Welcome! & Data Science Life-cycle

CVEN 5999 - Summer 2025

Lars Schöbitz

cven5999-ss25.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER

Welcome! 👋

cven5999-ss25.github.io/website/

Meet the lecturer

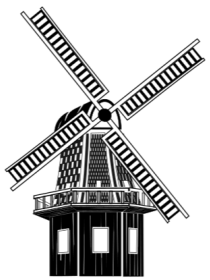
Lars Schöbitz (he/him)

- Environmental Engineer
- Open Science Specialist at ETH Zurich
- Independent Instructor for Data Science with R
- LinkedIn: Lars Schöbitz

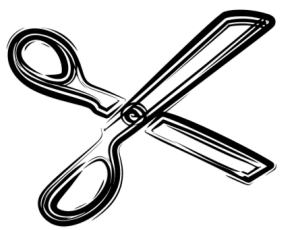
Learning Goals (for the course)

1. Be able to use a common set of data science tools **(R, RStudio IDE, Git, GitHub, tidyverse, Quarto)** to illustrate and communicate the utility of solutions for water, sanitation, air quality, and global health.
2. Learn to use the **Quarto file format** and the RStudio IDE visual editing mode to produce **scholarly documents** with citations, footnotes, cross-references, figures, and tables.

Why are you here?







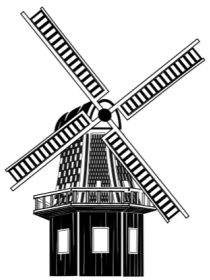

Images from: https://openclipart.org/

Pick an item

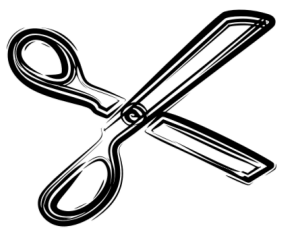
Take notes for 2 minutes:

What does the item you have picked have to do with the reason for you being here?

Why are you here?








Images from: https://openclipart.org/

In break-out rooms

Take 2 minutes each to share with your room partner:
What does the item you have picked have to do with the reason for you being here?

From which country are you joining us?

In the Zoom chat

Share with us from which country you are joining us.

Learning Objectives (for this week)

1. Learners can navigate the platforms (Posit Cloud, GitHub, Course Website) that are used to for the course.
2. Learners can render a Quarto file to an output file in HTML, PDF and DOCX format.
3. Learners can list the six elements of the data science lifecycle.
4. Learners can identify four components of a Quarto file (YAML, code chunk, R code, markdown).

cven5999-ss25.github.io/website/

Classroom tools

cven5999-ss25.github.io/website/

Live Coding Exercises

- Instructor writes and narrates code out loud
- Intstructor explains elements and principles that are relevant
- Code is displayed on second screen / split screen
- Learners join by writing and executing the same code
- Learners “code-along” with the instructor

Pair Programming Exercises

- Two learners work together in a break out session
- One person (the driver) shares the screen and does the typing
- The other person (the navigator) offers comments and suggestions
- Roles get switched

Platforms and Tools

- R
- Posit Cloud
- RStudio IDE
- tidyverse R Packages
- Quarto publishing system

cven5999-ss25.github.io/website/

Posit Cloud

Posit Cloud

posit.cloud/spaces/381404/content/6066891

Cven5837-Ss23 / course-material-rainbow-train

RAM

Rainbow Train

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.3.0

live-01a-setup.qmd | live-01b-data-science-lifecycle-so...

Render | Run

Source | Visual | Normal | Format | Insert | Table

3. calculating summary statistics (like counts or the mean)

Gapminder data

Goal: Calculate the median life expectancy at birth by continent for 2007.

```{r}

# before loading library, write code

library(dplyr)
```

(Top Level) | Quarto

Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...

Environment | History | Connections | Git | Tutorial

Import Dataset | 172 MiB | List

R | Global Environment

Environment is empty

Console | Terminal | Background Jobs

R 4.3.0 · /cloud/project/

```
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

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.
```

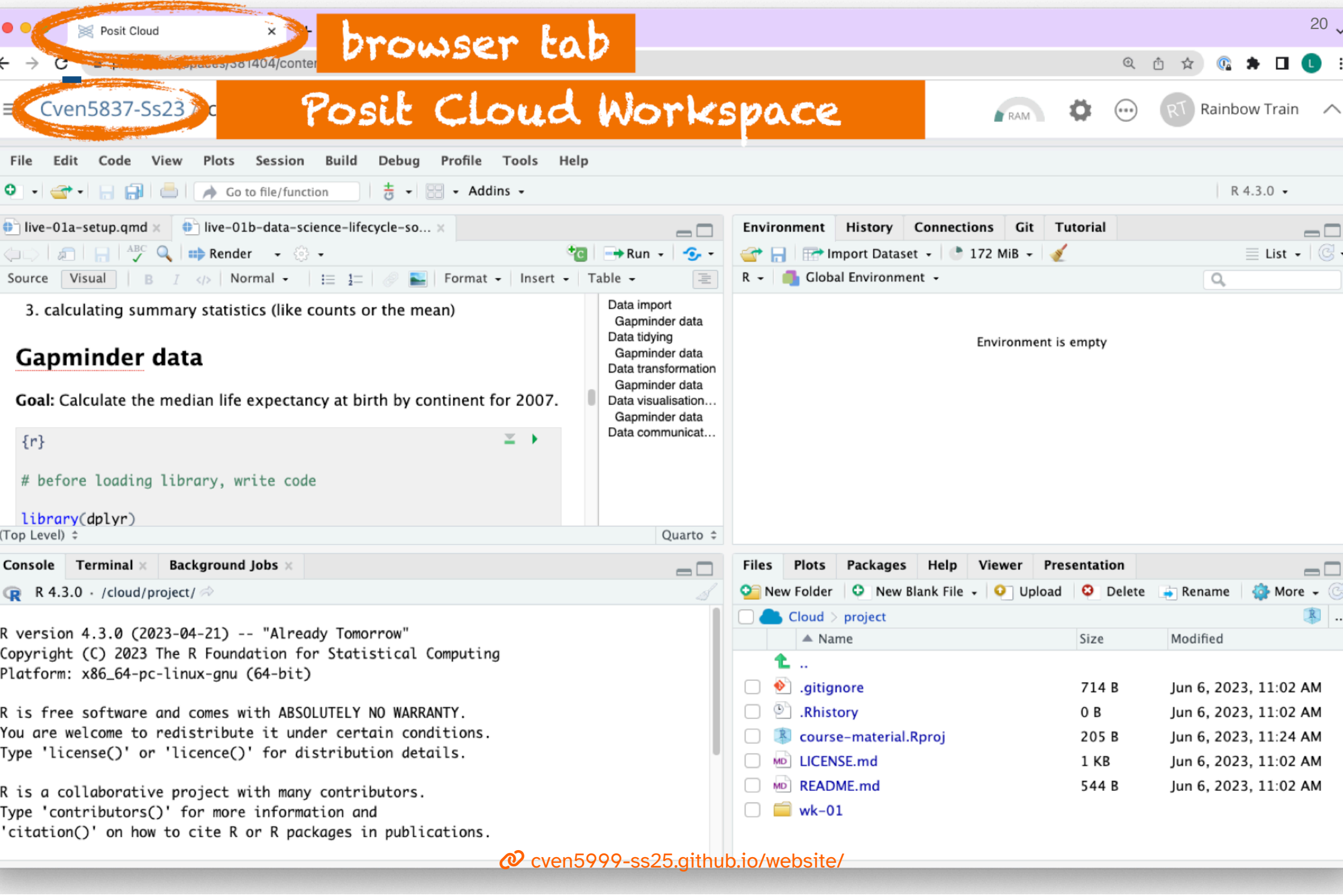
Files | Plots | Packages | Help | Viewer | Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
.gitignore	714 B	Jun 6, 2023, 11:02 AM
.Rhistory	0 B	Jun 6, 2023, 11:02 AM
course-material.Rproj	205 B	Jun 6, 2023, 11:24 AM
LICENSE.md	1 KB	Jun 6, 2023, 11:02 AM
README.md	544 B	Jun 6, 2023, 11:02 AM
wk-01		

cven5999-ss25.github.io/website/


Posit Cloud
browser tab
Cven5837-Ss23
Posit Cloud Workspace
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.3.0
live-01a-setup.qmd
live-01b-data-science-lifecycle-so...
Render
Run
Source Visual
Normal
Format Insert Table
3. calculating summary statistics (like counts or the mean)
Gapminder data
Goal: Calculate the median life expectancy at birth by continent for 2007.
{r}
# before loading library, write code
library(dplyr)
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
(Top Level)
Quarto
Environment History Connections Git Tutorial
Import Dataset
172 MiB
List
R
Global Environment
Environment is empty
Console Terminal Background Jobs
R 4.3.0 · /cloud/project/
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.
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.
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud > project
Name
Size
Modified
..
.gitignore 714 B Jun 6, 2023, 11:02 AM
.Rhistory 0 B Jun 6, 2023, 11:02 AM
course-material.Rproj 205 B Jun 6, 2023, 11:24 AM
LICENSE.md 1 KB Jun 6, 2023, 11:02 AM
README.md 544 B Jun 6, 2023, 11:02 AM
wk-01
cven5999-ss25.github.io/website/

Posit Cloud

posit.cloud/spaces/381404/content/6066891

Cven5837-Ss23 / course-material-rainbow-train

RAM

Rainbow Train

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Addins

RStudio IDE Menu

R 4.3.0

live-01a-setup.qmd

live-01b-data-science-lifecycle-so...

Render

Run

Source Visual

Normal

Format Insert Table

3. calculating summary statistics (like counts or the mean)

Gapminder data

Goal: Calculate the median life expectancy at birth by continent for 2007.

```{r}

# before loading library, write code

library(dplyr)
```

Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...

(Top Level)

Quarto

Environment History Connections Git Tutorial

Import Dataset

172 MiB

List

R

Global Environment

Environment is empty

Console Terminal Background Jobs

R 4.3.0 · /cloud/project/

```
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

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.
```

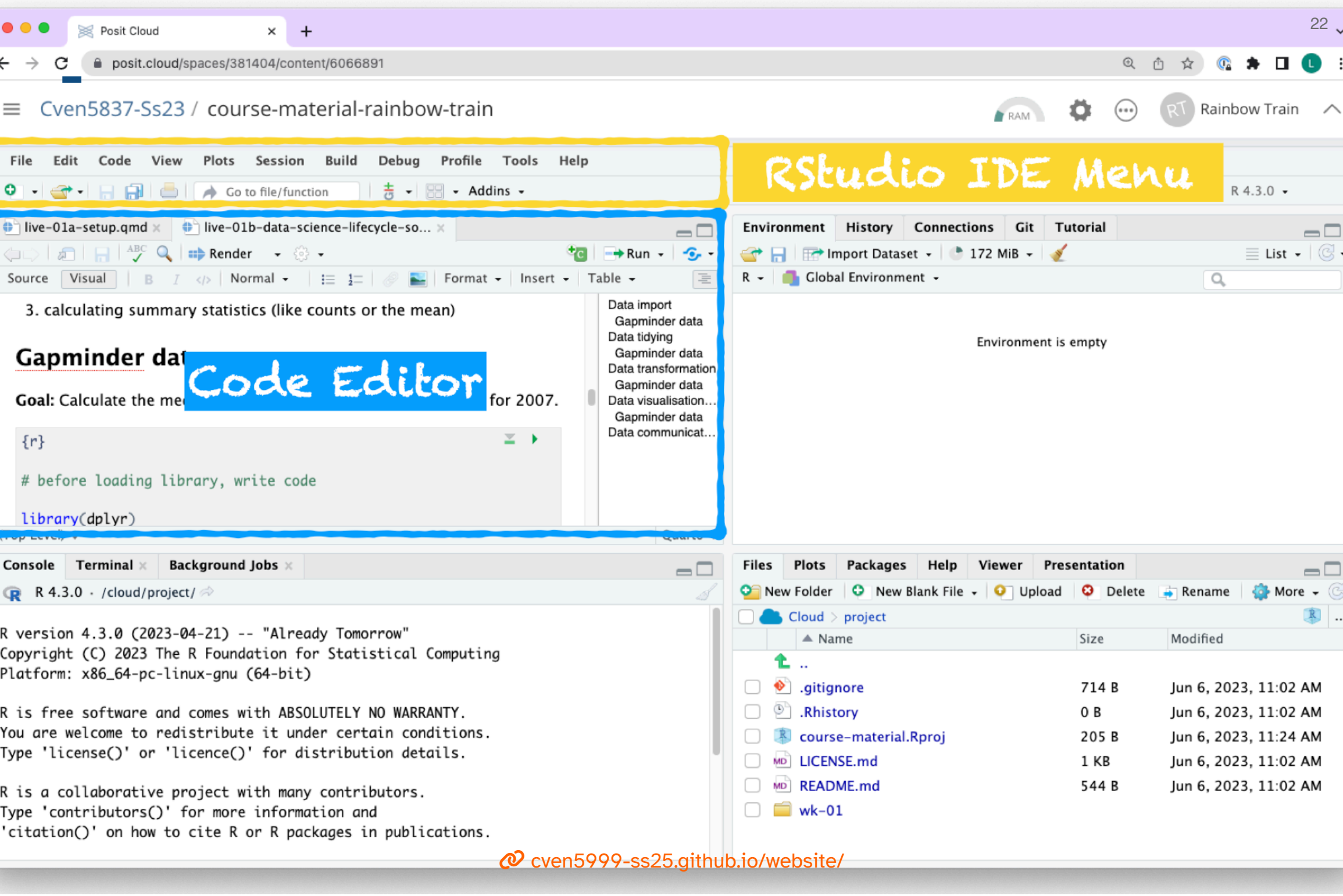
Files Plots Packages Help Viewer Presentation

