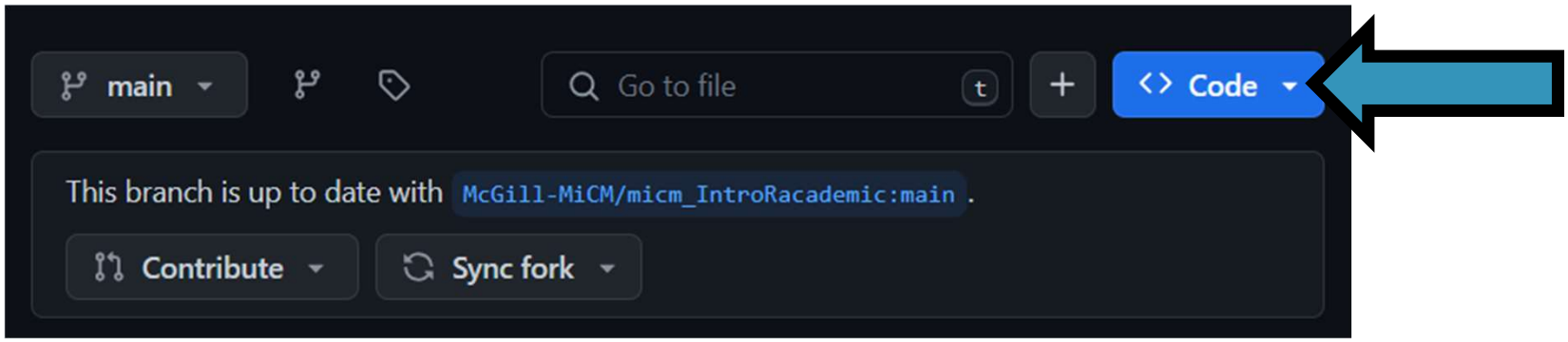
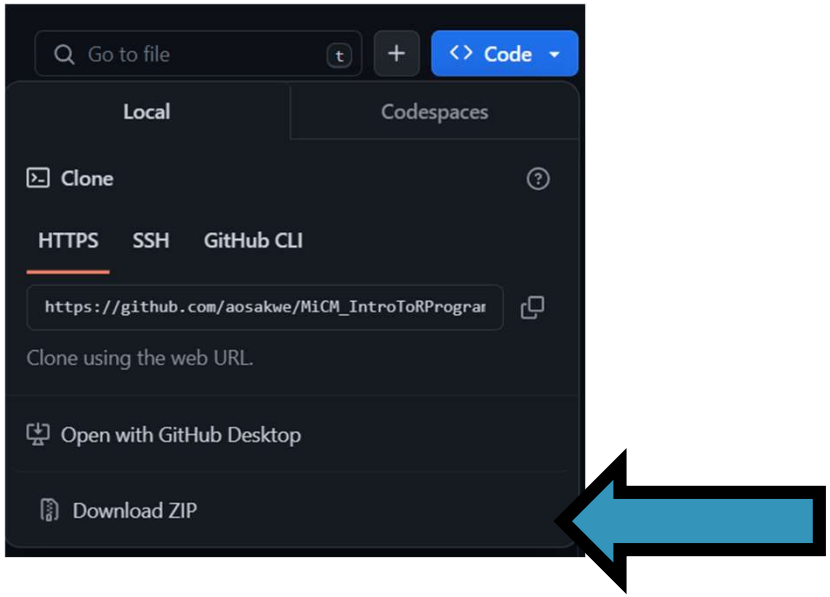
Download Workshop Materials

1. Go to https://github.com/aosakwe/MiCM_IntroToRProgramming
2. 



McGill initiative in Computational Medicine

Intro to programming in R

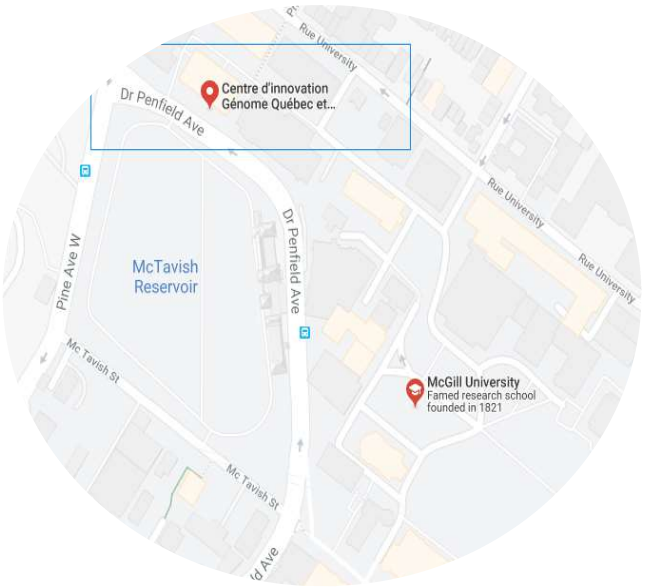
Lead: Adrien Osakwe
Facilitator: Bangli Cao

February 13, 2025
Materials adapted from Larisa M. Soto and Xiaoqi Xie

McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

McGill initiative in Computational Medicine

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



Location: 550 Sherbrooke West, Montreal, Quebec

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

Contact: workshop-micm@mcgill.ca

McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

Workshop	Date	Location	Registration
How to think in Code	Jan. 28 1PM-3PM	EDUC 133	Closed
Intro to Git & GitHub	Jan. 30 1PM-5PM	EDUC 133	Closed
Intro to Unix	Feb. 6 1PM-5PM	EDUC 133	Closed
Intro to Python (Part 1)	Feb. 11 1PM-5PM	EDUC 133	Closed
Intro to R (Part 1)	Feb. 13 1PM-5PM	EDUC 133	Closed
Exploring MATLAB	Feb. 18 1PM-5PM	EDUC 133	Open
Statistics in R (Part 2)	Feb. 20 1PM-5PM	EDUC 133	Open
Data Processing in Python	Feb. 25 1PM-5PM	EDUC 133	Open
Intro to Machine Learning	Mar. 13 1PM-5PM	EDUC 133	TBA
Intro to R (Part 1)	TBA	EDUC 133	TBA
Intro to Python (Part 1)	TBA	EDUC 133	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

