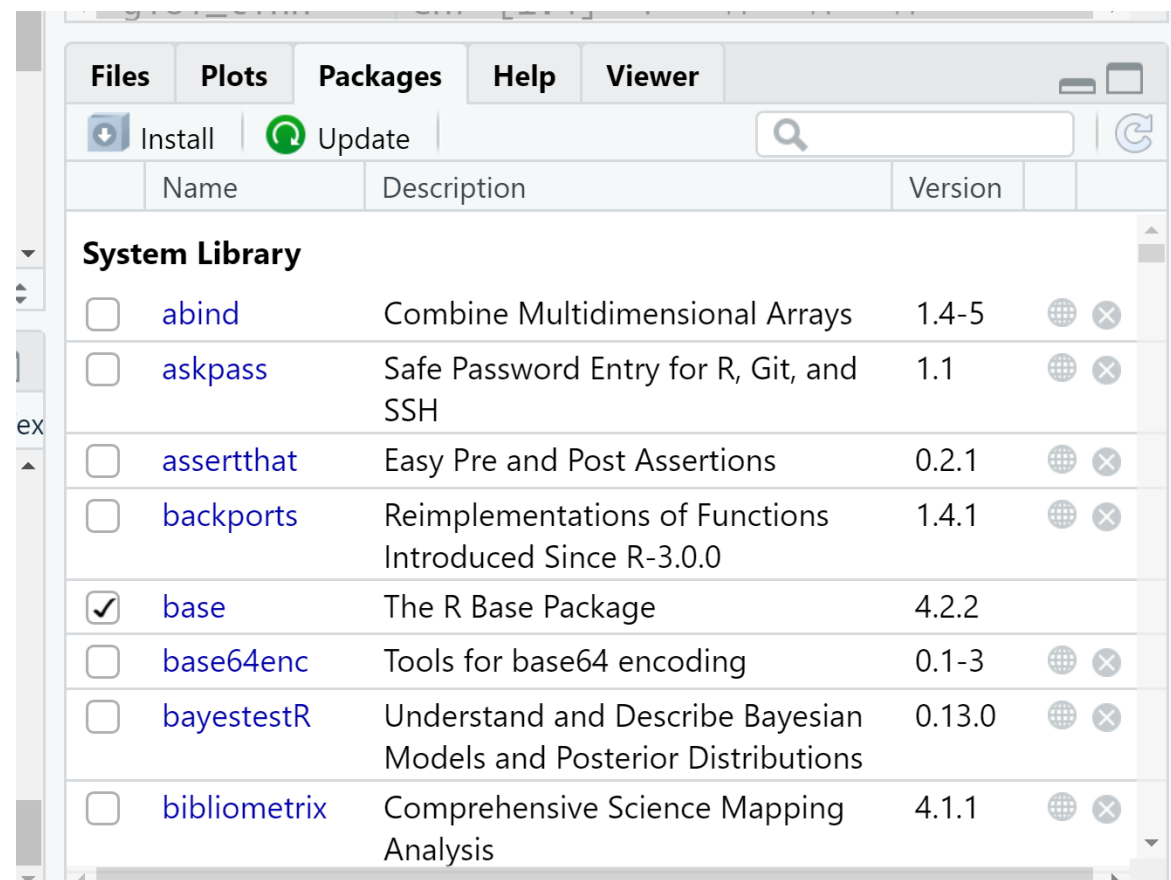
R Thực hành

- Tạo file R script và R markdown
- Chạy câu lệnh đầu tiên trong cửa sổ dòng lệnh dựa vào assignment 1 trong file discussion.

R Giao diện của phần mềm R

- Hiển thị biểu đồ nếu thực hiện câu lệnh trên file R scripts (VD2)
- Biểu thị tài liệu hướng dẫn sử dụng câu lệnh. Ví dụ để tìm hiểu chức năng lệnh print chúng ta sẽ gõ và chạy câu lệnh sau `?print` ở cửa sổ câu lệnh hay cửa sổ thực thi lệnh. Thông tin về cách sử dụng câu lệnh sẽ được hiển thị tại cửa sổ này. (VD2)
- Tiếp cận với các file dữ liệu Rmd, R scripts, R data, csv ...;
- Xác định các **packages** được cài đặt.