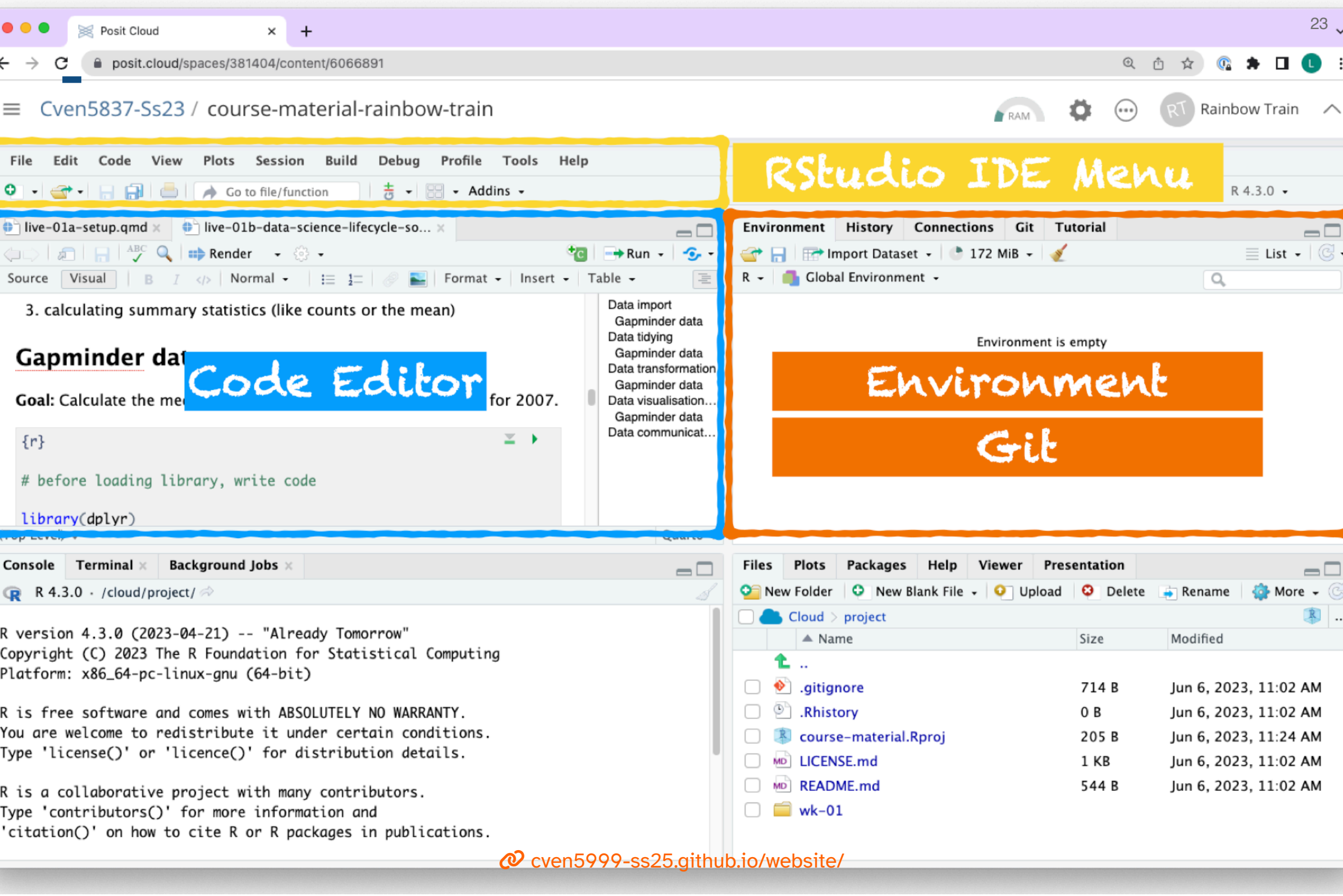
New Folder New Blank File Upload Delete Rename More

Cloud > project

Name	Size	Modified
..		
.gitignore	714 B	Jun 6, 2023, 11:02 AM
.Rhistory	0 B	Jun 6, 2023, 11:02 AM
course-material.Rproj	205 B	Jun 6, 2023, 11:24 AM
LICENSE.md	1 KB	Jun 6, 2023, 11:02 AM
README.md	544 B	Jun 6, 2023, 11:02 AM
wk-01		

cven5999-ss25.github.io/website/


Posit Cloud
posit.cloud/spaces/381404/content/6066891
Cven5837-Ss23 / course-material-rainbow-train
RAM
Rainbow Train
RStudio IDE Menu
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.3.0
live-01a-setup.qmd
live-01b-data-science-lifecycle-so...
Render
Run
Source Visual
Normal
Format Insert Table
3. calculating summary statistics (like counts or the mean)
Gapminder dat
Code Editor
Goal: Calculate the me
for 2007.
{r}
# before loading library, write code
library(dplyr)
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
Environment History Connections Git Tutorial
Import Dataset
172 MiB
List
R
Global Environment
Environment is empty
Console Terminal Background Jobs
R 4.3.0 · /cloud/project/
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.
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.
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud > project
Name
Size
Modified
..
.gitignore 714 B Jun 6, 2023, 11:02 AM
.Rhistory 0 B Jun 6, 2023, 11:02 AM
course-material.Rproj 205 B Jun 6, 2023, 11:24 AM
LICENSE.md 1 KB Jun 6, 2023, 11:02 AM
README.md 544 B Jun 6, 2023, 11:02 AM
wk-01
cven5999-ss25.github.io/website/


Posit Cloud
posit.cloud/spaces/381404/content/6066891
Cven5837-Ss23 / course-material-rainbow-train
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
RStudio IDE Menu
Go to file/function
Addins
R 4.3.0
live-01a-setup.qmd
live-01b-data-science-lifecycle-so...
Render
Run
Source Visual
Normal
Format Insert Table
3. calculating summary statistics (like counts or the mean)
Gapminder dat
Code Editor
Goal: Calculate the me
for 2007.
{r}
# before loading library, write code
library(dplyr)
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
Environment History Connections Git Tutorial
Import Dataset
172 MiB
List
R
Global Environment
Environment is empty
Environment
Git
Console Terminal Background Jobs
R 4.3.0 · /cloud/project/
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.
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.
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud > project
Name
Size
Modified
..
.gitignore 714 B Jun 6, 2023, 11:02 AM
.Rhistory 0 B Jun 6, 2023, 11:02 AM
course-material.Rproj 205 B Jun 6, 2023, 11:24 AM
LICENSE.md 1 KB Jun 6, 2023, 11:02 AM
README.md 544 B Jun 6, 2023, 11:02 AM
wk-01
cven5999-ss25.github.io/website/

Posit Cloud

posit.cloud/spaces/381404/content/6066891

Cven5837-Ss23 / course-material-rainbow-train

RAM

Rainbow Train

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Addins

RStudio IDE Menu

R 4.3.0

live-01a-setup.qmd

live-01b-data-science-lifecycle-so...

Render

Run

Source Visual

Normal

Format Insert Table

3. calculating summary statistics (like counts or the mean)

Gapminder dat

Code Editor

Goal: Calculate the me for 2007.

```{r}

# before loading library, write code

library(dplyr)
```

Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...

Environment History Connections Git Tutorial

Import Dataset

172 MiB

List

R

Global Environment

Environment is empty

Environment

Git

Console Terminal Background Jobs

R 4.3.0 · /cloud/project/

```
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gn

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.

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.
```

Console

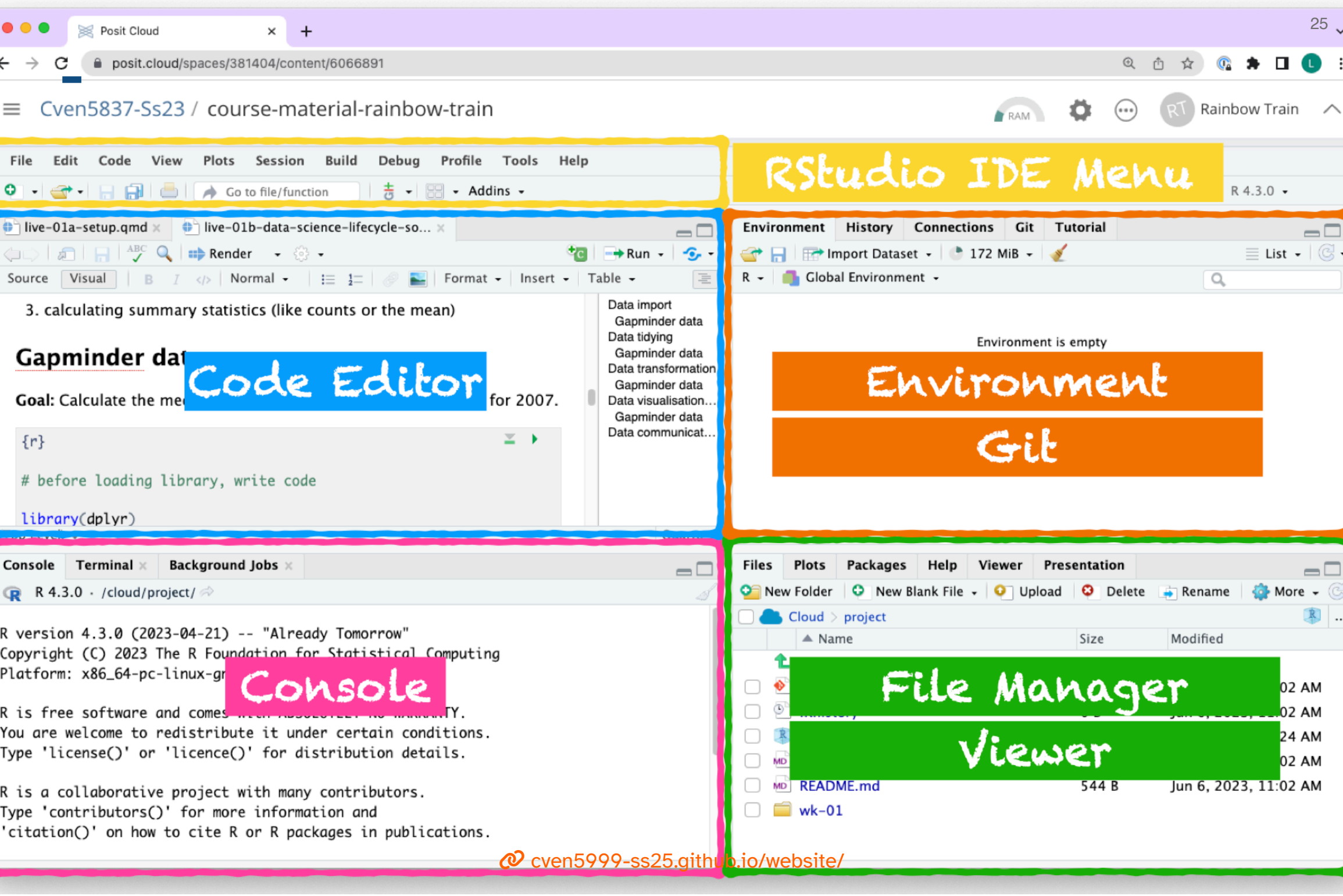
Files Plots Packages Help Viewer Presentation

