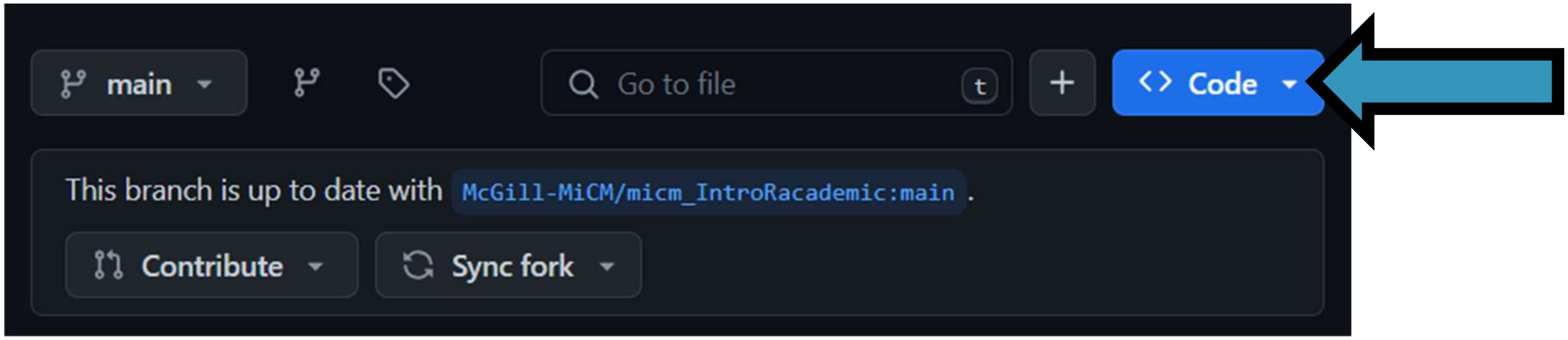
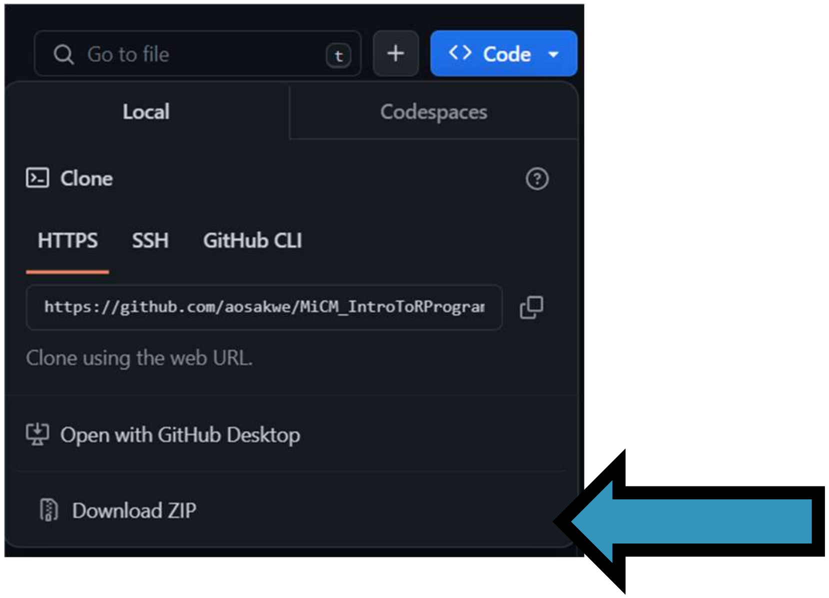
Download Workshop Materials

1. Go to https://github.com/aosakwe/MiCM_IntroToRProgramming

2.



This branch is up to date with McGill-MiCM/micm_IntroRacademic:main.



https://github.com/aosakwe/MiCM_IntroToRProgra

Clone using the web URL.

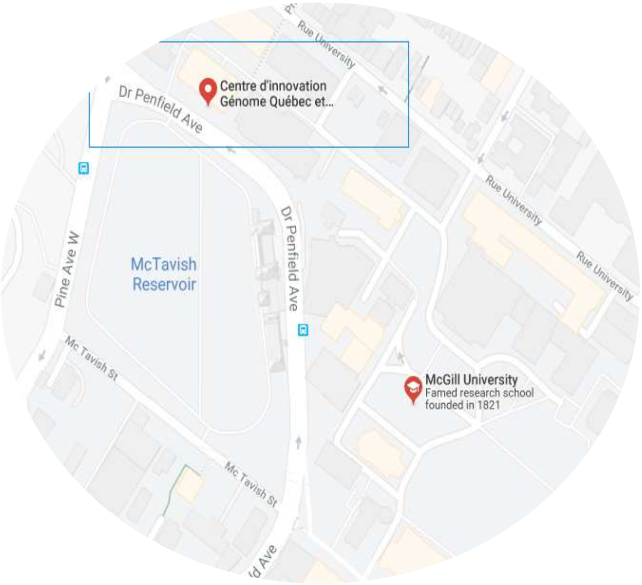
Intro to programming in R

Lead: Adrien Osakwe

February 14, 2024
Slides adapted from material by Larisa M. Soto

MiCM McGill initiative in Computational Medicine

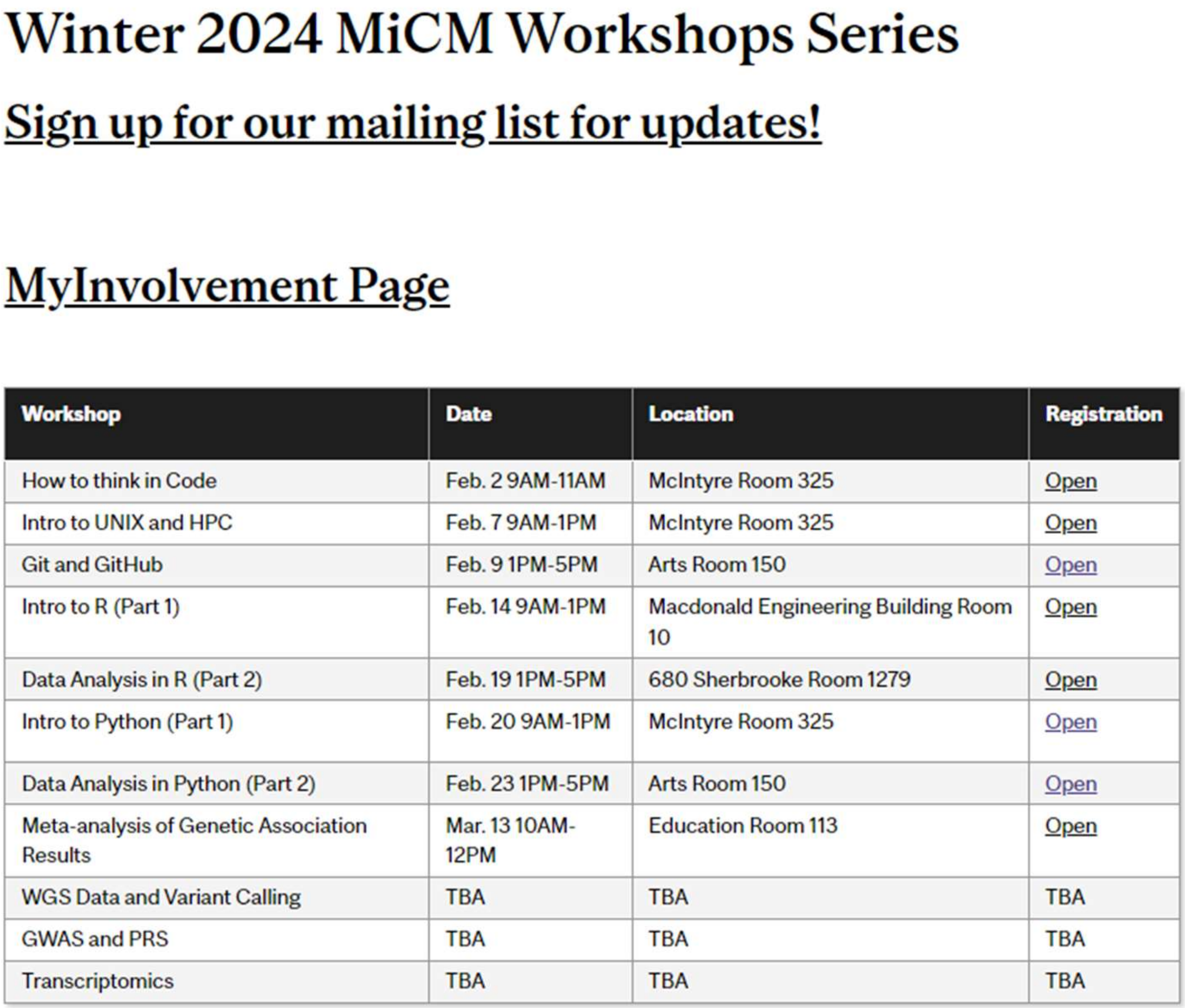
Mission statement: deliver quality workshops designed to help biomedical researchers develop the skills they need to succeed.



Location: 740 Dr. Penfield Avenue, Montreal, Quebec

Scan the QR code to sign up for our **mailing list**

Contact: workshop-micm@mcgill.ca



Winter 2024 MiCM Workshops Series

Sign up for our mailing list for updates!

MyInvolvement Page

Workshop	Date	Location	Registration
How to think in Code	Feb. 2 9AM-11AM	McIntyre Room 325	Open
Intro to UNIX and HPC	Feb. 7 9AM-1PM	McIntyre Room 325	Open
Git and GitHub	Feb. 9 1PM-5PM	Arts Room 150	Open
Intro to R (Part 1)	Feb. 14 9AM-1PM	Macdonald Engineering Building Room 10	Open
Data Analysis in R (Part 2)	Feb. 19 1PM-5PM	680 Sherbrooke Room 1279	Open
Intro to Python (Part 1)	Feb. 20 9AM-1PM	McIntyre Room 325	Open
Data Analysis in Python (Part 2)	Feb. 23 1PM-5PM	Arts Room 150	Open
Meta-analysis of Genetic Association Results	Mar. 13 10AM-12PM	Education Room 113	Open
WGS Data and Variant Calling	TBA	TBA	TBA
GWAS and PRS	TBA	TBA	TBA
Transcriptomics	TBA	TBA	TBA

https://www.mcgill.ca/micm/training/workshops-series

Workshop outline Part 1

1 **The language**

History
Foundation
Syntax
Logical ops
Help
Packages

2 **Data types**

Vectors
Factors
Lists
Data Frames
Arrays
Hands on

3 **Control Structures**

Functions
If statement
for loop
Hands on

Workshop outline Part 2

4 Basic data manipulation

Read & Write
Subset
Split
Join
Hands on

5 Advanced data Manipulation

dplyr
tidyr
plyr
DataTable
Hands on

6 Generating Outputs

Graphics
ggplot2
RMarkdown
Templates

7 Software development

Good coding practices
Documentation standards
Debugging

Workshop Components

- Theory
- Code Examples
- Hands-on Activities

1. The R programming language

Learning objectives

- Why Excel is not enough
- What is R
- What is an IDE
- Basic Operations

Why not Excel?

