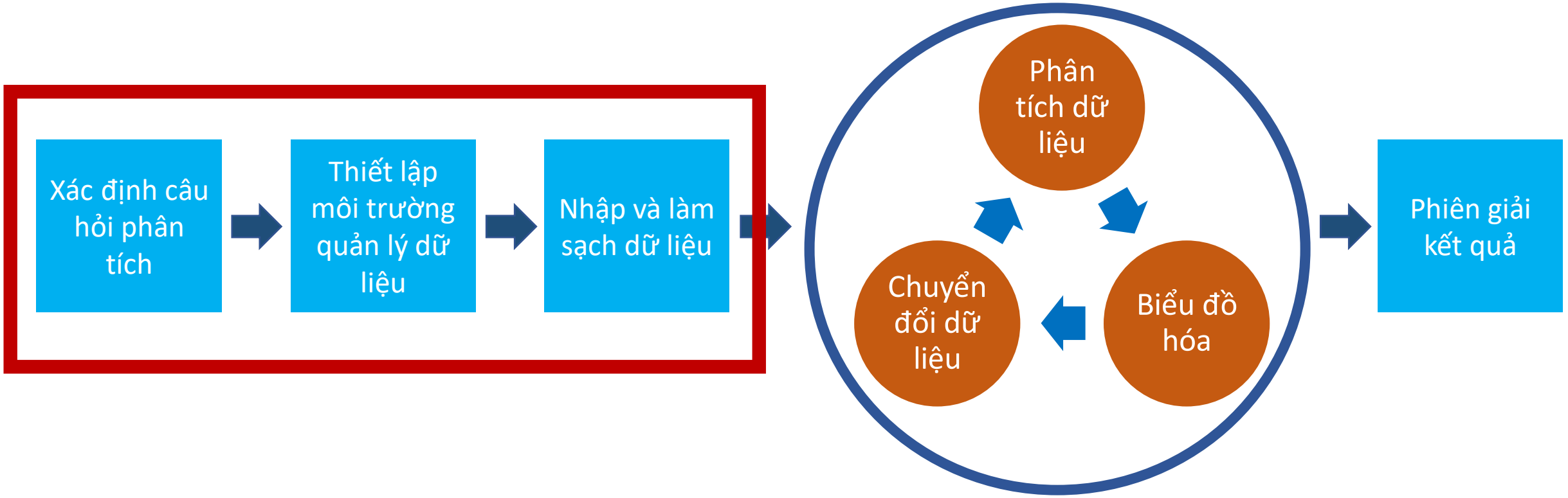
Bài 2. Giới thiệu quy trình phân tích dữ liệu cơ bản với R

Quy trình phân tích dữ liệu



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

B1. Xây dựng câu hỏi phân tích

Xác định được **2** thành tố:

- **Biến số:** Biến số kết quả, Biến số độc lập
  - 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)
- **Cấp độ phân tích:** mô tả, xây dựng mô hình, tìm kiếm mối quan hệ