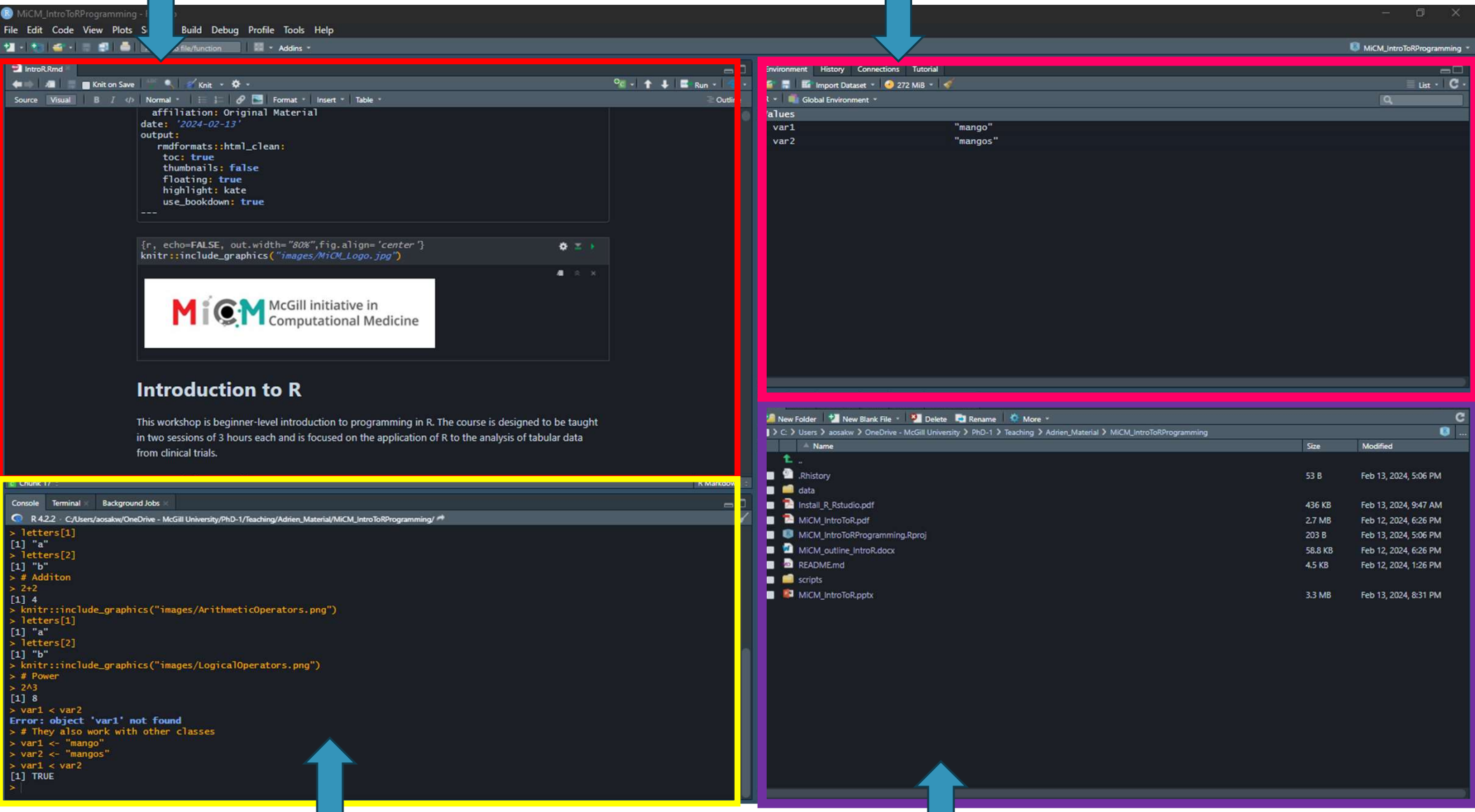
- R & RStudio are **different entities**

- 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



Console Pane

Files/Plot/Help Pane

McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

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

Arithmetic operators

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

Syntax operators

Comment line	#
Assignation	<-
Access content	$
Equal	=

Logical operators

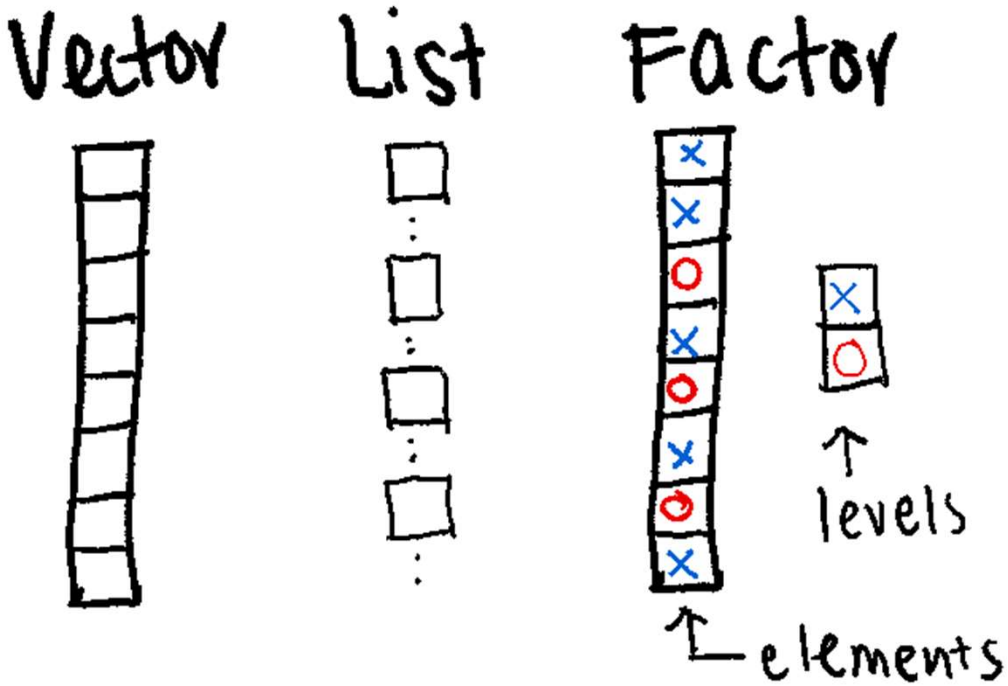
Equal	==
Not equal	!=
Greater than	>
Greater than or equal to	>=
Less than	<
Less than or equal to	<=
contains	%in%
x AND y	x & y
x OR y	x | y
NOT x	!x