Khu vực cửa sổ chức năng



R Giao diện của phần mềm R

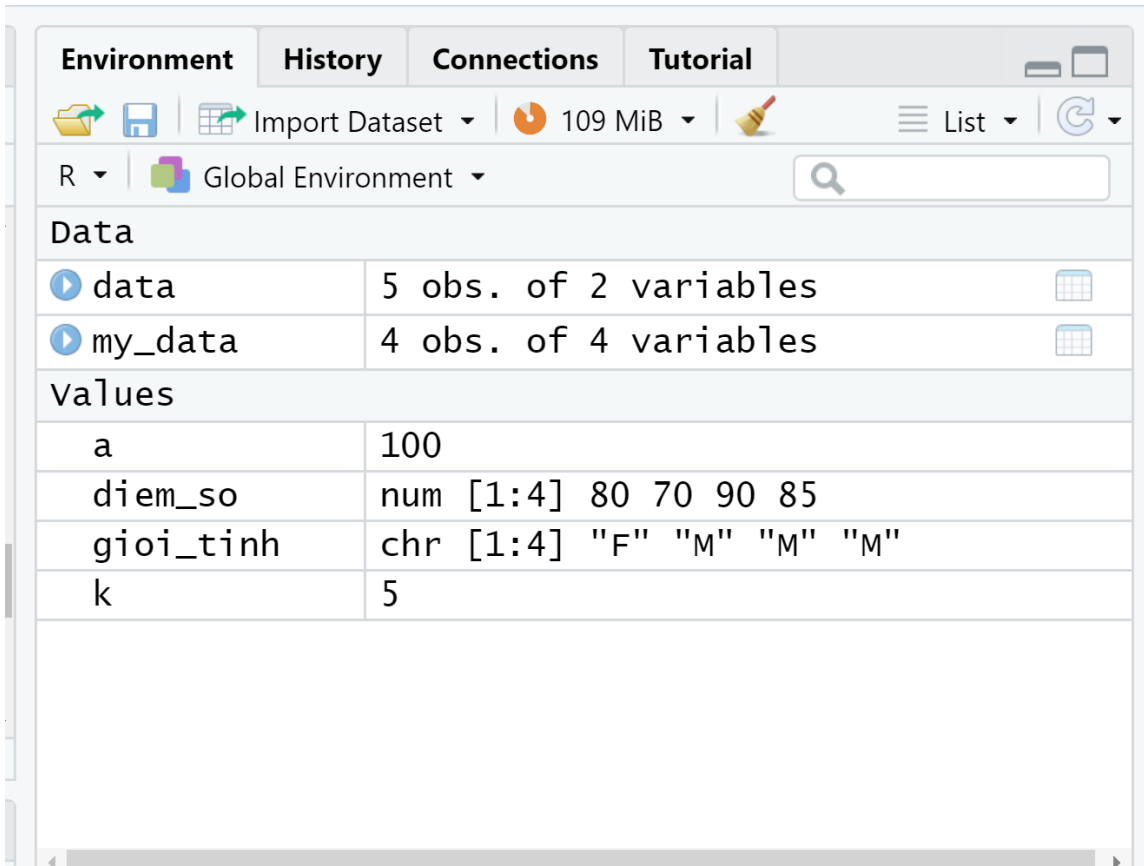
Khu vực cửa sổ chức năng

- **packages** là tập hợp các câu lệnh được phát triển bởi các lập trình viên hoặc nhà nghiên cứu trước đó nhằm giúp việc tiến hành phân tích nhanh, gọn và hiệu quả hơn.
 - Ví dụ nếu chúng ta sử dụng package tidyverse, trong đó sẽ có hàng loạt các câu lệnh như: `select()`, `mutate()`, `filter()`,...
- Cơ bản nhất, để cài đặt packages chúng ta sẽ dùng câu lệnh `install.packages()` và để nhập packages ta dùng lệnh `library()`. **VD3**
- Để cài đặt nhiều packages cùng một lúc có thể dùng `p_load()` (thuộc packages pacman) **VD3**
 - Ví dụ: `p_load(tidyverse, rio, here)`