New Folder New Blank File Upload Delete Rename More

Cloud > project

Name	Size	Modified
..		
.gitignore	714 B	Jun 6, 2023, 11:02 AM
.Rhistory	0 B	Jun 6, 2023, 11:02 AM
course-material.Rproj	205 B	Jun 6, 2023, 11:24 AM
LICENSE.md	1 KB	Jun 6, 2023, 11:02 AM
README.md	544 B	Jun 6, 2023, 11:02 AM
wk-01		

cven5999-ss25.github.io/website/


Posit Cloud
posit.cloud/spaces/381404/content/6066891
Cven5837-Ss23 / course-material-rainbow-train
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RStudio IDE Menu
R 4.3.0
live-01a-setup.qmd
live-01b-data-science-lifecycle-so...
Render
Run
Source Visual
Normal
Format Insert Table
3. calculating summary statistics (like counts or the mean)
Gapminder dat
Code Editor
Goal: Calculate the me
for 2007.
{r}
# before loading library, write code
library(dplyr)
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
Environment History Connections Git Tutorial
Import Dataset
172 MiB
List
R
Global Environment
Environment is empty
Environment
Git
Console Terminal Background Jobs
R 4.3.0 · /cloud/project/
R version 4.3.0 (2023-04-21) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gn
Console
R is free software and comes
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
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.
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud > project
Name
Size
Modified
File Manager
Viewer
README.md
544 B
Jun 6, 2023, 11:02 AM
wk-01
cven5999-ss25.github.io/website/

Screen setup

Who uses a setup with one screen?

“One screen” in the Zoom Chat

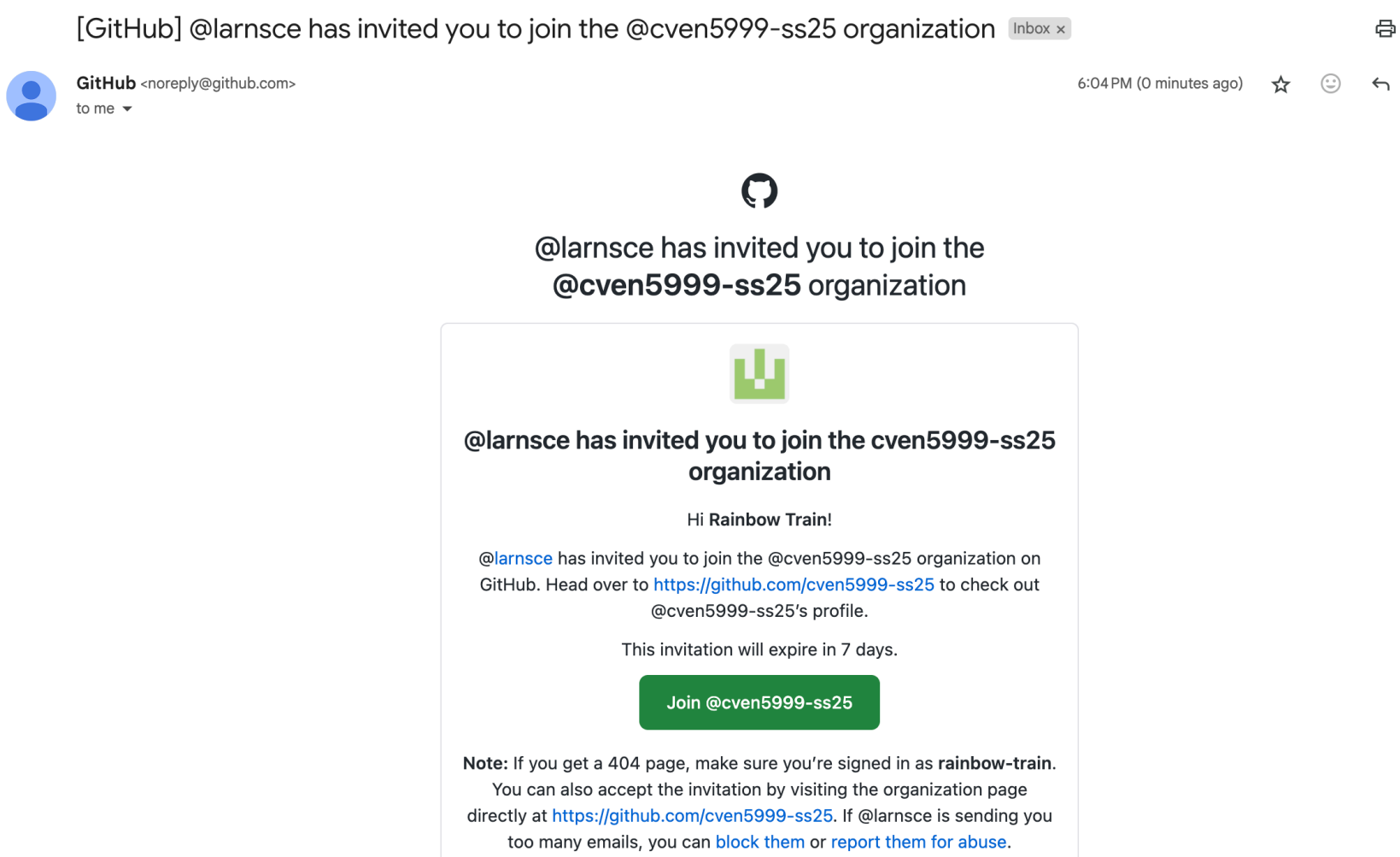
Screen setup

Who uses a setup with two screens?

“Two screens” in the Zoom Chat

Email from GitHub?

Please accept the invitation to the GitHub organisation for the course



cven5999-ss25.github.io/website/

Live Coding Exercise

cven5999-ss25.github.io/website/

git-configuration

Follow along on the screen

1. Open the GitHub organisation for the course: https://github.com/cven5999-ss25
2. You will find a repository titled: **wk-02-USERNAME** (with your GitHub Username)
3. You will “clone” this repository to Posit Cloud

Break

> **GitHub PAT from week 1**
>
> Do you have your **GitHub Personal Access Token** readily accessible?

10:00

Photo by Blake Wisz

cven5999-ss25.github.io/website/

Version Control

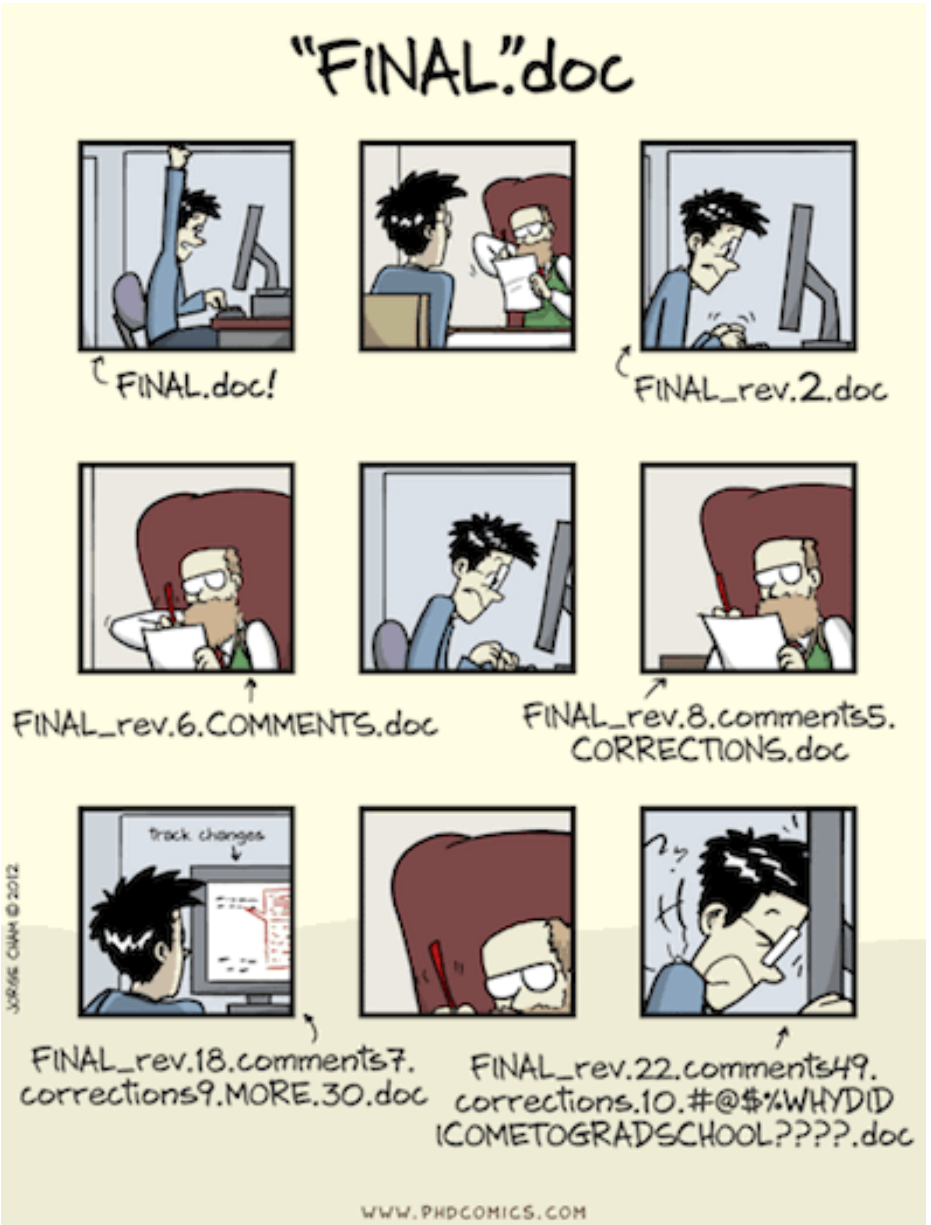
Version Control with Git and GitHub

A way to share files with others, so they can:

- download
- re-use
- contribute

You can view the history of files, and jump back in time to any point.

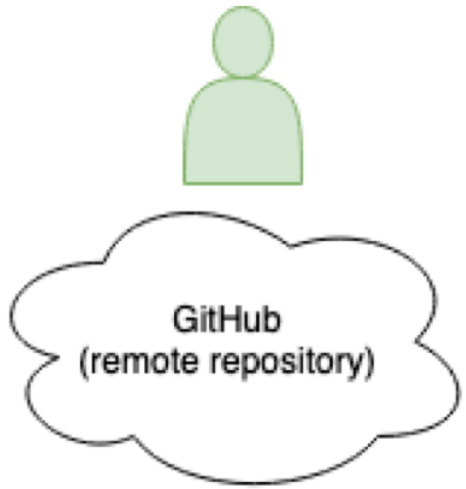
Why is it useful?

