Welcome to Data Analytics for Development

CVEN 5837 - Summer 2022

Lars Schöbitz

https://cven5837-ss22.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER

Welcome! 👋

https://cven5837-ss22.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
- Twitter: @larnsce

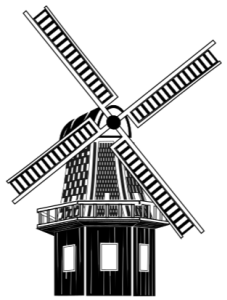
https://cven5837-ss22.github.io/website/

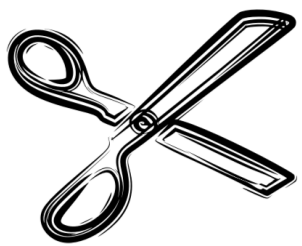
https://cven5837-ss22.github.io/website/

Learning Goals (for the course)

- Be familiar with the most commonly used qualitative and quantitative data collection methods and tools.
- Be able to employ remote sensing and in-situ data, and analysis tools to illustrate the utility of solutions for water, agriculture, disaster forecasting and relief, air quality, and global health.

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 that are used to for the course
2. Learners can render a file on RStudio Cloud in PDF 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://cven5837-ss22.github.io/website/

Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)

Classroom tools

https://cven5837-ss22.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

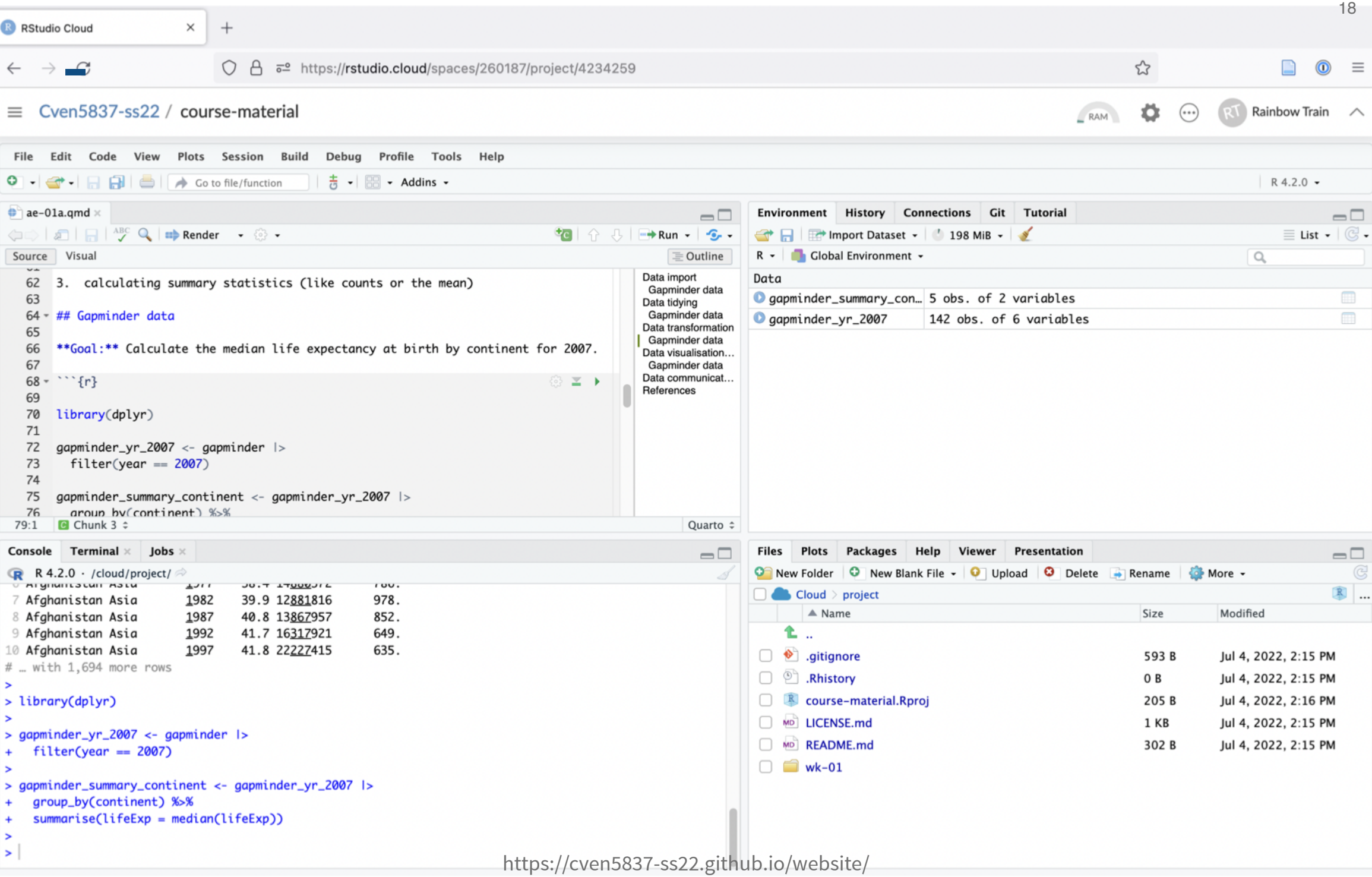
- Two learners work together in a break out session
- One person (the driver) shares the screen and does the typing
- The other person (the navigator) offers comments and suggestions
- Roles get switched

