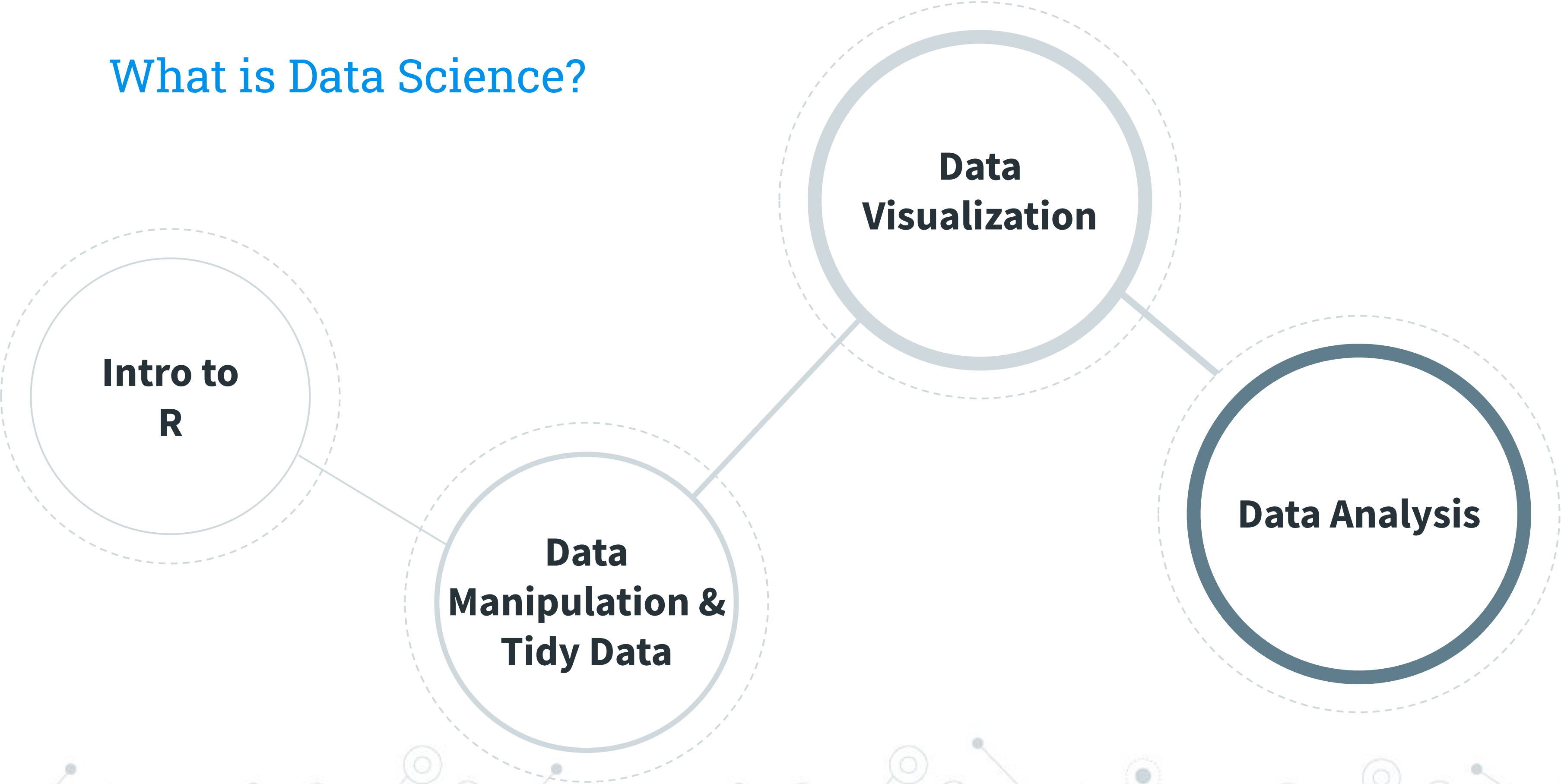
Data Science in R - Introduction

Fall 2021
Yama Chang

> *Data science is the problem-solving process to quantitatively formulate and rigorously answer questions that emphasizes clarity, reproducibility, and collaboration, and communicates the answer to a relevant audience.*

What is Data Science?