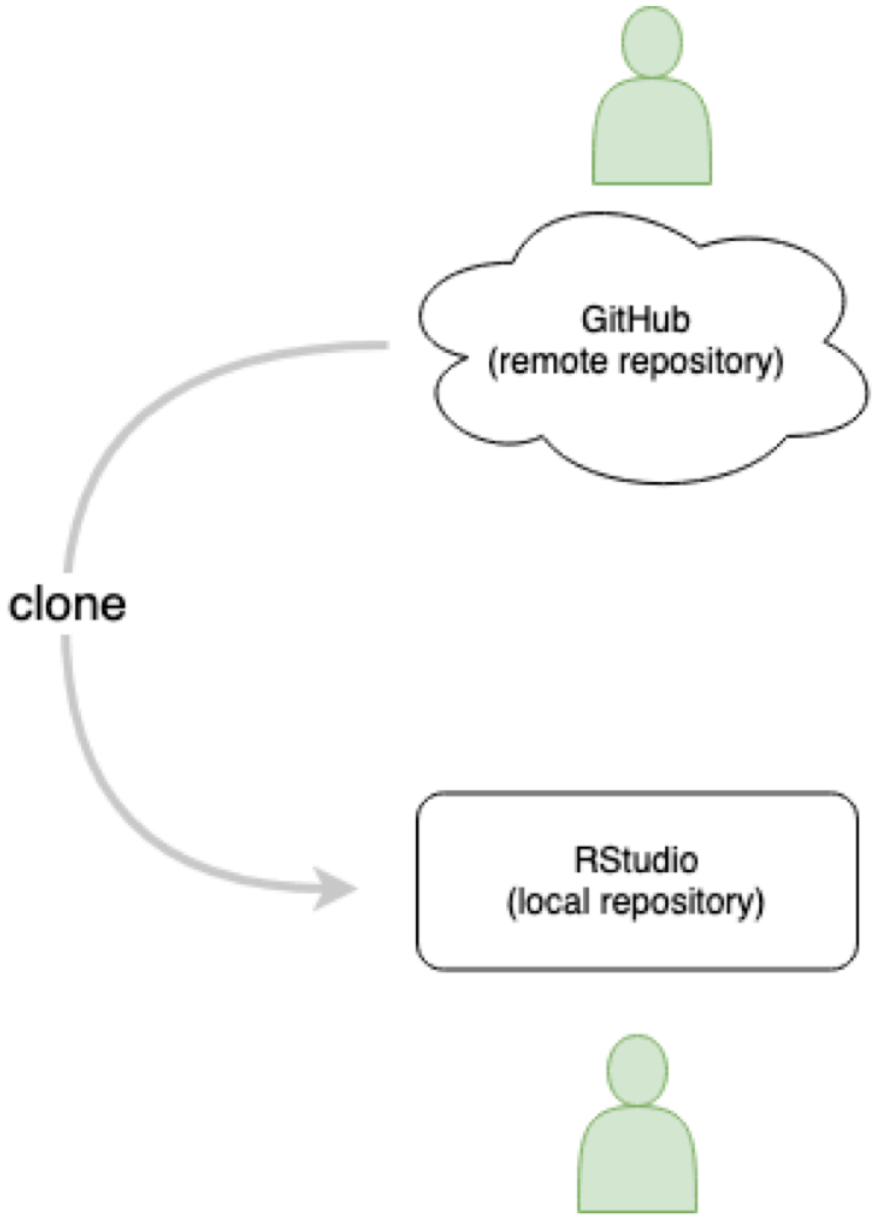


cven5999-ss25.github.io/website/

Git and GitHub

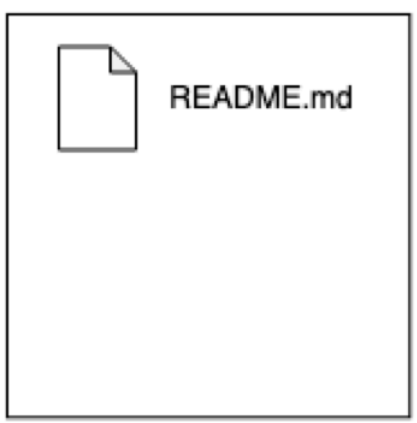
Version Control - Terminology

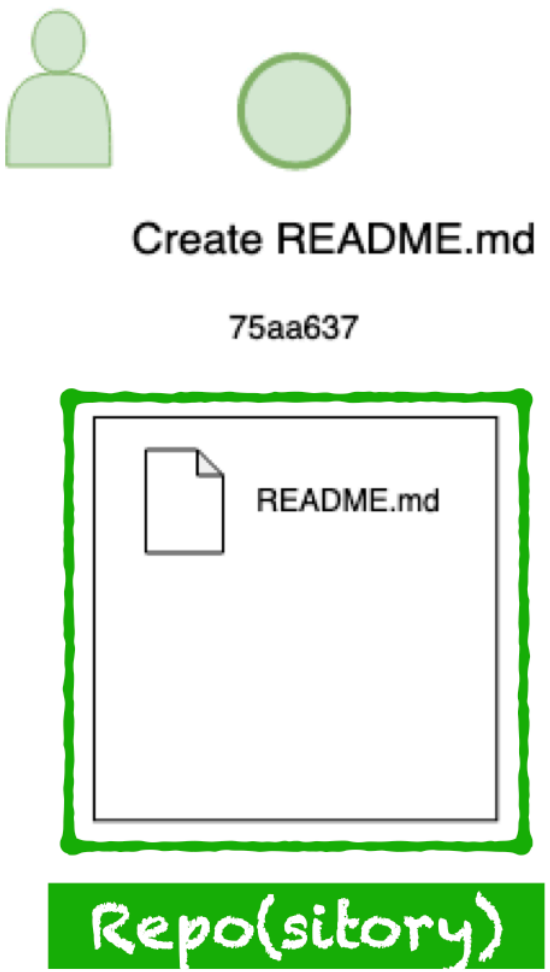

GitHub
(remote repository)

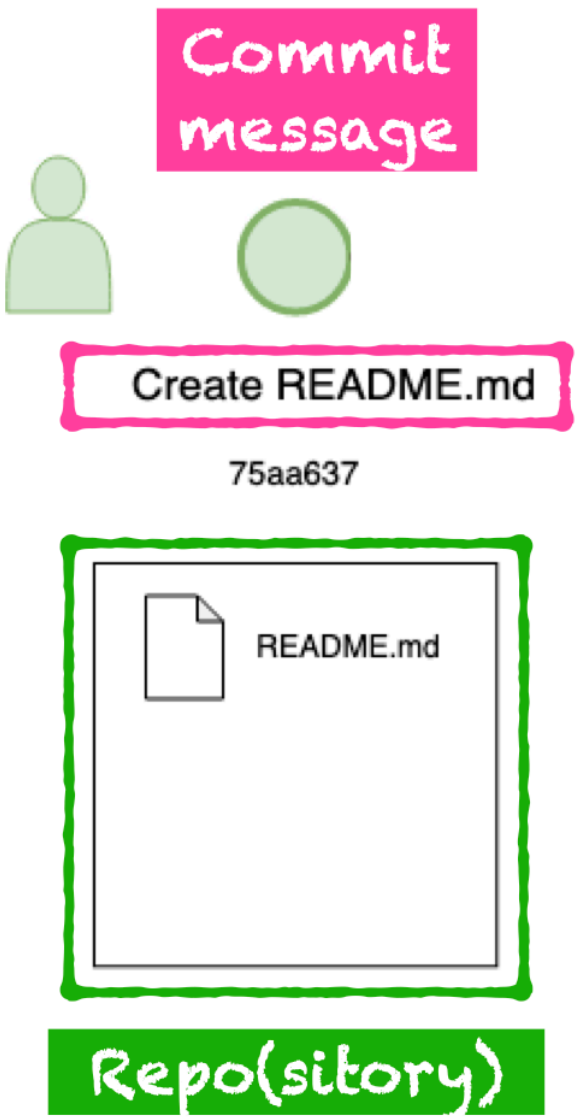

GitHub
(remote repository)
clone
RStudio
(local repository)