- Easy at first glance

- **Issues**
  1. Hard to automate
  2. Hard to reproduce
  3. Inflexible
  4. Slow!

- **Statistical** Programming Language
- Integrated suite for **data manipulation, analysis, and graphical visualization**
- Environment where **statistical tests can be performed**
- Its functionality can be easily extended with ***packages***

- GNU project of free software
- Users have the freedom to:
  - Run the program
  - View and modify the source code
  - Redistribute copies and
  - Distribute their modifications

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

R facts

- Interpreted language
- Object-oriented
- No spaces allowed in variable names
- Case sensitive
- 1-based indexing
- Allows user-defined functions
- Works with environments

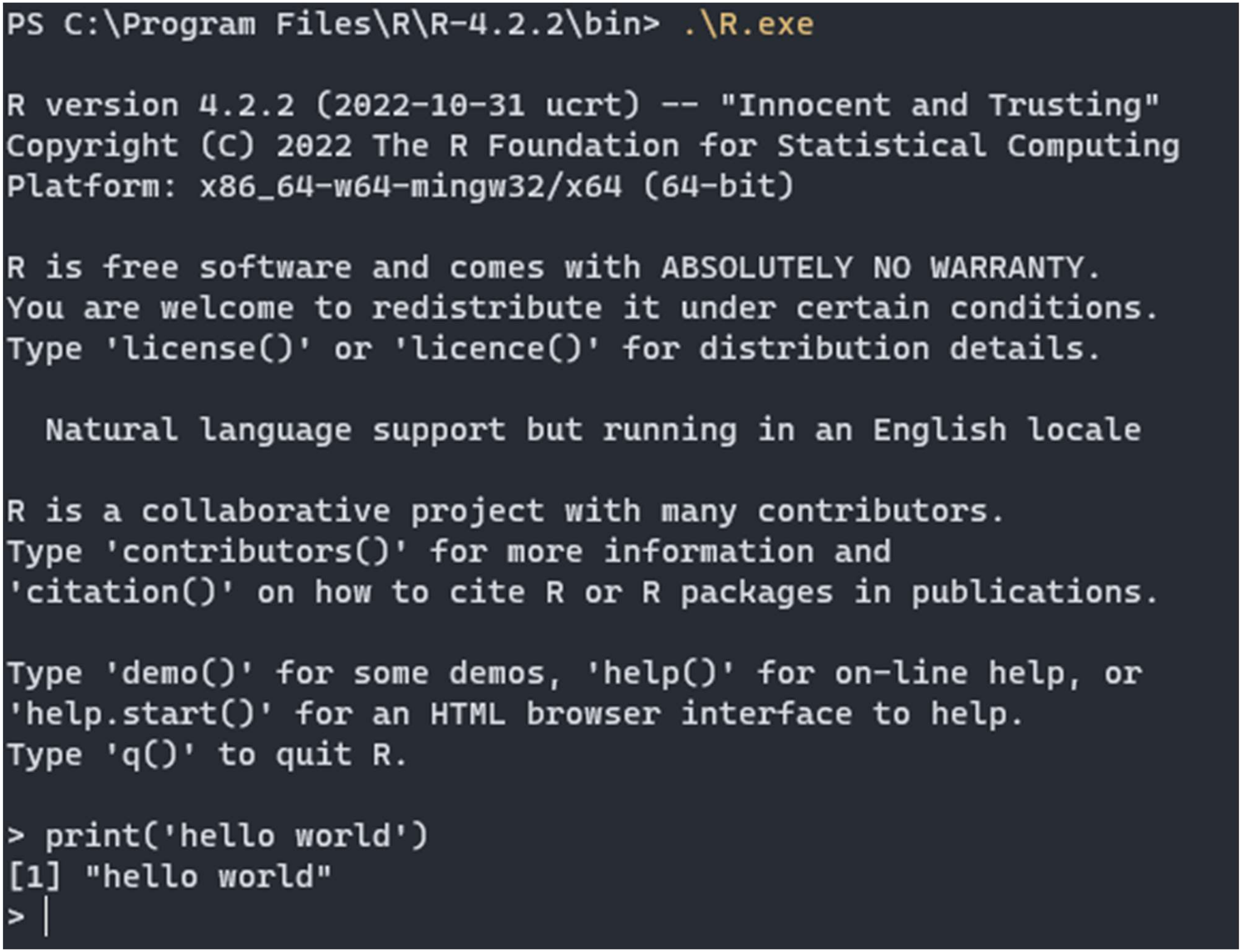
R Files

- Many types of files can contain R code
  - **.R ‘Script’**
  - **.Rmd ‘R Notebook’**
  - .qmd ‘Quarto Notebook’
  - .ipynb ‘Jupyter Notebook’
- Scripts
  - Automation & Portability
- Notebooks
  - Documentation
  - Accessibility

R & RStudio

- R & RStudio are **different entities**

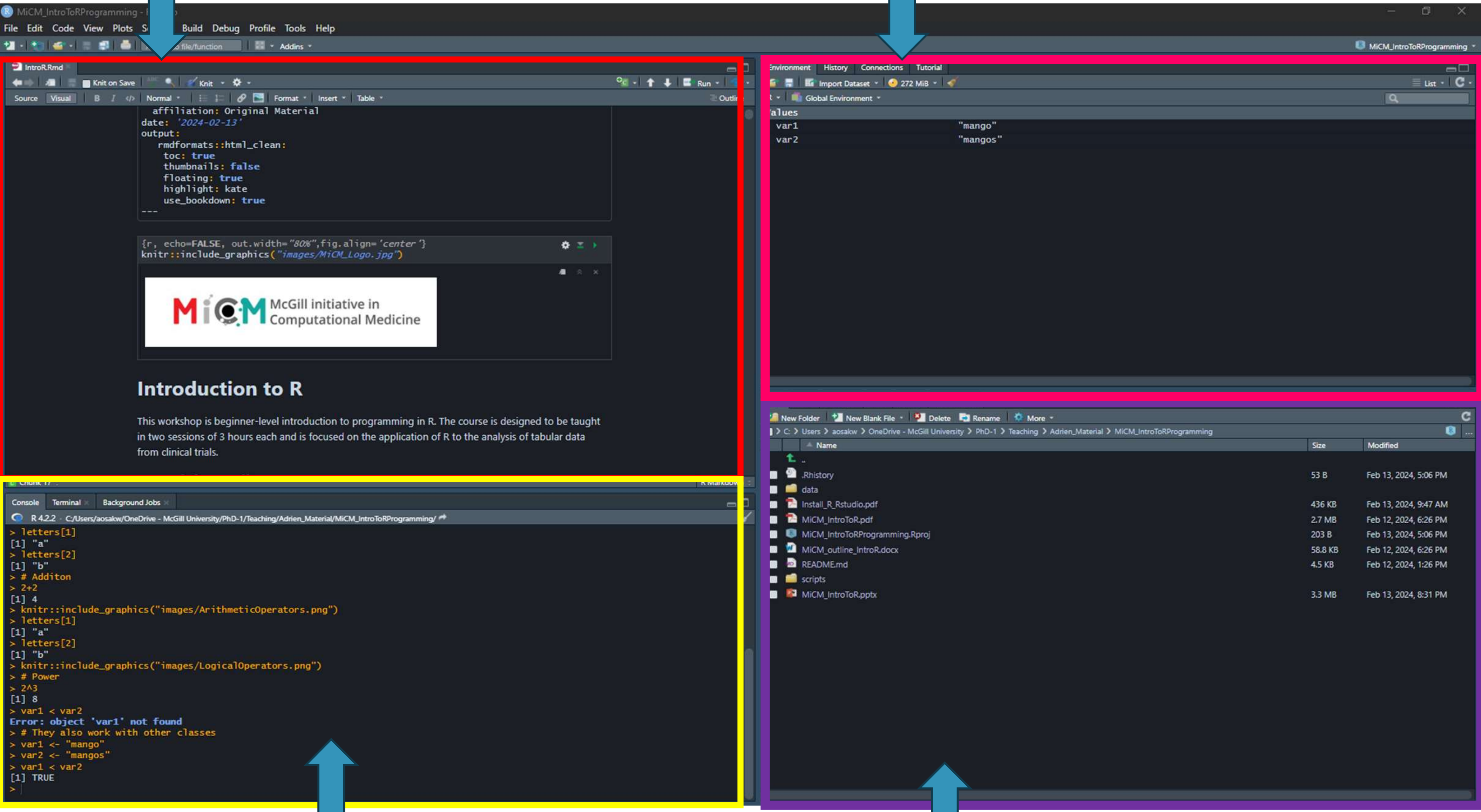
- R is the programming language
  - The actual code we execute
  - Developed at the University of Auckland
- RStudio is an **Integrated Development Environment (IDE)**
  - A GUI software to develop and execute R code
  - Developed by Posit



```
PS C:\Program Files\R\R-4.2.2\bin> .\R.exe

R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> print('hello world')
[1] "hello world"
> 
```

Source Pane

Environment Pane


MiCM_IntroToRProgramming
File Edit Code View Plots Build Debug Profile Tools Help
Addins
MiCM_IntroToRProgramming
IntroR.Rmd
Knit on Save
Knit
Run
Source Visual
Normal
Format Insert Table
Environment History Connections Tutorial
Import Dataset
272 MiB
List
Global Environment
New Folder New Blank File Delete Rename More
Console Terminal Background Jobs

```
  affiliation: Original Material
date: '2024-02-13'
output:
   rmdformats::html_clean:
    toc: true
    thumbnails: false
    floating: true
    highlight: kate
    use_bookdown: true
---
```

```
{r, echo=FALSE, out.width="80%",fig.align='center'}
knitr::include_graphics("images/MiCM_Logo.jpg")
```

McGill initiative in Computational Medicine

Introduction to R

This workshop is beginner-level introduction to programming in R. The course is designed to be taught in two sessions of 3 hours each and is focused on the application of R to the analysis of tabular data from clinical trials.

```
Values
 var1                           "mango"
 var2                           "mangos"
```

R 4.2.2 · C:/Users/aosakw/OneDrive - McGill University/PhD-1/Teaching/Adrien_Material/MiCM_IntroToRProgramming/

```
> letters[1]
[1] "a"
> letters[2]
[1] "b"
> # Additon
> 2+2
[1] 4
> knitr::include_graphics("images/ArithmeticOperators.png")
> letters[1]
[1] "a"
> letters[2]
[1] "b"
> knitr::include_graphics("images/LogicalOperators.png")
> # Power
> 2^3
[1] 8
> var1 < var2
Error: object 'var1' not found
> # They also work with other classes
> var1 <- "mango"
> var2 <- "mangos"
> var1 < var2
[1] TRUE
> 
```