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

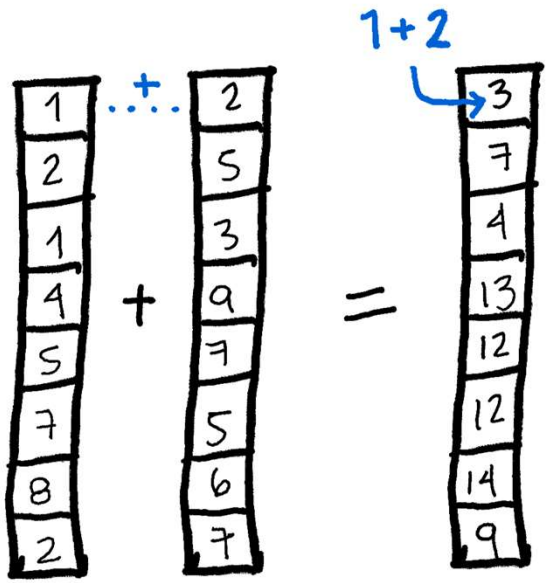
var1 ← 23

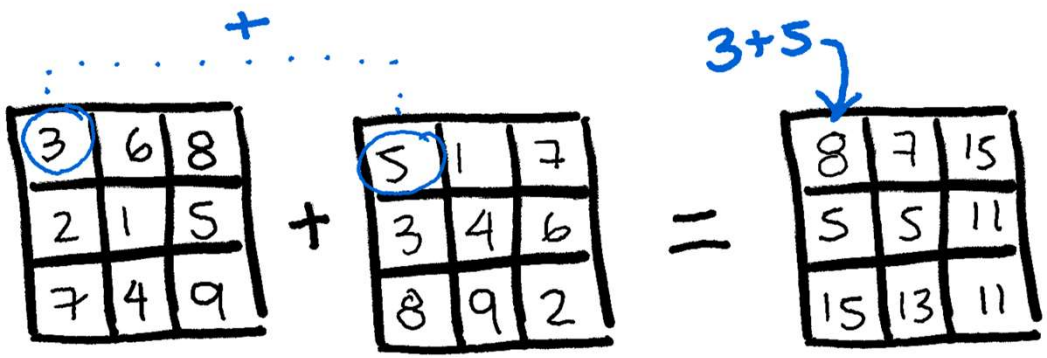
→ [23] vector of length 1

var2 ← "abc"

→ [abc] vector of length 1

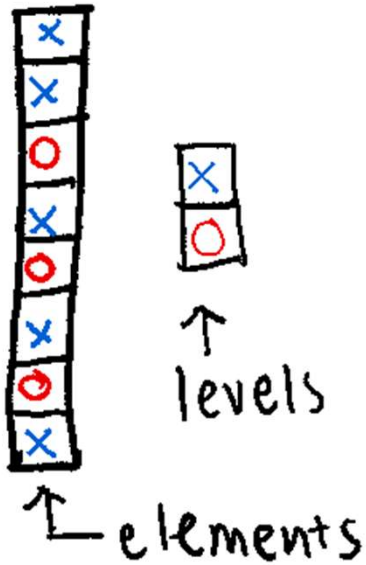
Vectorized operations





Factors

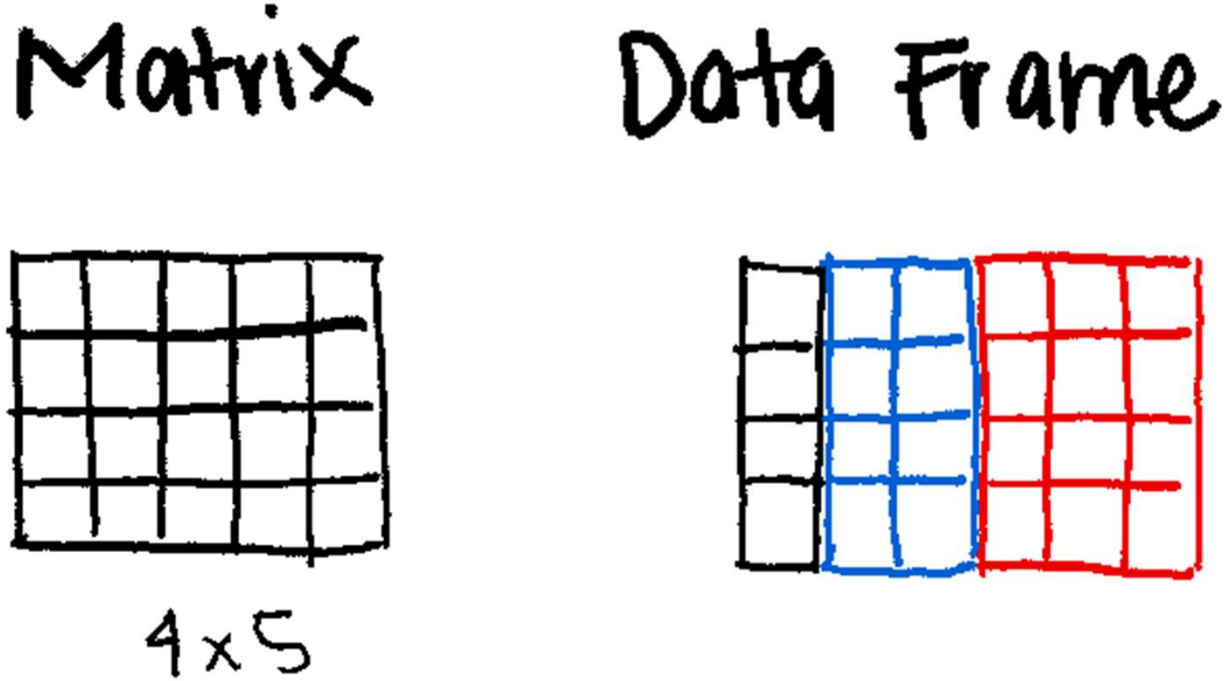
- Useful when for categorical data
- Can have implicit order, if needed
- Each **element** has a label or **level**
- They are important in statistical modelling and plotting with ggplot
- Some operations behave differently on factors