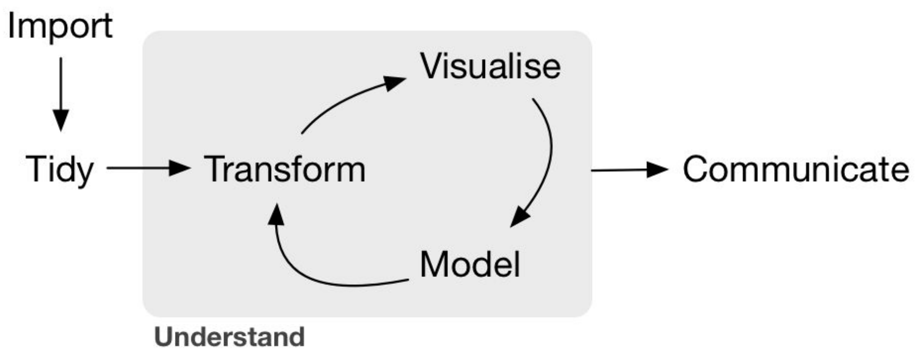
Intro to R

Data Manipulation & Tidy Data

Data Visualization

Data Analysis

A typical data project



Import

Tidy

Transform

Visualise

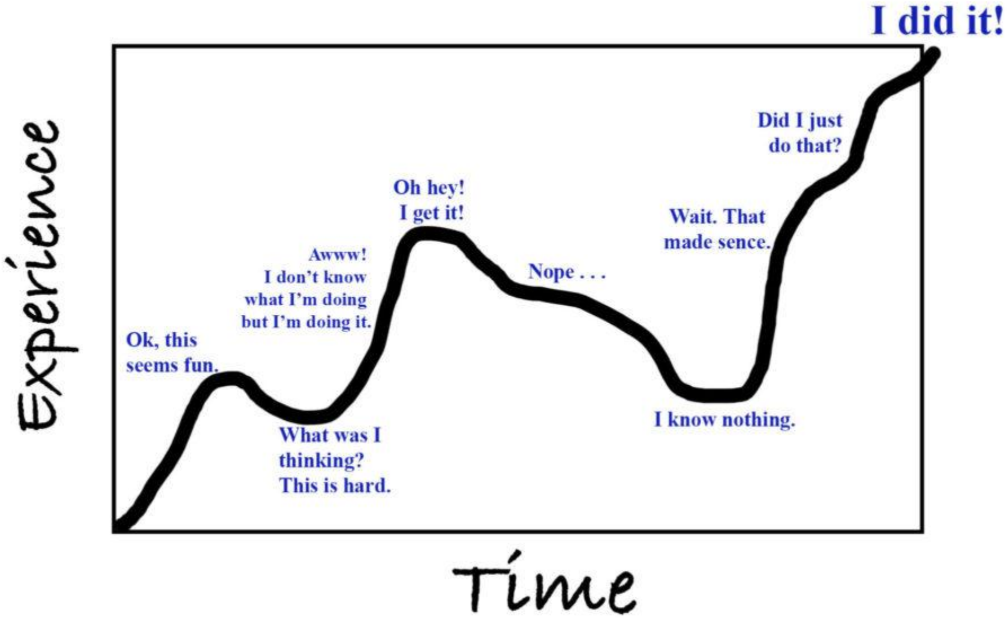
Model

Communicate

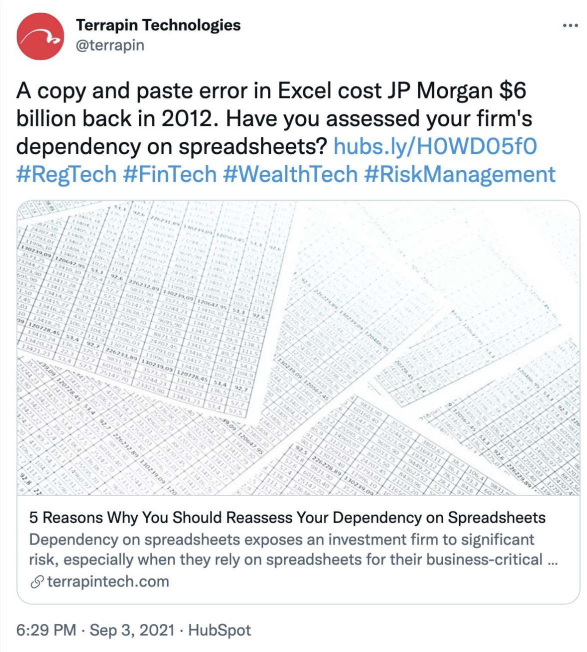
Understand

https://r4ds.had.co.nz/introduction.html

The Learning Curve



Why learn/use R?

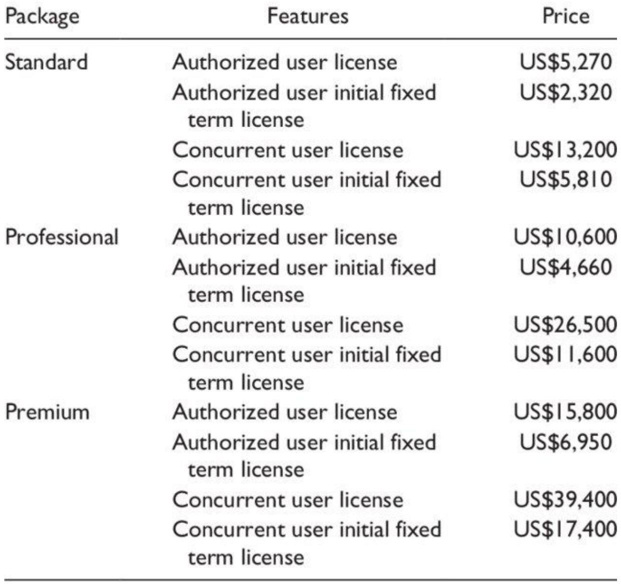


Terrapin Technologies
@terrapin

A copy and paste error in Excel cost JP Morgan $6 billion back in 2012. Have you assessed your firm's dependency on spreadsheets? hubs.ly/H0WD05f0 #RegTech #FinTech #WealthTech #RiskManagement

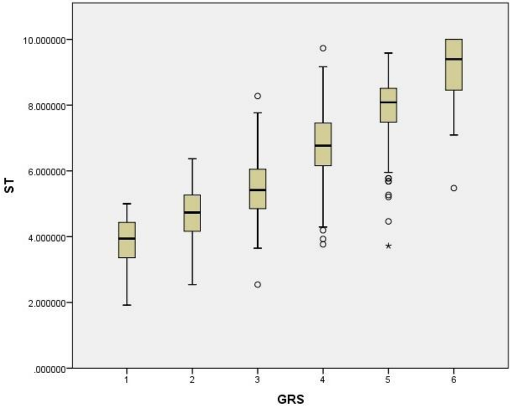
5 Reasons Why You Should Reassess Your Dependency on Spreadsheets
Dependency on spreadsheets exposes an investment firm to significant risk, especially when they rely on spreadsheets for their business-critical ...
terrapintech.com