R Giao diện của phần mềm R

Khu vực môi trường làm việc



The screenshot shows the RStudio Environment pane. At the top, there are tabs for 'Environment', 'History', 'Connections', and 'Tutorial'. Below the tabs, there are icons for file operations and a search bar. The 'Global Environment' is selected, and it contains two data objects: 'data' (5 observations of 2 variables) and 'my_data' (4 observations of 4 variables). Below the 'Data' section, there is a 'Values' section showing the values of the objects: 'a' is 100, 'diem_so' is a numeric vector [1:4] with values 80, 70, 90, 85, 'gioi_tinh' is a character vector [1:4] with values 'F', 'M', 'M', 'M', and 'k' is 5.

Data	
data	5 obs. of 2 variables
my_data	4 obs. of 4 variables

Values	
a	100
diem_so	num [1:4] 80 70 90 85
gioi_tinh	chr [1:4] "F" "M" "M" "M"
k	5

- Hiện thị các đối tượng, tệp dữ liệu, biến số được nhập hoặc tạo ra trong quá trình xử lý.
- Hỗ trợ làm việc với nhiều bộ dữ liệu cùng 1 lúc.
- Bấm chọn vào các tệp để hiển thị đối tượng được lưu trữ
- 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.
 - Để làm sạch môi trường làm việc dùng câu lệnh: `rm(list=ls())`

R Giao diện của phần mềm R

Khu vực môi trường làm việc

Các dạng dữ liệu trong R

Đối tượng (objects)

Bảng biểu
(data frame)

Chuỗi (List)

Ma trận
(Matrices)

Vector (Biến số)

Định lượng
(Numeric)

Định tính
(Characters)

Logic

Loại Biến số

Integer

Continuous

Factors

R Giao diện của phần mềm R

Khu vực môi trường làm việc

VD4 về các dạng dữ liệu

Tên	Tuổi	Chọn
Huy	28	T
Nga	18	F
Loan	23	F
Linh	29	F
Nam	36	T

Character Numeric Logic

data

Bảng dữ liệu

Tên	Tuổi	Chọn
Huy	28	T
Nga	18	F
Loan	23	F
Linh	29	F
Nam	36	T

Xác định cấu trúc dữ liệu: **str()**

R Thực hành

- Cài đặt packages, load nhiều packages cùng một lúc: `rio`, `here`, `tidyverse`, `lubridate`.
- Làm sạch môi trường dữ liệu `rm(list = ls())`
- Xác định cấu trúc dữ liệu: `str()`.

R Tổng kết bài 1

- R studio có **4** cửa sổ chính:
 - Cửa sổ ghi/ lưu trữ dòng lệnh: Mẫu cấu trúc câu lệnh, kí tự đặc biệt
 - Cửa sổ thực thi hiển thị câu lệnh
 - Cửa sổ môi trường làm việc: làm sạch môi trường
 - Cửa sổ chức năng: nhập packages

Bài 2. Giới thiệu quy trình phân tích dữ liệu cơ bản với



R Giới thiệu bộ số liệu

Bộ số liệu COVID-19 ở một số quốc gia châu Á

Tên biến	Định nghĩa biến	Loại biến	Ví dụ
iso_code	Mã số viết tắt quốc tế của mỗi quốc gia	Character	VNM: Việt Nam
location	Tên quốc gia	Character	Việt Nam
year	Năm cập nhật số liệu	Numeric	2022
total_cv_cases	Tổng số lượng ca mắc	Numeric	3212022
total_cv_death	Tổng số lượng ca tử vong	Numeric	3312
vax_per_100	% dân số tiêm ít nhất 1 mũi	Numeric	67%

