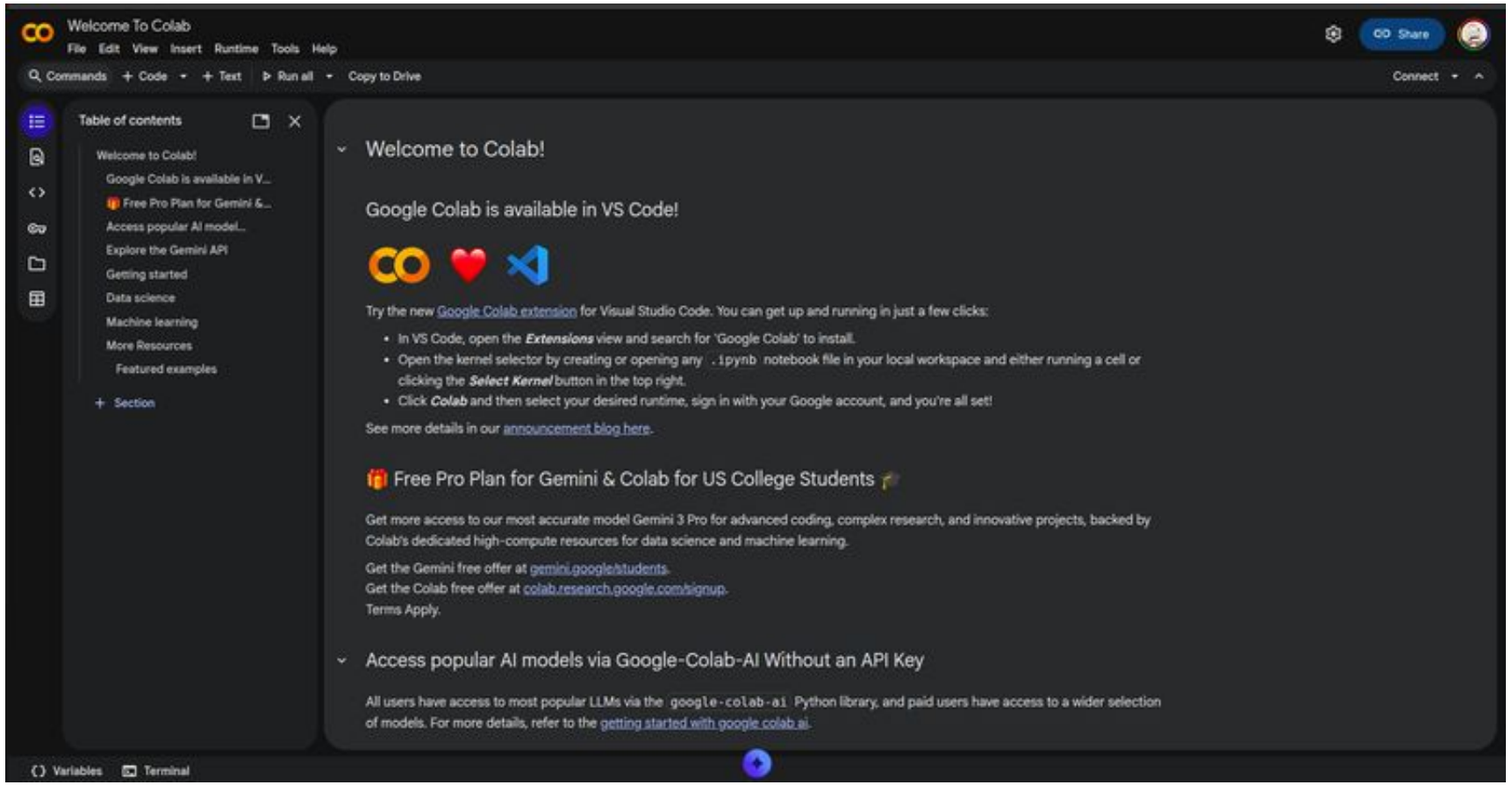
Giới thiệu Google colab và ứng dụng AI trong viết ngôn ngữ lập trình R

TS. Lưu Phúc Lợi,
Trưởng Phòng Nghiên cứu khoa học, Viện Ariha
Bệnh viện Thống Nhất