6:29 PM · Sep 3, 2021 · HubSpot



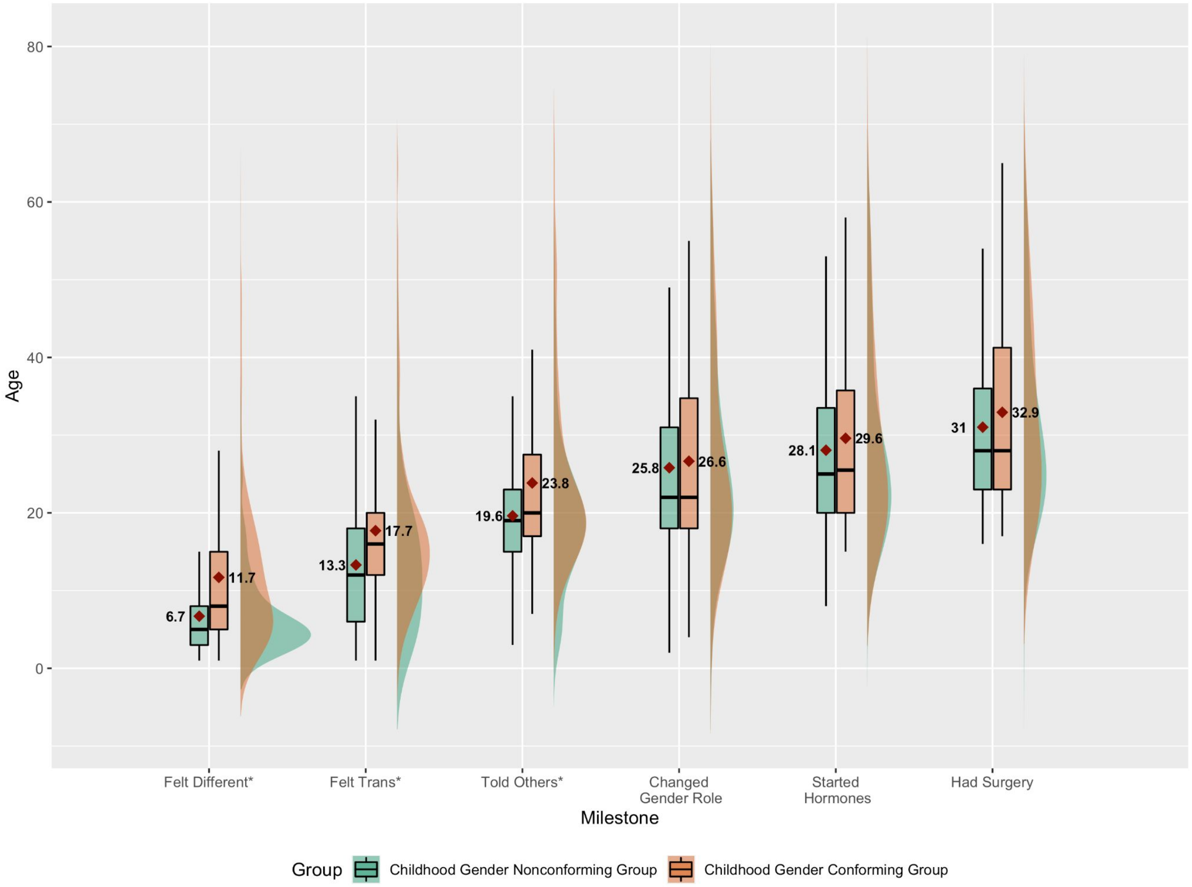
Package	Features	Price
Standard	Authorized user license	US$5,270
	Authorized user initial fixed term license	US$2,320
	Concurrent user license	US$13,200
	Concurrent user initial fixed term license	US$5,810
Professional	Authorized user license	US$10,600
	Authorized user initial fixed term license	US$4,660
	Concurrent user license	US$26,500
	Concurrent user initial fixed term license	US$11,600
Premium	Authorized user license	US$15,800
	Authorized user initial fixed term license	US$6,950
	Concurrent user license	US$39,400
	Concurrent user initial fixed term license	US$17,400

Source. IBM (2014a).