C: › Users › aosakw › OneDrive - McGill University › PhD-1 › Teaching › Adrien_Material › MiCM_IntroToRProgramming

Name	Size	Modified
..		
.Rhistory	53 B	Feb 13, 2024, 5:06 PM
data		
Install_R_Rstudio.pdf	436 KB	Feb 13, 2024, 9:47 AM
MiCM_IntroToR.pdf	2.7 MB	Feb 12, 2024, 6:26 PM
MiCM_IntroToRProgramming.Rproj	203 B	Feb 13, 2024, 5:06 PM
MiCM_outline_IntroR.docx	58.8 KB	Feb 12, 2024, 6:26 PM
README.md	4.5 KB	Feb 12, 2024, 1:26 PM
scripts		
MiCM_IntroToR.pptx	3.3 MB	Feb 13, 2024, 8:31 PM

Console Pane

Files/Plot/Help Pane

Arithmetic operators

Addition	+
Subtraction	-
Division	/
Power	^
Scalar multiplication	*
Matrix multiplication	%*%

Syntax operators

Comment line	#
Assignation	<-
Access content	$
Equal	=

Logical operators

Equal	==
Not equal	!=
Greater than	>
Greater than or equal to	>=
Less than	<
Less than or equal to	<=
contains	%in%
x AND y	x & y
x OR y	x | y
NOT x	!x

Data types and data structures

Learning objectives

- Understand the differences between classes, objects and data types in R
- Create objects of different types
- Subset and index objects
-
- Learn and use vectorized operations

Atomic Classes

Also called data types

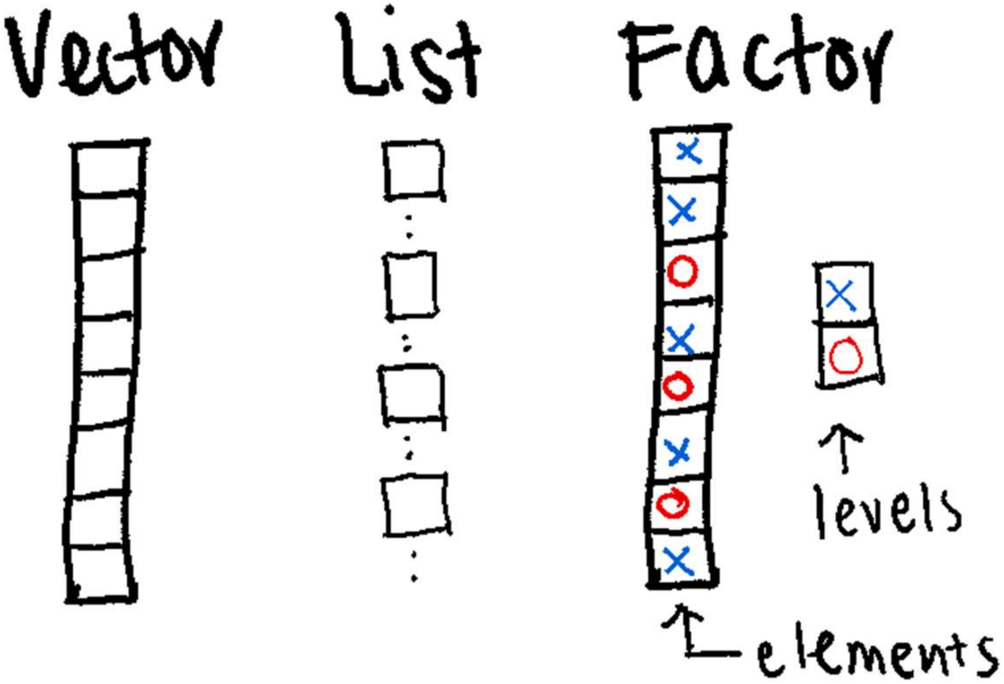
Character	A,b,c,d,e,..
Numeric (real numbers)	1.00,2.00,… Inf, NaN
Integer	1L,2L,3L,4L,….
Complex	2i
Logical (True/False)	TRUE,FALSE
Missing Value	NA

Objects

Also called data structures

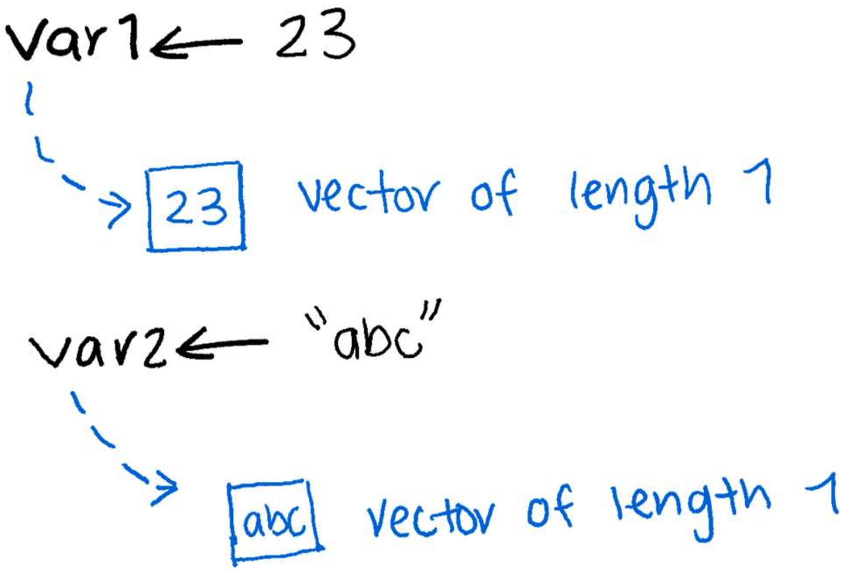
Vector	Only elements of the same class
List	Elements of any class
Factor	Categorical data
Matrix	Elements of the same class in 2D
Data frame	Elements of multiple classes in 2D
NULL	Empty object

One dimension

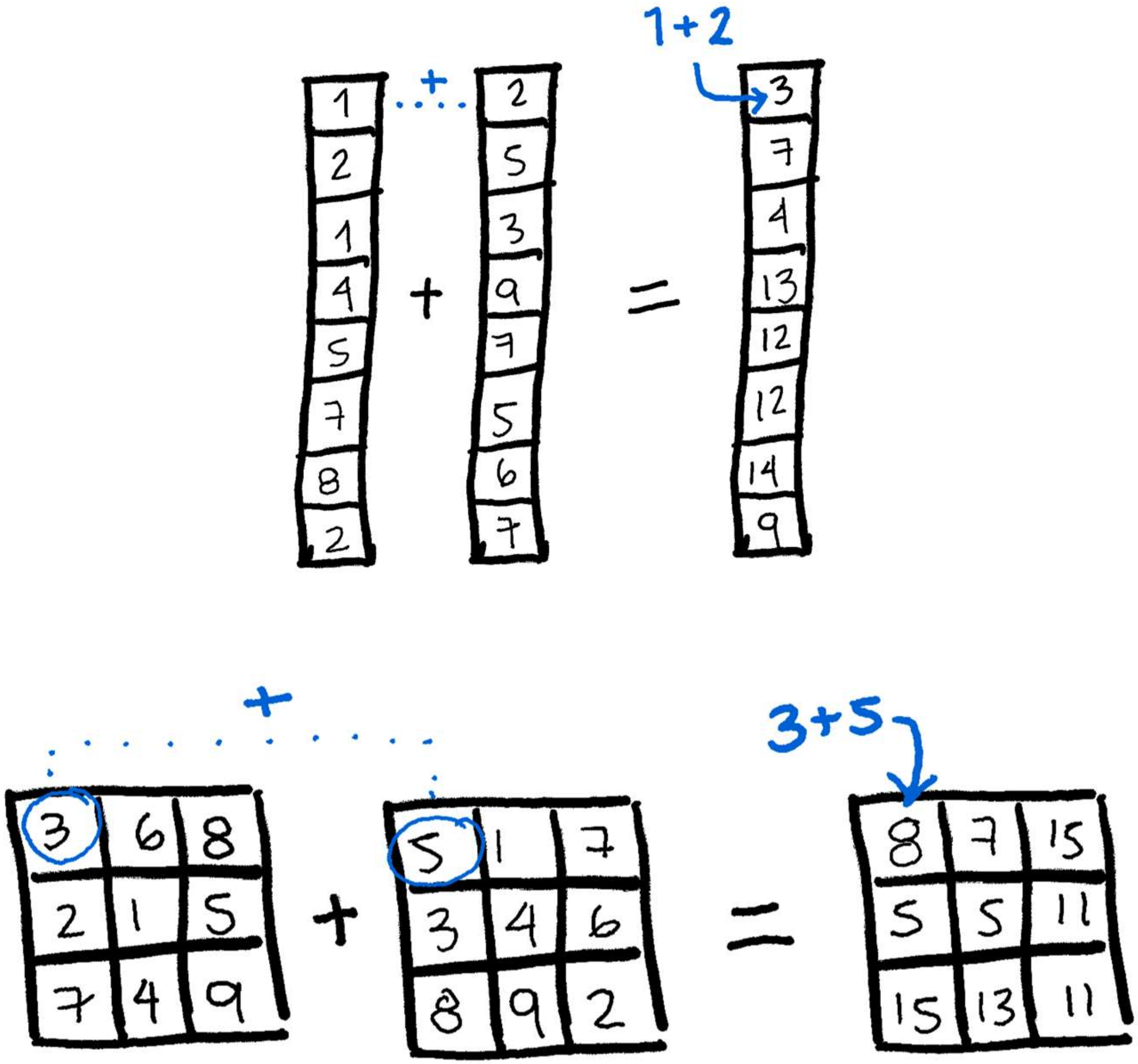


Vectors

- Can only contain objects of the **same class**
- Most basic type of R object
- Variables are vectors



Vectorized operations



$$\begin{bmatrix}1\\2\\1\\4\\5\\7\\8\\2\end{bmatrix} + \begin{bmatrix}2\\5\\3\\9\\7\\5\\6\\7\end{bmatrix} = \begin{bmatrix}3\\7\\4\\13\\12\\12\\14\\9\end{bmatrix}$$

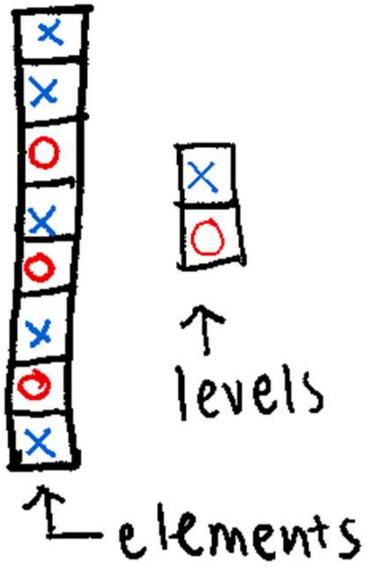
$$\begin{bmatrix}3&6&8\\2&1&5\\7&4&9\end{bmatrix} + \begin{bmatrix}5&1&7\\3&4&6\\8&9&2\end{bmatrix} = \begin{bmatrix}8&7&15\\5&5&11\\15&13&11\end{bmatrix}$$