Jan 11, 2026

GIỚI THIỆU GOOGLE COLAB

- Free, cloud-based tool that lets you write and run code
- “Google Docs for Code”
- runs entirely in your web browser—no software installation required



GIỚI THIỆU NGÔN NGỮ LẬP TRÌNH R

- Môi trường tính toán cho thống kê, mô hình hoá và trực quan hoá dữ liệu.
- Miễn phí, mã nguồn mở.
- Hỗ trợ phân tích tái lập (R Markdown/Quarto).
- Hệ sinh thái gói lớn: CRAN (đa lĩnh vực) và Bioconductor (omics/bioinformatics).
- Sử dụng rộng rãi trong lĩnh vực học thuật
- Có thể tích hợp vào các phần mềm khác
- Dễ học

https://www.r-project.org/

Thống kê & đồ hoạ | Tái lập nghiên cứu | Gói phong phú

ỨNG DỤNG NGÔN NGỮ LẬP TRÌNH R

1) Dữ liệu & trực quan

- readr/haven: nhập dữ liệu
- dplyr/tidyr: làm sạch & biến đổi
- ggplot2: biểu đồ chất lượng xuất bản

2) Thống kê lâm sàng

- Hồi quy (GLM), mô hình hỗn hợp
- Sống sót (survival), ROC/AUC
- Thiết kế & phân tích thử nghiệm

3) Omics & bioinformatics

- Bioconductor: RNA-seq, microarray, single-cell
- Hạ tầng dữ liệu chuẩn hoá (SummarizedExperiment, …)
- Vignettes giúp tái lập workflow

4) Báo cáo & triển khai

- R Markdown/Quarto: báo cáo & slide tái lập
- Shiny: dashboard/ứng dụng tương tác
- Chia sẻ: Git + dữ liệu + môi trường

Data Visualization with R

Starting from 13:30 to 15 Wed (19/11/2025 - 25/03/2026)

Location: Auditorium at level 6th

00.0 Introduction to Data Visualizaion [Loi] - 19/11/2025

- PDF
- Book

00.1 Practice [Dan] - 19/11/2025

- PDF
- Installation R & Rstudio
- Basic R
- Introduction to Google Colab
- Practice: Open in Colab

01.0 Boxplot and its Variants [Loi] - 10/12/2025

- PDF

01.1 Practice [Dan] - 10/12/2025

- PDF
- Data for Practice 1
- Practice 1: Open in Colab
- Data for Practice 2

```r
library(ggplot2)

df <- read.csv("/content/BÀI_TẬP_1_sheet2.csv",
               header = TRUE,
               check.names = FALSE)

# Tính p-value (t-test 2 nhóm)
test_res <- t.test(Weight ~ SEX, data = df)
p_val <- test_res$p.value

b <- ggplot(df, aes(x = SEX, y = Weight)) +
 geom_boxplot(staplewidth = 1, fill = "violet", color = "black") +
 geom_jitter(
   width  = 0,
   size   = 6,
   alpha  = 1,
   color  = "brown"
 ) +
 annotate("text",
          x = 1.5,
          y = max(df$Weight) * 1,
          label = paste0("p = ", signif(p_val, 3)),
          size = 8) +
 theme(
   axis.text.y = element_text(size = 20),
   axis.text.x = element_text(size = 20)
 )

b
```

CÁC THÀNH PHẦN CHÍNH TRONG NGÔN NGỮ R

- **Base R**: tập các hàm và cấu trúc dữ liệu mặc định sau khi cài đặt.
- R vận hành qua gõ lệnh (trong R, lệnh = hàm = function)
- **Packages**: mở rộng tính năng (vd: tidyverse, ggplot2, survival, Bioconductor).
- R vận hành dựa trên 'đối tượng': **object** (bao gồm dữ liệu, biến số, hình ảnh, bảng, v.v.)
- Dữ liệu (dataset) trong R gọi là **data.frame**
- Mỗi data frame có nhiều biến số gọi là **variable**
- **Function** phải có **arguments** / đối số (đầu vào)