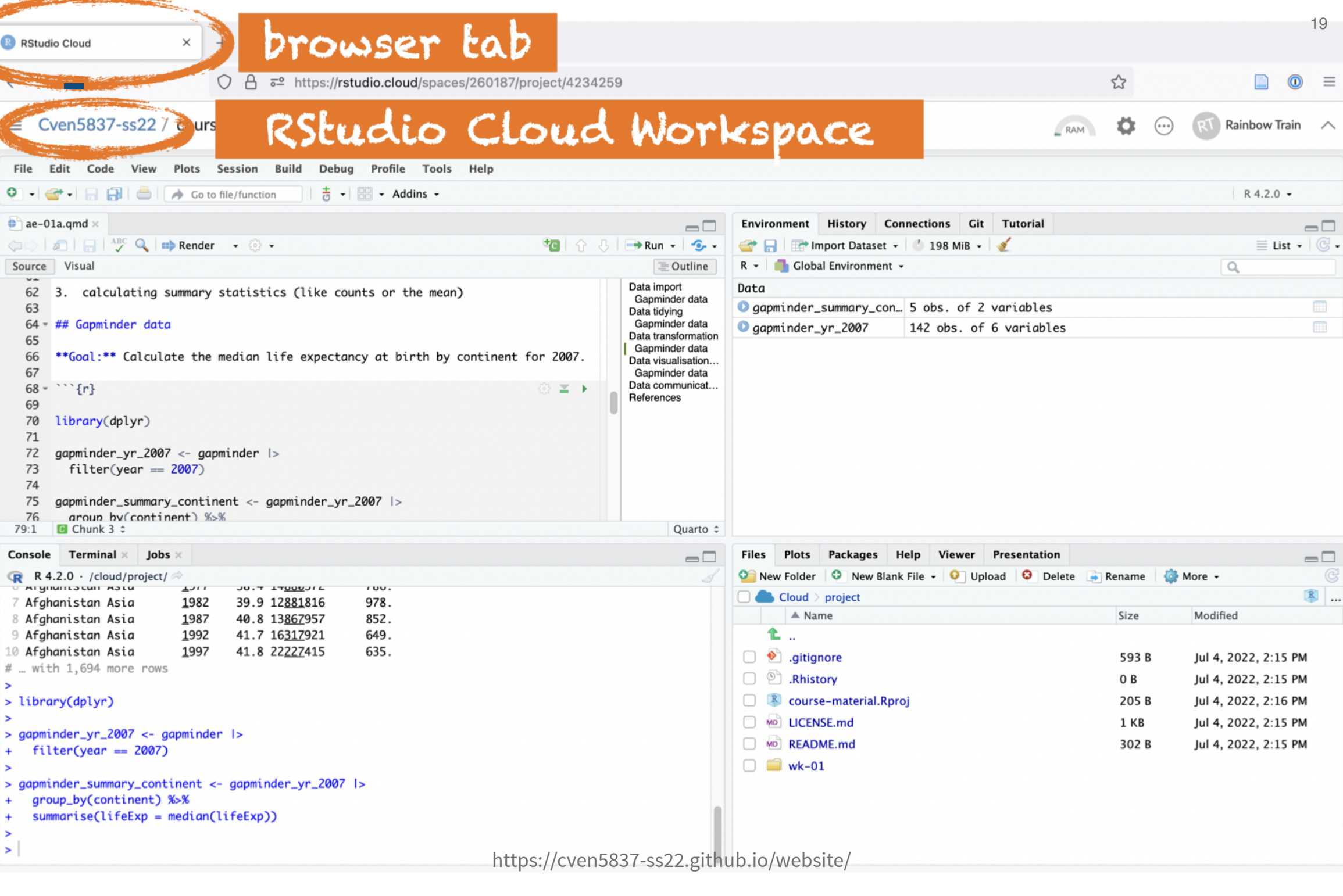
Platforms and Tools

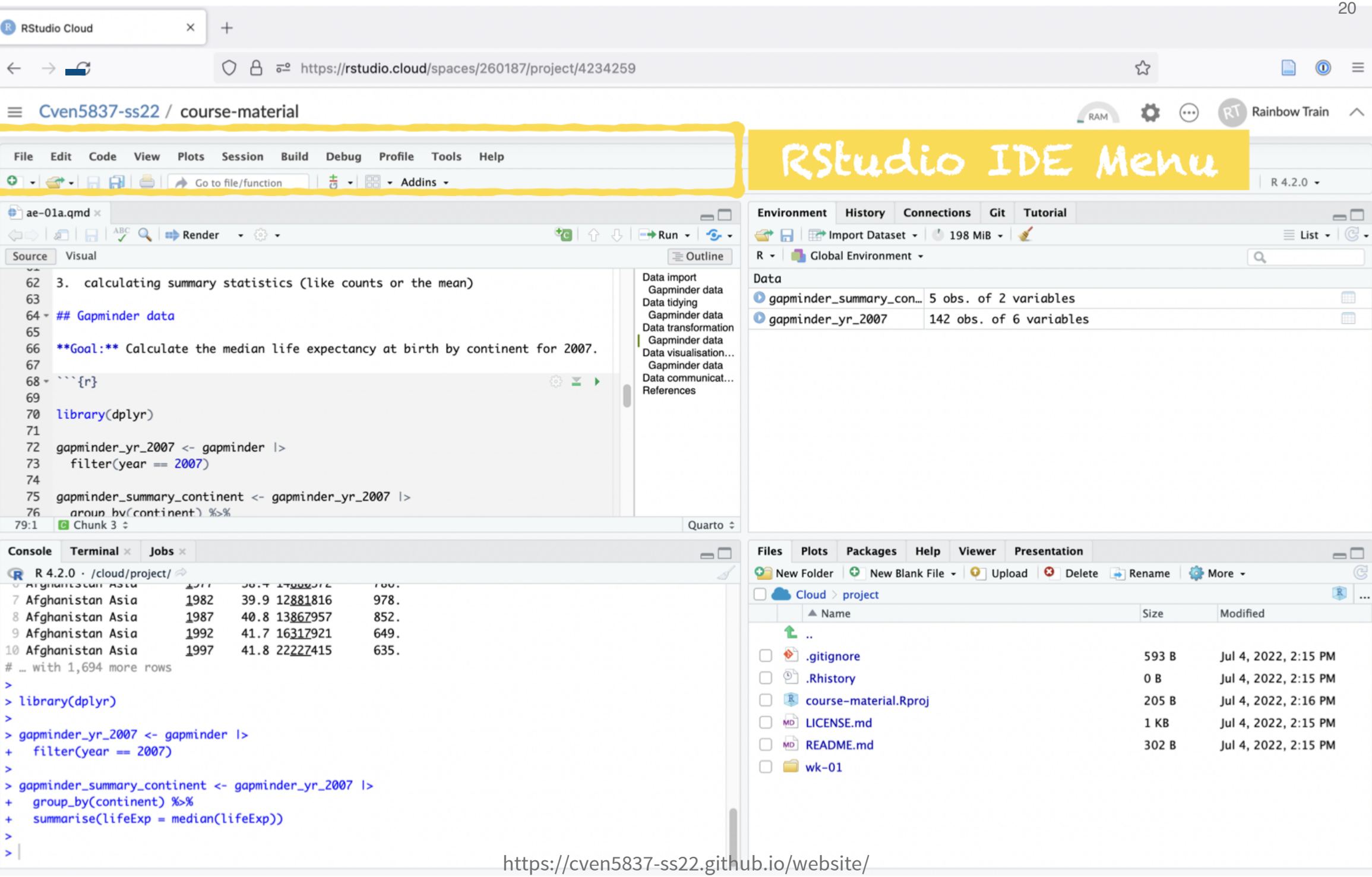
- R
- RStudio (Cloud)
- tidyverse R Packages
- Quarto publishing system