R Câu hỏi phân tích

- 1. Xác định 10 quốc gia có số lượng ca nhiễm/ tử vong trên 10 vạn dân nhiều nhất
- 2. So sánh số lượng ca nhiễm/ ca tử vong trung bình giữa quốc gia có tỷ lệ tiêm chủng đầy đủ trên 60%.
- 3. Liệu có mối liên quan giữa tỉ lệ số lượng ca tử vong/10k dân và phần trăm dân số được tiêm đầy đủ

R Nội dung bài học

- Quy trình phân tích dữ liệu
- Xây dựng câu hỏi phân tích
- Thiết lập môi trường quản lý dữ liệu bằng R project
- Nhập, làm sạch và xuất dữ liệu

R Quy trình phân tích dữ liệu

B1. Xác định câu hỏi phân tích/ mục tiêu phân tích. Xác định được dữ liệu đầu vào (input) và thông tin đầu ra (output).

B2. Thiết lập môi trường quản lý dữ liệu

B3. Nhập, làm sạch và xuất dữ liệu: Nhóm các câu lệnh phổ biến trong R

B4. Chuyển đổi dữ liệu

B5. Phân tích dữ liệu:

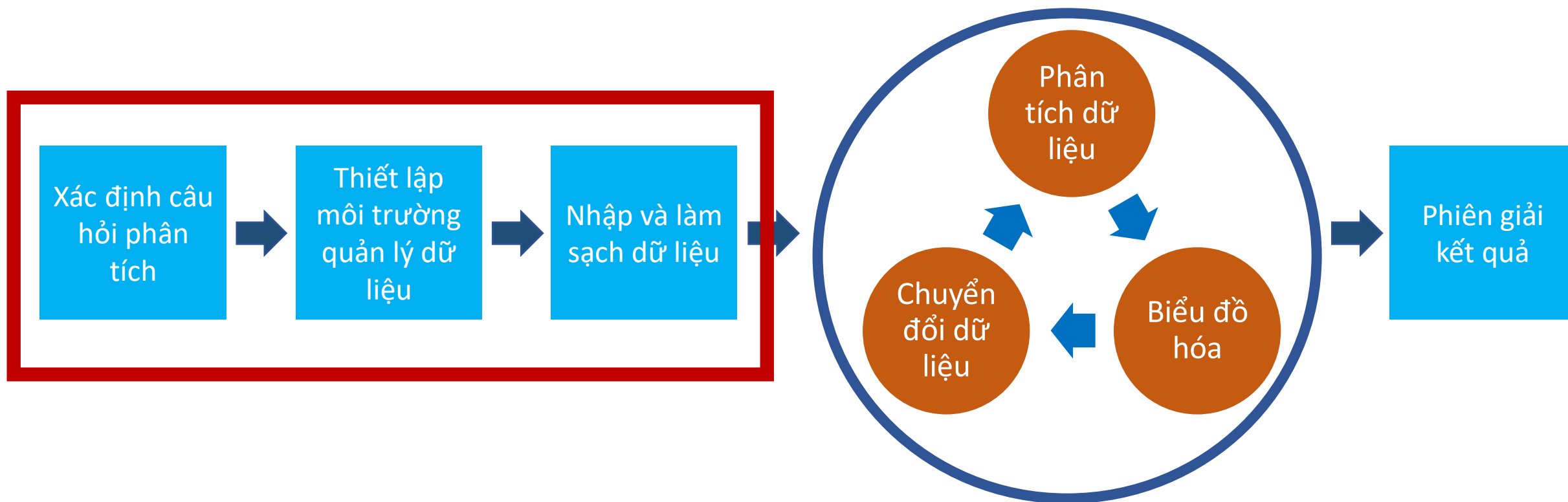
Phân tích mô tả: Kiểm tra phân bố dữ liệu (Có thể phải quay lại bước kiểm tra và làm sạch) đơn biến định tính (N (%)), biến định lượng (Mean(SD), Median(Q1,Q3)), bảng chéo 2xn

Phân tích hồi quy: Hồi quy tuyến tính, hồi quy logistic

B6. Biểu đồ hóa kết quả: Biểu đồ cột, biểu đồ scatter plot, biểu đồ barplot,...