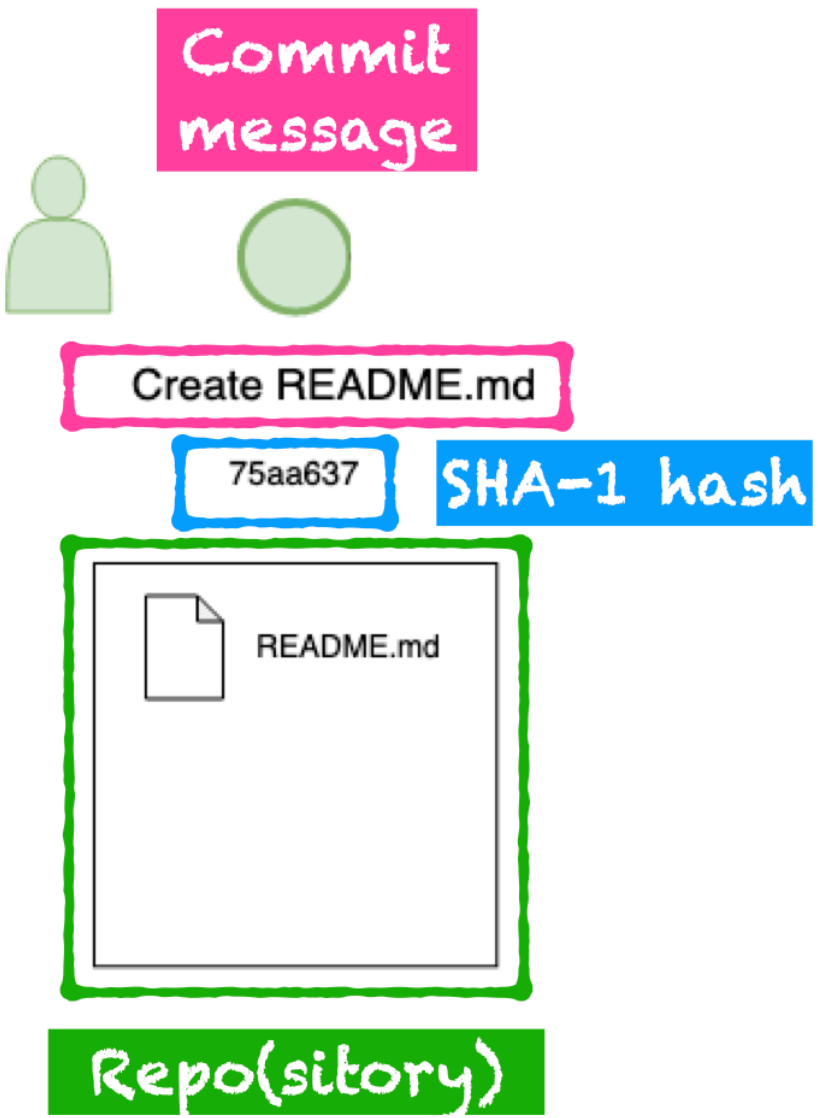
Create README.md

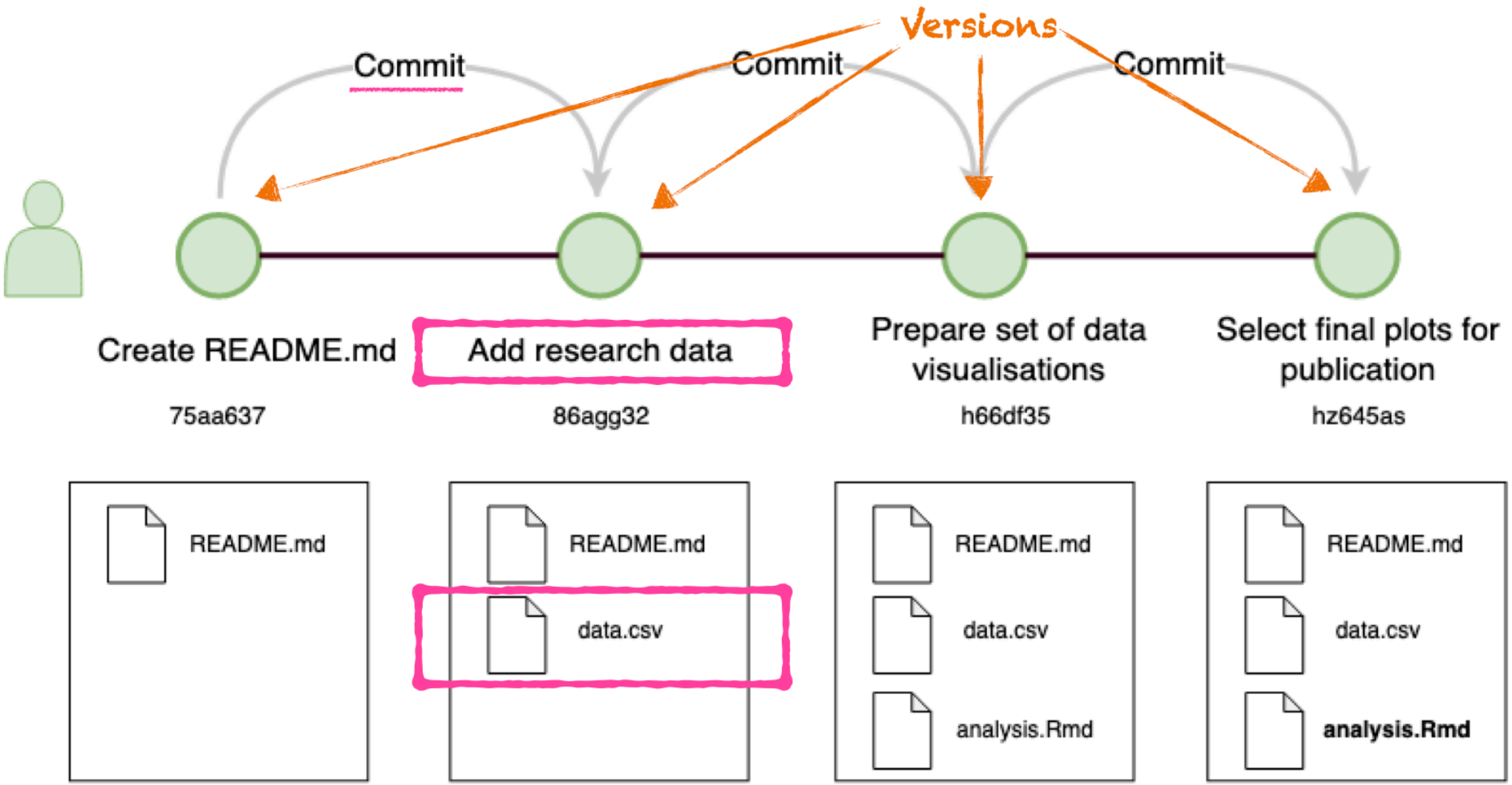
75aa637

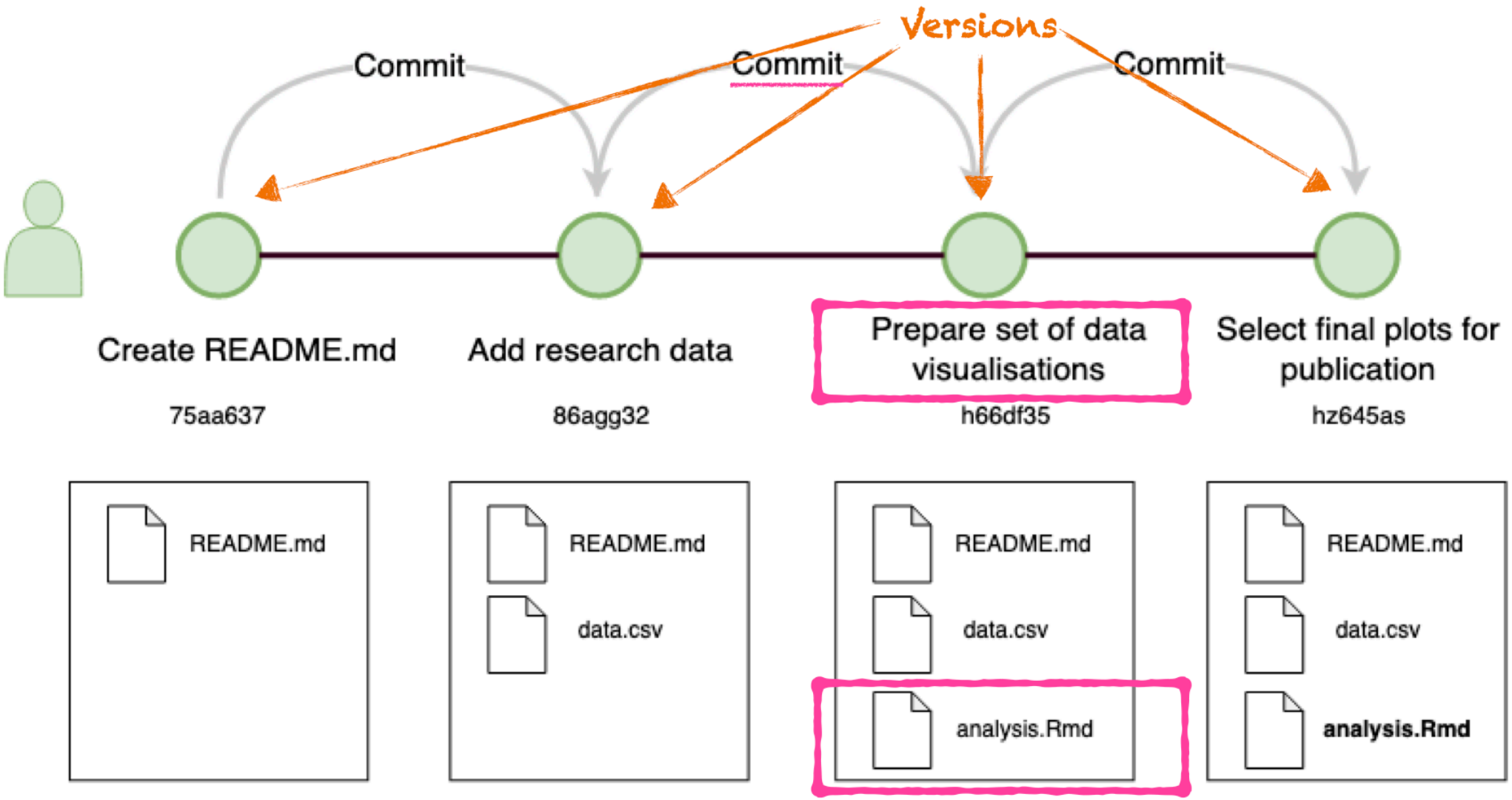


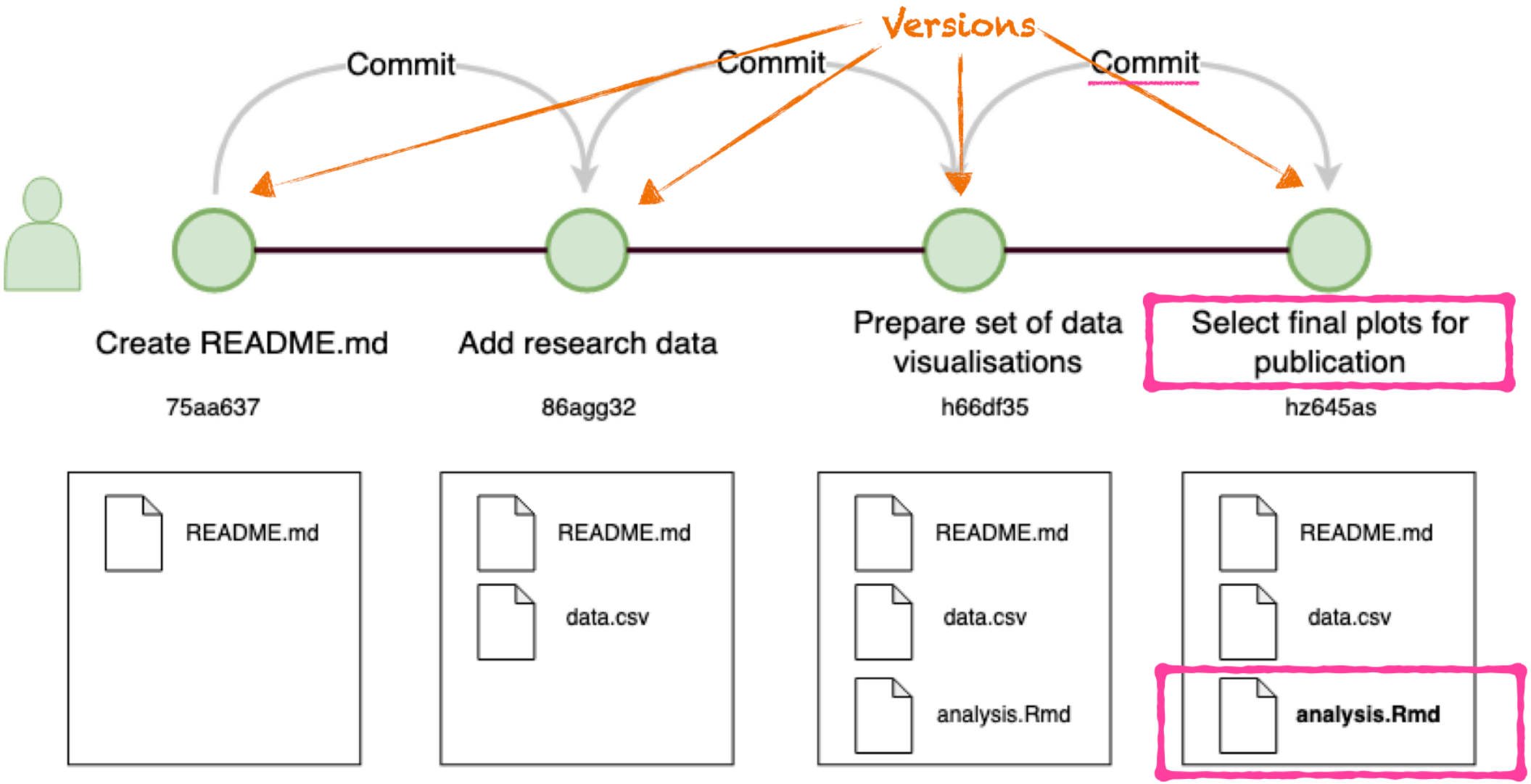

Create README.md
75aa637
README.md
Repo(sitory)

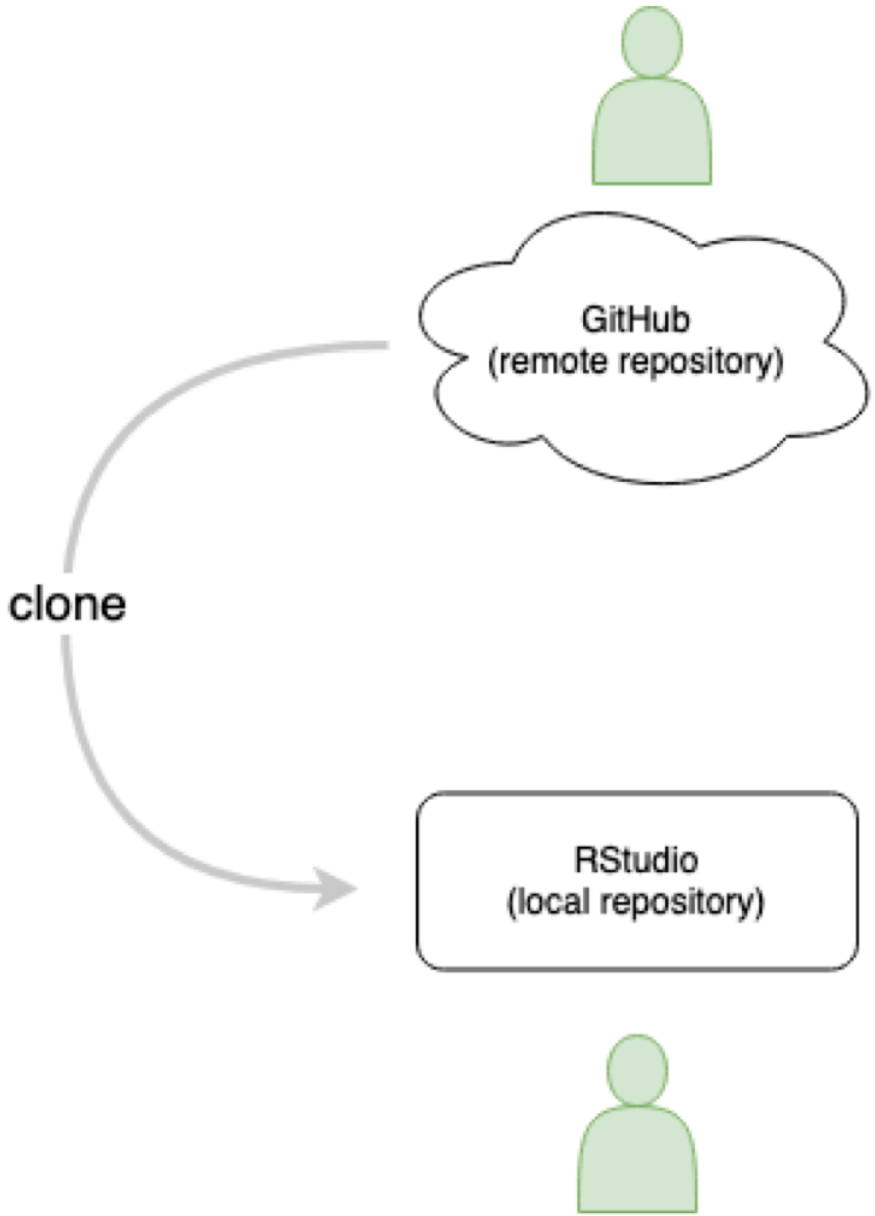

Commit message
Create README.md
75aa637
README.md
Repo(sitory)