cven5837-ss22.github.io/website/ 

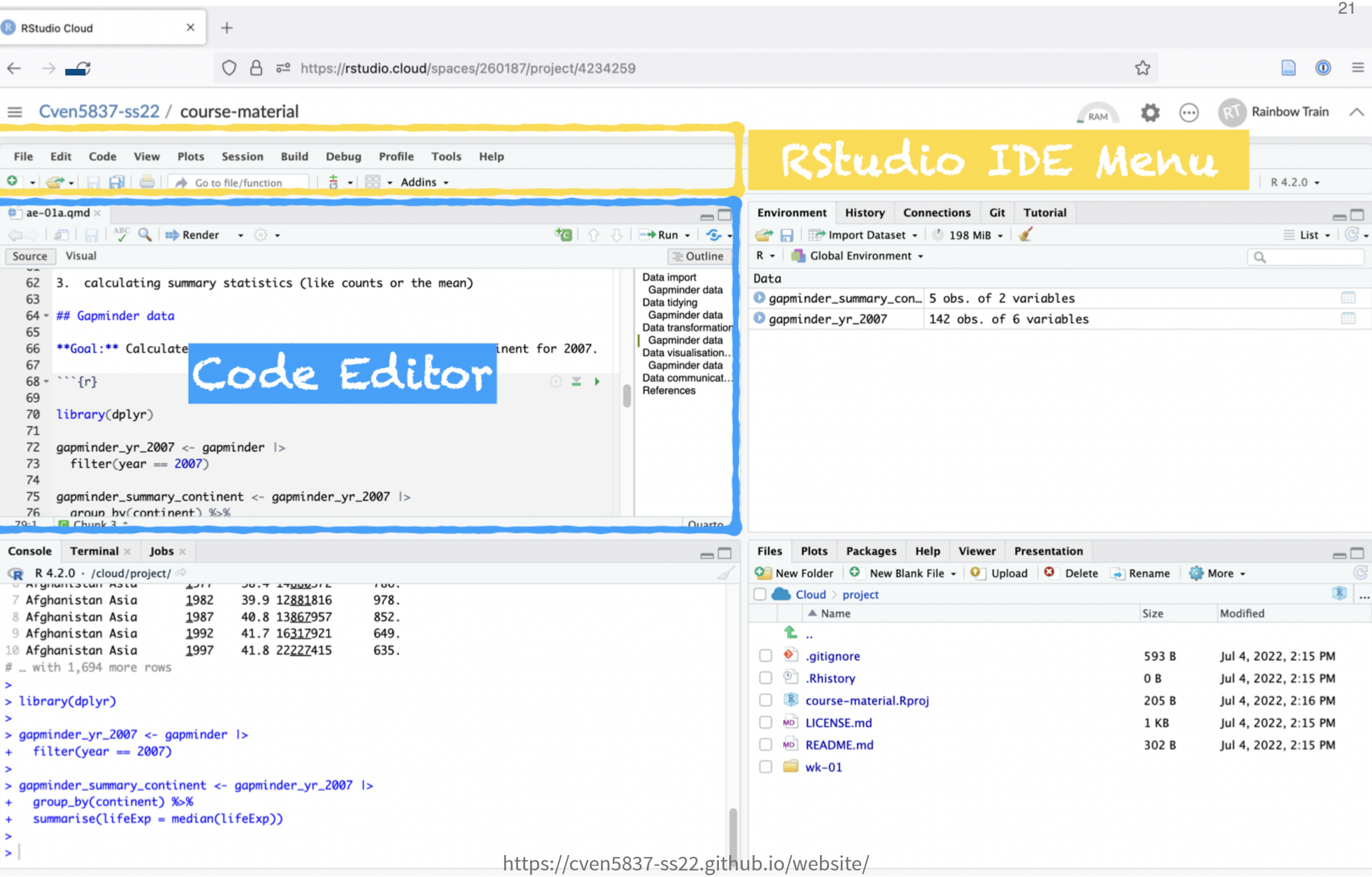
RStudio Cloud


RStudio Cloud
https://rstudio.cloud/spaces/260187/project/4234259
Cven5837-ss22 / course-material
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.0
ae-01a.qmd
Render
Run
Source Visual
Outline
3. calculating summary statistics (like counts or the mean)
Gapminder data
Goal: Calculate the median life expectancy at birth by continent for 2007.
```{r}
library(dplyr)
gapminder_yr_2007 <- gapminder |>
filter(year == 2007)
gapminder_summary_continent <- gapminder_yr_2007 |>
group_by(continent) %>%
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
References
79:1
Chunk 3
Quarto
Environment History Connections Git Tutorial
Import Dataset
198 MiB
List
R
Global Environment
Data
gapminder_summary_con... 5 obs. of 2 variables
gapminder_yr_2007 142 obs. of 6 variables
Console Terminal Jobs
R 4.2.0 · /cloud/project/
7 Afghanistan Asia 1982 39.9 12881816 978.
8 Afghanistan Asia 1987 40.8 13867957 852.
9 Afghanistan Asia 1992 41.7 16317921 649.
10 Afghanistan Asia 1997 41.8 22227415 635.
# … with 1,694 more rows
> library(dplyr)
> gapminder_yr_2007 <- gapminder |>
+ filter(year == 2007)
> gapminder_summary_continent <- gapminder_yr_2007 |>
+ group_by(continent) %>%
+ summarise(lifeExp = median(lifeExp))
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud project
Name Size Modified
.. 
.gitignore 593 B Jul 4, 2022, 2:15 PM
.Rhistory 0 B Jul 4, 2022, 2:15 PM
course-material.Rproj 205 B Jul 4, 2022, 2:16 PM
LICENSE.md 1 KB Jul 4, 2022, 2:15 PM
README.md 302 B Jul 4, 2022, 2:15 PM
wk-01
```