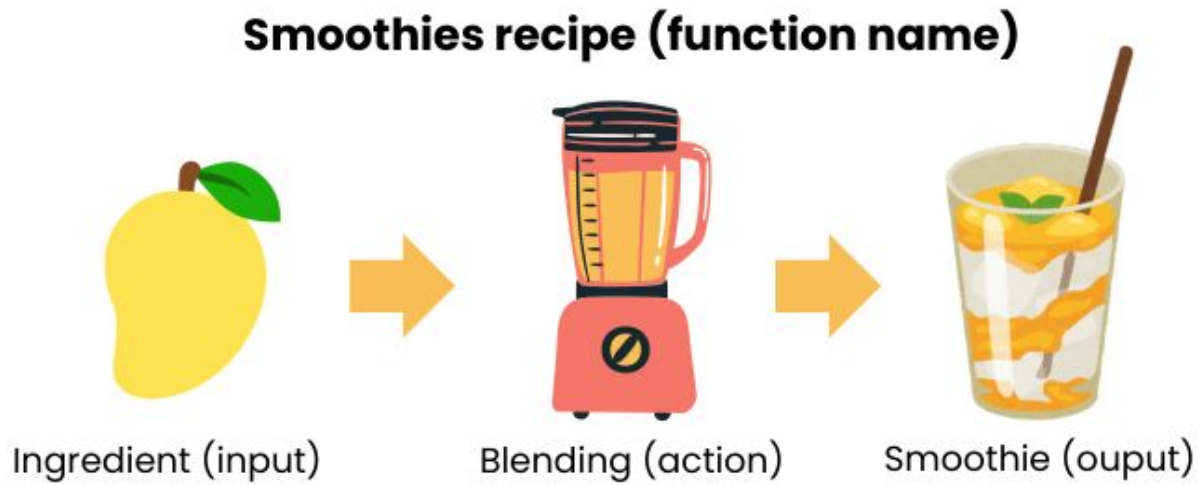
R = Base + Packages

Ví dụ: cài và gọi package để phân tích dữ liệu

```
install.packages("ggplot2")
library(ggplot2)
```

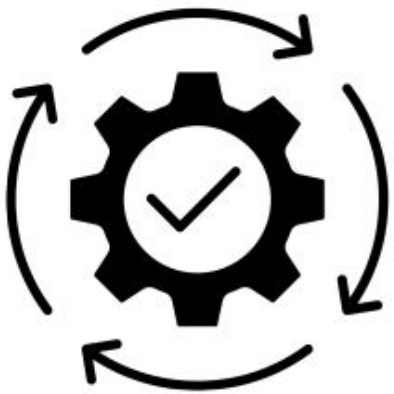
Functions

- Function: a group of instructions that takes inputs, computes other values, and returns a result.
- R has default functions:
  - mean ()
  - sum ()
  - read.table ()
  - ...
- R allows defining functions:
  - function_name ← function (input)
    {action
    return (output)}
- Why define a new function?
  - Highly reproducible
  - Don't have to write many lines of code



Why using functions?