Lists

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

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
}
```

While loop

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

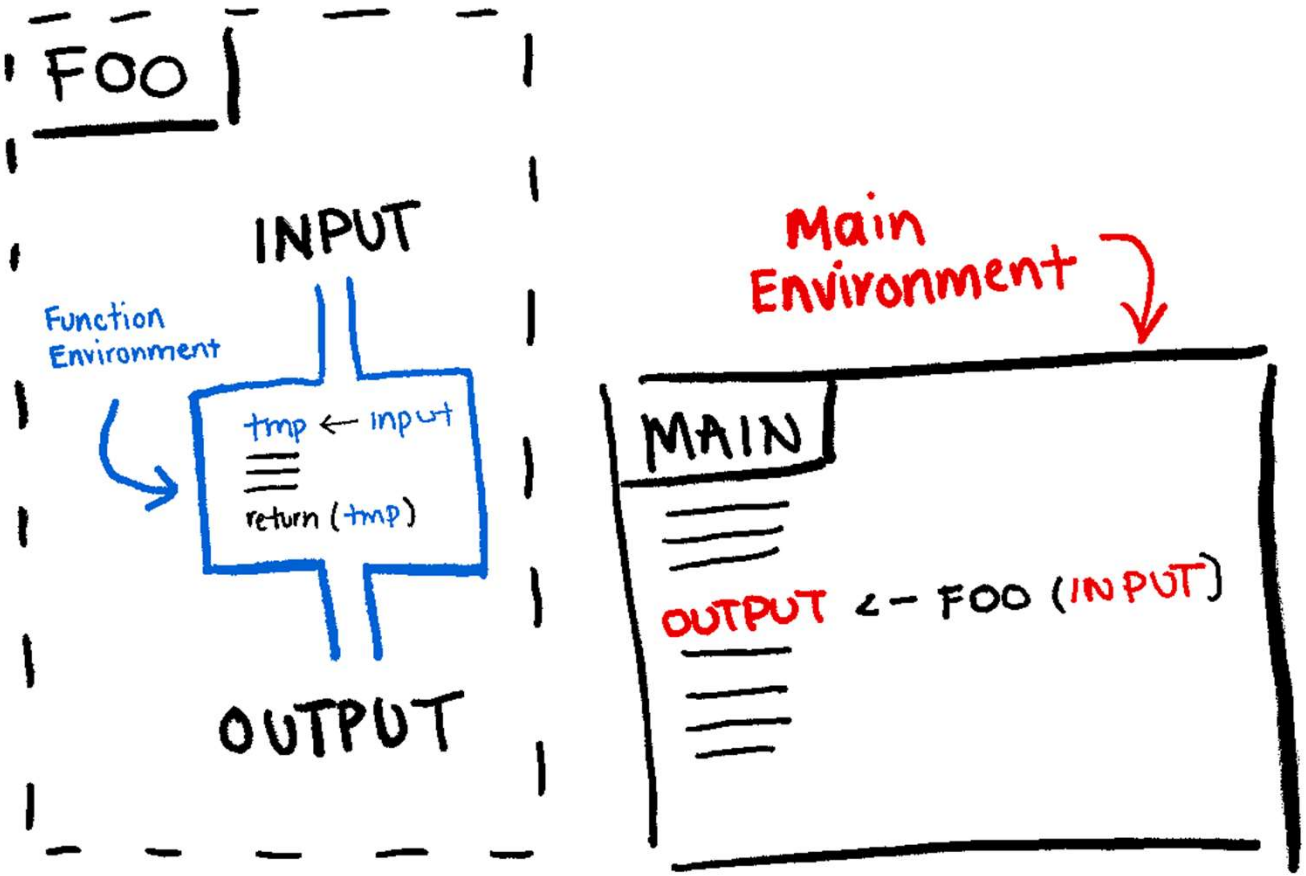
```
while(condition_is_true){
  do a thing
}
```

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



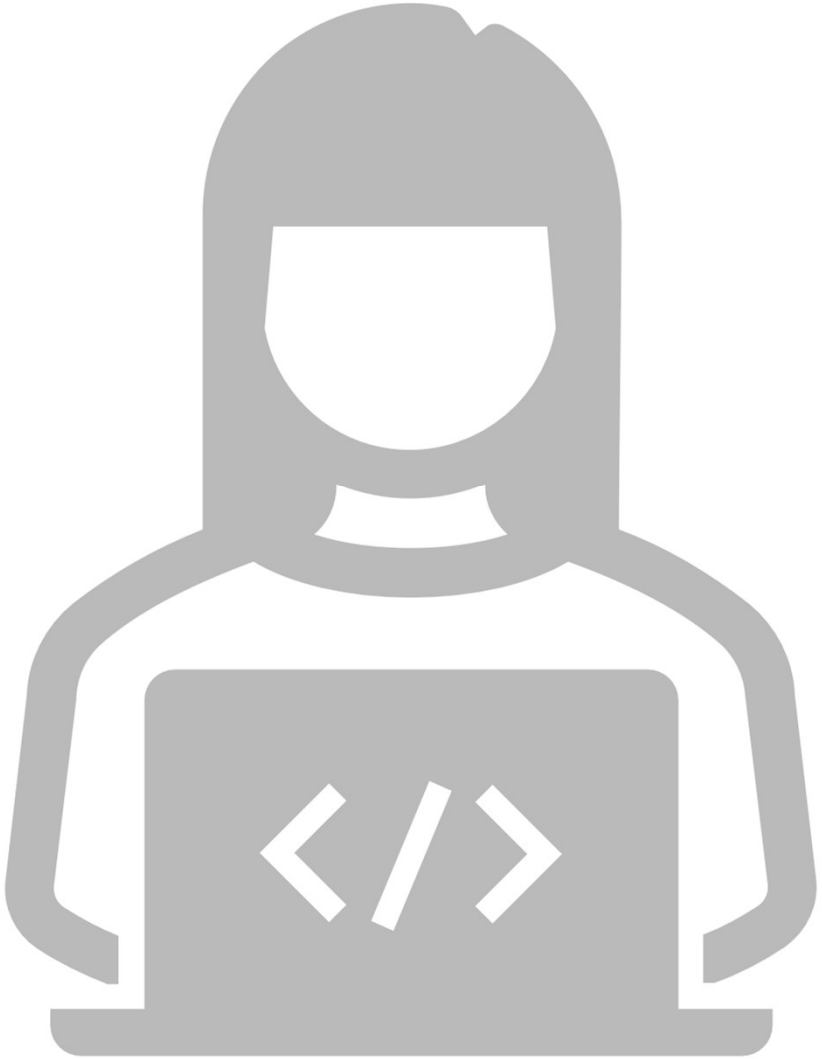
Pass by value and scope