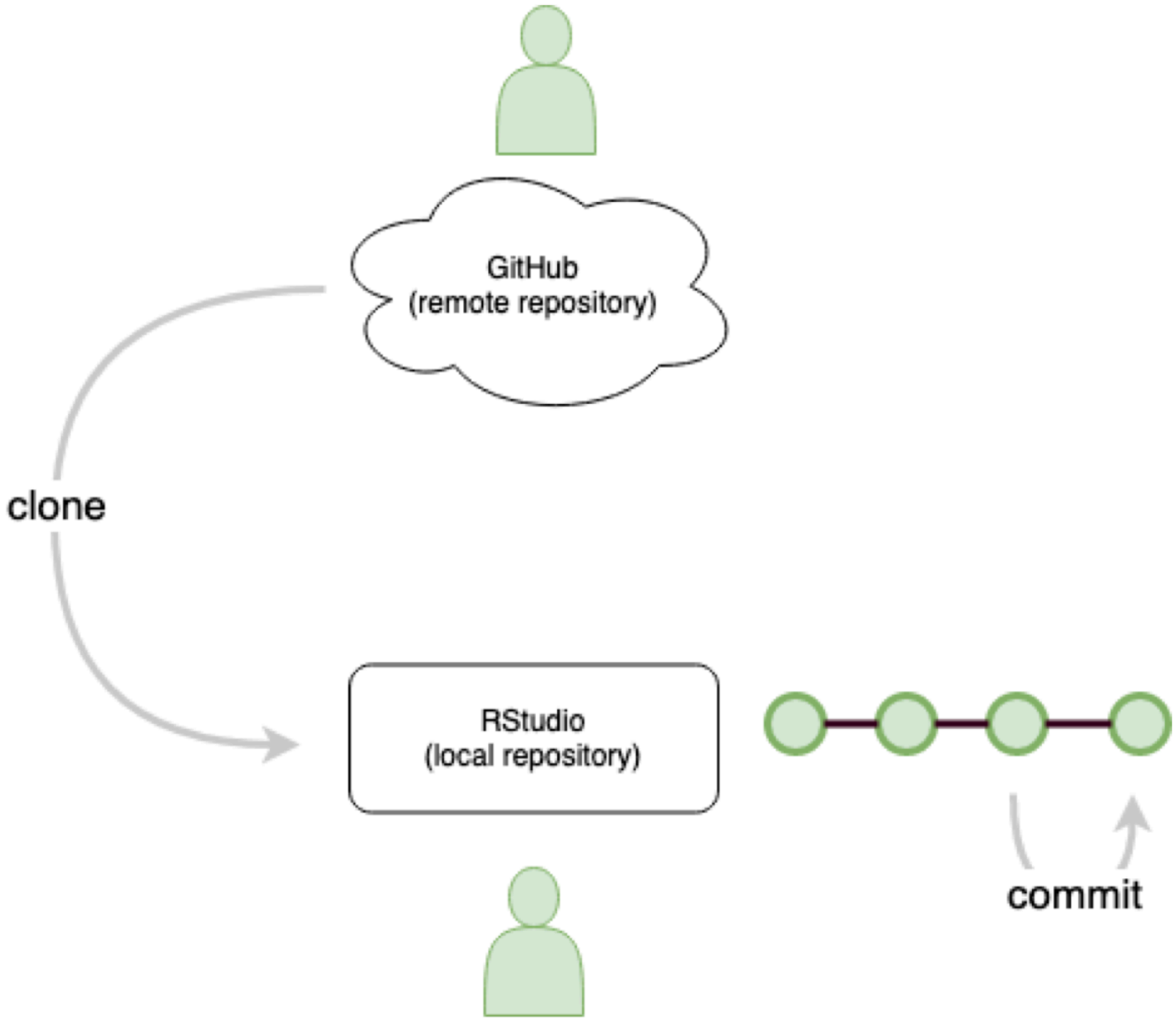

Commit message
Create README.md
75aa637
SHA-1 hash
README.md
Repo(sitory)

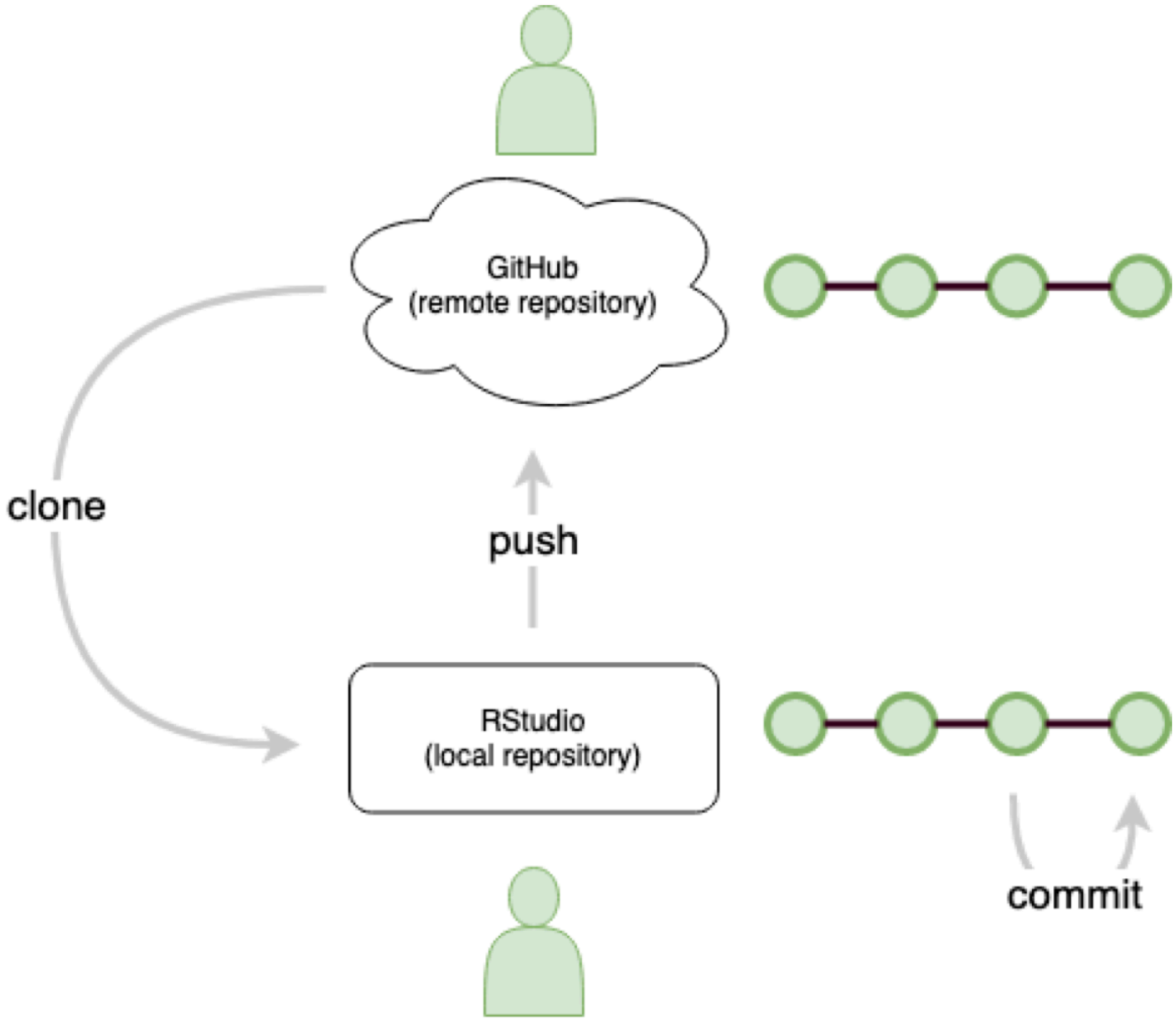

Versions
Commit
Commit
Commit
Create README.md
Add research data
Prepare set of data visualisations
Select final plots for publication
75aa637
86agg32
h66df35
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd


Versions
Commit
Commit
Commit
Create README.md
75aa637
Add research data
86agg32
Prepare set of data visualisations
h66df35
Select final plots for publication
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd


Versions
Commit
Commit
Commit
Create README.md
75aa637
Add research data
86agg32
Prepare set of data visualisations
h66df35
Select final plots for publication
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd


GitHub
(remote repository)
clone
RStudio
(local repository)