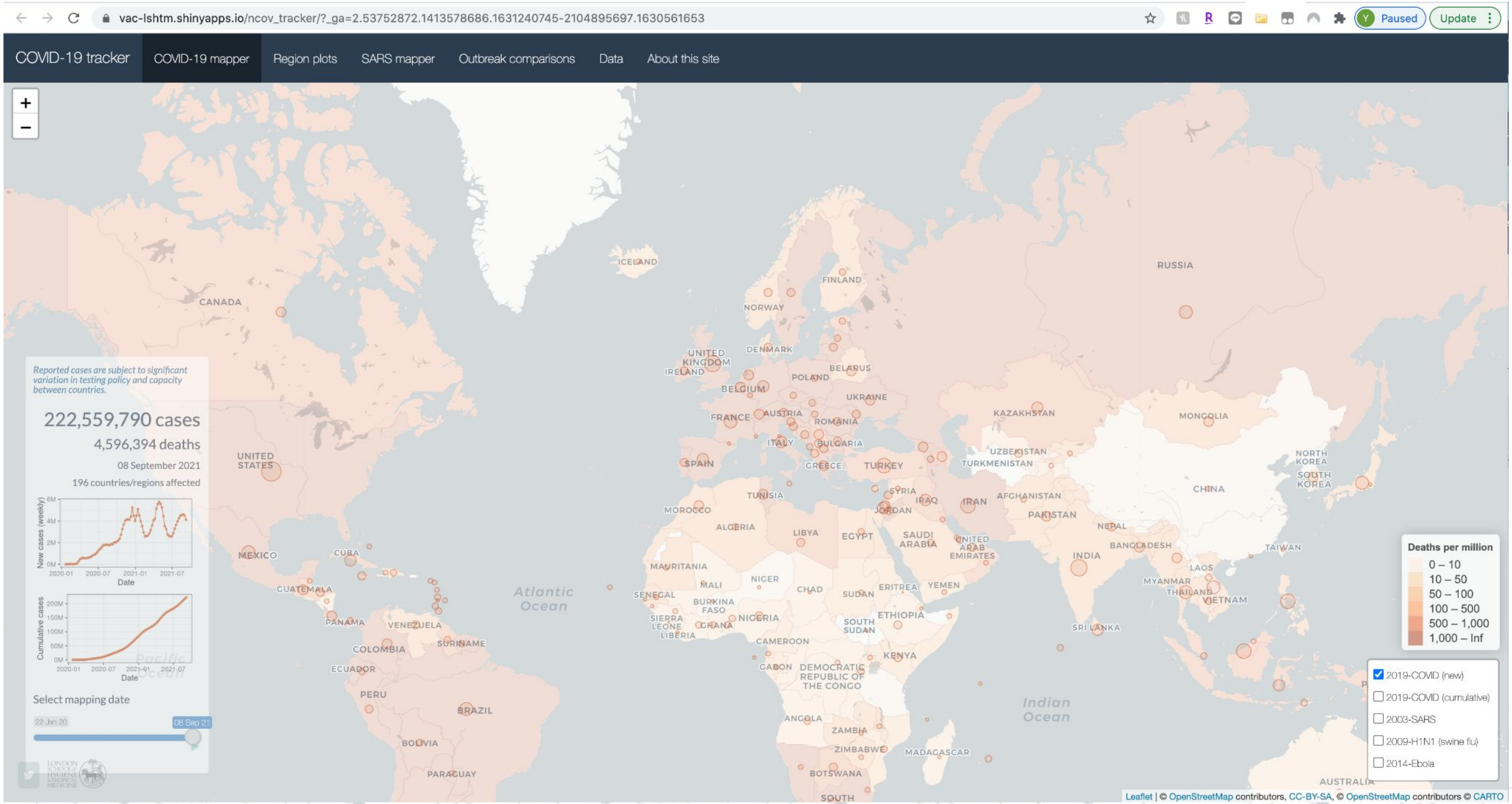
SPSS visualization

Figure 1. Developmental Trajectory of Milestones by Chilhood Gender Role Group (N = 330)



* Star denotes significant difference between two groups.

R Shiny apps



vac-lshtm.shinyapps.io/ncov_tracker/?_ga=2.53752872.1413578686.1631240745-2104895697.1630561653

COVID-19 tracker

Reported cases are subject to significant variation in testing policy and capacity between countries.

222,559,790 cases

4,596,394 deaths

08 September 2021

196 countries/regions affected

Select mapping date

22 Jan 20 — 08 Sep 21

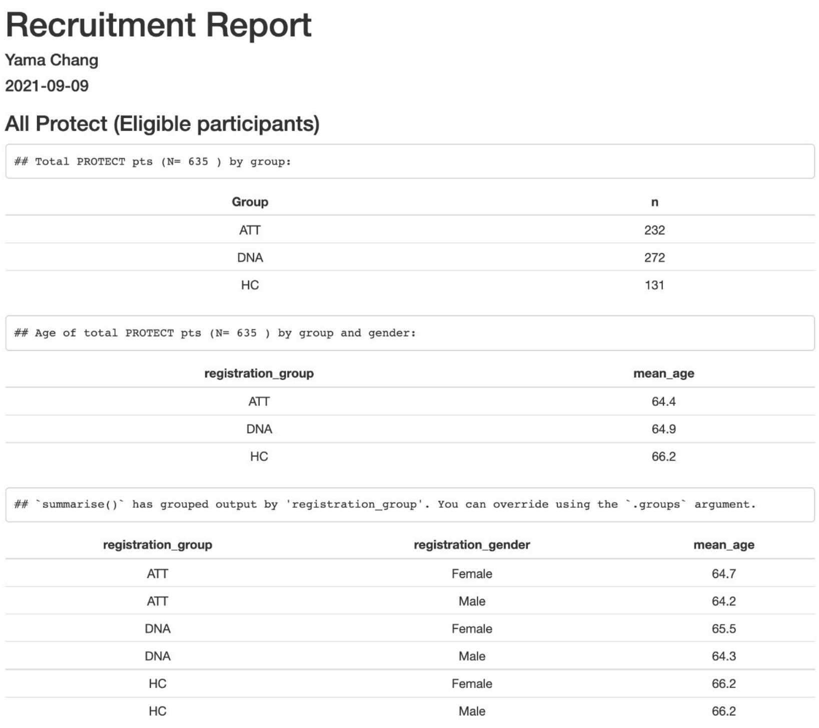
Deaths per million

- 0 – 10
- 10 – 50
- 50 – 100
- 100 – 500
- 500 – 1,000
- 1,000 – Inf

☑ 2019-COVID (new)
☐ 2019-COVID (cumulative)
☐ 2003-SARS
☐ 2009-H1N1 (swine flu)
☐ 2014-Ebola

Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap contributors © CARTO

R Markdown



Recruitment Report

Yama Chang

2021-09-09

All Protect (Eligible participants)

```
## Total PROTECT pts (N= 635 ) by group:
```

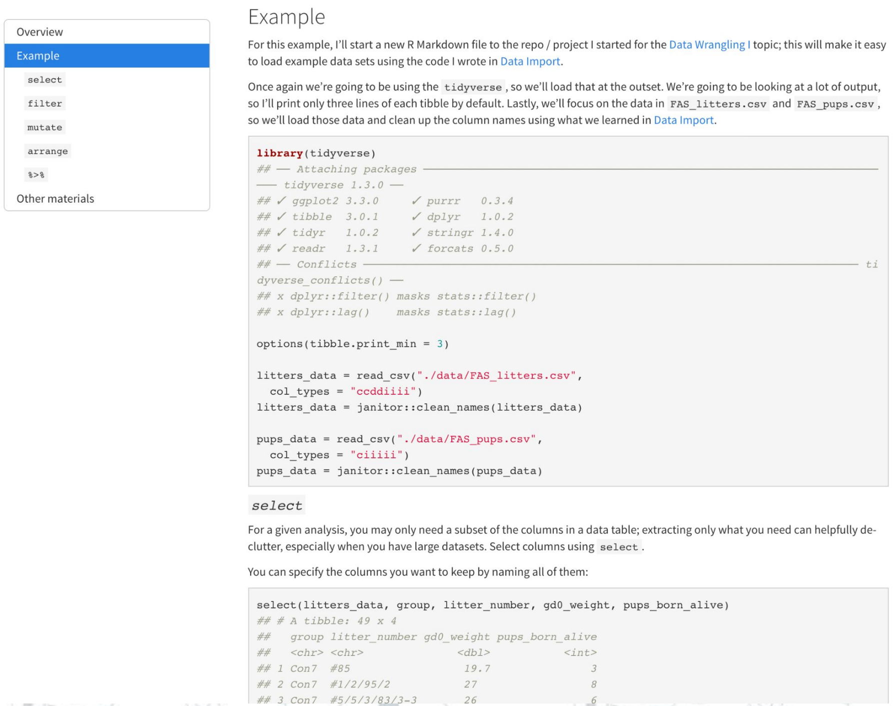
Group	n
ATT	232
DNA	272
HC	131

```
## Age of total PROTECT pts (N= 635 ) by group and gender:
```

registration_group	mean_age
ATT	64.4
DNA	64.9
HC	66.2

```
## `summarise()` has grouped output by 'registration_group'. You can override using the `.groups` argument.
```

registration_group	registration_gender	mean_age
ATT	Female	64.7
ATT	Male	64.2
DNA	Female	65.5
DNA	Male	64.3
HC	Female	66.2
HC	Male	66.2



Example

For this example, I'll start a new R Markdown file to the repo / project I started for the Data Wrangling I topic; this will make it easy to load example data sets using the code I wrote in Data Import.