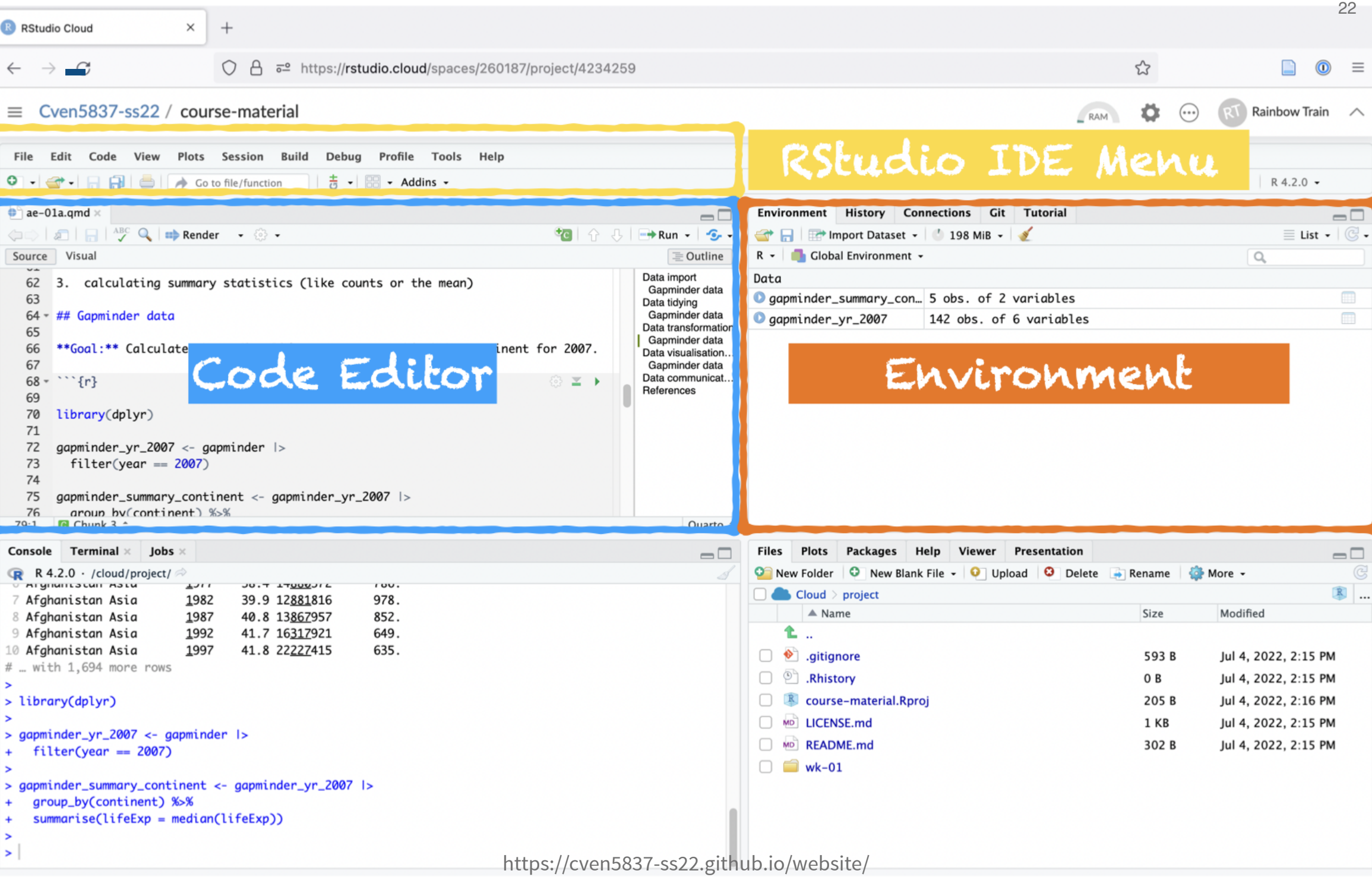

RStudio Cloud
browser tab
https://rstudio.cloud/spaces/260187/project/4234259
Cven5837-ss22
RStudio Cloud Workspace
Rainbow Train
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.0
ae-01a.qmd
Render
Run
Source Visual
Outline
62 3. calculating summary statistics (like counts or the mean)
64 ## Gapminder data
66 **Goal:** Calculate the median life expectancy at birth by continent for 2007.
68 ```{r}
70 library(dplyr)
72 gapminder_yr_2007 <- gapminder |>
73 filter(year == 2007)
75 gapminder_summary_continent <- gapminder_yr_2007 |>
76 group_by(continent) %>%
Data import
Gapminder data
Data tidying
Gapminder data
Data transformation
Gapminder data
Data visualisation...
Gapminder data
Data communicat...
References
79:1 Chunk 3
Quarto
Environment History Connections Git Tutorial
Import Dataset
198 MiB
List
R
Global Environment
Data
gapminder_summary_con... 5 obs. of 2 variables
gapminder_yr_2007 142 obs. of 6 variables
Console Terminal Jobs
R 4.2.0 · /cloud/project/
7 Afghanistan Asia 1982 39.9 12881816 978.
8 Afghanistan Asia 1987 40.8 13867957 852.
9 Afghanistan Asia 1992 41.7 16317921 649.
10 Afghanistan Asia 1997 41.8 22227415 635.
… with 1,694 more rows
> library(dplyr)
> gapminder_yr_2007 <- gapminder |>
+ filter(year == 2007)
> gapminder_summary_continent <- gapminder_yr_2007 |>
+ group_by(continent) %>%
+ summarise(lifeExp = median(lifeExp))
Files Plots Packages Help Viewer Presentation
New Folder
New Blank File
Upload
Delete
Rename
More
Cloud project
Name Size Modified
..
.gitignore 593 B Jul 4, 2022, 2:15 PM
.Rhistory 0 B Jul 4, 2022, 2:15 PM
course-material.Rproj 205 B Jul 4, 2022, 2:16 PM
LICENSE.md 1 KB Jul 4, 2022, 2:15 PM
README.md 302 B Jul 4, 2022, 2:15 PM
wk-01



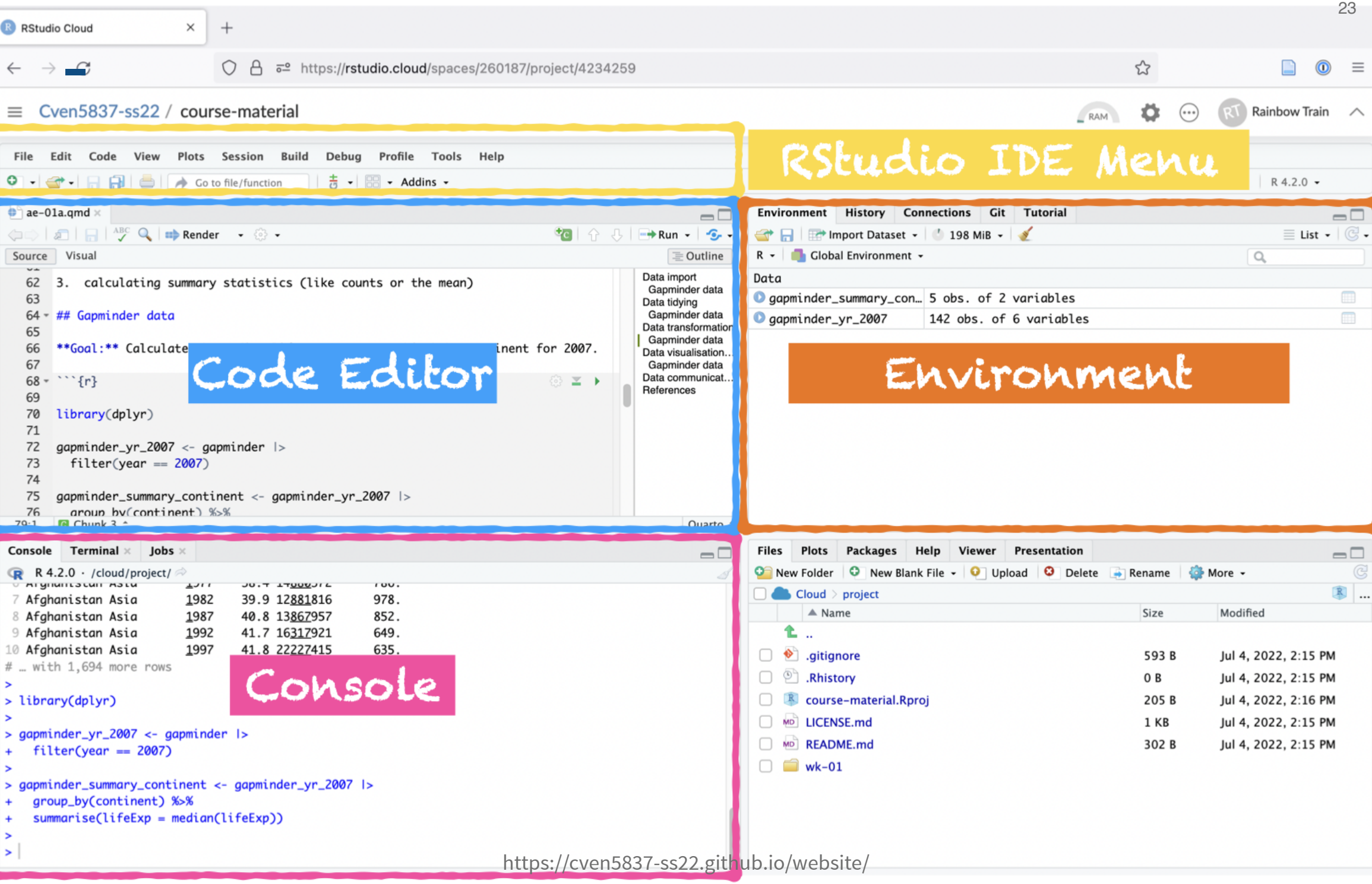
https://cven5837-ss22.github.io/website/