Lists

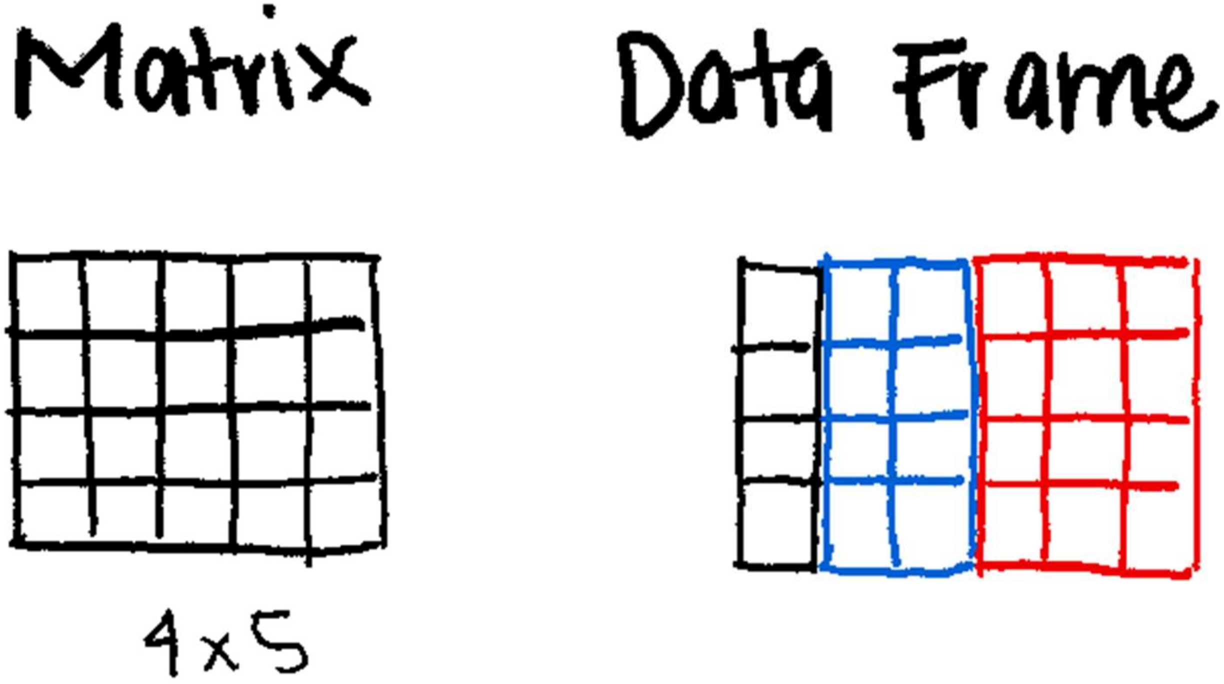
- Can contain objects of multiple classes
- Very important data type in R
- Extremely powerful when combined with some built-in functions

Factors

- Useful when for categorical data
- Can have implicit order, if needed
- Each **element** has a label or **level**
- They are important in statistical modelling and plotting with ggplot
- Some operations behave differently on factors



Multiple dimensions



Break

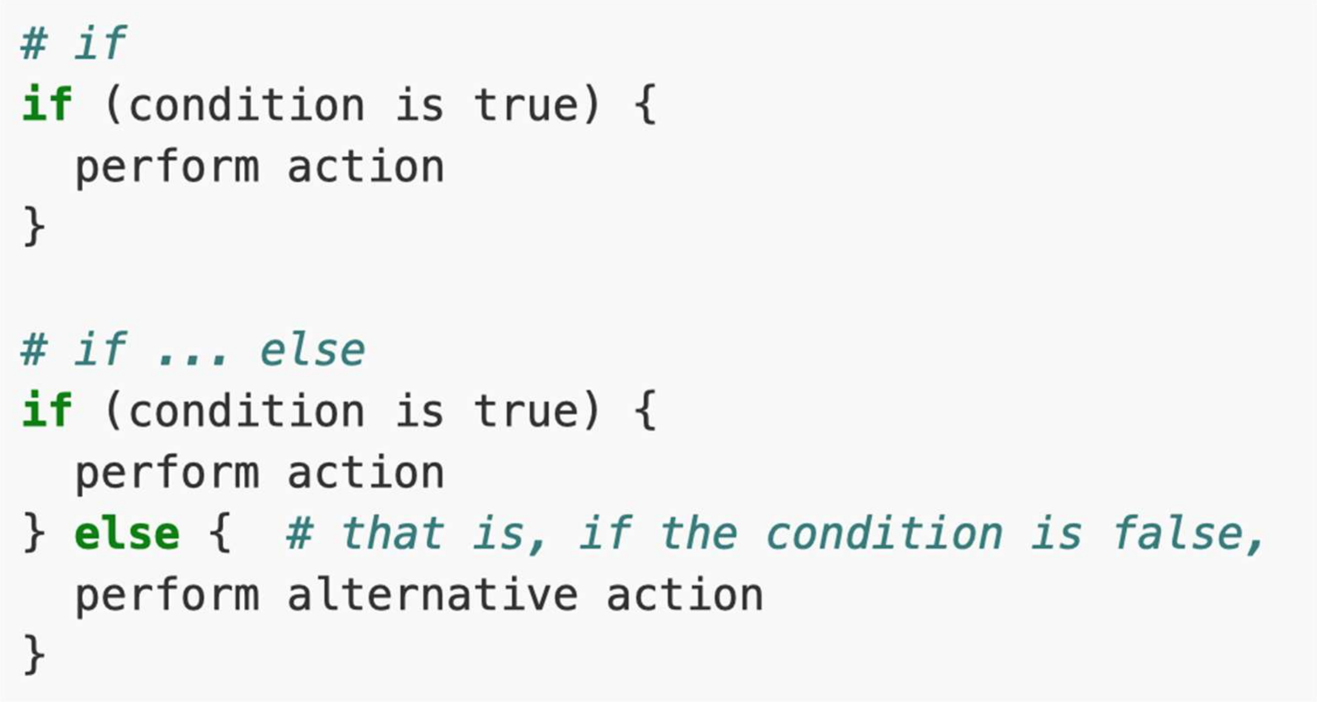
Control structures and functions

Learning objectives:

- Understand the concept of environments in R
- Create new functions
- Implement conditional statements
- Implement a for loop to iterate over a list of files

Conditional statements

- When we want a set of actions to be executed only if certain conditions are met

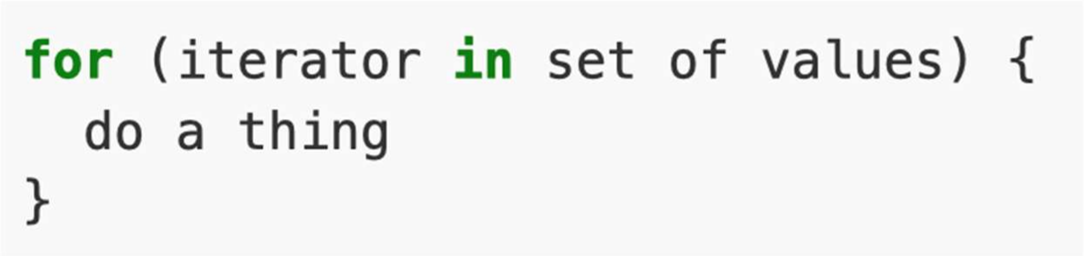


```
# if
if (condition is true) {
  perform action
}

# if ... else
if (condition is true) {
  perform action
} else {  # that is, if the condition is false,
  perform alternative action
}
```

https://swcarpentry.github.io/r-novice-gapminder/07-control-flow

For loop

- Repeat a set of operations a certain number of times

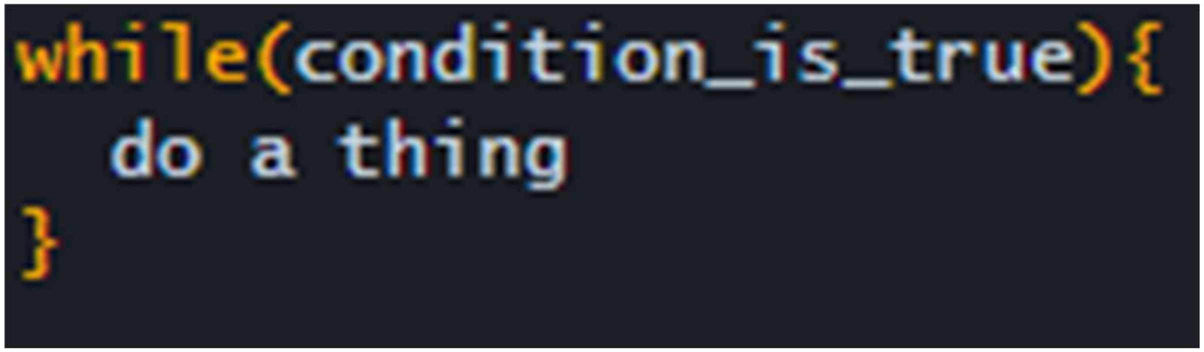


```
for (iterator in set of values) {
  do a thing
}
```

https://swcarpentry.github.io/r-novice-gapminder/07-control-flow

While loop

- Repeat a set of operations until a condition is no longer met



```
while(condition_is_true){
  do a thing
}
```