Once again we're going to be using the `tidyverse`, so we'll load that at the outset. We're going to be looking at a lot of output, so I'll print only three lines of each tibble by default. Lastly, we'll focus on the data in `FAS_litters.csv` and `FAS_pups.csv`, so we'll load those data and clean up the column names using what we learned in Data Import.

```
library(tidyverse)
## ── Attaching packages ──────────────────────────────────────────────────────────
─── tidyverse 1.3.0 ──
## ✓ ggplot2 3.3.0     ✓ purrr   0.3.4
## ✓ tibble  3.0.1     ✓ dplyr   1.0.2
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
## ── Conflicts ──────────────────────────────────────────────────────────────── ti
dyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

options(tibble.print_min = 3)

litters_data = read_csv("./data/FAS_litters.csv",
  col_types = "ccddiiii")
litters_data = janitor::clean_names(litters_data)

pups_data = read_csv("./data/FAS_pups.csv",
  col_types = "ciiiii")
pups_data = janitor::clean_names(pups_data)
```

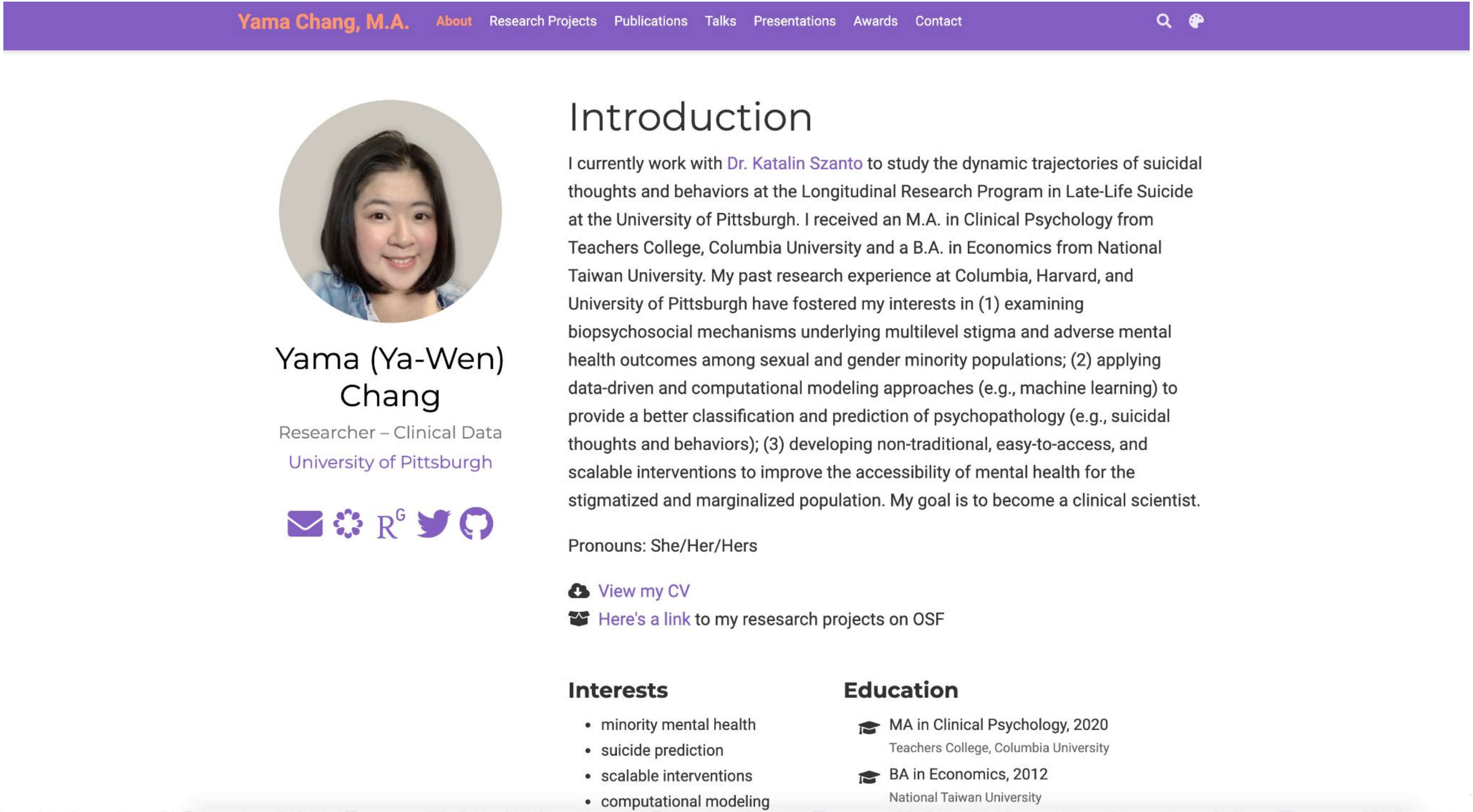
`select`

For a given analysis, you may only need a subset of the columns in a data table; extracting only what you need can helpfully de-clutter, especially when you have large datasets. Select columns using `select`.