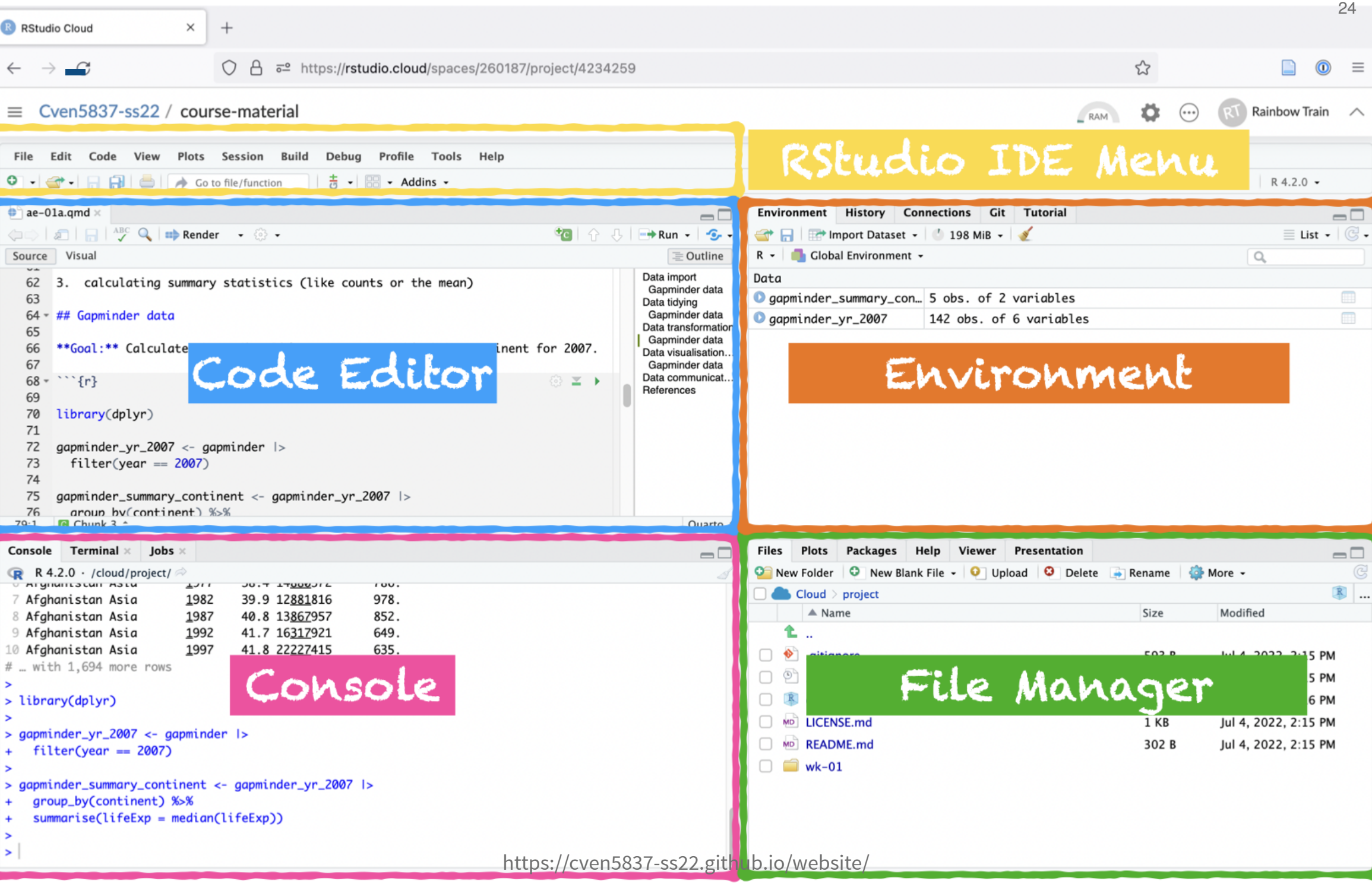
https://cven5837-ss22.github.io/website/



https://cven5837-ss22.github.io/website/



https://cven5837-ss22.github.io/website/



https://cven5837-ss22.github.io/website/

Screen setup

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

Live Coding Exercise

https://cven5837-ss22.github.io/website/

live-01a-setup - RStudio Cloud Setup

1. Head over to rstudio.cloud
2. Create a free account if you do not have one yet
3. Open the link that is posted to the Zoom chat
4. Accept the invitation to join the cven5837-ss22 workspace
5. Post “ready” to the Zoom chat when you are done

Break

10:00

https://cven5837-ss22.github.io/website/

Photo by Blake Wisz

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

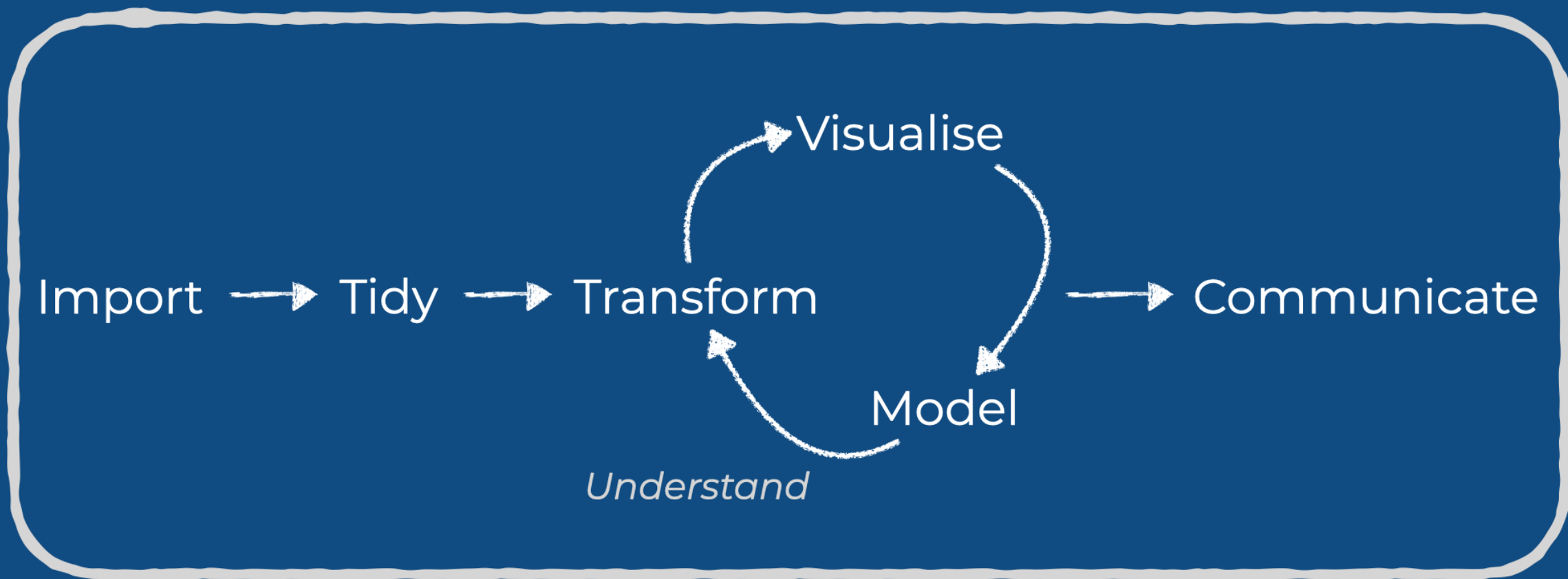
https://cven5837-ss22.github.io/website/

