Welcome to Data Analytics for Development

CVEN 5837 - Summer 2023

Lars Schöbitz

https://cven5873-ss23.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER

https://cven5873-ss23.github.io/website/

Welcome! 👋

https://cven5873-ss23.github.io/website/

Meet the lecturer

https://cven5873-ss23.github.io/website/

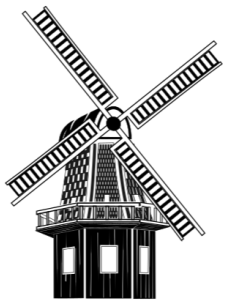
Lars Schöbitz (he/him)

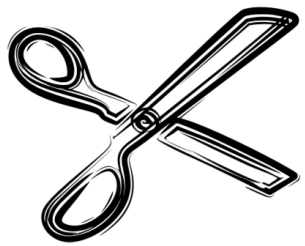
- Environmental Engineer
- Open Science Specialist at ETH Zurich
- Independent Instructor for Data Science with R
- Twitter: @larnsce

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

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

Topics

- Overview of qualitative and quantitative research methods and tools
- The data science life-cycle
- Data organization in spreadsheets
- Exploratory data analysis using visualization
- Concept of tidy data and data tidying
- Data transformation and descriptive statistics
- Data communication using the Quarto open-source scientific and technical publishing system

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).