Organization



Automation

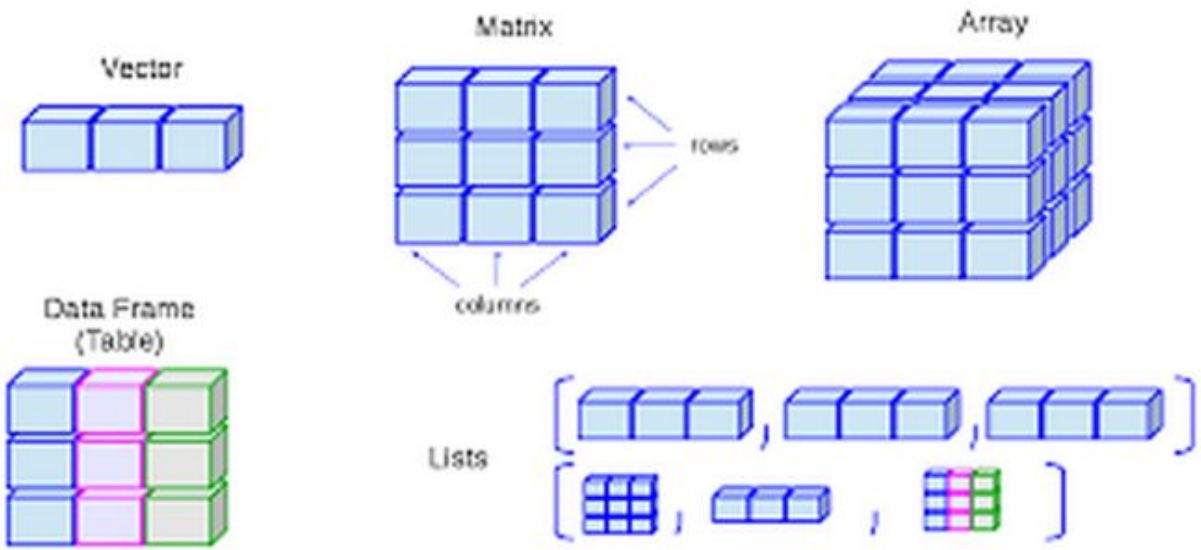
Without functions

```
> price <- 100
> tax <- price * 0.10
> total <- price + tax
> total
[1] 110
>
> price <- 1000
> tax <- price * 0.10
> total <- price + tax
> total
[1] 1100
>
> price <- 100000
> tax <- price * 0.10
> total <- price + tax
> total
[1] 110000
```

With functions

```
> calculate_total <- function(price) {
+   tax <- price * 0.10
+   total <- price + tax
+
+   return(total)
+ }
> calculate_total (100)
[1] 110
> calculate_total (1000)
[1] 1100
> calculate_total (1000000)
[1] 1100000
```