B7. Phiên giải kết quả và viết báo cáo: Phiên giải kết quả hồi quy.

R Quy trình phân tích dữ liệu



Biểu đồ 1. Quy trình phân tích dữ liệu trong R

R Quy trình phân tích dữ liệu

B1. Xác định câu hỏi phân tích/ mục tiêu phân tích. Xác định được dữ liệu đầu vào (input) và thông tin đầu ra (output).

B2. Thiết lập môi trường quản lý dữ liệu

B3. Nhập, làm sạch và xuất dữ liệu: Nhóm các câu lệnh phổ biến trong R

B4. Chuyển đổi dữ liệu

B5. Phân tích dữ liệu:

Phân tích mô tả: Kiểm tra phân bố dữ liệu (Có thể phải quay lại bước kiểm tra và làm sạch) đơn biến định tính (N (%)), biến định lượng (Mean(SD), Median(Q1,Q3)), bảng chéo 2xn

Phân tích hồi quy: Hồi quy tuyến tính, hồi quy logistic

B6. Biểu đồ hóa kết quả: Biểu đồ cột, biểu đồ scatter plot, biểu đồ barplot,...

B7. Phiên giải kết quả và viết báo cáo: Phiên giải kết quả hồi quy.

R B1. Xây dựng câu hỏi phân tích

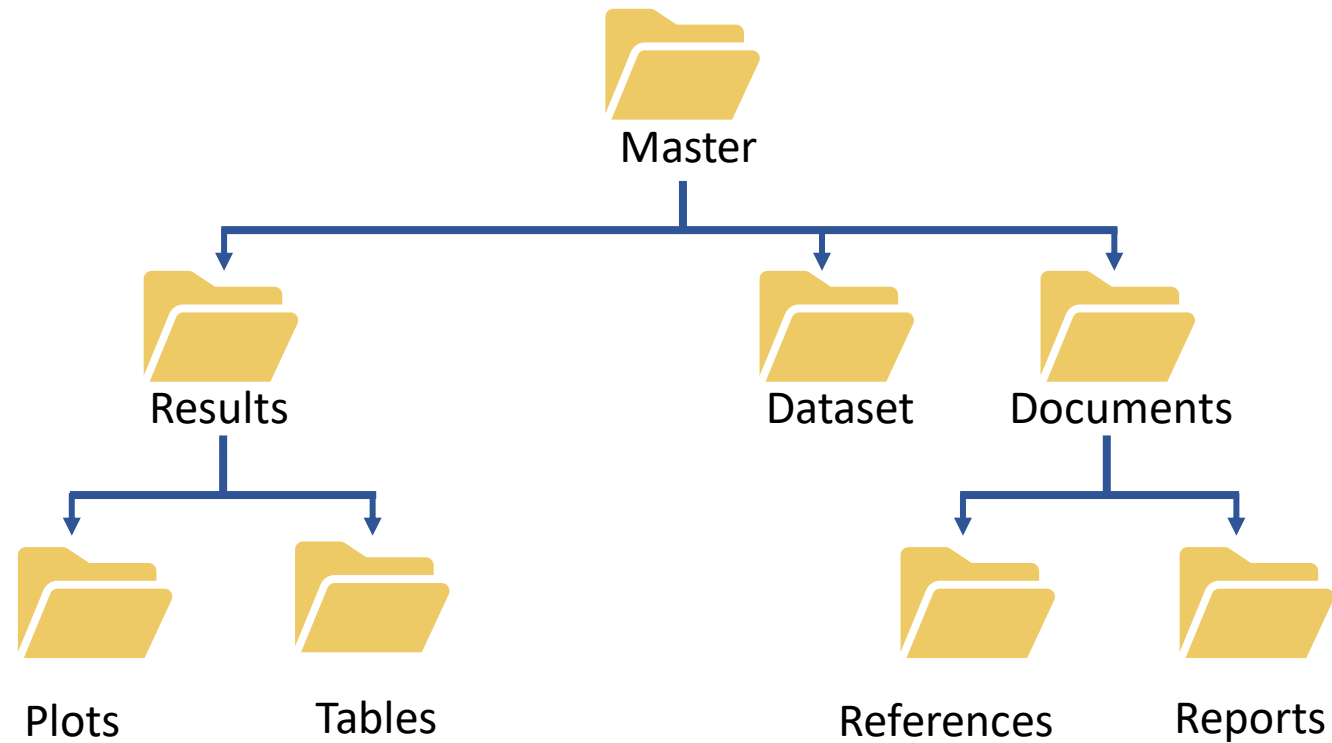
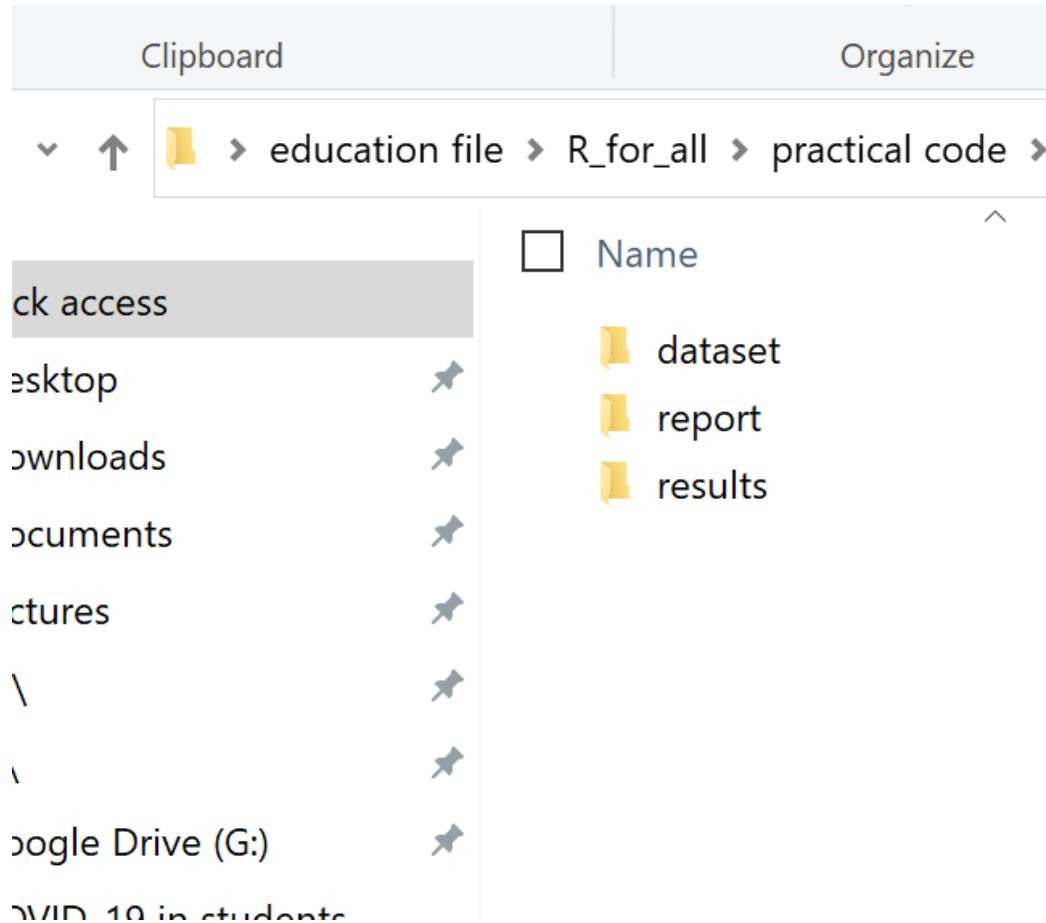
- Xác định được 2 thành tố:
 - Biến số kết quả
 - Biến số phụ thuộc
- Hướng phân tích dự kiến: các nguồn dữ liệu cung cấp biến số kết quả và biến số phụ thuộc.
- Có cần phải áp dụng công thức để tính được kết quả. (TLTK cho các công thức phức tạp)