Deep End

via GIPHY

https://cven5837-ss22.github.io/website/

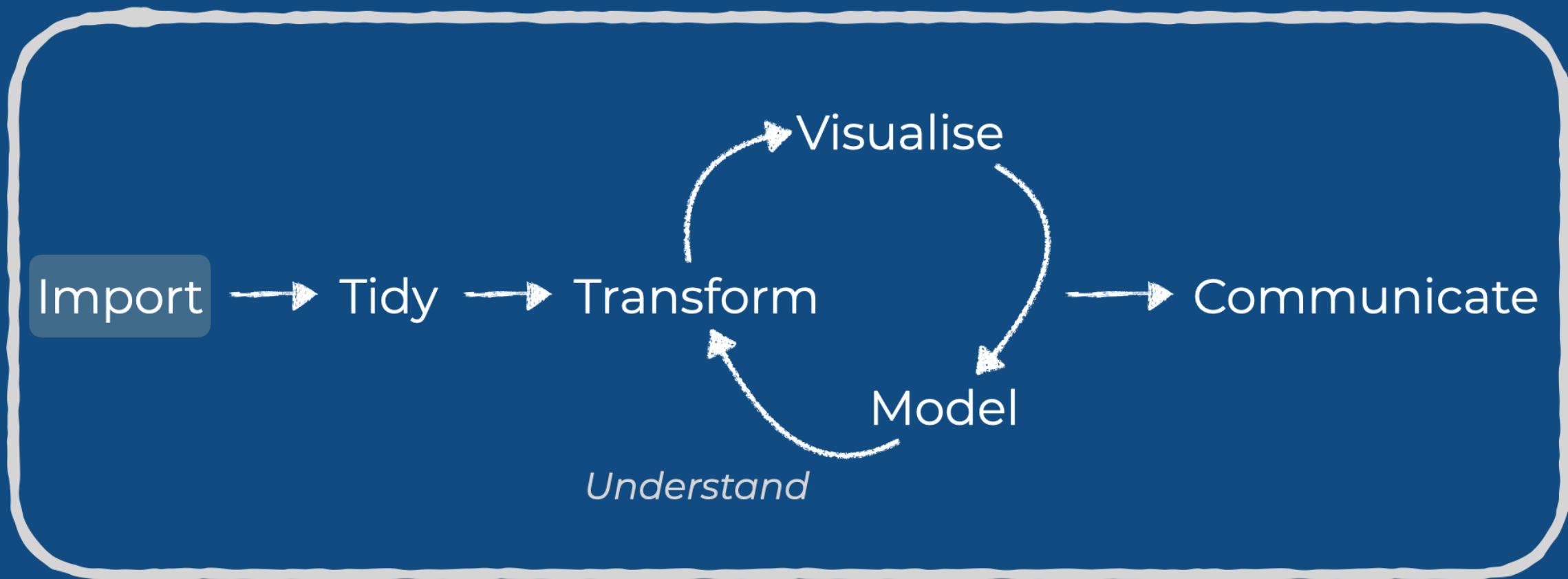
Data Science Lifecycle



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

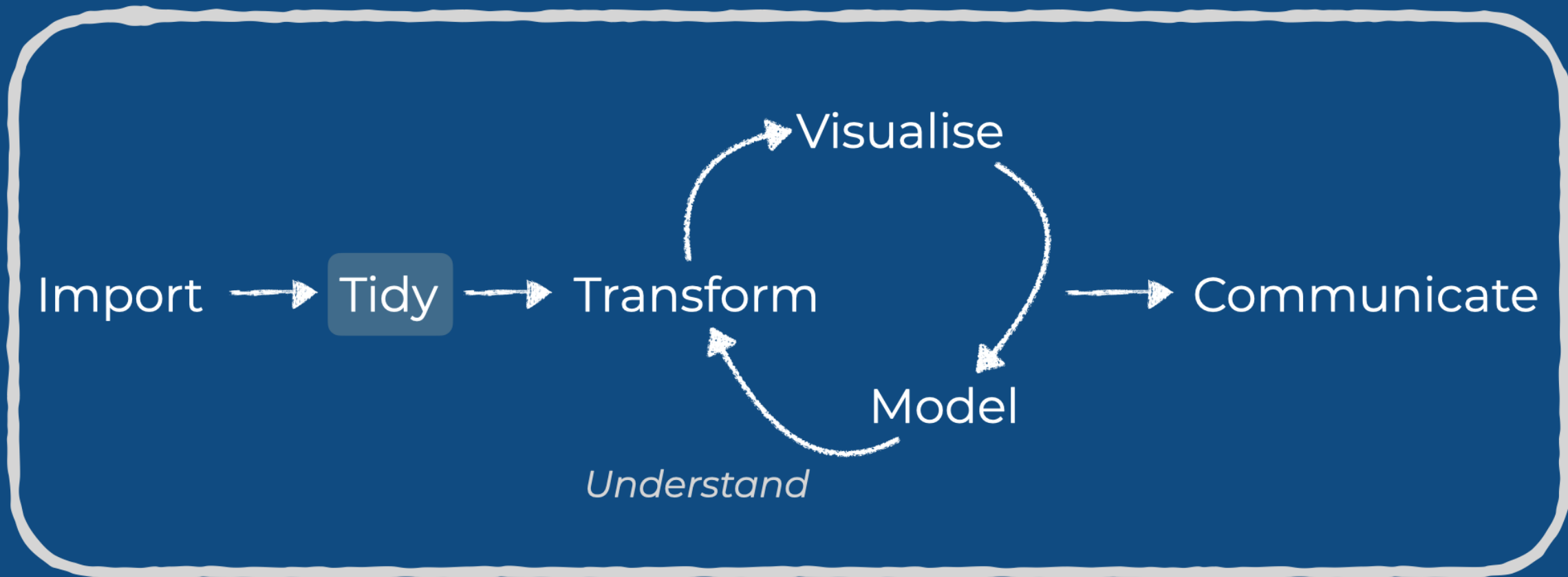
Data Science Lifecycle

Get your data into R



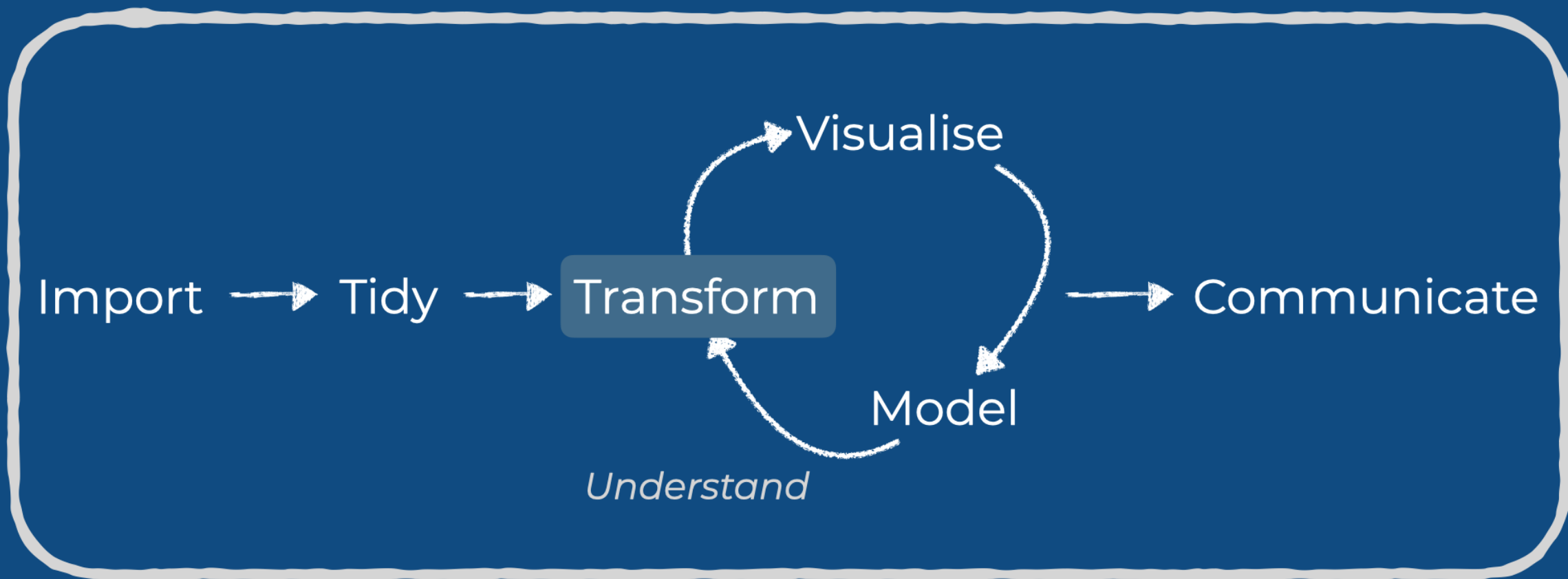
Data Science Lifecycle

Store your data in a consistent form



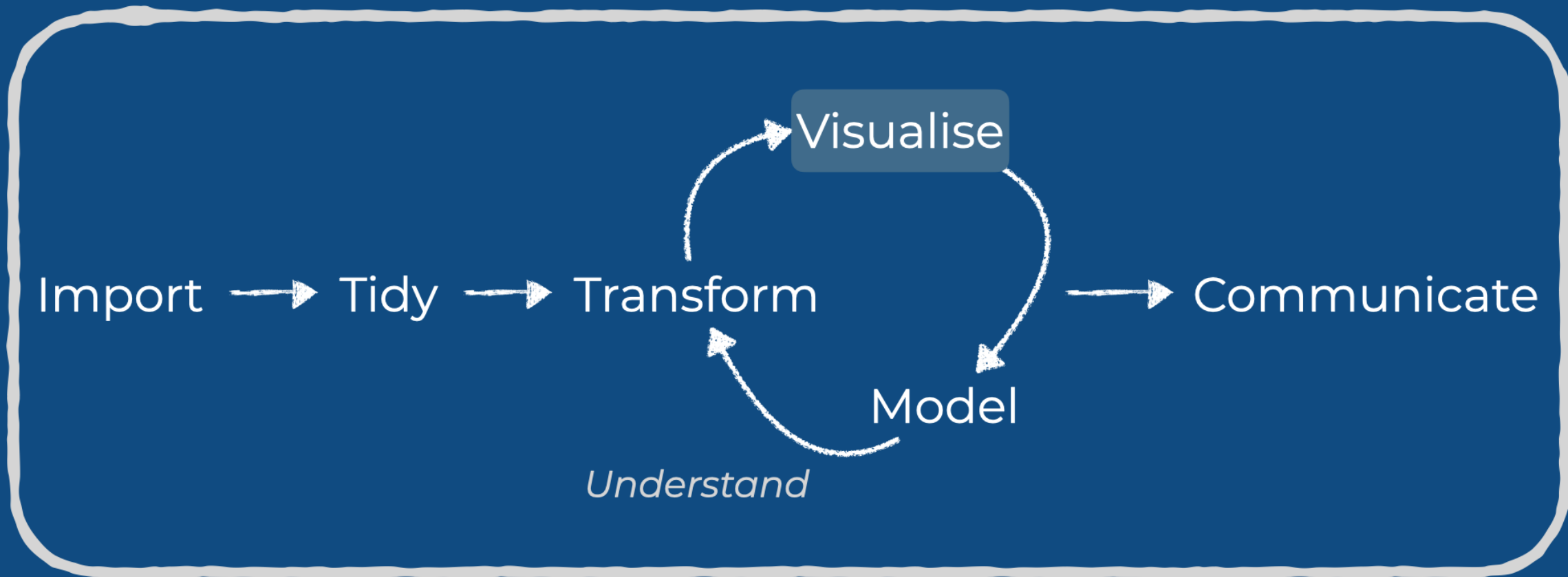
Data Science Lifecycle

Narrow down + Create new variables + Summary stats



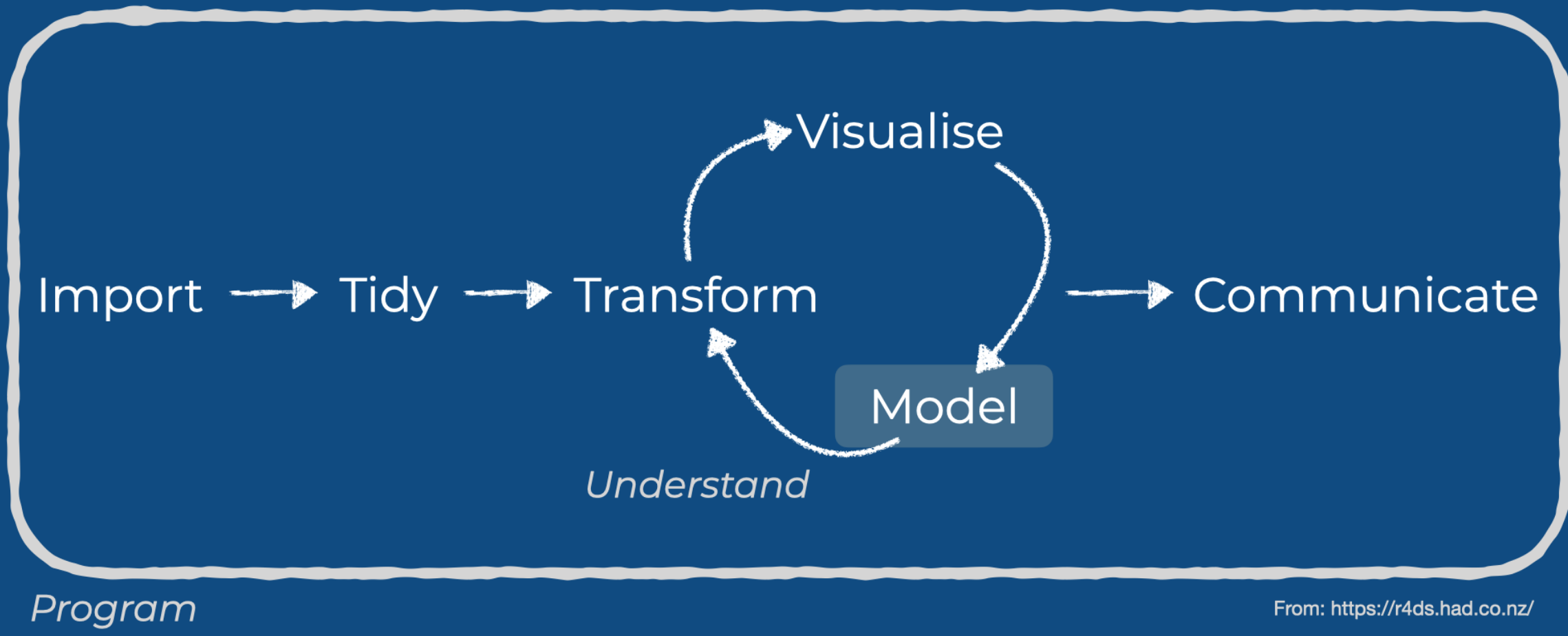
Data Science Lifecycle

Explore your with visual representations



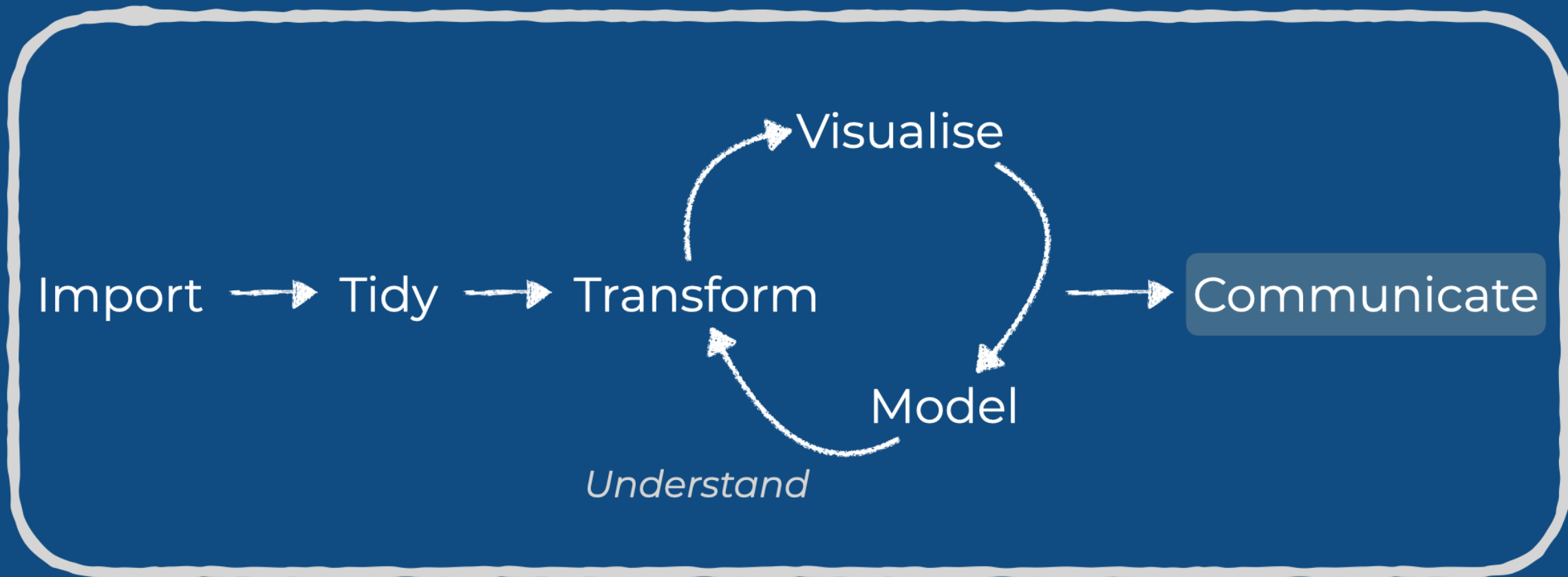
Data Science Lifecycle

Explore your with visual representations



Data Science Lifecycle

Share your findings with others