B1. Xây dựng câu hỏi phân tích

Xác định biến số và cấp độ 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

B2. Thiết lập môi trường phân tích

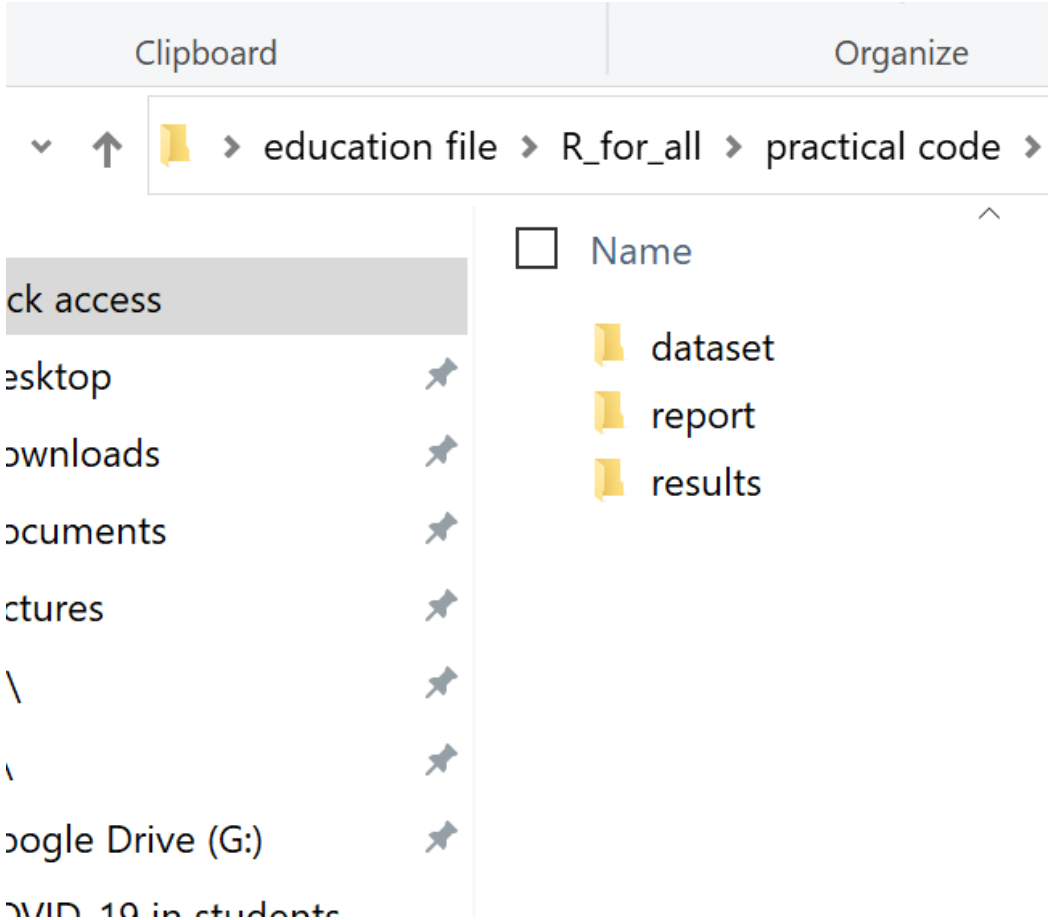
Why???

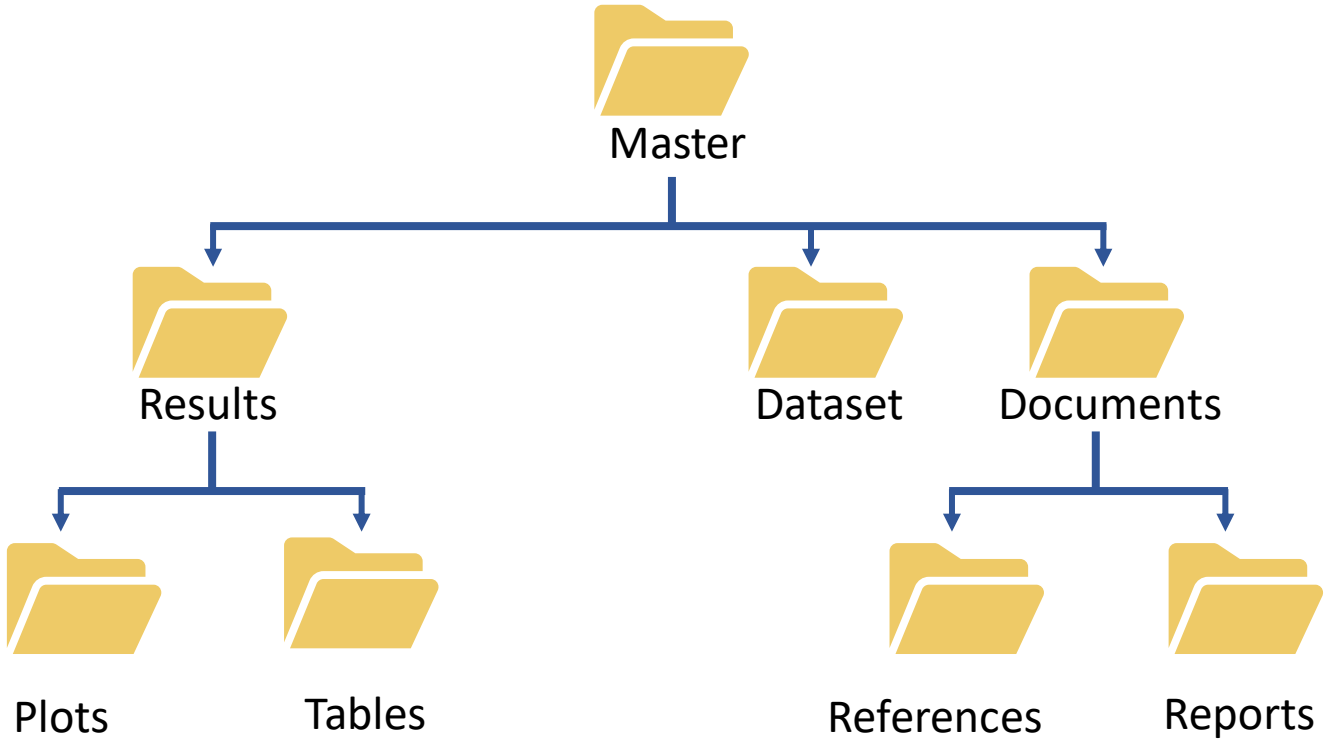
CHIA SẺ VÀ TÁI SỬ DỤNG

- 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].