https://cven5873-ss23.github.io/website/

https://cven5873-ss23.github.io/website/

Classroom tools

https://cven5873-ss23.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

cven5873-ss23.github.io/website/ 

Posit Cloud

https://cven5873-ss23.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

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.
```

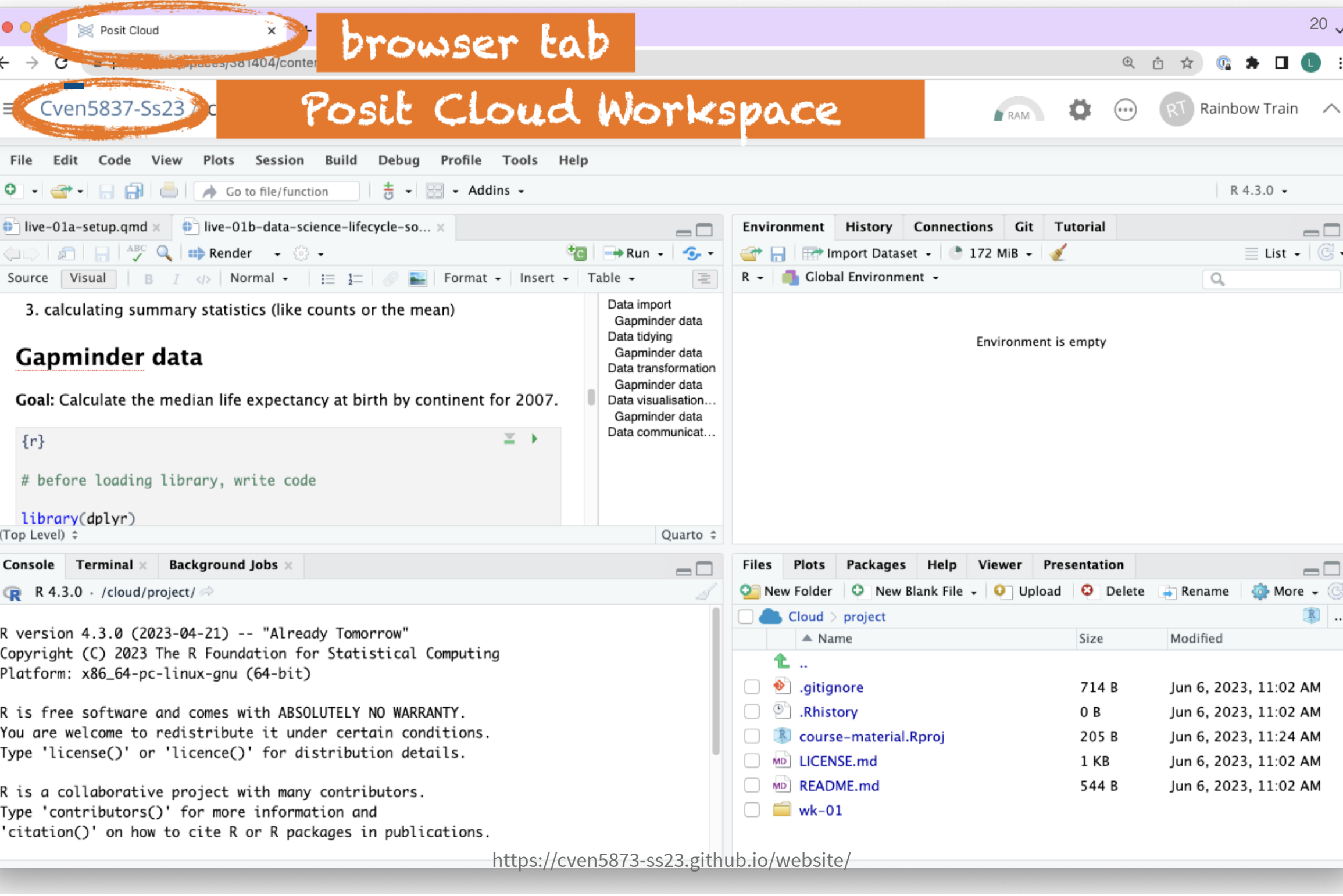
Files Plots Packages Help Viewer Presentation

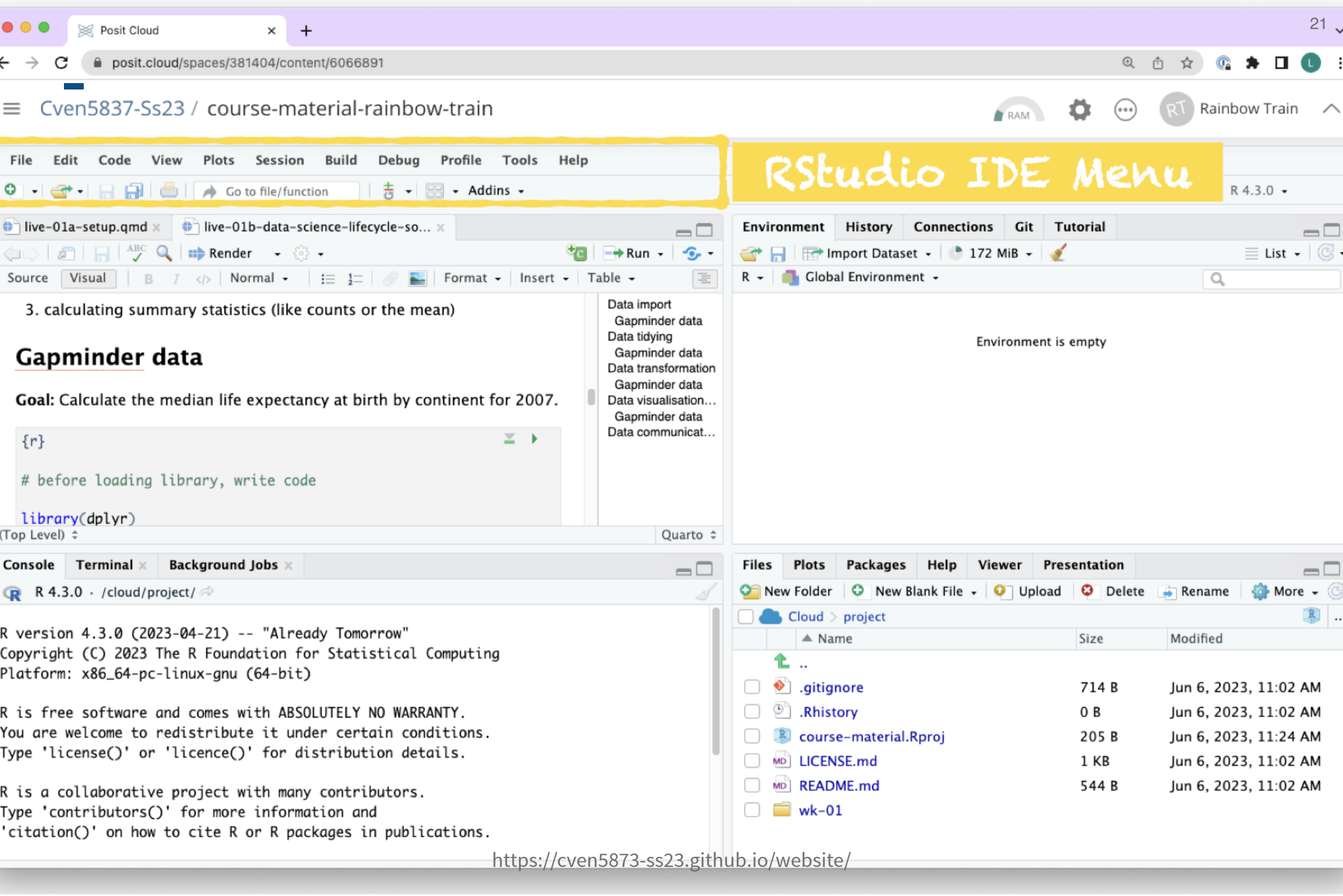
New Folder New Blank File Upload Delete Rename More

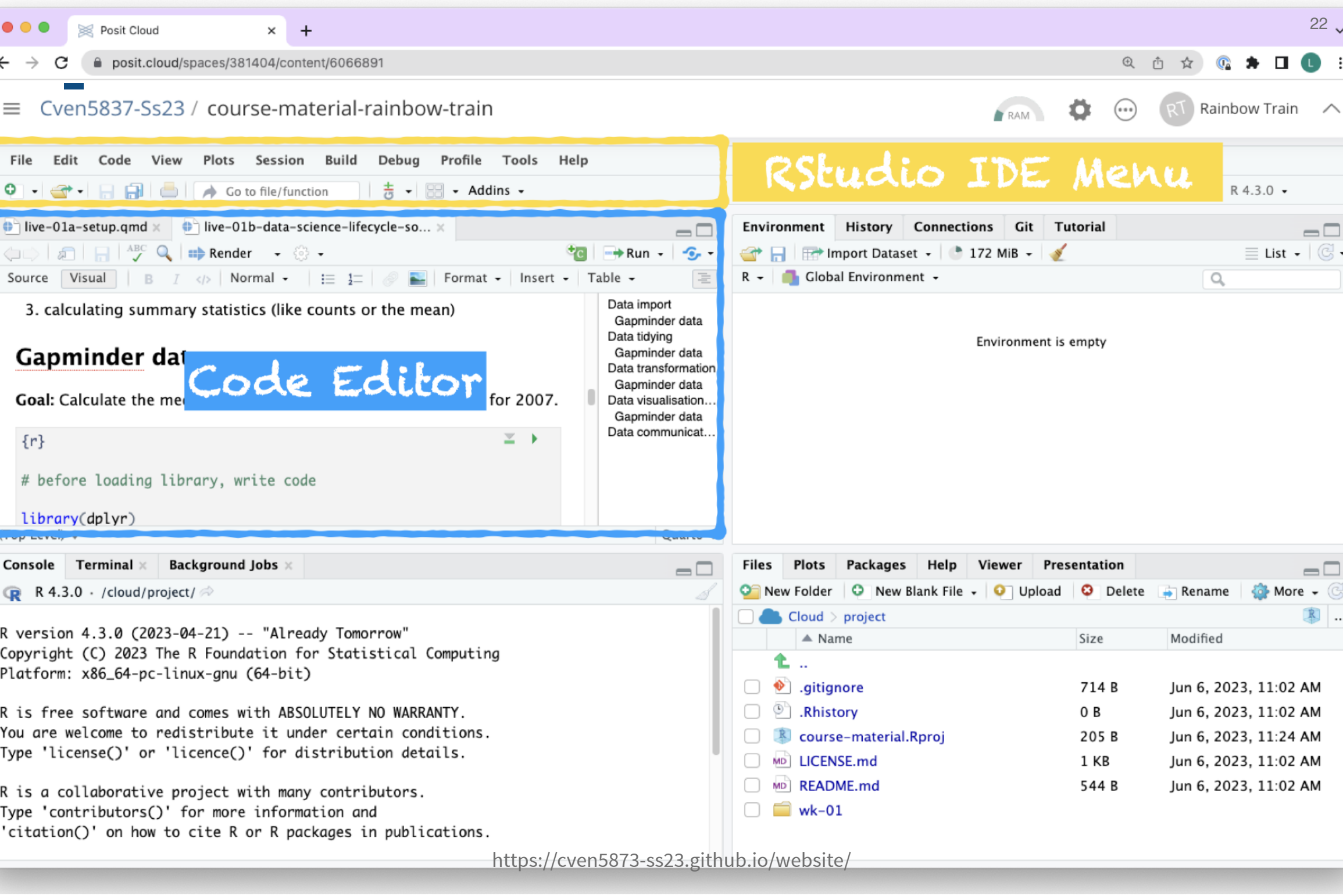
Cloud > project

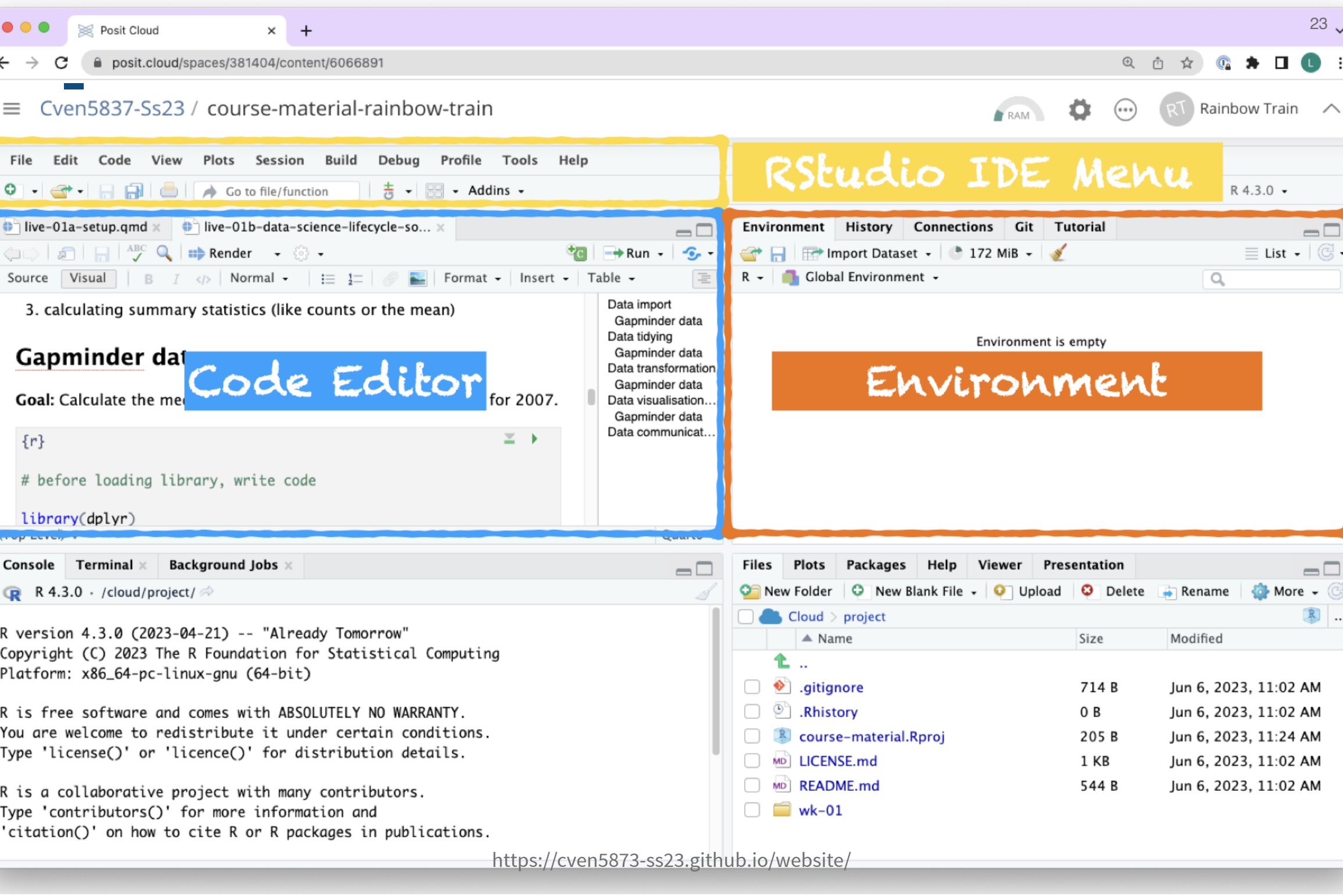
Name	Size	Modified
..		
.gitignore	714 B	Jun 6, 2023, 11:02 AM
.Rhistory	0 B	Jun 6, 2023, 11:02 AM