Live Coding Exercise

https://cven5837-ss22.github.io/website/

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

1. Head over to rstudio.cloud
2. Open the workspace for the course (cven5837-ss22)
3. Open “Projects”
4. Open the “course-materials” project
5. Follow along with me

Break

05:00

https://cven5837-ss22.github.io/website/

Photo by Blake Wisz

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://cven5837-ss22.github.io/website/

Weekly Structure

Monday	Learning reflections are due
Tuesday	Lecture
Wednesday	Feedback (grading) on assignment
Thursday	Student hours on Zoom (10 am to 12 pm CEST)
Friday	Homework assignment is due

Homework assignments

- Weekly programming assignments
- 75% of the total grade

Learning reflections

- Reflections on the different class elements (lecture, homework assignment, readings)
- minimum 200 words
- 25% of the total grade

Grading

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://cven5837-ss22.github.io/website/

Late work policy

- up to 2 working days after deadline (25% penalty for each day)
  - Tuesday for homework assignments (-50%)
  - Wednesday for learning reflections (-50%)
- work handed in more than two working days after due date with be graded 0%

Homework week 1

https://cven5837-ss22.github.io/website/

Homework due dates

- All material on course website
- Homework assignment due: Friday, 8th July
- Learning reflection due: Monday, 11th July

Thanks! 

https://cven5837-ss22.github.io/website/

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