GitHub
(remote repository)
clone
RStudio
(local repository)
commit

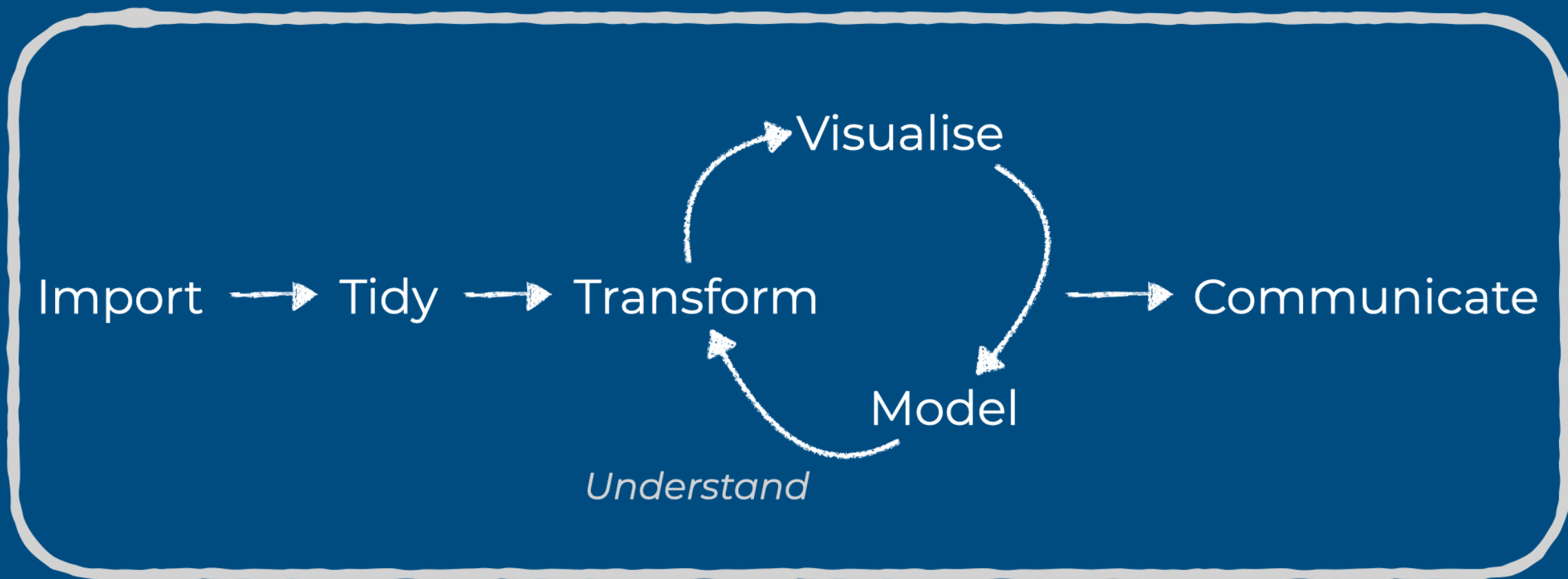

GitHub
(remote repository)
clone
push
RStudio
(local repository)
commit

Data Science Lifecycle

Deep End

via GIPHY

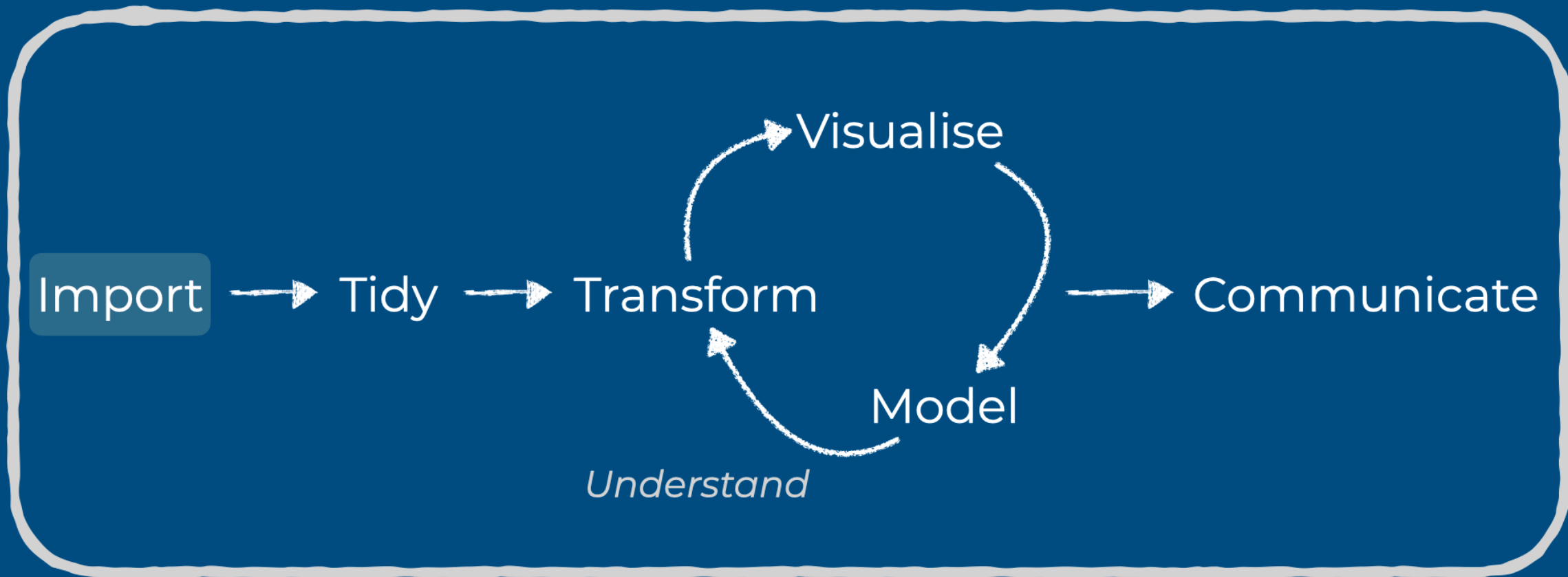
Data Science Lifecycle



From: https://r4ds.had.co.nz/

Data Science Lifecycle

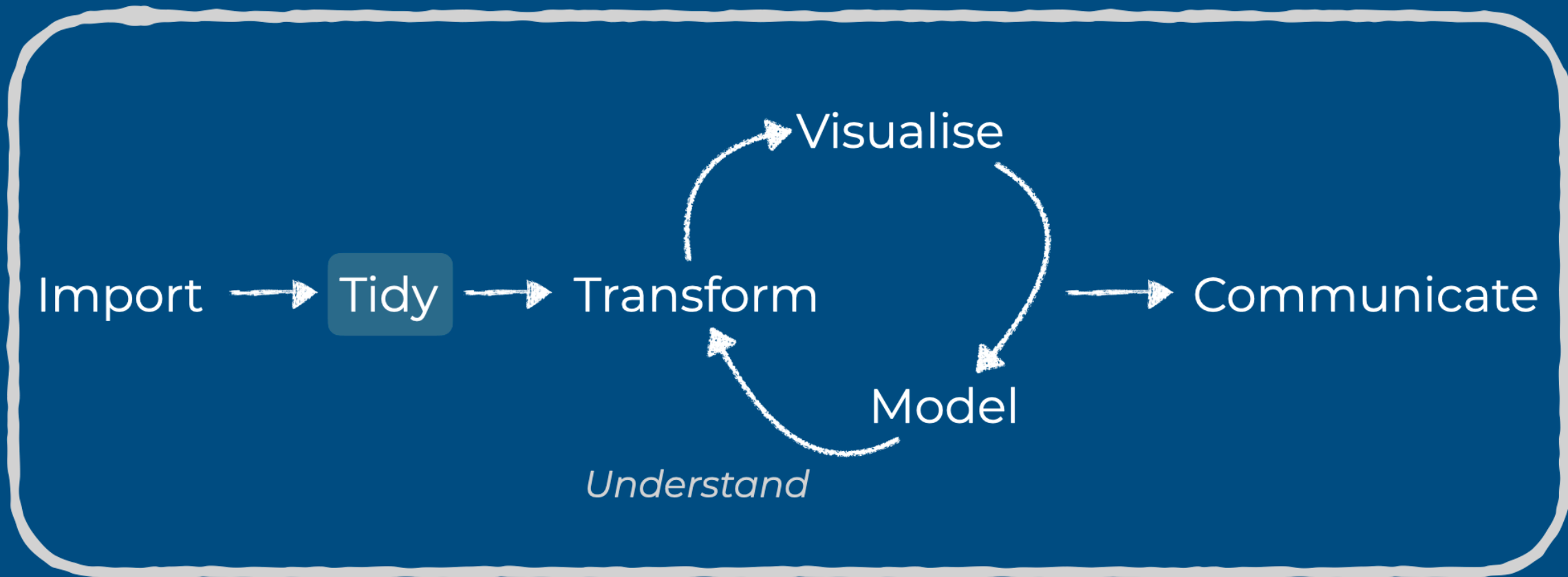
Get your data into R



From: https://r4ds.had.co.nz/

Data Science Lifecycle

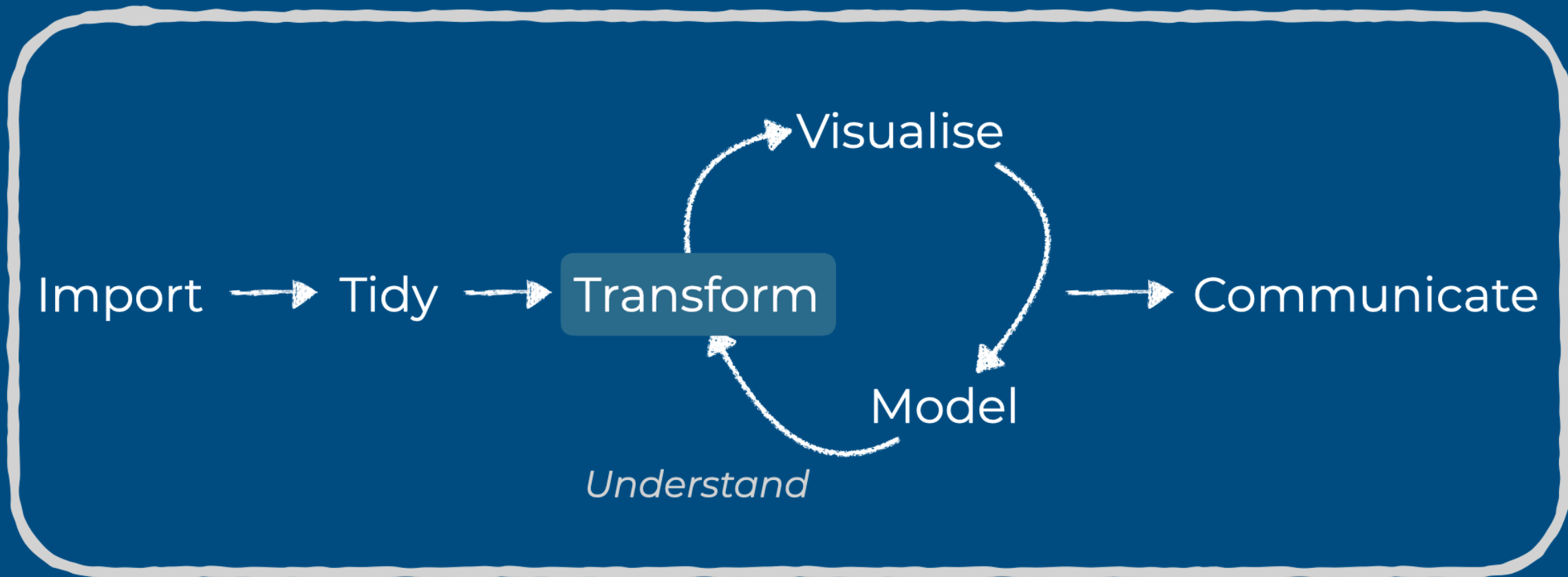
Store your data in a consistent form



From: https://r4ds.had.co.nz/

Data Science Lifecycle

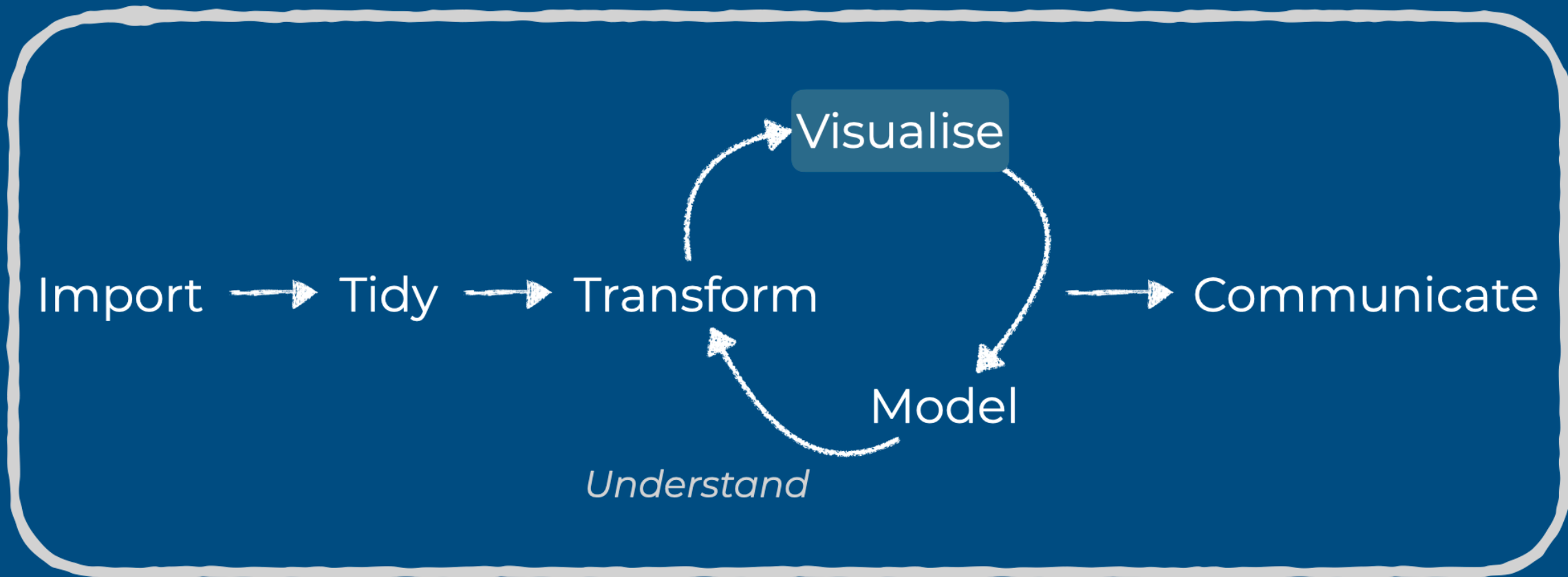
Narrow down + Create new variables + Summary stats