R B2. Thiết lập môi trường phân tích

- Tạo 3 nhóm thư mục: dataset (clean data) , results, scripts, reference documents,
- Đặt working directory cho việc nhập dữ liệu.
 - Đường dẫn tương đối: dễ dàng chia sẻ file scripts cho các máy tính khác
 - Đường dẫn tuyệt đối: rất dễ bị đứt gãy khi mở câu lệnh trên máy tính mới
- Quy tắc đặt tên file:

Tên file = [Nội dung file]-[Tên người tạo]-[ngày tạo].

R B2. Thiết lập môi trường phân tích



Các bước tạo file R project

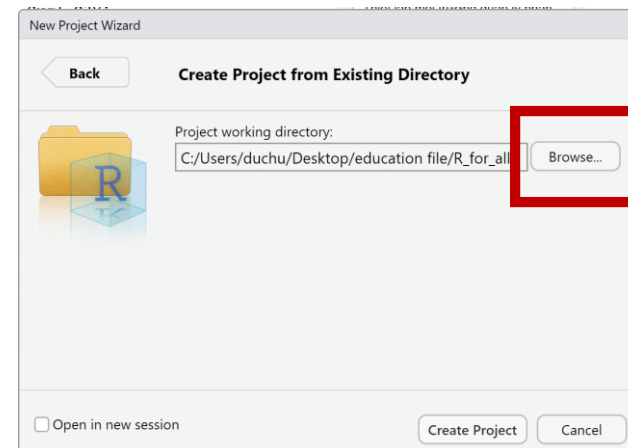
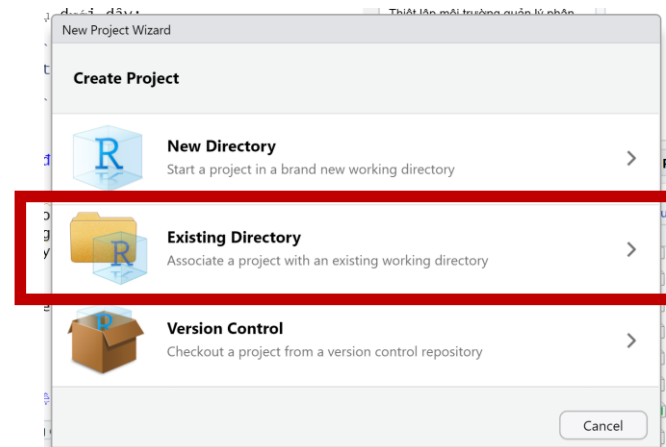
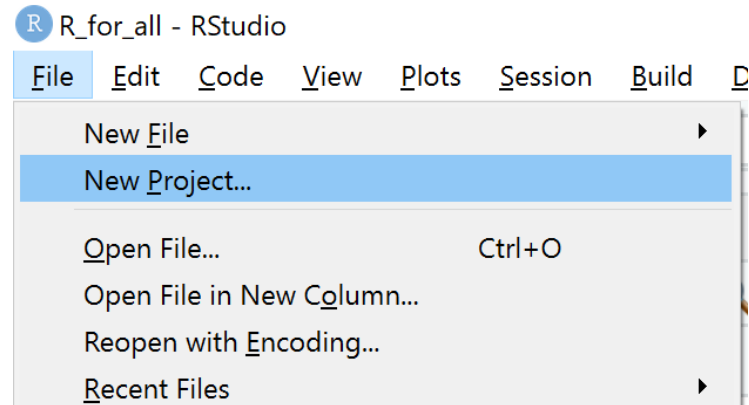
Bước 1: Bấm File > chọn New Project...