- 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.

McGill initiative in Computational Medicine



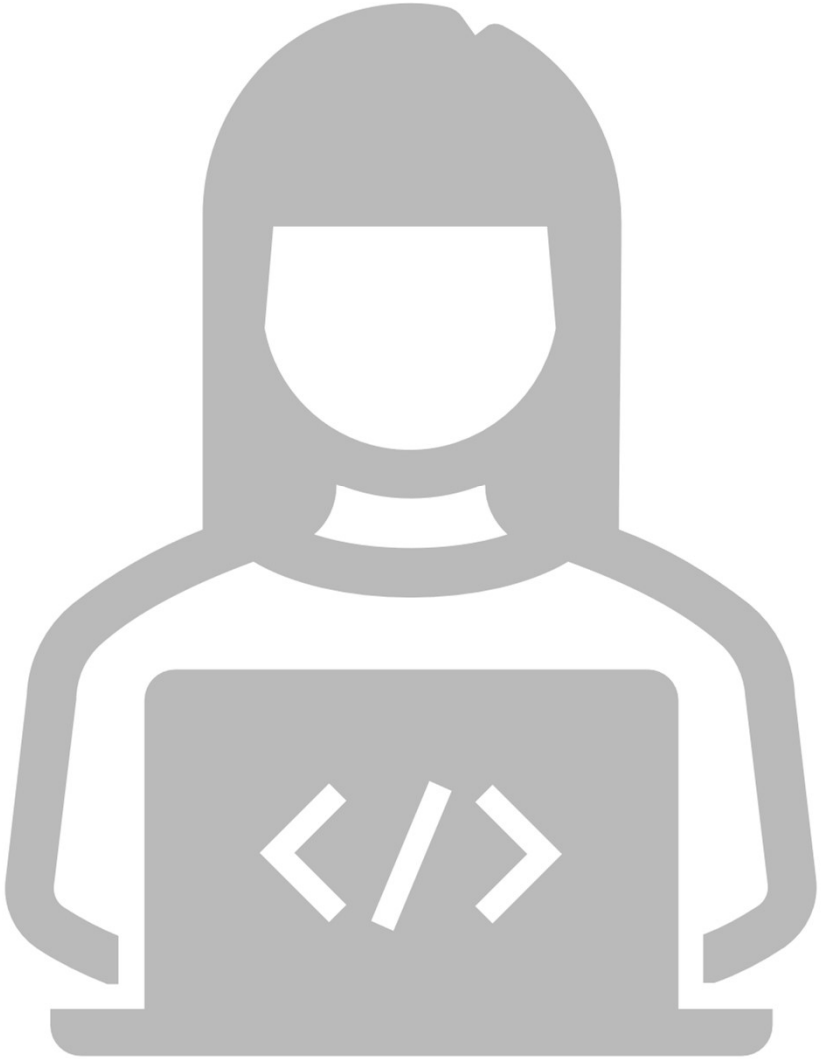
McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

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

McGill initiative in Computational Medicine



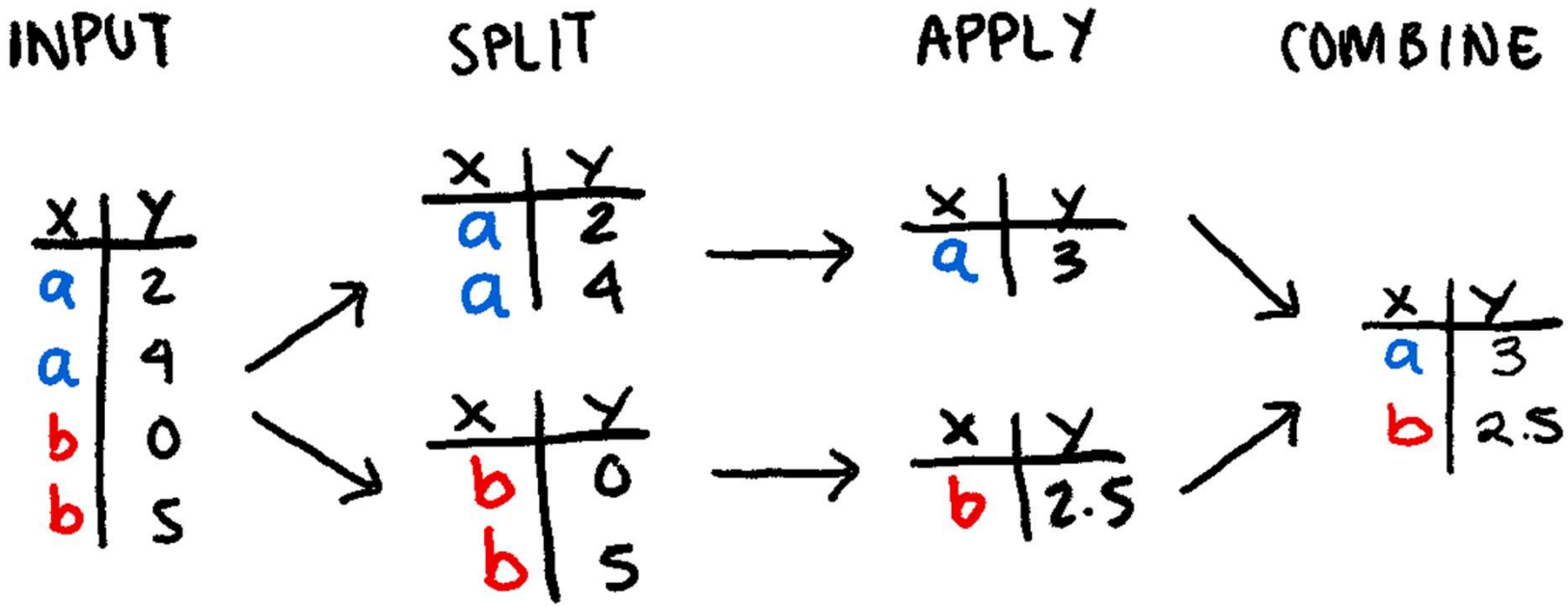
McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

Advanced data manipulation

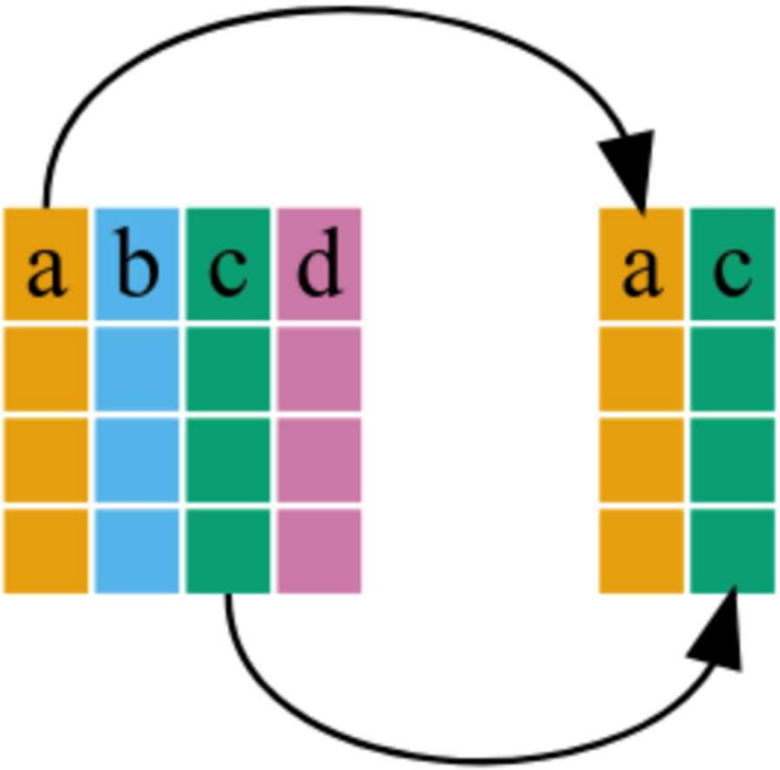