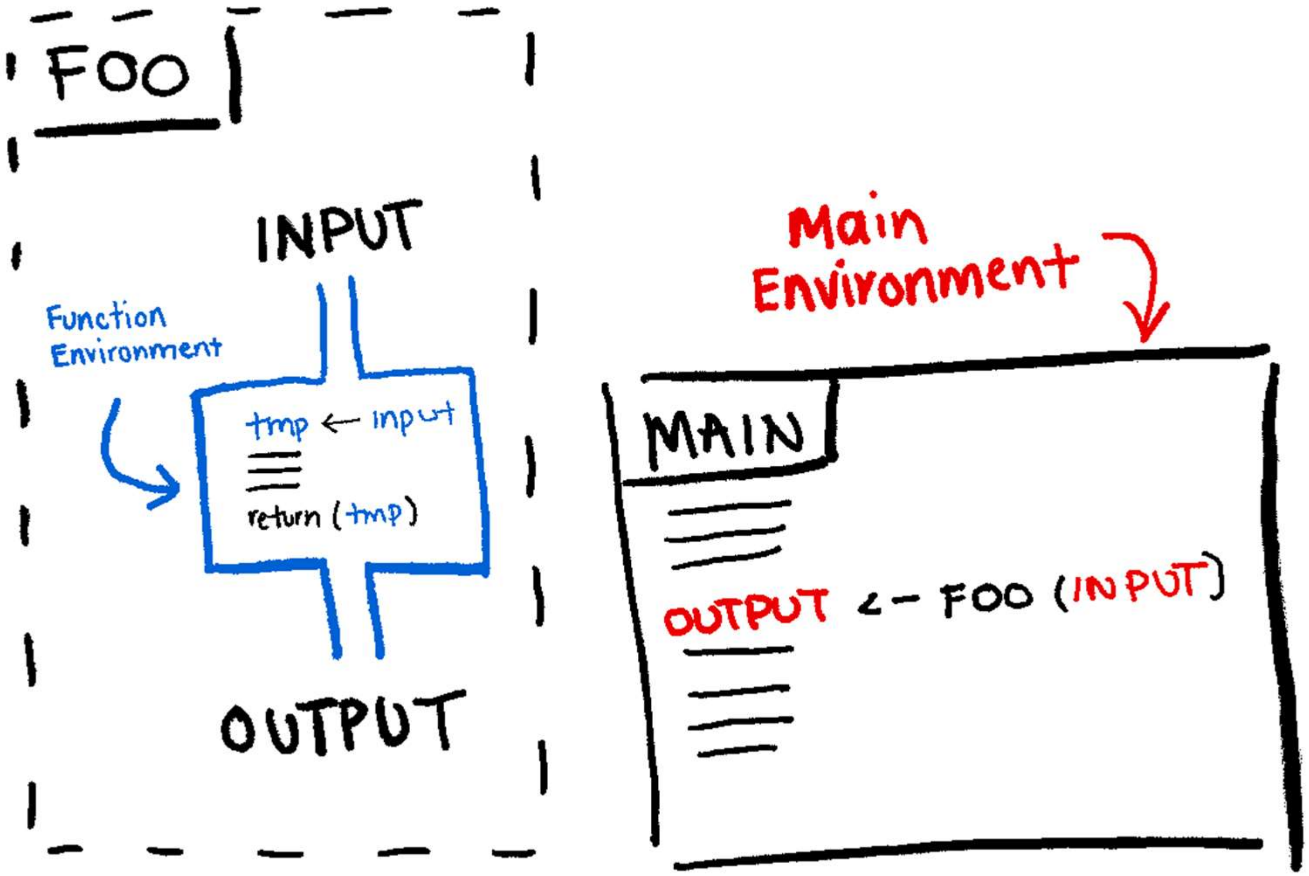
https://swcarpentry.github.io/r-novice-gapminder/07-control-flow

What if base R is not enough?

- Sometimes your analysis requires tools that are not available in base R
- Two options:
  1. **Create new functions**

  2. **Packages** provide a way to incorporate methods and functions from

Functions and environments

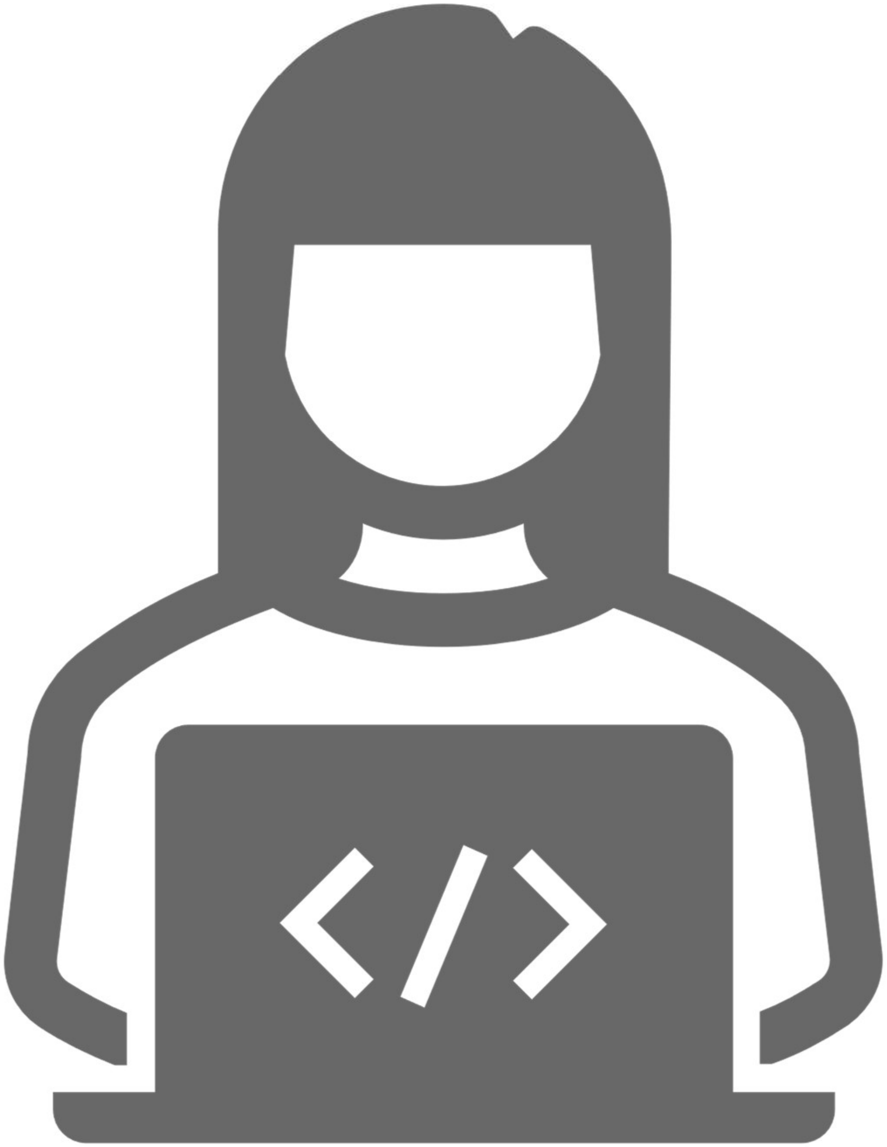


Pass by value and scope

- When we pass an object to a function, a copy of it is created internally
- The changes made inside the function won’t modify the original object we passed to it
- Any variables created inside the function will only exist during the function’s execution time

Packages

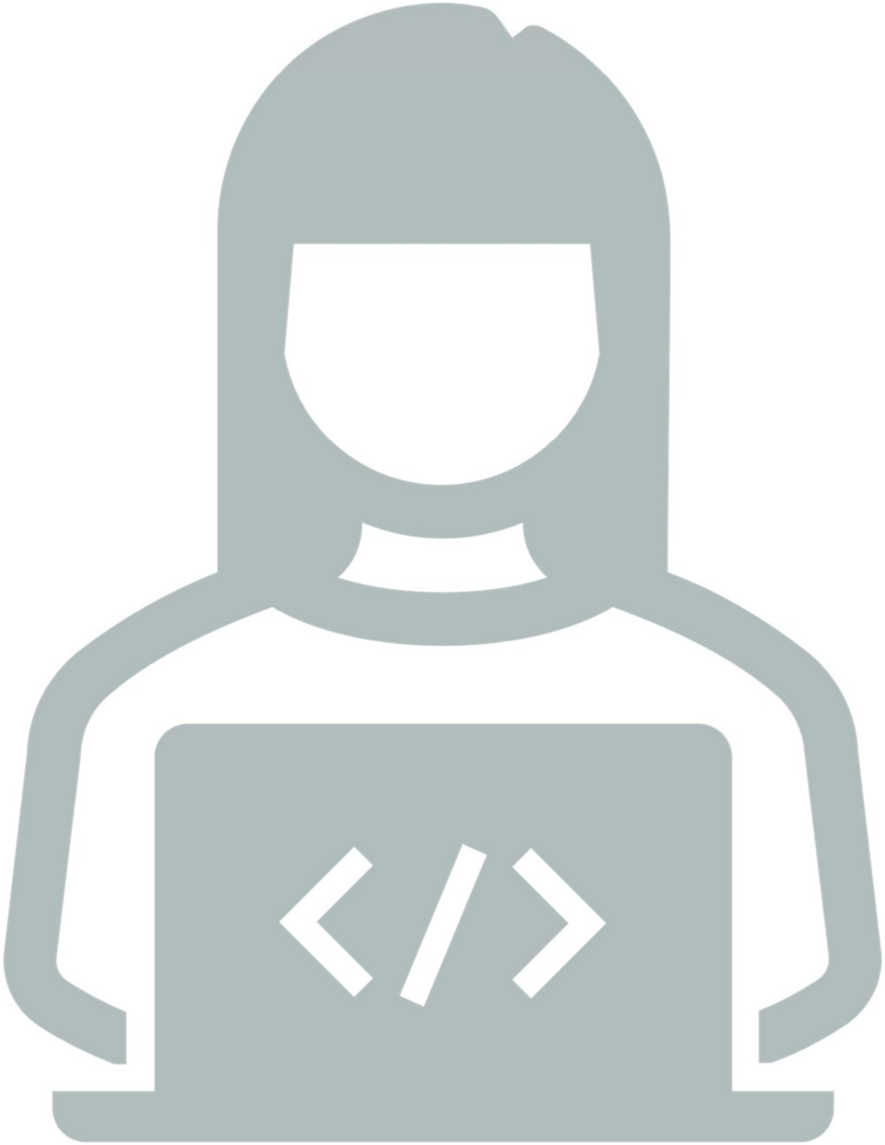
- Packages are a way for users to share methods they have developed
- Incorporate novel methods, datasets, or visualization tools
- Downloaded from many places:
  - **Comprehensive R Archive Network (CRAN)**
  - **Bioconductor**
  - GitHub, Bitbucket etc.