You can specify the columns you want to keep by naming all of them:

```
select(litters_data, group, litter_number, gd0_weight, pups_born_alive)
## # A tibble: 49 x 4
##   group litter_number gd0_weight pups_born_alive
##   <chr> <chr>              <dbl>           <int>
## 1 Con7  #85                 19.7               3
## 2 Con7  #1/2/95/2           27                 8
## 3 Con7  #5/5/3/83/3-3       26                 6
```

https://www.p8105.com/data_manipulation.html#Example



Yama (Ya-Wen) Chang

Researcher – Clinical Data

University of Pittsburgh

Introduction

I currently work with Dr. Katalin Szanto to study the dynamic trajectories of suicidal thoughts and behaviors at the Longitudinal Research Program in Late-Life Suicide at the University of Pittsburgh. I received an M.A. in Clinical Psychology from Teachers College, Columbia University and a B.A. in Economics from National Taiwan University. My past research experience at Columbia, Harvard, and University of Pittsburgh have fostered my interests in (1) examining biopsychosocial mechanisms underlying multilevel stigma and adverse mental health outcomes among sexual and gender minority populations; (2) applying data-driven and computational modeling approaches (e.g., machine learning) to provide a better classification and prediction of psychopathology (e.g., suicidal thoughts and behaviors); (3) developing non-traditional, easy-to-access, and scalable interventions to improve the accessibility of mental health for the stigmatized and marginalized population. My goal is to become a clinical scientist.

Pronouns: She/Her/Hers

View my CV

Here's a link to my resesarch projects on OSF

Interests

- minority mental health
- suicide prediction
- scalable interventions
- computational modeling

Education

MA in Clinical Psychology, 2020
Teachers College, Columbia University

BA in Economics, 2012
National Taiwan University

R blogdown

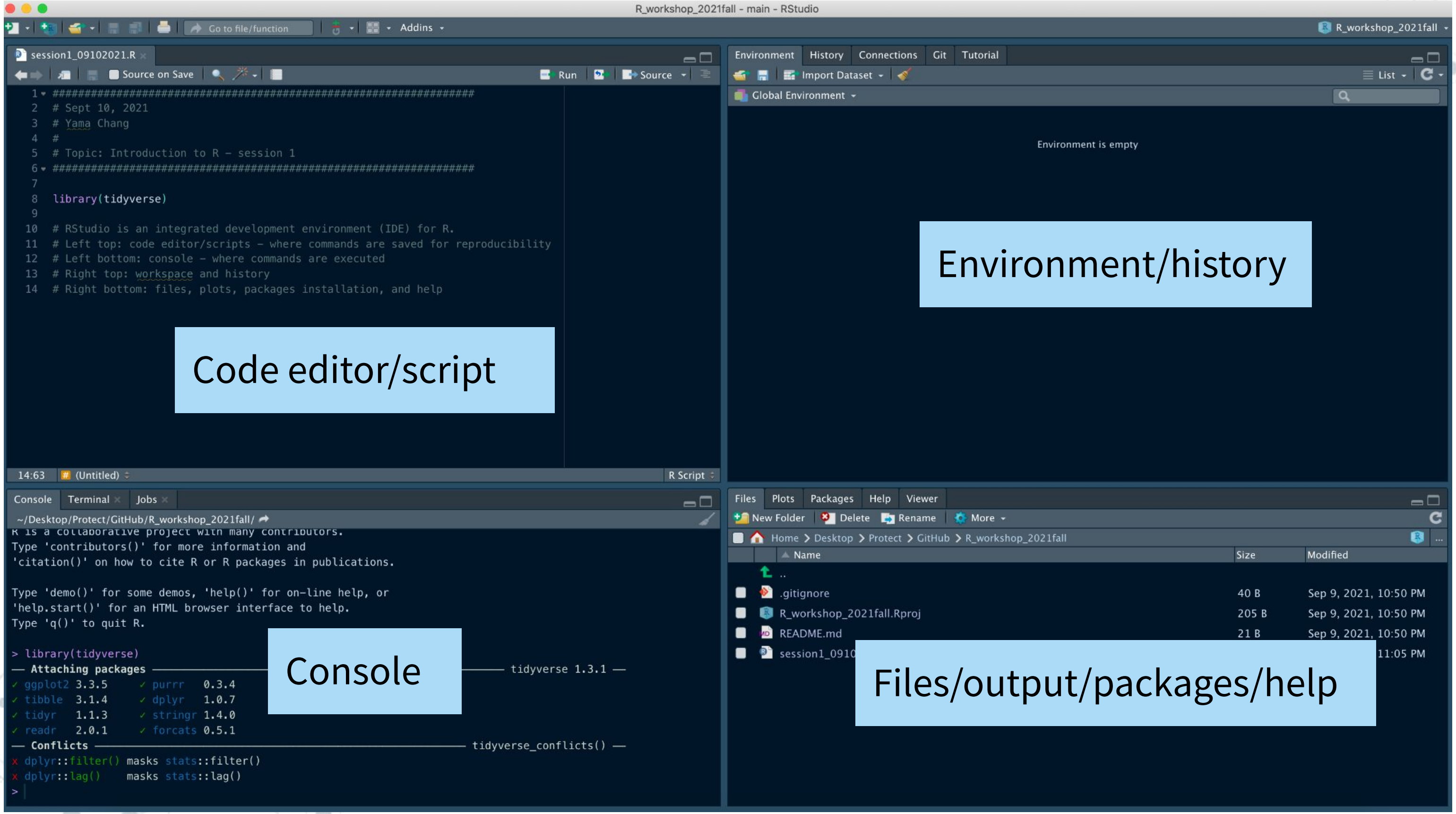
Overview: learning goal

- R Studio interface
- Establish good habits now (to make your life easier!)
- Reproducibility
- Get started
  - Create a new R script
  - Run codes
  - Install and load packages
  - Working Directory
- Let’s do some coding!
  - Computation - operation and objects
  - Data frames
  - Data structures

Before we actually start

- Installation of R and R Studio
  - R - programming language: https://www.r-project.org/
  - R Studio: an integrated development environment (IDE) for R.
    - https://www.rstudio.com/products/rstudio/download/