course-material.Rproj	205 B	Jun 6, 2023, 11:24 AM
LICENSE.md	1 KB	Jun 6, 2023, 11:02 AM
README.md	544 B	Jun 6, 2023, 11:02 AM
wk-01		

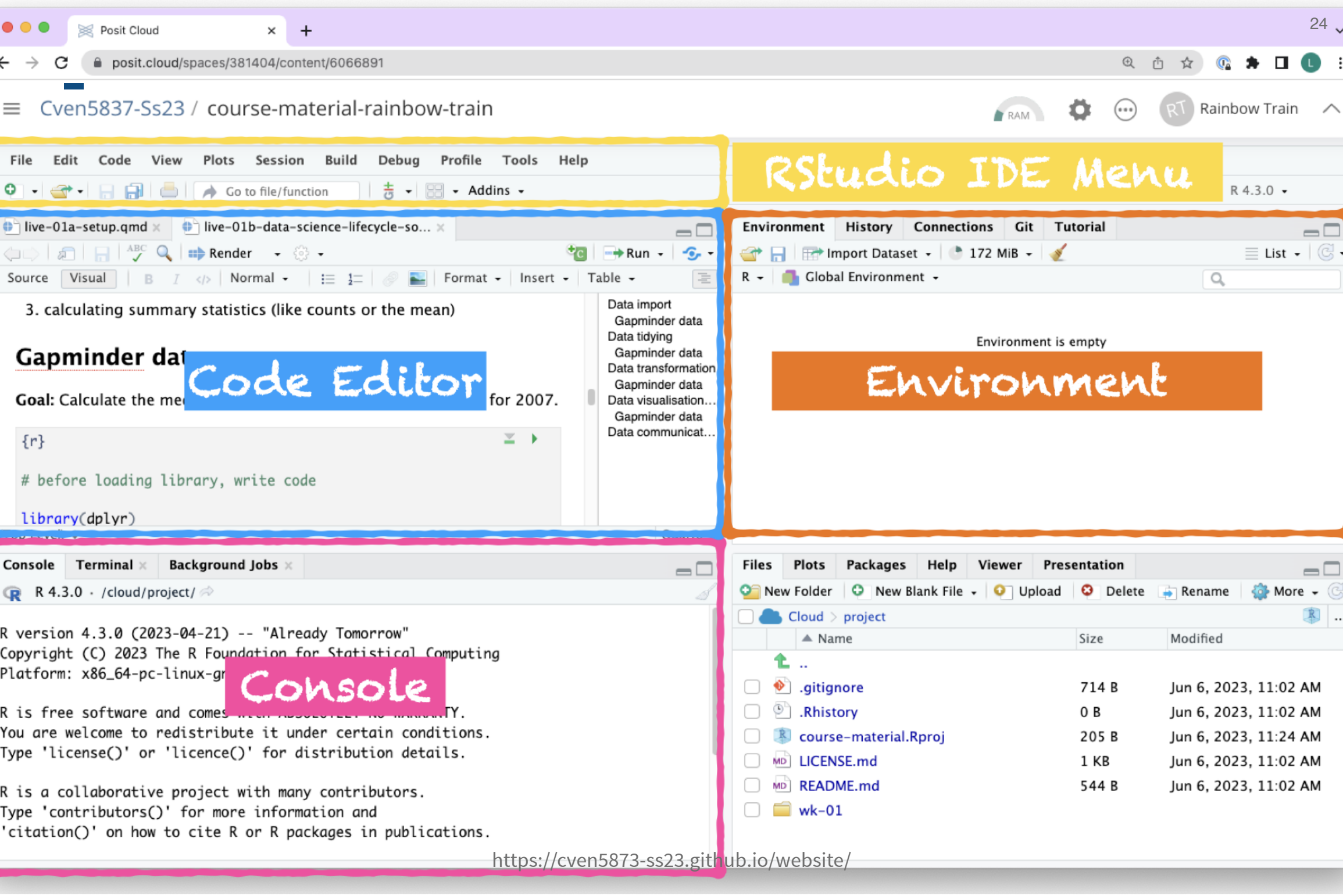
https://cven5873-ss23.github.io/website/

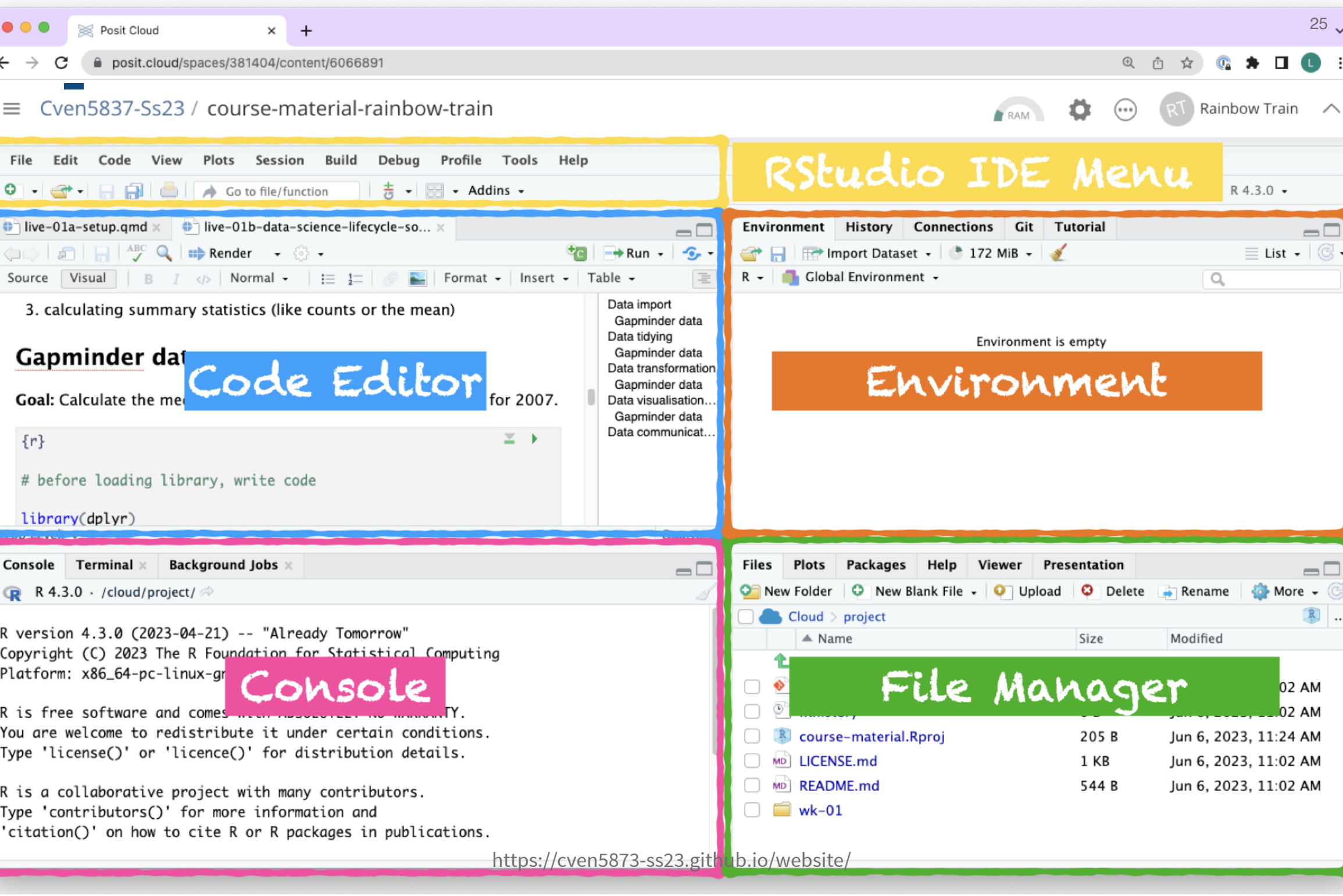

Posit Cloud
browser tab
Cven5837-Ss23
Posit Cloud Workspace
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.3.0
live-01a-setup.qmd
live-01b-data-science-lifecycle-so...
Render
Run
Source Visual
Normal
Format Insert Table
3. calculating summary 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
https://cven5873-ss23.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
https://cven5873-ss23.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