Basic data manipulation

Learning objectives:

- Learn how to read/write data to/from files with different formats (.tsv, .csv)
- Familiarize with basic operations of data frames
- Index and subset data frames using base R functions
- Manipulate specific data frame columns
- Joining by columns and rows

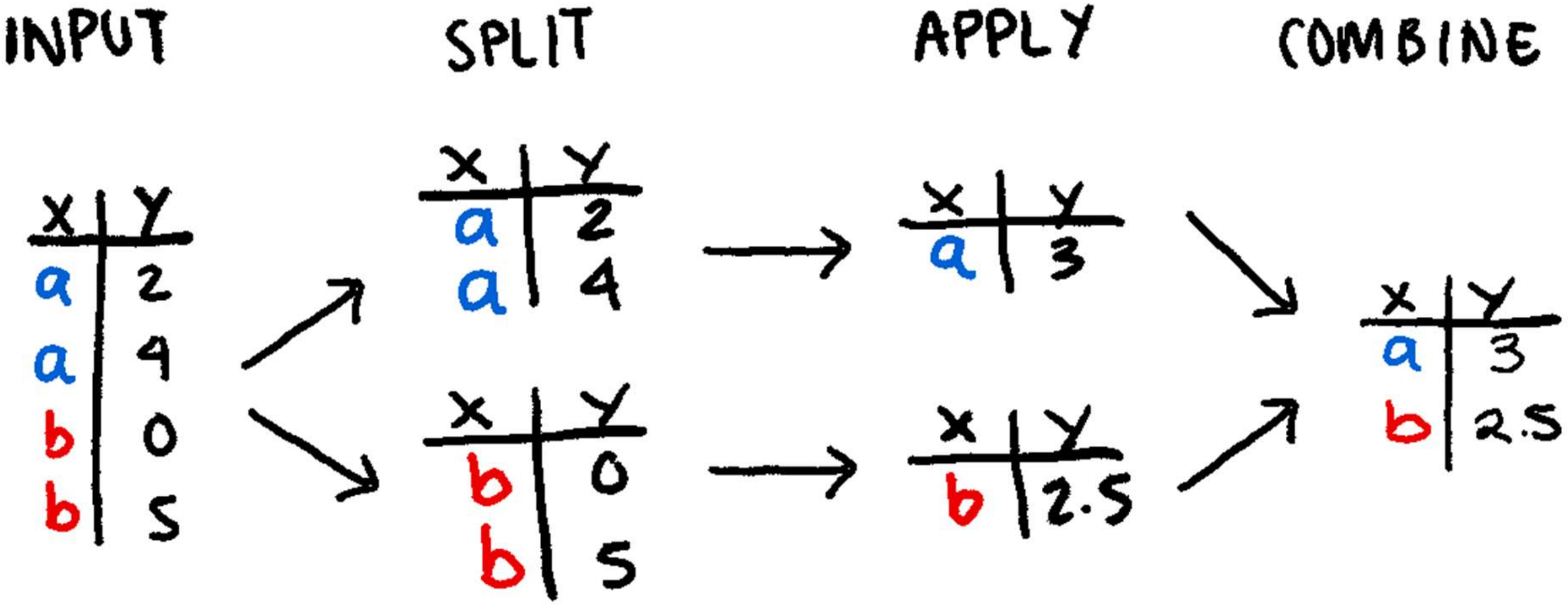


Advanced data manipulation

Learning objectives:

- Become familiar with the dplyr syntax
- Create pipes with the operator *%>%*
- Perform operations on data frames using dplyr and tidyr functions
- Implement functions from other external packages

Split-Apply-Combine problem



X	Y
a	2
a	4
b	0
b	5

X	Y
a	2
a	4

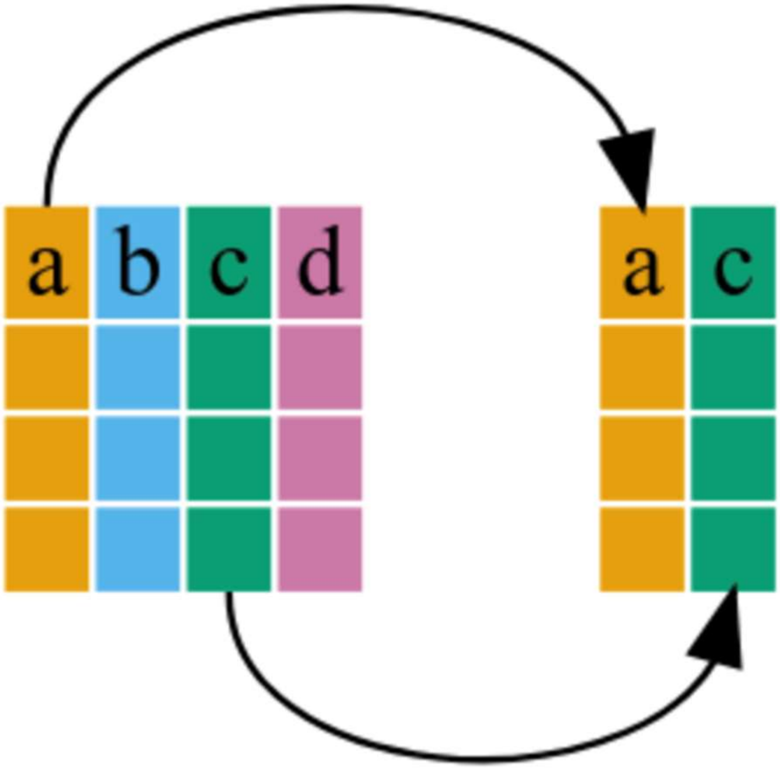
X	Y
b	0
b	5

X	Y
a	3

X	Y
b	2.5

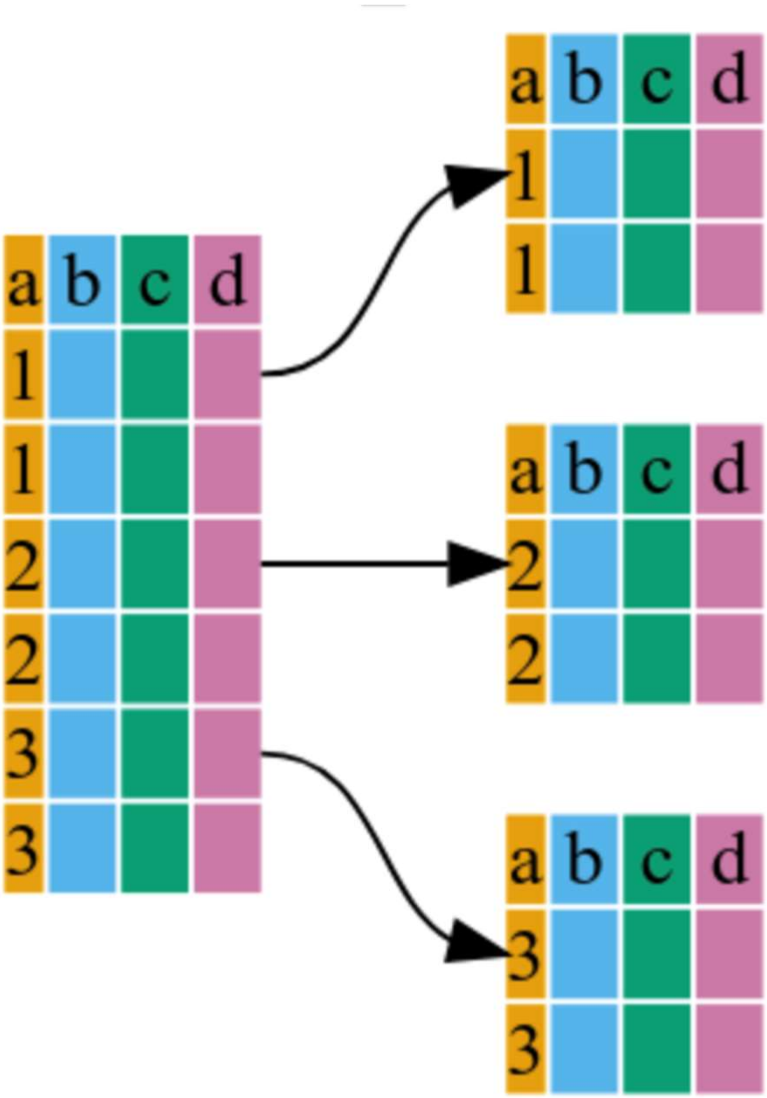
X	Y
a	3
b	2.5

Select



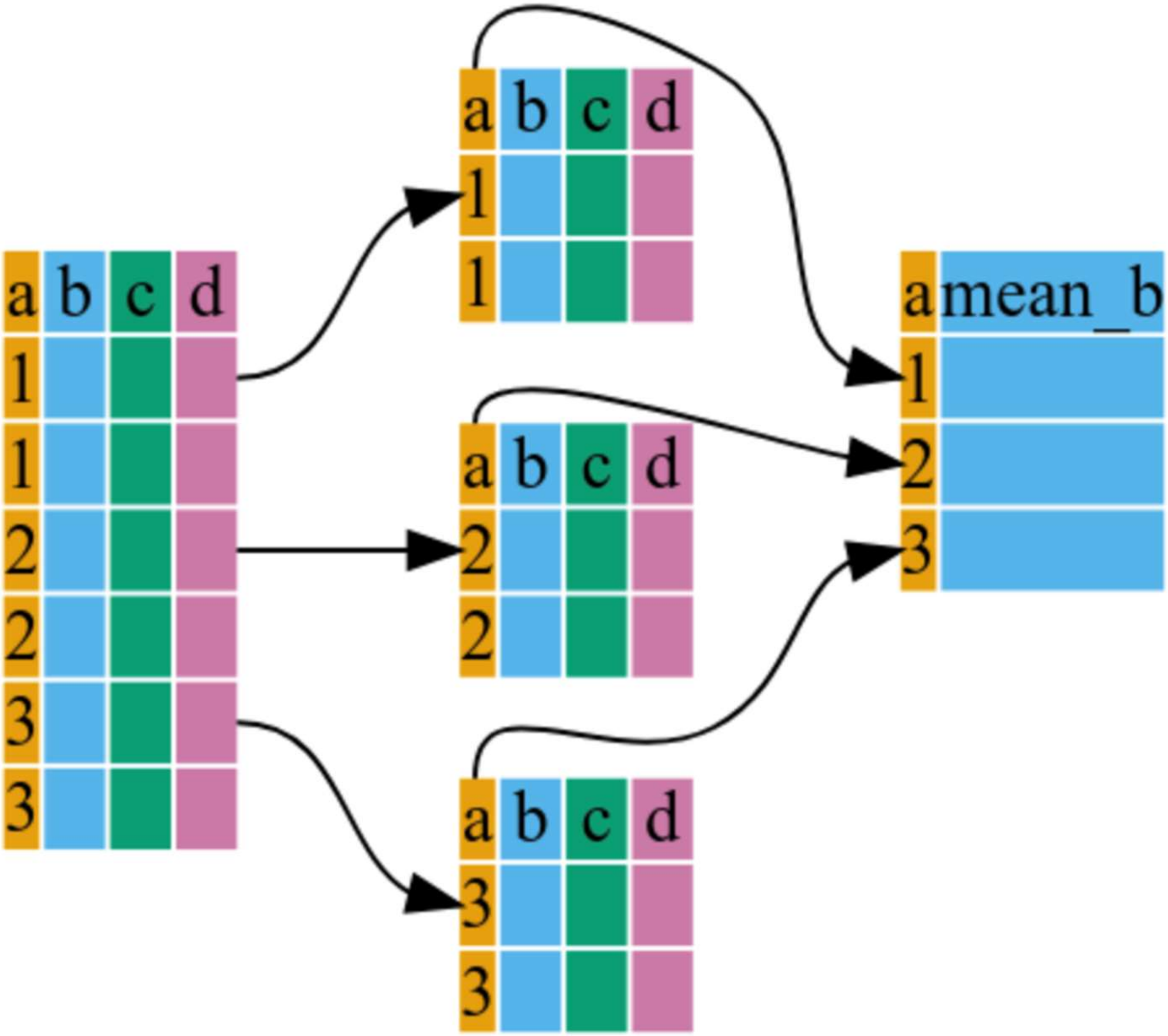
https://swcarpentry.github.io/r-novice-gapminder/13-dplyr/index.html