R Studio interface



```r
#############################################################################
# Sept 10, 2021
# Yama Chang
#
# Topic: Introduction to R - session 1
#############################################################################

library(tidyverse)

# RStudio is an integrated development environment (IDE) for R.
# Left top: code editor/scripts - where commands are saved for reproducibility
# Left bottom: console - where commands are executed
# Right top: workspace and history
# Right bottom: files, plots, packages installation, and help
```

```
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.

> library(tidyverse)
── Attaching packages ───                              ───── tidyverse 1.3.1 ──
✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.4     ✓ dplyr   1.0.7
✓ tidyr   1.1.3     ✓ stringr 1.4.0
✓ readr   2.0.1     ✓ forcats 0.5.1
── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
>
```

Name	Size	Modified
..		
.gitignore	40 B	Sep 9, 2021, 10:50 PM
R_workshop_2021fall.Rproj	205 B	Sep 9, 2021, 10:50 PM
README.md	21 B	Sep 9, 2021, 10:50 PM
session1_0910		11:05 PM

Environment/history

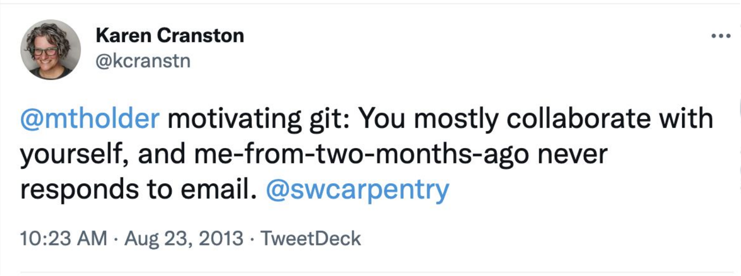
Code editor/script

Console

Files/output/packages/help

Establish good habits now

- Some R basics
  - Code is case sensitive
  - No autocorrect (which is good!)
- Some good habits
  - Establish a variable name convention
    - this_is_snake_case (preferable!)
    - this.is.period.case
    - ThIsNoTaNaMiNgCoNvEnTiOn
  - Comment your codes with # for reproducibility and save your headache
  - Make readable/beautiful codes



Karen Cranston
@kcranstn

@mtholder motivating git: You mostly collaborate with yourself, and me-from-two-months-ago never responds to email. @swcarpentry

10:23 AM · Aug 23, 2013 · TweetDeck

Reproducibility

- Give the same code and data, anyone should be able to reproduce each step of your work/analysis and show the same results
- One day someone will reproduce your work - be prepared!

Get started

- Create a new R script
    - File → New Files → R Script
- Run code: put your cursor at any place of a line of code
    - Command + enter (Mac)
    - Ctrl + enter (Windows)
- Autocompletion - start typing a variable name and click tab

Get started

- Install packages - collections of functions and data sets developed by the R user community
  - Currently, there’re 18149 available packages!
  - Only need to install **once** in your environment

  -
    ```
    install.packages("tidyverse")
    ```
- Load packages
  - Need to load **every time** in your environment

  -
    ```
    library(tidyverse)
    ```
- Working Directory - where you store this project/script/data/plot

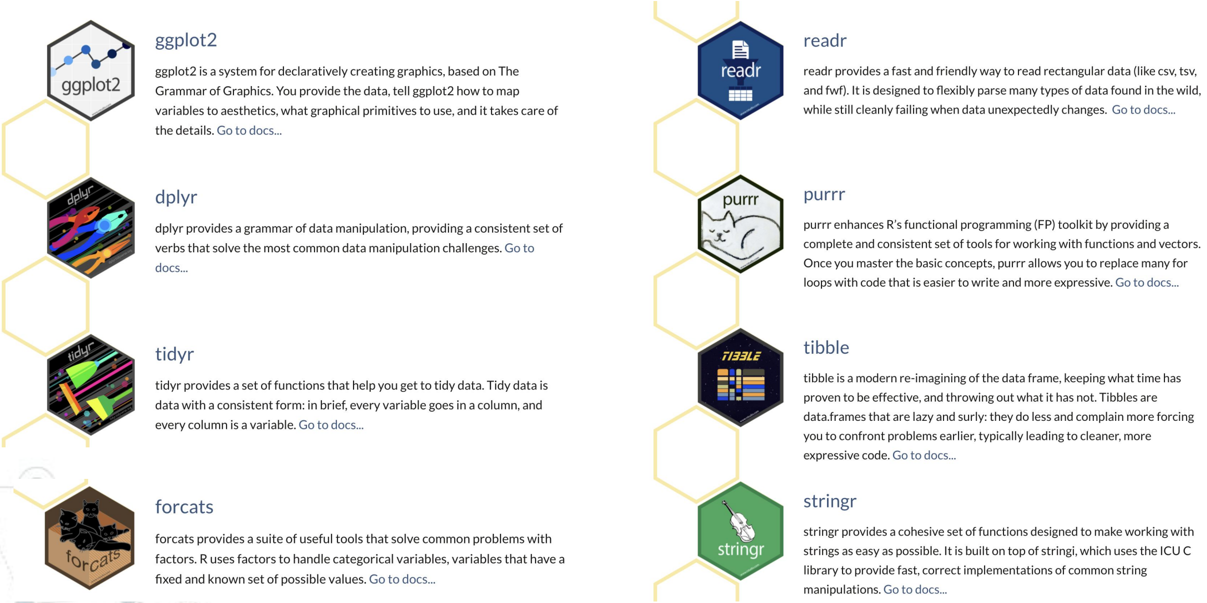
  -
    ```
    getwd()
    ```

  -
    ```
    setwd(“/Users/yama/Box/Yama/R workshop”)
    ```

https://cran.r-project.org/web/packages/

Tidyverse

- A package made for easier, faster, and more fun in coding
- You can basically use this package for everything in data science - tidy data, analysis, visualization, and analysis.



ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. Go to docs...

dplyr

dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges. Go to docs...

tidyr

tidyr provides a set of functions that help you get to tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable. Go to docs...

forcats

forcats provides a suite of useful tools that solve common problems with factors. R uses factors to handle categorical variables, variables that have a fixed and known set of possible values. Go to docs...

readr

readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. Go to docs...

purrr

purrr enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. Once you master the basic concepts, purrr allows you to replace many for loops with code that is easier to write and more expressive. Go to docs...