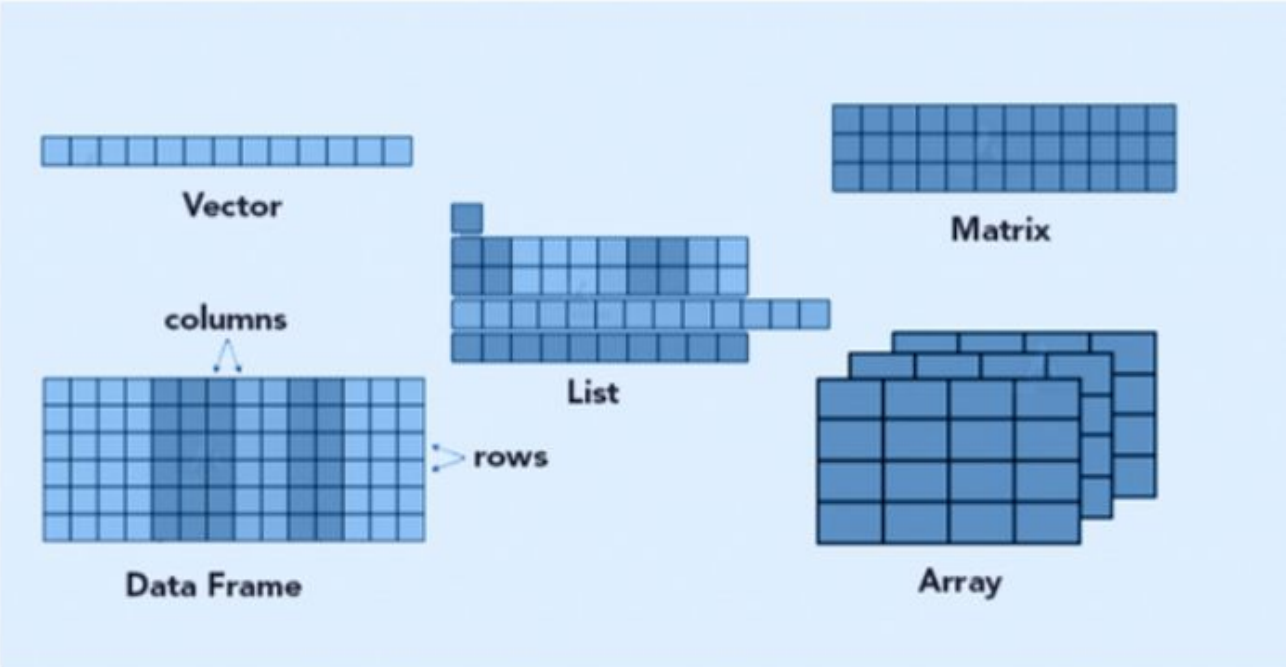
Data structure



Packages

Data structure

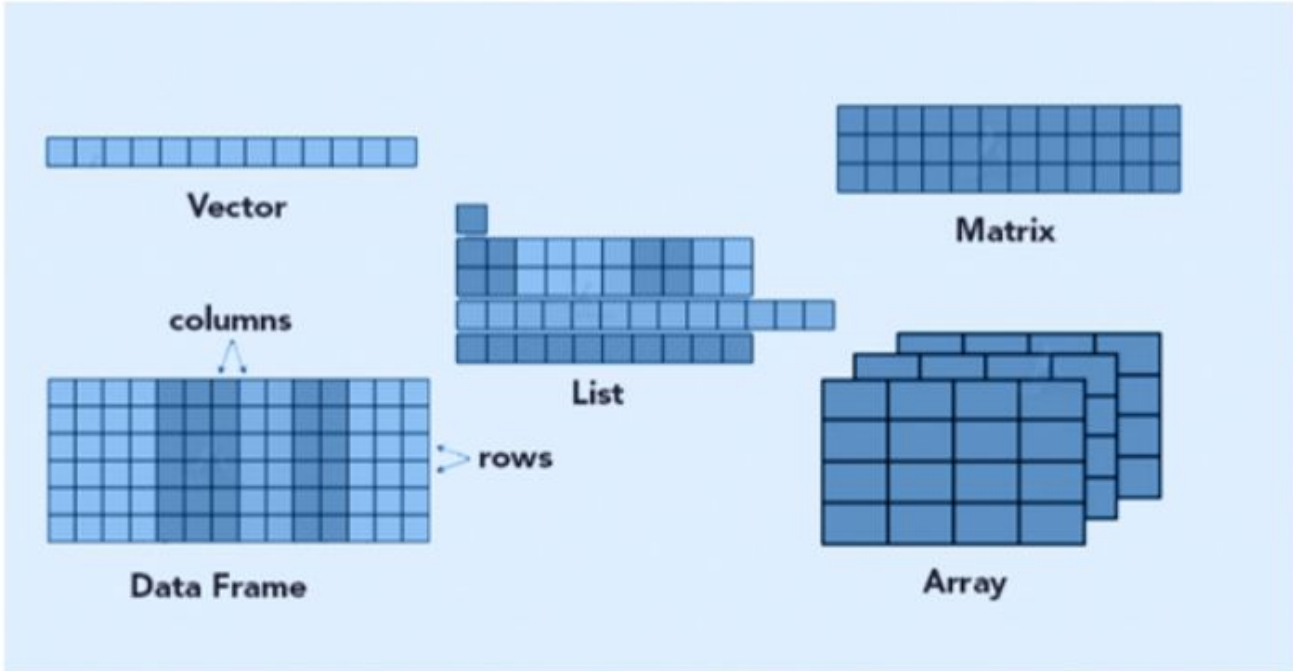
- Data structure is a container for storing and organizing data, designed to handle everything from simple lists of numbers to complex tables of mixed data types.
- Dimensionality (1D, 2D, or n-dimensional)
- Homogeneity (whether all elements must be the same type).