From: https://r4ds.had.co.nz/

Data Science Lifecycle

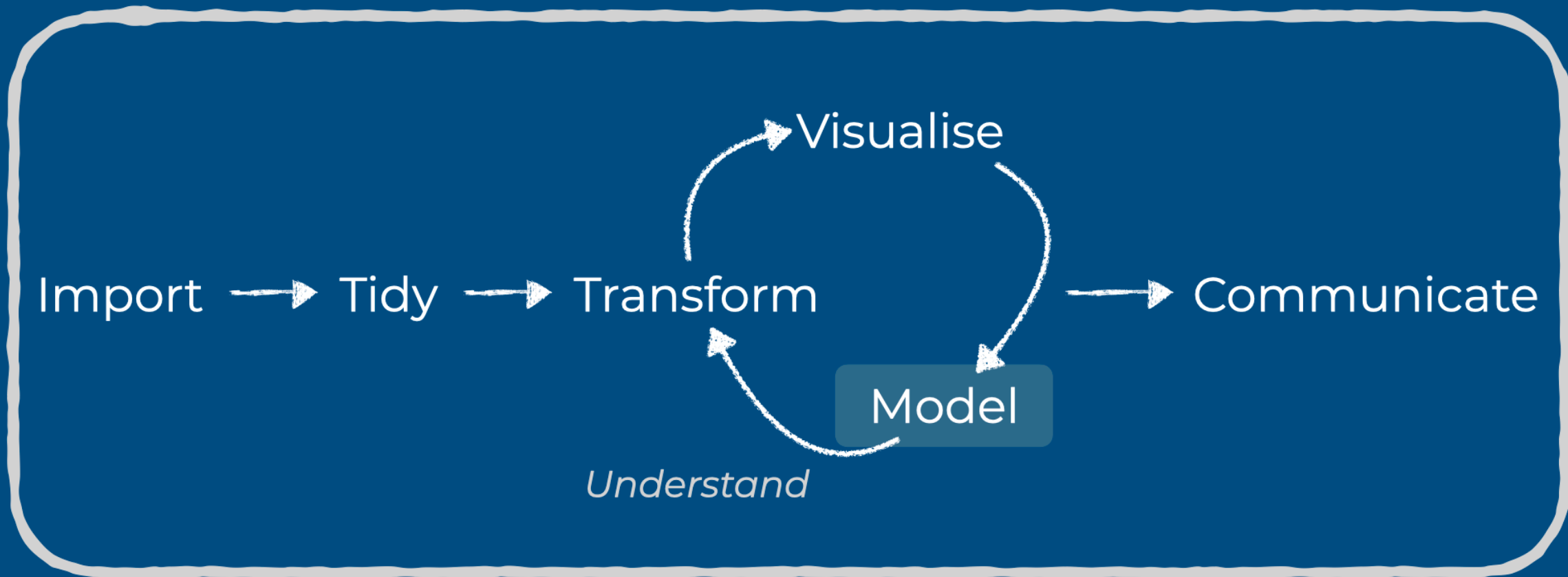
Explore your with visual representations



From: https://r4ds.had.co.nz/

Data Science Lifecycle

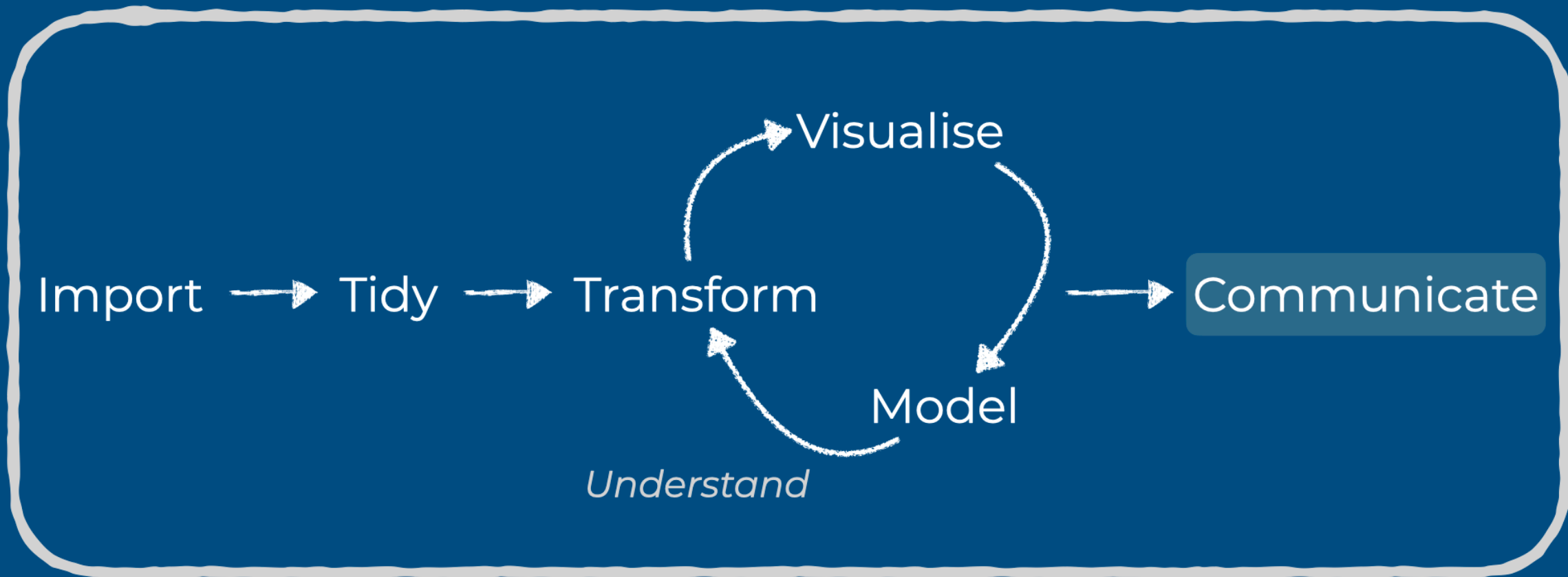
Explore your with visual representations



From: https://r4ds.had.co.nz/

Data Science Lifecycle

Share your findings with others



From: https://r4ds.had.co.nz/

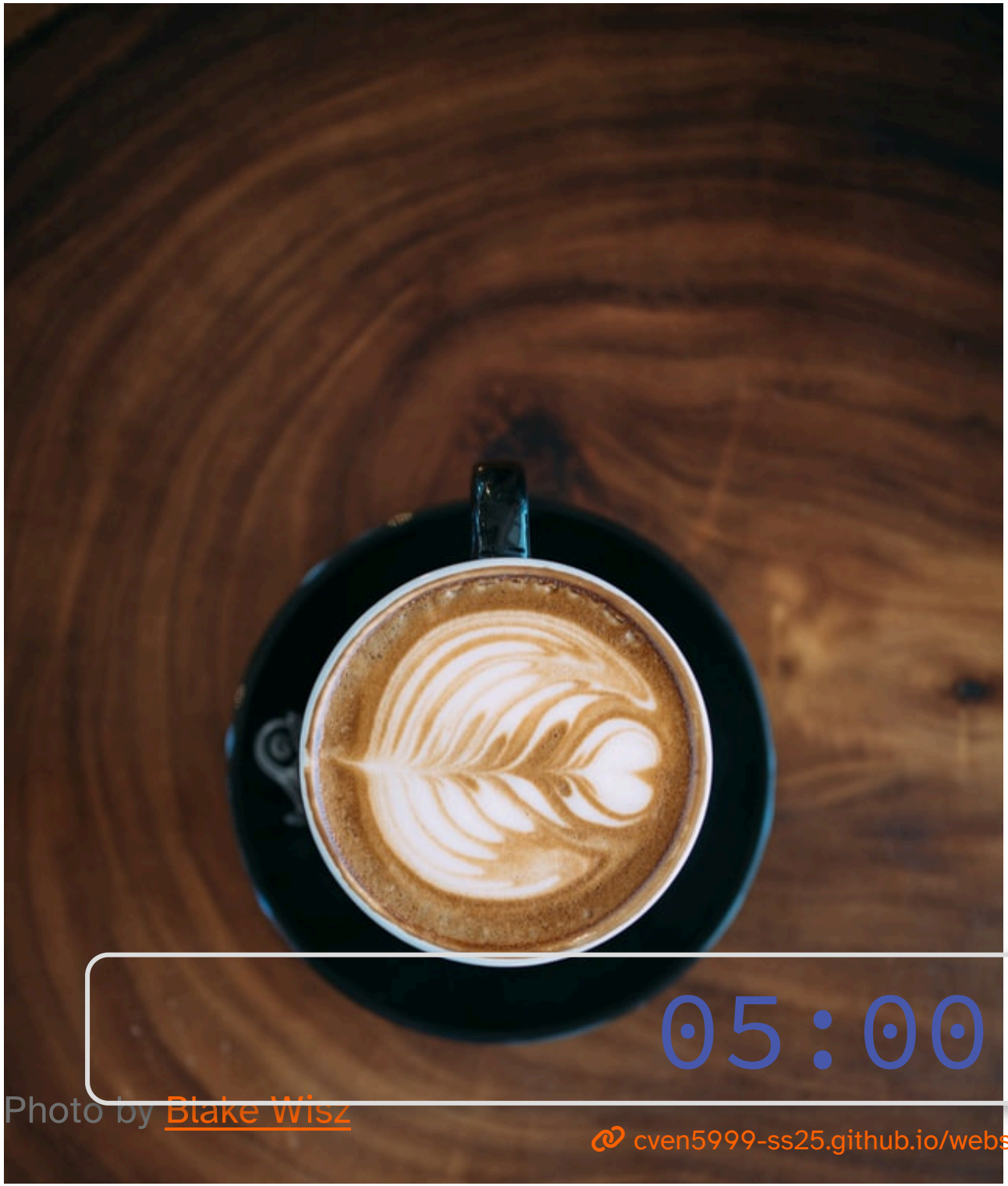
Live Coding Exercise

cven5999-ss25.github.io/website/

live-data-science-lifecycle

1. Head over to posit.cloud
2. Open the workspace for the course cven5999-ss25
3. Open “Projects”
4. Open the “wk-02-USERNAME” project
5. Follow along with me

Break



05:00

Photo by Blake Wisz

cven5999-ss25.github.io/website/

R

Packages

base R

```
1 sqrt(49)
2 sum(1, 2)
```

- Functions come with R

R Packages

```
1 library(dplyr)
```

- Installed once in the Console: `install.packages("dplyr")`
- Loaded per script

Functions & Arguments

```
1 library(dplyr)
2
3 filter(.data = gapminder,
4        year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**

Objects

```
1 library(dplyr)
2
3 gapminder_yr_2007 <- filter(.data = gapminder,
4                             year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**
- Object: `gapminder_yr_2007`

Operators

```
1 library(dplyr)
2
3 gapminder_yr_2007 <- gapminder |>
4   filter(year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**
- Object: `gapminder_yr_2007`
- Assignment operator: `<-`
- Pipe operator: `|>`

Rules

Rules of `dplyr` functions:

- First argument is always a data frame
- Subsequent arguments say what to do with that data frame
- Always return a data frame
- Don’t modify in place

Course information

cven5999-ss25.github.io/website/

Weekly Structure

Monday	Lecture
Tuesday	
Wednesday	
Thursday	Feedback (grading) on assignments from previous week
Friday	Homework assignment and learning reflection are due

Homework assignments

- Weekly programming assignments
- Graded as pass/fail (100 pts)
- Submitted as rendered Quarto documents on GitHub
- weighted at 40% of the total grade

Learning reflections

- Reflections on the different class elements (lecture, homework assignment, readings)
- Graded as pass/fail (100 pts)
- minimum 100 words
- Submitted as rendered Quarto documents on GitHub
- weighted at 20% of the total grade

Capstone Project

- Data analysis project report with a data set of your choice
- Graded as number of points out of 100 pts for pre-defined graded elements
- Submitted as rendered Quarto document on GitHub
- weighted at 40% of the total grade

Grading

Conversion from percent to grades.	
grade	percent
A+	97
A	93
A-	90
B+	87
B	83
B-	80
C+	77
C	73
C-	70
D+	67
D	63
D-	60
F	0

Late work policy

- due dates are set and all work is due on the stated date
- work not submitted by the due date will receive 0 pts
- the lowest score for each of the assignments or learning reflections is dropped

Homework week 2

cven5999-ss25.github.io/website/

Homework due dates

- All material on course website
- Homework assignment & learning reflection due: **2025-06-13**

Thanks! 

cven5999-ss25.github.io/website/

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/

Access slides as PDF on GitHub