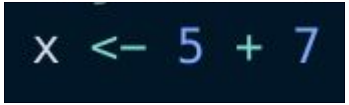
tibble

tibble is a modern re-imagining of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code. Go to docs...

stringr

stringr provides a cohesive set of functions designed to make working with strings as easy as possible. It is built on top of stringi, which uses the ICU C library to provide fast, correct implementations of common string manipulations. Go to docs...

Let’s do some coding!

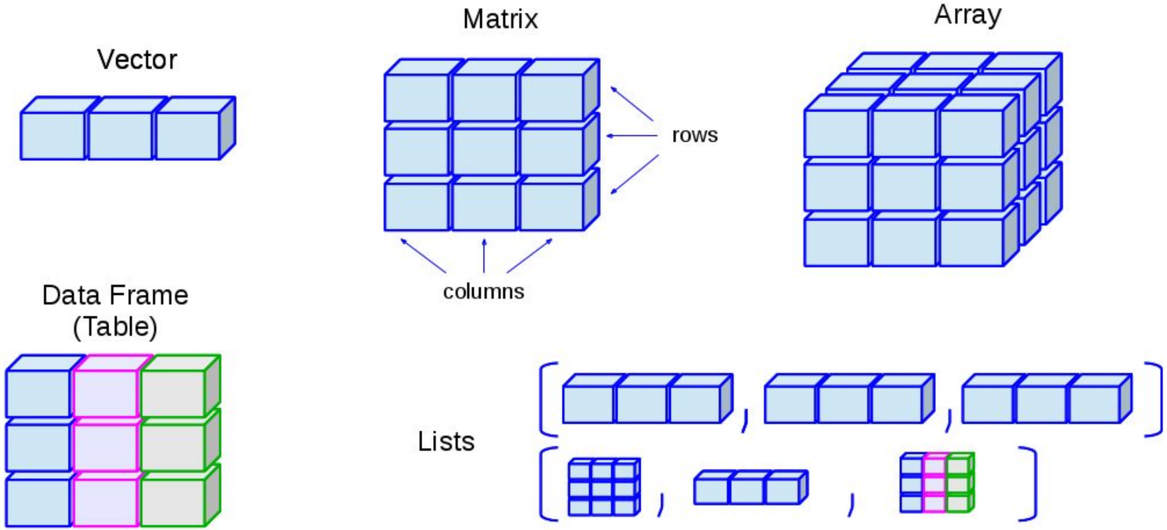


```
x <- 5 + 7
```

Object

Value

- Computation: operation and objects
- Data Structure

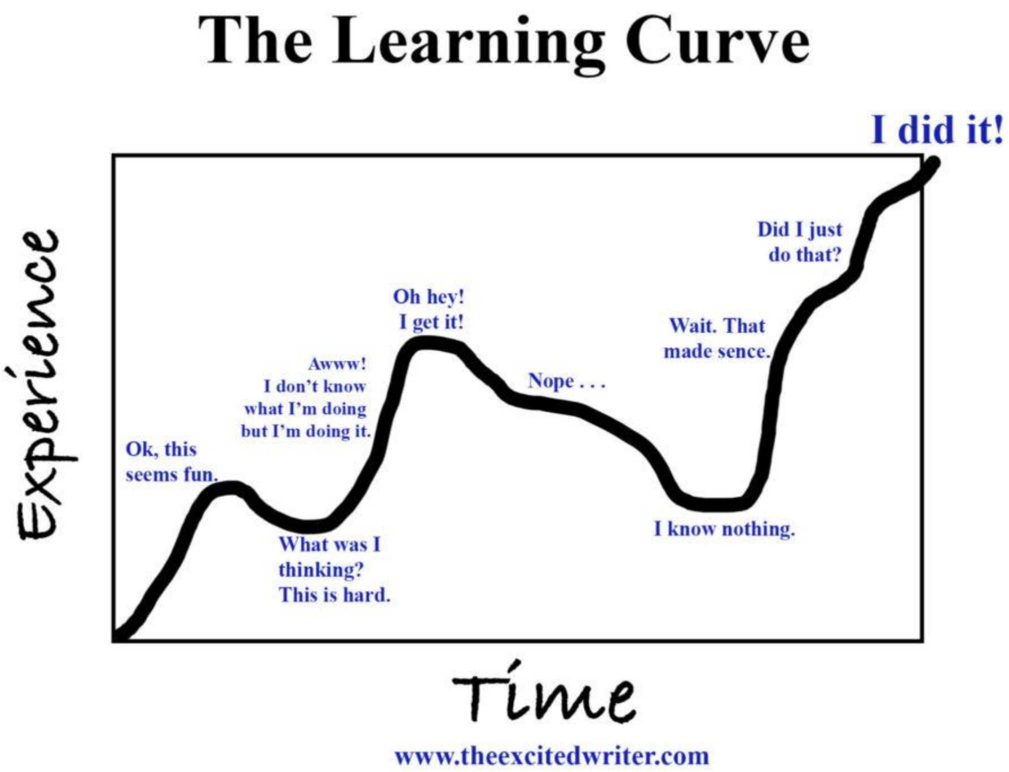


http://venus.ifca.unican.es/Rintro/dataStruct.html

Some mindset of coding

- Read the errors
  - A lot of googling!
  - You can basically find all solution at stackoverflow
- Plan for mistakes
  - It’s TOTALLY fine to make mistakes
  - Write codes that make it easy to fix - clean codes
- Learning curves

Some


The Learning Curve
I did it!
Did I just do that?
Oh hey! I get it!
Wait. That made sence.
Awww! I don't know what I'm doing but I'm doing it.
Nope . . .
Ok, this seems fun.
I know nothing.
What was I thinking? This is hard.
Experience
Time
www.theexcitedwriter.com

Helpful resources

- R Studio cheatsheet
- A (very) short intro to R
- R for Data Science
- Data Science I (P8105) at Columbia University School of Public Health

Thank you!

You can find my slides and codes at my GitHub

Also find me at: changy11@upmc.edu