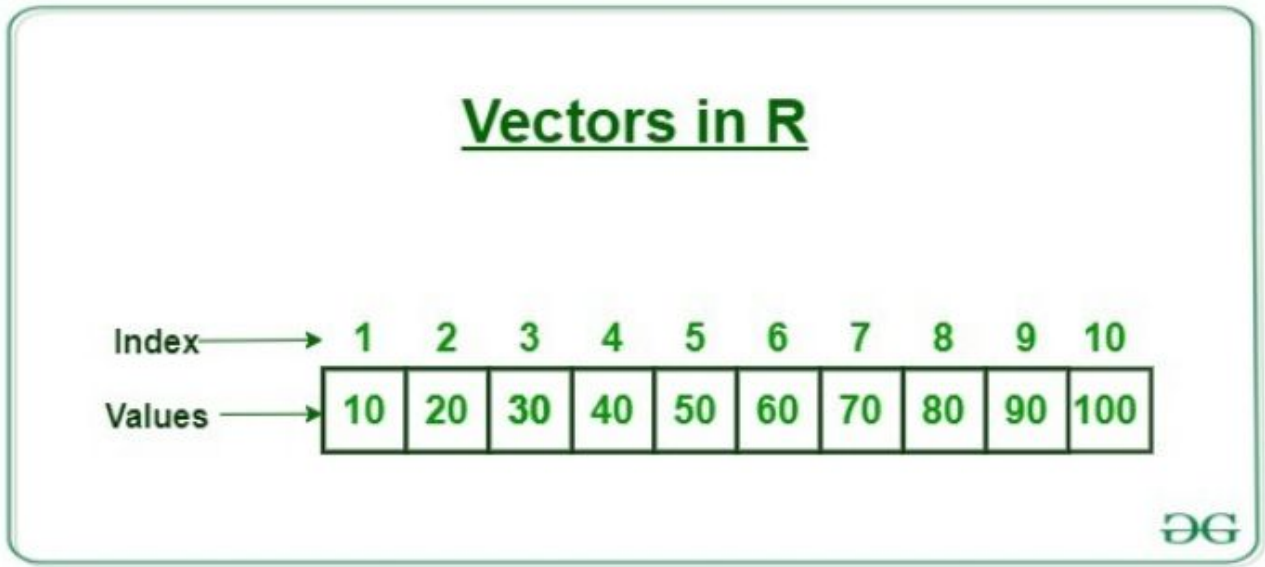


	Homogeneous	Heterogeneous
1-D	Atomic vector	List*
2-D	Matrix	Data Frame
n-D	Array	

Vector

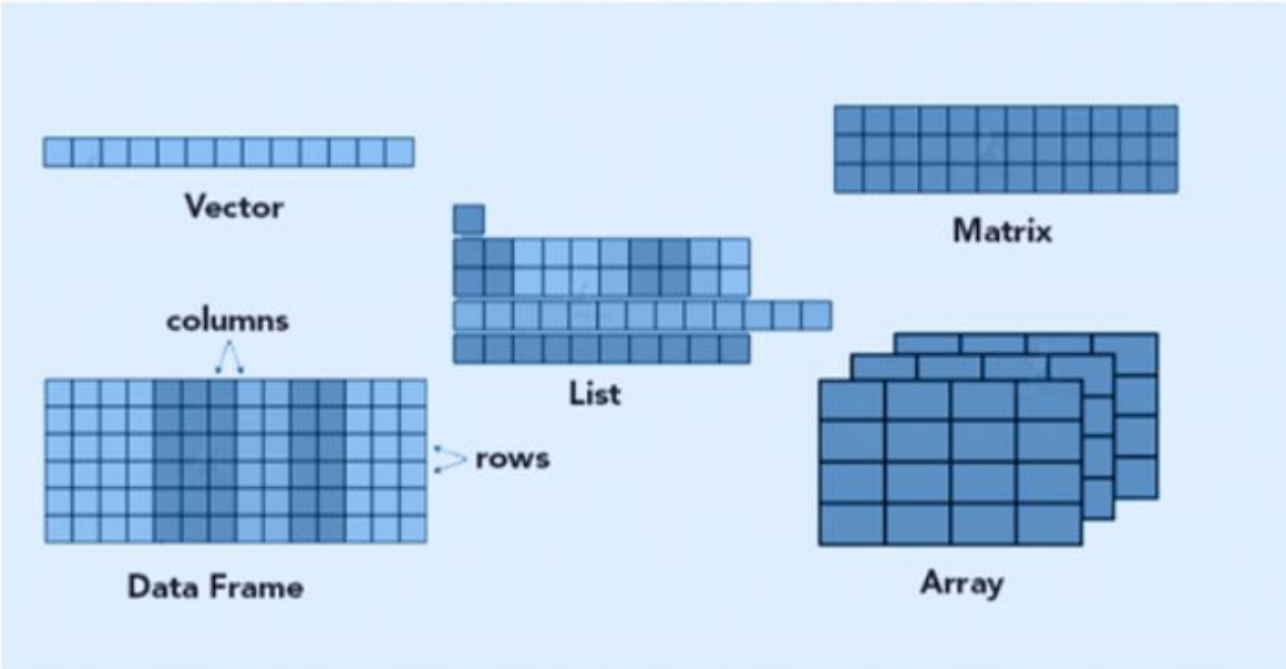
- The vector type is the heart of R
- The elements of a vector must all have the same mode, or data type. (Homogeneous)
- Dimension: 1-D
- Other data structures are just vectors with more or fewer attributes:
  - Matrices are vectors with rows and columns
  - Scalars (individual numbers) are vectors with a single component