Bước 2: Xuất hiện hộp thoại > chọn “Existing Directory”



Bước 3: Xuất hiện hộp thoại tiếp theo > chọn nút “Browse” > chọn thư mục Master project trước đó





B3. Nhập, làm sạch và xuất dữ liệu

Nhập dữ liệu

- Bước 1. Tạo file R project.
- Bước 2. Mở R project và tạo R markdown.
- Bước 3. Đưa file dữ liệu vào thư mục data
- Bước 4. Xác định được định dạng dữ liệu
- Bước 5. Dùng package trong R để nhập dữ liệu. Một số package dùng để nhập dữ liệu vào R bao gồm rio, here.
- Bước 6: Mở file dữ liệu và kiểm tra sơ bộ dữ liệu nhập vào có hoàn chỉnh.
- Bước 7. Ghép nối dữ liệu nếu cần



B3. Nhập, làm sạch và xuất dữ liệu

Nhập dữ liệu

Mẫu câu lệnh để nhập file bằng packages rio, here

- `df <- import(here(["tên thư mục"], ["tên file"]))`

Data có thể nhập vào với các định dạng sau

- Excel (.xlsx)
- csv
- Rdata, RDS
- Dta
- sav



B3. Nhập, làm sạch và xuất dữ liệu

Làm sạch dữ liệu

- Sau khi nhập dữ liệu, một số câu hỏi chúng ta cần đặt ra đó là:
 - Dữ liệu có nhiều missing data không?
 - Các biến số ghi nhận vào có định dạng đúng không?
 - Có nhiều giá trị outliers trong các nhóm biến không?
 - Có biến số nào chúng ta cần nhóm lại không?
 - Biến số nào cần tính toán lại?
 - Liệu chúng ta có cần dùng hết bộ dữ liệu hay chỉ cần giữ lại các biến số chúng ta thực sự quan tâm?



B3. Nhập, làm sạch và xuất dữ liệu

Làm sạch dữ liệu

- Bước 1. Đánh sơ bộ dữ liệu để kiểm tra các định dạng biến, các giá trị outliers,...
- Bước 2. Bảng tần số để đánh giá phân bố dữ liệu của biến số định tính...
- Bước 3. Dùng đồ thị để xác định phân bố dữ liệu của biến số định lượng...
- Bước 4. Kiểm tra tính logic giữa các biến...



B3. Nhập, làm sạch và xuất dữ liệu

Xuất dữ liệu đã làm sạch

Mẫu câu lệnh để xuất file bằng packages rio, here

- `export([tên data xuất] ,here(["tên thư mục"], ["tên file"]))`

Data có thể xuất dưới các định dạng sau

- Excel(.xlsx)
- csv
- RData, RDS
- dta
- sav

R Thực hành

- Thiết lập môi trường phân tích trước
- Nhập dữ liệu covid_df.rds (**Assignment 3**)
- Lưu dữ liệu dưới dạng file covid_df.xlsx (**Assignment 3**)

R Tóm tắt bài học 2

- Xây dựng được câu hỏi và mục tiêu thực hiện phân tích sẽ giúp xác định biến số chính, biến phụ thuộc và biến độc lập.
- Nhập, làm sạch và xuất dữ liệu tập trung chủ yếu việc sử dụng 3 packages tidyverse, rio, here.

R Một số tài liệu học tập tham khảo

- <https://r4ds.had.co.nz/>
- <https://epirhandbook.com/vn/index.html>