Learning objectives:

- 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

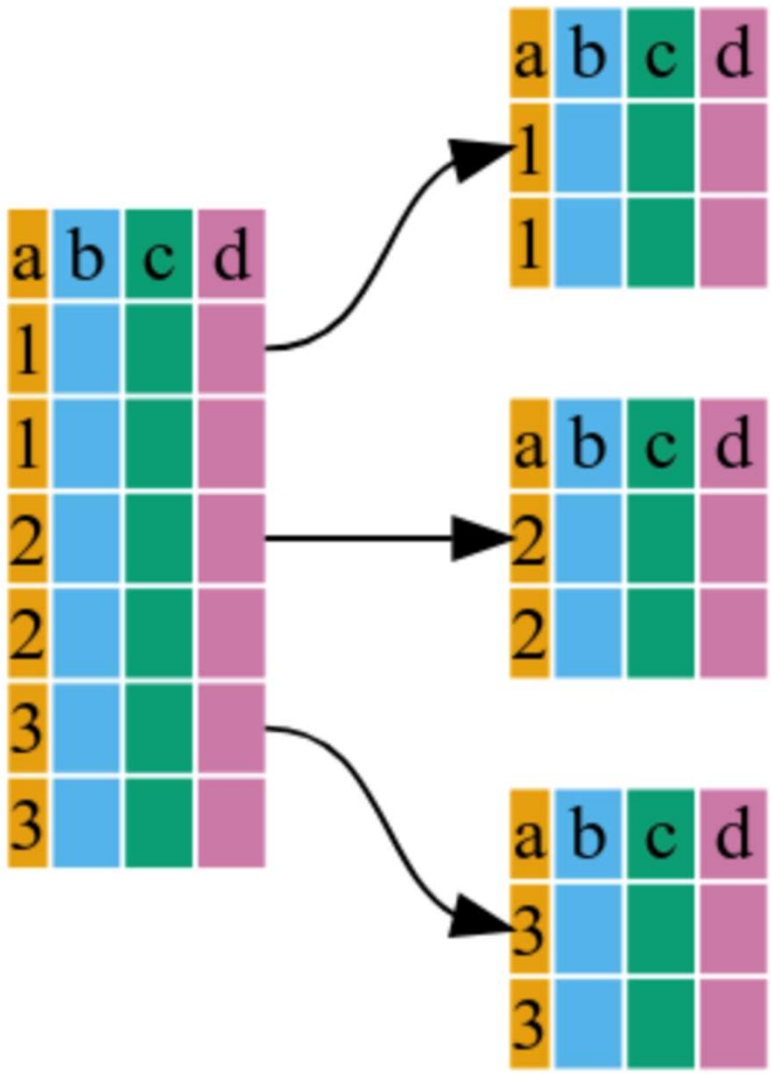


Select



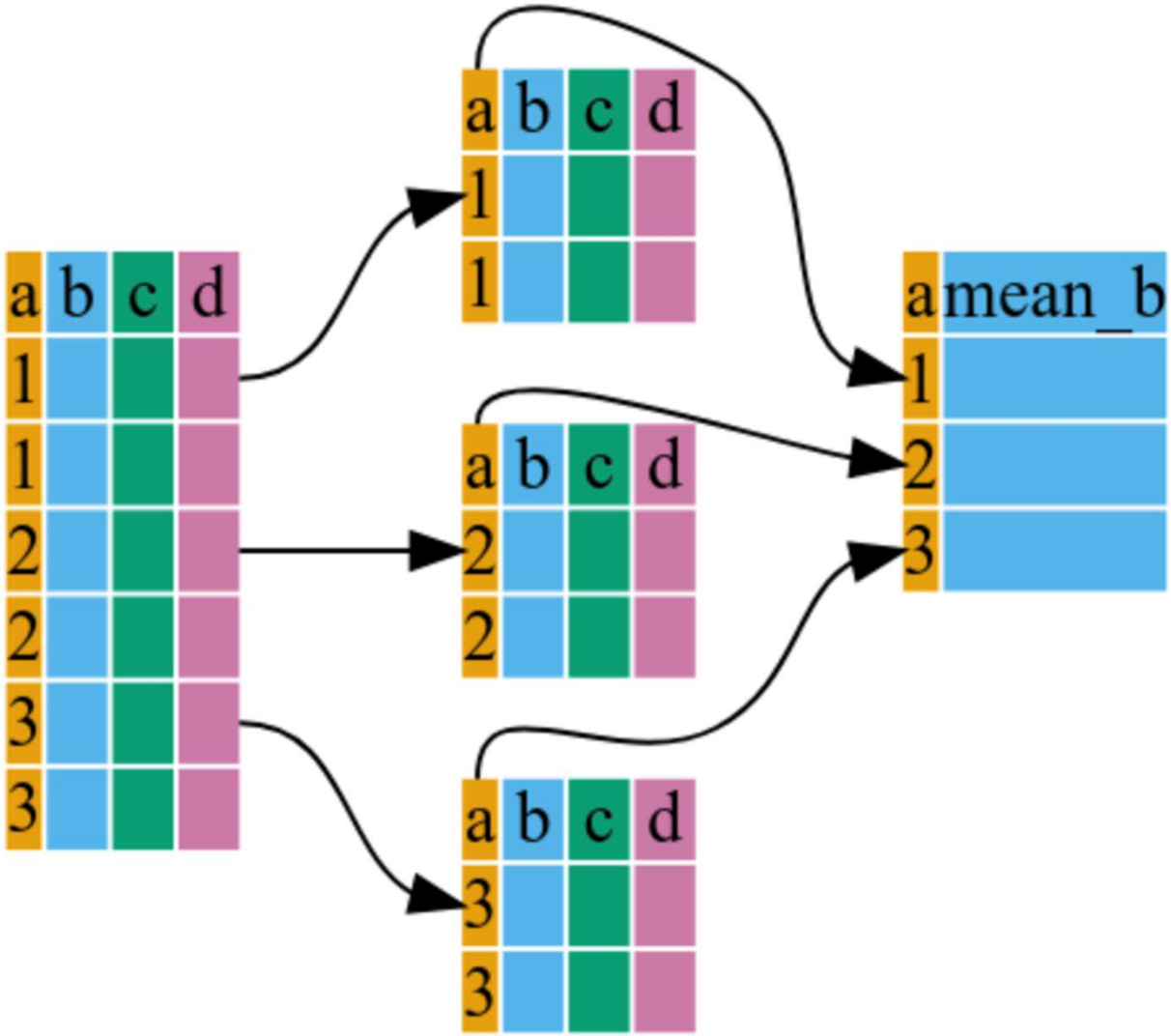
https://swcarpentry.github.io/r-novice-gapminder/13-dplyr/index.html

Group by



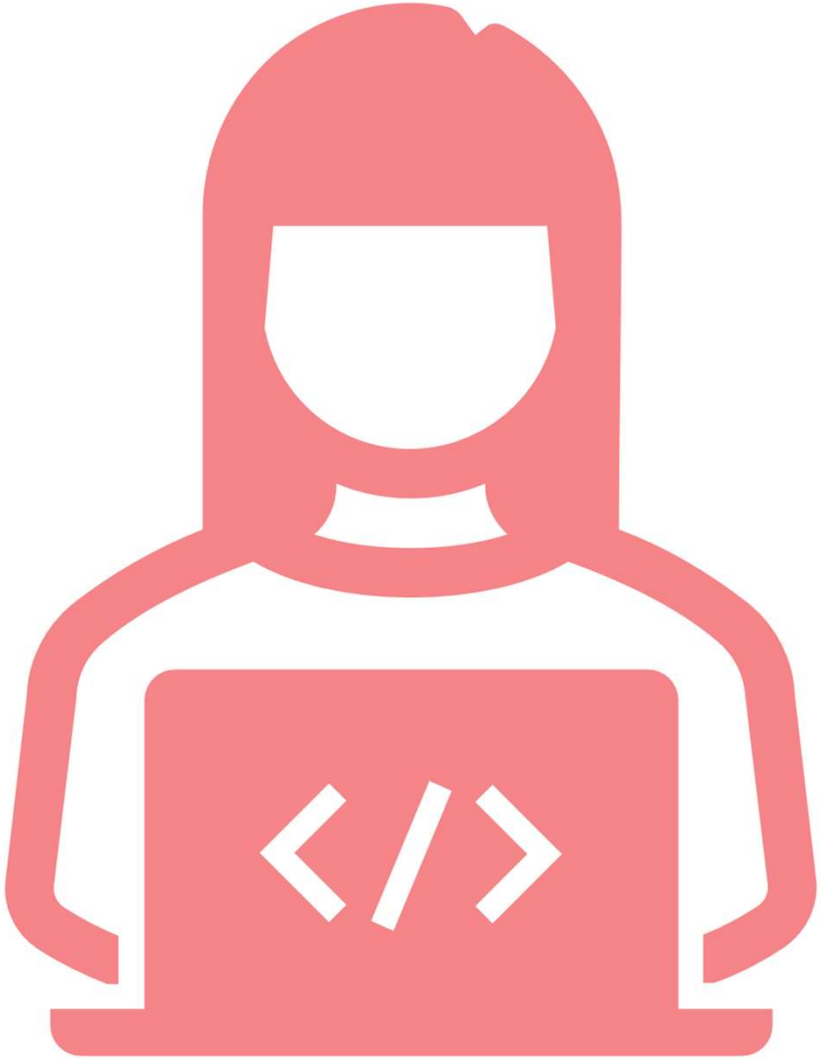
https://swcarpentry.github.io/r-novice-gapminder/13-dplyr/index.html

Summarize



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