Group by

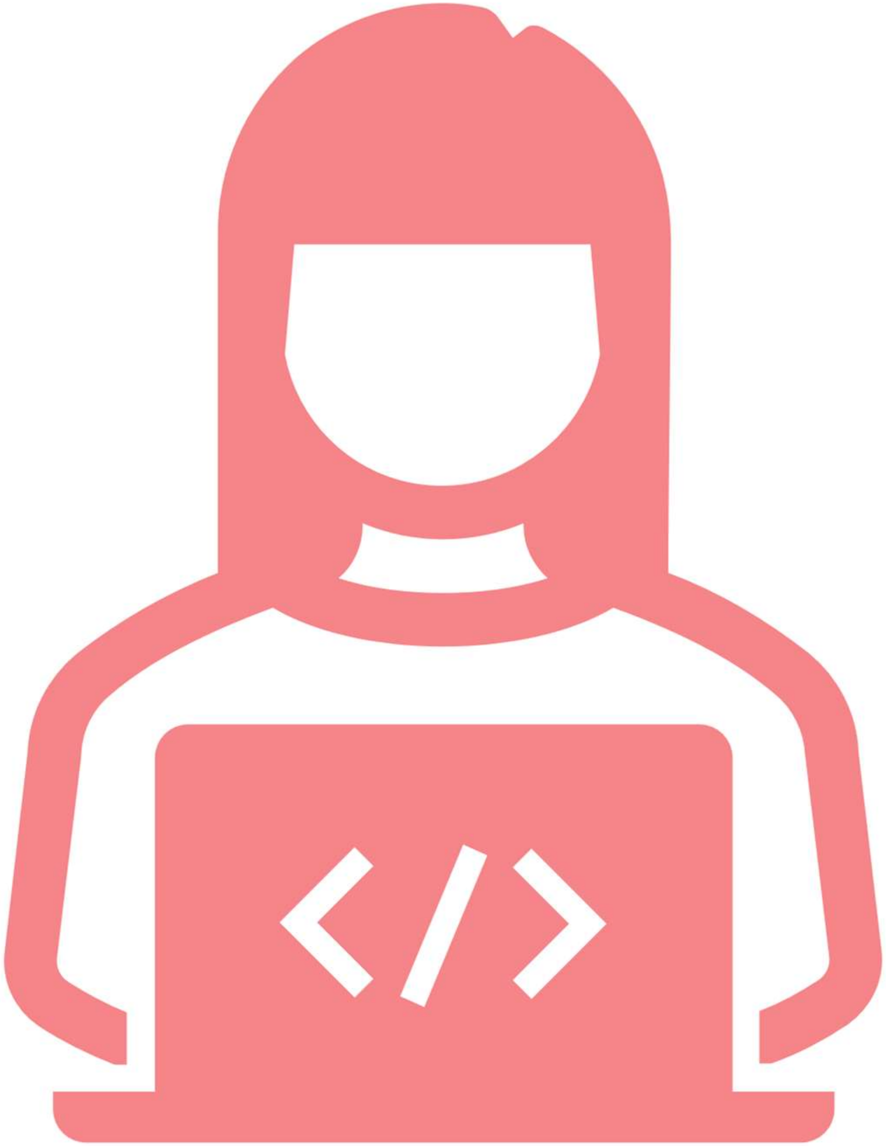


https://swcarpentry.github.io/r-novice-gapminder/13-dplyr/index.html

Summarize



https://swcarpentry.github.io/r-novice-gapminder/13-dplyr/index.html



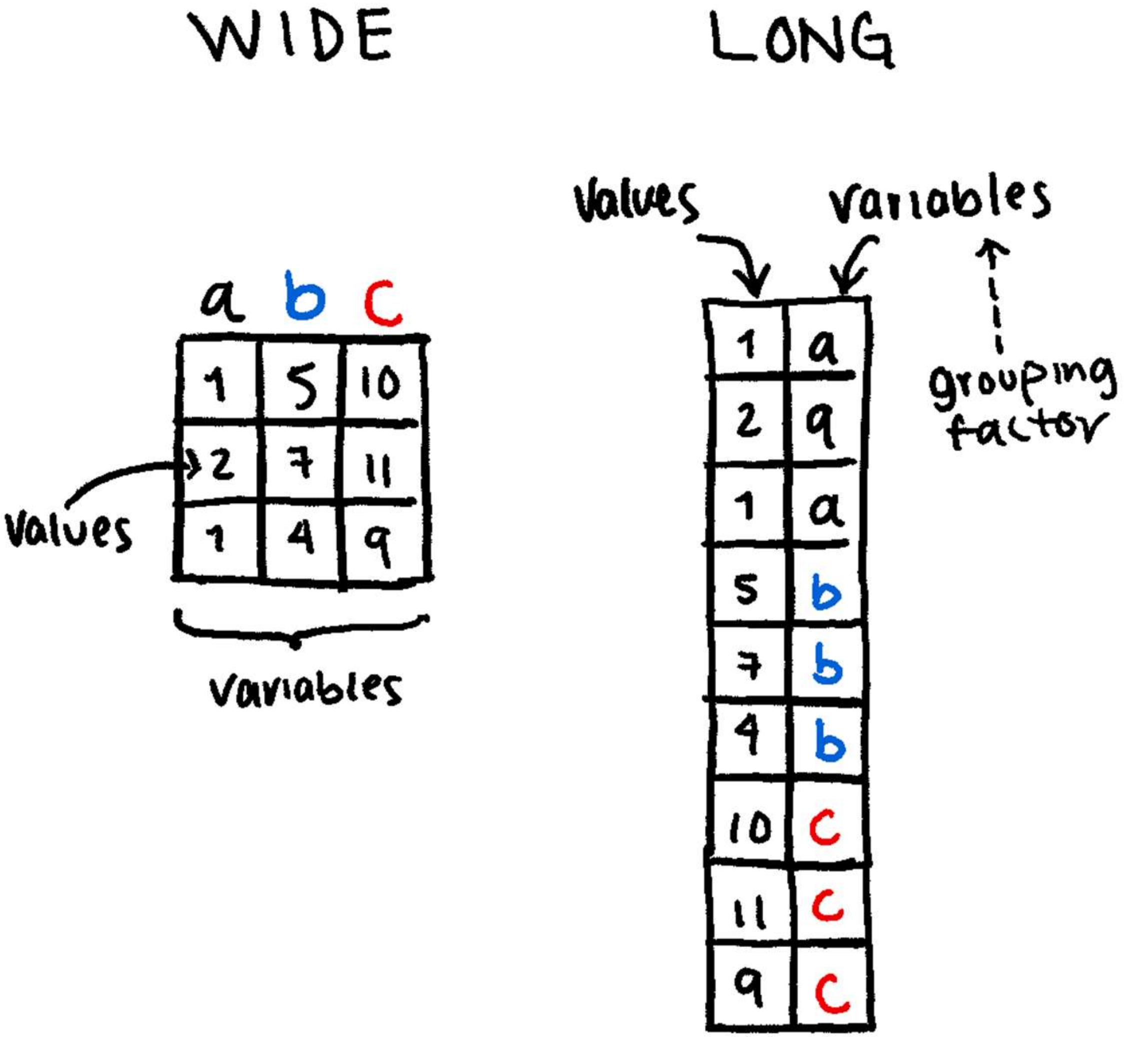
Break

Generating visual outputs

Learning objectives:

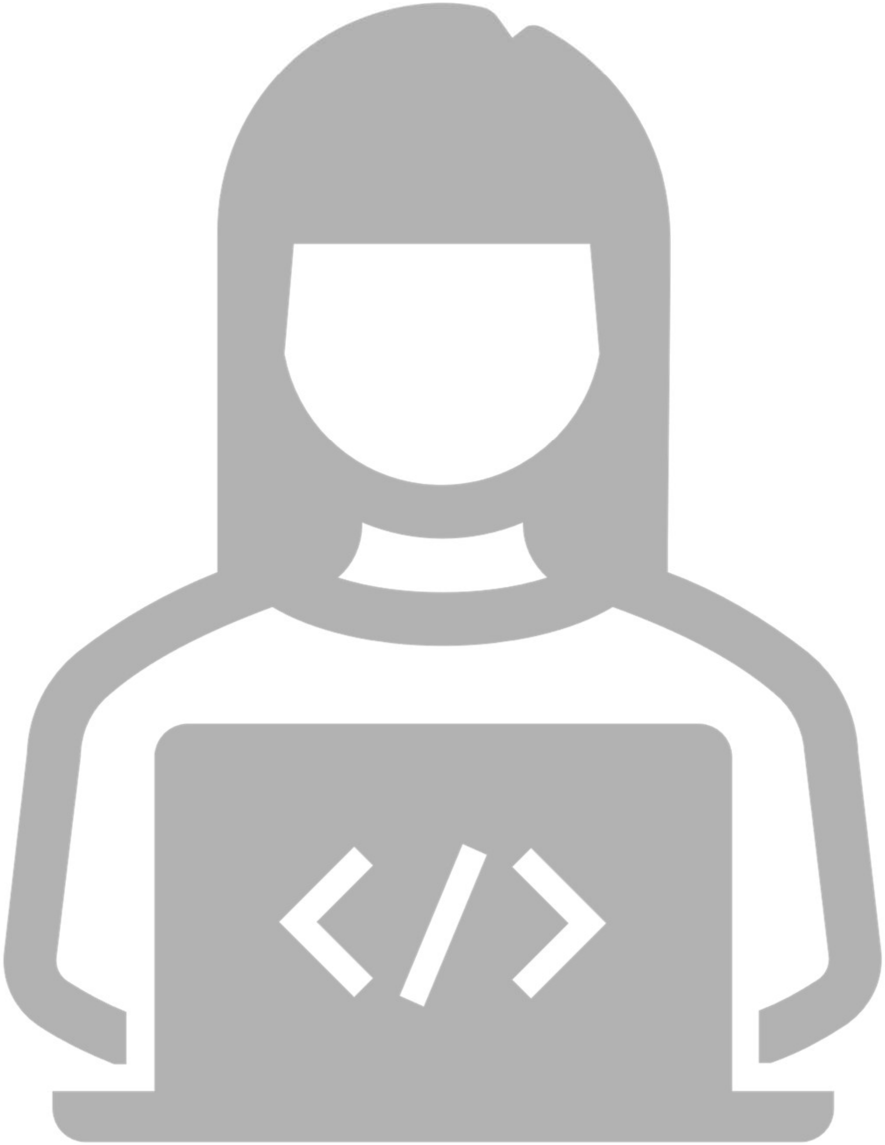
- Create basic plots using base R functions
- Understand the connection between data frames and ggplot2
- Create basic graphs with ggplot2
- Use factors to customize graphics in ggplot2
- Learn about RMarkdown syntax to create reports
- Get familiar with existing RMarkdown templates