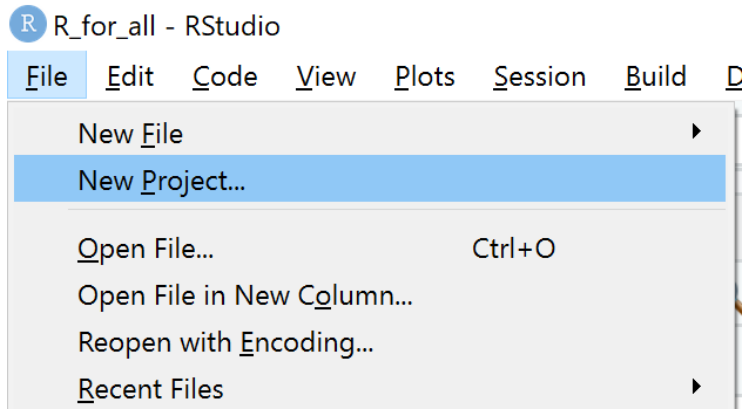
B2. Thiết lập môi trường phân tích



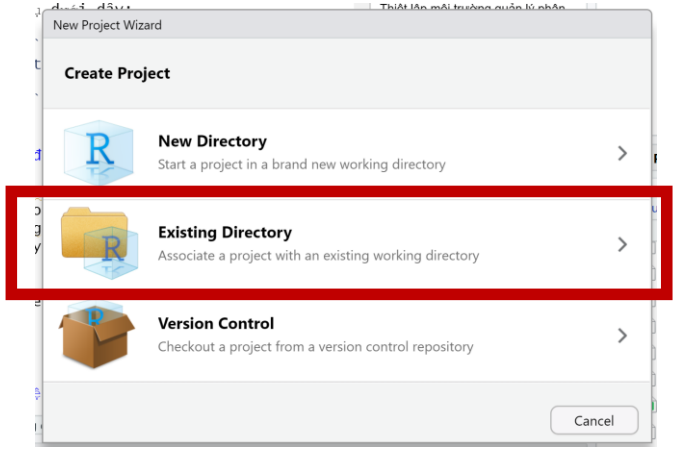


Các bước tạo file R project

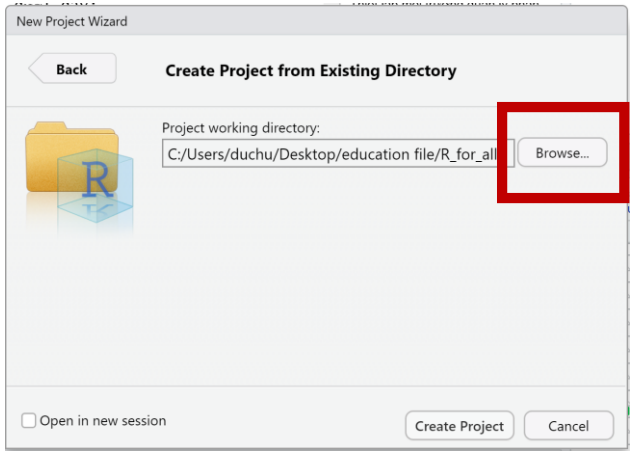
Bước 1: Bấm File > chọn New Project…



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



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

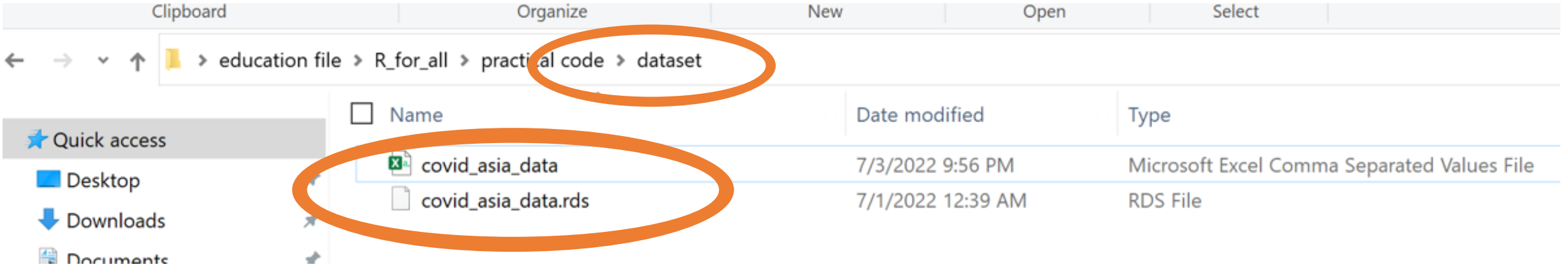


B3. Nhập, làm sạch và xuất dữ liệu

Nhập dữ liệu

- Bước 3.1. Xác định **vị trí** file dữ liệu
- Bước 3.2. Xác định được **định dạng** dữ liệu.
- Bước 3.3. **Đưa file dữ liệu vào thư mục** data
- Bước 3.4. Dùng **package** trong R để nhập dữ liệu.
- Bước 3.5: **Mở file dữ liệu và kiểm tra sơ bộ dữ liệu** nhập vào có hoàn chỉnh.
- Bước 3.6. **Ghép nối dữ liệu** (nếu cần)

B3. Nhập, làm sạch và xuất dữ liệu
Nhập dữ liệu=>B3.1.Vị trí - B3.2. Định dạng - B3.3. Sắp xếp



- csv (định dạng file các cột được chia cách bằng dấu phẩy),
- excel,
- Định dạng dữ liệu của R: RData (lưu trữ dữ liệu và các tệp đối tượng trên R), rds (lưu trữ dữ liệu trên R),
- Định dạng dữ liệu của các phần mềm thống kê khác ví dụ như `.sav` (spss), `dta` (STATA), `sas7bdat` (SAS). - Trong trường hợp ví dụ COVID data, định dạng dữ liệu xác định được là csv hoặc rds.

B3. Nhập, làm sạch và xuất dữ liệu

Nhập dữ liệu=> B3.4. Dùng package để nhập dữ liệu vào R

- Dùng hàm package trong R để nhập dữ liệu Một số package dùng để nhập dữ liệu vào R bao gồm rio,here

Mẫu câu lệnh để nhập file bằng packages rio, here

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

VD: df<- import(here(“dataset”, “covid_df.rds”))

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