McGill initiative in Computational Medicine



McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

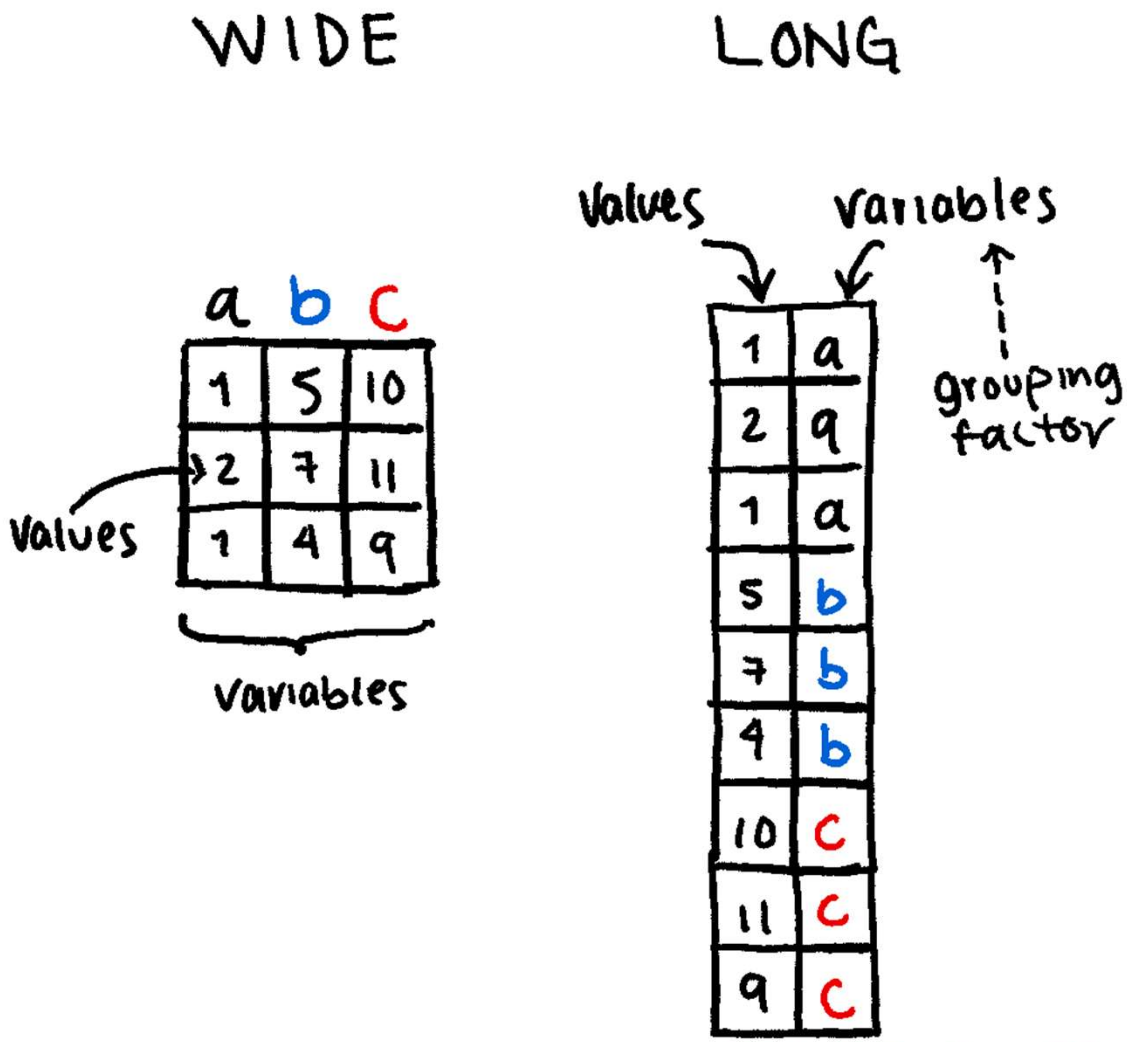
Break

Generating visual outputs

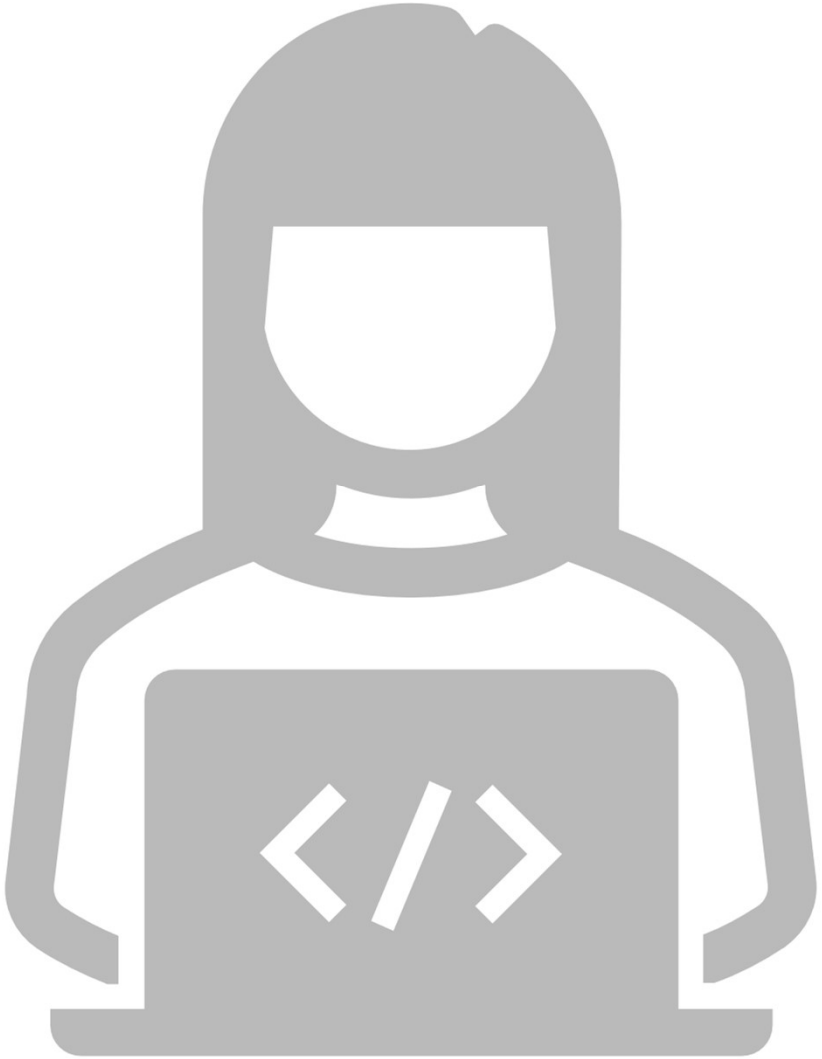
Learning objectives:

- 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



McGill initiative in Computational Medicine



McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

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?

Statistics in R Workshop (Part 2)

- Data Wrangling
- Regression
- Statistical Analysis

Statistics in R (Part 2)	Feb. 20 1PM-5PM	EDUC 133	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

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

McGill | Quantitative Life Sciences | Sciences quantitatives du vivant

Useful links

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