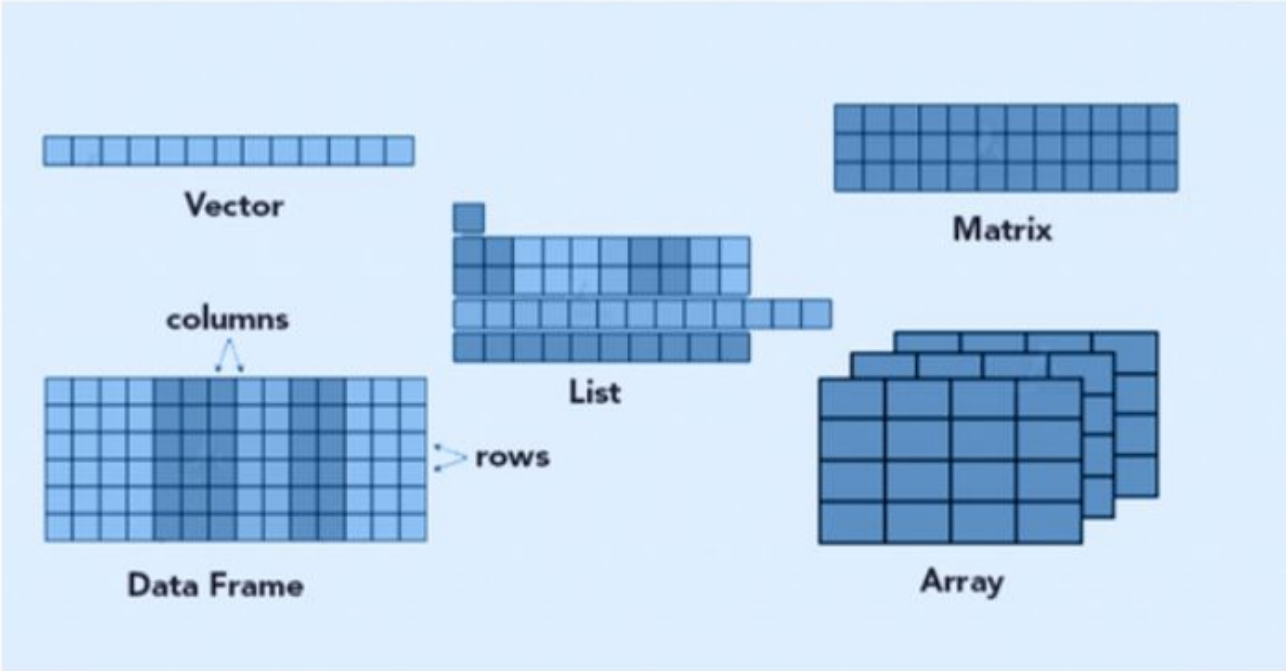
List

- List structure can combine objects of different types
- Lists are referred to as recursive vectors
- Dimension: 1-D
- Heterogeneous: contains many data types, modes