https://cven5873-ss23.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
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
https://cven5873-ss23.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
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
..
.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
https://cven5873-ss23.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
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-gr
Console
R is free software and comes
TY.
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
02 AM
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
File Manager
https://cven5873-ss23.github.io/website/

Screen setup

- Who uses a second external screen?
- “Yes” in the Zoom Chat

Live Coding Exercise

https://cven5873-ss23.github.io/website/

live-01a-setup - Posit Cloud Setup

Follow along on the screen

1. Open the GitHub organisation for the course: https://github.com/cven5873-ss23
2. You will find a repository titled: **course-material-USERNAME** (with your GitHub Username)
3. You will “clone” this repository to Posit Cloud

Break

https://cven5873-ss23.github.io/website/

https://cven5873-ss23.github.io/website/

Data Science Lifecycle

Think, Pair, Share

Question

1. What is your mental model of the Data Science Lifecycle?

- **Think** for 2 minutes
- **Pair** with your neighbour for 4 minutes
- **Share** your answer with the class

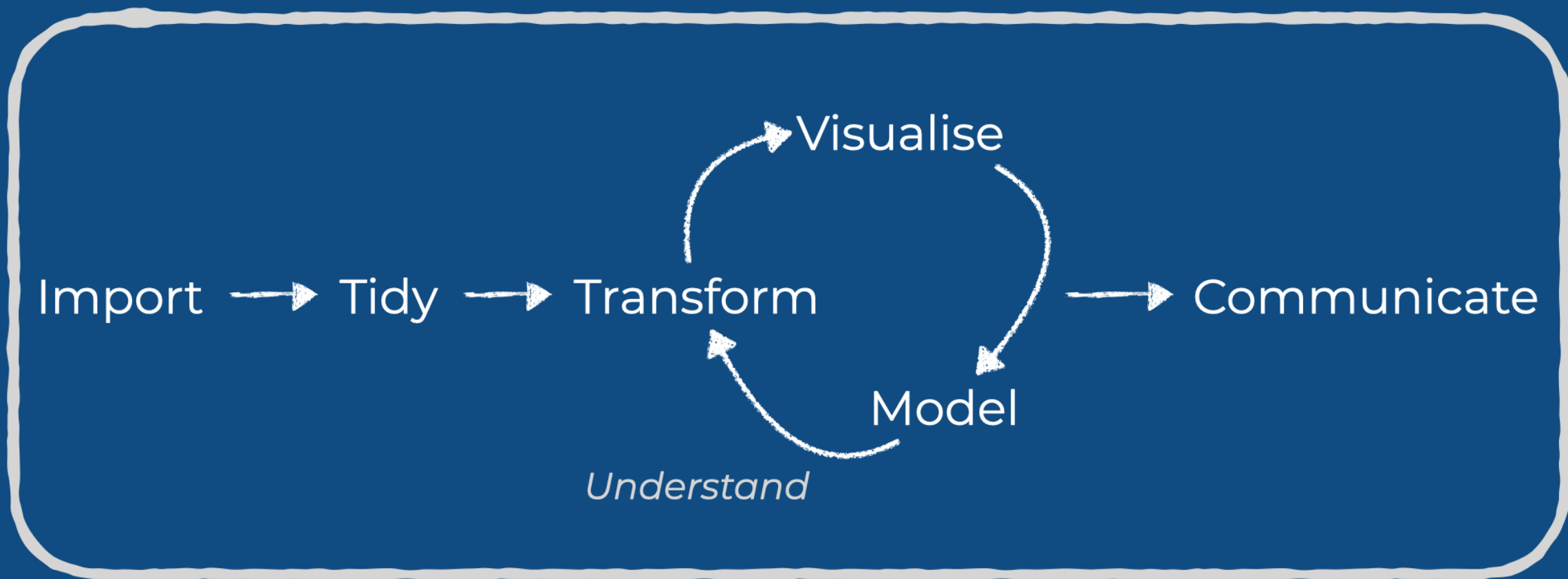