Nhập dữ liệu=> B3.5. Mở file dữ liệu và kiểm tra sơ bộ dữ liệu

- Kiểm tra hàng đầu tiên (tên biến)
- Kiểm tra số lượng hàng
- Kiểm tra dữ liệu missing

B3. Nhập, làm sạch và xuất dữ liệu

Nhập dữ liệu=> B3.6. Nối dữ liệu từ các nguồn dữ liệu khác

- Kết nối 2 bộ dữ liệu với nhau bằng câu lệnh sau:

+ merge: Nối 2 bộ dữ liệu có cùng cột id

Vd: df3 <- merge(df1, df2, by.x= “id”, by.y= “id”)

+ rbind: nối 2 data có cùng số hàng

Vd: df3 <- rbind(df1, df2)

+ cbind: nối 2 data có cùng số cột

Vd: df3 <- cbind(df1,df2)

B3. Nhập, làm sạch và xuất dữ liệu

Làm sạch dữ liệu

Câu hỏi	Giải pháp
Dữ liệu có nhiều missing data không?	summary()
Có nhiều giá trị outliers trong các nhóm biến không?	hist(), ggplot2()
Các biến số ghi nhận vào có định dạng đúng không?	str()
Có biến số nào chúng ta cần nhóm lại không?	case_when()
Biến số nào cần tính toán lại?	mutate()
Liệu chúng ta có cần dùng hết bộ dữ liệu hay chỉ cần giữ lại các biến số chúng ta thực sự quan tâm?	select(), filter()

B3. Nhập, làm sạch và xuất dữ liệu

Làm sạch dữ liệu

Lưu ý:

- 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"]))

Ví dụ:

Data có thể xuất dưới các định dạng sau

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

Tóm tắt bài học 2

- Xây dựng được câu hỏi hay mục tiêu phân tích rất quan trọng trong việc định hướng kế hoạch phân tích.
- Thiết lập môi trường quản lý và phân tích dữ liệu với R project.
- Nhập, làm sạch và xuất dữ liệu tập trung chủ yếu việc sử dụng 3 packages tidyverse, rio, here.

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)**

Một số tài liệu học tập tham khảo

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