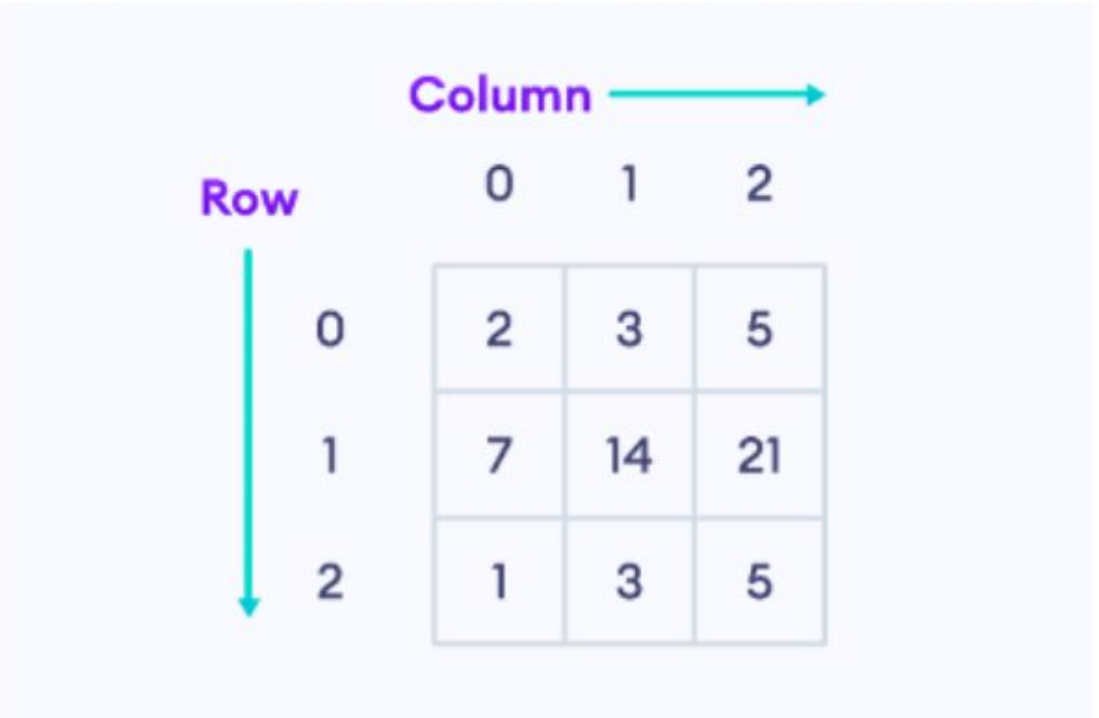
Matrix

- A matrix is a rectangular array of numbers.
- A matrix is a vector, but with two additional attributes: rows and columns
- Vectors are not one-column or one-row matrices.
- Dimension: 2-D
- **Homogeneous**



Data frame

- Data frames are matrices with different data types, modes
- Data frames are lists with rows and columns
- Dimension: 2-D
- **Heterogeneous**



Packages

- Packages store groups of related pieces of software
- Some packages are loaded automatically
- Does not load all available packages automatically

Dữ liệu nghiên cứu dạng excel (csv format)

Heart Failure Prediction Dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
2	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
3	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
4	37	M	ATA	130	283	0	ST	98	N	0	Up	0
5	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
6	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
7	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
8	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
9	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
10	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
11	48	F	ATA	120	284	0	Normal	120	N	0	Up	0
12	37	F	NAP	130	211	0	Normal	142	N	0	Up	0
13	58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
14	39	M	ATA	120	204	0	Normal	145	N	0	Up	0
15	49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
16	42	F	NAP	115	211	0	ST	137	N	0	Up	0
17	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0

Link project:
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

Làm sao đọc vào R?

Đọc dữ liệu vào R

– **Bước 1:** tìm đường link
– **Bước 2:** dùng lệnh/hàm `read.csv` và đặt tên cho dữ liệu (ví dụ: `df`)
– **Bước 3:** kiểm tra qua hàm `head(df)`

```
df <- read.csv("/content/Heart_Data.csv")
# hoặc
library(lessR)
df <- Read("/content/Heart_Data.csv")
head (df)
```

A data.frame: 6 × 12

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
	<int>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<int>	<chr>	<dbl>	<chr>	<int>
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

A data.frame: 918 × 12

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
<int>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<int>	<chr>	<dbl>	<chr>	<int>
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0.0	Up	0

Data frame (dataset) trong trường hợp này là **df**

Biến số (variables): **cột của df**: Age, Sex, ChestPainType, Cholesterol, …

df này có bao nhiêu variables?

có bao nhiêu samples?

CÁM ƠN SỰ LẮNG NGHE CỦA QUÝ ANH/CHỊ