02:00

https://cven5873-ss23.github.io/website/

Deep End

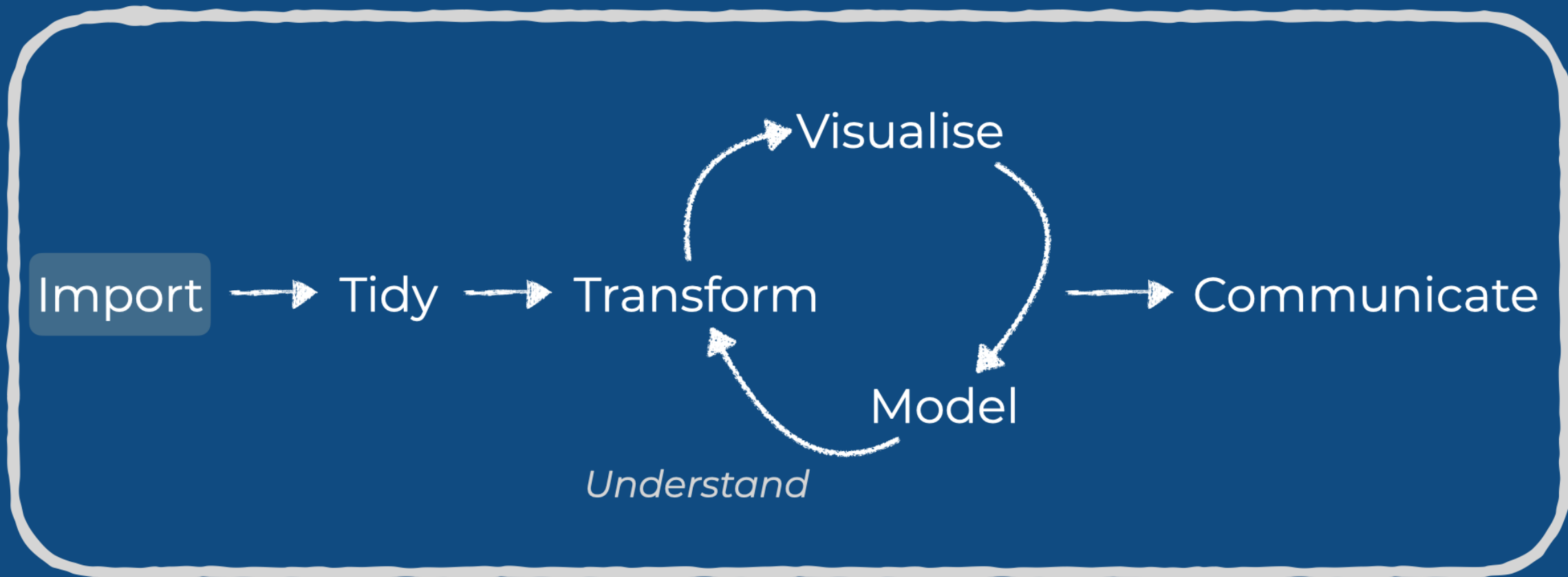
via GIPHY

Data Science Lifecycle



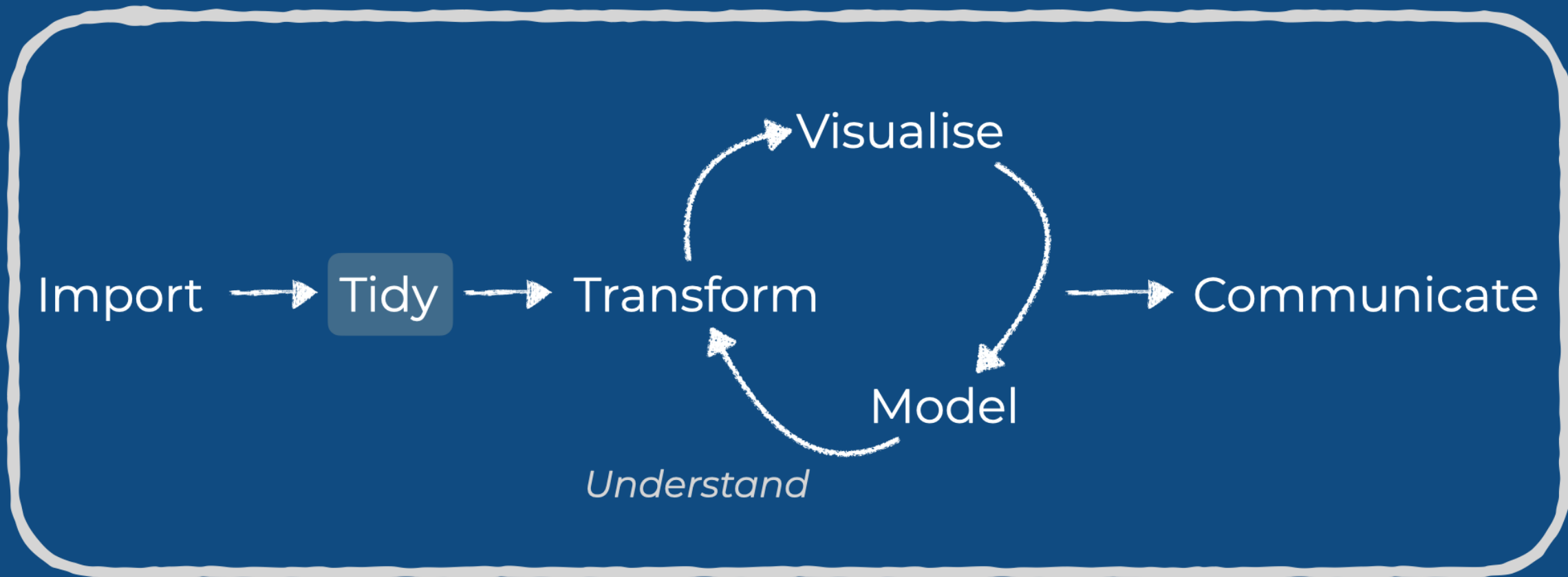
Data Science Lifecycle

Get your data into R



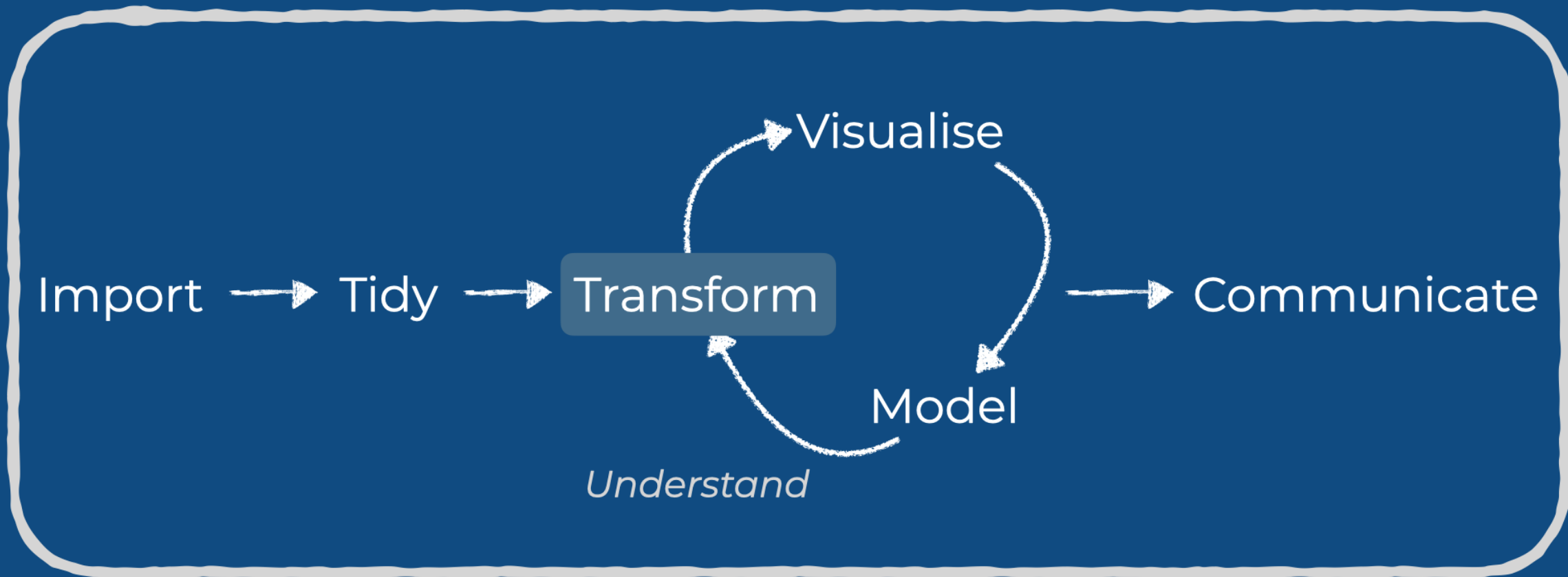
Data Science Lifecycle

Store your data in a consistent form



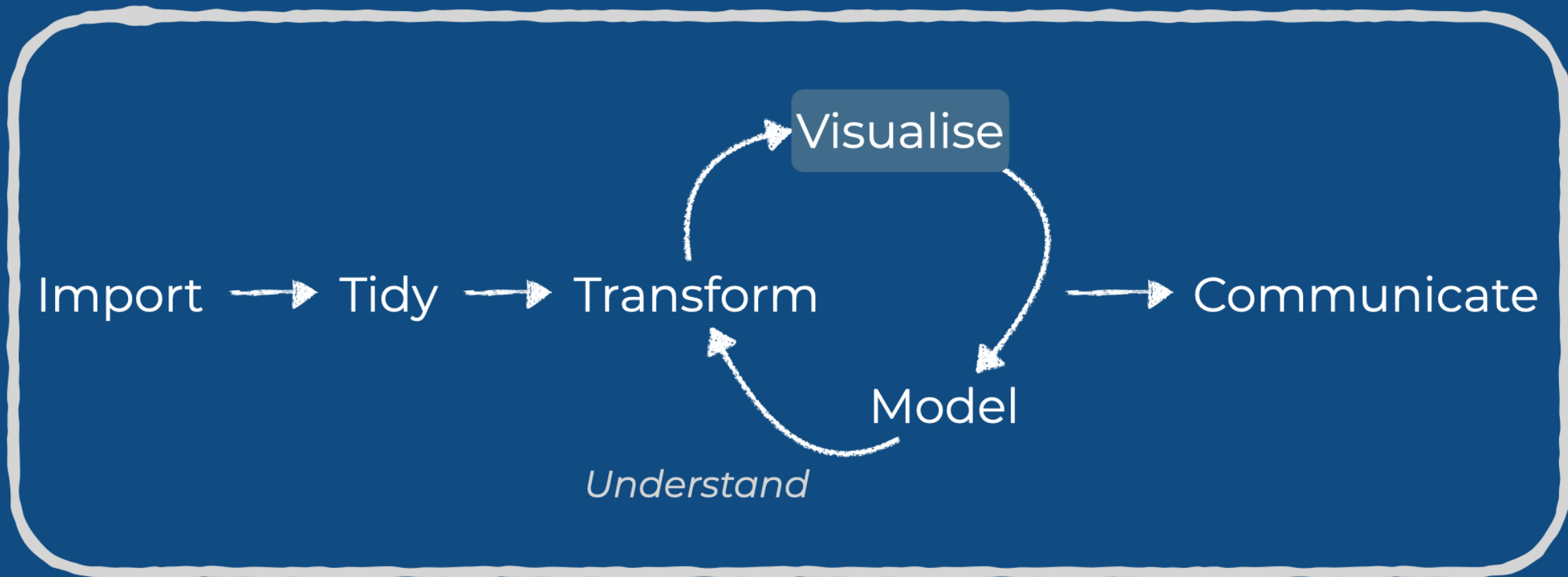
Data Science Lifecycle

Narrow down + Create new variables + Summary stats



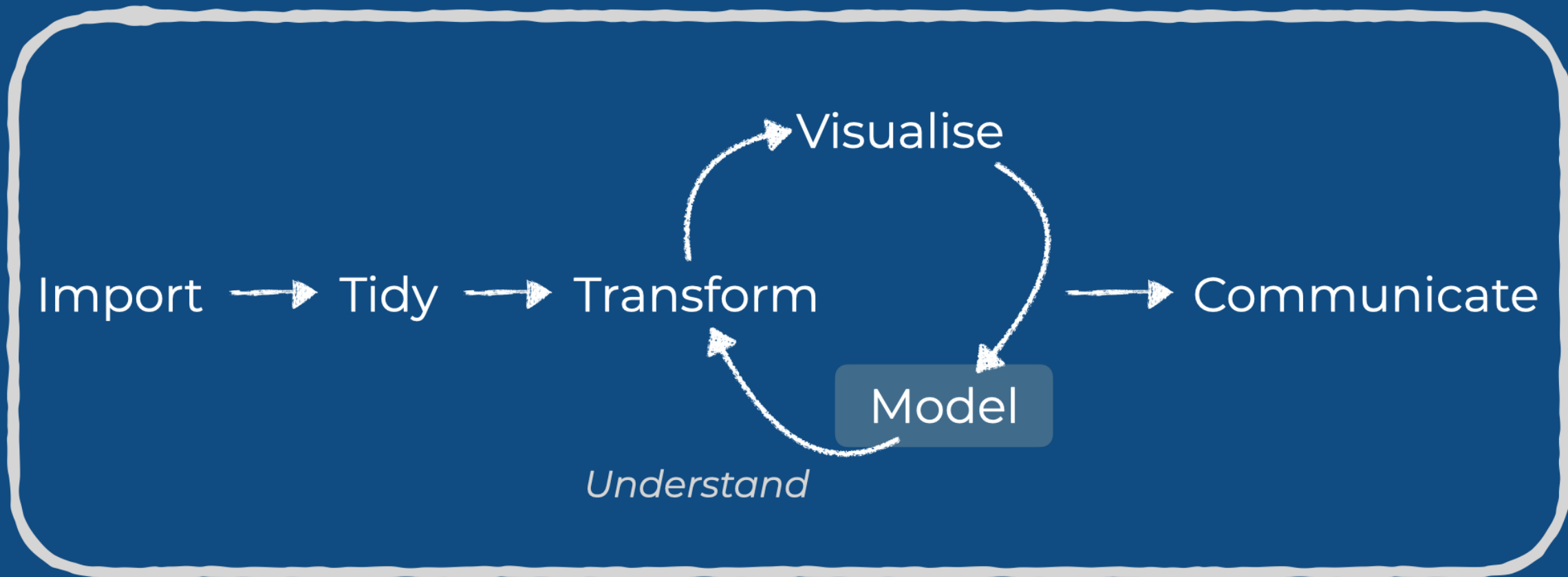
Data Science Lifecycle

Explore your with visual representations



Data Science Lifecycle

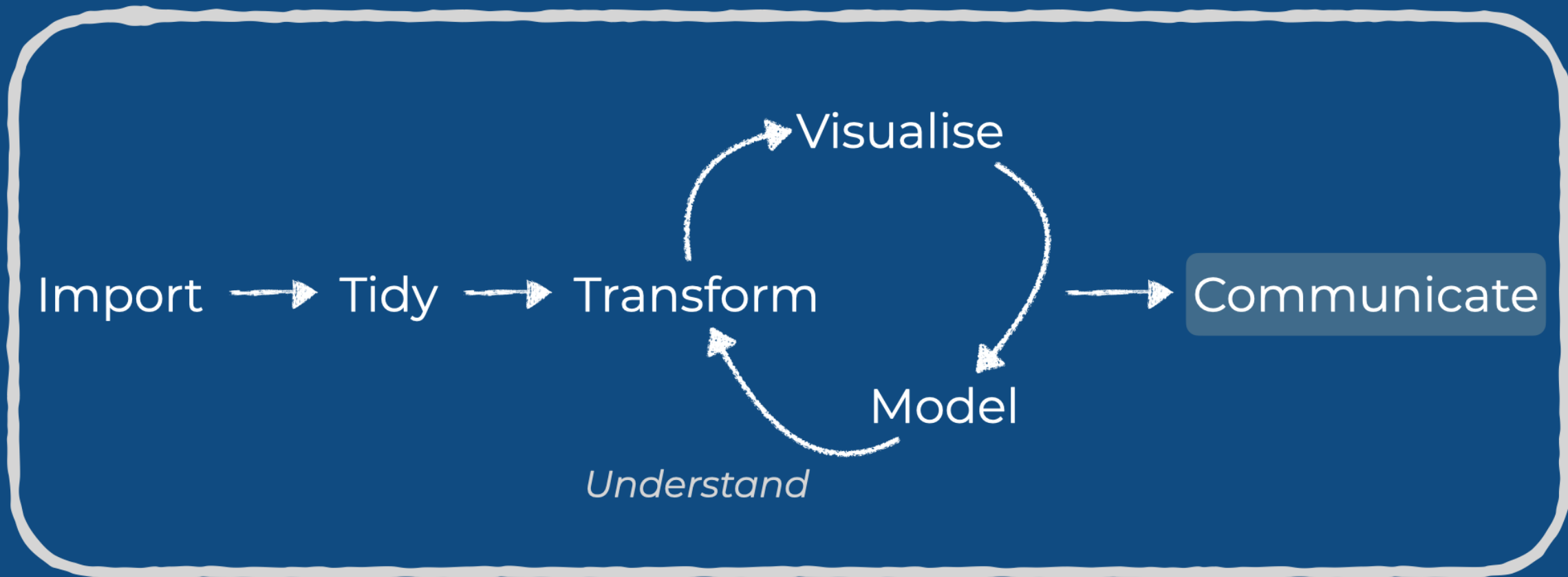
Explore your with visual representations



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

Data Science Lifecycle

Share your findings with others



Live Coding Exercise

https://cven5873-ss23.github.io/website/

live-01b-data-science-lifecycle - Data Science Lifecycle

1. Head over to posit.cloud
2. Open the workspace for the course (cven5837-ss23)
3. Open “Projects”
4. Open the “course-materials-USERNAME” project
5. Follow along with me

Break

https://cven5873-ss23.github.io/website/

https://cven5873-ss23.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

https://cven5873-ss23.github.io/website/

Weekly Structure

Monday	Lecture
Tuesday	
Wednesday	Feedback (grading) on assignments from previous week
Thursday	Student hours on Zoom (10 am to 12 pm CEST)
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 20% of the total grade

Exam

- 2-hour exam assessing the technical skills taught during the course
- Graded as number of points out of 100 pts for pre-defined graded elements
- Submitted as rendered Quarto document, but **not** on GitHub
- weighted at 20% 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

https://cven5873-ss23.github.io/website/

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 1

https://cven5873-ss23.github.io/website/

Homework due dates

- All material on course website
- Homework assignment & learning reflection due: **Friday, June 16th**

Thanks! 

https://cven5873-ss23.github.io/website/

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/ Access slides as PDF on GitHub