Formatting data for ggplot



a	b	c
1	5	10
2	7	11
1	4	9

1	a
2	a
1	a
5	b
7	b
4	b
10	c
11	c
9	c



Activity: Analyzing a medical data set

Learning objectives:

- Familiarize with a real-life use case of R
- Apply the knowledge from previous modules to create an analysis pipeline

COVID testing dataset

Details

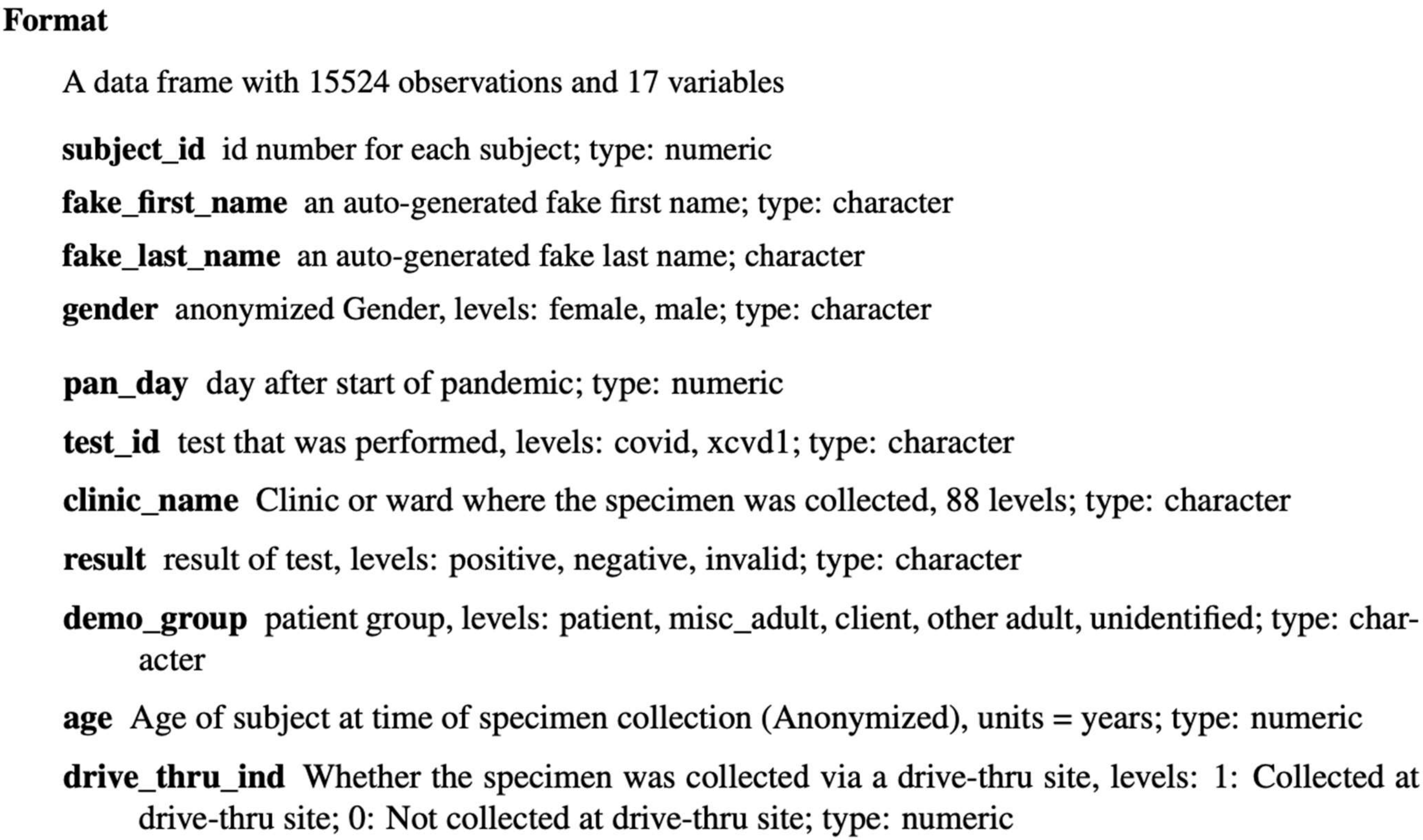
Data on testing for SARS-CoV2 from days 4-107 of the COVID pandemic in **2020**. CHOP is a pediatric hospital in Philadelphia, Pennsylvania, USA. These data have been anonymized, time- shifted, and permuted.

Text taken from https://cran.r-project.org/web/packages/medicaldata/medicaldata.pdf

The dataset

Documentation

- Part of the medicaldata package
- https://htmlpreview.github.io/?https://github.com/higgi13425/medicaldata/blob/master/man/description_docs/covid_desc.html
- https://htmlpreview.github.io/?https://github.com/higgi13425/medicaldata/blob/master/man/codebooks/covid_testing_codebook.html



Format

A data frame with 15524 observations and 17 variables

subject_id id number for each subject; type: numeric

fake_first_name an auto-generated fake first name; type: character

fake_last_name an auto-generated fake last name; character

gender anonymized Gender, levels: female, male; type: character

pan_day day after start of pandemic; type: numeric

test_id test that was performed, levels: covid, xcvd1; type: character

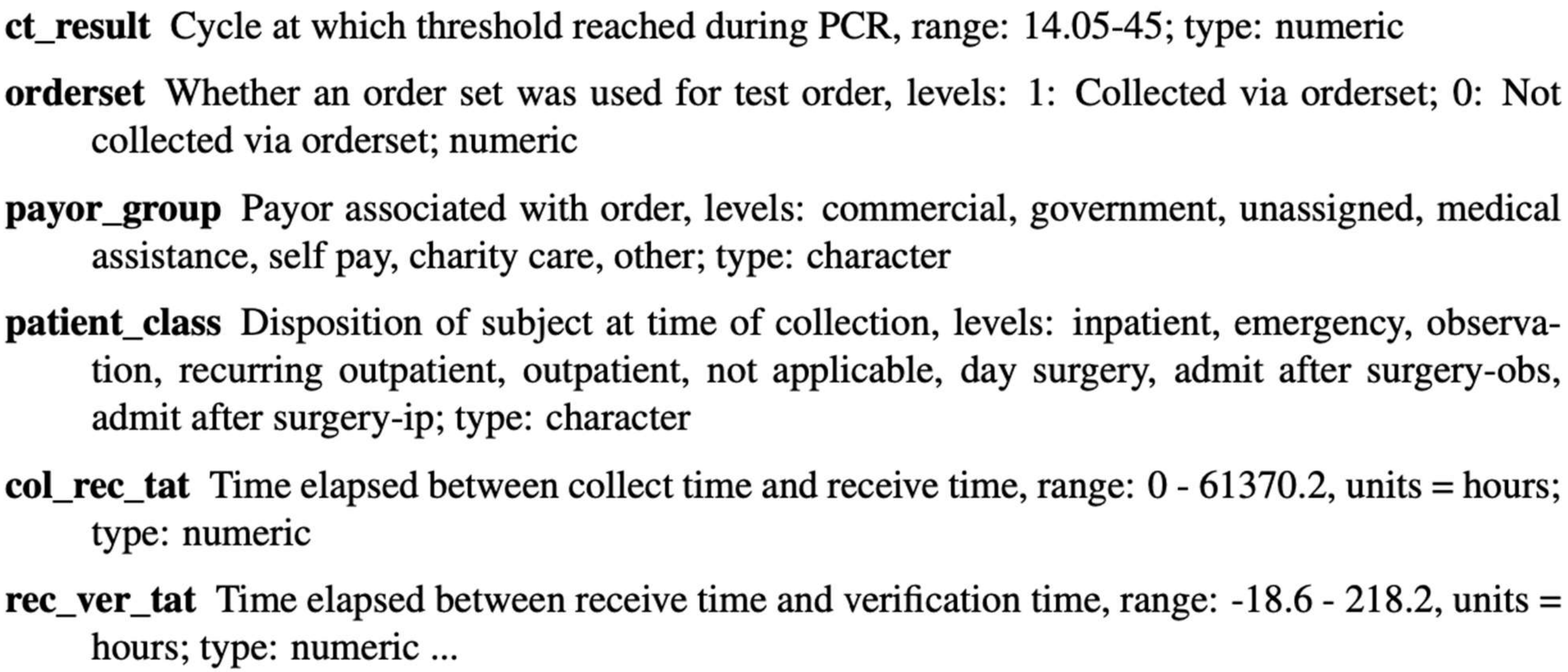
clinic_name Clinic or ward where the specimen was collected, 88 levels; type: character

result result of test, levels: positive, negative, invalid; type: character

demo_group patient group, levels: patient, misc_adult, client, other adult, unidentified; type: character

age Age of subject at time of specimen collection (Anonymized), units = years; type: numeric

drive_thru_ind Whether the specimen was collected via a drive-thru site, levels: 1: Collected at drive-thru site; 0: Not collected at drive-thru site; type: numeric



ct_result Cycle at which threshold reached during PCR, range: 14.05-45; type: numeric

orderset Whether an order set was used for test order, levels: 1: Collected via orderset; 0: Not collected via orderset; numeric

payor_group Payor associated with order, levels: commercial, government, unassigned, medical assistance, self pay, charity care, other; type: character

patient_class Disposition of subject at time of collection, levels: inpatient, emergency, observation, recurring outpatient, outpatient, not applicable, day surgery, admit after surgery-obs, admit after surgery-ip; type: character

col_rec_tat Time elapsed between collect time and receive time, range: 0 - 61370.2, units = hours; type: numeric

rec_ver_tat Time elapsed between receive time and verification time, range: -18.6 - 218.2, units = hours; type: numeric ...

Software development concepts

Learning objectives:

- Familiarize with general good coding practices
- Learn about documentation standards
- Things to avoid when programming in R
- Learn how to debug and troubleshoot

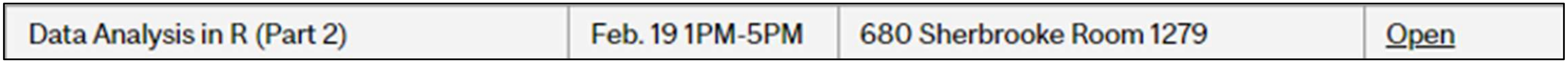
What we learned today:

- What is R
- Basic syntax, data types
- Data Manipulation and Visualization
- Package Installation

What’s next?

Data Analysis in R Workshop (Part 2)

- Data Wrangling
- Linear Regression & Statistical Analysis
- Classification



Data Analysis in R (Part 2)	Feb. 19 1PM-5PM	680 Sherbrooke Room 1279	Open

Thank you for attending!

1

Scan the QR code to confirm you attended today's workshop.

2

Fill out the feedback survey in the next 72h.

3

McGILL UNIVERSITY • CO-CURRICULAR RECORD

APPROVED

Get recognition for this workshop on your co-curricular record.

Useful links

- R software project
- RStudio Cheatsheet
- R ggplot2 Cheatsheet
- R dplyr Cheatsheet
- More resources